**Ivy：该文档来自**<https://www.1point3acres.com/bbs/thread-801430-1-1.html>

整理自Udacity的ABTesting的课程 三个老师都是google资深程序员和统计学家。这是我的笔记，分享给大家！

**一、design problem**

  When we conduct random sampling,

         how to avoid network effect?

         What is network effect (interference)?

An experiment to give a better engagement. User A and User B are friends. User B is assigned to test group while User A is at control group. User B has a better experience that make him engage more. And then his engagement leads to a better engagement for user A.

解决办法：A cluster-based experiment, where a whole cluster (i.e., community) of users is either in treatment or in control. In other words, if I am treated, a significant proportion of my connections are also treated. If I am part of the control group, a significant proportion of my connections are also under control. （依照community来安排分组， 同一个cluster 都在一个组里）

**2.  What will you do if your experiment take too long to run?**

Increase exposure from 10% to 20%.

Based on the formula of sample size in

Question 2:

Reduce the variance by Blocking –run experiment within sub-groups ..

3.What is your roll-out plan? 90-10. Starting with a small portion of their total audience and slowly increasing the sample size to a higher percentage of the audience until finally reaching 100%.

**二、implement problem**

1. What if your data is highly skewed or statistics is hard to approximate with CLT?. 1point 3 acres

Bootstrap. Shortage computational expensive.

2.Can you run an experiment and keep reading until the result is significant?No. Highly increase the type 1 error (false positive rate)

3.What is the chance of seeing at least one rejection having 10 tests simultaneously? follow up last question: 1-(1-0.05)^10 .

4.What if you have multiple test groups? multivariate? Multivariate experiment will increase the type 1 error.

Bonferroni Correction: Assume we have m test; Set significant level for each test = 0.05/m (Default significant level =0.05) FDR

Adjustments(Benjamini-Hochberg): rank p value of m test from low to high. find the largest k such that p\_k<=k/m\*alpha; reject 1…..,k, accept K+1……

5.What is bootstrap? Boosting?

A resampling method with replacement. It can be used to estimate sampling distribution of any statistics, commonly used in estimating CI & p-value & statistics with complex or no close-form estimator

6.The metrics of interest is 90th quantile of users‘spending. How to estimate sample mean and variance?Use bootstrap.. ----

**一、整体思路/一些基础知识和概念的先导（详细和重点从第三部分开始）**

1.metric： click through rate or click through probability? 目的不同 rate看button的作用（比同一个页面的button？how easy to find the button计算上直接click数/pageview数），probability看整体效果（到了funnel的下一步/通常是user或cookie维度）. check 1point3acres for more.

2.statistics：binomial distribution 二项分布 （success and failure） 计算confidence interval：用normal distribution算

3. hypothesis testing -baidu 1point3acres

4. pooled standard error： d服从个啥分布？

5.statistical significance is about repeatability. 但是也要business上practically significant(这个effect的程度要有意义)。并且statistical significance（α：if H0 is ture, the probability you happen to observe effect in your sample）和sample size有很大的关系：if the sample size is large enough, even tiny differences from the hypothesis value will be found statistically significant。sample size给了很大可信度。所以，你在size你的experiment的时候要确保：statistical significance bar要lower than practically significant bar(为啥？stats的要更加严格？比如business要看见ctr有2%的change算是significant，因为越小的change越难observe到，那么希望用statistical看到的就的change就要越小，比如1%。所以由此在决定：我想看见1%的change，我想显著性α=5%，我该用多大的sample size)

6. statistical power vs size trade-off. (The smaller change you want to detect, or the increased confidence that you want to have in the result, means that you need a larger experiment). The larger the sample size, cut-off for rejecting the H0 will be closer to zero. 下图：practical significant = 0.02，这种情况下you fail to reject the null，你认为无significant difference。β（type2 error/1-power)这时就很高，你会错过你care的difference。而增大sample size在不改变α的前提下，可以降低β。any larger changes（than your practical significant boundary) will have a lower β(你会更容易detect到这个significant difference) .1-β又叫做sensitivity。.

.--

7. 如何决定sample size?计算器https://www.evanmiller.org/ab-testing/sample-size.html 输入baseline，minimum effect you care about，α and β。α（how confident） .

8. 看结果（记得同时考虑statistical和practical significance）. 1point3acres.com

**二、policy and ethics（protect your participants）**

Risk

benefits

other available choices

Privacy

. 1point 3acres

**三、Metrics**

1.Define

Metrics两个uses cases: 1）invariant checking(randomisation checking?看population是否一样)sanity check(合理性/完整性)；2）evaluation of experiment（high-level/detailed)

. 1point 3acres

single or multiple? 都可以。也可以composite。要权衡。individual metrics们会一起变动互相影响。考虑how generally applicable: 最好可以有可以across the suite去用于比较一堆ABTest的metrics.

High-level metrics: Business funnel--stage--count/rate/probability/retention

difficult metrics（无data/take too long）--techniques to get a proxy(有correlation or causation)

techniques to validating metrics and brainstorming new metrics:

eg.: external data (market/industry research company--/ run survey of users company/academic research可以向他们学习建立metric/validation techniques的方法)

eg.: own data:

①existing data(logs...): retrospective/observational analysis (回顾性) 寻找baseline。（我们现在想找到一个合适的衡量experiment结果的metric，那么就需要这个metric是某种变化的结果，也就是causation关系）但是retrospective只能够得到correlation but not necessarily causation。所以我们actually需要run experiment to validate metrics???

②gather new data (questions that cannot be answered using existing data)(research/surveys/focus groups) 。可以结合retrospective：可以先通过focus group develop ideas about what you wanna look at，然后你到你的log里面去找data去做retrospective/observational analysis （without experiment structure的）

a) UER(user experience research类似于observe): go depth in a few users/brainstorming/use special equipment/validate results（see how that metrics change over time）

b) Focus group:

c) Surveys: 为了不太容易directly measure的指标

d) human evaluators??

Well-defined metric: fully realise the metrics(怎么算):

1) 用什么data/分母是啥分子是啥/要不要filter...filtering and segmenting（slice）: （你suspect有一些bias 需要把它们排除掉）external reason或者experiment需要/比较常见的是filter掉long, weird sessions/注意避免由filter带来的bias：注意filter掉奇怪session的时候确认一下这些奇怪不是你system的问题，比如compute metrics on different slices看你的filter是不是moving traffic disproportionately from one of these slices. 如果是并且make sense那就可以，但是如果是disproportionately但是找不出原因，那可能就会bias。/看week-over-week （是在做除法哦）traffic pattern changes比较有效识别unusual/比如你怀疑是不是Mobile设备会有JavaScript double count情况

2）用什么Summary Metrics: 有时候有很多可选择 比如per event measurement自己就是一个数字，比如video的load time。为你的metrics建立characteristic。（见第三pa- Characterise）四类：sum和counts/distributional的/prob and rate比如ctr/ratios

2. Build Intuition （about metric/data/system) 自己有数什么metric会根据什么change来变，所以你看到你的data的时候能判断出它是不是有问题，可不可以相信。

3. Characterise

sensitivity（你的metric要能够pickup你care的change） and robustness（还得对你不care的change没反应才行）How to measure？

1）run simple experiment；

2）retrospective analysis; 都是看你想选的metric是不是move in conjunction with those changes/ "a versus a experiment" 来确定这个metric没有过度敏感。你什么都不动，看你的metric会不会catch到伪差异（spurious difference）

distribution（retrospective analysis：画histogram，mean/median/25th percentile到底哪个合适）

absolute change or relative change(%,好处是stick with one practical significance boundary,但是你要很懂你metric的意思以及relative change的意思)?

Variability: more rigorous statistical definition (变异，指标一般是标准差、方差、变异系数。) metric的分布和variance。median和ratios或者两个rates的difference，的distribution 就比较难搞----用到non-parametric methods(在对它的分布不做什么假设的基础上analyse它，然后还能算出来non-parametric的confidence interval) 比如sign test(符号检验是一种统计方法，用于检验成对观察值之间的一致性差异，例如治疗前后受试者的体重。但是无法说出effect size有多大). (non-parametric statistics 非参数统计分析，或称非参数统计学，统计学的分支，适用于母群体分布情况未明、小样本、母群体分布不为常态也不易转换为常态。)

estimate the empirical variability of our metrics：for more complicated metrics, we might need to estimate the variance empirically instead of computing it analytically. Use A/A testing: 这时你measure到的any difference due to underlying variability/起到一个pin down the variability的作用//另外一个不用run太多AAtest的办法是在一个大的AAtest里面用bootstrap method。

empirical :经验的based on what is experienced or seen rather than on theory

Use of A/A test（做很多组 比如20组）:

  1) 如果你已经有了analytical的confidence interval：compare results to what you expect(sanity check);

2) 如果你不能analytically计算：那么就用AAtest去estimate variance and calculate confidence interval： 用empirical std （两个group的diff的Std）代替 analytical SE，就是在看diff的分布;

3) directly estimate confidence interval from your AA test.：从小到大排列取中间95%的点 1) 思考how many experiments you expect to see statistically significant difference at 95% level? out of 20, we expect to see

1. 还可以check 两个group之见之间的difference是不是正态分布，还有sample size和difference range的关系（size越大 range越tight）

Bootstrapping：关于metrics的总结： validating metrics非常重要。

四、Designing an experiment

1.Choose"subject" (unit of diversion) (顺序：userid/cookie/event比如pageview)

如何选择？1）consistency：看你想要consistency across what/看这个change的user visibility：user意识不到的时候用event是可以的，但是如果你发现了learning effect的现象，就要转用cookie。要求consistency across devices的话，需要用userid。通常IP diversion并不好做。

2）ethical considerations

. check 1point3acres for more.

3）variability: unit of analysis（你的分母是什么） 和 unit of diversion要统一，不统一的话，analytically variability和empirical variability会相差很多。原因：你在analytically算variability的时候，不只是assume了分布，还assume了什么是independent的。event是独立的，但是如果unit of diversion是cookie或者userid，event实际上时候correlation的，这时候variability会增加。cookie的variability明显更高

. Waral dи,

2. Choose"population" (who is eligible)

注意不要mismatch两个group的user。

inter-user experiment (同一拨人different time window) 但是问题很多，一般我们说ABtest还是指intra

intra-user experiment(different people on Agroup and Bgroup)

Target Population: 1）事先想好哪部分特定人群会被experiment affect，只在他们身上run. 这里要知道filtering traffic也会change variability和是否significant，因为unaffected的那一部分人会稀释你的结果。；2）事先想不到的，你也不知道会不会有unintentional effect on一部分traffic）

Population and Cohort: (我自己的理解是cohort是特定时间进入experiment的那一部分人,你"跟踪"他们,并不把他们和其他时间进入实验的人放在一起，用于一些特定的metric，比如user stability/learning effect/usage of site or device...一些相对于自身历史行为的变化)

3. Sizeby cookie的时候 analytical estimate of variabil

这里有一个Rcode一步一步讲如何算size（输入α，β，d\_min，还有一个standard error：s: The standard error of the metric with N=1 in each group，由一次5000个pageview empirically estimate的SE算出来）

http://video.udacity-data.com.s3 ... /empirical-sizing.r

有一个问题是：如果算出来的size太大了不太现实怎么办？怎样可以减小size？1）increase α，β或者d\_min；2）make unit of diversion and unit of analysis the same; 3) shrink it to specific traffic that is only affected排除掉那些会稀释你实验效果的traffic(这里注意你的d\_min可能会改变，以及你可能想要detect到smaller change)。

. 1point 3acres

最后在size上面还有两个问题：1）你事先不知道哪一部分人会被affected；2）有一些很明显的feature的trigger方式很容易被track（比如UI feature），但是有一些back end的feature，你怎么去知道他们是不是真的exposed to了。办法是：你可以run pilot或者用一小部分实验数据进行猜测。

4. Duration

1）Duration和每一天expose到user的数量(你不会想要一天run完全部的experiment get全部的experiment data可能需要的sample太大你也做不到 而是run them at the same time on smaller percentages of traffic,这样也有利于你run multiple tasks并且让他们comparable，还有sense到weekly variation，或者有一些有risk 安全起见先让小部分人看见change)

2）When时间点

\*learning effect：. 1point 3acres

1）choose userid or cookie; . check 1point3acres for more.

2) dosage或frequency会有影响，所以比起population最好用cohort，based on how long they've been exposed to the change or how many times; 3) maybe a high risk change, then you want to run it through a small proportion of users for a longer period of time. pre-period and post-period: 在ABtset前后分别的AAtest，pre确保两组可比，post衡量learning effect。

五、Analysing results

1.sanity check: check for invariant metrics:

1) population sizing metrics, make sure your control and experiment group are comparable;

2) metrics that should change in your experiments, check them 是不是真的没有change。

一般如果event是一个好的invariant metric那么cookie和useridu也会randomly分布的。

.1point3acres

然后来看如何check。我们randomly assign cookie到control group和experiment group的话，control组占比应该为0.5.我们建立confidence interval来看control组的cookie数目是不是过多了。

如果看完了daybyday发现这是歌overall的现象，你就要想想为什么会这样。比如talk to your engineer或者slice看看是不是某个particular slice在出问题。还可以去对比一下pre-period和post-period的数据，做做retrospective analysis看看是哪里出问题，也可以鉴别以下是不是learning effect。

2. Single metric

主要就是看是否有significant difference。如果结果是not statistically significant，很不符合你的预期，那么你要break/slice去看结果，一方面debug你的experiment setup，一方面可以get some new hypothesis. cross-check?

e.g.: CTR（possion distribution 而非CTP的binomial）

1）借助empirical SE来算experiment的SE，然后得出confidence interval，做出conclusion。（empirical SE在决定sizing的时候算的）

2）看day-by-day 来做sign test （p-value：7 out of 7 by chance的概率 <α说明不是by chance）

3）如果1）和2）得出的结果不一样，进一步探究原因在哪里，比如weekday和weekend分开看，是不是weekend明显是显著的但是weekday不是，解决办法要么和decision maker商量之影响weekend够不够，要么研究改进一下 如何也影响到weekday。

Simpson's Paradox:

..

分开department看到的时候都是Women更高，但是放在一块看就是Men更高了

**3. Multiple metrics**

最常见的问题：同时在test的metric越多，being significant just by chance的metric就也多。比如同时衡量3个metrics的时候，以5%的置信度去算，至少有一个metric by chance to be significant（False Positive）的概率为1-0.95\*0.95\*0.95 = 0.143(如果假设metric之间independent的话) 但是好在这种现象并不repeatable在同一个metric身上。-baidu 1point3acres

Solution: 每个metric都用higher confidence level: 常用方法就是 Bonferroni correction： [公式] 但是Bonferroni correction有个很大的问题就是，如果你选择的metric之间关联性比较大所以会move together，那么Bonferroni就会过于conservative而让你得出这些metric都不显著的结论，可能让你错过一些机会。

Solution：1）用一些less conservative的metrics：closed testing procedure, theBoole-Bonferroni bound, and theHolm-Bonferroni method. 2） 用false discovery rate (FDR：在所有你reject的假设里面 有多少是真的有difference) 代替 familywise error rate (FWER)：适用于你同时看上百种metric的情况。this articlecontains more information

但是，不管何种metric的组合 都要基于business goal。你要理解你的业务，并且有预期什么指标是好的，什么指标之间是有关联的，指标之间的权衡是什么，这样你得到结果的时候 你可以分辨出来任何地方是不是有bug，or你才能做出正确的判断：if you do see a change, do you understand the change? Do you want to launch the change? Worth it?

4. Gotchas

最后的一点Gotchas: 在ramp up your experiment (Gradually increase 的过程中，原先significant的不significant了。1） 也许受了seasonal or event driven impacts. 搞一个Holdback group去catch； 2）也许是他们adapt你的change，这时cohort analysis就很helpful； 3）没有很好的control advertise。