Trade-offs and Benchmarks

Between H_0 and $\overline{x}_\alpha$ the power goes from α to 0.5. Keeping to our simple test T+ will amply reward us here.

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:

$$\text{POW}(\text{T+}, \mu_0) = \Pr(\overline{X} \geq \overline{x}_\alpha; \mu_0), \text{ where } \overline{x}_\alpha = \mu_0 + z_\alpha \sigma_{\overline{X}}$$

Standardizing $\overline{X}$, we get $Z = [(\mu_0 + z_\alpha \sigma_{\overline{X}}) - \mu_0]/\sigma_{\overline{X}}$.

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

It's the *low power* against H_0 that warrants taking a rejection as evidence that $\mu > \mu_0$. This is desirable: we infer an indication of discrepancy from H_0 because a null world would probably have resulted in a smaller difference than we observed.

b. *The power of T+ for $\mu_1 = \overline{x}_\alpha$ is 0.5.* In that case, $Z = 0$, and $\Pr(Z \geq 0) = 0.5$, so

$$\text{POW}(\text{T+}, \mu_1 = \overline{x}_\alpha) = 0.5.$$

The power only gets to be greater than 0.5 for alternatives that exceed the cut-off $\overline{x}_\alpha$, whatever it is. As noted, $\mu^{.8} = \overline{x}_\alpha + 0.85\,\sigma_{\overline{X}}$ since $\text{POW}(\text{T+}, \overline{x}_\alpha + 0.85\sigma_{\overline{X}}) = 0.8$. Tests ensuring 0.9 power are also often of interest: $\mu^{.9} = \overline{x}_\alpha + 1.28\sigma_{\overline{X}}$. We get these shortcuts:

Case 1: POW(T+, μ) for μ between H_0 and $\mu = \overline{x}_\alpha$:
If $\mu_1 = \overline{x}_\alpha - k\sigma_{\overline{X}}$ then POW(T+, μ_1) = area to the right of k under N(0,1) ($\leq$ 0.5).

Case 2: POW(T+, μ) for μ greater than $\overline{x}_\alpha$:
If $\mu_1 = \overline{x}_\alpha + k\sigma_{\overline{X}}$ then POW(T+, μ_1) = area to the right of $-k$ under N (0,1) ($>$ 0.5).
Remember $\overline{x}_\alpha$ is $\mu_0 + z_\alpha\sigma_{\overline{X}}$.

Trade-offs Between the Type I and Type II Error Probability

We know that, 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). And as the probability of a Type II error goes down (and power goes up), the probability of a Type I error goes up, assuming we leave everything else the same. There's a trade-off between the two error probabilities. (No free lunch.) So 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! You'd know automatically they had made a mistake or were simply defining things in a way that differs from standard N-P statistical tests. Now you may say, "I don't care about Type I and II errors, I'm interested in inferring estimated effect sizes." I too want to infer magnitudes. But those will be ready to hand once we tell what's true about the existing concepts.

While $\mu^{.8}$ is obtained by adding 0.85 $\sigma_{\overline{X}}$ to $\overline{x}_{\alpha}$, in day-to-day rounding, if you're like me, you're more likely to remember the result of adding $1\sigma_{\overline{X}}$ to $\overline{x}_{\alpha}$. That takes us to a value of μ against which the test has 0.84 power, $\mu^{.84}$:

> The power of test T+ to detect an alternative that exceeds the cut-off $\overline{x}_{\alpha}$ by $1\sigma_{\overline{X}} = 0.84$.

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

Exhibit (i). Let test T+ ($\alpha = 0.025$) be H_0: $\mu = 0$ vs. H_1: $\mu \geq 0$, $\alpha = 0.025$, $n = 25$, $\sigma = 1$. Using the $2\sigma_{\overline{X}}$ cut-off: $\overline{x}_{0.025} = 2(1)/\sqrt{25} = 0.4$ (using 1.96 it's 3.92). Suppose you are instructed to decrease the Type I error probability α to 0.001 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+ will be $\overline{x}_{0.001}$. It must be $3\,\sigma_{\overline{X}}$ greater than 0 rather than only $2\sigma_{\overline{X}}$: $\overline{x}_{0.001} = 0 + 3(1)/\sqrt{25} = 0.6$. We decrease α (the Type I error probability) from 0.025 to 0.001 by moving the hurdle over to the right by $1\sigma_{\overline{X}}$ unit. But we've just made the power lower for any discrepancy or alternative. For what value of μ does this new test have 0.84 power?

$$\text{POW(T+, } \alpha = 0.001, \mu^{.84} = ?) = 0.84.$$

We know: $\mu^{.84} = 0.6 + (0.2) = 0.8$. So, POW(T+, $\alpha = 0.001$, $\mu = 0.8$) = 0.84. Decreasing the Type I error by moving the hurdle over to the right by $1\sigma_{\overline{X}}$ unit results in the alternative against which we have 0.84 power $\mu^{.84}$ also moving over to the right by $1\sigma_{\overline{X}}$ (Figure 5.1). We see the trade-off very neatly, at least in one direction.

Consider the discrepancy of $\mu = 0.6$ (Figure 5.2). The power to detect 0.6 in test T+ ($\alpha = 0.001$) is now only 0.5! In test T+ ($\alpha = 0.025$) it is 0.84. Test T+ ($\alpha = 0.001$) is less powerful than T+ ($\alpha = 0.025$).

Should you hear someone say that the higher the power, the higher the *hurdle* for rejection, you'd know they are confused or using terms in an

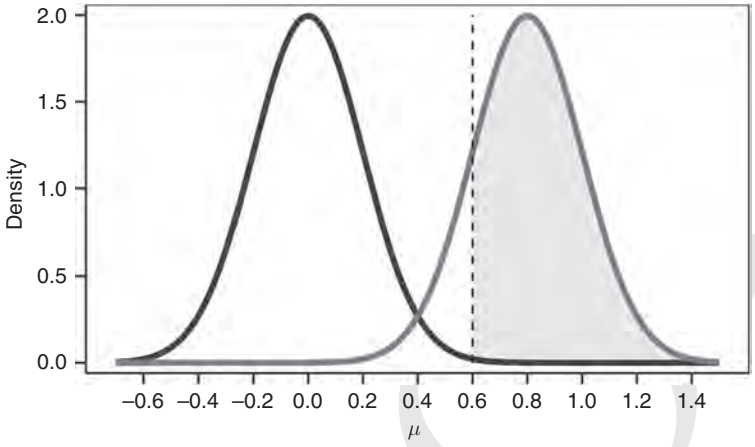

Figure 5.1 POW(T+, $\alpha = 0.001$, $\mu = 0.8$) = 0.84.

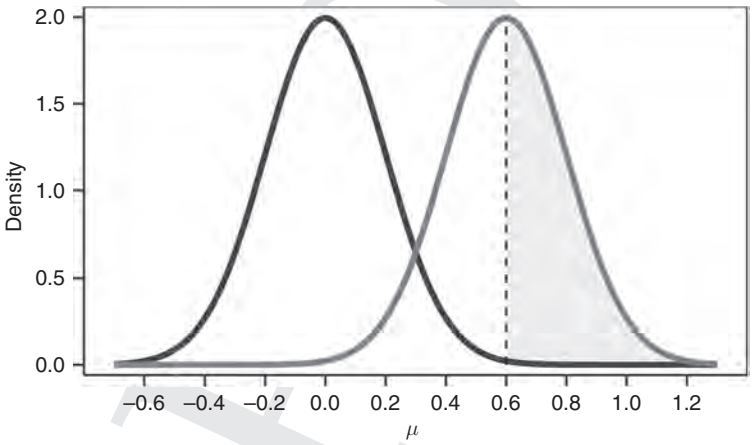

Figure 5.2 POW(T+, $\alpha = 0.001$, $\mu = 0.6$) = 0.5.

incorrect way. (The hurdle is how large the cut-off must be before rejecting at the given level.) Why then do Ziliak and McCloskey, popular critics of significance tests, announce: "refutations of the null are trivially easy to achieve if power is low enough or the sample is large enough" (2008a, p. 152)? Increasing sample size means increased power, so the second disjunct is correct. The first disjunct is not. One might be tempted to suppose they mean "power is high

enough," but one would be mistaken. They mean what they wrote. Aris Spanos (2008a) points this out (in a review of their book), and I can't figure out why they dismiss such corrections as "a lot of technical smoke" (2008b, p. 166).

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. What I think they want to say, or at any rate what would be correct, is

$$\Pr(\text{Test T+ } \textit{does not} \text{ reject the null hypothesis} \mid \mu^{.84}) = 0.16.$$

So the Type II error probability is 0.16. Looking at it this way, the flaw is in computing the complement of premise 1 by transposing (as we saw in the Higgs example, Section 3.8). Let's be clear about significance levels and hurdles. According to Ziliak and McCloskey:

> It is the history of Fisher significance testing. One erects little "significance" hurdles, six inches tall, and makes a great show of leaping over them, . . . If a test does a good job of uncovering efficacy, then the test has high power and the hurdles are high not low. (ibid., p. 133)

They construe "little significance" as little hurdles! It explains how they wound up supposing high power translates into high hurdles. It's the opposite.