LECTURE 1: CAUSALITY AND RANDOMIZED EXPERIMENTS

Guido Imbens – Stanford University

Economics 272, GSB 507, Spring 2025

OUTLINE

1. Introduction
2. Potential Outcomes and Realized Outcomes
3. Multiple Units
4. Stable Unit Treatment Value Assumption
5. The Assignment Mechanism
6. Other Perspectives on Causality
7. Fisher's Exact P-value Calculations
8. Neyman's Approach to Estimation of Average Treatment Effects

INTRODUCTION

- Three key notions underlying the general approach to causality taken in this course.
  1. Potential outcomes, each corresponding to the various levels of a treatment or manipulation.
  2. The presence of multiple units, and the related stability assumption.
  3. Central role of the assignment mechanism, which is crucial for inferring causal effects and serves as an organizing principle.

POTENTIAL OUTCOMES: THE SINGLE UNIT CASE

- There is an action/manipulation/treatment/cause $W \in \{C, T\}$
- We associate with each action a potential outcome.
  - $Y(C)$ denotes the outcome given the control treatment,
  - $Y(T)$ denotes the outcome given the active treatment.
- The causal effect of the action or treatment involves a comparison of these potential outcomes, *e.g.*,

$$\tau = Y(T) - Y(C)$$

POTENTIAL OUTCOMES AND THE REALIZED OUTCOME

- Only the potential outcome corresponding to the action actually taken at that time will be realized and observed.
- Observe treatment W and realized/observed outcome Y^{obs}:

$$Y^{\text{obs}} \equiv Y(W) = \begin{cases} Y(C) & \text{if } W = C \\ Y(T) & \text{if } W = T \end{cases}$$

- The other potential outcome is missing:

$$Y^{\text{mis}} \equiv \begin{cases} Y(C) & \text{if } W = T \\ Y(T) & \text{if } W = C \end{cases}$$

CAUSAL EFFECTS

- The causal effect is

$$\tau \equiv Y(T) - Y(C) = \begin{cases} Y^{\text{obs}} - Y^{\text{mis}} & \text{if } W = T \\ Y^{\text{mis}} - Y^{\text{obs}} & \text{if } W = C \end{cases}$$

- Holland's (1986) much quoted statement of "the fundamental problem of causal inference" refers to the fact that it is not possible to observe all components of the causal effects. (There is always a missing potential outcome.)

WHY IS IT USEFUL TO THINK IN TERMS OF POTENTIAL OUTCOMES?

- Potential outcome notion is consistent with the way economists traditionally think about demand/supply or production functions.
- Some causal questions are challenging: causal effect of immutable characteristics on economic outcomes. One solution is to make manipulation precise: change names on cv for job applications (Bertrand and Mullainathan, 2004) to change the perception of ethnicity.
- What is causal effect of weight, height or gender on earnings (Hamermesh and Biddle, 1993)? Strong statistical correlations, but what do they mean? Many manipulations possible (diet, exercise, surgery), corresponding to different causal effects.

MULTIPLE UNITS: PROBLEM

- Because we cannot learn much about causal effects from a single observed outcome, we typically rely on multiple units exposed to different treatments to make causal inferences.
- By itself, however, the presence of multiple units does not solve the problem of causal inference.
- Consider a drug (aspirin) example with two units—you and I—and two possible treatments for each unit—aspirin or no aspirin.
- There are now a total of four treatment levels:
  - you take an aspirin and I do not,
  - I take an aspirin and you do not,
  - we both take an aspirin,
  - or we both do not take an aspirin.

MULTIPLE UNITS: SOLUTION

- In many situations it may be reasonable to assume that treatments applied to one unit do not affect the outcome for another, SUTVA (Stable Unit Treatment Value Assumption, Rubin, 1978).
- Not Always Reasonable to Rule out Interference:
  - In agricultural fertilizer experiments, researchers have taken care to separate plots using "guard rows," unfertilized strips of land between fertilized areas.
  - In large scale job training programs the outcomes for one individual may well be affected by the number of people trained through increased competition (Crepon, Duflo et al, 2013)
  - In markets equilibrium effects imply spillovers: changing experience for one driver has effects on all other drivers in Uber/Lyft setting.
  - In the peer effects / social interactions literature these peer/interaction/spillover effects are the main focus.

AN EXAMPLE WITH SUTVA

Six Observations from the GAIN Labor Market Experiment in Los Angeles: Potential Outcomes, Treatments, Realized/Observed Outcomes

Individual	Potential Outcomes		Actual Treatment	Realized Outcome
	$Y_i(\text{C})$	$Y_i(\text{T})$	W_i	Y_i^{obs}
1	66	?	0	66
2	0	?	0	0
3	0	?	0	0
4	?	0	1	0
5	?	607	1	607
6	?	436	1	436

"?" indicates missing potential outcomes

The challenge, in one form or another, is to impute the missing potential outcomes.

THE ASSIGNMENT MECHANISM

- A key piece of information is how each individual came to receive the treatment level received: in our language of causation, the assignment mechanism.

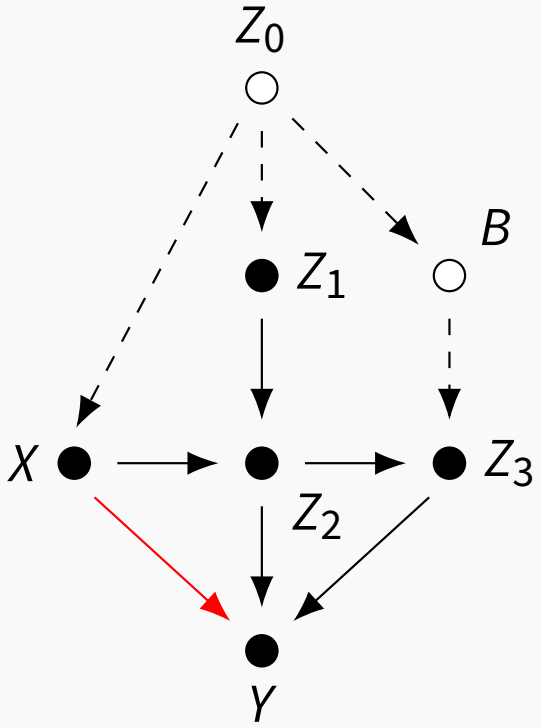
$$\Pr(\mathbf{W}|\mathbf{Y}(\text{C}), \mathbf{Y}(\text{T}), \mathbf{X})$$

  Three cases studied in this course:

  - Known, no dependence on $\mathbf{Y}(\text{C}), \mathbf{Y}(\text{T})$: Randomized experiment
  - Unknown, no dependence on $\mathbf{Y}(\text{C}), \mathbf{Y}(\text{T})$: Unconfounded assignment
  - Unknown, possibly dependent on $\mathbf{Y}(\text{C}), \mathbf{Y}(\text{T})$: most challenging case

- **Aside:** Compare with conventional focus on distribution of observed outcomes given explanatory variables. *E.g.*, $Y^{\text{obs}}|W_i \sim \mathcal{N}(\alpha + \beta W_i, \sigma^2)$. Here, other way around.

AN ALTERNATIVE TO POTENTIAL OUTCOME APPROACH: DIRECTED ACYCLICAL GRAPHS AND STRUCTURAL EQUATION MODELS



From Pearl (1995)

- X is fumigation, Y is yield, B is bird population, Z_t eelworm population: last season ($t = 0$), at beginning of season before fumigation ($t = 1$), after fumigation ($t = 2$) and end of season ($t = 3$), B bird population.
- How plausible is the DAG here? Can we identify the effect of X on Y?

A COMPLETELY RANDOMIZED EXPERIMENT

- Finite Population with N units,
- M are selected at random to receive the treatment, the remaining $N - M$ receive the control treatment,
- The assignment probability is

$$\text{pr}(\mathbf{W} = \mathbf{w}) = \binom{N}{M}^{-1}, \quad \text{for } \mathbf{w} \text{ s.t.} \sum_{i=1}^{N} w_i = M, \text{ and 0 elsewhere.}$$

- There are $\binom{N}{M}$ possible values for $\mathbf{w}$, all equally likely.
- Special case $M = N/2$, equal size treatment and control group.
- Slightly different: (and slightly less attractive!) Bernoulli trials with

$$\text{pr}(\mathbf{W} = \mathbf{w}) = 2^{-N}, \quad \forall \mathbf{w}.$$

A COMPLETELY RANDOMIZED EXPERIMENT

- Given data from a completely randomized experiment, Fisher (1935) was interested testing sharp null hypotheses
  - that is, null hypotheses under which all values of the potential outcomes for the units in the experiment are either observed or can be inferred.
- Notice that this is distinct from the null hypothesis that the average treatment effect across units is zero.
- The null of a zero average is a weaker hypothesis because the average effect of the treatment may be zero even if for some units the treatment has a positive effect, as long as for others the effect is negative.
- Question: when is the zero average versus the sharp null of interest?

BASICS OF FISHER P-VALUE CALCULATIONS

- If the null hypothesis is sharp we can determine the distribution of any test statistic T (a function of the stochastic assignment vector, $\mathbf{W}$, the observed outcomes).
- The test statistic is stochastic solely through the stochastic nature of the assignment vector, leading to the randomization distribution of the test statistic.
- Using this distribution, we can compare the observed test statistic, T^{obs}, against its distribution under the null hypothesis.
- The Fisher exact test approach entails two choices:
  - the choice of the sharp null hypothesis,
  - choice of test statistic.

THE SHARP NULL OF NO EFFECT

- We test the sharp null hypothesis that the program had absolutely no effect on earnings, that is:

$$H_0: \quad Y_i(\mathrm{C}) = Y_i(\mathrm{T}) \qquad \text{for all } i = 1, \ldots, 6.$$

- Under this null hypothesis, the unobserved potential outcomes are equal to the observed outcomes for each unit.
- Thus we can fill in all six of the missing entries using the observed data.
- This is the first key point of the Fisher approach:
  - under the sharp null hypothesis all the missing values can be inferred from the observed ones.

BACK TO EXAMPLE

SIX OBSERVATIONS FROM THE GAIN EXPERIMENT IN LOS ANGELES

Individual	Potential Outcomes		Actual	Observed Outcome
	Y_i(C)	Y_i(T)	Treatment	Y_i
1	66	(66)	0	66
2	0	(0)	0	0
3	0	(0)	0	0
4	(0)	0	1	0
5	(607)	607	1	607
6	(436)	436	1	436

"(·)" indicates values of missing potential outcomes imputed under the null hypothesis, in this case null hypothesis of no effect of treatment whatsoever.

LOGISTICS

- Now consider testing this null against the alternative hypothesis

$$H_a: \; Y_i(\mathrm{C}) \neq Y_i(\mathrm{T}), \text{ for some } i.$$

- We choose a test statistic:

$$T = T(\mathbf{W}, \mathbf{Y}^{\text{obs}}),$$

for example,

$$T_1 = \sum_{i=1}^{6} W_i Y_i^{\text{obs}} \qquad \text{or} \qquad T_2 = \frac{1}{3}\sum_{i=1}^{6} W_i Y_i^{\text{obs}} - \frac{1}{3}\sum_{i=1}^{6}(1 - W_i) Y_i^{\text{obs}}$$

EXAMPLE

- For the observed data the value of the test statistic is

$$T(\mathbf{W}^{\text{obs}}, \mathbf{Y}^{\text{obs}}) = \frac{Y_4^{\text{obs}} + Y_5^{\text{obs}} + Y_6^{\text{obs}} - Y_1^{\text{obs}} - Y_2^{\text{obs}} - Y_3^{\text{obs}}}{3} = 325.6.$$

- Suppose for example, that instead of the observed assignment vector $\mathbf{W}^{\text{obs}} = (0, 0, 0, 1, 1, 1)'$ the assignment vector had been $\tilde{\mathbf{W}} = (0, 1, 1, 0, 1, 0)$.
- Under this assignment vector the test statistic would have been different:

$$T(\tilde{\mathbf{W}}, \mathbf{Y}^{\text{obs}}) = \frac{-Y_1^{\text{obs}} + Y_2^{\text{obs}} + Y_3^{\text{obs}} - Y_4^{\text{obs}} + Y_5^{\text{obs}} - Y_6^{\text{obs}}}{3} = 35$$

- We can calculate that because we know all the missing potential outcomes under the null hypothesis

Randomization Distribution for six observations from GAIN data

W_1	W_2	W_3	W_4	W_5	W_6	levels	ranks
0	**0**	**0**	**1**	**1**	**1**	**325.6**	**1.00**
0	0	1	0	1	1	325.6	1.67
0	0	1	1	0	1	-79.0	-1.67
0	0	1	1	1	0	35.0	-1.00
0	1	0	0	1	1	325.6	2.33
0	1	0	1	0	1	-79.0	-1.00
0	1	0	1	1	0	35.0	-0.33
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
1	1	1	0	0	0	325.6	-1.00

P-VALUE

- Given the distribution of the test statistic, how unusual is this observed average difference ($T^{\text{obs}} = 325.6$), assuming the null hypothesis is true?
- One way to formalize this question is to ask how likely it is (under the randomization distribution) to observe a value of the test statistic that is as large in absolute value as the one actually observed.
- Simply counting we see that there are twelve vectors of assignments with at least a difference in absolute value of 325.6 between treated and control classes, out of a set of twenty possible assignment vectors.
- This implies a p-value of 8/20 = 0.40.

THE CHOICE OF NULL HYPOTHESIS

- The first question when considering a Fisher Exact P-value calculation is the choice of null hypothesis.
- Typically the most interesting sharp null hypothesis is that of no effect of the treatment: $Y_i(\text{C}) = Y_i(\text{T})$ for all units.
- Although Fisher's approach cannot accommodate a null hypothesis of an average treatment effect of zero, it can accommodate sharp null hypotheses other than the null hypothesis of no effect whatsoever, e.g.,

$$H_0: \quad Y_i(\text{T}) = Y_i(\text{C}) + c_i, \quad \text{for all } i = 1, \ldots, N,$$

for known c_i.

THE CHOICE OF STATISTIC

- The second decision, the choice of test statistic, is typically more difficult than the choice of the null hypothesis. First let us formally define a statistic:

  DEFINITION A statistic T is a known function $T(\mathbf{W}, \mathbf{Y}^{\text{obs}}, \mathbf{X})$ of assignments, $\mathbf{W}$, observed outcomes, $\mathbf{Y}^{\text{obs}}$, and pretreatment variables, $\mathbf{X}$.

- Any statistic that satisfies this definition is valid for use in Fisher's approach and we can derive its distribution under the null hypothesis.
- How do we choose a statistic?
- We look for a statistic whose distribution is different if an interesting version of the alternative hypothesis holds.

SOME COMMON CHOICES OF STATISTICS

- The most standard choice of statistic is the difference in average outcomes by treatment status:

$$T = \frac{\sum W_i Y_i^{\text{obs}}}{\sum W_i} - \frac{\sum (1 - W_i) Y_i^{\text{obs}}}{\sum (1 - W_i)}.$$

- An obvious alternative to the simple difference in average outcomes by treatment status is to transform the outcomes before comparing average differences between treatment levels, *e.g.*, by taking logarithms, leading to the following test statistic:

$$T = \frac{\sum W_i \ln(Y_i^{\text{obs}})}{\sum W_i} - \frac{\sum (1 - W_i) \ln(Y_i^{\text{obs}})}{\sum 1 - W_i}.$$

AN IMPORTANT CHOICE OF STATISTIC: THE RANK STATISTIC

- An important class of statistics involves transforming the outcomes to ranks before considering differences by treatment status.
- This improves robustness relative to difference in means or difference in ranks, without having to make an arbitrary choice such as taking logs.
- Not good with data with masspoint at zero.
- We also often subtract $(N+1)/2$ from each to obtain a normalized rank that has average zero in the population:

$$R_i(Y_1^{\text{obs}},\ldots,Y_N^{\text{obs}}) = \sum_{j=1}^{N} \mathbf{1}_{Y_j^{\text{obs}}<Y_i^{\text{obs}}} + \frac{1}{2}\left(1+\sum_{j=1}^{N}\mathbf{1}_{Y_j^{\text{obs}}=Y_i^{\text{obs}}}\right) - \frac{N+1}{2}.$$

- Given the ranks R_i, an attractive test statistic is the difference in average ranks for treated and control units:

$$T = \frac{\sum W_i R_i}{\sum W_i} - \frac{\sum (1-W_i)R_i}{\sum 1-W_i}.$$

COMPUTATION OF P-VALUES

- The p-value calculations presented so far have been exact. With both N and M sufficiently large, it may therefore be unwieldy to calculate the test statistic for every value of the assignment vector.
- In that case we rely on numerical approximations to the p-value.
- Formally, randomly draw an N-dimensional vector with $N - M$ zeros and M ones from the set of assignment vectors. Calculate the statistic for this draw (denoted T_1).
- Repeat this process $K - 1$ times, in each instance drawing another vector of assignments and calculating the statistic T_k, for $k = 2, \ldots, K$.
- We then approximate the p-value for our test statistic by the fraction of these K statistics that are more extreme than T^{obs} for the actual data.

NORMAL APPROXIMATION

- Comparison to p-value based on normal approximation to distribution of t-statistic:

$$t = \frac{\overline{Y}_1 - \overline{Y}_0}{\sqrt{s_0^2/(N-M) + s_1^2/M}}$$

where

$$s_0^2 = \frac{1}{N-M-1} \sum_{i:W_i=0} (Y_i^{\text{obs}} - \overline{Y}_0)^2, \qquad s_1^2 = \frac{1}{M-1} \sum_{i:W_i=1} (Y_i^{\text{obs}} - \overline{Y}_1)^2$$

and

$$p \approx 2 \times \Phi(-|t|) \qquad \text{where } \Phi(a) = \int_{-\infty}^{a} \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx$$

P-VALUES FOR FISHER EXACT TESTS: RANKS VERSUS LEVELS

		sample size		p-values		
Prog	Loc	controls	treated	t-test	FEP (levels)	FEP (ranks)
GAIN	AL	601	597	0.835	0.836	0.890
GAIN	LA	1400	2995	0.544	0.531	0.561
GAIN	RI	1040	4405	0.000	0.000	0.000
GAIN	SD	1154	6978	0.057	0.068	0.018
WIN	AR	37	34	0.750	0.753	0.805
WIN	BA	260	222	0.339	0.339	0.286
WIN	SD	257	264	0.136	0.137	0.024
WIN	VI	154	331	0.960	0.957	0.249

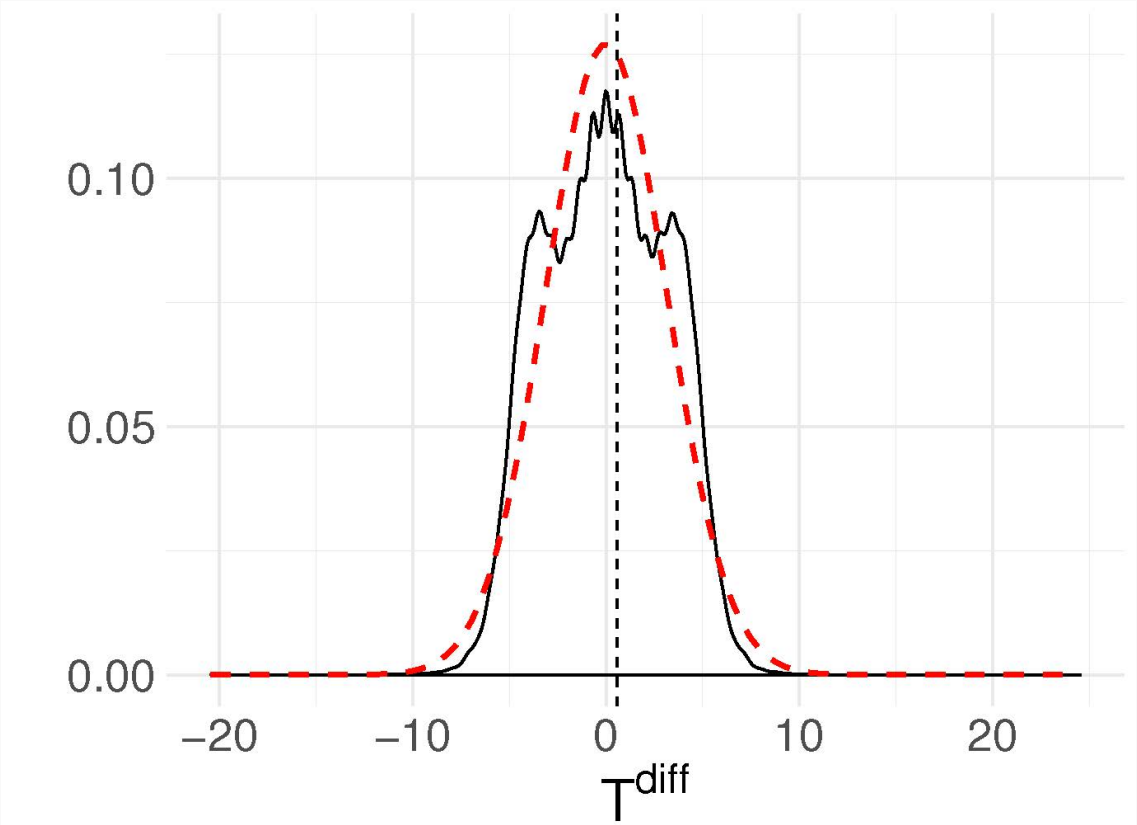
GAIN RIVERSIDE DATA

- Outcome: earnings in first quarter after randomization.
  - Skewness: 7.89
  - Kurtosis: 83.21
  - Fraction zeros: 0.7932
  - Fraction treated: 0.8090
- Statistic:

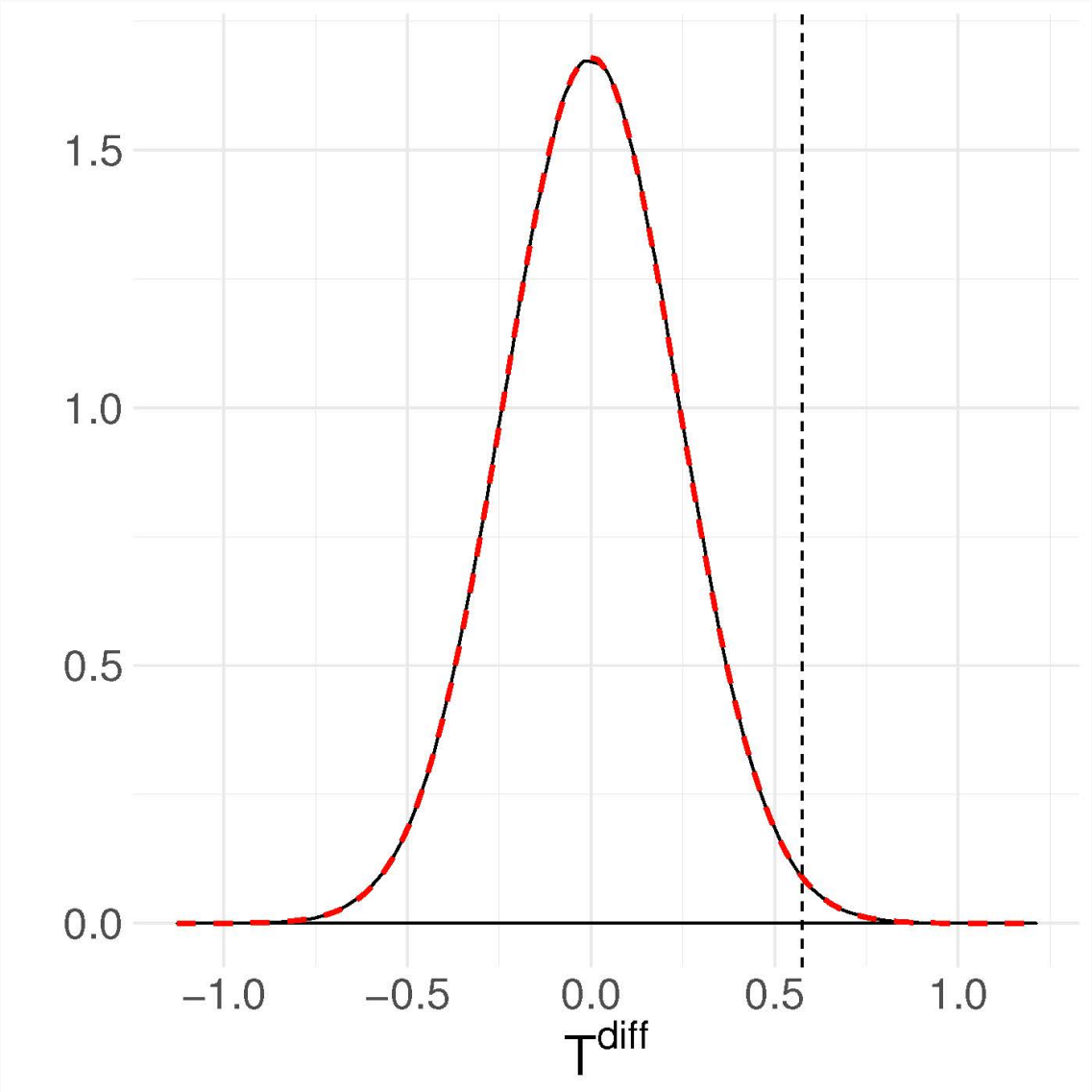
$$T = \overline{Y}_{\mathrm{T}} - \overline{Y}_{\mathrm{C}}$$

- p-value: 0.0014 (based on actual sample, 5,445 individuals)
- Simulations: Sample size: $N \in \{25, 100, 500, 5445\}$

EXACT P-VALUES: 25 UNITS



EXACT P-VALUES: 500 UNITS



EXACT P-VALUES: TAKE-AWAYS

- Randomization-based p-values underly tests for treatment effects.
- In practice using approximations to p-values based on normal approximation to the distribution of the t-statistic is often similar to exact p-values based on difference in averages test.
- With very skewed distributions rank-based tests are much better
  - but need to be careful with presence of masspoints.

NEYMAN'S REPEATED SAMPLING APPROACH

- During the same period in which Fisher was developing his p-value calculations for sharp null hypotheses, Jerzey Neyman was focusing on methods for estimating average treatment effects.
- His approach was to consider an estimator and derive its distribution (bias and variance) under repeated sampling by drawing from the randomization distribution of $\mathbf{W}$, the assignment vector.
- $\mathbf{Y}(C)$, $\mathbf{Y}(T)$ still fixed in repeated sampling thought experiment.

FISHER-NEYMAN EXCHANGE

(N) “So long as the *average* (italics in original) yields of any treatments are identical, the question as to whether these treatments affect *separate* yields on *single* plots seems to be uninteresting and academic ... ”

(F) “... It may be foolish, but that is what the *z* [FEP] test was designed for, and the only purpose for which it has been used. ...”

(N) “... I believe Professor Fisher himself described the problem of agricultural experimentation formerly not in the same manner as he does now. ...”

(F) “... Dr. Neyman thinks another test would be more important. I am not going to argue that point. It may be that the question which Dr. Neyman thinks should be answered is more important than the one I have proposed and attempted to answer. I suggest that before criticizing previous work it is always wise to give enough study to the subject to understand its purpose. Failing that it is surely quite unusual to claim to understand the purpose of previous work better than its author.”

FISHER-NEYMAN EXCHANGE

- I'm with Neyman all the way
  - Telling a decision make that the intervention has some effect without being able to tell them whether on average it goes up or down is not very useful.
  - Some exceptions: for some treatments knowing they do something can be interesting.

THE RANDOMIZATION DISTRIBUTION

Six Observations from the GAIN Experiment in Los Angeles

The Randomization Distribution

i	Potential Outc		Sample 1		Sample 2		Sample 3	
	$Y_i(C)$	$Y_i(T)$	W_i	Y_i^{obs}	W_i	Y_i^{obs}	W_i	Y_i^{obs}
1	66	28	0	66	1	28	1	28
2	0	0	0	0	0	0	1	0
3	0	101	0	0	1	101	0	0
4	592	0	1	0	0	592	0	592
5	350	607	1	607	0	350	1	607
6	0	436	1	436	1	436	0	0

Note: $(Y_i(C), Y_i(T))$ fixed for $i = 1, \ldots, 6$. $(W_1, \ldots, W_6)$ is stochastic, and so is Y_i^{obs}.

UNBIASED ESTIMATION OF THE AVERAGE TREATMENT EFFECT

- Neyman was interested in the population average treatment effect:

$$\tau = \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i(\mathrm{T}) - Y_i(\mathrm{C})\Big) = \overline{Y}(\mathrm{T}) - \overline{Y}(\mathrm{C}) \qquad \overline{Y}(\mathrm{T}) \equiv \frac{1}{N}\sum_{i=1}^{N} Y_i(T)$$

- Suppose that we observed data from a completely randomized experiment in which M units were assigned to treatment and $N - M$ assigned to control.
- Given randomization, the intuitive estimator for the average treatment effect is the difference in the average outcomes for those assigned to the treatment versus those assigned to the control:

$$\hat{\tau} = \frac{1}{M}\sum_{i:W_i=1} Y_i^{\text{obs}} - \frac{1}{N-M}\sum_{i:W_i=0} Y_i^{\text{obs}} = \overline{Y}_{\mathrm{T}}^{\text{obs}} - \overline{Y}_{\mathrm{C}}^{\text{obs}} \qquad \overline{Y}_{\mathrm{T}}^{\text{obs}} \equiv \frac{1}{M}\sum_{i:W_i=T} Y_i^{\text{obs}}$$

UNBIASEDNESS

- Let us show that this estimator,

$$\hat{\tau} = \overline{Y}_{\mathrm{T}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{C}}^{\mathrm{obs}}$$

  is an unbiased estimator of τ.

- First consider the statistic

$$T_i \equiv \left(\frac{W_i Y_i^{\mathrm{obs}}}{M/N} - \frac{(1 - W_i) Y_i^{\mathrm{obs}}}{(N - M)/N} \right).$$

- The average of this statistic over the population is equal to our estimator,

$$\hat{\tau} \equiv \overline{Y}_{\mathrm{T}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{C}}^{\mathrm{obs}} = \frac{1}{N} \sum_i T_i$$

UNBIASEDNESS

- Using the fact that Y_i^{obs} is equal to $Y_i(\text{T})$ if $W_i = 1$ and $Y_i(\text{C})$ if $W_i = 0$, we can rewrite this statistic as:

$$T_i = \left(\frac{W_i Y_i^{\text{obs}}}{M/N} - \frac{(1 - W_i) Y_i^{\text{obs}}}{(N - M)/N} \right) = \left(\frac{W_i Y_i(\text{T})}{M/N} - \frac{(1 - W_i) Y_i(\text{C})}{(N - M)/N} \right).$$

- The only element in this statistic that is random is the treatment assignment, W_i, with $\mathbb{E}[W_i] = M/N$.
- Using these results we can show that the expectation of T_i is equal to the unit-level causal effect, $Y_i(\text{T}) - Y_i(\text{C})$:

$$\mathbb{E}[T_i] = \left(\frac{\mathbb{E}[W_i] Y_i(\text{T})}{M/N} - \frac{(1 - \mathbb{E}[W_i]) Y_i(\text{C})}{(N - M)/N} \right) = Y_i(\text{T}) - Y_i(\text{C})$$

THE VARIANCE OF THE UNBIASED ESTIMATOR $\overline{Y}_{\mathrm{T}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{C}}^{\mathrm{obs}}$

- Neyman was also interested in the variance of this unbiased estimator of the average treatment effect
- This involved two steps:
  - deriving the variance of the estimator for the average treatment effect;
  - estimating this variance.
- In addition, Neyman sought to create confidence intervals for the population average treatment effect which also requires an appeal to the central limit theorem for large sample normality.

VARIANCE CALCULATION

- Consider a completely randomized experiment of N units, M assigned to treatment.
- To calculate the variance of $\overline{Y}_{\mathrm{T}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{C}}^{\mathrm{obs}}$, we need the second and cross moments of the random variable W_i, $\mathbb{E}[W_i^2]$ and $\mathbb{E}[W_i \cdot W_j]$.
- Second moment:

$$\mathbb{E}[W_i^2] = \mathbb{E}[W_i] = M/N.$$

- Cross-moment for $i \neq j$:

$$\mathbb{E}[W_i \cdot W_j] = \Pr(W_i = 1) \cdot \Pr(W_j = 1 | W_i = 1)$$

$$= (M/N) \cdot (M-1)/(N-1) \neq \mathbb{E}[W_i] \cdot \mathbb{E}[W_j],$$

(conditional on $W_i = 1$, $M - 1$ treated units remaining out of $N - 1$ total remaining.)

VARIANCE CALCULATION

- The variance of $\overline{Y}_{\mathsf{T}}^{\text{obs}} - \overline{Y}_{\mathsf{C}}^{\text{obs}}$ is equal to:

$$\text{Var}(\overline{Y}_{\mathsf{T}}^{\text{obs}} - \overline{Y}_{\mathsf{C}}^{\text{obs}}) = \frac{S_{\mathsf{C}}^2}{N-M} + \frac{S_{\mathsf{T}}^2}{M} - \frac{S_{\mathsf{CT}}^2}{N},$$

- Here, for $w = \mathsf{C}, \mathsf{T}$, S_w^2 is the variance of $Y_i(w)$ in the population, defined as:

$$S_w^2 = \frac{1}{N-1}\sum_{i=1}^{N}\Big(Y_i(w) - \overline{Y}(w)\Big)^2,$$

- S_{CT}^2 is the population variance of the unit-level treatment effect, defined as:

$$S_{\mathsf{CT}}^2 = \frac{1}{N-1}\sum_{i=1}^{N}\Big(Y_i(\mathsf{T}) - Y_i(\mathsf{C}) - \tau\Big)^2.$$

VARIANCE CALCULATION

- The numerator of the first term, the population variance of the potential control outcome vector, $\mathbf{Y}(\mathrm{C})$, is equal to

$$S_{\mathrm{C}}^2 = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i(\mathrm{C}) - \overline{Y}(\mathrm{C}))^2.$$

  An unbiased estimator for S_{C}^2 is

$$s_{\mathrm{C}}^2 = \frac{1}{N-M-1}\sum_{i:W_i=0}(Y_i^{\mathrm{obs}} - \overline{Y}_{\mathrm{C}}^{\mathrm{obs}})^2.$$

- Similar for $\mathrm{S}_{\mathrm{T}}^2$.

VARIANCE CALCULATION

- The third term,

$$S^2_{\text{CT}} = \frac{1}{N-1}\sum_{i=1}^{N}\Big(Y_i(\text{T}) - Y_i(\text{C}) - \tau\Big)^2$$

  is more difficult to estimate because we cannot observe both $Y_i(\text{T})$ and $Y_i(\text{C})$ for any unit.

- As noted previously, if the treatment effect is additive ($Y_i(\text{T}) - Y_i(\text{C}) = c, \forall i$), then this variance is equal to zero and the third term vanishes.

- Under this circumstance we can obtain an unbiased estimator for the variance as:

$$\hat{\mathbb{V}}\left(\overline{Y}^{\text{obs}}_{\text{T}} - \overline{Y}^{\text{obs}}_{\text{C}}\right) = \frac{s^2_{\text{C}}}{N-M} + \frac{s^2_{\text{T}}}{M}.$$

JUSTIFICATION FOR VARIANCE ESTIMATOR

- This estimator for the variance is widely used, even when the assumption of an additive treatment effect is inappropriate. There are two main reasons for this estimator's popularity.
  - Confidence intervals generated using this estimator of the variance will be conservative with actual coverage at least as large, but not necessarily equal to, the nominal coverage.
  - This estimator is unbiased for the variance of $\hat{\tau} = \overline{Y}_{\mathrm{T}}^{\mathrm{obs}} - \overline{Y}_{\mathrm{C}}^{\mathrm{obs}}$ when this statistic is interpreted as the estimator of the average treatment effect in the super-population from which the N observed units are a random sample. (we return to this interpretation later)

CONFIDENCE INTERVALS

- Given the estimator $\hat{\tau}$ and the variance estimator $\hat{\mathbb{V}}$, how do we think about confidence intervals?
- Let's consider the case where $M = N/2$, so $\mathbb{E}[W_i] = 1/2$, and define

$$D_i = 2W_i - 1, \qquad \text{so that } \mathbb{E}[D_i] = 0, \;\; D_i^2 = 1.$$

- Write

$$\hat{\tau} = \overline{Y}_{\mathsf{T}} - \overline{Y}_{\mathsf{C}} = \frac{1}{N/2}\sum_{i=1}^{N} W_i Y_i(\mathsf{T}) - \frac{1}{N/2}\sum_{i=1}^{N}(1-W_i)Y_i(\mathsf{C})$$

$$= \frac{1}{N}\sum_{i=1}^{N}\Big(Y_i(\mathsf{T}) - Y_i(\mathsf{C})\Big) + \frac{1}{N}\sum_{i=1}^{N} D_i\Big(Y_i(\mathsf{T}) + Y_i(\mathsf{C})\Big) = \tau + \frac{1}{N}\sum_{i=1}^{N} D_i\Big(Y_i(\mathsf{T}) + Y_i(\mathsf{C})\Big)$$

CONFIDENCE INTERVALS

- The stochastic part, normalized by the sample size, is

$$\sqrt{N}(\hat{\tau} - \tau) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} D_i \left(Y_i(1) + Y_i(0) \right)$$

- It has mean zero and variance

$$\mathbb{V} = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i(1) + Y_i(0) \right)^2.$$

- Under conditions on the sequence of $\sigma_i^2 = (Y_i(1) + Y_i(0))^2$, we can use a Lyapounov clt for independent but not identically distributed random variables.
- This leads to

$$\frac{\frac{1}{\sqrt{N}} \sum_{i=1}^{N} D_i \left(Y_i(1) + Y_i(0) \right)}{\sqrt{\frac{1}{N} \sum_{i=1}^{N} \sigma_i^2}} \xrightarrow{d} \mathcal{N}(0, 1)$$

STOCHASTIC ESTIMANDS

- So far we were interested in $\tau = \sum_i (Y_i(\mathrm{T}) - Y_i(\mathrm{C}))/N$.
- Suppose we are interested in the average effect for the treated:

$$\tau_t = \sum_i \mathbf{1}_{W_i=\mathrm{T}} (Y_i(\mathrm{T}) - Y_i(\mathrm{C}))/N_\mathrm{T}$$

- What changes:
  - The estimand is now stochastic: it varies with the assignment.
  - The standard estimator continues to be unbiased: $\mathbb{E}[\overline{Y}_\mathrm{T} - \overline{Y}_\mathrm{C} - \tau_t] = 0$
  - Its variance is smaller:

$$\mathbb{V}(\hat{\tau} - \tau) = \mathbb{E}[(\hat{\tau} - \tau)^2] = \mathbb{E}[(\hat{\tau} - \tau_t + \tau_t - \tau)^2]$$

$$= \mathbb{E}[(\hat{\tau} - \tau_t)^2] + \mathbb{E}[(\tau_t - \tau)^2] + 2\mathbb{E}[(\hat{\tau} - \tau_t)(\tau_t - \tau)] = \mathbb{E}[(\hat{\tau} - \tau_t)^2] + \mathbb{E}[(\tau_t - \tau)^2]$$

$$\text{so} \qquad \mathbb{E}[(\hat{\tau} - \tau_t)^2] \leq \mathbb{E}[(\hat{\tau} - \tau)^2]$$

REFERENCES

- Bertrand, Marianne, and Sendhil Mullainathan. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination." *American Economic Review* 94, no. 4 (2004): 991-1013.
- Carrell, Scott E., Bruce I. Sacerdote, and James E. West. "From natural variation to optimal policy? The importance of endogenous peer group formation." *Econometrica* 81, no. 3 (2013): 855-882.
- Crépon, Bruno, Esther Duflo, Marc Gurgand, Roland Rathelot, and Philippe Zamora. "Do labor market policies have displacement effects? Evidence from a clustered randomized experiment." *The Quarterly Journal of Economics* 128, no. 2 (2013): 531-580.
- Dehejia, Rajeev H., and Sadek Wahba. "Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs." *Journal of the American statistical Association* 94, no. 448 (1999): 1053-1062.

REFERENCES (CTD)

- Fisher, Ronald Aylmer. "The design of experiments." *The design of experiments.* 7th Ed (1960).
- Freedman, David A. "On regression adjustments in experiments with several treatments." *Annals of Applied Statistics* 2 (2008): 176-96.
- Hamermesh, Daniel S., and Jeff E. Biddle. "Beauty and the labor market." *American Economic Review* 84, no. 5 (1994).
- Holland, Paul W. "Statistics and causal inference." *Journal of the American statistical Association* (1986): 945-960.
- Imbens, Guido W., and Donald B. Rubin. *Causal inference in statistics, social, and biomedical sciences.* Cambridge University Press, 2015.
- LaLonde, Robert J. "Evaluating the econometric evaluations of training programs with experimental data." *The American economic review* (1986): 604-620.

REFERENCES (CTD)

- Li, Xinran, and Peng Ding. "General forms of finite population central limit theorems with applications to causal inference." *Journal of the American Statistical Association* 112, no. 520 (2017): 1759-1769.
- Pearl, Judea. *Causality*. Cambridge university press, 2009.
- Rubin, Donald B. "Assignment to treatment group on the basis of a covariate." *Journal of educational Statistics* 2, no. 1 (1977): 1-26.
- Rubin, Donald B. "Bayesian inference for causal effects: The role of randomization." *The Annals of statistics* (1978): 34-58.
- Splawa-Neyman, Jerzy, Dorota M. Dabrowska, and Terrence P. Speed. "On the application of probability theory to agricultural experiments. Essay on principles. Section 9." *Statistical Science* (1990): 465-472.