AB Test

# 1.什么是ab实验

AB实验（A/B Testing）是一种对比两种或多种版本的实验方法，通过随机分配用户到不同的实验组，以评估哪一个版本更有效。通常用于网页设计、用户界面优化、营销策略评估等领域。

# 2.ab实验的统计学原理

## 基本原理

AB实验利用控制变量法的思想，保证各个产品方案针对同质人群（特征分布相同）、同一时间进行实验，**确保除方案变量外其他变量一致，从而判断指标差异是方案不同造成的，从而选择优胜版本全流量上线，实现数据增长。**

## 实验思想

即假设检验。**先对总体的特征做出某种假设，然后通过抽样（即分组或者分桶）研究的统计推理，对此假设应该被拒绝还是接受做出推断（判断该改进是否有促进的效果）。**

既然AB Test的本质就是做一个假设检验，那么就必须要了解相关统计推断的　统计变量

显著性水平:表示为α，是一个概率值，原假设为真时，拒绝原假设的概率，常取值为0.05、0.01、0.10(也就是犯**第一类错误**的概率——去真)；同时置信度或者是置信水平即1-α

除了第一类错误之外，统计学中还有第二类错误，即原假设为假时，接受原假设——存伪

在统计学中，我们用**显著性水平α来衡量犯第一类错误的概率，β来衡量犯第二类错误的概率**。

统计功效：表示在假设检验中，拒绝原假设（H0）而接受备择假设（H1）的概率，所以统计功效等于1-β；即实际效应存在时能正确检测到效应的概率。统计功效越高，实验结果越可信。

*这有一个问题：为什么拒绝原假设（H0）而接受备择假设（H1）的概率就是实际效应存在时能正确检测到效应的概率？*

*答：*

*在假设检验中，我们通常设定两个假设：*

*• 原假设（H0）：通常表示没有效应或没有差异。例如，AB测试中，H0可能表示A组和B组之间的点击率没有差异。*

*• 备择假设（H1）：表示存在效应或存在差异。例如，H1可能表示B组的点击率显著高于A组。*

统计功效是实际效应存在时正确拒绝原假设的概率。它衡量的是在备择假设为真的情况下，我们能够检测到这个效应并拒绝原假设的能力。用公式表示，统计功效为1 - β，其中β是犯II类错误的概率。（**简言之就是统计功效衡量了这个效应存在的可能性究竟有多大**）

**一般来说当显著性水平在5%以下，统计功效在80%（或90%）以上，即β在0.2（或0.1）以下时，我们会认为实验结论是相对可信的。**

理论上，自然希望犯这两类错误的概率都很小。**当样本容量n固定时，α、β不能同时都小，即α变小时，β就变大；而β变小时，α就变大。一般只有当样本容量n增大时，才有可能使两者变小。**

# 3.ab实验的要点

1. **A组和B组**：

• **A组**：也称为对照组（Control Group），是保持现状或使用当前版本的组。

• **B组**：也称为实验组（Variant Group），是应用新变化或新版本的组。

• 有时实验会有多个实验组（B、C、D等）。

2. **随机分配**：

• 用户被随机分配到A组或B组，确保两组之间的其他条件尽可能一致，减少外部变量的影响。

3. **目标指标**：

• 明确实验的目标，如点击率（CTR）、转化率（Conversion Rate）、平均订单价值（Average Order Value）等。

# 4.ab实验的实施步骤

## 1确定目标和假设

确定实验的具体目标，如提高注册率、增加销售量等。

提出假设这里通常需要两步：

提出直观假设：如“更改按钮颜色可以提高点击率”。

定义统计假设：是为了在统计分析中明确你要检验的内容，包括原假设（H0）和备择假设（H1）。

原假设（H0）：表示没有效应或没有差异。即假设更改按钮颜色对点击率没有影响。

备择假设（H1）：表示存在效应或存在差异。即假设更改按钮颜色会影响点击率。

## 2确定指标

如点击率（CTR）、转化率（Conversion Rate）、平均订单价值（Average Order Value）等。这些指标应该是数据库中有的，没有的话需要埋点获取

## 3确定实验单位

常见的实验单位

1. 用户（User）：

适用于长期效果测试。例如测试某个功能的改变对用户长期使用行为的影响。

2. 会话（Session）：

适用于短期效果测试。例如测试网页某个元素的改变对当前访问会话内用户行为的影响。

3. 页面展示（Page View）：

适用于非常具体的页面级别的测试。例如测试页面加载速度对用户行为的影响。

4. 点击（Click）：

适用于细粒度的交互测试。例如测试按钮颜色对点击率的影响。

## 4计算样本量

最小样本量是指要想检测出预期的变化，至少需要多少样本才能满足。

需要的参数为：

表格

描述已自动生成

其计算公式为：

文本, 信件

描述已自动生成

举个例子：

假设我们进行一个实验，目标是检测新网页设计是否增加了用户停留时间。带入参数如下：

• 显著性水平（α）：0.05

• 统计功效：80%

• 数据的标准差（\sigma）：5分钟

• 希望检测到的效应大小（\Delta ）：2分钟

专门的样本量计算工具：https://www.evanmiller.org/ab-testing/sample-size.html

## 5分组并进行实验

## 分组

分流（也称为“随机分配”或“分桶”）是AB测试中的一个重要步骤，用于将用户随机分配到不同的实验组或对照组。具体来说，分流的目的是**确保每个用户都有相同的概率被分配到各个实验组，以减少选择偏差和其他外部因素的影响，从而确保实验结果的可靠性和有效性。**

根据确定好的实验单位（以uid为例）

1. uid：通常代表用户ID，每个用户都有一个唯一的标识符。

2. hash因子：可以是任何能够唯一标识用户的值，通常是用户ID。

3. 计算md5取模（每个用户都被随机且公平地分配到不同的实验组，从而确保实验的随机性和结果的可靠性。）：

* + 通过MD5算法对用户ID进行哈希计算，生成一个固定长度的哈希值。
  + 对哈希值进行取模运算（通常是对实验组的数量取模），以确定用户被分配到哪个实验组。
  + 例如，如果有两个实验组A和B，对哈希值进行模2运算（有n组就是模n运算），结果为0的用户分配到组A，结果为1的用户分配到组B。

当然实际操作过程中会有专门的数据收集工具和专门的AB测试平台（如Optimizely、VWO、Google Optimize）来管理实验的分组和流量分配。

## 实验

通知相关团队：确保所有相关团队（如产品、开发、数据分析等）知晓实验启动，以便及时处理潜在问题。

监控实验：定期监控实验进展，确保数据正常收集，用户体验不受影响。

在实验进行期间，实时收集数据并进行监控：

• 数据完整性：确保数据收集的完整性和准确性，避免数据丢失和错误。

• 异常检测：监控数据中的异常情况，如突发流量变化、技术问题等，并及时处理。

## 6分析实验结果

对于AB实验来说，我们要构造的是双总体均值之差的检验，对于双总体独立大样本，方差未知，用的是z分布来做检验，假设AB实验中A组的实验结果为\bar{t_{A}}，B组的实验结果为\bar{t_{B}}，我们要验证A组和B组实验结果是否有显著差异即\bar{t_{A}} - \bar{t_{B}}是否等于0，则其要检验的统计量为：

z = \frac{(\bar{t_{A}} - \bar{t_{B}} ) - 0}{ \frac{\sqrt{s_{A}^{2}}}{n_{A}} + \frac{\sqrt{s_{B}^{2}}}{n_{B}} }

（火山引擎AB测试用的是t检验，t检验对于小样本、大样本都试用）表格

描述已自动生成

# 7.得出结论

进行统计检验，计算检验统计量（如t值、z值、χ²值）和p值。

p值表示在原假设（H0）为真的情况下，观察到当前数据或更极端数据的概率（意味着p值不是预先设定的而是计算观察得到的）。p值（p-value），在统计学中是指在原假设（H0）为真的情况下，得到观测数据或更极端数据的概率。它是用来帮助我们判断实验结果是否具有统计显著性的一个关键指标。**几何上，p值是在分布图上计算出统计量值之后，两侧尾部区域的面积总和（对于双尾检验）或单侧尾部区域的面积（对于单尾检验）。**

**p值的范围**：

• p值的取值范围是从0到1。

• 小的p值（通常小于0.05）表示观测数据在原假设下出现的概率很低，因此我们倾向于拒绝原假设。

• 大的p值（通常大于0.05）表示观测数据在原假设下出现的概率较高，因此我们不能拒绝原假设。

2. **显著性水平（α）**：

• 在进行假设检验之前，研究人员通常会选择一个显著性水平（α），如0.05、0.01等。

• 这个显著性水平代表研究人员愿意接受的犯I类错误（错误地拒绝原假设）的最大概率。

3. **决策规则**：

• 如果p值 ≤ α，则拒绝原假设（H0），接受备择假设（H1），表示结果具有统计显著性。

• 如果p值 > α，则不拒绝原假设（H0），表示结果不具有统计显著性。

# 8.做ABTest时需要考虑的问题

1. 人群是否同质

设计合理的分流算法，确保分流到每个方案的都是同一特征分布的人群是AB实验结论可信的前提。

2. 实验是否同时

进行对比的各个实验版本一定要同时开启实验，实验过程中不能随意修改每个版本的流量，否则也会产生一些问题。

3. 是否需要AA实验

AA实验指的是实验中的各个方案都是一致的，目的是为了测试埋点、分流、实验统计的正确性，增加AB实验的实验结论可信度。

设命题1：如果实验的埋点、分流和统计都没有问题，那么AA实验中各个方案的数据表现一定一致。

则逆否命题2：如果AA实验中各个方案的数据表现存在显著差异，则实验的埋点、分流和统计至少一项有问题。

**严格意义上讲AA实验的通过并不能证明上述三项（埋点、分流、统计）绝对没有问题，但是AA实验不通过一定能证明上述三项至少一项存在问题。**

4. 实验结果是否反转

假如一个实验在线上第一天，方案A比方案B效果好，但并不代表第二天、第三天仍是这个结果。用户进入到新方案中，可能是因为好奇而表现的更加活跃，但随着时间推移，逐渐趋于冷静，数据可能回到本该有的水平，如果过早容易得出错误的结论。

另外一方面，如果实验样本量太少，实验结果的随机性也会比较强，随着实验天数增加，结果也可能出现反转。