MINI PROJECT

A/B Testing Application in Data Science

Controlled Experiment on Landing Page Optimization

Trần Tiến Đạt – 22110039
Nguyễn Thị Ngọc Anh – 23280037
Nguyễn Thái Hưng Thịnh – 23110209

Môn: Xử lý số liệu thống kê

Khoa Toán - Tin, VNUHCM-US

November 14, 2025

Nội Dung

Introduction to A/B Testing

A/B Testing, also known as Split Testing, is a research methodology where two versions of a variable (A and B) are compared simultaneously to determine which one performs better against a defined goal.

Role in Data Science

A/B testing is a foundational technique in data science that engineering and product teams use to validate decisions with real-world data. At its core, it's about understanding what changes improve user experience, conversion, or retention.

A/B Testing: Bridging UX and Data

- **Scientific Validation:** A/B Testing is a Randomized Controlled Experiment validating user experiences design changes based on objective evidence, not subjective opinion.
- **Core Function:** It measures the isolated impact of a single change (e.g., headline, button placement) on user behavior.
- **Primary Goal:** Maximize key performance indicators (KPIs) like **Conversion Rate** and **Revenue Per User**.

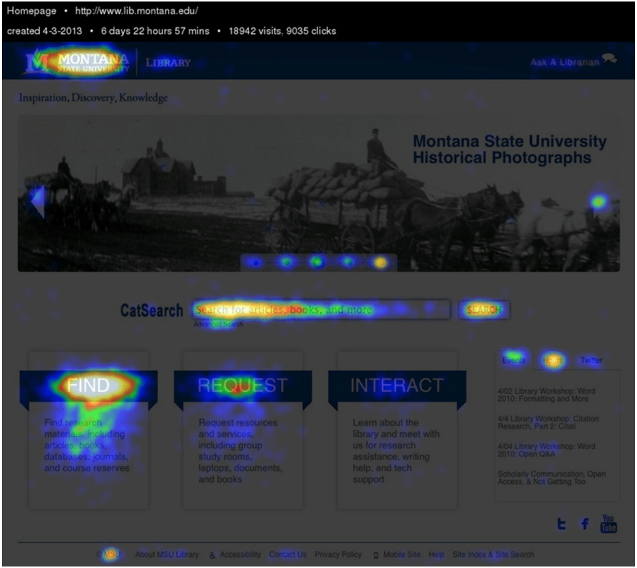

Figure: Library Homepage Click Data - April 3-April 10, 2013

Case study: Improving Library User Experience

- **Problem Identified:** The homepage category "**Interact**" had an extremely low **2% Click-Through Rate (CTR)**.
- **Research Question:** Will changing the confusing category title lead to a measurable increase in user engagement?
- **Refinement:** Used brief user interviews to select the most meaningful title variations for testing.
- **Hypothesis:** Replacing the title with "**Help**" or "**Services**" will generate significantly higher user engagement compared to all other options.

Case study: Improving Library User Experience

- **Set up and run experiment:** Users were randomly served one of the five variations (Control: Interact, Variations: Connect, Learn, Help, Services) over a set period. Tools used included Google Analytics and Crazy Egg.

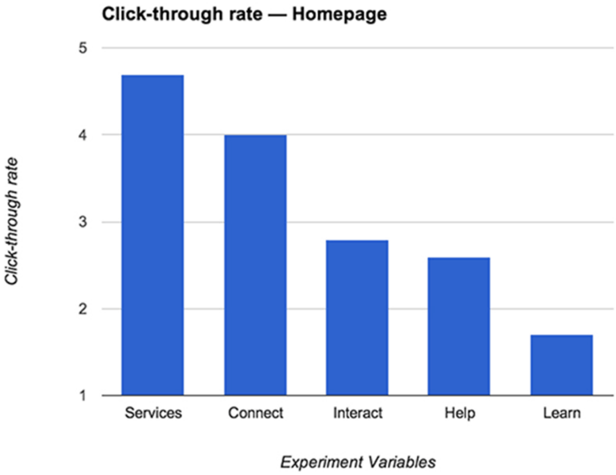

Figure: Click through rate by title variation

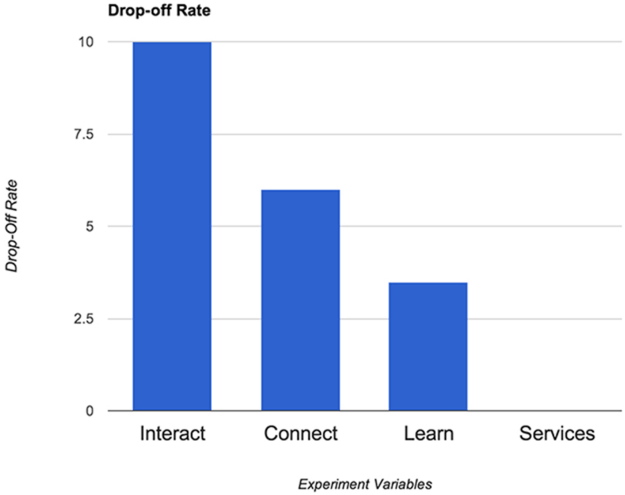

Figure: Drop off rates by title variation

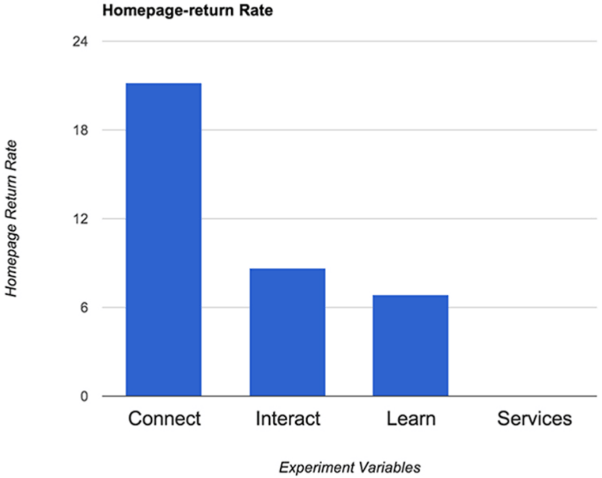

Figure: Homepages return rates by title variation

Case study: Improving Library User Experience

- **Winning Variation:** The title "**Services**" was the highest-performing option across all metrics (CTR, Drop-Off, Return Rate).
- **Unexpected Finding:** The internally favored title, "**Learn**," generated the **lowest user engagement**.
- **Validation:** This confirmed the value of A/B testing—relying on internal opinion would have resulted in a worse UX.

Data Collection Process Overview

- **Define research question:** Identify issues on a.
- **Conduct qualitative interviews:** Gather user insights to refine and validate variations to test.
- **Formulate hypothesis and metrics:** Decide what to measure (click-through rate, drop-off rate, etc.).
- **Set up experiment:** Deploy A/B or A/B/n variations randomly to users with controlled sampling.
- **Collect and analyze data:** Track defined metrics and compare the performance of variations.
- **Share results and decide:** Implement the winning variation based on the analysis.

Simulation Mini-Project

AB test to determine the effectiveness of a new landing page

Context & Experiment Design

- **Context:** The Design team developed a **new landing page** (updated layout, more relevant content) to attract new subscribers.
- **Objective:** To evaluate the **effectiveness** of this redesign compared to the original version.
- **A/B Test Design:**
  - **Total Users:** 100 users were randomly selected.
  - **Group Division:**
    1. **Control Group:** Shown the existing page (Old Page).
    2. **Treatment Group:** Shown the new version (New Page).
- **Data Collection:** User interaction data from both groups was collected and analyzed.

E-news Express Analysis Objectives

As a Data Scientist at E-news Express, we need to determine the effectiveness of the new landing page by answering two main questions:

1. **Engagement Time:**
   - Do users **spend more time** on the new landing page than on the existing page?
   - *Relevant Metric:* Time spent on the page (minutes).

E-news Express Analysis Objectives

As a Data Scientist at E-news Express, we need to determine the effectiveness of the new landing page by answering two main questions:

1. **Engagement Time:**
    - Do users **spend more time** on the new landing page than on the existing page?
    - *Relevant Metric:* Time spent on the page (minutes).
2. **Conversion Rate:**
    - Is the conversion rate (proportion of users who subscribe) for the new page **greater than** the conversion rate for the old page?
    - *Relevant Metric:* Converted (Binary: Yes/No).

Dataset Structure

The dataset includes 6 main columns:

Column	**Description**
`user_id`	Unique identifier for the user.
`group`	Whether the user belongs to the first group (control) or the second group (treatment)
`landing_page`	Which page they interacted with (`old`/`new`).
`time_spent_on_the_page`	Time (in minutes) spent by the user on the landing page.
`converted`	Binary variable: Whether the user **subscribed**.
`language_preferred`	Language chosen by the user to view the landing page.

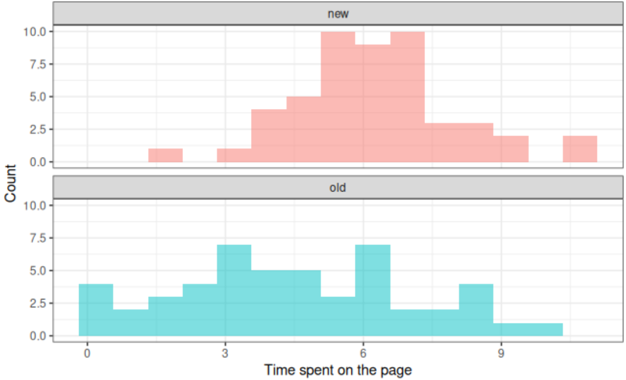

Figure: Distribution of Time Spent on the Page

For the new page, times are mostly concentrated around 4–7 minutes, while the old page shows more spread and more very short visits.

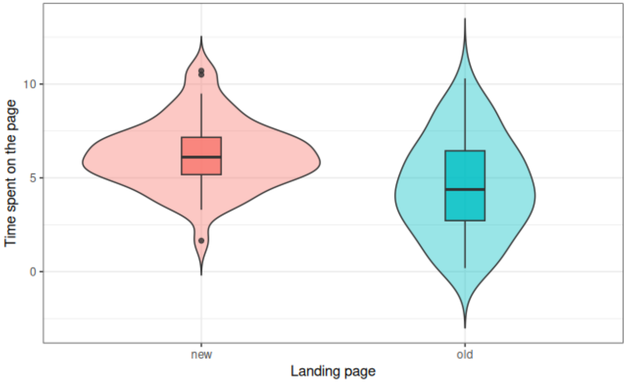

Figure: Violin + boxplot of time spent on the page for each landing page

The median and most of the mass for the new page are higher than for the old page.

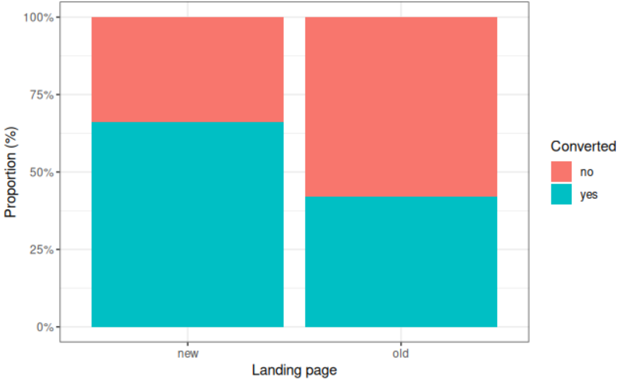

Figure: Stacked bar chart of conversion rates each landing page

The stacked bar chart shows a higher proportion of converted users for the new landing page, with more than a half, than for the old one.

Bootstrap test

- Compare **old** vs **new** page for:

$$\mu_{\text{time,new}} - \mu_{\text{time,old}}, \quad p_{\text{new}} - p_{\text{old}}.$$

- Test statistic for both metrics:

$$t_{\text{obs}} = \bar{x}_{\text{new}} - \bar{x}_{\text{old}}.$$

- Idea under H_0:
  - Make the two groups have the same mean (centering).
  - Resample many times to get t^* values.
  - Compare t_{obs} to the bootstrap *null* distribution of t^*.

Null distribution (time)

- Under $H_0 : \mu_{\text{new}} = \mu_{\text{old}}$:

$$\tilde{x_{\text{new}}} = x_{\text{new}} - \bar{x}_{\text{new}}, \quad \tilde{x_{\text{old}}} = x_{\text{old}} - \bar{x}_{\text{old}}.$$

- For each bootstrap sample:

$$t_b^* = \bar{x}^*_{\text{new},b} - \bar{x}^*_{\text{old},b}.$$

- The histogram of $\{t_b^*\}$ (time data) is the **null distribution** for time.
- p-value (idea):

$$p \approx \frac{\text{number of bootstrap } t_b^* \text{ with } |t_b^*| \geq |t_{\text{obs}}|}{R}$$

(or one tail for a one-sided test).

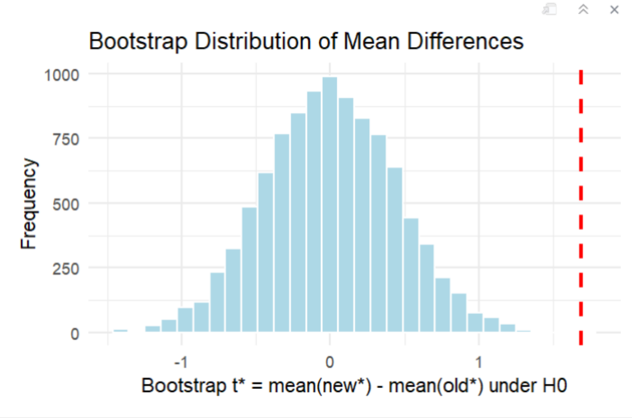

Figure: Time: bootstrap t^* with dashed line at t_{obs}.

Null distribution (conversion)

- Encode conversion as $x \in \{0, 1\}$, so $\bar{x} =$ conversion proportion.
- Apply the same bootstrap algorithm to:

  $x_{\text{new}} =$ conversion on new page, $\quad x_{\text{old}} =$ conversion on old page.

- For each bootstrap sample:

$$t_b^* = \bar{x}^*_{\text{new},b} - \bar{x}^*_{\text{old},b}.$$

- The histogram of $\{t_b^*\}$ gives the **null distribution** for the difference in conversion rates.

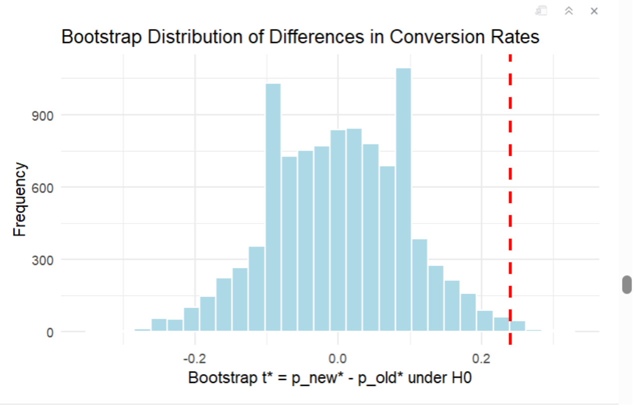

Figure: Conversion: bootstrap t^* for new − old.

Bootstrap CI idea

- For confidence intervals, we **do not** enforce H_0.
- Resample with replacement from the **original** data in each group.
- For each bootstrap sample:

$$\Delta_b^* = \bar{X}^*_{\text{new},b} - \bar{X}^*_{\text{old},b}.$$

- Collect many Δ_b^* values and sort them.
- Percentile CI:
  - lower endpoint = empirical $\alpha/2$ quantile,
  - upper endpoint = empirical $1 - \alpha/2$ quantile.

CI for time

- Apply the bootstrap CI procedure to:

$$\mu_{\text{time,new}} - \mu_{\text{time,old}}.$$

- 95% percentile CI for the difference in mean time (new − old) is obtained from the sorted Δ_b^* values.

A tibble: 2 × 2

endpoint <chr>	value <dbl>
Lower 95% CI	0.82559
Upper 95% CI	2.56446

2 rows

Figure: 95% bootstrap CI for mean time difference.

CI for conversion

- Apply the same CI procedure to:

$$p_{\text{new}} - p_{\text{old}},$$

  where p is the conversion proportion (0/1 data).
- 95% percentile CI for the difference in conversion rate (new − old).

A tibble: 2 × 3

quantity <chr>	endpoint <chr>	value <dbl>
p_new - p_old	Lower 95% CI	0.06
p_new - p_old	Upper 95% CI	0.42

2 rows

Figure: 95% bootstrap CI for conversion difference.

Limitations

- **Sample may not be representative**
  - Only 100 users; may come from a specific period, country or device type.
- **Small sample size in each group**
  - 50 old, 50 new; bootstrap distributions and CIs can be noisy.
- **Independence and simple structure**
  - Assumes i.i.d. users within each group.
  - Ignores clusters (same user, campaign, time-of-day); more advanced bootstrap would be needed there.

Permutation Test

1. **Analysis of Time-on-Page**
   - **Research Question:** Is there a significant difference in time spent?
   - **Null Hypothesis (H_0):** $\mu_{\text{new}} = \mu_{\text{old}}$
   - **Alternative (H_1):** $\mu_{\text{new}} \neq \mu_{\text{old}}$ (Two-sided)
2. **Analysis of Conversion Rate**
   - **Research Question:** Is there a significant difference in conversion rate?
   - **Null Hypothesis (H_0):** $p_{\text{new}} = p_{\text{old}}$
   - **Alternative (H_1):** $p_{\text{new}} \neq p_{\text{old}}$ (Two-sided)

Methodology

Both analyses will utilize permutation sampling (two-sided test).

Methodology: Generalized Permutation Test

The `perm_test()` function encapsulates a 6-step logic (from the diagram):

1. **Step 1: Compute t_{obs} (Observed Statistic):** Calculate the observed statistic from the original data.

$$t_{\text{obs}} = \bar{Y}_{\text{new}} - \bar{Y}_{\text{old}}$$

2. **Step 2: Pool Data:** Combine all observations $(Y_1, \ldots, Y_n)$ under the null hypothesis.
3. **Steps 3 & 5: Shuffle and Repeat:** Repeat $R = 10{,}000$ times: randomly shuffle the `group` labels.
4. **Step 4: Compute t^* (Permuted Statistic):** Recalculate the statistic (t^*) for each permuted dataset.

$$t^* = \bar{Y}^*_{\text{new}} - \bar{Y}^*_{\text{old}}$$

5. **Step 6: Calculate p-value (Two-sided):** Find the proportion of permuted statistics as extreme as the observed one.

Execution and Results ($R = 10{,}000$)

Output

```
> cat("Observed Mean Difference (Time):", res_time$observed, "\n")
Observed Mean Difference (Time): 1.6908

> cat("P-value (Time):", res_time$p_value, "\n")
P-value (Time): 1e-04
```

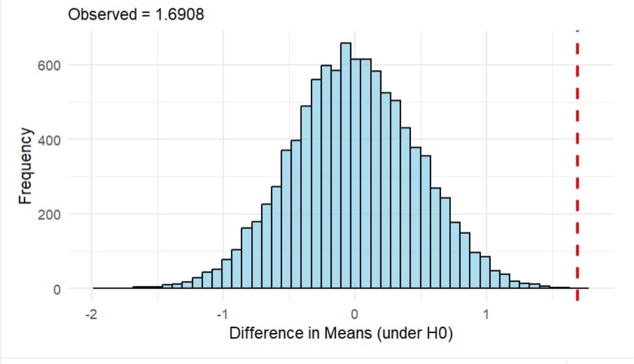

Figure: Permutation Distribution: Mean Time Difference.

Analysis 2: Conversion Rate

Test Statistic Function (Proportion Difference)

$$t = \hat{p}_{\text{new}} - \hat{p}_{\text{old}}$$

Execution and Results (R = 10,000)

```
> cat("Observed Proportion Difference (Conversion):", res_conv$observed, "\n")
Observed Proportion Difference (Conversion): 0.24

> cat("P-value (Conversion):", res_conv$p_value, "\n")
P-value (Conversion): 0.0147
```

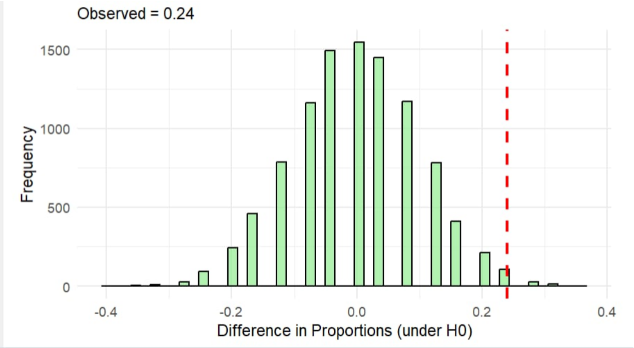

Figure: Permutation Distribution: Conversion Rate Difference (New - Old).

Interpretation and Conclusion

Analysis 1: Time-on-Page (H1: new $>$ old)

- The observed p-value is **0.0001**.
- Since the p-value (0.0001) is less than our significance level $(\alpha = 0.05)$, we **reject the null hypothesis (H_0)**.
- **Conclusion:** We have statistically significant evidence that users spend more time on the new landing page.

Analysis 2: Conversion Rate (H1: new $>$ old)

- The observed p-value is **0.0147**.
- Since the p-value (0.0147) is less than our significance level $(\alpha = 0.05)$, we **reject the null hypothesis (H_0)**.
- **Conclusion:** We have statistically significant evidence that the new landing page has a higher conversion rate.

Limitations

- **Sample may not be representative:**
  - Data (e.g., 100 users) may come from a specific period, country, or device type, not reflecting the entire user base.
- **Small sample size in each group:**
  - e.g., Only 50 in the Old group, 50 in the New group.
  - The permutation distribution can be unstable (noisy).
- **Independence Assumption:**
  - The permutation test assumes users are i.i.d (independent and identically distributed).
  - It ignores clusters (e.g., the same user, same campaign). More complex methods are needed if clustering is present.

Methodological Notes

- The analyses utilized $R = 10{,}000$ permutations. This number is generally sufficient for stable p-value estimation.

- This non-parametric (permutation) approach is robust as it avoids the assumptions of normality or homoscedasticity (equal variances) required by parametric counterparts (e.g., the two-sample t-test).

- The only core assumption is the **exchangeability** of observations under the null hypothesis (H_0).

Kết Thúc

Cảm ơn các bạn đã lắng nghe!