Statistical Inference in Experiments II

Ye Wang
University of North Carolina at Chapel Hill

Linear Methods in Causal Inference
POLI784

Review

- There are two consistent estimators for experimental analysis, the Horvitz-Thompson estimator and the Hajek estimator.
- The former is unbiased but the second is more efficient.
- We can conduct statistical inference in experiments with the analytic approach.
- First, we use the Neyman variance estimator to estimate the asymptotic variance.
- The variance captures the design-based uncertainty.
- The variance estimate is conservative unless the treatment effect is constant or the estimand is the PATE.
- Next, we construct confidence intervals using critical values from the normal distribution.

Resampling techniques

- The analytic approach is hard to work with.
- Deriving the variance is challenging, and proving asymptotic normality requires more technicalities.
- The CI may still perform poorly after all the labor.
- An alternative is to rely on resampling techniques.
- They approximate $F_N(\hat{\tau})$ with a direct estimate $\hat{F}_N(\hat{\tau})$ rather than $\mathcal{N}(0, N * Var(\hat{\tau}))$.
- They can be more efficient, and we don't even have to calculate the variance!
- But they do not work everywhere.
- We consider three methods: Fisher's randomization test, boostrap, and jackknife.

Fisher's randomization test

- We usually want to test the weak null hypothesis: $\tau_{SATE} = 0$.
- Fisher suggests that we may also test the sharp null hypothesis: $\tau_i = 0$ for any unit in the sample.
- What is the relationship between the weak null and the sharp null?
- Suppose the sharp null is true, then we actually know the counterfactual of each unit.
- Since the individualistic effect is zero, $Y_i(1) = Y_i(0)$ for any i.
- Now, we know the distribution of the potential outcomes in the sample!

Fisher's randomization test

- Remember that the potential outcomes are fixed quantities.
- Therefore, we can literally run the experiment repeatedly and obtain one ATE estimate from each experiment.
- This will be the true distribution of the estimates, $\hat{F}_N(\hat{\tau})$, under the sharp null.
- We reject the sharp null if the original estimate is an outlier in the distribution.
- This is called Fisher's randomization test (FRT).

Fisher's randomization test

- The true distribution of the potential outcomes is

Unit	$Y_i(1)$	$Y_i(0)$	D_i
1	3	2	
2	5	3	
3	4	5	

- The ATE equals to $(1 + 2 - 1)/3 = 2/3$.

Fisher's randomization test

- Our data is

Unit	Y_i	D_i
1	3	1
2	3	0
3	4	1

- And the ATE estimate is $(3+4)/2 - 3 = 0.5$.

Fisher's randomization test

- Under the sharp null

Unit	$Y_i(1)$	$Y_i(0)$
1	3	3
2	3	3
3	4	4

Fisher's randomization test

- Under the sharp null

Unit	$Y_i(1)$	$Y_i(0)$	D_i	Y_i
1	3	3	1	3
2	3	3	1	3
3	4	4	0	4

- The ATE estimate is $(3+3)/2-4=-1$.

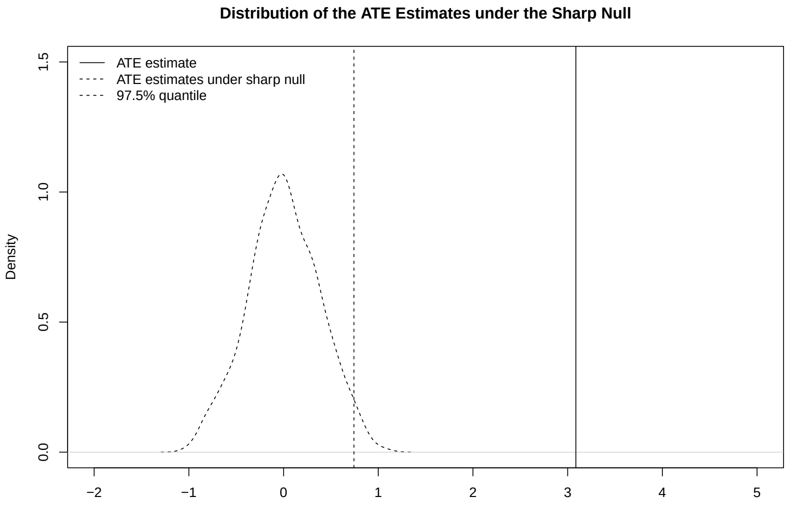
Fisher's randomization test

- Under the sharp null

Unit	$Y_i(1)$	$Y_i(0)$	D_i	Y_i
1	3	3	0	3
2	3	3	0	3
3	4	4	1	4

- The ATE estimate is $4 - (3 + 3)/2 = 1$.

Fisher's randomization test: simulation




```
## The 95% confidence interval is  -0.769 0.743
```

Fisher's randomization test: some theory

- The FRT is built upon a chosen test statistic T—usually our estimator.
- The statistic's value hinges on the observed outcome $\mathbf{Y}$ and treatment assignment $\mathbf{D}$: $T = T(\mathbf{Y}, \mathbf{D})$.
- Given $\mathbf{y}$ and $\mathbf{d}$ from the data, the observed value of the statistic is $T^* = T\left(\mathbf{y}, \mathbf{d}\right)$.
- Under the sharp null and an alternative treatment assignment $\tilde{\mathbf{d}}$, $T = T\left(\mathbf{y}, \tilde{\mathbf{d}}\right)$, which only depends on $\tilde{\mathbf{d}}$.
- We know the distribution of $\tilde{\mathbf{d}}$, thus the distribution of T (and T^*), $F_T(t)$, is also known.
- We will reject the null if $F_T(T^*) > 1 - \alpha$
- The size of the test equals

$$
\begin{aligned}
&\mathbb{P}_{H_0}\left(F_T(T^*) > 1 - \alpha\right) = 1 - \mathbb{P}_{H_0}\left(F_T(T^*) \leq 1 - \alpha\right) \\
=&1 - \mathbb{P}_{H_0}\left(T^* \leq F_T^{-1}(1 - \alpha)\right) = 1 - F_T(F_T^{-1}(1 - \alpha)) = \alpha.
\end{aligned}
$$

Fisher's randomization test: pros and cons

- The FRT works well under complex research designs.
- It ensures the correct coverage even in small samples.
- It circumvents regularity conditions in asymptotic analysis that are not satisfied in certain cases (Young 2019).
- If you know the assignment algorithm but not how to estimate the analytic variance, you can do FRT.
- Applying the FRT to test the weak null leads to anti-conservative results (Wu and Ding 2020).
- We can construct FRTs that have the correct coverage under the sharp null and remain asymptotically valid under the weak null (Cohen and Fogarty 2020).

Bootstrap

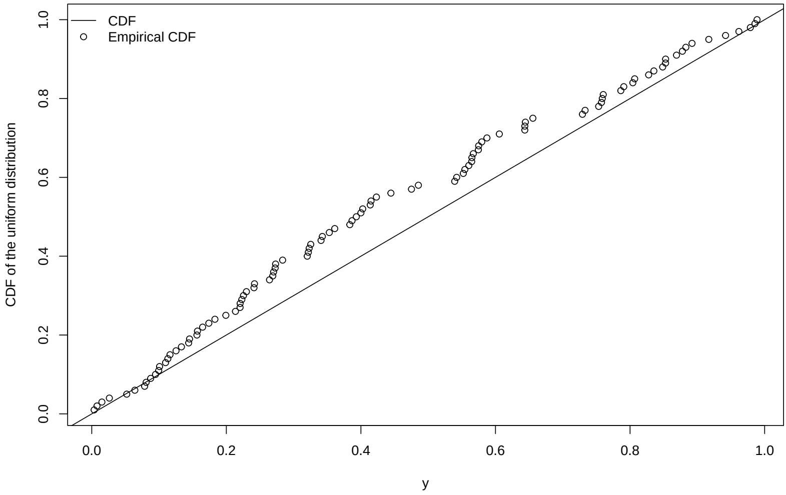
- Recall that an estimator maps the data to an number.
- If we know the distribution of the data, we can resample from it and construct the distribution of the estimate.
- That's what we did in our simulation for the sample average.
- We drew $1,000$ samples from the uniform distribution, $F(y)$, and approximate $\hat{\tau}$'s distribution with these $1,000$ estimates.
- In practice, we do not know the distribution of the data.
- The bootstrap approach suggests that we estimate this distribution with the empirical distribution of our data:

$$\hat{F}(y) = \frac{1}{N}\sum_{i=1}^{N} \mathbf{1}\{Y_i \leq y\}.$$

- The law of large numbers tells us that $\hat{F}(y) \to F(y)$ as $N \to \infty$.

Bootstrap

average-1.pdf



Bootstrap

- We sample from the empirical distribution as if we were sampling from the true distribution $F(y)$.
- The key is to rely on the same sampling strategy.
- This will be accurate when N is large.

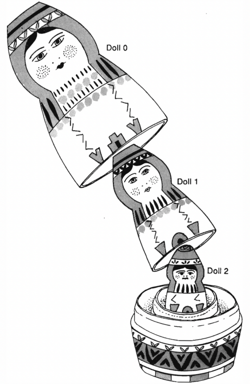


FIGURE 1.1. A Russian matryoshka doll.

Bootstrap: algorithms

- Sampling from the empirical distribution is equivalent to redraw a subsample from the data with replacement.
- The redrawn sample can have an arbitrary size.
- But statistical theory indicates that drawing a subsample with N units is the most efficient approach.
- Statistical inference for $\hat{\tau}$ proceeds by
  1. resampling N observations from the data with replacement,
  2. estimating $\hat{\tau}^*$ using the resampled data, and
  3. constructing confidence intervals from the distribution of $\hat{\tau}^*$.
- We do the same in the setting of causal inference.
- We need redraw Y_i and D_i simultaneously and stick with one estimator.

Bootstrap: algorithms

- Suppose we resample B times and obtain $\{\hat{\tau}_b^*\}_{b=1}^B$ and $\{\hat{\sigma}_b^*\}_{b=1}^B$.
- There are three variants to construct the 95% confidence intervals.
- The Efron method: we find the 2.5% and 97.5% quantiles of $\{\hat{\tau}_b^*\}_{b=1}^B$, $\hat{\xi}_{2.5\%}$ and $\hat{\xi}_{97.5\%}$, and $[\hat{\xi}_{2.5\%}, \hat{\xi}_{97.5\%}]$ will be the confidence interval.
- The percentile t-method: we find the 2.5% and 97.5% quantiles of $\frac{\hat{\tau}_b^* - \hat{\tau}}{\hat{\sigma}_b^*/\sqrt{N}}$ and construct the confidence interval using the original effect and variance estimates plus the bootstrapped quantiles.
- The percentile method: we find the 2.5% and 97.5% quantiles of $\hat{\tau}_b^* - \hat{\tau}$ and subtract them from $\hat{\tau}$.

Bootstrap: algorithms

- Clearly, if we can find numbers $z_{2.5\%}$ and $z_{97.5\%}$ such that

$$P\left(z_{2.5\%} \leq \frac{\hat{\tau} - \tau}{\hat{\sigma}/\sqrt{N}} \leq z_{97.5\%}\right) \geq 95\%,$$

  then the 95% confidence interval will be $[\hat{\tau} - z_{97.5\%} * \hat{\sigma}/\sqrt{N}, \hat{\tau} - z_{2.5\%} * \hat{\sigma}/\sqrt{N}]$.
- The percentile t-method estimates the critical values by finding $\hat{z}_{2.5\%}$ and $\hat{z}_{97.5\%}$ such that

$$P\left(\hat{z}_{2.5\%} \leq \frac{\hat{\tau}^* - \hat{\tau}}{\hat{\sigma}^*/\sqrt{N}} \leq \hat{z}_{97.5\%}\right) \geq 95\%,$$

- Therefore, the bootstrapped the 95% confidence interval is $[\hat{\tau} - \hat{z}_{97.5\%} * \hat{\sigma}/\sqrt{N}, \hat{\tau} - \hat{z}_{2.5\%} * \hat{\sigma}/\sqrt{N}]$.

Bootstrap: algorithms

- The percentile method resamples the centered estimate $\hat{\tau} - \tau$.
- The logic is similar and the bootstrapped the 95% confidence interval is $[\hat{\tau} - \hat{\eta}_{97.5\%}, \hat{\tau} - \hat{\eta}_{2.5\%}]$.
- Here $\hat{\eta}_{97.5\%}$ is an estimate of $z_{2.5\%} * \sigma/\sqrt{N}$.
- For the Efron method, we can see that $[\hat{\xi}_{2.5\%}, \hat{\xi}_{97.5\%}] = [\hat{\tau} + \hat{\eta}_{2.5\%}, \hat{\tau} + \hat{\eta}_{97.5\%}]$.
- Note that $P\left(\hat{\eta}_{2.5\%} \leq \hat{\tau}^* - \hat{\tau} \leq \hat{\eta}_{97.5\%}\right) \geq 95\%$ and $P\left(\hat{\xi}_{2.5\%} \leq \hat{\tau}^* \leq \hat{\xi}_{97.5\%}\right) \geq 95\%$.
- It works only when the true distribution is symmetric hence $\hat{\eta}_{2.5\%} = -\hat{\eta}_{97.5\%}$.
- In this case, the three variants have very similar performance.

Bootstrap: some theory

- The percentile t-method should provide us with a more accurate approximation of the true confidence interval.
- It resamples the t-statistic rather than the estimate.
- We call the transformation from the estimate to the t-statistic "studentization:"

$$t = \frac{\hat{\tau} - \tau}{\hat{\sigma}/\sqrt{N}}$$

- Note that the t-statistic converges to the standard normal distribution, which does not hinge on any parameter that has to be estimated.
- Such statistics are known as "pivotal" statistics.
- Bootstrap pivotal statistics gives us "asymptotic refinement," meaning the CI will be more accurately approximated.
- But of course it requires us to estimate the variance.

Jackknife

- Jackknife was invented before bootstrap.
- But now it is seen as another variant of bootstrap.
- We occasionally use it for variance estimation.
- We leave each unit out and conduct estimation with the rest $N-1$ units.
- We obtain N estimates: $\{\hat{\tau}_i^*\}_{i=1}^N$.
- Their variance is an approximation for the estimate's variance.
- We can also use bootstrap to approximate the estimate's variance.
- But critical values still need to be known.
- It got the name since "it is a rough-and-ready tool that can improvise a solution for a variety of problems."

Jackknife

Bootstrap: simulation

```
## 95% CI from the asymptotic method: 2.118 3.343
## 95% CI from the percentile t-method: 2.13 3.368
## 95% CI from the percentile method: 2.141 3.348
## 95% CI from the Efron method: 2.112 3.32
```

Bootstrap: caveats

- Bootstrap is not always valid.
- It requires the estimtor to be smooth for the empirical distribution.
- It thus fails when the estimator involves truncation or fixed quantities.
- We cannot use bootstrap to infer the extremum (e.g., $\hat{\tau} = \max Y_i$) or constrained estimators (e.g., $\hat{\tau} = \max\{\hat{\tau}^*, 0\}$).
- In causal inference, a well-known example is that bootstrap does not work for nearest-neighbor matching (Abadie and Imbens 2008).

Bootstrap: caveats

- Applying bootstrap to causal inference creates extra complexities.
- Note that we are resampling $\{Y_i, D_i\}_{i=1}^N$ not $\{Y_i(0), Y_i(1), D_i\}_{i=1}^N$.
- At most, we can approximate the marginal distribution of $Y_i(0)$ and $Y_i(1)$, but not their joint distribution.
- We thus ignore the variance caused by treatment effect heterogeneity by using bootstrap.
- The result will be similar to that from using the Neyman variance estimator.
- The problem is identified by Imbens and Menzel (2018).
- They provide a solution to increase the precision of estimation based on the idea in Aronow et al. (2014).

References I

Abadie, Alberto, and Guido W Imbens. 2008. "On the Failure of the Bootstrap for Matching Estimators." *Econometrica* 76 (6): 1537–57.

Aronow, Peter M, Donald P Green, Donald KK Lee, et al. 2014. "Sharp Bounds on the Variance in Randomized Experiments." *The Annals of Statistics* 42 (3): 850–71.

Cohen, Peter L, and Colin B Fogarty. 2020. "Gaussian Prepivoting for Finite Population Causal Inference." *arXiv Preprint arXiv:2002.06654*.

Imbens, Guido, and Konrad Menzel. 2018. "A Causal Bootstrap." National Bureau of Economic Research.

Wu, Jason, and Peng Ding. 2020. "Randomization Tests for Weak Null Hypotheses in Randomized Experiments." *Journal of the American Statistical Association*, 1–16.

References II

Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *The Quarterly Journal of Economics* 134 (2): 557–98.