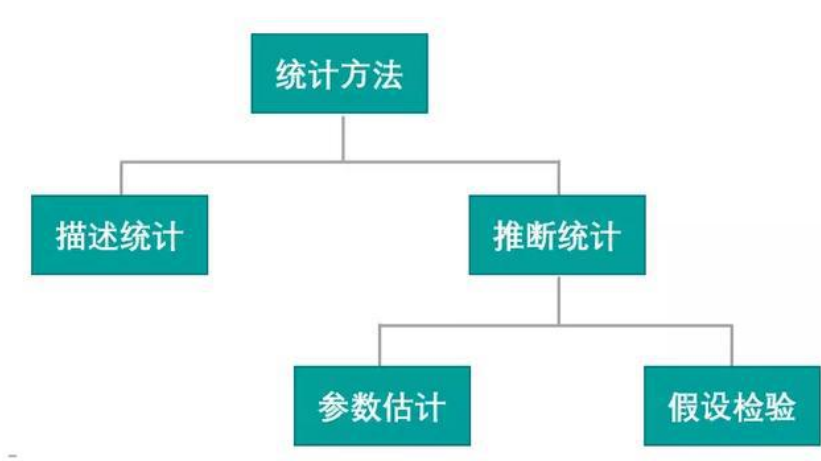


讲完概率分布，再来讲讲统计学的最后一个知识点——假设检验。

假设检验是数理统计学中根据一定假设条件由样本推断总体的一种方法。事先对总体参数或分布形式作出某种假设，然后利用样本信息来判断原假设是否成立，采用逻辑上的反证法，依据统计上的小概率原理。



为了更好的解释，这里举个例子。

假设我有一袋豆子，袋子里有红豆，也有黑豆，我想知道红豆和黑豆是不是一样多。若是一个个去看怕是要疯了。于是偷个懒，从袋子里拿了一把豆子，看看这把红豆多还是黑豆多。用这把豆子作为样本，去推断这袋豆子。既然是用样本推断总体，就有抽样误差的可能性。不管袋子里红豆多还是黑豆多，这一把不一定能真实反映这袋豆子，那怎么办呢？这就要用到假设检验了。

说假设检验之前，先要知道小概率事件。统计大牛觉得如果一件事情发生的可能性小于 0.05，就可以定义为小概率事件了，也就是说，在一次研究中该事件发生的可能性很小，如果只进行一次研究，可以视为不会发生。

回到豆子的话题。现在是想通过样本（一把豆子）去推断总体（一袋豆子）。先做一个假设，一般是我们心里特别不想承认的那一种可能，也称无效假设。和无效假设对立的是备择假设，是无效假设的对立面。

无效假设：袋子里红豆和黑豆是一样多的，如果观察到红豆黑豆不一样多完全是由抽样造成的。备择假设：袋子里红豆和黑豆的确不一样多。

假定袋子里有 100 个豆子，50 个红豆，50 个黑豆。拿的这把豆子有 3 个红豆，7 个黑豆。在无效假设成立的前提下，也就是说红豆黑豆一样多的基础上，能拿到 3 个红豆、7 个黑豆的概率为：

$$\frac{C_{50}^3 \times C_{50}^7}{C_{100}^{10}} = \frac{19600 \times 99884400}{17310309456440} = \frac{1957734240000}{17310309456440} = 0.113$$

这告诉我们，在红豆和黑豆一样多的假设下，拿到 3 个红豆 7 个黑豆的可能性为 0.11，是很常见的，说明所做的假设是可以成立的，还没有理由能拒绝无效假设。

假定袋子里有 100 个豆子，50 个红豆，50 个黑豆。拿的这把豆子有 1 个红豆，9 个黑豆。

在无效假设成立的前提下，能拿到 1 个红豆、9 个黑豆的概率为：

$$\frac{C_{50}^1 \times C_{50}^9}{C_{100}^{10}} = \frac{50 \times 2505433700}{17310309456440} = \frac{125271685000}{17310309456440} = 0.007$$

这告诉我们，在红豆和黑豆一样多的假设下，拿到 1 个红豆 9 个黑豆的可能性为 $0.007 < 0.05$ ，为小概率事件，在一次研究中是不应该发生的，而现在发生了，可能是所做的假设有问题，有理由拒绝无效假设。

简言之，假设检验的核心思想是小概率反证法，在假设的前提下，估算某事件发生的可能性，如果该事件是小概率事件，在一次研究中本来是不可能发生的，现在发生了，这时候就可以推翻之前的假设，

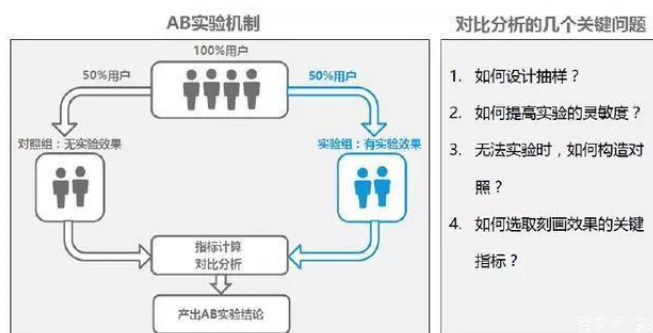
接受备择假设。如果该事件不是小概率事件，我们就找不到理由来推翻之前的假设，实际中可引申为接受所做的无效假设。

假设检验解决那些问题

互联网运营分析师与产品经理都很熟悉的方法：ABtest。适用方面：往往新版本或者新功能上线之前，需要验证新需求的是否有正收益，传统模式往往就是新版本发布，然后观察留存率、跳转率等核心指标，如果观察表现期内，流量正增长，那么则认可该版本，如果效果反馈不佳，则快速回滚旧版本。但在移动互联网时代，千人千面讲究用户差异化的时代，产品经理在面对海量用户流量，就可以实现ABtest，过程：抽样小规模流量，进行灰度测试，检验新需求是否有正收益，如果观察周期校验效果显著，则大规模上线测试迭代，直至全流量覆盖。

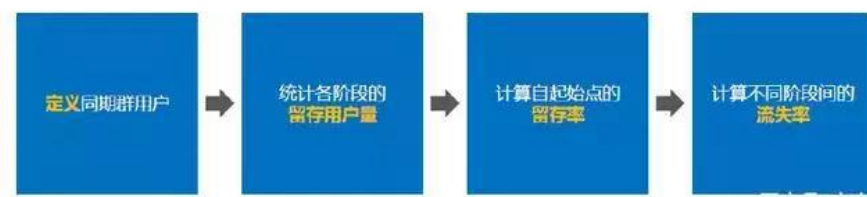
ABtest 其本质核心就是假设检验，但是实际操作中，这里有几个关键：

- 1、如何设计抽样，尽量实现对照实验两边公平；
- 2、如何提高灵敏度，即效果收益稳定性，多组交叉验证流量平衡；
- 3、对照组设置，避免引入干扰因素；
- 4、收益评价指标，正收益、用户体验，当然我们 T 检验也要通过。



实际应用中，我们会发现产品经理的脑洞实在太大，需求太多，能否实现少设置测试流量批次？能否缩短收益评价周期？这里就牵涉到上面第一个问题：如何抽样？抽样才是 ABtest 关键，根本原则是希望控制对照组两边客群一样，尽量公平。抽样方法也很多：简单抽样、分层抽样、正交抽样等等。这里为何提到一个正交抽样呢？正交抽样满足一次测试，多组对照的抽样方法。关于细节，大家可以关注并评论，后面我给相应解答。

实际上用户增长运营是件很有挑战工作，单纯数据流量 ABtest 是用户体验辨别方法，如果从运营监控角度去分析，也需要从不同群组去分析客群增长模型，学会从业务中抓取核心关键公式或者模型，然后细化至各个维度，交叉对比，隐藏冰山下问题。



假设检验实操

假设检验常用的方法有 u—检验法、t 检验法、 χ^2 检验法(卡方检验)、F—检验法，秩和检验等。以 t-假设检验举例。

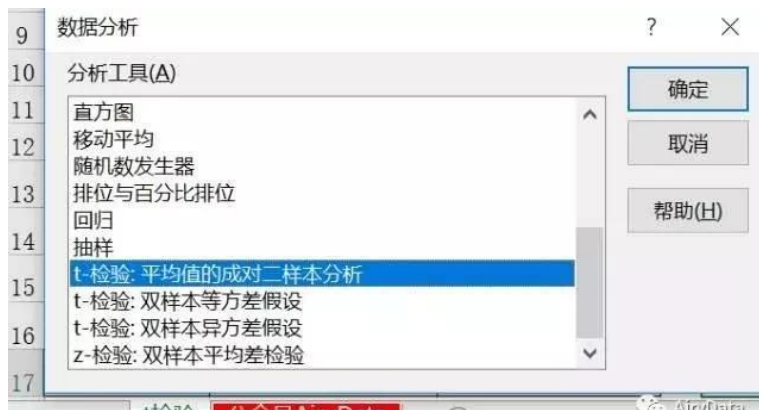
某药厂研发了一种能够降低血压的新药，现在为了了解该药的疗效，随机抽取了 15 名高血压患者，并得到他么在使用该药治疗前后的舒张压数据，如下图所示，现在需要判断：该药是否有效？如果有效，是否能够让高血压患者的舒张压平均降低 6.5mmHg？

	A	B	C
1	新药治疗前后的舒张压 (mmHg)		
2	病人号	治疗前	治疗后
3	1	101.4	95.9
4	2	110	90.4
5	3	95	88.3
6	4	115.6	116.8
7	5	102.5	96
8	6	113	110
9	7	108	100
10	8	102	90
11	9	100	95
12	10	108	88.5
13	11	120	115
14	12	118	118.6
15	13	98	92.5
16	14	99	90.5
17	15	102	102.5

根据上面的数据 ,我们可以使用 Excel 中的假设检验方法来判断(以前我用手算过 ,好累呀 ,有了 Excel 等工具 ,很简单方便 ,省时省力)。分析工具中的假设检验方法有多种 ,使用不同的方法 ,观察值在检验前后的关系就不同 ,所以需要选择合适的方案。

由于样本量较小 ,且样本值中的观察值存在治疗前后的配对关系 ,所以可先使用 “t-检验 :平均值的成对二样本分析” 方法判断该药的有效性 ,我们首先假设该药无效 (一般先否定 ,然后计算检验否定原假设 (如果正确的话)) , 然后进行假设检验。

第 1 步 : 选择分析工具 (Excel 老朋友了)。 “数据分析” —— “分析工具” —— “t-检验 :平均值的成对二样本分析” , 确定即可。如下图所示 :



第 2 步：设置相关参数。在“t-检验：平均值的成对二样本分析”对话框中，设置“输入”组中“变量 1 的区域”为“\$B\$2:\$B\$17”，“变量 2 的区域”为“\$C\$2:\$C\$17”，选择“标志”复选框，设置“ α ”值为“0.05”，在“输出选项”下单击“输出区域”，设置为“\$E\$2”，最后确定。如下图所示：



第 3 步：设置假设平均差。重复上一步骤，其他参数不变，这次设置“假设平均差”为“6.5”，“输出区域”为“\$I\$2”。这一步，“假设平均差”为期望中的样本均值的差值，如果该值设为 0，即假设样本均值相同。

t-检验: 平均值的成对二样本分析

输入

变量 1 的区域(1): \$B\$3:\$B\$17

变量 2 的区域(2): \$C\$3:\$C\$17

假设平均差(E): 6.5

☒ 标志(L)

$\alpha(A)$: 0.05

输出选项

☒ 输出区域(O): \$I\$2

☐ 新工作表组(P):

☐ 新工作簿(W)

确定

取消

帮助(H)

第 4 步：显示分析结果。做完之后，我们就可以看到 t-检验的结果：H 列左侧为第 2 步中检验该药是否有效的数据结果，H 列右侧为第 3 步中检验该药是否能让舒张压降低 6.5mmHg 的检验结果。

如下图所示：

t-检验: 成对双样本均值分析			t-检验: 成对双样本均值分析		
	治疗前	治疗后		治疗前	治疗后
平均	106.1666667	99.33333333	平均	106.1666667	99.33333333
方差	60.01095238	115.5995238	方差	60.01095238	115.5995238
观测值	15	15	观测值	15	15
泊松相关系数	0.817115539		泊松相关系数	0.817115539	
假设平均差	0		假设平均差	6.5	
df	14		df	14	
t Stat	4.211200001		t Stat	0.20542439	
P(T<=t) 单尾	0.000435683		P(T<=t) 单尾	0.420098765	
t 单尾临界	1.761310136		t 单尾临界	1.761310136	

上面的案例中，由于没有充分的理由判断该药治疗后的总体均数会大于或小于治疗前的舒张压均值，所以在检验过程中，前面的 t-检验我们采用的是双侧检验。

从分析结果看到：H 列左侧的检验结果中， $t_{\text{Stat}}=4.211$ ， $P_{\text{双尾}}=0.00087$ ， $t_{\text{双尾临界}}=2.145$ ，当 $t_{\text{双尾临界}}$ 时，假设成立，而这个案例中，检验结果 $t_{\text{Stat}} > t_{\text{双尾临界}}$ ，说明该结果拒绝原假设，也就是说该药有效，此外，我们还能看到 $P_{\text{双尾}}=0.00087 < \alpha=0.05$ ，这一比较结果也说明该结果拒绝原假设，同样说明该药有效。

由于已经确定该药有效，那么再判断该药能否将舒张压平均值降低 6.5mmHg，所以，后面的 t-检验采用的是单侧检验，这里我们设置了假设平均差，上图中红框内，检验结果 $t_{Stat}=0.205 < t_{\text{单尾临界}}=1.761$ ，说明该假设成立，即该药能够让高血压患者的舒张压平均降低 6.5mmhg。结果 $p_{\text{单尾}}=0.42 > \alpha=0.05$ 也说明了该假设成立。</t 单尾临界=1.761，说明该假设成立，即该药能够让高血压患者的舒张压平均降低 6.5mmhg。结果 $p_{\text{单尾}}=0.42 >$