Need to Reformulate Tests: P-values Don't Give an Effect Size

Severity function: SEV(Test T, data ***x***, claim *C*)

- Tests are reformulated in terms of a discrepancy γ from H_0

- Instead of a binary cut-off (significant or not) the particular outcome is used to infer discrepancies that are or are not warranted

An Example of SEV (3.2 SIST)

1-sided normal testing

H_0: μ ≤ 150 vs. H_1: μ > 150 (Let σ = 10, n = 100)

Reject H_0 whenever M ≥ 2SE: M ≥ 152

M is the sample mean (significance level = .025)

1SE = $\sigma/\sqrt{n}$ = *1*

Let M = 152, so I reject H_0.

H_0: μ ≤ 150 vs. H_1: μ > 150 (Let σ = 10, *n* = 100)

The usual test infers there's an indication of *some* positive discrepancy from 150 because

$$Pr(M < 152: H_0) = .97$$

SEV(M = 152, μ > 150) = 0.97

Not very informative

Are we warranted in inferring μ > 153 say?

- Recall the complaint of the Likelihoodist (p. 36)

- For them, inferring H_1: $\mu > 150$ means every value in the alternative is more likely than 150

- Our inferences are not to point values, but we block inferences to discrepancies beyond those warranted with severity.

consider **SEV(μ > 153)**

M = 152, as before, *C*: μ > 153

Pr(*"a worse fit"*; *C* is false)

Pr(M ≤ 152; μ ≤ 153)

Evaluate at μ = 153, as the prob is greater for μ < 153.

To get Pr(M ≤ 152: μ = 153), standardize:
Z = √100 (152- 153)/1 = -1

Pr(Z < -1) = .16 Terrible evidence

Now consider SEV($\mu > 150.5$)   (still with M = *152)*

Pr (A worse fit with *C*; claim is false) = .97

Pr($M < 152$; $\mu = 150.5$)

$Z = (152 - 150.5) / 1 = 1.5$

Pr ($Z < 1.5$)= .93   Fairly good indication $\mu > 150.5$

Table 3.1 Reject in test T+: H_0: $\mu \leq 150$ vs. H_1: $\mu > 150$ with $\bar{x} = 152$

Claim $\mu > \mu_1$	Severity $\Pr(\bar{X} \leq 152; \mu = \mu_1)$
$\mu > 149$	0.999
$\mu > 150$	0.97
$\mu > 151$	0.84
$\mu > 152$	0.5
$\mu > 153$	0.16

$\mu > 150.5$ → (between $\mu > 150$ and $\mu > 151$)

.093 → 0.97

FOR PRACTICE:

Now consider SEV($\mu > 151$)   (still with $M = 152$)

Pr (A worse fit with *C*; claim is false) = __

Pr($M < 152$; $\mu = 151$)

$Z = (152 – 151) /1 = 1$

Pr ($Z < 1$)= .84

MORE PRACTICE:

Now consider SEV($\mu > 152$) (still with M = 152)

Pr (A worse fit with *C*; claim is false) = __

Pr($M < 152$; $\mu = 152$)

$Z = 0$

Pr ($Z < 0$)= .5–important benchmark

Terrible evidence that $\mu > 152$

Table 3.2 has exs with M = 153.

(looks ahead) Compare *n* = 100 with *n* = 10,000

H_0: μ ≤ 150 vs. H_1: μ > 150 (Let σ = 10, *n* = 10,000)

Reject H_0 whenever M ≥ 2SE: M ≥ 150.2

M is the sample mean (significance level = .025)

1SE = $\sigma/\sqrt{n}$ = $10/\sqrt{10{,}000}$ = .1

Let M = 150.2, so I reject H_0.

Comparing n = 100 with n = 10,000

Reject H_0 whenever M ≥ 2SE: M ≥ 150.2

$SEV_{10,000}$(μ > 150.5) = 0.001

Z = (150.2 – 150.5) /.1 = -.3/.1 = -3

P(Z < -3) = .001

Corresponding 95% CI: [0, 150.4]

A .025 result is terrible indication μ > 150.5

When reached with n = 10,000

While SEV_{100}(μ > 150.5) = 0.93

Non-rejection. Let M = 151, the test does not reject H_0.

The standard formulation of N-P (as well as Fisherian) tests stops there.

We want to be alert to a fallacious interpretation of a "negative" result: inferring there's no positive discrepancy from μ = 150.

The data "accord with" H_0, but what if the test had little capacity to have alerted us to discrepancies from 150?

Condition (S-2) requires us to consider Pr(X > 151; 150), which is only .16.

Computation for M = 151

$Z = (151 - 150)/1 = 1$

$Pr(Z > 1) = .16$

SEV(T, M = 151, *C*: $\mu \leq 150$) = low (.16).

- So there's poor indication of H_0

Can they say M = 151 is a good indication that $\mu \leq 150.5$?

No, SEV(T, M = 151, C: $\mu \leq 150.5$) = ~.3.

[Z = 151 – 150.5 = .5]

But M = 151 is a good indication that $\mu \leq 152$

[Z = 151 – 152 = -1; Pr (Z > -1) = .84]

SEV($\mu \leq 152$) = .84

It's an even better indication $\mu \leq 153$ (Table 3.3, p. 145)

[Z = 151 – 153 = -2; Pr (Z > -2) = .97]

Frequentist Evidential Principle: FEV

FEV (i). ***x*** is evidence against H_0 (i.e., evidence of a discrepancy from H_0), if and only if, were H_0 a correct description of the mechanism generating ***x***, then, with high probability, this would have resulted in a less discordant result than is exemplified by ***x*** (Mayo and Cox 2006, p. 82; substituting ***x*** for ***y***).

FEV (i). ***x*** is evidence against H_0 (i.e., evidence of discrepancy from H_0), if and only if the P-value $\Pr(d > d_0; H_0)$ is very low (equivalently, $\Pr(d < d_0; H_0) = 1 - P$ is very high).

Contraposing FEV(i) we get our minimal priniciple

FEV (ia) ***x*** are poor evidence against H_0 (poor evidence of discrepancy from H_0), if there's a high probability the test would yield a more discordant result, if H_0 is correct.

Note the one-directional 'if' claim in FEV (1a)

(i) is not the only way **x** can be BENT.

P-value "moderate"

FEV(ii): A moderate *p* value is evidence of the absence of a discrepancy γ from H_0, only if there is a high probability the test would have given a worse fit with H_0 (i.e., smaller *P*- value) were a discrepancy γ to exist.

For a Fisherian like Cox, a test's power only has relevance pre-data, they 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 2006, p. 25)

$\Pi(\gamma)$: "sensitivity function"

Computing $\Pi(\gamma)$ views the P-value as a statistic.
$\Pi(\gamma) = \Pr(P < p_{obs}; \mu_0 + \gamma)$.

The alternative $\mu_1 = \mu_0 + \gamma$.

Given that P-value inverts the distance, it is less confusing to write $\Pi(\gamma)$

$\Pi(\gamma) = \Pr(d > d_0; \mu_0 + \gamma)$.

Compare to the power of a test:

$POW(\gamma) = \Pr(d > c_\alpha; \mu_0 + \gamma)$ the N-P cut-off c_α.

FEV(ii) in terms of $\Pi(\gamma)$

P-value is modest *(not small):* Since the data accord with the null hypothesis, FEV directs us to examine the probability of observing a *result more discordant from H_0* if $\mu = \mu_0 + \gamma$:

If $\Pi(\gamma) = \Pr(d > d_0; \mu_0 + \gamma)$ is very high, the data indicate that $\mu < \mu_0 + \gamma$.

Here $\Pi(\gamma)$ gives the severity with which the test has probed the discrepancy γ.

FEV (ia) in terms of $\Pi(\gamma)$

If $\Pi(\gamma)$ = Pr(d > do; μ_0 +γ) = moderately high (greater than .3, .4, .5), then there's poor grounds for inferring $\mu > \mu_0 + \gamma$.

This is equivalent to saying the SEV($\mu > \mu_0 + \gamma$) is poor.

FEV/SEV (for Excur 3 Tour III)

Test T+: Normal testing: H_0: $\mu \leq \mu_0$ vs. H_1: $\mu > \mu_0$

σ known

(FEV/SEV): If d(x) is statistically significant (P- value very small), then test T+ passes $\mu > M_0 - k_\varepsilon \sigma/\sqrt{n}$ with severity $(1 - \varepsilon)$.

(FEV/SEV): If d(x) is *not* statistically significant (P- value moderate), then test T+ passes $\mu < M_0 + k_\varepsilon \sigma/\sqrt{n}$ with severity $(1 - \varepsilon)$,

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

PRACTICE WITH P-VALUES

Let M = 151

$Z = (151 - 150)/1 = 1$

The P-value is $Pr(Z > 1) = .16$

SEV ($\mu > 150$) = .84 = 1 – P-value

PRACTICE WITH P-VALUES

Let $M = 150.5$

$Z = (150.5 – 150)/1 = .5$

The P-value is $Pr(Z > .5) = .3$

SEV ($\mu > 150$) = .7 = 1 – P-value

PRACTICE WITH P-VALUES

Let M = 150

$Z = (150 - 150)/1 = 0$

The P-value is $\Pr(Z > 0) = .5$

SEV $(\mu > 150) = .5 = 1 - $ P-value