第 1 章　观测研究和实验

1.1　引言

本书考虑回归模型及其一些变体，包括路径模型、联立方程模型、logit 和 probit 模型等. 回归模型能用于不同的目的：

（ⅰ）概括数据.

（ⅱ）预测未来.

（ⅲ）预测干预的结果.

第三点涉及因果推断，是最有意思和最难以捉摸的. 这将是我们的重点. 作为背景知识，这一节覆盖研究设计的一些基本原理.

因果推断是由观测研究（observational study）、自然试验（natural experiment）和随机化控制试验（randomized controlled experiment）组成. 当利用观测（非试验）数据来做因果推断时，关键问题在于混杂（confounding）. 有时通过划分研究的总体（称为分层（stratification）或者交叉制表（cross-tabulation））来处理这个问题，有时采用建模方法来处理这个问题. 这些策略各有其优缺点，需要探索.

在医学和社会科学中，基于随机化控制试验的因果推断是最可靠的，这时，研究者可通过掷硬币随机安排对象到处理组（treatment group）或控制组（control group）. 除了随机误差之外，掷硬币平衡了处理之外的全部有关因素. 因此，在处理组和控制组之间的差别就完全源于处理了. 这就是为什么因果关系容易由试验数据得到. 然而，试验往往是昂贵的，甚至由于伦理或实际原因而不可能实现. 于是统计学家转向观测研究.

在观测研究中，对象把自己安排到不同的组中. 研究人员仅仅观测发生了什么. 例如，吸烟效应的研究必须是观测性的. 然而，这里仍然使用处理-控制这一术语. 研究人员通过比较属于处理组（也称为暴露组（exposed group））的吸烟者及属于控制组的非吸烟者来确定吸烟的效应. 这些行话有些令人迷惑，因为“控制”这个词有两个意思：

（ⅰ）控制是没有得到处理的对象.

（ⅱ）控制试验是研究人员决定谁将在处理组的研究.

和非吸烟者比较，吸烟者结果很糟糕. 心脏病、肺癌等疾病在吸烟者中要更加常见. 在吸烟和疾病之间有很强的关联（association）. 如果香烟造成疾病，这就解释了这个关联，即吸烟者死亡率高是因为香烟有害. 一般来说，关联是因果关系的情况证据（circumstance evidence）. 然而，证明是不完全的. 可能会有某种隐藏的混杂因素，使得人们又吸烟又得病. 如果是这样，没有必要停止研究：这不会改变隐藏的因素. 关联和因果关系不同.

> 混杂意味着处理组和控制组之间的区别，是区别（而不是处理）影响着被研究的响应变量.

混杂因素一般来说是第三个变量，它和暴露相关联，并且影响着疾病的风险.

Joseph Berkson 和 R. A. Fisher 等统计学家不相信对香烟不利的证据，而且提出可能存在

混杂变量. 流行病学家（包括英格兰的 Richard Doll 和 Bradford Hill 及美国的 Wynder、Graham、Hammond、Horn 和 Kahn）做了认真的观测研究来表明这些另类的解释并不可信. 综合起来，这些研究强有力地证明了吸烟导致了心脏病和肺癌等疾病. 如果你放弃吸烟，你将更长寿.

流行病学研究经常对一些比较小的而且更加一致的群体分别做比较，假定在这些群体之中，对象如进行了随机化一样地分配到处理组或控制组. 例如，如果吸烟者不成比例地大部分是男性，那么，一个粗糙的关于吸烟者和不吸烟者的死亡率的比较可能被误导，因为男性比女性更可能得心脏疾病和癌症. 因此性别是一个混杂因素. 为了控制（又一次使用“控制”这个词）这个混杂因素，流行病学家比较男性吸烟者和男性不吸烟者，也在女性中做类似比较.

年龄是另一个混杂因素. 岁数较大的人有不同的吸烟习惯，而且患心脏病和癌症的风险更大. 因此在吸烟者和不吸烟者之间做比较时，应该依性别和年龄分别进行：比如，55～59 岁的男性吸烟者应该和同样年龄段的男性不吸烟者比较. 这是对性别和年龄的控制. 如果空气污染造成肺癌，而且吸烟者生活在更加污染的环境，空气污染则可能是一个混杂因素. 为了控制这个混杂因素，流行病学家在城市、郊区和乡下分别做比较. 最终，试图以混杂因素来解释吸烟对于健康的影响就变得非常不可能了.

自然，当我们以这种方式控制越来越多的变量时，研究的群体变得越来越小，使得机会本身发挥越来越大的影响. 这是用交叉制表方法来对付混杂因素的一个问题，也是使用统计模型的一个原因. 此外，大多数观测研究会不如对吸烟的研究那样令人信服. 下面的（稍微有些人工味的）例子说明了这个问题.

例 1 在跨国比较中，在一个国家，人均电话线路数量和它的乳腺癌死亡率有很强的相关性. 这并不是因为打电话造成癌症. 富国有更多的电话和较高的癌症病例. 对这种过多癌症风险的可能解释是，在富裕国家中，妇女有较少的子女. 怀孕——特别是较早的第一次怀孕——是起保护作用的. 在饮食和其他与生活方式有关的因素上，国与国之间的区别也可能起某些作用.

> 随机化控制试验使得混杂问题减到最小. 这也是从随机化控制试验得到的因果推断比从观测研究得到的结果更加有说服力的原因. 根据观测来做因果研究必须关心混杂. 处理组和控制组是什么？它们有什么区别（不是处理的区别）？做什么调整来对付这些区别？这些调整是否有意义？

这一章其余部分将讨论案例：乳房造影的 HIP 试验，Snow 对于霍乱的研究，贫穷的原因.

1.2 HIP 试验

乳腺癌是在加拿大和美国女性中最常见的恶性肿瘤之一. 如果该肿瘤发现得足够早——在它扩散之前发现，成功治愈的机会要大得多. “乳房造影”（mammography）意味着用 X 光对女性扫描以探测乳腺癌. 乳房造影能否及时探测到乳腺癌呢？首次大规模随机化控制试验为纽约的 HIP（Health Insurance Plan），接下去是瑞典的两县研究（Two-County study）. 还有 6

个其他的试验．某些结果是负面的（扫描没有帮助），但多数是正面的．但是到 20 世纪 80 年代后期，乳房造影已经被广泛接受．

HIP 试验是在 20 世纪 60 年代早期完成的．HIP 是一个群体医学实践，它在那时有约 70 万成员．试验的对象为 62 000 名年龄在 40～64 岁的妇女，均为 HIP 成员，她们随机地被分到处理组或控制组．“处理”由 4 次年度扫描的邀请组成，每次包括一次临床检查和乳房造影．控制组持续接受通常的健康护理．前 5 年的跟踪结果展示在表 1 之中．在处理组，有大约 2/3 的妇女接受邀请进行扫描，而有 1/3 拒绝．这里展示了死亡率（每 1000 名妇女）以便于比较不同样本量的组．

表 1　HIP 数据．组大小（四舍五入），5 年跟踪死亡数目，每千名妇女死亡率

	组大小	乳腺癌		所有其他原因	
		数目	死亡率	数目	死亡率
处理组					
扫描的	20 200	23	1.1	428	21
拒绝的	10 800	16	1.5	409	38
总数	31 000	39	1.3	837	27
控制组	31 000	63	2.0	879	28

哪些比率表明了处理的功效？在接受扫描和拒绝扫描的人之间做比较看来是自然的．然而，即使它出现在一个实验中，这也是一个观测比较．研究人员决定哪些对象将被邀请扫描，但是，由对象本人来决定是否接受扫描的邀请．较富裕和接受过较好教育的对象更有可能参加．而且乳腺癌（不像多数其他疾病）更容易袭击较富裕的人．社会地位因此成为与结果以及是否接受扫描的决定都有关联的一个混杂因素．

需要提请注意的是，表的最后一列由其他原因（非乳腺癌）造成的死亡率，在接受和拒绝扫描的人之间有很大的差别．拒绝者比接收者有几乎 2 倍的风险．由于扫描本身对这个风险不起作用，在接受和拒绝扫描者之间一定还有其他差别来说明由其他原因造成的死亡率的差异．

一个主要差别在于社会地位．较富有的妇女愿意扫描．富有女性不那么容易被其他疾病所攻击，但更易得乳腺癌．因此，对接受和拒绝扫描者所进行的比较是有偏的，而这个偏倚是不利于扫描的．

对那些接受扫描和拒绝扫描的人的乳腺癌死亡率的比较是*按所接受的处理所做的分析*（analysis by treatment received）．正如我们所看到的，这个分析是严重有偏的．这个试验比较是在整个处理组（所有那些被邀请扫描的成员，无论是否接受）和整个控制组之间进行的．这是*意向处理分析*（intention-to-treat analysis）．

意向处理分析是被推荐的分析．

作为一个进行得非常好的研究，HIP 做了意向处理分析．研究人员比较了整个处理组和控制组之间的乳腺癌死亡率，并且表明扫描有益．

邀请的效应在绝对数目上是小的：63－39＝24 个生命被挽救（表 1）．因为来自乳腺癌的绝对风险小，所以没有什么干预能对绝对数目有大的影响．另一方面，在相对意义上，从

乳腺癌的5年死亡率得到比例39/63＝62%．后来持续了18年的跟踪调查，挽救生命在这期间是持续的．两县研究是在瑞典实施的一个大规模的随机化控制试验，它证实了HIP的结果．在芬兰、苏格兰和瑞典进行的其他研究也得出同样结论．这就是乳腺造影被如此广泛接受的原因．

1.3 关于霍乱的研究

自然试验是观测研究，这里处理组或控制组的随机化好像是大自然安排的．1855年，在Koch和Pasteur奠定现代微生物学约20年之前，John Snow利用一个自然实验表明霍乱是一个水源性传染疾病．那时，疾病的细菌理论仅仅是许多理论之一．瘴气（腐烂的气味，特别是来自腐败有机体的气味）常常被认为是流行病的原因．体液（黑胆汁、黄胆汁、血液和粘液）的不平衡是疾病的较老式解释．土地里的毒素是稍后成为时尚的另一种解释．

Snow是伦敦的一个内科医生．根据观测疾病的过程，他得出霍乱是由一种小的有机生物造成的，它通过水或者食物进入人体，在人体内繁殖，并使得身体排出该生物的复制品的水液．然后这些排泄物污染了食物或重新进入水源，于是该生物继续感染其他牺牲者．Snow解释了在感染和发病之间的时间间隔（若干小时或若干天）为感染因素在牺牲者体内繁殖的时间．这种繁殖是生命的特征：无生命的毒素不会重新复制它们自己．（当然，毒素可能需要某些时间来为害：时间间隔不是强制性证据．）

Snow发展了一系列论据来支持他的细菌理论．比如，霍乱沿着人类贸易路线传播．还有，当一艘船来到霍乱流行的港口，水手们仅仅在他们接触了该港口的居民后才会得这种病．如果霍乱是传染病，那么这些事实很容易被解释，但很难用瘴气理论来解释．

在1848年，霍乱在伦敦流行．Snow识别了这个疾病的第一个“标志”病例：

> “一个刚刚从该疾病流行的汉堡乘Elbe号汽轮来的名叫John Harnold的海员．”[p. 3]

他还识别了第二个病例：一个名叫Blenkinsopp的海员，他在Harnold死后占用了其房间并且通过和床上用品接触而感染．其次，Snow还发现邻近的公寓建筑中，其中一幢霍乱流行，而另一幢没有．其中，被感染的建筑物使用被下水道污水污染了的水源，而另一个使用相对纯净的水源．这再次表明，如果霍乱是传染病而不是瘴气造成的，这些事实很容易被理解．

在1854年8月和9月，这种疾病再一次爆发．Snow做了一个“现场地图”，显示了受害者的位置．发病位置聚集在Broad街水泵附近．（Broad街在伦敦的Soho，那时，公共水泵用作饮用水源．）作为对照，在这个区域有一定数量的公共机构很少（甚至没有）人死亡，其中一处是酿酒厂．工人们宁愿喝淡色啤酒而不愿喝水，如果谁想喝水，在企业内有自己的水泵．另一个几乎没有霍乱的机构是一个贫民收容所，它也有自己的水泵．（贫民收容所一例将在1.4节再次讨论．）

在伦敦其他地区的人们也感染了该病．在大多数情况下，Snow能表明他们喝的水来自Broad街的水泵．比如，一位在Hampstead的女士如此喜欢这里的水的味道，以至于她找搬运工把Broad街水泵的水运到她家．

至此，已经有了令人信服的轶事证据表明霍乱是一种通过接触或水源传染的疾病．Snow

还使用了统计的思想．那时在伦敦有一些供水公司，其中有些从污染严重的泰晤士河流域取水，而另一些的水则相对来说没有被污染．

Snow 做了“生态学”的研究，把在伦敦各个地区的霍乱死亡率和水的质量相关联．一般来说，水被污染的地区死亡率较高．Chelsea 供水公司则是例外．这家公司也取污染了的水，但是使用一些较现代的方法来净化——用沉淀池，并认真过滤．因此，它所服务的地区霍乱死亡率较低．

在 1852 年，Lambeth 供水公司把它的取水管往上游迁移以得到较纯净的水．Southwark & Vauxhall 公司没有移动他们在污染严重的泰晤士河的取水管．Snow 进行生态分析，比较了在 1853—1854 年及更早的流行病时期这两家公司的服务区域．现在让他以自己的话来叙述．

> “虽然上面表［生态分析］中显示的事实提供了有深刻影响的强有力的证据，说明该疾病存在时，饮用含有城市污水的水助长了霍乱的扩散，但问题并不能到此为止；Lambeth 公司和 Southwark & Vauxhall 公司在伦敦相当大部分地区混杂供水的这一状况使得这个课题得以细查，以产生对结果的无可争议的证明．在上面表格所列举的由两家公司供水的区域中，混杂供应是最详尽的一类了．每家公司的水管遍布所有街道，进入几乎所有院子和小巷．有些房子由一家公司供水，而另一些由另一家公司供水，这完全依赖于在供水公司全力竞争时房主或住户所作的决定．在许多情况下，单独一所房子两边的供水公司可能都不同．每家公司都既为富人也为穷人供水，既为大房子也为小房子供水．接受不同公司供水的人在条件或职业上均没有区别．现在，很明显，在部分提供改良了的水的区域，霍乱的减少全凭了这种供水，被如此供水的房子享受减少该疾病的益处，而那些被提供来自 Battersea Fields 的［污染了的］水的房子就要承受像完全不存在改良水供应那样的死亡率．由于这些房子里的东西和接受这两家公司所提供的水的人，或者他们周围的物理条件，都没有什么不同，显然，在检验水的供应对霍乱进程的影响上，没有试验能够被设计得比这个更加彻底了，现成的情况就摆在观察家的眼前．”
>
> “该试验的规模也是最大的．这里有不少于 30 万人，这包括两种性别、各种年龄和职业，以及每种等级和身份，从名门世家到极度贫困者，他们没有选择地（大多数情况下他们并不知情）被分组，其中一组接受含有伦敦城市污水的水，水中含有可能来自霍乱病人的物质，而另一组则接受没有这些不纯物质的水．”
>
> “为了把这个宏大的试验变成论据，所需要的就是弄清楚发生霍乱的每个房子的水源供应．”［pp. 74—75］

Snow 的数据显示在表 2 中．分母数据为每家供水公司服务的房子数目，可在议会记录中得到．然而，分子数据则需要进行逐户调查以确定每一个霍乱死亡者居住地址的水源供应．（当时称为“bill of mortality”的死亡证书显示了地址，但没有说明每个死亡者的水源．）由 Southwark & Vauxhall 公司供水的死亡率大约 9 倍于 Lambeth 公司的．Snow 解释说，这个数据可以被当成随机化控制试验的结果来分析：两家供水公司的顾客之间除了水之外没有不同．数据分析很简单，就是死亡率的比较．是该研究的设计和效应的规模导致了结论．

表 2　按水源的霍乱死亡率，每千座房屋死亡率，伦敦，1854 年的流行病，Snow 的表Ⅸ

	房屋数目	霍乱死亡数	每千座房屋死亡率
Southwark & Vauxhall 公司	40 046	1 263	315
Lambeth 公司	26 107	98	37
伦敦其余地方	256 423	1 422	59

1.4　Yule 关于贫困原因的研究

Legendre（1805）和 Gauss（1809）发展了回归方法以拟合关于天体轨道的数据．有关的变量根据牛顿力学是已知的，联系它们的方程的函数形式也是已知的．度量可以是高精度的．在度量和方程中误差的性质也大都知道．此外，有大量的机会来比较预测和现实．一个世纪之后，研究人员把回归应用于社会科学数据，其中上述这些条件即使对粗糙的近似也并不成立，其后果是需要探索的（第 4～9 章）．

Yule（1899）研究贫困的原因．当时，英格兰的贫民㊀或者在称为贫民收容所的严酷的维多利亚式的机构内得到救济，或者在外面依赖于地方当局的政策救济．政策的选择是否影响贫民的数目？为研究这个问题，Yule 提出一个回归方程，

$$\Delta\text{Paup} = a + b \times \Delta\text{Out} + c \times \Delta\text{Old} + d \times \Delta\text{Pop} + \text{error}. \tag{1}$$

这个方程中，Δ 为随时间变化的百分比；Paup 为贫民数目；Out 为在外面救济比率 N/D，其中 N＝在贫民收容所外接受救济的数目，D＝在贫民收容所内接受救济的数目；Old 为 65 岁以上人口；Pop 为人口总数．数据来自于 1871、1881、1891 年的英格兰人口普查．这里有两个 Δ，一个是为 1871—1881，一个是为 1881—1891．（误差项后面将会讨论．）

救济政策分别在每个“区会”（union）中确定．（区会为救济贫民的教区联合组织，由几个教区组成．）那时，有约 600 个区会，而 Yule 把它们分为 4 类：乡下的、混合的、市内的、都市的．一共有 4×2＝8 个方程，每个相应于一类区会和一个时间段．Yule 用最小二乘法把他的方程拟合到数据，即确定 a，b，c，d，使得误差平方和

$$\sum(\Delta\text{Paup} - a - b \times \Delta\text{Out} - c \times \Delta\text{Old} - d \times \Delta\text{Pop})^2$$

最小．这是对所有已给类型的区会在已给时间段上求和，这里（实际上）假定了这些系数对所有那些地理和时间组合都是常数．

例如，考虑都市区会．拟合方程到 1871—1881 年的数据，Yule 得到

$$\Delta\text{Paup} = 13.19 + 0.755\Delta\text{Out} - 0.022\Delta\text{Old} - 0.322\Delta\text{Pop} + \text{误差}. \tag{2}$$

对于 1881—1991 年的数据，他的方程为

$$\Delta\text{Paup} = 1.36 + 0.324\Delta\text{Out} + 1.37\Delta\text{Old} - 0.369\Delta\text{Pop} + \text{误差}. \tag{3}$$

ΔOut 的系数相对较大，而且是正的．Yule 得出外面救济造成了贫困的结论．

㊀ 这里的贫民（pauper）是指当时在英国靠救济度日的人，他们或者到环境很差的贫民收容所得到救济，或者住在家里（即在收容所外面）得到救济．在 1834 年修正的济贫法把英国救济系统统一，并且把教区（perish）联合成区会来负责贫民收容所，又叫做工场（workhouse）．按照该法律，禁止能工作的人在家里接受救济，如果要想救济，必须去工场．而工场的条件故意弄得很恶劣以不鼓励贫民对救济的依赖．19 世纪末，工场条件得到改善．到 20 世纪初，社会福利服务和社会保险系统完全取代了工场．——译者注．

让我们看一下细节. 表 3 有关于贫困的 1881 年对 1871 年的比例、外面救济的比例、老年人口和总人口. 如果我们把表中每个数据减去 100，第 1 列给出回归方程 (2) 中的 ΔPaup，第 2、3、4 列给出了其他变量. 对于 Kensington (表中第一个区会)，

$$\Delta\text{Out} = 5 - 100 = -95, \quad \Delta\text{Old} = 104 - 100 = 4, \quad \Delta\text{Pop} = 136 - 100 = 36.$$

因此，根据 (2) 所得到的 ΔPaup 的预测值为

$$13.19 + 0.755 \times (-95) - 0.022 \times 4 - 0.322 \times 36 = -70.$$

实际的 ΔPaup 值为−73. 因此误差为−3. 如前所述，系数是 Yule 为了使均方误差减至最小而选择的. (第 4 章将说明如何实行.)

表 3　贫困、在外面救济比率、老年人口、总人口. 1881 年与 1871 年数据之比乘以 100. 伦敦市的区会. Yule (1899，表 XIX)

	Paup	Out	Old	Pop		Paup	Out	Old	Pop
Kensinton	27	5	104	136	Bethnal Green	46	19	102	106
Paddington	47	12	115	111	Whitechapel	35	6	93	93
Fulham	31	21	85	174	St. George's East	37	6	98	98
Chelsea	64	21	81	124	Stepney	34	10	87	101
St. George's	46	18	113	96	Mile End	43	15	102	113
Westminster	52	27	105	91	Poplar	37	20	102	135
Marylebone	81	36	100	97	St. Saviour's	52	22	100	111
St. John，Hampstead	61	39	103	141	St. Olave's	57	32	102	110
St. Pancras	61	35	101	107	Lambeth	57	38	99	122
Islington	59	35	101	132	Wandsworth	23	18	91	168
Hackney	33	22	91	150	Camberwell	30	14	83	168
St. Giles'	76	30	103	85	Greenwich	55	37	94	131
Strand	64	27	97	81	Lewisham	41	24	100	142
Holborn	79	33	95	93	Woolwich	76	20	119	110
City	79	64	113	68	Croydon	38	29	101	142
Shoreditch	52	21	108	100	West Ham	38	49	86	203

再回头看方程 (2). 系数 0.755 的因果解释是这样的：其他量都不变，如果 ΔOut 增加一个百分点，即行政区支持更多的在贫民收容所之外的人，那么 ΔPaup 将上升 0.755 个百分点. 这是一个定量推断 (quantitative inference). 在外救济造成了受救济的贫民数目的增长，这是一个定性推断 (qualitative inference). 把 ΔPop 和 ΔOld 引入到方程是为了控制可能的混杂因素，贯彻“其他量都不变”的想法. 对于 Yule 的论证，ΔOut 的系数为显著正数是重要的. 回归把定量和定性的方面编织在一起.

Quetelet (1835) 想要利用统计方法来揭示“社会物理学”，即人类行为的规律. Yule 利用回归来对贫困的社会物理进行推断. 但是这并不是很容易就能做到的. 混杂是一个问题. 按照 Yule 时代处于前列的福利经济学家 Pigou 的说法，管理较有效的区正在建造贫民收容所并正在减少贫困. 管理的有效性于是成为一个既影响假设中的原因又影响其效应的混杂因素. 经济状况可能是另一个混杂因素. Yule 有时把人口改变率描述为经济增长的替代. 然而，一般

来说，他很少关注经济状况．其解释为：

> “为了试验这个想法，花了大量时间和精力，但是结果证明是不满意的，最终该度量被完全放弃了．”[p. 253]

Yule 的方程形式有些任意，对于不同的时间和地点，系数并不一致：比较方程（2）和（3）可以看出，它们随着时间不同而不同．地点造成的不同在 Yule 文章的表 C 中报告了．这种不一致可能并不是致命的．然而，除非这些系数除了数据之外本身有存在的理由，否则它们怎么能预测可能改变数据的干预所造成的后果呢？对参数和估计予以区别是很基本的，我们在第 4～9 章将多次讨论这个问题．

还有其他的问题．充其量，Yule 建立了关联．以协变量为条件，在 ΔPaup 和 ΔOut 之间有正的关联．这个关联是因果关系吗？如果是，这个因果关系的箭头是指向哪个方向的？比如，针对贫民数目的短期增加，一个教区可能选择不去建设贫民收容所，这样，贫困造成了外部救济．类似地，一个区域的贫民数目可能被邻近区域的救济政策所影响．这样的问题并不能由数据分析来解决．相反，答案是事先假定的．和 Snow 关于霍乱的研究、HIP 试验或吸烟的流行病学相比，Yule 所做的实际上更有问题．

Yule 意识到这个问题．虽然他忙于对受救济贫民中的变化原因做分配：哪些归因于外部救济比率，哪些归因于其他变量，哪些归因于随机误差，但有一个老练的脚注（第 25 个）把所有的因果关系全收回了：“严格地说，‘归因于’意味着‘关联于’．”

除了没有包含星号来表示统计显著性的因果图之外，Yule 的研究非常现代．图 1 把他带到了今天．从 ΔOut 到 ΔPaup 的箭头说明 ΔOut 包含在解释 ΔPaup 的回归方程之中．“统计显著”由一个星号表示，而三个星号表示高度显著．其想法是：统计显著的系数是不同于零的，因此 ΔOut 对 ΔPaup 有因果影响．作为对照，不显著系数被认为是零：比如，ΔOld 对 ΔPaup 没有因果影响．我们在第 6 章还会讨论这些问题．

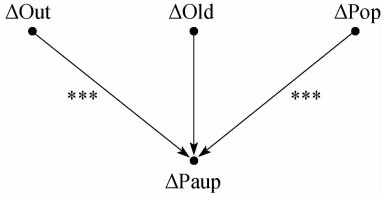

图 1　Yule 的模型．都市区会．1871—1881

Yule 本来可以利用回归来概括他的数据：对于已给的时间和某一类区会，对某种解释变量的值，被救济贫民的变化会是多少等等．换言之，在给定 ΔOut、ΔOld、ΔPop 时，他可以利用他的方程来近似 ΔPaup 的平均值．这假定了问题是线性的．关于预测，要有另一个假定：系统在时间变化时是稳定的．预测已经比描述更复杂了．另一方面，如果我们做一系列预测，并且用数据做检验，则有可能表明该系统足够稳定而使得回归有意义．

因果推断是不同的，因为系统中的一个改变是计划了的，是一个干预．描述统计告诉你关于你碰巧有的这个数据．因果模型断言，如果你故意改变一些数目，它将告诉你其他某些数目将会发生什么．这是一个值得研究的断言．有些东西在变化中必须保持不变．这是什么？为什么它必须是常数？第 4 章和第 5 章将解释如何拟合如（2）和（3）那样的回归方程．第 6 章讨论来自现代社会科学的一些例子，并且研究在变化之中的常数性假定，以证明用统计方法做因果推断的合理性．*响应方案*（response schedule）将用来把常数性假定形式化．

练习组 A

1. 在 HIP（表 1）试验中，有哪些证据表明处理对由其他原因造成的死亡没有影响？

2. 有人想要用接受扫描的妇女与控制组的比较来分析 HIP 数据．这个想法合适吗？

3. Snow 对 1853—1854 年的流行病研究（表 2）是随机化控制试验还是一个自然试验？为什么 Lambeth 公司在 1852 年移动其取水地点很要紧？简单解释．

4. Yule 的研究是随机化控制试验还是观测研究？

5. 在方程（2）中，假定 ΔOut 的系数是 -0.755．Yule 将会得到什么结论？如果该系数是 $+0.005$ 呢？

练习 6～8 为下一章做准备．如果对内容不熟悉，你可以读一下 Freedman-Pisani-Purves (2007) 的第 16～18 章，或者其他课本中的类似内容．记住

$$\text{方差}=(\text{标准误差})^2.$$

6. 假定 $X_1, X_2, \cdots, X_n$ 为独立随机变量，有共同的期望 μ 和方差 σ^2．令 $S_n = X_1 + X_2 + \cdots + X_n$．求 S_n 的期望和方差．再求 S_n/n 的期望和方差．

7. 假定 $X_1, X_2, \cdots, X_n$ 为独立随机变量，有共同的分布：$P(X_i=1)=p$ 及 $P(X_i=0)=1-p$，这里 $0<p<1$．令 $S_n = X_1 + X_2 + \cdots + X_n$．求 S_n 的期望和方差．再求 S_n/n 的期望和方差．

8. 大数定律是什么？

9. Keefe et al (2001) 概括他们的数据如下：

> “35 位风湿性关节炎患者记 30 天日记．参加者报告说有了精神上的体验，比如经常有和上帝结合的意愿．参加者对他们利用宗教应对方法来控制疼痛的能力评分为优，在那些日子里，他们很少会感觉到关节疼．”

该研究是否表明宗教应对方法在控制疼痛方面是有效的呢？如果不是，如何解释这个数据？

10. 按照许多教科书，关联不是因果关系．你多大程度上同意这个观点？简单讨论．

1.5 札记

实验设计（experimental design）本身就是一个主题．比如，许多实验把对象分成相对齐次的区组．在每个区组中，一些组被随机选为处理组，其余的为控制组．致盲（blinding）是另一个重要的主题．自然，实验可能出轨．关于这种情况的例子，参见由 Sundt (1987) 等人评论的 EC/IC Bypass Study Group (1985)．评论说，这个大规模多中心手术试验的管理和报告都失败了，这是许多可能通过手术康复的患者在试验之外动了手术，并且没有被统计到这份报告之中的结果．

流行病学（epidemiology）是研究医学统计的．更正式的说法为，流行病学是“在特定的人群中关于和健康相关的状态或事件的分布及影响因素的研究，以及这个研究在控制健康问题上的应用”．参见 Last (2001, p. 62) 及 Gordis (2004, p. 3)．

吸烟对健康的影响．参见 Cornfield et al (1959)，International Agency for Research on Cancer (1986)．从 Freedman (1999) 可看到简单的概括．已经有了一些关于停止吸烟的实验，但它们最多也是不确定的．类似地，能够做动物实验，但是从一个物种到另一个物种的外

推是困难的. 关于吸烟假设的批判评论包括了 Berkson (1955) 和 Fisher (1959). 后者的论点几乎有悖常理 (人无完人).

电话和乳腺癌. 包括了 165 个县, 相关系数为 0.74. 乳腺癌死亡率 (标准化的年龄) 来自 http://www-dep.iarc.fr/globocan/globocan.html 上的人口数据, 电话线数目 (还有其他的) 来自于 http://www.cia.gov/cia/publications/factbook.

HIP. 最好的原始资料为 Shapiro et al (1988). 实际的随机化机制包含了列表抽样. 表 1 中的差别在 18 年的追踪中一直持续着, 而且变得越来越显著, 这一点可以从取前 7 年而不是前 5 年的案例事件看出. 扫描停止于 4 或 5 年后, 而还需要一到两年效应才会显示出来, 因此 7 年可能是最好的可用时间段.

意向处理 (intention-to-treat) 度量分配的效应, 而不是扫描的效应. 扫描效应被人员转移所弱化, 因为仅有 2/3 的妇女来扫描. 当移动是单方面的, 即从处理组到控制组, 而没有相反的变动时, 很容易纠正这个弱化. 扫描减少了乳腺癌死亡率的效应 2 倍于不扫描的. 这个估计由两县研究所证实. 参见 Freedman et al (2004) 的回顾文章, 弱化的纠正在那里的第 72 页讨论, 另外参看 Freedman (2006b).

处理组中接受扫描的对象在非乳腺癌造成的死亡率上非常低 (表 1). 为什么? 一是, 接受者较富有并且受过较好的教育, 死亡率随着收入和教育水平的提高而下降. 此外, 一般来说接受者可能会更好地照顾自己. 参看 Freedman-Pisani-Purves (2007) 的 2.2 节及 Petitti (1994).

最近, 关于乳房造影的价值的问题又被提起, 但是来自扫描试验的证据是非常坚实的. 关于这方面的回顾, 参看 Smith (2003) 和 Freedman et al (2004).

Snow 关于霍乱的研究. 在 19 世纪末, 微生物学变得非常活跃. 在 1878 年, Pasteur 发表了 *La théorie des germes et ses applications à la médecine et à la chirurgie*[㊀]. 大约同样的时候, Pasteur 和 Koch 分离出了炭疽杆菌并发展了接种疫苗的技术. 接下去是结核杆菌. 在 1883 年, 在埃及和印度流行霍乱, 而 Koch 分离出霍乱弧菌 (先前 Filippo Pacini 在 1854 年的工作已经被人遗忘).

在 1892 年, 汉堡又暴发了流行病. 该城市的父老求助于当时德国卫生运动的首席人物 Max von Pettenkofer. 他不相信 Snow 的理论, 相反, 他坚持霍乱是由土地中的毒素造成的. 汉堡是屠宰工业的中心: 为了减少土地的污染, von Pettenkofer 把动物残骸挖出来拖走. 流行病一直持续到市民对 von Pettenkofer 失去信心, 并在绝望中求助于 Koch.

关于这段历史的参考文献包括 Rosenberg (1962)、Howard-Jones (1975)、Evans (1987) 和 Winkelstein (1995). 现在, 霍乱弧菌的分子生物学已经被相当好地理解了. 最近的回归文献有 Colwell (1996) 和 Raufman (1998). 对此的概要, 请参看 Alberts et al (1994, pp. 484, 738). 关于 Snow 工作的有价值的细节, 参看 Vinten-Johansen et al (2003). 还可以参看http://www.ph.ucla.edu/epi/snow.html.

在流行病学的历史上, 有许多类似于 Snow 在霍乱上所做工作的例子. 比如, Semmelweis (1860) 发现了产褥热的原因. Loudon (2000) 写的一本有趣的书描述了这段历史, 虽然

㊀ 细菌理论及其在医学和外科学的应用. ——译者注.

Semmelweis 大概能被更文雅地对待．另一个例子是在 1914 年左右，Goldberger 表明了糙皮病是饮食缺陷所造成的结果．Terris（1964）重新发表了许多 Goldberger 的文章，还可参见 Carpenter（1981）．脚气病研究的历史确实值得阅读（Carpenter，2000）．

Quetlet．几句话可以显示出他的工作风格．

> “把我的工作冠以社会物理学的标题没有别的目的，只是因为它以一个一致的秩序收集影响人类的现象，这类似于物理科学把属于物质世界的现象放到一起……在一个给定的社会状态，在某些原因的影响下，产生了一些有规则的效应，它们事实上围绕一个平均点震荡，不会经历任何可以察觉到的改变……
>
> “这个研究……有太多的吸引人的地方，它在太多的方面和科学的每一个领域相联系，而且所有都是哲学中最有意思的问题，该研究长期不会有热心的观察者来把它进行下去并且使它有科学的外观．”（Quetelet 1842，pp. vii，103．）

Yule．（1）和（2）中的“误差”在理论中扮演了不同的角色．在（1）中，它为随机误差，是统计模型无法观测的部分．在（2）中，它为残差，能够作为模型拟合的结果而被计算出来．方程（3）类似于（2）．细节在第 4 章介绍．关于这段历史具有同情心的记录，参看 Stigler（1986）和 Desrosières（1993）．Meehl（1954）提供了一些有名的通过回归进行预测的成功例子．在若干不同的背景下，制造一些真正的“事先”预测是预测有效性的最好证明：预测未来比拟合过去的回归方程要难得多（Ehrenberg and Bound 1993）．

John Stuart Mill．在试验和观测之间的对照及混杂的思想可追溯到 Mill（1843）．（在其第 7 版，参看 Book Ⅲ，第Ⅶ章和第Ⅹ章，特别是 423 和 503 页．）

试验和观测研究的对比．水果和蔬菜的流行病是试验和观测数据相矛盾的有名例子．简单地说，观测数据表明，那些吃了富含维生素的食物的人癌症发病率低，“因此”维生素防止癌症．实验表明，食物添加维生素既不能减少也不能增加患癌症的风险．

用观测数据研究的问题是，那些每天吃（比方）5 份水果和蔬菜的人在许多方面都不同于其余的人．用纯粹统计方法来确定所有这些区别是很难的（Freedman-Pisani-Purves，2007，p. 26 和 p. A6 的注 23）．研究文章包括 Clarke and Armitage（2002）、Virtamo et al（2003）、Lawlor et al（2004）和 Cook et al（2007）．Hercberg et al（2004）对男性而不是女性得到一个正面的效应．

激素取代疗法（HRT）是另一个例子（Petitti 1998，2002）．观测研究显示，HRT 防止了更年期后女性患心脏病．实验表明 HRT 没有好处．那些选择了 HRT 的妇女和其他女性有区别，在一些方面，这些区别在观测研究中被忽略了．第 7 章还将讨论 HRT．

Ioannidis（2005）表明利用各种干预来做实验比较，观测研究很少有可能给出可重复的结果．还可参看 Kunz and Oxman（1998）．

轶事证据（anecdotal evidence）基于个例，没有对不同的群体做系统比较，把它作为因果推断的基础是薄弱的．如果研究中没有控制组，特别在效应很小或很难度量的情况下，有理由对其产生强烈的怀疑．当效应非常引人注目时，比如用青霉素治疗伤口感染的情况，这些统计告诫可以放到一边．关于青霉素，参看 Goldsmith（1946）、Fleming（1947）、Hare（1970）和 Walsh（2003）．对效应较大的情况，Smith and Pell（2004）有一个很好的、非常幽默的关于因果推断的讨论．