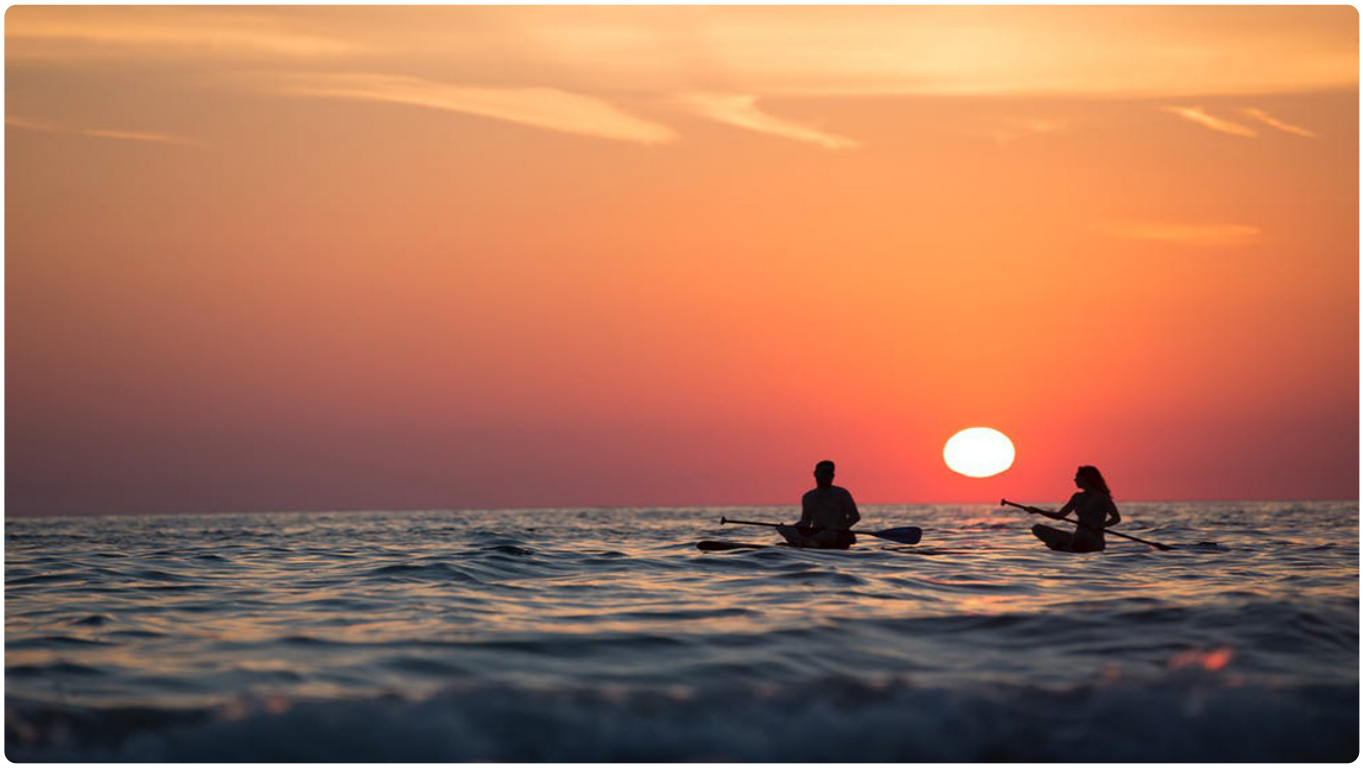
07 | 分析测试结果：你得到的测试结果真的靠谱吗？

2020-12-22 张博伟

A/B测试从0到1

进入课程 >



讲述：张博伟
时长 20:22 大小 18.66M

你好，我是博伟。

经过前面的确定目标和假设、确定指标、选取实验单位、计算所需样本大小后，我们终于来到了 A/B 测试的最后一站：分析测试结果。

在正式开始之前，我想先问你一个问题：拿到测试结果之后，就可以马上进行分析了吗？肯定不行。因为只有确定测试结果值得信赖之后，才可以进行分析。其实，分析 A/B 测试结果并不难，难的是如何得出值得信赖的结果，从而给业务以正确的指导。

为什么这么说呢？接下来，我就通过一个音乐 App 要提高用户升级率的例子，和你先拆解下导致测试结果不可靠的因素有哪些，然后再看看具体该怎么分析。

案例导入

通常情况下，音乐 App 有两种盈利模式，一种是提供免费音乐，但是会在 App 中加广告，通过广告赚钱；一种是让用户付费订阅 App，享受高品质的免广告音乐。

我们的这款音乐 App 是两种盈利模式都有，但是从长期盈利效果和用户体验来看，采用用户付费订阅的模式会更胜一筹。因此，我们计划在双十一前后，针对 App 里的免费用户做一次促销，吸引他们付费。

现在有这么两条广告语，为了通过 A/B 测试验证哪条更有效，将其分别放到实验组和对照组：

对照组广告语：千万曲库免广告无限畅听，用户升级，免费试用半年！

实验组广告语：即日起到 11 月 15 日，用户升级，免费试用半年！

现在，我们来完成 A/B 测试的整体设计方案。

确定目标：使更多的免费用户升级成为付费用户。

提出假设：通过在广告语中加入倒计时这种增加紧迫感的信息，能够提升免费用户的升级率。

确定实验单位：免费用户的用户 ID。

实验组 / 对照组：随机分配，50%/50%。

评价指标：用户升级率 = 点击广告升级的用户数 / 看到广告的用户数。

评价指标的波动范围：[1.86%，2.14%]。

设计好了 A/B 测试的框架，实施了 A/B 测试后，我们就可以等待分析测试结果了。那什么时候可以查看测试结果，停止 A/B 测试呢？这是保证测试结果可信赖要解决的第一个问题。

什么时候可以查看测试结果？

还记得我们上节课，在计算测试要达到显著性结果所需的最小样本量时，用到的一个公式吗？

A/B 测试所需的时间 = 总样本量 / 每天可以得到的样本量。

结合这个公式，再根据 App 中每天免费用户的流量，我们可以计算出这个测试在理论上需要跑 10 天。

其实，这个公式只是理论上推导，**具体到 A/B 测试的实践中，我们要确定测试时间，除了考虑样本量的大小外，还要考虑指标周期性变化的因素。**

如果指标具有强烈的周期性变化，比如周中和周末的变化很大，那么这时候的测试时间要包含至少一个周期的时间，以排除指标周期性变化的影响。

在音乐 App 这个案例中，我们通过历史数据发现，在周末升级的用户要比周中多。这就说明用户升级率这个评价指标，会以每周为单位形成周期性的变化，所以我们的测试至少要跑 7 天。而我们通过最小样本量已经算出了本次测试需要跑 10 天，包含了一个周期，所以我们可以放心地把测试时间定为 10 天。

我再多补充一句，如果计算出的测试时间小于一个周期的时间，那么最好也按照一个周期来算，这样做更为保险。

不过啊，在测试实际进行的过程中，有可能出现这样一种情况：在预计时间之前，评价指标出现了显著不同。这时候你就要小心了，如果提前结束测试，就会前功尽弃。我来给你具体解释下。

假设负责这个测试的数据分析师是第一次做 A/B 测试，所以特别激动兴奋，每天都在观测实验，计算测试结果。在实验进行到第 6 天的时候（样本量还没有达到预期），他发现实验组和对照组的评价指标出现了显著的不同。这位数据分析师就在想，**测试结果在预计时间之前达到了统计显著，这个实验是不是提前成功了呢?**

答案当然是否定的。

一方面，因为样本量是不断变化的，所以每次观测到的测试其实都可以算作新的实验。根据统计上的惯例，A/B 测试一般有 5% 的第一类错误率α，也就是说每重复测试 100 次，平均就会得到 5 次错误的统计显著性的结果。

这就意味着如果我们观测的次数变多的话，那么观测到错误的统计显著结果的概率就会大大提升，这是多重检验问题（Multiple Testing Issue）的一种体现。关于多重检验问题，我会在第 9 节课中详细讲解。

另一方面，提前观测到统计显著的结果，这就意味着样本量并没有达到事先估算的最小样本量，那么这个所谓的“统计显著的结果”就极有可能是错误的假阳性（False Positive）。“假阳性”是指，两组事实上是相同的，而测试结果错误地认为两组显著不同。

因此这位数据分析师还不能提前结束这次测试，仍然需要继续观测实验。

但**如果测试已经跑到了第 10 天，样本量也达到了之前计算的量，那是不是就可以开始分析 A/B 测试的结果了呢?**

答案依旧是不行。

俗话说心急吃不了热豆腐，为了确保实验在具体实施过程中按照我们预先设计的进行，保证中途不出现 Bug，那么在正式分析实验结果前，我们还要进行测试的合理性检验（Sanity Check），从而保证实验结果的准确性。

在第 3 和第 4 节课我们学过，为了确保在具体实施过程中不会出现破坏统计合理性的 Bug，我们可以用护栏指标来保证统计品质。这时，我们可以使用实验 / 对照组样本大小的比例和实验 / 对照组中特征的分布这两个护栏指标。这是保证测试结果可信赖，我们要关注的第二个问题。

保障统计品质的合理性检验

检验实验 / 对照组样本量的比例

我们预设的是，实验组和对照组的样本量各占总样本量的 50%，现在我们来看看实验过程中有没有发生什么变化。

各组样本量占总样本量的比例也是概率，也是符合二项分布的，所以具体的操作方法（参见第 4 节课指标波动性的相关内容）是：

首先根据二项分布的公式 $\sqrt{\frac{p(1-p)}{n}}$ 算出标准误差。

然后，以 0.5（50%）为中心构建 95% 的置信区间。

最后，确认实际的两组样本量的比例是否在置信区间内。

如果总的比例在置信区间内的话，就说明即使总的比例不完全等于 50%/50%，也是非常接近，属于正常波动，两组样本量大小就符合预期。否则，就说明实验有问题。那该如何确认和解决潜在问题呢？

回到我们的 A/B 测试上来，我们实验组的样本量 315256，对照组的样本量为 315174。通过公式我们求得标准误差为：

$$\sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0.5*(1-0.5)}{315256+315174}}$$

计算出来的结果是 0.06%，我们构建了 95% 的置信区间[50%-1.96*0.06%, 50%+1.96*0.06%] = [49.88%,50.12%]，也就是两组占比的波动范围，然后算出总体的实验组 / 对照组的样本量比例 =50.01%/49.99%。

可以看到，两组占比均在置信区间内，属于正常波动。也就是说，两组样本量符合均分的预期，成功通过了**实验 / 对照组样本量的比例**这个合理性检验。那我们接下来就可以进行实验 / 对照组中特征的分布这个合理性检验了。

检验实验 / 对照组中特征的分布

A/B 测试中实验组和对照组的数据要相似才具有可比性。这里的相似，我们可以通过比较两组的特征分布来判断。

常用的特征包括用户的年龄、性别、地点等基本信息，或者可能影响评价指标的特征。比如在音乐 App 这个案例中，我们还可以查看用户平时的活跃程度。如果这些特征在两组中分布比例相差较大，则说明实验有问题。

一旦从合理性检验中发现了问题，就不要急着分析实验结果了，实验结果大概率是不准确的。我们要做的就是找到出现问题的原因，解决问题，并重新实施改进后的 A/B 测试。

找原因的方法主要有以下两种：

和工程师一起从实施的流程方面进行检查，看看是不是具体实施层面上两组有偏差或者 bug。

从不同的维度来分析现有的数据，看看是不是某一个特定维度存在偏差。常用的维度有时间（天）、操作系统、设备类型等。比如从操作系统维度，去看两组中 iOS 和 Android 的用户的比例是否存在偏差，如果是的话那说明原因和操作系统有关。

通过数据分析发现这两组数据中重要特征的分布基本一致，说明两组数据是相似的。这就意味着我们已经通过了合理性检验，接下来我们就可以分析 A/B 测试的结果了。

最后，我还想跟你强调一下，这两个合理性检验是都要进行的，这是保障实验质量的关键。这两种检验如果没有通过的话都会使实验结果不准确，具体来说，实验 / 对照组样本量的比例和实验设计不相同时会出现样本比例不匹配问题（Sample Ratio Mismatch），实验 / 对照组的特征分布不相似则会导致辛普森悖论问题（Simpson Paradox），这两类问题我们会在第 11 节课中重点讲解。

如何分析 A/B 测试的结果？

其实，分析 A/B 测试的结果，主要就是对比实验组和对照的评价指标是否有显著不同。那怎么理解“显著”呢？其实，“显著”就是要排除偶然随机性的因素，通过统计的方法来证明两者的不同是事实存在的，而不是由于波动性造成的巧合。

那具体怎么做呢？

首先我们可以用统计中的假设检验（Hypothesis Testing）计算出相关的统计量，然后再来分析测试的结果。最常用的统计量是用 P 值（P value）和置信区间 (Confidence Interval) 这两种统计量。

你可能会说，假设检验中有各种各样的检验（Test），我应该选取什么检验来计算 P 值和置信区间呢？这里我们不需要理解这些检验的复杂理论解释，只要熟悉实践中常用的 3 种

检验方法的使用场景就可以了：

1. Z 检验（Z Test）

当评价指标为概率类指标时（比如转化率，注册率等等），一般选用 Z 检验（在 A/B 测试中有时又被称为比例检验（Proportion Test））来计算出相应的 P 值和置信区间。

2. T 检验（T Test）

当评价指标为均值类指标时（比如人均使用时间，人均使用频率等等），且在大样本量下可以近似成正态分布时，一般选用 T 检验来计算相应的 P 值和置信区间。

3. Bootstrapping

当评价指标的分布比较复杂，在大样本量下也不能近似成正态分布时（比如 70% 用户的使用时间，OEC 等），一般采用 Bootstrapping 的方法，从 P 值或者置信区间的定义来计算 P 值和置信区间（具体方法请参见第三节课指标波动性的相关内容）。

现在我们已经拿到了如下的测试结果：

实验组：样本量为 315256，升级的用户为 7566，升级率为 2.4%。

对照组：样本量为 315174，升级的用户为 6303，升级率为 2.0%。

因为评价指标的波动范围是[1.86%,2.14%]，所以我们可以得出实验组的升级率 2.4% 并不属于正常范围，很有可能显著不同于对照组。

接下来，我们就可以通过 P 值法和置信区间法来分析这个测试结果，验证我们的假设是否正确。

P 值法

首先我们可以采取 P 值法，借助一些计算工具，常见有 Python、R，还有网上的一些在线工具（比如这个网站），都可以计算 P 值。具体选择哪个工具，根据自己的喜好来就可以。我个人比较喜欢选用 R 来计算：

```
1 results <- prop.test(x = c(7566, 6303), n = c(315256, 315174))
```

复制代码

因为用户升级率这个评价指标属于概率类指标，所以我们选择了专门针对概率类指标的函数prop.test。

通过计算，我们可以得到 P 值 < $2.2e^{-16}$：

```
data:  c(7566, 6303) out of c(315256, 315174)
X-squared = 117.08, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.003273811 0.004728321
sample estimates:
    prop 1     prop 2
0.02399954 0.01999848
```

根据统计惯例，一般我们会把测试的显著水平（Significance Level）α定为 5%（统计上的约定俗成），再把计算出来的 P 值和 5% 相比。当 P 值小于 5% 时，说明两组指标具有显著的不同。当 P 值大于 5% 时，说明两组指标没有显著的不同。如果你对这块概念还不是很清楚，可以回顾下第二节课中假设检验的内容。

从上面的结果可以看出，P 值远远小于 5% 且接近于 0，说明两组指标具有显著的不同，这就意味着实验组的广告语确实能提升免费用户的升级率。

置信区间法

在第三节课介绍指标时，我们学习了该怎样构建置信区间。现在我们要比较实验组和对照组的评价指标是否显著不同，也就是看两者的差值是不是为 0。这时候，我们就要构建两组指标差值 $(p_{\text{test}} - p_{\text{control}})$ 的置信区间了。

置信区间的具体计算我们也可以借助 Python 和 R 等软件，当然你也可以使用我在第二讲时介绍过的具体函数，这里我们还是用 R 的prop.test这个函数。

其实当我们在上面用这个函数计算 P 值时，R 也顺便把 95% 的置信区间算出来了：

```
data:  c(7566, 6303) out of c(315256, 315174)
X-squared = 117.08, df = 1, p-value < 2.2e-16
alternative hypothesis: two.sided
95 percent confidence interval:
 0.003273811 0.004728321
sample estimates:
    prop 1     prop 2
0.02399954 0.01999848
```

由图可见，95% 的置信区间为[0.0033, 0.0047]。

接下来，我们需要比较一下两组指标是否有统计显著的不同，也就是要看看这个置信区间是否包括 0。

我们知道数值在置信区间内均为正常波动，如果置信区间包括 0 的话，就说明两组指标的差值也有可能为 0，两组指标是相等的。而如果置信区间不包括 0 的话，就说明两组指标的差值不为 0，两组指标是显著不同的。

显然，[0.0033, 0.0047]这个置信区间是不包括 0 的，也就是说我们的测试结果是统计显著的。那对应到业务上，与对照组的广告语（千万曲库免广告无限畅听，用户升级，免费试用半年！）相比，带有紧迫感的实验组广告语（实验组广告语即日起到 11 月 15 日，用户升级，免费试用半年！）能吸引更多用户升级，也就验证了我们最开始的假设是成立的。

学到这里，我们发现**无论是 P 值法还是置信区间法，都可以用来分析 A/B 测试结果是否具有统计显著性。那么，在实际应用中该如何选择呢？两者有什么差别吗？**

其实，在大部分情况下这两种方法是通用的，只要选择一种就可以。但如果需要考虑实施变化后的收益和成本的关系时，我们就要选择置信区间法了。

因为要考虑收益和成本的关系时，除了满足结果在统计上是显著的（两组指标不相同，差值的置信区间不包括 0）还不够，更要让结果在业务上也是显著的（两组指标不仅要不相等，而且其差值δ >= $\delta_{收支平衡}$，并且差值的置信区间的范围都要比 $\delta_{收支平衡}$ 大）。

小结

这节课我们主要讲解了 A/B 测试中如何分析结果，根据实践经验我给你总结了 3 个要点：

切莫心急，一定要等到达到足够样本量时再分析测试结果。

分析结果前一定要做合理性检验来确保测试的质量，否则一旦实施过程中出现 Bug，就会功亏一篑。

一定要根据指标和数据的特点，选择正确的分析方法来得出可以驱动业务的结论。

数据领域有一句名言：“Garbage in, garbage out”，意思就是“放进去的是垃圾，产出的还是垃圾”。这句话放在 A/B 测试中同样适用：如果 A/B 测试没有设置好，或者虽然计划得很好，但要是在实施过程中出现了问题，也会得到错误的结果和结论，从而给业务带来难以估量的损失。

所以，前面我们用 4 节课来讲怎么设置实验，今天又花了很多篇幅来介绍确保结果是可信赖的，都是在给“分析测试结果”做铺垫。

好了，今天这个音乐 App 的测试得到了显著的结果，皆大欢喜。但是如果结果不显著，又该怎么办呢？

关于这个问题，我们在第 9 节课再来好好讨论！

思考题

你觉得分析结果前的合理性检验还可以参考哪些护栏指标呢？为什么？

欢迎在留言区写下你的思考和答案，我们一起交流讨论。如果你觉得有所收获，欢迎你把这一讲分享给你的朋友，邀请他一起学习。

提建议

上一篇 06 | 选择实验样本量：样本量越多越好吗？

精选留言 (4)

写留言

西西
2020-12-22

合理性检验如果没有通过怎么办呢？比如样本量差异过大，应该重新做实验吗？或者在较大的样本量中随机选出和较小样本量差不多的样本进行比较可行吗？

展开

作者回复: 条件允许的话推荐重新做实验，因为样本差异过大可能是因为偶然的随机不够造成的，如果重新做实验还存在类似的情况的话如果实验流程没有错误的话，那么可能就是实施层面除了问题，需要和工程师来一起找bug了。当然如果时间紧迫没时间去跑新的实验在较大的样本量中随机选出和较小样本量差不多的样本进行比较，比如用倾向评分匹配(Propensity Score Matching)，这个我在第9节课和第12节课中都会讲解。

1

吴优秀同学
2020-12-22

对专栏唯一的抱怨就是更新太慢。在这里请教老师一个问题，在计算样本量的时候我们预估会有10%的提升，然后以此得出样本量进行abtest。最后发现a组和b组之间的比率之差没有达到10%只有3%，但是统计结果是显著的。那是不是意味着我们还要继续进行检验，增加样本量，让我们能够对3%的差异进行检测呢？还是说差异没达到10%，不用进行统计学检测，直接否定这个改变能带来10%的差异这个假设。

展开

编辑回复: 哈哈也是吴优秀同学学习吸收的速度太快了！

Geek_Libratus
2020-12-22

在 保障统计品质的合理性检验 这一节中老师提到， “各组样本量占总样本量的比例也是概率，也是符合二项分布的” ，这句话很不理解，为什么这个比例会服从二项分布呢？这里

面的总体、样本、样本点分别是什么呢？我理解二项分布是N次独立重复实现发生K次的概率，但是这个想法在这个例子中好像很难套进去。

展开

那时刻

2020-12-22

我们目前使用的护栏指标是DAU和留存，细想了下，感觉可以归入到老师提到的delta收支平衡。不知理解是否正确？

另外请教老师两个问题：
1. 因为要考虑收益和成本的关系时，让结果在业务上也是显著的（两组指标不仅要不相...

展开