Day 7 (A) Large n, CIs and Tests Excursion 3 Tour III, Excursion 4 Tour II

How Could a Group of Psychologists be so Wrong?
Morrison and Henkel's 1970 classic, *The Significance Test Controversy.*

- Notably, Rosenthal and Gaito (1963) discovered that statistical significance at a given level was often fallaciously taken as evidence of a greater discrepancy from the null the larger the sample size *n*.
- In fact, it is indicative of *less* of a discrepancy from the null than if it resulted from a smaller sample size.

> What is shocking is that these psychologists indicated substantially greater confidence or belief in results associated with the larger sample size for the same p values. According to the theory, especially as this has been amplified by Neyman and Pearson (1933), the probability of rejecting the null hypothesis for any given deviation from null and p values increases as a function of the number of observations. The rejection of the null hypothesis when the number of cases is small speaks for a more dramatic effect in the population…The question is, how could a group of psychologists be so wrong? (Bakan 1970, p. 241)

(Remember our convention: discrepancy refers to the parametric, not the observed, difference. Their use of “deviation” from the null alludes to our “discrepancy”.)

John Pratt : "the more powerful the test, the more a just significant result favors the null hypothesis" (1961, p. 166).

By contrast: "The thesis implicit in the [NP] approach, [is] that a hypothesis may be rejected with increasing confidence or reasonableness as the power of the test increases" (Howson and Urbach 1993, p. 209).

The fallacy is akin to making mountains out of molehills:

> *Mountains out of Molehills Fallacy* (large *n* problem): The fallacy of taking a (P-level) rejection of H_0 with larger sample size (*higher power*) as indicative of a greater discrepancy from the null than with a smaller sample size.

Recall the analogy with two fire alarms

It is true that a large enough sample size triggers the alarm with an observed mean that is quite “close” to the null hypothesis.

But, if the test rings the alarm (i.e., rejects H_0) even for tiny discrepancies from the null value, then the alarm is *poor* grounds for inferring larger discrepancies.

A test must have a large enough sample to satisfy model assumptions.

True, but our interpretive question can’t get started without taking the P-values as legitimate and not spurious.

4.3 *Significant Results with Overly Sensitive Tests: Large n Problem*.

The fact that the test would eventually uncover any discrepancy there may be, regardless of how small, doesn't mean there always is such a discrepancy.

Let's focus on the example of Normal testing,
T+ with H_0: μ ≤ 0 vs. H_1: μ > 0 letting σ= 1.
To bring out the effect of sample size many prefer to write the statistic as

$$d(\boldsymbol{x}_0) = \sqrt{n}\ (\bar{x} - 0)/\sigma.$$

rather than

$$d(\boldsymbol{x}_0) = (\bar{x} - 0)/\ \sigma_{\bar{X}},$$

where $\sigma_{\bar{X}}$ abbreviates ($\sigma/\sqrt{n}$).

T+ rejects H_0 (at the .025 level) iff $\bar{X} > 0 + 1.96(\sigma/\sqrt{n})$.
As n increases, a single $(\sigma/\sqrt{n})$ unit decreases.

The hypotheses are not point values, but discrepancies: each corresponds to an assertion: there's evidence of a discrepancy at least this large, but poor evidence it's as large as thus and so.

See Figure 4.3 p. 242 comparing 3 sample sizes

Consider the 2-SE cut-off for n = 25, 100, 400 in test T+, $\sigma = 1$
Let $\bar{x}_{.025}$ be the sample mean just statistically significant at the .025 level in each test.

With n = 25, $\bar{x}_{.025}$ = 2(1/5)
with n = 100, $\bar{x}_{.025}$ = 2(1/10),
with n = 400, $\bar{x}_{.025}$ = 2(1/20).

So the cut-offs for rejection are .4, .2, and .1 respectively.

Exhibit (v). *Responding to a Familiar Chestnut.*

Did you hear the one about the significance tester who rejected H_0 in favor of H_1 even though the result makes H_0 more likely than H_1 ?

Elliott Sober (2008, p. 56) (who is echoing Howson and Urbach (1993: 208-9), who are echoing Lindley (1957)).

I will allude to our test T+: H_0: μ = 0 vs. H_1: μ > 0 with $\sigma = 1$. "Another odd property of significance tests," says Sober, "concerns the way in which they are sensitive to sample size." Suppose you are applying test T+ with null H_0: μ = 0. If your sample size is *n* =25, and you choose α = .025, you will reject H_0 whenever $\bar{x} \geq 0.4$. If you examine *n* = 100, and choose the same value for α, you will reject H_0 whenever $\bar{x} \geq 0.2$. And if you examine $n = 400$, again with $α = .025$, you will reject H_0 whenever $\bar{x} \geq .1$. “As sample size increases” the sample mean $\bar{x}$ must be closer and closer to 0 for you to *not* reject H_0.

“This may not seem strange until you add the following detail. Suppose the alternative to H_0 is the hypothesis” H_1: $μ = .24$. “The law of likelihood now entails that observing” $\bar{x} < .12$ favors H_0 over H_1, so in particular $\bar{x} = .1$ favors H_0 over H_1.

Your reply: Hold it right at “add the following detail”.
You’re observing that the significance test disagrees with a law of likelihood appraisal to a point vs. point test:

H_0: $μ = 0$ vs. H_1: $μ = 0.24$.

- We require the null and alternative to exhaust the space of parameters, and these don't.

- Nor are our inferences to points, but rather to inequalities about discrepancies; let's consider your ex: $H_{0:}$ $\mu \leq 0$ vs. H_1: $\mu > 0$

- $\bar{x} = 0.1$, while indicating *some* positive discrepancy from 0, offers bad evidence for inferring μ as great as 0.24.

- Since $\bar{x} = .1$ rejects H_0, the result *accords* with H_1.

- The severity associated with inference $\mu \geq .24$ asks: what's the probability of observing $\bar{x} < .1$–i.e., a result *more discordant* with H_1 assuming $\mu = .24$.

SEV($\mu \geq .24$) with $\bar{x} = 0.1$, and $n = 400$ is

Pr($\bar{X} < .1$; $\mu = .24$).

Standardizing $\bar{X}$ yields Z = √400 (.1 - .24)/1 = 20(-.14) = -2.8.

So SEV($\mu \geq .24$) = .003!

Were μ as large as .24, we'd have observed a larger observed mean than we did with .997 probability! It's terrible evidence for H_1: μ = 0.24.

This is redolent of the Binomial example in discussing Royall (1.4).

To underscore the difference between the likelihoodist's comparative appraisal and the significance tester, you might consider an alternative that the likelihoodist takes as favored over H_0: μ = 0 with $\bar{x}$ = .1, namely, the maximum likely alternative H_1: μ = 0.1.

- This is one of our key benchmarks for a discrepancy that's poorly indicated!

- To the likelihoodist, inferring that H_1: μ = .1 "is favored" over H_0: μ = 0 makes sense, whereas to infer a discrepancy of .1 from H_0 is highly *un*warranted for a significance tester.

Our aims are very different.

Exhibit (vi). *Reforming the Reformers on Confidence Intervals*. SIST P. 244

Our one-sided test T+ (H_0: $\mu \leq 0$ vs. H_1: $\mu > 0$, and σ = 1) at α = .025 has as its dual the one-sided (lower) 97.5% general confidence interval: $\mu > \bar{X} - 2(1/\sqrt{n})$–rounding to 2 from 1.96.

With α = .05, we have $\mu > \bar{x} - 1.65(1/\sqrt{n})$.

Consider the three instances of test T+ with (i) n = 25 (ii) *n* = 100 and (iii) *n* = 400: (i) $\bar{x} = .4$, (ii) $\bar{x} = .2$ and (iii) $\bar{x} = .1$.

Form .975 confidence interval estimates for each:
(i) for *n* = 25, infer: $\mu > \hat{\mu}_{.975}$ that is, $\mu > \bar{x} - 2(1/5)$
(ii) for *n* = 100, infer: $\mu > \hat{\mu}_{.975}$ that is, $\mu > \bar{x} - 2(1/10)$
(iii) for *n* = 400, infer: $\mu > \hat{\mu}_{.975}$ that is, $\mu > \bar{x} - 2(1/20)$.

Substituting $\bar{x}$ in all cases, we get the same one-sided confidence interval:

$\mu > 0$.

Cumming writes them as (0, infinity].

How are the CIs distinguishing them?

- I don’t want to step too hard on the CI champion’s toes, since CIs are in sync with the severity critique I am mounting.

- Severity directs you to avoid taking your result as indicating a discrepancy beyond what’s warranted.

The same inference can be well indicated with *n* = 100, while poorly indicated when *n* = 400. (SIST p. 245)

The upper .975 bound would reflect the greater sensitivity with increasing sample sizes:

(i) n = 25: (0, .8] (ii) n = 100: (0, .4] (iii) n = 400: (0, .2]

But we cannot just deny one-sided tests, nor does Cumming. We need a justification for looking at the upper bound
p. 246:

There's an equivocation, or at least a potential equivocation, in Cumming's assertion "that for [2.5%] of replications the [lower limit] will exceed the true value" (Cumming 2012, p. 112 replacing 5% with 2.5%).

This is not a true claim if "lower limit" is replaced by a *particular* lower limit: $\hat{\mu}_{.025}(\bar{x})$, it holds only for the *generic* lower limit $\hat{\mu}_{.025}(\bar{X})$.

https://marcosjnez.shinyapps.io/Severity/