优达学城A/B测试项目（P4）

idong_20180310

项目概述

在本项目中，你所要考虑的是由优达学城运行过的一个真实试验。具体数字已做更改，但是模式并没有改变。你要将试验的想法变成一个完整定义的设计、分析其结果，并提供一个高层次的后续实验。项目背景是基于Udacity 的一次真实迭代，设计一个 A/B 测试，包括测试时使用的度量和测试的运行时长等。同时根据运行 A/B 测试后的结果，来判断是否全员执行此设计。

一、设计实验

1.1选择指标

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）

- 不变指标

(1)、Number of cookies（Cookie的数量）:首页的数据，按钮不变，对于实验组和对照组都一样。期望结果：不变。

(2)、Number of clicks（点击次数）:“开始免费试学”的按钮放在首页，对于实验组和对照组都一样，不会影响首页数据，期望结果：不变。

(3)、Click-through-probalility（点进概率）：因为点击「开始免费试学」点的次数以及 cookie 的数量都不变，所以点进概率也不变。期望结果：不变。

- 评估指标

(1)、Gross conversion（总转化率）：因为修改了点击「开始免费试学」按钮后，同时会根据用户所设的时间长短，走不同的用户路径。选择大于5小时的用户，会走付费流程，小于5小时的用户则会走完全免费的路径，那么这个环节里面相较于原方案，会导致登陆用户 id 数减少。同时因为点击开始「开始免费试学」按钮的用户不变，分母不变分子减少，则期望结果为：变小。

(2)、 Net conversion（净转化率）：因为实验期望「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」，且点击开始「开始免费试学」按钮的用户不变，分母不变，分子不变，那么期望结果应该为：不变。

(3)、Retention(留存率）：因为实验期望「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」，分子不变的情况下，分母变小，则期望结果为：变大。

用户id的数量（Number of user-ids）不被选择评估度量的理由是因为样本数量无法在试验组和控制组之间进行平均分配，因此只是用user id数量这一个绝对数值无法精确地对试验的效果进行评估。最好可以使用归一化后的度量，也就是总转化率（user id数量/点击的cookie数量）作为评估度量会更好。

1.2 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。）

Unique cookies to view page per day:	每天网页浏览量	40000
Unique cookies to click "Start free trial" per day:	“开始免费试用”的点击数	3200
Enrollments per day:	每天报名客户数量	660
Click-through-probability on "Start free trial":	“开始免费试用”的点入概率	0.08
Probability of enrolling, given click:	总转化率	0.20625
Probability of payment, given enroll:	留存率	0.53
Probability of payment, given click	经转化率	0.1093125

（1）Gross conversion（总转化率）：

p = 0.20625

N= 5000*0.08 = 400

Std dev = sqrt（0.20625 * （1-0.20625） / 400） = 0.0202

说明：因为总转化率 = 用户 id 的数量 / 点击「开始免费学」按钮的唯一 cookie 数，所以作为分母的 cookie 数既是转移单位也是分析单位，所以分析估计与经验变异类似。

（2） Net conversion（净转化率）：

p = 0.1093125

N= 5000*0.08 = 400

Std dev = sqrt（0.1093125 * （1-0.1093125） / 400） = 0.0156

说明：因为净转化率 = 14天期限后仍然参加课程的用户 id 数量 / 点击「开始免费学」按钮的唯一 cookie 数，所以作为分母的 cookie 数既是转移单位也是分析单位，所以分析估计与经验变异类似。

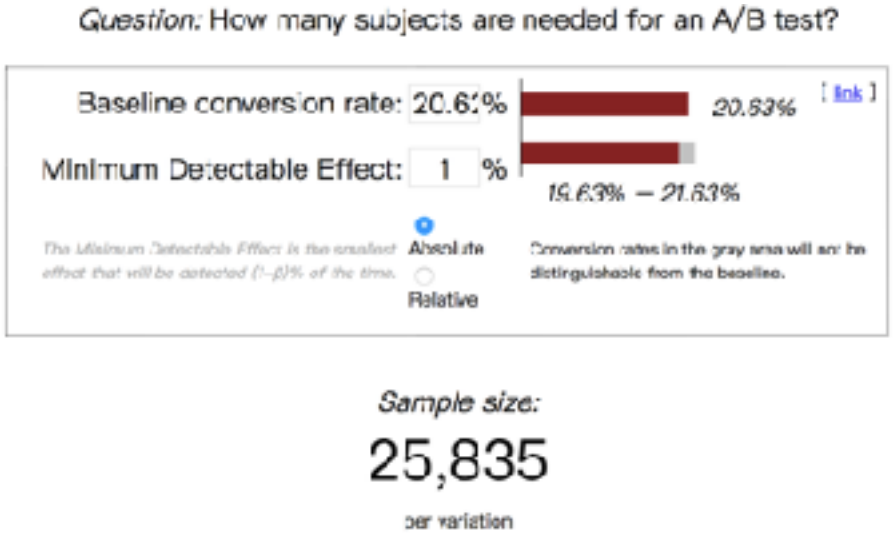
1.3 规模

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。）

1.3.1是否使用Bonferroni校正

不使用Bonferroni校正。因为本试验中仅选取了2个评估指标，且总转化率和净转化率彼此相互联系，可能同时移动，若使用Bonferroni校正会使试验结果过于保守。

因为 alpha = 0.05, beta = 0.2，根据在线计算器计算（http://www.evanmiller.org/ab-testing/sample-size.html ）所得：



1. 总转换率：

Baseline conversion rate = Probability of enrolling, given click = 0.20625 = 20.625%
Minimum Detectable Effect = 1%
样本数量 = 25835
总浏览量 = 25835/0.08 * 2 = 645875

2. 留存率：

Baseline conversion rate = Probability of enrolling, given click = 0.53 = 53%
Minimum Detectable Effect = 1%
样本数量 =39115
总浏览量 =39115/0.20625/0.08 * 2 =4741212

3. 净转换率：

Baseline conversion rate = Probability of enrolling, given click = 0.1093125= 10.93125%
Minimum Detectable Effect = 0.75%

样本数量 = 27413

总浏览量 =27413/0.08 * 2 = 685325

取上面最大的数字作为样本，留存率样本总数为 4741212。

1.3.2 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。）

（1） 曝光的流量部分：75%，试验持续时间：23天

（2）理由：

曝光流量比例要考虑两方面，风险和实验周期；

风险因素分析：

1、 对于用户来说，本试验只是询问用户每周能投入多少时间学习，不会对用户的身体、心理、财务等造成不良影响；收集的关于学习时间的数据，也不是敏感数据；

2、 对于网站来说，没有对数据库及后台改变，不用担心数据的丢失及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；因为只是多出了一个时间提示框，也没有对网页进行大的变化，因此也不必担心不同类型浏览器不符合的风险；因此试验风险较小，可以考虑给出较高比例的的流量。

实验周期因素分析：样本总数为 4741212，日流量为 40000，所以 100% 流量的话，实验天数为 119 天。但是该实验天数太长，考虑修改样本总数，取第二大浏览量 685325 。然后根据此数，75% 流量的话，实验天数为 23 天，此数较为合理。

二、试验分析

2.1 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。**

名称	Lower bound 下限	Upper bound 上限	Observed 观察值	Passes/Fail 是否通过
Number of cookies（Cookie的数量）	0.4988	0.5012	0.5006	通过
Number of clicks（点击次数）	0.4959	0.5041	0.5005	通过
Click-through-probalility（点进概率）	0.0812	0.0830	0.0821	通过

三、结果分析

3.1 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。）

	Lower bound 下限	Upper bound 上限	统计上的显著性	实际上的显著性
Gross conversion(总转化率)	-0.0291	-0.0120	具备	具备
Netconversion(净转化率)	-0.0116	0.0018	不具备	不具备

结论：总转化率具有统计显著性和实际显著性；净转化率同时不具有统计显著性和实际显著性

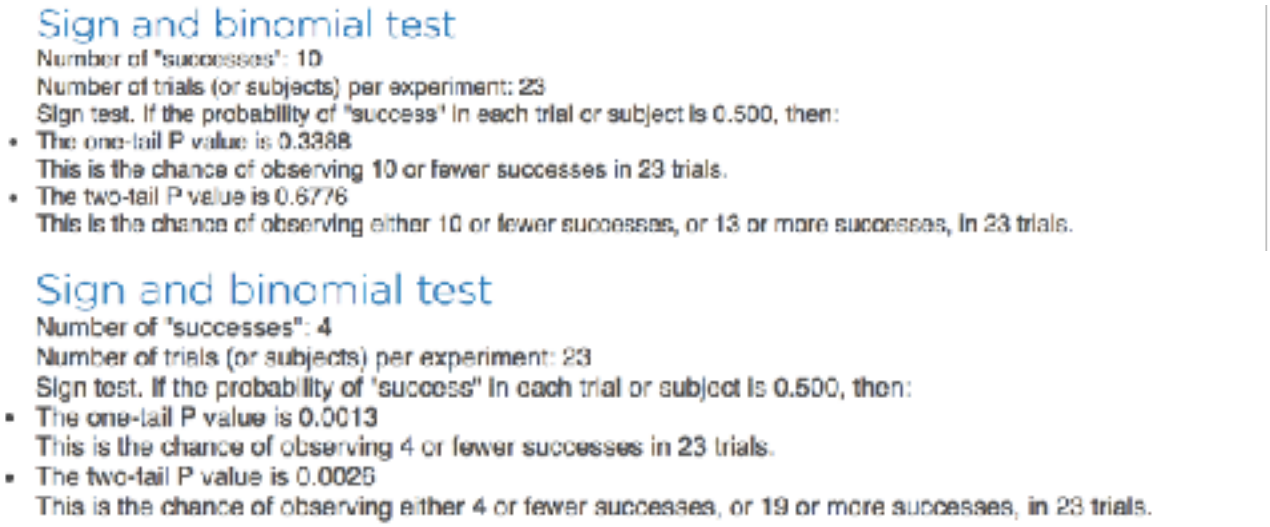
3.2符合检验

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。）

	P_value	Lower bound 下限	Upper bound 上限	统计上的显著性	实际上的显著性
Gross conversion(总转化率)	0.0026	-0.0291	-0.0120	具备	具备

Netconversion(净转化率)	0.6776	-0.0116	0.0018	不具备	不具备

通过网站https://www.graphpad.com/quickcalcs/binomial1.cfm 计算P值



结论：总转化率：双尾P值0.0026 小于 alpha 水平0.025，具有统计显著性；

净转化率：双尾P值0.6776 大于 alpha 水平0.025，不具有统计显著性。

3.3 汇总

对于每个说明你是否使用了 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；因为本试验中仅选取了2个评估指标，且总转化率和净转化率彼此相互联系，可能同时移动，若使用Bonferroni校正会使试验结果过于保守。

效应大小假设检验和符号检验之间无差异。

四、建议

暂时不实施该更改，还需要更深入的探究。原因如下：

1.从总转换率来看，我们有 95% 的信心实验结果落在（-0.0291，-0.0120）之间，说明总转化率是下降的，符合我们前面定下来的实验期望。

2.从净转换率来看，我们有理由相信它会落在（-0.01160，0.001857）之间,净转化率同时不具有统计和实际显著性，这说明增加时间提醒的实验有可能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；这不符合我们最初的假设。

五、后续试验

网络课程的一个很大短板是缺乏学习的互动性和学员的自制性，虽然有导师的答疑，有论坛互动，这些都是会有一些滞后性，没有现场的及时性，这些对于学员坚持节奏学习不利，对于大多数人，当用户想要点击试学的时候，就证明用户已经有意愿去学习，在这种情况可能会出现过于自信，而输入大于 5 小时，即使过滤出了那些小于 5 小时的用户，那些过于自信的用户也依然会流失。同时，当学员即使每周花费5小时在学习上，也会受到其它因素影响而放弃。

所以为了增大留存率，对于已经参加免费试学的学员，每周可以增加一些直播互动，直播内容可以涉及导师面对面、课程答疑、、优秀学员学习后的受益分享、每章小节测试奖励等。

- 假设：假设这种每周直播互动，能起到及时答疑、榜样鼓励、增加学员学习兴趣和参与互动的作用，增加已参加免费试学学员最终付费率。

- 测量指标：留存率。因为假设是提升最终付费率，所以选择留存率。

- 转移单位：用户id；因为试验对象是参加免费试学的用户，id便于跟踪，且 id 数量不受试验内容影响。