Katie Clayton
Mentor: James Honaker
Privacy Tools Project
August 3, 2017

Future Directions with J-PAL Dataverse

This summer, I worked on differentially private replications of published social science research data. My goal was to test how well PSI's differentially private algorithms perform across a variety of datasets and methods of analysis. Although I did not have time to replicate any of its analyses this summer, the Abdul Latif Jameel Poverty Action Lab (J-PAL)'s dataverse is a promising area for future replication work. J-PAL conducts research in poor regions around the world with the goal of reducing poverty by ensuring that scientific evidence informs policy. Most of their data and replication code is published online, which makes it possible for us to conduct replications. Moreover, the analyses that J-PAL conducts often have large sample sizes and include individual-level data on private or sensitive matters. Finally, the types of analyses tend to be relatively simple in nature (e.g., univariate statistics and means tests based on randomized controlled trials), so the PSI library already includes algorithms that could be applied to the J-PAL data. Below, I offer three examples of J-PAL studies that we could use to conduct future differentially private replications. All are included in the online replication corpus.

1. Banerjee, A., et al. (2016). The Long-term Impacts of a "Graduation" Program: Evidence from West Bengal. *Working paper*. Retrieved from https://economics.mit.edu/files/12031.

This analysis evaluates the long-term effectiveness of an anti-poverty program in West Bengal, India, by measuring outcomes across three waves of panel surveys. Participation in the program was randomly assigned to households, and the surveys measured a series of outcome variables including assets, food security, financial inclusion, income, physical health, mental health, political involvement, and stress. The data was collected at the household level and at the individual level (adult level), and the type of analysis was difference of means. The sample size for the individual level variables ranged from 1500 – 1900 across three waves of the panel survey.

If we were to replicate this analysis using differential privacy, it would be best to focus on the results for Table 2, which includes the individual-level data. Although the number of observations is on the smaller end of what we would consider ideal for PSI's differentially private algorithms, the three panels would allow us to compare the efficacy of PSI's algorithms across three different sample sizes over time (similarly to the Pew Research Center time-series example). In this case, we would use PSI's mean.release function to compare the true means to the differentially private means for the five individual-level variables.

The data and code for this analysis is available at:
https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi%3A10.7910/DVN/SYCXOS

2. Banerji, R., Berry, J., & Shotland, M. (2013). The Impact of Mother Literacy and Participation Programs on Child Learning: Evidence from a Randomized Evaluation in India. *3IE Draft Grantee Final Report.* Retrieved from

http://www.3ieimpact.org/media/filer_public/2013/10/25/the_impact_of_mother_literacy_programs_on_child_learning.pdf

Similarly to the previous example, this study ran a randomized controlled trial to evaluate the impact of mother literacy and participation programs on their children's learning outcomes in India. They randomly assigned households to receive one of three different types of intervention or no intervention, and measured a series of outcomes including the mothers' and children's test scores, women's empowerment, mothers' participation in their children's learning, and the presence of education assets in the home. The sample size was around 8000, and the main analysis type was difference of means.

Not all of this data would be appropriate for a differentially private replication, as some variables were collected at the household level, but we could conduct a differentially private replication on any of the outcome variables collected for mothers or children. Tables 5-10 and 12, for example, use individual-level data for mothers and children to compare the mean values for various outcome variables across the treatment and control conditions. Once again, we would use the mean.release function, and we could compare how well the algorithms function across a variety of outcome variables concerning different topics (some of which could be considered more private than others). Additionally, the large sample size of this study makes it a particularly promising example to use.

The data and replication code for this study are available at https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PXV79W

3. Beaman, L., et al. (2012). Female Leadership Raises Aspirations and Educational Attainment for Girls: A Policy Experiment in India. *Science 335*(6068), 582-586. doi: 10.1126/science.1212382

Finally, this J-PAL example, which was published in *Science*, reports the results of a natural experiment in which 495 randomly selected villages in India enacted a law reserving leadership positions for women in village councils. They measured adolescents' aspirations and educational outcomes (n = 3680), and the aspirations of parents for their children (n = 6140), in the treatment and control villages. Their method of analysis was difference of means tests. They also calculated the difference in means between boy and girl adolescents to determine the extent of a gender gap in aspirations and educational outcomes.

PSI's differentially private algorithms could be applied to the individual-level survey data collected in this analysis. Similarly to the two previous examples, we could use PSI's mean.release function to calculate the differentially private means in the treatment and control groups. The sample size would likely be large enough to produce accurate results, and we could compare the true and differentially private means across the several outcomes that the researchers measured.

The data and replication code for this study are available at: https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/PXV79W

All of these examples are similar in that they employ or exploit randomized assignment to a program or policy and use difference in means tests to measure the impact of the programs across treatment and control conditions. The individual-level survey data could be considered private in nature (and sometimes is indeed private, particularly the data on health outcomes, finances, or test scores), and most of the sample sizes are large enough for the mean.release functions to operate effectively. These examples, and others available on the J-PAL dataverse, are fruitful areas for future differentially private replications.