优达学城数据分析师纳米学位 P7

试验设计

优达学城在当前主页上有两个选项："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。

试验设计："试验组的学生"在点击"开始免费试学"后出现一个提示，问学生"一周有多少时间投入到课程中"，如果学生表明每周 5 小时或更多，则跳到输入信用卡信息的页面，之后所有的流程与上面提到的流程一致。如学生表明每周学不到 5 小时，则系统提示学生如果要完成优达学城的课需要更多的时间，并建议学生可以"先访问课程资料"；学生可以自己选择继续"免费试学"还是"访问课程资料"。

希望取得的试验结果：在"交费学生"不明显减少的情况下，"大幅"降低"总转化率 Gross conversion"，即减少免费试学学生的数量。一部分学生在得知需要大量时间学习后，知难而退，不进行试学，这样他们也没必要以后退课，也不会有挫败感；反过来说对优达学城是好事，因为免费试学也需要人工辅导和批改作业，如果学生以后真的坚持不下来，那么还不如不参加免费试学，这可以节省犹达学城的经费和精力。

度量选择

优达学城提供了 7 种度量：cookie 的数量、用户 id 的数量、点击次数、点进概率、总转化率、留存率、净转换率。

其中不变度量为：cookie 的数量、点击次数、点进概率

评估度量为：总转化率（Gross conversion）、留存率（Retention）、净转换率（Net conversion）

问题：对于每个度量，解释你为什么使用或不使用它作为不变度量和评估度量。此外，说明你期望从评估度量中获得什么试验结果。

答：-"cookie 的数量"、"点击次数"是不变度量，因为试验把"提示用户需要更多学习时间的信息"放在点击"开始免费试学"之后。所以不影响上述这两个指标。

-用户 id 的数量：即参与免费试学的用户数量，我们预计会减少，因为部分没有时间的学生会被"分流"到"访问课程资料"。但是它只是一个绝对数，它是"总转换率 Gross conversion"的分子，"总转换率 Gross conversion"的分母是"点击次数"（"点击次数"是不变度量）。所以"用户 id 的数量"和"总转换率 Gross conversion"反映的信息重复了，但是"总转换率 Gross conversion"在设置 d_{min} 更方便些，只需定 0.01 即可，而且"总转换率"是相对数，能反映"用户 ID 数量"和"点击次数"的比值，所以"用户 id 的数量"不作为评估度量，只作为评估度量"总转换率"的分子。

-点进概率：它等于"点击次数"/"cookie 的数量"，分子和分母都是不变度量，那么"点进概率"也是不变度量。

-总转化率（Gross conversion）：作为评估度量。当增加"需要更多时间的提示"后，该指标应降低，因为可能部分用户被分流到"访问课程资料",所以免费试学的人减少了，"总转化

率”的分子降低了，所以“总转化率”也降低了。
-留存率（Retention）：作为评估度量。当增加“需要更多时间的提示”后，该指标应该增加，因为登陆的学生的质量提高了（因为一部分学生知难而退，没有选择登陆），所以最后交钱的比例提高了。
-净转换率（Net conversion）：作为评估度量，当增加“需要更多时间的提示”后，我们期望该度量不降低，即期望不减少付费学生的数量。（但该指标可能会降低，因为一部分可能会交钱的学生，在得知需要更多的时间学习后，知难而退；如果没有提示，他们中的部分人最终也可能会付费。）

测量标准偏差

评估度量的标准差如下：
-总转化率（Gross conversion）：sqrt(0.20625*(1-0.20625)/(5000*3200/40000)) = 0.0202
-留存率（Retention）：sqrt(0.53*(1-0.53)/(5000*660/40000)) = 0.0549
-净转换率（Net conversion）：sqrt(0.1093125*(1-0.1093125)/(5000*3200/40000)) = 0.0156

问题：对于每个评估度量，说明你是否认为分析估计与经验变异是类似的，或者你是否期望它们是不同的（如果是这样，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。
答：首先“项目说明”中明确表示：“转移单位（unit of diversion）为 cookie”；“评估度量”的分母定义为“分析单位 Unit of analysis”，那么：

-总转化率（Gross conversion）：总转化率的分母是“cookie 的数量”，所以分析单位是“cookie 的数量”。分析单位与转移单位相同，**所以“分析变异性”与“经验变异性”相似。**

-留存率（Retention）：留存率的分母是“用户 id 数量”，所以分析单位为“用户 id 数量”。分析单位与转移单位不同，**所以“分析变异性”与“经验变异性”不相同，“分析变异性”要小于“经验变异性”。**在时间允许的情况下有必要进行经验估计。

-净转换率（Net conversion）：净转换率的分母是“cookie 的数量”，所以分析单位是“cookie 的数量”。分析单位与转移单位相同，**所以“分析变异性”与“经验变异性”相似。**

规模

样本数量和支持

我在分析阶段不使用 Bonferroni 校正。（原因见下面的“汇总”部分。）
我使用 http://www.evanmiller.org/ab-testing/sample-size.html 提供的在线计算器计算"网页访问数"：
总转化率（Gross conversion）: 2 × 25835 × 40000 ÷ 3200 = 645875
留存率（Retention）: 2 × 39115 × 40000 ÷ 660 = 4741212
净转换率（Net conversion）: 2 × 27413 × 40000 ÷ 3200 = 685325
我选择上述三个指标中网页访问数最大的 4741212 为样本数量。

持续时间和风险暴露

我选择 4741212 作为样本数量，这是一个非常大的数。流量系数是 1.0，即全部流量都用来做试验：

计算天数：4741212 ÷ (40000 × 1.0)= 118.5303，需要用 119 天来完成试验，**在测试中没通过**。

119 天作为试验太长了，也就是说试验暴露的时间太长了；而且把流量系数设置为 1.0 也会有很大的风险。因为我们不清楚增加“需要更多时间的提示”后，犹达学城的付费人数会下降多少，如果下降的很厉害，优达学城会有较大损失，所以我们需要对试验的时间和流量进行控制。

回过头来看上述 3 个指标所需的样本数量，留存率需要的样本数量要远大于“总转化率”和“净转换率”，即使用全部的流量测试“留存率”，也需要 119 天，AB 测试的成本太高，也不现实，所以我不得不剔除“留存率”这个度量，保留“总转化率”和“净转换率”。所以我选择上面计算的第二大的数字：**685325 作为样本数量**。流量系数为 0.5。

计算天数：685325÷ (40000 × 0.5) = 34.2662，需要 34 天来完成试验，**在测试中通过。**

问题：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？

答：我希望尽量减少测试时间，并尽量减少测试对优达学城现有体验的影响。34 天的时间并不太长，用流量的 50%用来测试，实际只影响了 25%的流量，因为对照组还是和以前一样的。在 25%的试验流量中，被影响的学生也是少数，因为大多数学生还是会选择一周超过 5 小时的学习，如果学生真的要学一门技能的话，这个要求也并不过分。所以用 50%的流量来测试，风险不大，而且测试时间只有 34 天，试验的暴露时间也不长，这是可以接受的。

试验分析

合理性检查

“cookie 的数量”完整性检查:

对照组 cookie 数量：345543

试验组 cookie 数量：344660

SD = sqrt(0.5*0.5/(345543+344660)) = 0.0006

m = SD*1.96 = 0.0012

上边界：0.5 - m = 0.4988

下边界：0.5 + m = 0.5012

观察到的值：p = 345543/(345543+344660) = 0.5006

p 在上下边界之间，完整性检查通过。

“点击次数”完整性检查:

对照组点击次数：28378

试验组点击次数：28325

SD = sqrt(0.5*0.5/(28378+28325)) = 0.0021

m = SD*1.96 = 0.0041

上边界：0.5 - m = 0.4959

下边界：0.5 + m = 0.5041

观察到的值：p = 28378/(28378+28325) = 0.5005

p 在上下边界之间，完整性检查通过。

***“点进概率”完整性检查：

P_{pool} = (28378+28325)/(344660+345543) = 0.0822

SE_{pool_1}=sqrt(P_{Pool_1}*(1- P_{Pool_1})*(1/344660+1/345543)) = 0.0007

m = SE_{pool_1}*1.96 = 0.0013
上边界：0 - m = -0.0013
下边界：0 + m = 0.0013
观察到的值：d = 0.0822 - 0.0821 = 0.0001
d 在上下边界之间，完整性检查通过。

问题：对于任何未通过的完整性检查，根据每日数据解释你猜测的最可能的原因。在所有合理性检查通过前，不要开始其他分析工作。
答：综上所述，完整性检查已通过。

结果分析效应大小检验
总转化率:
P_{Pool_1} = (3785+3423)/(17293+17260) = 0.2086
SE_{pool_1}=sqrt(P_{Pool_1}*(1- P_{Pool_1})*(1/17293+1/17260)) = 0.0044
评估度量的差异：d_1 = 0.1983-0.21887 = -0.0206
m_1 = SE_{pool_1}*1.96 = 0.0086
总转化率试验和对照组之间的差异的 95%的置信区间为：
[d_1 – m_1, d_1+m_1] = [-0.0291,-0.0120]
由于置信区间的上下边界均小于 0，所以具备统计显著性。
由于置信区间的上边界小于 d_{min} = -0.01，所以具备实际显著性。

净转换率:
P_{Pool_2} = (2033+1945)/(17293+17260) = 0.2086
S_{Epool_2}=sqrt(P_{Pool_2}*(1- P_{Pool_2})*(1/17293+1/17260)) = 0.0034
评估度量的差异：d_2= 0.112688-0.117562 = -0.0049
m_2 = SE_{pool_2}*1.96 = 0.0067
总转化率试验和对照组之间的差异的 95%的置信区间为：
[d_2 – m_2, d_2+m_2] = [-0.0116, 0.0019]
由于置信区间包含 0，所以不具备统计显著性，也不具备实际显著性。

符号检验
总转化率:
实验组和对照组“总转化率”每天的差值有 23 个，其中值为负的有 19 个。用课程提供的在线计算器 http://graphpad.com/quickcalcs/binomial1.cfm，计算得到 p-value = 0.0026，小于 0. 05，具备统计显著性。

净转换率:
实验组和对照组“总转化率”每天的差值有 23 个，其中值为负的有 13 个。用课程提供的在线计算器 http://graphpad.com/quickcalcs/binomial1.cfm，计算得到 p-value = 0.6776，大于 0. 05，不具备统计显著性。

汇总
问题：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。

答：我没有使用 Bonferroni 校正。首先我只有两个评估度量，每个度量的 alpha = 0.05，那么总的 alpha = 1-(1-0.05)^2 = 0.0975，并没有扩大太多。另外，Bonferroni 矫正有自身的缺点，就是 Bonferroni 太保守了；在学生看到“需要更多时间的提示”后净转换率和总转化率都是减少的，说明这两个度量是相关的。比如：极端情况是两个评估度量完全一样，本来 alpha= 0.05 下，这两个度量具备统计显著性和实际显著性，Bonferroni 则分配给每个度量 0.025，这样可能这两个度量都不具备统计显著性和实际显著性了，可能一个好的改变就让 Bonferroni 给否定了。

“净转换率”的下降虽然没有统计显著性和实际显著性，但是其置信区间包含负值，而且置信区间的范围更偏向负值，说明付费人数可能会减少，这仍然是我们不希望看到的情况。

建议

提供建议并简要说明你的理由。

答：我不建议发布“询问学生学习时间的提示”，原因如下：

（1）虽然“净转换率”的下降没有统计显著性和实际显著性，但是执行该项举措可能会使付费人数减少，这会降低优达学城的收入。

（2）虽然“总转化率”下降具有统计显著性和实际显著性，但是该措施并没有**“大幅”**降低“总转化率”，而且优达学城也需要评估“减少的 “辅导免费试学的学生”的工作量”能省多少钱，如果节省钱不多的话，这项措施就没有必要。

（3）如果该项举措确实能为优达学城节省很多钱，我建议再做一个更大强度的 A/B test（人数更多，持续时间更长），来验证该措施是否真正可行。

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，你将测量哪些度量，你的转移单位将是什么，以及做出这些选择的理由。

答：

（1）我设计的试验如下：

优达学城可以尝试在课程概述页面上列出“已毕业学员找到好工作的成功案例”，类似下图某学习网站的案例：

（2）试验结果假设：学生进入主页后，看到前辈的成功案例后可能会受到鼓励，增加学习的动力，预计可以提升点击“开始免费试学”或“访问课程资料”按钮的次数。这有助于提升优达学城用户的人数。零假设是“对照组”和“试验组”的评估度量（“点进概率”）的差等于 0，对立假设是二者的差不等于 0。

（3）不变度量是“cookie 的数量”,即访问课程概述页面唯一 cookie 的数量。唯一性按天决定。

（4）评估度量是“点进概率”，该度量是“点击“开始免费试学”或“访问课程资料”按钮的唯一 cookie 的数量”（也就是说，无论该学生是点击“开始免费试学”按钮还是“访问课程资料”按钮，或者二者都点击，该 cookie 只记录一次）除以“cookie 的数量”所得的比率。（d_{min} =0.01）唯一性按天决定。由于学生可能会受到成功案例的鼓励，所以该评估度量预计会提高。**选择此评估度量的理由：**无论用户点击了哪种按钮都说明用户与网站进行了更深层次的互动，如果该评估度量提高，说明有更多的人点击按钮进入下层的网页，这有助于提高优达学城的用户数。

（5）转移单位：“cookie 的数量”。因为分析单位为“cookie 的数量”（“点进概率”的分子），转移单位要与分析单位一致，这样 **“分析变异性”与“经验变异性”相似。**这样,我们通过计算得到的标准差与实际情况基本一致，保证了结果的正确性。

（6）试验的可行性： "cookie 的数量"和“点击按钮的唯一 cookie 数量”在记录和跟踪上都很简单，可在多数基础设施下进行测量。

（7）风险：

道德风险：对于参加试验组的学生来说没什么道德风险，在了解了别人成功的案例后，不会产生精神或者身体上的伤害，也不会暴露参与者的隐私。但是，该试验会暴露成功案例者的隐私，所以我们需征求“成功案例者”的同意，可能还需提供一些奖励作为补偿，比如可以免费学习课程多少小时。

试验的风险：参加试验的学生可能会怀疑案例的真实性，也可能会引起相反的结果。所以即使这个小内容的添加，也不宜开展大规模的试验。用 50%的流量参与 AB 试验可能就足够了（还需要我们计算样本的数量进行权衡）。另外验证“成功案例者的信息”也需要时间和经费。