Excursion 5: Power and Severity

Tour I: Power: Pre-data and Post-data

The power of a test to detect a discrepancy from a null hypothesis H_0 is its probability of leading to a significant result if that discrepancy exists. Critics of significance tests often compare H_0 and a point alternative H_1 against which the test has high power. But these don't exhaust the space. Blurring the power against H_1 with a Bayesian posterior in H_1 results in exaggerating the evidence. (5.1) A drill is given for practice (5.2). As we learn from Neyman and Popper: if data failed to reject a hypothesis H, it does not corroborate H unless the test probably would have rejected it if false. A classic fallacy is to construe no evidence against H_0 as evidence of the correctness of H_0. It was in the list of slogans opening Excursion 1. H is corroborated severely only if, and only to the extent that, it passes a test it probably would have failed, if false. By reflecting this reasoning, power analysis avoids such fallacies, but it's too coarse. Severity analysis follows the pattern but is sensitive to the actual outcome (it uses what I call *attained power*). (5.3) Using severity curves we read off assessments for interpreting non-significant results in a standard test. (5.4)

Tour I: keywords

power of a test, attained power (and severity), fallacies of non-rejection, severity curves, severity interpretation of negative results (SIN), power analysis, Cohen and Neyman on power analysis, retrospective power

Excursion 5 Tour II: How not to Corrupt Power

We begin with objections to power analysis, and scrutinize accounts that appear to be at odds with power and severity analysis.(5.5) Understanding power analysis also promotes an improved construal of CIs: instead of a fixed confidence level, several levels are needed, as with confidence distributions. Severity offers an evidential assessment rather than mere coverage probability. We examine an influential new front in the statistics wars based on what I call the diagnostic model of tests. (5.6) The model is a cross between a Bayesian and frequentist analysis. To get the priors, the hypothesis you're about to test is viewed as a random sample from an urn of null hypotheses, a high proportion of which are true. The analysis purports to explain the replication crisis because the proportion of true nulls amongst hypotheses rejected may be higher than the probability of rejecting a null hypothesis given it's true. We question the assumptions and the altered meaning of error probability (error probability$_2$ in 3.6). The Tour links several arguments that use probabilist measures to critique error statistics.

Excursion 5 Tour II: keywords

confidence distributions, coverage probability, criticisms of power, diagnostic model of tests, shpower vs power, fallacy of probabilistic instantiation, crud factors