Class 12 A/B Testing

Dr. Wei Miao

UCL School of Management

November 5, 2025

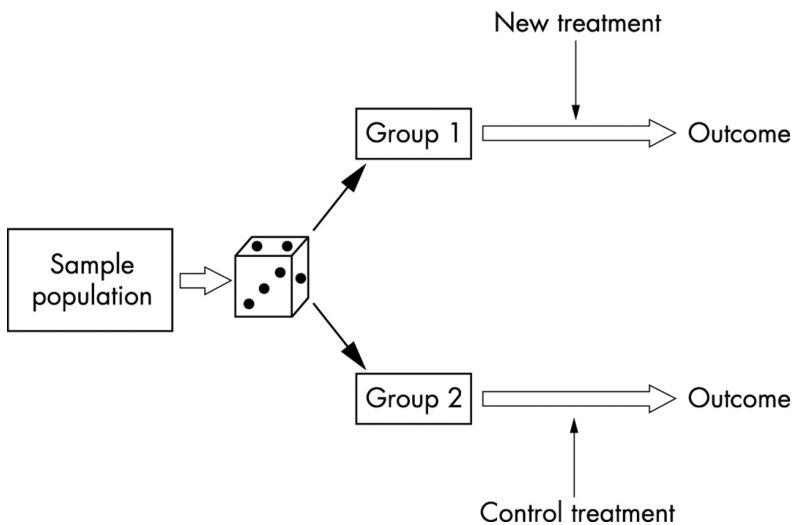
Section 1

Randomised Controlled Trials

Randomised Controlled Trials

Randomised Controlled Trials

A **randomised controlled trial** (RCT) is an experimental form of impact evaluation in which the population receiving the intervention is chosen *at random* from the eligible population, and a control group is also chosen *at random* from the same eligible population.



Types of RCTs: Based on Location

	Lab Experiment	Field Experiment
Location	In a controlled, laboratory environment	In the field
Internal validity	High	Low
External validity	Low (Hawthorne effect)	High

- Internal validity refers to the extent to which the experiment is free from selection bias.
- External validity refers to the extent to which the results can be generalised to the real world or other contexts.

Types of RCTs: Based on Treatment Design

- A/B testing (treatment group + control group)
  1. Loyalty programme
  2. No loyalty programme
- A/B/N testing (multiple treatment groups + control group)
  1. Point-based loyalty programme; points can be redeemed for price vouchers
  2. Point-based loyalty programme; points can be redeemed for gifts
  3. Point-based loyalty programme; points can be redeemed for free top ups
  4. No loyalty programme

Section 2

Procedures of A/B Testing

Motivating Example of Tom's Loyalty Programme

- Should we introduce a loyalty programme for our customers?
  - Cons: increased costs due to free drinks
  - Pros: it may increase spending and retention rate, and hence future CLV
- How to estimate the causal effect of introducing a loyalty programme on customer spending and retention?

Step 1: Design Experiment Conditions and Unit of Randomization

We decide **the granularity of the randomisation unit** based on the research question at hand.

- **individual**
- household
- store
- other levels more granular (e.g., device level) or even less granular (e.g., city level)

Potential Issues of Granularity

- **Crossover Effects**: A crossover occurs when an individual who was supposed to be assigned to one treatment is accidentally exposed to another or more treatments.
  - e.g., For online A/B testing, a notorious crossover effect is that when browsers reset the cookies, the same individual customer may be treated as a different new customer.
- **Spillover effects**: The behaviour of the treatment group can also affect the control group.
  - e.g., customers may share the promotions with family members and friends.

Question: How should we mitigate spillover and crossover effects?

Step 1: Decide on the Unit of Randomisation

Proposal 1: Randomise the treatment based on West London and East London.

- Do you expect this "randomisation" to be a true randomisation?[1]

Proposal 2: Randomise the treatment among individual customers.

- Is this a true randomisation?
- What problems might we still have?

[1] All answers to questions in the slides are on the webpage version of the slides.

Step 1: Pros and Cons of Granularity

Disadvantages of granularity:

- Costs and logistics
- Spillovers and crossovers

Advantages of granularity:

- Increase the chance of successful randomisation, thereby mitigating any systematic imbalance of covariates before the experiment.

Exercise:

- If we would like to randomise prices, how can we randomise individualised price discounts to customers?

Step 2: Decide on the Randomisation Allocation Scheme

- Individuals (or the relevant unit of randomisation) are allocated at random into a treatment condition based on some decision rules.
- Due to the high costs and potential risks of A/B testing, we often select a small percentage of customers into the treatment condition, while the remaining customer should do "business-as-usual".

Step 3: Decide on Sample Selection and Treatment Duration

- Any field experiment should be aware of the need for a sufficiently large sample size, or sufficient statistical power.
  - The larger sample size, the higher statistical power for the experiment; meanwhile, larger sample size brings higher costs and risks.
- Run a power calculation in R, e.g., using `pwr.t.test()` in the `pwr` package. We need to input the following:
  - the expected effect size (how big of a difference the treatment is expected to make, typically from previous studies or pilot studies)
  - significance level (the probability of a false positive, usually 0.05)
  - power level (the probability of detecting a true effect, usually 0.8)
  - the type of test (one-tailed or two-tailed, depending on the hypothesis)

Step 4: Collect Data

We need to collect data after the experiment is launched. We at least need to

- Collect data on the outcome variables of interest
- Collect consumer characteristics data for the purpose of the randomisation check and heterogeneity analysis.

Proposal: We need to collect customers' spending data and link the data with their treatment assignment.

Step 5: Interpreting Results from a Field Experiment

Step 5.1: Randomisation check

- We need to check if the treatment and control groups are well-balanced in terms of their **pre-treatment** characteristics, especially the outcome variables.

Step 5.2: Analyse the data and estimate the ATE

- **t-test** to examine the difference in the average outcome between the treatment and control groups. In R, we can use `t.test()`
- **Regression analysis** when analysing A/B/N testing or multivariate experiments.

After-Class Readings

- (optional) Varian, Hal R. 'Causal Inference in Economics and Marketing'. Proceedings of the National Academy of Sciences 113, no. 27 (5 July 2016): 7310–15.