AB Testing - Business Learning for Data Science

Avinash Barnwal

March 2019

1 Motivating Examples

1.1 Checkout page at Doctor FootCare

Conversion rate of users is the percentage of visits to website that include a purchase.
One variant had coupon code option and other variant didn't have it. Variant having coupon code had lower conversion rate as users started thinking twice about whether they were paying more.

1.2 Ratings of Microsoft Office help articles

Classical problem of putting the more ads comparing against users experience. This problem is categorized as Overall Evaluation Criteria(OEC). Page Views and Clicks are the key metrics considered and have been assigned monetary value.
Page Views have an assigned value based on ads and clicks to destination from the MSN are determined based on two ways:-

- **Gain** - Monetary value that the destination property assigned to a click from the MSN home page.
- **Cost** - The cost paid to search engines to generate leads to reach the destination links.

Idea is to first analyze the click through rate(CTR) and then convert the CTR to monetary value to decide on different variants.
Final result - A controlled experiment was run on 5% for 12 days.
Clickthrough rate decreased by 0.38% and p-value = 0.02.
Translating the lost clicks to their monetary value , it was higher than the expected ad revenue, so idea of adding more ads to MSN home page was scrapped.

1.3 Behavior-Based Search at Amazon

Its a new kind of recommendation algorithm where algorithm suggests item based on "People who searched for X bought item Y." It gave some spurious products but using controlled experiment it increased by revenenue by 3

2 Controlled experiments

2.1 Terminology

Overall Evaluation Criterion (OEC) - A quantitative measure of the experiment's objective. Metrics can be short-term(i.e. clicks) but we should also consider long-term aspects like predicting life-time value and repeat visits.

Factors - A controllable experimental variable that is thought to influence the OEC. Factors are assigned values, leading to single factor with two values : A and B.

Variant - A user experiences different features which we are testing, where one is Control and one which we are testing is the Treatment. In case of bug, for example, the experiment is aborted and all users should see the Control variant.

Experimental unit - The entity over which metrics are calculated. Sometimes called an item. The units are assumed to be independent.

Null hypothesis. - The hypothesis,that the OECs for the variants are not different and any observed differences during the experiments are due to random fluctuations.

Confidence level - The probability of failing to reject(i.e., retaining) the null hypothesis when it is true.

Power - Power measures the our ability to detect a difference when it indeed exists.

A/A test - Instead of an A/B test, you exercise the experimentation system, assigning users to one of two groups, but expose them to exactly the same experience. An A/A test can be used to (i) collect data and assess its variability for power calculation. (ii) test the experimentation system.

Standard-deviation - A measure of variability, typically denoted by σ.

Standard-Error - It is standard deviation of sampling distribution of sample statistics.

2.2 Hypothesis testing and sample size

Factors impacting the test:-

- Confidence Level :- Commonly set to 95%, this implies tht 5% of the

time we will incorrectly conclude that there is a difference when there is none(Type I error). All else being equal , increasing this decreases the power.

- Power :- The probability of determining that the difference is statistically significant.
- Standard error :- The smaller the Std-Err, the powerful the test. Three ways to reduce Std-Err
  - Increase the sample size.
  - Choose better OEC such as proportion.
  - Exclude users which are not impacted in the experiments.
- Effect :- The difference in OECs for the variants, i.e. the mean of Treatment minus the mean of the Control.

Two formulas are useful to share in this context. The first is t-test , used in A/B tests.

$$t = \frac{\bar{O}_B - \bar{O}_A}{\hat{\sigma_d}} \tag{1}$$

We assume throughout that the sample size is large enough to apply the Normal distribution even when it is skewed distribution.
Second formula is about the minimum sample size required for confidence level = 95% and power = 80%.

$$n = \frac{16\sigma^2}{\Delta^2} \tag{2}$$

A more conservative formula for sample size(for 90% power) has been suggested

$$n = \frac{4r\sigma^2}{\Delta^2} \tag{3}$$

2.2.1 Example: impact of lower-variability OEC on the sample size

Consider 2 types OECs - Purchasing amount and conversion rate.
Sample Size calculation - Purchasing amount For 5% change in revenue with 95% Confidence level and 80% power. Assume standard deviation is 30 , average spend is 3.75 and conversion rate is 5%.

$$n = \frac{1630^2}{3.75 * 0.05^2} n = 409,000 \tag{4}$$

Sample Size calculation - Conversion Rate

$$n = \frac{16 * (0.05) * (1 - 0.05)}{0.05^2} n = 122,000 \tag{5}$$

2.2.2 Example: impact of reduced sensitivity on the sample size

Δ^2 is inversely proportional to sample size (n), where if we increase the sensitivity then we less sample size and we can detect the bug early also.

2.2.3 Example: filtering users not impacted by the change

Exclude the users which are not impacted by change in the features. This can lead to lower sample size with same power.

2.2.4 The choice of OEC must be made in advance

It should be made in advance and with the corrections.

2.3 Confidence intervals for absolute and percent effect

2.3.1 Confidence intervals for absolute effect

$$CILimits = \bar{O_B} - \bar{O_A} + -1.96 * \hat{\sigma_d} \tag{6}$$

2.3.2 Confidence intervals for percent effect

The percent difference is calculated by =

$$PctDiff = \frac{\bar{O}_B - \bar{O}_A}{\bar{O}_A} * 100\% \tag{7}$$

$$CV_B = \frac{\hat{\sigma_B}}{\bar{O}_B} \tag{8}$$

$$CV_A = \frac{\hat{\sigma_A}}{\bar{O}_A} \tag{9}$$

C.I. for Percent Effect =

$$(PctDiff + 1) * \frac{1 + -1.96 * \sqrt{CV_A^2 + CV_B^2 - (1.92)^2 * CV_A^2 * CV_B^2}}{1 - 1.96 * CV_A^2} - 1 \tag{10}$$

2.4 Effect of robots on experimental results

Robots can skew the estimates, enough to render the assumptions invalid. Even it can have significant effect in A/A test. It is especially important to filter out robots, that interact with the user-id.

2.4.1 JavaScript versus server-side call

It is generally thought that very few robots will be included in the experiment if the treatment assignment is called by JavaScript so those experimental setups shouldn't be affected as much by robots.

2.4.2 Robots that reject cookies

It is recommended to exclude unidentified requests from analysis, so that robots that reject cookies will not be part of experimental results.

2.4.3 Robots that accept cookies

If a robot accepts cookies and doesn't delete them, the effect can be profound, especially if the robot has a large number of actions on the site. It can lead to large standard deviation of many metrics, thus reducing the power.
Thus, it is recommended to remove robots accepting cookies and have large number of actions.

2.5 Extensions for online settings

2.5.1 Treatment ramp-up

An experiment can be initiated with a small percentage of users assigned to the Treatment(s), and then that percentage can be gradually increased. For example, if you plan to run an A/B test at 50%/50%, you might start with a 99%/0.1% split, then rampup the Treatment from 0.1% to 0.5% to 2.5% to 10% to 50%. At each step, which could run for, say, a couple of hours, you can analyze the data to make sure there are no egregious problems with the Treatment as effect has inversely square relationship with sample size.

2.5.2 Automation

Once an organization has a clear OEC, it can run experiments to optimize certain areas amenable to automated search. Multi-armed bandit algorithms and Hoeffding Races can be for such optimization.

2.5.3 Software migrations

Experiments can be used to help with software migration. If a feature or a system is being migrated to a new back-end, a new database, or a new language, but is not expected to change user-visible features. A/B tests can be used to show the difference in software migration, helping identify bugs in the port.

2.6 Limitations

Following are the limitations :-

- Quantitative metrics, but no explanations. - "Why" is not known.
- Short term vs long term effects - Controlled experiments measure the effect on OEC during the experimentation period, typically a few weeks. Long-term should be part of the OEC. Like include some kind of penalty over not used advertisements.

- Primacy and newness effect - These are opposite effects that need to be recognized. It can bring "newness" bias to experienced users. Therefore, it is advised to run experiments only for new users.
- Features must be implemented - The feature may be a prototype that is being tested against a small portion, or may not cover all edge cases. Paper prototyping can be used for qualitative feedback and quick refinements of designs in early stages.
- Consistency - Users may notice they are getting different variant than their friends and family. It is relatively rare that users will notice the difference.
- Parallel experiments - Strong interactions are rare in practice. Pairwise statistical tests can also be done to flag such interactions automatically.
- Launch Events and Media Announcements. - If there is a big announcement made about a new feature, such that the feature is announced to the media, all users need to see it.

3 MultiVariable testing

Consider testing five factors on the MSN homepage in a single experiment. There are two primary benefits of a single MVT vs multiple sequential A/B tests to test the same factors:-

- Test as many factors in a short period of time.
- You can estimate interactions between factors.

There are three common limitations :-

- Some combinations of factors may give a poor user experience - Interaction can lead to worst user experience.
- Analysis and interpretation are more difficult - For an MVT you have the same metrics for many Treatment-Control comparisons (atleast one for each factor being tested) plus the analysis and interpretation of the interactions between the factors.
- It can take longer to begin the test.

3.1 Traditional MVT

Fractional factorial designs that are specific subsets of full factorial designs. These designs are popularized by Genichi Taguchi. The user must be careful to choose a design that will have sufficient resolution to estimate the main effects and interactions that are of interest.

With Five factors we can test it using full factorial, fractional factorial or a Plackett-Burman design.
For K=5, there are $2^5 = 32 user groups are required. K = 3, fractional factorial design exits. Generally, -1 denotes the control and 1 denotes the treatment.$
Plackett-Burman designs can be constructed where the factors are all at two levels with the number of users groups being a multiple of 4, so 4, 8, 12, 16, 20, etc. The number of factors that can be tested is number of groups minus 1.
Similar properties in Plackett-Burman design is also a fractional factorial.
Main goal is to reduce the number of user groups to test main effects and interactions with little or no confounding.
If we want to estimate all two factor interactions with five factors , we will need a fractional factorial design with 16 treatment combinations.

3.2 MVT by running concurrent tests

Fractions of the full factorial are used in offline testing because there is usually a cost to using more treatment combinations even when the number of experimental units does not increase. This doesn't happen with web where users are cheap. A side benefit is you can shut down the factors which are behaving egregious. The experiment that includes the remaining factors is not affected.
Factors not impacting Power :-

- Number of treatment combination cells.
- If sample size is same. It doesn't matter if you are using single factor or many or whether you are conducting eight run MVT.

Factors impacting Power :-

- Number of treatment combi.
- If sample size is same. It doesn't matter if you are using single factor or many or whether you are conducting eight run MVT.