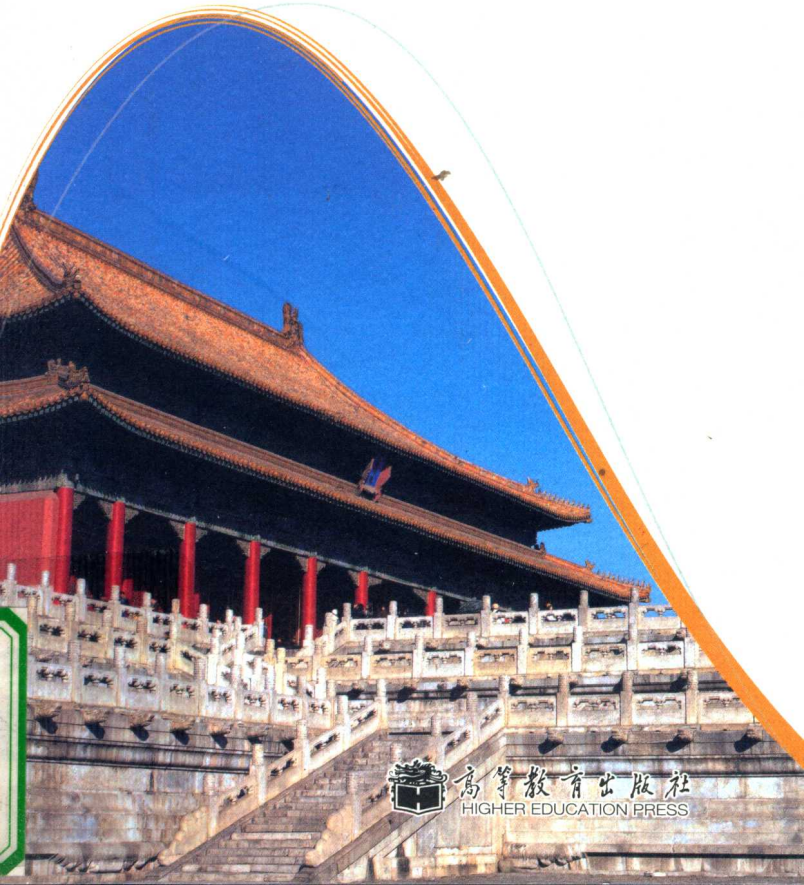
22

□ 数学文化小丛书

李大潜 主编

统计知玄妙

○ 杨 虎



高等教育出版社
HIGHER EDUCATION PRESS

数学文化小丛书

李大潜　主编

统计知玄妙

Tongji Zhi Xuanmiao

杨　虎

高等教育出版社·北京
HIGHER EDUCATION PRESS　BEIJING

图书在版编目(CIP)数据

统计知玄妙/杨虎编. —北京:高等教育出版社,2013.6
(数学文化小丛书/李大潜主编. 第3辑)
ISBN 978-7-04-037206-9

Ⅰ.①统… Ⅱ.①杨… Ⅲ.①统计分析-普及读物
Ⅳ.①O212.1-49

中国版本图书馆CIP数据核字(2013)第068075号

策划编辑 李 蕊　　责任编辑 胡 颖　　封面设计 张 楠
版式设计 王艳红　　插图绘制 邓 超　　责任校对 杨雪莲
责任印制 张泽业

出版发行 高等教育出版社
社　　址 北京市西城区德外大街4号
邮政编码 100120
印　　刷 北京机工印刷厂
开　　本 787mm × 1092mm　1/32
印　　张 2.25
字　　数 39千字
购书热线 010-58581118

咨询电话 400-810-0598
网　　址 http://www.hep.edu.cn
　　http://www.hep.com.cn
网上订购 http://www.landraco.com
　　http://www.landraco.com.cn
版　　次 2013年6月第1版
印　　次 2013年6月第1次印刷
定　　价 7.00元

本书如有缺页、倒页、脱页等质量问题,请到所购图书销售部门联系调换
版权所有 侵权必究
物料号 37206-00

数学文化小丛书编委会

顾　问：谷超豪（复旦大学）
项武义（美国加州大学伯克利分校）
姜伯驹（北京大学）
齐民友（武汉大学）
王梓坤（北京师范大学）
主　编：李大潜（复旦大学）
副主编：王培甫（河北师范大学）
周明儒（江苏师范大学）
李文林（中国科学院数学与系统科学研究院）
编辑工作室成员：赵秀恒（河北经贸大学）
王彦英（河北师范大学）
张惠英（石家庄市教育科学研究所）
杨桂华（河北经贸大学）
周春莲（复旦大学）
本书责任编委：赵秀恒

数学文化小丛书总序

整个数学的发展史是和人类物质文明和精神文明的发展史交融在一起的。数学不仅是一种精确的语言和工具、一门博大精深并应用广泛的科学，而且更是一种先进的文化。它在人类文明的进程中一直起着积极的推动作用，是人类文明的一个重要支柱。

要学好数学，不等于拼命做习题、背公式，而是要着重领会数学的思想方法和精神实质，了解数学在人类文明发展中所起的关键作用，自觉地接受数学文化的熏陶。只有这样，才能从根本上体现素质教育的要求，并为全民族思想文化素质的提高夯实基础。

鉴于目前充分认识到这一点的人还不多，更远未引起各方面足够的重视，很有必要在较大的范围内大力进行宣传、引导工作。本丛书正是在这样的背景下，本着弘扬和普及数学文化的宗旨而编辑出版的。

为了使包括中学生在内的广大读者都能有所收益，本丛书将着力精选那些对人类文明的发展起过重要作用、在深化人类对世界的认识或推动

此为试读，需要完整PDF请访问：www.ertong

人类对世界的改造方面有某种里程碑意义的主题，由学有专长的学者执笔，抓住主要的线索和本质的内容，由浅入深并简明生动地向读者介绍数学文化的丰富内涵、数学文化史诗中一些重要的篇章以及古今中外一些著名数学家的优秀品质及历史功绩等内容。每个专题篇幅不长，并相对独立，以易于阅读、便于携带且尽可能降低书价为原则，有的专题单独成册，有些专题则联合成册。

希望广大读者能通过阅读这套丛书，走近数学、品味数学和理解数学，充分感受数学文化的魅力和作用，进一步打开视野、启迪心智，在今后的学习与工作中取得更出色的成绩。

李大潜

2005 年 12 月

目　录

一、引言……1

二、乱箭齐射有多准……6

三、惊人的预测……15

四、莎士比亚新诗……25

五、生日问题……32

六、探寻运气的密码……40

七、艰难的选择……44

八、隐私问题调查……50

九、两极分化与回归……56

参考文献……61

致谢……62

一、引　言

提起统计，可以有两种理解. 一是统计学的简称，更多地存在于教育和学术研究中，目前已经发展成为一个独立的一级学科. 而在社会与生活中，人们提到统计，指的是一种行为，而统计——这种人类的行为几乎可以追溯到人类的起源，清点一下部落有多少人，有多少食物，就是在对资源进行统计. 这里的统计有笼统计数的意思，这其实是数据统计. 今天的统计部门主要进行的就是这项工作.

本书所说的统计是指后者，将告诉人们怎么去对随机现象进行数据统计和分析，并通过这种统计工作，去探知隐藏在随机现象背后的玄妙.

数据统计是统计学术研究的源泉. 当我们统计一个小规模团体的人数、年龄、收入、性别、身高、体重这些指标时，如果统计人员足够细致，由于总人数实在太少，也不用考虑统计误差，统计是派不上什么大用场的.

但随着这个团体人数的增加，如一个市的市民、一个国家的公民；统计指标的增多，如人们的健康状况、家庭经济情况、人的寿命等，仅仅要求统计人员认

真细致就远远不够了. 这时统计就会派上用场了.

今天, 问题驱动的应用数学研究已经逐渐成为学术界的共识, 统计学作为一门数据科学, 脱离数据就没有生命力了. 本书从一些基本的随机现象说起, 希望帮助读者体会进行数据统计的重要性, 进而发现统计规律. 有些问题本质上没有随机性, 但却表现出很强的随机特性, 如果片面理解, 就容易被迷惑.

下面先对统计学学科的发展作一简单回顾.

统计学 (statistics) 一词是 18 世纪中叶德国学者阿亨瓦尔 (G. Achenwall) 创设的, 词根是拉丁语中“国家” (status) 一词, 意即: “由国家来收集、处理和使用数据”. 但统计学成为一门学科的历史却很短, 至 1834 年创立英国皇家统计学会, 统计学才开始被承认为一门独立的学科. 当时认为统计学是

与人类有关的事实, 可以由数量来表示, 并且经过大量的累积重复可以导出一般规律.

统计起源于何时? 众说纷纭, 有说古巴比伦的, 有说古埃及的, 但记载最详实可信的是公元前 2000 年左右的夏朝, 中国进行了人口调查统计. 周朝 (公元前 1046—前 256 年) 为了管理统计工作还设立了“司书”职位, 这是中国国家统计局局长的鼻祖. 西方关于统计最早的记载是《旧约圣经》第四册, 该书引用了公元前 1500 年左右的早期人口统计结果, 这是摩西对以色列军队进行的调查统计.

早期的统计行为更多是统治者为了征兵、征税方便, 进行的人口统计. 这其实是今天统计局所做

的主要工作.

班上搞活动, 老师通知班长统计一下有多少人参加; 学校开运动会, 班主任需要统计一下本班学生取得的成绩; 考试结束了, 要对成绩进行统计和分析; 过春节要开展送温暖活动, 也要统计一下有多少贫困家庭, 等等. 经过统计, 才能获取最直接的第一手资料, 通过对资料的深入分析, 达到了解问题本质的目的.

生活中的统计其实很简单, 就是收集、整理和分析数据.

需要特别强调的是: 这里所说的数据通常是有随机性的, 统计整理的这些数据称为**样本** (sample). 提到随机性, 就不得不提到另外一个概念:**概率** (probability). 统计和概率联系紧密, 统计的过程离不开概率计算, 但和概率中对固定对象的概率分析不同, 统计总是需要处理大量的样本, 通过不断的**抽样** (sampling) 和分析, 才能发现隐藏在随机性表面现象背后的统计规律性.

随机性是一种不确定性, 但这种不确定性是有规律可循的. 比如投掷硬币, 比如掷一枚骰子, 比如股票的涨跌, 在结果显示前, 到底什么结果发生是无法预知的, 但人们都知道硬币出现正面的可能性是二分之一, 这种可能性就称为概率. 同样, 骰子掷出六点的概率是六分之一. 天气预报明天下雨的概率是 65%, 表示明天有 65% 的可能下雨, 从统计的角度, 如有 100 次这样的出行, 带雨伞会有 65 次左右是正确的, 另外 35 次雨伞则可能成为累赘.

上面所说的随机性包含两方面的内容, 一是所有可能的结果是已知的, 二是以固定的概率发生这些可能的结果, 后者又称为**概率分布** (probability distribution).

比如投掷硬币, 结果除了正面就是反面, 概率均为 $\frac{1}{2}$, 我们称它的概率分布为**两点分布**或**均匀分布** (uniform distribution). 投掷骰子, 所有可能的结果是 1~6 这 6 个数字, 概率均为 $\frac{1}{6}$, 概率分布也是均匀的. 通过统计获取概率是计算概率的方法之一. 除此之外, 概率论中还有很多计算概率的方法, 比如古典概率、几何概率、先验概率等. 本书涉及的概率都是通过样本统计出来的, 这样得到的概率俗称**频率** (frequency), 它其实是概率的一种近似.

通过大量观测和统计自然界的随机现象, 就能发现这些随机现象发生的频率, 进而获得概率 (作为频率的极限, 后面还要讲到). 比如, 通过天文观测, 我们发现了哈雷彗星的周期回归, 从而可以推算它的运行轨道; 通过水文观测, 我们可以预知潮起潮落; 对寿命与疾病的统计与分析, 可以得到人群中各种疾病的发生概率和平均寿命, 进而成为今天寿险保单的计算依据. 很多时候, 某件事发生的概率需要进行大量统计才能发现, 在通过计算得到概率的同时, 我们就掌握了相应的统计规律.

概率和统计的区别还是很明显的. 就第二部分介绍的生日问题而言, 我们计算 50 人中有两人生日相同的概率, 这完全是概率计算. 但当考虑的问题是

此为试读, 需要完整PDF请访问: www.ertong

一个班的同学中是否有两人生日相同, 这就是一个统计问题. 首先, 你需要进行客观的统计, 记录下班上每个同学的生日; 其次, 统计上来的生日数据仅仅是一个样本, 每个“读者”所在的班级不同, 样本不一样,结果就会不一样. 如果只做一次这样的统计, 我们无法印证前面计算的概率是否正确, 只有进行大量的统计, 才能检验概率计算的正确与否. 一般而言, 通过统计可以获得自然界中随机现象发生的真实概率, 这是人们进行理性思维与决策的一个基础.

在统计的意义下, 概率再小, 样本足够大, 也会存在统计的必然. 因此不要因为发生小概率现象大惊小怪: 由于样本很大 (这已经是统计必然的了), 碰巧被你遇上而已; 你不遇上, 也一定会有人碰上. 中国有句古话: 久走夜路遇到鬼. 放到今天, 一个喜欢冒险的人, 迟早会发生意外. 讲的都是同一个道理.

不要看媒体铺天盖地报道中彩票大奖的人, 甚至让他介绍成功经验, 以为他真的有什么独家秘笈可供大家分享. 其实中奖的人只有统计学上的意义, 是必然中的偶然个体. 同样, 有人迷信高考状元的学习经验, 以为照样学习就能中状元, 那也是无稽之谈. 高考状元也不具备统计学分类的特征, 只是一个偶然群体.

很多问题看起来都没什么规律, 因此很多人才会无奈地归结为命运, 经常有人感叹造化弄人. 面对纷纭的随机世界, 我们真的只能听天由命么?

且慢下结论, 让我们认真统计一下, 也许就能发现个中玄妙.

二、乱箭齐射有多准

我国大约在28000年前就出现了弓箭(图1),是世界上最早发明弓箭的国家.传说东夷人最早发明了弓箭,“夷”字就是由“大”和“弓”组成.《说文》记载:“夷,东方之人也.从大,从弓”.《说文通训定声》记载:“夷,东方之人也.东方夷人好战好猎,故字从大持弓会意,大人也”.至于具体的弓箭发明者,有史料可查的主要有三位:盘、张挥、后羿.后羿射日是脍炙人口的神话故事,这三人均是东夷人.

图1　骑射图

自弓箭产生以后,人类的活动范围迅速扩大,走出山洞巢穴,离开大树、森林,来到平坦广阔的平原草地安家.人类利用弓箭不但能够得到更多的猎物,

为自身的生存繁衍创造良好的物质条件, 而且也大大加强了自身的安全防御能力.

弓箭作为能够远距离射杀敌人的武器, 迅速成为古代战争的利器, 并被列为兵器之首. 古代战争常以军阵对垒的形式展开, 弓箭在军队中的作用非常显著. 唐代《太白阴经》一书“器械”篇记载: 唐一军编制 12500 人时, 装备“弓一万二千五百张, 弦三万七千五百条, 箭三十七万五千支”. 弓箭手们在战阵前一字排开, 控弓发箭, 千弓同张, 万箭齐发, 这场景在电影和连续剧中渲染得不少. 可是, 射箭总是有对象的, 而且箭的准头显然无法和现代的枪支相比, 距离越远, 射中的可能性越小, 战场上射箭的命中率究竟怎样呢?

古典小说中打不赢就放箭几乎是定式, 一些神箭手的故事更是脍炙人口, 但普通军士的射箭命中率有待考证. 奥运会的射箭比赛让人真的有点感慨人类的射箭水平怎么退化得那么快, 十环的圆圈很近, 也很大, 但各国的射箭运动员就是很难中的. 吕布辕门射戟, 薛仁贵能射张口雁, 养由基百步穿杨, 这些故事连小孩子都知道, 现在看来有些夸大也未可知.

《说唐》里面有一则让人落泪的故事, 后来也被搬上了戏剧舞台, 说的是罗成 (图 2) 匹马单枪大战后汉苏定方大军. 罗成虽然在隋唐英雄中武功排第七, 但到了那个时候, 排前面的好汉都死了, 论单打独斗, 罗成已经没有敌手, 但最终的结果是他被诱入淤泥河被乱箭射死.

图 2　罗成

二战时期, 没有导弹, 飞机在天上飞, 通常是用高射炮射击. 一门高射炮的命中率很低, 因此也采用了乱炮齐射的战术. 这样效果怎样呢?

假如每个士兵单独一次射击命中的概率为 0.1, 意味着射 10 箭才中一箭. 这样的弓箭手, 估计在冷兵器的古代没法在军队混饭吃吧? 曹操赤壁大战的上万弓箭手是这个水平么? 以曹操的治军和秉性, 估计会高于这个命中率. 出于统计分析的需要, 这里假定就是这个水平, 那么 100 次射击命中目标至少一次的概率有多大呢?

直接计算这个问题比较麻烦, 但考虑相反的问题就容易得多. 既然一次射击命中的概率是 0.1, 不能命中的概率就是 0.9, 两次独立射击都不中的概率应该是 $0.9\times0.9=0.81$, 以此类推, 100 次射击全都不能命中的概率就是 $0.9^{100} = 0.00003$. 可见, 100 次射击至少命中 1 次的概率就是 $1-0.00003 = 99.997\%$.

由此可见, 与其费劲去寻找神箭手, 冀望他一箭射中, 还不如让 100 个军士乱箭齐射的效果更好.

长坂坡能成就赵子龙 (图 3) 的神话, 曹操严令不许放箭、要抓活的是主要原因. 否则曹军的乱箭就不是 100 支, 而是上万支了. 赵子龙纵然武功盖世, 也难生还.

图 3　赵子龙

这个问题变一变, 放箭 100 次, 每箭的命中率仍然是 0.1, 那么仅中一箭的概率是多少呢? 也就是说: 这 100 箭中只有一箭射中, 而其余 99 箭都未射中. 这里的关键就是哪一箭射中的问题, 因为余下的都是不中的, 这 99 箭都未射中的概率就是 0.9^{99}, 而命中的一箭可以是 100 箭中的任何一箭, 故有 100 种可能. 于是, 若记 p_1 为 100 箭仅命中目标一箭的概率, 则

$$p_1 = 100 \cdot 0.1 \cdot 0.9^{99} = 0.0003.$$

用记号 C_{100}^{1} 表示 100 箭中有一箭命中的组合数, 则上面的公式又可以写成

$$p_1 = \mathrm{C}_{100}^{1} \cdot 0.1 \cdot 0.9^{99}.$$

排列组合早在中国西周后期就开始萌芽,《易经》中就有最早关于组合的内容. 公元 850 年印度数学家摩诃毗罗 (Māhāvira) 给出了 n 个物体每次取 r 个的取法数的完整组合公式. 现在我们都知道, 组合数 C_n^r 表示从 n 个不可区分的物体中抽取 r 个物体的取法总数. 如果这些物体是可区分的, 那么还需要考虑它们的抽取顺序, 从而构成排列数, 记为 P_n^r. 组合数 C_n^r 与排列数 P_n^r 其实就相差一个倍数 $r!$(r 的阶乘, $r! = 1 \cdot 2 \cdot \cdots \cdot r$), 即 $\mathrm{P}_n^r = r!\mathrm{C}_n^r$. 这里仅通过一个例子来说明排列组合公式. 假定有 n 个学生, 任选其中的 r 个学生排成一排, 有多少种排法呢? 如果依次选择, 选第一个学生的时候, 共有 n 种选择. 选第二个学生的时候, 由于只剩下了 $n-1$ 个学生, 故仅有 $n-1$ 种选择, 以此类推, 直到选出全部 r 个学生, 共有排法总数

$$\mathrm{P}_n^r = n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!}.$$

再根据组合数与排列数的关系, C_n^r 的公式如下:

$$\mathrm{C}_n^r = \frac{n!}{(n-r)!r!}.$$

回到原来的射箭问题, 如果射 100 箭命中目标 r 箭 ($0 \leqslant r \leqslant 100$), 这 r 箭可以是这 100 箭中的任何 r 箭, 共有 C_{100}^r 种可能, 因此, 记 p_r 为 100 箭中命中目标 r 箭的概率, 则有概率计算公式:

$$p_r = \mathrm{C}_{100}^r \cdot 0.1^r \cdot 0.9^{100-r}.$$

此为试读,需要完整PDF请访问: www.ertong