20世纪统计怎样变革了科学

# 女士品茶

David Salsburg 萨尔斯伯格（美）著
邱东 等译

中国统计出版社
China Statistics Press

还在督促自己每天进步一点吗？

还在坚持每天阅读的习惯吗？

还在为找不到自己喜欢的书籍烦恼吗？

那～

你愿意与我成为书友吗？

国内外当下流行书籍

各图书销量排行榜书籍

大量工具书籍

使我们受益终生的书籍

……

海量电子版、纸质版书籍及音频课程

还有贴心的“学习管家”服务哦！

微信：shuyou055

# The Lady Tasting Tea

## How Statistics Revolutionized Science in the Twentieth Century

第1章 女士品茶

第2章 偏斜分布

第3章 可爱的戈塞特先生

第4章 在“垃圾堆”中寻觅

第5章 收成变动研究

第6章 “百年不遇的洪水”

第7章 费歇尔获胜

第8章 致命的剂量

第9章 钟形曲线

第10章 拟合优度检验

第11章 假设检验

第12章 置信诡计

第13章 贝叶斯异论

第14章 数学界的莫扎特

第15章 “小人物”之见解

第16章 非参数方法

第17章 当部分优于总体时

第18章 吸烟会致癌吗？

第19章 如果您需要最佳人选

第20章 朴实的德克萨斯农家小伙

第21章 家庭中的天才

第22章 统计界的毕加索

第23章 处理有瑕疵的数据

第24章 重塑产业的人

第25章 来自黑衣女士的忠告

第26章 鞅的发展

第27章 意向治疗法

第28章 电脑随心所欲

第29章 “泥菩萨”

# 第1章 女士品茶

那是20世纪20年代后期，在英国剑桥一个夏日的午后，一群大学的绅士和他们的夫人们，还有来访者，正围坐在户外的桌旁，享用着下午茶。在品茶过程中，一位女士坚称：把茶加进奶里，或把奶加进茶里，不同的做法，会使茶的味道品起来不同。在场的一帮科学精英们，对这位女士的“胡言乱语”嗤之以鼻。这怎么可能呢？他们不能想象，仅仅因为加茶加奶的先后顺序不同，茶就会发生不同的化学反应。然而，在座的一个身材矮小、戴着厚眼镜、下巴上蓄着的短尖髯开始变灰的先生，却不这么看，他对这个问题很感兴趣。

他兴奋地说道：“让我们来检验这个命题吧！”并开始策划一个实验。在实验中，坚持茶有不同味道的那位女士被奉上一连串的已经调制好的茶，其中，有的是先加茶后加奶制成的，有的则是先加奶后加茶制成的。

写到这里，我可以想象，部分读者会对这种实验不以为意，认为它不过是一帮精英们于夏日午后的一个小消遣。他们会说：“这位夫人能不能区分两种不同的注茶方式，又有什么大不了的呢？这个问题并没有什么科学价值，这些大人物更应该把他们的天才用在对人类有所裨益的事情上去。”

不幸的是，不管外行对科学及其重要性怎么想象，从我个人的经验来看，大多数科学家之所以从事科研活动，只是因为他们对结果感兴趣，或者能够在工作中得到理性的刺激。好的科学家很少会想到工作的最终重要性，剑桥那个晴朗夏日的午后也是这种情景。那位夫人也许能、也许不能正确地品出不同的茶来，但这无关紧要，因为，实验的真正乐趣，在于找到一种判断该女士是对还是错的方案来。于是，在蓄着胡须先生的指导下，大家开始讨论应该如何进行实验判断。

接下来，在场的许多人都热心地加入到实验中来。几分钟内，他们在那位女士看不见的地方调制出不同类型的茶来。最后，在决战来临的气氛中，蓄短胡须的先生为那位先生为那位女士奉上第一杯茶，女士品了一小会儿，然后断言这一杯是先倒的茶后加的奶。　这位先生不加评论地记下了女士的说法，然后，又奉上了第二杯……

科学的合作性质

这个故事是我在20世纪60年代后期，从一个当时在场的先生那里听到的。这位先生就是休?史密斯（Hugh Smith），但他都是以H?费尔菲尔德?史密斯（H. Fairfield Smith）的名义发表科研论文。我认识他的时候，他在位于斯托尔斯（Storrs）的康涅狄格大学（the University of Connecticut）任统计学教授，而我则是两年以前在这个大学拿到了统计学博士学位。在宾州大学（the University of Pennsylvania）教了一阵子书后，我加入到了辉瑞公司（Pfizer Inc.）的临床研究部门。这是一家大型制药公司，它的研究园区坐落在格罗顿（Groton）,离斯托尔斯大约一个小时的车程。当时，我是那里唯一的统计学家。在辉瑞期间，我要处理许多疑难的数学问题，还要负责给他们讲解这些问题，并告诉他们，对这些问题，我个人的结论是什么。

在辉瑞工作期间，我发现，科研工作几乎不能独立完成，通常需要不同智慧的结合。因为，这些研究太容易犯错误了。当我提出一个数学公式作为解决问题的工具时，这个模型有时可能并不适合；或者我就所处理情况而引入的假设并不真实；或者我发现的“解”是公式中的失误部分推导出来的；甚至我可能在演算中出了错。

无论何时，我去斯托尔斯的大学拜访，与史密斯教授探讨问题，或者，与辉瑞的化学专家、药理专家坐在一起讨论，我提出的问题都会受到欢迎，他们对这种讨论充满兴趣和热情。对大多数科学家来说，工作中令他们最感兴趣的，就是解决问题时那种兴奋感。因此，在检验并试图理解问题时，他们期盼着与他人交流。

实验的设计

剑桥那个夏日午后的情形正是如此，那个留着短胡须的先生就是罗纳德?艾尔默?费歇尔（Ronald Aylmer Fisher），当时他只有三四十岁。后来，他被授予爵士头衔。1935年，他写了一本叫《实验设计》（The Design of Experiments）的书，书的第2章就描述了他的“女士品茶”实验。在书中，他把女士的断言视为假设问题，他考虑了各种可能的实验方法，以确定那位女士是否能做出区分。设计实验时的问题是，如果只给那位女士一杯茶，那么即使她没有区分能力，她也有50%的机会猜对。如果给两杯茶，她仍可能猜对。事实上，如果她知道两杯茶分别以不同的方式调制，她可能一下子全部猜对（或全部猜错）。

同样，即便这位女士能做出区分，她仍然有猜错的可能。或者是其中的一杯与奶没有充分地混合，或者是泡制时茶水不够热。即便这位女士能

做出区分，也很有可能是奉上了10杯茶，她却只是猜对了其中的9杯。

在这本书中，费歇尔讨论了这个实验的各种可能结果，他叙述了如何确定这样一些问题：应该为那位女士奉上多少杯茶？这些茶应该按什么样的顺序奉上？对所奉各杯茶的顺序应该告诉那位女士多少信息？依据那位女士判断的对错与否，费歇尔搞出了各种不同结果的概率。但在讨论中，他并没有指明这种实验是否真的发生过，也没有叙述这次实验的结果。

费歇尔书中有关实验设计的著述是科学革命的要素之一，这场革命在20世纪前半叶席卷了科学的所有领域。早在费歇尔出道以前，科学实验已经进行了几百年。在16世纪后期，英国的威廉?哈维（William Harvey）用动物做实验，他将不同动物静脉和动脉里的血液堵住，试图追踪血液从心脏到肺，回流到心脏，流向全身，再回到心脏的循环路线。

费歇尔没有发现实验是增长知识的方法。费歇尔之前，实验对每个科学家而言都是有其特性的。优秀的科学家可以做出产生新知识的实验，而二流的科学家常常从事的是积累数据的实验，但对知识增长没有什么用处。为说明这点，可以举发生在19世纪后期的一个例子。那时的科学家就测量光速做了许多无关要旨的努力，而直接到美国物理学家艾伯特?米切尔森（Albert Michelson）用光线和镜子建造了一个特别精巧的系列实验，才第一次得到好的估计。

在19世纪，科学家很少发表实验结果。他们所做的是论述自己的结论，并发表能证明结论真实性的数据。格雷戈尔?门德尔（Gregor Mendel）没有展示出他全部豌豆培育实验的结果，他叙述了他的系列实验，然后写道：“两组系列实验的前10个数据可以用来说明……”在20世纪40年代，费歇尔检验了门德尔用来说明结论的数据，发现这些数据过分完美，以至于失真，它们并没有表现出应该具有的随机程度。

尽管科学从审慎思考、观察和实验发展而来，但从来不清楚应该怎样从事实验，实验的全部结果通常也没有展现给读者。

19世纪末和20世纪初的农业研究中，上述情况尤为明显。20世纪早期费歇尔在农业实验站工作，在费歇尔去那儿工作之前，这个实验站已经进行了约90年的肥料构成（称之为人工肥料）实验。在一个典型的实验中，工人将磷肥和氮肥的混合物撒在整块田中，然后种植作物，测度收成和整个夏季的雨量。这里有精巧的公式用来“调整”某年或某块地的产

量，以便与另一块地、或同一块地的另一年产量相比，这被称为“肥力指数”。每一个农业实验站都有自己的肥力指数，而且都认为自己的指数是最精确的。

90年的实验结果不过是一堆未经发表、了无用处的混乱数据。看来某些品种的小麦对某种肥料反应优于其它品种，但只是在降雨过量的年份如此。其它实验似乎显示：第一年用钾硫化物，第二年用碳酸硫化物，会使某些品种的马铃薯增产，而对其它品种并非如此。因此，就这些人工肥料，充其量可以说，其中有些在有的时候，可能或大概有效。

作为一个卓越的数学家，费歇尔审视了农业科学家用来修正实验结果的肥力指数，这些指数是用来解释不同年份气象变化所造成的差异的，他还检查了其它农业实验站所用的同类指数。当简化为基本的代数式时，这些指数不过是同一公式的不同表现形式，换句话说，看似激烈争斗的两个指数，其实起着同样的修正作用。1921年，费歇尔在农业科学领域的领军期刊《应用生物学年报》(the Annals of Applied Biology)上发表了一篇论文，文中他指出了采用哪种指数并没有什么差异，并且，所有修正都不足以调整不同地块上的肥力差异。这篇非凡的论文终止了一场持续20多年的科学论战。

费歇尔接着检查了过去90年来的雨量和收成数据，指出年度间不同气候的影响远远大于不同肥力的影响。用费歇尔后来在他的实验设计理论里发明的一个词来说，“混合”（confounded）的，这意味着用已有的实验数据是不能将二者分开的。90年的实验和20年的科学论战几乎是无谓的浪费。

这使得费歇尔专注于实验和实验设计的思考。他的结论是：科学家需要从潜在实验结果的数据模型开始工作，这是一系列数据公式，其中一些符号代表实验中将被搜集的数据，其它则代表实验的全部结果。科学家从实验数据开始，并计算与所考虑科学问题相应的结果。

让我们考虑一个关于一个老师和某个学生的简单例子。这个老师非常想找出一些关于这个孩子学习情况的测试数据，为了达到这个目的，老师对孩子进行了一组考试，每一个考试都在0到100之间评分，任何一个单一的考试都不可能对孩子知识的掌握提供可靠的评估；这个孩子可能是没有学习多少考试所涉及的内容，但是知道不少考试以外的事情；可能是这个孩子在参加考试那天头疼；还可能是参加考试那天早上孩子与父母发生了争执。由于种种原因，单一考试不能对知识量提供好的估计，

所以老师进行了一组考试，然后计算出所有考试的平均分来评价孩子的知识量。这样的估计结果会更好，多少分是孩子知识量的实验结果，而每一个单独考试的分数则是数据。

那么老师应该如何组织考试？是搞那种只包括几天前所教授内容的系列考试，还是每次考试都从考试前所教授的全部内容中提取一部分？考试是一个星期搞一次，还是每天搞一次？或者在每个教学单元结束时搞？所有这些都是实验设计涉及到的问题。

如果农业科学家想知道某种人工肥料对小麦生长的效用，就要构建一个实验以取得效用估计时所需要的数据。费歇尔表明，实验设计的第一步是建立一组数学公式，用以描述待搜集数据与欲估计结果之间的关系，因此，任何有用的实验必须是能够提供估计结果的。实验必须是有效的，能够让科学家测定出气候的差异和不同肥料的使用对产量差别的影响。特别是，有必要包括同一实验中打算加以比较的实验处理（treatments）,即那些后来被称为“控制组件”（controls）的东西。

在他那本关于实验设计的书中，费歇尔提供了几个实验设计的范例，并导出优秀设计的一般原则。然而，费氏方法中所涉及到的数学非常复杂，多数科学家设计不了自己的实验，除非他们遵循费歇尔书中提出的实验设计中的某个模式。

农业科学家认识到费歇尔工作的伟大价值，在大多数说英语的国家中，费氏方法很快便成为农业科研的主流学派。从费歇尔的原创性工作出发，用来论述不同实验设计的完整科学文献发展起来。这些设计被应用到农业以外的领域，包括医学、化学和工业质量管理。在许多案例中，所涉及的数学高深且复杂，但此时此刻，我们不妨停下来想想，科学家不可能不假思索地动手实验，这通常需要长时间的审慎思考，而且，其中通常会有大量的、高难的数学。

至于前面所说的女士品茶——那个在剑桥晴朗的夏日午后所做的实验中，那位女士怎样了呢？费歇尔没有描述这项实验的结果，但史密斯教授告诉我，那位女士竟然正确地分辨出了每一杯茶！

# 第2章 偏斜分布

像人类思想史上的许多革命一样，要想找到统计模型成为科学组成部分的确切时刻，也是很难的。人们可以在19世纪初德国和法国数学家的工作中找到可能存在的特例，甚至在17世纪伟大的天文学家约翰尼斯?开普勒（Johannes Kepler）的论文中，也能找到某种启示。正像本书前言中所提到的那样，拉普拉斯（Laplace）发明了误差函数来说明天文学中的统计问题，但我仍然倾向于把统计革命的发生定位于19世纪90年代K?皮尔逊（Karl Pearson）的工作。查尔斯?达尔文（Charles Darwin）把生物变异认作生命的基本面，并将之作为适者生存理论的基础。然而，是他的英国伙伴K?皮尔逊首先认识到统计模型的根本性质，以及这种模型对19世纪科学中的决定论观点提供了哪些不同的东西。

当我在20世纪60年代开始学习数理统计时，K?皮尔逊的名字在课上很少被提到。当我与这一领域的大人物共同探讨一些问题时，也听不到对K?皮尔逊及其著作的参考。他或者是被忽略了，或是被视为行为早已出局的次要人物。例如，美国国家标准局（the U.S. National Burean of Standards）的邱吉尔?艾森哈特（Churchill Eisenhart）当时正在伦敦大学学院（University College,London）学习，那是K?皮尔逊人生的最后几年，艾森哈特记忆中的K?皮尔逊不过是一个精神头不足的老头儿。统计研究的步伐已经将他推出局外，他和他的工作被埋进故纸堆中，青年学生神采飞扬，集聚在新的大人物周围学步，其中之一，便是K?皮尔逊自己的儿子，但是没有人去拜见老皮尔逊，他的办公室孤零零地坐落在那里，远离着活跃的、振奋人心的新研究。

当然并不总是如此，在19世纪70年代，年轻的K?皮尔逊离开英国，到德去从事政治科学的研究生学习。在那里，他倾心于卡尔?马克思（Karl Marx）的著作，为了表达崇拜之情，他把自己名字的拼法从Carl改成Karl。带着政治学博士的学位，他回到了伦敦，并在这个领域写过两本值得重视的著作。在维多利亚时代的英国，伦敦的拘谨之风最甚，K?皮尔逊却大胆地效仿德国和法国上流社会的沙龙，组织了一个青年男女谈话俱乐部（Young Mens and Womens Discussion Club）。俱乐部的青年男女平等地聚焦在一起（未婚少女并没有人陪伴），讨论世界上重大的政治和哲学问题。K?皮尔逊正是在那种环境下与夫人相遇而结缘的，这个事实使人感到发起这类俱乐部可能另有动机。这个小小的社会冒险对我们进入K?皮尔逊的内心世界提供了帮助，可以见证他对已经建立起来

的传统是那样地不以为意。

尽管拿的是政治学博士学位，K?皮尔逊的主要兴趣还是在科学哲学和数学模型的性质上。19世纪80年代，他发表了《科学的法则》（The Grammar of Science），这本书后来再版了多次。在第一次世界大战之前的一段时间里，它被视为关于科学和数学性质最伟大的著作之一，其中充满了闪光的、原创性的、最具洞察力的见解，这使该书成为科学哲学的一本重要著作。同时，它又是以流畅、简单的风格写成，任何人都可以接受，你不必懂得数学就可以理解《科学的法则》。尽管从写作之日算起，这本书已经有100多年的历史了，但其中充满洞察力的见解和思想，对21世纪的数学研究，仍然是适用的。而它所提供的对科学性质的理解，至今也是真实的。

高尔顿的生物统计实验室

在人生的这个时段，K?皮尔逊感受到了英国科学家弗朗西斯?高尔顿（Francis Galton）爵士的影响。大多数人知道高尔顿这个名字，缘于他是指纹现象的“发现者”。高尔顿的贡献是认识到指纹对每一个人都是独特的，此外，还有通常用于识别和分类指纹的方法。指纹的唯一性存在于手指类型中出现的不规则标识和切面，这被称为“高尔顿标识”（Galton Marks）。高尔顿做的远比这多，作为一个只是将生物学算作其业余爱好的科学家，通过数字模型的研究，他寻求将数学的严密引入生物学，这同样是富有价值的。他所初创的各种调查当中的一项，是对天才遗传的研究。在这项研究中，他搜集了有关父子的信息，这些人因智商高而闻名。但由于当时对智力的测量没有什么好的办法，他发现研究这个问题特别困难，于是他决定转向诸如身高之类的遗传特性的研究，因为这更容易测量些。

高尔顿在伦敦成立了生物统计实验室（biometrical laboratory），并打广告动员不同的家庭来做测量。在这个实验室，他搜集身高、体重数据，测量特殊的骨骼和家庭成员的其它特性。他和他的助手将这些数据列成表格，并一再检验，他是在寻找利用父母测度数据来推断子女的某些办法。比如说，很明显，高个子父母很容易有高个子的小孩，但是不是存在某些数学公式，只用父母的身高就可以预测孩子将有多高呢？

相关与回归

高尔顿用这种方法，发现了他称之为“向平均回归”（regression to the

mean）的现象，这表现为：非常高的父亲，其儿子往往要比父亲矮一些；而非常矮的父亲，其儿子往往要比父亲高一些。似乎是某种神秘的力量，使得人类的身高从高矮两极移向所有人的平均值。不只是人类身高存在着向平均数回归的现象，几乎所有的科学观察都着了魔似的向平均值回归。在第5章到第7章，我们将看到，费歇尔如何能够将高尔顿向平均值回归的思想纳入统计模型，而这种模型现在支配着经济学、医学研究和工程学的很多内容。高尔顿仔细思考了他的惊人发现，而后认识到这必定是真实的，在进行所有观察之前这就是可以预言的。他说，假设不发生这种向平均值的回归，那么从平均意义上看，高身材父亲的儿子将与他们的父亲一样高，在这种情况下，一些儿子的身材必须高于他们的父亲，以抵消身材比父亲矮小者的影响，使平均值不变。高身材者这一代人的儿子也将如此，那么会有一些儿子身材更高。这个过程将一代一代延续下去。同样地，将会有一部分儿子身材比他们的父亲矮小，而且有一部分孙子将更加矮小，如此下去，不用多少代，人类种族就将由特别高和特别矮的两极构成。

上述的情形并没有发生，人类的身高在平均意义上趋向于保持稳定。只有当非常高的父亲其儿子平均身材变矮，而非常矮的父亲其儿子的平均身材变高，才能出现这种稳定。向平均值回归是一种保持稳定性的现象，它使得某给定物种代际之间大致相同。

高尔顿发现了这种关系的一种数学测度，他称之为“相关系数”（coefficient of correlation）。高尔顿给出了明确的公式，以计算这个系数，所用的资料则是在生物测量实验室搜集的。这是一个非常详细而明确的公式，它只计算了向平均值回归的一个方面，但没有告诉我们任何有关这种现象原因的信息。正是在这个意义上，高尔顿最先使用了“相关”这个字眼，这之后它演变进入了大众词汇。与高尔顿特定的相关系数相比，“相关”经常被用来表示更为模糊的东西，尽管“相关”本身有严格的科学含义。科学圈外的人经常说到这个词，似乎它描述了两种事物如何相联系，但除非你涉及到高尔顿的数学测量，否则，当你使用高尔顿用于特别目的的“相关”这个词时，它不必那么精确。

分布与参数

有了这个计算相关的公式，高尔顿实际上已经非常接近新的革命性观念了，这个观念革命在20世纪几乎修正了所有的学科。但却是他的弟子K?皮尔逊，在非常完整的意义上第一个规范地阐明了这个观念。

为了理解这个革命性的观念，你必须将已有的关于科学的成见抛开。通常我们被教导，科学就是测量，我们进行精心的测量，并用它来寻找描述自然的数学公式。在高中的物理课中我们学过，当时间给定时，一个自由落体的运行将遵循一个含有符号“g”的公式，这里的“g”是关于重力加速度的常量。我们学过可以用来确定“g”的值的实验。然而，当高中生们进行一系列确定值的实验时，顺着斜板滚动小球，并测量小球需要多长时间到达不同的位置时，发生了什么呢？这就是很少得出确切的结果。学生进行实验的时间越长，困惑就越多，因为不同的实验得出了不同的“g”值。老师仅凭自己优越的知识来审视学生的实验，并认定学生之所以得不到正确的结果，要么是因为工作草率，要么是因为不够细致，要么是抄错了数据。

老师没有告诉学生的是：所有的实验都是草率的，并且，即使是最精心的科学家，也很少得到确切的数值。不可预见和不可观察的小扰动在每一个实验中都有：室内的空气可能太潮湿，或者落体在滚动前卡住了一个微秒，旁边飞过的蝴蝶可能会有其影响：造成气流的轻微扰动。人们从一个实验中真正得到的是散乱的数据，其中没有一个单个数据是确切的，但所有这些数据可以用来对确切值进行近似的估计。

武装了K?皮尔逊的革命性观念，我们就不再将实验结果看作精心测量得出的数据，它们也不是本来就确切的，用更容易接受的术语来代替：它们是一组散布数据，或一个数据分布中的样本。数据的分布可以写成数学公式，它告诉我的数值是不可预测的，我们只能谈论概率值而不是确定值，单个实验的结果是随机的，在这个意义上看它们是不可预测的，然而，分布的统计模型却使我们能够描述这种随机的数学性质。

科学家花了一些时间才认识到观测值所固有的随机性质。在18和19世纪，天文学家和物理学家创造出描述他们观察值的数学公式，达到了可接受的精确程度，在为测量工具不够精确，所以观察值与预测值之间的是预料之中的，可以忽略不计。星体和其它天体的运动被假定遵循运动基本公式所确定的精确路径，其不确定性是由于简陋的测量工具造成的，并不是其固有的性质。

随着物理学中更为精确的测量工具的发展，随着将这种测量科学扩展到生物学和社会学的尝试，大自然所固有的随机性越来越明显了。怎么处理它？一种办法是坚持数学公式的精确性，将观测值与预测值之间的离差视为小的、无关紧要的误差。事实上，早在1820年，拉普拉斯的数学论文描述了第一个概率分布，即误差分布，那是一个与这些小的、无关

紧要的误差相联系的概率的数学公式。这个误差分布以钟形曲线（bell-shaped curve）或正态分布（the normal distribution ）的说法进入了大众的词汇。

这使K?皮尔逊比正态分布或误差分布更进了一步，审视生物学中积累的数据。K?皮尔逊认为，测量值本身，而不是测量的误差，就具有一种正态分布。我们所测量的，实际上是随机散布的一部分，它们的概率通过数学函数——分布函数被描述出来。K?皮尔逊发现了被他称为“偏斜分布”（skew distribution）的一组分布函数，他宣称，这组函数可以描述科学家在数据中可能遇到的任何散布类型，这组函数中的每一个分布由四个数字所确定。

用来确定分布函数的这些数字与测量中的数字不属于同一类型，这些数字决不会被观察到的，但可以从观测值散布的方式中推导出来。这些数字后来被称为参数（parameters——源自希腊语，意思是“几乎测量”（almost measurements））。能够完整地描述K?皮尔逊体系中数字的四个参数分别被称为：

1. 平均数（the mean）——测量值散布状态的中间值；

2. 标准差（the standard deviation）——测量值的散布与平均值偏离有多远；

3. 对称性（symmetry）——测量值在平均值一侧规程的程度；

4. 峰度（kurtosis）——个别的观测值偏离平均值有多远。

用K?皮尔逊偏斜分布体系去考虑问题，思路会有一种微妙的转移。在K?皮尔逊之前，科学所处理的事情都是真实的。开普勒试图发现行星如何在空间运行的数学规律；威廉?哈维的实验打算确定血液如何在某一特定动物的静脉和动脉中游动；化学则处理元素和由元素组成的化合物。然而，开普勒所试图追踪的“行星”实际上是一组数据，用来给地球上的观测者所看到的天空中微弱的光点定位。单匹马身上血液通过静脉流动的实际情形，也许与在另一匹马或者一个人身上所可能看到的不同。没有人能够生产出纯铁的样本，尽管谁都知道铁是一种元素。

K?皮尔逊提出，这些观测到的现象只是一种随机的映像，不是真实的，所谓的真实是概率分布。科学中真实的东西并不是我们所能观测到或能

把握到的，它们只是通过用来描述我们所观测事物随机性的数学函数来反应。科学调查中我们真正想确定的，是分布的四个参数。从某种意义上说，我们永远不能确定这四个参数的真实数值，而只可能从资料中估计它们。

K?皮尔逊并没有意识到这关键的一点，他以为，如果我们能够搜集到足够的数据去估计参数，就会得到参数的真实数值。而他的年轻对手费歇尔指出，K?皮尔逊的许多估计方法并不是最优的，在20世纪30年代末期，当K?皮尔逊临近他漫长生命的终点之际，一位杰出的波兰年轻数学家耶日?奈曼（Jerzy　Neyman）表明，K?皮尔逊的偏斜分布体系并没有包含所有可能存在的分布，许多重要问题不能用K?皮尔逊的体系解决。

还是让我们离开1934年那个被离弃的老皮尔逊吧。回到他三四十岁、精力充沛的时期，那时的他对自己所发现的偏斜分布充满了热情。1897年，他接管了高尔顿在伦敦的生物统计实验室，带领一支年轻的娘子军（被称为“计算员”），计算高尔顿所积累的人种测量数据的分布参数。在20世纪之交，高尔顿、K?皮尔逊和R?韦尔登（Rerhael　Weldon）共同努力，创办了一个新的科学期刊，这将使K?皮尔逊的观点应用到生物数据上。高尔顿用他的个人财富建立了一个信托基金支持这个期刊。在第一期，编辑们提出了一个雄心勃勃的计划。

生物统计计划

当时，英国科学家中有一位杰出的人物，他就是达尔文，同期的科学家们致力于探索达尔文富有洞察力的见解，高尔顿、K?皮尔逊和韦尔登便是其中相当热心的骨干。达尔文的进化理论认为，生命形式随着环境压力而变化，他提出，变化的环境会给更适应新环境的随机变化提供些许的优势，渐渐地，伴随着环境改变和生命形式继续发生随机转变，新物种将会出现并且更适于在新的环境中生存和繁殖。这一思想被简称为“适者生存”（survival of the fittest）。当恣意妄行的政治学家将其用于社会生活，宣称那些在经济竞争中取得胜利的富人比身陷贫困的穷人更为适于生存时，这一理论对社会就有不好的影响——适者生存理论成了猖狂的资本主义的辩护者，在那里，富人被授予了道义上的特权去鄙视穷人。

在生物科学中，达尔文的思想似乎很有道理。达尔文可以指出相关物种的相似性，作为现代物种从先前物种演化而来的佐证。达尔文表明，物种上些许不同的小型鸟类，即使是生活在孤岛上，也有许多解剖学上的

共性。他指出，不同物种胚胎之间的相似性，这包括人类的胚胎，在开始是有尾巴的。

有一件事是达尔文做不到的，那就是他不能给出人类历史的时间框架中，新物种实际出现的例子。达尔文设定新物种由于适者生存而出现，但没有证据，他不得不做的只是展示现代物种很好地适应了它们所处的环境。达尔文的说法似乎只是表明了已知的事情，而且理论本身有一个很吸引人的逻辑结构，但是如果套用犹太人的一句老话就是“举例并不是证明”（For instance is no proof）。

K?皮尔逊、高尔顿和韦尔登打算在他们的新期刊中将这事搞清楚。在K?皮尔逊看来，只有概率分布是真实的，达尔文的雀鸟（他在书中用到的一个重要例子）并不是科学调查的对象，而某一种雀鸟的总体随机分布才是这个对象。对某一给定雀鸟种类而言，如果能够测量其全体的喙长，这些喙长的分布函数将有四个参数，这四个参数将是这一种雀鸟的喙长。

K?皮尔逊说，假如存在着某种环境力量，通过提供优越的生存能力，使得某一物种产生某种特定的随机变化，我们也许不能生存得那么久，以看到新物种的出现，但我们能够看到分布的上个参数的变化。在他们期刊的创刊号上，三位编辑宣布：他们的新期刊将从全世界搜集数据，以确定这些分布的参数。最终期望表明，样本参数的变化与环境变化相关。

他们将新期刊定名为《生物统计》（Biometrika），高尔顿创建的生物统计基金会给予它慷慨资助。由于资金是这样地充裕，以至于该期刊成为世界上第一本印有全彩照片的期刊，甚至还带着画有复杂图画的下班纸折页。期刊以高品质的优质纸印刷，连最复杂的数学公式也展示了出来，尽管那意味着极端复杂和昂贵的排版工艺。

接下来的25年里，《生物统计》发表了通讯员们从各地发来的数据：有的深入非洲的丛林，测量原住民的胫骨和腓骨；有的从中美洲的雨林抓到奇特的热带鸟类，测量其喙长；还有的甚至偷盗古墓，揭开死人头盖骨灌铅，以测量其脑的容量。在1910年，该期刊发表了几幅全彩照片，画面是俾格米男人裸躺在地上，的生殖器旁还摆着量尺。

在1921年，一个年轻的女通讯员朱莉亚?贝尔（Julia　Bell）描述了她在试图对阿尔巴尼亚新兵进行人类形体测量时所遇到的困难。她离开维也

纳去阿尔巴尼亚一个边远的基地，本以为可以得到讲德语军官的帮忙，当她抵达时才发出，那里只有一个士官能说三句德语。她无所畏惧地拿出了测量所用的铜标尺，通过形体动作让那些年轻人理解她要干什么，直到他们按要求抬起手臂和脚。

对每一组这样的数据，K?皮尔逊和他的计算员们都计算出分布的四个参数，论文将展示最佳分布的图示，并评论该分布与其它相关数据的分布有何不同。回顾过去，很难看出所有这些行动怎样帮助证明了达尔文的理论。浏览《生物统计》的这些作品，我得到这样一种印象：这些工作不久就变成为自身原因而进行努力，除了给特定数据组估计参数外，没有实际目的。

在期刊中还夹杂着其它类型的论文，其中一些涉及理论数学，以处理发展概率分布时遇到的问题。比如在1908年，一个不知姓名的作者，以“学生”（“student”）为笔名发表了论文，提出了后来几乎在所有现代科学工作中都有作用的研究成果——“学生”的“t检验”。接下来的几章我们还会遇到这位匿名的作者，并将讨论他在K?皮尔逊与费歇尔之间作调解时的不幸角色。

高尔顿死于1911年，而韦尔登则于这之前死于阿尔卑斯山的一次滑雪事故。只剩下了K?皮尔逊这唯一的编辑和信托基金的支配者。在接下来的20年中，期刊成了K?皮尔逊个人的了，期刊发表什么完全以K?皮尔逊的判断为准，由他确定重要与否。K?皮尔逊为期刊写了很多社论，他让自己丰富的想象驰骋在各个领域。比如，在对一个古老的爱尔兰教堂翻修时，墙壁中发现了一副骨骼，K?皮尔逊通过对这些骨骼的测量和所涉及的数学推理，来确定它们事实上是不是某个中世纪圣徒的遗骨。再比如，一个据称是奥利弗?克伦威尔（Oliver Cromwell）的头骨被发现了，K?皮尔逊以一篇精彩的文章对其进行了研究。该文描述了所知的克伦威尔尸体的下落，并且还将对克伦威尔画像所做的测量结果和该头骨所做的测量进行了比较。在另外一些论文中，K?皮尔逊检验了古罗马各君主的统治期和贵族阶级的没落，还涉猎了社会学、政治学和植物学。所有这些，都带有复杂的数学解释。

就在去世之前，K?皮尔逊还发表了一篇题为“论犹太人与非犹太人关系”（On Jewish – Genlile Relationships）的短文。文中他分析了从世界各地收集到的犹太人与非犹太人的人体测量数据，最后得出的结论是：德国国家社会主义（the National Socialists）（正式的名称是纳粹（Nazis））的种族理论纯粹是胡说八道，根本就没有犹太种族（Jewish

race）或亚利安种族（Aryan race）那回事。这最后一篇论文与他以前的工作一样，组织清晰，有逻辑性，推理谨慎。

K?皮尔逊运用数学研究了人类思想的许多领域，而很少有人将这些领域视为科学的正宗地盘。浏览生物统计上他所写的社论，你仿佛看到了一个兴趣十分广泛的人，他具有直切问题核心的惊人能力，并能用数学模型去加以处理。还有浏览这些社论，你就像遇上一个意志坚定、主见鲜明的人。说实话，如果不需要与他争辩的话，我想我是很乐意与K?皮尔逊共处一天的。

K?皮尔逊他们是否证明了达尔文适者生存的进化论理论呢？也许是吧。通过将古墓中头骨的容量分布与现代男女的比较，他们设法证明：经历了几千年深化的人类种群保持了相当的稳定。他们表明：对澳洲原住民的人类学测量与对欧洲人的测量结果有着相同的分布，据此，他们推翻了某些澳洲人关于原住民不是人类的断言。K?皮尔逊从这些工作中发展了一种被称为“拟合优度检验“（goodness of fit test）的基本统计工具，这是现代科学所不可缺少的。它使科学家能够确定一组给定的观测值是否适合于某一特定的数学分布函数。在第10章我们会看到，K?皮尔逊的儿子E?皮尔逊（Eqon Pearson），是如何用这种拟合度检验是否定他父亲所完成的许多项工作的。

随着20世纪的来临，《生物统计》中讨论数理统计理论问题的文章越来越多，少量的文章仍停留在处理特定数据的分布。当K?皮尔逊的儿子E?皮尔逊接班成为编辑时，期刊的性质就完全转型为理论数学了。时至今天，《生物统计》仍是这个领域中卓越的刊物。

但他们到底有没有证明适者生存这个说法呢？20世纪初曾经有一个最接近的研究。韦尔登构想了一项宏大的实验：18世纪英格兰南部瓷器工厂的发展，导致了一些河道被粘土淤塞，普利茅斯（Plymouth）港和达特茅斯（Dartmouth）港也都受到了影响，近陆地区比近海地区淤得更为严重。韦尔登从这些港口抓了几百只螃蟹，分别放入广口瓶中，其中一半用内港的淤泥水，另一半用外港的较干净的水。一段时间后仍有螃蟹存活，韦尔登测量它们的壳，以确定两组螃蟹的分布参数。

正像达尔文所预言的那样，淤泥水中戚的螃蟹在分布参数上有了变化！这是不是证明了进化论呢？不幸的是，韦尔登在写出实验结果前就死了，K?皮尔逊对数据进行了粗略的分析，他描述了这个实验及其结果，但最后的分析却始终没有搞出来。为这项实验提供资助的英国政府要求

提供最终报告，但报告了无踪影，韦尔登死了，实验也夭折了。

就生命周期很短的生物，如细菌和果蝇而言，达尔文的理论最终被证明是真实的。用这些物种，科学家可以在较短的一个时间段里完成几千代的实验。现代的DNA研究，作为遗传的基石，已经为物种之间的关系提供了更为有力的证据。如果我们假定突变率在过去千万年或更长的时间里保持不变，那么DNA的研究可以用来估计灵长类和其它哺乳动物出现的时间框架，至少它经了几百万年。大多数科学家现在都把达尔文的进化论作为正确的东西接受下来。没有其它理论与所知数据吻合的如此之好，于是科学界满足了，原来人们认为需要通过确定分布参数转变来表明较短时间里的进化过程，一日三餐这种观念已经被放弃。

K?皮尔逊的革命所留下来的是这样一个观念：科学的对象并不是不可观测事物本身，而是数学分布函数，以描述与所观测事物相联系的概率。今天，医学研究运用精巧的分布数学模型来确定治疗方法对长期存活的可能效果；社会学家和经济学家用数学分布来描述人类社会的行为；物理学家用数学分布来描述次原子粒子。科学里没有哪一个方面从这场革命中逃脱。有的科学家宣称，概率分布的使用只是一时的权宜之中，最终我们会找到一种途径回到19世纪科学的决定论。爱因斯坦有句名言，他不相信上帝在和宇宙玩骰子，就是这种观点的例子。其他人则相信，大自然基本上是随机的，真实性只存在于分布函数之中。不管一个人的基本哲学是什么，事实仍然是，K?皮尔逊关于分布函数和参数的思想统治了20世纪的科学，并在21世纪初仍保持着优势。

# 第3章 可爱的戈塞特先生

爱尔兰都柏林的吉尼斯酿造公司（Guinness Brewing Company）是一个声誉卓著的老牌酿造公司，该公司于20世纪初开始投资于科学。年轻的吉尼斯刚刚继承这家企业，他就决定雇用牛津和合格大学在化学上顶尖的毕业生，以便将现代科学技术引进到公司的业务中来。在1899年，他招募威廉?西利?戈塞特（William Sealy Gosset）进入公司，那是个23岁的牛津大学新秀，拥有化学和数学两个学位。戈塞特的数学背景在当时是传统的，包括微积分、天文学和机械式宇宙观下的其它科学分支，K?皮尔逊的创新和后来成为量子力学的萌芽观念，还没有进入大学的课程。戈塞特是由于他的化学专长而被吉尼斯雇用的。对一个酿酒企业来说，要一个数学家又有什么用呢？

戈塞特成为吉尼斯一项很好的投资，他表明自己是一个很能干的管理者，最后他在公司里升任负责大伦敦区业务的主管。事实上，他对本行工艺做出了第一项主要贡献是以数学家的身份来完成的。几年前，丹麦电话公司（the Danish telephone company）是第一个雇用数学家的实业公司，但他们有一个明确的数学问题：制造多大的电话交换板？可制造啤酒又有什么数学问题需要解决呢？

戈塞特在1904年发表了第一篇文章，处理的是这样一个问题：麦芽浆准备发酵的时候，需要仔细地测量所用酵母的量，酵母是活的有机体，酵母培育需要保持鲜活，加入麦芽浆前它在瓶中的液体里系列。工人们得到测量清楚某个给定的瓶中有多少酵母，以便决定用多少液体，它们提取一定量的液体，在显微镜下检验，计量他们所看到的酵母细胞数。这种测量有多精确？了解这一点是很重要的，因为麦芽浆中所用的酵母数应该精确地控制。酵母太少，发酵不充分；太多了，啤酒又会发苦。

注意这个问题与K?皮尔逊对科学的观念是多么的吻合。测量的是样本中酵母细胞的量，但所寻求的真实“东西”是整个瓶中酵母细胞的浓度。由于酵母是活的，而细胞不断地分裂和繁殖，那个“东西”实际上并不存在，在某种意义上，真正存在的是单位液体中酵母细胞的概率分布。戈塞特检验了数据，确定酵母细胞的数量可以用所知的泊松分布（Poisson distribution ）来描述，这并不是K?皮尔逊偏斜分布家族中的一种概率分布。事实上，它是一种只有1个（而不是4个）参数的特殊分布。

确定了样本中的活酵母细胞数服从泊松分布，戈塞特就能够设计规则和

测量方法，从而得到对酵母细胞浓度更为精确的测量。用戈塞特的方法，吉尼斯能够生产质量更稳定的啤酒。

“学生”的诞生

戈塞特想找一份适合的期刊发表这个结果，泊松分布（或相应的公式）已经被发现100多年了，过去一直试图在现实生活中寻找实例，其中之一，便是计量普鲁士军队中被马踏死的士兵人数。在酵母细胞计量中，戈塞特有一个清楚的实例，还有对统计分布新观念的重要应用。然而，这违背了公司不准许雇员发表文章的政策。几年前，吉尼斯一位优秀的酿造师写了一篇文章，其中泄露了他们某个酿造过程的秘密成份。为了避免进一步损失，吉尼斯禁止它的雇员发表文章。

戈塞特成了当时《生物统计》编辑之一的K?皮尔逊的好朋友，而K?皮尔逊对戈塞特的数学能力印象很深。1906年，戈塞特说服了他的老板，数学的新思想对啤酒公司是很有用的，并到高尔顿生物统计室在K?皮尔逊门下脱产学习一年。这之前两年，当戈塞特描述他处理酵母的结果时，K?皮尔逊急于将之付印于他的期刊。他们决定用匿名的方式发表文章，于是，戈塞特的首次发现是仅是以“学生”的名义发表的。

在其后30年中，“学生”写了一系列极为重要的论文，几乎所有的都发表在《生物统计》上。从某些方面看，吉尼斯家族已经发现了他们“亲爱的戈塞特先生”违反了公司的规定，一直私下里撰写并发表科学论文。“学生”的数学活动大多是在家里进行，并且是在正常的工作时间之外。戈塞特在公司升迁到了负更多责任的位置，这表明他的副业并没有使吉尼斯公司受损。有这样一种不足为凭的说法：吉尼斯家族第一次知道这件事是在1937年，戈塞特突然死于心脏病，他数学界的朋友与吉尼斯公司探讨，想帮助支付其论文集的印刷成本。不管这事真实与否，美国统计学家哈罗德?霍特林（Harold Hotelling）的回忆录里清楚地记载，霍特林在20世纪30年代后期要与“学生”会谈，安排是秘密的，带有间谍小说的各种情节。这表明“学生”身份的真正确认，对吉尼斯公司仍是个秘密。“学生”在《生物统计》发表的论文涉及理论和实践的尖端问题，戈塞特将非常实际的问题带入有难度的公式，又把结论带回现实实践，后来者便照此办理。

尽管有很高的成就，戈塞特仍是个谦逊的人。在他的信中，人们经常可以发现这样的字眼：“我的研究只是提供了粗浅的想法”；或者，当他的某些发现被给予过多的荣誉，他会说：“费歇尔实际上已经能完成了整

个数学结构。”在人们的记忆中，戈塞特是一个和善的、体贴的同事，很在意别人的情感。他去世的时候61岁，离开了他的妻子马乔里（Majory）（一个精力充沛的运动员，曾经担任英国女子曲棍球队的队长）、一个儿子、两个女儿和一个孙子，当时他的父母还健在。

“学生”的t检验

如果不算别的，所有的科学家都受惠于戈塞特的一篇短文，该文的题目是“平均数的可能误差”（The Probable Error of the Meam），1908年发表在《生物统计》上。是费歇尔点出这篇杰出论文的一般性意义。对戈塞特来说，有一个特定的问题需要解决，一到晚上，他就习惯性地带着耐心和小心投入于这个问题。发现了结论，他就用其它资料来检查，重新验证他的结果，努力去确认是否遗漏了什么细微的差别，考虑他必须设定哪些假设，并一再重复计算自己的发现。他提前采用了现代计算机基础上才出现的蒙特卡罗技术（Monte Carlo techniques）,这是一种一再模拟的数学模型，以确定相关数据的概率分布。然而，当时他没有计算机，只能不辞辛苦地加总数据，从上百个样本中计算平均数，并绘制所得出频率的图表，所有这些都靠手工完成。

戈塞特所专注的特定问题是小样本（small　sample）问题。K?皮尔逊计算了某一分布的4个参数，这是在单一样本就积累了上千个测量数据的基础上完成的，因为使用了大样本，他设定所得到的参数估计是正确的。费歇尔要证明他的错误。根据戈塞特的经验，科学家很少能三八线以有如此大的样本，更为典型的实验通常能够看到10到20个观测数据，他还理解到，这种现象在所有的学科中都很普遍。在一封给K?皮尔逊的信中，他写道：如果我是你遇到的用小样本工作的唯一一人，那你太特异了，在这个题目上我与斯特拉顿（Stratton）（剑桥大学的一位研究员）相伴，他曾经用4个样本来做说明。

K?皮尔逊所有的工作都假定：样本足够大，以至于确定参数可以没有误差。戈塞特设问：如果是小样本会怎么样？我们将如何处理自己的计算中肯定会出现的随机误差？

晚间，戈塞特坐在自己的餐桌旁，取出一小组数据，算出平均值和标准差估计值，再将二者相除，并将结果绘在图纸上。他发现这个比率与K?皮尔逊的四个参数相关，并与K?皮尔逊的偏斜分布系列中的某一分布相配。他的伟大发现在于：你不必知道原始分布的4个参数的确切值。前两个参数估计值的比率有一个可以制表的概率分布，不管数据从哪里

来，或者标准差的真实值是多少，计算这两个样本估计值的比率，你就可以得到一个已知的分布。

正如弗雷德里克?莫斯特勒（Frederick Mosteller）和约翰?图基（John Tukey）所指出的那样，没有这一发现，统计分析注定要使用无限次的回归，没有“学生”的t检验 （这是该发现后来的称谓），分析者将不得不估计观测数据的4个参数，再估计这4个参数估计值的4个参数，接着估计4个新估计值的4个参数……这样继续下去，没有机会得到最终的结果。戈塞特表明，分析者可以在第一步就停止这种估计。

戈塞特的工作有一个基本的假设，即原始测量值服从正态分布。多年以来，科学家使用着“学生”的t检验，许多人渐渐相信，并不需要这项假设。他们经常发现：不管原始测量是否服从正态分布，“学生”的t检验都有相同的分布。在1967年，斯坦福大学（Stanford University）的布拉德利?埃弗龙（Bradley Efron）证明了这一点，更确切地说，他发现了不需要戈塞特假设的一般条件。

随着“学生”t检验的发展，我们不知不觉地习惯于统计分布理论的应用，这一理论在科学界广为流传，相伴而来的是更深层次的哲学问题，这就是我们所说的“假设检验”（hypothesis tests）或“显著性检验”（significance tests）的使用。后面我们会剖析这个问题，现在我们只想强调：“学生”提供了几乎每个人都使用的科学工具，尽管没有多少人真正理解它。

与此同时，“可爱的戈塞特先生”成了两个长期不和的超级天才——K?皮尔逊和费歇尔之间的中间人。尽管他经常对K?皮尔逊抱怨他看不懂费歇尔写给他的东西，他还是保持了与两个人的友谊。他与费歇尔的友谊开始于费氏在剑桥大学读本科的时候，那是在1912年，费歇尔刚刚成为剑桥大学数学学位甲等及格者（最高的数学荣誉），他的天文学导师 介绍两个人认识。当时费歇尔正在研究一个天文学问题，他写了一篇论文，在其中他重新发现“学生”在1908年得到的结果。年轻的费歇尔显然不大知晓以前戈塞特所做的工作。

在费歇尔给戈塞特看的这篇论文中，有一个小错误被戈塞特指了出来。当戈塞特回家的时候，他发现费歇尔写的两大页数学论证正等着他。这个年轻人把自己原先的工作又做了一遍，并加以扩充，还批评了戈塞特所犯的一个错误。戈塞特在给K?皮尔逊的信中写道：“附上一封信，它证明了我关于“学生”t检验的频率分布公式，您是否介意替我看一下。即

使我可以理解，超过三维空间我还是觉着不自在。”费歇尔用多维几何证明了戈塞特的成果。

在这封信中，戈塞特说明了自己的如何到剑桥去与朋友会面，而这个朋友恰巧在冈维尔与凯厄斯学院（Gonville and Caius College），是费歇尔的导师，他如何被介绍给这位22岁的学生。他接着写道：“费歇尔这小子写了一篇论文，提出概率的新标准或诸如此类的东西，看起来不错，但就我所能理解的，是一种不切实际且不大管用的认识事物方式。”

在描述了他在剑桥与费歇尔的讨论后，戈塞特写道：

对我们之间的讨论，他的回复是两大页书写纸，上面用最深的墨水写满了他所证明的数学（跟着是一组数学公式）……我看不大懂这些内容，回复他说等我闲下来时准备研究它，实际上我去湖区时随身带着它，可弄丢了。

现在他将这封信寄给我，我觉得如果它还可以的话，您也许愿意发表这个证明，它是这样的完美和数学化，对某些人也许有吸引力。

K?皮尔逊在《生物统计》上发表了费歇尔的短文，就这样，20世纪最伟大的天才之一面世了。3年以后，经过了一连串俯就的信件往来，K?皮尔逊发表了费歇尔的第二篇论文，但事先约定论文须以这种形式出现：它不过是对K?皮尔逊合作者之一所做工作的细微补充。K?皮尔逊再也没有允许他的期刊发表费歇尔的论文。费歇尔继续在K?皮尔逊许多最感自豪的成就中挑毛病，而K?皮尔逊则在稍后几期的《生物统计》中，以社论的方式点出“费歇尔先生”或“费歇尔先生的学生”在其它期刊所发表论文中的错误。这些都将是下一章介绍的内容，戈塞特会在以后几章中的某些地方再度出现，作为一个和蔼可亲的良师益友，他帮助年轻男女进入统计分布的新世界。他的许多学生和合作者都对新数学做出了重要贡献。尽管他本人谦逊地表示异议，但戈塞特的确做出了许多影响深远的贡献。

# 第4章 在“垃圾堆”中寻觅

1919年春天，费歇尔29岁，他带着妻子、三个孩子和小姨子，搬到了伦敦北部的一间旧农舍里，那儿靠近罗森斯特农业实验站（the Rothamsted Agricultural Experimental Station）。从许多方面来看，费歇尔的人生在别人眼里是失败的。他在孤单和多病的童年中长大，并有严重的视力损伤。为了保护他的近视眼，医生禁止他在人工灯光下阅读。但他很小就接触了数学和天文学，在6岁时他迷上了天文学，七八岁时，他就跑去听由著名天文学家罗伯特?鲍尔（Robert Ball）爵士主讲的通俗讲座。

费歇尔被著名的哈罗公学（Harrow Public School ）录取，在那里他的数学是出众的。由于不允许他使用电灯，他的数学导师在晚上教他时，不用铅笔、纸和任何其它视觉辅助品。久而久之，费歇尔发展了一种很强的几何直觉能力。在后来的岁月中，他那非凡的几何洞察力，使他得以解决许多数理统计中的难题。这种洞察力对他而言是那么明显，从而导致他经常不能被别人所理解。在他看来是显而易见的事情，别的数学家往往要花几个月甚至几年的时间去证明。

他于1909年进入了剑桥，在1912年获得了数学学位甲等及格者的头衔，对剑桥学生来说，这是一个很高的荣誉，要得到它必须通过一系列极为困难的口头和笔头数学考试，一般一年只会有一两个学生成功，有的年份甚至没有人能得到这种头衔。当费歇尔还是本科生时，他就发表了他的第一篇科学论文，其中复杂的迭代公式（iterative formulas）被转换成多维的几何空间形式。在这篇论文中，那些在人们眼里一直特别复杂的数学计算公式被转换成简单的几何形式。毕业后他花了一年时间，研究统计力学（statistical mechanics）和量子理论（quantum theory），到1913年，统计革命已经进入了物理学，而新观念已经较为系统地进入这两个领域，并成为正式的大学课程。

费歇尔的第一份工作是在投资公司的统计室，其后他突然离开那里，到加拿大去从事农场工作。后来又在第一次世界大战开始时突然离开农场，回到了英格兰。虽然他被批准入伍，但他那很差的视力使他免于军事服务。战争年代，他在许多公共学校教授过数学，但每一次的经历都比上一次更糟，他对学生们没耐心，因为他们都是不能理解在他看来很明显的事情。

费歇尔与K?皮尔逊

前一章提到，当费歇尔还是本科生时，就在《生物统计》发表了一篇短文。这使得费歇尔有机会见到K?皮尔逊，K?皮尔逊将一个困难的问题介绍给费歇尔：确定高尔顿相关系数的统计分布。费歇尔对此作了思考，用几何公式来处理它，不到一个星期就得出了完整的答案。他把结果交给K?皮尔逊，想在《生物统计》上发表。但K?皮尔逊不能理解其中的数学，把它转给了戈塞特，而戈塞特在理解上也有困难。K?皮尔逊知道如何就特定的案例得到问题的部分结论，他的方法涉及到大量的计算工作，于是便对生物统计实验室的工人做出安排，让他们去计算出这些明确的答案。在每一个案例中，所得到的答案都更加支持费歇尔的一般性结论。但K?皮尔逊仍然不发表费歇尔的论文，他要费歇尔做出修改，并降低费歇尔工作的一般性。K?皮尔逊将费歇尔的东西扣了一年多，同时让他的助手（计算员）计算一个庞大的扩展的表，以表明参数值的分布。最后，他发表了费歇尔的成果，但相对于K?皮尔逊及其助手展示分布表的大块文章来说，费氏的论文只是作为一个脚注。对不经意的读者来说，这样一个结果意味着：K?皮尔逊和他的合作者所做的工作更为重要，那里有大量的数据计算，而费歇尔的数学处理只是一个附属物。

费歇尔再也没有在《生物统计》上发表过文章，尽管它是这一领域的顶尖级期刊。在接下来的年份里，费歇尔的论文出现在《农业科学期刊》（the Journal of Agricultural Science）、《皇家气象学会季刊》（the Quarterly Journal of the Royal Meteorological）、《爱丁堡皇家学会会刊》（the Proceedings of the Royal Society of Edinburgh）、《心理研究学会会刊》（the Proceedings of the Society of Psychical Research）上，而所有这些期刊与数学研究通常都不怎么搭界。据知情者说，费歇尔作出这样的选择是因为K?皮尔逊和他的朋友们成功地将费歇尔逐出数学和统计研究的主流。根据其它人的说法，K?皮尔逊吹毛求疵的态度让费歇尔感到自身受到漠视，同时，他也没能够让类似的论文在《皇家统计学会期刊》（the Journal of the Royal Statistical Society，该领域另一份顶尖的期刊）上发表，于是他转而利用其它期刊，有时甚至付钱请他们发表自己的论文。

费歇尔这个“法西斯”！

费歇尔早期论文有一些是高度数学化的。他论述相关系数的文章，也就是K?皮尔逊最后同意发表的那篇，就充满了数学符号，一个标准页里有

一半甚至更多篇幅都是数学公式。但也有一些论文里面压根就没有数学。其中的一篇，他讨论了用达尔文的随机适应理论（Darwin’s theory of random adaptation）来说明最复杂的解剖学结构的方法。在另一篇论文中，他探讨了性别选择进化的问题。费歇尔在1917年加入了优生学运动（the eugenics movement），在《优生学评论》（the Eugenics Review）上发表了一篇社论，呼吁转变国民政策“以增加职业界人士和高技能工匠的生育率”，并抵制下层社会的生育率。他在这篇文章中质疑政府为贫民提供福利的政策，认为这会鼓励他们多生育，并将基因传给下一代，而中产阶级对经济安全的关注会导致他们推迟结婚，并节制生育。费歇尔担心，对整个国家来说最终的结果是：为后代选择了“最差的”而不是选择“较好的”基因。优生学问题是通过有选择的系列来改进人类基因库，这成为费歇尔的主要政治观念。在第二次世界大战期间，他被错误地指责为法西斯主义者，并被逐出了与战事有关的工作。

费歇尔的政治见解与K?皮尔逊不同，后者钟情于社会主义和马克思主义，他同情被压迫者，并喜欢挑战保守的优等阶层。但K?皮尔逊的政治观念对他的科学研究没有什么影响。费歇尔关注优生学，这导致他将相当大的精力投入到遗传学的数学研究中。当时有一种新观念，认为某种植物或动物的特性可能来自一个单个基因，这以两种形式中的一个就可表现出来。从这种观念出发，费歇尔将格雷戈尔?门德尔 的工作大大地推进了，他指出如何估计两个相信基因的彼此影响。

存在着控制生命性质的基因，这一观念是科学中广义统计革命的一个部分。我们观察植物和动物的我，专业上称之为“表型”（phenotypes）。但我们假设这些表形是基因之间交互作用的结果，而这些基因的交互作用又具有不同的概率。我们寻求以这些主要的和不可见的基因方式，来描述“表型”的分布。在20世纪后期，生物学家识别出这些基因，以确定它们让细胞制造什么样的蛋白质，我们说起这类事就像真的一样，但我们所观察到的还只是概率的分布，我们所说的基因，即DNA链，正是来自于这些分布。

我们这本书说的是总的统计革命，费歇尔在这场革命中起了很重要的作用。他对自己作为遗传学家所取得的成就感到自豪，他的一半以上的成果是与遗传学有关的。现在，我们不再把费歇尔当作一个遗传学家，而主要看他在一般统计技术和观念方面取得的进展。这些观念的萌芽在他的早期作品中就可以发现，但这些观念的全面发展，却是他在工作期间的事，那发生在20世纪20年代到30年代。

《研究工作者的统计方法》

虽然费歇尔在这段时间被数学界忽视了，但他所发表的论文和著作极大地影响了农学和生物学界科学家的工作。在1925年，《研究工作者的统计方法》（Statistical Methods for Research Workers）第一版面世。之后，这本书仅英文版就出了14个，此外，还有法文、德文、意大利文、西班牙文和俄文的译本。

《研究工作者的统计方法》与这之前的数学著作不同，通常数学著作都有许多定理及其证明，并展开抽象的概念将之一般化，与其它抽象概念联系。如果说这类书中有什么应用的话，也只是放在完整的数学描述和证明之后。《研究工作者的统计方法》从如何利用数据制图及如何读图开始，第3页就出现了第一个实例，展示一个婴儿生命头13周每一周的重量，这个婴儿就是费歇尔自己的头生子——乔治（George）。接下来的各章描述如何分析数据：费歇尔给出一些公式，列举一些实例，解读这些例子的结果，然后再转到其它公式。书中没有对公式的数学推导和证明，却带有详细的技术说明，并交待如何在机械计算器上应用它们。

尽管，或者说正是因为缺少理论数学，这本书迅速地被科学界采用。它顺应了现实需求，可以把这本书直接交给只受过有限的数学教育的实验室的技工，让他们自己应用。使用这本书的科学家认为费歇尔的主张是正确的，而评论这本书的数学家则对书中未加证明的大胆论述持怀疑态度，许多人弄不明白他是怎么得出这些结论的。

第二次世界大战期间，瑞典的数学家哈拉尔德?克拉美（Harald Cramér）被战争隔绝于国际科学界外，他花了相当多的时间来费歇尔的这本书和所发表的论文，补充了原来缺失的证明步骤，并推导出原来没有的证明。1945年，克拉美出版了一本书，书名叫作《统计的数学方法》（Mathematical Methods of Statistics），对费歇尔的许多著述给出了正式的证明。不过，克拉美只能对这位多产天才的论述进行选择性的证明，费歇尔的很多著述在克拉美的书中都没有包括进去。克拉美的书被用来教授新一代数学家和统计学家，他把费歇尔著述的“修注”编写成一个标准范式。在20世纪70年代，耶鲁大学（Yale University）的L?J?萨维奇（Savage）阅读了费歇尔最初的论文，发现里面有很多东西都被克拉美遗漏了。他还惊讶地看到，费歇尔对后人的工作早有预见，并且已经解决了在20世纪70年代被认为还没有解决的问题。

但所有这些对1919年的费歇尔来说都是未来的事情，当时他正打算放弃

不成功的学校老师职业。实际上他刚刚完成一项里程碑意义的工作：将高尔顿的相关系数与门德尔遗传学的基因理论结合在一起。但皇家统计学会和K?皮尔逊的《生物统计》都拒绝刊登这篇论文。费歇尔听说爱丁堡皇家学会正在寻找适于他们的《交流》（Transaction）上发表的论文，但期望由作者本人支付印刷成本，就这样，费歇尔自费将自己第二项伟大的成果交给这样一个当时并不起眼的期刊发表。

在当时，K?皮尔逊仍对年轻的费歇尔印象很深，他想聘请费歇尔到高尔顿生物统计实验室担任首席统计师，两个人之间的通讯来往是诚恳的，但对费歇尔来说，K?皮尔逊显然是一个主观意志很强并有支配欲的人，所谓首席统计师，充其量不过是在K?皮尔逊的指令下，从事细节的计算工作。

罗森斯特实验站与农业实验

当时，罗森斯特农业实验站（Rothamsted Agricultural Experimental Station）的所长约翰?罗素（John Russell）爵士也与费歇尔取得了联系。这个实验站是由一个英国的肥料制造商在一个旧农场里建立的。这个旧农场曾属于该肥料公司原来的主人。农场的粘土并不特别适于种植什么作物，但主人发现了如何将石头磨碎与酸混合，生产一种被称作“过磷酸石灰”（Super-Phosphate）的肥料的方法。从过磷酸石灰生产得到的利润用来建立一个实验站，以开发新的人工肥料。90年下来，这个站进行了许多实验，测试无机盐肥料与不同品第的小麦、黑麦、大麦和马铃薯的不同组合。这积累了一大仓库的数据，有雨量和温度准确的日记录、施肥追肥和土壤测量的周记录、收成的年度记录。所有这些都保存在皮面笔记本中。大多数这样的实验没有产生一致的结果，但这些笔记本被小心地存放在实验站的档案室中。

罗素先生看着积累下来这么多资料，想到也许应该雇个人来看看里边有什么东西，对这些资料进行一次统计整理。他四处询问，有的人推荐了费歇尔。罗素跟费歇尔签了一年的合同，给出了1000英磅的酬劳，他只能出这么多了，而且不能保证第二年续聘。

费歇尔接受了罗素的聘任，带着妻子、小姨子和三个孩子来到了伦敦北部的农区。他们租下了实验站旁边的一间农舍，妻子和小姨子打算在那里种种菜园，操持家务，而费歇尔则空上靴子，穿行在农业实验站的田间和90年的数据中，做起他后来称之为“在垃圾堆中寻觅”的工作。

# 第5章 收成变动研究

在我担任生物统计学家不久，一次去康涅狄格大学与休?史密斯教授讨论我所遇到的问题，他给了我一份礼物，那是一篇论文的复印件。论文有53页长，题目是《作物收成变动研究III：降雨量对罗森斯特小麦收成的影响》（Studies in Crop Variation. Ⅲ. The Influence of Rainfall on the Yield of Wheat at Rothamted）。这是一组杰出的数学论文的第三篇，其第一篇1921年发表在《农业科学期刊》第11郑上。产量变化是实验科学家的大忌，但却是统计方法研究的基本素材。在现代科学文献中，“变动”（variation）这个词已经很少被用到了，它已经被其它术语代替，比方说“方差”（variance），这个术语与特定的参数分布有关。“变动”对一般的科学用途来说过于含混，但对费歇尔而言，却是合适的，作物产量在年份之间、地块之间的这种变动，正是作者研究的起点，借此，他可以推导出新的分析。

大多数科学论文在结尾都有参考文献目录，一个长长的单子，以确认对所讨论问题曾经有过建树的论文。费歇尔系列论文的第一篇却只有三篇参考文献：其一，指明了1907年一次不成功的尝试，打算探讨降雨量与小麦生长的相关性；其二，1909年以德文写成的，描述了一种计算复杂数学公式最小值的方法；其三，是由K?皮尔逊发表的一组数表。先前没有什么论文涉足过这一杰出研究系列所涵盖的题目。《作物收成变动研究》是自成一格的，署名的地方写着：罗纳德?A?费歇尔，文学硕士，罗森斯特农业实验站统计实验室，哈盆登（Harpenden）。

1950年，出版商约翰?威利（John Wiley）征求费歇尔的意见，看他是否愿意从所发表的论文中挑选一些最重要的，好单独形成一本文集。后来这本文集的名称叫做《对数理统计的贡献》（Contributions to Mathematical Statistics）。一打开书，就是费歇尔当时的照片，他一头白发，双唇紧闭，领带稍微有点斜，白胡子梳理得不大好，书中标明费歇尔当时在剑桥大学遗传学系工作。《作物收成变动研究Ⅰ》是该文集中的第一篇文章，作者在文章前面加了一个序言，以明确该文的重要性及其在他全部成果中的地位：

早期在罗森斯特的工作中，作者对研究站多年积累下来的大量观察数据，如天气、收成、收成分析等，给予了极大的关注。气象记录在多大程度上能够提供来年收成的预测？对于这类问题，上述数据是有独特价

值的。现在这篇文章是用于此目的的系列研究的首篇。

这个系列研究最多有6篇论文，《作物收成变动研究Ⅱ》发表在1923年，而史密斯先生给我的那篇标号为“Ⅲ”，在1924年问世。《作物收成变动研究Ⅳ》则在1929年发表。标号为“Ⅴ”的论文没有出现在费歇尔的文集中。在科学史上还很少有这种事件：标题那么不起眼，而其内容却如此重要。在这些论文中，费歇尔开发了用于数据分析的原创性工具，建立了这些工具的数学基础，并描述了如何将它们应用到其它领域中去，包括如何应用到他在罗森斯特所遇到的“垃圾堆”上。这些论文表现了令人眩目的原创性，充满了奇妙的内涵，这足够理论家们在20世纪余下来的日子里忙乎的，也许那之后还会继续激发更多的研究。

《作物收成变动研究Ⅰ》

费歇尔系列研究的后两篇文章是有共同作者的，但《作物收成变动研究Ⅰ》却是他独立完成的，那需要大量的计算工作。他的唯一后援是一台名字叫“百万富翁”的计算器，那是一台原始的带有手摇曲柄的机械计算器。如果要算乘法，比方说算3342乘27,先要将转盘放在个位上，设定3342这个数字，摇动曲柄7次；再将转盘放在十位数上，设定3342这个数，摇动曲柄2次，计算方告结束。这架机械叫“百万富翁”，因为它的转盘大得足够容纳以百万计的数字。

为了体会到这篇论文所耗费的气力，我们来考虑一下《作物收成变动研究Ⅰ》中第123页的表7.如果完成一个多位数乘法需要1分钟，我估计费歇尔需要大概185个小时来完成这张表。这篇论文中有15张复杂程度相当的数表，还有4张更为复杂的图。只考虑体力劳动本身，准备这些图表至少需要耗去费歇尔8个月的时间，而且每天得工作12个小时！这还不包括其它工作所花费的时间。比方说：思考理论数学问题、整理数据、设计分析框架、修正不可避免的错误等等。

高尔顿回归思想的一般化

回顾一下高尔顿所发现的“向平均数回归”，他试图找到一个数学公式，将随机事件彼此联系在一起。费歇尔接过高尔顿“回归”（regression）这个词，建立了某个给定地块小麦收成与年份之间的一般数学关系，这个相当复杂分布的参数描述了小麦产量产业化的不同方面。要深入理解费歇尔的数学式，你得有坚实的微积分基础，得对概率分布理论有好的辨别力，还要对多维几何学有感觉，但理解他的结论并不那么难。

他将小麦产量的时间趋势分成几个部分，一个是由于土地退化导致产量稳定地整体性地下降；另一个是长期的缓慢的变化，每个阶段都要花几年时间；第三个是一组更快的移动变化，考虑的是气候在不同年份的差异。自从费歇尔开创性的尝试，时间序列的统计分析在他的思想和方法的基础上，建立了起来，现在我们有了计算机，可以用更巧妙的演算法进行大规模的计算，但基本的思想和方法仍然未变。给定一组随时间波动的数据，我们可以将之分解为不同来源导致的结果。时间序列分析用来检验：美国太平洋海岸拍激的海浪是不是印度洋风暴的起因。这些方法使研究人员能够区分地下核爆破与地震，能够精确地为病理学上的心中节律定位，能够确定环境管制对空气质量的影响，其应用范围还在继续扩大。

农场有一个名称叫“宽田硬”（Broadbalk）的地块，在分析其粮食收成时，费歇尔感到有些困惑，这块地只用了动物粪肥，所以不同年份收成的变动与人工肥料无关。当土壤得自动物粪肥的养分逐渐耗尽，地力退化的长期因素就可以得到解释，同时费歇尔还可以确定不同年份降雨类型不同所带来的影响。那么，什么是缓慢变化的原因呢？从缓慢变化的形态可以看出，在1876年产量开始下降，比从另两个因素所能预计的程度还要大，这种下降在1880年速度更快了；这种情形在1894年开始改善，持续到1901年，而后又是下降。

费歇尔发现了带有同样缓慢变化的另一种记录，不过形态是相反的，那是关于麦田里野草的。1876年后，野草蔓延得越发严重，而到了1894年突然开始消失，只是在1901年又开始茂盛起来。

后来发现，雇用小男孩到地里去拔草，在1876年以前是通告的做法。在英格兰的大地上，下午经常可以看到瘦弱的小男孩穿行于田间，不停地拔草。到了1876年，教育法（the Education Act）使得上学带有强制性，田间小男孩的大部队开始不见了。而1880年第二部教育法通过，对致使孩子辍学的家长施以罚款，田间剩下的男孩也离开了。没有了拔草的小手，那些野草就又茂盛起来了。

那么，在1894年又是什么事情发生，使得趋势逆转了呢？在罗森斯特附近有一所女子寄宿学校，新校长约翰?劳斯（John Lawes）相信，充满活力的户外活动有助于他那些年轻的被托管人的健康。他和实验站的头儿一起安排，让这些年轻姑娘在周六和傍晚出门，到地里去拔草。1901年劳斯去世后，这些小姑娘恢复久坐的习惯，多是在户内活动，野草也就又回到了“宽田埂”。

随机化控制实验

第二篇研究收成变动的论文也是发表在《农业科学期刊》上，时间是1923年。这篇论文并不处理罗森斯特过去实验所积累下来的数据，取而代之的是新实验：一组不同的人工肥料组合对不同品种马铃薯的影响。费歇尔到了罗森斯特后，实验有了明显的改善。不再将某种实验的人工肥料用于整个农场，现在他们把土地划成小的地块，每个地块进一步区分作物的行，地块中的每一行都给予不同的处理。

基本的想法是简单的，之所以简单，那是因为一经费歇尔提出后，它就简单了，但这之前却没有人想到它。任何人观察土地上的作物时，都会很明显地感到有的地块土质好于其它地块。在某些角落，作物长得又高又密，而其它角落，作物则又细又稀。这可能是由于排水方式、土壤类型的改变、未知养分的出现、多年生野草的抵制，或者一些其它未能预见的原因。如果农业科学家要测试两种人工肥料间的区别，他可以将一种施于地块的其它角。但这会将肥料的效应与土壤或者排水等的效应混淆在一起。如果试验在相同的地块不同的年份进行，又会把肥料的效应与气候变化的效应相混淆。

如果同一年里，在相同作物上进行肥料的比较，土壤的差别就会减到最低程度，但他们仍然存在，因为所处理的作物不会有绝对相同的土壤条件。如果我们使用足够多的成对比较，在某种意义上，土壤差异所造成的区别就会被平均掉。假定我们要比较两种肥料，其中一种磷肥的含量是另一种的两倍，我们将地分成小块，每一块有两行作物。我们总是将磷肥多的施于北边这行，南边的那行则施磷肥少的。做到这里，反对的声音就会出来了。如果土壤的肥力梯度（fertility gradient）由北向南，那么北边这行的土质就会比南边那行稍好一点，土壤差异的影响就不会被平均掉。

别急！我们正要做调整，在第一个地块，我们把磷肥多的施在北边，到了第二地块，它将被施在南边，就这样来回调整。我的读者中可能有的已经画出地块的草图，将施磷肥较多的行标上了记号。它会指出，如果肥力梯度从西北向东南，施以额外的磷肥的行将总是比别的行土质好。也会有人指出，如果肥力梯度从东北向西南，结论正好相反。好啦，另一个读者发问了，到底谁对了呢？肥力梯度究竟如何分布？我们的答案只能是：天晓得！肥力梯度这个概念是抽象的，当我们选择从北到南或从东到西时，肥力的真正形态可能以非常复杂的方式上下变动。

我可以想象得出来，当费歇尔提出小地块定型处理将得到更为细心的实验时，罗森斯特的科学家们之间也会有这样的讨论。我也可以想象，当讨论集中到如何确定土地的肥力梯度时，费歇尔笑咪咪地坐在一边，听任他们卷入复杂的争论。他已经考虑过这些问题，并有了简明的答案。了解他的人这样描绘费歇尔：即使是争论触及到他，他仍是静静地坐在那里，吞云吐雾，等等容他给出答案的时机。终于，他拿开嘴上的烟斗，说道：“用随机的方法吧！”

费歇尔的变异数分析

的确简单，科学家以随机的方式设计同一地块里不同行家作物的处理，由于随机处理没有固定模式，任何可能的肥力梯度结构都在平均意义上被抵消掉了。费歇尔猛地起身，兴奋地在黑板上写了起来，一行又一行数学符号，手臂在数学公式间挥来挥去，抵消公式两端相同的因子，最后出现的可能是生物科学中最为重要的工具了，在精心设计的科学实验中，如何分解各种不同处理的效应？费歇尔将这个方法称作“方差分析”（ananlysis of variance）。在《作物收成变动研究Ⅱ》中，方差分析第一次面世。

《研究工作者的统计方法》列出了方差分析某些例子的计算公式，但在这篇论文中，他给出了公式的数学推导，不过推导过程还没有详尽到学院派数学家满意的程度。所展示的代数式是为了这样一种特殊情形：比较三种类型的人工肥料、十种不同品种的马铃薯和四个地块。如果比较两种人工肥料、五种马铃薯，或者六种人工肥料、一种马铃薯，则需要几个小时的艰苦工作，以调整出新的代数式。至于搞出适合所有情形的一般公式，就需要更多的数学工作了，恐怕得出几头汗水吧！当然，费歇尔知道一般公式，对他来说，那是如此的明显，以至于没有必要展示它们。

难怪与费歇尔同时代的人对这个年轻人的成果感到困惑！

《作物收成变动研究Ⅳ》介绍了费歇尔年说的“协方差分析”（analysis of covariance），这是一种因素分解的方法，存在着并非由实验设计而来的条件，它们的效应是可以测量的。当时某医学期刊上发表了一篇论文，描写了针对性别和体重所做调整的治疗效应，用的实际上就是费歇尔在Ⅳ号论文中开创的方法。Ⅳ号论文提出了实验设计的精华，Ⅲ号论文，即史密斯教授推荐给我的那篇，将在本章后边一点儿再讨论。

自由度

1922年，费歇尔终于第一次在《皇家统计学会期刊》上发表了他的论文。那是一篇短文，适度地指出了K?皮尔逊公式中的一个错误，许多年后谈到这篇论文，费歇尔写道：

这个短文，尽管带着稚气，不那么完整，但却是破冰之举。它是带试验性质的，并且零零碎碎的，有的读者会因此而气恼，可他们不要忘了，它不得不在批判者中找到发表的渠道。对这些批判者来说，摆在第一位的就是绝不相信K?皮尔逊的成果需要改正，即使是承认了这一点，他们也觉得这事轮不到别的人。

1924年，费歇尔得以在《皇家统计学会期刊》发表别一篇论文，更长一些，更为一般化。后来在一份经济学期刊上，他对这篇论文及相关的另一篇做了如下的评论：“（这两篇论文）要借助于‘自由度’（degrees of freedom）这个新概念，来调和由不同作者观测到的有差异和表现异常的结果……”

自由度这个新概念是费歇尔的发明，这直接得益于他的几何洞察力和将数学问题置于多维几何空间的能力。所谓“异常的结果”出现在一本不大引人注目的书里，那是一个名叫T?L?凯利（T. L. Kelley）的人在纽约出版的。凯利发现有一些数据用K?皮尔逊的公式似乎不能得出正确的答案。看来只有费歇尔注意到了凯利的这本书，凯利的异常结果只是作为一个跳板，借此费歇尔彻底推翻了K?皮尔逊另一个最引以为自豪的成就。

《作物收成变动研究III》

《作物收成变动研究》第三篇发表在1924年的《伦敦皇家学会哲学学报》（the Philosophical Transactions of the Royal Society of London）上，它是这样开头的：

现在就气候对农作物影响而言我们知之甚少，尽管它对一个大的民族产业如此重要。课题的难解，部分地可以归于问题本身固有的复杂性，还有……缺少在实验或者自然产业条件下所取得的数据……

按下来就是长达53页的精彩论述，其中包含着现代统计方法的基础，任何学术领域，包括经济学、医学、化学、计算机科学、社会学、天文

学、药学，只要是需要建立大量相互关联原因的相关效应，就需要应用这些方法。论文中包含了特别精巧的计算方法（回想一下费歇尔只有那台手动的“百万富翁”用来工作），及如何为统计分析组织数据的良策。我将永远感激史密斯教授，他把这篇文章推荐给我，每次我读起它都会有新的收获。

《费歇尔文集》有五卷本，第1卷以1924年的论文作为结尾，靠近卷尾的地方，有一张费歇尔34岁时的照片，他双手交叉在胸前，胡子修理得挺整齐的，眼镜也没有以前照片中的那么厚，神情安详而自信。在这之前的5年里，它在罗森斯特建立了出众的统计部门，雇用了像弗兰克?耶茨（Frank Yates）那样的合作者。在费歇尔的鼓励下，耶茨将继续对统计分析的理论和实践做出贡献。除了少数例外，K?皮尔逊的学生大多默默无闻，当他们在生物统计实验室工作的时候，只能协助K?皮尔逊而不能超越他；反观费歇尔，他的多数学生响应了所得到的鼓励，独辟蹊径，赢得了辉煌。

1947年，英国广播公司（BBC）广播网邀请费歇尔做一个系列讨论，阐述科学的本质与科学研究，在其中一讲的开头，费歇尔这样说道：

科学生涯从某些方面看是奇异的，科学存在的理由，是要增加对自然知识的认知。有时候，虽然会有这种认知的增加，但是这个过程不是顺利的，并且是令人感到痛苦的。理由是：人们不可避免地会发现以前所得出的观点，至少在一定程度上，明显是过时的或者错误的。我想大多数人可以认识到这一点，如果已经教授了10年左右的东西需要修正，他们会以下面的态度加以接受。但有一些人绝对不能接受，就好像打击了他们的自尊心，甚至是对他们一直把持的私有领地的侵犯。他们必然做得像知更鸟和苍头燕雀寻亲残忍，在春天里我们可以看到，当自己的小巢被冒犯里，它们所表现出的愤怒反应。我并不认为能对此做什么补救。这是科学过程中所固有的特性。但年轻的科学家应该得到提醒和指导，当他们奉献出珍宝去丰富人类的宝库时，必然有人会拒绝他或排挤他。

# 第6章 “百年不遇的洪水”

有什么能比百年不遇的灌水更让人无法预料的呢？洪水奔腾肆虐，泛滥成灾，惨烈至极，确实是百年难得一遇。谁能为这样的突发事件制定防范计划呢？像这样罕见的洪水，我们又怎么能估计其洪峰会高达多少呢？如果说现代科学有统计模型能用来处理观测数据的分布，那么，对这种未曾发生过，或者即便发生，也是百年才发生一次的大洪灾，又该如何用统计模型来分析呢？伦纳德?亨利?凯莱布?蒂皮特（Leonard Henry Caleb Tippett）找到了答案。

L?H?C?蒂皮特1902年出生在伦敦，并在伦敦的帝国学院（Imperial College）读物理学，1923年他从帝国学院毕业。蒂皮特曾说过，他之所以被物理学所吸引，是因为物理学对“精确测量的坚持，……和当时科学辩论的那种学院式方法。回顾自己年轻时的激情，他继续说：“我们通常是把一个假设视为对或错，并把至关重要的实验当作加深认识的主要手段。”当他有机会做实验时，他发现实验的结果与理论预测的结果从未有过精确的一致。依据他自己的亲身体验，他说：“我发现最好是去改进抽样技术（这里他指的是统计分布），而不是丢弃理论。”蒂皮特认识到，他如此钟爱的理论所提供的信息仅仅是有关参数的，而不是具体的观测值。

这样，L?H?C?蒂皮特（当他因发表的文章而著称的时候）通过他自身对实验的理解，开始融入统计变革中来。从帝国学院毕业后，他在英国棉花工业研究协会任统计师。人们通常称这个研究协会为雪莉研究会（Shirley　Institute）。该研究会的研究目标主要是利用现代科学方法改进棉线与棉布的生产工艺，其中，他们所遇到的最棘手的问题之一是新纺棉线的强度。因为，即使是在相同条件下纺出来的棉线。其强度也存在很大的差异。蒂皮特非常仔细地做了一些试验，在显微镜下观察那些经过不同拉力抻拉后的棉线，结果他发现，棉线的断裂取决于棉线中最脆弱的纤维的强度。

居然是那些最脆弱的纤维！那么，怎样建立一个描述最脆弱的纤维强度的数学模型呢？由于无法解决这个难题，蒂皮特提出申请，并于1924年获准，到伦敦的大学学院高尔顿生物统计实验室（the Galton Biometrical Laboratory），在K?皮尔逊手下进修一年。关于这段经历，蒂皮特这样写道：

在大学学院度过的那段时光让我刻骨铭心。K?皮尔逊是位非常了不起的人物，并且我们也能深切地感受到他有多了不起。他工作勤奋、充满热情，而且关于激励他的下属和学生。我在那里进修的时候，K?皮尔逊依旧在做研究，并且经济热情洋溢、充满激情地出现在课堂上，讲解他刚刚研究出来的最新成果。那些年，虽然他的研究方式有点过时了，但他讲的课仍旧激动人心。……有一门他讲授的课程“17和18世纪的统计学史”，就是他研究兴趣广泛的一个典型代表。……他还是个精力充沛的辩手，……他出版了一套丛书，就叫做《一个好问者与他的问题》（Questions of the Day and of the Fray）……昔日充满活力与辩论的影响随处可见。系里的墙上装饰着格言与漫画，……有一幅关于“油嘴山姆”（Soapy Sam）的讽刺漫画，画的是那位大名鼎鼎的威尔伯福斯大主教（Bishop Wiberforce），漫画作者名为“间谍”。1860年在英国科学促进协会的会议上，这位大主教曾就达尔文的进化论与T?H?赫胥黎（T. H. Huxley）进行过一场短兵相接的舌战。此外，还陈列了一些在过去数十年内发表过的出版物，看这些出版物的题目会给人留下一个深刻的印象，那就是该系的研究兴趣十分广泛。如“人类遗传宝典（人的身体、精神与病理牲的谱系）”以及“达尔文进化论、医学发展与优生学”。在一次全系 的年度聚餐会上，K?皮尔逊用一种曾为高尔顿提供年度工作报告的方式来总结这一年的工作，就好像高尔顿依然健在，这让我们大家想起他与高尔顿之间非常密切的合作。于是我们共同举杯，“为已故去的生物统计学前辈干杯。”

这是K?皮尔逊一生中还活跃的最后几年，此后，他的科学成就大部分都被费歇尔和自己的儿子扫进了垃圾桶，成了被遗忘的思想。

尽管在K?皮尔逊在实验室里有那么多激励，尽管蒂皮特在进修期间学到很多数学知识，然而有关最不牢固的纤维强度的分布问题依然没有解决。回到雪莉研究所之后，蒂皮特发现了学期在最伟大的数学发现背后的一个简单的合乎逻辑的原理，他找到了一个看似简单的方程式，它能把样本数据的分布与极值（extreme values）的分布连在一起。

能写出方程式是一码事，解出这个方程则是另外一码事。为此，他去请教K?皮尔逊，但没有获得丝毫的帮助。在过去的75年里，工程学专业已经积累了大量的方程及其解，这些都能在那些大部头的概览中查到。然而，在这些概览中蒂皮特却找不到他的方程式。

于是，他采用了一个做法，就像一个可怜的高中生做代数题一样，先猜了一个答案，并把答案代进方程式，居然解出了这个方程。但是，对这

个方程式而言，这是唯一解吗？对他的问题而言，这恰好是“正确”答案吗？为此，他请教了费歇尔，费歇尔不仅能导出蒂皮特所猜的解，而且还给出了另外两个解，并指出，这些就是仅有的解。这就是所谓的“蒂皮特的三条极值渐近线”（Tippett’s three asymptotes of the extreme）。

极值分布

知道极值分布有什么用处？如果我们知道极值分布与正常值的分布之间的关系，就可以记录每年洪峰的高度，并预测百年不遇的洪灾发生时最有可能的洪峰高度。能够这样做的原因是，每年的灌水测量值给我们提供了足够的信息，用它就可以蒂皮特分布的参数。因此，美军工兵署（USACE）就能计算出在河上究竟该筑起多高的堤防，环保署就能规定气体排放标准来控制工业烟囱废气突然排放的极值，棉纺工业就能确定在棉线生产中究竟有哪些因素会对最脆弱的纤维强度的分布参数产生影响。

1958年，当时在哥伦比亚大学（Columbia　University）任工程学教授的埃米尔?J?冈贝尔（Emil　J.　Gumbel），出版了那本关于极值的权威教材，书名是《极值统计学》（Statistics of Extremes）。自那时起，由于他的思想已经扩展到许多相关的地方去，极值理论方面的建树就很少了。然而，冈贝尔的这本教材里包含了一个统计学家在处理这类问题时必备的一切知识，书中不仅包括蒂皮特的原创研究成果，而且还包括后来对该理论的精心的改进，其中有很多都是冈贝尔自己的研究成果。

政治谋杀

冈贝尔的一生富有传奇性。在20世纪20年代末至30年代初，他是德国一年大学里资历尚浅的一名教师。从他早期发表的论文中看得出来，他是个极具潜能的人，只是当时还没有机会得到一个令人尊敬的地位罢了。同样，他当时的职位也远算不上稳固，是否有能力养家糊口，还取决于政府那些权威的随心所欲。当时，纳粹在德国境内已经渐趋猖獗，国家社会主义工人党　虽然是正式的正常组织，实质上却是由一群歹徒纠集而成的。俗称“褐衫队”（Brown Shirts）的纳粹冲锋队是一个专门从事恐吓与胁迫、恣意暴力和谋杀来执行纳粹党意志的暴徒组织。任何批评纳粹党的人都会遭到暴力攻击，而且通常就发生在城市的大街上，以杀一儆百。冈贝尔有个朋友就是这样在光天化日之下曹到攻击并被公然杀害的。照理说，会有许多目击证人可以指认凶手，但法院往往宣称罪证不足而使纳粹突击队逍遥法外。

冈贝尔曾参加过一场审判，他亲眼目睹了法官全然无视任何证据，恣意裁决，纳粹党徒则在法庭上肆无忌惮地狂呼。对此，冈贝尔惊骇万分。于是，他开始着手调查那些凶手公然行凶的其他案例，结果没有一例被判有罪。最终他得出结论：司法部门已经被纳粹党人所控制，很多法官要么是纳粹的支持者，要么干脆就是纳粹所雇佣的。

冈贝尔搜集了许多案例，走访证人，证明判决那些凶手无罪是错误的。1922年，他出版了《四年的政治谋杀》（Four Years of Political Murder）一书，把他搜集调查的结果公之于众。由于发现很多书商根本不敢销售他的书，他不得不亲自去为自己的书安排发行分销。与此同时，他还在继续搜集案例，并于1928年又出版了《政治谋杀的原因》（Causes of Political Murder）一书。此外，他还设法成立一个反纳粹的政治团体，但是他的多数学术界同事太害怕了，甚至那些犹太籍的朋友们都吓得不敢参加。

1933年纳粹党取得了政权，当时冈贝尔正在瑞士参加一个数学会议。他本打算立即赶回德国去与这个新政权做斗争，但朋友们极力劝阻了他，因为只要他一越过边境，就会立刻遭到逮捕，并被处决。在纳粹掌权的最初阶段，在这个新政府还没来得及控制所有的出入境事务之时，少数犹太籍教授，如德国的顶尖的概率论大师里夏德?冯?米泽斯（Richard Vin Mises），他们已经预料到即将发生的灭顶之灾，提前逃离了德国。冈贝尔的朋友也趁这段有利的混乱时机，带着他的家人离开了德国。他们跑到法国暂避一时，但是，1940年纳粹又入侵了法国。

冈贝尔与家人继续逃往尚未沦陷的法国南部。当时统计法国的是纳粹扶植的傀儡政府，对德国惟命是从。像冈贝尔这样的德国民主党人已经是危在旦夕，因为他们都被列入了叛国者的黑名单，纳粹要求法国政府将这些人移交过去。除了冈贝尔，滞留在法国马赛的德国逃亡者还有德国作家托马斯?曼（Thomas Mann）的哥哥海因里希?曼（Heinrich Mann）、犹太裔小说家利翁?福伊希特万格（Lion Feuchtwanger）。当时驻马塞的美国领事海勒姆?宾厄姆四世（Hiram Bingham IV）违反美国国务院的规定，擅自给这批德国流亡者发了签证。宾厄姆为此受到华盛顿的谴责，最终由于此举而丢掉了他在马赛的职位，但宾厄姆毕竟尽他所能拯救了很多人，这些人如果留在纳粹统计下，将必死无疑。冈贝尔与家人到了美国之后 ，在哥伦比亚大学谋到一个职位。

数学著述有很多种不同的写法。有此所谓“权威”教科书，内容贫乏、苍白、毫无生气，提出一系列的定理及证明，却几乎引不起读者的任何兴

致；有此书通篇是从假设到结论的证明，玄虚而艰涩；而有此权威的教科书，则由始至终充满了精彩的证明，其中的数学推导过程被浓缩成看上去很简单的步骤，按照这些步骤可以毫不费力地得出最终结论；还有极少量的权威性的教科书，作者试图在书中把问题的背景和思想都交代清楚，不仅记述了学科的历史渊源，而且所举的例子也取自生动的现实生活。

最后一类所说的权威性教书的这些牲恰是对冈贝尔的《极值统计学》一书的真实描述。这本书提供了大量有关该学科发展的参考，是对一个高难学科的最为明晰的解释。该书的第1章“目录与手段”介绍了该书的主题以及在其他章节中必须理解的数学的发展。这一章本身就是对统计分布理论的数学知识的最卓越的介绍。它的设计思想是让那些只读过大学一年级微积分的学生能看得懂。我第一次读这本书的时候，尽管已经拿到了数理统计博士学位，还是从第一章中获准颇多。作者在前言中谦虚地说：“我期望，而决不是预料，本书的写作能使人类从中获益，哪怕是因为对科学进步的微不足道的贡献。”

这本书的贡献决不能称之为“微不足道”，它是由20世纪一位大师级的教师矗立的一座丰碑。集非凡的胆识与杰出的表达能力于一身，把最难理解的思想以条理清晰、简洁精炼的方式表达出来，埃米尔?J?冈贝尔正是这些极为罕见的杰出人才当中的一位。

# 第7章 费歇尔获胜

英国皇家统计学会（The Royal Statistical Society）拥有三种可以发表论文的学术期刊，每年学会还主办学术会议，会上邀请演讲者介绍他们最新的研究工作。论文要在这些期刊上发表是相当困难的，必须经过至少两位评阅人的审查，看内容是否正确，而且编辑与主编都必须认为该篇论文代表了当时在自然科学领域的显著进展。但是，与应邀在大会上演讲相比，在学会期刊上发表论文就显得容易多了。大会演讲，这只是留给那些在统计学领域里最杰出的研究人员的一种荣誉。

每一次应邀演讲结束之后，按照学会的惯例，都会组织一场与会者参加的讨论会。由于特邀的会议来宾已经预先拿到了将在大会上演讲的论文副本，因此他们的讨论常常不但详尽，而且一针见血。之后，这篇论文连同讨论会上对论文的评论意见都会发表在《皇家统计学会期刊》上。

这种讨论会，正如在期刊上所展现的，有一种非常程序化的英国风格。大会主席（或某个被指定的人）首先站起来提议向演讲人表示感谢，紧接着陈述他的评论。随后，一位事先指定的皇家统计学会的资深会员直立再次提议表示感谢，并随之发表他的评论。接下来，学会中一些最负声望的会员一个接一个地相继站起来发表他们的评论。除了学会的会员之外，大会还经常邀请一些来自美国、英联邦和其他国家的来宾，也请他们发表评论。演讲人再对所有的评论做出回应。最终，学会期刊允许评论人及主讲者对属于他们自己的那部分文字进行编辑之后才正式发表。

1934年12月18日，在学会会议上宣读这样一篇论文的无上荣誉赋予了理学博士、英国皇家学会会员费歇尔教授。经过了20世纪20年代事实上的孤立之后，费歇尔的天赋终于得到了公认。我们在前几章里读到他的时候，费歇尔的最高学位还只是个理学硕士（M.S.），他的“大学”也不过是伦敦郊外一个偏僻的农业试验站。到1934年，他又获得了一个理学博士学位，并且当选为威望很高的英国皇家学会的会员（缩写为F.R.S.）。直至此时，皇家统计学会才终于承认了他作为这个领域中的领军人物，应该占有一席之地。因为这项荣誉，费歇尔在大会上宣读了一篇论文，题为《归纳推理的逻辑》（The Logic of Inductive Inference）。大会主席是皇家统计学会当时的会长。皇家学会会员M?格林伍德（M. Greenwood）教授。费歇尔的论文印出来共计16页，另外还

呈上一份结构严谨、条理清晰的论文摘要，概括了他最新的研究工作。第一位发言的评论人是A?L?鲍利（A. L. Bowley）教授，他站起身来提议表达谢意，接着发表了他的感言：

我很高兴有这样一个机会向费歇尔教授表示感谢。不仅是因为他刚才为我们宣读的论文，更重要的是因为他对统计学的全面贡献。今天借此良机，我谨代表所有我熟悉的统计学家，对他带给统计学研究的无与伦比的热忱，对他提出的数学工具的威力，对他在这里、在美洲和在世界各地的广泛的影响力，以及对他深信做为数学的正确应用所发挥的激励作用表示钦佩之意。

K?皮尔逊当时不在讨论者之列。此前3年，他已从他任职的伦敦大学退休。在他的领导下，高尔顿生物统计实验室已经成长为大学里一个正式的生物统计学系。他退休后，该系一分为二，费歇尔受命担任其中之一的优生学系的系主任，另一个则是规模缩小了的生物统计学系，系主任由K?皮尔逊的儿子E?皮尔逊担任，同时他还负责高尔顿实验室的工作，并兼任《生物统计》杂志的编辑。

费歇尔与小皮尔逊的私交不大好，这完全是费歇尔的过错。他对E?皮尔逊的态度带着显而易见的敌意。小皮尔逊这位温文尔雅的先生，一则是代父受过，因为费歇尔不喜欢他的父亲老皮尔逊；二则是代合作伙伴耶日?奈曼受过，费歇尔特别讨厌奈曼（奈曼与E?皮尔逊的合作将在第10章介绍）。尽管如此，小皮尔逊倒是极其尊重并高度评价费歇尔的工作。多年后他曾写道，他早就习惯了费歇尔从不在著述中提到他的名字。但是，尽管两人之间关系紧张，尽管两系之间存在着争夺权限的纠纷，费歇尔和E?皮尔逊都清寒是派学生去听对方的课，竭力避免公开的冲突。

至于K?皮尔逊，此时的他已被学生们称之为“老家伙”了。他拥有一个研究生助手，并保留着一间办公室，但他的办公室无论离两个系的办公地点还是离生物统计实验室，都有一段距离。从美国来的邱吉尔?艾森哈特跟随费歇尔和E?皮尔逊进修一年，这期间他曾想去拜访K?皮尔逊，但他的同学和系里的同事都极力劝阻他。问他，为什么不去请教才华横溢的费歇尔，竟然想去看K?皮尔逊？去看那个老家伙能有什么新的收获？令艾森哈特万分遗憾的是，他在英国期间未曾去拜访K?皮尔逊，而就在那一年老皮尔逊去世了。

费歇尔学派与皮尔逊学派：两种统计观

哲学上的分歧使费歇尔与K?皮尔逊在研究统计分布的方法上分道扬镳。K?皮尔逊把统计分布视为对他所分析数据的集合的真实描述。而按照费歇尔的观点，真实分布只是一个抽象的数学公式，搜集的数据只能用来估计这个真实分布的参数。既然所有的估计都有误差，那么费歇尔提出来的一些分析的手段，可以把这种误差的程度降到最低，或者可以更经常地得出比其他任何手段都更接近真实分布的答案。

在20世纪30年代，看上去是费歇尔在这场辩论中获胜了，但到了70年代，皮尔逊学派的观点东山再起。直到写这本书时，统计学界在这个问题上已经分裂成两派，尽管K?皮尔逊本人几乎不接受他的天才继承者们的观点。费歇尔用他条理清晰的数学头脑廓清了残存在K?皮尔逊观点中大量的混淆，正是这些混淆使得K?皮尔逊没有意识到自己观点的深层本质，因此，后来东山再起的皮尔逊方法已经无法回避费歇尔的理论成果。当把统计模型应用于现实时，存在着一些很严重的问题。因此，本书打算在多处探讨这些哲学问题，这里就是其中的一处。

K?皮尔逊把测量值的分布视为一个真实的存在。在他的方法里，对于一个给定的情况，有一个庞大的然而却是有限的（finite）测量值的集合。在理想情况下，科学家会搜集所有的这些测量值，并确定其分布参数。如果无法搜集到全部测量值，那么就搜集一个很大的并且具有代表性的数据子集（subset）。由这些大量的、且具代表性的子集计算出来的参数会与完备集合的参数相同；此外，那些用来计算完备集合参数值的数学方法也适用于有代表性的子集的参数估计，而不会有严重的误差。

但依照费歇尔的观点，测量值是从所有可能出现的测量值中随机选取的，依据随机选取的数据计算得出的一个参数的任何估计值，其结果本身也具有随机性，因此，也会服从一种概率分布。为了能清楚地区分参数的估计值与参数本身这两个不同的概念，费歇尔把这个估计值称为“统计量”（statistic）；不过现代术语往往称其为“估计量”（estimator）。假设我们有两种不同的方法可以得到一个统计量，以估计某个特定的参数。例如老师想了解一个学生对知识掌握到什么程度（参数），就在全班进行了几次测验（测量），并且计算出测验的平均分数（统计量）。那么，究竟是用中位数（median）作统计量“更好”呢，或是取这几次测验中的最高分与最低分的平均值“更好”呢，还是去年最高分与最低分然后把其余的测验成绩加以平均“更好”？

既然统计量是随机的，那么讨论这个统计量的某个值的准确性到底有多大是毫无意义的。我们需要的是一个判别的准则，这个准则以统计量的

概率分布为依据，就像K?皮尔逊所指出的那样，对一组测量进行估计，必须根据它们的概率分布，而不是根据个别观测值。评判哪一个是好的统计量，费歇尔提出了如下三个准则：

一致性（consistency）：得到的数据越多，计算出来的统计量接近参数真值的概率就越大；

无偏性(unbiasedness):如果用很多组不同数据集多次测量某一特定的统计量，那么该统计量的这些测量值的平均数应该近似于这个参数的真值；

有效性（efficiency）：统计量的值不会完全等于该参数的真值，但是用来估计一个参数的大多数统计量应该与真值相去不远。这些阐述似乎有点含混不清，这是因为我在竭尽全力地把一些本来精确的数学公式，用一些一般性的文字表述出来。实际上，费歇尔的这些准则都可以用恰当的数学式来表达。

费歇尔之后的统计学家又提出了其他的准则，费歇尔自己也在后来的论文中提出了一些次要准则。剔除所有这些准则中的混乱不清的东西之后，剩下的最重要的元素就是，应该把统计量本身视为随机的，而好的统计量一定有好的概率特性。对于某一特定数据集，我们永远不知道一个统计量的值是否正确，只能说我们用一种方法得出来一个符合这些准则的统计量。

在费歇尔提出的三项基本准则中，“无偏性”准则最引人关注，这或许是由于“偏误”（bias）这个词带有某种贬义。一个有偏的（biased）统计量似乎是谁都不想要的某个东西。美国食品和药物管理局的正式指导准则就提出警告，要大家使用“避免有偏”的方法。有一种非常奇怪的分析方法（将在第27章里详细讨论），叫做“意向治疗”（intent to treat），已经成为占优势的医学试验法，因为，这种方法仍能保证结果是无偏的，尽管它忽略了有效性的准则。

事实上，一些有偏的统计量的应用常常极为有效。据费歇尔的研究，用来确定净化城市供水系统中氯浓度的标准方法，依据的就是一个有偏（但满足一致性与有效性）的统计量。所有这一切也是科学社会学(the sociology of science）中的一类研究课题——为准确定义一个概念而创造出来的一个词，怎样将情感好恶的包袱也带到了科学中来，并对人们的行为产生了影响。

费歇尔的极大似然法

当费歇尔研究了这些数学问题之后，他认识到，用K?皮尔逊的方法来计算分布参数所生成的统计量未必是一致的，而且经常是有偏的，他也认识到还存在着更加有效的统计量可以利用。为了得到一致且有效（但未必无偏）的统计量，费歇尔提出了被他称之为“极大似然估计量”（maximum likelihood estimator, MLE）的一个概念。

随后，费歇尔证明了MLE总是一致的，而且证明了如果人们认可几个被认为是“正则性条件”（regularity conditions）的假定，那么MLE是所有统计量中最有效的。此外，费歇尔还证明了，即便MLE是有偏的，也可以计算出其偏差的大小，然后将其从MLE的估计值中减掉，从而得到一个一致、有效且无偏的修正统计量 。

费歇尔的似然函数（likelihood function）席卷了整个数理统计学界，迅速成为估计参数的主要方法。极大似然估计只存在一个问题，就是在试图求解MLE时所涉及的数学问题，其难以对付的程度确实令人望而生畏。费歇尔的论文里写满了一行又一行的复杂代数式，用来说明不同分布的MLE数学公式的推导过程。他的方差分析和协方差分析的运算法则显示出他极高的数学造诣，去处过程中他设法在多维空间里利用巧妙的代入与变换，导出最终为使用者所需要的MLE的计算公式。

尽管费歇尔具有非凡的独创性，但在多数情况下，对于MLE的潜在使用者来说，仍然难以驾驭所必需的高深数学知识。20世纪后半叶的统计学文献中有许多非常睿智的文章，它们运用简化的数学方法，在某些实例中得到了相当理想的MLE的近似值。在我自己的博士学位论文里（大约写于1966年），我只能将就着不得不接受这样一个事实，即只有在能够得到非常多的数据时，我的问题的解才是好的。假定我有大量的数据，就能把似然函数简化到可以计算出挖MLE值的程度。

后来出现了电脑。电脑并非人脑的竞争对手，电脑只是一个巨大而有耐力的数字处理设备。它从不会厌烦，从不会困倦，也不会犯错误。它一而再、再而三地重复着做那些同样繁琐的计算，数百万次地一再重复。用所谓的“迭代算法”（iterative algorithms），它能算出MLE值。

迭代算法

最早的一种迭代数学方法好像出现在文艺复兴时期（虽然数学史学家大

卫?史密斯（David Smith）在他1923年出版的《数学史》（History of Mathematics）中声称，早在古埃及和中国的文字记载中就已经发现了这种方法的实例）。当资本主义曙光初露之时，在意大利北部刚刚建立起来的商业银行或商号中就碰到一个基本问题：每个小小的城邦或国家都有自己的倾向，所以商号必须能算出如何在各倾向之间兑换；比如说，如果汇率是雅典钱币14德拉克马（Athenian drachma）换一个威尼斯币达克特（Venetian ducat），那么用威尼斯的127达克特买来的一堆木材，价值多少雅典的德拉克马呢？如今，我们有能力用代数符号来解答这个问题。还记得高中的代数吗？若X等于雅典德拉克马的值，则……

尽管当时的数学家已经开始发展代数学，这种简单的计算方法仍不能为大多数人所用。银行家用的是一种叫做“试位法”（rule of false position）的计算方法。由于每家商号都确信自己的换算规则是“最好的”，所以每家商号都有自己的店员。罗伯特?雷科德（Robert Recorde，1510-1558），这位16世纪的英国数学家，在普及代数符号上功绩卓著。为了把代数的威力与试位法则相对照，他在1542年写了一本书“The Grovnd of Artes”，书中说明了试位法：

Gesse at this woorke as happe doth leade.

By chaunce to truthe you man procede.

And firste woorde by the question,

Although no truthe therein be don.

Suche falsehode is so good a grounde.

That truthe by it will soone be founde.

From many bate to many more,

From to fewe take to fewe also.

With to much ioyne to fewe againe,

To to fewe adde to manye plaine.

In crosswaied multiplye contrary kinde,

All truthe by falsehode for to fynde.

雷科德的这篇16世纪的英文说的是：你先猜一个答案，并把它代入问题中，由此你会得到一个结果，而它和你想要的结果之间会有些差异。有了这个差异，接着你可以用它再产生一个更好的猜测，再用这个新的猜测得到一个新的差异，这个差异又会产生出另一个新的猜测值。如果在计算这个差异的过程中，你做得足够聪明，这一连串的猜测值会最终接近正确的答案。对试位法来说，只要迭代计算一次，第二次猜测通常总能得到正确答案；而费歇尔的极大似然估计法，可能要迭代数千次甚至数百万次才能得到一个理想的答案。

然而，对一台任劳任怨的电脑，区区几百万次的迭代又算得了什么呢？在当今世界，不过是一眨眼的工夫。但在不久前，电脑的功能还不够强大，速度也很慢。在60年代末，我有个可以编写程序的台式计算机，是一种可以做加、减、乘、除的原始的电子工具。不过它还有个容易很小的内存，可以放进去一个程序，让它完成一系列的自述去处。这些运行的功能之一还能改写程序，因此，可以在这台可编程的计算机上运行迭代计算，只是要花很长的时间罢了。一天下午，我编好了计算机程序，检查了前几个步骤，确信我写的程序准确无误，然后，关掉办公室的灯就回家了。与此同时，这个编好了程序的计算机就开始了加减乘除的去处，静静地从它的电子结构内部发出喃喃的低语语，而且每隔一会儿就会按程序打印出一个计算结果。连接在计算机上的打印机是一个噪音很大的压缩设备，打印的时候会发出很响的“卡嗒、卡嗒”的声音。

那天晚上，保洁员到办公楼里清扫，其中一个人带着扫帚与废纸篓走进我的办公室。黑暗中，他听到了一种“嗡嗡嗡”的声音，他能看见在一遍又一遍进行加减的计算机上有只眼睛发出忽明忽暗的蓝光。突然，机器醒了过来，“卡”地响了一声，接着又“卡、卡、卡……卡嗒、卡嗒、卡嗒、”地响起来。后来他告诉我，那可真是一次让他毛骨悚然的经历。因此他要求我，如果下次计算机正在运行时，让我一定在办公室门口留一个提示纸条通知他们。

今天的电脑运行快得多了，甚至可以分析更加复杂的似然函数。哈佛大学的纳恩?莱尔德（Nan Laird）和詹姆斯?韦尔（James Ware）教授发明了一种异常灵活、功能异常强大、叫做“EM演算法”的迭代过程演算法。在我订阅的统计学期刊里，每一期新杂志都会介绍某人如何采用他或她的EM演算法解决了一度被认为无法解决的难题。另有一些算法，名字颇富想象力，像“模拟退火法”（simulated annealing）、“克利金

法”（kriging）等等，也不时地出现在文献中；还有“大都会”（Metropolis）算法或“侯爵”（Marquardt）算法，以及其他一些以发明者自己命名的算法。有一些很复杂的软件包，用成千上万行的程序编码，使这些迭代运算以“用户界面友好”的特点变得易于操作。

费歇尔的统计估计方法大获全胜，极大似然法统计了世界，而K?皮尔逊的方法则被尘封在被遗忘的历史角落里。然而，就在这个时候，20世纪30年代，当时费歇尔对数理统计理论的贡献终于得到了承认，他40多岁并且正值其事业鼎盛时期，就在那一刻，出现了一位名叫奈曼的年轻的波兰数学家，他对费歇尔一味遮掩却并没有真正解决的某些问题提出了质疑。

# 第**8**章 致死的剂量

每年的3月，生物统计学会都要在美国的南部城市召开一次春季会议，我们这些在北部生活和工作的人就借此机会南下，到路易斯维尔（Louisville）、孟斐斯（Memmphis）、亚特兰大（Atlanta）或新奥尔良（New　Orleans），在会议结束后回家前的几周，去呼吸春天的清新空气，观赏原野中盛开的鲜花和果园里花繁叶茂的果树。同其他的科学会议一样，会议期间会有三到五位论文作者在会上口头宣读他们的论文，然后与会者与演讲人就论文的内容展开热烈的讨论，询问某些思想的出版，或提出其他可以替代的方法。通常，上午的会议分成两个分会场同时进行。最后的会议一般在下午5点前后结束，与会者回到宾馆各自的房间。一个小时或一个半小时之后他们又会分头聚在一起，相约着出去找一家喜欢的餐馆共进晚餐。

开会的当天，一般人总能在会场上遇到一些朋友，并绝好了会后一同去吃晚饭。但是有一天我却错过了约人就餐的时机。我和那天下午的一位论文演讲者进行了一场长时间的且饶有兴趣的讨论，他是当地人，散会后可以直接回家，因此我没有邀他一起吃饭。我们的谈话结束的时候，大厅里已经空荡荡的，人都走光了。我联系不上任何人，就回到房间给太太打电话，与孩子们在电话上聊了几句，随后就下楼到宾馆的前大厅，心想说不定会碰上一伙我认识的人，可以和他们一道活动。

但是，大厅里几乎空无一人，只有一个身材高大的白头发男人，他独自坐在一张罩着椅套的椅子上。我认出他是切斯特?布利斯（Chester Bliss），我知道他发明了一些基本的统计模型。那天上午在我参加的那个分会场，他还宣读了一篇论文。我朝他走过去，做了自我介绍，并称赞他上午的发言。他邀请我坐下，我们就坐在那里聊了一阵子统计与数学。不错，我们的确是在聊着这样的话题，我们甚至可以用这个话题来开玩笑。显而易见，我们俩谁也没有晚餐的约会，于是我们决定一起去吃晚饭。他可真是个令人愉悦的就餐伙伴。那天的晚餐，我听他讲述了自己丰富的阅历。以后的几年，我们常在开会的时候碰面，有时还会相约一同用餐。他在耶鲁大学的统计系任教，所以，每当我参加由耶鲁大学统计系主办的研讨会时，就经常能见到他。

布利斯出身于美国中西部一个殷实而融洽的中产阶级家庭，父亲是医生，母亲掌管家务，有几个兄弟姐妹。他起初对生物学感兴趣，念大学

时学的是昆虫学。20世纪20年代末，他大学毕业后，以一个昆虫学家的身份供职于美国农业部，并且不久就参与　　了研制杀虫剂的工作。很快，他认识到，在田间试验杀虫剂会受到许多无法控制变量的干扰，使结果难以解释，于是，他把昆虫带到实验室里，做了一系列的实验。这时，有人把费歇尔所写的《研究工作者的统计方法》一书介绍给他，以此为起点，他一边努力去领悟费歇尔在这本书中介绍的许多统计方法的深层次内涵，一边又阅读了费歇尔更多数学论文。

概率单位分析

在费歇尔统计方法的引导下，不久，布利斯说开始了他在实验室内的实验。他把昆虫分成几组，养在广口玻璃瓶里，然后用不同成分和不同剂量的杀虫剂来实验。在他做这些实验的过程中，发现了一个值得关注的现象：无论他配制的杀虫剂尝试有多高，在用药之后总会有一两只昆虫还活着；此外，无论他怎么稀释杀虫剂，即便只是用了装过杀虫剂的容器，试验结果也总会有几只昆虫死掉。

有了这些显著的变异，如果能依据皮尔逊的统计分布建立一个数学模型来分析杀虫剂的作用，这将是非常有用的。但是如何建立这个模型呢？你很可能会回想起高中代数课上，当书本翻到解文字题时那令人头疼的时刻：A先生和B先生共同在静止的水中划船；或者在平稳流动的水中逆流而上；或者他们会把油与水混在一起；或者让他们来来回回地运球。无论哪一种问题，这种文字应用题总是给出一些数字，然后问一个问题，可怜的学生就必须把这些文字转换为数学公式，并解出未知数x。你或许能回想起当初是如何哗哗地翻查着教科书，拼命地寻找一个类似的并且已经解出答案的例题，然后把文字应用题的新数字塞进这道例题所用的公式中去。对高中的代数课而言，总有人已经把相关问题的数学公式列了出来，要么老师知道这些数学公式，要么能在与教科书配套的教师手册里找到这些公式。然而，试想有这样一个文字应用题，没有人知道如何将它转化为数学公式，没有人知道问题当中哪些数据是多余的，哪些应该是没用的，而一些至关重要的信息又常常缺失，况且教科书中也没有事先已经解出来的类似例题。这就是当你设法把统计模型应用到现实生活中去的时候所面临的情景，这也正是当布利斯打算采用概率分布这种新的数学思想来分析他的杀虫剂实验时所遭遇的困境。

为此，布利斯发明了一种他称之为“概率单位分析”（probit analysis）的方法，这项发明需要一种非凡跨越的原创性思想。这种方法中的任何思想，甚至哪怕是应该如何去做的启示，都未曾出现在费歇尔的“学

生”的、亦或其他什么人的著作中。他之所以使用“概率单位”（probit）这个词，是因为他的模型建立了“杀虫剂的剂量”与“使用该剂量时一只虫子会死掉的概率”这两者间的关系。他的模型中生成的最重要的参数谓之“半数致死剂量”（50 percent lethal does），通常用“LD-50”来表示，是指杀虫剂能以50%的概率杀死虫子的剂量。或者说，如果施用这种杀虫剂来对付大量的虫子，那么用“LD-50”的剂量，将有50%的虫子被杀死。布利斯模型的另一个推论则是：对一只特定的用做实验标本的虫子，要确定杀死它所需要的剂量是不可能的。

布利斯的概率单位分析已被成功地应用于毒物学（toxicology）。从某种意义上说，源于概率单位分析的认识已经形成了毒物学这门科学的主要基础。16世纪的医师P?A?帕拉赛瑟斯（P. A. Paracelsus,1493-1541）有一句名言：使用过量，什么都是毒药。概率单位分析为帕拉赛瑟斯首创的这个信条奠定了数学基础。按照帕拉赛瑟斯的这个信条，只要剂量足够大，任何东西都可能成为毒药；而只要剂量足够小，任何东西都是无害的。而布利斯则为了这个信条增加了与那些个案结果联系在一起的不确定性。

之所以会有那么多愚蠢的吸毒者，在古柯硷、海洛因或安非他命的作用下，或已毙命于街头，或变得极度虚弱，原因之一就在于，他们看到其他人同样服用这些毒品却没有死于中毒。他们就如同布利斯实验用的那些虫子，环顾四周，看到有些同伴依然活着。然而，即使知道某些个体还活着，也无法确定一个给定个体能否幸免于一死。我们根本没有任何办法能够预见某一独特个体对药物剂量的反应。就像皮尔逊统计模型里的那些个别观测值一样，它们都不是科学研究所关注的“事件”。惟有那些抽象的概率分布及其参数（如LD-50，半数致死剂量）才是能够估计的。

布利斯的概率单位分析一经提出　，其他研究人员也跟着提出了各种不同的数学分布。现代用来计算“LD－50”半数致死剂量的计算机程序，通常都会提供几种不同的模型让用户选择，这些模型都是在布利斯的原创基础上经过改进之后提出来的。用实际数据所做的研究表明，尽管在估计非常低的概率时，如“LD-10”，由这些不同模型得出的估计值是有差别的，但在“LD-50”上的估计值都非常接近。

我们完全可以运用概率单位分析或选择其他模型来分别估计一个不同的致死剂量，如“LD-25”或“LD－80”（25%的死亡剂量，或80%的死亡剂量）。不过，离50%点越远，就越需要更大规模的实验才能得到理想的

估计值。我自己就曾参与过一项研究，要确定某种能在老鼠身上致癌的化合物的LD-01（1%的致死剂量）是多少。我们的实验用了65000只老鼠，最终的分析结果表明，我们还是没能得到使1%老鼠致癌的化合物剂量的理想估计值。依据那项研究的数据推算，要想得到一个可接受的LD-01的估计值，我们得需要几亿只老鼠！

布利斯在列宁格勒

C?布利斯在概率单位分析上的开创性研究，到1933年却被迫中断了。那年，弗兰克林?D?罗斯福（Franklin D. Roosevelt）当选为美国总统。在竞选总统期间，罗斯福明确声称是联邦政府的赤字导致了经济萧条，并且保证他当选后会消减政府赤字，缩小政府部门的规模。虽然这并不是“新政”（the New Deal）最终的行为，却是竞选的诺言，因此这位新总统就职之后，他的一些内阁成员就遵照总统的竞选诺言，开始解雇一些非必要的政府工作人员。

那位协助农业部副部长负责研制新式杀虫剂工作的助理，当他在视察这个部门所做的工作时，发现有人居然不到有虫子的田间去做实验，反而无聊地躲在实验室里不厌其烦地用杀虫剂来做实验。于是，布利斯的实验室被关闭了，布利斯也被解雇了。当时正值严重的大萧条时期，他发现自己根本找不到工作。尽管布利斯曾发明了概率单位分析，但对于一个失业的昆虫学家，特别是一个与实验室的昆虫，而不是野外的昆虫打交道的昆虫学家来说，找不到工作实在是不足为奇。

布利斯与费歇尔取得了联系。费歇尔刚刚在伦敦得到一个新职位，他答案举荐布利斯，并给他一些实验设备，不过他不能给他一个工作岗位，因此也没有办法付给这位美国昆虫学家工作报酬。尽管如此，布利斯还是不得不去了英国。他与费歇尔及其家人一起住了几个月，并与费歇尔一起协作进一步完善了概率单位分析的方法论。费歇尔在布利斯的数学去处中发现了几处错误，并提出修改建议，得到的最终统计结果更为有效。布利斯按照费歇尔的修改建议，发表了一篇新论文。而费歇尔也把那个必不可少的统计表，补充编到他自己与弗兰克?耶茨（Frank Yates）联名写的有关统计表的那本书的新版中去。

布利斯在英国住了不到一年，费歇尔就为他找到了一份新工作，是在苏联的列宁格勒植物研究所（Leningrad Plant Institute）。试想一下，这个来自美国中西部地区中产阶级家庭、对政治漠不关心、而且永远不会学第二种语言的又高又瘦的家伙，随身带着只装了几件换洗衣服的一个小

行李箱，乘火车只身穿越欧洲大陆，终于到达列宁格勒火车站时的情景。而那时的俄国恰逢斯大林领导下的大清洗运动。

布利斯到达列宁格勒之后不久，聘请他来苏联的那个人的老板就被召到莫斯科去了，而且从此销声匿迹。一个月之后，那个聘请布利斯来苏联的人也被召到莫斯科去了，而且在返回途中“畏罪自杀”。负责布利斯旁边那个实验室的主管，也在某一天仓惶弃职，穿过拉脱维亚边境逃出了苏联。

就在这个时候，布利斯认真着手展开他的实验工作。他选了几组俄罗斯本地的害虫，用各种不同化合成分的杀虫剂来对这几组害虫进行试验，算出其概率单位极其“LD-50”半数致死剂量。他在研究所附近的房子里租了一个房间，他的俄罗斯女房东只会说俄语，而布利斯只会说英语。不过他告诉我，用各种手势加上亲切的微笑，他们相处得相当融洽。后来，布利斯遇见了一个来自美国的年轻女人，她为了投身于俄国伟大的共产主义实践，中断大学学业，满怀着年轻人的理想主义和马克思列宁主义的盲目崇拜来到苏联。她把可怜的只会说英语的布利斯当成好朋友，帮他购物、熟悉环境。此外，她还是当地的一个共产党员。党组织对布利斯的一切了如指掌，他们知道他何时受聘，何时抵达俄国，住在什么地方，以及在实验室里都做了些什么。

有一天，那女孩告诉他，党员里有些人已经认定他是美国间谍。她竭力为布利斯辩护，向他们解释布利斯是个单纯而又天真的科学家，只热衷于自己的实验。但是这些猜疑已经通报给了莫斯科当局，他们已经派出了一个委员会到列宁格勒来进行调查。

调查委员会就在列宁格勒植物研究所召开审查会，把布利斯叫来面对他们接受审问。当他走进审问室的时候，已经知道调查委员会里每个人的身份了，当然是他的女朋友透露的。他们几乎还没来得及调查完最初的几个问题，就在这时，布利斯对他们说：“我看到某某教授也坐在你们中间（告诉我这段往事的时候，布利斯已经不记得这位教授的姓名了），我一直在读他的论文。请告诉我，他提出的这种农业试验方法，是遵照圣人马克思和圣人列宁的绝对真理吗？“翻译踌躇着吞吞吐吐地把它这句话译了出来，刚一译完，审查委员会的委员们便一阵忙乱，他们要求布利斯对此做进一步的阐述。

“某某教授的方法”，布利斯接着又问：“就是正规的党的方式吗？就是按照党所要求的做法进行的农业试验吗？”

最终委员会给他的答案是，没错，这确实是做事情的正确方法。

于是布利斯说：“如果是那样的话，我就是违背了你们的信仰。”接着他进一步解释，如果按照这个教授提出来的做法，农业试验研究必须用很大面积的土地，而且所有这些农田都得用同样的实验方式来处理。布利斯说，他认为这样的试验是无益的，并且阐明他一直在倡导的方法，就是把农田分成很多小块地，以不同的方式处理相邻的地块。

审查工作没有再深入进行下去就结束了。那天傍晚，他的朋友告诉他，委员会已断定他不是间谍。他们认为他太率真了，透明得一眼就可以看穿，或许真是如她所说，他是一个头脑单纯、只关心他的实验的科学家。

其后，布利斯在列宁格勒植物研究所工作了几个月。他再也没有任何顶头上司了，他自己认为怎么做最好就怎么做。但是，他必须加入由实验室工作人员组成的工会组织，当时，每个在俄国工作的人都必须加入某个由政府控制的工会组织。除了这一点规定之外，他们就不管他了。在20世纪50年代，美国国务院还曾因为他一度属于一个共产党的组织，而拒绝给他签发美国护照。

突然有一天下午，他的女朋友冲进实验室，告诉他：“你必须马上离开。”他坚持说他的实验还没有做完，实验结果还没有详细记录下来，坚持要做完这些才肯离开。女友把布利斯从实验报告堆中拽出来，逼他赶紧穿上外套，告诉他刻不容缓，必须丢弃所有的一切，必须马上离开。刀子守候着催促着他，看着他装好那个小小的提箱，告别了女房东。女友把他送到火车站，临行前坚持要他在安全抵达里加（Riga，现拉脱维亚共和国的首都）时给她打个电话。

到了20世纪60年代，苏联的政治局势有了些微的松动，苏联的科学家重新回到国际科学团体中来。国际统计学会（International Statistical Institute, C?布利斯曾是该学会的会员）在列宁格勒召开了一次国际会议，会议期间，布利斯抽空去探访那些30年代的老朋友，但他们都已故去。他们当中，有的是在大清洗时期被杀，有的死于第二次世界大战，只有他当年的女房东还活着。见面时，他们不停地用各种手势，不断地点头，互致问候，并亲切拥抱，布利斯用英语低声地表达着对她的美好祝福，她则以俄语回应。

# 第9章 钟形曲线

读完这本书的前八章，你也许会以为统计革命只是发生在英国。从某种意义上说，这倒也是事实，因为最先将统计模型应用于生物研究和农业研究的，的确是在英国，还有丹麦。在费歇尔的影响下，统计学方法很快就传到了美国、印度、澳大利亚和加拿大。正当统计模型的实际应用在说英语的国家和地区推广之际，由于欧洲大陆长期形成的一种数学传统，使得欧洲的数学家正在研究与统计建模有关的理论问题。

这些理论问题中，最为重要的是中心极限定理（central limit theorem）。直到20世纪30年代初，这还是个未经证明的定理，或者说只是一个猜想（conjecture），因为许多人都信其为真，却没有一个人能证明它成立。费歇尔早在研究似然函数值的理论时，就曾假设这个定理是成立的；而回溯到19世纪初，法国数学家皮埃尔?西蒙?拉普拉斯也用这个推论证明了他的最小平方法（method of least squares）。此外，心理学这门新兴科学也是根据中心极限定理开创了智力测验技术与精神疾病量表。

什么是中心极限定理？

大量数据集合的平均数都有一个统计分布，而中心极限定理则阐明，无论初始数据是怎么来的，这个分布都可以用正态概率分布来逼近。这个正态概率分布与拉普拉斯的误差函数（Laplace’s error function）相同，有时也叫做高斯分布（Gaussian distribution），而在浅显通俗的普及书里，也常被称为“钟形曲线”（bell-shaped curve）。在18世纪晚期，亚伯拉罕?棣莫弗（Abraham de Moivre）已经证明，由机会博弈(games of chance)所得数字的简单集合符合中心极限定理。然而，在此之后的150年里，对这个猜想的证明没有丝毫的深入进展。

用正态分布来描述大部分数据都是正确有效的，因此，中心极限定理普遍被认为是一个正确的猜想。一旦假定数据服从正态分布，数学上的处理就容易多了。正态分布具备某些非常优良的性质：如果有两个随机变量服从正态分布，那么两变量之和也同样服从正态分布。就一般而言，

正态变量的各种类型的和与差也都服从正态分布。因此，由正态随机变量（variate）推演得出的许多统计量，其自身也服从正态分布。

正态分布只有K?皮尔逊四个参数中的两个——平均数和标准差，另外两个参数对称性偏度（symmetry）和峰度(kurtosis)均为零。因此，一量知道了平均数和标准差这两个参数值，其他的一切也就一清二楚了。费歇尔曾指出，由一组数据得出的平均数与标准差的估计值就是他所说的充分估计量（sufficient estimator），因为这两个参数值已经把这些数据中所有的信息都包括在内了。既然这两个参数值已经涵盖了能够从那些原始测量值中揭示出的一切，就根本没有必要去占有任何原始测量值了。如果有足够的测量值可以用来相当精确地估计出平均数与标准差，就不再需要其他任何测量值了，任何为搜集这些数据所做的努力，都不过是浪费时间而已。例如，有两个重要指标服从正态分布，如果你正打算得出这样一个正态分布的那两个参数，那么你只需要收集大约50个测量值就足够了。

正态分布的这种数学上便于处理的特性，使科学家能够构建一个复杂关系模型。只要其基本分布是正态的，费歇尔的似然函数通常就有了以简单代数进行处理的一种形式。即便模型复杂到必须用迭代运算法去解的程度，只要其分布是正态的，用纳恩?莱尔德（Nan Laird）和詹姆斯?韦

尔(James Ware)的EM演算法去解，就变得轻而易举了。由于正态分布在数学上的计算处理非常敏捷，因此在建模时，统计学家常常要假定所有的数据都服从正态分布。不过，做这样的假定就不能不援引中心极限定理。

但是，中心极限定理是否成立？说得更准确一点，它在什么条件下成立？

在20世纪20年代和30年代，斯堪的纳维亚地区、德国、法国和苏联的一批数学家，运用20世纪早期发明的一套新的数学工具，倾心于上述这些问题的研究。但就达个时候，整个人类文明都正面临着一场日益迫近的浩劫——那些极权主义的国家的恶性膨胀。

数学家并不有昂贵设备的实验室。在20世纪二三十年代，黑板和粉笔就是一个数学家最具代表性的实验设备。对数学研究而言，用黑板比用纸张更方便，因为数学研究过程的演算总免不了出错，而黑板上的粉笔字很容易擦掉。几乎没有数学家是关起门独自做研究的，只要你是一个数学家，你就必定要同其他的数学家一起讨论自己在研究的问题，你就必定要接受别人对你那些新想法的批评审视。在数学研究过程中太容易出错，或者太容易在研究中隐含着自己毫无察觉而在别人看来却是显而易

见的假设。有一个数学家的国际组织，在这个团体中，数学家们书信往来、开会、审阅彼此的论文，经常交换相互的批评和质疑，探究分歧所在。20世纪30年代初期，德国的威廉?费勒（William Feller）和里夏德?冯?米泽斯（Richard von Mises），法国的保罗?利维(Paul Lévy)，俄罗斯的安德烈?柯尔莫哥洛夫（Andrei Kolmogorov），斯堪的纳维亚的亚尔?瓦尔德马?林德伯格（Jarl Waldemar Lindeberg）和哈拉尔德?克拉美（Harald Cramer），奥地利的亚伯拉罕?沃尔德（Abraham Wald）和埃尔门?哈特利（Herman Hartley），意大利的圭多?卡斯泰尔诺沃（Guido Castelnuovo），还有许多其他数学家也都在这个团体中，其中不乏那些利用新工具来检验中心极限定理这个猜想的数学家。

然而，这种自由轻松、无拘无束的相互交流不久就将不复存在。它将毁于斯大林的肃反运动、纳粹的种族灭绝和墨索里尼的帝国梦。黑暗笼罩着欧洲。斯大林正把非法操纵的示众式的公开审讯同半夜里的秘密逮捕结合运用到了极致，处决、恐吓和威胁任何一个受到他偏执狂式的无端

猜疑的人。起初，希特勒及其罪大恶极有党羽把犹太裔教授从各大学里清洗出去，随后将他们关进惨无人道的集中营。墨索里尼则把国人强行禁锢在他所谓的“组合国”（Corporate state）所划定的各个社会等级中。

“死亡万岁！”

这一猖獗的、反理智主义（anti-intellectualism）的极端事件，就发生在西班牙内战时期。当时长枪党的党徒们（以西班牙的法西斯主义者闻名）已经占领了古老的沙拉曼卡大学（University of Salamanca）。该大学的校长是享誉世界的西班牙哲学家米格尔?德?乌纳穆诺（Miguel de Unamuno），当时他已经70岁出头了。长枪党的米连?阿斯特赖（Millan Astray）将军，一个在先前的战争中失去了一条腿、一只手臂和一只眼睛的残疾人，当时任这个刚以武力控制了西班牙的恶势力的宣传部长。他的座右铭就是：“死亡万岁！”如同莎士比亚笔下的国王理查德三世，阿斯特赖身体上的残缺不全恰恰映射出他扭曲的邪恶心灵。有一次，长枪党在沙拉曼卡大学的纪念大厅举行盛大的庆典，台上有新指派的省长、弗朗西斯科?佛朗哥（Francisco Franco）夫人、M?阿斯特赖、沙拉曼卡的大主教，还有年事已高的乌纳穆诺，他是被当作被征服的战利品拖到台上的。

“死亡万岁！”阿斯特赖高声狂呼，挤满了人群的大厅里随声附和着他的喊叫。又有人高呼：“西班牙！”大厅里的人也跟着喊。“西班牙！死亡

万岁！”穿着蓝色制服的长枪党的党徒们齐刷刷地站起来高呼，并朝着台上的佛朗哥肖像行法西斯的举手礼。就在这一浪高过一浪的叫嚣声中，乌纳穆诺站起身来，从容地走向讲台，镇静地开始演讲：

你们大家都记住我的话。你们都了解我，并且知道我不可能保持沉默，因为沉默也可以解释为默认，沉默中常常意味着谎言。我想对刚才的演讲做个评论，我们不防就叫它“M?阿斯特赖将军的演讲”吧……。就在刚才，我听见一种嗜尸成癖的愚蠢无知的叫嚣：“死亡万岁！”而我，

一个终生致力于各种悖论研究的人……我必须告诉你们，作为一个权威，这种荒诞怪异、语无伦次的谬论让我恶心。阿斯特赖将军是个残疾人……他是战争造成的一个残疾人……。不幸的是，眼下的西班牙这种残疾太多了。而且不久，如果上帝不能拯救我们，这种残疾人甚至还会更多……。

M?阿斯特赖把乌纳穆诺推到边上，厉声吼叫：“该死的臭知识分子！死亡万岁！”与他的叫嚣相呼应，那些长枪党徒们蜂拥而上，抓住乌纳穆诺。但是，老校长仍然继续说道：这里是知识的殿堂，而我才是这个殿堂的领袖。是你们亵渎了这个神圣的地方。你们可以凭借极其残暴的兽行获胜，但是你们无法得到人们的认可。因为要让人认可必须靠说服而不是压服，要达到说服的目的所必须具备的东西，恰恰是你们所没有的，那就是理智和正义……。

乌纳穆诺遭到软禁，不出一个月就被宣告“自然死亡”。

苏联的大清洗运动切断俄国数学家与欧洲其他地方的联系；希特勒的种族政策几乎毁掉了德国的大学，因为欧洲许多伟大的数学家要么是犹太人，要么是与犹太人联姻，而非犹太裔的那些数学家又大多是反纳粹的。结果，威廉?费勒去了美国的普林斯顿大学（Princeton University）

，亚伯拉罕?沃尔德到哥伦比亚大学（Columbia University）去了，埃尔门?哈特利和里夏德?冯?米泽斯去了英国伦敦，埃米尔?J?冈贝尔逃到了法国，埃米纳脱（Emmy Noether）在美国国宾夕法尼亚的布林莫尔学院（Bryn Mawr College）求得一个临时工作。但是，并非每个人都逃得脱。不能出示证明受聘到美国去工作的那些人，美国移民局对他们总是大门紧闭；而拉丁美洲那些国家的国门则由于那些小官僚的反复无常而

时开时关。纳粹军队占领了波兰首都华沙后，大肆搜捕能找到的华沙大

学所有的教授和学者，逮捕他们并惨绝人寰地将他们杀害，然后一起埋在一个巨大的坟墓里。在纳粹的种族主义世界里，波兰人和其他斯拉夫人只配做他们这些亚利安（Aryan）主人的奴隶，没有受教育的权利。欧洲那些历史悠久的大学里许多有培养前途的青年学生就这样被毁掉了。在苏联，大部分数学家都躲进了纯数学中去寻求庇护，而不敢在应用领域中做任何尝试。因为，那些从事应用研究的科学家，正受到斯大林令人不寒而栗的无端猜疑。

不过，在这些黑暗没有完全成为现实之前，欧洲的数学家们就已经解决了中心极限定理的证明问题。芬兰的亚尔?瓦尔德马?林德伯格和法国的保罗?利维分别发现了能够使中心极限定理这个猜想成立所必需的一组重叠的条件。这证明了至少存在三种解这个问题的方法，而且证明了中心极限定理不是只有一个单个的定理，而是有一组定理，其中每个中心极限定理都能从略有区别的一组条件中推导出来。到了1934年，中心极限定理（组）终于不再是猜想了，一个科学家必须要做的只是要证明林德伯格?利维条件（Lindeberg-Lévy Conditions）成立，那么中心极限定理就成立，于是，他就可以随意地把正态分布设为一个合适的模型。

林德伯格?利维条件与U统计量

然而，就一个特定情况而言，要证明林德伯格?利维条件成立很难。但在理解林德伯格?利维条件上倒有几分安慰，因为他们描述的条件看上去是合理的，而且在大多数情况下都是成立的。不过要证明其成立却是一个棘手的问题，这也正是战后远在北卡罗莱纳大学辛苦工作的瓦西里?霍夫丁（Wassily Hoeffding）在这个故事中竟会有如此重要地位的原因。1948年，霍夫丁在《数理统计年报》（Annals of Mathematical Statistics）上发表了一篇论文，题目是“渐近正态分布的一组统计量”。

回想费歇尔曾把统计量（statistic）定义为：从观察到的测量值得出的、可用来估计其分布参数的一个数值。费歇尔还建立了有用的统计量应该具备的一些准则，在这个过程中，他还指出了利用皮尔逊的许多方法导出的统计量不符合这些准则。有很多种计算统计量的不同方法，其中的很多统计量都能满足费歇尔提出的准则。一旦计算出统计量，为了要用它，我们必须知道它的分布。如果它服从正态分布，用起来就容易多了。霍夫丁提出了一种他所谓的“U-统计量（U-statistics），并指出一个统计量如果属于这种U-统计量，则满足林德伯格?利维条件。正因为如此，我们只须指出一个新的统计量是否与霍夫丁的定义相一致，而不必去解那些很困难的数学来证明林德伯格?利维条件成立。霍夫丁所做的

一切就是用一组数学必要条件取代另外一组。然而，霍夫丁的条件事实上很容易检查。因此，霍夫丁的论文发表之后，几乎所有的论文在证明一个新统计量服从正态分布的时候，都是通过证明该统计量是一个U统计量来完成的。

霍夫丁在柏林

第二次世界大战期间，霍夫丁处在一个不确定的微妙境况中。他1914年出生在芬兰，父亲是丹麦人，母亲是芬兰人。第一次世界大战之后，芬兰沦入俄罗斯帝国的统治，就在这个时候，霍夫丁随家人迁往丹麦，随后又迁往柏林，因此他拥有斯堪的纳维亚地区两个国家的双重国籍。1933年他高中毕业，随后开始在柏林攻读数学。就在那个时候纳粹开始在德国掌权。预料到以后可能发生的事，霍夫丁就读的那所大学的数学系的系主任R?冯?米泽斯早早地离开了德国，不久之后，为霍夫丁授课的其他许多教授，有的逃走了，有的被解除了职务。在动乱中，年轻的霍夫丁所选的课都是由一些低水平的教师来讲授的。即便如此，这些教师中的很多人也没能维持到把他们承担的课程教完，因为纳粹在持续不断地“净化”大学教师队伍，把大学教师中所有的犹太人和犹太人同情者全都清除出去。随同数学系里的其他学生一道，霍夫丁被迫去听路德维希?比贝尔巴赫（Ludwig Bieberbach）讲授的一堂课。比贝尔巴赫一直都是教师中的小角色，只是因为对纳粹党的狂热拥护，才合他成为数学系新的系主任。比贝尔巴赫这堂课讲的是“亚利安”数学与“非亚利安”数学的区别，他声称颓废的“非亚利安”（解读为犹太）数学家仰仗着复杂难解的代数符号做研究，相反，“亚利安”数学家则在更高贵、更纯粹的几何直觉领域里从事研究。结束了讲课的时候，他让学生提问题。坐在后排的一个学生问他，为什么偏偏是这个里夏德?库朗（Richard Courant，20世纪初德国伟大的犹太裔数学家之一）运用几何洞察力创建了他的实分析理论（theories of real analysis）。此后，比贝尔巴赫再也没有就这个题目上过公开课。但是他创办了《德国数学》（Deutsche Mathematik）杂志，这个杂志很快就成为当政者眼中居第一位的数学期刊。1940年，霍夫丁完成了他的大学学业，像他这个年龄的其他男青年都要应征到部队去服兵役，但由于他的双重公民身份，并且当时的芬兰已成为德国的一个盟国这样的事实，他因此不必服兵役。他找到一份工作，在一家跨校际的精算杂志社的办公室兼职。与比贝尔巴赫创办的那个杂志不同，这是一个很难约到论文，因此也很难定期出版发生的杂志。霍夫丁甚至连寻找一份教书的工作都不能，因为他必须申请到正式的德国公民身份才有资格去教书。

1944年德国政府宣布，具有“德国血统或相关血统“的非德国籍青年也要服兵役。不过，在霍夫丁体检的时候，发现他患有糖尿病而免于服兵役。这时他终于有了找工作的资格。他兼职的那家期刊的编辑哈拉尔德?格佩特（Harald Geppert）建议他从事某种军事应用方面的数学研究工作，他提这项建议的当时，期刊的另一个编辑赫尔曼?施密德（Hermann Schmid）也在场。霍夫丁犹豫了一下，然后，出于对格佩特的谨慎的依赖，他对格佩特说，任何一种与战争有关的工作都违背他的良心。施密德出身于一个普鲁士贵族家庭，霍夫丁希望他的家族荣誉感能让他对这次谈话守口如瓶。

随后的几天里，霍夫丁一直提心吊胆的，但什么事都没有发生，他得以继续他的研究。当俄国军队逼近柏林的时候，一天早上，格佩特在早餐里放了毒药喂给他年幼的儿子，随后他和他的太太也服毒自杀了。1945年2月，霍夫丁和他的母亲一起逃到汉诺威的一个小镇上，他们在那里的

时候，这个地方成为英军占领区的一部分。而他父亲仍滞留在柏林，在那里，他被俄国秘密警察以间谍罪逮捕，因为他一度曾为美国驻丹麦的商务参赞工作过。好几年时间，他杳无音信，直到他设法越狱，又历尽千辛万苦逃到了西方。在此期间，年轻的霍夫丁于1946年秋天到达纽约

，继续他的学业，后来应邀到北卡罗莱纳大学任教。

运筹学

纳粹的这种反理智主义、反犹太主义倒行逆施的结果之一，就是让第二次世界大战的同盟国因此而丰收了许多才华横溢的科学家与数学家，在他们的鼎立相助下打赢了这场战争。英国生物学家彼得?布莱克特（Peter Blackett）向海军部建议，武装部队应该请一些科学家来协助解决战略和战术上的问题。无论是哪个专业研究领域的科学家们，他们都训练有素，能够应用逻辑和数学模型来解决问题。他建议组成科学家的攻关小组，让这些小组从事有关战争问题的研究，由此诞生了一门新学科——“运筹学”（operational research，在美国称之为operations research）。从事不同领域研究的科学家组成的科研小组联合起来共同研究，决定用远程轰炸机对付潜艇的最佳使用方案；为防空武器提供射击表；决定靠近前线的军火补给站的最佳选址；甚至还要解决军队的食物补给问题。战争结束后，运筹学的应用由战场搬到了商场。那些在战争期间被征募到军队去服务的科学家已经证明了用数学模型和科学的思

维来解决战事中的战术问题是多么有用。同样的步骤和许多相同的方法也能用来组织工厂里的生产，找出仓库与销售部门之间的最优关系，解决许多别的商务问题，均衡有限的资源，或改进生产与提高产量。从那时候起，大公司里大部分都设立了作业研究部门，而这个部门所从事的多数工作都与统计模型有关。

我在辉瑞公司工作的时候所做的几个项目，其目的都是为了改善对药物研究进行控制和提取新产品进行测试的方法，在所有这些研究中涉及到的一个重要方面就是，当条件可以满足时，有能力用正态分布去处理问题。

# 第10 章 拟合优度检验

20世纪80年代，出现了一种新型数学模型，激起了公众的遐想，主要是因为这种数学模型的名字——混沌理论（chaos theory ）。这个名字提示着某种形式的统计建模明显带有杂乱无序特征的随机性。创造了这个名字的人有故意避开使用随机（random）这个词的嫌疑。实际上混沌理论是尝试着在一个更高端的层次上，通过复兴决定论（determinism）来动摇统计革命。

回想一下，在统计革命之前科学所处理的那些“事件”，要么是已有的测量，要么是生成这些测量值的自然事件。伴随着统计革命，科学的事件变成了能左右测量值分布的参数。在早期的确定性方法中，有一个信条是，越精确的测量，对所考察的自然客体的描述也就越精确；而在统计方法中，分布参数有时候不必有一个自然客体，无论多么精确的测量系统，分布参数的估计值终究是有误差的。例如，在确定性方法中，重力常数是描述物体如何向地球下落的一个恒定不变的值；而在统计方法中，我们对重力常数的测量值永远都不会是一样的。为了“通晓”落体的性质，这些测量值分布的离散状态才是我们想要确立的。

1963年，混沌理论专家爱德华?洛伦兹（Edward Lorenz）做了一个后来时常被引用的演讲，演讲题目为“巴西一只蝴蝶翅膀的翩翩舞动，会引起德克萨斯州的龙卷风吗？”洛伦兹的主要论点是，混沌的数学函数对初始条件非常敏感，初始条件的些微差异，经过多次迭代之后，中以导致全然不同的结果。洛伦兹相信，由于存在这种对初始条件微波差异的敏感性，以至于对所研究的问题不可能得出一个确定的答案。隐含在洛伦兹演讲中的是确定性假设，即理论上每一个初始条件都是促成某个最终结果的一个起因。这个被称之为“蝴蝶效应”（butterfly effect）的观念，已经被那些混沌理论的普及者们当作一个深邃而睿智的真理接受下来了。

然而，没有任何科学的证明揭示了这样一种因果关系的存在，也没有任何数学模型有准确的依据表明客观现实中存在着这一效应。它只是一种信念的表述而已，就其科学的有效性而言，它与关于鬼神的描述相去无几。而统计模型是用分布参数来对科学探索明确地进行解释，它们也是建立在对现实世界的一种信念所做的描述上。然而，我自己在科学研究上的经历让我确信，比起对信念的确定论的陈述，统计上的陈述更有可

能是真实的。

混沌理论与拟合优度

混沌理论源于这样的观察：一个固定不变的确定性公式生成的数字有可能看上去是一个具有随机性的模型。早在一批数学家处理相对简单的迭代公式并绘出其结果的时候，就曾经发现过这种现象。在第9章，我曾经把一个迭代公式描述为：首先得到一个数，接着把这个数代入方程式中得到另一个数，用第二个数又得到第三个数，如此等等。其实，早在20世纪的最初几年，法国数学家亨利?普安卡雷（Henri Poincaré）就尝试着把这些连续的成对数值绘在图上，用这种方式理解一组复杂的微分方程式。普安卡雷在图中发现了一些值得关注的图式，却因不知道如何对这些图式做进一步的研究而放弃了深入研究的想法。而混沌理论就是以普安卡雷的这些图式为起点发展起来的。当你在绘制一张普安卡雷图形（Poincaré plots）时，会发现图纸上出现的那些点起初好像完全不成形状，表面上这些点以一种偶然的方式出现在随便什么地方，但承受着绘在图上的点数的不断增加，图式开始显现出来，有时是几组平行线，有时也可能是一组相互交叉的线，或许是很多个圆，或是和直线相交的圆。

混沌理论的拥护者认为，现实生活中那些看上去是纯随机的测量值，实际上是由某个确定性的方程组生成的，这些方程可以从普安卡雷图形的模式推演出来。例如，有些混沌理论的拥护者记录下了人类心脏动脉搏动的间隔时间，并绘成普安卡雷图形。他们声称在这些图上看到了一些形状，并且已经发现一些似乎能产生同类形状的确定性生成方程。

直到写这本书时为止，以这种方式应用的混沌理论仍存在着一个严重的缺陷。根据数据绘出的图形与用一组特定方程组生成的图形，这两者之间的拟合度如何，并未测量。他们只是要求读者观察两种相似的图形，并以此为依据证明给出的生成方程是正确的。统计分析上已经证明这种用肉眼检验的方式难免出错。因为，用肉眼判断类似的或几乎完全相同的两个图形，如果改用为此目的创建的统计分析工具仔细检验之后会发现，两者往往是大不相同的。

皮尔逊的假使优度检验

这是K?皮尔逊在他的学术生涯早期就已经意识到的一个问题，K?皮尔逊最伟大的成就之一就是创造出第一个“拟合优度检验“（goodness of fit

test）。通过观测值与预测值的比较，皮尔逊构造出一种能对拟合优度进行检验的统计量，并称之为“χ2拟合优度检验”（chi-square goodness of fit test）。之所以用希腊字母χ（读作“kai”），是因为这个检验统计量的分布属于一组偏斜分布，而他称这组偏斜分布为χ家族(chi family)。实际上，这个检验统计量很像χ的平方，因此命名为“χ2”。在费歇尔看来，既然是一个统计量，就会服从一种概率分布。K?皮尔逊证明了无论用哪一种类型的数据，χ2拟合优度检验都服从相同的分布。也就是说，他能列出这个统计量的概率分布表。每一个检验都能用到同样的那套表。χ2拟合优度检验只有一个参数，费歇尔称之为“自由度”。费歇尔在1922年的那篇论文里，首次批评了皮尔逊的研究，指出在比较两种比例时，皮尔逊得出的那个参数值是错误的。

但是，没有任何理由只因为皮尔逊理论上的一个很小的错误，就贬低他的这项伟大成就。皮尔逊的拟合优度检验是现代统计分析中一个重要组成部分的先驱，这个重要组成就是“假设检验”（hypothesis testing）或“显著性检验”（significance testing），它允许分析人员提出用来模拟现实的两种（或多种）不一致的数学模型，然后利用数据来放弃其中的一个。假设检验应用得如此广泛，以至于很多科学家认为这是他们唯一能用的统计方法。在后面的章节中我们会发现，假设检验的应用甚至涉及到一些严肃的哲学问题。

检验女士是否真能品尝出茶的区别

假设我们要检验那位女士能否品尝出两杯茶的不同：是把牛奶倒进了茶水里，还是把茶水倒进牛奶里。我们给她两杯茶，告诉她一杯是茶水倒入牛奶里，另一杯是牛奶倒入茶水中。她尝了尝，正确区别开了这两杯茶。有可能她是凭猜测，猜对的机会是一半对一半。我们再给她同样的这样两杯茶，她又说对了。如果她仅仅靠猜测，那么连续两次都猜对的机会是四分之一。如果我们再给她两杯茶，假如她仍然能正确地分辨出来。若这人结果完全是猜出来的，此时猜对的机率则只有八分之一。我们继续两杯两杯地让她品尝更多杯茶，而她依然每次都能够正确地识别出来。某种意义上，我们就不得不相信她真的能品尝出其中的差别了。假定她说错了一次，假定说错的这一次就发生在第24组，而其他的全对，那么我们能否依然认为她真的有分辨不同奶茶的能力呢？假如她的错误是二十四分之四呢？或是二十四分之五呢？

假设检验（或者说显著性检验）是一种正规的统计方法，是在“待检验的假设为真”的假设前提下，用来计算以往观测到的结果发生的概率。

当观测结果发生的概率很低时，我们得出原假设不成立的结论。重要的一点是，假设检验提供了一种拒绝某个假设的工具。上述例子中，待检验的假设是：那位女士只是凭猜测。假设检验的目的不是让我们接受某个假设，即使与那个假设有关的概率非常高也不能接受。

在这个普遍被接受的概念发展的早期，“significant”（显著的）这个词只是用来指“概率低到足以拒绝的程度”，数据如果可以用来拒绝某个分布，则它就是显著的。在19世纪后期的英语里，这个词仅仅是指计算结果意味着或表明了什么意思。进入20世纪之后，英语“significant”这个词在原有含义的基础上又扩展了其他的解释意义，也指某些事情是非常重要的。在某个待检验的假设条件下，统计分析仍沿用“significant”这个词“显著的”含义来表示计算结果发生的概率很低，在这个层面上，“significant”这个词有一个精确的数学涵义。但令人遗憾的是，使用统计分析的人常把显著性检验统计量理解为某种更接近这个词的现代语意的东西。

费歇尔对P值的运用

现在运用的显著性检验方法，其中大部分都是费歇尔构造出来的。他把判定具有显著性的那个概率，称为“P值”（P-value）。他对P值的涵义和有效性坚信不疑。在《研究工作者的统计方法》一书中，很多地方都专门介绍了怎么计算P值。正如我在开头的时候谈到的，这是一本专门给想要应用统计方法的非数学专业人士写的书。在这本书中，费歇尔并未解释这些检验是如何推导出来的，也从没有明确指出究竟多大的P值才算是显著的。他只是举出一些计算实例，并说明结果是否显著。在一个例子中，他给出一个小于0.01的P值，并且说明“一百个值当中，只有一个值会偶然超过（计算出来的检验统计量），因此，很显然，计算结果之间的差异具有显著性。”

1929年，费歇尔在《心灵研究学会刊》（Proceedings of the Society for Psychical Research）上发表的一篇论文中，几乎等于定义了一个在任何情况下都将是显著的特殊的P值。“心灵研究”（psychical research）提到试图用科学的方法来证明“超视力”的存在。心理学的研究人员大量运用了统计学的显著性检验来证明，在受实验者完全随意猜测这种假设条件下，其结果是不可能的。费歇尔在他这篇论文中，先是谴责某些作者完全错误地使用了显著性检验，接着他申明说：

运用生物学的方法对生物界进行观察的时候，统计学的显著性检验是必

不可少的。其作用就在于防止我们被一些非主要的偶发事件所欺骗。并不是因为我们希望去研究或试图去查明这些偶发事件，而是因为它们与许多我们无法控制的其他境况联系在一起。一个观测的结果，倘若在我们正在寻找的真正原因根本不存在的情况下，几乎从未发生过，可以判断这个观测具有显著性。如果偶然发生的机率低于二十分之一，通常的做法是判断其结果具有显著性。对实际调查者来说，显著性水平的选择是任意的，但便于应用。不过，它并不意味着可以让自己每20次实验中就被骗一次。显著性检验只是告诉他什么是应该忽略掉的，也就是说应该把所有那些无法得到显著性结果的实验忽略掉。当他知道如何设计一个实验，而这个实验几乎一定能给出一个显著性的结果时，他也只能说明，这仅是一种实验上可以验证的现象。所以，对那些孤立的具有显著性的结果，他不知道如何才能让它们再现出来，只能留待以后再做进一步的调查研究了。

注意“……知道如何设计一个实验，而这个实验几乎一定能给出一个显著性的结果……”这句话，正是费歇尔使用显著性检验的核心之所在。对费歇尔而言，显著性检验只有在连续实验的相互联系中才有意义，所有这些实验的目的在于解释特定处理的作用。读过费歇尔的应用性论文之后，你会在他的引导下相信，使用显著性检验是为了得出三种可能的结论之一：如果P值很小（通常小于0.01），他断言某种结果已经显现出来；若P值很大（通常大于0.2），他宣称即便真的存在一个结果，也会因为该结果发生的可能性太小，所以不可能有任何显示出这个结果的大规模的实验；如果P值介于前两者之间，他讨论了应该如何设计下一个实验，才能得到一个更好的结果。除了上述情况，费歇尔从来没有明确说明科学家应该怎么解释P值。对费歇尔而言，看上去是如此显而易见的事，对读者来说可能并不清楚。

我们将在第18章回过头来重新审视费歇尔对显著性检验的态度。费歇尔始终坚持，从来都没有显示过吸烟有害健康，这也正是他的一个较大错误的核心之所在。费歇尔对有关吸烟和健康的证据做了犀利的分析，我们暂且把它放下，以后再谈。现在把话题转到1928年，看看当时35岁的耶日?奈曼。J?奈曼的数学教育

当第一次世界大战在东欧爆发，奈曼的祖国陷于战火之中的时候，他还是一个在数学系读书的非常有发展前途的大学生。他被迫搬到俄国，就读于卡尔可夫大学（University of Kharkov）——一个远离数学活动的视野偏狭的地方。学校缺少具有当代最新数学知识的合格老师，而且由于

受到战争的影响，他是在学期中途才入学的，因此，在卡尔可夫，他只学到一些最基础的数学知识。奈曼只能寄希望于那些能得到的数学期

刊，从中查找论文文献。可想而知，奈曼受到的正规的数学教育只相当于19世纪学生学到的内容，20世纪的数学知识则是他通过自学掌握的。

对奈曼来说，可利用的数学期刊仅限于卡尔可夫大学的图书馆和后来在当地的波兰学校图书馆里能找到的。偶然的机会，他发现了亨利?勒贝格（Henri Lebesgue 1857-1941）的一套论文集。20世纪的最初几年，勒贝格提出许多现代数学分析的基本思想，但是他的论文晦涩难懂。后来的数学家把勒贝格积分、勒贝格收敛定理以及这个伟大数学家的其他一些创见简化并整理成更容易理解的形式。现在已经没有人再去读勒贝格的原著了，学生们都是通过阅读这些新版的文章来学习勒贝格的思想。

所谓的“没有人”当然是除了奈曼之外的，当时他只有勒贝格的原文可以读，他苦读这些原文，从中感受到了这些全新的（对他而言）伟大创见所蕴含的辉煌。此后的许多年，奈曼一直非常景仰勒贝格，20世纪30年代末在法国的一次数学研讨会上，终于得以与勒贝格见面。据奈曼所

说，勒贝格表现得态度生硬、粗鲁无礼。当奈曼热情洋溢地表达对他的仰慕时，他阴郁冷淡地回应了一句，就转身离开了正在喜出望外地等待与他交谈的奈曼。

这种冷淡让奈曼深受伤害，并且，奈曼可能把这次经历当作了反面教训，他对青年学生一直都格外的亲切有礼，仔细地倾听他们的谈话，并对他们的热情给予鼓励和回应。奈曼正是这样的一个人。所有认识他的人都对他的亲切和蔼、富于同情心的为人记忆犹新。他与人为善、体贴入

微、待人真实宽厚。当我见到他的时候，他已经80多岁了，一个身材瘦小、举止高贵、衣着讲究、蓄着整洁白胡须的老人。他在听别人讲话和别人深入交谈的时候，蓝眼睛神采奕奕地闪烁着，对每个人都同样地全神贯注，无论对方是谁。在他的职业生涯之初，奈曼好不容易才找到工作，成为华沙大学（the University of Warsaw）的一个年轻的教师。当时，刚刚独立的波兰因资金短缺，没钱资助学术研究，也很少有给数学家的职位。1928年，他在伦敦的生物统计实验室呆了一个暑假，并认识了E?皮尔逊和他的太太艾琳（Eileen）以及他们的两个女儿。E?皮尔逊是K?皮尔逊的儿子，但是父子两人在个性上的天壤之别可谓绝无仅有：

K?皮尔逊精力充沛，有支配控制他人的欲望；E?皮尔逊却腼腆谦虚。K?皮尔逊喜欢追逐新观念，常在数学概念还相当模糊，甚至还存在某些错误的时候，就忙着发表论文；E?皮尔逊则极其小心谨慎，甚至为每一步计算的细枝末节担忧。

E?皮尔逊与奈曼的深厚友谊长存在两人1928-1933年间的通信中。这些信件展示了他们对社会科学卓越的洞察力，以及两颗富于独创精神的心灵是如何提出各自的想法，或批评对方的想法，并共同解决难题的。E?皮尔逊踌躇地指出奈曼的提议或许不可行，这时他表现出谦逊的一面；奈曼巧妙地剖析复杂的问题，并抓住每个难题的重要本质，这时展现出他的独创力。有人如果想知道数学研究为什么是需要经常进行合作的事业的话，我建议他看看奈曼与E?皮尔逊的通信。

E?皮尔逊对奈曼提出的第一个问题是什么呢？回想K?皮尔逊的χ2拟合优度检验，他创立这种方法来检验观测数据是否与理论分布相符。但事实上根本不存在像χ2拟合优度检验的这种东西。分析人员有无数种方法可用来对给定的一组数据进行检验，似乎没有任何准则能够判定如何在这么多的选择中挑选出“最好的”。每次用到检验的时候，分析人员必须做出一个相当随意的选择。对此，E?皮尔逊问了奈曼以下的问题：如果我用了χ2拟合优度来检验一组服从正态分布的数据，但我没能得到一个显著的P值，那么我怎么知道这组数据确实服从正态分布呢？也就是说，我怎么知道至今尚未发现的另一种χ2检验或者另一种拟合优度检验不会已经产生了一个显著的P值，而允许我在拟合数据的时候拒绝这个正态分布呢？

奈曼的数学风格

奈曼把这个问题带回华沙，并由此而开始了两人之间的书信往来。奈曼与小皮尔逊都对费歇尔建立在似然函数基础上的估计概念印象深刻。通过检查与拟合优度检验联系在一起的似然函数，他们开始了调查研究。两人联名发表的第一篇论文介绍的就是那些研究的结果。这是他们撰写的三篇顶尖论文当中最难的一篇，它几乎彻底变革了关于显著性检验的全部思想。当他们继续探索这些问题时，奈曼极度清晰的洞察力使问题在蒸馏中不断提纯，精炼出最基本的元素，使他们的研究成果变得更为清晰，也更容易理解。虽然读者对此可能不太相信，但在数学研究领域，一个人写文章的风格确实发挥着很重要的作用。有些数学文献的作者似乎写不出让人容易理解的文章；有些人则似乎以写成一行又一行的数学符号与注释为乐事，一篇论文中充斥着无比繁琐的细节，以至于把

总的思考都迷失在了微不足道的细节中。与之相反，有些作者却总是有能力用非常简单而有说服力的方式表达复杂的思想，数学的发展在他们的表达中显得如此的鲜明而平实。只有在回顾已经学到些什么时，读者才会确实认识到结果的伟大力量。奈曼就是这样的作者，读他的论文是件令人愉快的事，数学观点自然地展开，使用的符号简单得令人无法相信，结论的显现竟如此的自然，以至于让人感到难以理解，不禁要问，为什么很久以来居然没有人发现这项结论？

我在辉瑞的研究中心工作了27年，该中心每年都赞助康涅狄格大学举办一次学术年会。该校的统计系通常会邀请一位生物研究方面的重要人物来一天，与学生们见面聊聊，随后，会在下午的晚些时候发表演讲。由于我曾经参与负责一年一度的研讨会的资金事宜，因此有幸会见统计学

界的一些大人物，奈曼就是应邀者之一。在一次研讨会前，奈曼想让他的演讲以一种特殊的方式进行，他先介绍一篇论文，随后组织一个专题组来评判他的论文。由于是大名鼎鼎的奈曼，研讨会的组织者联系了美国新英格兰地区著名的资深统计学家组成了这个专题讨论组。在研讨会开幕前的最后一记得，有位专题组成员无法出席，于是会议安排我代替他。奈曼事先已经把他打算演讲的论文印发给了我们。那真是篇激动人心的论！论文中奈曼利用他1939年完成的研究成果，去解决一个天文学上的难题。我知道1939年的那篇论文。几年前，当我还是个研究生的时候就看到了它，并留下了深刻的印象。论文中阐释了奈曼已经发现的一类新的分布，他称之为“散播分布”（contagious distribution）。论文中所提到的问题，开始是试着模拟土壤里昆虫幼虫的分布情形：即将排卵的母昆虫带着满肚子的卵在田野里四处飞，然后随机选取一个地点排卵，一旦排完卵，幼虫孵化出来，就从那个地点钻出地面。现在，从田野里取一个土壤样本，那么，在这个样本里发现的幼虫数量的概率分布是什么？散播分布描述了这种情形。奈曼1939年的论文，运用一系列看似简单的方程，导出散播分布。推导的过程看上去明显而自然。显然，看完论文之后，读者会觉得除了奈曼的做法之外，再没有更好的推导方法了。但这只是在读了奈曼的文章后才清楚的。自从1939年那篇论文发表之后，人们发现奈曼的散播分布适用于相当多的领域，如医学研究、冶金术、气象学、毒物学，以及解决宇宙中星系的分布问题（就像奈曼在辉瑞的那个研讨会介绍论文所描述的）。演讲结束，奈曼坐下来听专题小组的讨论。讨论组的其他成员都是著名的统计学家。由于太忙，不能提前阅读他的论文，他们把辉瑞的研讨会作为对奈曼荣誉的肯定。他们的“讨论”包括对奈曼的学术生涯和以往建树的评论。我作为最后一记得

的替补者加入到这个专题组中，并且被告之不能提及我先前和奈曼相处的经历（其实我根本没有这种经历）。因此，我就应他的本意，直接评论奈曼那天演讲的东西。我提到在几年前是如何发现了1939年的那篇论文，以及为了准备参加座谈会，重读了论文。我尽一切所能描述论文的内容，谈到奈曼创立的分布参数其意义的巧妙方式时，我显出极大的兴趣。奈曼对我的评论显得非常高兴。之后，我们俩热烈地讨论了散播分布以及它的用法。几周以后，我收到寄来的一个大包裹，是一本加州大学出版社（The University of California Press）出版的《J?奈曼早期统计论文选》（A Selection of Early Statistical Papers of J. Neyman），在书的内封有一行题词："致大卫?萨乐斯伯格（David Salsburg）博士，衷心感谢他在1974年4月30日对我演讲的有趣讲评。J?奈曼。"

我把这本书视为珍宝，一是由于奈曼的题字，二是因为书中那一系列精美绝伦、文笔极佳的论文。从那时起，我有机会与奈曼的很多学生和同事交谈，得知这个我在1974年碰到的、友善的、风趣的、有感召力的人，也是他们深知并崇敬的人。

# 第11章 假设检验

在他们一开始合作的时候，E?皮尔逊就问耶日?奈曼，在检验一组数据是否为正态分布时，如果没能得到一个显著性的P值，那么怎样才能看这组数据是正态分布的呢？他们的合作从这个问题开始，然而，E?皮尔逊最初的这个问题，却打开了一扇通往更广阔领域的大门。在显著性检验中，如果得到的是一个不显著的结果，那么它的涵义是什么呢？如果我们找不到拒绝一个假设的证据，我们能做结论说这个假设为真吗？

费歇尔其实已经间接地回答了这个问题。费歇尔把比较大的P值（代表没有找到显著性证据）解释为：根据该组数据不能做出充分的判断。依据费歇尔的解释，我们绝对不会得出这样的推理，即没有找到显著性的证据，就意味着待检验的假设为真。这里引用费歇尔的原话：

相信一个假设已经被证明是真的，仅仅是由于该假设与已知的事实没有发生相互矛盾，这种逻辑上的误解，在统计推断上是缺乏坚实根基的，在其它类型的科学推理中也是如此。当显著性检验被准确使用时，只要显著性检验与数据相矛盾，这个显著性检验就能够拒绝或否定这些假设，但该显著性检验永远不能确认这些假设一定是真的，……如果显著性检验真的被人们理解到这种程度，那么就说明显著性检验的道理已被人们认识清楚了……

在这之前，K?皮尔逊常常利用他的卡方拟合优度检验来“证明”某些数据符合某些特定的分布。在费歇尔把更精确的方法引入到数理统计之后，K?皮尔逊的方法就不再为人接受了。但问题仍然存在。为了知道应该估计哪些参数，为了确定这些参数与所研究的科学问题之间有何关系，我们必须假设该数据符合某一特定的分布。统计学家们常常会利用显著性检验来证明数据符合何种分布。

在他们的通信往来中，E?皮尔逊与奈曼经常探讨一些由显著性检验中浮现出来的悖论，不假思索地使用一项显著性检验，可能会把一个显然为真的假设拒绝掉。但费歇尔从未陷入这种尴尬，因为对他来说，显著性检验怎样被误用他是非常清楚的。奈曼问：用什么标准来判断一项显著性检验的应用是正确的还是不正确的呢？逐渐地，随着E?皮尔逊与奈曼的书信往来，加上奈曼在暑期到英国的几次访问以及E?皮尔逊的几次波兰之旅，假设检验的基本思想已经浮出水面 。

现在，在所有基础统计学的教科书中，都可以发现一个简化的奈曼－皮尔逊假设检验理论公式。该公式结构简单，我发现大部分的大学一年级学生很容易看懂，因为已经被编纂整理过，所以这个公式很精确，也很有说服力。假设检验理论必须这样来写，当然这也是教科书所需要的写法，也只能这样来写。这种直接表述假设检验的方法已经被一些政府和社会机构所接受，如美国食品及药品管理局、美国环保署，许多医学院在给将来做医学研究的人授课时，采用的也是这一套方法。此外，这种方法也逐渐地被应用到了司法界，当法院处理某些需要鉴别的歧视性案子时，就经常会用到这种方法。

当由奈曼和E?皮尔逊创建起来的这种理论以奈曼的这种直接而简化的方式来讲授时，由于集中于公式中有错误的一面，从而曲解了他的发现。奈曼的主要发现是，除非至少有两个可能的假设，否则显著性检验根本就没有意义。也就是说，你不可能检验一组数据是否服从正态分布，除非你认为该组数据也可能会被其它的一些分布或分布集来拟合。这些备择假设的选择，决定了显著性检验的执行方式。当一个备择假设为真时，该备择假设被接受的概率奈曼称之为该检验的效力（power）。在数学里，要清晰阐述一种思想，通常要给某一特定的概念赋予清楚明确的定义。为了区别被用来计算费歇尔P值的假设与其它可能的一个或多个假设，奈曼和E?皮尔逊把被检验的假设称为“零假设”（null hypothesis），称其它可能的假设为“备择假设”（alternative hypothesis）。在他们的理论公式中，计算P值是为了检验零假设，而检验的效力则是指在备择假设为真的条件下P值的表现效果。

奈曼由此得出两个结论。第一个结论是，检验的效力是用来测量一个检验方法好坏的指标，两种检验方法中效力较强的方法就是较好的方法；第二个结论是，备择假设不能太多。统计分析师不能这样来表述，某一组数据来自于一个正态分布（零假设），或者它来自于任何其它可能的分布。这种备择假设集涵盖的范围太广了，没有哪种检验方法会有那么强的效力能处理所有可能的备择假设。

在1956年，芝加哥大学的L?J?萨维奇与拉杰?拉克?巴哈杜尔（Raj Raghu Bahadur）证明，对于一个零假设未通过的情形，并不一定要求有很多的备择假设。他们构建了一个相对较小的备择假设集，除此之外的所有检验的效力均为零。在20世纪50年代，奈曼就发展出了有限制的假设检验的想法，其中的备择假设集被定义得非常狭窄。他证明得出了这样的结论：这种检验方法比那些处理较多备择假设的检验方法效力更强。

在很多情况下，假设检验的目的是用来推翻零假设的，而这个零假设就好比我们所要攻击的稻草人。举例来说，当我们比较两种药的临床效果时，待检验的零假设是两种药的效果一样。但是，如果真是如此，研究工作就永远不必进行了。所以，“两种处理的效果相同”这一零假设，就是我们所要攻击的稻草人，应该被我们研究的结果来推翻。因此，根据奈曼的思想，该项研究的设计必须使最终数据有最大的检验效力，这样才能推倒这个稻草人，即表明这两种药的效果有多大的不同。

什么是概率？

遗憾的是，为了对具有内部一致性的假设检验设计出一种数学方法，奈曼必须处理一个已被费歇尔扫到地毯下的问题。这是一直困扰假设检验的一个问题，尽管奈曼的纯数学解非常简洁巧妙。这也是统计方法应用到一般的科学领域中通常会碰到的问题。从更一般的意义讲，这个问题可以这样来概括：在现实生活中，概率的意义是什么？

统计学的数学公式可用来计算概率。而这些计算出来的概率可使我们应用统计方法解决科学中的问题。就所用到的数学而言，概率的定义很明确。但这种抽象的概念怎样和现实相联系呢？当科学家试图决定什么为真、什么不为真时，他该如何解释统计分析的概率陈述呢？在本书的最后一章，我将讨论这个一般性的问题，并分析长久以来设法解答这些问题所做的努力。但现在，我们将分析促使奈曼找到他的答案的特殊情况。

前面我们谈过，费歇尔利用显著性检验产生了一个他称为P值的数字。这是一个计算出来的概率，是在零假设为真假定下，与观测数据有关联的一个概率。例如，假定我们要检验一种新药，对做过乳房切除手术的妇女来说，这种药可以防止乳腺癌的复发。我们把这种药的效果与一种安慰剂作比较。此时的零假设（那个稻草人）就是，该新药不比安慰剂好。现在，假定5年之后，用安慰剂的妇女有一半乳腺癌复发，但用新药的完全没有复发，这样能证明新药“有效”吗？答案当然得看这个50%代表多少病人。

如果在这项研究中，两组各仅有4名病人，也就是总共有8名病人，而其中2人在5年后复发。假定我们任选一个8人团体，把其中两人做上标记，接着把人随机分成两组，每组4人，那么做标记的两个被分在同一组的概率大约是0.30.因此，如果每组只有4名妇女，“所有复发的妇女都落在安慰剂组”是不显著的。如果该项研究中每一组包含500名妇女，且

乳腺癌复发的所有250名妇女都落在安慰剂姐，这是极度不可能的，除非新药真的有效。如果新药并不比安慰剂有效，这250名妇女都落在同一组的概率就是P值，计算出来的结果将小于0.0001.

P值是一个概率，它就是这样被计算出来的。既然P值被用来表明一个假设（P值就是在该假设下计算出来的）为假的概率，那它的实际意义又是什么呢？答案是，P值是在极可能为假的条件下，与观测值相关联的一个理论概率。P值与现实没什么联系，它是一种对似是而非问题的间接测量。它不是我们错误理解的新药有效的概率，它也不是出现任何一种类型误差的概率。但是，为了决定哪一种检验方法比别的检验方法更好，奈曼必须想出一种办法把假设检验放进一个架构里，使得与根据检验所做出的决策相联系的概率能够计算出来的。因此，他需要将假设检验的P值与现实生活联系起来。

概率的频数定义

1872年，英国哲学家约翰?维恩（John Venn）提出了一个数学概率的公式。这个公式使得概率在现实生活中有了含义。他把一个重要的概率定理转了一个方向，这个定理就是大数定律（law of large numbers）。大数定律指出，如果某事件有给定的概率（比如掷一个骰子，得到六点这一事件的概率是六分之一），而且如果我们重复地进行相同的试验时，该事件发生的次数的比率就会越来越接近这个概率值。

维恩指出，与一个给定事件相联系的概率，是该事件从长期来看所发生的次数的比率。按照维恩的意见，概率的数学理论并没有隐含大数定律，反而是大数定律隐含了概率的思想。这就是以频数为基础对概率的定义。1921年，约翰?梅纳德?凯恩斯（John Maynard Keynes ）推翻了这种定义方式，认为它不是一种有用的或有意义的解释，并指出这种定义具有根本性的矛盾，因而无法在许多要求计算概率的情况不应用概率的频数定义。

在用正规的数学方法来构造假设检验时，奈曼又重新回到了维恩的概率的频数定义上。奈曼利用这个定义来证明他在假设检验中对P值解释的合理性。在奈曼－皮尔逊的公式中，科学家设定一个固定的值，比如0.05，之后，当显著性检验的P值小于或等于0.05时，就拒绝零假设。按照这种理解，从长期来看，该科学家会正好有5%的机会拒绝一个正确的零假设。假设检验当前就是这样来讲授的，奈曼所采用的频数方法被得到强调。我们太容易把奈曼－皮尔逊的假设检验公式看作是概率的频

数方法的内容，因而太容易忽略奈曼所提的观点中更重要的见解，即为了检验零假设这个“稻草人”，必须要有一组定义明确的备择假设。

费歇尔误解了奈曼的见解。他把注意力集中到了显著性水平的定义上，但却忽略了检验效力和需要定义一组备择假设这些重要的思想。在批评奈曼时费歇尔写到：

奈曼认为他自己修正并改善了我早期所做的关于显著性检验的工作，结果“改进了自然知识”，不过实际上他只是用技术性与商业性的形式，也就是大家所熟知的接收程序，重新解释了这些检验方法罢了。现在，在当代世界里，这种接收程序变得十分重要。例如，当英国海军总部接到某工程公司的大批材料时，我认为要安排很仔细的检查与检验，以降低残次品被接收的频率，……不过在我看来，这种管理运作与透过物理或生物实验的科学发现工作相比，它们之间的逻辑上有很大的差别，所以拿这两者做类比是没有多大帮助的，而把它们当成是同一回事，更是一种决定性的误导。

尽管存在对奈曼基本观点的这些扭曲，假设检验还是成为科学研究中应用得最多的统计工具。奈曼提出的精巧数学构思，在科学的很多领域中都占有一席之地，变成了一种固定的观念。大部分的科学期刊都要求论文的作者在做数据分析时要采用假设检验方法，甚至连科学期刊之外的领域也开始这么做。美国、加拿大与欧洲的药物管理机构，纷纷把假设检验方法的使用列为对药品检查的强制性要求，就连法庭允许原告用这种方法证明自己受到就业歧视。假设检验已经渗透到统计学的所有分支学科中。

奈曼－皮尔逊的理论攀升到统计学的巅峰地位，一路上也不是没有挑战的。费歇尔从一开始就攻击它，而且在他有生之年一直在攻击这个理论。1955年，费歇尔在《皇家统计学会期刊》上发表一篇文章，题目是“统计方法与科学归纳”，而在他的最后一本书《统计方法与科学推论》（Statistical Methods and Scientific Inference）里，更进一步详述了他的看法。在20世纪60年代晚期，不久之后就出任《生物统计》期刊主编的大卫?考克斯（David Cox），发表了一篇分析清晰的文章，分析了假设检验在科学中的实际用途，同时也证明了奈曼的关于频数的解释不符合实际状况。在20世纪80年代，W?爱德华兹?戴明（W. Edwards Deming）攻击了假设检验的整个思想，认为假设检验的整个思想都是荒谬的（第24章还会再提到戴明对统计学的影响）。年复一年，在统计学文献中一直有相关文章发表，指出在教科书中已成定格的奈曼－皮尔

逊理论中发现了新的毛病。

不过，在奈曼－皮尔逊假设检验理论的神圣化过程中，奈曼本人并没有参与。早在1935年，他在《法国数学学会会刊》《bulletin de la Société Mathématique de France》上就用法文发表过一篇文章，对是否能找到最佳的假设检验方法提出严厉的质疑。在他后来的文章里，奈曼很少直接使用假设检验方法，他的统计方法通常是由理论原则导出概率分布，然后再由数据来估计参数。

其他一些人则捡取藏在奈曼－皮尔逊理论背后的观点来进一步发展。在第二次世界大战期间，亚伯拉罕?沃尔德扩展了奈曼利用维恩关于频数的定义，发展成了一个叫统计决策理论（statistical decision theory）的领域。埃里希?莱曼（Erich Lehmann）给出了用来判断一个好的假设检验可供选择的标准，后来在1959年，他还写了一本有关假设检验问题的权威性的教科书，这本书至今仍然是该领域对奈曼－皮尔逊假设检验理论描述得最完整的一部著作。

就在希特勒入侵波兰，将邪恶之幕笼罩欧洲大陆之前，奈曼就到了美国，并在加州大学的伯克利分校开始创建统计系。在那里他一直工作到1981年去世，这期间，他把该系创建成全世界最重要的学术性统计学系之一。他把一些统计学界赫赫有名的人物引入该系，同时也提拔了一些默默无闻的人，这些人正致力取得卓越的成就。例如，大卫?布莱克韦尔（David Blackwell）原来只是只身孤单地在霍华德大学（Howard University）工作，没有数理统计同行与他来往。由于他的种族原因，他一直没能在“白人”学校谋得一职，尽管他很有潜能。奈曼把他请到了伯克利。此外，奈曼还招了一位出身法国农民家庭的研究生吕西安?勒卡姆（Lucien Lecam），他后来成为世界领先的概率学家。

奈曼总是非常和善地对待他的学生和同事。他们常常津津乐道的是系里每天下午茶歇的欢乐时光，这是由奈曼主持的他与职员亲近接触的一个重要场合。他总是亲切地鼓励学生和同事谈谈自己最新的研究成果，同时很和蔼地提出他自己的思路和见解，给出评论，加入大家的讨论。他常常在下午茶歇即将结束时举起茶杯说“为尊敬的女士们！”他特别关照女士，鼓励她们在学术生涯上不断进步。在他的女弟子当中，伊丽莎白?斯科特（Elizabeth Scott）博士是较为杰出的，她与奈曼一起做研究，共同发表论文，范围从天文学到致癌物研究，甚至动物学。还有伊夫琳?菲克斯（Evelyn Fix）博士，她在流行病学的研究上有很重要的贡献。

直到费歇尔于1962年去世，奈曼一直受到这位天才的尖刻批评。奈曼每做一件事都会遭到费歇尔的批评。如果奈曼成功地证明出了费歇尔某项非常难解的叙述，费歇尔就说奈曼误解了他写的东西；要是奈曼扩充了费歇尔的某个观点，费歇尔就批评奈曼说他把好端端的理论用错了地方。对比，不论是付诸笔端，还是在私人场合，奈曼从不回应（如果我们相信奈曼同事的说法）。

在奈曼去世前的一次访谈中，奈曼说了一件发生在20世纪50年代的往事。当时他准备在一次国际研讨会上公开发开一篇用法语写的论文。当他步上讲台时，意识到费歇尔也坐在听众席上。在演讲论文时，他知道一场激辩难免，于是开始武装自己，他预计费歇尔会抓住论文里某个无关紧要的小地方，将论文和他本人攻击得体无完肤。奈曼讲完之后，等待听众提问，结果只有几个问题。费歇尔相当平和，一言未发。后来奈曼才知道，费歇尔不会讲法语。

# 第12章 置信诡计

当20世纪80年代出现了艾滋病（AIDS）这种传染病时，有若干问题需要回答。一旦传染源HIV（human immunodeficiency virus，即人体免疫缺损病毒）确定了，卫生官员需要知道有多少人受到感染，以便安排需要的资源来应付这种传染病。幸运的是，在此之前的20至30年所开发出来的流行病学 数学模型，在这里可派上用场。

从传染病的现代科学观点来看，某些个体病人接触到传染源，其中有些人会被传染，而在经过一段所谓的“潜伏期”之后，那些被传染的人会显现该疾病的症状。一旦被传染，这个人就会成为其他还没有被传染人的潜在传染源。我们没有办法预测谁会与传染源接触，谁会被传染，或谁会传染他人。我们所能做的，只是处理相关的概率分布，并估计这些分布的参数。

参数之一是平均潜伏期，也就是从被传染到症状产生的平均时间。就艾滋病这种传染病来说，平均潜伏期对卫生官员是特别重要的参数。他们没有办法知道究竟有多少人被传染，又有多少人最终会得上这种疾病，但如果能知道平均潜伏期，他们就能根据已经患有这种疾病的人数，估计出受感染的人数。不仅如此，由于艾滋病传染模式的不寻常特征，卫生官员拥有一组患者，并知道这组患者感染的时间和他们的发病时间。有一个小的血友病患者群体由于使用了被污染的血液制剂而感染上HIV，他们提供的数据可以用来估计平均潜伏期这一参数。

这个估计值的准确性如何？流行病学家可以说，他们使用的是费歇尔意义上的最佳估计量。因为他们所得的估计值是一致的，又是最有效的。他们甚至还可以修正可能的偏差，并宣称他们的估计值是无偏的。但是，如果我们在前面章节里指出的，我们没有办法知道某一个具体的估计是否正确。

如果我们不能够说某个估计值是绝对准确的，那么我们还有没有办法可以说这个估计值与参数的真值之间有多接近呢？这个问题的答案在于使用区间估计（interval estimate）。点估计（point estimate）是一个单一的数字。例如，我们可能利用从血友病研究那里得到的数据，估计出平均潜伏期是5.7年。而一个区间估计会这样表述：平均潜伏期在3.7年至12.4年之间。在很多情况下，有区间估计的数字就够了，因为所需要的公共政策对区间估计的两端边界值来说是一样的。但有些时候，区间估

计值显得太宽了，对最小的边界值和最大的边界值需要制定不同的公共政策。根据一个很宽的区间估计值所能得出的结论是，利用已有的信息不足以做出充分的决策，应寻求更多的信息，可以通过扩大调查的范围或进行一系列其它的实验来得到。

举例来说，如果艾滋病的平均潜伏期长达12.4年，则感艾滋病毒的人当中约有五分之一的人在感染之后要存活20年以上；如果平均潜伏期是3.7年，那么几乎每一个被感染的人在20年内都会发病。这两个结果相差太大。没有任何一种最佳的公共政策可以兼顾，因此需要更多的信息。

在20世纪80年代末期，美国国家科学院（National Academy of Science）如今国内一批顶尖的科学家组成一个委员会，讨论臭氧层破洞的问题。臭氧层可保护人类不受紫外线辐射的伤害，但由于人类使用的喷雾剂中含氟氯碳化物，可能破坏外层空间的臭氧层。这个委员会（主席为约翰?图基（John Tukey），是本书第22章讨论的主角）不做是或否的二分法回答，而是决定以概率分布的形式建立氟氯碳化物对臭氧层的影响模型。于是，他们计算出了臭氧层每年平均变化的区间估计值。虽然使用的数据量不是很多，但他们发现，该估计区间的下边界值暗示，每年臭氧层将以一个较大的幅度减少，而这将使人类的生命在50年内受到严重的威胁。

区间估计现在已经普及到几乎所有的统计分析领域。当一项民意调查指出44%的一般民众认为总统干得不错时，通常会加上一个附注，说明这个数字“具有正负3个百分点的误差”。上述民意调查结果的意思是，44%被调查的民众认为总统干得不错。由于这是个随机的调查，所求的参数是全国所有的民众中认为总统干得不错的人数的百分比。由于样本的容量较小，因此一个合理的猜测是，总体的参数值应落在41%（44%－3%）与47%（44%+3%）之间。

怎样计算区间估计值？怎样解释一个敬意估计值的涵义？我们能对一个区间估计值做出相应的概率表述吗？我们有多大的把握确信总体参数的真值会落在所估计的区间里？

奈曼的解

1934年，耶日?奈曼在皇家统计学会做了一个演讲，题目是“论代表性方法的两个不同方面”（On the Two Different Aspects of the Pepersentative

Method）。他的论文是关于抽样调查分析的。正如奈曼作品的一贯风格，这篇文章非常优美，导出了形式简单具直观易懂的数学表达式（当然是经过奈曼的推导之后才会如此）。但全文最重要的部分却在附录里，奈曼在这个附录中提出了一个很直接的方法，用来创建区间估计，并确定所得的区间估计值有多准确。奈曼称这个新的方法为“置信区间”（confidence intervals），而把置信区间的两端称为“置信界限”（confidence bounds）。

G?M?鲍利（G. M. Bowley）教授是大会的主席，起身致谢辞。他先用几段话讨论了论文的主要部分。接着就说到了附录：

我不太确定是否应该要求给出一个说明，或者直接提出质疑。论文的字里行间暗示，论文很难读懂，而我可能是被这个暗示误导的人之一（在这段话之后，他举出一个例子，表明他完全理解了奈曼提出的方法）。我只能说，从我一看到这篇论文开始，我就很认真地读它，而且昨天我还很仔细地读了奈曼博士对这篇论文的补充资料。我指的是奈曼博士的置信界限。我不太有把握地说，这里的“置信”是不是一个“置信诡计”。

鲍利接着举了一个例子说明奈曼的置信区间，然后继续说道：

这个方法真的会将我们引向深入吗？我们会比艾萨克?托德亨特（Isaac Todhunter，一位19世纪末的概率学家）知道的更多吗？它会让我们超越K?皮尔逊和埃奇沃思（Edgeworth，数理统计发展早期的先驱之一）吗？它真的会引领我们到我们所需要的地方去吗？就是说我们所从中抽取样本的总体其比重会正好落在这些界限内吗？我看并不见得，……我不知道我是否已把我的想法表达清楚了，……自从我看到这个方法，我就觉得它是个难题。其理论陈述没有说服力，除非有人能说服我，否则我还是怀疑它的有效性。

鲍利对置信区间这个新方法的疑惑，是自从置信界限的概念被提出来以后大家对它的普遍迷惑之一。显然，奈曼在推导其结果过程中所用的四行优美的微积分式子，在抽象的概率数学理论上是正确的。它也确实能算出一个概率值。但这个概率值究竟代表什么则并不清楚。数据是观测得来的，参数是固定的值（尽管是未知的），因此参数取某个特定值的概率只有两个结果，或者是100%（如果它就是那个值），或者是0（如果它根本不是那个值）。然而，一个95%的置信区间涉及的是95%的概率。这个概率指的是什么？奈曼在此绕过了这个问题，把他的创造称为置信区间，回避使用概率这个词。但是鲍利及其他同行一眼就看穿了这

个手法。

费歇尔也在批判者之中，不过他没有抓住这个要点。他所讨论的内容空洞又含混，而且根本不是奈曼论文里的内容。因为费歇尔根本没有完全弄清楚区间估计值的计算过程。在他的评论里，他所指的是“信念概率”（fiducial probability），而奈曼的论文里并没有这个词汇。长久以来，费歇尔一直试图解决这个问题——怎样确定与一个参数的区间估计相关联的不确定度？费歇尔从一个很复杂的角度来解决这个问题，有点像他的似然函数。不过他很快就证明，用这种方式研究这个公式并不符合概率分布的要求。费歇尔称这个函数为“信念分布”（fiducial distribution），但他后来又违反了他自己的思路，使用了其他人在处理适当概率分布时可能会用到的相同数学方法。费歇尔所希望的结果，是从观测数据中得到参数的一组合理的值。

这也正是奈曼所得的结果，而且如果该参数为正态分布的平均数时，两个方法会得到相同的答案。据此费歇尔认为奈曼窃取了他的偏偏分布的思想，只是换了个名字而已。费歇尔对他的信念分布的研究从来没有取得进一步的发展，因为他的方法在遇到更复杂的参数（比如标准差）时就不管用了。奈曼的方法对处理任何类型的参数都是有效的。费歇尔似乎从未理解这两种方法之间的差异，直到死前他还坚持认为，奈曼的置信区间最多只是他的信念区间（fiducial intervals）概念的推广。他坚信，在碰到足够复杂的问题时，奈曼的显然是推广的方法也不会奏效——就像他自己的信念区间方法一样。

概率与置信水平

不管碰到的问题有多复杂，奈曼的方法没有失败，这也是该方法在统计分析中得到广泛应用的原因之一。奈曼置信区间中的真正问题，倒不是费歇尔所提出的那个，而是鲍利在一开始讨论时就点出来的问题，即这个方法中的概率到底指的是什么？奈曼的回答又回到了现实生活中概率的频数定义上。正如他在这篇论文里所说的（他在稍后的另一篇探讨置信区间的论文里，对这一点做了更清楚的解释），不应该从每一个结论的角度看待置信区间，而应该其视为一个过程。从长期来看，对于一直计算95%的置信区间的统计学家来说，他们将发现，在总次数中，参数的真值将有95%的机会落在所计算的区间内。请注意，对奈曼来说，与置信区间相联系的概率并不是我们“答对”的概率，而是统计学家使用某种方法从长期来看做出正确陈述的频率。这个数字与当前的估计值有多“准确”根本没有任何关系。

尽管奈曼定义这个概念时非常仔细，尽管许多像鲍利这样的统计学家也都非常小心，力图保持对概率概念的清晰理解并使其不被误用，但在科学领域中对置信区间的普遍应用却导致了许多草率的思维。举例来说，有人使用95%的置信区间来表示他有“95%的把握”保证参数的真值会落在这个区间里，这是很普遍的。我们在13章会碰到：L?J?萨维奇和布鲁诺?德费奈蒂（Bruno de Finetti），并介绍他们对个人概率的研究，他们的研究结果证明了使用上述陈述的合理性。但是，计算某人对某一件事的把握程度，与计算一个置信区间完全是两回事。统计文献里有很多文章都谈到，根据一组相同的数据，以萨维奇和德费奈蒂的方法所推导出的参数范围，和以奈曼的方法为基础推导出的置信界限，两者之间是截然不同的。

尽管在奈曼的方法中人们对概率的涵义仍存有疑问，但是奈曼的置信界限已经成为计算区间估计值的标准方法。许多文学家计算90%或95%的置信界限，而且看上去好像他们有把握认为，该区间包含了参数的真值。

时至今日，已无人再谈论或在写作中涉及费歇尔的“信念分布”的话题了。该思想已随费歇尔的去世而消失。费歇尔竭力让他的思想能发挥作用，他做了大量的相当聪明而且非常重要的研究工作，其中有些研究成果已成为当今的主流，而其它部分则仍停留在费歇尔搁笔时的不成熟状态。

在费歇尔的研究过程中，他曾有好几次差点儿就建立一门统计学业的分支学科，也就是他所称的“逆概率”（inverse probability），但每次他都半途而废。逆概率的思想起源于18世纪的一位业余数学家雷韦朗?托马斯?贝叶斯（Reverend Thomas Bayes），贝叶斯与很多同时代的顶尖科学家都有密切的书信往来，并经常提出一些很复杂的数学问题给他们。有一天，他随意玩弄一些概率的标准数学公式，用简单的代数把其中两个式子结合在一起，竟发现一些令他很惊讶的结果。

下一章，我们来谈谈贝叶斯异论（Bayesian heresy），并且看看为什么费歇尔拒绝使用这种逆概率。

# 第13章 贝叶斯异论

从8世纪的早期，威尼斯共和国是地中海一带的一个主要的强权国家。在其政权鼎盛时期，威尼斯控制了大部分的亚得里亚海岸，以及克里特岛和赛浦路斯岛，同时还垄断了东方通往欧洲的商业贸易路线。威尼斯共和国由一群贵族家族所统治，这些家族之间保持着某种民主的程序。整个国家名义上的领袖是总督，从公元697年该共和国成立起，到1797年被奥地利吞并，总共有150余任总督，有的任期很短，只有1年或不到1年，也有的任期长达34年。在在的总督去世之后，该共和国会遵守一项很复杂的选举程序，他们先从贵族家族的长者当中，以抽签的方式选出一小群元老，这些被选出的元老还会再挑选一些人加入到他们之中，之后再从这一扩大的元老群中以抽签方式选出一小群人。这样的程序进行几次之后，会选出一群最后的总督候选人，总督就在这群人当中产生。

在威尼斯共和国历史的早期，每阶段的抽签都要准备一批大小相同的蜡球，有的蜡球里什么都没有，有的蜡球里面却有一张小纸条，上面写着“元老”二字。到了17世纪，最后几个阶段用的道具是大小完全相同的金球与银球。公元1268年，当多杰?拉伊涅里?泽诺（Doge Rainieri Zeno）总督去世时，在第二阶段有30位元老，于是准备了30个蜡球，其中9个蜡球内藏有“元老”纸条。一个小孩被带过来，他从装有蜡球的篮子中取出一个蜡球，交给第一位元老候选人，这位元老候选人就打开蜡球，看看自己是否能够成为下一阶段的元老候选人。接着，小孩从篮子中取出第二个蜡球，交给第二位元老候选人，第二位再打开蜡球，以此类推。

在小孩选出第一个蜡球前，候选人群中的每个成员被选为下个阶段元老的概率是9/30。如果第一个蜡球是空的，剩下的候选人中每个人有9/29的概率成为下坠估摸元老。但如果第一个蜡球里有纸条，则其余人被选中的机会就剩下8/29。一旦第二个蜡球被选定且被打开，则下一个人被选中成为元老的概率同样会减少或增加，是减少还是增加取决于前次的抽球结果。这样继续抽下去，直到所有的9个纸条都被抽出为止。而在这时，剩下的候选人下一阶段成为元老的概率就降为零。

这是条件概率的一个例子。某一特定候选人被选为下一阶段元老的概率，取决于在他的选择之前被选出的蜡球。J?M?凯恩斯曾指出，所有的

概率都是条件概率。用凯恩斯所举的一个例子：从他的图书室的书架上随机地选择一本书，而选中的书是精装本的概率，也是一种条件概率，其条件取决于他的图书室里究竟有多少书，以及他怎样“随机”地选取。一个病人患小细胞肺癌的概率，是以该病人的吸烟史为条件的。对一个控制实验，检验没有处理效果这一零假设所计算出来的P值，是以该实验的设计为条件的。条件概率的重要方面是，某些已知事件（例如在彩票发行过程中，某一组特定数字能赢）的概率，会随前提条件的不同而不同。

在18世纪，为处理条件概率而导出的公式都是根据以下的思想做出的，即条件事件要发生在所研究的事件之前。但是到了18世纪后期，R?T?贝叶斯在摆弄条件概率的公式时，忽然有个惊人的发现，这些公式都是内部对称的！

假设有两个事件在一段时期内发生，就像先洗牌，再发出5张扑克牌。我们称这两个事件分别为“前事件”（the events before）和“后事件”（the events after）。以“前事件”为条件讨论“后事件”的概率是有意义的。如果牌没有洗好，当然会影响玩家得到一对A的概率。贝叶斯发现，我们也可以“后事件”为条件计算“前事件”发生的概率。这是没有道理的。就像玩家已经拿到一对A之后，再来确定整副牌里有4张A的概率。或是已知一个病人已患了肺癌，再来计算他是吸烟者的概率。或者是已经知道了有个叫C?A?史密斯的人是唯一得到大奖的人，然后再计算州立彩票游戏公平不公平的概率。

贝叶斯把这些计算结果搁置起来，没有发表。在他死后，这些论文才被发现，而后才被发表出来。从那里起，贝叶斯定理　就困扰着许多统计分析数学家。绝对不是毫无道理，贝叶斯将条件概率倒转过来反倒很有意义。当流行病学家试图想找出某种罕见医学病状的可能原因时，例如雷氏症候群（Reye’s　syndrome），他们通常是利用病例控制研究方法（case-control　study），在这种研究中，他们首先搜集一组患有该病症的病人，然后拿去与控制组的病人做比较，控制组的病人没有患这种疾病，但在其他方面与患有这种疾病的病人类似。于是，流行病学家在已知控制组病人已患有该疾病的条件下，计算某些先前治疗或先前条件导致该病的概率。吸烟对心脏病和肺癌都有影响，就是这样首次被发现的。镇静剂对新生儿畸形的影响，也是从这种病例控制研究中发现的。

直接应用贝叶斯定理，可以把条件概率反转过来，比这更为重要的，是使用贝叶斯定理估计分布的参数。有一种建议，可以把一项分布的参数

本身看作是随机的，然后计算与这些参数相关的概率。例如，我们可能想要比较两种癌症治疗方法，并希望得到结论说“我们有95%的把握认为使用治疗方法A会比使用治疗方法B的5年期存活率高”。我们只要应用贝叶斯定理一两次就可以解决这个问题。

关于“逆概率”的问题

有很多年，以这种方式使用贝叶斯定理被认为是一种不适当的作法。当用于参数时，关于概率代表什么涵义有很多质疑。毕竟皮尔逊革命（Pearsonian revolution）的整个基础在于，科学的测量结果本身不再是我们所感兴趣的问题，相反，正如K?皮尔逊所指出的那样，我们所感兴趣的是这些测量结果的概率分布，而科学的调查研究的目的就是要估计出控制这些分布的那些参数值（固定的但却是未知的）。所以，如果这些参数被视为是随机的（而且以观测的测量结果为条件），那么这种方法就不再有这样清楚的意义了。

在20世纪的早些年，统计学家非常谨慎，避免使用人们所说的“逆概率”。有一次在皇家统计学会上，对费歇尔的一篇早期论文进行讨论时，就有人质疑他使用了逆概率，他坚定地为自己辩护，否认这项可怕的指控。在第一篇关于置信区间的论文里，奈曼似乎使用了逆概率的概念，但只是作为一个数学方法，用来得到一个计算结果，而在他的第二篇论文里，他证明不了不用贝叶斯定理也能得到相同的结果。到了20世纪60年代，为种方法的潜在力量与用途已开始吸引越来越多的研究者跟踪研究，这个贝叶斯异论变得越来越受尊重了。到了20世纪末，它已经达到了如此高的接受水平，如今在一些期刊像《统计年报》（Annals of Statistics）和《生物统计》上，几乎半数以上的文章现在都使用贝叶斯方法。不过，贝叶斯方法的应用仍然会经常遭到质疑，尤其是在医学领域。

在解释贝叶斯异论时碰到的一个困难是，目前有好几种不同的分析方法，而这些方法的应用又至少有两种完全不同的哲学基础。长期以来，看上去好像完全不同的思想却经常贴着相同的标签——贝叶斯。后面我将说明贝叶斯异论的两个种理论：贝叶斯层次模型（Bayesian hierarchal model）和个人概率（personal probability）。

贝叶斯层次模型

20世纪70年代早期，由于弗雷德里克?莫斯特勒（Frederidck Mosteller）

和大卫?华莱士（David　Wallace）早期的工作和贡献，原文分析的统计方法有了很大的进展，他们俩人曾运用统计方法来判定《联邦主义论文集》（Federalist）中一些匿名文章的作者。自1787年，在纽约州带头鼓动通过新的美国宪法期间，詹姆士?麦迪逊（James　Madison）、亚力山大?汉密尔顿（Alexander　Hamilton）和约翰?杰伊（John　Jay）写了大约70篇文章，支持通过宪法。但这些文章都是匿名发表的。19世纪初，汉密尔顿与麦迪逊两人开始确认这些两个人都声称有著作权的论文，其中有12篇文章他们都认为是自己写的 。

在用统计方法对这些署名有争议性的文章进行分析时，莫斯特勒与华莱士找出了几百个无“特定内容”的英文词汇，
如“if”、“when”、“because”、“over”、“whilst”、“as”、“and”等。这些字在句子里只有语法上的意义，本身并没有什么特定的含义，这些字的使用主要取决于作者的语言使用习惯。在这上百个没什么特定含义的字里，他们发现，大约有30个字在这两位作者的其他著作中使用频率不同。

例如，麦迪逊使用“upon”这个字的频率，是每千字平均0.23次，但汉密尔顿对这个字的使用频率很高，平均每千字高达3.24次（在12篇署名有争议的文章里，有11篇根本没有用“upon”这个字，而在剩下的那一篇文章中，平均每千字就出现1.1次）。这些平均的频率并不是描述一千字中任何特定组合。这些数值本身并不是整数，这就意味着这些频率并不是在描述任意一个观测的文字序列。这些数值其实是两位不同作者在写作时用字分布的其中一个参数的估计值。

对于某篇文章著作权的争议，所要解决的问题是：这些文章中用词的分布形态，是来自与麦迪逊相联的概率分布呢？还是来自与汉密尔顿相联的概率分布？这些分布各有各有参数，其中能够定义出各自作品的特定参数各不相同。参数值只能根据他们的论文来估计，而且这些估计可能是错的。因此，要想区分哪个分布可应用在一篇署名有争议的文章上，充满了这种不确定性。

估计这种不确定性水平的一种方法是，这两个人的分布参数的确切值，是来自于描述18世纪晚期所有北美洲有教养的人用英文写作时用字习惯的参数分布。例如，汉密尔顿每千字中用到“in”这个字24次，麦迪逊则是每千字用23次，而同时代的其他作家，使用“in”这个字的频率在每千字22至25次之间。

由于受到当时和当地一般用字分布形态的制约，每个人分布的参数是随机的，并且具有一个概率分布。这样一来，制约汉密尔顿和玫迪逊使用这些无特定含义的字的参数本身也有参数，我们可以称之为“超参数”（hyper-parameter）。根据当时和当地其他作者发表的文章来分析，我们就能估计出这些超参数。

英语语言总是随着时间和地域的变化而变化。例如在20世纪的英语文学里，使用in的频率通常是每千字少于20次，这表明从汉密尔顿和麦迪逊的时代到现在的200多年里，英语的用字型态已经稍微有所转变。我们可以把这些定义18世纪北美用字习惯参数分布的越参数，看作是它们本身也有一个相对于所有时间与空间的概率分布。因此，除了用18世纪的北美作品，我们还可以搜集其它地区和其它时期的英语文献，来估计这些超参数的参数，我们可以称这些参数为“超－超参数”（hyper-hyperparameter）。

通过重复使用贝叶斯定理，我们就能决定这些参数的分布，然后再决定这些超参数的分布。从原则上来说，我们可以用超－超－超参数求出超－超参数的分布，进而把这种层次分析引向深入，依次类推。但在我们的例子里，显然没有必要进一步分析，以免增添更多的不确定性。利用超参数与超－超参数的估计值，莫斯特勒与华莱士就能算出与下面这个陈述有关的概率：是麦迪逊还是汉密尔顿写了这篇文章。

自20世纪80年代早期以来，贝叶斯层次模型已经成功地解决了许多工程上和生物学上的难题。比如，一些数据看上去似乎是来自于两个或两个以上不同的分布，这个问题就属于这类难题。分析家可以建议，有一个未观测到的变量存在，而这个变量可以定义已知的一个观测结果究竟来自于哪个分布。这个差别标识本身是个参数。但它还有一个概率分布（含有超参数），这个概率分布可以纳入到似然函数当中来进行分析。莱尔德和韦尔的EM演算法特别适合于解决这类问题。

统计文献中对贝叶斯方法的广泛使用充满了混淆与争议。大家可以提出得出不同结果的不同方法，但却没有明确的标准来决定哪个是对的。通常，保守肖像统计学家反对使用贝叶斯定理，而贝叶斯学派的人彼此对他们模型的细节看法也不一致。这种混乱的状况亟需另一个像费歇尔这样的天才出现，找出一个统一的原则来解决这些争议。当我们进入21世纪的时候，还没有这样的天才出现。因此，相关的问题还是像在200多年前的贝叶斯时代一样，令人困惑。

## 个人概率

另外一种贝叶斯方法其基础看上去要坚实得多。这就是个人概率（personal probability）的概念。个人概率的意思自从17世纪贝努里一开始研究概率时就已经产生了。实际上，概率（probability）这个英文字创造的初衷，就是用来处理主观不确定性的。

L?J?萨维奇和布鲁诺?德费奈蒂在20世纪60年代和70年代，推导出了个人概率背后的许多数学模式。我在20世纪60年代末期曾参加一场在北卡罗来纳大学举办的统计学会议，会上萨维奇在演讲中曾阐述他的一部分想法。萨维奇认为，世界上并没有“已被证明的科学事实”这样的事情。有的只是一些陈述，而那些自认为是科学家的人对这些陈述持有很高的赞成概率。他举例说，在场听他演讲的人对“地球是圆的”这项陈述一定持有很高的认同概率，但若我们有机会对全世界的人做一次普查，则我们很可能发现在中国中部的许多农民对上述陈述持有很低的概率。讲到这里的时候，萨维奇不得不被迫停下来，因为校园晨一群学生正在会堂外游行通过。他们还高喊着口号“停止上课！罢课！罢课！停止上课！”这些学生在要求全校的学生罢课，以抗议越南战争。等到他们走远，四周又恢复平静，萨维奇才看看窗外，然后说：“看来，我们可能是认为地球是圆的人中的最后一代。

个人概率有许多不同的版本。其中一个极端是萨维奇－德费奈蒂的方法，该方法认为每个人都有其自己独特的一套概率。而另一个极端则是凯恩斯的观点，他认为概率是一种信仰程度（the degree of belief），这种信仰是一个在特定的文化环境中一个有教养的人可能期望持有的信念。按照凯恩斯的观点，一个特定文化环境中的所有人（萨维奇所说的科学家或中国中部的农民）对某一特定的陈述，会持有一个一般的概率水平。由于这个概率水平取决于文化和时间，因此从某种绝对的意义上为说，很有可能这个适当的概率水平是错的。

萨维奇和德费奈蒂则主张每个人都有自己特定的一套个人概率，他们还描述怎样运用一种叫做“标准赌博”（standard gamble）的技巧把这种人人概率求出来。为了让整个文化中的人能共享既定的一套概率，凯恩斯不得不弱化相关的数学定义，概率不再是一个精确的数字（例如67%），而是一种将想法排序的方法（例如，明天可能下雨的概率大于可能下雪的概率）。

不管个人概率的概念是如何被准确定义的，贝叶斯定理在个人概率中的

应用方式，看上去与大多数的想法相吻合。贝叶斯方法一开始是假设在一个人的头脑中有一组先验概率（a prior set of probabilities），接下来这个人经过观测或实验产生了数据，然后再拿这组数据来修正先验概率（prior probability），生成一组后验概率（a posterior set of probabilities）：

先验概率 → 数据 → 后验概率

假设这个人想确定是否所有的大乌鸦都是黑的。她首先存有一些关于“这个陈述是真的”概率的先验知识。例如，起初她可能对大乌鸦一无所知，对“所有大乌鸦都是黑的”这句话半信半疑，相信比例是50：50。数据则包括她对大乌鸦的观测。假如她看到了一只大乌鸦，而且这只大乌鸦是黑色的，她的后验概率就会增加。因此下一次她再观测大乌鸦时，她的新的先验概率（也就是上一次的后验概率）就会大于50%，如果她继续观测大乌鸦而且都是黑的，这个概率还会继续上升。

另一方面，一个人也有可能在进行观测之前就已经带着非常强的事前主见，其程度非常强，需要有很大量的数据才能改变这个事前主见。在20世纪80年代，美国宾夕法尼亚州的三里岛核电厂发生了近乎是灾难性的事故。反应炉的操作员面对一个很大的操作盘，通过上面的各种仪表和指示灯来了解反应炉的运转情况。这些指示灯当中有一些是警告灯，其中有的出过问题，以前曾经发出过假的警告。当时操作员有个事先的成见，当他们看见任何一个新的警告灯亮时，总是认为它是假的信号。结果，即使当警告灯的型态及相关的指示器都一致显示反应炉的水位过低时，他们仍然置之不理。他们的先验概率太强了，以至于新的数据也无法使后验概率产生多大的改变。

假定只有两种可能性，就像前面署名有争议的联邦主义论文的例子：它不是麦迪逊写的就是汉密尔顿写的。于是，在应用了贝叶斯定理之后，就会得到了一个先验胜率（prior odds）与后验胜率（posterior odds）之间的简单关系，这里的数据可以归纳成一种称为“贝叶斯因子”（Bayes factor）的东西。这是一种根本不用参考先验胜率来刻画数据的一种数学计算。有了这个计算工具，分析家就可以告诉读者，插入任何他想要的先验胜率，乘以计算出来的贝叶斯因子，再计算后验胜率。莫斯特勒与华莱士对12篇署名有争议的文章，每篇都是这样处理的。

此外，他们对文章里的那些无特定含义的字出现的频率，还进行了两种非贝叶斯分析。这样他们有了四种方法来判断有争议文章的作者：层次

贝叶斯模型，计算的贝叶斯因子，以及两个非贝叶斯分析方法。结果如何呢？所有12篇文章都压倒性地指向麦迪逊。实际上，如果使用计算的贝叶斯因子，那么对某几篇文章来说，读者认为是汉密尔顿写的先验胜率可能要大于100000：1才有办法让后验胜率为50：50。

样章到此结束

需要完整版

扫下面二维码

或加微信：shuyou055

免费领取

# 第27章 意向治疗法

在20世纪80年代初，英国杰出的生物统计学家雷沙尔?皮托（Rechard Peto）遇到了一个难题，当时他正在分析比较不同癌症治疗方法的临床试验结果。根据费歇尔实验设计规定，典型临床实验研究要求确定需要治疗的病人群体，并且采用随机的方法分配给病人不同的治疗实验方法。

数据的分析应该是相当直接的，用费歇尔方法，只要在不同治疗方法的组别间，比较病人的5年存活率即可。另外还可以进行更加精确的比较，就是用奥伦（Aalen）的鞅方法（martingale approach），分析从开始研究到每个病人死亡的时间，以此作为衡量治疗效果的基本标准。不论是哪种方法，分析结果的准确性取决于最初分配给病人采用治疗方法的随机选择。根据费歇尔定律，指定病人采取何种治疗方法与研究的结果是完全不相关的，假设检验的P值是可以计算出来的。

皮托的难题是所有病人的治疗方法并不是随机指定的。这些病人也是人，正饱尝病痛的折磨，而且很多人得的是绝症，因此医生沉得有责任放弃实验性的治疗，或者如果觉得对于病人来讲是最好的选择的话，至少也要进行方案的调整。盲目地照搬某种治疗方法而不考虑病人的需要和反应是不首先的。与费歇尔的实验设计要求相矛盾，在这些实验中的病人经常变换治疗方法，而对治疗方法的选择主要取决于病人的治疗效果，如果效果好可能会继续采用这种方法，一旦觉得治疗效果不理想就会改变治疗方法。

这是癌症研究中的一个典型问题。从20世纪50年代人们刚刚开始研究癌症起，这就一直是一个令人困扰的问题，直到皮托涉入此领域研究之前，通常的做法只是去分析那些坚持采用随机分配治疗方法的病人，而其他的病人不在分析的范围之内。皮托认为这会导致严重的错误。例如，假设我们正在比较两种治疗方法，一种是有效的治疗，另一种只是给病人服用安慰剂，即一种没有生物作用的药物。如果病人对治疗无反应，就会转而使用常规的治疗。服用安慰剂、没有效果就转而使用别的治疗方法的病人不能做为研究对象，只有那些继续服用安慰剂、因为某些原因有反应的病人才是研究的对象。如果在研究分析中的研究对象只有那些继续服用安慰剂并且有反应的病人，那么研究的结果必然是：安慰剂治疗方法与有效的治疗具有同样的疗效，甚至可能疗效更好。

德克萨斯州安德森医院（M. C. Anderson Hospital）的埃德蒙?吉亨（Edmund Gehan）比皮托更早发现了这个问题。他当时的办法只是提出：因为这些研究不符合费歇尔实验的条件，所以不能够作为比较不同治疗方法的有效实验，只能算是研究中通过对采用不同治疗方法病人仔细观察而取得的记录，最多只是对实验结果的一种总体描述，为以后的治疗提供了一些思路。后来，吉亨也考虑了解决这个问题的不同方法，但是他的第一个结论让人非常气馁，竭力想在一个设计和执行都不好的实验中运用统计分析方法看来是不可能的。

皮托提出了一个直截了当的解决方法：当比较不同的治疗方法的疗效时，病人采用哪种治疗方法应该是随机的，否则不可能在假设检验中计算出P值。他建议在分析过程中假定每个接受治疗的病人采用治疗方法是随机分配的，否则不可能在假设检验中计算出P值。他建议在分析过程中假定每个接受治疗的病人采用治疗方法是随机分配的，忽略研究中治疗方法的调整。如果一个病人随机采用方法A，但在研究结束前改变了方法，这个病人视为采用A方法的病人进行研究；如果病人随机采用方法A只治疗了一个星期，病人当作采用方法A来分析；如果病人随机采用A方法治疗，却根本没有吃一粒A方法的药，就采用了另外一种治疗方法，这个病人仍被视为采用方法A的病人。

乍一看这种方法是愚蠢的。人们可以假设一种情形：对一个实验治疗方法和一个标准治疗方法进行比较，病人采用的实验治疗方法一旦失败就会转而使用标准方法。如果实验治疗方法是无用的，那么，所有的或者大多数被随机指定使用实验治疗方法的病人就会转而使用标准方法，分析将会发现这两种治疗方法效果是一样的。正如皮托在他的假设中指出的，这种分析研究结果的方法不能用于比较疗效相同的治疗方法，只有当疗效“不同”时才可使用。

皮托的方法后来被称为“意向治疗”（intert to treat）分析方法。这样命名的理由及其用途是：如果我们对医疗政策的总体结果感兴趣的话（该政策通常会推荐使用某个治疗方案），就得授权引而伸之医生，让他可以按照他的判断去调整治疗方法。用皮托的方法，临床实验的分析可以判断：建议使用一个给定的方法作为治疗的起点，是不是一个好的公共政策。“意向治疗”分析方法最被认为是一种很好的方法，适合用于那些政府资助的、为制定好的公共政策而进行的大型研究。

很不幸的是，有些科学家往往在并不了解和理解其背后数学含义的情况下，随意地把一些统计方法拿过来就用，这在临床研究中是司空见惯

的。皮托早就指出了他的方法的局限性，但是意向治疗方法不但已经成为许多大学里的医科教条，并且被认为是临床实验唯一正确的统计分析方法。在许多临床实验中，尤其是对癌症的研究实验，实验设计是为了证明新的治疗方法至少与标准治疗方法效果相同，同时副作用较小。很多的实验目的是为了显示新疗法的等效性。正如皮托指出的，他的方法只能用来找出差别，但是，如果没有找出差异也并不代表两种方法的疗效相同。

某种程度上，这个问题的产生主要是因为奈曼－皮尔逊理论的刚性。在基础统计学的教科书里都可找到奈曼－皮尔逊理论的标准版本，假设检验往往被介绍为一种固定的程序，方法中许多完全随意的方面也被描述成不变的。

尽管许多这些随意的元素并不适用于临床研究　，但是一些医学家在研究中不得不用“正确”的方法，这种需求使得他们视奈曼－皮尔逊理论为最严格的信条，除非通过统计程序事先确定了P值，并且使之保持不变，否则没有任何事是可接受的。这是费歇尔反对奈曼－皮尔逊理论的原因之一，他认为P值和显著性检验的应用程序不应该受如此严格条件的限制，他特别反对奈曼事先竟然确定了错误概率的存在，并且只有在P值小于这个事先确定的值时才有效。费歇尔在《统计方法和科学推论》（Statistical Methods and Scientific Inference）一书中建议，对于P值多大才有意义，最后结果应视情况而定。在这里我用了“建议”的字眼，是因为费歇尔从没有很明确地说明他怎么使用P值，他只是提供一些例子。

考克斯的理论

1977年，大卫?R?考克斯（即第23章里提到的博克斯和考克斯中的一位）开始研究费歇尔的论点，并对它们加以发展。为了区分费歇尔所用的P值和奈曼－皮尔逊理论，他称费歇尔的方法为“显著性检验”（significance testing），而称奈曼－皮尔逊的理论为“假设检验”（hypothesis testing）。在考克斯撰写他的论文的时候，统计显著性（通过计算P值）的计算已经是应用最广泛的科学研究方法，因此，考克斯断言，这种方法已经证明了其在科学研究中的作用，尽管存在费歇尔与奈曼之间的尖锐争执，尽管存在W?爱德华兹?戴明这样的统计学家坚持认为假设检验毫无用途，尽管出现了根本不需要计算P值、不需要考虑显著性的贝叶斯统计学……总之，尽管在数理统计学家之间存在着上述这些争论，显著性检验和P值一直被使用着。考克斯就问了：科学

家真的在使用这些检验吗？他们怎么会知道这些检验的结果是真的还是有用的呢？他发现，在实践中，科学家用假设检验主要是通过消除不必要的参数，来提高其对现实的了解程度，或是用来在两个不同的现实模型间进行选择。

博克斯的研究方法

博克斯（博克斯和考克斯中的另一位）从稍微不同的角度来研究这个问题。他认为，科学研究不只是做一个简单的实验，科学家在进行实验前，已经掌握了大量的知识，或者至少对实验的结果已经有了一个期望值，研究是为了提升知识、实验设计取决于你要提升的知识类型。在这一点上，博克斯和考克斯具有很多共同之处。对于博克斯来说，一次实验是一系列实验的一部分，将这次的实验数据与其它实验的数据进行比较，那么早先的知识就会在新的实验中和对以往实验的重新分析中得到重新审视。科学家从未停止过对以往研究的回顾，并从较新的研究视角去提升过去的认识。

举一个关于博克斯方法的例子。假设一个造纸厂引进了博克斯的一个主要创新方法——调优运算（evolutionary variation in operations，EVOP），按照博克斯的方法，这个工厂在生产过程中引入了一系列的实验，用不同的方法在温度控制、速度、硫磺处理过程以及温度控制等环节进行了微调，结果发现纸张的强度变化不大。如果要生产的产品仍然可销售的话，这种变化是不能大的。然而，根据费歇尔的方差分析（analysis of variance），用这些微弱的差别可以进行另外一个实验，在这个新的实验中，纸的平均强度稍微增大，这样，这个新的实验就可以用来确定可以提高纸张强度的工作方向。在过程操作改进中每个步骤的结果都与先前步骤的结果进行比较，当得到的结果看起来比较反常时，实验要重新做，这个过程周而复始——永远没有所谓最终“正确”的结论。在博克斯的模型里，这个不断进行着数据检验和再检验的科学实验是没有尽头的——没有最后的科学真相。

戴明的观点

戴明和其他许多统计学家坚决否定假设检验的作用。他们坚持认为费歇尔的估计方法才是统计分析的基础，认为真正应该估计的是统计分布的参数，而通过P值和武断的假设间接地处理这些参数而进行的分析是毫无意义。这些统计学家继续使用奈曼的置信区间去衡量他们研究结论的不确定性，但是他们却认为奈曼－皮尔逊的假设检验就象K?皮尔逊的矩

法（method of moments）一样已经过时了。有趣的是，奈曼自己也很少在他的应用性论文里用到P值与假设检验。

对假设检验的拒绝以及博克斯与考克斯对费歇尔显著性检验定义的重新诠释，使得人们可能对于皮托在癌症临床研究中解决问题的方法提出质疑。但是他面对的这个根本问题始终没有解决。当接受治疗的病人改变治疗方法，实验因此被动地做了调整时你能怎么做？亚伯拉罕?沃尔德（Abraham Wald）已经指出在实验中怎样的调整是可以接受的，那就是序贯分析（sequential analysis）。但是在皮托的问题中，肿瘤学家不会采用沃尔德的序贯分析法，一旦他们察觉到必要时，他们就会采用不同的治疗方法。

科克伦的观测研究

从某种方面来说，皮托的问题也是约翰?霍普金斯大学的威廉?科克伦在20世纪60年代研究的问题。巴尔地摩（Baltimore）市政府想知道，公共住宅是否影响低收入人群的社会态度和生活水平的提高。他们联系了约翰?霍普金斯大学的统计小组，请求他们帮助设计一个实验。按照费歇尔的方法，约翰?霍普金斯大学的统计学家建议寻找一群人，不论他们是否申请了公共住宅，随机分配公共住宅给其中一部分人，而对其中的另外一些人不提供公共住宅。这个建议吓坏了市政官员，以往，在公布安置公共住宅时，他们通常的做法是先到先受理，这是惟一公平的做法，他们不能拒绝那些先提出申请而却是因为计算机的随机抽取而没有选中的人。但是约翰?霍普金斯大学的统计学家指出，不管使用何种方法，那些最先申请的人通常都是最积极并且有野心的人，如果这种说法是对的，那么住在公共住宅里的人本来就比另外一些人干得好，这与提供住宅本身无关。

科克伦的结论是，如果他们不能够采用已经设计好的科学实验，那么通过追踪那些住进公共住宅以及那些没有住进的家庭，他们可以采用观察研究的方法来替代。这些家庭有很多因素不同，如年龄、受教育程度、宗教信仰以及家庭的稳定状况。他对这类观察研究的统计分析提出了许多方法，在各种方法中，他会考虑不同家庭的上述因素对测量结果进行调整，建立一个数学模型，其中包括年龄、是否是单亲家庭、宗教信仰等因素的影响力。一旦代表这些因素的影响力参数估计出来了，剩下的影响就应该是由公共住宅造成的。

如果临床研究声称，治疗效果的差异已经根据病人年龄和性别的差异进

行了调整，那就是说研究人员在估计治疗方法的主要效果时，已经应用了科克伦的方法，并且考虑了在治疗中为病人指定方法不平衡性的影响。几乎所有社会学研究都采用了科克伦的方法，但有些研究的作者可能没有认识到他们用的方法来自科克伦，而且认为其中很多特殊技术通常比科克伦的研究还要早。然而，科克伦为这些方法建立了稳定的理论基础，他写的关于观察研究的论文已经影响了医学、社会学、政治科学和天文不，在这些领域里“治疗方法”的随机指派，既不可能，也不道德。

鲁宾模型

在20世纪80年代和90年代，哈佛大学的唐纳德?鲁宾（Donald Rubin）提出了不同的方法，来解决皮托的问题。在鲁宾的模型中，假设每个病人对每个治疗方法都有一个可能的反应，也就是说，如果有两个治疗方法A和B，我们可以只观察采用其中一种治疗方法的病人，这些病人采用的方法是已经确定的。我们可以建立一个数学模型，在这个模型的公式中用一个符号来表示每种病人可能会有的反应。鲁宾界定了这个数学模型的使用条件，而在估计病人转而使用其它治疗方法会有什么样的反应时，这些条件是必需的。

鲁宾模型和科克伦的方法可以应用于现代统计分析中，因为应用计算机可以处理大量的数据。这些方法即使在费歇尔时代有人想到了，也是不可能实现的，因为这个数学模型涉及的数据太多，计算非常复杂，必须要借助于计算机。这个方法经常要求进行迭代计算，计算机要进行上万甚至百万次的计算，最后才会收敛于一个最终的答案。

科克伦和鲁宾的方法是高度依赖特定模型的，也就是说，除非所用的这个复杂的数学模型能非常准确地描述现实，否则就不会得出正确的答案。如果使用他们的方法，就要求分析人员要建立一个能够全面或近似全面描述事实各个方面的数学模型，如果事实与模型不符，那么分析的结论就不成立。像科克伦和鲁宾这些方法的一个伴生部分，已经成为去确定事实与模型怎样的拟合度下，结论是稳健的一种尝试。目前，数学界正在致力于研究：在结论不再成立之前，事实与模型之间可以有多大偏差。科克伦在直到1980年去世以前的日子里，一直在研究这些问题。

统计分析方法可以看作是一个连续过程，一端是高度依赖模型的方法，如科克伦和鲁宾的方法；另外一端则是一些非参数方法，采用最普通的方式检查数据。正如计算机的出现使模型模拟的方法得以实现一样，在

使用非参数方法时，也发起了一场计算机革命，这种方法极少或根本不用设计数学结构，数据不必放在一个预想的模型中就可以展现它们的含义。这些方法在使用中都有一些奇怪的名字，像“解靴带”（“boot-strap”，我们称为“自助法”——译者注）。这是下一章要叙述的内容。

# 第28章 电脑随心所欲

圭多?卡斯泰尔诺沃（Guido　Castelnuovo）出生于显赫的意大利犹太家庭，他的家庭背景可以追溯到古罗马最早的凯撒时代。1915年，卡斯泰尔诺沃当时是罗马大学（University of Rome）的数学教授，他正在进行一场孤独的战争，他想在研究生项目中引入一些有关概率和精算数学的课程。当时，安德烈?柯尔莫哥洛夫还没有建立起概率论的基础，数学家认为概率只是一个使用了复杂计算技术的众多方法的集合，是数学中的一个有趣的花絮，经常作为代数课里的一个部分来教授，在纯数学美丽的微光尚待关注的时候，没有人认为值得在研究生项目中开设这种课程。就精算数学而言，这段时间是应用数学最低迷的时期，人的寿命及意外事故发生频率的计算都只是采用简单算术，所以，系里其他的数学教授都认为没有开设这个课程的必要。

卡斯泰尔诺沃不仅在代数几何学这个抽象领域做了许多开创性工作，他对数学应用也有着浓厚的兴趣，他还劝说系里的其他人允许他开设这个课程。作为教学的成果，他在1919年出版了第一本关于概率与统计应用的教科书《概率运算与应用》（Calcolo della probabilità e applicazioni），这本书被意大利其它一些大学用于类似课程的教学中。到了1927年，卡斯泰尔诺沃已经在罗马大学成立了统计与精算科学学院（The School of Statistics and Actuarial Sciences），而且在整个20年代和30年代，意大利学校里致力于精算研究的统计学家越来越多，他们与瑞典该领域的专家进行极其活跃的交流。

1922年，贝尼托?墨索里尼（Benito　Mussolini）在意大利实行法西斯主久，利用强权控制人民的言论自由，对大学里的学生和教职工都进行调查，以驱逐所谓的“国家的敌人”。在这次驱逐行动中，因为没有提及种族问题，所以卡斯泰尔诺沃是犹太人这件事没有被考虑进去　。所以最初的7年里他能够继续在法西斯政府的统计下工作。到了1935年，意大利法西斯与德国纳粹的联合导致在意大利实行反犹太的法律，70岁的卡斯泰尔诺沃失去了工作。

但是，这些并没有使这位不知疲倦的人停止工作，直到1952年去世。随着纳粹种族政策的实施，许多有前途的犹太研究生也被逐出大学。卡斯泰尔诺沃就在他和其他犹太教授的家里设立了特殊的课堂，坚持授课，以帮助这些犹太研究生继续他们的学业。卡斯泰尔诺沃除了写一些关于

数学历史的书外，还在他87岁时的最后日子里，研究决定论和机遇之间的哲学关系，并试图去说明因果的概念——这些我们已经在前面的章节中接触过了，在本书的最后一个章节我将作进一步的探讨。

由于卡斯泰尔诺沃的努力而建立起来的意大利统计学派，拥有稳定的数学基础，但大多数研究都是以在实际应用中遇到的困难作为出发点。而与卡斯泰尔诺沃同时代的年轻人科拉多?基尼（Corrado Gini）则带领罗马中央统计研究所（Istituto Centrale Statistica in Rome）进行了在精算方面的深入研究。罗马中央统计研究所是一家由保险公司设立的私人研究机构。基尼对所有应用课题的极大兴趣促使他在20世纪30年代期间与活跃在数理统计领域大部分年轻的意大利数学家保持着密切的联系。

格利文科－坎泰利引理

在这些意大利数学家中有一位叫弗朗切斯科?保罗?坎泰利（Francesco Paolo Cantelli，1875－1966），他差不多先于柯尔莫哥洛夫就建立了概率论的基础。坎泰利对基础理论研究（如研究概率的意义是什么？）不感兴趣，没有像柯尔莫哥洛夫那样更深入地研究概率论，他只是满足于用概率运算的各种方法去推导出一些基本的数学定理，而这些概率运算的方法都是自18世纪数学家亚伯拉罕?棣莫弗将微积分引入概率计算后就存在的。1916年，坎泰利发现了我们所称的数理统计的基本原理。尽管它非常重要，却起了一个不起眼的名字“格利文科－坎泰利引理”（the Glivenko-Cantelli Lemma ）。坎泰利是第一个证明了这个定理的人，并且，他非常理解它的重要性。至于柯尔莫哥洛夫的学生——约瑟夫?格利文科（Joseph Glivenko）对此定理也做出了贡献，他采用一种新的数学符号，即斯蒂尔切斯积分（Stieltjes integral）概括了这一结果，他的论文在1933年发表于一本意大利的数学期刊。格利文科所采用的数学符号是现代教科书中使用最多的一个符号。

格利文科－坎泰利引理是那种直观上显而易见的，但是，只有当别人发现后，你才会意识到，否则看不出来。如果有一些数，我们对它们的概率分布一无所知，那么数据本身可以用来构造一个非参数分布，这是一个不那么好看的数学函数，其间有许多断点，怎么看都不优美，尽管它的结构不雅观，坎泰利还是可以通过增大观测值的数量，来使不那么美的经验分布函数（empirical distribution function）越来越接近真实的分布函数。

格利文科－坎泰利引理的重要性立刻得到了承认，在这之后的20年里，

这个引理被用来还原并证明了许多重要的定理，它是一种经常用于证明中的数学研究工具之一。为了用这个引理，数学家在20世纪初，不得不想出一些计算方法的简便算法，如果没有小窍门，在大量的数据样本中用经验分布函数来进行参数估计，就需要有一部在一秒钟内可以进行数百万次计算的超强计算机。在20世纪50年代、60年代乃至70年代都还没有这样的机器，到了80年代，才有这样的计算机用于这样的计算。格利文科－坎泰利引理成为新统计方法的基础，而这种新统计方法只能生存在高速计算机的世界里。

埃弗龙的“解靴带”法

在1982年，斯坦福大学的布拉德利?埃弗龙（Bradley Efron）发明了所谓“解靴带”（Bootstrap）（我们称为“自助法”）的方法，它基于格利文科－坎泰利引理的两种简单应用。这两种应用方法的原理很简单，但是它们要求用电脑进行大量的计算、再计算，……如果对一组数量适中的数据进行典型的“解靴带”分析，即使是利用最好的计算机也需要花好几分钟的时间。

埃弗龙把这种方法称为“解靴带”，是因为整个计算过程是一个数据自身模拟提升的过程，就像是解靴带一样，一个接一个地被解开。计算机不会介意重复单调的工作，它一遍又一遍地做着同样的工作，从不抱怨。由于使用了现代的晶体管芯片，计算机可以在不到万分之一秒内完成这些工作。在埃弗龙的“解靴带”背后还有一些复杂的数学理论，他最初的论文中证明了，如果对真实的数据分布做出了恰当的假设，这个方法与标准方法是等同的。这个方法的应用非常广泛，从1982年开始，几乎在每个数理统计期刊上都刊载一篇或更多的与“解靴带”相关的文章。

重复抽样和其它运算密集方法

还有其它一些与“解靴带”类似的方法，总称为重复抽样（resampling）。事实上埃弗龙已经阐述了费歇尔的许多标准统计方法都可以看作是重复抽样，而且，重复抽样方法属于范围更广的统计方法的一种，我们称之为“运算密集”（computer-intensive）。运算密集法充分利用现代计算机，对相同的数据不断地重复进行大量的运算。

20世纪60年代，美国国家标准局（the National Bureau of Standards）的琼?罗森布拉特（Joan Rosenblatt）和德州农工大学（Texas A&M University）的伊曼纽尔?帕仁（Emmanuel Parzen）发展了这种运算密集

的程序，他们的方法被称为“核密度估计”（kernel density estimation），而且，由此产生了“核密度回归估计”（kernel density-based regression estimation）。这两种方法涉及到两个任意参数，一个是“核”（kernel），另一个是“带宽”（bandwidth）。这些方法出现不久，1967年（远在计算机可以解决这些问题之前）哥伦比亚大学的约翰?范里津（John van Ryzin）利用格利文科－坎泰利引理确定了参数的最优配置。

当数理统计学家们还在研究理论，并在他们自己的期刊发表文章时，罗森布拉特和帕仁的核密度回归已经被工程界独立地发现了，在计算机工程师中，它被称为“模糊近似值”（fuzzy approximation）。它用了范里津所称的“非最优核”（nonoptimal kernel），并且，只是非常随意地选用了一个“带宽”。工程实践不是为了寻找理论上最佳的可能方法，而是在于追求可行性。当理论家们还在为抽象的最优标准而大费周折时，工程师们已经走出去，到了真实的世界，用模糊近似值的概念建立了以计算机为基础的模糊系统。模糊工程系统应用于傻瓜相机，可以自动对焦和调整光圈。这一系统还应用于新建筑物中，根据不同房间的不同需要调整和保持舒适的恒定室温。

巴特?科什科（Bart Kosko）是工程界一个私人咨询师，是模糊系统推广者中最成功的一位。当我读他书中列出的参考书目时，可以找到关于19世纪一些主流数学家，像戈特弗里德?威廉?冯?莱布尼茨（Gottfried Wilhelm von Leibniz）等的参考资料，还有对随机过程理论及其在工程领域的应用方面做出贡献的数理统计学诺伯特?维纳（Norbert Wiener）的一些资料。但我找不到罗森布拉特、帕仁、范里津或核回归理论（the theory of kernel-based regression）任何后来贡献者的资料。这表明，尽管模糊系统和核密度回归的计算机运算法则基本一致，但它们各自完全独立地得到了发展。

统计模型的胜利

运算密集法在标准工程实践中的扩展，是20世纪末统计革命已经渗透到科学界各个角落的一个实例。数理统计学家们已经不再是统计方法发展唯一的、甚至已经算不上是最重要的参与者了。在过去的70年中，科学家和工程师们并不知道那些刊载于他们期刊中最重要的理论经常一次次地被重新发现 。

有时，应用者应用基础定理时没有进行重新论证，仅仅凭直觉上以为是

对的就假定它是正确的。还有的情况是，使用者使用了已经被证明是错误的定理，仅仅是因为这些定理直观上看起来是正确的。存在这种问题的原因，是因为在现代科学教育中概率分布的概念已经根深蒂固，以至于统计学家和工程师们思考问题的方式也是基于概率分布的角度。一百多年前，K?皮尔逊认为，所有的观测都来自于概率分布，而科学的目的就在于估计这些分布的参数。在这之前，科学界相信宇宙遵守着某些规律，如牛顿运动定律，而观测到的任何差异都是因为误差的存在。逐渐地，皮尔逊的观点占据了优势，其结果，每个在20世纪接受科学方法训练的人都理所当然地接受了皮尔逊的观点。这种观点深深地植根于现代数据分析的科学方法之中，几乎没有人去考虑其所以然。很多科学家和工程师使用这些方法，但从不考虑K?皮尔逊观点的哲学含义。

然而，当科学研究的真正“主体”是概率分布这一观念被广为接受时，哲学家和数学家发现了许多严重的基本问题，我已经在以上的章节中概略地列举了一些，在下一章节将详细论述。

# 第29章 “泥菩萨”

1962年，芝加哥大学的托马斯?库恩（Thomas Kuhn）出版了《科学革命的结构》（The Structure of Scientific Revolutions）一书。这本书深刻地影响了哲学家们和实践者们如何去看待科学。库恩指出，现实是复杂的，是绝对不可能由一个有组织的科学模型来完全描述出来的。他认为科学就是试图模拟建立一个描述现实的模型，符合可用的数据，并且可以用来预测新实验的结果。因为没有任何一个模型是完全真实的，所以，数据越来越多，要求不断地配合新的发现去修正模型以修正对现实的认知。这样，模型因为带有特例的直觉上难以置信的延伸，变得越来越来复杂，最终，这个模型不再适用了。这时，有创新精神的人将会考虑建立一个全新的模型，一场新的革命在科学领域即将展开。

统计革命就是模型变换的例子。用19世纪决定论的科学观，牛顿物理学已经成功地描述了行星、月球、小行星和彗星等天体的运动，运动都是遵守几个明确的运动和引力定律；在寻找化学规律方面也取得了一些成功；并且达尔文的自然选择学说为理解进化提供了有利的依据；甚至有些人试图将这种寻找科学规律的模型研究引入社会学、政治科学以及心理学等领域。那时，人们相信寻找规律的难点在于测量不准确。

19世纪初，一些数字家如皮埃尔?西蒙?拉普拉斯认为，天文测量存在微小误差，可能是因为大气状况和测量的人为因素。他提出，这些误差也应该存在一个概率分布，从而开启了统计革命的大门。按照库恩的观点，这就是在获得新的数据后对机械式宇宙观进行的修正。19世纪，比利时学者兰伯特?阿道夫?雅克?凯特莱（Lambert Adolphe Jacques Quételet）最早开创了统计革命，他认为人类行为的规律也具有概率论的性质。他没有用皮尔逊的多参数方法，并且也不知道最佳估计方法（optimum estimation），他的模型是极其朴素的。

最终，人们发现，更加精确的测量反倒使模型预测值和实际观测值之间的差异变得更大，关于科学的决定论观点彻底崩溃，测量的越加精确，不但没有按照拉普拉斯的想法去消除误差，反而降低了人们观测行星真实运动的能力，而且表现出的差异越来越大。基于这一点，科学界已经做好了接受皮尔逊及其参数分布的准备。

本书前面的章节已经介绍了皮尔逊的统计革命是怎么逐渐改变整个现代科学的，尽管分子生物学遵循这种决定论（基因会决定细胞产生特殊的

蛋白质），但是，在该科学中产生的实际数据充满了随机性，而且基因事实上就是这些随机数据分布的参数。现代药物对人体功能的影响是绝对的，1毫克或2毫克药物就可能对血压或精神有很大的影响，这一点是确定无疑的。但是证明了这一影响力的药理研究过程，却是按照概率分布来设计和分析的，影响力就是这些分布的参数。

同样，经济计量学的统计方法被用来模拟一个国家或者一个企业的经济活动。我们确信的电子的质子这些次原子粒子在量子力学中都是作为概率分布描述的。社会学家用总体的加权算术平均数来描述个体的交互作用，但这只能按照概率分布的方式进行。在许多类似的科学领域里，统计模型的应用在它们的方法论中非常广泛。当谈及分布的参数时，好像它们是真的并且是可测量的一样。多变且不确定的数据集合，就是这些科学的起点，计算结果则是隐藏在大量计算中，以参数形式来表示，这些参数是永远不能通过直接观测得到的。

统计学家失去控制权

现代科学中的统计革命如此彻底，以致于统计学家已经失去了对过程的控制。在数理统计文献的基础上，分子遗传学家已经独立发展了自己的概率计算方法。计算机对大量数据的处理能力，和人们对整理并搞清楚这些巨大信息库含义的需求，促使信息科学这一新学科的诞生。在信息科学新期刊的文章中已经很少提到数理统计学家的工作，而且，在《生物统计》或《数理统计年报》中刊登过的许多分析方法，都正在被重新发现。统计模型在公共政策问题研究中的应用，已经演变成了一个被称为“风险分析”（risk analysis）的新学科，并且风险分析的新期刊也忽视数理统计学家的工作。

现在几乎所有新学科的期刊，要求在结论中有一个结果表，列出对统计结论产生影响的不确定因素的测量值。统计分析的标准方法已经成为大学中这些学科的研究生课程，通常，课程的讲授还不必同一个学校的统计系参与。

自K?皮尔逊发现偏斜分布的一百多年里，统计革命不仅扩展到大多数的科学领域中，而且其许多思想已经传播到了一般的文化当中。当电视新闻主持人宣布，某项医学研究已经表明被动吸烟的人的死亡风险比不吸烟的人高一倍时，几乎每个听众都认为他或她明白主持人的意思；当一个公众民意调查说65%的公众对总统表示满意，上下误差3%时，我们大多数人都认为我们都明白这个65%和3%的含义；当我们听到气象播报员

预测明天下雨的概率为95%时，大多数人出门都会带上一把雨伞。

除了这些我们自以为理解的可能性和比例问题外，统计革命对流行思潮和文化，有更深刻的影响力。即使实际测量的数据不够精确地与这些结论吻合，我们还是接受基于估计参数的科学研究结果。我们愿意根据众多数据算出的数来制定公共政策和安排我们的个人计划。我们认为搜集人口出生和死亡的数据，不仅是一个正当的程序，更有必要的工作，我们不必担心数人数会惹怒了上帝。从语言描述方面，我们用“相关”（correlation）或“相关的”（correlated）这两个词，好像它们意味着什么，也好像我们知道其含义。

写这本书的初衷是为了向那些没有数学专业背景的人士解释这场统计革命，我已经尽力描述了在这场革命背后的基本思想，它将如何应用于其他科学领域？它将如何最终主导几乎所有科学领域？我也尽力用语言和实例解释了一些数学模型，使大家不用再去研究抽象的数学符号就能够理解。

统计革命走到尽头了吗？

深邃未及的这个世界是一个集情感、事件与骚动的复杂混合体。我同意库恩的观点，我不相信人类的头脑能够构造一个理想的结构去解释、甚至不能挖地描述这个世界的真实情况。任何这种努力都存在根本的缺陷，最终，这些缺陷会变得非常明显，以至于科学模型必须不断地被修正，最终将走到它的终点，取而代之的是其它的什么东西。

随着统计方法应用的扩展，越来越多地应用到了人类生活的很多领域，哲学问题就显现出来。因此，我认为以讨论哲学问题作为本书的结尾是个好主意。接下来的将是在哲学领域中的一次冒险经历。读者可能想知道哲学究竟对科学信现实生活起到了什么作用。我的答案是，哲学并不是一些被称为哲学家的怪人们所做的神秘学术练习，哲学关注的是我们日常文化思想和活动的基本假设（underlying assumption）。我们的世界观来自于我们的文化，是受许多微妙的假设影响的，甚至很少有人会意识到它们。学习哲学会让我们揭开这些假设，并去检查它们的有效性。

我曾经在康涅狄格大学的数学系教过一门课程，这门课程有一个正式的名称，但是系里的人却更愿称之为“给诗人开的数学”。这门课只开一个学期，是为艺术专业的学生设计的，目的是向他们介绍基本的数学观念。在学期的开始，我向学生们介绍了16世纪意大利数学家吉罗拉莫?

卡尔达诺（Girolamo Cardano）的一本书《高等艺术》（Ars Magna），在这本书中，第一次描述了代数的方法。与他的大部头著作相呼应，卡尔达诺在该书的介绍中写道：代数不是新东西。他暗示他不是无知的傻子，他认为自人类产生以来，人类对知识的掌握一直在减少，亚里士多德所拥有的知识远远要多于卡尔达诺那个时代的任何一个人。他断言不可能有新的知识。然而，由于他的无知，他没能在亚里士多德的著作中找到关于代数思想的参考书目，所以他就把代数——这个看起来像是新东西的概念介绍给读者，他确信一些更加有知识的读者会从古人的著作中找到出处，这看起来是新东西的观念一定会被找出来的。

坐在我教室里的这些学生，生活在一个不同的文化环境中，他们不但相信后人会发现新事物，而且事实上，还鼓励创新。他们被卡尔达诺震惊了。写这些是多么愚蠢的呀！我告诉他们，在16世纪的时候，因为当时的一些基本哲学假设，欧洲人的世界观具有局限性，他们的世界观中，一个重要的部分就是人类的堕落以及随之而产生的道德、知识、工业等所有事物的持续退化，这些在当时是如此的真实，以至于很少有人去探寻究竟。

我问学生们，他们的世界观的基本假设中，哪些可能在500年后看起来是很荒谬的？他们一个都想不出来。

因为统计革命的表面观念已经传播到现代文化中，越来越多的人相信所谓的真实性，而不考虑它的基本假设，所以，让我们用统计的宇宙观来考虑下面三个哲学问题：

1. 可以用统计模型来做决策吗？

2. 当概率应用于现实生活中时其含义是什么？

3. 人们真的懂得什么是概率吗？

可以用统计模型来做决策吗？

牛津大学的L?乔纳森?科恩（L. Jonathan Cohen）是被他称之为“帕斯卡式”（“Pascalian”）观点的尖锐批评家，所谓“帕斯卡式”观点就是认为可以用统计分布去描述现实。1989年他写了《归纳和概率的哲学导论》（An Introduction to the Philosophy of Induction and Probability）一书，书中他提出了一个关于彩票的悖论，他认为那是康涅狄格州卫斯理大学

（Wesleyan University in Middletown Connecticut）的西摩?屈贝里（Seymour Kyberg）教授发明的。

假定我们接受假设或者显著性检验的观点，我们赞同如果现实中该假设的相应概率非常小，就可以拒绝这个假设。为了更进一步说明，假设0.0001就是一个非常小的概率，让我们组织一次公正的10000张彩票的抽彩活动。按这个假设，1号彩票中奖的概率，我们也可以拒绝这种假设，依次类推，我们可以拒绝类似的任何针对某号彩票的假设。按照这一逻辑规则，如果A不为真，B和C都不为真，那么A、B、C的集合也不为真。也就是说，按照这一逻辑规则，如果每一张彩票都中不了奖，那么就没有彩票可中奖（而事实却是总会有中奖的彩票）。

在科恩较早写的《可能与可证》（the Probable and the Probable）一书中，基于普遍的法律实践，他提出了这种悖论的一个变形。在习惯法（common law）中，一个涉及民事诉讼的原告提供了“有利”证据，其陈述看起来是真的，那么他就会胜诉，法庭接受原先诉求的概率高于50%。科恩还提出了一个关于“无票入场者”（gate crashers）的悖论：假设在一个有1000个席位的音乐厅里举办一场摇滚音乐会，主办单位只售出499张票，但是当音乐会开始的时候，1000个席位都坐满了，根据英国的习惯法，主办单位有权在音乐会上向每个现场的人收票钱，因为他们每个人无票入场的概率都是50.1%，这样，虽然音乐厅只有1000个席位，但是主办单位却将会有1499张门票的收入。

这两个悖论都说明了，以概率为依据所得到的决策是不合逻辑的，逻辑和概率是矛盾的。费歇尔在设计良好的实验基础上，利用显著性检验来证明科学研究中的归纳推理是可取的，但是科恩的悖论则表明，这样的归纳推理是不合逻辑的。杰里?科恩菲尔德根据积累的大量证据来判断吸烟会导致肺癌这个说法，但连续的研究表明，除非你假设吸烟是致癌的原因，否则这个结论是极不可能的。相信吸烟致癌是不合逻辑的吗？

以逻辑推理和统计为基础所得出决策上的不一致，是不能靠在科恩提出的悖论中找到错误的假设来解决的。这种不一致的深层次原因存在于逻辑的含义中（科恩认为概率模型可以由一种我们称为“模型逻辑”（model logic）的复杂数学逻辑结构来代替，但是我认为这个方法会产生更多的问题，比它所解决的问题还要多）。在逻辑上，一个命题是对还是错，我们是完全不同的。但是概率引入的观念却是说一些命题“可能”或者“多数”是对的。就是结果的这一点点不确定性，就使我们在分析原因和结果时，难以应用事物实质蕴涵的冷酷的精确性。在临床

实验中，处理这类问题的方法，是把每个临床研究看作是对某个治疗方案的效果提供资料。这些资料的价值取决于这个研究的统计分析，但则无也取决于研究的质量。研究质量这一额外的测量决定了哪些研究对结论起决定作用。但是，质量的概念含糊不清而且难以计算，悖论依然存在，而且吞噬着统计方法的核心。这种不一致的毛病是否需要在21世纪发起一场新的革命？

当概率应用于现实生活中时，其含义是什么？

柯尔莫哥洛夫建立了概率的数学定义：概率是一个抽象空间里对一事件集合的一种测量。所有概率的数学特征都可由这个定义导出。当我们希望在现实中使用概率时，我们需要确定眼前特定问题事件的抽象空间。当气象播音员说明天降雨的概率为95%时，什么是所测量的抽象事件的集合？是指明天要外出的所有的人吗？其中有95%的人会淋雨？还是指可能逗留在外面的时间？其中有95%的时间我会淋雨？或是说在一个1平方英寸大的地方，有95%的面积会下雨？当然这些解释都不对，那么到底是什么意思呢？

柯尔莫哥洛夫之前的K?皮尔逊认为概率分布是可以通过收集到的数据观察得出的，我们已经看到了使用这个方法存在的问题。

威廉?S?戈塞特试图为一个设计好的试验描述其事件空间。他说事件空间就是试验得出所有可能结果的集合。这听起来可能是对的，但是在实践中却是无用的。在实验中，我们必须相当精确地描述出结果的概率分布，才能计算出统计分析中需要用到的概率值。“所有可能实验结果的集合”的概念非常含糊，我们怎样才能得到一个精确的概率分布呢？

起初费歇尔同意戈塞特的想法，继而他发展了一个更好的定义。在他的实验设计中，治疗方案是随机分配给各个实验单位的。如果我们想在肥老鼠身上做实验，比较两个治疗动脉硬化的方案，我们就随机地在一些老鼠身上使用A方法，而在其余的老鼠身上使用B方法。实验开始进行，我们开始观察结果。假设两种治疗方案具有同样的效果，因为动物是随机使用治疗方法的，所以另外一些分配治疗的效果应该是同样的。随机治疗方法的标签是不相关的，只要治疗效果是一样的，我们就可以在动物间随意调换。因此，对于费歇尔，事件的空间是所有可能随机分配的治疗方案的集合。这是一个事件的有限集合，所有的事件都是等概率发生的。在所有治疗方法的效果是相等的零假设（null hypothesis）条件下，实验结果的概率分布是可以计算出来的，这就是我们所说的排列

检验（permutation test）或随机检验。当费歇尔提出这一检验方法时，还不能计算出所有可能的随机实验分配方式，费歇尔证明了，他的方差分析公式可以求得一个非常理想的排列检验的近似值。

那时还没有高速计算功能的计算机，而现在进行排列检验是可能的，因为电脑可以不知疲倦地进行计算，这样费歇尔的方差分析公式就不再需要了，而且很多数理统计学家经过多年求证得出的非常聪明的定理也不再需要了。只要数据结果是来自于一个随机控制的实验，就可以在计算机上用排列检验来进行所有的显著性检验。

如果对观测数据用一个显著性检验，那就不可能了。这是费歇尔反对吸烟与健康问题研究的主要原因。一些论文的作者使用统计检验方法证明他们的例子。费歇尔认为，除非他们研究的是随机化的实验，否则统计显著性检验就是不合适的。在美国法院中的歧视性案件就常常是根据统计的显著性检验来裁决的。美国最高法院（The U. S. Supreme Court）规定，统计显著性检验是一种可以在裁决中使用的方法，可以用来判定是否因为性别或种族歧视的原因而造成了影响。费歇尔如果知道，他一定会强烈反对。在20世纪80年代后期，美国国家科学院（The U. S. National Academy of Science）赞助了一项研究，研究在法院中使用统计方法作为裁决依据是否合理。这项研究的主持者是卡内基梅隆大学（Carnegie Mellon University）的斯蒂芬?菲恩伯格（Stephen Fienberg）和明尼苏达大学（the University of Minnesota）的塞缪尔?克里斯洛夫（Samuel Krislov）。这个研究小组在1988年发表了他们的研究报告。研究报告中的许多论文批判了将显著检验用于歧视性案件的作法，所持的论点类似于费歇尔在反对吸烟导致癌症的证据时所使用的理由。如果最高法院想在诉讼中使用显著性检验，它必须确定产生概率的事件空间。

如何找出柯尔莫哥洛夫事件空间？第二种方法来自于样本调查理论。当我们希望通过一个随机样本去判断整个群体的某些事时，我们要精确地确定要研究的人群总体，确立一个选取样本的方法，并且根据该方法进行随机抽样。在实验的结论中存在不确定性，我们可以使用统计方法来量化这一不确定因素。不确定性产生的原因，是因为我们处理的是样本而不是所有人群。我们研究的宇宙现象的真实数值是固定不变的，例如，支持总统施政政策的美国选民的百分数是确定的，只是他们不知道。能够使用统计方法的事件空间，是所有可能的随机样本的集合，同样，这是一个有限集合，它的概率分布是可以计算出来的。概率在现实

生活中的含义清楚地建立在抽样调查之上。

当统计方法应用于天文学、社会学、流行病学、法律或者天气预报等观测研究中时，事件空间就不好确定。在这些领域之中的很多争论，通常都是因为不同的数学模型会产生不同的结论。如果我们不能确定可进行概率计算的事件空间，那么就不能说某种模型比另外一种更适用。就像在很多法律案件中所显示的那样，两个统计专家分析同一组数据却得不到统一的结论。当统计方法越来越多地被政府和社会团体应用到观察研究和解决社会问题时，这个基本问题的存在，即不可能算出确切概率的事实，将使人们对这些统计方法的有效性产生怀疑。

人们真的懂得什么是概率吗？

概率在现实生活中还有一个含义是“个人概率”。美国的L?J?萨维奇和意大利的布鲁诺?德费奈蒂是倡导这种观点的先驱。其先驱地位的确定是因为萨维奇1954年出版的《统计学基础》（The Foundations of Statistics）一书。在这种观点下，概率是一个广泛的概念，人们很自然地使用概率来支配生活。在进行冒险前，人们总会本能地根据可能产生结果的概率根据可能产生结果的概率进行决策，如果预想危险的概率很高，人们就会采取回避的态度。对萨维奇和德费奈蒂来说，概率是一个普通的概念。人们不必去联系柯尔莫哥洛夫的数学概率，我们所要做的就是建立一些一般性的规则，将个人概率与生活联系起来，因此，我们只要假设人们在判断事件的概率时所遵照的规则是一致的就可以了。萨维奇在这一假设下提出了一些关于内部一致性的规则。

按照萨维奇和德费奈蒂的方法，个人概率对每个人来讲是独特的。对同样的数据进行同样的观察，有的人会判断降水概率是95%，有的人则会判断是72%，这样的事情是极有可能发生的。利用贝叶斯定理，萨维奇和德费奈蒂向人们展示了具有相同个人概率的两个人如果分析的是同一序列数据，最终他们会得到相同的概率估计。这是一个令人满意的结论：人看起来都是不同的，但却都是理性的。如果提供了足够的数据，理性的人们会最终求得共识，哪怕最初他们是存在意见分歧的。

约翰?梅纳德?凯恩斯在1921年发表的题为《关于概率的讨论》（A Treatise on Probability）的博士论文中，对个人概率提出了不同的看法。凯恩斯认为，概率是在某一文化教育背景下的人们，对其既定情况的不确定性的测量，概率的判断不仅是个人内心的直觉，还与个人的文化背景有关系。如果我们想在72%和68%之中作出哪一个更准确的选择，用

凯恩斯的方法就会很困难，因为人们的总体文化水平很难达到精确的同一程度。凯恩斯指出，如果只是为了做决定，我们很少或根本不必去知道这些事件确切的概率数值，只要将事件进行排序就足够了。根据凯恩斯的理论，我们只要知道哪一事件更可能发生就可以了。明天下雨比下冰雹的可能性要大，或者说明天下雨的可能性是下冰雹可能性的两倍。凯恩斯指出，概率可以是部分排序（partial ordering）。不必要把每件事与其它事情进行比较。我们可以忽视某些概率关系，如根本不必要把扬基队得总冠军的概率与明天下雨的概率联系起来。

照这样，关于概率含义的两个结论取决于人类对不确定性量化的愿望，或者至少是大致的量化的要求。在凯恩斯的《关于概率的讨论》中，他为他的个人概率的部分序列设计出了一个正式的数学结构。他的做法比柯尔莫哥洛夫为数学概率建立基础理论还要早。他所做的工作没有借鉴柯尔莫哥洛夫的理论。凯恩斯声称，他的概率的定义有别于1921年提出的概率数学的一系列数学计算公式。为了使凯恩斯的概率定义得到应用，使用者还必须符合萨维奇的一致性原则。

凯恩斯的定义提供了关于概率的一种观点，它是用统计方法进行决策的基础。这种观点认为概率不再以事件空间为基础，而是产生于所涉及人员的个人感觉的数值。接着希伯来大学（Hebrew University）的两个心理学家——丹尼尔?卡内曼（Daniel Kahneman）和阿莫斯?特韦尔斯基（Amos Tversky）开始了他们关于个人概率的心理学研究。

在20世纪70年代和80年代间，卡内曼和特韦尔斯基研究了个体理解概率的方式。他们的研究成果编入了由P?斯洛维奇（P. Slovic）编辑的《不确定情况下的判断——启发与偏见》（Judgment under Uncertainty: Heuristics and Biases）一书中。他们为大学生、大学教员和一般的市民提出了许多概率场景，他们发现没有人符合萨维奇的一致性原则，相反，大多数人对不同概率数值的含义甚至没有一个一致的观点。他们所发现最好的一点就是人们对50：50和“几乎肯定”的含义有着一致的认识。通过卡内曼和特韦尔斯基的研究，我们可以得出结论：天气预报员尽力想区分降雨概率90%和75%间的不同，但实际上他们根本不可能说清楚，而那些预报的收听者也不可能真的说清楚这两者间的区别。

1974年，特韦尔斯基在皇家统计学会的一次会议上宣布了他的研究结果。在随后的讨论中，斯坦福大学的帕特里克?苏佩斯（Patrick Suppes）提出了一个简单的概率模型，符合柯尔莫哥洛夫的公理，并且也模拟卡内曼和特韦尔斯基的发现。这意味着用这个模型的人在他们的

个人概率方面应该是一致的，在苏佩斯的模型中只有五个概率值：

必然为真

为真的可能性大

为真的概率为一半

为真的可能性小

必然为假

这导出了一个很无趣的数学理论。大概只有六个理论可由此模型导出，并且它们的论证几乎是不言而喻的。如果卡内曼和特韦尔斯基是对的，那么惟一有用的个人概率将对奇妙的抽象数学理论十分不利，并且由此产生的统计模型极基有限。事实上，如果苏佩斯的模型是惟一适合个人概率的模型，许多标准统计分析方法就毫无用处了，因为它们算出的差异水平低于人类感觉的水平。

概率真的必要吗？

统计革命背后的基本观点是：科学真实的主体是数字的分布，这个分布可以通过参数来描述。将概念溶入概率理论并处理概率分布，这是数学的方便之处。将数字的分布看作是概率数学理论的元素，这样就可以建立参数估计量的最优化标准，然后，去解决用数据描述分布时遇到的数学问题。因为概率看起来与分布的概念的关系是与生俱来的，许多人做了很多工作，试图让人们理解概率的含义，努力将概率的含义与现实生活联系起来，并且使用条件概率这一工具去解释学实验和观测的结果。

分布的思想可以存在于概率理论之外。事实上，许多“非正常分布”（improper distributions）（因为这些分布不符合概率分布的所有要求）已经应用于量子力学和一些贝叶斯方法中。排队论（queuing theory）（指两次排队间的平均间隔时间等于在队伍中等候的平均时间）的发展，推导出一个非正常的分布——描述一个人加入队伍必须要等候的时间。这正是一个将概率论的数学理论应用于实际生活，同时却将我们带离概率分布集合的一个例子。

21世纪将会发生什么事？

柯尔莫哥洛夫表现出来的最后的聪明才智，是他用一组有限符号序列的特性来描述概率。在这个描述中，信息理论不是概率计算的结果，而是概率本身的起源。也许在将来，某个人会继续他的工作，并且发展一个新的分布理论，而在新的分布理论中数字计算机的特性会被带入哲学理论的范畴。

谁知道呢？也许在什么地方有另外一个费歇尔，正工作于科学的最前沿，并在不久的将来，会以其前所未有的见识和观念打破目前的书面？也许在中国的内地，另一个吕西安?勒卡姆已经在一个没有文化的农家出生了；或者在北美，另一个乔治?博克斯只上了初中就休学了，现在正在做机修工，正在努力自学；也许另一个格特鲁德?考克斯将要放弃当传教士的愿望，被科学和数学的谜团深深吸引；或者另一位威廉?S?戈塞特正在努力寻找方法去解决啤酒发酵问题；或者另一个奈曼或皮特曼正在印度某个偏远的地方学院里教书，并且思考着深奥的问题。谁知道下一个伟大的发现将发生在什么地方？

当我们进入21世纪的时候，统计革命在科学领域取得了胜利，除了极少数的角落，它已经征服了科学界几乎所有领域的决定论观点。统计观点的应用如此广泛，以至于其基本假设已经成为西方世界通俗文化的一部分，就如同一尊泥菩萨一样立在那里，洋洋得意，而在未来的某个隐蔽的角落，另一场科学革命正在孕育，而那些即将发起这场革命的男男女女，可能正生活在我们中间。

作者后记

在写这本书之前，我已经将那些对统计发展有贡献的女士和先生们分成了两组，一组是我在书中提及到的，一组是我没有提及的。第一组人可能对我在书中只提及他们一小部分的工作而感到不满意，第二组人可能会因为我根本就没有提及他们的工作而表示抗议。。为了表达我对他们的敬意，我有必须解释一下我取舍的原则。

对第一组取舍的原因在于：现代科学的范畴太大了，任何人都不可能知道它所有的支派。因此，在有些研究领域，统计方法的应用可能非常广泛，但是我却不知道。在20世纪70年代，我曾查找过关于计算机在医学诊断中应用的资料。在查找过程中，我发现有三个互相独立的支派，在任何一个支派内人们互相引述论文，并且都发表在同一份期刊内，但是，不同派别的科学家却很少了解其他派别的人在做什么。这还只是在医学界这样一个小小的相关领域中的情形，在更广阔的科学界，可能有

很多人群在应用统计方法，并且可能有一些成果在我从来没听过的期刊中发表。我对统计革命结果的认识，来自于对一些数理统计主流期刊的阅读。不阅读这些主流期刊或者不在这些期刊中发表文章的统计学家，就像发展模糊集合论（fuzzy set theory）的工程师，他们可能做了很多值得记载的工作，但是因为他们不在我知道的科学或数学期刊上发表文章，那么他们的工作就不会被包括进来。

有些东西我是知道的，但还是被省略了。我不想写一本关于统计方法论发展的全面的历史书，因为这本书的读者定位是一些不懂或者略懂数学的人，所以我不得不选择一些能用文字而不是用数学符号来解释的例子，这就更进一步限定了我的选择。另外，我还想让这本书读起来比较流畅，如果我用了数学符号，我可能就可以说明了众多主题间的关系了。但是没有数学符号，这本书很容易退化为一种观念的介绍，这些观念间没有什么关系。这本书需要一条主线将各个主题组织起来，我所选择的贯穿20世纪统计学复杂理论的主线是与别人不一样的，一旦这条主线确定了，我就不得不忽视了统计学的很多方面，而实际上，我对它们同样非常感兴趣。

在我的书中，很多人我都没有提及到，这并不代表他们的工作不重要，更不代表我认为他们的工作不重要。仅仅是因为本书的结构限制，我没有办法将他们的研究写进来，只好放弃。

我希望读者读了本书后能有所启发，去进一步了解统计革命的内涵。我希望有人在读后甚至能钻研这个题目，加入统计研究的行列。在参考书目中，我选择了一些供没有数学学习背景的人阅读的图书和文章。在这些书中，其他许多统计学家尝试向我们解释了统计所学带给他们的乐趣，那些想进一步了解统计革命的读者将会喜欢其中的一些书。

我要感谢W. H. Freeman出版的公司相关人员在本书出版过程中所做的工作。感谢Don Gecewicz细致的校对与编辑；感谢Eleanor Wedge和Vivien Weiss最后文字定稿和进一步的校对；感谢Patrick Farace对本书潜在价值的肯定；感谢Victoria Tomaselli、Bill Page、Karen Barr、Meg Kuhta和Julia Derosa对本书的美术制作工作。

# 作者后记

在写这本书之前，我已经将那些对统计发展有贡献的女士和先生们分成了两组，一组是我在书中提及到的，一组是我没有提及的。第一组人可能对我在书中只提及他们一小部分的工作而感到不满意，第二组人可能会因为我根本就没有提及他们的工作而表示抗议。。为了表达我对他们的敬意，我有必须解释一下我取舍的原则。

对第一组取舍的原因在于：现代科学的范畴太大了，任何人都不可能知道它所有的支派。因此，在有些研究领域，统计方法的应用可能非常广泛，但是我却不知道。在20世纪70年代，我曾查找过关于计算机在医学诊断中应用的资料。在查找过程中，我发现有三个互相独立的支派，在任何一个支派内人们互相引述论文，并且都发表在同一份期刊内，但是，不同派别的科学家却很少了解其他派别的人在做什么。这还只是在医学界这样一个小小的相关领域中的情形，在更广阔的科学界，可能有很多人群在应用统计方法，并且可能有一些成果在我从来没听过的期刊中发表。我对统计革命结果的认识，来自于对一些数理统计主流期刊的阅读。不阅读这些主流期刊或者不在这些期刊中发表文章的统计学家，就像发展模糊集合论（fuzzy set theory）的工程师，他们可能做了很多值得记载的工作，但是因为他们不在我知道的科学或数学期刊上发表文章，那么他们的工作就不会被包括进来。

有些东西我是知道的，但还是被省略了。我不想写一本关于统计方法论发展的全面的历史书，因为这本书的读者定位是一些不懂或者略懂数学的人，所以我不得不选择一些能用文字而不是用数学符号来解释的例子，这就更进一步限定了我的选择。另外，我还想让这本书读起来比较流畅，如果我用了数学符号，我可能就可以说明了众多主题间的关系了。但是没有数学符号，这本书很容易退化为一种观念的介绍，这些观念间没有什么关系。这本书需要一条主线将各个主题组织起来，我所选择的贯穿20世纪统计学复杂理论的主线是与别人不一样的，一旦这条主线确定了，我就不得不忽视了统计学的很多方面，而实际上，我对它们同样非常感兴趣。

在我的书中，很多人我都没有提及到，这并不代表他们的工作不重要，更不代表我认为他们的工作不重要。仅仅是因为本书的结构限制，我没有办法将他们的研究写进来，只好放弃。

我希望读者读了本书后能有所启发，去进一步了解统计革命的内涵。我希望有人在读后甚至能钻研这个题目，加入统计研究的行列。在参考书目中，我选择了一些供没有数学学习背景的人阅读的图书和文章。在这些书中，其他许多统计学家尝试向我们解释了统计所学带给他们的乐趣，那些想进一步了解统计革命的读者将会喜欢其中的一些书。

我要感谢W. H. Freeman出版的公司相关人员在本书出版过程中所做的工作。感谢Don Gecewicz细致的校对与编辑；感谢Eleanor Wedge和Vivien Weiss最后文字定稿和进一步的校对；感谢Patrick Farace对本书潜在价值的肯定；感谢Victoria Tomaselli、Bill Page、Karen Barr、Meg Kuhta和Julia Derosa对本书的美术制作工作。

# 大事年表

年份 事件 人物

1857 卡尔?皮尔逊出生 K?皮尔逊（Karl Pearson）

1865 圭多?卡斯泰尔诺沃出生 G?卡斯泰尔诺沃（Guido Castelnuovo）

1866 格雷戈尔?门德尔从事植物杂交实验 G?门德尔（Gregor Mendel）

1875 弗朗切斯科?保罗?坎泰利出生 F?P?坎泰利（Francesco Paolo Cantelli）

1876 威廉?西利?戈塞特出生 W?S?戈塞特（“学生”）（William Sealy Gosset）

1886 保罗?利维出生 P?利维（Paul Lévy）

1890 罗纳德?艾尔默?费歇尔出生 R?A?费歇尔（Ponald Aylmer Fisher）

1893 普拉桑塔?钱德拉?马哈拉诺比斯出生 P?C?马哈拉诺比斯（Parasanta Chandra Mahalanobis）

1893 哈拉尔德?克拉美出生 H?克拉美（Harald Cramér）

1894 耶日?奈曼出生 J?奈曼（Jerzy Neyman）

1895 发现偏斜分布 K?皮尔逊

1895 埃贡?S?皮尔逊出生 E?S?皮尔逊（Egon S. Pearson）

1899 切斯特?布利斯出生 C?布利斯（Chester Bliss）

1900 格特鲁德?M?考克斯出生 G?M?考克斯（Gertrude M. Cox）

1900 重新发现格雷戈尔?门德尔的成果 W?贝特森（W. Bateson）

续1

年份 事件 人物

1902 《生物统计》（Biometrika）第1期出版 F?高尔顿（F. Galton）、K?皮尔逊、R?韦尔登（R. Weldon）

1903 安德烈?尼古拉耶维奇?柯尔莫哥洛夫出生 A?N?柯尔莫哥洛夫（Andrei Nikolaevich Kolmogorov）

1906 塞缪尔?S?威尔克斯出生 S?S?威尔克斯（Samuel S. Wilks）

1908 《平均数的可能误差》（“The probable Error of the Mean”）“学生”t检验（student’s t-test） W?S?戈塞特

1909 弗洛伦斯?南丁格尔?大卫出生 F?N?大卫（Florence Nightingale David）

1911 弗朗西斯?高尔顿爵士去世 F?高尔顿（Francis Galton）

1911 《科学的法则》（The Grammar of Science） K?皮尔逊

1912 杰尔姆?科恩菲尔德出生 J?科恩菲尔德（Jerome Cornfield）

1912 费歇尔发表第一篇论文 R?A?费歇尔

1915 相关系数（correlation coefficient）的分布 R?A?费歇尔

1915 约翰?图基出生 J?图基（John Tukey）

1916 格利文科－坎泰利引理（Glivenko-Cantelli lemma）首次出现 F?P?坎泰利

续2

年份 事件 人物

1917 L?J?萨维奇出生 L?J?萨维奇（L. J. （“Jimmie”） Savage）

1919 《概率运行与应用》（Calcotlo della probabilità…）出版 G?卡斯泰尔诺沃（G. Castelnuovo）

1919 费歇尔在罗森斯特实验站（Rothamsted Experimental Station） R?A?费歇尔

1920 关于勒贝格积分（Lebesgue integration）的第一篇论文发表 H?勒贝格（H. Lebesgue）

1921 《关于概率的讨论》（A Treatise on Probability） J?M?凯恩斯（J. M. Keynes）

1921 《作物收成变动研究Ⅰ》（Studies in Crop Variation. Ⅰ） R?A?费歇尔

1923 《作物收成变动研究Ⅱ》（Studies in Crop Variation. Ⅱ） R?A?费歇尔

1924 《作物收成变动研究Ⅲ》（Studies in Crop Variation. Ⅲ） R?A?费歇尔

1924 《消除心智缺陷》（The Elimination of mental Defect）——费歇尔关于优先学的第一篇文章 R?A?费歇尔

1925 《研究工作者的统计方法》（Statistical Methods for Research Workers）第一版出版 R?A?费歇尔

续3

年份 事件 人物

1925 统计估计理论（极大似然估计（ML Estimation）） R?A?费歇尔

1926 关于农业实验设计的第一篇论文 R?A?费歇尔

1927 《作物收成变动研究Ⅳ》（Studies in Crop Variation. Ⅳ） R?A?费歇尔

1928 奈曼－皮尔逊（Neyman－Pearson）关于假设检验（hypothesis testing）的第一篇论文 J?奈曼、E?S?皮尔逊

1928 三条极值渐近线 L?H?C?蒂皮特（Tippett）、R?A?费歇尔

1928 《作物收成变动研究Ⅵ》（Studies in Crop Variation. Ⅶ） R?A?费歇尔

1930 《数理统计年报》（Annals of Mathematical Statistics）第一期出版 H?卡弗（H. Carver）

1930 《自然选择的遗传理论》（The Genetical Theory of Natural Selection） R?A?费歇尔

1931 印度统计研究所（Indian Statistical Institute）成立 P?C?马哈拉诺比斯（P. C. Mahalanobis）

1933 概率的公理化 A?N?柯尔莫哥洛夫

1933 《印度统计年报》（Sankhya）第一期出版 P?C?马哈拉诺比斯

续4

年份 事件 人物

1933 概率单位分析（probit analysis）成果完成 C?布利斯（C.Bliss）

1933 塞缪尔?S?威尔克斯到达普林斯顿（Princeton） S?S?威尔克斯（Samuel S. Wilks）

1934 奈曼的置信区间（confidence intervals） J?奈曼

1934 中心极限定理（central limit theorem）的证明 P?利维、J?林德伯格

1934 切斯特?布利斯在列宁格勒植物保护研究所（Leningrad Institute for Plant Protection） C?布利斯（Chester Bliss）

1935 鞅理论（martingale theory）的首次发展 P?利维

1935 《实验设计》（The Design of Experiments）出版 R?A?费歇尔

1936 卡尔?皮尔逊去世 K?皮尔逊

1937 利用随机抽样对美国失业普查进行数字检查 M?汉森（M.

Hansen）、F?斯蒂芬（F. Stephan）

1937 威廉?西利?戈塞特去世 W?S?戈塞特（“学生”）

1938 《生物、农业与医疗研究统计表》（Statistical Tables for Biological，Agricultural，and Medical Research） R?A?费歇尔、F?耶茨（F. Yates）

1940 《统计方法》（Statistical Methods）教科书 G?W?斯内德克（G. W. Snedecor）

续5

年份 事件 人物

1941 亨利?勒贝格去世 H?勒贝格（Henri Lebesgue）

1945 在《统计的数学方法》（Mathematical Methods of Statistics）中对费歇尔的成果进行修订 H?克拉美

1945 威尔科克森关于非参数检验的第一个出版物 F?威尔科克森

1947 在出版物中第一次出现序贯估计理论（sequential estimation theory） A?沃尔德

1947 曼－惠特尼（Mann－Whitney）对非参数检验的表述 H?G?曼、D?R?惠特尼

1948 皮特曼在非参数统计推断方面的成果 E?J?G?皮特曼（E. J. G. Pitman）

1949 科克伦关于观测研究的成果 W?G?科克伦（W. G. Cochran）

1950 科克伦和考克斯关于实验设计的著作出版 W?G?科克伦、G?M?考克斯

1952 圭多?卡斯泰尔诺沃去世 G?卡斯泰尔诺沃（Guido Castelnuovo）

1957 费歇尔关于吸烟假定危险的辩论 R?A?费歇尔

1958 《极值统计学》（Statistics of Extremes）出版 E?J?冈贝尔（E?J? Gumbel）

1959 博克斯使用“稳健”（“rebust”）这一术语 G?E?P?博克斯（G. E. P. Box）

续6

年份 事件 人物

1959 假设检验的最终表述 E?L?莱曼（E. L. Lehmann）

1960 《组合机遇》（Combinatorial Chance） F?N?大卫、D?E?巴顿（D. E. Barton）

1962 萨维奇一德费奈蒂（Savage-de Finetti）个人概率理论的表述 L?J? 萨维奇、B?德费奈蒂（B. de Finetti）

1962 费歇尔关于遗传学中性别差异的最后论文 R?A?费歇尔

1962 罗纳德?艾尔默?费歇尔去世 R?A?费歇尔

1964 塞缪尔?S?威尔克斯去世 S?S?威尔克斯

1964 《变换分析》（An analysis of transformations） G?E?P?博克斯、D? R?考克斯（D. R. Cox）

1966 弗朗切斯科?保罗?坎泰利去世 F?P?坎泰利

1967 哈耶克秩检验的表述 J?哈耶克（J. Hájek）

1969 全国性三氟溴氯乙烷研究（包括对数线性模型的成果） Y?M?M? 毕晓普（Y. M. M. Bishop）及其他人

1970 南希?曼关于可靠性理论（reliability theory）和威布尔分布（Weibull distribution）的第一个出版物 N?曼（Nancy Mann）

1970 《赛局、上帝与赌博》（Games，Gods，and Gambling） F?N?大卫

续7

年份 事件 人物

1971 保罗?利维去世 P?利维

1971 L?J?萨维奇去世 L?J?萨维奇

1972 普林斯顿稳健估计研究（普林斯顿稳健性研究） D?F?安德鲁（D. F. Andrews）、P?J?比苛尔（P. J. Bickel）、F?R?汉佩尔（F. R. Hampel）、P?J?休伯（P. J. Huber）、W?H?罗杰斯（W. H. Rogers）、J?W?图基（J. W. Tukey）

1972 普拉桑塔?钱德拉?马哈拉诺比斯去世 P?C?马哈拉诺比斯

1975 斯特拉?坎利夫当选皇家统计学会（Royal Statistical Society）会长 S?V?坎利夫（Stella Cunliffe）

1976 “科学与统计学”，显著性检验应用的一个观点 G?E?P?博克斯

1977 考克斯对显著性检验的表述 D?R?考克斯

1977 《探索性数据分析》（Exploratory Data Analysis）出版 J?图基

1978 格特鲁德?M?考克斯去世 G?M?考克斯

1979 切斯特?布利斯去世 C?布利斯

1979 杰尔姆?科恩菲尔德去世 J?科恩菲尔德

续7

年份 事件 人物

1979 珍妮特?诺伍德被任命为劳工统计局（Bureau of Labor Statistics）局长 J?诺伍德（Janet Norwood）

1980 埃贡?S?皮尔逊去世 E?S?皮尔逊

1981 耶日?奈曼去世 J?奈曼

1982 混沌理论（chaos theory）的现代表述 R?亚伯拉罕（R. Abraham）、C?肖（C. Shaw）

1983 表明个人概率局限性的研究 A?特韦尔斯基（A. Tversky）、D?卡内曼（D. Kahneman）

1985 哈拉尔德?克拉美去世 H?克拉美

1987 安德烈?尼古拉耶维奇?柯尔莫哥洛夫去世 A?N?柯尔莫哥洛夫

1987 将核回归（Kernel－based regression）应用到调焦照相机（“模糊系统”） T?山川（T. Yamakawa）

1989 L?J?科恩对统计模型和方法的批评 L?J?科恩（L. J. Cohen）

1990 《观测数据的样条模型》（Spline Models for Observational Data） G?沃赫拜（G. Wahba）

1992 鞅方法用于医学研究得到了充分发展 O?奥伦（O. Aalen）、E?安德森（E. Anderson）、R?吉尔（R. Gill）

续7

年份 事件 人物

1995 弗洛伦斯?南丁格尔?大卫去世 F?N?大卫

1997 将科克伦方法（Cochran’s methods）扩展到序贯分析（sequential analysis） C?詹尼森（C. Jennison）、B?W?特恩布尔（B. W. Turnbull）

1999 使EM演算法适用于有关奥伦－安德森－吉尔鞅模型的问题 R?A?比滕斯凯（R. A. Betensky）、J?C?林赛（J. C. Lindsey）、L?M?瑞安（L. M. Ryan）

2000 约翰?图基去世 J?图基

# Table of Contents

封面

女士品茶

目录

第1章 女士品茶

第2章 偏斜分布

第3章 可爱的戈塞特先生

第4章 在“垃圾堆”中寻觅

第5章 收成变动研究

第6章 “百年不遇的洪水”

第7章 费歇尔获胜

第8章 致死的剂量

第9章 钟形曲线

第10 章 拟合优度检验

第11章 假设检验

第12章 置信诡计

第13章 贝叶斯异论

第14章 数学界的莫扎特

第15章 “小人物”之见解

第16章 非参数方法

第17章 当部分优于总体时

第18章 吸烟会致癌吗？

第19章 如果您需要最佳人选……

第20章 朴实的德克萨斯农家小伙

第21章 家庭中的天才

第22章 统计学界的毕加索

第23章 处理有瑕疵的数据

第24章 重塑产业的人

第25章 来自黑衣女士的忠告

第26章 鞅的发展

第27章 意向治疗法

第28章 电脑随心所欲

第29章 “泥菩萨”

作者后记

大事年表