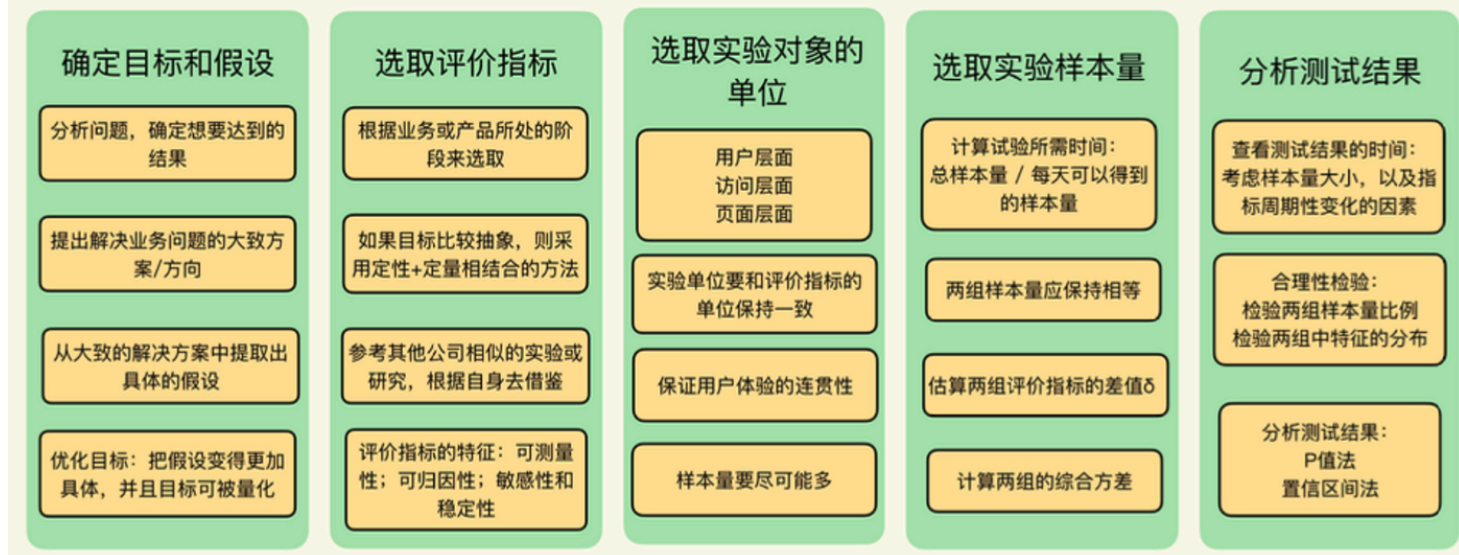


AB实验框架与流程

1. 实验框架

A/B测试的规范框架



1. 从业务问题出发，确定AB测试的目标和假设

2. 确定AB测试的评价指标

- 确定触发条件
- 定义实验中的用户
- 确定用户使用功能的时间窗口期
- 指标类型：
 - 核心指标：用以监控评估实验效果；
 - 辅助指标：辅助核心指标，用以评估实验效果
 - 不变指标：用以评估实验环境变化；
 - 负面指标：用于观测实验风险；

3. 选取实验对象的单位

- 用户ID，IP地址，Cookies ID等

4. 明确置信水平、统计功效、想要观测的最小变化量来计算所需的样本大小

- 确定统计量，及用来计算最小样本量大小：
 - 显著性水平
 - Power

- 实验组、对照组的综合方差
- 实验组、对照组的评价指标差值
 - 可以根据经验确定

5. 确定分流策略和实验所需的时间

- 确定好样本量之后便可以开始正式分组了。
- 需要注意多实验开启时，需要正交分流，防止实验之间的交互影响。

6. 分析测试结果

- 合理性检验：保证测试的质量、确保AB测试具体实施过程符合预期设计
 - 使用 **护栏指标** 检验：
 - 实验组、对照组样本大小比例
 - 实验组、对照组中的特征分布是否相似

7. 正式分析结果

- 计算P值、置信区间

2. AB实验场景

AB实验又称为受控实验（Controlled Experiment）或者对照实验。AB实验的概念来自生物医学的双盲测试，双盲测试中病人被随机分成两组，在不知情的情况下分别给予安慰剂和测试用药，经过一段时间的实验后，比较这两组病人的表现是否具有显著的差异，从而确定测试用药是否有效。

2.1 AB测试中不适用的场景

1. 当没有办法 **控制想要测试的变量** 时：AB测试是控制变量实验，而控制变量前提是我们能够人为控制，例如一些需要用户自己个人选择决定的变量
2. 当有 **重大事件发布** 时：例如新产品、新业务的发布，或者产品形象的变化
 - 例如产品代言人、公司的商标
3. 当 **用户数量很少** 时：当流量很少时，很难在短时间内达到所需要的样本量
4. 不适用对一些初期不成熟想法的验证

2.2 AB实验无法使用时的替代方法

2.2.1 用户研究

用户研究适用于 AB 测试无法进行时，比如新产品业务发布前的测评，我们就可以通过直接或间接的方式，和用户交流沟通来获取信息，从而判断相应的变化会对用户产生什么影响。

1. 深度用户体验研究 (Deep User Experience Research): 通过 选取几个潜在用户进行深度的信息提取，比如通过用户眼球的运动来追踪用户的选择过程的眼动研究，或者用户自己记录的日记研究

2. 焦点小组 (Focus Group): 有引导的小组讨论, 由主持人把潜在的用户组织起来, 引导大家讨论不同的话题, 然后根据大家在讨论中发表的不同意见, 综合得出反馈意见。
3. 调查问卷 (Survey): 通过事先设计好的问题, 选择题或开放性问题, 将问题做成问卷发给潜在用户

2.2.2 因果推断

当AB测试不适用时, 因果推断方法可以帮助评估因果关系或比较不同组之间的效果, 见后续。

3. 实验指标 - 选取具体的实验指标

选取评价指标的规则:


- 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。
- 评价指标需要满足一下特征:
 - 可归因性:
 - 可测量性: 需要可以被量化
 - 敏感性: 指标要能敏感地反映出实验中变量的变化
 - 稳定性: 其他因素变化了, 指标要能保持相对的稳定

具体步骤:

1. 清楚业务产品当前所处的阶段, 根据阶段的目标, 确定评价指标: 例如起步阶段、发展阶段、成熟阶段
2. 如果目标较抽象, 则采用定性+定量结合的方法: 问卷调查、用户调研等定性方法, 将定性的调研结果与定量的用户使用行为分析结合
3. 有条件, 则可以通过公开或非公开的渠道, 参考其他公司相似的实验或研究, 根据自身实际情况借鉴他们使用的评价指标

当需要综合考虑多个指标时, 我们需要综合考虑改动所带来的好处和潜在的损失, 结合多个指标, 构建一个总体评价标准 (Overall Evaluation Criteria, 简称 **OEC**)。

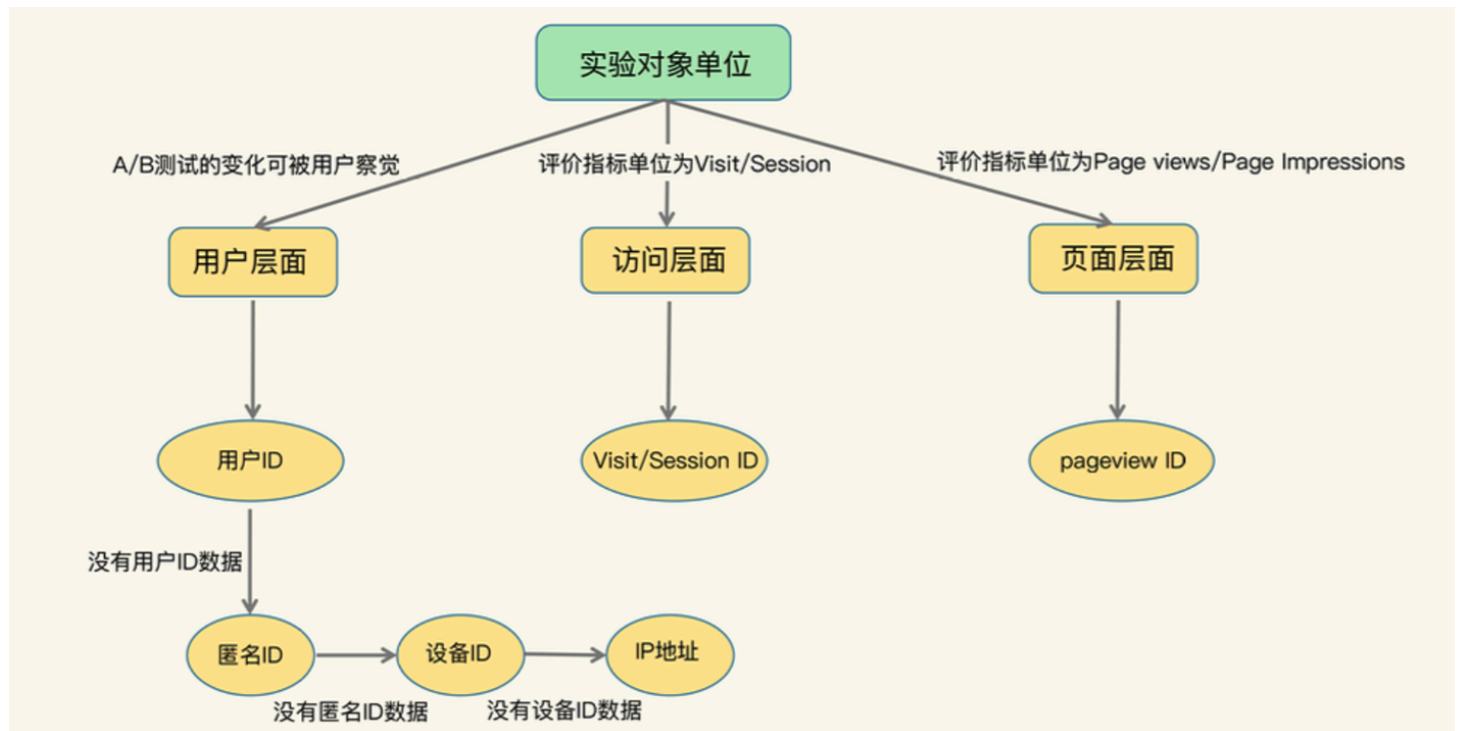
当要考察的事物包含多个方面时, 只有综合各方面的指标, 才能把握总体的好坏。这也是使用 OEC 最明显的一个好处。最常见的一类 OEC, 就是亚马逊的这种结合变化带来的潜在收益和损失的 OEC。需要注意的是, 这里的“损失”还有可能是护栏指标, 也就是说 OEC 有可能会包含护栏指标。

 使用 OEC 的另一个好处就是可以避免 **多重检验问题(Multiple Testing Problem)**。如果我们不把不同的指标加权结合起来分析, 而是单独比较它们, 就会出现多重检验的问题, 导致 A/B 测试的结果不准确。多重检验问题是 A/B 测试中一个非常常见的误区。

4. 实验对象 - 如何选取实验单位

理解误区：实验单位不就是用户吗？

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。



4.1 用户层面

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。

- **用户ID：** 用户注册、登录时的用户名、手机号、电子邮箱，等等

特点：稳定，不易改变

- **匿名ID：** 用户浏览网页时产生的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。
- 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。
- 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。

- **设备ID：**

- 和设备绑定的，一旦出厂就不可改变。
- 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。
- **IP地址：**
 - 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。
 - 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。

4.2 访问层面

把**用户的每次访问**作为一个最小单位。

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。
- 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。

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。

4.3 页面层面

页面层面指的是把每一个**新的页面浏览**为最小单位。

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。

4.4 三种层面的对比

1. 从变化是否易被察觉考虑：

- 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。不然可能会使得同一个用户，一下体验到好用的新功能，一下功能又没了，会让用户感到困惑、甚至沮丧，影响体验。

2. 实验单位的细粒度：

- 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。
- 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。

在改动容易被察觉的情况下，以用户层面的实验单位的实验可能短时间没法获取足够样本量，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

4.5 总结选取实验单位的原则

1. 保证用户体验的连贯性。

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. 实验单位应与评价指标的单位保持一致。

A/B 测试的一个前提是**实验单位相互独立且分布相同的**，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。

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。

5. 样本量的选择

样本量越大，样本所具有的代表性才越强：

因为当样本数量很少的时候，实验容易被新的样本点带偏，造成了实验结果不稳定，难以得出确信的结论。相反的，样本数量变多，实验说服力也更强。

但在实际业务中，样本量其实是越少越好。

A/B 需要做多长时间的一个公式: $A/B \text{ 测试所需的时间} = \text{总样本量} / \text{每天可以得到的样本量}$ 。

而在实际操作时，需要考虑以下因素：

1. 流量有限：公司流量有限，不合理分配流量，产品迭代速度会大大降低

从公式就能看出来，样本量越小，意味着实验所进行的时间越短。在实际业务场景中，时间往往是最宝贵的资源，毕竟，快速迭代贵在一个“快”字。

2. 试错成本大

如果使用50%的流量进行实验，一周后结果表明实验组的总收入下降了20%。算下来，实验在一周内给整个公司带来了10%的损失。试错成本太高。

实验范围越小，样本量越小，试错成本就会越低

实践和理论上对样本量的需求，其实是一对矛盾。所以，我们就要在统计理论和实际业务场景这两者中间做一个平衡：在 A/B 测试中，既要保证样本量足够大，又要将实验控制在尽可能短的时间内。

需要计算满足实验要求的最小样本量，最小样本量是根据统计功效进行计算的

主要分两类：绝对值类（例如：UV）和比率类（例如：点击率）

5.1 最小样本量计算公式-绝对值类

$$n = 2 \times \frac{\left(Z_{\frac{\alpha}{2}} + Z_{1-\beta}\right)^2}{\left(\frac{\Delta}{\sigma_{\text{pooled}}}\right)^2} = 2 \times \frac{\left(Z_{\frac{\alpha}{2}} + Z_{\text{power}}\right)^2}{\left(\frac{\Delta}{\sigma_{\text{pooled}}}\right)^2}$$

理解：

- $Z_{1-\frac{\alpha}{2}}$ 为 $(1 - \frac{\alpha}{2})$ 对应的 Z Score。
- Z_{Power} 为 Power 对应的 Z Score。
- Δ 为实验组和对照组评价指标的差值。
 - 如果两个版本的均值差别巨大，也不太需要多少样本，就能达到统计显著
 - 两组数值的差异，如点击率1%到1.5%，那么 Δ 就是0.5%（这个差异需要根据实施变化后所需成本和收益是否达到预期来估算）
- σ_{pooled}^2 为实验组和对照组的综合方差 (Pooled Variance)。
 - 组间的标准差越小，代表两组差异的趋势越稳定。越容易观测到显著的统计结果

在公式中，样本量主要由统计显著性 α , 统计功效 Power, 评价指标差值 Δ 和 综合方差 σ_{pooled} 决定。

因此样本量大小的调整依靠这四个因素。

5.2 实践中计算

绝大部分的 A/B 测试都会遵循统计中的惯例：**把显著水平设置为默认的 5%，把 Power 设置为默认的 80%**，这样的话我们就确定了公式中的 Z 分数，而且四个因素也确定了两个(α 、Power)，因此样本量计算公式可以简化为：

$$n \approx \frac{16 * \sigma_{\text{pooled}}^2}{\Delta^2}$$

那么，样本量大小就主要取决于剩下的两个因素：

- 实验组和对照组的综合方差 σ_{pooled}^2
- 两组评价指标的差值 Δ

5.2.1 综合方差的计算

综合方差的计算公式如下：

$$\sigma_{\text{pooled}}^2 = \frac{(n_{\text{treat}} - 1) \cdot s_{\text{treat}}^2 + (n_{\text{control}} - 1) \cdot s_{\text{control}}^2}{n_{\text{treat}} + n_{\text{control}} - 2}$$

其中，实验组的样本方差为 s_{treat}^2 ，对照组的样本方差为 s_{control}^2 ，而实验组和对照组的样本大小分别为 n_{treat} 和 n_{control}

这个公式基于自由度为 $n_{\text{treat}} + n_{\text{control}} - 2$ 的 t 分布，用于将两组样本的方差合并成一个总体方差。

综合方差在统计假设检验和置信区间估计中非常有用，因为它能更好地估计总体方差，从而提高了统计分析的准确性。

5.2.2 两组评价指标的差值

两组评价指标的差值 Δ 的计算公式如下：

$$\Delta = \overline{X}_{treat} - \overline{X}_{control}$$

对于每个组，计算评价指标的平均值，这可以通过将所有数据点相加并除以样本大小来完成。对于实验组，表示为 \overline{X}_{treat} ，对于对照组，表示为 $\overline{X}_{control}$ 。

5.3 比例类最小样本量

可以使用以下公式计算比例类AB实验的最小样本量：

$$\begin{aligned} n &= 2 \times \frac{(Z_{\frac{\alpha}{2}} + Z_{power})^2}{\frac{\Delta^2}{\sigma_{pooled}^2}} \\ &= 2 \times \frac{(Z_{\alpha/2} + Z_{1-\beta})^2}{\frac{(p_1 - p_2)^2}{(p_1 \cdot (1 - p_1) + p_2 \cdot (1 - p_2))}} \end{aligned}$$

6. 流量分配

通常网站会利用分层和分流的机制保证本站的流量高可用，原因有以下几点：

1. 网站的流量是有限的
2. 实验的对象是多层的或同一层内互不干扰的
 - 多层：例如网站不仅仅有UI层（界面），通常还有算法层等。
 - 同一层内互不干扰：例如网站的推荐位有多个（首页推荐位、商详页推荐位）。
3. AB 实验的需求是大量的

6.1 分层规则

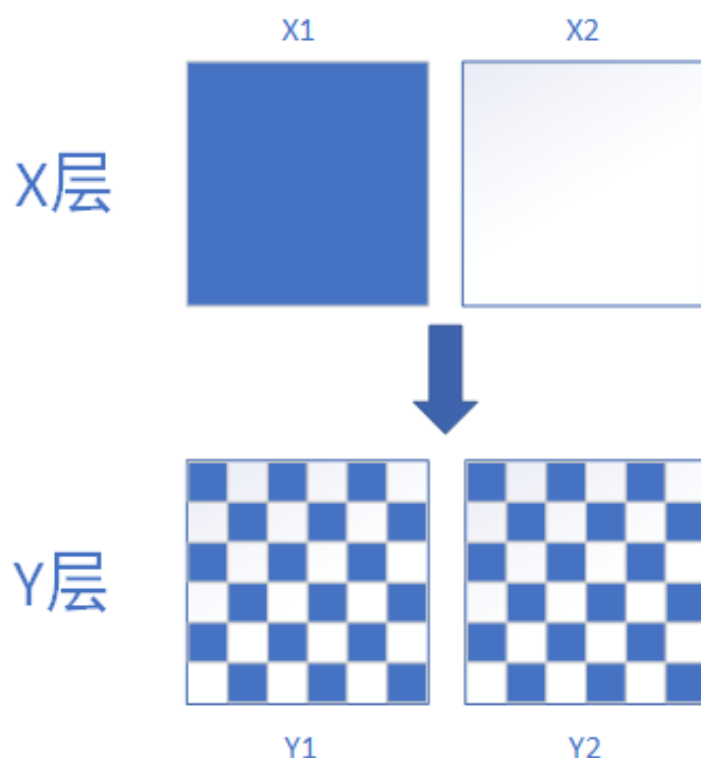
根据不同的实验共享流量的情况下，不同的实验之间是否会产生干扰，将实验类型分为 **正交实验** 和 **互斥实验**

为了更充分、更高效的使用流量，实际往往是多组试验同时存在，既有正交，又有互斥。

6.1.1 正交实验

正交实验：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。

正交是指用户进入所有的实验之间没有必然关系。比如进入X层的用户再进入Y层也是均匀分布的，而不是集中在某一块区间内。



如何理解正交？

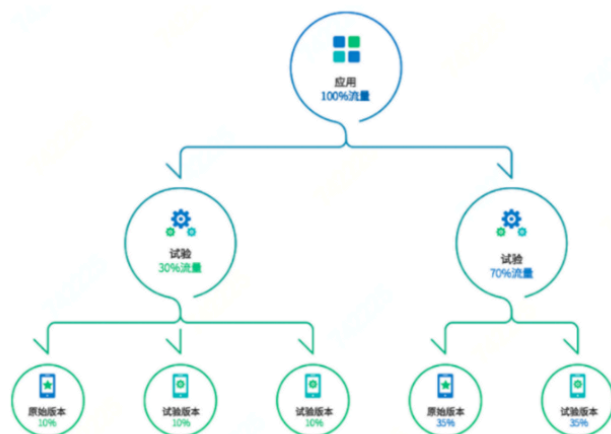
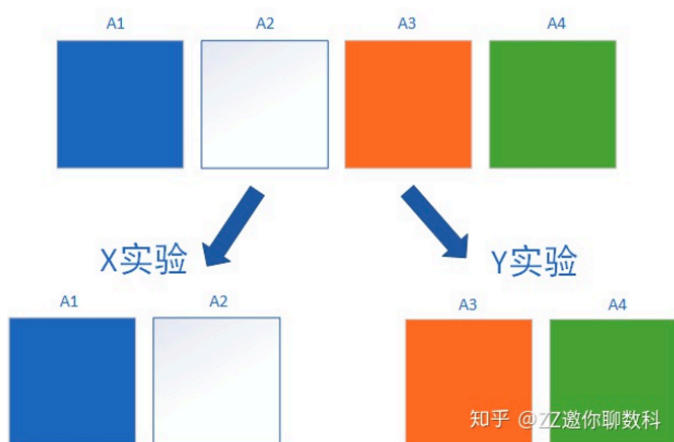
例如：我们有100个乒乓球，随机拿出来50个染成蓝色，50个染成白色，则我们有蓝色、白色乒乓球各50个，现在我们把这100个乒乓球重新放在袋子中摇匀，随机拿出50个乒乓球，那么这50个乒乓球颜色蓝色和白色各25。

正交实验的意义：各分层之间的流量是正交的，可以保证不同流量层的实验不会互相影响。将一个实验A的实验组和对照组的流量随机均匀分给另一个实验B的实验组和对照组，由于分配是均匀的，所以实验A对实验B的影响被均匀打散，从而避免实验A对实验B的结果产生影响。

6.1.2 互斥实验

互斥实验：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。

指两个实验流量独立，用户只能进入其中一个实验。比如进入X实验的用户就不能进入Y实验。



如何理解互斥？

例如：我们有100个乒乓球，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。

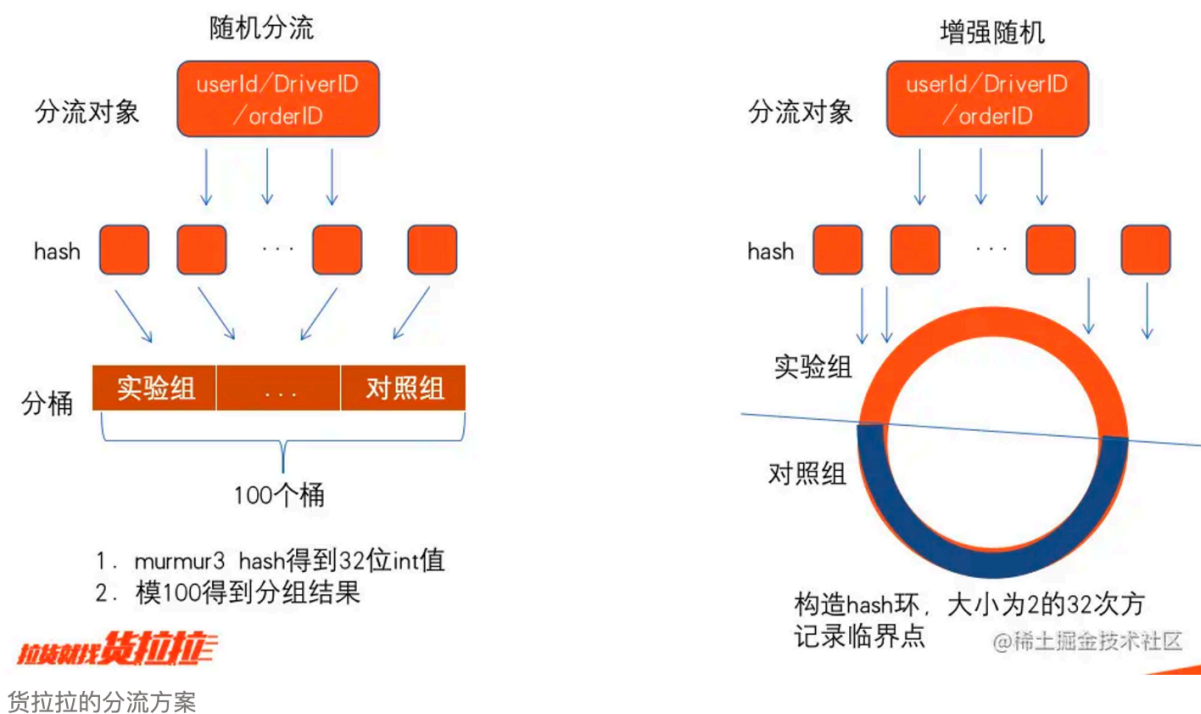
6.2 随机分流

所谓“**同质性**”就是保证对可能影响实验指标的因素，实验组和对照组应该是一致的。

为了保证不同的组的样本的“同质性”，最简单的办法就是随机，将总样本完全随机分成不同组，基本上能保证不同组样本的同质。这也是业内目前最常见的A/B实验分流算法——**随机分流**。

随机分流算法是业内最常见的A/B实验分流算法，能够满足大部分A/B实验场景，比如广告投放、UI样式、营销派券等场景对A/B实验的诉求。随机分流算法的设计也较为简单，通过**将分流ID（用户ID/司机ID等）随机分到不同组的任意一组**即可。

但需要特别注意的是，A/B实验一般都需要满足：进组的用户不再出组，即：同一个分流id，在不调整分组流量占比的情况下，无论多少次进入A/B实验，都应该在同一个分组。



6.3 Summary

在 AB 测试中，流量分割策略是指 **如何将流量平均分配到控制组（control group）和实验组（experiment group）中**。这是非常重要的，因为如果流量分配不均匀，就可能对结果造成影响。

下面是一些设计流量分割策略的建议：

1. 尽量平均地分配流量：尽量将流量平均分配到控制组和实验组中，这样可以减少因流量分配不均匀对结果的影响。
2. 考虑流量来源：如果流量来源不同，则可能会对结果产生影响。因此，应尽量使流量来源相似，例如，如果有多个渠道，则应尽量使每个渠道的流量比例相似。
3. 考虑流量特征：如果流量中存在某些特征，例如地区、设备类型、浏览器等，则应尽量使控制组和实验组的流量特征相似。
4. 使用随机分配：可以使用随机分配的方法将流量分配到控制组和实验组中，这样可以最大程度地减少因人为因素对结果的影响。

7. 分析实验结果前的检查

7.1 实验有效天数

实验的有效天数的确定需要考虑两个因素：

- 试验进行多少天能达到流量的最小样本量
- 同时还要考虑到用户的行为周期和适应期

用户的行为周期

部分行业用行为存在周期性，例如电商用户购买行为，周末与工作日有显著差异。故实验有效天数应覆盖一个完整的用户行为周期。

用户适应期

如果进行的样式改版一类的实验，新版本上线用户会因为新奇效应而存在一定得适应期。故应考虑适应期在实验有效天数内，然后再分析实验结果。适应期的长短通常以足量用户流量参与试验后的2到3天为宜。

7.2 完整性检查

进行实验结果评估前，应该先进行完整性检查（Sanity Checks），确保已经恰当地完成了实验。

实验中有太多环节可能导致实验结果是无效的。例如实验分配失误，导致实验组和对照组无法进行对比；又或者数据收集过程出错了。在评估前，应该对实验检查。

如果完整性检查都失败，则可能背后的实验设计、基础设施或数据处理都是有问题的。

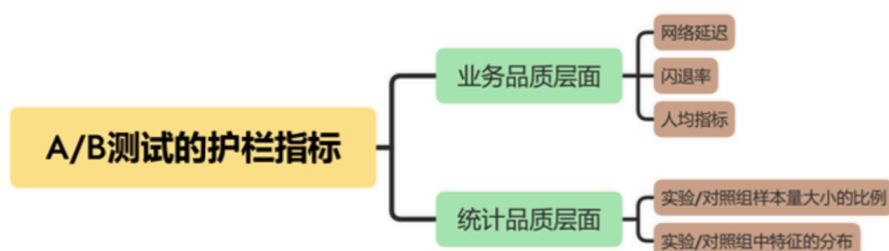
- 检查一些不变量是不是在实验中真的没有发生变化。例如实验单元数量是否在实验前后大致相同。
- 如果检查失败了，则不再对实验结果分析，而是直接分析为什么完整性检查失败了。
 - 可能是技术上的错误，与工程师一起检查，可能是实验架构出错
 - **!!** 回顾性分析，尝试从数据采集方面重新进行实验分组
 - 利用在实验前后、实验中对比, 可能是实验设置出错

7.3 合理性检验 - 护栏指标

保证测试的质量、确保AB测试具体实施过程符合预期设计，一般使用护栏指标进行检验

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。



7.3.1 业务品质层面

保证用户体验的同时，兼顾盈利能力和用户参与度。

通常会用到的护栏指标主要是三个:网络延迟(Latency)、闪退率(Crash Rate)和人均指标。

- 网络延迟：**网页加载时间、App 响应时间等，都是表征网络延迟的护栏指标**。增加产品功能可能会增加网页或 App 的响应时间，而且用户可以敏感地察觉出来。
- 闪退率：闪退发生的概率虽然不大，但是会严重影响用户体验。
- 人均指标：
 - 收入角度，比如人均花费、人均利润等。→ 反应产品的盈利能力
 - 用户参与度，比如人均使用时长、人均使用频率等。→ 反应用户的满意程度

7.3.2 统计品质层面

统计方面主要是尽可能多地消除偏差，使实验组和对照组尽可能相似，比如检测两组样本量的比例，以及检测两组中特征的分布是否相似。

造成偏差的原因：

- 随机分组的算法出现bug
- 样本量不够大
- 触发实验条件的数据出现延迟等

问题类型

1. 实验组、对照组样本大小比例：

- 实验组和对照组样本大小的比例，预期是 1:1=1。但有的时候，当实验结束后却发现两者的比例并不等于 1，甚至也没有很接近 1。这就说明这个实验在具体实施的过程中出现了问题，导致实验组和对照组出现了偏差。

2. 实验组、对照组中特征的分布

- A/B 测试中一般采取随机分组，来保证两组实验对象是相似的，从而达到控制其他变量、只变化我们关心的唯一变量(即 A/B 测试中的原因)的目的。

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。

8. *实验前：AA实验

A/A实验通常是在AB实验的早期阶段或设计阶段进行的。A/A实验是一种控制实验，其目的是验证实验系统的稳定性和确保实验的有效性。在A/A实验中，两个或多个组被随机分配到相同的处理条件，也就是相同的A条件，以确保在实验中没有任何预期的效果或差异。

A/A实验的主要目标包括：

1. 检查实验系统是否正常运行：通过比较相同条件下不同组的结果，可以确定实验系统是否能够生成一致的基准数据。
2. 检查随机分组的有效性：A/A实验有助于验证随机分组是否均匀，并且各组之间没有显著的差异。
3. 验证测量工具的可靠性：A/A实验还可以用来评估用于收集数据的测量工具的可靠性，以确保它们能够准确地捕捉结果。

一旦通过A/A实验确认实验系统的稳定性和有效性，就可以继续进行AB实验，比较不同处理条件（A和B）的效果。这有助于确保实验结果的可信度，并减少由于实验系统问题而引起的误导性结果。

总之，A/A实验是AB实验的前期步骤，旨在确保实验的有效性和可靠性。

9. 结果分析

9.1 结论分析过程(Recall)

做结论的过程：

1. 一般来说，ab 测试有四类指标（不变指标）
2. 选择参数：选择显著性水平，统计力量 and 实际意义水平
3. 计算所需的样本量（例如主题测试，人口测试）
4. 为对照/治疗组取样本并进行测试（持续时间，曝光程度和学习效果）
5. 分析结果并得出结论

- 完整性分析检查：检查你的不变指标是否已更改
- 分析结果：第一轮检查是否真的没有显著差异，第二轮利用不同方法进行交叉检查

6. 得出结论

- 你明白这个改变吗？你想推出改变吗？我该如何决定是否启动更改？
- 问自己，我了解实际对我们的用户体验所做的更改吗？我是否具有统计意义和实际意义的结果，以证明变更的正确性？
- 最后但并非最不重要的是，它是否值得冒险？

9.2 问题：测试结果不显著，如何解决

9.2.1 为什么会出现实验结果不显著

- A/B 测试中的变化确实没有效果，所以两组的指标在事实上是相同的。
- A/B 测试中的变化有效果，所以两组的指标在事实上是不同的。但是由于变化的程度很小，测试的灵敏度，也就是 Power 不足，所以并没有检测到两组指标的不同。

如果是第二种原因，那我们可以从 A/B 测试的角度进行一些优化和调整。

具体来说就是，通过提高 Power 来提高 A/B 测试检测到实验结果不同的概率。Power 越大，越能够准确地检测出实验组与对照组的不同。所以当我们提高了 Power 之后，如果仍然发现测试结果不显著，这样才能得出“两组指标事实上是相同的”的结论。

9.2.2 如何提高统计功效

$$n = 2 \times \frac{(Z_{\frac{\alpha}{2}} + Z_{1-\beta})^2}{\left(\frac{\Delta}{\sigma_{\text{pooled}}}\right)^2} = 2 \times \frac{(Z_{\frac{\alpha}{2}} + Z_{\text{power}})^2}{\left(\frac{\Delta}{\sigma_{\text{pooled}}}\right)^2}$$

从计算样本量的公式来看，影响Power的因素有：

1. 样本量：样本量和 Power 成正比。即通过增大样本量就可以提高 Power。
2. 方差：方差和 Power 成反比。即通过减小方差就可以提高 Power。

具体来说，实践中：

- **增加样本量**

在有条件获得更大样本量的情况下，可以选择增大样本量的方法来提高 Power，相对简单易操作。

- **减小方差**

如果受流量或时间限制，没有条件获得更多的样本量，此时可以通过减小方差来提高 Power。

9.2.2.1 如何通过增加样本量来提高 Power

1. 延长测试时间

每天产生的可以测试的流量是固定的，那么测试时间越长，样本量也就越大。所以在条件允许的情况下，可以延长测试的时间。

2. 增加测试使用流量在总流量中的占比

假设某个产品每天有 1 万流量，如果我要做 A/B 测试，并不会用 100% 的流量，一般会用总流量的一部分，比如 10%，也就是测试使用流量在总流量中的占比。

1. 考虑试错成本：使用的流量越少，试错成本越低，也就越保险。
2. 考虑产品的媒体效应：在大数据时代，对于互联网巨头来说，由于本身就拥有巨大的流量，那么产品本身做出的任何比较明显的改变，都有可能成为新闻。

3. 多个测试使用同一对照组

1. 增加每组的流量利用率
2. 在同一个基础上想同时验证多个变化，也就是跑多个 A/B 测试有相同的对照组的时候，我们可以把对组合并，减少分组数量，这样每组的样本量也会增加。这种测试又叫做 A/B/n 测试。

9.2.2.2 如何通过减小方差来提高Power

1. 减小指标的方差

- 保持原指标不变，通过剔除离群值 (outlier)来减小方差
 - 通过设定封顶阈值(Capping Threshold)的方法把离群值剔除掉。
- 选用方差较小的指标

2. 倾向得分匹配(Propensity score matching): 因果推断的一种方法，目的是解决实验组和对照组分布不均匀的问题。

- 两组的各个特征越相似，就说明两组的方差越小。
- 倾向评分越接近，说明两个数据点越相似。
- PSM具体算法：
 - 把我们要匹配的两组中每个数据点的各个特征(比如用户的性别，年龄，地理位置，使用产品 / 服务的特征等)放进一个逻辑回归(Logistics Regression)中。
 - 计算得到的logistics得分即为1. 每个数据点的倾向评分

- 通过最近邻等方法对相近的倾向得分对应的样本点匹配
- 最后我们只需要比较匹配后的两组相似的部分即可。
- **PSM 能够有效地减少两组的方差。通过比较倾向评分匹配后的两组的相似部分，我们可以来查看结果是否显著。**

3. 在触发阶段计算指标：在 A/B 测试中我们把实验单位进行随机分组的这个过程叫做分配 (Assignment)。我们要测试的变化是需要满足一定条件才能触发的。

- 变化不需要条件触发。所有用户在被分配到实验组后，就都可以体验到 A/B 测试中的变化。
- 变化需要条件触发。在被分配到实验组的所有用户中，只有满足一定条件的用户才会触发 A/B 测试中的变化。