可信实验白皮书

（方法指南篇）

美团履约 & 外卖团队资深数据科学家撰写
根据多年 AB 实验设计与评估经验
系统阐述了 AB 实验的基础原理与应用案例

美团

前言

为什么要写 AB 实验白皮书?

增长与优化是企业永恒的主题。面对未知的策略价值，数据驱动的 AB 实验已经成为互联网企业在策略验证、产品迭代、算法优化、风险控制等方向必备的工具。越来越多的岗位，如数据科学家、算法工程师、产品经理以及运营人员等，要求候选人了解 AB 实验相关知识。然而，许多从业者由于缺乏有效的学习渠道，对 AB 实验的理解仍停留在初级阶段，甚至存在一些误解。我们希望通过系统性地分享和交流 AB 实验的理论基础、基本流程、核心要素及其应用优势，能够帮助更多相关人员深入了解实验，提升实验文化的普及度，最终辅助企业在更多领域做出精确数据驱动决策。

除了广泛传播实验文化外，该白皮书在深度上也可给实验研究人员，提供复杂业务制约下进行可信实验设计与科学分析评估的参考经验和启发。从美团履约技术团队、美团外卖业务的实践来看，实验者常常面临多种复杂的实验制约和难题，例如，在美团履约业务中，实验往往需要应对小样本、溢出效应（即实验单元间互相干扰）以及避免引发公平性风险等多重约束，需设计科学复杂的实验方案以克服相应挑战。通过撰写白皮书，我们系统性地总结和分享应对复杂实验约束的研究经验，进而能够促进实验技术的传播与升级，推动实验科学持续进步。

本白皮书以 AB 实验为中心，涵盖 AB 实验概述与价值、实验方法基础原理与案例剖析以及配套 SDK 代码分析等，内容丰富且易于理解和应用。适合从事 AB 实验研究

的数据科学家、系统开发人员，以及需要实验驱动策略决策的业务和产研团队，同时也适合对数据驱动增长和数据科学等领域感兴趣的读者。若本白皮书存在不当或者错误之处，欢迎大家批评指正，我们将不断完善与丰富内容，跟大家一起理解 AB 实验和数据科学，推动技术进步。

目录

第一部分 AB 实验概述

第一章：走进 AB 实验

1.1 了解 AB 实验

工欲善其事，必先利其器。在这个数据驱动决策的时代，AB 实验已经成为洞察用户行为、优化产品体验的不可或缺的工具。AB 实验，又称为在线对照实验（Online Controlled Experiment），其概念源自生物医学中的“双盲测试”，即将病人随机分为两组，在不知情的情况下分别给予安慰剂（或旧药物）和新药治疗，经过一段时间实验后再比较两组病人是否有显著差异，从而确定新药的有效性。自 2000 年 Google 将 A/B 实验应用于互联网产品测试以来，这一方法已在包括美团在内的各大互联网公司得到了广泛应用。

假设美团履约侧在可为某些（用户，商家）提供配送服务时，想验证在 App 的 C 端产品上弹窗以及展示某标签是否能促进用户下单意愿。此时，AB 实验提供了理想的解决方案。如图 1-1 所示，其做法为通过圈选一部分用户并随机分配为实验组和对照组（随机分流可确保两组在诸多特征上无差异），实验组用户施加新功能 / 新版本策略，而对照组用户继续使用旧功能 / 旧版本策略。一段实验周期后基于日志系统和业务系统收集的用户指标数据进行分析，比较实验策略与对照策略是否有显著收益，并以此为依据判断新策略是否应推广到全部用户。

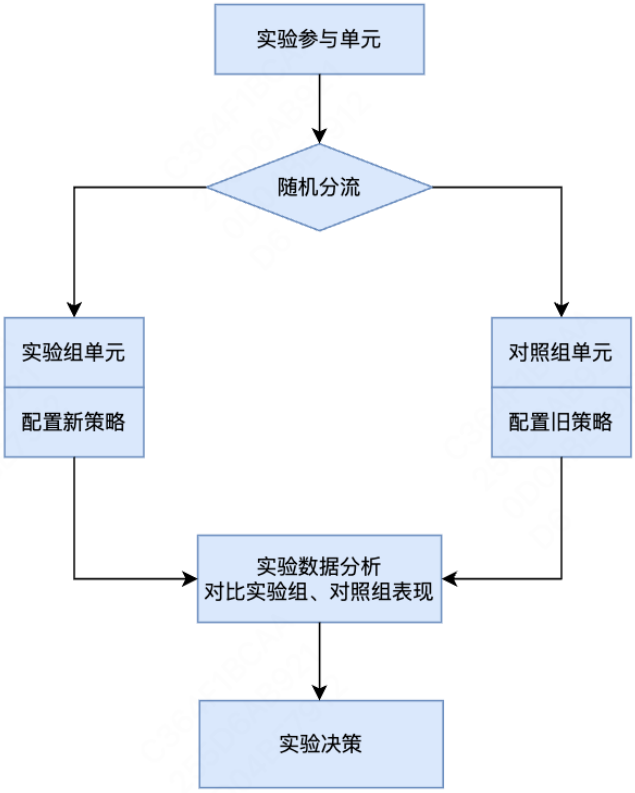


图 1-1: AB 实验流程

AB 实验之所以能迅速成为工业界数据驱动决策的黄金标准，主要归功于其能定性验证因果关系以及定量评估增长价值。某个策略的改变是否会导致产品指标的改变，本质上需要的是一种因果关系的判断，即“策略迭代优化”的因是否会带来“产品质量改变”的果。单凭经验以及相关性分析难以做成正确的决策，Google 和 Microsoft 相关统计表明，即使很有经验的相关人士正确判断产品策略的概率也只有 1/3。依赖相关性同样可能导致错误的决策，例如提供订阅服务的微软 Office 365 观测到看到错误信息并遭遇崩溃的用户有较低的流失率，这是因为高使用率用户往往看到更多错误信息以及流失率更低。但这并不意味着 Office 365 应该显示更多的错误信息或者降低代码质量使得频繁崩溃。

另一个著名的相关性案例为国家的巧克力消耗量与获得诺贝尔奖的数量相关性高达

0.79，但这并不意味着通过提高巧克力消耗量可以提高诺贝尔奖数量。实际产品迭代过程中往往应透过相关性寻找真正的因果关系。而 AB 实验作为目前已知的快速、低成本、科学验证因果关系的最有效手段，其可以通过随机化过程等可有效控制除干预策略外，实验组、对照组间其他混杂变量与影响特征是均衡的，最终的结果差异可归因于完全由干预贡献。同时借助假设检验等统计理论，能够科学、定性地验证策略迭代是否会带来业务的真实提升。因此，在产品迭代中通常采用 AB 实验识别正确的因果关系，保障迭代优化朝着正确方向前进。

AB 实验同样可通过精确量化策略收益、产品风险和成本，定量评估增长价值。例如，当某业务希望准确评估新补贴策略带来的下单规模提升时，最理想的方案是面对同一拨用户，假设存在两个完全相同的平行时空，平行时空一中所有用户体验新补贴策略 B，类似的平行时空二中所有用户体验旧补贴策略 A，通过直接对比 2 个平行空间的用户行为的平均表现（例如人均单量），则可观测新补贴策略相比旧补贴策略的提升效果。然而现实世界中不存在两个平行时空，针对同一用户，我们只能观察到其接受策略 A 或策略 B 下的一种表现，在此约束下，AB 实验可为我们提供了理想平行时空的一个近似替代。

具体的仍如图 1-1 所示，现实世界中通过随机实验手段可将用户随机均匀的分为实验组和对照组 2 个足够相似群体，并分别施加新策略以及旧策略。由于在随机分配机制下理论上实验组和对照组用户的平均表现可以分别代表 2 个平行时空下所有用户的平均表现（可参阅第 2 章实验基础原理），因此通过对比实验组、对照组间差异可以有效估计策略迭代带来的具体收益、风险与成本，帮助实验者做出更为理性的决策。

1.2 深入 AB 实验——以到家可信实验为例

1.2.1 错综复杂的实验陷阱与挑战

以美团到家业务实验为例，如图 1-2 所示，实验者可能会经常面临各种各样复杂的陷阱与挑战，处理稍有不当则可能损失实验的可信度，甚至带来错误的实验结论。

图 1-2：到家实验难题示例

具体的，以下是到家几个常见实验难题的简要介绍，这些问题也经常出现在其他业务实验中，更多案例与解决方案可详见后面章节。

案例一：小样本和溢出效应是制约履约场景下进行可信实验的两大难题。一方面，履约配送场景下样本量稀少与地域差异明显的现状，使得随机对照实验下难以保证分组的业务同质性以及很难有效地检测出实验提升效果。受自身业务形态和空间维度限制，部分配送策略的最小作用单元为区域 / 区域组（一个配送区域可以理解为某个地域空间）。因此在实验设计上，我们必须考虑区域或者更粗颗粒维度的分流。然而大部分城市区域 / 区域组很少，仅几十个左右。并且同城市各地域间的差异也往往比较显著，这在数据上体现为区域间指标波动剧烈。严峻的小样本与地域间差异显著的问题，导致随机分流下通常难以检测到策略小的提升效果，并且与结果变量相关的特征在实验组、对照组的分布差距可能较大，放大业务上实验组对照组不同质问题的同时给实验结果带来质疑。

另一方面，溢出效应（Spillover effects）引发的实验组、对照组间的不独立性，也会导致一些履约实验效果估计不够精确，甚至带来显著的估计偏差。AB 随机实验中关键的个体处理稳定性假设（SUTVA）假定实验单元的结果不受到其他单元分组的影响，简而言之，实验单元间相对独立，然而美团履约业务策略通常会涉及用户、商家和骑手等多方协同以及各方的相互依赖，特别是用户订单和骑手存在多对一耦合关系，且骑手可以跨越多个区域甚至整个城市进行接单和配送，在这种场景下无论运单还是区域等粒度的实验，实验单元间都往往存在溢出、干扰，进而造成实验估计不准

确。关于小样本与溢出效应更多案例与解决方案将在第 3 ~ 5 章重点介绍。

案例二：不可忽视的方差与 P 值计算陷阱，以及求和型统计量、ROI 指标等高阶评估方法诉求。AB 实验主要是通过在某个设定的抽样机制下，观察抽样的样本来推断总体的提升效果，并通过显著性检验辅助判断实验组、对照组之间差异是真实策略还是抽样噪音带来的。在该过程中通常需涉及大量统计学理论，包括方差、检验方式和 P 值计算等，稍有不慎容易掉入统计陷阱，难以得出可靠的实验结论。例如当分流单元与分析单元不一致时，错误的方差计算方式容易低估实际方差，导致假阳性。如图 1-3 左侧所示，在真实策略没有任何提升的情况下，分析单元细于分流单元时出现错误判别策略有效的概率接近 50%。正确的做法应该是先聚合到分流单位，再应用 Delta 技术推导的正确方差计算公式，如图 1-3 右侧所示，在正确方差计算下如果真实策略没有任何提升，P 值近似服从均匀分布，以及假阳性错误率基本控制在指定的显著性水平 5% 以内。

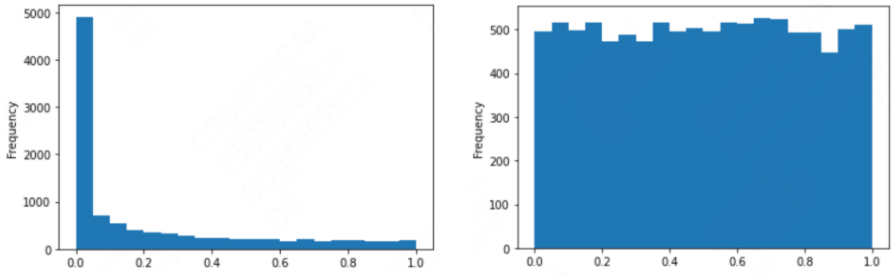


图 1-3：10000 次 AA 模拟下 P 值分布图

许多场景同样存在求和型统计量、ROI 指标等高阶评估方法诉求。例如假设策略可能影响实验组和对照组间用户的活跃度（留存）。如果实验组策略优于对照组，边缘用户可能从对照组流失，而实验组会吸引新用户。这种情况下，尽管实验组的下单量提升，但由于转入实验组的是非活跃用户，其均值可能低于对照组均值。基于均值统计量的显著性分析会拉低策略效果，甚至出现相反结论，不再适用，需引入求和型评估统计量。不同于非营销场景下关注策略的绝对提升（实验组观测值 − 对照组观测值）与相对提升（实验组观测值 / 对照组观测值 −1），营销场景下有时关注 ROI：（实验

组观测值 − 对照组观测值) / (实验组成本 − 对照组成本)。无论是求和型统计量还是 ROI 统计量，都需要重新推导和适配正确的方差计算和 P 值计算公式，以确保实验结论的准确性。更多详情可参阅第 3 章。

案例三：受限于公平性风险等与产品形态无法采用传统 AB 实验，需引入准实验或者观察性研究工具评估。当运营策略或产品升级涉及实验对象公平性等风险，或者产品分流与干预不受实验者控制时，通常需要在整城范围内施加策略，并采用观察性研究进行评估。例如，在某个城市推广线下广告策略时，由于无法控制部分用户看到广告的同时部分用户看不到，无法进行用户随机 AB 实验。

同样的，即使可在实验城市内干预分组，但受限于产品形态、运营管理难度甚至溢出效应，部分实验也只能运行准实验。例如考虑在保障整体覆盖范围不变的情况下，对所有不重叠的区域进行边界优化（新配送区域边界划分规则）甚至合并。此时显然不能考虑按区域随机分流，因为 2 个相邻的区域，在保持覆盖范围（并集）不变且不重叠约束下，优化 A 区域边界必然会导致 B 边界跟随变化，从产品形态上无法实现 A 区域边界变更但 B 区域边界维持不变。此时一种退而求其次的做法可以考虑将整个城市拆分为 2 个半城，在实验半城内部调整优化区域边界，对照半城维持不变，然后再利用 DID 等准实验手段评估新区域划分规则带来的提升效果。关于准实验与观察性研究基础原理与更多应用案例可参阅第 5 ~ 6 章。

1.2.2 零门槛运行可信实验范式与流程

为了让任何人都能摆脱 AB 测试重重困境，零门槛自主运行科学可信的实验，美团履约技术团队制定了一套数据科学家、数仓开发、系统开发多方协调保障的实验接入与运营机制，通过科学的实验方案、规范的实验流程和正确的指标数据保证实验可信度。对于新业务场景实验，尤其是重点或复杂实验，数据科学团队全程参与，前置深入实验场景，明确实验痛点，攻克置信难题，制定匹配的实验方案，并在实验平台配置实验模板。数仓开发为对应场景订阅和维护关注的实验指标数据集，保障指标定义规范与准确。与此同时数据科学家与系统研发人员共同规范化、模块化平台实验流程，允许对应算法场景后续可零门槛自主运行可信实验。

规范的实验流程和匹配的平台能力帮助实验者快速验证策略并科学决策。整个实验流程实验者只需选择实验场景模板新建实验设计、配置实验变体参数并查看实验报告。在实验设计环节，实验者可自助选择评估指标以及圈选流量，并可通过 MDE 分析与样本量预估功能辅助判断圈流样本量是否足够以及选择实验周期。完成实验设计后直接输出分流表达式，帮助用户轻松完成分流配置，同时可查看同质性、MDE（实验可有效检测出的提升效果）等关键信息。实验者可直接基于实验设计快速创建、管理实验，实验结束后自动输出显著性、趋势图等实验报告，用户无需再担心包括异常值陷阱、方差计算陷阱、P 值计算陷阱和多重比较陷阱在内的各种统计陷阱对实验结论的影响。同时平台还提供实验监控与诊断结果衡量实验有效性，以及实验探究功能支持实验者按维度、日期、指标等下钻与查看实验结果，辅助实验者进行决策。

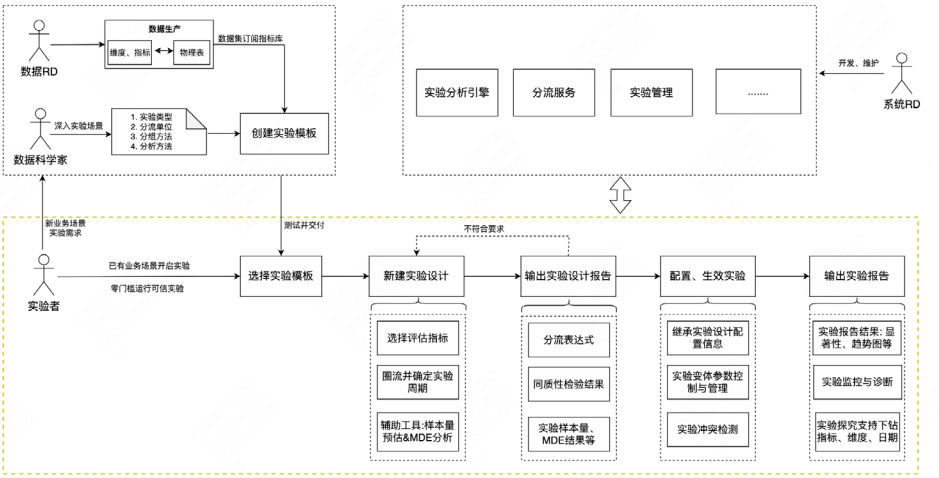


图 1-4：单个实验流程图

在上述实验流程中，不难看出，即使没有复杂的实验背景与专家知识的实验者也可零门槛自主运行可信实验。这不仅归功于数据科学家前置制定实验模板，还得益于构建了体系化的实验分析引擎，为用户提供标准化的流程和多样化的方法，并帮助用户避开各类实验陷阱。分析引擎作为一个中心方法库，整合了数科同学的所有优秀的实践，并涵盖学业界绝大部分实验方法。同时分析引擎也旨在促进知识共享，它可以像“积木”一样接入各种实验平台，服务不同角色的用户。对于具有专家级统计理解的

用户，可以提供原子化工具组件，帮助他们在业务场景约束下综合权衡偏差和方差，制定适合其业务场景的实验方案。对于普通用户，可以使用实验平台，轻松避开各类实验陷阱并输出实验报告，零门槛运行可信实验。

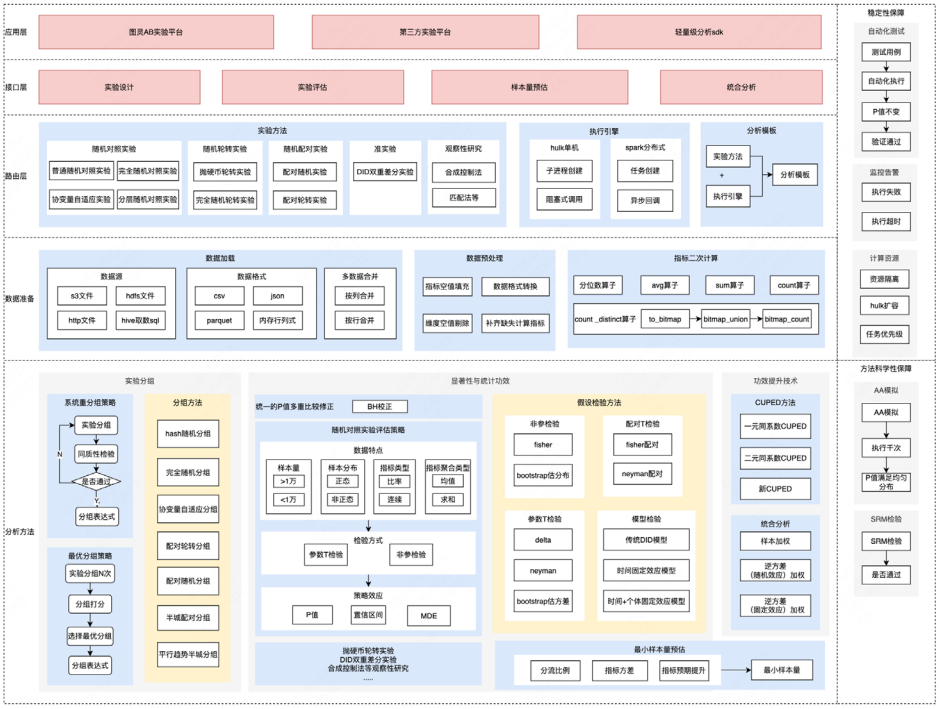


图 1-5：分析引擎架构图

1.2.3 实验方法选择指南

考虑到各类评估方法的复杂度和准确性上各有千秋，我们基于实验理论与实践经验，沉淀了一套大体的实验方法选择流程图，如图 1-6 所示，总体而言从可信度等级上优先选择随机实验（包括随机对照实验和随机轮转实验），其次是准实验，最后是观察性研究。

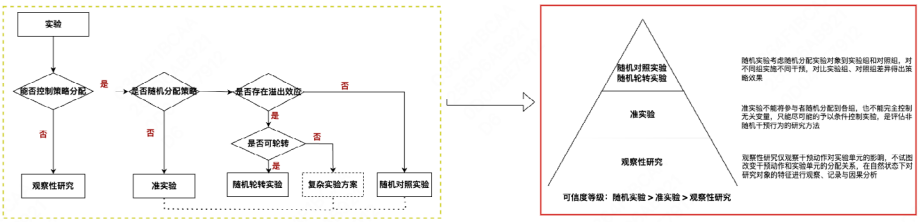


图 1-6：实验方法选择流程图

在上述大体流程下部分实验场景同样存在方法升级，下表简要展示实验方法库及其适用场景，详细方法内容将在本白皮书后续第 3 ~ 7 章节中重点展开与讨论。同时大部分方法也已集成于履约 SDK 分析引擎，线上调用与线下分析详情，大家可参阅白皮书的第 8 章节。

沉淀方法库	方法类	应用场景
随机对照实验	普通随机对照实验	经典随机对照实验，支持连续型、比率型指标，以及绝对提升、相对提升、ROI评估、SUM求和等评估类型。
	方差削减技术下的随机对照实验	Cuped、MLRATE等方差削减技术，适用于中小样本下提高实验检测灵敏度。
	提升同质性的高阶随机对照实验	包括分层随机对照实验、配对随机对照实验、协变量自适应设计、重随机化等，适用于中小样本下提升分组同质性等。
	解决溢出效应的复杂随机对照实验	溢入溢出建模解决地域级溢出效应，随机饱和实验、双边实验设计识别与消除订单等颗粒度实验溢出效应。
随机轮转实验	完全随机轮转	全城存在强溢出效应下，采用单城按天轮转实验以及多城分层轮转实验可完全消除溢出效应，但轮转实验不适用于用户感知明显的实验策略，因为其会严重干扰用户的自然体验。
	抛硬币随机轮转	适用于短轮转片场景，或者可在随机对照实验基础上搭配轮转增加实验样本量。选择轮转时间片过短时需警惕或考虑携带效应（上一时间片策略影响下一时间片表现），该轮转方式同样不适用于用户感知明显的实验策略。
	配对轮转实验	相比完全随机轮转，配对轮转可节约实验资源，同时更适合需聚焦于天气等不可控因素下的实验评估场景。缺点为配对轮转可能存在极轻微溢出效应，同样不适用于用户感知明显的实验策略。
准实验	双重差分法	适用于实验者可干预分组，但受限于产品形态、运营管理难度甚至溢出效应等无法随机分组，退而求其次采取半城分组等设计的实验。
观察性研究	合成控制法	在受限于公平性等风险需整个实验城市（极少量实验城市）施加实验策略，或者干预措施不受实验者控制等场景下，需酌情引入观察性研究工具评估策略迭代带来的收益。
	Causal impact	
	匹配(含PSM)	
高阶工具	统合分析	用于综合同一策略的多次独立实验结果（可不同时空），以得出更全面和可靠的实验结论。特别是在难以同时开展大规模实验，或者单次实验功效不足需通过重复实验提高功效场景中，具有极大的应用价值。
	多重比较	在同时进行多个统计假设检验时，例如检测实验组和对照组多个指标显著性，需采用多重比较方法来控制整体假阳性错误率（即将无效策略错误判为有效的概率）
	异质性因果	在不同个体或群体间策略效果存在差异时，细究不同特征条件群体下的提升效果，通常被用于算法建模训练弹性等。
	序贯分析	适用于在实验过程中查看实验结论并根据显著性决定继续/关闭实验，但通常需要确保前后时间段实验单元满足独立性。
	MAB实验	用于优化和探索多种策略的实验方法，可以动态调整各组流量以快速收敛到最优策略，但会损失一定的因果结论。

第二部分 基础原理与案例剖析

第二章：AB 实验基础

2.1 实验基础原理概述

AB 实验原理源于统计学中经典的 Rubin 潜在结果模型（也称反事实因果推断框架）。考虑最简单的情况，当我们想要比较两个策略的差异以获得更优策略时。如图 2-1 所示，最理想的方案是面向同一拨用户或者全部用户，假设存在两个完全相同的平行时空，平行时空一中所有用户体验实验策略 B，类似的平行时空二中所有用户体验对照策略 A，那么直接对比 2 个平行空间用户行为指标表现，则可决定哪个策略胜出以及观测真实的平均实验效应。

具体的，如果记 $Y_i(1), Y_i(0)$ 分别为第 i 个个体在实验策略 B（平行空间一）以及对照策略 A（平行空间二）下的指标表现，则显然可定义 Individual causal effects: $\tau_i = Y_i(1) - Y_i(0)$，以及策略真实平均提升效果：

$$ATE = \sum_{i=1}^{n} \tau_i / n = \sum_{i=1}^{n} Y_i(1) / n - \sum_{i=1}^{n} Y_i(0) / n \text{ or } E(Y_i(1)) - E(Y_i(0)).$$

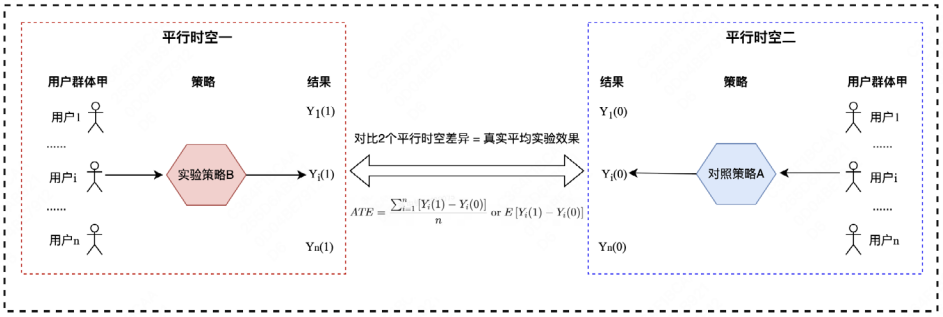


图 2-1: Rubin 潜在结果模型

然而，现实世界中不存在两个平行时空，针对同一用户，我们只能观察到其接受策略 A 或策略 B 下的一种表现。因此，现实世界中通常考虑先通过随机实验手段，将用户随机均匀地分为实验组和对照组 2 个足够相似的群体，并分别施加实验策略 B 以及对照策略 A。

如图 2.2 所示，在这种随机分配下理论上实验组和对照组用户的平均表现（在数学期望意义下）可以分别代表 2 个平行时空下所有用户的平均表现，因此通过对比实验组、对照组间差异可以有效估计策略迭代带来的具体收益、风险与成本，帮助实验组精细成本收益，结合业务做出更为理性的决策。然而在单次实验中，尽管理论上实验组和对照组来自同一总体，但实际上每次随机分配下 2 组间业务指标通常存在一定的差异（样本量越多差异越小）。这种差异可以理解为由抽样机制或者是分组机制的随机性贡献，即每次随机分配下实验组、对照组个体未施加策略时的平均差异在真值 0 附近波动。为准确识别单次 AB 实验中两组差异观测值是由分组的随机波动还是真实策略效果贡献，通常需借助假设检验、置信区间等统计工具进行判断和论证（相关内容可参考 2.2 章节）。

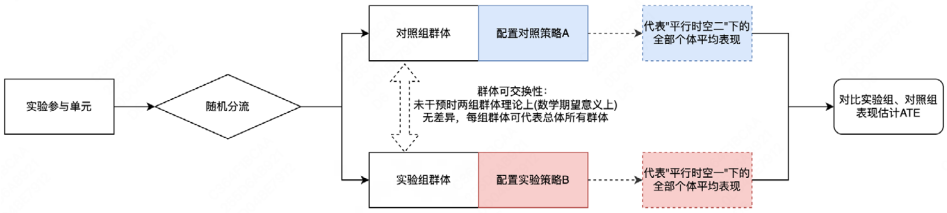


图 2.2：随机对照实验原理

然而，随机对照实验准确刻画策略因果效应存在 2 大关键前提：

1. **个体处理稳定性假设（SUTVA）：**实验单元的行为结果不受到其他单元分组的影响，即实验单元间相对独立，不会因为直接关联（如社交网络）或者间接关联（如共享资源）而互相产生干扰或者溢出。SUTVA 被破坏的典例包括：某打车 App 想要测试不同的溢价算法时，如果效果很好以至于实验组乘客更愿意打车，则路上可供搭乘的司机数量会减少，进而可能导致对照组难打上车，从而打车的对照乘客减少。又例如某通信工具上线增加通话时长的新功能时，如果实验组用户通话时长增加，而实验用户通话对象包括对照用户，从而也会提高对照组用户的通话时长。（信息源自：Ron Kohavi, Diane Tang, Ya Xu 著作《关键迭代 —— 可信赖的线上对照实验》）
2. **分组随机性：**实验单元进入实验组、对照组可完全由实验者随机分配，不受限于实验单元自身行为选择与表现。分组随机性破坏的案例包括例如在测试吃药是否对治疗感冒有效时，吃药行为可能完全由病人自行决定，且感冒更严重的人更加偏向于吃药，而不是随机选择。SUTVA 假设以及分组随机性的破坏会导致实验组（对照组）平均表现并不代表平行空间一（平行空间二）——全部个体接受实验（对照）策略下的平均表现，因此对比实验群体与对照群体的表现不能准确反映策略的真实效果。需引入更高阶实验方法或因果推断技术来解决，详情请参阅后面章节。

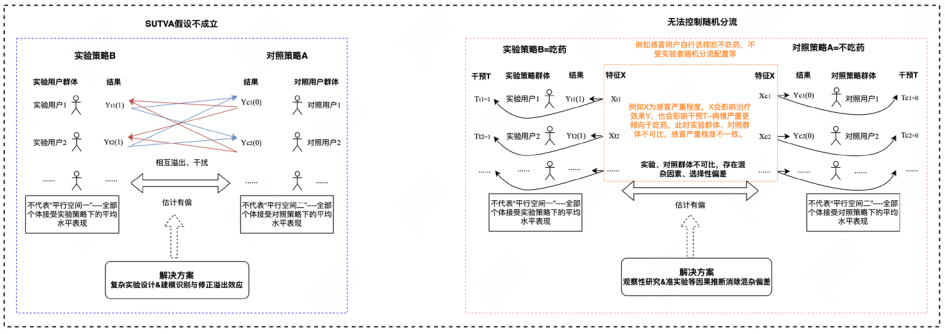


图 2-3：随机对照实验不可用情形

2.2 AB 实验统计学基础

2.2.1 参数估计

参数估计是数理统计中通过样本数据推断或估计总体未知参数的基本方法，在众多实际领域中被广泛应用。例如基于某批产品的随机抽样检查结果来估计总体废品率；又或者在 AB 实验中基于实验组、对照组样本表现差异去估计真实策略提升效果。大体而言，参数估计可划分为两大类：点估计和区间估计。

点估计（Point Estimation）

点估计，简而言之是使用样本数据计算一个单一的数值来估计总体参数。例如为了调查某批产品的废品率 c，可以从该批产品中随机抽取 n 个产品进行检查，记 a 为检查产品中为废品的个数，则可考虑用 a/n 估计总体废品率 c。常用的构造点估计的方法包括矩估计、极大似然估计、贝叶斯估计等，在此不详细展开介绍。点估计作为明确告知“未知参数是多少”的基本手段，那么现实中怎么评估点估计准不准？进一步的对于同一参数，不同估计方法求出的估计量可能不一样，那么如何判断不同的估计量之间的优劣。相合性、无偏性和有效性是常用的 3 个标准。相合性指当样本量无限增加时，点估计值趋近于总体参数值，即大样本下估计量能够准确反映总体参数。无偏性指从样本中得到的估计量的期望与总体参数相等，而有效性则指在样本量相同情况下，点估计 A 方差 < 点估计 B 方差则代表估计量 A 更有效。

实际上如果不失一般性，记 $\hat{\theta}$ 为参数 θ 的点估计，那么估计量 $\hat{\theta}$ 与总体参数真实值 θ 的均方误差 MSE(Mean Squared Error) 可以拆解为偏差的平方与方差。其中偏差 $Bias = E[\hat{\theta}] - \theta$：

$$MSE = E\left[(\hat{\theta} - \theta)^2\right] = \left(E[\hat{\theta}] - \theta\right)^2 + E\left[\left(\hat{\theta} - E\left[\hat{\theta}\right]\right)^2\right] = Bias^2 + Var$$

从上式中不难看出一个好的估计需要满足无偏性或者渐进无偏性，即偏差 Bias 等于 0 或者随着样本量增加趋于 0。与此同时在无偏条件下方差越小则点估计与参数真值越接近。通常而言，基于极大似然估计等方法构造的点估计的方差项 Var 通常以 1/n

阶速度趋于 0，其中 n 为样本量。

回到 AB 实验，实验者通常感兴趣策略总体提升效果 ATE，旨在通过实验收集样本构造 ATE 的点估计。在 SUTVA 假设成立的随机对照实验下直接对比实验组、对照组表现的点估计满足相合性和无偏性 / 渐进无偏性，并且随着样本量的增长点估计值趋近于总体参数值，因为方差（抽样 / 分组随机性贡献）随着样本量增加也趋向于 0。然而对于 SUTVA 假设以及分组随机性的破坏，会导致偏差 Bias 存在或者说不收敛到 0。因此此时需要一些复杂实验设计、建模分析与因果推断技术着重消除、避免偏差项，从而保证点估计的准确性。

置信区间（Confidence Interval）

对于总体的未知参数，在有限样本下点估计总存在一定的波动或误差，一个取而代之的自然想法为：兼顾波动性考虑估计参数落在哪个区间范围内，这便是统计学中经典的置信区间模块。置信区间顾名思义指的是总体参数的一个区间估计，以 95% 置信区间 [a,b] 为例，其表明区间 [a,b] 包含参数真值的概率在 95% 左右。例如假设我们要估计某城市中所有居民的平均收入。我们从这个城市中随机抽取了一部分样本，并计算了 95% 的置信区间结果为 [5000 元 , 7000 元]。这意味着我们有 95% 的信心认为，整个城市中所有居民的平均收入在 5000 元到 7000 元之间。又例如在对比新 App 页面设计与旧页面设计 AB 实验中，考虑到单次实验下随机分组波动性，转化率提升值点估计 0.03 与真实效果理论值存在一定的波动，此时可进一步参考 95% 置信区间估计 [−0.00136,0.06136]，即判断置信区间 [−0.00136,0.06136] 包含真实策略效果理论值的把握在 95% 以上，或者说有 95% 以上信心判断真实提升效果在 −0.00136 ~ 0.06136 之间。通常而言在置信水平固定情况下区间长度越短越好，学业界最经典的 95% 置信区间构造方式为 $\hat{\theta} \pm 1.96 * \sqrt{Var(\hat{\theta})}$，即在点估计基础上增加一个波动范围。从置信区间构造形式上也不难看出随着样本量的不断增加，置信区间变得越来越窄并收敛到参数真值点。

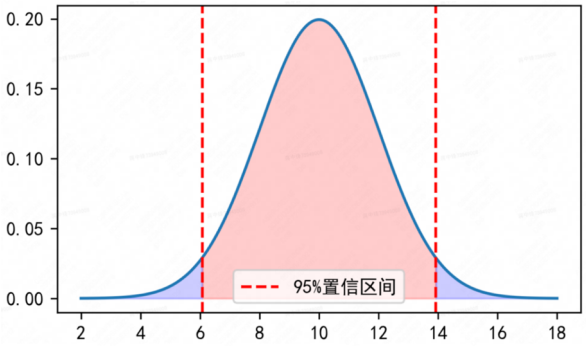


图 2-4：95% 置信区间图示

2.2.2 假设检验

假设检验（Hypothesis testing）是统计学中用数据论证某假设是否成立的方法，在工程、医学、社会科学等多个领域广泛应用。假设检验本质可理解为反证法，有点类似于法庭的评理，想象法庭上有一名被告，在开始无信息时假设被告是清白的（原假设），而检察官必须要提出足够的证据去证明被告的确有罪。如果没有足够的信息和证据证明被告有罪，那么判定原假设：被告清白成立。除非检察官提供足够的证据才判定被告有罪。统计学家 Fisher 提过一个女士品茶的假设检验著名例子，一名女士声称其可以品尝出奶茶制作过程中是先加入茶还是先加入牛奶。Fisher 提议给她八杯奶茶，并告知其中四杯先加茶，四杯先加牛奶，但随机排列，需要女士说出这八杯奶茶中，哪些先加牛奶，哪些先加茶。原假设是该女士无法判断奶茶中的茶先加入还是牛奶先加入，根据猜中的次数判断该假设是否成立。结果女士测试结果为八杯品尝都正确。在原假设下若单纯以概率考虑，八杯都正确的概率为 1/70（因为 8 选 4 的组合数是 70），约 1.43%，即原假设成立下统计上完全猜对可能性极小，单次测试基本上不会发生，即几乎排除女士完全盲猜正确的可能，因此我们有理由去拒绝“该女士无法判断奶茶中的茶先加入还是牛奶先加入”的假设。

类似的，假设检验在 AB 实验中通常被作为基本工具论证新策略是否相对旧策略会带来业务收益。例如当测试一个新的 App 广告设计是否能提高用户点击率时，通常原假设新策略相对旧策略无效，然后收集现有证据 -- 样本数据去论证实验组和对照

组之间是否具有显著的差异，如果拥有足够证据——实验组对照组差异很大（这在新策略无效下基本上不太可能出现），则推翻“新策略相对旧策略无效”的假设，否则认为在现有证据——样本信息下接受原假设成立，除非收集更多证据（样本数据）再“重新开庭论证”。一个完整的假设检验主要包括以下几个步骤：

1. 提出假设

- 原假设（Null Hypothesis，通常选择为默认结论或者需推翻的结论）H0：实验组与对照组无差异，表示策略无效果。
- 备择假设（Alternative Hypothesis，通常为想被证明的结论）H1：实验组与对照组有差异，也可考虑单边备择假设 H1：实验组 > 对照组，或者 H1：实验组 < 对照组。但在 AB 实验中为同时兼顾收益和风险通常默认选择双边备择假设。

2. 选择显著性水平

显著性水平（α）指能容忍的犯第一类错误的概率，其中第一类错误是指在原假设为真时，拒绝原假设的犯错，又称假阳性。显著性水平是人为定义或指定的概率值，学业界常见的显著性水平为 0.05。

3. 构造检验统计量

根据样本数据和假设类型，选择合适的检验统计量，AB 实验中最常用的方式为双样本 t 检验。例如在探索某策略是否会带来单量增长时，按用户随机对照试验可考虑构造检验统计量：

$$T = \frac{\bar{Y}_t - \bar{Y}_c}{\sqrt{\mathrm{Var}\left(\bar{Y}_t\right) + \mathrm{Var}\left(\bar{Y}_c\right)}} \text{ or } \frac{\bar{Y}_t / \bar{Y}_c - 1}{\sqrt{\mathrm{Var}\left(\bar{Y}_t / \bar{Y}_c\right)}}$$

其中，$\bar{Y}_t$，$\bar{Y}_c$分别为实验用户人均单量、对照用户人均单量。

其中方差计算常用算法包括 Delta 方法、Bootstrap、Jackknife 方法等，当然检验方式也包括参数检验、非参数检验等。

4. 计算拒绝域和 p 值

拒绝域是指在假设检验中拒绝原假设的检验统计量的取值范围，其通常依赖于显著性水平等。尽管可通过判断检验统计量观测值是否落在拒绝域决策拒绝 / 接受原假设，假设检验实际应用中通常考虑一个更常用的标准——P 值。P 值表示在原假设为真时，比所得到的统计量观察结果更极端的概率。其计算逻辑为先推导出在原假设 H0 成立条件下检验统计量的概率分布（在 AB 实验场景可以想象为，在策略无效场景下，假设允许做无数次实验，每次实验独立执行分组机制，并且得到一个检验统计量，基于若干次实验得到的若干个检验统计量观测值画图，即得到 H0 下且在对应实验分组机制下的检验统计量的概率分布。现实中可通过一些极限理论等统计定理性质来基本近似获得原假设 H0 成立条件下检验统计量的概率分布），然后再计算观察到比当前样本下检验统计量观测值更极端的概率，直观上也可理解为在原假设成立情况下，出现当前观测值及更极端场景的概率，如果很小则意味着原假设成立下单次实验不太能出现的小概率事件发生了，需质疑甚至拒绝原假设。

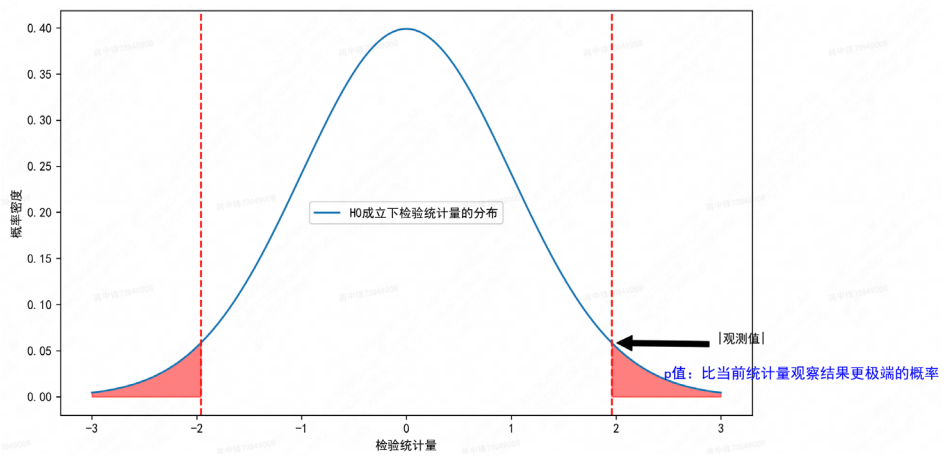


图 2-5: P 值图示

5. 作出决策

假设检验的核心思想反证法，理论上小概率事件在一次实验中几乎不可能发生，如果发生了则说明原假设不合理。因此可通过比较 p 值与显著性水平 α：

- 如果 p 值 ≤ α，拒绝原假设，支持备择假设。
- 如果 p 值 > α，接受原假设，拒绝备择假设。

2.2.3 极限理论

极限理论是假设检验与置信区间等过程中构建统计量分布的理论基础，是统计学中一个庞大且内容丰富的关键模块。由于主题和篇幅的限制，本白皮书将不对其进行深入探讨，仅简要介绍几个常用的原理。读者也可选择跳过本部分内容。

大数定律（Strong Law of Large Numbers）： 假设$X_1, X_2, \ldots, X_n$是一组独立同分布的随机变量，每个变量的期望值为μ且方差有限。根据强大数定律，当样本量n趋于无穷大时，样本均值几乎必然收敛于总体均值：

$$\frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mu$$

其中：$\frac{1}{n}\sum_{i=1}^{n} X_i$是样本均值，a.s. 表示几乎处处收敛（almost sure convergence），μ是总体均值。强大数定律描述了独立同分布随机变量的样本均值几乎必然收敛于总体均值的现象。

中心极限定理（Lindeberg–Levy Central Limit Theorem）： 假设$X_1, X_2, \ldots, X_n$是一组独立同分布的随机变量，每个变量的期望值为μ和方差为σ^2。则当n趋于无穷大时，样本均值的标准化形式收敛于标准正态分布：

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{d} N(0,1)$$

其中：$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$是样本均值，$N(0,1)$表示均值为 0，方差为 1 的标准正态分布。上述中心极限定理表明样本量足够大时，样本均值的分布可以近似为正态分布，即使原始数据的分布不是正态的。

Delta 定理： Delta 方法是统计学中用于近似计算函数的随机变量的分布的一种方法。它通常用于推导复杂函数的渐近分布，尤其是在处理非线性变换时。Delta 方法的核

心思想是使用泰勒展开来近似函数的变化。例如假设X_n是一组随机变量的样本均值（样本量n），且$\sqrt{n}(X_n-\theta)$收敛于某个正态分布，其中θ是参数。对于一个可微函数$g(\cdot)$：

$$\sqrt{n}(g(X_n)-g(\theta)) \xrightarrow{d} N(0,\sigma^2[g'(\theta)]^2)$$

其中：$g'(\theta)$是函数g在θ处的导数。σ^2是X_n的方差。该结论同样可推广到多元场景：假设我们有一个$k-$维随机向量$\mathbf{X}_n=(X_{n1},X_{n2},\ldots,X_{nk})^T$，其均值为$\boldsymbol{\theta}=(\theta_1,\theta_2,\ldots,\theta_k)^T$，并且$\sqrt{n}(\mathbf{X}_n-\boldsymbol{\theta})$的分布收敛于一个正态分布。对于一个可微的向量值函数$\mathbf{g}(\mathbf{X})$，Delta 方法的多元版本可以表示为：

$$\sqrt{n}(\mathbf{g}(\mathbf{X}_n)-\mathbf{g}(\boldsymbol{\theta})) \xrightarrow{d} N(\mathbf{0},\mathbf{G}\Sigma\mathbf{G}^T)$$

其中：$\mathbf{g}(\mathbf{X})$是一个从$\mathbb{R}^k$到$\mathbb{R}^m$的可微函数。$\mathbf{G}$是在$\boldsymbol{\theta}$处的雅可比矩阵（Jacobian matrix），其元素为$\frac{\partial g_i}{\partial \theta_j}$。$\Sigma$是$\mathbf{X}_n$的协方差矩阵。$\mathbf{0}$是$m-$维零向量。

Slutsky 定理： Slutsky 定理是概率论和统计学中的一个重要定理，它描述了在某些条件下随机变量的极限行为。下面仅简单介绍涉及的以下三种情况：

1. 和的极限

如果$X_n \xrightarrow{d} X$（即X_n分布收敛于X），并且$Y_n \xrightarrow{p} c$（即Y_n以概率收敛于常数c），那么$X_n+Y_n \xrightarrow{d} X+c$。

2. 积的极限

如果$X_n \xrightarrow{d} X$并且$Y_n \xrightarrow{p} c$，那么$X_nY_n \xrightarrow{d} cX$。

3. 商的极限

如果$X_n \xrightarrow{d} X$并且$Y_n \xrightarrow{p} c$，那么$\frac{X_n}{Y_n} \xrightarrow{d} \frac{X}{c}$。

其中分布收敛$(\xrightarrow{d})$指的是随机变量的分布函数收敛于某个极限分布函数。概率收敛$(\xrightarrow{p})$指的是随机变量依概率收敛于某个常数。

2.3 常用实验术语

常见术语	介绍	备注
目标指标	目标指标又称北极星指标，是实验关注的核心指标，用来决策实验功能是否符合预期的「直接效果指标」或「成功指标」。目标指标通常是一个指标或者极少数个指标的集合	实验指标体系，选择评估指标时需尽量确保可计算、可归属（归属到不同的实验分组）和及时性，另外兼顾灵敏性、业务解释性等。
护栏指标	护栏主要存在2种类型，一是保护业务的指标，用来限制新策略带来的负面影响。其可以帮助实验者在达到业务目标的同时，确保不会违背重要的限制。二是与策略无因果关联的指标，额外用于辅助保障 A/B 测试的质量，即监控不受策略影响的指标是否发生显著变化以保障实验质量。	
驱动指标	驱动指标也称Sign Post指标、代理指标、间接指标，通常用来反映策略是通过影响那些中间因素来最终提升目标指标。驱动指标相比目标指标通常更短期、微观以及能灵敏反应业务变化的指标	
绝对提升	绝对提升=实验组指标值−对照组指标值，代表某新策略相对基准策略带来的绝对数值变化。	策略量化手段，其中连续型、比率型指标通常仅用来判断实验组/对照组指标与方差计算形式等。
相对提升	相对提升=（实验组指标值−对照组指标值）/对照组指标值×100%，代表某新策略相对基准策略所带来的相对百分比变化。	
ROI	常用于营销场景，通常为（实验组指标值-对照组指标值）/（实验组成本-对照组成本），可用于衡量当前策略下多花1元钱能带来多少提升效果等	
连续型指标	连续型指标通常指直接根据样本求平均/求和类型指标，例如完单量、GMV等规模指标。	
比率型指标	比率型指标主要指相对准时率（准时完成单/完单量）、完成单人效（完单量/有单骑手数）等Y/Z型指标，计算时通常采取整实验组Y/整实验组Z形式。	
同质性检验	同质性检验用于监控和保证随机对照实验下实验组和对照组在实验前没有明显差异，如果未通过同质性检验，则说明实验组、对照组实验前已存在显著差异，当前实验结论科学性无法得到保证。	实验监控模块
AA测试	AA测试类似于AB测试，仅B组与A组策略一致，即已知策略无效果。常用于测试实验方法与流程的正确性，具体做法包括例如1000次模拟观测p值是否服从均匀分布等。	
SRM检验	即样本比率不匹配检验，通过检验实验组、对照组分流比例是否满足设定标准，监控线上分流过程是否正确。	
显著性水平	显著性水平（α）为能容忍犯第一类错误的概率，其中第一类错误（Type I Error）是指在假设检验中，当原假设为真时，拒绝原假设的错误，也被称为“假阳性”。	假设检验相关知识
统计功效	统计功效是指在假设检验中，备择假设成立（策略真实有效果）情况下拒绝原假设的概率，即能检测到实际策略效果的能力。统计功效通常用符号 1−β 表示，其中 β 是第二类错误（即备择假设成立情形下未能拒绝原假设）的概率，统计功效通常依赖于样本量与实际策略效果等。	
P值	p值（p-value）表示在原假设为真时，观测到至少像当前极端统计量值的概率。	
MDE	MDE反映实验的灵敏度，代表在当前样本条件下能有效检测（检出概率大于等于80%）的指标提升幅度。即如果真实提升效果>MDE，则有80%以上概率能检测出显著性	
最小样本量	最小样本量一般在实验前确定，在达到预定功效（通常0.8）和显著性水平（通常0.05）下为检测出x效果（用户指定）所需的最少样本量。	
触发式分析	当实验策略仅影响或触发实验组和对照组中的部分个体时，针对这些触发群体进行的分析被称为触发式分析。触发式分析通常能够有效提高实验检测的灵敏度。	高阶实验知识
溢出效应	由于实验单元间的直接关联（社交网络）或者间接关联（竞争共享资源等），参与AB实验的实验组与对照组之间可能并不独立，这种实验组、对照组间的相互干扰影响通常被称为溢出效应。	
携带效应	携带效应（Carryover Effect）是指上一时间段策略对下一时间段存在直接与间接影响，在轮转实验中需要特别警惕携带效应以避免实验偏差。	
长期效应	长期效应通常指需要花费数周、数月甚至数年等较长时间跨度后才能累积的实验效应，在短期实验（1~2周）效果难以稳定代表长期效果时，例如提价商品或服务价格可能会增加短期营收，但由于用户放弃产品或服务，长期收入会减少。可考虑长周期实验、代理指标建模法等技术策略长期效应。	

第三章：随机对照实验

在美团到家业务场景中，经常会碰到随机分流的实验场景，比如全城 AOI（Area of Interest，可以是小区、学校等点位，是按照社会功能定位，在地图上将特定区域绘制成一个个电子围栏的面状地理信息）随机分流或者订单随机分流。在随机对照实验中，我们可以定量判断 A、B 两个策略是否有显著的差异，如果有差异则进一步探究哪个更有效，并依次对更优的策略进行推广。因此，随机对照实验是帮助业务和算法探索并迭代策略的重要工具。

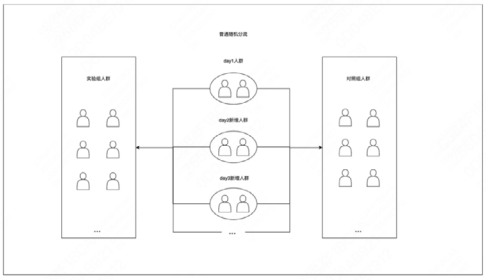


图 3-1：普通随机分流示意图

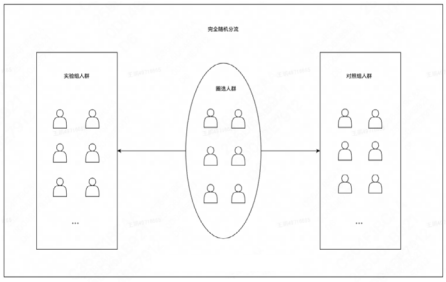


图 3-2：完全随机分流示意图

3.1 经典随机对照实验

随机对照实验是 AB 实验最基础且最重要的实验方式。对于施加实验策略的对象，理想情况下，我们想要在完全相同的时间与外部环境下将其与不施加实验策略的对象进行对比。但是我们没有穿梭时空的超能力去直接观测另一个平行时空中这些对象的表现。而随机对照实验就是连通现实与平行世界的一个桥梁，使得我们可以人为模拟出平行世界中的情形。

具体来说，以两组为例，我们可以通过随机分流的手段，将全体实验对象随机分为实验组和对照组两部分，其中实验组受到干预，对照组未受到干预。根据反事实结果框架，随机对照实验的一个重要前提是可交换性，即对于任意一个个体i，其在未受干

预和受到干预的两个平行时空的指标表现 $Y_i(0)$ 和 $Y_i(1)$，与其实际是否受到干预 T 无关。这意味着在随机对照实验中，除了是否受到干预，两组在各个特征上都能从统计上完全代表总体。即使把实验组和对照组内的对象进行互换再进行实验，也并不会影响最后的实验结果。因此，根据第二章的框架，我们可以通过下列方式来近似计算总体的策略平均提升效果：

$$\mathrm{ATE} = E(Y_i(1)) - E(Y_i(0)) \approx \sum_{T_i=1} Y_i(1) / n_t - \sum_{T_i=0} Y_i(0) / n_c = \sum_{T_i=1} Y_i / n_t - \sum_{T_i=0} Y_i / n_c$$

即策略平均提升效果实际可以用实验组的平均值 − 对照组的平均值来替代，其中对于个体 i，$Y_i(0)$ 和 $Y_i(1)$ 是未干预和受到干预的两个反事实结果，Y_i 为实际结果，T_i 为施加在个体上的干预情况。此外，n_t 为实验组样本量，n_c 为对照组样本量。

另一个重要假设是个体处理稳定性假设，即 SUTVA 假设（Stable Unit Treatment Value Assumption）。它要求实验单位的表现是独立的，且干预效果稳定，实验单元的行为结果不受到其他单元分组的影响，不会因为实验组和对照组的关联而产生干扰或者溢出。

3.1.1 随机对照实验的限制与挑战

随机化分组能使所有可能的混杂变量（包括未观测到的混杂变量）在实验组和对照组之间呈均匀分布，消除混杂变量带来的影响，提升结果可信度。因此，只要实验条件允许，随机对照实验就是我们的首选选择。在理想情况下，同一个个体在两个平行时空完全一样。但现实生活往往不如人所愿，在有限的样本量下，随机分出的两部分对象会存在一定差异，也即可交换性无法严格满足。此时，我们需要一些定量标准来刻画两组之间的差异是否可以被忽略，即同质性检验。在随机对照实验中，我们会选取一段实验前周期，对实验组和对照组两组的需要考察的一些指标值进行差异是否显著的检验。当两组结果没有检验出显著差异时，我们可以认为同质性检验通过，也即可交换性近似满足，此时使用随机对照实验得到的结果是可信的。

尽管随机对照实验的可信性最高，我们也常常会面临很多客观上的限制与挑战：

1. **公平性：**在一些特殊业务场景，考虑到对用户以及骑手等群体的公平性，无法对考察群体进行随机分组
2. **溢出效应：**实验单元之间存在相互影响与干扰，造成结果偏差。例如，在调度算法等场景，分别在实验组和对照组的两个区域往往会召回相同的骑手，即存在实验组和对照组两组之间的相互干扰。
3. **小样本量情形：**美团履约业务中有很多通过地理单元分流的随机对照实验。对于使用配送城市、配送区域、配送站点等面积较大单元的实验，在可用流量有限的情况下，样本量一般较少（几十个甚至十几个）且地域差异明显，分组难以保证同质且难以检测出显著的策略提升效果。
4. **业务影响：**在诸多业务场景会考量留对照组对实际业务影响的情况。如果对照组流量过多，可能存在影响当前线上策略效率的风险，从而对体验指标造成影响，造成用户端客诉。为了不影响正常业务，一些场景的实验组比对照组会采用 95:5 等极端的分组比例，实验功效较低难以检测出显著的策略提升。
5. **流量未全部触发策略：**在履约业务中，存在很多圈选流量与实际策略触发流量不完全一致的情况。为了准确评估策略效果，我们应该考察实际被策略触发的流量。此时的同质性需要进一步重新验证。

在美团的实验应用中，经典的随机对照实验通过普通随机分组和完全随机分组两种方式来实现，并相应配套有同质性检验和显著性检验的评估方式。通常来说，我们会取实验前一段周期的实验组和对照组两组指标表现，来进行同质性检验以验证分组特征的均衡性，也即近似保证随机对照实验的可交换性。而在实验完成后，我们会取实验期间的指标数据进行显著性检验，来判断策略效果是否显著有效。同质性检验和显著性检验实际上使用的都是下面同一套流程与方法，区别在于：我们希望同质性检验结果不显著，则可以认为两组表现相似，而希望显著性结果显著，则可以认为策略有效。本文主要详细讨论两组的情况，多组情况下相应的分组与评估方式可以类似推广，这里不再过多地进行阐述。

3.1.2 普通随机分组

正交的 AB 实验，需要保证流量足够的均匀分散，这就需要一个性能高、效果好的 Hash 算法来支撑，这里我们选用了 MurmurHash3_32。

1. 分组机制

从分组机制上来说，Hash 分流可以理解为伯努利实验。以两组为例，对于实验中总的 n 个实验单元，其中$n = n_t + n_c$。我们可以事先设定一个干预概率为$p = n_t / n$，对于每个样本，其在实验组的概率为p，在对照组的概率为$(1-p)$，满足如下分组机制：

$$P(\mathbf{T} = \mathbf{t} \mid \mathbf{Y}(1), \mathbf{Y}(0)) = \prod_{i=1}^{n} p^{t_i} (1-p)^{1-t_i}$$

其中$\mathbf{t} = (t_1, \ldots, t_n)$，满足$\sum_{i=1}^{n} t_i = n_t$和$\sum_{i=1}^{n} (1-t_i) = n_c$。

2. 适用场景

- 实验单位之间相互独立；
- 尤其适用于样本量较大，随着实验不断进行，可能有新的实验单位不断进入实验的场景。比如订单分流、用户分流、AOI 分流等实验场景。

3. 评估方式

我们的评估建立在假设检验的理论之上，原假设为实验组和对照组的均值相等，即$H_0 : \mu_t = \mu_c$。我们可以依此构造统计量，并计算实验结果的 p 值，当 p 值小于 0.05 时即为有显著差异，当 p 值大于等于 0.05 时认为没有显著差异。在显著性评估中，我们主要采用 Delta 和 Bootstrap 两大类方法（大规模数据情景还可考虑 Group Jackknife，但需注意 Jackknife 不适合分位数评估等）。

Delta 方法是统计学中一种用于计算极限方差的有名方法，其基本原理为若随机向量Y依分布地满足$\sqrt{n}(\mathbf{Y} - \boldsymbol{\mu}) \to N(0, \boldsymbol{\Sigma})$，则对于任意可微函数$g$，利用泰勒展开式和随机变量的渐近分布可得$\sqrt{n}(g(\mathbf{Y}) - g(\boldsymbol{\mu})) \to N(0, (\nabla g(\boldsymbol{\mu}))^T \boldsymbol{\Sigma} (\nabla g(\boldsymbol{\mu})))$。而 Bootstrap 方法是一种广泛应用于统计学的重采样技术，用于估计样本相关的统计量

（比如均值、标准差、分位数等）以及经验分布，尤其在小样本或者难以假设数据分布情况的时候使用。

实际应用中，我们会先区分需要考察的指标类型，基本的主要分为连续型指标和比率型指标，我们会在后面讨论求和型指标等特殊的指标类型。其中连续型指标主要指规模指标，比率型指标主要指型 Y / Z 的指标。

（1）连续型指标

Delta 方法

对于连续型指标 Y 的评估，如果考虑绝对提升可以直接使用经典的 Welch T 检验进行评估，可以按如下公式计算 p 值：

$$p值 = 2 * \left(1 - \Phi \left(\frac{|\bar{Y}_t - \bar{Y}_c|}{\sqrt{\frac{\sigma_t^2}{n_t} + \frac{\sigma_c^2}{n_c}}} \right) \right)$$

其中 Φ 为标准正态分布的分布函数。在业务层面，我们更多会关注实验组相对对照组的相对提升率：

$$\Delta_{\text{rel}} = \frac{\sum_{i=1}^{n} Y_i T_i / \sum_{i=1}^{n} T_i}{\sum_{i=1}^{n} Y_i (1 - T_i) / \sum_{i=1}^{n} (1 - T_i)} - 1 = \frac{\bar{Y}_t}{\bar{Y}_c} - 1$$

使用 Delta 方法可以推出其渐近方差形式为：

$$\text{var}(\Delta_{\text{rel}}) = \frac{1}{\mu_c^2} \text{var}(\bar{Y}_t) + \frac{\mu_t^2}{\mu_c^4} \text{var}(\bar{Y}_c) = \frac{1}{n_t} \frac{\sigma_t^2}{\mu_c^2} + \frac{1}{n_c} \frac{\mu_t^2 \sigma_c^2}{\mu_c^4}$$

我们可以通过如下估计量来估计这个方差：

$$\widehat{\text{var}}(\Delta_{\text{rel}}) = \frac{1}{n_t} \frac{\hat{\sigma}_t^2}{\bar{Y}_c^2} + \frac{1}{n_c} \frac{\bar{Y}_t^2 \hat{\sigma}_c^2}{\bar{Y}_c^4}$$

其中$\bar{Y}_t = \sum_{T_i=1} Y_i / n_t$，$\bar{Y}_c = \sum_{T_i=0} Y_i / n_c$分别为实验组和对照组的样本均值，$\hat{\sigma}_t^2 = \sum_{T_i=1} (Y_i - \bar{Y}_t)^2 / (n_t - 1)$，$\hat{\sigma}_c^2 = \sum_{T_i=0} (Y_i - \bar{Y}_c)^2 / (n_c - 1)$分别为实验组和对照组的样本方差。

进一步我们可以使用如下公式来计算p值：

$$p值 = 2 * (1 - \Phi(|\Delta_{\text{rel}}| / \sqrt{\widehat{\text{var}}(\Delta_{\text{rel}})}))$$

其中Φ为标准正态分布的分布函数。

Bootstrap 方法

Step1： 对于$b = 1, \ldots, B$做B次 Bootstrap 抽样，每次从n个样本里有放回抽取n个样本$(Y_{b1}^*, T_{b1}^*), \cdots, (Y_{bn}^*, T_{bn}^*)$，其中$Y_{bi}^*$为在第$b$次 Bootstrap 抽样中的第$i$个样本，$T_{bi}^*$为此样本受到干预的情况。

Step2： 对于$b = 1, \ldots, B$，计算$\Delta_{\text{rel},b}^* = \dfrac{\sum_{i=1}^{n} Y_{bi}^* T_{bi}^* / \sum_{i=1}^{n} T_{bi}^*}{\sum_{i=1}^{n} Y_{bi}^* (1 - T_{bi}^*) / \sum_{i=1}^{n} (1 - T_{bi}^*)} - 1$。

Step3： 再基于$\Delta_{\text{rel},b}^*$，$b = 1, \ldots, B$计算方差估计量为：

$$\hat{\sigma}_{\text{cont,boot}}^2 = \frac{1}{B} \sum_{b=1}^{B} (\Delta_{\text{rel},b}^* - \frac{1}{B} \sum_{i=1}^{B} \Delta_{\text{rel},i}^*)^2$$

Step4： 如果使用 Bootstrap 估分布的方式，我们可以根据 Bootstrap 的抽样结果进一步给出p值：

$$p值 = \sum_{b=1}^{B} 1_{\{|\Delta_{\text{rel,b}}^* - \Delta_{\text{rel}}| \geq |\Delta_{\text{rel}}|\}} / B$$

（2）比率型指标

Delta 方法

对于比率型指标Y的评估，我们主要关注实验组相对对照组的绝对提升率：

$$\Delta_{\text{abs}} = \frac{\sum_{i=1}^{n} Y_i T_i}{\sum_{i=1}^{n} Z_i T_i} - \frac{\sum_{i=1}^{n} Y_i (1-T_i)}{\sum_{i=1}^{n} Z_i (1-T_i)} = \frac{\bar{Y}_t}{\bar{Z}_t} - \frac{\bar{Y}_c}{\bar{Z}_c}$$

在显著性评估中，我们主要采用 Delta 和 Bootstrap 两大类方法。Delta 方法是统计学中一种用于计算极限方差的有名方法，其基本原理为若随机向量Y依分布地满足$\sqrt{n}(\mathbf{Y}-\boldsymbol{\mu}) \to N(0,\boldsymbol{\Sigma})$，则对于任意可微函数$g$，利用泰勒展开式和随机变量的渐近分布可得$\sqrt{n}(g(\mathbf{Y})-g(\boldsymbol{\mu})) \to N(0,(\nabla g(\boldsymbol{\mu}))^T \boldsymbol{\Sigma}(\nabla g(\boldsymbol{\mu})))$。使用 Delta 方法可以推出其实验组方差估计形式为：

$$\text{var}\left(\frac{\bar{Y}_t}{\bar{Z}_t}\right) = \frac{1}{\mu_{t,Z}^2}\text{var}(\bar{Y}_t) + \frac{\mu_{t,Y}^2}{\mu_{t,Z}^4}\text{var}(\bar{Z}_t) - 2\frac{\mu_{t,Y}}{\mu_{t,Z}^3}\text{cov}(\bar{Y}_t,\bar{Z}_t) =$$
$$\frac{1}{n_t}\frac{\sigma_{t,Y}^2}{\mu_{t,Z}^2} + \frac{1}{n_t}\frac{\mu_{t,Y}^2\sigma_{t,Z}^2}{\mu_{t,Z}^4} - 2\frac{1}{n_t}\frac{\mu_{t,Y}\sigma_{t,YZ}}{\mu_{t,Z}^3}$$

我们可以通过如下估计量来估计这个方差：

$$\widehat{\text{var}}\left(\frac{\bar{Y}_t}{\bar{Z}_t}\right) = \frac{1}{n_t}\left(\frac{\hat{\sigma}_{t,Y}^2}{\bar{Z}_t^2} + \frac{\bar{Y}_t^2\hat{\sigma}_{t,Z}^2}{\bar{Z}_t^4} - 2\frac{\bar{Y}_t\hat{\sigma}_{t,YZ}}{\bar{Z}_t^3}\right)$$

其中$\bar{Y}_t = \sum_{T_i=1} Y_i / n_t$，$\bar{Z}_t = \sum_{T_i=1} Z_i / n_t$分别为分子和分母的样本均值，$\hat{\sigma}_{t,Y}^2 = \sum_{T_i=1}(Y_i-\bar{Y})^2/(n_t-1)$，$\hat{\sigma}_{t,Z}^2 = \sum_{T_i=1}(Z_i-\bar{Z}_t)^2/(n_t-1)$分别为分子和分母的样本方差，$\hat{\sigma}_{t,YZ} = \sum_{T_i=1}(Y_i-\bar{Y}_t)(Z_i-\bar{Z}_t)/(n_t-1)$为分子分母的样本协方差。对照组方差也同理可得，最终由实验组和对照组之间的独立性，可以如下计算最后的方差估计量：

$$\widehat{\text{var}}(\Delta_{\text{abs}}) = \widehat{\text{var}}\left(\frac{\bar{Y}_t}{\bar{Z}_t}\right) + \widehat{\text{var}}\left(\frac{\bar{Y}_c}{\bar{Z}_c}\right)$$

进一步我们可以使用如下公式来计算p值：

$$p\text{值} = 2*(1-\Phi(|\Delta_{\text{abs}}|/\sqrt{\widehat{\text{var}}(\Delta_{\text{abs}})}))$$

其中Φ为标准正态分布的分布函数。

Bootstrap 方法

Step1： 对于$b=1,\ldots,B$做B次 Bootstrap 抽样，每次从n个样本里有放回抽取n个样本$(Y_{b1}^*,Z_{b1}^*,T_{b1}^*),\cdots,(Y_{bn}^*,Z_{bn}^*,T_{bn}^*)$，其中$Y_{bi}^*$，$Z_{bi}^*$为在第$b$次 Bootstrap 抽样中的第$i$个样本指标的分子与分母，$T_{bi}^*$为此样本受到干预的情况。

Step2： 对于$b=1,\ldots,B$，计算

$$\Delta_{\text{abs},b}^* = \frac{\sum_{i=1}^{n} Y_{bi}^* T_{bi}^* / \sum_{i=1}^{n} T_{bi}^*}{\sum_{i=1}^{n} Z_{bi}^* T_{bi}^* / \sum_{i=1}^{n} T_{bi}^*} - \frac{\sum_{i=1}^{n} Y_{bi}^* (1-T_{bi}^*) / \sum_{i=1}^{n} (1-T_{bi}^*)}{\sum_{i=1}^{n} Z_{bi}^* (1-T_{bi}^*) / \sum_{i=1}^{n} (1-T_{bi}^*)}。$$

Step3： 再基于$\Delta_{\text{abs},b}^*$，$b=1,\ldots,B$计算方差估计量为：

$$\hat{\sigma}_{\text{abs,boot}}^2 = \frac{1}{B}\sum_{b=1}^{B}(\Delta_{\text{abs},b}^* - \frac{1}{B}\sum_{i=1}^{B}\Delta_{\text{abs},i}^*)^2$$

Step4： 如果使用 Bootstrap 估分布的方式，我们可以根据 Bootstrap 的抽样结果进一步给出：

$$p\text{值} = \sum_{b=1}^{B} 1_{\{|\Delta_{\text{abs},b}^* - \Delta_{\text{abs}}| \geq |\Delta_{\text{abs}}|\}} / B$$

3.1.3 完全随机分组

由于互联网很多涉及订单的实验有几十万以上的海量数据，这种大样本情况下会广泛使用哈希函数来进行普通随机分组。而在美团的履约配送业务当中，常常会涉及人群分流以及配送城市、区域、站点等地理单元的分流，圈选出的样本量相对较少。例如，人群分流涉及的样本量较少时在 1000 左右，且由于业务约束一般只能允许留有较少的对照组，会采用相对极端的分流比例（例如 95:5）。此时如果采用普通随机分组方式，一定概率会出现 1000 人的分组中对照组只有 30 ~ 40 人的情况，实际会较大影响实验的检验功效。同样的，对于较大面积的地理单元分流，通常样本量在

100 以下，即使采用 5:5 分流，也可能出现分组较不均匀的情况。因此，在这种情况下，我们会采用完全随机分组的方式，以事先严格保证最终分组的比例与实验设定的比例一致，使实验符合预期设定。

1. 分组机制

以两组为例，对于实验中总的n个实验单元，其中$n = n_t + n_c$。通俗来说，对于给定的n个实验样本，和根据实验比例要求得到的实验组样本量n_t和对照组样本量n_c，从中随机挑选n_t个样本施加干预，剩下n_c个样本不施加干预。由于从n个样本中选取个n_t样本共有$\binom{n}{n_t}$种取法，且每种取法概率相同，满足如下分组机制：

$$P(\mathbf{T} = \mathbf{t} \mid \mathbf{Y}(1), \mathbf{Y}(0)) = 1 / \binom{n}{n_t}$$

其中$\mathbf{t} = (t_1, \ldots, t_n)$，满足$\sum_{i=1}^{n} t_i = n_t$和$\sum_{i=1}^{n} (1 - t_i) = n_c$。

2. 适用场景

- 实验单位之间相互独立；
- 适用于实验前能够确定全部进入实验的实验单元的场景；
- 对于小样本的实验推荐采用，以确保分组比例与实验功效，尤其是分组比例不均衡的情形。

3. 评估方式

评估方式与完全随机轮转的实验方式相同，都可以通过 Fisher 方法和 Neyman 方法来计算，其中 Fisher 对小样本情形的显著性计算更为准确但计算成本相对高，Neyman 方法在大样本情形中计算更为便捷。具体方法原理可以参见第四章随机轮转实验。

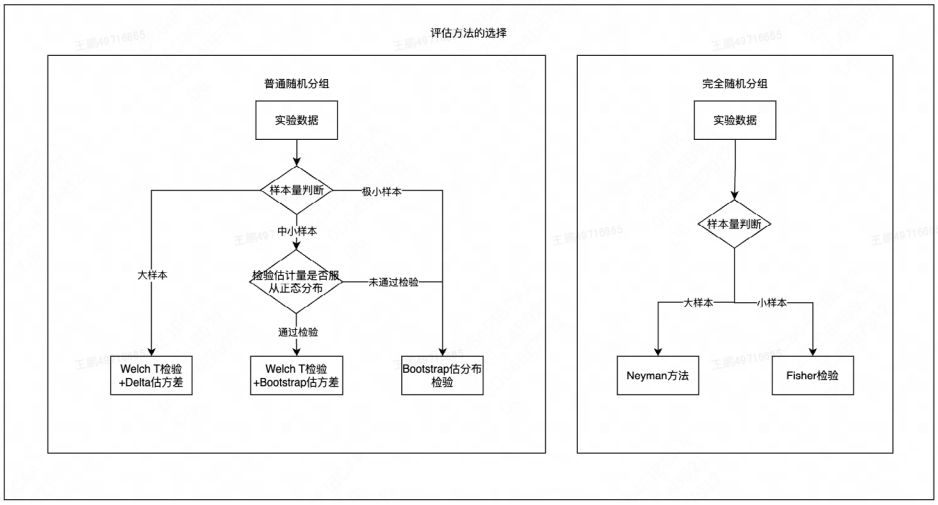


图 3-3：评估方法选择流程图

3.1.4 评估中的统计陷阱

在实验的评估中，常用的显著性计算公式并不是放之四海而皆准的，需要结合实际场景与使用方式精细判断。实验者需要关注一些潜在的统计陷阱，防止得出错误的显著性结果：

1. **分配机制陷阱：**忽视样本在实验组或对照组的分配机制，可能会导致方差计算的错误。例如业务上有时由于产品限制，会采用对流量 id 奇偶分流进行实验，这时实际上没有任何随机性，且与其他随机实验的流量不正交，容易影响其他进行随机分流实验的结果。又例如一些业务方可能会对实验单位进行分层分组以确保各层表现相似，又或者通过多次分组来使两组指标差异小于一定的阈值。这时实际上已经对分流的随机性进行了限制，使用常规公式进行显著性计算时会高估方差。在本章后续 3.3 节中会讨论分层随机分组相关内容，在 3.5 中会提及重随机化的显著性计算方式。
2. **计算口径陷阱：**不同的指标类型，比如连续型指标、比率型指标、求和型指标，或者不同的指标差值口径，比如计算绝对差值、相对差值或者 ROI 差值，其显著性计算的方式都有所不同。如果忽视这些差异，可能会导致方差

计算的错误。

3. **检验方法陷阱：**对于不同的样本量和数据分布特性，应该选用合理的分析方法。当样本量比较大时，我们根据中心极限定理可以认为数据的均值近似服从正态分布，从而可以使用 Delta 方法评估；而当样本量很小或者数据分布离正态分布差异较大时，此时使用 Delta 方法评估可能会导致方差估计不准，我们需要采用更为稳健的非参数检验方式，如 Bootstrap 估分布等方式。
4. **多重比较陷阱：**当指标个数较多时或者有多个实验组时，此时会涉及同时进行多组假设检验。单个假设检验可以控制第一类错误为α，而多个假设检验中至少一个被错误拒绝的概率却是大于α的。因此如果不考虑使用多重比较对 p 值进行修正，可能出现假阳性，影响对策略结果的判断。在第七章的高阶工具中我们会详细论述多重比较的用法。
5. **独立性陷阱：**分析单位与分流单位的不同，可能会带来错误的方差计算。通常来说能使用随机对照实验的情况中，分流单位之间是独立的，但更细的分析单位无法保证独立性，例如分流单位是用户，但我们期望分析每个用户下的订单，这时订单之间相互并不是独立的。我们在方差计算时需要注重单位之间的独立性。

3.1.5 特殊指标类型的评估方式

1. 求和型指标

在一些特殊的实验场景中，会存在无法圈选或定义实际受策略影响的实验单位，只能获取产生事实的实验单位，因此如果使用常规的均值计算方式是不合理的。例如假设在一些 uuid 随机分流实验中，我们只能取到下单用户的数据，实验组策略使部分不会下单的用户下了少量单，只取下单数据分析很可能导致实验组单量均值降低，但单量的总和是增加的。对于连续型指标的这种情况，我们采用求和计算来评估是更加符合常理的。与均值计算相比，主要差异体现在相对提升以及方差的计算上。

以两组为例（多组情况下可简单推广），假设实验中设置的实验组与对照组分流比例为 $p:(1-p)$，我们可以定义求和型指标的相对提升为：

$$\Delta_{\text{rel}} = \frac{\sum_{i=1}^{n_t} Y_{ti} / p}{\sum_{i=1}^{n_c} Y_{ci} / (1-p)} - 1$$

Delta 方法

我们同样也可以使用 Delta 方法来计算求和型指标相对提升的方差，由于推导过程相对复杂，我们这里直接给出方差公式为：

$$\sigma^2_{\text{sum,delta}} = \frac{\mu_c^2\mu_t^2 + p\mu_t^2\sigma_c^2 + (1-p)\mu_c^2\sigma_t^2}{np(1-p)\mu_c^4}$$

因此，我们可以采用下面的方差估计量：

$$\hat{\sigma}^2_{\text{sum,delta}} = \frac{\hat{\mu}_c^2\hat{\mu}_t^2 + p\hat{\mu}_t^2\hat{\sigma}_c^2 + (1-p)\hat{\mu}_c^2\hat{\sigma}_t^2}{np(1-p)\hat{\mu}_c^4}$$

其中p为分流时实验组流量占比，$\hat{\mu}_t$和$\hat{\mu}_c$为样本均值，$\hat{\sigma}_t^2$和$\hat{\sigma}_c^2$为样本方差。

Bootstrap 方法

Step1： 对于$b = 1,\ldots,B$做B次 Bootstrap 抽样，每次从n个样本里有放回抽取n个样本$(Y_{b1}^*, T_{b1}^*),\cdots,(Y_{bn}^*, T_{bn}^*)$，其中$Y_{bi}^*$为在第$b$次 Bootstrap 抽样中的第$i$个样本，$T_{bi}^*$为此样本受到干预的情况。

Step2： 对于$b = 1,\ldots,B$，计算$\Delta^*_{\text{rel},b} = \frac{\sum_{i=1}^{n} Y_{bi}^* T_{bi}^* / p}{\sum_{i=1}^{n} Y_{bi}^*(1-T_{bi}^*)/(1-p)} - 1$。

Step3： 再基于$\{\Delta^*_{\text{rel},b}\}$，$b = 1,\ldots,B$计算方差估计量为：

$$\hat{\sigma}^2_{\text{sum,boot}} = \frac{1}{B}\sum_{b=1}^{B}(\Delta^*_{\text{rel},b} - \frac{1}{B}\sum_{i=1}^{B}\Delta^*_{\text{rel},i})^2$$

2. ROI 型差值

在履约涉及花费的业务中，除了考虑常规的指标提升，还需要考察效率。从指标定义上来说，所针对的指标本质也是比率型指标，但计算的不是绝对差值，即实验组分子

/ 实验组分母 – 对照组分子 / 对照组分母，而是计算 ROI 型差值，即：

$$\Delta_{\text{roi}} = \left(\frac{\sum_{i=1}^{n} Y_i T_i}{\sum_{i=1}^{n} T_i} - \frac{\sum_{i=1}^{n} Y_i (1-T_i)}{\sum_{i=1}^{n} (1-T_i)}\right) / \left(\frac{\sum_{i=1}^{n} Z_i T_i}{\sum_{i=1}^{n} T_i} - \frac{\sum_{i=1}^{n} Z_i (1-T_i)}{\sum_{i=1}^{n} (1-T_i)}\right) = \frac{\bar{Y}_t - \bar{Y}_c}{\bar{Z}_t - \bar{Z}_c}$$

即口径为（实验组指标 – 对照组指标）/（实验组花费 – 对照组花费）。在显著性方面，我们也可以通过 Delta 方法和 Bootstrap 方法两种方式来进行计算。

Delta 方法

类似的，在原假设 $\mu_{Y,t} = \mu_{Y,c}$ 的情况下，我们可以给出如下方差的估计量：

$$\hat{\sigma}^2_{\text{roi,delta}} = \frac{\hat{\sigma}^2_{Y,t} / n_t + \hat{\sigma}^2_{Y,c} / n_c}{(\hat{\mu}_{Z,t} - \hat{\mu}_{Z,c})^2}$$

其中 $\hat{\mu}_{Z,t}$ 和 $\hat{\mu}_{Z,c}$ 分别为分母 Z 的实验组和对照组的样本均值，$\hat{\sigma}^2_{Y,t}$ 和 $\hat{\sigma}^2_{Y,c}$ 分别为分子 Y 的实验组和对照组样本方差。实际上，抛开原假设，我们也能在一般情况下推导 Delta 公式，但由于相对复杂我们在这不再详述，使用原假设下的方差公式已经能较好的控制住第一类错误。

Bootstrap 方法

Step1： 对于 $b = 1,\ldots,B$ 做 B 次 Bootstrap 抽样，每次从 n 个样本里有放回抽取 n 个样本 $(Y^*_{b1}, Z^*_{b1}, T^*_{b1}), \cdots, (Y^*_{bn}, Z^*_{bn}, T^*_{bn})$，其中 Y^*_{bi} 为在第 b 次 Bootstrap 抽样中的第 i 个样本，T^*_{bi} 为此样本受到干预的情况；

Step2： 对于 $b = 1,\ldots,B$，计算

$$\Delta^*_{\text{roi},b} = \left(\frac{\sum_{i=1}^{n} Y^*_{bi} T^*_{bi}}{\sum_{i=1}^{n} T^*_{bi}} - \frac{\sum_{i=1}^{n} Y^*_{bi} (1-T^*_{bi})}{\sum_{i=1}^{n} (1-T^*_{bi})}\right) / \left(\frac{\sum_{i=1}^{n} Z^*_{bi} T^*_{bi}}{\sum_{i=1}^{n} T^*_{bi}} - \frac{\sum_{i=1}^{n} Z^*_{bi} (1-T^*_{bi})}{\sum_{i=1}^{n} (1-T^*_{bi})}\right);$$

Step3： 再基于 $\{\Delta^*_{\text{roi},b}\}$，$b = 1,\ldots,B$ 计算方差估计量为：

$$\hat{\sigma}^2_{\text{roi,bootstrap}} = \frac{1}{B}\sum_{b=1}^{B}(\Delta^*_{\text{roi},b} - \frac{1}{B}\sum_{i=1}^{B}\Delta^*_{\text{roi},i})^2$$

3.1.6 随机对照实验配套功能

1. 验证样本量均衡的 SRM 检验

验证实验中的样本分布是否与预期一致的检验，被称为 SRM（Sample Ratio Mismatch）检验。如果 SRM 检验不通过，那么除非我们能够诊断 SRM 的原因在哪里，否则结果是不可信的。因为 SRM 检验不通过时，可能由于一些潜在原因导致分组的随机性被破坏，从而违反随机对照实验的基本假设。SRM 的成因多种多样，原因大致可分为五类：

1. 实验分配阶段，例如流量未正确分桶、随机分组方法有问题等；
2. 策略实质性阶段，例如各组的准入条件发生了变化、数据传递丢失等；
3. 数据处理阶段，例如没有对未发生事实的单位补零等；
4. 实验分析阶段，例如使用了错误的分析时间周期，使用了错误的过滤条件；
5. 其他干预手段，例如遭受黑客攻击。

假设实验共分为k个组，共有n个样本，其中满足限制条件：

$$\sum_{i=1}^{k} n_i^o = \sum_{i=1}^{k} n_i = n$$

其中n_i^o为第i组的实际观测到的样本量，n_i为第i组的按设计比例预计的样本量，可以考虑如下卡方统计量：

$$\chi^2 = \sum_{i=1}^{k} \frac{(n_i^o - n_i)^2}{n_i}$$

当样本量n足够大时，这个卡方统计量将会趋近于自由度为$(k-1)$的卡方分布。当卡方检验不通过时，我们认为出现了 SRM 的情况，需要进一步排查原因。

2. MDE 与最小样本量计算

实验在当前条件下能有效检测的指标差异幅度即为 MDE。在实验报告分析阶段计算 MDE 来判断不显著的指标结论是否是由于样本量不足所导致，避免实验在灵敏度不足的情况下得到非显著结论，而做出认为策略没有效果的误判。

具体的双边假设检验下 MDE 的计算公式如下：

$$\mathrm{MDE} = (z_{1-\alpha/2} + z_{1-\beta}) \cdot \sigma_{\mathrm{est}}$$

其中$z_{1-\alpha/2}$，$z_{1-\beta}$分别为标准正态的$1-\alpha/2$，$1-\beta$分位数，当显著性水平为$\alpha = 0.05$时，$z_{1-\alpha/2} \approx 1.96$，当功效$1-\beta = 80\%$时，$z_{1-\beta} \approx 0.84$，而$\sigma_{\mathrm{est}}$为估计量（比如连续型指标的相对提升或者比率型指标的绝对提升）的标准差。

对于每个实验我们可以根据策略制定一个合理的预期提升值，如果实验 MDE< 预期提升值，则说明策略效果不显著；如果 MDE> 预期提升值，则说明当前的样本量不足以检测出预期提升值，建议增大样本量，还有检测出显著的可能。在实验前我们可以根据实验前同质性检验的 MDE 作为实验 MDE 一个粗糙的估计量，再通过最小样本量计算公式给出实验建议样本量。对于比率型指标，具体最小样本量计算公式如下：

$$n_t = k n_c$$

$$n_c = \left(1 + \frac{1}{k}\right)\left(\frac{\hat{\sigma}(z_{1-\alpha/2} + z_{1-\beta})}{\Delta_{\text{预期绝对提升}}}\right)^2$$

其中$\hat{\sigma}$为实验前指标的标准差。对于连续型指标，具体最小样本量计算公式如下：

$$n_t = k n_c$$

$$n_c = \left(1 + \frac{1}{k}\right)\left(\frac{\hat{\sigma}(z_{1-\alpha/2} + z_{1-\beta})}{\hat{\mu}\Delta_{\text{预期绝对提升}}}\right)^2$$

其中$\hat{\sigma}$为实验前指标的标准差，$\hat{\mu}$为实验前指标的均值。

3.2 提高实验功效的办法

在线上 AB 实验中，常常会出现实验功效不足而检测不出显著性的情况。一种最常用的方式是增加样本量来提高实验功效，但这会增大实验成本。另一种提高检测灵敏度的方式是创建一个方差更小并能捕捉相同信息的评估指标。方差缩减的方式有很多，例如 CUPED、分层分析、回归调整、配对实验等。在这节我们主要介绍 CUPED（Controlled Experiment Using Pre-Experiment Data）在履约和外卖实验中的一些应用，在下节中我们会讨论分层随机分组和配对随机分组。在履约和外卖的实验场景中，CUPED 能够降低 50% 左右的策略效果估计量方差，大大提升检验灵敏度并减少实验所需样本量。

3.2.1 CUPED 降方差原理

对于连续型指标Y，Deng et al. (2013)[1] 考虑引入协变量来减少在线实验中估计量方差。在通常的实验评估中，我们一般采用样本均值（即求平均数）作为其中一组的均值估计量。CUPED 的方法主要思路是，使用指标Y和与干预无关的协变量X，来构造一个新的无偏估计量使其比样本均值的方差更小。具体构造如下：

$$\tilde{Y} = \bar{Y} - \theta(\bar{X} - \mathbf{E}(X))$$

其中θ取如下取值时，可以使$\tilde{Y}$的方差最小：

$$\theta = \frac{\mathrm{cov}(X,Y)}{\mathrm{var}(X)}$$

这时，可以借助Y与X的相关系数ρ来表示新构造指标$\tilde{Y}$与原指标Y方差之间的关系：

$$\mathrm{var}(\tilde{Y}) = \mathrm{var}(\bar{Y})(1-\rho^2)$$

直观上来说，指标Y与协变量X越相关，越能捕捉到个体的差异信息，方差降低效果也就越明显。在随机对照实验的实际应用中，我们通常将实验前的指标作为协变量X，对于实验组和对照组来说，X的期望是相同的，因此我们可以构造 ATE 的一个无偏估计：

$$\Delta_{\text{cuped}} = (\bar{Y}_t - \hat{\theta}(\bar{X}_t - \bar{X})) - (\bar{Y}_c - \hat{\theta}(\bar{X}_c - \bar{X}))$$

其中：

$$\hat{\theta} = \left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} \right) / \left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1} \right)$$

这实际上也就是指标 Y 对协变量 X 做回归的最小二乘估计量。

我们可以总结出，CUPED 降方差主要有以下的适用条件：

1. **有实验前可用的数据：**例如用户、地理单元等实验单位，都有较长周期的实验前历史数据可用。对于订单等在实验中新产生的实验单元，没有历史数据可用，使用 CUPED 意义不大。
2. **指标数据表现稳定：**由上述的 CUPED 降方差思想可以看出，当使用的协变量（一般是实验前的指标数据）与实验数据相关性很高时，降方差效果越好。因此，当指标数据相对稳定，不会随时间变化或者开展实验而出现剧烈波动时，会呈现实验后数据与实验前数据有较好相关性的情况，从而能更多降低方差，提升实验功效。
3. **选取的协变量对于实验组和对照组期望一致：**当协变量受到干预变量影响的时候，此时导出的 CUPED 估计量不再是策略效果提升的无偏估计，可能存在偏差。因此实验前的指标值作为天然的满足不受干预影响的协变量，常常是 CUPED 协变量的首选。

回归调整是统计中降低实验效果估计量方差的一种常用方法，其主要思想是利用实验单位相关的协变量信息，通过回归等技术构造一个比常规均值相减方差更小的 ATE 估计量。CUPED 本质上是一种带协变量的回归调整方法，上述 Δ_{cuped} 实际上类似于求解下列带协变量回归得到的 β_1 的最小二乘估计 $\hat{\beta}_1$：

$$Y_i = \alpha_0 + \beta_1 T_i + \beta_2 X_i$$

实际上，我们可以进一步考虑带协变量与干预变量交互项回归中β_1的最小二乘估计$\hat{\beta}_{1,\text{interact}}$：

$$Y_i = \alpha + \beta_1 T_i + \beta_2 X_i + \beta_3 T_i (X_i - \bar{X}) + \epsilon_i$$

这个估计量的另一种等价表达形式为：

$$\Delta_{\text{cuped}} = (\bar{Y}_t - \hat{\theta}_t(\bar{X}_t - \bar{X})) - (\bar{Y}_c - \hat{\theta}_c(\bar{X}_c - \bar{X}))$$

其中$\hat{\theta}_t$为实验组指标Y对协变量X做回归的最小二乘估计量，$\hat{\theta}_c$为对照组指标Y对协变量X做回归的最小二乘估计量。理论上 Lin（2013）[2] 指出$\hat{\beta}_{1,\text{interact}}$的方差小于等于$\hat{\beta}_1$的方差，在流量设置为 50%：50% 或者$\hat{\theta}_t$与$\hat{\theta}_c$渐近相等（例如是在 AA 实验阶段）时$\hat{\beta}_{1,\text{interact}}$与$\hat{\beta}_1$的方差渐近相等。这意味着带干预变量 * 协变量的交互项的回归调整方法能保证在各种流量分配机制下可更有效的降低方差，由此我们可以对于实验组和对照组分别引入不同的回归系数。下面我们将分别介绍连续型指标和比率型指标中我们对 CUPED 方法的应用。

3.2.2 连续型指标和比率型指标 CUPED 方法的应用

在实际业务中，我们需要考虑拓展 CUPED 方法来降低连续型指标相对提升Δ_{rel}的估计量方差。具体来说，定义连续型新指标：

$$\tilde{Y} = Y - \hat{\theta}(X - \bar{X})$$

则可考虑：

$$\Delta_{\text{rel,cuped}} = \frac{\bar{Y}_t - \hat{\theta}(\bar{X}_t - \bar{X})}{\bar{Y}_c - \hat{\theta}(\bar{X}_c - \bar{X})} - 1$$

来估计相对提升率，并可以直接使用基于新指标$\tilde{Y}$的 Bootstrap 或者 Delta 方法来估计方差。由于在随机对照实验中实验组和对照组是独立的，因此分别降低分子分母的方差我们即可以降低分子分母相除后随机变量的方差，在实际业务应用中很多场景我们验证都能够实现方差的降低。

在业务中会遇到很多比率型指标的评估，这时我们无法直接使用经典的 CUPED 降方差方法。我们在常规 CUPED 方法的基础上，进一步建立了比率型指标的降方差方式。对于比率型指标，我们探索了如下三种基于回归调整的 CUPED 降方差的方法，其中二元回归系数调整 CUPED 方法和新 CUPED 方法都通过严格证明可以降低比率型指标的方差。

1. 一元回归系数调整 CUPED 方法

（1）一元同系数：对实验组和对照组的分子分母加入相同的回归系数进行调整，可考虑估计量：

$$\Delta_{\text{abs,cuped}} = \frac{\bar{Y}_t - \hat{\theta}(\bar{X}_t - \bar{X})}{\bar{Z}_t - \hat{\gamma}(\bar{U}_t - \bar{U})} - \frac{\bar{Y}_c - \hat{\theta}(\bar{X}_c - \bar{X})}{\bar{Z}_c - \hat{\gamma}(\bar{U}_c - \bar{U})}$$

这里U为分母Z的对应的实验前数据，γ代表实验组 + 对照组的Z关于U的带截距项的回归系数。注意到此时实验组和对照组的调整回归系数$\hat{\theta}$和$\hat{\gamma}$是相同的。应用上，我们可以定义新的指标分子分母：

$$\tilde{Y} = Y - \hat{\theta}(X - \bar{X})$$

$$\tilde{Z} = Z - \hat{\gamma}(U - \bar{U})$$

再来进行随机对照实验相关的 Delta 或者 Bootstrap 评估。

（2）一元不同系数：我们可以考虑干预变量 * 协变量加入回归，即对实验组和对照组的分子分母加入不同的回归系数进行调整，可考虑估计量：

$$\Delta_{\text{abs,cuped}} = \frac{\bar{Y}_t - \hat{\theta}_t(\bar{X}_t - \bar{X})}{\bar{Z}_t - \hat{\gamma}_t(\bar{U}_t - \bar{U})} - \frac{\bar{Y}_c - \hat{\theta}_c(\bar{X}_c - \bar{X})}{\bar{Z}_c - \hat{\gamma}_c(\bar{U}_c - \bar{U})}$$

这里U为分母Z的对应的实验前数据，γ_t代表实验组Z关于U带截距项的回归系数，γ_c代表实验组Z关于U带截距项的回归系数。可以看出，此时实验组和对照组的回归调整系数是不相同的。应用上我们可以定义新的指标分子分母，对于实验组有：

$$\tilde{Y} = Y - \hat{\theta}_t(X - \bar{X})$$

$$\tilde{Z} = Z - \hat{\gamma}_t(U - \bar{U})$$

而对于对照组有：

$$\tilde{Y} = Y - \hat{\theta}_c(X - \bar{X})$$

$$\tilde{Z} = Z - \hat{\gamma}_c(U - \bar{U})$$

再来进行随机对照实验相关的 Delta 或者 Bootstrap 评估。

2. 二元回归系数调整 Cuped 方法

（1）二元同系数：有一个很自然的想法是将做回归时的协变量拓展为分子分母两者的实验前数据，可考虑估计量：

$$\Delta_{\text{abs,cuped}} = \frac{\bar{Y}_t - \hat{\theta}_1(\bar{X}_t - \bar{X}) - \hat{\gamma}_1(\bar{U}_t - \bar{U})}{\bar{Z}_t - \hat{\theta}_2(\bar{X}_t - \bar{X}) - \hat{\gamma}_2(\bar{U}_t - \bar{U})} - \frac{\bar{Y}_c - \hat{\theta}_1(\bar{X}_c - \bar{X}) - \hat{\gamma}_1(\bar{U}_c - \bar{U})}{\bar{Z}_c - \hat{\theta}_2(\bar{X}_c - \bar{X}) - \hat{\gamma}_2(\bar{U}_c - \bar{U})}$$

这里X为Y对应的实验前数据，U为Z对应的实验前数据，$\hat{\theta}_1$，$\hat{\gamma}_1$分别代表实验组 + 对照组的Y关于协变量X，U的（带截距项）对应的X与U的回归系数，$\hat{\theta}_2$，$\hat{\gamma}_2$分别代表实验组 + 对照组的Z关于协变量X，U的（带截距项）对应的X与U的回归系数。此时实验组和对照组的回归系数是相同的。应用上我们可以定义新的指标分子分母：

$$\tilde{Y} = Y - \hat{\theta}_1(X - \bar{X}) - \hat{\gamma}_1(U - \bar{U})$$

$$\tilde{Z} = Z - \hat{\theta}_2(X - \bar{X}) - \hat{\gamma}_2(U - \bar{U})$$

再来进行随机对照实验相关的 Delta 或者 Bootstrap 评估。

（2）二元不同系数：类似的，使用拓展带干预变量与协变量交互项的回归调整方法，可考虑估计量：

$$\Delta_{\text{abs,cuped}} = \frac{\bar{Y}_t - \hat{\theta}_{t1}(\bar{X}_t - \bar{X}) - \hat{\gamma}_{t1}(\bar{U}_t - \bar{U})}{\bar{Z}_t - \hat{\theta}_{t2}(\bar{X}_t - \bar{X}) - \hat{\gamma}_{t2}(\bar{U}_t - \bar{U})} - \frac{\bar{Y}_c - \hat{\theta}_{c1}(\bar{X}_c - \bar{X}) - \hat{\gamma}_{c1}(\bar{U}_c - \bar{U})}{\bar{Z}_c - \hat{\theta}_{c2}(\bar{X}_c - \bar{X}) - \hat{\gamma}_{c2}(\bar{U}_c - \bar{U})}$$

这里X为Y对应的实验前数据，U为Z对应的实验前数据，$\hat{\theta}_{t1}$，$\hat{\gamma}_{t1}$（$\hat{\theta}_{c1}$，$\hat{\gamma}_{c1}$）分别

代表实验组（对照组）的Y关于协变量X，U的（带截距项）对应的X与U的回归系数，$\hat{\theta}_{t2}$，$\hat{\gamma}_{t2}$（$\hat{\theta}_{c2}$，$\hat{\gamma}_{c2}$）分别代表实验组（对照组）的Z关于协变量X，U的（带截距项）对应的X与U的回归系数。此时实验组和对照组的回归系数是不同的。应用上我们可以定义新的指标分子分母，对于实验组有指标变换：

$$\tilde{Y} = Y - \hat{\theta}_{t1}(X - \bar{X}) - \hat{\gamma}_{t1}(U - \bar{U})$$

$$\tilde{Z} = Z - \hat{\theta}_{t2}(X - \bar{X}) - \hat{\gamma}_{t2}(U - \bar{U})$$

而对于对照组有指标变换：

$$\tilde{Y} = Y - \hat{\theta}_{c1}(X - \bar{X}) - \hat{\gamma}_{c1}(U - \bar{U})$$

$$\tilde{Z} = Z - \hat{\theta}_{c2}(X - \bar{X}) - \hat{\gamma}_{c2}(U - \bar{U})$$

再来进行随机对照实验相关的 Delta 或者 Bootstrap 评估。

3. 新 CUPED 方法

与前面部分对比率型指标的分子分母分别进行降方差操作不同的是，新 CUPED 方法直接对比率型指标整体进行降方差。核心思想参考了 Deng et al. (2013) 对于用户随机流时实验单元和分析单元不一致的情况，将其拓展到一般比率型指标上的应用。具体构造的无偏估计量如下：

$$\tilde{R} = \frac{\bar{Y}}{\bar{Z}} - \theta\frac{\bar{X}}{\bar{U}} + \theta\mathbf{E}\left(\frac{\bar{X}}{\bar{U}}\right)$$

类似推导可知上述估计量方差最小时取θ为：

$$\begin{aligned}
\theta &= \operatorname{cov}\left(\frac{\bar{Y}}{\bar{Z}}, \frac{\bar{X}}{\bar{U}}\right) / \operatorname{var}\left(\frac{\bar{X}}{\bar{U}}\right) \\
&\approx \operatorname{cov}\left(\frac{\bar{Y}}{\mu_Z} - \frac{\mu_Y \bar{Z}}{\mu_Z^2}, \frac{\bar{X}}{\mu_U} - \frac{\mu_X \bar{U}}{\mu_U^2}\right) / \operatorname{var}\left(\frac{\bar{X}}{\mu_U} - \frac{\mu_X \bar{U}}{\mu_U^2}\right) \\
&= \frac{(1/\mu_Z, -\mu_Y/\mu_Z^2, 0, 0)\boldsymbol{\Sigma}(0, 0, 1/\mu_U, -\mu_X/\mu_U^2)^T}{(0, 0, 1/\mu_U, -\mu_X/\mu_U^2)\boldsymbol{\Sigma}(0, 0, 1/\mu_U, -\mu_X/\mu_U^2)^T}
\end{aligned}$$

我们可以基于此考虑绝对提升的估计量：

$$\Delta_{\text{abs,cuped}} = \left(\frac{\bar{Y}_t}{\bar{Z}_t} - \theta\frac{\bar{X}_t}{\bar{U}_t}\right) - \left(\frac{\bar{Y}_c}{\bar{Z}_c} - \theta\frac{\bar{X}_c}{\bar{U}_c}\right)$$

对于新 CUPED 估计量的方差计算或者假设检验，一方面可以根据独立性有：

$$\text{var}(\Delta_{\text{abs,cuped}}) = \text{var}\left(\frac{\bar{Y}_t}{\bar{Z}_t} - \theta\frac{\bar{X}_t}{\bar{U}_t}\right) + \text{var}\left(\frac{\bar{Y}_c}{\bar{Z}_c} - \theta\frac{\bar{X}_c}{\bar{U}_c}\right)$$

再使用 Delta 方法分别计算两项的方差，以实验组为例，有：

$$\text{var}\left(\frac{\bar{Y}_t}{\bar{Z}_t} - \theta\frac{\bar{X}_t}{\bar{U}_t}\right) = \frac{1}{n}[(1/\mu_Z, -\mu_Y/\mu_Z^2, 0, 0)\boldsymbol{\Sigma}(1/\mu_Z, -\mu_Y/\mu_Z^2, 0, 0)^T + \theta^2(0,0,1/\mu_U,$$
$$-\mu_X/\mu_U^2)\boldsymbol{\Sigma}(0,0,1/\mu_U, -\mu_X/\mu_U^2)^T - 2\theta(1/\mu_Z, -\mu_Y/\mu_Z^2, 0, 0)\boldsymbol{\Sigma}(0,0,1/\mu_U, -\mu_X/\mu_U^2)^T]$$

对照组结果类似可得，其中 θ 可以替换为：

$$\hat{\theta} = \frac{(1/\bar{Z}, -\bar{Y}/\bar{Z}^2, 0, 0)\hat{\boldsymbol{\Sigma}}(0,0,1/\bar{U}, -\bar{X}/\bar{U}^2)^T}{(0,0,1/\bar{U}, -\bar{X}/\bar{U}^2)\hat{\boldsymbol{\Sigma}}(0,0,1/\bar{U}, -\bar{X}/\bar{U}^2)^T}$$

而 $\mu_Y, \mu_Z, \mu_X, \mu_U, \boldsymbol{\Sigma}$ 也可以由 $\bar{Y}, \bar{Z}, \bar{X}, \bar{U}, \hat{\boldsymbol{\Sigma}}$ 替代，分别为实验组 + 对照组所有样本的关于 Y, Z, X, U 的样本均值以及 (Y, Z, X, U) 的样本协方差矩阵。

3.3 进一步保证同质性的实验方式

同质性检验是一种用于确保在实验前通过随机分流后，实验组和对照组之间没有显著差异的手段。如果同质性检验未通过，则意味着在实验开始前，两组之间存在系统性差异。通常，对于那些在实验前后高度相关的指标来说，如果在实验前未能达到同质性，这些指标在实验后也可能表现出系统性差异，从而影响实验结论的准确性，无法真实反映策略效果。

随机对照实验是 AB 实验中最基本且可信度最高的方式，在样本量充足的情况下，能够有效平衡两组之间的协变量分布，从而通过同质性检验。然而，在特定的实验条件和业务需求下，简单的随机分流可能仍然无法完全满足实验需求。例如，在样本量较

小的情况下（如几百甚至几十），单次随机分流难以轻松获得同质的分组，即使同质，组间差异可能仍然较大。

此外，业务上不仅关注实验组和对照组的整体同质性，还常常进一步关注按某些重要特征分层后的同质性，并对各层进行深入分析以获得精细化的实验结果。在有限样本量下，简单随机分流往往难以同时保证各层的同质性，尤其是在分层较多的情况下。对于一些不可预测和不可控的因素，我们也无法通过实验前的同质性验证来确保两组在实验期间的这些因素相似，而策略的触发条件或使用效果往往依赖于这些特殊因素。为了应对样本量有限、分层分析以及不可控因素等挑战，我们探索并制定了一些更加精细化的分组策略，以进一步保证同质性，从而提高实验的科学性和结果的可信度。在实际业务场景中，我们已经积累了分层随机分组、配对随机分组、协变量自适应分组等方法，以进一步确保同质性。接下来，我们将逐一介绍这些方法的原理和适用场景。

3.3.1 分层随机分组

在美团的实验场景中，分层随机分组被广泛应用于验证不同分层的运营策略的效果。通过这种方法，我们可以根据特定特征进行分层，在每一层进行完全随机分组，确保层内样本在主要特征上具有相似分布以满足同质性，从而减少潜在的混杂因素对每一层实验结果的影响，便于探查策略在不同层的效果，进行精细化分析。此外分层随机分组在数学上实际等价于 CUPED 中将协变量设置为分层协变量的示性函数的情形，在层间差异显著时能够有效降低方差，提高实验功效。分层抽样的核心思想是将总体样本根据分层协变量情况（如年龄、性别、城市规模等）分为若干独立的层，然后在每层内分别进行随机分组，并聚合各层结果以获得最终估计。以下是关于分层随机分组的一些应用经验与建议。

1. 优点

- **提高同质性：**通过在层内随机化，保证了实验组和对照组在分层变量上的相似性，从而提高了实验的同质性。

- **减少偏差：**进一步有效控制潜在的混杂变量，尤其减少了分层协变量对实验结果的影响。
- **提高统计功效：**由于减少了组间差异，分层随机分组通常能够提高统计分析的功效。

2. 局限性

- **分层变量选择的挑战：**选择合适的分层变量需要深入了解研究对象，使得分层后各层群体差异明显，且不当选择可能导致分层效果不明显或过多。
- **层数和样本量的限制：**分层过多可能导致每层样本量偏少，影响统计分析的有效性，尤其在样本总量有限时。
- **实施复杂性：**分层随机分组增加了实验设计和实施的复杂性。当分层层数很多时，大规模的实现通常费时费力，并且会带来分组表达式繁冗等缺点。

3. 适用场景

使用分层随机分组实验设计需要满足以下条件：

- **独立性：**实验组和对照组必须是相互独立的，分层结果与实验策略的实施相互独立。
- **存在分层差异：**实验单元在某个协变量上有较为明显的差异。
- **样本量要求：**每一层下的实验组和对照组样本量应满足最小样本量要求，如果层内人数过少，缺乏某一组，则需重新划分层。

分层随机分组是一种有效的实验设计方法，旨在通过控制层内同质性和层间异质性来进一步保证同质性，提高实验结果的准确性和统计功效。需要注意的是，分组机制应当与评估方式对应，分层分组的实验设计在评估时采取分层评估分析。一个常见的误区是采用分层随机分组却采用常规的随机对照评估方式，这样会导致显著性计算错误。如下图左边是使用分层随机但采用完全随机的方式进行评估，会较为严重的高估方差，AA 模拟中 p 值不再服从均匀分布，实验功效降低（但能控制第一类错误），而右边是使用分层评估方式，AA 模拟中p值表现为正常的均匀分布。

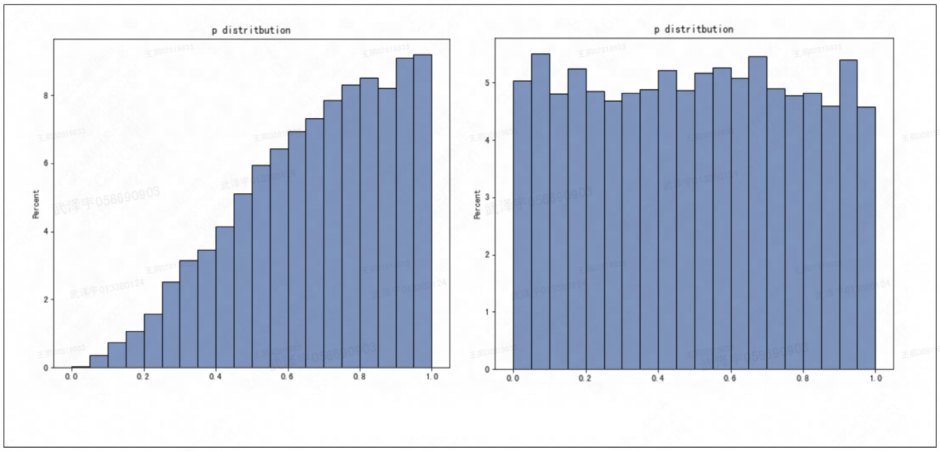


图 3-4：错误的评估方式导致方差高估

在每一层使用随机对照实验评估方式的基础上，我们可以采用 Neyman 方法计算每一层的指标差异和方差，并使用统合分析工具进行加权，从而获得整体效果的评估。这种方法也能够有效支持比率型指标的分层随机分组实验设计与评估。以下是其基本原理和实现过程。

1. 分组机制

所谓分层随机分组，就是将N个样本根据定义好的分层变量（如配送能力），（不重不漏地）划分为$J \geq 2$层，每层的样本量为$N(j)$。然后在每一层中，也制定好总共恰有$N_t(j)$个实验单元在实验组，剩余都在对照组，并保证每一层内实验组与对照组同质。也就是在每层中分别做完全随机分组。具体来说，用数学公式表示N个实验单元分层随机分配机制$W = \{W_1, \ldots, W_N\}$为：

$$N = \sum_{j=1}^{J} N(j), \quad N_t = \sum_{j=1}^{J} N_t(j), \quad W^+ = \left\{ W \mid \sum_{i:B_i=j}^{N} W_i = N_t(j) \quad \text{对于} j = 1, 2, \ldots, J \right\},$$

$$P(W \mid X, Y(0), Y(1)) = \begin{cases} \prod_{j=1}^{J} \binom{N(j)}{N_t(j)}^{-1}, & \text{如果 } W \in W^+ \\ 0, & \text{如果 } W \notin W^+ \end{cases}$$

其中W_+表示所有可能的分配方式的集合，B_i表示第i个实验单元所处的层，$W_i \in \{0,1\}$表示第i个实验单元是否落在实验组，即$W_i=1$时落在实验组，$W_i=0$时落在对照组，X表示考虑的协变量，$Y(0),Y(1)$分别表示实验单元落在对照组、实验组的反事实结果，在给定$X,Y(0),Y(1)$的情况下，每种可能的分配方式都是等概率的，为$\prod_{j=1}^{J}\left(\begin{matrix}N(j)\\N_t(j)\end{matrix}\right)^{-1}$。

2. 评估方式

Neyman 方法计算方差与p值

在实验评估阶段，我们可以通过 Neyman 方法，分别计算各层平均因果效应及其方差估计：

$$\hat{\tau}^{\text{dif}}(j)=\bar{Y}_t^{\text{obs}}(j)-\bar{Y}_c^{\text{obs}}(j)$$

其中，$\bar{Y}_t^{\text{obs}}(j)$表示观测到的第$j$层的实验组目标指标均值，$\bar{Y}_c^{\text{obs}}(j)$表示观测到的第$j$层的实验组目标指标均值。记$\hat{V}^{\text{Neyman}}(j)$是根据 Neyman 方法计算得到的第层样本的方差。

以样本量加权为例，可以加权求和得到 Population Average Effect（PAE）[3] 为：

$$\hat{\tau}^{\text{strat}}=\sum_{j=1}^{J}\lambda_j\hat{\tau}(j),\text{且}\hat{V}^{\text{Neyman}}=\sum_{j=1}^{J}(\lambda_j)^2*\hat{V}^{\text{Neyman}}(j)\text{其中}\lambda_j=\frac{N(j)}{N}$$

更多加权方式可以参见第七章高阶实验工具。再基于$\hat{\tau}^{\text{strat}}$近似正态分布来构造置信区间与$p$值等。

Fisher 方法计算p值

类似地，在用 Fisher 精确p值算法计算p值时，在抽样、计算等环节也会考虑分层：

Step1： 给定抽样次数K（通常为 1000 次）；

Step2： 对于轮次k，从 1 到K进行循环；

基于分层实验设计环节相同的抽样机制给出一个分配方式W^k：即对于$J\geq 2$个层，每

层的$N(j)$个样本中选取$N_t(j)$个实验单元在实验组，剩余都在对照组。

计算统计量：

$$T^{\text{dif},\lambda}=\sum_{j=1}^{J}\lambda(j)\left(\overline{Y}_t^{\text{obs}}(j)-\overline{Y}_c^{\text{obs}}(j)\right)\text{或}\,T^{\text{dif},\lambda}=\left|\sum_{j=1}^{J}\lambda(j)\left(\overline{Y}_t^{\text{obs}}(j)-\overline{Y}_c^{\text{obs}}(j)\right)\right|$$

Step3： 近似计算p值：

$$\hat{p}=\frac{1}{K}\sum_{k=1}^{K}1_{\left\{T^{\text{dif},k}\geq T^{\text{dif},\text{obs}}\right\}}\text{或}\hat{p}=\frac{1}{K}\sum_{k=1}^{K}1_{\left\{\left|T^{\text{dif},k}\right|\geq\left|T^{\text{dif},\text{obs}}\right|\right\}}$$

3.3.2 配对随机分组

配对随机分组是指通过将实验对象根据某些关键特征进行配对，每对中的一个对象被随机分配到实验组，另一个则进入对照组，以确保实验组和对照组在这些特征上的均衡性。此方法通过控制个体间的差异来进一步保证同质性，提高实验功效，特别适用于样本间存在显著异质性的研究。当样本量较少时，或者实验策略的触发因素不可控，受到外部环境因素（如地理位置、时间段等）影响时，配对可以尽可能控制实验组和对照组的外部环境因素相似，也能够保证其他关注特征尽可能同质，减少这些因素对于实验结论的影响。以下是关于配对随机分组的应用经验。

1. 优点

配对随机分组的实验方式在许多方面表现出色，但也可能有一些局限性。我们详细总结了其优缺点，首先其优点很明显：

- **控制个体差异，进一步保证同质性：**通过配对相似的个体，配对随机分组能够有效控制个体差异，进一步保证实验组、对照组的同质性，提高实验结果的可信度。
- **提高统计功效：**在样本量有限的情况下，配对可以提高统计功效，使得实验更容易观察到显著效果。
- **应对小样本量：**在样本量较小的情况下，配对随机分组仍能有效地平衡组间特征。

2. 局限性

- **配对复杂性：**找到合适的配对对象可能需要额外的时间和资源，尤其在样本特征复杂或数量庞大时。
- **样本利用率降低：**如果无法为某些对象找到合适的配对，这些对象可能无法被纳入实验，导致样本利用率降低。

3. 适用场景

- **个体差异显著的场景：**当实验对象之间存在显著的个体差异时，配对随机分组有助于控制这些差异，保证同质性。
- **样本量有限的场景：**在样本量较小的研究中，配对可以提高实验的统计功效。
- **外部环境影响大的场景：**在实验结果可能受到外部因素（如地域差异）显著影响的情况下，配对有助于提高结果的内部有效性。
- **复杂多因素干扰的场景：**当涉及多个干扰因素时，配对可以通过对这些因素的匹配来提高实验设计的复杂性和精确性。

下面主要介绍配对随机分组的基本原理。

1. 分组机制

配对随机分组是指，首先（根据某些特征）将N个样本划分为$J = N / 2$个层 / 对（Pair），对于每层恰好只有两个样本，且一个样本被随机分配到实验组，另一个落在对照组，以保证实验组和对照组的同质性。记B_i为第i个个体所处的层 / 对，则数学公式表示配对后的随机分配机制为：

$$W_+ = \left\{ W \,\middle|\, \sum_{i:B_i=j} W_i = 1, \text{for } j = 1,2,...,N/2 \right\},\ P\left(W \,\middle|\, X, Y(0), Y(1)\right) = \begin{cases} 2^{-N/2}, & W \in W_+ \\ 0 & W \notin W_+ \end{cases}$$

其中W_+表示所有可能的分配方式的集合，B_i表示第i个实验单元所处的层 / 对（Pair），$W_i \in \{0,1\}$表示第i个实验单元是否落在实验组，即$W_i = 1$时落在实验组，$W_i = 0$时落在对照组，X表示配对时考虑的协变量，$Y(0), Y(1)$分别表示实验单元落在对照组、实验组的反事实结果，在给定的$X, Y(0), Y(1)$情况下，每种可能的分配方

式都是等概率的，为$2^{-N/2}$。

2. 评估方式

(1) 连续型指标评估

考虑连续型指标绝对提升$\bar{Y}_t - \bar{Y}_c = \sum_i^{n_t} Y_{ti} / n_t - \sum_i^{n_c} Y_{ci} / n_c$即【实验组指标的平均值 - 对照组指标平均值】的评估情形，其中$\bar{Y}_t$代表实验组指标的平均值，$\bar{Y}_c$代表对照组指标的平均值，Y_{ti}为实验组第i个实验单元的指标值，Y_{ci}为对照组第i个实验单元的指标值，n_t为实验组的样本量，n_c为对照组的样本量，配对随机分组实验下$n_t = n_c$。

连续型指标相对提升之 Fisher 方法：

相对提升点估计可考虑用$T^{\text{lift,obs}} = \hat{\tau}^{\text{dif}} / \bar{Y}_c^{\text{obs}}$关于统计推断优先考虑 Fisher 精确$p$值方法：

1. 给定抽样次数K（通常为 1000 次）；

2. 对于轮次k，从 1 到 K 进行循环；

采用配对随机分组机制产生新的$(W_{1,A}^k, ..., W_{j,A}^k, ..., W_{J,A}^k)$，即对于每个配对$(A_j, B_j)$随机选择一个进入实验组，另一个为对照组，计算相对提升：

$$T^{\text{lift,k}} = \frac{\frac{1}{J}\sum_{j=1}^{J}\left(Y_{j,t}^k - Y_{j,c}^k\right)}{\frac{1}{J}\sum_{j=1}^{J} Y_{j,c}^k} = \frac{\frac{1}{J}\sum_{j=1}^{J}\left(W_{j,A}^k \cdot \left(Y_{j,A}^{\text{obs}} - Y_{j,B}^{\text{obs}}\right) + \left(1 - W_{j,A}^k\right) \cdot \left(Y_{j,B}^{\text{obs}} - Y_{j,A}^{\text{obs}}\right)\right)}{\frac{1}{J}\sum_{j=1}^{J}\left(\left(1 - W_{j,A}^k\right) \cdot Y_{j,A}^{\text{obs}} + W_{j,A}^k \cdot Y_{j,B}^{\text{obs}}\right)}$$

$$\text{or} \frac{\frac{1}{J}\sum_{j=1}^{J}\left(W_{j,A}^k \cdot Y_{j,A}^{\text{obs}} + \left(1 - W_{j,A}^k\right) \cdot Y_{j,B}^{\text{obs}}\right)}{\frac{1}{J}\sum_{j=1}^{J}\left(\left(1 - W_{j,A}^k\right) \cdot Y_{j,A}^{\text{obs}} + W_{j,A}^k \cdot Y_{j,B}^{\text{obs}}\right)} - 1$$

3. 近似计算p值$\hat{p} = \frac{1}{K}\sum_{k=1}^{K} 1_{\{T^{\text{lift},k} \geq T^{\text{lift,obs}}\}}$或$\hat{p} = \frac{1}{K}\sum_{k=1}^{K} 1_{\{|T^{\text{lift},k}| \geq |T^{\text{lift,obs}}|\}}$。

连续型指标相对提升之 Neyman 方法：

如果想要计算连续型指标相对提升的 MDE，我们在一定的假设前提 [4] 下考虑 Neyman 方法，推导了方差近似公式 $\mathrm{var}(\hat{\tau}^{\mathrm{dif}} / \bar{Y}_c^{\mathrm{obs}}) \approx \mathrm{var}(\hat{\tau}^{\mathrm{dif}})\left(\frac{1}{4}\cdot\frac{1}{\bar{Y}_c^2}+\frac{1}{4}\cdot\frac{\bar{Y}_t^2}{\bar{Y}_c^4}+\frac{1}{2}\cdot\frac{\bar{Y}_t}{\bar{Y}_c^3}\right)$，并通过 AA 模拟验证了基本可行。

（2）比率型指标评估

从业务理解出发往往需要考虑比率型指标，例如实验组 GMV 完成率定义为（实验组所有区域完成 GMV）/（实验组所有区域 GMV），而不是实验组每个区域 GMV 完成率的平均值。针对配对随机分组实验下比率型指标评估的理论研究几乎没有，在此探讨部分我们通过 AA 模拟来保证方法的科学性。

Fisher 方法

在强零假设 $H_0 : Y_i(0) = Y_i(1),\ \text{for all } i = 1,...,N$ 下，例如各个区域在实验状态和对照状态下 GMV 完成率相等的原假设下，基于 $J = N/2$ 个配对 (A_j, B_j) 的观测值。（（实验组 $X_{j,t}^{\mathrm{obs}}$，实验组 $Z_{j,t}^{\mathrm{obs}}$）（对照组 $X_{j,c}^{\mathrm{obs}}$，对照组 $Z_{j,c}^{\mathrm{obs}}$））例如 X 为区域完成 GMV，Z 为区域总 GMV，计算对应的策略绝对提升为：

$$T^{\mathrm{dif,obs}} = \frac{\sum_{j=1}^{J} X_{j,t}^{\mathrm{obs}}}{\sum_{j=1}^{J} Z_{j,t}^{\mathrm{obs}}} - \frac{\sum_{j=1}^{J} X_{j,c}^{\mathrm{obs}}}{\sum_{j=1}^{J} Z_{j,c}^{\mathrm{obs}}} = \frac{\sum_{j=1}^{J}\left(W_{j,A}\cdot X_{j,A}^{\mathrm{obs}} + \left(1-W_{j,A}\right)\cdot X_{j,B}^{\mathrm{obs}}\right)}{\sum_{j=1}^{J}\left(W_{j,A}\cdot Z_{j,A}^{\mathrm{obs}} + \left(1-W_{j,A}\right)\cdot Z_{j,B}^{\mathrm{obs}}\right)} - \frac{\sum_{j=1}^{J}\left(\left(1-W_{j,A}\right)\cdot X_{j,A}^{\mathrm{obs}} + W_{j,A}\cdot X_{j,B}^{\mathrm{obs}}\right)}{\sum_{j=1}^{J}\left(\left(1-W_{j,A}\right)\cdot Z_{j,A}^{\mathrm{obs}} + W_{j,A}\cdot Z_{j,B}^{\mathrm{obs}}\right)}$$

其中 $W_{j,A}$ 代表第 j 个对中个体 A_j 是否落在实验组，$Y_{j,A}^{\mathrm{obs}}$ 为个体 A_j 的指标观测值。上述公式代表的实际上就是【实验组所有区域完成 GMV)/(实验组所有区域 GMV)- 对照组所有区域完成 GMV)/(对照组所有区域 GMV)】。进一步可通过 Fisher 精确 p 值算法计算 p 值。

1. 给定抽样次数 K（通常为1000次）；

2. 对于轮次 k，从1到 K 进行循环；

采用配对随机分组机制产生新的 $(W_{1,A}^k,...,W_{j,A}^k,...,W_{J,A}^k)$，也即对于每个配对 (A_j,B_j) 随机选择一个进入实验组，另一个为对照组，计算统计量：

$$T^{\text{dif,k}}=\frac{\sum_{j=1}^{J}\left(W_{j,A}^k\cdot X_{j,A}^{\text{obs}}+\left(1-W_{j,A}^k\right)\cdot X_{j,B}^{\text{obs}}\right)}{\sum_{j=1}^{J}\left(W_{j,A}^k\cdot Z_{j,A}^{\text{obs}}+\left(1-W_{j,A}^k\right)\cdot Z_{j,B}^{\text{obs}}\right)}-\frac{\sum_{j=1}^{J}\left(\left(1-W_{j,A}^k\right)\cdot X_{j,A}^{\text{obs}}+W_{j,A}^k\cdot X_{j,B}^{\text{obs}}\right)}{\sum_{j=1}^{J}\left(\left(1-W_{j,A}^k\right)\cdot Z_{j,A}^{\text{obs}}+W_{j,A}^k\cdot Z_{j,B}^{\text{obs}}\right)}$$

3. 近似计算 p 值 $\hat{p}=\frac{1}{K}\sum_{k=1}^{K}1_{\{T^{\text{dif},k}\geq T^{\text{dif,obs}}\}}$ 或 $\hat{p}=\frac{1}{K}\sum_{k=1}^{K}1_{\{|T^{\text{dif},k}|\geq|T^{\text{dif,obs}}|\}}$。

Neyman 方法

对于连续型指标，配对实验场景下通常考虑配对 t 检验（即每一对计算 diff 后基于各个对的 diff 计算方差），但这在比率型指标下显然是不可行的，因此我们探讨了比率型指标的 Neyman 双样本 t 检验，在部分假设 [5] 下推导方差近似公式：

$$\text{Var}\left(\frac{\bar{Y}_t}{\bar{Z}_t}-\frac{\bar{Y}_c}{\bar{Z}_c}\right)\approx\left(\frac{1}{4\bar{Z}_t^2}+\frac{1}{4\bar{Z}_c^2}+\frac{2}{4\bar{Z}_t\bar{Z}_c}\right)\cdot\frac{S_\tau^2}{n}+\left(\frac{\bar{Y}_t^2}{4\bar{Z}_t^4}+\frac{\bar{Y}_c^2}{4\bar{Z}_c^4}+\frac{2\bar{Y}_t\bar{Y}_c}{4\bar{Z}_t^2\bar{Z}_c^2}\right)\cdot\frac{S_\nu^2}{n}-\left(\frac{2\bar{Y}_t}{4\bar{Z}_t^3}+\frac{2\bar{Y}_c}{4\bar{Z}_c^3}+\frac{2\bar{Y}_c}{4\bar{Z}_t\bar{Z}_c^2}+\frac{2\bar{Y}_t}{4\bar{Z}_t^2\bar{Z}_c}\right)\cdot\frac{S_{\tau,\nu}}{n}$$

其中 $S_\tau^2=\frac{\sum_{j=1}^{n}(\hat{\tau}_j-\bar{\tau})^2}{n-1}$，$S_\nu^2=\frac{\sum_{j=1}^{n}(\hat{v}_j-\bar{v})^2}{n-1}$，$S_{\tau,\nu}=\frac{\sum_{j=1}^{n}(\hat{\tau}_j-\bar{\tau})(\hat{v}_j-\bar{v})}{n-1}$，$\bar{\tau}=\frac{\sum_{j=1}^{n}\hat{\tau}_j}{n}$，$\bar{v}=\frac{\sum_{j=1}^{n}\hat{v}_j}{n}$，$n$ 为配对数目，$\hat{\tau}_j$ 为第 j 个对实验组 $Y-$ 对照组 Y，$\hat{v}_j$ 为第 j 个对实验组 $Z-$ 对照组 Z，并通过 AA 模拟验证基本可行。

在美团履约侧的实验场景中，许多策略（如调度优化等）的作用单元为区域粒度等较大的地理单元，且策略的触发因素受到外部环境的影响。这意味着无法保证触发策略的实验组、对照组区域的外部环境相似，这使实验设计面临很多独特的挑战，特别是

在确保实验组和对照组的同质性方面。

1. **实验粒度限制：**由于一些策略的最小作用单元是区域，这限制了实验无法在更细粒度上（如订单）分流。而城市下的区域数量较少且策略的触发因素受到外部环境的影响时，随机分组难以保证触发策略的实验组、对照组同质。
2. **其他影响因素：**区域内的交通状况、订单密度等因素也可能在实验期间发生变化，进一步增加了实验组和对照组之间的异质性。

3.3.3 协变量自适应分组

在美团履约侧的实验场景下，**调度等场景下样本量稀少与地域差异明显的现状使得随机对照实验下难以保证分组的同质性以及很难有效地检测出实验提升效果**。受自身业务形态和空间维度限制，调度等算法的最小作用单元为区域，且通常情况下以区域组作为策略施加的单位，受限于策略的最小作用单元，在实验设计上只能考虑区域或区域组维度的分流，这就导致参与实验的样本量较少，在这样的中小样本场景下（样本量几十到几百），协变量自适应分组可以通过减小组间指标分布不平衡性，使得分组更同质，从而有效提升点估计的准确度，在保证同质性和提升统计功效等方面普遍优于经典的随机对照实验分组。

协变量自适应分组（Covariate-Adaptive Randomization）是一种在实验设计中用来进一步保证同质性的实验方式，旨在通过减小实验组和对照组之间重要协变量分布的不平衡性，使得分组更同质，来提高实验结果的准确性和可靠性。在随机对照试验中，协变量自适应分组通过序贯分配的方式，使得各组之间在关键协变量上的分布更加相似，减少了协变量对试验结果的混杂影响。下面是一些协变量自适应分组的应用经验。

1. 优势

协变量自适应分组的优点在于它能：

- **进一步保证同质性：**通过平衡重要协变量，进一步保证了实验组对照组的同质

性，使得分组更同质。

- **提高统计功效：**平衡协变量后，由于减少了组间差异，协变量自适应分组通常能够提高统计分析的功效。
- **适用于小样本：**在小样本场景下（如几十到几百个样本），协变量自适应分组能够显著提升同质性和估计精度。

2. 局限性

- **计算成本增加：**相比于经典随机对照实验的分组方式，协变量自适应分组需要更多的计算时间。
- **需要先验知识：**需要对重要协变量有较好的理解和认知能力，以选择合适的协变量进行平衡。

3. 适用场景

- **个体差异显著的场景：**在实验对象之间存在显著个体差异的情况下（如区域间地理差异等），协变量自适应分组可以有效平衡这些差异。通过在实验组和对照组之间平衡重要协变量，确保个体差异不会显著影响实验结果的解读。
- **样本量有限的场景：**在样本量较小的研究中，协变量自适应分组有助于提高实验的统计功效。小样本量可能导致组间协变量的不平衡，协变量自适应分组通过动态调整分组策略，确保在小样本条件下，实验组和对照组的协变量分布尽可能相似，从而提高结果的可靠性。

平衡协变量对于实施可靠 A/B 实验是极为重要的，可以有效地降低处理效应估计的偏差，而协变量自适应设计则是为实现这一目标最为常用的方法。在协变量自适应随机化过程中，往往以每一步最小化某特定的不平衡测度为目标，序贯地（Sequential）将实验个体逐个（或逐对）分为试验组和对照组，其中不平衡测度的选择包括但不限于协变量均值（Covariate Means）、马氏距离（Mahalanobis Distance）、离散化加权处理、核方法等。以下是其基本原理和实现过程。

1. 不平衡测度的刻画标准

以“区域”粒度随机对照实验为例，协变量自适应分组的目的是尽量保障划分后实验组、对照组特征足够相似，此时如何刻画一个区域分配到实验组或对照组时的组间差异至关重要。令 T_i 为第 i 个区域的分配示性函数，即第 i 个区域在实验组则 $T_i=1$，第 i 个区域在对照组则 $T_i=0$。令 $X_i=(X_{i1},\ldots,X_{ip})^T$ 为第 i 个区域的协变量。采用 Imb (Imbalance measure) 作为刻画上述组间差异的标尺。根据 Ma et al. (2022), 可将 Imb 定义为：

$$Imb=\left\|\sum_{i=1}^{n}(2T_i-1)\phi(X_i)\right\|^2$$

其中 $\phi(X)$ 为纳入 Imb 考虑的协变量特征函数。在履约场景下我们考虑用马氏距离（Mahalanobis Distance）作为衡量组间差异的工具，它可以看作是欧氏距离的一种修正，修正了欧式距离中各个维度尺度不一致且相关的问题。其它的常用度量包括协变量均值、离散型协变量等方式。假定 $\mathrm{cov}(X_i)$ 是已知的，其特征分解为 $\mathrm{cov}(X_i)^{-1}=VDV^T$，定义 $\phi(X_i)=D^{1/2}V^TX_i$，其中 V 是特征向量矩阵，D 是特征值对角矩阵，则不平衡度 Imb 可以表示为：

$$Imb=\left\|\sum_{i=1}^{n}(2T_i-1)D^{1/2}V^TX_i\right\|^2=\left(\sum_{i:T_i=1}X_i-\sum_{i:T_i=0}X_i\right)^T\mathrm{cov}(X_i)^{-1}\left(\sum_{i:T_i=1}X_i-\sum_{i:T_i=0}X_i\right)$$

在这里之所以考虑马氏距离 (Mahalanobis Distance) 来定义不平衡度 Imb，是基于以下几点考量：

- **形式简洁、计算成本低：**马氏距离的计算相对简单，且符合统计学分析的直觉和业界常用评估指标的习惯。
- **不受数据线性变换的影响：**协变量之间的量纲差异可能会对不平衡性的衡量过程带来麻烦。马氏距离不受数据线性变换的影响，省略了数据预处理的必要性，使得计算更加简便和科学。
- **优良的统计学性质：**马氏距离具有减少估计处理效应方差的优良性质。在处理效应的估计中，马氏距离能够提供最优的渐进方差，使得实验结果更加可靠。

2. 协变量自适应分配方式

在进行协变量自适应分组时，学业界主要以完全序贯分配和配对序贯分配作为协变量自适应设计的分配方式。协变量自适应分配的主要思想是，逐个或者逐对分配实验单元，其通过倾向于使不平衡测度差异最小来判断将实验单元分在实验组还是对照组，在分组过程中动态调整以确保实验组和对照组在关键协变量上的平衡。

(1) 完全序贯分配

Step1: 初始化，将第一个个体以 1/2 的概率分配到实验组；

repeat

Step2: 在前 $(n-1)$ 个个体的分配完成之后，并记第 n 个个体落在 $(k_1^*,\ldots,k_I^*)$ 层；

Step3: 若第 n 个个体分配到实验组，计算此时的 Imb，记为 $Imb_n^{(1)}$；若第 n 个个体分配到对照组，计算此时的 Imb，记为 $Imb_n^{(2)}$。

Step4: 在已知前 $(n-1)$ 个个体的分配情况下，第 n 个个体的分组将基于以下准则：

$$\mathbb{P}(T_n=1\,|\,\mathbf{Z}_n,\mathbf{T}_{n-1})=\begin{cases}q, & Imb_n^{(1)}<Imb_n^{(2)};\\ 1-q, & Imb_n^{(1)}>Imb_n^{(2)};\\ 0.5, & Imb_n^{(1)}=Imb_n^{(2)}.\end{cases}$$

其中 $0.5<q<1$。

until 全部个体的分组完成。

(2) 配对序贯分配

Step1: 初始化：将全体 n 个待分配个体随机排序为 $x_1,\ldots,x_n$；

Step2: 将序列中第一个和第二个个体分别分配到实验组和对照组中，即 $T_1=1,T_2=0$；

repeat

Step3：

- 在前$2i$个个体的分组完成后，对于第$(2i+1)$和第$(2i+2)$个个体而言；
- 若第$(2i+1)$个个体被分配到实验组，第$(2i+2)$个个体被分配到对照组，在此情况下，定义并计算“潜在”不平衡测度$Imb_1(2i+2)$；
- 若第$(2i+1)$个个体被分配到对照组，第$(2i+2)$个个体被分配到实验组，定义并计算“潜在”不平衡测度$Imb_2(2i+2)$。

Step4：第$(2i+1)$和第$(2i+2)$个个体的分组将基于以下准则：

$$\mathbb{P}(T_{2i+1}=1\,|\,x_{2i},\ldots,x_1,T_{2i},\ldots,T_1)=\begin{cases}q, & Imb_1(2i+2)<Imb_2(2i+2);\\ 1-q, & Imb_1(2i+2)>Imb_2(2i+2);\\ 0.5, & Imb_1(2i+2)=Imb_2(2i+2).\end{cases}$$

until 全部n个个体的分组全部完成。若n为奇数，那么最后一个个体将以等概率分配到两组中。

概率q通常选为 0.85（也可 0.75、0.9 等等）。两种分配流程各有优劣，完全序贯分配的随机性更强，但很难保证分配到试验组和对照组的样本量相同，而配对序贯分配则可以用部分随机性换取试验组和对照组的样本量相同的结果。

协变量自适应分组通过协变量回归调整降低方差，在这个分组机制下的评估中，方差计算同经典随机对照实验中的 CUPED 方差削减技术，具体可参考经典随机对照实验部分，这里不再赘述。

下面我们介绍一个使用协变量自适应分组得到更为同质区域分组的案例，我们比较了某业务场景中完全随机分组和协变量自适应分组的多个指标情况。其中，CR 代表完全随机分组，CAR-c，CAR-d，CAR-m 分别代表使用平衡均值，平衡离散型协变量，平衡马氏距离的协变量自适应方法。表中结果显示，在相同的检验显著性水平为 0.05 下，协变量自适应设计下的各协变量的拒绝原假设概率更低，即更不易拒绝同质性检验，且能生成更为同质的分组，降低了分组后的组间差异均值和方差波动。

		CR	CAR-c	CAR-d	CAR-m
指标1	均值（标准差）	2.72（546.40）	-3.12（268.03）	0.26（393.71）	-1.90（350.11）
	拒绝原假设概率	5.34%	0.10%	0.42%	0.24%
指标2	均值（标准差）	2.28（161.92）	0.61（97.37）	1.36（118.07）	0.03（104.29）
	拒绝原假设概率	5.30%	0.18%	0.46%	0.38%
指标3	均值（标准差）	11.92（1443.21）	-4.60（621.60）	2.64（1171.11）	-6.83（933.58）
	拒绝原假设概率	5.34%	0.06%	1.82%	0.38%

3.4 解决溢出效应难题的实验方式

在 AB 实验中，一个很重要的假设是 SUTVA (Stable Unit Treatment Value Assumption，个体处理稳定性假设)，即实验中每个实验参与单元的行为是相互独立的，然而实践中由于实验单元间的直接关联或者间接关联，参与 AB 实验的实验组与对照组之间可能并不独立，我们通常称这种实验组、对照组间的相互干扰影响为溢出效应。

溢出效应的存在往往会引发实验效果的估计偏差，进而损失实验结论的可信度。例如通信工具 Skype 电话测试提升通话质量策略时，由于实验组呼叫可以拨给实验组或对照组的用户，从而对照组用户使用 Skype 电话频率也会增加，因此实验组和对照组之间的差值会被低估。溢出效应难题仍是目前学界与业界的重点研究领域，现有的实验设计与解决方案主要有时间片轮转实验、聚类随机试验、双边实验以及随机饱和实验等。其中时间片轮转实验在美团的实验场景下已经落地应用，会在第四章中详细介绍，这里我们将分别重点介绍在美团履约侧实验场景下，如何通过区域溢入溢出效应模型以及随机饱和实验来解决溢出效应的问题。

3.4.1 区域溢入溢出效应模型

在美团履约的实际业务背景下，例如调度策略，由于混合调度下在分别在实验组和对照组的相邻区域可以召回相同骑手，导致实验组与对照组之间共享骑手运力资源，使得实验组和对照组区域单元的独立性假设难以满足。如何在总体效应中有效识别溢出

效应，并将最为关心的直接效应分离出来，是互联网企业在双边（多边）市场背景下研究网络因果推断问题的关键技术难点。为此，我们通过与中国人民大学进行校企合作，引入了区域溢入溢出效应模型，在部分场景下解决了溢出效应问题。

在区域层面，我们建模刻画了目标变量在不同区域间流动规模与方向，以及量化变动幅度，引入了“溢出权重”、“溢出强度”、“溢入权重”与“溢入强度”四大指标。其中“溢出权重”主要基于对目标变量在不同区域间流动规模与方向的考量，通过计算各区域间目标变量的流动量和方向，构建出反映流动规模与方向的综合指标，直观展示目标变量在区域间的流动情况，为决策提供参考。而“溢出强度”侧重于衡量目标变量在某一区域内的变动幅度，通过计算某一时期内目标变量在该区域内的变化量，结合“溢出权重”，得出反映变动幅度的量化指标，帮助研究者与决策者快速识别变动幅度较大的区域，进行针对性分析与应对。同样，“溢入权重”和“溢入强度”则反映与“溢出权重”和“溢出强度”相反方向的变动幅度。

区域溢入强度和区域溢出强度可以理解为周围不同组别区域对自身的影响强度，最简单的定义是周围不同组别的区域个数比例（比如某单元自身为实验组，周围为对照组的比例越大，认为影响越强）。但是这种定义没有考虑到周围区域的大小及单量规模，一般认为大区域对自身的影响会比小区域对自身的影响强。因此，我们结合溢入（溢出）运单量来设计溢入（溢出）强度的定义。

描述	基本含义
溢出权重 W_{ij}^{OUT}： 用以计算“溢出强度”e_i^{OUT}的中间指标，它衡量所有从i区域溢出的目标单元中溢出至j区域的比例。既衡量了跨区域流动单元的总体规模，也刻画出了溢出单元的流出区域和流入区域。需要说明的是，受统计的“溢出”情况定义为跨区域组的区域间溢出，在同一区域组的区域间溢出因所受处理相同而不被计算。	$W_{ij}^{OUT}=\frac{\text{区域}i\text{溢到区域}j\text{的单位数}}{\sum_{k=1}^{K}\text{区域}i\text{溢出到区域}k\text{的单位数}}$， 其中区域i溢出到区域k的单位数可以理解为：进行中的区域i的订单对后续接起的区域k的订单的影响。
溢出强度 e_i^{OUT}： 综合考虑了“溢出权重”w_{ij}^{OUT}和标记区域i为试验组或控制组的二值变量T_i。通过计算溢出强度e_i^{OUT}，可以精确地观察某一目标区域的单元流动特征。	$e_{i(t)}^{OUT}=\sum_{j=1}^{K}W_{ij}^{OUT}\cdot I_{\{T_i=t,(1-T_j)\cdot t=1\}}+\sum_{j=1}^{K}W_{ij}^{OUT}\cdot I_{\{T_i=t,T_j\cdot(1-t)=1\}}$， 当区域i属于试验组时，$e_{i(1)}^{OUT}$可以衡量出试验区域i对其他所有跨组对照区域的影响；而当区域i属于对照组时，$e_{i(0)}^{OUT}$则衡量的是对照区域i被试验区域影响的程度。
溢入权重 W_{ij}^{IN}： 用以计算“溢入强度”e_i^{IN}的中间指标，它衡量所有溢入到i区域的目标单元中从j区域溢入的比例。既衡量了跨区域流动单元的总体规模，也刻画出了溢入单元的流出区域和流入区域。需要说明的是，受统计的“溢入”情况定义为跨区域组的区域间溢入，在同一区域组的区域间溢入因所受处理相同而不被计算。	$W_{ij}^{IN}=\frac{\text{区域}j\text{溢到区域}i\text{的单位数}}{\sum_{k=1}^{K}\text{区域}k\text{溢入到区域}i\text{的单位数}}$， 其中区域k溢入到区域i的单位数可以理解为：进行中的区域k的订单对后续接起的区域i的订单的影响。
溢入强度 e_i^{IN}： 综合考虑了“溢入权重”w_{ij}^{IN}和标记区域i为试验组或控制组的二值变量T_i。通过计算溢入强度e_i^{IN}，可以精确地观察某一目标区域的单元流动特征。	$e_{i(t)}^{IN}=\sum_{j=1}^{K}W_{ij}^{IN}\cdot I_{\{T_i=t,(1-T_j)\cdot t=1\}}+\sum_{j=1}^{K}W_{ij}^{IN}\cdot I_{\{T_i=t,T_j\cdot(1-t)=1\}}$， 当区域i属于试验组时，$e_{i(1)}^{IN}$可以衡量试验区域i对其他对照区域要素的影响程度；而当区域i属于对照组时，$e_{i(0)}^{IN}$则衡量的是对照区域i受其他所有跨组试验区域的影响。

溢入强度与溢出强度的现实意义：

参照下图，在某次实验中，对实验组区域施加正向激励的策略（如升级调度策略使实验组区域单和骑手更匹配，实验组单被骑手更好更快接起），则在全城运力和总订单量无显著变化的情况下，实验组区域存量订单快速被消耗，实验组区域骑手倾向于跨区接起地理层面相邻且没有正向激励策略的对照组的订单。

此时，除了策略对实验组区域的直接效应外，有对照组向实验组区域的订单溢入影响，这种影响的量化指标为溢入强度$e_{i(1)}^{IN}$，用于度量衡量实验组区域影响对照组区域订单的强度大小；同样地，在正向激励的策略动作下，若某对照组区域周围多是实验组区域，则该对照区域的运力倾向流入实验区域（因为骑手倾向去实验组区域送单）。为了衡量出这种对照区域的受影响程度，我们可以使用溢出强度$e_{i(0)}^{OUT}$。类似的，可解释$e_{i(1)}^{OUT}$与$e_{i(0)}^{IN}$的现实意义。

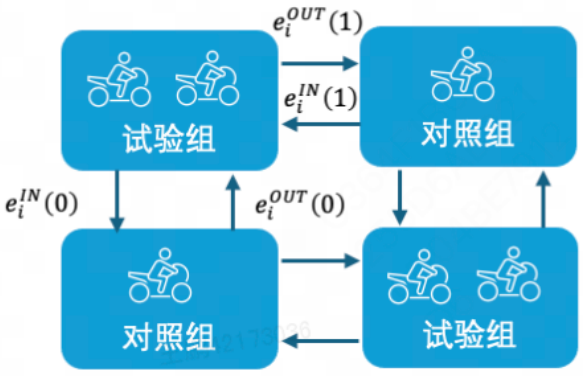


图 3-5：溢出效应建模示意图

在这四个指标的基础上，以比率型指标为例，我们可以进行溢入溢出效应的建模，模型如下：

$$Y_i = T_i(\mu_{Y1} + e_{i(1)}^{IN}\gamma_{Y11} - e_{i(1)}^{OUT}\gamma_{Y12}) + (1-T_i)(\mu_{Y0} + e_{i(0)}^{IN}\gamma_{Y01} - e_{i(0)}^{OUT}\gamma_{Y02})$$

$$Z_i = T_i(\mu_{Z1} + e_{i(1)}^{IN}\gamma_{Z11} - e_{i(1)}^{OUT}\gamma_{Z12}) + (1-T_i)(\mu_{Z0} + e_{i(0)}^{IN}\gamma_{Z01} - e_{i(0)}^{OUT}\gamma_{Z02})$$

$e_{i(t)}^{IN}$和$e_{i(t)}^{OUT}$分别刻画了溢入和溢出的强度，相应的，$\gamma_{Yt1},\gamma_{Yt2},\gamma_{Zt1}$和$\gamma_{Zt2}$则代表着相应溢出溢入强度对比率型指标分子$Y$和分母$Z$的影响力度。这种建模方式，能够将实验组和对照组的溢出效应有效地从Y和Z中剥离出，此时的μ_{Yt}和μ_{Zt}是实验组或对照组中Y和Z真实的均值。因此，比率型指标的处理效应（Treatment Effect）定义为$\frac{\mu_{Y1}}{\mu_{Z1}} - \frac{\mu_{Y0}}{\mu_{Z0}}$，估计量是$\frac{\hat{\mu}_{Y1}}{\hat{\mu}_{Z1}} - \frac{\hat{\mu}_{Y0}}{\hat{\mu}_{Z0}}$，然后可以通过回归系数的标准差以及 Delta 方法估计出该比率型指标的标准差，从而得到其置信区间和p值。

3.4.2 随机饱和实验

随机饱和实验（Randomized Saturation Design）源于两阶段随机实验（Two-staged Randomized Experiment）的理念。与传统的随机对照实验中仅有一个固定的实验组与对照组比例不同，随机饱和实验通过将样本划分为多个簇，在每个簇中设置不同的实验组与对照组比例。理论上，实验组比例较高的簇对对照组的溢出效应更强，从而可以通过分析不同簇内实验组、对照组的表现，检测出真实的实验效应和溢

出效应。

随机饱和实验设计：

Step1：将实验对象预先划分为多个簇（Cluster），并设置一个饱和度集合，为每组按照某个概率分布（如离散型均匀分布）随机分配一个饱和度，饱和度即簇内实验单元占比；

Step2：对于每个簇，根据被分配到的饱和度，按此比例随机分配簇内的个体划分为实验组和对照组。

假设我们将n个实验单元划分成C个簇，每个簇$c=1,\cdots,C$根据某个规则（概率分布）f分配到一个饱和度$\pi_c \in \Pi$，如果某个簇的饱和度为 0，称这个簇为纯对照簇（Pure Control Cluster）。

一个随机饱和实验ω由饱和度集合Π和概率分布f决定。对于每个单元，我们有如下的定义：

符号	描述
Y_{ic}	簇c内的第i个个体的潜在效应
X_{ic}	簇c内的第i个个体的协变量
T_{ic}	个体(c,i)的实验状态，若$T_{ic}=1$代表此个体接受策略，若$T_{ic}=0$代表此个体没接受策略
S_{ic}	个体(c,i)是否是簇内对照状态 (within-cluster control)，$S_{ic}=1(T_{ic}=0,\pi_c>0)$，即非纯对照簇内的对照单元
ϵ_{ic}	噪声项

根据这些定义易得$\mathrm{P}(T_{ic}=1 \mid \pi_c=\pi)=\pi$。在随机饱和实验中，个体可能处于三种状态：

1. **纯对照（Pure Control）：**$\pi_c=0$的簇中全部的个体均处于纯对照状态，这些个体满足$T_{ic}=0, S_{ic}=0$；

2. **簇内对照（Within-Cluster Control）：** 在有实验状态的簇中的对照个体，即$T_{ic}=0, S_{ic}=1$；
3. **实验状态：** 接受了策略的个体，即$T_{ic}=1, S_{ic}=0$。

并且我们可以计算出个体(i,c)处于实验状态、纯对照状态、簇内对照状态的边际概率，分别为：

$$\text{Treatment: } \mathrm{P}(T_{ic}=1)=\sum_{\pi\in\Pi}\mathrm{P}(T_{ic}=1\mid \pi_c=\pi)\mathrm{P}(\pi_c=\pi)=\sum_{\pi\in\Pi}\pi f(\pi)=\mathrm{E}(\pi),$$

$$\text{Pure Control: } \mathrm{P}(S_{ic}=0,T_{ic}=0)=\sum_{\pi\in\Pi}\mathrm{P}(S_{ic}=0,T_{ic}=0\mid \pi_c=\pi)\mathrm{P}(\pi_c=\pi)=\mathrm{P}(\pi_c=0)=f(0),$$

$$\text{Within-cluster Control: } \mathrm{P}(S_{ic}=1)=1-\mathrm{P}(T_{ic}=1)-\mathrm{P}(S_{ic}=0,T_{ic}=0)=1-\mathrm{E}(\pi)-f(0).$$

对于随机饱和实验，我们需要以下一些基本假设。

假设 1： 随机饱和实验放宽了 SUTVA 的限制，允许簇内各个单元有相互影响，但是要求簇间没有相互干涉，即溢出效应仅存在于簇内而不流动于簇间；

这个假设保证了我们可以基于不同簇间的不同饱和度检测出真实的实验效应和溢出效应。在实际应用中，我们可以通过地理单元对区域进行或者时间上对天进行簇的隔离来实现簇间没有相互干涉。

假设 2： 在饱和度不为 0 的簇中，个体(i,c)的效应仅取决于其是否是实验单元和所在簇分配到的饱和度π，而不取决于簇内其他个体实验状态的排列情况。

这一简化了分析的难度，并允许我们在不清楚每一个簇背后的网络结构的情况下进行统计推断。这样，我们可以将Y_{ic}表示为：

$$Y_{ic}=g(T_{ic},\pi_c;X_{ic})+\epsilon_{ic}$$

假设 3： 考虑用随机效应结构来建模不同簇的噪声项$\epsilon_{ic}=v_c+w_{ic}$，簇c内的个体拥有相同部分$v_c\sim(0,\tau^2)$，个体部分为$w_{ic}\sim(0,\sigma^2)$。

首先我们定义几种因果效应与溢出效应：

描述	定义
因果效应-Intention to Treat (ITT)： 饱和度为π的簇中实验组个体的预期结果与纯对照簇中个体的预期结果之间的差异。	$\mathrm{ITT}(\pi)=\mathrm{E}(Y_{ic}\|T_{ic}=1,\pi_c=\pi)-\mathrm{E}(Y_{ic}\|T_{ic}=0,\pi_c=0)$
因果效应-Treatment on the Uniquely Treated (TUT)： 饱和度为$\frac{1}{n}$的簇（只有一个实验个体）中唯一的实验组个体的预期结果与纯对照簇中个体的预期结果之间的差异。	$\mathrm{TUT}=\mathrm{E}\left(Y_{ic}\mid T_{ic}=1,\pi_c=\frac{1}{n}\right)-\mathrm{E}\left(Y_{ic}\mid T_{ic}=0,\pi_c=0\right)=\mathrm{ITT}\left(\frac{1}{n}\right)$
溢出效应-Spillover on Non-Treated (SNT)： 饱和度为π的簇中对照组个体的预期结果与纯对照簇中个体的预期结果之间的差异。	$\mathrm{SNT}(\pi)=\mathrm{E}(Y_{ic}\|T_{ic}=0,\pi_c=\pi)-\mathrm{E}(Y_{ic}\|T_{ic}=0,\pi_c=0)$
溢出效应-Spillover on the Treated (ST)： 饱和度为π的簇中实验组个体的预期结果与饱和度为$\frac{1}{n}$的簇中唯一的实验组个体的预期结果之间的差异。	$\mathrm{ST}(\pi)=\mathrm{E}(Y_{ic}\|T_{ic}=1,\pi_c=\pi)-\mathrm{E}(Y_{ic}\|T_{ic}=1,\pi_c=\frac{1}{n})$
因果效应-Total Causal Effect (TCE)： 饱和度为π的簇中所有个体的预期结果与纯对照簇中所有个体的预期结果之间的差异。	$\mathrm{TCE}(\pi)=\mathrm{E}(Y_{ic}\|\pi_c=\pi)-\mathrm{E}(Y_{ic}\|\pi_c=0)=\pi\mathrm{ITT}(\pi)+(1-\pi)\mathrm{SNT}(\pi)$

根据上述定义易得$\mathrm{ITT}(\pi)=\mathrm{TUT}+\mathrm{ST}(\pi)$。

上述效应都是对于某个特定的饱和度定义的，我们也可以求出整个随机饱和设计ω的平均效应。以 ITT 为例，随机饱和设计ω的$\overline{\mathrm{ITT}}_\omega$的定义为：

$$\begin{aligned}\overline{\mathrm{ITT}}_\omega &= \sum_{\Pi\setminus\{0\}} \mathrm{E}(Y_{ic}\mid T_{ic}=1,\pi_c=\pi)\frac{f(\pi)}{1-f(0)}-\mathrm{E}(Y_{ic}\mid T_{ic}=0,\pi_c=0)\\ &= \sum_{\Pi\setminus\{0\}} \mathrm{ITT}(\pi)\frac{f(\pi)}{1-f(0)}\end{aligned}$$

其他如$\overline{\mathrm{SNT}}_\omega,\overline{\mathrm{TCE}}_\omega,\overline{\mathrm{ST}}_\omega$的定义同理。因此，如果存在一个$\pi$满足$\mathrm{SNT}(\pi)\neq 0$或$\mathrm{ST}(\pi)\neq 0$时，则认为存在溢出效应。一个常用的溢出效应的检测手段是检测$\overline{\mathrm{SNT}}\neq 0$或$\overline{\mathrm{ST}}\neq 0$。

对于随机饱和实验的评估，我们常常有如下的建模方式。

简单线性回归模型建模

$$Y_{ic}=\delta_0+\beta_1 T_{ic}+\beta_2 S_{ic}+\phi X_{ic}+\epsilon_{ic}$$

给定随机饱和实验设计ω，该模型可识别 ITT 与 SNT：$\widehat{\overline{\mathrm{ITT}}}_\omega=\hat{\beta}_1,\widehat{\overline{\mathrm{SNT}}}_\omega=\hat{\beta}_2$，可以通过检验$\beta_1=0$来反映实验是否有效果，通过检验$\beta_2=0$来检测是否存在溢出效应。

仿射模型（Affine Model）建模

$$Y_{ic} = \delta_0 + \delta_1 T_{ic} + \delta_2 S_{ic} + \delta_3 (T_{ic} \times \pi_c) + \delta_4 (S_{ic} \times \pi_c) + \phi X_{ic} + \epsilon_{ic}$$

给定随机饱和实验设计 ω：

1. 该模型可识别 TUT：$\widehat{\mathrm{TUT}} = \hat{\delta}_1$，可以通过检验$\delta_1 = 0$来反映实验是否有效果；
2. 由$\mathrm{ITT}(\pi) = \mathrm{TUT} + \mathrm{ST}(\pi)$，有$\frac{\mathrm{dST}(\pi)}{\mathrm{d}\pi} = \frac{\mathrm{dITT}(\pi)}{\mathrm{d}\pi} = \hat{\delta}_3, \frac{\mathrm{dSNT}(\pi)}{\mathrm{d}\pi} = \hat{\delta}_4$，可以通过检验$\delta_3 = \delta_4 = 0$来判断实验是否存在溢出效应（因为溢出效应是与饱和度有关的，如果$\delta_3 = \delta_4 = 0$，说明溢出效应不随饱和度发生改变，也就不存在溢出效应了）；
3. $\hat{\delta}_2$ 为饱和度为 0 时的溢出效应 SNT，根据定义，饱和度为 0 时应该不存在溢出效应，故可用 $\delta_2 = 0$ 来验证溢出效应的线性关系是否合理（如果不合理，可通过添加二次项等修正为非线性关系等）。
4. 在实验中，我们更关注策略带来的真实效应，即$\mathrm{E}(Y_{ic} \mid \pi_c = 1) - \mathrm{E}(Y_{ic} \mid \pi_c = 0)$，可以理解为某个纯对照组在被全部施加策略后的平均效应的增量。在实验中我们设置了纯对照组的簇，而全部施加策略的纯实验组簇的效应可以通过仿射模型进行预测：当时$\pi_c = 1$，这个簇里不存在对照个体，只有实验个体，也就是$S_{ic} = 0, T_{ic} = 1, \forall i$；纯对照组满足$\pi_c = 0$时，这个簇里的每个个体均满足$S_{ic} = 0, T_{ic} = 0, \forall i$。代入到仿射模型中并做差，易得$\delta_1 + \delta_3$，可以用这个预测值来估计策略的真实效应。

非参数模型建模

$$Y_{ic} = \beta_0 + \sum_{\Pi \setminus \{0\}} \beta_{1\pi} T_{ic} * 1_{\{\pi_c = \pi\}} + \sum_{\Pi \setminus \{0\}} \beta_{2\pi} S_{ic} * 1_{\{\pi_c = \pi\}} + \phi X_{ic} + \epsilon_{ic}$$

1. **该模型可识别 ITT、SNT、TCE：**$\widehat{\mathrm{ITT}}(\pi) = \hat{\beta}_{1\pi}, \widehat{\mathrm{SNT}}(\pi) = \hat{\beta}_{2\pi}$ and $\widehat{\mathrm{TCE}}(\pi) = \pi\hat{\beta}_{1\pi} + (1-\pi)\hat{\beta}_{2\pi}$ for each $\pi \in \Pi \setminus \{0\}$；
2. **考虑假设检验：**$\beta_{1\pi_j} = \beta_{1\pi_k}$可反映是否对于非纯对照簇中的实验组个体存在溢出效应，因为由定义$\beta_{1\pi_j} - \beta_{1\pi_k} = \mathrm{ITT}(\pi_j) - \mathrm{ITT}(\pi_k) = \mathrm{ST}(\pi_j) - \mathrm{ST}(\pi_k)$，该

假设检验实际检测溢出效应 ST 是否会随着饱和度π变动；

3. **考虑假设检验：**$\beta_{2\pi_j} = \beta_{2\pi_k}$可反映是否对于非纯对照簇中的对照组个体存在溢出效应，因为由定义$\beta_{1\pi_j} - \beta_{1\pi_k} = \mathrm{SNT}(\pi_j) - \mathrm{SNT}(\pi_k)$，该假设检验实际检测为溢出效应 SNT 是否会随着饱和度π变动。

3.5 拓展与展望

在随机对照实验的业务应用中，**触发式分析**具有重要作用。在某些特定业务场景中，实验组可能并未全部受到实验干预，导致直接比较实验组和对照组的策略效果时，结果可能被稀释，从而难以获得显著结论。这种情况通常是由于策略对全部实验单元施加时，但仅一部分实验单元被实际触发策略，并且哪些单元被触发通常由实验单元选择，实验设计者并不可控。

在策略触发的背景下，还会出现对照组的群体可能“偷偷”受到策略干预，或者实验组个体不遵守规则的场景，即依从者问题。如果感兴趣测试药物治疗对某项疾病治疗效果，随机对照实验考虑将病人随机分为实验组、对照组并且实验组病人推荐药物治疗以及对照组不吃药，实验组病人有可能没按要求吃药，以及对照组病人有可能“偷偷”吃药。此时可以采用 **CACE**（Complier Average Causal Effect）估计与推断来评估服从干预群体的策略效果。

在降方差方面，CUPED 方法的协变量选取不仅限于实验前的协变量。在实验前不存在实验单元或实验前数据与实验后数据相关性较差的情况下，也可考虑使用实验中及实验后的协变量进行调整。在 CUPED 方法的基础上，学界和业界衍生出其他降方差的方法，如 **CUPAC、MLRATE、STATE** 等。

CUPAC 和 MLRATE 都使用不受实验干预影响的协变量特征训练机器学习模型，以预测目标评估指标，并将预测值作为回归调整中的协变量进行降方差。MLRATE 在回归调整中加入了干预变量和机器学习预测变量的交互项，并使用交叉拟合减小过拟合带来的偏差。STATE 方法结合机器学习回归调整与t分布，针对厚尾数据分布进一步提升降方差效果。我们通过线下模拟和实际数据验证发现，CUPAC 和 MLRATE

能进一步减少约 10% 的方差，而 STATE 能降低接近 50% 的方差，但估计量会存在一定偏差。因此，在选择降方差方法时，建议根据具体场景验证后使用。

我们还尝试了一些其他实验方式以进一步保证同质性，未来也将考虑建设。**重随机化**是一种实验设计方法，用于在实验分组之间实现更好的平衡，进而提高实验的功效。重随机化通过多次随机分配实验单元，直到在关键协变量上达到预设的平衡标准，例如实验组和对照组的差异小于一定阈值或同质性检验中的p值大于一定阈值。这有助于确保实验组和对照组在重要特征上更加相似，从而减少混杂因素的影响。重随机化可以从两个方面提高实验功效：

1. 改善组间平衡，降低实验结果的方差，使得在相同样本量下更容易检测到真实效应；
2. 减少协变量不平衡，使实验组和对照组在关键特征上更相似，使结果更能反映策略的真实效果而非其他混杂因素。

在实践中，需要事先定义需要平衡的协变量及可接受的分流程度。当对分流均衡性要求严苛时，可能需要多次随机分流才能达到预期结果，增加了计算和时间成本。在实验设计阶段，应充分考虑这些因素，在结果准确性和计算成本之间找到最佳平衡，以确保实验的科学性和可行性。确保实验组和对照组的同质性是提高实验结果可靠性和有效性的关键步骤，通过合理设计和实施减少混杂因素的影响，能使实验结果更具可信度与推广性。

对于实验领域的溢出效应难题，我们当前主要考虑通过区域等地理单元以及订单单元之间的溢出机制建模来解决。未来，我们将进一步探索使用**马尔科夫决策过程**等方法解决无法物理隔离情况下分流溢出效应等难题。

参考资料

[1] Deng et al. (2013): Deng et al. (2013), Improving the sensitivity of online controlled experiments by utilizing pre-experiment data, In proceedings of the 24th International Conference on World Wide Web, ACM, 123–132.

[2] Lin（2013）: Winston, Lin(2013), Agnostic notes on regression adjustments to experimental data:Reexamining Freedman’s critique, The Annals of Applied Statistics, 295-318.

[3] Population Average Effect（PAE）: 在整个目标人群中某种处理或干预的平均效果。

[4] 假设前提：假设每个个体实验效果为常数，或者每个个体的实验效果虽然不同但相对提升很小。（AA 模拟情形下实验效果均为 0，假设成立）。

[5] 部分假设：假设每个个体实验效果为常数（AA 模拟情形下实验效果均为 0，假设成立）。

第四章：随机轮转实验

时间片轮转实验（Switchback Experiment）是一种基于时间随机化的实验设计，其核心思想是将实验单元在实验时间段多次进行实验组与对照组模式之间来回切换，通过比较某些指标在多时间段内实验状态与对照状态的表现差异来检测实验效应。其被广泛用于应对 AB 实验中**空间维度溢出效应**（Spillover Effects）[1] 干扰和样本量不足的问题。

- **溢出效应：**AB 实验的个体干预稳定性假设（SUTVA）假定实验单元的结果不受到其他单元分组的影响，然而实际中由于实验单元的直接关联（社交网络）或者间接关联（共享资源等），使得无法保证实验组与对照组个体之间彼此独立，进而可能导致估计的实验效应存在偏差，影响实验结论的可信度。为解决这一问题，可考虑对同一个城市进行时间片轮转实验，例如在为期 14 天的实验中，随机分配 7 天为实验组日期、7 天为对照组日期，分别施加实验策略、对照策略，以彻底消除空间溢出效应带来的估计偏差。
- **样本量不足：**当随机对照实验样本量存在不足时，例如，以单元 A 为实验单位的随机对照实验功效不足，适当的结合时间片轮转，采用实验单元 * 时间片的分流轮转实验可在相同实验时间内获得更多的样本量，进而提高实验的效率。

由于上述特点，时间片轮转实验在履约场景中被广泛应用，成为验证履约业务策略的重要工具。然而，需要注意的是，轮转实验不适用于用户感知明显的实验策略，因为这可能会严重干扰用户的自然体验。在下面章节中我们将重点介绍抛硬币随机轮转、完全 / 分层随机轮转以及配对随机轮转实验，更多轮转实验可详见拓展与展望。

方法类	应用场景
抛硬币随机轮转	适用于短轮转片场景，或者可在随机对照实验基础上搭配轮转增加实验样本量。选择轮转时间片过短时需警惕**携带效应（Carryover Effect）**[2]（可理解为上一时间片策略影响下一时间片表现）。
完全随机轮转	全城存在强溢出效应下，采用单城按天轮转实验以及多城分层轮转实验可完全消除溢出效应。
配对随机轮转	相比全城完全随机轮转，半城配对随机轮转可节约实验资源，同时更适合需聚焦于天气等不可控因素下的实验评估场景。缺点为半城配对轮转时可能存在轻微溢出效应。

4.1 抛硬币随机轮转

4.1.1 方法概述

在普通的随机分组实验中经常会面临样本量不足的问题，这可能导致无法有效检测出目标预期的提升效果。在这种情况下通常可以考虑加入时间片轮转以增加独立的样本量，具体而言，可以采用实验单元 * 时间片粒度的抛硬币随机轮转实验。在这种设计中，对于每个实验单元i和时间片t，通过独立的伯努利试验随机决定第i个实验单元在第t个时间片分配到实验组还是对照组。对于落在实验组（对照组）的实验单元i * 时间片t施加实验策略（对照策略），最后通过对比实验组和对照组的表现来估计策略的提升效果。抛硬币随机轮转实验设计比较简单，通过加入时间片轮转增加样本量通常能够显著降低方差。但其不太适用于样本量极少的场景（例如 1 个城市 14 天实验周期下的全城按天轮转实验），由于在样本量较少时容易出现实验组与对照组样本量差异明显的情况，因此建议在独立的实验个体较多或者时间片较短时的情况下考虑使用这种方法。

4.1.2 分组机制

抛硬币轮转分组是指对每个实验单元在每个轮转时间片以固定的概率随机分配该实验单元在该时间片是实验或对照状态的分组方式。对于N个实验单元（在时间片轮转实验中实验单元i * 时间片t即为一个实验单元），其分配机制满足：

$$\Pr\left(\mathbf{W} \mid \mathbf{X}, \mathbf{Y}(0), \mathbf{Y}(1)\right) = \prod_{i=1}^{N}\left[e(X_i)^{W_i} \cdot (1 - e(X_i))^{1-W_i} \right]$$

其中 $W_i \in \{0,1\}$ 代表第 i 个实验单元是否落在实验组，即 $W_i = 1$ 时为实验组，$W_i = 0$ 时为对照组，$e(X_i) = \Pr(W_i = 1 | X_i, e(X_i)) = \Pr(W_i = 1 | e(X_i)) = \Pr(W_i = 1 | X_i)$ 表示倾向得分（Propensity Score），即给定协变量 X_i 的情况下，个体 i 接受处理（即落在实验组）的概率，$e(x) \in (0,1)$，在抛硬币轮转实验实际场景中，$e(x)$ 通常选择为不依赖于协变量的固定值，即 $e(x) = q$，例如 $q = 1/2$。$\mathbf{W} = \{W_1, ..., W_i, ..., W_N\}$，$\mathbf{W} \in \mathbb{W}^+$，$\mathbb{W}^+ = \{0,1\}^N$，实验组和对照组样本量分别为 $N_t = \sum_{i=1}^{N} W_i$ 和 $N_c = N - N_t = \sum_{i=1}^{N} (1 - W_i)$，$\mathbf{X}$ 表示处理前变量或协变量，$\mathbf{Y}(1)$ 和 $\mathbf{Y}(0)$ 分别表示实验组个体和对照组个体的潜在结果向量。例如 AOI 按天抛硬币轮转实验，如图 4-1 所示每个 AOI 每天以 1/2 的概率随机分配到实验组或对照组，具体可通过一种 Hash 算法将实验单位随机分到各组，目前实验平台默认使用 MurmurHash3，分组表达式示例如下：

- AOI 按天抛硬币轮转实验分组表达式示例
- 对照组分组表达式：(murmur332(murmur332(aoi_id, 随机种子 A)+murmur332(dt, 随机种子 B), 随机种子 C)%2) in (0)
- 实验组分组表达式：(murmur332(murmur332(aoi_id, 随机种子 A)+murmur332(dt, 随机种子 B), 随机种子 C)%2) in (1)

图 4-1：AOI 按天抛硬币随机轮转示意图

4.1.3 评估原理

抛硬币随机轮转实验本质上与普通随机对照实验无差异，因此可直接引用第三章 3.1.2 普通随机分组的评估方法。类似地，抛硬币随机轮转实验同样可应用 CUPED 方法降方差，例如 AOI* 天抛硬币随机轮转实验选择对应 AOI 实验前对应周几的数据作为协变量，如若是 AOI* 天 * 小时抛硬币轮转实验选择对应 AOI 实验前对应周几对应小时的数据作为协变量，以尽量提高实验前后数据相关性，从而最大限度降低方差。

在使用抛硬币轮转实验时，同样需注意：实验单元与分析单元不一致时，错误的方差计算方式容易低估方差，导致假阳性的问题。例如某实验在分流时，将所有 AOI 分为两部分，这两部分 AOI 集合每天随机分到实验组或对照组，这时实验单位是 AOI 集合 * 天，而实验者评估时却采用 AOI* 天粒度的数据计算方差，AOI 集合下的 AOI 分组不独立，若直接套用随机化分组下的方差计算公式可能会导致低估方差，导致假阳性。如下图 4-2 左图所示，在策略没有效果的情况下，误判策略有效的概率超过 25%。正确的计算方法是将实验数据汇到 AOI 集合 * 天粒度计算方差，此时如图 4-2 右图 P 值近似服从均匀分布，且假阳性的概率控制在 5% 以内。

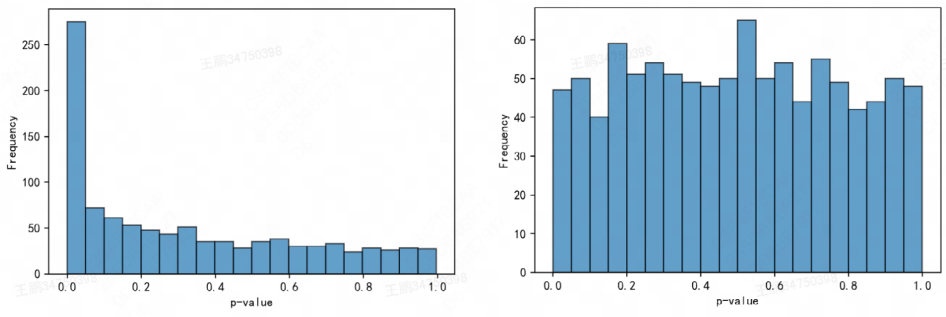


图 4-2: 1000 次 AA 模拟下的 P 值分布

4.2 完全随机轮转

4.2.1 方法概述

在全城存在强溢出效应，且小时级时间片轮转存在携带效应的情况下，一种可行的做法是采用城市按天随机轮转实验。例如在具有强 LBS[3] 业务属性的履约实验场景下，通常会存在溢出效应问题，超过 1/3 的履约实验场景采用全城按天轮转实验。然而，由于实验周期有限，城市按天轮转实验设计下的样本量（即某个城市的实验天数）通常较少。在这种情况下，若采用抛硬币方式进行轮转分组可能导致实验组和对照组天数不平衡，例如 14 天的实验可能出现 5 天实验组和 9 天对照组的情况。这种不平衡通常不符合业务方对实验状态和对照状态天数相等或相近的预期，某组天数非常少时也很难准确反映策略的效果，并可能损失实验检测功效。因此，在设计全城按天轮转实验时，通常需要特别注意组间天数的平衡，以确保实验结果的可靠性和有效性。

完全随机轮转分组为整城按天轮转实验提供了一种合适的实验设计，并具备科学的因果推断评估理论。其通常可在实验前的实验设计阶段，预先指定或固定实验组和对照组的天数，从而实现实验组和对照组天数相等或接近，甚至按需指定实验状态天数等。例如，在 14 天的按天轮转实验中，完全随机轮转允许指定恰好分配 X 天进入实验组，剩下的 14−X 天进入对照组（例如 X=7）。在此基础上，可进一步结合分层技术进行分层完全随机轮转，即先按照某些特征属性划分为多个层 / 类，再在每层分别采用完全随机轮转。例如为期 14 天实验中，按照是否周末分层，在 10 个工作日中随机分配 5 天作为实验组、5 天作为对照组，4 个周末日期随机分配 2 天为实验组、2 天为对照组。类似地，对于多个实验城市，可考虑按城市分层，在每个层（即城市）内应用完全随机轮转，以提高实验组和对照组的同质性。

4.2.2 分组机制

完全随机轮转分组本质上为对时间片分配采用完全随机分组机制，即对于N个实验单元，随机分配恰好N_t个单元于实验组中，剩下的$N - N_t$个单元分配到对照组。用数学公式表示N个实验单元完全随机分配机制为：

$$\mathbb{W}^{+}=\left\{\mathbf{W}\in\mathbb{W}\,\middle|\,\sum_{i=1}^{N}W_i=N_t\right\},\quad and\quad \Pr\left(\mathbf{W}\,\middle|\,\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1)\right)=\begin{cases}1\Big/\binom{N}{N_t}, & \mathbf{W}\in\mathbb{W}^{+}\\ 0 & \mathbf{W}\notin\mathbb{W}^{+}\end{cases}$$

其中$\mathbf{W}=\{W_1,...,W_i,...,W_N\}$，$N_t\in\{1,2,...,N-1\},\mathbb{W}=\{0,1\}^N,W_i\in\{0,1\}$代表第$i$个实验单元是否落在实验组，即$W_i=1$时为实验组，$W_i=0$时为对照组，$\mathbf{X}$表示处理前变量或协变量，$\mathbf{Y}(1)$和$\mathbf{Y}(0)$分别表示实验组个体和对照组个体的潜在结果向量。此时是按照一个组合数的倒数的概率进行分配的，每种可能性都是等概率的。对于履约最常用的城市按天完全随机轮转实验，若实验周期为 14 天，完全随机分组机制可以确保 7 天分配到实验组，另外 7 天分配到对照组。此外考虑到实际业务中周中和周末之间的差异，分组时可先对每个城市按照周中周末分层（原理见 4.2.4 分层随机轮转）。

例如，如果实验周期内共有 4 个周末天数，可以确保 2 天分配到实验组，另外 2 天分配到对照组，以此控制因周中和周末差异引起的潜在偏差，从而提高实验结果的准确性。在涉及多个城市的按天轮转实验时，建议按城市进行分层，并在每个城市内分别采用完全随机轮转。通过此方案确保每个城市实验组和对照组的同质性，同时可降低因城市之间差异带来的方差，提高检验灵敏度。

4.2.3 评估原理

考虑到完全随机轮转分组通常在样本量较小时应用，在评估时建议采用非参 Fisher 精确检验计算p值，Neyman 方法计算方差 /MDE（Fisher 无法计算方差等），具体计算逻辑如下表所示：

		连续型指标（Y）	比率型/比例型指标（X/Z）
符号说明	指标值	Y_i^{obs}表示连续型指标指标值	X_i^{obs}，Z_i^{obs}分别表示比率型或比例型指标分子和分母
	样本量	N，N_t，N_c 分别表示总样本量、实验组样本量和对照组样本量	
	分组标记	W_i代表第i个样本是否落在实验组，即$W_i=1$时为实验组，$W_i=0$时为对照组	
	指标均值	$\overline{Y}_t$，$\overline{Y}_c$分别表示实验组和对照组指标均值	$\overline{X}_t$，$\overline{Z}_t$，$\overline{X}_c$，$\overline{Z}_c$ 分别表示实验组指标分子均值和分母均值、对照组指标分子均值和分母均值，则$\overline{X}_t/\overline{Z}_t$, $\overline{X}_c/\overline{Z}_c$ 分别代表实验组、对照组观测值。
	绝对提升	$T^{\text{dif,obs}}=\overline{Y}_t-\overline{Y}_c$	$T^{\text{dif,obs}}=\dfrac{\overline{X}_t}{\overline{Z}_t}-\dfrac{\overline{X}_c}{\overline{Z}_c}$
	相对提升	$T^{\text{lift,obs}}=\dfrac{\overline{Y}_t}{\overline{Y}_c}-1$	$T^{\text{lift,obs}}=\dfrac{\overline{X}_t}{\overline{Z}_t}\Big/\dfrac{\overline{X}_c}{\overline{Z}_c}-1$
P值计算原理		采用Fisher精确检验计算P值，具体步骤如下： 1. 给定抽样次数K； 2. $for\ k\ in\ 1:K\ do$ 采用完全随机分组机制产生新的 $\mathbf{W}^k$，即从N个样本中（不放回）重新完全随机抽样N_t个实验组与N_c个对照组，计算**相对提升**统计量 $T^{\text{lift,k}}=\dfrac{\sum_{i:W_i^k=1}Y_i^{\text{obs}}/N_t-\sum_{i:W_i^k=0}Y_i^{\text{obs}}/N_c}{\sum_{i:W_i^k=0}Y_i^{\text{obs}}/N_c}=\dfrac{\sum_{i:W_i^k=1}Y_i^{\text{obs}}/N}{\sum_{i:W_i^k=0}Y_i^{\text{obs}}/N}$ 3. 近似计算p值 $\hat{p}=\dfrac{1}{K}\sum_{k=1}^{K}1_{\{\lvert T^{\text{lift,k}}\rvert\geq\lvert T^{\text{lift,obs}}\rvert\}}$	采用Fisher精确检验计算P值，具体步骤如下： 1. 给定抽样次数K； 2. $for\ k\ in\ 1:K\ do$ 采用完全随机分组机制产生新的 $\mathbf{W}^k$，即从N个样本中（不放回）重新完全随机抽样N_t个实验组与N_c个对照组，计算**绝对提升**统计量 $T^{\text{dif,k}}=\dfrac{\sum_{i:W_i^k=1}X_i^{\text{obs}}}{\sum_{i:W_i^k=1}Z_i^{\text{obs}}}-\dfrac{\sum_{i:W_i^k=0}X_i^{\text{obs}}}{\sum_{i:W_i^k=0}Z_i^{\text{obs}}}$ 3. 近似计算p值 $\hat{p}=\dfrac{1}{K}\sum_{k=1}^{K}1_{\{\lvert T^{\text{dif,k}}\rvert\geq\lvert T^{\text{dif,obs}}\rvert\}}$
方差计算原理		$var(T^{\text{lift,obs}})=var(\dfrac{\overline{Y}_t}{\overline{Y}_c})\approx\dfrac{s_t^2}{N_t}*\dfrac{1}{\overline{Y}_c^2}+\dfrac{s_c^2}{N_c}*\dfrac{\overline{Y}_t^2}{\overline{Y}_c^4}$ 其中s_t^2，s_c^2 分别表示实验组和对照组样本方差， $s_t^2=\dfrac{1}{N_t-1}\sum_{i:W_i=1}(Y_i^{\text{obs}}-\overline{Y}_t)^2$ and $s_c^2=\dfrac{1}{N_c-1}\sum_{i:W_i=0}(Y_i^{\text{obs}}-\overline{Y}_c)^2$	$var(T^{\text{dif,obs}})=var(\dfrac{\overline{X}_t}{\overline{Z}_t}-\dfrac{\overline{X}_c}{\overline{Z}_c})\approx\dfrac{s_{x,t}^2}{N_t}*\dfrac{1}{\overline{Z}_t^2}+\dfrac{s_{z,t}^2}{N_t}*\dfrac{\overline{X}_t^2}{\overline{Z}_t^4}-2*\dfrac{s_{xz,t}}{N_t}*\dfrac{\overline{X}_t}{\overline{Z}_t^3}+\dfrac{s_{x,c}^2}{N_c}*\dfrac{1}{\overline{Z}_c^2}+\dfrac{s_{z,c}^2}{N_c}*\dfrac{\overline{X}_c^2}{\overline{Z}_c^4}-2*\dfrac{s_{xz,c}}{N_c}*\dfrac{\overline{X}_c}{\overline{Z}_c^3}$ 其中$s_{x,t}^2$，$s_{z,t}^2$，$s_{x,c}^2$，$s_{z,c}^2$ 分别为实验组指标分子、实验组指标分母、对照组指标分子、对照组指标分母样本方差，$s_{xz,t}$，$s_{xz,c}$ 分别为实验组、对照组指标分子分母样本协方差。

4.2.4 分层随机轮转

当采用随机轮转实验时，如果涉及多个独立区域或城市，或者需要根据某些特征进行分层，可以进一步采用分层随机轮转实验。分层是指将N个样本按照定义好的分层变量（例如根据城市、星期几分层等），不重不漏地划分为$J\geq 2$层，每层的样本量为$N(j)$，然后在每一层中，随机分配恰好$N_t(j)$个实验单元在实验组，剩余都在对照组，即在每层中分别进行完全随机分组。用数学公式表示N个实验单元分层随机分配机制为：

$$\mathbb{W}^{+}=\left\{\mathbf{W}\in\mathbb{W}\,\middle|\,\sum_{i:B_i=j}^{N}W_i=N_t(j)\quad \text{for}j=1,2,...,J\right\}\text{and}\Pr\left(\mathbf{W}\,\middle|\,\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1)\right)=\begin{cases}\prod_{j=1}^{J}\binom{N(j)}{N_t(j)}^{-1}, & \mathbf{W}\in\mathbb{W}^{+}\\ 0 & \mathbf{W}\notin\mathbb{W}^{+}\end{cases}$$

其 中$\mathbf{W}=\{W_1,...,W_i,...,W_N\}$，$N=\sum_{j=1}^{J}N(j)$，$N_t=\sum_{j=1}^{J}N_t(j)$，$\mathbb{W}=\{0,1\}^N$，$W_i\in\{0,1\}$ 代表第i个实验单元是否落在实验组，即$W_i=1$时为实验组，$W_i=0$时为对照组，$\mathbf{X}$ 表示处理前变量或协变量，$\mathbf{Y}(1)$和$\mathbf{Y}(0)$分别表示实验组个体和对照组个体的潜在结果向量。类似地，对于分层随机轮转同样建议采用非参 Fisher 精确检验计算p值，Neyman 方法计算方差 /MDE。具体计算方法可参考 7.1 统合分析章节。

4.3 配对随机轮转

4.3.1 方法概述

为避免溢出效应而采用城市按天完全随机轮转实验时，由于实验组和对照组处于不同天，若天之间存在较大差异时，往往会导致样本数据波动较大，难以检测出策略效果。此时一种可行的做法是采用半城配对随机轮转实验，具体来说，可以事先将整个城市按照地理位置和其它相关特征划分为两个特征足够相似的半城[4]，记为（半城 A，半城 B），然后每天随机选择一个半城进入实验组，另一个半城作为对照组。例如第一天随机分配半城 A 为实验组，半城 B 为对照组；第二天再进行类似的随机分配。

在此实验设计下，同一城市中的实验组和对照组每天同时存在，相似的两个半城控制天之间差异对实验组和对照组的影响是相似的，从而有效减少随机误差。加之配套的配对评估理论，配对轮转实验通常可显著提升实验检测灵敏度。然而，由于半城配对轮转在空间上无法完全隔离实验组和对照组，仍可能存在轻微的溢出效应（主要在两个半城交界处）。与全城按天随机轮转相比，半城配对按天随机轮转实际上是通过接受一定的溢出效应偏差来换取更小的随机误差。当随机误差大于溢出效应带来的偏差时，这种实验设计能够提供更精确的结果。

4.3.2 分组机制

配对随机轮转分组使用配对随机分组，配对随机分组是指根据一些关键特征将N个实验单元划分为$J = N/2$个层 / 对（Pair），对于每层恰好只有两个实验单元，确保每对中的两个实验单元在这些特征上尽可能相似，随机将一个单元分配到实验组（分配的概率为$1/2$），另一个落在对照组。用数学公式表示配对随机分配机制为：

$$\mathbb{W}^{+}=\left\{\mathbf{W}\in\mathbb{W}\,\middle|\,\sum_{i:G_i=j}W_i=1,\text{for } j=1,2,...,N/2\right\},\quad \Pr\left(\mathbf{W}\,\middle|\,\mathbf{X},\mathbf{Y}(0),\mathbf{Y}(1)\right)=\begin{cases}2^{-N/2}, & \mathbf{W}\in\mathbb{W}^{+}\\ 0 & \mathbf{W}\notin\mathbb{W}^{+}\end{cases}$$

其中G_i表示第i个实验单元所处的层 / 对，$G_i\in\{1,...,N/2\}$，$W_i\in\{0,1\}$代表第i个实验单元是否落在实验组，即$W_i=1$时为实验组，$W_i=0$时为对照组，X表示处理前变量或协变量，$\mathbf{Y}(1)$和$\mathbf{Y}(0)$分别表示实验组个体和对照组个体的潜在结果向量。

配对的设计旨在控制实验组和对照组之间的差异，从而减少混杂变量的影响。配对随机轮转实验在配对随机基础上引入了时间片轮转机制，使得每个个体都有机会进入实验组或对照组，从而进一步控制潜在的混杂变量。对于履约最常用的半城配对随机按天轮转实验，如图 4-3 所示，其将整个城市基于地理位置和其它相关特征（协变量）划分为两个特征相似的半城，即每天两个半城为一对，每天随机选择一个半城分配到实验组，另一个半城到对照组。

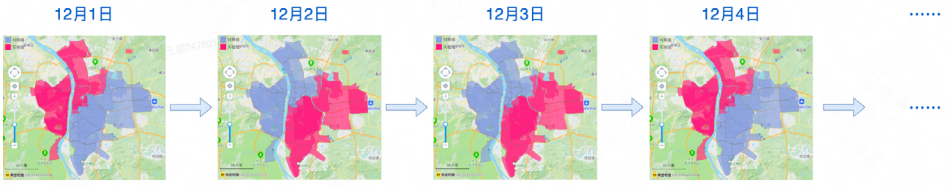


图 4-3：半城配对随机按天轮转示意图

4.3.3 评估原理

配对随机轮转实验和配对随机实验使用相同的评估方法，采用 Fisher 精确检验计算 P 值，Neyman 方法计算方差，具体计算逻辑如下表：

		连续型指标（Y）	比率型/比例型指标（X/Z）
符号说明	指标值	$Y_{j,t}^{\text{obs}}$，$Y_{j,c}^{\text{obs}}$ 分别表示实验组和对照组指标值	$X_{j,t}^{\text{obs}}$，$Z_{j,t}^{\text{obs}}$，$X_{j,c}^{\text{obs}}$，$Z_{j,c}^{\text{obs}}$ 分别表示实验组指标分子和分母、对照组指标分子和分母
	配对数目	J 表示配对数目，即为实验组或对照组样本量	
	分组标识	$W_{j,A}$ 代表第 j 个配对 (A_j, B_j) 中个体 A_j 是否落在实验组，即 $W_{j,A}=1$ 时，A_j 为实验组，B_j 为对照组；$W_{j,A}=0$ 时，A_j 为对照组，B_j 为实验组	
	指标均值	$\overline{Y}_t$，$\overline{Y}_c$ 分别表示实验组和对照组指标均值	$\overline{X}_t$，$\overline{Z}_t$，$\overline{X}_c$，$\overline{Z}_c$ 分别表示实验组指标分子均值和分母均值、对照组指标分子均值和分母均值
	绝对提升	$T^{\text{dif,obs}} = \overline{Y}_t - \overline{Y}_c$	$T^{\text{dif,obs}} = \dfrac{\overline{X}_t}{\overline{Z}_t} - \dfrac{\overline{X}_c}{\overline{Z}_c}$
	相对提升	$T^{\text{lift,obs}} = \dfrac{\overline{Y}_t}{\overline{Y}_c} - 1$	$T^{\text{lift,obs}} = \dfrac{\overline{X}_t}{\overline{Z}_t} \Big/ \dfrac{\overline{X}_c}{\overline{Z}_c} - 1$
P值计算原理		采用Fisher精确检验计算P值，具体步骤如下： 1. 给定抽样次数 K； 2. $for\ k\ in\ 1:K\ do$ 采用配对随机分组机制产生新的 $(W_{1,A}^k, ..., W_{J,A}^k, ..., W_{J,A}^k)$，即对于每个配对 (A_j, B_j) 随机选择一个进入实验组，另一个为对照组，计算**相对提升**统计量 $T^{\text{lift},k} = \dfrac{\frac{1}{J}\sum_{j=1}^{J}\left(Y_{j,t}^k - Y_{j,c}^k\right)}{\frac{1}{J}\sum_{j=1}^{J} Y_{j,c}^k} = \dfrac{\frac{1}{J}\sum_{j=1}^{J}\left(W_{j,A}^k * Y_{j,A}^{\text{obs}} + \left(1-W_{j,A}^k\right) * Y_{j,B}^{\text{obs}}\right)}{\frac{1}{J}\sum_{j=1}^{J}\left(\left(1-W_{j,A}^k\right) * Y_{j,A}^{\text{obs}} + W_{j,A}^k * Y_{j,B}^{\text{obs}}\right)} - 1$ 其中 $Y_{j,A}^{\text{obs}}$，$Y_{j,B}^{\text{obs}}$ 分别表示第 j 个配对中单元 A_j 和 B_j 的指标值； 3. 近似计算p值 $\hat{p} = \dfrac{1}{K}\sum_{k=1}^{K} 1_{\{|T^{\text{lift},k}| \geq |T^{\text{lift,obs}}|\}}$	采用Fisher精确检验计算P值，具体步骤如下： 1. 给定抽样次数 K； 2. $for\ k\ in\ 1:K\ do$ 采用配对随机分组机制产生新的 $(W_{1,A}^k, ..., W_{J,A}^k, ..., W_{J,A}^k)$，即对于每个配对 (A_j, B_j) 随机选择一个进入实验组，另一个为对照组，计算**绝对提升**统计量 $T^{\text{dif},k} = \dfrac{\sum_{j=1}^{J}\left(W_{j,A}^k * X_{j,A}^{\text{obs}} + \left(1-W_{j,A}^k\right) * X_{j,B}^{\text{obs}}\right)}{\sum_{j=1}^{J}\left(W_{j,A}^k * Z_{j,A}^{\text{obs}} + \left(1-W_{j,A}^k\right) * Z_{j,B}^{\text{obs}}\right)} - \dfrac{\sum_{j=1}^{J}\left(\left(1-W_{j,A}^k\right) * X_{j,A}^{\text{obs}} + W_{j,A}^k * X_{j,B}^{\text{obs}}\right)}{\sum_{j=1}^{J}\left(\left(1-W_{j,A}^k\right) * Z_{j,A}^{\text{obs}} + W_{j,A}^k * Z_{j,B}^{\text{obs}}\right)}$ 其中 $X_{j,A}^{\text{obs}}$，$Z_{j,A}^{\text{obs}}$，$X_{j,B}^{\text{obs}}$，$Z_{j,B}^{\text{obs}}$ 分别表示第 j 个配对中单元 A_j 和 B_j 的指标的分子和分母； 3. 近似计算p值 $\hat{p} = \dfrac{1}{K}\sum_{k=1}^{K} 1_{\{|T^{\text{dif},k}| \geq |T^{\text{dif,obs}}|\}}$
方差计算原理		相对提升方差： $var(T^{\text{lift,obs}}) = var\left(\dfrac{\bar{r}^{\text{dif}}}{\overline{Y}_c}\right) \approx var(\bar{r}^{\text{dif}})\left(\dfrac{1}{4} * \dfrac{1}{\overline{Y}_c^2} + \dfrac{1}{4} * \dfrac{\overline{Y}_t^2}{\overline{Y}_c^4} + \dfrac{1}{2} * \dfrac{\overline{Y}_t}{\overline{Y}_c^3}\right)$ 其中 $\bar{r}^{\text{dif}} = \overline{Y}_t - \overline{Y}_c$	绝对提升方差： $var(T^{\text{dif,obs}}) = var\left(\dfrac{\overline{X}_t}{Z_t} - \dfrac{\overline{X}_c}{Z_c}\right) \approx \left(\dfrac{1}{4\overline{Z}_c^2} + \dfrac{1}{4\overline{Z}_c^2} + \dfrac{2}{4\overline{Z}_t\overline{Z}_c}\right) * s_r^2 + \left(\dfrac{\overline{X}_t^2}{4\overline{Z}_t^4} + \dfrac{\overline{X}_c^2}{4\overline{Z}_c^4} + \dfrac{2\overline{X}_t\overline{X}_c}{4\overline{Z}_t^2\overline{Z}_c^2}\right) * s_v^2 - \left(\dfrac{2\overline{X}_t}{4\overline{Z}_t^3} + \dfrac{2\overline{X}_c}{4\overline{Z}_c^3} + \dfrac{2\overline{X}_c}{4\overline{Z}_t\overline{Z}_c^2} + \dfrac{2\overline{X}_t}{4\overline{Z}_t^2\overline{Z}_c}\right) * \dfrac{s_{rv}}{n}$ 其中 n 为配对数目，$\bar{r}_j = X_{j,t} - X_{j,c}$，$\bar{v}_j = Z_{j,t} - Z_{j,c}$，$\bar{r} = \dfrac{\sum_{j=1}^{n}\bar{r}_j}{n}$，$\bar{v} = \dfrac{\sum_{j=1}^{n}\bar{v}_j}{n}$， $s_r^2 = \dfrac{\sum_{j=1}^{n}(\bar{r}_j - \bar{r})^2}{n-1}$，$s_v^2 = \dfrac{\sum_{j=1}^{n}(\bar{v}_j - \bar{v})^2}{n-1}$，$s_{rv} = \dfrac{\sum_{j=1}^{n}(\bar{r}_j - \bar{r})(\bar{v}_j - \bar{v})}{n-1}$，

4.4 拓展与展望

4.4.1 异常场景处理

在按天轮转实验中，若实验期间出现突发性外部干扰，可能导致指标波动剧烈，影响策略效果的检测。针对此类场景，可根据实际需求选择以下处理方式。

方式 1：异常值剔除

对于非目标场景或无需关注特定干扰下策略效果的情况，可采用异常值剔除方法。支持自定义业务场景中反映异常状态的指标，通过统计分析识别并剔除异常值。若选择多指标，则对每个指标剔除的天数取并集。具体流程如下：

① 取过去 45 天 [5] 的数据来进行正态性检验，并用于估计 3-sigma 准则中的方差和 IQR 准则中的分位点，以此为依据进行实验数据的剔除。

② 当数据的正态性较好时，采用常用的 3-sigma 方法；当数据正态性较差时，通常会出现厚尾情况，这时采用更为激进的 IQR 方法来进行剔除。

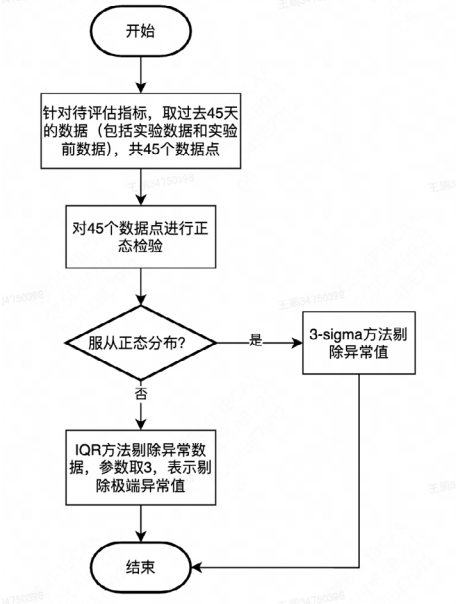


图 4-4：异常值剔除流程图

方式 2：协方差分析 +CRSE

对于需评估特定干扰场景下策略效果的情况，直接剔除数据可能引入偏误，此时可考虑采用协方差分析的方法消除混杂因素对分析指标的影响。协方差分析是用于在检验两组或多组修正均数之间有无差异时，消除混杂因素对于分析指标影响的一种分析方法。例如某实验采用城市按天完全随机轮转设计，因突发性外部干扰导致指标波动较大，这时可以以运单为个体建立回归模型，在模型中加入环境干扰等级作为协变量，以消除混杂因素影响。

由于实验采用全城按天轮转的方式，同一城市同一天的运单可能是相关的，导致模

型的误差项之间可能相关，这时使用普通最小二乘法（Ordinary Least Squares，OLS）估计的标准误是有偏的。为了解决这一问题，可以使用 CRSE（Cluster Robust Standard Error，聚类调整标准误），放宽独立同分布的假设，允许组内个体存在相关性，不同组之间个体彼此不相关，将分流维度（即城市 * 天）作为 Cluster 来评估策略效果。

4.4.2 小时级轮转下的携带效应

需要特别注意的是，尽管时间片粒度越细在实验总时长不变情况下样本量越大，通常可带来实验功效的提升。然而由于时间维度的相依性往往会导致细粒度时间片的轮转实验中存在携带效应，即上一时刻策略会影响下一时刻的表现。例如，在交通信号灯优化实验中，假设某路口每十分钟切换一次绿灯时长策略以优化车辆通行效率。若前一时间片采用缩短绿灯时长的策略（如绿灯 30 秒），可能导致车辆排队积压；即使下一时间片恢复为原有时长（绿灯 60 秒），积压的车辆仍需额外时间疏散，此时通行效率指标（如平均等待时长）仍受前一阶段策略的滞后影响。

这种跨时间片的策略干扰会导致因果效应估计偏误，影响实验结论的准确性。这时需要科学的方法消除携带效应的影响。目前对连续型指标的携带效应估计模型已在履约有所应用，但对于履约场景最常见的比率型指标下的携带效应估计还未有落地方案。

经过对学业界理论方案的调研，对于携带效应，通常有以下三种解决思路：

① 利用模型估计携带效应并辅助调整消除真实效果偏差；

② 利用消除时长（wash-out/burn-in period）去除携带效应的影响；

③ 利用时间序列模型进行优化设计。

上述的三种方案均具有一定的局限性，尽管方案 ① 易于操作，但在实际问题中可能存在模型错误等问题，影响评估效果。其次，由于模型中包含携带效应，对于携带效应的估计也会影响处理效应的估计精度。因此方案 ① 并不是一种最为理想的分析方案。相较于方案 ①，方案 ② 不依赖于模型，因此更加稳健，但需要预估携带效应影

响时长以构建合理的分配时长。在分析数据的过程中，方案 ② 还需利用消除时长去除受携带效应影响的部分数据，再对处理效应进行估计，以此实现消除携带效应。然而，目前对于如何预估携带效应时长，以及如何确定消除时长尚无明确的解决方案。

此外，方案 ② 未能将数据的时间序列特征纳入考虑，因此尽管方案 ② 具有一定的优越性，但仍不能保证该方案能较好的降低估计的方差。方案 ③ 考虑从数据的时间序列特征出发，将时间序列模型与因果推断问题结合，利用最优实验设计的想法提升处理效应的估计精度和检验功效。目前，该方案仅考虑了 ARMA(p,q) 模型，因此对于实际问题中可能存在的非平稳过程并不完全适用。因此尽管已线下落地上述部分方法，但如何针对美团的履约问题开发最合理的按小时轮转实验方案还有待进一步的研究。我们已通过校企合作，针对美团的业务场景开展研究，为美团履约平台开发具有优良性质的按小时轮转实验以减小携带效应以及时间混淆效应的影响。

4.4.3 其他轮转实验设计

交替轮转实验是另一类重要的时间片轮转实验设计，其特点是在连续的时间片中交替分配实验组和对照组，例如上一时间片为实验组，下一时间片为对照组，再下一时间片为实验组这种交替改变分组的方式。通常而言这种实验设计往往更符合实际业务诉求，尤其是在每天的各时间片具有明显周期性且各时间片差异显著的场景下。然而，在评估方面（尤其是 p 值计算）通常需要模型 / 条件假设，这对评估的科学性具有一些挑战。

例如，业界的一些应用案例，DoorDash 在评估广告效果时采用按天的时间片交替轮转实验，通过使用历史数据 +Bootstrap 抽样来近似构造统计量在原假设下的分布，从而进行 t 检验，但该方式需要假设历史基线与实验期间保持一致。

国内某互联网公司考虑小时级交替轮转实验，并在评估时引入未考虑携带效应的 VCM（Varying Coefficient Model）模型，或者考虑携带效应的 VCDP（Varying Coefficient Decision Process）模型，由于理论细节较多，暂不在此做详细介绍，感兴趣的读者可以进一步查阅相关文献。

但需要注意的是，按天交替轮转实验的分组机制是比较偏向于非随机的：一旦确定实验开始的第一天属于实验组或对照组，后续天的分组将相继确认。这时如果实验者采用完全随机轮转实验等方法计算方差，忽略分配机制可能导致方差计算的错误。此外，交替轮转基于某些假设的建模分析，通常需要较大的样本，在按天轮转样本量较少的场景下通常不适用。

解释说明

[1] 溢出效应（Spillover Effects）：AB 实验中关键的个体干预稳定性假设（SUTVA）假定实验单元的结果不受到其他单元分组的影响，然而实践中由于实验单元间的直接关联（社交网络）或者间接关联（竞争共享资源等），参与 AB 实验的实验组与对照组之间可能并不独立，我们通常称实验组、对照组间的干扰影响为溢出效应。

[2] 携带效应（Carryover Effect）：可理解为时间维度的溢出，指某一时刻的策略效果或影响延续到后续时刻，影响后续时刻的策略或效果。携带效应的阶为 m 代表时刻 t 最多会影响 t+m 时刻的结果，不会影响更后时刻的潜在结果。

[3] LBS：Location-Based Services，基于位置的服务

[4] 半城：半城在此是指同一个城市中由地理位置相邻的具有经纬度信息的单元（例如配送区域、AOI 等）组成的集合。

[5] 45 天：45 天的考虑在于，历史数据的时间跨度不应取的太长，因为不同季节的异常天气影响可能差别较大，较久前的数据刻画跟当前季节的实际情况有差距；同时天数不应取的太少，否则会导致正态性检验，方差和分位点估计的不准确。

第五章：准实验

准实验（Quasi-experiment）适用于“实验设计者”可干预分组，但无法随机分配实验单元至实验组和对照组的场景。经典随机对照实验通过随机分配实验单元，保证了实验组和对照组的可观测特征和不可观测特征分布都是相同的，差异仅在于样本是否受策略影响，因此两组观测结果的差异可以归因于策略影响。然而，在一些无法随机分配样本的场景下，实验组和对照组的特征分布往往不一致，进而导致两个组在未施加策略时就存在差异，此时需在满足部分特定条件假设的前提下使用准实验评估方法，才能够比较准确地估计策略的效果。

以美团履约业务场景为例，以下几个因素可能阻碍进行时空粒度的随机实验。

溢出效应 + 小样本等多重约束下无法开展时空随机实验：

- **溢出效应：**履约业务是一个典型的多边场景，容易造成实验单元间相互依赖和影响，而简单的随机对照实验，通常会违背个体处理稳定性假设（SUTVA），进而造成实验偏差。在这种存在溢出效应的履约业务场景中，实验有时需要在地理上隔离样本，以避免或者减少溢出效应，一种典型的做法是依据地理位置将一个城市划分为两个半城，将实验组和对照组之间的运力溢出等限制在半城交界处，将溢出效应的影响尽量降至最低。
- **小样本：**履约策略大多以配送区域为基本单元，即使是区域溢入溢出效应模型也通常要求配送区域数量至少超过 20 个。但是部分城市规模较小，可供分析建模的配送区域数量达不到该要求，因此也无法采用随机分组 + 溢出效应建模的实验方案。
- **策略和产品的特殊性：**部分策略和产品的特殊性限制了随机分组。例如，配送区域优化策略考虑在保障整体覆盖范围不变且区域之间不重叠的约束下，对区域进行边界优化甚至合并，然而对于 2 个相邻的区域，在该约束下，优化 A 区域边界必然会导致 B 边界跟随变化，因此从产品形态上无法实现 A 区域边

界变更但 B 区域边界维持不变，此时不能考虑按区域随机分流。

综上所述，考虑到美团履约业务场景的特殊性，许多实验无法采用随机对照实验准确量化策略效果，因此发展一套标准的准实验设计与评估流程尤为必要。接下来，我们着重介绍经典的准实验方法——双重差分法，关于双重差分法的衍生和其他准实验方法，请参考文末的拓展部分。

5.1 双重差分法

5.1.1 方法概述

双重差分法（Difference in differences，简称 DID）的基本思想，就是用**实验后的实验组、对照组差异减去实验前的实验组、对照组差异**，来估计策略在实验组上的效果（ATT），图 5-1 直观展示了该思想。下面先从单重差分开始，逐步解析双重差分法。

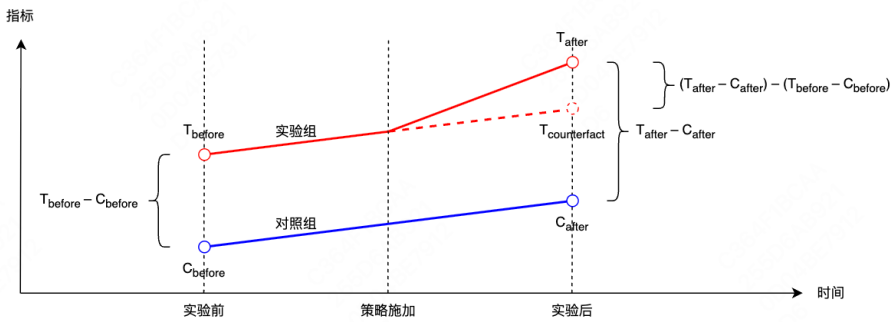


图 5-1：双重差分法示意图

在无法实施随机对照实验的情况下，我们可以通过收集实验组群体、对照组群体在实验前后时间段的面板数据，来分析策略效果。不失一般性，假设实验前实验组个体平均值记为 T_{before}，实验前对照组个体平均值记为 C_{before}，实验后实验组个体平均值记为 T_{after}，实验后对照组个体平均值记为 C_{after}。我们先看 2 种 Naive 的评估方法。

1. **实验组 – 对照组（横截面单重差分法）：**即用实验后实验组的观测值T_{after}减去实验后对照组的观测值C_{after}得到$T_{after} - C_{after}$，来估计策略效果。但是在无法随机分配实验组和对照组的情况下，两组之间往往存在固有差异，因此简单地使用实验组减对照组的估计结果可能会存在偏差。
2. **实验后 – 实验前（时间序列单重差分法）：**即用实验后实验组的观测值T_{after}减去实验前实验组的观测值T_{before}得到$T_{after} - T_{before}$，来估计策略效果。但是随着时间推移外部条件发生变化，即使不施加策略，实验组指标也可能会随时间自然变化，因此使用实验后减实验前的估计结果往往也存在偏差。

双重差分法在上述两种单重差分法基础上进行了改进：其基本思想为假设实验组和对照组在不施加策略时差异固定不变，用实验组对照组在实验后的差异减去实验组对照组在实验前的差异，得到：

$$(T_{after} - C_{after}) - (T_{before} - C_{before})$$

消除了两组之间的固有差异，这就是双重差分法的基本原理。

	实验组	对照组	实验组-对照组 （横截面单重差分）
实验前	T_{before}	C_{before}	$T_{before} - C_{before}$
实验后	T_{after}	C_{after}	$T_{after} - C_{after}$
实验后-实验前 （时间序列单重差分）	$T_{after} - T_{before}$	$C_{after} - C_{before}$	$(T_{after} - C_{after}) - (T_{before} - C_{before})$

5.1.2 评估原理

本节我们将详细介绍双重差分法的数学模型和原理，包括传统 DID 模型、固定效应模型、平行趋势假设合理性检验等。

传统 DID 模型

基本双重差分法模型的形式为：

$$Y_{it} = \beta_0 + \beta_1 \times treat_i + \beta_2 \times after_t + \beta_3 \times treat_i \times after_t + e_{it}$$

其中，i表示个体，t表示时间；Y_{it}是个体i在时间t的指标值；$treat_i$是分组虚拟变量，如果个体i属于实验组，则$treat_i = 1$，否则$treat_i = 0$；$after_t$是分期虚拟变量，如果时间t在实验后，则$after_t = 1$，否则$after_t = 0$；e_{it}是期望为 0 的干扰项；β_j是待估计的模型系数。

在该模型下，实验组对照组在实验前后的均值以及差分可以写成下表中展示的形式，并以此解读各项系数：β_0表示实验前对照组的均值；β_1表示实验前实验组均值和对照组均值的差异；β_2表示对照组在实验前后均值的差异；β_3表示策略的效果。

	实验组	对照组	实验组-对照组 （横截面单重差分）
实验前	$T_{before} = E(Y_{it} \mid treat_i = 1, after_t = 0) = \beta_0 + \beta_1$	$C_{before} = E(Y_{it} \mid treat_i = 0, after_t = 0) = \beta_0$	$T_{before} - C_{before} = \beta_1$
实验后	$T_{after} = E(Y_{it} \mid treat_i = 1, after_t = 1) = \beta_0 + \beta_1 + \beta_2 + \beta_3$	$C_{after} = E(Y_{it} \mid treat_i = 0, after_t = 1) = \beta_0 + \beta_2$	$T_{after} - C_{after} = \beta_1 + \beta_3$
实验后-实验前 （时间序列单重差分）	$T_{after} - T_{before} = \beta_2 + \beta_3$	$C_{after} - C_{before} = \beta_2$	$(T_{after} - C_{after}) - (T_{before} - C_{before}) = \beta_3$

使用传统 DID 模型对实验前后面板数据进行回归计算，可以得到策略效果的估计量$\hat{\beta}_3$、估计量方差、p值、置信区间、MDE 等一系列统计量。此处需要强调一点，上述对于β_3的估计其实是对于指标绝对提升的估计，但在实际场景的应用中还会希望计算某些指标的相对提升，此时需要额外计算实验组的反事实：

$$T_{counterfact} = T_{before} + (C_{after} - C_{before}) = C_{after} + (T_{before} - C_{before})$$

表示实验组在实验后假如未施加策略时的均值，并进一步计算相对提升。

固定效应模型

为了进一步提高模型的精度，可以引入时间和个体固定效应，对不同时间和不同个体

的效应做进一步细化，可以得到时间固定效应模型和时间 + 个体固定效应模型。固定效应模型可以消除时间和个体上的差异，估计系数的方差也会降低，对于策略效应的检测也更加灵敏，在实践中会优先使用固定效应模型。引入了时间固定效应 ϕ_t 的时间固定效应模型形式如下：

$$Y_{it} = \alpha + \beta \times treat_i \times after_t + \alpha_1 \times treat_i + \phi_t + e_{it}$$

在对系数进行估计时，进一步引入时间虚拟变量 T_t，将模型写为以下形式，再使用最小二乘法进行计算，此即最小二乘虚拟变量估计法：

$$Y_{it} = \alpha + \beta \times treat_i \times after_t + \alpha_1 \times treat_i + \sum_{t=1}^{T-1} \gamma_t \times T_t + e_{it}$$

引入了时间固定效应 ϕ_t 和个体固定效应 λ_i 的时间 + 个体固定效应模型形式如下：

$$Y_{it} = \alpha + \beta \times treat_i \times after_t + \lambda_i + \phi_t + e_{it}$$

在对系数进行估计时，仍可使用最小二乘虚拟变量估计法，进一步引入时间虚拟变量 T_t 和个体虚拟变量 D_i，将模型写为以下形式再计算：

$$Y_{it} = \alpha + \beta \times treat_i \times after_t + \sum_{i=1}^{N-1} \delta_i \times D_i + \sum_{t=1}^{T-1} \gamma_t \times T_t + e_{it}$$

在个体数目较多时，虚拟变量过多可能会导致计算性能的问题，此时可以使用个体内差分估计法：

① 首先将模型写作以下形式，其中 $\mathbf{X}_{it}$ 由交叉项 $treat_i \times after_t$ 和时间虚拟变量 T_t 组成：

$$Y_{it} = \alpha + \mathbf{X}_{it}^T \boldsymbol{\beta} + \lambda_i + e_{it}$$

② 对每个个体的变量取平均值，得到：

$$\overline{Y}_i = \alpha + \overline{\mathbf{X}}_i^T \boldsymbol{\beta} + \lambda_i + \overline{e}_i$$

③ 相减得到下式，其中 $\widetilde{Y}_{it} = Y_{it} - \overline{Y}_i$，$\widetilde{\mathbf{X}}_{it} = \mathbf{X}_{it} - \overline{\mathbf{X}}_i$，$\tilde{e}_{it} = e_{it} - \overline{e}_i$：

$$\widetilde{Y}_{it} = \widetilde{\mathbf{X}}_{it}^T \boldsymbol{\beta} + \tilde{e}_{it}$$

④ 再使用最小二乘法计算该模型，得到对系数$\boldsymbol{\beta}$的估计，可以证明使用个体内差分估计法和最小二乘虚拟变量估计法得到的结果一致。

平行趋势假设合理性检验

平行趋势假设是使用双重差分法估计策略效果的关键假设。平行趋势假设要求，在没有策略影响的情况下，实验组和对照组的差异不随时间变化是恒定的，即实验组和对照组的趋势保持平行。一种简单的平行趋势检验方法是通过画图观察平行趋势是否满足，但是这种方法比较粗糙。为了得到更加严谨的量化结果，可以使用模型进行平行趋势检验。在此基础上一种方法是将 DID 模型拓展为以下形式：

$$Y_{it} = \beta_0 + \beta_1^1 \times treat_i + \beta_1^2 \times treat_i \times time_2 + \ldots + \beta_1^{T_0} \times treat_i \times time_{T_0} + \beta_2 \times after_t + \beta_3 \times treat_i \times after_t + e_{it}$$

其中$time_t$是时间虚拟变量，$t \le T_0$为实验前没有施加策略的时期。如果平行趋势成立，那么在没有施加策略的每个时间点，实验组和对照组的差异没有显著变化，即有$\beta_1^2 = \ldots = \beta_1^{T_0} = 0$。因此，只需检验$\beta_1^2 = \ldots = \beta_1^{T_0} = 0$是否成立。若系数均不显著（$p$值大于 0.05），则认为通过平行趋势检验。

另一种常用方法是，指定一个实验前时间为虚构的策略开始施加的时间，然后使用 DID 模型对实验前数据进行回归分析，若系数不显著（p值大于 0.05），则认为通过平行趋势检验。这种方法又称作安慰剂检验，虽然严格性不如第一种方法，但是胜在简单好用，该检验方法和实验评估方法一致，在实践中更加常用。

5.1.3 平行趋势分组

不难看出，平行趋势假设是影响双重差分实验结论可信度的关键。因此，为了尽量保证实验结论的可信度，我们建议采取下述平行趋势分组，以尽量保障“实验组”、“对照组”平行趋势假设的合理性：

1. 随机划分 2 个半城为实验组和对照组；
2. 使用实验前数据，对所有目标指标和护栏指标做平行趋势检验，根据通过检

验的模型和实验组对照组差异对本次分组进行打分（通过固定效应模型平行趋势检验的分组得分更高，两组差异更小的分组得分更高）；

3. 重复步骤 1 和步骤 2 若干次，选取得分最高的分组作为最终分组。

尽管采取平行趋势分组的做法在实验设计上尽量保障平行趋势假设的合理性，但在实际场景中仍存在以下潜在风险，因此在实践中优先考虑随机实验，随机实验不可行时才考虑双重差分实验：

1. 平行趋势是一个比较强的假设，在样本量较少时，有时难以划分满足平行趋势的实验组和对照组；
2. 平行趋势检验只能检验实验前的平行趋势以证明假设的合理性，实验后的平行趋势是否满足是无法得知的，并且无法得到完全保证，在某些情况下平行趋势假设会受到挑战：

   a. 有不可控的外部因素影响时，平行趋势假设可能被打破，此时可考虑适当剔除不可控因素影响日期再进行评估分析；

   b. 评估指标的数值限定范围，可能影响到平行趋势。在履约场景中准时率指标时常被关注，准时率的数值范围在 0 ~ 100% 之间并且通常处于较高水位，在某些极端情况下如果平行趋势成立，实验组准时率的反事实结果可能会超过其上限 100%，这时平行趋势假设与实际情况会略有出入。

5.1.4 实验案例

实验案例：配送区域优化实验

实验背景：为解决现有配送区域划分畸形、切割商户热力等问题，提升配送效率，通过算法智能规划对各城市配送区域进行重新规划。

实验目标：降低运单超出配送区域范围占比，提高配送效率。

实验指标：

- **目标指标：**XXXX；
- **护栏指标：**XXXX。

实验难点及约束：

- **策略和产品的特殊性：**配送区域优化策略考虑在保障整体覆盖范围不变且区域之间不重叠的约束下，对加盟区域进行边界优化甚至合并，然而对于 2 个相邻的区域，在该约束下，优化 A 区域边界必然会导致 B 边界跟随变化，因此从产品形态上无法实现 A 区域边界变更但 B 区域边界维持不变，此时不能考虑按区域随机分流。这种情况下可以考虑将城市划分为两个半城，在实验半城内部调整优化区域边界，对照半城维持不变。

实验方案：考虑到实验难点及约束，采用半城平行趋势分组，并使用双重差分法进行评估。

实验设计：采用半城划分 + 平行趋势检验的实验设计机制，对城市中配送区域进行分组，详细流程可见图 5-2：

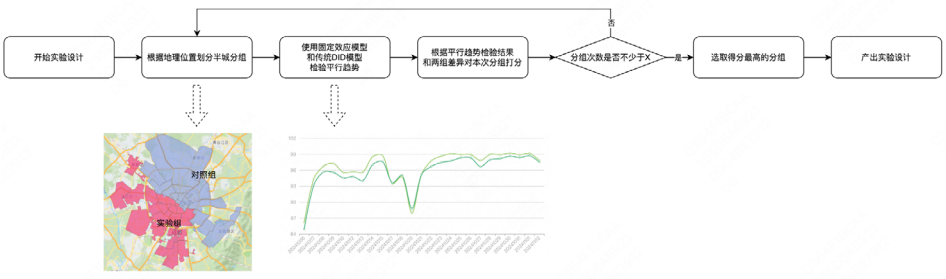


图 5-2：实验设计流程

实验评估：根据实验前通过哪个模型的平行趋势检验来决定用哪个模型来评估实验后策略效果，详细流程可见图 5-3，评估结果以下表为例：

指标名	结论	差异相对值	差异绝对值	p-value	MDE	置信区间	实验前实验组指标值	实验后实验组指标值	对照组实验前指标值	对照组实验后指标值
xxxx 检验方式：时间固定效应模型检验	显著提升	xx %	xx pp	xx	xx pp	[xx, xx]	xx %	xx %	xx %	xx %

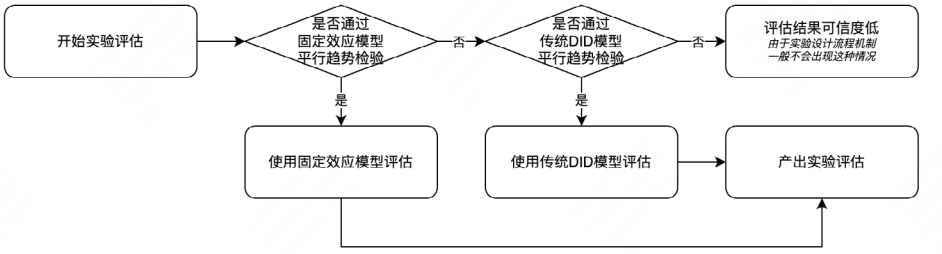


图 5-3：实验评估流程

5.2 拓展与展望

5.2.1 双重差分法拓展

在传统 DID 模型设定中，一个隐含假设是，实验组的所有个体开始实验的时间均相同。但有时我们也会遇到每个个体的实验时间不完全一致的情形（Staggered Timing），比如有的实验经过逐步放量，一部分个体从实验第 1 天就开始接受策略处理，而另一部分个体则等到放量之后，第 8 天才开始接受策略。这时我们就可以用**多时点 DID 模型**来同时考察多次实验的效果，模型设定如下：

$$Y_{it} = \alpha + \lambda_i + \phi_t + \beta \times effect_{it} + e_{it}$$

其中对于某个个体的$effect_{it}$，如果个体i是从t_0时开始接受实验，那么t_0之前时间点$effect_{it}$取 0，t_0及之后时间点$effect_{it}$取 1，对于从来没有接受策略的对照组，$effect_{it}$始终取 0。同样 5.1 节介绍的处理效应模型一般假设“同质性处理效应”，即所有个体的处理效应都相同。对此，**异质性双重差分模型**在传统 DID 基础上引入“异质性处理效应”，即允许每位个体（或群体）的处理效应不尽相同。具体而言是对双重差分模型中交互项（treat*post）的调整，即引入在组别上的交互项（treat*post*group）。

此外，在固定效应模型中可进一步考虑加入**其他可观测的随时间变化的协变量**X_{it}，但前提为协变量不可受实验策略影响。引入新的协变量通常存在两个重要作用，一方面通常可以更进一步的降低估计值的方差；另一方面在实验前后环境存在较大变化时，添加相应的观测协变量有助于控制因协变量差异导致的估计误差，缓解因环境变更对平行趋势假设带来的不合理风险。例如拓展后的模型形式如下：

$$Y_{it} = \alpha + \beta \times treat_i \times after_t + \gamma \times X_{it} + \lambda_i + \phi_t + e_{it}$$

在实践中当出现**平行趋势不成立**的情况时（建议尽量在实验设计上采取更合理的分组，如果现实中已经结束实验并平行趋势检验表明假设不合理时），通常可以尝试如下做法：

1. **放宽平行趋势假设：**例如学界的 Honest DID 为一种在平行趋势假设可能不成立的前提下，进行稳健推断（Robust Inference）和敏感性分析（Sensitivity Analysis）的方法。与直接假设平行趋势成立不同，Honest DID 允许实验后平行趋势的违背，但是限制违背程度与实验前趋势（pre-trends）的违背并不存在太大差异或至少有迹可循。
2. **条件平行趋势假设：**通过匹配等方法寻找满足平行趋势的群体，例如基于实验群体 PSM 匹配合适的对照组群体，再应用 DID 进行评估等。
3. **三重差分法：**在双重差分基础上引入第三个差异维度（不受干预影响）更精确评估政策或干预措施影响的计量经济学方法，但也增加了数据需求和模型复杂性。

5.2.2 其他准实验方法

本文在准实验上着重介绍了双重差分法，此外还有一些断点回归、中断时间序列等类准实验方法可供读者参考。

1. 断点回归（Regression Discontinuity Design, RDD）根据某个可观测变量的阈值（断点）划分为实验组和对照组，分析主要集中在断点附近的样本上。断点附近可以认为有局部随机性，即断点附近的样本是否受处置是随机的，

并且在是否处置之外的特征上没有系统性差异。

2. 中断时间序列（Interrupted Time Series Analysis, ITSA）具体做法为在干预之前，使用不同时间的多次测量来创建一个模型（例如时间序列分析 ARIMA 模型），该模型可以估计干预介入后的相关指标的虚拟事实。干预后，再进行多次测量，并将关注指标的实际值和模型的预测值之间的平均差作为实验效应的估计。当然中断时间序列同样可应用于多个实验对象并且各个实验对象可在不同时间点接受实验干预（即设计上类似于多基线实验）。此外简单中断时间序列的一种拓展是引入实验变动然后将其反转，并可以选择多次重复此过程。

第六章：观察性研究

观察性研究常用于解决无法进行控制实验的问题。在美团的到家履约业务场景中，由于法律约束以及实际操作成本等多种限制，我们通常无法直接进行控制实验。因此，**观察性研究成为一种重要的替代方法，它允许我们在不进行控制实验，且不影响用户体验的情况下，评估不同策略和措施的业务效果**。

著名统计学家 Cochran（1965）总结了观察性研究的两个常见特征：**一是目标是阐明因果关系，二是使用控制实验不可行**。第一个特征与随机对照实验或准实验相同，但第二个特征与其有根本性的不同：随机对照试验和准实验的干预是外生的，不受实验个体自身控制，不存在自选择问题，而在观察性研究中，干预是不可控的，即我们无法通过实验的方式控制一部分实验个体分配到实验组和对照组，这可能存在**选择性偏差问题**（由于样本的选择方式不当，使得样本不能代表总体，导致评估结果具有偏差）。选择合适的观察性研究方法，能够帮助我们在无法进行控制实验分组的情况下，尽可能消除选择性偏差，得到较为科学的评估结果。

接下来，我们将介绍一些具体的观察性研究方法，包括合成控制法、匹配方法以及 Causal Impact 等。各方法的基本思想和适用场景简单总结如下表，每个方法的具体细节可参考对应章节，一些其他观察性研究方法的简单介绍可见拓展部分。

方法类型	基本思想	适用场景
合成控制法	通过对若干与干预地区相似的未干预对照组进行线性加权，构造出一个虚拟对照组，用以近似干预地区在未受干预情况下的情形。	1.干预单元数量有限； 2.丰富的对照组单元； 3.多期面板数据。
匹配(PSM等)	通过匹配方式，为实验组中的样本找到高度相似的未接受干预的反事实样本，从而有效控制协变量的影响，以更精准地估计因果效应。	1.多个协变量影响； 2. 在随机化不可行的情况下减少选择偏差。
Causal Impact	基于贝叶斯结构时间序列（Bayesian Structural Time Series, BSTS）模型，构建一个“虚拟对照组”，用于预测在没有干预措施情况下目标变量的可能表现。	1.只有少数单元受到干预； 2.存在多个与这些受干预单元相关的协变量，但这些协变量并未受到干预。

6.1 合成控制法

6.1.1 概述

2024 年，北京发布了《餐饮外卖流通绿色包装评价要求》，这项规定对美团北京地区外卖履约业务会有多大影响？为了评估这类事件或政策的影响，根据潜在因果框架理论，我们需要为受政策影响的地区构建“反事实”结果，即如果该地区未受干预会如何。通常，这需要选择一个在各方面与受干预地区相似的对照组，然而，干预政策通常只发生在特定地区，由于美团外卖履约业务的特殊性，我们很难找到一个业务特征[1]相似的对照组。

为此，我们可以考虑为干预地区构建一个未受干预且特征相似的对照组，具体而言，通过对若干与干预地区相似的未干预对照组进行线性加权，构造出一个虚拟对照组，用以近似干预地区在未受干预情况下的情形，这便是 Abadie 和 Gardeazabal（2003）[2] 提出的“合成控制法”。

基本思想

合成控制法（Synthetic Control Method, SCM）的基本思想是通过从其他相似地区的数据中学习权重，构建一个加权平均的“合成对照组”来估计政策或干预对一个处理单元（如一个城市、国家或公司）的因果效应，该方法特别适用于个案研究，尤其是在随机对照试验不可行的情况下，其主要流程可以见下图 6-1：

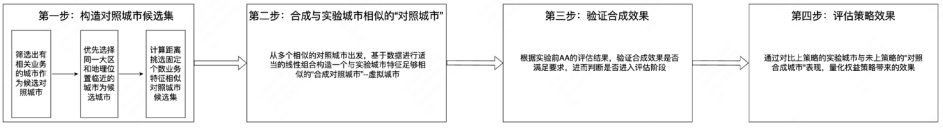


图 6-1：合成控制法

适用场景与优缺点

在实际应用中，合成控制法具有其独特的优势，尤其是在以下业务场景中尤为适用：

- **无法进行随机对照实验或准实验：**由于法律约束以及实际操作成本等多种限制，无法实施随机对照实验或准实验时，合成控制法提供了一种有效的替代方法，通过构建合成对照组来模拟对照实验的效果。
- **干预单元数量有限：**适用于评估单个或少量干预单元的影响。这种情况下，合成控制法通过利用多个对照单元的数据来创建一个合成对照组。
- **丰富的对照组单元：**需要有足够数量和多样性的对照组单元，以便从中选择并加权组合，创建一个合成对照组，使其在未受干预时表现与干预单元相似。
- **多期面板数据：**合成控制法依赖于多期面板数据，以观察干预前后干预单元和对照单元的表现。这种数据结构允许更准确地捕捉时间趋势，并验证合成对照组在干预前的适用性。

随着合成控制法被广泛应用，优缺点也逐渐明显，其优点如下：

- **适用于个案研究：**特别适合评估只有单个城市、地区或特定市场的政策或策略影响。
- **数据驱动的对照组构建：**通过加权组合多个对照组，创建一个合成对照组，模拟处理组在未受干预时的表现，可以减少单个对照城市的偏差。
- **减少模型依赖：**减少对复杂模型假设的依赖，更加依赖于观测数据的实际表现。
- **直观的可视化：**结果通常可以通过图形表示，便于干预政策影响的直观理解和解释。

然而，合成控制法也存在一些局限性，这些限制在特定情况下可能影响其应用效果：

- **数据要求高：**需要足够的对照单元和多期面板数据来构建合成对照组，对数据质量要求较高，存在较多缺失数据或者对照单元较少时可能难以评估。
- **复杂性：**合成控制的权重计算和假设检验的 p 值计算可能较为复杂，特别是存在多个处理单元时，需要计算多个权重。
- **外推性限制：**结果的外推性可能有限，由于处理组的特殊性可能并不能代表总体情况，无法轻易推广到其他场景或城市。

- **处理组和对照组的相似性要求：**要求合成的对照组能很好地模拟处理组在未受干预时的表现，但异质性较大时，合成对照组和实验组在未受干预时差别可能会较大。

6.1.2 原理

本节我们将详细介绍合成控制法的数学原理。

基本假定

假设我们有$J+1$个实验单元，其中有J个对照组和 1 个处理组，不失一般性，假设第一个单元接受干预，其余的J个地区是潜在的对照组。令Y_{it}^N为对于实验个体i在每个时间点t上未接受干预后的潜在结果，个体$i=1,\ldots,J+1$，时间周期$t\in\{1,2,\cdots,T_0\}$。令T_0为干预前的时间周期数量，其中$1\le T_0<T$，Y_{it}^I为在T_0+1到T时期内，实验个体i在每个时间点t上接受干预后的潜在结果。我们假设在实施干预之前，干预对结果没有影响，因此对于$t\in\{1,2,\cdots,T_0\}$和所有$i\in\{1,2,\cdots,J+1\}$，有有$Y_{it}^I=Y_{it}^N$。

定义个体i在时间周期t的因果效应为：$\alpha_{it}=Y_{it}^I-Y_{it}^N$，此时个体$i$在时间周期$t$的观测结果为：$Y_{it}=Y_{it}^N+\alpha_{it}D_{it}$，因为只有第 1 个实验个体被干预，且在时间周期$T_0$之后被干预，因此：

$$D_{it}=\begin{cases}1 & \text{if } i=1 \text{ and } t>T_0,\\ 0 & \text{otherwise.}\end{cases}$$

且对于$j=1$，我们只能观测到Y_{1t}^I，观测不到Y_{1t}^N；相反，对$j=2,\ldots,J+1$，我们只能观测到Y_{jt}^N，观测不到Y_{jt}^I。

假设Y_{it}^N满足如下的因子模型：

$$Y_{it}^N=\delta_t+\boldsymbol{\theta}_t\mathbf{Z}_i+\boldsymbol{\lambda}_t\boldsymbol{\mu}_i+\varepsilon_{it}$$

其中，δ_t是随时间变化的影响Y_{it}^N的公共因子，$\mathbf{Z}_i$是一个$(r\times1)$维的可观测协变量向量（不受干预影响），$\boldsymbol{\theta}_t$是一个$(1\times r)$维的未知参数，$\boldsymbol{\mu}_i$是一个$(F\times1)$维的不可观测协

向量，$\boldsymbol{\lambda}_t$是一个$(1\times F)$维的未知公共因子，在各个体中具有不同的因子载荷$\boldsymbol{\mu}_i$，ϵ_{it}是未观测到的随机误差项，对所有个体i，其期望为零。

权重计算

在合成控制法中，我们关心的是 ATT（Average Treatment Effect of the Treated），也就是说我们希望推断的参数是$\tau_{1t}=Y_{1t}^{I}-Y_{1t}^{N}$，因此如何估计$Y_{1t}^{N}$是我们要解决的问题。前文讲到，合成控制法中我们会对对照组中的城市进行线性组合，组成一个虚拟的和干预个体 1 相似的对照单元。也就是说，对每个对照组个体$j=2,\cdots,J+1$，给定权重$\mathbf{W}=(w_2,w_3,\ldots,w_{J+1})$，满足$w_j\geq 0,\sum_{j=2}^{J+1}w_j=1$，此时有：

$$\sum_{j=2}^{J+1}w_jY_{jt}=\delta_t+\boldsymbol{\theta}_t\sum_{j=2}^{J+1}w_j\mathbf{Z}_j+\boldsymbol{\lambda}_t\sum_{j=2}^{J+1}w_j\boldsymbol{\mu}_j+\sum_{j=2}^{J+1}w_j\varepsilon_{jt}$$

假设对干预前每个t，以及可观测的协变量Z，存在权重$\mathbf{W}=(w_2^*,...,w_{J+1}^*)$使得：

$$\sum_{j=2}^{J+1}w_j^*Y_{j1}=Y_{11},\quad \sum_{j=2}^{J+1}w_j^*Y_{j2}=Y_{12},\quad \ldots,\quad \sum_{j=2}^{J+1}w_j^*Y_{jT_0}=Y_{1T_0},\quad \text{and}\quad \sum_{j=2}^{J+1}w_j^*\mathbf{Z}_j=\mathbf{Z}_1$$

对于$t\in\{T_0+1,\ldots,T\}$，我们关心的t期的因果效应估计量$\hat{\tau}_{1t}$就可以被表示为：

$$\hat{\tau}_{1t}=Y_{1t}-\sum_{j=2}^{J+1}w_j^*Y_{jt}$$

于是干预后的处理组平均因果效应就是：

$$\hat{\tau}_1=\sum_{t=T_0+1}^{T}\hat{\tau}_{1t}/(T-T_0)$$

然而，上述方程很难同时完全成立，我们对权重W选取的要求是尽量近似满足上式。设$\mathbf{X}_1=(\mathbf{Z}_1',Y_{11}^N,\ldots,Y_{1T_0}^N)'$是干预前第一个实验单元的协变量向量，类似地，$\mathbf{X}_0$为干预前对照组实验单元的协变量矩阵，其中第$j$列对应一个处理组单元$(\mathbf{Z}_j',Y_{j1}^N,\ldots,Y_{jT_0}^N)'$。我们可以选取权重$\mathbf{W}^*$使得：

$$\min\|\mathbf{X}_1-\mathbf{X}_0\mathbf{W}\|_{\mathbf{V}}=\sqrt{(\mathbf{X}_1-\mathbf{X}_0\mathbf{W})'\mathbf{V}(\mathbf{X}_1-\mathbf{X}_0\mathbf{W})}$$

其中V是一个半正定对称矩阵，代表不同协变量的重要性权重。在关于V的选择上，我们可以考虑选择使得在干预前阶段观测结果Y的预测均方误差最小化的正定对角矩阵，从而得到相应的权重$\mathbf{W}=(w_2^*,...,w_{J+1}^*)$。

显著性评估

在得到了具体的因果效应估计后，我们自然希望知道其效果是否显著，此时我们可以考虑 Fisher 精确检验方法计算p值：将对照组个体依次作为处理组，计算其效应值，然后确定这些效应值中有多少比例高于处理组个体的效应值，具体计算步骤：

1. 实验个体作为处理组，使用合成控制法计算效应值$\hat{\tau}_1$；
2. 对照个体逐个作为处理组，分别使用合成控制法并计算效应值$\hat{\tau}_2,\hat{\tau}_3,...,\hat{\tau}_{J+1}$；
3. 计算p值：$P=\frac{1}{J}\sum_{j=2}^{J+1}1\{\hat{\tau}_j>\hat{\tau}_1\}$或者$P=\frac{1}{J+1}\sum_{j=1}^{J+1}1\{\hat{\tau}_j\geq\hat{\tau}_1\}$。

考虑到我们得到的因果效应可能并非完全由干预引起，可能存在一些随机因素，我们需要通过稳健性检验来排除随机因素的影响，此时可以考虑改变干预时间节点进行稳健性检验：即通过提前或延后干预时间，创造一个虚拟干预时间节点，观察在这种情况下得到的平均因果效应，与真实干预时间点的平均因果效应是否存在显著差异。

合成控制法的拓展

近年来，针对前述局限性，众多研究者在 Abadie 和 Gardeazabal（2003）提出的合成控制法基础上进行了改进。我们对这些改进方法进行了简要总结，如下表所示，具体细节可参考原文：

方法	核心原理	适用场景	原文信息
广义合成控制法	基于线性交互固定效应模型，利用对照组信息为每个处理单元推断反事实结果，能够处理多个处理单元和可变干预期的情况。	存在多个处理单位和可变干预期	Xu, Yiqing. "Generalized synthetic control method: Causal inference with interactive fixed effects models." *Political Analysis* 25.1 (2017): 57-76.
稳健合成控制法	通过奇异值阈值化来去噪协变量数据矩阵，仅保留信息较为有效的部分，具有稳健性，可在干预前数据存在缺失值的情况下使用。	存在干预前的缺失数据	Amjad, Muhammad, Devavrat Shah, and Dennis Shen. "Robust synthetic control." *Journal of Machine Learning Research* 19.22 (2018): 1-51.
增强合成控制法	类似于对不精确匹配进行偏差校正，通过模型估计因处理前拟合不完善而产生的偏差，并调整原始合成控制法的估计以消除这些偏差。	干预前的拟合效果较差	Ben-Michael, Eli, Avi Feller, and Jesse Rothstein. "The augmented synthetic control method." *Journal of the American Statistical Association* 116.536 (2021): 1789-1803.
合成控制实验设计	通过求解优化问题，指导如何选择处理组和对照组，并选择合适的权重，使加权后的处理组与对照组能够代表总体情况。	需要选择合适的处理组和对照组，并能够代表总体情况	Abadie, Alberto, and Jinglong Zhao. "Synthetic controls for experimental design." *arXiv preprint arXiv:2108.02196* (2021).

评估模型选择

我们介绍了很多合成控制法的拓展，在面对复杂场景时，我们应该如何选择合适的模型呢？一方面，可以结合具体业务和经验进行判断，例如，当存在多个实验单元时，可以考虑使用广义合成控制法；当存在较多干预前缺失数据时，可以考虑使用稳健合成控制法。另一方面，也可以采用数据驱动（Data-Driven）的方式，利用实验前数据进行模型评估：通过不同模型预测实验前几周的 AA 数据，如果预测值与真实值接近（以 MAPE 衡量，即 Mean Absolute Percentage Error，平均绝对百分比误差），则说明模型的预测较为准确，实验期间预测值的参考价值较高。此外，还可以计算实验前 AA 结果的 p 值，p 值越大，说明该模型的 AA 结果越不显著，因果效应估计值更接近 0，这也意味着在实验期间预测值的参考价值更高。

6.1.3 实验案例

案例背景：美团履约运营团队设计了一种新的运营策略，希望验证该策略能否实现数量和效率的可控性，使得运力和用户需求更匹配，从而提高骑手和用户的体验。

评估难点：受限于多方面的业务约束情况，不适合采用分组实验的方式进行验证。新模式需要通过长期运营来观察和评估用户的接受度，无法实现每日切换，因此也不适合采用时间轮转的实验设计。此外，也难以找到业务特征高度相似的单一城市，作为实施新策略城市的对照组。

解决方法：考虑“全城灰度”策略，即在整个城市范围内实施新策略一段时间（如一个月），然后利用合成控制法，从一些还没有实施该新策略的城市中拟合一个虚拟的对照组进行评估。

评估指标：**

评估周期：**

评估结果：

评估指标	相对增幅	显著性	拟合结果图
**	**pp	p值=**	**

6.2 匹配方法

6.2.1 概述

上文提到，在美团履约和外卖的实验中，部分场景由于法律约束以及实际操作成本等诸多限制，无法开展控制实验。例如，在“评估购买优惠券对订单量增量效果”的研究中，我们无法控制用户是否实际购买优惠券。因此，若要评估整体人群中购买优惠券对订单量的提升效果，随机对照实验并不适用。

通常，评估购买优惠券对订单量的影响最直接的方法，是比较“购买优惠券”与“不购买优惠券”用户的订单量差异。然而，**现实中多种因素都会影响购买优惠券的行为和订单量，购买优惠券的用户与不购买优惠券的用户在某些协变量特征上也往往存在天然差异**，直接比较这两类人群的订单量差异会存在选择性偏差问题。

为此，我们可以采用匹配方法，通过匹配购买优惠券与不购买优惠券用户的协变量特征来控制这些干扰因素，减少因选择偏差导致的估计误差，从而更准确地估计实验效果。

基本思想

匹配是因果推断中常用的一种方法，其核心思想是通过平衡处理组和对照组之间的协变量分布，从而消除混杂因素的影响。具体而言，在多维协变量空间中，匹配方法尽量模拟随机分配的情境，为每个处理组个体找到一个或多个相似的对照组个体，作为其反事实结果，从而减少样本间协变量（非处理因素）差异对效果评估的干扰，其基本流程如下图 6-2 所示：

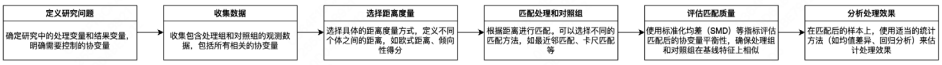


图 6-2：匹配流程

适用场景与优缺点

匹配方法在观察性研究中被广泛应用，尤其适用于以下场景：

- **无法进行控制试验：**出于法律约束以及实际操作成本等原因，无法实施控制试验时，匹配方法成为因果推断的重要工具。
- **处理组与对照组存在相似个体：**匹配方法适用于处理组和对照组中存在相似个体的情况，通过确保这两组在协变量上的分布尽量一致，从而减少因组间差异带来的偏差。
- **观测的协变量特征较为全面：**当评估中涉及多个协变量且需要在这些协变量上达到平衡时，匹配方法能够有效控制混杂因素，提高因果效应估计的准确性。

在应用匹配方法进行因果效应分析时，我们需要详细了解其优势和局限性，从而确保评估的准确性。首先，匹配方法具有以下优点：

- **减少选择偏差：**匹配方法通过平衡处理组和对照组的协变量分布，显著减少了由于非随机分配导致的选择偏差，从而提高因果效应估计的准确性。
- **易于理解与实施：**相较于其他复杂的因果推断方法，匹配方法直观且易于理解，解释性强。我们可以通过匹配后直接比较处理组和对照组的结果，步骤清晰。

- **灵活性高：**匹配方法可以与多种统计模型和技术结合使用，如不同的倾向得分模型、距离度量方法等，适应不同研究需求和数据特点。

不过，匹配方法也存在以下局限性：

- **数据需求较高：**为了有效匹配个体，处理组和对照组需要有足够的重叠区域（Overlap or Common Support），即处理组和对照组中需要存在相似个体，但在某些场景中，可能并不满足该条件，这会限制匹配的有效性。
- **无法控制未观测到的混杂因素：**匹配方法仅能控制已观测到的协变量，对于未被包含在匹配过程中的潜在混杂变量，匹配方法可能无法完全消除选择偏差，这可能导致因果效应估计的偏差。

6.2.2 原理

由上述匹配的基本流程可知，匹配主要包括：选择协变量特征、定义距离度量、选择匹配方法等步骤。在本节，我们将详细介绍这些步骤和一些注意点。

基本假定

匹配方法灵活且易于实施，但其评估结果的有效性会依赖于以下两个假定条件：

- **条件独立假设**（Conditional Independence Assumption）：在给定观测协变量的条件下，处理的分配与潜在结果独立，其数学表达如下：

$$(Y_i(1), Y_i(0)) \perp T_i \mid X_i$$

其中 $Y_i(1)$ 和 $Y_i(0)$ 分别表示个体 i 在处理组和对照组中的观测结果，T_i 表示处理变量，X_i 表示协变量。在该假设下，我们只要控制了可观测变量 X，就控制了所有会影响处理变量 T 和观测结果 Y 的混杂因子，处理组与对照组之间就不存在未观测的差异。

重叠性假设（Overlap 或 Common Support）：在所有协变量的取值下，对应个体分配到处理组和对照组概率都大于 0，即 $0 < P(T=1 \mid X) < 1$。这意味着，对于每一个协变量的组合，既有接受处理的个体，也有不接受处理的个体。该假设确保了每个处理组个体都有相似的对照组个体可供匹配，保证了因果效应估计的有效性。

协变量特征选择

在确定匹配过程中应选择哪些协变量时，关键概念是**条件独立假设**（Conditional Independence Assumption）。匹配方法以及大多数观察性研究方法都依赖于该假设，该假设认为在已观测协变量的条件下，处理组与对照组之间不存在未观测的差异。为了满足可忽略性假设，重要的是在匹配过程中包含**所有已知与处理分配和结果相关的变量**。

通常，使用相对较少的**便利预测变量**（Predictors of Convenience）的匹配方法表现较差。在使用倾向得分匹配（Propensity Score Matching, PSM，下文将详细介绍）时，包含与处理分配无关的变量几乎没有成本，因为它们对倾向得分模型的影响极小。虽然包含与结果无关的变量可能会略微增加方差，然而，排除潜在的重要混杂变量往往会导致较大的偏差。因此，我们在选择协变量特征时，**应采取宽松的态度，尽可能包含可能与处理分配和结果相关的变量，以提高因果效应估计的准确性**。

此外，匹配过程中不应包含那些可能受到处理影响的变量，当协变量、处理变量和结果变量同时收集时，这一点尤为重要。如果确实需要控制受处理影响的变量，应该在匹配之后，通过回归调整或其他适当的统计方法在分析模型中进行控制。

距离度量

在匹配时，我们需要定义个体之间的距离，用来衡量两个个体的相似性。定义个体i和个体j之间的距离D_{ij}，有以下几种方法，我们总结如下表所示：

距离定义	计算公式D_{ij}	适用场景	优点	缺点
精确距离	$D_{ij}=\begin{cases}0 & \text{if } X_i=X_j\\ \infty & \text{if } X_i\neq X_j\end{cases}$	协变量较少，且均为离散型变量特征	可以使得匹配对个体的协变量特征完全一致	协变量较多、存在连续型变量时，无法匹配到个体
欧式距离	$D_{ij}=\|\|X_i-X_j\|\|_2$	协变量独立、简单数据结构	易于理解与计算；适用广泛；基础性强	尺度敏感；忽略协变量关系
马氏距离	$D_{ij}=(X_i-X_j)\Sigma^{-1}(X_i-X_j)$	协变量较少，且协变量服从多元正态分布	考虑协变量相关性；尺度不敏感；适用于复杂数据	计算复杂；数据需求高
倾向性得分距离	$D_{ij}=\|e(X_i)-e(X_j)\|$	高维协变量、复杂混杂、估计ATT	降维优势；处理高维数据；灵活性高；降低偏差	模型依赖；复杂性高；与变量选择相关

在上述距离定义中，除了倾向得分距离之外，其他距离类型都较为常见且易于理解。接下来，我们将对倾向得分距离进行详细介绍。首先我们先简单介绍倾向性得分的定义：倾向性得分（Propensity Score）是指在给定协变量的条件下，个体接受处理的概率。

具体而言，对于个体i及其协变量向量X_i，倾向得分$e(X_i)$定义为：$e(X_i)=P(T_i=1|X_i)$，其中，T_i是处理指示变量，$T_i=1$表示个体i接受处理，$T_i=0$表示未接受处理。为了更好地理解倾向性得分的定义，我们考虑上文评估整体人群中购买优惠券对订单量的提升效果的例子，假如协变量X只有 AB 两种选择，取值为 1 表示 A 类用户，取值为 0 表示 B 类用户，购买优惠券表示用户接受处理，当用户协变量$X_i=1$时，倾向性得分$e(X_i=1)=P(T_i=1|X_i=1)$表示 A 类用户购买优惠券的概率，当$X_i=0$时，倾向性得分$e(X_i=0)=P(T_i=1|X_i=0)$表示 B 类用户购买优惠券的概率。为什么我们可以用倾向性得分定义距离并进行匹配呢？这是因为 Rubin[3] 证得：

在条件独立假设和重叠性假设下，有：

$$(Y_i(1),Y_i(0))\perp T_i\,|\,e(X_i)$$

该结论表明，倾向性得分$e(X_i)$包含了协变量特征X_i的所有信息，只要我们控制了倾向性得分$e(X_i)$，那么处理组与对照组之间不存在未观测的差异。由于我们并不知道每个个体的真实倾向性得分，我们可以通过二元响应模型（如 Logistic 回归或 Probit 回归）进行估计。

具体而言，我们使用处理指示变量T作为因变量，协变量X作为自变量，拟合 Logistic 回归模型：

$$P(T_i=1|X_i)=\frac{e^{\beta_0+\beta_1X_{i1}+\cdots+\beta_kX_{ik}}}{1+e^{\beta_0+\beta_1X_{i1}+\cdots+\beta_kX_{ik}}}$$

然后基于该拟合模型预测计算每个个体的倾向得分估计值。

匹配方法

在匹配方法中，最常见且最容易实施和理解的方法之一是**最近邻匹配**（Nearest

Neighbor Matching）。最近邻匹配几乎总是能估计出处理组的平均处理效应（ATT），因为它将对照组个体匹配到处理组，并丢弃未被选为匹配的对照组个体。在其最简单的形式中，**1:1 最近邻匹配**为每个处理组个体选择距离最近的一个对照组个体，这也是我们最常用的形式。在使用匹配方法时，经常会存在一些细微问题，我们简单总结如下，具体细节可参考 Stuart（2010）[4] 的综述论文：

- **一对一匹配与一对多匹配：**最常见的形式是使用一对一匹配，但该方式丢弃的对照组个体可能会比较多，检验功效会降低，此时可以考虑一对多匹配，但对应地，其计算复杂度会增加，且匹配效果会依赖于超参数的调整。
- **有放回匹配与无放回匹配：**我们一般使用有放回匹配，但一些研究者更倾向于无放回匹配。当对照组的样本量较大时，这两种方法在最终结果上通常不会有太大差异。有放回匹配在计算上更为简便，而无放回匹配则涉及计算密集的离散优化过程。有放回匹配通常能够获得更高质量的匹配，但由于需重复使用相同的样本，可能会引入依赖性。相比之下，无放回匹配的优势在于确保匹配样本的独立性，并简化后续的数据分析过程。
- **匹配限制：**在匹配方法中，一个常见的担忧是缺乏限制可能导致不良匹配。例如，某处理组个体的倾向得分（Propensity Score）与对照组中任何个体的相似度不足，无法找到合适的匹配对。为避免此类问题，可以实施卡尺（Caliper），即仅选择匹配距离在预设范围内的对照组个体。虽然这可能导致部分处理组个体无法找到匹配对，增加因果效应解释的难度，但有助于确保匹配质量，减少估计偏差。
- **匹配方法的选择：**目前有各种各样的匹配方法可供选择，但相关的指导却相对较少。迄今为止，学术界主要的建议是选择能够实现最佳平衡的方法，例如 Ho 等人（2007）[5] 的研究。然而，定义“最佳平衡”是复杂的，因为这涉及在多个协变量之间进行权衡。选择匹配方法的可能方式包括：(1) 在最多协变量上实现最小标准化均差的方法；(2) 最小化少数特别具有预测性协变量的标准化均差的方法；(3) 产生最少“大”标准化均差（大于 0.25）的方法等。这些方法各有侧重，我们可能需要根据具体的研究需求和数据特点选择最合适的匹配方法。

评估与检验

在匹配完成后，需要评估匹配的质量，确保处理组和控制组在协变量上的平衡。常用的方法包括：标准化均差（SMD）和分布图。

- **标准化均差（SMD）：** 评估匹配后协变量的平衡性，确保处理组和对照组在基线特征上相似。标准化均差（Standardized Mean Difference, SMD）是用于衡量两组之间均值差异的标准化效应量。SMD 的公式如下：

$$\text{SMD} = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}}$$

  其中$\bar{X}_1$是处理组的均值，$\bar{X}_2$是对照组的均值，s_1和s_2分别是处理组和对照组的标准差，n_1, n_2分别为处理组和对照组的样本量。SMD 是无单位的，因此可以用于不同尺度的变量之间的比较，通常用于评估组间平衡性。在倾向评分匹配（PSM）中，SMD 小于 0.1 或 0.2，通常被认为是组间平衡良好的标志。

- **分布图：** 绘制协变量的分布图或倾向得分的分布图，检查匹配前后的变化。

在匹配完成并验证平衡性后，可以估计处理效应。常见的处理效应估计方法包括：

- **平均处理效应**（ATE）：估计总体的处理效应。
- **处理组平均处理效应**（ATT）：估计处理组的平均处理效应。
- **控制组平均处理效应**（ATC）：估计控制组的平均处理效应。

处理效应的估计通常通过比较匹配后的处理组和控制组的结果变量均值来进行。

匹配的拓展

在上文中，我们主要介绍了最常用且最经典的匹配方法。然而，在处理一些复杂情形时，这些方法可能无法充分满足评估需求，因此需要对其进行扩展。我们对此进行了简要总结，具体细节可参考相关文献。

方差估计：Badie 和 Imbens（2008）[6] 首次表明，仅通过对原始数据进行重抽样的简单自助法（Bootstrap）无法有效估计匹配估计量的方差，但他们提出的方差估计方法实施起来并不容易。Otsu 和 Rai（2017）建议对估计量在线性展开中进行 Bootstrap，Otsu 和 Rai（2017）[7] 的 Bootstrap 本质上产生了方差估计量。得到方差估计后，便可以计算p值。

距离组合：在某些场景下，我们希望匹配的个体在某些关键协变量特征上（如身份、归属城市）保持完全一致，然后再在这些子组内进行匹配，此时我们可以考虑将上文介绍的距离度量进行组合。例如，我们可以考虑类似粗糙精确匹配（Coarsened Exact Matching，CEM）的距离：

$$D_{ij}=\begin{cases}|e(X_i)-e(X_j)| & \text{if } X_{ik}=X_{jk}\\ \infty & \text{if } X_{ik}\neq X_{jk}\end{cases}$$

其中X_{ik}, X_{jk}表示个体i与个体j的关键协变量。此外，我们也可以考虑 Rubin 和 Thomas（2000）[8] 提出的结合马氏距离和倾向性得分卡尺的距离：

$$D_{ij}=\begin{cases}(X_i-X_j)'\Sigma^{-1}(X_i-X_j) & \text{if } |\text{logit}(e(X_i))-\text{logit}(e(X_j))|\leq c\\ \infty & \text{if } |\text{logit}(e(X_i))-\text{logit}(e(X_j))|>c\end{cases}$$

存在多个处理组：上文我们讨论的都是一个处理组和一个对照组的情形，但是，在很多实际场景下，往往会面临多个处理组的情况，此时往往会更复杂。在面对多个处理组时，我们可以考虑广义倾向性得分（Generalized Propensity Score），利用多项逻辑回归模型（Multinomial Logistic Regression Model）预测每个个体的广义倾向性得分，再利用向量匹配方法（Vector Matching，VM）进行匹配，具体细节可参考 Scotina 和 Gutman(2019)[9] 的工作。

共同支撑问题：匹配方法中普遍存在共同支持（Common Support）的问题。迄今为止，我们假设两组的倾向得分分布具有明显重叠，但在某些情况下，分布可能不完全重叠。例如，许多对照组个体与处理组成员差异较大，不适合作为估计平均处理效应（ATT）的比较对象。使用卡尺（caliper）的最近邻匹配方法仅匹配位于或接近共同支

持区域的个体，而子分类（subclassification）和加权（weighting）方法则通常使用所有个体，无论分布是否重叠，具体细节可参考的 Dehejia 和 Wahba(1999)[10] 的工作。

协变量缺失问题：大多数关于匹配和倾向性得分的文献都假设协变量是完全观测的，但实际上大多数研究至少存在一些缺失数据。一种可能性是使用广义提升模型（Generalized Boosted Models）来估计倾向得分，因为它们不需要完全观测的协变量。另一种推荐的方法是进行简单的单一插补（Single Imputation）来填补缺失的协变量，并在倾向得分模型中包含缺失数据指示变量，具体细节可参考 Greenland 和 Finkle(1995)[11] 的工作。

6.2.3 实际案例

案例背景：美团神会员是美团推出的综合权益卡，用户可通过免费领取或者支付一个很低的价格成为“美团神会员”。用户成为神会员用户，可以享受到平台的各种优惠权益。神会员项目中售卖的无门槛券包称为省钱包，目前用户可以通过在美团神会员 Tab 页直接购买。业务方需要对用户在不同行业中购买省钱包后的下单行为变化进行定量分析，以评估用户购买省钱包对业务的影响。

评估难点：实验观察的行为（是否购买券包）不满足随机对照条件，无法进行随机 AB 实验评估效果。由于业务特性，影响用户下单行为的协变量较多，需要考虑如何进行匹配，能够减少选择偏差。

解决方法：采用倾向分匹配（PSM）进行观察性研究，以计算策略效果，具体流程如下：

1. 圈选购买省钱包的用户作为实验组；
2. 圈选未购买省钱包的用户作为候选的对照组；
3. 计算用户特征作为倾向分计算的协变量，包含用户历史交易相关数据、访问相关特征、用户分层等，训练倾向分模型；
4. 使用可放回的抽样，根据倾向分得分，从候选的对照组中为实验组的用户进

行匹配，得到对照组；

5. 计算实验组和对照组的目标指标，评估实验的效果。

评估指标： **

评估周期： **

评估结果： **

评估指标	实验组	对照组	相对增幅	匹配效果	显著性
**	**	**	**%	SMD=**%	P值=**

6.3 Causal Impact

6.3.1 概述

在美团履约和外卖业务中，部分策略由于无法进行随机实验，同时为了避免影响用户体验，需要在城市粒度上进行实施和评估。这些策略包括线下广告投放、冬夏季城市战和时段场景营销等。然而，常用的评估方法在处理这些局部全量策略效果时存在一定的局限性：首先，单重差分法假设功能或策略是唯一的影响因素，但现实中市场环境复杂，影响因素多样，使得这一假设难以成立。其次，倾向分匹配法（PSM）虽然在特征选择和匹配质量上有其优势，但难以消除未观测的混杂因素。此外，合成控制方法（SCM）要求协变量及目标变量均相似的对照组，这在实际应用中难以获得。最后，双重差分法（DID）假设干预组和对照组在没有干预的情况下会有相同的趋势，这一假设在实践中较难成立。

为了解决这些问题，Causal Impact 方法 [12] 应运而生。该方法基于贝叶斯结构时间序列（BSTS）模型，通过构建“虚拟对照组”来更准确地评估干预效果。Causal Impact 能够有效捕捉时间序列中的长期趋势和周期性变化，从而提供稳健的因果效

应估计，为企业提供可靠的决策支持。

基本思想

Causal Impact 方法的基本思想是通过贝叶斯结构时间序列（Bayesian Structural Time Series, BSTS）模型来评估干预措施的因果效应。其核心在于构建一个“虚拟对照组”，用于预测在没有干预措施情况下目标变量的可能表现。然后，将该预测值与实验组的真实值进行对比，从而评估策略效果。

以城市粒度实验为例，具体步骤如下图 6-3：

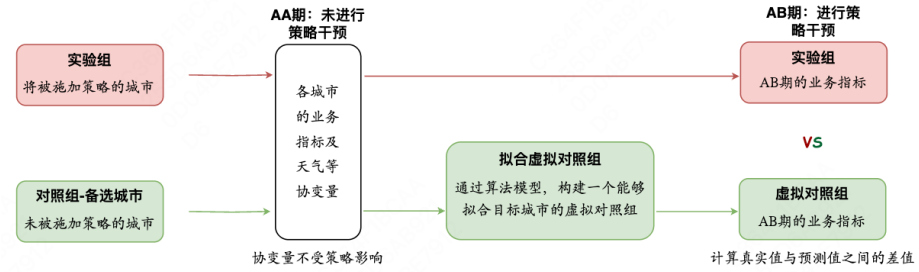


图 6-3：Causal Impact

适用场景与优缺点

Causal Impact 方法的有效性依赖于若干重要的前提条件和假设，这些条件共同构成了确保因果效应估计准确性和可靠性的基础。首先，需要有可用且平稳的时序数据，即足够的历史数据，涵盖完整的周期性模式，且时间序列中的趋势和季节性成分是平稳的。其次，须有相似、独立、稳定的对照组，与目标市场行为模式相似且未受干预影响，并在实验期间保持稳定。此外，时间序列数据需符合状态空间模型的基本假设，包括线性关系、正态分布误差和马尔可夫性质。模型中应包含所有重要的控制变量，确保没有遗漏关键的影响因素。最后，数据中应没有显著的异常值或极端情况，或已妥善处理这些问题。

在应用 Causal Impact 方法进行因果效应分析时，了解其优势和局限性对于确保分

析的准确性和可靠性至关重要。Causal Impact 方法结合了贝叶斯结构时间序列模型和反事实预测技术，能够在复杂的时间序列数据中提供稳健的因果效应估计，具体而言，其优势在于：

- **灵活的时间序列建模：**适用于复杂时间依赖结构的数据集，能够捕捉数据中的趋势、季节性和异常值。
- **无需随机对照试验：**能够在没有随机实验的情况下估计因果效应，通过构建“虚拟对照组”来进行因果推断。
- **不确定性量化：**提供完整的后验分布，能够量化不确定性，从而提供更为全面的因果效应评估。
- **动态适应性：**支持动态回归系数，能够根据时间变化动态调整模型，增强模型的灵活性和适应性。

尽管 Causal Impact 方法具有显著的优势，但在实际应用中也需注意其局限性，以确保分析结果的可靠性。这些局限性包括：

- **依赖高质量对照组：**方法的准确性高度依赖于对照组的选择。如果对照组选择不当，可能导致估计偏差。
- **假设严格：**方法假设目标市场和对照组的行为模式相似，且对照组不受干预影响。这一假设在实际应用中可能不完全成立。
- **难以处理复杂因果关系：**对于复杂的多因素交互作用或长期滞后效应，方法可能不够准确。
- **需要足够长的历史数据：**需要足够长的历史数据来训练模型，以捕捉数据中的长期趋势和季节性变化。

通过明确这些优势和局限性，可以更好地应用 Causal Impact 方法进行因果效应分析，从而确保分析的准确性和可靠性。

6.3.2 原理

在本节，我们将详细介绍 Causal impact 的基本原理。

模型设定

Causal Impact 通过采用贝叶斯结构时间序列（Bayesian Structural Time Series，BSTS）模型，结合状态空间模型（State-Space Models）与贝叶斯推断（Bayesian Inference）方法来构建反事实预测模型，从而估计在没有干预措施的情况下结果变量的预期表现。考虑一个常规的 BSTS 模型：

$$\begin{cases} y_t = Z_t^T z_t + \epsilon_t \text{ with } \epsilon_t \sim \mathcal{N}(0, \sigma_t^2), & \text{观测方程（1）} \\ z_{t+1} = T_t z_t + R_t \eta_t \text{ with } \eta_t \sim \mathcal{N}(0, Q_t), & \text{状态方程（2）} \end{cases}$$

其中Z_t是设计矩阵，z_t是潜在状态向量，$\eta_t \sim N(0, Q_t)$表示状态噪声，T_t是状态转移矩阵，R_t是控制噪声影响的矩阵，通过改变矩阵Z、T、R和Q，可以为时间序列建模几个不同的行为（包括著名的$ARMA$或$ARIMA$）。

在很多情况下，我们对于将要评估的时间序并没有模型的先验认知，此时我们可以构建一个默认 Local Level 的模型，并在状态方程中加入协变量X，此时y_t可表示为：

$$\begin{cases} y_t = \mu_t + \gamma_t + \beta X_t + \epsilon_t \\ \mu_{t+1} = \mu_t + \delta_t + \eta_{\mu,t} \\ \delta_{t+1} = \delta_t + \eta_{\delta,t} \\ \gamma_{t+1} = -\sum_{s=0}^{S-2} \gamma_{t-s} + \eta_{\gamma,t} \end{cases}$$

在上述模型中，各项含义如下：

- μ_t 代表一个自回归的过程，任何给定的时间点首先由随机游走分量建模，反应的是局部水平（Local Level）；
- 分量βX_t是协变量的线性组合（目前考虑静态回归系数，如有需要可考虑随t变化的动态的$\beta_{j,t+1} = \beta_{j,t} + \eta_{\beta,j,t}$，例如对照城市作为协变量，对照城市与实验城

市关系会发生变化，当相对稳定时考虑使用静态协变量）；

- δ_t 是 μ 在 t 和 $t+1$ 期之间期望的增量，同样是一个自回归过程，反应的是局部趋势（Local Trend）；
- γ_t 表示季节效应（如果不考虑季节性则无该项），S 表示季节的周期数；
- $\eta_{\mu,t} \sim N(0,\sigma_\mu^2)$、$\eta_{\delta,t} \sim N(0,\sigma_\delta^2)$ 以及 $\eta_{\gamma,t} \sim N(0,\sigma_\gamma^2)$ 是噪声项。

贝叶斯推断

在 Causal Impact 中，对于上述介绍的 BSTS 模型，我们通常会使用贝叶斯后验推断来估计反事实预测值，即：

$$p(\tilde{y}_{n+1},\ldots,\tilde{y}_m \mid y_1,\ldots,y_n,x_1,\ldots,x_m)$$

具体步骤如下：

先验选择：在贝叶斯模型中，我们需要对各参数设置合理的先验分布。对于方差参数 σ^2，可以采用共轭先验伽玛分布：$\frac{1}{\sigma^2} \sim G\left(\frac{\nu}{2},\frac{s^2}{2}\right)$，对于协变量系数 β，我们可以考虑 Spike-and-Slab 先验 [13]，该先验可以帮我们自动选择重要的变量，并剔除那些不重要的变量。

后验推断：一般情况下，因为模型的复杂性，我们无法直接得到反事实预测值后验分布的显示表达式。因此，我们可以考虑利用 MCMC（Markov Chain Monte Carlo）方法，通过构建一个马尔可夫链（Markov Chain），使得该链的极限分布（平稳分布）为目标后验分布，从而实现从后验分布中的有效采样进行后验推断。以上面的 Local Level 模型为例，一个完整的 Causal Impact 过程可见图 6-4：

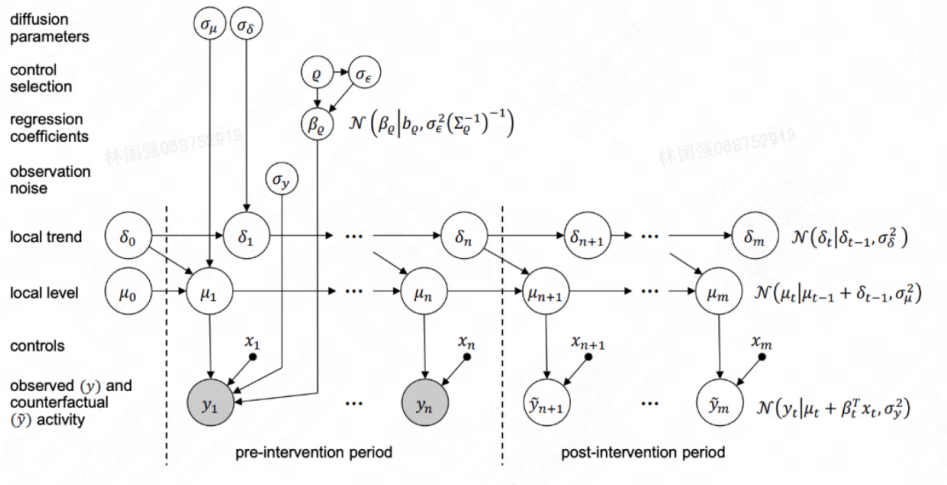


图 6-4

因果效应评估

通过贝叶斯后验推断，我们可以得到反事实预测（Counterfactual Prediction）结果，即：

$$p(\tilde{y}_{n+1},\ldots,\tilde{y}_m \mid y_1,\ldots,y_n,x_1,\ldots,x_m)$$

此时我们可以计算每一期的因果效应估计 $\phi_t := y_t - \tilde{y}_t$，也可以得到累积效应（Cumulative Effect）：$\sum_{t'=n+1}^{t} \phi_{t'}^{(\tau)}$，以及运行平均效应（Running Average Effect）：$\sum_{t'=n+1}^{t} \phi_{t'}^{(\tau)} / (t-n)$。最后，通过验证因果效应的 95% 后验置信区间是否包含 0 来评估显著性。

6.3.3 实际案例

为了更直观地展示 Causal Impact 方法的运行机制，这里举一个外卖一体化营销的例子。

背景介绍：以往在城市维度进行营销时，业务主要依赖站内补贴资源来推动城市交易

额的增长，而站内外、线上线下资源的协同效应相对较弱。前几年，美团外卖推出了一种全新的一体化营销模式，通过组织统筹和综合效应，促进站外广告营销、一线运营的协同作用。当前，美团外卖在一体化营销城市战中投入了大量人力和物力，这样的投入是否值得？为此，我们需要构建一个评估方法来衡量一体化营销策略对业务的影响。

评估难点：由于涉及站外和站内、线上和线下的多策略组合，评估面临一些挑战，无法通过 A/B 测试和倾向评分匹配（PSM）进行有效评估。同时，不同城市的天气等外部因素差异显著，难以找到满足平行趋势的对照组城市，这也使得双重差分法不适用。此外，为避免影响用户体验，策略不能频繁变更，因此时间片轮转也不可行。

解决方法：考虑在全城范围内实施站内站外、线上线下的组合策略，可以利用 Causal Impact 方法进行评估。具体做法是，从暂未实施该策略的城市中选择一些作为候选城市，并结合天气等外生变量，拟合出一个虚拟的“对照城市”进行评估。

评估指标：**

评估周期：**

评估结果：

评估指标	相对增幅	拟合效果	显著性
**	**%	MAPE=**%	P值=**

6.4 展望与拓展

在上文中，我们主要介绍了合成控制法、匹配方法以及 Causal Impact 等方法。此外，还有许多广泛应用于观察性研究的方法值得进一步探讨，尤其是在上述方法不满足评估需求时，可以考虑使用以下方法：

- **逆概率加权**（Inverse Probability Weighting, IPW）：通过为每个样本分配权

重来调整样本分布，以有效控制混杂变量的影响，从而更准确地估计处理效果。

- **双重稳健估计**（Doubly Robust Estimation）：结合倾向得分模型和结果模型的优点，即使其中一个模型不完全正确，依然能够提供一致的因果效应估计。
- **工具变量法**（Instrumental Variable, IV）：通过引入一个工具变量（IV），该变量与处理变量相关但与结果变量无关（仅通过处理变量影响结果），从而解决内生性问题，准确估计因果效应。工具变量法特别适用于处理变量与误差项相关的情况，例如遗漏变量偏差或测量误差。
- **双重机器学习**（Double Machine Learning, DML）：结合了机器学习与因果推断方法，旨在高维数据环境下准确估计因果效应。该方法通过使用机器学习模型分别估计处理变量和结果变量与协变量之间的关系，并通过残差化（residualization）与交叉验证（cross-fitting）技术，有效控制潜在的混杂因素，减少模型误差带来的偏差。

这些方法各具特色，为我们提供了多样化的评估工具。如果能够合理选择并使用这些方法，我们可以在复杂的业务环境中更好地进行效果评估，得到科学的评估结果，进而为决策提供科学依据。

参考资料

[1] 业务特征：各运力线的承托比、骑手规模、总完单量、拼好饭单占比、跑腿单占比、推订单完成率等等。
[2] Abadie 和 Gardeazabal（2003）: Abadie, A., & Gardeazabal, J. (2003). The economic costs of conflict: A case study of the Basque Country. American economic review, 93(1), 113-132.
[3] Rubin 证得: Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. Biometrika, 70:41-55.
[4] Stuart（2010）: Stuart, E. A. (2010). Matching methods for causal inference: A review and a look forward. Statistical science: a review journal of the Institute of Mathematical Statistics, 25(1), 1.
[5] Ho 等人（2007）: Ho DE, Imai K, King G, Stuart EA. Matching as nonparametric preprocessing for reducing model dependence in parametric causal inference. Political Analysis 2007;15(3):199-236.

[6] Abadie 和 Imbens（2008）: Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. Econometrica, 76:1537–1557.
[7] Otsu 和 Rai（2017）: Otsu, T. and Rai, Y. (2017). Bootstrap inference of matching estimators for average treatment effects. Journal of the American Statistical Association,112:1720–1732.
[8] Rubin 和 Thomas（2000）: Rubin, Donald B., and Neal Thomas. Combining propensity score matching with additional adjustments for prognostic covariates."Journal of the American Statistical Association 95.450 (2000): 573–585.
[9] Anthony 和 Gutman(2019)：Scotina, Anthony D., and Roee Gutman. Matching algorithms for causal inference with multiple treatments. Statistics in medicine 38.17 (2019): 3139–3167.
[10] Dehejia 和 Wahba(1999)：Dehejia, Rajeev H., and Sadek Wahba. Causal effects in nonexperimental studies: Reevaluating the evaluation of training programs. Journal of the American statistical Association 94.448 (1999): 1053–1062.
[11]Greenland 和 Finkle(1995)：Greenland S, Finkle WD. A critical look at methods for handling missing covariates in epidemiologic regression analyses. American Journal of Epidemiology 1995;142:1255 - 1264.
[12] Causal Impact 方法: Brodersen, K. H., Gallusser, F., Koehler, J., Remy, N., & Scott, S. L. (2015). Inferring causal impact using Bayesian structural time-series models.
[13] Spike-and-Slab 先验：结合了“尖峰”（Spike）和“平板”（Slab）两个部分，尖峰（Spike）部分是一个集中在零附近的分布，表示某个参数可能为零或接近零，反映了变量不被选择或对模型贡献很小的情况；平板（Slab）部分是一个较为宽松的分布，允许参数有较大的值，表示该变量可能对模型有显著贡献。

第七章：高阶实验工具

在前面的几个章节中，我们已经详细讨论了许多实验方法的适用场景以及实验设计与评估流程，然而在实际操作中，实验者仍会面临一些常见的困难和疑问。例如由于业务约束，实验者常常无法在单个城市选取足够流量进行实验，即单次实验的样本量难以达到检测出预期提升的功效，从而无法得到显著的实验结论。为此，实验者可能会在多城进行实验或者在同一个城市多次进行实验，以期积累样本量使得能够检测出显著的实验效果。此时，实验者需寻求新手段，科学地整合多次实验的结果，以最终确定策略的有效性。

此外，在一些在线实验中，实验者可能需要考察十几甚至几十个指标的变化情况，或者分多个实验组以同时考察多个策略的效果，甚至实验者有时还会倾向于在完整实验周期结束之前监控实验结果，提前查看显著性。这些操作本质上都涉及到多重的假设检验，当实验指标、组别数量增多和查看结果的频率提高时，假设检验的次数也随之增加。虽然单次假设检验能将第一类错误率控制在 5%，但在多次假设检验中犯第一类错误（无效策略错误判为有效）的概率却不再是 5%，而是可能远大于这个概率。

即如果继续采用原始逻辑进行显著性判断，往往会发现更容易出现一些误判策略显著的结果。因此，如何在多重比较的情况下防止假阳性带来的错误判断，也是实验者需关注的问题。针对单次实验功效不足、假阳性、策略调优等实验中面临的问题，我们也针对性的探索并建设了一些高阶实验工具予以解决。

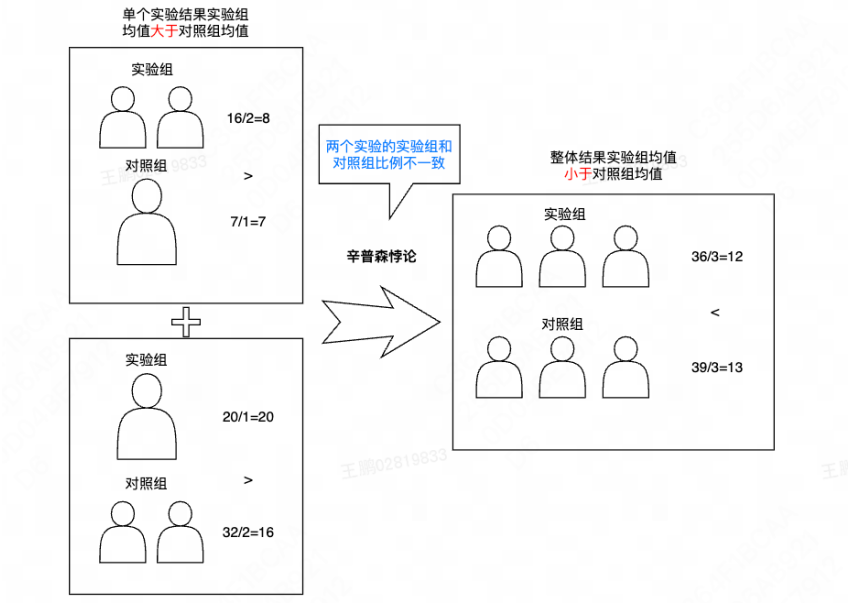
7.1 统合分析

7.1.1 统合分析概述

在实际业务中，可能在不同城市等（可不同时）开展同一个实验，或者在同一城市进行正交随机化后的重复实验，亦或是两者皆有。这些实验相互之间可以认为是独立

的。统合分析旨在综合考虑多个考察同一策略的独立实验，对这些实验的实验结果进行统合分析，从而给出对于这些实验整体效果的评估结果。一种业务上常用的整体评估方式是打包分析，即将多次实验的实验单位数据放在一起进行计算。

但这会遇到两个问题：(1) 其一是辛普森悖论，当不同实验各组的分组比例不同时，可能出现整体结果与单次实验结果截然相反的情况。这通常与业务的常规认知相悖。(2) 其二是对于同一城市进行不同周期多次正交打散的实验时，可能会存在实验单位在多次实验中的实验组和对照组中都出现。如果实验可以认为是独立的（即不存在前序实验对后序实验的结果造成影响），这种情况下的同一单位在两次实验中应当被当作两个独立的单位进行处理，这在数据分析时需要注意。



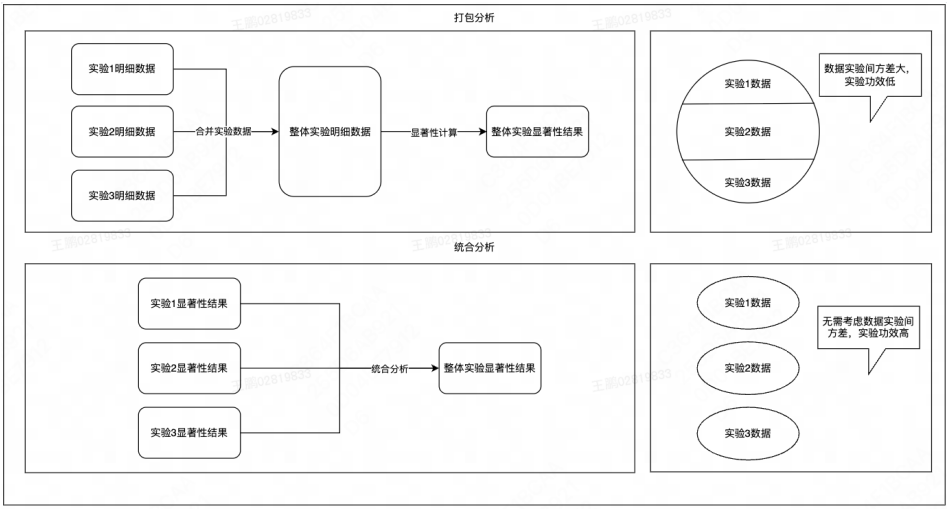
统合分析相比直接打包分析有几个主要优势：

1. 统合分析提高实验功效。例如在多个城市分别开展多次实验时，如图 7-2 所示，城市间的差异可能导致打包分析的方差较大，而统合分析实际上考虑了按子实验的分层，降低子实验之间差异带来的方差，提高检验灵敏度。而从估计

量业务口径上来说，统合分析在很多时候也能与打包分析的估计结果对齐。

2. 统合分析快捷方便。当在不同时间进行多次实验得到各自单个实验的结果后，打包分析需要拿到所有实验的明细数据进行整体的显著性计算，而统合分析只用在各个实验的结果基础上进行二次处理即可。
3. 当使用逆方差等加权方式来进行统合分析时，能够有效避免辛普森悖论对分析结果的影响，得到业务上较好解释的整体结果。

在统合分析的具体应用时，我们同样可根据具体场景与用户诉求来确定具体使用的统合分析加权方式，产出的结果包括实验估计量的加权结果，以及 MDE 估计量的加权结果，最终给出统合分析的 P 值以及显著性结论。



7.1.2 统合分析原理

假设我们有k个独立的实验。记统合分析中每个实验的加权系数为w_i，估计量为Δ_i，方差为s_i^2，其中$i=1,\ldots,k$。根据这些量我们可以得到统合结果，其中统合估计量：$\Delta_{\text{meta}}=\sum_{i=1}^{k} w_i\Delta_i$，方差$s_{\text{meta}}^2=\sum_{i=1}^{k} w_i^2 s_i^2$，MDE：

$$\text{MDE}_{\text{meta}}=(1.96+0.84)*s_{\text{meta}}$$

p值：

$$p_{\text{meta}} = 2 * (1 - \Phi(|\Delta_{\text{meta}}|/s_{\text{meta}}))$$

其中Φ为正态分布的分布函数。对于各个实验考察的连续性指标或者比率型指标Δ_i，我们先给出以下几种待选取的加权方式来确定每个实验的w_i。

1. 逆方差加权（固定效应模型）

逆方差加权（固定效应模型）提高实验功效的效果是最佳的，因为它在所有加权方式中选取了使得统合后方差最小的加权方式。从统计学直观上来说，对于方差较大的实验，我们可以认为其估计结果相对不太精确，会给予较低的权重。反之对于方差较小的实验，我们可以认为其估计结果相对精确，会给予较高的权重。但同样逆方差加权的使用所依赖的假设条件也是最强的，需要假设每个实验的实验效果都能认为是相同的。此外需要注意的是，使用固定效应模型的逆方差加权在解释意义上与传统的打包实验意义会有不同，导致口径存在区别（在每个实验样本独立同分布时往往比较类似）。

具体的，当考虑固定效应模型，权重可取为$w_i = \dfrac{1/s_i^2}{\sum_{i=1}^{k} 1/s_i^2}$，其中$s_i^2$表示第$i$个实验的方差，可得统合估计量$\Delta_{\text{meta}} = \dfrac{\sum_{i=1}^{k} \Delta_i / s_i^2}{\sum_{i=1}^{k} 1/s_i^2}$，统合方差$s_{\text{meta}}^2 = \dfrac{1}{\sum_{i=1}^{k} 1/s_i^2}$。依此两变量，结合上述统合公式可得最后的$p$值与显著性结论。

2. 逆方差加权（随机效应模型）

逆方差加权（随机效应模型）与逆方差加权（固定效应模型）的核心思想类似，但在假设上相对较为宽松，认为多个实验的策略效果估计量实际上是在一个平均效果附近波动的随机变量，通过正态分布来刻画多个实验结果。同样在解释意义上，会与打包分析的口径存在一定区别。具体的对于随机效应模型，权重可以取为：

$$w_i = \frac{1/(s_i^2 + \tau^2)}{\sum_{i=1}^{k} 1/(s_i^2 + \tau^2)}$$

其中s_i^2表示第i个实验的方差，τ^2表示随机效应的方差，其中τ^2用 DerSimonian and Laird 方差估计量τ_{DL}^2来进行估计，即：

$$\tau_{\mathrm{DL}}^2 = \max\left\{0, \frac{\sum_{i=1}^{k} s_i^{-2}(\Delta_i - \Delta_F)^2 - (k-1)}{\sum_{i=1}^{k} s_i^{-2} - \sum_{i=1}^{k} s_i^{-4} / \sum_{i=1}^{k} s_i^{-2}}\right\}$$

其中$\hat{\theta}_F$时在固定效应模型下的逆方差加权估计量，即：

$$\Delta_F = \frac{\sum_{i=1}^{k} \Delta_i / s_i^2}{\sum_{i=1}^{k} 1 / s_i^2}$$

由此，我们可以得到统合估计量：

$$\Delta_{\mathrm{meta}} = \frac{\sum_{i=1}^{k} \Delta_i / (s_i^2 + \tau_{\mathrm{DL}}^2)}{\sum_{i=1}^{k} 1 / (s_i^2 + \tau_{\mathrm{DL}}^2)}$$

统合方差为：

$$s_{\mathrm{meta}}^2 = \frac{1}{\sum_{i=1}^{k} 1 / (s_i^2 + \tau_{\mathrm{DL}}^2)}$$

依此两变量，结合上述统合公式可得最后的p值与显著性结论。

3. 分母求和加权（比率型指标）

对于比率型指标，从业务解释性上，分母求和加权是业务最好理解且与打包分析的口径对齐的计算方式。记$\Delta_i = \frac{\overline{Y}_{ti}}{\overline{Z}_{ti}} - \frac{\overline{Y}_{ci}}{\overline{Z}_{ci}}$，即第$i$个实验数据中得到的比率型指标提升。我们可以考虑使用实验组和对照组总体的分母和进行加权，即第i个实验的权重为$w_i = \frac{n_i \overline{Z}_i}{\sum_{j=1}^{k} n_j \overline{Z}_j}$，其中：

$$\bar{Z}_i = \frac{n_{ti}\bar{Z}_{ti} + n_{ci}\bar{Z}_{ci}}{n_i}$$

这个权重取法的假设前提是：

$$\frac{n_{ti}\bar{Z}_{ti}}{\sum_{j=1}^{k} n_{tj}\bar{Z}_{tj}} \approx \frac{n_i\bar{Z}_i}{\sum_{j=1}^{k} n_j\bar{Z}_j} \approx \frac{n_{ci}\bar{Z}_{ci}}{\sum_{j=1}^{k} n_{cj}\bar{Z}_{cj}}$$

即实验组的分母权重，对照组的分母权重比例与总体的分母权重比例三者近似相等。这种加权方式从业务解释上比较直观。这时有：

$$\frac{\bar{Y}_t}{\bar{Z}_t} - \frac{\bar{Y}_c}{\bar{Z}_c} = \frac{\sum_{i=1}^{k} n_{ti}\bar{Y}_{ti}}{\sum_{j=1}^{k} n_{tj}\bar{Z}_{tj}} - \frac{\sum_{i=1}^{k} n_{ci}\bar{Y}_{ci}}{\sum_{j=1}^{k} n_{cj}\bar{Z}_{cj}} = \sum_{i=1}^{k}\left(\frac{\bar{Y}_{ti}}{\bar{Z}_{ti}}\frac{n_{ti}\bar{Z}_{ti}}{\sum_{j=1}^{k} n_{tj}\bar{Z}_{tj}} - \frac{\bar{Y}_{ci}}{\bar{Z}_{ci}}\frac{n_{ci}\bar{Z}_{ci}}{\sum_{j=1}^{k} n_{cj}\bar{Z}_{cj}}\right) \approx \sum_{i=1}^{k}\frac{n_i\bar{Z}_i}{\sum_{j=1}^{k} n_j\bar{Z}_j}\left(\frac{\bar{Y}_{ti}}{\bar{Z}_{ti}} - \frac{\bar{Y}_{ci}}{\bar{Z}_{ci}}\right)$$

与将所有实验打包分析得到的估计量是对齐的。根据权重，进一步我们同样可以得到统合估计量、统合方差以及p值与显著性结论。

4. 样本量加权（连续型指标）

对于连续型指标，样本量加权是业务解释性最好的加权方式。实际上样本量加权可以认为是分母求和加权的一种退化形式，即分母固定为 1。对于连续性指标，记$\Delta_i = \bar{X}_{ti} - \bar{X}_{ci}$。我们可以按各个实验的样本量占全部实验样本量的权重进行加权：$w_i = \frac{n_i}{\sum_{j=1}^{k} n_j} = \frac{n_i}{n}$，其中$n_i$是第$i$个实验的样本量，有总样本量$n = \sum_{i=1}^{k} n_i$，实验组样本量$n_t = \sum_{i=1}^{k} n_{ti}$，对照组样本量$n_c = \sum_{i=1}^{k} n_{ci}$。当样本量比例近似相等，即$\frac{n_{ti}}{n_t} \approx \frac{n_{ci}}{n_c} \approx \frac{n_i}{n}$时，这时有：

$$\overline{X_t} - \overline{X_c} = \sum_{i=1}^{k}\left(\frac{n_{ti}}{n_t}\bar{X}_{ti} - \frac{n_{ci}}{n_c}\bar{X}_{ci}\right) \approx \sum_{i=1}^{k}\frac{n_i}{n}(\bar{X}_{ti} - \bar{X}_{ci})$$

与将所有实验打包分析得到的估计量是对齐的。由于连续型指标通常需要考虑相对提

升，在样本量加权时需要进行一定特殊的处理。我们对实验组绝对提升估计量、对照组绝对提升估计量按权重统合分别得到$\Delta_{\text{meta},t}$，$\Delta_{\text{meta},c}$，可以计算相对提升的估计量：

$$\Delta_{\text{meta}} = \frac{\Delta_{\text{meta},t}}{\Delta_{\text{meta},c}} - 1$$

对实验组均值方差、对照组均值方差按权重统合分别得到$s^2_{\text{meta},t}$，$s^2_{\text{meta},c}$，然后可以如下计算相对提升的统合方差：

$$s^2_{\text{meta,relative}} = \frac{s^2_{\text{meta},t}}{\Delta^2_{\text{meta},c}} + \frac{s^2_{\text{meta},c}\Delta^2_{\text{meta},t}}{\Delta^4_{\text{meta},c}}$$

7.1.3 统合分析的实际选取逻辑

面对多种统合分析权重，以履约实验为例，在兼顾业务解释意义以及功效角度下，建议实际选取逻辑为：

Step1： 先判断是同城多实验统合，还是多城实验统合。如果是同城多实验统合，则直接使用逆方差加权（固定效应模型），否则进入下一步。

Step2： 判断是比率型指标还是连续型指标。

(1) 如果是比率型指标，先计算分母求和加权的前提条件，即实验组的分母权重比例，对照组的分母权重比例是否超出总体的分母权重比例的正负 20% 区间范围。如果有超出，则进入下一步；如果均未超出，如果有填预期提升则判断是否分母求和加权的 MDE 小于预期提升量或者p值显著，如果未填则判断是否p值显著，若是则采用分母求和加权，否则进入下一步。

(2) 如果是连续型指标，先计算样本量加权的前提比例条件，即实验组的样本量比例，对照组的样本量比例是否超出总体的样本量比例的正负 20% 区间范围。如果有超出，则进入下一步；如果均未超出，如果有填预期提升则判断是否样本量加权的 MDE 小于预期提升量或者p值显著，如果未填则判断是否p值显著。若是，则采用样本量加权，否则进入下一步。

Step3：先使用逆方差加权（随机效应模型），如果有填预期提升则判断是否 MDE 小于预期提升量或者p值显著，如果未填则判断是否p值显著，则采用随机效应模型下的逆方差加权，否则进入下一步。

Step4：使用逆方差加权（固定效应模型）。

7.2 多重比较

7.2.1 多重比较概述

多重比较问题（Multiple Comparison Problem）是统计分析中常见的一个挑战，特别是在同时进行多个假设检验时。随着检验数量的增加，出现假阳性结果（即错误地拒绝原假设）的概率也显著增加。这会导致结果的不可靠性和科学发现的误导性。

例如，我们进行 20 个指标的独立检验，每次的显著性水平为 0.05，那么至少出现一次假阳性的概率为$1-(1-0.05)^{20}\approx 0.64$。因此在同时进行多个假设检验时，我们会调整检验的思路，从将至少出现一次假阳性的概率控制在 5% 以下变更为：控制在多重假设检验中被错误拒绝的原假设的比例在 5% 以下。具体来说，FDR（False Discovery Rate，假发现率）控制的是在所有被拒绝的原假设中，实际为真（即错误拒绝）的比例，以便在进行多个统计检验时减少假阳性结果的比例，而不是控制每个单独检验的错误率。

7.2.2 二阶段 Benjamini-Hochberg 方法规避假阳性

在业界与学界当中，对于多重比较情况有很多方法可以来纠正 P 值，例如 Bonferroni、Holm、Conditional Calibration BH 等技术，受限于篇幅这里一一列举。我们主要依赖的理论是 Benjamini-Hochberg 方法。

在实际应用中，我们对于同时检验的多个指标，会采用二阶段 Benjamini-Hochberg 方法来进行p值的修正。二阶段 BH 方法在每个阶段动态调整 FDR 阈值，以适应数据中的实际显著性模式。这种自适应调整允许在控制 FDR 的同时，尽量减少漏

掉真正显著假设的可能性。通过结合宽松的初步筛选和严格的确认检验，二阶段 BH 方法在提高统计功效的同时，有效控制 FDR。二阶段 Benjamini–Hochberg 方法除了能够处理多个独立的假设检验，对于多个假设检验正相关或弱负相关的情况也能较好的应对，能够有效防止业务对于在多重比较情况下出现的显著结果而做出错误决策。

Benjamini–Hochberg 方法

具体步骤如下：

1. **多个假设检验：**假设我们有 m 个独立的假设检验，每个检验都有一个对应的 p 值。我们记这些 p 值为 $p_1,\ldots,p_m$。
2. **排序值：**将所有的 p 值按从小到大的顺序进行排序。假设排序后的 p 值为 $p_{(1)},\ldots,p_{(m)}$，对应的原假设分别为 $H_{(1)},\ldots,H_{(m)}$。
3. **选择 FDR 阈值：**设定一个较为宽松的 FDR 阈值 α，这通常是一个较高的值，以便在第一阶段识别更多的可能显著的假设。
4. **计算临界值：**对于每个排序后的 p 值 $p_{(i)}$，计算其对应的临界值：临界值= $\frac{i}{m}\alpha$。这里，i 是当前 p 值的排序位置。找到最大的 k，使得：$p_{(k)} \le \frac{k}{m}\alpha$。这意味着从第一个到第 k 个 p 值都被认为是显著的。
5. **筛选显著假设：**将所有满足 $p_{(i)} \le \frac{i}{m}\alpha$ 的假设 $H_{(i)}$ 标记为显著。记录不显著的假设数量为 m_0。

二阶段 Benjamini–Hochberg 方法

上面 Benjamini–Hochberg 方法可以证明严格控制 $FDR \le \frac{m_0}{m}\alpha$，其中 m_0 为零假设正确的个数。为调整实现恰好 $FDR \le \alpha$，二阶段 BH 方法在传统一阶段 BH 方法的基础上，通过初步估计真假设数量，然后在第二阶段对临界值进行更精细的调整，以控制 FDR。在一阶段应用 BH 方法以初步筛选出不显著的 m_0 个假设后，我们在二阶段进入循环过程：

1. **调整临界值：**对于一阶段中得到的α阈值，我们进行调整得到新的阈值$\alpha^* = \frac{m}{m_0}\alpha$。
2. **筛选显著假设：**根据这个新阈值可以重新筛选假设，仍记录新的不显著假设数量为m_0。

重复以上过程，直至不显著假设的数量不再发生变化。根据最后筛选出的显著假设的数量（另可直方图或者阈值算法），对各个指标的p值进行修正，具体修正方式为：

a. **初步修正：**将所有指标得到的p值进行$p_{修正} = p * \frac{m_0}{i}$，其中$i$是$p$值从小到大的排序位置。

b. **累积最小化：**将初步校正的p值序列反转，从序列的末尾开始进行累积最小化得到新的修正后p值序列。其中累积最小化指的是对于每个位置i，计算从位置i到数组末尾的所有值的最小值。这一步有两个目的，一是保证p值的大小排序与修正之前的一致性，二是防止由于过度修正导致p值过大。

c. **上界限制：**将所有大于1的p值限制为1。

7.3 拓展与展望

在互联网的线上实验当中，实验者往往期望在实验运行期间就不断监控实验结果，来观测实验的走势是否符合策略预期以及样本量是否能满足需求。这里常常会有一个误区，例如在传统的随机对照实验中，实验者在实验中期观察到有策略显著的情况时，会认为策略有效从而提前结束实验，以缩减实验周期并加快策略迭代频率。然而，学界有不少研究指出，在实验期间不断偷窥实验结果会带来假阳性问题，因为直观来说，实验者每看一次结果都相当于进行了一次假设检验，多次查看即会有多重比较问题。理论上如果实验周期足够长，并且在实验期间不断进行数据收集和分析，那么几乎一定观察到一次显著的情况，这显然是不符合实验的初衷的。在这种情况下，为了兼顾实验者缩减实验周期提前观测结果的诉求与多重比较情况下的显著性结果科学性，我们探索了混合序贯概率比检验、成组序贯分检验等序贯分析的方式，能够在控

制第一类错误的情况下进行中期分析，一旦统计学上足够显著即可立即停止实验，节省实验成本。一般来说，这些序贯分析的方式通常要求在不同时间进入实验的实验单元相互是独立的，因此通常较为适用于订单随机分流等实验单元只会随着时间唯一出现的情形。

此外，一般实验者在策略中会涉及到很多参数的选择，如何对合适的策略对象选择效果最优的参数也是实验者十分关心的问题。**异质性因果效应**估计即 HTE 方法会关注不同子群体对同一策略的不同反应。传统的随机对照实验通常假设所有实验单元对策略的响应是均匀的，但在实际情况中，不同的用户群体可能对同一策略有不同的反应。在实验中，HTE 方法允许实验者在实验过程中进行更细粒度的分析，识别不同子群体的反应差异。在参数寻优方面，**MAB**（多臂老虎机）是一种动态分配策略，旨在在实验过程中不断调整资源分配，以最大化策略的整体收益。它模拟了赌博机的操作，试图在不同策略（臂）之间找到最优选择。MAB 适用于需要在实验过程中快速迭代和调整策略的场景。通过动态分配资源，MAB 能够在保证探索和利用之间取得平衡。MAB 方法在实验中期允许实验者根据实时反馈调整策略分配，减少资源浪费并提高实验效率。这在资源有限或时间紧迫的情况下尤其有用。

在搜索、广告和推荐等排序场景中，为解决溢出效应以及提高实验灵敏度，**Interleaving（交错式）实验设计**不失为一种可行的解决方案。与传统的 A/B 实验不同，Interleaving 实验考虑将 A、B 两种策略的推荐结果依次随机交织到同一个推荐列表后再展示给用户，而不是将用户分成不同组分别展示不同算法的结果。通过观察用户在这些混合结果中的行为（如点击行为），可以更快速和精确地评估哪种算法更优。该方法的优势在于使用较少的样本量就能区分出两种策略的优劣，然而其无法直接给出具体的差异值，并且工程实现成本较高。在使用时也可考虑先通过该方法快速筛选出较优的策略，如有需要再使用其他实验方法得到具体的提升幅度。

除此之外，学术界还存在贝叶斯实验评估等高阶实验技术，受限于白皮书篇幅，目前暂不做大规模详细介绍。对于这些方法，我们进行了线下小范围探索与应用，未来也计划成体系的进行建设，然后进行实践与应用。

第三部分 SDK 代码应用

第八章：开放式分析引擎

为了帮助任何用户轻松摆脱 A/B 测试中的各种挑战，让没有复杂实验背景和专家知识的人也能零门槛自主进行可信、高效的实验。同时实现实验方法库与实验平台各基础设施（例如流量配置、数据生产）的解耦，以方便各专业团队能在各自领域内发挥专长，提高平台功能与方法的迭代效率。美团履约技术团队孵化了 AB 实验分析方法库——实验分析引擎 BETA（Banma Experimentation and Testing Analysis）。该库涵盖数十种先进的实验技术，支持实验设计、评估、诊断等环节所需的多样化核心功能以及标准化流程。并在工程层面统一解决了实验过程中大量统计理论导致的实验分析复杂化问题、过多统计陷阱引起的实验不可信难题，使实验者能够无门槛地以科学、高效的方式开展实验，并轻松获得可信的实验报告。

目前，该实验分析引擎作为核心方法库，已经面向美团内部开放，方便实验者和数据科学家的按需取用以及灵活探索需求。同时可避免重复开发相同解决方案的工作浪费，促进跨团队的知识共享和能力提升，推动整体实验能力的提升。关于履约平台实验分析引擎的更多思考，可参阅美团技术博客《[新一代实验分析引擎：驱动履约平台的数据决策]》。

8.1 产品特性

在履约技术团队，运行着大量实验，我们希望赋能所有团队以速度、严谨和信心进行改进。为此，秉承着可信、开放、敏捷、易用的原则，打造了新一代实验引擎。该实验分析引擎通过良好的封装以及设计，包括但不限于以下特性。

1. **丰富的实验方法：**涵盖白皮书中提到的所有实验方法，包括随机对照实验、随机轮转实验、准实验、观察性研究 4 大类别，其中提供了 11+ 种实验方法、7+ 种分组方法、10+ 种假设检验方法等等。不仅覆盖单边、多边实验场景多样实验用例，还提供了业界领先的小样本解决方案，如协变量自适应分组来解决小样本同质性问题，轮转实验和双重差分实验来应对溢出效应问题，以及合成控制法等观察性研究技术等。
2. **方便易用：**标准化实验分析请求参数，分析引擎会依据实验方法、指标类型、样本分布等上下文自动选择最为合适的检验方法，确保分析过程的鲁棒性。同时整个实现流程标准化，可自动执行数据预处理、策略效应估计、方差计算、P 值计算（假设检验），最终得出分析结论。如果进行的是实验设计，系统会根据实验方法选择与之相匹配的实验分组方法。
3. **性能高效：**分析方法执行期间会充分利用向量化运算、并行化技术来提升分析效率，其中随机对照实验支持分布式计算，亿级别数据可在分钟级完成分析。
4. **多重比较修正：**检验结果自动进行多重比较修正，解决了对于多实验组、多指标检验引起的第一类错误增大问题，确保实验结果科学性。
5. **功效提升：**例如随机对照实验场景下支持通过 CUPED 来进行方差缩减以提升检验灵敏度，支持选择一元同系数 CUPED、二元同系数 CUPED、新 CUPED 等多种方差缩减方法。
6. **统合分析：**突破单次实验样本量限制，支持对相同目的、相互独立的多个实验进行综合分析，以提升检验的统计功效。有助于在小流量场景下检测策略效果，且无需单个实验有大量样本就能获得可信结果。支持的统合分析技术包括样本加权、逆方差加权等。
7. **功效测算：**提供最小样本量预估、MDE 等计算方法，方便用户在实验前判断实验样本量是否充足，以避免实验白做。同时在实验后帮助用户分析策略不显著原因，判断是由于样本量不足，还是策略无效或未达预期导致，从而支持科学的实验决策。

8.2 系统设计

如图 8-1 所示，分析引擎提供标准化的分析流程以及多样化方法，采用模块化和分层设计原则来提升实验方法的迭代、拓展效率，以及实现像积木一样灵活应用，服务不同角色的用户。具体每层作用如下：

- **应用层：**上层实验平台入口，包括到家履约团队孵化的图灵实验平台，通过接口接入分析引擎的第三方实验平台。此外还可通过 Python SDK 线下便捷使用实验分析引擎进行实验分析。
- **接口层：**面向应用层提供的标准化的实验设计与实验评估接口。通过抽象出通用的实验分析参数，如数据集、分析指标、指标元数据、实验分组信息、扩展参数等信息，进而标准化了整个实验分析流程。这种方式提升了系统扩展性，方便我们快速集成新的实验方法，降低运维成本。
- **路由层：**实验分析模板是通过原子分析方法库编排出的对应实验方法的确定的分析流程。路由层会根据实验方法、执行引擎信息，路由至不同的实验分析模板。特别的对于面向海量数据场景下的普通随机对照实验，我们通过抽象出一些关键聚合算子（如协方差、方差、均值等）的计算逻辑，适配了一套基于 PySpark 的分布式执行引擎。利用到 Spark 对于批量聚合算子的处理优化技术，做到了分钟级完成亿级以上的海量数据评估。
- **数据准备层：**实验分析流程之前引擎层统一进行数据处理流程来准备实验数据集，包括：数据加载、数据预处理、指标二次计算等。这里的数据处理流程同时支持了单机与分布式两种方式。
  - **数据加载：**单机分析方式支持 S3（美团内部存储服务）文件、HTTP 文件作为数据源，通过 pandas.DataFrame 方式表征数据集。分布式分析方式支持 HDFS 文件、Hive 取数 SQL 两种方式定义数据源，通过 pyspark.DataFrame 来表征数据集。实验数据集定义方式支持多源合并策略，包括按列合并、按行合并。
  - **数据预处理：**引擎侧对异常数据进行统一的预处理，以获取有效、准确的实

验数据，整个流程包括：① 空值填充，对于指标空值进行补零填充；② 数据格式转换，指标类型统一转换为 Float32 类型；③ 异常值剔除：支持实验单元为空值时的自动剔除，支持配置 3σ 、IQR 等统计方法对天级异常数据的剔除，并将剔除信息展示在最终的实验报告中。④ 补齐缺失计算指标，基于指标计算公式以及对应原子指标，补齐缺失的计算指标。

- **指标二次计算：**满足个性化指标计算诉求，通过将更细粒度的数据按照预定义的指标聚合算子，上卷至实验单元粒度的数据。如：xx 指标 90% 分位点 −10% 分位点。

- **分析方法层：**实验分析引擎所集成的核心方法库，这一层通常由数据科学家负责，涵盖实验分组方法、假设检验方法、功效提升技术、最小样本量预估、MDE 计算等核心方法。每种实验方法的实验分组方法、显著性检验方法均编排至对应的分析模板中。其中显著性检验方法必须经过 AA 模拟验证才能上线，以确保实验方法的科学性。
- **实验分组：**一般情况下，一个分组方法确定了一个实验方法。整体来看实验分组方法主要分为两类，① 随机实验分组，支持对于单次分组不同质时的系统重分组（最多 5 次），以尽可能获取一个满足实验条件的分组。这类实验方法包括：随机对照实验、随机轮转实验。② 最优实验分组，通过随机多次产生多个分组，选取 Diff 最小的一个分组作为最终分组。这类实验方法包括：随机配对实验、DID 双重差分实验。
- **显著性检验：**通过实验方法、数据特点选择合适的假设检验方法，产出显著性检验结果。这里检验方法主要包括四大类：非参检验、参数检验、配对检验、模型检验。对于随机对照实验，默认会通过 CUPED 方法来提升检验灵敏度。完成显著性检验后也会通过统一的 P 值多重比较修正，以解决指标多重比较之后带来的假阳率升高问题。

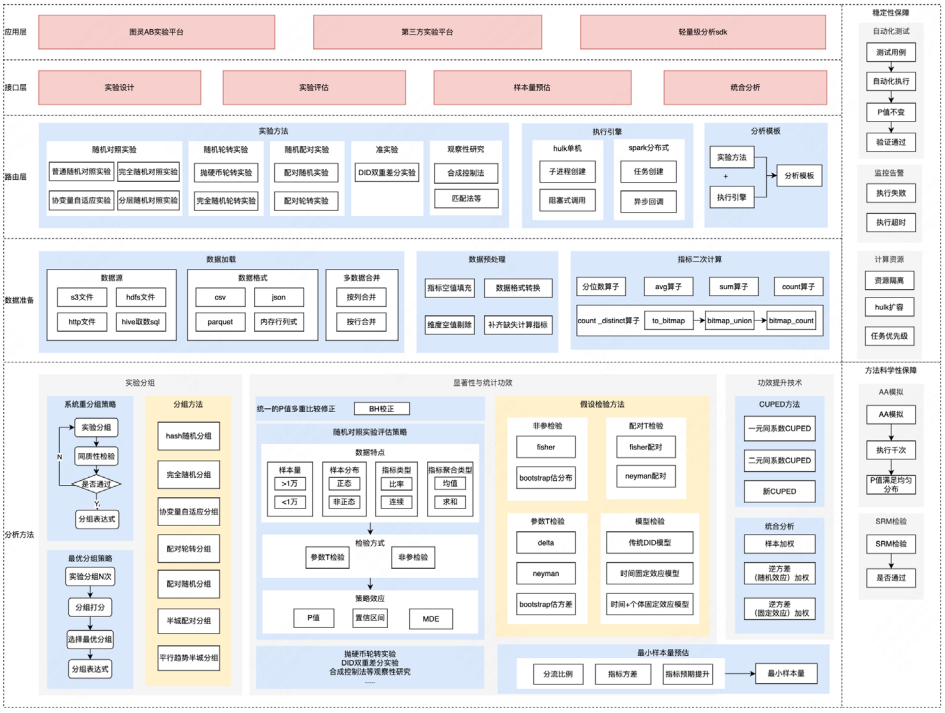


图 8–1：分析引擎架构图

8.3 系统接入

作为核心分析方法库，当前分析引擎已面向美团内部全体成员开放，并提供多种接入途径，方便实验者自行根据自身场景选择最佳实验接入方式。

接入方式	详情	面向场景
平台整体开放	**平台全托管**：实验由图灵实验平台统一管理。 • 平台侧统一维护实验信息，实验整个生命周期在平台侧进行统一管理。可提供一站式的完成算法迭代、实验设计、实验下发、实验报告、实验推全等能力。 • 平台侧可统一进行流量管理、冲突判定、策略配置，同时使用平台侧提供的分流框架确定线上流量分配。	面向实验平台用户 1. 已是图灵平台的用户，业务也集成在图灵平台中，需要AB实验能力。 2. 在其他业务平台中开了实验，但是想在图灵平台中看实验效果报告。
服务化接口开放	**平台半托管**：实验设计与实验评估能力的开放。 • 以thrift接口方式为第三方提供稳定的分析服务，包括开放实验设计、实验评估能力等。 • 业务方自行进行流量管理、冲突判定、策略配置。	面向平台工程，需要闭环在业务平台看实验效果报告 1. 业务平台具备了实验开展能力但是缺少全部或部分实验评估能力。 2. 自建了实验平台但是感觉分析方法不够置信。
线下分析SDK	**分析引擎开放**：原子分析方法粒度的开放 • 通过python sdk方式提供线下的分析能力，提供了分析Demo一键本地运行，可由业务侧自行根据情况选择适合的方法进行分析。	面向线下分析评估场景 1. AB实验暂未流程化但是需要看报告的场景，主要面向数据科学家、算法、商分等同学。 2. 已通过其它渠道产出了一份实验报告，但不确定可信性，使用SDK对结果进行线下对比。

8.4 线下分析实战

案例：履约 xx 算法计划开展一次随机对照实验以对比验证策略效果，实验流量选取若干城市，实验单元为 AOI，实验指标：订单量、完单用户数，人均完成单量，其中人均完成订单量指标为比率型指标，计算公式 = 订单量 / 完单用户数

预期分三个实验分组，分组流量配比分别为 2:3:5。额外要求采用 CUPED 方法以提升检验灵敏度，通过 murmurhash 哈希函数来生成随机分组表达式。

方案：由于实验或指标暂时并未接入实验平台，为了快速开展实验，需采用线下分析的方式来进行实验设计，以寻找满足实验要求的分组划分方式。通过如下四个步骤即可在线下完成一个完整的随机对照实验设计（实验后的评估流程类似），具体流程如下：

步骤 01：引入分析包

通过 pip 命令安装线下分析 SDK，引入核心分析客户端 AbAnalyzeClient 及相关类。

```
# 安装线下分析SDK
!pip3 install bm_exp_analyse_client==1.0.75.dev3 -i http://pypi.sankuai.com/simple --trusted-host
```

找Bug | 重构 | 解释 | 注释 | 单测 | Chat

```
import pandas as pd
import numpy as np
from exp_analyse.sdk.AbAnalyzeEngine import AbAnalyzeClient
from exp_analyse.util.domain import AnalyzeRequest, KpiAttentionType, ExpGroupingType

# 定义ab分析client，分析引擎说明参考：https://km.sankuai.com/collabpage/2216636907
ab = AbAnalyzeClient()
```

```
Last executed at 2024-12-30 10:55:48 in 50ms
```

步骤 02：定义分析参数

1. 通过定义数据集、实验分组、分析指标等信息来构造分析请求，请求参数对应于数据结构 AnalyzeRequest。
2. 可以通过扩展参数 extArgs 来控制指定具体的分析行为，本案例中通过设置方差估计方法为 delta、CUPED 方法为二元同系数回归调整来指定随机对照实验的分析行为。

找Bug | 重构 | 解释 | 注释 | 单测 | Chat

```python
# rdf=... # 实验数据集，随机化单元粒度明细数据，包含实验前、实验前前两个周期的数据

# 定义计算指标
kpi_list = ["订单量", "完单用户数", "人均完成单量"]

# 定义指标元数据，参考数据结构：KpiInfo
kpi_infos = [
    {
        "kpiId": "订单量",
        "calculateFormula": "k1",
        "aggrType":"AVG", #指标值聚合类型，一般连续性指标可定义，可选值为SUM/AVG，默认为AVG
        "kpiAttentionType": KpiAttentionType.TARGET ##指标关注类型，分为目标指标、驱动指标、护栏指标、其它指标，用以按照指标类型进行多重比较修正。
    },
    {
        "kpiId": "完单用户数",
        "calculateFormula": "k2",
        "kpiAttentionType": KpiAttentionType.TARGET
    },
    {
        "kpiId": "人均完成单量",
        "calculateFormula": "k1/k2",
        "kpiAttentionType": KpiAttentionType.TARGET
    }
]
request = {
    "expInfo": {  # 描述实验基本信息
        "expName": "普通随机对照实验DEMO",
        "expUnits": ["aoi"], # 实验单元
        "groupMethod": ExpGroupingType.NORMAL_RANDOM.value, # 实验方法，可枚举
        "startDt": "20241015", # 实验开始、结束时间，非必填
        "endDt": "20241028"
    },
    "expData": [ # 描述实验数据集
        {
            "dataframe": rdf.query('period==1'),
            "dataPeriodFlag": 1 # 实验期
        },
        {
            "dataframe": rdf.query('period==0'),
            "dataPeriodFlag": 0 #历史期，做CUPED，如未提供历史期数据，则默认不启用CUPED
        }
    ],
    "expGroupDescriptions": [ # 描述实验分组信息
        {
            "groupId": "A",
            "groupName": "对照组",
            "flowRatio": 0.2
        },
        {
            "groupId": "B",
            "groupName": "实验组B",
            "flowRatio": 0.3
        },
        {
            "groupId": "C",
            "groupName": "实验组C",
            "flowRatio": 0.5
        }
    ],
    "kpiInfos": kpi_infos, # 描述指标元数据信息
    "kpiList": kpi_list,
    "extArgs": {
        "groupHashMethod":"matrixhash", #定义分组hash函数，支持murmurhash和matrixhash两种方式，默认为murmurhash
        "groupBucketCount": 10, # 指定分桶数量，默认为100
        "significance_method": "delta", # 指定使用delta方法估计方差
        "ratio_cuped_method": "binary_homology_coefficient" #使用二元同系数回归调整方式对于比率型指标进行方差缩减，提升检验指标灵敏度
    }
}
```

步骤 03：发起分析请求

1. 通过函数式调用发起分析请求，这里会自动执行远程调用操作，直至返回分析响应结果，响应结果对应数据结构 [AnalyzeResponse]
2. 如果某次分组不同质，后端系统会自动进行重试以获取一个目标指标满足同质性的分组结果。

3. 发起实验设计请求

找Bug | 重构 | 解释 | 注释 | 单测 | Chat

```
[14]: analyzeRequest = AnalyzeRequest.parse_obj(request)
      rs = ab.design_main(analyzeRequest)  #产出实验设计分组结果
```

Last executed at 2024-12-30 10:56:52 in 664ms

步骤 04：实验设计报告

1. 通过 show_report 函数以表格方式展示同质性检验结果信息，该报告支持复制。
2. 业务可自行根据 MDE、P 值来判断该次分组是否达到实验开展的前提条件，如果达到满足实验条件，即可使用最终分组表达式来开展实验。

4. 展示实验设计报告

```
show_report(rs, show_group_expression=True, ndigits=3)
```

Last executed at 2024-12-30 11:26:52 in 23ms

```
分组：A，分组表达式：(matrixhash(aoi, 1735529211083)%10) in (0,1)
分组：B，分组表达式：(matrixhash(aoi, 1735529211083)%10) in (2,3,4)
分组：C，分组表达式：(matrixhash(aoi, 1735529211083)%10) in (5,6,7,8,9)
```

分组	指标ID	tMean	cMean	绝对提升	相对提升	mde	置信区间	p值	结论	检验方法
B	订单量	[illegible]	[illegible]	8.187	0.594%	8.585%	[-5.412%,6.6%]	0.846	不显著	delta估计方差T检验
	完单用户数	[illegible]	[illegible]	6.046	4.802%	9.593%	[-1.909%,11.513%]	0.161	不显著	delta估计方差T检验
	人均完成单量	[illegible]	[illegible]	-0.439	-4.015%	0.641	[-0.888,0.009]	0.055	不显著	delta估计方差T检验
C	订单量	[illegible]	[illegible]	9.134	0.663%	7.568%	[-4.631%,5.957%]	0.806	不显著	delta估计方差T检验
	完单用户数	[illegible]	[illegible]	4.205	3.34%	8.315%	[-2.477%,9.157%]	0.260	不显著	delta估计方差T检验
	人均完成单量	[illegible]	[illegible]	-0.283	-2.591%	0.581	[-0.69,0.123]	0.172	不显著	delta估计方差T检验

总结与展望

本白皮书基于美团履约与外卖策略的实验实践，系统梳理了随机对照实验、随机轮转实验、准实验、观察性研究四大类方法以及高阶实验工具，涵盖数十种实验技术，构建了完整的实验科学方法体系。为提高实操性，同步提供了配套分析引擎工具的使用指南，可助力用户快速上手。展望未来，我们将持续追踪实验方法的前沿进展，分享其在履约等场景的落地经验与最佳实践。同时逐步开放可信实验分流与计算架构等实验知识，推动实验能力的规模化赋能。通过共建科学、高效的实验体系与文化生态，致力于为组织的创新突破与可持续增长提供坚实支撑。

致谢

首先衷心感谢美团履约与外卖数据科学团队，特别是主要作者以及同事们对本白皮书的辛勤付出。同时感谢履约与外卖算法、数据、业务和产品等团队的鼎力支持，正是多部门背后的协同实践探索，以及对实验科学的信赖与支持，才使得我们不断深化对实验方法的理解与应用。最后诚挚感谢每一位读者的关注与阅读，希望本文对您有所启发，也欢迎和期待与您分享交流，共同成长。