Discussion Paper Economic E-Journal 2019-65 (December 2019)

Inference in economic experiments

Response to the comments of referee 1

This is a very useful paper, reminding the reader what the requirements are for establishing causation in experiments. The paper also contains a good discussion of external validity. This paper is timely, given the awarding of the 2019 Nobel Prize in Economic Sciences to three pioneers of using randomised controlled trials (RCTs) in development economics. I have only a couple of minor comments on the paper.

- *The discussion of randomised controlled trials versus non-randomised controlled trials, on pages 2 and 3, could give the impression that RCTs do not involve before and after comparisons, when typically they do.*

As regards statistical inference, the important thing is to clearly distinguish **between-subject designs** (RCTs) from **within-subject designs**. The essential feature of former is the random assignment of treatments, i.e. the causal (ceteris paribus) argument in RCTs is based on comparing (independent) treatment groups across which confounders are balanced. In within-subject designs, in contrast, no probabilistic assignment procedure is used and all subjects are subjected to the treatment under investigation, i.e. the causal (ceteris paribus) argument is based the researcher's capacity to hold everything else but the treatment constant before and after the treatment ("over time"). We use "within-subject designs" and "before-and-after-treatment comparison" as interchangeable terms to highlight the fact that the crucial "without-treatment" vs. "with-treatment" comparison in these experiments coincides with the before-and-after comparison (two measurements per person), while in RCTs (one measurement per person) it does not. To clarify things, we have added a footnote that explicitly explains our usage of terms.

- *The paper correctly makes the point that a low p-value tells us nothing about the external validity of the results. This is presumably also true of studies based on observational data. It would have been useful to include a discussion of why external validity is more problematic (assuming it is) for experimental studies than for studies using observational data.*

The focus of our paper is on inference in experiments. We make the point that the assessment of internal validity (causality) can be aided by statistical inference based on *p*-values when there was randomization – even when the subjects under study were not randomly recruited from a larger population. However, we also make the point that using *p*-values as an aid for assessing external validity and generalizing from experimental subjects to a broader population requires that they are randomly drawn from this population. The same probabilistic requirement applies in observational studies. Non-compliance with the "empirical commitment" of random sampling precludes a meaningful use of inferential statistics in the inductive exercise of generalizing from a sample to its parent population, except when deviations from random-sampling are adequately corrected for (e.g. through sample selection models). This is indeed an important issue in observational studies since analysed data are often obtained from convenience samples that do not meet the underlying probabilistic assumptions and that might therefore be biased in unknown ways. Nonetheless, many studies tacitly proceed as if they had a random sample and follow a misguided routine of always displaying *p*-values. This is, however, an issue beyond the scope of this paper, in which we focus on inference in experiments and which we intend to be short.