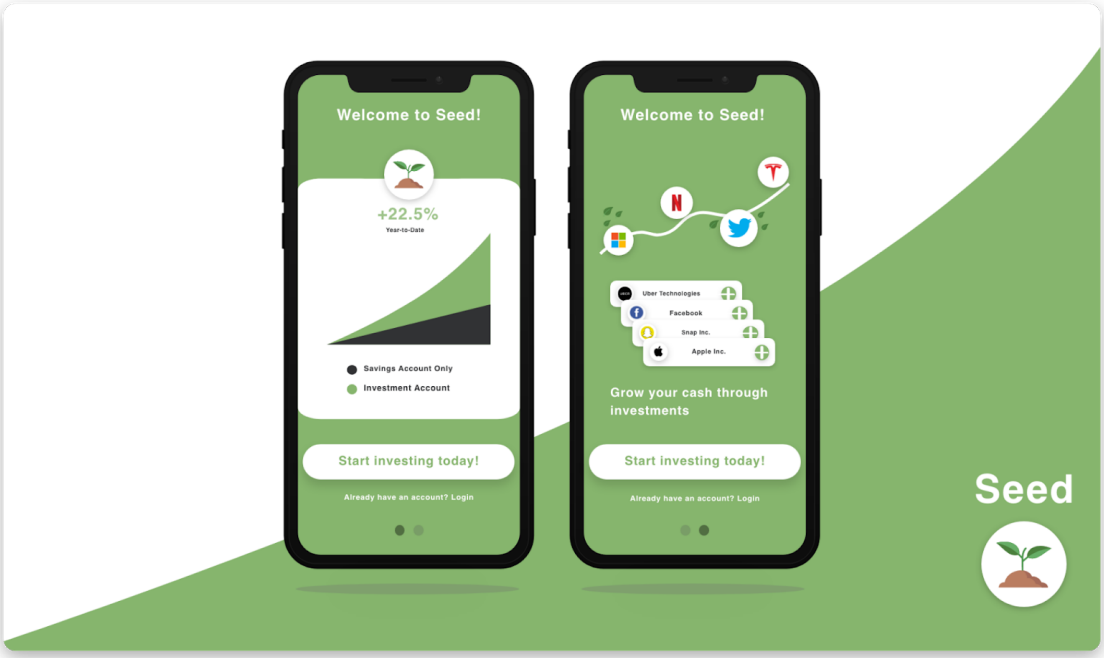
A/B testing at Seed Investing (Case-Study)



Overview

Seed is a **fictional** fast-growing fintech company, growing to 100K users in the first 6 months since their app launched. While new users have been signing up, there has been low conversion to commit to their first investment deposit. This is the more important metric they track since it increases the assets under management (AUM). Seed charges 1% of the total AUM in each user account annually.

The app currently records a **22.68%** conversion rate and a standard deviation of 0.42 for new users opening an account and submitting their first deposit into an investment account. They call this ***New User First Deposit Rate (NUFD Rate)***. The company believes that customers would be more inclined to open an investment account if they could interact with the app before providing personal details like SSN, contact information, and bank details.

After running user surveys with new and current users, the team found a strong interest in improving the user experience by adding more content and accessibility. The goal was to provide as much transparency about the investment app before users needed to commit any financial capital.

Goal

The goal of this report is to provide suggestions for experimental variations and assess the impacts of the Control (A) vs. Test (B) groups. Given the size of Seed's company and lack of excessive financial resources, the team can only provide experimentation to a portion of the users. This report starts by interpreting results from **Experiment #1** to improve the ***New User First Deposit Rate*** **(NUFD Rate)**, then continues to **Experiment #2**, concentrating on **Average First Deposit Value ($).**

Experiment #1 - A/B Test Results

To address this issue, I developed an experiment scenario for new users to experience in the onboarding process. Working with the design and engineering team to implement these changes, and test the results with a 50/50 split between the control (A) and experiment (B). Below are descriptions of the two groups.

- A → Control group and the current onboarding experience
- B → Test group that allows users to demo the app before signing up

Option A → Current Situation (Control Group)
When users download the app, they **must** sign up with their email and provide information to open their investment account, including their SSN, banking information, and personal income. The user can't interact with the app without providing this information to open their account.

Option B →Demo Mode (Test Group)
Allow the users to download the app and interact with it on a demo screen. This demo screen gives the new users a hypothetical look at their accounts and investments before they need to provide personal information. It is essentially giving them the ability to browse without committing. A small dropdown option would allow the user to sign up for the account whenever ready.

The experiment was live for one week to collect data. The results have been collected, and they have seen marginal improvements across the two test groups compared to control A. Since we know the population's standard deviation (0.42), this will be a one-tail z-test with an alpha (α) level of 5% (0.05).

Below are the results from the A/B test.

Hypothesis:

- **H_0 : μ = 22.68% → (Platform New User First Deposit Rate)**
- **H_1 : μ > 22.68%**

Experiment #1 - A/B Test Results

Group	Sample Size	Sample Proportion	Converted Users	NUFD Rate	Lift %
A	9,753	49.3%	2,218	22.68%	-
B	10,042	50.7%	2,381	23.78%	+4.85%

1. $$Z = \frac{(\bar{x} - \mu)}{(\sigma/\sqrt{n})}$$
2. $$Z = \frac{(0.2378 - 0.2268)}{(0.42/\sqrt{10042})} = 2.6241$$
3. $$p - value = (1 - 0.9956) = 0.004$$
4. $$\alpha = 0.05 \longrightarrow 0.004 < 0.05 = Reject\ H_0$$

After one week of testing, they can see that experiment group B recorded a **23.78% NUFD Rate**, a +4.85% uplift from the control group. From the observed results, they ended up with a z-score of 2.62, and when referencing the z-table, they obtained a **p-value of 0.004** (1 - 0.9956). Thus the results from the experiment are statistically significant since **0.004 < 0.05**. As a result, they can reject the H_0 and claim the changes for experiment group B has improved the NUFD Rate.

Experiment #2 - A/B Test Results

The team has officially deployed group B (Demo Mode) into production, and 100% of users are exposed to the new experience. The Product Managers were pleased with Experiment #1 results, so they wanted to run a **new A/B (Experiment #2)** test to see if they can continue improving the product.

The team wants to test a new metric **Average First Deposit Value ($)**, which the average dollar value a new user deposits into their investment account. The team wants to introduce an additional step before a user enters the Demo mode and sign-ups for an account, which provides them with Pre Sign-Up educational content on how the app works. The team believes that by providing more information about investment options, users will consider increasing their **Average First Deposit Value ($)**.

This new A/B test is financially costly for the company to have ongoing for too long. Therefore a four-week testing period was set to collect data since it generally takes a new user to submit their first deposit within two weeks of signing up.

Option A → Current Situation (Control Group)

Option A in this new A/B test would represent the changes from Experiment #1. We know that the group B, from Experiment #1 was implemented and deployed to the entire population. Now in Experiment #2, this group would represent the control group.

Option B → Pre Sign-Up Content (Test Group)

Replace the "sign up" screen a new user sees with educational resources about what exactly the app does. This step will have multiple swiped screen options explaining each feature and benefit the new user will receive if they signed up. The user will not need to provide any personal information to view the educational content. It is purely for informational purposes, intending to provide as much content as persuading the user to sign up.

Hypothesis:

- **H_0 : μ = $72.43 → (Platform Average First Deposit Value ($))**
- **H_1 : μ > $72.43**

Experiment #2 - A/B Test Results

Group	Sample Size	Sample Proportion	Total Deposits	Avg Deposit	Lift %
A	9,467	49.3%	$685,676	$72.43	-
B	9,745	50.7%	$767,185	$78.73	+8.70%

1. $$Z = \frac{(\bar{x} - \mu)}{(\sigma/\sqrt{n})}$$
2. $$Z = \frac{(78.72 - 72.42)}{(387.03/\sqrt{9745})} = 1.61$$
3. $$p - value = (1 - 0.9463) = 0.054$$
4. $$\alpha = 0.05 \longrightarrow 0.054 > 0.05 = Fail\ to\ Reject\ H_0$$

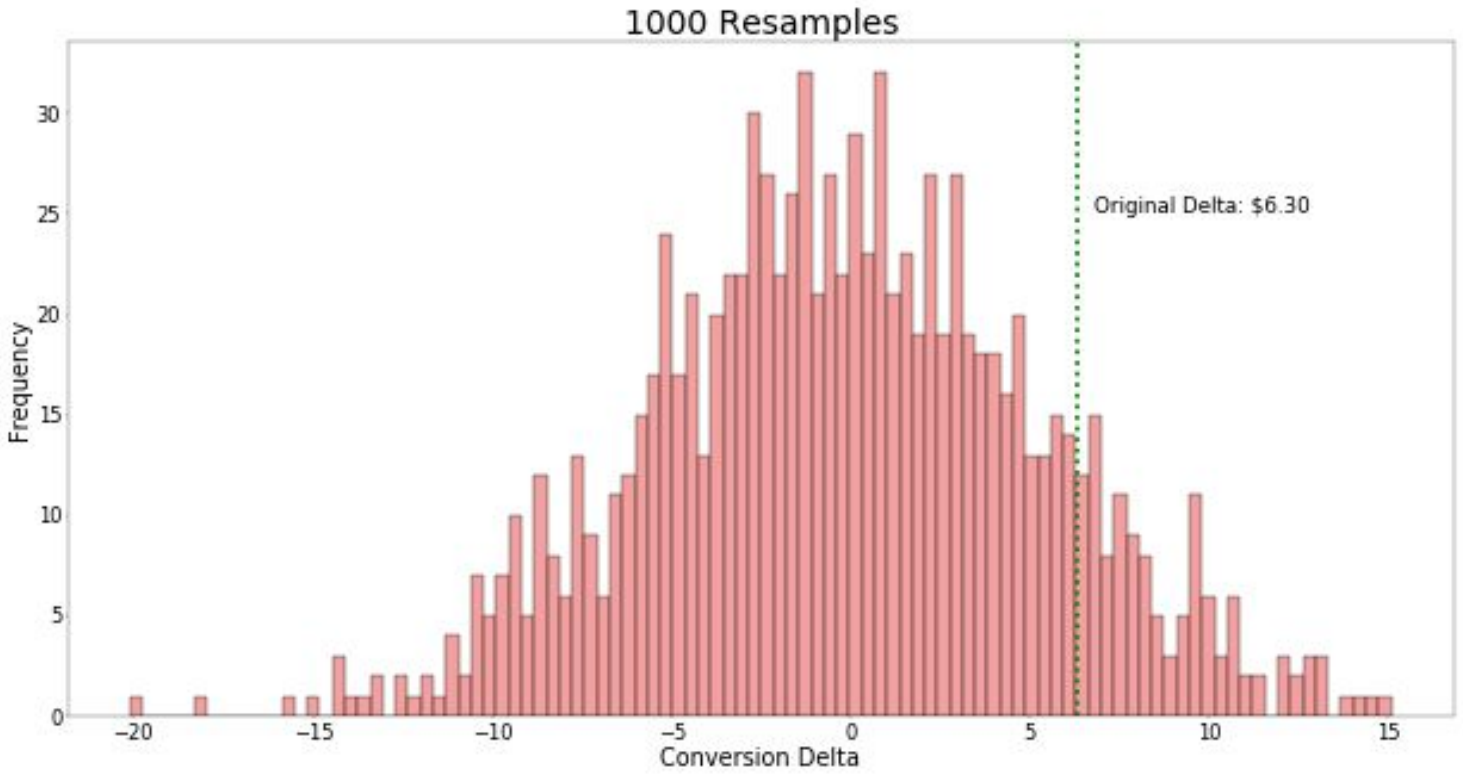
We can see from the results that the B group performed better with an average deposit of $78.73, a +8.7% uplift from A. However, running through the same process to validate these results is statistically significant; they saw that our p-value was 0.054, so they fail to reject the Null hypothesis. The results are not strong enough to indicate that the results from the change are not random.

However, since a **8.7%** increase seems relatively good, I wanted to ensure if these results would remain true if they ran some resampling through permutation testing. Remember that this change would be too costly to the company financially, so the Product Managers wanted to be 100% sure this would be beneficial. they could not run the test for a more extended period to collect new samples, so resampling via permutation testing would help validate if there was, in fact, NO randomness in group B's results.

A permutation test is a type of statistical significance test. The focus is on capturing the deltas' distribution between the variant and control through numerous rearrangements/resampling of the available data points. In this case, it is the delta in the Avg Deposit.

Since $78.73 (group B) was higher than $72.43 (group A), the original delta from the experiment was +$6.30 and will be used as the benchmark for the resampling results. Running 1000 resamples and storing each delta, they can see the distribution below.

Experiment #2 - Permutation Testing



The results show us that 121 of 1000 resamples recorded a delta ≥ $6.30, which would result in a p-value of 0.121. Therefore they can not reject the Null even after running through the permutation tests, and the results are not statistically significant.

$$p - value = \frac{121}{1000} = 0.121$$

Before the permutation test, we saw that the p-value was very close to significance, and because of the +8.7% (+$6.30) in average deposit, it can be tempting to conclude the results are robust. However, since there is a strict time window to obtain results, and we can't increase the sample size to collect more information, running the permutation test was the best option to validate the numbers. Nonetheless, the company has a tight budget and limited engineering bandwidth to develop a feature that doesn't provide significant results, so we can not confidently conclude that group B made a difference.

Conclusion

From Experiment #1 A/B test, we saw that test group B (Demo Change) performed +4.85% better than the Control. It resulted in a **New User First Deposit Rate (%)** of 23.78% compared to the original 22.68%. These numbers were proven to be statistically significant, and after sharing results with the company and stakeholders, group B was developed and deployed to 100% of Seed's users.

After the success from Experiment #1, the team wanted to continue growing the product, prompting further exploration to develop Experiment #2 A/B test to see if they can move a different metric **Average First Deposit Value ($)**. I suggested adding an educational content section before the sign-up screen → group B in the new test.

Experiment #2 was very costly from an engineering and design perspective. Hence, the test had to run for a maximum of four weeks. Despite group B showing positive results by improving the average deposit value by +$6.30, results were not significant. In order to ensure that results were not significant, I conducted additional validation steps through permutation testing. Unfortunately, the permutation test resulted in numbers not being significant. Since the financial burden of developing this new feature was high, the team decided not to pursue these changes.

Appendix

- Python Code - Jupyter Notebook
- Experiment #1 dataset
- Experiment #2 dataset