Gov 2002: 3. Randomization Inference

Matthew Blackwell

September 10, 2015
Where are we? Where are we going?

- Last week:
  - What can we identify using randomization?
  - Estimators were justified via unbiasedness and consistency.
  - Standard errors, test, and CIs were asymptotic.
  - Neyman’s approach to experiments

- This week:
  - Condition on the experiment at hand.
  - Get correct p-values and CIs just relying on randomization.
  - Fisher's approach to randomized experiments.
Effect of not having a runoff in sub-Saharan African

- Glynn and Ichino (2012): is not having a runoff ($D_i = 1$) related to harassment of opposition parties ($Y_i$) in sub-Saharan African countries.
- Without runoffs ($D_i = 1$), only need a plurality $\leq$ incentives to suppress turnout through intimidation.
- With runoffs ($D_i = 0$), largest party needs wider support $\leq$ courting of small parties.
## Data on runoffs

<table>
<thead>
<tr>
<th>Unit</th>
<th>No runoff?</th>
<th>Intimidation</th>
<th>$Y_i(0)$</th>
<th>$Y_i(1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cameroon</td>
<td>1</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>Kenya</td>
<td>1</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>Malawi</td>
<td>1</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>Nigeria</td>
<td>1</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>Tanzania</td>
<td>1</td>
<td>0</td>
<td>?</td>
<td>0</td>
</tr>
<tr>
<td>Congo</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>Madagascar</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>Central African Republic</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>Ghana</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>Guinea-Bissau</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
</tbody>
</table>

- **Clear difference-in-means:** 0.8
- **Very small sample size** → can we learn anything from this data?
Ho & Imai (2006): 2003 CA gubernatorial recall election there were 135 candidates.

Ballot order was randomly assigned so some people ended up on the first page and some did not.

Can we detect an effect of being on the first page on the vote share for a candidate?
What is randomization inference?

- Randomization inference (RI) = using the randomization to make inferences.
- Null hypothesis of no effect for any unit $\leadsto$ very strong.
- Allows us to make exact inferences.
  - No reliance on large-sample approximations.
- Allows us to make distribution-free inferences.
  - No reliance on normality, etc.
- $\leadsto$ truly nonparametric
Brief review of hypothesis testing

RI focuses on hypothesis testing, so it’s helpful to review.

1. Choose a null hypothesis:
   ▶ $H_0 : \beta_1 = 0$ or $H_0 : \tau = 0$.
   ▶ No average treatment effect.
   ▶ Claim we would like to reject.

2. Choose a test statistic.
   ▶ $Z_i = (X_i - \bar{X}) / (s / \sqrt{n})$

3. Determine the distribution of the test statistic under the null.
   ▶ Statistical thought experiment: we know the truth, what data should we expect?

4. Calculate the probability of the test statistics under the null.
   ▶ What is this called? $p$-value
Sharp null hypothesis of no effect

Sharp Null Hypothesis

\[ H_0 : \tau_i = Y_i(1) - Y_i(0) = 0 \quad \forall i \]

- Motto: “No effect means no effect”
- Different than no *average* treatment effect, which does not imply the sharp null.
- Take a simple example with two units:

\[ \tau_1 = 1 \quad \tau_2 = -1 \]

- Here, \( \tau = 0 \) but the sharp null is violated.
- This null hypothesis formally links the observed data to all potential outcomes.
We can use the sharp null \((Y_i(1) - Y_i(0) = 0)\) to fill in the missing potential outcomes:

<table>
<thead>
<tr>
<th>Unit</th>
<th>No runoff?</th>
<th>Intimidation</th>
<th>(Y_i(0))</th>
<th>(Y_i(1))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cameroon</td>
<td>1</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>Kenya</td>
<td>1</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>Malawi</td>
<td>1</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>Nigeria</td>
<td>1</td>
<td>1</td>
<td>?</td>
<td>1</td>
</tr>
<tr>
<td>Tanzania</td>
<td>1</td>
<td>0</td>
<td>?</td>
<td>0</td>
</tr>
<tr>
<td>Congo</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>Madagascar</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>CAR</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>Ghana</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
<tr>
<td>Guinea-Bissau</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>?</td>
</tr>
</tbody>
</table>
Life under the sharp null

We can use the sharp null \((Y_i(1) - Y_i(0) = 0)\) to fill in the missing potential outcomes:

<table>
<thead>
<tr>
<th>Unit</th>
<th>No runoff?</th>
<th>Intimidation</th>
<th>(Y_i(0))</th>
<th>(Y_i(1))</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cameroon</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Kenya</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Malawi</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Nigeria</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Tanzania</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Congo</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Madagascar</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>CAR</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Ghana</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Guinea-Bissau</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
</tbody>
</table>
Comparison to the average null

- Sharp null allows us to say that $Y_i(1) = Y_i(0)$
  - $\implies$ impute all potential outcomes.
- Average null only allows us to say that $\mathbb{E}[Y_i(1)] = \mathbb{E}[Y_i(0)]$
  - $\implies$ tells us nothing about the individual causal effects.
- Don’t need to believe either hypothesis $\implies$ looking for evidence against them!
- Stochastic version of “proof by contradiction.”
Other sharp nulls

- Sharp null of no effect is not the only sharp null of no effect.
- Sharp null in general is one of a constant additive effect:
  \[ H_0 : \tau_i = 0.2. \]
  - Implies that \( Y_i(1) = Y_i(0) + 0.2. \)
  - Can still calculate all the potential outcomes!
- More generally, we could have \( H_0 : \tau_i = \tau_0 \) for a fixed \( \tau_0 \)
- Complications: why constant and additive?
Test Statistic

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

- Typically measures the relationship between two variables.
- Test statistics help distinguish between the sharp null and some interesting alternative hypothesis.
- Want a test statistic with high statistical power:
  - Has large values when the null is false
  - These large values are unlikely when then null is true.
- These will help us perform a test of the sharp null.
- Many possible tests to choose from!
Null/randomization distribution

- What is the distribution of the test statistic under the sharp null?
- If there was no effect, what test statistics would we expect over different randomizations?
- **Key insight of RI:** under sharp null, the treatment assignment doesn’t matter.
  - Explicitly assuming that if we go from \( \mathbf{D} \) to \( \mathbf{\tilde{D}} \), outcomes won’t change.
  - \( Y_i(1) = Y_i(0) = Y_i \)
- **Randomization distribution:** set of test statistics for each possible treatment assignment vector.
Calculate p-values

- How often would we get a test statistic this big or bigger if the sharp null holds?
- Easy to calculate once we have the randomization distribution:
  - Number of test statistics bigger than the observed divided by total number of randomizations.

\[
\Pr(t(d, Y) \geq t(D, Y) | \tau = 0) = \frac{\sum_{d \in \Omega} \mathbb{I}(t(d, Y) \geq t(D, Y))}{K}
\]

- These are **exact tests**:
  - p-values are exact, not approximations.
  - with a rejection threshold of \( \alpha \), RI test will falsely reject less than 100\(\alpha\)% of the time.
1. Choose a sharp null hypothesis and a test statistic,
2. Calculate observed test statistic: $T = t(D, Y)$.
3. Pick different treatment vector $\tilde{D}_1$.
4. Calculate $\tilde{T}_1 = t(\tilde{D}_1, Y)$.
5. Repeat steps 3 and 4 for all possible randomization to get $\tilde{T} = \{\tilde{T}_1, ..., \tilde{T}_K\}$.
6. Calculate the p-value: $p = \frac{1}{K} \sum_{k=1}^{K} \mathbb{I}(\tilde{T}_k \geq T)$
Difference in means

- Absolute difference in means estimator:

\[ T_{\text{diff}} = \left| \frac{1}{N_t} \sum_{i=1}^{N} D_i Y_i - \frac{1}{N_c} \sum_{i=1}^{N} (1 - D_i) Y_i \right| \]

- Larger values of \( T_{\text{diff}} \) are evidence against the sharp null.
- Good estimator for constant, additive treatment effects and relatively few outliers in the the potential outcomes.
Example

- Suppose we are targeting 6 people for donations to Harvard.
- As an encouragement, we send 3 of them a mailer with inspirational stories of learning from our graduate students.
- Afterwards, we observe them giving between $0 and $5.
- Simple example to show the steps of RI in a concrete case.
# Randomization distribution

<table>
<thead>
<tr>
<th>Unit</th>
<th>Mailer</th>
<th>Contr.</th>
<th>$Y_i(0)$</th>
<th>$Y_i(1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Donald</td>
<td>1</td>
<td>3</td>
<td>(3)</td>
<td>3</td>
</tr>
<tr>
<td>Carly</td>
<td>1</td>
<td>5</td>
<td>(5)</td>
<td>5</td>
</tr>
<tr>
<td>Ben</td>
<td>1</td>
<td>0</td>
<td>(0)</td>
<td>0</td>
</tr>
<tr>
<td>Ted</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>(4)</td>
</tr>
<tr>
<td>Marco</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(0)</td>
</tr>
<tr>
<td>Scott</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>(1)</td>
</tr>
</tbody>
</table>

$$T_{\text{rank}} = |8/3 - 5/3| = 1$$
Randomization distribution

<table>
<thead>
<tr>
<th>Unit</th>
<th>Mailer $\bar{D}_i$</th>
<th>Contr. $Y_i$</th>
<th>$Y_i(0)$</th>
<th>$Y_i(1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Donald</td>
<td>1</td>
<td>3</td>
<td>(3)</td>
<td>3</td>
</tr>
<tr>
<td>Carly</td>
<td>1</td>
<td>5</td>
<td>(5)</td>
<td>5</td>
</tr>
<tr>
<td>Ben</td>
<td>0</td>
<td>0</td>
<td>(0)</td>
<td>0</td>
</tr>
<tr>
<td>Ted</td>
<td>1</td>
<td>4</td>
<td>4</td>
<td>(4)</td>
</tr>
<tr>
<td>Marco</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>(0)</td>
</tr>
<tr>
<td>Scott</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>(1)</td>
</tr>
</tbody>
</table>

\[\tilde{T}_{\text{diff}} = |12/3 - 1/3| = 3.67\]

\[\tilde{T}_{\text{diff}} = |8/3 - 5/3| = 1\]

\[\tilde{T}_{\text{diff}} = |9/3 - 4/3| = 1.67\]
## Randomization distribution

| $D_1$ | $D_2$ | $D_3$ | $D_4$ | $D_5$ | $D_6$ | |Diff in means||
|---|---|---|---|---|---|---|---|
| 1  | 1  | 1  | 0  | 0  | 0  | | 1.00||
| 1  | 1  | 0  | 1  | 0  | 0  | | 3.67||
| 1  | 1  | 0  | 0  | 1  | 0  | | 1.00||
| 1  | 0  | 1  | 1  | 0  | 0  | | 1.67||
| 1  | 0  | 1  | 0  | 1  | 0  | | 2.33||
| 1  | 0  | 1  | 0  | 0  | 1  | | 1.67||
| 1  | 0  | 0  | 1  | 1  | 0  | | 0.33||
| 1  | 0  | 0  | 1  | 0  | 1  | | 1.00||
| 1  | 0  | 0  | 0  | 1  | 1  | | 1.67||
| 0  | 1  | 1  | 1  | 0  | 0  | | 1.67||
| 0  | 1  | 1  | 0  | 1  | 0  | | 1.00||
| 0  | 1  | 1  | 0  | 0  | 1  | | 0.33||
| 0  | 1  | 0  | 1  | 1  | 0  | | 1.67||
| 0  | 1  | 0  | 1  | 0  | 1  | | 2.33||
| 0  | 1  | 0  | 0  | 1  | 1  | | 0.33||
| 0  | 0  | 1  | 1  | 1  | 0  | | 1.67||
```r
library(ri)
y <- c(3, 5, 0, 4, 0, 1)
D <- c(1, 1, 1, 0, 0, 0)
T_stat <- abs(mean(y[D == 1]) - mean(y[D == 0]))
Dbold <- genperms(D)
Dbold[, 1:6]
```

```
# [ ,1] [ ,2] [ ,3] [ ,4] [ ,5] [ ,6]
# 1   1   1   1   1   1   1   1
# 2   1   1   1   1   0   0   0
# 3   1   0   0   0   1   1
# 4   0   1   0   0   1   0
# 5   0   0   1   0   0   1
# 6   0   0   0   1   0   0
```
Calculating means

```r
dist <- rep(NA, times = ncol(Dbold))
for (i in 1:ncol(Dbold)) {
  D_tilde <- Dbold[, i]
  rdist[i] <- abs(mean(y[D_tilde == 1]) - mean(y[D_tilde == 0]))
}
dist
```

```r
table(distr)
```

```
P-value

Histogram of rdist

# p-value
mean(rdist >= T_stat)

## [1] 0.8
Order of the candidates on the ballots was randomized in the following way:

1. Choose a random ordering of all 26 letters from the set of 26! possible orderings.
   
   R W Q O J M V A H B S G Z X N T C I E K U P D Y F L

2. In the 1st assembly district, order candidates on the ballot from this order.

3. In the next district, rotate ordering by 1 letter and order names by this.

   W Q O J M V A H B S G Z X N T C I E K U P D Y F L R

4. Continue rotating for each district.
CA recall election with RI

1. Pick another possible letter ordering.
2. Assign 1st page/not first page based on this new ordering as was done in the election.
4. Lather, rinse, repeat.
Other test statistics

- The difference in means is great for when effects are:
  - constant and additive
  - few outliers in the data
- Outliers \(\sim\) more variation in the randomization distribution
- What about 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.
- \( \rightarrow \) transform the observed outcome using the natural logarithm:

\[
T_{\log} = \left[ \frac{1}{N_t} \sum_{i=1}^{N} D_i \log(Y_i) - \frac{1}{N_c} \sum_{i=1}^{N} (1 - D_i) \log(Y_i) \right]
\]

- Useful for skewed distributions of outcomes.
To further protect against outliers can use the differences in quantiles as a test statistics.

Let use $Y_t = Y_i; i : D_i = 1$ and $Y_c = Y_i; i : D_i = 0$.

Differences in medians:

$$T_{med} = |\text{med}(Y_t) - \text{med}(Y_c)|$$

Remember that the median is the 0.5 quantile.

We could estimate the difference in quantiles at any point in the distribution: (the 0.25 quantile or the 0.75 quantile).
Rank statistics

- Rank statistics transform outcomes to ranks and then analyze those.
- Useful for situations
  - with continuous outcomes,
  - small datasets, and/or
  - many outliers
- Basic idea:
  - rank the outcomes (higher values of $Y_i$ are assigned higher ranks)
  - compare the average rank of the treated and control groups
Rank statistics formally

- Calculate ranks of the outcomes:

\[ \tilde{R}_i = \tilde{R}_i(Y_1, \ldots, Y_N) = \sum_{j=1}^{N} \mathbb{I}(Y_j \leq Y_i) \]

- Normalize the ranks to have mean 0:

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

- Calculate the absolute difference in average ranks:

\[ T_{\text{rank}} = |\bar{R}_t - \bar{R}_c| = \left| \frac{\sum_{i:D_i=1} R_i}{N_t} - \frac{\sum_{i:D_i=0} R_i}{N_c} \right| \]

- Minor adjustment for ties.
Randomization distribution

<table>
<thead>
<tr>
<th>Unit</th>
<th>Mailer $D_i$</th>
<th>Contr. $Y_i$</th>
<th>$Y_i(0)$</th>
<th>$Y_i(1)$</th>
<th>Rank</th>
<th>$R_i$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Donald</td>
<td>1</td>
<td>3</td>
<td>(3)</td>
<td>3</td>
<td>4</td>
<td>0.5</td>
</tr>
<tr>
<td>Carly</td>
<td>1</td>
<td>5</td>
<td>(5)</td>
<td>5</td>
<td>6</td>
<td>2.5</td>
</tr>
<tr>
<td>Ben</td>
<td>1</td>
<td>0</td>
<td>(0)</td>
<td>0</td>
<td>1.5</td>
<td>-2</td>
</tr>
<tr>
<td>Ted</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>(4)</td>
<td>5</td>
<td>1.5</td>
</tr>
<tr>
<td>Marco</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(0)</td>
<td>1.5</td>
<td>-2</td>
</tr>
<tr>
<td>Scott</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>(1)</td>
<td>3</td>
<td>-0.5</td>
</tr>
</tbody>
</table>

$$T_{rank} = |1/3 - -1/3| = 0.67$$
Effects on outcome distributions

- Focused so far on “average” differences between groups.
- What about differences in the distribution of outcomes? \( \xrightarrow{\sim} \) Kolmogorov-Smirnov test
- Define the empirical cumulative distribution function:

\[
\hat{F}_c(y) = \frac{1}{N_c} \sum_{i:D_i=0} \mathbb{1}(Y_i \leq y) \quad \hat{F}_t(y) = \frac{1}{N_t} \sum_{i:D_i=1} \mathbb{1}(Y_i \leq y)
\]

- 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

- eCDFs are functions, but we need a scalar test statistic.
- Use the maximum discrepancy between the two eCDFs:

\[ T_{KS} = \max \left| \hat{F}_t(Y_i) - \hat{F}_c(Y_i) \right| \]

- Summary of how different the two distributions are.
- Useful in many contexts!
KS statistic
KS statistic
KS statistic

- Frequency
- Treated
- Control

- y

Graph showing the distribution of treated and control groups with KS statistic.
KS statistic

The KS statistic is a measure of the maximum distance between two empirical distribution functions. In this graph, we compare the distribution of treated and control groups. The top graph shows the density functions, while the bottom graph displays the empirical cumulative distribution functions (eCDFs).

The treated group (blue) and the control group (red) are close in density, suggesting similar distributions. The eCDFs also indicate a similar pattern, with the treated group slightly higher at lower values of y, and the control group slightly higher at higher values of y.
KS statistic

- The top graph illustrates the KS statistic distribution with Frequency on the y-axis and y on the x-axis.
- The middle graph shows the eCDF of Treated and Control groups.
- The bottom graph further details the eCDF with Frequency on the y-axis and y on the x-axis.
KS statistic

The KS statistic is used to compare the empirical cumulative distribution functions (eCDFs) of two samples. The top graph shows the frequency distribution of the treated and control groups, while the bottom graph depicts the eCDFs. The KS statistic is calculated as the maximum difference between the two eCDFs, which in this case is approximately 1.0. This indicates a significant difference between the treated and control groups.
KS statistic
KS statistic

- Frequency

- Treated
KS statistic

The graph shows two distributions: "Control" and "Treated". The distribution for the "Treated" group is more peaked and has a higher frequency at the peak compared to the "Control" group. The frequency axis ranges from 0.00 to 0.35, and the y-axis ranges from -10 to 10.
KS statistic

The upper graph shows the frequency distribution of treated and control groups. The KS statistic is calculated based on the difference between the cumulative distribution functions (CDFs) of the two groups. The lower graph displays the empirical cumulative distribution function (eCDF) for the treated group, indicating the cumulative proportion of observations up to each value of the variable y.
KS statistic

The graph shows the comparison of treated and control distributions using the Kolmogorov-Smirnov (KS) test. The top graph displays the frequency distribution, while the bottom graph illustrates the empirical cumulative distribution functions (eCDFs) for both groups. The KS statistic helps in assessing the dissimilarity between the two distributions.
KS statistic

![Graph showing the Kolmogorov-Smirnov statistic for treated and control groups. The graph compares the empirical cumulative distribution functions (eCDF) for the two groups, highlighting the differences in their distributions. The KS statistic is used to determine if the null hypothesis that the two samples are drawn from the same distribution can be rejected. The plot includes a visual representation of the null distribution, which is crucial for understanding the significance of the observed differences.]
Two-sided or one-sided?

- So far, we have defined all test statistics as absolute values.
- ✈ testing against a two-sided alternative hypothesis:
  \[ H_0 : \tau_i = 0 \quad \forall i \quad H_1 : \tau_i \neq 0 \text{ for some } i \]

- What about a one-sided alternative?
  \[ H_0 : \tau_i = 0 \quad \forall i \quad H_1 : \tau_i > 0 \text{ for some } i \]

- For these, use a test statistic that is bigger under the alternative:
  \[ T^*_{\text{diff}} = \bar{Y}_t - \bar{Y}_c \]
Computing the exact randomization distribution not always feasible:

- \( N = 6 \) and \( N_t = 3 \) \( \leadsto \) 20 assignment vectors.
- \( N = 10 \) and \( N_t = 5 \) \( \leadsto \) 252 vectors.
- \( N = 100 \) and \( N_t = 50 \) \( \leadsto \) \( 1.0089134 \times 10^{29} \) vectors.
- Workaround: simulation!
  - take \( K \) samples from the treatment assignment space.
  - calculate the randomization distribution in the \( K \) samples.
  - tests no longer exact, but bias is under control!
  (increase \( K \))
Confidence intervals via test inversion

- CIs usually justified using Normal distributions and approximations.
- Can calculate CIs here using the duality of tests and CIs:
  - A $100(1 - \alpha)\%$ confidence interval is equivalent to the set of null hypotheses that would not be rejected at the $\alpha$ significance level.
- 95% CI: find all values $\tau_0$ such that $H_0 : \tau = \tau_0$ is not rejected at the 0.05 level.
  - Choose grid across space of $\tau$: $-0.9, -0.8, -0.7, ..., 0.7, 0.8, 0.9$.
  - For each value, use RI to test sharp null of $H_0 : \tau_i = \tau_m$ at 0.05 level.
  - Collect all values that you cannot reject as the 95% CI.
Testing non-zero sharp nulls

- Suppose that we had: \( H_0 : \tau_i = Y_i(1) - Y_i(0) = 1 \)

<table>
<thead>
<tr>
<th>Unit</th>
<th>Mailer</th>
<th>Contr.</th>
<th>( Y_i(0) )</th>
<th>( Y_i(1) )</th>
<th>Adjusted ( Y_i - D_i\tau_0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Trump</td>
<td>1</td>
<td>3</td>
<td>(2)?</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Carly</td>
<td>1</td>
<td>5</td>
<td>(4)?</td>
<td>5</td>
<td>4</td>
</tr>
<tr>
<td>Ben</td>
<td>1</td>
<td>0</td>
<td>(-1)?</td>
<td>0</td>
<td>-1</td>
</tr>
<tr>
<td>Ted</td>
<td>0</td>
<td>4</td>
<td>4</td>
<td>(5)?</td>
<td>4</td>
</tr>
<tr>
<td>Marco</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>(1)?</td>
<td>0</td>
</tr>
<tr>
<td>Scott</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>(2)?</td>
<td>1</td>
</tr>
</tbody>
</table>

- Assignments will now affect \( Y_i \).
- Solution: use adjusted outcomes, \( Y_i^* = Y_i - D_i\tau_0 \).
- Now, just test sharp null of no effect for \( Y_i^* \).

\[
\begin{align*}
Y_i^*(1) &= Y_i(1) - 1 \times 1 = Y_i(0) \\
Y_i^*(0) &= Y_i(0) - 0 \times 1 = Y_i(0) \\
\tau_i^* &= Y_i^*(1) - Y_i^*(0) = 0
\end{align*}
\]
Notes on RI CIs

- CIs are correct, but might have **overcoverage**.
- With RI, p-values are discrete and depend on $N$ and $N_t$.
  - With $N$ and $N_t$, the lowest p-value is 1/20.
  - Next lowest p-value is 2/20 = 0.10.
- If the p-value of 0.05 falls "between" two of these discrete points, a 95% CI will cover the true value more than 95% of the time.
Point estimates

- Is it possible to get point estimates?
- Not really the point of RI, but still possible:
  1. Create a grid of possible sharp null hypotheses.
  2. Calculate p-values for each sharp null.
  3. Pick the value that is “least surprising” under the null.
- Usually this means selecting the value with the highest p-value.
Including covariate information

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

\[ T_{\text{gain}} = |(\bar{Y}_t - \bar{Y}_c) - (\bar{X}_t - \bar{X}_c)| \]

- If $X_i$ is strongly predictive of $Y_i(0)$, then this could have higher power:
  - $T_{\text{gain}}$ will have lower variance under the null.
  - \( \Rightarrow \) easier to detect smaller effects.
Using regression in RI

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

\[
(\hat{\beta}_0, \hat{\beta}_D, \hat{\beta}_X) = \arg\min_{\beta_0, \beta_D, \beta_X} \sum_{i=1}^{N} (Y_i - \beta_0 - \beta_D \cdot D_i - \beta_X \cdot X_i)^2.
\]

- Then, our test statistic could be \( T_{\text{ols}} = \hat{\beta}_D \).
- RI is justified even if the model is wrong!
  - OLS is just another way to generate a test statistic.
  - If the model is “right” (read: predictive of \( Y_i(0) \)), then \( T_{\text{ols}} \) will have higher power.