聊聊AA实验的波动性

转载

weixin_34268753 于 2018-10-22 02:59:50 发布 1847 收藏 13
原文链接：https://juejin.im/post/5bcd3cfcf265da0b001f69a9

> **当我们在实验评估系统上开启一个实验组和对照组配置一摸一样的实验时，我们称之为AA实验**。AA实验通常用来辅助观察指标在产品不做改变时的偏差范围。我们通常会在实验里加一个和对照组一模一样的实验组来观察这个偏差，而如果这个偏差很大，通常你的AB实验也容易结果不置信。

本文的目标受众是需要在实验评估系统上做实验，发现AA实验的指标差异很大，又懒的再回去翻大学概率论课本的同学们。以最低的学习成本使用实验评估系统拿到高效的产出，也是实验评估组的愿景，所以我们会用尽量通俗的语言展开描述，如果你看不懂，随时拿着你的水杯来砍我（别忘了先把热水倒掉）；当然，记得先留言指出没有说清楚的地方，我们会改正。

闲话少说，我们开始：

波动？啥叫波动？为啥我的AA实验指标会有波动？

举个例子。假设我在实验评估系统上开了一个AA实验。实验开启一段时间之后我们去看产出的实验指标 Read/U（平均每个用户每天会有多少次阅读），虽然分配到两个组的用户使用的是完全一样的产品，但是两个组汇总到的 Read/U 均值总是有多多少少的差别。如果你重复开这个实验很多次，你会发现每次两个组上的差别都不太一样。

这种出现在AA实验上的不稳定的指标差，就是我们说的波动

产生波动的原因很好理解，一句话来说就是“随机性”。下一秒打开头条的那个用户今天会读几篇文章这完全是随机的，不可预知的。所以当你开两个完全相同的实验组的时候，因为每个组里的用户今天会读的文章数完全随机，所以最终我们拿到的两个 Read/U 指标的差别也是随机的。

怎么描述AA实验指标的波动呢？

描述波动的方法很多，对应AB实验这个应用场景，我们用置信度和置信区间来描述波动性。如果你忘掉这两个统计学概念的话，就不要去网上查了，简单说就是：

你做无数多次AA实验，指标的差落在某个范围内（置信区间）的概率有多大（置信度）

假如我们知道头条主app的 Read/U 指标，200W入组用户的AA实验在置信度为95%的时候上下波动0.62%，说明大概率下，我们做一个AA实验，Read/U 指标的变化比例会在正负0.62%以内。

如果你做的AB实验预期 Read/U 会上升 1%，那么恭喜你，做实验验证去吧；如果你做的AB实验预期 Read/U 会上升 0.1%，那么不好意思，这个变化太不明显了，假如最终实验结果真的上升了0.1%，我们很难判断这是策略生效导致的还是波动导致的。

那么问题来了：

> 【问】你告诉我的波动在0.62%，为啥我的AA波动出现了0.78%？ 【答】因为你有95%的概率波动在0.62%以内，还有5%的概率你会遇到指标超过0.62%。出现这种意外的概率（5%）还是要比买彩票中奖高太多。最简单的办法就是重新再开一次实验。

【问】5%的意外概率我无法承受，怎么办？【答】那就看看置信度为99%的波动值吧。当然，这个数字一定会比95%的波动值大，比如说0.81%。也就是说只有1%的概率，你的AA实验波动会超出0.81%。

【问】可是我的预期变化只有0.68%，不要说0.81%，就是0.62%，变化也不够明显啊！【答】加流量吧。试想一下你在掷硬币，你抛硬币的次数越多，拿到正面的次数越接近0.5, 这说明**实验的越多（进组用户数越大），指标的结果越稳定（波动越小）**。当入组用户数升高到800W时，你会发现95%的置信度下，波动会降低到0.31%。【是的，你猜对了，波动与用户数的平方根呈反比，所以用户数升4倍，波动会降一半，如果感兴趣，回去复习概率论吧】

分流不是均匀的吗？怎么入组用户数也有波动？

用户的潜台词是：“你们在逗我吗？”还真没有。再举抛硬币的例子，分流的时候一个用户会进入哪个组就好比抛硬币时会出现哪个面。因为进入哪个组和出现哪个面一样，都是**随机的**。所以无论分流策略多么完美，入组用户数和其他指标一样，都存在波动性。

什么影响波动性？不同产品的相同指标波动为何不同？

入组用户数

入组用户数对波动的影响前面说过，这个不难理解。入组用户数越多，波动性越小。所以当做实验的同**学发现AA波动很大时，可以考虑一下提高实验流量来提高入组用户数数量，从而降低AA波动。**

指标稳定性（标准差或方差）

指标标准差描述的是指标取值的稳定程度。举一个射箭的例子。如果有小张小王两个人射箭，平均都拿8环，小张比较稳定，大多数时候都射中8环，少数时候射中7环和9环；小王发挥很不稳定，大部分时候要么10环要么6环。如果小张先射100次算平均分，再射100次算平均分（等价于我们在小张这里做了一次AA实验），两个平均分的差别体现的就是波动性。很显然，小王指标的波动性要大很多，因为他本身射箭不稳定。

类似的，例如 Comment/U 指标，因为它的取值稳定性特别差，所以我们总是能看到这个指标的波动要大于 Read/U 这些稳定性稍好的指标。怎么描述指标的稳定性呢？算个标准差吧。

有同学曾经问过为什么相同的指标相近的入组用户数，在同一个产品的国内版本和国外版本波动不一样。可以简单的看看这个指标在两个版本上的标准差，如果不出意外，波动大的那个版本的标准差也会较大。

你们是怎么计算波动性的呢？

先辈们提出并证明了一条统计学公式，简单且不严谨的说就是，如果已知母本的期望与方差，那么从该母本上的任意样本数量为N的采样得到的期望满足正态分布；正态分布的参数与母本的期望，方差和样本数量N有关。

如果我们把某个app一整天全量的日志数据作为母本，AA实验不过是在考察两个采样样本的期望的变化比。期望和方差我们都有，套入公式，我们就能得到发生在这一天的所有指标的波动性，并以此推测明天这些指标在相同app下的波动性。

其他资料

如果周末不需要陪男/女朋友，而且上面那些看得不过瘾的话，请阅读《概率论与数理统计》，随便谁出版的哪个版本都可以。