TURING 图灵新知

THE SEVEN PILLAR OF STATISTICAL WISDOM

统计学七支柱

[美] Stephen M. Stigler 著
高蓉 李茂 译

追溯统计学来龙去脉，阐释统计推理核心思想

知名统计学家、哈佛大学文理研究生院院长 | **孟晓犁**
美国国家科学院院士、斯坦福大学教授 | **Bradley Efron**
斯坦福大学数学与统计学教授 | **Persi Diaconis**
联合推荐

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

Stephen M. Stigler

著名统计学家、统计学史研究家，芝加哥大学教授。其父是诺贝尔经济学奖得主George J. Stigler。除本书外，还著有《统计探源》等统计学著作。

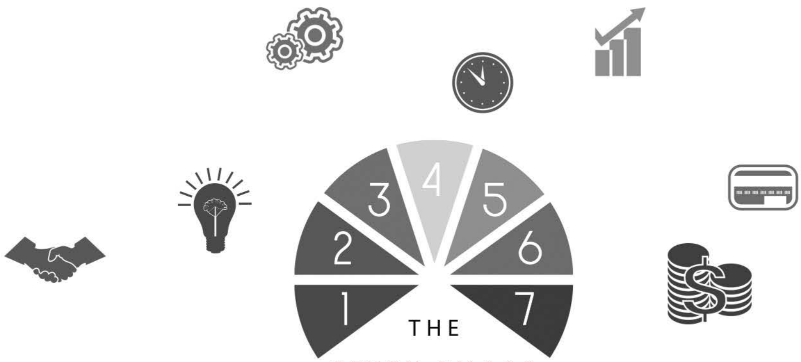
高蓉

博士，任教于杭州电子科技大学，研究领域包括资产定价、数据科学应用。已出版教材《实验投资学》《数据科学入门》，并有《量化金融R语言高级教程》等多部译著。本书翻译感谢杭州电子科技大学2016年高等教育研究资助项目YB201631“投资学教学与R软件应用”的支持。

李茂

毕业于北京师范大学，任教于天津理工大学数学系。热爱数据科学，从事统计、数据分析相关的教学和研究工作。

图灵新知



THE SEVEN PILLAR OF STATISTICAL WISDOM

统计学七支柱

[美] Stephen M. Stigler 著
高蓉 李茂 译

人民邮电出版社
北京

图书在版编目（C I P）数据

统计学七支柱 / (美) 史蒂芬·斯蒂格勒 (Stephen M.Stigler) 著 ; 高蓉, 李茂译. -- 北京 : 人民邮电出版社, 2018.1
（图灵新知）
ISBN 978-7-115-46997-7

Ⅰ. ①统… Ⅱ. ①史… ②高… ③李… Ⅲ. ①统计学 Ⅳ. ①C8

中国版本图书馆CIP数据核字(2017)第240876号

内 容 提 要

本书介绍了统计学的七个基本思想——聚合、信息、似然、相互比较、回归、设计、残差，从其由来到引入，从基本概念到对“统计”这门学科的深远影响，并由此深入阐述统计学的科学本质。

◆ 著　　[美] Stephen M. Stigler
译　　高　蓉　李　茂
责任编辑　陈　曦
责任印制　彭志环

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路11号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京　　　　　印刷

◆ 开本：880×1230　1/32
印张：4.5
字数：113千字　　　　2018年 1 月第 1 版
印数：1 – 3 500册　　　　2018年 1 月北京第 1 次印刷
著作权合同登记号　图字：01-2016-6686号

定价：39.00元
读者服务热线：(010)51095186转600　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

版权声明

送给我亲爱的孙女、孙子——Ava和Ethan。

前　　言

“统计学是什么？”早在1838年就有人提出过这个问题（与英国皇家统计学会有关），此后这个问题又被反复提起。多年来，铁打的问题和流水的答案已成为该讨论的特点。综合问题和答案可以看出，持续的疑问源于，统计学并不是一个单一学科。自诞生至今，统计学的工作内容经历了翻天覆地的变化：从极端强调“统计学家仅收集数据而不分析”，转变为从计划到分析的所有研究阶段皆积极寻求与科学家的合作。并且，统计学工作者面对不同的科学领域时，需要相应调整自身角色：在某些应用中，我们接受基于数学理论推导的科学模型；而某些应用中，我们构建如牛顿力学体系一样稳定的模型。在一些应用中，我们既是积极的计划者，又是消极的分析师；而在另一些应用中，我们的角色则恰恰相反。统计学工作者除了角色众多，还需要为了避免失误、保持角色平衡而面对种种挑战。这就难怪“统计学是什么”的老问题，无论面对哪个时代的新挑战，总会被重复提起。“统计学的挑战”在19世纪30年代指经济统计，在20世纪30年代指生物问题，而目前指定义模糊的“大数据”问题。

统计学有各种各样的问题、方法和解释，那到底有没有自己的核心科学呢？如果统计学工作者总是致力于在诸多科学领域工作——从

公共政策到验证希格斯玻色子的发现——甚至有时候只被视为服务人员，那统计学还能真正合理地被大家视为统一的学科吗？它能被视为我们统计学工作者自己的科学吗？这个问题就是我想在本书中解决的。我不打算告诉你统计学是什么或不是什么，而是尝试制定七个原则，即支撑统计学领域的七根支柱。它们在过去曾以不同方式支撑统计学，我保证，它们一定还会在无限的未来继续起到这样的作用。我会尽力使你相信，每根支柱的引入都是革命性的，并对统计学的发展产生了深远影响。

本书书名借鉴了托马斯·劳伦斯（即阿拉伯的劳伦斯）完成于1926年的回忆录《智慧七柱》。这部回忆录的名称源于《旧约·箴言》，《箴言》9:1写道："智慧建造了房舍，雕琢了七根支柱。"根据《箴言》，建造智慧的房屋是为了欢迎寻求知识的人。此外，本书还有一个目的：阐释统计推理的核心思想。

将这七个原则称作"统计学的七大支柱"之前，我先强调，它们是七根"支撑"的柱子，是统计学的学科基础，而不是完整的体系。一方面，这七根支柱都有古老的起源；另一方面，现代学科通过自身结构的伟大独创性，以及华丽承诺不断产生的精彩的新思想，将统计学构建为多元化的科学。在不脱离现代工作的前提下，我希望在统计学核心中建立跨时代和跨应用领域的连接和统一。

第一根支柱称为聚合（Aggregation）。我们也可以使用它在19世纪的名称"观测的组合"，甚至使用最简化的名称：均值。名字太简单可能误导读者，其实，虽然它现在看来已不新鲜，但在早年却真正地具有革命性，并且时至今日依然如此——无论它在何时进入新的应用领域。那么，它如何体现革命性？按照规定，给定一些观测值，你可以通过丢弃信息而真正获得信息！我们对观测值取简单的算术平均值，丢弃观测值的个别特征，而将其都纳入汇总值进行考虑。目前，这在

重复测量中很常见，比如观测恒星在太空中的位置。然而在 17 世纪，可能需要忽略这样一些信息，比如法国是个酒鬼观测员做出的观测，俄罗斯人是用旧仪器做出的观测，英国是个很靠谱的朋友做的观测。事实上，抹去个体观测的细节比任何单个观测都能给出更棒的指示。

根据记录，算术平均值的使用最早出现在 1635 年；而其他形式的统计汇总的历史则更为悠久，可以追溯到美索不达米亚文明最初出现文字的时代。当然，第一根支柱最近的重要实例更为复杂。最小二乘法及其衍生方法的本质都是均值，它们通过对数据进行加权汇总而抹去数据的个体特性——指定的协变量除外。甚至核密度估计和各类现代平滑器在本质上也是均值。

第二根支柱叫作信息（Information），更具体地说是“信息度量”，也是说来话长又很有意思。我们什么时候有足够的证据证明一种药物的疗效？这个问题可以追溯到古希腊。而研究信息积累率的时代则要近很多。18 世纪早期，人们发现在很多情况下，一个数据集的信息量仅与观测个数 n 的平方根成正比，而不与 n 本身成正比。这也是革命性的思想。假设你试图说服一名宇航员，如果他想将研究精度提高一倍，那么他需要用 4 倍数目的观测；又或者，第二组 20 个观测值与前 20 个观测值尽管同样精确，但第二组的信息量并不像第一组的那么大。我们将这个思想称为“根号 n 规则”。它需要一些很强的假设，并且在很多复杂的情形中使用时需要修正。无论如何，1900 年就明确建立了这样的思想：数据中的信息可以测量，而测量的精度与数据量有关，某些情形下可以精确刻画相关性的形式。

我将第三根支柱命名为似然（Likelihood），意味着使用了概率的推理的校准。显著性检验和普通的 P 值都是最简单的似然形式，但诚如其名，与“似然”有关的方法丰富多彩，其中许多方法或者与费舍尔推断的参数族有关，或者与贝叶斯推断的参数族有关。各种各样的

检验可以追溯到至少一千年前，但最早使用概率的检验则出现在 18 世纪早期。许多例子出现在 18 世纪~19 世纪，而系统性处理则出现在 20 世纪罗纳德·费舍尔的工作，以及耶日·奈曼和伊冈·皮尔逊的工作中。从那时起，统计学家开始认真发展了一整套似然理论。人们最熟悉的检验可能是用概率校准推断，但一个概率数字无论作为置信区间还是贝叶斯后验概率，都必须完全附属于一种推断。事实上，250 年前发表的“托马斯·贝叶斯定理”就是为了完成这个目标。

第四根支柱的名字是相互比较（Intercomparison）。这个名称借鉴了弗朗西斯·高尔顿的一篇论文，它表达了一个过去激进但现在普通的思想：统计比较常常可以采用数据自身的内部标准，而不必采用外部标准。相互比较最常见的例子是学生 t 检验和方差分析的检验。一方面，在复杂设计中，变化的划分可能错综复杂；另一方面，复杂设计允许区组设计、裂区设计，或完全根据手头数据评价的层次设计。这种思想非常激进，而且在“有效”的检验中，这种思想有着与最强大的工具一样的问题：可能由于忽略外部科学标准而导致错误方式的滥用。我们可以将自助法视为相互比较在假设弱化后的现代版本。

第五根支柱叫作回归（Regression）。这个名称源于高尔顿 1885 年发表的论文，这份文献基于二元正态分布解释了什么是回归。达尔文的自然选择理论存在内部矛盾：选择需要增加多样性，但定义物种需要群体外观稳定。高尔顿尝试为这个理论设计一个数学框架，并成功地克服了这组矛盾。

回归现象可简单解释为：假设有两个不完全相关的观测变量，你选择了其中极值远离均值的变量，那么可以预期另一个（以标准差为单位）不会那么极端。高个子的父母平均会孕育身高稍矮的子女，而高个子的子女平均会有身高稍矮的父母。但这一现象涉及的不只是一个简单的悖论：真正新奇的思想在于，提问的方式不同，答案就完全

不同。事实上，这项工作引入了现代多元分析和任何推断理论都需要的工具。引入这个条件分布的工具前，真正一般化的贝叶斯定理无法使用。因此，这根支柱与因果、推断一样，是贝叶斯学派的核心内容。

第六根支柱是设计（Design）。类似于在“实验设计”中的含义，但“设计”的范围更广泛，它的目标是：先设定观测的权重相同，再训练我们的思想。设计的某些要素历史悠久，《旧约全书》和早期的阿拉伯医学提供了相应的例子。从19世纪晚期，随着查尔斯·皮尔斯和费舍尔先后发现随机化在推断中的巨大作用，统计学出现了对设计主题的新理解。费舍尔认识到结合严谨的随机化方法将会带来好处，于是在实验法则中引入激进的改变。这些改变一反几个世纪以来的实验哲学和实践，将这一主题提升到了一个新的高度。多因素现场试验中，费舍尔的设计允许效应的分离和相互作用的估计；实施随机化后，有效推断不再需要正态性或者材料的均匀性的假设。

第七根也是最后一根支柱称为残差（Residual）。“残差”表示“其他的一切”，你也许会怀疑这是一种托词，但我想表达一种更具体的思想。从19世纪30年代开始，有关残差现象的概念在关于逻辑的书籍中就很常见。正如一位作者所说：“复杂的现象……可以通过减去已知原因的影响进行简化……留下……需要解释的残差现象。通过这样处理……科学……得到了极大的促进。”而后，这种思想总体上归入古典的范围，却以一种新方式在统计学中得到使用。这种新方式结合了结构化模型族，并通过概率计算和统计逻辑在族内做选择，从根本上强化和规范了方法。模型诊断（画出残差）在统计学中极为常见，但通过拟合和比较嵌套模型探索高维空间的方法更具重大意义。每个对回归系数显著性的检验都体现了这种思想，针对时间序列的每一个探索亦是如此。

我重新概括了七根支柱，用七种基本统计思想的作用来表达——

尽管这样做也许会导致过度简化的风险。

(1) 定向减少或压缩数据的价值。

(2) 数据量上升，价值会减少。

(3) 如何使用概率测量我们做的事？

(4) 如何使用数据中的内部变化帮助分析？

(5) 从不同角度提问可以产生有启发性的不同答案。

(6) 规划观测的重要作用。

(7) 所有这些思想如何用于科学探索和比较彼此矛盾的解释。

但是，无论这些思想出现于过去还是现在，以上平淡的陈述都没有表现出这些思想出现时的革命性。在当时，这些思想——从放弃数据值的个体特点到降低新数据和等价值数据的权重，再到克服障碍使用概率测量博弈外部的不确定性——已经丢弃或推翻了既有的牢固的数学和科学信念。世界产生了数据，那么数据自身的变化如何能够测量世界的不确定性？高尔顿的多元分析向科学家揭示，科学家依赖的比例规则（流传自欧几里得时代的比例规则）不适用于数据有变化的科学世界。这推翻了 3000 年来的数学传统。费舍尔的设计直接否定了实验科学家和逻辑学家几个世纪以来深信的内容，他的模型比较方法对实验科学来说绝对新颖,而接受这种方法则需要几代人的思维改变。

想知道以上所有思想的革命性和影响力有多大吗？只要考虑一下这些思想持续受到的强烈批评便可知一二。这些批评常常攻击那些我认为价值很大的地方，列举如下。

- 批评统计将人视为纯粹的统计量，而忽略人作为个体的特性。
- 批评大数据仅仅可以回答那些默认基于规模基础的问题。
- 批评显著性检验会忽略问题的科学内容。
- 批评回归分析会忽略问题中更重要的内容。

这些批判本身也有缺陷。虽然不乏正确之处，并且在某些极端的

例子中直击要害，但是，这些批判常常只瞄准方法，而非方法在例子里的运用方式。1927年，爱德华・B.威尔逊对此做了一番精彩的评论："就像没有接受过工具训练的人会害怕仓库中的任何一件工具一样，缺乏统计学知识的人会相信科学方法论中的统计工具都非常危险。"

我将讲述这七根支柱，并简单介绍它们的历史。这七根支柱都是优秀的工具，但人们需要足够的智慧和训练才可以有效使用它们。这些思想不是数学的一部分，也不是计算机科学的一部分，它们是统计学的核心内容。另外，我现在需要承认，虽然在本书开始直接否认了我的目的是解释统计学是什么，但到本书结尾，其实我已经完成了这个目标。

现在，我要简短地回应一个未了结的问题：《箴言》9:1究竟说了什么？它是这样一条古语："智慧建造了房舍，雕琢了七根支柱。"为什么一间房屋需要七根柱子？这种结构无论在古代还是在当代似乎都鲜有人知。最近的一项我比较信服的研究表明，那些负责为日内瓦[①]和詹姆斯王[②]翻译圣经的16世纪学者们，因为不太了解早期的苏美尔神话，错误地翻译了这一节。七根支柱根本不是建筑的结构，而是大洪水之前美索不达米亚的七个伟大王国。七位智者向国王进谏建立了七个城市，七个王国正是建立在这七个城市基础之上的。因此，智慧的房屋建立在这七位智者的意见之上。时代更近的学者提出了新的翻译："智慧建造了房舍，七位智者奠定了其基础。"

正是由于远远多于七位的智者的不懈努力，我得以将他们的成果总结为七根支柱。其中一些智者的姓名已经淹没在历史的长河之中，在本书的相关部分，我们会读到他们的智慧成果。

① 指1570年在日内瓦出版的圣经译本。——译者注

② 英王詹姆斯一世下令将圣经译为英文，于1611年出版。——译者注

电子书

扫描如下二维码，即可购买本书电子版。

目　　录

第 1 章　聚合：从表格和均值到最小二乘 1

1.1　指针的变化 3
1.2　古代的聚合 10
1.3　平均人 14
1.4　聚合与地球的形状 17

第 2 章　信息：度量与变化率 23

2.1　铸币检查试验 24
2.2　亚伯拉罕·棣莫弗 26
2.3　优化、扩展、悖论 30

第 3 章　似然：概率尺度上的校准 35

3.1　阿布斯诺特和显著性检验 36
3.2　休谟、普莱斯和贝叶斯归纳 41
3.3　拉普拉斯检验 43
3.4　似然理论 46

第 4 章　相互比较：作为标准的样本内变异 ······ 51

4.1　戈塞特和费舍尔的 *t*-检验 ······ 52
4.2　弗兰西斯・埃奇沃思和方差成分的双因素分析 ······ 58
4.3　相互比较的一些陷阱 ······ 61

第 5 章　回归：多元分析、贝叶斯推断和因果推断 ······ 65

5.1　发现之路：从达尔文到高尔顿 ······ 68
5.2　高尔顿的解释 ······ 79
5.3　达尔文问题的解决 ······ 80
5.4　影响 ······ 81
5.5　多元分析和贝叶斯推断 ······ 82
5.6　贝叶斯推断 ······ 85
5.7　收缩估计 ······ 87
5.8　因果推断 ······ 88
5.9　三分律：愿你安息 ······ 92

第 6 章　设计：实验方案和随机化的作用 ······ 95

6.1　可加模型 ······ 97
6.2　随机化 ······ 100

第 7 章　残差：科学逻辑、模型比较以及诊断展示 ······ 109

结论 ······ 125

第 1 章

聚合：从表格和均值到最小二乘

第一根支柱——聚合，不仅最古老，也最激进。在 19 世纪，它被称为“观测的组合”。这种说法表达的思想是：把数据集中的个体值进行统计汇总，概括出的信息可以超越个体。统计学的整体概括大于各部分的加总。样本均值就是这样一个例子，它是较早就被大家重视的一门技术，同样的思想也反映在其他一些汇总指标上，比如加权均值，甚至最小二乘在本质上也是一种基于个体数据值的特征进行加权或调整的平均值。

在分析中，对数据以任何形式取均值都是一个相当激进的步骤，因为取均值会丢弃数据中的信息，让每个观测值失去个性：测量的顺序和不同的产生环境，包括观测者的身份。1874 年曾有一次万众瞩目的“金星凌日”，是 1769 年以来的第一次，因此许多国家都向最佳观测位置派遣了远征队。获知凌日开始与结束的确切时间，可以帮助精准确定太阳系的规模。不同城市的观测人员提供的观测报告能相似到使均值有意义吗？这些观测是由技术水平不同的人，使用不同的设备，在不同的地点和稍有不同的凌日发生时间做出的。就此而言，如果单个观测者连续观测一颗恒星的位置，切实感受每次抖动、停顿和心烦

意乱，是否足以拿来取均值呢？在古代甚至现代，对每个观测环境过于熟悉会打消组合观测的意愿，人们忍不住要去选择那个认为是最好的观测，而不会用其他疑为较差的观测值来跟它求均值。

即便在取均值的方法变得司空见惯之后，人们也不见得总能接受“信息少即是多”的想法。19 世纪 60 年代，威廉姆·斯坦利·杰文斯提出，通过价格指数来测量价格水平的变动，也就是采用不同商品价格变动的百分比的均值，就有批评人士认为，把生铁和胡椒的价格放在一起取均值非常荒谬。并且，一旦讨论到某个商品，这些历史知识渊博的研究者们总会认为，他们可以借助某个特定事件发生的缘由故事“解释”这个商品的每个变动和波动。1869 年，杰文斯强烈谴责了这种理由：“如果每个波动都需要复杂的解释，那么不仅这个主题的所有相关探索都没有希望，而且还得放弃那些依赖数值事实的完整统计和社会科学。”这并不是说讲述数据的故事错了，而是说数据（以及单独观测的个体特点）需要置于背景之中。如果需要揭示一般性的趋势，那么必须将观测视为一个集合，必须把它们组合起来。

豪尔赫·路易斯·博尔赫斯理解这一点。他于 1942 年出版了奇幻短篇小说《博闻强识的富内斯》，其中描述了一个叫作伊雷内奥·富内斯的人。一次事故后，富内斯发现自己几乎能记住所有事情。他能以最微小的细节重新建构每一天，甚至以后能再重复这次重构，但他缺乏理解能力。博尔赫斯写道：“思维是忘却差异，是归纳，是抽象化。而富内斯的拥塞世界中仅仅充斥着触手可及的细节。”汇总产生的益处大于个体。富内斯正是没有经过统计处理的大数据。

算术均值是什么时候开始用于概括数据集的？又是在什么时候受到广泛采用的？这两个问题相当不同。第一个问题也许没有答案，理由随后会讲。第二个问题似乎在 17 世纪的某段时间得到了答案，但无法确定精准日期。为了更好地理解测量和涉及的这种报告问题，我们

来看一个有趣的例子，它的内容包括了可能最早使用“算术平均”这种说法的出版例子。

1.1　指针的变化

到 1500 年，热爱冒险的水手日益增多，他们把磁罗盘或“指针”当作必备工具。无论在任何地方和任何天气情况下，指南针都可以读出“磁北”。更早的一个世纪以前，人们就已经公认“磁北”与真正的北方有差异；而 1500 年，人们还认识到，真正的北方和“磁北”之间的差异会随着地点变动。差异数量通常比较可观——10°，也许偏东，也许偏西。当时，人们相信原因是海边缺乏磁引力，所以指南针的偏差指向大陆而偏离海洋。因此，需要通过指南针的修正找到真正的北方，这称为“指针的变化”。那时，一些航行地图会在关键位置，比如通航的海峡或者海上可见的显著标志，标注这种修正的已知大小，水手们信任这些记录的偏差。威廉・吉尔伯特 1600 年出版了地磁学经典著作《论磁》，其中给出报告：只要地球稳定，就可以信赖每个位置的变化的恒定性，“因为磁针总是偏向东或者偏向西，所以即使在今天，无论在任何地点或区域，无论是海洋或陆地，变化弧度都保持相同。因此，除非发生大陆崩塌和国家湮灭，就像柏拉图和其他古代作家所讲的亚特兰蒂斯地区那样，否则它将永远不变”。

唉！水手们和吉尔伯特“痴心错付”了。1653 年，亨利・盖里布兰德比较了伦敦同一地点、时间相隔 50 多年的一系列磁针变化的测定，他发现这些变化已经发生了相当大的变动。1580 年，真正的北方需要向东修正 11°；而在 1634 年，修正已经减少到大约向东 4°。这些早期测量结果的每一个都是基于几个观测值进行的计算，仔细分析后可以发现，这些观测者们也都各自在朝着“使用算术平均”摸索前进，

但从未明确说明要这么做。

1581 年，威廉・布劳出版了一本题为《指南针或磁针变化的一个讨论》的小册子，是早期测定指针变化的一份最佳记录。在第三章中，他描述了测定变化值的一种方法，这种方法不需要预先知道真正北方在观测点的什么方位，他还实际演示了一次，在伦敦东部尽头的港口区莱姆豪斯，那地方距离格林尼治子午线不远。他提出使用一个星盘——其实就是一个标有刻度的铜盘，垂直悬挂的同时用一个取景器观测太阳并记下它的高度。每当太阳到达一个新的高度角（在中午之前上升，在中午之后下降），他就观察并记录罗盘表面一条阴影线的方向，这样可以取得太阳与磁北偏差的一个读数。太阳抵达子午线时，高度角会达到最大值，那时它在真正的北方（如图 1-1 所示）。

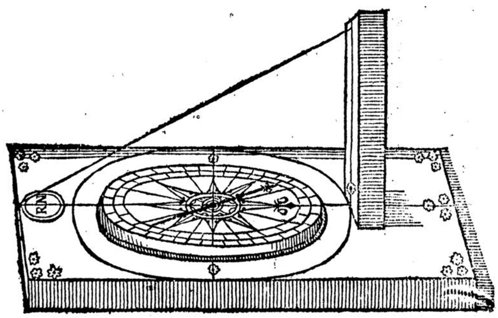


图 1-1　布劳使用的罗盘。垂直的柱子在罗盘北端，标记着一朵鸢尾花。附加的首字母 R. N. 指罗伯特・诺曼（Robert Norman），他预定了布劳的小册子。布劳在文中所指的罗盘上的“点”不是显示的 8 个点，而是由分割线将其间隔再分割为 4 部分，因此整个圆划分为 32 个部分，每个部分的大小为 11°15′（参见 Norman 1581）

布劳会考虑在同样的太阳高度角进行的每一对观测结果，一次在上午（图 1-2 中的 Fornoone，以下命名为 AM），另一次在下午（图 1-2

中的 Afternoone，以下命名为 PM）。一方面，如果莱姆豪斯的真正北方与磁北一致，那么共同值应该是（接近）两个测量的中点。因为太阳经过了一个对称弧度，在子午线（“正午”）的角度最大。另一方面，如果磁北位于真正北方以东 10°，那么早晨的阴影应该向西偏 10°，并且下午的影子也一样。无论哪种情况，成对观测的平均值应该就是磁针的变化。布劳 1580 年 10 月 16 日观测的数据表格如图 1-2 所示。

¶ In Limehouse the sixteenth of October. Anno, 1580.

Fornoone.			Afternoone.				
Elevation of the Sunne.	Variation of the shadow from the North of the Needle to the Westwardes.		Elevation of the Sunne.	Variation of the shadow from the North of the Needle to the Eastwards.		Variation of the Needle from the Pole or Axis.	
Deg.	Degr.	Min.	Deg.	D.	M.	D.	M.
17	52	35	17	30	0	11	17 ½
18	50	8	18	27	45	11	11 ½
19	47	30	19	24	30	11	30
20	45	0	20	22	15	11	22 ½
21	42	15	21	19	30	11	22 ½
22	38	0	22	15	30	11	15
23	34	40	23	12	0	11	20
24	29	35	24	7	0	11	17
25	22	20	25	Fró N. to w. 0. 8		11	14

图 1-2 布劳 1580 年在伦敦附近的莱姆豪斯对磁针变化的观测数据（参见 Norman 1581）

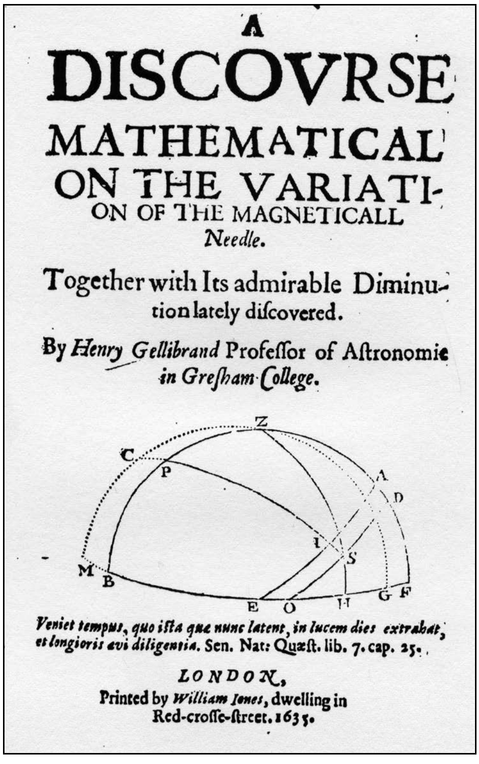
他有 9 对数据，取自高度角从 17°~25°的上午变化（偏西的角度）和下午变化（偏东的角度，因此和早晨的符号相反；除了 25°的下午测量，它稍微有点偏西）。因为上午和下午的符号不同，可以发现，图 1-2 右侧栏中的变化是变化差除以 2 的结果。例如，对于在太阳高度角 23°的观测对，我们有

$$\begin{aligned}(AM + PM)/2 &= (34°\ 40' + (-12°\ 0'))/2\\ &= (34°\ 40' - 12°\ 0')/2\\ &= (22°\ 40')/2 = 11°\ 20'\end{aligned}$$

计算出的这 9 个测定值相当接近，但又不完全相同。布劳是如何决定报哪一个数字的呢？在前统计时代，很明确是需要数据报告的，但又因为没有一致认可的一套概括方法，故而也就不需要描述什么概括方法，实际上也没有先例可循。布劳的答案很简单：参考右侧栏，他写道，“经过综合考虑，我确信莱姆豪斯真正的磁针或罗盘的变化是 11 又 1/4°或 11 又 1/3°，这刚好是罗盘上的一个点或稍多一点儿的值”。他给出的值是 11°15′，不能对应到任何现代的概括度量上——它小于均值、中位数、中点以及众数。它符合 22°高度角的值，并有可能就是这么选出的。但为什么 23°高度角的数字会给出 11°20′呢？也或许和他四舍五入到与“罗盘上的一个点”相一致，即 11°15′，罗盘上每相邻 2 个点之间的距离。无论如何，布劳认为没有必要给出正式的折中。他可以取上午和下午同一个高度角的平均值，但他用了一种聪明的做法：使用观测的对比得到结果，而不使用基本相等的观测的组合。“平均”就是一种“前减后”的对比。

半个多世纪之后的 1634 年，格雷山姆学院天文学教授盖里布兰德重温了这个问题（如图 1-3 所示）。12 年前，他在格雷山姆学院的前任埃德蒙・甘特在莱姆豪斯重复了布劳的实验，得到磁针变化的 8 个测定值。结果范围大约是 6°，与布劳的 11 又 1/4°相去甚远。甘特是一位杰出的观测者，但他缺乏将这个结果推广成一项发现的想象力，而将这个矛盾之处归结为布劳的错误。盖里布兰德对布劳极为尊敬，因此并不支持这种观点，他遗憾地写道：“这种巨大的差异使得我们当中某些人过早地中伤了布劳先生的观测（虽然某些仅仅是借口）。”盖里布兰德试着调整布劳关于太阳视差的数字，使用了第谷・布拉赫的一个

方法，布劳时代还没有这个方法，但是发现影响可以忽略（比如，布劳对于 20°高度角的值是 11°22 又 1/2′，变成了大约 11°32 又 1/2′）。于是，盖里布兰德开始使用昂贵的新设备（包括一个代替星盘的六英尺四分仪）在德特福德进行观测，这里是泰晤士河南岸，与莱姆豪斯隔河相望，并位于同一经度。



A
DISCOVRSE
MATHEMATICALL
ON THE VARIATI-
ON OF THE MAGNETICALL
Needle.

Together with Its admirable Diminu-
tion lately diſcovered.

By *Henry Gellibrand* Profeſſor of Aſtronomie
in Greſham College.

Veniet tempus, quo iſta quæ nunc latent, in lucem dies extrahat, et longioris ævi diligentia. Sen. Nat: Quæſt. lib. 7. cap. 25.

LONDON,
Printed by *William Iones*, dwelling in
Red-croſſe-ſtreet. 1635.

图 1-3　盖里布兰德小册子的封面（参见 Gellibrand 1635）

1634 年 6 月 12 日，盖里布兰德采用基于第谷表格的方法，分别做出了磁针变化的 11 个测定：5 个在上午，6 个在下午（如图 1-4 所示）。最大的是 4°12′，最小的是 3°55′。他总结说：

这些一致的观测产生的变化不会大于4°12′或小于3°55′，算术平均将其限制在大约4°4′。（此处的“度”的原文为gr.，简单地指代degree，是当时的一种“有刻度”的标尺单位。18世纪90年代法国大革命时期，标尺用gr.指代grad，表示直角的1/100。）

Obſervations made at Diepford An. 1634 Iunij 12 beforeNoone

Alt: ⊙ vera		Azim: Mag		Azim. ⊙		variatio	
Gr.	Min.	Gr.	M.	Gr.	M,	Gr.	M.
44,	45.	106,	0	110	6	4.	6
46,	30,	109,	0	113	10	4,	10
48,	31,	113,	0	117	1	4.	1
50,	54,	118	0	122,	3	4.	3
54,	24,	127	0	130	55	3	55

After Noone the ſame day.

Alt. ⊙ vera		Azi. Mag		Azim. ⊙		Variation	
Gr.	Min.	Gr:	M.	G,	Mn.	Gr,	Min
44	37	114:	0	109.	53.	4:	7
40	48	108:	0	103,	50	4:	10
38	46	105.	0	100,	48	4.	12
36	43	102,	0	97.	56	4.	4
34	32	99,	0	95,	0	4:	0
32	10	96:	0	91.	55	4:	5

Theſe Concordant Obſervations can not produce a variation greater then 4 gr. 12 min. nor leſſe then 3 gr. 55 min. the Arithmeticall meane limiting it to 4 gr. and about 4 minutes.

图1-4　盖里布兰德的数据和“算术平均值”的出现（参见Gellibrand 1635）

那么，盖里布兰德报告的“均值”并不是所有11个观测的算术平均值——4°5′。相反，他给出了最大值和最小值的平均值，也就是后

世统计学家们所说的“中点”，所以并不引人注目。尽管这是两个观测值的算术平均，但好像也没有其他方式可以生成两个数中间的数值了。事实上，早年的天文学家们面对两个值只需要取一个时，已经使用了算术平均或者类似的计算。可以确定，第谷和开普勒在 17 世纪早期，甚至阿尔–比鲁尼可能在公元 1000 年前后就使用了算术平均或者类似的计算。不过，盖里布兰德给所用的方法起了一个名字，这个术语是他工作的新颖之处。古人其实也了解这个名词，但看来没有人认为真有必要把它用于自己的著作中。

观测的统计分析确已进入新阶段，一个标志是 1668 年英国《皇家学会会刊》中的一个简短注记，其内容还是与磁针的变化有关。编辑亨利 · 奥尔登伯格刊登了某位姓名简写为 D. B. 的人的信件摘录，其中给出了布里斯托尔附近的某个位置磁针变化产生的 5 个值（如图 1-5 所示）。

An Extract

Of a Letter, written by D. B. *to the Publiſher, concerning the preſent* Declination *of the* Magnetick Needle, *and the* Tydes, *May* 23. 1668.

SIr, I here preſent you with a Scheme of the *Magnetical Variations*, as it was ſent me by Capt. *Samuel Sturmy*, an experienced Seaman, and a Commander of a Merchant Ship for many years; who (as he aſſures me) took the Obſervations himſelf in the preſence of Mr. *Staynred*, an antient Mathematician, & others, in *Rownham*-Meadowes by the water-ſide, in ſome ſuch approach, I think, to *Briſtol*, as *Lime-houſe* or the Fields adjoyning are to *London*. This (as the *Table* ſhews) was taken *June* 13. 1666. They obſerved again in the ſame day of the next year, *viz. June* 13. 1667; and then they found the Variation increaſ'd about *6.* minutes *Weſterly*.

Obſerved June 13. 1668.							
Sun's Obſerv'd Altitude.		Magnetical Azimuth.		Suns true Azimuth.		Variat. Weſterly.	
Gr.	M.	Gr.	M.	Gr.	M.	G.	M.
44	20	72	00	70	38	1	22
39	30	80	00	78	24	1	36
31	50	90	00	88	26	1	34
27	42	95	00	93	36	1	24
23	20	103	00	101	23	1	23

图 1-5　D. B. 信件公开的部分（参见 D. B. 1668）

D. B. 报告了斯特米船长的总结："采用这张**表格**的时候，他注意到最大距离或差异是 14′。因此，他对真正的**变化**取**均值**，并推断在**当时当地**，即 1666 年 6 月 13 日的变化，仅为 1°27′。"尽管真实的均值是 1°27.8′，并且斯特米船长（或者数学家斯特恩莱德）做了向下舍去，但无论如何都很明显，到 17 世纪的最后三十多年，算术平均值已经受到正式认可，成为组合观测的一种方法。它的诞生时间也许永远是个谜，但其诞生事实似乎无可辩驳。

1.2 古代的聚合

统计概括与书写一样拥有悠久的历史。图 1-6 是一块大约公元前 3000 年（与书写的起源时间很接近）的苏美尔人的泥板文书复原品，由芝加哥大学东方研究所的同事克里斯·伍兹向我展示。

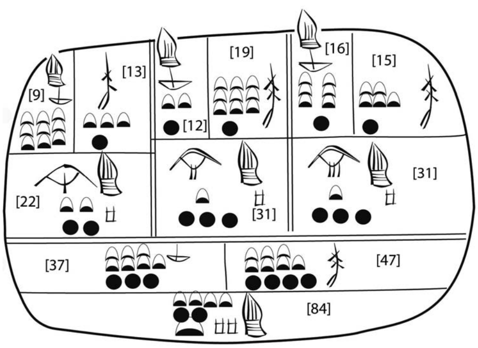


图 1-6　一块大约公元前 3000 年的苏美尔人泥板文书重现，添加了现代的数字（由罗伯特·英格伦复原，参见 Englund 1998, 第 63 页）

这块泥板代表的内容相当于一个 2 × 3 的列联表，显示了两种类型的商品计数，可能是两种作物 3 年内的产量（加上了现代的数字）。

顶上一行显示了 6 个单元格，商品符号显示在相应的计数之上。第二行是年份或者列的总计，第三行是两种作物行的总计，底部是全体的合计值。今天我们会以不同方式重列这些数字，如表 1-1 所示。

表 1-1　苏美尔人泥板文书数字记录的列联表形式

	年份 1	年份 2	年份 3	总计
作物 A	9	12	16	37
作物 B	13	19	15	47
合计	22	31	31	84

统计分析没有保存下来，但可以确定其中不包括卡方检验。我们能说的是，这块泥板展现了那个时代的高水平统计智慧，但它没有离个别数据值走得太远：不仅表格主体展现了每年所有作物的计数，泥板背面还给出了这些计数依赖的原始数据、个体生产者的个数。甚至 5000 年前就有人认为公开原始数据是有用的！

数据统计的科学分析始于何时呢？算术平均值的使用是什么时候变为统计分析的一个正式组成部分的？真的没有在 17 世纪以前很久吗？为什么更早的时代没有用均值对天文、调查和经济进行组合观测？古代的均值数学是众所周知的。毕达哥拉斯学派在公元前 280 年已经了解均值的 3 种类型：算术平均值、几何平均值和调和平均值。公元 1000 年时，哲学家波伊修斯将均值数量提高到了至少 10 种，包括毕达哥拉斯的 3 种在内。不可否认，这些均值是在哲学意义下展开的，主要用于讨论线段的比例，以及音乐，而非用于数据总结。

我们当然可以期待，古希腊人、古罗马人或者古埃及人早在 2000 多年前日复一日的生活中，就已经摸索出对数据取均值。又或者他们并没有这样做，但可以肯定的是，早在 1000 年前的阿拉伯科学的杰出天文研究中，就可以找到均值。但是，哪怕只是想在这些来源中找到

一个证据充分的例子，费尽心血广泛搜索之后，也总是免不了落空。

针对早期使用均值的历史，最坚定的搜索者是不屈不挠的研究者邱吉尔・艾森哈特，他在国家标准局度过了大部分职业生涯。数十年间，艾森哈特一直追踪均值的历史应用，并在 1971 年美国统计学会的主席演讲中总结了自己的研究。他热情洋溢地演讲了近 2 小时，但他发现的对于所有均值的相关使用工作、有证据表明使用均值的最早工作等，就是我前面提到过的由 D. B. 和盖里布兰德做出的。艾森哈特发现，希帕克（大约公元前 150 年）以及托勒密（大约公元 150 年）对自己的统计方法默不作声，而阿尔–比鲁尼（大约公元 1000 年）则使用通过二分最小值和最大值之差产生的数——并不接近均值。均值很早就出现在印度的应用几何中，婆罗摩及多在公元 628 年写的一本关于测量的小册子中有这样的建议：处理挖掘问题时，要使用与挖掘平均规模相一致的长方体当作不规则挖掘量的近似值。

所有这些年代的历史证据表明，人们收集了许多类型的数据。某些情况下，不可避免需要概括。如果不使用平均值，人们需要做什么以进行总结呢？选定单个数字进行报告吗？我们先看几个例子，其中运用了类似于均值的概念，看完之后也许会更好地理解前统计时代人们是怎样看这些问题的。

修昔底德讲过一个关于攻城梯的故事，发生在公元前 428 年：

> “一方为了达到敌人城墙的高度，需要制造一批梯子。因为城墙面向他们的一面粉刷不仔细，所以可以根据测量砖的层数计算城墙的高度。许多人同时数砖的层数，尽管有些人可能会数错，但大多数人会数对，尤其是多次计数之后。并且他们距离城墙也不远，完全可以看清楚。计算砖块的厚度后，就可以进一步推算梯子要求的长度了。”

修昔底德描述了所谓“众数”（mode，最频繁报告的值）的使用。因为计数过程缺失独立性的预期，众数并不非常精确。但如果报告非常接近，那它就和任何其他概括一样好。修昔底德并没有给出数据。

另一个很晚的例子来自 16 世纪早期，由雅各布·科贝尔在一本关于测量的图文并茂的书中提到。科贝尔说，那个时代土地测量的基本单位用一根 16 英尺长的木棒来确定。而且，当时的 1 英尺（foot）真的表示一只脚长，但是谁的脚呢？肯定不是国王的脚，也不是每次上台都会要求重新商定土地合约的新君主的脚。科贝尔说到的解决方案简单而优雅：在教堂礼拜之后留下 16 位市民代表（那时都是男性），他们鞋头对着鞋跟，站成一条线，这条线的长度就是那根 16 英尺木棒的长度。科贝尔的图片由他自己蚀刻，是一幅解释艺术的杰作（如图 1-7 所示）。

图 1-7　科贝尔关于确定一根合法木棒的描述（Kobel 1522）

这真是一根“社区的”木棒！而且，这根木棒确定以后，又细分为 16 个相等的部分，每个部分都表示这根公共木棒中单只脚（即 1 英尺）的度量。从功能角度讲，这就是 16 个人的脚长的算术平均值，但“均值”这个术语在任何地方都未提及。

这两个例子相隔大约 2000 年，但它们都涉及一个共同问题：如何概括一组相似但不完全相同的测量。每种情况中，解决问题的方式反映了组合涉及的智力困难，这种困难到今天依然存在。在古代和中世纪，每当需要概括不同数据时，人们便选择个别的例子。修昔底德的故事中，被选中的个别例子是最主流的情形——众数。而在其他示例中，也可以选择那个最突出的例子；对数值数据而言，甚至可以选择最大的那个记录值。每个社会都希望宣扬它们最好的部分以代表整体社会，或者选择的情形也可以是基于不明确的理由而选择的“最佳”个体或值。天文学中，“最佳”值的选择可能反映了观测者的个人知识或观测的天文条件。但无论做了什么，这意味着要保持至少一个数据值的个别特征。科贝尔的记述中，重点是 16 只个体的脚，甚至可以在图片中认出那时的人们。无论如何，“由个体共同决定木棒长度”，这种思想是一个强有力的观点，因为这没有抛弃它们的个性。这是木棒合法性的关键，甚至也决定了单独的英尺标志是真正意义的平均。

1.3 平均人

到了 19 世纪，均值已经广泛运用于天文学与测地学。19 世纪 30 年代，它还在社会中开辟了更广阔的应用空间。那时，比利时统计学家阿道夫·凯特勒开创了他所谓的“社会物理学”。为了可以在不同人群之间进行比较，他引入了“平均人”的概念。最初，凯特勒将这个概念当作人类种群之间的比较工具，或用来刻画单个种群随时间发

生变化的情况。有了这一工具，便可以比较英国人和法国人的平均身高；也可以随着时间的变化记录某一年龄的平均身高，由此导出一条种群生长曲线。社会中不存在单个的“平均人”，每个种群都有自己的“平均人”。另外，凯特勒只关注男性，女性不在考虑之内。

19 世纪 40 年代，一位批评家开始攻击这种思想。安东尼·奥古斯丁·库尔诺认为，“平均人”必然身体畸形：任何一个种群中，真正出现具有平均身高、体重和年龄的人的可能性非常低。库尔诺指出，对一组直角三角形相应的边进行平均再组成新的三角形，得到的图形不会是直角三角形（除非这些三角形都是彼此成比例的）。

另一位批评家是生理学家克劳德·伯纳德，他在 1865 年写下这样一段话：

> “数学在生物学中的另一个频繁应用是平均值的使用，可以说这在医学和生理学中必然导致错误……如果我们收集一个人 24 小时内的尿液，混合起来分析平均值，那么得到的是对一种根本不存在的尿液的分析。禁食时的尿液不同于消化时的尿液。一位生理学家发明了诸如此类的一个惊人实例，他选择了一座各国人都会经过的火车站，从小便池取出尿液，并相信自己能据此提出一份针对普通欧洲人的尿液分析！”

这种批评没有吓倒凯特勒，他坚称“平均人”可以作为一组人中的一个“典型”样本。这个样本抓住了“类型”，可以作为一组人的代表用于比较分析。因此，这个概念高度成功并经常受到滥用。“平均人”及其衍生概念发展出一套理论体系，使一些物理科学方法得以运用于社会科学。

19 世纪 70 年代，弗朗西斯·高尔顿分析非定量数据时进一步采用了均值的思想。他花费大把时间和精力，根据肖像的组合构建所谓

的“一般性肖像”。其中，通过叠加一组中若干成员的图像，本质上生成了这一组中男士或女士的平均图像（如图 1-8 所示）。高尔顿发现，从姐妹和其他家庭成员之间的面容相似之处可以提取家族特征。他也用了其他群组进行实验，生成了亚历山大大帝的勋章组合（希望能揭示出更逼真的画像），以及罪犯群组和相同疾病的患者群组。

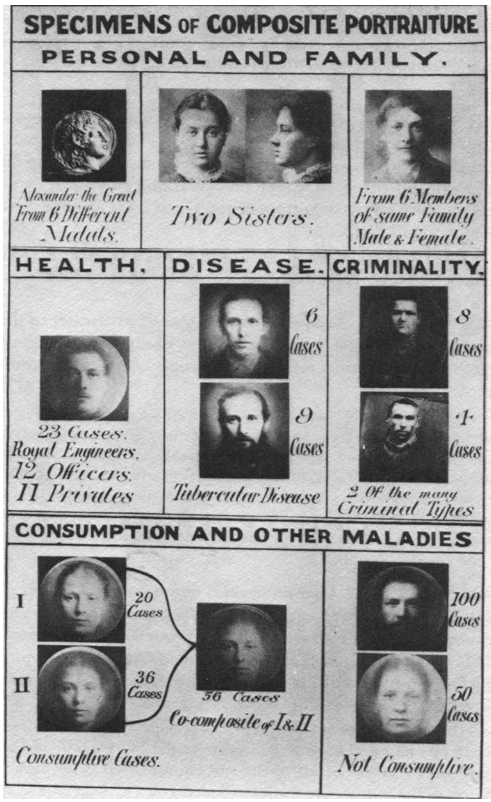


图 1-8　高尔顿的一些复合肖像（Galton 1883）

高尔顿合成照片时施加了一些约束，他很清楚这种一般性肖像的局限性。正如他自己的解释：“没有哪位统计学家会梦想着组合那些同

属一个种群但没有共同的中心聚集目标的对象。我们不应再用异质元素组合一般性肖像，如果这样做，结果会很可怕而且毫无意义。”他的一些追随者并没有这样谨慎。一位名叫拉斐尔·庞佩利的美国科学家于 1884 年 4 月参加美国国家科学院会议时为一些与会者拍摄了照片，第二年，他发表了图片合成的结果。图 1-9 是其中一个例子，这是由 12 位数学家（这个称呼在当时还包括天文学家和物理学家）的肖像叠加生成的“平均”数学家的合成图片。除了这张图片里的人看起来和高尔顿合成的那些罪犯一样阴险之外，我们还会注意到，将胡子剃干净的人、一些络腮胡子的人以及更多一些蓄小胡子的人的肖像组合后，产生的类型看起来更像是某个一周没刮胡子的人。

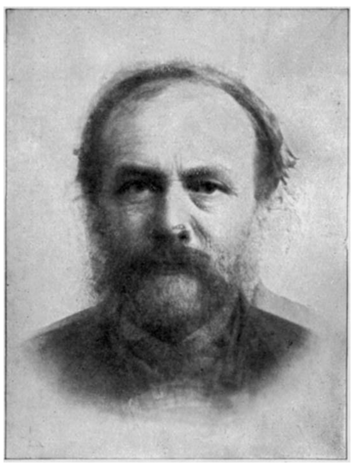


图 1-9　庞佩利的 12 位数学家的复合肖像（Pumpelly 1885）

1.4　聚合与地球的形状

到 18 世纪中期，统计聚合的运用已经扩展至许多场合，其测量是

在迥异环境中做出的。事实上，这也是环境使然。一个最简单的例子是 18 世纪关于地球形状的研究。初步估计，地球是一个球体；但随着航海和天文学精度的增加，问题随之而来。艾萨克·牛顿出于动态角度的考虑，提出地球是个略扁的球体（在两极处压缩，在赤道处膨胀）。法国天文学家多美尼科·卡西尼则认为，地球是一个扁长的球体——在两极拉长。要想解决这个问题，可以比较在不同纬度的地面做出的测量。从赤道到北极的几个不同地点，可以测量出一个相对较短的弧长——A。这条弧的方向垂直于赤道，是由北极到赤道的所谓子午线 1/4 圆的一段。可以先测量沿着地面的弧长，再除以两个端点的纬度差，结果是单位纬度的弧长。纬度可通过仪器观测北极星与水平线的夹角测得。观察这个 1°的弧是如何随着到赤道的距离变动的，即可解决这个问题。

椭圆积分给出了球体的弧长与纬度之间的关系，但一个简单公式就足以计算短距离（而且实际上，只有比较短的距离才可以测量）。令 A = 沿着地面测量的 1°弧长，L = 弧中点的纬度——这也是通过观测北极星决定的，则赤道有 $L = 0°$，北极有 $L = 90°$。因此，$A = z + y\sin^2 L$ 可以良好地近似每个测量的短弧度：

> 如果地球是一个完美的球体，那么 $y = 0$ 并且所有的 1°弧具有相同的长度 z。
>
> 如果地球是扁平的（牛顿），那么 $y > 0$ 并且弧长在赤道为 z（$\sin^2 0° = 0$），而在北极点变为 $z + y$（$\sin^2 90° = 1$）。
>
> 如果地球是扁长的（卡西尼），那么 $y < 0$。

y 的值可以认为是一种极地超额（如果值为负，那就是不足）。“椭圆率”（从球形形状出发的偏离测量）可以近似地计算为 $e = y/3z$，一

个稍微改进的近似计算式是 $e = y/(3z + 2y)$（有时会用到）。

计算需要数据。这个问题听起来很容易：测量任意两个度数，也许一个在赤道，另一个在罗马附近。那个年代的长度采用“突阿斯”（Toise）计量，是“米”制单位出现之前的单位，一突阿斯约合 6.39 英尺[①]。1°的纬度大约有 70 英里[②]那么长，实地测量太长、太难，因此需要测量一个短的距离再做推断。1736 年，皮埃尔·布格率领一个法国考察队，在今天厄瓜多尔的基多附近进行测量，那里可以在南北方向上测量接近赤道的较长距离。他测出的长度是 A = 56 751 突阿斯以及 $\sin^2 L = 0$。1750 年，耶稣会学者鲁杰罗·朱塞佩·博斯科维奇发现罗马附近的测量值是 A = 56 979 突阿斯以及 $\sin^2 L = 0.4648$。这给出了两个方程：

$$56\,751 = z + y \cdot 0$$

$$56\,979 = z + y \cdot 0.4648$$

这些方程很容易解出，得到 $z = 56\,751$ 和 $y = 228/0.4648 = 490.5$，以及 $e = 490.5/(3 \cdot 56\,751) = 1/347$，那时人们喜欢这么写计算结果。

但在 18 世纪 50 年代末，到博斯科维奇写出关于这个问题的报告为止，已经存在 5 个而不是 2 个获得肯定的弧长记录：基多（In America）、罗马（In Italia）、巴黎（In Gallia）、拉普兰（In Lapponia）以及一路南下直到非洲最南端的好望角（Ad Prom. B. S.）。其中任何两个都会给出一个结果，因此博斯科维奇面临数据的窘境：共有 10 个解，且它们各不相同（如图 1-10 和图 1-11 所示）。

① 约为 1.95 米。——译者注

② 约 112.7 千米。——译者注

Locus observationis	*Latitudo* °	′	*½ sin. vers.rad.* 10000	*Hexapedæ*	*Differ. a primo*	*Differ. computata*	*Error*
In America	0	0	0	56751	0	0	0
Ad Prom. B. S.	33	18	2987	57037	286	240	−46
In Italia	42	59	4648	56979	228	372	144
In Gallia	49	23	5762	57074	323	461	138
In Lapponia	66	19	8386	57422	671	671	0

图 1-10　博斯科维奇的 5 个弧长数据。各列（用我们的记号）对 $i=1\sim5$ 的每个弧给出了维度 L_i（用°表示）、$\sin^2 L_i$ [$=1/2\cdot(1-\cos L_i)=1/2\cdot\text{versin}\,L_i$]、$A_i$（hexapedae，用突阿斯表示的长度）、$A_i-A_1$ 的差、使用弧 1 和 5 的解的差，以及这些差之间的差。好望角的 $\sin^2 L$ 值应该是 3014，而不是 2987（参见 Boscovich 1757）

Binarium	*Differ. in pol., & æqu.*	*Ellipticitas*	*Binarium*	*Differ. in pol., & æqu.*	*Ellipticitas*
1, & 5	800	$\frac{1}{213}$	2, & 4	133	$\frac{1}{128}$
2, 5	713	$\frac{1}{239}$	3, 4	853	$\frac{1}{200}$
3, 5	1185	$\frac{1}{144}$	1, 3	491	$\frac{1}{347}$
4, 5	1327	$\frac{1}{128}$	2, 3	−350	$-\frac{1}{486}$
1, 4	549	$\frac{1}{314}$	1, 2	957	$\frac{1}{78}$

图 1-11　博斯科维奇计算的 10 对弧，对应每对弧都给出极地超额 y 以及椭圆率 $e=3y/z$。(2, 4)和(1, 3)的椭圆率印刷有误，应该是 1/1282 和 1/178。(1, 4)的数字有误，应该是 560 和 1/304（Boscovich 1757）

博斯科维奇现在进退两难。5 个测量的弧不一致，他应该随便选择其中一对并接受这个结果吗？恰恰相反，他创造了一种真正新颖的聚合方法，给出了综合 5 种结果后的客观答案。博斯科维奇认为，数据中最不可靠的要素就是弧的测量。这些弧需要在极端困难的环境下仔细测量，从巴黎和罗马附近的森林到非洲之角，再到拉普兰的冰冻

苔原，以及世界另一端的厄瓜多尔平原。而且，几乎不可能为了进行检查而重复这些测量。根据方程 $A = z + y\sin^2 L$，博斯科维奇进行了如下推理：z 和 y 的每个选择都隐含着 A 的一个对应值，并且这个值和观测值的差可以认为是一种调整，需要对观测的 A 进行这种调整以使测量匹配方程。所有可能的 z 和 y 中，隐含着“寻找调整绝对值总和的最小值”的目的，还假定选出的 z 和 y 与各个 A 的均值和各个 L 的均值相一致。博斯科维奇给出了一种聪明的算法求解最佳值，就是现在所谓“线性规划问题”的早期实例。对于这 5 个弧，根据他的方法求出的答案为：$z = 56\ 751$、$y = 692$、$e = 1/246$。

接下来的半个世纪，人们提出了多种方法，通过某种聚合形式整合不同条件下不一致的测量。最成功的方法是最小二乘法，它在形式上是观测的加权平均，而优势是很容易扩展为其他更复杂的形式，从而决定多个未知量。1805 年，阿德里安–玛丽・勒让德首次公布了这种方法——在一本解释如何确定彗星轨道的书中。勒让德给出了说明测定地球椭圆率的例子，采用的测量是法国大革命之后定义“米”的长度的方法。这些数据给出的椭圆率是 1/148，这个数值很大。但由于弧的范围更短（只有 10°的纬度，全在法国之内），并且与其他值不一致，因此人们认为它还不如早期从赤道到拉普兰范围内的测量。所以，最终的“米”是基于不同考察的混合值而决定的。

聚合具有多种形式——从简单的加总到不透明抽检的现代算法。但是，使用概括取代完全枚举个体观测的原则，和通过选择性地丢弃信息以获取信息的原则，都是一脉相承的。

第 2 章

信息：度量与变化率

第二根支柱——信息度量——从逻辑上与第一根支柱紧密相关：如果我们通过组合观测值获取信息，那么获取的信息与观测个数有何联系？我们应该如何度量信息的价值和获取过程？这个问题的历史悠久又有趣，可以追溯到古希腊。

在古希腊，“沙堆悖论” 非常著名：一粒沙不成沙堆；向一滩沙子添加一粒沙，也不会使它变成沙堆。但是，每个人也都同意，无论如何，沙子确实累积成了沙堆。一般认为，在公元前 4 世纪，哲学家米利都的欧布里德提出了这个悖论。5 个世纪以后，生理学家、哲学家盖伦将其作为一个统计问题再次提出，他展示了一个经验主义者和一个教条主义者之间的争论。

这个教条主义者是个早期的医学理论家，他会使用逻辑开药方：这种症状说明患者体质偏寒还是偏热？相应地判断应该为患者升温还是降温。患者身体里是否存在某种有毒元素？如果有，那么通过放血或其他方法排毒。

而经验主义者是证据论医学的支持者。如果对治疗有怀疑，就查看记录：放血或升温有多少次是有效的？这种疗法以前有用吗？以前

失败过吗？充分积累支持这种疗法的有利证据后，它就会作为一种标准被采用，而在那之前都要保持怀疑的态度。

教条主义者使用“沙堆悖论”反击：仅仅一个有利的证据当然不足以得出一般性结论；而且，当你处在任何不确定的阶段，多积累一个有利证据又怎会打破平衡？你会因此不再信任单一证据吗？但如此一来，你又如何信任证据的积累？可是，正如不可否认“沙堆”的存在，令人信服的医疗记录也是存在的。盖伦支持经验主义者，并认为必须充分重视医学史上积累的证据。但问题依然存在：证据的确越多越好，但究竟好了多少？这在很长一段时间里没有确切答案。

2.1 铸币检查试验

“缺乏答案导致成本增加”的一个例子是铸币检验的试验问题。12世纪的英格兰没有单一、强大、中央集权的权威，这对实施货币政策构成了挑战。虽然实际上有国王在位，但几位势力强大的贵族制衡了他的权威，这些贵族甚至通过 1215 年的《大宪章》强迫约翰王放弃权力。就在同时（或稍早一些时——早期的历史稍有点模糊），社会中产生了对普遍认可的货币的商业需求，这种货币可以获得广泛承认。伦敦铸币厂是英国硬币的主要来源，而且它在 1851 年变为皇家铸币厂，同时交由王国政府独立运作。国王和贵族向铸币厂提供金银锭，再收回硬币作为回报。为了保证该流程完美无缺，国王的契约发挥了巨大作用，它详细规定了硬币的重量和成色。为了监督铸币厂成功达到规定的标准，这份契约专门指出，应该通过试验检验铸币厂的产品。这是在生产过程中监测保证质量水平的早期实例。

这种铸币厂试验至少始于 13 世纪晚期，或许还要更早一个世纪。此后的时代有了试验过程的更详细描述，但之前应该也不可能有太多

变化。生产中，工人每天都会选择一些货币放入“货币检验箱”（Pyx），以便日后检验。选择并不是严格随机的，但账目中运用的一些词汇（比如“公正地”或者“随意取得”）说明它和随机样本差不多。货币检验箱会当着评审人的面在不同的时间间隔（在 14 世纪是每 3 个月）打开，这些评审人代表了与硬币精度利益攸关的各方团体。接着再进行一次选择，对一些硬币进行化验以检验金子的成色，用另外一些以检验重量。后者吸引了统计界的兴趣。

大家都能理解，不同硬币的重量不可避免会有不同，而契约专门指定了目标重量（用 T 指代）以及一个称为“公差”（remedy，用 R 指代）的可接受限度。如果重量低于 $T-R$，铸币厂的厂主就必须支付相应的罚金，经常会根据自前一次铸币检验之后的所有硬币支付现金罚金。早期的检验处罚很严厉，会威胁砍断厂主的手，甚至更糟。此外，硬币太重也是个问题，它们会被滑头的商家们从流通的货币中挑出，并融铸成金锭。但这对铸币厂毫无利润可言，检验的重点在于硬币太轻。

硬币会按批称重，这可能反映了一种模糊的理解：个体硬币的称重需要付出更多劳动，误差比例也会变得更大。比如，如果一次称一批 100 枚金币，目标显然是 $100T$。那公差应该是多少？他们的选择很有启发性：公差在这种情况下仅仅是 $100R$。只有这批硬币的重量小于 $100T-100R$ 时，才意味着铸币厂未能通过检验。但现代统计学告诉我们，这种观点错了。这个条件对铸币厂太过慷慨。这个标准如此之低，精明的铸币厂厂主会放低目标，比如将硬币的铸造目标设为 $T-0.5R$，甚至 $T-0.8R$，事实上也不会有测试失败的风险。如果硬币的重量变化是独立的（个体硬币的变化在统计上彼此无关），对一批 100 枚硬币，合适的公差是 $10R$ 而不是 $100R$。统计上独立的硬币重量的变化会随着个数的平方根增加。独立变化是个朴素的假设，给出的结果也确实更

接近事实，而不接近 $100R$。1866 年的一些数据说明，单个硬币的公差大约设定为两倍的标准差，这意味着一批 100 枚硬币的公差错误地设定为距离目标 20 个标准差。一些试验会涉及 1000 个甚至更多硬币的称重,铸币厂的测试结果可以满足任何官僚机构希望的安全标准，哪怕他们的目标比 $T-R$ 还高一点。

19 世纪的一次英国议会调查中，皇家铸币厂的官员受到了“是否降低了铸币标准”的质问，而他们让调查者相信自己从来都没做过此类行为，尽管他们相信法国人会这么干。当然，在铸币检验试验的最初几年，即使最优秀的数学家也不会知道今天称为“根号 n 规则”的知识，其中 n 是硬币个数。当然，有一位铸币厂的厂主比普通的数学家更加杰出：艾萨克・牛顿。1696~1727 年，他先任铸造厂的监管，接着成为厂主。1727 年牛顿去世，留下了一笔可观的财富。但很明显，他的财富归功于投资之道，不能捕风捉影地怀疑他发现铸币厂生产过程中的缺陷，并利用这一点中饱私囊。

2.2 亚伯拉罕・棣莫弗

直到 18 世纪初，人们才首次认识到，数据和的变化并不随着相加的独立项个数成比例上升（并且均值的标准差也不会随着项数的增加反比例减少）。精度信息不会随着数据的增加而产生线性积累，这种新颖的观点出现在 18 世纪 20 年代，由亚伯拉罕・棣莫弗提出，当时他正尝试通过大量试验寻找二项分布概率计算的精确方法。1733 年，他的尝试产生了著名结果，即现在的“二项分布的正态近似”。棣莫弗在 1730 年已经注意到，分布的一个重要特点是与根号 n 倍的离差相关。如果将二项频率函数视为一条曲线，拐点（可以认为它控制着散布）会出现在 $\pm\sqrt{n}/2$ 的区间内。

这个含义即使对于棣莫弗也很明确。他在关于正态近似的第 5 个推论中提出了标准差（尽管没有命名），在 *The Doctrine of Chances* 中稍有扩展——这部著作的拉丁文版本在 1738 年首次出版。当时，棣莫弗注意到，对于大的 n，拐点之间区间的总概率为 0.682688（大约 28/41），而介于 $\pm\sqrt{(2n)}/4$ 之间的较短区间的胜败机会相等（如图 2-1 和图 2-2 所示）。无论选择哪种确定的标准，是 68%还是 50%（或是 90%、95%，或者如同最近希格斯玻色子的发现，大约 99.9999998%），估计的精度都根据试验个数的平方根变动。

The DOCTRINE *of* CHANCES. 239

COROLLARY 5.

And therefore we may lay this down for a fundamental Maxim, that in high Powers, the Ratio, which the Sum of the Terms included between two Extreams diſtant on both ſides from the middle Term by an Interval equal to $\frac{1}{2}\sqrt{n}$, bears to the Sum of all the Terms, will be rightly expreſs'd by the Decimal 0.682688, that is $\frac{28}{41}$ nearly.

Still, it is not to be imagin'd that there is any neceſſity that the number *n* ſhould be immenſely great; for ſuppoſing it not to reach beyond the 900th Power, nay not even beyond the 100th, the Rule here given will be tolerably accurate, which I have had confirmed by Trials.

But it is worth while to obſerve, that ſuch a ſmall part as is $\frac{1}{2}\sqrt{n}$ in reſpect to *n*, and ſo much the leſs in reſpect to *n* as *n* increaſes, does very ſoon give the Probability $\frac{28}{41}$ or the Odds of 28 to 13; from whence we may naturally be led to enquire, what are the Bounds within which the proportion of Equality is contained; I anſwer, that theſe Bounds will be ſet at ſuch a diſtance from the middle Term, as will be expreſſed by $\frac{1}{4}\sqrt{2n}$ very near; ſo in the caſe above mentioned, wherein *n* was ſuppoſed $= 3600$, $\frac{1}{4}\sqrt{2n}$ will be about 21.2 nearly, which in reſpect to 3600, is not above $\frac{1}{169}$th part: ſo that it is an equal Chance nearly, or rather ſomething more, that in 3600 Experiments, in each of which an Event may as well happen as fail, the Exceſs of the happenings or failings above 1800 times will be no more than about 21.

图 2-1　棣莫弗的第 5 个推论，来自 *The Doctrine of Chances* 第二版。最后一段是在更早时期私下传阅的 1733 年版的文本基础上添加的（De Moivre 1738）

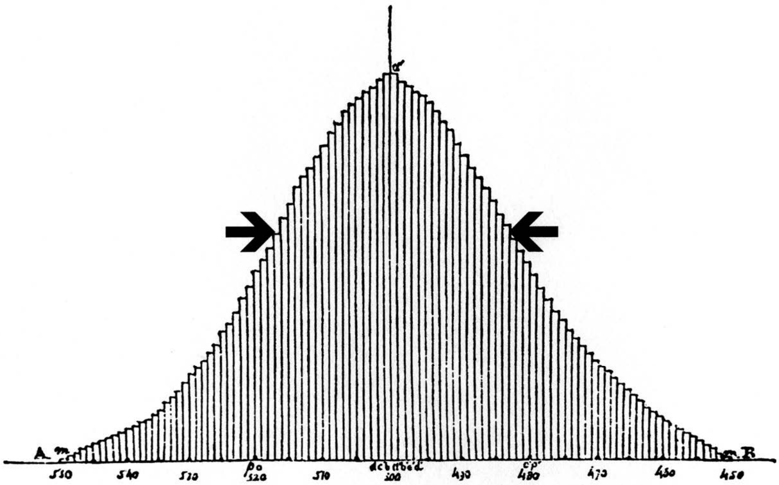
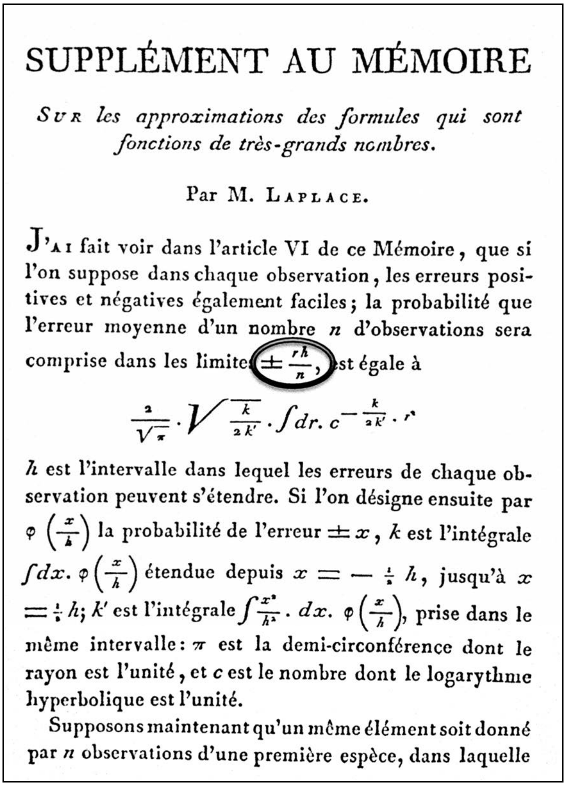


图 2-2　一张显示棣莫弗两个拐点的图片（和 1730 年讨论的一样），叠加在一张 $n=999$ 的对称二项分布图上，由凯特勒绘于 19 世纪 40 年代

1810 年，皮埃尔·西蒙·拉普拉斯证明了棣莫弗结果的一个一般形式，现在称为“中心极限定理”。其中，棣莫弗已经推断出，二项分布的成功个数 n 的近似变化会像正态曲线。拉普拉斯针对观测的总数或均值（比如一个硬币样本的重量）得到了同样的结论，其中个体观测（或者观测中的误差）几乎可以服从任何分布。但这个证明并不严谨。1824 年，西莫恩·德尼·泊松注意到一个例外情况，我们现在称为“柯西分布”。但是，这个结论对广泛的情况而言是成立的，并且数学家们迅速注意到了这个现象。



SUPPLÉMENT AU MÉMOIRE

Sur les approximations des formules qui sont fonctions de très-grands nombres.

Par M. Laplace.

J'ai fait voir dans l'article VI de ce Mémoire, que si l'on suppose dans chaque observation, les erreurs positives et négatives également faciles; la probabilité que l'erreur moyenne d'un nombre n d'observations sera comprise dans les limites $\pm\frac{rh}{n}$, est égale à

$$\frac{2}{\sqrt{\pi}}\cdot\sqrt{\frac{k}{2k'}}\cdot\int dr.\,c^{-\frac{k}{2k'}\cdot r^2}$$

h est l'intervalle dans lequel les erreurs de chaque observation peuvent s'étendre. Si l'on désigne ensuite par $\varphi\left(\frac{x}{h}\right)$ la probabilité de l'erreur $\pm x$, k est l'intégrale $\int dx.\,\varphi\left(\frac{x}{h}\right)$ étendue depuis $x = -\frac{1}{2}h$, jusqu'à $x = \frac{1}{2}h$; k' est l'intégrale $\int\frac{x^2}{h^2}.\,dx.\,\varphi\left(\frac{x}{h}\right)$, prise dans le même intervalle: π est la demi-circonférence dont le rayon est l'unité, et c est le nombre dont le logarythme hyperbolique est l'unité.

Supposons maintenant qu'un même élément soit donné par n observations d'une première espèce, dans laquelle

图 2-3　拉普拉斯关于中心极限定理的第一次清晰表述。这里他所用的 c 是我们现在常用的 e，k'/k 表示方差，积分表示等于平均误差不超过给定限制的概率（画圈的部分），但分母中的 n 应该是 $\sqrt{n}$（Laplace 1810）

讽刺的是，第一次出版的拉普拉斯的结果中有一个打印错误，它用 n 代替了 $\sqrt{n}$（如图 2-3 所示）。两年后图书付梓时，这个错误才得以纠正。

根号 n 规则的含义引人注目：如果你希望一项研究的精度翻倍，双倍的努力是不够的，必须增加到 4 倍的努力。学到更多的代价比通常的

看法要费力很多。雅各布·伯努利以名著《猜测术》（*Ars conjectandi*）封笔，他在其中提到了自己的发现，实际上要做 26 000 次试验才能获得他认为的可接受精度水平。那时候大家还不知道根号 n 规则，而且他也不可能知道自己寻找的精度水平无法实际实现。随着时间的流逝，统计学家不得不接受现实，满足于较少的精度，并相应调整自己的期望，同时继续寻找对误差或方差累积的更好理解。这和长期以来的数学实践形成了直接的对比。在数学运算的序列中，数学家们会记录每一个步骤引起的最大误差，这个量会随着序列的增长而增大。而统计学家们可以接受一个可能的误差补偿，相对来说，这个补偿会随着序列的增长而收缩。

2.3 优化、扩展、悖论

这个规则到 19 世纪中期得到了改进。英国天文学家乔治·艾里 1861 年出版了一本小教材，题为《观测误差与观测组合的代数与数值理论》，其中一节叫作“纠缠的观测”。“纠缠的”意思是说几个观测有共同的部分，因此，正如我们现在所说，是相关的。艾里确实展示了这种关系对估计方差的影响，更深地理解了相关性与数据信息量的关系。

美国哲学家、博学者查尔士·皮尔士在这个方向上走得更远。他在 1879 年发表了一份简短的笔记，内容关于他所谓的“经济研究理论”。皮尔士这样描述自己的目标，“一般来说，经济原则处理效用和成本之间的关系。经济学的研究分支关心的是，减少可能的知识偏差的成本效应关系。主要问题是，给定了金钱、时间和能量的花费，如何使知识得到最有价值的增加。”

皮尔士将其视为一个效用理论的问题：考虑标准差不同的两个实

验，都是艾里考虑过的混合类型（本质上是方差成分模型），它们两个都提供了要紧的信息，关键是你如何优化自己的努力。在一个测量重力的可倒摆试验的具体例子中，你该如何分配摆锤向上停顿和向下停顿之间的时间？这是一个优化问题，很明显，其中的优化规则度量了相关观测的获取信息。皮尔士认为，这个试验需要以相同时间周期针对每个位置进行，而且，试验的持续时间还需要“与支撑点到质心的距离成比例”。他以这样的忠告作为笔记的结束语：“需要注意，这里给出的理论依据假定了研究目的在于发现真相。而如果一项研究是出于实现个人荣誉的目的来做的话，这个问题的经济学就会完全不同。但这也可以很好地理解那些人为什么会投入这种研究类型。”这种评论讽刺的对象大概不难对号入座。

无论如何，直到 20 世纪才明确建立这种思想：数据中的信息可以度量，其精度在某种程度上与数据的数量有关，在某些情况下能够精确处理。不过并不能预设这种思想不曾受到挑战。可以想象，许多人仍然相信第二组 20 个观测与前 20 个至少一样有价值。但一些权威还有另一个有趣的说法，这个说法更为极端地走向了相反的观点。它认为，如果你有两个观测，那么舍弃一个要好于取两者的均值！而且更糟糕的是，这种观点居然正确！

牛津的逻辑学家约翰·维恩在 1878 年的《普林斯顿评论》上发表了一篇文章，设想了这样的情况：一位舰长计划夺取敌人的一座要塞，他派了两名间谍潜入要塞，并要求其返回报告要塞中加农炮的口径，如此就可以准备尺寸合适的加农炮弹，以确保夺取要塞后可以加强防守。一名间谍报告口径是 8 英寸，另一名报告是 9 英寸，那么舰长应该配置 8.5 英寸的加农炮弹吗？当然不会，无论哪种情况，这个炮弹都不能用。哪怕扔硬币决定取两种尺寸中的某一种，都好过注定失败的平均值。

问题在于标准化分析,这种分析隐藏于本章所有其他例子的背后,都默认了假定精度的恰当度量是均方根误差，或是它常用的替代方式——估计的标准差。如果观测是围绕目标值的正态分布，那么精度的所有合理度量都会一致，并与均方根误差相一致。但维恩的例子并不属于这种类型。他的例子中，残差超过这个闭区间时，估计的惩罚不增加，因此这个区间越紧密，产生的成本会越高。对维恩来说，合适的度量是估计值落在目标旁非常小的数字 ε 邻域内的概率。用于估计口径 C 的估计值 E 的选择是要最大化 $\text{Prob}\{|E - C| \leqslant \varepsilon\}$。弗朗西斯・埃奇沃思对此表示赞同，在 1883 年的一份简短笔记中，他表示，这种解决方案“扔掉了一个随机选择的观测”，除了维恩离散案例的某些情况，一般优于均值。对于那些表现出矛盾的误差分布，甚至一些所有矩都有限的连续单峰密度的误差分布，埃奇沃思给出了一类例子。与通常的情况相比，它们在众数位置更尖，但也不总是如此。很明显，信息的度量需要关注研究目的。

在 20 世纪其他更平凡的例子中，根号 n 规则的失效吸引了关注。有一类是时间序列模型，其中的序列相关性产生了明显的模式，欺骗了数据分析者，除非你意识到序列相关减少了有效的样本规模，使其远低于画出的点数。这会导致错误地发现周期性。如果只研究了有限的几次循环，即使非周期的机制也会表现出周期性。20 世纪 80 年代，两位著名的地理学家发现了地球上小型海洋生命灭绝率中一个 2600 万年周期的证据。如果这个证据成立，它预示存在着地球之外的某些原因。有一种假设是，我们的太阳有一个伴星，现在观察不到它，但每隔 2600 万年就向我们发出辐射。

这个假设掀起了人们的狂热兴趣，还登上了《时代周刊》的封面，热切的科学家们立刻投入到寻找其他统计证据的工作中。“寻找就能找到”的观点在不受控制的数据分析中发挥了巨大作用，许多论文得到

发表，声称有一个类似于地球磁场的逆转周期，和其他尚不清晰的周期。最后，人们证明，在灭绝率中发现的第一个信号确实稍有周期性，但原因是数据的人工处理，而不是一颗路过的死星。这个数据为过去2 亿 5000 万年间地质纪元的测定提供了线索。对这一半时期来说，纪元已经很好地确定下来，但不包括时间。这个时间线的提出者只好将这一段 1 亿 2500 万年的时间划分为 20 个子区间，每个子区间的平均长度为 625 万年。但端点的分数表达会夸大精度，因此他们把这个周期分割成长度为 6、6、6、7、6、6、6、7、6、6、6、7、6、6、6、7、6、6、6、7 的片段。这个人工周期在分析中发挥了作用，给出了周期性的表象，引爆了最初的激情。

尽管所有测量的精度都相同，但最后 10 个测量的价值远不及前10个，这就是信息积累的悖论，并因为统计和科学中“信息”术语的不同使用方式而增强（还有某种程度的误导）。一个合适的例子是统计理论中的术语“费舍尔信息量”。在参数估计问题的最简单形式中，费舍尔信息量 $I(\theta)$是得分函数平方的期望值，定义为数据概率密度函数对数的导数，即 $I(\theta) = E[\mathrm{d}\log f(X_1, X_2, \cdots, X_n; \theta)/\mathrm{d}\theta]^2$。（直观上，如果数据概率迅速地随着 θ 改变，这个导数趋于变大，那么概率对于 θ 越敏感，数据中的信息就越丰富。）这是一种非凡的统计构造，在非常广泛的适用条件下，它的倒数给出了所能找到的方差的最佳估计。对于这些情况下通过聚合可以获得的结果，它设定了一个黄金标准。但从评估信息积累的角度看，它有误导，因为它使用了错误的单位表达——一个平方单位的尺度。这是一种附加的度量，并且可以说，等长的数据片段具有等量的信息。费舍尔信息量和根号 n 规则是一致的，使用时只要取平方根即可。

20 世纪 40 年代，克劳德·香农提出了另一种信息的附加度量。香农关心一个相当不同的问题——编码和信号处理问题。在这个问题

里，他把下面这一点看作不证自明的：因为传送的信号没有限制，并且在传递过程中经过编码成为等信息量的，因此信息度量是可加的。统计学家考虑的自然和人文科学问题中，可加性仅在数据集规模非线性的尺度上成立。

评估信息积累是一项相当复杂的任务，但通过仔细关注相关性和科学目标，数据的信息度量——不同信息集中的比较信息和随着数据增长而产生的信息增长率——已经成为统计学的一根支柱。

第 3 章

似然：概率尺度上的校准

没有背景的测量只是一个数字，没有意义的数字。背景提供了尺度，可帮助校准，并允许比较。当然，我们每天都会看到没有背景的数字，比如报纸上的一个常见栏目叫作“数字事实”，会列举一些试图震惊或娱乐读者的例子。即便高高在上的《科学》杂志也使用过这种手段，有一个条目（2011 年 8 月 5 日）这样说：

> 42 000 根据 *PLoS ONE* 上的一项研究，世界上有 42 000 名儿童死于腹腔疾病。

表面上看，这个统计数字令人不安。但果真如此吗？多长时间的？一周、一年还是 20 年？这个数字是大还是小？毕竟世界上有 70 亿人口，其中儿童大约有 20 亿。有全球儿童的其他死亡原因的数字可参照吗？不同国家的疾病流行程度一样吗？而且，42 000 是一个很整的数字，肯定不精确，那么可能的误差是多少？10%还是 50%？可能是 100%吗？进一步研究 *PLoS ONE* 可以找到一些答案，也能发现有点乱。给出的数字是年度数字，但仅仅是根据一个数学模型推测出来的数字，这个模型是想估计未确诊人数等其他数据的。并且，这个条目所引的

文章中没说“有 42 000 人死亡”，而是说“42 000 人可能死亡”，这两种说法非常不同。这个数字也不是基于数据给出的，事实上，*PLoS ONE* 的文章提醒过读者，当前“严重缺乏有全球代表性的流行病学数据”。它通过模型试探发现了一个取值范围（±15%），但没有说明模型失败导致的范围。*PLoS ONE* 给我们提供了背景，但《科学》不仅没有这么做，而且还严重地误导了我们。

测量只有用于比较才是有用的。背景提供了比较的基础，或是一条基线、一个基准，或是一组用于相互比较的测量。有时基线是隐含的、基于常识的，比如报告当天天气时，会自然地联系到本地的知识和过去的经验。但更常见的情形——比如源于腹腔疾病的儿童死亡情况——缺乏一般的常识。并且无论如何，科学需要更多基线：真实的数据、明确的来源，以及衡量差异水平的测量尺度。这种差异是显著的还是不显著的？

常规物理测量的最早的一个例子，是第 2 章讨论过的铸币检验试验。该试验中，即使从最开始算起——大约公元 1100 年——也由契约给出了重量的基线，这是一种契约基线。试验托板给出了金属成色方面的基线，一个托板样品现在还保存在伦敦塔中。铸币检验试验也有一个评价差异的尺度——“公差”，给出了我们今天称为“容忍度”的概念。它通过谈判得出，但没有迹象表明，它是以任何正式的方式从数据导出的，或来自于铸币过程变化性的正式评价。并且，我们注意到，有一种缺陷妨碍了它的运用。

3.1 阿布斯诺特和显著性检验

现代统计学使用“概率度量”作为差异评价的基本组成部分，它通常按照统计检验的形式，其根源可以追溯到几个世纪之前。检验的

结构问题看起来简单而直接：对一种理论或者假设，手边的数据是支持它还是否定它?“似然”的概念是回答这个问题的关键，因此它不可避免地涉及了统计检验的构造。检验的问题会通过比较不同假设下的数据概率进行回答。最早的例子中只计算一个概率，并隐晦地进行比较。

约翰·阿布斯诺特作为一位极具鼓动性的作家而家喻户晓，他1712年出版了一部讽刺文学作品，题为 *Law Is a Bottomless Pit*。其中，阿布斯诺特创造了“约翰牛”这个典型的英国人形象。阿布斯诺特是乔纳森·斯威夫特和亚历山大·蒲柏的密友。蒲柏给他的朋友写过一封信，也是一篇著名的讽刺文章——《致阿布斯诺特的一封信》。他在信中批评约瑟夫·爱迪生时，引入了短语“明褒实贬”（damn with faint praise）。阿布斯诺特也接受过数学和医学的训练（1705~1714 年作为私人医生为安妮女王服务），作为数学家，他有过两个著名的贡献。第一个贡献是 1692 年出版的概率小册子，大部分是克里斯蒂安·惠更斯1657 年出版的拉丁文小册子的翻译，但这本书是该主题的最早的英文出版物之一。第二个贡献是他于 1710 年在伦敦皇家学会宣读的一份简短笔记，随后发表在学会的《学报》上。笔记标题是“对于神圣天意的一个论点——来自两性出生率恒定规律的观察”。直到今天，这份笔记依然作为显著性检验的早期经典例子被大量引用。

阿布斯诺特认为，现在观察到的男性（M）和女性（F）人数的平衡，原因不是概率，而必然是神圣天意的结果。原因有二：第一，用数学语言来说，如果像投掷一枚公平的双边骰子那样设计性别，那么极不可能得到精确的平衡结果（或者即使是非常接近的平衡）。他计算了两个人的性别精确平衡的概率（即两个人成为 MF 或 FM 的概率，就是 $1/4 + 1/4 = 1/2$），6 个人的概率是 $20/64 = 0.3125$，以及 10 个人的是 $63/256 < 1/4$。他同时表示，通过对数可以将这种计算运用于人数很

大的情形。很明显，结果会得到一个非常小的概率。所有这些计算都正确，将一枚公平的硬币投掷 $2n$ 次，正反面个数会恰好平衡的机会近似为 $c/\sqrt{n}$ ，其中 $c=\sqrt{(2/\pi)}=0.8$ ，如表 3-1 所示。

阿布斯诺特使用较低的精度，并且声称，即使将平衡的定义从精确平衡放宽到近似平衡，这种概率依然很小。在那样的情况下，“近似（approximate）意味着什么”这样的问题就非常关键，而且做这个计算需要的数学在未来若干年之后才出现。但无论如何，他表达的这第二个观点终于使其留名青史。

表 3-1 硬币投掷次数与精准平衡的概率关系

投掷次数	精确平衡的概率
2	0.50
6	0.31
10	0.25
100	0.08
1000	0.025
10 000	0.008

在长达 82 年的死亡率报表（bills of mortality）中，阿布斯诺特检查了出生男性相对于出生女性的超额数量（如图 3-1 所示）。他发现，这样一种运行机制会发生的概率只有 $1:2^{82}$，小到可以忽略：

1/4 836 000 000 000 000 000 000 000。

这里的“随机分布”——即对每种情况独立地指派相等概率的性别——与神圣天意的结果比较后，这种假设对这些数据中男性多于女性给出了更大的概率。为什么呢？因为存在这样一些观点：考虑到“（那些必须冒险寻找食物的）男性主体面临着外部的意外”，“大自然极富远见，它的创造者聪明又睿智，经过他的处置，创造出的男性比女性更多，而且比例接近常数”。阿布斯诺特没有计算这种选择。

Christened.			Christened.		
Anno.	Males.	Females.	Anno.	Males.	Females.
1629	5218	4683	1648	3363	3181
30	4858	4457	49	3079	2746
31	4422	4102	50	2890	2722
32	4994	4590	51	3231	2840
33	5158	4839	52	3220	2908
34	5035	4820	53	3196	2959
35	5106	4928	54	3441	3179
36	4917	4605	55	3655	3349
37	4703	4457	56	3668	3382
38	5359	4952	57	3396	3289
39	5366	4784	58	3157	3013
40	5518	5332	59	3209	2781
41	5470	5200	60	3724	3247
42	5460	4910	61	4748	4107
43	4793	4617	62	5216	4803
44	4107	3997	63	5411	4881
45	4047	3919	64	6041	5681
46	3768	3395	65	5114	4858
47	3796	3536	66	4678	4319

B b　　Christened.

Christened.			Christened.		
Anno.	Males.	Females.	Anno.	Males.	Females.
1667	5616	5322	1689	7604	7167
68	6073	5560	90	7909	7302
69	6506	5829	91	7662	7392
70	6278	5719	92	7602	7316
71	6449	6061	93	7676	7483
72	6443	6120	94	6985	6647
73	6073	5822	95	7263	6713
74	6113	5738	96	7632	7229
75	6058	5717	97	8062	7767
76	6552	5847	98	8426	7626
77	6423	6203	99	7911	7452
78	6568	6033	1700	7578	7061
79	6247	6041	1701	8102	7514
80	6548	6299	1702	8031	7656
81	6822	6533	1703	7765	7683
82	6909	6744	1704	6113	5738
83	7577	7158	1705	8366	7779
84	7575	7127	1706	7952	7417
85	7484	7246	1707	8379	7687
86	7575	7119	1708	8239	7623
87	7737	7214	1709	7840	7380
88	7487	7101	1710	7640	7288

图 3-1　阿布斯诺特的数据（Arbuthnot 1710）

相似地，丹尼尔·伯努利发表于 1735 年的一份有奖征文中，考察了当时其他 5 个已知行星的轨道平面与地球轨道平面之间令人惊讶的接近程度。这 6 个轨道平面并非完美一致，但它们都落在一个很小的角度差异中：轨道平面的相互倾斜都在 6°54′之内。伯努利判断，在随

机分布的假设下，这种接近的一致性极不可能被接受。他的一个计算中，将 6°54′取作近似 90°的 1/13。并且，他判断其他 5 颗行星位于一个给定的、包含地球轨道的 6°54′区间内的概率是$(1/13)^5 = 1/371\ 293$。伯努利认为，这给出了囊括所有轨道平面需要的最小角度。

阿布斯诺特和伯努利都对数据集设置了概率尺度，特别在本质上都使用了后来罗纳德 · A. 费舍尔逻辑分明地提出的原则："这个推断的支持力量在逻辑上出于一种简单的分离：要么发生了一个很少见的机会，要么随机分布的理论不是真的。"如果数据不是随机分布的结果，那么某些其他的规则必然起作用。阿布斯诺特和伯努利的例子中，似然的比较是大家默认的。至少一个其他假设（神圣天意或牛顿力学）会导致观测数据发现的概率会高于在"概率"假设之下发现的概率，大家认为这是理所当然的。

比较问题还简单时，只需要考虑两个不同的概率，答案也比较简单：计算一个概率，如果它非常小，就推断另一个概率。乍一看，阿布斯诺特和伯努利的问题就属于这种类型,但即使如此也出现了困难。阿布斯诺特认为平衡无需精确，只需近似时，发生了他的第一个讨论。这带来的问题是，近似的程度怎样才足够好？因此，他转向伦敦的出生数据。根据这些数据，他发现了 82 年间出生的男性多于女性，并就此计算了一个概率，给出了"其他"的推断。阿布斯诺特的计算具备了现代检验的一些要素，但它仅仅处理了一种极端情况：在可获得的 82 年样本中的每一年，出生的男性都超过女性。如果 82 年中只有 81 年发生这种超额,他会做什么？他会在一个概率假设下评估 82 年中有 81 年发生的精确概率吗？又或者（就像现代检验通常会做的），他会发现 82 年中至少发生 81 年的概率吗？这两个概率都很小，但是更中间的情况如何？比如 82 年中发生 60 年，或者 82 年中有 48 年出生的男性超过女性的情况，其中不同的方法会产生非常不同的答案吗？他

会怎样做，我们不得而知。

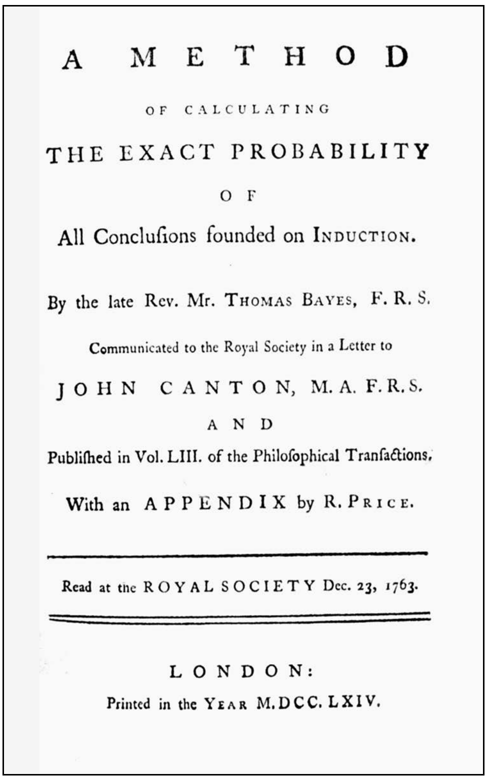
如果数据记录在连续的尺度（或近似连续的尺度）上，这个问题会更严重。所以在最合理的假设之下，每个数据值都有一个小概率。出生性别可能性均等的总体中，每种性别决定都独立时，如果有 1 000 000 人出生，那么出生的男性个数多于女性的概率不可能大于 1/1000。这是否意味着我们拒绝了一种自然随机平衡的假设，即对每种性别数据显示一种完美相等的个体数量？很明显，单个的概率计算并不是所有问题的答案。概率本身是一种度量，因此需要一个比较的基础。显然，也需要在可允许的假设上施加一些约束，否则这种“数据天定”的自我实现的假设会对任何数据集给出为 1 的概率。

3.2　休谟、普莱斯和贝叶斯归纳

并非所有似然观点都以数值精确表示。一个著名的例子是大卫·休谟对基督教神学某些基本原则发表的看法。1748 年，休谟发表了论文《论神迹》。这篇文章其实早已完成，但他希望该文能产生轰动，因此推迟了发表。休谟主张，不应该信任那些以基督复活作为主要例子所报告的神迹。他把神迹描述为“自然法则的违逆”，因此更是极不可能发生的。事实上，这些基督复活的神迹也的确不可能，报告者要么撒了谎，要么搞错了，相比之下，那些个不准确的神迹报告倒像是更有可能的。

休谟准确预料到会有一场论战，但没能预计到其中一种回应的数学特征。那时，也许是作为对休谟的回应，托马斯·贝叶斯至少写出了他的著名论文中的精彩部分。无论如何，在 1764 年早期，理查德·普莱斯考虑出版贝叶斯的论文时，毫不犹豫地认为论文的目的是回应休谟的论文。直到最近大家才注意到，普莱斯为贝叶斯的论文拟定的

标题（很可能也是贝叶斯的意思）——《建立在归纳基础上计算所有推断的精确概率的一种方法》（如图 3-2 所示）。这个标题的野心太大，还没有文本可以完全证明。这篇文章对此问题给出了一种数学处理：如果一个事件在 n 次独立试验的每次都以未知的概率 p 发生，并且发现它发生了 x 次，那么在 p 的所有值等可能的先验假设之下，寻找 p 的后验分布。这是贝叶斯定理首次出现的特殊形式，并且，正如普莱斯下一本重要出版物所表明的，他是冲着休谟来的。


A METHOD
OF CALCULATING
THE EXACT PROBABILITY
OF
All Concluſions founded on INDUCTION.

By the late Rev. Mr. THOMAS BAYES, F. R. S.

Communicated to the Royal Society in a Letter to
JOHN CANTON, M. A. F. R. S.
AND
Publiſhed in Vol. LIII. of the Philoſophical Tranſactions,

With an APPENDIX by R. PRICE.

Read at the ROYAL SOCIETY Dec. 23, 1763.

LONDON:
Printed in the YEAR M.DCC.LXIV.

图 3-2　普莱斯选定的贝叶斯论文单行本标题页（Watson 2013）

1767 年，普莱斯的书 *Four Dissertations* 问世，书中一部分在一个很有煽动性的标题下明确引用了贝叶斯的论文，直接反驳了休谟。它包括了普莱斯的贝叶斯定理应用（贝叶斯论文以外的最早应用），一个详尽计算表明了反对那些被视为自然法则的内容——并不像休谟所主张的那样。休谟坚持认为，既然神迹的存在仅仅是基于经验的，那么保卫神迹也同样需要基于经验。假定支持自然法则的是相同的事件（例如涨潮或者每天的日出），无一例外地接连发生了 1 000 000 次。将其视为一个观测次数 $n = 1\ 000\ 000$ 的二项试验，并且发现神奇的例外个数是 $X = 0$。这隐含了“p = 下一次神迹的概率是零”吗？不是的。普莱斯使用贝叶斯定理对此做了计算，在这种情况下，神迹发生概率大于 1/1 600 000 的条件概率是 $\text{Prob}\{p > 1/1\ 600\ 000 | X = 0\} = 0.5353$，好于 50%的概率。不可否认，1/1 600 000 确实太小了，但还没有到不可能的地步。反之，将其作为单次试验中神迹发生的概率时，普莱斯发现，接下来的 1 000 000 次试验中至少发生一次神迹的概率为

$$1.0 - (1\ 599\ 999/1\ 600\ 000)^{1\ 000\ 000} = 0.465$$

概率接近一半！神迹发生的可能性远远大于休谟的设想。

贝叶斯的文章出版后基本上受到了半个世纪之久的忽略，毫无疑问，这要归咎于印刷刊物上出现的乏味标题（《求解机遇理论的一个问题》）。直到 20 世纪，后验概率才在校准推断中开始发挥重要作用，这个大胆的标题才走向现实，但当时还不可能。我们会在第 5 章再次讨论这个主题。

3.3　拉普拉斯检验

整个 19 世纪不断出现大量可称为“显著性概率”的特别计算。这些计算通常遵循了丹尼尔·伯努利的计算思想，使用数据定义一组值

的极限，再在某些随机性的假设之下找到这组值的概率。

1827 年，皮埃尔 · 西蒙 · 拉普拉斯在巴黎天文台看到了一个长序列的气压表读数，是为大气中的太阴潮提供证据而记录的。他发现了一个 $x = 0.031758$ 的效应。接着，因为缺少任何实际效应，拉普拉斯计算了发生一个效应（按绝对值）不大于 0.3617 的概率。这会对应到一个现代的 $1-0.3617=0.6383$ 的双边 P 值，他还给出判断，0.3617 太小（即 P 值太大），不足以支持潮汐的存在。拉普拉斯写道：

> 如果这个概率（0.3617）非常接近 1，它会以极大的似然表明 x 的值并不单单受到概率的不规则性影响，也会受到一种恒定原因的部分影响，这种原因只能是月亮对大气的作用。但这个概率和表示为 1 的确定性之间的差异相当大，这就意味着，尽管运用的观测个数非常大，这种作用依然仅仅表明了很弱的似然。因此，可以把它在巴黎的这种可察觉的存在性看作不确定的。

拉普拉斯的解释经受住了时间的考验。太阴潮在巴黎的影响太弱，无法用可获得的观测进行检测。相比之下，他能找到气压计变化上的季节影响的证据（上午 9 点到下午 3 点的气压平均改变）。这里，拉普拉斯给出了现代的 P 值，同时阐述称：没有任何季节影响的情况下，可以计算出，这种规模的或者规模更大一些的偏差发生的概率是 0.0000015815。这太小了，不能归结为机遇。

1840 年，朱尔斯 · 加瓦雷特用“合法”出生的性别比与“非法”出生性别比进行了比较，男性的比例分别是 0.51697 和 0.50980，相差 0.00717。出生的数目很大（合法 1 817 572 人，非法 140 566 人，如图 3-3 所示），并且他遵循了一条借鉴泊松思想的指导原则——用 0.00391 和这个差做比较，发现了我们今天描述的 $2\sqrt{2}$（$=2.828$）乘以差异的

估计的标准差，相当于一个以概率 0.0046 偶然发生的绝对偏差。因为观测到的差异接近这个阈值的 2 倍，加瓦雷特对此解释，这表明离差大于可归属于试验变化的部分。我们当然想提醒他，这个检验不能告诉我们，差异是缘于社会因素还是生物因素。

1824-1825.

Enfants légitimes.

$m = 939641$ = le nombre de garçons.
$n = 877931$ = le nombre des filles.
$\mu = 1817572$ = le nombre des naissances.

D'où résulte que la chance moyenne de naissance d'un garçon en France dans l'état de mariage, est représentée par le rapport

$$\frac{m}{\mu} = \frac{939641}{1817572} = 0{,}51697$$

En poussant l'approximation jusqu'à la cinquième décimale.

Enfants illégitimes.

$m' = 71661$ = le nombre des garçons.
$n' = 68905$ = le nombre des filles.
$\mu' = 140566$ = le nombre des naissances.

D'où résulte que la chance moyenne de naissance d'un garçon en France hors l'état de mariage, est représentée par le rapport

$$\frac{m'}{\mu'} = \frac{71661}{140566} = 0{,}50980$$

En poussant l'approximation jusqu'à la cinquième décimale.

图 3-3　加瓦雷特的出生数据（Gavarret 1840，274）

1860 年，美国天文学家西蒙·纽康从新的视角重新思考了一个老问题。比如昴星团的情况，“第五星等的 6 颗亮星在天球的单个小正方形（1°正方形）内被发现”应该是异常现象吗？或者说，即使这些星星在苍穹中随机散布，我们可以认为这会以合理的概率发生吗？可见星的亮度按等级确定，其中最暗淡的星星分类为第六星等，亮度超过一个范围的星星来自第五星等，以此类推。一个好的近似估算是，第五或更高的星等有 $N = 1500$ 颗已知的星星，而天球共包括 41 253 平方度。于是 p = 单颗随机的星星会落在天球任意平方度的概率是 1/41 253。纽康分析的原创之处在于，将星星的分布当作一个泊松过程，以 $\lambda = Np = 1500/41\,253 = 0.0363$ 作为空间过程的比率，即每平方度星星的期望数目。那么，在一个平方度中发现 s 颗星星的概率是

$$e^{-\lambda}\lambda^{s}/s!$$

对于 $s=6$，这给出的概率是 0.000000000003。因为这涉及的只是单个指定的平方度，而选择关注昴星团是因为它的平方度最紧密。纽康知道这不是个合适的概率，所以相反地，他找到了 41 253 个平方度上会成为六等星的所在地的期望个数。也就是说，用 41 253 乘以这个小概率 0.00000013，结果依然是一个微乎其微的小数字。事实上，他也知道这并非一个正确的数字，而需要的是这样的概率，为了能够包括最多的星星而允许平方度小幅调整。但纽康认为，他没能计算出的那个答案不会非常大。他确实注意到，为了使包含六等星的区域的预期数为 1，需要把目标空间从 1 平方度扩展到 27.5 平方度。

3.4 似然理论

在我已经给出的例子里可以看到一种日益明显的诡辩，但这段时期也开始了一种更正式的理论发展。18 世纪中期，一些人开始用数学问题表示观测的组合和误差的分析。他们中的某些人，包括托马斯·辛普森（1757，如图 3-4 所示）、约翰·海因里希·朗伯（1760，如图 3-5 所示）、约瑟夫·路易斯·拉格朗日（1769）、丹尼尔·伯努利（1769，1776，如图 3-6 所示）、皮埃尔·西蒙·拉普拉斯（1774 及以后，如图 3-7 所示）以及卡尔·弗里德里希·高斯（1809）等，都描述了一个对称的单峰误差曲线或者密度。作为其分析的一部分，他们寻求选择一个“最有可能”概括符合心中曲线的数据。

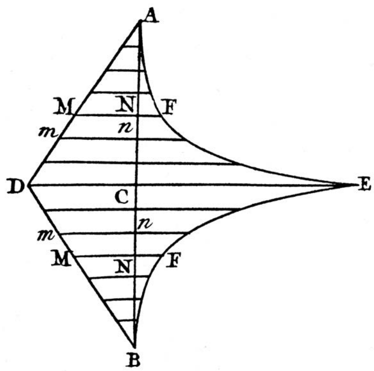


图 3-4　辛普森 1757 年的曲线（垂线左侧的对称三角形），AB 右侧的曲线意图表示 6 个观测的平均值的密度，但这是根据他的想象画出的，并不精确（Simpson 1757）

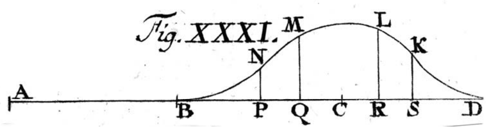


图 3-5　朗伯 1760 年的曲线（Lambert 1760）

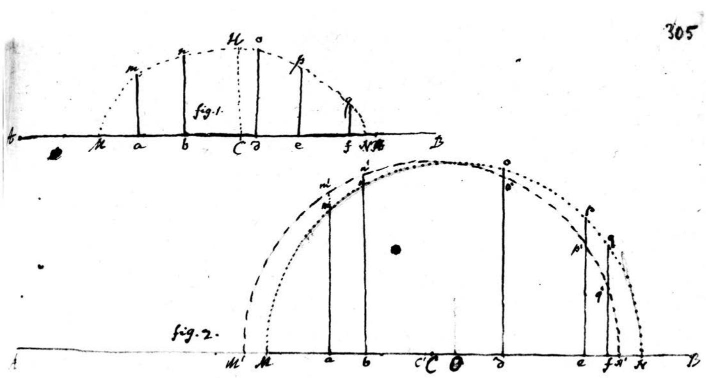


图 3-6　伯努利 1769 年绘制的曲线，其中 figure 1 使用的权重基于 figure 2 的曲线（Bernoulli 1769）

这些早期分析中的一部分可以视为我们今天所谓的“极大似然估计”的先驱。这些理论变得越来越精细：拉普拉斯提出，后验中位数以尽量缩小后验期望误差。高斯在这个问题的第一个工作中采用了贝叶斯方法，给出一个平坦的先验；而当误差是正态分布时，给出了“最有可能的”答案，这引出了最小二乘法（没有概率内容的最小二乘法已在 4 年前由勒让德发表）。但 20 世纪以前尚未出现似然的完整理论。

20 世纪 20 年代，在卡尔·皮尔逊的一些早期工作的基础上（包括 1900 年皮尔逊对卡方检验的有影响力的介绍），费舍尔宣布了一种相当大胆又可以理解的理论：如果 θ 代表科学的目标，X 代表数据，两者都可以是多维的，那么似然函数可以定义为 θ 函数的观测数据 X 的概率或者概率密度。我们根据习惯，将这个记法中的 X 记为固定的观测，写作 $L(\theta) = L(\theta \mid X)$。他会取能使 $L(\theta)$最大化的 θ，在某种意义上，这个值是在所有看起来可能的 θ 中，使观测数据 X 最有可能发生的值。并且，费舍尔描述了这种选择方法，称它为“θ 的极大似然估计”。到此，除了术语以外，他和丹尼尔·伯努利、朗伯以及高斯都是一致的。但费舍尔还声称，当最大值作为一个光滑函数的最大值，并通过对 θ 求导且设置求导结果为 0 而求解时，可以发现精度（估计的标准差）是来自 L 最大值点的曲率（二阶导）的优良近似。我们还可以发现，这个估计表达了数据提供的所有相关信息，也不可能再通过其他的一致估计方法改进性能。如此，它就是所有统计学家期盼的答案：理论上最优，程序简单，而且不费什么代价就可以得到精度的完整描述。

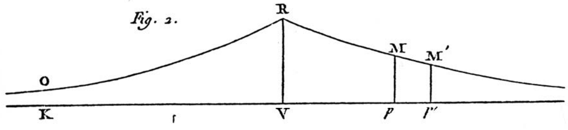


图 3-7 拉普拉斯 1774 年的曲线，现在称为“双指数密度”（Laplace 1774）

费舍尔程序的最初思想在应用中没有这么一般化，也没有这么严密和完整。严格证明太困难，因此借助了一些反例。反例中的大多数是病态的——例子真实，但对实践没有困扰。一个显然的例外是毛里斯·巴特莱特在 1937 年和亚伯拉罕·瓦尔德在 1938 年分别发现的（同年，他还与杰日·奈曼就此进行了通信），奈曼和伊丽莎白·斯科特在 10 年后发表了这个反例的简单版本，基本内容如下。假设你的数据包含 n 个独立的正态分布的数据对(X_i, Y_i)，其中每一对 X 和 Y 是独立的测量，期望 μ_i 相同，但所有 X 和 Y 都有相同的方差 σ^2。于是，有 $n+1$ 个需要估计的量。μ_i 的最大似然估计是数据对的均值，$(X_i+Y_i)/2$，而且方差 σ^2 的最大似然估计是 $\Sigma(X_i-Y_i)^2/4n$，一个期望为 $\sigma^2/2$ 的估计量仅仅是它应该成为的数值的一半。困难出现了，因为这是对每两个样本的 n 个个别的方差估计的平均。正态情况下，样本大小为 m 的方差的极大似然估计是有偏差的，等于方差乘以$(m-1)/m$。如果 m 很大，这个结果会接近 1；但对于 $m=2$，则是 1/2。现在把它作为一个大数据问题考虑，其中记录的数据个数大约等于目标的数量。全部样本中的信息必须分散在大量的目标上，而且不能把任务的各部分都做好。这个例子可以视为最大似然估计在大数据应用上的困难或挑战：它确实能预期均值，而且容易补偿方差的问题（只要乘以 2），但这也确实表示处理高维问题时要小心。

尽管有这些周折，但费舍尔程序不仅为这个世纪的其余大部分时间设定了研究议程，而且他支持的似然方法或与其非常相关的其他方法，在许多可行的领域也占据了主导的实践地位。

虽然费舍尔做了许多显著性检验的应用，但他表达成原假设的检验，并没有明确的备选假设。这就轮到了奈曼和伊冈·皮尔逊，他们基于似然的直接比较以及备选假设的明确介绍，发展出假设检验的一种形式化理论。检验的思想——无论在费舍尔的意义上还是在奈曼和

皮尔逊的意义上——显然已经有了巨大影响。它的应用相当普遍，还在某些方面或某些运用方式上受到了批评、引起了关注，比如不加批判地接受和常规使用的 5%的水平。这些都证明了它的巨大影响。与似然相联系的思想作为一种方法校准推断——在统计背景中放入数据变差和我们会放置在观测差异上的置信度——已经成为现代统计的支柱。

第 4 章

相互比较：作为标准的样本内变异

第四根支柱——相互比较——指这样一种思想：统计比较可以严格遵循数据的内部变异进行，而无须参考甚至依赖外部准则。这种思想的大意很古老，但我能想到的精确表述直到 1875 年才出现，由弗朗西斯·高尔顿提出。使这种思想发扬光大并成为统计学主流的扩展工作，则发生在高尔顿的文章发表之后的 10 年、33 年和 50 年，分别由弗朗西斯·埃奇沃思、威廉·希利·戈塞特和罗纳德·A. 费舍尔完成。

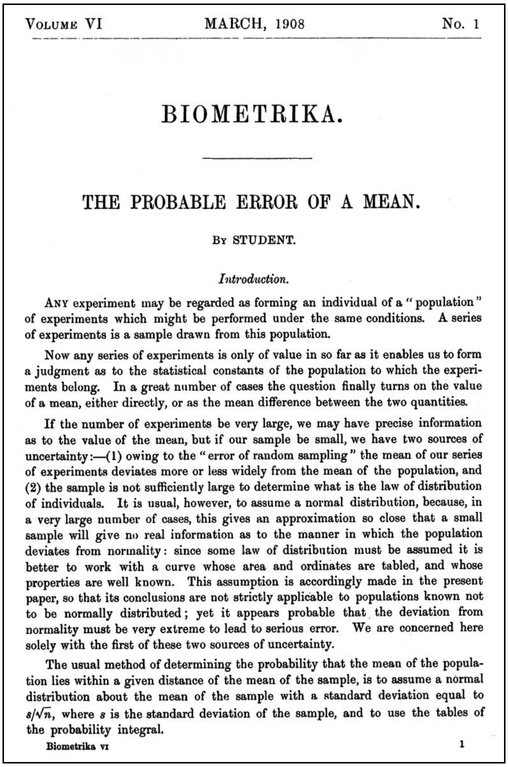
1875 年发表的文章《相互比较的统计》中，高尔顿给出了一种有几个理想属性的比较方法，其中包括，在比较时，“我们可以省去参考标准，而可以通过共同接受的说法创建并间接定义它们。……（它们）完全受到**相互比较**影响，不需要借助任何外部标准”。这个定义适用于后来发展出的概念，但高尔顿自己的应用局限于百分位数的使用，特别是（但不完全是）中位数和两个四分位数。这些方法只要通过排序数据就可以完成，不需要比计数更复杂的算术计算。而且，即使在某些情况下，测量是描述性的、排序的而非数值时，百分位数的表现也很好。其实早在 1896 年，高尔顿就第一次使用了百分位数。在其著作

Hereditary Genius 中，高尔顿使用了几组生物学中的专门术语排列并比较群体中的天才，而没有用到任何有关天才的数值测量。这本书至今一点名气也没有，但它的统计方法完全正确。

4.1 戈塞特和费舍尔的 *t*-检验

带着历史的“后见之明”，我们可以说，相互比较的数学深入应用的第一粒种子，在 1908 年由一位看似不可能的父亲播下。1899 年起，戈塞特受到都柏林的吉尼斯公司雇用，成为一名化学技师。他曾在牛津的新学院接受数学训练（1897 年获得数学学科第一等学位）和化学训练（1899 年获得第一等学位），而且很快看出了统计对啤酒厂的巨大作用。1904~1905 年，他阅读了伦敦大学学院的卡尔·皮尔逊实验室的最新工作报告，总结了误差理论和相关系数的使用，写下了一组内部备忘录（实际上是内部指导文本）。这些备忘录的第一份中，戈塞特有一个陈述，表达了希望将 *P* 值与数据相联系的想法：“我们遇到一些困难，没有哪本书曾提过这种可能性（the odds）。它易于接受以建立任何结论，这件事去咨询某位数学物理学家对我们来说或许有益。”这位物理学家当然就是皮尔逊。

1906~1907 年，吉尼斯公司同意戈塞特离职，去皮尔逊的实验室访问两个学期，以学习更多知识。在那里，戈塞特写出了文章《均值的可能误差》。通过这篇文章，他树立了其统计学家的声望（如图 4-1 所示）。



VOLUME VI　　MARCH, 1908　　No. 1

BIOMETRIKA.

THE PROBABLE ERROR OF A MEAN.

BY STUDENT.

Introduction.

ANY experiment may be regarded as forming an individual of a "population" of experiments which might be performed under the same conditions. A series of experiments is a sample drawn from this population.

Now any series of experiments is only of value in so far as it enables us to form a judgment as to the statistical constants of the population to which the experiments belong. In a great number of cases the question finally turns on the value of a mean, either directly, or as the mean difference between the two quantities.

If the number of experiments be very large, we may have precise information as to the value of the mean, but if our sample be small, we have two sources of uncertainty:—(1) owing to the "error of random sampling" the mean of our series of experiments deviates more or less widely from the mean of the population, and (2) the sample is not sufficiently large to determine what is the law of distribution of individuals. It is usual, however, to assume a normal distribution, because, in a very large number of cases, this gives an approximation so close that a small sample will give no real information as to the manner in which the population deviates from normality: since some law of distribution must be assumed it is better to work with a curve whose area and ordinates are tabled, and whose properties are well known. This assumption is accordingly made in the present paper, so that its conclusions are not strictly applicable to populations known not to be normally distributed; yet it appears probable that the deviation from normality must be very extreme to lead to serious error. We are concerned here solely with the first of these two sources of uncertainty.

The usual method of determining the probability that the mean of the population lies within a given distance of the mean of the sample, is to assume a normal distribution about the mean of the sample with a standard deviation equal to $s/\sqrt{n}$, where s is the standard deviation of the sample, and to use the tables of the probability integral.

Biometrika VI　　1

图 4-1　1908 年论文的第一页，引入了会以 *t*-检验而闻名的“学生”（Gosset 1908）

1908 年，这篇文章以“学生”的笔名发表在皮尔逊的刊物 *Biometrika* 上。这反映了当时吉尼斯公司坚持的一项政策：员工的外部出版物不得标识他们的公司来源。此文没有突出这个方法在啤酒酿造质量控制中的潜在应用，而且，那时大家把它视为皮尔逊团队的一个不起眼的产品。除了其中一点之外，本文在其他方面的表现也确实平平。一个世纪以来，科学家们在天文学中例行公事一样地使用算术平均值，并根据“可能的误差”或简写 p.e.描述其准确性。对于正态分布的数据，

p.e.被定义为中位数误差。1893 年，皮尔逊引入了替代性的标尺“标准差”，记为 SD 或 σ，它与 p.e.成比例（$\text{p.e.} \approx 0.6745\sigma$）。而且，皮尔逊的方法很快成为常规方法。对于大样本，如果无法计算它的值，统计学家会毫不犹豫地采用 $\sqrt{\frac{1}{n}\sum(X_i-\bar{X})^2}$ 代替 σ（或者根据高斯的偏好，采用 $\sqrt{\frac{1}{n-1}\sum(X_i-\bar{X})^2}$）。戈塞特的那篇文章的目标是试图理解，样本不是很大且这些估计的准确性有限时，这种近似的不足需要什么样的补偿。他特别了解，各个 X_i 服从均值为 0 的正态分布时，$\bar{X}/\sigma$ 服从均值为 0 且标准差为 $1/\sqrt{n}$ 的正态分布。但是，σ 被 $\sqrt{\frac{1}{n}\sum(X_i-\bar{X})^2}$ 代替时，会发生什么？$z=\dfrac{\bar{X}}{\sqrt{\frac{1}{n}\sum(X_i-\bar{X})^2}}$ 的分布是什么？后来，费舍尔把这个标尺更改为现在我们都很熟悉的单样本 t 统计量，它们之间的联系是 $t=\sqrt{n-1}z$。

借助没有严格证明支持的出色猜测以及这些猜测产生的合理分析，戈塞特导出了正确的结果——我们现在（根据费舍尔的标尺）称为“自由度为 $n-1$ 的学生 t 分布”。这个过程包含了数学上的一些运气，戈塞特隐含地假设，样本均值和样本标准差之间缺乏相关性，意味着它们相互独立。这在他的正态情况下是真的，但在其他任何情况下都不是。图 4-2 显示了 9 个自由度的 z 的分布（实线），与标准差相同（在这个尺度上是 $1/\sqrt{7}=0.378$）的正态分布进行比较。他注意到，双方的一致性还不错，但标准差更大时，正态分布会给出一种“虚假的安全感”。

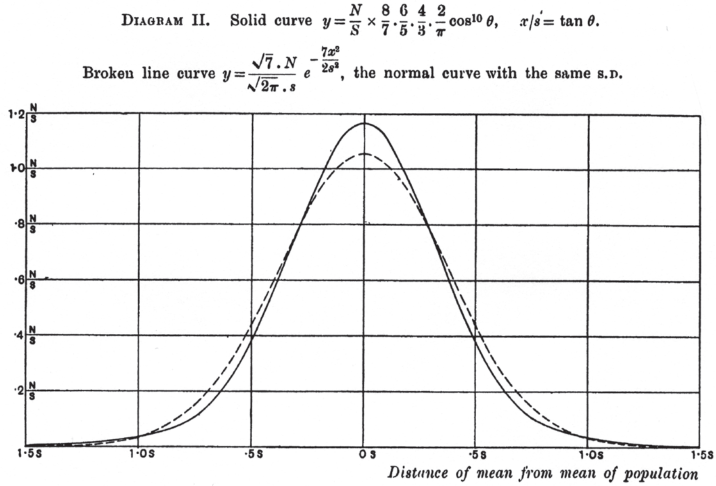


图 4-2　1908 年论文示意图，比较正态分布和 t 分布（自由度为 9）的密度（Gosset 1908）

n 越大，曲线就越接近正态分布。戈塞特附上了一个表格，对 $n=4$、5、…、10 计算显著性概率，并且使用了一些例子说明这个表格的使用。这些例子中最著名的是 Cushny-Peebles 数据①（如图 4-3 所示）。利用图 4-3 表格中最后一列的成对差异，戈塞特发现 $z=1.58/1.17=1.35$。就是说，均值差与 0 相距 1.35 倍的标准差。这会给出 $t=1.35\sqrt{(n-1)}=1.35\times 3=4.05$。由此，他说："这张表格可以得出，概率是 0.9985，或概率大约为 666 对 1，所以第二种是更好的催眠药剂。""学生"的 t-分布由此诞生，并初试牛刀。我们可以在结论中发现没有根据的贝叶斯推论的缺陷，在来源中发现错误引用（这篇文章发表在 1905 年，而非 1904 年），在药物中发现错误识别（他错误标记了列，而且事实上，他复制的数据并非用于催眠），以及不恰当的分析（个体的数据事

① 这个数据集显示了两种催眠药物对 10 位患者的治疗效果。——译者注

实上意味着大小不同的样本，因此方差非常不同，并且通过使用一个共同的缩放比例使它们看起来相关）。但至少数值工作是清晰且正确的，并且其他内容正确遵循了逻辑。

SECTION IX. *Illustrations of Method.*

Illustration I. As an instance of the kind of use which may be made of the tables, I take the following figures from a table by A. R. Cushny and A. R. Peebles in the *Journal of Physiology* for 1904, showing the different effects of the optical isomers of hyoscyamine hydrobromide in producing sleep. The sleep of 10 patients was measured without hypnotic and after treatment (1) with D. hyoscyamine hydrobromide, (2) with L. hyoscyamine hydrobromide. The average number of hours' sleep gained by the use of the drug is tabulated below.

The conclusion arrived at was that in the usual dose 2 was, but 1 was not, of value as a soporific.

Additional hours' sleep gained by the use of hyoscyamine hydrobromide.

Patient	1 (Dextro-)	2 (Laevo-)	Difference (2–1)
1.	+ ·7	+ 1·9	+ 1·2
2.	− 1·6	+ ·8	+ 2·4
3.	− ·2	+ 1·1	+ 1·3
4.	− 1·2	+ ·1	+ 1·3
5.	− 1	− ·1	0
6.	+ 3·4	+ 4·4	+ 1·0
7.	+ 3·7	+ 5·5	+ 1·8
8.	+ ·8	+ 1·6	+ ·8
9.	0	+ 4·6	+ 4·6
10.	+ 2·0	+ 3·4	+ 1·4
	Mean + ·75	Mean + 2·33	Mean + 1·58
	S. D. 1·70	S. D. 1·90	S. D. 1·17

图 4-3 1908 年论文的 Cushny-Peebles 数据。第一列中的–1 是–0.1 的印刷错误（Gosset 1908）

对于当前的目的，要点是比较带有样本标准差的样本均值——在没有利用任何外部参考的情况下做出的——既没有参考“真实的”标准差，也没有参考科学研究领域通常接受的阈值。更直接地说，比值 t 的分布绝不包含 σ ，因此任何包括比值 t 的概率统计量——比如 P 值——都可以在数据内部做出。如果这个比值的分布随 σ 变化，t 证据的使用也会随着 σ 变化。学生 t 推断是一种纯粹的内部数据分析。这种相互比较的使用将自身从输入中解放出来，威力强大。它也将自己置于批评面前，这些批评在 1919 年已经很常见，直到今天也未减弱：

统计显著性不需要反映科学的显著性吗？这个差异断言了有关催眠方法的任何的实际显著性吗？戈塞特对此未置一词。但是，当误导性的陈述依然是一个问题时，关注手边数据的能力成为一种威力，一种不可否认的好处。

戈塞特的文章发表后就几乎被忽略了。发表的刊物十分出名，一些调查也对这篇文章做了例行引用，但直到 20 世纪 20 年代，似乎没人真正在出版物中使用这个检验。1914 年出版的《统计学家和生物统计学家用表》中，皮尔逊介绍了戈塞特的检验和表格，并给出了 1908 年论文中的例子，包括不正确的 Cushny-Peebles 数据和贝叶斯推断。但根据我的检索，1925 年以前没有一个成功使用 t-检验的例子。我在都柏林的吉尼斯档案馆待了一个下午，检索 1908~1924 年的科学备忘录，没有任何发现：戈塞特自己在实践工作中忽视了这个检验。我找到好几个使用了统计的例子，都是根据均值距 0 有多少个标准差描述差异的。但实践中都没有 t-检验，也没有对这篇论文的引用。

即便如此，这篇论文还是产生了深远的影响，这都要归功于一位特殊的读者看出了它结果中的魔力。费舍尔很可能在 1912 年从剑桥毕业时就已经阅读了这篇论文，他看到文章里没有证明，但看出从多维几何的角度思考这个问题会得到简单和严格的证明。他写信给戈塞特（不知怎么他已经了解到“学生”的真实身份），解释了这个证明，但戈塞特没能理解。当戈塞特把它转寄给皮尔逊时，皮尔逊也没能理解。这封信遗失了，而且可能从未被回复过。1915 年，费舍尔在 *Biometrika* 上以一篇短小精悍的文章介绍了这个证明，他在文中还提到自己发现了一个更复杂的统计量——相关系数 r 的分布。

戈塞特的检验依然没有受到关注。到 20 世纪 20 年代早期，那时的费舍尔正在罗萨姆斯泰德试验站研究农业问题，他已经看出，将学生 t 分布从对 σ 的依赖中解放出来的数学魔法只是冰山一角。费舍尔

创造了双样本 t-检验，并推导出了用于回归系数的分布理论以及方差分析的完整步骤。

要想了解戈塞特工作对统计实践的历史影响，需要追溯到费舍尔的教科书对其的介绍。这本开创性的书出版于 1925 年，名为 *Statistical Methods for Research Workers*。戈塞特自己的论文提出了一种杰出的思想，但程度仅仅到单样本检验为止。这种检验除了应用于样本的成对差异外，很少有其他用处。费舍尔吸收了这种思想，并把它扩展到两个甚至多个样本的情况。这种情形下，该方法显现出了真正强大的作用。费舍尔的方差分析是真正的变异分析，他按照一种前人从未尝试的方法分解变异。好吧，这些并不是都是真的，埃奇沃思早在 40 年前就已经做出了某些卓越的工作。

4.2 弗兰西斯·埃奇沃思和方差成分的双因素分析

19 世纪 80 年代，埃奇沃思致力于将概率尺度的应用扩展至社会科学。为此，他发展了一种统计表格的分析方法，这种方法早于费舍尔之后所做的部分工作。在英国科学促进会 1885 年 9 月于阿伯丁举办的科学进展会议上，埃奇沃思结合两个例子给出了他的方法，一个是有意设计的，另一个则受到社会科学的广泛承认。对于第一个例子，他用表格列出了维吉尔的作品《埃涅阿斯纪》的一个片段中的扬抑抑格（dactyl，一个长音节后面跟随两个短音节的音步，如图 4-4 所示）。对于第二个例子，他采用了 1883 年户籍总署署长报告的英国 6 个郡 8 年间的死亡率（如图 4-5 所示）。这两个例子中，他都提供了每行、每列的和以及均值，还有所谓的“波动”（fluctuation），两倍于我们现在所称的“经验方差”，即相应的行和列满足 $2\Sigma(X_i - \bar{X})^2 / n$ 。

Æneid, XI, 1–75.	Lines 1—5	6—10	11—15	16—20	21—25	26—30	31—35	36—40	41—45	46—50	51—55	56—60	61—65	66—70	17—75	Sums.	Means.	Fluctuations.
First foot	3	3	5	5	4	4	2	2	2	1	2	4	3	2	4	46	3·06	2·8
Second ,,	1	4	0	3	3	3	5	2	2	4	3	1	2	3	2	38	2·5	3·2
Third ,,	1	2	4	2	5	2	1	2	2	2	0	2	2	0	1	28	1·86	3·1
Fourth ,,	2	2	1	0	3	1	2	0	2	1	1	2	1	1	0	19	1·26	1
Sums ..	7	11	10	10	15	10	10	6	8	8	6	9	8	6	7	131	8·68	10 1·0
Means ...	1·75	2·76	2·5	2·5	3·75	2·5	2·5	1·5	2	2	1·5	2·25	2	1·5	1·75	33	2·17	2·5 0·9
Fluctuations	1·5	2·5	9	7	3	3	5	2	0	3	3	2·5	1	3	2·5	2·08 48	0·9 3·2	

图 4-4　埃奇沃思 1885 年对维吉尔的《埃涅阿斯纪》所做的数据分析。这里和图 4-5 中的一些数值误差在文献 Stigler(1999)的第 5 章中做了修正（Edgeworth 1885）

	1876.	1877.	1878.	1879.	1880.	1881.	1882.	1883.	Sums.	Means.	Fluctuations.
Berks...........	175	172	187	186	181	153	169	166	1,389	173½	224
Herts	174	165	185	184	176	186	163	188	1,401	175	176
Bucks	182	171	186	195	179	162	177	183	1,435	179½	172
Oxford	179	182	194	183	180	169	167	166	1,420	177½	162
Bedford	196	174	203	195	198	171	181	184	1,502	188½	246
Cambridge	173	177	190	191	187	165	171	181	1,435	179½	158
Sums	1,079	1,041	1,145	1,134	1,101	986	1,028	1,068	8,582	1,073	1,138
Means	180	173½	191	189	183½	164	171	178	1,630	179	190 146
Fluctuations	124	55	77	50	107	68	73	152	—	846 88	—

图 4-5　埃奇沃思 1885 年关于郡死亡率的数据分析（Edgeworth 1885）

埃奇沃思的分析有一种潜台词。这两种情况下，数据都是计数的，要么是直接的计数（维吉尔），要么单位化到每万人的计数（死亡率）。当时，威尔赫姆 · 莱克希斯发展了一种方法，尝试在某种二项变异上

进行分析。埃奇沃思将这种方法称为“组合的”（combinatorial），并明确地想要避免这样一种假设。他的分析仅仅是建立于数据的内部变异，用高尔顿的术语来说就是“相互比较”。莱克希斯的方法将二项变异作为外部标准。使用简单的掷硬币模型作为基准是一种古老的诱惑——想想阿布斯诺特和出生性别数据。但在更复杂的情况中，这需要付出代价。对一个有 n 次试验的二项分布，并且单次试验成功概率为 p，它的均值 np 和方差 $np(1-p)$严格相关，而且不是所有数据都能反映这种联系。事实上，阿布斯诺特的出生数据是少有的二项分布起作用的情况，从那之后，大部份数据都存在分析家们所谓的“过度分散”：变异大于简单的二项分布，有可能因为 p 在试验中有随机变化。埃奇沃思想要避免莱克西斯受到的约束。用今天的术语来说，无论数据是不是二项分布的，只要变异是近似正态的，埃奇沃思就可以处理。

埃奇沃思的分析框架是我们今天所称的“方差分量”的一种估计。例如，将所有死亡率放入一组，我们可以考虑把全部的“波动”当作 3 个分量之和：$C^2+C_t^2+C_p^2$，其中第二个成分代表时间（年到年）变异，第三个分量代表地点（郡）变异，而第一个分量代表独立于时间和地点的随机变异。如果分析师想比较同一个郡随时间变化的死亡率，可以使用行的均值或者汇集的波动（图 4-5 中的 190）作为 $C^2+C_t^2$ 的估计以评估准确性；而为了比较同一年中各郡的情况，可以相应地使用 $C^2+C_p^2$ 进行估计（图 4-5 中的 88）。为了估计随机波动 C^2，可以考虑平均的行波动减去均值行的波动的差，$190-146=44$；也可以考虑平均的列波动减去均值列的波动的差，$88-46=42$。因为这是数值工作而非代数工作，他没有意识到，除了计算中的误差，这些差应该精确相等，都等于 $2SSE/IJ$。其中 I 和 J 分别是行数和列数，SSE 是拟合可加模型的残差平方和。类似地，他能发现，维吉尔往往对不同音步或不同诗段采用不同韵律频率。

埃奇沃思的工作可以说是一系列错失的机会，元凶是数值的误差和代数的笨拙。他的估计是方差分析中平方和的简单线性函数，而且其某些计算现在可以认为与费舍尔后来用在类似分析中的某些 F 统计量大致相同。但埃奇沃思没有用分布理论补充它们。当费舍尔在 20 世纪 20 年代中期面对这个问题时，他显然看到了完整的代数结构和正交性，比如多元正态分布的数学逻辑、允许行效应和列效应的统计分解、可以仅仅基于数据内部变异的显著性独立检验进行测量。

20 世纪后半叶，计算机的使用迅速普及，由此产生了大量使用计算机的过程的更多应用，其中包括几个可以认为是使用了相互比较的过程。20 世纪 50 年代，毛里斯 · 昆努利以及之后的约翰 · 图基发展了一种估计标准差的估计方法：通过相继删除每个观测以考虑估计变化了多少。图基将这个过程命名为“刀切法”（jackknife）。与此相关，一些人提出了在名为交叉验证的方法下研究变异，即在交叉验证中对数据的子集执行一个过程并比较结果。20 世纪 70 年代晚期，布拉德利 · 埃弗龙引入了他口中的“自助法”（bootstrap），现在已得到广泛应用。该方法有一个以随机替换方式重抽样的数据集，和一个每次都会计算的目标统计量，这个“自助样本”的变异性可以用于判断那些没有统计模型可借助的统计量的变异性。所有这些方法在估计变异性时都涉及了相互比较。

4.3　相互比较的一些陷阱

如果仅仅依靠一种数据内部变异的分析作为指导，将会遇到许多陷阱。模式似乎已经出现，接下来就是解释模式的故事。数据集越大，故事越多。有些故事是有用的，或具有深刻意义的，但许多故事两者皆非，甚至有些最优秀的统计学家也会受到这些差异的蒙蔽。

威廉·斯坦利·杰文斯并不是第一个在经济时间序列中分辨出商业周期的人，也不是第一个在商业周期和太阳黑子行为周期之间看到存在可能联系的人。但 19 世纪 70 年代后期，他痴迷于这种思想，甚至面对公众和专家的嘲讽依然如此。他的痴迷超过了前辈，而对后来者则意味着警告。

人们对周期的深入研究引发了天文学历史上的一些最伟大的发现，但社会科学中的周期类型与此不同。商业周期事实上是“周期性的”，但通常会有调整，一位赛马比赛的分析师将其称为“不断变化的周期”。19 世纪六七十年代，杰文斯对经济数据的多个序列做了仔细研究，最终得出这样的结论：有一种规律的商业周期，大约每 10.5 年被一次主要的商业危机打断。其他人也已看到这种规律，是一种真正的现象。他使用了 18 世纪晚期到 19 世纪 70 年代来自多个出版源的数据，甚至最终将这个记录回溯至包括“南海泡沫”的 1720 年。杰文斯第一次查看数据的时候，没有提供接近预测时间的恰当危机，但却发现能在适当时间至少发现小型危机。但原因是什么？

威廉·赫歇尔很久以前就提到，太阳黑子行为的主要爆发规则可能与商业周期有联系，但那些追随这个观点的人受到一种匹配差错的阻碍。太阳黑子周期每 11.1 年出现一个高峰——至少大家是这么认为的——那么几十年之后，这个 11.1 年和 10.5 年的商业周期就会变得逐渐不同步。但 19 世纪 70 年代中期，J. A. 布朗做了研究，将太阳黑子周期从 11.1 年修正为 10.45 年。杰文斯对这个问题的态度由兴趣变为痴迷，他甚至“改良”了太阳黑子序列。当序列中出现一个间隔，通常能够发现一个最小的最大值，那就是杰文斯能够接受的。

有些联系看起来确信无疑，而且那时杰文斯也发现，德里粮食的价格统计中存在长度大约相同的近似周期，这就是例子。太阳行为可能影响气象，这貌似有道理。而且，这种效应被宣布时，看起来也符

合印度的情况，是可能会对贸易产生影响的。毕竟，格拉斯哥城市银行的失败引起了 1878 年的英国危机，而这个失败源于最近的印度饥荒导致的贸易萧条。因为英国危机看起来跟随着德里价格序列变动了好几年，所以这个理论享受到了解释危机滞后的红利。图 4-6 取自杰文斯最后的论文，这篇论文发表在 1882 年 7 月的《自然》杂志上（同年 8 月 13 日，杰文斯死于一次意外的溺水，享年 46 岁）。这张图显示了太阳黑子序列（沃尔夫数，Wolf's numbers）、德里的粮食价格（"谷物"，corn），并像杰文斯标出的那样标识了英国的主要商业危机。杰文斯并未有意筛选和寻找，以便选择并重新计算这些数据，这给人们留下了深刻的印象——即使未能吸引所有观众。

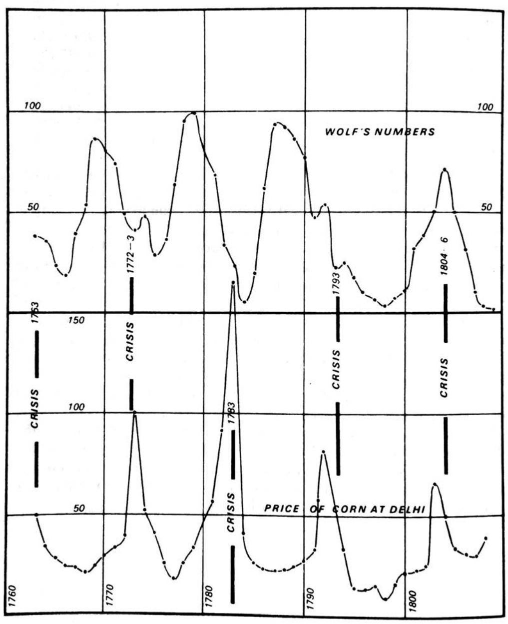


图 4-6　杰文斯 1882 年的图表，显示了太阳黑子行为和商业危机之间的联系（Jevons 1882）

杰文斯那时面临着嘲笑。他在 1879 年的伦敦统计协会会议上发言，预期经济会在未来一两年内复苏，“假如，太阳真的展现出适当的黑子数量”。听众们听到后都抑制不住地大笑。同年，协会《会刊》刊登了两篇匿名短文，一篇关于太阳黑子的高计数如何帮助剑桥大学在年度赛艇比赛中击败牛津大学，另一篇关于死亡率和木星运动之间可能存在的联系。1863 年，高尔顿写道，“行使偶尔抑制和轻微修正的权利，可以看到一个数量有限的观测在具有先入为主想法的人手中，如何塑造出想要的结论，这真是太荒谬了。”当然，即使完全良性的时间序列也能够表现出欺骗性的模式。1926 年的一篇题为“为什么我们有时候得到时间序列之间无意义的相关？”的挑衅性论文中，乔治·尤德尼·约尔展示了简单的自相关序列如何在有限的时间跨度上表现周期性。也许是出于好意，他没有提到杰文斯。

第 5 章

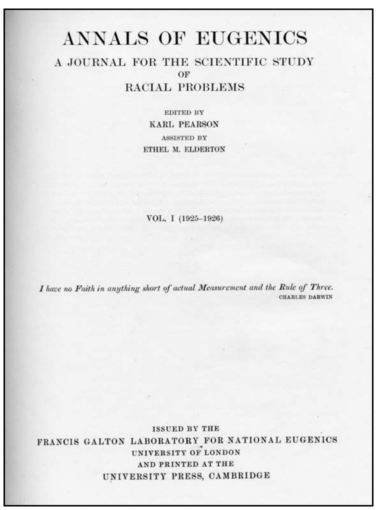
回归：多元分析、贝叶斯推断和因果推断

查尔斯·达尔文很少使用高等数学。1855 年，他给老朋友（也是他的表兄）威廉·达尔文·福克斯写了一封信。在信中，他用一种说法总结了自己的观点："我不相信任何缺乏真实测量和三分律（Rule of Three）的事情。"这种说法后来因为卡尔·皮尔逊而声名鹊起。1901 年，皮尔逊采用这句话作为新杂志 *Biometrika* 的座右铭；而 1925 年创办 *Annals of Eugenics* 时，他甚至把这句话放在每一期的标题页上（如图 5-1 所示）。这是皮尔逊能在达尔文的著作中找到的、最接近的对数学的认可。

在评价真实测量方面，达尔文确实是对的，但他对三分律的信念却是错误的。每一个学习过欧几里得五卷本的学童都很熟悉达尔文引用的三分律，这是一个简单的数学命题：如果 $a/b = c/d$，那么 a、b、c、d 中的任何三个都足以决定第四个。对于达尔文来说，这个规律是一个方便的外推工具，对在他之前的许多人也是如此（如图 5-2 所示）。17 世纪，约翰·格兰特和威廉·配第曾用这样的比例估计人口和经济

行为。18 世纪和 19 世纪初期，皮埃尔·西蒙·拉普拉斯和阿道夫·凯特勒也曾这样做。

无论是达尔文还是他之前的人，都没有意识到三分律的分析基础多么薄弱。在分派商业交易和欧几里得的数学问题中，这条规则尚可发挥作用。但对于存在变化和测量误差的有趣科学问题，这个规则就会失效。这些情况下，三分律会给出错误的答案。结果中会有系统性偏差，误差或许还会特别大，而其他方法可以减小误差。这个事实的发现出现在达尔文逝世 3 年之后，是统计学的第五根支柱。发现者是达尔文的表弟弗兰西斯·高尔顿，他 1885 年 9 月 10 日在苏格兰阿伯丁宣布了这项影响深远又令人惊讶的发现。在那里，他把这个发现命名为“回归”。1885 年之后的半个多世纪的时光中，这个基本概念成为统计发展的主线。这个发现的故事极具启发性，但在解释之前，有必要先说明欧几里得的错误是什么，以及为什么这个错误居然可以持续几千年。


ANNALS OF EUGENICS
A JOURNAL FOR THE SCIENTIFIC STUDY
OF
RACIAL PROBLEMS

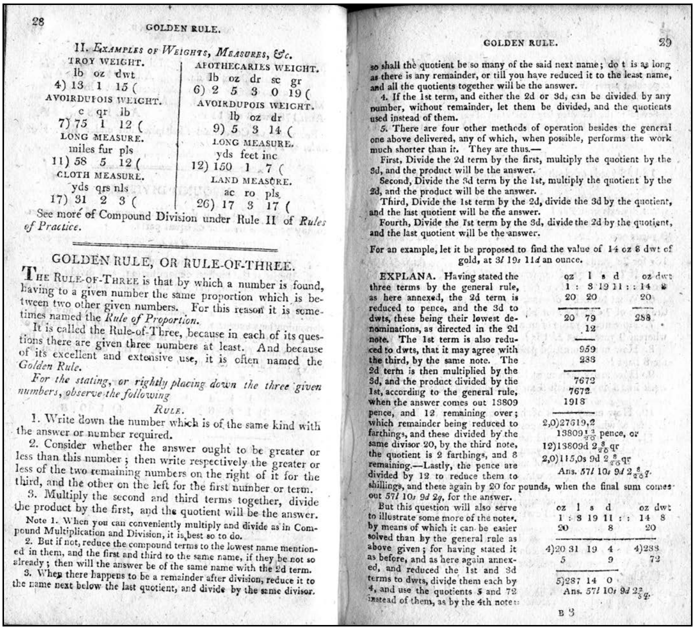
EDITED BY
KARL PEARSON
ASSISTED BY
ETHEL M. ELDERTON

VOL. I (1925–1926)

I have no Faith in anything short of actual Measurement and the Rule of Three.
CHARLES DARWIN

ISSUED BY THE
FRANCIS GALTON LABORATORY FOR NATIONAL EUGENICS
UNIVERSITY OF LONDON
AND PRINTED AT THE
UNIVERSITY PRESS, CAMBRIDGE

图 5-1 *Annals of Eugenics* 第一期标题页



28　GOLDEN RULE.

II. *Examples of Weights, Measures, &c.*

TROY WEIGHT.
lb oz dwt
4) 13 1 15 (

AVOIRDUPOIS WEIGHT.
c qr lb
7) 75 1 12 (

LONG MEASURE.
miles fur pls
11) 58 5 12 (

CLOTH MEASURE.
yds qrs nls
17) 31 2 3 (

APOTHECARIES WEIGHT.
lb oz dr sc gr
6) 2 5 3 0 19 (

AVOIRDUPOIS WEIGHT.
lb oz dr
9) 5 3 14 (

LONG MEASURE.
yds feet inc
12) 150 1 7 (

LAND MEASURE.
ac ro pls
26) 17 3 17 (

See more of Compound Division under Rule II of *Rules of Practice.*

GOLDEN RULE, OR RULE-OF-THREE.

The Rule-of-Three is that by which a number is found, having to a given number the same proportion which is between two other given numbers. For this reason it is sometimes named the *Rule of Proportion.*

It is called the Rule-of-Three, because in each of its questions there are given three numbers at least. And because of its excellent and extensive use, it is often named the *Golden Rule.*

For the stating, or rightly placing down the three given numbers, observe the following

Rule.

1. Write down the number which is of the same kind with the answer or number required.

2. Consider whether the answer ought to be greater or less than this number; then write respectively the greater or less of the two remaining numbers on the right of it for the third, and the other on the left for the first number or term.

3. Multiply the second and third terms together, divide the product by the first, and the quotient will be the answer.

Note 1. When you can conveniently multiply and divide as in Compound Multiplication and Division, it is best so to do.

2. But if not, reduce the compound terms to the lowest name mentioned in them, and the first and third to the same name, if they be not so already; then will the answer be of the same name with the 2d term.

3. When there happens to be a remainder after division, reduce it to the name next below the last quotient, and divide by the same divisor, so shall the quotient be so many of the said next name; do t is as long as there is any remainder, or till you have reduced it to the least name, and all the quotients together will be the answer.

GOLDEN RULE.　29

4. If the 1st term, and either the 2d or 3d, can be divided by any number, without remainder, let them be divided, and the quotients used instead of them.

5. There are four other methods of operation besides the general one above delivered, any of which, when possible, performs the work much shorter than it. They are thus.—

First, Divide the 2d term by the first, multiply the quotient by the 3d, and the product will be the answer.

Second, Divide the 3d term by the 1st, multiply the quotient by the 2d, and the product will be the answer.

Third, Divide the 1st term by the 2d, divide the 3d by the quotient, and the last quotient will be the answer.

Fourth, Divide the 1st term by the 3d, divide the 2d by the quotient, and the last quotient will be the answer.

For an example, let it be proposed to find the value of 14 oz 8 dwt of gold, at 3*l* 19*s* 11*d* an ounce.

EXPLANA. Having stated the three terms by the general rule, as here annexed, the 2d term is reduced to pence, and the 3d to dwts, these being their lowest denominations, as directed in the 2d note. The 1st term is also reduced to dwts, that it may agree with the third, by the same note. The 2d term is then multiplied by the 3d, and the product divided by the 1st, according to the general rule, when the answer comes out 13809 pence, and 12 remaining over; which remainder being reduced to farthings, and these divided by the same divisor 20, by the third note, the quotient is 2 farthings, and 8 remaining.—Lastly, the pence are divided by 12 to reduce them to shillings, and these again by 20 for pounds, when the final sum comes out 57*l* 10*s* 9*d* 2*q*, for the answer.

oz　l s d　oz dwt
1 : 3 19 11 : : 14 8
20　20　20
20　79　288
12
959
288
7672
7672
1918
2,0)27619,2
$13809\frac{12}{20}$ pence, or
12)13809d $2\frac{8}{20}$qr
2,0)115,0s 9d $2\frac{8}{20}$qr
Ans. 57*l* 10*s* 9*d* $2\frac{8}{20}$*q*.

But this question will also serve to illustrate some more of the notes, by means of which it can be easier solved than by the general rule as above given; for having stated it as before, and as here again annexed, and reduced the 1st and 3d terms to dwts, divide them each by 4, and use the quotients 5 and 72 instead of them, as by the 4th note:

oz　l s d　oz dwt
1 : 3 19 11 : : 14 8
20　8　20
4)20 31 19 4　4)288
5　9　72
5)287 14 0
Ans. 57*l* 10*s* 9*d* $2\frac{2}{5}$*q*.

B 3

图 5-2　来自一本 1825 年的教科书内页，达尔文或许学过类似的教科书。注意所用的案例：“举例来说，按 3 英镑 19 先令 11 便士一盎司的价格，求 14 盎司 8 英担黄金的价值。”英国式的测量将一个简单的外推变成需要不小篇幅来解决的问题（Hutton ca. 1825）

我们选择高尔顿考虑过的一个案例，这个案例直到今天依然是人类学中常见的问题。考古中发现了一个人的部分骨骼，但仅仅是一个长为 T 的大腿骨，人类学家还想知道这个人的身高 H。他们有一些可以用来比较的完整骨骼，给出一组配对值(T_i, H_i)，可以从中计算算术平均值 m_T 和 m_H。高尔顿的计划是使用三分律，从这些均值、已知的 T 以及关系 $m_T/m_H = T/H$ 推断未知的 H。如果这个关系与欧几里得考虑过的问题一样，在数学上是严格的，且对所有 T_i/H_i 都有一个常数比值，三分律就会起作用。但在这里，和所有有趣的科学问题一样，

效果会发生变化。高尔顿发现，这决定了三分律并不适用于回归现象。在一种极端的情况下，如果 T 和 H 变化，但仍不相关，H 的最佳估计会忽略 T，是 m_H。只有它们完美相关时，欧几里得的答案才正确。在一种普通的情况下，高尔顿发现存在一个普通的解。说来奇怪，从 T 预测 H 的关系明显不同于从 H 预测 T 的关系，而且两者都不符合欧几里得的规律。

5.1 发现之路：从达尔文到高尔顿

达尔文 1859 年发表的物种起源理论是不完整的，而且直到 1882 年他去世的时候，也依然不完整。所有理论出现时也都是不完整的，因为一旦理论有所突破，就会出现更多需要解决的事情。因此，一种理论越丰富就越不完整。但达尔文理论的缺憾以一种更加基础的方式表现了出来——留下了一个可争议的问题。如果它受到广泛的注意，问题就出来了。这个问题很微妙，直到达尔文逝世 3 年以后，高尔顿才刚刚发现一种解决方案，这才出现了问题的完整思考和表达。

这个问题涉及达尔文观点的基本结构。为了使进化符合自然选择，这种理论需要这样的基础：物种内必须有足够的可遗传的变异性。一个亲代的后代的遗传方式必须有一定差异，否则连续的世代之间不会发生任何改变。《物种起源》的第一章以一种令人信服的方式建立了这种理论，对象是驯养的和自然的动植物种群。但同时，达尔文无意中又产生了一个问题——理论中的一个明显矛盾。

毫无疑问，达尔文在世时只有两位读者注意到了这个问题。工程师弗莱明 · 詹金在 1867 年的一篇书评中谈到了这个问题，另一位就是 10 年后的高尔顿。詹金的评论仅仅认识到了这个问题的一部分，之后就将注意力分散到了其他不相关的一组问题的介绍。1877 年，高尔

顿完全阐明了这个问题，并把它视为一个严重的挑战。高尔顿的构想可以用图形表示。达尔文已经令人信服地确定：可遗传的变异通过代际转移传递给后代（如图 5-3 所示）。

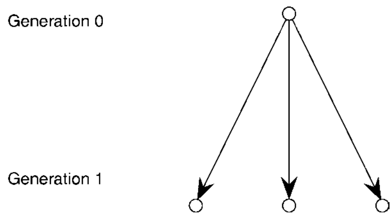


图 5-3　在一代的后代中产生的变异

同一个亲代可以生育遗传特征不同的后代。尽管高尔顿只是泛泛地考虑了这个问题,但将身材作为特征来考虑的想法或许是有帮助的。成年身高只是高尔顿广泛研究的一个特征，考虑到特征的已知性别差异，女性的身高会放大 1.08 倍。但是，如果从父母到子女有增加的变异，那后代会怎么样？如果随后每一代中变异都增加，这个相同的模式会不会持续（如图 5-4 所示）？

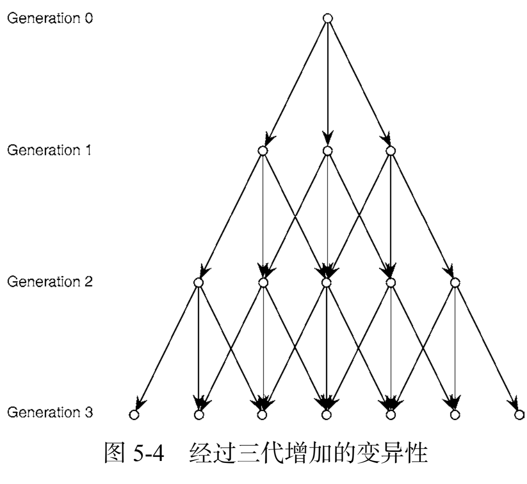


图 5-4　经过三代增加的变异性

但是，短期内无法观察到一代代累积起来的增加的变异。一个物种内，种群密度在随后的世代中几乎相同（如图 5-5 所示）。种群的散布在短期内是稳定的，事实上，这种稳定性对定义一个物种来说是必不可少的。

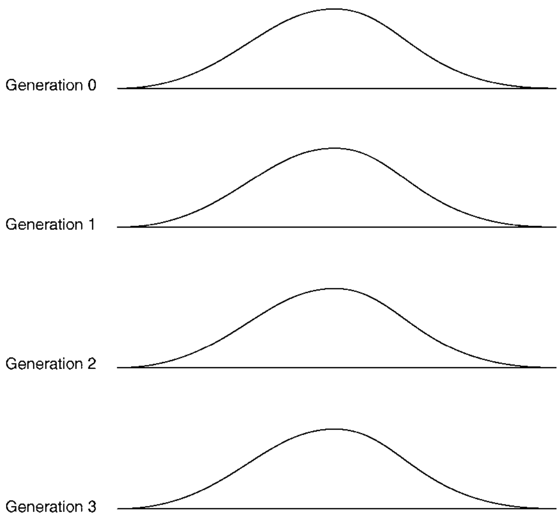


图 5-5　种群多样性经过三代后是稳定的

无论在自然界还是在（缺乏积极育种控制的）种植生产中，今年庄稼的大小和颜色的多样性看起来与去年差不多，前年也是。没有饮食上的显著改变时，任何人类种群从一代到下一代在身材上都表现出相同的变化。

高尔顿提出的观点并非关于物种的长期进化，他认为，根据达尔文给出的理由，显著性的改变已经发生，或者即将发生。高尔顿关心的是短期。他考虑达尔文理论的含义，甚至称它为“典型的遗传”。当时间尺度足够短时，至少可以认为环境中有一个近似的均衡，不会发生任何突然的改变。即使在近似均衡中，达尔文要求并论证过的变异

性的存在也与种群中观测到的短期稳定性互相冲突。除非能够发现某些力量可以抵消增加的变异性，并符合遗传的代际变异，否则达尔文的模型行不通。发现这种力量之前，高尔顿研究了 10 年。事实上，也是他的成功才拯救了达尔文的理论。

尽管高尔顿的解决方案以一系列模拟的模型为框架，但依然明显是纯数学式的。因此，它在早期生物学中非常独特。威廉·哈维基于算术计算发现了血液循环，但这项研究还是偏经验主义的。许多早期的科学家（例如洛伦佐·贝利尼和阿奇博尔德·皮特凯恩）试图创造一种数理医学，但没有成功的记载。重新发现孟德尔工作的 20 年前，高尔顿在不了解遗传学的条件下，实际上独立得出了孟德尔遗传学的一些应用结果。

1873 年，高尔顿开始设计使用“钉板”（quincunx）这种装置，用以表达代际变异。铅粒在钉板中经过一行行钉有偏置销的靶子降落，每一行中，铅粒随机偏离向左或向右，直到落入底部的几个隔间之一（如图 5-6 所示）。

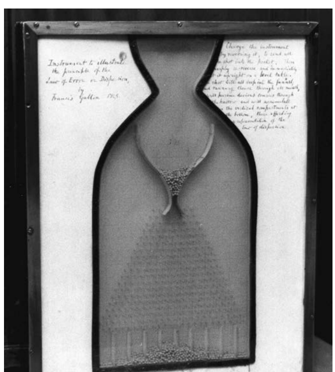


图 5-6　最初的高尔顿钉板，为了用于 1874 年的一次公开演讲而建造于 1873 年。底部的铅粒给出了钟形曲线的效果（Stigler 1986a）

1877 年，他进一步发展了这种想法，展示了连续种群分布之间发生变异性的效应。图 5-7 中，上层代表种群分布。就是说，对于第一代，矮小的身材在左侧，高大的身材在右侧，大致上呈钟形正态分布。

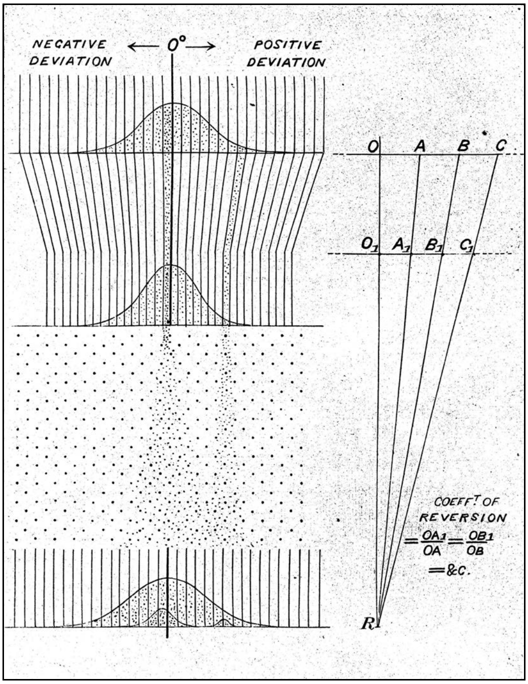


图 5-7 1877 年版本的高尔顿“钉板”，显示出为了保持恒定的种群分散，靠近顶部的倾斜的滑道如何补偿了下方增加的分散，以及位于上层的两个身高群体的后代如何通过这个过程追踪到下层

为了保持恒定的种群分散，高尔顿发现有必要引入他所谓的“倾斜的滑槽”(inclined chutes)，以便使其在服从代际变异之前压缩分布。在中部，高尔顿以图示展现了这种变异在两个身高相似的代表性群体中的效应，一个在中部，一个在右侧。每一个群体会立即在下面产生

一个后代的小分布，表现为一个小的正态曲线，其面积与母群体的大小成比例。为了保持代际平衡，他精确地计算了滑道倾斜的样子（他称为“恢复系数”：coefficient of reversion）。但是，他很难解释它们的形成原因。1877 年，他所能做的最好的事情就是，暗示它们可能代表了一个较小的生存倾向。对于那些远离群体均值的曲线来说，这意味着降低的适应性。一个退而求其次的借口是，为了给出一种精确的平衡，似乎要求存在一定水平的巧合。这个情节连好莱坞也无法接受，高尔顿自己也没再提起。

为了理解高尔顿最终给出的解决方案，以及带他到达解决方案的那个奇妙装置，可参考图 5-8（先关注左侧的图）。这是他发表于 1889 年的一幅图的修饰，这幅图显示了一个中部（A）有截断的钉板。现在，铅粒在半路就停止了。如果不受打断，它们可能产生的轮廓显示在底部（B）。在水平 A 和 B 的两个分布的轮廓是相似的，它们的区别仅仅在于，中间水平（A）画得更粗糙一些（这是我加上去的），而且比低水平（B）更紧凑一些。一切正在预料之中，水平 A 的铅粒只有大约一半的变异。

高尔顿观察到了下面的悖论。如果你在单一的中层隔舱释放铅粒，如图 5-8 左图箭头所示，它们会随机落到左侧或右侧，但平均会直接落下。有些会落到左边，有些落到右边，但不会有朝向某一个方向的明显倾向。但如果我们看更低的隔间，比如左侧的，释放所有中间层的铅粒让它们落到底部。问问底层隔间的这些“居民们”，它们是从哪里落下来的。答案不是“直接从上面落下的”，而是来自平均更靠近中部的地方（如图 5-8 右图所示）！原因很简单。比起那些位于层 A 左侧但会落到右侧的铅粒，更多层 A 中部的铅粒会落到隔间的左侧。因此，从不同立场提问这两个问题会获得根本不同的答案。或许，我们天真期待过的简单相互作用还没有找到。

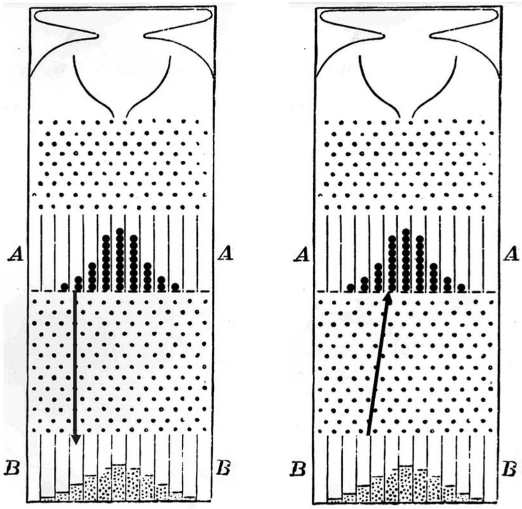


图 5-8　1889 年示意图的修饰版本。左图显示铅粒从上层隔间落下的平均最终位置，右图显示铅粒抵达更低的隔间的平均起始位置（Galton 1889）

下面考虑高尔顿“钉板”和他收集的数据之间的联系。图 5-9 显示了高尔顿的表格，给出来自 205 组父母的 928 个子女的成年身高的交叉分类。父母的身高概括为“中亲”（mid-parent）身高，是父亲身高和放大 1.08 倍的母亲身高的平均。女性后代的身高同样放大 1.08 倍。看“成年的儿童总个数”（Total Number of Adult Children）这个列。考虑将其作为钉板中 A 层的群组大小的计数，对应最左侧一列标签描述的分组。表格中的行给出每个群组内后代变异性的历史。例如，在身高标签为 72.5 英寸的行中有 6 个中亲，对于他们生育的 19 个子女，其成年身高范围从 68.2 英寸到“以上”（above），位于一个与高尔顿显示在图 5-7 底部的小正态曲线之一类似的模式里。于是，每一行显示（原则上）一个这样的小正态曲线，然后，“总和”（Totals）行给出加总计数，即显示在钉板底层隔间的计数（图 5-8 中的层 B）。

NUMBER OF ADULT CHILDREN OF VARIOUS STATURES BORN OF 205 MID-PARENTS OF VARIOUS STATURES.
(All Female heights have been multiplied by 1·08).

Heights of the Mid-parents in inches.	Heights of the Adult Children.														Total Number of		Medians.
	Below	62·2	63·2	64·2	65·2	66·2	67·2	68·2	69·2	70·2	71·2	72·2	73·2	Above	Adult Children.	Mid-parents.	
Above ..	..	..	..	..	..	..	..	..	..	..	..	1	3	..	4	5	..
72·5	..	..	..	..	..	..	..	1	2	1	2	7	2	4	19	6	72·2
71·5	..	..	..	..	1	3	4	3	5	10	4	9	2	2	43	11	69·9
70·5	1	..	1	..	1	1	3	12	18	14	7	4	3	3	68	22	69·5
69·5	..	..	1	16	4	17	27	20	33	25	20	11	4	5	183	41	68·9
68·5	1	..	7	11	16	25	31	34	48	21	18	4	3	..	219	49	68·2
67·5	..	3	5	14	15	36	38	28	38	19	11	4	..	..	211	33	67·6
66·5	..	3	3	5	2	17	17	14	13	4	..	..	..	..	78	20	67·2
65·5	1	..	9	5	7	11	11	7	7	5	2	1	..	..	66	12	66·7
64·5	1	1	4	4	1	5	5	..	2	..	..	..	..	..	23	5	65·8
Below ..	1	..	2	4	1	2	2	1	1	..	..	..	..	..	14	1	..
Totals ..	5	7	32	59	48	117	138	120	167	99	64	41	17	14	928	205	..
Medians ..	..	..	66·3	67·8	67·9	67·7	67·9	68·3	68·5	69·0	69·0	70·0	..	..	..	..	..

NOTE.—In calculating the Medians, the entries have been taken as referring to the middle of the squares in which they stand. The reason why the headings run 62·2, 63·2, &c., instead of 62·5, 63·5, &c., is that the observations are unequally distributed between 62 and 63, 63 and 64, &c., there being a strong bias in favour of integral inches. After careful consideration, I concluded that the headings, as adopted, best satisfied the conditions. This inequality was not apparent in the case of the Mid-parents.

图 5-9　关于家庭身高的高尔顿数据（Galton 1886)

如果人的身高确实与“钉板”显示的趋势相似，后代的身高应该从中亲身高水平直线下降。表格右侧的列“中位数”（Medians）给出了每一个中亲群组的子女中位数身高。（每一个像小正态曲线的群组，毫无疑问，要从未分组的数据中计算。）高尔顿注意到这些中位数并不是直接下降的，相反，它们往往更接近整体的平均——清晰的迹象表明倾斜槽必然在那里！可以肯定，这些倾斜槽是看不见的，但它们正在以一种神秘的方式执行着高尔顿在1877年示意图中为它们安排的任务。高尔顿给出一张示意图（如图5-10所示），清楚地展示了这个想法。

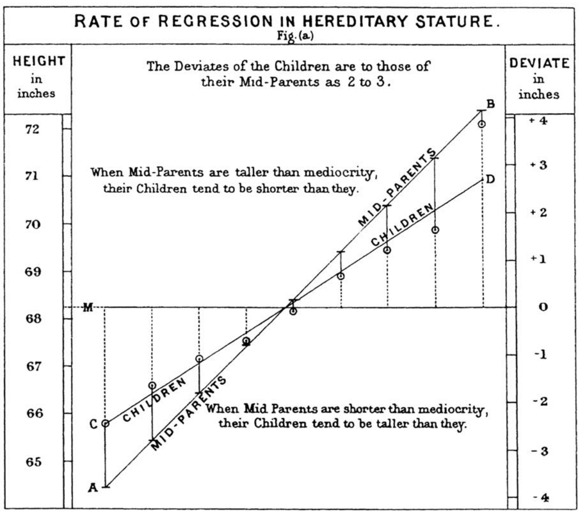


图5-10 高尔顿在此画出图5-9的最左侧列和最右侧列的数字，表明子女身高的趋势更接近人群的平均身高，而不是他们中亲的加权平均身高。这是一种“向着平庸的回归”（Galton 1886）

看了列的中位数，他注意到了相同的现象：每个后代的身高群组都有一位比他们更接近中间值（平庸）的平均中亲。高尔顿还认为，这里有更多可以生育超常身高的后代的普通身高的父母，而不是更多可以生育更少极端身高的子女的极端身高的中亲。但是，斜槽起了什

么作用？它的解释是什么？

到 1885 年，高尔顿有了更多证据支持这个现象，这些证据闪耀着新的光芒。因为前一年做过一项研究，他现在拥有许多家庭的数据。于是，高尔顿考虑参照已经研究过的中亲及其子女，以同样的方式考查兄弟之间的数据。结果极其相似，相联系的模式是相同的（家庭的高度在变动）。但最引人注目的是，他在这里又发现了“回归”。看图 5-11 右侧的列，这里的中位数比可能的预期更加系统性地接近“平庸”。

他的表格中，兄弟之间没有指向性，即没有哪位兄弟从其兄弟那里继承身高。这个原因很简单，结果也很特别。高尔顿曾使用他的各种“钉板”捕捉排序的定向流动，而这里没有他曾经寻求的那种定向流动。兄弟数据明显对称。除了拼写错误，他必须把每对兄弟的数据重复两次，每个排序中各列出一次。图 5-11 的左上角和右下角给出的“低于 63 和 74 以上”（below 63 and 74 and above）的对，必须是相同的两个个体。在这里，“倾斜的滑槽”如何起作用？事实上，这里似乎甚至排除了“遗传”。高尔顿认为，这必然是明确的。问题的解释必须是统计的，而不是生物的。

高尔顿返回中亲和子女的数据，对计数进行了光滑处理，对 4 个格子的分组计数计算均值并四舍五入，以便更好地发现模式。他看到，一个粗糙的椭圆形轮廓出现在表格最密集的部分。高尔顿写出了刻画钉板作用的一个数学方程（中亲总体可以视为一个正态分布，而且每个子群的后代都可以看作一个离差相同的正态分布），而且借助了一位剑桥数学家的一些帮助，他发现了这张表格的一种理论表述，就是我们现在确认的二元正态分布。它带有主轴和副轴，更重要的是，带有两条“回归线”（如图 5-12 所示）。一条是理论上的线，描述了后代的期望身高（线 *ON*），看作中亲身高的函数；另一条是中亲的期望身高（线 *OM*），看作子女身高的函数。

Relative number of Brothers of various Heights to Men of various Heights, Families of Five Brothers and upwards being excluded.

Heights of the men in inches.	Heights of their brothers in inches.													Total cases.	Medians.
	Below 63	63·5	64·5	65·5	66·5	67·5	68·5	69·5	70·5	71·5	72·5	73·5	Above 74		
74 and above	1	1	..	..	..	..	..	1	1	..	5	3	12	24	
73·5........	..	..	..	..	..	1	3	4	8	3	3	2	3	27	
72·5........	..	..	..	..	1	1	6	5	9	9	8	3	5	47	71·1
71·5........	..	1	..	1	2	8	11	18	14	20	9	4	..	88	70·2
70·5........	..	..	1	1	7	19	30	45	36	14	9	8	1	171	69·6
69·5........	..	1	2	1	11	20	36	55	44	17	5	4	2	198	69·5
68·5........	..	1	5	9	18	38	46	36	30	11	6	3	..	203	68·7
67·5........	2	4	8	26	35	38	38	20	18	8	1	1	..	199	67·7
66·5........	4	3	10	33	28	35	20	12	7	2	1	..	..	155	67·0
65·5........	3	3	15	18	33	36	8	2	1	1	..	..	..	110	66·5
64·5........	3	8	12	15	10	8	5	2	1	..	..	..	..	64	65·6
63·5........	5	2	8	3	3	4	1	1	..	1	..	..	1	20	
Below 63	5	5	3	3	4	2	..	..	..	..	..	..	1	23	
Totals	23	29	64	110	152	200	204	201	169	86	47	28	25	1329	

图 5-11　高尔顿 1886 年关于兄弟身高的数据（Galton 1886）

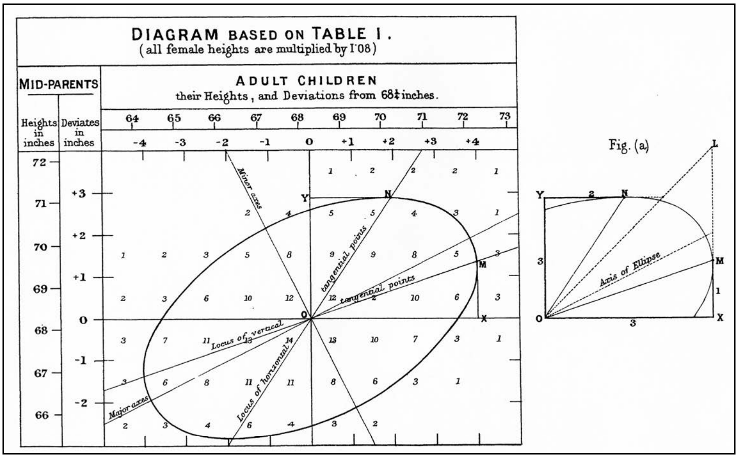


图 5-12　高尔顿 1885 年的示意图，展示了两条回归线 *ON* 和 *OM*。“三分律”线在右侧的小示意图中给出，即线 *OL*（Galton 1886）

这种统计现象的本质逐渐变得清晰。根据这些线，无论从二元密度的理论版本出发，还是从表格的数值版本出发，通过对两个不同方向取均值都会发现，除非所有数据都位于表格的对角线上，否则这两条线不可能一致。除非这两个特征的相关系数为 1（这里使用了高尔顿在 1888 年底为这些数据引入的术语“相关系数”），否则这两条线必须有区别，并且每条线必须在完美相关情况（椭圆的主轴）与零相关情况（穿过中心的水平线或垂直线）之间做出某种妥协。有趣的是，达尔文的示意图也给出了三分律可能给出的线。它是右图中的 *OL* 线，既不符合回归线，也不符合任何特殊的统计解释。这种情况下，它就是简单的 45°对角线，反映了中亲和子女总体相等的平均身材。

5.2　高尔顿的解释

高尔顿 1889 年出版了著作 *Natural Inheritance*，他在其中总结了

这个研究，用文字表达了自己的想法。高尔顿认为，如果 P 是总体的平均身材，那么给定一位兄弟的身材，则“未知的兄弟有两种不同的倾向，一种是与已知的兄弟相一致，而另一种是与他的种族相一致。一种倾向是对 P 的偏离和他的兄弟一样多，而另一种趋势则是毫无偏离。结果就是妥协”。按照现代术语，我们用 $S1$ 和 $S2$ 表示两兄弟的身材，其中每人的身材包含两个成分，$S1 = G + D1$，$S2 = G + D2$。其中 G 是一个不可观测的持续性成分（他们彼此共同持有的一种基因成分），对两兄弟共同起作用；而 $D1$ 和 $D2$ 是不可观测的暂时或随机成分，与 G 不相关而且彼此也不相关。高尔顿的 P 会代表总体中所有 G 的均值。

于是，我们可以把回归思想表达为一种选择效应。如果观察到第一位兄弟的身材 $S1$ 超过了总体平均 P，那么平均而言，$S1$ 可能的偏离缘于这两种原因的某种平衡，因为一个个体的 G 的变化略高于 P，同时因为 $D1$ 的变化略高于 0。转向第二位兄弟时，他的 G 将会和他的兄弟一样，但平均起来，$D2$ 的贡献将是 0，所以 $S2$ 的期望高于 P，但只会高到 G 的程度，而不是 $G + D1$，因此没有 $S1$ 那么多。而且颠倒 $S1$ 和 $S2$ 的位置后，同样的观点也会发挥作用。

5.3 达尔文问题的解决

高尔顿已经发现，向均值的回归不是生物上改变的结果，相反，这是亲代和后代之间不完美相关的结果。并且，达尔文认为，完美相关的缺失是一个必要条件，否则不会有代际变异，也不会有自然选择。之前我给出的理论表示（图 5-4）的绘图最好包含回归，如图 5-13 所示。图 5-13 和图 5-4 的差异在于，前者认为观测到的身材并非是完全遗传的，而是由两个成分组成，其中暂时性的成分是不遗传的。于是，这与一个近似的进化均衡中种群的分散相一致——从种群中心向极端

值的运动被反向运动所平衡。因此，朝向极端的大多数变异是缘于人口稠密的中间部分的短暂运动。高尔顿确认的问题归根结底不是一个问题，只不过由于统计效应导致之前无人确认。种群均衡和代际变异并不冲突。

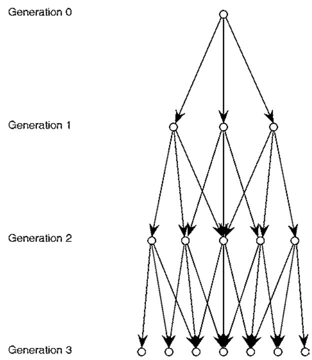


图 5-13　图 5-4 的重绘，目的是为了允许回归

5.4　影响

高尔顿对达尔文问题的研究发挥了重大影响。接纳达尔文理论的过程中，他扮演了重要角色。高尔顿解决了一个似乎没有人完全意识到的问题，而且展示出完全正确的理解，他发展的方法在 20 世纪早期的生物学中非常重要。高尔顿介绍了相关系数和简单的方差成分模型。事实上，他仅通过统计的方法就发现了一些结果，这些结果在 1900 年因为孟德尔的工作而重获发现。比如亲代对后代贡献的定量程度，再比如兄弟之间的联系要比父母子女间的联系更紧密。1918 年，罗纳德·A. 费舍尔使用了一种难度极高的数学技巧，扩展了孟德尔分配的方差计算，从相关到偏相关，再到所有各种关系。这样就产生了现代数量遗传学。

影响不仅发生在生物领域。方差成分思想成为定量心理学和教育心理学的关键。高尔顿的分离持久效应和暂时效应的思想，是经济学家米尔顿·弗里德曼 1957 年出版的著作 *The Theory of the Consumption Function* 中提出的模型的核心，后者因此获得了 1976 年诺贝尔经济学奖。弗里德曼主张，个人消费主要依靠个人收入的持久成分，而个人消费对诸如约翰·梅纳德·凯恩斯提出的暂时增长（“经济刺激计划”）是相对不敏感的。他由此得出结论，基于“政府临时支出会产生持久效应”的相反假设而得到的经济政策，是错误的。

5.5 多元分析和贝叶斯推断

历史学家们忽略了高尔顿这个发现在某一方面的影响，而这个方面不但深远，甚至可以说更有影响力。在 1885 年高尔顿的工作之前，没有工具可以用于进行真正的多元分析。早期的研究者们考虑过多维统计分布。图 5-14 展示了一些早期二维误差分布的例子，比如发生在打靶射击中的误差，图 5-15 展示了多于一个变量的早期密度公式。另外，首次出版的分析多于一个未知量的公式可以追溯到 1805 年及其以前的最小二乘。

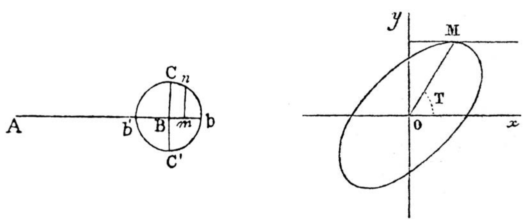


图 5-14　二元密度轮廓，来自罗伯特·阿德里安（左侧）、奥古斯特·布拉菲（右侧）和伊西多尔·迪迪翁（下页）（参见 Adrain 1808, Bravais 1846, Didion 1858）

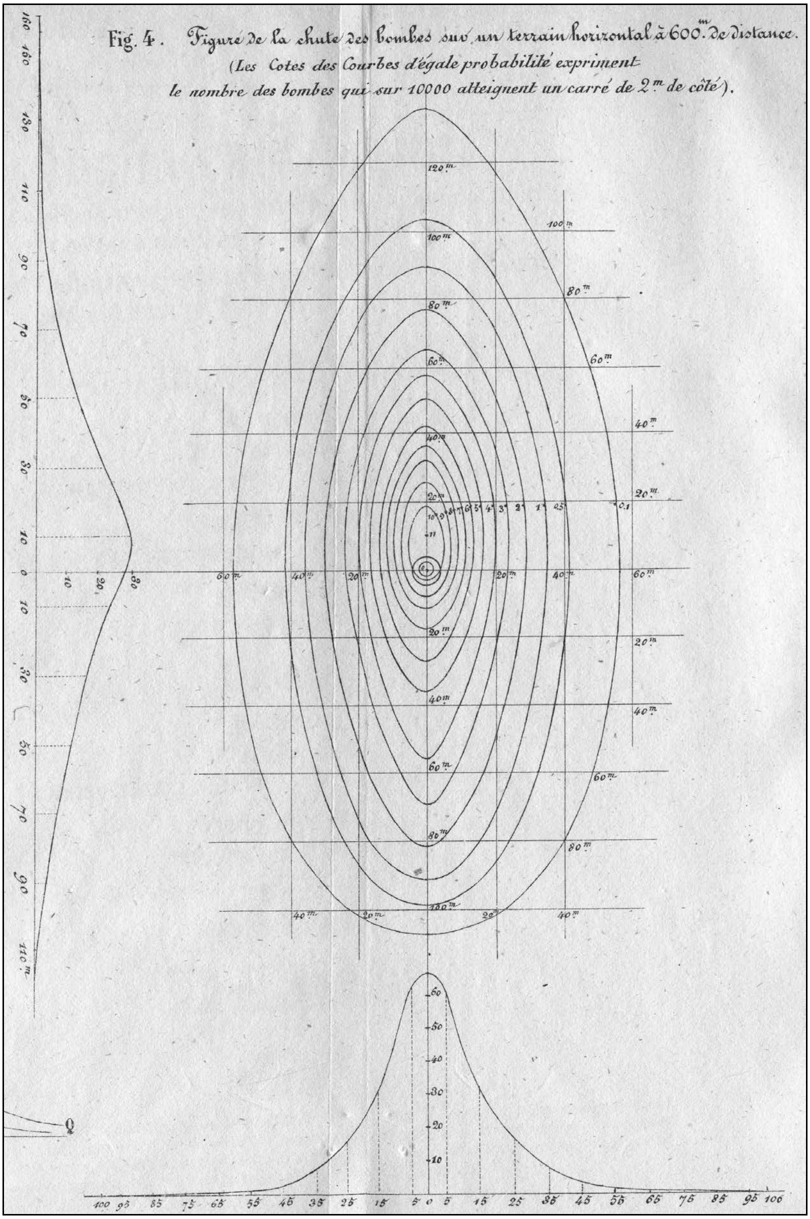


Fig. 4. Figure de la chute des bombes sur un terrain horizontal à 600m. de distance.
(Les Cotes des Courbes d'égale probabilité expriment
le nombre des bombes qui sur 10000 atteignent un carré de 2m de côté).

Soit maintenant $x = \xi\sqrt{n}$, $y = \psi\sqrt{n}$ $z = \zeta\sqrt{n}$ &c., & $\frac{\alpha}{n} = A, \frac{\beta}{n} = B, \frac{\gamma}{n} = C$&c. on aura $\xi + \psi + \zeta +$ &c. $= 0$ & $A + B + C +$ &c. $= 1$; donc,

$$P = \frac{1}{(\pi n)^{\frac{m-1}{2}} \sqrt{(ABC\ldots)}} \quad \&$$

$$V = e^{-\frac{1}{2}\left(\frac{\xi^2}{A} + \frac{\psi^2}{B} + \frac{\zeta^2}{C} + \&c.\right)}$$

Or, comme l'incrément où la différence des quantités x, y, z &c. eſt $= 1$, la différence des variables ξ, ψ, ζ &c. ſera $= \frac{1}{\sqrt{n}}$ &, par conſéquent, infiniment petite; de ſorte que, ſi on appelle cette différence $d\theta$, on aura

$$P = \frac{d\theta^{m-1}}{\sqrt{(\pi^{m-1} ABC\ldots)}}$$

Donc $-\frac{1}{2}\left(\frac{\xi^2}{A} + \frac{\psi^2}{B} + \frac{\zeta^2}{C} + \&c.\right)$

$$P\,v = \frac{e \qquad d\theta^{m-1}}{\sqrt{(\pi^{m-1} A.A.C\ldots)}}$$

$$E = S.m^{(i)2}.S.n^{(i)2} - (S.m^{(i)}n^{(i)})^2;$$

la double intégrale précédente devient

$$c^{-\frac{k}{4k''a^2.E}[l^2.S.n^{(i)2} - 2ll'.S.m^{(i)}n^{(i)} + l'^2.S.m^{(i)2}]}$$
$$\times \iint \frac{dt.dt'}{4\pi^2.a^2}.c^{-\frac{k''t^2}{k}.S.m^{(i)2} - \frac{k''t'^2.E}{k.S.m^{(i)2}}}.$$

En prenant les intégrales dans les limites infinies positives et négatives, comme celles relatives à $a\varpi$ et $a\varpi'$, on aura

$$\frac{1}{\frac{4k''\pi}{k}.a^2\sqrt{E}}.c^{-\frac{k}{4k''a^2}.\frac{l^2.S.n^{(i)2} - 2ll'.S.m^{(i)}n^{(i)} + l'^2.S.m^{(i)2}}{E}}. \qquad (o)$$

Il faut maintenant, pour avoir la probabilité que les valeurs de l et de l' seront comprises dans des limites données, multiplier cette quantité par $dl.dl'$, et l'intégrer ensuite dans ces limites. En nommant X cette quantité, la probabilité dont il s'agit sera donc

图 5-15 多元正态密度的发表公式，来自拉格朗日（上方）和拉普拉斯（下方）（参见 Lagrange 1776, Laplace 1812）

但 1885 年前，没有人考虑过分割连续的二元分布，比如寻找 X 和 Y 的密度。给定 Y 的 X 的条件分布和给定 X 的 Y 的条件分布，在正态

情况下求出条件均值和条件方差。这个数学步骤很简单，但显然，在高尔顿之前没有人有动力去实现。高尔顿通过遗传研究，考虑了不同条件下二元关系的一般问题。

正当高尔顿等待着自己 1889 年的著作步入出版流程时，他意识到，如果 X 和 Y 的标准差相等，那么 Y 对 X 的回归斜率与 X 对 Y 的回归斜率相等，并且，它们共同的值可以用作联合关系的测量。相关系数就此诞生。几年之内，弗朗西斯・埃奇沃思、G. 尤德尼・约尔和皮尔逊把这种思想发展到了更高的维度，使用与偏相关、多维最小二乘以及方差的主成分有关的联合测量。统计学已经跳出数字的二维表格，开始作为一种处理更复杂问题的技术而崭露头角。

5.6　贝叶斯推断

对推断来说，这个新发现具有深刻的含义。从根本上讲，推断是结合手边数据做出的条件性陈述，通常根据的是之前概括的公式。贝叶斯推断是一个典型的例子，在最简单的形式中，它相当于对统计学家关心的未知值 θ 指定一个先验的概率分布 $p(\theta)$，以及对给定 θ 的数据 X 也指定一个分布，也就是似然函数 $L(\theta)=p(x\mid\theta)$。然后找到(X,θ)的二元概率分布，以及从它得出给定 $X=x$ 时 θ 的条件概率分布，即后验分布 $p(\theta\mid x)$。至少我们现在是这么处理的。这些简单的步骤在 1885 年以前还是不可行的。高尔顿的“回望”为了在给定成年子女的身高时找到中亲身高的分布，或者在给定一个兄弟的身高时找到另一个兄弟身高的分布，这是一种真正的贝叶斯计算，而且似乎是第一次以这种简单形式做出的贝叶斯计算。

当然，这种或那种的逆概率有着很长的历史，至少可以追溯到托马斯・贝叶斯（发表于 1764 年）和拉普拉斯（发表于 1774 年）。但无

论是他们还是其他人，在随后的年代中，都没有给出符合现代约定的形式。对于连续变化的量，他们中间没有人使用条件分布，所有人都用了假定本质相等并平坦的（均匀的）分布。贝叶斯只考虑了推断 n 次独立试验中的二项概率 θ（在单次实验中成功的概率）的情况。其中，他假定唯一可得的信息已经成功发生（按照我们的术语）X 次，并失败了 $n-X$ 次。严格地说，那里没有"先验"，但他调整了自己的分析，说它相当于相信（缺乏任何经验证据）X 的所有值的可能性是相同的。就是说，对于所有 $k=0$、…、n 的值，$\text{Prob}(X=k)=1/(n+1)$。这与 θ 的均匀先验相一致，并且他得以推理出一个正确的结论，而无需求助于高尔顿的技术支持，但这些只能在二项分布这样狭窄的情形中成立。拉普拉斯的处理方式更加一般化，他使用了一个毫无掩饰的假设——$p(\theta|x)$ 必须和 $p(x|\theta)$ 成比例，这与均匀的先验分布相一致。拉普拉斯的方法尽管没有使他在简单的问题上误入歧途，但却让他在更高的维度上犯了严重错误，尽管这一点他也许从未意识到。贝叶斯在整个 19 世纪都受到了忽视，大多数人不加鉴别地追随着拉普拉斯。

大多数人认为，贝叶斯推断是理想的推断形式。它提供了科学家想要的精确答案：根据手头的数据，完整描述一项研究目标的不确定性。而且，许多人相信，就像大多数的理想一样，这在实践中是达不到的。因为手边的构成要素通常并不清晰，特别是先验概率的指定。1885 年后，数学的发展可以支持更一般的思想表述，但依旧困难重重。从 20 世纪 20 年代起，哈罗德・杰弗里主张使用某些所谓的"参考先验"，这种先验分布带有一定的不确定性，对测量尺度并不敏感，并合理地（至少对某些人）反映了缺乏的信息。20 世纪 50 年代，布鲁诺・德・费奈蒂和吉米・萨维奇愿意拥护一种个人化的贝叶斯推断。在他们那里，每位统计学家都寻求自己的真实信念评价作为先验，哪怕每个人的先验不同会导致不同结论。最近，其他人主张一种"客观

的”贝叶斯方法，在这种方法中再次使用先验的参考以表示先验信息的缺乏。统计学家寻求可以得到这样知识的其他方法作为补偿，不基于强大的先验信息，至少也会得到在质量上相似的结论。高维的问题更严重，在一维或二维中看似自然的假设力量到高维中就是难以辨别的。并且，尽管作用在结论上的效果会很强且难以预见，却有可能获得赞同。

5.7　收缩估计

高尔顿工作的主要成就是引入了多元分析。而回归悖论的解释却不太重要，也不太成功，即所谓高个的父母的子女不会太高，而高个子女的父母却不太高。缘于回归的错误理解而发生的失误一直都有，并且无处不在。

1933 年，西北大学经济学家贺拉斯・塞奎斯特出版了 *The Triumph of Mediocrity in Business* 一书，这本书完全建立在统计错误之上。例如，他观察到，如果你在 1920 年列出利润率最高的前 25%的百货公司，并且跟踪这些公司的平均表现到 1930 年为止，那么会发现它们的业绩表现不断趋于行业平均值，走向平庸。即使塞奎斯特知道回归，他也没有理解它。塞奎斯特这样写道：“在商业中，走向平庸的趋势不仅是统计的结果，更表现了普遍的行为关系。”他浑然不觉的是，如果根据 1930 年的利润选择前 25%的公司，效果将会发生逆转。1920~1930 年，业绩表现会稳定地远离平庸。贯穿这本 468 页的书，他对几十个经济部门分别重复了这种错误言论。

20 世纪 50 年代，查尔斯・斯坦暴露了另外一个有关的悖论。假定有一组独立的测量 $X_i = 1$、⋯、k，每个都是一个分别的均值 μ_i 的估计。这些 μ 可以是完全无关的，但出于简化考虑，假定每个 X_i 有一个

正态(μ_i, 1)分布。这些 X_i 可能是 k 个不同人标准化的考试得分，或者不同行业的 k 家公司利润的标准化估计。在那时，应该根据对应的 X_i 估计每个 μ_i，这被认为是太过明显而无需证明的。斯坦表明这是错误的。如果采用估计的误差平方和最小化这个全局目标，尤其通过将所有 X_i 向 0 的方向收缩，就可以发现一组更好的估计。收缩的方法是利用一个只依赖于各个 X_i 的量，例如，使用 $\left(1-\dfrac{k}{S^2}\right)X_i$，其中 $S^2=\sum_{i=1}^{k}X_i^2$。

斯坦的悖论可以解释为某种版本的回归。考虑 k 对观测(μ_i, X_i)，考虑所有 X_i 的简单线性函数的可能的估计，即任何形如 bX_i 的估计。“明显”的估计恰好是 $b=1$ 的这个。但将误差平方和最小化作为目标时，如果我们有成对(μ_i, X_i)作为数据，那么 b 的最佳选择会是在 X_i 上 μ_i 的线性回归的最小二乘估计。这会是 $b=\dfrac{\Sigma\mu_iX_i}{\Sigma X_i^2}$。但我们不知道 μ_i，如何确定它们是分析的要点。再一次地，我们能估计 b 的分子。这是一个基本的练习，以显示 $\mathrm{E}(\mu_iX_i)=\mathrm{E}(X_i^2-1)$，所以用 $\Sigma(X_i^2-1)$ 替换 $\Sigma(\mu_iX_i)$，并仅仅使用 $b=\dfrac{\Sigma(X_i^2-1)}{\Sigma X_i^2}=1-\dfrac{k}{S^2}$。斯坦证明，这给出了（在此所用的假设之下）比起“明显的”估计更小的期望的误差平方和。无论 μ_i 是什么，只要 k 不太小即可（$k\geqslant 4$ 将适合这里给出的估计）。高尔顿不会对此感到惊讶：“明显的”估计 X_i 落在线 $\mathrm{E}(X_i\mid\mu_i)=\mu_i$ 上，这是他已经识别出来的错误的回归线。对应 X 在 μ 上，而不是 μ 在 X 上。

5.8 因果推断

在今天，“相关关系并不隐含因果关系”，统计学家普遍赞同这样的陈述。这个版本的陈述甚至早于 1888 年“相关系数”的发明。哲学

家乔治·伯克利 1710 年这样写道："思想的联系并不隐含着因和果的关系，它们只是标记事物的一个记号或符号。"现代技术的版本似乎出现在 19 世纪 90 年代晚期。在一个调查中，皮尔逊发现了令他吃惊的事情。男性头盖骨的长度和宽度本质上不相关，女性的头盖骨也是如此。但如果将男性和女性的头盖骨混在一起，情况就变了。对于这个合并的组，由于两组均值不同，因此那些相同的测量表现出显著的正相关性。在他的描述下，平均而言，男性在两个维度上都要更大一些。考虑（在一种极端的情况下）把合并的组画成两个不相交的圆簇，分开显示时没有关系，但合起来就会显示出它们的中心决定的关系。

皮尔逊将这称为"虚假相关"，他写道：

> 这种相关性完全可以称为"虚假的"，但保证任何社会的绝对同质性几乎是不可能的，因此我们的相关结果总会倾向于犯错，其数量无法预言。而那些坚持把所有相关关系都看成因果关系的人，通过将两个密切同源的种群人工地混合在一起，可以在两个非常不相关的特征 A 和 B 之间产生相关性，这样的事实必然引来相当大的震动。

长期以来，人们广泛接受这个问题，却也同时强烈希望接受相反的观念，认为找到一种相关就有助于支持某种程度的因果联系推断。当然，其中的一些愿望是自欺欺人，尤其当一位对因果关系抱有强烈先验信念的科学家对于自己的结论做出发现相关性的不严谨陈述的时候。但这些年来，一系列统计技术发展起来，通过潜在允许这样的陈述——诸如"如果这些假设可以满足，那么相关性并不隐含因果关系"——为因果推断提供了方法，继而衍生了一系列随着使用方法变动的条件。

其中的某些条件，与其说是数学的，不如说是哲学的。1965 年，

奥斯汀·布莱德福·希尔给出了一个系列条件，包含 7 个一般性陈述。他认为这些足够应付流行病学中的因果推断。这些条件都是实际可用的，并没有试图做出严格的定义，而使用了“力量”和“关系的一致性”“可能”和“联系的继承性”这样的术语。这 7 个条件中，有一个被他称为“时序性”，在本质上陈述了声称的原因必须先于结果。但是，这些条件尽管看起来在生物学或物理学中是合理的，但在社会科学中的作用尚不明确。西蒙·纽康在一本政治经济学教科书中给出过一个反例，比希尔的系列条件发表还要早 80 年：

> 这种看待经济现象的方法（即假定的时序性）如此自然，因此需要举证它的危险性的某些表现。假定有这样一位研究者，他通过统计观察试图了解奎宁和公共健康之间的关系。这位研究者可能会这样推理：“如果奎宁有益于间歇热的治疗，那么在人们摄入最多奎宁的那些地区，间歇热会最少。随着每一次新的奎宁进口，将会带来公共健康的发展。但实际上，我们的发现恰恰相反。沿密西西比河谷的低地以及海湾国家湿地生活的人们，比国内其他地区的人摄入更多的奎宁。但他们远远谈不上更健康，比起其他人，他们遭受了更多的间歇热。不仅如此，我们还发现奎宁的大规模进口发生在每年的夏天，接着到了秋天，间歇热发病的频率就规律性地增加。”

我们的预见能力确实在使事情复杂化。

其他更严格的方法包括数据相互依赖的假设结构，比如，某些偏相关是 0；或给定某些变量，其他一些变量是条件独立的；再或者，引入反映假设因果性的“结构方程”。1917 年，休厄尔·赖特发现，构建不同变量之间的有向图将相依的每个方向用一个箭头表示，只要图

中没有环，就能轻松计算成对相关系数（如图 5-16 所示）。随后，他把这个方法称为“路径分析”。赖特最初的工作是研究孟德尔的遗传结构，这在本质上是数学的，而非因果的。但是，他在后来的工作例子中引入了因果推断。最早的一项应用是在 1917 年，可以用来解决皮尔逊讨论过的问题。令 L 为头盖骨长度，W 为头盖骨宽度，S 为性别（M 或者 F）。那么，他的方法应用于混合的头盖骨，则产生下面的协方差关系：

$$\mathrm{Cov}(L, W) = \mathrm{E}\{\mathrm{Cov}(L, W \mid S)\} + \mathrm{Cov}(\mathrm{E}\{L \mid S\}, \mathrm{E}\{W \mid S\})$$

皮尔逊认为，右侧的第一项会接近 0，两个子群均值之间的关系（第二项）占据支配性的地位。

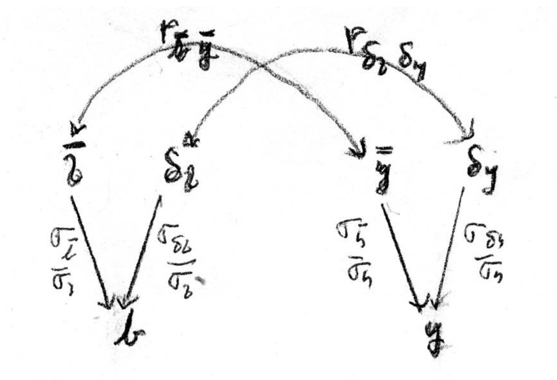


图 5-16　休厄尔·赖特 1917 年的第一个路径分析，根据 1975 年 4 月赖特和作者的私人通信重新画出（Wright 1975）

赖特的方法成为后来很多工作的先驱，包括无环图的因果模型，以及经济学家的结构方程模型。许多现代工作都是沿着这个脉络组织的——从强大的假设推导严格的结论——然而也面临着强烈的警告，这种假设通常不像孟德尔的遗传例子中那样明显真实。

5.9　三分律：愿你安息

到 19 世纪末期，三分律已经被丢入数学史的垃圾箱。今天，无论数学专业的学生还是教师，都基本上不了解由它命名的定律。而且，这个名字偶尔会用到其他无关的地方，从未收获过多少人追随。那个经常受到引用的问题，“耶稣会士会三人一组地去死吗？”（Do Jesuits die in threes? 受到这样一个事实的鼓舞，随机事件似乎只会随机地聚集在一起）已经是如今大多数统计学家最接近术语“三分律”的时候了。只有 *Annals of Eugenics* 封面上，皮尔逊对达尔文 1855 年陈述的重复还在。杂志在 1954 年更名为 *Annals of Human Genetics*，并在 1994 年重新设计了封面，这个重复引用才终于消失。在这个对三分律不再关心的世界上，无人哀悼它的消失，甚至其消失受到了无条件的欢迎。但这个责任不需要由高尔顿承担。数学自身的成长和发展使得三分律逐渐成为代数中一个次要部分，甚至不值得获得一个命名。它曾是每本数学教材指定的部分，是英国公务员考试的要求。但即使这个名称已经死去，这种思想的统计误用依然不断出现在常见的考虑不周和天真的外推法之中。

即使在三分律的全盛时期，公众对此也抱有一种悲观的看法。而且，它可能驱使许多学童离开了数学，就像今天很难讲授的三角和微积分。1850 年，约翰·赫歇尔已经在一本书评中承认了三分律应用的局限性，即使他当时还不知道高尔顿的看法：“三分律已经不再是政治数学家们的最后希望，也不再是——在规则由来已久的领域中通过给出武断的或不必要的假设——促使内容简化从而得到解决的问题。”

1859 年，一部戏剧被弗兰西斯·塔尔福德冠以“三分律”的名字，在伦敦短暂上演。这是一部单幕喜剧，剧中男子希瑟伯尔怀疑自己可爱的妻子玛格丽特容易受到其他男人的诱惑，他筹划着暗中破坏一个

事实上根本不存在的关系。希瑟伯尔的密谋使他几乎失去自己的妻子，但最终皆大欢喜。这部戏以玛格丽特为他朗诵的一首诗结束，这首诗反映出，即使在作为人生指南的年代，三分律具有的可信度也比较低：

> 你会做得很好，只要将诡计丢在一边，
> 女人的荣誉是心灵的最佳指南。
> 要可靠，也要信任，永远铭记这一点，
> 怀疑另一半之前，或许她已经消失不见。
> 什么填补了空缺的位置？啊，谁人可以诉说？
> 背信弃义者的不义之举终将平息，
> 全部的婚后生活永远不能
> 由任何三分律来制定，来检验。

第 6 章

设计：实验方案和随机化的作用

第六根支柱是设计，例如实验中的设计。但这个术语受到了更广泛的解释，包括一般的观测计划、决策结果分析以及执行方案。设计包括积极实验的计划、研究规模的决定、问题的设计以及处理的安排，还包括田野试验和抽样调查、质量监督和临床试验，以及在实验科学中的政策和策略评价。所有这些情况下，方案受到预分析的指导。设计甚至会在被动观测的科学中发挥关键作用，那些领域很少或不会控制数据的生成。如果你有能力生成数据以解决手边的主要问题，那么你会寻求什么样的数据？思考这样的问题可以使任何观测研究的关注变得更清晰。因此，设计可以在任何统计问题上训练我们的思维。

一些设计的例子很古老。《旧约全书·但以理书》中，但以理拒绝食用尼布甲尼撒王赏赐的丰盛酒肉，提出按照犹太教的规定食用洁食。国王的代表接受了但以理的建议，这在本质上是一种临床试验：整整 10 天，但以理和他的三个同伴只吃素食、饮清水，接着比较他们和另一组人的健康——那组人只吃国王的丰盛饮食。10 天后通过外观判定两组人的健康，最后但以理这一组赢了。

阿拉伯医学家阿维森纳（伊本·西纳）在公元 1000 年前后写过

Canon of Medicine，这本书中讨论了有计划的医学试验。随后的 6 个世纪，阿维森纳的论著在医学中一直占据领先地位。在这部书的第二卷中，他列出了医学试验的七个原则。这些原则根据亚里士多德的古老思想进行提炼，将药物作用归因于四项基本特性（热、冷、湿、干），阿利斯泰尔·克隆比将这七项原则翻译如下。

(1) 药物必须避免受到任何外来的、偶然特性的影响。例如，我们不能在水被加热时测它的作用，而应该先等它冷却。

(2) 进行实验的疾病必须选择简单的而非混合的。因为在第二种情况中，不可能从药物的治疗中推断治愈原因。

(3) 必须通过两种类型相反的疾病测试药物。因为，有时药物通过其本身的特性治愈一种疾病，而通过偶然的特性治愈另一种，所以不能因为它治愈了某一类型的疾病，就简单推断这种药物必然具有某种特性。

(4) 药物的药力必须匹配疾病的力量。例如，有一些药物的“热”少于某类疾病的“冷”，那么这些药物对此类疾病无效。因此，实验应该先对较弱类型的疾病进行，再对强度逐渐增加的疾病进行。

(5) 必须观测活动的时间，这样才不会混淆本质和偶然性。例如，热水因为获得了外在的偶发事件，会暂时具备加热效应。但一段时间之后，它就会回到寒冷的本质。

(6) 必须看到，药效会常常发生，或在许多情况下发生。因为如果没有出现这种效果，那就说明这种作用是偶然的。

(7) 实验必须对人体完成。因为对狮子或马测试药物不会证明对人可能有任何作用。

我们可以使用现代眼光这样解读这些规则：它们强调控制和复制的必要性，强调混合效应的危险性，以及强调对多个不同因子水平效应的观测智慧。人们甚至可以把这些规则看成一般因果推理的早期表

达方式。从阿维森纳开始，有什么事情改变了吗？或者对同样的问题，从亚里士多德开始算起呢？好吧，老鼠已经取代狮子，成为实验室动物的首选。但再来看看阿维森纳的第二条准则。他说，实质上一次只要做一个因素的试验。对此，威廉姆·斯坦利·杰文斯在 1874 年的《科学原理》中有一个更现代的版本。他这样写道：

> “实验中最重要的预防措施之一是，每次只能改变一种情况，同时要严格保证所有其他情况不变。”

现在读一下费舍尔，他 1926 年这样写：

> “我们必须每次只向大自然提很少的问题，或者可能的话，只问一个问题。没有与田间试验有关的哪一句格言比这一句会更被频繁地重复。但笔者（费舍尔）确信这个观点完全错误。他认为，大自然会对一个逻辑严密、思虑周全的调查问卷提供最好的答案。事实上，如果我们只向她问一个问题，大自然通常会拒绝回答，直到其他一些主题也被一起讨论。”

6.1　可加模型

费舍尔放弃了 2000 年来大部分实验的哲学与实践观点，使用了一种具有伟大独创性的统计观点。费舍尔的多因子设计孕育于他在罗森斯塔特实验站的农业研究经验，彻底改变了实验过程。在相同的农业地块上，他会同时改变种子、肥料和其他因素，种植在拉丁方阵或希腊–拉丁方阵这样的阵列上，并回答所有这些因素对单个种植季、单个种植阵列的问题。

多年来，农业实验者已经尝试过各种安排。1770 年，阿瑟·杨提

出了同时运用于相同地块的比较方法。例如，比较播种与畦植，同时需要避免受到其他变化因素的误导。但是，杨几乎没提该如何分割地块，只是说了“相等的”。到 1919 年费舍尔抵达罗森斯塔特时，那里的研究人员使用的是棋盘格和夹层设计，目的是使两种不同的处理地块彼此相邻，从而最小化土地差异。可以说，这是区组的一种尝试，尽管这些实验缺乏允许这种区组效应的分析。费舍尔提供的分析和实验逻辑的组合是崭新的。他不仅认识到，农业中的统计变异产生了必要的新方法；还看到，如果执行有经济效率的试验，考查变异也会指向答案。而费舍尔的思想中，令人印象最深刻的部分在于他使用的复杂设计——既结合了层次结构，还允许估计交互作用，而收益已经存在于一个更简单的水平上。

下面考虑对一块实验田上的作物产量使用一个可加模型。田地分割成一些（比如说 $I \times J$ 个）地块，每个地块安排不同处理的组合。可以使用代数方法表达模型，用 Y_{ij} 表示地块（i, j）的作物产量，并假设产量是总体均值的简单加总——各种处理单独的效应以及每个地块的随机变动。对 $i=1$、⋯、I 和 $j=1$、⋯、J，令 $Y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}$，其中 μ 代表整块田地的平均收益，α_i 代表种子的种类 i 的效应，β_j 代表了肥料水平 j 的效应，地块（i, j）的随机变动 ε 则归因于不可控的因素。

1885 年，弗朗西斯·埃奇沃思对这种可加模型提出一个相关的口头描述。该描述的数学化程度没这么高，但表达得既雄辩又明晰，拥有一种形式化模型无法企及的美。埃奇沃思写道：

> “由于温和的地质作用，若干台地产生了一个城市的选址。台地从东至西彼此平行。火山岩的位移产生山脊，山脊垂直分割台地。我们或许可以猜测，火山以恒定的速率从西向东运动，每年产生一座宽度相同的山脊。在观测之前，我

们尚不知道，（任何一块或所有台地）一年的位移是否与近几年的位移差不多。类似地，我们也不知道，一块台地的位移是否与相邻的台地趋同。于是，地面上建起了纵横交错的各式房屋。每个屋顶的海拔高度都可以根据气压或其他方法确定。每英亩的屋顶平均海拔高度都有登记。”

在这里，Y_{ij} 是英亩(i, j)上房屋的平均海拔高度，μ 是整个城市的房屋平均海拔高度，α_i 是台地 i 上的地面位移效应，β_j 是山脊 j 的地面位移效应，ε 是在英亩(i, j)中距离该英亩的房屋高度均值的变差。

其要点是，这个模型以结构化的方式结合了 3 种变异的来源。这种方式允许它们分别受到处理，即便没有哪两个地块的处理组合会完全相同。在一个模型中综合所有这些变异，会产生巨大的优势。如果在某种方法中，数据遗漏了某个因素（例如肥料或山脊），那么这种遗漏引起的变异会使其他因素和不可控因素引起的变异变小，从而使其他因素（例如种子种类或台地）的检测或者估计不可行。但是，如果同时综合考虑两种因素（在某些应用中，费舍尔称为“区组”），二者的效应可能超越行均值或列均值及其变异,因而两者的效应清晰可辨。即使在一个基本的可加效应例子中，结果也可能很惊人。而在更复杂的情况下，结果可能带来史诗级的震撼。

接下来，我们举一个例子，清楚展示可能错失什么。考虑一个著名的数据集，它由拉迪斯劳斯·冯·鲍特凯维茨在 19 世纪 90 年代提出，鲍特凯维茨将这个数据集放在自己 1898 年发行的简短小册子《小数定律》中。收集这个数据集耗费了大量精力，是根据海量的普鲁士国家统计编辑而成的。数据给出 14 个骑兵团 20 年间被马踢死的骑兵人数（如图 6-1 所示）。鲍特凯维茨希望论证，如果数字这么小并且不可预测，那么其中大的变异可以掩盖真实效应。而且，根据他的展示，

一起观察这 280 个数可以发现，它们很好地符合一组独立分布的泊松变量的集合——也确实如此。但是，当时的鲍特凯维茨没有掌握可加模型的技术。如果在此运用可加模型（使用带泊松变异的一般线性模型），那么不仅可以清晰展示团与团的变异，还可展现年与年的变异。骑兵团和年份的变异并不大，但可加模型允许 14 + 20 个单独效应捕捉它们。利用 240 个单独观测，这个分析可以检测这些效应。在埃奇沃思的城市里，如果只看屋顶，肉眼即可看到山脊和台地的变异；而如果沿着山脊和台地观察，则可看到房屋的变异。鲍特凯维茨似乎已经预计到存在团和团的变异，但随机变异掩盖了这种变异——毕竟各团规模不同。年度变化很可能出人意料。

	75	76	77	78	79	80	81	82	83	84	85	86	87	88	89	90	91	92	93	94
G	—	2	2	1	—	—	1	1	—	3	—	2	1	—	—	1	—	1	—	1
I	—	—	—	2	—	3	—	2	—	—	—	1	1	1	—	2	—	3	1	—
II	—	—	—	2	—	2	—	—	1	1	—	—	2	1	1	—	—	2	—	—
III	—	—	—	1	1	1	2	—	2	—	—	—	1	—	1	2	1	—	—	—
IV	—	1	—	1	1	1	1	—	—	—	—	1	—	—	—	—	1	1	—	—
V	—	—	—	—	2	1	—	—	1	—	—	1	—	1	1	1	1	1	1	—
VI	—	—	1	—	2	—	—	1	2	—	1	1	3	1	1	1	—	3	—	—
VII	1	—	1	—	—	—	1	—	1	1	—	—	2	—	—	2	1	—	2	—
VIII	1	—	—	—	1	—	—	1	—	—	—	—	1	—	—	—	1	1	—	1
IX	—	—	—	—	—	2	1	1	1	—	2	1	1	—	1	2	—	1	—	—
X	—	—	1	1	—	1	—	2	—	2	—	—	—	—	2	1	3	—	1	1
XI	—	—	—	—	2	4	—	1	3	—	1	1	1	1	2	1	3	1	3	1
XIV	1	1	2	1	1	3	—	4	—	1	—	3	2	1	—	2	1	1	—	—
XV	—	1	—	—	—	—	—	1	—	1	1	—	—	—	2	2	—	—	—	—

图 6-1　鲍特凯维茨的数据收集自卷轶浩繁的普鲁士国家统计（这段时期每年出版三大卷）。他列入了 14 个团 20 年的数据（Bortkiewicz 1898）

6.2　随机化

大卫·考克斯曾这样描述统计中的三个随机化规则：“一个是消除偏差的策略，例如，在不可观测的解释变量和选择效应中可以见到；

一个是估计标准差的基础；最后一个是用于规范精确的显著性检验的基础。”其中第一条规则受到了最广泛的认可，甚至可以在主流文化中发现它的踪迹。漫画书《功夫大师》1977 年 7 月号中有这样的情节：大师根据思想气球随机选择了一张音乐专辑，“我从丰富的专辑中偶然选择了一张。它毫无偏见，令人耳目一新，尽管无知是唯一的救星。”但是，另两个原则（它们彼此相关）更微妙，并且统计重要性更大。通过这些规则，随机化在许多方面对推断而言变得很基础，特别是在设计相关的问题上，甚至在某些真正定义了推断目标的情况下。

19 世纪晚期，查尔斯 · S. 皮尔斯认识到这样一个事实：随机样本使得推断变得可行。他甚至对“归纳”给出这样的定义：“随机抽样，从一个样本直到许多样本，再对整批样本进行推理。”19 世纪 80 年代早期，约翰 · 霍普金斯大学刚刚创立，皮尔斯在这里研究实验心理学。可以说，实验心理学就是实验设计创建的领域。古斯塔夫 · 费希纳在 1860 年前后完成了举重法对刺激和感觉影响的早期研究，先定义了实验的方案，并给出了实验目的的意义。这个实验的内容是，实验者（或助手）轮流举起两个小容器中的每一个，每个容器的基本重量为 B，其中一个（并且唯一）还包含一个不同的重物 D；举起容器的人需要猜测，根据举重的感觉判断哪个容器更重，即哪一个的重量是 B，哪一个是 $B+D$？实验重复数百甚至上千次，B 和 D 不断变化，并且使用不同的手和不同的举重顺序。有人把这种方法称为“正误法”。收集到的数据允许实验者估计正确猜测的概率如何随着 D、B 以及举起重物的手而变化，并使用现在所谓的 probit 模型。这个模型假定 $D=0$ 时概率为 0.5，D 变大时概率逐渐提高，D 非常大时概率渐近于 1。当时，提高的速度常常用于测量与实验条件相关的灵敏度。如果没有这个实验，理论就会空泛，或者至少无法量化。相似地，在 19 世纪 70 年代，赫尔曼 · 艾宾浩斯进行了与短期记忆强度有关的大量实验，使用了一

个精心制定的实验计划，运用了无意义音节。

1884~1885 年，皮尔斯研究了一个更加微妙的问题，对实验方法的要求更进了一步。早期的心理学家曾经猜想，两种感觉的差异有一个阈值，可称为“最小可觉差”（just noticeable difference）。如果差异低于这个阈值，就无法区分两种刺激。皮尔斯与约瑟夫·贾斯特罗一起设计了一个实验，说明这种猜测是错的。他们改进了举重试验，设计了一个新版本，其中两个重物相差质量 D 非常小，这样一个重物只比另一个稍重一点。皮尔斯和贾斯特罗表明，随着重量的比例逐渐接近 1，正确判断的概率平稳接近 1/2（但保持了可以察觉的差异）。没有迹象表明存在 jnd 理论声称的离散阈值效应。

很明显，这个实验非常精细。轻微的偏差或提前知道了重物展示的顺序，都会让这项实验泡汤。皮尔斯和贾斯特罗使用了庞杂的防备措施，并做了记录，以确保在不知情的情况下做出判断。同时，他们还引入了一种全面并严格的随机化方式——使用一副完全洗好的卡片决定重物展示的顺序（先是重的还是轻的）。此外，整个研究过程中，对每一次判断的准确性信心 C，他们都做了一个主观的记录：$C = 0$（最低的信心）、1、2、3（最高的信心）。皮尔斯和贾斯特罗发现，正确猜测的概率 p 越大，信心越强，近似于 $\log(p/(1-p))$的倍数，这个对数是猜测正确胜率的对数。他们还发现，一个人对概率的看法是对数胜率尺度的线性函数。这个实验的有效性以及归纳的结论都很关键地依赖于随机化。

20 世纪早期，费舍尔把这个主题推得更深远。我之前提过，他认识到在多因子设计中使用组合方法进行设计的优势。1925~1930 年的 5 年间，费舍尔扩展了这个设计的复杂度。他看出，在那些复杂的设置中也可以验证推断随机化的行为。最简单的情景中，在成对对象内，随机指派的处理和控制有效推断了处理效应，而无需任何分布假设

（超越了对不同成对对象独立性的依赖）。随机分布本身诱导出一个关于情况个数的二项分布，在那里的“无差异”原假设下，处理超过控制的概率是 1/2。

现在，回头思考约翰·阿布斯诺特的洗礼数据。82 年的数据中，每一年都按性别分类（见第 3 章）。可以肯定，这些数据不会是设计好的实验产物。把洗礼视为出生数据的替代，并假设两种性别的出生概率相等。由此，阿布斯诺特计算出，接连 82 年中，男婴出生的概率都超过女婴的概率是 $1/2^{82}$。用这个检验评估出生频率可能会受到批评：与女婴相比，男婴是否更不容易记录在洗礼的教区记录中？而且即便如此，两种性别的婴儿死亡率在洗礼之前一样吗？他的数据无法回答这些问题。对于观察性的研究，如果没有其他选择，我们通常会接受这些假设。但设想一下（违背所有逻辑），为了解决“男婴和女婴的洗礼无差别”的假设，可以通过随机指派出生时性别的方法，作为实验的一部分。于是，在这个假设下，人们设计的随机化会保证在一年中更多男性受洗的概率是 0.5（少于平局机会的一半），并且阿布斯诺特的 $1/2^{82}$ 会在这个假设之下成为数据概率的评估。我们甚至可以回想第 1 章，意识到一个更好的检验会聚合更深。对于 82 年间的 938 223 个洗礼记录，其中 484 382 个为男婴，标准差为 31.53，高于性别概率相等的期待值，概率与 $1/2^{724}$ 相距不太远。当然，阿布斯诺特没有使用随机化。而且对他来说，不相等的出生频率和不相等的记录混淆在一起显得毫无希望。但如果有可能，随机化本身可以提供推断的基准。

多因素田间实验中，费舍尔的随机化设计实现了多个目标。完全随机化的行动（例如拉丁方阵设计中的随机选择）不仅允许交互作用的分离和估计，还可能进行有效推断，其方式不依赖于正态假设或者材料匀质性的假设。费舍尔意识到他的检验——多种 F 检验——在原假设之下，只需要球对称就可以发挥作用。正态性和独立性隐含着球

对称，但这些不是必要条件。设计的随机化本身就能引起离散的球对称特性，对看似要求更强条件的过程保证了近似的有效性，正如随机化处理引起皮尔斯举重实验中的二项变化。这个微妙的要点并没有受到广泛的掌握。即使像威廉·西利·戈塞特（“学生”）这样聪明的统计学家，直到 1937 年生命终结时，依然坚持“系统化的田间实验（比如夹层设计 ABBABBABBA…）会比随机试验提供更好的估计”，并且认为两者都需要正态性。费舍尔很尊敬戈塞特，但在 1939 年的讣告文章中，他这样写：“尽管无疑的是（戈赛特）实践了它，但他并不总是意识到随机化的必要性，或获得系统化实验的理论不可能性。这种实验中，真实误差和估计误差都应小于同一地块随机化时给出的值。这种特殊的失败也许仅仅表现了他对同事们的忠诚——这些同事正在这方面承受批评。”

这些方法的广泛普及不但缓慢，而且通常只是部分性地受到使用。近似随机化的模糊形式很早就有过实践。从大约公元 1100 年开始，用于铸币检验的硬币就是“随意地”选择的，或至少没有故意的偏差。1895 年，挪威统计学家安德斯·凯尔推广了一种方法，他称之为“代表性抽样”，有针对性地选择创建样本的目标。从不完全的意义来说，这个样本是总体的一个缩影。

1934 年，杰日·奈曼为皇家统计学会宣读了一篇有影响力的论文（这是他和费舍尔最后一次享有表面的同事关系）。奈曼的论文发展了随机抽样的理论，这种方式严格地实现凯尔的目标。在讨论中，费舍尔赞同论文的这个部分。他注意到，只用于样本选择的随机化在社会科学中的应用与费舍尔在农业中运用的方法通常不同。在费舍尔那里，不同的处理随机地强加于实验单位。“不幸的是，这个实验的随机化过程不能在社会调查中被效仿。如果能，那么人类事务中的因果关系一定比已知的要多得多。”在接下来的 20 年，社会科学中的随机抽样（以

非扩散性的方式）方法腾飞了——经常带着各种变异，比如关注子总体（分层抽样）的，或作为序列过程的一部分（例如“雪球抽样”）。

费舍尔的随机化方法极具扩张性，在一个领域中取得了重大进展——医学或临床试验。在那里，随机指派处理是可行的，就像皮尔斯在举重实验和费舍尔在罗森斯塔特做过的那样。而且，费舍尔的工作吸引了奥斯丁・布拉福德・希尔的注意。通过希尔的大力宣传，这种方法在那些抗拒改变的医疗机构中缓慢但稳步地推进着。今天，随机临床试验被视为医学实验的“黄金标准”，尽管在一些情况中，研究人员觉得他们担负不起“黄金”。

还有一个领域广泛实践了随机化设计，但这里从来不提那些术语，并且经常受到谴责——彩票。彩票将随机化引入社会过程，并对那些自愿选择的人指派了处理。对有些人来说，彩票是一种娱乐；对另一些人来说，彩票是对他们收取的“智商税”。但彩票的历史悠久，并毫无消失的迹象。另外，需要注意，它们会产生一些科学的红利。举一个例子就够了。

法国的彩票设立于 1757 年，它模仿了早期的热那亚彩票，很像今天的“乐透”。法国彩票一直到 1836 年才被废除，之前只在 1794~1797 年的法国大革命恐怖时期有所间断。在正规场合，从一组 90 个记有 1~90 的球当中不重复地取出 5 个，选择在本质上是随机的。玩家可以通过指定所有 5 个数字下注（quine），或者 4 个数字（quaterne）、3 个数字（terne）、2 个数字（an ambe)或者 1 个数字（extrait）的列表。如果一位玩家选择的数字以任何顺序、任何位置出现在抽出的 5 个数字中，则为获胜。有时候不允许 quine 赌注，因为存在欺诈风险（因为受贿的代理可能会在知道抽出的数字后卖出一张彩票）。但当允许时，它支付 1 000 000 比 1（一个公平的赌注会支付大约 44 000 000 比 1）。胜率优于发生更频繁的情形，extrait 会支付 15 比 1（18 比 1 是公平的），

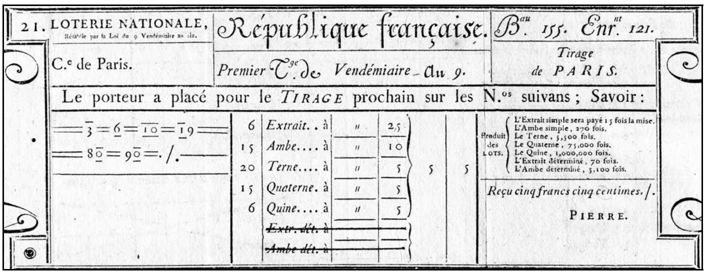
ambe 是 270 比 1，terne 是 5500 比 1，quaterne 是 75 000 比 1。

玩家通常会一次投下多个赌注。例如，对图 6-2 所示的投注单，玩家赌了 6 个数字：3、6、10、19、80 和 90，6 个对应的 extraits 每个下注 25 生丁，15 个 ambes 每个下注 10 生丁，20 个 ternes 每个下注 5 生丁，15 个 quaternes 每个下注 5 生丁，6 个 quines 每个下注 5 生丁，总共下注 5 法郎 5 生丁。所有收益都安排在一个固定的时间表内，并且国王会提供保证。那时没有彩金池保护国王（现在有）。在指定的日期，真正抽出来的数字是 19、26、51、65 和 87。这张票只赢了一个单个的 extrait（19），付给 $25 \times 15 = 375$ 生丁，或者 3 法郎 75 生丁，损失 1 法郎 30 生丁。如果抽出来的数字是 2、6、19、73、80，展示的这张票会收到支付一个 terne（6、9、80）、3 个 ambes（6 和 19、80 以及 19 和 80）和 3 个 extraits，总收益为 $5500 \times 5 + 3 \times 270 \times 10 + 3 \times 15 \times 25 = 367$ 法郎 25 生丁。

早些年，彩票收入用于支持圣西尔军校。到 1811 年，彩票的净利润提供了多达 4%的国家预算，比邮政或关税还多。1810 年前后，彩票达到销售高峰，超过 1000 家地方办事处发售彩票，而且在 5 座城市每月抽奖 15 次（但巴黎人可以对任何城市的彩票下注）。整个法国大革命期间，即使在路易十六和玛丽・安托马内特被处决时，抽奖依然没有间断。仅仅在恐怖统治达到某个高度后，玩家都丧失得到偿付的信心时，彩票才暂停。但只停了两年多，少量抽奖又恢复了。即使经过了拿破仑战争，抽奖也持续不断。直到 1836 年基于道德的考虑，彩票才被禁止。这是一场规模宏大的、真正的随机化实验。

纵览其整个历史，获胜数字被广泛公布，这就允许检查抽奖的随机性。彩票通过了所有可用的检验，包括数字联合发生的测试，其中 6606 种可用的抽奖集合可以允许合理的检验。这不奇怪！任何明显的偏差只会有利于参与者而伤害彩票。

彩票的一种社会后果是提高了数学的教育水平。为了评估下注，参与者学习组合分析，而彩票也为当时的教科书提供了许多例子。彩票管理者不得不培训大量本地操作员，以便这些销售代理能够准确地对类似显示的那种多重赌注定价。（为了达到这个目的，他们制作了特殊的文本，如图 6-2 所示。）



21. LOTERIE NATIONALE, Rétablie par la Loi du 9 Vendémiaire an six. | République française. | B.au 155. Enr.nt 121.

C.e de Paris. | Premier T.ge de Vendémiaire - An 9. | Tirage de PARIS.

Le porteur a placé pour le TIRAGE prochain sur les N.os suivans ; Savoir :

3 = 6 = 10 = 19 = 80 = 90 = ./.

6	Extrait. . à	"	25		
15	Ambe. . . . à	"	10		
20	Terne. . . . à	"	5	5	5
15	Quaterne. à	"	5		
6	Quine. . . . à	"	5		
	~~Extr. dét.~~ à				
	~~Ambe dét.~~ à				

Produit des LOTS.
L'Extrait simple sera payé 15 fois la mise.
L'Ambe simple, 270 fois.
Le Terne, 5,500 fois.
Le Quaterne, 75,000 fois.
Le Quine, 1,000,000 fois.
L'Extrait déterminé, 70 fois.
L'Ambe déterminé, 5,100 fois.

Reçu cinq francs cinq centimes. /.

PIERRE.

图 6-2　为彩票工作人员准备的说明手册，1800 年，用于演示的填写好的彩票样本，说明了如何明确无误地书写数字、如何记录，还有如何对一张组合的彩票定价。在彩票中，消费者指定 6 个数字，并对所有 5 个数字的子集投注，在不同奖项上投注不同数量（Loterie An IX）

另一种好处是，彩票无意中执行了可能是最早的科学随机化社会调查。法国大革命之后的时期，有关机构不仅公布获胜的数字，还公布 quaterne 及其以上水平的所有获胜的赌注。这些记录给出了支付的金额（从中可以发现赌注的规模）、售卖彩票的位置和机构个数。那些获胜者真的随机选择那些放置的 quaterne 赌注，因此，这个调查给出了一幅图景，刻画了法国的哪些地方对彩票最有兴趣（当然是巴黎，虽然全国的兴致都很高），以及彩票的吸引力如何随着时间变化。关于最后这个问题，结果显示，在最后的 20 年间，投注稳步下降。这让人们相信，只有利润下降到低于维持大型的运作所必需的水平时，才导致“道德问题”主导了政策。

第 7 章

残差：科学逻辑、模型比较以及诊断展示

我把第七根，即最后一根支柱叫作残差，这个名称暗示了它是标准数据分析的一部分。而在我的想法里——尽管这并非完全错误——其中有一个更大、更经典的科学逻辑主题。

约翰·赫歇尔的父亲威廉·赫歇尔发现了天王星，约翰追随父亲的脚步进入天文学领域。尽管父亲的第二职业是音乐，约翰却对数学和科学哲学感兴趣，并成为那一代最受尊敬的科学家之一。他 1831 年出版了著作《自然哲学研究初探》，这本书得到广泛阅读并产生很强的影响。在这本书中，约翰讨论了科学发现的过程。对自己所谓的残差过程，赫歇尔给予了特别强调：

“复杂现象往往存在几种相同的、相反的或者彼此完全独立的原因，它们糅合在一起，产生了复合作用。为了简化，可以排除已知原因的影响与这种情况允许的特点，也可以通过演绎推理或诉诸经验。留下的就是要解释的残差现象。事实上，科学正是通过这样的过程受到了极大的促进，表现出目前的先进状态。绝大多数现象的类型极其

复杂，而估计出所有已知因素的精确影响并排除后，残存的事实会不断呈现全新的现象形式，进而导出最重要的结论。”

从历史角度说，赫歇尔选择了一个不幸的案例：他把以太（aether，后来称为 luminiferous ether）的“发现”归功于这种推理。当时的人们认为，以太是一种充满了外层空间的物质，会发光，并且对牛顿理论中的一些异常现象负责。我们仍在寻找以太。但科学原则是有效的，也是重要的：我们尝试进行解释，然后看还有什么需要解释，这样能够学习新知识。

查尔斯 · 达尔文受到了赫歇尔的著作影响。据说，他因为受到这本书的鼓励而成为科学家。约翰 · 斯图尔特 · 穆勒 1843 年出版了《逻辑学体系》，这本书提出了赫歇尔的思想可作为实验探究的 4 种方法中最重要的一种。穆勒稍微修改了赫歇尔给出的名字，将之称为“残差的方法”。穆勒还写道，“在自然法则的所有研究方法中，这在意外的结果中最富有想象力。它常常提醒我们注意那些表现突出到足以吸引观察者注意的序列，虽然它们既不是原因，也不是影响。”

这种思想大致上是古典的，但这种思想的统计学发展进化出一种崭新而强大的科学方法，进而改变了科学实践。这种思想的统计学解释和相关的科学方法向它赋予了一种新的纪律性力量。统计方法描述了这样一个过程：根据假设模型生成数据，继而通过非正式的（例如通过图形或者图表演示）或正式的统计检验比较数据和模型的偏差，比较简单的模型和复杂的模型（比较两个“嵌套”模型，其中一个模型是另一个的特例）。

最早的例子包括小的、集中的嵌套模型，比较其中一个理论和另一个稍微更复杂的理论。一个最简单类型的好例子是 18 世纪的地球形状研究，我们曾在第 1 章讨论过。那里的基本（更小的）模型是将地球看作一个球形。为了检验这个模型是否成立，我们构造了稍微复杂

一些的模型：把地球看作椭球体——一个在两极被压扁或者拉长的球体。我们从地球中减去球体，看看残差表示的偏离，以及如果残差存在，方向会向着哪里。但是需要怎么做呢？对这样一种检验应该运用什么样的地球测量？18 世纪采用的途径是沿着一条子午线测量一系列的短弧长。如果 A 是纬度 L 上的 1°弧度的长度，那么假设地球是一个球体，A 应该在所有纬度上都相同，即 $A = z$ 。但如果地球是一个椭球体，那么为了获得良好的近似，应该使用 $A = z + y\sin^2 L$ 。如果地球在两极是扁的或长的，可以得到 $y > 0$ 或者 $y < 0$。所以，给定不同纬度的弧度测量集合后，问题就成了：如果 $\sin^2 L$ 增加，拟合 $y = 0$（即 $A = z$）产生的拟合残差是向上偏还是向下偏，或者两者都不是？更靠近极点的弧度相比于那些靠近赤道的弧度是更短还是更长？

图 7-1 展示了罗杰・约瑟夫・博斯科维奇自己的图，记为 XY 的线表示球体（$A = z$，其中 z 是数据的平均弧长），5 个数据点 a、b、c、d 和 e 表示 XY 的残差。水平轴是 AF，给出 $\sin^2 L$ 从 $A = 0$ 到 $L = 1$ 的值，垂直轴是 AX，给出弧长 A 的值。线 GV 是博斯科维奇通过自己的算法计算的，即这条线穿过 5 个点的重心 G，同时最小化了它们与这条线的绝对垂直离差之和。XY 的残差计算出斜率为正，这意味着地球是扁的。

这是嵌套成对模型的可能例子中最简单的情况，$A = z$ 是 $A = z + y\sin^2 L$ 的特例。这是最早对一个回归方程与另一个加入新的“预测变量”—— $\sin^2 L$ ——的回归方程进行比较的例子。尽管博斯科维奇并未提及概率（而且因此也没有给出属于 y 估计的不确定性的估计），但从那以后，他使用的基本方法已经成为模型适当性的统计解释的主流方法。

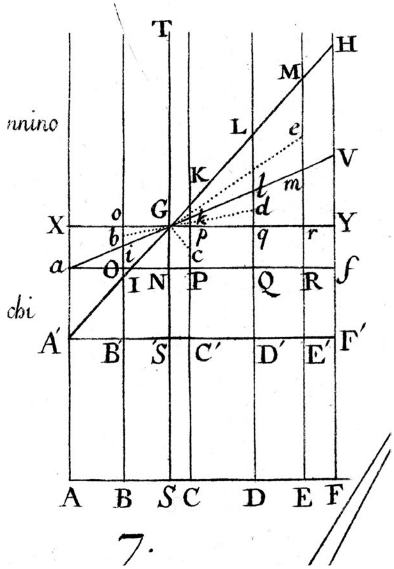


图 7-1　博斯科维奇自己的图，来自 1770 年的一份出版物，展示了表示球体的线 *XY*，和表示为 *a*、*b*、*c*、*d* 和 *e* 的 5 个 *XY* 的残差点。博斯科维奇通过自己的算法发现的线是 *GV*，是 *XY* 残差计算的斜率（Maire 与 Boscovich 1770）

自然地，这些嵌套模型出现在物理科学中。一个简单的模型越来越复杂，并通过局部逼近进行线性化时，往往需要在方程中添加一个或者多个项。根据方程的不同点取测量，会受到实验误差的影响。但 1805 年后，不断发展的最小二乘技术使计算更简单。18 世纪，莱昂哈德·欧拉、约瑟夫·路易·拉格朗日以及皮埃尔·西蒙·拉普拉斯尝试不同的方式，试图在行星运动的两体牛顿模型中引入三体引力时，需要面对新的问题：增加的项可不可以是 0？他们的研究基于观察到的残差效应而展开。随着人们更好地确定了木星和土星的轨道，以及很长一段时间内更仔细地检查数据，似乎可以认为，过去几个世纪以来，木星一直在加速，而土星在减速。如果这些巨大星体的轨道不稳

定，那对太阳系来说不是个好兆头。尽管这种针对行星速度改变的怀疑只是三体引力产生的结果——太阳、木星和土星。欧拉和拉格朗日都对这个研究做出了重要推进，但最终完成的是拉普拉斯。他找到了扩展运动方程的一种新方法：先在方程中引入更高阶的项，再检验明显的运动是否与这些高阶项的影响一致。拉普拉斯成功揭示了观察到的改变速度只是星体运动周期改变的一部分，周期长度大致为 900 年，原因是这两个行星平均运动的比率接近 5∶2。残差分析挽救了太阳系。

这些研究后来成为其他人学习的模范案例。19 世纪 20 年代，人们会通过检验显著性进行比较，或者至少比较未知差异和额外系数的可能误差（或 p.e.）的估计。拉普拉斯在 1825~1827 年关于大气层太阴潮的研究就属于这种类型的比较（见第 3 章）。但这种方法未能给出一种容易的实施路径，用以比较无嵌套的模型，哪个模型都没有只比其他模型相差一个残差。另外，本质更复杂的哲学问题并没有普适的方法，而模型的陈述中也不易得到“更简单”的含义。

从 20 世纪初开始，物理科学的线性方程开始用于社会科学，并且很自然地遵循着同样的路线。先指定一组“解释”变量，再多增加一些线性变量，最后考查它们是否导致显著性差异，或这些项是否在统计上与 0 无差异。乔治·尤尔 1899 年对“济贫法”（英国福利系统的一部分）进行了检查，这是最早使用该方法进行深入研究的实例之一。一次调查中，他考查了贫困水平与福利救济金额之间的关系，增加福利救济会导致贫困水平上升还是下降？尤尔比较了 1871 年和 1881 年的数据，根据一个市辖区 10 年来的“贫困”（贫困水平）变化，探寻和确定这个地区接受的“贫民救济”（福利）变化的比例效应（如图 7-2 所示）。这个简单回归的样本的地区不同，数据点也不同，尤尔还知道，其他经济因素也在变化。因此，他把这个问题重定义为与残差现象有关的新问题。先根据其他经济因素校正所有数据，再研究关系。

经过漫长而细致的分析，他发现这两个变量之间的一个正相关关系。

从那时到现在，解释这个结果的问题以及从相关关系推断因果关系的困难一直是个麻烦（而且尤尔很仔细地注意到了这一点）。但是，这仍然标志着社会科学研究开启了一个新时代。使用这种方法会面临诸多困难，比如，线性方程对于问题是开放的（尽管尤尔注意到，无论关系多么复杂，他都估计了最接近的线性近似方程）；再比如，“解释”变量之间的关系可能严重模糊问题的解释。如果小心使用，它会展示出技术解释的强大。不过，如果没有出现一项重要进展，该技术似乎难逃束缚于线性最小二乘的命运。这项重要进展就是参数化模型的引入。

used. Then suppose a characteristic or regression equation to be formed from these data, in the way described in my previous paper, first between the changes in pauperism and changes in proportion of out-relief only. This equation would be of the form—

$$\left.\begin{array}{l}\text{change in pauperism}\\ =\text{A}+\text{B}\times\text{(change in proportion of out-relief)}\end{array}\right\}-$$

where A and B are constants (numbers) (1)

This equation would suffer from the disadvantage of the possibility of a double interpretation, as mentioned above: the association of the changes of pauperism with changes in proportion of out-relief might be ascribed *either* to a direct action of the latter on the former, *or* to a common association of both with economic and social changes. But now let all the other variables tabulated be brought into the equation, it will then be of the form—

change in pauperism=

$$\left.\begin{array}{l} a+b\times\text{(change in proportion of out-relief)}\\ \quad +c\times\text{(change in age distribution)}\\ \quad \left.\begin{array}{l}+d\times\\ +e\times\\ +f\times\end{array}\right\}\text{changes in other economic, social, and moral factors}\end{array}\right\}(2)$$

Any double interpretation is now—very largely at all events—excluded. It cannot be argued that the changes in pauperism and out-relief are both due to the changes in age distribution, for that has been separately allowed for in the third term on the right; $b\times$(change in proportion of out-relief) gives the change due to this factor *when all the others are kept constant.* There is still a certain chance of error depending on the number of factors correlated both with pauperism and with proportion of out-relief which have been omitted, but obviously this chance of error will be much smaller than before.

图 7-2　尤尔描述贫困变动的多元回归方程（Yule 1899）

明确使用参数模型是罗纳德·A. 费舍尔的一项精妙创新，这项创新很容易受到忽视，因为他自己都不认为需要注意这一点。但费舍尔 1922 年发表了一篇基础性论文，介绍了一种新的理论数理统计。“参数”这个词看似无处不在，但在费舍尔或更早其他人的统计工作中，它几乎是缺席的。费舍尔通过分布族 $f(x,\theta)$——可能的多维参数 θ 的光滑函数——替换了卡尔·皮尔逊的总体分布 $f(x)$，这个分布非常一般，并未特别指定。由此，费舍尔对他研究的估计或检验问题施加了约束和结构，支持了之前不可能的数学分析。回首过往，我们可以把线性模型的早期最小二乘视为参数模型的特例，而费舍尔进行了非常一般化的扩展，并且挖掘了出乎意料的理论成果。

杰日·奈曼和卡尔·皮尔逊的儿子爱冈·S. 皮尔逊在 1928~1933 年发表的文章中，采用了费舍尔的创新，并把它与假设检验的方法相结合。从一开始，设计假设检验就是为了检验模型。其最强的结果就是奈曼–皮尔逊引理,它甚至解答了如何比较两个非嵌套模型——尽管只能在它们完全指定的情况下。这里没有参数需要估计，只有一个直接的问题：数据来自抽样分布 A 还是抽样分布 B？更一般的情形是广义似然比检验，其中检验是明确的残差类型。这个检验可以视为一种竞争：一个参数化的分布族对抗一个更广泛的分布族，后者包括前者。当然，分布族更大缘于它的灵活性更大，拟合的结果会更接近。但是，这种额外的灵活性提高足以证明它的使用合理吗？尤其是，这种提高会大于单个机遇给我们的期待吗？

举个简单例子，可以考虑卡尔·皮尔逊 1900 年研究过的一个问题。为了更好地理解概率，弗兰克·韦尔登付出了非凡的努力。他设计了一个骰子投掷实验，每次投掷 12 个骰子，同时记录每组 12 个骰子中有多少骰子显示 5 或者 6。这个实验总共重复了 26 306 次，骰子投掷的总个数为 $12 \times 26\ 306 = 315\ 672$。皮尔逊的表格在“观测频率”

（Observed Frequency）这一列给出了结果。

为了使你了解这其中包含的劳动，我可以告诉你几年前我课上的一位学生扎克·拉比对这个实验的重复。他发明了一种机械方法做这个实验：先摇动一个有 12 个普通骰子的盒子，骰子落下，计算机随即拍下结果照片机并进行处理，然后整理数据。一次实验大约需要 20 秒。机器日夜运转，大约花费了一周的时间，运行了 26 306 次。想象一下，一个人何以亲手完成这一切。韦尔登在一个报告中表示过，他的妻子帮助了他。人们有些好奇这桩婚姻面临的压力。

韦尔登实验的重点是想考查真实世界与理论世界的接近程度。每次试验有 13 种可能结果：0、1、…、12。如果骰子完全公平，同时每次投掷完全独立，那么理论会认为，每次投掷 12 个骰子的试验中，出现 5 或者 6 的概率是 1/3。所以 12 个骰子当中有 k 个是 5 或者 6 的概率会是二项概率 $\text{Prob}\{\#\text{“5 or 6”} = k\} = \binom{12}{k}\left(\frac{1}{3}\right)^k\left(\frac{2}{3}\right)^{12-k}$。“理论频率”这一列，就是这一系列的每个值乘以 26 306 的结果。

1900 年之前（甚至在 1900 年中），这个超级难题很不容易得到正确处理。数据点位于 13 维空间（图 7-3 中皮尔逊的 m'），当时的统计世界并不熟悉这样的结构。聚集 26 306 次实验的集合距离理论点这么远（皮尔逊的 m），那么应该放弃二项模型吗？而且，如果放弃这个假设，那么替代假设是什么？那个时代更加杰出的分析者（例如弗朗西斯·埃奇沃思和皮尔逊）认为，单独考察一个维度——13 个维度中的每个维度（表中的 13 行）——是不正确的方式。但是，然后该做什么呢？皮尔逊通过卡方检验的离差解决这个问题，这种方法在 1900 年的统计学发展中是相当具有变革性的一步。统计学上第一次出现同时进行多个检验的方法——将 13 个维度纳入一个检验。它不仅考虑了“13 个问题都包括”的事实，还考虑了 13 个维度之间明显的不独立关系。如

果点集远离一个维度的理论，那么它也会远离其他维度的理论。如果根本没有出现 5 或 6，那么其他一些结果必然会频频发生。卡方检验通过一个简单的模型做检验，只需要这 26 306 个试验相互独立，就可以用于其他广泛的解释，包括其他可能的所有二元分布以及非二元分布。

No. of Dice in Cast with 5 or 6 Points.	Observed Frequency, m'.	Theoretical Frequency, m.	Deviation, e.
0	185	203	− 18
1	1149	1217	− 68
2	3265	3345	− 80
3	5475	5576	−101
4	6114	6273	−159
5	5194	5018	+176
6	3067	2927	+140
7	1331	1254	+ 77
8	403	392	+ 11
9	105	87	+ 18
10	14	13	+ 1
11	4	1	+ 3
12	0	0	0
	26306	26306	

图 7-3　韦尔登的骰子数据，由皮尔逊给出（Pearson 1900）

皮尔逊的检验发现，数据与简单模型并不一致，有些东西不对劲。而且数据表明，“5 或 6”发生的频率相比于每 3 个骰子掷一次的情况更高。数据本身给出的 5 或 6 的比例函数为 106 602/315 672 = 0.3377，略大于 1/3。接下来，皮尔逊尝试检验了一般的二项假设，其中 $\text{Prob}\{\#\text{“5 or 6”} = k\} = \binom{12}{k}(\theta)^k(1-\theta)^{12-k}$，但未坚持 $\theta = 1/3$，而使用了 $\theta = 0.3377$ 计算理论值，提升了拟合度（如图 7-4 所示）。

事实上，数据对新计算的列 m' 进行了卡方检验。费舍尔根据自己对参数模型提出的新概念，在 20 世纪 20 年代早期表示，皮尔逊在这里犯了错误。他没有对事实做出考虑，而实际上让数据选择了理论。但这种情况下的这种错误并不严重，并且，皮尔逊的结论受到了费舍

尔对损失“自由度”的修正。

Group.	m'.	m.	e.	e^2/m.
0 …………	185	187	− 2	·021,3904
1 …………	1149	1146	+ 3	·007,8534
2 …………	3265	3215	+ 50	·777,6050
3 …………	5475	5465	+ 10	·018,2983
4 …………	6114	6269	−155	3·991,8645
5 …………	5194	5115	+ 79	1·220,1342
6 …………	3067	3043	+ 24	·189,2869
7 …………	1331	1330	+ 1	·000,7519
8 …………	403	424	− 21	1·040,0948
9 …………	105	96	+ 9	·841,8094
10 …………	14	15	− 1	·666,6667
11 …………	4	1	+ 3	9
12 …………	0	0	0	0

图 7-4　韦尔登的骰子数据，皮尔逊使用 $\theta = 0.3377$ 重新计算了 m 理论值（Pearson 1900）

皮尔逊猜测，$\theta = 0.3377$ 更大的理由是，骰子的每一个点是通过在相应位置上挖去少量材料形成的，那时如此，以后亦是如此。于是点数 5 和 6 的面是 6 个面中最轻的，哪怕数量差得极其微小。这种猜想一直盛行于接下来的这个世纪，似乎说服了所有听过的人。但当拉比以同样种类的骰子重复这个实验时，一些惊人的事情发生了。通过计算机的计数处理，他可以得到每个骰子的 6 种可能结果中的每一个的计数；而不像韦尔登和他的妻子，只注意“5 或 6”或者“不是 5 或 6”。拉比发现（就像皮尔逊），结果并不符合简单的假设。他发现骰子发生“5 或 6”的比例是 0.3343，而且最频繁出现的面确实是 6。但另一个意外出现了：第二频繁出现的面是 1。这就引发了另一种解释：6 和 1 在每个骰子上都是相对的两面。或许骰子不是正立方体，或许这两面比其他相对的面更接近。或许骰子就像厚厚的方块硬币，1 和 6 是硬币的正反面，其他的面都是硬币的边缘。当拉比用游标卡尺测量骰子时，结果非常支持这个猜想。他的骰子上，1–6 这条轴短了大约 0.2%。

皮尔逊的检验为残差方法的实践打开了新的统计前景。经过费舍尔对自由度问题的修正，以及复杂参数模型的多种扩展，现在可以在一个检验中对更复杂的综合性备选假设,并检验一些相当复杂的假设。20 世纪 70 年代，统计学引入了广义线性模型，它吸收了似然的思想，以一种非常灵活的方式扩展至多种形式的参数化计数数据，以及标准的线性模型和方差分析，甚至跨平台的嵌套模型的内部检验使用。

扩展并不局限于参数方程。大卫・考克斯洞察到，我讨论的残差检验必需的全部要素只是附加部分的参数化，而基本模型可以是非参数的。换句话说，比较一个复杂模型和一个简单模型时，当后者是前者的特例，那么“简单模型”可以并不简单，甚至可以相当复杂，也甚至可以并不明确。附加部分是真正需要参数化的，其目的是允许使用强大的参数方法，以对解释能力的增益做出严格检验。考克斯的方法有时称为“部分似然方法”，经过生存数据分析以及医学的其他应用对考克斯回归的使用，它已经产生了巨大影响。

诊断和其他图形

“残差分析”是统计学最常见的术语，出现在模型诊断（“画出残差”）中。先拟合回归方程，再画出“残差”（=观测的因变量减去拟合值），进而对拟合进行评价，以及看看是否存在模式有可能提示下一阶段的建模，这种做法在统计学家中已经很常见。例如，图 7-5 展示了两张残差图，第一张图对应回归 $S = a + bA + cE$，来自加拉帕戈斯群岛的 $n = 23$ 的一组数据，记录了特有物种个数 S、面积 A，以及每个岛最高点海拔 E，目标是了解 A 和 E 如何影响物种多样性 S。这个拟合模型产生 23 个残差，是 $S-\hat{s}$ 的差，其中每一个 $\hat{s}$ 是 a、b、c 为最小二乘估计时的岛屿值 $a+bA+cE$ 。$S-\hat{s}$ 对 $\hat{s}$ 的图表明，$\hat{s}$ 更大，变

异的关系也会随之增强。它建议，对变量进行对数转换，这引出了更合理的模型，$\log S = a + b\log A + c\log E$，这个模型的残差如图 7-5 下图所示。第二个模型描述了一种乘数关系，$S \propto A^b E^c$。这个分析的一个结果是使我们认识到，虽然科隆群岛也许是达尔文进行探索的最佳地点，但对于分离面积和海拔对物种多样性的各自影响而言，它们不大合适。这些岛屿大体由火山岩构成，而 E 大致与 A 的平方根成比例。而且，在实践中，$\log A$ 和 $\log E$ 这两个对数尺度的变量近似于线性关系。为了分离这种效应，我们还需要其他数据。

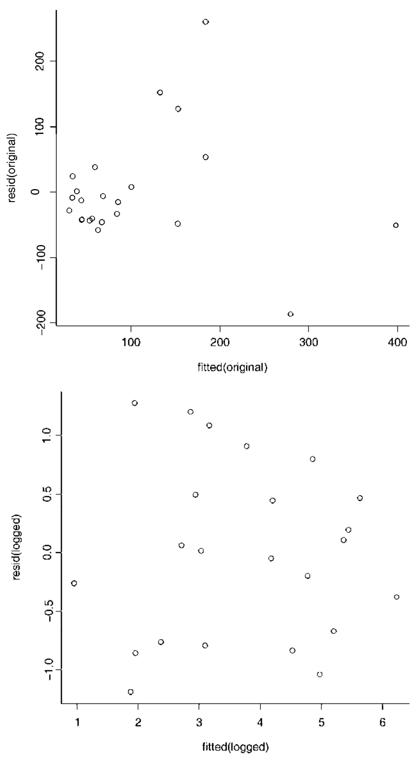


图 7-5　加拉帕戈斯数据的残差图，（上图）原始尺度与（下图）变换为对数尺度

统计图形历史悠久。人们在 18 世纪初就已发现很多有趣的应用，但到 20 世纪，它们才真正繁荣起来。并且，在计算机时代，它们的使用产生了爆炸式发展，甚至偶尔会出现 1000 张图只值一个字的情况。如果我们先不考虑那些只用作装饰的图形（这占了它们今天应用的一个显著比例），那么简单地说，其余所有图形要么是修饰工具，要么是诊断与发现的工具。残差图的类型可以归入后者，但事实上，所有诊断图都使用了残差的扩展定义——都是某种程度的残差图。比如，饼图通过图形对圆的等分基线的偏离，显示各部分不相等的程度。这时，本身低水平的饼图也显现出超越修饰的价值（如图 7-6 所示）。

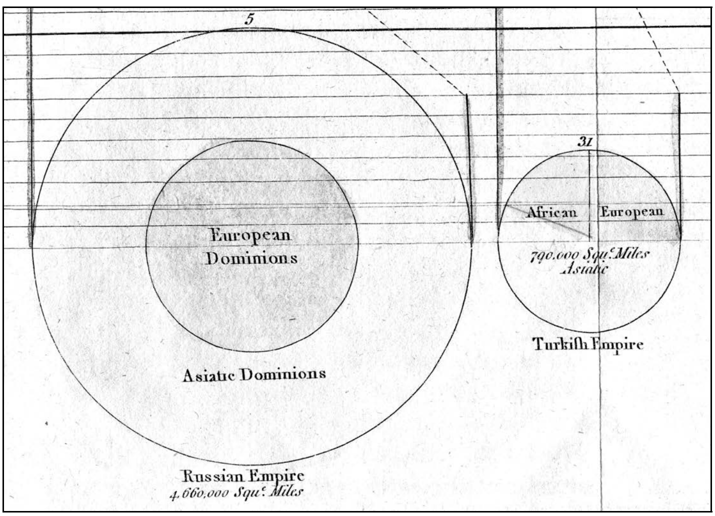


图 7-6　第一张饼图（Playfair 1801）

1852 年，威廉 · 法尔发表了一个类型不同的环形图，这幅图是对 1848~1849 年肆虐英格兰的霍乱所做的部分研究。他希望发现霍乱传播的机制，对每年的数据绕着圆圈画出几个变量。图 7-7 显示了 1849 年的数据。（原始图是彩色印刷的。）外圆显示死亡率，这个圆表示没

有瘟疫流行时，平均年死亡率的周基线——将每周的死亡总人数标记为这一周到中心的距离。从图上可以看到，1849 年 7~9 月，死亡人数相当多。5 月和 11 月的死亡率低于平均值，用更浅的阴影显示，位于外圆之内。1849 年 7~9 月的瘟疫大流行似乎跳离了页面。法尔怀疑，空气传播是致病原因，但这些图并不能提供答案。直到后来根据非图形的理由，他才相信水是这种疾病的传播方式。内圆作为基线显示了平均周气温，这个圆本身提供了年度平均，6~9 月更高的气温用轻微的铅灰色追踪了霍乱流行。总之，疾病和气候之间有着明显的关系，这个图形的两个部分显示了残差现象。

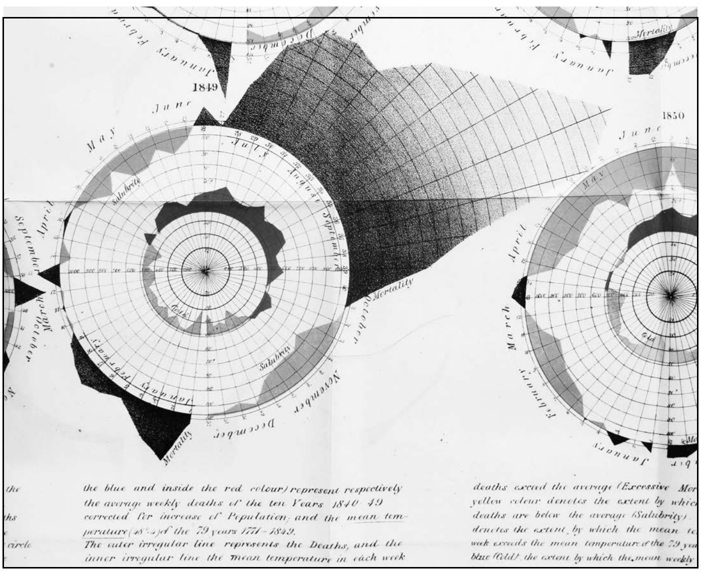


图 7-7　法尔的 1849 年霍乱疫情的环形图（Farr 1852）

法尔的图还有另一个重要结果。弗洛伦斯·南丁格尔采用了法尔的思想，并应用于她对英国野战医院卫生措施进行的改革中，由此产生了巨大的修饰效果（如图 7-8 所示）。

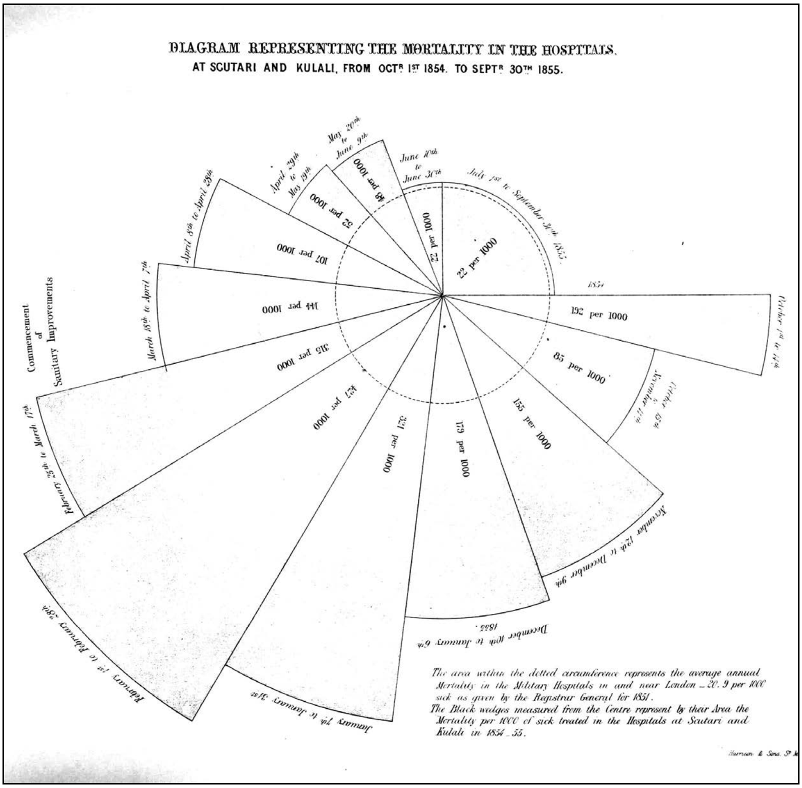


图 7-8　南丁格尔的示意图显示了克里米亚战争中野战医院的死亡人数（Nightingale 1859）

克里米亚战争期间，南丁格尔服务于斯库塔兵营（位于土耳其，靠近克里米亚）的陆军医院。她得知，高死亡率并非由于在战斗中受伤，而是霍乱这样的疾病以及卫生政策的漏洞造成的。相比之下，英国的军队医院更有能力处理这样的问题，并有更好的记录。她返回英格兰，决心为野战医院争取更高的标准。南丁格尔采用了法尔图形的形式，用大的楔形显示了斯库塔和库拉里的死亡率，用虚线环形表示位于伦敦及其附近的军队医院的平均死亡率基线，对两者做对比，基

线比楔形小得惊人。南丁格尔只对法尔的图形做了一处改动，使它产生更震惊的效果。对于死亡数字，法尔用中心距离表示，而南丁格尔用楔形面积表示。换句话说，她把死亡人数的平方根画成中心距离。（在当时，构建这样的图形需要大量劳动。后来，我根据数据重新计算了图形的高度，发现她确实是照自己所说的方式去做的。）

法尔的图会给人欺骗性的印象。在他的系统中，如果死亡率加倍，面积会增大到原来的 4 倍，视觉效果更夸张。南丁格尔确实避免了这一点，给出了一张没有误导性的图像。矛盾的是，法尔心怀探索的希望去绘图，却产生了误导性的图像；而南丁格尔为了修饰来绘图，却并未产生误导。两者都强调了平均值的残差。

结论

这七根支柱是统计智慧的主要支撑，而它们自身并不构成智慧。七根支柱可以至少追溯到 20 世纪上半叶，其中有些还会更古老。经过长期使用，它们已经证明了自己，但还需根据新的需要适应新的应用。它们是统计科学的基础，也是最早的、卓越的数据科学。这七根支柱可以视为统计这门科学的一个智力分类系统，它们可以很好地与其他信息科学合作，比如计算机科学，以及其他有着新名字但尚未获得完整身份的新学科。但是，这些支柱的思想依然相当激进，一旦误用会相当危险，而且闯入陌生领域时会引发对抗性反应。它们当中没有哪个已经过时，但我们依然会问，当代是否需要更多支柱？我们应该开辟第八根支柱吗？如果答案是“是”，原因是什么？接下来，我们使用一种统计方法回答这个问题。下面回顾数据，看看这七根支柱是否透露了答案。

第一根支柱——“聚合”，本质上在讲信息的放弃，这是一种“创造性毁灭”的行为。这个术语是约瑟夫·熊彼特提出的，用于描述一种经济重组的形式，是看待这种行为的一种方式。正如类似的其他应用，“聚合”必须遵守原则进行，放弃不利于（甚至可能减损）最终科

学目标的信息。不过，即便如此，它也会受到指责：隐匿了其他观点下的个别特征。而没有个别特征将如何开发“个人医疗信息系统”？一些统计问题会用到“充分统计量”的概念，汇总了损失无关信息数据的变量。但在大数据时代，这常常不可行，或者背后的假设不稳定。因此，要支持统计智慧，必须平衡这些问题。

第二根支柱——“信息及其测量”，在统计中的含义与在信号处理中的含义不同。它和聚合一起，帮助识别信息中的增益率递减如何同预期的使用相联系，以及这如何帮助设计实验和聚合两者的形式。信号处理中，传递的信息可以无限期保持一个恒定速率；而在统计中，信号的信息累积率则必须递减。看似相等的信息块在统计分析中并不等价，这种现实依然矛盾。

第三根支柱——“似然”，使用概率校准推断，并为度量不确定性提供尺度。危险性很高，价值也很大，要求有极大的细心和理解进行明确运用，但回报也是丰厚的。其中最简单的应用是显著性检验，它的误导性应用已经展现出来，似乎它们成了“诅咒”企业的证据，而不是为了支持特殊用途的证据。过去一个世纪中，显著性检验的使用不断增长，证明了需要为支持或反对某个命题而校准证据。而使用不当时，汇总会产生误导，但这不该使我们更倾向于盲目接受口头总结的误导，因为这种口头总结甚至缺乏对公认标准尝试校准的认同。而似然不仅向我们的结论提供了度量，还对分析、聚合的方法以及信息积累的速率提供了指导。

第四根支柱——“相互比较”，给出内部标准，以及根据手边数据判断内部效应及其显著性的一种方法。它是一把双刃剑，因为不需要外部标准会使结论移除所有相关性。因此，一方面，使用时要小心明智；另一方面，与第六根支柱“设计”一起，在某些高维情形下，可以产生近乎神奇的路径，通向对问题的理解。

第五根支柱——“回归”，极其巧妙。对统计分析而言，它是一种相对性原则。从不同出发点提出问题，这种思想不仅产生了意外的洞察力，也产生了框架性分析的新方法。19 世纪 80 年代，迟到的发现证明了这种精妙。它是使用多元对象的方式，以一种纯正的多元分析，取其中一部分并重新组合。基本形式的逆概率比较古老，但 19 世纪 80 年代前，还没有机制可以用于进行一般性描述推断，特别是贝叶斯推断。早期的尝试可以用滑翔机飞行来比喻。它最多只是在缓慢地下降，但在理想的环境下，它会在某些有限的地带给出飞行的错觉。随着在 19 世纪 80 年代的发展，我们的动力飞行在原则上可以在所有环境中高飞，并且避免了事故或不可能性。对一些早期探险家来说，这已证明为是致命的。经过 20 世纪的再次充分发展，这种理解产生的新方法可以允许高度更高甚至维度更高的旅行，这是更普通的运输方法还未达到的关键。

第六根支柱——“设计”，也极其巧妙。构造结构模型，可以对同时考虑多个因素的高维数据模型进行探索。通过基础推断生成的随机化，可以只依赖于最低程度的建模。

最后一根支柱——“残差”，将复杂模型的比较逻辑作为探索高维数据的途径，图分析中也使用了同样的科学逻辑。现在，这是我们面对的最大需求。数百年之后，我们面临的问题依然存在，最不可能提供广泛的解答。在这里，我们或许看到了对第八根支柱的潜在需求。

现在，数据集越来越大，待解答的问题和焦虑也越来越多，人们担心现代计算中固有的灵活性将会超过答案确定性的校准和判定能力。如果我们可以把注意力限于少数备选的或结构良好的参数化模型，就能舒舒服服地待在家里。但许多情况下，这种舒适正在消失，或者成为泡影。例如，考虑这样三类问题：(1)大数据的预测或者分类器的公式化，数据与许多观测样例有关，每种情况有许多维度的测量；

(2)大型的多重比较问题；(3)关注的问题至少部分是科学研究最后阶段中的探索性分析。

我们面临的第一个问题是任何高维探索都必然面对的。假定现在需要根据 20 个特征构建某种测量的一个响应预测。20 维的空间有多大？如果将每个预测变量的范围划分成四分位数，20 维空间就被划分为 4^{20} 个不同的部分。如果有 10 亿个观测样例，平均每 1000 个部分才会有一个样例。这个经验基础几乎不可能建立置信度！因此，任何合理的分析必须（哪怕只是含蓄的）做出高度限制的假设：或是通过一个低维的参数模型，或至少假设数据接近某个低维子空间。在这样的假设下，机器学习领域的许多优秀算法已经设计出来。通常，优秀的算法在某些情况下受到了成功应用的有限支持，但几乎没有普遍适用的证据。其中一种情况是所谓的“支持向量机”，统计学家格蕾丝·沃赫拜曾表示，可以将其视为近似特定的贝叶斯处理。通过揭示为什么以及何时会这么好，极大地增加其扩展知识，但一般性的问题依然非常困难。

第二类问题多重比较中,我们面临的是需要进行数目庞大的检验。方差分析中，通过用大量成对比较的置信区间，比较许多个因子的效应。基因组学研究中，数千个点位分别放到彼此不独立的假设检验中。概率校准，即置信区间或显著性检验，对只有一对或一个样例的情况是有效的；但在 50 万个样例的更极端的值中选择样例时，它们就没那么有用了。即使在 20 世纪 60 年代，大家就已经知道，W. 图基和亨利·谢弗所设计的过程，通过弱化结果的陈述，对上面所说的选择做了补偿。它和更大的置信区间一样，并非完整的答案。1965 年，大卫·考克斯看到了困难的一部分：对大量陈述同时成立的正确性计算一个概率事实，通常并不能保证这个概率与其中一个陈述的不确定性的度量相关。考克斯注意到（比如根据图基或者谢弗的），整体修正

并不以手边数据的特殊性作为条件，而这或许是因为太保守了。更现代的概念——比如错误发现率——正处在发展阶段，但问题依然困难。

第三类问题的焦点问题出现在分析后期，与第一、二类问题有关，但更加一般化。即使在小数据的问题上，可以使用的解决途径也很多。但途径如此之多——甚至从某种角度看——事实上处于一种大数据的窘境。早在 1885 年，阿尔弗雷德·马歇尔就意识到了这一点，他写道："所有理论家中，最鲁莽和狡诈的是那些自称让事实和数字为他们说话的人。这些人他在幕后扮演着自己的角色，有意无意地对事实和数字进行选择和重组，并提出'后发者因之而发'的观点。"安德鲁·格尔曼从路易斯·博尔赫斯 1941 年的小说标题中借用了一个合适的术语描述这个问题——小径分叉的花园。一个结论经过的曲折、涉及许多选择（关于数据、方向、问题类型）后被合理确定时，没有考虑最终的显著性评价。通常，大数据就是这样的花园。在花园的每个分叉所关注的问题上，我们的校准依然有用，但它们能成功转移到外部的观点吗？

我已经确认了第八根支柱的位置，但没说它是什么。它是一个领域，其中大量过程结合着对某些特定科学问题的部分回答，已经得到发展。这根支柱可能存在，但没有整体结构，还没有获得需要的普遍认同以确立其存在。历史表明，这种普遍认同不会轻易出现，或者不会一步到位。每种有生命力的科学都有它自身的奥秘：天文学是暗能量和暗物质，物理学是弦理论和量子理论，计算机科学是 P-NP 难题，数学是黎曼假设。甚至对于最困难的情形，现存的七根支柱也能支持至少部分的回答。统计是一门活跃的科学，这七根支柱提供了强大的支持。我们进入了一个充满挑战性的时代，与我们相伴的是其他领域的强大盟友以及胜任挑战的强烈期望。

“最难的科学思考通常关注的是一个领域的基础部分，Stigler则为他针对更有影响力的统计学领域的潜在调查带来了光明的信号。”

——Bradley Efron，美国国家科学院院士、斯坦福大学教授

“提取数世纪统计学研究精华并以智慧装点，作者精心准备的七道‘大菜’对那些想要探寻数据的人来说无疑是一次思维盛宴——无论他们想探寻的是大数据还是小数据。”

——孟晓犁，知名统计学家、哈佛大学文理研究生院院长

“统计学具有思考的核心集合，能够覆盖我们生活的方方面面。Stigler对这些思考进行了开发，并让它们焕发了生机。”

——Persi Diaconis，斯坦福大学数学与统计学教授

ISBN 978-7-115-46997-7

图灵社区：iTuring.cn
新浪微博：@图灵教育

分类建议 科普/统计学

人民邮电出版社网址：www.ptpress.com.cn

ISBN 978-7-115-46997-7
定价：39.00元

看完了

如果您对本书内容有疑问，可发邮件至 contact@turingbook.com，会有编辑或作译者协助答疑。也可访问图灵社区，参与本书讨论。

如果是有关电子书的建议或问题，请联系专用客服邮箱: ebook@turingbook.com。

在这可以找到我们:

微博 @图灵教育 : 好书、活动每日播报
微博 @图灵社区 : 电子书和好文章的消息
微博 @图灵新知 : 图灵教育的科普小组
微信 图灵访谈 : ituring_interview，讲述码农精彩人生
微信 图灵教育 : turingbooks