

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

Severity function: $\text{SEV}(\text{Test } T, \text{data } \mathbf{x}, \text{claim } C)$

- 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

$H_0: \mu \leq 150$ vs. $H_1: \mu > 150$ (Let $\sigma = 10$, $n = 100$)

let significance level $\alpha = .025$

Reject H_0 whenever $M \geq 2\sigma/\sqrt{n}$: $M \geq 152$

M is the sample mean, its value is M_0 .

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

Rejection rules:

Reject iff $M > 150 + 2SE(N-P)$

In terms of the P-value:

Reject iff $P\text{-value} \leq .025$ (Fisher)

(P-value a distance measure, but inverted)

Let $M = 152$, so I reject H_0 .

PRACTICE WITH P-VALUES

Let $M = 152$

$$Z = (152 - 150)/1 = 2$$

The P-value is $\Pr(Z > 2) = .025$

PRACTICE WITH P-VALUES

Let $M = 151$

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

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

$$\text{SEV} (\mu > 150) = .84 = 1 - \text{P-value}$$

PRACTICE WITH P-VALUES

Let $M = 150.5$

$$Z = (150.5 - 150)/1 = .5$$

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

PRACTICE WITH P-VALUES

Let $M = 150$

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

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

Frequentist Evidential Principle: FEV

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 principle

FEV (ia) \mathbf{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 \mathbf{x} can be BENT.

$H_0: \mu \leq 150$ vs. $H_1: \mu > 150$ (Let $\sigma = 10$, $n = 100$)

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

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

Not very informative

Are we warranted in inferring $\mu > 153$ 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: How severely has $\mu > 153$ passed the test?

SEV($\mu > 153$) (p. 143)

$M = 152$, as before, claim $C: \mu > 153$

The data “accord with C ” but there needs to be a reasonable probability of a worse fit with C , if C is false

$\Pr(\text{“a worse fit”}; C \text{ is false})$

$\Pr(M \leq 152; \mu \leq 153)$

Evaluate at $\mu = 153$, as the prob is greater for $\mu < 153$.

Consider: How severely has $\mu > 153$ passed the test?

To get $\Pr(M \leq 152: \mu = 153)$, standardize:

$$Z = \sqrt{100} (152 - 153)/1 = -1$$

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

Consider: How severely has $\mu > 150$ passed the test?

To get $\Pr(M \leq 152: \mu = 150)$, standardize:

$$Z = \sqrt{100} (152 - 150) / 1 = 2$$

$$\Pr(Z < 2) = .97$$

Notice it's $1 - \text{P-value}$

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

$\Pr(\text{A worse fit with } C; \text{ claim is false}) = .97$

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

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

$\Pr(Z < 1.5) = .93$ 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	Severity
$\mu > \mu_1$	$\Pr(\bar{X} \leq 152; \mu = \mu_1)$
$\mu > 149$	0.999
$\mu > 150$	0.97
$\mu > 151$	0.84
$\mu > 152$	0.5
$\mu > 153$	0.16

$\mu > 150.5$



.093

FOR PRACTICE:

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

$\Pr(\text{A worse fit with C; claim is false}) = \underline{\hspace{1cm}}$

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

$Z = (152 - 151) / 1 = 1$

$\Pr(Z < 1) = .84$

MORE PRACTICE:

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

$\Pr(\text{A worse fit with } C; \text{ claim is false}) = \underline{\hspace{1cm}}$

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

$Z = 0$

$\Pr(Z < 0) = .5$ —important benchmark

Terrible evidence that $\mu > 152$

Table 3.2 has exs with $M = 153$.

Using Severity to Avoid Fallacies:

Fallacy of Rejection: Large n problem

- Fixing the P-value, increasing sample size n , the cut-off gets smaller
- Get to a point where \mathbf{x} is closer to the null than various alternatives
- Many would lower the P-value requirement as n increases-can always avoid inferring a discrepancy beyond what's warranted:

Severity tells us:

- an α -significant difference indicates *less* of a discrepancy from the null if it results from larger (n_1) rather than a smaller (n_2) sample size ($n_1 > n_2$)
- What's more indicative of a large effect (fire), a fire alarm that goes off with burnt toast or one that doesn't go off unless the house is fully ablaze?



- [The larger sample size is like the one that goes off with burnt toast]

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

$H_0: \mu \leq 150$ vs. $H_1: \mu > 150$ (Let $\sigma = 10$, $n = 10,000$)

Reject H_0 whenever $M \geq 2SE$: $M \geq 150.2$

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 $M \geq 2SE$: $M \geq 150.2$

$$\mathbf{SEV_{10,000}(\mu > 150.5) = 0.001}$$

$$Z = (150.2 - 150.5) / .1 = -.3 / .1 = -3$$

$$P(Z < -3) = .001$$

Corresponding 95% CI: $[0, 150.4]$

A .025 result is terrible indication $\mu > 150.5$

When reached with $n = 10,000$

$$\mathbf{While SEV_{100}(\mu > 150.5) = 0.93}$$

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 $\mu = 150$.

Is there evidence of compliance? $\mu \leq 150$?

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

No evidence against H_0 is not evidence for it.

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

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)

Computation for SEV(T, M = 151, C: $\mu \leq 150$)

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

$$\Pr(Z > 1) = .16$$

SEV(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, $\text{SEV}(T, M = 151, C: \mu \leq 150.5) = \sim .3$.

$[Z = 151 - 150.5 = .5]$

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

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

$\text{SEV}(\mu \leq 152) = .84$

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

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

$\Pi(\gamma)$: “sensitivity function”

Computing $\Pi(\gamma)$ views the P-value as a statistic.

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

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:

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

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 $\Pi(\gamma) = \Pr(d > d_0; \mu_0 + \gamma)$ = moderately high (greater than .3, .4, .5), then there's poor grounds for inferring $\mu > \mu_0 + \gamma$.

This is equivalent to saying the $\text{SEV}(\mu > \mu_0 + \gamma)$ 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 $P(d(X) > k_\varepsilon) = \varepsilon$.