Tour I Power: Pre-data and Post-data

> A salutary effect of power analysis is that it draws one forcibly to consider the magnitude of effects. In psychology, and especially in soft psychology, under the sway of the Fisherian scheme, there has been little consciousness of how big things are. (Cohen 1990, p. 1309)

So how would you use power to consider the magnitude of effects were you drawn forcibly to do so? In with your breakfast is an exercise to get us started on today's shore excursion.

> Suppose you are reading about a statistically significant result $\boldsymbol{x}$ (just at level α) from a one-sided test T+ of the mean of a Normal distribution with n IID samples, and known σ: H_0: $\mu \leq 0$ against H_1: $\mu > 0$.
>
> Underline the correct word, from the perspective of the (error statistical) philosophy, within which power is defined.
>
> - If the test's power to detect μ' is very low (i.e., POW(μ') is low), then the statistically significant $\boldsymbol{x}$ is poor/good evidence that $\mu > \mu'$.
> - Were POW(μ') reasonably high, the inference to $\mu > \mu'$ is reasonably/poorly warranted.

We've covered this reasoning in earlier travels (e.g., Section 4.3), but I want to launch our new tour from the power perspective. Assume the statistical test has passed an audit (for selection effects and underlying statistical assumptions) – you can't begin to analyze the logic if the premises are violated.

During our three tours on Power Peninsula, a partially uncharted territory, we'll be residing at local inns, not returning to the ship, so pack for overnights. We'll visit its museum, but mostly meet with different tribal members who talk about power – often critically. Power is one of the most abused notions in all of statistics, yet it's a favorite for those of us who care about magnitudes of discrepancies. Power is always defined in terms of a fixed cut-off, c_α, computed under a value of the parameter under test; since these vary, there is really a *power function*. If someone speaks of the power of a test *tout court*, you cannot make sense of it, without qualification. First defined in Section 3.1, the *power* of a test against μ' is the probability it would lead to rejecting H_0 when $\mu = \mu'$:

$$\text{POW(T, } \mu') = \Pr(\text{d}(\boldsymbol{X}) \geq c_\alpha; \mu = \mu'), \text{ or } \Pr(\text{test T rejects } H_0; \mu = \mu').$$

If it's clear what the test is, we just write POW(μ'). Power measures the capability of a test to detect μ' – where the detection is in the form of producing a $d \geq c_\alpha$. While power is computed at a point $\mu = \mu'$, we employ it to appraise claims of form $\mu > \mu'$ or $\mu < \mu'$.

Power is an ingredient in N-P tests, but even practitioners who declare they never set foot into N-P territory, but live only in the land of Fisherian significance tests, invoke power. This is all to the good, and they shouldn't fear that they are dabbling in an inconsistent hybrid.

Jacob Cohen's (1988) *Statistical Power Analysis for the Behavioral Sciences* is displayed at the Power Museum's permanent exhibition. Oddly, he makes some slips in the book's opening. On page 1 Cohen says: "The power of a statistical test is the probability it will yield statistically significant results." Also faulty is what he says on page 4: "The power of a statistical test of a null hypothesis is the probability that it will lead to the rejection of the null hypothesis, i.e., the probability that it will result in the conclusion that the phenomenon exists." Cohen means to add "computed under an alternative hypothesis," else the definitions are wrong. These snafus do not take away from Cohen's important tome on power analysis, yet I can't help wondering if these initial definitions play a bit of a role in the tendency to define power as 'the probability of a correct rejection,' which slips into erroneously viewing it as a posterior probability (unless qualified).

Although keeping to the fixed cut-off c_α is too coarse for the severe tester's tastes, it is important to keep to the given definition for understanding the statistical battles. We've already had sneak previews of "achieved sensitivity" or "attained power" [$\Pi(\gamma) = \Pr(\text{d}(\boldsymbol{X}) \geq \text{d}(\boldsymbol{x}_0); \mu_0 + \gamma)$] by which members of Fisherian tribes are able to reason about discrepancies (Section 3.3). N-P accorded three roles to power: the first two are pre-data, for planning and comparing tests; the third is for interpretation post-data. It's the third that they don't announce very loudly, whereas that will be our main emphasis. Have a look at this museum label referring to a semi-famous passage by E. Pearson. Barnard (1950, p. 207) has just suggested that error probabilities of tests, like power, while fine for pre-data planning, should be replaced by other measures (likelihoods perhaps?) after the trial. What did Egon say in reply to George?

> [I]f the planning is based on the consequences that will result from following a rule of statistical procedure, e.g., is based on a study of the power function of a test and then,

having obtained our results, we do not follow the first rule but another, based on likelihoods, what is the meaning of the planning? (Pearson 1950, p. 228)

This is an interesting and, dare I say, powerful reply, but it doesn't quite answer George. By all means apply the rule you planned to, but there's still a legitimate question as to the relationship between the pre-data capability or performance measure, and post-data inference. The severe tester offers a view of this intimate relationship. In Tour II we'll be looking at interactive exhibits far outside the museum, including N-P post-data power analysis, retrospective power, and a notion I call shpower. Employing our understanding of power, scrutinizing a popular reinterpretation of tests as diagnostic tools will be straightforward. In Tour III we go a few levels deeper in disinterring the N-P vs. Fisher feuds. I suspect there is a correlation between those who took Fisher's side in the early disputes with Neyman and those leery of power. Oscar Kempthorne being interviewed by J. Leroy Folks (1995) said:

Well, a common thing said about [Fisher] was that he did not accept the idea of the power. But, of course, he must have. However, because Neyman had made such a point about power, Fisher couldn't bring himself to acknowledge it (p. 331).

However, since Fisherian tribe members have no problem with corresponding uses of sensitivity, *P*-value distributions, or CIs, they can come along on a severity analysis. There's more than one way to skin a cat, if one understands the relevant statistical principles. The issues surrounding power are subtle, and unraveling them will require great care, so bear with me. I will give you a money-back guarantee that by the end of the excursion you'll have a whole new view of power. Did I mention you'll have a chance to power the ship into port on this tour? Only kidding, however, you will get to show your stuff in a Cruise Severity Drill (Section 5.2).

5.1 Power Howlers, Trade-offs, and Benchmarks

In the Mountains out of Molehills (MM) Fallacy (Section 4.3), a rejection of H_0 just at level α with a larger sample size (higher power) is taken as evidence of a greater discrepancy from H_0 than with a smaller sample size (in tests otherwise the same). Power can be increased by increasing sample size, but also by computing it in relation to alternatives further and further from H_0. Some are careful to avoid the MM fallacy when the high power is due to large n, but then fall right into it when it is due to considering a very discrepant μ'. For our purposes, our one-sided T+ will do.

Mountains out of Molehills (MM) Fallacy (second form). Test T+: The fallacy of taking a just statistically significant difference at level α (i.e., $d(\boldsymbol{x}_0) = d_\alpha$) as a better indication of a discrepancy μ' if the POW (μ') is high, than if POW(μ') is low.

Two Points Stephen Senn Correctly Dubs Nonsense and Ludicrous

Start with an extreme example: Suppose someone is testing H_0: the drug cures no one. An alternative H_1 is it cures nearly everyone. Clearly these are not the only possibilities. Say the test is practically guaranteed to reject H_0, if in fact H_1, the drug cures practically everyone. The test has high power to detect H_1. You wouldn't say that its rejecting H_0 is evidence of H_1. H_1 entails it's very probable you'll reject H_0; but rejecting H_0 doesn't warrant H_1. To think otherwise is to allow problematic statistical affirming the consequent – the basis for the MM fallacy (Section 2.1). This obvious point lets you zero in on some confusions about power.

Stephen Senn's contributions to statistical foundations are once again spot on. In drug development, it is typical to require a high power of 0.8 or 0.9 to detect effects deemed of clinical relevance. The clinically relevant discrepancy, as Senn sees it, is the discrepancy "one should not like to miss" (2007, p. 196). Senn labels this delta Δ. He is considering a difference between means, so the null hypothesis is typically 0. We'll apply severity to his example in Exhibit (iv) of this tour. Here the same points will be made with respect to our one-sided Normal test T+: H_0: $\mu \leq \mu_0$ vs. H_1: $\mu > \mu_0$, letting $\mu_0 = 0$, σ known. We may view Δ as the value of μ of clinical relevance. (Nothing changes in this discussion if it's estimated as s.) The test takes the form

Reject H_0 iff $Z \geq z_\alpha$ (Z is the standard Normal variate).

"Reject H_0" is the shorthand for "infer a statistically significant difference" at the level of the test. Though Z is the test statistic, it makes for a simpler presentation to use the cut-off for rejection in terms of $\overline{x}_\alpha$: Reject H_0 iff $\overline{X} \geq \overline{x}_\alpha = (\mu_0 + z_\alpha \sigma/\sqrt{n})$.

Let's abbreviate the alternative against which test T+ has 0.8 power by $\mu^{.8}$, when it's clear what test we're talking about. So POW($\mu^{.8}$) = 0.8, and let's suppose $\mu^{.8}$ is the clinically relevant difference Δ. Senn asks, what does $\mu^{.8}$ mean in relation to *what we are entitled to infer* when we obtain statistical significance? Can we say, upon rejecting the null hypothesis, that the treatment has a clinically relevant effect, i.e., $\mu \geq \mu^{.8}$ (or $\mu > \mu^{.8}$)?

"This is a surprisingly widespread piece of nonsense which has even made its way into one book on drug industry trials" (ibid., p. 201). The reason it is

nonsense, Senn explains, is that $\mu^{.8}$ must be in excess of the cut-off for rejection, in particular, $\mu^{.8} = \overline{x}_\alpha + 0.85\ \sigma_{\overline{X}}$ (where $\sigma_{\overline{X}} = \sigma/\sqrt{n}$). We know we are only entitled to infer μ exceeds the *lower bound* of the confidence interval at a reasonable level; whereas, $\mu^{.8}$ is actually the upper bound of a 0.8 (one-sided) confidence interval, formed having observed $\overline{x} = \overline{x}_\alpha$. All we are to infer, officially, from just reaching the cut-off $\overline{x}_\alpha$, is that $\mu > 0$.

Granted, as Senn admits, the test "lacks ambition" (ibid., p. 202), but with more data and with results surpassing the minimal cut-off, we may uncover a clinically relevant discrepancy. Why not just set up the test to enable the clinically relevant discrepancy to be inferred whenever the null is rejected?

$$H_0\colon \mu \leq \Delta \text{ vs. } H_1\colon \mu > \Delta.$$

This requires redefining Δ. "It is no longer 'the difference we should not like to miss' but instead becomes 'the difference we should like to prove obtains'" (ibid.). Some call this the "clinically irrelevant difference" (ibid.). But then we can't also have high power to detect $H_1\colon \mu > \Delta$.

> [I]f the true treatment difference is Δ, then the observed treatment difference will be less than Δ in approximately 50% of all trials. Therefore, the probability that it is less than the critical value must be greater than 50%. (ibid., p. 202)

Indeed, it will be approximately $1 - \alpha$. So the power – the probability the observed difference exceeds the critical value under H_1 – is, in this case, around α. The researcher is free to specify the null as $H_0\colon \mu \leq \Delta$, but Senn argues against doing so, at least in drug testing, because "a nonsignificant result will often mean the end of the road for a treatment. It will be lost forever. However, a treatment which shows a 'significant' effect will be studied further" (ibid.). This goes beyond issues of interest now. The point is: Δ cannot be the value in H_0 and also the value against which we want 0.8 power to detect, i.e., $\mu^{.8}$.

If testing $H_0\colon \mu \leq 0$ vs. $H_1\colon \mu > 0$, then a just α-significant result is poor evidence for $\mu \geq \mu^{.8}$ (or other alternative with high power). To think it's good evidence is *nonsense*. Senn's related point is that it is *ludicrous* to assume the effect is either 0 or a clinically relevant difference, as if we are testing

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

"But where we are unsure whether a drug works or not, it would be ludicrous to maintain that it cannot have an effect which, while greater than nothing, is less than the clinically relevant difference" (ibid., p. 201). That is, it is ludicrous to cut out everything in between 0 and Δ. By the same token, it would seem odd to give a 0.5 prior probability to H_0, and the remaining 0.5 to H_1. We will have plenty of occasions to return to Senn's points about what's nonsensical and ludicrous.

Trade-offs and Benchmarks

Between H_0 and $\overline{x}_\alpha$ the power goes from α to 0.5. Keeping to our simple test T+ will amply reward us here.

a. *The power against H_0 is α.* We can use the power function to define the probability of a Type I error or the significance level of the test:

$$\text{POW}(\text{T+}, \mu_0) = \Pr(\overline{X} \geq \overline{x}_\alpha; \mu_0), \text{ where } \overline{x}_\alpha = \mu_0 + z_\alpha \sigma_{\overline{X}}$$

Standardizing $\overline{X}$, we get $Z = [(\mu_0 + z_\alpha \sigma_{\overline{X}}) - \mu_0]/\sigma_{\overline{X}}$.

The power at the null is $\Pr(Z \geq z_\alpha; \mu_0) = \alpha$.

It's the *low power* against H_0 that warrants taking a rejection as evidence that $\mu > \mu_0$. This is desirable: we infer an indication of discrepancy from H_0 because a null world would probably have resulted in a smaller difference than we observed.

b. *The power of T+ for $\mu_1 = \overline{x}_\alpha$ is 0.5.* In that case, $Z = 0$, and $\Pr(Z \geq 0) = 0.5$, so

$$\text{POW}(\text{T+}, \mu_1 = \overline{x}_\alpha) = 0.5.$$

The power only gets to be greater than 0.5 for alternatives that exceed the cut-off $\overline{x}_\alpha$, whatever it is. As noted, $\mu^{.8} = \overline{x}_\alpha + 0.85\,\sigma_{\overline{X}}$ since $\text{POW}(\text{T+}, \overline{x}_\alpha + 0.85\sigma_{\overline{X}}) = 0.8$. Tests ensuring 0.9 power are also often of interest: $\mu^{.9} = \overline{x}_\alpha + 1.28\sigma_{\overline{X}}$. We get these shortcuts:

Case 1: POW(T+, μ) for μ between H_0 and $\mu = \overline{x}_\alpha$:
If $\mu_1 = \overline{x}_\alpha - k\sigma_{\overline{X}}$ then POW(T+, μ_1) = area to the right of k under N(0,1) ($\leq$ 0.5).

Case 2: POW(T+, μ) for μ greater than $\overline{x}_\alpha$:
If $\mu_1 = \overline{x}_\alpha + k\sigma_{\overline{X}}$ then POW(T+, μ_1) = area to the right of $-k$ under N (0,1) ($>$ 0.5).
Remember $\overline{x}_\alpha$ is $\mu_0 + z_\alpha\sigma_{\overline{X}}$.

Trade-offs Between the Type I and Type II Error Probability

We know that, for a given test, as the probability of a Type I error goes down the probability of a Type II error goes up (and power goes down). And as the probability of a Type II error goes down (and power goes up), the probability of a Type I error goes up, assuming we leave everything else the same. There's a trade-off between the two error probabilities. (No free lunch.) So if someone said: As the power increases, the probability of a Type I error *decreases*, they'd be saying, as the Type II error

decreases, the probability of a Type I error decreases. That's the opposite of a trade-off! You'd know automatically they had made a mistake or were simply defining things in a way that differs from standard N-P statistical tests. Now you may say, "I don't care about Type I and II errors, I'm interested in inferring estimated effect sizes." I too want to infer magnitudes. But those will be ready to hand once we tell what's true about the existing concepts.

While $\mu^{.8}$ is obtained by adding 0.85 $\sigma_{\overline{X}}$ to $\overline{x}_{\alpha}$, in day-to-day rounding, if you're like me, you're more likely to remember the result of adding $1\sigma_{\overline{X}}$ to $\overline{x}_{\alpha}$. That takes us to a value of μ against which the test has 0.84 power, $\mu^{.84}$:

> The power of test T+ to detect an alternative that exceeds the cut-off $\overline{x}_{\alpha}$ by $1\sigma_{\overline{X}} = 0.84$.

In test T+ the range of possible values of $\overline{X}$ and μ are the same, so we are able to set μ values this way, without confusing the parameter and sample spaces.

Exhibit (i). Let test T+ ($\alpha = 0.025$) be H_0: $\mu = 0$ vs. H_1: $\mu \geq 0$, $\alpha = 0.025$, $n = 25$, $\sigma = 1$. Using the $2\sigma_{\overline{X}}$ cut-off: $\overline{x}_{0.025} = 2(1)/\sqrt{25} = 0.4$ (using 1.96 it's 3.92). Suppose you are instructed to decrease the Type I error probability α to 0.001 but it's impossible to get more samples. This requires the hurdle for rejection to be higher than in our original test. The new cut-off for test T+ will be $\overline{x}_{0.001}$. It must be $3\,\sigma_{\overline{X}}$ greater than 0 rather than only $2\sigma_{\overline{X}}$: $\overline{x}_{0.001} = 0 + 3(1)/\sqrt{25} = 0.6$. We decrease α (the Type I error probability) from 0.025 to 0.001 by moving the hurdle over to the right by $1\sigma_{\overline{X}}$ unit. But we've just made the power lower for any discrepancy or alternative. For what value of μ does this new test have 0.84 power?

$$\text{POW(T+, } \alpha = 0.001, \mu^{.84} = ?) = 0.84.$$

We know: $\mu^{.84} = 0.6 + (0.2) = 0.8$. So, POW(T+, $\alpha = 0.001$, $\mu = 0.8$) = 0.84. Decreasing the Type I error by moving the hurdle over to the right by $1\sigma_{\overline{X}}$ unit results in the alternative against which we have 0.84 power $\mu^{.84}$ also moving over to the right by $1\sigma_{\overline{X}}$ (Figure 5.1). We see the trade-off very neatly, at least in one direction.

Consider the discrepancy of $\mu = 0.6$ (Figure 5.2). The power to detect 0.6 in test T+ ($\alpha = 0.001$) is now only 0.5! In test T+ ($\alpha = 0.025$) it is 0.84. Test T+ ($\alpha = 0.001$) is less powerful than T+ ($\alpha = 0.025$).

Should you hear someone say that the higher the power, the higher the *hurdle* for rejection, you'd know they are confused or using terms in an

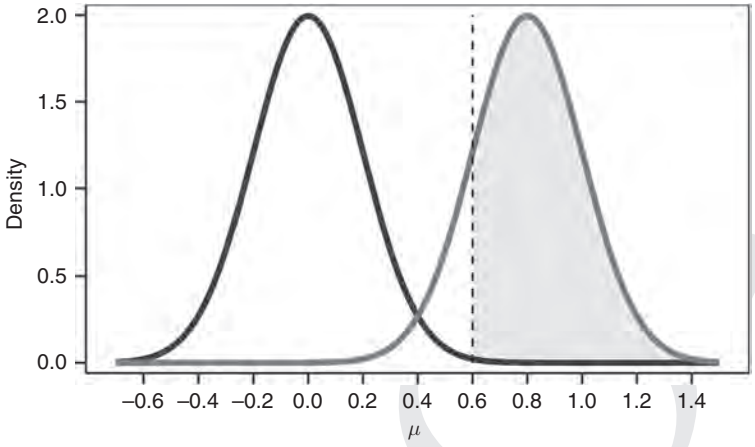

Figure 5.1 POW(T+, $\alpha = 0.001$, $\mu = 0.8$) = 0.84.

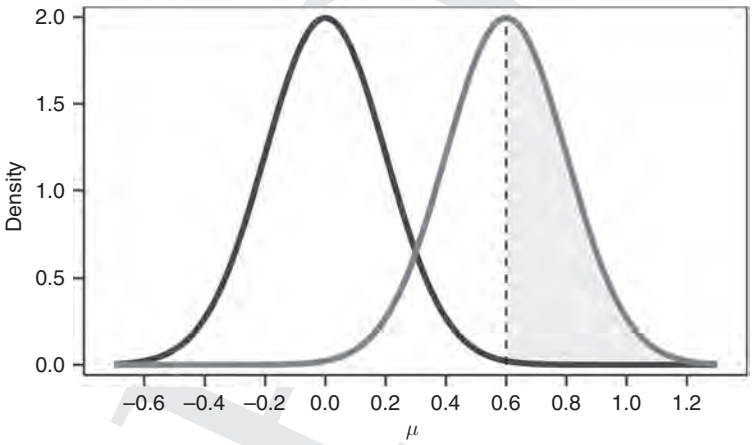

Figure 5.2 POW(T+, $\alpha = 0.001$, $\mu = 0.6$) = 0.5.

incorrect way. (The hurdle is how large the cut-off must be before rejecting at the given level.) Why then do Ziliak and McCloskey, popular critics of significance tests, announce: "refutations of the null are trivially easy to achieve if power is low enough or the sample is large enough" (2008a, p. 152)? Increasing sample size means increased power, so the second disjunct is correct. The first disjunct is not. One might be tempted to suppose they mean "power is high

enough," but one would be mistaken. They mean what they wrote. Aris Spanos (2008a) points this out (in a review of their book), and I can't figure out why they dismiss such corrections as "a lot of technical smoke" (2008b, p. 166).

Ziliak and McCloskey Get Their Hurdles in a Twist

Still, their slippery slides are quite illuminating.

> If the power of a test is low, say, 0.33, then the scientist will two times in three accept the null and mistakenly conclude that another hypothesis is false. If on the other hand the power of a test is high, say, 0.85 or higher, then the scientist can be reasonably confident that at minimum the null hypothesis (of, again, zero effect if that is the null chosen) is false and that therefore his rejection of it is highly probably correct. (Ziliak and McCloskey 2008a, p. 132–3)

With a wink and a nod, the first sentence isn't too bad, even though, at the very least, it is mandatory to specify a particular "another hypothesis," μ'. But what about the statement: if the power of a test is high, then a rejection of the null is probably correct?

We follow our rule of generous interpretation to try to see it as true. Let's allow the ";" in the first premise to be a conditional probability "|", using $\mu^{0.84}$:

1. Pr(Test T+ rejects the null | $\mu^{.84}$) = 0.84.
2. Test T+ rejects the null hypothesis.

   Therefore, the rejection is correct with probability 0.84.

Oops. The premises are true, but the conclusion fallaciously transposes premise 1 to obtain conditional probability Pr($\mu^{.84}$ | test T+ rejects the null) = 0.84. What I think they want to say, or at any rate what would be correct, is

$$\Pr(\text{Test T+ } \textit{does not} \text{ reject the null hypothesis} \mid \mu^{.84}) = 0.16.$$

So the Type II error probability is 0.16. Looking at it this way, the flaw is in computing the complement of premise 1 by transposing (as we saw in the Higgs example, Section 3.8). Let's be clear about significance levels and hurdles. According to Ziliak and McCloskey:

> It is the history of Fisher significance testing. One erects little "significance" hurdles, six inches tall, and makes a great show of leaping over them, . . . If a test does a good job of uncovering efficacy, then the test has high power and the hurdles are high not low. (ibid., p. 133)

They construe "little significance" as little hurdles! It explains how they wound up supposing high power translates into high hurdles. It's the opposite.

The higher the hurdle, the more difficult it is to reject, and the lower the power. High hurdles correspond to insensitive tests, like insensitive fire alarms. It might be that using "sensitivity" rather than power would make this abundantly clear. We may coin: The high power = high hurdle (for rejection) fallacy. A powerful test does give the null hypothesis a harder time in the sense that it's more probable that discrepancies from it are detected. That makes it easier for H_1. Z & M have their hurdles in a twist.

5.2 Cruise Severity Drill: How Tail Areas (Appear to) Exaggerate the Evidence

The most influential criticisms of statistical significance tests rest on highly plausible intuitions, at least from the perspective of a probabilist. We are about to visit a wonderfully instructive example from Steven Goodman. It combines the central skills gathered up from our journey, but with a surprising twist. As always, it's a canonical criticism – not limited to Goodman. He happens to give a much clearer exposition than most, and, on top of that, is frank about his philosophical standpoint. Let's listen:

> To examine the inferential meaning of the p value, we need to review the concept of inductive evidence. An inductive measure assigns a number (a measure of support or credibility) to a hypothesis, given observed data. ... By this definition the p value is not an inductive measure of evidence, because it involves only one hypothesis and because it is based partially on unobserved data in the tail region.
>
> To assess the quantitative impact of these philosophical issues, we need to turn to an inductive statistical measure: mathematical likelihood. (Goodman 1993, p. 490)

Well that settles things quickly. Influenced by Royall, Goodman has just listed the keynotes from the standpoint of "evidence is comparative likelihood" seen as far back as Excursion 1 (Tour II). Like the critics we visited in Excursion 4 (Sections 4.4 and 4.5), Goodman finds that the P-value exaggerates the evidence against a null hypothesis because the likelihood ratio (or Bayes factor) in favor of a chosen alternative is not as large as the P-value would suggest. He admits that one's assessment here will turn on philosophy. On Goodman's philosophy, it's the use of the tail area that deceitfully blows up the evidence against the null hypothesis. Now in Section 3.4, Jeffreys' tail area criticism, we saw that considering the tails makes it harder, not easier, to find

evidence against a null. Goodman purports to show the opposite. That's the new twist.

Three Steps to the Argument

Goodman's context involves statistically significant results – he writes them as z values, as it is a case of Normal testing. We're not given the sample size or the precise test, but it won't matter for the key argument. He gives the two-sided α value, although "[w]e assume that we know the direction of the effect" (ibid., p. 491). I am not objecting. Even if we run a two-sided test, once we see the direction, it makes sense to look at the power of the relevant one-sided test, but double the α value. There are three steps. *First step*: form the likelihood ratio of the statistically significant outcome $z_{0.025}$ (i.e., 1.96) under the null hypothesis and some alternative (where Pr is the density):

$$\Pr(z_\alpha; H')/\Pr(z_\alpha; H_0).$$

But which alternative H'? Alternatives against which the test has high power, say 0.8, 0.84, or 0.9, are of interest, and he chooses $\mu^{.9}$. He writes this alternative as $\mu = \Delta_{0.05,\ 0.9}$, "the difference against which the hypothesis test has two-sided $\alpha = 0.05$ and one-sided $\beta = 0.10$ (power = 0.90)" (ibid., p. 496). We know from our benchmarks that $\mu^{.9}$ is approximately 1.28 standard errors from the one-sided cut-off: $(1.96 + 1.28)\sigma_{\overline{X}} = 3.24\sigma_{\overline{X}}$. The likelihood for alternative $\mu^{.9}$ is smaller than if one had used the maximum likely alternative (as in Johnson).[1]

I'll follow Goodman in computing the likelihood of the null over the alternative, although most of the authors we've considered do the reverse. Not that it matters so long as you keep them straight.

"Two likelihood ratios are compared, one for a 'precise' p value, e.g., $p = 0.03$, and one for $\ldots$ $p \leq \alpha = 0.03$. The alternative hypothesis used here is the one against which the hypothesis test has 90 percent power (two-sided $\alpha = 0.05$) $\ldots$" (ibid., pp. 490–1). The precise and imprecise P-values correspond to reporting $z = 1.96$ and $z \geq 1.96$, respectively.[2] The likelihood ratio for the precise P-value "corresponds to the ratio of heights of the two probability densities" (ibid., p. 491): Number this as (1):

$$(1)\ \Pr(Z = 1.96; \mu_0)/\Pr(Z = 1.96; \mu^{.9}) = 0.058/0.176 = 0.33.$$

[1] We saw this in Section 4.5. Goodman also shows the results for the maximum likely alternative.
[2] He writes the two descriptions as $p = \alpha$ vs. $p > \alpha$, but I think it's clearer using corresponding z values.

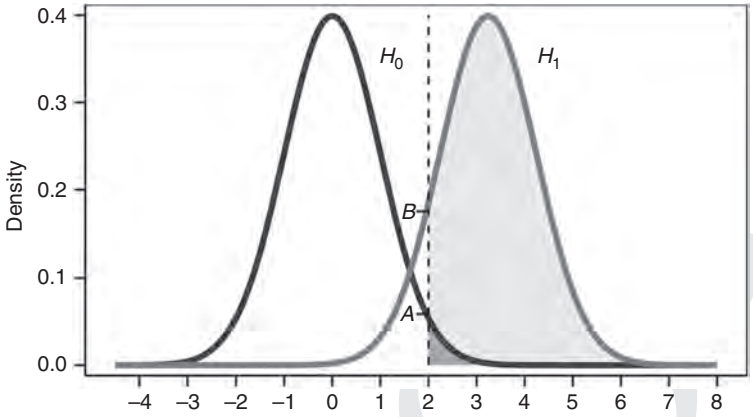

Figure 5.3 Comparing precise and imprecise *P*-values. The likelihood ratio is *A*/*B*, the ratio of the curve heights at the observed data. The likelihood ratio associated with the imprecise *P*-value ($p \leq \alpha$) is the ratio of the small darkly shaded area to the total shaded area (adapted from Goodman 1993, Figure 1 p. 492).

These are the ordinates not the tail areas of the Normal distribution.

That's the first step. The *second* step is to consider the likelihood ratio for the imprecise *P*-value, where the result is described coarsely as $\{z \geq 1.96\}$ rather than $z = 1.96$ (or equivalently $p \leq 0.025$ rather than $p = 0.025$):

(2) $\Pr(Z \geq 1.96; \mu_0)/ \Pr(Z \geq 1.96; \mu^{.9})$.

We see at once that the value of (2) is

α / POW($\mu^{.9}$) = 0.025/0.9 = 0.03.

The comparative evidence for the null using (1) is considerably larger (0.33) than using (2) where it's only 0.03. Figure 5.3 shows what's being compared.

The difference Goodman wishes to highlight looks even more impressive if we flip the ratios in (1) and (2) to get the comparative evidence for the alternative compared to the null. We get 0.176/0.058 = 3.03 using $z = 1.96$, and 0.9/0.025 = 36 using $z \geq 1.96$. Either way you look at it, using the tail areas to compare support exaggerates the evidence against the null in favor of $\mu^{.9}$. Or so it seems.

Now for the *third* step: Assign priors of 0.5 to μ_0 and to $\mu^{.9}$:

With this alternative hypothesis [$\mu^{.9}$], '$p \leq 0.05$' represents *11 times* (= 0.33/0.03) less evidence in support of the null hypothesis than does '$p = 0.05$.' Using Bayes' Theorem, with initial probabilities of 50 percent on both hypotheses (i.e., initial odds = 1), this means that after observing $p = 0.05$, the probability that the null hypothesis is true falls only to *25 percent* (= 0.33/(1 + 0.33)). When $p \leq 0.05$, the truth probability of the null hypothesis drops to 3 percent (= 0.03/(1 + 0.03)). (ibid., p. 491)

He concludes:

When we use the tail region to represent a result that is actually on the border, we misrepresent the evidence, making the case against the null hypothesis look much stronger than it actually is. (ibid.)

The posterior probabilities, he says, are assessments of credibility in the hypothesis. Join me for a break at the coffeehouse, called "Take Type II", where we'll engage a live exhibit.

Live Exhibit (ii): Drill Prompt. Using any relevant past travels, tours, and souvenirs, critically appraise his argument. Resist the impulse to simply ask, "where do significance tests recommend using error probabilities to form a likelihood ratio, and treat the result as a relative support measure?" Give him the most generous interpretation in order to see what's being claimed.

How do you grade the following test taker? Goodman actually runs two distinct issues together: The first issue contrasts a report of the observed *P*-values with a report of whether or not a predesignated cut-off is met (his "imprecise" *P*-value); the second issue is using likelihood ratios (as in (1)) as opposed to using tail areas. As I understand him, Goodman treats a report of the precise *P*-value as calling for a likelihood analysis as in (1), supplemented in the third step to get a posterior probability. But, even a reported *P*-value will lead to the use of tail areas in Fisherian statistics. This might not be a major point, but it is worth noting. For the error statistician, use of the tail area isn't to throw away the particular data and lump together all $z \geq z_\alpha$. It's to signal an interest in the probability the method would produce $z \geq z_\alpha$ under various hypotheses, to determine its capabilities (Translation Guide, Souvenir C). Goodman's hypothesis tester only reports $z \geq z_\alpha$ and so he portrays the Bayesian (or Likelihoodist) as forced to compute (2). The error statistical tester doesn't advocate this.

Let's have a look at (1): comparing the likelihoods of μ_0 and $\mu^{.9}$ given $z = 1.96$. Why look at an alternative so far away from μ_0? The value $\mu^{.9}$ gets a low likelihood, even given statistically significant z_α supplying a small denominator in (1). A classic issue with Bayes factors or likelihood ratios is the ease of

finding little evidence against a null, and even evidence for it, by choosing an appropriately distant alternative. But $\mu^{.9}$ is a "typical choice" in epidemiology, Goodman notes (ibid., p. 491). Sure it's a discrepancy we want high power to detect. That's different from using it to form a comparative likelihood. As Senn remarks, it's "ludicrous" to consider just the null and $\mu^{.9}$, which scarcely exhaust the parameter space, and "nonsense" to take a single statistically significant result as even decent evidence for the discrepancy we should not like to miss. $\mu^{.9}$ is around 1.28 standard errors from the one-sided cut-off: $(1.96 + 1.28)\ \sigma_{\overline{X}} = 3.24\ \sigma_{\overline{X}}$. We know right away that any μ value in excess of the observed statistically significant z_α is one we cannot have good evidence for. We can only have decent evidence that μ exceeds values that would form the CI lower bound for a confidence (or severity) level at least 0.7, 0.8, 0.9, etc. That requires *subtracting* from the observed data, not adding to it. Were the data generated by $\mu^{.9}$, then 90% of the time we'd have gotten a larger difference than we observed. Goodman might reply that I am using error probability criteria to judge his Bayesian analysis; but his Bayesian analysis was to show the significance test exaggerates evidence against the test hypothesis H_0. To the tester, it's *his* analysis that's exaggerating by giving credibility 0.75 to $\mu^{.9}$. Perhaps it makes sense with the 0.5 prior, but does *that* make sense?

Assigning 0.75 to the alternative $\mu^{.9}$ (using the likelihood ratio in (1)) does not convey that this is terrible evidence for a discrepancy that large. Granted, the posterior for $\mu^{.9}$ would be even higher using the tail areas in (2), namely, 0.97 – and that is his real point, which I've yet to consider. He's right that using (2) in his Bayesian computation gives a whopping 0.97 posterior probability to $\mu^{.9}$ (instead of merely 0.75, as on the analysis he endorses). Yet a significance tester wouldn't compute (2), it's outside significance testing. Considering the tails makes it harder, not easier, to find evidence against a null – when properly used in a significance test.

He's right that using $\alpha/\text{POW}(\mu^{.9})$ as a likelihood ratio in computing a posterior probability for the point alternative $\mu^{.9}$ (using spiked priors of 0.5) gives some very strange answers. The tester isn't looking to compare point values, but to infer discrepancies. I'm doing the best I can to relate his criticism to what significance testers do.

The error statistician deservedly conveys the lack of stringency owed to any inference to $\mu^{.9}$. She finds the data give fairly good grounds that μ is less than $\mu^{.9}$ (confidence level 0.9 if one-sided, 0.8 if two-sided). Statistical quantities are like stew ingredients that you can jumble together into a farrago of

computations, but there's a danger it produces an inference at odds with error statistical principles. For a significance tester, the alleged criticism falls apart; but it also leads me to question Goodman's posterior in (1).

What's Wrong with Using (1 − β)/α (or α/(1 − β)) to Compare Evidence? I think the test taker from last year's cruise did pretty well, don't you? But her "farrago" remark leads to a general point about the ills of using $(1 - \beta)/\alpha$ and $\alpha/(1 - \beta)$ to compare the evidence in favor of H_1 over H_0 or H_0 over H_1, respectively. Let's focus on the $(1 - \beta)/\alpha$ formulation. Benjamin and J. Berger (2016) call it the pre-data *rejection ratio*:

> It is the probability of rejection when the alternative hypothesis is true, divided by the probability of rejection when the null hypothesis is true, i.e., the ratio of the power of the experiment to the Type I error of the experiment. The rejection ratio has a straightforward interpretation as quantifying the strength of evidence about the alternative hypothesis relative to the null hypothesis conveyed by the experimental result being statistically significant. (p. 1)

But does it? I say no (and J. Berger says he concurs[3]). Let's illustrate. Looking at Figure 5.3, we can select an alternative associated with power as high as we like by dragging the curve representing H_1 to the right.

Imagine pulling it even further than alternative μ'^{9}. How about μ'^{999}? If we consider the alternative $\mu = \overline{x}_{0.025} + 3\sigma_{\overline{X}}$ then POW(T+, μ_1) = area to the right of −3 under the Normal curve, which is a whopping 0.999. For some numbers, use our familiar T+: H_0: $\mu \leq 0$ vs. H_1: $\mu > 0$, $\alpha = 0.025$, $n = 25$, $\sigma = 1$, $\sigma_{\overline{X}} = 0.2$. So the cut-off $\overline{x}_{0.025} = \mu_0 + 1.96\,\sigma_{\overline{X}} = (0 + 1.96(0.2)) = 0.392$. Thus,

$$\mu'^{.999} = 4.96\,\sigma_{\overline{X}} = 4.96(0.2) = 0.99 \simeq 1.$$

Let the observed outcome just reach the cut-off to reject H_0, $z_0 = 0.392$. The rejection ratio is

$$\text{POW(T+, } \mu = 1)/\alpha = 40 \text{ (i.e., 0.999/0.025)!}$$

Would even a Likelihoodist wish to say the strength of evidence for $\mu = 1$ is 40 times that of H_0? The data $\overline{x} = 0.392$ are even closer to 0 than to 1.

How then can it seem plausible, for comparativists, to compute a relative likelihood this way? We can view it through their eyes as follows: take $Z \geq 1.96$

[3] In private conversation.

as the lump outcome and reason along Likelihoodist lines. The probability is very high that $Z \geq 1.96$ under the assumption that $\mu = 1$:

$$\Pr(Z \geq 1.96; \mu^{0.999}) = 0.999.$$

The probability is low that $Z \geq 1.96$ under the assumption that $\mu = \mu_0 = 0$:

$$\Pr(Z \geq 1.96; \mu = \mu_0) = 0.025.$$

We've observed $z_0 = 1.96$ (so $Z \geq 1.96$).

> Therefore, $\mu^{0.999}$ (i.e., 1) makes the result $Z \geq 1.96$ more probable than does $\mu = 0$.

Therefore, the result is better evidence that $\mu = 1$ than it is for $\mu = 0$. But this likelihood reasoning only holds for the specific value of z. Granted, Bayarri, Benjamin, Berger, and Sellke (2016) recommend the prerejection ratio before the data are in, and "the 'post-experimental rejection ratio' (or Bayes factor) when presenting their experimental results" (p. 91). The authors regard the pre-data rejection ratio as frequentist, but it turns out they're using Berger's "frequentist principle," which, you will recall, is in terms of error probability$_2$ (Section 3.6). A creation built on frequentist measures doesn't mean the result captures frequentist error statistical$_1$ reasoning. It might be a kind of Frequentstein entity!

Notably, power works in the opposite way. If there's a high probability you should have observed a larger difference than you did, assuming the data came from a world where $\mu = \mu_1$, then the data indicate you're *not* in a world where $\mu > \mu_1$.

> If $\Pr(Z > z_0; \mu = \mu_1) =$ high, then $Z = z_0$ is strong evidence that $\mu \leq \mu_1$!

Rather than being evidence *for* μ_1, the just statistically significant result, or one that just misses, is evidence against μ being as high as μ_1. POW(μ_1) is not a measure of how well the data fit μ_1, but rather a measure of a test's capability to detect μ_1 by setting off the significance alarm (at size α). Having set off the alarm, you're not entitled to infer μ_1, but only discrepancies that have passed with severity (SIR). Else you're making mountains out of molehills.

5.3 Insignificant Results: Power Analysis and Severity

We're back at the Museum, and the display on power analysis. It is puzzling that many psychologists talk as if they're stuck with an account that says nothing about what may be inferred from negative results, when Cohen, the leader of power analysis, was toiling amongst them for years; and a central role

for power, post-data, is interpreting non-significant results. The attention to power, of course, is a key feature of N-P tests, but apparently the prevalence of Fisherian tests in the social sciences, coupled, some speculate, with the difficulty in calculating power, resulted in power receiving short shrift. Cohen's work was to cure all that. Cohen supplied a multitude of tables (when tables were all we had) to encourage researchers to design tests with sufficient power to detect effects of interest. He bemoaned the fact that his efforts appeared to be falling on deaf ears. Even now, problems with power persist, and its use post-data is mired in controversy.

The focus, in the remainder of this Tour, is on negative (or non-statistically significant) results. Test T+ fails to reject the null when the test statistic fails to reach the cut-off point for rejection, i.e., $d(\boldsymbol{x}_0) \leq c_\alpha$. A classic fallacy is to construe no evidence against H_0 as evidence of the correctness of H_0. A canonical example was in the list of slogans opening this book: Failing to find an increased risk is not evidence of no risk increase, if your test had little capability of detecting risks, even if present (as when you made your hurdles too high). The problem is the flip side of the fallacy of rejection: here the null hypothesis "survives" the test, but merely surviving can occur too frequently, even when there are discrepancies from H_0.

Power Analysis Follows Significance Test Reasoning

Early proponents of power analysis that I'm aware of include Cohen (1962), Gibbons and Pratt (1975), and Neyman (1955). It was the basis for my introducing severity, Mayo (1983). Both Neyman and Cohen make it clear that power analysis uses the same reasoning as does significance testing.[4] First Cohen:

> [F]or a given hypothesis test, one defines a numerical value **i** (for *i*ota) for the [population] ES (effect size), where **i** is so small that it is appropriate in the context to consider it negligible (trivial, inconsequential). Power $(1 - \beta)$ is then set at a high value, so that β is relatively small. When, additionally, α is specified, n can be found. Now, if the research is performed with this n and it results in nonsignificance, it is proper to conclude that the population ES is no more than **i**, i.e., that it is negligible . . . (Cohen 1988, p. 16; α, β substituted for his **a, b**).

Here Cohen imagines the researcher sets the size of a negligible discrepancy ahead of time – something not always available. Even where a negligible **i** may be specified, it's rare that the power to detect it is high. Two important points

[4] A key medical paper is Freiman et al. (1978).

can still be made: First, Cohen doesn't instruct you to infer there's no discrepancy from H_0, merely that it's "no more than **i**." Second, even if your test doesn't have high power to detect negligible **i**, you can infer the population discrepancy is less than whatever γ your test *does* have high power to detect. Some call this its *detectable discrepancy size*.

A little note on language. Cohen distinguishes the population ES and the observed ES, both in σ units. Keeping to Cohen's ES_s for "the effect size *in the sample*" (1988, p. 17) prevents a tendency to run them together. I continue to use "discrepancy" and "difference" for the population and observed differences, respectively, indicating the units being used.

Exhibit (iii): Ordinary Power Analysis. Now for how the inference from power analysis is akin to significance testing. Let $\mu^{1-\beta}$ be the alternative against which the null in T+ has high power, $1 - \beta$. Power analysis sanctions the inference that would accrue if we switched the null and alternative, yielding the one-sided test in the opposite direction, T–, we might call it. That is, T– tests H_0: $\mu \geq \mu^{1-\beta}$ vs. H_1: $\mu < \mu^{1-\beta}$ at the β level. The test rejects H_0 (at level β) when $\overline{X} \leq \mu_0 - z_\beta \sigma_{\overline{X}}$. Such a significant result would warrant inferring $\mu < \mu^{1-\beta}$ at level β. Using power analysis doesn't require making this switcheroo. The point is that there's essentially no new reasoning involved in power analysis, which is why members of the Fisherian tribe manage it without mentioning power.

> *Ordinary Power Analysis*: If data $\boldsymbol{x}$ are not statistically significantly different from H_0, and the power to detect discrepancy γ is high, then $\boldsymbol{x}$ indicates that the actual discrepancy is no greater than γ.

A simple example: Use $\mu^{.84}$ in test T+ ($H_0: \mu \leq 0$ vs. H_1; $\mu > 0$, $\alpha = 0.025$, $n = 25$, $\sigma_{\overline{X}} = 0.2$) to create test T–. Test T+ has 0.84 power against $\mu^{.84} = 3\sigma_{\overline{X}} = 0.6$ (with our usual rounding). So, test T– is H_0: $\mu \geq 0.6$ vs. H_1: $\mu < 0.6$, and a result is statistically significantly *smaller* than 0.6 at level 0.16 whenever the sample mean $\overline{X} \leq 0.6 - 1\sigma_{\overline{X}} = 0.4$. To check, note that $\Pr(\overline{X} \leq 0.4;\ \mu = 0.6) = \Pr(Z \leq -1) = 0.16 = \beta$.

It will be useful to look at the two-sided alternative: test $T^{\pm}$. We'd combine the above one-sided test with a test of H_0: $\mu \geq -\mu^{1-\beta}$ vs. H_1: $\mu < -\mu^{1-\beta}$ at the β level. This will be to test H_0: $\mu \geq -0.6$ vs. H_1: $\mu < -0.6$, and find a result statistically significantly smaller than -0.6 at level 0.16 whenever the sample mean $\overline{X} \leq -0.8$ (i.e., $-0.6 - 1\sigma_{\overline{X}}$). If both nulls are rejected (at the 0.16 level), we infer $|\mu| < 0.6$ but the two-sided test has double the Type I error probability: 0.32.

How high a power should be regarded as high? How low as low? A power of 0.8 or 0.9 is common, we saw, in "clinically relevant" discrepancies. To anyone

who complains that there's no way to draw a cut-off, note that we merely need to distinguish blatantly high from rather low values. Why have the probability of a Type II error exceed that of the Type I error? Some critics give Neyman and Pearson a hard time about this, but there's nothing in N-P tests to require it. Balance the errors as you like, N-P say. N-P recommend, based on tests in use, first to specify the test to reflect the Type I error as more serious than a Type II error. Second, choose a test that minimizes the Type II error probability, given the fixed Type I. In an example of testing a medical risk, Neyman says he places "a risk exists" as the test hypothesis since it's worse (for the consumer) to erroneously infer risk absence (1950, chapter V). Promoters of the precautionary principle are often surprised to learn this about N-P tests. However, there's never an automatic "accept/reject" in a scientific context.

Neyman Chides Carnap, Again

Neyman was an early power analyst? Yes, it's in his "The Problem of Inductive Inference" (1955) where we heard Neyman chide Carnap for ignoring the statistical model (Section 2.7). Neyman says:

> I am concerned with the term 'degree of confirmation' introduced by Carnap ... We have seen that the application of the locally best one-sided test to the data ... failed to reject the hypothesis [that the n observations come from a source in which the null hypothesis is true]. The question is: does this result 'confirm' the hypothesis [that H_0 is true of the particular data set]? (ibid., p. 40)

> The answer ... depends very much on the exact meaning given to the words 'confirmation,' 'confidence,' etc. If one uses these words to describe one's intuitive feeling of confidence in the hypothesis tested H_0, then ... the attitude described is dangerous ... the chance of detecting the presence [of discrepancy from the null], when only [n] observations are available, is extremely slim, even if [the discrepancy is present]. Therefore, the failure of the test to reject H_0 cannot be reasonably considered as anything like a confirmation of H_0. The situation would have been radically different if the power function [corresponding to a discrepancy of interest] were, for example, greater than 0.95. (ibid., p. 41)

Ironically, Neyman also criticizes Fisher's move from a large P-value to inferring the null hypothesis as:

> much too automatic [because] ... large values of P may be obtained when the hypothesis tested is false to an important degree. Thus, ... it is advisable to investigate ... what is the probability (probability of error of the second kind) of obtaining a large value of P in cases when the [null is false ... to a specified degree]. (1957a, p. 13)

Should this calculation show that the probability of detecting an appreciable error in the hypothesis tested was large, say 0.95 or greater, then and only then is the decision in favour of the hypothesis tested justifiable in the same sense as the decision against this hypothesis is justifiable when an appropriate test rejects it at a chosen level of significance. (1957b, pp. 16–17)

Typically, the hypothesis tested, $[H_0]$ in the N-P context, could be swapped with the alternative. Let's leave the museum where a leader of the severe testing tribe makes some comparisons.

Attained Power Π

So power analysis is in the spirit of severe testing. Still, power analysis is calculated relative to an outcome just missing the cut-off c_α. This corresponds to an observed difference whose *P*-value just exceeds α. This is, in effect, the worst case of a negative result. What if the actual outcome yields an even smaller difference (larger *P*-value)?

Consider test T+ ($\alpha = 0.025$) above. No one wants to turn pages so here it is: H_0: $\mu \leq 0$ vs. H_1: $\mu > 0$, $\alpha = 0.025$, $n = 25$, $\sigma = 1$, $\sigma_{\overline{X}} = 0.2$. So the cut-off $\overline{x}_{0.025} = \mu_0 + 1.96\,\sigma_{\overline{X}} = (0 + 1.96(0.2)) = 0.392$, or, with the $2\,\sigma_{\overline{X}}$ cut-off, $\overline{x}_{0.025} = 0.4$. Consider an arbitrary inference $\mu \leq 0.2$. We know POW(T+, $\mu = 0.2$) = 0.16 ($1\sigma_{\overline{X}}$ is subtracted from 0.4). A value of 0.16 is quite lousy power. It follows that no statistically insignificant result can warrant $\mu \leq 0.2$ for the power analyst. Power analysis only allows ruling out values as high as $\mu^{.8}$, $\mu^{.84}$, $\mu^{.9}$, and so on. The power of a test is fixed once and for all and doesn't change with the observed mean $\overline{x}$. Why consider every non-significant result as if it just barely missed the cut-off? Suppose, $\overline{x} = -0.2$. This is $2\sigma_{\overline{X}}$ lower than 0.2. Surely that should be taken into account? It is: 0.2 is the upper 0.975 confidence bound and $\text{SEV}(\text{T}+,\ \overline{x} = -0.2,\ \mu \leq 0.2) = 0.975$.[5]

What enables substituting the observed value of the test statistic, $\text{d}(\boldsymbol{x}_0)$, is the counterfactual reasoning of severity:

> If, with high probability, the test would have resulted in a larger observed difference (a smaller *P*-value) than it did, if the discrepancy was as large as γ, then there's a good indication the discrepancy is no greater than γ, i.e., that $\mu \leq \mu_0 + \gamma$.

That is, if the *attained power* (att-power) of T+ against $\mu \leq \mu_0 + \gamma$ is very high, the inference to $\mu \leq \mu_0 + \gamma$ is warranted with severity. (As always, whether it's a mere indication or genuine evidence depends on whether it passes an audit.)

[5] To show, as crude power analysis does not, that $\text{SEV}(\text{T}+,\ \overline{x} = -0.2,\ \mu \leq 0.2) = 0.975$ when $\overline{x} = -0.2$: We standardize $\overline{x}$ to get $z = (-0.2 - 0.2)/0.2 = -2$ and so $\Pr(\overline{x} \geq -0.2;\ 0.2) = 0.975$.

If you elect to use the term attained power, you'll have to avoid confusing it with animals given similar names; I'll introduce you to them shortly.

Compare power analytic reasoning with severity (or att-power) reasoning from a negative or insignificant result from T+.

> *Power Analysis*: If $\Pr(d(\boldsymbol{X}) \geq c_\alpha; \mu_1)$ = high and the result is not significant, then it's an indication or evidence that $\mu \leq \mu_1$.
>
> *Severity Analysis*: If $\Pr(d(\boldsymbol{X}) \geq d(\boldsymbol{x}_0); \mu_1)$ = high and the result is not significant, then it's an indication or evidence that $\mu \leq \mu_1$.

Severity replaces the predesignated cut-off c_α with the observed $d(\boldsymbol{x}_0)$. Thus we obtain the same result if we choose to remain in the Fisherian tribe, as seen in the Frequentist Evidential Principle FEV(ii) (Section 3.3).

We still abide by the logic of power analysis, since if $\Pr(d(\boldsymbol{X}) \geq d(\boldsymbol{x}_0); \mu_1)$ = high, then $\Pr(d(\boldsymbol{X}) \geq c_\alpha; \mu_1)$ = high, at least in a test with a sensible distance measure like T+. In other words, power analysis is conservative. It gives a sufficient but not a necessary condition for warranting an upper bound: $\mu \leq \mu_1$. But it can be way too conservative as we just saw.

(1) $\Pr(d(\boldsymbol{X}) \geq c_\alpha; \mu = \mu_0 + \gamma)$: Power to detect γ.

Ordinary power analysis requires (1) to be high (for non-significance to warrant $\mu \leq \mu_0 + \gamma$).

Just missing the cut-off c_α is the worst case. It is more informative to look at (2):

(2) $\Pr(d(\boldsymbol{X}) \geq d(\boldsymbol{x}_0); \mu = \mu_0 + \gamma)$: Attained power ($\Pi(\gamma)$).

(1) can be low while (2) is high. The computation in (2) measures the severity (or degree of corroboration) for the inference $\mu \leq \mu_0 + \gamma$. The analysis with Cox kept to $\Pi(\gamma)$ (or "attained sensitivity"), keeping "power" out of it.

As an entrée to Exhibit (iv): Isn't severity just power? This is to compare apples and frogs. The power of a test to detect an alternative is an error probability of a method (one minus the probability of the corresponding Type II error). Power *analysis* is a way of using power to assess a statistical inference in the case of a negative result. Severity, by contrast, is always to assess a statistical inference. Severity is always in relation to a particular claim or inference C, from a test T and an outcome $\boldsymbol{x}$. So with that out of the way, what if the question is put properly thus: If a result from test T+ is just statistically insignificant at level α, then is the test's power to detect μ_1 equal to the severity for inference C: $\mu > \mu_1$? The answer is no. It would be equal to the severity for inferring the *denial* of C! See Figure 5.4 comparing $\text{SEV}(\mu > \mu_1)$ and $\text{POW}(\mu_1)$.

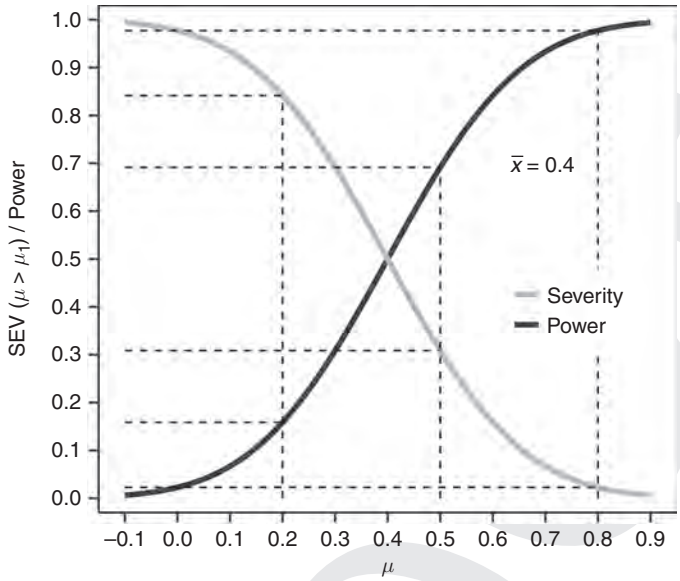

Figure 5.4 Severity for ($\mu > \mu_1$) vs power (μ_1).

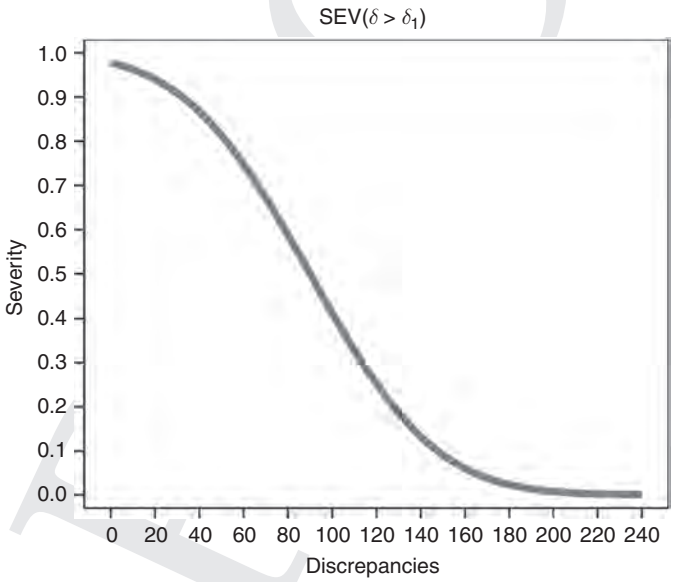

Figure 5.5 The observed difference is 90, each group has $n = 200$ patients, and the standard deviation is 450.

Exhibit (iv): Difference Between Means Illustration. I've been making use of Senn's points about the nonsensical and the ludicrous in motivating the severity assessment in relation to power. I want to show you severity for a difference between means, and fortuitously Senn (2019) alludes to severity in the latest edition of his textbook to make the same point. An example is a placebo-controlled trial in asthma where the amount of air a person can exhale in one second, the forced expiratory volume, is measured. "The clinically relevant difference is presumed to be 200 ml and the standard deviation 450 ml" (p. 197). It's a test of

$$H_0: \delta = \mu_1 - \mu_0 \leq 0 \text{ vs. } H_1: \delta > 0.$$

He will use a one-sided test at the 2.5% level, using 200 patients in each group yielding the standard error (SE) for the difference between two means equal to $(450\sqrt{2}/\sqrt{n})$.[6] The test has over 0.99 power for detecting $\delta = 200$. The observed difference $d = \left(\overline{X} - \overline{Y}\right) = 90$, which is statistically significant at the 0.025 level (90/SE = 2), but it wouldn't warrant inferring $\delta > 200$, and we can see the severity for $\delta > 200$ is extremely low. The observed difference is statistically significantly different from H_0: in accord with (or in the direction of) H_1, so severity computes the probability of a worse fit with H_1 under $\delta = 200$:

$$\text{SEV}(\delta > 200) = \Pr(d < 90; \delta = 200).$$

While this is strictly a Student's t distribution, given the large sample size, we can use the Normal distribution:[7]

$$Z = \left\{\frac{\sqrt{200}(90 - 200)}{450\sqrt{2}}\right\} = (90 - 200)/45 = -2.44.$$

SEV($\delta > 200$) = the area to the left of −2.44 (See Figure 5.5) which is 0.007,

For a few examples, SEV($\delta > 100$) = 0.412 the area to the left of $z = (90 - 100)/45 = -0.222$. SEV($\delta > 50$) = 0.813 the area to the left of $z = (90 - 50)/45 = 0.889$.

SEV($\delta > 10$) = 0.962 the area to the left of $z = (90 - 10)/45 = 1.778$.

[6] From the general case, we get the case where $n_1 = n_2$,

$$\sigma_{M_1 - M_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{\sigma^2}{n} + \frac{\sigma^2}{n}} = \sqrt{\frac{2\sigma^2}{n}} = \frac{\sqrt{2}\sigma}{\sqrt{n}}$$

[7] $$\frac{\sqrt{n}\left(\overline{X} - \overline{Y}\right)}{s\sqrt{2}}$$

Senn gives another way to view the severity assessment of $\delta > \delta'$, namely "adopt $[\delta = \delta']$, as a null hypothesis and then turn the significance test machinery on it (2019). In the case of testing $\delta = 200$, the P-value would be 0.999. Scarcely evidence against it. We first visited this mathematical link in touring the connection between severity and confidence intervals in Section 3.8. As noted, the error statistician is loath to advocate modifying the null hypothesis because the point of a severity assessment is to supply a basis for interpreting tests that is absent in existing tests. Since significance tests aren't explicit about assessing discrepancies, and since the rationale for P-values is questioned in all the ways we've espied, it's best to supply a fresh rationale. I have offered the severity rationale as a basis for understanding, if not buying, error statistical reasoning. The severity computation might be seen as a rule of thumb to avoid misinterpretations; it could be arrived at through other means, including varying the null hypotheses. It's the idea of viewing statistical inference as severe testing that invites a non-trivial difference with probabilism.

5.4 Severity Interpretation of Tests: Severity Curves

We visit severity tribes who have prepared an overview that synthesizes non-significant results from Fisherian as well as ("do not reject") results from N-P tests. Following the minimal principle of severity:

> (a) If data d($\boldsymbol{x}$) are not statistically significantly different from H_0, but the capability of detecting discrepancy γ is low, then d($\boldsymbol{x}$) is not good evidence that the actual discrepancy is less than γ.

What counts as a discrepancy "of interest" is a separate question, outside of statistics proper. You needn't know it to ask: What discrepancies, if they existed, would very probably have led your method to show a more significant result than you found? Upon finding this, you may infer that, at best, the test can rule out increases of that extent.

> (b) If data d($\boldsymbol{x}$) are not statistically significantly different from H_0, but the probability to detect discrepancy γ is high, then $\boldsymbol{x}$ is good evidence that the actual discrepancy is less than or equal to γ.

We are not changing the original null and alternative hypotheses! We're using the severe testing concept to interpret the negative results – the kind of scrutiny in which one might be interested, to follow Neyman, "when we are faced with . . . interpreting the results of an experiment planned and performed by someone else" (Neyman 1957b, p. 15). We want to know how well tested are claims of form $\mu \leq \mu_1$, where $\mu_1 = (\mu_0 + \gamma)$, for some $\gamma \geq 0$.

Why object to applying the severity analysis by changing the null hypothesis, and doing a simple *P*-value computation? *P*-values, especially if plucked from thin air this way, are themselves in need of justification. That's a major goal of this journey. It's only by imagining we have either a best or good test or corresponding distance measure (let alone assuming we don't have to deal with lots of nuisance parameters) that substituting different null hypotheses works out.

Pre-data, we need a test with good error probabilities (as discussed in Section 3.2). That assures we avoid some worst case. Post-data we go further.

For a claim H to pass with severity requires not just that (S-1) the data accord with H, but also that (S-2) the test probably would have produced a worse fit, if H were false in specified ways. We often let the measure of accordance (in (S-1)) vary and train our critical focus on (S-2), but here it's a best test. Consider statistically insignificant results from test T+. The result "accords with" H_0, so we have (S-1), but we're wondering about (S-2): how probable is it that test T+ would have produced a result that accords *less* well with H_0 than $\boldsymbol{x_0}$ does, were H_0 false? An equivalent but perhaps more natural phrase for "a result that accords *less* well with H_0" is "a result *more discordant.*" Your choice.

Souvenir W: The Severity Interpretation of Negative Results (SIN) for Test T+

Applying our general abbreviation: SEV(test T+, outcome $\boldsymbol{x}$, inference H), we get "the severity with which $\mu \leq \mu_1$ passes test T+, with data $\boldsymbol{x}_0$":

$$\text{SEV}(\text{T+}, \text{d}(\boldsymbol{x}_0), \mu \leq \mu_1),$$

where $\mu_1 = (\mu_0 + \gamma)$, for some $\gamma \geq 0$. If it's clear which test we're discussing, we use our abbreviation: SEV($\mu \leq \mu_1$). We obtain a companion to the severity interpretation of rejection (SIR), Section 4.4, Souvenir R:

SIN *(Severity Interpretation for Negative Results)*

(a) If there is a very *low* probability that $\text{d}(\boldsymbol{x_0})$ would have been larger than it is, even if $\mu > \mu_1$, then $\mu \leq \mu_1$ passes with *low* severity: SEV($\mu \leq \mu_1$) is low.
(b) If there is a very *high* probability that $\text{d}(\boldsymbol{x}_0)$ would have been larger than it is, were $\mu > \mu_1$, then $\mu \leq \mu_1$ passes the test with *high* severity: SEV($\mu \leq \mu_1$) is high.

To break it down, in the case of a statistically insignificant result:

$$\text{SEV}(\mu \leq \mu_1) = \Pr(\text{d}(\boldsymbol{X}) > \text{d}(\boldsymbol{x}_0); \mu \leq \mu_1 \text{ false}).$$

We look at $\{d(\boldsymbol{X}) > d(\boldsymbol{x}_0)\}$ because severity directs us to consider a "worse fit" with the claim of interest. That $\mu \leq \mu_1$ is false within our model means that $\mu > \mu_1$. Thus:

$$\text{SEV}(\mu \leq \mu_1) = \Pr(d(\boldsymbol{X}) > d(\boldsymbol{x}_0); \mu > \mu_1).$$

Now $\mu > \mu_1$ is a composite hypothesis, containing all the values in excess of μ_1. How can we compute it? As with power calculations, we evaluate severity at a point $\mu_1 = (\mu_0 + \gamma)$, for some $\gamma \geq 0$, because for values $\mu \geq \mu_1$ the severity increases. So we need only to compute

$$\text{SEV}(\mu \leq \mu_1) > \Pr(d(\boldsymbol{X}) > d(\boldsymbol{x}_0); \mu = \mu_1).$$

To compute SEV we compute $\Pr(d(\boldsymbol{X}) > d(\boldsymbol{x}_0); \mu = \mu_1)$ for any μ_1 of interest. Swapping out the claims of interest (in significant and insignificant results), gives us a single criterion of a good test, severity.

Exhibit(v): Severity Curves. The severity tribes want to present severity using a standard Normal example, one where $\sigma_{\overline{X}} = 1$ (as in the water plant accident). For this illustration:

$$\text{Test T+}: H_0: \mu \leq 0 \text{ vs. } H_1: \mu > 0,\ \sigma = 10,\ n = 100, \sigma/\sqrt{n} = \sigma_{\overline{X}} = 1.$$

$$\text{If } \alpha = 0.025, \text{ we reject } H_0 \text{ iff } d(\boldsymbol{X}) \geq c_{0.025} = 1.96.$$

Suppose test T+ yields the statistically insignificant result $d(\boldsymbol{x}_0) = 1.5$. Under the alternative $d(\boldsymbol{X})$ is $N(\delta, 1)$ where $\delta = (\mu - \mu_0)/\sigma_{\overline{X}}$.

Even without identifying a discrepancy of importance ahead of time, the severity associated with various inferences can be evaluated.

The severity curves (Figure 5.6) show $d(\boldsymbol{x}_0) = 0.5$, 1, 1.5, and 1.96.

$$\text{How severely does } \mu \leq 0.5 \text{ pass the test with } \overline{X} = 1.5\ (d(\boldsymbol{x}_0) = 1.5)?$$

The easiest way to compute it is to go back to the observed $\overline{x}_0$, which would be 1.5.

$$\text{SEV}(\mu \leq 0.5) = \Pr(\overline{X} > 1.5;\ \mu = 0.5) = 0.16.$$

Here, $Z = [(1.5 - 0.5)/1] = 1$, and the area under the standard Normal distribution to the right of 1 is 0.16. Lousy. We can read it off the curve, looking at where the $d(\boldsymbol{x}) = 1.5$ curve hits the bottom-most dotted line. The severity (vertical) axis hits 0.16, and the corresponding value on the μ axis is 0.5. This could be used more generally as a discrepancy axis, as I'll show.

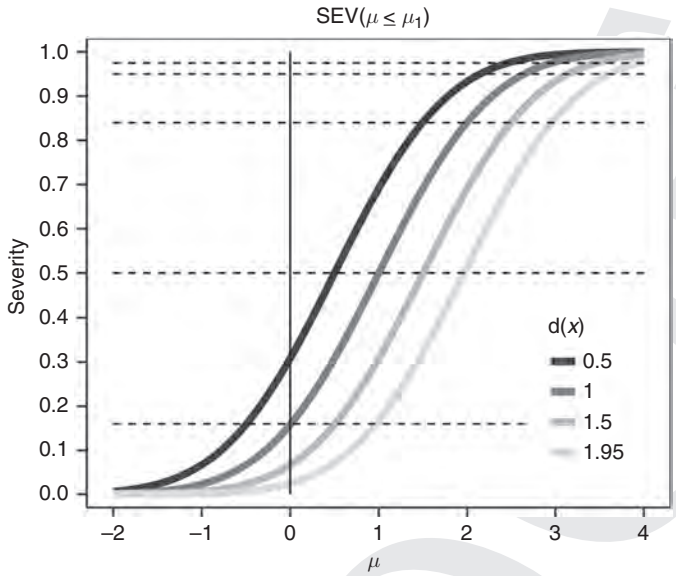

Figure 5.6 Severity curves.

We can find some discrepancy from the null that this statistically insignificant result warrants ruling out at a reasonable level – one that very probably would have produced a more significant result than was observed. The value $d(\boldsymbol{x}_0) = 1.5$ yields a severity of 0.84 for a discrepancy of 2.5. $SEV(\mu \leq 2.5) = 0.84$ with $d(\boldsymbol{x}_0) = 1.5$. Compare this to $d(\boldsymbol{x}) = 1.95$, failing to trigger the significance alarm. Now a larger upper bound is needed for severity 0.84, namely, $\mu \leq 2.95$. If we have discrepancies of interest, by setting a high power to detect them, we ensure – ahead of time – that any insignificant result entitles us to infer "it's not that high." Power against μ_1 evaluates the worst (i.e., lowest) severity for values $\mu \leq \mu_1$ for any outcome that leads to non-rejection. This test ensures any insignificant result entitles us to infer $\mu \leq 2.95$, call it $\mu \leq 3$. But we can determine discrepancies that pass with severity, post-data, without setting them at the outset. Compare four different outcomes:

$d(\boldsymbol{x}_0) = 0.5$, $SEV(\mu \leq 1.5) = 0.84$; $d(\boldsymbol{x}_0) = 1$, $SEV(\mu \leq 2) = 0.84$;

$d(\boldsymbol{x}_0) = 1.5$, $SEV(\mu \leq 2.5) = 0.84$; $d(\boldsymbol{x}_0) = 1.95$, $SEV(\mu \leq 2.95) = 0.84$.

In relation to test T+ (standard Normal): If you add $1\sigma_{\overline{X}}$ to $d(\boldsymbol{x}_0)$, the result being μ_1, then $SEV(\mu \leq \mu_1) = 0.84$.[8]

We can also use severity curves to compare the severity for a given claim, say $\mu \leq 1.5$:

$d(\boldsymbol{x}_0) = 0.5$, $SEV(\mu \leq 1.5) = 0.84$; $d(\boldsymbol{x}_0) = 1$, $SEV(\mu \leq 1.5) = 0.7$;

$d(\boldsymbol{x}_0) = 1.5$, $SEV(\mu \leq 1.5) = 0.5$; $d(\boldsymbol{x}_0) = 1.95$, $SEV(\mu \leq 1.5) = 0.3$.

Low and high benchmarks convey what is and is not licensed, and suffice for avoiding fallacies of acceptance. We can deduce SIN from the case where T+ has led to a statistically significant result, SIR. In that case, the inference that passes the test is of form $\mu > \mu_1$, where $\mu_1 = \mu_0 + \gamma$. Because $(\mu > \mu_1)$ and $(\mu \leq \mu_1)$ partition the parameter space of μ, we get $SEV(\mu > \mu_1) = 1 - SEV(\mu \leq \mu_1)$.

The more devoted amongst you will want to improve and generalize my severity curves. Some of you are staying the night at Confidence Court Inn, others at Best Bet and Breakfast. We meet at the shared lounge, Calibration. Here's a souvenir of SIR, and SIN.

Souvenir X: Power and Severity Analysis

Let's record some highlights from Tour I:

First, ordinary power analysis versus severity analysis for Test T+:

Ordinary Power Analysis: If $\Pr(d(\mathbf{X}) \geq c_\alpha; \mu_1)$ = high and the result is not significant, then it's an indication or evidence that $\mu \leq \mu_1$.

Severity Analysis: If $\Pr(d(\mathbf{X}) \geq d(\boldsymbol{x}_0); \mu_1)$ = high and the result is not significant, then it's an indication or evidence that $\mu \leq \mu_1$.

It can happen that claim $\mu \leq \mu_1$ is warranted by severity analysis but not by power analysis.

[8] • If you add $k\sigma_{\overline{X}}$ to $d(\boldsymbol{x}_0)$, $k > 0$, the result being μ_1, then $SEV(\mu \leq \mu_1)$ = area to the right of $-k$ under the standard Normal (SEV > 0.5).
• If you subtract $k\sigma_{\overline{X}}$ from $d(\boldsymbol{x}_0)$, the result being μ_1, then $SEV(\mu \leq \mu_1)$ = area to the right of k under the standard Normal (SEV ≤ 0.5).

For the general case of Test T+, you'd be adding or subtracting $k\sigma_{\overline{X}}$ to $(\mu_0 + d(\boldsymbol{x}_0)\sigma_{\overline{X}})$. We know that adding $0.85\sigma_{\overline{X}}$, $1\ \sigma_{\overline{X}}$, and $1.28\sigma_{\overline{X}}$ to the cut-off for rejection in a test T+ results in μ values against which the test has 0.8, 0.84, and 0.9 power. If you treat the observed $\overline{x}$ as if it were being contemplated as the cut-off, and add $0.85\sigma_{\overline{X}}$, $1\sigma_{\overline{X}}$, and $1.28\sigma_{\overline{X}}$, you will arrive at μ_1 values such that $SEV(\mu \leq \mu_1) = 0.8$, 0.84, and 0.9, respectively. That's because severity goes in the same direction as power for non-rejection in T+. For familiar numbers of $\sigma_{\overline{X}}$'s added/subtracted to $\overline{x} = \mu_0 + d_0\sigma_{\overline{X}}$:

Claim	$(\mu \leq \overline{x} - 1\sigma_{\overline{X}})$	$(\mu \leq \overline{x})$	$(\mu \leq \overline{x} + 1\sigma_{\overline{X}})$	$(\mu \leq \overline{x} + 1.65\sigma_{\overline{X}})$	$(\mu \leq \overline{x} + 1.98\sigma_{\overline{X}})$
SEV	0.16	0.5	0.84	0.95	0.975

Now an overview of severity for test T+: Normal testing: H_0: $\mu \leq \mu_0$ vs. H_1: $\mu > \mu_0$ with σ known. The severity reinterpretation is set out using discrepancy parameter γ. We often use μ_1 where $\mu_1 = \mu_0 + \gamma$.

> Reject H_0 (with $\boldsymbol{x}_0$) licenses inferences of the form $\mu > [\mu_0 + \gamma]$, for some $\gamma \geq 0$, but with a warning as to $\mu \leq [\mu_0 + \kappa]$, for some $\kappa \geq 0$.
>
> Non-reject H_0 (with $\boldsymbol{x}_0$) licenses inferences of the form $\mu \leq [\mu_0 + \gamma]$, for some $\gamma \geq 0$, but with a warning as to values fairly well indicated $\mu > [\mu_0 + \kappa]$, for some $\kappa \geq 0$.

The severe tester reports the attained significance levels and at least two other benchmarks: claims warranted with severity, and ones that are poorly warranted.

Talking through SIN and SIR. Let $d_0 = d(\boldsymbol{x}_0)$.

SIN *(Severity Interpretation for Negative Results)*

(a) *low*: If there is a very *low* probability that d_0 would have been larger than it is, even if $\mu > \mu_1$, then $\mu \leq \mu_1$ passes with *low* severity: SEV($\mu \leq \mu_1$) is low (i.e., your test wasn't very capable of detecting discrepancy μ_1 even if it existed, so when it's not detected, it's poor evidence of its absence).

(b) *high*: If there is a very *high* probability that d_0 would have been larger than it is, were $\mu > \mu_1$, then $\mu \leq \mu_1$ passes the test with *high* severity: SEV($\mu \leq \mu_1$) is high (i.e., your test was highly capable of detecting discrepancy μ_1 if it existed, so when it's not detected, it's a good indication of its absence).

SIR *(Severity Interpretation for Significant Results)*

If the significance level is small, it's indicative of some discrepancy from H_0, we're concerned about the magnitude:

(a) *low*: If there is a fairly high probability that d_0 would have been larger than it is, even if $\mu = \mu_1$, then d_0 is not a good indication $\mu > \mu_1$: SEV($\mu > \mu_1$) is low.[9]

(b) *high*: Here are two ways, choose your preferred:

- (b-1) If there is a very high probability that d_0 would have been smaller than it is, if $\mu \leq \mu_1$, then when you observe so large a d_0, it indicates $\mu > \mu_1$: SEV($\mu > \mu_1$) is high.

[9] A good rule of thumb to ascertain if a claim C is warranted is to think of a statistical *modus tollens* argument, and find what would occur with high probability, were claim C false.

- (b-2) If there's a very low probability that so large a d_0 would have resulted, if μ were no greater than μ_1, then d_0 indicates $\mu > \mu_1$: SEV($\mu > \mu_1$) is high.[10]

[10] For a shorthand that covers both severity and FEV for Test T+ with small significance level (Section 3.1):

(FEV/SEV): If d($\boldsymbol{x}_0$) is not statistically significant, then $\mu \leq \overline{x} + k_\varepsilon \sigma/\sqrt{n}$ passes the test T+ with severity $(1 - \varepsilon)$

(FEV/SEV): If d($\boldsymbol{x}_0$) is statistically significant, then $\mu > \overline{x} - k_\varepsilon \sigma/\sqrt{n}$ passes test T+ with severity $(1 - \varepsilon)$,

where $\Pr(d(\boldsymbol{X}) > k_\varepsilon) = \varepsilon$ (Mayo and Spanos (2006), Mayo and Cox (2006).)