**Part One A/B test introduction**

1. Product feature test: control group and experiment group
2. Limitations of A/B tests:

* new experience:
* learning effect:

1. change aversion：users hate feature changes
2. novelty effect: users are curious to try out everything new
3. How to detect? Pre-period analysis + Post-period analysis

* other situations: (1) apartment rental website: short-term A/B test cannot collect useful data to make decisions; (2) Camera review website: cannot tell you are missing something; (3) change main logo can be emotional, so not suitable for A/B test; low frequency of user behavior is not suitable for A/B test;
* Complimentary for A/B test: different ways to gather data of users; qualitative method to decide which mountain to climb;

1. Process of A/B test:

* refine the hypothesis;
* choose metrics to use;
* statistics: variations between samples; distribution; confidence interval; Two sample comparison -- pool standard errors
* design the experiment: power analysis (1 - beta), sensitivity (effect size, to decide on the minimum sample size that we need)
* analyze (statistical significance and practical significance);
* make decisions

1. Two sample comparison -- pool standard errors

A close up of a piece of paper

Description automatically generated

1. Click-through-rate & click-through-probability

* Click through rate = number of total clicks/number of page views; eg: 10 users visit the website for 100 times, among these users, 5 clicked on the button for 20 times, so click through rate = 20/100 = 0.2
* Click through probability: number of unique clicks/number of visitors; eg: 10 users visit the website for 100 times, among these users, 5 clicked on the button for 20 times, so click through probability = 5/10 = 0.5

**Part Two Ethics in A/B test**

1. Tuskegee & Milgram experiment & Facebook: bad examples

* Tuskegee experiment: not treat patients properly in order to track the progression of syphilis
* Milgram experiment: to understand the conflicts between obedience to authority and personal conscience; administer electronic shock each time the participants make a mistake
* Facebook: to study how negative posts will affect people’s mood

1. Principles:

* Risk: minimal risks, if exceeds minimal risks, participants’ consent is needed
* Benefits: try to illustrate the benefit of the study
* choice (participants’ choice): do participants have the freedom to choose?
* Privacy: do we need user consent for the experiment?

1. Internal ethical training for employees conducting A/B test

**Part Three Metrics to use for A/B test**



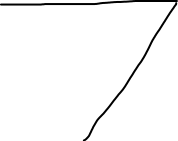
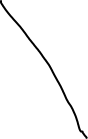
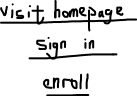
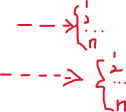
1. Invariant metric:

* Purpose: for sanity check; invariant metrics are supposed to be the same across control group and experiment group, if not, then the setup of experiment is wrong

1. Evaluation metric:

* How to brainstorm evaluation metrics?

1. Start with a high-level concept, for example, business object;
2. expand the funnel and each stage of the funnel can be a metric:



1. take it to smaller granularity
2. Summarize detailed metrics: compound metric;

* What to do when A/B tests not working well: data takes too long to gather the data or do not have access to data

1. Finding external data: companies collecting data you want; companies who run survey on users; academic researches
2. Internal data:

* **User Experience Research (UER)**: the process of discovering the behaviors, motivations and needs of your customers through observation, task analysis, and other types of user feedback; cons: more depth, less participants, validate with retrospective analysis; good for brainstorming
* **Focus group**: A focus group is a gathering of deliberately selected people who participate in a facilitated discussion intended to elicit consumer perceptions about a particular topic or area of interest. Cons: less depth but more participants than UER; pros: get feedback on hypothesis but run the risk of group thinking
* **Survey**: cons: useful for metrics you cannot directly measure, such as whether students attending coursera.com course get a job; pros: people do not always tell the truths in survey
* **Human evaluation**:
* Four categories of metrics to consider

1. Sums or counts
2. Distribution (mean, median, percentiles)
3. Probability and rates
4. Ratios: any two numbers divide by each other

*Note: If the ideal metrics are not accessible due to data constraints, use proxy*

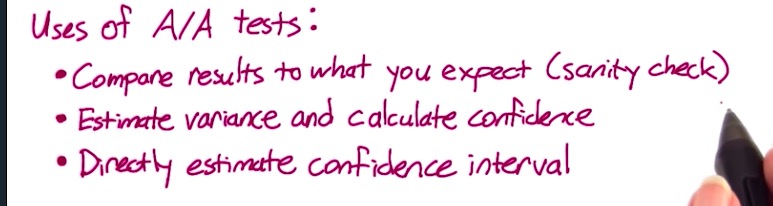
* Principles when choosing evaluation metrics

1. Generality
2. Filter and separate: filter spam or fraud; separate traffic based on different platforms, languages, etc.
3. Time interval is needed
4. Sensitivity: metrics can capture changes in metrics even the changes are tiny

* How to measure: change factors that impact metrics to see if these metrics capture the change

1. Robustness: metrics don’t change when nothing relevant happens

* How to measure: run a A/A test to see if the metrics differ in control group and experiment group



*Note: retrospective analysis can be used to test the validness of metrics too*

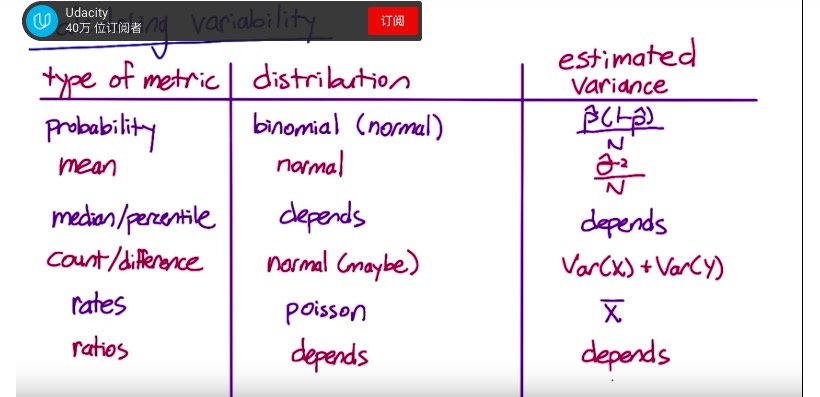
* How to compute comparisons?

1. absolute difference: subtraction between two numbers you want to compare
2. relative difference: ratio of two numbers you want to compare

* why we need to compute variability for metrics and how?

1. Why: because we want to know whether our metrics varies greatly even in normal circumstances, if so, then the metrics are not meaningful for our experiments
2. How to compute?

* Analytical variability and confidence interval



* Empirical variability and empirical confidence interval

1. Use empirical standard error (the standard error of you’re the difference results of A/A test) to replace the analytical standard error
2. Two ways of computing empirical confidence interval: calculate empirical standard error to replace analytical standard error and make assumptions about the distribution of the data; make no assumptions of the distribution of the data, use average difference as the point estimate: simply remove the max and min, and the confidence interval will be (2nd smallest, 2nd largest)

*Note: what if you don’t want to run so many A/A tests? use Bootstrapping, which means repeatedly sample the full sample data and compute the variability of metrics*

**Part Four Design an A/B test**

1. choose unit of diversion

* Definition: the individual subject that you assign to different group of the experiment
* Choices of unit of diversion:

1. User-id: Logins;



1. Anonymous id: Cookies;



1. Event;
2. Device id;



1. IP address;

* Principles of choosing unit of diversion

1. User consistency: users have consistent user experience; cookies and user-id guarantee high level of consistency
2. User visibility: if changes not visible to users, event-based unit of diversion is enough; if changes are visible to users, if users reload the page, it is possible for them to see both changes, so user-ids or cookies should be used instead;
3. Variability: When unit of analytic = unit of diversion, then variability will be low; *Note: Unit of analytic: denominator of your metrics: eg: click-through rate, pageviews is the unit of analytic*
4. Ethics: is new data being collected? Can identity of users be tracked?
5. Choose a population

* A/B test is almost always different people in different groups so called inter-user experiment, paired experiment is not likely to happen in internet
* Restrict your target population

1. Cohorts vs Population: cohorts is defined within a population, meaning people that entering the experiment at the same time; choose cohorts when you need user utility; factors to create cohorts are various, time, region, language

* a subset of behavioral analytics that takes the data from a given data set (e.g. an EMRS, an e-commerce platform, web application, or online game) and rather than looking at all users as one unit, it breaks them into related groups for analysis

1. Reduce the required sample size

* Increase significance level alpha
* Reduce power of analysis: 1 – beta
* Change unit of diversion (if unit of diversion is not the same as unit of analysis)

*Note: Practical significance means the how much change the test detects that will actually convince you to launch that change*

1. Decide on the duration of your experiment

* Definition: How long to run the experiment;

1. Decide on when to run your experiments

* Weekends versus weekdays
* Holidays versus non-holidays

1. Decide on the exposure of your experiment

* Definition: proportions of traffic to send through the experiments;
* Is there any risk if you expose all traffic?

**Part Five Analyze the Results**

1. Sanity check:

* What to check

1. Check population size of treatment group and experiment group to make sure that they are comparable

* How? Run a 2-sample proportions z-test, 50% as point estimate and compute confidence interval

1. Check for invariant metrics to make sure that they don’t change, otherwise the experiment setup is not right

* Why sanity check will fail?

1. Experiment infrastructure
2. Experiment setup
3. Choose inappropriate unit of diversion
4. Data capture issues (most frequently seen)

* What to do if sanity checks fail

1. Retrospective analysis: looks backwards and examines exposures to suspected risk or protection factors in relation to an outcome that is established at the start of the study
2. Look into whether there are learning effects or not
3. Analyze the results

* Single metric:

1. Check if the change is statistically significant; run more experiments
2. if non-parametric sign test not the same with hypothesis tests, beware of Simpson’s paradox

*Note: Simpson’s paradox: a trend appears in several different groups of data but disappears or reverses when these groups are combined; why? unbalanced sample size between two groups; there are different patterns within one group, which will drag the overall performance down*

* Multiple metrics

1. Concerns: Type I error will be high if having more metrics, so the change you observe might due to chance
2. How to address these problems?

* A screenshot of a cell phone

  Description automatically generated

A screenshot of a cell phone

Description automatically generated

1. How to balance the importance of different metrics;

* Use an OEC (overall evaluation criteria) that balance long term and short-term goals

1. How to make a decision:

* Do I have a statistically significant change?
* Do I have a practically significant change?
* Do I understand how the change impacts user experience?
* Is it worth launching the change?

*Note: Ramp-up experiment is a good idea. Ramp-up means gradually include more traffic to your experiments by removing filters, etc.*