TURING 图灵新知

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

TURING 图灵新知

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（CIP）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
三河市中晟雅豪印务有限公司印刷

◆ 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月河北第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