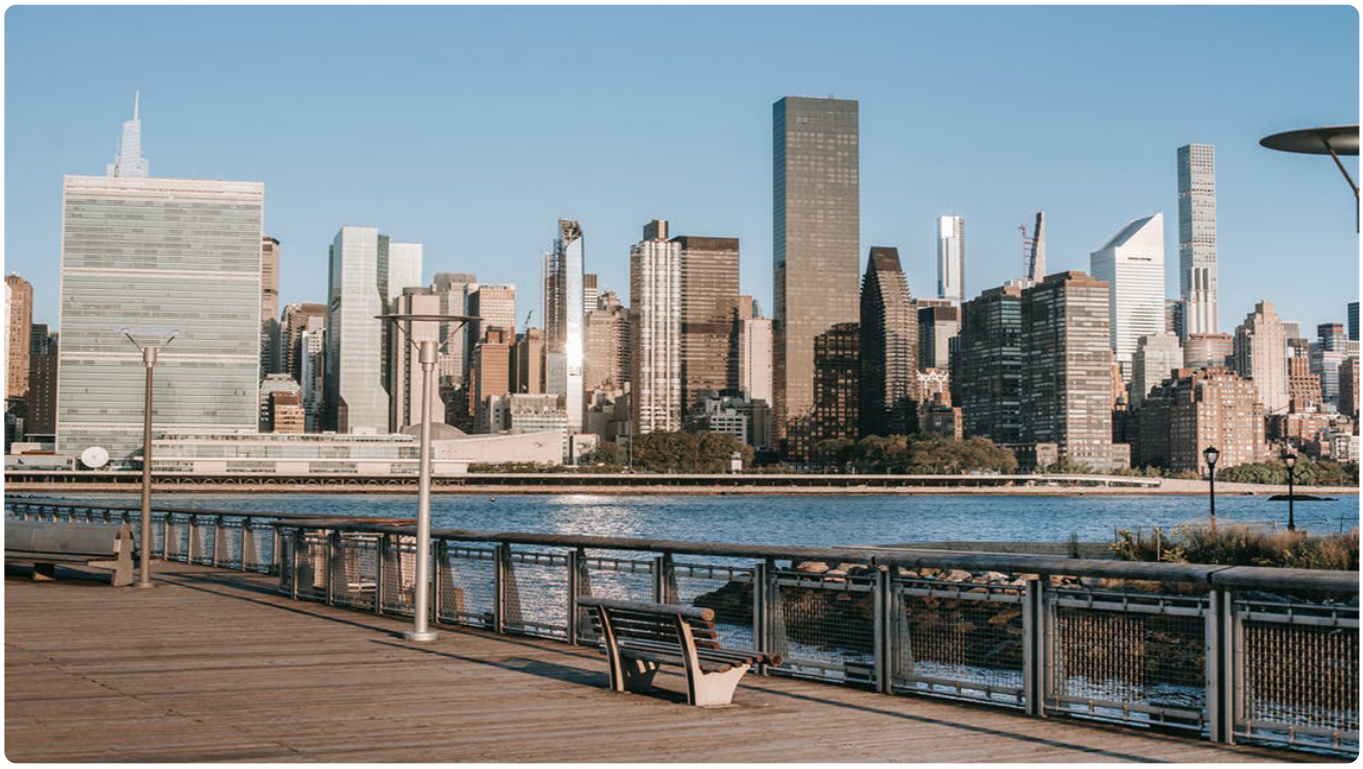
14｜举一反三：A/B测试面试必知必会（下）

2021-01-07 张博伟

A/B测试从0到1

进入课程 >



讲述：张博伟

时长 15:05　大小 13.82M

你好，我是博伟。

今天这节面试课，在学习的过程中你会发现考察的知识点都已经掌握得差不多了。不过我想要强调的是，知识是你业务精进的基础，也是面试时考察的一个重要方面。但更为关键的，是你能够把知识举一反三，知道在不同的场景中如何应用，这也正是把知识转化为解决问题的能力，是你的面试竞争力。

好了，那我们就趁热打铁，继续来讲面试这个主题，帮你夯实基础，做到面试不慌！

面试应用一

假设你现在负责跑一个 A/B 测试，根据样本量计算测试需要跑 2 周。但是业务上的同事会每天关注测试结果，一周之后就观察到显著结果了，这时候他觉得既然结果已经显著了，就想让你停止测试，然后实施测试中的变化。由于业务上的同事对统计不是很熟悉，所以你该怎么用直白的语言来给他解释现在还不能停止实验呢？

考点：

1. 多重检验问题
2. 统计原理的通俗解释

解题思路：

其实这道题，我在第 7 节讲分析测试结果时就给出了一个类似的实践背景，由于在样本量还没有达到规定前不断查看结果，就会造成多重检验问题。而一旦出现多重检验，我们之前花费的功夫就会功亏一篑。所以在这道题中，你需要首先指出多重检验问题，接着说明出现的原因，以及可能造成的具体后果（得到假阳性的概率增大，实验结果不准确）。

你也能看出来，如果只是考多重检验问题，那就太简单了。斟酌一下题目中的问题，就能知道面试官想要考察的是其实是你的表达与理解能力，也就是说你该怎么用**通俗直白的语言**来给业务同事解释复杂难懂的统计概念和原理。在实际工作中，很多时候需要和没有统计背景的同事去沟通交流 A/B 测试的相关内容，所以面试官也非常喜欢考察面试者这方面的能力。

不仅如此，这其实也是在变相考察面试者是不是真正内化了相关统计知识。毕竟如果只是死记硬背概念，肯定是不能在实践中灵活运用这些原理的，更别说再把这些原理用直白的语言去讲给没有统计背景的人听。

你可能会问，通俗直白的语言到底是什么呢？其实也很简单，就是说人话。我的经验就是“一个避免，两个多用”。

一个避免，指的是尽量避免使用统计术语（P 值、第一类错误、假设检验等）。

一方面，专业统计术语会加大你们沟通的时间成本和沟通障碍。业务同事是不懂这些术语的，当你用专业术语去向他说明时，你就需要花更多的时间来解释术语，不仅对方难以理

解，而且你们的沟通目的也没有达到。

另一方面，仔细想想，你为什么要去给业务同事解释呢？主要就是为了告诉对方，现在还不能停止实验。所以啊，说清楚为什么不能在此时停止实验就可以了，术语能少则少。

两个多用指的是多打比方、多举例子。尤其是通过日常生活中的事物来打比方、举例子，这会是非常好的一种方式。

比如，A/B 测试其实是比较两组的表现，既然有比较，那就有好坏输赢的概念。那**你就可以选择生活中任何有好坏输赢结果，但是每次发生结果都有可能不同的事件来打比方。**

我比较喜欢拿体育比赛来打比方，比如篮球，就和 A/B 测试非常类似。每场 NBA 篮球比赛都会有事先规定的时间：48 分钟。而且篮球比赛的结果是以比赛结束后的最终结果为准。如果在比赛结束前的任何时间查看比分，任何一方都有可能领先，但是我们并不会以比赛中间的结果作为最终结果。

同理，回到 A/B 测试当中来，如果我们还没有到达规定的时间，看到显著的结果就宣布实验已经完成，从而停止实验，这就和在比赛中看到一方领先就宣布领先的一方获胜、比赛结束是一样的道理。

再回到我们的面试场景中，多重检验问题在工作中其实是很常见的。尤其是在业务上的同事没有很强的统计背景的情况下，可能只是依靠 P 值来做决定，不会考虑样本量是否充足这个前提，所以用通俗的语言来解释这些统计原理尤为重要。

面试应用二

某产品现在想改变商标，所以想衡量新商标对业务的影响，该如何做？

考点：

A/B 测试的适用范围及替代方法

解题思路：

如果你对第 12 节课讲“什么情况下不适合用 A/B 测试”的知识足够熟悉，就知道这里并不能用 A/B 测试来衡量商标改变的影响性。我讲过，“当有重大事件发布时”，是不适合去做 A/B 测试的，商标即是其中之一。毕竟商标代表了产品和公司的形象，如果一个产品有多个商标同时在市场流通，就会给用户带来困扰，从而会对产品形象有不利的影响。

还记得 A/B 测试的两种替代方法吗？分别是非实验的因果推断方法和用户研究。不过啊，在这个情境下非实验的因果推断方法也行不通，因为这个商标是全新的，并没有历史的相关数据。所以用户研究就是我们最终选定的方法。

在这个案例中，我们只需要收集用户对新商标的看法如何，所以就需要的样本尽可能大一些，这样意见才有代表性，但是并不会涉及到用户体验等很有深度的问题。那么我们就可以选用调查问卷的方式来收集用户反馈，从而给我们一些方向性的指导。让我们知道相较于现有的商标，用户对新商标偏正面反馈，还是更偏负面反馈。

如果从调查问卷中得到总体正面的反馈后，团队决定在市场中废除现有商标，推出新商标。这时候就可以来衡量更换商标后的影响，相对于比较推出新商标前后产品的北极星指标的变化来计算出差值去推断出新商标影响，一个更加准确的方法是建立模型。

我们可以用历史数据建立起对北极星指标的时间序列模型，用推出新商标前的数据去训练这个模型，它也可以预测出没有新商标的北极星指标的走势，然后我们可以把模型的预测数据和推出新商标后的实际数据进行比较，从两者的差值来推断出新商标的影响。

总结一下，这道题的答题思路即：说明题中场景下 A/B 测试不适用及其原因，然后再给出用户研究和模型的办法来作为替代解决方法。

面试应用三

某社交网站准备给用户推荐好友，在首页的右上角推出“你可能认识的人”这个新功能，怎么设计 A/B 测试才能真正衡量这个功能底层的推荐算法的效果呢？假设这里没有网络效应。

考点：

A/B 测试分组设计

解题思路：

当拿到题目一看到社交网站，你会立马想到网络效应，但是读完题发现这里假设没有网络效应。

你可能会想，想要推出一个新功能，而且还不考虑网络效应，那肯定就是常规的 A/B 测试设计了呗。所以就把用户随机均分成两组，对照组的用户没有“你可能认识的人”这个新功能，实验组的用户有这个新功能。最后比较两组的指标，来确定推荐新功能的推荐算法的效果如何。

你看，这没有什么难的！如果真的这么想，那你就在不知不觉中掉进面试官给你设的坑了。

我们再仔细读题中的场景描述，就会发现这个新功能是在页面的右上角，这意味着增加这个新功能还涉及到用户交互界面的改变。

如果按照我们刚才所说的实验分组进行设计，把实验组和对照组相比，其实是既增加了推荐算法，又改变了交互界面，是同时改变了两个因素。以此来看，即使实验组的指标相对于对照组有所提升，我们也无法确定究竟是哪个因素在起作用。

所以这道题的关键点就是如何分离这两个潜在的影响因素。在实践中，解决的方法一般是设计多个实验组，每个实验组只改变一个因素，同时共用一个对照组，也就是改变前的状态。

是不是觉得这个方法有点熟悉呢？没错儿，这就是我在第 9 节课中提到的 A/B/n 测试。不过这个案例的情况比较特殊，因为要增加推荐算法的话，肯定会改变交互界面，也就是说其中一个因素必须依赖另一个因素，不能单独存在。

但是如果反过来想，其实改变交互界面并不一定要增加推荐算法，所以我们可以把各个分组设计成递进关系：

对照组：改变前的原始版本。

实验组 A: 增加“你可能认识的人”这个新功能， 其中推荐的内容随机产生。

实验组 B: 增加“你可能认识的人”这个新功能， 其中推荐的内容由推荐算法产生。

我们可以发现，实验组 A 相对于对照组只是改变了交互界面，因为它的推荐内容是随机产生的。而实验组 B 相对于实验组 A，则是只增加了推荐算法，而二者的交互界面是相同的。这样我们就可以通过比较对照组和实验组 A 来衡量改变交互界面是否有影响，比较实验组 A 和实验组 B 来判断新功能的底层推荐算法是否有效果。

面试应用四

某社交平台开发出了一个新的交互界面，希望能增加用户的点赞次数。团队通过把一部分用户随机分组进行 A/B 测试，发现用了新界面的实验组的用户平均点赞次数，比对照组高出了 5%，结果也是显著的。那么如果把新界面推广给所有用户，你认为用户的平均点赞次数会提升多少呢？是大于 5% 还是小于 5%？为什么呢？在这个案例中，我们假设没有学习效应的影响。

考点：

网络效应

解题思路：

看到“社交平台”就要想到“网络效应”，经过前面的学习，你应该对这一点形成肌肉记忆。

这道题其实难度不大，考察的是网络效应及其形成原因。不过我想通过这道题，一方面让你清楚网络效应的具体场景，另一方面，也想让你知道在有网络效应的影响下，社交平台开发新交互界面后的真实提升效果和实验结果之间的关系。

在没有学习效应的情况下，因为是社交平台，存在网络效应，所以随机分组并不能保证实验组和对照组的独立性，意味着两组的独立性被破坏了。

具体而言，即：如果实验组的用户 A 因为用了的新界面点赞了一个内容，那么这个被点赞的内容也会被 A 的好友，在对照组的 B 看到，B 也有可能点赞这个内容。所以这个新界面改动既影响了实验组，还会通过网络效应影响对照组，即实验组的用户平均点赞次数提升，对照组的也会提升。

以此来看，这里的 5% 的提升其实是受到网络效应影响后的结果，真实的提升效果应该会更大（即只有实验组的指标提升而对照组的指标不变），即大于 5%。

所以当我们把这个新的交互界面推广到所有用户，也就是在没有对照组的情况下，那么和旧版本相比，真实的提升效果应该是大于 5% 的。

小结

我们两节课的 A/B 测试的面试之旅，到这里也就告一段落了。你应该也能发现，这些常见的考点我们在前面的课程中都有讲解过，只要你认真学习了专栏的内容，是不会有太大的问题的。

在最后呢，我还想强调一点。我们在这两节课讲的面试题大都是题目中直接提到 A/B 测试的，在面试中，A/B 测试的考查形式是多种多样的。有时题目中并没有明确提到 A/B 测试，但是 A/B 测试是这些题目中答案的有机组成部分，比如让你衡量产品新功能的好坏，是不是应该推进这个产品变化这种问题，你的答案中肯定会有要如何定义目标和指标去表征新功能的影响，如何设计 A/B 测试去验证新功能是否有效。总之，只要让你进行因果推断，需要量化改变带来的影响时，A/B 测试都是你的好帮手！

思考题

这里呢我们开动脑筋，如果让你用直白通俗的语言（不用统计上的定义，不引用其他术语）解释 A/B 测试的相关术语的话，你会怎么解释呢？选取一两个尝试着解释下。

欢迎把你的解释分享在评论区，我们一起交流、讨论。同时如果你有所收获，也欢迎你把这节面试课分享给你有需要的朋友。

提建议

上一篇 13 | 融会贯通：A/B测试面试必知必会（上）

下一篇 15 | 用R/Shiny，教你制作一个样本量计算器

精选留言 (2)

写留言

qinsi
2021-01-07

辛普森悖论：去年新房成交5万套，均价3万每平方米，二手房成交1万套，均价1万每平方米；今年新房成交1万套，均价5万每平方米，二手房成交5万套，均价2万每平方米；自媒体：《今年房价比去年下跌6%》

作者回复: 标题党!

那时刻
2021-01-07

请问老师，用历史数据建立起对北极星指标的时间序列模型，建立时间序列模型，有木有推荐的python或者R的库呢?

展开

作者回复: 比较推荐Python的statsmodels:
https://www.statsmodels.org/stable/tsa.html
和prophet:
https://facebook.github.io/prophet/