**五、AB Test**

在实际业务场景中，为了优化用户体验，产品会有许多功能、版本、页面的迭代更新。但如何评判更新是否有效，ABTest是利用数据验证更新是否带来正向收益的核心方式。

AB Test的核心思路是利用统计学中假设检验的思想，对用户进行分流，部分用户使用1号功能，部分用户使用2号功能，随后对1号和2号功能带来的观测指标进行对比，来评判1号功能与2号功能的好坏。

举例：某电商公司产品经理小A对电商的支付页面进行更新，希望提高用户支付步骤的转化率。数据分析师小B可以采用AB Test，部分用户仍然沿用原页面，部分用户使用新页面，随后对比两部分用户的转化率情况，来评判新页面是否对用户转化率有提升。

AB Test是应届生数据分析面试中常考题目，属于必须掌握项目。

**核心知识点**

**1、AB Test的整体流程**

* 确认AB Test的测试需求，判断更新的功能是否适合使用AB Test测试。
* 讨论观测指标，即使用新功能主要用来提升的指标。如更新支付页面，提高用户支付转化率。更新某搜索算法，提高用户点击率。
* 对用户进行分流，确认投放人群和投放量，将用户划分为对照组和实验组，保证分流后用户特征相同。
* 配置实验策略，同一公司有不同的AB Test策略同时进行，部分用户流量存在于其他AB Test实验中，需要判断新的AB Test测试功能是否对原AB Test测试产生影响，从中划分流量。
* 评估AB Test效果。对主要的观测指标，使用T检验/Z检验等方式，判断是否存在显著差异，即新功能是否有效。

**2、AB Test的适用场景**

AB Test可以应用的功能非常广泛 ，包括算法的更新、页面的更新、策略的更新等等。在采用AB Test之前要明确更新的功能是否适用于AB Test的场景。

一般而言，以下几种场景不适合用AB Test:

* 变量不可控制。AB Test的核心需要保证用户除了更新的功能外其他因素全部相同，部分场景中无法控制。如，想做一个策略验证是否能够提高用户使用了A产品，再去使用B产品的概率。这种是不支持AB测试的，因为用户关闭一个APP后，非常多的不可控因素。
* 样本量较小：AB Test从统计学假设检验角度来说需要一定的样本数量，若实际业务场景中这个功能运用的数据量较小，则无法使用。
* 全量投放：AB Test必须对用户进行分流，分别使用不同功能进行对比。如若某一功能对全体用户投放，则无法使用AB Test。
* 用户感知性强的功能。部分功能在线上对用户进行分流，如若感知性较强，可能对分流用户造成负面影响。

1. **AB Test投放人群分组**

在AB Test中，投放人群=流量，选择投放人群有两种形式：

* 随机:随机从大盘用户里抽样。例如userid单双尾号抽取、随机种子打乱数据再抽样等。例如对用户userid hash后取模，然后将流量分成多份。
* 定向: 针对特定用户群体下发策略。例如结合用户画像标签或自定义规则筛选用户。

**4、AB Test样本量计算**

AB Test实验需要多少样本才能保证实验的准确性，是许多公司的笔试常考知识。一般而言样本量越大，越能够保证实验的准确性，但是实际业务中，使用过大的样本量会造成流量浪费，且部分功能可能对用户感知效果存在影响，因此需要计算最小的AB Test所需样本量。

1. **样本量计算参数**

假设检验中定义原假设H0：实验组和对照组无显著差异。备择假设H1：实验组和对照组存在显著差异。根据实验结果是否拒绝原假设H0，会出现四种情况。



* 第一种情况（**统计学中的第一类弃真错误**）：当更新功能没有提升，而实验结果有提升，就是一类错误、这种情况会导致最终的实验结果出问题：功能本身没有提升但公司认为有提升，从而使用新功能浪费公司资源。**所以，第一类错误非常严重，要尽量避免，我们需要给这种错误出现的频率设置一个最大值，也就是Significance Level （α）。一般情况下永远不能超过5%**
* 第二种情况：当真实数据没有差异，预测也没有差异，证明更新功能无效。
* 第三种情况：实际值有差异，实验结果也有差异，**实际上这种情况就是我们需要的**，这个概率就是判断正确的概率，叫做Statistical Power。如果power过低，也就是说本来有效果的，而实验却得出来没有效果，就会导致实验白做，所以power越高越好。
* 第四种情况：本来有差异，实验却展示出来没有差异。这种用β来表示，当确定了事实是有差异的情况下：power = 1 - β

**这两个指标：**

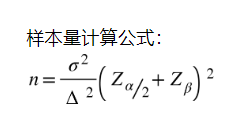
Significance level（α）：越小越好，一般不大于5%；

Statistical power（1-β）：越大越好，一般不小于80%，也就是说β需要小于20%；

从两个错误的容忍度来看，β是要比α大很多的，也就是说，宁可错杀一个好产品，也不能放过一个坏产品。

1. **非比例类指标所需样本量**

非比例类指标指pv、uv等、非分数形式的比率指标。



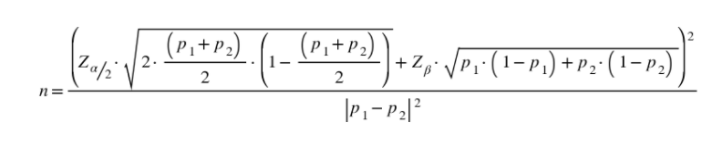
其中：

* n是每组所需样本量，因为A/B测试一般至少2组，所以实验所需样本量为2n；
* α和β分别称为第一类错误概率和第二类错误概率，一般分别取0.05和0.2
* Z为正态分布的分位数函数；
* Δ为两组数值的差异，如目前点击率1%，目标将点击率提高至1.5%，那么Δ就是0.5%；
* σ为标准差，是数值波动性的衡量，σ越大表示数值波动越厉害。

**（3）比例类数值所需样本量的计算**

实际业务中关注较多的一类是比例类的数值，如点击率、转化率、留存率等。这类比例类数值的特点是，对于某一个用户（样本中的每一个样本点）其结果只有两种，“成功”或“未成功”；对于整体来说，其数值为结果是“成功”的用户数所占比例。如转化率，对于某个用户只有成功转化或未成功转化。

比例类数值的假设检验在统计学中叫做两样本比例假设检验。其最小样本量计算的公式为：



其中：

* p1为基础值，是实验关注的关键指标现在的数值（对照组）；
* p2为目标值，是希望通过实验将其改善至的水平；
* α和β第一类错误和第二类错误上边已经提到，通常设定α为0.05，β为0.2 。

**5、AB Test分流机制**

公司实际业务中，需要进行的AB Test数量较多，但流量总量有限，因此对公司流量需要有科学的管理分流机制，来确保多个AB Test的同时运行，且互相不影响效果。

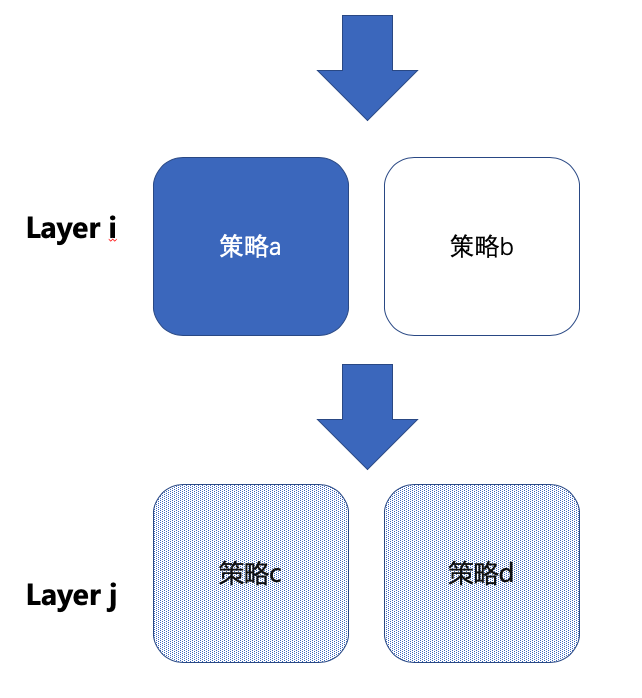
举例：现在公司正运行一个搜索算法的AB Test，其中50%的流量保持原算法，50%的流量使用新算法。而此时公司又要更新某一个界面，对界面效果使用AB Test，但流量已经全部在运行搜索算法AB Test了，那么如何分配流量做新的界面AB Test，并且两个AB Test之间不互相影响？业内使用最多的是**可重叠分层分桶。**

1. **流量正交与互斥**

**流量正交：**：每个独立实验为一层，层与层之间流量是正交的，一份流量穿越每层实验时，都会再次随机打散，且随机效果离散。**正交实验的每一层都是独立的，业务关联较低的AB Test实验。**

例如：Layer1正在使用搜索算法AB，其中50%流量用于原算法，50%用于新算法。而Layer2需要使用界面效果AB，则可以将Layer1中的流量打乱，各取50%流量使用原界面，50%流量使用新界面。此时Layer2中的50%原界面流量与50%新界面流量中，都各自包含50%使用原算法和50%新算法的流量，则可以保证实验的独立性，算法和界面AB未互相干扰。

**一个简单通俗的例子：**我们有 100 个兵乓球，随机拿出来 50 个染成蓝色，50 个染成白色，则我们有蓝色、白色兵乓球各 50 个，现在我们把这 100 个兵乓球重新放在袋子中摇匀，随机拿出 50 个兵乓球，那么这 50 个兵乓球颜色蓝色和白色各 25 。

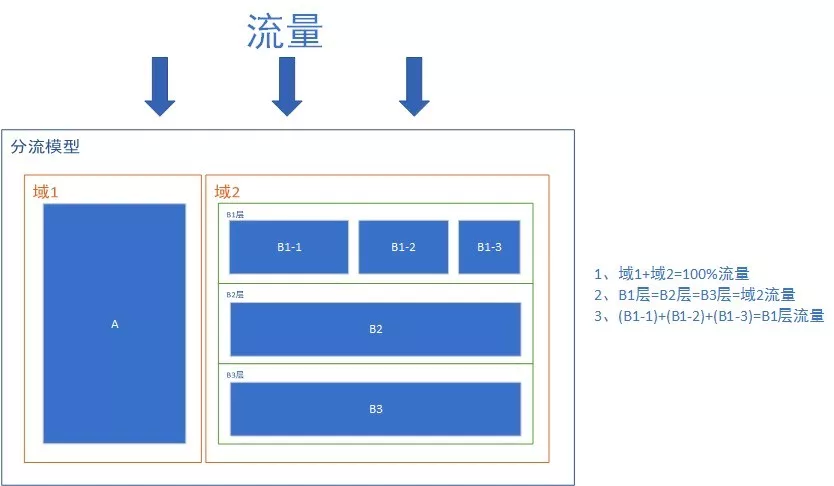


**流量互斥：**实验在同一层拆分流量，且不论如何拆分，不同组的流量是不重叠的。

例如：现要进行界面效果AB，第一个方案是将界面中一个按钮改成圆形，第二个方案是将界面中一个按钮改成红色。修改一个按钮的颜色和形状，两个方案业务关联度过高。必须使用互斥的流量方案。即流量在同一个Layer中，一部分流量用户修改颜色的AB ，一部分用于修改形状的AB，不同时进行。

一个简单通俗的例子：我们有 100 个兵乓球，每 25 个为一组，分别染成蓝、白、橘、绿。若 X 实验拿的是蓝色、白色则 Y 实验只能拿橘色和绿色，我们说 X 实验的和 Y 实验是互斥的。

1. **可重叠分层分桶**



* 域 1 和域 2 拆分流量，此时域 1 和域 2 是互斥的。
* 流量流过域 2 中的 B1 层、B2 层、B3 层时，B1 层、B2 层、B3 层的流量都是与域 2 的流量相等。此时 B1 层、B2 层、B3 层的流量是正交的。
* 扩展：流量流过域 2 中的 B1 层时，又把 B1 层分为了 B1-1 ，B1-2 ，B1-3 ，此 B1-1 ，B1-2 ，B1-3 之间又是互斥的。
* 根据以上规则我们可以不断的在此模型中增加域、层，并且可以互相嵌套。这要与实际的业务相匹配，拆分过多的结构可能会把简单的业务复杂化，拆分过少的结构又可能不满足实际业务。

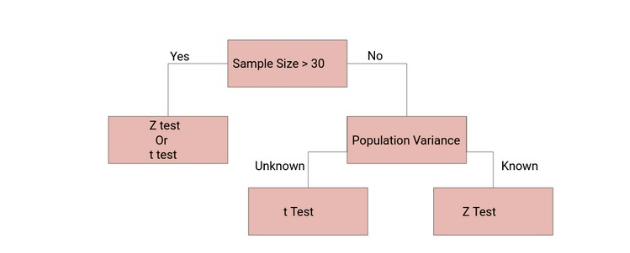
1. **使用场景**

* B1 层、B2 层、B3 层可能分别为：UI 层、搜索结果层、广告结果层，这几层基本上是没有任何的业务关联度的，即使共用相同的流量 ( 流量正交 ) 也不会对实际的业务造成结果。
* 如果不同层之间所进行的试验互相关联，如 B1 层是修改的一个页面的按钮文字颜色，B2 层是修改的按钮的颜色，当按钮文字颜色和按钮颜色一样时，该按钮已经是不可用的了。
* 建议同一类型的实验在同一层内进行，并且需要考虑到不同实验互相的依赖。

**6、AB Test效果评估**

**（1）AB Test检验方式选择**

AB Test的效果评估部分基本就是统计学中假设检验部分内容。公司一般采用t检验或Z检验。从统计学原理来说，当样本数量大于30或样本数量大于30且总体方差已知时，都采用Z检验，若样本数量小于30且总体方差未知则采用t检验。但实际业务中，短期的AB Test实验时间很短，都在1个月内，大多选择t检验。只有长期实验采用Z检验。



**（2）建立假设**

**AB Test一般采用的假设为：**

零假设(H0):即假定策略AB没显著差异。

备择假设(H1):即策略AB有显著差异。**即双边假设检验。**

**（3）设置决策标准**

假设检验的决策标准有两种:

* p值: p值是用来判定假设检验结果的参数，如果p值很小，说明原假设情况发生的概率很小，我们就有理由拒绝原假设。我们一般会认为显著性水平值为0.05，即如果p值<0.05，则有超过95%的信心拒绝零假设。
* 检验统计值: 例如t值。当我们定了显著性水平后，可以透过查表，得到对应的t临界值，若t真实值(Test Statistic) > t临界值(Critical Value)，我们会拒绝零假设。

**（4）假设检验得出结论**

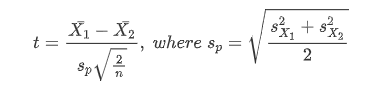
**举例：**

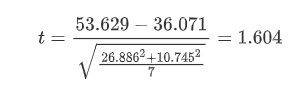
策略A 7天的转换率: [64.2, 28.4, 85.3, 83.1, 13.4, 56.8, 44.2]，平均值为53.629，标准差为26.886。

策略B 7天的转换率: [45, 29.5, 32.3, 49.3, 18.3, 34.2, 43.9]，平均值为36.071，标准差为10.745。

由于样本量为7个，所以选择T检验。而T检验在实际应用中有三种：单样本T检验: 总体样本vs抽样样本、配对样本T检验: 同一对象接受两种不同的处理、双独立样本T检验: 两组独立样本。

正常情况下，AB Test用的是两组独立样本，控制变量进行对比实验，所以我们这里采用双独立样本T检验：





**7、AA Test**

AA Test的操作步骤与AB Test一致，唯一的区别在于AB Test的AB两组采用的是不同的功能，而AA Test采用的是相同的功能，即AA Test的两组用户面临的版本、功能完全一致。使用AA Test的意义为：

* **验证AB Test结果的正确性。**AB Test得出结论很容易，但得出正确的AB Test解决相当困难，在AB Test过程中，流量分配不均匀、实验配置错误、统计计算和推理不科学都会导致出现错误结论。因此采用AA Test，如若两组用户使用的是相同的功能，理论上观测指标应当一致，如若检测出不一致的话，则可能出现上述情况，AB Test结果有误。
* **设定观测指标参考值。**AB Test的目标是通过新功能提高观测指标，但具体观测指标提高多少是真的提升，可能事先无法确定。因此采用AA Test，例如， 在主页上开始A / A测试，目标是填写联系表格。比较结果时，得到的结果几乎相同（这是正常的）：5.01％和5.05％的转换。现在，可以先确定使用该数据，它代表当前数据的转换率波动，然后激活A / B测试以尝试超过该转换率（超过AA Test的转换率）。如果A / B测试的结果是更新的功能只实现了5.05％的转化（上述AA的转化率结果），这实际上意味新功能并未带来提升。
* **寻找更好的样本。**通过AA Test，可以查看两组样本中，表现最相似的用户样本，这部分样本相似性更高，可以只挑选这部分样本用于AB Test，从而减少AB中因样本差异导致的误差。

**面经真题：**

1、ABTest实验过程的阐述？（Oppo）

2、AB Test 大致怎么做，显著性、统计功效 怎么描述（携程）

3、简单说下假设检验中p\_value的含义吧？（H0假设的参数前提下，出现样本估计参数发生的概率。）（领英）

4、abtest最小样本量怎么计算（字节跳动）

5、abtest的ab组一定是1:1吗（字节跳动）

6、AB Test有什么缺点？（滴滴）

7、怎么确定AB测试的样本（网易）

8、ABtest检验有效性（京东）

9、如何保证ab实验两组样本差异不大（触宝）

10、ab分析怎么做的？aa阶段实验组和对照组出现差异怎么处理的？(滴滴)

11、A/Btest的配置流程以及样本量如何计算？（莉莉丝）

**参考资料：**

1. **ABTest适用场景**

原文链接：<https://blog.csdn.net/u011487470/article/details/121739858>

1. **样本量计算：**

https://blog.csdn.net/weixin\_43287568/article/details/116357645

https://blog.csdn.net/weixin\_41744624/article/details/109840263

**3、AB Test分流机制**

<https://www.infoq.cn/article/BuP18dsaPyAg-hflDxPf>

**4、AB Test效果评估**

微信公众号：数据挖掘工程师。

2022年3月16日发布文章：一文看懂AB Test流程规范

**5、AA Test**

<https://blog.csdn.net/weixin_41744624/article/details/109990705>

<https://baijiahao.baidu.com/s?id=1691552349437631082&wfr=spider&for=pc>