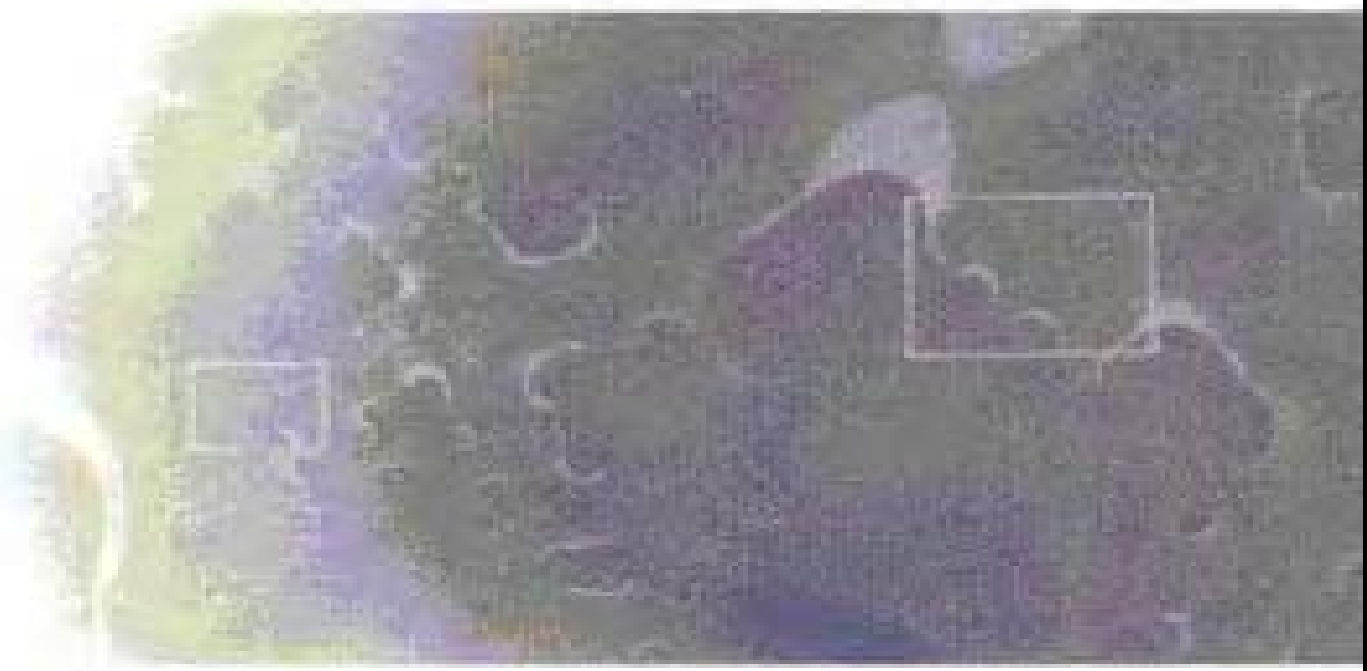
上海市教育委员会组编
张莫宙

daxuesheng renwen suyang jiangzuo

大学生人文素养讲座

shuju yu ren sheng

数据与人生



49

上海交通大学出版社

责任编辑／柳义生　封面设计／顾　风

ISBN 7-313-02415-0
9 787313 024152 >

ISBN7－313－02415－0/G·342

全套定价：72.50 元

本册定价：5.50 元

355 O1-49
Z31

上海市教育委员会组编

大学生人文素养讲座

数 据 与 人 生

张奠宙

上海交通大学出版社

内 容 提 要

"数字地球"的口号风靡世界。人类面临着处处充满着"数字"的环境。如何处理汹涌而来的"数字"大潮,是新世纪摆在每个中国公民面前的课题。我国王选院士实现了中文排版印刷技术上的革命,这归功于数学上的数据压缩技术。美国在数字电视上击败日本,是数据处理上的成功范例。本书还将揭露一些"数据欺诈"的事例,包括广告、中奖、伪科学等诸多社会生活中的数据问题。世界已经充满数据,人生必须处理数据。

图书在版编目(CIP)数据

数据与人生/张莫宙. —上海:上海交通大学出版社,2000
(大学生人文讲座/王铁仙主编,叶敦平、陈卫平副主编)
ISBN 7-313-02415-0

Ⅰ.数… Ⅱ.张… Ⅲ.数据-人生-青年读物 Ⅳ.G.416

中国版本图书馆 CIP 数据核字(2000)第 15837 号

大学生人文素养讲座
数据与人生
张莫宙
上海交通大学出版社出版发行
(上海市番禺路 877 号 邮政编码 200030)
电话:64071208 出版人:张天蔚
常熟市文化印刷厂印刷 全国新华书店经销
开本:787mm×960mm 1/32 总印张:46.5 总字数:912 千字
2000 年 10 月第 1 版 2000 年 10 月第 1 次印刷
印数:1—3050
ISBN 7-313-02415-0/G·342 全套定价:72.50 元
本册定价:5.50 元

版权所有 侵权必究

前　言

自从有了人类，便有了“数字”，因为猎物是需要计数的。不过，直到20世纪前半叶，人们应付日常生活，还只需要扩大了的算术——知道数的加减乘除，以及一些几何图形知识就够了。数据，还只是数学家和科学家关注的对象。

1945年出现的电脑一步一步地在改变世界。身份证是一串数据，超市里的商品标是“条形码”的数据，人事处电脑里储存着“个人的一列数据”。20世纪90年代，美国的数字电视打败了日本的模拟电视，中国的中文排版技术占领了世界的汉字印刷业大半江山。1998年，美国副总统戈尔在“信息高速公路”的基础上，又提出“数字地球”的概念，整个地球都要数字化了。人们终于知道，社会人生离不开数据，任何人必须处理数据，不得不和数据打一辈子的交道。

在数据背后的是数学。统计学是处理数据的，数理统计学更揭示出数据后面的数学规律。与20世纪50年代相比，中学生已经减少了对平面几何的学习，数学课却增加了“统计”内容。当电视上播出“去掉一个最高分，去掉一个最低分”的时候，数据处理就在我们的身边了。

用数据说话往往是最有说服力的。可是，数据也可以造假。原苏联的李森科，就用假数据来制造

伪科学。广告上的数据欺诈不可不防，但是合理的数据表示又是时代的要求。

本书是为“非数学专业”的大学生们写的，里面没有数学公式，也没有大段的推理。它只向读者表明，世界在数字化，人生将面对数据。“预则立”，让我们更多地关注那些“枯燥的数据”，把它们处理得富有人情味，具有社会感，在新世纪中面对从未有过的“数据人生”。

1 数字化与数学观

数字化生存，预示着数字化的未来。当世间万物都数字化的时候，没有数学的生活是不可想象的。信息时代要求人们掌握基本的数学知识，具备良好的数学能力，更为重要的是树立正确的数学观。数学不再被人们只当做“升学”的敲门砖。数学将浸透在人们的日常生活之中，成为谋生立业的基础。数学，将是为发展生产和兑现经济效益的技术。数学技术体现了新的数学文化现象。

1.1 数学技术：从几则新闻说起

1978年，徐迟的报告文学《哥德巴赫猜想》，风靡大江南北。新闻媒介使陈景润成为中国青少年的科学偶像。数学，在人们的心目中几乎等于陈景润和他的哥德巴赫猜想研究。陈景润的科学业绩已经载入中国的史册，无需赘言。20年过去了，数学正在发生悄悄的变化。除了华罗庚、陈景润等的数论研究之外，数学新闻正向数学技术倾斜。请看以下的新闻和事例。

(1) 1998年9月7日《文汇报》报道：8月16日，沙市水位从44.88米涨到晚上11时45分的

45.08 米。估计到凌晨 5 时，水位可能涨到 45.20 米。只等中央一声令下，早已准备好的 20 吨炸药将立即在荆江大堤分洪。次日凌晨 5 时 30 分，中央命令的传真文件下达，要求继续严防死守。作出这一重大决策的根据有：

> “由多方专家组成的水利专家组用‘有限元素法’对荆江大堤的体积渗透进行了测算，确定出一个安全系数。照这一系数推定，即使沙市水位涨到 45.30 米，也可以坚持对大堤严防死守，不用分洪。”

这“有限元素法”便是一种求解偏微分方程的数学方法。我国已故数学家冯康在 1965 年曾做出前驱性的贡献。

(2) 1998 年底，中央电视台、香港凤凰卫视先后播送王选院士的访问记。方正集团，中国高科技产业的“大哥大”，已成为改革开放时代的成功代表。但是，这项告别“铅与火”的印刷业革命，其核心技术之一是数学技术。方正创始人王选，1962 年毕业于北京大学数学系，从事汉字排版技术研究。20 世纪 70 年代，直接采取“数字化”方案，运用自己发明的“汉字”信息压缩技术，取得了世界领先的水平。台湾地区《中央日报》也买方正的技术。

技术的核心是数字信息的压缩。一个汉字包含许多笔画，形状各异，如何使用较少的信息加以刻划，是一个核心问题，当然也是核心商业机密。

(3) 1996 年 8 月，英国的格拉斯哥地区法庭审理一桩工业诉讼案。原告是一所剧院，起诉某建筑设计院的设计不好，影响票房收人。而证人是一台电脑。法庭诉讼中，将剧院设计的各种设计参数输人电脑，并用直观的动画模拟剧院内空气流动的情况。结果是剧院业主胜诉。电脑中存储的空气动力学软件其实是一组数学方程式。此举开创了电脑模拟作为法庭证据的先例。

(4) 美国对冲基金 1998 年损失了 100 亿美元，原因是运用的金融数学技术出现了失误，一个小概率事件导致了错误的决策，损失惨重。

(5) 周光召著文谈中国科学的发展，提到美国现在掌握着未来电视的技术，日本的电视王国地位受到动摇。原因是美国掌握了数字化技术，其核心是“小波”数学技术。

此外，密码编制和破译，CT 扫描技术，海湾战争模拟技术，喷气机和航天器的控制技术，说到底，其核心都包含着一种特殊的数学技术。

因此，人们对数学的认识发生了重大变化。如果说，陈景润在解决哥德巴赫猜想上的贡献是激励人们攻克数学难题，那么王选的贡献表明数学是可以成为直接创造经济效益的技术。数学技术可以成为企业技术的核心。

那么，请读者审视一下我们流行的世界观，回答一个基本的问题：什么是数学？前苏联的加里宁有一句名言：“数学是思想的体操”，数学通过逻辑证明、严密推理培养人们的良好思维习惯。这句名言尤其成为数学教育的指针。极端的说法是把数学等

同于逻辑。另一种说法是，“数学是科学的语言”。数学是物理学、工程学、天文学等等学科的工具。科学规律都是用数学公式表示的。因此，不懂数学也就无法掌握其他科学知识。这当然是完全正确的论断。但是，20 世纪下半叶数学的进展，特别是电脑技术的推动，数学已经从科学的后台走到幕前，成为一门能够直接产生经济效益的技术，前面的例子说明了这一点。

大百科全书的“数学”条目是这样写的：“数学是反映现实世界空间形式和数量关系的科学”。它仍然是数学的最好概括。不过，我们的理解应当加深。数学从反映现实到服务于现实，从“思想体操”和“科学工具”，进一步提出“数学技术”，乃是数学观的一项革命性进步。

1.2 数字化：电视大战的历史经验

《纽约时报》畅销书，尼葛洛庞帝的《数字化生存》，于 1997 年由海南出版社出版。一年内三次印刷，一时供不应求。封面上写着：“计算不再只和计算机有关，它决定着我们的生存。”

数字化将是未来社会的主要特征。世间万物都会数字化。地球、气象、资源、海洋甚至交通指挥，都会数字化。未来卫星的分辨率将是一米。当人们用卫星导航驾驶汽车的时候，数字信号的传送是我们生存空间的灵魂。一个明显的事实是：一切信息都将以数字的方式而存在，数字的计算和信息交换将密不可分，那么，人们离开数学怎么行？

现在让我们回顾在 20 世纪 90 年代发生的电视大战。这场媒介革命的历史经验证明了数字意识和数学技术是何等重要。

电视是 20 世纪最重要的发明，它对社会发展的影响怎样估计都不会过高。那么未来的电视技术将是什么样子呢？

日本是世界电视工业的领头羊。索尼、松下等公司一向领导电视技术的新潮流。早在 1972 年，几位富于前瞻的日本人对自己提出问题，下一步的电视往何处走？他们的答案是："用高分辨率提高清晰度。"确实，高清晰度电视是每个人都会提出的要求。

大家知道，现在的电视使用的是模拟技术。先用摄像机将画面的光学特性转变为电的特性（光电效应），即将图像上明亮和色彩不同的光点，逐点、逐行、逐帧地转变为一串电信号。这一转变是模拟性质的，例如光的亮度和电信号的强度一致。然后电视台将此电信号发射出去，电视机再将它们逐点、逐行、逐帧地同步恢复原来各个光点，重构原来的画面。中国现行标准是每秒传送 25 幅画面，每个画面上有 400 条水平线，每条线上有 400 个点，屏幕宽高比是 4:3。显然，每幅画面分解的行数和点数越多，图像就会越清晰。用通俗的语言来说，"电视屏幕上的点子越细越好。"计算机屏幕上有 1150 条线，当然就比电视画面要清晰得多。由于光信号转变为电信号的方法多种多样，于是有美国的 NTSC 制，法国的 SECAM 制，德国和亚洲的 PAL 制，互不统一。

日本人的超前意识，意味着更多的帧频（每秒 60 幅），更高的行数（每帧 1125 条或 1250 条），每行

有 800 个点，以及更合理的屏幕宽高比 16:9，更科学的电子技术设置等等。他们花了 12 年的时间，提出了各种制式，希望能够得到国际的公认，而且最好是全世界统一的制式。然后按新制式生产各种新的电视摄像机、发射机、接收机，从而获得新一轮电视贸易的主动权。但是，这是一个落后的技术意识，数字化电视把日本的“高清晰度”模拟技术彻底打垮了。

大约在 1981 年，欧洲警觉到日本可能独霸电视市场，也研究起“高清晰度”电视。从贸易保护主义的立场出发，1986 年欧洲共同体立法通过了自己的“高清晰度”模拟技术。20 世纪 80 年代以前的美国，几乎没有电视工业，所有的电视设备全由国外进口，特别是日本。在日本和欧洲电视战的初期，美国的一些研究机构支持日本的方案。后来，美国又反对日本和欧洲方案。可笑的是，一些厂商拾人牙慧，所提出的方案仍然是模拟技术。1987 年，美国通讯委员会(FCC)决定开发新一代电视，争夺市场。当时美国有四家公司，另有日本的方案作为第五家共同竞争。

1991 年，美国通用仪器公司(General Instrument Corporation)和麻省理工学院联合宣布实行电视的全数字化技术方案。几乎在一夜之间，美国所有研究电视技术的计划全部改弦易辙，抛弃模拟思想的陈旧温床，站到“数字化”的大旗之下。有充分的证据显示，数字信号的处理有更好的成本效益。美国的四家公司都提出全数字化方案，日本的模拟方案明显落后，不得不于 1993 年退出竞争。

1991 年 9 月，《数字化生存》作者尼葛洛庞帝在一次午餐会上向法国总统密特朗和内阁成员建议：放弃他们自称的“领先地位”——电视模拟技术。当时法国没有接受。英国的撒切尔夫人听从了尼葛洛庞帝的建议。1992 年末，英国的梅杰首相否决了给“高清晰度”电视计划补贴 8 亿美元的提案。到了 1993 年，欧洲共同体终于承认了“数字化”方案的优越性，宣布放弃模拟技术的发展计划。

1992 年，尼葛洛庞帝向日本首相宫泽喜一说明“高清晰度电视没有前途”，这使宫泽喜一大吃一惊。1994 年，日本邮政省放送行政局长讲山晃正提议日本跨入“数字世界”时，遭到日本产业领袖的围攻。日本人在“高清晰度”电视上投入的钱实在太多了，不肯轻言放弃是可以理解的。但是，当美国和欧洲相继放弃模拟技术之后，世界市场的大部分已经在“数字化”方案的控制之下，日本在“高清晰度”电视模拟技术上的近 20 年努力，终于宣告失败。

1.3 烽火台：“比特”的胜利

数字信息的单位是“比特”，上节所述的电视大战以数字化方案获胜而告终，因此被称为“比特的胜利”。

那么，什么是“数字化”呢？它何以有如此大的威力？

以电话为例，过去的老式电话是将声波模拟成电波传送，现在的数字式电话是把声波变换成一组数字。接受者将这串数字还原为声音，传送即完毕。

再如，如果要将你的基本情况告诉别人，只要把你的身份证号码传过去就行，对方按这串数字就可以读出你的住址、性别和出生年月。传一串数字比传一组汉字要方便多了。

数字有0，1，2，3，4，5，6，7，8，9共十个。从技术上说，只传送两个数字0、1就够了。这是因为有以下的对应关系：

0 ⟶ 000
1 ⟶ 001
2 ⟶ 010
3 ⟶ 011
4 ⟶ 100
5 ⟶ 101
6 ⟶ 110
7 ⟶ 111
8 ⟶ 1000
9 ⟶ 1001

因此，所谓数字化，实际上是把信息化为一串由0、1组成的数串。身份证不过15个数字，转换成0、1数串就要几十个数字。如果要传声音、图像等信息量很大的信息，那就需要传送几万几亿以至更高数量级的0、1数串了。

那么，用什么方法来衡量信息量的大小呢？这便是“比特”(Bit)。这要从中国古老的烽火台说起。

我国古代传递军事信息，烽火台十分有用。平日边关太平无事，烽火台便没有动作。一旦敌人来犯，烽火台上便燃起烽烟，瞭望哨发现烽烟，立即通报大本营决策迎敌。这是最原始的通讯，只有“无

烟”、“有烟”两种情况(分别用0、1表示)。因为这种0、1通讯最原始、最基本、最简单,我们很自然地将它定为信息量的单位,即定义为一个信息量。信息论的奠基人申农(Shannon)将这样的单位信息量称为1比特。

定义:在通讯过程中,如果传送的信息只有0、1两种情况,则称该信息的信息量为1比特。$\log_2 2 = 1$。一般地,如果信息有N种情况,则定义它的信息量为$\log_2 N$。以2为底的对数,在这里很自然地用上了。

例如,在烽火台通讯中,如果某前沿阵地有甲、乙两座烽火台。甲台表示有无敌人来犯,乙台表示是否需要补给。于是有以下四种情况:

甲台	乙台	
0	0	(没有敌人,不必补给)
0	1	(没有敌人,需要补给)
1	0	(敌人进犯,不必补给)
1	1	(敌人进犯,需要补给)

这时的信息量显然应该定为2比特:

$$\log_2 4 = 2\log_2 2 = 2(\text{比特})。$$

比特的定义,使原来无法确切衡量的信息得到数量表示。如果说烽火台所传送的早期信息不过是几个比特,那么传送一封信,或一篇文章,那就需要几千甚至几万比特。过去使用电报传送信息,用滴滴嗒嗒的短声长声(分别是0和1),可以表示数字、英文字母以及汉字(四个数字表示一个汉字)。到了无线电通讯和广播时代,信息量又成倍增加。后来,发明了电视,一个画面的数字化信息量相当于100

万比特。现在市面上通用的 5 英寸光盘可存储 100 亿比特。以美国为代表的工业化国家正在建立“信息高速公路”，就是要在全国畅通无阻地传送这类极大容量的信息。

1.4 现在的计算机运算速度仍然是太慢了

电子计算机的运算速度，以令人眩晕的方式在发展。从每秒运算几千次，到几万次，乃至上亿次，都曾经令人激动万分，把它看做是人类智慧的伟大胜利。

计算速度是障碍人类科学前进步伐的主要原因之一。许多科学问题，有了可靠的理论，也有实行的社会需要，就是因为计算速度的制约，理论只能束之高阁。

气象预报是一个典型的例子。大气环流的动力学理论，包括它的数学方程，在 19 世纪已经完成了。无线电通讯技术使得气象数据的收集和传输成为可能。但是，用这些数据，求解这些方程，至少需要几天甚至数月的时间。天有不测风云，气象预报必须在几小时之内发出才有实际价值。明日的天气要几天之后才能算出来，成了事后诸葛亮，还有何用？1950 年 4 月，电子计算机设计方案的创始人冯·诺依曼，和气象学家合作，运用早期的计算机成功地进行了世界上第一次“数值天气预报”，成为计算机解决科学问题的一个里程碑。

现在的计算机每秒运算 1 亿次，已经不是新闻。但是，从数学计算的要求来说，这个速度还是太慢

了。现用著名的货郎担问题为例加以说明。

若有一货郎想走遍 N 个村庄，试问怎样走可使路线最短？这是一个世界性的难题，至今未获得满意解决。所谓满意解决，是指找到一种算法，使得能用计算机在可以承受的时间内得到结果。遗憾的是，现有算法都需要太长的计算机时间——几年，因而事实上做不到。

为便于说明起见，我们用最笨的算法来计算。第一步，找出所有的路线。第二步，将所有的路线长度两两比较，长的弃去，短的留下。这样，把所有的路线都比较过，最短路线自然就出来了。我们取 $N=31$。31 个村庄，对货郎担来说，不算多。

先计算路线数。设货郎在某村庄出发。他的第一站有 30 种选法。第二站有 29 种选法，第三站有 28 种选法。依此类推，全部可以走的货郎路线是 30！条，其数量级是 10 的 32 次方：

$$30! \approx 2.6 \times 10^{32}$$

然而，1 天 = 86400 秒。一年 $= 3.2 \times 10^7$ 秒。倘若我们用每秒一亿次（即 10 的 8 次方）电子计算机，不停地运转一年，可以运算

$$10^8 \times 3.2 \times 10^7 \approx 3.2 \times 10^{15}\text{次。}$$

因此，要完成 30！次，需要大约 0.8×10^{17}年，即 8 亿亿年！即使用每秒 1 亿亿次的电子计算机（现在还没有），用这一算法来计算，还得 8 亿年，也是办不到的事。

当然，大家会说，这个办法太笨了，有些一眼就看出是兜圈子的路线，根本不需要拿来比，及早把它们剔除就是了。确实，数学家正在努力这样做。他

们定的标准是：找一种算法，能在 N 的若干次方的运算时间内完成货郎担最短路线问题。这种算法，称为“多项式算法”。大家知道，$N!$ 的数量级和 10 的 N 次方差不多，我们把这种计算次数等于某数的 N 次方的算法，称为指数式算法。上述的货郎担问题笨算法就是一种指数式算法。至于货郎担问题是否有多项式算法，至今尚未找到答案。记得 1979 年，《纽约时报》刊登消息，说苏联的哈奇扬找到了货郎担问题多项式算法。其他各国的报纸也多方报道，包括中国的《参考消息》。后来经过验证，原来是哈奇扬提出了一种求解线性规划问题“椭球算法”，这是多项式算法。但椭球算法不能用于货郎担问题。即便如此，哈奇扬也获得了很高的国际声誉。可以想像，如果有朝一日找到了货郎担的多项式算法，一定会轰动世界，为世界各大报纸争相报道。

1.5 数据压缩与方正集团

数学是思维的体操，数学是科学的工具。如果说数学是智慧的源泉，那么人们把数学奉为“科学的女王”。如果说数学推动社会的发展，那么数学是一门服务性的科学，一直是幕后英雄，“科学的侍女”。

但是，到了 20 世纪的下半叶，数学发生了重大变化：从幕后走向前台，成为能够直接创造财富的数学技术。且不说 CT 扫描、密码破译、军事模拟、最优控制等广为人知的数学技术，只就最火红的计算机软件来说，数学也在其中发挥重要作用。反过来，软件事业也为数学的发展带来了生机。这里我们介

绍“微软公司”的数学研究，以及中国方正集团依靠数学的成功。

比尔·盖茨(Bill Gates)创建的微软公司，已成为世界的首富。1997 年，它的总收益是 113.5 亿美元。1998 年，美国数学会的记者报道了微软公司对数学的投资，题目是“理论进入利润：微软在数学投资”。①

报道说，1991 年微软公司成立“微软研究”(Micrisoft Research)实验室，现在计划每年投入研究的费用是 26 亿美元。1998 年聘用了 250 名科学家，到 2000 年这个数字将是 600 名。管理这个实验室的是三个人：

Richard Rashid　微软公司副总裁　计算机科学家，

Daniel Ling　实验室主任　原 IBM 工程师，

James Kajiya　实验室副主任　资深图论专家。

许多人认为，这是美国继 Bell 实验室、IBM 实验室之后的又一信息科学的实验室。不过，前两个实验室研究硬件、系统和软件三部分，而微软研究实验室只研究软件领域：计算机图，数据库，程序语言，运算系统，网络系统，等等。

微软研究实验室有一个“理论组”，负责的是一对夫妇：原加州大学洛杉矶分校数学教授 Jennifer Chayes，和她的丈夫 Christian Borgs，原德国莱比锡大学统计物理学教授。1998 年，大名鼎鼎的 Michel

---

① Ally Jackson: Theory into Profit: Microsoft Invest in Mathematics. Notices of AMS. Vol.45. No.6. 1998。

Freedman加入微软研究实验室的理论组。Freedman以解决四维的庞加莱猜想而获1986年的菲尔兹奖，近年来对理论计算机研究感兴趣。现在他仍是加州大学圣地亚哥分校的教授，以后逐步转入微软公司。另外一位数学家是Jeong Han Kim，他是图论专家，用图论的半随机方法解决了一些60年未决的图论问题，获得1997年的Fulkerson奖。

据报道，微软研究实验室在数学方面的主要目标是研究 $P=NP$ ？Freedman注意到，物理学家威顿(Witten)和琼斯(Jones)的纽结多项式和 $P=NP$ 问题有关，并已经着手研究威顿工作和理论计算机的联系。

微软研究实验室有充足的经费。Chayes说，我们几乎没有预算。只要需要，开支就是了。现在每年有5～10位理论计算机科学家来访问，4～5位博士后，50位左右的短期访问者。访问者除报销旅费之外，每天可获得300美元的生活补助，这是相当舒服的生活了。

可以预料，由于像Freedman这样的名家加盟微软，特别在 $P=NP$ 问题上加强研究，这也许会开辟数学研究的一个新时代。

以下是一位中国数学家兼企业家的故事。①②③

王选，中国科学院院士，中国工程院院士，北京

---

① 《中国科学院院士自述》，上海教育出版社1996年，732页。

② “王选访谈录”，香港凤凰电视台，杨澜工作室，1998年12月。

③ “王选言商”，中央电视台，经济半小时，1999年1月7日。

大学教授，激光照排实现汉字印刷术革命的“方正”集团的创造者。如果说陈景润是 80 年代中国知识分子的代表，其特征是个人忘我地进行科学研究，那么王选就是新时期的知识分子代表，其特征是把科学转化为生产力，使科学和企业联姻。碰巧的是，这两位都是数学家。

让我们来叙述王选教授的业绩。

王选，1937 年出生，江苏无锡人。曾就读于上海南洋模范中学，1958 年毕业于北京大学数学系，从事计算机科学的研究。

1975 年，国家有一个“758 工程”，即汉字信息处理系统工程，其中有三个子项目：汉字通讯，汉字情报检索，汉字精密照排。当时国际上的电子分色机技术状况是：光学机械式的二代照排机前途不大，字模管式三代机和飞点扫描式三代机正在走下坡路。数字存储的第四代技术将占主要地位。当时全国有五个单位在从事这项研究，但是有两家选择了二代照排机方案，另外三家选择了飞点扫描、字模管、全息模拟存储的技术途径。王选毅然选择了“数字存储”的方案，跳过二三代，直接研制第四代技术。这样做，困难虽大，却将合了数字化的信息技术的发展方向。

汉字字形信息量太大，数字化的困难是西方文字照排所无法相比的。王选说：“由于我是数学系毕业，所以很容易想到信息压缩，即用轮廓描述和参数描述相结合的方法描述字形，并于 1976 年设计出一套把汉字轮廓快速复原成点阵的算法。”但那时的计算机速度很慢，复原点阵需要的时间很长。如果王

选只有数学和软件的知识，他也就打退堂鼓了。然而他有微程序和软硬件的知识，就采用专门硬件配合微程序，使速度提高了几十倍，初步克服了困难。1976 年夏，终于决定跳过二三代机，直接研制“激光照排系统”。

到了 20 世纪 80 年代，轮廓描述西方文字的做法在国外大为流行，当时已由轮廓发展为三次曲线轮廓。由于王选的研究起步早，很快吸收了国外的一些经验，发明了高分辨率字形的高倍率信息压缩和复原技术，用于印刷照排系统。以后又设计专用的超大规模集成电路实行复原算法，显著改变了性能价格比。这一技术已经领先于国际先进水平，他所领导研制的华光和方正系统开始在全国的报社和出版社使用。

1980 年 2 月 21 日，当时任电子工业部长的江泽民同志，写信给党中央，报告王选的研究成果，建议不再向国外进口同类产品。邓小平同志批示表示支持。同年，《光明日报》在头版头条报道了这一科技成就。1985 年，中央电视台新闻联播也报道了这一消息。印刷业告别“铅与火”的历史，就此开始。

1985 年，方正系统列入全国十大科技发明成果；1986 年，获日内瓦科技发明奖；1987 年，获国家科技进步一等奖。

这些成果应该令人满足了，但王选却没有笑起来。他说：“国家花了这么多的钱，只是得奖却没有创造财富，如何对得起国家和人民？”

方正集团在继续前进。一些关键设备得到不断地改进。输出系统的一项重要设备，王选领导了前

五代，年青博士阳振钟领导了后两代。方正集团的技术研究院副院长、34岁的邹维发明了包括动画制作在内的一系列新技术。

1992年1月21日，《澳门日报》的彩色版用方正激光照排系统排版印刷，轰动海内外。台湾的印刷业同行在报纸上以“眼见为实”加以报道，原来用电子分色机三四个小时的工作，现在只要20分钟，乃至9分钟就可以完成。这样的技术是一种“挡不住的诱惑”。现在，台湾省国民党的机关报《中央日报》，最大报业集团《联合报》系，最大发行量的《自立晚报》都已采用方正的激光照排系统。一位台湾人士曾认真地说过，你们大陆的方正技术帮助台湾的国民党机关报，江泽民先生大概会不高兴罢。王选的回答是：“江泽民总书记一定会高兴的。”果然，王选在元宵夜谈到这消息时，江泽民同志非常高兴，认为这是两岸正常的科技交流。1995年，方正集团在香港股票市场上市，王选成为名副其实的科学家兼企业家。

现在，方正集团及其伙伴的产品占有国内排版系统的市场份额的99%，国外市场的70%以上。计算机技术是西方发达国家领先的领域，中国科学家要在某一方向上占据一个领先地位，非常不容易。王选的成功，具有很不平常的意义。其中，数学技术是关键的因素之一。

中国的科学家很多，但能赚钱的科学家不多，王选是杰出的一位。他给我们带来的启示是十分深刻的。让我们来看看王选自己的说法。

“科技顶天，市场立地”。王选的这一“顶天立

地”说，揭示了新时代科学技术发展的一种方向。

“可怕的满足感”。王选提到，在80年代初，科技人员最时髦的是出国，写论文，评职称。以为这样就是对国家做贡献，于是就满足了。那时埋头搞迎头赶上的新技术，做没有把握的技术攻关，有了成果也不能发表，没有文章不能升教授，所以那是很不受人欢迎的工作。据说当时的聂荣臻元帅对如此评教授也忧心忡忡。医生不看病，急着发论文。这种情形，是到改变的时候了。

“没有先进水平的填补国内空白是没有意义的”。我们常常见到报道说，某项科技成果达到80年代的国际水平，填补了国内的空白。这样的成果只能在展览会上陈列，当外国新产品进来时，“成果”已经是过时产品，除了报废别无用处。这样的科研“从立项之初便注定要失败”。“技术不超前，等于浪费国家资源。”这是何等宝贵的忠告啊！

“知识分子长期价廉就不会物美了”。一位领导人曾说中国知识分子是“价廉物美”。王选说，作为中国知识分子之一，能够为国家作无私贡献，值得自豪。但是，凡事都有一个限度。如果长期“价廉”，就不会“物美”了。

“方正集团要在未来造就5～10名院士，几十个百万富翁”。王选说，知识经济时代，知识更新非常之快。比尔·盖茨也许要在55岁退休。我已经61岁了，不可能再在技术上领导方正集团，必须交给年青人。他们做得比我好。这些人中间，应当出现中国科学院和中国工程院的院士。他们所创造的价值，也应该得到应有的回报，成为百万富翁。中国未

来的数学技术，将会百花齐放。但是，王选的榜样，将会永远激励后来者。

2 日常生活中的数据处理

数据伴随着我们终生。从出生的那天起，每个人的生日就是数据。待成人之后，每人都有一张身份证，上面打着独有的号码。银行的账号、信用卡、电话磁卡等等，都是一组数据。学校考试，结果是数据。到商店买东西，货物条形码、购货的数量、价格也都是数据。如前所说，数字化的世界，一切信息都已经数据化了。

大众数学，曾是人们的一种教育理想。在15世纪，文艺复兴之后，算术是人人都要学会的数学技能。一个人不会算账，就无法在社会上生存。到了18、19世纪，大家以为在工业化社会里，代数、几何、三角大概是人人必须掌握的。可是，时至今日，似乎并没有做到。那么，在21世纪，人人都必需的数学究竟是哪些？对现代公民来说，最重要的数学知识和技能是什么？美国芝加哥大学的Z. 尤什斯金认为，数据处理是最基本的内容之一，它比解方程和证几何题的要求更为基本。

2.1 数据的代表数：去掉最高分和最低分

大千世界，到处出现数据。例如，班上每人的数

学测验成绩；排球赛各个队员的得分数；某工厂工人的工资；各种牌号牙膏的日销量；100 只电灯各自的寿命；等等，摆在我们面前的是一串数字。但是光靠这一串数字还不能直接得出结论，往往需要进一步分析，从中提取有用的信息。

在上述例子中，我们常说："我班数学成绩平均为 82 分"，"上场的 6 名排球队员每人平均得分为 6.2 分"，"某工厂职工的平均工资为 67 元"，"大约 10 人中有 5 人喜欢某牌牙膏"，"灯炮寿命约为 1000 小时"。这里的 82，6.2，67，5/10，1000 等数字，就是根据原始数据（许多数字的集合）选出来的一个"代表数"。这些代表数，反映并且标志这些数字集合具有的"水平"，供我们作决断时参考。

那么怎样选取代表数呢？通常我们选取平均数。例如全班 30 人，第 k 个同学的数学成绩为 x_k 分，这时全班同学的数学考试水平就用算术平均数

$$\bar{x} = \frac{1}{30}(x_1 + x_2 + \cdots + x_k + \cdots + x_{30})$$

来代表。

算术平均数虽然具有很高的代表性，但是决不是唯一的，而且还有一些缺点和局限。我们还可以引进其他的代表数及方法。

先看体操比赛的例子。在体操比赛中，当一个运动员做完一套动作后，四个裁判分别给出四个评分。那么如何根据这四个评分，定出一个足以代表该运动员水平的分数呢？如果四个裁判的评分都一样，例如都是 10 分，那么该运动员无疑应得 10 分。但是通常四个裁判的评分是不一样的，例如四个评

分依次为

$$8.6,9.4,9.6,9.9。$$

按常规，我们取它们的平均数，此时应为 9.375 分。但是这时比赛场记分牌显示的却是 9.5 分。这是什么原因？

原来，按体操竞赛规则，评分时应将四个裁判的给分依小到大排列，删除首尾两数，然后将中间两数取平均数。这个数称为这四个数据的中位数。在上述例子中，中位数是 9.5。这就是记分牌上显示得 9.5 的来历。这种评分方法，可以避免个别裁判误判以及有意偏袒或压低的情况。例如在上述数据中，8.6 分是异常值，过分压低了，9.9 分则可能反映个别裁判的爱好。将它们删除后，评分就比较公正了。

中位数的一般定义如下：设有 n 个数据，将它们从小到大依次排列为

$$x_1, x_2, \cdots, x_k, \cdots, x_n。$$

如果 n 是奇数，则第 $\frac{n+1}{2}$ 项的值即为中位数；如果 n 是偶数，则取第 $\frac{n}{2}$ 项和第 $\frac{n}{2}+1$ 项的值作算术平均，以此数作为中位数。例如：

(1) 有 9 个数据。如 6，6.5，6.6，6.6，6.65，6.7，6.8，6.8，8，则取第 5 个数据 6.65 作为中位数。

(2) 有 10 个数据。设除上述 9 个数据外，再增一个数据 9，则应取第 5 个和第 6 个数据的平均值 $\frac{1}{2}(6.65+6.7)=6.675$ 作为中位数。

中位数的特征是它处于 n 个数据依大小次序

排列后的中间位置，即大于此数的数据个数和小于此数的数据个数一样多。中位数的优越性是不受个别特异值的影响。此外，与求算术平均数相比，不需要作繁琐的计算。

中位数是真正代表“中等水平”的，而一般的平均数有时并不能反映出“中等水平”来。例如某班有30人，有两位同学成绩很差，只得了10分，5个同学得90分，22个同学得80分，某同学得了78分。这30个人的平均分数是

$$\frac{1}{30}(2 \times 10 + 5 \times 90 + 22 \times 80 + 78)$$

$$= \frac{1}{30} \times 2308 = 76.9。$$

如果以平均数76.9分为平均数，那么某同学得了78分岂不是中等偏上，属中上水平了吗？但是，这位同学在班上是倒数第3名！如果取中位数，那么这30个数据中第15位和第16位都是80分，中位数是80，得78分只能是中等偏下，这就比较真实地反映了客观情况。

在资本主义国家，常常把大资本家的收入和普通工人的收入作平均，表明该国人均收入很高，其实，大多数人的收入在该平均数以下。因此在资本主义国家，用平均数来掩盖一些不平等现象是屡见不鲜的事。

除了中位数外，还可选择众数作为代表数，即数据中重复出现次数最多的那个数据。例如全班30人所穿鞋子的尺码为

尺码号	33	34	35	36	37
穿该号码的人数	5	6	15	3	1

用什么数来代表全班所穿鞋子的尺码？如取平均数，得到平均尺码约为 34.63 码。如取众数，则取班上同学穿得最多的那个尺码——35 码（有 15 个人穿）。显然，35 比 34.63 更有代表性。这是因为市场上并不生产 34.63 码的鞋，此数值参考意义不大，而 35 码这个数却有重要意义。例如一家鞋店如果拿 35 码的鞋到该班来推销，自然是最有销路的。

众数往往作为反映一般水平的标志。例如本班同学成绩大多在 70～80 分这一段；在各种牙膏中 A 型最受人喜爱；新发行的 10 张唱片中，第×号最受欢迎；一场篮球赛中，得 10 分的队员最多等等都是。通常的“最佳”、“最受欢迎”、“最畅销”等等都是用投票法取众数得到的，它反映人们最普遍的倾向，因而有广泛的应用。

上面介绍了三种代表数：平均数、中位数和众数。让我们来考虑下面的例子，不同的人从不同的角度会选取不同的代表数。

例　设在某资本主义国家里某厂职工的月工资支付情况为：

月工资数（美元）	得此工资的人数
10000	1（总经理）
8000	2（副经理）
5000	2（助理）
2000	5
1000	12
900	18

800	23
700	5
500	2

那么如何来选取代表数，即如何用一个数字来标志该厂职工的月工资？经计算，平均数约为 1387 美元，中位数为 900 美元，众数为 800 美元。工厂主为了说明本厂工资水平高，自然用平均数 1387 美元。但是达到此水平以上的人数只有 10 人，少数人的高工资提高了平均工资数的水平。工会领导人则说，本厂工人月工资为 800 美元，因为拿 800 元的人最多，最有代表性。而税务官则采用中位数，他认为征收所得税时应当了解目前的税率对多数职工有利还是不利，因而中位数最有代表性。这样一来，三种代表数各有各的用处，各有各的目的。我们在研究某些统计数据的时候，不可不注意到这一点。

现在让我们总结一下这三种代表数的优缺点。

算术平均数通常使用最广，它可以用公式表示，且考虑每个因子提供的数量信息，因而最能标志数据集中的趋势。其缺点是容易受异常数值的影响，以至失去代表性。

中位数只和数据所处的位置有关，不受特大或特小的异常数值的影响，计算简单，往往一眼就看出来，且最能反映中等水平。不过如果数据明显地在中位数附近密集，而且重复次数很高，那会形成中位数不在中间的情形。例如，1，1，1，2，2，2，2，2，4 这 9 个数据，中位数是 2，这时就很难说它代表中间状态了。另外中位数没有公式可计算，不能反映许多数值提供的全部数量信息，这是优点中埋伏着的缺点。

众数由于出现频率最高，常被人们采用。尤其在评“最佳”的时候，总以得票最多为依据。不过众数只讲频率，不讲数值大小，有和中位数一样的缺点。特别是当有多个数据出现频率一样的时候（并列第 1 名），就很难选择众数了。

最后，我们列出一批数据，请大家根据需要，挑选适当的代表数。

（1）在一个 20 人的班级中，他们在某学期出勤的天数是：7 人未缺课，6 人缺课 1 天，4 人缺课 2 天，2 人缺课 3 天，1 人缺课 90 天。试确定该班学生该学期的缺课天数。

（2）确定你所在班级中同学身高的代表数，如果是为了：①体格检查，②服装推销。

（3）一个生产小组有 15 个工人，每人每天生产某零件数目是 6，6，7，7，7，8，8，8，8，8，9，11，12，12，18。欲使多数人有产可超，每日生产定额（标准日产量）应为多少？

（提示：（1）取中位数；（2）①取平均数，②取众数；③取中位数。）

2.2 数据的运用：“公说公有理，婆说婆有理”

香港的中学数学教材里，有这样一节，名为“公说公有理，婆说婆有理”。数学教材里有这样的标题，在大陆十分罕见。在一些人看来，数学应该板起面孔才是。且让我们来看其内容。

某企业有 5 位股东，100 名工人。1990、1991、

1992 三年的利润分配情况是：

年份	工资总额	股东红利
1990	10 万元	5 万元
1991	12.5 万元	7.5 万元
1992	15 万元	10 万元

在企业从业人员大会上，股东老板上台画了一张图（见图 2.1(a)），标题是“有福共享，有难同当”。图上是两条平行线。三年来工资和股红都增加了 5 万元，劳资双方的利益同步增加。

但是，工会的负责人说，我也来画一张图（见图 2.1(b)）。大家都以 1990 年为基础，定为 100%。三年来，工资总额从 10 万元到 15 万元，增长了 50%，而股红却从 5 万元到 10 万元，增长了 100%，翻了一番。所以，工资增长速度赶不上股红速度，今后应当更多加工资。

一位工人的发言指出，每个工人的平均工资从 1000 元增加到 1500 元，股东红利从 1 万元增至 2 万元。我也画一个图（见图 2.1(c)）。标题是一个太高，一个太低。工人工资应该多多增加。

请大家注意，香港特别行政区实行“一国两制”，过去和现在都是实行资本主义制度。所以，这节数学课的标题是“公说公有理，婆说婆有理”。大家都有理。至于如何处理不同意见，那是劳资协商的问题了。

令人深思的是，在江苏某地师范学校附中高一的测验，全班 40 人所画的图，竟然全都是图 2.1(a)：两条平行线。这难怪学生，因为我们的数学教材、数学教学中从来只有把图表转换为函数图像的一种画

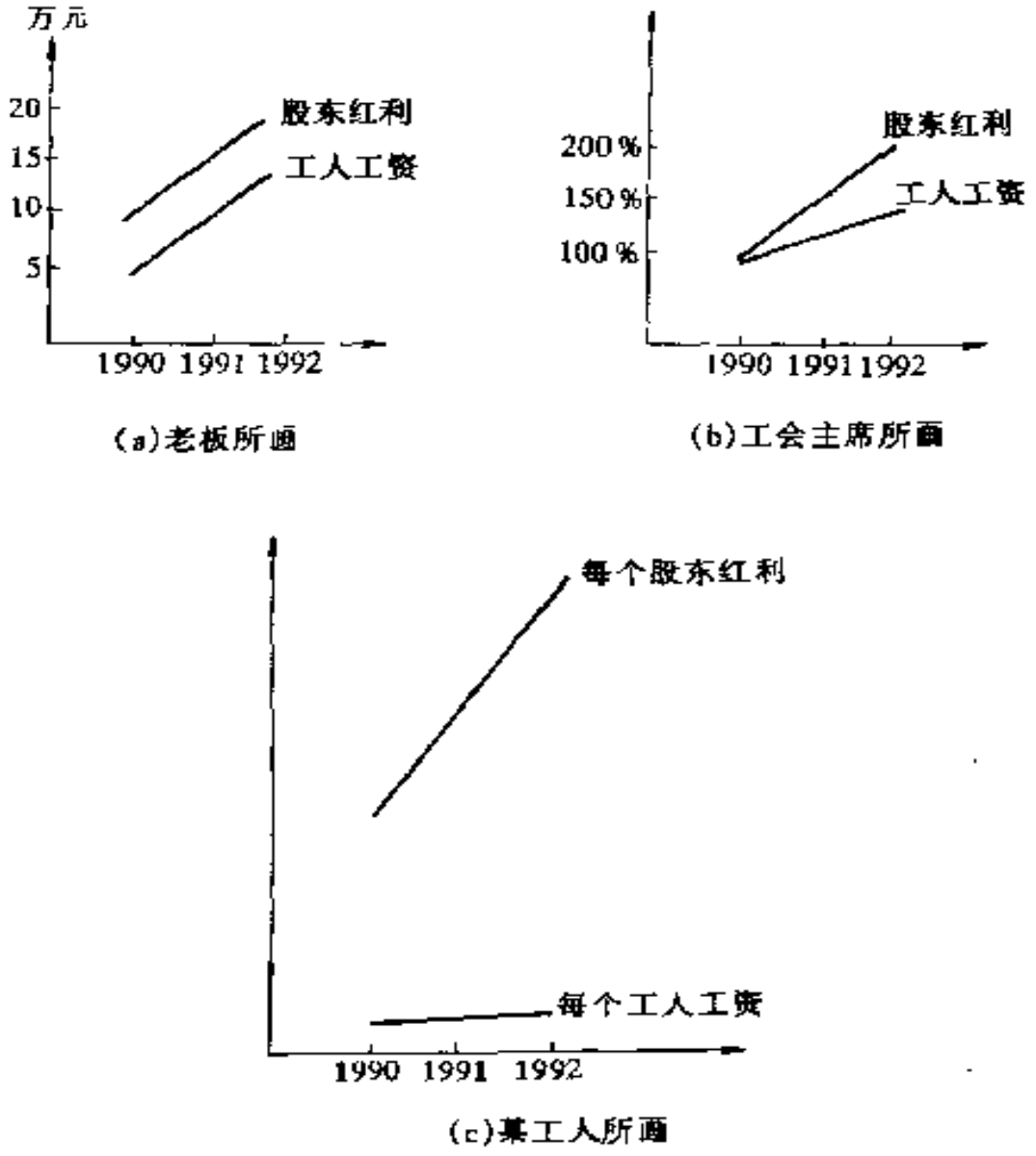


图 2.1

法，从来不教数据处理。

记得前些年，河南有位孙天帅在珠海某老板面前不下跪，深受传媒关注。但是，在受了 9 年义务教育之后的打工仔和打工妹，如果老板画出“有福共享，有难同当”的图 2.1(a)，他们只会点头同意。这岂不在精神上也是一种下跪。

2.3 把报纸上的数据和图表读懂

以下是一组从报章杂志和现实生活中选来的问题。

2.3.1 房地产广告

《新民晚报》1993 年 1 月 24 日登载一则泰信和(无锡)房地产广告，如图 2.2 所示。请回答下列问题：

(1) 大约在哪几年，日本和中国台湾的价值变化率相同？

(2) 1980 年后，日本和中国台湾的价值上升率何者快？

(3) 台湾的价值变化率上升最快的是哪几年？

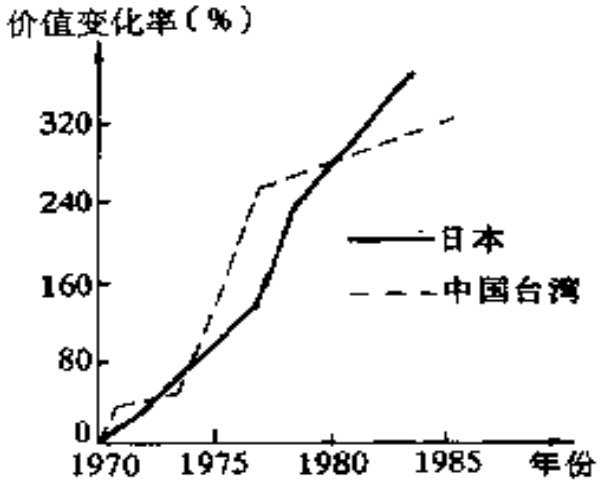


图 2.2 房地产价值变化示意图

解 (1) 1973、1974 和 1980 这三年。

(2) 日本。

(3) 1974 年至 1977 年之间。

说明 社会主义市场经济的发展，要求我们经常从报刊、杂志、电视等新闻媒介中获取大量有用信息，而数据、表格和图表是最常见的形式。

2.3.2 价格与成交量趋势图

图 2.3 所示为上海物资贸易中心信息部提供的一份上海金属期货市场 1# 电解镍价格与成交量趋势图，原载 1993 年 4 月 5 日《解放日报》。请从图中读出下列数字(近似数)：

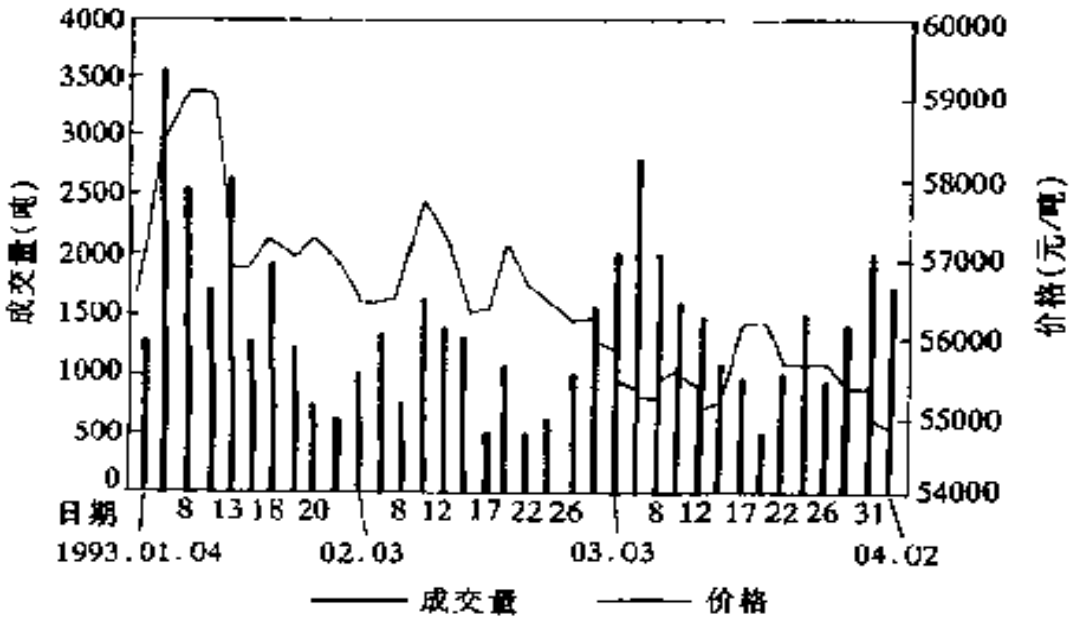


图 2.3 上海金属期货市场 1# 电解镍价格与成交量趋势图

(起始时间:1993 年 1 月 4 日;终止时间:1993 年 4 月 2 日)

(1) 最大成交量和最小成交量；

(2) 最高价格和最低价格；

(3) 从图中判断以下断言是否正确：

A. 平均成交量大于 2500 吨，

B. 平均价格低于 58000 元，

C. 1 月 8 日至 1 月 13 日价格下跌最快，

D. 成交量大时价格必然高，

E. 1 月 23 日春节期间成交量最低，

F. 成交量的第二个高峰产生于低价格时，

G. 价格的总趋势呈下降态势，

H. 成交量的趋势呈波浪形。

解　(1),(2)由读者据图 2.3 读出。

(3) A. 错；B. 对；C. 对；D. 错；E. 错；F. 对；G. 对；H. 对。

2.3.3　股市走势图

1993 年 1 月 24 日《文汇报》登载了一幅根据深圳大学股票分析智能系统 TSAS3.0 绘制的上海股市走势图(图 2.4)，请从图中提供下列信息：

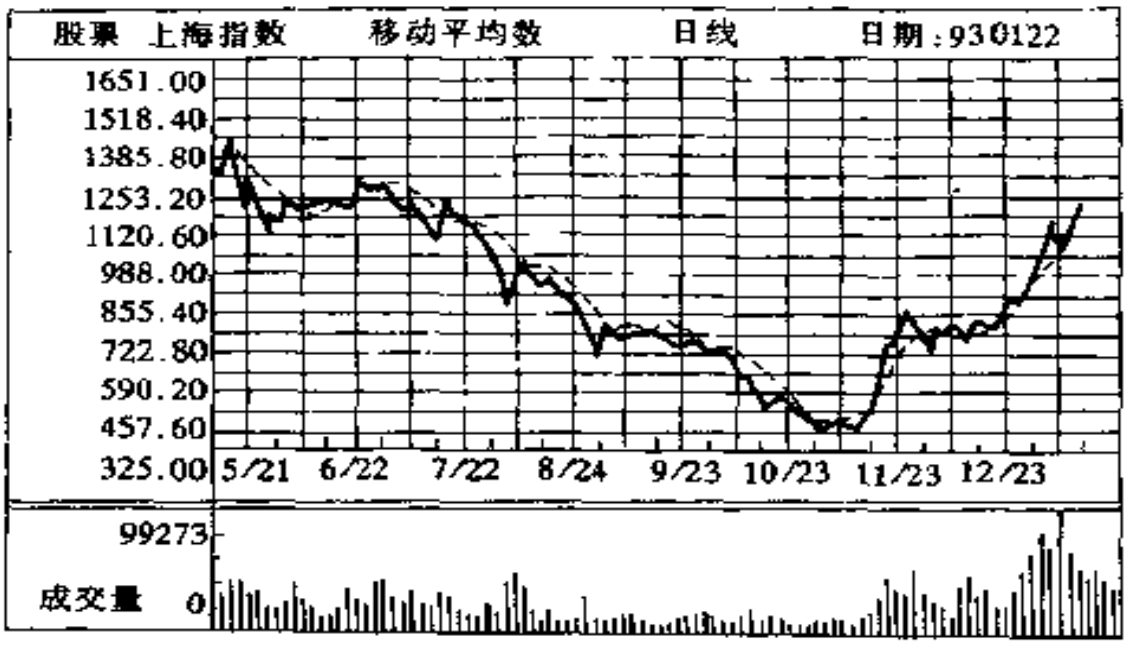


图 2.4　上海股市走势图

(1) 从 1992 年 5 月 21 日到 1993 年 1 月 22 日之间，大约在何时上海指数处于最高峰？最高峰值是多少？大约何时处于最低谷？最低指数是多少？

(2) 成交量何时最高，其成交股数是多少？指数最高的时候其成交量是多少？

(3) 11 月 23 日到 12 月 8 日的半个月内指数上升最快，试测算这段时期的日平均增长率。

解 (1) 约在 5 月下旬达到最高指数约 1440 点，在 11 月 23 日左右达到最低指数约 440 点。

(2) 成交量在 12 月 25 日至年底的某天达到最大值 99273 股。在指数最高的 5 月下旬，成交量约为 4.3 万股。

(3) 从图中估算，11 月 23 日为最低指数 391.30 点，12 月 8 日约为 830 点。故日增长率为：

$(830 - 391.30) \div 15 \approx 29.25$(点／天)。

说明 股票将会成为未来经济生活中重要的一部分。股市走势图公之于报刊，乃是“大众数学”的组成部分。

图 2.4 直接取自计算机荧屏，不甚清晰，但可供大致地估算、了解股市的走向，由图得到信息，应注意单位，善于捕捉各种信息来源，提出有价值的问题。此题是开放性的，读者可根据图提出其他问题进行研究。

我国教材中的函数图像，都是有解析式的，对这种实际的函数图很少接触，但实际应用很多，值得重视。

2.3.4 有奖销售

1993 年 4 月 1 日上海《新民晚报》刊登一则有奖销售广告(见下页)。计算奖品总金额占销售总额的比例，并与该公司若实行 9.8 折的销售方法相比较。

解 如全部奖券都是批发销售，则这 1 万张奖

券共销售 1000 万元，奖金额只占销售额的 0.51%。若 1 万张奖券都以零售额发出，则销售额为 400 万元，奖金额占总销售额 1.275%。

若该公司以 9.8 折销售 1000 万元，实际让利给顾客的金额占 2%。因此，有奖销售让利给顾客的金额比打 9.8 折少。

上海橡胶塑料五金公司

'93 系列有奖销售活动

一、有奖销售活动起讫日：1993 年 4 月 1 日起，奖券 10000 张发完为止。

二、凡累计批发额满 1000 元或累计零售额满 400 元，发奖券一张。

三、开奖日期：详见 5 月中旬的《新民晚报》。

四、本活动由黄浦区公证处公证，并请顾客代表参加当天的开奖仪式。

五、奖品设立：

特等奖 2 名 2000 元(奖品)，
一等奖 10 名 800 元(奖品)，
二等奖 20 名 200 元(奖品)，
三等奖 50 名 100 元(奖品)，
四等奖 200 名 50 元(奖品)，
五等奖 1000 名 20 元(奖品)，
奖品总金额 51000 元，中奖率 12.82%。

上海橡胶塑料五金公司
北京东路 697～711 号

折扣售货是平均地让利给顾客，而有奖销售则

是将利润的一小部分集中给几名中奖顾客。这种中奖的刺激能促使销售额增加，但实际让利并不多。

说明 此题所用的数学知识很少，仅仅是四则运算和百分比，但是，在市场经济大潮面前，了解销售方法，特别是目前盛行的有奖销售活动的数学背景，仍是有现实意义的。此题选自现实广告，里面有很多数据，需要精心选择，并加以评估。

2.4 彩票中奖和交通事故：哪种概率高

1998 年的上海福利彩票可中 100 万元大奖，摇奖的镜头，刺激了许多上海市民想碰碰运气。确实，发行彩票是筹集福利基金的有效措施，世界各国多有采用。1999 年 3 月，上海电视台有这样的镜头：一位青年男子准备拿出 1000 元买福利彩票，说是要“搏一记”，中奖了好买房子。旁边一位老伯伯身穿“福利彩票”的工作服，却诚恳相劝：“你买彩票我们欢迎，但花 1000 元想‘搏一记’买房子，我们劝你不要买。因为中 100 万大奖的可能性实在是太小了。”这位老伯伯的话是理性的。买福利彩票的目的是为了“献爱心”，同时碰碰运气。如果一门心思要中大奖，那是缺乏概率知识的表现。中国数学教学中没有概率，实在是一大失误。

1997 年 1 月 13 日《文汇报》报道，1996 年中国因交通事故，“7 万冤魂葬身车轮之下”。让我们来计算一下概率。我国人口为 13 亿。所以，每人在一年内因交通事故死亡的概率，近似于 7 万/14 亿，即 2 万分之一。但是，福利彩票、储蓄等的中奖机会都

在十万分之一，或更低。

以下是--件真实的事：

某糖果厂为新产品问世举办1994年春节促销活动，方式是买一份糖果摸一次彩，摸彩的器具是绿、白两色的乒乓球，这些乒乓球的大小和质地完全相同。该厂拟按中奖率1%设大奖，其余99%则为小奖。厂方请教了数学老师，问数学老师应提供怎样的设计？

数学老师首先验看了厂方提供的器具：棱长约为30cm的立方体形木箱，密封良好，不透光，木箱的上方可容一只手伸入，另备足够多的白色乒乓球和少量绿色乒乓球。数学老师提供了五种方案供厂方选择。

方案一：在箱内放置100个乒乓球，其中1个为绿色乒乓球，其余99个均为白色乒乓球。顾客一次摸出1个乒乓球，如果为绿色乒乓球，即中大奖，否则中小奖。

方案二：在箱内放置14个乒乓球，其中2个为绿色乒乓球，其余12个均为白色。顾客一次摸出2个乒乓球（或分两次摸，每次摸出1个乒乓球，不放回），如果摸出的2个乒乓球均为绿色，即中大奖；如果摸出的2个乒乓球均为白色，或1个白色、1个绿色则中小奖。

数学老师告知厂方，采用这个方案，大奖的中奖率为1/91，略高于1%，所以必须稍多准备一些大奖的奖品。

方案三：在箱内放置15个乒乓球，其中2个为绿色，其余13个均为白色。顾客摸球和中奖的办法

与方案二相同。

数学老师告知厂方，采用这个方案，大奖的中奖率为 1/105，略低于 1%，所以可略少准备一些大奖的奖品。

方案四：在箱内放置 25 个乒乓球，其中 3 个为绿色乒乓球，其余 22 个均为白色。顾客一次摸出 2 个乒乓球（或分两次摸，每次摸出 1 个乒乓球，不放回），如果摸出的 2 个乒乓球均为绿色，即中大奖；如果摸出的 2 个均为白色，或 1 个白色、1 个绿色，则中小奖。

数学老师告知厂方，采用这个方案，大奖的中奖率恰好是 1%。

方案五：在箱内放置 10 个乒乓球，其中 3 个为绿色，其余 7 个均为白色。顾客一次摸出 3 个乒乓球（或分几次摸，一次摸 1 个或 2 个，共摸出 3 个，不放回），如果摸出 3 个乒乓球均为绿色，即中大奖；否则，只要摸出的乒乓球有 1 个为白色，就中小奖。

这个方案大奖的中奖率为 1/120，比方案三的中奖率还要小。

同时数学老师认为这个方案还可以加些花点子，即当顾客摸出的 3 个乒乓球中，如 2 个为绿色，只有 1 个为白色，在顾客再购买 10 份糖果的条件下，允许顾客在所剩余的 7 个乒乓球（其中 6 个白色、1 个绿色）中再摸 1 个，如果摸出的是绿色乒乓球，仍得大奖，否则就得 1 个小奖。这种花点子也许能提高顾客的兴趣。

厂方派出销售代表并邀请几位商业营销心理学家对数学老师所提供的五种方案进行了评估。

在评估会上，数学老师介绍了设计每个方案的数学依据。

设 $P_i(i=1,2,3,4,5)$ 是第 i 个方案中大奖的概率，则

$$P_1 = \frac{1}{C_{100}^1} = \frac{1}{100};$$

$$P_2 = \frac{1}{C_{14}^2} = \frac{1}{91};$$

$$P_3 = \frac{1}{C_{15}^2} = \frac{1}{105};$$

$$P_4 = \frac{C_3^2}{C_{25}^2} = 3 \div \frac{25 \times 24}{1 \times 2} = \frac{1}{100};$$

$$P_5 = \frac{1}{C_{10}^3} = \frac{1}{120}。$$

关于在方案五中加的花点子，顾客中大奖的概率为

$$P_5' = \frac{1}{C_7^1} = \frac{1}{7}。$$

考虑到厂方原来拟定的中大奖概率为 1%，顾客似应购买 15 份糖果（至少 14 份）才允许他补摸一次，为什么只要求购买 10 份糖果，就允许顾客补摸一次，这样厂方不是吃亏了吗？数学老师说，这里有个计算问题，对厂方和顾客来说都是公平的，厂方并不吃亏，并不需要多支出奖品的费用。原来每个大奖的奖品的价值是 400 元，小奖的奖品是 2 元。于是按 1%的大奖中奖率，摸一次彩的期望值为

$$400 \times 1\% + 2 \times 99\% \approx 6(元)。$$

现在顾客购买 10 份糖果，本来可以摸 10 次彩，其期望值应为

$$6 \times 10 = 60(元)。$$

而实际上顾客仅补摸一次，按得奖的办法，其期望值为

$$400 \times \frac{1}{7} + 2 \times \frac{6}{7} = \frac{412}{7} \approx 59(\text{元})。$$

两相比较，期望值几乎相同，所以对厂方和顾客来说都是公平的。

评估的结果，大家认为：

关于方案一：中大奖的概率为 1/100，恰好符合厂方的要求，但缺点是箱内装的乒乓球个数太多，不便顾客查验，一个个地数又不大可能，所以顾客容易产生“中大奖太难”、“是不是真的装了 100 个乒乓球？”等心理。

关于方案二：中大奖的概率为 1/91，较厂方的要求略高，但还是很接近的。由于箱内装的乒乓球个数不多，便于顾客当面查验，容易产生信任感，也可能会产生“中大奖很容易”的心理，增强了顾客购买糖果的欲望。

关于方案三：中大奖的概率为 1/105，略低于厂方的要求。这个方案也和方案二一样，便于顾客当面查验；并且这个方案的大奖中奖率虽较方案一为低，如果顾客不精于计算的话，仍然可能产生“中大奖不难”的心理，尤其是当顾客大部分可能是儿童的时候。皮亚杰的实验告诉我们，10～12 岁的儿童形成数量守恒的概念要经过 1～2 年的训练，何况这里的守恒概念要涉及到求组合数的计算。在数学上，两个方案何者概率较大，这是不容欺骗的事实，但在商业心理上顾客的心态则是不可忽视的一种考虑。

关于方案四：中大奖的概率为 1/100，恰好符合

厂方的要求，也具有方案二和方案三的优点，但是在顾客估算中大奖的可能性时可能会感到困惑。

关于方案五：中大奖的概率为 1/120，低于厂方的要求。箱内装的乒乓球个数很少。顾客查验时一目了然。但一次摸出 3 个乒乓球，对于儿童顾客来说，手太小，可能有困难，这样就要分几次摸，当摸彩的顾客很多的时候，大家可能等得不耐烦。

亲爱的读者，如果你是厂方的话，将采纳哪个方案呢？

说明 本题是浙江教育学院戴再平教授遇到的一件事，它只要用到少量数学知识，希望读者能够解决这类问题。

2.5 超市里的数据和向量

在物理课中学过向量，即具有大小和方向的量称为向量。例如，力，有大小和方向，因而是向量。然而，数据不过是一组数，怎么会和向量搭界呢？

大家知道，向量可以用坐标来表示。在中学里有复数一节，说的是复数 z，平面向量 Oz，和一对实数 (a,b) 是一一对应的，如图 2.5 所示。

$$z = a + bi \leftrightarrow (a,b) \leftrightarrow Oz$$

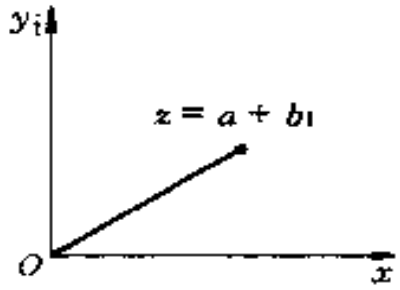


图 2.5

请大家注意，(a,b) 正是两个有序的实数，无非是有前后顺序的两个数据而已。推而广之，我们把 n 个有序的数据 $a_1, a_2, \cdots, a_n$ 称做一个 n 维向量 A，它有 n 个坐标：

$$A = (a_1, a_2, \cdots, a_n)。$$

这样的 n 维向量，我们随处可见，并不陌生。看看你每个月的工资单：

职务工资 岗位津贴 书报费 午餐补贴 奖金 总数

(800　230　40　100　350　1520)

就是一个 6 维的向量。所谓“有序”，无非是指每个位置都有特殊的意义，不可随意颠倒。

向量的好处是可以作加减运算，把它们的对应项做加(减)法就是了：

$$\begin{aligned} A + B &= (a_1, a_2, \cdots, a_n) + (b_1, b_2, \cdots, b_n) \\ &= (a_1 + b_1, a_2 + b_2, \cdots, a_n + b_n)。 \end{aligned}$$

你要计算两个月工资收入的明细数目，就需要把两个月的工资向量相加

$$A = (800, 230, 40, 100, 350, 1520),$$

$$B = (800, 160, 40, 0, 0, 1000),$$

$$A + B = (1600, 390, 80, 100, 350, 2520)。$$

一个企业的财务科，当然需要把每个员工的“工资向量”保存在计算机里，要计算全体员工的领取的工资总额、奖金总额、书报费总额等等，只需将各人的工资向量统统加起来，就一清二楚了。

向量不仅可以相加，还可以相乘。现在让我们来看一个超市里买东西的情形。

现在设超市内有一家食品商店，拥有多种商品，为了简单起见，不妨说只有 5 种：蛋，糖，奶粉，藕粉和香肠。我们将它们编上号码，依次为 1 号、2 号、3 号、4 号和 5 号。这时来了顾客 A，他买了 4 斤蛋、1 斤糖和 10 斤藕粉。于是我们便得到顾客 A 的购货向量为

(4,1,0,10,0)。

请看,这不是实实在在的5维向量吗?每一个顾客来买东西,对于这个商店来说,只要用一个向量就能表示出该顾客的全部购货信息。同样,对于商店经理来说,店中货物的库存量也是一个向量,每日销售量也是一个向量。各种商品的价格也是一个向量,以上述5种商品为例,蛋每斤1元,糖每斤0.70元,奶粉每斤3元,藕粉每斤0.20元,香肠每斤4元,于是我们有价格向量

(1,0.7,3,0.20,4)。

那么顾客A应付款多少?显然应该是

$$(4,1,0,10,0)\cdot(1,0.7,3,0.20,4)$$
$$=4\times1+1\times0.7+0\times3+10\times0.20+0\times4$$
$$=4+0.7+2=6.70(\text{元})$$

我们不妨把6.70看作购货向量和价格向量的数量积。

综上所述,管理一个商店,所接触的数学信息无非是向量及其数量积。大型超级市场的商品种类成千上万,我们就考虑一千维或一万维的向量。人工书写或计算不方便,就把这些机械的记忆和计算规则由计算机去执行。所谓用电子计算机管理大商场,其最基本的一个数学工具,就是一组有序的数,一个向量!

数的加减乘除,我们都很熟悉。对向量来说,我们通常只讲加法和减法。中学里一般不涉及向量乘法。其实向量的乘法也不难掌握,常用的有两种:数量积和向量积。这里我们只谈数量积。

两个向量的数量积,得出的是一个数,它是对应

坐标乘积之和。对二维向量来说

$$(a_1,a_2)\cdot(b_1,b_2)=a_1b_1+a_2b_2。$$

一般地

$$(a_1,a_2,\cdots,a_n)\cdot(b_1,b_2,\cdots,b_n)=\sum_{i=1}^{n}a_ib_i。$$

上面的购货向量和价格向量的数量积，正是顾客 A 的付款数。因此，这样定义数量积是有客观依据的。

与向量有密切关系的还有矩阵，矩阵的概念在中学讲线性方程组时已经引入过。一个线性方程组，可以看成是系数矩阵在未知向量(x,y)上作用的结果。具体说来，我们有

$$\begin{pmatrix}a_1 & b_1\\ a_2 & b_2\end{pmatrix}\begin{pmatrix}x\\ y\end{pmatrix}=\begin{pmatrix}c_1\\ c_2\end{pmatrix}。\qquad(*)$$

将矩阵的第一行(a_1,b_1)(这是一个向量)和(x,y)作数量积即为a_1x+b_1y，而等式右边的向量(c_1,c_2)的第一个坐标为c_1，于是我们得到一个线性方程式

$$a_1x+b_1y=c_1。$$

同样，用矩阵的第二行(a_2,b_2)和(x,y)作数量积应为c_2，即

$$a_2x+b_2y=c_2。$$

因此，线性方程组

$$\begin{cases}a_1x+b_1y=c_1,\\ a_2x+b_2y=c_2。\end{cases}$$

可以写成矩阵和向量的形式(*)。也就是说，矩阵不仅是一个数字的阵列，还可以作用在向量(x,y)上使它变为另一个向量(c_1,c_2)。如果说函数是把数映射为数，那么矩阵就是一种把向量映成向量的

映射，它比函数概念更为宽广。

对于三维向量的情形，本质上是一样的，我们可以这样写

$$\begin{pmatrix} a_1 & b_1 & c_1 \\ a_2 & b_2 & c_2 \\ a_3 & b_3 & c_3 \end{pmatrix} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} d_1 \\ d_2 \\ d_3 \end{pmatrix},$$

它相当于

$$\begin{cases} a_1x + b_1y + c_1z = d_1, \\ a_2x + b_2y + c_2z = d_2, \\ a_3x + b_3y + c_3z = d_3。 \end{cases}$$

我们只需把矩阵的每个行向量(a_i, b_i, c_i)和(x, y, z)作数量积，并把它作为新向量的第 i 个坐标d_i，这里 $i=1,2,3$。很自然地我们可以令 $i=1,2,\cdots,n$，就得到 n 元线性方程组的矩阵表示。

那么，矩阵除了描写线性方程组以外，还有没有别的用处？矩阵的用处多得很，试举两例：

（1）购货矩阵

在上面食品商店管理的例子中，如果有四个特别购买户，购货数量如下：

	蛋	糖	奶粉	藕粉	香肠
A	4	1	0	4	0
B	7	10	5	0	0
C	10	8	7.5	100	31
D	2.5	7	6.4	232	9

那么，它就是一个特别购买户的数量矩阵，它作用于价格向量上，就得出付款向量

$$\begin{pmatrix} 4 & 1 & 0 & 4 & 0 \\ 7 & 10 & 5 & 0 & 0 \\ 10 & 8 & 7.5 & 100 & 31 \\ 2.5 & 7 & 6.4 & 232 & 9 \end{pmatrix} \begin{pmatrix} 1 \\ 0.7 \\ 3 \\ 0.20 \\ 4 \end{pmatrix} = \begin{pmatrix} 5.50 \\ 29.00 \\ 182.10 \\ 109.00 \end{pmatrix}$$

可见矩阵并不神秘，不过是日常生活中一些有序数组之间的变换关系而已。

（2）路径矩阵

设在南京、苏州、上海、杭州四城市之间有铁路相连，杭州与上海之间还有航空线。我们把这四城市间从一城市直接通往另一城市的交通路径（一级

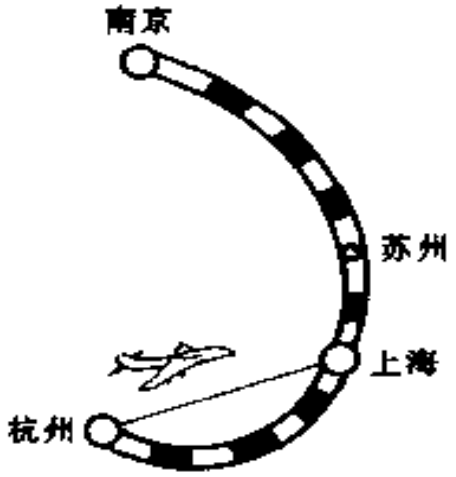


图 2.6

路径）数目写成矩阵形式：

$$\begin{array}{c} \\ \text{南京} \\ \text{苏州} \\ \text{上海} \\ \text{杭州} \end{array} \begin{array}{c} \begin{array}{cccc} \text{南京} & \text{苏州} & \text{上海} & \text{杭州} \end{array} \\ \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 2 \\ 0 & 0 & 2 & 0 \end{pmatrix} \end{array}$$

这一矩阵就成为路径矩阵。它能一目了然地表示出各城市之间直接联系的交通线数目，比其他形式更

为醒目，更突出数字。这一表示的好处远不止此，例如要问这四个城市之间，跨过另一个城市到某城市的不同路径(二级路)有几条？比方问杭州到苏州有几条路可走，我们说有两条，即杭州乘飞机到上海，上海乘火车到苏州，也可以从杭州坐火车到上海，再坐火车到苏州。上海经另一城市仍回上海的路径则有五条：①上海乘火车到杭州，再乘火车返上海；②上海乘飞机到杭州，仍乘飞机返上海；③上海乘火车去杭州，乘飞机返上海；④上海乘飞机到杭州，坐火车返回；⑤上海坐火车到苏州，仍坐火车返上海。仅从这样简单的地图出发，二级路就很难数得清了，再复杂一些的能不能看得明算得清，这就要依靠矩阵的乘法。

两个矩阵 A 和 B 相乘，得到新的矩阵 C，以 A 第 i 行的行向量和 B 的第 j 列的列向量作数量积，构成 C 在 i 行 j 列的元素 c_{ij}。

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & & & \vdots \\ a_{i1} & a_{i2} & \cdots & a_{in} \\ \vdots & & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1j} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2j} & \cdots & b_{2n} \\ \vdots & & & & & \vdots \\ b_{i1} & b_{i2} & \cdots & b_{ij} & \cdots & b_{in} \\ \vdots & & & & & \vdots \\ b_{n1} & b_{n2} & \cdots & b_{nj} & \cdots & b_{nn} \end{pmatrix}$$

$$= \begin{pmatrix} c_{11} & c_{12} & \cdots & c_{1j} & \cdots & c_{1n} \\ c_{21} & c_{22} & \cdots & c_{2j} & \cdots & c_{2n} \\ \vdots & & & & & \vdots \\ c_{i1} & c_{i2} & \cdots & c_{ij} & \cdots & c_{in} \\ \vdots & & & & & \vdots \\ c_{n1} & c_{n2} & \cdots & c_{nj} & \cdots & c_{nn} \end{pmatrix}$$

为了清晰起见，我们将上述路径矩阵自乘，看看得出什么结果：

$$\begin{pmatrix}0&1&0&0\\1&0&1&0\\0&1&0&2\\0&0&2&0\end{pmatrix}\begin{pmatrix}0&1&0&0\\1&0&1&0\\0&1&0&2\\0&0&2&0\end{pmatrix}=\begin{pmatrix}1&0&1&0\\0&2&0&2\\1&0&5&0\\0&2&0&4\end{pmatrix}$$

显然结果是二级路矩阵，矩阵乘法解决了我们的难题。

读者不妨自己考虑一个交通图，还可考虑三级路四级路的情形，那是饶有兴味的。

3 数据欺诈和数据滥用

在数字化的社会里，数据无处不在。人们用数据标志事物，用数据传达信息，以准确的数据报告科学结论，以诱人的数据吸引公众注意。这一切，织成了社会上的数据之网，包容着天下万象。但是，数据可以为人造福，也可以构成陷阱。数据欺诈，成为现代社会的一颗毒瘤，警惕数据陷阱，揭穿数据欺诈，应当是现代公民的基本素质之一。

3.1 “郭大王溜溜卷”名裂南京路

1995 年 11 月 6 日，上海《文汇报》记者王伟如、李浩明相继报道：“经商赢利岂能不顾社会责任——莫让‘溜溜卷’卷去孩儿心”，拆穿了台商“郭大王溜溜卷”的促销骗局。

所谓“郭大王溜溜球”，是台商郭某推销的一种玩具。出于新奇，上海的一些孩子十分痴迷。1995 年下半年，上海乔丰实业公司见有利可图，推出一种“郭大王溜溜卷”，其实不过是一袋五根的蛋卷，共重 120 克，售价 1.50 元。换言之，每千克高达 75 元。但同样的沪产蛋卷，每千克不过 8 元而已。

如果仅仅是贵，愿者购买，也就是了，出格的是

这种蛋卷的促销术。每包“郭大王溜溜卷”中附有一张彩色奖券，上面画有五角星（红）、航天飞机（白）、荡秋千（绿）、剪刀（黄）以及东方明珠（蓝）溜溜球造型的图案。集满红、黄、蓝、白、绿五种颜色的对奖券，可到指定地点兑换一只“郭大王霹雳龙珠”溜溜球（每只市场价格为 25 元）。此举一出，不少小学生卷入了收集对奖券的旋涡，都想比比谁的运气好。“郭大王”对销售网点的承诺是“中奖率是 20%”。

据一些小学校长反映，对奖券大部分是蓝、黄、绿、白四种颜色，红色很难收集。虹口区某小学，9 月拆启的 800 多包“溜溜卷”中，红色五角星只有 22 包，中奖率只有 2.8%。《文汇报》记者于 1995 年 11 月 8 日，来到上海南京路的第一食品商店实地调查。记者当场购买了 100 袋“郭大王溜溜卷”，打开一看，共有白色航天飞机 39 张，黄色剪刀 28 张，绿色荡秋千 22 张，蓝色东方明珠 10 张，而红色五角星只有一张！中奖率仅为 1%！这样，记者用 150 元买下的 100 包“溜溜卷”，只得到一张“溜溜球”的兑奖券。在食品一店被采访的一位顾客说：“开始我们都相信‘郭大王’关于 20% 的承诺。结果 4 个孩子买回 150 袋“溜溜卷”，花了 200 多元才换回一只溜溜球，‘郭大王’的心太黑了。”家住旧仓街 118 弄的刘女士说：“我的 7 岁儿子买了 112 袋‘溜溜卷’，竟然一张红五角星也没有！”显然，购买溜溜卷的中奖率，充其量不过百分之一、二、三。这场“郭大王 20% 中奖率”的承诺，是不折不扣的数据欺诈！

其实，我们可以简单地作些计算。假使“郭大王”不做假，红黄绿蓝白五种对奖券平均分配，出现

的概率都是 1/5。那么买五包恰好是五种花色的可能性是：

第一包：随便哪一种颜色：概率显然是 5/5＝1。

第二包：要求不是第一包的颜色，其概率为(1－1/5)＝4/5。

第三包：要求不是第一包和第二包的颜色，其概率是 1－1/5－1/5＝3/5。

第四包：要求不是前三包的颜色，剩下还有两种颜色可供选择：

$$1 - 1/5 - 1/5 - 1/5 = 2/5。$$

第五包：必须和前四包的颜色都不同，只有一种选择了：

$$1 - 1/5 - 1/5 - 1/5 - 1/5 = 1 - 4/5 = 1/5。$$

这样，买五包的对奖券的分概率为：

$$(5/5)(4/5)(3/5)(2/5)(1/5)$$
$$=(1/5)(1/5)(1/5)(1/5)(1/5)\cdot 5\cdot 4\cdot 3\cdot 2\cdot 1$$
$$=24/625 = 0.0384 \approx 0.04。$$

这就是说，若“郭大王”老老实实做生意，将五种花色平均投放市场，即 100 包内红黄绿蓝白五种对奖券各放 20 张，让顾客随意挑选购买，那么承诺近似的获奖率是 4% 是可以的。这就是说，用 150 元买 100 包“溜溜卷”(总重 2 千克)，一般可以有四套中奖，用它换走四只溜溜球，值市价 100 元，2 千克“溜溜卷”不过值 15～20 元，超额利润还有 30 元以上。这已经是一笔好生意了。但是，这位善“溜”的郭大王，在 100 包“溜溜卷”内只放一张红五角星，小学生自然大上其当了。

一般地，设红黄绿蓝白对奖券在总体中所占的

份额，即出现的概率分别为 P_1, P_2, P_3, P_4, P_5。那么在所买五包中恰好依次出现红黄绿白蓝的概率是五个概率的乘积：$P_1P_2P_3P_4P_5$。但是，对奖只要颜色不同就行，而不管颜色出现的次序。所以其他次序（如依次为蓝红黄绿白）同样可得奖。五种颜色出现的次序共有 5！种，因此得奖总概率为

$$A = (P_1P_2P_3P_4P_5)(5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)。\quad (*)$$

如果每种颜色出现的概率均等，即在 100 包中，5 种颜色各占 20 张：P_1, P_2, P_3, P_4, P_5 都是 1/5，那么前面已经算出来了：0.0384。

假如“郭大王”红的为 10 张，黄的为 30 张，绿、白、蓝各为 20 张，那么

$$P = 10/100 = 1/10,\ P = 30/100 = 3/10,$$

$$P = P = P = 20/100 = 1/5。$$

$$A = (1/10)(3/10)(1/5)(1/5)(1/5)(5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)$$

$$= 0.0288。$$

这时每 100 包中奖的概率不到 3 包。

若红的只有 5 张，黄的 35 张，其余仍为 20 张，则

$$A = (5/100)(35/100)(1/5)(1/5)(1/5)(5 \cdot 4 \cdot 3 \cdot 2 \cdot 1)$$

$$= 0.0168$$

若红的减至 3 张，黄的 37 张，其余各 20 张，则 $A = 0.01068$，非常接近 0.01。这时，100 包中出现不同颜色的机会只有 1/100，也就是记者在上海食品一店发生的那一幕了。

从数学上看，5 个数 P_1, P_2, P_3, P_4, P_5 的几何平均总小于代数平均，只在 5 个数相等时达到极大。也就是说，上述 5 个数的乘积在彼此相等时达到最

大。但这 5 个数之和必须是 1。它们彼此相等时，即都是 1/5 时，A 达到最大值。

看来，对付“郭大王”等欺诈商招，还得用些数学。

《文汇报》记者报道发表的次日，上海第一食品商店宣布，停销“郭大王溜溜卷”，保护消费者利益。这场闹剧就此收扬。

据上述报道，“郭大王”的这种欺诈行业，居然得到了有关部门的批准，完成了销售的“全套手续”。“有关部门”如此批准，若非“腐败行为”，便是对数据欺诈缺乏警惕。这从一个侧面反映了我国干部群众的数据意识十分薄弱，反数据欺诈能力不够强。至于中小学的数学教育大做其题，却很少涉及这类现实问题，实在也该好好改一改了。

3.2 美国的一桩“科学数据造假案”①

纽约洛克菲勒大学校长巴尔的摩，是一位诺贝尔奖获得者，因合作者的科学数据有假，受牵连后终于辞去校长职务，曾经轰动一时。

美国波士顿市的剑桥，坐落着两所著名的学府：哈佛大学和麻省理工学院。1996 年初夏，正是大学生放暑假离校、暑期班尚未开学之际。照例校园应该十分寂静。但是这些天来，大学走廊上、实验室里却是人声嘈杂，议论纷纷。原因是：历时十年之久的一件科学实验数据被篡改的案件即将公开宣判。

---

① [英]《Economist(经济学家)》，1996 年 6 月 22 日。

这就是著名的“巴尔的摩事件”。

戴维·巴尔的摩(David Baltimore)教授,美国最受尊敬的免疫学权威之一。80 年代时,他是麻省理工学院的“怀特海研究所(Whitehead Institute)的所长。麻省理工学院还有一个研究所,由 T. 伊曼尼什-卡利博士领导。两个实验室有一个合作项目,研究免疫系统的机制。那时,他们联合发表了一篇论文,刊登在极负盛名的《细胞》杂志上,署名者有卡利和巴尔的摩。

此后不久,卡利博士的实验室来了一位初级的研究人员 M. 奥'图勒博士(Dr. Margot O'Toole)。她计划研究上述论文如何进一步扩展的课题。她照着卡利博士在论文中所说的那样进行工作,但是原来说可以实现的结果,却一直做不出来,甚至连简单地重复原始工作也做不到。这自然引起奥'图勒博士的怀疑。于是她找出来一本原始纪录,那是一位和卡利博士有密切合作关系的技术员所使用的。这本只有 17 页的笔记本,显示了卡利博士有篡改数据的嫌疑。

此后,事件发生了一系列的变化。开始时,奥'图勒的说法没有什么人来关注。恰巧,卡利博士此时正在向麻省理工学院和也在剑桥地区的土富兹大学(Tuftz University)申请新的职位。两校都在进行一种非官方的调查。此事自然引人注意,经过媒体报道,遂广为人知。卡利博士断然否定有意作假的行为,各种说法不一而足。与此同时,奥'图勒博士在麻省理工学院的任职期满,不再聘用。事情又涉及到人事关系。

巴尔的摩教授是怀特海实验室的负责人，也是论文的署名者之一，名声特别大。尽管他不是直接当事人，也难辞其咎。90 年代，巴尔的摩教授已经离开麻省理工学院，担任纽约洛克菲勒大学的校长。奥'图勒博士的离职，也和他没有关系。但是人们还是把这一事件称为"巴尔的摩事件"。

由于事关美国科学的声誉，美国国会组织专门委员会举行听证活动，并设立了一个"科学公正办公室"(Office of Scietific Integrity)调查此事。由于事情越闹越大，巴尔的摩教授不得不辞去洛克菲勒大学校长职务。由于"科学公正办公室"不得力，遂被废除，代之以"研究公正办公室"(Office of Reaserch Integrity)主持调查。1996 年 6 月，研究公正办公室终于散发了一份报告：卡利博士有篡改科学数据的过失。卡利博士立即宣布要上诉。

不管卡利博士的上诉是否成功，巴尔的摩事件的影响是深远的。科学数据的作假会毁掉许多人的科学生命。这一事件的当事人，在科学声誉和学术生涯上都受到了深刻的伤害，包括无罪的奥-图勒博士。

3.3 原苏联李森科的"伪科学"数据[①]

原苏联的科学界，因为政治气候不正常，曾出现过许多冤案，也培植了一些伪科学。其中，李森科自

---

① Z.Y. 麦德维杰夫著，李宝恒、赵寿元译，《李森科浮沉录》，上海译文出版社，1980 年。

诽为“米丘林遗传学派”，把基因理论说成是“反动的”、“唯心的”，打击和迫害一些生物学家，造成了严重的后果。

李森科的“科学”工作，虽然一时十分“火爆”，被吹得神乎其神，但是纸包不住火，最后终于暴露真相，以至身败名裂。他最恶劣的一手便是伪造数据，欺世盗名，以保持他的学阀地位。

李森科（Николай В. Лысенко），1898 年生于乌克兰。1925 年毕业于基辅农业专科学校。1935 年任敖得萨植物遗传育种研究所所长。1939 年出任列宁全苏农业科学院院长。1965 年被解除职务。1976 年去世。

李森科的主要“工作”是提出植物阶段发育理论，冬小麦向春小麦转化的获得性理论，以及一个物种飞跃为另一个物种的进化理论等。这些理论后来被证明完全错误。这里我们不涉及他的生物学观点，只谈他的数据欺诈现象。

李森科能够青云直上成为全苏农业科学院院长，主要是冬小麦春化处理可以提高产量的研究工作。但是，在他的论文中所使用的数据全部是集体农庄填报的。在 1932 年《春化》第 2、第 3 期上的论文中，春化处理的增产结果是从 59 个集体农庄和国营农场的报告估计的。这些农场和农庄接到了关于春化处理的指示和几种格式的调查表，填好后由农庄主席和农艺师送交上级。调查表上必须填写的有：播种春化处理种子的面积和增产幅度。这就是李森科论文的主要数据。后来，这一方法推广到全苏联的几千个农庄和农场，导致了国家从春化处理

中获得了几百万公斤谷物的耸人听闻的官方报道。

李森科的研究报告从来没有统计分析所必需的重复试验，也没有用起码的春化种子和一般种子的对照实验，更没有考虑土壤肥力、田间管理等其他因素的影响。许多集体农庄都没有合格的农业实验人员进行科学工作。集体农庄主席不负责任地填写上级领导发下来的几十份表格，其中有一份是春化种子的增产数字。在把“反春化种子处理者”当作富农处理的时刻，表中“因春化种子处理所增产”的数字，其真实性便可想而知了。李森科及其追随者的论文，经常使用的数字是：“来自几千个集体农庄的田野的报告表明——”；“几百个集体农庄的产量试验报告表明——”；等等。

1936 年，育种专家康士坦丁诺夫在全苏农业科学院会议报告中说：“春化种子处理小麦产量是每公顷 960 千克，对照组的产量是 956 千克。4 千克的差别在统计上是没有意义的。”李森科对此十分恼怒，威胁要把康士坦丁诺夫“清除出去”。除了要把康士坦丁诺夫的“错误资料”清除出去以外，还要把“坚持保留这些资料的人清除出去”。

到了 20 世纪 50 年代，春化种子处理工作再也搞不下去了。李森科主义者的解释是：“高度技术性的农业装备已经能够在很短时间内播种完毕，不必再做春化种子处理了。”

1961 年 8 月，李森科在遗传研究所所长就职演说中提到，他的“用泽西种奶牛同地方品种杂交提高乳脂含量 6% ～7% 的试验”，来源于他的“生物学物种生存定律”——物种里的较弱个体会为了全体的

利益而自愿死去，即自我稀疏。这种“自愿死去”的“定律”，自然遭到生物学家的唾弃。1964 年，支持李森科的赫鲁晓夫下台，李森科主义也受到批判。11 月，科学院和媒体相继报道了用李森科方法提高乳脂含量的问题。其中最严厉的批评是哥罗定斯基的文章《事实胜于捏造》，其中提到：根据会计的报告，李森科在实验报告中的关于乳脂含量的数据至少夸大了 29% ～45% ，与 1954 年相比，每头牛的牛奶产量下降了 2660 千克。

李森科的伪科学并不是唯一的例子。1949 年波希扬宣布“病毒起源于细菌，反之，细菌也起源于病毒”。李森科是这一理论的支持者，封之为“米丘林主义者”。最后，由著名科学家组成的 18 个评审委员会揭露了这场骗局，剥夺了波希扬的博士学位。李森科竭力支持勒柏辛斯卡娅关于“细胞起源于‘活质’”的理论。勒柏辛斯卡娅宣布在干草叶浸泡的肉汤里自发产生了纤毛虫，卵的蛋白形成了细胞，蛋黄形成了血管等奇谈怪论。当然，勒柏辛斯卡娅的“理论”，终于成为过眼烟云，落得个欺世盗名的下场。

1965 年 2 月，李森科被解除苏联科学院遗传研究所所长的职务。1976 年 11 月 21 日，这个曾经显赫一时的“学阀”悄无声息地死去。

被李森科迫害至死的 N. I. 瓦维洛夫曾写道：

我们将走向火葬场
我们将被火化，但我们
决不放弃我们的信念

科学的信念是“真实”。一切科学欺诈的下场，

都会和李森科一样。

3.4 媒体信息中的数据欺诈和滥用

在现代信息社会中，社会竞争日趋激烈。由于媒体传播的强大社会效能，商品质量的数据化，科学数据日渐渗入人们的日常生活。这一切都表明，在未来世纪，数据的作用会更加强大。与此同时，由于数据的利用涉及太多的利益，数据欺诈也就会不择手段地出现。数据欺诈不仅仅在科学研究中会发生。“用数据说谎”会出现在宣扬政绩的报表中，推销产品的广告里，乃至一些貌似科学的文字中。当然，有一些并非恶意欺诈，只是不恰当地滥用数据而已。

对付数据欺诈的利器当然是法律。反对不正当竞争的法律当然会包括惩罚“数据欺诈”的内容，而且会随着数据欺诈的严重性和多变性，不断地进行修改和完善。另一方面，科学道德、商业道德、职业道德旗帜的扬张，人们的社会生活更加有序，也会使得数据欺诈难售其奸。但是，既然是欺诈，它总带着美丽的假面具。如果人们提高了对虚假数据的认识能力，善于揭穿数据欺诈的鬼把戏，那就会比法律和道德更有意义。没有人上当了，数据欺诈也就不存在了。

因此，提高社会公众的数据识别能力，加强教育措施，则是根本的。我们在这里揭供美国的一些案

例作为借鉴。[①] 应当说，以下的数据欺诈和数据滥用问题，在当前我国广告、文章、生活中都有不同程度地出现，值得我们严重注视。

3.4.1 取样的样本有误

* 100 年前有一则报道说："美国某大学的女生有 33% 和该校的教师结婚"。因此男女合校是不好的。但是，那年该大学只招收了 3 名女生，其中有一名和一位教师结婚。原来如此：样本太小了。

* 1949 年，纽约《太阳》杂志报道："1924 年耶鲁大学的毕业生一般年收入为 25111 美元"。这一数字来源于大学对校友的一次问卷调查。著名的《时代》杂志评论说，25 年之后，和母校保持联系的多半是那些事业有成的毕业生。许多人早已不知去向，或者是收入平平甚至穷困潦倒不愿和母校进行联系。因此，25111 美元的数字明显地被夸大了。

* 一则广告说："据用户报告，使用都克斯牙膏可以使蛀牙减少 23%。"这是一个十分模糊的说法。那么厂家是否信口开河？不。他们拿出一份调查报告备查。可是，报告经不起细究。首先，样本有多大？据厂家出示的调查报告，用户只有 12 户。再看 23% 从何而来，原来，这 12 户只是许多小组中的一个小组。其他小组的蛀牙数或者没有减少，或者减少数很小（只有 2%），只有这一组减少 23%，所以就选用了这一组。

---

① Darrell Huff，《怎能利用统计撒谎》，中国统计出版社，1989 年。

＊有一位精神病学者曾在一份报告中说，事实上每个人都有点精神病。可是这位学者所说的“人”的样本，居然是他所接触的病人群。这群人怎么能作为“人”的样本呢？

＊1936 年，美国《文学文摘》杂志在预测“总统选举”中惨败。事因是杂志根据 1000 万电话用户和从该杂志订户所收回的意见，断定 A. 兰登将以 370∶161 的优势击败 F. 罗斯福。结果，F. 罗斯福当选，《文学文摘》大丢其脸。原因何在？事实上，1936 年时，能装电话或订阅《文学文摘》杂志的人，在经济上都相对富裕，并有大量的共和党人。这些人选举兰登是可以理解的。然而，收人不太高的大多数选民选择的是 F. 罗斯福。

＊有位调查员说，我喜欢在火车站作调查，因为在那里可以遇到各种各样的人。可是，在火车站很少能见到带婴儿的妈妈。你如果在火车站调查有关“奶粉”的事情，结论就会出偏差。

＊广告上称：“在许许多多著名医生的大样本中，抽 XX 牌香烟的占 27%——比其他牌号的香烟多。”这个数字的真实性暂且不谈。即使它是真的，那又怎样呢？医生喜好难道就一定正确吗？他们有特别的消息知道烟草公司的内部消息吗？显然没有。登这样的广告当然只是一种商品诱导。

3.4.2 选择的代表数有误

我们曾经谈到过数据的代表数有很多种：算术平均数，中位数，众数。如果不注意观察分析，那是容易上当的。

* 一年前，某房产商向顾客介绍说某地区每户居民的年收入是 15000 美元。这在 20 世纪 30 年代是相当高的。确实，那里居住着几家大富豪。这个数字是当地居民年收入的算术平均数。一年后，房产商向地方当局申请降低税率，报告中说的是每户居民收入是 3500 美元。这是因为只有一半家庭的收入高于 3500 元，其余一半则低于 3500 元。这一中位数对确定征税面有很大关系。可是，当一位零售商打算在该地区开超市时，房产商向他提供的数据是每户居民的年平均收入是 2000 美元。因为这里的工人家庭很多，以 2000 美元年收入居民最多。这一众数是统计和分析购买力的主要数据。

*《时代周刊》有一篇“出版者的话”。它对新订户介绍说：“本杂志订户年龄的中位数是 34 岁，他们家庭年收入是 7270 元。”又一年的介绍则是：“本杂志订户年龄的中位数是 42 岁，他们家庭年收入是 9535 元。”他们对年龄指出是中位数，而收入数字则不指明。这位出版者是否用平均数来炫耀他们拥有富裕的读者呢？

3.4.3 用模糊的字眼误导

* 一篇实验报告，用大字在报纸上公布：“用 X 药 XX 克便能在 11 秒内杀死试管里的 31108 个细菌。”挑选一个奇特的实验室作背景，配一穿白大褂的医务人员的照片做陪衬。于是一份药品广告就完成了。但是，这种药在试管里有效，在人的喉咙里依然有效吗？药经过稀释之后，还能杀死细菌吗？对这种不负责任的宣传，不信为好。

＊美国《星期日》周报上刊有一份“盖塞标准”，其中提到“一个婴儿到第 X 个月就能坐直”。许多父母看到这则消息，马上联想到自己的孩子。如果他们的孩子到这个月份还坐不直，就会怀疑孩子“弱智”、“软骨”、“发育不正常”等等。这个标准是什么意思呢？据了解，这是孩子出生到能坐直时间的“中位数”。一般的孩子在 X 月时一定是坐不直的。没有什么可担心的。“标准”一词，意味着达不到此数据就不合格，可是中位数是不能作为标准的。

＊1952 年美国《柯里尔》杂志刊登两张图表，并说：“要知道你的孩子成年时的身高，根据现在的身高查一查图表就行了。”这两张图表是千真万确的。错误的是那句说明词。经检查，两张图表制作时所使用的样本很大，抽样很科学，数据也很准确。然而，调查是根据大样本的平均身高为依据的。要知道，孩子的生长规律并不一样，有的早早长高，有的很晚才“窜”起来。所以这张表只能用于一群人（例如 100 个孩子）平均身高的预测，而不可以对“单个”孩子作身高预测。这次误导的结果是造成许多人的烦恼和失望。

＊一篇文章为“书价太高”作辩解：书价是由成本上升而造成的。10 年来，原材料上涨了 9%，印刷费增加了 12%，推销广告费用上升了 10%，这样，公司成本就增加了 31%，书价怎能不涨？天知道，如果每个项目都增加 10%，那么总成本也只增加 10%，而不是 30%，这种数据欺诈是利用人们的数学常识不够。

3.4.4 忽视数据的随机误差

采取数据的时候，不可避免地要产生误差。这种误差是随机因素产生的，不表示实质意义。例如，在某班级的一次考试中，99 分和 98 分根本没有实质性的能力差别。其差异是考题选择，学生临考时的身体状况，学生回答时的心理表现，乃至气候、考卷清晰度等等都有关系。些微的差距在统计学上是没有意义的。但是，在现实生活中，因一分之差而落榜，两分之优受父母重奖。更有甚者，还因几分之差受奚落而走绝路的，每年都有。我们在考试管理时，“从高分到低分”的录取办法也许是不得已而为之，但是，考试部门、学校领导、父母家长可否不要强化那些没有统计意义的差别，给孩子们以比较宽松的环境呢？能否正确对待这一随机误差，当是社会数学素质强弱的标志之一。

统计上常常用标准差来表示，例如，考查一个学生的“智商”，得到的数据应该是 98 ± 5，或者 101 ± 4。在这个范围内的差异是正常的，没有本质的差别。让我们来看一些例子。

*20 世纪 50 年代，著名的《读者文摘》杂志组织了一些人将不同牌号的香烟进行随机抽查，检验其尼古丁含量。最后的结果是各种牌号的香烟所含的尼古丁含量只有非常微小的差别，在统计上看是相同的。杂志社刊登了各牌号的尼古丁含量数据。结论是“香烟有害，都不要抽”。

但是，这些牌号中所含的尼古丁含量毕竟有微小的差异。排在最后一位的“老金牌”香烟，抓住机

会大做文章，一时电报满天飞，报纸大做宣传，说“老金牌”香烟的尼古丁含量最小，大捞商机。后来，生产“老金牌”香烟厂家被禁止用此数据做广告，但是广告效应已经存在。“不正当”竞争已成事实。

*某杂志进行读者调查，发现有40%的读者喜欢A类文章，35%的读者喜欢B类文章。于是编辑们约写了大量A类文章，结果喜欢B类文章的读者纷纷退订，遭受了一次重创。其实，从杂志的问卷调查之类的结果来看，读者反应的随意性很大，误差是不可忽视的。40%和35%并没有实质性的差别。

3.4.5 使用不同的比较基数

*一则广告中说：“X牌号榨汁机能多榨出26%的果汁”，“它已经由实验室所证实，且由某著名家政学院作担保”。这听起来很诱人。一只苹果可多榨出26%的果汁，当然是一件好事。不过，它没说参照物，比什么多26%呢？仔细一看，原来是比老式的手摇机多26%而已。这和你原来的想象差之千里，也许这是现代同类产品中最差的。这26%除了显示一种精确的科学数据之外，毫无意义。

*一篇文章谈到，如果早上以每小时70英里的速度驾车，那么你活下来的机会是晚上7时的4倍。论据是晚上车祸的比例是早上的4倍。这是毫无意义的论证。晚上车多，自然出事的机会多，基数不一样，怎能比较？照这种推理，我们可以说，晴天驾车比雾天危险，因为晴天的车祸比雾天多。其实，晴天的汽车数会比雾天的汽车多许多倍。你也可以说，现在乘飞机比20世纪20年代更危险，因为现在的

飞机失事比20世纪20年代多。但是，现在的飞机乘客比20世纪20年代多千万倍。

＊一份报告说，在美国和西班牙战争时，美国海军的死亡比例是千分之九，而纽约市民的死亡率则是千分之十六。结论是参加海军比较安全。这是毫无意义的比较。海军中都是健康的青壮年，纽约市民则包括病人、老人等等。

3.4.6 图表有意诱导

同样的数据，用不同的图表加以表示，可以产生一种有意的诱导。

＊某公司两个车间，生产相同产品，条件也类似。结果也相同，都从去年的1000件增至今年的2000件。可是，两车间主任送来的图却不一样。图3.1是正常的，图3.2则做了手脚。香港的中学数学教材中使用标题是“独具匠心还是别有用心?”十分传神。

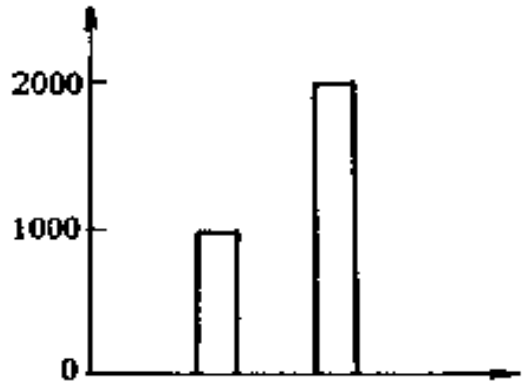


图3.1

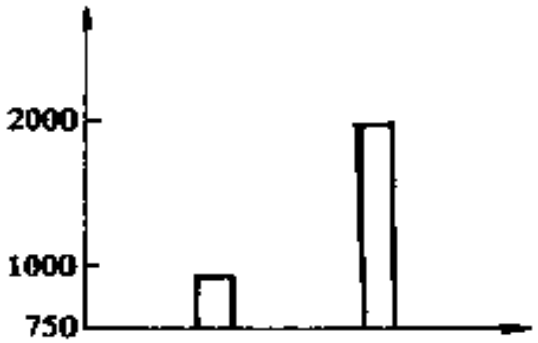


图3.2

总之，用一句俗语说：“成也数据，数也数据”。数据给我们带来科学、信息、愉悦。但也会被滥用，带来虚假、欺诈、烦恼。愿我们多多学习，提高数据意识，迎接未来世纪的数据时代。

4 数学与社会:民主投票

投票是民主政治的表现之一,可是很少有人知道数学和投票有着非常密切的联系。候选名额的分配,选举者选票的投法,判断当选的规定,党派联合的多数组成等等,一系列的问题都涉及数据的处理。因此,必要的数学知识已成为民主意识的组成部分。

4.1 不同的选举方案

民主选举的最基本的原则是"多数选举",即获得多数票的候选人当选。不过,投票的程序可以有很多种。假定候选人是 5 名,只有 1 人当选,常见的方案有:

(A) 简单多数票。

(B) 若无人过半数,则在两位获票最多的候选人之间增加一场决定性竞选。

(C) 选举人可以对 5 名候选人中的一名投赞成票(+1),对另一位不喜欢的人投反对票(-1),最后以得票的代数和最高者当选(支持票数减去反对票数)。

(D) 赞成选举法,即可以对 5 名候选人中喜欢者都投票,选举者可以投 1～5 票,最高得票的一人

当选。

(E) 排序方案。选举人可以将 5 名候选人排队,即对 5 人中最满意的投 5 票,其次的投 4 票,再次的投 3 票,排第 4 位的只投 2 票,最不满意的投 1 票,不得弃权。

让我们来评论这 5 种方案。方案(A)最简单易行;缺点是当选者的得票数未必过半,也许是得票数第 2 的人因为和得票数第 3 的人票数分散而落选。方案(B)较(A)为好,保证最后的当选者得票数过半,只是要选 2 次,比较麻烦。方案(C)允许投反对票,可以充分表达选举人的意愿,但是可能是缺点和优点都不突出的"中间人物"当选。方案(D)为赞成投票法,能够充分表达选民的倾向,使得当选者能够获得较多的选民拥护,不会出现提出相似竞选纲领的多名候选人选票分散的情形;缺点是实行比较困难,因为选举人有 5 票投票权,如何分配给这 5 个人有多种选择,没有投票经验的人往往会乱投一起,反而失去真实。方案(E)可能出现非常复杂的情况:人们共有 $5\times4=20$ 种不同的选择,容易失控。

例 假设某团体打算出外旅游,有 3 种选择:西湖,太湖,泰山。请 21 位成员投票决定。

可能有以下情况出现:

(1) 按方案(A),以简单多数确定。此时的票数为:西湖 5 票,太湖 6 票,泰山 10 票。结果泰山以 10 票中选。

(2) 按方案(B),令太湖和泰山对决。结果支持西湖的选票全部转向太湖,共 11 票,超过泰山的 10 票。于是太湖中选。

(3) 按方案(C),大家对去西湖不反对。有 5 票赞成去西湖,其中有 1 人反对去泰山。但想到太湖的 6 人投太湖 6 张赞成票之外,因非常不愿爬山,对泰山投了 6 张反对票。而在愿意去泰山的 10 票中,对太湖投了 7 张反对票。这样,到太湖和泰山的人互相投了反对票,结果是:

西湖　5－0＝5,

太湖　6－7＝－1,

泰山　10－6－1＝3 票。

于是,西湖中选,太湖竟然得了负票,不受欢迎。

(4) 按方案(D),每人可投 1～3 张选票。

原投而湖的 5 人,也喜欢太湖,但不愿支持爬泰山,这组人的票数分布为 5,5,0。

原投太湖的 6 人中,只投自己喜欢的太湖,不投其他的票。票数分布为 0,6,0。

原喜欢泰山的 10 人中,7 人反对去太湖,有 4 人不反对游西湖,甚至也赞成。这组人的票数分布为 4,3,10。

现在的结果是:

西湖 5＋0＋4＝9,

太湖 5＋6＋3＝14,

泰山 0＋0＋10＝10。

结果在方案(3)中最不受欢迎的太湖以 14 票中选。

(5) 按方案(E),可能出现的情况是:

5 人选择　　西湖＞太湖＞泰山

6 人选择　　太湖＞而湖＞泰山

10 人选择　　泰山＞西湖＞太湖

(总共有 6 种排列,为简单起见,只列出三种选

择的情形)

于是,西湖票数　$3\times5+2\times6+10\times2=47$,

太湖票数　$2\times5+3\times6+10\times1=38$,

泰山票数　$1\times5+1\times6+10\times3=41$。

结果是在简单多数时最不受欢迎的西湖中选。

我们不妨做些试验,看看这五种选举方法的选举结果是否都相同。例如,在巩俐、潘虹、姜文、葛优、陈佩斯等 5 人中选出一人为最佳影星。

4.2　代表的名额分配

选举课题的涉及面很广,简单的如名额分配。美国宪法指出:"众议院议员的名额……将根据各州的人口比例分配。"但是这么一句话,实行起来却问题不少。20 多年来,关于"如何公正合理"的争论,一直没有平息。

名额分配的数学的描述方法是:

设众议院议员数为 N,共有 S 个州,各州的人口数为 P_i, $i=1,\cdots,S$。问题是找一组数 $N_1, N_2, \cdots, N_n$,使得 $N_1+N_2+\cdots+N_n=N$,其中 N_i 表示第 i 州的议员数,并要求尽可能满足人口比例份额 Q_i,即尽量使得 N_i 和 Q_i 接近。这里,

$$Q_i = NU = N[P_i/(P_1+P_2+\cdots+P_n)]。$$

最简单的方案是四舍五入法,即 Q_i 不是整数时,小数部分按四舍五入法调整。但此时可能出现因四舍去的州多,导致名额空余,或者因五而入的州多,导致名额不够。

1791 年,当时的美国财政部长 A. 哈密顿(Hamilton)

提出的方案如下：

（1）先取各州人口份额 Q_i 的整数部分$[Q_i]$。

（2）按 $R_i = Q - [Q_i]$的大小排列，将剩余的名额由 R_i 的大小顺序分配，分完为止。

例 设学生会代表会议由各系学生数按比例分配代表。现有甲、乙、丙三系。代表数目为 20。甲系学生数为 100，乙系为 60，丙系为 40。此时，

$$N = Q = 10, N = Q = 6, N = Q = 4。$$

分配成功。后来，丙系各有 3 名学生分别转到甲、乙两系学习。于是：

$$P = 103, P = 63, P = 34。$$

$$U = 51.5, U = 31.5, U = 17。$$

$$Q = 10.3, Q = 6.3, Q = 3.4。$$

$$R = 0.3, R = 0.3, R = 0.4。$$

按四舍五入法，将多出一个名额。按哈密顿方法，以 R 为大，即将剩余的一个名额给丙系。最后的代表数仍是 10，6，4。虽然丙系学生数减少了 6 个，代表数没有变化。甲、乙两系稍微有些怨言，但还是接受了。

后来，情况发生了意想不到的变化。由于其他原因，代表总数增加到 21 名。这时按哈密顿方法处理有如下结果：

系别	学生数	人数比例	按比例代表数	最终代表数
甲系	103	51.5	10.815	11
乙系	63	31.5	6.615	7
丙系	34	17	3.570	3

这一结果自然引起丙系学生大哗。代表名额总数增加一个，丙系名额反而减少一个，未免说不过

去。

如果只是为学生会的代表数目发生争论，适当调整也就算了。但是在美国众议员数目分配上也出现此类状况，当然会引起轩然大波。1880 年，美国阿拉巴马州对哈密顿方法提出异议，因为当议员总数增加时，该州议员席位反而减少。

另外，当某州的人口增加率比其他州高时，议员数目也可能反而减少。这使得哈密顿方法不得不放弃。为了说明问题，请看以下数据。原来按哈密顿方法丙州可以用尾数 0.375 较乙州的 0.365 和甲州的 0.26 大，获得一个名额。但当人口增加时，丙州的人口增加率最高，达到 20%，却因尾数 0.41 恰小于乙州的 0.42，乙州获得一个剩余名额，丙州却连一个代表也没有了。

州别	人口数	比例分配数	实际名额	人口增长率	新人口数	新比例分配数	新实际分配数
甲	420	1.26	1	2.38	430	1.17	1
乙	455	1.365	1	14.39	520	1.42	2
丙	125	0.375	1	20	150	0.41	0
总和	1000	3	3		1100	3	3

这个例子说明了哈密顿方法存在着“人口悖论”。哈密顿方法于 1791 年提出后，为美国总统华盛顿所否定，但在 1851 年以美国国会议员 S.F. 文东(Vinton)的名字为国会所采用。鉴于后来出现的悖论，于 1910 年被废止。1941 年起至今，美国国会分配名额采用的是亨丁顿方法。E.V. 亨丁顿(Huntington)是哈佛大学教授，20 世纪 20 年代提出

用不公平度概念来处理。但这一方法比较复杂。①

不过，亨丁顿方法也有不足之处，并非十全十美。到了 1974 年，两位美国人 L. 巴林斯基(Balinsky)和 H.P. 杨(Young)提出了名额分配公理：

公理 1. 人口增加不会失去席位。

公理 2. 平均说，每州应得到自己应分摊的名额。

公理 3. 总名额增加不会使某州名额减少。

公理 4. 最后名额不会偏离比例份额。

公理 5. 从一个州到另一州的名额转让，不会使得这两个州都接近它们应得的份额。

1982 年，巴林斯基和杨证明：不存在一个方案能满足以上五条公理。这样，争论也就不会有结果了。亨丁顿的分配名额的相等比例方案尽管还有毛病，相比之下，仍较为合理，所以现在还在使用。

4.3 投票实力分析

西方国家实行多党制，在议会里各政党都有不同数额的席位。各政党未必可以单独取得多数而上台执政。因此，在没有一个政党获得绝对多数的情况下，几个党拼成多数联合执政的情况屡见不鲜。因此如何估计各党的投票实力，是一件颇为实用的事情。我国在政治上不搞多党制。但是在经济领域

① 卢卡斯，《政治及有关模型》(中译本)，国防科技出版社，1996，第 14 章。

内，股票持有者的份额为多数时，将决定企业的发展前途，所以也是需要借鉴研究的。

让我们看一个实例。1972 年，加拿大选举行时的形势是：

自由党	109 席
保守党	107 席
新民主党	31 席
其　他	17 席
总　计	264 席

超过 132 席为多数。这时，只拥有 31 席的新民主党，与拥有 100 席以上的两个大党处于同样有利的地位。而“其他”的 17 席，在构成多数的联合中，有你不多，无你不少，处于无足轻重的位置。因此，一个政党在国会中拥有的席位数目当然重要，但是各党派席位形成格局同样需要研究。

设有 n 个政党，w_i 为第 i 个政党在国会中拥有的席位。全部席位数是

$$W = \sum_{i=1}^{n} w_i。$$

用 S 表示 $N=1,2,\cdots,n$ 的一个子集，借助 S 描述若干个政党的联合。Q 是一个定额，通常是 $W/2$ 或 $2W/3$，意为半数，2/3 多数，表示执政党所需的最低票数。现在用$[Q;w_1,w_2,\cdots,w_n]$表示 n 个政党分别具有 $w_1,w_2,\cdots,w_n$ 的一次竞争。（政党理解为股东，国会理解为公司董事会，也是一样的）

若有集合 S，使得

$$\sum_{i\in S} w_i > Q,$$

则由 S 中各党派的联合可以执政。在股份公司的

决策上，几个股东若联合持有公司51%的份额，则有最大的决策权。例如：有 A、B、C、D 四股东，投票的格局为：

	Q；A，B，C，D
模型Ⅰ	[51；28，24，24，24]
模型Ⅱ	[51；26，26，26，22]
模型Ⅲ	[51；40，25，20，15]

在模型Ⅰ中，A 的地位特别有利，B、C、D 处于同等地位。模型Ⅱ中，A、B、C 之间有两者联合即可持有多数，彼此同等地位，D 只少4票，却无足轻重，肯定受冷落。模型Ⅲ中，A 仍据优势地位，B、C、D 三派虽然票数有多少，但作用却是一样的。

有些格局，表面上不同，其实是等价的。如[3；2,2,1]，[8；7,5,3]，[51；49,48,3]，[2；1,1,1]表示的格局没有本质区别。[51；49,48,3]中，3票的小股东和49票的大股东，投票决策时具有同等的效力。

上面只是一种直观的观察。观在我们设法给出衡量某投票集团实力的数量化指标。

采用如上的记号。再用 P 表示 N 的所有可赢子集，即

$$P = \{S \subset N; \sum_{i \in S} w_i > Q\}。$$

用 M 表示最小可赢联合（这一联合中再少任何一派便赢不到决定权）：

$$M = \{S \subset N; \sum_{i \in S} w_i > Q,$$

$$且 \sum_{S \setminus i \in \{j\}} w_i < Q, 对任何\ j \in S\}。$$

对第 i 个竞争者(政党,股东,集团),定义

$$M(i) = |S \subset M; i \in S|$$

用以表示 i 竞争者能够参与最小可赢联合全体,即找出在哪些可赢联合中少了 i 不行。于是我们定义 i 的实力

$$L(i) = (1/|M|)(\sum_{S \in M(i)} 1/|S|)。$$

这里 $|M|$ 表示所有可赢联合的数目,$|S|$ 表示 S 中元素个数。它反映了在所有可赢联合中,有多少含 i 的最小可赢联合,而且考虑到在含 i 的最小可赢联合中是几家联合而成。

例如,有 A、B、C、D、E 五派竞争,格局为[5;4,2,1,1,1]则

$M = AB,\ AC,\ AD,\ AE,\ BCDE;$

$M(A) = AB,\ AC,\ AD,\ AE;$

$L(A) = (1/5)(1/2 + 1/2 + 1/2 + 1/2) = 2/5$

$= 8/20;$

$M(B) = AB,\ BCDE;$

$L(B) = (1/5)(1/2 + 1/4) = 3/20。$

同理 $L(C) = L(D) = L(E) = 3/20。$

5 数据与社会科学

在自然科学的发展史上，实验数据是最有说服力的证据。重力加速度 g、光速 c、分子量、地球半径、天文距离等等，何处没有数据？但是，长久以来，社会科学研究中，却很少运用数据。社会科学的数量化进程，步伐缓慢而沉重。许多社会科学的专家，对于用数学方法得来的结果，采取半信半疑的态度。这是否和中国传统文化中缺乏"数学论证"有关？初初想来，却又不是。看那训诂考证之学，举证是何等翔实，立论又是处处严密，颇有"严密逻辑"之精神。当今中国的数学教育偏重操练，力求严格，缺乏想象，倒是这种考据之学的遗风。

然而，中国社会科学界借助数学方法的不足，却也是不争的事实。请看：形式逻辑难于上升为数理逻辑，一般统计学不能发展为数理统计学，一般经济学看不惯数理经济学。计量历史学、数学心理学、数理语言学、数理政治学散在各处不成气候。数学和社会科学的结合，似乎还停留在马克思写资本论时代的水平。如果有超越，也只是个别情形，整体上似乎还没有大的突破。

让我们来看一些个别的例子。

5.1 《红楼梦》作者判断中的数据分析

关于《红楼梦》一书的作者，研究者颇多，意见也各异。但自胡适于1921年发表《红楼梦考证》以来，断言前80回为曹雪芹所著，后40回则为高鹗续成。这种意见广为流传。晚近，人民文学出版社的1982年本的署名仍是：曹雪芹、高鹗著。这种意见对不对呢？数理语言学似乎可以帮一点忙。

用数学方法判断一项文学作品的作者，在国外早有先例。当年，苏联萧洛霍夫是否创作了《静静的顿河》，也曾经闹得满城风雨。最后据说是用了统计方法，才确认了萧洛霍夫的作者地位。

1987年，复旦大学数学系的数理统计专家李贤平，发表"《红楼梦》成书新说"的论文。① 作者用现代数学方法，以电子计算机为工具，对《红楼梦》的语言作了统计分析，获得了大量的新发现。他的做法是：

1．将120回看做一个整体，不再先验地分成前80回和后40回，由平等对待的客观态度，用统计数据来下判断。

2．从统计语言学的角度建立识别特征。主要用47个虚字的出现频率，有时也用到句长分布。47个虚字是：

（1）13个文言虚字：之，其，或，亦，方，于，即，皆，因，仍，故，尚，乃。

---

① 《复旦大学学报》社科版，1987年第5期。

(2) 9 个句尾虚字：呀，吗，咧，罢咧，啊，罢，罢了，么，呢。

(3) 13 个常用白话虚字：了，的，着，一，不，把，让，向，往，是，在，别，好。

(4) 10 个表示转折、程度、比较等意思的虚字：可，便，就，但，越，再，更，比，很，偏。

(5) 后缀词 2 种：儿(用于名词)，儿(用于副词和形容词)。

3. 用各种统计方法(主成分分析，典型相关分析，多维尺度法，广义线性模型，类 Chi 方距离，相关分析)，探索各回的写作风格接近度，并用三种层次聚类方法对各回目进行分类。

李贤平的研究使用了上海纺织工业学校陈大康先生用两年半时间精心统计而得到的字频数据。这是一个十分基础的工作。以上 47 个虚字的出现频率，乃是统计分析的出发点。

然后，李贤平使用美国威斯康星大学麦迪逊分校的计算机进行统计分析，获得约 300 张图表。这里附有一张 chi 方距离多维点云的正视图，如图 5.1 所示。图中的号码表示回目，从 1 到 120。每个点(回目)的位置由该回文字的 47 个虚字的频率所决定，各点之间的距离就是 Chi 方距离。

由图可以看出各回目之间的关系。如果从左下角到右上角画一对角线，则除了 67 回以外，1～80 回全在对角线之上，后 40 回在对角线之下。这一现象和过去红学家的判断是一样的。说明前 80 回和后 40 回是两种手笔。这个貌似平凡的结论，反衬了数学方法的有效性，也说明这 47 个虚字是有区别能

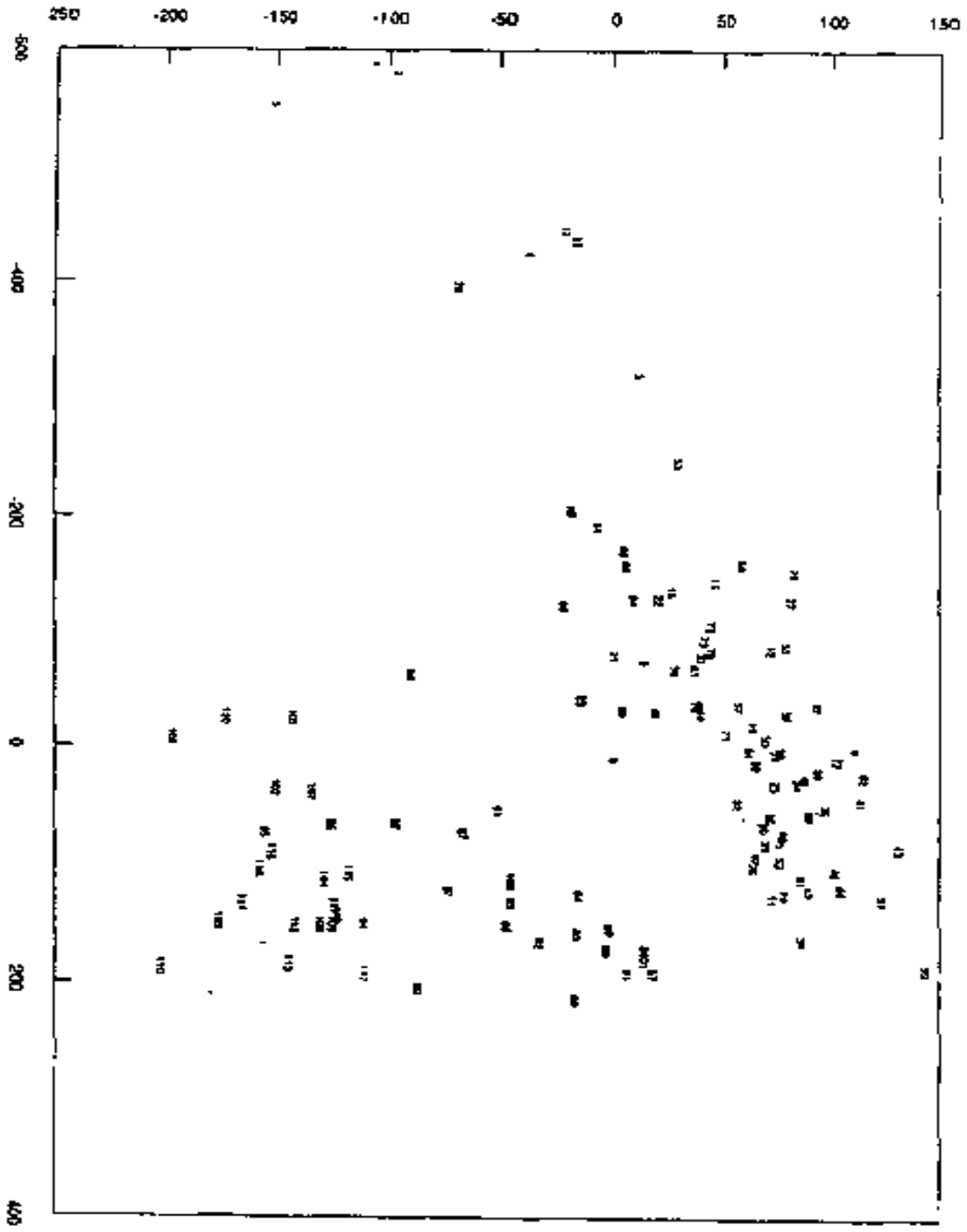


图 5.1　正视图(以 47 个虚字为识别特征)

力的。图中的这 120 个点能够区分开前 80 回和后 40 回，反映了《红楼梦》的特征，不是偶然碰上的。要凑得这么好，其概率是一万亿亿亿亿分之一，事实上不可能发生。但是，更为重要的是，有许多特征为过去红学家们从来没有注意到。比如：

* 第 1、2、3、4、5 回，第 17、18 回，第 78 回，在图的左边，其他各回多数集中在右边。

* 另有 9，12～16，21～23，58，63～69(除 67)，74～80(除 78)处于中间过渡地带。

* 后 40 回大体可分为两类，前 20 回偏上，后 20 回偏下。

李贤平先生在作了大量数据分析之后，得出他自己关于《红楼梦》的新见解(其中包括作者的判断)：

* 不能笼统认为，前 80 回为一人所写，后 40 回为另一毫不相关的人所写。因为彼此之间有交叉。

* 前 80 回是曹雪芹据《石头记》写成，中间插入《风月宝鉴》，还有一些深刻的增加成分。80 回在图中的位置，可分为三部分：右边几回叙述宝黛爱情，中间几回则是取材于《风月宝鉴》的部分；左边的几回则为“梦幻”或涉贾府之外的社会事件。

* 后 40 回是曹雪芹亲友将曹雪芹的各种草稿经过相当长的时间整理而成。后 40 回，前半是宝黛爱情的继续，与前 80 回的宝黛爱情故事似为一人所写。在图中的距离相近。后半则是贾府衰败情景，在图中聚在右下角，当为另外一人所写。

李贤平先生的论文得出的结果，远不止这些。我们在这里引述一小部分，无非是说明数学方法的

价值。它印证了红学家们的科学研究成果，同时又提供了新的视角，应该是很有说服力的。当然，数学描述只是一种参考，47 个虚字作为指标，也是一种尝试，历史和文学不能简化为“虚字”统计。但是，数学的统计毕竟是客观存在，其结果是人人可以重复的。从一个特定的角度看问题，自有其特殊的价值。李贤平教授在研究工作中得到周汝昌、赵冈、周策纵等红学家的支持，表明红学家是欢迎数学家介入红学研究的。

在数学方法日益进步、计算机科学手段大显身手的信息时代，研究方法总要有所前进，有所突破。21 世纪即将到来，我们预言数学方法在文学研究中会有更大的作为，大概不至于过分罢。

5.2 数据与历史：计量历史学

从 20 世纪 50 年代以来，国外发展的“计量历史学”，获得了大量进展，但在中国却一直不能登堂入室，处于“受冷落”的地位。这种情况，应该是有所改变的时候了。

历史学的研究，历来集中于政治领域，于是，历史研究方法多半会使用定性的政治分析，往往忽视甚至排斥定量的分析。然而，政治是经济的集中反映。对于古代社会的户籍、土地、雇佣、生产、贸易等经济发展的定量研究，能够更好地揭示阶级斗争、党派利益、中央与地方关系、文化艺术价值取向等深层次的问题。比如，我国历代的农民起义，都有深刻的经济背景。过去的研究也运用一些阶级剥削的史

料，但总是不系统。一些形容词如“强烈的”、“迅猛的”、“缓慢的”等评价，总使人感到有些含糊。如果有定量的分析，自然会更有说服力了。

计量历史学的基本方法仍然是数据处理。从史料中收集样本数据，推测社会总体情况，分析社会财富的分布形态。研究某些量的相关程度，判断历史上的因果关系是否成立。构建数学动态模型，观察是否合乎历史进程。这是一门单独的学问，在此无法作全面的介绍。下面是一些例子。

计量历史学的早期工作，多在人口统计史，经济史等方面。一些典型的工作有：“英国收入不平等的长期变化”，“《末日裁判书》中教区的土地面积”，“1860 年以来英国的工资和收入”，“对 19 世纪工业扩张的一个解释模型：在美国生铁业中的应用”，“1840～1870 年英国铁路和经济增长的关系”，“从欧洲中世纪教区登记簿研究家庭和婚姻状况”。这类工作，中国有十分优越的条件。正统的 25 史中大多有专门的记载——“食货志”，其中详细记录了税赋、田亩、人口以及其他经济情况。在浩如烟海的地方志、历史典籍中，数量化的资料不计其数。这些资料为中国的计量历史学提供了良好的研究条件。

原苏联的 D.V. 迪奥皮克，有一个工作是“用计量方法分析古代东方编年史《春秋》的经验”。他将《春秋》（含《左氏》、《公羊》、《谷梁》三传）分为：①外交；②军事；③帝王生活；④内政；⑤经济；⑥宗教仪式；⑦自然现象，共 7 部分进行数量化研究。例如，战争事件，全书共有 18000 处提到。然后按主客体、时间地点列出资料，分成早期晚期，依 14 个主要国

家及其部族关系形成系统。通过这些战争的数量分析，可以得出许多宏观的认识，为过去局部研究所未见。比如，牢固军事联盟早期为 27 组，晚期为 38 组，两个国家联合行动的次数，早期比晚期多 11 倍。这表明，早期的同盟不牢固，晚期的同盟大多数是牢固的。这反映了大量中小国家混战的旧体制逐步瓦解，出现了比较统一的社会政治综合体，即统一的社会趋势。①

用模糊数学方法来研究历史分期也是有意义的工作。例如，古代中国社会从奴隶制向封建制的过渡是我国史学界长期争论的问题。素有“春秋战国封建制”、“汉代封建制”、“魏晋封建制”等多种观点。其实，从模糊数学的角度看，从奴隶制到封建制是一个长期缓变的过程。封建社会是一个“模糊集合”，其边缘是模糊的，正如年青人是一个模糊集一样。模糊集合都有一个隶属函数。“年青人”集合 Y 的隶属函数 $M(Y)$ 是

$$M(Y)=\begin{cases}1, & 0<Y<25,\\ \{1+[(n-25)/10]\}^{-1}, & Y>25。\end{cases}$$

这表示当一个人的年龄小于 25 岁时，属于“年青人”集合的隶属度是 1。在 25 岁以后，属于“年青人”集合的程度要下降。当 35 岁时，$M(35)=0.50$。当 50 岁时，年青程度为 $M(50)=0.13$。这样来理解“年青人”，比 25 岁是“年青人”，而 26 岁就不是“年青人”似乎要合理得多了。

---

① I.D. 科瓦利琴科，《计量历史学》，四川人民出版社，1987 年。

殷商时期是奴隶制的隶属度是 1，没有问题。魏晋以后是封建制，即隶属“奴隶制”集合的程度为 0，这也没有问题。那么在春秋至魏这几百年间，用一个隶属函数来描写不是很好吗？这当然要从“殉葬制度”、“农奴数目”、“生产水平”等多种因素加以测定。这是一个很复杂的工作，需要很多的历史资料才能完成。

再如，我国从半封建半殖民地社会过渡到社会主义阶段，也不能用某一天作为分界线。早在老解放区，就有“公有制的经济成分”。1949 年 10 月 1 日，中华人民共和国成立，公有制成分还不占优势。1956 年，资本主义工商业公私合营之后，私营经济仍然存在。现在，我们认识到，社会主义初级阶段是一个很长的时期。21 世纪开始时的中国社会其“社会主义”隶属度是多少，应该是一个可以研究的问题。

5.3 人尽其才的数学

实际工作中常常要遇到“最优决策”问题，即在许多可能的选择中，选取最好的方案，使得决策能产生最好的效果。这在工程计算、商业运作、调度安排时，当然十分重要，即使在社会科学中，最优决策也是有用武之地的。这里我们举人事安排的一个例子。

设某工厂来了三名新工人：甲，乙，丙。工厂的空缺岗位有电子维修工、图书管理员、木工三个。工厂为了充分发挥每人的特长，对三人有针对性地举

行了三门课的考试，得分如下：

	1．电子学	2．语文	3．制图	总分
1．甲	8	7	5	20
2．乙	8	6	7	21
3．丙	5	10	7	22

按照通常的依总分从高到低的录用方法，派丙去从事文化要求高的电子维修，派总分第二的乙去做图书管理员，派总分第三的甲去做木模工。但是这种指派方法不够好。因为丙的电子学知识只有5分，乙的语文知识只有6分，甲的制图知识仅为5分。每人工作的对口才能，实际上只用了5＋6＋5＝16分。现在如果改变一下，让甲去做电子工，乙做木模工，丙做图书管理员，则对口技能分可达8＋10＋7＝25分。当然，让乙去做电子工，也能利用电子学的分数8，但这时甲将去做木模工，利用指标仅为5，不如乙做木模工可得7分为好，这一方案的能力利用指标为8＋10＋5＝23。因此，从企业的整体利益看，还是采用能力利用指标达到25分的方案为最好。

以上例题是用直观算得的。一般而言，如果职务数和聘用人的数目都很大时，直接估算将不可能。这时就要借助一般的数学方法——变量只取0，1两值的线性规划。

以本例来说，我们令 $x_{ij}(i,j=1,2,3)$ 表示第 i 个人是否分配做第 j 项工作，$x_{ij}=0$ 表示不这样分配，$x_{ij}=1$ 表示如此分配。于是我们得出如下的一系列关系式

$$\left.\begin{aligned} x_{11}+x_{12}+x_{13}&=1\\ x_{21}+x_{22}+x_{23}&=1\\ x_{31}+x_{32}+x_{33}&=1 \end{aligned}\right\}\text{(1)}$$

（表示每一人分配一件且只一件工作）

$$\left.\begin{aligned} x_{11}+x_{21}+x_{31}&=1\\ x_{12}+x_{22}+x_{32}&=1\\ x_{13}+x_{23}+x_{33}&=1 \end{aligned}\right\}\text{(2)}$$

（表示每一工作分配且只分配给一个人）

x_{ij} 只取 0,1 两个值之一

在上述约束条件下，我们欲使下列目标函数：

$$8x_{11}+7x_{12}+5x_{13}+8x_{21}+6x_{22}+7x_{23}+5x_{31}+10x_{32}+7x_{33}$$

取得极大值。

前已看出当 $x_{11}=x_{23}=x_{32}=1$，$x_{12}=x_{13}=x_{22}=x_{23}=x_{31}=x_{33}=0$ 时，目标函数即能力利用指标取得最大值 25。

一般而言，线性规划的数学提法是：有 n 个变数 $x_1,x_2,\cdots,x_n$，满足 m 个（$m\leqslant n$）线性不等式或等式，而使得某一线性函数 $l(x_1,x_2,\cdots,x_n)$ 取得极大(小)值。

在人事指派的例中，变数为 $x_{11},x_{12},x_{13},x_{21},x_{22},x_{23},x_{31},x_{32},x_{33}$ 共 9 个，约束条件是(1)(2)两组等式，目标函数是能力利用指标，这时是求极大值。

线性规划的详尽论证可查阅专著。为了提高读者的兴趣，这里写下两题，用通常观察法可以解出。

问题一　设某机械加工车间需加工甲、乙两种零件，这两种零件可在三种不同机床（车床、刨床、铣床）上加工。其生产效率如下表：

机床种类	机床数	机床生产效率(件/日)	
		甲	乙
车床	4	15	20
刨床	4	25	30
铣床	2	20	40

试问如何合理安排机床加工任务，使得在甲乙两种产品保持 2:1 的配套比例情况下，成套产品数量达到最大值？

问题二　某工厂生产甲、乙两种产品。已知生产甲产品 1 公斤耗煤 10 吨，电力 5 千瓦，钢材 4 公斤；生产乙产品 1 公斤耗煤 4 吨，电力 4 千瓦，钢材 6 公斤。1 公斤甲产品的利润是 600 元，1 公斤乙产品的利润是 1000 元。按计划，国家给该厂煤 300 吨，电力 200 千瓦，钢材 300 公斤。问如何确定生产计划，即甲乙两产品各生产多少，才能使利润额达到最大？

5.4　聪明人之间斗智的学问：对策论

田忌赛马是一个为人熟知的故事。传说战国时期，齐王有一天要和他的臣子田忌赛马，双方约定：①各出三匹马；②从上、中、下三等级中各出一匹马；③每匹马都参加比赛，且只参加一次；④每次比赛后，由输者付给胜者千金。看来田忌要输了，因为三种不同等级的马，都是齐王的比田忌的强。然而，田忌手下有一个聪明的谋士名叫孙膑，他给田忌出了一个主意。田忌照他的意思办，结果赢了一千金。其中奥妙何在？马还是那几匹马，比赛规则也并未修改，取胜的秘诀在于选取最佳的比赛策略。试看

以下两种赛法

第一种

齐王	田忌
胜[上]	[上]负
胜[中]	[中]负
胜[下]	[下]负

齐王马强三场皆胜，田忌三场皆输，共输三千金

第二种

齐王	田忌
胜[上]	[下]负
负[中]	[上]胜
负[下]	[中]胜

田忌马虽弱，但[上]马可胜[中]马，[中]马可胜[下]马，三场二胜，赢得一千金

可见，与其说是赛马，还不如说是人的斗智。

我们的数学研究如果到此为止，那就太浅薄了。实际上，田忌所以能赢是因为齐王“愚蠢”，齐王老老实实地告诉田忌，第一场出的[上]马，第二场是[中]马，第三场是[下]马，于是田忌可从容地想对策。如果齐王也有很机灵的谋士，针对田忌可能出的“鬼点子”，用新策略去对付，那么田忌也许就占不到便宜了。例如

第三种方案

齐王	田忌
胜[中]	[下]负
胜[上]	[上]负
负[下]	[中]胜

齐王二胜一负，田忌用孙膑老办法，输一千金

第四种方案

齐王	田忌
胜[中]	[中]负
胜[下]	[下]负
胜[上]	[上]负

齐王、田忌各变换策略，可是碰巧成了同等马之间比赛，田忌仍输三千金

由此可见，如果是聪明人之间的斗智，那么情况就会复杂得多。一门新的数学学科——对策论由此诞生。

现在我们比较一般地来提问题。设有甲、乙双方进行某种对抗，甲可出策略 $A_1, A_2, \cdots, A_m$，乙可出策略 $B_1, B_2, \cdots B_n$，而在甲取策略 A_i，乙出策略 B_j 时，甲方的得分为 C_{ij}（C_{ij} 可正可负），乙方的得分为 D_{ij}（D_{ij} 可正可负），而且有 $C_{ij} + D_{ij} = 0$，即甲方得分则乙方失分，反之亦然。这种对策，数学上称为有限两人零和对策。我们来看两个例子。

例 1 田忌赛马，这时齐王和田忌可取的策略各有 6 种，即三匹马的出场先后次序 $3! = 1 \cdot 2 \cdot 3 = 6$，设齐王的策略 α_i 和田忌的策略 β_j 分别是

α_1 和 β_1 均指（上，中，下）　α_2 和 β_2 均指（上，下，中）

α_3 和 β_2 均指（中，上，下）　α_4 和 β_4 均指（中，下，上）

α_5 和 β_2 均指（下，中，上）　α_6 和 β_6 均指（下，上，中）

按赛马规则和马的实力，C_{ij} 和 D_{ij} 可以定出来。例如 α_2 与 β_4 所对应的齐王得分数是 1（千金），即 $C_{24} = 1$。这时的阵势是齐王的上马对田忌中马，齐王下马对田忌下马，这两场齐王赢，另一场齐王的中马对田忌的上马，齐王输。故齐王总计赢一场，得一千金。这也可以用来说明 $D_{24} = -1$。

现在，我们可以写出齐王的得分矩阵和田忌的得分矩阵。

齐王得分 / 田忌策略 / 齐王策略	β_1	β_2	β_3	β_4	β_5	β_6
α_1	3	1	1	1	1	−1
α_2	1	3	1	1	−1	1
α_3	1	−1	3	1	1	1
α_4	−1	1	1	3	1	1
α_5	1	1	−1	1	3	1
α_6	1	1	1	−1	1	3

即齐王的得分矩阵为

$$A=\begin{pmatrix} 3 & 1 & 1 & 1 & 1 & -1 \\ 1 & 3 & 1 & 1 & -1 & 1 \\ 1 & -1 & 3 & 1 & 1 & 1 \\ -1 & 1 & 1 & 3 & 1 & 1 \\ 1 & 1 & -1 & 1 & 3 & 1 \\ 1 & 1 & 1 & -1 & 1 & 3 \end{pmatrix}$$

田忌的得分矩阵可写成(用 $D_{24}=-C_{24}$)

$$B=\begin{pmatrix} -3 & -1 & -1 & -1 & -1 & 1 \\ -1 & -3 & -1 & -1 & 1 & -1 \\ -1 & 1 & -3 & -1 & -1 & -1 \\ 1 & -1 & -1 & -3 & -1 & -1 \\ -1 & -1 & 1 & -1 & -3 & -1 \\ -1 & -1 & -1 & 1 & -1 & -3 \end{pmatrix}$$

例 2 设有一场海上遭遇战，敌我双方对彼此的实力都很清楚，而且各方已设想了几种策略。我方估计在各种情况下能杀伤敌舰的数目为：

杀伤敌舰数 / 我方策略 \ 敌方策略	B_1	B_2	B_3
A_1	5	6	8
A_2	9	7	8
A_3	9	6	10

$$A=\begin{pmatrix}5 & 6 & 8\\ 9 & 7 & 8\\ 9 & 6 & 10\end{pmatrix}$$

由于双方都是“聪明人”，所以这张表格并非为我方秘密，而是双方都预料到的。现在形势是我强敌弱，敌方在不得不打的情况下寻求损失最小的策略，而我方则制订杀伤敌方最多的策略。

现在我们分析矩阵 A。由 A 可见，若敌方出策略 B_3，我们则取策略 A_3，将得最大杀伤数 10。但敌方是聪明的，决不肯出 B_3。敌方当然希望我方取策略 A_1，他出 B_1，损失数取得最小值 5。但显然预料我方不会出 A_1，而出 A_2，A_3。而这时损失数都是 9，又不合算。所以，敌方为求稳妥，必然取 B_2。在敌方取 B_2 的情形下，我方当然取 A_2，此时可达到 7，比取另外两个策略得分 6 要高。因此(A_2, B_2)就是这场对抗中，聪明的双方都能接受的最优策略。

从数学上看，敌方的最佳策略是使各列最大数取最小的那个策略 B_2，我们的最佳策略是使各行最小数取最大的那个策略 A_2（如下表）。这时(A_2, B_2)为最优局势，杀伤值 7 为对策值。

	B_1	B_2	B_3	各行最小数
A_1	5	6	8	5
A_2	9	7	8	$\boxed{7}$
A_3	9	6	10	6
各列最大数	9	$\boxed{7}$	10	(A_2, B_2)最优 $V=7$

现在让我们回到齐王赛马的故事上来。我们能不能套用例 2 中那种求最优策略来求解呢？试试看：

						各行最大
3	1	1	1	1	−1	3
1	3	1	1	−1	1	3
1	−1	3	1	1	1	3
−1	1	1	3	1	1	3
1	1	−1	1	3	1	3
1	1	1	−1	1	3	3
各列最小 −1		−1	−1	−1	−1	−1

这时，各列最小值都是 −1，求最大值仍是 −1，无法确定这 6 个策略中哪一个最优。同样，各行最大都是 3，也无法确定何者为最优策略。因此例 2 的方法不可用。那么我们应如何来解这一问题呢？

我们要用概率的思想来处理。齐王和田忌出马的阵势虽有 6 种，但取哪一种阵势的机会是均等的，即出每种策略的可能性都是 1/6。例如出 A_1 的可能性为 1/6，那么当田忌分别取 $B_1, B_2, \cdots, B_6$ 六种策略时，齐王的可能得分是

$$\frac{1}{6}(C_{11} + C_{12} + C_{13} + C_{14} + C_{15} + C_{16})$$

$$= \frac{1}{6}(3 + 1 + 1 - 1 + 1 + 1) = 1。$$

用矩阵语言来说，矩阵 A 的第一行各元素之和为 6，但因处于第一行 A_1 的可能性为 1/6，故齐王出 A_1 时可能得分数为 6/6＝1。

由于矩阵 A 的各行之和均为 6，故不论齐王出

何种策略，它可能的得分数都是$\frac{1}{6}\times 6=1$。同样，不论田忌出何种策略 B_j，齐王的可能得分数也都是$\frac{1}{6}\times 6=1$(因为 A 的各列中元素之和也是 6)。

总而言之，在齐王与田忌赛马的过程中，如果双方都是聪明的，那么只好随便出一种阵势(因各种机会都是 1/6)，齐王的可能得分数是 1。这里说“可能”是指齐王也许得 3 分，也许得 1 分，也许输 1 分，但平均来说，可能得到的分数是 1，即赢一千金。田忌出各种策略的机会也是均等的，即没有哪一种策略特别占便宜或吃亏，因而也是随便出一种。出了以后，可能输 3 分，输 1 分，但也可能赢 1 分。但平均而言，田忌总要输 1 分，即输一千金。

到现在为止，我们可以归纳如下：

(1) 如齐王宣布自己出场阵势是(上，中，下)，那么田忌如果是笨蛋，也出(上，中，下)，必输三千金。

(2) 如齐王宣布自己出场阵势是(上，中，下)，那么田忌采用孙膑的聪明办法，出(下，上，中)，则可赢一千金，这是最理想的。若出(上，下，中)等其他策略，不赢，但只输一千金。

(3) 其实，只要齐王宣布任何一种策略，例如(下，上，中)，田忌总能找到一种策略，如(中，下，上)，获得两胜一负的战绩，赢得一千金。

(4) 如果齐王不宣布自己的策略，也估计到田忌不会死守某种策略(即双方都聪明)，那么采取各种策略的机会均等，没有最优或最差策略之分。这时双方都可随便出一种策略，齐王可能得 3 分和 1

分，也可能输 1 分。但是总计来看，赢一千金的可能较大。也就是说，由于齐王的马有较强的实力，平均能赢田忌一千金。这时“聪明”将不起作用，完全靠机会(运气)来决定了。

请读者注意，我们在这里决没有宣传“宿命论”，而是客观的数量分析。在双方都“聪明”的条件下，确实没有一个“必赢的策略”可供选取。可能赢，也可能输，必须冒风险。但是估计双方实力后，齐王是强的，他举行比赛，赢一千金的可能性很大，是值得冒风险的。对田忌来说，如果估计齐王墨守成规，用老办法，当然可以用最佳策略击败齐王，赢一千金。但是实力毕竟不如齐王，输的可能性很大，输多少，平均说来是一千金，而不会是原来那样输三千金。有了这样的数量估计，大家心中基本有底。至于结局究竟如何，只能看临场“机会”了。

5.5 数据绘画

用计算机绘画，已经不是什么新鲜事了，连小孩子都会。我们在这里探讨的是用一个公式，通过迭代方法获得的数据自动画出来的一种图案。这一 20 世纪 70 年代刚刚出现的数学学科，称为“分形几何”。欣赏“分形”艺术，当是信息时代的一种时尚。

这门学科的奠基人是法国数学家 B. 曼德勃劳特(Mandelbrot)。第一张分形几何图形，就称为曼德勃劳特集。

数学问题很简单。一个复数值的二次函数：

$$w = z^2 + C,$$

其中 C 是一个常数。如果我们用 $z=C$ 为值作迭代，即

$$C_1 = C^2 + C,$$
$$C_2 = C_1^2 + C,$$
$$C_3 = C_2^2 + C,$$
$$\vdots$$
$$C_n = C_{n-1}^2 + C,$$
$$\vdots$$

得到的这一个无限数列 $C_1, C_2, \cdots C_n, \cdots$ 若是有界的（指都含在某半径为 M 的圆之内），则令 C 点为黑色点。反之，若这一数列不能被任何一个圆所盖住，则称为白色点。

例如，取 $C=0.1$，则可以得到：

$$C_1 = (0.1)^2 + 0.1 = 0.11,$$
$$C_2 = (0.11)^2 + 0.1 = 0.1121,$$
$$C_3 = (0.1121)^2 + 0.1 = 0.11256641,$$
$$\vdots$$
$$C_n = C_{n-1}^2 + 0.1,$$
$$\vdots$$

它的极限是 0.112701665…，所以数列含在半径为 1 的圆之内。这表明 0.1 是黑色点。若取 $C=1$，则有 $C=1+1=2$，$C=2^2+1=5$，$C=5^2+1=26$，…，如此继续，这一数列会趋向无穷大，不可能被一个圆所围住。所以 1 是白色点。照此办理，平面上每个点都可以分辨出是黑色点或白色点。黑色点形成的图形便是曼德勃劳特集。如果把黑色点所联系的圆半径 M 再分档次：M 小于 100 的为蓝色，$M=100\sim500$

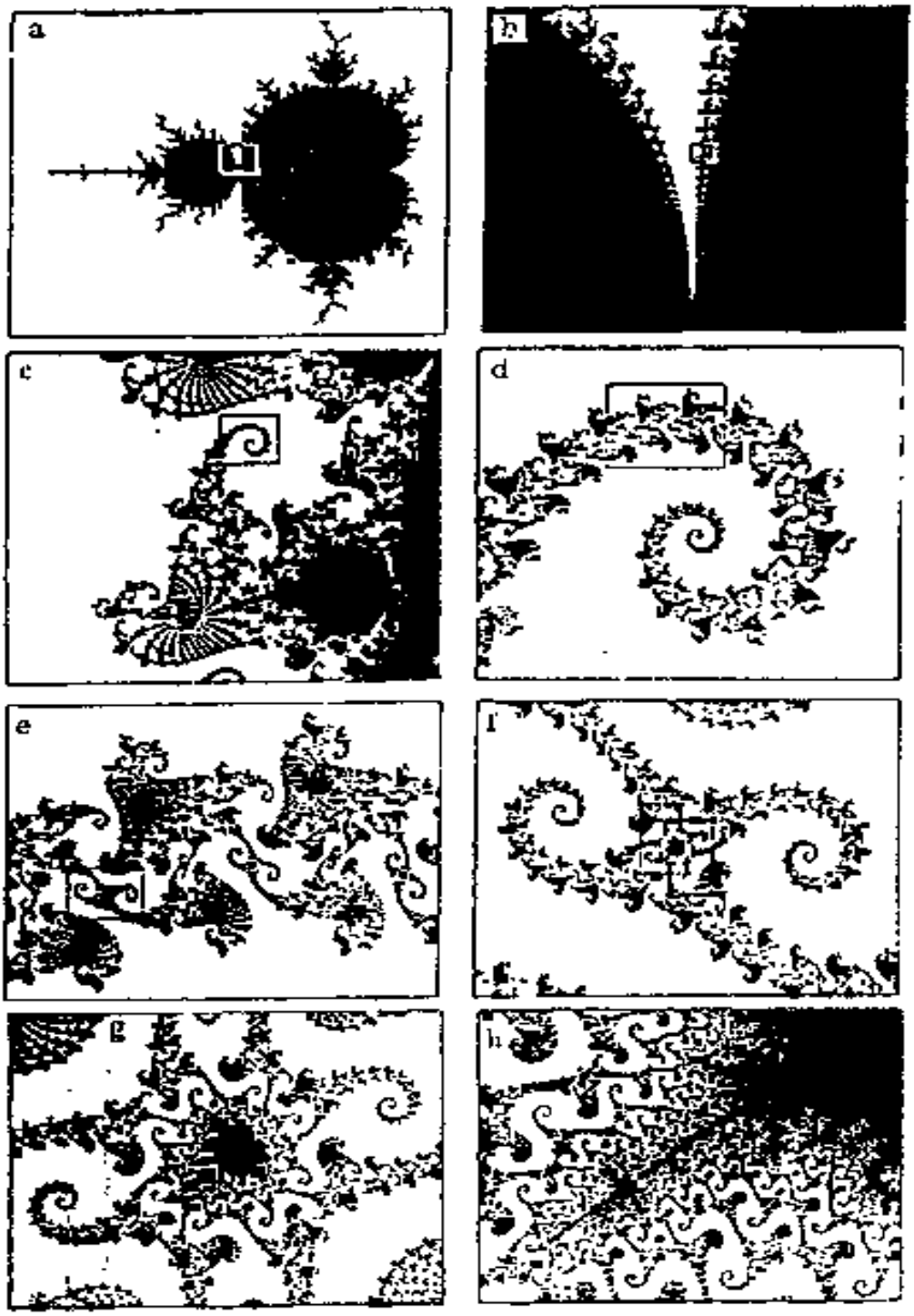


图 5.2　黑白的曼德勃劳特图形

图 a 中，黑色区域内的每一点 c 都对应于属于 A 类的复数，其黑色区域边缘的发状结构是最有趣的部分；图 b 对应于图 a 中的小方块，其黑色区域内的每一点 c 都对应于 A 类的复数；图 c 对应于图 b 中的小方块，其黑色区域内的每一点 c 也都对应于 A 类的复数。这个模式在直到图 h 的整个系列中重复出现。

的为绿色，$M=500\sim1000$ 之间的点为黄色，M 在1000以上的为红色。那么曼德勃劳特集便成为彩色的了。

曼德勃劳特集是一个很美丽的图形，它是用一个公式，借助迭代所获得数列的特性画出来的图案，对高速电子计算机来说，这种机械的迭代计算并不困难。我们说的数据绘画，就是这样画出来的。

曼德勃劳特集有一个非常引人注目的特征：自相似性。每一个局部放大之后，仍和整体相似，一层一层没有完结。（参见图 5.2）

分形几何是新兴学科，目前还在发展之中。它的制作方法多种多样，但都是用迭代数列而画出来的。参见封面图。最普通、最熟悉的迭代方程，只要变动了参数、尺度范围、逃逸值(escape value)的色谱搭配，也会产生好像全然陌生的景象。所以，在可以飞霜的六月时节，让我们重访由经典函数 $z_{n+1}^2=z_n^2+c(c=0.3+0.6i)$ 在 z-面所产生的殷红与碧蓝相交错泼的朱利亚集图像。左图的左上和右下两角是 $z=-0.263+0.802i$ 和 $z=-0.279+0.814i$；左图白框部分放大后依次得中图和右图。