Excursion 5 Tour II: Shpower and Retrospective Power

"There's a sinister side to statistical power" (**SIST** 354) I call it *Shpower analysis* because it distorts the logic of ordinary power analysis (from insignificant results).

Because ordinary power analysis is also post data, the criticisms of shpower are wrongly taken to reject both.

Shpower evaluates power with respect to the hypothesis that the population effect size (discrepancy) equals the observed effect size, e.g., the parameter μ equals the observed mean $\bar{x}_0$, i.e., in *T*+ this would be to set μ = $\bar{x}_0$).

The Shpower of test T+: $(\bar{X} < \bar{x}_\alpha; \mu = \bar{x}_0)$.

The Shpower of test T+: Pr($\bar{X} < \bar{x}_\alpha$; μ = $\bar{x}_0$)

The thinking is since we don't know the value of μ, we might use the observed $\bar{x}_0$ to estimate it, and then compute power in the usual way, except substituting the observed value.

Can't work for the purpose of using power analysis to interpret insignificant results. Why?

Since alternative μ is set = $\bar{x}_0$, and $\bar{x}_0$ is given as statistically insignificant, we are in Case 1 from 5.1 (Exhibit i): the power can never exceed .5.

In other words, since shpower = POW(T+, μ = $\bar{x}_0$), and $\bar{x}_0 < \bar{x}_\alpha$,the power can't exceed .5.

Between H_0 and $\bar{x}_\alpha$ the power goes from α to .5.

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:

POW(T+, μ_0) = Pr($\bar{X} > \bar{x}_\alpha; \mu_0$), $\bar{x}_\alpha = (\mu_0 + z_\alpha \sigma_{\bar{X}})$,

$\sigma_{\bar{X}} = [\sigma/\sqrt{n}]$

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

But power analytic reasoning is all about finding an alternative against which the test has *high* capability to have obtained significance. Shpower is always "slim" (to echo Neyman) against such alternatives.

Unsurprisingly, Shpower analytical reasoning has been criticized in the literature: But the critics think they're maligning power analytic reasoning.

The severe tester uses attained power Pr(d(**X**) > $d(\boldsymbol{x}_0)$; μ') to evaluate severity, but to address criticisms of power analysis, we have to stick to ordinary power (**SIST** 355).

Ordinary Power POW (μ'): Pr(d(**X**) > c_α; μ')
Shpower: Observed or retro-power: Pr(d(**X**) > c_α; μ = $\bar{x}_0$)

An article by Hoenig and Heisey (2001) ("The Abuse of Power") calls power analysis abusive. Is it? Aris Spanos and I say no (in a 2002 note)

Power-analytic reasoning: *High* power to get significance when μ = μ', together with your *not getting significance* indicates μ < μ'

But if you replace μ' by $\bar{x}_0$,it will never be high.

Exhibit (vii) (SIST, p. 359): Gelman and Carlin (2014) appear to be at odds with the upshot of quiz on p. 323, start of Tour I.

From our mountains out of molehill fallacies, if POW(μ') is high then a just significant result is *poor* evidence that μ > μ'; while if POW(μ') is low it's good evidence that μ > μ'.

A way to make sense of their view is to see them as saying if the observed mean is so out of whack with what's known, that we suspect the assumptions of the test are questionable or invalid.

(See **SIST** pp. 360-1)

5.6 Positive Predictive Value: Fine for Luggage (SIST 361)

To understand how the *diagnostic screening* criticism tests really took off, go back to a paper by John Ioannidis (2005).

> Several methodologists have pointed out that the high rate of nonreplication (lack of confirmation) of research discoveries is a consequence of the convenient, yet ill-founded strategy of claiming conclusive research findings solely on the basis of a single study assessed by formal statistical significance, typically for a p-value less than 0.05. Research is not most appropriately represented and summarized by p-values, but, unfortunately, there is a widespread notion that medical research articles should be interpreted based only on p-values. …

> It can be proven that most claimed research findings are false (p. 0696).

However absurd such behavior sounds, 70 years after Fisher exhorted us never to rely on “isolated results,” let’s suppose Ioannidis is right.

Worse, even the single significant result is very often the result of the cherry-picking and data-dredging.

> Commercially available ‘data mining’ packages actually are proud of their ability *to yield statistically significant results through data dredging* (ibid., p. 0699).

Diagnostic-Screening Model of Tests

- If we imagine randomly selecting a hypothesis from an urn of nulls 90% of which are true
- *Consider just 2 possibilities:* H_0: no effect
  H_1: meaningful effect, all else ignored,
- Take the prevalence of 90% as
  $\Pr(H_0) = 0.9$, $\Pr(H_1) = 0.1$
- Reject H_0 with a single (just) 0.05 significant result, with cherry-picking, selection effects

9

Then it can be shown most "findings" are false

Diagnostic Screening (DS) model of Tests

- **Pr(H_0|Test T rejects H_0) > 0.5**

really: prevalence of true nulls among those rejected at the 0.05 level > 0.5.

Call this: False Finding rate FFR

- **Pr(Test T rejects H_0 | H_0) = 0.05**

Criticism: N-P Type I error probability ≠ FFR

(Ioannidis 2005, Colquhoun 2014)

FFR: False Finding Rate: Prev(H_0) = .9

$$\Pr(H_0|\text{T rejects } H_0) =$$

$$\frac{\Pr(\text{T rejects } H_0|\, H_0)\Pr(H_0)}{\Pr(\text{T rejects } H_0|H_0)\Pr(H_0)+\Pr(\text{T rejects } H_0|H_1)\Pr(H_1)}$$

$$= \frac{\alpha \Pr(H_0)}{\alpha \Pr(H_0) + \text{POW}(H_1)\Pr(H_1)}$$

α = 0.05 and (1 – β) = .8, FFR = 0.36, the PPV = .64

Pr(H_0 / findings with a P-value of .05) ≠ Pr (reject at level .05; H_0)

Only the second one is a Type 1 error probability)

Positive Predictive Value (PPV) (1 – FFR). Apply Bayes' rule using the given relative frequencies (or prevalences):

$$\text{PPV: } \Pr(\text{D}|+) = \frac{\Pr(+|\text{D})\ \Pr(\text{D})}{[\Pr(+|\text{D})\ \Pr(\text{D}) + \Pr(+|\sim\text{D})\ \Pr(\sim\text{D})]}$$

$$= \frac{1}{(1+\text{B})}$$

where $$\text{B} = \frac{\Pr(+|\sim\text{D})\ \Pr(\sim\text{D})}{\Pr(+|\text{D})\ \Pr(\text{D})}$$

Sensitivity

SENS: Pr(+ | D).
H_1: D: Dangerous bag
(~ power)

H_0: ~D: no danger

Specificity

SPEC: Pr(– | ~D);
(1 – α)

Even with Pr(D) = .5, with Pr(+|~D) = .05 and Pr(+|D)= .8, we still get a rather high

$$PPV = \frac{1}{\left[\frac{1+Pr(+|\sim D)}{Pr(+|D)}\right]}$$

1/(1 + 1/16) = 16/17

With Pr(D) = .5, all we need for a PPV greater than .5 is for Pr(+|~D) to be less than Pr(+|D).

With a small prevalence Pr(D) e.g., < Pr(+|~D) (< α)

- We get PPV < .5 even with a maximal sensitivity Pr(+|D) of 1. In There is still a boost from the prior prevalence.

Recall absolute vs relative confirmation (B – boost)

- Chart SIST 365

- What is prevalence? (bott 366)

Probabilistic instantiation fallacy (367)

The outcome may be ***X*** = 1 or 0 according to whether the hypothesis we've selected is true.

- The probability of ***X*** = 1 is .5, it does not follow that a specific hypothesis we might choose–say, your blood pressure drug is effective–has a probability of .5 of being true, for a frequentist-

- Other problems arise is using the terms from significance tests for FFR or PPV assessments: Pr(+|D) and Pr(+|~D) in the DS criticism.

The DS model of tests considers just two possibilities "no effect" and "real effect".

H_0: 0 effect ($\mu = 0$),
H_1: the discrepancy against which the test has power $(1 - \beta)$.

It is assumed the probability for finding any effect, regardless of size, is the same.

$[\alpha/(1 - \beta)]$ used as the likelihood ratio to get a posterior of H_1

If the H_1 for which $(1 - \beta)$ is high, they take it as high likelihood for H_1

That's why this is on a chapter on power.

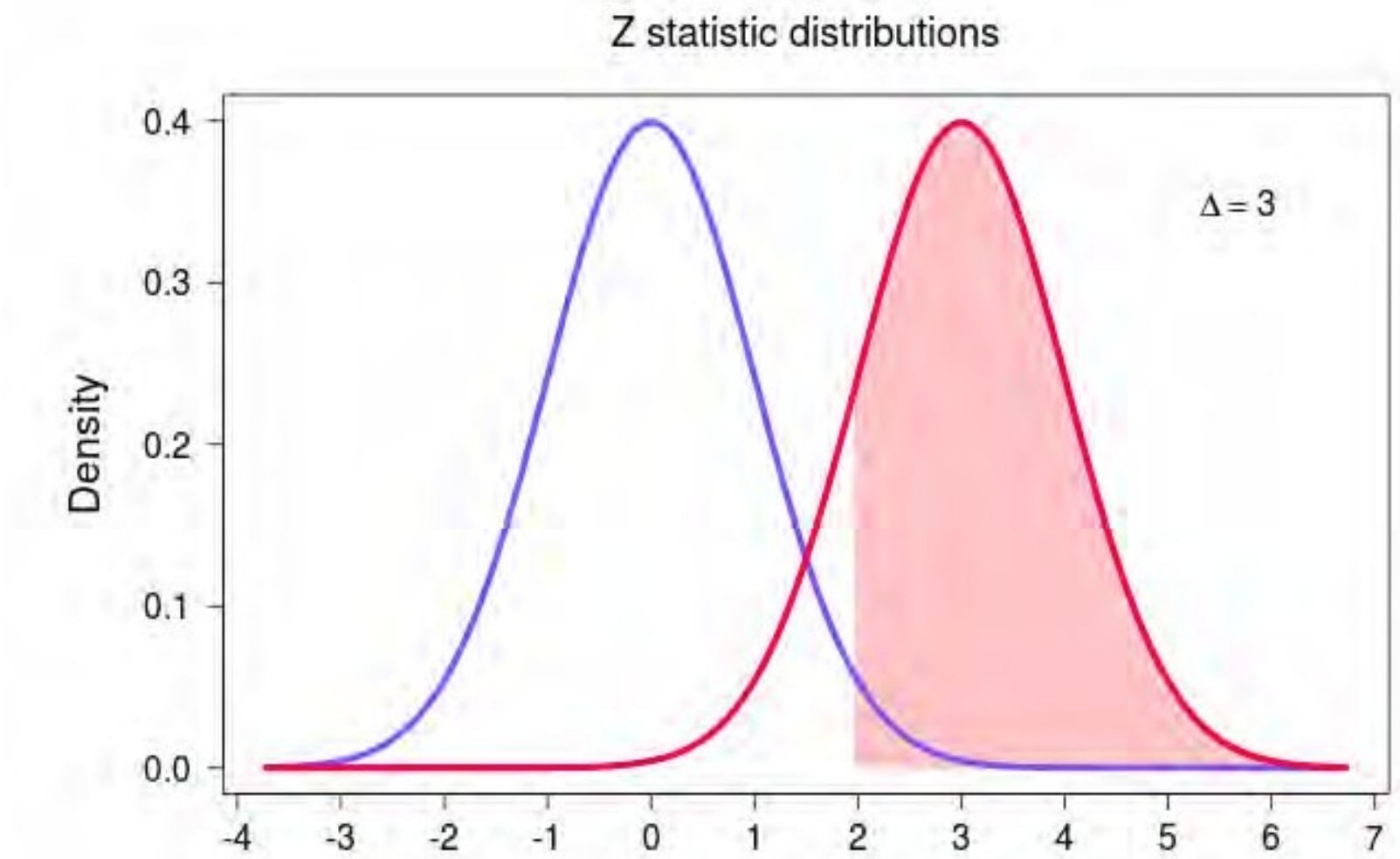
For an H_1 where (1 – β) is high, take our H_1

H_1: $\mu \geq \mu_{.84}$

$\mu_{.84}$ is the alternative against which the test has .84 power.

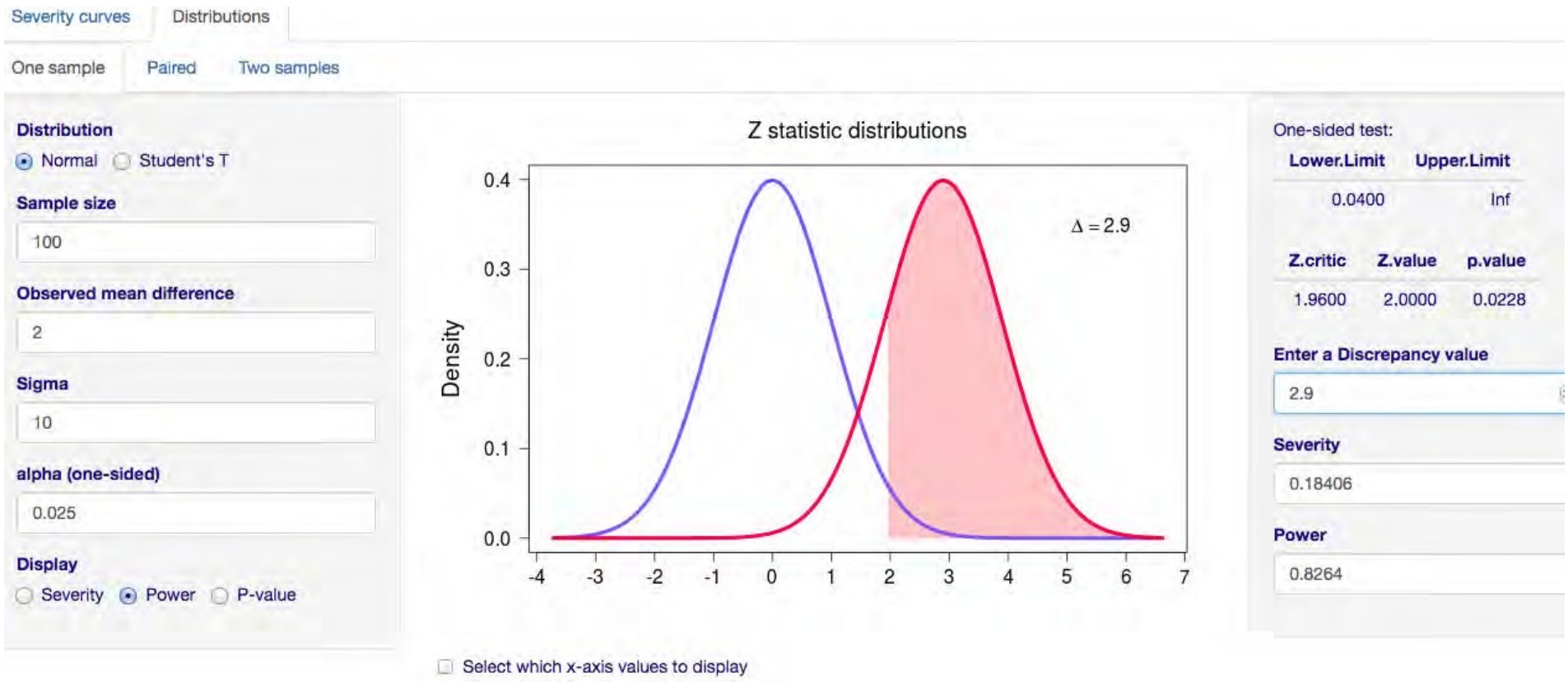
But now the denial of the alternative H_1 is not the same null hypothesis used to get *Type I error probability of .05.*

Instead it would be high, nearly as high as .84.



alternative is $\mu_{.84}$ (3, in our example)

e.g., let alternative be 2.9, *Type I error probability .82*



Likewise if the null $\mu \leq \mu_0$ is to have low α, its denial won't be one against which the test has high power (it will be close to α).

High power requires a μ exceeding the cut-off for rejecting *at level* α

> *We have* to assume they have in mind a test between a point null H_0, or a small interval around it, and a *non-exhaustive* alternative hypothesis H_1: $\mu = \mu_{.84}$

Problem: To infer $\mu_{.84}$ based on α = .025 (one-sided) is to be wrong 84% of the time.

We'd expect a more significant result 84% of the time were $\mu_{.84}$.

> Same problem as with Johnson.

Back to the more general problem with the DS model

Is the PPV computation *relevant* to what working scientists want to assess: strength of the *evidence* for effects or its degree of corroboration?

Crud Factor. In many fields of social and biological science it's thought nearly everything is related to everything: "all nulls false".

> These relationships are not, I repeat, Type I errors. They are facts about the world, and with N – 57,000 they are pretty stable. Some are theoretically easy to explain, others more difficult, others completely baffling. The 'easy' ones have multiple explanations, sometimes competing, usually not. (Meehl, 1990, p. 206).

He estimates the crud factor at around .3 or .4.

High prior prev gives high posterior prev

Will we be better able to replicate results in a field with a high crud factor?

By contrast: Even in a low prevalence situation, if I've done my homework, went beyond the one P-value, developed theories, I may have a good warrant for taking the effect as real.

Avoiding biasing selection effects and premature publication is what's doing the work, not prevalence.

The PPV doesn't tell us how valuable the statistically significant result is for predicting the truth or reproducibility of *that effect*.

We want to look at how well tested the particular hypothesis of interest is.

Suppose we find it severely tested.

Granted, we might assess the probability with which hypotheses pass so stringent a test, if false.

We have come full circle to evaluating the severity of tests passed. *Prevalence has nothing to do with it.*

The Dangers of the Diagnostic Screening Model for Science

Large-scale evidence should be targeted for research questions where the pre-study probability is already considerably high, so that a significant research finding will lead to a post-test probability that would be considered quite definitive (Ioannidis, 2005, p. 0700).

The DS model has mixed up the probability of a Type I error (often called the “false positive rate”) with the posterior probability: False Finding Rate FFR: Pr(H_0|H_0 is rejected).

In frequentist tests, reducing the Type II error probability results in *increasing* the Type I error probability: there is a trade-off.

In the DS model, the trade-off disappears: reducing the Type II error rate also reduces the FFR.

Diagnostic Screening, Probabilistic instantiation, base rates and all that

The computations for the DS model stem from Berger and Sellke (1987).

They claim it's just a heuristic, not that you'd use prevalences to assign priors.

FALLACIOUS ARGUMENT:

Pr(the randomly selected null hypothesis is true) = .5

The randomly selected null hypothesis is H_{51}

Pr(H_{51} is true) = .5

Each null either is true or not! My selecting it from an urn by means of a chosen selection procedure does not give evidence for its truth or probable truth

Equivocal:

I can model an experiment of selecting hypotheses from an urn: if it satisfies a Bernouilli model, I might say, the probability a (generic) outcome has the property (true) = the % true.

But the event (of being red, being true) isn't a statistical hypothesis; so isn't what you need for the likelihoods.

A statistical hypothesis H assigns probabilities to all possible outcomes

Consider this in relation to some criticisms of severity

A “hypothesis” that consists of asserting that a sample possesses a characteristic such as “having a disease” or “being college-ready.”

The point is to give it a frequentist prior.

Students From the Wrong Side of Town

Isaac, has passed comprehensive tests of mastery of high school subjects regarded as indicating college readiness S…

The battery of tests is assumed to be very capable of uncovering lack of readiness, so that such high scores *S* could very rarely result among high school students who are not sufficiently prepared to be deemed ‘college ready’.

Take S to be good evidence

H(I): Isaac is not deficient but is college ready.

And against

H'(I): Isaac's mastery of high school subjects is deficient, i.e., he is not college-ready.

Pr(*S*|*H(I)*: Isaac is college ready) ≈ 1, (practically 1)

Pr(*S*|*H*'(I): not college ready (i.e., deficient)) = .05 (very low)

- We should really consider degrees of readiness, but here I keep to the supposed counterexample.

Note: Thse numbers do not by themselves lead us to say H(I) has passed severely.

- Need to know of selection effects
- Not to check that they translate into a process that probes readiness, not an "isolated result"

Suppose a case where *H*(I) is warranted by dint of scores S.

"But wait a minute!" says the critic, Isaac was randomly selected from a population wherein college-readiness is exceedingly rare, Fewready Town where only 1 in 1000 are college ready. e.g.,

(*) P(*H(I)*) =.001.

Thus the posterior probability for *H(I)* is still low and *H*' (I)(deficient), the posterior is high.

e.g., Pr(*H'*(I)|S) = .95.

$$\text{PPV: } \Pr(D|+) = \frac{\Pr(+|D)\,\Pr(D)}{[\Pr(+|D)\,\Pr(D) + \Pr(+|\sim D)\,\Pr(\sim D)]}$$

$$= \frac{1}{(1+B)}$$

where $$B = \frac{\Pr(+|\sim D)\,\Pr(\sim D)}{\Pr(+|D)\,\Pr(D)}$$

D: ready ~D: not-ready Pr(D) = .001 Pr(~D) = .999
Pr(e| D) = 1 Pr(e| ~D) = .05. (D is null hyp in Howson)

Pr(~D| e) = 1/(1 + B). = 1/51 = .02

B = (.05)(.999)/(1)(.001) =.05/.001 = 50

Fallacy of probabilistic instantiation

The critic – for example, Howson, Achinstein – sees the conclusion as problematic for the severity account as, it's assumed the frequentist would also accept ($*$) P(*H(I)*) = .001.

Although the probability of college readiness in a randomly selected student from high schoolers from Fewready Town is .001, it doesn't follow that Isaac, the one we happened to select, has a probability of .001 of being college-ready

To suppose it does is to commit a kind of a fallacy of division:

The prevalence of readiness in Fewready Town is low. Isaac comes from Fewready Town

Thus, there's a low probability that Isaac is ready

We need not preclude that H(I) has a legitimate frequentist prior; it might refer to generic and environmental factors that determine the chance of his deficiency

Achinstein's "response to the probabilistic fallacy charge is to say that it would be true if the probabilities in question were construed as relative frequencies. [but] I am concerned with epistemic probability."

***Achinstein's Rule for Objective Epistemic Probabilities*:** If (we know only that) Isaac is randomly selected from a population where p% have property C, then the objective epistemic probability that Isaac has C equals p.(2010, p. 187)

"If all we know is that Isaac was chosen at random from a very disadvantaged population, very few of whose members are college ready, say one out of one thousand,then we would be justified in believing that it is very unlikely that Isaac is college-ready"

(i.e., Pr(H(I)) = .001 and, hence Pr(H(I)|S) is very low)

Even though Pr(*H(I)*|S) has increased from P(*H(I)*)

For Achinstein, unless the posterior reaches a threshold of a fairly high number, he claims, the evidence is “lousy.”

The example considers only two outcomes: reaching the high scores or not, i.e., S or ~S.

Clearly a lower grade gives even less evidence of readiness; that is, $Pr(H(I)|\sim S) < Pr(H(I)|S)$

Therefore, whether Isaac scored a high score or not, Achinstein’s epistemic probabilist reports justified high belief that Isaac is not ready.

The probability of Achinstein finding evidence of Isaac's readiness even if in fact he is ready (H is true) is zero.

Therefore, Achinstein's account violates what we have been calling the most minimal principle for evidence:

- The weak severity principle: Data ***x*** fail to provide good evidence for the truth of *H*' if the inferential procedure had very little chance of providing evidence against *H*', even if *H*' is false.

Reverse discrimination?

- If Isaac had been selected from a population where college-readiness is common, Manyready suburbs, the same set of passing scores *S* would be regarded as strong evidence for *H(I)*, Isaac being ready.
- Using this way of *evaluating evidence*, a high school student would have to have scored quite a bit higher on these tests than one selected from the affluent neighborhood in order for his scores to be considered evidence for his readiness!

I consulted with Lehmann after the first round of examples in 1996-7

- I was visiting him in Princeton where his wife J. Shaffer was at the Institute for Educational Testing Service, and this type of case could arise in policy-making
- He said: the test hasn't done its job if it can't make distinction in cases of rare diseases or rare assets

Actually if you actually had a large proportion of unready students amongst those who get passing scores S, there would be many reasons to deem the tests too lax for severity to be satisfied for Isaac.

So the prior enters, and is grounds to question those numbers

Severity for Problem-Solving (p. 300): Souvenir U

Note that there's no reason the problem at hand can't be providing an ordinary conditional probability

Severity then enters to assess if there is adequate warrant to take the problem as solved

The Case of General Hypotheses

When we move from hypotheses like “Isaac is college-ready” (which are really events) to generalizations – which Achinstein makes clear he regards as mandatory the difficulty for obtaining epistemic probabilities via his frequentist straight rule become more serious

The percentages “initially true” will vary considerably, and each would license a distinct “objective epistemic” prior.

The problems with the diagnostic screening.

Fisher: The Function of the p-value is Not Capable of Finding Expression…..

Discussing a test of the hypothesis that the stars are distributed at random, Fisher takes the low p-value (about 1 in 33,000) to "exclude at a high level of significance any theory involving random distribution" (Fisher (1956), p. 42).

Even if one were to imagine that H_0 had an extremely high prior probability, Fisher continues,-- never minding "what such a statement of probability a priori could possibly mean "—the resulting high posteriori probability to H_0, would only show that "our reluctance to accept a hypothesis strongly contradicted by a test of significance"… (Fisher (1956), p. 45) "is not capable of finding expression in any calculation of probability a posteriori" (Fisher (1956), p. 43).

Indeed, if one were to consider the claim about the a priori probability to be itself a hypothesis, Fisher suggests, it would be rejected by the data.

So, if there were a case where H severely passes test T with x, and yet the posterior of H given x is low, we are free to take this as evidence that the posterior fails to do the job demanded by the severity requirement.

Conflicts Between Posteriors & P-values

David Bickel (2021) argues that "if the *p*-value is sufficiently small while the posterior probability according to a model is insufficiently small, then the model will fail a model check" and may need revision (p. 249). In this view, akin to Fisher (1956), conflicts between posteriors and p-values may be resolved by revising the Bayesian model.

1. The priors given in all the examples I've seen are not legitimate frequentist priors for the statistical hypothesis being tested;

2. Using the procedures recommended by those who advocate using those priors, we would endorse inferences that fail on severity grounds.

3. If there is a legitimate frequentist prior for the *H* under test, then, we would assign those posteriors in the same way we assign probability to events—values of random variables.

4. If the posterior probability of "not-*H*" is high even though I regard *H* as having passed a severe test, it might well be seen as showing the inability of the posterior probability concept to capture the notion of evidence held by a severe tester! (R.A. Fisher's position).

I think many people who think they want a probability of a hypothesis really just want ordinary probabilities of events

The way to get them on the error statistical account is to obtain good evidence for the statistical hypothesis (or model) that assigns these probabilities!

Take a common linear regression

model M: $Y = a + bX_1 + cX + u$

This might let you predict the expected value of Y given a value for X, e.g., salary, given numbers of years of training, sex, etc. given the model has passed severely.

SIST p. 351 SIN & SIR

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.

- (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]

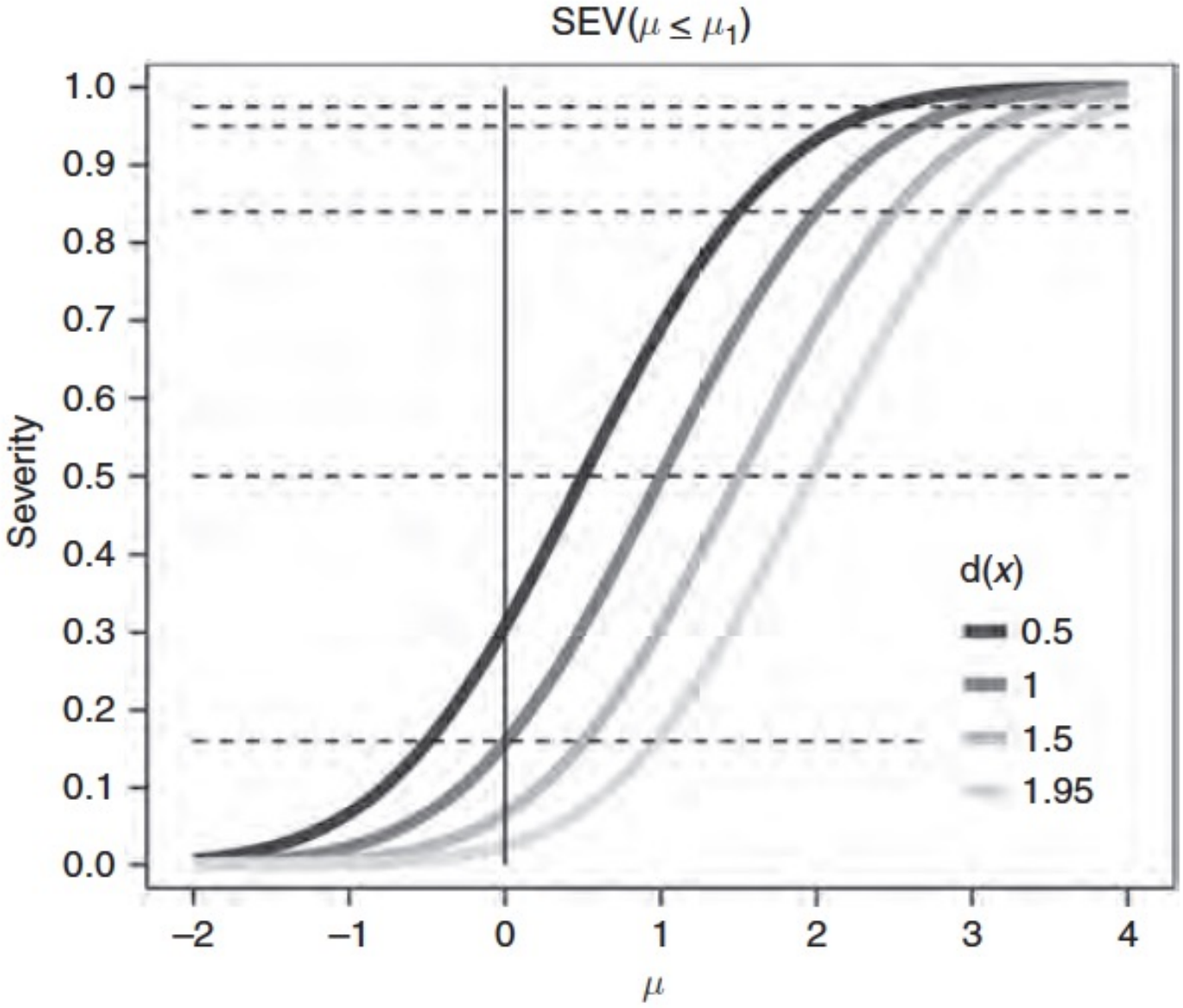


Figure 5.6 Severity curves.

REVIEW: Rigging EC from Assignment 1

1. What's the difference between weak and strong severity? (14, 23). You might wish to give an example of each from Excursion 1, or discuss key concepts such as: linked vs convergent arguments, arguing from coincidence, arguing from error.

Extra Credit for question #1: Choose one:

(i) Although one could stop after weak severity, the severe tester accepts strong severity as well. Why?

(ii) While an ultra-skeptic can always invent "rigged" hypotheses, how might this be criticized using weak severity alone?(108)