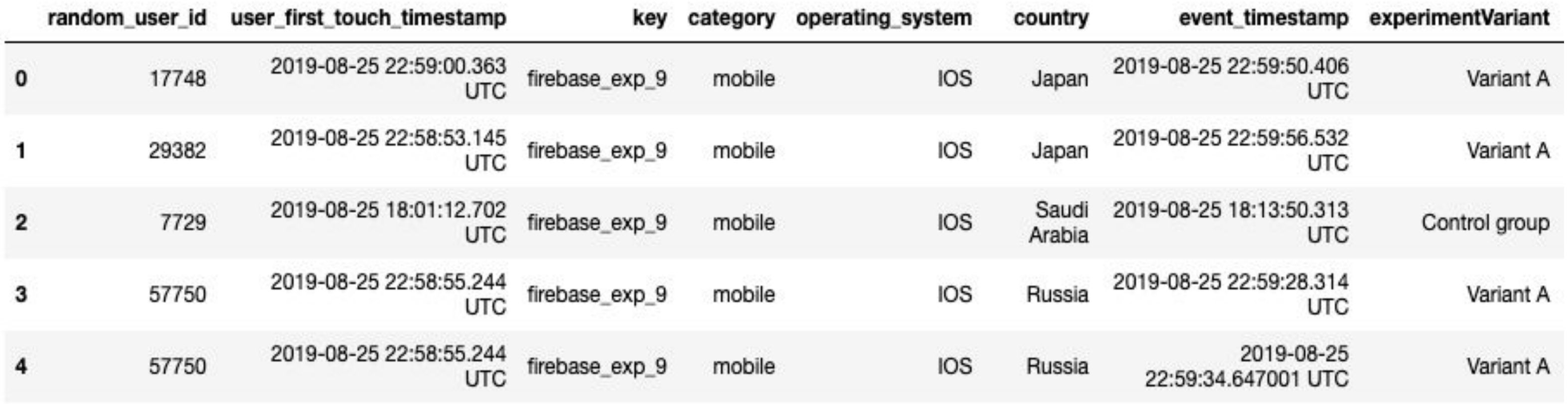
Mobile App

AB Testing Case Study

Si Li

Data Available

As more new users download the app, the company needs a refresher on the on-boarding experience, due to the fact that new users don't necessarily have prior knowledge of the tech. To respond, the company has been building a new feature to better the onboarding experience and is conducting a AB testing to quantify the impact of such a feature on the App.



	random_user_id	user_first_touch_timestamp	key	category	operating_system	country	event_timestamp	experimentVariant
0	17748	2019-08-25 22:59:00.363 UTC	firebase_exp_9	mobile	IOS	Japan	2019-08-25 22:59:50.406 UTC	Variant A
1	29382	2019-08-25 22:58:53.145 UTC	firebase_exp_9	mobile	IOS	Japan	2019-08-25 22:59:56.532 UTC	Variant A
2	7729	2019-08-25 18:01:12.702 UTC	firebase_exp_9	mobile	IOS	Saudi Arabia	2019-08-25 18:13:50.313 UTC	Control group
3	57750	2019-08-25 22:58:55.244 UTC	firebase_exp_9	mobile	IOS	Russia	2019-08-25 22:59:28.314 UTC	Variant A
4	57750	2019-08-25 22:58:55.244 UTC	firebase_exp_9	mobile	IOS	Russia	2019-08-25 22:59:34.647001 UTC	Variant A

Experiment design

As more new users download the app, the company needs a refresher on the on-boarding experience, due to the fact that new users don't necessarily have prior knowledge of the tech. To respond, the company has been building a new feature to better the onboarding experience and is conducting a AB testing to quantify the impact of such a feature on the App.

Actual experiment duration: 2019-08-20 ~ 2019-09-01

Actual number of qualified unique users: Control (37,581) vs Variant (37,269)

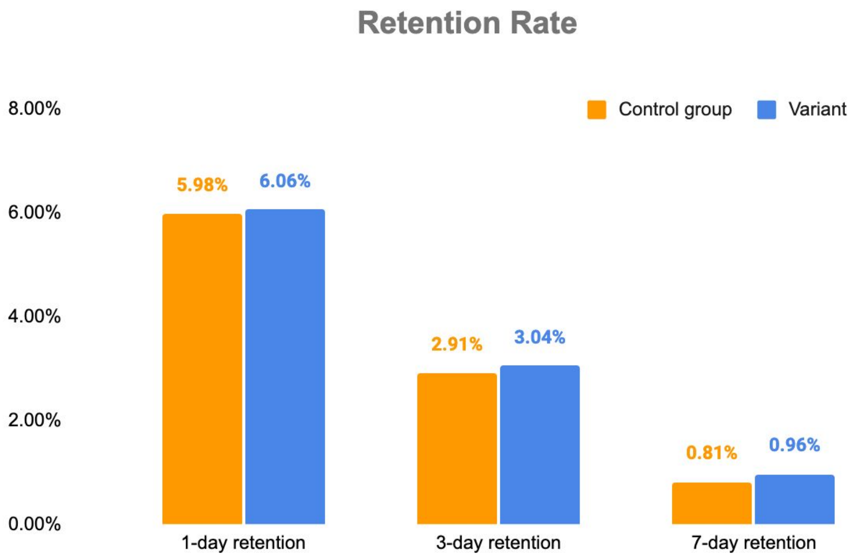
Null Hypothesis: there is no statistical difference between control and variant groups

Statistical test: two-tailed chi square test, t-test, and bootstrapping difference t-test of **alpha value** 0.05

Key metric*: 1-day retention, 3-day retention, 7-day retention

* This AB test could more metrics such as session length. But in the dataset provided, the only relevant variable is timestamp as it can be used to derive retention rate

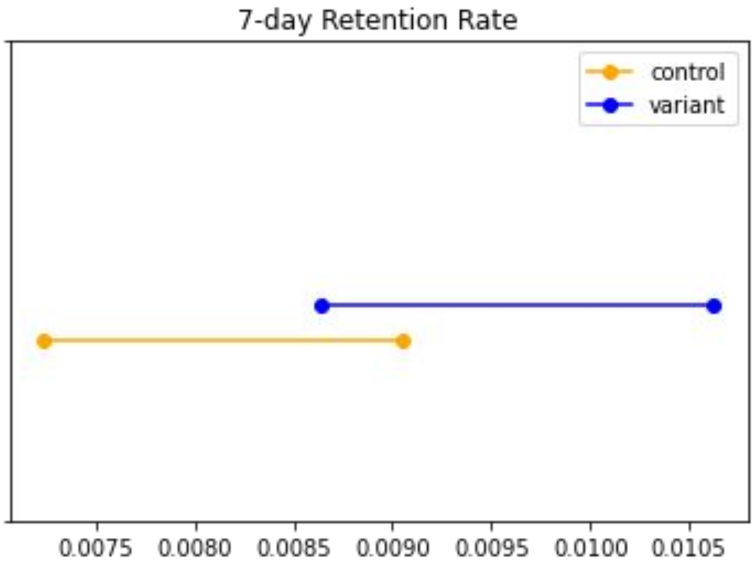
Data Analysis - Retention & P value



	1-day retention	3-day retention	7-day retention
P value	0.67	0.29	0.03

- Retention rates of variant group are higher, which is good news. It signals that the new feature potentially has improved retention rates
- However, only **7-day retention** is statistically significant given less than 0.05 p value
- It shows that there is **not enough sample size** to this experiment given there are consistent positive signals but only one of them is statistically significant

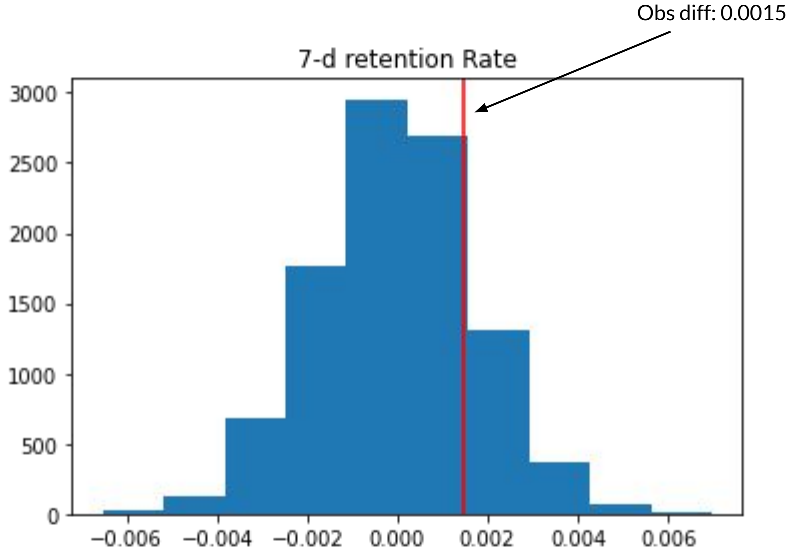
Data Analysis - Confidence Interval



	control	variant	**lift**
7-d retention	0.81%	0.96%	0.15%

- There is a small overlap in the confidence interval even the p-value is less than 5% and the lift is very small - only 0.15%
- Statistics is all about estimation given the obtained samples. It's possible that the true values lie within the overlapping area. Therefore, it's risky to declare a winner in this scenario and measure the lift
- It strengthen the belief that there is **not enough sample size** to this experiment

Data Analysis - Bootstrap



	1-day retention	3-day retention	**7-day retention**
P value	0.34	0.22	0.19

- As noted previously, the **ideal** sample size is **156,978** for each group (aiming for 1% lift) but the **actual** sample size of each group is **37,581** - a third of the ideal size
- Traditional significant test works best when sample size is **large enough** and follows the underlying **normal distribution** (or alike), which is not met. Therefore, to verify, bootstrapping approach, which does not assume any underlying distribution of the data, is performed
- Unfortunately, **none** of the metrics is tested statistically significant using bootstrap approach

Experiment design

As more new users download the app, the company needs a refresher on the on-boarding experience, due to the fact that new users don’t necessarily have prior knowledge of the tech. To respond, the company has been building a new feature to better the onboarding experience and is conducting a AB testing to quantify the impact of such a feature on the App.

Actual experiment duration: 2019-08-20 ~ 2019-09-01

Ideal experiment duration: 2019-08-20 ~ 2019-09-19

Actual number of qualified unique users: Control (37,581) vs Variant (37,269)

Ideal sample size of each group given 1% lift: 156,978

Null Hypothesis: there is no statistical difference between control and variant groups

Statistical test: two-tailed chi square test, t-test, and bootstrapping difference t-test of **alpha value** 0.05

Key metric*: 1-day retention, 3-day retention, 7-day retention

Ideal key metrics: retention, session length, number of sessions

* This AB test could more metrics such as session length. But in the dataset provided, the only relevant variable is timestamp as it can be used to derive retention rate

Conclusions

- The good news is that variant group consistently performs better than control group in retention rate
- However, only the 7-day retention rate is test statistically significant and the lift is marginal
- Due to the insufficient sample size, the lift is hard to measure definitively
- It might be hard to justify introducing such a feature that would only **potentially and marginally** improve the app as there are lots of cost, risks and uncertainties associated if rolling out the feature to every user

Recommendations

- The experiment could be better designed by clearly defining a **sufficient sample size**, **experiment duration**, **metrics** to measure as it would substantiate the ab testing results and allow rigorous and robust analysis
- Also, it might be worthwhile to **talk to product team** on the hypotheses of introducing such a feature and develop more metrics to tell a comprehensive story. Right now, only retention rate can be measured with the dataset provided

Thank You