优达学城数据分析师纳米学位

A/B 测试项目

试验设计

指标选择

不变指标 1：Cookie 的数量，即查看课程概述页面的唯一 cookie 的数量。

不变指标 2：点击次数：即点击“开始免费试用”按钮（在免费试用屏幕触犯前发生）的唯一 cookie 的数量。

不变指标 3：点击概率，即点击“开始免费试学”按钮的唯一 cookie 的数量除以查看课程概述 页的唯一 cookie 的数量所得的比率。

评估指标 1：总转化率，即完成登录并报名参加免费试用的用户 id 的数量除以点击“开始免费试用”按钮的唯一 cookie 的数量所得的结果。

评估指标 2：留存率：即在 14 天期限后仍保持参加（并进行了至少一次支付）的用户 id 的数量除以完成登录的用户 id 的数量。

评估指标 3：净转化率，即在 14 天期限结束后仍然参加（并至少进行了一次支付）的用户 id 的数量除以点击“开始免费试用”按钮的唯一 cookie 的数量所得的结果。

对于 **Cookie 的数量**，因为查看课程概述页面行为发生在本次试验行为之前，不会受试验影响，因此作为不变指标。

对于 **点击次数**，同样的，点击“开始免费使用”按钮的行为也发生在试验行为之前，理论上不会受到试验所作改变的影响，因此同样也应作为不变指标。

对于**点击概率**，由于其是点击次数和 Cookie 数量两个不变指标的比值，因此也可以看作是不变指标。

对于 **总转化率**，本次试验预期的改变就是减少参加免费试用的用户数量，总转化率理论上会产生相应的变化，而且很变化能直接反映出我们试验的效果，因此应作为评估指标。

对于 **留存率**，本次试验的一个重要考察目标是，不会显著减少继续试学和通过的用户数量，因此留存率也是一个重要的应考察的评估指标。

对于 **净转化率**，与留存率类似的，继续学习的用户数量理论上会受到试验的影响而变化，同时该指标可反映出我们关切的重要试验目标，因此应作为评估指标。

对于参与免费试学的**用户 id 的数量**，由于免费试学发生在试验之后，因此它会受到试验的影响。但作为评估指标时，我们无法判断其在对照和实验组中的差异是试验的影响还是由于分配到两组的 cookie 数量不同造成的，因此不适合作为评估指标。

综上所述，本次试验的预期目标反映在评估指标上，应该为 **总转化率** 有显著下降，对应为对时间不够的同学的提前过滤；而 **净转化率** 应没有显著变化，对应为不会在很大程度上减少继续通过免费试学和最终完成课程的学生数量。对于**留存率**，由于预期申请试学的用户数量会有显著下降，而付费人数不会显著变化，因此留存率的预期会有显著上升。

测量标准偏差

计算标准偏差的公式为 $SE = \sqrt{P*(1-P)/N}$，试验样本的 Cookie 数为 5000，对应到评估指标中，对应的 N 分别为 $N_G = 5000*0.08 = 400$，$N_r = 400*0.20625$ =82.5，$N_N = 5000*0.08 = 400$。

总转化率标准偏差 $SE_G = \sqrt{0.20625*(1-0.20625)/400} = 0.0202$

留存率标准偏差 $SE_R = \sqrt{0.53*(1-0.53)/82.5} = 0.0549$

净转化率标准偏差 $SE_N = \sqrt{0.1093125*(1-0.1093125)/400} = 0.0156$

对于**总转化率**和**净转化率**，因为本试验的转移单位是 cookie，因此分析估计和经验变异是类似的。对于**留存率**，由于考察的对象都是用户 id，因此分析变异性不能匹配经验变异性。

规模

样本数量和功效

因为分析估计和经验变异是类似的，Bonferroni 校正就显得过于保守，因此在分析阶段不会使用 Bonferroni 校正。

我们这里借助课程中提供的在线计算器进行浏览量估计计算。α 值为 0.05，β 值为 0.2。

对于**总转化率**，计算器中的 Baseline conversion rate 设置为 20.625%，Minimum Detectable Effect 设置为 1%，得到的样本数量为 25835。这是点击 cookie 的单组的样本数量，要的到页面浏览量，需要 25835/0.08*2 = 645875 。

对于**留存率**，将 Baseline conversion rate 设置为 53%，Minimum Detectable Effect 设置为 1%，得到登录用户 id 样本数量 39115 。页面浏览量为 39115/0.08/0.20625*2=4741212。

对于**净转化率**，将 Baseline conversion rate 设置为 10.93125%，Minimum Detectable Effect 设置为 0.75%，得到的点击 cookie 样本量为 27413。页面浏览量为 27413/0.08*2=685325 。

持续时间和暴光比例

要确定运行时间，需要先按百分比取每日流量中的对应浏览量数目，然后再用总页面浏览量来除。但在第一次运算中发现即使取用 100%的流量，用 4741212/40000=118.5303 天。这个持续时间过于漫长，因此第二次运算选取了总浏览量 685325，75%的日均流量进行试验，计算的结果为 685325/30000 = 22.8，取整为 23 天。这个测试时间长于 14 天的免费试听时间，又不会过于漫长，因此 23 天，75%的流量转入试验，是一个比较合适的选择。

关于试验的风险：

1-首先，我们的试验本身不涉及用户的生理、心理、情感、经济等方面，用户可以说没有面临什么风险；

2-其次，这项试验对帮助用户规划自己的学习计划是有益的；

3-我们的试验不是强迫用户改变学习行为，只是建议，用户仍可以按试验前的方式去进行学习；

4-试验需要的数据都是敏感性很低的网站访问数据，不会对用户的个人隐私和信息安全造成不良影响，试验的改变也不需要对数据库等底层基础进行什么变动。

因此我们的试验即使选择 100%的用户流量去进行，也是可行的。

试验分析

合理性检查

我们将是否转入试验组的期望几率定为 50%，取 95%的置信区间，计算方法为

$0.5 \pm 1.96 * \mathrm{SE}$。

对于 **Cookie 的数量**，对照组的总数为 345543，实验组总数 344660，$\mathrm{SE} = \sqrt{0.5 \times 0.5 /（345543 + 344660）} = 0.0006018$，期望置信区间为 $0.5 \pm 0.0011795288 = [0.4988, 0.5012]$，而实际观察值为 $345543/（345543 + 344660）= 0.5006$，在期望范围内，因此通过合理性检查。

对于**点击次数**，同样的计算方法，对照组总数 28378，实验组总数 28325，$\mathrm{SE} = \sqrt{0.25/(28378 + 28325)} = 0.0021$，期望区间[0.4959,0.5041]，观察值 0.5005，同样处于期望范围内，通过合理性检查。

对于**点击概率**，计算公式为 $\mathrm{SE} = \sqrt{\mathrm{P}_{control} * (1 - P_{control}) / N_{\exp}} = 0.000468$,则期望区间为[0.0812,0.0830]，对比试验组的点击概率 0.0822，处于期望范围内，所以也通过合理性检查。

结果分析

效应大小检验

对于评估指标差异的置信区间，首先要计算出两组的汇总标准误差，计算公式为 $\mathrm{SE}_{pooled} = \sqrt{P_{pooled} * (1 - \mathrm{P}_{pooled}) * (1/N_{control} + 1/N_{\exp})}$，其中 $\mathrm{P}_{pooled} = (X_{control} + X_{\exp}) / (N_{control} + N_{\exp})$。下面来具体计算每个评估指标。

对于**总转化率**，我们取前 23 天的数据，得到 $\mathrm{P}_{\mathrm{pooled}} = (3785 + 3423) / (17293 + 17260) = 0.2086$，进而得到 $\mathrm{SE}_{pooled} = \sqrt{0.2086 * (1 - 0.2086) * (1/17293 + 1/17260)} = 0.0044$。而差异 $\mathrm{d} = X_{\exp} / N_{\exp} - X_{control} / N_{control} = 0.0206$，最终得到的置信区间为 $\mathrm{d} \pm 1.96 * SE_{pooled} = [-0.0292, -0.012]$。这个区间中不包含 0，因此具有统计显著性，而且也不包含显著性边界 0.01,-0.01,因此可以说具有实际显著性。

对于**留存率**，由于实验持续时间不足这里就不做计算。

对于净转化率，用同样的公式我们可以计算得到的置信区间为[-0.0116,0.0019],包含了 0，且包含了显著性边界-0.0075,因此可以说不具有统计显著性和实际显著性。

符号检验

对于每个评估指标，使用每日数据进行符号检验，然后报告符号检验的 p 值以及结果是否具有统计显著性。（这些应是“符号检验”小测试中的答案。）

符号检验借助课程中提供的在线计算器进行。

对于**总转化率**，我们观察到实验数据中有 19 天的总转化率都比对照组的低，因此在“success”栏填 19，在总数中填 23，计算得出的双尾 p-value 为 0.0026，结果小于 α 值 0.05，因此可以说具有统计显著性。

对于**净转化率**，我们观察实验数据未下降的天数为 10，总数仍为 23，计算的结果 p-value 为 0.6776，大于 α 值 0.05，因此不具有统计显著性。

汇总

这里并没有采用 Bonferroni 校正，因为我们对主要评估指标之间是否具有相关性的判断，和实际相关性，会影响到 Bonferroni 校正判断结果的准确性。很大几率我们使用 Bonferroni 校正会得到过于保守的结论。

建议

基于以上试验结果，我认为这项变更能够达到预期目的，即减少了因为没有足够的时间而离开免费试学，并因此受挫的学生数量，同时不会在很大程度上减少继续通过免费试学和最终完成课程的学生数量。但净转化率的置信区间的分布(-0.0116,0.0019)，显示净转化率很有可能会减少，且减少的幅度大于实际显著性边界 0.0075，可以说很大几率对净转化率产生负面影响。因此不建议应用这项变化。

后续试验

关于后续试验，我认为可以为每周学习时间大于 10 小时的学员推荐一些与所选课程相关的其他课程，预期可以在不影响当前课程完成率的前提下，提高其他相关课程的参与率。衡量的指标有当前课程的净转化率（预期不会产生显著变化），推荐课程介绍页面的点击率和净转化率（预期会有显著增长）。不变量为参与当前课和推荐课的用户 Id 的数量。转移单位是登录学员的 user id。