A Unified Approach to AB Testing and Campaign Evaluation

Tamás Barczikay

August 1, 2024

Goals

How does a common statistical testing approach helps us?

- It makes different results **comparable**.
- **Saves time**, since the procedure is coded and focuses on credit spending/purchase, only minor modifications need to be done. Results can be obtained within days and not weeks.
- We will see the **true average effect** of a change in the product on members' behavior, separated from the additional effect of confounding variables (like their salary which we have no information on).
- We include **statistical tests** to verify if a change or new feature in the product made any difference, or if it is only due to chance in the sampling process.
- **Note**: this might hurt many times. On the **positive side**, if we know that one of our approach doesn't work, we will not face increasing production costs investing in them.

Different Cases Need Different Approaches

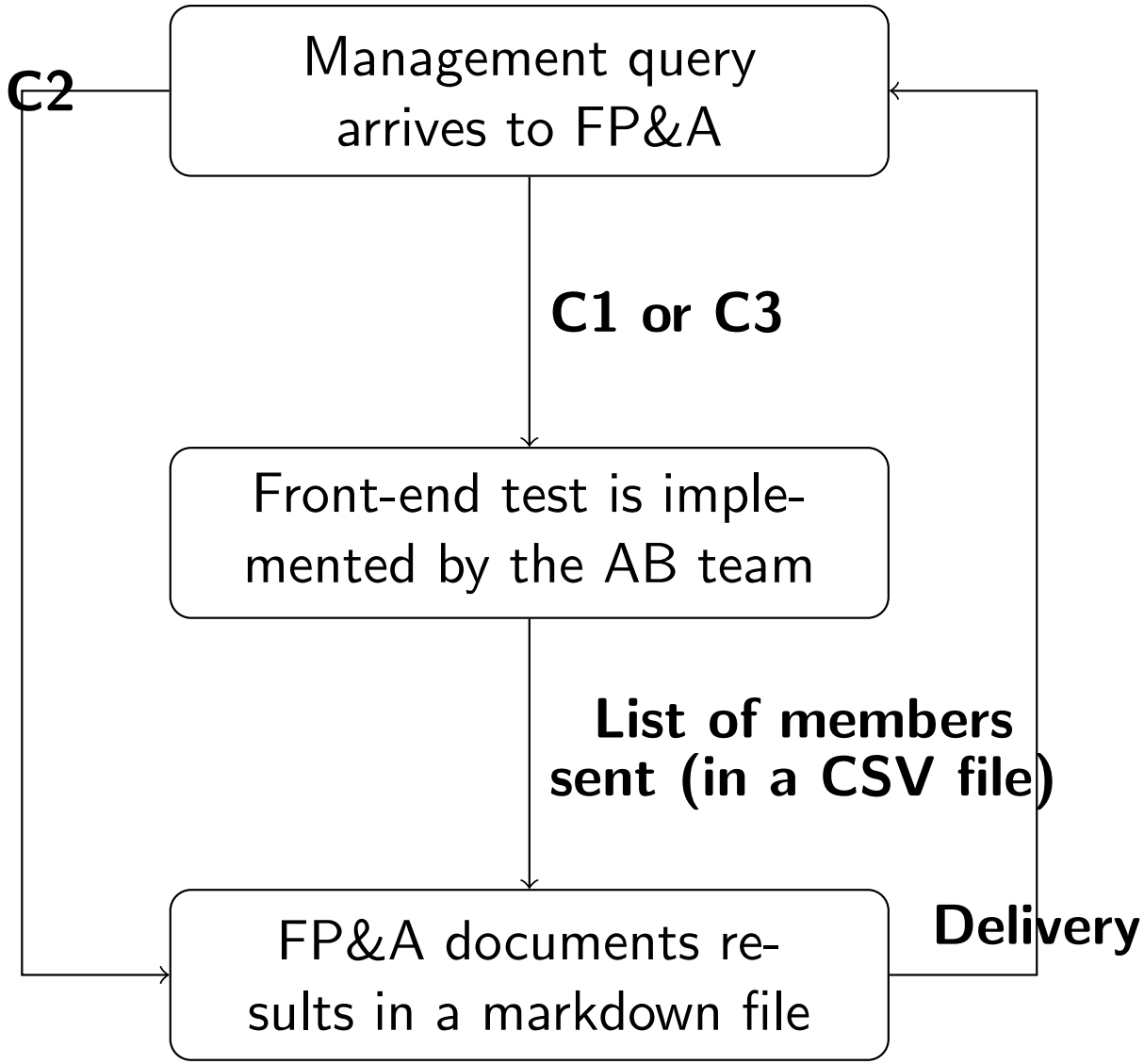
- **AB testing on new members (C1)**: during an AB test we modify the rules of the game for specific **new members**, meaning they have different conditions than the rest of the new members. They **cannot decide** to opt-in since they are not aware that others are treated differently.
- **Camping evaluation on ALL existing members (C2):** a new seasonal campaign is being launched and every member can opt-in by buying discounted credits, meaning they **can decide**.
- **AB testing on SELECTED members (C3)**: some of the existing members are randomly chosen - together with a control group of comparable size - and given the right to opt-in in a new discount type or feature **if they decide**.

Criteria	C1 - New members	C2 - All members	C3 - Selected members
Before/after data avbl.	No	Yes	Yes
Can opt-in	No	Yes	Yes
Randomized	Yes	No	Yes
Example	Cheaper credit price in a minor country.	Seasonal campaign (Valentines day, Christmas).	New vibration type OR cheaper credit price in a major country for a month using wheel of fortune.
FE/RE	Not possible	Must be used	Can be used

Two pieces of info are important. Whether we can randomize (C1 & C2) and if we have data from before the test (C2 & C3).

All statistical models run on the data should be dependent on these conditions, due to decision on opting-in!

Workflow


Management query arrives to FP&A
C2
C1 or C3
Front-end test is implemented by the AB team
List of members sent (in a CSV file)
FP&A documents results in a markdown file
Delivery

Tests Used During AB testing

Aspect	Pairwise t-test	Mann-Whitney U Test	Chi-Square Test
Type of Data	Continuous (parametric)	Ordinal/Cont. (non-parametric)	Categorical
Assumptions	Independence, normal distribution, homogeneity of variance	Independence, similar shape distributions	Independence
Purpose	Compare means	Compare two samples	Test relationships
Examples	Drug treatments	Satisfaction scores	Gender vs. preference

The Selection Problem and Causal Inference

Suppose we are interested in the causal question: "Do hospitals make people healthier?"

Group	Sample Size	Mean Health Status	Std. Error
Hospital	7774	2.79	0.014
No Hospital	90049	2.07	0.003

Table: Comparison of Health Status by Hospitalization

The difference in means is 0.71, suggesting that hospitalization correlates with worse health (higher number means worse health in the table).

However, people who go to the hospital are likely less healthy initially.

Selection Problem is solved by Randomization

To formalize this, let $d_i \in \{0, 1\}$ represent hospital treatment, and y_i denote the health status. The potential outcomes framework defines:

$$y_i = \begin{cases} y_{1i} & \text{if } d_i = 1 \\ y_{0i} & \text{if } d_i = 0 \end{cases}$$

The observed outcome can be written in terms of potential outcomes:

$$y_i = y_{0i} + (y_{1i} - y_{0i})d_i$$

The causal effect of hospitalization for individual i is $y_{1i} - y_{0i}$.

We can link the comparison of average health conditional on hospitalization status to the average causal effect:

$$\mathbb{E}[y_i \mid d_i = 1] - \mathbb{E}[y_i \mid d_i = 0] = \underbrace{\mathbb{E}[y_{1i} \mid d_i = 1] - \mathbb{E}[y_{0i} \mid d_i = 1]}_{\text{Average treatment effect on the treated}} + \underbrace{\mathbb{E}[y_{0i} \mid d_i = 1] - \mathbb{E}[y_{0i} \mid d_i = 0]}_{\text{Selection bias}}$$

In the selection bias term **randomization** grants us the **independence** of y_{0i} and d_i allows us to swap $\mathbb{E}[y_{0i} \mid d_i = 1]$ for $\mathbb{E}[y_{0i} \mid d_i = 0]$, which means the selection bias term is cancelled out and we have the the average treatment effect.

$$\mathbb{E}[y_i \mid d_i = 1] - \mathbb{E}[y_i \mid d_i = 0] = \underbrace{\mathbb{E}[y_{1i} - y_{0i} \mid d_i = 1]}_{\text{Average treatment effect on the treated}} + \underbrace{\mathbb{E}[y_{0i} \mid d_i = 0] - \mathbb{E}[y_{0i} \mid d_i = 0]}_{\text{Selection bias} = 0}$$

Example of an Aggregated Table for C1

mbr_id	crd_spent	AB
1	150	0
2	245	1
3	320	1
4	395	0

- We will automatically aggregate our transaction level data to the format in the table, and join the names to it from the received CSV file.
- **crd_spent** is our target variable and **AB** is the so called treatment indicator.
- Member 2 and 3 are randomly selected **controls**.
- Since we do not have other relevant covariates, we must implement the test using only the AB dummy indicator.

Implementation

The theoretical equation is the following:

$$\text{crd\_spent}_i = \beta_0 + \beta_1 \text{AB}_i + \epsilon_i$$

Where:

- crd_spent_i is the outcome variable.
- AB_i is the treatment indicator.
- β_0 is the intercept (i.e. the mean spending of those who were in the control group).
- β_1 is the treatment effect (i.e. the additional average effect of those in the treatment group).
- ϵ_i is the error term (containing all effects that we did not account for).

The estimated equation is as follows:

$$\text{crd\_spent}_i = 272.5 + 10.0 \times \text{AB}_i + \epsilon_i$$

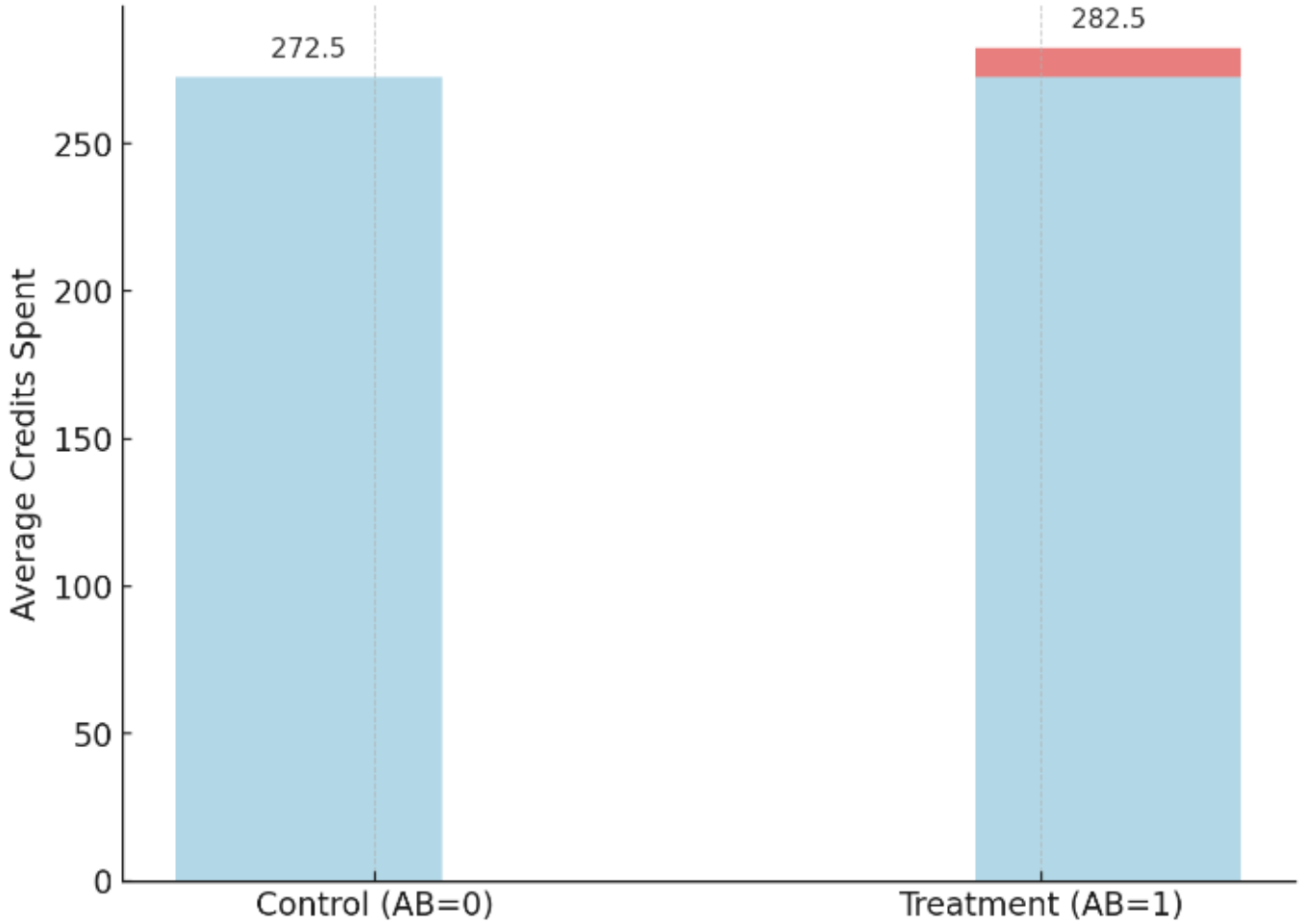


Figure: Effect of the feature change

Example of an Aggregated Table for C2

mbr_id	crd_spent	during_test
1	100	0
1	150	1
2	200	0
2	230	1
3	310	0
3	320	1
4	400	0
4	395	1

- This is technically **not** an AB test, since it doesn't include randomization!
- In the context of C2, we are evaluating the impact of a **new seasonal campaign** on all existing members, where each member has the option to opt-in by purchasing discounted credits. This scenario involves the entire member base and allows for self-selection, meaning we cannot get rid of the **selection bias** term so easily!

Estimated Regression Model for C2

Using the example data, the estimated parameters are:

$$\text{crd\_spent}_i = 252.50 + 21.25 \times \text{during\_test}_i + \epsilon_i$$

This estimated "treatment effect" would mean that b/c of the campaign on average 21.25 more credits were spent by customers.

- **Selection Bias**: since members chose to opt into the campaign, this self-selection can introduce biases because the characteristics of those who opt-in may **systematically differ** from those who do not.
- **Confounding Variables**: without randomization, other variables that influence credit spending might also correlate with the decision to participate in the campaign. These confounders can skew the results, making it **difficult to isolate** the effect of the campaign itself.
- **Internal Validity**: without randomization any observed differences in outcomes could be due to **pre-existing differences** between the groups rather than the treatment.

Example of an Aggregated Table for C2 with Salary

mbr_id	crd_spent	during_test	salary
1	100	0	2900
1	150	1	2900
2	200	0	3950
2	230	1	3950
3	310	0	5050
3	320	1	5050
4	400	0	5900
4	395	1	5900

$$\text{crd\_spent}_i = -188.50 + 17.50 \times \text{during\_test}_i + 0.098 \times \text{salary}_i + \epsilon_i$$

1. Parameter value **decreases** on the "treatment variable".
2. Notice that salary doesn't change in the **short-run**!
3. New var correlated with dependent - should be in the model.
4. Assumption 1: during test **correlates** with salary - use **FE**.
5. Assumption 2: during test **does not correlate** with salary - use **RE**.

How to Treat the Latent Salary's Effect?

Fixed Effects (FE) Transformation:

$$\begin{aligned} \text{crd\_spent}_{it} - \overline{\text{crd\_spent}_i} &= \beta_1(\text{during\_test}_{it} - \overline{\text{during\_test}_i}) \\ &+ (\epsilon_{it} - \bar{\epsilon}_i) \\ &+ (u_i - \overline{u_i}) \longrightarrow \text{salary}_i - \overline{\text{salary}}_i = 0 \end{aligned}$$

Random Effects (RE) Transformation:

$$\begin{aligned} \text{crd\_spent}_{it} - \theta\overline{\text{crd\_spent}_i} &= \beta_1(\text{during\_test}_{it} - \theta\overline{\text{during\_test}_i}) \\ &+ (\epsilon_{it} - \theta\bar{\epsilon}_i) \\ &+ (u_i - \theta\overline{u_i}) \longrightarrow \text{salary}_i - \theta\overline{\text{salary}}_i \end{aligned}$$

$$\theta = 1 - \sqrt{\frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + T\sigma_u^2}}$$

- σ_ϵ^2: Var of the idiosyncratic error term (ϵ_{it}).
- σ_u^2: Var of the random effects (u_i).
- T: Time periods.

Calculating Necessary Values for FE and RE

Member 1:

$\overline{\text{crd\_spent}_i} = \frac{100+150}{2} = 125$

$\overline{\text{salary}_i} = \frac{2900+2900}{2} = 2900$

Member 2:

$\overline{\text{crd\_spent}_i} = \frac{200+230}{2} = 215$

$\overline{\text{salary}_i} = \frac{3950+3950}{2} = 3950$

Member 3:

$\overline{\text{crd\_spent}_i} = \frac{310+320}{2} = 315$

$\overline{\text{salary}_i} = \frac{5050+5050}{2} = 5050$

Member 4:

$\overline{\text{crd\_spent}_i} = \frac{400+395}{2} = 397.5$

$\overline{\text{salary}_i} = \frac{5900+5900}{2} = 5900$

$$\overline{\text{during\_test}_i} = \frac{0+1}{2} = 0.5$$

From the table, we have $\sigma_\epsilon^2 = 122.77$, $\sigma_u^2 = 0$, and $T = 2$.

$$\theta = 1 - \sqrt{\frac{\sigma_\epsilon^2}{\sigma_\epsilon^2 + T\sigma_u^2}} = 1 - \sqrt{\frac{122.77}{122.77 + 2 \times 0}} = 1 - \sqrt{1} = 1 - 1 = 0$$

So, θ is 0, indicating that there is no between-individual variation in the random effects model (i.e., salary needs a minimal variation).

Transformation of the Input Values

mbr_id	**crd_spent** (DM)	**during_test** (DM)	**salary** (DM)
1	100 - 125 = -25	0 - 0.5 = -0.5	2900 - 2900 = 0
1	150 - 125 = 25	1 - 0.5 = 0.5	2900 - 2900 = 0
2	200 - 215 = -15	0 - 0.5 = -0.5	3950 - 3950 = 0
2	230 - 215 = 15	1 - 0.5 = 0.5	3950 - 3950 = 0
3	310 - 315 = -5	0 - 0.5 = -0.5	5050 - 5050 = 0
3	320 - 315 = 5	1 - 0.5 = 0.5	5050 - 5050 = 0
4	400 - 397.5 = 2.5	0 - 0.5 = -0.5	5900 - 5900 = 0
4	395 - 397.5 = -2.5	1 - 0.5 = 0.5	5900 - 5900 = 0

Using these values we can estimate the real effect of the campaign.

$$\text{crd\_spent}_i = 8.88 \times 10^{-16} + 21.25 \times \text{during\_test}_i + \epsilon_i$$

Example of an Aggregated Table for C3

mbr_id	crd_spent	during_test	AB
1	100	0	1
1	150	1	1
2	200	0	1
2	230	1	1
3	310	0	0
3	320	1	0
4	400	0	0
4	395	1	0

- **crd_spent** is our target variable and **AB** is the so called treatment indicator.
- **during_test** shows whether the aggregation was made during the 28 days of the experiment or from the previous 28 days.
- Member 3 and 4 are randomly selected **controls**.

Measuring Treatment Effect

The regression model used to determine the treatment effect in the A/B test is specified as follows:

$$\text{crd\_spent}_{it} = \beta_0 + \beta_1 \text{AB}_{it} + \beta_2 \text{during\_test}_{it} + \beta_3 (\text{AB}_{it} \times \text{during\_test}_{it}) + \epsilon_{it}$$

The **treatment effect** is captured by the coefficient β_3 of the interaction term ($\text{AB}_{it} \times \text{during\_test}_{it}$). This coefficient indicates how the treatment (AB) affects credit spent (crd_spent) during the test period compared to the pre-test period.

These parameters can be estimated which results in the following equation:

$$\begin{aligned} \text{crd\_spent}_{it} = 355.00 - 205.00 \cdot \text{AB}_{it} + 2.50 \cdot \text{during\_test}_{it} \\ + 37.50 \cdot (\text{AB}_{it} \times \text{during\_test}_{it}) + \epsilon_{it} \end{aligned}$$

	Control	Treated
Before	$(310 + 400)/2 = 355.0$	$(100 + 200)/2 = 150.0$
After	$(320 + 395)/2 = 357.5$	$(150 + 230)/2 = 190.0$

Table: Average Spending Calculations

	Control	Treated
Before	$\beta_0 = 355.00$	$\beta_0 + \beta_1$ $= 355.00 - 205.00$ $= 150.00$
After	$\beta_0 + \beta_2$ $= 355.00 + 2.50$ $= 357.50$	$\beta_0 + \beta_1 + \beta_2 + \beta_3$ $= 355.00 - 205.00 + 2.50 + 37.50$ $= 190.00$

Table: Parameter-based Calculations

Why not only compare averages?

- For the **AB tested group**, the average credit spent increased from 150.0 before the treatment to 190.0 after the treatment.
- For the **control group**, the average credit spent slightly increased from 355.0 before the treatment to 357.5 after the treatment.
- These averages, however, do not properly account for both the time effect and the treatment effect simultaneously.
- Although in this case, the difference in average spending for the control group was only **2.75**, it highlights the potential for much larger discrepancies in other scenarios.
- Not accounting for the separate effects we can even get a result that shows that the treatment had **negative effect!**

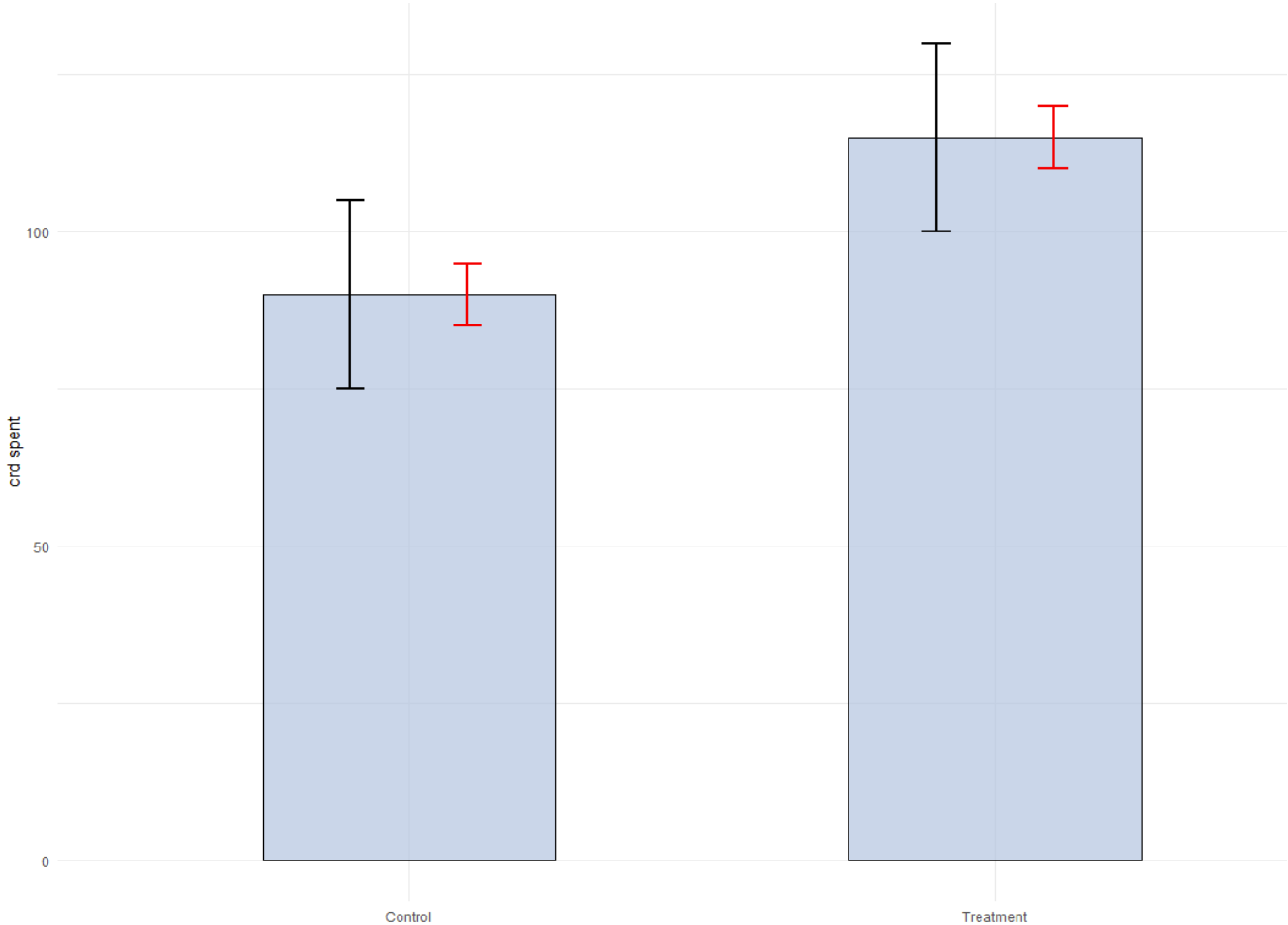
Time Invariant Effects Can Also Be Treated

id	crd_spent	drg_test	AB	crd_pur	salary	edu	fetish
1	100	0	1	10	50000	12	1
1	150	1	1	15	50000	12	1
2	200	0	1	20	60000	16	0
2	230	1	1	25	60000	16	0
3	310	0	0	30	70000	17	1
3	320	1	0	35	70000	17	1
4	400	0	0	40	80000	16	0
4	395	1	0	45	80000	16	0

- We can include new explanatory variables to refine our estimates like **crd_pur**, since we have data.
- And it is also possible to filter out the effect of time invariant variables with **FE/RE**.
- FE will average out the effect of AB, but the interaction term still shows the treatment effect.

Why do we need robust standard errors?

- Randomization eliminates selection bias but does not address **heteroscedasticity** (due to outliers for example).
- Implementing **robust standard errors** ensures reliable and valid statistical inferences about treatment effects.



Comparison of Standard Errors and p-values

mbr_id	crd_spent	crd_spent_outlier	AB
1	100	100	0
2	160	160	1
3	180	180	1
4	140	140	0
5	200	5000	1

Table: Data for AB Test with and without Outlier

Method	Data	Standard Error	p-value
Standard t-test	No Outlier	24.49	0.056
Standard t-test	With Outlier	1264.13	0.247
Robust SE	No Outlier	20.00	0.048
Robust SE	With Outlier	900.00	0.037

Table: Comparison of Standard Errors and p-values

Feedback 1

- **Borbála Fehér**:
  - In C2 cases (before vs after calculations), unknown variables strongly influence the dependent variable, affecting the observable differences between Before and After.
  - For example, simultaneous site changes or spring breaks may overlap differently for members in the before and after periods.
  - It's essential to filter out the effects of these unknown variables with differing values in the before and after periods.
- **Noémi Gönczi**:
  - The main question for campaigns is how bonuses/discounts affect subsequent member behavior and purchases, impacting campaign success.
  - What is gained from the campaign considering that bonuses might delay the next purchase?
  - Would the member have purchased without the discount?

Feedback 2

- **Balázs Klujber**:
  - Types of random and directed sampling (brief summary + representativity: `https://valstat.ktk.pte.hu/mintavetel.html`)
  - Representativity of samples (spending habits of Indian users cannot be directly applied to American users)
  - We do not use certain latent variables because previous tests showed negligible or no effect (However, we are unaware of which variables these are).
- **Gábor Vecsei**:
  - Netflix's AB testing should be checked as they have solutions for the length of data collection of which we do not.
  - Manage to have a meeting with the AB team and simplify the presentation's content.