Design an AB Test

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）

- 不变指标：Number of cookies, Number of clicks,Click-through-probability
- 评估指标：Gross conversion, Net conversion

> 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。

不变指标：

- **Number of cookies**：访问课程概述页面的唯一 cookie 的数量。这个指标平均地分布在试验组和控制组当中。并且在访问者看到页面更改之前，cookies 已经生成了，所以 Number of cookies 是个很好的不变指标。
- **Number of clicks**：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）。选择它作为不变指标的原因与 Number of cookies 类似，在点击“开始免费试学”按钮前时，访问者还看不到该试验的变化。
- **Click-through-probability**: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。它是两个不变度量的比值，也应当是不变度量。

评估指标：

- **Gross conversion**：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。Gross conversion 可以作为评估指标，因为 该试验组的用户会看到与对照组不一样的地方在于，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。 但对照组的用户就看不到这个提示。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。如果试验组的总转化率要低于对照组的总转化率，则该假设成立；若试验组总转化率并不比对照组低，则假设不成立。
- **Net conversion**：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。 因为我们假设这项更改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，而这项指标可以衡量该更改是否会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。如果试验组的净转化率并不比对照组少很多，那么该假设成立；若试验组的净转化率明显低于对照组的，则假设不成立。

其他指标：

- **Number of user-ids**：它发生于试验之后，会受到试验的影响，因此它是一个ok的评估度量。但是，由于实验组和对照组的cookie数量不一定相同，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的数量作为评估度量。
- **Retention**：留存率也发生于试验之后，也是比较不错的评估指标。不过经过后续的计算，我们会发现它需要过多的页面浏览量和试验运行时间，因此在规定的时间内我们无法采集足够的样本数据，也不适合作为评估度量。

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。）

- Gross conversion: 0.0202
- Net conversion: 0.0156

> 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。

对于上述两个指标，分析估计与经验变异都是类似的。因为两个指标都是基于 Number of cookies，他们也是分组单元，所以分组单元和分析单元是一致的。所以分析估计与经验变异是类似的。

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，并给出试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。）

我不会在分析阶段使用 Bonferroni 校正，因为本次试验的指标具有高度相关性，而 Bonferroni 校正可能太过保守，会影响试验结果。

页面浏览量：685325

> 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。）

页面浏览量：685325

曝光的流量部分：1

试验持续时间：18

> 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？

我认为可以转移全部流量到该试验中。

在该试验中，系统会问实验组的用户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。虽然增加了询问的环节，但学生选择免费试学还是访问课程资料，是自由的。所以该试验并不会对学生造成任何伤害。

在此试验中，我们虽然用了cookies来判断用户的一致性，但并没有收集时间戳、年龄、性别、住址等个人信息，因此也没有涉及隐私问题。

对于优达学城来说，该试验也没有造成过多的改变，不会产生太多的数据，也不会显著增加网站的负担，所以对优达学城没有明显伤害。

另外，如果仅小将部分流量分流至该试验，则会导致试验周期边长，也可能会导致由于时间变化引起的不可控因素增加。

所以，我认为可以将 100% 流量用于该试验，18天完成。

试验分析

完整性检查

> 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）

两项指标均通过完整性检查，结果如下：

- **Number of cookies**：

  Confidence Interval: [0.4988,0.5012]

  Observed: 0.5006

- **Number of clicks**：

  Confidence Interval: [0.4959,0.5041]

  Observed: 0.5005

- **Click-through-probability**:

  Confidence Interval: [0.0812 , 0.0830]

  Observed: 0.0822

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。）

- **Gross conversion**： 95% 置信区间: [-0.0291,-0.0120]

  具有统计显著性 (置信区间不包括 0)

  具有实际显著性 (置信区间不包括 dmin)

- **Net conversion**：

  95% 置信区间: [-0.0116,0.0019]

  不具有统计显著性 (置信区间包括 0)

  不具有实际显著性(置信区间包括 dmin)

> 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。）

- **Gross Conversion**：

  Success: 4

  Total: 23

Probability: 0.5

p-value : 0.0026

具有统计显著性

- **Net conversion**：

  Success: 10

  Total: 23

  Probability: 0.5

  p-value : 0.6776

  不具统计显著性

> 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。

Bonferroni 校正是对于多重检验问题进行校正的方法之一。是否要进行Bonferroni 校正取决于我们的测试是什么类型的，如果测试指标之间的关系是 OR，指标间的相关性低，用多个指标去验证一个假设，则可以用Bonferroni 校正；如果测试指标之间的关系是 AND，指标间相关性高，则不需要进行Bonferroni 校正。

在该试验中，总转化率和净转化率两个指标相关性很高，且是为了检验不同的问题，都具有重要的意义，所以不需要进行Bonferroni 校正。

建议

我们可以看到，总转化率具有负面的统计显著性和实际显著性，说明测试的修改起到了作用，新的提示使学生更加慎重地考虑自己是否有足够的时间投入到课程的学习中，而这些没有足够时间的学生，显然也不会成为付费用户。这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的。

而净转化率不具有实际显著性，也不具有统计显著性，但是净转化率的置信区间包含负数，置信区间的含义是”我们有95%的信心试验结果会落在这个区间“，根据此处的计算结果(-0.0116, 0.0019)，也就是说有很大的概率净转化率会减少，并且有一定的概率净转化率的减少会超过实际显著性0.0075。也就是说此试验可能“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”。

因此，不建议发布此免费试学筛选器。

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。

为了实现“减少提前终止”这个目的，我计划设计一个试验：在学生点击“开始免费试学”的第 8 天，向学生发放 300 元的课程优惠券，学生可以在付费时抵扣使用。

假设：为学生提供优惠券，会减轻学生的经济负担，也就减少了学生付费的阻力，会使更多有需求的学生进行后续课程学习。

- 转移单位：用户 id

  此试验关注的是已经点击“开始免费试学”并注册的用户，所以参与该试验的用户都具有用户 id，使用用户 id作为转移单位，非常适合统计用户的行为。

- 不变指标：用户 id 的数量

  即参与免费试学的用户数量。 此试验关注的是用户注册后的付费行为，所以用户 id 的数量是一个很好的不变指标。

- 评估指标：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。选择留存率作为评估指标的原因是，它可以反映注册用户进一步成为付费用户的比例，非常合适用来做此次试验，如果实验组的留存率高于对照组，则假设成立，如果实验组不高于甚至低于对照组，则假设不成立。