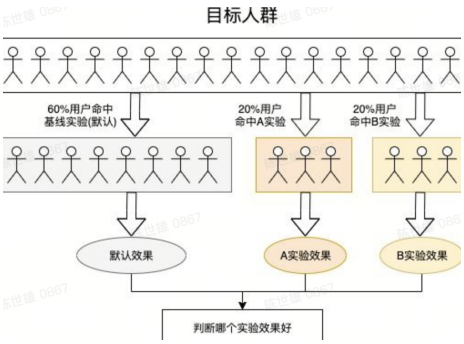
一.什么是AB实验

AB实验通俗讲就是在线上可以切出一部分用户（降低风险），完全随机的分成两组或多组（确保人群一 致），一组保持现有的方案叫对照组（或控制组），另外一组使用改进的方案叫实验组，使用统计的方法对两组 之间指标差异进行分析，评估是否符合预期的一种方法。



二.AB实验的起源

最早的AB测试本身是起源于医学。当一个药剂被研发后，医学工作人员需要评估药剂的效果。一般就会 选择两组用户（随机筛选的用户），构建实验组和对照组。用这两组用户来“试药”。也就是实验组用户给真的 药剂，对照组用户给安慰剂，但是用户本身不知道自己是什么组，都有医生指导。之后，在后期的观察中，通 过一些统计方法，验证效果的差异性是否显著，从而去校验药剂是否达到我们的预期效果。

三.为什么要做AB实验

a. 降低风险

b. 科学验证想法

c. 数据化驱动决策

d. 快速试错

四.AB适合场景

a. UI内容优化

b. 算法优化

c. 收益优化

d. 新功能效果评估

五.什么情况不适用

a.变量多：AB测试只适用于单变量

b. 产品不成熟：用户量少，实验效果可信度不高

c. 缺乏统计思维：最终的评价结论不严谨，不科学＜分析师兜底>

六.AB实验的基本原理

a. 抽样理论：流量分配

i. 完全随机

b. 假设检验：效果评估

七.AB实验的基本流程

a. 明确实验背景和目的

b.实验设计

i. 收集指标，建立指标体系

1. 核心指标：核心关注指标（依赖背景和目的）

2. 护栏指标：当护栏指标显著负向时，即使核心关注指标显著正向，我们依然不推荐推全实验（比 如：ltv,dau 用户充值，消费等等）

3. 是否需要增加埋点以支持以上的指标？如果有走埋点需求流程.

i. 实验条件确定：唯一可以加人群过滤条件的地方；推全固化的条件要跟实验条件一致

ili. 确定实验组逻辑和变量（变量和开发确认）：实验组不能加其他的限制条件 iv. 确定实验周期：（一般为7天），可以通过用户使用频率来判断产品周期

v. 确定实验所需样本量：

c. 实验开发：

i. 实验需要注意（九.常见问题．b）

d. 实验测试：

i. 添加白名单进行测试

e. 运行实验：

f. 数据回收：

g. 效果评估

i. 计算统计功效：

i. 决策方案：给出结论

ii. 实验报告：（如果实验有效果）

h. 推全固化：推全就需要代码固化（当前），固化时需要注意包含实验条件逻辑；

八.AB实验的分流

流量分配：如果有其他的人群条件需要先判断条件再调用AB；不然AB入组会有问题，影响最终的效果评价

i. 客户端：

1. app启动后，每5分钟调用一次AB分流服务（getUserExpList），得到该用户有可能进入的所有AB实 验组列表；

2. 当用户进到某个实验场景时（知道这个实验的code），这根据上一步获取到的AB实验组结果（通常根 据下发的变量）做对应的逻辑处理，如果有入组需要调用用户真实入组实验接口（reportUserExp）

3. 后续的埋点上报都需要带上用户的这个实验信息（所有入组的实验）

i.服务端：

1. 调用AB分流服务获取该实验的入组信息（getUserExp）

2. 根据入组信息做对应的逻辑处理

九.常见的问题

a.单变量原则：实验组之间的只能存在一个变量，就是分组的执行的逻辑不同；

i. 有问题的例子：同一个实验，对照组（展示蓝色）：条件是新用户（当天注册），实验组（展示黄色）：条件是 注册：2~7天

ii. 正确的是：实验条件：注册7天内 对照组（展示蓝色）实验组（展示黄色）

b.AB实验条件：

例子：实验人群条件注册时间7天内，对照组（展示蓝色） 实验组（展示黄色）

i. 错误的：客户端或服务先调用AB分流服务接口，再判断是否满足注册时间7天内，如果不是7天内的走 默认逻辑

ii. 正确的：先判断实验条件：即先判断是否7天内注册，如果是，再调用AB分流服务接口，再走对应的展示逻辑

c.开发提示：使用实验组变量来控制对应的逻辑而不是通过实验组id来判断，可以达到解耦的目的（AB实验 可以直接控制各个组的逻辑），注意：上报AB埋点需要用实验组id（目前的做法）.

i. 可以做AA实验

ii. 可快速重建实验，不需要发版.

d.实验推全：

i. 目前的AB平台推全逻辑是需要服务端固化对应的代码：例如实验需要推全 实验组（展示黄色）

1.服务端实验：需要移除从AB分流服务获取AB的代码，直接写死（展示黄色），后在AB平台上点击推全实验组按钮

2. 客户端实验：需要确认AB平台上的客户端版本是做实验的最高版本，然后在AB平台上推全对应的 实验组；客户端在下个版本写死（展示黄色）的逻辑

3.例如：某客户端实验在推全的时候，客户端版本最高：769；那在推全之前需要在AB平台上确认该 实验的版本为769，如果不是需要改成769后，再推全实验；之后的版本可能需要固化代码（即770 以后版本需要把推全的逻辑写在客户端里

ii. 实验推全后人群条件不能变

1. 例如：做实验的时候条件是注册7天内的用户，

a. 推全或固化代码后去掉了实验条件，变成全量用户；这个会使人群发生变化，导致效果的不确定 性.

b. 如果推全或代码固化需要的是全量用户，那么在做实验的时候就应该去掉实验条件.

e. 实验组逻辑变更：

i. 一个AB实验运行了一段时间，且有一定的入组用户，这时候调整实验组的逻辑：然后再观察实验组和 对照组的实验效果；这个是不严谨的，不科学的；实际上实验组的用户是经过了两个实验逻辑，而最终 推全后，所有的用户只有实验组的最后那个逻辑；从而导致最终推全的变量和实验的变量不一致，影响 了实验的科学性

i. 严谨的做法：关闭掉该实验，重开一个实验，用新逻辑做实验组；

f. 实验组关闭后重开：＜系统上应该禁止实验组关闭后重开＞

i 也是不科学的，原因和实验组逻辑变更一个道理：关闭后会走对照组的逻辑，重开又走实验组逻辑，把整 个过程当作是一个变量，与最终推全的实验组逻辑不一致，因此这样实验不严谨

i. 举个夸张一点的例子：实验组第一次的逻辑是给每天用户补贴50块钱，运行了3天，改成了补贴2块钱， 这样之前补贴50块的用户现在只有2块的那批用户由于心理预期落差，有可能会导致留存大幅下降， 而一直享受2块钱补贴的，留存相对比较稳定；

g. 实验评估有效果，上线后就一定有效果?

i. 不一定.只能说大概率有效果；

十.扩展知识

实验层

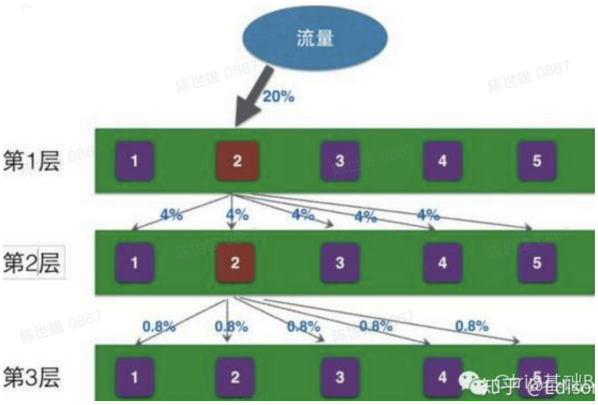
i. 背景：解决流量不够用的，流量需要复用

1. 同一时间需要做很多的AB实验
2. 每个AB实验都不互相干扰
3. 每个AB实验都能用到100%的流量

ii. 同层互斥（同一个用户，最多只能进入同层上多个实验的一个实验），层间正交（就是一份流量穿越每层 实验时，都会再次随机打散，且随机效果离散）

ii.互斥实验，就是让这两个实验放在同一个实验层上内容相同或相关、可能会彼此影响的实验

iv.如何实现：分流hash函数足够均匀，完全随机



隔离域

i. 背景：如何评估某个业务在一段时间内所有AB实验的效果总和

ii. 分类：（在AB实验平台需要去配置对应的隔离域）

1. 全局隔离域：全平台的所有效果评估

2. 业务隔离域：某个业务模块的效果评估

ii 基本原理：

1. 从全局流量中预留一部分流量，所有的（全局隔离域）或某业务（业务隔离域）的流量都走对照组逻辑

2. 在效果评估的时候用隔离域流量和对应（全局隔离域对所有；业务隔离域对该业务隔离域对应的所 有实验）实验域的流量做效果评估

辛普森悖论

在某个条件下的两组数据，分别讨论时都会满足某种性质，可是一旦合并考虑，却可能导致相反的结果

i. 例子：

某大学由两个学院组成。1号学院的男生录取率是75%，女生录取率49%，男生录取率高于女生；2号学院男生录取率 10%，女生录取率5%，男生录取率同样高于女生。问：综合两个学院来看，这所大学的总体录取率是否男生高于女 生?

学院	女生申请数	女生录取数	女生录取率	男生申请数	男生录取数	男生录取率	申请总数	录取总数	总录取率
1号学院	100	49	49%	20	15	75%	120	64	53.3%
2号学院	20	1	5%	100	10	10%	120	11	9.2%
总计	120	50	42%	120	25	21%	240	75	31.3%

ii. 如何避免?

1. 如果我们觉得某两个变量对试验结果都有影响，那我们就应该把这两个变量放在同一层进行互斥试验，不要让一个变量的试验动态影响另一个变量的检验。

a. 例子：如果我们觉得一个试验可能会对新老客户产生完全不同的影响，那么就应该对新客户和 老客户分别展开定向试验，观察结论。

b.i的例子就应该对男女用户分别展开定向实验，观察结论

场景：测试新推荐算法（B组）与旧日算法（A 组）的转化率。整体数据显示 B 组转化率（5.6%）低于A组 （7.3%），但按用户活跃度细分后发现

原因：低频用户占比高达 70%，其转化率绝对值低但权重高，拉低了B组整体表现。

根本原因：抽样流量有问题，实验中存在未发现的变量干扰项。