Section 2: Randomization Inference

Valentine Gilbert

September 16, 2021

Overview

Sampling Uncertainty and Design-Based Uncertainty

Sampling Uncertainty

- We are used to thinking of variability in our estimates as coming from sampling uncertainty
- For example, if we are trying to estimate a population mean based on a random sample

$\Longrightarrow$ We observe only one of many potential sample means

- But this isn't always the most relevant source of uncertainty
- What if we have data on the population of interest and run a randomized experiment? Where does variability in our causal estimate come from?

TABLE I

SAMPLING-BASED UNCERTAINTY (✓ IS OBSERVED, ? IS MISSING)

	Actual Sample			Alternative Sample I			Alternative Sample II			...
Unit	Y_i	Z_i	R_i	Y_i	Z_i	R_i	Y_i	Z_i	R_i	...
1	✓	✓	1	?	?	0	?	?	0	...
2	?	?	0	?	?	0	?	?	0	...
3	?	?	0	✓	✓	1	✓	✓	1	...
4	?	?	0	✓	✓	1	?	?	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	...
n	✓	✓	1	?	?	0	?	?	0	...

Source: Abadie et al. (2020 ECMA)

Sampling Uncertainty and Design-Based Uncertainty

Design-Based Uncertainty

- In randomized experiments, variability in our causal estimates comes from randomization, not sampling
  - We observe outcomes associated with only one of many potential randomization vectors
- Unlike with sampling uncertainty, we know the underlying probability distribution of randomization vectors
- For certain kinds of null hypotheses, we can generate the **exact distribution** of the test statistic under the null hypothesis
- Then we don't have to rely on asymptotic approximations!

$\Longrightarrow$ Can be especially useful when N is small and asymptotic approximations may be poor

TABLE I

SAMPLING-BASED UNCERTAINTY (✓ IS OBSERVED, ? IS MISSING)

	Actual Sample			Alternative Sample I			Alternative Sample II			…
Unit	Y_i	Z_i	R_i	Y_i	Z_i	R_i	Y_i	Z_i	R_i	…
1	✓	✓	1	?	?	0	?	?	0	…
2	?	?	0	?	?	0	?	?	0	…
3	?	?	0	✓	✓	1	✓	✓	1	…
4	?	?	0	✓	✓	1	?	?	0	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	…
n	✓	✓	1	?	?	0	?	?	0	…

TABLE II

DESIGN-BASED UNCERTAINTY (✓ IS OBSERVED, ? IS MISSING)

	Actual Sample			Alternative Sample I			Alternative Sample II			…
Unit	$Y_i^*(1)$	$Y_i^*(0)$	X_i	$Y_i^*(1)$	$Y_i^*(0)$	X_i	$Y_i^*(1)$	$Y_i^*(0)$	X_i	…
1	✓	?	1	✓	?	1	?	✓	0	…
2	?	✓	0	?	✓	0	?	✓	0	…
3	?	✓	0	✓	?	1	✓	?	1	…
4	?	✓	0	?	✓	0	✓	?	1	…
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	…
n	✓	?	1	?	✓	0	?	✓	0	…

Sharp and Dull Nulls

- Randomization inference allows us to test **sharp null hypotheses**
- A sharp null hypothesis is any hypothesis that lets us fill in counterfactual outcomes (i.e. the question marks)
- For example, the null hypothesis that the treatment effect is 0 *for everyone* is sharp
- The null hypothesis that the treatment effect is 5 for everyone is also sharp
- The null hypothesis that the treatment effect is 0 for men and 5 for women is sharp
- **Question**: Why isn't the null hypothesis of 0 average treatment effect sharp?

id	D_i	$Y_i(0)$	$Y_i(1)$
1	1	?	12

Example with N = 4

Suppose you observe the following data:

id	D_i	$Y_i(0)$	$Y_i(1)$
1	1	?	12
2	1	?	15
3	0	3	?
4	0	7	?

- Under a sharp null of a constant treatment effect of 0, what are the unobserved potential outcomes?

Example with N = 4

Suppose you observe the following data:

id	D_i	$Y_i(0)$	$Y_i(1)$
1	1	12	12
2	1	15	15
3	0	3	3
4	0	7	7

- Under a sharp null of a constant treatment effect of 0, what are the unobserved potential outcomes?

Example with N = 4

Suppose you observe the following data:

id	D_i	$Y_i(0)$	$Y_i(1)$
1	1	?	12
2	1	?	15
3	0	3	?
4	0	7	?

- Under a sharp null of a constant treatment effect of 5, what are the unobserved potential outcomes?

Example with N = 4

Suppose you observe the following data:

id	D_i	$Y_i(0)$	$Y_i(1)$
1	1	7	12
2	1	10	15
3	0	3	8
4	0	7	12

- Under a sharp null of a constant treatment effect of 5, what are the unobserved potential outcomes?

Hypothesis Testing

- So a sharp null hypothesis is any hypothesis that lets us fill in unobserved potential outcomes (i.e. counterfactual outcomes)
- How does this help us test if an observed difference in means is due to a causal effect or due to chance?
- Calculate the test statistic you would have calculated under different possible randomization draws
  - With $N = 4$, we can do this for all possible randomization vectors
  - When N is large, we can do this for a random sample of randomization vectors
- Compare the test statistic you calculated with the observed data to the distribution of test statistics under the null
  - If the observed test statistic is very extreme, it's unlikely to be due to chance

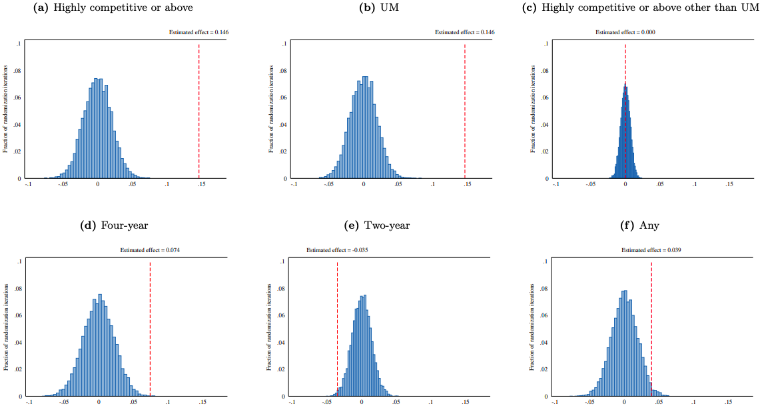
All Potential Randomizations and Differences in Means Under $H_0 : \beta_i = 0$ for all i

D_1	D_2	D_3	D_4	$\bar{Y}(1)$ - $\bar{Y}(0)$
0	0	1	1	-8.5
0	1	0	1	3.5
0	1	1	0	-.5
1	0	0	1	.5
1	0	1	0	-3.5
1	1	0	0	8.5

Question: What's the implied one-sided p-value of the observed difference in means? What about the two-sided p-value?

Example: Dynarski et al. (2018)

Figure 8
Randomization-Based Inference



Notes: Each simulated treatment effect comes from first randomly assigning schools to treatment using the same randomization algorithm used for true assignment, then running a regression of the outcome on "treatment" status, including controls for strata. Exact p-value is calculated as the number of simulated effects greater in absolute value than the estimated effect.

Example: Chetty, Looney, and Kroft (2009)

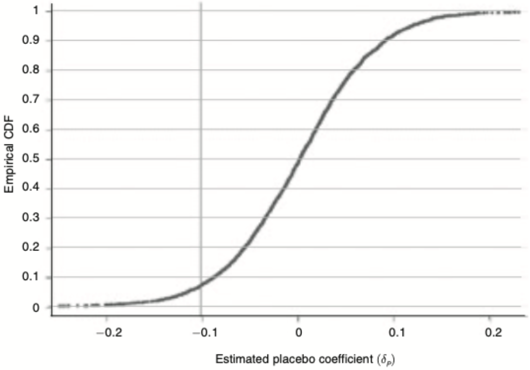


FIGURE 1. DISTRIBUTION OF PLACEBO ESTIMATES: LOG QUANTITY

Notes: This figure plots the empirical distribution of placebo effects (G) for log quantity. The CDF is constructed from 4,725 estimates of δ_p using the specification in column 3 of Table 4. No parametric smoothing is applied: the CDF appears smooth because of the large number of points used to construct it. The vertical line shows the treatment effect estimate reported in Table 4.

Testing $H_0 : \beta_i = c$ for all i

- How do we use randomization inference to test the null that $\beta_i = c$ for all i?
- Just like before, fill in the unobserved potential outcomes
- Then calculate an appropriate test statistic under different randomization vectors
  - e.g. $T^{dif} = |\bar{Y}_1 - \bar{Y}_0 - c|$
  - Question: Why does this make sense as a test statistic?
- Finally, compare the test statistic calculated under the observed randomization vector to the distribution of test statistics under the null and calculate a p-value
  - The p-value is the share of randomization vectors that give you a test statistic *at least* as extreme as the observed one

Aside: Rank Statistic for $H_0 : \beta_i = c$ for all i

- Imbens and Rubin formally define an observation's rank as:

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

- The rank statistic for the null that $\beta_i = 0$ for all i is

$$T^{rank} = |\bar{R}_t - \bar{R}_c|$$

- How would we modify this rank statistic to accommodate a different null hypothesis?
- Imbens and Rubin suggest the following:
  1. Calculate the implied value of $Y_i(0)$ under the null for all units in the data
  2. Convert those $Y_i(0)$ into ranks R_i, so each observations rank is a function of all the *untreated outcomes* in the data (not all *observed outcomes*)
  3. Calculate the test statistic $T = |\bar{R}_t - \bar{R}_c|$

Building Confidence Intervals

- Now that we know how to test different kinds of null hypotheses, it's straightforward to construct a 95% confidence interval
- Recall that we can think of the 95% confidence interval as the set of all null hypotheses that cannot be rejected at the 5% level
- We can therefore construct a 95% confidence interval by searching over a grid of null hypotheses and retaining the ones we fail to reject (just like on problem set 2!)
  - E.g., test the null hypotheses that $\beta_i = c$ for all i for $c \in \{-0.50, -0.49, ..., 0.49, 0.50\}$

Trivia: The Lady Tasting Tea

- Randomization inference is also known as Fisher's exact test, after Ronald Fisher
- Fisher was chatting with colleagues when he offered the phycologist Muriel Bristol a cup of tea. She declined, stating that she preferred tea with the milk poured in first.
- Fisher scoffed at the notion that she could tell the difference, but she insisted she could. So Fisher and a colleague devised a test.
- They brewed 8 cups of tea, 4 of which had milk added first and 4 with tea first.
- They then presented Bristol with the cups in random order and asked her to identify which of the 8 had milk added first.
- Bristol correctly identified all 8 cups. With 70 possible ways of choosing 4 cups out of 8, the implied p-value is $p = 1/70 = 0.014$