Abstract Title Page
Not included in page count.

Title: Using Propensity Score Matching Methods to Improve Generalization from Randomized Experiments

Author: Elizabeth Tipton

Abstract Body

Limit 5 pages single spaced.

Background / Context:

Experiments are the gold standard for research in education. The randomization of units to treatment conditions affords experiments high internal validity. That is, on average the only difference between the treatment and control groups is the *causal* effect of the treatment. This is in comparison to observational studies, in which it is difficult to disentangle the effect of a treatment and the mechanism by which units select their treatment conditions.

Unfortunately, experiments do not completely circumvent *selection*, since units rarely enter the experiment randomly. The number of social experiments that have drawn their experimental samples using random sampling methods is very very small (Shadish, Cook and Campbell, 2002). This means that units – e.g. individuals, schools, districts, or even states –enter an experiment through a non-random selection process. For example, researchers may target a small subset of schools in a state for recruitment, and of these, only certain types of schools may agree to the experimental protocols. If we are certain that the effect of a treatment is constant or additive, then this sort of non-random selection is not problematic. However, we argue that if we believe that the effect of a treatment may vary, then this selection process matters.

The main result of an experiment is typically an estimate of the average treatment effect (ATE) and its standard error. To see how non-random selection can introduce bias into an estimate of the population ATE, assume the simple case in which a single variable X, as a result of non-random selection, has a different distribution in the experiment and population. If the estimate of the ATE is the difference $\hat{\tau} = \overline{Y}_T - \overline{Y}_C$, $Y_{ti} = f_t(X_i)$ and $Y_{ci} = f_c(X_i)$, then it can be shown that

$$E(\text{bias}(\hat{\tau} \mid \tau)) = (\beta_T - \beta_C)(\overline{X}_E - \overline{X}_P) .$$

Here the expectation is over the within-study randomization of units to treatment and control groups. The point is that non-random selection leads to bias in the estimator of PATE when $\beta_T \neq \beta_C$, as happens when X is also a treatment effect moderator. In most experiments, the number of covariates that may be moderators is large.

One way we typically skirt this issue is by interpreting the ATE as the average effect for "some" population. Cornfield and Tukey (1956) famously explained the process of generalization as involving two bridges. The first bridge, they argued, is statistical, from the sample in hand to "some" putative population like it. The second bridge, then, is a subject-matter span from this putative population to the one truly of interest. Indeed, this is how experimental results are interpreted in the policy context. Policy makers take an estimate of the ATE and its accompanying standard error and ask how "like" the population they are concerned with is to the one in the study. This reasoning is generally qualitative, and may involve comparing a few univariate statistics on a small number of variables. In comparison to the first bridge, this second bridge is largely astatistical.

Purpose / Objective / Research Question / Focus of Study:

The focus of this paper is to develop a method for making this second bridge in generalization a statistical bridge. The goal is to formalize the process of moving from a non-random sample in hand to making inferences about the estimate and standard error of a treatment effect in a particular, policy relevant population. The method we propose is an extension of propensity score matching, which is commonly used in observational studies to adjust for the

process of selection into treatment. Here we propose using propensity scores to adjust for the process of selection into the experiment. This application of propensity scores was first proposed by Hedges and O'Muircheartaigh (under review), but with a slightly different focus.

The theory and method of propensity scores was introduced by Rosenbaum and Rubin (1983). In this first paper, they define the propensity score $e(\mathbf{X}) = Pr(Z_t = 1|\mathbf{X})$, where Z_t is an indicator for a unit selecting the treatment condition, and **X** is a vector of covariates that are related to the propensity to be in the treatment. They show that the propensity score has two important characteristics. First, it is a balancing score, in the sense that two units with the same $e(\mathbf{X})$ value on average will have the same values on all the variables in **X**. Second, they show that $e(\mathbf{X})$ can be interpreted as a probability of being assigned to the treatment condition, much like would be found in a randomized experiment. Additionally, they show that these propensity scores can be estimated using a logistic regression model.

As an extension to Cochran (1968), Rosenbaum and Rubin (1984) introduces a subclassification estimator that uses the propensity score. The most widely cited result of this work is that by stratifying the units in the observational study into five equally sized strata based on the distribution of $e(\mathbf{X})$ in the treatment group, approximately 90% of the bias in the original estimate of the ATE can be removed. They show that this result is true under all of the conditions studied by Cochran, when there is a monotonic relationship between $e(\mathbf{X})$ and the outcome in the two groups.

This paper develops a propensity score based method for generalization that builds upon this previous work. In addition to data from the experiment, the method we propose requires information on the population of interest. For example, a population data set could consist of a state administrative data system, a census, or a probability survey. Our method requires that either the units in the experiment can be located in the population data set, or that variables collected in the experiment can be matched to variables collected in the population. Note that the outcome variable only needs to be collected for those units in the experiment. Here we let $e(\mathbf{X}) = Pr(Z_e = 1|\mathbf{X})$ be the probability that a unit in the population is in the experiment. We focus on the subclassification estimator, since it is easy to use and works under a variety of conditions.

Significance / Novelty of study:

While propensity score methods are well developed for the treatment selection problem – indeed there is over 30 years of research in this area – the experimental selection process is different enough that this theory cannot be directly applied without some adjustments. This is particularly true for the subclassification estimator, for which many results hinge on the distributional and functional form assumptions found in Cochran. We argue that these assumptions are easier to accept in the treatment selection case than in the generalization case. At the level of distributions, we develop an extension to Cochran (1968) and Rosenbaum and Rubin (1984) for the generalization case, with regards to the expected amount of bias reduction and variance inflation or deflation for the subclassification estimator.

Additionally, we carefully construct the assumptions needed to generalize from an experiment, those needed for using propensity score methods for this purpose, and finally, we introduce new estimands that become important in the generalization case, but that are not as important in observational studies. We focus on these at the level of distributions, but make clear the connection to the level of observed samples as well.

Statistical, Measurement, or Econometric Model:

Model and Assumptions

In this paper, we focus on three potential relationships between the sample (experiment) and population: (1) when the sample is a *subset* of the population data set; (2) when the sample and population data sets *intersect*, as happens if the population data set is itself from a sample survey; and (3) when the population and sample data sets are *disjoint*, for example if an experiment is conducted in Texas but we wish to find an estimate of the treatment effect in Florida. Note that in observational studies, we are nearly always in case (1), since often the data set containing both treatment and control units itself contains the population of interest. Additionally, we define the propensity to be in the experiment as follows.

Definition: Sampling Propensity Score (variation of Rosenbaum and Rubin, 1983)

We define the *propensity score* to be

$e(\mathbf{X}) = Pr(Z=1|\mathbf{X})$

where we assume $Pr(Z_1,\ldots,Z_N|X_1,\ldots,X_N) = \Pi_1^N e(X_i)^{Zi} (1-e(X_i))^{1-Zi}$.

Note that here by propensity we refer to the propensity to be *sampled into the experiment* not to be assigned to treatment. Furthermore, the propensity score is a balancing score, in the sense that $\mathbf{X} \perp Z \mid e(\mathbf{X})$, i.e. the conditional distribution of $\boldsymbol{X}$ given $e(\boldsymbol{X})$ is the same for the sampled ($Z=1$) and the non-sampled ($Z=0$).

**

In order to generalize from an experiment to a population, we propose that 4 assumptions are needed. Additionally, in order to use the particular method presented here, we will need an additional assumption (A5). In our paper, we carefully lay each of these out. Briefly, these are:

(A1) *Random Assignment to Treatment*: Within the experimental sample, the treatment must be assigned randomly.

(A2) *Stable Unit Treatment Value Assumptions for the Sample and Population:* SUTVA (Rubin 1986) must be met not just for all units in the experiment, but also for all units in the population of interest.

(A3) *Treatment Applicability*: A unit can have a zero probability of being in the experiment only if it has a non-zero probability of receiving the treatment in the population. It can have a unitary probability of being in the experiment only if the sample is a *subset* of the population.

(A4) *Ignorability:* Conditional on the covariates used in the model, the unit-level treatment effect is independent of the sampling mechanism, i.e. $(Y(1) - Y(0)) \perp Z \mid e(\boldsymbol{X})$.

(A5) Monotonic conditional treatment effects (Optional): The conditional treatment effect distribution is a monotone function of $e(\boldsymbol{X})$.

Note that when the sample is a *subset* of the population (Case 1 above), in many cases A5 will hold. For example, if a study is conducted in Texas and the population is the state of Texas, then it may be realistic to assume that the units with larger treatment effects were more likely to be included in the study. However, A5 is less likely to hold in the *intersect* or *disjoint* cases. In our paper, we investigate a few situations for these cases in which this assumption may be made. For the remainder of this paper, we assume that A5 holds.

Subclassification Estimator

While there are many propensity score matching methods, the focus of this paper is on creating a subclassification or stratification estimator for the treatment effect. In particular, we focus on this type of estimator, since the number of experimental units is much smaller than the number of population units, which makes close matching difficult. We define such an estimator for this purpose in generalization as follows.

Estimator: General subclassification estimator of PATE

Assume that there are *2n* units in the sample, and that *S* is a sample (experiment) and *P* is a population. Based on the distribution of $e(\mathbf{X})$, divide the units in the population into *k* strata; for each stratum calculate a stratum conditional estimated average treatment effect, and combine these using weights w_{pi} from the population P (such that $\Sigma w_{pi} = 1$). Then the stratification estimator is defined as follows,

$$\hat{\tau}_S = \sum_{i=1}^{k} w_{pi}\left(\bar{Y}_{Ti} - \bar{Y}_{Ci}\right).$$

**

Rosenbaum and Rubin (1984) *Theorem A.1*. proves that stratification on the propensity score is equivalent to stratification on the outcome. In this paper, we show that under assumption A5, stratification on the propensity score is equivalent to stratification on the conditional treatment effects. In the remainder of our paper, we focus on determining the proportion of bias reduction that can be expected for a stratification estimator in the generalization case.

For observational studies, Cochran investigated the bias and variance reduction that can be expected for stratification estimators under a series of statistical distributions. These distributional results, however, are not adequate for the generalization case. In the generalization case, we are likely to find that the population distribution of $e(\mathbf{X})$ is skewed (e.g. chi-squared or log-normal), whereas in the experiment the distribution of $e(\mathbf{X})$ is more likely to be normal. For example, if propensity scores follow a logistic model, and if many cases are not likely to be sampled, then in the logit scale there will be a long tail. Similarly some covariates may be highly skewed in the population (e.g. income), yet in an experiment in which the units are chosen for being "modal" (by some standard), the experiment will not contain any of these extreme cases.

Finally, by focusing on cases with skewed distributions in the population and normal distributions in the experiment, we are able to provide tables comparing the subclassification based estimator with *k=2,3,4,* and *5* strata to the unadjusted estimator in terms of both bias reduction and variance inflation or deflation. We find that in some cases, the variance inflation can be very large. In these cases, it may be more desirable to generalize the experiment to a sub-population of the larger population. For example, if the population is the set of schools in the state of Texas, it may be that by focusing on a particular sub-set of schools, the estimate of the PATE is more precise. In these cases, we can report an additional parameter, *p**, which is the proportion of the population contained in this sub-set, and which we refer to as the *area of generalization*.

Preliminary results are reported in the Table and Figure in the Appendix for three cases. For these cases, we assume that the population distribution is Chi-Squared with 3 degrees of freedom, and that the experimental distribution is normal, with different means and variances. For each of these cases, we report two analyses, one in which the full population is used ($p^* = 1$) and one in which the full population (Chi-squared (3)) is truncated at the value $x = F^{-1}(.99)$, where *F* is the CDF of the normal distribution used in that particular case. We do not truncate the

experimental distribution. In both cases, we report the proportion of average bias removed with *k=2,3,4,* or *5* strata, and the variance inflation as a function of the correlation ρ between the propensity score $e(\boldsymbol{X})$ and the conditional treatment effect.

The Table reveals a few important trends. First, with *k=5* strata, the average bias reduction is between 50 and 70%. However, the price in variance here can be quite large, and in all cases explored leads to variance inflation factors (VIF's) of between 2 and over 10,000 depending on the model and ρ. Importantly, for *k=2* strata, the average bias is reduced between 20 and 40%, with much smaller costs in terms of variance; in some cases, these VIF's are close or less than 1. Furthermore, by focusing on a smaller portion of the population, the subclassification estimator performs even better. In the cases studied here, this amounts to p^* values between .40 and .80, which lead to much larger reductions in average bias and smaller VIF's. Note finally, that in the last case studied, the bias appears to increase in this case. This is because by truncating the population – without even subclassifying – the average bias reduces to close to zero.

Usefulness / Applicability of Method:

In order to show the usefulness of this method, we briefly present an example using data from an experiment conducted by SRI in Texas (Roschelle *et al*, 2010). The intervention was a math curriculum for middle schoolers and the experiment took place in 78 schools across Texas. Here the focus is the school ATE for middle schools in Texas. We use the state AEIS administrative data system and match the schools in the experiment to those in the state using a set of 30 covariates, including student and teacher demographics, school structure, and prior year tests scores. Briefly, we are able to show that:

(1) The unadjusted effect is 3.26 (.35).

(2) If we generalize using a subclassification estimator with *k=5* strata, the estimate is 2.58 (2.30).

(3) If we focus instead on the sub-population that is easy to generalize to, we find $p^* = 0.70$, meaning the results of the experiment can generalize to 70% of the population. In this case, a subclassification estimator is not needed, and thus for this sub-population the estimate is 3.26 (.35).

It is important to note that one outcome of this analysis is that it clearly defines which regions of the state and types of schools that are not well represented by the experiment. Future experiments could be conducted to better represent these areas.

Conclusions:

The method and theory we present here is aimed at making the process of generalization from an experiment to a population a statistical process. One of its main virtues is that it allows the assumptions necessary for generalization to be clearly defined, and allows the experiment and population to be balanced on a large number of covariates. In addition to providing an estimator of the treatment effect that has less bias than the standard estimator, we have shown that under certain conditions there are no costs in terms of variance for using this method. Furthermore, it allows the identification of the proportion of the population that the experiment cannot easily generalize to ($1-p^*$), which can help researchers clearly define sub-populations for future experimentation. Finally, it should be noted that a weakness of this method is that the level of analysis for the treatment effect is restricted to that found in the population data set, which may or may not coincide with the level at which randomization to treatment occurred.

Appendices

Not included in page count.

Appendix A. References

Cochran, W. G. (1968). The effectiveness of adjustment by subclassification in removing bias in observational studies. *Biometrics*, *24*(2), 295-313.

Cornfield, J., & Tukey, J. W. (1956). Average values of mean squares in factorials. *The Annals of Mathematical Statistics*, *27*(4), 907-949.

Hedges, L.V. and O'Muircheartaigh, C.A. (*under review*) Improving generalization from designed experiments.

Ho, D. E., Imai, K., King, G., & Stuart, E. A. (2006). Matching as Nonparametric Preprocessing for Reducing Model Dependence in Parametric Causal Inference. *Political Analysis*, *15*(3), 199-236. doi: 10.1093/pan/mpl013.

Rosenbaum, P. R., & Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, *70*(1), 41-55. doi: 10.1093/biomet/70.1.41.

Rosenbaum, P. R., & Rubin, D. B. (1984). Reducing bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association*, *79*(387), 516-524. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/20199225.

Roschelle, J., Shechtman, N., Tatar, D., Hegedus, S., Hopkins, B., Empson, S., et al. (2010). American Educational Research Journal. *American Educational Research Journal*. doi: 10.3102/0002831210367426.

Rubin, D. B. (1986). Statistics and Causal Inference : Comment : Which Ifs Have Causal Answers. *Journal of the American Statistical Association*, *81*(396), 961- 962.

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). Experimental and Quasi-Experimental Designs for Generalized Causal Inference. Boston: Houghton-Mifflin.

Appendix B. Tables and Figures

Table: Comparison of bias reduction and variance inflation, by number of (equal) strata and proportion *p*, for a population with Chisq(3)

Experiment distribution	p*	Bias reduction					Variance Inflation			
		k equal population strata					k equal population strata			
		2	3	4	5	ρ	2	3	4	5
N(1,.5)	1	0.378	0.522	0.620	0.678	0.1	65.047	>10,000	>10,000	>10,000
						0.2	57.738	>10,000	>10,000	>10,000
						0.4	43.120	>10,000	>10,000	>10,000
						0.6	28.502	>10,000	>10,000	>10,000
						0.8	13.884	>10,000	>10,000	>10,000
	0.461	0.630	0.743	0.877	0.917	0.1	0.910	0.967	1.024	1.043
						0.2	0.843	0.877	0.920	0.932
						0.4	0.708	0.697	0.712	0.710
						0.6	0.573	0.517	0.504	0.488
						0.8	0.439	0.338	0.296	0.266
N(2,.5)	1	0.229	0.316	0.435	0.505	0.1	1.208	27.655	4078.230	>10,000
						0.2	1.111	24.557	3595.594	>10,000
						0.4	0.917	18.361	2630.322	>10,000
						0.6	0.722	12.165	1665.050	>10,000
						0.8	0.528	5.969	699.778	>10,000
	0.633	0.682	0.846	0.883	0.916	0.1	1.620	3.708	5.871	7.985
						0.2	1.481	3.321	5.229	7.093
						0.4	1.202	2.548	3.946	5.309
						0.6	0.923	1.775	2.664	3.525
						0.8	0.644	1.001	1.381	1.741
N(2,1)	1	0.232	0.353	0.474	0.544	0.1	0.947	1.507	3.454	8.525
						0.2	0.876	1.361	3.085	7.573
						0.4	0.734	1.071	2.347	5.670
						0.6	0.592	0.780	1.609	3.766
						0.8	0.450	0.490	0.871	1.863
	0.772	1.794	2.006	1.650	1.527	0.1	0.893	0.913	0.923	0.929
						0.2	0.828	0.829	0.829	0.829
						0.4	0.696	0.659	0.642	0.631
						0.6	0.565	0.490	0.455	0.433
						0.8	0.434	0.320	0.267	0.235

Note: The proportion of the population that can be generalized to is p*; ρ is the correlation between e(X) and the conditional average treatment effects.

Figure: Population (Chisq(3)) and Experiment (N(-,-)) Distributions