Day 12 (A) Aug 8
Power, Shpower, Attained Power, Diagnostic Screening

Negative results: $d(\boldsymbol{x}_0) \leq c_\alpha$:

(**SIST** 339, from Excur 5 Tour I)

A classic fallacy is to construe no evidence against H_0 as evidence of the correctness of H_0.

A canonical example in the list of slogans opening this book:

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

(improves on Cohen a little, he imagines we can identify a "negligible discrepancy)

Here we infer: discrepancy < γ

Probem: Too Coarse

Consider test T+ ($\alpha = .025$): H_0: $\mu = 0$ vs. H_1: $\mu \geq 0$, $\alpha = .025$, $n = 100$, $\sigma = 10$, $\sigma_{\bar{X}} = 1$. Say the cut-off must be $> \bar{x}_{.025} = 2$.

Consider an arbitrary inference $\mu < 1$.

We know POW(T+, $\mu = 1$) = .16 ($1\sigma_{\bar{X}}$ is subtracted from 2). .16 is quite lousy power.

It follows that no statistically insignificant result can warrant $\mu < 1$ for the power analyst.

Suppose, $\bar{x}_0 = -1$. This is $2\sigma_{\bar{X}}$ lower than 1. That should be taken into account.

We do. SEV(T+, $\bar{x}_0$ = -1, μ < 1) = .975.

Z = (-1 -1)/1 = -2
SEV (μ < 1) = Pr (Z > z_0; μ = 1) = .975

It would be even larger for values of μ smaller than 1

μ < 1 is also the upper one-sided CI bound at level .975

But it's not the one-sided CI corresponding to the test:
H_0: μ = 0 vs. H_1: μ ≥ 0

That's the one-sided <u>lower</u> (.975) bound

It matters because some people say you don't need to consider power if you've got CIs (discussed in the readings), since CIs have a duality with tests. (356-8)

Yes, but the test T+ corresponds to forming the one-sided lower bound.

One can look at the upper bound, but there needs to be a rationale for doing so.

(1) $P(d(X) > c_\alpha; \mu = \mu_0 + \gamma)$ Power to detect γ

- Just missing the cut-off c_α is the worst case
- It is more informative to look at the probability of getting a worse fit than you did

(2) $P(d(X) > d(x_0); \mu = \mu_0 + \gamma)$ "attained power"

a measure of the **severity** (or degree of corroboration) for the inference $\mu < \mu_0 + \gamma$

(1) can be low while (2) is high

Not the same as something called "retrospective power" or "ad hoc" power! (There μ is identified with the observed mean)

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 a 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 $\mu = \bar{x}_0$).

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

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

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:

$$POW(T+, \mu_0) = Pr(\bar{X} > \bar{x}_\alpha; \mu_0), \bar{x}_\alpha = (\mu_0 + z_\alpha \sigma_{\bar{X}}), \sigma_{\bar{X}} = [\sigma/\sqrt{n}])$$

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

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

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

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 μ' 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.

You have grounds to question the low power computation because you question the reported error probabilities, be it due to selective reporting, publication bias, or violated statistical assumptions. (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.

But it gets worse. Even the single significant result is very often the result of the cherry-picking, and barn-hunting we are all too familiar with.

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

The DS criticism of tests shows that if:

1. you publish upon getting a single P-value < .05,
2. you dichotomize tests into "up-down" outputs rather than report discrepancies and magnitudes of effect,
3. you data dredge, and cherry-pick and/or
4. there is a sufficiently low prevalence of genuine effects in your field

then the proportion of true nulls among those found statistically significant– (FFR)–differs from and can be much greater than the Type I error set by the test.

For the severe tester, committing #3 alone is suspect, unless we adjustment to get proper error probabilities

High prevalence of true hypotheses in your field should not atone for this sin.

Diagnostic Screening

- *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 you picked)= .9, Pr(H_1) = .1*
- *Rejecting H_0 with a single (just .05) significant result, cherry picking to boot*

The unsurprising result is that most "findings" are false:
$Pr(H_0|$ 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). To get the (PPV) we are to apply Bayes' rule using the given relative frequencies (or prevalences):

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

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

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

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

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

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.

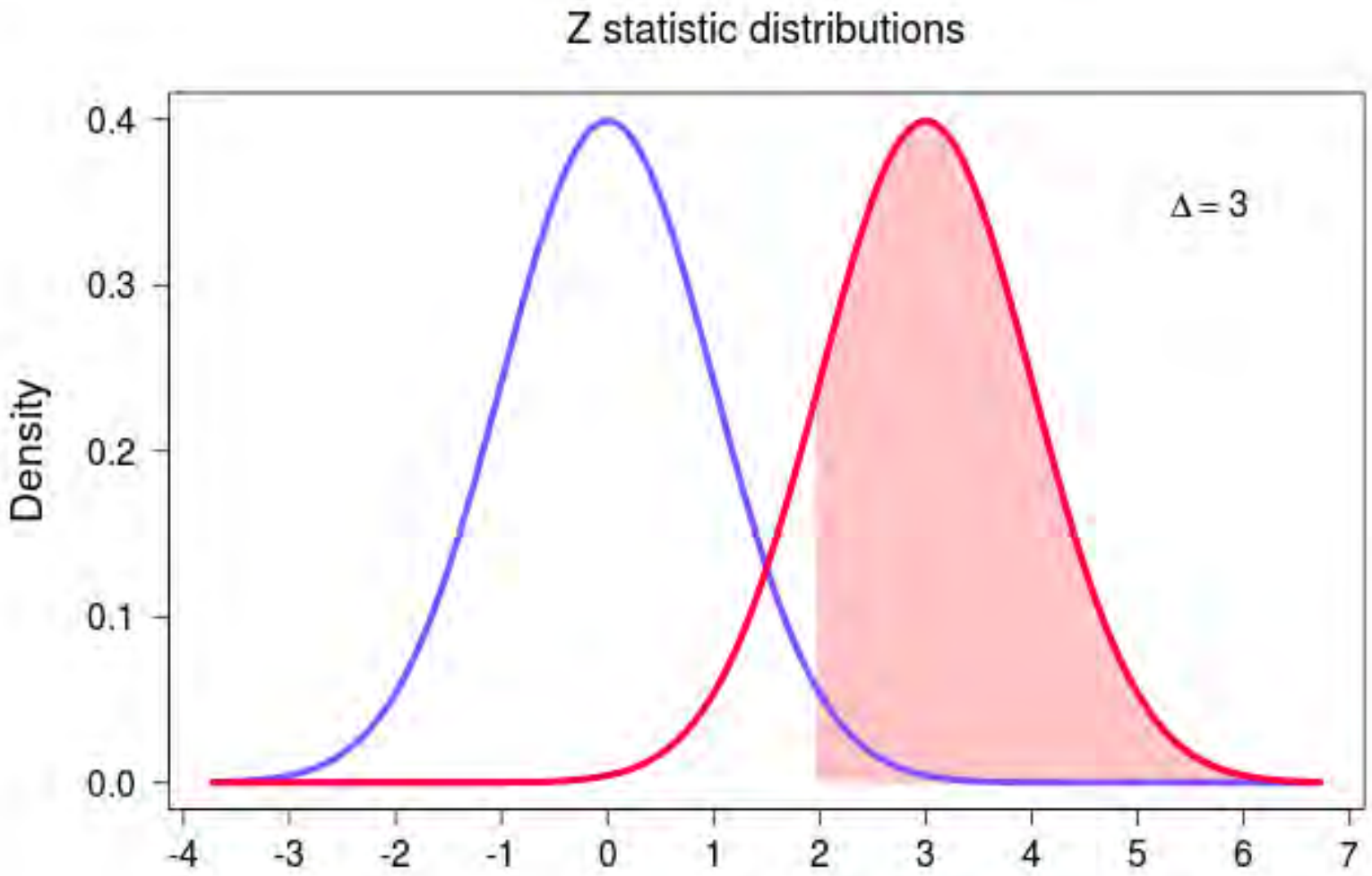
For an H_1 where (1 – β) is high, take our H_1

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

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

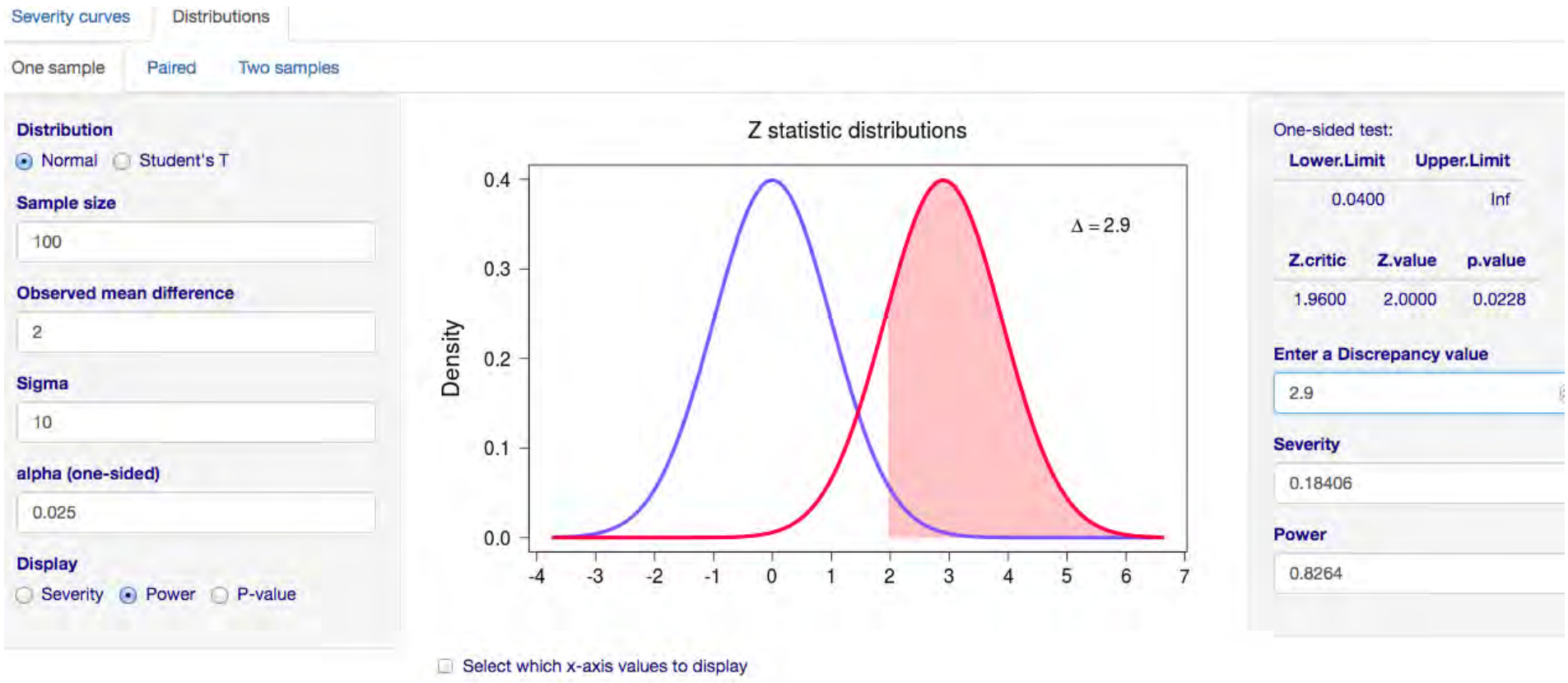
But now the denial of the alternative H_1 is not the same null hypothesis used to get *Type I error probability of .05.*

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 $\alpha = .025$ (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.