Lecture 15: Analyzing Randomized Experiments,

Prof. Esther Duflo

14.310x

Analyzing Completely Randomized Experiments

- The Fisher exact test
- Neyman's approach: average treatment effects
- Regression method (to be seen after the regression chapter!)

Fisher: Can we reject that the treatment has no effect on *anyone*

- Fisher was interested in the sharp Null hypothesis:
  $H_o : Y_i(0) = Y_i(1)$ for all i
- Note that this sharp null hypothesis is very different from the hypothesis that the average treatment effect is zero.
- The sharp null allows us to determine for each unit the counterfactual under H_o.
- The beauty of it is that we can calculate, for any test statistics we are interested in, the probability of the observed value under the sharp null
- So, suppose we choose as our statistic the absolute difference in means by treatment status:
  $|T^{ave}(W, Y^{obs})| = |\bar{Y}_t^{obs} - \bar{Y}_c^{obs}|$

Fisher exact test

- We can calculate the probability, over the randomization distribution, of the statistic taking on a value as large, in absolute value, as the actual value given the actual treatment assigned.
- This calculation gives us the p-value for this particular null hypothesis:

$$p = Pr(|T^{ave}(W, Y^{obs}|) > Pr(|T^{ave}(W^{obs}, Y^{obs})|)$$

Example: Cough and Honey

- A randomized study where children were given honey or nothing.
- Main outcome: cough severity the night after the assignment (from 1 to 6)
- Imbens and Rubin (2015) use it to illustrate Fisher exact test
- First, assume we have the data for the first 6 children

the first 6 observations

Table 5.4: First Six Observations on Cough Frequency from Honey Study

Unit	Potential Outcomes		Observed Variables		
	$Y_i(0)$	$Y_i(1)$	W_i	X_i (`cfp`)	Y_i^{obs} (`cfa`)
1	?	3	1	4	3
2	?	5	1	6	5
3	?	0	1	4	0
4	4	?	0	4	4
5	0	?	0	1	0
6	1	?	0	5	1

Reference: Imbens and Rubin "Causal inference for statistics, social and biomedical sciences"

$T^{obs} = 8/3 - 5/3 = 1$

Filling out the counterfactual under the Sharp null

Table 5.5: First Six Observations from Honey Study with missing potential outcomes in brackets filled in under the null hypothesis of no effect

Unit	Potential Outcomes		Observed Variables			
	$Y_i(0)$	$Y_i(1)$	Treatment	X_i	Y_i^{obs}	$\text{rank}(Y_i^{\text{obs}})$
1	(3)	3	1	4	3	4
2	(5)	5	1	6	5	6
3	(0)	0	1	4	0	1.5
4	4	(4)	0	4	4	5
5	0	(0)	0	1	0	1.5
6	1	(1)	0	5	1	3

Reference: Imbens and Rubin "Causal inference for statistics, social and biomedical sciences"

All the possible assignment vector, and associated statistic

W_1	W_2	W_3	W_4	W_5	W_6	levels	ranks
0	0	0	1	1	1	-1.00	-0.67
0	0	1	0	1	1	-3.67	-3.00
0	0	1	1	0	1	-1.00	-0.67
0	0	1	1	1	0	-1.67	-1.67
0	1	0	0	1	1	-0.33	0.00
0	1	0	1	0	1	2.33	2.33
0	1	0	1	1	0	1.67	1.33
0	1	1	0	0	1	-0.33	0.00
0	1	1	0	1	0	-1.00	-1.00
0	1	1	1	0	0	1.67	1.33
1	0	0	0	1	1	-1.67	-1.33
1	0	0	1	0	1	1.00	1.00
1	0	0	1	1	0	0.33	0.00
1	0	1	0	0	1	-1.67	-1.33
1	0	1	0	1	0	-2.33	-2.33
1	0	1	1	0	0	0.33	0.00
1	1	0	0	0	1	1.67	1.67
1	1	0	0	1	0	1.00	0.67
1	1	0	1	0	0	3.67	3.00
1	1	1	0	0	0	**1.00**	**0.67**

Reference: Imbens and Rubin "Causal inference for statistics, social and biomedical sciences"

p value?

All the possible assignment vector, and associated statistic

W_1	W_2	W_3	W_4	W_5	W_6	levels	ranks
0	0	0	1	1	1	-1.00	-0.67
0	0	1	0	1	1	-3.67	-3.00
0	0	1	1	0	1	-1.00	-0.67
0	0	1	1	1	0	-1.67	-1.67
0	1	0	0	1	1	-0.33	0.00
0	1	0	1	0	1	2.33	2.33
0	1	0	1	1	0	1.67	1.33
0	1	1	0	0	1	-0.33	0.00
0	1	1	0	1	0	-1.00	-1.00
0	1	1	1	0	0	1.67	1.33
1	0	0	0	1	1	-1.67	-1.33
1	0	0	1	0	1	1.00	1.00
1	0	0	1	1	0	0.33	0.00
1	0	1	0	0	1	-1.67	-1.33
1	0	1	0	1	0	-2.33	-2.33
1	0	1	1	0	0	0.33	0.00
1	1	0	0	0	1	1.67	1.67
1	1	0	0	1	0	1.00	0.67
1	1	0	1	0	0	3.67	3.00
1	1	1	0	0	0	**1.00**	**0.67**

Reference: Imbens and Rubin "Causal inference for statistics, social and biomedical sciences"

p value? $\frac{16}{20} = 0.8$

Simulation based p value

- If we have more observations we may not be able to do all the permutations (N choses 2)
- Can do it as simulation: draw an assignement. Compute the statistics. repeat K times, compute the probability that the statistics is above the observed statistics.
- Example for cough and honey study (35 honey, 37 control).

Number of Simulations	p-value	(s.e.)
100	0.010	0.010
1,000	0.044	0.006
10,000	0.044	0.002
100,000	0.042	0.001
1,000,000	0.043	0.000

Neyman: The average treatment effect

- Neyman was interested in the difference $\tau = \frac{1}{N}\sum_{i=1}^{N} Y(1) - Y(0) = \overline{Y(1)} - \overline{Y(0)}$ (average treatment effect in the sample).
- And also in constructing confidence interval for this difference
- And in particular to test whether this difference was 0.
- Neyman and Fisher where really upset with each other!
- Note that it is not the same as the sharp null: the treatment effect could be zero on average even if it is not zero for some or even all units.
- Suppose we have a completely randomized experiment, with N_t treatment unit, and N_c control units
- What would seems a reasonable estimator for the object of interest?

Neyman: The average treatment effect

- Neyman was interested in the difference $\tau = \frac{1}{N}\sum_{i=1}^{N} Y(1) - Y(0) = \overline{Y(1)} - \overline{Y(0)}$ (average treatment effect in the sample).
- And also in constructing confidence interval for this difference
- And in particular to test whether this difference was 0.
- Neyman and Fisher where really upset with each other!
- Note that it is not the same as the sharp null: the treatment effect could be zero on average even if it is not zero for some or even all units.
- Suppose we have a completely randomized experiment, with N_t treatment unit, and N_c control units
- What would seems a reasonable estimator for the object of interest?
- $\hat{\tau} = \frac{1}{N_t}\sum_{i:W_i=1} Y_i^{obs} - \frac{1}{N_c}\sum_{i:W_i=0} Y_i^{obs} = \overline{Y_t^{obs}} - \overline{Y_c^{obs}}$ (the difference in outcomes by observed treatment status)

The difference in sample means is an unbiased estimator of average treatment effect

$$\hat{\tau} = \frac{1}{N_t} \sum_{i:W_i=1} Y_i^{obs} - \frac{1}{N_c} \sum_{i:W_i=0} Y_i^{obs}$$

$$= \frac{1}{N} \sum \frac{W_i Y_i(1)}{\frac{N_t}{N}} - \frac{1}{N} \sum \frac{(1 - W_i) Y_i(0)}{\frac{N_c}{N}}$$

Only W_i and $1 - W_i$ are random. The potential outcomes are fixed.

$$E[\hat{\tau}] = \frac{1}{N} \sum \frac{E[W_i] Y_i(1)}{\frac{N_t}{N}} - \frac{1}{N} \sum \frac{E[1 - W_i] Y_i(0)}{\frac{N_c}{N}}$$

$E[W_i] = P(W_i) = \frac{N_t}{N}$ and $E[W_i] = 1 - P(W_i) = \frac{N_c}{N}$
therefore $E[\hat{\tau}] = \overline{Y(1)} - \overline{Y(0)}$

Constructing the variance of $\hat{\tau}$

It can easily (if tediously) be shown that $V(\hat{\tau}) = \frac{S_c^2}{N_c} + \frac{S_t^2}{N_t} - \frac{S_{tc}^2}{N}$.[1] where S_c^2 is the variance of $Y_i(0)$ in the sample and S_t^2 is the variance of $Y_i(1)$ in the sample:

$$S_c^2 = \frac{1}{N_c - 1} \sum_i (Y_i(0) - \overline{Y_i(0)})^2$$

and S_{tc}^2 is the variance of the unit level treatment effects:

$$S_{tc}^2 = \frac{1}{N - 1} \sum (Y_i(1) - Y_i(0) - (\overline{Y_i(1)} - \overline{Y_i(0)}))^2$$

If the treatment effect is constant this term is zero. If it is not constant, this term is positive, and therefore *reduces* the sampling variance.

[1] see Imbens and Rubin , "causal inference", chapter 6, appendix B

Estimating the variance of $\hat{\tau}$

1. $s_c^2 = \frac{1}{N_c - 1} \sum_{i:W_i=0} (Y_i(0) - \overline{Y_c^{obs}})^2$

Estimating the variance of $\hat{\tau}$

1. $s_c^2 = \frac{1}{N_c - 1} \sum_{i:W_i=0} (Y_i(0) - \overline{Y_c^{obs}})^2$
2. $s_t^2 = \frac{1}{N_t - 1} \sum_{i:W_i=1} (Y_i(1) - \overline{Y_t^{obs}})^2$

Estimating the variance of $\hat{\tau}$

1. $s_c^2 = \frac{1}{N_c - 1} \sum_{i:W_i=0} (Y_i(0) - \overline{Y_c^{obs}})^2$
2. $s_t^2 = \frac{1}{N_t - 1} \sum_{i:W_i=1} (Y_i(1) - \overline{Y_t^{obs}})^2$
3. What about the third term? We cannot easily estimate it because we never see $Y_i(1)$ and $Y_i(0)$.
4. Neyman proposed to ignore it and use as estimator of the sampling variance: $V_{Neyman} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}$ and this is typically what we do today.
5. Three justifications:

Estimating the variance of $\hat{\tau}$

1. $s_c^2 = \frac{1}{N_c - 1} \sum_{i:W_i=0} (Y_i(0) - \overline{Y_c^{obs}})^2$
2. $s_t^2 = \frac{1}{N_t - 1} \sum_{i:W_i=1} (Y_i(1) - \overline{Y_t^{obs}})^2$
3. What about the third term? We cannot easily estimate it because we never see $Y_i(1)$ and $Y_i(0)$.
4. Neyman proposed to ignore it and use as estimator of the sampling variance: $V_{Neyman} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}$ and this is typically what we do today.
5. Three justifications:
   - if the treatment effect is constant, it is correct
   - if the treatment effect is not constant, it is conservative (since accounting for variance of the treatment effect would in fact reduce the sampling variance of the estimator of the average treatment effect in the sample).
   - it turns out that if we were interested in the best estimate of the average treatment effect *for the population that the sample is drawn from*, the difference in observed outcome would still be an unbiased estimate, and the third term would drop in the variance.

Confidence intervals

- Recall our prior definition of confidence intervals: We want to find function of the random sample A and B such that $P(A(X_1 \dots X_N) < \theta < B(X_1 \dots X_N)) > 1 - \alpha$
- In this case, this becomes: $P(C_L(X^{obs}, W) < \tau < C_U(X^{obs}, W))) > 1 - \alpha$
- And the only reason why the lower and upper bounds are random is through their dependance on W, the assignment.
- $CI^{\tau}_{1-\alpha} = (\hat{\tau} - t_{crit} * \sqrt{(\widehat{V_{neyman}})}, \hat{\tau} + t_{crit} * \sqrt{(\widehat{V_{neyman}})})$
- with small samples take t_{crit} from a table of t-distribution (as we saw in the CI lecture)
- with larger samples, we can use the normal approximation and take the critical value from the standard normal tables, e.g. 1.645 for $\alpha = 0.1$, and 1.96 for $\alpha = 0.05$.

Hypothesis testing

Let's start with the Neyman hypothesis:

$$H_o : \frac{1}{N}\sum_{i=1}^{N} Y(1) - Y(0) = 0$$

$$H_a : \frac{1}{N}\sum_{i=1}^{N} Y(1) - Y(0) \neq 0$$

Natural test statistics (following our discussion last week): $t = \frac{\overline{Y_t^{obs}} - \overline{Y_c^{obs}}}{\sqrt{\hat{V}_{Neyman}}}$ Follows a t distribution with $N-1$ degrees of freedom, or with N large enough, a normal distribution. Associated p value for two sided test : $2 * (1 - \Phi(t))$ [for the normal approximation]

Neyman approach: An example

- Duflo, Hanna, Ryan: a teacher incentive experiment
- 54 Control schools, 53 treated schools, where teachers are paid more if they show up to school.
- Questions: effect on school presence, kids test scores.

Summary statistics

Table 6.1: Summary Statistics for Duflo-Hanna-Ryan Teacher-Incentive Data

	Variable	Control ($N_c = 54$) avg	(s.d.)	Treated ($N_t = 53$) avg	(s.d.)	min	max
pretreatment	`pctprewritten`	0.19	0.19	0.16	0.17	0.00	0.67
posttreatment	`open`	0.58	0.19	0.80	0.13	0.00	1.00
	`pctpostwritten`	0.47	0.19	0.52	0.23	0.05	0.92
	`written`	0.92	0.45	1.09	0.42	0.07	2.22
	`written_all`	0.46	0.32	0.60	0.39	0.04	1.43

Important components of the estimation

Table 6.2: Estimates for Effect of Teacher Incentives on Proportion of Days that School is Open

Estimated Means	$\overline{Y}_c^{\text{obs}}$	0.58
	$\overline{Y}_t^{\text{obs}}$	0.80
	$\hat{\tau}$	0.22
Estimated Variance Components	s_c^2	0.19^2
	s_t^2	0.13^2
	s^2	0.16^2
	N_c	54
	N_t	53
Sampling Variance Estimates	$\hat{\mathbb{V}}_{\text{neyman}} = \frac{s_c^2}{N_c} + \frac{s_t^2}{N_t}$	0.03^2

Confidence intervals for various variables

Table 6.3: Estimates of, and Confidence Intervals for, Average Treatment Effects for Duflo-Hanna-Ryan Teacher-Incentive Data

		ate	(s.e.)	95% c.i.
pretreatment	`pctprewritten`	-0.03	(0.04)	[-0.10,0.04]
posttreatment	`open`	0.22	(0.03)	[0.15,0.28]
	`pctpostwritten`	0.05	(0.04)	[-0.03,0.13]
	`written`	0.17	(0.08)	[0.00,0.34]
	`written_all`	0.14	(0.07)	[0.00,0.28]

Design of experiments

- Contrary to ex-post analysis where data is given to us, we have the freedom to *design* experiments to answer the question we are interested in.
- There are many interesting design questions and we will get back to them later in the semester, but for now we are equipped to ask a simple question: suppose we are interested in designing a simple randomized experiment to test the Neyman hypothesis. How large should our sample be?
- This calculation is called a *power calculation*.

Power calculations

- For a sample of size N, we will observe $W_1 \ldots W_N$, and $Y_1^{obs} \ldots Y_N^{obs}$
- Suppose we are interested in testing: $H_o = E[Y_i(1) - Y_i(0)] = 0$ against $H_a : E[Y_i(1) - Y_i(0)] \neq 0$
- A reminder:

	H_o true	H_o false
accept H_o	No error	Type II error
reject H_o	Type I error	No error

The significance level of the test, α, is the probability of type I error.

The operating characteristic of the test, β, is the probability of type II error.

We call $1-\alpha$ the confidence level. We call $1-\beta$ the power.

What ingredients goes into the power calculation?

- We tend to pick α low because society does not want to conclude that some treatment work when it fact it really does not.
- Following Fisher, it is often $\alpha = 0.05$
- We want to pick $N = N_c + N_t$ such that , if the average treatment effect is in fact some value τ, the power of the test will be at least $1 - \beta$ for some β, given that a fraction γ of the units are assigned to the treatment group.
- In addition we must assume (know) something about the variance of the outcome in each treatment arm: for simplicity we often assume it is the same, and some parameter σ^2.
- In summary we know, impose, or assume $\alpha, \beta, \tau, \sigma$, and γ, and we are looking for N.
- Alternatively, we could be interested in the power for a given sample size: we know $\alpha, \gamma, \tau, \sigma$, and N and look for —β.

Guess work

- α and β are imposed and we can decide γ (if this was just power what would we pick?)
- Problem: how do we know/determine τ and σ ?
  - τ: could be known from a pilot, from a previous study, or could be picked as a value of interest.
    - For example: the lowest value such that, if we could reject zero when the effect is really τ, the program would be worth doing.
    - This is more about optics than about statistics... (rejecting zero is not "accepting" the point estimate...)
    - But it has the merit to remind us that we may be interesting in 'detecting' a small effect, we will work with large sample. If the program is very expensive such as it won't be adopted unless the effect are very large anyway, we can go with a smaller sample.
  - σ: Need to get that from prior data, with similar outcomes.
  - Some item it is wide guess work!

Now for the formulas

- This is of course in practice the easy part: many software will give you power curves as you tinker with the parameters and the sample size.
- But it is worth working through the logic.
- $T = \frac{\overline{Y_t^{obs}} - \overline{Y_c^{obs}}}{\sqrt{\hat{V_{Neyman}}}} \approx \frac{\overline{Y_t^{obs}} - \overline{Y_c^{obs}}}{\sqrt{\frac{\sigma^2}{N_t}} + \sqrt{\frac{\sigma^2}{N_t}}}$
- We reject this hypothesis if $|T| > \Phi(1 - \frac{\alpha}{2})$, e.g. if $\alpha = 0.05$, if $|T| > 1.96$
- What is the probability that this occurs?
- By the central limit theorem, the difference in means minus the true treatment effect, scaled by the true standard error of that difference, has distribution that is approximately $N(0, 1)$:
- $\frac{\overline{Y_t^{obs}} - \overline{Y_c^{obs}} - \tau}{\sqrt{\frac{\sigma^2}{N_t} + \frac{\sigma^2}{N_t}}} \approx N(0, 1)$ and hence $T \approx N(\frac{\tau}{\sqrt{\frac{\sigma^2}{N_t}} + \sqrt{\frac{\sigma^2}{N_t}}}, 1)$

So

$$P\big(|T| > \Phi(1-\frac{\alpha}{2})\big) \approx \Phi\big((-\Phi^{-1}(1-\frac{\alpha}{2}) + \frac{\tau}{\sqrt{\frac{\sigma^2}{N_t}+\frac{\sigma^2}{N_c}}}\big) +$$

$$\Phi\big(-\Phi^{-1}(1-\frac{\alpha}{2}) - \frac{\tau}{\sqrt{\frac{\sigma^2}{N_t}+\frac{\sigma^2}{N_c}}}\big)$$

The second term is very small, so we ignore it.
So we want the first term to be equal to β, which requires:

$$\Phi^{-1}(\beta) = -\Phi^{-1}(1-\frac{\alpha}{2}) + \frac{\tau\sqrt{N}\sqrt{\gamma(1-\gamma)}}{\sigma}$$

Which leads to the required sample size:

$$N = \frac{(\Phi^{-1}(\beta) + \Phi^{-1}(1-\frac{\alpha}{2}))^2}{\frac{\tau^2}{\sigma^2}.\gamma.(1-\gamma)}$$

Other considerations to take into account when you do power calculations

- If you have stratified or not: with stratified design, variance of estimated treatment effect is lower.
- If you have clustered or not: with clustered design, variance of estimated treatment effect is larger

Analysis of a stratified design

- Take the difference in means within each strata
- Take a weighted average of the treatment effect with the weight the size of the strata $\sum_g (\frac{N_g}{N}) \widehat{\tau_g}$
- This will be an unbiased estimate of the average treatment effect
- And the variance will be calculated as $\sum_g (\frac{N_g}{N})^2 \widehat{V_g}$
- Special case: probability of assignment to control group stays the same in each strata. Then this coefficient is equal to the simple difference between treatment and control, but the variance is always weakly lower.
- Stratification will lower the required sample size for a given power.

Analysis of a clustered design

- The opposite happens with a clustered design (all the unit within a same unit are either treated or control).
- We need to take into account the fact that the potential outcomes for units within a randomization clusters are not independent.
- Conservative way to do this: just average the outcome by unit, and treat each as an observation (like we did for classrooms in the Duflo-Hanna data.
- Then the number of observations is the number of clusters, and you can analyze this data exactly as a completely randomized experiment but with clusters as the unit of analysis.
- For example, this tells you that a randomization with two clusters is unlikely to go very far!!

References

- Imbens and Rubin *Causal Inference for Statistics Social and biomedical Sciences*
- Duflo, Hanna, Ryan "Incentives work". American Economic Review