Udacity A/B Testing

Introduction

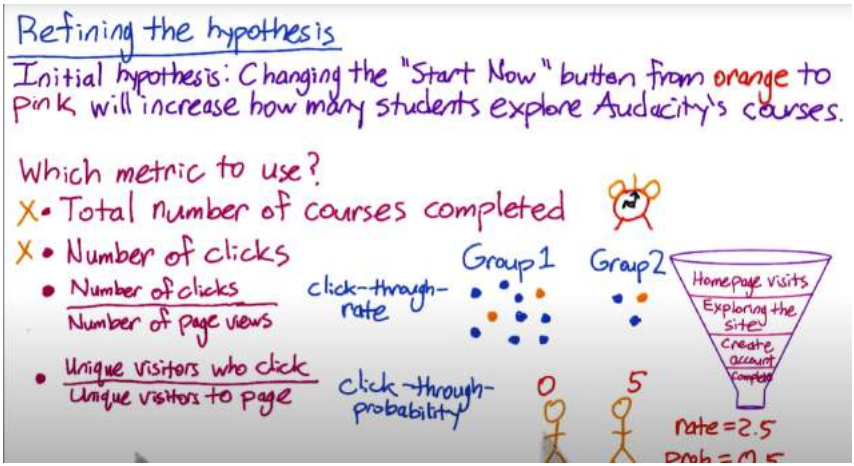
What you cannot do with ab testing

e.g. add premium service- users need to opt-in to the service, cannot do randomly assigning (the willing in control and test group need to be the same)

referral- will take too long

update logo – surprisingly emotional, user need some time to get used to it

Business Example



- Customer Funnel (created based on User Flow)
- CTR (measure Usability) and CTP (measure the impact - work with engineers to merge with trail clicks)
- Calculating Margin of Error (binominal, z- test)

> You can use this z-score table.
> If you don't recall how to read a z-score table, this page contains instructions.

Standard deviation of binomial

If you look up a binomial distribution elsewhere, you may find that it has a mean of np and a standard deviation of $\sqrt{np(1-p)}$.

This is for a binomial distribution defined as the total number of successes, whereas we will use the fraction or proportion of successes throughout this class. In this case, the mean is p and standard deviation is $\sqrt{\frac{p(1-p)}{n}}$.

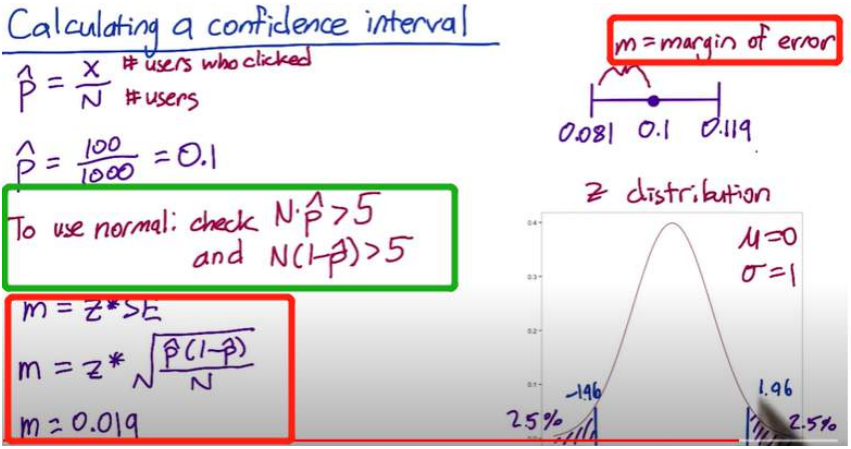
Useful equations

You may find these equations helpful in solving the quiz:

$$\hat{p} = \frac{X}{N}$$

$$SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{N}}$$

$$m = z^* \cdot SE$$



Overview of the Design

Goal – if change the start button's color

Metric – click-through-probability (measure impact) instead of CTR (measure usage)

Review statistics:
Distribution – Binomial Statistics > ^5283fa

Confidence Intervals – Use N and X to calculate P-hat, then use P-hat to calculate SE, then CI

Design:
Hypothesis testing – H0:P control = P experiment , cut-off=0.05

Two-tailed vs. one-tailed tests

The null hypothesis and alternative hypothesis proposed here correspond to a two-tailed test, which allows you to distinguish between three cases:

1. A statistically significant positive result
2. A statistically significant negative result
3. No statistically significant difference.

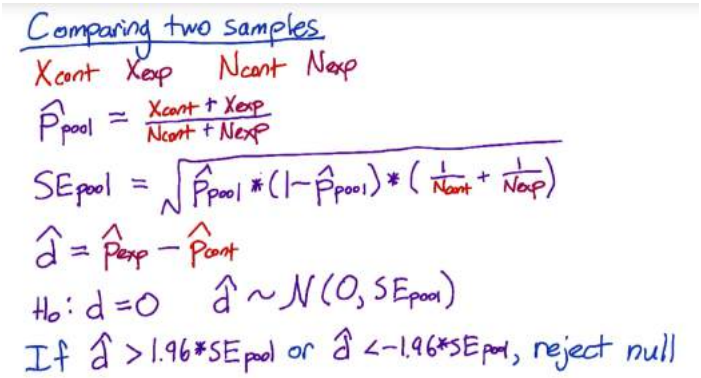
Sometimes when people run A/B tests, they will use a one-tailed test, which only allows you to distinguish between two cases:

1. A statistically significant positive result
2. No statistically significant result

Which one you should use depends on what action you will take based on the results. If you're going to launch the experiment for a statistically significant positive change, and otherwise not, then you don't need to distinguish between a negative result and no result, so a one-tailed test is good enough. If you want to learn the direction of the difference, then a two-tailed test is necessary.

Comparing Two Samples:

Pooled Probability



$$SE_{pooled} = S_p \sqrt{(1/n_1 + 1/n_2)}$$

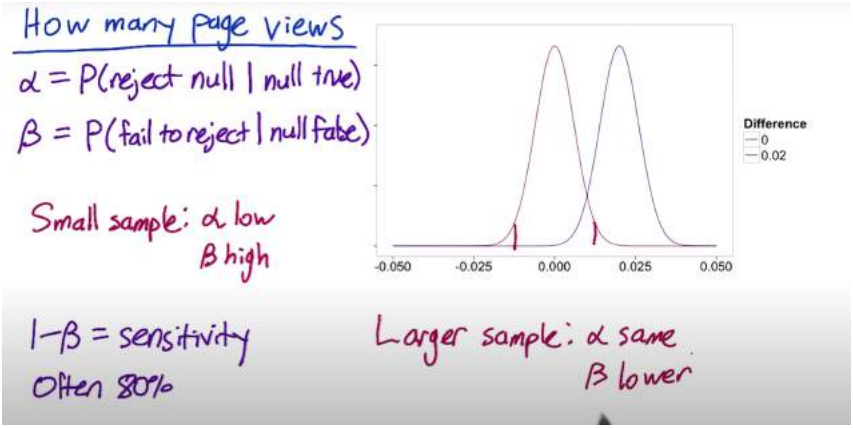
Pick significance boundary- practically significant(business) vs statistically significant (repeatedly)

medicine 10~15% change, for online world, 2% is important

Statistical power- Size vs power trade off

Decide how many page views needed (Sample size) to get a statistically significant result.

The smaller changes you want to detect, or the increase confidence you wanna have in the result which means a larger experiment.
In the well-behaved distribution, such as normal distribution, as the true difference gets larger and larger, Beta will go down. So you typically consider beta at your practical significance boundary (20
5, and 1-beta sensitivity = 80%), since you don't care about any smaller changes and any larger changes will have a lower beta - that is a lower chance of (type II) error.



Calculate Sensitivity:

Baseline Conversion Rate (e.g CTP before make the change)
-当baseline很大（接近1）或者很小（接近0）的时候，实验更容易检

测出差别（beta变小，power变大），如果保持power不变，那么所需要的样本数量变小。

Minimum Detectable Effect (Practical significance level)
-参数越大（比如10%），说明我们期望实验能够检测出10%的差别即可。检测这么大的差别当然比较容易（beta变小，power变大），所以保持power不变的情况下，所需要的样本量会变小。

Calculate result

Policy and Ethics for Experiments

4 principles of IRB's (For Participants)

1. Risk

> The main threshold is whether the risk exceeds that of “minimal risk”. Minimal risk is defined as the probability and magnitude of harm that a participant would encounter in normal daily life. The harm considered encompasses physical, psychological and emotional, social, and economic concerns. If the risk exceeds minimal risk, then informed consent is required.

2. Benefits

> ***what benefits might result from the study***? Even if the risk is minimal, how

might the results help? In most online A/B testing, the benefits are around improving the product.

3. Choice/Alternatives

In online experiments, the issues to consider are what the other alternative services that a user might have, and what the switching costs might be, in terms of time, money, information, etc.

4. Privacy/Data Sensitivity

Identified data means that data is stored and collected with personally identifiable information. This can be names, IDs such as a social security number or driver's license ID, phone numbers, etc. HIPAA is a common standard, and that standard has **18 identifiers (see the Safe Harbor method)** that it considers personally identifiable.

Anonymous data means that data is stored and collected without any personally identifiable information. This data can be considered **pseudonymous** if it is stored with a randomly generated id such as a cookie that gets assigned on some event, such as the first time that a user goes to an app or website and does not have such an id stored.

Choosing and Characterizing Metrics

Two types of checks

· Invariant checking (Sanity): Metrics that shouldn't change between your test and control
e.g is the sample size the same to both control and exp groups? Is sample distribution similar?

· Evaluation: High level business metrics, user experience with the product

==========

How do we go about making a definition of a metric?

1. High level concept of metrics - Overall business metrics (e.g active users, CTR)

2. Details (e.g. how do you define user activity - 80%? or)

3. Summary Metrics (e.g. overall evaluation criterion (OEC) - weighted function combine all the metrics as used by Microsoft)

High-level concept for metrics

some metrics may be difficult to measure due to :

- Don't have access to data
- Takes too long to get the data

Supporting Materials
additional_techniques.pdf

In such case, we can compare with external data or use internal data. With internal data, we can do:

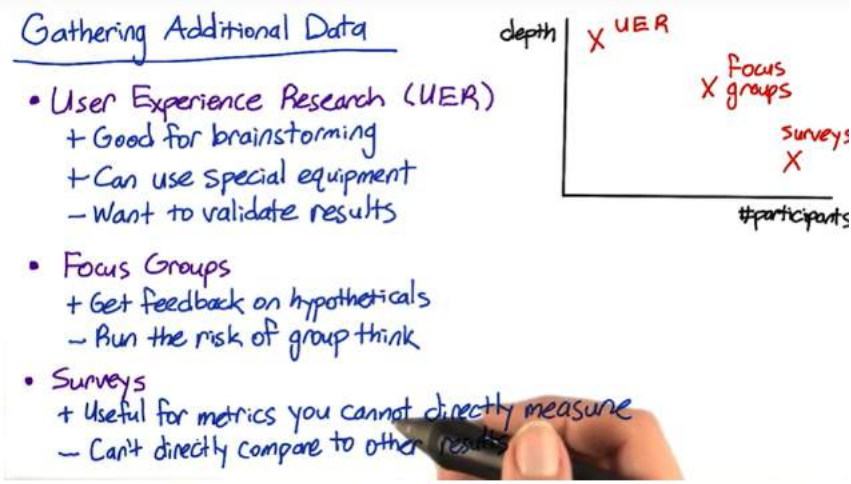
· Retrospective analysis: Look at historic data to look at changes and see the evaluation
· Surveys and User experience research: This helps you develop ideas on what you want to research

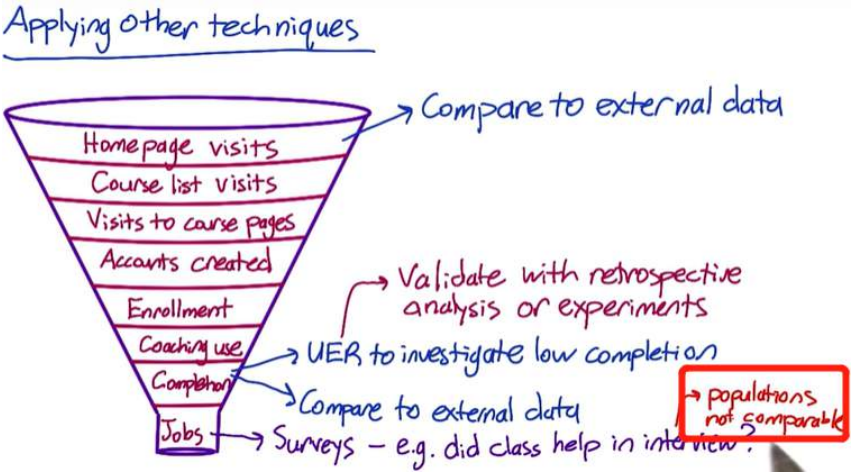
You can gather additional data by:

1. _User Experience Research (UER) - high depth on a few users. This is good for brainstorming. You can also use special equipment in a UER (e.g. eye movement camera) that you cannot use on your website. You may want to validate the results using retrospective analysis_

2. _Focus groups: Medium depth and # of participants. Get feedback on hypotheticals, but may run into the issue of groupthink_

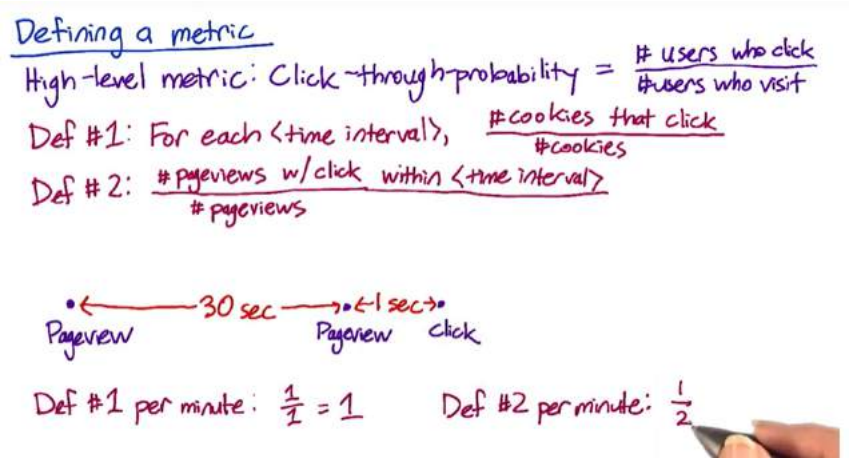
3. _Surveys have low depth but high # of participants: Useful for metrics you cannot directly measure. Can't directly compare with other metrics since population for survey and internal metrics may be different._





Metric Definition

Examples :



> Def #1 (Cookie probability): For each < time interval>, number of cookies that click divided by number of cookies
>
> Def #2 (Pageview probability): Number of pageviews with a click within < time interval> divided by number of pageviews
>
> Def #3 (Rate): Number of clicks divided by number of pageviews

Filter the traffic and Segment

The goal is to filter out those bots and crawler traffic.

To avoid bias, we need to find the ***baseline***. we can do slicing by country, or other categories, and look at the ***traffic pattern changes*** regularly i.e. week over week, day over day, or year over year.

All these are good for evaluate metrics definition and building an intuition on your data and system.

Summary Metrics

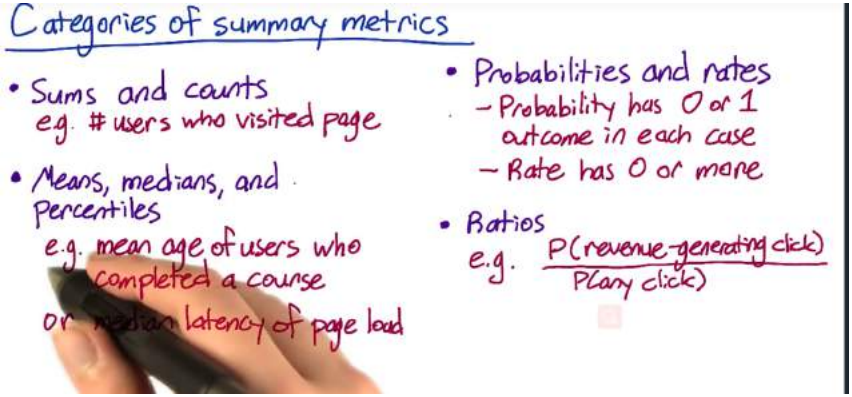
4 categories of metrics

· Sums and counts (e.g. # of users who visited a page)

· Distributional metrics - means, medians, percentiles (do a retrospective analysis, if normal distribution, then mean or median, if skewed distribution, percentiles)

· Probabilities and rates

· Ratios (e.g. revenue/click)



Common distributions in online data

Let's talk about some common distributions that come up when you look at real user data.

For example, let's measure the rate at which users click on a result on our search page, analogously, we could measure the **average staytime** on the results page before traveling to a result. In this case, you'd probably see what we call a **Poisson distribution**, or that the stay times would be **exponentially distributed**.

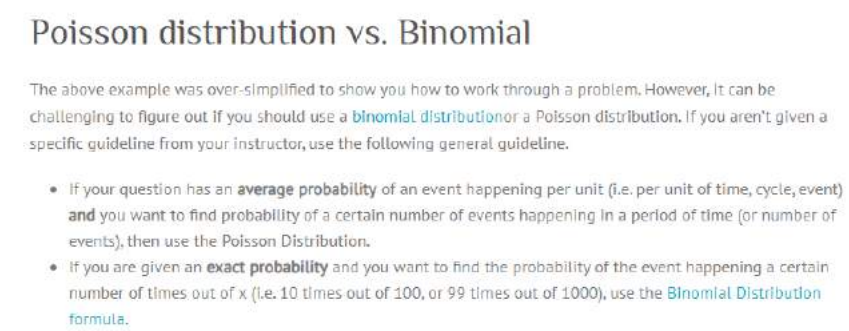
Another common distribution of user data is a "power-law," **Zipfian or Pareto distribution**. That basically means that the probability of a more extreme value, z, decreases like 1/z (or 1/z^exponent). This distribution also comes up in other rare events such as the frequency of words in a text (the most common word is really really common compared to the next word on the list). ****These types of heavy-tailed distributions are common in internet data.

Finally, you may have data that is a composition of different distributions - latency often has this characteristic because users on fast internet connection form one group and users on dial-up or cell phone networks form another. Even on

mobile phones you may have differences between carriers, or newer cell phones vs. older text-based displays. This forms what is called a **mixture distribution** that can be hard to detect or characterize well.

The key here is not to necessarily come up with a distribution to match if the answer isn't clear - that can be helpful - but to choose summary statistics that make the most sense for what you do have. If you have a distribution that is lopsided with a very long tail, choosing the mean probably doesn't work for you very well - and in the case of something like the Pareto, the mean may be infinite!

Reference:



Poisson distribution vs. Binomial

The above example was over-simplified to show you how to work through a problem. However, it can be challenging to figure out if you should use a binomial distributionor a Poisson distribution. If you aren't given a specific guideline from your instructor, use the following general guideline.

- If your question has an **average probability** of an event happening per unit (i.e. per unit of time, cycle, event) **and** you want to find probability of a certain number of events happening in a period of time (or number of events), then use the Poisson Distribution.
- If you are given an **exact probability** and you want to find the probability of the event happening a certain number of times out of x (i.e. 10 times out of 100, or 99 times out of 1000), use the Binomial Distribution formula.

Sensitivity and Robustness

e.g. mean is sensitive to outliers so not robust, median is robust but not sensitive to changes to small group of users.

This can be measured by using prior experiments

(assume the videos are comparable)

Another alternative is to do A/A tests to see if the metric picks up any spurious differences

(assume 5 has the highest resolution)

Absolute vs Relative Differences

The simplest way to compare metrics for test and control is to take a difference.

If you are running a lot of experiments you want to use the relative difference i.e the percentage change. The main advantage of computing the percentage change is that you only have to choose one practical significance boundary to get stability over time. If you are running a lot of experiments over time, your metrics are probably changing over time. Using relative difference helps here by having to use one practical significance boundary rather than change it as the system changes.

The main disadvantage is variability, relative differences such as ratios are not as well behaved as absolute differences, so if you do not

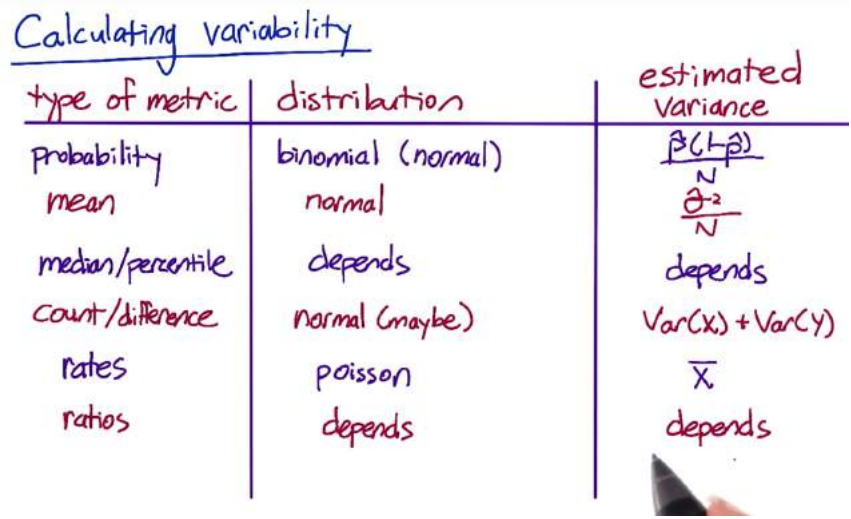
understand the metrics well, may start off absolute differences.

Calculating variability

To calculate the confidence interval, you need

· Variance (or standard deviation)

· Distribution



Depends here means depends on distribution of original data, if the data is too complex, may not be able to get analytical estimate of variance.

Non-parametric methods

Empirical Variability

通常对样本进行统计分析的时候，首先要假设他们来自某个分布，然后用样本中的数据去estimate这个分布对应的参数，之后再做一些test之类。比如你假设某个样本来自同一个正态分布，然后用样本数据估算和，再用估算出来的这两个值做test。non-pararmetric则不然，不对总体分布做假设，自然也就不必estimate相应的参数。一个比较简单的例子是Sign test，可以用来检验两个分布X和Y的中位数是否相等。在这里不必介意X和Y分别是什么分布，只在意样本中每对x, y的相对大小。如果中位数相等（），那么 P(X>Y) = 0.5，所以去除刚好相等的情况之后，x > y 的样本数量应该是符合二项分布的。像这样就可以在在不对X和Y的分布做任何假设的情况下检验X和Y的中位数是否相等。

A/A Test - both are control group then measure differences which indicate underlying variabilities. The underlying variability may caused by your system, or user population etc.

At Google, it was observed that the analytical estimates of variance was often under-estimated, and therefore they have resorted to use empirical measurements based on A/A test to evaluate sensitivity and robustness, and variance.

If you see a lot of variability in a metric in an A/A test, it is probably too sensitive to be used. Rather than do several multiple A/A tests, one way is to do a large A/A test, and then do `#bootstrap` to generate small samples and test the variability.

With A/A tests, we can do:

- Compare result to what you expect (sanity check)
- Estimate variance empirically and use your assumption about the distribution to calculate confidence
- Directly estimate confidence interval without making any assumption of the data

Empirical Confidence Interval:

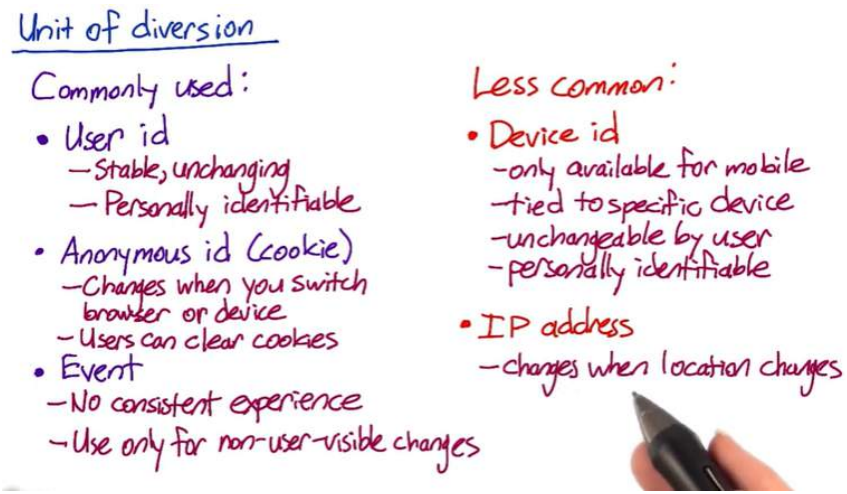
Second large difference, and last second difference from result of 40+ AA testing experiments or one large AA testing with 40+ bootstrapping samples.

Designing an Experiment

Choose Subject

Unit of Diversion

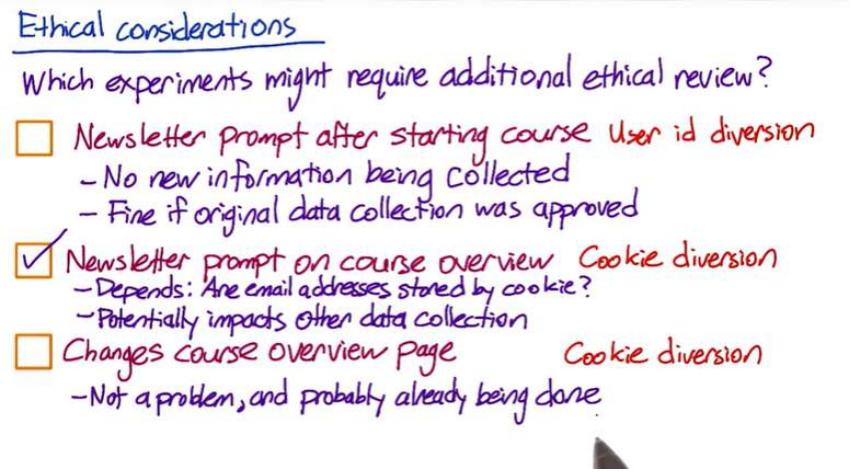
User-id based, cookie-based, event-based (such as query, or invisible changes for users)...etc.



Considerations of choosing an unit of diversion

1)User consistence –users visibility or Measure learning effect (if load is slower, will users use less, need to see across time)

2)Ethical consideration



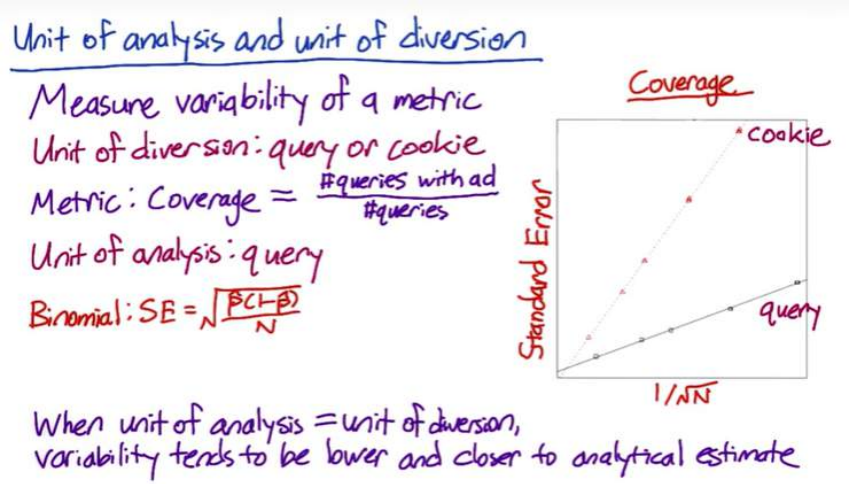
Chose two because the anonymous-based diversion such as cookie diversion here could be linked with email addresses and become identifiable.

3)variability – empirical variability may be much higher than analytical variability.

Because when unit of analysis (the denominator of metrics) is different from unit of diversion. E.g. unit of analysis is pageview and

unit of diversion is cookie.

When you're doing event-based diversion, every single event is different random draw. When you're using cookie- based diversion, you are doing grouped events, they are correlated, which increase the variability.



Standard error graph

The graph Caroline discusses is Figure 4 in [this paper](#).

Options: Inter- vs. Intra-User Experiment

User-Id, cookie -- these are proxy for users; while event-based diversion may contain users from both sides (control and test), so have to be careful

Intra-user experiment:

Expose same users to the feature on/off over time, and you actually analyze how they behave in different time window.

Pitfalls: Risk caused by choosing different time windows; learning effects

Inter-user experiment:

Ranked-order list based (Such as Preference, search ranking), expose same users at A/B side at the same time, and typically this worked in the cases of reordering a list.

AB testing is a kind of inter-user experiment, with different users at different groups

Target Population and Cohort

Choose population: What population are you going to use (US only?)

Size - unit of diversion 不同， size 不同

Targeting the experiment to the appropriate traffic can reduce the size of experiment

But we need a pilot first to determine which cohort will be impacted mostly

#Cohort people who enter the experiment at the same time; harder to analyze, use basically only when you measure user stability, otherwise stay with population.

Sizing

iterative process.

Sizing calculation code

You can find the R code Caroline used to calculate the experiment size in the Downloadables section.

Supporting Materials

- empirical_sizing.R

Duration

How long to run the experiment, when to run?

The duration of the experiment is related to the proportion of how many traffics you sending to your experiment. But also, weekends, holidays different time will also affect traffics.

Handling Learning Effects:
It takes time -

- Chose correct unit of diversion, need stable unit of diversion to test in long run
- Using a cohort instead of just population

- Risk management - run on small proportion of users for a long period of time

--- requires advanced techniques and experience

Analyzing result

Sanity check

Two main types:

1. Population sizing invariant metrics, based on your unit of diversion, to make sure control group population and exp group population are actually comparable
2. Invariant metrics

Pre-period and Post-Period experiments, not only can be used to measure learning effects, but also can be used in sanity check prior to analyze the results. (also to debug with engineers)

Checking Invariants

checking population sizing invariant:
For example,
if we chose cookie as unit of diversion, cookies are going to control group or exp group are randomly selected; therefore, it should be a binominal distribution, with a large sample size, we can assume it as normal distribution.
P = 0.5 ; SE = sqrt(0.5 *0.5 / (Ncont + Nexp)), m = SE*1.96 (95% confidence level), IC = 0.5(True mean - theoretically) +- m
Then check, if P-hat = Ncont / N-total falls in the IC at 95% confidence level.

Single Metric

What NOT to do if your results aren't significant

Carrie gave some ideas of what you can do if your results aren't significant, but you were expecting they would be. One tempting idea is to run the experiment for a few more days and see if the extra data helps get you a significant result. However, this can lead to a much higher false positive rate than you expecting! See this post for more details. Instead of running for longer when you don't like the results, you should be sizing your experiment in advance to ensure that you will have enough power the first time you look at your results.

When metrics use CTR as a high-level concept, Cookie as an unit of diversion

Calculation Process

1. Sanity Check - checking invariants
2. Effect Size Test:

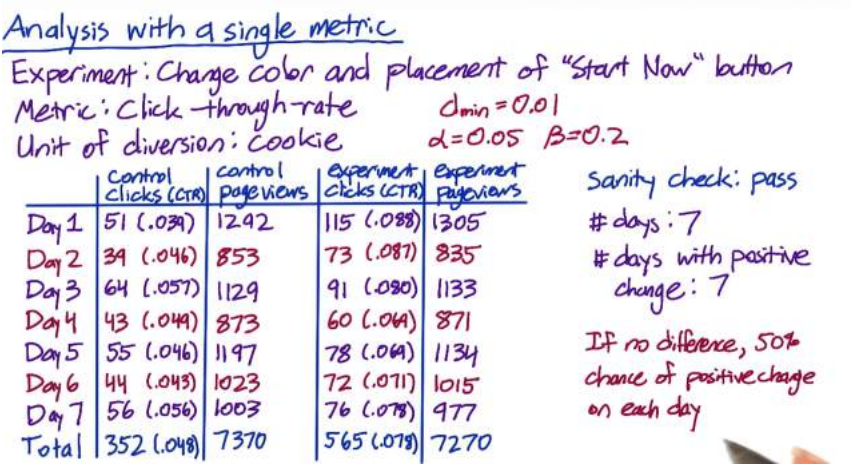
- calculate estimate difference d-hat
- Calculate SE by using empirical SE (ideally Udacity AB Testing Notes > ^f3f7bd)
  (Because CTR distribution is more like Poisson other than binomial)
- Compare if 0 is included in the CI (if it is not included, statistically significant), and if d-min (Practical significance boundary) is included in the CI, then determine if the change is what we care about practically.

3. Sign Test (non-parametric) to double check: (check on day-to-day data, if d-hat >0), how frequent it is d>0 in past seven days, check if it is random by using the online calculator:this online calculator

Sign test has lower power than effect

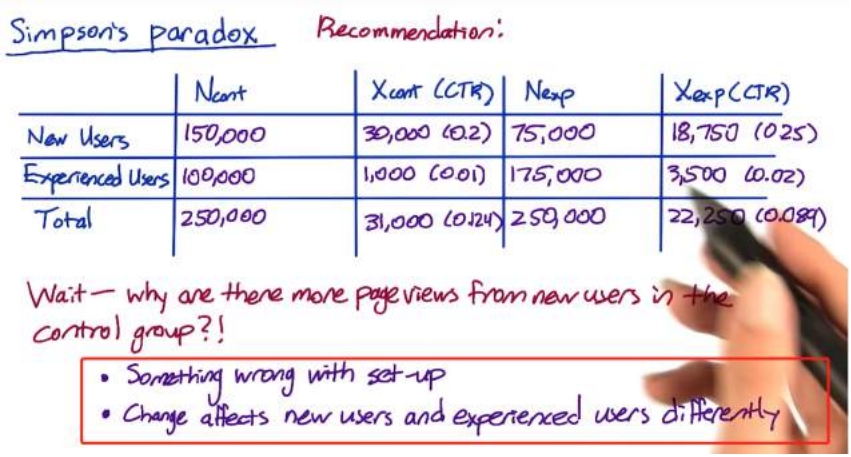
size test --- that's the pay for non-parametric test that you are not making any assumptions about distribution

Sign test can be an alarm that need you dig deeper on your data



	control clicks (CTR)	control pageviews	experiment clicks (CTR)	experiment pageviews
Day 1	51 (.039)	1242	115 (.088)	1305
Day 2	39 (.046)	853	73 (.087)	835
Day 3	64 (.057)	1129	91 (.080)	1133
Day 4	43 (.049)	873	60 (.069)	871
Day 5	55 (.046)	1197	78 (.069)	1134
Day 6	44 (.043)	1023	72 (.071)	1015
Day 7	56 (.056)	1003	76 (.078)	977
Total	352 (.048)	7370	565 (.078)	7270

Simpson's paradox



e.g. new users - refresh page more than experienced users when we use user-id as unit of diversion, and changed the environment

Multiple Metrics

Multiple Comparison (FWER, FDR...etc.)

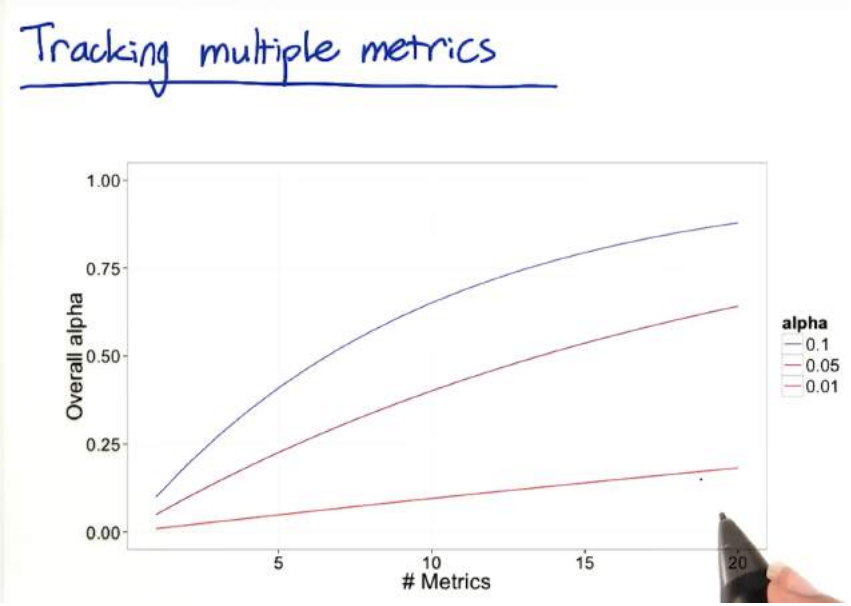
Multiple comparisons this article.

The more things you test, the more likely you are to see significant difference just by chance. This is a problem, but since it is not repeatable for the same metric across multiple attempts, there is a way out. One can do multiple runs of the experiment, or alternately bootstrap. There is another technique called multiple comparison that adjusts your significance levels that accounts for how many metrics or tests you are doing.

False Positive Rate (Type I Error) of at least 1 metrics

As showed below, as number of metrics increase, the P(FP>=1) will

increase too.



Tracking multiple metrics

Problem: Probability of any false positive increases as you increase number of metrics

Solution: Use higher confidence level for each metric

Method 1: Assume independence

$\alpha_{overall} = 1 - (1 - \alpha_{individual})^n$

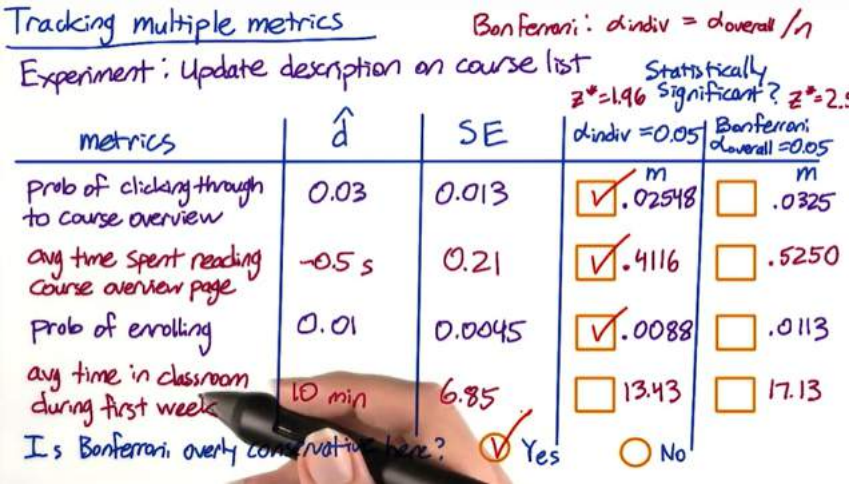
Method 2: Bonferroni correction

- simple
- no assumptions
- conservative – guaranteed to give $\alpha_{overall}$ at least as small as specified

$\alpha_{individual} = \frac{\alpha_{overall}}{n}$

$\alpha_{overall} = 0.05$

$n = 3$ $\alpha_{individual} = 0.0167$



Less conservative multiple comparison methods

The **Bonferroni correction** is a very simple method, but there are many other methods, including the **closed testing procedure**, the **Boole–Bonferroni bound**, and the **Holm–Bonferroni method**. **This article** on multiple comparisons contains more information, and **this article** contains more information about the **false discovery rate (FDR)**, and methods for controlling that instead of the familywise error rate (FWER).

Judgment calls are needed as well for lots of situations in multiple comparison.

About OECUdacity AB Testing Notes > ^697fc2, we come up with an OEC, but we never actually using OEC to make a final decision, because

we also wanna know what specific metrics have changed so that to have a better understanding (also retrospective analysis could help here as well); but it is good to have OEC, so that we know what metrics/business we actually care about.

When you analyze the result here are few basic questions to ask yourself:

- do I have statistically significant and practical significant result in order to justify the changes?
- do you understand the changes?
- do you want to launch the changes?
  The final goal is - did you recommend to launch or not, and what's the right recommendation to the business questions asked at the beginning.

Ramp up percentage of your traffic (removing filters - i.e. country, platform...etc.) until your feature is fully launched

Gotcha: The launch effect may be flatten as ramping up on your users.

(due to, for example, seasonality)
Solution: add holdbacks on certain groups of people, continue comparing their behavior to control...track it overtime, until you are confident that result will be repeatable, and it will help a lot in capturing seasonal or even-driven impacts.

(or, due to - changes on some human behaviors)
solution: pre- and post- periods design + cohort analysis, will be helpful.