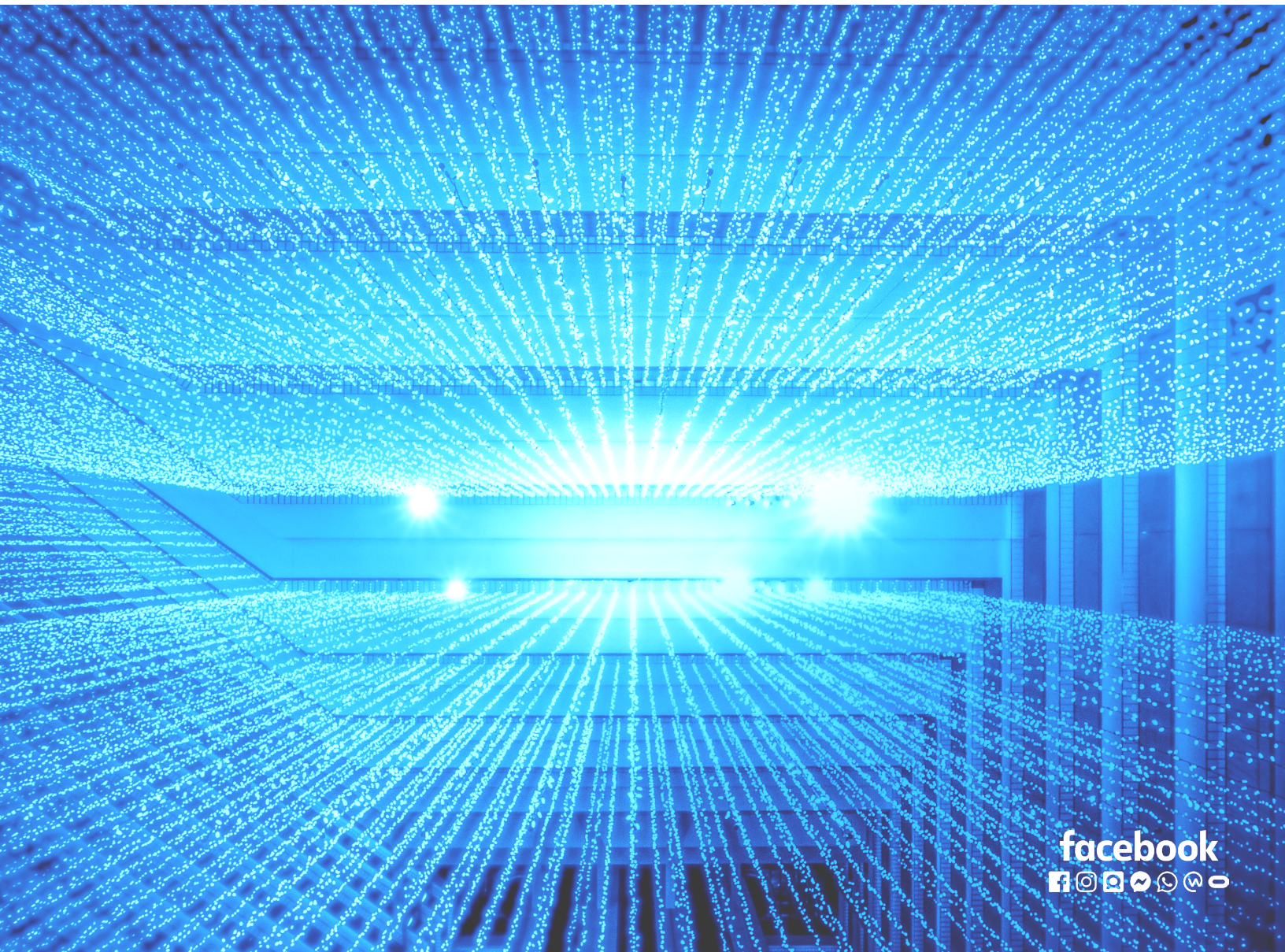
2019 年 3 月

未雨绸缪

增效衡量测试和经验总结指南



目录

编著者

Alok Gupta，Lyft
数据科学部总监

Matthew Gerrie，Booking.com
营销科学与公关部高级总监

Stephan McBride，Netflix
科学与分析部及营销与经济学部总监

Tony Flanery-Rye，eBay
增长分析部高级总监

Dan Johns，Facebook
产品营销经理

Jesse Goranson，Facebook
客户衡量部总监

Maggie Burke，Facebook
客户委员会主管

Sophia Lin，Facebook
项目经理

简介

您现在或许已经准备好开始测试不同的营销策略，但是否已准备好应对测试中可能出现的所有意外情况？

在“衡量营销成效：增效衡量应用指南”中，我们介绍了如何衡量营销策略的真正价值。借助增效衡量，营销人员可以通过隔离其他营销策略、业务因素和变量的方式来衡量特定营销策略的真正价值。

在营销高手看来，实验法是衡量策略价值的最有效方法。从本质上讲，实验会受到外部因素和内部假设的双重制约。同时，实验结果也是难以预测的，有时甚至会出错。即便如此，这些测试的结果仍然证明是有用的，能够提供有价值的经验。

我们与业内顶尖的成效衡量专家合作编写了这份指南，旨在介绍营销实验中可能出现的错误并提供相应的解决思路。这样，即便在测试过程中可能遇到困难，营销人员仍然可以充分利用增效衡量的益处。

> **Netflix 科学与分析部及营销与经济学部总监 Stephan McBride** 说：“很多事情会出错，我们需要始终做出这样的假设和预期。一定要在这方面做好准备，因为增效衡量的成功意味着实验证据将在业务运营中发挥巨大的影响力。”

好消息是，实验中可能出现的错误往往是可以预测的，这意味着营销人员可以提前准备好应对措施。更令人欣慰的是，实验中遇到的典型问题对营销人员具有指导意义，甚至还能扭转乾坤。事实上，在过去几年里大获成功的品牌都推崇一条共同的理念：积极建立“从测试中学习”的企业文化。他们发现这是推动企业发展和提升竞争力的最有效方式。

在准备应对可能出现的错误时，营销人员需要知道无论发生什么情况，一些实验都是有价值的，这样才能做好应对各种挑战的准备。

> **eBay 增长分析部高级总监 Tony Flanery-Rye** 说：“我经手的大型实验几乎都出现过失误，零失误的情况少之又少。但出现错误很正常。我们需要积极去适应。”

不断测试并从经验中学习，将这一准则纳入您业务的核心部分，能为您的企业创造卓越的营销表现。

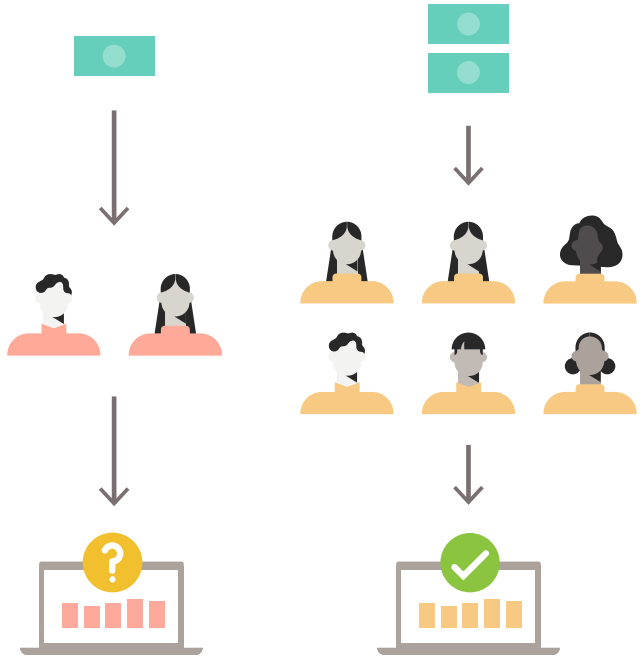
问题 1

测试的统计功效不足

开展增效衡量测试时，拥有合适的数据量至关重要。

数据样本容量太少可能会有损结果的准确性，而太多又可能导致成本高昂。做出商业判断时，必须平衡测试所需的信息和开展更大规模、时间更长的营销活动所需的成本。归根到底，测试需要有足够的统计功效。

例如，在开展实验时，应确保设置的测试能足够准确地衡量您要检测的变量。揭示较小差异所需的数据更多，反之则更少。即使您预先安排了一项测试来收集足够的数据，但依据历史数据制定的计划在实际测试期间也可能行不通。事实上，并非所有的测试都具有足够的统计功效来检测营销效果的变化。



为什么会这样呢？

导致统计功效不足的因素有很多：

测试组和对照组之间的差异不足。
有时候，您所设置的测试组和对照组之间根本就没有区别或差异。这种情况下，即使您收集了再多数据，都不太可能发现变化。

实验结果与根据以往实验所预期的结果不同。
测试出的结果可能不及您根据以往经验所预想的结果。出现这种情况，可能是由于您之前预计所需的数据较少，导致测试没有收集到足够的样本来检测实际的效果。这意味着实验可能无法检测到一些实际存在的影响。

外部因素导致无法收集到预期的数据量。
外部因素可能导致广告无法按预期投放或影响关键表现指标（如转化率、销量、应用安装量或店铺回访量）。这可能会影响您所收集的数据量，并导致测试功效不足。

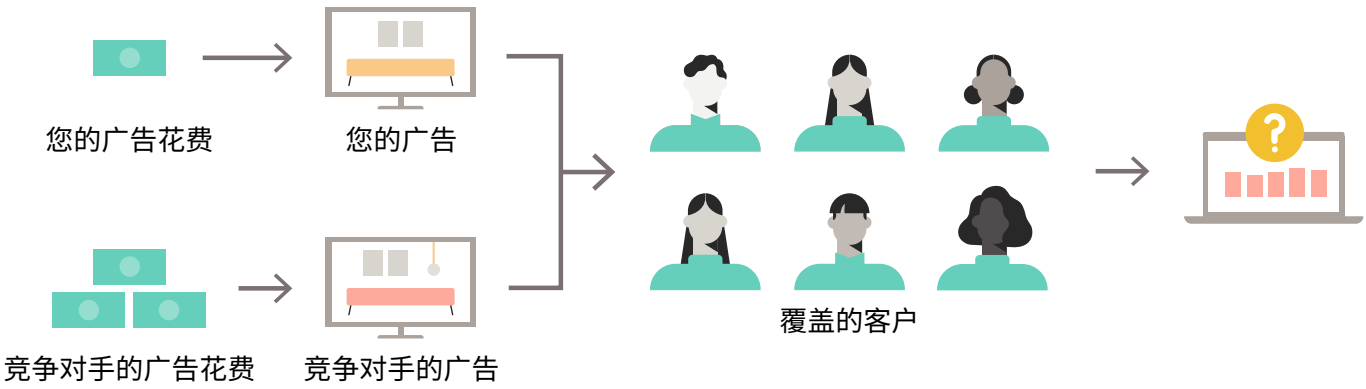
缺乏足够数据开展特定实验。
如果您在针对特定的受众（如客户列表或使用某项功能的用户）进行测试，可能永远无法找到足够多的人来检测效果。

测试的结果并不经常发生。
有些结果发生的频率很低，因此很难收集足够的数据来获得具有统计学意义的结果。

漫长的购买周期导致难以收集足够的数据。
对于需要较长考虑周期才能得到的结果，实验的时间长度可能未跟上，导致无法检测到实际影响。诸如 Cookie 污染等问题和其他形式的错误也会削弱测试效果的准确性，从而降低统计功效。

示例

某家初露头角的床垫直销商建立了一个客户数据库，其中的数据主要来自于社交媒体广告。他们现在正准备首次测试电视广告花费，以衡量其对销售业绩的影响。但就在该直销商准备在一些城市推出新的电视广告，并在对应市场开展成效测试之际，另一家初创的床垫销售商也开始投放电视广告，他们的广告投入要高得多，因此覆盖的消费者也明显更多。看到测试结果后，这家直销商认识到他们的测试并没有足够的统计功效来检测营销效果的变化。电视广告有用吗？是否被竞争对手的广告淹没了？



如何应对这些问题？

在上述示例中，缺乏统计功效并不一定意味着测试完全无用。事实上，这些测试结果仍然可能具有利用价值。确定这些测试结果是否有用的方法之一是功效计算，即计算调研能够检测出实际提升效果的概率。这项重要指标能够表明是否有足够的数据来报告可靠的结果。

功效计算的目的是在一定程度上可靠地检测超过特定比例的效果（例如检测大于 x% 的效果）。但没有检测到效果并不意味着没有效果。如果没有收集到足够的数据，则需要相应地更改阈值。即使没有如预期中检测到超过特定数值的成效，也并不意味着没有比这数值更小的成效。

除了衡量您是否有足够的数据来确定影响之外，这样的测试还有其他指导价值。例如，您需要得出特定比例的成效才能确定某个营销策略能带来利润，而实际的测试结果却并不能满足这个要求，那么这一结果同样对您有指导作用，您仍然可以据此做出决策。

或者，如果您在比较营销方案，则测试结果可以让您做到心中有数。虽然调研的功效可能没有您想的那么高（一般来说应该大于 80%），但您仍可以依据这些信息制定决策。当在两个或多个策略执行方案中进行选择时，在其他条件都相同的情况下，建议您选择最有可能带来更多增量的方案。

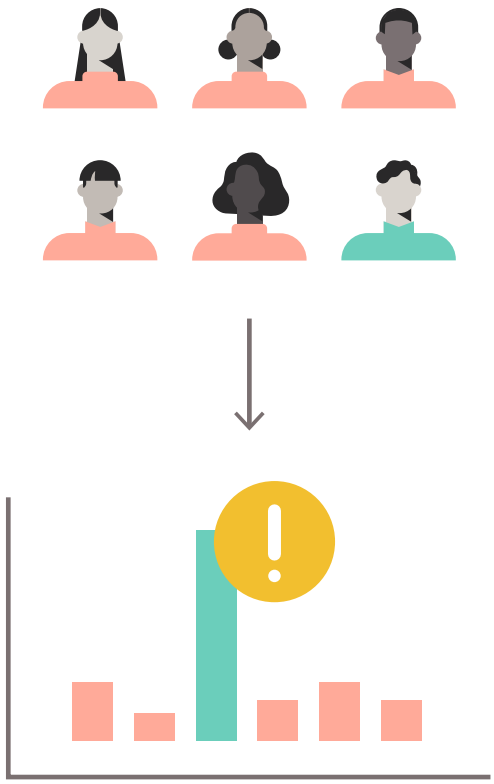
问题 2

测试组出现异常值

增量测试的效果本质上依赖于对比两个或多个试验组。

营销实验中的随机分配有助于确保各个试验组在统计方面具有相似性。然而，异常值会使对比出现问题。异常值可以定义为表现远高于或远低于标准水平的测试对象。

异常值会导致被测试的两个试验组存在明显差异，因而不具备可比性。虽然异常值是一个正常现象，且在任何数据集合中都会存在，但测试中的异常值却可能会显著影响您的成效评估。对于规模较小的测试，如果其中一个试验组中存在一个显著异常值，便会导致结果出现偏差。



为什么会这样呢？

首先，请先确定业务中存在这些异常值的原因：

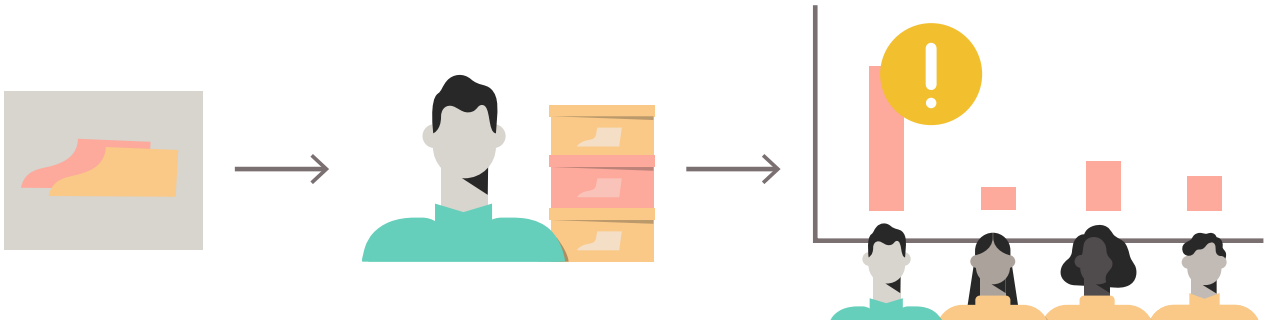
异常值是业务的驱动因素。
许多采用免费增值模式运营的企业，比如游戏公司和一些 SaaS 公司，都是围绕着低频大客户建立起来的。游戏公司通常称这些客户为“鲸鱼”，因为他们对营收的贡献占比远高于普通玩家。虽然 SaaS 公司可能会以月费的形式向个人销售软件，但是企业客户可能会为数百名员工购买该软件的访问权限，单月花费便达到数十万美元。

您可能同时拥有普通消费的个人客户和大额消费的企业客户。
您的客户群中包括低频低价值交易的个人客户和频繁或高价值交易的企业客户，这些数据集合之间存在显著的分配不平衡。

销售数据会受到随机事件的影响。
测试期间发生的随机事件会导致人们的行为与平时不同。例如，某旅行社可能有一个普通的企业客户要召开销售会议，一下子向他们预订 7,000 名与会者的机票。

示例

一家销售定制运动鞋的小型零售商正在开展节日促销，他们希望在获得大量新客户的同时，衡量促销活动的效果。某公司的首席执行官看到广告后，决定为全体员工订购运动鞋作为节日礼物。乍一看，这场营销活动似乎取得了巨大成功。然而，经过进一步的调查，这家零售商意识到，他们的销量主要来自同一个买家。



如何应对这些问题？

从一开始就要保持警惕。确保让您的分析或数据科学团队调查是否存在异常值。虽然您可以在计算统计功效前为异常值做好准备，但我们还是建议您制定一套策略，以便即时发现异常值并为之作出相应调整。

“缩尾处理”是一种常用的统计策略，通过输入更常见的值来手动调整异常值。这种方法通过将异常值的数据隔离到指定的百分位数来区分异常值。

举例来说，发现营销测试被异常值数据扭曲后，公司可能会选择使用销售额中值来代替前 0.01% 的客户的实际销售额，因为这些数额太大，会对结果产生显著影响。

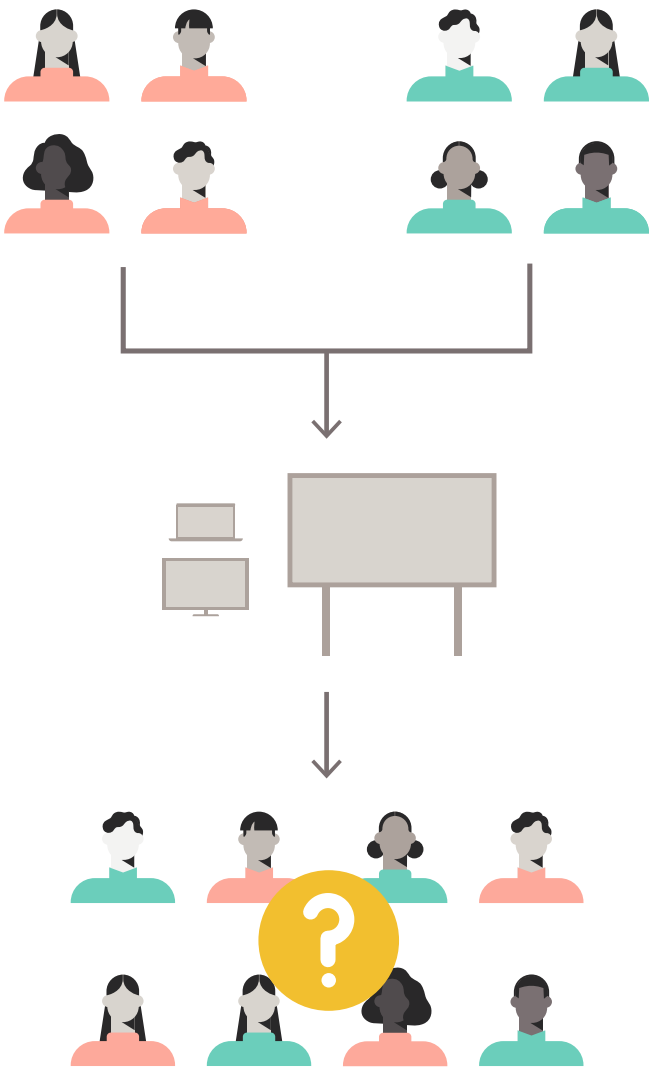
当异常值不是业务的常见情形或并非由错误引起时，这种方法很有效。如果异常值不是由合理的业务原因引起的（例如因错误而导致异常值），则可能需要使用其他方法，完全剔除异常值。

问题 3

测试中的变量不独立

隔离变量能让您准确理解在实验中观察到的变化是由哪些原因引起的。

但有些变量是可以控制的，有些是无法控制的，还有一些是无法预见的。尽管您费尽心血，人们也可能在测试过程中接触到其他相关信息，导致您的变量不独立。



为什么会这样呢？

导致特定测试变量无法被恰当隔离的原因有很多：

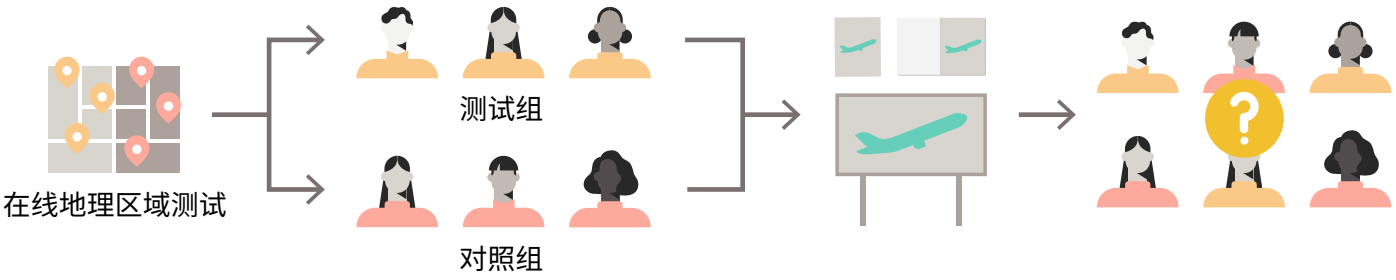
由于没有统一协调测试而导致错误。
参与投放媒体广告的各个团队可能没有在整个实验中保持所有其他变量相同。他们可能调整了某个试验组的竞价、预算、创意或其他执行细节，而没有调整其他试验组的同一变量。

未能控制公司的所有媒体广告计划。
由于伙伴团队出乎意料地投放与您的试验组不同的媒体广告或推广项目，导致您的试验组中的人群可能同时接触到多个变量。值得注意的是，您不应该在开展测试时停止其他渠道的营销。事实上，我们建议您在与平时相同的情况下开展测试，这意味着，您要在正常情况下进行客户关系管理、开展效果类营销及其他活动。关键是确保这些活动在对照组和测试组中保持一致。

如果测试组中的成员能影响对照组中的成员，就会降低测试的受控程度。
如果一个试验组中的用户会影响另一个试验组中的用户的行为，就不能真正体现测试未受到干扰。例如，某本图书数量有限，测试组中的一位客户购买了一本，这可能会导致图书缺货，令对照组中一个也想购书的人无法购买。

示例

一家旅游公司正在投放在线视频广告 (Video Ads)，吸引人们关注他们的夏季促销活动。为了测试这次营销活动的效果，该公司的营销人员决定开展一项在线地理区域测试，排除特定邮政编码或 DMA（指定营销地区）的人群。由于不知道正在开展的测试，该公司的户外媒体经理面向某些城市投放了户外广告。这使得对照组中的人群也接触到了夏季促销广告信息，而测试的原计划是不让他们看到这些信息。由于测试中的变量不再独立，因而难以了解这次营销活动的真实效果。



如何应对这些问题?

虽然变量可能没有完全隔离，但这并不意味着测试结果没有用。如果差异很小（如执行方面的细微变化），而您所观察的影响足够大，就可以放心地忽略这个误差。

如果误差很大，测试结果仍然可以利用，但是您可能必须改变对测试的解读，或者将导致偏差的因素隔离。因为您无法有把握地将变化的原因仅仅与您在实验中最初设计的变量直接关联。

如果您想要重新测试，那么此时您已经更充分地了解了预期的效果或可能出现的执行误差。

问题 4

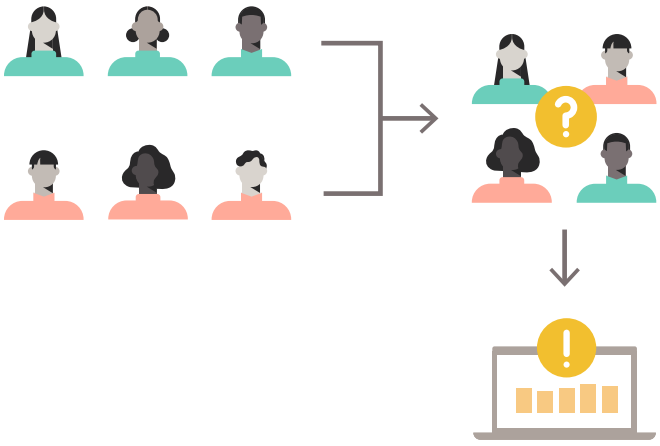
测试组和对照组中的人员出现交叉

对于任何实验而言，关键要素之一是在衡量的所有设备和平台中，确保每个测试对象在整个测试过程中都处于被指定的测试组中。

如此一来，不应该接触测试内容（例如创意测试）的人员就不会看到，而应该看到的人员则能按照您所计划的节奏看到相应内容。

但在现实操作中并非总能达到这种理想状态。人们的行为是不可预测的。例如，孩子可能会用父母的手机看视频，这时原本向其父母投放的新车广告却被他们看到了。或者快餐连锁店在某些市场推广一种新的早餐三明治，并测试新的当地媒体宣传策略，但该国其他地区的一些人正好到这里来旅游，并看到了这些测试广告，即使这种产品尚未在他们居住的地方推出。

在测试过程中，有些人可能会接触到多种测试内容。或者，对照组中的人员也可能接触到测试内容。最后，还有些人可能因为被移到对照组而没有看到内容，从而减弱了信号。



为什么会这样呢？

下面这些因素可能导致测试组和对照组在不经意间出现交叉：

识别客户的方法不可靠或不稳定。
您可能使用的是某种形式的标识符，如登录或网站 Cookie，而这种标识符（对个人或家庭）不是唯一的，同一名客户可能会有多个标识符。如果同一用户的身份随着时间发生变化，则他 / 她可能会看到多种测试内容。例如，某人可能被随机分配到桌面端的测试组中，但稍后则有可能被随机分配到移动端的对照组中。因此，在没有设置跨设备匹配的情况下，同一个人可能接触到多种实验条件。

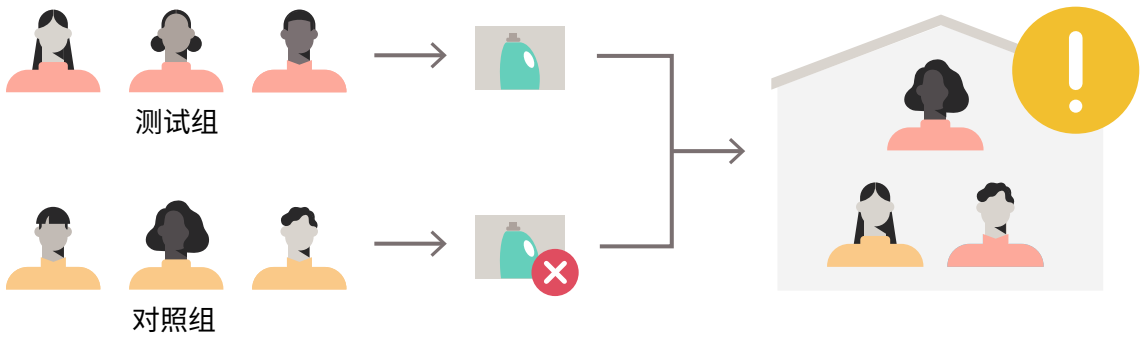
对照组中的人员与组外人员有交流。
参与测试的人员在看到您的广告后，可能会将相关的商品或服务告诉他们的朋友。这种情况对于娱乐类或其他以活动为主的营销而言尤为常见。这种口碑营销意味着，那些正常情况下本不应该看到广告的人员仍可能会以测试中没有衡量的方式接触到广告内容。

其他人为接受测试的对象代买商品或服务。
您可能按照某种层级（例如按个人）来划分试验组，但有些产品是在家庭层级购买的（例如包装消费品或保险产品）。家庭成员可能被分在多个试验组中，因此可能会看到多种测试内容。

有时您会看到一些意外的变量。
测试的分组可能是公平的，但一些您无法控制的情况可能只会发生在其中一个组内的人员身上。例如在开展地理区域测试的地区发生了自然灾害。

示例

一家销售各种家居用品的包装消费品公司正在进行一项测试，以观察某新品牌洗衣液的在线广告效果。他们把参与测试的每个人员都作为一个独立的研究对象，并划分了测试组和对照组。随着广告的投放，营销团队很快意识到洗衣液实际上是在家庭层面上购买的，而且测试组和对照组中的人员可能是一家人。



如何应对这些问题？

在评估实验时，需要了解测试组和对照组发生交叉时，会以何种方式以及在多大程度上影响实验结果。通常而言，交叉现象会干扰观察到的效果，使测试的增量不明显，或弱化试验结果之间的差异。

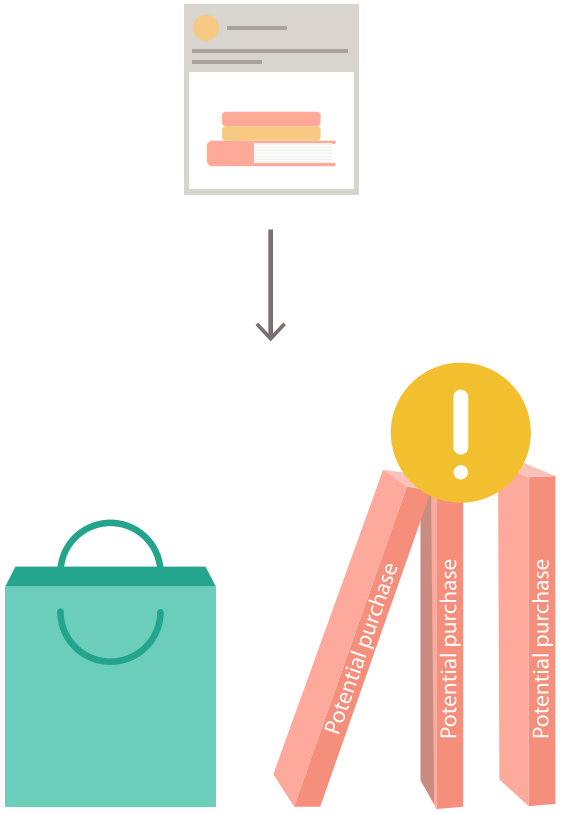
如果您仍然发现存在很强的增量结果或组间差异，则可以利用试验结果。您可以将这些数据视为测试产生的实际效果的下限。如果您认为受不稳定性影响的人数太多而无法观察到效果，请考虑使用其他测试方案。

问题 5

一些测试的影响超出最初的用户互动

在分析实验时，一个常见的问题是没有考虑用户在看到广告后的行为产生的效应，以及这种后续效应对其他潜在购买行为产生的影响。这种现象被称为二阶效应。

除了衡量直接的用户操作之外，评估这些二阶效应也非常重要。例如，因参加测试而采取的操作可能会影响客户的终生价值，或者影响他们是否购买其他相关的产品类别。这些场景可能改变您评估营销成效的方式。



为什么会这样呢？

很多场景都可能发生二阶效应：

测试可能对客户的终生价值产生连锁反应。
一次实验可能影响客户的综合价值和期望价值。未来的购买量、客户的质量或回报质量都可能影响您判断哪种营销策略能够为公司带来更多价值。

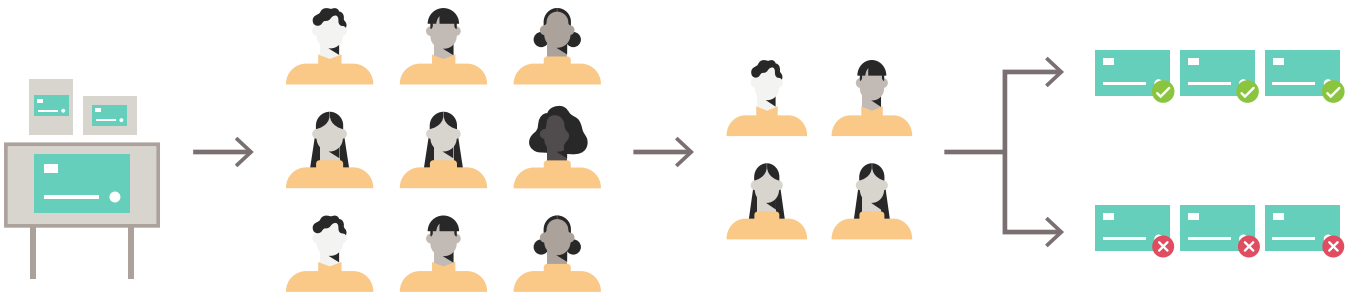
测试还可能影响其他产品类别。
测试可能会影响客户购买您销售的其他产品。如果不将公司的产品组合考虑进来，则可能会导致您高估或低估测试的价值。

测试可能引导客户进入特定的购买渠道。
某些测试可能吸引人们通过特定渠道购买商品，例如在线下实体店而不是线上渠道。虽然这从单一的渠道来看是成功的，但从整体上看，可能并没有创造多少价值。

对于一些迟早都会转化的客户，测试只是提前吸引他们发生转化而已。
您的测试可能会让那些迟早都会转化为客户的人更早完成转化。通过促销活动、抵用券和优惠券吸引的客户尤其如此。这会“抬高”广告的实际效果。

示例

某金融服务公司正在测试一种推广信用卡的新营销策略。在测试过程中，营销团队变得紧张起来，因为批准的信用卡数量看起来非常少，这让他们认为这场营销会“打水漂”。不过，他们很快发现一开始没有考虑到，即使消费者积极响应了营销信息，其中一部分人却无法通过信用卡审批。



新的营销策略　已覆盖的消费者　申请者

如何应对这些问题？

要应对这类问题，您应该认真考虑将什么指标视为成效（例如是“申请量”还是“批准量”），并将这些关键表现指标纳入测试设计和数据评估中。处理二阶效应的最佳方法是理解您所做的假设，尽量直接衡量下游的影响，同时确保全面考虑与功效有关的问题。事实是，即使测试表明某个策略或广告是成功的，仍然会存在一些您无法衡量的“死角”。大多数业内领先的成效衡量公司每次进行实验时，都会采用上百个指标来衡量测试对其他关键表现指标的影响。

第一步，了解您的假设。您是否假设客户的终生价值是相同的？您是否假设客户会在同一渠道购买商品？您是否考虑了最主要的客户消费流程，以便确定您需要做的假设？这有助于您了解测试中可能存在的盲点，以及消除这些盲点的方法。

在可能的情况下，您也应该在测试期间衡量二阶效应。例如，您想提升线上销量，但同时也关注线下销量，那就可以同时衡量二者（尽管二阶效应的统计功效可能较低）。如果您的交易具有不同的价值，那么除了交易数量之外，还应评估订单大小或终生价值。如果您认为测试可能让客户更快完成转化，则应确保衡量周期足够长，并寻找稳定的指标变化。

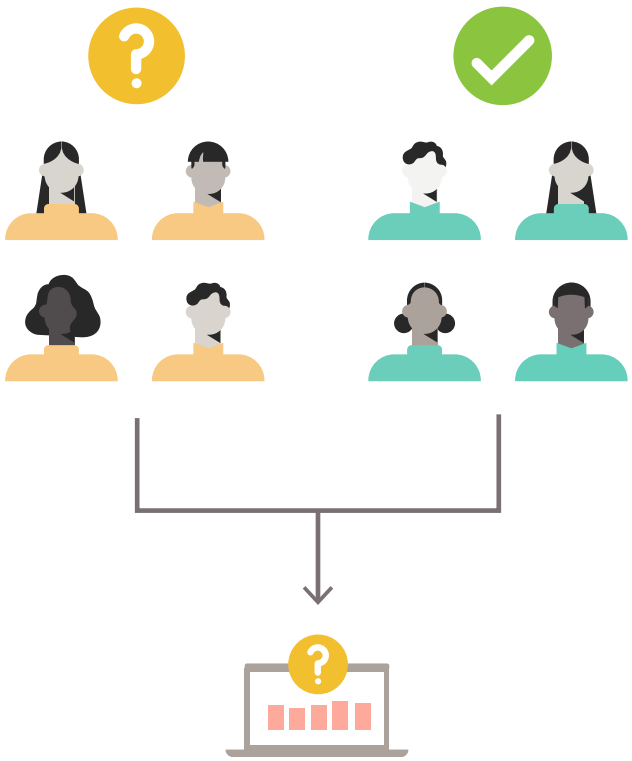
问题 6

无法追踪想要了解的各项指标

正如一些测试可能会受到不稳定因素的干扰一样，也有一些测试存在不可避免的未知因素。

人们可能并不总是活跃在您开展测试的平台上，而且不会总是按照您的愿望或需求来提供数据。

例如，当您开展实验时，在分析阶段您需要知道某位用户位于哪一个试验组。如果相关结果以匿名方式呈现，或者在实验平台之外的其他平台上出现，这可能就会非常棘手。



为什么会这样呢？

测试通常会受到难以追踪的变量的影响：

客户用现金购物，或存在其他无法追踪的购买行为。
有些结果无法与个人或其他形式的身份关联起来。例如在没有会员计划的零售店内进行的现金交易。

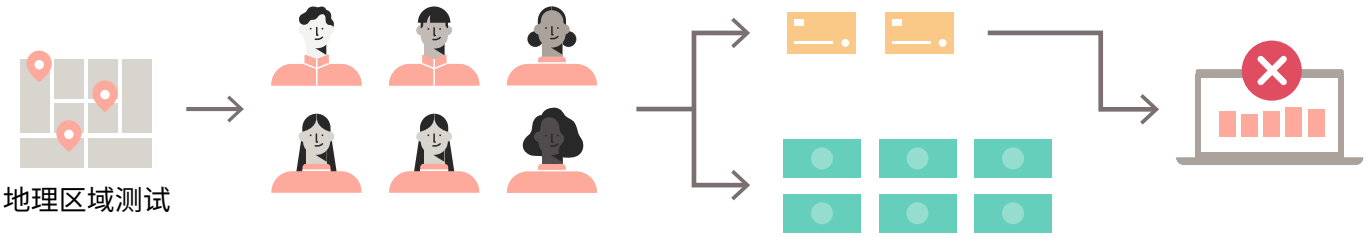
将所有用户编号一一匹配是不可能实现的。
将数据导入某个平台进行衡量时，您可能会依赖于采用不同标识符的匹配系统。由于诸多原因，比如使用多个电子邮箱或缺少用于匹配的通用标识符，您可能无法在平台之间对结果进行 100% 匹配。

人们的浏览行为使匹配难以实现。
将结果与各试验组进行在线匹配需要稳定的标识符。用户行为（如切换使用设备）或浏览器行为（如删除或禁用 Cookie）会使标识符变得不稳定。

您很难吸引足够多的人响应品牌调研。
要获得某些形式的结果（如品牌影响），采用问卷调查方式最为合适。然而，问卷调查的回复（即便是用户为获得奖励而作出的回复）可能是不完整的，因而无法追踪。

示例

某家快餐店正在开展地理区域层级的测试，以便了解其媒体广告对周围社区的影响。然而，随着测试的开展，营销人员很快意识到，由于没有可用的数据，占比较高的现金交易很难追踪。



地理区域测试

如何应对这些问题？

首先，务必要从两个方面来理解这对您的业务有多大的影响。

1. 无法追踪的交易占比有多大？

2. 测试组和对照组中的比例是否相同？

如果无法追踪的交易占比很低，而且这种影响在各个试验组中大致相同，则可以忽略是否可追踪带来的影响。这种情况通常会发生在基于测试结果制定优化决策时。而确定策略价值或投资回报时则不常发生。

解决这种问题最常见的方法是：通过调整来纠正覆盖范围不足所产生的偏差。这样一来，您将能获得与可追踪情形下相似的数据。这种策略对于诸如问卷调查之类的追踪方式而言很重要，也可以用于调整其他结果。虽然这种情况下，您必须假设可追踪人群和不可追踪人群之间的相似性，以及他们的匹配率，但该方法能让您更加准确地了解整体效果。

如果大多数交易都无法追踪，那么建议您采用实验性策略，并将之与笼统一些但可以追踪的测量单元相匹配。这些高一阶层的测量单元是您退而求其次的最佳选择。例如，如果店内以现金交易为主，则您可以开展地理区域测试，以店铺为测量单元，收集店铺层级的可追踪数据。再举一例，当某个测试难以收集个人数据时，品牌可以选择追踪区域影响数据，例如某个邮政编码归属地。

总结

测试中出现差错或误差在所难免，重点是您在这个过程中能不断学习，并总结经验教训从而优化实践。毕竟，这才是您开展测试的最终目的。

正如我们在本指南中所讲的，由于营销测试中的许多问题都是常见并且可预测的，因此您将能提前知道需要注意的问题，以及这些问题的妥善解决方法。此外，如果处理得当，即使是有缺陷的测试也可以产生价值。最成功的营销人员也会经常遇到问题，但他们能在这个过程中不断地测试和学习，继而提升自我，精进技能，获得成长。

> **Booking.com 营销科学与公关部高级总监** Matthew Gerrie 说：“仅仅依靠实验并不能保证能得到完美的答案，因为现实世界总是会与理想的实验环境有些许差异。重要的是知道哪些地方可能出错，明白您在开展测试时愿意放弃或忽略哪些因素，以及哪些情况会严重到需要您开始或重启测试。有两种情况一定要分清楚：一是在测试中可以忽略的错误，二是让您必须停止测试的错误。”

在开始测试之前制定一份妥善计划，预见一些可能出错的关键方面，将会为您的企业奠定成功的基础。

> **Lyft 数据科学部总监 Alok Gupta** 说：“不要害怕出错，因为认真从错误中学习和总结经验教训是公司所乐见的负责任行为。不断学习对公司发展至关重要。”

实验方法能为企业的营销带来深远影响，您将发现，为此遇到任何困难和挫折都是十分值得的。