Lesson 1

Introduction of A/B Testing

问题一：什么是 A/B Testing

A/B Testing can be used to test new features. You have two sets of users, one is control set who can only see existing product, the other set is experiment set, who can see your new product. A/B testing is to see how users respond differently and help you to decide which kind of feature is better.

A/B Testing is helpful for you to climb to the peak of the mountain, but if you want to know whether you are in this mountain and in the other mountain, it is not useful.

理解：AB Testing 能够帮助我们去研究新的且基于原有框架改进的变动是否会提高 performance，但是不能比较一个新的框架和旧的框架哪一个更好。换句话说对照组必须存在。

问题二：A/B Testing 可以做什么

You can test new features, different look of your website etc.

e.g.1: User Visible Changes
Amazon first customer recommendations, and they can see significantly revenue changes after using AB Testing.
这里的实验组就是得到推荐系统推荐的用户，对照组就是原来的未接触推荐系统的用户，比较两者之间的消费差异。

e.g.2: User invisible changes
LinkedIn shows whether news article or an encouragement can basically add new contacts.
这个是属于 ranking changes， 是新闻文章还是鼓励更能够增加用户的联系

e.g.3: User may not notice
Amazon 2007 proved that every 100 millionsecond they added to the page, 1% decrease of revenue.
这种 AB Testing 是用户几乎难以注意到的

问题三： A/B Testing 不可以做什么

Conclude: A/B Testing isn’t as useful testing out new experiences

A. We need to take time into consideration!

- changer version: I like the previous one, I don’t want it to be changes.
- novelty effect: Oh, It’s new! I like new items!

In AB testing, we need to know what’s our baseline of comparison. We also need to know how much time we need in order for the users to adapt to the new change so that we can actually say what is going to be the plateaued experience so that we can actually make a robust decision.

B. AB Testing cannot really tell you if you are missing something
e.g. If you have a digital camera review site, AB Testing can show you this camera review should above that one. But it can’t tell you what will happen if you miss entire other camera review you should be reviewing but you aren’t

理解：

这里我认为 new experience 有两层含义

- 时间上的新， 即由于 changer version 和 novelty effect 两种现象的存在，我们不能马上确定新的变化会如何影响。比如房地产租赁 APP，你设置了部分用户可以 refer 其他人，但是由于本身浏览这类 APP 的人相对较少，这个作用可能要 3-6 个月才能体现出来。
- 产品或内容的新， 即如果缺少或新增了一个完全没有的框架或内容。比如我们有一个相机评论网站，我们可以用 AB Testing 来做哪一个评论放在上面更好，但是它不能告诉你如果你完全缺失某一个相机的评论（一个你应该有但你没有的）会怎样。这个就是结构上的新

练习：以下哪些情形可以使用 AB Testing
1. Online shopping company: is my site complete?
不能，因为只能 try specific product 但是不能回答整个问题
2. Add premium service
不能，因为用户需要自己选择要不要 premium，随机分配没有意义，所以缺乏对照组。但是 AB Testing 这里确实可以帮助我们获取一些信息，比如用户点击量，多少用户对这个感兴趣等。
3. Movie recommendation site, new ranking algorithms
可以
4. Change back-end page load time, results etc.
可以
5. Website selling cars: will a change increase repeat customers or referrals?
不能， take too long time and no data to see whether they recommend the site to their families and friends.

6. Update brand, including main logo
不能， surprisingly emotional
7. Test layout of initial page
可以！

Other Techniques

在遇到 AB Testing 不能解决的问题时有两种方案

- 方案一：运用其他的数据来进行回溯分析（retrospectively）或其他分析然后看看能否构建假设检验，来研究是什么来导致用户的变化等等。最终可能会演变成一个 AB testing，或作为其结果的重要辅助
- 方案二：直接采用其他分析方式，user experience research, focus groups, surveys, human evaluations. 一般 AB Testing 会给你大量的 broad quantitative data， 而其他方法会给你 deep and quanlitative data。

Which Metric to use?

Metric 1:
click through rate = Number of clicksNumber of page views
Metric 2:
click through probability = Unique visitors who clickUnique visitors to page
二者的区别

- rate: When you want to measure the usability of the site --- how often users find that button
- probability: When you want to measure the impact --- how often people went to the next page on your site.

procedure:

1. work with engineer to change your website
2. capture data
3. compute metrics

Binominal Distribution

Three key elements:

- 2 types of outcomes
- independent events
- Identical distributions (p for all)

exercise:

1. drawing 20 cards from a shuffled deck (outcome red and black) No different probability
2. roll a die 50 times (outcome 6 or other) Yes

3. clicks on a search results page (outcome click or not) No (not independent, didn’t get result and change words to search again)
4. student completion of course after 2 months.(complete or not) Yes
5. Purchase of items within one week(outcomes purchase or not) No different probability

Confidence interval & Hypothesis Testing

miu: p
stddev: p(1-p)N

Hypothesis Testing

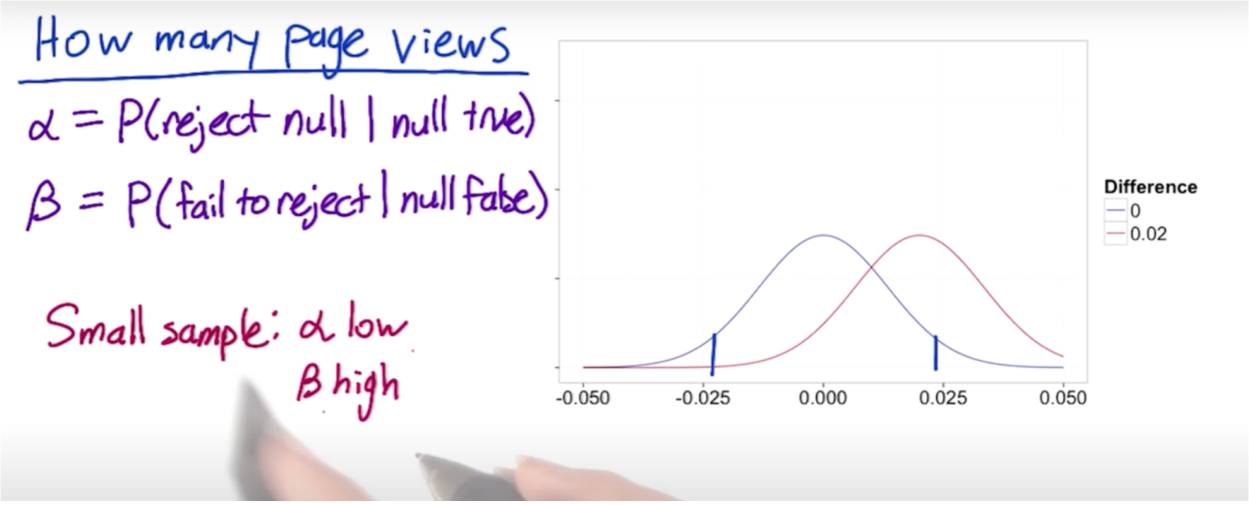
- We need to know what Null hypothesis is, what alternative hypothesis is
- compare two samples, compute confidence interval, decide whether it’s in it
- define a practical significant bar to see whether it’s practical

Ppool = Xcoun+XexpNcoun+Nexp, SEpool=Ppool*(1-Ppool)*(1Ncon+1Nexp)
d = Pexp - Pcont
H0:d = 0, d N(0, SEpool)
If d > 1.96*SEpool or d < -1.96*SEpool reject
Here, we use hypothesis testing to see whether it’s likely that the result we get, the difference we observed could have occured by chance or if it would be extremely unlikely to have occurred if two sides were actually the same.

However, in business perspective, what change in the metrics should be both statistically significant and practically significant. Because we also care about what size of the change really matters to us, we want to make it clear that it worths, so, we need a higher level of practical significance to justify. (Note that in different cases, magnitude of this change could be quite different!)

I practice, we want to observe repeatability. Thus statistical significant bar is lower than practical significance bar.

How to decide the size of the experiment
Here is an inverse trade off, the smaller change that you want to detec or the increased confidence that you basically want to have in the result, means that you have to run a larger experiment.



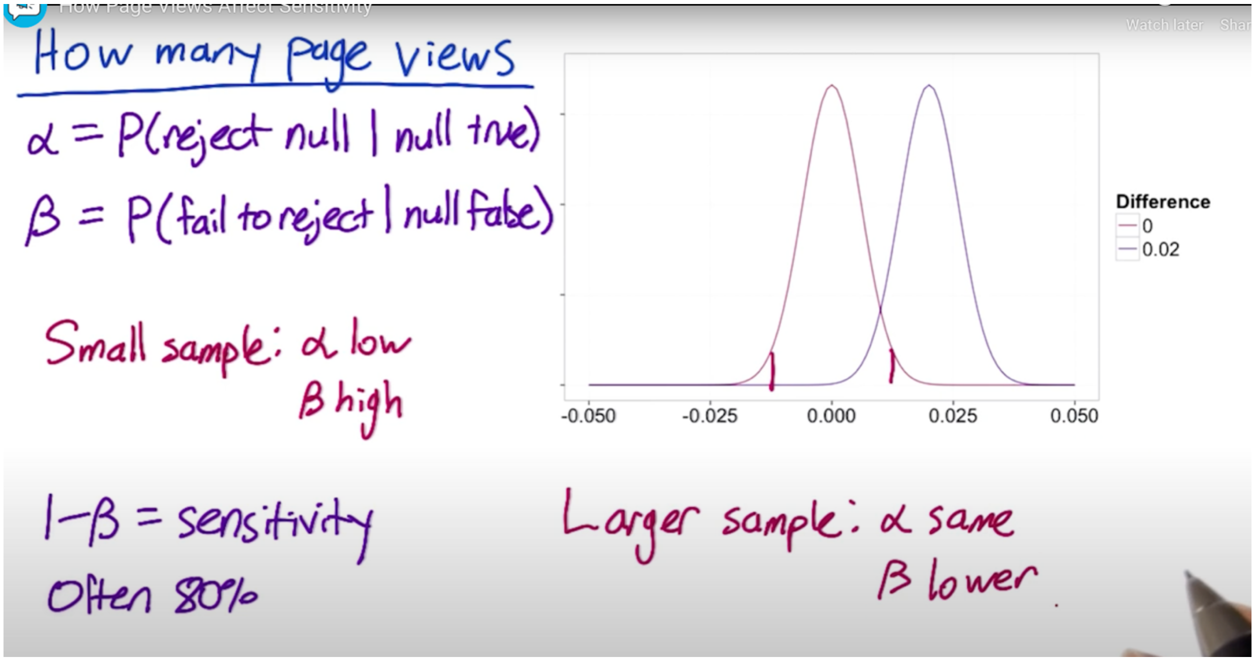
How many page views

$\alpha = P(\text{reject null} \mid \text{null true})$

$\beta = P(\text{fail to reject} \mid \text{null false})$

Small sample: α low

β high



How Page Views Affect Sensitivity

How many page views

$\alpha = P(\text{reject null} \mid \text{null true})$

$\beta = P(\text{fail to reject} \mid \text{null false})$

Small sample: α low

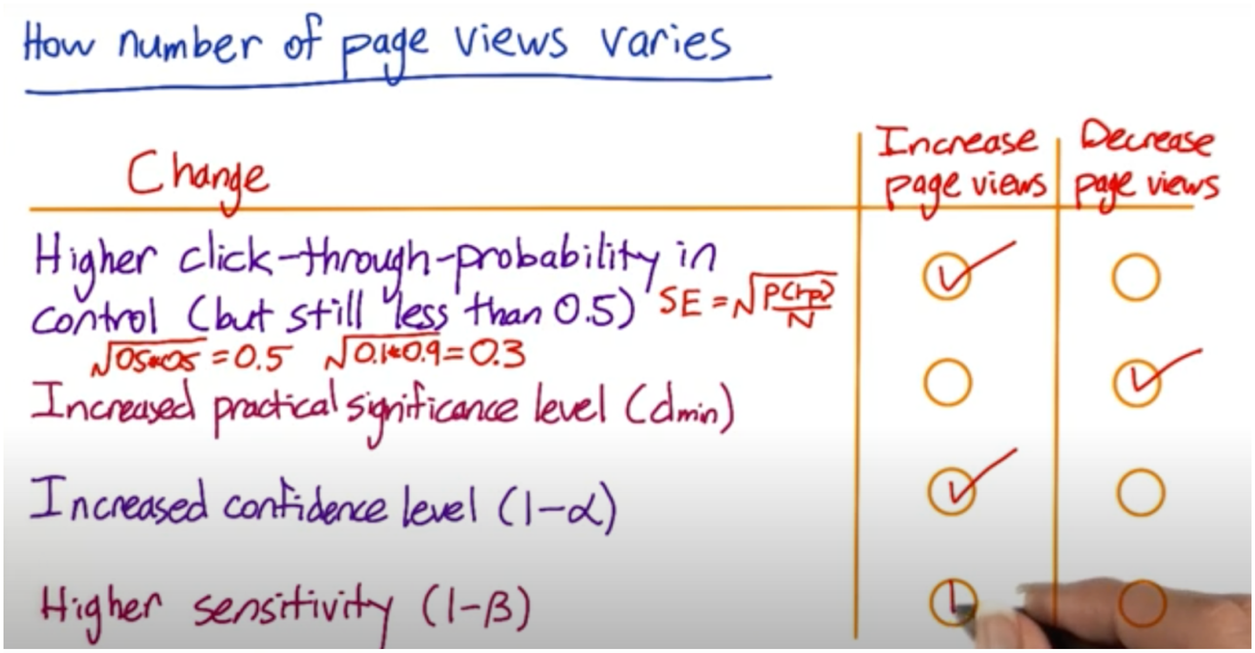
β high

$1-\beta$ = sensitivity

Often 80%

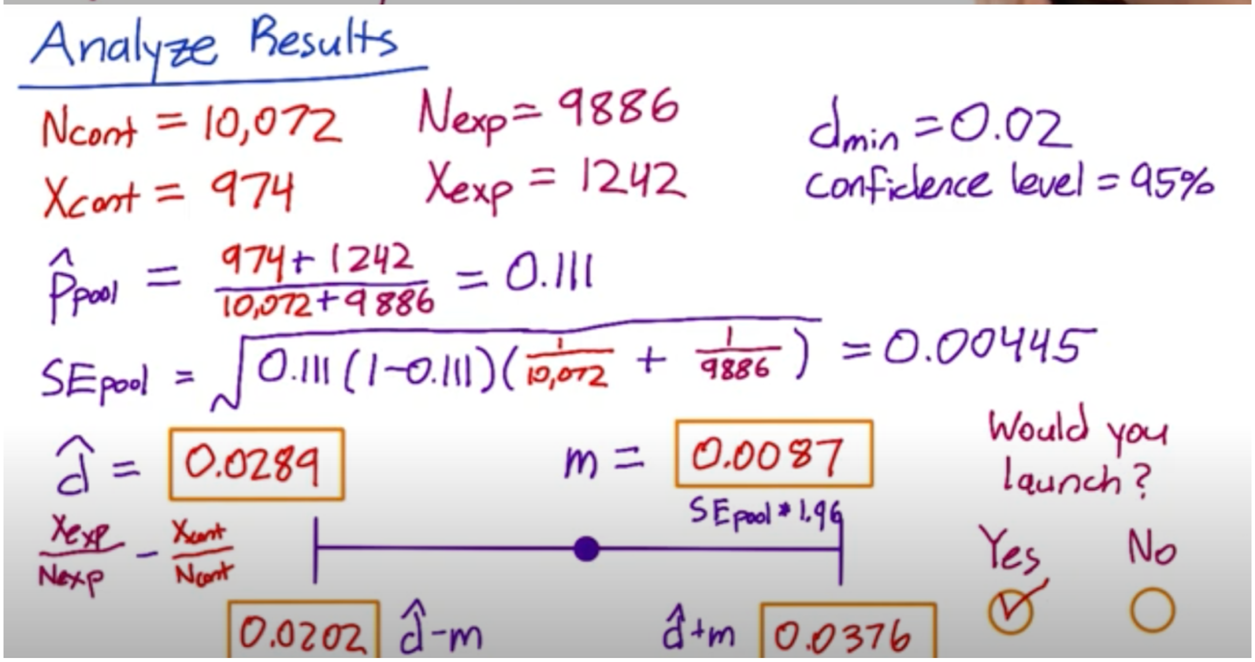
Larger sample: α same

β lower



How number of page views varies

Change	Increase page views	Decrease page views
Higher click-through-probability in control (but still less than 0.5) $SE = \sqrt{\frac{p(1-p)}{N}}$ $\sqrt{0.5*0.5} = 0.5$ $\sqrt{0.1*0.9} = 0.3$	✓	○
Increased practical significance level (d_{min})	○	✓
Increased confidence level ($1-\alpha$)	✓	○
Higher sensitivity ($1-\beta$)	○	○



Analyze Results

$N_{cont} = 10{,}072$ $N_{exp} = 9886$ $d_{min} = 0.02$

$X_{cont} = 974$ $X_{exp} = 1242$ confidence level = 95%

$$\hat{p}_{pool} = \frac{974 + 1242}{10{,}072 + 9886} = 0.111$$

$$SE_{pool} = \sqrt{0.111(1-0.111)\left(\frac{1}{10{,}072} + \frac{1}{9886}\right)} = 0.00445$$

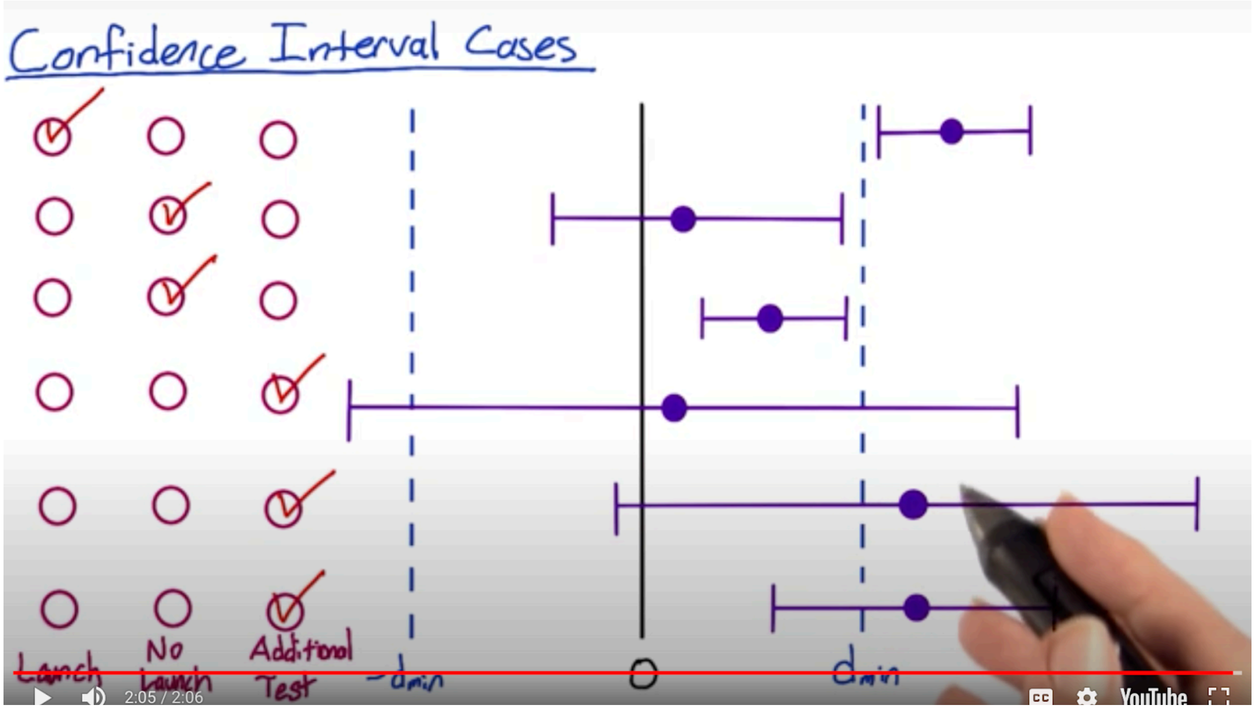
$$\hat{d} = \boxed{0.0289} \qquad m = \boxed{0.0087}$$

$\frac{X_{exp}}{N_{exp}} - \frac{X_{cont}}{N_{cont}}$ $SE_{pool} * 1.96$

$\boxed{0.0202}$ $\hat{d} - m$ $\hat{d} + m$ $\boxed{0.0376}$

Would you launch?

[✓ Yes] No ○



What if the last three cases happens and you don’t have time to take additional test?

- talk to decision maker and talk about data uncertainty
- they should use other factors such as strategic business issues.