HZ BOOKS
华章教育

CAMBRIDGE

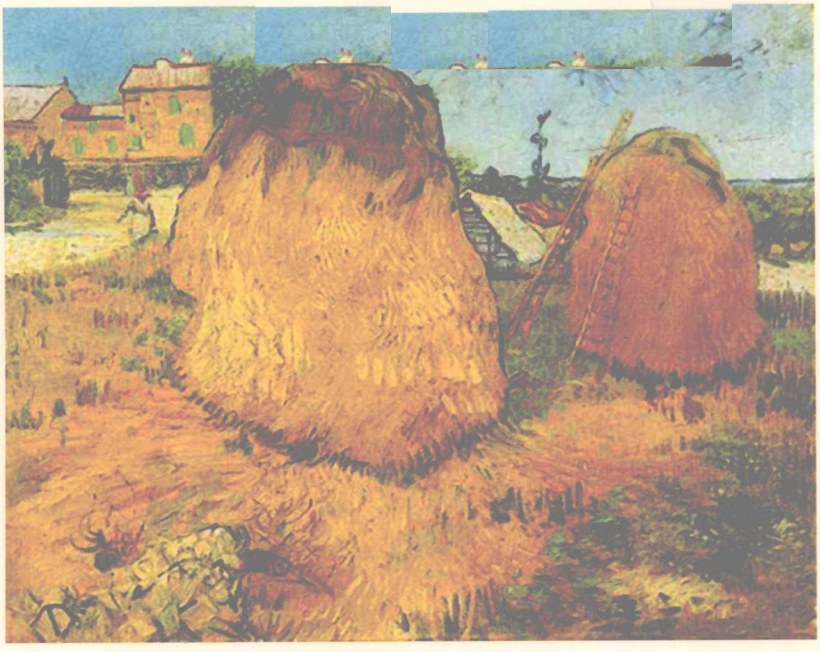
统计学精品译丛

统计模型

理论和实践（原书第2版）

Statistical Models Theory and Practice

(Revised Edition)



(美) David A. Freedman 著
加州大学伯克利分校

吴喜之 译

机械工业出版社
China Machine Press

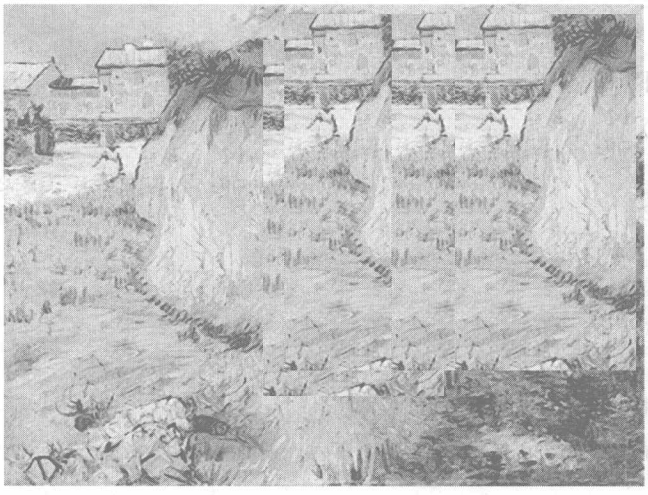
统计学精品译丛

统计模型

理论和实践（原书第2版）

Statistical Models *Theory and Practice*

(Revised Edition)



(美) David A. Freedman 著
加州大学伯克利分校

吴喜之 译

机械工业出版社
China Machine Press

本书是一本优秀的统计模型教材，着重讲解线性模型的应用问题，包括广义最小二乘和两步最小二乘模型，以及二分变量的 probit 及 logit 模型的应用. 本书还包括关于研究设计、二分变量回归及矩阵代数的背景知识. 此外，本书附有大量的练习，并且其中多数练习题在书后都有答案，便于读者学习、巩固和提高.

本书适合作为统计专业高年级本科生和低年级研究生线性模型课程的教材，同时也适合作为相关领域研究人员的参考书.

Statistical Models: Theory and Practice, Revised Edition (ISBN 978-0-521-74385-3) by David A. Freedman, first published by Cambridge University Press 2009.

本书版权登记号：图字：01-2010-3477

图书在版编目（CIP）数据

统计模型：理论和实践（原书第 2 版）/（美）弗里曼（Freedman, D. A.）著；吴喜之译. —北京：机械工业出版社，2010.6

（统计学精品译丛）

书名原文：Statistical Models: Theory and Practice, Revised Edition

ISBN 978-7-111-30989-5

Ⅰ. 统…　Ⅱ. ①弗…　②吴…　Ⅲ. 统计模型-研究　Ⅳ. C81

中国版本图书馆 CIP 数据核字（2010）第 111139 号

机械工业出版社（北京市西城区百万庄大街 22 号　邮政编码　100037）

责任编辑：王春华

北京京北印刷有限公司印刷

2010 年 9 月第 1 版第 1 次印刷

186mm×240mm · 18.25 印张

标准书号：ISBN 978-7-111-30989-5

定价：45.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

客服热线：（010）88378991；88361066

购书热线：（010）68326294；88379649；68995259

投稿热线：（010）88379604

读者信箱：hzjsj@hzbook.com

译 者 序

读这本书是一种完完全全的享受。自从伯克利加利福尼亚大学统计系郁彬教授在2008年向我推荐这本书之后，我一开始期望的是一本数学味很强的标准回归分析教材。后来，完全出乎意外，这本书竟然是我多年来企图寻找却又不可得的涉及回归分析甚至统计领域核心问题的一本以不寻常的清楚明白方式写的传奇式的读物。一眼就可以看出该书是出自大家的手笔。在应用统计于科学、医学和社会科学等领域方面，几十年来，本书作者David Freedman都一直被誉为统计的良心。该书是他在研究生命最旺盛的时期写的，代表了当代应用回归教科书的最高水平。作者不仅在伯克利，而且在世界许多高校都使用该教材讲过回归。多年的应用经验和教学实践使得该书内容丰富，语言平易近人，易教易学。该书的实际例子和练习题是精心挑选的，对掌握该书的内容不可或缺。

通常的回归或统计模型教材，无论标以理论或是应用的标签，往往对模型附加了很多假定，但又从来不解释如果这些假定不满足，则会发生什么问题或灾难。这本书不但不回避这些一般教师避之唯恐不及的关于模型的设置和假定等敏感问题，而且专门对各个领域最著名的、最有影响的文章的模型设置及各种假定进行认真的剖析。读这本书对于教师、学生，特别是实际工作者皆是一种心灵的震撼。我相信，任何有心人读了这本书之后，都会在未来涉及回归的课题上倍加小心，避免发生各种根本意想不到的错误。这本书会使许多人受益不浅，功德无量。

我对这本书的翻译是在2009年David Freedman去世之后，当时还不知道他在去世前已经定稿了修订版。因此，我先翻译了初版，后来又翻译了这一版。我恐怕是本书最忠实的中国读者之一。我希望那些在中国大学教本科生或者研究生回归模型课程的教师，能够以本书作为教材或者主要参考书，使得学生能够直接受益于国际一流统计大师的经验与智慧。

吴喜之

2010年4月

引　言

这本生动而又迷人的教科书解释了在阅读社会科学和医疗卫生领域的经验文章时所必须知道的东西，以及建立你自己的统计模型所需要的方法. 作者 David A. Freedman 解释了关联和回归的基本思想，并且带领你理解当前连接这些思想和因果关系的模型.

该书聚焦于对线性模型，包括广义最小二乘和两步最小二乘模型，以及二分变量的 probit 及 logit 模型的应用. 自助法是为了估计偏倚和计算标准误差而发展的方法. 对于统计推断的原理特别予以了专注. 此外，还包括了关于研究设计、二分变量回归及矩阵代数的背景知识. 为了学会方法，还有包含计算机样例程序的计算机实验. 该书有大量的练习，其中多数都有答案.

本书的目标读者包括统计专业高年级本科生和低年级研究生，以及社会科学和医疗卫生领域的学生和专业人员. 该书的讨论以及许多练习是围绕着发表过的研究所组织的. Freedman 对于这些文章及各种其他例子中的统计方法做了详尽的评价. 他描述了建模的原理以及陷阱. 这些论述为读者展示了如何思考一些重要的问题，包括在统计模型和实际现象之间的联系（或缺乏联系）.

本书的特点

- 在教学、研究及咨询方面有丰富经验的著名作者给出的权威指导.
- 将使得与应用统计打交道的任何人感兴趣.
- 直接而简洁的风格.
- 对主要在社会科学和卫生科学中的真实应用提出的统计问题做了认真分析.
- 可以作为教材或者自学的参考书.
- 在伯克利经过了多年课堂教学的检验.
- 包含回归和矩阵代数的背景材料.
- 含有大量的练习.
- 为教师提供额外材料，包括数据集及实验课题的 MATLAB 代码（发电子邮件至 solutions@cambridge. org 索取）.

关于作者

David A. Freedman（1938—2008）是加州大学伯克利分校的统计学教授. 他是杰出的数理统计学家，其研究范围包括鞅不等式分析、Markov 过程、de Finetti 定理、贝叶斯估计的相合性、抽样、自助法，以及关于因果推断方法的模型检验和评估的方法.

Freedman 发表了范围广泛的关于统计在社会科学中的应用和误用的研究，涉及流行病学、人口统计学、公共政策及法律. 他强调对标准方法背后的假定进行揭示和核对，以及当这些假定不成立时理解这些方法会如何表现，比如，当拟合来自随机化实验的数据时，回归模型会如何表现. 正如本书所表明的那样，他有把认真打磨的统计论述与引人注目的经验应用和例证整

合起来的非凡天赋.

Freedman 是美国科学院（American Academy of Art and Sciences）院士. 在 2003 年，他获得了美国科学院（National Academy of Sciences）授予的 John J. Carty 科学进步奖（John J. Carty Award for the Advancement of Science），以表彰他“对统计的理论及实践做出的意义深远的贡献”.

第 2 版序

有些书是正确的，有些书是清楚的，有些书是有用的，有些书是给人以享受的．即使是上面的两个优点，也很少有书全部具备，而这本书具有全部上面 4 项优点．本书明晰、公正而且具有深刻见解，读起来令人愉快．幸运的是，David Freedman 在 2008 年末去世之前完成了这个新的版本．我们为他的逝世深感哀悼，并非常钦佩他在最后的日子里带给这本书及许多其他计划的活力和振奋．

这本书清楚地介绍了应用统计中最常用的 6 种工具，这里没有难懂的行话及夸张之言．它解剖实际应用：该书的四分之一篇幅重印了依赖于统计模型的社会和生命科学的文章．它清楚地阐明了使这些工具正常运作所必需的假定，并且确定了这些假定的作用．这种清楚的表达使得学生及实际工作者可以较容易地看到：这些方法在什么情况下会是可靠的；在什么情况下有可能失败，并且有多么糟糕；在什么情况下另一种方法可能行得通；在什么情况下，无论用何种被人试图推销的工具，都不可能做出推断．

很多这个层次的教科书比“方法大全”好不到哪里，展示了几十种工具，缺乏说明及见解，像一本菜谱，是一种数目仅仅是数目的方式．“如果左边是连续的，利用线性模型，用最小二乘法来拟合．如果左边是离散的，利用 logit 或 probit 模型，用最大似然法来拟合．”以这种方式来教统计，诱使学生相信得到的参数估计、标准误差及显著性检验是有意义的，甚至可能揭开复杂的因果关系．他们教导学生把科学推断看成纯粹的运算法则．代入数字，就是科学了．这既低估了实体知识，也低估了统计知识．

选择一个适当的统计方法实际上需要认真思考数据收集的方法及其所度量的对象．数据并不“仅仅是数目”．在背后的假定是错误的情况下使用统计方法，既能产生金子，也能产生渣滓，但多半是渣滓．

本书通过展示有重大影响的研究所使用的好的和有问题的统计工具来给出上面的信息．这些研究包括：关于麦卡锡（McCarthy）时代的政治不容忍性的研究，就学于天主教学校对完成中学学业和进入大学的影响，生育力和教育之间的关系，政府机构在重组社会资产中的角色．其他例子来源于医学和流行病学，包括 John Snow 的关于霍乱病因的经典之作，这是简单统计工具加上实质性的知识及脚踏实地的工作而获得成功的闪闪发光的例子．这些实际应用给予理论以活力并给练习以动力．

高年级本科生和低年级研究生均能读懂本书．高年级研究生和成熟的研究工作者还会得到新的收获．我们三个人从阅读和教授这本书的过程中确实学到了不少东西．

仅仅读这本书并不能完全覆盖 Freedman 在这个领域的所有可以找到的研究．他的许多研究文章收集在《Statistical Models and Causal Inference：A Dialogue with the Social Sciences》（Cambridge University Press，2009）之中，它是本书有用的补充．该文集对本书提到的某些应用进行了更深入的探讨，比如霍乱病因以及激素替代疗法的健康效果等．此外，还涵盖从调整不足的普查到量化地震的风险等应用．有些文章涉及本书提到的一些理论问题．比如，实验

中的随机分配不足以说明回归是正当的：没有更进一步的假定，处理效果的多元回归估计是有偏的. 该文集还包括了统计的哲学基础和本书没有的一些方法，比如生存分析.

本书展示了重要应用和背后的理论，但没有丧失掉清晰易懂的特点. Freedman 以其智慧和明白的表述说明了统计分析如何能够揭示知识以及如何能够行骗. 这本书与众不同，它是一个宝藏：它是一本入门的书，具有做出可靠统计推断所要求的某些智慧. 它是 Freedman 传奇的一个重要部分.

David Collier，Jasjeet Singh Sekhon，Philip B. Stark
加利福尼亚大学，伯克利

前　言

这本书主要是为统计学专业的高年级本科生和低年级研究生准备的．社会科学和医疗卫生领域的学生和专业人员也会对本书感兴趣．虽然我把它写得像一本教科书，但是它其实自成体系．本书着重讲解线性模型的应用问题，包括广义最小二乘和两步最小二乘模型，以及二分变量的 probit 及 logit 模型．自助法是作为估计偏倚和计算标准误差的方法来讲解的．

恰当地说，要想开始阅读利用统计模型的经验性文章，本书的内容是必须知道的．全书所强调的是在模型和实际现象之间的联系或缺乏联系．多数讨论是围绕着已发表的研究成果进行的，为了易于参考，关键的文章重新印在书后．一些读者可能发现作者以怀疑的态度作为本书讨论的基调．若您也在这部分读者之中，那么我会做出一个不同寻常的建议，即在您读完本书之前，请保持这种怀疑态度．（一般来说，作者都要求读者暂时相信书中的结论，但本书不做如此要求．）

第 1 章对比了观测研究和实验研究，并引进了回归方法，这种方法有助于理清观测研究中的繁杂关系．本书中，有一章用来解释回归线，而另一章快速地复习了矩阵代数（在伯克利，半数主修统计的学生需要学习这些章节），知道这些内容，学生们会轻松很多．另外一个重要的附加课程是坚实的概率论和统计基础知识．

方法是通过实践来发展的．在伯克利，我们有实验室上机环节，在那里，学生利用计算机来分析数据．书后面有 13 个这样的实验（lab），一些我们给出了要点，此外，还附上了几个计算机程序样例．若想获取数据以及程序代码，教师可发邮件至 solutions@cambridge.org 索取．

好课本应该有好的练习，书中有大量的课后练习．这些练习题中有些是关于数学的，有些是假想的，它们是对一些引理和传统方法中的反例的模拟练习．另一方面，许多练习题都是基于实际研究．这里有数据的概括和分析，还有特别的一点：你如何下手？多数练习题的答案附在本书后．除了做练习和实验外，伯克利的学生在学期中还要完成一些课题的研究报告．

作为教材，一方面要确定选择什么来讨论，而另一方面要确定选择什么来忽略．无论一本书有多厚，都无法覆盖所有感兴趣的内容．我的目标是解释实际工作者如何从关联中推断出因果关系，而自助法则用来替代通常使用的渐近方法．检查该领域的逻辑性是至关重要的，而且需要时间．如果我们忽视了一种广受欢迎的方法，或许这种检查可以对比做出修正．

本书的内容足够用于本科生 15～20 周或研究生 10～15 周的课程和讨论．对大学期的本科生课程，我讲授第 1～7 章，并同时介绍 9.1～9.4 节．这通常需要 13 周．如果时间允许，我还会讲自助法（第 8 章）和第 9 章的例子．在 10 周的小学期，我将跳过学生的演示和第 8～9 章，以及第 7 章中二分变量的 probit 模型．

在学期的最后两周，学生展示他们的课程，或者在答疑时间和我讨论这些课题．我常常在最后一次课中总结一下．对于研究生课程，我增加了附加的案例分析和方法讨论．

本版的内容在安排上与前版稍有不同，这样使得教学更容易．我已经以某些其他方式对内容讲解做出了改进，（希望）没有引进新的困难．本版增加了许多新的例子和练习．

致谢

多年来，基于本书内容，我在伯克利，也在斯坦福和雅典教授过研究生和本科生课程. 这些课上的学生给予了我很大的帮助和支持. 我还要感谢 Dick Berk、M'aire N'ıBhrolch'ain、Taylor Boas、Derek Briggs、David Collier、Persi Diaconis、Thad Dunning、Mike Finkelstein、Paul Humphreys、Jon McAuliffe、Doug Rivers、Mike Roberts、Don Ylvisaker、Peng Zhao 及多位匿名的评审人的非常有益的意见. Ross Lyons 和 Roger Purves 是本书的合作者. David Tranah 是位出色的编辑.

目　录

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.