Product Analyst: Case Interview

Optimizing User Engagement for Client

Subhra Mukherjee

Experiment Design

Firstly, decide on the proportion of users to be Randomly assigned to Test vs Control. Typically 50:50 but if the Feature could be potentially disruptive then 80:20 (or other proportions) can be used. Product Manager in consultation with Analyst usually decides

Control Group

- Randomized Users who visited pages on which the "Free Cancellation" badge is present (but hidden/not shown to them). Note, any arbitrary user who hasn't visited such pages are not considered in the experiment
- If a user converts post visiting such pages within a predefined time period (say 7 days & not necessarily in the same session) then record a Conversion
- **Conversion Rate:** No of Converting Users/Total Number of Users

Test Group

- Randomized Users who saw the "Free Cancellation" badge in a Session. This is a **User Level Experiment** & not a Session level Experiment
- If a user converts post seeing the "Free Cancellation" badge within a predefined time period (say 7 days & not necessarily in the same session) then record a Conversion. (Every user is a **bernoulli random variable** i.e. a coin flip essentially)

Null Hypothesis (Strictly Superiority Test): **Conversion Rate (Test) - Conversion Rate (Control) <= δ** **(Delta is the Superiority Margin ,typically set to 0)**
Alternative Hypothesis: **Conversion Rate (Test) - Conversion Rate (Control) > δ**

Guard Rails for correct Experimental Setup

- Depending on the Control vs Test Proportion (say it is 60:40) setup every user has 60% chance to be not shown the badge vs a 40% chance for it to be shown. Every user who lands on such a page (a test subject) thus gets **randomly assigned** to either Control or Test. However, doing this systematically (e.g., using a counter to show 6 users the Control version and Test to the next 4) induces significant chances of Selection Bias

- For a User Level Experiment, it's important to keep showing one user the **same version** across multiple Sessions that they may be visiting for

- Experiments assume that except the Feature everything else is Ceteris Paribus (i.e., everything else remaining the same). While this may seems obvious to suggest in reality whilst running multiple A/B tests Users/Sessions intersect in unpredictable ways giving rise to **Interaction Effects** which is difficult to detect. So try to minimise running multiple experiments

- The most common 'error' when setting up a Test is a Sample Ratio Mismatch. Say Control vs Test Proportion is set as 60:40, and after running the test for 14 days we see 6100 users in control vs 3900 users in test. Even though this seems close to the Original proportion we have a **Sample Ratio Mismatch** here (based on a Fisher's exact test or a Chi-Squared Test). Typically, incorporation of a new feature (like a "Free Cancellation" badge through a Gtag) alters page load speeds which impacts time taken for the firing of the experimentation tag

- This Test can be extended to guard against particular confounders. For Example, Test group is showing a higher conversion value but that could be because we allocated more repeat users in Test vs Control. One can run a Chi-Squared Test of Independence for Test vs Control against Returning & New users to see if the factors are Causal. Such tests however are feasible once Test is live

Sample Size Calculations

- Sample Size Calculations are important because firstly we want to get a result to the Stakeholders as soon as possible to minimise the Opportunity Costs of implementing vs not implementing the Feature
- As explained in Kolmogorov's Law of Iterated Logarithms and as A/B testing practitioners may have intuitively noticed, too large sample sizes will make minute differences in Metrics significant (this is true for both Bayesian vs Frequentist hypothesis testing)
- I will **employ a frequentist Sample Size Calculation** even though i will **employ a Bayesian Framework** later when actually running the Test. The Statistical justification is that Standard Frequentist Frameworks are usually much more restrictive than Bayesian Frameworks , basically meaning that they ask for more data. So if one has enough data to decide via a Frequentist framework one can assume that they have sufficient data for a Bayesian framework
- **Inputs for Sample Size Calculations: a) Minimum Detectable effect (MDE)** typically coming from the stakeholder who generated the hypothesis. For example, say **20%** is the current conversion rate and the hypothesis is that introduction of the **"Free Cancellation"** badge takes it to **22%** (MDE is 0.02). **b) Significance** & **c) Power** of the test are other two inputs but usually set as **95%** & **80%**
- Based on the fact that we are running a Strictly Superiority Test, and Control vs Test Proportion is set as 60:40, we would need **425 users in Control vs 283 in Test** for having a sufficient sample to determine a **0.02%** lift on a **0.2%** Base Conversion Rate*
- Smaller an MDE, or larger the significance or power, larger the sample size and longer one needs to run tests. Typically one runs tests for Full Weeks (so 7 , 14, 21 days) to avoid any daily fluctuations

*Based on Sample Size Calculation for Comparing Proportions, Hansheng Wang, Shein-Chung Chow

Finalizing the Results

- The Final Bayesian Calculation (**implemented here**) estimates the **Probability of {Conversion Rate (Test) > Conversion Rate (Control)}.** Any value above 90% is usually accepted as sufficient proof that the Test proportion is indeed better
- We also want to capture **No of Booking per User (Test/Control), price per booking (Test/Control), no of users interacting with the Badge** vs **no of users who were shown the badge but didn't interact with it** etc.
- **Estimating the Business Impact:** Next we want to be be able to estimate the impact of incorporating this feature. We want to see the revenue impact of incorporating the change. So **Revenue per User** (Test/Control) = **Conversion Rate (Test/Control)** * **No of Booking per User (Test/Control)** * **price per booking (Test/Control)**
- Using the Delta Method on **Revenue per User** (Test) vs **Revenue per User** (Control) one estimate increase in the actual **Revenue per User** from implementing the **"Free Cancellation" badge** with some certainty (for example, one may conclude with **90% confidence** that the increase in Revenue is at least **$3 per user**)
- Point here is that the A/B Test may be successful but it **doesn't warrant** a feature implementation. For example the increase **Conversion Rate (Test)** may be accompanied by a reduction in **price per booking (Test)**
- Meta analyses for validating reinforcing hypotheses, like if the **"Free Cancellation"** badge did indeed improve conversion rates then amongst the Test group, those who interacted with the badge should have a higher conversion rate than those that didn't could also be tested with the same Bayesian Framework for further validation