一个个生活化的故事，读懂日常科学思维
一条条暗藏的脉络，入门AI统计基础

Broadview
www.broadview.com.cn

统计之美

人工智能时代的科学思维

李舰 海恩◎著

中国工信出版集团
電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

统计之美

人工智能时代的科学思维

李舰 海恩◎著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书基于经典统计学的知识体系，结合数据科学的应用经验，使用历史经典故事、网络热点事件、行业真实案例等素材进行介绍，聚焦于科学思维的训练，并对应到具体的理论和技术点，能够帮助读者轻松掌握各种分析方法的背景和思想，并能快速地将相关知识应用到实际的工作中去。

本书深入浅出，所举例子通俗有趣，有助于读者理解人工智能时代的思维模式，应对这迅速变化的世界。

图书在版编目（CIP）数据

统计之美：人工智能时代的科学思维/李舰，海恩著. —北京：电子工业出版社，2019.3
ISBN 978-7-121-35404-5

Ⅰ. ①统… Ⅱ. ①李… ②海… Ⅲ. ①数据处理－通俗读物 Ⅳ. ①TP274-49

中国版本图书馆 CIP 数据核字（2018）第 253094 号

策划编辑：张月萍
责任编辑：刘　舫
印　　刷：三河市鑫金马印装有限公司
装　　订：三河市鑫金马印装有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：720×1000　1/16　　印张：14.25　　字数：312 千字　　彩插：4
版　　次：2019 年 3 月第 1 版
印　　次：2019 年 3 月第 1 次印刷
印　　数：8000 册　　定价：59.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

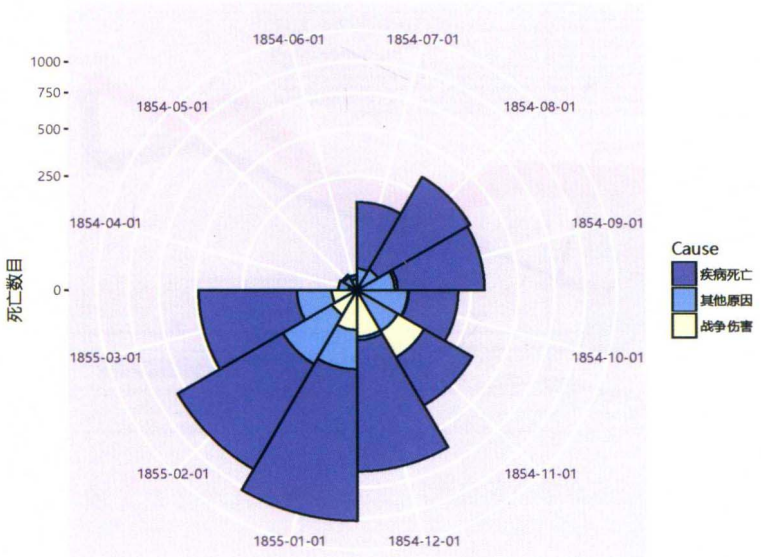

图 3.4 用现代技术绘制的玫瑰花瓣图

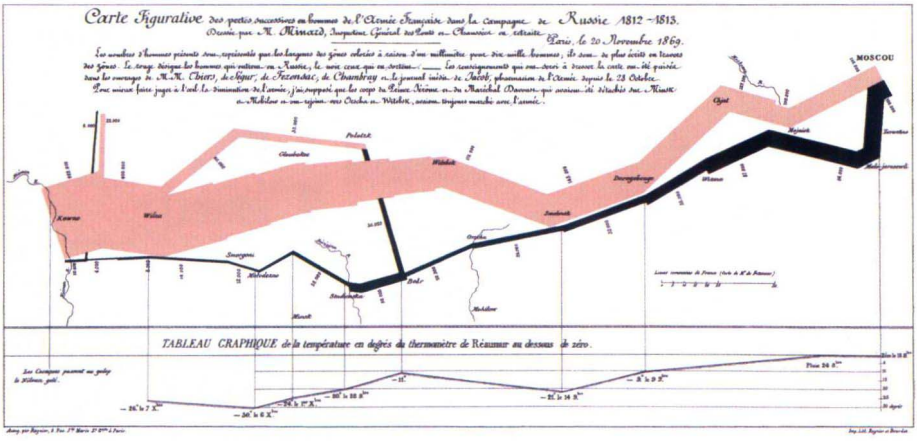

图 3.5 Minard 绘制的拿破仑远征图

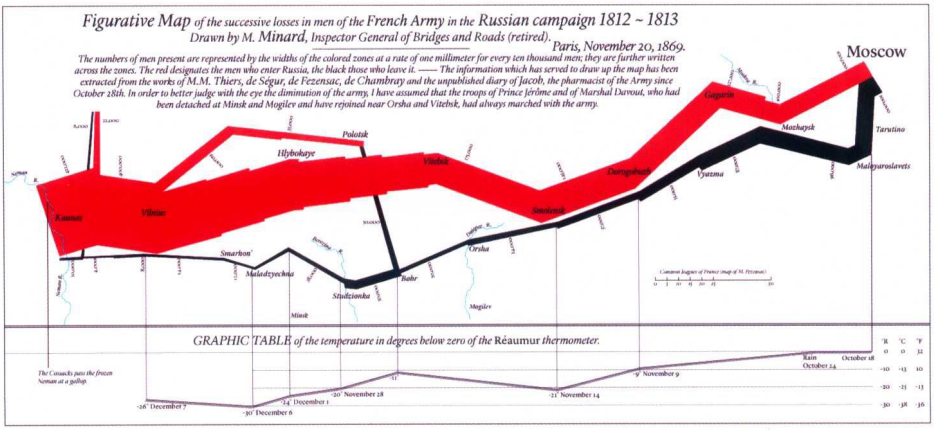

图 3.6 用现代技术重绘的拿破仑远征图

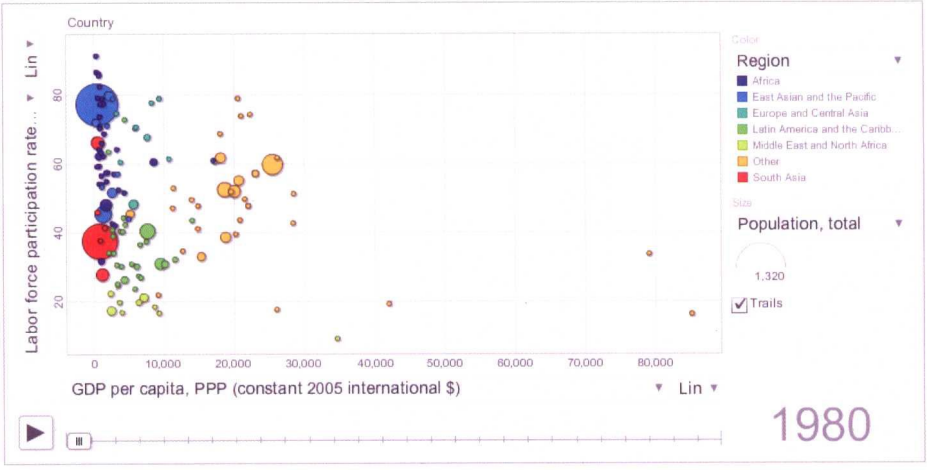

图 3.21 人均 GDP 与劳动参与率的动态气泡图

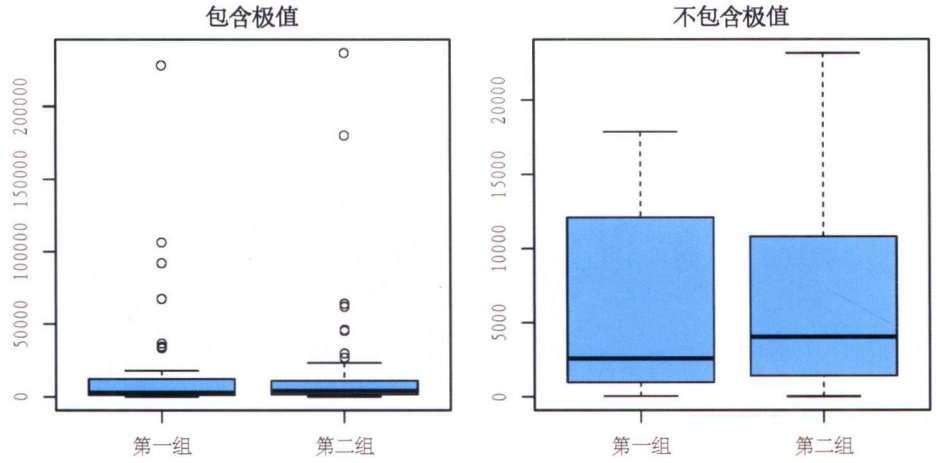

图 3.24 收入数据的箱线图

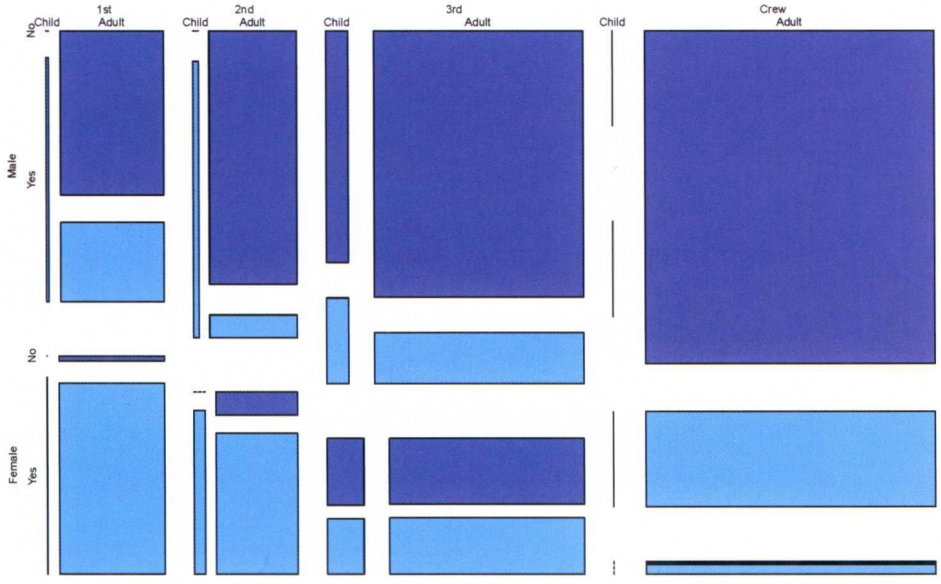

图 3.25 泰坦尼克号的幸存者

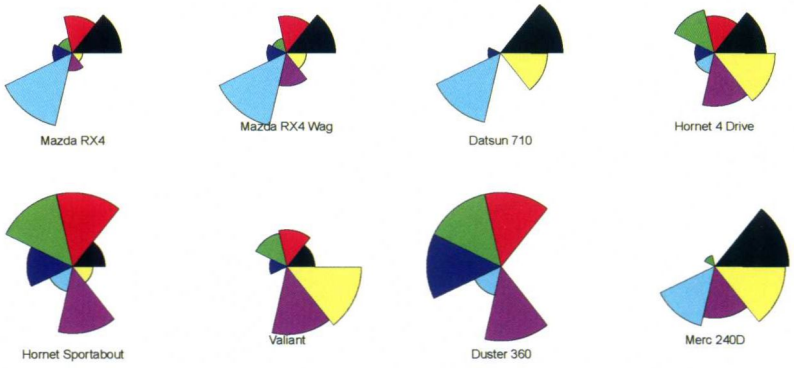

图 3.27 汽车型号的星形图

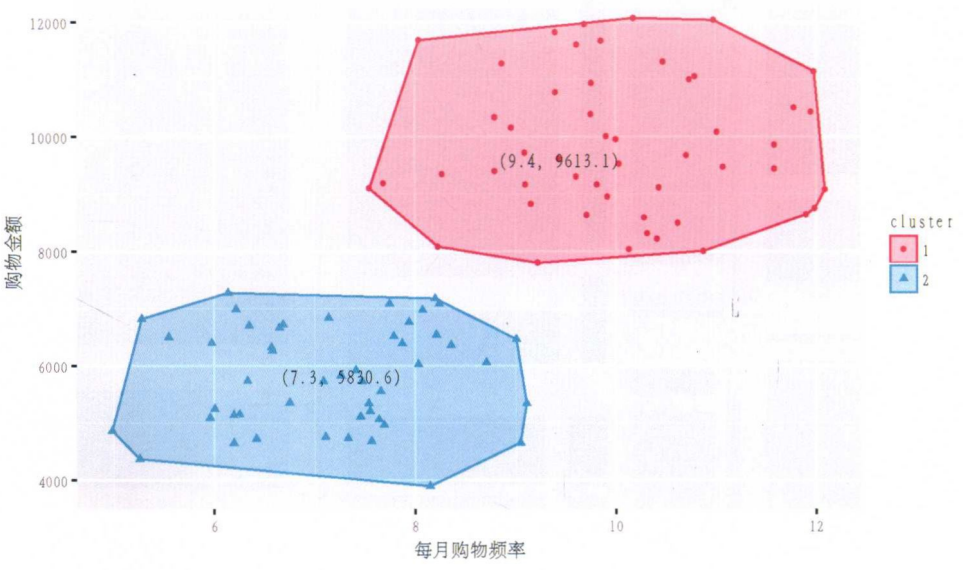

图 4.17 K-Means 聚类结果

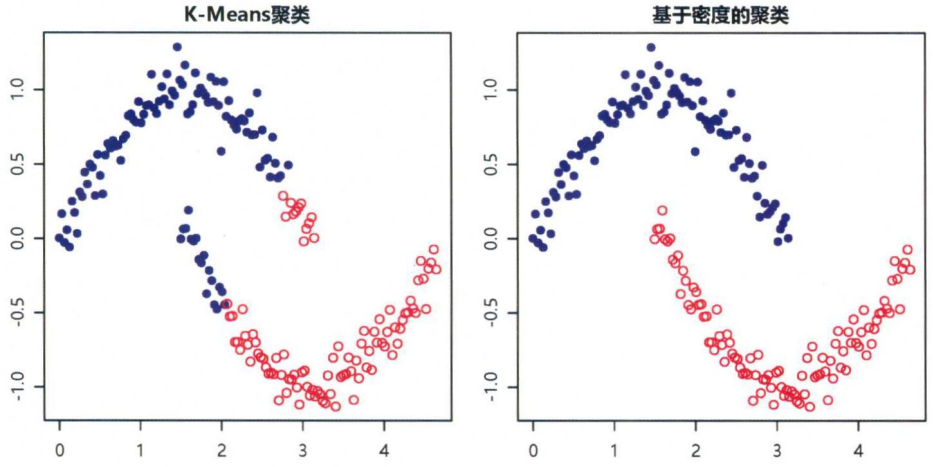

图 4.18 K-Means 和基于密度的聚类

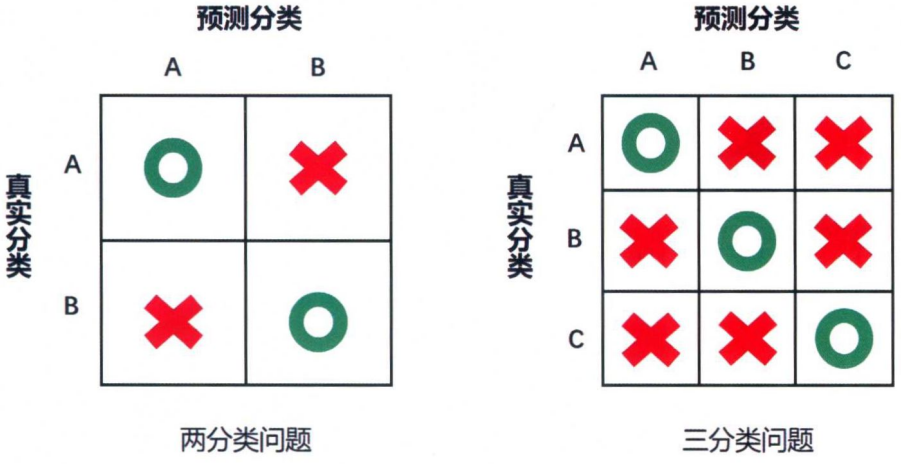

图 4.20 分类的评价

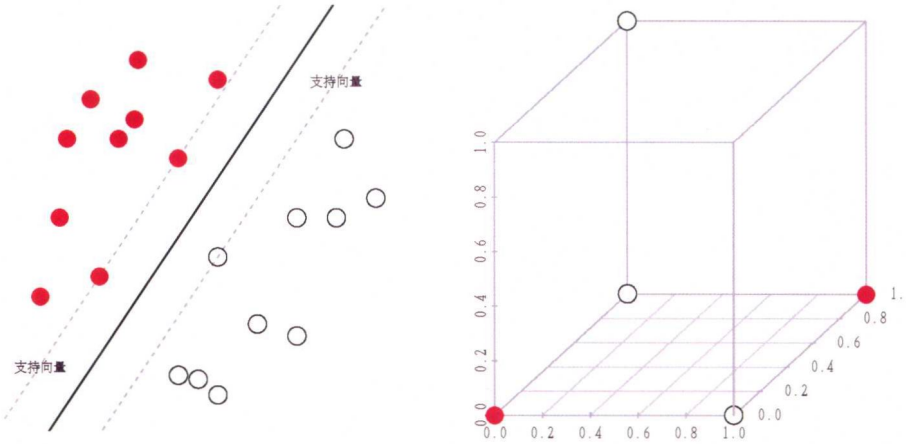

图 4.22 支持向量机简介

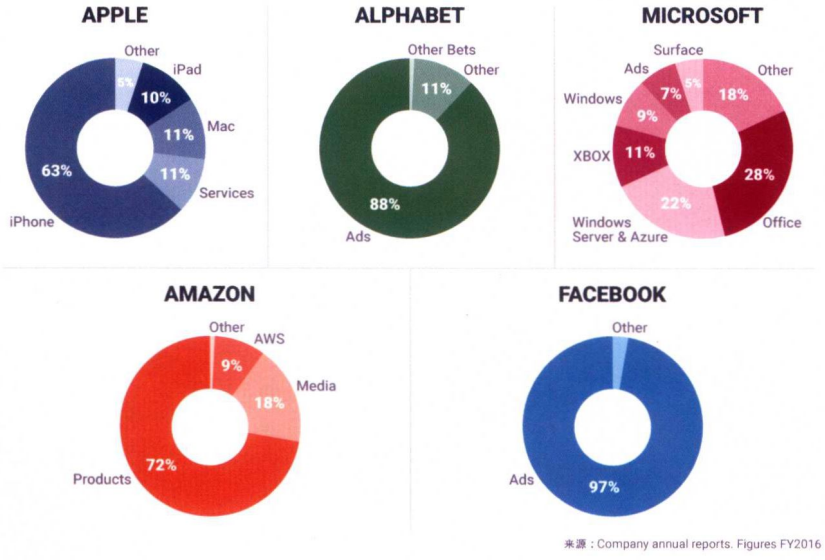

图 5.6 各大互联网公司利润的构成

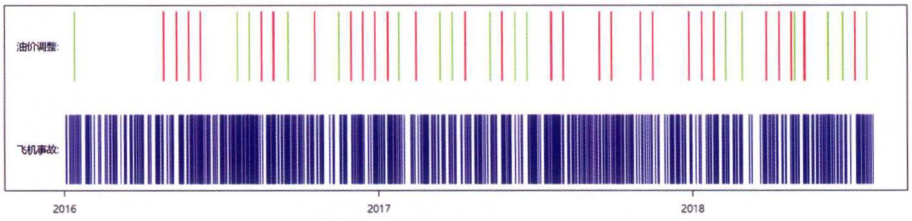

图 6.2 飞行事故和油价

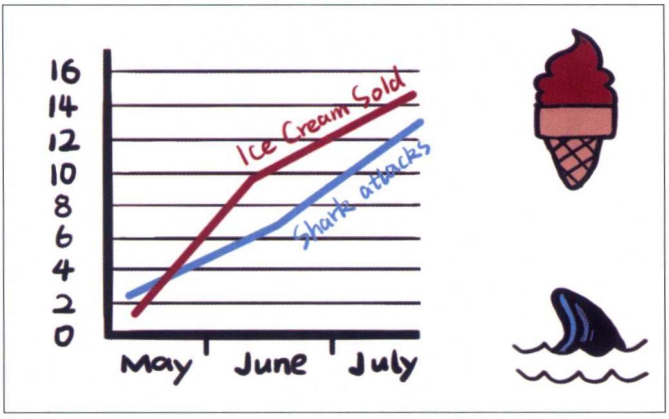

图 6.4 冰淇淋和鲨鱼袭击

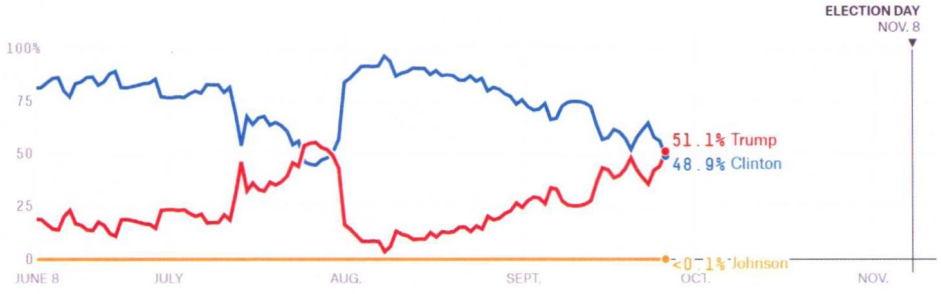

图 6.7 2016 年美国大选预测

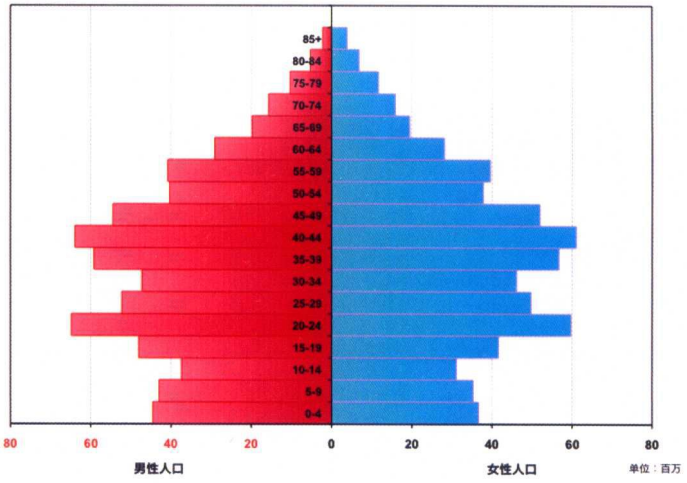

图 6.8 人口男女比例分布

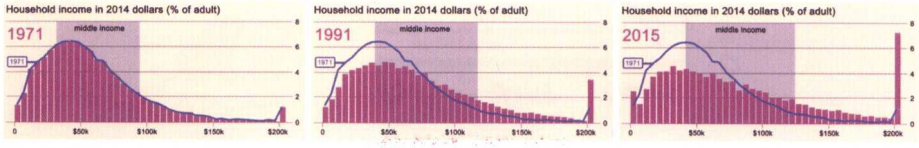

图 6.13 收入分布的变化

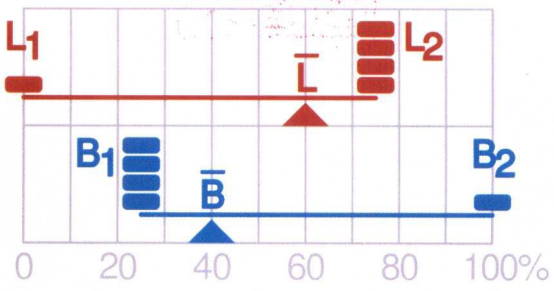

图 6.20 辛普森悖论

序　言

统计学可真是一个尴尬的存在，常常处在各种鄙视链的低端。从数学的角度看，统计学中的数学原理太肤浅，撑死也就一个大数定律，一个中心极限定律，这也能算数学？从应用学科的角度看（例如，计算机、管理学），统计学太数学，一点也不应用。分析数据就好好分析数据，还要整什么大数定律，什么中心极限定律，这也能算应用？作为一名统计学的工作者，对这样的观点虽然并不苟同，但确实很惭愧。常常为此，反省自问：问题到底出在哪里？我辈应该如何作为？

我有一个朴素的信念，任何东西只要是美的，就一定会被大家接受甚至追捧。但是，这里有两个非常具有挑战性的问题。第一、统计学的美到底是什么？第二、她那独特的美如何被大众感知？这是两个非常朴素的问题，作为一名统计学教员，我常常用这两个问题来检讨自己。统计学作为一个历史非常短的学科，在它的发展历史中，有不少杰出的学者做出了卓著的贡献。是他们的卓著努力为统计学建立了扎实的理论基础，为统计学的应用开疆拓土。在这个过程中，产生了很多有用的统计学思想，闪烁着统计学智慧的光芒，解决了太多的实际问题。因此，统计学的美是毋庸置疑的，但为什么大家感受不到？对此，作为一名统计学教员，我没有理由去埋怨大众，而应该做自我批评，自我检讨。如果，我们有能力把统计学中最闪光的智慧，用最朴素而有趣的语言，结合生动而有趣的故事表述出来，那世界又会怎样？如果能够做到，统计学的美就可以被大家感受到。届时，统计学是不是数学重要吗？统计学是不是应用重要吗？统计学就是统计学，她既有理论，又有应用。关键是，她很美，她真的很美，美得令人窒息，美得令人流连忘返，而我们都陶醉于她那独特的美。请问，到哪里去找这样一本书，专攻统计学之美？

要写这样一本书的难度可想而知。首先，你要对统计学的宏观理论框架，从历史到最新前沿，非常熟悉。说来惭愧，我做不到。其次，你要对推动统计学理论发展的重要故事、案例，甚至历史性事件如数家珍。太惭愧，我也做不到。还有，你需要很强的语言文字能力。让文字和数学公式自由穿插，流畅而优美地交织在一起。这对我来说也很难。这样一本书一定是跨学科的。与数据科学相关的领域可不仅仅是统计学，还有计算机科学、经济学、管理学等。不同学科的交叉融合，也极大地促进了统计学的发展。要对这么多学科有所研究，也不是一件简单的事情。

正当我一筹莫展的时候，突然老天眷顾，统计之都大侠舰哥送给我他的新书《统计之美》。首先，我一下子就被目录吸引了。从统计学科学入手，谈到数据与数学，讨论了数据可视化，当然也有模型与方法，还有大数据技术，以及数据的陷阱。每一章的立意都是如此独特，跟任何现有专著或者教材完全不同。这些章节的框架设计恢宏壮美，讨论的问题深刻而朴素，覆盖的内容从过去到未来。这样的框架设计，散发着强烈的舰哥独有的大侠风范。从微观处看，每个章节下面都是一个又一个短小、精炼、经典而深刻的小故事或者案例。这些小故事（或案例）有：上帝掷骰子、女士喝茶、寻找失踪的核潜艇等。每个小故事都突出讲述了一个统计学的智慧，一个知识点。通过这样精炼而经典的小故事，让人们感受到统计学之美，她美在哪里？她美就美在统计学的智慧上，这些智慧变成了统计学思想，统计学思想变成了统计学理论，统计学理论变成了统计学的模型算法。噢，这个路途太长了，难怪当人们看到模型算法的时候，实在是想不起她原来的美了。

不过，别着急，没关系。舰哥的《统计之美》为你揭开这层面纱，让你重新领略统计学的独特之美！为舰哥鼓掌，为《统计之美》点赞，我辈加油！

王汉生

北京大学光华管理学院教授

2018 年 12 月

前　言

英国学者李约瑟研究中国科技史时提出了一个问题："尽管中国古代对人类科技发展做出了很多重要贡献，但为什么科学和工业革命没有在近代的中国发生?"这就是著名的李约瑟难题（Needham's Grand Question）。具体地说，是问"为什么近代科学没有产生在中国，而是在 17 世纪的西方，特别是文艺复兴之后的欧洲?"李约瑟通过对中国科学技术史的研究[1]，在社会制度和地理环境中寻找答案。但这个问题一直被国人拿来反思自己的文化和传统，很多人都分析出了各种原因，大多数人认为中国的传统文化中缺少科学精神、甚至没有能够产生现代科学的基因，再结合现实生活中的各种乱象，无不痛心疾首，都想治病救人。

让我们把时间拉回到百年前的中国，轰轰烈烈的新文化运动已经开始，"德先生"和"赛先生"进了中国。国人深切地认识到了科学的威力，无数仁人志士立志向学，1923 年的"科玄之争"更是加速了科学在全民中的普及。当时"科学派"的观点不仅仅是科学在实业中的价值，更是要全面介入人们的生活。当然，当时的"玄学"也不是指魏晋那套老庄玄学和今天人们认为的旧中国玄学，而是指"在欧洲鬼混了二千多年的无赖鬼"[2]，也就是形而上学。这次科玄之争可以说力度非常大，当时国人对科学的信仰程度超乎今天人们的想象。中华人民共和国成立后，对全民进行科学教育的成就更是有目共睹，中国的科技水平也是发展神速，但是如今国民科学素质的情况似乎仍然不容乐观，很多科普作者越科普越心焦，质疑中国科学精神的言论也仍然甚嚣尘上。

国民的科学素养真的这么差吗？科学素养的缺失真的是传统文化带来的吗？我看都不见得。梁启超在东南大学时，学生罗时实认为国粹将亡，因为读经的人太少了，梁启超闻声大怒，拍案道："从古就是这么少"[3]。当然，科学相比于经学更值得普及，但是对普通民众缺乏专业的科学知识不应苛责，这是正常现象，不同科学领域、不同知识内容的科普是一项漫长而有意义的事业，更需要普及的可能是科学思维。科学思维虽然与任何形式的玄学都水火不容，但也并不等于"死理性派"，也不是"死的机械论"，不能说演绎法是科学而归纳法就不是科学，也不能说理性主义是科学而经验主义就不是科学。不同的历史文化可能侧重不同，我们不能因为中国历史上三百年的特殊时期就质疑整个历史的科技成就，也不能因为中国传统公理体系的缺失就否认整个文化的科学精神，这是不

科学的做法，也属于没有文化自信的表现。

卢瑟福曾说过“如果你的实验需要统计学，那么你应该再做一个更好的实验”，波普尔强烈排斥归纳逻辑[4]并力求以可证伪性为划界的标准，乔姆斯基高举理性主义的大旗并自创“笛卡尔语言学”[5]，这些观点曾经都是主流并且影响了很多人。但是需要指出的是，如今大数据时代下已经充分证明了经验主义、归纳推理的强大之处，即使是如日中天的人工智能实际上也是大数据加上深度学习的归纳方法的成功。我们无意对大师们进行臧否，也不参与具体路线的争论。实际上，无论是倾向于经验主义还是理性主义、归纳主义还是演绎主义，都不会动摇科学的根基。库恩认为，科学很重要的特点在于其独特的范式，在科学领域里大部分时间并没有竞争学派在质问彼此的目的和标准，因此相比其他领域能够取得明显的进步[6]。在不同的领域，大家遵循公认的科学范式进行研究，不管认识论和推理逻辑方面有何不同的倾向，都是科学的。但是由于欧几里得、笛卡儿那一类的完美体系实在太迷人，容易导致很多人忽视了一种重要的科学思维方式，也就是统计思维。

巧合的是，当年科玄论战中“科学派”的主要理论基础就是统计学大宗师卡尔·皮尔逊早期的代表作《科学的规范》[7]。当年的皮尔逊还没有发展出后来的很多统计学经典理论，该书是一本科学哲学著作，坚定地表达了对科学的信仰，他认为科学的领域是无限的，科学方法是通向整个知识区域的唯一门径。但是他也认为无论在哪种情况下科学都不能证明任何固有的必然性，也不能以绝对的确定性证明它必须重复，科学对过去是描述，对未来是信仰。有些精密科学靠明晰的定义和逻辑可以发展，有些问题要靠近似的测量来解决，需要测量理论、误差理论、概率论、统计理论来实现。后来随着统计学的发展成熟，直到今天大数据和人工智能成为显学，都验证了皮尔逊当年的观点。

也许是因为科学这个词听起来太高端，也可能是科学比较接近真理，现在很多科普过于强调精确科学或者“硬”科学，有时候站在了普通人直觉或者经验的对立面，更侧重理性主义和演绎推理。这种精神放在一百年前的蒙昧期是合适的，放在今天全民教育水平不低的情形下可能有些矫枉过正，我觉得还是允执厥中比较好。能够在概念世界和知觉世界[7]中达到和谐、能够在演绎法与归纳法中达到平衡，统计学可能是一个很好的桥梁。如今无论是自然科学还是社会科学都离不开统计学，尤其在应用领域，直接掀起了大数据的热潮，技术层面的威力已经深入人心，但是思维方面的普及还有所不足。实际上，对中国人来说，理解统计思维似乎是一件非常轻松的事，无论是上古伏羲观天法地的归纳精神，或者神农尝百草的试验精神，还是后世天人合一的整体思维、观过知仁的结果导向、未战而庙算的预测习惯，都是深合统计之道的。

很多人受到各种原因的误导之后对中国的文化不自信，易于走向崇洋媚外的极端，这是不对的。即使是作为很多科学基础的数学，也不止一种思维方式。数学家吴文俊院士说

过“我国古代数学并没有发展出一套演绎推理的形式系统，但却另有一套更有生命力的系统”，这个生命力就是“从实际中发现问题，提炼问题，进而分析问题和解决问题”[8]，完全不同于希腊几何学纯逻辑推理的形式主义道路，中国数学的经典著作大都是以问题集的形式出现的，对结果不是用定理来表达的，而是用“术”来表达的，用现代的话来讲就是程序，与近代计算机的使用融合无间[9]。可见中国传统的数学思维是非常适合现在这个算法时代的。算法与统计的结合造就了机器学习、人工智能的大爆发，甚至可以说是主导了这个时代的科技应用方向。统计学家约翰·图基 1962 年的文章[10] 中指出，任何数理统计学工作都应该在纯数学或者数据分析的实践中二选一，两个标准都不符合的工作必然只是一时的过客。陈希孺院士也曾预测“新一轮的突破性进展正在孕育中，它也许就是数据分析？”[11] 如今大师们的论断都已言中，统计学与算法结合解决实际问题，已经渐成主流，甚至发展出了一门新的学科——数据科学。

卡瓦列里原理在西方数学史中被认为是微积分发明前的重要基础，而中国的祖暅原理与之等价[12]。莱布尼茨在提出二进制的那篇著名文章[13] 里直接引用了伏羲八卦，他还认为“如果说我们（欧洲人）在手工技能上与他们（中国）不分上下、在理论科学方面超过他们的话，那么，在实践哲学方面……我不得不汗颜地承认他们远胜于我们”[14]。在这里我们无意比较中西的优劣，也并不是为了说明中国有多厉害（如果是这个目的的话，可以举更多例子或者写另一本书），仅仅只是为了澄清一些误解，这些误解既是对中国传统的某种误读，同时也是科学思维上的某类误区。我们追求理性和完美的体系，也希望能止于至善，但我们也不应忽视经验主义和观察、试验、归纳、计算的力量，这些都是科学，不应偏颇。尤其对于普通人来说，多从观察身边的小事、解决实际问题的角度训练科学思维，可能效果更好，毕竟“刻鹄不成尚类鹜，画虎不成反类狗”。

在如今这个理性与经验、理论与实践、演绎与归纳、公理体系与算法程序和谐统一的大好时代里，我们多了解一些统计学，关注一下数据科学在新时代的发展，类比一下我们祖先的思维方式，是很有必要的。作者不敢妄图进行全面的科普，只能摘录一些平时读书、工作、看新闻时注意到的例子，尝试介绍统计学的发展历程、理论方法和应用实务。受本人的经验和学识所限，很多例子并不是最好的，也肯定存在各种疏漏，但是希望能做一些尝试，和更多的人一起探索统计中的美，分享科学思维中比较人性化的一面。

本书假设读者具有中学的数学基础，如果从书中介绍的概率与随机的角度去理解统计的基本方法，可以作为统计学的入门参考。另外，结合作者的行业经验，比较偏重统计思维方式和大数据应用实务的介绍，如果完全避开书中的所有公式，也不大影响阅读，可以作为这个大数据或者人工智能时代下的统计学科普资料。本书对于基础的数学尽量用最简单的公式来描述，对于更深入的知识提供了参考资料，可以通过正文中类似“[1]”的符号对应到图书最后的“参考文献”中查找。全书中重要的概念和人名也可以到书末

的“索引”中查找相应的页码。

这本书计划了很久，也拖延了很久，感谢本书的策划人成都道然科技有限责任公司的姚新军先生，帮助我们谋篇布局、规划时间以及处理各种杂事。也感谢“统计之都”和“狗熊会”的各位朋友，本书中的很多案例都来自社区中的各种线上线下的交流与讨论。还要感谢我的宝贝女儿从动笔之初就开始的陪伴。当然，最需要的是提前感谢读者的宽宏大量，本人才疏学浅，难免或有所遗漏或偏颇，希望能多多海涵和多多指正。

李舰

2018 年 8 月