PSC 504: Fisher's Randomization Inference

Matthew Blackwell

1/31/2013

Why randomization inference?

Effect of not having a runoff in sub-Sarahan African

- The data below comes from Glynn and Ichino (2012) and shows how not having a runoff ($A_i = 1$) is related to harrassment of opposition parties (Y_i). To test this, they collected data on 10 sub-Sahara African countries.

Unit	$Y_i(0)$	$Y_i(1)$	A_i	Y_i
Cameroon	?	1	1	1
Kenya	?	1	1	1
Malawi	?	1	1	1
Nigeria	?	1	1	1
Tanzania	?	0	1	0
Congo	0	?	0	0
Madagascar	0	?	0	0
Central African Republic	0	?	0	0
Ghana	0	?	0	0
Guinea-Bissau	0	?	0	0

- The central idea behind this study is that when there are no runoff elections, incumbents are able to win elections with only a plurality of votes. Thus, they don't need to court any smaller or opposition parties. In fact, without a runoff, there are incentives to suppress the turnout through itimidation. When there is a runoff ($A_i = 0$), the incumbent/largest party needs to gain wider support and is more likely to court smaller parties as opposed to itimidating them.
- Even though there is a strong difference in means between the control and treated group (0.8), it is unclear if we can say anything causal about these data at all. There is a very small sample size we have little idea about the distribution of the outcome.
- For today, we are going to assume that electoral institutions were randomly assigned to these countries, though, note that the original goal of the Glynn/Ichino paper is to use case-study information in place of such a randomization assumption.
- The great thing about randomization inference is that it will help us assess causal claims in experimental or observational data without any need to appeal to a model or large samples.

What is randomization inference?

- Randomization inference is all about using nothing but the act of physical randomization to make inferences about causal effects. Most of the time, RI is about hypothesis testing. Fisher was the first to demonstrate the usefulness of physical randomization in this way. This differs from last week's discussion of average effects because here we focus on formulating null hypotheses that allow us to fill in the missing potential outcomes for each unit.
- In general, there are three components to randomization inference: a null hypothesis, a test statistic, and a measure of extremeness.

Brief review of hypothesis testing

- Remember the idea behind hypothesis testing: we want to formulate a null hypothesis that usually represents a fact about about the data we would like to refute. Usually this comes in the form of a "no effect" hypothesis. Last week we might have written an a hypothesis such as $H_0 : E[Y_i(1)] - E[Y_i(0)] = 0$, which is the null hypothesis of no average treatment effect. The crucial part of hypohtesis testing is that when we assume the null hypothesis is true, it is usually straightfoward to derive the distribution of the data and statistics of the data. Once we know the distribution of some statistic, we can compare that statistic in the observed data to see how likely or unlikely that observation was, under the null hypothesis. And we take this as evidence for or against the null hypothesis.
- Usually we test a hypothesis of no effect, but we're not limited to that. In fact, a useful way to construct a $100 * (1 - \alpha)\%$ confidence interval is calculate all null hypotheses that where we cannot reject the null at the α level.

Sharp null hypothesis of no effect

- No effect means no effect: $H_0 : \tau_i = Y_i(0) - Y_i(1) = 0$ for all units.
- Note that this is different than the null hypothesis of no **average** treatment effect, which does not imply the sharp null. Under the null of no average effect, there still could be positive effects for some units and negative effects for other units.
- This null hypothesis formally links the observed data to all potential outcomes. In fact, it allows us to fill in the missing potential outcomes of our data:

Unit	$Y_i(0)$	$Y_i(1)$	A_i	Y_i
Cameroon	1	1	1	1
Kenya	1	1	1	1
Malawi	1	1	1	1
Nigeria	1	1	1	1
Tanzania	0	0	1	0
Congo	0	0	0	0
Madagascar	0	0	0	0
Central African Republic	0	0	0	0
Ghana	0	0	0	0
Guinea-Bissau	0	0	0	0

- Now we have a complete dataset under the null hypothesis. Remember that the potential outcomes are fixed in this setup, so all we are doing is using the sharp null hypothesis to assume knowledge of these fixed, but unobserved quantities.
- Note that we have chosen the sharp null hypothesis to be that there is no effect so that $\tau_i = \tau = 0$, but we could easily choose another hypothesis such as an additive effect: $H_0 : \tau_i = 0.2$. This implies that for each treated unit we can calculate their potential outcomes under control easily: $Y_i(0) = Y_i(1) - 0.2$. More generally, we have null hypotheses of the form $H_0 : \tau_i = \tau_0$ for some fixed value τ_0.
- When we have a non-o null, then note that the observed outcomes will change if the treatment assignments change. Thus, it is easier to calculate the test statistic for a quantity that does not varying with the treatment assignment. This quantity is called the **adjusted outcome** and it $Y_i - A_i\tau_0$ where τ_0 is the value of the constant treatment effect under the null distribution. Under the null, this is also equal to $Y_i(0)$ and thus doesn't vary with A_i.

Test statistic

- To assess the evidence for or against the sharp null hypothesis, we need to specify a test statistic. The test statistic is just a function of the treatment assignment and the response: $t(A_i, Y_i)$. In general, this is just some measure of the relationship between these two variables, but under the null hypothesis, it becomes a causal quantity because $t(A_i, Y_i) = t(A_i, Y_i(0))$. Thus, under the sharp null we have a test of a causal quantity.
- This test depends on what test statistic we use and this choice will affect the power of our test against various alternatives. If we have some idea about the type of effect we are likely to see, this can guide us. Suppose we thought that the effect will shift some part of the distribution of the potential outcomes (that is, $Y_i(1)$ will have a somehow different distribution than $Y_i(0)$), then we should choose a test statistic that will measure those shifts. We might also want to choose test statistics that are robust against outliers if that might be a problem in our data.
- Note that the test statistic need not estimate a direct causal effect of any sort, since their only purpose in Fisher's RI is to test the sharp null.

Difference in means

- This is the usual difference in means estimator

$$T_{diff} = \frac{1}{N_t}\sum_{i=1}^{N} A_i Y_i - \frac{1}{N_c}\sum_{i=1}^{N}(1 - A_i)Y_i$$

- This is a good estimator when there is a constant, additive treatment effect and there are relatively few outliers in the frequecy distributions of the potential outcomes. We can always perform a transformation of the observed outcomes (by the natural logarithm, for instance) if the effect is multiplicative or if the distributions are skewed. This transformed estimator would be:

$$T_{log} = \frac{1}{N_t}\sum_{i=1}^{N} A_i \log(Y_i) - \frac{1}{N_c}\sum_{i=1}^{N}(1 - A_i)\log(Y_i)$$

Difference in median/quantiles

- To further protect against outlier, we might use the differences in quantiles as a test statistics. The most obvious is the median, which is the quantile at 0.5. Here we use $Y_t = Y_i; i : A_i = 1$ and $Y_c = Y_i; i : A_i = 0$.

$$T_{med} = med(Y_t) - med(Y_c)$$

- Of course, the median is only one quantile. We could estimate the difference in quantiles at any point in the distribution (say, the 0.25 quantile or the 0.75 quantile).

Rank statistics

- In situations with continuous outcomes, small datasets and/or many outliers, it is useful to use what are called rank statistics. In general, these statistics add together the ranks (higher rank means higher values of Y_i) of the treated units. This statistic will take its maximum when all of the treated units are ranked above all of the control units and vice versa for the minimum. The rank for a given unit is just the number of units (including that unit) that have the same or lower value of Y_i(where we use $\mathbb{I}()$ as the idenity function that takes a value of 1 when the argument is true):

$$R_i(Y_1, \ldots, Y_N) = \sum_{j=1}^{N} \mathbb{I}(Y_j \leq Y_i)$$

- Here is the definition of a particular rank statistic, the **Wilcoxon rank sum**:

$$T_{wilcoxon} = \sum_{i=1}^{N} A_i R_i$$

- Note that we have to change the definition of the ranks when there are ties in the data. Thus, these statistics are most useful for continuous variables.

- In general, there are many, many other test statistics that might be more or less appropriate for a given situation. For instance, with pair matching or stratified randomization, there are different test statistics that may be more appropriate. See Rosenbaum (2002) for more details.

Measure of extremeness

- We need to make a choice about what directions we would consider violation of the null hypothesis. For instance, do violations come with large values of the test statisic, small/negative values of the test statistic, or both? The first two are one-sided tests and the last one is a two-sided test.

- This choice affects how we calculate the p-value below.

Null/randomization disitribution

- Once we specify the null hypothesis and test statistic we can figure out what the distribution of some test statistic would be under that null.
- Under the null distribution of no effect, then it doesn't matter how the treatment was assigned. We could take any pair and flip the treatments and this wouldn't change the observed outcomes. If $Y_i(1) = Y_i(0)$ and $Y_i = A_iY_i(1) + (1 - A_i)Y_i(0)$, tne $Y_i = Y_i(1) = Y_i(0)$ no matter the value of A_i. That is, in our data we could switch all of the treatment variables and this would have changed the outcomes at all. It would change the value of the test statistic, though. If we were using the difference in means, then the test statistic under the inverse treatment assignment would be -0.8.
- Thus, under the null we have calculated 2 different test statistics. But of course we could go further and figure out the test statistic under *every possible treatment assignment vector*. In this case there is a 5 treated in an experiment of 10 units, therefore there are $\binom{10}{5}$ possible treatment assignments. Let Ω be the set of all possible treatment vectors (size K) with representative vector $\mathbf{a}$. Under different assumptions about the randomization there might be a different number of possible assignment vectors. For instance, suppose that we had used a pair-randomized design for the above data with the following pairs: (Cameroon, Congo), (Kenya, Madagascar), (Malawi, CAR), (Nigeria, Ghana), (Tanzania, Guinea-Bissau). Now there are no longer $\binom{10}{5}$ possible treatments since both Nigeria and Ghana cannot receive the treatment in any pair-randomized design.
- Once we know each of the possible values that the treatment vector could take, we can calculate the test statistic for each of these assignments. The distribution of the test statistic across all treatment assignments is called the null or randomization distribution. When calculate all of these, it becomes easy to calculate a p-value by finding the proportion of the randomization distribution that is larger than our observed test statistic:

$$\Pr(t(\mathbf{a}, \mathbf{Y}) \geq t(\mathbf{A}, \mathbf{Y}) | \tau = 0) = \frac{\sum_{\mathbf{a} \in \Omega} \mathbb{I}(t(\mathbf{a}, \mathbf{Y}) \geq t(\mathbf{A}, \mathbf{Y}))}{K}$$

- In the above, we can replace $\tau = 0$ with $\tau = \tau_0$ for whatever null hypohtesis we are interested in. When we do that, we just replace the observed outcome $t(A_i, Y_i)$ with the adjusted outcome, $t(A_i, Y_i - A_i\tau_0)$.
- The tests that are done this way are valid in the sense that if you choose some test rejection threshold α, the randomization test will falsely reject the null less than $100\alpha\%$ of the time. We don't have to rely on large samples or approximations to achieve this, though sometimes we can approximate the randomization distribution with parametric distributions such as the χ^2, Normal, and F.
- In general, this procedure might be very computationally instensive for large values of N and N_t. We can instead take K samples of the treatment assignment vectors and calculate the p-value with that sample, which should give accurate approximations. The amount of approximation error will be in the control of the researcher, by choosing the number of simulations K.

Beyond hypothesis testing

Confidence intervals

- Confidence intervals are usually justified using Normal distributions and approximations, but it turns out that you can create valid confidence intervals anytime you have a valid hypothesis test. This is because there is a duality between confidence intervals and hypohtesis tests. A $100(1-\alpha)\%$ confidence interval is equivalent to the set of null hypotheses that **would not be rejected** at the α significance level. Thus, we can construct a 95% confidence interval on the constant, additive treatment effect τ by finding all of the null hypotheses τ_0 such that $H_0 : \tau = \tau_0$ is not rejected at the 0.05 level.
- How would we do this? We would pick some grid of possible treatment effects: $-0.9, -0.8, -0.7, \ldots, 0.7, 0.8, 0.9$. Then for each of these values, use the randomization distribution to calculate a p-value for the test statistic under that null hypothesis. Then, find the lowest value with $p > 0.05$ and the highest value with $p > 0.05$ and that will formulate a 95% confidence interval.
- Note that exact confidence intervals (where the coverage is exactly 95%, say) won't be available in small samples because the p-values are discrete. Only so many p-values can be observed with an N of 10 (0.1, 0.2, 0.3, etc). Thus, for these the confidence will be conservative: the coverage will be at least $100(1-\alpha)\%$.
- Also, a two-sided confidence interval requires a two-sided hypothesis test.

Point estimates

- Up to this point we have talked about hypothesis tests and confidence intervals, but not point estimates. The best way to do this with randomization inference is to find the null hypothesis value that is the "least surprising"; that is, the one that sets the test statistic equal to its expectation under the null.
- In practice, this means you can find the value of the null hypohtesis that gives the largest p-value.
- Nicely, this point estimate inherits the properties of the test statistic on which it is based. That is, the estimator is consistent if the test is consistent. A *consistent test* is one in which the probability of rejecting false hypotheses tends to 1 as the sample size increases.

Including covariate information

- Last week we talked about how including covariates in a regression of the outcome on the treatment could make the estimates of the treatment effect more precise. In much the same way, we can adjust the outcomes in randomization inference to shrink the size of the confidence intervals.
- To do this, we first have to a vector of covariates X_i that we think are predictive of Y_i. In a normal regression, we would just include those covariates, but that doesn't quite work in this case. Instead, we will use an approach from basic linear modeling. Remember that one way to control for a set of covariates is run a regression of Y_i on X_i, then calculate the residuals of that regression, ϵ_i. Then, run a regression of ϵ_i on our treatment indicator, A_i. Covariate adjustment in randomization inference works in a similar manner.

- First, we define a function that will produce residuals, $\epsilon(Y(0), X) = e$. This could be a crazy machine learning alogrithm or a simple regression, but the important point is that it is predefined and does not involve the treatment.
- Next, calculate the potential outcomes under control for all units using some hypothesis about the treatment effect, τ_0, $Y_i(0) = Y_i - A_i\tau_0$. Use these values to calculate the residuals under that null: $e_i(0) = \epsilon(Y_i - A_i\tau_0, X_i)$.
- Last, use these residuals in place of the outcome in calculating the test statistics in the randomization distribution: $t(A_i, e_i(0))$. Then, use this test statistic and its randomization distribution to calculate p-values in the same way as above. Go nuts, calculate confidence intervals as well.
- As long as we do all of the data exploration to fit Y_i and X_i **before** we calculate the test statistics, this procedure is still valid.