03｜确定目标和假设：好的目标和假设是什么？

2020-12-12 张博伟

A/B测试从0到1

进入课程 >

讲述：张博伟

时长 19:31　大小 17.88M

你好，我是博伟。

今天这节课我们就进入到“基础篇”模块了，通过前面的学习，你已经清楚了做 A/B 测试的基本流程，接下来呢，我会带你去看看在实践中确定目标和假设、确定指标、选取实验单位、估算样本量大小，以及分析测试结果这 5 步，具体应该怎么操作。

我们知道，确定目标和假设、确定指标这两步决定了测试的方向，可谓至关重要。那么，如何一步步地把业务问题转化为 A/B 测试的目标和假设呢？又如何根据目标来选择合适的指标呢？在接下来的两节课，我会通过大量的案例来给你解答这两个问题。在讲解案例的同时，我也会结合我的实践经验，给你一些可落地执行、切实可操作的建议，让你知道该如何规避坑点。

确定目标和假设

首先，我们要明确，做 A/B 测试肯定是为了解决业务上遇到的问题，而绝不是为了做而做。所以，找到了要解决的业务问题，也就基本找到了 A/B 测试目标。为什么这么说呢？

让我们来回顾下开篇词中讲的 A/B 测试解决的常见业务问题，看看 A/B 测试可以用在什么领域，解决什么问题：

A/B测试可以解决哪些常见的业务问题?

产品迭代	算法优化	市场营销
如何改变用户的交互界面来提升用户体验?	如何通过提高推荐系统算法的准确度来提升用户粘性?	如何确定最优的营销内容?
如何优化新用户的注册流程来提高转化率?	如何通过提高搜索排名算法的准确度来提升结果的点击率?	如何确定最优的营销时间?
如何确定产品优惠券的最有价值?	如何通过提高广告显示算法的精确度来提升广告的点击率?	如何确定最精准的受众群体?
如何增加产品功能来提升用户留存?		如何衡量市场营销的效果?

总结一下这些业务问题，我们就会发现一些共性：

1. 所有的业务问题都会有一个目标，比如提升用户粘性是业务问题的目标，同时我们也把这个目标称作“**结果**”。
2. 有的业务问题会有明确的努力方向，比如，通过改变外观来提升点击率，这里的“改变外观”就是明确的努力方向，同时我们也把“改变外观”等变化称作“**原因**”。不过有的业务问题没有明确的努力方向，这时候我们需要根据具体的情况去发现原因。比如对于“如何确定最优的营销时间”这个业务问题，我们分析发掘之后会发现，周五晚上的营销效果会比较好。那么这里的“原因”就是大家结束了一周忙碌的工作，就会比较有时间。

你看，**把产品 / 业务的变化作为原因，把业务目标变成结果，我们就把业务问题转换成了因果推断**。而对于做 A/B 测试来说，把业务问题转换成因果推断，也就意味着找到了测试

的目标。**所谓的假设，在 A/B 测试的语境下，就是既包含了想要做出的改变，又包含了期望达到的结果。**

接下来，我就以一款按月付费的音乐 App 要提高营收为例，带你看看该如何确定目标和假设。

首先，分析问题，确定想要达到的结果。

想要提高营收，我们首先得清楚问题出在哪里。这个时候，我们可以进行数据分析。比如，和竞品进行对比分析后发现，我们 App 的用户留存率低于行业平均水平。因此，用户留存率就是我们这款 App 目前存在的问题。

其次，提出解决业务问题的大致方案。

影响用户留存的原因有很多种。比如，内容是否足够丰富，能满足不同用户的音乐需求？产品是否有足够多的便利功能，可以给用户更好的使用体验？App 的开启和运行速度是否足够流畅？

通过进一步的分析发现，我们的产品在歌曲库的内容和丰富程度上，都在行业平均水平之上，而且 App 的运行也十分流畅，但是缺少一些便利的产品功能。所以，我们提出的大致解决方案就是，要通过增加产品功能来提升用户留存。

最后，从大致的解决方案中提取出具体的假设。

那针对这款音乐 App，可以增加什么具体的产品功能呢？你可能会想到，**在每个专辑 / 歌单播放完成后增加“自动播放下一个专辑 / 歌单”的功能，以此来提升用户留存。**

这样一来，我们就通过三个步骤基本确定了目标和假设。

为什么说是“基本确定”了呢？因为确定目标和假设到这里还没有完全完成。要注意了，我们在上面确定目标和假设的时候其实还忽略了一个隐形的坑：这个假设中的“提升用户留存”还不能算是一个好的目标。因为这个假设还不够具体，目标没有被量化，而没有量化就没有办法提升。所以在这里，我们还需要做的就是量化“用户留存率”这个概念。

在按月付费的音乐 App 这个案例中，用户只要每个月按时付费续订，就是留存。所以，我们可以把用户留存定义为下个月的续订率，这样我们就把假设变得更加具体，并且目标可被量化。

那我们优化后，这个 A/B 测试的假设就变成了：**在每个专辑 / 歌单播放完成后增加“自动播放下一个专辑 / 歌单”的功能，可以提升用户下个月的续订率。**

为了帮你理解怎样才能做出好的假设，我根据自己的经验，把到底啥是好的假设，啥是不好的假设归纳到了一张图中，你一看就明白了：

好的A/B测试的假设是什么？

	好的假设	不好的假设
来源	用户调研，数据挖掘，观察经验等	不基于事实或观察的猜测
因果	明确包含可能的原因和结果	可能的原因和结果不明确
可证伪性	可被证伪	模糊，很难被证伪
可测量性	定量的指标	定性的结果
例子	在每个专辑/歌单播放完成后增加“自动播放下一个专辑/歌单”的功能，可以提升用户下个月的续订率	我们的产品可以打入高端市场

以上就是确定目标和假设的核心内容，你只要记住以下两点就够了：

A/B 测试是因果推断，所以我们首先要确定原因和结果。

目标决定了结果（用户留存），而假设又决定了原因（增加自动播放的功能），所以目标和假设对于 A/B 测试来说，是缺一不可。

有了测试目标和假设，我们就可以进入 A/B 测试的第二步了：确定指标。具体该如何确定指标呢？在解答这个问题之前，我们还需要先熟悉下指标的分类。

A/B 测试的指标有哪几类？

一般来说， A/B 测试的指标分为评价指标（Evaluation Metrics）和护栏指标（Guardrail Metrics）这两类。

评价指标，一般指能驱动公司 / 组织实现核心价值的指标，又被称作驱动指标。评价指标通常是短期的、比较敏感、有很强的可操作性，例如点击率、转化率、人均使用时长等。

可以说，评价指标是能够直接评价 A/B 测试结果的指标，是我们要重点关注的。

那有了评价指标，就可以保证 A/B 测试的成功了吗？显然不是的。很多时候，我们可能考虑得不够全面，忽略了测试本身的合理性，不确定测试是否会对业务有负面效果，因此很可能得出错误的结论。

举个例子。如果为了优化一个网页的点击率，就给网页添加了非常酷炫的动画效果。结果点击率是提升了，网页加载时间却增加了，造成了不好的用户体验。长期来看，这就不利于业务的发展。

所以，我们还需要从产品长远发展的角度出发，找到**护栏指标**。概括地说，护栏指标属于 A/B 测试中基本的合理性检验（Sanity Check），就像飞机起飞前的安全检查一样。它的作用就是作为辅助，来保障 A/B 测试的质量：

衡量 A/B 测试是否符合业务上的长期目标，不会因为优化短期指标而打乱长期目标。

确保从统计上尽量减少出现各种偏差（Bias），得到尽可能值得信任的实验结果。

到这里我们小结一下。在确定指标这一步，其实就是要确定评价指标和护栏指标。而护栏指标作为辅助性的指标，需要在选好了评价指标后才能确定。

那么问题来了，什么样的指标才能作为评价指标呢？

什么样的指标可以作为评价指标？

既然 A/B 测试的本质是因果推断，那么我们选择的业务指标的变化（结果）必须要可以归因到实验中的变量（原因）。所以，**评价指标的第一个特征，就是可归因性。**

比如，我们要测试增加“自动播放”功能，是否可以提升 App 的续订率。那么，这里的评价指标续订率的变化，就必须可以归因于增加了“自动播放”功能。在测试中我们控制其他可能影响续订率的因素都相同的情况下，增加了“自动播放”功能的变化就成了续订率的唯一影响因素。

刚才我们提到了，好的假设要能够被量化，否则就没有办法进行实验组和对照组的比较。这也就是**评价指标要有的第二个特征：可测量性。**

比如，对于音乐 App 来说，像用户满意度这个指标就不是很好量化。但是像用户续订率这样的指标，就可以量化。所以，我们就可以把“用户满意度”转化成“用户续订率”这种可以量化的指标。

可测量性和可归因性这两个特征都比较容易判断，除此之外，评价指标还具有第三个特征：**敏感性和稳定性**。那怎么理解呢？我用一句话来解释下：如果实验中的变量变化了，评价指标要能敏感地做出相应的变化；但如果是其他因素变化了，评价指标要能保持相应的稳定性。

看一个例子吧。还是在音乐 App 中，如果我想测试某一个具体内容的推送效果，比如推送周杰伦的新专辑，那么续订率会是一个好的指标吗？答案是否定的。

因为具体的推送是一次性的，而且推送只会产生短期效果（比如增加用户对杰伦新专辑的收听率），但不太会产生长期效果（比如增加续订率）。所以，续订率这个指标就对杰伦的推送不是很敏感。相反，短期的收听率是对单次推送更加敏感且合适的指标。

从这个例子中，我们可以得出两个结论：

用 A/B 测试来检测单次的变化时（比如单次推送 / 邮件）一般选用短期效果的指标，因为长期效果目标通常对单次变化并不敏感。

用 A/B 测试来检测连续的、永久的变化时（比如增加产品功能），可以选用长期效果的指标。

可见，如果选取的评价指标对 A/B 测试中的变化不敏感，或者对其他变化太敏感，我们的实验都会失败。那么，具体该如何测量评价指标的敏感性和稳定性呢？业界通常采用 A/A 测试来测量稳定性，用回溯性分析来表征敏感性。我来给你具体解释一下。

和 A/B 测试类似，A/A 测试（A/A Test）也是把被测试对象分成实验组和对照组。但不同的是，A/A 测试中两组对象拥有的是完全相同的体验，如果 A/A 测试的结果发现两组的指标有显著不同，那么就说明要么分组分得不均匀，每组的数据分布差异较大；要么选取的指标波动范围太大，稳定性差。

如果没有之前实验的数据，或者是因为某些原因（比如时间不够）没有办法跑新的实验，那我们也可以通过分析历史数据，进行**回溯性分析（Retrospective Analysis）**。也就是在分析之前不同的产品变化时，去看我们感兴趣的指标是否有相应的变化。

比如，我们选取续订率作为衡量增加“自动播放”功能是否有用的指标，那么我们就要去分析，在过去增加其他有利于用户留存的产品功能前后，续订率是不是有明显的变化。

好了，知道了应该选择什么样的指标作为评价指标之后，我们就可以开始选取适合我们自己业务的指标了。

如何选取具体的评价指标?

正像我们今天所看到的，确定评价指标的方法林林总总，但到底哪些是好用的，是真正可落地的呢？经过这些年的实践，我逐步总结积累了 3 种经验证确实简单、可落地的方法。

我还是以音乐 App 为例，和你解释下。

第一，要清楚业务或产品所处的阶段，根据这个阶段的目标，来确定评价指标。

这是因为，不同的业务 / 产品，甚至是同一个业务 / 产品的不同阶段，目标不同评价指标也会差别较大。

拿音乐 App 来说，在起步阶段，我们一般把增加新用户作为主要目标，把在拉新过程中的各种点击率、转化率作为评价指标；在发展和成熟期，一般会重点关注现有用户的使用和留存情况，把用户的平均使用时间和频率、产品特定功能的使用率，以及用户的留存率等作为评价指标。

比如要提高留存，首先要明确什么是留存：用户只要每个月按时付费续订，就是留存。那么这个时候，我们可以把用户留存的评价指标定义为下个月的续订率。

第二，如果目标比较抽象，我们就需要采用定性 + 定量相结合的方法了。

对于一些比较抽象的目标，比如用户的满意度，我们可以使用一些定性的方法，确定一些假设和想法，像**问卷调查、用户调研**等。同时，我们还可以利用用户使用产品时的各种数据，进行定量的数据分析，来了解他们的使用行为。

最后，我们把定性的用户调研结果和定量的用户使用行为分析结合起来，找出哪些使用行为和用户的满意度有着强烈的关系。

对于音乐 App 来说，我们具体可以这么做：

首先，通过定性的用户调研，来确定哪些用户满意、哪些用户不满意，完成分组。

接着，我们对每组用户（满意的用户和不满意的用户）分别做定量的用户使用习惯的数据分析，发现把音乐收藏到自己曲库的用户有较高的满意度，说明收藏音乐这个行为和用户满意度有强烈的正相关性。这时候，我们就可以把收藏音乐作为评价指标（比如收藏音乐的数量）。更进一步，我们还可以通过数据分析确定“收藏 X 首以上音乐的用户非常满意”中 X 的最优值是多少。

第三，如果有条件的话，你还可以通过公开或者非公开的渠道，参考其他公司相似的实验或者研究，根据自己的情况去借鉴他们使用的评价指标。

公开的渠道，是指网络上公开的各个公司关于 A/B 测试的文章或者论文。我经常会看的大公司的博客是Facebook、Google、Twitter，也推荐给你，你可以重点看 Facebook 中 Measurement 相关的文章，都是介绍评价广告效果的指标。

另外，你还可以去看一下《精益数据分析》这本书。在这本书里，你几乎可以找到所有重要互联网商业模式（电商，社交网络，移动 App 等）在各个阶段的典型指标。

为什么其他公司的评价指标有借鉴意义呢？原因很简单，To C 的产品用到 A/B 测试的场景都很相似。比如，我们想要通过 A/B 测试提升音乐 App 中广告的效果，那么 Facebook 在广告业务上的经验就能给我们很大的启发。

相应地，非公开的渠道，是指你的从事 A/B 测试并愿意和你分享经验的朋友，以及 A/B 测试相关的行业峰会。

在实践中，大部分的指标是根据产品 / 业务发展阶段的目标来确定的；如果实验的目标比较抽象或者比较新，通过经验和数据分析无法产生，你就可以采用定性 + 定量的方法了。

小结

今天这一讲，我们解决了下面两个问题。

第一，确定目标和假设，其实就是三大步：分析问题，确定结果；找出大致的解决方案；确定假设。

第二，确定指标，就是要确定评价指标和护栏指标。这节课主要讲了评价指标，其中关键的是我们要从目标入手，把目标量化。

最后，我要再和你强调一下，在 A/B 测试中确定目标和假设的重要性。A/B 测试是和业务紧密相关的，但我们往往会忽视业务中的目标，把注意力过多地放在选取评价指标上。在我看来，这就是本末倒置，就像一个不知道终点在哪里却一直在奔跑的运动员，如果能先明确终点，朝着终点的方向努力，会更快地取得成功。所以，你一定要按照今天学的内容，在做 A/B 测试时先试着找出你的目标和假设。

实际的业务场景大多比较复杂，很多时候单一的评价指标不足以帮助我们达成目标，而且指标也有波动性。所以，下节课，我会给你讲一讲综合多个指标建立总体评价标准的方法，以及指标的波动性。同时，我还会具体给你介绍护栏指标，保证你的 A/B 测试在业务和统计上的品质和质量。

思考题

根据生活和工作中的经历，结合今天所学内容，说说你认为有哪些指标是不适合做 A/B 测试的评价指标的？为什么呢？

欢迎在留言区写下你的思考和想法，我们可以一起交流讨论。如果你觉得有所收获，欢迎你把课程分享给你的同事或朋友，一起共同进步！

提建议

上一篇 02 | 统计基础（下）：深入理解A/B测试中的假设检验

下一篇 04 | 确定指标：指标这么多，到底如何来选择？

精选留言 (1)

写留言

皓昊 置顶

2020-12-19

老师，您给的三个博客，进去了不知道怎么看，能大概讲件吗？

作者回复: 你好，对于Facebook的你重点看下Measurement标签下的文章，里面有关于广告的测试方法和指标,有的文章好友whitepaper也可以参考；对于Google重点看Data&Measurement https://www.thinkwithgoogle.com/marketing-strategies/data-and-measurement/ 里面的内容.