05｜选取实验单位：什么样的实验单位是合适的?

2020-12-17 张博伟

A/B测试从0到1

进入课程 >

讲述：张博伟

时长 15:14　大小 13.96M

你好，我是博伟。

上节课我们确定了实验的目标、假设以及各类指标，那么今天我们就来讲一讲 A/B 测试的第三步：如何选取合适的实验单位。

前面我提到，A/B 测试的本质就是控制变量实验。既然是实验，那就要有实验单位。毕竟，只有确定了实验单位，我们才能在这个单位层面进行合理的样本分配（Assignment），从而决定哪些样本在实验组（Treatment/Test Group），哪些样本在对照组（Control Group）。

谈到实验单位，你可能会问，这有什么难理解的，实验单位不就是用户吗?

其实，这是一个非常常见的认知误区。除了测试系统的表现外，在绝大部分情况下，准确地说，实验单位都是用户的行为。因为**我们在产品、营销、业务上所做的调整，本质上都是为了观察用户的行为是否会有相应的变化**。

那么问题就来了，很多单位都可以表征用户的行为。那到底是以用户为单位，以用户的每次浏览、访问为单位，还是以用户浏览的每个页面为单位呢？

这节课，我们就来学习下常用的实验单位有哪些，以及实践中选择实验单位的三大原则。

实验单位有哪些？

虽然可以表征用户行为的实验单位有很多，但综合来看，我们可以从用户层面、访问层面和页面层面这三个维度来分别学习。

用户层面（User Level）

用户层面是指，把单个的用户作为最小单位，也就是以用户为单位来划分实验组和对照组。

那么，具体到数据中，用户层面都包括什么呢？其实，主要是 4 种 ID。

第一种 ID 是**用户 ID**，也就是用户注册、登录时的用户名、手机号、电子邮箱，等等。

这类 ID 包含个人信息，它的特点就是稳定，不会随着操作系统和平台的变化而变化。用户 ID 和真实的用户一般是一一对应的关系，也是代表用户的最准确的 ID。

第二种 ID 是**匿名 ID**，一般是用户浏览网页时的 Cookies。

Cookies 是用户浏览网页时随机生成的，并不需要用户注册、登录。需要注意的是，用户使用的 iOS 和安卓操作系统也会随机生成 Cookies，但是这些 Cookies 仅限于该操作系统内部，和用户浏览时使用的设备或者浏览器有很大关系。所以，综合来看，Cookies 一般不包含个人信息，而且可以被抹除，因此准确度不如用户 ID 高。

第三种 ID 是**设备 ID**。它是和设备绑定的，一旦出厂就不可改变。设备 ID 虽然不会被抹除，但是如果用户和家人、朋友共享上网设备的话，它就不能区分用户了。所以，设备 ID

的准确度也低于用户 ID。

第四种 ID 是 **IP 地址**，它和实际的地理位置以及使用的网络都有关系。

同一个用户，即使用同一个设备，在不同的地方上网，IP 地址也是不同的。同时，在一些大的互联网提供商中，很多用户往往共享一个 IP 地址。所以，IP 地址的准确度是最差的，一般只有在用户 ID、匿名 ID 和设备 ID 都得不到的情况下，才考虑使用 IP 地址。

这就是用户层面的 4 个实验单位，它们的准确度从高到低的顺序是：

用户 ID > 匿名 ID（Cookies）/ 设备 ID > IP 地址。

为什么我要强调这 4 种 ID 类型的准确度呢？这是因为，**实验单位的准确度越高，A/B 测试结果的准确度才会越高。**

因此，当我们确定了选择用户层面的实验单位时，如果数据中有用户 ID，就优先选择用户 ID；如果数据中没有用户 ID，比如用户出于对隐私的考虑没有注册和登录，或者是测试网页的功能无需用户注册和登录，那么就可以选用匿名 ID 或者设备 ID；当这些 ID 都没有时，再选择准确度最低的 IP 地址。

访问层面（Visit/Session Level）

访问层面是指把用户的每次访问作为一个最小单位。

当我们访问网站或者 App 的时候，都会有后台系统来记录我们的每次访问动作。那么，我们怎么定义一次访问的开始和结束呢？

访问的开始很好理解，就是进入到这个网站或者 App 的那一瞬间。但难点就在于怎么定义一次访问的结束。在一次访问中，我们可能会点开不同的页面，上下左右滑动一番，然后退出；也有可能只是访问了一下没有啥操作，甚至都没有退出，就进入了其他的页面或者 App。

因此，考虑到用户访问的复杂性，通常情况下，如果用户在某个网站、App 连续 30 分钟之内没有任何动作，系统就认定这次访问已经结束了。

如果一个用户经常访问的话，就会有很多个不同的访问 ID。那在进行 A/B 测试的时候，如果以访问层面作为实验单位，就可能会出现一个用户既在实验组又在对照组的问题。

比如，我今天和昨天都访问了极客时间 App，相当于我有两个访问 ID，如果以访问 ID 作为实验单位的话，我就有可能同时出现在对照组和实验组当中。

页面层面（Page Level）

页面层面指的是把每一个新的页面浏览（Pageview）作为最小单位。

这里有一个关键词“新的”，它指的是即使是相同的页面，如果它们被相同的人在不同的时间浏览，也会被算作不同的页面。举个例子，我先浏览了极客时间的首页，然后点进一个专栏，最后又回到了首页。那么如果以页面浏览 ID 作为实验单位的话，这两个首页的页面浏览 ID 就有可能一个被分配到实验组，一个被分配到对照组。

到这里，我们就可以对比着理解下这三个层面了。

1. 访问层面和页面层面的单位，比较适合变化不易被用户察觉的 A/B 测试，比如测试算法的改进、不同广告的效果等等；如果变化是容易被用户察觉的，那么建议你选择用户层面的单位。
2. 从用户层面到访问层面再到页面层面，实验单位颗粒度越来越细，相应地可以从中获得更多的样本量。原因也很简单，一个用户可以有多个访问，而一个访问又可以包含多个页面浏览。

看到这儿，你可能觉得信息量有些大，这么多单位，具体操作时到底怎么选呢？不用担心，下面我就通过一个“视频 App 增加产品功能来提升用户留存率”的具体案例，来带你一步步地选出合适的实验单位。

一个案例：如何选择实验单位？

某视频 App 最近收到了不少用户反馈，其中很大一部分用户希望在没有网络或者网络不好的情况下也能看视频。于是，产品经理希望增加“离线下载”的功能，来提高用户的留存率。

现在，产品经理要通过 A/B 测试，来看看增加“离线下载”的功能是否真的能提升留存。那应该怎么选取实验单位呢？

如果把用户层面的 ID 作为实验单位的话（即把每个用户作为最小单位来分组），由于收集样本的时间比较紧迫，可能收集到的样本量就不够。因此，我们要去寻找颗粒度更细的实验单位，来产生更大的样本量。所以，我们可以选择访问层面或者页面层面作为实验单位。

数据分析师通过查看发现数据中有访问 ID，但没有 pageview ID，所以这里选择访问层面，把每一次访问作为最小单位来分组，因为一个用户可以产生多次访问。

这样一来样本量是足够了，但是我们分析计算实验结果之后发现，实验组的用户的留存率不仅没有上升，反而低于对照组。

这就很奇怪了，难道是因为“离线下载”功能导致用户体验变差了吗？这不是和之前用户反馈的结果相反了吗？

于是，我们再次对这些用户进行采访调研，得到的结论确实是用户体验确实变差了，但并不是因为用户不喜欢新增加的功能。那么问题究竟出在哪儿了呢？

其实，这里的问题就在于选择了不恰当的实验单位。在刚才的实验中，我们把每一次访问作为最小单位来分实验组和对照组，就造成了同一个用户因为有多个访问而被分到了不同的组。

所以，用户在实验组时可以使用新功能，但是被分到对照组时就会发现没有新功能，让用户很困惑。就好比，昨天你还在用一个很好用的功能今天突然消失了，是不是很沮丧呢？

所以，当业务的变化是用户可以察觉的时候，我建议你一定要选择用户层面作为实验单位。

在这种情况下，如果样本量不足，那就要和业务去沟通，明确样本量不足，需要更多的时间做测试，而不是选取颗粒度更小的单位。如果不能说服业务方增加测试时间的话，我们就要通过其他方法来弥补样本量不足会给实验造成的影响，比如增加这次 A/B 测试使用的

流量在总流量中的比例，选用波动性（方差）更小的评价指标等方法（我会在第 9 节课和你讲这些方法）。

回过头我们再看看这个案例，是不是可以提炼些选取实验单位的经验和坑点呢？没错儿，我将其归纳为了三个原则：

1. 保证用户体验的连贯性。
2. 实验单位应与评价指标的单位保持一致。
3. 样本数量要尽可能多。

掌握了这三条原则，你就能根据实际情况去选择最佳的实验单位啦！

确定实验单位的三大原则

1. 保证用户体验的连贯性

保证用户获得最好的体验几乎是所有产品的目标之一，用户体验的连贯性尤其重要，视频 App 的例子告诉我们：**如果 A/B 测试中的变化是用户可以察觉到的，那么实验单位就要选择用户层面。**

否则，同一个用户同时出现在实验组和对照组，就会体验到不同的功能、得到不同的体验。这种体验的不连贯性，就会给用户带来困惑和沮丧，很容易导致用户流失。

2. 实验单位要和评价指标的单位保持一致

为什么这么说呢？我们还得从统计学上入手去理解。

A/B 测试的一个前提是实验单位相互独立且分布相同的（Independent and identically distributed），简称 IID。如果两个单位不一致，就会违反相互独立这一前提，破坏了 A/B 测试的理论基础，从而导致实验结果不准确。

举个例子。如果用 A/B 测试来检测音乐 App 推送新专辑的效果，评价指标为用户的新专辑收听率（收听新专辑的用户数量 / 收到推送的用户数量），这里评价指标是建立在用户层面上的，那么实验单位一般也要为用户。

假如我们把实验单位变为新专辑页面层面，由于每个用户可以多次浏览该页面，所以对于同一个用户的多次页面浏览，每次页面浏览其实并不是独立的，IID 的假设前提就被破坏了，那么实验结果也就变得不准确了。

所以，在选择实验单位时，你一定要记住：**A/B 测试中的实验单位应与评价指标的单位保持一致。**

3. 样本数量要尽可能多

在 A/B 测试中，样本数量越多，实验结果就越准确。但增加样本量的方法有很多，我们绝对不能因为要获得更多的样本量，就选择颗粒度更细的实验单位，而不考虑前面两个原则。

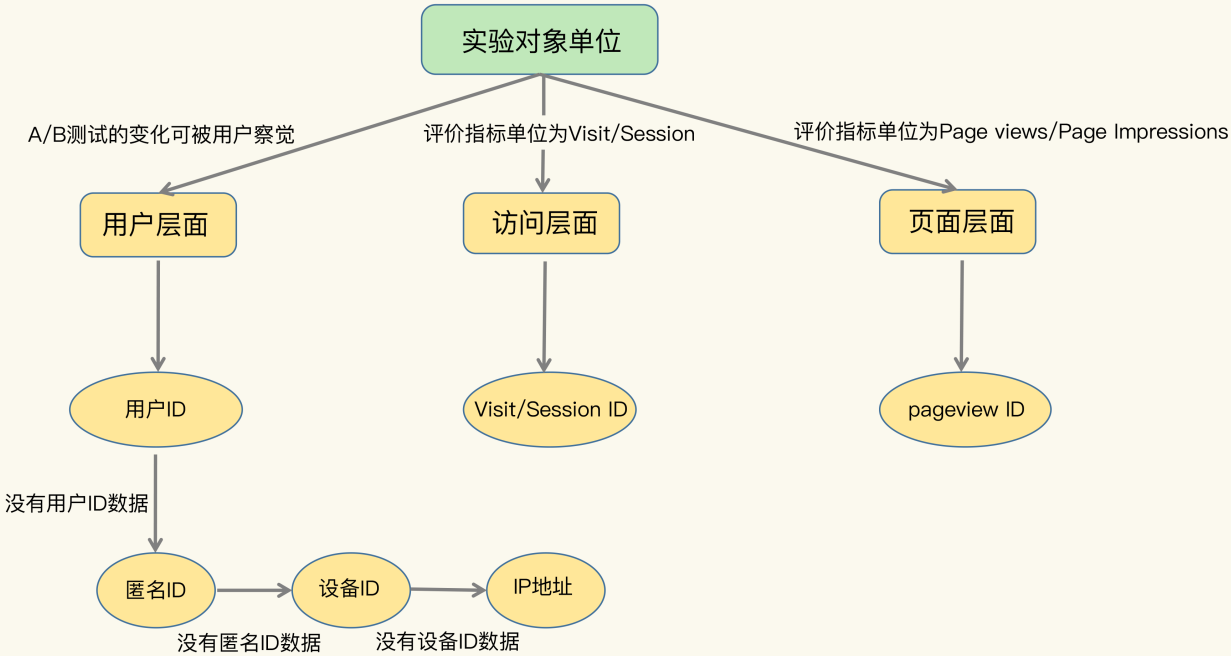
所以我们选取实验单位的第三个原则就是：在保证用户体验连贯性、实验单位和评价指标的单位一致的前提下，可以尽可能地选择颗粒度更细的实验单位来增加样本量。

那么现在三个原则就讲完啦，我来给你总结下：前两个原则是一定要考虑和满足的，第三个原则则是锦上添花，有条件的情况下可以考虑。

小结

这节课，我详细讲解了实践中常用的实验单位及其适用范围，也结合我的实际经验，给你总结了选取不同单位时需要考量的主要因素，让你真正理解并掌握背后的逻辑，从而帮助你在将来的实践中做出正确的判断。

我还给你总结了一个简化版的决策图，便于你回顾和记忆：



在实践中，我们要考虑的最重要的两点就是：**用户体验的连贯性、实验单位和评价指标单位的一致性**。毕竟用户是上帝，维持好的用户体验适用于所有的业务 / 产品。所以，针对用户可见的变化（比如 UI 的改进），大部分的实验都是把用户作为最小的实验单位（用户 ID/ 匿名 ID/ 设备 ID），同时也把用户作为评价指标的单位。

如果你想要更多的样本量，同时 A/B 测试的变化是用户不易察觉到的（比如推荐算法的提升），可以用比用户颗粒度更细的访问或者页面作为实验单位。与此同时，也要让评价指标与实验单位保持一致。

思考题

你平时做 A/B 测试时，是不是都以用户为单位的？学完了这节课以后，你可以再回想一下，有些 A/B 测试是不是可以用其他单位？为什么？

欢迎在留言区写下你的思考和想法，我们一起交流讨论。如果你有所收获，也欢迎你把今天的内容分享给你的朋友，一起共同进步！

提建议

上一篇 04 | 确定指标：指标这么多，到底如何来选择？

下一篇 06 | 选择实验样本量：样本量越多越好吗？

精选留言 (4)

写留言

那时刻 置顶

2020-12-17

我们一般都采用用户为实验单元，为了保持老用户体验的连贯性，大都在新用户做AB测试，然后把效果好的体验推广到老用户。不知老师如何处理这种情况呢？也就是保持老用户体验连贯性。
因为使用新用户，会遇到老师提到样本量不足，需要更多的时间做测试。请问老师，这个测试时间如何来把控呢？我们目前采用的是尽量拉长测试时间。

展开

作者回复: 你好，这个具体要看实验中改变的什么，对用户的体验影响到底有多大，只用新用户做实验的可能会有bias,因为新用户和老用户对变化的反应可能是不同的，一般只要不是特别显眼的变化都可以在全体用户上来做的，如果变化比较显眼比较大的话，可以先从总体流量的一小部分来做实验，然后慢慢扩大流量，在扩张的过程中就可以监控这个变化有没有什么明显的负面效果，及时发现问题即使解决。
至于你的第二个问题，可以到时候看一下第6节课如何计算样本量和第9节课中如何提高测试的Power.

1

xiaomin 置顶

2020-12-20

请问老师，以用户层面做AB实验，实验持续多天，那用户多天的行为是要累积起来比较吗？还是可以以一人天作为一个样本？此时用户分组是不变的，但以一人天做样本样本量会更多一些，不知道这种做法是否有问题？

展开

作者回复: 你好，如果在用户层面上做实验，实验持续多天，每个用户的行为是要累积起来比较的，不能按照user/day这种结合做为单位，你的前提是用户层面所以我假设你的评价指标也是用户层面，在这个前提下每个user/day这种结合其实是不独立的，昨天的用户A和今天的用户A都是用户A的行为。

1

西西
2020-12-21

想不出来visit level和page level为实验单元的实验有什么样合适的案例，page里文案或者按钮颜色的改变算用户可察觉吗？

展开

作者回复: 你好，很多算法（比如推荐算法，排序算法等）的改进是以visit level和page level为实验单位的，因为用户无法明显察觉；page里文案或者按钮颜色的改变用户一般是可以察觉的，推荐用户层面的单位

金hb.Ryan 冷空氣...
2020-12-19

原先应该就是用用户或者cookie来做A/B。
有没有场景是用访问层或者页面层来做A/B分组的？

作者回复: 你好，测试用户不易察觉的的变化很多都使用访问或者页面层面的，比如各种算法的改进，这样的好处是样本量比用户层面的更大。

1