Excursion 5: Power and Severity

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? (p. 323)

Power is one of the most abused notions in all of statistics

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.

If someone speaks of the power of a test *tout court*, you cannot make sense of it, without qualification.

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

POW(T, μ') = Pr(d(***X***) > c_α; μ = μ'), or Pr(Test T rejects H_0; μ = μ').

Power measures the capability of a test to detect μ'–where the detection is in the form of producing a $d > c_\alpha$.

Power is computed at a point μ = μ', we use it to appraise claims of form μ > μ' or μ < μ'.

Power is an ingredient in N-P tests, but even Fisherians invoke power

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

Two errors in Jacob Cohen's definition in his (1969/1988) *Statistical Power Analysis for the Behavioral Sciences* (**SIST** p. 324)

Keeping to the fixed cut-off c_α is too coarse for the severe tester

We will see why in doing power analysis today.

The data-dependent version was in (3.3), but now we'll focus on it.

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

Achieved sensitivity" or "attained power"

$$\Pi(\gamma) = \Pr(d(\boldsymbol{X}) > d(\boldsymbol{x}_0); \mu')$$

$\mu' = \mu_0 + \gamma$

N-P accorded three roles to power: first two are pre-data, for planning, comparing tests; the third for interpretation post-data.

(I broke Tours I and II at the last minute)

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).

It's too performance oriented, Fisher claimed ~ 1955.

5.1 Power Howlers, Trade-offs and Benchmarks

In the Mountains out of Molehills (MM) Fallacy (4.3), an α-level rejection with a larger sample size (higher power) is taken as evidence of a greater discrepancy from the null hypothesis than with a smaller sample size (in tests otherwise the same).

Power can also be increased by computing it in relation to alternatives further and further from the null.

> *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.

(SIST 326)

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

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 statistical affirming the consequent–the basis for the MM fallacy.

Stephen Senn. In drug development, it is typical to set a high power of .8 or .9 to detect effects deemed of clinical relevance. Test T+: Reject H_0 iff $Z > z_\alpha$ (Z is the standard Normal variate).

A simpler presentation to use the cut-off for rejection in terms of $\bar{x}_\alpha$: Reject H_0 iff $\overline{X} > \bar{x}_\alpha = (\mu_0 + z_\alpha \sigma \sqrt{n})$.

Abbreviate: the alternative against which test T+ has .8 power by $\mu_{.8}$.
So POW($\mu_{.8}$) = .8.

Suppose $\mu_{.8}$ is the clinically relevant difference.

Can we say, upon rejecting the null hypothesis, that there's evidence the treatment has a clinically relevant effect, i.e., $\mu \geq \mu_{.8}$?

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

$\mu_{.8}$ > the cut-off for rejection, in particular, $\mu_{.8} = \bar{x}_{\alpha} + .85\ \sigma_{\bar{X}}$ (where $\sigma_{\bar{X}} = \sigma/\sqrt{n}$).

An easy alternative to remember: (**SIST** 329): $\mu_{.84}$:

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

The result of adding $1\sigma_{\bar{X}}$ to $\bar{x}_\alpha$: That takes us to a value of μ against which the test has .84 power: $\mu_{.84}$:

Trade-offs and Benchmarks

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

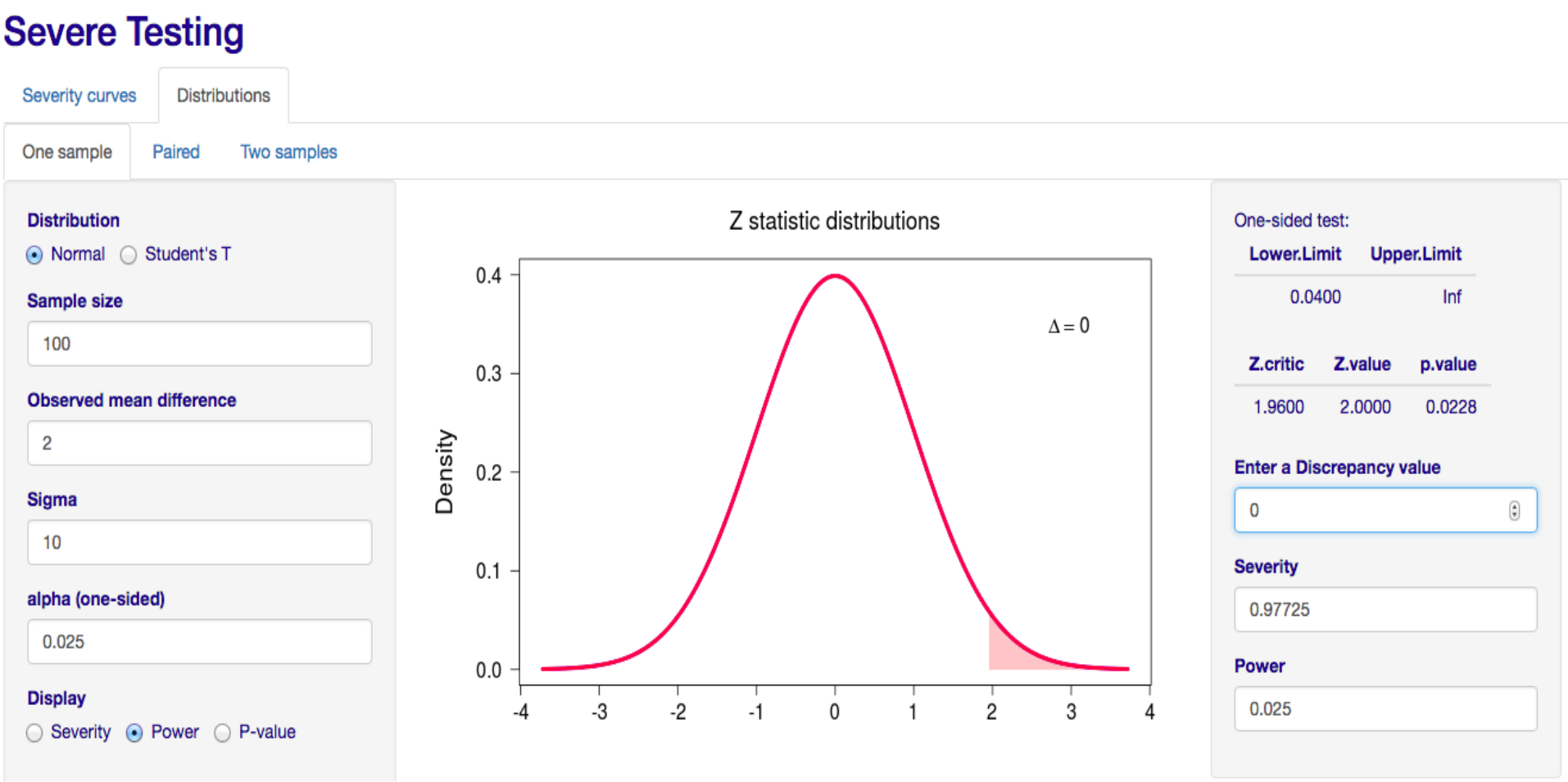
a. *The power against H_0 is α.* We can use the power function to define the probability of a Type I error or the significance level of the test:

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

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

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.

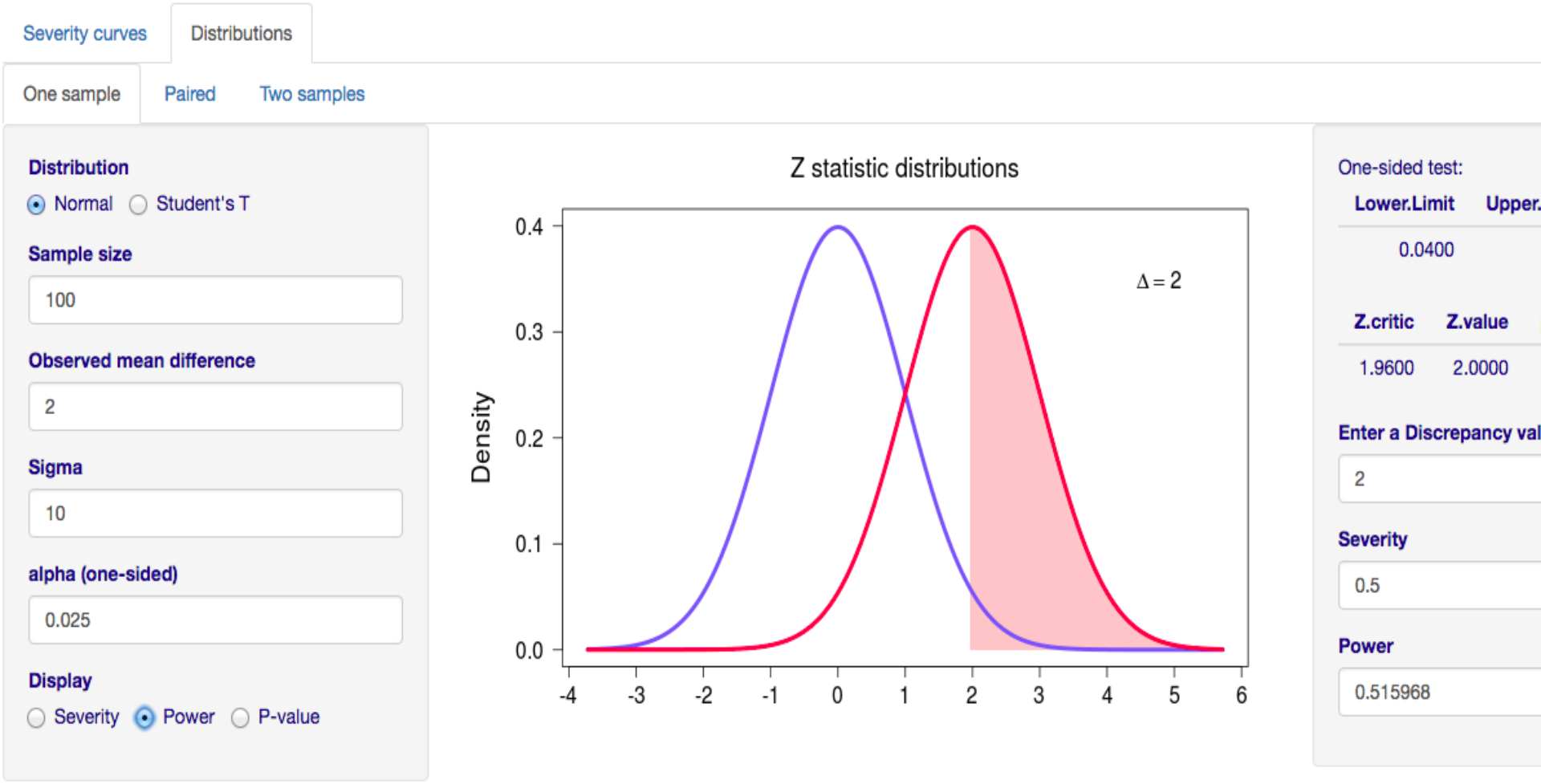


Example 1: Left Side: Sample size: 100; Observed mean difference (from null): 2; Alpha: 0.025

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

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

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

We know 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).

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!

Many current reforms do just this! After this class, you can readily be on the look-out, and refuse to be fooled.

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.

Exhibit (i). Here I let n = 25 in Test T+ (α = .025)
H_0: μ = 0 vs. H_1: μ ≥ 0, α = .025, n = 25, σ = 1.

But keep to n = 100
Say you must decrease the Type I error probability α to .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+ (α = .001), will be $\bar{x}_{.001}$.

Old cut off was 2, new cut-off is 3, it must be $3\sigma_{\bar{X}}$ greater than 0 rather than only $2\sigma_{\bar{X}}$:

$$\mu_{.5} = \bar{x}_{\alpha},$$

With $\alpha = .025$, the smallest alternative the test has 50% power to detect is $\mu_{.5} = 2$

With $\alpha = .001$, the smallest alternative the test has 50% power to detect is $\mu_{.5} = 3$

Decreasing the Type I error by moving the hurdle over to the right by $1\sigma_{\bar{X}}$ unit results in the alternative against which we have .5 power $\mu_{.5}$ *also moving over to the right by* $1\sigma_{\bar{X}}$.

We see the trade-off very neatly, at least in one direction.

Ziliak and McCloskey get their hurdles in a twist **SIST** p. 330-1,

Their slippery slides are quite illuminating.

> If the power of a test is low, say, .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, .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 2013, p. 132-3).

If the power of a test is high, then a rejection of the null is probably correct?

We follow our rule of generous interpretation (**SIST** 331)

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 are detected. That makes it easier for H_1.

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

(**SIST** 339)

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:

Power analysis 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 Chides Carnap, Again (**SIST** 341)*

In his "The Problem of Inductive Inference" (1955) where he chides Carnap for ignoring the statistical model (2.7).

"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]. The question is: does this result 'confirm' the hypothesis that H_0 is true of the particular data set]? ".

Ironically, Neyman (1957a,b) 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 (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 .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)

"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; M the sample mean

$$H_0: \mu \leq \mu_0 \text{ against } H_1: \mu > \mu_0.$$

Test Statistic $d(\mathbf{X}) = (M - \mu_0)/\sigma_{\mathbf{x}}$,

$\sigma_{\mathbf{x}} = \sigma/\sqrt{n}$

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 $< \gamma$

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

(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 $\mu < \mu_0 + \gamma$

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

Mayo and Spanos (2006, p. 337):

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

(SEV): If d(x) is <u>not</u> 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.

1.1. If one wants a post-data measure, one can write:

SEV($\mu < \mathbf{M_0} + \gamma\sigma_x$) to abbreviate:

The severity with which

($\mu < \mathbf{M_0} + \gamma\ \sigma_x$).

passes test T

It's computed $\mathbf{Pr(}\boldsymbol{d(X)} > \boldsymbol{d(x_0)}\mathbf{;}\ \mu = \mu_0 + \gamma\mathbf{)}$

Severity has 3 terms: SEV(Test, outcome, inference)

One can consider a series of upper discrepancy bounds…

$SEV(\mu < \mathbf{M_0} + 0\sigma_x) = .5$

$SEV(\mu < \mathbf{M_0} + .5\sigma_x) = .7$

$SEV(\mu < \mathbf{M_0} + 1\sigma_x) = .84$

$SEV(\mu < \mathbf{M_0} + 1.5\sigma_x) = .93$

$SEV(\mu < \mathbf{M_0} + 1.96\sigma_x) = .975$

This seems to relate to work by Min-ge Xie and others 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

Probability gives the wrong logic for "how well-tested"
(or "corroborated") a claim is

(there may be a confusion of ordinary language use of "probability":
belief is very different from well-testedness)

Note: low severity is not just a little bit of evidence, but bad evidence, no test (BENT)

The severity construal is different from what I call the

Rubbing off construal: The procedure is rarely wrong, therefore, the probability it is wrong in this case is low.

Still too much of a *performance* criteria, too *behavioristic*

The long-run reliability of the rule is a necessary but not a sufficient condition to infer *H* (with severity)

The reasoning instead is counterfactual:

$$H: \quad \mu \leq \mathbf{M_0} + 1.96\sigma_{\mathbf{x}}$$

$$(\text{i.e., } \mu \leq CI_u)$$

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:

What enables substituting the observed value of the test statistic, 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* of T+ against $\mu \leq \mu_0 + \gamma$ ($\Pi(\gamma)$) is very high, the inference to $\mu < \mu_0 + \gamma$ is warranted with severity.

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