面试常见问题

1. 实验原理、实验前设计

1.1 在没有做AB实验的前提下，如何评估策略迭代的优劣

B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。

1.2 AB实验的原理和本质

1. 原理：来源于假设检验。有两个同质的样本组，对其中一个组做出某种改动，然后来观测这个改动对于我们关注的指标是否有显著的影响
    - 原假设：这项改动不会对我们关注的指标有显著的影响
    - 如果在做完试验后发现P值足够小，则推翻原假设，证明改动有影响
2. 同质样本组的对照实验

1.3 第一列错误和第二列错误的概念？实际含义？

1. 第一类错误：原假设是正确的，却拒绝了原假设。
    - 原假设 $\mu 1 = \mu 2$,原假设是正确的，意思是 $\mu 1 = \mu 2$ ，即这个功能的改变不能带来收益，但是误判，以为能带来收益
2. 第二类错误：原假设是错误的，却没有拒绝原假设。
    - 举例：原假设 $\mu 1 = \mu 2$,原假设是错误的，意思是 $\mu 1 \neq \mu 2$ ，即这个功能的改变能带来收益，但是误判，以为不能带来收益

1.4 第一类错误和第二类错误哪个更不能接受

在实际工作中，第一类错误时我们更加不能接受的。换一句更加直白的话说，就是**我们宁愿砍掉4个好的产品，也不让一个坏的产品上线**。因为一个坏的产品上线，对用户体验影响很大。

1.5 做留存率的AB Test，选择什么检验?

A留存率是指用户在使用产品之后还会继续使用产品的比率。在进行ABTest时，如果要比较两组用户的留存率，可以使用卡方检验来进行比较。

卡方检验是一种用于比较两组分类数据的差异的统计检验方法，它可以用来检验两组数据是否有显著差异。在进行留存率的ABTest时，可以将两组用户的留存情况分成“留存”和“不留存”两类，然后使用卡方检验来比较两组用户的留存率的差异是否显著。

在使用卡方检验时，需要注意的是，卡方检验的前提是两组数据的分布独立，如果两组数据的分布不独立，则卡方检验的结果可能会失真。此外，卡方检验的结果受样本大小的影响较大，当样本数量较小时，结果的可信度较低。

2. 指标选择

2.1 AB测试如何确定目标和假设

> 在 A/B 测试中确定目标和假设的重要性。A/B 测试是和业务紧密相关的，但我们往往会忽视业务中的目标，把注意力过多地放在选取评价指标上。
>
> 在我看来，这就是本末倒置，就像一个不知道终点在哪里却一直在奔跑的运动员，如果能先 明确终点，朝着终点的方向努力，会更快地取得成功。

1. 分析问题，确定想要达到的结果
2. 提出解决业务问题的大致方案
3. 从大致的解决方案中提取出具体的假设

假设要避免：

1. 拒绝模糊、太空
2. 拒绝太主观推断，要基于事实、经验等
3. 拒绝问题的原因、结果不明确
4. 拒绝定性问题

好的A/B测试的假设是什么？

	好的假设	不好的假设
来源	用户调研，数据挖掘，观察经验等	不基于事实或观察的猜测
因果	明确包含可能的原因和结果	可能的原因和结果不明确
可证伪性	可被证伪	模糊，很难被证伪
可测量性	定量的指标	定性的结果
例子	在每个专辑/歌单播放完成后增加“自动播放下一个专辑/歌单”的功能，可以提升用户下个月的续订率	我们的产品可以打入高端市场

3. 样本选择

3.1 AB实验的样本选择时，应该注意什么

在选择AB实验的样本时，应该注意以下几点：

1. 可比性、均衡性：AB实验的两个样本应该尽可能相似均衡，两个样本在重要的特征方面应该尽可能相似，以便能够更好地对比结果。
2. 样本数量：AB实验的样本数量应该足够大，以便能够得到较为准确的结果。
3. 随机性：AB实验的样本应该用随机的方式选择，以便能够更好地控制其他可能影响结果的因素。
4. 独立性：两组样本应该组间样本应该相互独立，不能产生影响，否则容易破坏SUTVA假设
5. 排除外部干扰因素：AB实验的样本应该尽可能排除外部的干扰因素，以便能够得到更准确的结果。

注意，AB实验的样本选择是影响实验结果的重要因素。 因此，在选择样本时应该认真考虑以上几点，以便能够得到较为可靠的结果。

4. 流量分配

4.1 如何确定分流样本不倾斜

- AA test，或者说分流之后先空跑一段时间看是否显著
- 检查各个协变量指标，要么和关注指标无相关性，要么在实验组和对照组分布相近。
- 或许可以用所有特征对分组结果建立逻辑回归模型，如果变量不显著，可以认为在该变量上分组均匀。（比如男女，不需要1：1，只要不显著即可）
- 使用随机分流：这是最常用的方法，即通过随机数生成器将用户随机分配到实验组或对照组。这样可以确保样本分布在两个组中是平均的，从而减少偏差。

4.2 实验组按照5%流量随机分流的依据是什么

实验组按照5%流量随机分流的依据可能是为了在进行 A/B 测试时，对实验组和对照组的流量进行平衡。

在使用 A/B 测试时，随机分流是很重要的。这意味着将参与者随机分配到实验组和对照组，以确保两组人群的差异不会影响测试结果。如果不使用随机分流，那么两组人群可能会存在明显的差异，这会对测试结果产生影响。

根据 5% 的流量随机分流的依据，意味着将参与者随机分配到实验组和对照组，使得两组人群的大小比例相差不超过 5%。这种方法可以帮助确保两组人群之间的差异较小。

4.3 AB实验中辛普森悖论

1. 在某种条件下，我们关注的两组数据，如果分别讨论，则这两组数据都会满足某种同样的性质，而当我们把两个子数据集合并再观察整体，却会得出截然相反的结论
2. 原因 ： 我们把“值”与“量”两个维度的数据合并成了“值”， 即我们划分数据集时，并没有对流量进行一个合理的分割，导致我们所选取的实验组并不具有一定的代表性
3. 影响 ： 在互联网案例中，我们使用1%的用户数据去跑实验，结论是改动效果好，结果上线后，全量用户结果却是新版本带来的用户体验下降
4. 避免方法 ： 保证对样本量进行一个合理的分配，并保证我们选取的样本量具有相似的特征，都能代表总体的特征

5. 结果分析 - 不显著

5.1 AB实验的结果在统计上显著，而在实际中却不显著

1. 可能是AB实验中选取的样本量过大，导致实验样本和总体数据量差异很小，这样即使我们发现细微的差别，在统计意义上是显著的，在实际案例中也会变得不显著了

   > 💡 举个栗子，对应到我们的互联网产品实践当中，我们做了一个改动，APP的启动时间的优化了0.001秒，这个数字可能在统计学上对应的P值很小，也就是说统计学上是显著的，但是在实际中用户0.01秒的差异是感知不出来的。

2. 根据t检验统计量公式，如果样本量过大，只需要很小的区别就会造成显著
3. 可能出现了第一类错误：将这种误导性的结论称为误报；而在统计学中，则称之为“I 类错误”（即，当原假设正确时，您错误地拒绝了该假设）
4. 只考虑了短期的测试，没有考虑长期效应（短期新奇效应）

5. 可能在测试期间更改了流量分配策略
6. 过早的停止了测试，或者在测试过程中多次观测实验

> 测试完成后做假设检验时构造置信区间受到以下三个关键因素影响：
>
> - 测试样本量： 唯一可控的因素
> - 总体标准偏差
> - 显著性水平

5.2 在AB实验中发现我们选取的指标在统计意义上都不显著，那如何判断这个实验的收益

弄清楚统计显著性和实际显著性，有时候在一些app上即使是没有达到统计显著性，例如Google某些功能上1%，2%的实际显著性也是非常高的。

1. 拆分到天 ： 将指标拆分成每一天去观察，如果 在实验中每一天实验组都高于对照组，即使在统计意义上不显著，我们也认为在这个观测周期内，实验组的关键指标的表现是优于对照组的 ，然后再比较这个效益的 增幅与我们预期增幅对比 ，若达到预期，最终也可以得出优化可以上线的结论
2. 拆分人群 ： 通过假设实验对每个人效果都一致的，然而这个前提往往不成立， 例如三四线人群和一二线人群对同一促销活动感知不一样 。在整体实验组和对照组结果不显著前提下，可以将实验组和对照组中的人群按特征（年龄、城市）单独拆分出来进行分析，在某些特征维度上可能结果是显著的。

如果实际效果不错，但是AB测试不显著？ （比如指标提升了10%但不显著）

- 指标方差过大&样本量太少——对策：增大样本量或者随机区组试验
- 第二类错误

5.3 如何判断实验组、对照组的某个指标是否有显著差异

1. 在实验开始前，对实验组和对照组先进行数据指标的监测，实验前两组指标没有显著差异（AA实验，检验两组实验人群不存在明显差异）
2. 实验结束，观测实验后的结果，根据假设检验原理设置显著性水平，在该水平下判断两组的指标是否有显著差异
3. 若实验前两组就存在差异，可采用 DID（双重差分） 的方法，查看两组的指标差距在设定显著性水平下实验前后是否有显著差异。

5.4 什么时候需要考虑实际显著性?

> 如果是需要改变现有的状态且改变有一定成本时需要是要考虑实际显著性的，但是 对于探索性质的且改变成本不高的情况下可以不考虑.

5.5 AB测试中的方差很大，如何解决

AB实验场景下，如果一个指标的方差较大表示它的波动较大，那么实验组和对照组的显著差异可能是因为方差较大即随机波动较大。

- 解决方法有:
  - PSM方法
    - PSM倾向值匹配方法（Propensity Score Matching)：观测性研究有时无法人为控制干扰因素，因此可能会导致因果推断的偏差。
    - 常规的解决思路是

      尽量模拟随机试验

      , 这样实验组与对照组在结果变量上的差异就可归因与实验条件的改变而非干扰因素或协变量施加的影响。PSM基于反事实因果模理论发展而成，属于因果推断的一种，相当于人为去造一个理想的实验环境
      - 倾向得分匹配，此时对照组和实验组对象间除了处理外，没有其他不同
  - CUPED方差缩减方法（Controlled-experiment Using Pre-Experiment Data）：
    - 先分层计算后汇总，举个例子，我们计算对照组和实验组的用户平均使用时长，可以分别按照城市划分，先计算每个城市的用户平均使用时长，然后再按照权重(各城市实验用户)计算总的。(前提是城市这个特征与用户平均使用时长高度相关)

机器学习场景下，特征的方差反而越大越好，因为如果一个特征方差为0，那么其实这个特征对于模型来说没有什么意义，所以特征方差大对于模型的训练才是有帮助的。

6. 结果分析 - 实际业务

6.1 算法部门上线了新的推荐算法，在ab-test中败给了老算法，让你找出其中的原因，需要说出具体思路和框架

在发现新的推荐算法在 AB-test 中败给了老算法后，我会采取以下思路和框架来寻找原因：

1. 分析 AB-test 的设置情况 ： 首先，我会查看 AB-test 的设置情况或者是实施中出现了一些问题，要检查包括 样本选择是否合理、AB-test 时间是否足够长、控制组和实验组的分布情况是否均衡等 。如果发现 AB-test 设置存在问题，则可能是 AB-test 的结果并不具有可信度，需要进一步调整 AB-test 的设置。
2. 分析推荐算法的工作原理：其次，我会深入了解新的推荐算法的工作原理， 并与老算法进行比较。 如果发现新算法在某些方面的表现不如老算法 ，则可能是新算法的工作原理存在问题，需要进一步调整和优化。
3. 分析数据的特征：再者，我会分析 AB-test 中使用的 数据的特征 ，包括 数据的质量、数据的分布情况等 。 如果发现数据存在某些特殊的特征，则可能会对新算法的表现产生影响，需要进一步分析。（数据对算法的质量是有很重要的影响）
4. 分析用户的行为数据 ： 最后，我会分析 AB-test 中，可能算法中的一些改动让用户明显察觉到不同，导致部分用户体验下降；可以分析 A 算法和 B 算法在处理用户行为数据时的表现情况，看看哪个算法能够更好地根据用户行为数据进行推荐。
5. 其他因素分析：在上述步骤中，如果还没有找到新算法较差的原因，则可能需要考虑其他因素。例如，可能是 新算法的实现存在bug问题 ，或者是 新算法的调参过程存在问题 。 这时，需要进一步分析新算法的实现情况和调参过程，并尝试纠正可能存在的问题。
6. 总结并形成建议：在完成上述步骤后，我会总结所有可能影响新算法表现的因素，并根据分析结果形成建议。 例如，可以提出修改新算法的工作原理、调整 AB-test 的设置、优化数据的质量等建议，以期望在下一次 AB-test 中取得更好的结果。

6.2 在AB实验中发现关注的核心指标有一个显著的提升（显著正向），那么这个优化一定能上线吗?

不一定

> 从性能上看，一方面的优化可能会导致另一方面的劣化，要多综合评估所有方面的一些指标变动，同时对收益和损失进行评估，再来确认优化是否能上线

7. 其他

7.1 实验有效性

对内部有效性的威胁：实验还没有推广到其他人群或时间段情况下的实验结果的正确性。

1. 违法了个体处理稳定性假设
2. 幸存者偏差：只分析了特定的用户
3. 意向性分析
4. 样本比率不匹配(Sample ratio mismatch, SRM)

对外部有效性的威胁：对照试验可以沿不同维度，如人群、时间的推广程度

基于时间的外部有效性威胁

1. 初始效应： 新功能的适应需要时间
2. 新奇效应：无法持续的效应，新功能可能吸引用户尝试，但是若功能无用，用户重复使用次数会减少

7.2 关于线上对照实验的几个关键主题：

- 一个想法的价值很难被预估。 在这个案例中，一个价值超过每年1亿美金的简单的产品改动被耽搁了好几个月。
- 小改动可以有大影响。 一个工程师几天的工作就能带来每年1亿美金的回报。当然这样极端的投资回报率（return-on-investment, ROI）也很罕见。
- 有很大影响的实验是少见的。 必应每年运行上万个实验，但这种小改动实现大增长的案例几年才出一个。
- 运行实验的启动成本要低。 必应的工程师可以使用微软的实验平台ExP，来便利地科学评估产品改动。
- 综合评估标准（overall evaluation criterion，OEC）必须清晰。 在这个案例中，营收是OEC的一个关键组成，但仅营收本身不足以成为一个OEC。以营收为唯一指标可能导致网站满是广告而伤害用户体验。必应使用的OEC权衡了营收指标和用户体验指标，包括人均会话数（用户是否放弃使用或者活跃度增加）和其他一些成分。关键宗旨是即使营收大幅增长，用户体验指标也不能显著下降。
- 两组实验执行上的时间要具有一致性
- 两组数据在各个维度上特征分布的一致性
- 实验分组流量分配上药均匀
- 实验周期中也要避免外部因素的影响，尽量在平稳时期进行，减少外部因素的干扰；有时候为了保证实验效果的置信，防止小流量分布不均匀，可以在试验过程中， 逐步增大流量分配 ，同时监控关键指标的数据走势，从而得到置信的结论；

7.3 AB实验的开设流程

AB实验经常运用在活动策略是否有效的问题上，进行实验的步骤是：实验的流程：

确定目标和假设->确定指标->确定实验单位->计算样本量->实施测试->分析实验结果

AB实验的流程:

1. 与相关PM沟通确定验证点：确定实验所要验证的功能、改动点在哪里，确定目标和假设
2. 分析师确认在实验中需要观测的一些核心指标
3. 确认实验的样本流量：一般为20w → 确定实验单位, 以及计算样本量
4. 实施测试：
    - 发邮件和相关PM以及开发同学确认可以开启实验，确认实验配置
    - 在发版正式实验前，一般会通过小流量开启一段时间的灰度实验，来验证我们的实验并不会造成一些特别极端的影响
    - 正式发版，一般执行1周左右时间
5. 在实验分析平台，整理实验数据，产出实验报告 → 分析实验结果

📣 注意点：

- 其中确定指标中比较关键的是要确定评价指标和护栏指标
    - 评价指标就是驱动公司实现核心价值的指标，要具有可归因性、可测量性、敏感性和稳定性；
    - 护栏指标也就是辅助指标
- 确定实验单位有从用户层面、访问层面和页面层面进行考虑的情况:
    - 用户层面适用于易被用户察觉的变化实验，访问和页面层面适用于不易被用户察觉的变化实验；
    - 从用户层面到页面层面实验粒度越来越细，累计的样本量也越来越多
- 计算样本量，需要预先确认以下数值：显著性水平、功效、实验组和对照组的综合方差以及期望的最小差值
    - 实验组和对照组数据量最好均分，非均分的时候只有相对较小的组达到最小样本量，实验结果才可能显著
    - 并不是说实验组越大越好，因为 瓶颈是在样本量较小的对照组上 ，所以实验组和对照组的样本量最好相同 (避免出现辛普森悖论？)
- 分析测试结果的时候要注意辛普森悖论等问题，而且要保证样本达到足够的量、检验是否在正常的波动范围内

7.4 若每次改动都要进行一次AB实验测试，这样成本不会变高吗?

要考虑成本投入的问题。

1. 如果是验证一个小按钮或小改动，我们可以在界面上设置一个开关，用户可以通过开关的形式自行决定采用哪种方式，最终我们可以通过开关的相关指标去判断用户对哪种形式有更大的倾向性
2. 或者可进行一些用户调研，如访谈或问卷形式来收集

AB测试缺陷:

1. 无法测试新的体验
2. 无法告诉你是否遗漏内容
3. 无法对时间跨度较长的实验

在测试新的体验时，AB测试的用途就没那么大了。

还有一些时间跨度要求比较大的实验，也不适合进行AB测试。例如酒店订购app的推荐，大家住酒店的频率不会像是日常活动，可能这次订了酒店，即使用户将该APP推荐给其他人，他人用到这个APP的时间可能要过去一段时间。

AB测试无法确切告诉你，你是否遗漏了什么东西。

7.5 谈谈对AB测试的理解

AB测试本质上是假设检验。

而在AB测试中，在 A/B 测试的语境下，零假设指的是实验组和对照组的指标是相同的，备择假设指的是实验组和对照组的指标是不同的。

原假设：两个组的指标没有差异，备择假设：两个组的指标存在显著差异。

AB测试有助于快速迭代产品功能，促进业务的持续增长。