Nonperfect Randomization: What Do We Do When Logistics Get in the Way of an Experiment?

Hunter Busick

October 25, 2024

One of the first concepts that any introduction to statistics course will cover is simple random sampling. The simple random sample is the backbone of conducting research and experiments in which you need a group of participants or observations to be split up to test effects and differences amongst groups. At its core the simple random sample gives every member of a specific population an equal probability of being selected for a given sample. Once you cover this you then likely move onto more complex types of simple random sampling such as clustering or stratifying groups. These techniques are more advanced than the simple random sample and allow us to fine tune our experiments by adjusting our proportions of our sample or selecting entire groups at random instead of individuals. A student would then likely be given a series of scenarios and be asked to give which sampling technique makes the most sense given a set of highly fictionalized and unrealistic circumstances. This would then conclude the intro to sampling and would then prompt another topic.

What is not accounted for in these fictional examples of simple random sampling are the logistical issues that arise when performing an experiment or study in practice. All sorts of biases begin to present themselves once we get into the wild. Voluntary response bias appears when we reach out to a large group and only those who have a reason to respond do. Others can taint an experiment with overly extreme or overly neutral responses, leaving the data unrepresentative of what a participant believes. Performing a sample experiment is much more difficult than intro to statistics had us believe. This is why oftentimes our data collection methods are not perfectly randomized, and we must instead do the best we can with the logistics that we are presented with. But then how are we supposed to interpret results when getting a perfectly randomized sample can seem so far-fetched?

This paper examines the setting, assumptions, testing procedures, results, and conclusions found in "Dealing with imperfect randomization: Inference for the highscope perry preschool program" by James Heckman, Rodrigo Pinto, and Azeem Shaikh. The HighScope Perry Preschool study was a study performed in Ypsilanti, Michigan on preschool aged children. This was a longitudinal study in which the participants were followed up with at multiple points throughout their life to gather data on certain life metrics. The 123 students were at the time identified as being at high risk for school failure and assigned into either a treatment group or a control group. The treatment group consisted of a 2.5-hour preschool program

during the week, supplemented by in home visits from a teacher of the program. The control group did not have access to this program.

So far this sounds simple, take each child within our entire sample population and assign them a 50% chance of being assigned to the control group, and a 50% chance of being assigned to the treatment group. To keep gender and socioeconomic variables consistent within the two groups there were children who were randomly shifted from one group to the other to make these match. However, we then have a logistical conflict that causes a shift in our ratio. There were numerous students that were assigned to the treatment group however since the mother was working a job that did not allow for the in-home visit, these students were shifted into the control group for no reason other than the mother was working. This is the exact moment where our sample is no longer perfectly random as we have assigned participants to a group based on an external variable (whether the mother was working a job that conflicted with in home visits) rather than exclusively leaving it to probability.

Despite this error in random sampling, the study went on to have significant findings consistent with the idea that introducing academically at-risk children to early education programs was correlated with more positive life outcomes such as greater wealth and less incarceration. These findings were all significant under the assumption that the groups had been perfectly randomly assigned, but what if we wanted to consider the fact that groups were nonrandomly reassigned after initial assignment?

After giving a similar introduction to the program as the one I have given above, Heckman and colleagues go onto define their setup and key assumptions for their methods. They define the treatment status of an entire family as a binary classifier equal to one if a certain sibling of a particular family was placed in the program. This model consists of gender, socio-economic status, and Stanford-Binet IQ score along with whether any of that given child's siblings had been selected for the program in the prior cohort. They then go on to define a series of assumptions regarding assignment. These consist of making sure only variables used to define treatment status are the ones that can affect the potential outcomes on whether the child has a mother that can or cannot attend the in-home visits, which were earlier defined in the setup. Additionally, they lay out assumptions that ensure the ordering in which participants are assigned is irrelevant and that the two groups remain proportional in size. Finally, an assumption is stated that makes it clear that when a family has a single mother that is working, it is then evident that the family has a mother that is working

The authors then move onto testing procedures in which they develop methods for testing a single null hypothesis which states that the program has no effect on a defined subset of outcomes. To test this, they define a test statistic which provides evidence against the null for larger values of the statistic. They then calculate a critical value for this test statistic that controls the type 1 error rate.

The authors then perform testing of multiple null hypotheses, each of which state that the program has no effect on the previously mentioned life outcomes. They use stepwise testing algorithms to attempt at controlling the familywise error rate. They familywise error rate was chosen here because of an increased importance in ensuring any changes in life outcomes are because of the program and not from external factors. The familywise error rate is the probability of making more than 0 type 1 errors which in our case a type 1 error suggests

finding a cause for change in life outcomes not correlated with the effects of the treatment program defined in the setup of the methods. Each null hypothesis is then tested, and critical value subsequently adjusted to the point where we can no longer reject additional null hypotheses.

This developed methodology is then directly applied to the data from the HighScope Perry original study. Using the methodology to control for imperfections in the randomization process revealed still that the effects of subjecting academically at-risk children to a preschool program are greatly in favor of improving those individuals later success in life. Outcomes were broken down into "blocks" which consisted of IQ metrics, achievement metrics, educational metrics, criminal activity metrics, and 3 metrics per participant for employment activity at the ages of 19, 27, and 40. The findings and tables of accounting for imperfect randomization vs. the original study were similar. An interesting result is found for the educational outcomes metric where the program proved to be significant in affecting schooling outcomes for females in the study, however not for males. The other six "blocks" proved to be consistent with the original study even after accounting for imperfect randomization of groups.

There are many key applications and takeaways from this paper that contribute to the current literature as well as give guidance towards future use of this model. The authors note that this method simply lays a framework that held under weak assumptions, however also state that if further improved upon, could reject statements that criticize the validity of an experiment when its randomization techniques are imperfect due to logistical issues. We should aim to have a model such that for a given number of imperfectly randomized participants in a sample we can perform a series of analysis which allow us to adjust our results in an appropriate manner. This model would allow us to revisit prior experiments performed with sampling errors, as well as continue with future experiments that may present logistical issues, knowing that we will be able to present our results in a way that accounts for these sampling errors produced from logistical issues. This will result in an increase in efficiency in the community as less samples are thrown out for being non-representative. A new problem this will present is keeping track of what variables and logistical issues are causing an experiment to face a nonrepresentative sample population. Future guidance for improving this model likely involves subjecting it to highly nonrepresentative datasets in which the factors causing the sampling error are known. I look forward to working with this model and subjecting it to different data and adjustments to generalize its use case beyond the findings of this paper.

References

Heckman, James, Rodrigo Pinto, and Azeem Shaikh. "Dealing with Imperfect Randomization: Inference for the Highscope Perry Preschool Program." Journal of Econometrics, 1, 243 (December 2023). https://doi.org/10.3386/w31982.