

Statistical Analysis of IOI Scoring

Gordon V. Cormack

I am using regression and bootstrap techniques to try to determine the sensitivity of IOI results to chance, to particular test cases, and to scoring variants such as the batching of test cases.

The main aim is to develop a methodology for evaluating hypotheses such as: would eliminating the 50% rule introduce more than a chance variation in the rankings; do particular test cases have more substantive effect on rankings; what variance in rankings is due entirely to chance?

In preliminary analysis I have used linear regression to recover the scoring weights of particular test cases. Although this was done as just a preliminary exercise - of course the score is a linear combination of the test cases - I discovered that a handful of the cases are totally linearly dependent; i.e. *every* participant received the same score on multiple cases.

Predicting rank, rather than raw score, is perhaps more enlightening. Rank by itself has poor algebraic properties, so I have used backward linear regression on $\text{logit}(\text{percentile rank})$, which spreads rank over the range $-\infty \dots +\infty$ and yields an equal step size between first and second, and also between last and second-to-last.

While this investigation is incomplete, it appears that a handful of test cases have much more impact ($p < .001$) than the rest. These include some, but not all, of the *hard* cases, as well as some of the easy cases. It remains to be evaluated how stable the reverse stepwise elimination process was. That is, whether minor perturbations would identify a different set of significant test cases. It also remains (assuming the selection is stable) to perform qualitative analysis of what distinguishes these cases.

Bootstrapping is used to try to determine the repeatability of any given ranking. One may assume that *all* the test cases are drawn at random from a source population of possible tests, and determine the expected correlation, with confidence intervals, among the different possible test cases that might have been selected. One may also stratify the results, for example, by fixing the number of test cases for each problem, or for the 50% rule, and so on.

I would like to present some preliminary results and to solicit feedback during the workshop on these and other ways of *measuring* the impact of test case design (and more generally, on scoring).