项目四：设计 A/B 测试

Udacity 数据分析（入门）纳米学位

实验概述

“在进行此试验时，优达学城当前的主页上有两个选项：“开始免费试学”和“访问课程资料”。如果学生点击“开始免费试学”，系统将要求他们输入信用卡信息，然后他们将进入付费课 程版本的免费试学。14 天后，将对他们自动收费，除非他们在此期限结束前取消试用。若学生点击“访问课程材料”，他们将能够观看视频和免费进行小测试，但是他们不会获得导师指导支持或验证证书，无法提交最终项目来获取反馈。

在此试验中，优达学城测试了一项变化，如果学生点击“开始免费试学”，系统会问他们有多少时间投入到这个课程中。如果学生表示每周 5 小时或更多，将按常规程序进行登录。如果他们表示一周不到 5 小时，将出现一条消息说明优达学城的课程通常需要更多的时间投入才能成功完成，并建议学生可免费访问课程资料。在这里，学生可选择继续进行免费试学，或免费访问课程资料。我们的假设是这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，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。如果这个假设最后为真，优达学城将改进整体学生体验和提高导师为能够完成课程的学生提供支持的能力。分组单位（unit of diversion）为 cookie，尽管学生参加的是免费试学，但在登录后他们的用户 id 便被跟踪。同一个用户 id 不能两次参加免费试学。对于不参加免费试学的用户，他们的用户 id 不会在试验中被跟踪，即使他们在访问课程概述页面时登录了网站。”[1]

实验设计

所有指标

[1] ″优达学城数据分析师纳米学位最终项目说明.″ 优达学城. Accessed February 4, 2018. https://s3.cn-north-1.amazonaws.com.cn/static-documents/nd002/FinalProjectInstructions_zh.pdf.

- cookie 的数量：即访问课程概述页面的唯一 cookie 的数量。（d 最小 = 3000）
- 用户 id 的数量：即参与免费试学的用户数量。（d 最小 = 50）
- 点击次数：即点击“开始免费试学”按钮的唯一 cookie 的数量（在免费试学筛选器触发前发生）。（d 最小 = 240）
- 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（d 最小 = 0.01）
- 总转化率：即完成登录并参加免费试学的用户 id 的数量除以点击“开始免费试学” 按钮的唯一 cookie 的数量所得的比率。（d 最小 = 0.01）
- 留存率：即在 14 天的期限过后仍参加课程（因此至少进行了一次付费）的用户 id 数量除以完成登录的用户 id 的数量。（d 最小 = 0.01）
- 净转换率：即在 14 天的期限后仍参与课程的用户 id 的数量（因此至少进行了一 次付费）除以点击了“开始免费试学”按钮的唯一 cookie 的数量所得的比率。（d 最小 = 0.0075）

不变量指标

- cookie 的数量： 由于变化发生在学生点击“开始免费试学”之后，对照组和实验组分配的访问课程页面 cookie 的数量不应该存在统计学上的差异。
- 点击次数：理由同上，由于用户点击“开始免费试学”这一动作在免费试学筛选器触发前发生，对照组和实验组分配的访问课程页面 cookie 的点击次数不应该存在统计学上的差异。
- 点击概率：由于点击概率 = 点击次数/cookie 的数量，且 cookie 数量和点击次数都属于此次实验中的不变量，因此点击概率也为此次实验的不变量指标。

测量指标

- 总转化率：此次实验的目标是“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”，而我们的预测是通过增加一个询问学习时间页面，那些缺少学习时间的学生不会选择继续注册课程，即完成登陆并参加免费试学的用

户 id 的数量会减少。由于总转换率 = $\frac{\text{完成登陆并参加免费试学的用户 id 的数量}}{\text{点击“开始免费试学”按钮唯一 cookie 的数量}}$，总转换率属于此次实验的测量指标。

- 留存率：由于增加询问学习时间页面后我们预计完成登陆并参加免费试学的用户 id 的数量会减少，而留存率 = $\frac{\text{在期限过后仍参加课程的用户 id 数量}}{\text{完成登录的用户 id 的数量}}$，因此留存率也属于此次实验的测量指标。
- 净转换率：此次实验的目标除“减少因为没有足够的时间而离开免费试学的学生数量”之外还包括“且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”，即希望期限后仍参与课程的用户 id 的数量不会在统计学角度有明显减少。由于净转换率 = $\frac{\text{在期限后仍参与课程的用户 id 的数量}}{\text{点击了“开始免费试学”按钮的唯一 cookie 的数量}}$，净转换率同样属于此次实验的测量指标。

非不变量/非测量指标

- 用户 id 的数量：实验中增加询问页面这一变化可能会导致选择免费试学用户 id 的数量发生变化，因此不能作为不变量指标。然而没有将此指标选作测量的原因是，由于对照组和实验组被分配的 cookie 数量不一定相同，即使发现对照组和实验组的用户 id 数量不一致，也无法肯定这是询问页面导致的还是分配的 cookie 数量之间的差异导致的。上面部分列出的三个指标相比用户 id 的数量是更好的测量指标选择。

实验测量指标的期待变动

综上所述，根据实验目标中的“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”，此次实验期待的测量指标的变化有总转换率的下降和存留率的上升。以及由于目标还包括“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”，我们希望净转换率不会有很大程度的下降。

测量变异性

计算标准偏差

此电子表格包含了度量基准值的粗略估计（从优达学城的真实数字变化而来），下面计算假设有 5000 个 cookie 访问课程概述页面时，先前选择的三个测量指标的标准偏差。

cookie 数量 = 5000

点击次数 = cookie 数量 × 点击概率 = 5000 × 0.08 = 400

用户 id 的数量 = 点击次数 × 总转换率 = 400 × 0.20625 = 82.5

总转换率的标准偏差 = $\sqrt{\frac{p\times(1-p)}{n}}$ = $\sqrt{\frac{0.20625\times(1-0.20625)}{400}}$ = 0.0202

留存率的标准偏差 = $\sqrt{\frac{p\times(1-p)}{n}}$ = $\sqrt{\frac{0.53\times(1-0.53)}{82.5}}$ = 0.0549

净转换率的标准偏差 = $\sqrt{\frac{p\times(1-p)}{n}}$ = $\sqrt{\frac{0.1093125\times(1-0.1093125)}{400}}$ = 0.0156

分析变异性与经验变异性

由于总转换率和净转换率的分析单位为 cookie，此次实验的分组单位也是 cookie，因此预计总转换率和净转换率的分析变异性和经验变异性不会出现较大差异。而存留率的分析单位为用户 id，不完全等于此次实验的分组单位 cookie，因此需要计算经验变异性。

规模

分析阶段的 Bonferroni 校正

在此次实验的分析阶段中不会使用 Bonferroni 校正，原因是对于相互独立的多个测量指标，Bonferroni 校正可以避免多重比较谬论，然而对于非独立关系的测量指标，比如此次实验中的三个相互关联的测量指标，Bonferroni 校正过于保守，易致第二类错误。

计算页面浏览量

alpha = 0.05

beta = 0.2

以下页面浏览量结果均通过 http://www.evanmiller.org/ab-testing/sample-size.html 得到。

利用总转换率计算所需页面浏览量：

总转换率基准值：0.20625

最小实际显著性比例：0.01

通过网页计算器得到每组点击次数：25825

每组所需页面浏览 = 25835 / 0.08 = 322937.5

所需总浏览量 = 322937.5 × 2 = 645875

利用留存率计算所需页面浏览量：

留存率基准值：0.53

最小实际显著性比例：0.01

通过网页计算器得到每组参与试学用户 id 数量：39115

每组所需页面浏览 = 39115 / 0.20625 / 0.08 = 2370606.061

所需总浏览量 = 2370606.061 × 2 = 4741212.121

利用净转换率计算所需页面浏览量：

总转换率基准值：0.1093125

最小实际显著性比例：0.0075

通过网页计算器得到每组点击次数：27413

每组所需页面浏览 = 27413 / 0.08 = 342662.5

所需总浏览量 = 342662.5 × 2 = 685325

最后选择通过三个测量指标得到的所需页面总浏览量的最大值，即 4741212 个页面浏览。

持续时间和暴露比例

根据上一部分分析结果，所需页面总浏览量为 4741212。

流量的暴露比例设置为 0.8，原因是此次实验不涉及用户隐私或道德问题，也不存在新闻舆论压力，总体风险较低，且同期没有其它需要进行的实验，因此让大部分流量进入实验可以有效缩短实验的持续时间。

当所需页面总浏览量为 4741212 以及流量的暴露比例为 0.8 时，根据每日页面浏览量基准值，实验持续时长为 4741212 / (40000 × 0.8) = 148.16。这一持续时间相对较长，会影响到未来可能进行的其它实验，因此放弃根据存留率得到的所需总页面浏览量转而选择根据净转换率得到的所需总页面浏览量。

最后计算出的持续时长为 685325 / (40000 × 0.8) = 21.42。因此，当暴露比例为 0.8 时，实验需要 22 天。

实验结果分析

完整性检查

在完整性检查过程中，需要计算所有不变量指标在对照组和实验组中是否没有差异。实验结果分析部分的所有数据来自 https://s3.cn-north-1.amazonaws.com.cn/static-documents/nd002/Final+Project+Results.xlsx。

首先对 cookie 的数量进行检查。

对照组的总 cookie 数量为 345543，实验组的总 cookie 数量为 344660。因此，对照组的 cookie 数量占比为$\frac{345543}{345543+344660} = 0.5006$。

标准偏差 = $\sqrt{\frac{p\times(1-p)}{n}}$ = $\sqrt{\frac{0.5\times(1-0.5)}{345543+344660}} = 0.0006$

置信区间 = $0.0006 \times 1.96 = 0.0012$

上界 = $0.5 + 0.0012 = 0.5012$

下界 = $0.5 - 0.0012 = 0.4988$

由于对照组的 cookie 数量占比落在了上界和下界之间，cookie 的数量通过完整性检查。

对点击次数的检查过程同上，得到对照组的点击次数占比为 0.5005。由于上界和下界分别为 0.5041 和 0.4959，点击次数通过完整性检查。

最后对点击概率进行完整性检查。对照组的点击率为$\frac{28378}{345543} = 0.0821$，标准偏差为$\sqrt{\frac{0.0821\times(1-0.0821)}{345543}} = 0.0005$。因此，置信区间为$0.0005 \times 1.96 = 0.0009$，上界和下界分别为 0.0830 和 0.0812。

实验组的点击概率为$\frac{28325}{344660}=0.0822$，由于其落在上界和下界之间，点击概率也通过完整性检查。

效应量检验

在持续时间和暴露比例部分，留存率这一测量指标由于其所需实验持续时间过长而被放弃，因此在效应量检验部分只需检验总转换率和净转换率的效应量。

首先检查总转换率。对照组的总转换率为$\frac{3785}{17293}=0.2189$，实验组的总转换率为$\frac{3423}{17260}=0.1983$，那么总转换率的差为 0.1983 - 0.2189 = -0.0206。

合并概率为$\frac{3423+3785}{17293+17260}=0.2086$，对应的合并方差为

$\sqrt{0.2086\times(1-0.2086)\times(\frac{1}{17293}+\frac{1}{17260})}=0.0044$。因此，置信区间为

$0.0044\times1.96=0.0086$，得到的上界为-0.0120，下界为-0.0291。由于上界和下界包围的区间不经过 0，且区间里的所有数字均小于-0.01，这一结果既具有统计显著性也具有实际显著性。

接下来检查净转换率。对照组的净转换率为$\frac{2033}{17293}=0.1176$，实验组的净转换率为$\frac{1945}{17260}=0.1127$，那么净转换率的差为 0.1127 - 0.1176 = -0.0049。

合并概率为$\frac{2033+1945}{17293+17260}=0.1151$，对应的合并方差为

$\sqrt{0.1151\times(1-0.1151)\times(\frac{1}{17293}+\frac{1}{17260})}=0.0034$。因此，置信区间为

$0.0034\times1.96=0.0067$，得到的上界为 0.019，下界为-0.0116。由于上界和下界包围的区间经过 0，且区间里的并非所有数字都小于-0.01，这一结果既不具有统计显著性也不具有实际显著性。

符号检验

接下来通过符号检验查看其结果是否和上一部分效应量检验得到的结果相符。

通过对比对照组和实验组从 10 月 11 日至 11 月 2 日（包括界限日）期间每一天的总转换率，可以发现在 23 天中有 19 天对照组的总转换率都小于实验组的总转化率。根据 https://www.graphpad.com/quickcalcs/binomial1/的网页计算， p 值为 0.0026。由于这一结果小于我们设定的 alpha 值 0.05，总转换率具有统计显著性，符合效应量检验得到的结果。

通过对比对照组和实验组从 10 月 11 日至 11 月 2 日（包括界限日）期间每一天的净转换率，可以发现在 23 天中有 13 天对照组的净转换率小于实验组的净转化率。根据 https://www.graphpad.com/quickcalcs/binomial1/的网页计算， p 值为 0.6776。由于这一结果远大于我们设定的 alpha 值 0.05，净转换率不具有统计显著性，也符合效应量检验得到的结果。

结果汇总

在此次实验结果的分析过程中，Bonferroni 校正没有被使用，原因和实验前的分析阶段相同：对于相互独立的多个测量指标，Bonferroni 校正可以避免多重比较谬论，然而对于非独立关系的测量指标，比如此次实验中的三个相互关联的测量指标，Bonferroni 校正过于保守，易致第二类错误。

由于符号检验的结果与效应量检验一致，说明总转换率有非常大的概率具有统计显著性，同时净转换率有非常大的概率不具有统计显著性。

建议

通过以上结果分析，由于实验组的总转换率相比对照组大大降低，且结果既具有统计显著性也具有实际显著性，增加询问页面这一变化达到了“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”的目标。但是，对于“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”这一目标，由于净转换率的计算结果不具有统计显著性，我们还无法确定增加询问页面这一变化究竟会不会大幅度地减少继续通过免费试学和最终完成课程的学生数量。由于以上原因，我的建议是先不要启动这一变化，而是设计并实施其它实验以确定净转换率是否会大幅度下降。

后续实验

为提高学员在注册课程之后的参与积极度，我会考虑在学员的“我的课程”页面增加一个“你这周累计学习了…分钟，打败了…%学员”的显示块。

我的假设是课程页面有这个显示块的学员参与课程的积极度会更高，也因此在试用期限到期前取消课程的概率更低。

测量指标为留存率，即在试用期限后仍参与课程的用户 id 的数量除以点击“开始免费试学”按钮的唯一 cookie 的数量所得的比率。

分组单位为用户 id，因为学员只能在登陆以后才能进入“我的课程页面”。