Wisdom In Statistics

统计中的智慧

朱 莹/著

复旦大学出版社

统计中的智慧

朱　莹／著

复旦大学出版社

图书在版编目(CIP)数据

统计中的智慧/朱莹著. —上海:复旦大学出版社,2014.1
ISBN 978-7-309-10130-0

Ⅰ.统… Ⅱ.朱… Ⅲ.统计学 Ⅳ.C8

中国版本图书馆 CIP 数据核字(2013)第 243189 号

统计中的智慧
朱 莹 著
责任编辑/范仁梅

复旦大学出版社有限公司出版发行
上海市国权路 579 号 邮编:200433
网址:fupnet@fudanpress.com http://www.fudanpress.com
门市零售:86-21-65642857 团体订购:86-21-65118853
外埠邮购:86-21-65109143
江苏省句容市排印厂

开本 890×1240 1/32 印张 6.5 字数 155 千
2014 年 1 月第 1 版第 1 次印刷

ISBN 978-7-309-10130-0/C·275
定价:20.00 元

内容提要

全书分为基础篇、趣味篇、应用篇3部分，共25个专题.基础篇主要介绍数据中的代表数、大数规律、正态分布、假设检验、抽样调查等数理统计学的基础知识和思想方法.趣味篇精选扑克牌魔术揭秘、色盲遗传问题、假账识别秘诀、投票悖论、文学名著作者的鉴定等统计案例，在展示统计学魅力的同时，激发读者学习统计知识、思考统计问题的兴趣.应用篇通过密码破译、美国总统选举预测、高考按“总分”录取是否科学、癌症检查准确度探秘、聚类分析方法、用统计观点看世界等专题介绍，让读者领略统计学应用的广泛性和解决问题的独特思想，引领读者学习运用统计方法分析工作和生活中遇到的问题，提升自己的生存智慧和竞争能力.

与常见的数理统计学书籍不同，本书取材新颖有趣，内容深入浅出，特别着力于统计学思想和其独特思维方式的阐述.全书尽量避免艰深的理论演绎和公式推导，力求熔知识性、趣味性、科学性于一炉.

本书可用作大学生的通识教育读物和中学数学教师的进修用书，也可供其他社会人士阅读.

前　言

大千世界，数据无所不在. 例如，各种食品的价格，班级学生的考试成绩，公司职员的工资，股票指数的升降，100 台电视机的使用寿命……呈现在我们面前的往往是一大堆数据. 但是，仅凭这一堆数据并不能直接得出结论，需要经过进一步分析，才能从中提取有用的信息. 统计学正是对数据进行收集、整理和分析的一门学问.

虽然以收集和整理数据为特征的描述性统计学早已存在，可是，在概率论基础上，以统计推断为主要内容的现代意义的数理统计学，直到 20 世纪才告成熟. 本书所讨论的统计学即指数理统计学. 随着统计学的发展，统计应用的领域日益广泛，目前统计方法已深入到自然科学、社会科学和日常生活的各个方面. 作为一个现代公民，经常会在工作和生活中受到大量数据的困扰，只有具有一定的统计学知识，才能洞察隐藏在纷繁的数据后面的规律. 尤其是统计学独特的思想方法闪烁着智慧的光芒：在传统数学无能为力的领域，它常常能提供大胆、新奇、有趣的结论；在面对有争议的判断时，它能帮助我们作出接受或拒绝的明智选择. 如同向阳的树木亦有其阴暗的一面，统计也常常被误用和滥用，甚至被一些人作为设计骗局的工具. 人们在接触统计结果时，应该多一些批判的眼光和具备一定的识别能力. 从正反两个方面看来，学一点统计知识都很有必要. 在美国，数理统计已成为所有的大学一年级学生的一门必修课.

全书分为基础篇、趣味篇、应用篇3部分.基础篇主要介绍数据中的代表数,方差,大数规律,正态分布,假设检验和随机抽样等数理统计学的基础知识和思想方法,是阅读全书的知识基础.在趣味篇将会读到:利用数字的统计规律可以有效识别造假的账目,大家熟悉的投票选举会出现悖论,在解决此悖论的过程中竟促成多位诺贝尔奖得主产生,多年未了的文学公案却可以用统计方法一锤定音;还有小世界现象,魔术揭秘,色盲遗传问题,传递性的丢失,揭露伪造证词案,以及电视收视率造假等也都是饶有趣味的统计案例,在展示统计学风采的同时,能激发读者学习统计知识、思考统计问题的欲望.应用篇则通过密码破译,美国总统选举预测,高考按"总分"录取是否科学,癌症检查准确度探秘,利用回归模型进行预测,聚类分析方法,用统计的观点看世界等专题介绍,让读者领略统计应用的广泛性和用统计分析问题、解决问题的思路,引领读者运用统计学知识和思想方法分析工作和生活中遇到的问题,提升自己的生存智慧和竞争能力.

与常见的数理统计学书籍不同的是,本书通过有趣的统计案例介绍,着力于统计思想和其独特思维方式的阐述,文字深入浅出,尽量避免艰深的理论演绎和公式推导.全书力求融知识性、趣味性、科学性于一体.

本书可作为大学生的通识教育读物和中学数学教师的进修用书,也可供其他社会人士阅读.

在写作过程中,作者参考了大量的书籍和报刊杂志(书后附有主要的参考文献),谨向这些著作的作者和译者表示真诚的感谢.

限于水平,书中一定会有不足或错误之处,恳请读者朋友批评指正.您如果有什么意见或建议,欢迎通过电子信箱 yingyzhu@gmail.com 与我联系.

朱　莹

目　录

基础篇

统计学是关于收集和分析数据的科学和艺术.

英国《不列颠百科全书》

1 统计的起源

统计离不开数字，统计的历史要从数字的起源说起. 早在远古时期，我们的祖先就用在绳子上打结的办法记录一笔笔数字和一件件事，这就是史书上记载的“结绳记事”的故事.

中国最早的统计大约起源于公元前 21 世纪的夏朝. 春秋时期的著名政治家管仲(前 725—前 645)认为：“不明于计数，而欲举大事，犹无舟楫而欲经于水险也.”杰出的历史学家司马迁(约前 145 或前 135—?)约于公元前 91 年著的《史记》中编制的 10 个历史年表，是中国的第一批统计表. 古埃及在公元前 30 世纪已有人口和财产的统计数字. 古希腊公元前 6 世纪已进行过人口普查. 古罗马公元前 4 世纪就建立了人口出生、死亡登记制度. 随着社会生产的不断发展，人们计算的范围由人口、土地逐渐发展到社会经济生活的各个方面. 只不过人们尚未使用“统计”这个术语.

统计作为专用名词来源于英语的“Statistics”，是从表示国家的“state”一词演化而来. 其意思是，由国家来收集、处理和使用数据. 统计，在语言学中就是数据的意思. 人们将进行各种统计、计算、科学研究和技术设计时依据的数值称为数据.

早期的统计学称为政治算术. 威廉·配第(William Petty，1623—1687)是政治算术的主要创始人，英国古典政治经济学的奠基人，在科学史上具有重要地位. 马克思(Karl Marx，1818—1883)称他为“政治经济学之父，在某种程度上也可以说是统计学的创始人.”恩格斯(Friedrich Engels，1820—1895)在《反杜林论》

中则说:“配第创造‘政治算术’,即一般所说的统计.”如同牛顿(Isaac Newton, 1642—1727)一样,配第也是贫苦家庭出生,后来因其学术上的成就被册封为爵士,成为贵族.他还是英国皇家学会的创始人之一,并于1673年被选为副会长.

配第的《政治算术》成书于1671—1676年,但是在他去世后的1690年才在伦敦出版.当时的英国在政治上陷入非常不利的境地,荷兰、法国都对英国构成强大的威胁,与荷兰的几次战争使英国元气大伤,加上瘟疫流行,使英国人对国家的前途十分担心,整个英伦三岛笼罩着一种悲观的气氛.配第写《政治算术》的目的,就是为了给新兴的英国资产阶级鼓气加油,努力消除悲观失望.他通过对各种统计资料的比较分析,论证英国完全可以超过荷兰和法国在全世界称霸,建立英国殖民帝国.

配第的政治算术具有两个主要特点:

(1) 一切论说都通过数量来表达,全面排斥形而上学的和思辨的空洞议论.

(2) 只重视诉诸感觉的论证,只对那些在自然中有可见的根据的原因进行考察.

在17世纪,配第就坚持认为,社会科学必须像物理学一样定量化,这的确难能可贵.他把“政治算术”定义为“利用数字处理与政治相关的问题的推理艺术”.他认为,所有的政治、经济问题都是政治算术的分支.配第作为统计学的先驱,在统计学的研究对象及研究方法等方面都颇有贡献.尽管他当时没有采用统计学之名,但是已有统计学之实.

在英国兴起的政治算术学派很快跨越英吉利海峡,传播到欧洲大陆.于是政治算术也在整个欧洲得到蓬勃发展,并很快形成人口统计派和经济统计派两大支派.

在政治算术发展的过程中,最值得称道的是两性平衡规律的发现.该规律在文化史上常常被称为“神意”或“神定规则”.1741

年，一位德国牧师出版了《神定秩序》一书，指出男女出生率基本平衡：在100个婴儿中，男婴为51，女婴为49. 由于男性一生中遇到的危险要大于女性，因此男婴多于女婴. 男女出生率所以会基本平衡，他认为是上帝为了在人世间实行一夫一妻制而特意作出的安排，所以叫做“神意”.

拉普拉斯(Pierre Simon de Laplace，1749—1827)在《关于概率的哲学探讨》一书中，则给出了一个关于“神意”的经典例子. 他根据很多地方的统计资料得出结论：男婴出生数与全体出生数的比值几乎完全相同，在10年间的这些比值都保持在$\frac{22}{43}=51.16\%$左右. 但是，巴黎在40年间(1745—1784)的资料却显示，这个比值是$\frac{25}{49}=51.02\%$. 他无法解释为什么会出现这种显著的差异. 后来经过调查，终于发现在巴黎某些地方当时有丢弃男婴的习俗，因而使比值发生偏差. 经过修正以后，这个比值也稳定地接近51.16%.

大约在18世纪80年代以后，英国才逐渐用“统计学”的名称代替了政治算术. 1885年6月24日国际统计学会正式诞生.

19世纪以前的统计学主要是根据较少的抽样统计资料对总体进行推算. 但是由于政治算术学派使用的统计方法很不完善，计算结果常常具有很大误差. 如果不能很好解决误差问题，就不能准确地说明问题. 这就形成了促使统计学向更高阶段发展的客观需要.

概率论被引进统计学是统计学发展史上的一个重要里程碑. 关于概率论的研究在16世纪就已开始，最初是为了解决赌博输赢问题才得以发展. 17世纪、18世纪的许多数学家都对概率问题进行了研究，尤其是瑞士的贝努里家族作出了很大贡献. 19世纪中叶，概率论已发展成为数学上的一个重要分支.

将概率论正式引入统计学的功劳当属比利时人凯特勒

(Adolphe Quetelet, 1796—1874). 凯特勒是一位知识渊博、多才多艺之人. 他不仅是数学家、物理学家、天文学家、统计学家,而且是诗人、歌剧作家,还是一位社会活动家. 他长期担任比利时统计委员会主席,并且主持国际统计会议. 他一生著作很多,其中与统计学有关的就有 60 多种. 凯特勒特别强调数学对于社会科学的重要性,认为:"要更多地促进科学发展,必须使之更多地进入数学领域,这是一种必然的趋势. 我们判断一门科学的精密程度,就是看它利用数学的多少."他提出正态分布可用于各种科学,而正态分布规律只有利用概率论才能给予确切的说明. 他还指出任何现象都有误差,并且符合误差规律;任何现象通过大量观察都可以发现规律,并且符合大数规律. 他将概率论、误差理论和大数规律看作统计学的理论基础,并通过自己提出的方法对数量变化的规律进行研究. 特别值得指出的是凯特勒将统计方法拓展成为既可适用社会现象研究、又可适用自然现象研究的一种通用方法. 自此,统计学就不再是单纯的社会科学.

由于凯特勒在统计学上的杰出贡献,引进概率论的统计学发生了质的飞跃,从此走上了近代科学的道路. 凯特勒也因此作为近代统计学的奠基人而被载入统计学史. 由于数理统计学是在概率论和统计学的基础上发展起来的,因此人们又把凯特勒视为数理统计学的奠基人. 有趣的是,凯特勒却从未将他的统计学称为数理统计学.

数理统计学这个名称最早是在韦特斯坦(T. Wittstein)1867 年发表的论文"关于数理统计学及其在政治经济学和保险学中的应用"中出现的. 论文的用意是将概率论应用于经济学和保险学. 令人意想不到的是,数理统计学这个术语从此被广泛使用,并发展成自成体系的数理统计学学科和数理统计学派.

近代数理统计的创始人是皮尔逊(Karl Pearson, 1857—1936),他原先是数学物理学家,曾任英国伦敦大学学院的应用数

学力学教授，后来从事生物统计学研究. 1901 年，皮尔逊创办《生物计量学》(*Biometrika*)杂志，使数理统计有了自己的阵地，并形成一个数理统计学派.

现代数理统计的奠基人是英国的费歇尔(Ronald Aylmer Fisher，1890—1962). 他的理论工作主要有：研究如何测量数据中的信息，如何缩减数据而不损失信息，以及如何估计模型中的参数等. 尤其是关于实验设计的研究，能大大节省人力和物力，使工作效率提高很多倍. 费歇尔的工作领域主要在农业、水利、遗传等方面，获得的成果具有极大的实用价值，并产生了巨大的影响.

统计学是收集数据、分析数据的科学. 哪里有试验，哪里有数据，哪里就会有统计学. 人口统计、生物统计、医学统计、天文统计、气象统计、水文统计、工程统计、金融统计、教育统计，各种专业统计的不断涌现足以说明统计应用之广泛. 在大学里，统计系已纷纷从数学系里独立出来. 作为一个现代社会的公民，不懂一些统计恐怕已难以应付每天扑面而来的大量的数据和信息.

2 数据中的代表数

我们每天都会遇到各种数据：班上每个学生的考试成绩，单位里每个员工的工资，各种股票的价格，100 台电脑的使用寿命……面临一串大小不一的数字，人们希望能用一个数字来代表这一串数字，概括地反映其一般水平. 这是人类追求简单的天性使然，也是一种科学的思维方式. 以下介绍几种常用的代表数.

2.1 平均数

例如，我们接触的中国人和日本人有各种不同的身高，有的日本人比中国人高，有的中国人比日本人高，很难判断哪一个国家的人高. 可是，如果我们算出中国人的平均身高和日本人的平均身高，经过比较就可以知道，中国人比日本人要高一些.

一组数据的**算术平均数**（简称**平均数**）就是用这组数据的总数除以这组数据的个数所得的商.

设有数据 x_1，x_2，x_3，…，x_n，则这组数据的平均数为

$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n}. \tag{2.1}$$

平均数在计算过程中用到了一组数据中的每一个数，反映并且标志这组数据的集中趋势和“水平”. 平均数在日常生活、工农业生产和社会经济活动中使用非常频繁. 例如，上海市无人售票的公共汽车实行单一票价：每人一次乘车 2 元. 一天中有很多人乘公共

汽车，有人只乘两三站，有人要乘八九站. 对公共汽车公司来说，一天运营的收入与原先实行多级票价比较，既没有减少，也不会增加. 这是因为每人 2 元的单一票价是根据人们乘车的平均数确定的. 可是，汽车公司却因为省去售票员而减少一大笔开支. 对乘客而言，只要你经常外出乘车，在一个月或一年内也不会觉得吃亏或占便宜，而且乘车更加方便.

在上述算术平均数的计算过程中，各个数据出现的次数都是一次，所占的分量是相等的. 可是，实际情况有时并非如此，若要区分不同数据的不同分量，就需要使用**加权平均数**.

表 2.1　一组数据

数据(x_i)	出现次数(f_i)
x_1	f_1
x_2	f_2
x_3	f_3
⋮	⋮
x_n	f_n

设有一组数据如表 2.1 所示，则该组数据的加权平均数为

$$\bar{x}=\frac{x_1f_1+x_2f_2+\cdots+x_nf_n}{f_1+f_2+\cdots+f_n}=\frac{\sum\limits_{i=1}^{n}x_if_i}{\sum\limits_{i=1}^{n}f_i}. \tag{2.2}$$

例 1　某研究团队共有 60 岁的研究人员 1 人，55 岁的研究人员 4 人，46 岁的研究人员 10 人，32 岁的研究人员 5 人. 求该研究团队研究人员的平均年龄.

解　因为团队中不同年龄的人所占的分量不相等，所以应用加权平均数计算：

$$\bar{x}=\frac{60\times1+55\times4+46\times10+32\times5}{1+4+10+5}=45.$$

所以该研究团队的平均年龄为 45 岁.

以上计算式中的频数(即某年龄的人数)，在计算过程中对平均数的大小起着一种权衡轻重的作用，频数大的数值对平均数的