《Understanding and misunderstanding randomized controlled trials》

文献归纳

刘娟伶、王裕

随机控制实验是因果推断的常用方法，本篇文章作者通过提出两个论断来阐述了对随机控制实验的理解和误解。首先，在第一部分中，作者认为从随机对照试验估计出的 ATE 可能更接近于以其他方式估计的事实真相。作者以熟悉的统计术语，包括偏差和精度、效率或预期损失，探讨了随机对照试验的优缺点。其中，无偏意味着反复试验得出的估计值在平均意义上是正确的，但是对任何一个估计值与事实的距离没有加以限制。而精确意味着与事实的距离差距。无论在哪里进行 RCT 或使用 RCT 结果，这些区别应该是众所周知的，如果这些统计推断的问题不被重视，尤其是当研究人群中存在个体治疗效果的不对称分布时，将会对显著性造成威胁。第二部分探讨了如何使用 RCT 的结果。作者描述了使用 RCT 证据的几种不同方法，将 RCT 的结果运用在不同学科中具有相似性，作者在该部分强调的是明确研究假设和目的的重要性。在通常的文献中，强调将 RCT 的结论的推广运用，但是在作者看来这既是不足又是过度的。因为推广 RCT 结果需要大量额外信息，而这些信息无法从 RCT 中获得，所以是过度的；又由于 RCT 可以服务于更多的目的，而不是预测在试验人群中获得的结果将会在其他地方发挥作用，因此说是不足的。

1. 对 ATE 的估计

从试验样本出发，给定随机分配的处理组和对照组，F1（Y1）和 F0（Y0）分别是来自处理组和对照组结果 Y1 和 Y0 的两个（边际）分布，估计的正是这两种分布均值的差异，这也是社会科学和医学中大部分文献的焦点。然而政策制定者不仅仅是对 ATE 感兴趣，他们也会对这两种分布的特征感兴趣，例如，如果 Y 是疾病负担，医疗卫生人员关心治疗是否能减少疾病负担的不平等，或者它对分布的第 10 或第 90 百分位数的影响，希望比较两种分布下治疗和未治疗的预期效用，并在控制受试者特征的情况下，考虑最佳预期效用最大化治疗规则。

作者通过以下线性因果模型评估 ATE，虚拟变量 Ti 是表示（0,1）的 treatment，Bi 是个体 i 的 treatment effect，r 表示其他因素 x 对结果 y 的影响，假定这里的 r 不随个体差异变化。如果 T 和 x 之间存在交互关系，即 x 会导致 x 对 y 的影响发现变化，那么 X 称为 effect modifiers。

$$Y_i = \beta_i T_i + \sum_{j=1}^{J} \gamma_j x_{ij} \qquad (1)$$

2. 对 RCT 的理解和误解

进一步的，作者将（1）式和反事实研究联系起来。对于每一个个体而言，T 对 Y 的影响只有一种潜在结果可以被观察到，yi0 或 yi1，前者表示没有受到 treatment，后者表示受到 treatment，当其中一个结果出现时，另一个结果就是反事实结果，我们将无法观测到。因此，Bi 代表的每个个体的 treatment effect， yi1 -yi0，也将无法观测。潜在结果不能全都被观察到，这是进行因果分析所面临的最根本的问题。 Rubin(1974)认为这一问题实质上就是一个缺失数据的问题。因此要进行因果分析，就需要把缺失的潜在结果填补起来。一些个体的背景特征（协变量）往往能够帮助我们进行缺失潜在结果的预测。这值得注意的是，

不是所有变量都可以作为协变量，协变量必须满足一个条件：个体是否得到处理被先验地认为不会对协变量造成影响。个体永久性的特征，或者发生在处理之前的变量，都可以作为协变量，它们也被称为处理前变量。协变量的有用性体现在三个方面：使估计更加精确、提供特定群体的因果分析、非混淆性。

因果效应的估计必须依赖能够被观察到的潜在结果。定义因果效应只需一个个体就够了，而由于一个个体只能带来一个可被观察到的结果，所以因果效应的估计需要多个个体，而且这些个体被施加的行动需要有所不同，即随机分为处理组和控制组，ATE 就是两组均值之差，而且是无偏估计。但是，反事实方法受到很多经济学家的争论，Imbens 和 Wooldridge（2009）强调了个体在治疗效果方面具有几乎无限的异质性，Heckman 和 Vytlacil（2007）也指出反事实方法常常使我们对治疗的确切性质一无所知。不管是随机还是非随机，我们从实验可以得到处理组的平均结果减去对照组的平均结果。当所有其他原因的均值在两组中相同时，或者当两组中其他因素的净平均余额之和为零时，即完美平衡的情况，两组之间的差异恰好等于处理组中处理效应的平均值。在随机试验中，通过多次重复抽样，处理组和对照组来自相同的分布，因此能够保证第二项的期望为零

$$\overline{Y_1} - \overline{Y_0} = \overline{\beta_1} + \sum_{j=1}^{J} \gamma_j((\overline{x}_{1ij} - \overline{x}_{0ij})) = \overline{\beta_1} + ((\overline{S_1} - \overline{S_0})) \qquad (2)$$

我们知道，Fisher 的 P 值方法、Neyman 的重复抽样法、回归方法和以模型为基础的推断方法是四种处理随机试验的方法。经济学的实证工作中因种种限制而较少进行经典随机试验，尽管经济学者更多时候手头上只有观测性数据，但是对随机试验的分析可以为观测性数据的研究提供一个 benchmark，我们使用诸如 DID、PSM、IV 等方法就是为了使因果分析的可信度接近于随机试验。随机试验的好处就在于，研究人员不需要了解影响结果的所有因素。一个著名的例子是 Lalonde（1986）对劳动力市场培训计划的研究，该论文对观察性研究样本进行了大规模的重新检验，试图将它们纳入考虑范围，尽管差异在于似乎不同的研究结果适用于不同的人群这一事实（Heckman et al,1999）。从随机控制实验中得到 ATE 就好像是普遍的事实一样，而不是只在试验样本中。

综合考虑样本量、平衡和精确度时，会得出什么结论呢？当样本量很大时，试验更可能得到平衡。由于样本量趋于无穷大，处理组和对照组中其他原因 x 的平均值将变得任意接近，然而，这在有限样本中却无法达到。另外，即使样本量非常大，如果 x 的数量很大，每种原因的平衡可能也是不可行的。Vandenbroucke（2004）指出，人类基因组中有 30 亿个碱基对，其中许多或全部可能是我们试图影响的生物学结果的相关预后因素。正如（2）明确指出的那样，我们不需要单独地对每个原因进行平衡，只需要根据它们的净效应，即 S1-S0。但考虑人类所有这些数十亿之中基因组碱基对，只有一个可能是重要的，如果那个是不平衡的，单个试验的结果可能会“随机混淆”而且远非事实。因此，如果不了解其他原因以及它们如何影响结果，就无法判断大样本下是否能达到平衡。

那么，如何使 RCT 对平均处理效应有一个很好的估计呢？我们可以通过 loss function 或者 utility function 等方式来权衡 bias 和 precision。我们可以先考虑这些函数取决于实验者如何受到 ATE 估计值与事实的偏差的影响，然后我们选择最大限度地减少预期损失或最大化预期效用的估计值或实验设计，从而从 RCT 中找到较好的估计结果。

3. 对 RCT 结果的运用

在做因果分析时，RCT 的主要优点在于，如果进行得很好，可以在试验样本中对 ATE 进行无偏估计，从而提供证据表明 treatment 在该样本中导致某些个体的结果，明确该样本的因果关系。但我们往往只能对某些群体的平均效果有一些了解，而不能确保当结果推广到其他群体上时也会具有相同的效果。

第二部分的论点是随机化的重要性以及从随机试验中给予估计的 ATE 的解释有何意义？首先，作者应该确保对试验人群的 ATE 的无偏估计足以运行的可能的试验成本。其次，由于随机化不能确保正交性，因此得出结论估计是无偏的，需要保证没有显著的随机化与 treatment 相关。 第三，这里要研究的因果关系不能仅仅只是被从样本中推测出来，更重要的是总体的因果关系。当存在显着的异质性时，即使从该群体中随机选择试验，试验样品中的 ATE 也可能与感兴趣的群体中的 ATE 也完全不同。而且，在实践中，试验样本与总体之间的关系通常是模糊的（Longford.

Nelder,1999）。第四，除此之外，在许多情况下，统计推断将是很好的方法，但应该注意处理效应的异常值的可能性，处理组和对照组的边际分布都可能提供信息。

随机控制试验 RCT 是试验样本中对平均处理效应非参数估计的最终结果，因为它对异质性、因果结构、变量选择和函数形式等做出很少的假设，它通常是引入实验者控制方差的便捷方式。但是结果的可信度，可能受到不平衡协变量和过度异质性的影响。RCT 中可信度的代价是我们只能获得处理效应分布的均值，而且只能用于试验样本。然而，在存在处理效应或协变量的异常值的情况下，对因果关系的可靠推断往往是困难的。因此，我们必须根据之前的研究产生的先验知识，对要研究的样本背后的因果机制以及 effect modifier 的分布有一个全面的了解，才能知道 ATE 的作用范围。