Food For Thought: Inference in Randomized Experiments

Erich Battistin
University of Maryland, CEPR, FBK-IRVAPP and IZA

AREC 829
Policy Design and Causal Inference for Social Science

Taxonomy of Randomized Experiments

Randomized Experiments

- The **assignment mechanism** describes how each of N units in the population is assigned to treatment, and can be characterized by:

$$P\left(\mathbf{D}|\mathbf{Y}_0, \mathbf{Y}_1, \mathbf{X}\right),$$

  where $\mathbf{Y}_0$, $\mathbf{Y}_1$, $\mathbf{D}$ and $\mathbf{X}$ are the N-component vectors of potential outcomes, treatment status and covariates with i-th element equal to Y_{0i}, Y_{1i}, D_i and X_i, respectively.

- In randomized experiments this probability is **known** and **independent of potential outcomes**.

- In a **full randomization** of $N_1 \equiv \sum_{i=1}^{N} D_i$ units we have that:

$$P\left(\mathbf{D}|\mathbf{Y}_0, \mathbf{Y}_1, \mathbf{X}\right) = \binom{N}{N_1}^{-1},$$

  where $N_0 \equiv \sum_{i=1}^{N}(1 - D_i)$ and $N_0 = N - N_1$.

- Think of a **coin** or a **lottery** for which the unit-level probability is $\frac{N_1}{N}$ and independent of covariates.

Randomized Experiments

- **Stratification** on blocks B_j defined from covariates ensures variability in treatment assignment within each stratum, for $j = 1, \ldots, J$.
- In **paired randomized experiments**, the units included in the randomization (e.g., villages) are matched into pairs, and then one unit of each pair is randomly assigned to treatment.
- In a **stratified randomization** of $N_1(j) \equiv \sum_{i \in B_j} D_i$ units we have:

$$P(\mathbf{D}|\mathbf{Y}_0, \mathbf{Y}_1, \mathbf{X}) = \prod_{j=1}^{J} \binom{N(j)}{N_1(j)}^{-1},$$

  where $N_0(j) \equiv \sum_{i \in B_j} (1 - D_i)$ and $N(j) \equiv N_0(j) + N_1(j)$.
- Think of different coins or lotteries across strata, where the unit level probability is $e(X_i) \equiv \frac{N_1(j)}{N(j)}$.
- Can be used to treat a fixed number of units of each type.

Why Experiments?

- Experiments can't solve for the problem of missing counterfactual outcomes, but ensure that **missing values occur by chance**.
- Randomization is welcome because it yields **exchangeability**, $(Y_0, Y_1) \perp D$, or conditional versions of it, $(Y_0, Y_1) \perp D | X \in B_j$. **Mean exchangeability** is sufficient for some causal parameters.
- The groups $D = 1$ and $D = 0$ are exchangeable if their outcomes would be the same whenever they were subjected to the identical exposure history.
- The identification problem is solved, and the focus must be on **estimation** and **precision**.

Rethinking the Thought Experiment

Sampling-Based vs. Design-Based Uncertainty

- Let $R_i \in \{0, 1\}$ denote if unit i was sampled and $D_i \in \{0, 1\}$ if unit i was treated.
- Let n be the total sample size from a population of size N. Size by treatment status is defined similarly (e.g., n_1 and N_1).
- Consider the difference-in-sample-means estimator:

$$\hat{\tau} \equiv \underbrace{\frac{1}{n_1}\sum_{i=1}^{n_1} Y_i}_{\text{treatment}} - \underbrace{\frac{1}{n_0}\sum_{i=1}^{n_0} Y_i}_{\text{control}} = \underbrace{\frac{1}{N}\sum_{i=1}^{N} \frac{NR_iD_i}{n_1} Y_i - \frac{1}{N}\sum_{i=1}^{N} \frac{NR_i(1-D_i)}{n_0} Y_i}_{\text{re-weighting}}.$$

- What is the origin of uncertainty in estimates?
  - Sampling variation from selection of units in the sample.
  - Variation induced by the particular treatment allocation.
  - The population may be finite.
- Random sampling assumptions are not natural when considering states or counties as units of observation.

Sampling-Based vs. Design-Based Uncertainty

- Imagine you're tasked with evaluating a specific policy through a Randomized Controlled Trial (RCT).
- You have a population size of N, but due to budget constraints, you can only work with a sample of size n. After drawing this sample at random, you administer a baseline survey to all n statistical units.
- Then, you randomly assign n_1 of these units to the intervention group. After the intervention concludes, you conduct a follow-up survey. Since there is:

$$\begin{aligned} P\left(R_i D_i = 1\right) &= P\left(D_i = 1 | R_i = 1\right) P\left(R_i = 1\right) = \frac{n_1}{n}\frac{n}{N} = \frac{n_1}{N}, \\ P\left(R_i(1 - D_i) = 1\right) &= P\left(D_i = 0 | R_i = 1\right) P\left(R_i = 1\right) = \frac{n_0}{n}\frac{n}{N} = \frac{n_0}{N} \end{aligned}$$

the comparison of outcome averages from the follow-up survey is numerically equal to:

$$\hat{\tau} = \frac{1}{N}\sum_{i=1}^{N}\left[\frac{R_i D_i}{P\left(R_i D_i = 1\right)} - \frac{R_i(1 - D_i)}{P\left(R_i(1 - D_i) = 1\right)}\right] Y_i.$$

Sampling-Based vs. Design-Based Uncertainty

- **What is the specific causal parameter you aim to identify through this comparison**?
  - Average causal effect in the **sample** (S-ATE):

$$\frac{1}{n}\sum_{i=1}^{N} R_i \left(Y_{1i} - Y_{0i}\right),$$

    where n is the sample size from a population of size N.
  - Average causal effect in the **population** (P-ATE):

$$\frac{1}{N}\sum_{i=1}^{N} \left(Y_{1i} - Y_{0i}\right).$$

- Internal vs external validity: the distinction between sampling and design-based uncertainty helps clarify these concerns.

Randomization Inference

Randomization Inference

- To fix ideas, assume that **all the population is considered** so that uncertainty follows from the assignment mechanism.
- Aim: derive the randomization distribution of a test statistic, i.e., the distribution induced by the random assignment of treatment given fixed potential outcomes under a sharp null hypothesis.
- Consider testing the hypothesis $Y_{1i} = Y_{0i}$ for all units i.
- **Randomization inference** allows us to obtain the **exact p-value**.
  - **Potential outcomes and covariates are fixed traits of units**.
  - Treatment assignment is imagined to reveal those traits, which can't change by changing the treatment status.
  - A thought experiment follows from the coin flips that assign units to treatments **over all possible combinations**.
- Define the test statistic of interest, like:

$$\frac{1}{N_1}\sum_{i=1}^{N_1} Y_i - \frac{1}{N_0}\sum_{i=1}^{N_0} Y_i.$$

Randomization Inference

- The exact distribution under the null is derived by calculating values of the statistic for each possible combination of randomized assignment: if $N = 6$ and $N_1 = 3$ one has $\binom{6}{3} = 20$ different assignment vectors.
- In randomization-based inference, uncertainty arises from the random assignment to treatment rather than sampling from a large population.
- Permutation methods are enjoying a healthy computer-age revival...
- In practice, a random sub-sample of permutations can generate close-to-exact p-values.
- The **Fisher exact p-value** is the proportion of treatment assignments that yield a test statistic greater than or equal to the observed test statistic.
- Does not require a large-sample approximation, and the approach extends to any statistic (for example, the **Wilcoxon's rank** or the **Kolmogorov-Smirnov** statistics).

Randomization Inference

- The hypothesis $Y_{1i} = Y_{0i}$ is **sharp** because it allows to **fill missing values of potential outcomes**. Can also accommodate the variant $Y_{1i} = Y_{0i} + \kappa$ for some known κ's.
- "**Parallel distributions**": the marginal distribution functions of potential outcomes $F(y_0|D = 0)$ and $F(y_1|D = 1)$ are a constant shift apart: $F(y|D = 1) = F(y - \kappa|D = 0)$.
- This null is not unreasonable: for example, it states that the experimental treatment is irrelevant, a benchmark worth examining.
- Permutations can be done within blocks B_j.
- Building on this idea, an alterative statistics is the regression coefficient on D_i in:

$$Y_i = \beta_0 + \beta_1 D_i + \underbrace{\beta_2 X_i}_{\textit{stratification}} + \varepsilon_i.$$

Neyman's Approach

Neyman's Approach

- When interested in **testing hypotheses** for the P-ATE:

$$\tau \equiv \frac{1}{N}\sum_{i=1}^{N}(Y_{1i} - Y_{0i}) = \bar{Y}_1 - \bar{Y}_0,$$

  one can't fill out missing potential outcomes under the null.
- Assume random sampling and random assignment to treatment.
- Taking the expectation **over both the assignment and the sampling** one can show that:

$$E\left[\frac{1}{n_1}\sum_{i=1}^{n_1} Y_i - \frac{1}{n_0}\sum_{i=1}^{n_0} Y_i | \mathbf{Y}_0, \mathbf{Y}_1, \mathbf{X}\right] = \tau,$$

  so that $\hat{\tau}$ is unbiased for τ.
- What is the **variance** of this estimator when one considers only the variation induced by the particular treatment assignment?
- Note that the states of units i and j are not independent.

Neyman's Approach

- The variance of $\hat{\tau}$ is known since 1923! Specifically:

$$V\left[\hat{\tau}|\mathbf{Y}_0, \mathbf{Y}_1, \mathbf{X}\right] = \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1} - \frac{S_{01}^2}{N_0 + N_1},$$

where the terms involved are the **population variances of potential outcomes** and of the **unit level treatment effect**:

$$\begin{aligned} S_\omega^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{\omega i} - \bar{Y}_\omega\right)^2, \qquad \omega = 0, 1 \\ S_{01}^2 &= \frac{1}{N-1}\sum_{i=1}^{N}\left(Y_{1i} - Y_{0i} - \tau\right)^2. \end{aligned}$$

- The term S_{01}^2 is zero with constant effects, so the largest variance of the estimator comes with constant effects.
- A finite population reduces the variance since, for fixed n_0 and n_1, if $N = N_0 + N_1 \to \infty$:

$$V\left[\hat{\tau}|\mathbf{Y}_0, \mathbf{Y}_1, \mathbf{X}\right] \simeq \frac{S_0^2}{n_0} + \frac{S_1^2}{n_1}.$$

Neyman's Approach

- One can construct unbiased estimators for S_ω^2 using group-level empirical analogues:

$$s_0^2 \equiv \frac{1}{n_0 - 1} \sum_{i=1}^{N} R_i(1 - D_i) \left[Y_i - \left(\frac{1}{n_0} \sum_{i=1}^{N} R_i(1 - D_i) Y_i \right) \right]^2,$$

$$s_1^2 \equiv \frac{1}{n_1 - 1} \sum_{i=1}^{N} R_i D_i \left[Y_i - \left(\frac{1}{n_1} \sum_{i=1}^{N} R_i D_i Y_i \right) \right]^2,$$

- Since S_{01}^2 cannot be estimated, **conservative standard errors** can be obtained as follows:

$$\hat{V}\left[\hat{\tau} | \mathbf{Y}_0, \mathbf{Y}_1, \mathbf{X}\right] = \frac{s_0^2}{n_0} + \frac{s_1^2}{n_1}.$$

- Equivalent to standard errors for two sample t-test with unequal variances or robust standard errors in regressions of Y on D.

Neyman's Approach

- 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.

Conditioning on Covariates

- Covariates $\mathbf{X}$ can be used to define G strata, and stratum-specific estimates of causal effects $\hat{\tau}_g$ can be obtained.
- The new estimator from this stratification is constructed as follows:

$$\hat{\tau}_S = \sum_{g=1}^{G} \frac{N_g}{N} \hat{\tau}_g$$

- If the primary interest is in the **most precise estimate of the average effect**, stratified randomization has lower expected squared error than completely randomized design:

$$\frac{V(\hat{\tau}_S)}{V(\hat{\tau})} \leq 1.$$

- Strictly lower if the covariates predict potential outcomes, irrespective of sample size.
- Fisher's exact p-value can be computed by randomization inference within-strata.

Large Scale Testing

False Positive Rate in Multiple Testing

- The econometric analysis of experiments often involves testing multiple null hypotheses simultaneously.
  - Treatment effects are zero simultaneously for **multiple outcomes**.
  - Treatment effect heterogeneity across **multiple groups**.
  - Effect heterogeneity across **multiple treatments**.
- Consider the case of K hypotheses to test, each with a significance level of α: the probability of a type I error, or false positive, is α.
- The **false positive rate (FPR)** is the expected number of false positives out of all hypothesis tests conducted.
- Assuming independence, the probability of observing at least one significant result just due to chance is $1 - (1 - \alpha)^K$.
- If $K = 20$ and $\alpha = 0.05$, you'll have a 64% chance to be fooled into significant effects (i.e., null "features" called significant).

Randomization Based Testing

- Randomization inference for multiple comparisons can be considered in null hypothesis testing.
- In the case of two outcomes Y_i^A and Y_i^B, the exact distribution of the χ^2 statistic can be computed under the null $\beta_1^A = \beta_1^B = 0$ from the following **system of equations**:

$$\begin{aligned} Y_i^A &= \beta_0^A + \beta_1^A D_i + \beta_2^A X_i + \varepsilon_i^A, \\ Y_i^B &= \beta_0^B + \beta_1^B D_i + \beta_2^B X_i + \varepsilon_i^B. \end{aligned}$$

- A similar idea applies to one equation with coefficients for multiple treatments or groups.
- Alternative approaches adjust α in some way, so that the probability of observing at least one significant result due to chance remains below the desired significance level.

Family-Wise Error Rate in Multiple Testing

- Multiple comparison procedures control for the **family-wise error rate (FWER)**, which is the probability of having any (i.e., one or more) false positives out of all the cases considered.
- The **Bonferroni correction** strengthens the threshold of evidence required to declare an individual case significant, and sets the significance level of each hypothesis at α/K.
- In the example above with $K = 20$ tests and $\alpha = 0.05$, you'd only reject a null hypothesis if the p-value is below 0.0025.
- With a Bonferroni adjustment, we would multiply the observed p-values by the number of outcomes ($K = 20$) and then cap at one.
- The probability of observing at least one significant result just due to chance is $1 - (1 - \alpha/K)^K = 4.8\%$: we guarantee that FWER is less than α.

False Discovery Rate in Multiple Testing

- In applications, depending on the correlation structure across tests, the Bonferroni adjustment can be massively conservative, leading to many missed significant findings.
- The **false discovery rate (FDR)** has become a standard criterion when K is large (e.g., when testing effect heterogeneity along very many dimensions).
- It is defined as the rate that "features" called significant are truly null. An FDR of 5% means that, among all "features" called significant, 5% of these are truly null.
- Controlling for FDR is a way to identify as many significant "features" as possible while incurring a relatively low proportion of false positives.
- Just as we set α as a threshold for the p-value to control the FPR, we can also set a threshold for the q-value, which is the FDR analog of the p-value.

False Discovery Rate in Multiple Testing

- How are **sharpened FDR q-values** computed?.
  - Sort "features" in order of decreasing significance (p-values from smallest to largest):

$$p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(K)},$$

    with $H_{0(i)}$ denoting the corresponding null hypotheses.
  - Define i_{max} to be the largest index i for which:

$$p_{(i)} \leq \frac{i}{K} q,$$

    with q a pre-chosen value between zero and one.
  - Let $\mathcal{D}_q$ be the rule that rejects $H_{0(i)}$ for $i \leq i_{max}$, accepting otherwise.
- It can be demonstrated that the rule $\mathcal{D}_q$ controls FDR at level q.