1. AB-Test 的概念

AB-Test 是支持产品和业务决策，实现业务持续和快速增长的必备要素。 其核心价值在于准确地预知新方案上线后对业务的影响，支撑做出正确的业务决策。

AB-Test 的核心思想是在总体样本中，先采样少量样本来估计策略的影响，并估计出策略上线后对总体的影响。 它在本质上是基于统计的假设检验过程，核心要解决在于先验性和准确性。先验性在于能否在策略整体上线前评估出对整体系统的影响。准确性在于能否在样本的数据准确评估出总体的数据。

1.1. AB-Test 的分类

AB-Test 按照实验场景的不同，可以分为不同的类型。从细分粒度来看从细到粗有流量维度、设备维度、用户维度、地域维度。

- **圈流量实验。**圈流量实验是指采样的粒度为流量维度或者说 PV (Page View) 维度，其假设是每个 PV 都当做独立的。该实验方法的优势是样本量大，数据相对稳定。其劣势是对用户的建模不足，因为相同用户的不同 PV 并非独立。圈流量实验在广告算法领域应用广泛，因为广告拍卖的维度是 PV 维度，衡量方式是 CPM (Cost Per Mille)。
- **圈用户实验。**圈设备实验是指采样粒度为用户维度或者 UV (Unique Vistor)，其假设是每个设备都当做独立的。圈设备试验的数据更能够反映对用户的影响，适合评估 App 产品或者用户策略对用户的影响。圈用户试验的样本量比圈流量实验略小，用户数据也相对不稳定，需要一些显著性统计的方法来辅助判断实验效果。
- **圈地域实验。**圈地域实验是指采样粒度为某个地域，其假设整个平台的生产者和消费者在不同地域是独立的。圈地域实验能够评价独立市场下策略对生产者消费者市场生态的影响。圈地域实验通常用于长期观察生态影响，需要更丰富的统计方法来评估对生产者和消费者双方的影响。

2. 分层实验方法

分层实验的设计初衷是为了能够同时容纳非常多的实验，从而提高产品研发和策略上线的效率。其设计思想是当每个实验策略互相独立或者影响可以忽略的时候，可以通过不同的实验之间做流量正交的方式来实现。实验策略互相独立的意思是任意两个实验不会发生直接的参数冲突，能够在同一个用户上同时命中两个实验。流量正交的意思是实验 A 的任意用户分组在实验 B 上也是随机分配用户分组的，实验 B 各个策略对实验 A 的任意用户分组都是白噪声，可以忽略。

引出分层实验的概念：就是一份流量穿越每层实验时，都会再次随机打散，且随机效果离散。

参考文献 Overlapping Experiment Infrastructure More, Better, Faster Experimentation

3. 实验数据分析工具

3.1. 统计显著性——p value

p value 是假设检验中的一个概念，而 AB-Test 也是假设检验的一个应用，p value 一般用于评估实验效果的显著性，作为实验效果是否置信的参考之一。

3.1.1. p value 的概念

p-value（简称 P 值），又称“显著性水平”。在广义上，p-value 是对某个假设的检验错误的概率，可以用来评估该检验方法的结果是否置信。在狭义上，是指在零假设为真的条件下，样本数据拒绝零假设事件发生的概率。

总结 P 值的概念和用处如下：

1. P 值可以表达的是数据与一个给定模型（也就是零假设下的模型）不匹配的程度；
2. P 值并不能衡量某条假设为真的概率，或是数据仅由随机因素产生的概率；
3. 科学结论、商业决策或政策制定不应该仅依赖于 P 值是否超过一个给定的阈值；
4. 合理的推断过程需要完整的报告和透明度；
5. P 值或统计显著性并不衡量影响的大小或结果的重要性；
6. P 值就其本身而言，并不是一个非常好的对模型或假设所含证据大小的衡量。

通常意义上的 P 值的计算方法取决于假设检验的具体方式，常用的假设检验方法有 z 检验、t 检验和卡方检验等，不同的方法有不同的适用条件和检验目标。下面分步骤介绍在 AB-Test 中如何计算 p value。

AB-Test 中是用对照组和实验组两个样本的数据来对这两个总体是否存在差异进行检验，适合使用 t 检验中的独立双样本检验。

1. 首先通过独立双样本检验的统计检验量公式计算 z score，这个 z score 就是正态分布图中横轴上的点。 假设实验组和对照组各自均值为 $\bar{x_1}, \bar{x_2}$，标准差为 s_1, s_2 ，样本量为 n_1, n_2

$$z = \frac{\bar{x_1} - \bar{x_2}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

2. 然后用 z score 查找标准正态分布表得到 p value，p value 就是正态分布大于或者于 z score 的累积概率。

从 z score 和 p value 的计算中可得，实验组和对照组的差异越大，显著性越大；样本量越充足，显著性越大；样本标准差越大，显著性越小。

3.1.2. p value 的使用误区

P 值统计结果显著=效果显著=效果的商业价值?

P 值只能代表样本数据是否与实验预期效果有多么不一致，并不能反映效果是否显著，更不能描述效果的商业价值。

例如，我们通过A/B测试对一个资源耗费10倍以上的推荐算法进行优化，得到 p 值=0.001，说明这次的试验结果是显著的。而试验的效果，只对收入提升了万分之一。这种优化实验结果显著，但是效果并不显著，商业价值同样不显著。

如果 $p < \alpha$，是否立刻能够得到结论?

p value 在部分场景下并不一定稳定，这种情况下一旦 p value 小于显著性水平，就立刻停止实验并得出结论，是有可能对实验效果判断错误的。p value 与实验效果持续稳定的情况可以相对较快的得到结论，否则应该持续观察才能得到相对准确的结论。

比如，当某 App 的某一个A/B测试为例，当实验开始后持续每天都观察 P 值。实验前 3 天 P 值并不稳定，当试验运行到第 4 天时，P 值第一次小于显著性水平，但一直到第 10 天 P 值并没有稳定下来，甚至一度增大到实验结果显示不显著。也就是说，单纯凭借 P 值来判定实验结果的显著与否是不可靠的，尤其是 P 值和实验效果并不稳定的情况下。

3.2. 置信区间

置信区间的概念

置信区间是用来对一个概率样本的总体参数进行区间估计的样本均值范围，它展现了这个均值范围包含总体参数的概率，这个概率称为置信水平。

置信水平代表了估计的可靠度，一般而言，我们采用 95% 的置信水平进行区间估计。

置信区间的计算方法

根据统计学的中心极限定理，样本均值的抽样分布呈正态分布。根据两个总体参数的假设检验方法，可以根据统计检验量公式反推获得置信区间。

AB-Test 中是用对照组和实验组两个样本的数据来对这两个总体是否存在差异进行检验。 通过独立双样本检验的统计检验量公式，反推如果显著性水平如果在 α 的双侧检验的情况下，两个样本均值之差应该在哪个范围之内。

假设实验组和对照组各自均值为 $\bar{x_1}, \bar{x_2}$，标准差为 s_1, s_2 ，样本量为 n_1, n_2。 $Z_{\alpha/2}$ 其中 α 为显著性水平，假设 $\alpha = 0.05$ ，查 z score 表得到 $Z_{0.025} = 1.96$。

$$(\delta_1, \delta_2) = (\bar{x_1} - \bar{x_2}) \pm Z_{\alpha/2} * \sqrt{s_1^2/n_1 + s_2^2/n_2}$$

3.3. 统计功效

统计功效的概念

统计功效在 AB-Test 中的意义在于，当实验组确实有比较显著统计差异的时候，发现了这个差异的概率。 在假设检验中，统计功效是指零假设为假时拒绝了零假设的概率。 为了提高零假设为假时我们做出正确判断的概率，使结果更加可靠，统计功效的值越大越好。 一般来说，当统计功效取到80%~95%时，结果就是比较可信的了。

统计功效的计算

统计功效的计算依然有假设检验的基础公式推导而来。

$$Power = \Phi(-Z_{1-\alpha/2} + \frac{|\bar{x_1} - \bar{x_2}|}{\sqrt{s_1^2/n_1 + s_2^2/n_2}})$$

其中，$\Phi(x)$ 为正态分布的累积概率密度，$-Z_{1-\alpha/2}$ 为 z score 查询得到置信区间最左侧的取值，而 $\frac{|\bar{x1} - \bar{x2}|}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$ 的意义是实际得到的 z score 取绝对值。

这样得到的统计功效在 0-1 之间，取值越大，实验的统计显著性越大。

4. 假设检验与 AB-Test

4.1. 中心极限定理

- 描述一：从总体中随机抽取一个样本量为n的样本，当n充分大时，样本均值的抽样分布近似服从正态分布。
- 描述二：多个相互独立的随机变量，他的均值（和）的分布是以正态分布为极限，也就是逼近正态分布，与随机变量的具体分布无关。

5. 假设检验的定义

假设检验是先对总体的参数提出某种假设（比如说转化率的平均值），然后利用样本数据判断假设是否成立的过程。在逻辑上，假设检验采用了反证法，即先提出假设，再通过适当的统计学方法来计算这个假设可能性的大小。

一些假设检验所需要的概念理解。

5.1. 假设检验的假设

- **零假设**：是试验者想收集证据予以反对的假设，记为H0；对比试验中的原假设就是试验版本的总体均值等于对照版本的总体均值。

- **研究假设**：是试验者想收集证据予以支持的假设，记为H1；对比试验中的备择假设就是试验版本的总体均值不等于对照版本的总体均值。

5.2. 假设检验的检验

- **检验统计量**：是据以对零假设和研究假设作出决策的某个样本统计量。
- **弃真错误**：零假设为真却拒绝零假设，错误概率记为α
- **取伪错误**：零假设为假却未拒绝零假设，错误概率记为 β
- **拒绝域**：弃真错误和取伪错误的概率，成为拒绝域。
- **统计显著性**：弃真错误的概率，称为统计显著性，也叫 p value。
- **统计功效**：不犯取伪错误的概率，称为统计功效。
- **显著性水平**：是指检验统计量落在拒绝域的概率，错误概率越低，显著性水平越小。
- **双侧检验**：是指研究假设没有特定方向，形式为$\neq$的检验假设。
- **左侧检验**：是指研究假设统计量小于某个值，形式为$<$的检验假设。
- **右侧检验**：是指研究假设统计量大于某个值，形式为$>$的检验假设。
- **置信区间**：弃真错误在一定概率范围时，样本参数的区间范围。

5.3. 假设检验的步骤

1. 提出零假设与研究假设
2. 从所研究总体中出抽取一个随机样本
3. 构造检验统计量
4. 根据显著性水平确定拒绝域临界值
5. 计算检验统计量与临界值进行比较

5.4. 假设检验的分类

- Z 检验：一般用于大样本（即样本容量大于30）平均值差异性检验的方法。它是用标准正态分布的理论来推断差异发生的概率，从而比较两个平均数的差异是否显著。
- T 检验：主要用于样本含量较小（例如n < 30），总体标准差σ未知的正态分布。T 检验是用 t 分布理论来推论差异发生的概率，从而比较两个平均数的差异是否显著。
- 卡方检验：卡方检验是统计样本的实际观测值与理论推断值之间的偏离程度，实际观测值与理论推断值之间的偏离程度就决定卡方值的大小，如果卡方值越大，二者偏差程度越大；反之，二者偏差越小；若两个值完全相等时，卡方值就为0，表明理论值完全符合。

6. 4 种主要假设检验的方法

6.1. 假设检验的分类方法

按照采样总体的不同，分为**一个总体的假设检验**和**两个总体的假设检验。** 按照统计量的不同，分为**参数的假设检验**和**成数的假设检验**

一个总体参数的假设检验：某药企为了监测某特效药对新冠肺炎是否有治愈作用，做以下检验。如果新冠肺炎的自愈时间为 72 小时，标准差 8 小时；采样 100 个患者，使用某药品之后，治愈时间为 69.6 小时。检验该药物是否有效，那么这种检验就是一个总体参数的检验。

两个总体参数的假设检验：某药企为了监测某特效药对新冠肺炎是否有治愈作用，做以下检验。将患者分为实验组和对照组分别 100 人，对照组服用安慰剂，实验组复用特效药，其他变量不变。实验组的治愈时间为 69 小时，标准差为 6 小时；对照组的治愈时间为 72 小时，标准差为 8 小时。检验该药物是否有效，那么这种检验就是一个总体参数的检验。

一个总体成数的假设检验：某 App 灰度发布新版本，为了监测某 App 新版本是否提升注册率，做以下检验。假设该 App 共有 10000 新用户安装，3000 人注册。该 App 新版本有 1000 人安装，320 人注册。检验该 App 新版本是否有效提升了注册率。

两个总体成数的假设检验：某 App 灰度发布新版本，为了监测某 App 新版本是否提升注册率，做以下检验。假设该 App 版本 1 近期有 5000 新用户安装，其中 1000 人完成注册；版本 2 近期有 1000 名新用户安装，其中 260 人完成注册。检验版本 2 是否比版本 1 有更高的注册率。

6.2. 一个总体参数的假设检验

6.2.1. 一个总体参数的假设检验的方法

一个总体参数的大样本（$n \geq 30$）假设检验方法：

假设形式

- **双侧检验**：H0 ($\mu = \mu_0$); H1 ($\mu \neq \mu_0$)
- **左侧检验**：H0 ($\mu \geq \mu_0$); H1 ($\mu < \mu_0$)
- **右侧检验**：H0 ($\mu \leq \mu_0$); H1 ($\mu > \mu_0$)

检验统计量

- σ已知，使用 z 检验

$$z = \frac{\bar{x} - \mu_0}{\sigma / \sqrt{n}}$$

- σ未知，使用 t 检验

$$z = \frac{\bar{x} - \mu_0}{s / \sqrt{n}}$$

其中，z 服从自由度为 n 的 t 分布，$z \sim t(n)$；$\bar{x}, s$ 分别为样本均值和样本标准差；μ, σ 分别为假设的总体均值和总体标准差；n 为样本量。

显著性与拒绝域

- 双侧检验: $|Z| > Z\alpha/2$
- 左侧检验: $Z < -Z\alpha$
- 右侧检验: $Z > Z\alpha$

P 值决策

如果 $p < \alpha$ ，拒绝H0

置信区间

若 σ 已知

$$\mu_0 \pm Z_{\alpha/2} * \sigma / \sqrt{n}$$

若 σ 未知

$$\mu_0 \pm Z_{\alpha/2} * s / \sqrt{n}$$

6.2.2. 一个总体参数假设检验的示例

> 某药企为了监测某特效药对新冠肺炎是否有治愈作用，做以下检验。如果新冠肺炎的自愈时间为 72 小时，标准差 8 小时；采样 100 个患者，使用某药品之后，治愈时间为 69.6 小

> 时。检验该药物是否有效，那么这种检验就是一个总体参数的检验。

1. 提出零假设与研究假设。零假设 (H0): 药物无效；研究假设 (H1)：药物有效。

$$H0: \mu \leq 72; H1: \mu > 72$$

2. 从所研究总体中出抽取一个随机样本。随机样本为进行临床试验的 100 人，平均治愈时间为 69.6 小时。
3. 构造检验统计量。平均自愈时间为 72 小时，标准差为 8 小时；样本大小为 100，平均治愈时间为 69.6 小时。

$$z = \frac{69.6 - 72}{8/\sqrt{100}} = -3$$

4. 根据显著性水平确定拒绝域临界值。求解 P 值 $p = 0.0013$，拒绝临界值 $\alpha = 0.05$。这意味着，在 H0 成立时，出现 69.6 小时以及更极端情况的概率为 0.3%
5. 计算检验统计量与临界值进行比较。$p < \alpha$，因此拒绝零假设，药物有效的概率为 99.87%。

6.3. 两个总体参数的假设检验

6.3.1. 两个总体参数的假设检验的方法

一个总体参数的大样本（$n \geq 30$）假设检验方法： **假设形式**

- **双侧检验**：H0 ($\mu_1 - \mu_2 = 0$); H1 ($\mu_1 - \mu_2 \neq 0$)
- **左侧检验**：H0 ($\mu_1 - \mu_2 \geq 0$); H1 ($\mu_1 - \mu_2 < 0$)
- **右侧检验**：H0 ($\mu_1 - \mu_2 \leq 0$); H1 ($\mu_1 - \mu_2 > 0$)

检验统计量

$$z = \frac{(\bar{x_1} - \bar{x_2}) - (\mu_1 - \mu_2)}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$$

其中，$\bar{x_1}, \bar{x_2}, s_1, s_2$ 分别为样本均值和样本标准差；$\mu_1, \mu_2, \sigma_1, \sigma_2$ 分别为假设的总体均值和总体标准差；n 为样本量。

显著性与拒绝域

- 双侧检验: $|Z| > Z\alpha/2$
- 左侧检验: $Z < -Z\alpha$
- 右侧检验: $Z > Z\alpha$

p 值决策

如果 $p < \alpha$ ，拒绝H0

置信区间

$$(\mu_1 - \mu_2) \pm Z_{\alpha/2} * \sqrt{s_1^2/n_1 + s_2^2/n_2}$$

6.3.2. 两个总体参数的假设检验的实例

> 某药企为了监测某特效药对新冠肺炎是否有治愈作用，做以下检验。将患者分为实验组和对照组分别 100 人，对照组服用安慰剂，实验组复用特效药，其他变量不变。实验组的治

> 愈时间为 69 小时，标准差为 6 小时；对照组的治愈时间为 72 小时，标准差为 8 小时。检验该药物是否有效，那么这种检验就是一个总体参数的检验。

1. 提出零假设和研究假设。假设 μ_1和μ_2 分别为实验组和对照组的期望治愈时间。

$$H0: \qquad \mu_1 \geq \mu_2$$

$$H1: \qquad \mu_1 < \mu_2$$

2. 随机样本分别是实验组和对照组。

$$\bar{x_1} = 69; s_1 = 6; \bar{x_2} = 72; s_2 = 8; n_1 = 100; n_2 = 100;$$

3. 计算检验统计量。

$$z = \frac{(69-72)-(72-72)}{\sqrt{6^2/100+8^2/100}} = -3$$

4. 计算显著性；拒绝域临界值确定为 0.05

$$p = 0.0013$$

5. 比较显著性和拒绝域临界值，决策检验结果

$$p < 0.05$$

拒绝零假设。该特效药有效的概率为 99.87%

6.4. 一个总体成数的假设检验

6.4.1. 一个总体成数的假设检验的方法

样本成数：它是指样本中具有某一相同标志表现的单位数占样本容量的比重,记为 p.

总体成数：它是指总体中具有某一相同标志表现的单位数占全部总体单位数的比重,一般用 π 表示.

假设形式

- **双侧检验**：H0 ($\pi = \pi_0$); H1 ($\pi \neq \pi_0$)
- **左侧检验**：H0 ($\pi \geq \pi_0$); H1 ($\pi < \pi_0$)
- **右侧检验**：H0 ($\pi \leq \pi_0$); H1 ($\pi > \pi_0$)

检验统计量 T 检验

$$z = \frac{p - \pi_0}{\sqrt{\pi_0(1-\pi_0)/n}}$$

其中，p 为样本成数，π_0 为总体成数，n 为样本量。

显著性与拒绝域

- 双侧检验: $|Z| > Z\alpha/2$
- 左侧检验: $Z < -Z\alpha$
- 右侧检验: $Z > Z\alpha$

P 值决策

如果 $p < \alpha$ ，拒绝H0

置信区间

$$\pi_0 \pm Z_{\alpha/2} * \sqrt{\pi_0(1-\pi_0)/n}$$

6.4.2. 一个总体成数的假设检验的实例

> 某 App 灰度发布新版本，为了监测某 App 新版本是否提升注册率，做以下检验。假设该 App 共有 10000 新用户安装，3000 人注册。该 App 新版本有 1000 人安装，320 人注册。检验该 App 新版本是否有效提升了注册率。

1. 提出零假设和研究假设。零假设为该 App 版本对注册率没有显著提升；研究假设为该 App 版本对注册率有显著提升。假设总体注册率为 $\pi_0 = 3000/10000 = 0.30$，样本注册率为 p。

$$H0: \qquad \pi \leq 0.3$$

$$H1: \qquad \pi > 0.3$$

2. 随机样本为新版本的 1000 个用户，注册人数为 320。

$$p = 320/1000 = 0.32$$

3. 计算检验统计量。

$$z = \frac{0.32 - 0.30}{\sqrt{0.3 * (1-0.3)/1000}} = -1.3801$$

4. 计算显著性。

$$p = 0.0838$$

5. 比较显著性和拒绝域临界值，决策检验结果

$$p > 0.05$$

因此，接收零假设，该 App 版本对注册率没有显著提升。

6.5. 两个总体成数的假设检验

6.5.1. 两个总体成数的假设检验的方法

假设形式

- **双侧检验**：H0 ($\pi_1 = \pi_2$); H1 ($\pi_1 \neq \pi_2$)
- **左侧检验**：H0 ($\pi_1 \geq \pi_2$); H1 ($\pi_1 < \pi_2$)
- **右侧检验**：H0 ($\pi_1 \leq \pi_2$); H1 ($\pi_1 > \pi_2$)

检验统计量

$$z = \frac{(p_1 - p_2) - (\pi_1 - \pi_2)}{\sqrt{p(1-p)/(1/n_1 + 1/n_2)}}$$

其中，

$$p = \frac{p_1 * n_1 + p_2 * n_2}{n_1 + n_2}$$

p_1, p_2 分别为两个样本成数，n_1, n_2 分别为两个样本量，p 为两个样本的合并成数，

显著性与拒绝域

- 双侧检验: $|Z| > Z\alpha/2$
- 左侧检验: $Z < -Z\alpha$
- 右侧检验: $Z > Z\alpha$

P 值决策

如果 $p < \alpha$ ，拒绝H0

置信区间

$$(\pi_1 - \pi_2) \pm Z_{\alpha/2} * \sqrt{p(1-p)/(1/n_1 + 1/n_2)}$$

6.5.2. 两个总体成数的假设检验的实例

> 某 App 灰度发布新版本，为了监测某 App 新版本是否提升注册率，做以下检验。假设该 App 版本 1 近期有 5000 新用户安装，其中 1000 人完成注册；版本 2 近期有 1000 名新用户安装，其中 260 人完成注册。检验版本 2 是否比版本 1 有更高的注册率。

参考答案：

1. 提出零假设与研究假设。零假设 (H0): B 组实验没有显著好于 A 组；研究假设 (H1)：B 组实验显著好于 A 组。假设实验组和对照组的注册率均值为 μ_1, μ_2

$$H0: \qquad \mu_1 \leq \mu_2$$

$$H1: \qquad \mu_1 > \mu_2$$

2. 从所研究总体中出抽取一个随机样本。随机样本为 A 组和 B 组分别 1000 人，样本成数分别为 0.3 和 0.32，总体成数为 0.31
3. 构造检验统计量。

$$z = \frac{0.32 - 0.3}{0.31 * (1 - 0.31) / \sqrt{1/1000 + 1/1000}} = 0.0042$$

4. 根据显著性水平确定拒绝域临界值。求解 P 值 $p = 0.4840$，拒绝临界值 $\alpha = 0.05$。这意味着，在 H0 成立时，转化率大于或等于 0.32 的概率为 48.4%
5. 计算检验统计量与临界值进行比较。$p > \alpha$，因此无法拒绝零假设，实验显著性水平不足，无法证明 B 组实验显著好于 A 组。