Contents

15

Guest lectures and hybrids of experimental and observational data

We had two guest lectures by people using experimental design and developing new methods to handle the new problems. We also had a lecture on methods to mix some randomization in with what would otherwise be an observational causal inference.

15.1 Guest lecture by Min Liu

We had a lecture by Min Liu from LinkedIn. Min Liu has an M.S. in Statistics from Stanford where she took this course. Her talk was entitled "Online Experimentation at LinkedIn". They face great challenges in measuring the causal impact of changes to their product.

People with LinkedIn accounts (members) are connected to each other. Changes to the experience of one member might affect behavior of others. That may then produce al SUTVA violation.

Very small and hard to detect effect sizes can be economically meaningul because of the scale (675 million members at the time of that presentation).

There are over 3000 different metrics to track when deciding whether to launch a new feature or not. It is not reasonable to expect a change that improves some metrics to not be detrimental to some others.

They like where possible to get an experiment to completion within two weeks.

They need to go beyond mean responses and they find that quantiles are very useful. For instance a variable like page load time is important to the user experience. Raising all page load times by 0.5 seconds is meaningfully

different from raising 10% of them by 5 seconds. They compare 50'th and 90'th percentiles within the A and B pupulations. Comparing quantiles is more complicated than comparing means and bootstrapping is too slow at scale.

Some networks can be chopped up in to pieces that barely overlap at all and then treatments can be randomized to those pieces. This becomes very difficult in networks of people where some may have thousands of neighbors.

A second SUTVA violation arises in two-sided markets visualized as bipartite graphs. Think of links between advertisers and members. An experiment on one side of the graph can affect participants on the other and indirectly spill over to the first side. What that means is that, for instance, a difference observed between members in treatment and control groups might not end up as the real difference seen when making the change for all members.

15.2 Guest lecture by Michael Sklar

We had a lecture by Michael Sklar, a PhD candidate at Stanford working with Professor T.-L. Lai, entitled "Trial Design for Precision Medicine + Applications to Oncology". His lecture focused on the high and rising costs of pharmaceutical research in the US and how this is spurring the development of new complex experimental trial designs. There is an especially great need for new designs for cancer drugs because drug development for oncology has an unusually low success rate (3 percent versus 20 percent outside of oncology).

One method he described is the ***basket trial*** where for intance one drug is tested against multiple cancer type within one experiment. Another is the ***umbrella trial*** in which multiple drugs are tested against one cancer type. The third kind was the ***platform trial*** where, similarly to a bandit method, the protocol calls for algorithmic addition or exclusion of new treatment arms over time. A platform trial might also be a basket trial or an umbrella trial. The term ***master protocol*** is used to describe basket, umbrella and platform trials.

15.3 First hybrid

Sometimes we have observational and experimental data on the same phenomenon. It would be worthwhile to use them both together, especially if the resulting method is better than either of them on their own.

In other settings we might face resistance to doing an experiment. It may then still be possible to inject a small amount of randomness into a plan to gather data.

This lecture presented results from Rosenman et al. (2018) on merging a small experimental data set with a larger observational one. The motivating setting is that a large insurer or national health organization might have enormous observational records along with a small randomized clinical trial on the same disease.

One of the methods was based on a causal inference approach involving ***propensity scores***. The propensity $e(\boldsymbol{x}) = \Pr(W = 1 \,|\, \boldsymbol{x})$ is simply the chance of getting an experimental treatment given the covariates $\boldsymbol{x}$. See Imbens and Rubin (2015) for an explanation of how propensity methods can be used to estimate a causal claim as well as the additional assumptions one must make in order for the causal interpretation to be justified. One approach to estimating the causal effect of a treatment is to stratify a population based on their values $e_i = e(\boldsymbol{x}_i)$. The treatment effect in each stratum is estimated by the simple difference between average Y values for control and treated stratum members. The overall treatment effect is a weighted average of stratum values.

The first proposal in Rosenman et al. (2018) is to simply find ***counterfactual propensities*** $e(\boldsymbol{x}_i)$ for subjects i in the randomized trial. Those subjects are then added to the corresponding propensity strata of the observational data and contribute to the averages there. This is called the ***spike-in*** method. There are several other proposed methods some designed to fix possible biases in the spike-in method.

The ***Women's Health Initiative*** has data of this type relating hormone therapy to coronary heart disease (among other responses). It was a good test case for these methods because it had both observational and experimental studies of this issue. Furthermore, the experiment was large enough that it could be split into two subsets, with one of them held out to define the true treatment effect and the other combined with the observational data to estimate that effect. The spike-in method turned out to have less bias than simply using the large observational data set and less variance than using just a smallish experiment and less mean squared error than either study had on its own.

15.4 Second hybrid

The second hybrid method from that lecture was about the tie-breaker design as analyzed by Owen and Varian (2020). That design inserts some randomness into a ***regression discontinuity design*** or RDD. In an RDD we have an assigment variable x with a threshold t. Subjects with $x_i > t$ get the treatment while subjects with $x_i \leqslant t$ get the control. In an observational setting we might suppose that subjects with x_i barely larger than t are almost the same as subjects with x_i barely smaller than t at least in terms of how they would respond to the treatment. An RDD then looks for a discontinuity in the regression function $\mu(t) = \mathbb{E}(Y \,|\, x)$ at the point $x = t$. The size of the discontinuity may have a causal interpretation. See Imbens and Rubin (2015) for more.

In a tie-breaker design there are potentiallytwo thresholds A and B with $A \leqslant t \leqslant B$. If $x_i \leqslant A$ then subject i gets control. If $x_i \geqslant B$ then subject i gets treatment. If $A < x_i < B$ then subject i gets the treatment with probability 1/2. Tie-breaker designs have been used to award scholarships (!).

The paper Owen and Varian (2020) was motivated by loyalty reward programs that companies might offer to their best customers. For instance they might offer an upgrade to the top 10% of customers ranked in some way appro-

priate to the business. In a tie-breaker they could offer it instead to the top 5% of customers and randomly to half of the next 10% of customers.

The analysis in Owen and Varian (2020) shows that statistical efficiency is monotonically increasing in the amount of experimentation. Of course there is a cost issue preventing one from just making all of the awards at random. The paper analyzes that tradeoff. It also shows that there is no benefit to making the award probability take values other than 0%, 50% or 100%, perhaps on a sliding scale.

Bibliography

Imbens, G. W. and Rubin, D. B. (2015). *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press.

Owen, A. B. and Varian, H. (2020). Optimizing the tie-breaker regression discontinuity design. *Electronic Journal of Statistics*, 14(2):4004–4027.

Rosenman, E., Owen, A. B., Baiocchi, M., and Banack, H. (2018). Propensity score methods for merging observational and experimental datasets. *arXiv preprint arXiv:1804.07863*.