Complimentary to AB testing:

Analyze logs of what users do on the website retrospectively/observationally🡪

Design randomized experiments🡪develop hypothesis

AB testing:

Click through rate: number of clicks/total number of page views(usability)

button

Click through probabilities: unique visitors who click at least once/unique visitors(total impact)

More data

Statistical significance bar/practical significance bar

When we don’t have enough data for modeling

Design:

Low alpha: unlikely to launch a bad experiment

High beta: fail to launch an experiment that did have a difference you care about

1-beta: sensitivity

Dmin=the minimal difference in rate we care about(practical significance level)

Metric:

Sanity check metrics

Evaluation metrics: overall business metrics/detailed experiment metrics

Multiple metrics->composite metric(objective function/overall evaluation criterion): weighted functions combined by multiple metrics

Summary metrics:

Mean, median, percentiles,

How to choose: sensitivity(pick up the change we care about) and robustness(robust to changes we don’t care about), look at the distribution of your metrics:

1. Sums and counts : how many cookies visited the page
2. Distribution metrics: Mean(sensitive to outliers), median(robust but not sensitive), percentiles(bootstrapped inference)
3. Probabilities and rates
4. Ratios: hard to characterize but useful

Assess sensitivity/robustness: retrospective analysis, old experiments, A vs. A experiment(no change, two traffic flows)

Compare the experiment to the control:

1. Difference(absolute change)
2. Relative change(percent change): don’t have to change the significance boundary as the system changes over time

Variability :

CLT, the SE of the metrics

Nonparametric methods:

No assumptions on the distribution

Empirical variability: theoretical variance usually underestimates the variability

Run A/A test to examine the system and detect variability

Bootstrap estimates

Uses of A/A testing:

1. Compare to the results from analytical answers(sanity check)
2. Calculate variability empirically: based on the metrics got from multiple experiments
3. Empirical CI: empirical CI/SD from multi experiments and assume normality

Experiment Design:

Unit: each individual in a test

Experimental/treatment variable

Control variable: influential on modeling, but are not predictor variables(don’t have enough data for modeling). Used to make sure the treatment units are similar to the control units

List of potential variables; test correlation

Lurking variables/confounders

Target variable

Duration**: full cycle** of data reduces the chance of bias in responses

**Unit of diversion**: what an individual subject is in an experiment

User consistency, ethical considerations, variability

If Unit of diversion matches the unit of analysis

1. User id: stable, personally identifiable (user level)
2. Anonymous id: cookie, specific to a browser and a device, can be cleared by users
3. Event: for users non-visible changes

**Population**

**Duration**

**Size**

**Randomized design:** phone/website-based experiments(little control of other variables), high volume

**Matched pair design**: low volume observations, concern for bias, cost per observation is high, unit by unit match, must be set up ahead of time, all units must be known

**Choose a population:**

Inter/intra user experiments

Cohort: define an entering class in the population, harder to analyze, lose data and takes longer, appropriate if we look for the learning effects

Analyze results:

1. Sanity check: invariant metrics(population sizing checks, other metrics that should not change)
2. Family-wise error rate: overall alpha(FWER), too conservative, usually related comparisons all move in the same direction
3. False discovery rate(number of false positive/all rejections): more lenient than FWER

Difficulty in measuring a metric:

Takes too long/ don’t have access to data

Proxy of difficult metrics/ Gather additional data:

Get in-depth information vs. get a larger number of participants

1. User experience research: in-depth
2. Focus groups: get feedback on hypothetical questions.
3. Surveys: get data hard to measure

Filtering the traffic: bias or debias data, build intuition about data

Compute a baseline value for the metric: compute metric on a bunch of disjoint sets

Identify spam and fraud

   
Always calculate the needed sample size ahead of time, and make sure you have at least that many people in your experiment

 Make sure you have enough representativeness in your sample, run it full weeks at a time, at least 2 business cycles

 No or minimal overlap in difference intervals

 Only look at statistical significance (95% or higher) once the 2 previous conditions have been met

A/B Testing

Summary: In an A/B test, a change is applied to a treatment group, and its performance is compared against a control group to estimate the impact of the change.

STEP 1: SELECT A PERFORMANCE METRIC

It’s important to understand the metric used to evaluate the results of the test. Whether the goal is to increase sales, profit, conversion rate, etc., this should be specified at the upfront.

STEP 2: SELECT THE EXPERIMENT DESIGN

Matched pair - when the sample size is small and/or the data is difficult to collect, a matched pair experiment should be used.

Randomized design - when the sample size is large and the data is easy to collect, then a randomized experiment should be used. Randomized experiments are very common for web-based AB tests.

STEP 3: SELECT TREATMENT AND CONTROL UNITS

Each individual in the test is considered a unit. The unit can be a person, store, etc. In a test, units are split into two groups, the treatment group and control group. Treatment and control units are compared against each other

Useful Alteryx Tool: Formula

STEP 4: SELECT EXPERIMENTAL AND CONTROL VARIABLES

Experimental variable - The experimental, or treatment, variable(s) is the variable that is different between treatment and control units. For example, if you are testing a new price point, the experimental variable would be price.

Control Variables - The control variables are the variables that should remain constant between test and control groups. These variables ensure that the treatment and control groups are representative of each other and that the results will apply to the population. Control variables are used to match each treatment unit to one or more control units.

Useful Alteryx Tool: AB Controls

STEP 5: DETERMINE TEST DURATION AND SAMPLE SIZE

These two go hand in hand and contribute most directly to statistical significance. You can improve statistical significance by either increasing the sample size or test duration. Generally the duration of a test should be at least as long enough to capture a representative sample.

STEP 6: RUN THE TEST AND PREPARE THE DATA

Now it’s time to run the test and collect the data. Preparing the data includes filtering for the dates of the test, ensuring there are no duplicate records, removing records with incomplete data, and removing outliers.

STEP 7: ANALYZE RESULTS

Lift - Compare the average performance between the two groups. It can also be useful to understand the distribution of the performance of the units.

Statistical Significance - Performing a t-test provides a p-value. P-values below 0.05, indicate statistically significant results. Paired t-test are used for matched pair experiments and unpaired t-test for randomized experiments.

Impact Estimation - In order to provide an expected impact of broad implementation of the treatment, apply the lift calculation to the entire population.

Useful Alteryx Tool: AB Analysis

To leverage the AB Trend tool, we need at least a full year of data prior to the test, plus an additional six periods (weeks) in our example, to calculate trend and seasonality. Therefore, we need fifty-eight weeks of historical data for trending purposes.

