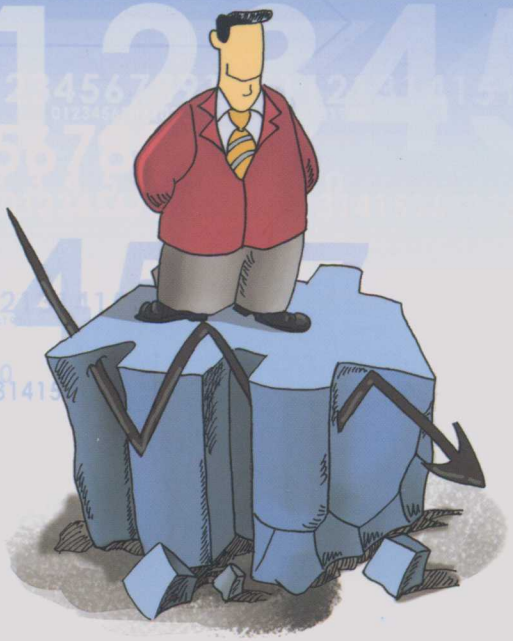
统计通俗读本

漫话信息时代的统计学

——兼话诺贝尔经济学奖与统计学

韦博成 编著



中国统计出版社
China Statistics Press

统 计 通 俗 读 本

漫话信息时代的统计学

——兼话诺贝尔经济学奖与统计学

韦博成 编著

中国统计出版社
China Statistics Press

(京)新登字 041 号

图书在版编目(CIP)数据

漫话信息时代的统计学 / 韦博成编著. —北京 : 中国统计出版社, 2011.10

ISBN 978-7-5037-6383-0

Ⅰ. ①漫… Ⅱ. ①韦… Ⅲ. ①统计学—通俗读物 Ⅳ. ①C8-49

中国版本图书馆 CIP 数据核字(2011)第 201913 号

统计通俗读本:漫话信息时代的统计学

作　　者/韦博成
责任编辑/胡文华
装帧设计/杨　超
出版发行/中国统计出版社
通信地址/北京市西城区月坛南街 57 号　邮政编码/100826
办公地址/北京市丰台区西三环南路甲 6 号
电　　话/邮购(010)63376907　书店(010)68783172
印　　刷/河北天普润印刷厂
经　　销/新华书店
开　　本/710×1000mm　1/18
字　　数/105 千字
印　　张/9.25
印　　数/1—3000 册
版　　别/2011 年 10 月第 1 版
版　　次/2011 年 10 月第 1 次印刷
书　　号/ISBN 978-7-5037-6383-0/C・2577
定　　价/22.00 元

前 言

本书为一本普及型读物，其目的是希望使更多的人了解统计学。在当今信息时代，统计学是非常重要的学科。因为绝大多数信息都是经过量化由数字表达出来，所以数据就是信息的载体。而统计学作为分析数据、从数据中寻找规律性的学科，在当今信息社会就必然会发挥越来越重要的作用。本书第 1 章列举的 3 个事例有力地说明了这一点。因此，向社会公众传播统计学知识是非常必要的，本书的目的就是向大家通俗地介绍和正确地解释统计学的意义与价值。为此，我们安排了两方面的内容：第一，通过众多的实例说明统计学在各个领域的广泛应用；第二，介绍现代统计学的形成与发展概况，并通过其历史发展进程的介绍，进一步阐明统计学的意义与价值。

本书为一本通俗读物，第 2、第 3 章介绍了 20 多个实例，说明统计学在各方面的应用价值。这些实例大体有以下几类：(1). 趣味性比较强，但也有较好的统计意义，诸如：文学著作的统计分析方法；量化投资与统计学；美国总统选举结果预测的抽样调查及反例；二次大战时期统计学的应用实例等。(2). 影响比较大，但也有一定的趣味性，诸如：诺贝尔经济学奖与统计学 (单列为第 3 章)；基因学说的形成、发展与统计学；宇宙起源的大爆炸理论与统计学；六西

格玛 (6σ) 管理与统计学等。(3). 介绍一些很有特色的统计学应用, 诸如: 统计机器翻译方法; 软件工程与统计学; 脑功能成像数据的统计分析; 生物医学中的等价性检验和药物运行的动力学系统等。本书所列举的实例在统计学的广泛应用中虽然只是沧海之一粟, 但是由此也能看到它在国民经济、科学技术以及人类生活等各个方面所起的重要作用。

本书第 4 章介绍现代统计学的形成与发展概况。其中第 4.1–4.4 节简要介绍现代统计学的形成过程及代表人物, 说明统计学到上世纪 50 年代已经发展成为一门成熟的学科。第 4.5–4.8 节介绍近年来统计学的新进展, 说明由于计算机的高速发展, 统计学在一切有数据的地方都有了用武之地, 大大扩展了它的应用范围, 因而统计学在上世纪 70–80 年代就逐步发展成为当今最重要的科学技术之一。最后, 第 4.10 节结合笔者的观点介绍了我国改革开放以前和改革开放以后统计学的发展情况, 以及与国际接轨的问题。说明作为发展中国家 (又深受前苏联的影响), 统计学在我国还远远没有普及。很多人都不了解统计学 (其中也包括不少数学工作者与经济工作者), 他们往往只是把统计学看成从属于数学或经济学的一个小分支, 而没有意识到“在人类活动范围内的一切领域都要求统计学的专业知识和技术”(引自参考文献 [6])。本书的目的就是希望使更多的人对统计学的意义与价值有更全面、更深切的了解, 在我国积极发展科学的统计学, 为早日实现四个现代化服务。

本书是描述型的科普书, 读者只要有大专水平即可阅读其中大部分内容; 若对统计学有所了解, 则更有参考价值。本书也有少数几节用到较多统计学或数学知识, 这几节标上 * 号, 供有兴趣的读者选读。本书亦可作为:(1) 在校大学生和研究生的课外阅读书; (2) 统计学专业或相关专业的“统计学学科概论”课程的教学参考书。

在本书写作过程中, 参考了国内外许多图书资料以及互联网上

的许多材料，受益匪浅，一并对这些作者表示衷心的感谢！同窗好友陈荣昭教授仔细校阅了全书，并且提出许多宝贵意见，也表示衷心的感谢！本书的出版一直得到中国统计出版社的大力支持与帮助，特别要感谢教材编辑部陈悟朝主任的精心策划与安排，他对本书的编辑、审定与出版都倾注了大量心血，特此表示衷心的感谢！

由于作者水平有限，本书难免有不妥之处，恳请同行专家和广大读者提出批评和建议。

作者

2011 年 4 月于东南大学

目　录

1. 引子 — 信息时代的统计学

在当今信息社会, 绝大多数信息都是经过量化由数字表达出来, 所以信息时代就是充满数据的时代, 数据就是信息的载体。而统计学作为分析数据、从数据中寻找规律性的学科, 就必然会发挥越来越重要的作用。现在, 越来越多的理论与应用工作者都意识到: 统计学对于国民经济和科学技术的发展以及人类生活的各个方面都是必不可少的工具与方法。以下首先通过几个事例予以说明, 更详细的介绍可参见后面各章。

1.1 当今研究生的首选: 统计学 — 2009 年 8 月 5 日《纽约时报》

美国《纽约时报》于 2009 年 8 月 5 日曾经刊登一篇文章: "当今研究生的首选: 统计学" (原文为: "For Today's Graduate, Just One Word: Statistics", 见 [1])。该文介绍了 Google、IBM 等大公司争相聘

用统计学家的情况, 同时还报道了 Google 首席经济学家 Hal Varian 的观点: "在下一个 10 年, 统计学将是最有吸引力的工作, 刚毕业的统计学博士, 其年薪可达 12.5 万美元"。这说明, 统计学研究生的前途是非常看好的。该文还指出: "数据就是新知识的素材 (data is merely the raw material of knowledge)", 并且引用了美国麻省理工学院数据产业中心 (Center for Digital Business.) 主任 Erik Brynjolfsson 的观点: "我们正在快速地进入任何事情都可以用数字来度量和操控的时代, 这是对人类的巨大挑战, 我们必须有能力去利用, 分析和解释这些数据"。对于这一挑战, 首当其冲的就是以数据分析为己任的统计学家, 当然这也是他们难得的机遇。因此, 《纽约时报》的这篇文章充分说明, 美国媒体十分看重统计学对于未来科学技术的发展所起的重要作用。这也应该引起我国有关人士的高度重视。

1.2 全球九大开拓性新兴科技领域之一 —— 贝叶斯 (Bayes) 统计技术

我国《科技日报》于 2004 年 2 月 12 日曾经刊登一篇文章: "全球九大新兴科技展望" (见 [2], 亦可见: 央视国际 – 科技频道, 2004 年 2 月 13 日)。该文报道: 美国《技术评论》杂志根据 2003 年的调查, 介绍了全球九大开拓性新兴科技领域, 其中第 4 项为贝叶斯 (Bayes) 统计技术 (其它为: 个人基因学; 合成生物学; 微射流光纤; 纳米导线; T 射线等)。以下为美国《技术评论》杂志对贝叶斯统计技术的部分评价:

- 贝叶斯统计作为数学的一个古老分支正在焕发青春, 将是下一波软件开发的基本工具。
- 它可能使外语翻译、微型芯片制造、药物发现、基因管控问

题、新的生物医学技术等领域发生巨大进步。

• 英特尔、微软、IBM、Google 等大公司都已挤入这一新领域的研发。微软 2003 年版 Outlook 就包括了 Bayes 办公室助手软件; 英特尔开发了基于 Bayes 技术的程序, 可解释半导体晶片的质量测试数据; Google 正在应用 Bayes 技术寻找网上大量互相关联的数字化图形, 并予以开发利用。

把统计学的一个分支 (即贝叶斯统计) 作为全球九大开拓性新兴科技之一, 虽然这只是美国《技术评论》的一家之言, 但是也在一定程度上说明了统计学对于未来科技发展所起的重要作用。

1.3 1991−2001 年期间数学论文引用率 —— 统计学家遥遥领先

据多家刊物报道 (如见 [3]), 根据编制科学引文索引 (SCI) 和相关文献的美国科学信息研究所 (ISI) 的统计, 在 1991−2001 这十年期间, 全世界数学论文引用率最高的前25名数学家中, 有18名是统计学家 (其中华人统计学家范剑青名列第 6, 孟晓犁名列第 20); 引用率最高的前10名数学家中, 有8名是统计学家。这说明, 统计学家论文的引用率远远超过其他方向数学家论文的引用率。另外, 引用次数最高的前 15 本数学杂志中, 有 5 本是统计学杂志, 其中 JASA (即 Journal of the American Statistical Association) 的引用次数遥遥领先于其他数学杂志。虽然引用率只是评价学术成就的指标之一, 而且有一定的局限性 (见《数学文化》2010 年第 1 期, p.85−91), 但是正如文 [3] 所述: "统计文献相对于整体数学的高引用率是与它广泛的科学影响分不开的"。统计学毕竟仅仅是数学几十个分支中的一个, 其遥遥领先的引用率在一定程度上说明了当今统计学的重要

性以及它在数学学科中的特殊性 (事实上, 统计学已逐步发展成为一门独立的学科, 见 [3] 及本书 4.7–4.8 节)。

以上三个事例也许出乎不少人士的意料之外, 但是它恰恰向人们展示了统计学强大的生命力和影响力。说明统计学是当今最有吸引力、最有发展前途的学科之一 (本书 §4.7 将有进一步的阐述)。我们应该充分重视统计学的发展, 切不可错过时机。

本书的目的就是希望使更多的人对统计学的意义与价值有更全面、更深切的了解, 从而在我国更多地应用统计学; 更快地发展统计学。除本章引言外, 第 2 章通过十几个实际案例说明统计学在各个领域的广泛应用; 特别, 作为对于经济学的应用, 第 3 章专门介绍统计学对于经济学的发展以及对于获得诺贝尔经济学奖的意义与作用, 并系统介绍经济学家在计量经济学和金融经济学方向的获奖情况。第 4 章简要介绍统计学在国内外的发展概况, 并通过其历史发展进程的介绍, 进一步阐明统计学的意义与价值; 也列举更多的论据说明, 近代统计学是当今信息时代最重要的科学技术之一。同时结合笔者的观点, 介绍改革开放以前和改革开放以后我国统计学的发展情况, 以及与国际接轨的问题。

2. 什么是统计学 — 实际应用案例

什么是统计学? 根据《大英百科全书》的定义: 统计学是“用以收集数据、分析数据和由数据得出结论的一组概念、原则和方法”。更确切地说, 统计学是“研究如何获取数据、如何分析数据、如何解释数据, 从数据中提取信息, 寻找规律性的学科”。即统计学的研究对象是数据; 研究任务是分析数据, 提取信息 (见 [4]–[6]), 因此, 有数据的地方就需要统计学。而在当今信息社会, 数据无处不在, 因而统计学也就无处不在。另一方面, 无论在科学探索的哪个领域, 一个好的研究都包括以下要素: 创新的想法或问题 (及其理论推导); 合理的实验设计和实验方法; 以及有效的数据处理。这最后一个要素主要就是基于现代计算机的统计学方法。因此任何学科都离不开统计学。正如著名统计学家劳 (C. R. Rao) 在他的名著《统计与真理》(见 [6]) 所指出的: “人类活动范围内的一切领域都要求统计学的专业知识和技术”; “统计思维总有一天会像读与写一样成为一个有效率公民的必备能力”。现在, 统计学已经成为当今英、

美等发达国家最热门的职业之一 (详见第 4 章)。

统计学是数学的一个分支, 它置根于概率论与数学, 同时也受到现代化的计算机科学的强烈影响 ([3])。其处理数据的特点是通过对局部样本进行统计推断, 从而了解总体的规律性。但是, 以数据为基础的统计学与其他数学分支很不一样, 现在更接近于一门独立的学科 (详见 [3] 及本书 4.7–4.8 节)。

以下通过十几个实际案例说明统计学在各个领域的广泛应用, 其中有些案例在统计书刊中鲜有介绍, 诸如: 量化投资与统计学、统计机器翻译方法、净室软件工程与统计学、脑功能成像数据的统计分析等, 希望引起读者的兴趣。有些案例的分析比通常的报道要全面详细, 以加强说服力, 如孟德尔豌豆杂交实验的统计分析等。虽然本章的实例只是统计学广泛应用中的沧海之一粟, 但是在一定程度上也能看出它的意义与价值。

2.1 文学著作的统计分析方法

在统计学应用的诸多领域中, 文学著作的统计分析是一个饶有兴趣的分支, 国内外在这方面都有卓有成效的工作 (见 [6]–[13] 及其引用的文献)。本节介绍国内外学者应用统计学方法研究莎士比亚著作 (和一些相关著作) 、前苏联名著《静静的顿河》以及红学 (《红楼梦》研究) 的一些情况。

2.1.1 莎士比亚新诗鉴定 —— 一曲统计学的赞歌

美国斯坦福大学统计学家埃弗龙 (B. Efron) 教授和他的学生曾经对莎士比亚的著作进行过相当深入的统计分析 (见 [8], [9])。另一著名统计学家劳 (Rao) 在文献 [6] 中对 Efron 的工作曾经做过生

动的介绍。

1985 年 11 月 14 日, 学者泰勒 (G. Taylor) 在专门保存莎士比亚著作的保代林 (Bodelian) 图书馆中, 发现了写在纸片上的一首从未见过的新诗, 该诗仅有 9 节 429 字, 但是无年代, 无作者可考。问题是: 该无名诗是否为莎士比亚所作? Efron 和他的学生 R. Thisted 在他们以前研究的基础上, 对此无名诗的用词风格进行了深入的分析研究, 迅速回答了这一难题, 并发表在著名的统计学杂志Biometrika上 (见 [9])。

根据他们的统计, 莎士比亚全部著作的用词总数为 884647 个, 其中不同的词数为 31534 个。而在这些不同的词中, 仅用一次就弃之不用的词高达 14376 个 (占 45.8%); 仅用二次的词有 4343 个 (占 13.8%)。这说明, 莎士比亚善用新词。他们比较了莎士比亚著作中不同单词使用的频数分布以及刚发现的无名新诗中不同单词使用的频数分布 (见 [6], [9]), 通过深入的统计分析, 并应用非参数经验贝叶