

特殊问题

1. 辛普森悖论

辛普森悖论出现的次数可能会很多，每次做AB测试结果的细分分析时，最好都要先检查下细分领域在两组的比例是否符合两组整体的比例，以确保实验结果的准确性。

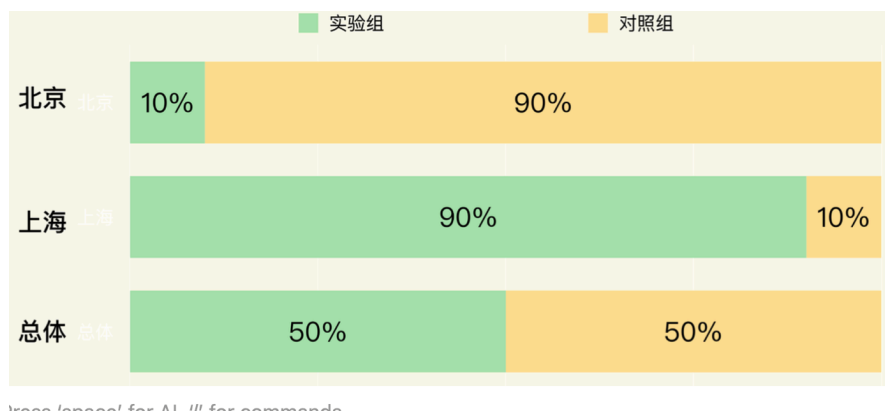
	北京 <small>(1M用户，对照组0.9M，实验组0.1M，对照/实验组=90%/10%)</small>	上海 <small>(1M用户，对照组0.1M，实验组0.9M，对照/实验组=10%/90%)</small>	总体 <small>(2M用户，对照组1M，实验组1M，对照/实验组=50%/50%)</small>
对照组 <small>(旧注册流程，1M用户)</small>	19710/900000=2.13%	1010/100000=1.01%	(19710+1010)/1000000=2.02%
实验组 <small>(新注册流程，1M用户)</small>	2560/100000=2.56%	11790/900000=1.31%	(2560+11790)/1000000=1.44%

辛普森悖论：指当多组数据内部组成分布不均匀时，从总体上比较多组数据和分别在各个细分领域中比较多组数据可能会得出相反的结论。

从数学的语言来看，可以解释为： $\frac{a}{b} < \frac{A}{B}, \frac{c}{d} < \frac{C}{D}$, 也可能会出现 $\frac{a+c}{b+d} > \frac{A+C}{B+D}$ 。

原因：存在我们尚未观测或不知道的潜在变量在对实验也有影响，以及多组数据中各个细分领域的分布不均匀。辛普森悖论其实是理论上无法避免的，因为我们永远不知道哪些维度/特征也在发挥作用，而这些维度或许没被觉察到，或许没有被数据采集到。

在这个例子当中，其实是因为实验组和对照组虽然在总体上实现了我们在设计实验时要求的样本量均分。但是在北京和上海这两个细分市场中却分布不均匀，没有实现样本量均分。



为了避免辛普森悖论，除了要保证总体多组样本量要均匀，在各个细分领域上的多组样本分布也要均匀。

1.1 解决方法

1. 在分析测试结果前做好 **合理性检验**，那出现辛普森悖论的几率就会大大减小
2. 如果我们在进行总体分析和细分分析时发现了辛普森悖论，最好的解决办法就是重新跑实验，看看两组在不同细分领域的分布不均会不会消失。

如果分布不均的情况还是没有消失，那就说明这很可能不是偶然事件。需要检查是否是工程或者实验实施层面出现问题，并针对性地解决

3. 如果时间紧迫，没有时间重跑实验和检查问题原因，则就以细分领域的结果为准，因为总体结果出现了辛普森悖论会变得不准确。

但如果比较多个细分领域的结果，也可能又造成多重检验的问题

4. 因为维度根据分法不同可能有无穷多个，实践中我们能做的是 **重点关注对我们有意义的维度，尽量减少它的影响**。

1.2 启示

1.2.1 相关性、因果性

统计只告诉你相关性，而我们真正有兴趣的因果性，却必须依靠其他的考虑来确定。大到人类社会，小到产品功能，被研究的内容有着太多的维度，到底哪一个是因，哪一个果不能简单论断。即使将能想到的维度按相关性强弱一一考虑，也只是减小出错的概率，更不能只挑选易得的、自己感兴趣的变量任意拿来研究并得出结论。

一个典型的社会学例子是疫情初始时，《经济学人》用新冠死亡率和“民主程度”统计，发现有正向相关性，然后断言“民主”有利于防疫，当然按照现在的情况，他们怕是不会再做一次“统计”。影响防疫的因素有太多，政治优先的《经济学人》只选择了自己感兴趣的那个维度，然后将其相关联，得到完全没有根据的结论。而在一个具体的产品中，普适型的数据（如粗暴的对比IOS和Android总体情况）也是没有多大参考意义的，最起码需要将用户、设备、场景等细分足够清楚去看，才能得到接近事实的结论。

1.2.2 分层分类，定性定量

分层分类是将事实有关的内容区分的足够清楚明白，找到足够多的相关因素；而定性定量则是分层分类不可分割的下一步骤，数量与性质（权重）是不对等的，但是往往数量比性质更容易获得；因此在得到分层分类的数量数据后，需要斟酌个别分组的权重，以一定的系数去消除以分组资料基数差异所造成的影响，同时必须了解该情境是否存在其他潜在因素而综合考虑。

1.2.3 AB测试

工作中经常遇到的分组数据寻求结论场景就是AB测试，很多时候我们用不到5%的用户进行小规模的测试，发现效果很好，就直接一步跨到全量，最终效果可能并不如预期。我们可以通过三个方法避免这种情况的发生：

1. 首先是设计用户组时，尽量保证用户特征一致，并能代表产品的目标用户或者核心用户；
2. 其次当我们觉得某两个变量对试验结果都有影响时，需要将这两个变量放在同一层进行互斥试验，不要让一个变量的试验动态影响另一个变量的检验；
3. 最后分析结论时除了看大组的效果，也要针对一些细分用户进行效果回收，保证结果的可靠性。

- 准确的用户分层在数据分析中是非常重要的，尤其是在免费产品当中，平均用户不仅不存在，而且是误导研发的因素之一，所以关键在于利用特征将用户进行合理划分。
- 在一个具体的产品中，普适型的数据（如粗暴的对比IOS和Android总体情况）是没有多大参考意义的，一定要细分到具体设备、国家、获取渠道、消费能力等等再进行比对才有价值。
- 斟酌个别分组的权重，以一定的系数去消除以分组资料基数差异所造成的影响，同时必需了解该情境是否存在其他潜在要因而综合考虑。

2. 多重假设检验问题

多重检验问题：当存在多个假设检验且选择了其中最低的P值作为结果，这样会对P值和效应大小的估算出现偏差。

多重比较问题常出现：

1. 查看多个指标
2. 查看跨时间的P值
3. 查看受众细分群
4. 查看实验的多次迭代

多重检验问题，又叫多重测试问题或多重比较问题(Multiple Comparison Problem)，指的是当同时比较多个检验时，第一类错误率 α 就会增大，而结果的准确性就会受到影响这个问题。

2.1 多重检验为什么会是一个问题

要搞清楚多重检验为什么会是一个问题，我们还得先从第一类错误率 α (又叫假阳性率，显著水平，是测试前的预设值，一般为 5%)说起。第一类错误率指的就是当事实上两组指标是相同的时候，假设检验推断出两组指标不同的概率，或者说由于偶然得到显著结果的概率。而且，它在统计上的约定俗成是 5%。

5% 看上去是个小概率事件，但是如果我们同时比较 20 个检验(测试)呢?你可以先思考一下，如果每个检验出现第一类错误的概率是 5%，那么在这 20 个检验中至少出现一个第一类错误的概率是多少呢?

要直接求出这个事件的概率不太容易，我们可以先求出这个事件发生情况的反面，也就是在这 20 个检验中完全没有出现第一类错误的概率，然后再用 100% 减去这个反面事件的概率。

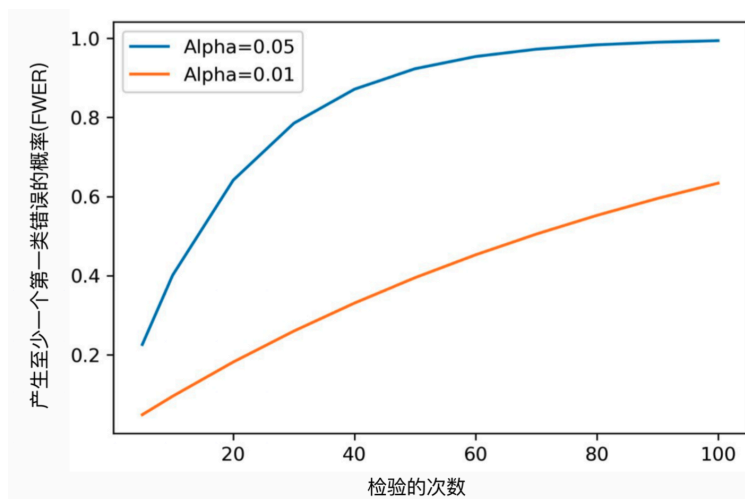
这里我们用 $P(A)$ 来表示出现事件 A 的概率。 $P(\text{每个检验出现第一类错误})=5\%$ ，那么 $P(\text{每个检验不出现第一类错误}) = (1-5\%)=95\%$ ，所以 $P(20 \text{ 个检验中完全没有第一类错误})= 95\%$ 的 20 次方。

这样我们就可以求得这个概率：

$$\begin{aligned} P(\text{至少出现一个第一类错误}) &= 1 - P(20 \text{ 个检验中完全没有第一类错误}) \\ &= 1 - (1 - 0.05)^{20} \\ &= 64\% \end{aligned}$$

这里的 $P(\text{至少出现一个第一类错误})$ 的概率又叫做 **FWER** (Family-wise Error Rate)。

通过计算得出来的概率是 64%。这就意味着当同时比较 20 个检验时，在这 20 个结果中，至少出现一个第一类错误的概率是 64%。看看，这是不是个很大的概率了呢?事实上，随着检验次数的增加，这个概率会越来越大。



根据这个图我们可以得出两个结论：

1. 随着检验次数的增加，FWER，也就是出现第一类错误的概率会显著升高。
2. 当 α 越小时，FWER 会越小，上升的速度也越慢。

第一个结论讲的就是多重检验带来的问题。第二个结论其实为我们提供了一种潜在的解决方法：降低 α 。

当我们同时比较多个检验时，就增加了得到第一类错误的概率(FWER)，这就变成了一个潜在的多重检验问题。

2.2 什么时候会遇到多重检验问题

2.2.1 第一种形式，当 A/B 测试有不止一个实验组时

当我们想要改变不止一个变量且样本量充足时，我们可以不必等测试完一个变量后再去测试下一个，而是可以同时测试这些变量，把它们分在不同的实验组当中。

每个实验组只变化一个变量，在分析结果时分别用每个实验组和共同的对照组进行比较，这种测试方法也叫做 **A/B/n 测试**。

比如我想要改变广告来提升其效果，那么想要改变的变量包括内容、背景颜色、字体大小等等，这个时候我就要有相对应的 3 个实验组，然后把它们分别和对照组进行比较。这就相当于同时进行了 3 个检验，就会出现多重检验问题。

2.2.2 第二种形式，当 A/B 测试有不止一个评价指标时

这个很好理解，因为我们分析测试结果，其实就是比较实验组和对照组的评价指标。如果有多个评价指标的话，就会进行多次检验，产生多重检验问题。

2.2.3 第三种形式，当你在分析 A/B 测试结果，按照不同的维度去做细分分析时

当我们分析测试结果时，根据业务需求，有时我们并不满足于只把实验组和对照组进行总体比较。比如对于一个跨国公司来说，很多 A/B 测试会在全球多个国家同时进行，这时候如果我们 想要看 A/B 测试中的变化对于各个国家的具体影响时，就会以国家为维度来做细分的分析，会分别比较单个国家中的两组指标大小，那么此时分析每个国家的测试结果就是一个检验，多个国家则是多个检验。

2.2.4 第四种形式，当 A/B 测试在进行过程中，你不断去查看实验结果时

因为当测试还在进行中，所以每次查看的测试都和上一次的不一樣，每查看一次结果都算是一次检验，这样也会产生多重检验问题。

不要在 A/B 测试还在进行时就过早地去查看结果，一定要等样本量达到要求后再去计算结果

2.3 如何解决多重检验问题

四种形式的多重检验问题的解决方案：

鉴于多重检验问题的普遍性，在统计上有很多学者提出了自己的解决方法，大致分为两类：

1. 保持每个检验的 P 值不变，调整 α 。
2. 保持 α 不变，调整每个检验的 P 值。

用 P 值来判断假设检验的结果是否显著时，是用检验中计算出的 P 值和 α 进行比较的。当 $P < \alpha$ 时，我们才说结果显著。所以，要么调整 α ，要么调整 P 值。

- 降低 α 是一种解决办法，最常用的调整 α 的方法是 **Bonferroni 校正** (Bonferroni Correction)，其实很简单，就是把 α 变成 α/n 。
 - Bonferroni 校正由于操作简单，在 A/B 测试的实践中十分流行，但是这种方法只是调整了 α ，对于不同的 P 值都采取了一刀切的办法，所以显得有些保守，检测次数较少时还可以适用。
- 根据实践经验，在检测次数较大时(比如上百次，这种情况在 A/B 测试中出现的情况一般是做不同维度的细分分析时，比如对于跨国公司来说，有时会有上百个 markets)，Bonferroni 校正会显著增加第二类错误率 β ，这时候一个比较好的解决办法就是去调整 P 值，常用的方法就是通过 **控制 FDR(False Discovery Rate)** 来实现。
 - 最常用的是 **BH 法(Benjamini-Hochberg Procedure)**。
 - BH 法会考虑到每个 P 值的大小，然后做不同程度的调整。
 - 大致的调整方法就是把各个检验计算出的 P 值从小到大排序，
 - 然后根据排序来分别调整不同的 α 值，

- 最后再用调整后的 P 值和 α 进行比较。

3. 学习效应


当我们想通过 A/B 测试检验非常明显的变化时，比如改变网站或者产品的交互界面和功能，那些网站或者产品的老客户往往适应了之前的交互界面和功能，而新的交互界面和功能对他们来说需要一段时间来适应和学习。所以往往老用户在学习适应阶段的行为会跟平时有些不同，这就是学习效应。

3.1 学习效应在实践中的形式

老用户的两种学习适应期的反应：

- 第一种是积极的反应，一般也叫做新奇效应(Novelty Effect)，指的是老用户对于变化有很强的好奇心，愿意去尝试。
- 第二种是消极的反应，一般也叫做改变厌恶(Change Aversion)。指的是老用户对于变化比较困惑，甚至产生抵触心理。

3.2 如何检测

- 第一种方法是表征实验组的指标随着时间(以天为单位)的变化情况：在没有学习效应的情况下，实验组的指标随着时间的变化是相对稳定的。但是当有学习效应时，因为学习效应是短期的，长期来看慢慢会消退，那么实验组(有变化的组)的指标就会有一个随着时间慢慢变化的过程，直到稳定。
 - 如果是新奇效应，实验组的指标可能会由刚开始的迅速提升，到随着时间慢慢降低。
 - 如果是改变厌恶，实验组的指标可能会由刚开始的迅速降低，到随着时间慢慢回升。
 -  不需要每天都去比较实验，否则容易出现多重比较问题。只有达到样本量之后才可以去比较两组大小，分析测试结果。
 - 第二种方法是只比较实验组和对照组中的新用户。
 - 学习效应是老用户为了学习适应新的变化产生的，所以对于新用户，也就是在实验期间才第一次登录的用户来说，并不存在“学习适应新的变化”这个问题，那么我们可以先在两组找出新用户(如果是随机分组的话，两组中新用户的比例应该是相似的)，然后只在两组的新用户中分别计算我们的指标，最后再比较这两个指标。
 - 如果我们在新用户的比较中没有得出显著结果(在新用户样本量充足的情况下)，但是在总体的比较中得出了显著结果，那就说明这个变化对于新用户没有影响，但是对于老用户有影响，那么大概率是出现了学习效应。
-

想真正排除学习效应的影响，得到准确的实验结果，还是要延长测试时间，等到实验组的学习效应消退再来比较两组的结果。