very significant result (it exceeds .5 by 4 standard deviations) in a sample of 100 trials. In a sample of 10,000 trials, an equally statistically significant result requires only 52 percent successes! An alleged paradox is that a significance test with large enough sample size rejects the null with outcomes that seem very close to, and by a Bayesian analysis are supportive of, the null hypothesis. This might be called the Jeffreys-Good-Lindley paradox, after those Bayesians who first raised it.

I discuss this paradox at length in Mayo 1985a and elsewhere, but here I just want to show how easy it is to get around a common criticism that is based on it. The criticism of NP tests results only by confusing the import of positive results. The fallacious interpretation results from taking a positive result as indicating a discrepancy beyond that licensed by RR. Howson and Urbach give a version of this criticism (along the lines of an argument in Lindley 1972). Their Binomial example is close enough to the one above to use it to make out their criticism (their p is equal to the proportion of flowering bulbs in a population). The criticism is that in a test with sample size 10,000, the null hypothesis $H: p = .5$ is rejected in favor of an alternative J, that p equals .6 even though .52 is much closer to .5 (the hypothesis being rejected) than it is to .6. And yet, the criticism continues, the large-scale test is presumably a better NP test than the smaller test, since it has a higher power (nearly 1) against the alternative that $p = .6$ than the smaller test (.5).[23]

The authors take this as a criticism of NP tests because "The thesis implicit in the [NP] 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" (Howson and Urbach 1989, 168). Not only is this thesis not implicit in the NP approach, but it is the exact reverse of the appropriate way of evaluating a positive (i.e., statistically significant) result. The thesis that gives rise to the criticism comes down to thinking that if a test indicates the existence of some discrepancy then it is an even better indication of a very large discrepancy!

Looking at RRii makes this clear. Let us compare the import of the two 4-standard-deviation results, one from a test with sample size $n = 100$, the second from a test with sample size $n = 10{,}000$. In the experiment with 10,000 trials, the observation of 52 percent successes is an extremely *poor* indicator that p is as large as .6. For such a result is very probable even if the true value of p is actually less than .6,

23. I am calculating power here with the cutoff for rejection set at .6—the 2-standard-deviation cutoff.

say, if $p = .55$. Indeed, it is practically certain that such a large result would occur for p as small as .55. Were one to take such a result as warranting that p is .6, one would be wrong with probability very near one.

In contrast, the observation of 70 percent successes with $n = 100$ trials is a very good indication that p is as large as .6. The probability of getting so large a proportion of successes is very small (about .03) if μ is less than .6. The severity of a test that passes "p is as large as .6" with 70 percent successes out of 100 trials is high (.97).

Howson and Urbach's criticism, and a great many others with this same pattern, are based on an error to which researchers have very often fallen prey. The error lies in taking an α-significant difference (from H) with a large sample size as more impressive (better evidence of a discrepancy from H) than one with a smaller sample size.[24] That, in fact, it is the reverse is clearly seen with rule RR. The reasoning can be made out informally with an example such as our ultrasound probe. Take an even more homey example. Consider two smoke detectors. The first is not very sensitive, rarely going off unless the house is fully ablaze. The second is very sensitive: merely burning toast nearly always triggers it. That the first (less sensitive) alarm goes off is a *stronger* indication of the presence of a fire than the second alarm's going off. Likewise, an α-significant result with the *less* powerful test is *more* indicative of a discrepancy from H than with the more powerful test.[25] Interpreting the results accordingly, the authors' criticism disappears.

To be fair, the NP test, if regarded as an automatic "accept-reject" rule, only tells you to construct the best test for a small size α and then accept or reject. A naive use of the NP tools might seem to license the problematic inference. Rule RR is not an explicit part of the usual formulation of tests. Nevertheless, that rule, and the fallacious interpretation it guards against, is part of the error statistician's use of these tests.[26]

24. Rosenthal and Gaito (1963) explain the fallacy as the result of interpreting significance levels—*quite illicitly*—as E-R measures of the plausibility of the null hypothesis. In this view, the smaller the significance level, the less plausible is null hypothesis H, and so the more plausible is its rejection. Coupled with the greater weight typically accorded to experiments as the sample size increases, the fallacy emerges.

25. See Good 1980, 1982 for a Bayesian way of accommodating the diminishing significance of a rejection of H as the sample size increases.

26. The probabilities called for by RR would be obtained using the usual probability tables (e.g., for the Normal distribution). A good way to make use of rule RR without calculating exact severity values for each result is to substitute certain