

Causal Inference I

MIXTAPE SESSION



Roadmap

Randomization inference

- Lady tasting tea

- Fisher's sharp null

- Alternative test statistics

Randomization inference and causal inference

- “In randomization-based inference, uncertainty in estimates arises naturally from the random assignment of the treatments, rather than from hypothesized sampling from a large population.” (Athey and Imbens 2017)
- Athey and Imbens is part of growing trend of economists using randomization-based methods for doing causal inference
- Unclear (to me) why we are hearing more and more about randomization inference, but we are.
- Could be due to improved computational power and/or the availability of large data instead of samples?

Lady tasting tea experiment

- Ronald Aylmer Fisher (1890-1962)
 - Two classic books on statistics: *Statistical Methods for Research Workers* (1925) and *The Design of Experiments* (1935), as well as a famous work in genetics, *The Genetical Theory of Natural Science*
 - Developed many fundamental notions of modern statistics including the theory of randomized experimental design.

Lady tasting tea

- Muriel Bristol (1888-1950)
 - A PhD scientist back in the days when women weren't PhD scientists
 - Worked with Fisher at the Rothamsted Experiment Station (which she established) in 1919
 - During afternoon tea, Muriel claimed she could tell from taste whether the milk was added to the cup before or after the tea
 - Scientists were incredulous, but Fisher was inspired by her strong claim
 - He devised a way to test her claim which she passed using randomization inference

Description of the tea-tasting experiment

- Original claim: Given a cup of tea with milk, Bristol claims she can discriminate the order in which the milk and tea were added to the cup
- Experiment: To test her claim, Fisher prepares 8 cups of tea – 4 **milk then tea** and 4 **tea then milk** – and presents each cup to Bristol for a taste test
- Question: How many cups must Bristol correctly identify to convince us of her unusual ability to identify the order in which the milk was poured?
- Fisher's sharp null: Assume she can't discriminate. Then what's the likelihood that random chance was responsible for her answers?

Choosing subsets

- The lady performs the experiment by selecting 4 cups, say, the ones she claims to have had the tea poured first.

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

- “8 choose 4” – $\binom{8}{4}$ – ways to choose 4 cups out of 8
 - Numerator is $8 \times 7 \times 6 \times 5 = 1,680$ ways to choose a first cup, a second cup, a third cup, and a fourth cup, in order.
 - Denominator is $4 \times 3 \times 2 \times 1 = 24$ ways to order 4 cups.

Choosing subsets

- There are 70 ways to choose 4 cups out of 8, and therefore a 1.4% probability of producing the correct answer by chance

$$\frac{24}{1680} = 1/70 = 0.014.$$

- For example, the probability that she would correctly identify all 4 cups is $\frac{1}{70}$

Statistical significance

- Suppose the lady correctly identifies all 4 cups. Then ...
 1. Either she has no ability, and has chosen the correct 4 cups purely by chance, or
 2. She has the discriminatory ability she claims.
- Since choosing correctly is highly unlikely in the first case (one chance in 70), the second seems plausible.
- Bristol actually got all four correct
- I wonder if seeing this, any of the scientists present changed their mind

Null hypothesis

- In this example, the null hypothesis is the hypothesis that the lady has no special ability to discriminate between the cups of tea.
- We can never prove the null hypothesis, but the data may provide evidence to reject it.
- In most situations, rejecting the null hypothesis is what we hope to do.

Null hypothesis of no effect

- Randomization inference allows us to make probability calculations revealing whether the treatment assignment was “unusual”
- Fisher’s sharp null is when entertain the possibility that no unit has a treatment effect
- This allows us to make “exact” p-values which do not depend on large sample approximations
- It also means the inference is not dependent on any particular distribution (e.g., Gaussian); sometimes called nonparametric

Sidebar: bootstrapping is different

- Sometimes people confuse randomization inference with bootstrapping
- Bootstrapping randomly draws a percent of the total observations for estimation; “uncertainty over the sample”
- Randomization inference randomly reassigns the treatment; “uncertainty over treatment assignment”

(Thanks to Jason Kerwin for helping frame the two against each other)

6-step guide to randomization inference

1. Choose a sharp null hypothesis (e.g., no treatment effects)
2. Calculate a test statistic (T is a scalar based on D and Y)
3. Then pick a randomized treatment vector \tilde{D}_1
4. Calculate the test statistic associated with (\tilde{D}, Y)
5. Repeat steps 3 and 4 for all possible combinations to get $\tilde{T} = \{\tilde{T}_1, \dots, \tilde{T}_K\}$
6. Calculate exact p-value as $p = \frac{1}{K} \sum_{k=1}^K I(\tilde{T}_k \geq T)$

Pretend experiment

Table: Pretend DBT intervention for some homeless population

Name	D	Y	Y^0	Y^1
Andy	1	10	.	10
Ben	1	5	.	5
Chad	1	16	.	16
Daniel	1	3	.	3
Edith	0	5	5	.
Frank	0	7	7	.
George	0	8	8	.
Hank	0	10	10	.

For concreteness, assume a program where we pay homeless people \$15 to take dialectical behavioral therapy (DBT). Outcomes are some measure of mental health 0-20 with higher scores being improvements

Step 1: Sharp null of no effect

Fisher's Sharp Null Hypothesis

$$H_0 : \delta_i = Y_i^1 - Y_i^0 = 0 \forall i$$

- Assuming no effect means any test statistic is due to chance
- Neyman and Fisher test statistics were different – Fisher was exact, Neyman was not
- Neyman's null was no average treatment effect (ATE=0). If you have a treatment effect of 5 and I have a treatment effect of -5, our ATE is zero. This is not the sharp null even though it also implies a zero ATE

More sharp null

- Since under the Fisher sharp null $\delta_i = 0$, it means each unit's potential outcomes under both states of the world are the same
- We therefore know each unit's missing counterfactual
- The randomization we will perform will cycle through all treatment assignments under a null well treatment assignment doesn't matter because all treatment assignments are associated with a null of zero unit treatment effects
- We are looking for evidence *against* the null

Step 1: Fisher's sharp null and missing potential outcomes

Table: Missing potential outcomes are no longer missing

Name	D	Y	Y^0	Y^1
Andy	1	10	10	10
Ben	1	5	5	5
Chad	1	16	16	16
Daniel	1	3	3	3
Edith	0	5	5	5
Frank	0	7	7	7
George	0	8	8	8
Hank	0	10	10	10

Fisher sharp null allows us to **fill in** the missing counterfactuals bc

Step 2: Choosing a test statistic

Test Statistic

A test statistic $T(D, Y)$ is a scalar quantity calculated from the treatment assignments D and the observed outcomes Y

- By scalar, I just mean it's a number (vs. a function) measuring some relationship between D and Y
- Ultimately there are many tests to choose from; I'll review a few later
- If you want a test statistic with high statistical power, you need large values when the null is false, and small values when the null is true (i.e., *extreme*)

Simple difference in means

- Consider the absolute SDO from earlier

$$\delta_{SDO} = \left| \frac{1}{N_T} \sum_{i=1}^N D_i Y_i - \frac{1}{N_C} \sum_{i=1}^N (1 - D_i) Y_i \right|$$

- Larger values of δ_{SDO} are evidence *against* the sharp null
- Good estimator for constant, additive treatment effects and relatively few outliers in the potential outcomes

Step 2: Calculate test statistic, $T(D, Y)$

Table: Calculate T using D and Y

Name	D	Y	Y^0	Y^1	δ_i
Andy	1	10	10	10	0
Ben	1	5	5	5	0
Chad	1	16	16	16	0
Daniel	1	3	3	3	0
Edith	0	5	5	5	0
Frank	0	7	7	7	0
George	0	8	8	8	0
Hank	0	10	10	10	0

We'll start with this simple the simple difference in means test statistic,
 $T(D, Y): \delta_{SDO} = 34/4 - 30/4 = 1$

Steps 3-5: Null randomization distribution

- Randomization steps reassign treatment assignment for every combination, calculating test statistics each time, to obtain the entire distribution of counterfactual test statistics
- The key insight of randomization inference is that under Fisher's sharp null, the treatment assignment shouldn't matter
- Ask yourself:
 - if there is no unit level treatment effect, can you picture a distribution of counterfactual test statistics?
 - and if there is no unit level treatment effect, what must average counterfactual test statistics equal?

Step 6: Calculate “exact” p-values

- Question: how often would we get a test statistic as big or bigger as our “real” one if Fisher’s sharp null was true?
- This can be calculated “easily” (sometimes) once we have the randomization distribution from steps 3-5
 - The number of test statistics ($t(D, Y)$) bigger than the observed divided by total number of randomizations

$$Pr(T(D, Y) \geq T(\tilde{D}, Y | \delta = 0)) = \frac{\sum_{D \in \Omega} I(T(D, Y) \leq T(\tilde{D}, Y))}{K}$$

First permutation (holding N_T fixed)

Name	\tilde{D}_2	Y	Y^0	Y^1
Andy	1	10	10	10
Ben	0	5	5	5
Chad	1	16	16	16
Daniel	1	3	3	3
Edith	0	5	5	5
Frank	1	7	7	7
George	0	8	8	8
Hank	0	10	10	10

$$\tilde{T}_1 = |36/4 - 28/4| = 9 - 7 = 2$$

Second permutation (again holding N_T fixed)

Name	\tilde{D}_3	Y	Y^0	Y^1
Andy	1	10	10	10
Ben	0	5	5	5
Chad	1	16	16	16
Daniel	1	3	3	3
Edith	0	5	5	5
Frank	0	7	7	7
George	1	8	8	8
Hank	0	10	10	10

$$T_{rank} = |36/4 - 27/4| = 9 - 6.75 = 2.25$$

Sidebar: Should it be 4 treatment groups each time?

- In this experiment, I've been using the same N_T *under the assumption* that N_T had been fixed when the experiment was drawn.
- But if the original treatment assignment had been generated by something like a Bernoulli distribution (e.g., coin flips over every unit), then you should be doing a complete permutation that is also random in this way
- This means that for 8 units, sometimes you'd have 1 treated, or even 8
- Correct inference requires you know the original data generating process

Randomization distribution

[illegible]

Step 2: Other test statistics

- The simple difference in means is fine when effects are additive, and there are few outliers in the data
- But outliers create more variation in the randomization distribution
- A good test statistic is the one that best fits your data.
- Some test statistics will have weird properties in the randomization as we'll see in synthetic control.
- What are some alternative test statistics?

Transformations

- What if there was a constant multiplicative effect: $Y_i^1/Y_i^0 = C$?
- Difference in means will have low power to detect this alternative hypothesis
- So we transform the observed outcome using the natural log:

$$T_{log} = \left| \frac{1}{N_T} \sum_{i=1}^N D_i \ln(Y_i) - \frac{1}{N_C} \sum_{i=1}^N (1 - D_i) \ln(Y_i) \right|$$

- This is useful for skewed distributions of outcomes

Difference in medians/quantiles

- We can protect against outliers using other test statistics such as the difference in quantiles
- Difference in medians:

$$T_{median} = |\text{median}(Y_T) - \text{median}(Y_C)|$$

- We could also estimate the difference in quantiles at any point in the distribution (e.g., 25th or 75th quantile)

Rank test statistics

- Basic idea is rank the outcomes (higher values of Y_i are assigned higher ranks)
- Then calculate a test statistic based on the transformed ranked outcome (e.g., mean rank)
- Useful with continuous outcomes, small datasets and/or many outliers

Rank statistics formally

- Rank is the domination of others (including oneself):

$$\tilde{R} = \tilde{R}_i(Y_1, \dots, Y_N) = \sum_{j=1}^N I(Y_j \leq Y_i)$$

- Normalize the ranks to have mean 0

$$\tilde{R}_i = \tilde{R}_i(Y_1, \dots, Y_N) = \sum_{j=1}^N I(Y_j \leq Y_i) - \frac{N+1}{2}$$

- Calculate the absolute difference in average ranks:

$$T_{rank} = |\bar{R}_T - \bar{R}_C| = \left| \frac{\sum_{i:D_i=1} R_i}{N_T} - \frac{\sum_{i:D_i=0} R_i}{N_C} \right|$$

- Minor adjustment (averages) for ties

Randomization distribution

Name	D	Y	Y^0	Y^1	Rank	R_i
Andy	1	10	10	10	6.5	2
Ben	1	5	5	5	2.5	-2
Chad	1	16	16	16	8	3.5
Daniel	1	3	3	3	1	-3.5
Edith	0	5	5	5	2.5	-2
Frank	0	7	7	7	4	-0.5
George	0	8	8	8	5	0.5
Hank	0	10	10	10	6.5	2

$$T_{rank} = |0 - 0| = 0$$

Effects on outcome distributions

- Focused so far on “average” differences between groups.
- Kolmogorov-Smirnov test statistics is based on the difference in the distribution of outcomes
- Empirical cumulative distribution function (eCDF):

$$\hat{F}_C(Y) = \frac{1}{N_C} \sum_{i:D_i=0} 1(Y_i \leq Y)$$

$$\hat{F}_T(Y) = \frac{1}{N_T} \sum_{i:D_i=1} 1(Y_i \leq Y)$$

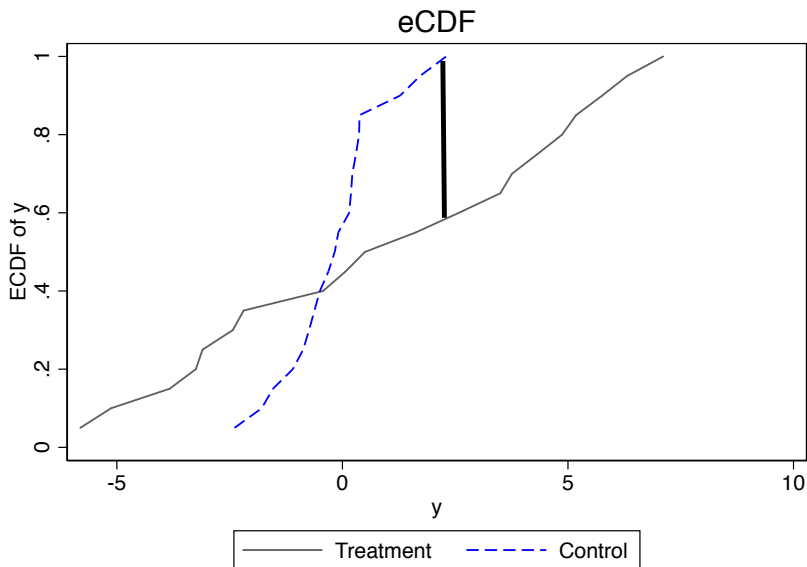
- Proportion of observed outcomes below a chosen value for treated and control separately
- If two distributions are the same, then $\hat{F}_C(Y) = \hat{F}_T(Y)$

Kolmogorov-Smirnov statistic

- Test statistics are scalars not functions
- eCDFs are functions, not scalars
- Solution: use the maximum discrepancy between the two eCDFs:

$$T_{KS} = \max |\hat{F}_T(Y_i) - \hat{F}_C(Y_i)|$$

eCDFs by treatment status and test statistic



Small vs. Modest Sample Sizes are non-trivial

Computing the exact randomization distribution is not always feasible
(Wolfram Alpha)

- $N = 6$ and $N_T = 3$ gives us 20 assignment vectors
- $N = 8$ and $N_T = 4$ gives us 70 assignment vectors
- $N = 10$ and $N_T = 5$ gives us 252 assignment vectors
- $N = 20$ and $N_T = 10$ gives us 184,756 assignment vectors
- $N = 50$ and $N_T = 25$ gives us 1.2641061×10^{14} assignment vectors

Exact p calculations are not realistic bc the number of assignments explodes at even modest size

Approximate p-values

These have been “exact” tests when they use every possible combination of D

- When you can't use every combination, then you can get *approximate* p-values from a simulation (TBD)
- With a rejection threshold of α (e.g., 0.05), randomization inference test will falsely reject less than $100 \times \alpha\%$ of the time

Approximate p values

- Use simulation to get approximate p -values
 - Take K samples from the treatment assignment space
 - Calculate the randomization distribution in the K samples
 - Tests no longer exact, but bias is under your control (increase K)
- Imbens and Rubin show that p values converge to stable p values pretty quickly (in their example after 1000 replications)

Thornton's experiment

ATE	Iteration	Rank	p	no. trials
0.45	1	1	0.01	100
0.45	1	1	0.002	500
0.45	1	1	0.001	1000

Table: Estimated p -value using different number of trials.

Including covariate information

- Let X_i be a pretreatment measure of the outcome
- One way is to use this as a gain score: $Y^{d'} = Y_i^d - X_i$
- Causal effects are the same $Y^{1i} - Y^{0i} = Y_i^1 - Y_i^0$
- But the test statistic is different:

$$T_{gain} = \left| (\bar{Y}_T - \bar{Y}_C) - (\bar{X}_T - \bar{X}_C) \right|$$

- If X_i is strongly predictive of Y_i^0 , then this could have higher power
 - Y_{gain} will have lower variance under the null
 - This makes it easier to detect smaller effects

Regression in RI

- We can extend this to use covariates in more complicated ways
- For instance, we can use an OLS regression:

$$Y_i = \alpha + \delta D_i + \beta X_i + \varepsilon$$

- Then our test statistic could be $T_{OLS} = \hat{\delta}$
- RI is justified even if the model is wrong
 - OLS is just another way to generate a test statistic
 - The more the model is “right” (read: predictive of Y_i^0), the higher the power T_{OLS} will have
- See if you can do this in Thornton’s dataset using the loops and saving the OLS coefficient (or just use `ritest`)

Concluding remarks

- Randomization inference is very common, particularly useful you don't want to make strong assumptions (parametric free)
- It's an area of continual examination by statisticians and econometricians, both in the experimental design and the quasi-experimental design
- We will use it primarily in my workshops with synthetic control, but it's going to be one you encounter and valued because of the non-parametric nature of it