Small Sample Inference

Jorge Luis García[*,†]

First Draft: October 15, 2014
This Draft: October 24, 2014

Abstract

I summarize the procedure in Heckman et al. (2010) formalized in Lehman and Romano (2005); Romano and Wolf (2005) and I accompany it with an implementation in Stata.

In the following lines we write-up a methodology accounting for the following usual threads to internal validity in the evaluation of Early Childhood Interventions. Specifically, we tailor our inference methodology to account for the following: (i) compromised randomization; (ii) small sample size; (iii) attrition.

Let $\mathcal{I}$ be the index for participants of a generic intervention with typical element i. The cardinality of $\mathcal{I}$, $\#\mathcal{I}$, is the number of participants of the program, N. The observed outcome of individual i is:

$$Y_i = D_i Y_i(1) + (1 - D_i) Y_i(0) \tag{1}$$

where D_i indicates treatment status. $D_i = 1$ means that individual belongs to the treatment group and $D_i = 0$ to the control group. $(Y_i(0), Y_i(1))$ are the outcomes of participant i when her treatment status is *fixed* at control and treatment statuses, respectively. Fixing means that the value D_i takes a value exogenously. For instance, $Y_i(0)$ is the value that the outcome takes when the treatment is exogenously set to control status.[1]

Heckman et al. (2010) explain how randomization solves potential problems of selection bias. They argue that it induces independence between the counter-factual outcomes, $(Y_i(0), Y_i(1))$, and the treatment status indicator, D_i, conditional on pre-program or background variables, which we denote by X, used in the randomization protocol. Mathematically, randomization validates Assumption 1:

Assumption 1 *(Counterfactual Outcomes-Treatment Conditional Independence)* $(Y_i(0), Y_i(1)) \perp\!\!\!\perp D|X$*, where* $X, D, Y(0), Y(1)$ *are* N*-dimensional vectors. Each of these vectors' entries are associated with each participant of the program, e.g.* $X = (X_i : i \in \mathcal{I})$*.*

By design, X contains information on pre-treatment variables.

Our objective is to test the null hypothesis of no treatment effect. Put differently, we want to test if the outcome vectors, conditioned on pre-program variables, have the same distribution:

Hypothesis 2 *(Counterfactural Outcomes Equivalent Distribution)* $Y(1) \overset{d}{=} Y(0)|X$*, where* $\overset{d}{=}$ *means distribution equivalence.*

[*] Department of Economics, the University of Chicago (jorgelgarcia@uchicago.edu).

[†] I thank Sneha Elango for helpful comments.

[1] Heckman and Pinto (2013) discuss what *fixing* means in Economics and how does it differ from *conditioning* in Statistics; they discuss the link of this concepts to causality.

Actually, in this context, we can test Hypothesis 2 in a more tractable way; Assumption 1 and Hypothesis 2 imply

Hypothesis 3 *(Outcome-Treatment Independence)* $Y \perp\!\!\!\perp D|X$.

Testing Hypothesis 3 has some methodological challenges that we attempt to solve. First, the interventions we focus on have small sample sizes. This casts doubts on inference that relies on the asymptotic behavior of test statistics. We use exact permutation tests to address this problem. We tailor the tests to account for the exact randomization protocol of the program and account for the second challenge: compromised randomization. Third, non-random attrition biases our estimates. We use an inverse probability weighting scheme to consider this issue. Our weights are based on the probability of attrition conditioned on pre-program variables.

Small Sample Inference

Our permutation test is valid for small sample sizes. To construct it, we explore the invariance of the joint distribution of (Y, D) under permutations that swap elements of the treatment status vector, D. This arises from two statistical properties: (i) randomization guarantees that D is exchangeable for a set of selected permutations. Put differently, the distribution of D remains the same for a set of selected swaps of elements in it; (ii) Hypothesis 3, under which we construct the inference method.

Let us develop the first property more. Colloquially, "scrambling" the order of the participants who share the same values of X does not change the underlying distribution of the treatment assignment vector, D. Actually, X limits the set of permutations that do not alter the distribution of D, i.e. there is a set of admissible permutations.

Let $\mathcal{G}_X$ be the set of admissible permutations–that is, the set of permutations for all individuals who share the same value of X. Formally,

Definition 4 *(Admissible Permutations)*

$$\mathcal{G}_X = \{\pi_g : \mathcal{I} \rightarrow \mathcal{I} \mid \ \pi_g \textit{ is a bijection and } (\pi_g(i) = j) \Rightarrow (X_i = X_j)\ \forall\, i \in \mathcal{I}\}.$$

Thus, we can write:

Property 5 *(Exchangeability)* $D \overset{d}{=} gD\ \forall g \in \mathcal{G}_X$ *where* $gD = (D_{\pi_g(i)} : i \in \mathcal{I})$.

An appealing feature of Property 5 is that it relies on limited information about the randomization protocol. It does not require a full specification of the distribution of D or about the assignment mechanism. The variables that the protocol uses, X, are sufficient. Actually, this makes the inference method hold when the randomization compromises are based on X. In other words, the method considers the randomization process and builds on its structure.

Hypothesis 3 states that the vector of outcomes is independent of the vector of treatment assignment, D. Together with Property 5, this implies that the joint distribution of outcomes and treatment status is invariant under the set of admissible permutations, $\mathcal{G}_X$. Theorem 6 summarizes this. We follow the literature and call this theorem the Randomization Hypothesis.

Theorem 6 *(Randomization Hypothesis) Assume that Hypothesis 3 holds. Then, the joint distribution of outcomes, Y, and treatment assignment, D, are invariant under the set of admissible permutations, $\mathcal{G}_X$. This is,*

$$(Y, D) \overset{d}{=} (Y, gD)\ \forall g \in \mathcal{G}_X.$$

Proof. By Property 5, $D \overset{d}{=} gD \ \forall g \in \mathcal{G}_X$. But we know that $Y \perp\!\!\!\perp D|X$ by Hypothesis 3. Thus $(Y, D) \overset{d}{=} (Y, gD) \ \forall g \in \mathcal{G}_X$. ■

A consequence of Theorem 6 is that a statistic based on Y and D is distribution invariant under any admissible permutation, i.e. $\forall g \in \mathcal{G}_X$. Moreover, under Hypothesis 3, the exact distribution of a statistic is given by the collection of the values generated by all the permutations in $\mathcal{G}_X$ (see Lehman and Romano, 2005).

We use Theorem 6 to construct a permutation test. Let $T(Y, D)$ be a statistic associated with Y and D which large values provide evidence against Hypothesis 3. Let $c \in \mathbb{R}$ be a critical value such that we reject Hypothesis 3 if $T(Y, D) \geq c$. Hence, if we want a test with significance level α, the following must hold:

$$\begin{aligned}&\Pr(\text{Reject Hypothesis 3} \mid \text{Hypothesis 3 holds}) \\ &\qquad\qquad = \Pr(T(Y, D) \geq c | \text{ Hypothesis 3 holds}) \leq \alpha. \qquad (2)\end{aligned}$$

We take the α-quartile of the set $\{T(Y, gD) \ : g \in \mathcal{G}_X\}$ to compute the critical value. The theoretical justification of this is that the critical value can be computed by the fact that the distribution of $T(Y, D)$ is given by the set of values that $T(Y, gD)$ takes as different admissible permutations are allowed (see Romano and Wolf, 2005, Theorem 15.2.2).

It is important to note that $T(Y, D)$ is uniformly distributed across the values of $T(Y, gD)$ as different admissible permutations are allowed, i.e. with each $g \in \mathcal{G}_X$. Uniformity in this context means that each value that $T(Y, gD)$ takes is equally likely. This implies that the critical value can be computed through the α-quantile of the set $\{T(Y, gD) \ : g \in \mathcal{G}_X\}$.

In practice, permutation tests compare a test statistic computed on the original (not permuted) data with a distribution of test statistics obtained from a series of admissible permutations. The measure of evidence against the *Randomization Hypothesis*, the $p - value$, is computed as the fraction of permuted data which yields a test statistic greater than the test statistic obtained with the not permuted data. In order to generate the distribution of test statistics and compute the critical value c, we assume full enumeration of the set of admissible permutations, $\mathcal{G}_X$. However, when the method is executed, it is common to take a random sample of admissible permutations, which size is arbitrary.

This method has the following benefits: (i) it relies on assumptions that are consequences of randomized trials; (ii) it does not require full information on the randomization protocol: the knowledge of which variables were used to randomize are enough, which we denote by X; (iii) it is fully non-parametric; (iv) it does not rely on any particular choice of a test statistic; (v) it is valid for any distribution of the data so that it is not based on particular asymptotic behaviors of the test statistic.

This method is robust to the choice of any test statistic that has direct relation with evidence against the null hypothesis. The choice of the particular statistic is associated with the power of the inference, not with its validity to contrast a hypothesis. We opt for the t-statistic associated with the mean difference between the treatment and the control groups to obtain results that are comparable with standard program evaluations.

References

Heckman, J., S. H. Moon, R. Pinto, P. Savelyev, and A. Yavitz (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the highscope perry preschool program.

Quantitative economics 1(1), 1–46.

Heckman, J. J. and R. Pinto (2013). Causal Analysis after Haavelmo. University of Chicago, Unpublished Manuscript.

Lehman, E. L. and J. P. Romano (2005). *Testing Statistical Hypotheses*. Springer Texts in Statistics.

Romano, J. P. and M. Wolf (2005). Exact and Approximate Stepdown Methods for Multiple Hypothesis Testing. *Journal of the American Statistical Association 100*(469), 94–108.