ARE 213 **Applied Econometrics**

UC Berkeley Department of Agricultural and Resource Economics

Statistical Inference:

Part 2, Randomization Inference

All of the statistical tests that we have discussed so far have relied upon precise distributional assumptions (for small sample inference) or asymptotic theory and large sample approximations. Suppose, however, that we are unwilling to make distributional assumptions and that our sample is small, so that we are unsure whether it is safe to apply asymptotic approximations.[1] In that case we may want to apply a nonparametric test that relies neither on distributional assumption nor asymptotic theory. In these notes, we consider one such class of nonparametric tests: randomization inference, or permutation tests.

1 The Lady Tasting Tea and Fisher's Exact Test

Fisher (1935) presents a statistical test of a lady's claim that she can discriminate whether milk was added prior to the tea or after the tea simply by tasting a cup of tea.[2] The experiment consists of mixing eight cups of tea, four of which have had the milk added before the tea, and four of which have had the milk added after the tea. The lady is told of the experimental design and knows that there are exactly four cups of each type. She is instructed to taste each cup of tea, which are presented to her in random order, and to place them into two groups of four (milk before and milk after).

To conduct the statistical test, note that there are 70 ways to choose 4 objects out

[1] In practice, the central limit theorem takes hold shockingly quickly for most distributions – e.g., by the time you have a dozen observations, the distribution of the mean strongly resembles a normal distribution. In principle, however, you might have a really bizarre distribution. More importantly, randomization inference can be useful in cases with clustering.

[2] This episode is apparently based on an historical event involving R.A. Fisher. The lady in question reportedly passed the test with flying colors – the explanation is that pouring hot tea into cold milk causes the milk to curdle, but pouring cold milk into hot tea does not.

of 8, assuming that order does not matter (which, in this case, it does not). Formally, $\begin{pmatrix} 8 \\ 4 \end{pmatrix} = \frac{8!}{4!4!} = \frac{8 \cdot 7 \cdot 6 \cdot 5}{4 \cdot 3 \cdot 2} = 70$. Thus, there are 70 ways in which the lady could potentially divide the 8 cups of tea. If the lady has no ability to discriminate between early-milk and late-milk cups, which constitutes our null hypothesis, then the probability of dividing them such that all four early-milk cups end up in the early-milk group is 1 in 70, or 0.014.[3] This result would be statistically significant at conventional levels. Now consider the probability that three or more early-milk cups end up in the early-milk group. There are $\begin{pmatrix} 4 \\ 3 \end{pmatrix}$ ways to place 3 early-milk cups in the early-milk group, and $\begin{pmatrix} 4 \\ 1 \end{pmatrix}$ ways to place one early-milk cup in the late-milk group. Hence there are 16 ways for three early-milk cups to end up in the early-milk group, plus the one way that four early-milk cups and can end up in the early-milk group. The probability that three or more early-milk cups end up in the early-milk group is thus $17/70 = 0.24$.[4] This result is not statistically significant at conventional levels, so the lady must place all four early-milk cups in the early-milk group in order to prove to us beyond a reasonable doubt that she can discriminate between early-milk and late-milk cups.[5]

Note that this test makes absolutely no assumptions about the parametric distribution of residuals (we didn't even define a random variable ε_i) or the independence of various cups of tea. In fact, it may well be the case that cups are correlated with each other in that some came from one pot of tea while the rest came from another pot. But this is irrelevant from our perspective – all that matters is that the cups were sorted randomly before the decision to add the milk before or after the tea was made. Nor does it matter what strange form the distribution of tea taste may take. The key insight is that we take the outcomes for what they are under the null hypothesis and then map out the distribution of the test statistic, the number of cups correctly classified, that arises due to the randomization procedure. That is

[3]Formally, there are 4 choose 4 (1) ways to place 4 early-milk cups in the early-milk group and 4 choose 0 (1) ways to place 0 early-milk cups in the late-milk group.

[4]We could likewise calculate the probability of getting two cups right and two cups wrong as 4 choose 2 ways to place two early-milk cups in the early-milk group and 4 choose 2 ways to place two early-milk cups in the late-milk group.

[5]Alternatively, if she fails but shows some promise, e.g. places 3 out of 4 correctly, then we may rerun the experiment with a larger number of cups, increasing the power of the experiment.

to say, the variability in the test statistic under the null hypothesis comes from the random assignment mechanism itself rather than sampling variability that arises because we do not observe the entire population of cups of tea.

This test generalizes to a permutation test known as Fisher's Exact Test. This non-parametric test can be applied in any scenario in which there is a binary outcome and a binary treatment.[6] Let N be the number of observations. In general the test looks like:

	Treated	Control	Row Total
High	N_{TH}	N_{CH}	N_H
Low	N_{TL}	N_{CL}	N_L
Column Total	N_T	N_C	

The probability of observing any realization of this table is $p = \begin{pmatrix} N_H \\ N_{TH} \end{pmatrix} \begin{pmatrix} N_L \\ N_{TL} \end{pmatrix} / \begin{pmatrix} N \\ N_T \end{pmatrix}$. Thus, if we want to compute whether the realization we observe is too improbable to be due to chance, we simply calculate p for the realization we observed and all realizations that are "more extreme" (generally, less probabilistic) than the one we observed.[7] We then sum these probabilities to get a p-value.

For a variety of reasons, Fisher's Exact Test is not often applied in economics.[8] Nevertheless, it is very useful in demonstrating the advantages of randomization tests. Because randomization forms the basis for inference – no distributional or independence assumptions are necessary, nor do we need to use asymptotic approximations. The only thing we need to do is model the assignment procedure correctly.

[6] In the tea case the treatment might be defined as "early-milk" while the outcome is defined as "classified as early-milk."

[7] If we expect that the treatment increases the probability of a "high" outcome, then higher values of N_T correspond to more extreme realizations.

[8] Among the reasons: The outcomes must be binary – many times are not. Furthermore, if the outcomes are binary, then the CLT will take hold very quickly, so an exact test is unnecessary unless the samples are much smaller than is typical in economics. Finally, the exact test does not accommodate covariates.

2 Randomization Tests

Anderson (2008) applies a randomization test in the context of randomized trials of preschool programs. The samples in this study are as small as a dozen individuals (the control group for the smallest study); thus there is some concern that asymptotic theory may not apply. Because we know the assignment procedure (children were randomly assigned to treatment and control), it is easy to simulate the distribution of the test statistic (the difference in means divided by its standard error) under the null hypothesis.

The intuition behind the test is as follows. The preschool children, who were recruited from at-risk families in Ypsilanti, MI during the 1960s, cannot reasonably be thought of as a random sample from any larger population. In fact, we observe the entire population of at-risk children recruited for the Perry Preschool Program in Ypsilanti from 1962 to 1967. But that does not mean that any observed difference between the treatment and control groups must represent a treatment effect. Even with random assignment, the treatment and control groups will not be perfectly balanced. The question is whether the observed differences are large enough to be reasonably due to chance (randomness in the assignment procedure) or whether they represent a true treatment effect.

What we would like to do is run the experiment thousands of times under the null hypothesis of no treatment effect and record the distribution of our estimator through all of these runs. To achieve this, we impose the null hypothesis, i.e., we assume that $Y_i(1) = Y_i(0)$ for all individuals. Under this hypothesis, we can simulate the experiment as many times as we would like by randomly assigning placebo treatment indicators and recording the difference in means between "treated" and "control" groups. Using these results, we can map out the null distribution of our test statistic. We can then compare the observed test statistic from the real data to this null distribution. Note that we make no distributional assumptions at all – the variation in the null distribution of the test statistic arises from the randomization procedure which we, the experimenters, designed (or, in this case, we know how it was implemented).

For a given sample size N, the procedure is implemented as follows:

1. Draw binary treatment assignments Z_i^* from the empirical distribution of the original treatment assignments without replacement.
2. Calculate the t-statistic for the difference in means between treated and untreated groups.
3. Repeat the procedure 100,000 times and compute the frequency with which the simulated t-statistics – which have expectation zero by design – exceed the observed t-statistic.

If only a small fraction of the simulated t-statistics exceed the observed t-statistic, reject the null hypothesis of no treatment effect. This procedure tests the sharp null hypothesis of no treatment effect, so rejection implies that the treatment has some distributional effect. Formally, only two assumptions are required:

1. Random Assignment: Let $Y_i(0)$ be the outcome for individual i when untreated and $Y_i(1)$ be the outcome for individual i when treated (we only observe either $Y_i(0)$ or $Y_i(1)$). Random assignment implies $\{Y_i(0), Y_i(1) \perp Z_i\}$.
2. No Treatment Effect: $Y_i(0) = Y_i(1) \; \forall \; i$

Note that no assumptions regarding the distributions or independence of potential outcomes are needed. This is because the randomized design itself is the basis for inference (Fisher 1935), and pre-existing clusters cannot be positively correlated with the treatment assignments in any systematic way. Even if the potential outcomes are fixed, the test statistic will still have a null distribution induced by the random assignment. Since the researcher knows the design of the assignment, it is always possible to reconstruct this distribution under the null hypothesis of no treatment effect, at least by simulation if not analytically. Thus, this test always controls Type I error at the desired level (Rosenbaum 2007).

For binary Y_i, this test generally converges to Fisher's Exact Test. However, it differs slightly from Fisher's Exact Test in that Fisher's test rejects for small p-values while this test rejects for large t-statistics. This test is also similar to bootstrapping under the assumption of no treatment effect (Simon 1997); the only difference is that the resampling is done without replacement rather than with replacement. This highlights the fact that the variance in the test statistic's null distribution arises from the randomization procedure itself rather than from unknown variability in the potential outcomes.

In the paper, the randomization test produced p-values that were relatively close to those from standard t-statistics – the CLT takes hold quite quickly. Nevertheless, randomization tests can be useful in addressing doubts from people who find it hard to believe that you can definitely disprove a null hypothesis with only 30 or 40 observations. They can also be useful in challenging situations involving clustering. It is important to keep in mind, however, that they only apply to testing the null hypothesis. You should not use them to test alternative hypotheses or create confidence intervals because the null distribution is just that, the null distribution (it's not the alternative distribution).

3 Randomization Tests with Clustering

The nice thing about randomization tests is that they are immune to clustering issues *as long as the randomization correctly forms the basis for inference.* In other words, as long as you model the randomization process correctly, you are guaranteed to have a valid test of the null hypothesis. It doesn't matter if the potential outcomes are dependent or even if they are fixed numbers, because the null distribution of the test statistic is assumed to arise from the randomization in the assignment procedure. Thus, as long as you can model the randomization in the assignment procedure correctly, you can conduct a test of the null hypothesis that will have the correct size.

Consider first our case with the preschool projects. Obviously the potential outcomes are correlated across different students within the study. For example, some students will attend

Elementary School A, while others attend Elementary School B. The students attending School A will experience common shocks that the students in School B do not, and vice versa. None of this, however, affects the validity of our randomization test. One way to see this is to note that the treatment is randomly assigned, i.e., there is no serial correlation in D_i. We know from our clustering lectures that serial correlation in outcomes is not a problem unless it is accompanied by serial correlation in the treatment variable. Alternatively, however, you could simply note that we know exactly how the treatment was assigned (i.e., randomly by individual), and once we know that we don't have to worry about the distributions of the potential outcomes. Under the null hypothesis, there is no missing data problem ($Y_i(0) = Y_i(1)$ for all units), and we can construct all the counterfactual estimates of $\hat{\beta}$ under different distributions of the treatment assignment that could have arisen in alternative universes.

Now consider a more challenging case, such as a diffs-in-diffs study in which one state is affected by a treatment and a group of comparison states is not. The test that Abadie, et al. (2007) propose in this setting is a randomization test. We know how the assignment mechanism works in this case – it is turned on indefinitely for one state at some point in time and never turns on for the other states. Under the null hypothesis we observe all of the untreated outcomes for every state. We can test whether the "effect" we observe for the treated state (e.g., California in the Abadie, et al. example) is large or small compared to other realizations of the test statistic under other potential treatment assignments. The important thing is that we model the treatment assignment correctly, which in this case means that we turn it on at a given point for the "treated" state and then leave it on. If we randomly switched the treatment on and off for a given state, we would not be reproducing the original assignment procedure. In this case, we would tend to over-reject because we are assuming that the treatment was randomly assigned within a state (i.e., the treatment was not serially correlated) when in fact it was randomly assigned across states but serially correlated within states.

Finally, consider a case in which it is infeasible to cluster the standard errors using the

standard panel techniques. Aker (2010) tests the impact of the introduction of cell phones on grain markets in Niger. She uses monthly price data on 31 markets over several years; from these data she constructs 433 market pairs. She then examines the effect of a cell phone dummy, equal to unity if both market pairs have cell phone coverage and zero otherwise, on price dispersion between markets using a diffs-in-diffs type strategy. In a normal panel setting we might cluster at the market-pair level, but in fact it is impossible to construct two or more independent clusters using the market-pair data.[9] One solution is to try to implement the multi-way clustering technique clustering at both the year level and the market-pair level; however, the assumption that any common shock to market-pair 1-2 and market-pair 1-3 does not persist over time seems unlikely.[10]

Alternatively, we can implement a randomization test of the null hypothesis in which we assign placebo cellular towers to markets in random order and simulate the treatment effect hundreds of times. We "build" cellular towers in the same way that they are built in the data – i.e., there are no cellular towers initially, then one is built in an initial market and persists indefinitely, then a second is built in another market and persists indefinitely, etc. – and estimate the test statistic of interest (e.g., a t-statistic for a regression coefficient) for each simulation. This should give us an accurate test of the null hypothesis without requiring unknown and/or unrealistic assumptions on the structure of the variance-covariance matrix. Inference is based on the structure of the treatment rollout, combined with the null hypothesis of no treatment effect.

4 Additional References

Aker, J. "Information from Markets Near and Far: The Impact of Mobile Phones on Grain Markets in Niger." *AEJ: Applied Economics*, 2010, 2, 46-59.

[9]For example, consider clustering at the market level. The first cluster contains all market pairs with market 1. But by definition it also contains pairs with markets from every other market cluster, so we can't construct non-overlapping clusters.

[10]Remember, multi-way clustering assumes that if two observations are different along both of the cluster dimensions, then there is zero correlation. In this case, that implies that if you are looking at two different market pairs in two different time periods, they should have zero correlation.

Fisher, R.A. *The Design of Experiments*, 1935, Oliver and Boyd: Edinburgh and London.