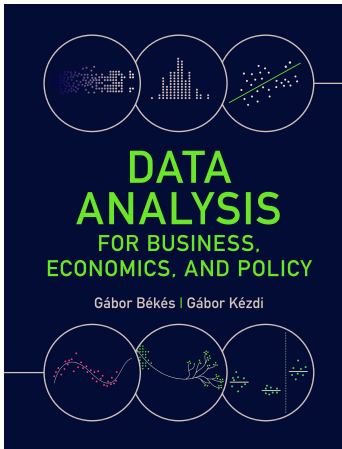# VII: Experimental causal analysis
(randomized field experiments in practice I)

Alexandra Avdeenko

Evaluation

2021

Literature

▶ Cambridge University Press, 2021

▶ gabors-data-analysis.com
  ▶ Download all data and code: gabors-data-analysis.com/data-and-code/

▶ Slides for Chapter 20

Controlled experiments

- ▶ The most powerful way to identify cause and effect relationships is conducting controlled experiments.
- ▶ Controlled experiments allows for controlling variation in the causal variable
- ▶ Variation in the causal variable x is controlled by assigning values of x to the observations.
- ▶ This practice is called controlled assignment.
- ▶ By controlling assignment:
  - ▶ No self-selection: the value of x observations "receive" is not affected by the decisions of people who may be interested in the outcome.
  - ▶ No reverse causality by not letting the outcome y affect x in any way.
- ▶ If binary treatment x, observations are assigned to a treated and an untreated ("control") group by the experimenter.

Controlled experiments

▶ Well-controlled assignment of x results in very similar features for observations that are different in x.
▶ With a binary treatment, observations in the treatment group and observations in the control group are expected to have (on average)
  ▶ the same values of $Y(0)$
  ▶ same values of the treatment effect $Y(1)-Y(0)$

▶ Independence assumption of the potential outcomes framework is met.
▶ ATE will be identified and we can estimate ATE

Controlled experiment types

- ▶ field experiments: aim to be as similar as possible to real-world decision situations
    - ▶ Test the impact of small loans in rural areas
- ▶ lab experiments: are carried out in an artificial environment, usually a computer lab
    - ▶ Test how people play games, react to incentives
- ▶ A/B tests aim to evaluate different versions of the same product
    - ▶ online presentation of an advertisement or a website.

Randomization in controlled experiments

▶ Controlled assignment often involves randomization.

▶ Randomization is an assignment rule: it assigns different values of x to different observations.

▶ Random assignment rule is independent of all potential confounders.
  ▶ A good assignment rule makes sure assignment is independent of everything that may affect the outcome.

▶ This independence is by design.
  ▶ A well-executed randomization guarantees that features are very similar in groups with different values of the causal variable x, such as treated and control groups.
  ▶ remember: ideal would be measuring the average TE, but instead we can have control and treated groups to be similar (on average)

The Experimental Setup

- ▶ Well-controlled experiments - average difference in y identify the average effect of x.
- ▶ With a binary treatment, the average difference in y between the treated and untreated group identifies the average treatment effect.
- ▶ Simple regression of $y^E = \alpha + \beta x$ plus a well-controlled experiment: Estimated ATE$= \beta$

Randomization in controlled experiments

▶ Randomized controlled trial – RCT: controlled experiment with randomization as assignment rule.

▶ RCT are run to estimate the effect of an intervention

▶ Randomized assignment (treated units vs. control units) ensures independence.

▶ So average observed outcomes approximate well potential outcomes

Random assignment and checking balance

▶ Random assignment should make the distribution (incl. the expected value) of all variables the same in assignment groups

▶ Binary: in the treated and control groups
  ▶ Quantitative intervention, at all values

▶ This is called balance. The variables are said to be balanced across groups if their distribution is the same.

▶ Must check it - process of random assignment may be imperfect.

▶ A random rule leads to independence of potential outcomes in expectation.

Case study: Working from home

- ▶ Working from home four days a week makes employees more or less likely to quit the firm and whether it affects their performance.
- ▶ Data is from employees of a travel agency in China.
- ▶ Bloom, Liang, Roberts, Ying "Does Working from Home Work? Evidence from a Chinese Experiment".
- ▶ The company is a large travel agency in China. The experiment took place in its call center in Shanghai that dealt with booking hotels and airfare.
- ▶ Background: commuting time for employees was 80 minutes per day. Employees who were subjects of the experiment work in cubicles.

Case study: Working from home

- This is a field experiment.
- The intervention is making employees work from home four days a week.
- The subjects are employees of the firm.
- Two outcome variables:
  - quit firm (yes/no)
  - performance (number of phone calls).
- About half of the subjects were order takers - answer calls from customers and administer those calls. Performance is easy to measure by counting the number of phone calls they processed.

Case study: Working from home

- 503 people volunteered for the experiment. Of them 249 qualified for the experiment (they worked for more than six months at the company and had broadband internet access and independent workspace at home).
- 131 were assigned to work from home four days a week,
- 118 employees were assigned to continue work in the office.
- Selection was based on birthdays:
  - even birth date assigned to work from home,
  - odd birth date were assigned to work from the office.
- Consider the assignment rule random.

Noncompliance and intent-to-treat

▶ Suppose we have random assignment - RCT.

▶ The intervention may not reach all subjects that were assigned to it.

▶ The intervention may reach some subjects that were not assigned to it.

Noncompliance and intent-to-treat

- ▶ When assignment to treatment and actual treatment do not conform we have noncompliance in the experiment.
    - ▶ assignment: what a subject should do
    - ▶ treatment: what a subject actually does
- ▶ The extent of compliance indicates how closely assignment and treatment are - on both directions
- ▶ Compliance is perfect if assignment and actual treatment are the same
- ▶ Compliance is imperfect if there is some noncompliance
- ▶ Non-compliance - broader concept: all cases when assignment and actual treatment differ, any reason.
    - ▶ Intervention does not get done is some locations

Noncompliance and intent-to-treat

- ▶ With imperfect compliance, we distinguish two kinds of average treatment effects.
- ▶ The effect of the treatment itself -> ATE
- ▶ The effect of being assigned to the treatment –> average intent-to-treat effect, $AITTE$, or simply intention-to-treat effect, ITT
- ▶ Estimate AITTE = average observed outcomes among subjects assigned to treatment - subjects assigned to non-treatment
- ▶ When compliance is perfect the $AITTE = ATE$
- ▶ When compliance is not perfect the $AITTE \neq ATE$

Noncompliance and intent-to-treat

▶ Unfortunately, imperfect compliance makes it hard to estimate the average treatment effect even if assignment to treatment is random.

▶ compliance is rarely random.

▶ Most common issue is self-selection: compliance is often a decision made by the subjects of the intervention.

    ▶ Those that decide to comply with the assignment are likely to be different from those that decide not to comply

    ▶ ...in ways that are related to potential outcomes, such as how much they think they would benefit from the treatment.

▶ With random assignment but imperfect compliance we can get a good estimate of the average intent-to-treat effect but we can't get a good estimate of the average treatment effect.

Noncompliance and intent-to-treat

▶ If non-compliance - what can we do?

▶ If small, acknowledge it

▶ If substantial, ATE cannot be safely estimated

▶ Often, we care about AITTE - that is closer to what we shall expect from a similar intervention.

▶ We can also calculate the average effect on the subjects who comply with the assignment.

▶ Local average treatment effect (LATE) a measures besed in AITTE which we can compute

    ▶ scaled-up from AITTE, in which the scaling factor is the inverse of the proportion of those that comply.

    ▶ Because this is what we observe

    ▶ More in Under the Hood section.

Estimation and statistical inference

▶ Random assignment = difference between the average outcome in the treatment group and the non-treatment group is a good estimate of ATE
  ▶ ATE if compliance is perfect
  ▶ AITTE if compliance is imperfect.

▶ The regression approach uncovers the exact same difference as comparing means by a t-test. Just easier.

$$y^E = \alpha + \beta x \tag{1}$$

  ▶ here α is average y in the non-treatment group;
  ▶ the estimate of β is the difference in the average outcome (y) between the treatment group ($x = 1$) and the non-treatment group ($x = 0$).

Including Covariates in a Regression

▶ With random assignment there is no need to include any variable for getting an unbiased estimate of ATE.

▶ Often estimate the effects with multiple regressions that include other variables.

▶ Including them may serve as an indirect way of checking balance. Especially problematic vars.

▶ Get more precise estimates of the average treatment effect (smaller SE of β_1).
  ▶ (Chapter 10) depends on R^2_{xz} = R-squared of the regression of x on the z variables.

▶ A regression with outcome y, treatment variable x and other variables $z_1, z_2, \ldots$ can be formulated in the following way:

$$y^E = \beta_0 + \beta_1 x + \beta_2 z_1 + \beta_3 z_2 + \ldots \tag{2}$$

Case study: Working from home

- ▶ In principle, all variables that affect potential outcomes need to be balanced.
- ▶ In practice, we can check balance only for variables that are measured. Typically: mean values.
- ▶ If very similar, we say the variables are balanced.
- ▶ Hypothesis test: difference mean in the treatment vs control group.
- ▶ Look at magnitude. Say, 1/10 SD difference = small.
- ▶ If some variables worrisome, we will add them as controls, but can carry on.
- ▶ If many variables, large differences: something went wrong or - if small sample - we were very unlucky

Case study: Working from home - balance

	Treatment mean	Control mean	Std.dev.	p-value of test of equal means
Number of observations	131	118	249	
Prior performance z-score	-0.03	-0.04	0.58	0.87
Age	24	24	4	0.85
Male	0.47	0.47	0.50	0.99
Secondary technical school	0.46	0.47	0.50	0.80
High school	0.18	0.14	0.36	0.38
Tertiary	0.35	0.36	0.48	0.94
University	0.02	0.03	0.15	0.91
Prior experience (months)	19	17	26	0.48
Tenure (months)	26	28	22	0.45
Married	0.22	0.32	0.44	0.07
Children	0.11	0.24	0.38	0.01
Age of youngest child	4.60	3.00	3.35	0.14
Rent apartment	0.24	0.20	0.42	0.44
Cost of commute (yuan)	7.89	8.34	6.96	0.63
Own bedroom	0.99	1.00	0.06	0.13
Internet	0.97	0.99	0.14	0.00
Base wage (yuan monthly)	1540	1563	161	0.23
Bonus (yuan monthly)	1031	1093	625	0.43

Case study: Working from home: balance

- A few cases p-values below 10% - (e.g. having internet access) no big deal
- Magnitudes of the differences - small.
- Most importantly, the two groups are virtually identical in terms of their performance score measured before the experiment.
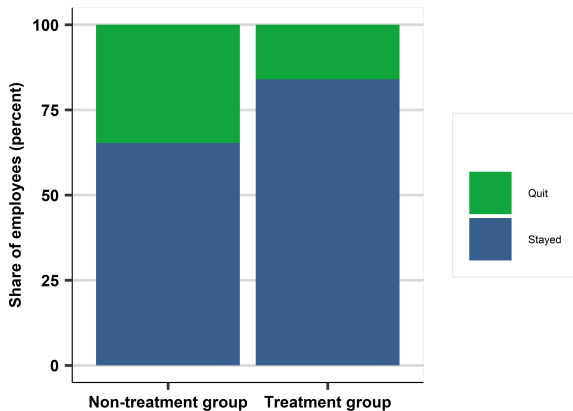- But there are some larger differences: married, have children. Need to pay attention.

Case study: Working from home: compliance

▶ A small fraction (less than 20%) of employees in the treatment group had to be re-assigned to work in the office because of changes in their circumstances (failure to establish internet connection, loss of private work space, etc.).

▶ All employees assigned to the non-treatment group worked from the office through the duration of the experiment.

▶ Compliance in this experiment was imperfect.

▶ We can get the average intent-to-treat effect (AITTE).

▶ The degree of noncompliance was small.

▶ Thus AITTE is likely to be just a bit smaller than ATE.

Case study: Working from home

- ▶ Two outcomes: worker retention and worker performance.
- ▶ 16% in the treatment group quit, compared to 35% in the non-treatment group. The difference is 19 percentage points.
- ▶ Measure the performance of these 134 "order-takers" by the number of telephone calls they take.
- ▶ Number of phone calls taken, measured in thousands. The mean in the treatment group is 14, the mean in the non-treatment group is 10. The difference is 4 thousand.

Case study: Employee retention and quit rates in treatment and control group

Share of employees (percent) chart showing Non-treatment group and Treatment group. Legend: Quit (green), Stayed (blue).

- ▶ Employee retention and quit rates in the treatment group (working from home) and the non-treatment group (working from the office).
  - ▶ Stayed: remained employed for eight months;
  - ▶ quit: quit the company within eight months
- ▶ Source: working-from-home dataset. N=249

Case study: Working from home - effects

VARIABLES	(1) Quit job	(2) Phone calls (thousand)
Treatment group	-0.19**	4.04**
	(0.055)	(0.99)
Constant	0.35**	10.06**
	(0.044)	(0.75)
Observations	249	134
R-squared	0.047	0.113

Source: `working-from-home` dataset.

Case study: Working from home - Effects extended

VARIABLES	(1) Quit job	(2) Phone calls (thousand)
Treatment group	-0.19**	4.06**
	(0.056)	(0.96)
Married	-0.13	-5.44*
	(0.074)	(2.17)
Children	0.11	3.87
	(0.097)	(2.41)
Internet at home	0.18**	
	(0.036)	
Constant	0.19**	10.65**
	(0.056)	(0.76)
Observations	249	134
R-squared	0.055	0.168

Source: working-from-home

Case study: Working from home: internal validity

▶ Based on all the information, we can judge internal validity well.

▶ Assignment was random.

▶ Compliance was imperfect, but only in the treatment group, and even here more than 80% of the subjects complied with the treatment.

▶ Spillovers are unlikely to be important in this experiment

Case study: Working from home: external validity

- ▶ Based on all the information, we can somewhat judge external validity .
- ▶ It had an actual impact, management changed practices
- ▶ Would it work for other employees? Yes for those who are like the ones in the experiment = applied
- ▶ Not necessary for all
- ▶ What can we say about other companies?

How to do an experiment? Practical advice.

Experiments in practice

- ▶ Design can be complicated
    - ▶ Multiple intervention "arms" (beyond 0,1)
    - ▶ "placebo" interventions
    - ▶ intensity of treatment, intention to treat versus actual treatment
    - ▶ Etc.
- ▶ Experiments need careful planning
    - ▶ Expertise needed
    - ▶ Results uninformative or misleading if poorly designed
- ▶ But analysis is relatively simple
    - ▶ Easy to carry out
    - ▶ Easy to communicate
- ▶ Practical problems = threats to internal validity

Spillover effects

▶ Intervention has spillover (or external) effect: its assignment to individual i affects some other individual j.

▶ As a result of spillovers the overall effect of the intervention may be different from the sum of its effects on treated individuals.

▶ In the presence of spillover effects analyzing the effects of interventions is more complicated.

▶ Sometimes it's easy to see potential spillover, in other cases not
  ▶ Examples?

Issues: Hawthorne, Placebo, John Henry

- ▶ Hawthorne effect
  - ▶ Both treated and non-treated individuals change behavior due to being observed
  - ▶ Are workers would become more productive in higher or lower levels of light - in the Western Electrics Hawthorne factory
    - ▶ Later proved that the effect was not there at all

- ▶ Placebo effect
  - ▶ Treated individuals change behavior due to being in treated (and not the treatment effect)
    - ▶ Important in many medical interventions
    - ▶ Placebo treatment arm - in medical experiments: both treated and comparison individuals receive pill - Only comparison individuals receive placebo

- ▶ John Henry effect
  - ▶ Non-treated individuals change behavior due to being in comparison group
    - ▶ E.g., they work harder
    - ▶ John Henry, a legendary US steel driver in the 1870s. When heard his output was being compared with that of a steam drill, worked so hard to outperform the machine.

Issues: Compliance (recap)

▶ Non-compliance
  ▶ Not all units assigned to treatment actually finish treatment
  ▶ Some units assigned to non-treatment end up being treated
▶ Intent-to-treat effect
  ▶ Comparing those assigned to treatment and those assigned to non-treatment
    ▶ Average intent-to-treat effects smaller than average treatment effects
▶ Cannot estimate treatment effect by comparing treated and non-treated
  ▶ Even if assignment random
  ▶ Actual treatment status affected by selection

Issues: Efficiency of Randomization

▶ Benchmark: simple randomization
  ▶ Simple random draw of treatment indicator for each unit
▶ Cluster randomization
  ▶ Clusters of units selected
    ▶ E.g., villages
  ▶ Lose efficiency / save on costs
  ▶ May minimize spillovers
▶ Stratified randomization
  ▶ First create strata, randomize treatment within strata
    ▶ E.g., within same geographic unit, income category,
  ▶ Gain efficiency / same cost

Experimenting in Business

- ▶ Some firms do it
  - ▶ Most don't
  - ▶ Even though potentials are great
- ▶ Internet firms experiment a lot
  - ▶ randomly assign features of their service to different customers
  - ▶ A/B testing to study the effects of those features
- ▶ Few firms experiment with promotion activities
  - ▶ To see what increases sales most at lowest cost
- ▶ Even fewer firms experiment with incentive structures
  - ▶ To study what makes employees perform better

Number of Subjects and Proportion Treated, Power

▶ Interested to run an experiment that can prove a difference or the lack of it.

▶ Need to design the experiment and make decisions.

▶ Two decisions that are relevant for all experiments:
  ▶ (1) the number of subjects
  ▶ (2) the proportion of subjects assigned to treatment.

▶ Affect the precision of the effect estimate.
  ▶ More precision = smaller SE and a narrower CI.
  ▶ More likely to detect a nonzero ATE by hypothesis testing.

Number of Subjects and Proportion Treated

▶ Testing = see if there is enough evidence in the data to reject a null hypothesis = a zero ATE. (Chapter 06)

▶ Decision to reject the null = avoid both a false positive and a false negative
  ▶ In the general pattern represented by our data

▶ The likelihood of avoiding a false positive = level of significance;

▶ the likelihood of avoiding the false negative = level of power.

▶ A more precise effect estimate helps avoid both the false positive and the false negative

Number of Subjects and Proportion Treated

▶ To determine the precision we need, we fix a level of significance, (say 5%) and we want as high a power as possible.
  ▶ Often 80% – not straightforward to interpret
▶ Low level of significance and a high power both require as precise an effect estimate as possible

Number of Subjects and Proportion Treated

► How many subjects to include in the experiment, needs more work, and balance:
  ► precision of the effect estimate
  ► costs of carrying out an experiment.

► Power calculation - formula with some assumptions we make

► Binary outcome - need to assume the proportions in the two groups.

► Quantitative outcome, need to assume the difference in the means in the two groups and the standard deviation among all subjects

Number of Subjects and Proportion Treated

▶ It is a hypothesis testing setup
  ▶ Mean outcome in the treated group be m_1,
  ▶ Mean outcome in the untreated group be m_0.
  ▶ Standard deviation of the outcome variable is σ.

$$H_0 : m_1 - m_0 = 0 \tag{3}$$
$$H_A : m_1 - m_0 \neq 0 \tag{4}$$

▶ Pick a 5% level of significance, and power at 80%

▶ Set the number of treated and untreated observations be same: $n_0 = n_1$. Total number of observations is $n = n_0 + n_1$

Number of Subjects and Proportion Treated

▶ We need some values from the normal distribution
  ▶ 1.96 corresponds to 5% level of significance we set
  ▶ 0.84 corresponds to the 80% level of power we set
▶ Standard deviation of the outcome variable is σ.
▶ Formula is

$$n = 4\sigma^2 \times \left(\frac{1.96 + 0.84}{m_1 - m_0}\right)^2 \tag{5}$$

▶ Can be used with any number of significance and power
▶ Has more complicated versions depending on distributional assumptions.

Number of Subjects and Proportion Treated

- ▶ Binary outcome, average=proportion.
    - ▶ Proportion of 1 among treated: r_1,
    - ▶ untreated observations r_0.
- ▶ The simplified test:

$$H_0 : r_1 - r_0 = 0 \tag{6}$$
$$H_A : r_1 - r_0 \neq 0 \tag{7}$$

- ▶ The overall proportion of 1s is $r = r_1/2 + 1_0/2$
- ▶ $\sigma = \sqrt{r(1-r)}$.

$$n = 4r(1-r) \left(\frac{1.96 + 0.84}{r_1 - r_0} \right)^2 \tag{8}$$

Case Study: Fine Tuning Social Media Advertising

- ▶ A/B testing: Social media site, two versions of the ad
- ▶ A: CEU MS in Business Analytics is Number 1 in Central Europe
- ▶ B: CEU MS in Business Analytics is Top 50 in the World

Case Study: Fine Tuning Social Media Advertising

- Stat: $p = 0.05$ (95% CI) $\alpha = 0.8$ for power (i.e., 20% chance of a false negative)

- Good case assumptions. :
  - Click through rate is 1%. Show ad 100 people, to get 1 click.
  - Conversion rate is 10%. Need 10 people to get an action (lead).
  - = Need 1000 impressions (show ad 1000 times) to get 1 action.
  - Assume one of the ads will be 30% higher conversion rate.

- Use power calculator, this means 400 thousand impressions needed.

- A worse case assumption
  - 20% difference, and 2000 impressions for one lead, so 0.0005 vs 0.0006
  - We would need 1.7 million impressions - appr 2000 USD.

Case Study: Fine Tuning Social Media Advertising

▶ Reality: Ads were shown to 660,000 people, total impressions = 2 million).

	Cost (dollars)	Target no. of impressions	No. of clicks	No. of actions	Cost per action (dollars)
A	1000	1 million	3323	32	31.25
B	1000	1 million	3128	21	47.62
Percent difference			6.23%	52.4%	
p-value			0.015	0.131	

Source: Summary data from the social media campaign.

Case Study: Fine Tuning Social Media Advertising

- ▶ 2 million impressions, 6461 unique visitors on the landing site
- ▶ 53 of them acted on it (ie gave some info)

- ▶ In terms of clicks, there is a small difference of 6.2% in favor of version A.
  - ▶ This difference is statistically significant at the 5% level
- ▶ The difference between the number of actions is large in relative terms – 52% in favor of version A.
  - ▶ = implied costs per action are lower for version A (47.62 vs 32.25.)
  - ▶ The number of actions is very small.
  - ▶ The difference not statistically significant = cannot reject the hypothesis that the two numbers are equal

Case Study: Fine Tuning Social Media Advertising

▶ We tested two messages, small difference. No1 in CEE wins in this region.
▶ Take-home message is that
  ▶ If starting now, pick "A"
  ▶ If already doing "B", we do not strong evidence to change to "A"

▶ Why this uncertainty?

Validity

- ▶ Internal validity
  - ▶ Whether RCT identifies causal effect of treatment
  - ▶ Here and now, in the actual context
  - ▶ If yes we say RCT has high internal validity
- ▶ External validity
  - ▶ The extent to which results of RCT generalize to other situations
  - ▶ Similar treatments not here or not now
  - ▶ If yes we say RCT has high external validity

- ▶ Sometimes trade-off
  - ▶ High internal validity more likely achieved in controlled situation
- ▶ Yet hard to generalize if low internal validity

Summary – controlled experiments

▶ Controlled experiments assign different values of x to observations in ways to
  ▶ avoid selection and reverse causality,
  ▶ control other aspects of the situation
  ▶ avoid other confounders.

▶ With binary treatment assignment decides
  ▶ which subjects in the experiment receive the treatment (the treatment group)
  ▶ and which don't (the control group).

▶ Randomization
  ▶ the most widely used assignment in business, policy and health applications
  ▶ use of a rule that is known to be independent of all aspects of selection and other influences on outcomes.