Excursion 5 Tours I Power: Pre-data, Post-data & How not to corrupt power

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)

- You won't find it in the ASA P-value statement.

- Power is one of the most abused notions in all of statistics (we've covered it, but are doing a bit more today)

- Power is always defined in terms of a fixed cut-off c_α, computed under a value of the parameter under test

These vary, there is really a power function.

- The *power* of a test against μ', is the probability it would lead to rejecting H_0 when μ = μ'. (3.1)

$POW(T, \mu') = Pr(d(\boldsymbol{X}) > c_\alpha; \mu = \mu')$

Fisher talked sensitivity, not power:

Oscar Kempthorne (being interviewed by J. Leroy Folks (1995)) said (SIST 325):

> "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 abut power, Fisher couldn't bring himself to acknowledge it" (p. 331).

Errors in Jacob Cohen's definition in his *Statistical Power Analysis for the Behavioral Sciences* (SIST p. 324)

Power: POW(T, μ') = Pr(d(***X***) > c_α; μ = μ')

- Keeping to the fixed cut-off c_α is too coarse for the severe tester—but we won't change the definition of power

"

N-P gave three roles to power:

- first two are pre-data, for planning, comparing tests; the third for interpretation post-data—to be explained in a minute

(Hidden Neyman files, from R. Giere collection).
Mayo and Spanos (2006, p. 337)

5.1 Power Howlers, Trade-offs and Benchmarks

Power is increased with increased *n*, but also by computing it in relation to alternatives further and further from the null.

- ***Example.*** A test is practically guaranteed to reject H_0, the "no improvement" null, if in fact H_1 the drug cures practically everyone. (SIST p. 326)

It has high power to detect H_1
But you wouldn't say that its rejecting H_0 is evidence H_1 cures everyone.

To think otherwise is to commit the second form of MM fallacy (p. 326):

> *Mountains out of Molehills (MM) Fallacy (second form)* Test T+*:* The fallacy of taking a just 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.

"This is a surprisingly widespread piece of nonsense which has even made its way into one book on drug industry trials" (ibid., p. 201).(bott SIST, 328)

Trade-offs and Benchmarks

a. The power against H_0 is α.

POW(T+, μ_0) = Pr($\bar{X} > \bar{x}_\alpha$; μ_0), $\bar{x}_\alpha = (\mu_0 + z_\alpha \sigma_{\bar{X}})$, $\sigma_{\bar{X}} = [\sigma/\sqrt{n}]$)

The power at the null is: Pr($Z > z_\alpha$;μ_0) = α.

It's the low power against H_0 that warrants taking a rejection as evidence that $\mu > \mu_0$.

We infer an indication of discrepancy from H_0 because a null world would probably have yielded a smaller difference than observed.

Trade-offs and Benchmarks

Let $\bar{x}_\alpha$ = the cut-off for rejection at the α level

$\bar{x}_\alpha$ =($\mu_0 + z_\alpha \sigma_{\bar{X}}$), . $\sigma_{\bar{X}}$ = [$\sigma/\sqrt{n}$]) i.e.,SE

Our usual test T+: $\mu = \mu_0$ versus $\mu > \mu_0$

IN GENERAL POW(μ') = $Z > z_\alpha$

Z = $\overline{(x_\alpha}$ **– μ')/SE**

The power at the null μ_0 is: $Pr(Z > z_\alpha; \mu_0) = \alpha$.

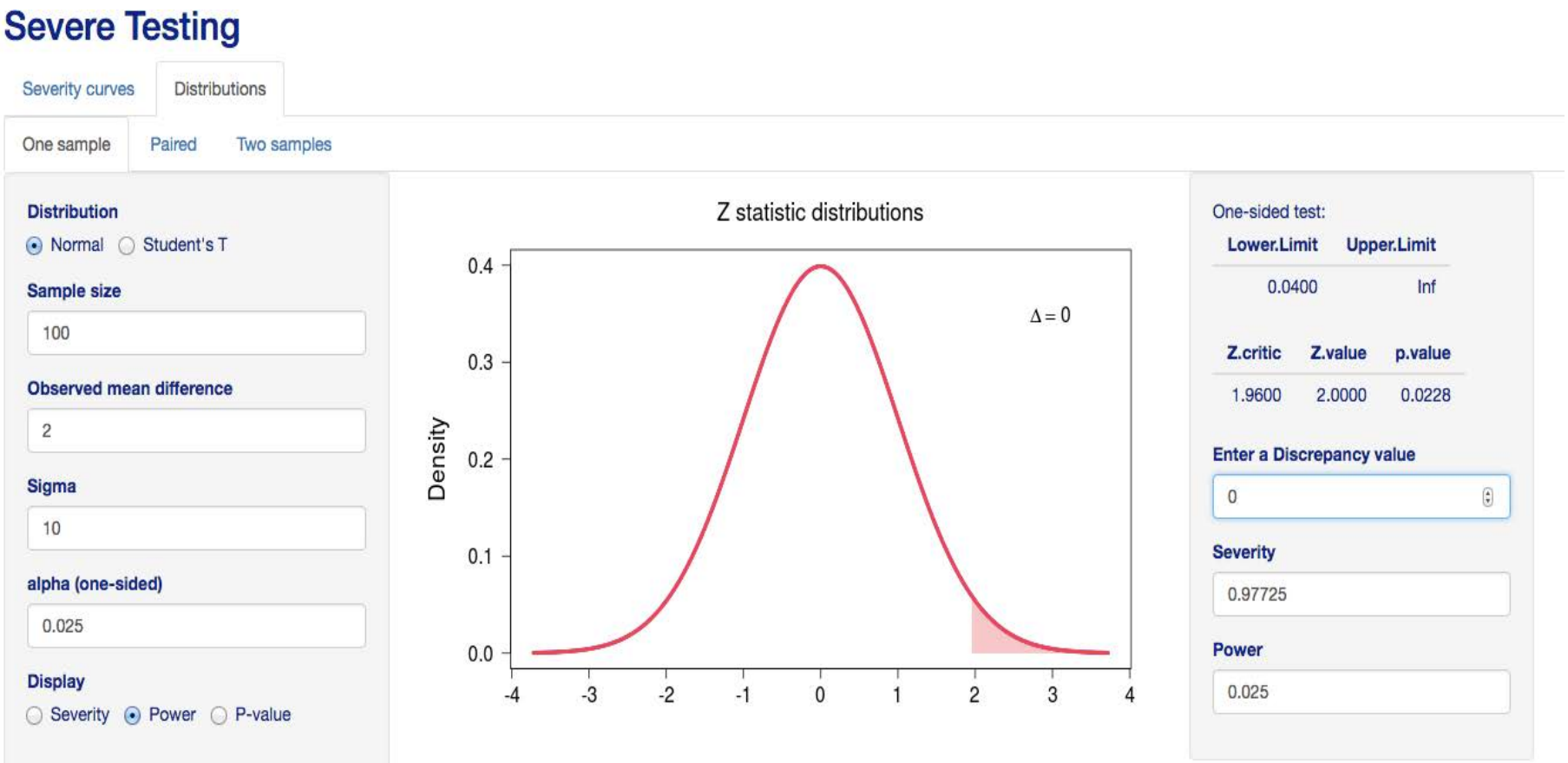
Let $\bar{x}_\alpha$ = the cut-off for rejection at the α level

$\bar{x}_\alpha$ =($\mu_0 + z_\alpha \sigma_{\bar{X}}$), $\sigma_{\bar{X}}$ = [$\sigma/\sqrt{n}$]) i.e., SE

POW(μ_0)

Z = $\overline{(x_\alpha} - \mu_0$)/SE

The power at the null μ_0 is: Pr($Z > z_\alpha; \mu_0$) = α.



Example 1: Left Side: Sample size: 100; Observed mean difference (from null): 2; α: 0.025

Right side: "discrepancy value" is 0. Power is .025 (same as α)

POW $(\bar{x}_{\alpha})$ **?**

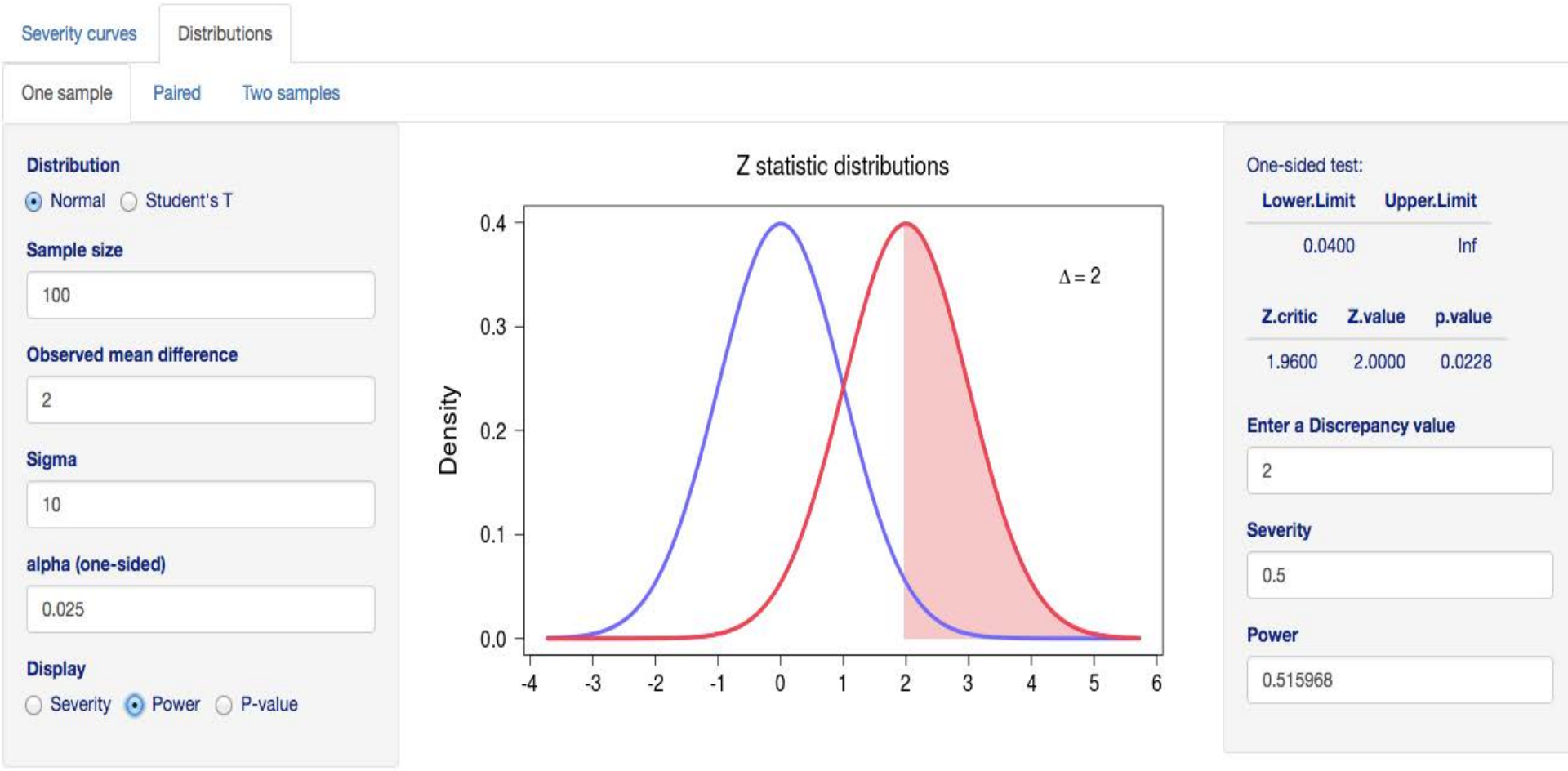
Let $\bar{x}_{\alpha}$ = the cut-off for rejection at the α level

Z = $(\bar{x}_{\alpha} - \bar{x}_{\alpha}$ **)/SE**

The power at $\bar{x}_{\alpha}$ is Pr(Z > 0) = .5

b. *The power of T+ for* $\mu_1 = \bar{x}_\alpha$ *is .5.* Here, Z = 0, and Pr(Z > 0) = .5, so:

POW(T+, $\mu_1 = \bar{x}_\alpha$) = .5.



discrepancy = 2,power is ~0.5

b. The power > .5 only for alternatives that exceed the cut-off $\bar{x}_\alpha$,

Remember $\bar{x}_\alpha$ is ($\mu_0 + z_\alpha \sigma_{\bar{X}}$).

The power of test T+ against $\mu = \bar{x}_\alpha$ is .5.

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

An easy alternative to remember with reasonable high power (SIST 329): $\mu_{.84}$:

Abbreviation: the alternative against which test T+ has .84 power by $\mu_{.84}$:

The power of test T+ to detect an alternative that exceeds the cut-off $\bar{x}_\alpha$ by $1\sigma_{\bar{X}}$ =.84.

Other shortcuts on SIST p. 328

The power > .5 only for alternatives that exceed the cut-off $\bar{x}_{\alpha}$,
We get the shortcuts on **SIST** p. 328

Remember $\bar{x}_{\alpha}$ is ($\mu_0 + z_{\alpha}\sigma_{\bar{X}}$).

marcosjnez.shinyapps.io/Severity/

Trade-offs Between α, the Type I Error Probability and Power

As the probability of a Type I error goes down the probability of a Type II error goes up (power goes down).

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! So they're either using a different notion or are wrong about power.

Many current reforms do just this!

Criticisms that lead to those reforms also get things backwards

Ziliak and McCloskey “refutations of the null are trivially easy to achieve if power is low enough or the sample is large enough” (2008a, p. 152)?

They would need to say power is high enough raising the power is to lower the hurdle, they get it backwards (SIST p. 330)

More howlers on p. 331

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.

338 Power analysis arises to interpret negative results: $d(x_0) \leq c_\alpha$:

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

- "Researchers have been warned that a statistically nonsignificant result does not 'prove' the null hypothesis (the hypothesis that there is no difference between groups or no effect of a treatment …)".

Amrhein et al., (2019) take this as grounds to "**Retire Statistical Significance**"

- No mention of power, designed to block this fallacy

It uses the same reasoning as significance tests. Cohen:

> [F]or a given hypothesis test, one defines a numerical value **i** (or *i*ota) for the [population] ES, where **i** is so small that it is appropriate in the context to consider it negligible (trivial, inconsequential). Power (1 – β) 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**).

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

Neyman an early power analyst

In his “The Problem of Inductive Inference” (1955) where he chides Carnap for ignoring the statistical model (p. 341).

“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 26 observations come from a source in which the null hypothesis is true]**”.**

"Locally best one-sided Test T

A sample **X** = (X_1, …,X_n) each X_i is Normal, N(μ,σ^2), (NIID),
σ assumed known; $\bar{X}$ the sample mean

H_0: $\mu \leq \mu_0$ against H_1: $\mu > \mu_0$.

Test Statistic $d(\mathbf{X}) = (\bar{X} - \mu_0)/\sigma_{\mathbf{x}}$,
$\sigma_{\mathbf{x}} = \sigma / \sqrt{n}$

Test fails to reject the null, $d(\boldsymbol{x}_0) \leq c_\alpha$.
"The question is: does this result 'confirm' the hypothesis that H_0 is true [of the particular data set]? " (Neyman).

Carnap says yes…

Neyman:
"….the attitude described is dangerous.
…the chance of detecting the presence [of discrepancy γ from the null], when only [this number] of observations are available, is extremely slim, even if [γ is present]."

"One may be confident in the absence [of that discrepancy only] if the power to detect it were high". (power analysis)

If $Pr(d(\boldsymbol{X}) > c_\alpha; \mu = \mu_0 + \gamma)$ is high

$d(\boldsymbol{X}) \leq c_\alpha$;

infer: discrepancy < γ

Problem: Too Coarse

Consider test T+ (α = .025): H_0: μ = 150 vs.
H_1: μ ≥ 150, α = .025, n = 100, σ = 10, $\sigma_{\overline{X}}$ = 1.
The cut-off = 152.

Say $\bar{x}_0$ =151.9, just missing 152, treated the same as smaller ones, say 149

Consider an arbitrary inference μ < 151.

We know POW(T+, μ = 151) = .16
(1$\sigma_{\overline{X}}$ is subtracted from 152).
.16 is quite lousy power.

It follows that no statistically insignificant result can warrant μ< 151 for the power analyst.

Problem: Too Coarse

Consider test T+ (α = .025): H_0: μ = 150 vs. H_1: μ ≥ 150, α = .025, n = 100, σ = 10, $\sigma_{\bar{X}}$ = 1. The cut-off = 152.

We know POW(T+, μ = 151) = .16

Z = (152 – 151)/1 = 1

Pr(Z > 1) = .16

.16 is quite lousy power.

It follows that no statistically insignificant result can warrant μ< 151 for the power analyst.

We should take account of the actual result:

SEV(T+, $\bar{x}_0$ = 149,μ < 151) = .975.

Z = (149 -151)/1 = -2

SEV (μ < 151) = Pr (Z > z_0; μ = 1) = .975

$SEV(\mu \leq 151) = 1 - SEV(\mu > 151)$

$Z = (149 - 151)/1 = -2$

$SEV(\mu < 151) = Pr(Z > z_0; \mu = 1) = .975$

If the test fails to reject, we look at Pr (test would have resulted in a larger difference than it did)

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

- Just missing the cut-off $\mathbf{c_\alpha}$ 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" $\Pi(\gamma)$

Here it measures the **severity** for the inference $\mu < \mu_0 + \gamma$

Not the same as something called "retrospective power" or "ad hoc" power!

Π(γ) = “sensitivity achieved” p. 151 (Mayo and Cox 2006)

Here it measures the **severity** for the inference

$\mu < \mu_0 + \gamma$

Not the same as something called “retrospective power” or “ad hoc” power!

Sensitivity Achieved or Attained

For a Fisherian like Cox, a test's power only has relevance pre-data, in planning tests, but, like Fisher, he can measure "sensitivity":

> In the Neyman–Pearson theory of tests, the sensitivity of a test is assessed by the notion of *power*, defined as the probability of reaching a preset level of significance . . . for various alternative hypotheses. In the approach adopted here the assessment is via the distribution of the random variable P, again considered for various alternatives. (Cox 2006a, p. 25)

This is the key: Cox will measure sensitivity by a function we may abbreviate as $\Pi(\gamma)$. Computing $\Pi(\gamma)$ may be regarded as viewing the P-value as a statistic. That is:

$$\Pi(\gamma) = \Pr(P \leq p_{\text{obs}}; \mu_0 + \gamma).$$

The alternative is $\mu_1 = \mu_0 + \gamma$. Using the P-value distribution has a long history and is part of many approaches. Given the P-value inverts the distance, it is clearer and less confusing to formulate $\Pi(\gamma)$ in terms of the test statistic d. $\Pi(\gamma)$ is very similar to *power* in relation to alternative μ_1, except that $\Pi(\gamma)$ considers the observed difference rather than the N-P cut-off c_α:

$$\Pi(\gamma) = \Pr(d \geq d_0; \mu_0 + \gamma),$$

$$\text{POW}(\gamma) = \Pr(d \geq c_\alpha; \mu_0 + \gamma).$$

Π may be called a "sensitivity function," or we might think of $\Pi(\gamma)$ as the "attained power" to detect discrepancy γ (Section 5.3). The nice thing about

The only Time Severity equals Power for a claim

$\bar{X}$ just misses $\bar{x}_\alpha$ and you want SEV(μ **<** μ')

Then it equals POW(μ')

For claims of form μ **>** μ' it's the reverse:

(the ex on p. 344 has different numbers but the point is the same:)

n=25, σ =1 SE = 1/sq root of 25 = .2

Power vs Severity for $\mu > \mu_1$

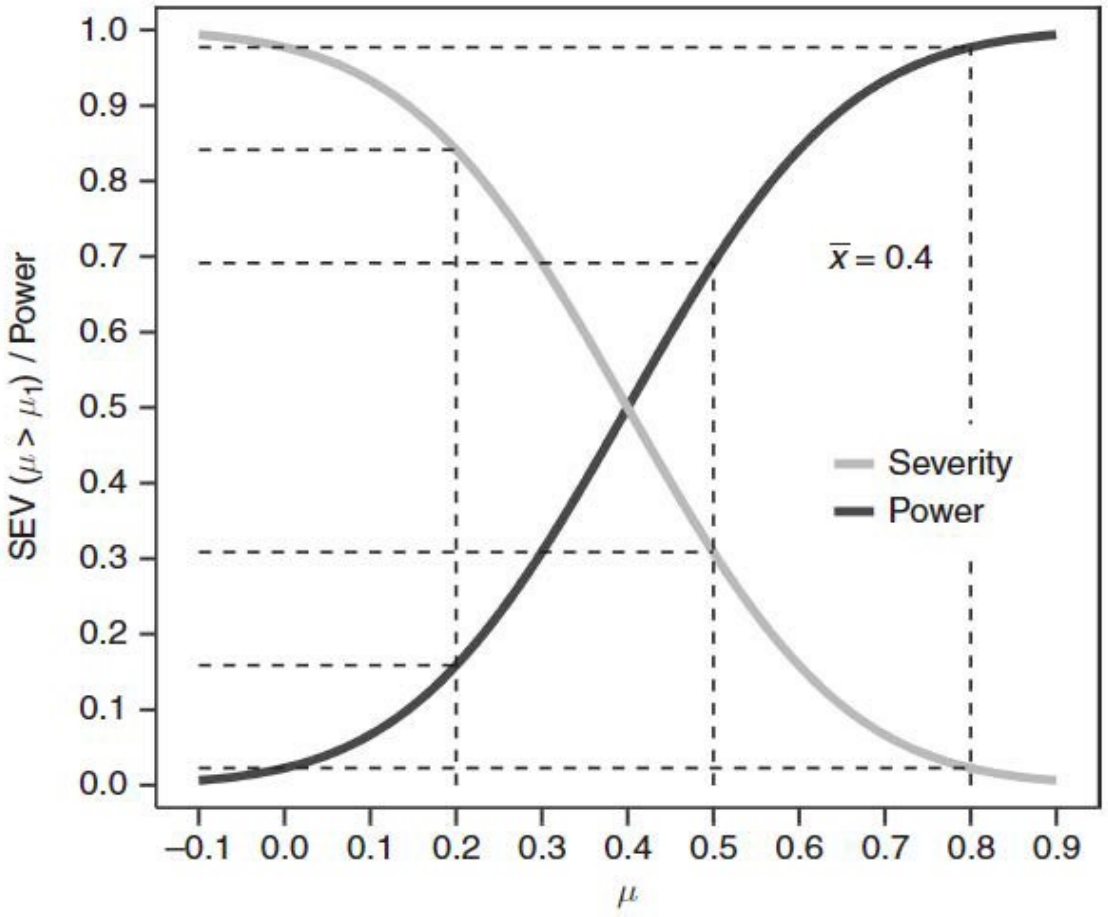


Figure 5.4 Severity for ($\mu > \mu_1$) vs power (μ_1).

Severity for (nonsignificant results) and confidence bounds

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

(SEV): If d(x) is not statistically significant, then test T+ passes $\mu < M_0 + k_\varepsilon \sigma / n^{.5}$ with severity (1 – ε),

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

The connection with the upper confidence limit is obvious.

One can consider a series of upper discrepancy bounds…

SEV($\mu < \bar{x}_0 + 0\sigma_x$) = .5

SEV($\mu < \bar{x}_0 + .5\sigma_x$) = .7

SEV($\mu < \bar{x}_0 + 1\sigma_x$) = .84

SEV($\mu < \bar{x}_0 + 1.5\sigma_x$) = .93

SEV($\mu < \bar{x}_0 + 1.96\sigma_x$) = .975

This relates to work on confidence distributions.

But aren't I just using this as another way to say how probable each claim is?

No. This would lead to inconsistencies (famous fiducial feuds)

(Excursion 5 Tour III: Deconstructing N-P vs Fisher debates

The reasoning instead is counterfactual:

H: $\mu \leq \bar{x}_0 + 1.96\sigma_{\bar{x}}$

(i.e., $\mu \leq CI_u$)

H passes severely because were this inference false, and the true mean $\mu > CI_u$ then, very probably, we would have observed a larger sample mean

Power vs Severity analysis for non-significant results

Power Analysis (ordinary): If Pr(d(**X**) > c_α; μ') = high and the result is not significant, then it's an indication or evidence that μ < μ' (or μ ≤ μ'.)

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

If Π(γ) is high it's an indication or evidence that μ < μ.'

Excursion 5 Tour II
Focus just on ordinary power analysis

"There's a sinister side to statistical power" (SIST, p. 354)

I've seen otherwise excellent books, say "Power analysis? Don't!"

I call it shpower analysis because it distorts ordinary power analytic reasoning from large P-values—negative results.

Excursion 5 Tour II
Shpower and Retrospective Power

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

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

Since alternative μ is set = $\bar{x}_0$, and $\bar{x}_0$ is given as statistically insignificant, the power can never exceed .5.

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

But power analytic reasoning is about finding an alternative against which the test has *high* capability to have obtained significance.

Neyman and Cohen focus on cases where there's high power to detect an effect deemed negligible, so you can infer evidence of "a negligible effect"

The logic lets you infer $\mu < \mu'$—the discrepancy or ES that probably would have led to a significant result is absent in shpower Tour II.

Else, just report you cannot rule out a non-negligible effect

Tiny illustration of Power & sample size

H_0: μ ≤ 150 vs. H_1: μ > 150
(Let σ = 10, ***n* = 100**)
let α = .025
POW(T+, μ_1) = Pr(Test T+ rejects H_0; μ_1),

Consider μ_1 = 153

POW(T+, 153) Pr($\bar{X}$ > 152; μ =

153) Z = (152 – 153)/ $\sigma_{\bar{X}}$ = -1
Pr (Z > -1) = .84

(Test has .84 power to detect 153 but observing observing M = 152 is poort's poor evidence μ > 153)

Illustration with ordinary power and sample size

H_0: μ ≤ 150 vs. H_1: μ > 150

(Let σ = 10, ***n* = 25**) Now = SE=2 (i.e., 10/5) let α = .025

POW(T+, μ_0) = Pr(Test T+ rejects H_0; μ_0),

Again consider μ_1 = 153

2SE cut-off is 154

POW(T+, 153) Pr($\bar{X} \geq$ 154; μ = 153)

Z = (154 – 153)/ $\sqrt{2}$ = .5

Pr (Z > .5) = .3

Illustration with ordinary power and sample size
H_0: μ ≤ 150 vs. H_1: μ > 150

(Let σ = 10, ***n* = 10,000**) Now = SE=.01

let α = .025
POW(T+, μ_0) = Pr(Test T+ rejects H_0; μ_0),

The 2SE cut-off is now 150.02!

Again consider POW(μ_1 = 153)

Z ~-3. so the POW ~1

***But reaching stat sig with n* = 10,000 is horrible evidence that** ($\mu_1 = 153$)

Now = SE=.01

let $\alpha = .025$
POW(T+, μ_0) = Pr(Test T+ rejects H_0; μ_0),

The 2SE cut-off is now 150.02!

Again consider POW($\mu_1 = 153$)

Z ~-3. so the POW ~1

Explaining a strange remark you may hear:
The higher the power the less likely a statistically significant result exaggerates the true mean.

p. 360 SIST refers to Gelman and Carlin

If you use the observed mean M_0 to estimate μ, then the larger the sample size, the smaller the stat sig M is, so it's not as big as with a smaller sample size

(you should use the lower CI bound to estimate μ

But when the sensitive smoke alarm goes off its less indicative of a fire than when the insensitive (small power) alarm goes off!