因果推断（Casual Inference）

阅读笔记

October 26, 2020

主要的参考是 UCB 丁鹏老师的材料《因果推断简介》。

统计上的因果推断问题来源于相关性的缺陷 Yule-Simpson Paradox(YSP)。比较经典的例子如患者的双盲（?）实验：药物实验中，患者总体接受吃药/安慰剂的康复率对比，与按性别拆开后的康复率对比恰好相反（如材料表 1），其出现的原因在于康复情况不仅由吃药/吃安慰剂决定，也可能与性别存在联系，用因果图（一种有向无环图，Direct Acyclic Graph，DAG）表示的话，就是材料图 1 的样子。

YSP 的存在表明，统计学上的相关并不能作为因果关系的证据，而证明因素之间具有因果关系（即因果推断）是比较困难的。这篇材料讲解了 Rubin 因果模型（RCM）、观测性研究下的因果推断、不依从和死亡删失问题（频率学派、贝叶斯学派）、因果图问题。其中前四章主要用 RCM 讨论，最后一章简要讨论一下因果图模型。

1 Rubin 因果模型

探究因果问题的根本阻碍是什么？对于一个实验体，人们只能观测其中的一个结果，而不能同时得到两种情况的结果。例如一个人头疼，吃了阿司匹林好了，那么阿司匹林是治好头疼的原因吗？我们没法知道，因为这个人不能回退到没吃阿司匹林的时候了。

就某一个具体的实验者来说，重复观测是不可行的，但是随机抽样实验可以在大量样本的基础上观测总体的因果效应。Rubin 因果模型(Rubin Causal Model, RCM) 应运而生，这是 Rubin(1974,1977,1978)[10–12] 所做工作的集合。简单来说，它将总体基本均分为实验组和对照组，观测两组人接受（实验组）/未接受（对照组）处理后的期望之差，由此获得平均因果效应（Average Causal Effect,ACE）的估计[1]。

设 Z_i 表示个体 i 是否接受处理（实验/对照组），实验组取 1，对照组取 0；个体 i 的结果被表示为 Y_i，个体 i 接受处理/成为对照的潜在结果可以被表示为 $(Y_i(1), Y_i(0))$ 这个数值对。RCM 就将 $Y_i(1) - Y_i(0)$ 作为个体 i 接受处理的个体因果作用，但就像刚才说的，一个人不能同时观测两个潜在结果，因而个体的因果作用无法被识别。但 i 作为一个个体，可以成为样本空间 Ω 中的一个样本点。因而，在 Z 在总体范围内随机化[2]的前提下，可以识别总体的平均因果作用：

$$ACE(Z \to Y) = E(Y_i(1) - Y_i(0)) \tag{1}$$

这么定义的原因是：

$$ACE(Z \to Y) = E(Y_i(1)) - E(Y_i(0)) = E[Y_i(1)|Z = 1] - E[Y_i(0)|Z = 0] \tag{2}$$

$$(Z \text{ 与 } (Y(1), Y(0)) \text{ 数值对互相独立}) = E[Y_i|Z = 1] - E[Y_i|Z = 0] \tag{3}$$

对于我们来说，$Y_i(1)$ 和 $Y_i(0)$ 是无法直接观测到的，那么上式就让我们通过能观测的 $E[Y_i|Z = 1]$ 和 $E[Y_i|Z = 0]$（对照实验的两个结果）来估计 ACE。

2 观测性研究下的因果推断

RCM 的理想是很美好的，但是这个模型没有解决 YSP 的问题。RCM 本身带着**总体中所有的个体（除实验/对照外）完全同质化**的假设。但在现实中，由于个体之间存在各种条件（不妨设这些条件为 X，以区别于实验/对照的 Z）上的差异，RCM 提出的 $E[Y_i|Z = 1] - E[Y_i|Z = 0]$ 是**无法识别**因果作用的。怎么解决？(1) 如果我们收集足够多的个体信息 X 之后，能以 X 为条件重新做到 Z 与 Y_i 独立，那最好不过，从而提出了**可忽略性**问题（ignorability）；(2) 如果没有办法直接用 X，比如 X 的维度很多，那么能不能给出一个指标来压缩 X 的维度，然后根据指标的取值来分层，从而达成对信息 X 的忽略？从而有了**倾向得分**方法（propensity score）；(3) 此外，还可以用回归方式，将 ACE 作为一个参数直接估计出来，从而出现了传统回归法和 Heckman 选择模型。

[1] 更进一步，RCM 还要求两个条件：(1) “因” 是能够控制的，比如研究一个人的身高/体重的因果关系，我们没有办法让一个人长高，然后判断这个人的身高更高以后，体重会不会变化，所以这个因果关系没法判断；(2) 固定单位处理值 (Stable Unit Treatment Value Assumption，SUTVA)，一个实验个体的观察结果不应受到其他实验个体所受处理的任何影响，比方说讨论药物对血压的影响，那么张三的血压就不应该取决于李四吃不吃药。

[2] 一个人被划入实验组、对照组是完全随机的，这暗含了 Z 与 $Y_i(1), Y_i(0)$ 相互独立的条件，下面的推导会用到。

2.1 可忽略性

这个概念来自于 Rosenbaum & Rubin(1983)[9]。Hernan & Robins(2010)[5] 将其拆分为两部分：可交换性 (exchangeability) 和 positivity。其中下面要讲到的“强可忽略”对应“完全可交换”与 positivity 的组合，而“弱可忽略”则是“可交换性”和 positivity 的组合。

回到前面 RCM 的内容，在随机化的实验中隐含着 $Z \perp (Y(1), Y(0))$ 的假定，即强可忽略性。一个人被分在实验/对照组所产生的潜在结果对 $(Y(1), Y(0))$ 与其具体被分在哪一组无关。尽管现实中的随机试验会受到各式各样信息的影响，但如果将信息作为条件后仍有 $Z \perp (Y(1), Y(0))|X$ 成立，那么信息的加入其实并不会破坏强可忽略性，这时的 ACE 在加入信息作为条件后仍然可以识别。

强可忽略性下可识别的证明：

$$ACE = E(Y(1)) - E(Y(0)) = E[E(Y(1)|X)] - E[E(Y(0)|X)] \tag{4}$$

$$= E[E(Y(1)|X, Z=1)] - E[E(Y(0)|X, Z=0)] = E[E(Y|X, Z=1)] - E[E(Y|X, Z=0)] \tag{5}$$

举个例子，一个论坛有各种各样的用户，广告商可能会更倾向于给新用户（历史登录天数少）投送广告，因而 Z（是否投送广告）与 X（历史登录天数）负相关；而对于历史登录天数多的老用户来说，他们通常更频繁的浏览论坛，因而会更容易看到广告，广告的效果更好，也就是 X 与 $Y(1), Y(0)$ 正相关。这两个因素一撮合，就发现 Y 和 Z 不独立。但是对于登录天数差不多的用户来说，是否投送广告就可以被看作是随机的了。

只要匹配上信息相同的个体，那么再去做观测性实验来估计 ACE 就可以了。但是，就算数据量很大，能找到无关信息完全相同的个体也是很难的，接下来要讲的倾向得分法，就是一种放宽的匹配方式。

第一段讲到，可忽略性是两段内容的组合：条件概率为正 (positivity)、可交换性 (exchangeability)。**正条件概率**要求 $\forall X, 0 < P[Z=1|X] < 1$。用上面的例子来解释，如果对于某类人来说，他们永远收不到广告，那么这类人就没法通过历史数据分析广告效果，对他们做因果推断是没意义的。而**可交换性**要求一个接受实验的个体，应当可以在实验/对照组自由变换，而不会使得其预期结果有任何改变。如 Hernan & Robins(2010) 书中表 3.1 的数据，实验组 69% 的个体面临严重情况，而对照组里只有 43%，因而两组数据并不平衡，不能满足可交换性。

2.2 倾向得分

由于大多数情况下，直接做精确匹配是很难的，需要做点数据上的压缩，凑合着匹配起来，但目的永远只有一个——**消除要比较的两组人群之间的不同质问题**。因此，我们引入倾向性得分，其定义为：一个实验个体属于实验组的倾向性，即 $e(x) = P[Z=1|X=x]$。Imbens & Rubin(2015) 给出了倾向性得分的两个性质:(1)balancing;(2)unconfoundedness[8]。

(1)Balancing Property（平衡得分的性质）:$Z_i \perp X_i|e(X_i)$，即已知倾向性得分的前提下，随机分组 Z_i 与此人个人信息 X_i 独立，下面给出证明[3]，我们发现待证结论等价于：

$$P[Z_i = 1|X_i, e(X_i)] = P[Z_i = 1|e(X_i)] \tag{6}$$

首先，考虑这个式子的左半边，因为倾向得分 $e(X)$ 是 X 的函数，所以：

$$P[Z_i = 1|X_i, e(X_i)] = P[Z_i = 1|X_i] = e(X_i) \tag{7}$$

左边变成了倾向得分的形式，再看式 *6* 右边，因为 Z_i 是 *0-1* 变量，所以这个概率即为条件期望，然后第二步反用 *LIE*，有：

$$P[Z_i = 1|e(X_i)] = E[Z_i|e(X_i)] = E[E[Z_i|X_i, e(X_i)]|e(X_i)] = E[e(X_i)|e(X_i)] = e(X_i) \tag{8}$$

从而左右相等，因而证出这个性质。

(2)Unconfoundedness（无混杂性）：若强可忽略性成立，即有 $Z \perp (Y(1), Y(0))|X$，则 $Z_i \perp Y_i(0), Y_i(1)|e(X_i)$。给定倾向得分后，处理方式与潜在结果之间互相独立，下面给出证明[4]，这个结论等价于：

$$P_Z[Z_i = 1|Y_i(0), Y_i(1), e(X_i)] = P_Z(Z_i = 1|e(X_i)) \tag{9}$$

对于左边，由 Z_i 是 *0-1* 变量，可以将概率写为期望的形式：

$$P_Z[Z_i = 1|Y_i(0), Y_i(1), e(X_i)] = E_Z[Z_i|Y_i(0), Y_i(1), e(X_i)] \tag{10}$$

$$= E[E_Z[Z_i|Y_i(0), Y_i(1), X_i, e(X_i)]|Y_i(0), Y_i(1), e(X_i)] \tag{11}$$

引用强可忽略性，则里面那层期望 $E_Z[Z_i|Y_i(0), Y_i(1), X_i, e(X_i)] = E_Z[Z_i|e(X_i)]$，从而使上式变为：

$$E[E_Z[Z_i|e(X_i)]|Y_i(0), Y_i(1), e(X_i)] = E[Z_i|e(X_i)] = P(Z_i = 1|e(X_i)) \tag{12}$$

从而有这个性质成立，第二步也是 *LIE* 的应用。

[3]对应 Imbens&Robin(2015) Lemma 12.1 (p.266)。
[4]对应 Imbens&Rubin(2015) Lemma 12.2 (p.267)。

2.2.1 基于倾向分的匹配

由这两个性质，我们发现，给定 X 如果可以使得处理机制能够忽略 X 的影响，那么如果给定一个降维后的 $e(X)$ 也能起到同样的作用。这样做的步骤为：

- **(1) 估计倾向性得分模型**。通常是 Logit/Probit 这种因变量为 0-1 的模型，估计每个个体的倾向性得分。计算机方向的会用逻辑回归配合 LightGBM 等算法一起做，然后挑拟合性好的模型做，但是工作量很大。
- **(2) 匹配倾向性得分**。常见的方式是分层法，即利用估计出的倾向性得分 $\hat{e}(X_i)$ 来对个体做分层，再每一层中估一个 ACE，然后对总体加权平均。除了这种方式之外，还有差异匹配、Mahalanobis 指标、最近邻等匹配方式，参见 Wikipedia。匹配时可以使用原分 $e(X_i)$，也可以尝试用其 Logit 即 $l(X) = ln[e(X)/(1-e(X))]$；通常计算出倾向分后，要修剪掉极端值，保留实验组/对照组的倾向分交集，或者选取 [5%,95%] 区间内的倾向分数据；实际工作中常用最近邻方式做 1 对 K 的无放回/有放回抽样，也可以用阈值法，即匹配所有与个体倾向分差距小于阈值的个体；有时还会设定倾向分差异上限，在最近邻等算法强行找邻居的时候，如果“强扭的瓜不甜”，就放弃这个尝试。更多的内容可以参照 Caliendo & Kopeinig(2008) 的工作[3]。
- **(3) 平衡性检查**。在第二步完成匹配后，最理想的结果是新的实验组、对照组在倾向分上的分布完全相同，而二者分布越相似，说明“配平”的效果越好，考察这个配平效果的方式通常是看倾向得分分布图或 QQ 图。Austin(2009) 给出了一种比较精确的方式，即 Standarized Mean Difference（SMD，标准均值差异），其计算方式为**（实验组均值 - 对照组均值）/实验组标准差**，对于一个变量来说较好的配平效果要求 $SMD \leq 0.1$，如果 SMD 过大，应当考虑这个变量的显著性[2]。
- **(4) 因果效应估算**。这个时候直接比较配平后的实验组和对照组即可，也就是逐组估计 ACE 后加权求出总体 ACE 估计。
- **(5) 鲁棒性分析**。简单的方式抽出一个或几个 X 中的变量，再走一遍上述过程，如果 ACE 变化极大，则说明信息变量的 unconfoundedness(无混杂性）要么依赖 X 全体变量，要么就不存在。

2.2.2 基于倾向分的加权

以上的匹配法是一种方式。此外，还可以采用对倾向分做加权的方式做处理。有人做了如下的图：

Figure 1: 倾向分加权法

一共有 *20* 人，*10* 男 *10* 女，其中男女被投广告的概率分别为 *20%* 和 *80%*，由此得到投广告（实验组）的比例是 *2* 男 *8* 女，对照组的比例是 *8* 男 *2* 女。其中，投广告组的人转化为客户的有 *1* 男 *6* 女，对照组则是 *1* 男 *1* 女，由此得到左上图中的转化率表。由表可知，对于整体人群（比例 *1*: *1*），广告的投放使得转化率提高了 *31.25* 个百分点 *(62.5% - 31.25%)*，而对于实验组人群（比例 *1*: *4*），则提高了 *27.5* 个百分点：

$$(50\% \times 0.2 + 75\% \times 0.8) - (12.5\% \times 0.2 + 50\% \times 0.8) = 27.5\% \tag{13}$$

上面是开了上帝视角看到的东西，但是有人会对于这样的对照数据，得到打广告会把转化率从 *20%* 提高到 *70%* 的错觉（图二）。这还是那个老问题，对照组和实验组的人群信息不同（性别比例有区别），除了像之前那样做匹配外，还可以做加权，把数据的信息拉到同等的水平上。讨论总体处理效应如图 *3*，将处理组和对照组的比例拉成 *1*: *1*，可以给处理组的男生赋予 *5* 倍的权重，给女生 *1.25* 倍权重，从而使得处理组

“有” *10* 个男生 *10* 个女生；对应的，对对照组赋予 *1.25* 和 *5* 倍权重，然后求加权转化率，从而估计 *ACE*。类似的思路，可以对处理组的比例进行模拟，如图四。

意思是这么个意思，如果维度低还挺简单的，但是 X 通常维度很高，因而加权法也不会很简单，真正的加权法仍然遵照上面提到的各项基本假设（可忽略性及其附属的各项假设）。倾向分加权的具体定义是：**(1)** 对处理组属性为 X 的用户权值为 $1/e(X)$；**(2)** 对照组属性 X 的用户权值为 $1/(1-e(X))$。由此，开始推导 ACE，详细推导过程参照Rebecca Barter 的文章。

首先，ACE 的定义是：

$$\hat{ACE} = \frac{1}{n_T}\sum_{i:Z_i=1} y_i - \frac{1}{n_C}\sum_{i:Z_i=0} y_i \tag{14}$$

其中 n_T 与 n_C 分别是实验组、对照组人数。经过加权，则修正后的 ACE 估计值为:

$$\hat{ACE}^{IP} = \frac{1}{n_T}\sum_{i:Z_i=1} \frac{y_i}{e(X)} - \frac{1}{n_C}\sum_{i:Z_i=0} \frac{y_i}{1-e(X)} \tag{15}$$

$$= \frac{1}{n}\sum_{i}^{n} \frac{Z_i Y_i}{e(X_i)} - \frac{1}{n}\sum_{i}^{n} \frac{(1-Z_i)Y_i}{1-e(X_i)} \tag{16}$$

$e(X)$ 是倾向分的**估计值**。式 16 的前项为潜在结果 $Y(1)$ 的期望，后者为潜在结果 $Y(0)$ 的期望。此式由 Hirano et.al.(2003) 提出，并证明其半参数有效[6]。如果讨论处理组比例的平均因果效应（Average Causal Effect on Treated,ACT) 时，则是：

$$\hat{ACT}^{IP} = \frac{1}{n_T}\sum_{i=1}^{n}[Z_i Y_i - \frac{e(X_i)(1-Z_i)Y_i}{1-e(X_i)}] \tag{17}$$

2.3 Heckman 选择模型

在观测性实验中，ACE 的定义是 $ACE = E(Y|X, Z=1) - E(Y|X, Z=0)$，即两个条件矩的估计，所以可以使用回归模型，比如下面的模型：

$$E(Y|Z, X) = \alpha + \beta Z + g(X, \gamma) \tag{18}$$

那么:

$$\hat{ACE} = E(Y|X, Z=1) - E(Y|X, Z=0) = \beta(1-0) = \beta \tag{19}$$

换句话说，估计出 β 就是估计出了 ACE。而 Greene(2002) 提出了一种 Heckman 模型的改进型：

$$Z^* = \delta + \theta X + v \tag{20}$$

$$Z = I(Z^* \leq 0) \tag{21}$$

$$Y = \alpha + \beta Z + \gamma X + u \tag{22}$$

并设定 $(u, v) \sim MN$。回归法问题在于假设较强（甚至假设了分布），因而较少被使用。

3 随机化实验的特殊情况：不依从、死亡删失问题

“不依从” 问题是指实验者可能不服从实验安排，做出与实验不同的行为；而死亡删失则更常见于医学实验，即实验结果出炉之前实验者已经死亡，实验被迫中止。

3.1 不依从——主分层法与工具变量

设总体有 N 个个体，对于某个个体 i，$Z_i = 1$ 表示其被分配到了实验组，而 $D_i = 1$ 表明其实际接受了处理，而 $D_i = 0$ 代表其实际上接受了对照措施，那么 $Z_i \neq D_i$ 时，即代表了此人不服从原实验安排的措施，即所谓的 “不依从”（noncompliance）。与上面相同，此人的结果变量仍然表示为 Y_i。此类问题的讨论方式有两种，一种直接忽略不依从现象，直接估计 $ACE(Z \to Y) = E(Y(1)) - E(Y(0))$，那么他讨论的就只是 Z（实验分配/随机化）的作用，而不能考虑实验者最终所做处理（D）的作用。

另一种方式是抛开原始随机化结果（Z），只讨论 D 对 Y 的影响：$E(Y|D=1) - E(Y|D=0)$，它的问题在于，影响实验者是否执行处理（D）的因素，和影响实验者结果（Y）的因素可能相同，也就是说 Y 和 D 可能混杂 (confound)，那么得到的这个类似 ACE 的结果用处不大。

面对这种问题，Frangakis & Rubin(2002) 提出了 “主分层” 的分析框架[4]：

$$C_i = \begin{cases} c, & \text{if } D_i(0)=0 \quad \text{and } D_i(1)=1 \\ n, & \text{if } D_i(0)=0 \quad \text{and } D_i(1)=0 \\ a, & \text{if } D_i(0)=1 \quad \text{and } D_i(1)=1 \\ d, & \text{if } D_i(0)=1 \quad \text{and } D_i(1)=0 \end{cases} \tag{23}$$

C 可以被看作是每个个体的个人信息（一种特殊的 X），他不受任何处理、处理后变量的影响。用处理后变量的潜在结果做分层的方式就是主分层方式。但并不是简单的将 D 的取值作为分层去求 ACE，因为这是有问题的 (Rubin,2004)[13]，如材料图 2。这种分析方式将 D 作为条件，而已知 D 的情况下，Y 和 C 不再互相独立[5]，而 Z 和 C 会同时作用于 Y，就带来了混杂问题。这个问题至今无法解决，Imbens & Rubin(1997) 只是将 $C = c$ 时的 ACE 估计出来[7]：

$$CACE(Z \to Y) = E[Y(Z = 1, D(Z = 1)) - Y(Z = 0, D(Z = 0))|C = c] \tag{24}$$

由于 $C = c$ 时实验者完全依从，因而随机化对结果的 ACE 就是实际操作对结果的 ACE，但其他组就无法得到类似的结果。由于在其他组别中，Z 与 C 存在混杂，因而其它组别的因果度量是混乱的——无法清晰表明实际的处理对结果的随机作用——而且可证明其它组别的因果作用不可识别。而对于 CACE，频率学派证明其在特定假定下可以识别：

- 单调性假定 $D(1) \geq D(0)$。即 $C = d$（完全不依从）的人不存在。
- Exclusion Restriction（排除假设）。对于 $C = n$ 或 $C = a$ 的人群，其具有性质 D(Z = 1) = D(Z = 0)。现在加入第二条限制，即 Y(Z = 1,D(Z = 1)) = Y(Z = 0,D(Z = 0))，体现在图中就是去掉了 $Z \to Y$ 的直接影响关系，而 Z 只能通过 D 向 Y 施加影响。

Angrist et.al. (1996) 证明了在以上两个假定下 CACE 可以识别[1]，其表达式为[6]：

$$CACE = \frac{ACE(Z \to Y)}{ACE(Z \to D)} = \frac{E(Y|Z = 1) - E(Y|Z = 0)}{E(D|Z = 1) - E(D|Z = 0)} \tag{25}$$

这个估计量属于 Wald(1940) 提出的 Wald 估计量。而这个估计与计量中的工具变量类似。

3.1.1 从计量模型开始说

通常的线性模型为 $Y = \alpha + \beta D + \epsilon$，而且要求外生性假设：$E(\epsilon|D) = 0$，这暗含了 $cov(D, \epsilon) = 0$。但很多时候这个条件无法成立，这在因果推断的例子中很好解释。抽烟的人可能比不抽烟的人有更坏的生活习惯，从而使得包含个体其他信息的随机误差项 ϵ 不再与 D 完全独立，也就是内生性问题，那么这样估计出来的 $\hat{\beta}$ 就不再是 β 的相合估计。

放在原来的模型里，就是由于 C 的存在，$C \to D$ 和 $C \to Y$ 的链条同时存在，所以 $D \to Y$ 的因果关系存在混杂。但是可以加入一个工具变量，来做这个因果关系的识别，体现在模型里就是 Z。Z 要满足的条件是：$Z \perp C, Z \not\perp D, Z \perp Y|(D, U)$（也就是上面讲的 exclusion restriction）。有了这三个条件后，去看协方差：

$$cov(Z_i, Y_i) = \beta cov(Z_i, D_i) \tag{26}$$

从而立刻得到 β 的估计值：

$$\hat{\beta} = \frac{cov(Z_i, Y_i)}{cov(Z_i, D_i)} = \frac{\sum(Y_i - \overline{Y})(Z_i - \overline{Z})}{\sum(D_i - \overline{D})(Z_i - \overline{Z})} \tag{27}$$

由大数定律知，$\hat{\beta}_{IV}$ 为 β 的相合估计，如果 Z_i 为 0-1 变量，则：

$$\hat{\beta}_{IV} = \frac{\overline{Y}_1 - \overline{Y}_0}{\overline{D}_1 - \overline{D}_0} \tag{28}$$

3.1.2 回到因果推断

套到因果模型里，对应工具变量法的几个要求，对应提出了几个假定：

- $Z_i \perp \{D_i(1), D_i(0), Y_i(1), Y_i(0)\}$（随机化假设）；
- $D_i(0) \leq D_i(1)$（单调假设，被安排到实验组的实验者更可能做出实验组的处理）；
- $Y(Z = 1, D(Z = 1)) = Y(Z = 0, D(Z = 0))$（排除假设，即 Z 与 Y 没有直接的作用）。

从而得到：

$$ACE(Z \to Y) = E(Y_i(1)) - E(Y_i(0)) \tag{29}$$

$$= P[C = c]E[Y_i(1) - Y_i(0)|C = c] + P[C = n]E[Y_i(1) - Y_i(0)|C = n] + P[C = a]E[Y_i(1) - Y_i(0)|C = a] \tag{30}$$

$$= P[C = c]E[Y_i(1) - Y_i(0)|C = c] \tag{31}$$

对于 C = n 或 a 的人，其 $E[Y_i(1) - Y_i(0)|C = n] = E[Y_i(1) - Y_i(0)|C = a] = 0$，且 Z 对 D 没有任何影响，那么 Z 对 D 的影响体现在 $C = c$ 的人，即有：$ACE(Z \to D) = 1 \times P[C = c]$，所以有：

$$CACE = E[Y_i(1) - Y_i(0)|C = c] = \frac{ACE(Z \to Y)}{ACE(Z \to D)} = \frac{\overline{Y}_1 - \overline{Y}_0}{\overline{D}_1 - \overline{D}_0} = \hat{\beta}_{IV} \tag{32}$$

CACE 是可以识别的，但局限性在于，我们往往不知道个体在哪个子总体中，或者说 C_i 的取值对我们是不可观测的。

[5] X，Y 互相独立，但如果已知 X+Y = 1，那么就不独立了。
[6] 可参考 Hernan & Robins(2010) Technical Point 16.3, p.199。

3.2 死亡删失

死亡删失的问题是在原始问题上加入一个变量 S，其意义是实验者是否仍然存活（$S=1$ 代表存活，$S=0$ 为死亡），而对于那些已经死亡的个体，其实验结果已经没有意义了。与 3.1 一样的是，我们这里仍然采用主分层的思想，不过分层的依据略有不同：

$$C_i = \begin{cases} DL, & \text{if } S_i(0)=0 \quad \text{and } S_i(1)=1 \\ DD, & \text{if } S_i(0)=0 \quad \text{and } S_i(1)=0 \\ LL, & \text{if } S_i(0)=1 \quad \text{and } S_i(1)=1 \\ LD, & \text{if } S_i(0)=1 \quad \text{and } S_i(1)=0 \end{cases} \tag{33}$$

对于 LL，即无论如何都不会死的人，他们的情况和原始模型没有区别，即 $SACE = E[Y(1)-Y(0)|C=LL]$ 是可识别的。但对于其他的情况，结果变量都无法定义（有人死了），死了的人是没有结果的，你得到的所有的结果都是活下来的人才有的（也就是所谓的“幸存者偏差”），因而不能明确因果作用，具体的细节可以参照 Wang et.al.(2017) 的文章[14]。此外，Zhang et.al.（2008）利用参数模型来推导 SACE，如材料第 13 页所示，与之前的回归法类似，它的识别性严重依赖于正态分布[15]。亦有人尝试引入工具变量来考察识别性问题，如材料第 14 页，但死亡删失问题大多用于医学实验统计上，与我们关系不大，不再展开描述。

4 贝叶斯观点的 RCM

回过头去，把实验重新定义一下：设一次实验中有 T 种处理方案，总体有 N 个实验个体。此外，还有个体已被观测的其他信息（称之为协变量）$X=(X_1,...,X_C)$；个体接受处理的类型 $W \in (0,1,...,T)$，其中 $W=0$ 表示划入对照组。设潜在的结果变量 $Y=(Y^0,...Y^T)$，其中 $Y^t=(Y_1^t,...,Y_d^t)^T$（每个结果会有 d 个分量）。那么，处理 1 和处理 2 对于个体 i 结果向量中第 k 个参量的因果作用可以表示为 $Y_{ki}^1 - Y_{ki}^2$。这里的讨论继承了 RCM 的基本假设，如 SUTVA。

但是因果推断的根本问题仍然存在——有 $T+1$ 种处理，但只能观察到其中一种的结果，其他 T 种结果永远也观测不到，所以 Rubin 引入了缺失值的指示变量：对于某一个个体 i，若有 $W_i=t$，则 $M_{ki}^j=0, \forall j \neq t, k=1,...,d$，而 $M_{ki}^t=1, k=1,...,d$。

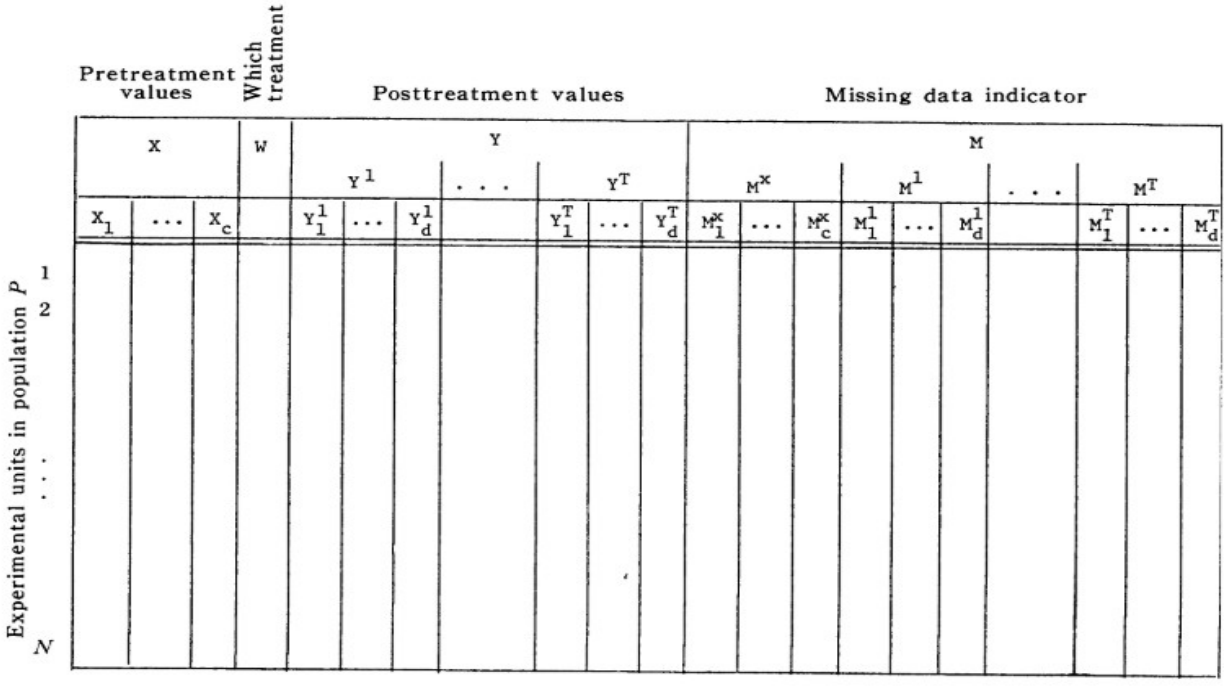

Figure 2: 一次实验中出现的所有数据

那么，所有的变量就是 (X,Y,W,M)，其中 (X,Y) 只能观测一部分，而 (W,M) 可以完全被观测。对于一次特定的实验来说，我们规定，观测到的 (W,M) 记作 $\tilde{W}$ 和 $\tilde{M}$，而 $\tilde{X}=(X_0,\tilde{X}_1)$，$\tilde{Y}=(Y_0,\tilde{Y}_1)$，$X_0$ 与 Y_0 是缺失数据（观测不到的），$\tilde{X}_1$ 和 $\tilde{Y}_1$ 是被观测到的数据，那么这些数据的联合分布可以写作：

$$f(X,Y,M,X) = f(X,Y|\pi)k(W|X,Y,\pi)g(M|W,X,Y,\pi) \tag{34}$$

第一项是给定未知参数 π 后 (X,Y) 的联合分布；第二项则是安排如何处理（分配实验）的机制；第三项则表示了给定处理之后，哪些数据被记录了 $(M=1)$，哪些则没有被观测到 $(M=0)$。做因果推断的一个目的就是要建立起观测数据与缺失数据的关系，从而通过观测数据来推断缺失数据（特别是 Y_0），为了完成这个目的，需要做出几个假定：

- (Assume 1) 给定参数 π 后，(X,Y) 的各行独立同分布，即有 $f(X,Y|\pi)=\prod_{i=1}^{N} f[(X_i,Y_i)|\pi]$。

- (Assume 2) 安排处理方式的机制具有可忽略性[7]：$k(\tilde{W}|\tilde{X},\tilde{Y},\pi)=k(\tilde{W}|\tilde{X}_1,\tilde{Y}_1)$。
- (Assume 3) 数据记录的机制具有可忽略性：$g(\tilde{M}|\tilde{X},\tilde{Y},\tilde{M},\pi)=g(\tilde{M}|\tilde{X}_1,\tilde{Y}_1,\tilde{W})$。

无论是 Bayes 方法还是频率学派方法，其最终的目的是将 Y_0，那些没有被观测到的数据含有的信息揭示出来，比如说预测 Y_0，参照 Rubin(1978) 的工作[12]，下式用于预测 Y_0 的分布：

$$Pre(Y_0|\tilde{X}_1,\tilde{Y}_1,\tilde{W},\tilde{M}) \tag{35}$$

$$=\frac{\int\int p(\pi)f(\tilde{X},\tilde{Y}|\pi)k(\tilde{W}|\tilde{X},\tilde{Y},\pi)g(\tilde{M}|\tilde{X},\tilde{Y},\tilde{W},\pi)d\pi dX_0}{\int\int\int p(\pi)f(\tilde{X},\tilde{Y}|\pi)k(\tilde{W}|\tilde{X},\tilde{Y},\pi)g(\tilde{M}|\tilde{X},\tilde{Y},\tilde{W},\pi)d\pi dX_0dY_0} \tag{36}$$

这个模型中存在的随机变量只有三个：π, X_0, Y_0。分子是考虑了所有未知参数 π 和所有未被观测到的协变量 X_0 后的期望实验结果。分母除了两者外，还考虑了所有的未被观测的实验结果 Y_0。因而这个式子就是表示了在所有已知条件的支持下，估计 Y_0 的分布情况。基于 Assume 2 和 Assume 3，有：

$$Pre(Y_0|\tilde{X}_1,\tilde{Y}_1,\tilde{W},\tilde{M}) \tag{37}$$

$$=\frac{\int\int p(\pi)f(\tilde{X},\tilde{Y}|\pi)k(\tilde{W}|\tilde{X}_1,\tilde{Y}_1)g(\tilde{M}|\tilde{X}_1,\tilde{Y}_1,\tilde{W})d\pi dX_0}{\int\int\int p(\pi)f(\tilde{X},\tilde{Y}|\pi)k(\tilde{W}|\tilde{X}_1,\tilde{Y}_1)g(\tilde{M}|\tilde{X}_1,\tilde{Y}_1,\tilde{W})d\pi dX_0dY_0} \tag{38}$$

$$=\frac{\int\int p(\pi)f(\tilde{X},\tilde{Y}|\pi)d\pi dX_0}{\int\int\int p(\pi)f(\tilde{X},\tilde{Y}|\pi)d\pi dX_0dY_0} \tag{39}$$

假定 2、3 介入后，得到的新 k 和 g 是定值，因而可以上下约去。所以说，这两个可忽略性假设的结果是 Y_0 预测的结果（即因果推断的结果）与处理的安排方式、数据记录的方式均无关系。接下来，Rubin 证明了贝叶斯观点下的因果推断结果只与三项内容有关：(1) 观测到的数据 $\tilde{X}_1,\tilde{Y}_1,\tilde{M}$；(2) 在给定 $\tilde{X}_1$ 和 π 时，Y 的分布：

$$h(Y|X_1,\pi)=\frac{\int f(X,Y|\pi)dX_0}{\int\int f(X,Y|\pi)dX_0dY} \tag{40}$$

(3) 给定 $\tilde{X}_1$ 时，π 的分布：

$$q(\pi|X_1)=\frac{p(\pi)\int\int f(X,Y|\pi)dX_0dY}{\int\int\int p(\pi)f(X,Y|\pi)dX_0dYd\pi} \tag{41}$$

将式 40、41 带回 39，有：

$$Pre(Y_0|\tilde{X}_1,\tilde{Y}_1,\tilde{W},\tilde{M})=\frac{\int h(\tilde{Y}|\tilde{X}_1,\pi)q(\pi|\tilde{X}_1)d\pi}{\int\int h(\tilde{Y}|\tilde{X}_1,\pi)q(\pi|\tilde{X}_1)d\pi dY_0} \tag{42}$$

A 附录：数学推导

References

[1] Joshua D Angrist, Guido W Imbens, and Donald B Rubin. Identification of causal effects using instrumental variables. *Journal of the American statistical Association*, 91(434):444–455, 1996.

[2] Peter C Austin. Balance diagnostics for comparing the distribution of baseline covariates between treatment groups in propensity-score matched samples. *Statistics in medicine*, 28(25):3083–3107, 2009.

[3] Marco Caliendo and Sabine Kopeinig. Some practical guidance for the implementation of propensity score matching. *Journal of economic surveys*, 22(1):31–72, 2008.

[4] Constantine E Frangakis and Donald B Rubin. Principal stratification in causal inference. *Biometrics*, 58(1):21–29, 2002.

[5] Miguel A Hernán and James M Robins. Causal inference, 2010.

[6] Keisuke Hirano, Guido W Imbens, and Geert Ridder. Efficient estimation of average treatment effects using the estimated propensity score. *Econometrica*, 71(4):1161–1189, 2003.

[7] Guido W Imbens and Donald B Rubin. Bayesian inference for causal effects in randomized experiments with noncompliance. *The annals of statistics*, pages 305–327, 1997.

[8] Guido W Imbens and Donald B Rubin. *Causal inference in statistics, social, and biomedical sciences*. Cambridge University Press, 2015.

[7] 即单凭手中已经观察到的信息，就可以估计出原始的函数，原函数里那些不可观测的参数实际上并不会对估计产生任何影响。

[9] Paul R Rosenbaum and Donald B Rubin. The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55, 1983.

[10] Donald B Rubin. Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of educational Psychology*, 66(5):688, 1974.

[11] Donald B Rubin. Assignment to treatment group on the basis of a covariate. *Journal of educational Statistics*, 2(1):1–26, 1977.

[12] Donald B Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of statistics*, pages 34–58, 1978.

[13] Donald B Rubin. Direct and indirect causal effects via potential outcomes. *Scandinavian Journal of Statistics*, 31(2):161–170, 2004.

[14] Linbo Wang, Xiao-Hua Zhou, and Thomas S Richardson. Identification and estimation of causal effects with outcomes truncated by death. *Biometrika*, 104(3):597–612, 2017.

[15] Junni L Zhang, Donald B Rubin, and Fabrizia Mealli. Evaluating the effects of job training programs on wages through principal stratification. *Advances in Econometrics*, 21:117–145, 2008.