【从0到1】AB实验（待补充）

什么是AB实验

> 为了验证一个新的产品交互设计、产品功能或者策略、算法的效果，在同一时间段，给多组用户（一般叫做对照组和实验组，用户分组方法统计上随机，使多组用户在统计角度无差别）分别展示优化前（对照组）和优化后（实验组，可以有多组）的产品交互设计、产品功能或者策略、算法，并通过数据分析，判断优化前后的产品交互设计、产品功能或者策略、算法在一个或者多个评估指标上是符合预期的一种实验方法。
> --王晔博士（*耶鲁大学计算机科学博士*）

简单来说，AB实验就是针对需要验证的新feature，随机分流出两拨统计可比的用户，一组用户展示新feature，另一组用户不展示该feature（控制变量），经过一段时间观察其数据表现，结合统计学方法评估新feature的效果是否符合预期。

为什么做AB实验

> - 收益：希望能够**分离并量化**每个特性的影响，计算收益
> - 风险：无法承担任何一个错误特性**影响上亿用户体验**的严重后果；希望以**较小的风险**对新特性进行评估，积极试错积累经验；

对于性能实验，我们想通过ab实验关注的问题主要是：加载速度到底有没有提升？bug有没有减少，成功率是否提升？以及会不会引入其他bug，暴露风险？

AB实验背后的统计学原理

> 如何判断新feature是否有效？--> 怎么判断新feature带来了指标变化（数据正向/负向）？

假设检验

AB实验的核心统计学理论是（双样本）假设检验：首先做出假设，然后运用数据来检验假设是否成立；但需要注意的是，我们在检验假设时，逻辑上采用**反证法**；

- **原假设**：$H_0：\mu_1 = \mu_2$
  - 一般来说，产品ab实验的原假设是：新feature无效
- **备择假设**：$H_1：\mu_1 \neq \mu_2$
  - 一般来说，产品ab实验的备择假设是：新feature有效

将反证法迁移到AB实验中，我们需要证明原假设：新feature无效，是错误的（伪），并借此证明备择假设：新feature有效，是正确的（真）。

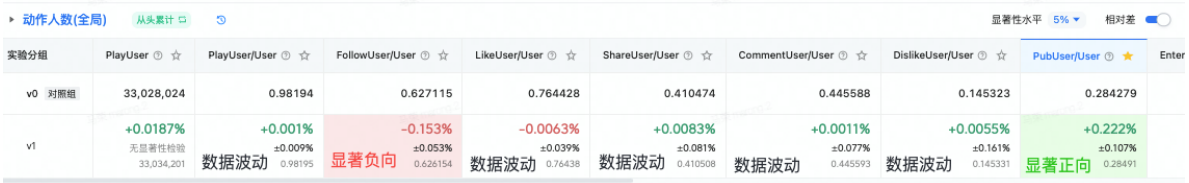
那么如何证明原假设是正确/错误的？ --> 两组指标表现是否相等

两类错误

在AB实验中，通常我们只随机抽取部分用户进行实验，因此AB实验天然存在抽样误差，即两组用户天生存在差异；那么如果看到了两组的数据表现存在差异，有两个可能：（1）抽样误差（2）新feature带来的变化；所以问题变成了：差异是来自抽样，还是新feature？

统计学上认为，概率低于5%的事件称为“不可能事件”，如果两种策略的指标变化值落在5%之后，那么统计学上被认为是小概率事件，有理由拒绝H0：两组均值不等，由此证明H1是正确的。

- **第一类错误 & 显著性水平(α)：** $P(\text{reject } H_0 \mid H_0 \text{ is correct})$ 假阳性
  - "实验显示改动有效，但其实无效"的概率，Libra默认取0.05；
  - **置信度：** “改动是有效的”的概率，置信度 = 1- α。
- **第二类错误 β：** $P(\text{accept } H_0 \mid H_0 \text{ is wrong})$ 假阴性
  - "做了有效的改动，但指标上不显著"的概率，默认取0.2；
  - **统计功效(Power)：** “改动是有效的，有多大概率能够被检出”，power = 1-β。
- **P_value：** 在原假设H_0正确的情况下，检验统计量出现当前值或者更极端值的可能性。



数据波动：底色透明	显著正向：底色绿色	显著负向：底色红色
抽样误差带来数据波动	95%的概率数据变化可信，新策略有效	95%的概率数据变化可信，新策略有效

常见名词解释

统计概念

名词	解释
一类错误	实验结论显示我的新策略有用，但实际上我的新策略没有用。这种被称为一类错误，假阳性。
显著性水平(α)Significance level	显著性水平是可能发生一类错误的概率，用α表示。在根据业界标准，α取0.05。
置信度Confidence level	置信度 = 1-α。在A/B实验中，如果发生误差的概率小于0.05，我们即称实验结论已经“统计显著/显著/可置信”。这意味着你采取的新策略大概率（A/B实验中意味着大于95%）是有效的。相反，如果这一事件的概率大于0.05，则称实验结论“不显著/不置信”。
置信区间 Confidence interval	置信度区间是用来对一组实验数据的总体参数进行估计的区间范围。 **例子**：我们开了一个ab实验，实验组采取新策略：推荐页分发特效卡片，投稿率提升**均值**为0.222%，置信区间为[0.115%, 0.328%]。 由于在AB实验中我们采取小流量抽样的方式，样本不能完全代表总体，那么实际上新策略如果在总体流量中生效，不见得会获得AB实验中的相应增长。如果我们设新策略在总体流量中推行所导致的真实增长率为μ，那么在上述例子中，μ的真实取值会在[0.115%, 0.328%]之间。在计算置信区间的过程中，我们会先取一个置信水平，计算这一置信水平下的置信区间是多少，AB实验中我们通常计算95%置信度下的置信区间。回到刚刚的例子，我们就可以得知，μ的真实取值有95%的可能落在[0.115%, 0.328%]之间。

名词	解释
二类错误	我的策略其实有用，但是没有检测出来。这是统计学中第二类错误的表现。这种错误的概率被记为β。
统计功效Power	统计功效（power，也被称为检验效力）= 1-β，表示的是“假设我的新策略是有效的，我有多大概率在实验中检测出来。”
MDE(Minimum Detectable Effect)	检验灵敏度，是指在当前条件下，我们所关心的指标，在实验中可检测出来的最小提升值。 这个提升值越小，也就意味着检验越灵敏。 MDE是基于多天累计数据计算所得，仅「多天累计」的指标可以查看MDE 用途：实验不显著，**我要结束实验吗？**如果指标的灵敏度比预期提升值大，那么我们可以将实验延长几天，再观察一段时间；如果灵敏度已经比预期提升值小了，那么很遗憾，我们的实验结果没有置信的可能了，另起炉灶吧。

检验灵敏度&最小可检测效果（MDE）

```
灵敏度和MDE对比：
-----------------------------------------------------
| MDE        | 1%                |<--- 我们希望检测到的效果
| 灵敏度      | 1.5%              |<--- 当前实验可检测的最小效果
-----------------------------------------------------
```

> **延长实验时间：**
>
> - 灵敏度高于MDE，延长时间或增加样本量来提高检测能力。
>
> **终止实验：**
>
> - 灵敏度小于或等于MDE，但未显著，考虑重新设计实验。

检验灵敏度

检验灵敏度是在特定条件下，实验能够检测到的最小效果大小（提升值）。它表示实验在多大程度上可以识别出我们关心的指标的细微变化。灵敏度越高，能够检测到的最小变化就越小。

影响灵敏度的因素

1. **样本量：** 样本量越大，灵敏度越高。随着样本量增加，检测出微小效果的能力增强。
2. **基线转化率：** 基线转化率（control group的平均指标值）越高，检测出相对变化的能力越强。
3. **方差：** 数据的方差越小，灵敏度越高，因为样本间差异小，可以更容易识别出真实效果。

4. **显著性水平：** 设定的显著性水平（如0.05）和统计功效（如0.8）也影响灵敏度。显著性水平越低，检测的标准越严格。

最小可检测效果（MDE）

MDE 是实验能够可靠检测到的最小效果大小。MDE 是一个预先设定的目标，反映了我们希望检测到的最小效果。例如，如果我们希望检测到至少1%的转化率提升，MDE就是1%。

MDE的计算

MDE的计算考虑了显著性水平、统计功效、样本量、基线转化率等因素。通常，MDE公式为：

$$\text{MDE} = \text{z} \times \sqrt{\frac{p \times (1-p)}{n} + \frac{p \times (1-p)}{m}}$$

其中：

- z 是标准正态分布的临界值（与显著性水平和统计功效有关）。
- p 是基线转化率。
- n 和 m 分别是实验组和对照组的样本量。

应用情境

在实际应用中，当我们进行AB实验时，如果实验结果在预期的时间内不显著，我们可以通过评估当前实验的灵敏度和MDE来决定是否继续实验。

实验不显著时的决策：

1. **灵敏度 > 预期提升值：** 如果当前实验的灵敏度高于我们所希望检测到的最小效果（MDE），即使结果不显著，也表明实验尚未有足够的能力检测到期望的效果。此时，我们可以选择延长实验时间或增加样本量，以期能够检测到微小的效果。
2. **灵敏度 < 预期提升值：** 如果灵敏度已经达到了甚至小于我们希望检测到的效果，结果仍然不显著，这说明即使延长时间或增加样本量也不太可能获得有意义的效果。此时，可以考虑终止当前实验，重新设计实验或尝试其他策略。

案例分析

假设我们在一个电商平台上进行AB测试，目标是提高转化率，我们希望检测至少1%的提升（MDE=1%）。实验进行了两周，但结果并不显著。我们计算当前实验的灵敏度（假设为1.5%）。

- **灵敏度 = 1.5% (> MDE = 1%)：** 实验当前只能检测到1.5%的最小提升，因此我们未能达到足够的灵敏度去检测1%的提升。这时，应该考虑增加样本量或延长实验时间以提高灵敏度。
- **灵敏度 = 0.8% (< MDE = 1%)：** 实验已经能够检测到小于1%的效果提升，但仍然不显著，这意味着实验可能没有产生所需的效果，继续实验也很可能不会改变这一结果。可以终止实验并尝试新的策略。

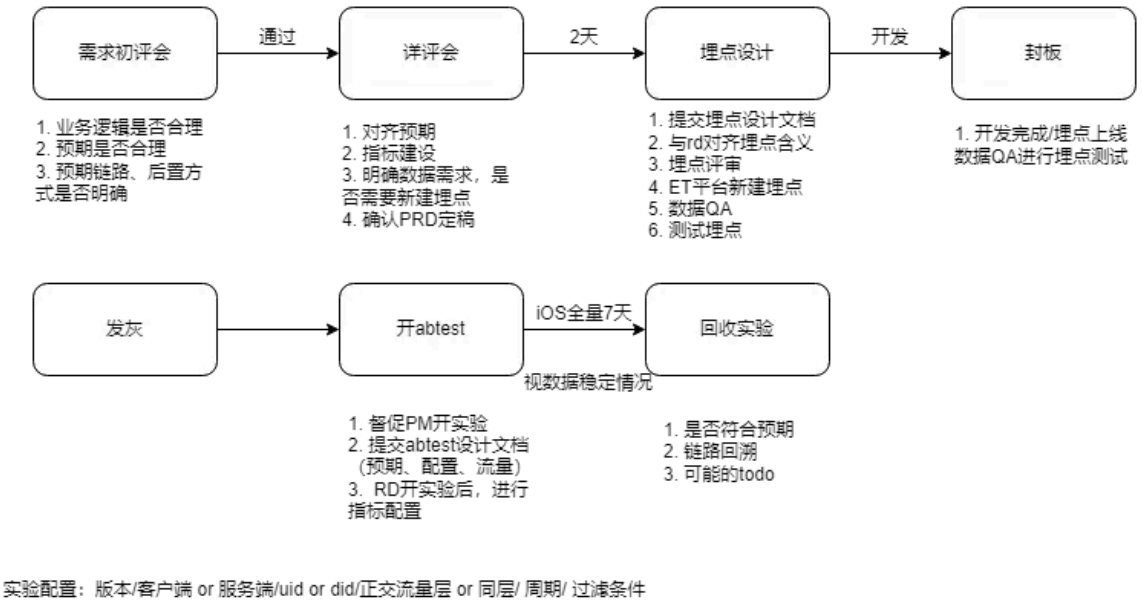
反转实验

反转实验（Reverse Experiment）是一种通过重新验证实验结果的策略，旨在了解改动的长期表现。反转实验涉及将实验组和对照组的角色对换，以观察长期效果的变化。它尤其有助于分辨短期收益是否会持续、扩大、或收敛。

什么情况下需要开反转实验？需要评估改动长期的表现，这里包括几种情况：

1. 短期有收益/观测不到收益，预期长期收益能够放大；
2. 短期有收益，可能是新奇效应，长期收益会收敛；

AB实验流程



实验配置：版本/客户端 or 服务端/uid or did/正交流量层 or 同层/ 周期/ 过滤条件