Parents on FB FB家长的出现对teenager的影响

1. 如果年轻小朋友的父母也有脸书账号，这对年轻小朋友在脸书上有什么影响？

1) 首先说FB的goal 是什么。AARRR模型，这道题可以关联到New Users，Retention，Engagement（session time per week/post & interactions per week），referral，revenue
2) 内部找问题：AA test或者AB test。
    a. AA的话就找teenager有了家长之后跟没有之前的对比，看各项metrics的变化。quasi-regression (difference in difference) 父母是为加入作为一个0 or 1 的dummy variable， fit 父母加入前后的用户的sessions per week. 比较难以做长时间的对比（时间越长，时间和其他因素影响越大），所以比较难搞定learning affect。
    b. ABtest 比较好，找类似的两组人，一个有家长一个没家长。然后看各项指标。这里主要看session time per week和post comment per week。因为最大可能影响这两点（AB test要注意network effect）。
3) 短期vs长期要把时间线拉长，可能短期内不好，但长期内是好事情。Learning affect。teenager可能学会怎么回避了。
4) Simpson Paradox：一定要把问题分解到不同的user segment，不同类型的人肯定反应是不一样的。比如印度人和美国人，西海岸和东海岸，高中生和大学生，甚至学生的兴趣专业。所以绝对不可以看总体metrics，要分开看怎么影响

* AB test的问题：怎么控制一组有家长一组没有家长？
家长加入不加入Facebook，experimenter是不能控制的。所以这里存在着selection bias。也许家长加入的家庭孩子根本不在意这个，或者家长取得了孩子的同意之类的，你看的lift in metric 就不会significant。ab testing不是最好的办法。把它当作observational studies 用一些econometrics来解决会比较好（propensity score matching + DiD)

具体的测试方法（DiD)：
此类问题就是如何用短期feature预测长期metrics。所以主要答清楚短期feature都有什么。之后的问题就是去historical data里面找training set
Feature：

- 用户的profile信息。所有注册信息。
- 用户的browsing信息。location，语言，device等等
- 用户的behavior信息，interest，浏览过什么，关注什么
- 用户的network信息，朋友圈都是什么interest什么profile，地理位置。
- 用户的使用时常，频率。

定义Training set：

- 找用户父母加入前和加入一年后的Y。Y可以是session time per week，post per month，like per month等等
- 收集所有X feature。其中一个feature就是父母加入。
- 再找一群没有父母加入的数据，用来帮助train 其他feature的影响

模型可以用random forest，xgboost，neural network。结果做好generalization，不要overfit。然后你就得到了不同情况下父母加入对用户的影响。可以用你的模型预测。

- 最后就是有了家长带来的好处和坏处，就可以综合考虑对fb的影响，然后通过simpson paradox分析和短期长期的分析，我们可以改变产品，取长补短，让FB越来越好。

* stratified sampling在online ab testing上比较难实现，因为它需要数据采样好了，才能进行stratify。但是tech company ab testing都是来一个assign一个。所以需要比较complex infrastructure来支持。 如果直接做random sampling 如果数据量不是特别大 很有可能导致selection bias, 所以通过在一些对结果影响比较大的characteristic上采样然后分配到control和treatment 来保证sample 的probablistic equivalence

2. 如何确定父母和子女的关系
   1) 已有的label信息
   2) last name，location, pictures together, 称呼in message/comment 找
   3) 用survey填写

* 不管是加入的时候做survey自己self report，还是用ml predict，都能产生不少问题. 比如survey，如果很多人选择不reveal，那这样treatment triggering traffic要小很多。一般response rate有个10%就不错了。而且本身父母的traffic就要小很多。再者也容易把没有self-report的父母分到control. 还是影响% lift in oec. ml的问题是cold start的问题。刚加入的话，什么信息都没有，如果predict？

* 如果单纯靠用户自己标识出来的话，有什么问题？这样就需要挑出来那些没自主标识出家长的账户，就有了建模的后半部分。fb上用户可以把好友关系打标为父母兄弟姐妹亲戚，在个人about页面里面会显示出来。如果我们默认这是“正确答案”的话，就可以建模预测另外一些没有标注这项信息，但是父母也在平台上的用户是否好友里有家长。这是个标准的建模流程，从开始的output定义，特征收集，到建模后的交叉检验，然后得到理想的recall水平。最后还可以利用其它手段来检验模型质量。

3. 父母来了，子女走了，如何在父母和子女之间取得平衡？
   1) 从用户活跃的角度：父母在fb上contribute的sessions vs 子女离开fb失去的sessions
   2) 从Revenue的角度：父母在fb上的ad clicks vs 子女离开失去的ad clicks
   3) 找到问题就可以提出假设，设计ab testing，看怎么改进。比如session time没变，但是post少了，可以给他们屏蔽家长post的功能。这里面试官跟我说，我们fb是改变news feed algorithm来实现的，家长比较难以看到孩子们的的news feed。
   4) 然后我们聊了AB testing的细节。就是有俩news feed algorithm怎么ab。
      - AB能不能独立（这里没问题，因为algorithm变化用户无法察觉）
      - 怎么设计control variable（对结果correlation大的variable。可以通过correlation check来找。Variable有很多很多，最科学的方法是全部看一遍，按照correlation排序，然后设计experiment。面试的话你要提一下你认为什么有high correlation。比如这道题就是Teenager age, parent age, location, language, teenager interest, city/village, gender 等等）
      - metrics: session time）
      - sample size（effect size, alpha, sensitivity, variability这里跟季节，population稳定性，去掉outlier）
      - 得到结果先sanity check（population metrics, control variable)
      - 然后novelty affect还好，因为algorithm变化很难detect。基本上就可以通过结果判断哪个好了。

* self selection bias 不会因为traffic大，就会mitigate。它是endogeneity的问题, 需要统计方法来解决，不是central limit theory能解决的。就好比traffic再大，也可能会存在confounders,

影响ate。test两个algorithm，treatment应该是尽量不给父母看子女的post。这个就有效的规避了self-selection bias,因为randomization criteria 不再是是否是父母，两组都是有父母和非父母。理论上父母和非父母的proportion应该comparable，任何差别应该是randomly distributed across groups, 可以a/a test一下。

4. 接下来还有时间，面试官就想听听家长本身的因素

- 家长本身自己的ad revenue就比一般人高，而且给我提供了新的targeting customer（家长），可以看这类segment给我带来的revenue。通过Simpson paradox的分解，找到哪类家长比较好。
- 家长的referral factor，他们会带动同时，亲戚什么的来fb，我们可以evaluate他们network能给我们带来多少engagement，session，和广告收入
- 建ml模型，看家长进入对我们metrics长期的影响。

FB Memory page

类似人人网过往的今天功能, 在newsfeed顶端的位置，用户每一天可以看到一次memory, friendship, anniversary这三个feature中间的一个，怎么evaluate memory page performance

- metrics：
    - product: ctr，comment和其他interaction
    - ecosystem: total time spent on FB
    - trade-off: reduction in other features

- 如果ctr 20% 你觉得可以吗？

需要跟其他两个feature对比，draw一个baseline。以user为出发点，跟另外两个(Friendship, anniversary)横向比较，如果user的CTR比另外两个feature高的话，证明是effective的。这样同位置的post相比较就不会出现higher ranked post is more likely to be shared 这样的bias。而且这三个feature是互相代替的关系，所以这样横向比较应该是make sense的。

- 如果看fb整体的timespent变化，是一个好的metric吗？

既然这是个newsfeed feature, 是related to content generation, 所以用overall time_spent change 是合理的。

- 如果2019年的帖子在2020年被推荐了，怎么确保2021年不会继续推荐。

1) Clarify: 被推荐是指用户看到了，但是用户有没有分享？
2) The easiest approach: label recommendation and use it as a flag
3) 看comment time的distribution，如果有两个peak，那可能就是这个帖子引发了两个时间段的discussion，之后就没有必要再推荐了
4) Why study this question: distinguish whether the user dislike the function or the recommended content
    - if the user click the memory but does not share it, we can pump up a survey asking whether the user likes the content/whether the memory is important
    - collect features (view, click, share, comments) and build a model to improve the recommendation system

怎么evaluate FB里有多少fake news?

Fake News

1) 如果facebook现在没有很多content reviewer，让你一天之内来measure fake news的impact，你要怎么做。

Stratified sampling一些post by category (sports news, politics news, financial news, etc) 然后一个一个看是不是fake news.
这个题要看面试官怎么问你，感觉地里有两种问法，一种是让你估计X%，一种是问impact。如果是问impact，不要直接开始就说用sampling来做estimation，最好给一个框架，从几个角度来说impact，然后再深入的说怎么用sampling（stratified），为什么可以用sampling（CLT）

2) 如果你sample了1000个结果有10%是fake的,你怎么report给executive说FB总共有多少个?

提示了答案说是一个binominal distribution, either是fake new或者不是, p = 10%,用公式估计mean和confidence interval之类的

3) 如果时间充足，资源也充足，你要怎么来measure这个fake news的prevalence。

- 这个我先说了用sampling的局限在哪里，如果时间足建模会更准确detect fake news，最后用metrics来measure impact
- 重点说一说怎么找feature
    - from new itself: similarity score to any detected fake news based on NLP (title, content), link and photo included. Title: call capitalized? Content: any grammar mistakes?
    - from user interaction: user reported fake, comments suggests they don't believe it, shared by lot robot users
    - from source: is it a trustable website? whether the pages that shared the content have a history of sharing fake news

- Linguistic features + machine learning approaches (first find some real news and fake news by hand, and then use SVM classifier):
    - Punctuation
    - Psycholinguistic features (the proportions of words that fall into psycholinguistic categories. LIWC is based on large lexicons of word categories that represent psycholinguistic processes (e.g., positive emotions, perceptual processes), summary categories, as well as part-of-speech categories. Categorized into summary categories (e.g., analytical thinking, emotional tone), linguistic processes (e.g., function words, pronouns), and psychological processes (e.g., affective processes, social processes)
    - Readability. We also extract features that indicate text understandability. These include content features such as the number of characters, complex words, long words, number of syllables, word types, and number of paragraphs, among others content features.
    - Syntax

- Check by id: Key with fraud is, not happening only once. People who commit fraud would like to repeat it if not being caught. all variables are really about something

that should be unique but is not or extreme values. Hence two main ways to capture fraud:
    - Same device IP/Bank account/phone number as existing accounts;
    - Anomaly detection-find outliers (extremely low price)
        - More specifically with market place posting, we can address the listing and seller. For listing, pictures cannot be stolen from elsewhere/descriptions cannot be copied/resolution should not be too low/price should not be two
        - With fake profile (say fake school): using ML algorithm or anomaly detection to find outliers. For instance, you may include the percentage of connections went to the same school/interaction with people from the same school/acceptance rate for the same school request as variables. In order to minimize the fake profile, you may want to use 2-step verification for risky users (minimize false negative you may not apply this to all users).

- Use the existing fact-check database

4) 如果我们发现这个model没能很好的capture所有的fake news,你会怎么改进?
    a) 是不是说recall不够高? 我们需要discussion怎么balance precision和recall,看哪个对user影响更大.通过分析和跟PM讨论利弊决定. 先看看users如果看了fake news有什么不好的影响,如果fake news的negative影响大,我们应该尽可能detect多fake news,increase recall. 如果fake news的negative影响不大,但如果一个不是fake news却被删掉/derank的影响很大,我们应该保证precision
    b) 知道了我们需要focus什么metric之后, 根据上线后的实际反馈,收集更多的数据,再retrain model并优化

5) impact of problem，就是对用户的影响。
最直接的就是看损失多少活跃用户，先定义什么叫活跃什么叫churn，一个月不上线就算churn了，所以我们需要一个月的记录来获取label (1: churn, 0: active)，然后把一个月之前的数据提炼features，features种类分demographic, behavioral就不多说了，但这里重要的是要加上他们和fake news互动的feature，比如comment, share, like news that turned out to be fake之类的。怎么分别用户的损失是fake news导致的？先run full model，预测那些人会churn，然后把fake news feature去掉，预测哪些人会churn，两者之差就是fake news导致的用户损失。这个最好用tree ensemble model实现，课用logistic regression做baseline，面试官说logistic regression is fine，又问了我如何interpret coefficient，直接影响log odds，正数正相关，负数反向关。

【8小时以内怎么衡量fake news的impact】
这问题的背景，和商业要求是什么？我在答案结尾也会强调，面试**是对话**，而**不是考试**。你和面试官的关系，是一起解决一个复杂问题的同事。对问题的关键参数了解越充分，越能给出最优解。主动询问问题，不但是自信的体现，也是面对复杂问题时负责任的态度。当然，问题要问在点子上，问完问题也要保持互动。

比方说，你问了面试官这问题的背景，面试官告诉你，假定这是因为小马哥接到上头的询问，要他给个汇报。那你这时候可以猜测，（1）时间不能商量，（2）资源很可能是很有弹性的，（3）对谣言的定义要尽可能符合 "上头" 的定义，而不是按照我们自己的想法来定义。接下

来，你应该马上把这些猜测和面试官进行沟通和确认。一方面可以框定解决方案的范围与取舍，另一方面，在沟通的过程中，你也体现了很多优秀的职业素质。

举例
用我们之前 "谣言" 的例子：通过跟面试官的沟通，我们知道，短期需求是，我们需要知道谣言的百分比，用于汇报。长期需求是，我们需要降低微信上谣言的比例。这时，我们可以把问题需求分解成 "总体" vs."个体"。前者意味着只需要知道整个微信层面的平均比例就行了，而后者要对每一篇个体文章做一个判断，究竟是不是谣言。前者可以用于汇报，后者可以用于反谣言。前者，用抽样 + 人工标记的方法就可以得到一个大概的估计，不需要再分解下去，小马哥也可以拿着这个估计去汇报了。
后者，想判断每篇文章是不是谣言，我们可以把 "谣言" 的定义进一步分解，区分为 "内容" 和 "来源" 两类。解决方案也可以归类，比如人工 vs. 自动化；自动化可以分解为专家系统或机器学习。先把大框架列出来，再跟面试官沟通，给出自己对于不同方案复杂度、成功率、资源要求，的估计和相应的建议，和面试官选择一个方案展开进行重点探讨。到这一步，就可以穿插自己对数据和技术的理解了。比如，我们选择先解决 "来源" 类的定义：
根据我过往的*经验*，我*猜想*，"谣言" 的 "来源" 很有可能遵循 power law，也就是少部分网站 / 作者贡献了大部分的谣言。我第一步想通过现有网站数据和人工标记的 "谣言 "标签来*验证*这个猜想。如果猜想成立，那我们通过重点监督 / 禁止这些网站，就可以有效减少谣言的数量。
如果面试官对这个答案满意，我们就可以去解决下一个问题。如果面试官有进一步的问题，比如，"你的第一步我还是不太明白，可不可以详细阐述一下你具体的检验方法？" 我们也把问题限定在了一个有限的范围，并且完全理解背景和目的，使这个问题更接近一个纯粹的技术问题，更容易回答。
当然，这个例子只是管中窥豹。想提高自己系统分析问题的能力，是需要很多练习的。在我看来，准备面试时，比较有效的方法是多做 mock interview。关于 mock interview 的心得，可以参考下我 [之前的回答](https://www.zhihu.com/question/20544832/answer/374701116)，有空还可以读读 "Case Interview Secret" 这本书。

FB Group

1. FB group的health怎么measure？

问health其实就是问performance。可以分两个部分，内部和外部。万物皆可内外。另外任何公司的goal都是growth，fb的话就是connect people。从aarrr模型出发，就是要考虑，新用户增长，用户retention，用户engagement，用户referral。FB group可以从各个方面影响这几个metrics。

- 内部：这个group本身的health。我们可以看用户（Acquisition, Activation, Retention, Referral, Revenue）：新用户增长率，retention率，月活（定义一个月内在group的session时间>30mins)。这一点可以看这个group的增长和活跃度。第二个角度是用户的engagement细节：post数量/月，post总阅读量，点赞，comment，分享；第三个是referral affect：有多少用户来自referral。根据我们的目标，我们会用不同的metrics。有些group，总人数很重要，比如information类，找工作啥的，就看用户增长，retention，月活，referral。有些group engagement更重要，就看post，like，comment等等。
- 外部：就是这个group对我们整个fb的影响。这是比较难直接看出来的。最好的方法是，先定义一个大metrics，比如我们希望fb group可以让用户在fb上总体来说更active。那我们就可以用session time per month来衡量，这个group对用户的影响。然后，两种方法，AAtest或者ABtest。AA就是看用户加入group之后和之前的时间engagement

变化。当然要考虑nevalty affect，这个从retention rate可以看；AB就是找两个很像的用户(Profile, interest, Location, Device, Style...), 一个加入了group，一个没加入，看他俩的变化。可以建立ml模型，得到generalized的结果。就是加入这个group，对session time是有多大的正相关。

- Trade-off: any significant loss from other FB features

(weighted sum of # interaction) / (# groups)。interaction包括评论，post等。小哥指出可能有问题，我说没考虑group大小，一视同仁了。小哥说没错，怎么解决。我提议把分母改成(# members)，但这样的metric很难interpret，因为估计会很接近0了，而且同一个用户加入多个group会double count。于是提议按group size分strata，大组和大组比，小组和小组比。

2. 大组vs 小组？

- 首先看你是要quality还是quantity。如果考虑quantity，那么大组可能会影响更多的人。另外很多大组目的就是要reach out 更多人，比如job posting。那他们关于用户增长，和DAU % 的metrics就很重要。组能reachout 很多人，而且大家都能长期active就是他们的目的。另外metrics的总量会比较重要，比如总活跃，总session，总发帖。
- 如果考虑quality，就是小组比较需要的。很多小组人很少，比如朋友圈，学校的某个班级组。这里就要引入private的组还是public的组的概念，会影响我们的metrics选择。private组往往更注重quality。quality其实就是所有metrics换一个方法scale的个人level。比如人均发帖数，人均阅读，点赞，comment，还有人均referral等等。用户增长并不是这种组的目的，所以可以不考虑。

3. 一个comment新feature，怎么判断好不好？/ 用Reddit的hierarchy comment，是否提高engagement。

1) 首先搞清楚goal和metrics。这里可以narrow down，选择comment数量/post或者(weighted sum of # interaction) / (# groups)。前提当然是post总数不变，用户总量尽量不变，所以要选择一个好的区间和ab。然后做ab testing。
    - Sample size：effective size，power，alpha，然后最重要的variablility。
    - Variability：选择的group尽量metrics稳定，人数稳定。选择的season尽量普通。去掉Outlier FB group。
    - Control Variable：两个ab group，interest 类似，人员构成类似，总活跃度，注册时间各方面都类似。注，不可以同一个group内分ab，因为是不能独立分开的。
2) 然后就run test啦。比如run了2周，咱们分析结果：
    - Sanity check：保证Control variable没变化。看类似其他group没有受season影响等等
    - Novelty effect：新用户和current用户对新变化反映类似；FB很多历史数据可以帮我们adjust
    - 假如结果significant，那就是可以扩大test到其他类型fb group。跟上一题很相关，不同大小，不同类型的group反应肯定不一样。
3) AB testing的其他问题：
    - Learning Effect：人们可能会适应新变化，反而comment增加更多
    - 同质化问题：AB testing 永远只能根据现有用户preference调整产品，而这一可能会讲少部分用户拒之门外，导致用户越来越趋向于大部分。不同类型的用户就被排斥了。
    - 每当有大变化的时候，就可以考虑做以前做过的ab testing。比如产品变动。或者用户组成变动。一个简单测试就是，改变上次ab testing的用户成分，看会不会很大影响结果。

- 分experiment & control groups要按cluster来分，不能按用户来分，而且cluster之间用户的friends overlap要最低，不然无法保证independent。

4. 还问了ads之类的，衡量经济效益 / facebook 'recommend other groups' 这个feature在news feeds里头，如何test the balance/trade off between showing ads & 'recommend other groups'

- balance ads matching algorithm:
- Compare the ROI
    - 'recommend other groups': engagement increase vs. cost of recommendation system
    - ads: ads revenue vs. loss on user engagement/retention

5. If FB Group is so good that a lot of people start to only use it but no other FB products such as story, what would you think?

- Novelty effect: monitor long-term performance
- To analyze the trade-off effect between the two products, we can compare the cost and benefit/ROI, especially in the long-term. Also compare the level of AARRR for the two, so that we can understand what is the total loss/benefit
- Whether there are any overlaps between the two features and how people use it. Analyze user's shared behavior and characteristics of the two products, and determine what are some factors that may influence the probability of using the two. Then determine new features to attract new users / differentiate the two products
- Try to link the two products together

6. We found posts in Group normally get more comments than regular posts, why? How do you verify your hypothesis?

Hypothesis: People are more engaged/know each other better than regular FB friends

- More relevant content: recruit some users in an experiment and ask them to post relevant and irrelevant posts in the same group, and they compare the number of the comments
- More close relationships/networks: recruit some users in an experiment and ask them to post the same posts in the group, as well as in the regular section. Compare the number of the comments

Messenger 支付

1) 这样的好处坏处;

- To FB

A: it can bring some new users because if many of my friends are using it to transfer money, I will also register and install the APP to user it

A: active more users because many users would need this feature and they will user it to transfer money to their friends. If more users are aware of this feature, they will be active in messenger APP

R: users would feel messenger is more useful and they will occasionally use this feature

R: if I like this feature and user it a lot, I will referral it to my friend and invite them to use it, so that the level of referral would also increase

R: there can be some service fee; or interest revenue because it takes time for the money

transfer; there can also be some online financial products that the user can purchase

Cons: need to work with bank or third-party financial services; need to work more on how to protect user's privacy and safety

- To users

Can transfer money more easily
Small business users can run business without cash
Worries of privacy and safety

2) 问说有什么framework可以evaluate这个feature吗？
Metrics: engagement (total # of msg sent per week, total time spent on msg per week)

- Product
  - Acquisition: # of new users for messenger & for the new feature
  - Activision: number of transactions/ratio of users who use the feature
  - Retention: Retention rate, Conversion rate (click the transfer icon – add bank/card info – find the friend to transfer – enter the amount – finish transfer
  - Referral: How many new users acquired by referral
  - Revenue: service fee, interest, ads revenue
- Ecosystem
  - FB engagement & retention
  - referral effect on FB
  - connected with other relevant features (marketplace, group)
- Trade-off

Test:

- AB test: not so good here
  - compliance issue: ATE is biased
  - network effect: need to randomly sample community/cluster
- Propensity score + DiD
  - use propensity score to match the real treatment group (with compliance) and people in control group
  - use DiD to test the change in engagement
  - also need to randomly sample community/cluster

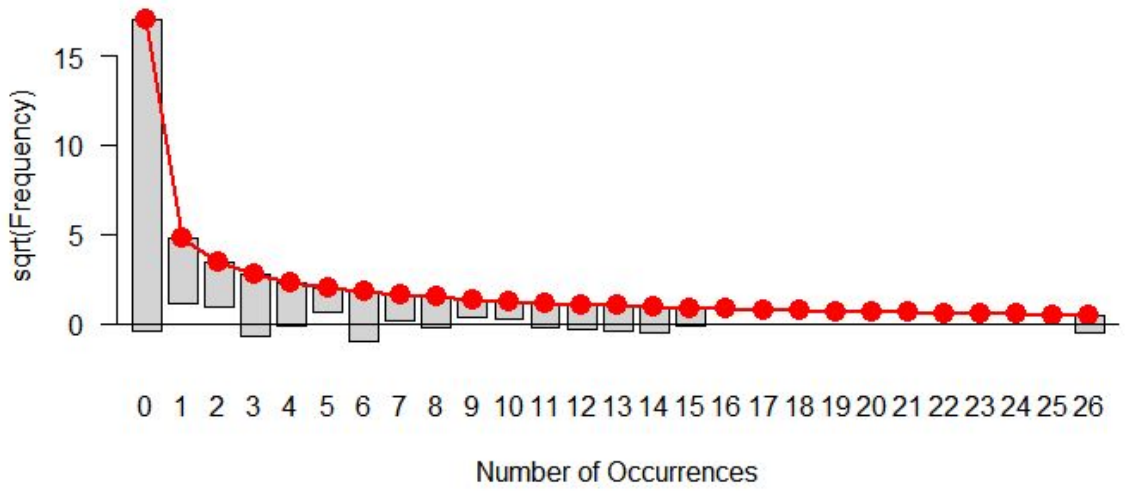
3) estimate what percentage of users will use this feature

- market research & analysis on competitive products: the size/scale of the market and the feature of the user
- experiment and prediction:
  - randomly pick some users to try the new feature (randomly select a community of the user instead of individual because 只有我很多朋友都用我才会用，只找几个人试试我可能并不想用)
  - if they use the feature, label as 1. Collect other features to run a logistic regression, and then predict
  - feature: demographic, location, level of engagement, social network (# of friends, # of friends using payment)

4) estimate distribution of number of payments in a month (x轴是number of payments, y是

number of users); 求这个distribution的mean和median, 是不是 会有偏差？
distribution with a long right tail / right skewed distribution
* 指数分布不行，因为是连续的
* 泊松分布：方差和平均数必须一样(the probability of a given number of events occurring in a fixed interval of time or space. If these events occur with a known constant mean rate and independently of the time since the last event)

* zero inflated negative binomial: modeling count variables with excessive zeros and it is usually for overdispersed count outcome variables. Furthermore, theory suggests that the excess zeros are generated by a separate process from the count values and that the excess zeros can be modeled independently.



mode < median < mean

- Median: if more than 50% of users use zero times, then it would be zero
- Mean: Poisson distribution for number of payments larger than zero (assume we have the data on the time interval of two payments of one user, e.g. three days, that is, the lambda is 10, which is 1 over 1/10 months. The parameter of Poisson distribution is lambda times t, which is 10 here. The mean should be lambda + 1 for users with non-zero payments. The mean for all users is p*0+(1-p)*(lambda + 1)

5) distinguish small business and suspicious accounts
- small business:
    - large number of payments received
    - frequently have a large number of transactions
    - the amount of each transaction may not be very big
    - receive transaction on work time (business time)
- suspicious accounts:
    - large number of payments
    - abnormal large number of transactions
    - the amount of one transaction can be very large
    - may have abnormal transaction time

INS切换账号

Instagram为了使用户方便 launch了可以迅速切换账号的button，以前想要切换必须退出当前账号再登陆，现在简化了这个步骤，点屏幕右下方就可直接切换 / instagram now have feature let users to switch accounts with one button

1) Clarify: why users want to have different accounts?
    a) the original accounts have too many followers
    b) be anonymous & post different content in different groups
    c) business runners with their personal accounts

2) 怎么识别几个账号来自同一用户

- Same device id, ip address, registration phone number
- Similar user name, geographic data (but user name & demographic features can be different if the user want to be anonymous on the second account)
- Shared friends, shared follower, all following some accounts

3) 做了实验后发现numebr of account increased，但是avg time spent 没有变化问为什么？What is your hypothesis about why this happened? What data do you need and how to testify your hypothesis?

- novelty effect，一开始推出这个feature大家觉得有趣就去多建了几个账号但 是the way ppl interact with Instagram并没有变化所以avg time spent没有变。 我觉得是因为人们没有足够的精力细致的管理多个账户，或者对于使用多个账户还在学习中。
- Total level of engagement does not change: users can spend more time on one account, which causes them to spend less time on the other. 回归模型或DiD to measure the time change in original account (x would be the interaction between 0/1 having a new account and the level of engagement)
- 对用户进行分层，simpson paradox
- data是否correct？test有没有错，有没有population selection bias
- AA test

4) 是否要launch这个feature? How do you determine if this feature is a successful launch?

Before launch: AB test on engagement (total time spent on Ins per week)

- network effect
- novelty effect

After launch: No ab testing at all since the feature already launched.
问清楚这个feature的goal然后做cost-benefit analysis。结合opportunity size 以及你想的一些metrics来决定是否有practical impact，impact有多大来决 定。

5) 哪些用户会抵制？哪些用户会不喜欢这个feature？

- 父母会抵制，因为子女会建小号屏蔽他们
- people who don't like changes

Ins Checkout

1) 为什么要推出这个feature? 有什么好，要从fb和商户两个方面讲

- Facebook:
    - A: new business users with large number of potential followers
    - A: create new interactions between brands and users
    - R: good interaction would increase retention
    - R: may also increase the referral effect
    - R: monetization
- Business: large size of user (1B MAU); especially good and convenient for Early-stage online retailers, E-commerce newbies, Retailers with limited products, Established Instagram-first retailers, and Retailers with a lot of social traffic because
    (i) they do not need to own their e-commerce platform;
    (ii) they don't have the revenue to create a custom shopping cart or pay for a full-scale solution;
    (iii) link to your product every time you post

2) How to measure success?

metric: engagement/retention rate + Ads revenue

3) 有什么坏处，有谁会不喜欢这个feature

- Customer: do not like Ads & sponsored contents; worried about privacy & safety; believe and prefer brand identity
- Instagram: how to handle large transaction / limited-inventory drops / prevent bots and resellers
- Business: diverse category of goods; lose customer data; already have their own inventory, payment and shipping system

4) 怎么样去改善这个feature

- Goal & any feedback/complaint/reports
- Recommendation system: Find a way to differentiate customers who like the feature & do not
- Business tool: tracking inventory & monitor business in a more systematic way

5) 怎么看我们是不是应该推出这个feature

- AB test
- Cost-benefit analysis / ROI analysis
- Qualitative research on user's experience

Ins Story

1) 怎么评估Instagram 增加 story好不好？

AB test (with network effect):

- product metric (revenue from ads)
- ecosystem metric (ins & other FB product)
- trade-off metric

2) 如何决定是否launch story这个功能

AB test: engagement
Cost: loss on feed post (engagement & revenue) + cost of design & PR
Benefit: engagement (session time, interaction) + revenue

3) 如果launch 了以后发现，用户减少 feed post了怎么办？
- novelty effect: long-term analysis能否长期从story弥补？
- 用户有没有大约等额的增加使用了story？
- 如果是monetization的问题， 我们要算两种feature的roi和增长率

4) facebook 上头 Stories 的 view 只有 instagram 的一半，问这个是什么原因？如何 design experiment 确认是这个原因
- 用户画像：年龄，性别
- 使用目的：
    - 是否有原创内容还是新闻

In recent years, Facebook users have been creating and sharing less original content, and opting instead for sharing news and information from other media websites like Buzzfeed and Clickhole.
In fact, some users pointed out that they are using Facebook far less to post personal updates and far *more* to share third-party content and news.

- 是否好友圈子过大

The reasoning makes sense: as your network of friends grows beyond your intimate circle, you feel less inclined to share the more personal aspects of your life. The phenomenon is referred to as “context collapse”-- in face-to-face interactions, we read our environment and adjust our behavior to suit the context, but online, there’s no single context that you’re catering to. So there are an infinite number of potential contexts collapsing in on themselves, and we struggle to reconcile our self-representation in the face of such a diverse audience. And consequently, we avoid the dilemma altogether, and stop sharing the more intimate details of our lives on platforms like Facebook.
“friend” group was built up very broadly and they don’t want to share real-time moments with this mixed group of people.

- 是否使用ins

already quite content using either Instagram Stories or Snapchat Stories—They don’t want to use (or switch to) yet another platform to post “stories.”

UI Design

目前当你在手机上刷FB的newsfeed的时候你会发现一条post几乎占据了你整个手机频幕，其中有post的主体内容，别人留下的comments，还有一个comment box等等。现在打算缩小每个post的大小，让一个手机屏幕可以fit下3条post

1. 怎么去evaluate? / Post 的UI size 减小15%，怎么evaluate？用什么metric？有什么影响？A/B test步骤？

1) goal of the function:

    - increase engagement, attractive users to spend more time on posts & may also reduce ads impressions / CTR
    - fewer Ads impressions
2) metrics: engagement + ad impression
3) network effect: randomize by cluster/community

2. 如果launch之后，US 的revenue增加，泰国的revenue减少怎么办？
Users from different countries/languages may have different habits and preferences. It makes sense to differentiate UI design based on country/language

3. 怎么知道放多少广告合适？
- Goal: increase CTR & Ads revenue
- Baseline: 使用历史数据了解变化前主页中广告的点击率有多少
- AB Test: 插入更多广告之后ctr是否有显著变化（和baseline比较），可以同时check engagement level有没有明显的变化
- Causal Inference:
    - logistic regression on ads & user features
    - DiD: 插入更多广告之后ads revenue/engagement的变化

Notification

1) 怎么measure quality of notification？
- product metric: open rate/views, CTR, unsubscribe rate
- ecosystem: engagement, interaction

2) 如何improve？客户complain太多push了你怎么解决?
- 找user may like 的content to push：Recommendation system
    - feature: level of interest (based on event/location); interaction with friends; time of push; level of concerns for big newspaper
    - target variable: CTR
- find the proper number of notification
- unsubscribe feature
- UX research

3) 怎么选push 多少条notification的threshold？定threshold和定push多少条notification 是不是the same thing？
- goal of the threshold: prevent the users to unsubscribe or churn
- metric: label of unsubscribe or churn
- model: logistic regression (treat the number of notifications as a categorical variable, and create an interaction between number of notifications and other features, e.g. engagement level)
- the significance of the coefficients can tell when to stop
- some users may have a higher level of tolerance for notifications, may slice the users into different categories and run the model again

Friend Nearby

用nearby friends里的location信息建新的功能或者产品

- friend recommendation
- event recommendation
- ads recommendation (local/small business)

首先搞清楚是什么，在问如何。Location data可能的形式有：

- 用户打开FB，当他们选择分享地址的时候会有用户的一个location；
- 用户签到的时候会有location；
- 用户Post东西的时候会配location然后可以点到面分析

可以分为大到小分析

- 点的分析：每个用户的location数据点有什么用？
    - country level可以帮我们确认语言，demographic；
    - city level可以帮我们界定用户属性，可以推广local event和local business的广告；如果突然city变了很有可能是travel或者搬家了
    - 具体location level可以知道用户的activity（工作，学习，散步，shopping，运动，等等），可以知道用户在哪，推附近的广告（饭店了推饭，shopping可以推促销）；用户的历史数据-去过哪里-可以帮我们判断用户的属性（比如常去某个店，这个店有打折活动要通知他）
- 面的分析：千千万万location在一起，我们可以绘制人口流动图，可以作为一个人群在哪里的分析，可以用来预测流量，帮助广告商进行地毯式宣传，或者covid情况避免人流，帮忙找停车位，或者某个event的location范围，比如某个游行，我们可以看到地理上的规模，波及范围。可以看到这个用户跟朋友们是不是隔得很远，推各种远程交友的东东，或者鼓励用户参加各种线下社交，结交更多当地的朋友；如果离的很近推各种线下活动相关feature。
- 特殊情况：比如用户post的时候配location，说明用户想告诉别人我在哪里，可以作为一个”想被知道在哪indicator“，我们可以看这种炫耀行为对产品有没有积极影响，若有可以鼓励他们。也可以作为feature在其他模型里出现。附近的人可以成为一个社交手段类似于微信

Identify the location at home/work/shopping?

- 有label好的data，那可以直接run模型，预测。如果没有label好的就要找比较靠谱的indicator，然后用少量的靠谱label train模型，产生generalized大模型
- 可能用到的判断标准
    - 时间
    - 发post的内容，照片，定位
    - 用户画像：年龄、性别、职业、爱好
- 判断场景
    - Shopping的比较简单，直接看是否在某个shopping mall，或者在某个商场。如果更细化，你想判断用户今天来商场是shopping的就需要其他indicator：比如今天是不是节假日或者非工作时间，可以判断用户是不是有闲暇的时间；看用户有没有在post跟shopping有关的内容（比如今天要跟闺蜜去shopping之类的）；用户本身的feature，比如gender，年龄，收入，平时的爱好，是不是经常shopping；爱好跟来的shopping的地方是不是有高相关性；来的地方是不是经常来，而且来了基本就是shopping

- Home看地址，是不是match。看用户睡觉的地方在哪，起床的地方在哪，可以说大部分用户每天睡觉和起床的地方大部分时间都在家。看用户有没有带locaiton的post/picture，内容是在说我在家。这个其实也可以用来看有没有搬家什么的。所以也要留意用户有没有搬家倾向。
- Work：首先这个用户要有work。然后判断用户的工作时间：可以通过工作内容，跟工作相关Post，或者不怎么发post的时间，大致判断这个用户是白天工作还是晚上工作。然后就比较容易找到他工作时间基本在哪，就是工作地点了。还有个有趣的是，可以通过工作地点和用户的FB activity反推用户每天几点上班，几点出来吃饭，几点出来喝咖啡，几点下班。然后我们可以掌握这个pattern，给他推各种相关内容，比如上班时间，推各种新闻，杂志，排解路上的无聊，或者各种self help书籍推荐；午饭时间推各种饭店；咖啡时间推各种咖啡相关；下班的时候推餐厅，推各种下班后可以进行的活动等等。

Restaurant recommendation

1. 为什么要这个feature ,会带来用户行为怎么样的变化

- engagement
- ads revenue

2. 需要什么data做这个feature

- demographic: 出生地，性别，学校，年龄
- behavioral: like过的page，check in过的餐厅, 点赞记录，阅读记录
- network: 朋友喜欢的餐厅, 和朋友的互动, 朋友的记录也可以作为参考，特别是close friend
- outcome: conversion rate (CTR of recommended restaurant page)

3. 怎么evaluate你的模型是有效的

evaluate with labeled data (correctness of prediction on clicks)

4. 现在算法有了，选什么merics衡量结果

- product: dau, ctr, revenue
- ecosystem: overall engagement on FB (total time spent per week)

5. potential negative impact of this feature?

- Ads experience

6. Restaurant recommendation和people you may know的区别？

- goal: ads revenue vs. engagement
- training data:
    - travel & restaurant specific vs. all features
    - sensitive to location vs. not so sensitive to location
- recommendation: need to be precise vs. can be not 100% precise

Auto Play

facebook video auto play vs play video when users click on video (why time spent went down but dau went up)

- novelty effect: try new version.feature, but may not like it much
- learning effect: it takes time for users to
- bad trend: user is not interested in the change. although they are still using FB but are annoyed at the auto play and log out quickly
- good trend: they can find the content they want easily, so they spend less time looking for the content.

Ads Revenue

如果看到total ads revenue下降，你该怎么办？如何figure out原因？怎么给Senior management汇报

- 突然增加：对于revenue来说不太可能
- 缓慢增长

内部问题：

- 大化小，simpson paradox：看具体哪个部分有变化：哪个country？哪个market？哪一类demographic用户？哪一类内容的comment？新出现的retention是哪一类内容？哪一类产品？等等，都可以给我们不同的答案。
- 大化小，长期vs短期：可以建立一个模型，Y是DAU或者comment，X是我们能收集的所有的features，train到我们历史数据，看看哪个feature大大影响了我们的Y。这些feature就是短期的metrics，咱们的Y就是一个长期的metrics比如一年的retention。Features有：新产品新功能Binary，用户profile，用户browsing behavior，channel，用户network属性，时间因素等等。
- NLP模型可以用于feature engineering：用户的说话风格，topic 的热度，用户的sentiment变化等等
- 产品本身的变化：比如新功能。看看他们有没有给我们带来显著增长：AA或者AB testing。看使用前使用后短期内变化；看同类型用户，使用和不使用的区别

外部问题

- 市场变化，大事件（covid），竞争对手倒闭。

Local Business

1) 如何定义Local Business

provides goods or services to a local population

metric:

- locations listed in business page description
- ads targeting group (the location of users who buy services)
- ad objectives: direct response vs. brand advertising
- size of business/business start date
- price range
- product selection

2) fb会在那些small business 的news feed上放一些ads，这些ads的内容是那些small business比较popular的post，如果他们买了fb的广告，this is what it'll look like on your followers' news feed。请问用什么metric衡量这个ads的效应。

goal：让small business多买广告。
metric: engagement + ads revenue
downside: 没办法catch users，比如他看到这个ad，当时没点，但是后来又自己主动的买了广告，这个时候我们就没办法找到他了。

3) 实验发现广告CTR低，为什么？how to verify?

因为在一些国家FB不popular，大家（small business）觉得ads less relevant，质量不高，看到太多重复的广告，买了这个广告也没什么用，conversion很低
metrics: #impressions before creating an ad and use AB Test to measure

Meaningful Metric

现在别的组定义了一个metric A针对news feed产品怎么样看这个metric A是不是meaningful？

metric是不是meaningful 看它是不是够sensitive和robust，比如说这个metric的goal是测试engagement的变化，那如果产品没什么change或者不应该对engagement有影响时，它就应该不动（robustness），它在相似的用户之间也应该不变，但是如果我们有一点对engagement应该有影响的变化，它有能体现出来（sensitivity），这样的metric比较有意义

robustness可以用AA test, 比如说你对两个背景相同的用户群，放一样的feature，然后看这个metric是不是不变，我觉得是可以用hypothesis test的；sensitivity就可以用对相同用户组放效果不同的feature（你得知道这个feature带来的效果是啥）然后测这个metric能不能测出来你想要的feature之间的difference，有时候你不一定要做AB test,回去查看user log也可以，做retrospective analysis也行。

Report System

date | content_id | owner_id | reporter_id | is_valid (binary)
Q1: 被report的人中有百分之多少的人有至少一个的validate report
Q2: 你怎么来看有没有人abuse这个report system？算什么metrics，用sql写出来。

1) 如何具体定义那些滥用report system的user

没有被validate的report的绝对值 + 百分比（5%，10%）

2) 定这个threshold要考虑哪些问题，比如我后续的action是什么，让reviewer去confirm的话就要考虑一下人力审查的capacit

3) 这类行为为什么不好？

注册信息

脸书上有很多高中生注册。他们会被要求写学校。我们如何知道他们写的学校名字是不是正确的？你可以用任何脸书上的数据。

- check existing data: location, how long they stay in school
- social network: mutual friend, groups, events
- post content: NLP
- survey
- report system

Impression

X people, 一共有Y Impression随机分配

1) expected impression per audience？
   For one impression, the prob one person will see = 1/X (Each ad has uniform 1/X prob to be seen). n次实验成功了m次的binomial distribution
   p=1/X, n=Y
   E = np = Y/X

Let's check all the requirements of binomial distribution are valid:
1. Trials are independent (because we can assign an impression to any people irrespective of whether they already saw an impression or not)
2. Fixed number of trials (Y)
3. P(success) is the same across trials (1/X for every impression assignment we have to do)

2) probability of each person have at least one impression
   For one user
   - the prob of seeing no impression = (1-1/X)^Y
   - the prob that he/she will see at least 1 ad = 1-(1-1/X)^Y

   ~= 1-(1-Y/X) = Y/X
   这里用到了tyler expansion on binomial coefficient: (1 + x)^a ~= 1 + ax
   需要X很大

3) expected number of audience have at least one impression
   the number of people who will see at least one impression = X * (1-(1-Y/X)^Y)

插入广告

是关于脸书的广告。假如有两种不同方法来往信息流里插入广告。
第一种是：每个信息位子，我们以4%的概率来替换成一个广告。
第二种是：每25个信息，把其中一个变成广告。

Q1: For each option, what is the expected number of ads shown in 100 news stories? Variance?

1) binomial distribution: p=0.04, n=100, E=4, V=3.84
2) E=4, V=0

Q2: Probability of seeing over than twice of expected value

1) 1-p(x<=8)

normal approximation: P(x>8) = P(z > (8-4)/3.84) = 1-pnorm(1.04)

2) 0

Q3: Max number of back to back ads, which one is more likely

1) Expected value of back-to-back ads

expected number of ads collision, 就是连续两个post都是ads。等于是有了个新的Bernoulli process，概率p=0.04^2. 考虑collision是否independent，写了个模拟发现自己没算错，相邻两个点是否出现collision是同时决定的，所以independent

99*(0.04^2)

3*(0.04^2)

2) Max number of back-to-back ads

99 vs 2

3) Probability of seeing back-to-back ads

- 100个里面有k个广告，k>50则有连续。算k>50个广告的概率
  插空法：把100-k个广告插入k个广告中间
  求和k from 0 to 50（100-k+2 选 k）* 0.04^k *0.96^(100-k)
- 3*(0.04^2) - (0.04^2)^2

Conference Room

1) 一共有N个conference room from no.1 to no.N。有K个meeting独立随机分配到这N个conference room。现在已知1号conference room里面被schedule了一个meeting,问1号conference room里面被schedule的总共的meeting的数量（已知1号房存在一个meeting，也就是1号房不为空。在这个条件下求1号房总的meeting数的期望。把meeting的集合写成 M1,M2,...,Mk，对任意的i, 利用Bayes公式计算条件概率）

We can think of this as a binomial trial:
Success = Meetingi is scheduled in Room1
P(success) = 1/N
Let's check all the requirements of binomial distribution are valid:
1. Trials are independent (because we can schedule a meeting in any room irrespective of whether they have meetings scheduled or not)
2. Fixed number of trials (k)
3. P(success) is the same across trials (yes, this is 1/N for every meeting assignment we have to do)

P(room1 not empty) = 1 - (1-1/N)^k
P(Mi in room 1 | room 1 not empty) = P(Mi in room 1) / P(room 1 nonempty)
= (1/N) / 1 - (1-1/N)^k
E(number of meetings in room1) = np = k (1/N) / 1 - (1-1/N)^k = (k/N) / [1 - (1-1/N)^k]

2) 两个会议室，1/3的可能性是两个会议室都有人，1/3是一个有人一个没人，1/3是两个都没人。

- 你现在去了一个会议室，里面有人的概率多少
  1/3+1/2*1/3 = ½

- 如果你现在去了一个会议室，发现里面有人，那么另外一个有人的概率是多少（这个小问需要贝叶斯）
  P(B有人 | A有人) =（1/3）/（1/2）=2/3

选人

1. 1000 people, each time select 10 (w/o replacement)， 每个人on average在第多少次会被抽到？

一个人第一次被抽中的概率是1/100
一个人第二次抽签才被被抽中的概率是99/100*1/99=1/100
第n次才被抽中的概率是1/100，and n in [1,100]
E(#days)=1*P(D1)+2*P(D2)+...+100*P(D100)=(1+100)*100/2 * (1/100) = 50.5

2. 1000 people, each time select 10, (w/ replacement), 每个人on average在第多少次会被第一次抽到？

每个人每轮被抽到的概率是p=10/1000=0.01
Define success: 一轮中抽到这个人
Context: 在成功（第一次被抽到）之前每个人经历了多少次失败（没抽中）, 服从几何分布
mean=1/p=100

Bad Actor/贝叶斯

5%是bad user，95% good user。有一个good user go through模型的话，95%概率模型会预测是good user， bad user go through模型的话 95%的概率会预测是bad user，问如果模型预测一个人是bad user，那么真是情况下多大概率这个人是真的bad user
P(real bad | test bad) = P(test bad | real bad)*P(real bad) / [P(test bad | real bad)*P(real bad) + P(test bad | not real bad)*P(not real bad)]
= (0.95*0.05) / (0.95*0.05 + 0.05*0.95) = 0.5

Bayes's theorem
Bayes's theorem is stated mathematically as the following equation

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}$$

where A and B are events and P(B) is not zero.

- $P(A \mid B)$ is a conditional probability: the likelihood of event A occurring given that B is true
- $P(B \mid A)$ is also a conditional probability: the likelihood of event B occurring given that A is true.
- $P(A)$ and $P(B)$ are the probabilities of observing A and B respectively; they are known as the marginal probability.

- A and B must be different events.
- $P(A)$, the *prior*, is the initial degree of belief in A.
- $P(A \mid B)$, the *posterior*, is the degree of belief after incorporating news that B is true.
- the quotient $P(B \mid A) / P(B)$ represents the support B provides for A.

Reviewer

Careful reviewer and Lazy reviewer:
20% Lazy reviewer, always give good review
80% Careful reviewer, 60% good review, 40% bad review

1) Probability of a review being good
   20%*1 + 80%*60% = 0.68

2) If an ad gets a negative review, what is the probability that it's reviewed by a lazy reviewer?
   P(lazy | negative) = P(lazy, negative) / P(negative) = 0

3) Expected number of good reviews in 100 reviews
   100*0.68 = 68

4) If we have very few labeled data, how can we build a model to distinguish between careful and lazy reviewers?
   clustering: 这一类当中哪个label更多就是哪一类
   Bayes probability to calculate the probability of being a lazy reviewer given the number of good reviews.

5) A reviewer gave 3 good reviews, probability of him/her being lazy reviewer
   P(lazy | 3 good) = P(3 good | lazy)*P(lazy) / [P(3 good | lazy)*P(lazy) + P(3 good | not lazy)*P(not lazy)]
   = (1^3)*0.2 / [(1^3)*0.2 + (0.6^3)*0.8] = 0.53648

6) What would the above probability change as number of good reviews (N) approach infinity
   1

7) Based on the results, how would you design an approach to classify lazy and careful reviewer
   上面算出来的probability > 0.5那就classify as lazy
   criteria: P(lazy | data) = P(x|lazy)P(lazy) / P(x) > 0.5 then lazy
   x: data with # of reviews & # of reviews
   P(x|lazy)P(lazy) > 0.5P(x)
   P(x) = 0.5P(x|lazy)P(lazy) + 0.5P(x|not lazy)P(not lazy)
   So if P(x|lazy)P(lazy) > P(x|not lazy)P(not lazy) then lazy
   Indicator(P(x|lazy)P(lazy) > P(x|not lazy)P(not lazy))

   * 4) 中的probability is very dependent on the size of

8) What would your false positive/false negative rate be if use your approach

false positive: label as lazy but the user is careful
P(>0.5 | careful)
E(indicator) when the reviewer is careful

false negative: label as careful but the user is lazy
P(<0.5 | lazy)
E(indicator) when the reviewer is lazy

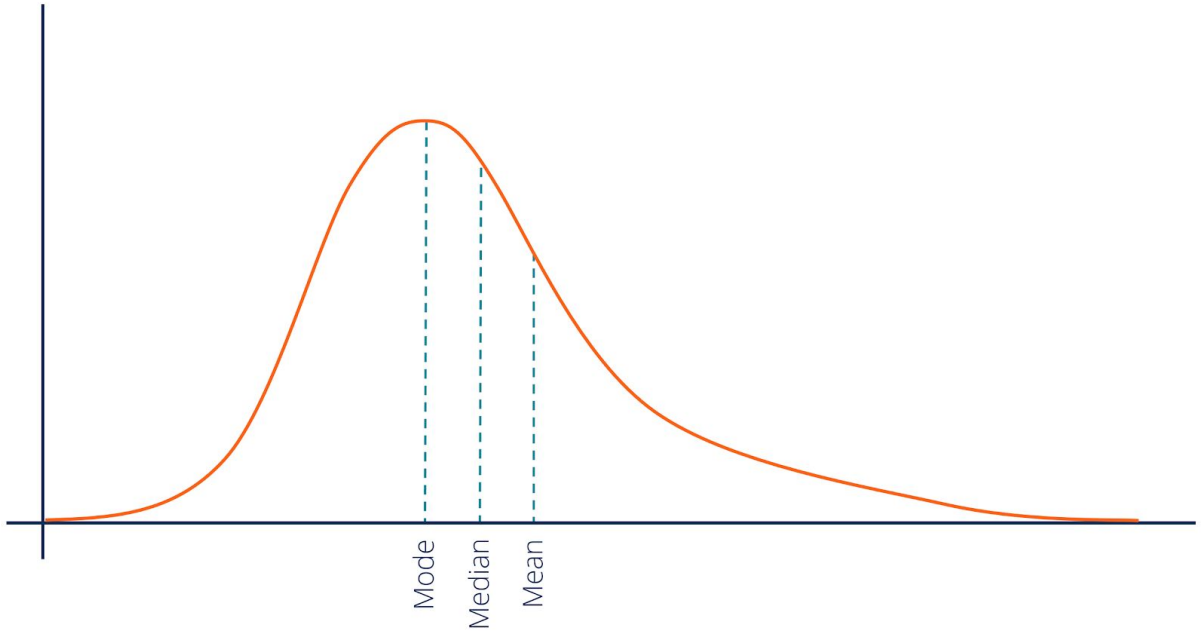
9) So what we should do is to make those reviewers do the same number of reviews (like 100), and then classify them. How would type I/II change now.

Distribution

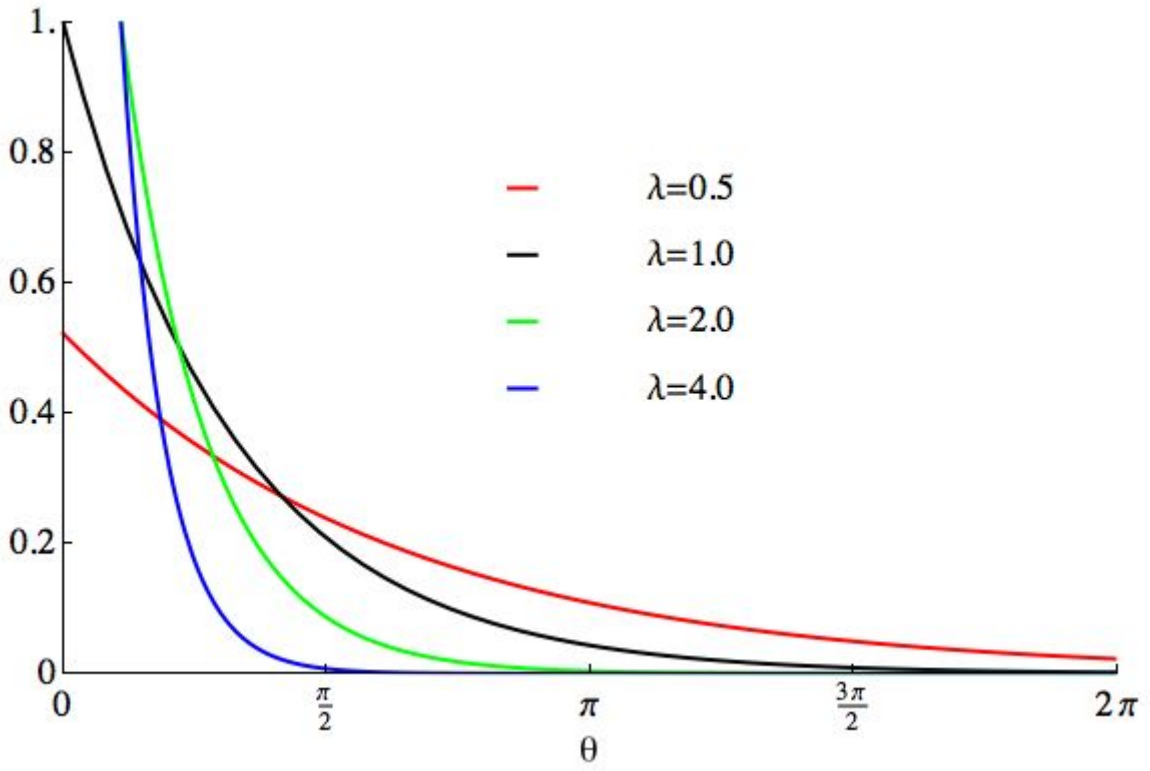
1. 画distribution of page shared (#users vs # page share)
    a. 标mean, median, p1, p99, 问根据经验mean是多少，median是多少

* 10% users would share content once per day, 18% more than once daily
* videos are the most shared content

- right skewed distribution (long tail):



- exponential distribution:



p1: 0
median: 1
mean: 2
p99: 20

b. 取出今天p50 和p95的两组用户，两周以后，你觉得这两组人分别的mean是多少？

如果取出分布在5% 的用户， 或者分布在95% 或者分布在mean周围的用户，他们未来三周，time spend distribution 还是right skewed / exponential。每天都记录5% 的那个数据，比如第一天是2.7，记录三周，这个5%数据应该是正态分布，mean就在这个5%数值附近。总之这个distribution 很稳定，5% 那个值估计没啥变化。同理类推median, mean.

c. 取出day 1所有page share=2的population, 问trends for day 2-30

如果是一个trend：a fluctuating curve with larger value during weekends and smaller value during weekdays. If we are running any campaigns or suppose FB is healthily growing and increasing user engagement, there might be an increase in the values in general.
如果是其中某一天的dist：波动比较大的长尾分布
如果是average daily dist：比较稳定的长尾分布

d. 取出day 1 所有page share=5的population, 问同样的trend。和c比哪个方差大，是什么分布？

a similar fluctuating curve with larger average # of sharing and smaller variance/smaller gap between weekends and weekdays. c has larger variance because users are not as engaged as users with 5 shares per day - they can just share two posts by chance.

2. dist of time spent per user，一般是一个right skewed dist

a. 画出5%，median，mean,95%，以及他们分别在未来三周的dist

* average time spent on FB is around 35min per day
p5: 0min
median: 30min
mean: 35min
p95: 100min
未来三周的dist: a fluctuating curve with larger value during weekends and smaller value during weekdays. Fluctuation is largering in p5, smaller in p95.

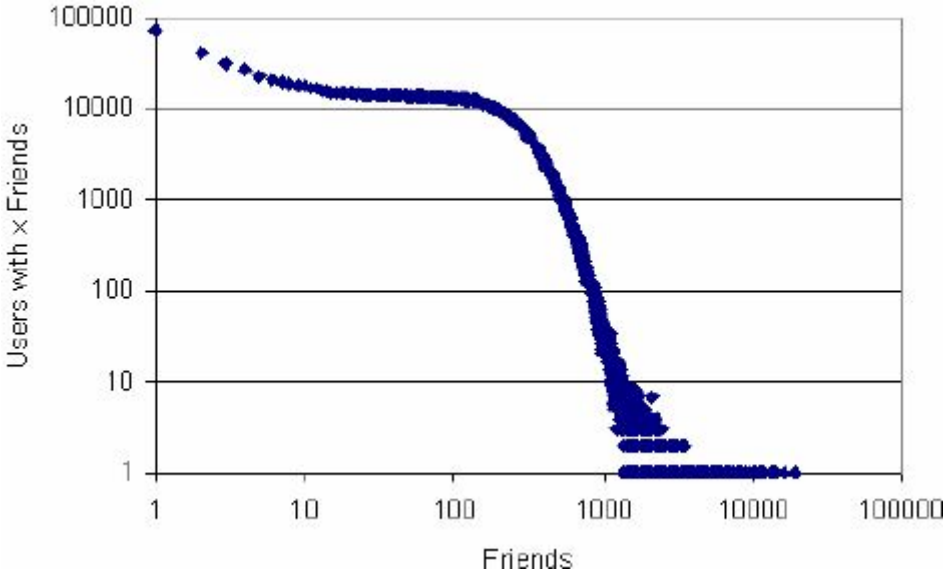
b. 取出今天分布在mean的用户，他们在未来30天的time spent dist应该是怎样的？Will the mean still be same or larger or smaller?

A fluctuating curve with larger value during weekends and smaller value during weekdays. The mean will remain the same.

Distribution of time spent/DAU
假设所有人同时在线，每人下线几率独立于已在线时间，则服从exponential

Distribution of #comments/DAU
假设每人每看到一个post有独立的p的几率comment，p较小，impression较多，则服从poisson
帖子的总回复数、转发数之类的分布，不是确定时间内的
社交网络的这些metric通常都是exponential 因为绝大部分帖子都没人回复 没人转发 一些名人的回复转发非常多 长尾巴 左边0的特别多

Distribution of #friends/user
假设每个friend request有独立的q的几率通过，则服从binomial或者poisson



comment per user
'daily-comment trend of median/99% cohort:
Think about Expected Value and Variance:
Expected Value: how usual for both cohort to achieve same number of comments
Variance: the size of the cohorts

CLT

用sample来估计population的计算区间，列一下公式 再用大白话解释下什么CLT

The central limit theorem states that if you have a population with mean μ and standard deviation σ and take sufficiently large random samples from the population with replacement , then the distribution of the sample means will be approximately normally distributed. For the random samples we take from the population, we can compute the mean of the sample means:

$$\mu_{\bar{X}} = \mu$$

and the standard deviation of the sample means:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$$

* 每一次抽出来的random sample需要独立

This will hold true regardless of whether the source population is normal or skewed, provided the sample size is sufficiently large (usually n > 30). If the population is normal, then the theorem holds true even for samples smaller than 30. In fact, this also holds true even if the population is binomial, provided that min(np, n(1-p))> 5, where n is the sample size and p is the probability of success in the population. This means that we can use the normal probability model to quantify uncertainty when making inferences about a population mean based on the sample mean.

Having the statistics for a random sample (u, sigma^2), to calculate the probability that the mean would be greater/smaller than a, we can use the formula of Z-score and look it up in the z-table or calculate in R by pnorm. For greater, it is 1-pnorm.

$$z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}$$

Confidence Interval

1. What is a confidence interval?

We can be 95% certain that the Mean of our sample lies within these two values.

DEFINITION:

The Confidence Interval expresses a range of values in which we are fairly certain that a value lies. The Confidence Interval has two parts. The Upper Confidence Interval and the Lower Confidence Interval.

CALCULATION:

In this example we will calculate the Confidence values for a single Mean. More often than not in Clinical Trials we are asked to present 95% Confidence intervals:

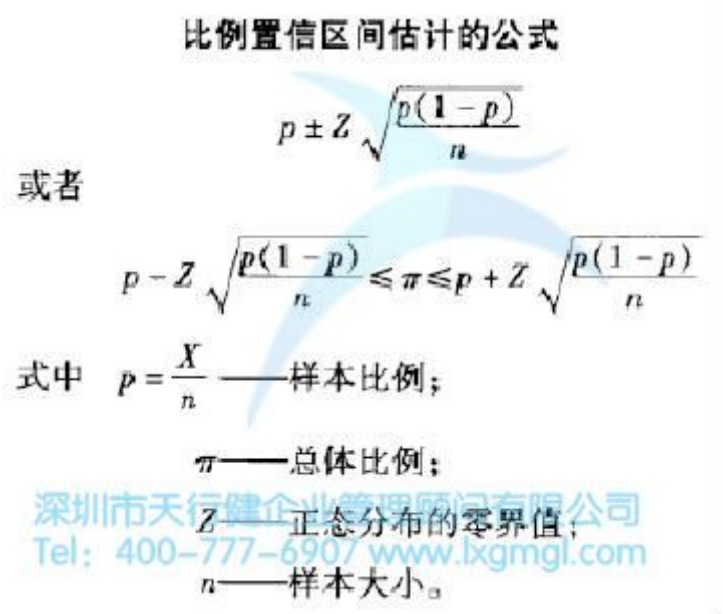
- Lower 95% Confidence Interval

$$Mean - \left(1.96 \times \frac{SD}{\sqrt{n}}\right)$$

- Upper 95% Confidence Interval

$$Mean + \left(1.96 \times \frac{SD}{\sqrt{n}}\right)$$

2. how to calculate the CI of X/Y？given they are normally distributed?


比例置信区间估计的公式

$$p \pm Z\sqrt{\frac{p(1-p)}{n}}$$

或者

$$p - Z\sqrt{\frac{p(1-p)}{n}} \leqslant \pi \leqslant p + Z\sqrt{\frac{p(1-p)}{n}}$$

式中 $p = \frac{X}{n}$ ——样本比例；

π——总体比例；

Z——正态分布的零界值；

n——样本大小。

3. interpret ab test result，treatment effect for each group as below：
   treatment1: -5 （-7.5，-2.5）
   treatment2: -15 （-17，-13）
   treatment3: -12（-28，-4）

how to interpret ci？which treatment to choose？increase test power/accuracy？分析这个AB test的结果，哪个group可以launch？ CI 更宽为什么？ 从user case角度分析user的反应

第二个最好，因为效果最好而且ci也不宽。第一个明显效果不如后面那两个，第三个虽然均值接近第二个，但是ci宽，说明variability比较大。综合看来，第二个最好.

practical significance也是需要看的，我刚才说的只是纯粹比较这三个ci而言，但是很有可能最好的那个带来的效益也是微乎其微的，打个比方，这三个药物都是降低血糖的，最好的那个可以降低15，那这个15有没有意义呢？那就得看practical significance了。 你要知道，只要sample size足够大，任何差距都可以变得显著，但这种显著不一定有实际意义。我觉得这个题吧，你要是最后提一下practical，会加分，但是不提呢，可能也不会减分

interval很宽是为什么，因为standard error高，为啥高？user segment对treatment的反应不一样，两极分化呗。会有什么outlier：bot user, inactive screentime。怎么处理：truncate tails, 或者用Winsorized mean.

linear regression, logistic regression, what are the metrics for goodness of fit.

How to select features?

a. stepwise methods
b. regularization method
c. PCA
random forest(variable importance) .
各个的优缺点。我从bias-variance tradeoff 的角度回答的。

从bias and variance (complexity and generalization)的角度：
1）stepwise regression model 一般都是很complexity很底的model, bias 会很高，variance 较低，比较容易under-fitting 在 training set上面，所以generalization on new data set不是很好，也就是predictive accuracy 不好。
2）regularization method and PCA method: regularization method can use the learning rate(shrinkage parameter) to strike a balance between complexity and generalization, and and at the same time select features. This method enjoys relatively lower bias compared to stepwise model, and also has low variance, therefore not easily prone to over-fitting. PCA is a more of a feature engineering method, which generates new features based on existing ones to capture as much variance as possible in the training dataset via SVD method..

3)Feature importance from tree-based model: it depends on which kind of tree model you use. Random forest is more towards to reducing variance and hence less prone to overfitting, but bias is not reduced significantly if any. Boosted trees is more towards to reducing bias and more prone to overfitting.