1. Consider Figure 2 of Mayo and Spanos p. 343. Let’s read off some values of the power of
the test (the solid curve) specified at the start of Section 4, p. 336. (The test rejects Ho iff \( d(x_0) \) is at
least 1.96 standard deviation units, i.e., iff the observed sample mean is greater than or equal to
12.39---can round to 12.4).
A. Use both calculations and Figure 2 to determine:
The power of the test to detect 12.1, 12.2, to detect 12.4, to detect 12.6, to detect 12.8, to detect 13.
(ANSWERS: .07, .16, .5, .84, .97, ~1)

B. Imagine that “compliance” with a given regulatory code requires showing evidence that the
population mean is greater than 12.8, and suppose a company just rejects the null, Ho = 12, at the
cut-off, say the data yield 12.4. Does this statistically significant result provide evidence of
compliance?

2. Consider a highly informal and qualitative example, just to eek out some reasoning,
concerning a hypothesis Ho that certain toxic substances are harmful even in low doses (perhaps
that there’s a linear dose-response). Suppose a TEST T rejects Ho and infers
J: there are no harms, and there might even be benefits (which may be quantified), from exposure to
low doses of toxic substances,
so long as a single study (perhaps found from a literature review) has observed a
((nominally)statistically significantly) lower incidence of a given disease among a given group of
people exposed to a given toxin, as opposed to a comparative non-exposed group. Hypothesis J is
sometimes called a hormetic hypothesis. Now, TEST T has a high probability of rejecting Ho and
inferring J, under the assumption that the hormetic hypothesis J is true. That is, were the Hormetic
hypothesis generally true, the test has an extremely good chance of locating at least one such study.
So the POWER of TEST T to reject Ho and infer J is high---even though we’re considering all of
this qualitatively.
i.e., POW(TEST T, against J) is high.

EVALUATE THE FOLLOWING ARGUMENT:
(1.) There is a (very) high probability that TEST T would reject Ho (toxic substances are harmful)
and infer J (hormesis hypothesis) were in fact J true.
(2.) I applied TEST T and rejected Ho with data x
(3. ) Therefore , x is good evidence for J.

3. What are the two main ways that “fallacies of rejection” can occur (Mayo and Spanos,
section 5, pp. 341-2)? Why does the N-P requirement that the null and alternative hypotheses
exhaust the parameter space (strictly speaking) preclude the former version of the fallacy (5.1).
Can the severity principle prevent the second variation of the fallacy of rejection (5.2)? How? (5.3)
Recall the connection between this fallacy and the “large n problem” we’ve discussed.
4. Using your reasoning from example (2) evaluate the following claim by Howson and Urbach (hereafter H & U):

“The thesis implicit in the [N-P] approach, that a hypothesis may be rejected with increasing confidence or reasonableness as the power of the test increases, is not borne out in the example, which signals the reverse trend.” (H&U, p. 208)

5. Howson and Urbach criticize N-P tests using an example (that of the tulip bulbs) where the null and alternative hypotheses are “point” hypotheses (p = .4 vs. p = .6), violating the exhaustive requirement of tests. However, we can consider H & U’s concern (which is related to what they say regarding a proper one-sided version of this test, with composite alternative, p = .4 vs. p > .4). Now the tulip example (p. 208) concerns discrete binomial distributions, but we can apply our familiar calculations using the Normal approximation to the Binomial for, say, n = 100. The standard deviation (for this statistic) is ~ .5 divided by the square root of n, so it would be .05 when n = 100. We are given that the significance level of the test is .05, so the test rule is:

- A. So what’s the cut-off for rejection?
- B. Calculate the Power (to reject the null) when the true value is .6, and see if you get their answer (.99)

Suppose the actual outcome is .48 red flowers, so the null is just rejected. What is the SEV associated with inferring p > .6?

ANSWER: P(would observe less than .48; p = .6)
You can consider some benchmarks, SEV(p > .4), SEV (p > .48), SEV (p > .5)
Using these results, evaluate the following claim:
The outcome .48 is evidence that p > .4 but is very poor evidence that p is as great as .48, or .5...and that much worse that there’s evidence that p is as great as .6.

6. In H & U section d, p. 215 we get a legitimate N-P test (our one-sided friend) but they declare “there is no determinable probability of a type II error in tests of composite hypotheses.” Evaluate!

7. Ponder: H & U criticize N-P tests as being subjective (indeed more subjective than those based on subjective degrees of belief!) because there is discretion in test specification, e.g., determining the hypothesis to test, the sample size, significance level and power. I would argue that objectivity is attained in error statistical tests because, however you specified the test, I can objectively scrutinize which inferences are, and which are not, warranted by the data.