AB常见考点

# 一、AB概述

## 题目1：为什么要进行AB实验？

√ 一个想法的价值是很难进行预估的

√ 小改动可以有大影响

√ 大增长的实验很罕见

√ 实验运行的成本要低

A方案和B方案主观上各有优劣，无法判断，选取一部分用户样本，做AB测试，根据上线之后的数据表现，来决定哪个更好。

## 题目2：AB实验的优缺点？

**⮚ 优点：**

√ 高概率确立因果关系的最佳方法

√ 能够检测微小变动

√ 能够检测出意想不到的变动

**⮚ 缺点：**

√ 额外的A/B测试代码以及数据采集链路，增加了开发工作量与后期维护成本

√ 由于增加了实验分桶计算，增加了用户请求响应时间

## 题目3：AB实验的目标可以分成哪几类？

**⮚ 策略评估：**

最常见的AB实验目标，如：字体大小、颜色的改变

**⮚ ROI评估：**

如：双十一大促，发券的ROI评估

# 二、统计学

## 题目4：在AB实验中，以下哪些属于随机事件？

● 淘宝每日人均成交额：属于随机事件

● 抖音首页视频曝光点击率：属于随机事件

● 西瓜视频用户的人均播放时长：属于随机事件

● 西瓜视频app用户的观看时长：属于随机事件

● 京东app黑屏的占比：属于随机事件

● 安卓用户安装支付宝的比例：属于随机事件

## 题目5：请列举在互联网工作中有什么必然和不可能事件？

**必然事件：**比如：事件1：淘宝订单填写页面报错，无法跳转支付，事件2：成交金额是0；这事件1与事件2是必然事件

**不可能事件：**比如：事件1：淘宝平台最高商品单价100万，事件2：某天平台上只有一个客户下单，仅下单一件商品，金额1亿。事件1与事件2是不可能事件。

## 题目6：请判断以下随机事件属于什么分布？

● 淘宝用户每日支付的订单数：泊松分布

● 腾讯视频用户是否为会员用户：0-1分布

● 腾讯视频用户人均观看视频的时长：正态分布

## 题目7：请计算淘宝用户成交金额小于100的概率？

根据中心极限定理，每个淘宝用户的成交金额都是独立的，当淘宝用户充分大时（n>30），平均成交金额X服从正态分布N( μ, (δ^2)/n )， X~N( μ, (δ^2)/n )

P(X<100) = P( (X-μ)/ sqrt(δ^2/n) < (100-μ)/ sqrt(δ^2/n) )

## 题目8：请计算腾讯用户观看时长小于1小时的概率？

根据中心极限定理，每个腾讯用户的观看时长都是独立的，当腾讯用户充分大时（n>30），平均观看时长X服从正态分布N( μ, (δ^2)/n )， X~N( μ, (δ^2)/n )

P(X<1) = P( (X-μ)/ sqrt(δ^2/n) < (1-μ)/ sqrt(δ^2/n) )

## 题目9：腾讯视频中《扫黑风暴》在首页推荐的曝光点击率是否服从正态分布？（提示：一个用户可以多次打开视频app。且每次打开均会曝光《扫黑风暴》）

不服从正态分布，样本不独立

## 题目10：腾讯视频中《扫黑风暴》在首页推荐的uv点击率是否服从正态分布？

服从正态分布，中心极限定理

## 题目11：腾讯视频中《扫黑风暴》的人均观看时长是否服从正态分布？

服从正态分布，中心极限定理

## 题目12：结合自己的工作和生活经验，可以想想身边有哪些符合假设检验的例子？

传统保险理赔业务中，理赔发票人工录入理赔系统，录入界面的调整是否能对录入效率提升。

使用新录入界面作为实验组，使用旧录入界面作为对照组

H0：实验组的人均录入票据数量 = 对照组的人均录入票据数量

H1：实验组的人均录入票据数量 <> 对照组的人均录入票据数量

## 题目13：结合具体的业务背景说明一下什么是中心极限定理？

**中心极限定理：**由n的独立同分布的变量（n>30）的均值是一个均值为μ、方差为 (σ^2)/n的正态分布

腾讯视频用户人均观看视频的时长符合正态分布

## 题目14：请结合具体的业务背景举例说明如何构造假设检验？

**背景：**

传统保险理赔业务中，理赔发票人工录入理赔系统，录入界面的调整是否能对录入效率提升。

**实验设计：**

实验组：使用新录入界面

对照组：使用旧录入界面

**建立假设：**

原假设H0：实验组的人均录入票据数量 = 对照组的人均录入票据数量

备择假设H1：实验组的人均录入票据数量 <> 对照组的人均录入票据数量

**选择统计量：**

双样本z统计量x

**选择显著性水平：**

一般选择第一类错误5%，来看检验统计量z的值的绝对值是否大于1.96，如果大于，拒绝H0，有差异；否则，接受H0，无差异。

## 题目15：如何理解第一类错误和第二类错误？

第一类错误：当H0为真，拒绝H0，接受H1；拒真错误

第二类错误：当H0为假，接受H0，拒绝H1；取假错误

## 题目16：P值是什么？

当原假设H0为真时，抽出来的样本值发生的概率最大是多少。

## 题目17：什么是MDE（Minimun Detectable Effect，可检测的最小差异）？

实验组的指标与对照组的指标，最小出现多少偏差，才能监测出有差异。

实验能有效检测出来的指标差异幅度。一般在实验过程中，通过MDE来评判实验的灵敏度，通俗来讲，在评估实验结论的过程中：

⦁ MDE<真实提升率：指标评估置信。

⦁ MDE>真实提升率：指标评估缺乏置信。可以通过扩大样本量的方式，降低可有效检测的MDE值。

# 三、实验分析

## 3.1实验指标

### **题目18：实验的指标可以分成哪几类，请结合具体的业务场景进行举例说明**？

**⮚ 目标指标：**

目标指标，也称为北极星指标，它体现了产品最关心的是什么。短期内不易改变，并需要遵循以下原则：

√ **简单的：**易理解和可以被利益相关者广泛接受

√ **稳定的：**不需要每次都为了一个新功能而更新目标指标

例如：**平台总GMV**

**⮚ 驱动指标：**

也称为评价指标或者间接指标，比目标指标更短期、变化更快、更灵敏，并具有以下特性：

√ **与目标统一的：**也就是说驱动指标的提示确实可以为公司目标带来增长

√ **可行动的和相关的：**产品可以为了提供驱动指标制定相应的方案

√ **灵敏的：**驱动指标需要足够灵敏的检测到相应功能改变带来的效果

√ **不容易被操纵的：**驱动指标和目标指标都是用来衡量公司目标的，必须保障指标不容易被操纵

例如：**人均支付转化率、人均浏览时长、商品点击次数、曝光点击率、用户评价数**

**⮚ 护栏指标：**

√ 护栏指标是指我们不会违背的假设，保障AB测试中基本的合理性假设

例如：**页面成功加载率，页面加载时长**

### **题目19：什么是综合性评估指标（OEC），什么情况下需要使用该指标？**

当存在多个目标和驱动指标，将多个指标组合成一个综合性评估指标

### **题目20：什么是累计口径，什么是非累计口径？**

**举个例子**

第一日来了100个用户，第二日来了100个用户，两日中有50个用户是重复的。

两日累计去重口径用户数 = 100+100-50=150人；

两日非累计去重口径用户数 = 100+100=200人。

### **题目21：实验评估是用累计口径还是非累计口径，为什么？**

首先，解释一下什么是累计去重口径（多日累计去重口径）？什么是非累计去重口径（多日非累计去重口径）？

**举个例子**

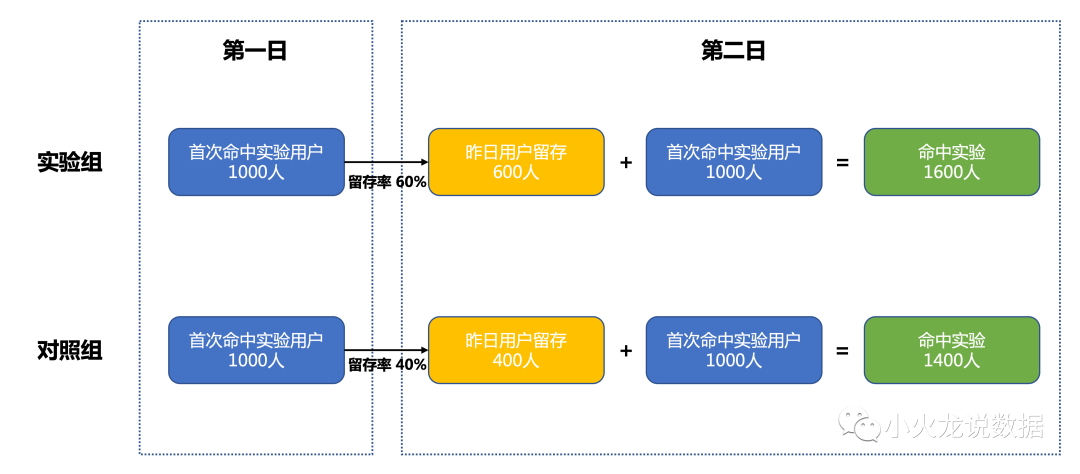
第一日来了100个用户，第二日来了100个用户，两日中有50个用户是重复的。

两日累计去重口径用户数 = 100+100-50=150人；

两日非累计去重口径用户数 = 100+100=200人。

回到指标上来，假设实验上线2日，评估指标为「人均时长=总时长/总人数」，分子总时长直接加和即可，分母总人数选择累计还是非累积口径更为科学呢？

答案是「累计去重口径」，在分组用户均衡的情况下，**累计去重口径可以保证样本量的均衡，不会受到实验策略对留存的干扰，避免用户出现有偏的情况。**如果觉得不好理解，可以看下面这张图：



由于实验组与对照组是1：1流量，因此每日首次进入分组的用户量级一致，均为1000人，但受到策略影响，实验组留存高于对照组，两种口径计算的累计用户数如下：

**采取「非累积去重口径」，两日用户数求和：**

实验组：1000+1600=2600

对照组：1000+1400=2400

**采取「累积去重口径」，两日用户数求和：**

实验组：1000+1000=2000

对照组：1000+1000=2000

**随着实验的进行，「非累计去重口径」用户数偏移会越来越严重，因此在AB实验评估中要采用「累计去重口径」更为科学。**

### **题目22：实验评估指标是否越多越好？**

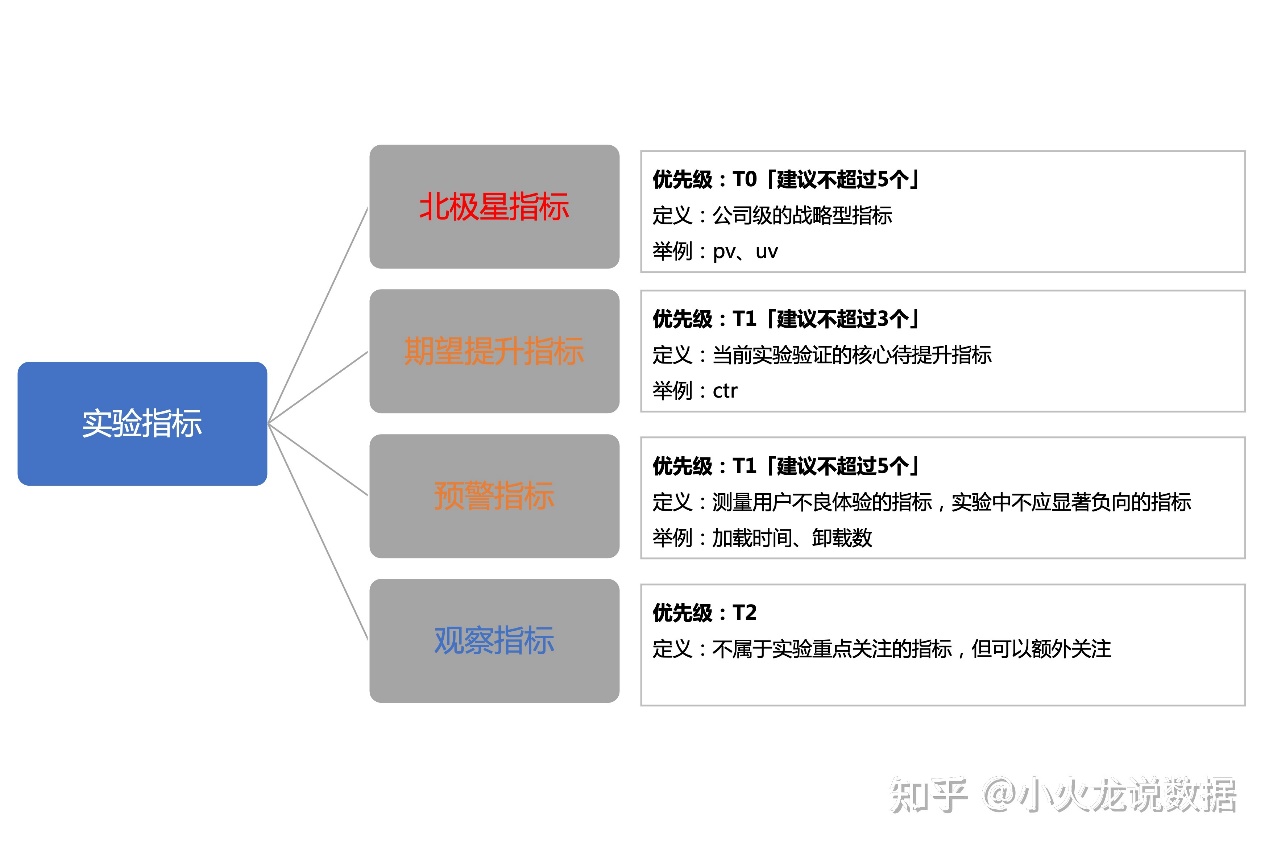
指标选择不宜过多，当指标过多时，实验犯第一类错误的概率增加，即：实验本身无差别，而误判为有差别。

**举个例子**

假设犯第一类错误的概率为5%，实验选择3个指标进行评估，则至少有一个指标犯第一类错误的概率为：1-(1-5%)×(1-5%)×(1-5%)=14.2%。

随着指标的增加，这个值会逐步增大，用一个形象的成语描述“言多必失”。

根据评估方向，一般选择4个方向的指标进行实验评估，如下图：



主要评估「实验自身局部的影响」以及「实验对大盘的影响」。

### **题目23：什么是多重检验，如何解决多重检验问题？**

指标数量越多，检验的次数就越多，至少有一个指标显著的概率就越高。

**解决方案：**重点关注与本次实验最为密切的指标，是否出现显著性结果。

## 3.2实验指标

### **题目24：实验的测试单元有哪几种**？

**⮚ 用户级别：**

√ 用户级别是最常见的一种测试单元。常见的测试单元有**用户ID、设备ID**。用户ID就是我们通常登录的账号。设备ID可以理解为同一个设备，比如手机。同一个手机可以登录不同的账号，同一个账号也可以在不同的设备上登录

**⮚ 会话级别：**

√ 每一次访问可以作为一次会话，不同的访问可以理解为多次进入和打开app。比如用户A上午10点打开抖音，并持续浏览20分钟后退出app。下午2点又重新打开了抖音，浏览10分钟后退出app。这就是2次访问记录了。

**⮚ 页面级别：**

√ 页面级别可以理解为每次页面的浏览都是一个单元。这里需要注意的是，同一个页面访问多次，即为多个单元

### **题目25：如何选取测试单元，需要遵循什么原则**？

√ **连续性：**需要保证用户体验的连贯性。比如某个页面的功能改动，最好是以用户级别进行测试单元的划分，否则可能会出现同一个用户时而看到该功能，时而看不见该功能。这样会让用户体验上变差，影响整个app的使用。

√ **一致性：**测试单元需要和驱动指标保持一致。假设去的指标为用户营收，那么测试单元就应该以用户为颗粒度进行测试。

### **题目26：什么是网络效应（提示：微信社交）**？

产品可以划分为「单边产品」和「多边产品」。

**单边产品，**例如：搜索引擎，用户使用之间是不会相互干扰的；

**多边产品：**例如：聊天软件，用户之间存在互通，一方影响会传导到另外一方。

而这种「多边产品」用户之间相互影响的情况，我们称之为网络效应。

**那么，网络效应为什么会影响实验呢？**

**举个例子**

微信在进行策略迭代时，随机选择两拨人群作为实验组A和对照组B。新策略在A组中执行，假设策略使得A组用户聊天频次显著提升，而社交性产品由于存在交互性，往往会导致B组用户的聊天频次也相应提升。这种AB组用户直接或间接干扰的情况，则是网络效应对实验的影响。

### **题目27：存在网络效应的实验，如何进行实验设计**？

**步骤一：**将用户按照地理位置、社会群体等方式进行用户群体切割，AB桶分别选择相对独立的两个群体。

**步骤二：**由于群体之间是存在差异性的，因此需要尽可能选择相似的用户群体，结合PSM等方式进行评估。例如：按照地理位置划分，假设北京和上海在重点关注的特征上比较相近，则可以作为实验的AB组。

## 3.3实验周期

### **题目28：实验样本量是不是越多越好**？如何确认实验样本量？

**⮚ 实验角度，样本量越多越好：**

样本数量变多，实验则有了更多的“证据”，实验的“可靠性”也就越强。

**⮚ 业务角度，样本量越少越好**

**（1）试错成本大**

假设我们拿50%用的户来跑实验，但不幸的是，1周后结果表明实验组的总收入下降了20%。算下来，你的实验在一周内给整个公司带来了10%的损失。这个试错成本未免高了一些...

**（2）其它风险增加**

移动端例子，假设B方案崩溃率增长，1%流量我们可以从容处理，50%流量会对业务造成严重影响，甚至事故定责。

**（3）流量有限**

流量总数是确定的，同类型的实验不能重叠，实验流量更小，就可以同时运行更多的实验。

满足最小样本量 =

β=20%

1-β=80%

Z(1-β)=0.84

α=5%

Z(1-α/2)=1.96

SA为对照组标准差， 实验组标准差SB=SA

δ：提升量（提前预估，如：会提升0.2%）

### **题目29：如何确认实验周期，应该考虑什么因素**？

**（1）满足最小样本量**

实验周期 = 根据差异预估的对应每组最小样本量 / 每天进入实验的流量

**（2）周期效应：**用户的操作行为存在很大的周期性变化

√ 周内效应：用户行为以周为单位进行变化

√ 季节性效应：对有明显季节性/节假日表现差异的实验，比如酒店预订

**（3）新奇效应：**

√ 实验可能会存在刚开始有比较强烈/微小的反应，用户要经历一段适应的过程，因此需要一定适应期逐渐建立用户基础，然后再考察实验结果。

### **题目30：实验如何进行分流，一般使用何种算法**？

（1）确认测试单元Target之后，再确认分层Layer

（2）选用一个hash函数，比如md4

（3）确认分组数量N，比如随机分成100组

（4）通过hash(Layer,Target)%N确认每一个测试单元对应的分组信息

### **题目31：如何应用hash算法实现不同实验的正交**？

hash( layer, uid )

### **题目32：如何应用hash算法实现不同实验之间的互斥**？

hash( layer, uid ) % N ，假设N=4，会有0、1、2、3，这4个桶。

实验1分配0号桶

实验2分配1号桶

实验3分配2号桶

实验4分配3号桶

### **题目33：如何控制实验用户的比例**？比如实验组和对照组只需要各10%的用户

hash( layer, uid ) % N，N=10，会有0、1、2、3、4、5、6、7、8、9，这10个桶

实验组分配1号桶

对照组分配0号桶

### **题目34：最小样本量的计算公式是？为什么要计算最小样本量？**

最小样本量 = ，用最小的流量满足实验

β=20%

1-β=80%

Z(1-β)=0.84

α=5%

Z(1-α/2)=1.96

SA为对照组标准差， 实验组标准差SB=SA

δ：提升量（提前预估，如：会提升0.2%）

### **题目35：如何判断是否存在新奇效应？如何剔除新奇效应的影响？**

实验可能会存在刚开始有比较强烈/微小的反应，用户要经历一段适应的过程，因此需要一定适应期逐渐建立用户基础，然后再考察实验结果。

## 3.4 AA校验

### **题目36：AA测试的定义是什么**，目标是什么？

**⮚ 定义**

像AB测试一样将用户分成两组，但两组的策略是一样的，因此命名为AA测试，有时也成为空测试

**⮚ 目的**

√ 确保第一类错误控制在预期范围内

√ 评估指标的波动性

√ 确保实验组和对照组的用户之间没有偏差

√ 确保实验组和对照组的分流比例符合预期

### **题目37：如果实际分流后发现用户分流比例不符合预期该怎么办**？

（1）检验分流代码是否有问题

（2）发生了小概率事件，重新分层信息重新分流

### **题目38：如何验证AA阶段第一类错误在预期范围内**？

重新分流，计算显著性p值。重复操作n次，将每次计算出来的p值<0.05，进行累计，n次操作结束后，将累计值/n，是否0.05左右小幅波动，是的话就是符合预期范围，否则不符合预期。

### **题目39：AA出现异常的原因有哪些**？

AA实验由于是在相同策略下进行的，因此分桶指标差异理论上是不显著的。但如果某些指标差异较大，出现显著情况，则可能有以下几点原因：

**原因1：下线实验惯性所导致**

**出现原因：**当分桶流量不足时，新分桶配置的流量恰巧是之前实验释放的流量，导致用户有偏，引起AA波动较大。

**解决方案：**hash分桶数量提升，尽可能将用户打散。

**原因2：打标上报异常所导致**

**出现原因：**该原因是AA存在差异最可能的原因，由于SDK调用方式不当、上报错误所引起的用户非完全随机。

**解决方案：**对实验进行样本均衡性检验，提早发现用户不平的问题。

**原因3：碰到小概率事件所导致**

**出现原因：**目前小流量实验一般显著性水平设置为5%，即：100个实验当中，大约有5个实验在AA实验过程中会出现指标显著的情况。

**解决方案：**如果业务流量足够，可以将显著性水平下降为1%或0.5%（微软建议0.01%），通过空跑期判断实验桶和对照桶是否有显著差异，并在AB上线之前，为实验桶和对照桶分配差异较小的流量。

**原因4：多重检验问题所导致**

**出现原因：**指标数量越多，检验的次数就越多，至少有一个指标显著的概率就越高。

**解决方案：**重点关注与本次实验最为密切的指标，是否出现显著性结果。

**原因5：极端用户所导致**

**出现原因：**各个产品，均存在某些极端的高销用户，例如：电商行业的土豪用户。这类用户的消费情况远远高于一般用户，而量级又很小，很难均匀的分配到分桶当中，导致干扰分桶的均衡性。

**解决方案：**有两种解决方式。其一：设置实验黑名单，将该类用户直接加入黑名单当中，不参与实验分桶；其二：根据百分比截断用户，对于指标值大于95%/98%的用户给予剔除。

### **题目40：异常值如何定义**？是否需要处理？不处理可能导致什么问题？

异常值（outlier）是指一组测定值中与平均值的偏差超过两倍标准差的测定值，与平均值的偏差超过三倍标准差的测定值，称为高度异常的异常值。

需要处理异常值，处理方式：（1）剔除处理：将这些异常用户剔除；（2）抹平处理：比如将成交金额大于5千的用户处理为5千；（3）抹0处理：比如将成交金额大于5千的用户处理为0

如不处理异常值，会使方差变大，可能导致实验结果不准确

## 3.5 AB分析

### **题目41：实验指标正向提升，是否可以上线实验**？

不可以。实验指标正向提升，不代表假设检验结果显著。

（1）显著，逐步分阶段实验放量

（2）不显著，加长实验周期，在持续观察

### **？题目42：实验指标显著提升N%，能否说明全量后也能提升N%**？

不能。

### **题目43：实验显著后可以直接扩量到100%吗**？

不可以。实验放量需要综合考虑「效率、质量、风险」三个因素，因此需要阶段性的放量，保障线上策略不会出现bug等情况。

一般实验放量分为三个阶段，分别为「小流量阶段、放量阶段、长期存放阶段」。

### **题目44：实验第一天就显著了，可否直接全量**？

不可以。可能还没达到最小样本量的要求，实验结果不可信。

# 四、其他

## ？题目45：长期实验是什么？作用是？

实验周期长。

## 题目46：实验分析的整体流程是什么？

（1）确认目标和假设

（2）确认实验指标

（3）确认测试单元

（4）确认实验周期

（5）实验分流

（6）AA测试

（7）AB分析

## 题目47：结合具体的业务场景，请说明哪些情况下不适合做AB实验？

提升长期指标，不能用AB测试；

流量较小的产品也不适合AB测试；

不能上线的策略也不适合做AB测试，比如手机启动时长对用户的留存

## 题目48：在没有AB实验的情况下，如何进行策略评估？

AB实验的本质是为了解决「因果问题」，在没有做AB实验的情况下，可以通过：DID（双重拆分法）、传递熵、因果森林等方式进行替代。

不过总体来说，AB实验仍然是处理因果问题最简单、最直接的方式。

## 题目49：什么是双边实验？并举例说明

比如：短视频主播美颜功能，进行AB实验，同时进行了主播侧和观众侧的分流：

|  |  |  |
| --- | --- | --- |
| 分组 | 实验组主播 （有美颜功能） | 对照组主播 （没有美颜功能） |
| 实验组观众 （可以看见美颜） | 看见美颜A | 看不见美颜B |
| 对照组观众 （不可以看见美颜） | 看不见美颜C | 看不见美颜D |

**观众侧溢出（外溢）：**

√ B = D，则没有差异，也就是不存在溢出

√ B > D（溢出），可能是因为实验组的观众不喜欢看美颜，更爱素颜，导致实验组的用户更倾向于无美颜功能的对照组主播

√ B < D（溢出），则可能是因为实验组的观众更喜欢看美颜，导致原本应该观看无美颜功能（对照组主播）的用户跑去看有美颜功能（实验组主播）

## 题目50：什么是非参检验？

非参数检验是在总体方差未知或知道甚少的情况下，利用样本数据对总体分布形态等进行推断的方法。