arXiv:2302.02570v1 [cs.AI] 6 Feb 2023

算法资源分配随机试验的改进策略评估

Aditya Mate [1 2] † **Bryan Wilder** [3] **Aparna Taneja** [2] **Milind Tambe** [1 2]

摘要

我们研究通过随机对照试验（RCT）评估算法资源分配策略的任务。这类策略旨在优化有限干预资源的利用，以最大化由此带来的收益。通过随机对照试验评估此类分配策略并非易事，无论试验规模多大皆是如此，因为资源约束决定的策略决策使个体的结果彼此密不可分。我们的主要贡献是提出一种新的估计量，该估计量利用了我们提出的新颖概念：在随机对照试验结束时，对不同实验组中的参与者进行回顾性重新分配。我们确定了此类重新分配可行且可用于构建反事实试验的条件；这些反事实试验的结果可以免费准确地确定。我们从理论上证明，与基于样本均值的常见估计量相比，这种估计量更准确——我们表明，它既能给出无偏估计，又能同时降低方差。我们通过合成数据、半合成数据以及真实案例研究数据开展实证实验，展示了我们方法的价值，并表明该方法在各种情况下都能提高估计准确性。

1. 引言

我们研究一类随机对照试验（RCT），其目标是评估算法资源分配策略的效能。这类策略会推荐一种分配（行动），通常会利用强化学习等工具（Galstyan 等人，2004；Tesauro 等人，2006）、多臂老虎机框架的各种变体（Jamieson 等人，2014）、网络优化——（Wilder 等人，2021年；Vaswani 等人，2015年）等。随着机器学习日益广泛地应用于对社会至关重要的领域，这类策略已被用于在各种领域分配有限资源，包括活动优化（Leskovec 等人，2007；Eagle 等人，2010年）、改善孕产妇医疗保健（Biswas 等人，2021年）、肝细胞癌筛查（Lee 等人，2019年）、结核病患者监测（Mate 等人，2021年）等。这类策略可以对群体中的所有个体进行排序，并将家访等稀缺资源提供给群体中排名最高的个体，例如由卫生工作者进行家访。

评估这些资源分配策略的随机对照试验由 M 标记为 $\{1, \ldots, M\}$的实验组构成，每个实验组包含 N 名独一无二、随机分配的个体。每个实验组中的策略都规定了资源分配方式，同时遵守一定的资源约束。这一过程既可以是一次性的（只进行一次分配），也可以是序贯的（在多个轮次中自适应地做出新的分配决策）。每种策略都规定了自己的资源分配方案，并在试验结束时通过分析相应实验组的结果数据进行评估。资源分配策略在群体层面做出决策，这给其实验分析带来了新的挑战。在标准随机对照试验中，目标是评估处理效应——由于不存在资源约束，所有参与者都会接受治疗——针对的是平均参与者（Angrist 与 Pischke，2009）。由于所有参与者的结果彼此独立，分析人员只需比较治疗组与对照组的结果。然而，资源分配策略试验旨在评估策略所作用的一组个体的群体结果（其中一些人获得资源，另一些人则没有）。即使按常规随机对照试验的标准来看，试验中的个体数量很多，分配策略的随机试验仍可能存在较高方差，从而导致处理效应的估计值噪声较大且可能有误。

这种高方差源于两个方面。首先，给定臂中的个体，其结果彼此相关，因为分配策略通常会在做出分配决策时联合 考虑所有个体（例如，为了遵守关于可接受干预的个体总数的预算约束）。这种相互依赖意味着

† 作者在谷歌研究院担任研究实习生期间完成的工作 [1]美国哈佛大学 [2]印度谷歌研究院 [3]美国卡内基梅隆大学。联系作者：Aditya Mate <aditya mate@g.harvard.edu>。

我们在每次随机对照试验中只能观察到某一策略表现的一个独立样本，而不是许多个（与标准随机对照试验中每位参与者对应一个样本不同）。其次，对于许多个体，不同待比较策略所做的分配决策可能完全一致。例如，无论采用何种策略，许多个体可能始终会被筛选进入，或始终会被排除在接受资源之外。由于这些个体不会受到所采用策略的影响，并且在不同测试策略下获得相同的分配，他们的结果实际上从未受到所采用策略的影响。然而，由于其结果具有随机性，他们的存在会在群体平均结果中产生额外方差，而这种方差与所评估的策略无关。据我们所知，尽管基于机器学习的策略在社会关键领域中的应用日益广泛，但此前尚无研究提出任何策略来缓解资源分配算法随机对照试验中的这些方差来源。

一种直观的方差缩减初步方法，是在各试验组之间共享信息。例如，我们可以将所有测试策略中未获得资源分配的个体汇集起来，对其结果的波动取平均。不幸的是，这种简单的估计量存在偏差，因为各试验组中个体获得资源（或未获得资源）的分布并不相同——事实上，评估这些分布的相对效能正是试验的目的。更复杂的策略是通过逆倾向加权估计量，利用不同策略之间的重叠情况，重新加权某一试验组中的参与者，使其匹配另一策略原本会采取的动作（*Kuss* 等人，2016年；Austin 和 Stuart，2015年）。然而，这类估计量本身也有方差极高的恶名，尤其是在策略之间重叠程度较低时（Zhou 等人，2020年）。此外，正如我们在第4.1节所讨论的，在序贯干预场景中，倾向得分无法计算，因为未观测到的状态使我们无法评估针对另一策略的个体层面倾向。

我们为解决这些问题所做的核心贡献，是提出一种新颖的估计量，用于估计资源分配策略的处理效应。该估计量以严谨的方式利用试验组之间的重叠，并保证降低方差，且适用于单阶段或序贯场景。其核心思路是寻找满足以下性质的个体子集：如果我们进行一项假设试验，将这些个体的试验组分配互换，那么所有个体获得的资源分配仍保持不变。我们的估计量能够识别出具有这一性质的个体集合，并对所有相应假设试验的结果取平均（由于分配决策未偏离原试验，这些结果均可观测）。我们的贡献如下：（1）提出这一新颖的估计量，并证明与标准估计量相比，该估计量能够产生处理效应的无偏估计，并保证降低方差，这意味着其平均误差严格更小；（2）展示如何以计算高效的方式实现该估计量，适用于一类策略，其中涵盖了真实世界分配决策中最常用的策略；（3）我们在三个领域开展实验，使用合成、半合成和真实世界数据。我们将这些技术应用于真实世界案例研究数据，展示了其在社会关键领域中的用途。总体而言，与现有估计值相比，我们的估计量大幅降低了误差（最高可达 $\sim 70\%$）。

2. 问题表述

随机对照试验设置：我们考虑一组 N 个体。每个个体都有一个特征向量 $\mathbf{x} := [\mathbf{x}_o, \mathbf{x}_u]$，其中 $\mathbf{x}_o \in \mathcal{R}^o$ 表示该个体的可观测特征（例如人口统计特征），而 $\mathbf{x}_u \in \mathcal{R}^u$ 表示尽管不可观测、却会影响结果的特征。资源分配策 π 联合考虑所有 N 个体及其可观测特征的联合矩阵 $\mathbf{X}_o \in \mathcal{R}^{N \times o}$，并规定一个分配（动作）向量 $\mathbf{a} := \pi(\mathbf{X}_o) \in \mathcal{A}^N$，其中 $\mathcal{A}$ 表示每个个体可能采取的动作空间。分配 $\mathbf{a}$ 必须遵守该领域特定的资源约束。一个例子是预算约束 $||\mathbf{a}|| \leq B$，用于限制所分配动作的总成本。随后，每个个体根据未知函数 $P^*(\mathbf{x}\ a, s)$ 随机产生一个结果状态 $s \in \mathcal{S}$，该函数表示该个体获得结果 s 的概率。我们用 $r(s, a)$ 表示政策制定者从该结果中获得的奖励。需要指出的是，这一记法等价于标准的潜在结果框架（Rubin，2005年）：其中 $r(s, a)$ 是给定动作 a 时该个体潜在结果的实现值，而 P 决定 $\mathbf{x}$ 与潜在结果之间的联合分布。我们采用当前记法，以便将其轻松推广到序贯场景。

随机对照试验的目标是比较 M 此类资源分配策略，方法是使用 M 随机构建的实验组 $C_1, \ldots C_M$ 其中 C_i 表示分配到 i^{th} 臂的个体集合。我们用 $\mathfrak{C} := \{C_1, \ldots C_M\}$ 表示这一特定的个体实验组分配，它是从 $\mathcal{C}$中均匀随机选取的；$\mathcal{C}$表示所有可能分配的全集。随机对照试验结束时，分析人员可以比较 M 分配策略的性能，具体方法是比较各臂参与者获得的奖励总和，记为 $\text{Eval}(\pi_m) := \sum_{i \in C_m} r\big(s(i)\ a(i)\big)$。

序贯随机对照试验： 我们还考虑涉及多阶段资源分配设置的随机对照试验。这类随机对照试验总共运行 T 轮（上述单阶段设置对应于 $T = 1$），其中每个时间步 $t \in [T]$ 都必须进行分配$\mathbf{a}_t \in \mathcal{A}^N$，并受可能随时间变化的资源约束限制。在每个时间 t，每个个体都具有一个属于状态空间 $\mathcal{S}$的状态 s 。我们

用 $\mathbf{s}_t$ 表示时间步 t时个体的状态 N 。策略 π 可以考虑先前状态和动作的完整历史，以生成新的行动分配 $\mathbf{a}_t \coloneqq \pi(\mathbf{X}_o,\ \mathbf{s}_{0:t-1},\ \mathbf{a}_{1:t-1})$。类似地，个体根据概率函数 $P^*(\mathbf{x}\ s_{0:t},\ a_{1:t+1},\ s_{t+1})$转移到新状态 s_{t+1} 。所有 N 参与者的状态和动作轨迹，按每个 M实验组记录为矩阵： $S_1,\ \ldots,\ S_M \in \mathcal{S}^{N\times T+1}$和 $A_1,\ \ldots,\ A_M \in \mathcal{A}^{N\times T}$，从而便于按照单次设置中的定义，进行类似的 $\text{Eval}(\pi_m) \coloneqq \sum_{i\in C_m} r(S[i],\ A[i])$ 计算。

基于指标的策略：我们定义了一类特定的策略——称为“基于指标的策略”——这类策略可以推导出计算效率较高的估计量。基于指标的策略依据某种评分规则对个体进行排序，该规则决定了资源分配时个体之间的优先顺序。由于透明且易于实施，这类策略涵盖了实践中常用的大多数相关资源分配策略（Ustun & Rudin，2019年）。请注意，这类策略还包括一些看似不太可能的候选策略，例如按固定顺序循环向个体分配资源，或“对照组”等基准策略（详见附录 B）。形式上，对于二元行动空间 $\mathcal{A} \coloneqq \{0,1\}$，基于指标的策略定义为满足预算约束的策略，即最多允许 B 个体接受动作 $a=1$。该策略在每个时间步 t，仅根据个体的可观测特征 $\mathbf{x}_o$、状态轨迹 s 以及已接受动作的历史记录 a，为每个个体计算一个时变指标$\Upsilon(\mathbf{x}_o, s_{0:t-1}, a_{1:t-1})$。该策略 π 将动作 $a=1$ 分配给指标 Υ 值最大的前 B 个体。在任意给定时间 t，我们假设每个个体的 Υ 值均不相同，即该策略会产生一个全序。基于指标的策略的关键特征是， Υ 仅根据每个个体自身的特征和历史记录独立计算。

问题陈述： $\text{Eval}(\pi_m)$ 提供一个随机样本，用于估计分配策略 π_m 的期望 性能。该估计结合了分配 $\mathfrak{C} \in \mathcal{C}$ 中的随机性，以及状态转移的随机性所引发的每个 $\mathfrak{C}$ 内结果的随机性。令 $\text{Eval}^*(\pi_m)$ 表示 π_m 的性能期望值：

$$\text{Eval}^*(\pi_m) \coloneqq \mathbb{E}_{S_m\sim P^*}\mathbb{E}_{\mathfrak{C}\sim\mathcal{C}}\Big[\text{Eval}(\pi_m)\Big] \qquad (1)$$

我们假设数据仅来自一次随机对照试验。我们的目标是构建一个估计量，仅凭这一个随机对照试验实例就能准确估计 $\text{Eval}^*(\pi_m)$。值得指出的是，我们的结果不对试验中的个体集合做任何分布假设，例如，不要求这些个体独立同分布地来自某个分布。相反，我们将观测到的个体视为固定的（非随机）对象，并提出仅利用随机对照试验分配过程中的随机性的技术。唯一需要的假设是，个体的状态转移彼此独立，这与因果推断中标准的稳定单元治疗值假设（SUTVA）类似（Hern´an &Robins，2010年）。这与因果推断中日益受到重视的基于设计的推断相一致（参见（Hudgens & Halloran，2008；Mukerjee 等人，2018年；Abadie 等人，2022年）），这使我们能够提出所需建模假设极少的方法。

3. 相关工作

离策略评估： 与本文最相关的既有工作是离策略评估（OPE）领域的研究，其目标是利用在某个基线策略下收集的样本来评估新策略（Sutton 等人，2008；Jiang & Li，2016；Wang 等人，2017）。OPE 经常采用逆倾向加权估计量（Liu& Zhao，2010年；Li 等人，2015年）；我们将在第 4.1 节讨论构造此类估计量的策略及其缺点。迄今为止，OPE 领域的研究主要集中于个体层面决策，而非我们研究的群体层面资源分配。一个例外是列表级离策略评估，其中策略会向用户推荐一个排序后的项目列表（Swaminathan 等人，2017）。列表级 OPE 与我们的单步情形最为相似，而我们则开发了可扩展至多步设置的方法。此外，列表级 OPE 的研究动机是个体层面奖励不可观测；而我们假设个体奖励可观测，单步设置面临的挑战则是策略可能是确定性的（因此无法使用他们的方法）。

个体治疗规则： 统计学领域的近期研究探讨了个体治疗规则（ITR）的实验设计与分析，这与我们研究的基于指标的策略类别相似（Imai &Li，2021年；Athey & Wager，2021年）。两者的关键区别在于，ITR 会针对每个个体独立做出决策，而在我们的设定中，策略会综合考虑整个个体群体，这对于精确落实预算等约束是必要的。我们的技术旨在解决这样一个需求：当策略只能在群体层面进行评估时，如何降低方差。

整群随机化与干扰： 我们的研究设置与一类称为整群随机试验的随机对照试验有关（概述请参见（Hayes&Moulton，2017））。在这类试验中，治疗是在群组层面分配的（例如，在学校内分配教室，而不是分配学生），这与我们的设置中将策略分配给个体群组的做法相同。然而，整群随机试验的动机在于可能存在溢出效应，即一个单元的结果可能影响其他单元。相比之下，在我们的设置中，给定策略的动作后，个体的结果仍然相互独立。因此，无需像干扰文献中那样考虑潜在的相关性；相反，我们利用这种结构化的独立性来开发方差更低的估计量。

4. 方法论

我们的目标是利用多个资源分配策略在决策上的重叠，改进对部署每种策略所获奖励的估计。我们首先开发一种采用逆倾向加权的估计量——这是一种通常用于此类任务的常用方法——并展示如何将其自然地应用于单阶段设置。然而，这种自然的估计量面临两项挑战。首先，逆倾向评分估计量可能存在众所周知的高方差问题，我们在第6节通过实证确认了这一现象。其次，在多阶段设置中，由于数据缺失（详见下文），倾向评分无法计算，因此该方法完全失效。为解决这些挑战，我们开发了一种更稳定的"分配置换"估计量，适用于这两种设置，并保证能够降低估计误差。

4.1. 倾向得分方法

在典型的离策略评估场景中，逆倾向加权（IPW）方法会根据观测到的动作被某一策略采取的概率，对样本重新加权。由于我们并不假设策略是随机的，这些方法无法直接应用于我们的问题——事实上，现实世界中的政府、健康系统等所实施的策略很少明确采用随机化。当策略是确定性的，策略产生另一种动作的概率恰好为零，导致标准的倾向得分估计量无法定义。

我们展示了如何在单步设置中利用另一种随机性来源来规避这一问题：试验本身的分配。在分配的随机性范围内计算，每个人被分配到某一给定动作的概率各不相同，记为 $Pr_{\mathfrak{C},\,\pi}[a(i) = a]$（直观而言，他们是否获得资源取决于该策略将他们与哪些其他人进行比较）。形式上，交换期望的顺序后，我们可以将 Eval*(π) 写为：

$$\sum_{i=1}^{N} Pr[i \in C_m] \sum_{a \in \mathcal{A}} Pr_{\mathfrak{C},\pi}[a(i) = a | i \in C_m] \mathbb{E}_s[r(s,a)]$$

由于分配 $\mathfrak{C}$ 是随机的，$Pr[i \in C_m] = \frac{1}{M}$。此外，在内层项中，以 $i \in C_m$ 为条件后，C_m 的其他成员仍服从均匀随机分布。因此，对于任意策略 π，我们可以通过反复抽取分配 $\mathfrak{C}$ 的样本并运行 π，来估计 $Pr_{\mathfrak{C},\,\pi}[a(i) = a | i \in C_m]$，从而揭示在给定包含个体 i 的群组 C_m 时，π 是否会将 $a(i) = a$。令 $\hat{p}(i,\ a|\pi)$ 表示这些样本中满足 $a(i) = a$ 的比例。用于估计 Eval(π) 的标准逆概率加权估计量为

$$\frac{1}{M} \sum_{m=1}^{M} \sum_{i \in C_m} \frac{\hat{p}(i, a(i)|\pi)}{\hat{p}(i, a(i)|\pi_m)} r(s(i), a(i)). \qquad (2)$$

在序贯设置（$T > 1$）中，倾向得分方法因两个原因而完全不适用。首先，标准的多时步逆倾向加权估计量要求策略具有随机性，而我们假设策略可以是确定性的。我们不能在多个时间步中采用上述替代方法（利用分配中的随机性），因为个体 i 在未来时间步获得动作 a（i）的边际概率取决于所有其他个体的状态，而我们没有反事实分配下这些未来状态的样本。其次，即使我们仅考虑随机策略，标准的离策略方法也需要计算在观测到的状态下恰好采取观测到的动作序列的概率。在我们的情形中，这需要计算 π 为每个个体选取其分配的动作 向量 的概率；也就是说，在每个时间步内，我们有一个 N维的行动空间。众所周知，多步逆概率加权估计量的方差会随 T呈指数级增长，常常因此变得不切实际（Li 等人，2015年）。在我们的情形中，其方差最坏情况下也会随 N 呈指数级增长。

4.2. 主要贡献：分配置换

我们提出了一种新颖的方法，同时应对这两项挑战，从而计算出稳定、准确的估计量。该估计量的核心思路是，识别具有反事实实验组分配的假设试验；利用原始试验中给定的结果，可以准确确定这些假设试验的奖励结果。我们利用这样一个事实：状态转移仅取决于所接收的分配，而与选择这些分配的策略 π 无关。

作为热身，考虑一个单次试验 $\mathcal{T}$，其中两个个体 i 和 j 被分配到策略 π_i 和 π_j，这两个策略为两名个体作出完全相同的资源分配 a，分别产生结果 s_i 和 s_j。现在考虑一个假设试验 $\mathcal{T}^\dagger$，其运行方式与 $\mathcal{T}$完全相同，唯一不同的是，i 和 j 的分配互换了。如果在 $\mathcal{T}^\dagger$中，i 和j 都获得与 $\mathcal{T}$中相同的分配 a，那么对其他个体的分配也不会受到影响，因此所有个体在 $\mathcal{T}^\dagger$ 和 $\mathcal{T}$中看到的输入都完全相同。由此，$\mathcal{T}$中的实际结果样本 s 与由 $\mathcal{T}^\dagger$诱导的分布相同。将这一想法推广，考虑一个序贯随机对照试验 $\mathcal{T}$，其中将一部分个体的组别分配进行置换，以构造一个假设试验 $\mathcal{T}^\dagger$，该试验在时间 t作出新的分配 $\mathrm{a}_t^\dagger$。如果某个个体截至时间 $t-1$ 经历的状态$s_{0:t-1}$和动作 $a_{1:t-1}$ 子轨迹在 $\mathcal{T}^\dagger$ 和 $\mathcal{T}$中完全相同，并且其收到的新分配 $a_t^\dagger$ 也与 a_t 完全相同，那么在 $\mathcal{T}$中观察到的原始状态样本 s_t，同样是 $\mathcal{T}^\dagger$中的有效样本，且来自相同的分布，$P^*(\mathrm{x}\ s_{0:t-1},\ a_{1:t}, s_t)$。此外，通过归纳，若 $\forall\, t \in [T]$ 的输入子轨迹 $s_{0:t-1}$产生的新分配 $a_t^\dagger$ 与 a_t完全相同，则 $\mathcal{T}$的整个原始状态轨迹 $s_{0:T}$ 可视为 $\mathcal{T}^\dagger$ 的有效样本。

我们利用这一概念，回溯检查所有此类可能的重新分配：在每个时刻，给定状态-动作轨迹中相同的输入子序列，这些重新分配会产生相同的动作输出序列。这意味着，我们可以在估计给定测试策略的性能时，找出并汇总多个此类额外的“可观测反事实分配”（下文定义）所产生的结果。算法 1 概述了适用于一般M-臂情形的思路。随后，在算法 2 中，我们将介绍一种专为处理基于指标的策略而设计的高效算法。

定义 4.1（可观测反事实分配）。对于实际分配 $\mathfrak{C}$，我们将 $\mathfrak{C}^\dagger$ 定义为可观测反事实分配：在一次假设试验中，若分配为 $\mathfrak{C}^\dagger$，则对于每个 $t = 1...T$，在状态和动作历史（$\mathbf{s}_{0:t-1}$，$\mathbf{a}_{1:t-1}$）截至时间 $t-1$ 之前匹配的条件下，各策略会分配给每个个体的动作，都与其最初接受的动作相同。

通过分配置换进行估计： 令 $\mathcal{C}^\dagger(\mathfrak{C})$ 表示由单个实际试验分配 $\mathfrak{C}$ 产生的所有“可观测反事实分配”构成的集合。我们提出的估计量对所有此类可观测反事实的结果取平均：

$$\text{Eval}^\dagger(\pi_m) := \frac{\sum_{\mathfrak{C}\in\mathcal{C}^\dagger} \text{Eval}(\pi_m|\mathfrak{C})}{|\mathcal{C}^\dagger|} \tag{3}$$

在定理 4.3中，我们证明这是对真实期望 Eval* 的无偏估计量。关键的技术步骤（引理 A.1）是证明 $\mathcal{C}^\dagger(\mathfrak{C})$ 在 $\mathcal{C}$（所有可能分配的集合）上定义了一个划分，其中两个分配 $\mathfrak{C}_1$，$\mathfrak{C}_2$ 当且仅当 $\mathcal{C}^\dagger(\mathfrak{C}_1) = \mathcal{C}^\dagger(\mathfrak{C}_2)$时属于同一部分。直观地说，这意味着我们的估计量不会对任何特定的反事实分配“过度加权”；它保持每个分配在 Eval* 中所占的相同权重。

算法 1通过分配置换进行估计**输入**：状态：$\{S_1^{N\times T+1},\ldots,S_M^{N\times T+1}\}$，动作：$\{A_1^{N\times T},\ldots A_M^{N\times T}\}$，分配，$\mathfrak{C}$：$\{C_1,\ldots,C_M\}$

输出：Eval†

1：计算 $\mathcal{C}^\dagger$，即 $\mathfrak{C}$ 的可观测反事实分配集合。2：计算 Eval†(π_m)：$= \frac{\sum_{\mathfrak{C}\in\mathcal{C}^\dagger}\text{Eval}(\pi_m|\mathfrak{C})}{|\mathcal{C}^\dagger|}$3：**返回**Eval†($\pi_m$)

4.3. 理论结果

在本节中，我们从理论上证明，Eval†(.) 能够给出更准确的估计，因为它是无偏的，同时还能降低方差。令 † 是定义在 $\mathcal{C}$ 上的齐次关系，定义如下：$\dagger = \{(\mathfrak{C}_1,\mathfrak{C}_2) \in \mathcal{C}\times\mathcal{C}$：$\mathfrak{C}_2 \in \mathcal{C}^\dagger(\mathfrak{C}_1)\}$。直观而言，† 表示存在一种有效的重新排列方式，能够从 $\mathfrak{C}_1$ 得到反事实分配 $\mathfrak{C}_2$。

引理 4.2。关系 † 是一个等价关系，而由 $\mathcal{C}^\dagger(\cdot)$ 定义的集合族构成 $\mathcal{C}$ 上的一个划分。

所有证明均可在附录中找到。我们利用这一性质来证明无偏性：

定理 4.3. $Eval^\dagger(\pi_m)$ 是性能期望值的无偏估计，$Eval^*(\pi_m)$ $\forall m \in [M]$，定义见方程 *1*。即

$$\mathbb{E}_{S_m\sim P^*}\mathbb{E}_{\mathfrak{C}\sim\mathcal{C}}[Eval^\dagger(\pi_m)] = Eval^*(\pi_m)\ \forall m\in[M]$$

定理 4.4。我们的估计量 $Eval^\dagger(\pi)$的样本方差小于标准估计量 $Eval(\pi)$：

$$\text{Var}(Eval(\pi)) - \text{Var}(Eval^\dagger(\pi)) =$$

$$\frac{1}{|\mathcal{C}|}\cdot\sum_{j\in[\eta]}\left[\sum_{\mathfrak{C}\in\mathcal{P}_j} Eval^2(\pi|\mathfrak{C}) - \frac{\left(\sum_{\mathfrak{C}\in\mathcal{P}_j} Eval(\pi|\mathfrak{C})\right)^2}{|\mathcal{P}_j|}\right]$$

≥ 0，其中 $\{\mathcal{P}_1,\ldots,\mathcal{P}_\eta\}$ 是由 †诱导的 $\mathcal{C}$ 划分。

证明概要。我们首先以划分 $\mathcal{P}_j$（由 †定义的等价集合构成）为条件，计算样本方差；该划分确定了分配 $\mathfrak{C}$ 的一个实例所属的集合，然后再考虑划分内候选分配 $\mathfrak{C}$ 所带来的方差。最后，我们利用柯西-施瓦茨不等式证明定理 4.4 右侧的表达式非负。 □

定理 4.4 中的方差收缩表达式当且仅当 Eval($\pi|\mathfrak{C}$)与 $\forall\mathfrak{C}\in\mathcal{P}_j,\ \forall\, j \in[\eta]$完全相同时才为零；如果不同的分配对应不同的奖励，那么我们的估计量就能严格降低方差。

5. 高效交换算法

穷举识别 $\mathcal{C}^\dagger$ 算法1 中的所有可能情况，需要遍历 $\mathcal{C}$ 中的每一种可能分配，并运行策略以确定该分配是否属于 $\mathcal{C}^\dagger$。然而，可能的分配数量 $|\mathcal{C}|$ 会随着 N呈指数级增长，使得完全枚举不可行。我们展示了如何通过修改估计量（记为 Eval）来绕过这一计算瓶颈，该估计量适用于基于指标的策略$\dagger_\Upsilon(\cdot)$。此估计量隐式地对 $\mathcal{C}^\dagger$中的部分可能排列取平均，以牺牲部分方差缩减为代价换取计算效率。

为便于说明，这里我们考虑具有两个实验臂的随机对照试验（$M = 2$），分别采用分配策略 π_0 和 π_1。我们的目标是估计 Eval*(π_j)$_{j=\{0,1\}}$。直观而言，我们不再处理所有可能分配构成的空间，而是考虑参与试验的所有个体，并识别出满足某些理想属性的不重叠个体组 $\{\mathbf{G}_k\}\subset(C_0\cup C_1)$。具体而言，我们旨在找到由“兼容”个体构成的集合，使得集合中任意个体子集

每个组 $\{\mathbf{G}_k\}$ 中的个体都可以相互交换，从而得到不变的分配，或在 $\mathcal{C}^\dagger(\mathfrak{C})$ 中得到有效且可观测的反事实分配。我们的意图是通过将每个个体 $i \in \mathbf{G}_k$ 的原始奖励替换为 $\mathbf{G}_k$中所有个体奖励的平均值，为每个这样的组 $\mathbf{G}_k$ 计算估计值（定理 5.1 中给出依据）。我们通过检查与个体可交换性相关的两个资格条件来识别这些组。

交换两个个体 i 和 j 的第一个资格条件是，它们原始的分配 $a(t)$必须完全相同 $\forall t$，这样交换后两个臂中的资源约束才能继续得到满足。为检查这一条件，我们将个体划分为超组_ _{$\mathbf{G}_1$，……，$\mathbf{G}_\kappa$}，将经历相同动作向量 $a(t) \in \mathcal{A}^T$的所有个体归入同一超组，其中 κ 表示此类超组的数量。每个 $\mathbf{G}_k$ 中的所有个体都满足被纳入组 $\mathbf{G}_k$的第一个资格条件。为方便起见，我们令ϕ： $C_0 \cup C_1 \to [\kappa]$ 表示一个多对一映射，用于标识个体 i 所属的超组 $\mathbf{G}^{\phi(i)}$ 。

交换个体的第二个资格条件是：对于与原始试验相同的一系列输入状态，其在新策略下的新分配 $a^\dagger(t)$必须与原始分配 $a(t)$完全相同。对于每个个体 i，我们使用一个二值变量 $\Lambda_i \in \{0,1\}$ 来表示其是否满足这一第二个条件。此处我们引入并利用“索引阈值”性质，以高效地验证该条件。我们将索引阈值 $\tau_j(t)$定义为时刻 t时，集合 C_j 中所有依据策略 π_j被选中接受分配 $a = 1$ 的个体的索引 $\Upsilon(t)$中的最小值。为实现高效计算，我们只允许在组 $\mathbf{G}_k$ 内进行交换，且交换后索引阈值 $\tau_j(t)$保持为与原始值相同。我们通过将索引阈值处所有个体的 $\Lambda_i = 0$设为相应值来实施这一约束。此外，对于其他个体 i，只需验证索引 $\Upsilon_i^{\pi_j}(t)$是否位于阈值 $\tau_j(t)$ $\forall\, t \in [T]$ 的同一侧以及是否满足 $j \in \{0,1\}$，即可轻松确定 Λ_i 。总而言之

$$\Lambda_i = \begin{cases} 1 & \text{if } \prod_{j=0}^{j=1} (\Upsilon_i^{\pi_j}(t) - \tau_j(t)) > 0 \;\forall\, t \in [T] \\ 0 & \text{otherwise} \end{cases} \tag{4}$$

直观地说，$\Lambda_i = 1$ 意味着$a_1^\dagger : T(i) = a_{1:T}(i)$，并表明个体 i 满足第二个资格条件。取两个条件的交集后，我们通过纳入所有具有$\Lambda_i = 1$的个体 $i \in \mathbf{G}_k$ 来构成¯组$\mathbf{G}_k$ 。对于每个组$\mathbf{G}_k$，我们计算代表性平均个体的奖励，如下所示：

$$\tilde{r}_k \coloneqq \frac{1}{|\mathbf{G}_k|} \sum_{i \in \mathbf{G}_k} r(S[i], A[i]) \tag{5}$$

在计算最终估计值 $\text{Eval}_\Upsilon^\dagger(\pi_j)$ 时，我们考虑臂 C_j 中的所有个体，但将以下个体的奖励替换为其中的可交换个体（即其 $\Lambda_i = 1$ 的个体）替换为$\tilde{r}_{\phi(i)}$。其他个体的奖励保持不变。最后，我们通过求和计算 $\text{Eval}_\Upsilon^\dagger(\pi_j)$，如下所示：

$$\text{Eval}_\Upsilon^\dagger(\pi_j) = \sum_{i \in C_j} \Big(\Lambda_i \tilde{r}_{\phi(i)} + (1 - \Lambda_i) r(S[i], A[i]) \Big) \tag{6}$$

我们对该估计量的理论分析表明，它对应于一般置换估计量的一个实例，该估计量对 $\mathcal{C}^\dagger$ 中的一部分分配进行平均（而不是对整个集合进行平均）。其主要思想是，我们可以将 $\mathcal{C}^\dagger$ 进一步划分为索引阈值取值相同的集合（具体形式见引理A.2）。我们将 $\mathcal{C}^\dagger$ 中索引阈值与实际试验相同的部分记为 $\mathcal{C}_\Upsilon^\dagger$。组内个体的每一种置换 $\{\mathbf{G}_k\}$ 都对应于 $\mathcal{C}_\Upsilon^\dagger$中的一个分配（$\mathfrak{C}$），最终的估计量对所有此类分配取平均：

定理 5.1。 *$Eval_\Upsilon^\dagger$（·）根据公式6计算得出，等于 $\mathcal{E}val$ 的平均值$_\mathcal{C}$ π $\mathcal{E}$ 在 $\mathcal{C}_\Upsilon^\dagger$中的所有分配。*

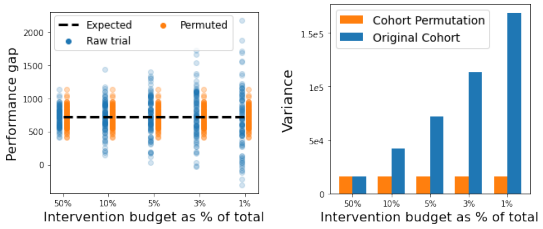
i.e. $Eval_\Upsilon^\dagger(\pi) = \frac{\sum_{\mathfrak{C} \in \mathcal{C}_\Upsilon^\dagger} Eval(\pi|\mathfrak{C})}{|\mathcal{C}_\Upsilon^\dagger|}$

由此，$\text{Eval}_\Upsilon^\dagger$（$\pi_j$）很容易证明继承了一般置换估计量的理想性质，例如，推论 A.3证明了它同样是 Eval 的无偏估计量*（π_j）。其代价是在方差收缩方面略有牺牲，因为它产生了更小的划分 $\mathcal{P}_j$，即 $\mathcal{C}$（如定理 4.4中所定义），这是因为我们舍弃了索引阈值不同的分配。然而，使用 $\mathcal{C}_\Upsilon^\dagger$ 可以为索引策略实现计算高效的算法，从而避免一般估计量的指数运行时间。

算法 2 索引策略之间的重新置换 输入： $\mathfrak{C} := \{C_0, C_1\}$，状态： S_0、 $S_1 \in \mathcal{S}^{N \times T+1}$，动作：$A_0$、$A_1 \in \{0,1\}^{N \times T}$，索引： Υ^{π_0}、 $\Upsilon^{\pi_1} \in \mathbb{R}^{2N \times T}$输出：估计值：$\text{Eval}_\Upsilon^\dagger$1：根据动作向量将个体分组为_{$\mathbf{G}_k$}，并确定 $\phi(i)$ $\forall$ 个体 $i \in C_0 \cup C_1$。 2：确定 $\forall t \in [T]$，两条臂上的索引阈值为：$\bar{\tau}_j(t)$：$=\min \{\Upsilon_{\pi_j}(i,t) \;|A_j[i,t] = 1$、 $i \in [N]\}$3：根据公式 4 确定 Λ_i $\forall$ 个体 i。4：计算组 $\mathbf{G}_k$： $\{i \in \mathbf{G}_k \,|\Lambda_i = 1\}$、 $\forall k \in [\kappa]$5：对于每个组 $\mathbf{G}_k$，根据公式 5 计算代表性个体的平均奖励。6：根据公式 6 计算 $\text{Eval}_\Upsilon^\dagger(\pi_j)$

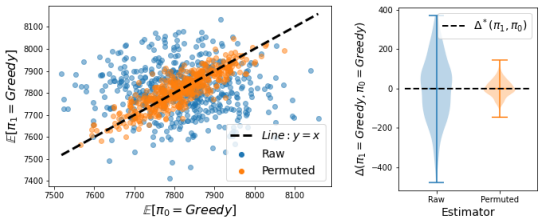
6. 实证评估

我们通过多个数据集对所提出的方法进行实证检验：(1) 合成示例；(2) 半合成结核病用药依从性监测数据；以及 (3) 来自孕产妇医疗保健干预项目的真实世界现场试验数据。



(a) 总参与人数 (b) 样本方差

图 *1*。合成领域中的估计值和样本方差。



(a) (b)

图*2*。多步设置。(a) 置换后的估计值（橙色）比原始估计值（蓝色）更接近真实期望（黑线）。(b) 分配置换降低了方差。

我们考虑一个状态空间 $\mathcal{S} = \{0, 1\}$，分别表示（健康、项目参与等方面的）“不理想”状态和“理想”状态。行动空间 $\mathcal{A} = \{0, 1\}$ 表示“不提供”或“提供”干预。我们假设存在预算约束，以限制每个时间步的干预总次数。奖励函数定义为$r(s_{0:T},\ a_{1:T}) = \sum_t s_t$，对应的目标是最大化个体处于状态 $s = 1$的总时间。

6.1. 合成数据集

该设置包含三种类型的个体，其特征由它们的 $P-$ 矩阵刻画。 P_1 和 P_2 经过设计，使得始终最优的做法是持续对 P_1 个体进行干预，而对 P_2 个体进行干预则严格次优（详情见附录 C.1）。 P_3 个体不受干预影响，其转移动态与所接收的动作无关。我们考虑两种测试策略π_1 和 π_2，其设计使得策略π_j 在 P_j 类型个体可用时总是选择对其进行干预，从而使 π_1 成为最优策略。我们模拟 300 P_1个体、 300 P_2 个体和（300*η P_3）个体，并将每个时间步的干预预算设为 300，持续 $T = 20$ 个时间步。我们衡量 π_1相对于 π_2 的性能提升，定义如下： $\Delta := \Delta(\pi_1, \pi_2) = \text{Eval}(\pi_1) - \text{Eval}(\pi_2)$。

在图1a中，我们改变横轴上的预算。散点图中的每个蓝点代表一个独立的随机对照试验实例，并测量奖励的原始差值 Δ （纵轴）。应用分配置换后，每个蓝点都会映射为一个橙点。黑色虚线表示性能提升的期望值。从视觉上看，两种颜色都以黑线为中心，因为这两种估计量都是无偏的。分配置换后的估计值比原始估计值更接近期望值，表明其样本方差更小。具体而言，我们测量图 Δ 中纵轴上1b的样本方差。应用分配置换后，方差大幅降低——例如，当预算水平为3%时，我们的方法将方差降低了 7×，从 11.3×10^4 降至 1.6×10^4。其直观解释是， π_1和 π_2 在决定不干预 P_3 个体时存在重叠。然而，它们的最终奖励部分取决于哪个 P_3 个体被（随机）分配到哪个组，与底层策略无关。分配置换通过对这些 P_3 个体的其他分配方式取平均，抵消这种随机性。

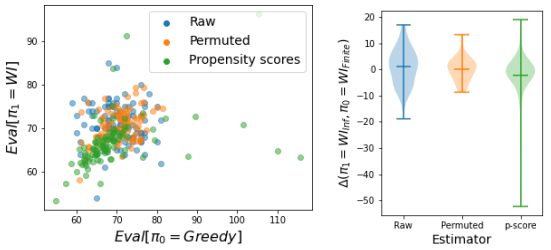
6.2. 使用结核病数据集进行半合成评估

我们使用真实的结核病用药依从性监测数据，其中包含印度孟买患者的每日记录，数据取自（Killian 等人，2019年），并通过估计 P矩阵来模拟患者行为。更多详情请参阅附录。

我们考虑两种策略：“Whittle 指数”策略（Whittle，1988），该策略旨在最大化长期奖励；以及贪心策略，它优化下一步奖励的估计值。我们在每个臂中模拟 $N = 1000$ 名患者，并改变预算约束。我们同时考虑多步设置（$T = 10$）和单步设置（$T = 1$）。我们比较三种估计方法：“原始”是各臂结果的朴素平均值，“置换”是我们提出的估计量，“逆概率加权”是第4.1节中的逆倾向评分估计量（仅适用于 $T = 1$）。在实践中，我们发现有必要将逆概率加权的倾向得分截断到Zhou 等人，2020年提出的范围 [0.01,0.99],，因为极端值会导致方差非常大。这会引入轻微偏差，可在图3b中观察到。

表1 展示了各方法所得估计值的样本方差。对于无偏方法（原始、置换），方差也等于其均方误差；由于进行了截尾，这一结论对逆概率加权近似成立。表 1 还包含一列，标题为“$n-$值”。为评估我们的方法所带来的改进，该列给出了达到分配置换（通过模拟计算）所得样本方差所需运行（并取平均）的最少相互独立的随机对照试验数。

在所有比较和参数设置下，我们发现分配置换估计量显著降低了方差。事实上，若使用朴素原始估计量，要达到同等精度的估计，需要开展 2 到 *13* 项独立的随机对照试验。这凸显了方差缩减的重要性——开展随机对照试验的成本极高，而且会将许多



(a) (b)

图3. 单步设置结果示意图

表1： 测得的性能提升的样本方差

T	B	π_1 V π_0	RAW	置换	IPW	n-VAL
1	3%	π_{WI} V π_{GR}	49.09	4.94	**0.48**	9
1	10%	π_{WI} V π_{GR}	49.86	15.11	**6.66**	3
1	25%	π_{WI} V π_{GR}	49.45	**19.94**	78.12	2
10	3%	π_{WI} V π_{WI}	2381	**916**	NA	3
10	3%	π_{WI} V π_{GR}	2348	**728**	NA	4
10	3%	π_{GR} V π_{GR}	26356	**1860**	NA	13
10	10%	π_{GR} V π_{GR}	25983	**3808**	NA	7
10	25%	π_{GR} V π_{GR}	23619	**5477**	NA	5

个体分配到次优策略；改进分析方法则能让我们以大幅降低的成本得出同等精确的结论。

图2 通过一个示例说明了这一改进：两个试验组均采用贪心策略，因此奖励的预期差异恰好为零（其中 $B = 3\%$和 $T = 10$）。图 2a中的每个点都对应一次试验实例，横轴和纵轴分别表示两个试验组的总参与度。黑色虚线（$y = x$）表示期望值 $\Delta = 0$。代表原始测量值的蓝点在黑线周围分布得比通过分配置换得到的橙点更分散。图 2b展示了两个试验组之间奖励样本差异的小提琴图。两个小提琴图都以零线为中心，表明这些估计量是无偏的。对应分配置换的小提琴图更紧凑，说明样本方差更低。

在单步设置中，逆概率加权估计量的表现喜忧参半：预算较小时，其方差最低；但预算较大时，表现实际上不如朴素原始估计量。其根本原因是，随着预算增加，两种策略之间的重叠程度会降低，因为它们只在少数优先级最高的个体上达成一致。重叠程度低会导致倾向得分极端，从而增大方差。图 3b 和 3a展示了一个示例：逆概率加权通常能带来改进，但也容易受到较大离群值的影响。不过，当重叠程度较高且仅处于单步设置时，逆概率加权仍是一个值得考虑的选项。

6.3. 案例研究：真实世界试验

我们的方法可直接应用于真实世界场景；为此，我们考察了（Mate 等人，2022年）报告的一项实际大规模随机对照试验，该试验评估了一种基于静息多臂老虎机的资源分配算法在妇幼医疗保健中的应用。数据包含 23,000 名真实世界受益人，他们在试验中被随机分为三组：RMAB算法组、基线算法组和对照组；对照组不接受任何干预。真实世界的健康工作者实施了算法推荐的干预措施。

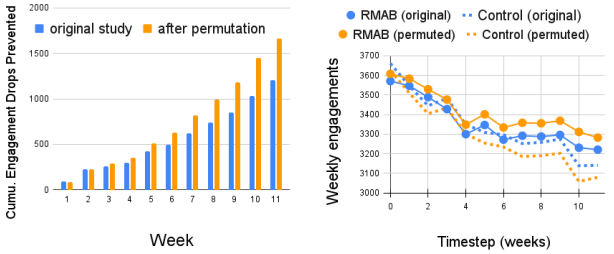


图4。置换估计量对真实世界数据的影响。

我们考察了 RMAB算法相较于对照组在提升项目参与度方面的性能提升，并将我们提出的置换算法应用于最初报告的原始结果。图 4（左）以横轴表示时间（周），纵轴表示参与度总数。图 4 （右）根据（Mate 等人，2022年）中的定义，在纵轴上计算 RMAB算法带来的性能提升。我们的研究结果表明，RMAB算法的性能提升高于最初报告的结果；到第 7 周时，估计 RMAB可避免 815 次参与度下降，而最初报告的数字为 622。由于这是实测数据，真实数值未知。不过，这项案例研究表明，我们的估计量所带来的方差缩减在实际应用中可能具有显著效果。

7. 结论

我们致力于弥补一个关键缺口：通过随机对照试验评估资源分配策略时，如何降低误差。我们提出了一种新的估计量，其依据是一项新颖理念：将参与者回顾性地重新分配至实验组。我们证明，该估计量在保持无偏的同时降低了样本方差，从而减少误差。通过在多个数据集上进行实证检验——其中包括一个来自社会关键领域的真实世界数据集——我们表明，该方法最多可将误差降低 70%；而且仅凭一项随机对照试验，就能实现相当于并行开展多达 13项独立的随机对照试验所带来的效益。

参考文献

Abadie, A., Athey, S., Imbens, G. W. 和 Wooldridge, J.M. 何时应针对聚类调整标准误？《经济学季刊》，138(1):1–35，2022年。

Angrist, J. D. 和 Pischke, J.-S. 基本无害的经济学计量：一位经验主义者的伴侣。普林斯顿大学出版社，2009年。

Athey, S. 和 Wager, S. 基于观察性数据的策略学习。《计量经济学》，89(1):133–161，2021年。

Austin, P. C. 和 Stuart, E. A. 使用倾向得分估计观察性研究中的因果治疗效应时，改进逆概率治疗加权（iptw）的最佳实践。《医学统计学》，34(28):3661–3679，2015年。Biswas, A.、Aggarwal, G.、Varakantham, P. 和 Tambe, M. 学习干预：一种用于预防性医疗的restless多臂老虎机自适应学习策略。载于《第*30*届国际人工智能联合会议论文集》，2021年。Eagle, N.、Macy, M. 和 Claxton, R. 网络多样性与经济发展。《科学》，328(5981):1029–1031，2010年。doi: 10.1126/science.1186605。Enderton, H. B. 集合论基础。学术出版社，1977年。Galstyan, A.、Czajkowski, K. 和 Lerman, K. 使用强化学习在电网上进行资源分配。载于《第三届国际自主智能体与多智能体系统联合会议论文集，*2004*年。*AAMAS 2004*。》，第1卷，第1314–1315页。IEEE计算机学会，2004年。Hayes, R. J. 和 Moulton, L. H. 整群随机试验。Chapman and Hall/CRC，2017年。Hern´an, M. A. 和 Robins, J. M. 因果推断，2010年。Hudgens, M. G. 和 Halloran, M. E. 迈向存在干扰情况下的因果推断。《美国统计学会杂志》，103(482):832–842，2008年。Imai, K. 和 Li, M. L. 个体化治疗规则的实验评估。《美国统计学会杂志》，第1–15页，2021年。Jamieson, K.、Malloy, M.、Nowak, R. 和 Bubeck, S. lil'ucb：一种针对多臂老虎机的最优探索算法。载于学习理论会议，第423–439页。PMLR，2014年。Jiang, N. 和 Li, L. 强化学习的双重稳健离策略价值评估。载于国际机器学习会议，第652–661页。PMLR，2016年。

Killian, J. A., Wilder, B., Sharma, A., Choudhary, V., Dilk- ina, B., and Tambe, M. 利用数字依从性数据为结核病患者开具干预措施处方的学习方法。载于*KDD*，2019年。Kuss, O., Blettner, M., and B¨orgermann, J. 倾向得分：一种分析治疗效果的替代方法：科学出版物评估系列第23篇。¨《*Deutsches Arzteblatt International*》，113(35-36): 597，2016年。Lee, E., Lavieri, M. S., and Volk, M. 肝细胞癌的最优筛查：一种休眠多臂老虎机模型。《制造与服务运营管理》，21 (1):198–212，2019年。Leskovec, J., Adamic, L. A., and Huberman, B. A. 病毒式营销的动态。《*ACM* 网络汇刊》，1(1): 5–es，2007年5月。ISSN 1559-1131。doi: 10.1145/1232722.1232727。

U

R

L

h

ttps://doi.org/10.1145/1232722.1232727。

Li, L., Munos, R., and Szepesv´ari, C. 探索极小极大离策略价值估计。载于人工智能与统计，第608–616页。PMLR，2015年。Liu, K. and Zhao, Q. 休眠多臂老虎机问题的可指标性及 Whittle 指数在动态多信道接入中的最优性。《*IEEE* 信息论汇刊》，第5547–5567页，2010年。Mate, A., Killian, J. A., Xu, H., Perrault, A., and Tambe, M. 坍缩多臂老虎机及其在公共卫生干预中的应用。载于神经信息处理系统进展（*NeurIPS*），2020年。Mate, A., Perrault, A., and Tambe, M. 公共卫生中的风险感知干预：使用休眠多臂老虎机进行规划。自主智能体与多智能体系统（*AAMAS*），2021年。Mate, A., Madaan, L., Taneja, A., Madhiwalla, N., Verma, S., Singh, G., Hegde, A., Varakantham, P., and Tambe, M. 部署休眠多臂老虎机的实地研究：协助非营利组织改善妇幼健康。2022年。Mukerjee, R., Dasgupta, T., and Rubin, D. B. 运用有限总体抽样的标准工具，改进复杂实验中的因果推断。《美国统计学会杂志》，113(522):868–881，2018年。Rubin, D. B. 运用潜在结果进行因果推断：设计、建模与决策。《美国统计学会杂志》，100(469):322–331，2005年。

Sutton, R. S., Maei, H., and Szepesv´ari, C. 一种用于线性函数逼近离策略学习的收敛型$o(n)$时序差分算法。神经信息处理系统进展，21，2008年。Swaminathan, A., Krishnamurthy, A., Agarwal, A., Dudik, M., Langford, J., Jose, D., and Zitouni, I. 列表推荐的离策略评估。神经信息处理系统进展，30，2017年。Tesauro, G., Jong, N. K., Das, R., and Bennani, M. N. 一种用于自主资源分配的混合强化学习方法。载于*2006*年 *IEEE*自主计算国际会议，第65–73页。IEEE，2006年。Ustun, B. and Rudin, C. 学习优化风险评分。《机器学习研究杂志》，20(150):1–75，2019年。Vaswani, S., Lakshmanan, L., Schmidt, M., et al. 利用多臂老虎机实现影响力最大化。*arXiv*预印本 *arXiv:1503.00024*，2015年。Wang, Y.-X., Agarwal, A., and Dudık, M. 上下文多臂老虎机中的最优自适应离策略评估。载于国际机器学习会议，第3589–3597页。PMLR，2017年。Whittle, P. 休眠多臂老虎机：变化世界中的活动分配。《应用概率杂志》，25(A):287–298，1988年。Wilder, B., Onasch-Vera, L., Diguiseppi, G., Petering, R., Hill, C., Yadav, A., Rice, E., and Tambe, M. 一项针对无家可归青年的艾滋病毒预防人工智能增强干预临床试验。载于*AAAI*人工智能会议论文集，第35卷，第14948–14956页，2021年。Zhou, Y., Matsouaka, R. A., and Thomas, L. 在重叠有限且模型设定错误情况下的倾向评分加权。医学研究统计方法，29(12):3721–3756，2020年。

A. 理论结果的完整证明

A.1. 引理 A.1 的证明A.1

引理 A.1. 关系 † 是一个等价关系，而由 $\mathcal{C}^{\dagger}(\cdot)$ 定义的集合族构成了 $\mathcal{C}$ 上的一个划分。

证明。为了证明 † 是一个等价关系，我们证明它具有自反性、对称性和传递性。根据定义，† 具有自反性，因为 $\forall \mathfrak{C} \in \mathcal{C}$，$\mathfrak{C} \in \mathcal{C}^{\dagger}(\mathfrak{C})$。此外，† 显然也具有对称性，因为如果 $\mathfrak{C}_2 \in \mathcal{C}^{\dagger}(\mathfrak{C}_1)$，那么根据定义，在 $\mathfrak{C}_1$ 和 $\mathfrak{C}_2$ 下，所有个体在所有时刻获得的分配都相同。因此，$\mathfrak{C}_1 \in \mathcal{C}^{\dagger}(\mathfrak{C}_2)$。最后，† 也具有传递性，因为如果所有个体在所有时刻获得的分配在 $\mathfrak{C}_1$ 和 $\mathfrak{C}_2$ 中相同，并且在 $\mathfrak{C}_2$ 和 $\mathfrak{C}_3$ 中也相同，这意味着这些分配在 $\mathfrak{C}_1$ 和 $\mathfrak{C}_3$ 中也相同。因此，形式上，如果 $\mathfrak{C}_2 \in \mathcal{C}^{\dagger}(\mathfrak{C}_1)$ 且 $\mathfrak{C}_3 \in \mathcal{C}^{\dagger}(\mathfrak{C}_2)$，那么 $\mathfrak{C}_3 \in \mathcal{C}^{\dagger}(\mathfrak{C}_1)$。由此，† 是 $\mathcal{C}$ 上的一个等价关系，因此将 $\mathcal{C}$ 划分为一个等价类集合族$\mathcal{C}^{\dagger}(\cdot)$，使得每个元素 $\mathfrak{C} \in \mathcal{C}$ 恰好属于一个划分（Enderton，1977年）。 □

A.2. 定理 4.3 的证明

定理 4.3。 $Eval^{\dagger}(\pi_m)$ 是性能的期望值 $Eval^{*}(\pi_m)$ $\forall m \in [M]$的无偏估计，定义见公式 *1*。即

$$\mathbb{E}_{S_m \sim P^*} \mathbb{E}_{\mathfrak{C} \sim \mathcal{C}}[Eval^{\dagger}(\pi_m)] = Eval^{*}(\pi_m) \, \forall m \in [M]$$

证明。

$$\begin{aligned}
\text{Eval}^{*}(\pi_m) =& \mathbb{E}_{S_m \sim P^*}\Big[\mathbb{E}_{\mathfrak{C} \sim \mathcal{C}}\big[\text{Eval}(\pi_m)\big]\Big] \\
=& \mathbb{E}_{S_m \sim P^*}\Big[\sum_{\mathfrak{C} \in \mathcal{C}} \text{Prob}[\mathfrak{C}] \cdot \text{Eval}(\pi_m|\mathfrak{C})\Big] \\
=& \mathbb{E}_{S_m \sim P^*}\Big[\frac{1}{|\mathcal{C}|}\sum_{\mathfrak{C} \in \mathcal{C}} \text{Eval}(\pi_m|\mathfrak{C})\Big] && (\because \text{all } \mathfrak{C} \text{ equally likely}) \\
=& \mathbb{E}_{S_m \sim P^*}\left[\frac{1}{|\mathcal{C}|}\Big[\sum_{\mathfrak{C} \in \mathcal{P}_1} \text{Eval}(\pi_m|\mathfrak{C}) + \cdots + \sum_{\mathfrak{C} \in \mathcal{P}_\eta} \text{Eval}(\pi_m|\mathfrak{C})\Big]\right] \\
& \text{where } \{\mathcal{P}_1, \ldots, \mathcal{P}_\eta\} \text{ defines partition of } \mathcal{C} \text{ induced by } \dagger. \\
=& \mathbb{E}_{S_m \sim P^*}\left[\sum_{j \in [\eta]} \frac{|\mathcal{P}_j|}{|\mathcal{C}|} \cdot \frac{1}{|\mathcal{P}_j|} \cdot \Big[\sum_{\mathfrak{C} \in \mathcal{P}_j} \text{Eval}(\pi_m|\mathfrak{C})\Big]\right] \\
=& \mathbb{E}_{S_m \sim P^*}\left[\sum_{j \in [\eta]} \frac{|\mathcal{P}_j|}{|\mathcal{C}|} \cdot \Big[\text{Eval}^{\dagger}(\pi_m|\mathfrak{C})\Big]\right] && (\forall \mathfrak{C} \in \mathcal{P}_j) \\
=& \mathbb{E}_{S_m \sim P^*}\left[\sum_{j \in [\eta]} \text{Prob}[\mathcal{P}_j] \cdot \Big[\text{Eval}^{\dagger}(\pi_m|\mathfrak{C})\Big]\right] && (\forall \mathfrak{C} \in \mathcal{P}_j) \\
=& \mathbb{E}_{S_m \sim P^*}\left[\sum_{j \in [\eta]} \sum_{\mathfrak{C} \in \mathcal{P}_j} \text{Prob}[\mathfrak{C}] \cdot \Big[\text{Eval}^{\dagger}(\pi_m|\mathfrak{C})\Big]\right] && \Big(\because \text{Prob}[\mathcal{P}_j] = \sum_{\mathfrak{C} \in \mathcal{P}_j} \text{Prob}[\mathfrak{C}]\Big) \\
=& \mathbb{E}_{S_m \sim P^*} \mathbb{E}_{\mathfrak{C} \sim \mathcal{C}}[\text{Eval}^{\dagger}(\pi_m)]
\end{aligned}$$

□

A.3. 定理 4.4 的证明

定理 4.4。T　我们的估计量的样本方差 $Eval^{\dagger}(\pi)$ 小于标准估计量 $Eval(\pi)$：

$$\begin{aligned}
&\text{Var}(Eval(\pi)) - \text{Var}(Eval^{\dagger}(\pi)) = \\
&\frac{1}{|\mathcal{C}|} \cdot \sum_{j \in [\eta]} \left[\sum_{\mathfrak{C} \in \mathcal{P}_j} Eval^2(\pi|\mathfrak{C}) - \frac{\Big(\sum_{\mathfrak{C} \in \mathcal{P}_j} Eval(\pi|\mathfrak{C})\Big)^2}{|\mathcal{P}_j|}\right]
\end{aligned}$$

≥ 0，其中 $\{\mathcal{P}_1, \ldots, \mathcal{P}_\eta\}$ 是由 †诱导的 $\mathcal{C}$ 划分。

证明。 我们首先对实例所属的分配 $\mathfrak{C}$ 的划分 $\mathcal{P}_j$（由 †定义的等价集合）进行条件化，再计算划分内候选分配 $\mathfrak{C}$ 所产生的方差，从而求得样本方差。因此，我们得到：

$$\begin{aligned}
\mathrm{Var}(\mathrm{Eval}(\pi)) &= \frac{1}{|\mathcal{C}|}\sum_{\mathfrak{C}\in\mathcal{C}}\big(\mathrm{Eval}(\pi|\mathfrak{C}) - \mathrm{Eval}^*(\pi)\big)^2 \\
&= \frac{1}{|\mathcal{C}|}\sum_{j\in[\eta]}\sum_{\mathfrak{C}\in\mathcal{P}_j}\big(\mathrm{Eval}(\pi|\mathfrak{C}) - \mathrm{Eval}^*(\pi)\big)^2 \\
&= \frac{1}{|\mathcal{C}|}\sum_{j\in[\eta]}\left(\sum_{\mathfrak{C}\in\mathcal{P}_j}\big(\mathrm{Eval}(\pi|\mathfrak{C})\big)^2 - 2\,\mathrm{Eval}^*(\pi)\sum_{\mathfrak{C}\in\mathcal{P}_j}\mathrm{Eval}(\pi|\mathfrak{C}) + |\mathcal{P}_j|\big(\mathrm{Eval}^*(\pi)\big)^2\right) \qquad (7)
\end{aligned}$$

类似地，我们计算估计量 $\mathrm{Eval}^\dagger$ 的方差如下：

$$\begin{aligned}
\mathrm{Var}(\mathrm{Eval}^\dagger(\pi)) &= \frac{1}{|\mathcal{C}|}\sum_{j\in[\eta]}\sum_{\mathfrak{C}\in\mathcal{P}_j}\big(\mathrm{Eval}^\dagger(\pi|\mathfrak{C}) - \mathrm{Eval}^*(\pi)\big)^2 \\
&= \frac{1}{|\mathcal{C}|}\sum_{j\in[\eta]}\left(|\mathcal{P}_j|\cdot\left\{\frac{\sum_{\mathfrak{C}\in\mathcal{P}_j}\mathrm{Eval}(\pi|\mathfrak{C})}{|\mathcal{P}_j|} - \mathrm{Eval}^*(\pi)\right\}^2\right) \\
&= \frac{1}{|\mathcal{C}|}\sum_{j\in[\eta]}\left(|\mathcal{P}_j|\cdot\left\{\left(\frac{\sum_{\mathfrak{C}\in\mathcal{P}_j}\mathrm{Eval}(\pi|\mathfrak{C})}{|\mathcal{P}_j|}\right)^2 - 2\,\mathrm{Eval}^*(\pi)\left(\frac{\sum_{\mathfrak{C}\in\mathcal{P}_j}\mathrm{Eval}(\pi|\mathfrak{C})}{|\mathcal{P}_j|}\right) + \big(\mathrm{Eval}^*(\pi)\big)^2\right\}\right) \\
&= \frac{1}{|\mathcal{C}|}\sum_{j\in[\eta]}\left\{\frac{\left(\sum_{\mathfrak{C}\in\mathcal{P}_j}\mathrm{Eval}(\pi|\mathfrak{C})\right)^2}{|\mathcal{P}_j|} - 2\,\mathrm{Eval}^*(\pi)\sum_{\mathfrak{C}\in\mathcal{P}_j}\mathrm{Eval}(\pi|\mathfrak{C}) + |\mathcal{P}_j|\big(\mathrm{Eval}^*(\pi)\big)^2\right\} \qquad (8)
\end{aligned}$$

将式8中的表达式减去式7 中的表达式，得到：

$$\mathrm{Var}(\mathrm{Eval}(\pi)) - \mathrm{Var}(\mathrm{Eval}^\dagger(\pi)) = \frac{1}{|\mathcal{C}|}\sum_{j\in[\eta]}\left(\sum_{\mathfrak{C}\in\mathcal{P}_j}\big(\mathrm{Eval}(\pi|\mathfrak{C})\big)^2 - \frac{\left(\sum_{\mathfrak{C}\in\mathcal{P}_j}\mathrm{Eval}(\pi|\mathfrak{C})\right)^2}{|\mathcal{P}_j|}\right) \qquad (9)$$

由柯西-施瓦茨不等式可直接得出，式 9 中推导出的方差收缩表达式非负。柯西-施瓦茨不等式指出，对于两个向量 u 和 v，$|\langle u, v\rangle|^2 \leq \langle u, u\rangle \cdot \langle v, v\rangle$，其中 $\langle\cdot,\cdot\rangle$ 表示内积。令 $u = \left[\mathrm{Eval}(\pi|\mathfrak{C}_1)\ldots, \mathrm{Eval}(\pi|\mathfrak{C}_{|\mathcal{P}_j|})\right]$，且令 $v = \left[1, \ldots, 1\right]$，即可得到所需的$|\mathcal{P}_j|$项$|\mathcal{P}_j|$ $1's$

结果：$\mathrm{Var}(\mathrm{Eval}(\pi)) - \mathrm{Var}(\mathrm{Eval}^\dagger(\pi)) \geq 0$。 □

A.4. 高效算法

引理 A.2。 关系 $\dagger_\Upsilon$ 既是 $\mathcal{C}$ 上的等价关系，也是集合族 $\mathcal{C}^\dagger(\cdot)$ 中每个集合上的等价关系，而由 $\mathcal{C}^\dagger_\Upsilon(\cdot)$ 定义的集合族构成了 $\mathcal{C}$上的划分。

证明。与引理 A.1 类似，我们通过证明 $\dagger_\Upsilon$ 具有自反性、对称性和传递性，来说明它是一个等价关系。根据定义， $\dagger_\Upsilon$ 具有自反性，因为 $\forall\mathfrak{C}\in\mathcal{C}, \mathfrak{C}\in\mathcal{C}^\dagger_\Upsilon(\mathfrak{C})$。此外， † 显然也具有对称性，因为如果$\mathfrak{C}_2 \in \mathcal{C}^\dagger(\mathfrak{C}_1)$，那么根据定义，在 $\mathfrak{C}_1$ 和 $\mathfrak{C}_2$下，索引阈值以及所有个体在所有时刻获得的分配都相同。因此， $\mathfrak{C}_1 \in \mathcal{C}^\dagger(\mathfrak{C}_2)$。最后， † 也具有传递性，因为如果在 $\mathfrak{C}_1$ 和 $\mathfrak{C}_2$ 中，以及在 $\mathfrak{C}_2$ 和 $\mathfrak{C}_3$中，所有索引阈值和所有个体在所有时刻获得的分配都相同，那么在 $\mathfrak{C}_1$ 和 $\mathfrak{C}_3$中，这些也都相同。因此，形式化地说，如果 $\mathfrak{C}_2 \in \mathcal{C}^\dagger(\mathfrak{C}_1)$且 $\mathfrak{C}_3 \in \mathcal{C}^\dagger(\mathfrak{C}_2)$，那么 $\mathfrak{C}_3 \in \mathcal{C}^\dagger(\mathfrak{C}_1)$。因此， $\dagger_\Upsilon$ 是定义在 $\mathcal{C}$ 上的等价关系，进而将 $\mathcal{C}$ 划分为一族等价类 $\mathcal{C}^\dagger(\cdot)$，使得每个元素 $\mathfrak{C}\in\mathcal{C}$ 都属于

且仅属于一个划分（Enderton，1977年）。类似的推理也表明，$\dagger_\Upsilon$ 是定义在该族中每个集合上的等价关系 $\mathcal{C}^\dagger(\cdot)$。

□

推论 A.3. *$Eval^\dagger_\Upsilon(\pi_m)$ 是对式 $^*(\pi)$中定义的性能期望值的无偏估计，其中 1。即 $\mathbb{E}_{S_m\sim P^*}\mathbb{E}_{\mathfrak{C}\sim\mathcal{C}}[Eval^\dagger(\pi_m)] = Eval^*(\pi_m)\ \forall m \in [M]$*

证明。 利用引理 A.2，我们对由 $\dagger_\Upsilon$ 定义的划分采用与定理 4.3 中类似的论证，以证明 $\dagger_\Upsilon$会产生无偏估计。

□

定理 5.1. *$Eval^\dagger_\Upsilon(\cdot)$ 按公式 6 计算，得到所有分配中的 $(\pi|\mathfrak{C})$ 的平均值，其中 $\mathcal{C}^\dagger_\Upsilon$。即 $Eval^\dagger_\Upsilon(\pi) = \frac{\sum_{\mathfrak{C}\in\mathcal{C}^\dagger_\Upsilon} Eval(\pi|\mathfrak{C})}{|\mathcal{C}^\dagger_\Upsilon|}$*

证明。 要证明两者等价，关键在于将个体的（修正后）奖励求和，理解为对各项分配的群体平均奖励进行求和。从数学上说，从 $\text{Eval}^\dagger_\Upsilon(\pi_m)$ 的定义出发，关键是将对 $\mathcal{C}^\dagger_\Upsilon$ 中各项分配的求和运算，从 $\text{Eval}^\dagger_\Upsilon()$ 项外移至项内，并分别作用于每个作出贡献的参与者。形式上，我们可以将 $\text{Eval}^\dagger_\Upsilon(\pi_m)$ 的表达式重写为：

$$\text{Eval}^\dagger_\Upsilon(\pi_j) = \frac{\sum_{\mathfrak{C}\in\mathcal{C}^\dagger_\Upsilon} \text{Eval}(\pi_j|\mathfrak{C})}{|\mathcal{C}^\dagger_\Upsilon|} \tag{10}$$

$$= \frac{\sum_{\mathfrak{C}\in\mathcal{C}^\dagger_\Upsilon}\sum_{i\in C_j} r(S[i],A[i])}{|\mathcal{C}^\dagger_\Upsilon|} \tag{11}$$

$$= \sum_{\mathfrak{C}\in\mathcal{C}^\dagger_\Upsilon} \frac{\sum_{k\in\kappa}\sum_{i\in\mathbf{G}_k}\mathbb{1}_{\{i\in C_j\}}\cdot r(S[i],A[i]) + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i],A[i])}{|\mathcal{C}^\dagger_\Upsilon|} \tag{12}$$

（将求和拆分为$\mathbf{G}_k$ 组，以及其他无法互换的个体） (13)

$$= \sum_{k\in\kappa}\sum_{i\in\mathbf{G}_k}\left[\sum_{\mathfrak{C}\in\mathcal{C}^\dagger_\Upsilon}\frac{\mathbb{1}_{\{i\in C_j\}}\cdot r(S[i],A[i])}{|\mathcal{C}^\dagger_\Upsilon|}\right] + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i],A[i]) \tag{14}$$

$$= \sum_{k\in\kappa}\sum_{i\in\mathbf{G}_k}\left[\sum_{\mathfrak{C}\in\mathcal{C}^\dagger_\Upsilon}\frac{\mathbb{1}_{\{i\in C_j\}}}{|\mathcal{C}^\dagger_\Upsilon|}\cdot r(S[i],A[i])\right] + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i],A[i]) \tag{15}$$

$$= \sum_{k\in\kappa}\sum_{i\in\mathbf{G}_k}\left[\Pr(i\in C_j|\mathcal{C}^\dagger_\Upsilon)\cdot r(S[i],A[i])\right] + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i],A[i]) \tag{16}$$

$$= \sum_{k\in\kappa}\sum_{i\in\mathbf{G}_k}\left[\frac{|\{\iota:\iota\in(\mathbf{G}_{\phi(i)}\cap C_j)\}|}{|\mathbf{G}_{\phi(i)}|}\cdot r(S[i,0:T],A[i,1:T])\right] + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i],A[i]) \tag{17}$$

$$= \sum_{k\in\kappa}|\{\iota:\iota\in(\mathbf{G}_{\phi(i)}\cap C_j)\}|\cdot\sum_{i\in\mathbf{G}_k}\left[\frac{1}{|\mathbf{G}_{\phi(i)}|}\cdot r(S[i,0:T],A[i,1:T])\right] + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i],A[i]) \tag{18}$$

$$= \sum_{k\in\kappa}|\{\iota:\iota\in(\mathbf{G}_{\phi(i)}\cap C_j)\}|\cdot\tilde{r}_k + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i],A[i]) \tag{19}$$

$$= \sum_{i\in C_j}\tilde{r}_{\phi(i)}\cdot\Lambda_i + \sum_{i\in C_j}(1-\Lambda_i)\cdot r(S[i],\ A[i],\ \mathbf{1}:\ T) \tag{20}$$

□

表1. 测得的性能提升的样本方差

T	B	π_1 V π_0	RAW	置换	IPW	n-VAL
1	3%	π_{WI} V π_{GR}	49.09	4.94	**0.48**	9
1	10%	π_{WI} V π_{GR}	49.86	15.11	**6.66**	3
1	25%	π_{WI} V π_{GR}	49.45	**19.94**	78.12	2
10	3%	π_{WI} V π_{WI}	2381	**916**	NA	3
10	3%	π_{WI} V π_{GR}	2348	**728**	NA	4
10	3%	π_{GR} V π_{GR}	26356	**1860**	NA	13
10	10%	π_{GR} V π_{GR}	25983	**3808**	NA	7
10	25%	π_{GR} V π_{GR}	23619	**5477**	NA	5

B. 将资源分配策略构造为索引策略

对照组： 对于不接受任何干预措施的对照组，可以使用任意有限取值的随机生成索引矩阵来处理。设置索引阈值 $\Upsilon_i = \infty \, \forall \, i$ 可确保没有被分配到对照策略的个体会被选中接受干预。

轮转： 诸如“轮转”这类常见策略，按照固定顺序循环选择个体进行干预，也可以表示为索引策略。每个个体在每个时间步的索引可分两个阶段确定。首先，我们考虑用于对 N 个体进行排名的特征，并先设置 $\Upsilon_i(t) := r$，其中$r \in \{1, \dots N\}$ r表示个体的优先级排名（优先选择排名最高者）。接下来，每当某个个体获得一次动作 $a = 1$ 时，我们希望将其移至队列末尾，因此从其之后所有时间步的索引中减去 N，并对每次 $a = 1$都重复这一过程。

C. 补充实验结果

第 6.1 节中合成数据生成的6.1说明

$$P^{p,1} = \begin{bmatrix} 0.97 & 0.03 \\ 0.03 & 0.97 \end{bmatrix} \quad P^{a,1} = \begin{bmatrix} 0.96 & 0.04 \\ 0.01 & 0.99 \end{bmatrix}$$
$$P^{p,2} = \begin{bmatrix} 0.25 & 0.75 \\ 0.03 & 0.97 \end{bmatrix} \quad P^{a,2} = \begin{bmatrix} 0.23 & 0.77 \\ 0.01 & 0.99 \end{bmatrix}$$

图 5. 构成矩阵 P_1 和 P_2的概率值

我们复现了模拟中使用的转移概率 P_1 和 P_2，这些概率采用自（Mate 等人，2020年）的数值，见图5。每个 P 由两种动作下的一组概率组成（$a = 0$，用“p”表示被动，以及 $a = 1$，用“a”表示主动）。

直观来看， P_1 的$P^a_{0,1}$非常小，而$P^p_{0,1}$较大，因此一旦进入状态 $s = 0$，就很难恢复，即使采取干预（$a = 1$）也是如此，所以持续进行干预以防止个体进入 $s = 0$至关重要。另一方面， P_2 的$P^p_{0,1}$较大，因此具有自我纠正能力，也就是说，即使没有干预，个体也很可能迅速回到 $s = 1$ 。

C.2. 单次随机对照试验

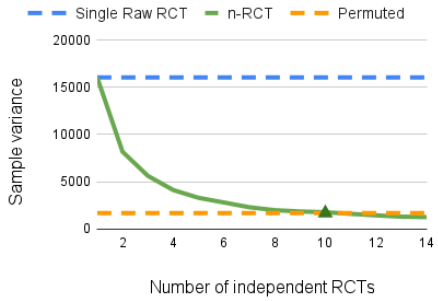


图6. 通过运行并对 n次独立试验取平均来降低方差。

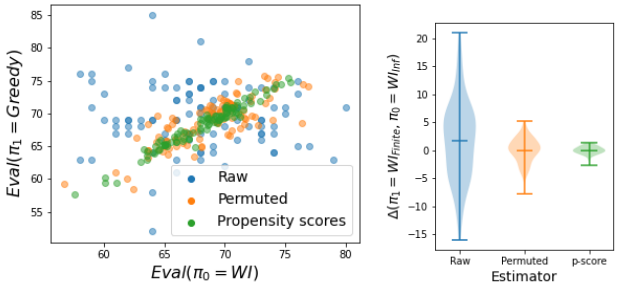


图7. 单次试验设置

C.3. 序贯随机对照试验

我们进行了更多比较，每个臂模拟 $N = 100$ 个体，运行500 个试验实例，每个持续 $T = 10$ 个时间步。其n–数值列于表1。

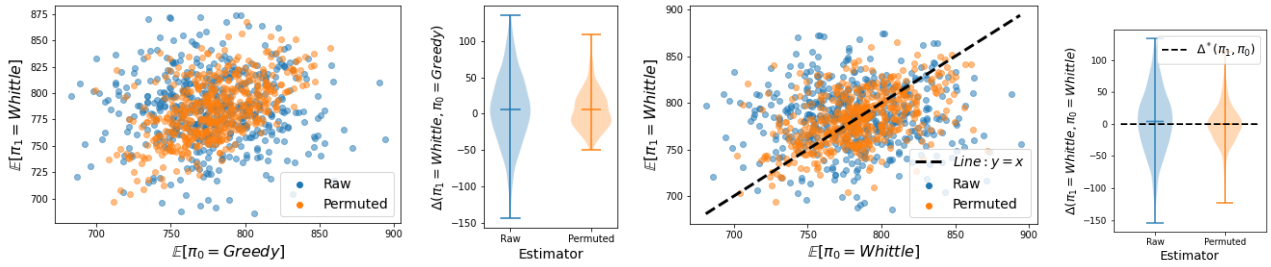


图8. Whittle 与贪心算法（左侧两个面板）以及 Whittle 与 Whittle（右侧两个面板）的比较

设置：贪心算法对比贪心算法 为了进行更细致的分析，我们考察参与试验的个体的状态轨迹。该分析每个臂使用 $N = 1000$ 个体，并模拟 30 个试验实例，持续 $T = 10$ 个时间步。我们发现，橙色轨迹（置换后）比蓝色轨迹（蓝色）更接近期望值

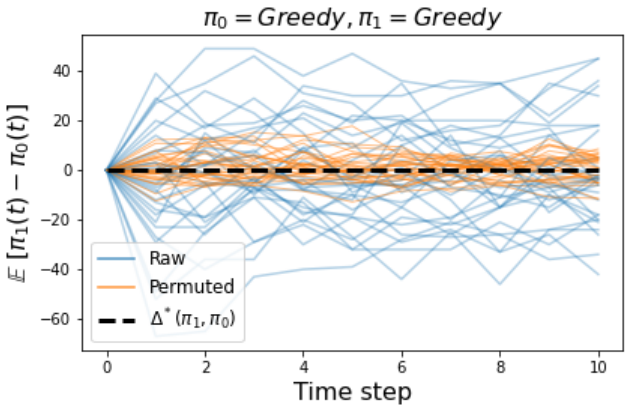


图*9.* 贪心算法与贪心算法的时间轨迹