AB-test方法论报告

刘铮源

1.理论依据

AB测试是为Web或App界面或流程制作两个（A/B）或多个（A/B/n）版本，在同一时间维度，分别让组成成分相同（相似）的访客群组（目标人群）随机的访问这些版本，收集各群组的用户体验数据和业务数据，最后分析、评估出最好版本，正式采用。方便起见，本报告中仅讨论两个版本且**样本量足够大**的情况。

中心极限定理：大量相互独立的随机变量的均值的分布以正态分布为极限，也就是趋近正态分布，与随机变量的具体分布无关。

$$\sum_{i=1}^{n} X_i \sim N(n\mu, n\sigma^2) \tag{1}$$

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N(\mu, \frac{\sigma^2}{n}) \tag{2}$$

特别的，我们有：

$$\bar{X_1} \sim N(\mu_1, \frac{\sigma_1^2}{n}), \bar{X_2} \sim N(\mu_2, \frac{\sigma_2^2}{n}) \tag{3}$$

$$\bar{X_1} - \bar{X_2} \sim N(\mu_1 - \mu_2, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}) = N(0, \frac{2\sigma^2}{n}) \tag{4}$$

其中，$\mu_1 = \mu_2, \sigma = \sigma_1 = \sigma_2$，因为我们假设A/B组样本均值的分布是无差异的，有时候我们会放宽一些假设，只假设A/B组样本均值分布的期望是无差异的，即$\mu_1 = \mu_2$，这时原假设下的事件分布是下面的形式：

$$\bar{X_1} - \bar{X_2} \sim N(0, \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{n}) \tag{5}$$

中心极限定理给出了样本均值的抽样分布形式，但是我们注意到，样本均值的抽样分布涉及到总体的均值和方差，总体的均值由样本均值给出估计，总体的方差由S^2给出：

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{6}$$

由此我们得到总体均值μ的无偏估计$\bar{X}$，总体方差σ^2的无偏估计S^2。

讨论双边情况，记零假设$\mathbf{H0}$：$\bar{X_1} = \bar{X_2}$，备择假设$\mathbf{H1}$：$\bar{X_1} \neq \bar{X_2}$，显著性水平α。

我们有 z 统计量和置信区间：

$$z = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}} \tag{7}$$

$$\sqrt{\sigma^2 \quad \sigma^2}$$

$$(\bar{x_1}-\bar{x_2})\pm z_{\alpha/2}\sqrt{\frac{\sigma_1}{n_1}+\frac{\sigma_2}{n_2}} \tag{8}$$

单边情况下置信区间为

$$\begin{gathered}[(\bar{x_1}-\bar{x_2})-z_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}},+\infty] \\ [0,(\bar{x_1}-\bar{x_2})+z_\alpha\sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}]\end{gathered} \tag{9}$$

2. 适用场景

i UI优化

这是比较常见的场景。

不像功能的设计，存在着很多逻辑上的思路，经常还是可以确定哪种方案好，哪种方案不好。UI的优化，往往是很"艺术"层面的。往往看到真实数据前，谁也难以说明哪种设计能带来更好的数据效果。如下图：

ii 文案变化

这个其实和UI层面的优化很类似。同一个按钮，叫【现在申请】还是【立刻申请】呢？

iii 页面布局

页面布局，主要指的是同页面中的不同元素的排列方式。

iv 算法优化

算法优化，也是AB测试的一个重要场景。

上线前的算法，基本都是基于历史数据进行算法模型的训练、搭建。在本地模型效果再好，上线后也不一定有良好的表现。只有线上才是检验算法效果的决定性标准。

但谁也不能确保上线后的效果吧？那这时先用小流量做一些AB测试，是非常不错及通用的选择。

3. 执行步骤

i 调查

在制定A/B测试计划之前，需要对网站当前的运行情况进行彻底的研究。你必须收集所有相关的数据，包括有多少用户访问网站，哪个页面驱动了最大的流量，不同页面的不同转换目标，等等。这里使用的A/B测试工具可以包括谷歌analytics、Omniture、Mixpanel等定量网站分析工具，它可以帮助你找出你访问次数最多的页面、花费时间最多的页面或跳出率最高的页面。例如，你可能想从有最高收入潜力或最高每日流量的候选页面开始。

ii 观察并提出假设

从本部分开始，将用一个简单的实操例子来说明具体的执行步骤，数据集来自Kaggle的ABtest**数据集**。

数据集包含两个40*4的dataframe,分别为控制组和对照组。每组四个变量为，impression（网页流量），click（点击数），purchase（购买数），和earning（消费总金额）。

```
## read data
control_Group=pd.read_excel('ab_testing.xlsx',sheet_name='Control Group')
test_Group=pd.read_excel('ab_testing.xlsx',sheet_name='Test Group')
control_Group.head()
```

✓ 0.7s

	Impression	Click	Purchase	Earning
0	82529.459271	6090.077317	665.211255	2311.277143
1	98050.451926	3382.861786	315.084895	1742.806855
2	82696.023549	4167.965750	458.083738	1797.827447
3	109914.400398	4910.882240	487.090773	1696.229178
4	108457.762630	5987.655811	441.034050	1543.720179

由于数据量=40>30，所以认为是足够大的数据集，因此可以用z分布近似t分布。

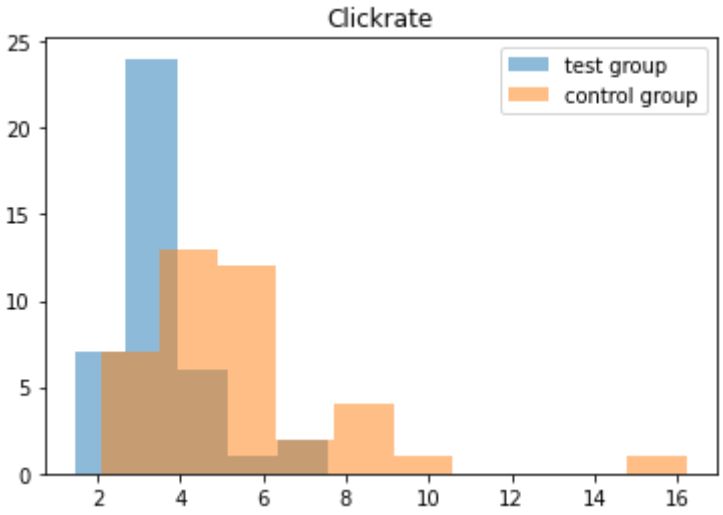
$$\text{clickrate} = \frac{\text{click}}{\text{impression}} \tag{10}$$

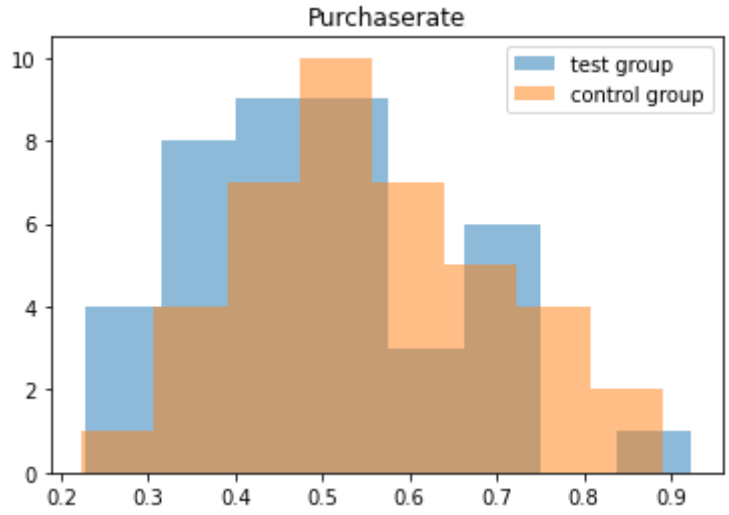
$$\text{purchaserate} = \frac{\text{purchase}}{\text{impression}} \tag{11}$$

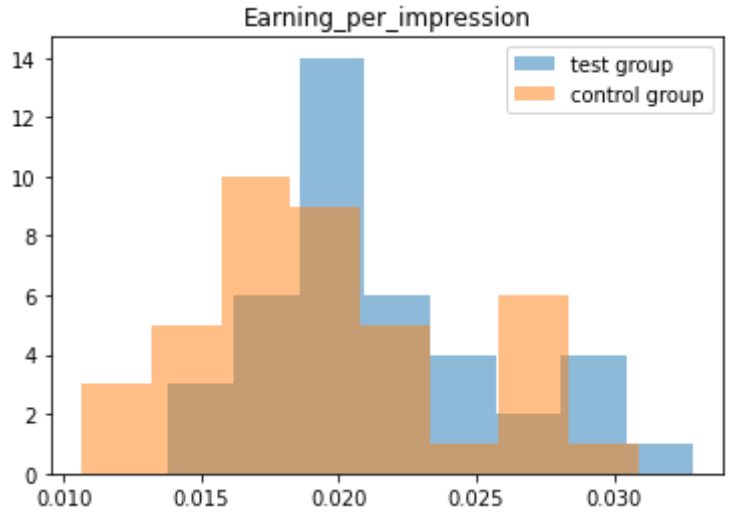
$$\text{earning per impression} = \frac{\text{earning}}{\text{impression}} \tag{12}$$

接下来通过画柱状图对整个数据集有一个大致了解,我们可以发现，测试集点击率，购买率都相比于对照集低，但是收入高。由此我们分别对三个数据提出单边零假设：

$$\begin{aligned} H_{01}&：\text{测试组点击率} >= \text{对照组} \\ H_{02}&：\text{测试组购买率} >= \text{对照组} \\ H_{03}&：\text{测试组消费金额/网页流量} <= \text{对照组} \end{aligned} \tag{13}$$







iii 进行AB-test

这里使用python中statsmodels包做z-test。

```python
diff=np.mean(test_Group['Clickrate'])-np.mean(control_Group['Clickrate'])
pvalue=sw.ztest(test_Group['Clickrate'],control_Group['Clickrate'],alternative='smaller')
if pvalue<0.05:
    print('p-value<0.05, so we reject null hytheposis')
else:
    print('p-value>=0.05, so we fail to reject null hytheposis')
```

0.5s Python

```
p-value<0.05, so we reject null hytheposis
```

```python
diff=np.mean(test_Group['Purchaserate'])-np.mean(control_Group['Purchaserate'])
pvalue=sw.ztest(test_Group['Purchaserate'],control_Group['Purchaserate'],alternative='smaller')
if pvalue<0.05:
    print('p-value<0.05, so we reject null hytheposis')
else:
    print('p-value>=0.05, so we fail to reject null hytheposis')
```

0.4s Python

```
p-value<0.05, so we reject null hytheposis
```

```python
diff=np.mean(test_Group['Earning_per_impression'])-np.mean(control_Group['Earning_per_impression'])
z,pvalue=sw.ztest(test_Group['Earning_per_impression'],control_Group['Earning_per_impression'],alternative='larger')
if pvalue<0.05:
    print('p-value<0.05, so we reject null hytheposis')
else:
    print('p-value>=0.05, so we fail to reject null hytheposis')
```

0.7s Python

```
p-value<0.05, so we reject null hytheposis
```

三个零假设均被拒绝。

iv 分析结果

虽然统计学意义上我们推翻了零假设，但是我们并不能说这具有实际或者商业上的价值，因此我们还要计算两组的平均收入并比较。

```
diff=np.mean(test_Group['Earning_per_impression'])-np.mean(control_Group['Earning_per_impression'])
per=diff/np.mean(test_Group['Earning_per_impression'])*100
print('The earning in test group is '+str(round(per,2))+'%'+' lager than that in control group ')
```

✓ 0.5s

```
The earning in test group is 8.99% lager than that in control group
```

发现测试组每单位点击的收入相比对照组有9%的提高，结合iii中的统计学结论，我们得出:

虽然测试组的点击率购买率均较低，但是总收入提升了。可能是吸引了人数较少的消费意愿更强的顾客或者展示了价格更高的商品，使得平均每单单价上升以达到增加消费额的目的。

4. 我的思考

在AB-test基础上我们可以分析用户画像。例如我是一个20岁男性城市青年，我对于产品的只会有刚需，也就是说我不想用的软件，商品都不会使用/购买，也就是说如果对我做AB-test，到最后购买率或者长时间留存率AB效果大概率是一样的。基于此，有一种人，我们称他们为"白嫖者"，会被精美且合理的网页布局吸引但不会消费，我们就可以无视他们的数据。但是或许会有一批人被网页布局吸引后更容易消费，我们称他们为"消费者"。对这两类人做用户画像，我们会倾向于做更多利于"消费者"（如刚刚工作几年的年轻人）点击的页面布局而不是迎合"白嫖者"（如上有老下有小的中年人）。