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

列出你将在项目中使用的不变指标和评估指标。（这些应与你在“选择不变指标”和“选择评估指标”小测试中使用的指标一样）

对于每个指标，解释你为什么使用或不使用它作为不变指标或评估指标。此外，说明你期望从评估指标中获得什么样的试验结果。

一、不变指标：

1） Number of cookies:

原因：首先，cookie 数量是本次实验的分组单元；其次，本次实验的变化——用户点击“开始免费试学”按钮后询问用户每周有多少时间投入学习并作出相应的学习建议，该项变化并不会影响首页，所以并不会导致 cookie 的变化。

2） Number of clicks：

原因：点击数量描述的是点击“开始免费试学”按钮的次数，而点击按钮在该项变化开始前就已经存在，点击动作发生在变化开始前，所以变化并不会影响点击数量的改变。

3） Click-through-probalility：

原因：点击概率是由点击数量除以 cookie 数量得来的，在分子和分母不变的情况下，点击概率本身也不会发生改变。

二、评估指标：

1) Gross conversion:

原因：总转化率由完成登录并参加免费试学的用户 id 数量除以点击数量所得，同时，由于实验的前提假设是：这项变化会为学生预先设定明确的期望，从而减少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。所以该项指标中分子的用户 id 数量会受到变化影响而可能会有所减小，而点击数量则不会受到该项变化影响，使得该项指标可以用作评估指标。

期望结果：减小。因为实验假设是：建议学生每周投入 5 小时的学习时间，这项变化会减少因为没有足够的时间而离开免费试学的用户，所以这项变化会导致该指标的分子减小，而同时分母点击数量不变，所以该指标应该受到变化的影响而有所减小。

2) Net conversion：

原因：净转换率指标是由登录并付费的用户 id 数量除以点击数量所得，实验的前提假设是变化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。即实验希望通过免费试学和最终完成课程的用户数不变，这就意味着付费用户的 id 数量应该尽可能地不受变化影响，所以净转化率可以作为评估指标。

期望结果：不变。由于实验希望付费用户 id 数量不变，而同时，变化并不会影响点击数量的变化，即净转换率的分子分母不变，净转换率的期望结果应该不变。

三、无关指标：

1) Number of user-ids：

原因：首先，由于实验假设变化会减少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，即这项改变可能会导致实验前后用户 id 数量的变化，所以其并不适合作为不变指标。其次，由于在实验中，实验组和对照组的 cookie 数量不一定完全相同，也就是说两组中用户 id 数量不同可能是由于实验的影响，也可能是由于两组 cookie 的不同。所以使用用户 id 数量的区别不能够很好的评估试验的效果。在一个比例化的评估度量（总转化率）存在的情况下，我们可以不选择用户 id 数量作为评估指标。

2) Retention：

原因：首先，由于实验的变化会影响用户 id 数量的变化，而留存率由付费用户 id 数量除以用户 id 数量所得，所以留存率也会受到实验的变化的影响，不能用作不变指标。其次，在关于样本大小和持续时间的计算时，由于留存率所需要的样本数量太多，而在本次 A/B 试验持续的时间内，我们无法采集到足够数量的样本来对留存率进行评估，因此无法用作评估指标。

测量标准偏差

列出你的每个评估指标的标准偏差。（这些应是来自“计算标准偏差”小测试中的答案。）

对于每个评估指标，说明你是否认为分析估计与经验变异是类似还是不同（如果不同，在时间允许的情况下将有必要进行经验估计）。简要说明每个情况的理由。

1) Gross conversion：

由实验数据中得出：P = 0.20625

同时，N=5000*0.08=400

$$\text{所以 } SD = \sqrt{\frac{0.20625*(1-0.20625)}{400}} = 0.0202$$

其中，由于总转化率的分析单位是 cookie 数量，分组单位也是 cookie 数量，分析单位=分组单位，所以总转化率的分析估计和经验变异类似。

2) Net conversion：

由实验数据中得出：P=0.1093125

同时， N=5000*0.08=400

$$\text{所以 } SD = \sqrt{\frac{0.1093125*(1-0.1093125)}{400}} = 0.0156$$

其中，由于净转换率的分析单位是 cookie 数量，分组单位也是 cookie 数量，分析单位=分组单位，所以净转换率的分析估计和经验变异类似。

规模

样本数量和功效

说明你是否会在分析阶段使用 Bonferroni 校正，并给出实验正确设计所需的页面浏览量。（这些应是来自“计算页面浏览量”小测试中的答案。）

不使用 Bonferroni 校正。因为作为仅有的 2 个评估指标，总转化率与净转换率的分母相同，即它们是相互关联的，此时如果使用 Bonferroni 校正会让实验结果过于保守。

以下页面浏览量计算是通过在线计算器计算所得

1) Gross conversion：

由实验数据所得：baseline conversion rate=20.625%

Minimum detectable effect=1%

1- β =80%

α = 5%

通过计算得出样本的点击数量所需为：25835

而实验数据中，点击概率为 0.08，所以每一组的页面浏览量所需：25835/0.08=322938

又由于有实验组和对照组两组，所以总的页面浏览量所需：322938*2=645875

2) Net conversion：

由实验数据所得：baseline conversion rate=10.93125%

Minimum detectable effect=0.75%

1- β =80%

α = 5%

通过计算得出样本的点击数量所需为：27413

而实验数据中，点击概率为 0.08，所以每一组的页面浏览量所需：27413/0.08=342663

又由于有实验组和对照组两组，所以总的页面浏览量所需：342663*2=685325

两个评估指标所需的页面浏览量不相同，所以两者中选最大值作为最终结果，即 685325 为实验所需的页面浏览量。

持续时间和暴光比例

说明你会将多少百分比的页面流量转入此试验，以及鉴于此条件，你需要多少天来运行试验。（这些应是来自“选择持续时间和曝光”小测试中的答案。）

说明你选择所转移流量部分的原因。你认为此试验对优达学城来说有多大风险？

曝光比例：75%

实验持续时间：23 天

原因：因为实验的持续时间由曝光比例和样本数量所决定，而曝光比例主要是通过综合评估实验的风险和实验的周期来决定的。

风险评估：1、实验只是询问用户的每周投入学习的时间并作出相应建议，并没有收集用户的敏感数据，也不会对用户的身体、心理、情感、社会和经济方面造成影响，所以风险并没有超过“最低风险”。

2、实验仅收集用户在该网站上学习投入时间等相关数据，并不涉及用户的健康或财务数据，该部分数据并非敏感性数据。

3、由于本实验需要收集用户的 cookie 数据，该数据在一定特殊条件下可以用来识别个人身份，但敏感性极低。

4、由于实验本身只是在网站原有的基础上添加一个子页面，不用更改网站原有的数据库和后台结构，所以不用担心实验会引起数据的丢失或由于后台的失误导致网页奔溃用户无法访问网页等大问题，对网站的风险小。

实验周期：由于实验最少需要转移 50%的流量，为此所需的实验持续时间为 35 天；最多可以转移 100%的流量，此时所需的的实验持续时间仅为 18 天。同时还需要考虑到节假日周末等因素的影响。

所以，在实验整体风险较小的条件下，选择曝光 75%的流量，进行持续 23 天的实验，可以在尽量避免节假日等因素影响的同时尽可能地缩短实验时间方便做出决策。

试验分析

合理性检查

对于每个不变指标，对你在 95%置信区间下期望观察到的值、实际观察的值及指标是否通过合理性检查给出结论。（这些应是来自“合理性检查”小测试中的答案）

对于任何未通过的合理性检查，根据每日数据解释你觉得最有可能的原因。**在所有合理性检查通过前，不要开始其他分析工作。**

由实验数据可得：

1）Number of cookies：

对照组页面总量：345543

实验组页面总量：344660

页面总量：690203

Cookie 分布概率：0.5

$$SE = \sqrt{\frac{0.5*0.5}{345543+344660}} = 0.0006018$$

m = SE * 1.96 = 0.0011796

置信区间=[0.5-0.0011796, 0.5+0.0011796] = [0.4988, 0.5012]

观察值 = 344660 / 690203 = 0.5006

观察值落在置信区间内，所以通过合理性检查。

2） Number of clicks：

对照组总量：28378

实验组总量：28325

总量：56703

Cookie 分布概率：0.5

$$SE = \sqrt{\frac{0.5*0.5}{28378+28325}} = 0.0021$$

m = SE * 1.96 = 0.0041

置信区间=[0.5-0.0041, 0.5+0.0041] = [0.4959, 0.5041]

观察值 = 28378 / 56703 = 0.5005

观察值落在置信区间内，所以通过合理性检查。

3) Click-through-probalility:

对照组点击概率：0.0821258

实验组页面总量：344660

$$SE = \sqrt{\frac{0.0821258 * (1 - 0.0821258)}{344660}} = 0.000468$$

m = SE * 1.96 = 0.00092

置信区间=[0.0821258-0.00092, 0.0821258+0.00092] = [0.0812, 0.0830]

观察值 = 0.0821824

观察值落在置信区间内，所以通过合理性检查。

所有不变指标都已通过合理性检查。

结果分析

效应大小检验

对于每个评估指标，对试验和对照组之间的差异给出 95% 置信区间。说明每个指标是否具有统计和实际显著性。（这些应是来自“效应大小检验”小测试的答案。）

由实验数据所得：

1) Gross conversion：

对照组点击数量 Ncont=17293

实验组点击数量 Nexp=17260

对照组用户 id 数量 Xcont= 3785

实验组用户 id 数量 Xexp= 3423

对照组总转化概率 Pcont=3785/17293= 0.2188

实验组总转化概率 Pexp=3423/17260= 0.1983

合并概率 Ppool=(3785+3423)/(17293+17260)= 0.2086

实验组与对照组总转化概率差异 d=0.1983-0.2188= -0.0205

合并误差:

$$SE_{pool} = \sqrt{\frac{P_{pool} * (1 - P_{pool})}{1/N_{cont} + 1/N_{\exp}}} = \sqrt{\frac{0.2086 * (1 - 0.2086)}{1/17293 + 1/17260}} = 0.0044$$

m=1.96*0.0044= 0.0086

置信区间=[-0.0205-0.0086,-0.0205+0.0086]=[-0.0291, -0.0120]

dmin=0.01

置信区间不包括 0，所以具有统计显著性；

置信区间不包含[-dmin, dmin]，所以具有实际显著性

2) Net conversion：

对照组点击数量 Ncont=17293

实验组点击数量 Nexp=17260

对照组付费用户 id 数量 Xcont= 2033

实验组付费用户 id 数量 Xexp= 1945

对照组净转换率 Pcont=2033/17293= 0.1176
实验组净转换率 Pexp=1945/17260= 0.1127
合并概率 Ppool=(2033+1945)/(17293+17260)= 0.1151
实验组与对照组净转换率差异 d=0.1127-0.1176= -0.0049
合并误差:

$$SE_{pool} = \sqrt{\frac{P_{pool} * (1 - P_{pool})}{1/N_{cont} + 1/N_{\exp}}} = \sqrt{\frac{0.1151 * (1 - 0.1151)}{1/17293 + 1/17260}} = 0.0034$$

m=1.96*0.0034=0.0067
置信区间=[-0.0049-0.0067,-0.0049+0.0067]=[-0.0116, 0.0019]
dmin=0.0075
置信区间包括 0，所以不具有统计显著性；
置信区间包含[-dmin, dmin]，所以不具有实际显著性

总结：总转化率同时具有统计显著性和实际显著性；而相反，净转换率则是同时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。）

以下结果均使用在线计算器结合实验数据计算得出：

1) Gross conversion：

成功数量：4
试验次数：23
概率：0.5
双尾 P 值：0.0026
由于双尾 P 值 0.0026 小于 α 水平 0.025，所以总转化率指标具有统计显著性

2) Net conversion：

成功数量：10
试验次数：23
概率：0.5
双尾 P 值：0.6776
由于双尾 P 值 0.6776 大于 α 水平 0.025，所以净转换率指标不具有统计显著性

汇总

说明你是否使用了 Bonferroni 校正，并解释原因。若效应大小假设检验和符号检验之间存在任何差异，描述差异并说明你认为导致差异的原因是什么。

不使用 Bonferroni 校正。原因：作为仅有的 2 个评估指标，总转化率与净转换率的分母相同，即它们是相互关联的，此时如果使用 Bonferroni 校正会使实验结果过于保守。

效应大小假设检验和符号检验之间不存在差异，都表明总转化率指标具有统计显著性而净转换率指标不具有统计显著性。

建议

建议还需要深入探究，暂时不发布该项更改措施。

实验使用总转化率和净转换率作为评估指标。而总转化率作为评估指标具有统计显著性和实际显著性，且置信区间为[-0.0291, -0.0120]，说明该项更改措施确实会减少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，这个符合实验假设。但净转换率作为评估指标则同时不具有统计和实际显著性，而且其置信区间为[-0.0116, 0.0019]，包含了实际显著性，表明净转换率很有可能朝着负方向移动，即这项更改措施可能会导致付费用户数量的减少，且无法确定付费用户数量减少的幅度，这并不符合实验的预期。

由上可见，在目前的评估指标判断下，该项更改措施可以显著减少因为没有足够时间而离开免费试学并因此受挫的学生数量，但同时，也有可能会导致付费用户数量的减少，而且无法确定付费用户数量的减少幅度。而由于用户缴费学习是优达学城的主要收入来源，所以付费用户数量的减少无疑会对公司的业务利润造成不利影响，而且在这一不利影响有多大尚不能确定的情况下，贸然发布该项更改措施有可能会造成较为严重的经济损失。

综上所述，还需要深入探究该项措施，建议暂时不发布。

后续试验

对你会开展的后续试验进行概括说明，你的假设会是什么，你将测量哪些指标，你的转移单位将是什么，以及做出这些选择的理由。

1、后续试验概述：线上教育网站的主要经济来源还是付费用户，所以为了增加付费用户数量，可以试着在原有的基础上，当用户点击“开始免费试学”按钮后，将会在未来的 14 天里享受和付费用户一样的体验：进入试用群并配有专门的助教在线及时解答疑惑。在试用期结束后再撤销相关付费体验服务，这样使用户切身对比感受到付费的好处，增加他们付费的可能性。

2、假设：这种体验付费服务可以增加用户的学习动力和增强用户学习体验，从而提高最终的付费率。

3、不变指标：cookie 数量、用户 id 数量、点击次数和点击概率

原因：由于这项更改是发生在用户登录网站并点击“开始免费试学”按钮后，所以 cookie 数量、用户 id 数量、点击次数和点击概率将不会受到该项更改的影响。

4、评估指标：留存率

原因：由于该项更改假设是会增加付费用户数量，目的是提高付费率，所以留存率是个不错的评估指标

5、分组单位：用户 id

原因：因为实验对象是登录并点击“开始免费试学”的用户，实验的更改并不会引起用户 id 数量的变化，而且评估指标留存率的分母是用户 id 数量，即分析单位也是用户 id，所以分组单位使用用户 id 有利于减少差异性。