普通高中教科书

数学

SHUXUE

必修

第四册

湖北教育出版社

普通高中教科书

数学

必修

第四册

湖北教育出版社

ISBN 978-7-5564-3143-4

普通高中教科书

数学

SHUXUE

必修

第四册

主　编　彭双阶

湖北教育出版社

主　　编：彭双阶

副 主 编：徐胜林　胡典顺　郭熙汉

本册主编：胡典顺

主要编者：余锦银　陈应保　左国新　胡典顺　徐胜林

郭熙汉　彭树德

Shuxue

致高中生

高中数学是一门非常重要的课程. 数学以其卓越的智力成就被人们尊称为“科学的皇后”. 数学是人类最高超的智慧活动，是人类心灵最独特的创造，是形成人类文化的主要力量，是人类文明的核心部分，是认识世界和创造世界的一把关键钥匙.

我们需要数学，因为作为人类文明发展标志的数学，是人类文化的重要组成部分. 数学既是一种睿智的文化、一种思想的体操，更是现代科技进步中理性文化的核心.

我们需要数学，因为数学在形成人类理性思维和促进个人智力发展的过程中发挥着独特的、不可替代的作用. 数学素养是现代社会公民应该具备的一种必备品格.

我们需要数学，因为数学是刻画自然规律和社会现象的特殊语言和有力工具，是自然科学、技术科学的基础，在经济科学、社会科学、人文科学的发展中发挥越来越强大的作用.

我们需要数学，因为数学已经渗透到现代社会和人们日常生活的各个方面. 学好数学是提升生活质量、优化生活品质的重要保证.

本套教科书以《普通高中数学课程标准（2017 年版）》为依据来编写，遵循了现代数学教与学的规律，着眼于 21 世纪现代生活和未来发展，力求提升同学们的数学核心素养，更快地适应未来社会的发展.

教科书是教与学的一种重要资源. 在使用本套教科书的同时，我们还应该多关注现实生活，关注社会进步和科技发展，用数学眼光观察世界，用数学思维思考世界，用数学语言表达世界. 现代社会是信息社会，又是终身学习的社会. 在这个大数据时代，我们可以根据实际条件，选择利用计算机与互联网，丰富学习资源，提高学习效率. 积极参与数学活动，勤于思考，敢于质疑，乐于合作交流，克难奋进，砥砺前行，养成良好的数学学习习惯，让数学学习变得更加生动活泼、富有情趣.

亲爱的同学们，插上快乐的翅膀，带着青春的梦想，在浩瀚的数学海洋扬帆奋进吧！

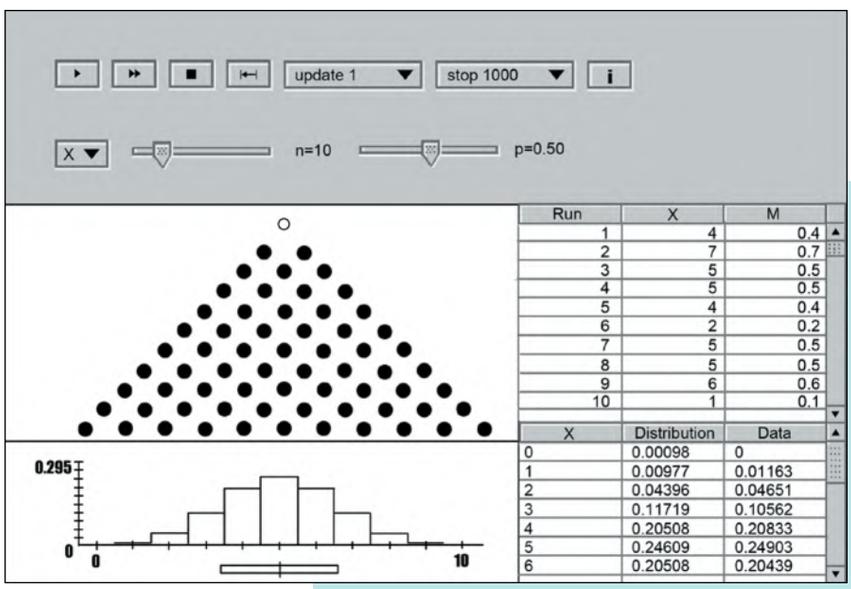
Mulu

目 录

第1章 概率

第2章 统计

第1章 概 率



验证频率稳定性的高尔顿板实验

乒乓球比赛时，裁判用投掷硬币猜正反面的方法，让双方决定谁先要场地，谁先要发球权，这种做法公平吗？

气象台发布的天气预报说：“预计我市明天降雨的概率为 80%.”这句话应如何理解？

某人购买了一张彩票，他中奖的概率有多大？如果他已经知道了以前各次开奖的信息，他能设计一个相对有利的选号方案吗？

我们通常所说的“百年一遇”，指的是“任意一百年只会发生一次”或者“如果已经发生一次，在未来若干年内不可能再发生”吗？

一些偶然现象，虽然表面上看来杂乱无章、毫无规律可循，然而如果大量重复观察这些偶然现象，就有可能发现其中的规律性. 对这种规律性的研究就是概率论的重要内容之一.

20 世纪 30 年代，概率论作为一门数学学科，在理论研究和实际应用方面得到了极大的发展，如今已经广泛地渗透到了物理、化学、生物等其他学科，在天文、生态、经济、工程技术等领域也有着广泛的应用.

本章将介绍随机现象，随机事件的关系及运算，频率与概率的关系，古典概型，独立事件概率的简单计算等相关知识.

1.1 随机事件及其概率

1.1.1 随机事件

在人类社会和自然界中，经常会遇到两类不同的现象．一类称为**确定性现象**，即在一定条件下必然发生或必然不发生的现象．例如，在地球上向空中投掷一颗石子，由于地球引力的作用，它必然会落到地面，或者说它必然不会飞向太空；在1个标准大气压下，水加热到100℃必然会沸腾；在所受的合外力为零时，做匀速直线运动的物体必然继续做匀速直线运动．这类现象都是确定性现象．另一类称为**随机现象**，即在一定条件下，可能出现这种结果，也可能出现那种结果，并且事先不能确定会出现哪一种结果的现象．我们来看下面的几个问题．

如果没有特别说明，本章中所说的骰子、硬币等都是均匀的．

问题 1　投掷一枚均匀的骰子，考虑朝上一面的点数，投掷前无法确定．

如果没有特别说明，本章中所说的球，都是指大小和质量完全相同的球．

问题 2　一个不透明的盒子中，装有1个白球和1个红球．随机取出一个，可能是白球也可能是红球，取出前无法确定是什么颜色的球．

问题 3　为了了解某农作物种子的质量，更好地指导生产，确定播种量，常要先做种子的发芽试验．若取100粒种子进行发芽试验，你能确定有多少粒种子发芽吗？大家容易想到，种子的发芽粒数可能为0，1，2，…，100，但事先并不知道到底会有多少粒种子发芽．

这些问题中所描述的现象，虽然有不同的背景，但它们有一个共同的特点，即在基本条件不变的情况下，一系列试验或观察可能得到不同的结果．换句话说，就个别的试验或观察而言，它会时而出现这种结果，时而出现那种结果，呈现出事先不能确定出现哪种结果的偶然性．

经过长期实践并深入研究之后，人们发现这类现象在大量

重复试验或观察下，这些偶然性的结果呈现出某种必然的规律性. 例如：多次重复投掷一枚硬币，得到正面朝上的次数大致有一半，这种大量重复试验或观察中所呈现出的固有规律性，就是我们通常所说的统计规律性.

概率论就是研究和揭示随机现象的科学.

概率统计中把对客观现象进行观察或进行科学试验统称为试验. 进行一次试验，如果其所得结果不能预知，但其全体可能结果是已知的，则称此试验为**随机试验**，简称**试验**. 在概率论中，我们通过研究随机试验来研究随机现象. 一般地，一个随机试验要具有下列特点：

(1) 可重复性：试验能够在相同的条件下重复进行；

(2) 可观察性：每次试验的可能结果不止一种，这些可能结果是什么，在试验之前是明确的；

(3) 随机性：每次试验必能而且只能出现这些可能结果中的一种，但是事先不能确定会出现何种结果.

我们把在随机试验中每一种可能结果叫作**样本点**(sample point). 所有样本点组成的集合叫作**样本空间**(sample space)，记为 Ω.

问题 1 中的样本空间为 $\Omega=\{1, 2, 3, 4, 5, 6\}$，数字表示骰子朝上一面的点数，每个数字为一个样本点. 若记“出现 1 点”为事件 A，则它由单个样本点“1”组成；若记“出现偶数点”为事件 B，则它由三个样本点“2”“4”“6”组成.

问题 2 中的样本空间为

$$\Omega=\{\text{白球，红球}\},$$

其中“白球”“红球”为样本点.

问题 3 中的样本空间为

$$\Omega=\{0, 1, 2, \cdots, 100\},$$

数字表示可能出现的发芽粒数，每个数字为一个样本点.

在一定条件下必然发生的事件称为该条件下的**必然事件**(certain event)，用样本空间 Ω 表示. 在一定条件下一定不会发生的事件称为该条件下的**不可能事件**(impossible event)，记为 $\varnothing$. 不可能事件 $\varnothing$ 不含任何样本点. 必然事件与不可能事件统称为**确定事件**.

从集合的角度看，$\varnothing$ 是空集，Ω 是全集.

问题 1 中，若记“出现的点数大于 6”为事件 C，则 Ω 中的任意样本点都不在 C 中，所以 C 是不可能事件；若记“出现的点数不超过 7”为事件 D，则 Ω 中的任意样本点都在 D

中，所以 D 是必然事件.

在一定条件下可能发生也可能不发生的事件称为在该条件下的**随机事件**(random event)，简称**事件**，用大写拉丁字母 A，B，C，…表示.

例如，过马路交叉口时可能遇上各种颜色的交通信号灯，这是一个随机现象，而“遇到红灯”则是一个随机事件.

随机事件是样本空间的子集. 事件 A 发生当且仅当 A 所含的某个样本点在试验中发生.

随机事件 A 在一次试验中发生的可能性大小的度量，叫作此随机事件发生的**概率**(probability)，记作 $P(A)$. 必然事件 Ω 的概率 $P(\Omega)=1$，不可能事件 $\varnothing$ 的概率 $P(\varnothing)=0$.

例 1 设袋中有编号为 1，2 的 2 个黑球，编号为 3，4，5 的 3 个白球. 现从袋中依次不放回地取出 2 个球，试用集合的符号表示下列事件：

(1) 必然事件 Ω；

(2) A：第一次取到黑球；

(3) B：第二次取到黑球.

解 为了方便描述，我们用一个两位数来表示样本点，两位数的十位数字表示第一次取到的球的编号，个位数字表示第二次取到的球的编号，则

(1) $\Omega=\{12, 13, 14, 15, 21, 23, 24, 25, 31, 32, 34, 35, 41, 42, 43, 45, 51, 52, 53, 54\}$.

(2) $A=\{12, 13, 14, 15, 21, 23, 24, 25\}$.

(3) $B=\{21, 31, 41, 51, 12, 32, 42, 52\}$.

你还能给出其他的表示吗？若考虑“有放回地取球”，则答案如何？

例 2 将一枚骰子依次投掷两次，写出样本点和样本空间.

解 用 1，2，3，4，5，6 表示投掷出的点数，用 (i, j) 表示“第一次投掷出 i 点，第二次投掷出 j 点”，则相继投掷两次骰子的所有可能结果如下：

$(1, 1), (1, 2), (1, 3), (1, 4), (1, 5), (1, 6),$

$(2, 1), (2, 2), (2, 3), (2, 4), (2, 5), (2, 6),$

$(3, 1), (3, 2), (3, 3), (3, 4), (3, 5), (3, 6),$

$(4, 1), (4, 2), (4, 3), (4, 4), (4, 5), (4, 6),$

$(5, 1), (5, 2), (5, 3), (5, 4), (5, 5), (5, 6),$

(6，1)，(6，2)，(6，3)，(6，4)，(6，5)，(6，6).

注意到这里(1，2)和(2，1)是不同的样本点，分别表示“第一次投掷出 1 点，第二次投掷出 2 点”和“第一次投掷出 2 点，第二次投掷出 1 点”这两个随机事件，因此，样本空间共有 36 个样本点. 样本空间为

$$\begin{aligned}\Omega=&\{(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),\\&(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),\\&(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),\\&(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),\\&(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),\\&(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)\}\\=&\{(i,j)\mid i,j=1,2,3,4,5,6\}.\end{aligned}$$

练习

1. 下列现象中，哪些是随机现象？哪些是确定性现象？
   (1) 一天内进入某超市的顾客数；
   (2) 明天的最高气温；
   (3) 射击训练中，某运动员每次命中的环数；
   (4) 太阳从东方升起；
   (5) 投掷一枚骰子，出现的点数；
   (6) 两人各买一张彩票，中奖的情况.
2. 依次投掷三枚硬币，写出下列事件的集合表示：
   (1) 样本空间 Ω；
   (2) 第一次出现正面；
   (3) 前两次都出现正面.
3. 从字母 a，b，c，d 中任意取出两个不同字母的试验中，有哪些样本点？
4. 一只口袋内装有 3 个编号为 1，2，3 的白球和 2 个编号为 4，5 的黑球，从中一次取出两个球.
   (1) 共有多少个样本点？
   (2) “两个球都是白球”这一事件包含几个样本点？

1.1.2 事件的关系及运算

我们知道，集合之间存在一定的关系，并且可以进行运算. 把样本空间看成全集，则事件就是子集. 因此，作为一类

特殊的集合问题，对事件也可以讨论它们的关系和运算，并且沿用集合中的运算符号.

在投掷一枚骰子的试验中，用 1，2，3，4，5，6 表示投掷出的点数，可以定义许多事件，如记：

$C=\{1，2，3\}$；　　$D=\{2，3，4\}$；

$E=\{$点数不大于 2$\}$；　　$F=\{$点数大于 3$\}$；

$G=\{$点数是偶数$\}$；　　$H=\{$点数是奇数$\}$.

一般地，若某事件发生当且仅当事件 A 发生且事件 B 发生，则称此事件为事件 A 与事件 B 的**积事件**(或**交事件**)，记为 AB 或 $A\cap B$.

AB 由同时属于 A 和 B 的样本点组成.

例如，在投掷骰子的试验中，$CD=\{2，3\}$.

若某事件发生当且仅当事件 A 发生或事件 B 发生，则称此事件为事件 A 与事件 B 的**和事件**(或**并事件**)，记为 $A+B$ 或 $A\cup B$.

$A+B$ 由属于 A 或属于 B 的样本点组成.

例如，在投掷骰子的试验中，$C+D=\{1，2，3，4\}$，它表示事件 C 与事件 D 至少有一个发生.

一般地，我们称事件 A 不发生的事件为事件 A 的**对立事件**(complementary events)，记为 $\overline{A}$. A 与 $\overline{A}$ 互为对立事件.

$\overline{A}$ 由所有不属于 A 的样本点组成.

显然有 $A\overline{A}=\varnothing$，$A+\overline{A}=\Omega$，在任何一次试验中，事件 A 与事件 $\overline{A}$ 有且仅有一个发生.

例如，在投掷骰子的试验中，$GH=\varnothing$，$G+H=\Omega$，所以 G 与 H 互为对立事件.

一般地，若两个事件 A，B 在任何一次试验中都不会同时发生，则我们称事件 A 与事件 B 为**互斥事件**(exclusive events)，也称为**互不相容事件**(incompatible events).

例如，在投掷骰子的试验中，事件 E 与事件 F 互斥，事件 G 与事件 H 互斥.

一般地，若事件 A 发生必然导致事件 B 发生，我们称事件 B 包含事件 A，记为 $A\subseteq B$.

$A\subseteq B$ 时，属于 A 的样本点都属于 B.

例如，在投掷骰子的试验中，事件 C 包含事件 E.

容易证明下列结论：

(1) 对任意事件 A，有 $\varnothing\subseteq A\subseteq\Omega$；

(2) 若 $A\subseteq B$，则 $AB=A$，$A+B=B$；

(3) 若 $A\subseteq B$ 且 $B\subseteq A$，则 $A=B$.

对于事件间的关系及运算，比起用式子或者语言描述，用画图的方法更能让人一目了然，图 1-1 是事件关系及运算的

Venn 图表示：

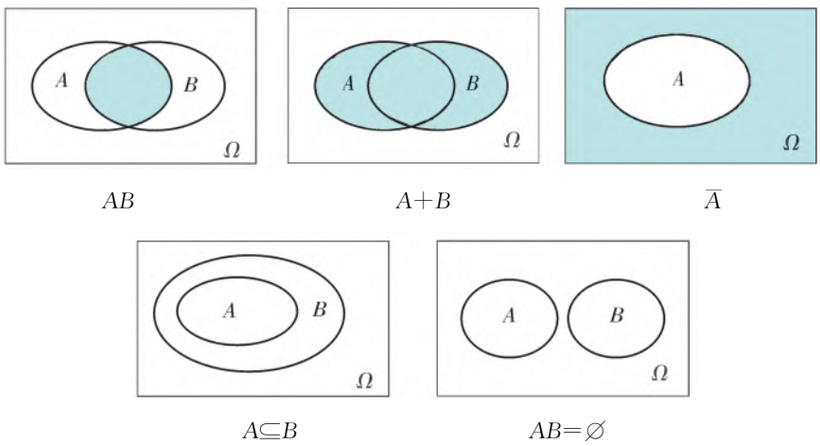


图 1-1

分析事件之间的关系，可以帮助我们更加深刻地认识事件的本质，借助 Venn 图来思考事件的关系及其运算，就更易于厘清事件关系，从而简化一些复合事件的概率计算.

请根据集合运算的性质，给出事件运算的性质.

例 1 设 A，B 为任意两个事件，试用它们的运算关系表示下列事件：

（1）A 发生，B 不发生；

（2）A，B 中恰有一个发生；

（3）A，B 中至多有一个发生；

（4）A，B 中至少有一个发生.

解 （1）A 发生，B 不发生可以表示为

$$A\overline{B}.$$

（2）A，B 中恰有一个发生可以表示为

$$A\overline{B}+\overline{A}B.$$

（3）A，B 中至多有一个发生可以表示为

$$\overline{A}\,\overline{B}+\overline{A}B+A\overline{B}.$$

（4）A，B 中至少有一个发生可以表示为

$$A+B \text{ 或 } A\overline{B}+\overline{A}B+AB.$$

例 2 依次投掷两枚骰子，若记“两枚骰子朝上的点数之和小于 7”为事件 A，记“第一枚骰子朝上的点数小于 3”为事件 B，分别写出试验的样本空间，$A+B$，AB，$A\overline{B}$.

解 用一个两位数的十位数字表示第一枚骰子朝上的点

数，个位数字表示第二枚骰子朝上的点数，则试验对应的样本空间为

$$\Omega=\{11，12，13，14，15，16，21，22，23，24，25，26，31，32，33，34，35，36，41，42，43，44，45，46，51，52，53，54，55，56，61，62，63，64，65，66\}，$$

共 36 个样本点组成.

$$A=\{11，12，13，14，15，21，22，23，24，31，32，33，41，42，51\}，$$

共 15 个样本点组成.

$$B=\{11，12，13，14，15，16，21，22，23，24，25，26\}，$$

共 12 个样本点组成.

$$A+B=\{11，12，13，14，15，16，21，22，23，24，25，26，31，32，33，41，42，51\}，$$

共 18 个样本点组成.

$$AB=\{11，12，13，14，15，21，22，23，24\}，$$

共 9 个样本点组成.

$$A\overline{B}=\{31，32，33，41，42，51\}，$$

共 6 个样本点组成.

练习

1. 运动员进行一次射击，试判断下列事件哪些是互斥事件？哪些是对立事件？
   记“命中环数大于 7 环”为事件 A；
   记“命中环数为 10 环”为事件 B；
   记“命中环数小于 6 环”为事件 C；
   记“命中环数为 6，7，8，9，10 环”为事件 D.
2. 投掷一枚骰子，观察其朝上的点数. 记“朝上的点数为奇数”为事件 A；记“朝上的点数为偶数”为事件 B；记“出现点数 4，5，6”为事件 C；记“出现点数 1，2，3”为事件 D. 请写出下列事件：

$$A，B，AC，B+D，\overline{A}.$$

1.1.3 频率与概率

若某项试验重复进行 N 次，其中事件 A 发生了 n 次，则称

$$f_N(A)=\frac{n}{N}$$

为这 N 次试验中事件 A 发生的**频率**(frequency).

为了认识频率与概率的关系，历史上人们做过大量的研究. 下表给出的是历史上著名的投掷硬币的试验：

试验者	投掷硬币次数	出现正面次数	频率
棣莫弗(De Moiver)	2 048	1 061	0.518 1
蒲丰(Buffon)	4 040	2 048	0.506 9
费勒(Feller)	10 000	4 979	0.497 9
卡尔·皮尔逊(Pearson)	12 000	6 019	0.501 6
卡尔·皮尔逊(Pearson)	24 000	12 012	0.500 5

试验结果表明，在大量重复投掷硬币试验中，正面朝上的频率稳定在常数 0.5 附近.

一般地，在大量重复某一试验时，事件 A 发生的频率 $\frac{n}{N}$ 总是接近于某个常数，并在它附近摆动. 试验次数越多，这种摆动的幅度就越小，这就是频率的稳定性. 这个常数就是事件 A 发生的概率. 频率的稳定性告诉我们，做大量试验后，可用频率近似代替概率. 频率是概率的一种表现形式.

投掷一枚骰子，投掷出 1 点的概率为 $\frac{1}{6}$. 这个概率 $\frac{1}{6}$ 该如何理解呢?

事实上，对于一个随机事件来说，它发生的概率是由它自身决定的，而且是客观存在的，它不随试验次数的变化而变化. 也就是说，如果我们投掷 6 次骰子，即使 1 点一次也没有出现过或者连续出现过两次，都与上述概念毫不相悖，而当投掷次数足够大时，“出现 1 点”的频率会非常接近 $\frac{1}{6}$. 因此，概率可以通过频率来“测量”，概率是频率的稳定值.

例1 在甲、乙两名围棋选手近期的 10 局比赛中，甲胜 6 局. 预计下一局甲胜的概率有多大?

解 每局比赛均为一次随机试验. 设“甲获胜”为事件 A，那么，在近期重复的 10 次试验中，事件 A 发生的频率 $f_{10}(A)=\frac{6}{10}=0.6$. 根据频率估计概率的思想，可预计下一局比赛甲获胜的概率约为 0.6.

例2 渔民有什么方法能方便且快速地知道自己鱼池中有多少鱼呢? 有经验的渔民常用一种称为“标记后再捕”的方法. 先从鱼池中随机捕捉一些鱼上来，比如说捕到1 000条鱼，在每条鱼的身上作记号(不影响其存活)，又放回鱼池中. 经过一段时间以后，又从鱼池中随机捕捉一些鱼，比如说第二次捕到 200 条，看其中有记号的鱼有多少条，如果 10 条有记号，那么渔民就会估计出鱼池里的鱼大约为20 000条. 你知道渔民是怎样估计出来的吗?

解 200 条鱼中有 10 条是有记号的，如果鱼池中鱼的分布是均匀的，那么每条有记号的鱼被捕到的可能性的大小是相等的. 根据频率估计概率的思想，每条有记号的鱼被捕捉到的概率是

$$P\approx\frac{10}{200}=\frac{1}{20}.$$

若湖中有 n 条鱼，其中 1 000 条是有记号的，则每条有记号的鱼被捕到的概率是

$$P\approx\frac{1\,000}{n},$$

所以

$$\frac{1\,000}{n}\approx\frac{1}{20},$$

解得

$$n\approx 20\,000.$$

练习

1. 事件 A 发生的频率 $f_N(A)$是不是不变的? 事件 A 发生的概率 $P(A)$是不是不变的? 它们之间有什么区别和联系?
2. 如果某种彩票的中奖概率为 0.001，那么买 1 000 张彩票一定能中奖吗?

习题 1.1

1. 甲、乙两个盒子中分别装有标号为 1，2，3，4 的四个小球，现从甲、乙两个盒子中各取出 1 个小球，有哪些样本点？用集合符号写出试验的样本空间.

2. 设 A，B 为两个事件，用它们的运算关系表示下列事件：

（1）A，B 同时发生；

（2）A 不发生且 B 发生；

（3）A，B 都不发生.

3. 某班选一名学生当代表. 记“选到的学生为男生”为事件 A，记“选到的学生喜欢体育运动”为事件 B，记“选到的学生喜欢数学”为事件 C. 请给出下列事件的含义：

（1）AB；　　（2）$A\overline{B}$；　　（3）$A+C$；　　（4）$\overline{A}$.

4. 甲、乙两支足球队在近 10 场比赛中的成绩如下：

	胜	平	负
甲	4	3	3
乙	3	3	4

不考虑球员变动因素，预计下一场比赛中甲队胜、平、负的概率分别为多大.

课题学习

数学探究

——同时投掷两枚硬币试验

随机事件在一次试验中是否发生具有不确定性，但是，在相同条件下的大量重复试验中，它发生的频率会呈现出一定的规律性. 为了加深大家对这一问题的认识，我们开展如下探究活动.

探究课题

同时投掷两枚硬币，会出现“两个正面”“两个反面”“一正一反”三种可能的结果，小明认为出现“一正一反”

的概率为$\frac{1}{3}$. 你觉得小明的说法正确吗?

探究目的

通过投掷硬币试验，加深大家对概率的认识.

动手操作

同时投掷两枚硬币，大量重复试验时，我们观察出现“一正一反”的频率的变化情况.

(1) 每人重复 20 次，记录下“一正一反”出现的次数.

(2) 汇总每个人所得的数据，并将每个人的数据进行编号，分别得出前 20 次、前 40 次、前 60 次……试验出现“一正一反”的频率.

(3) 在平面直角坐标系中，横轴表示同时投掷两枚硬币的次数，纵轴表示以上试验得到的频率，将上面的计算结果表示在坐标系中.

(4) 从图上观察出现“一正一反”的频率的变化趋势，出现“一正一反”的频率是否稳定在小明认为的$\frac{1}{3}$附近? 你能从中得出什么结论呢?

通过上面的试验，我们可以看出，出现“一正一反”的频率是一个变化的量，但是，在大量重复试验时，它又具有“稳定性”——在一个“常数”附近摆动.

思考交流

在上面的同时投掷两枚硬币的试验中，随着试验次数的增加，出现“一正一反”的频率在某个“常数”附近的摆动幅度是否一定越来越小?

结论概括

(1) 在大量重复试验的情况下，出现“一正一反”的频率会呈现出稳定性，即频率在一个“常数”附近摆动. 随着试验次数的增加，摆动的幅度具有越来越小的趋势.

(2) 有时候试验也可能出现频率偏离“常数”较大的情形，但总的来说，随着试验次数的增加，频率偏离“常数”的幅度呈减小趋势.

1.2 古典概型

1.2.1 古典概型

我们先看以下两个问题：

问题 1 一批产品有 N 件，其中有 M 件不合格产品. 从中随机取出一件，求取到不合格产品的概率.

分析 从中随机取出一件，有 N 个样本点，其中每件产品被取出的可能性是相等的，即每个样本点都等可能出现. 于是可以得到所求概率为$\frac{M}{N}$，即不合格产品数 M 与产品总数 N 之比.

问题 2 投掷一枚骰子，求出现奇数点的概率.

分析 根据经验，投掷一枚骰子，试验对应的样本空间为

$$\Omega=\{1，2，3，4，5，6\}，$$

共 6 个样本点，且每个样本点都等可能出现，因此，所求概率为$\frac{3}{6}$，即投掷骰子试验中所有可能出现奇数点的个数与所有可能出现的点数的个数之比.

分析上述问题中的随机试验，我们可以发现，它们具有如下两个方面的共同点：

(1) 试验中所有可能出现的样本点只有有限个(有限性)；

(2) 每个样本点出现的可能性相等(等可能性).

我们称这种样本点总数有限且每个样本点等可能发生的概率模型为**古典概型**(classical model of probability).

古典概型是概率论中最早讨论的一类最基本的概率模型，它在实际生活中有着广泛的应用. 仔细体会问题 1、问题 2 中求概率的计算过程，不难发现，对古典概型中某事件发生的概率计算可用一个公式来描述.

设 Ω 是古典概型所对应的样本空间，它包含 n 个样本点. $A\subseteq\Omega$ 是一随机事件，它包含 m 个样本点，则事件 A 发生的

概率为

$$P(A)=\frac{m}{n}=\frac{A\text{ 所含的样本点数}}{\Omega\text{ 所含样本点总数}}.$$

例1 投掷两枚骰子，分别求出现的点数之和为 7，10，11 的概率.

解 由 1.1.1 节的例 2，我们得到对应试验的样本空间 Ω 包含 36 个样本点. 因为骰子是均匀的，所以每个样本点发生的可能性相同，这是一个古典概型的问题.

分别记出现的点数之和为 7，10，11 的事件为 A，B，C，则

$$A=\{16，61，25，52，34，43\}，$$
$$B=\{46，64，55\}，$$
$$C=\{56，65\}.$$

由古典概型的计算公式，得

$$P(A)=\frac{6}{36}=\frac{1}{6}，$$
$$P(B)=\frac{3}{36}=\frac{1}{12}，$$
$$P(C)=\frac{2}{36}=\frac{1}{18}.$$

例2 从不超过2 000的正整数中任取一数，

（1）求此数能被 8 整除的概率；

（2）求此数能同时被 4 和 6 整除的概率.

解 从不超过2 000的正整数中任取一数的可能结果有2 000个，且每一个数被取到的可能性相同，因此这是一个古典概型的问题.

记“所取到的数能被 8 整除”为事件 A，“所取到的数能同时被 4 和 6 整除”为事件 B.

（1）因为$\frac{2\,000}{8}=250$，得到能被 8 整除的数有 1×8，2×8，…，250×8，共 250 个，所求概率为

$$P(A)=\frac{250}{2\,000}=\frac{1}{8}.$$

（2）能同时被 4 和 6 整除等价于能被它们的最小公倍数 12 整除.

因为 $166<\frac{2\,000}{12}<167$，所以，能被 12 整除的数有 12×1，12×2，…，12×166，共 166 个，所求概率为

$$P(B)=\frac{166}{2\,000}=\frac{83}{1\,000}.$$

例 3 将一部四册的文集(编号为 1，2，3，4)按任意次序放到书架上，问：各册自右向左或自左向右恰成 1，2，3，4 的顺序的概率是多少?

解 用四位数表示自左向右排列的书的编号，则所有可能的排法组成的样本空间为

$\Omega=\{1234，1243，1324，1342，1423，1432，2134，$
$2143，2314，2341，2413，2431，3124，3142，$
$3214，3241，3412，3421，4123，4132，4213，$
$4231，4312，4321\}.$

样本空间 Ω 包含 24 个样本点，是一个有限样本空间. 由于我们是按任意次序把书放到书架上的，即对每种放法没有任何特别的关注，因此，所有放法(样本点)是等可能发生的，这是一个古典概型的问题.

记“各册自右向左或自左向右恰成 1，2，3，4 的顺序”为事件 A，则 $A=\{1234，4321\}$，含 2 个样本点，所以

$$P(A)=\frac{2}{24}=\frac{1}{12}.$$

练习

1. 有一本书，甲、乙两位同学都想看. 甲同学提议：在一个不透明的箱子里放 4 个标号为 1，2，3，4 的球，充分搅拌后随机取出一个球，若取到标号为偶数的球则甲先看，否则乙先看. 而乙同学提议：投掷一枚骰子，若出现点数 1，2，3 则甲先看，否则乙先看. 他们的提议是否公平?
2. 连续投掷一枚硬币三次，写出此试验的样本空间，并求恰好出现一次正面朝上的概率.
3. 从不超过 100 的正整数中任取一个，求取到的数含有数字 5 的概率.
4. 已知 5 件产品中有 2 件次品和 3 件合格品，现从这 5 件产品中任取 2 件，求恰有一件次品的概率.
5. 袋中装有 6 个球，其编号分别为 1，2，3，4，5，6. 现每次随机地摸出 1 球.
   (1) 若有放回地摸 2 次，求摸出的球的编号之和为 7 的概率；
   (2) 若不放回地摸 2 次，求摸出的球的编号之和为 7 的概率；
   (3) 若不放回地摸 2 次，求摸出的球的编号的奇偶性不同的概率.

1.2.2 应用实例

我们常常看到：在围棋比赛中棋手比赛前的猜先，球类比赛中分组的抽签及比赛前场地的挑选，电视挑战赛的选题抽签等. 现实生活中，人们处理一些机会性的棘手问题时也常常采用这种抓阄、抽签的方法. 这种抽签的方法能够被人们认同和采用，一定有它的合理性. 下面我们利用概率的知识从理论上证明这种合理性(即对各方来说机会是公平的).

我们先来看一种简单情况.

抽签问题中的“好签”和“坏签”可以代表不同的内容. 如：在产品抽样问题中，“好签”可以理解为“合格品”，“坏签”可以理解为“不合格品”；在抽取题目的问题中，“好签”可以理解为难度低的题目，“坏签”可以理解为难度高的题目；在彩票问题中，“好签”可以理解为中奖的号，“坏签”可以理解为没有中奖的号等.

在抽签过程中抽到好签的概率与抽签顺序无关，因此对每个人都是公平的，这也从理论上说明了现实生活中抽签和抓阄方式被广泛应用的原因.

例1 设袋中有 2 支好签，2 支坏签，4 个人依次从袋中不放回地任取一签，分别求他们取到好签的概率.

解 分别用数字 1，2 表示两支好签，用数字 3，4 表示两支坏签，则 4 个人依次抽签对应的样本空间与 1.2.1 节例 3 的样本空间一样.

因为每人抽取是随机的，所以每个可能结果发生的可能性相同，这是一个古典概型的问题.

分别记第 1 个人、第 2 个人、第 3 个人、第 4 个人抽到好签为事件 A_1，A_2，A_3，A_4，则

$$\begin{aligned}A_1=\{&1234，1243，1324，1342，\\&1423，1432，2134，2143，\\&2314，2341，2413，2431\}，\end{aligned}$$

共 12 个样本点.

$$\begin{aligned}A_2=\{&3142，3124，2134，2143，\\&4132，4123，3241，3214，\\&1234，1243，4213，4231\}，\end{aligned}$$

共 12 个样本点.

类似地，我们可列举出 A_3，A_4，也都包含有 12 个样本点.

由古典概型的计算公式，可得

$$P(A_1)=P(A_2)=P(A_3)=P(A_4)=\frac{12}{24}=\frac{1}{2}.$$

计算结果说明抽到好签的概率与抽签顺序无关，抽签方法是公平的.

从上面的计算可以看出，对于第 $i(i=1，2，3，4)$个人来说，他抽 4 支签中的每支签都有可能，因此，我们也可以将样本空间定为

$$\Omega=\{1，2，3，4\}，$$

则

$$A_i=\{1，2\}(i=1，2，3，4)，$$

所以

$$P(A_i)=\frac{2}{4}=\frac{1}{2}.$$

利用同样的思路，同学们可以探究一般的情况.

例2 银行一般要求储户为自己的银行卡设置一个密码，这样可以防范一旦银行卡丢失或被盗后存款被冒领. 一般密码为从 0，1，2，3，…，9 这 10 个数字中选取若干个组成. 若已设置一个六位数字的密码，一旦银行卡丢失，求密码在一次试验中就被破译的概率.

解 从 0 到 9 这 10 个数字中选取 6 个数字组成密码，一个密码相当于一个基本事件，对应的样本空间为 $\Omega=\{000000，000001，000002，000003，\cdots，999998，999999\}$，共有 10^6 个样本点. 输入一次密码，即从这 10^6 种可能的密码中任选 1 种，每种可能的密码会以相同的可能性被选到，因此这是一个古典概型的问题. 事件“密码在一次试验中就被破译”由一个基本事件构成，即由正确的密码构成，所以密码在一次试验中就被破译的概率为 $\frac{1}{10^6}$.

显然，密码的数字位数设置越多，一旦银行卡丢失，密码被破译的可能性就越小. 我国银行卡的密码以前只要求选取 4 位数字组成，现在则要求选取 6 位数字组成，这样就大大增加了破译的难度，起到了更好的保密效果.

以上的讨论也可以用来考虑电话号码问题. 电话号码的升位一方面可以增加用户容量，另一方面也可以减少用户被骚扰的可能性.

例3 有 6 名同学利用假期参加义工活动，已知参加义工活动次数为 1，2，3 的人数分别为 1，2，3，现从这 6 人中随机选出 2 人作为代表参加座谈会.

(1) 记事件 A 为“选出的 2 人参加义工活动次数之和为 4”，求事件 A 发生的概率；

(2) 记事件 B 为“选出的 2 人参加义工活动次数之差的绝对值为 1”，求事件 B 发生的概率.

解 分别用 a，b，c，d，e，f 表示这 6 名同学，其中 a 参加义工活动 1 次，b，c 参加义工活动 2 次，d，e，f 参加义

工活动 3 次.

从 6 人中随机选出 2 人，则所有的选法组成的样本空间是

$$\Omega=\{ab，ac，ad，ae，af，bc，bd，be，bf，cd，ce，cf，de，df，ef\}，$$

共有 15 个样本点.

（1）事件 A 发生，有两种情形：选出的 2 人参加义工活动的次数都是 2；选出的 2 人参加义工活动的次数分别为 1，3. 故 $A=\{bc，ad，ae，af\}$，含有 4 个样本点，所以

$$P(A)=\frac{4}{15}.$$

（2）事件 B 发生，有两种情形：选出的 2 人参加义工活动的次数分别为 1，2；选出的 2 人参加义工活动的次数分别为 2，3. 故 $B=\{ab，ac，bd，be，bf，cd，ce，cf\}$，含有 8 个样本点，所以

$$P(B)=\frac{8}{15}.$$

练习

1. 在 3 张奖券中，有一等奖、二等奖、无奖各 1 张. 甲、乙两人各抽取 1 张，两人都中奖的概率是(　　).

A. $\frac{1}{2}$　　B. $\frac{2}{3}$　　C. $\frac{3}{4}$　　D. $\frac{1}{3}$

2. 从 1，2，3，4 四个数字中任取三个组成一个三位数，求该数是偶数的概率.

习题 1.2

1. 同时投掷一枚硬币和一枚骰子，求下列事件的概率：
（1）硬币出现正面；　　（2）硬币出现正面且骰子出现偶数点；
（3）骰子出现的点数不小于 3；　　（4）硬币出现反面且骰子出现的点数大于 4.
2. 从不超过 1 000 的正整数中任取一个，求：
（1）该数能被 5 整除的概率；　　（2）该数能同时被 2 和 3 整除的概率.
3. 从数字 1，2，3 中可重复地任选 3 个，组成一个三位数. 写出试验的样本空间 Ω，并分别求下列事件的概率：
（1）可组成一个无重复数字的偶数；（2）可组成一个有重复数字的偶数.
4. 从 2 位女生、3 位男生中选 2 人参加志愿者活动，求至少有一位女生入选的概率.
5. 在分别写有 2，4，6，7，11，12 的六张卡片中，任意取出两张，分别作为分数的分子和分母，求所取的两个数组成的分数不是最简分数的概率.
6. 甲、乙两人玩“剪刀、石头、布”的游戏，求：
（1）平局的概率；　　（2）甲赢的概率；　　（3）乙赢的概率.

1.3 概率的加法公式

本节我们讨论概率的一些简单运算性质.

例1 投掷两枚骰子，求出现点数之和为 7 或 11 的概率.

解 用一个两位数的十位数字表示第一枚骰子出现的点数，个位数字表示第二枚骰子出现的点数，则试验所对应的样本空间为

$$\Omega=\{11,\ 12,\ 13,\ 14,\ 15,\ 16,\ 21,\ 22,\ 23,\ 24,\ 25,\ 26,\ 31,\ 32,\ 33,\ 34,\ 35,\ 36,\ 41,\ 42,\ 43,\ 44,\ 45,\ 46,\ 51,\ 52,\ 53,\ 54,\ 55,\ 56,\ 61,\ 62,\ 63,\ 64,\ 65,\ 66\},$$

共 36 个样本点组成.

分别记点数之和为 7 和 11 的事件为 A 和 B，则

$$A=\{16,\ 61,\ 25,\ 52,\ 34,\ 43\},$$

$$B=\{56,\ 65\},$$

$A+B$ 表示出现点数之和为 7 或 11 的事件，即

$$A+B=\{16,\ 61,\ 25,\ 52,\ 34,\ 43,\ 56,\ 65\},$$

共 8 个样本点组成.

由于骰子是均匀的，每种可能结果的发生是等可能的，这是一个古典概型问题. 由古典概型概率公式，可得

$$P(A+B)=\frac{8}{36}=\frac{2}{9}.$$

而 $P(A)=\frac{6}{36}=\frac{1}{6}$，$P(B)=\frac{2}{36}=\frac{1}{18}$，所以

$$P(A)+P(B)=\frac{6}{36}+\frac{2}{36}=\frac{2}{9}.$$

在一次试验中出现的点数之和不可能同时为 7 和 11，即事件 A，B 是互斥事件.

这个简单的例子表明：两个互斥事件的和的概率等于它们各自概率的和，即

$$P(A+B)=P(A)+P(B).$$

这个性质对古典概型都是正确的.

如果 A_1，A_2，A_3 中任意两个都是互斥事件，那么 $P(A_1+A_2+A_3)=P(A_1)+P(A_2)+P(A_3)$ 是否成立？如果 A_1，A_2，…，A_n $(n\geqslant 2)$ 中任意两个都是互斥事件，则 $P(A_1+A_2+\cdots+A_n)=P(A_1)+P(A_2)+\cdots+P(A_n)$ 是否成立？

在古典概型中，设 A，B 是两个互斥事件，则

$$P(A+B)=P(A)+P(B). \qquad \text{（公式 1）}$$

公式 1 叫作**两个互斥事件的概率加法公式.**

对于两个对立事件 A，$\overline{A}$，有如下重要结论：

$$P(\overline{A})=1-P(A). \qquad \text{（公式 2）}$$

公式 2 给出了一种计算概率的方法，若直接计算事件 A 的概率较困难时，我们可以考虑 A 的对立事件 $\overline{A}$，先计算出事件 $\overline{A}$ 的概率，然后由这个公式求得 $P(A)$.

例 2 设袋中有 4 个白球、3 个黑球、1 个红球、2 个蓝球，从袋中取出一个球，求该球是白球或黑球的概率.

解 记“从袋中取出一个球，是白球”为事件 A，“从袋中取出一个球，是黑球”为事件 B，则

$$\begin{aligned}P(A+B)&=P(A)+P(B)\\&=\frac{4}{4+3+1+2}+\frac{3}{4+3+1+2}\\&=\frac{7}{10}.\end{aligned}$$

例 3 投掷两枚硬币，求至少出现一个正面的概率.

解 投掷两枚硬币共有四个等可能样本点：

（正，正），（正，反），（反，正），（反，反）.

记“至少出现一个正面”为事件 A，事件 A 包含其中前三个样本点，故

$$P(A)=\frac{3}{4}.$$

我们再来考察事件 A 的对立事件 $\overline{A}$，它的含义是：

投掷两枚硬币，都出现反面.

它只含有一个样本点(反，反)，于是，$P(\overline{A})=\frac{1}{4}$，故

$$P(A)=1-P(\overline{A})=1-\frac{1}{4}=\frac{3}{4}.$$

由此可见，两种不同思路获得相同的结果，但相比之下，后一种思路更容易实现，因为事件 $\overline{A}$ 所含的样本点比事件 A 所含的样本点要少一些，从而计算也更容易些，这种思路在投

用 1 减去对立事件的概率：大多数情况下，计算“至少有 n 次”的事件的概率，只要用 1 减去事件“至多 $n-1$ 次”的概率即可.

掷更多枚硬币的时候，更容易显现出其方便的特点.

例如，在投掷五枚硬币时，至少出现一个正面(记为事件 B)的概率为

$$P(B)=1-P(\overline{B})=1-\frac{1}{2^5}=\frac{31}{32}.$$

因为对立事件 $\overline{B}$ 表示五枚硬币都出现反面，它在 $2^5=32$ 个等可能结果中仅含其中一个，故很容易算得 $P(\overline{B})=\frac{1}{2^5}$.

练习

1. 袋中有 1 个白球，1 个红球，2 个黑球，从中随机取出一个球，记“从中取出一个球，是白球”为事件 A，记“从中取出一个球，是黑球”为事件 B. 请问：事件 A 和 B 是否为互斥事件？是否为对立事件？
2. 投掷编号为 1，2 的两枚骰子，记“1 号骰子出现 2 点”为事件 A，记“2 号骰子出现 3 点”为事件 B. 判断下列各题中的两个事件是否为互斥事件：
   (1) A 与 B；　(2) A 与 $\overline{A}B$；　(3) B 与 $A\overline{B}$；　(4) A 与 AB.
3. 从编号为 1，2，3，4，5 的 5 个球中取出 4 个球，求 1 号球被取出的概率.

习题 1.3

1. 袋中有编号为 1，2，3，4，5，6 的 6 个球，从中有放回地取出 3 个球，求取出的球的最大编号为 4 的概率.
2. 从不超过 1 000 的正整数中任意取一个，求：
   (1) 该数能被 2 或 3 整除的概率；
   (2) 该数不能被 6 整除的概率.
3. 某班要进行一次辩论比赛，现将 4 位男生和 2 位女生随机分成甲、乙两个辩论小组，每组 3 人. 考虑甲组的人员组成情况，记“甲组有 k 个女同学”为事件 A_k($k=0$，1，2)，
   (1) 写出样本空间及事件 A_0，A_1，A_2；
   (2) 写出甲组至少有一位女生的事件；
   (3) 计算 $P(A_0)$，$P(A_1)$，$P(A_2)$及甲组至少有一位女生的概率.

阅读与讨论

高 尔 顿 板

古典概型试验的简单工具——硬币和骰子，我们已多次谈到过. 通过很多次的投掷硬币或骰子，能直观地帮助我们认识概率的概念，它们的基础是等可能性. 这里我们再介绍一种有趣的装置——高尔顿板. 它是以英国数学家、统计学家高尔顿(1822—1911)的名字命名的，它比硬币或骰子结构复杂些，却能使我们对大量重复试验的统计规律有更深的体会.

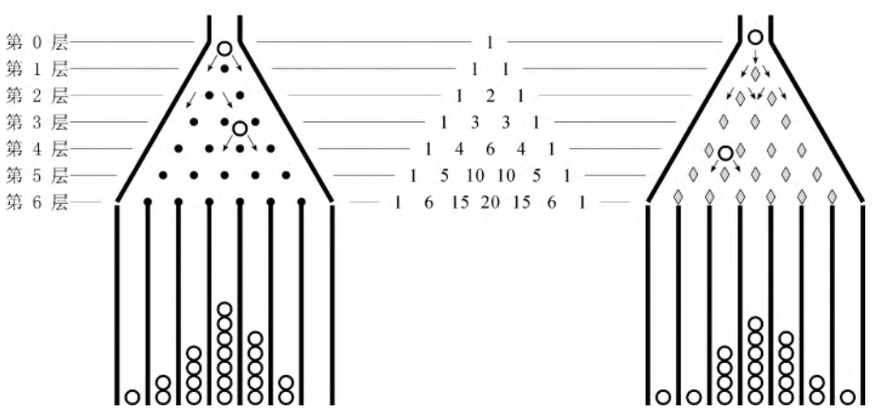


图 1

图 1 是高尔顿板的两种常见形式. 设想一个球从装置上端入口处投入，当这个球碰到板上任何一颗钉(如左图)或任何一个菱形的朝上的顶端(如右图)，它向左边或右边通道滚下的可能性相等. 当我们投入大量的球，多次重复这样的试验(可用计算机模拟)，试验结果表明，从高尔顿板下端各出口滚出球的个数的分布大致相同，具有一定的规律性.

下面我们作一些讨论. 设想投入了很多球，那么从第 1 层的两个通道滚出球的个数之比接近 1∶1. 对从任何一颗钉或菱形上顶端的两通道滚下的球的数目比均如此.

为便于理解，下面我们假设总是出现“理想”的情况(即碰到同一颗钉或菱形上顶端时，每 2 个球中各 1 个分别从左、右通道滚下).

如果从入口处投入 2 个球，那么从第 1 层左右通道各滚出 1 个.

如果从入口处投入 4 个球，那么从第 1 层各通道滚下的球各 2 个;其中左通道中 2 个球又各 1 个分别从第 2 层左起第 1，2 通道滚下，而右通道中 2 个球又各 1 个从第 2 层的左起第 2，3 通道滚下. 于是从第 2 层左起各通道滚出的球各为 1，2，1 个.

如果从入口处投入 8，16，32，64，…个球，“理想”情况分别如何？分析一下，你的结论与图 1 中的结论一致吗？

图 1 中的那个数字组成的三角形叫作杨辉三角，它有许多奇妙的性质，最简单而易被发现的是:从第 2 层起，各层左起第 2 个数到右起第 2 个数，每一个数等于它们左右斜上方两个数的和.

如果我们把杨辉三角的第 n 层除以 2^n，得到如下的数字三角形：

第 0 层——————————————————— 1
第 1 层——————————————————— 0.5　　0.5
第 2 层——————————————————— 0.25　　0.5　　0.25
第 3 层——————————————————— 0.125　　0.375　　0.375　　0.125
第 4 层——————————————————— 0.062 5　　0.25　　0.375　　0.25　　0.062 5
第 5 层——————————————————— 0.031 25　　0.156 25　　0.312 5　　0.312 5　　0.156 25　　0.031 25
第 6 层——————————————————— 0.015 625　　0.093 75　　0.234 375　　0.312 5　　0.234 375　　0.093 75　　0.015 625

图 2

这个数字三角形的各行恰与小球从该行自左至右各出口滚出的概率分布一致. 例如，一个球从第 3 层左起第 3 通道滚出的概率是 0.375.

讨论题

通过高尔顿板试验，你对随机现象有何认识？

1.4 随机事件的独立性

在 1.1.1 节例 2 投掷两枚骰子的试验中，记“第一枚骰子出现 1 点”为事件 A，“第二枚骰子出现偶数点”为事件 B. 很明显，第一枚骰子出现的点数是否为 1 对第二枚骰子出现的点数是否为偶数的概率没有影响.

事件 A 和事件 B 发生的概率分别是

$$P(A)=\frac{6}{36}=\frac{1}{6},\ P(B)=\frac{18}{36}=\frac{1}{2},$$

而这两个事件同时发生的事件 AB 含有三个样本点：

$$(1,\ 2),\ (1,\ 4),\ (1,\ 6),$$

故 $P(AB)=\frac{3}{36}=\frac{1}{12}$，于是有等式 $P(AB)=P(A)P(B)$.

一般地，对任意两个事件 A 与 B，若有

$$P(AB)=P(A)P(B),$$

则称事件 A 与 B 相互独立，简称 A **与** B **独立**. 否则称事件 A 与 B 不独立.

可以证明如下性质：

若事件 A 与 B 独立，则事件 A 与 $\overline{B}$ 独立，$\overline{A}$ 与 B 独立，$\overline{A}$ 与 $\overline{B}$ 独立.

当事件 A 与 B 相互独立时，事件 A(或 B)是否发生对事件 B(或 A)发生的概率没有影响.

例 1 制造一种零件，甲机床的正品率是 0.96，乙机床的正品率是 0.95，从它们制造的产品中任意各抽一件. 求：

(1) 两件产品都是正品的概率；

(2) 恰有一件产品是正品的概率.

解 (1) 记“从甲机床制造的产品中抽得正品”为事件 A，“从乙机床制造的产品中抽得正品”为事件 B，则两件都是正品的概率为

$$\begin{aligned}P(AB)&=P(A)P(B)\\&=0.96\times0.95\\&=0.912.\end{aligned}$$

(2) 记“恰有一件是正品”为事件 C，则 $C=A\overline{B}+\overline{A}B$.

由题意知，A，B 是相互独立事件，故恰有一件是正品的概率为

$$\begin{aligned}P(C)&=P(A\overline{B})+P(\overline{A}B)\\&=P(A)P(\overline{B})+P(\overline{A})P(B)\\&=0.96\times(1-0.95)+(1-0.96)\times0.95\\&=0.086.\end{aligned}$$

例2 甲、乙两人约定在某地会面，记“甲准时到达”为事件 A，记“乙准时到达”为事件 B，并假设事件 A 与事件 B 相互独立. 若 $P(A)=0.95$，$P(B)=0.70$，求：

(1) 甲、乙都准时到达的概率；

(2) 甲、乙都未准时到达的概率；

(3) 甲、乙仅有一位准时到达的概率.

解 记“甲、乙都准时到达”为事件 C，“甲、乙都未准时到达”为事件 D，“甲、乙仅有一位准时到达”为事件 E，则事件 $C=AB$，事件 $D=\overline{A}\overline{B}$，事件 $E=A\overline{B}+\overline{A}B$.

事件 A 与 B 相互独立，且

$$P(A)=0.95,\quad P(B)=0.70.$$

(1) 甲、乙都准时到达的概率为

$$\begin{aligned}P(C)&=P(AB)\\&=P(A)P(B)\\&=0.95\times0.70\\&=0.665.\end{aligned}$$

(2) 甲、乙都未准时到达的概率为

$$\begin{aligned}P(D)&=P(\overline{A}\overline{B})\\&=P(\overline{A})P(\overline{B})\\&=(1-0.95)(1-0.70)\\&=0.015.\end{aligned}$$

(3) 甲、乙仅有一位准时到达的概率为

$$\begin{aligned}P(E)&=P(A\overline{B}+\overline{A}B)\\&=P(A\overline{B})+P(\overline{A}B)\\&=P(A)P(\overline{B})+P(\overline{A})P(B)\\&=0.95\times0.3+0.05\times0.7\\&=0.32.\end{aligned}$$

例3 甲、乙两人进行乒乓球比赛，每人各局取胜的概率均为$\frac{1}{2}$. 现采用五局三胜制，胜 3 局者赢得全部奖金 C 元. 若前两局比赛均为甲获胜，此时因某种原因比赛终止，两人应如何分配奖金比较合理?

解法一 分配奖金的依据是甲、乙两人获胜的概率.

记“一局比赛甲获胜”为事件 A，则“一局比赛乙获胜”为事件 $\overline{A}$，则 $P(A)=\frac{1}{2}$，$P(\overline{A})=\frac{1}{2}$.

记“继续比赛，最终甲获胜”为事件 B，则

$$B=A+\overline{A}A+\overline{A}\overline{A}A,$$

显然 A 与 $\overline{A}A$、$\overline{A}\overline{A}A$ 互斥，所以

$$\begin{aligned}P(B)&=P(A+\overline{A}A+\overline{A}\overline{A}A)\\&=P(A)+P(\overline{A}A)+P(\overline{A}\overline{A}A)\\&=P(A)+P(\overline{A})P(A)+P(\overline{A})P(\overline{A})P(A)\\&=\frac{1}{2}+\frac{1}{2}\times\frac{1}{2}+\frac{1}{2}\times\frac{1}{2}\times\frac{1}{2}\\&=\frac{7}{8}.\end{aligned}$$

由 $P(B)=\frac{7}{8}$知，乙获胜的概率为 $1-\frac{7}{8}=\frac{1}{8}$.

所以，甲、乙两人应按概率比 7∶1 来分配全部奖金.

解法二 若继续比赛三场，则一定可以决出胜负. 用 a 表示一局比赛中甲获胜，b 表示一局比赛中乙获胜，则剩下 3 局比赛的所有可能结果构成的样本空间为

$$\Omega=\{aaa，aab，aba，abb，baa，bab，bba，bbb\}.$$

由于两人水平相当，因此每种可能结果发生的概率相同，这是一个古典概型的问题.

甲获胜的情况有：aaa，aab，aba，abb，baa，bab，bba.

乙获胜的情况有：bbb.

由古典概型的计算公式，甲、乙获胜的概率分别为$\frac{7}{8}$和$\frac{1}{8}$.

所以，甲、乙两人应按概率比 7∶1 来分配全部奖金.

练习

1. 从一副去掉大小王的扑克牌中任取一张，记“出现黑桃”为事件 A，“出现 K”为事件 B，则 A 与 B 相互独立吗？

2. 甲、乙两人独立地破译同一份密码，他们各自破译出密码的概率分别为 $\frac{1}{3}$，$\frac{1}{4}$，求：

   （1）两人都能破译出密码的概率；

   （2）两人都破译不出密码的概率；

   （3）恰有一人破译出密码的概率；

   （4）至多有一人破译出密码的概率；

   （5）至少有一人破译出密码的概率.

习题 1.4

1. 甲、乙两人各进行一次射击，如果两人击中目标的概率都是 0.6，计算：

   （1）两人都击中目标的概率；

   （2）恰有一人击中目标的概率；

   （3）至少有一人击中目标的概率；

   （4）至多有一人击中目标的概率.

2. 甲、乙两个篮球运动员互不影响地在同一位置投球，命中率分别为 $\frac{1}{2}$ 与 p，且乙投球两次均未命中的概率为 $\frac{1}{16}$.

   （1）求乙投球的命中率 p；

   （2）求甲投球两次，至少命中 1 次的概率；

   （3）若甲、乙两人各投球一次，求两人共命中一次的概率.

3. 某班有两个课外活动小组，其中第一小组有足球票 6 张，排球票 4 张；第二小组有足球票 4 张，排球票 6 张. 甲从第一小组的 10 张票中任抽 1 张，乙从第二小组的 10 张票中任抽 1 张. 求：

   （1）两人都抽到足球票的概率；

   （2）两人中至少有 1 人抽到足球票的概率.

阅读与讨论

彩票中的概率问题

我国发行了“中国福利彩票”和“中国体育彩票”两大系列彩票，为社会福利事业、全民健身场所的修建和我国竞技体育事业募集了大量的资金，同时也“造就”了一批百万元户(指彩票大奖获得者). 同学们可能从电视中看到过多种彩票的抽奖过程，从报纸上了解过抽奖的结果，那么，你知道中大奖的概率有多大吗？在购买彩票时有相对有利的选号方法吗？

有一种7位数体育彩票，从0000000～9999999中选择任意7位自然数进行投注，如果投注号码与开奖号码全部相同且排列一致，就可以中“特等奖”. 下表是这种7位数体育彩票连续100期中奖号码的统计结果(数位的次序从左到右依次为第一位、第二位、……、第七位)：

	号码为0的频数	号码为1的频数	号码为2的频数	号码为3的频数	号码为4的频数	号码为5的频数	号码为6的频数	号码为7的频数	号码为8的频数	号码为9的频数
第一位	10	9	14	11	9	11	9	10	4	13
第二位	7	11	6	8	12	12	11	12	8	13
第三位	13	9	5	10	7	8	16	7	12	13
第四位	7	12	9	13	10	9	10	12	4	14
第五位	10	8	10	9	15	9	12	10	5	12
第六位	13	6	12	7	11	9	7	10	13	12
第七位	11	14	10	7	12	10	9	6	12	9

这种7位数体育彩票的中奖号码是一个七位数，每一位的数字通过摇奖机从0～9这10个数字中随机产生，因此，在正

常情况下，每个数字出现的概率都是$\frac{1}{10}$. 那么，一注彩票中 7 个位置上的数字都选对的概率是$\frac{1}{10^7}$，所以，购买一注彩票，中特等奖的概率为$\frac{1}{10^7}$.

根据频率与概率的关系，大量试验后，频率应接近于概率. 因此，若某个数字在前面出现的频率明显小于其概率，则在后面的摇奖中，频率应倾向于变大，即该数字应倾向于出现. 如上统计表中，第一数位上的数字 8 在统计的 100 期中出现的频率为$\frac{4}{100}<\frac{1}{10}$，我们可以认为在后面的摇奖中第一数位上选择数字 8 应更合理. 类似地，可分析其他位置上的数字. 但这也只是建立在大量重复试验基础上的一个相对合理的选择. 虽然我们可以给出一个相对有利的选号方案，但能中大奖的可能性仍然非常非常之小，因此，每个人都应有一个正常的心态，在为社会福利和体育事业作贡献且不影响正常生活的前提下，留给自己一点期望.

讨论题

仿照上述购买彩票时选号的方法，举一个利用频率的稳定性进行某项决策的实例.

复习题

A 组

1. 有一批含有质量为 5 kg，10 kg，15 kg，20 kg，25 kg，30 kg 的物品，每种质量的物品有 2 件. 从这批物品中依次任选 2 件. 用(x, y)表示选出的第一件物品的质量为 x kg，第二件物品的质量为 y kg 的事件. 写出试验的样本空间及以下事件：

(1) 两件物品的质量相等；

(2) 第一件物品的质量大于第二件物品的质量；

(3) 第二件物品的质量是第一件物品质量的两倍；

(4) 第一件物品的质量比第二件物品的质量少 10 kg.

2. 设样本空间 $\Omega=\{1，2，3，4，5，6，7，8，9，10\}$，事件 $A=\{2，3，4\}$，事件 $B=\{3，4，5\}$，求事件 $\overline{A}\,\overline{B}$.
3. 从不超过 100 的正整数中任取一个，求该数中含有数字 9 的概率.
4. 某同学将写给 4 个人的 4 封信随意装入已写好收信人的信封，求他恰好都装对的概率.
5. 有 3 个兴趣小组，甲、乙两位同学各自参加其中一个小组，每位同学参加各个小组的可能性相同，求这两位同学参加同一个兴趣小组的概率.
6. 从不超过2 000的正整数中任取一个，求：
   （1）该数能被 7 整除的概率；
   （2）该数能同时被 9 和 6 整除的概率；
   （3）该数能被 9 或 6 整除的概率.
7. 从 A，B，C，D 这四名学生中任选两名，分别担任班长、副班长.
   （1）写出试验的样本空间；
   （2）若 A，B 是女生，C，D 是男生，求班长是女生的概率.
8. 甲、乙两名水平相当的乒乓球选手进行七局四胜制的比赛，先胜四局者可赢得全部奖金 a 元. 现已进行了三局(甲胜两局而乙胜一局)，比赛因故中止，问应如何分配奖金？
9. 从 1，2，…，10 这 10 个数中任取 2 个，求：
   （1）取到数字 5 的概率；
   （2）其中一个大于 5 而另一个小于 5 的概率.
10. 一栋楼房有 4 个单元，甲、乙都住在此楼内，求两人恰好住在同一单元内的概率.
11. 某家庭电话，在家里有人时，打进的电话响第 1 声时被接的概率为 0.1，响第 2 声时被接的概率为 0.3，响第 3 声时被接的概率为 0.4，响第 4 声时被接的概率为 0.1，求电话在响前 4 声内被接的概率.
12. 在甲、乙两个盒子中分别装有标号为 1，2，3，4 的四个小球，现从甲、乙两个盒子中各取出一个小球. 求：
    （1）取出的两个小球上的标号为相邻整数的概率；
    （2）取出的两个小球上的标号之和能被 3 整除的概率.

B 组

1. 设样本空间 $\Omega=\{1，2，\cdots，10\}$，事件 $A=\{3，4，6，9\}$，事件 $B=\{1，2，5\}$，
   （1）A，B 是否为互斥事件？
   （2）写出事件 A 的对立事件；
   （3）写出事件 B 的对立事件；
   （4）写出事件 $\overline{A}+B$ 的对立事件.
2. 一个盒子中装有号码为 1，2，…，7 的 7 张标签，现分别按如下两种方式随机地取出两张标签：
   （1）标签的选取是不放回的；
   （2）标签的选取是放回的.
   分别求出上述两种情况下两张标签上的数字为相邻数的概率.
3. 把一个表面涂有颜色的立方体等分为 1 000 个小立方体，在这些小立方体中随机取出一个，求它有两面涂有颜色的概率.
4. 袋中装有 4 个形状大小完全相同的球，编号分别为 1，2，3，4.

（1）从袋中随机取出两个球，求取出的球的编号之和不大于 4 的概率；

（2）先从袋中随机取出一个球，该球的编号为 m，将球放回袋中，然后再从袋中随机取出一个球，该球的编号为 n，求 $n<m+2$ 的概率.

5. 有编号为 A_1，A_2，…，A_{10} 的 10 个零件，测量其直径（单位：cm），得到如下数据：

编号	A_1	A_2	A_3	A_4	A_5	A_6	A_7	A_8	A_9	A_{10}
直径	1.51	1.49	1.49	1.51	1.49	1.51	1.47	1.46	1.53	1.47

其中直径在区间［1.48，1.52］内的零件为一等品.

（1）从上述 10 个零件中，随机抽取 1 个，求这个零件为一等品的概率；

（2）从一等品零件中，随机抽取 2 个，求这 2 个零件直径相等的概率.

6. 某商场推出二次开奖活动，凡购买一定价值的商品可以获得一张奖券. 奖券上有一个兑奖号码，可以分别参加两次抽奖方式相同的兑奖活动. 如果两次兑奖活动的中奖概率都是 0.05，求两次抽奖中以下事件的概率：

（1）都抽到某一指定号码；

（2）恰有一次抽到某一指定号码；

（3）至少有一次抽到某一指定号码.

7. 袋中装有 4 个球（1 个白球，1 个红球，2 个黄球），随机地摸两次，每次摸出 1 个球. 在“不放回”和“有放回”这两种摸球方式下，分别求以下事件的概率：

（1）摸出的 2 球都是黄球；

（2）摸出的 2 球中恰有 1 个是黄球；

（3）摸出的 2 球中至少有 1 个是黄球；

（4）摸出的 2 球颜色不同.

思考与实践

1. 上网查询“概率”的含义，并以小论文的形式谈谈你对“概率”的认识.

2. 搜集诸如抓阄、抽签模型在现实生活中应用的例子，并从概率的角度说明它们的公平性.

第2章 统 计

南丁格尔(1820—1910)

我们经常面临如下的问题：工业生产中，要判断一批产品(如灯泡、弹药)的质量是否符合指定的要求；农业生产中，农技部门在推广一种农作物种子之前必须要了解这批种子的质量；经济活动中，经营管理者需要知道某种商品的定价与供应量之间、定价与需求量之间的关系，决策部门需要为某项决策或措施提供依据；等等.

显然，我们不可能将每个灯泡用来测试，也不可能将所有的种子试种之后再来确定这批种子是否值得推广，经营管理者也不可能拿自己的资金盲目地投资. 可行的办法是：拿一部分灯泡进行测试，拿一部分种子进行试验，对市场进行抽样调查，将测试、试验或调查收集来的数据进行整理、分析，然后进行推断，得出合理的结论.

这种利用部分数据来估计总体的思想正是统计学的基本思想. 如何进行数据的收集、整理、分析与判断，以及如何评价这一判断的合理性，是统计学研究的主要内容.

本章将介绍统计学中几种最基本的数据收集、整理，以及利用收集到的数据对总体进行估计的方法. 通过本章的学习，同学们将会进一步体会到统计在现实生活中的意义和作用，并为以后更深入的学习奠定基础.

2.1 数据获取

2.1.1 总体与样本

统计学是一门关于数据资料的收集、整理、分析、推断和评价的科学，它在社会生活中有着广泛的应用. 下面，我们结合具体问题介绍统计学中的一些基本概念.

问题 1 工厂生产了一批产品，若这批产品的次品率大于某个事先给定的常数 p_0，就视为不合格而不能出厂，若产品的次品率不超过 p_0，就认为是合格的而可以出厂. 因此，在产品出厂前，需要了解这批产品的次品率是多少.

问题 2 某农场在确定某种农作物种子的单位面积播种量之前，需要了解这批种子的发芽率.

问题 3 一批炮弹存放一段时间后，国防部门需要了解炮弹的效果是否起了变化.

问题 4 县教育局希望了解全县中学生对公共卫生知识的了解程度.

在上述问题中，我们所关心的问题都存在一个“全体”，如“一批产品”“这批种子”“一批炮弹”“全县中学生”；也都与其中的具体对象有关，如“一件产品”“一粒种子”“一发炮弹”“一个中学生”. 在统计学中，把所研究对象的全体叫作**总体**(population)，而每一个具体的研究对象叫作**个体**(individual).

问题 1 中，该批产品构成一个总体，而每一件具体的产品即为一个个体；问题 2 中，该批种子构成一个总体，而每一粒种子即为一个个体；问题 3 中，所存放的该批炮弹构成一个总体，而每发炮弹即为一个个体；问题 4 中，该县所有中学生构成一个总体，而每个中学生即为一个个体.

对于总体中的每个个体，我们并不关心其所有的特征，而只是根据研究的目的，关心它的一个或几个方面的特征. 如果

我们将这些特征数量化，就把这些数量叫作描述这些特征的**数量指标**(quantitative index).

例如，在问题 1 中，对每一件产品，我们只关心它是否合格，合格可记为 1，不合格可记为 0；在问题 2 中，我们只关心种子是否发芽，发芽可记为 1，不发芽可记为 0；在问题 3 中，我们只关心炮弹是否失效，失效可记为 1，没失效可记为 0. 上述 0 与 1 均为对应的数量指标；在问题 4 中，关于学生对公共卫生知识了解的程度可按 5 分制表示，也可按 100 分制表示，我们关心的是学生得分的多少，所对应的数也是描述学生了解程度的数量指标. 我们把所关心的数量指标在总体中的概率结构叫作**总体分布**.

如问题 1 中，若产品的次品率为 0.1，从中任取一件产品，取到次品记为 0，取到合格品记为 1，则所关心的数量指标为“0”和“1”. 其概率结构为：取“0”的概率为 0.1，取“1”的概率为 0.9.

前面几个问题中的次品率、发芽率、失效率、学生对公共卫生知识的了解程度的成绩在各个分数段的概率，事先都是未知的，也就是说，总体分布未知. 为了了解总体分布，我们需要做试验、作调查，这就是统计学的一项重要任务.

这里的调查有两种方法：一是全面调查，即对构成总体中的每个个体都进行调查；二是抽样调查，即从总体中抽取一部分进行调查. 全面调查可以获得总体的完全信息，对总体有一个完整的认识，但它常常需要大量的时间和费用，有时可能没有多大意义，甚至是不可能完成的.

如问题 2 中，若对所有种子做试验，则试验完后，这批种子也就不存在了；问题 3 中，若对库存的所有炮弹都做试验，则这批炮弹也就不存在了. 又如一些时间性很强的产品，花大量的时间做全面调查，可能结果出来后，这批产品已经过时，这样，全面调查就失去了意义. 因此，实际工作中常常做的是抽样调查.

从总体中抽取的一定数量的个体组成**样本**(sample)，抽取的个体数叫作**样本容量**(sample size). 为了使样本具有很好的代表性，抽样时必须遵循随机性原则，即抽样过程中每个个体按事先给定的概率被抽到. 如本教材中所研究总体中的每个个体在抽样时都等可能地被取到(等概率抽样).

问题 5　1936 年，美国《文学文摘》杂志根据 1 000 万电话用户和该杂志订户所收回的意见，断言兰登将以 370∶161 的

绝对优势在总统选举中击败罗斯福，但结果是罗斯福当选了. 这使《文学文摘》大丢面子. 究其原因在于，1936 年能装电话和订阅《文学文摘》的人是经济上相对富裕的少数人，而大多数收入不高的选民并未被征询意见. 这违背了抽样的随机性原则，所获得的样本不具备广泛的代表性.

某些保健品的宣传中常有类似如下的广告词：通过追踪调查，该保健品的有效率达到 80%，显著有效率达到 70%. 其实，这种所谓的调查对被调查对象的选取常带有明显的倾向性，例如仅仅考虑对宣传有利的对象. 这违反了抽样的随机性原则，不能反映真实情况. 你的身边还有类似的例子吗？收集一下，并与同学交流.

问题 5 说明，随机性原则是抽样的基本原则，违反这个原则，所进行的抽样调查将会毫无意义.

随机抽样所得样本中包含有总体的未知信息. 统计学的另一重要任务是，通过对样本中包含总体的有关信息进行加工整理(即将分散的信息进行集中，如求样本平均数、众数、中位数、方差、标准差等)，然后对总体进行统计推断. 这一过程我们用图 2-1 表示如下：

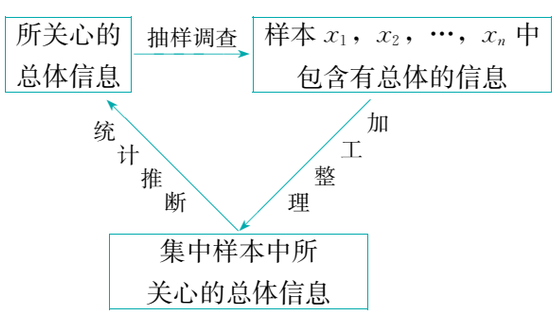


图 2-1

练习

1. 举一个实际生活中的例子，说明该例中的总体、个体、样本、样本容量各指的是什么.
2. 实际调查时为什么常采用抽样调查的方法？
3. 抽样的基本原则是什么？

2.1.2 获取数据的基本途径

各种数据广泛地存在于我们的日常生活中. 新闻媒体中经常出现国家公布的一些经济指标，比如月度居民消费价格指数(Consumer Price Index，CPI)，年度国内生产总值(Gross Domestic Product，GDP)等. 当你对某一问题感兴趣时，想了解与其相关的数据，此时你该怎么办？下面我们就来简单地介绍一下常见获取数据的基本途径和实例，例如行政管理部门的统计报表和年鉴、社会科学研究中的社会调查、人口普查和抽

样、科学实验中的试验设计、互联网数据获取等.

统计报表是指各级政府部门及企事业单位按规定的格式、内容和时间要求，自下而上提供统计资料的一种统计调查方式. 前面提到的定期公布的经济指标通常就是通过统计报表的形式得到的.

另外，各级政府每年都会对上一年的各项数据进行统计汇总. 以年鉴的形式保存各种数据以供政府部门决策和理论研究等各种应用. 年鉴是全面、系统、准确地记录上一年度的重要时事、文献和统计资料，且按年度连续出版的工具书. 年鉴有综合性统计年鉴，也有行业类统计年鉴. 像《中国 2016 年统计年鉴》《湖北省 2016 年统计年鉴》就属于综合性统计年鉴. 而《中国体育年鉴》《湖北教育年鉴 2016 卷》就属于行业类统计年鉴.

社会调查是指由非官方的社会力量(企业、高校、科研单位等)针对社会研究中某些特定的研究问题和研究目标直接收集社会资料或数据的过程与方法. 社会调查的主要对象是社会事务和社会信息. 比如，调查公司关于电视节目收视率的调查.

我国每十年会做一次全国人口普查，最近的一次全国人口普查于 2010 年进行. 人口普查主要关注的是人口总量、家庭户规模、性别构成、年龄构成、民族构成、各种受教育程度人口比例、城乡构成、地区分布、人口的流动等与人口相关的数据和相关的统计指标. 一般情况下普查是指政府为详细调查某项重要的国情和国力而专门组织的全面调查. 普查一般是调查属于一定时间点上的社会经济现象的总量，但也可以调查某段时期现象的总量，乃至调查一些并非总量的指标. 普查涉及范围广，统计指标多，工作量大，时间性强，通常要动用大量的人力、物力以及财力. 因此，普查不可能是常用的数据获取方式. 在大多数情形下，为了便捷地获取相关的数据信息，通常采取抽样的形式. 抽样是从所有个体中选取部分个体进行调查，获取相关数据，据此对整体规律进行推断. 我国在两次人口普查之间会开展 1%人口抽样调查.

试验设计(也称为实验设计)是主要用于经济地、科学地安排试验的一项技术. 试验是人为控制条件下有目的地进行的一种实践活动. 例如比较不同训练方法对运动员的生理指标的影响，不同的教学方法对学生接受新知识的影响. 比较不同配方对产品质量的影响. 在农业生产中比较几种施肥方法的好坏，比较某种农作物不同品种的优劣. 试验的目的是为了获取数

据，然后对数据进行统计分析，从而得到客观且适宜的结论．而用统计方法来分析数据都是在试验数据满足一定条件和假设时才有效．因而试验需要事前进行周密而审慎的设计，使其满足统计方法的要求．试验设计的另一个目的就是用尽可能少量的试验获取尽可能多的信息．

随着信息技术的高速发展，互联网已经融入我们的日常生活．在互联网中，计算机、平板电脑、手机及其他终端设备通过网络互相联接在一起，每个终端设备可以相互共享计算和数据资源．通过互联网在线获取数据是目前非常便捷的常用数据获取方式．通常在互联网上获取数据只需要通过一些网站提供的搜索功能查到相关数据的位置信息，即可直接或者间接地获取数据．

练习

1. 了解你所在地区近五年生产总值的情况．
2. 了解上一年度你所在地区居民的消费水平及结构．

2.1.3 简单随机抽样

前面我们介绍了抽样调查的重要性和必要性，介绍了抽样所必须遵循的基本原则——随机性原则．本节介绍一种最基本的抽样方法——简单随机抽样．

问题 假定我们研究的总体中含有 10 个个体，通过逐个抽取的方法从中抽取容量为 3 的样本．第 1 次随机抽取一个个体，每个个体被抽到的概率都是$\frac{1}{10}$；第 2 次抽取时，在余下的个体中任取一个，每个个体被取到的概率都是$\frac{1}{9}$；第 3 次抽取时，余下的每个个体被抽到的概率都是$\frac{1}{8}$．

问题中，每次抽取时每个个体被抽到的概率都相同．一般地，设一个总体含有 N 个个体，从中逐个不放回地抽取 n 个个体作为样本（$n \leqslant N$），且每次抽取时，总体中余下个体被抽到的概率相等，这样的抽样方式叫作**简单随机抽样**（simple random sampling）．

在问题中，若把先后抽取 3 个个体看成一个完整的过程，那么我们关心的是，在整个抽样过程中，每个个体被抽到的概率是否相等？

例 1 从包含 10 个个体的总体中，按简单随机抽样的方法从中抽取容量为 2 的样本，求总体中任意指定的个体 a 被抽到的概率.

解 记“第 i 次抽取时取到指定的个体 a”为事件 $A_i(i=1,2)$，则由概率中抽签与顺序无关的结果，有

$$P(A_1)=P(A_2)=\frac{1}{10},$$

且 A_1 与 A_2 是互斥事件.

由互斥事件概率的加法公式，有

$$P(A_1+A_2)=P(A_1)+P(A_2)=\frac{1}{10}+\frac{1}{10}=\frac{2}{10}.$$

这个结果与指定的个体 a 无关.

所以，在容量为 2 的样本中，10 个个体每个被抽到的概率都是$\frac{2}{10}$.

一般地，设总体含有 N 个个体，采用简单随机抽样的方法从中抽取容量为 n 的样本，则总体中每个个体被抽到的概率都是$\frac{n}{N}$.

上面的讨论说明了简单随机抽样符合抽样的随机性原则，而且这种抽样方法简单，是其他较为复杂的抽样方法的基础.

怎样实施简单随机抽样而获得样本呢？下面介绍两种常用的方法.

1. 抽签法

先将总体中的个体编号，N 个个体编号为 1，2，…，N. 再把号码写在形状、大小相同的号签上(号签可以用小球、卡片、纸条、棋子等制作)，然后将这些号签放在同一个箱子里，搅拌均匀. 抽签时，按照简单随机抽样的方法，每次从中抽出 1 个号签，连续抽取 n 次. 抽到的签号对应总体中的个体号，这样就得到一个容量为 n 的样本.

对个体编号时，可利用已有的编号. 例如，从全班学生中抽取样本时，可以利用学生的学号、座位号等.

抽签法简单易行，它适用于总体个数不多，且抽取的样本容量较小的情形.

2. 随机数表法

为了简单随机抽样方法的使用方便，统计工作者编制了随机数表. 本书后面的附录是一个随机数表. 表中每个位置上出现 0，1，2，…，9 这十个数码的概率是相等的. 下面举例说明随机数表的使用方法.

例2 从全班 50 位同学中按简单随机抽样的方式抽出 11 位同学作代表. 利用随机数表应如何抽取?

解 利用随机数表，抽取样本可按下面的步骤进行：

第一步，先将 50 位同学编号(也可按座位号或学号)，编为

00，01，02，…，48，49.

第二步，在随机数表中任选一个数作为开始，例如从第 8 行第 9 列的数 5 开始. 为便于说明，此处将附录中随机数表的第 6 行至第 10 行摘录如下：

16 22 77 94 39	49 54 43 54 82	17 37 93 23 78	87 35 20 96 43	84 26 34 91 64
84 42 17 53 31	57 24 55 06 88	77 04 74 47 67	21 76 33 50 25	83 92 12 06 76
63 01 63 78 59	16 95 55 67 19	98 10 50 71 75	12 86 73 58 07	44 39 52 38 79
33 21 12 34 29	78 64 56 07 82	52 42 07 44 38	15 51 00 13 42	99 66 02 79 54
57 60 86 32 44	09 47 27 96 54	49 17 46 09 62	90 52 84 77 27	08 02 73 43 28

总体中的个体编号可以从 0 开始，例如 $N=100$ 时，编号可以是 00，01，…，98，99. 这样总体中的个体均可用两位数号码表示，便于运用随机数表.

第三步，从选定的数 5 开始向右读下去，得到一个两位数字的号码 59，由于 $59>49$，不取，继续向右读，得到 16，将它取出. 如此读下去，相继得到 19，10，12，07，44，39，38，33，21，接下去是 12. 由于 12 前面已经出现，不取，再继续读下去，得到 34. 至此 11 个样本号码已经取满，所得容量为 11 的样本号是

16　19　10　12　07　44　39　38　33　21　34

按这 11 个号就可以抽取 11 位同学作代表.

当随机地选定开始的读数后，读数的方向可以向右，也可以向左、向上或向下.

练习

1. 举例说明简单随机抽样方法如何进行.
2. 将全班学生按座位编号，利用随机数表抽取容量为 10 的样本.

2.1.4 分层抽样

我们来看一个具体问题.

问题 设某大型商场有部门经理及以上人员 20 人，一般员工 180 人. 部门经理及以上人员月平均工资 20 000 元，一般员工月平均工资 3 000 元. 现有某求职者希望到该商场工作，为了了解商场员工的工资状况，随机抽取 10 名人员进行调查. 应如何抽样?

显然，若按简单随机抽样的方法，求职者可能调查到的 10 人全部是部门经理及以上的人员，也可能调查到的 10 人全部是一般员工，这样会得出与实际相差甚远的结论. 那么，应如何设计抽样方法呢?

针对上述问题，下面介绍分层抽样的方法.

若已知总体由差异明显的几部分组成，为了避免可能抽到的样本过多地来源于某一特定部分，从而影响样本的代表性，我们可将总体中的个体按某共同特性分成几个部分，然后按各部分所占的比进行抽样. 这种抽样方式叫作(等比)**分层抽样** (stratified sampling)，其中所分成的各部分叫作层.

上面的问题中，个体总数为 200，抽取的样本容量为 10，样本容量与总体中个体数之比(抽样比)为 $10 : 200 = 1 : 20$，所以在各层抽取的个体数依次为 $20\times\frac{1}{20}$，$180\times\frac{1}{20}$，即 1，9. 于是我们得到抽样方案：按简单随机抽样的方法分别从部门经理及以上人员中抽取 1 名，从一般员工中抽取 9 名.

一般地，设总体中包含的个体数为 N，按总体中个体的某共同特性分成 k 个层，各层分别含个体数为 N_1，N_2，…，N_k. 现从总体中抽取容量为 n 的样本，则抽样比为 $\frac{n}{N}$，按分层抽样的思想，各层的抽样数分别为

$$N_1 \cdot \frac{n}{N},\ N_2 \cdot \frac{n}{N},\ \cdots,\ N_k \cdot \frac{n}{N},$$

即为

$$\frac{N_1 n}{N}=n_1,\ \frac{N_2 n}{N}=n_2,\ \cdots,\ \frac{N_k n}{N}=n_k.$$

第 i 层中任一个个体被抽到的概率为

$$\frac{n_i}{N_i}=\frac{n}{N},\ i=1,\ 2,\ \cdots,\ k.$$

称$\frac{n_i}{N_i}$为第 i 层的抽样比. 此时

$$\frac{n_i}{n}=\frac{N_i}{N},\ i=1,\ 2,\ \cdots,\ k.$$

即总体各层规模之比等于样本各层规模之比.

这一结果表明，各层中的任一个体被抽到的概率都为$\frac{n}{N}$，这也是按简单随机抽样方法从总体中抽到某一个体的概率，所以分层抽样方法符合抽样的随机性原则.

分层抽样充分利用了我们对总体所掌握的信息，有效地避免了简单随机抽样可能带来的抽样误差(如样本中所有个体均来自某个极端的层)，使所抽取的样本有较好的代表性，能更好地反映总体的情况. 而在各层抽样时，可以根据具体情况采用不同的抽样方法(如简单随机抽样方法或其他抽样方法). 所以，分层抽样在实践中有着非常广泛的应用.

下面列表对以上两种抽样方法进行简单的比较：

类别	共同点	各自特点	相互关系	适用范围
简单随机抽样	抽样过程中每个个体被抽取的概率相等	从总体中逐个抽取		总体中的个体数较少，或个体间差异不显著，或不知道个体间差异情况而无法给出分层
分层抽样(等比例)		将总体中个体按某共同特征分成几层，再分层抽取. 总体抽样比与各层抽样比相等	各层抽样时采用简单随机抽样或其他等概率随机抽样	对总体有更多了解，且总体由差异明显的几部分组成

初中阶段已学习过计算样本平均数和样本方差，对于分层抽样的情形，我们既需要计算整体的样本平均数和样本方差，还需要计算各层的样本平均数和样本方差.

例 1 某高校 2017 年招收统计专业新生 56 人，档案显示来自农村家庭学生 32 人，来自城镇家庭学生 24 人. 为了解学生家庭经济情况，班主任按分层抽样方式随机抽取农村家庭学生 4 人，其家庭 2016 年人均可支配收入分别为(单位：万元)：0.93，1.41，1.03，1.71；城镇家庭学生 3 人，其家庭 2016 年人均可支配收入分别为(单位：万元)：3.51，3.07，3.32. 求分层抽样的样本平均数和样本方差.

解 设被抽到的 4 名农村家庭学生家庭 2016 年人均可支配收入分别记为 x_{11}，x_{12}，x_{13}，x_{14}，被抽到的 3 名城镇家庭学生家庭 2016 年人均可支配收入分别记为 x_{21}，x_{22}，x_{23}. (单位：万元)

农村家庭(第一层)：

$$\bar{x}_1=\frac{1}{4}\sum_{i=1}^{4}x_{1i}=1.27,$$

$$s_1^2=\frac{1}{4}\sum_{i=1}^{4}(x_{1i}-\bar{x}_1)^2=0.096\,6.$$

城镇家庭(第二层)：

$$\bar{x}_2=\frac{1}{3}\sum_{i=1}^{3}x_{2i}=3.30,$$

$$s_2^2=\frac{1}{3}\sum_{i=1}^{3}(x_{2i}-\bar{x}_2)^2\approx 0.032\,5.$$

分层抽样的样本平均数为：

$$\begin{aligned}\bar{x}&=\frac{1}{7}(x_{11}+x_{12}+x_{13}+x_{14}+x_{21}+x_{22}+x_{23})\\&=\frac{1}{7}\left[4\times\frac{1}{4}(x_{11}+x_{12}+x_{13}+x_{14})+3\times\frac{1}{3}(x_{21}+x_{22}+x_{23})\right]\\&=\frac{1}{7}(4\bar{x}_1+3\bar{x}_2)\\&=\frac{1}{7}(4\times 1.27+3\times 3.30)\\&=2.14.\end{aligned}$$

分层抽样的样本方差为：

$$\begin{aligned}s^2&=\frac{1}{7}\left[\sum_{i=1}^{4}(x_{1i}-\bar{x})^2+\sum_{i=1}^{3}(x_{2i}-\bar{x})^2\right]\\&=\frac{1}{7}\left[\sum_{i=1}^{4}(x_{1i}-\bar{x}_1+\bar{x}_1-\bar{x})^2+\sum_{i=1}^{3}(x_{2i}-\bar{x}_2+\bar{x}_2-\bar{x})^2\right]\\&=\frac{1}{7}\left[\sum_{i=1}^{4}(x_{1i}-\bar{x}_1)^2+4(\bar{x}_1-\bar{x})^2+\sum_{i=1}^{3}(x_{2i}-\bar{x}_2)^2+\right.\\&\quad\left.3(\bar{x}_2-\bar{x})^2\right]\\&=\frac{1}{7}\left[4s_1^2+3s_2^2+4(\bar{x}_1-\bar{x})^2+3(\bar{x}_2-\bar{x})^2\right]\end{aligned}$$

$\sum_{i=1}^{n}x_i=x_1+x_2+\cdots+x_n.$

$$=\frac{1}{7}[4\times 0.096\ 6+3\times 0.032\ 5+4(1.27-2.14)^2+$$
$$3(3.30-2.14)^2]$$
$$\approx 1.078\ 3.$$

若将总体分成 k 个层，获得分层抽样样本

$$\begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1n_1} \\ x_{21} & x_{22} & \cdots & x_{2n_2} \\ & \cdots\cdots & & \\ x_{k1} & x_{k2} & \cdots & x_{kn_k} \end{bmatrix}.$$

由上表第 j 行可得，第 j 层样本平均数和样本方差为

$$\overline{x}_j=\frac{1}{n_j}\sum_{i=1}^{n_j}x_{ji},$$

$$s_j^2=\frac{1}{n_j}\sum_{i=1}^{n_j}(x_{ji}-\overline{x}_j)^2.$$

则分层抽样样本的样本平均数和样本方差可表示为

$$\overline{x}=\frac{1}{n}\sum_{j=1}^{k}n_j\overline{x}_j,$$

$$s^2=\frac{1}{n}\sum_{j=1}^{k}n_js_j^2+\frac{1}{n}\sum_{j=1}^{k}n_j(\overline{x}_j-\overline{x})^2.$$

以上给出了利用各层样本平均数和样本方差计算分层抽样的样本平均数和样本方差的关系式，不需要知道各层样本的个体取值，这体现出对任务的分解、综合的思想，且其中 $\frac{1}{n}\sum_{j=1}^{k}n_js_j^2$ 反映了各层内数据的分散程度，$\frac{1}{n}\sum_{j=1}^{k}n_j(\overline{x}_j-\overline{x})^2$ 反映了层间数据的分散程度.

例2 高一年级有学生 500 人，其中男生 320 人，女生 180 人. 按比例 $\frac{50}{500}=\frac{1}{10}$ 从所有学生中抽取了容量为 50 的分层样本，其中男生 32 人，女生 18 人. 已知男生组样本的平均身高为 173.5 cm，女生组样本平均身高为 163.83 cm，男生组样本方差为 17，女生组样本方差为 30.03. 求分层抽样的样本平均数和样本方差.

解 此时

$k=2$，$n_1=32$，$n_2=18$，

$\overline{x}_1=173.5$，$\overline{x}_2=163.83$，

$s_1^2=17$，$s_2^2=30.03$.

由公式得

$$\begin{aligned}\bar{x}&=\frac{32\bar{x}_1+18\bar{x}_2}{32+18}\\&=\frac{32\times173.5+18\times163.83}{50}\\&\approx170.02,\\s^2&=\frac{1}{50}[32s_1^2+18s_2^2+32(\bar{x}_1-\bar{x})^2+18(\bar{x}_2-\bar{x})^2]\\&=\frac{1}{50}(32\times17+18\times30.03+32\times3.48^2+18\times6.19^2)\\&\approx43.24.\end{aligned}$$

分层随机抽样的关键是如何分层，分层的原则是层内个体间所关心指标的差异尽可能小，层间所关心指标差异尽可能大. 分层可依据与总体中个体所关心指标具有密切联系、易于了解的辅助指标来进行. 如：调查某地区某农作物的亩产量，可按山地、丘陵和平原地貌来分层，也可按自然村、镇或其他自然区域来分层；调查居民的年收入，可按城镇和乡村来分层，也可按性别分层；调查某年级学生的体重情况，可按性别来分层，也可按身高来分层，或按多个辅助指标交叉分层(多级分层). 总的样本由各层样本组成，总的样本平均数、样本方差(样本参数)由各层样本平均数、样本方差汇总给出.

初中时我们学过用样本平均数来估计总体平均数. 对分层抽样，自然想到用层样本平均数估计层总体平均数，从而用 $N_i\bar{x}_i$ 来估计总体第 i 层的总值，用 $\sum\limits_{i=1}^{k}N_i\bar{x}_i$ 来估计总体的总值. 由此，总体平均数一个好的估计为

$$\hat{\mu}=\frac{1}{N}\sum_{i=1}^{k}N_i\bar{x}_i=\sum_{i=1}^{k}\frac{N_i}{N}\bar{x}_i.$$

在等比分层抽样下

$$\hat{\mu}=\sum_{i=1}^{k}\frac{N_i}{N}\bar{x}_i=\sum_{i=1}^{k}\frac{n_i}{n}\bar{x}_i=\frac{1}{n}\sum_{i=1}^{k}n_i\bar{x}_i=\bar{x}.$$

在一般分层抽样下，总体平均数一个好的估计为各层样本平均数关于总体层规模的加权平均.

练习

1. 举例说明分层抽样的特点及抽样步骤.
2. 某单位有职工 500 人，其中 35 岁以下的有 125 人，35～50 岁的有 280 人，50 岁以上的有 95 人. 为了了解这个单位职工的身体状况，要从中抽取一个容量为 100 的样本，问：应如何抽取?
3. 举例说明简单随机抽样和分层抽样在实际生活中的应用.

习题 2.1

1. 2017 年，某医院有 2 000 名新生婴儿，市卫生局打算从中抽取 150 名婴儿进行某项健康指标调查. 请指出这个问题中的总体、个体、样本、样本容量.
2. 某份试卷由 10 道选择题、5 道解答题组成，每道题均从试题库中抽出. 现从试题库中的 100 道选择题中随机选出 10 道，100 道解答题中随机选出 5 道. 请用随机数表法抽出考题序号，并写出抽样过程.
3. 从你所在班的同学中随机抽取 10 人去参加一项活动. 请采用不同的编号方法，并分别用抽签法和随机数表法抽取，写出抽样过程.
4. 现调查一个城市的 300 家企业，其中国有企业 50 家，股份合作企业 150 家，私营企业 100 家. 为了了解企业的经营状态，要从中抽取一个容量为 30 的样本. 若按照分层抽样的方法抽取，各类企业分别要抽取多少家? 请写出抽样过程.
5. 某市有普通高中 60 所，中等专业技术学校 12 所，职业高中 48 所. 为了了解该市高中阶段素质教育的有关情况，需抽取一个容量为 20 的样本. 问：应采用何种抽样方法? 并请写出抽样过程.
6. 某林业局欲估计植树面积，对该局所辖 240 个林场按面积大小分为四层. 现按比例从各层中共抽出 40 个林场，调查植树面积得数据资料如下表(单位：公顷)，求样本平均数和样本方差.

第一层	第二层	第三层	第四层
$N_1=84$ $n_1=14$	$N_2=72$ $n_2=12$	$N_3=54$ $n_3=9$	$N_4=30$ $n_4=5$
$\bar{x}_1=71$ $s_1^2=132$	$\bar{x}_2=210$ $s_2^2=118$	$\bar{x}_3=315$ $s_3^2=205$	$\bar{x}_4=520$ $s_4^2=80$

2.2 数据整理

我们知道，通过随机抽样从总体中抽取一定数量的个体，然后可以通过做试验、观察或问卷调查等形式获得样本的有关数据(简称为样本数据). 由于抽样是随机的，所获得的样本数据中包含有未知总体的信息. 但是数据往往是枯燥的，含于样本数据中关于总体的信息往往是分散的，我们不能一眼就看出. 因此，我们需要对样本数据进行适当的处理，采用丰富多彩的图形、简洁明了的表格呈现对样本数据进行适当处理后的结果，让数据变得生动起来，让数据说话，使得我们能够快速、清晰地从中获得所关心的总体的信息. 本节主要介绍数据的图形表示——扇形图与折线图，数据的分组处理——分组数据统计表与频率直方图这两种数据整理方法.

2.2.1 扇形图与折线图

扇形图(fan chart)也称为扇形统计图，它用整个圆表示总数，用圆内各个扇形的大小表示各部分数量占总数的份额或者百分比. 扇形图可以很清楚地表示出各部分数量同总数之间的关系. 比如从国家统计局网站获取的 2007—2016 年国内生产总值及各产业增加值数据表中，每年的国内生产总值由第一产业增加值、第二产业增加值和第三产业增加值的总和组成. 如图 2-2 中分别是 2012 年、2014 年和 2016 年国内生产总值各产业构成比例的扇形图.

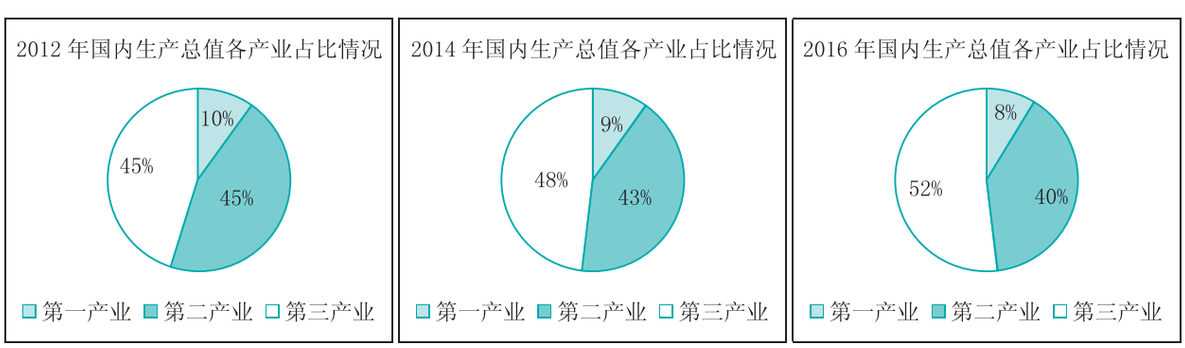


图 2-2　我国 2012 年、2014 年和 2016 年国内生产总值的构成比例扇形图

从上面的扇形图可直观感受 2012 年、2014 年和 2016 年国内生产总值各产业构成比例及变化.

折线图(line chart)是用直线段将各数据点连接起来而组成的图形，以折线方式显示数据的变化趋势. 折线图可以显示随时间(根据常用比例设置)而变化的连续数据，因此非常适用于显示在相等时间间隔下数据的变化趋势. 在折线图中，类别数据沿水平轴均匀分布，所有值数据沿垂直轴均匀分布.

图 2-3 给出了基于 2007—2016 年国内生产总值及各产业增加值数据的折线图. 其中，第一幅图为国内生产总值的折线图，第二幅图为第一产业、第二产业和第三产业增加值的折线图. 从第一幅图能理解 2007—2016 年我国国内生产总值的变化趋势和快慢；从第二幅图能读出三产业增加值的变化趋势和增速对比.

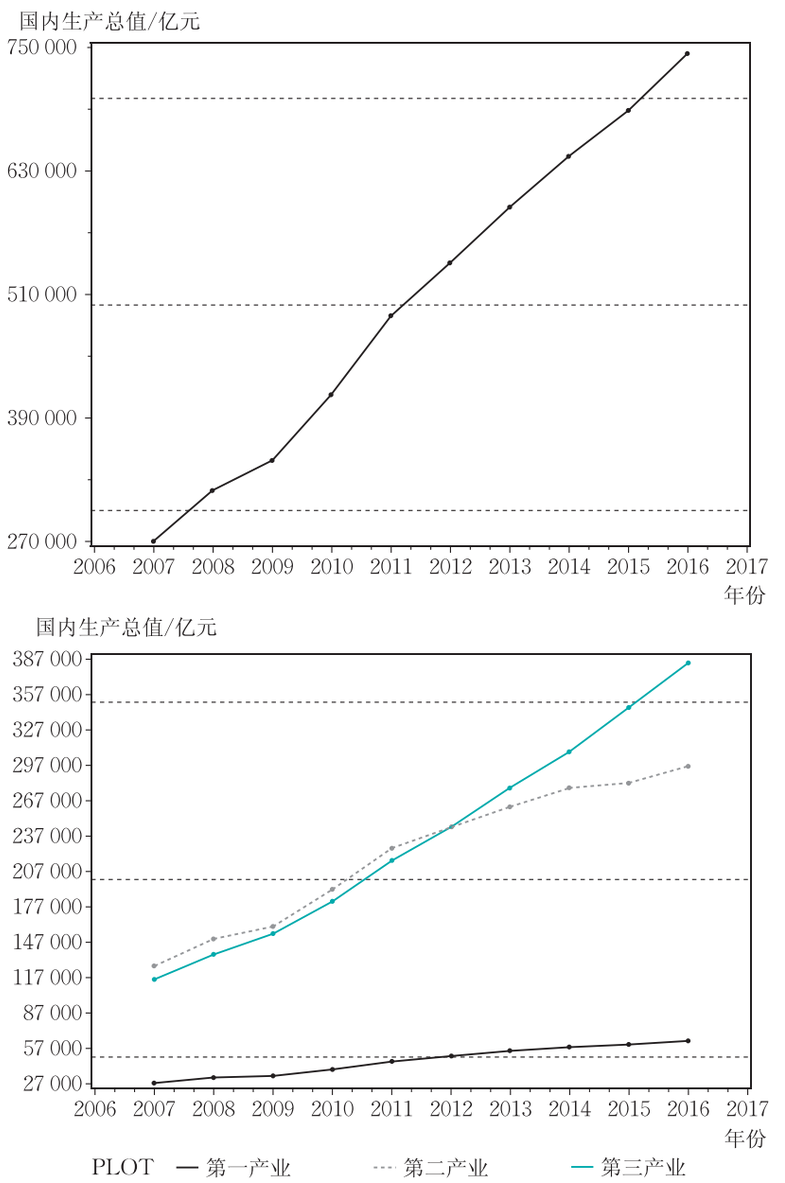


图 2-3 2007—2016 年我国国内生产总值折线图

扇形图主要从直观上描述总量的构成和分解，而折线图可以有效描述所关心指标随时间或空间的变化状况，它能直观呈现其变化趋势.

练习

1. 举出生活中用扇形图和折线图表示数据的例子，并体会两种图示方法的特点.

2.2.2 分组数据统计表与频率直方图

在这一节中，我们将通过例题介绍分组处理数据的主要步骤，分组数据统计表的制作以及频率直方图的画法，体会频率直方图的作用.

例1 县教育局为了了解在青少年中进行公共卫生知识普及的效果，从某中学抽出 50 名学生参加测试，其测试成绩如下(百分制，单位：分)，试对测试情况进行统计分析.

98	71	82	65	45	95	75	80	73	84
43	56	67	84	92	83	76	78	87	89
88	57	74	78	82	79	86	87	74	92
66	94	54	83	78	78	87	64	69	67
75	81	89	75	77	65	68	74	69	63

解 由于测试成绩的数据比较分散，不便于我们从整体上了解这 50 名学生的测试情况，因此，有必要对以上测试成绩(数据)进行适当的整理. 对数据进行分组处理是数据整理的一个有效方法. 下面给出对数据分组处理的具体做法.

第一步：找出数据中最大值(max)和最小值(min). 本例中，max=98，min=43.

第二步：确定组数 m，计算组距 c. 这里我们分成 6 个组，即 $m=6$，组距 $c=\dfrac{b-a}{m}$，其中 a 是一个略小于或等于 min 的数，b 是一个略大于或等于 max 的数，而且 a，b 的选择应尽可能使得组距 $c=\dfrac{b-a}{m}$ 比较“简单”. 本例中，取 $a=40$，$b=100$，则 $c=\dfrac{100-40}{6}=10$.

第三步：明确各组的边界，计算数据落入各组的频数 v_i

(个数)和频率 f_i(频数/总数)，形成分组数据统计表. 下面是测试成绩分组数据统计表.

组号	分组	组频数 v_i	组频率 f_i
1	[40，50)	2	4%
2	[50，60)	3	6%
3	[60，70)	10	20%
4	[70，80)	15	30%
5	[80，90)	15	30%
6	[90，100]	5	10%

由此表，我们可以得到这 50 名学生的测试成绩的基本情况为：及格(分数不少于 60)的频率为 90%，优良(分数不少于 80)的频率为 40%，优秀(分数不少于 90)的频率为 10%.

为了对以上的分析结果有一个更直观的了解，依据分组数据统计表，可以画出**频率直方图**. 具体画法如下：

在平面直角坐标系的横轴(x 轴)上，以每个分组区间(长度均为组距 c)为底边，画出高为$\frac{f_i}{c}$的矩形. 本例中$\frac{f_i}{c}=\frac{f_i}{10}$.

上表对应的频率直方图如图 2-4 所示.

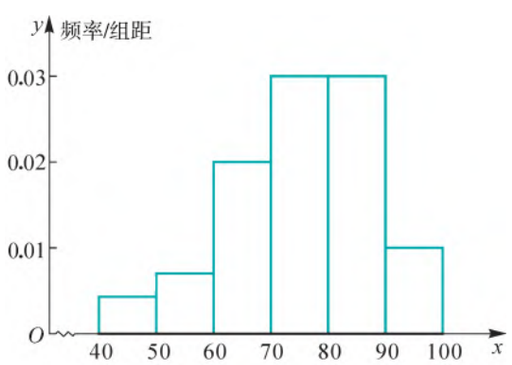


图 2-4　频率直方图

由频率直方图的画法可知，以每个分组区间为底边的矩形面积恰好为数据落入该区组的频率. 由频率直方图我们大致可以了解被测试学生成绩的分布情况：中间分数段(70～89)居多，两边分数段(70 分以下和 89 分以上)较少. 这比较符合测试成绩的分布特点.

对数据进行分组统计是数据处理的一个初步方法. 分组过程中，事先确定分多少组很重要. 若分组过多，则达不到数据整理的效果，看不出数据变化的明显规律；若分组太少，则掩盖了数据的变动情况，同样不易看出数据的变化规律，同时，由于分组

后各组中数据没有什么区别，会造成数据中所包含信息的损失. 那么究竟应该如何分组呢？下表给出了分组数的参考值.

数据个数	分组数
50 以下	5～6
50～100	6～10
100～250	7～12
250 以上	10～20

例2 某地区为了了解新生婴儿的体重情况，随机抽查了 40 名新生婴儿，并测得体重数据如下(单位：g)：

2 540　3 120　3 300　3 560　3 320　3 710　2 560　2 750
3 250　2 930　3 050　3 950　4 050　4 250　3 510　2 740
2 850　2 950　2 970　3 120　3 280　3 340　3 460　3 580
2 670　3 050　2 880　3 320　3 180　3 260　3 620　3 650
2 930　3 080　3 750　3 040　3 300　3 210　3 120　3 780

(1) 请列出分组数据统计表；

(2) 画频率直方图.

解 (1) 按照分组数据列统计表的步骤，数据中的最小值为 2 540，最大值为 4 250，取 $a=2\,500$，$b=4\,300$. 因样本容量为 40，取分组数 $m=6$，则组距为 $\dfrac{4\,300-2\,500}{6}=300$(g).

将数据范围分成

$[2\,500, 2\,800)$，$[2\,800, 3\,100)$，$[3\,100, 3\,400)$，
$[3\,400, 3\,700)$，$[3\,700, 4\,000)$，$[4\,000, 4\,300)$，

共6 组.

统计出数据落入各组的次数(可通过计算机对数据由小到大排序，以便于分组的统计)，各组的中值称为**组中值**，列如下分组数据统计表：

组号	分组	组中值	组频数 v_i	组频率 f_i
1	$[2\,500, 2\,800)$	2 650	5	12.5%
2	$[2\,800, 3\,100)$	2 950	9	22.5%
3	$[3\,100, 3\,400)$	3 250	14	35%
4	$[3\,400, 3\,700)$	3 550	6	15%
5	$[3\,700, 4\,000)$	3 850	4	10%
6	$[4\,000, 4\,300)$	4 150	2	5%

(2) 在平面直角坐标系的横轴上，以每个分组区间为底边，画出高为 $\dfrac{f_i}{c}=\dfrac{f_i}{300}$ 的矩形，即得分组数据的频率直方图，

a，b，c，m 的意义同例 1.

能取 $m=8$ 吗？若能取 $m=8$，求出 c 的值.

如图 2-5.

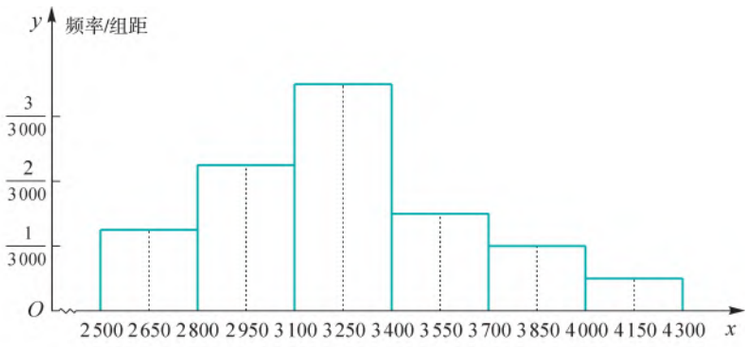


图 2-5　频率直方图

频率直方图能直观地表现样本数据的分布特点，如例 2 中，新生婴儿体重基本呈对称形态，且体重在 3 100～3 400 (单位：g)之间居多；频率直方图表达了样本数据的分布情况，它可以作为总体指标——新生婴儿体重分布的一个良好估计. 据此可以估计新生婴儿体重在任一范围内的概率——对应范围内直方图的面积.

练习

1. 举例说明数据分组的主要步骤.
2. 叙述作频率直方图应注意的事项.
3. 下表是我国 30 个地区 1995 年的人口死亡率数据(资料来源《中国人口统计年鉴(1996)》，单位：‰)：

5.12，6.23，6.32，6.12，6.70，6.15，6.09，5.33，7.05，6.56
6.75，6.41，5.90，7.28，6.47，6.28，6.91，7.15，5.70，6.53
5.61，7.21，7.60，8.03，8.80，6.57，6.49，6.89，5.49，6.45

取 $m=6$，$a=4.9$，$b=9.1$，作分组数据统计表，画频率直方图.

4. 就本节例 2 的数据表中前 20 名新生婴儿的体重数据，列分组数据统计表，画频率直方图.

习题 2.2

1. 某校高三有 195 名学生，为了调查学生的身高情况，根据男女生人数按分层抽样的方法获得一个容量为 30 的样本如下(单位：cm)：

164	151	162	156	167	172	178	176	163	152
171	174	163	166	168	169	158	157	153	154
164	161	166	165	166	167	161	159	156	157

(1) 取 $a=150$，$b=180$，$m=6$，列出分组数据统计表；
(2) 根据分组数据统计表画出频率直方图；
(3) 根据上述初步统计的结果，谈谈你对该校高三学生身高的看法.

2. 请调查你所在班级同学的视力情况，列出分组数据统计表，画出频率直方图，并进行初步分析.

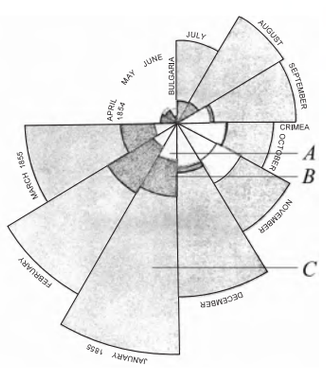
3. 请依据本节中例 2 的数据表中后 20 名新生婴儿的体重数据，列分组数据统计表，画频率直方图，并进行初步分析.

阅读与讨论

南丁格尔

同学们可能很吃惊，为什么在一本数学教科书里提到南丁格尔(1820—1910). 南丁格尔是一位伟大的女性，热情的人道主义者，现代护理事业的首创者. 护士的最高国际荣誉奖就是以她的名字命名的南丁格尔奖. 南丁格尔虽然没有上过大学，但却受过良好的数学训练，她的成就很大程度上获益于统计学.

1854 年，南丁格尔奉召参加克里米亚战争，在土耳其的乌斯库达尔(今伊斯坦布尔的一部分)主持一家军医院的护理工作. 那里恶劣的卫生条件使她大为震惊，她看到因病而死的受伤士兵人数远远超过直接战死的士兵人数. 她把丧生的士兵分为三类，A："死于战场"的，B："死于其他原因"的，C："死于可以防止的疾病"的. 它的统计表明：死于可以防止的疾病的伤兵人数远远超过死于战场的士兵人数. 她首次使用扇形统计图(图 1)，十分形象地说明了她的结论.



A：死于战场的士兵人数
B：死于其他原因的士兵人数
C：死于可以防止的疾病的士兵人数

图 1

在她大声疾呼和不懈努力下，英国政府对军医院进行了认真的改革，使得"死于可以防止的疾病"的死亡率由 40% 下降到 2%. 南丁格尔的努力得到了广泛的认可，战后英国的护士制度及护士教育制度都是依照她的方法建立起来的. 人们称她为"上帝的使女"，现代统计学创始人之一的皮尔逊(1857—1936)称她为"统计学应用的女预言家"，其传记作者则称她为"激情的统计学家". 由于她对统计学的贡献，她被选为英国皇家统计学会的第一位女会员. 在她的倡议下，牛津大学设立了统计学教授职位.

讨论题

南丁格尔的事例给了我们什么样的启示?

2.3 用样本估计总体

上节我们讨论了样本数据的初步整理，由分组数据统计表、频率直方图可以初步看出数据的分布特点．但这些整理还不能很集中地反映出数据中我们所关心的某些方面或某个方面的信息．实际问题中，我们常常希望能给出反映总体或样本数据整体水平的量化指标．

为此，我们介绍两组量化指标——反映数据集中趋势的参数(平均数、中位数、众数)和分散程度的参数(方差、标准差、极差)．

2.3.1 平均数、中位数和众数

首先我们看下面两个实际问题．

问题 1 谈到某地某种农作物的产量时，常用“平均亩产”这样一个量来说明．也就是说，农作物的平均亩产量是描述当地这种农作物种植水平的一个整体指标．

问题 2 某厂有一批电视机准备投放市场，出厂前需要对该批电视机的质量状况有一定的了解．应如何衡量这批电视机的质量水平呢?

作为厂家和消费者来说，除了通过试看了解电视机的清晰度、外观等情况外，所关心的另一重要指标是该批电视机的使用寿命，我们初中学习过的平均数、中位数和众数都能提供电视机使用寿命的整体信息．

初中学过，对一组数据 x_1，x_2，…，x_n，称

$$\overline{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$$

为该组数据的**平均数**(average)．

将该组数据 x_1，x_2，…，x_n，由小到大排列成 $x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$，则**中位数**（median）$m_{0.5}$定义为

$$m_{0.5}=\begin{cases}x_{(\frac{n+1}{2})}, & n\text{ 为奇数},\\ \dfrac{1}{2}\left[x_{(\frac{n}{2})}+x_{(\frac{n}{2}+1)}\right], & n\text{ 为偶数}.\end{cases}$$

众数(mode)即 x_1，x_2，…，x_n 中出现次数最多的数.

对于分组数据：

区组号	分组	组中值	组频数 v_i	组频率 f_i
1	$[a,\ a+c)$	$a+\frac{c}{2}=y_1$	v_1	f_1
2	$[a+c,\ a+2c)$	$a+\frac{3}{2}c=y_2$	v_2	f_2
…	…	…	…	…
m	$[a+(n-1)c,b]$	$\frac{a+b+(n-1)c}{2}=y_m$	v_m	f_m

在计算平均数、众数时同组数据同等看待，取组中值为代表值，组频数为重复次数，因此 $\bar{x}=\frac{1}{n}\sum\limits_{i=1}^{m}v_iy_i$，而众数即为最大组频数对应的组中值.

平均数反映了数据数值大小的平均水平，中位数反映了将数据按数值大小分成两个相同数量部分的界值，众数反映了数据最聚集的位置，它们都是从不同侧面反映数据集中趋势的参数.

当 x_1，x_2，…，x_n 为样本数据或样本分组数据时，我们得到样本平均数、中位数、众数的计算方法；当 x_1，x_2，…，x_n 为总体数据或总体分组数据时，我们得到总体平均数、中位数、众数的概念.

分组数据是对原始数据整理的结果，它损失了数据的原始信息，但能反映出数据的分布规律，且实际问题中我们时常要面对的就是分组数据.

例 1 (续 2.2.2 节例 1)甲、乙两位教师分别从该校 50 份试卷中随机各抽取 10 份. 分别得测试成绩如下：

教师甲　98　75　67　78　74　87　54　64　89　68

教师乙　82　80　67　78　78　92　78　67　77　63

分别计算所得样本数据的平均数、中位数、众数.

解　由样本数据平均数的计算公式，可得教师甲所抽取的 10 份试卷测试成绩的样本平均数为

$$\bar{x}_1=\frac{1}{10}(98+75+67+\cdots+89+68)=75.4,$$

且因为

$$54<64<67<68<74<75<78<87<89<98,$$

样本中位数为 $\frac{1}{2}(74+75)=74.5$，样本众数不存在.

教师乙所抽取的 10 份试卷测试成绩的样本平均数为

$$\bar{x}_2=\frac{1}{10}(82+80+67+\cdots+77+63)=76.2,$$

且因为

$$63<67=67<77<78=78=78<80<82<92,$$

样本中位数为 78，样本众数为 78.

计算结果表明，在给定的样本容量下，不同的人或不同的时间可能会抽到不完全相同的样本数据，从而可能得到不同的样本参数. 即在给定样本容量下，样本参数是可以变化的，它与抽取的样本取值有关. 这种仅与样本取值有关的量，统计上称为统计量.

平均数是反映数据整体水平最常用的参数，但它易受极端值的影响. 如：某部门有员工 6 人，其中部门经理年薪 30 万元，一般工作人员 5 人平均年薪 6 万元. 该部门在年初招聘宣传中声称该部门员工平均年薪 10 万元，这则宣传理论上没错，但招录的一般工作人员有被平均的事实. 这实际是一种带有欺骗性的宣传. 实事求是地宣传应该用到层平均，若招录一般工作人员，则平均年薪为 6 万元.

练习

1. 举例说明平均数的意义和利用平均数反映整体水平时的优缺点.
2. 对于 2.2.2 节例 1 中前 20 个数据，计算样本平均数、样本中位数和样本众数.

2.3.2 方差、标准差和极差

前面我们介绍了数据平均数的意义及其计算，知道平均数是衡量数据整体水平的一个重要指标. 但实际问题中，我们还常常需要了解数据取值的分散状态，这就是方差或标准差能够反映的内容.

对一组数据 x_1，x_2，…，x_n，其平均数为 $\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i$，称

$$s^2=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2,$$

$$s=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^2},$$

为该组数据的**方差**(variance)和**标准差**(standard deviation).

可以证明，方差还可以用如下公式计算：

$$s^2=\frac{1}{n}\sum_{i=1}^{n}x_i^2-\bar{x}^2.$$

请同学们自己证明.

若将数据 x_1，x_2，…，x_n 由小到大排列成 $x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$，则称 $R=x_{(n)}-x_{(1)}$ 为该组数据的**极差**.

对于由 2.3.1 节给出的一般分组数据，方差为

$$s^2=\frac{1}{n}\sum_{i=1}^{m}v_i(y_i-\bar{x})^2.$$

方差和标准差都反映了数据相对于平均数的分散程度，它们只是单位不同. 一般地，方差或标准差越小，则表示数据越集中在平均数附近. 极差给出了数据的变化幅度.

请同学们体会样本(总体)方差和标准差的概念.

例 1 （续 2.3.1 节例 1）求教师甲和教师乙各自随机抽取的 10 份测试卷的测试成绩的样本方差、样本标准差和样本极差.

解 （1）计算可得教师甲所抽取的 10 份试卷成绩的有关计算结果如下表：

样本号 i	1	2	3	4	5	6	7	8	9	10	$\sum$
x_i	98	75	67	78	74	87	54	64	89	68	754
x_i^2	9 604	5 625	4 489	6 084	5 476	7 569	2 916	4 096	7 921	4 624	58 404

由公式，有

$$s_1^2=\frac{1}{10}\sum_{i=1}^{10}x_i^2-\bar{x}_1^{\,2}=\frac{58\,404}{10}-\left(\frac{754}{10}\right)^2=155.24,$$

$$s_1=\sqrt{s_1^2}=\sqrt{155.24}\approx12.46,$$

$$R_1=98-54=44.$$

（2）计算可得教师乙所抽取的 10 份试卷成绩的有关计算结果如下表：

样本号 i	1	2	3	4	5	6	7	8	9	10	$\sum$
x_i	82	80	67	78	78	92	78	67	77	63	762
x_i^2	6 724	6 400	4 489	6 084	6 084	8 464	6 084	4 489	5 929	3 969	58 716

由公式，有

$$s_2^2=\frac{1}{10}\sum_{i=1}^{10}x_i^2-\bar{x}_2^{\,2}=\frac{58\,716}{10}-\left(\frac{762}{10}\right)^2=65.16,$$

$$s_2=\sqrt{s_2^2}=\sqrt{65.16}\approx8.07,$$

$$R_2=92-63=29.$$

计算结果表明：

（1）教师甲和教师乙虽然都是从这 50 名学生中抽取容量相等($n=10$)的样本数据，但却得到不同的样本方差和标准差、样本极差. 这说明样本方差、样本标准差和样本极差不仅与样本容量有关，还与从总体中抽取的个体有关. 在给定的样本容量和抽样方法下，样本方差、标准差和样本极差可能随着抽取的具体样本的变化而变化. 它们都是统计量.

（2）教师乙抽取的样本不仅平均分数高，且 10 名学生测试成绩比较集中.

（3）样本标准差和样本极差有明显的关系.

例 2 国家射击队要从过去成绩不相上下的甲、乙两名选手中选派一名选手参加男子 10 米气手枪奥运会资格赛. 在近期的 8 次内部训练中(每次射击 10 发子弹)，两名选手的训练成绩分别为(单位：环)：

甲选手　100　104　98　93　102　96　95　105

乙选手　99　100　101　98　100　98　99　102

若排除比赛时的心理因素，仅就近期的训练成绩来看，应选派哪位选手参赛？

解　从两选手近 8 次的训练成绩很难一眼就看出两选手的技术水平高低，因此有必要对数据进行统计分析，提炼出反映两选手整体水平(平均数)和技术稳定性(标准差)的指标.

由样本数据，利用计算器可得如下计算结果：

甲选手：

样本号 i	1	2	3	4	5	6	7	8	$\sum$
x_i	100	104	98	93	102	96	95	105	793
x_i^2	10 000	10 816	9 604	8 649	10 404	9 216	9 025	11 025	78 739

$$\overline{x}_1=\frac{793}{8}=99.125,$$

$$s_1^2=\frac{1}{8}\times 78\,739-99.125^2\approx 16.61,$$

$$s_1\approx 4.08,$$

$$R_1=105-93=12.$$

乙选手：

样本号 i	1	2	3	4	5	6	7	8	$\sum$
x_i	99	100	101	98	100	98	99	102	797
x_i^2	9 801	10 000	10 201	9 604	10 000	9 604	9 801	10 404	79 415

$$\overline{x}_2=\frac{797}{8}=99.625,$$

$$s_2^2=\frac{79\,415}{8}-99.625^2\approx 1.73,$$

$$s_2\approx 1.32,$$

$$R_2=102-98=4.$$

由此可见，若从近 8 次训练成绩看，则乙选手的平均成绩(99.625 环)比甲选手的平均成绩(99.125 环)高，且技术水平的稳定性指标 $s_2=1.32<s_1=4.08$，这表明乙选手的技术较甲选手稳定. 又因为波动幅度 $R_1>R_2$，一般来说应选派乙选手参赛.

但若凭过去经验，需要 101 环以上的成绩方有可能拿到参加奥运会的资格，则应选派甲选手参赛.

理解科学的决策是有前提条件的，你能从概率的角度对这种选派方式给予解释吗？

练习

1. 对同一个总体来说，为什么样本方差、极差是可以变化的？
2. 举例说明样本标准差的实际意义.
3. 说明样本极差的优缺点.

2.3.3 用样本估计总体

本章开头我们讲过，统计问题的核心就是利用随机抽样所获得的样本数据对总体中我们所关心的问题进行数据分析和推断.

按照随机抽样原则，从总体中抽取的样本数据，能够很好地反映总体的概率结构(常常未知)，包含有总体分布的信息. 对样本中所包含的总体的信息进行加工整理，把我们所感兴趣的有关总体的信息集中起来，这正是我们在前几节所介绍的内容. 自然我们会想到用样本的信息来估计总体的相应信息. 例如：用样本平均数、众数、中位数来估计总体平均数、众数、中位数；用样本方差(标准差)来估计总体方差(标准差)；用样本的频率直方图来估计总体分布的形态，进而用样本取值的频率估计总体取值的概率.

在 2.3.2 节例 2 中，对甲、乙两名选手技术水平的评价，依据的就是用样本平均数估计总体平均数，用样本方差(标准差)估计总体方差(标准差)的统计思想. 即用近期的 8 次内部训练成绩(容量为 8 的样本)的平均数 $\overline{x}_1=99.125$、$\overline{x}_2=99.625$ 来估计甲、乙两名选手的整体技术水平；用 $s_1=4.08$、

$s_2=1.32$ 来估计甲、乙两名选手技术水平的稳定性. 然后利用整体水平和稳定性作出综合评断.

在 2.3.1 节例 1 和 2.3.2 节例 1 中，我们分别计算出教师甲和教师乙从 50 份试卷中各随机抽取 10 份试卷的样本平均数(平均成绩)、标准差、极差、中位数和众数为：

教师甲：$\bar{x}_1=75.4$，$s_1=12.46$，$R_1=98-54=44$，$m_{0.5}=74.5$，样本众数不存在；

教师乙：$\bar{x}_2=76.2$，$s_2=8.07$，$R_2=92-63=29$，样本中位数和样本众数都是 78.

若按样本参数估计相应总体参数的思想，则可知：

教师甲估计 50 名学生测试成绩的平均数为 75.4，中位数为 74.5，标准差为 12.46，极差为 44；

教师乙估计 50 名学生测试成绩的平均数为 76.2，中位数和众数都是 78，标准差为 8.07，极差为 29.

这样，对同一总体我们得到两个不同的估计结果，究其原因在于：样本数据只反映了总体的部分信息，并不能完全代表总体. 样本平均数、样本中位数、样本众数、样本极差和样本标准差可能随着抽取样本的变化而变化，因此，从不同的样本数据可能会得出对总体不同的估计结果. 这样基于样本数据得出的对总体估计的任何统计结果都可能存在误差，甚至错误，但这并不影响统计的作用. 统计是在总体信息未知、只有样本数据所提供的部分信息(没有其他更多信息)的情况下，对总体作出的一个合乎情理的推断，在诸多对总体的可能选择中作出的一个相对好(合理)的选择，即找出犯错可能性小的选择.

例 1 某制药公司拥有一种新药的专利权，它可以卖掉专利，从而获得 50 万元收入，也可以保留专利，自己生产销售. 为此，公司对该药是否有效随机征求了业内 10 位专家的意见. 其中，有 5 位专家认为该药有效，有 4 位专家认为该药无效，有 1 位专家认为说不清. 若该药有效，则公司可获利 130 万元(已扣除试验成本)；若该药无效，则公司要损失 10 万元(试验费用). 根据专家意见，公司应如何作出决策?

解 根据用频率估计概率的思想，可以认为该药有效的概率为$\frac{5}{10}$，无效的概率为$\frac{4}{10}$，说不清是否有效的概率为$\frac{1}{10}$. 对于不知是否有效的情况，我们可认为它处于中间状态，取

$\frac{130+(-10)}{2}=60$(万元)作为公司的可能获利的对应值，由此对公司保留专利自己生产销售的获利情况如下表：

药物状况	有效	不明	无效
公司获利（单位：万元）	130	60	−10
频率	$\frac{5}{10}$	$\frac{1}{10}$	$\frac{4}{10}$

10 位专家的意见相当于抽取容量为 10 的样本．公司获利的样本平均数为

$$\frac{1}{10}[130\times5+60\times1+(-10)\times4]=\frac{670}{10}=67.$$

根据用样本平均数估计总体平均数的思想，可预计公司将保留专利、自己生产和销售，以获取期望收益 67 万元．因为 67 万元＞50 万元，所以根据专家意见及对药效不明时公司获利的看法，从平均收益的角度公司应保留专利，自己生产销售．当然，这样做公司也会承担一定的风险．

在实际生活中，我们还经常用到百分位数这个概念．

分位数(Quantile)也叫作分位点，是指将一个随机变量的概率分布范围分为几个等份的数值点，常用的有二分位数（即中位数）、四分位数、百分位数等．如果将一组数据从小到大排序，并计算相应的累计百分位，则某一百分位所对应数据的值就称为这一百分位的**百分位数**.

在用样本估计百分位数时，同组数据看成均匀地分布在组区间内．下面我们通过例题介绍如何计算百分位数.

例 2 食品厂用一台自动包装机装包，机器正常工作时每包标准重量为 100 g．为了解该机器工作是否正常，生产组从某日生产的产品中抽出 130 包进行测试，得分组数据如下：

分组编号	1	2	3	4	5	6	7	8
分组界限	[96，97)	[97，98)	[98，99)	[99，100)	[100，101)	[101，102)	[102，103)	[103，104)
组频数	1	5	16	34	40	22	9	3

（1）列出分组数据统计表，画出频率直方图；

（2）计算分组数据的样本平均数和样本方差；

（3）分别计算样本众数、中位数、25%分位数、75%分位数、95%分位数；

（4）分别估计包装误差不超过 1 g 和不超过 2 g 的概率.

解 （1）依据分组数据得分组数据统计表如下：

编号	分组	组中值	组频数 v_i	组频率 f_i
1	[96，97)	96.5	1	$\frac{1}{130}$
2	[97，98)	97.5	5	$\frac{5}{130}$
3	[98，99)	98.5	16	$\frac{16}{130}$
4	[99，100)	99.5	34	$\frac{34}{130}$
5	[100，101)	100.5	40	$\frac{40}{130}$
6	[101，102)	101.5	22	$\frac{22}{130}$
7	[102，103)	102.5	9	$\frac{9}{130}$
8	[103，104)	103.5	3	$\frac{3}{130}$

由分组数据统计表得频率直方图(如图 2-6).

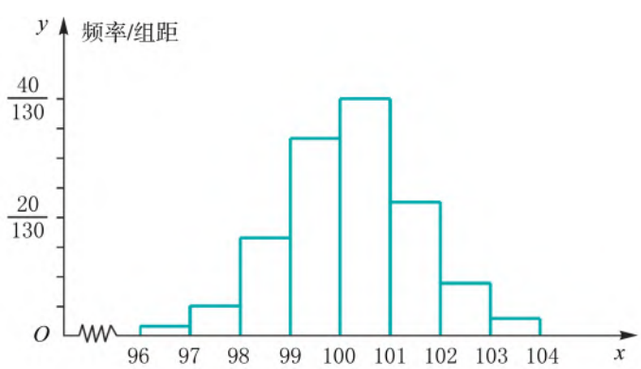


图 2-6 频率直方图

（2）根据分组数据样本平均数和样本方差的计算公式，得分组数据样本平均数

$$\bar{x}=\frac{1}{130}(1\times 96.5+5\times 97.5+16\times 98.5+\cdots+9\times 102.5+3\times 103.5)$$
$$\approx 100.2,$$

分组数据的样本方差

$$s^2=\frac{1}{130}[1\times(96.5-100.2)^2+5\times(97.5-100.2)^2+\cdots+9\times(102.5-100.2)^2+3\times(103.5-100.2)^2]$$
$$\approx 1.8.$$

（3）众数为第 5 个区组的中点 100.5.

记 $S_k=v_1+v_2+\cdots+v_k$ （$k=3$，4，5，6，7，8），则

$S_3=22$，$S_4=56$，$S_5=96$，$S_6=118$，$S_7=127$，$S_8=130$.

因为 $130\times0.5=65$，$S_4<65<S_5$，所以中位数 $m_{0.5}$ 在第 5 区组内，

$$\begin{aligned}m_{0.5}&=100\times\frac{S_5-65}{S_5-S_4}+101\times\frac{65-S_4}{S_5-S_4}\\&=100\times\frac{96-65}{96-56}+101\times\frac{65-56}{96-56}\\&=100.225.\end{aligned}$$

因为 $130\times0.25=32.5$，$S_3<32.5<S_4$，所以 25%分位数 $m_{0.25}$ 在第 4 区组内，

$$\begin{aligned}m_{0.25}&=99\times\frac{S_4-32.5}{S_4-S_3}+100\times\frac{32.5-S_3}{S_4-S_3}\\&=99\times\frac{56-32.5}{56-22}+100\times\frac{32.5-22}{56-22}\\&\approx99.31.\end{aligned}$$

因为 $130\times0.75=97.5$，$S_5<97.5<S_6$，所以 75%分位数 $m_{0.75}$ 在第 6 区组内，

$$\begin{aligned}m_{0.75}&=101\times\frac{S_6-97.5}{S_6-S_5}+102\times\frac{97.5-S_5}{S_6-S_5}\\&=101\times\frac{118-97.5}{118-96}+102\times\frac{97.5-96}{118-96}\\&\approx101.07.\end{aligned}$$

因为 $130\times0.95=123.5$，$S_6<123.5<S_7$，所以 95%分位数 $m_{0.95}$ 在第 7 区组内，

$$\begin{aligned}m_{0.95}&=102\times\frac{S_7-123.5}{S_7-S_6}+103\times\frac{123.5-S_6}{S_7-S_6}\\&=102\times\frac{127-123.5}{127-118}+103\times\frac{123.5-118}{127-118}\\&\approx102.61.\end{aligned}$$

（4）由频率直方图可知，包装误差不超过 1 g，即重量在 99～101 g 范围内的概率约为

$$\frac{34}{130}+\frac{40}{130}=\frac{74}{130}\approx0.57,$$

包装误差不超过 2 g，即重量在 98～102 g 范围内的概率约为

$$\frac{16}{130}+\frac{34}{130}+\frac{40}{130}+\frac{22}{130}=\frac{112}{130}\approx0.86.$$

中位数 $m_{0.5}$ 也可以这样来求：

因为 $130\times0.5=65$，可知中位数 $m_{0.5}$ 在第 5 区组内，设 $m_{0.5}=100+x$，则 $\frac{S_5-S_4}{130}\cdot x=\frac{65}{130}-\frac{S_4}{130}$，即 $\frac{40}{130}x=\frac{65}{130}-\frac{56}{130}$，解得 $x=0.225$.

故 $m_{0.5}=100+0.225=100.225$.

由频率直方图可以看出，自动包装机装包产品的重量关于标准重量 100 g 呈基本对称的形态，且装包重量集中在标准重量 100 g 附近，由样本方差的计算结果可以看出集中程度较高. 又样本平均数 $\bar{x}=100.2$，与理想的标准重量 100 g 相差很小，包装误差在 2 g 范围内的概率约为 0.86. 因此，可以认为包装机当日工作正常.

练习

1. 为什么由样本数据所得到的对总体的统计结果可能存在误差甚至错误?
2. 对本节例 3，
   (1) 估计包装机当日装包重量不超过 100.25 g 的概率；
   (2) 估计包装机当日装包重量不少于 98.25 g 的概率.

2.3.4 分位数应用案例——阶梯电价

本案例的目的在于通过实际案例理解百分位数的统计含义及其应用，让学生体会统计解决问题的全过程.

为了实现绿色发展，避免浪费，节约能源，某地政府计划改变居民用电单一电价收费方案，采用阶梯式递增电价收费方案. 为此，需要制定切实可行的收费标准，考虑既要满足居民基本用电需求，又要提高能源的利用效率. 为做到决策尽可能地科学合理，首先需要了解该地居民家庭用电量的现状(用电量的分布)，然后制定切实可行的阶梯式递增电价收费方案.

该地区决定采用三档递增电价，75%的用户按照最低一档电价缴费(保证基本需求)；20%的用户用电量超出一档电价的临界值、不超过二档电价的临界值，超出一档电价临界值的用电量按二档电价缴费；5%的用户用电量超过二档电价的临界值，超出二档电价临界值的用电量按三档电价缴费(体现方案的先进性,达到节约用电的目的). 两个临界值(总体 75%分位数和总体 95%分位数)的确定依赖于当地居民用电量的状况.

由于该地区六月份有一些天需要使用空调，用电量在一年 12 个月中处于中等偏上水平，为简单起见，以六月份居民户月用电量为基准，居民户年用电量近似等于该月的用电量乘 12. 为此，该地区政府部门随机调查了 200 户居民六月份(具有一定合理性)的用电量，数据如下(单位：kW·h)：

107	101	78	99	208	127	74	223	31	131
214	135	89	66	60	115	189	135	146	127
203	97	96	62	65	111	56	151	106	8
162	91	67	93	212	159	61	63	178	194
194	216	101	98	139	78	110	192	105	96
22	50	138	251	120	112	100	201	98	84
137	203	260	134	156	61	70	100	72	164
174	131	93	100	163	80	76	95	152	182
88	247	191	70	130	49	114	110	163	202
265	18	94	146	149	147	177	339	57	109
107	182	101	148	274	289	82	213	165	224
142	61	108	137	90	254	201	83	253	113
130	82	170	110	108	63	250	237	120	84
154	288	170	123	172	319	62	133	130	127
107	71	96	140	77	106	132	106	135	132
167	82	258	542	51	107	69	98	72	48
109	134	250	42	320	113	180	144	116	530
200	174	135	160	462	139	133	304	191	283
121	132	118	134	124	178	206	626	120	274
141	80	187	88	324	136	498	169	77	57

我们要用以上样本数据的信息确定上述两档电价临界值，该如何处理？

对于该样本数据，依据用样本估计总体的思想，可以用样本分位数估计相应的总体分位数，即：以 75%样本分位数作为一档电价临界值的估计，以 95%样本分位数作为二档电价临界值的估计.

利用 Excel 软件，对这组样本数据进行排序，排序结果如下：

8	18	22	31	42	48	49	50	51	56
57	57	60	61	61	61	62	62	63	63
65	66	67	69	70	70	71	72	72	74
76	77	77	78	78	80	80	82	82	82
83	84	84	88	88	89	90	91	93	93
94	95	96	96	96	97	98	98	98	99
100	100	100	101	101	101	105	106	106	106
107	107	107	107	108	108	109	109	110	110
110	111	112	113	113	114	115	116	118	120
120	120	121	123	124	127	127	127	130	130
130	131	131	132	132	132	133	133	134	134
134	135	135	135	135	136	137	137	138	139
139	140	141	142	144	146	146	147	148	149
151	152	154	156	159	160	162	163	163	164
165	167	169	170	170	172	174	174	177	178
178	180	182	182	187	189	191	191	192	194
194	200	201	201	202	203	203	206	208	212
213	214	216	223	224	237	247	250	250	251
253	254	258	260	265	274	274	283	288	289
304	319	320	324	339	462	498	530	542	626

样本数据总共有 200 个，其中最小值是 8，最大值是 626，说明 200 户居民六月份的最小用电量为 8 kW·h，最大用电量为 626 kW·h，极差为 618 kW·h.

75%样本分位数的计算：因为 200×75%=150，所以，75%样本分位数(即第一个临界值的估计)为有序样本中第 150 个数(178)和第 151 个数(178)的平均数，仍然是 178.

95%样本分位数的计算：因为 200×95%=190，所以，95%样本分位数(即第二个临界值的估计)为有序样本中第 190 个数(289)和第 191 个数(304)的平均数，即 296.5.

依据上述计算结果，确定该地居民家庭用电量三档阶梯电价方案如下：

按月结算：户用电量不超过 178 kW·h 按最低档电价收费，户用电量超过 178 kW·h 但不超过 297 kW·h 部分按第二档电价收费，户用电量超过 297 kW·h 部分按第三档电价收费；

月户用电量/(kW·h)	(0，178]	(178，297]	>297
电价收费档	一	二	三

按年结算：户用电量不超过 178×12=2 136(kW·h)按最低档电价收费，户用电量超过2 136 kW·h 但不超过 296.5×12=3 558(kW·h)部分按第二档电价收费，户用电量超过 3 558 kW·h 部分按第三档电价收费.

年户用电量/(kW·h)	(0，2 136]	(2 136，3 558]	>3 558
电价收费档	一	二	三

分位数可能不同. 按分位数的一般定义，上面的案例中，样本的 25%分位数可以是[93，94]中的任何数.

练习

1. 谈谈上述方案可改进的地方.
2. 列举分位数方法在现实生活中的应用.

习题 2.3

1. 高三(1)班学生李兵从教室步行 150 步到寝室，测得他步行 5 步的距离分别为 0.61m，0.58m，0.62m，0.59m，0.60m. 试估计教室到寝室的距离，并说明理由.

2. 根据 2.2.2 节例 2 中的数据，随机抽取两个容量为 10 的样本，分别计算其样本平均数、样本方差和标准差，并进行比较.

3. 某炼钢厂生产一种合金钢 25 MnSi. 由于受各种因素的影响，各炉钢的含 Si 量是有差异的. 下面给出的是 30 炉正常生产的合金钢 25 MnSi 的含 Si 量的数据(%)：

0.86　0.83　0.77　0.81　0.81　0.80　0.79　0.82　0.82　0.81
0.81　0.87　0.82　0.78　0.80　0.81　0.87　0.81　0.77　0.78
0.77　0.78　0.77　0.77　0.77　0.71　0.95　0.78　0.81　0.79

(1) 列出分组数据统计表；

(2) 画出分组数据频率直方图.

4. 下面是抽到的 48 株某品种小麦的穗长(单位：mm)的数据：

127	118	121	113	145	125	87	94	118	111	102	72
113	76	101	134	107	118	114	128	118	114	117	120
128	94	124	87	88	105	115	148	89	141	114	119
150	107	126	95	137	108	129	136	98	121	91	111

(1) 列出分组数据统计表；

(2) 画出频率直方图；

(3) 计算样本平均数和样本标准差；

(4) 估计穗长在 100.5 mm～120.8 mm 之间的概率；

(5) 求 20%分位数，75%分位数.

5. 某商场经理在最近 10 周内收到的顾客投诉次数如下表：

周　次	1	2	3	4	5	6	7	8	9	10
投诉次数	13	15	10	9	12	3	8	4	9	7

(1) 计算周平均投诉次数及方差；

(2) 估计周投诉次数在 5～8 次间的概率.

6. 根据 2.3.4 节案例中数据，

(1) 列出分组数据统计表，画出分组数据频率直方图(建议取 $a=7, b=627$)；

(2) 依据分组数据求 25%样本分位数、75%样本分位数、95%样本分位数，并与案例结果比较.

“百年一遇”的含义

修建三峡水利枢纽工程，最重要的目的之一是防洪．工程完工后，受长江洪水威胁最大的江段——荆江的防洪能力将提高到能抵御百年一遇洪水的水平．那么，“百年一遇的洪水”是什么意思呢？难道说某个大流量的洪水一定是一百年一次吗？抑或是荆江的防洪能力能够抵御未来一百年内的洪水出现吗？

下面是一条河流在某个水文站17年的记录：

年最大洪峰流量/(1 000 $m^3 \cdot s^{-1}$)	出现次数 v_i	出现频率 f_i	累积频率
[1.00，2.00)	1	5.9%	5.9%
[2.00，3.00)	4	23.5%	29.4%
[3.00，4.00)	7	41.2%	70.6%
[4.00，5.00)	4	23.5%	94.1%
[5.00，6.00]	1	5.9%	1

组频率累加，即得累积频率．

以年最大洪峰流量为横轴，累积频率为纵轴作上面分组数据的累积频率直方图，见图1：

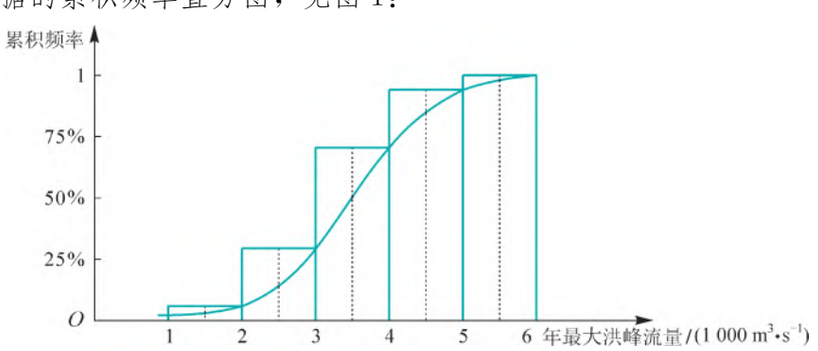


图1

以每个分组区间为底边，画出高为相应累积频率值的矩形，即得分组数据累积频率直方图．

有了累积频率直方图，连接有关的点成一条平滑的曲线即可得年最大洪峰流量的累积频率曲线．这条累积频率曲线在水文资料的分析中十分有用，它可以用来估计年最大洪峰流量在某个给定范围内的概率．例如，年最大洪峰流量大于5 000 m^3/s的概率为 $p=1-94.1\%=5.9\%$．

为便于理解，人们经常采用比较通俗的说法——重现期. 重现期就是概率的倒数. 例如，概率 $p=\frac{1}{100}$，则重现期 $T=100$(年)，通俗的说法就是“百年一遇”.

按照上面的解释，如果我们记年最大洪峰值大于某个数的概率为 p，则年最大洪峰流量不大于这个数的概率为 $1-p$，规定：$p<50\%$ 时用 p 报告洪水，即洪水的重现期 $T=\frac{1}{p}$；$p>50\%$ 时用 $1-p$ 报告枯水，即枯水的重现期 $T=\frac{1}{1-p}$. 因此利用年最大洪峰流量的累积频率曲线既可报告洪水，也可报告枯水. 如下表所示：

p	T/年	意　义
1%	100	百年一遇的洪水
10%	10	十年一遇的洪水
90%	10	十年一遇的枯水
99%	100	百年一遇的枯水

上面是对“百年一遇”最基本的统计解释. 它将各年的年最大洪峰流量同等看待，这相当于是重复试验的不同观察值. 若考虑到年最大洪峰流量的变化受不同年代的影响(如环境因素，周期因素等)，则需用更复杂的模型. 另外年最大洪峰流量的累积频率曲线与记录资料的多少有关，一般记录资料越多，则报告的可信程度越大.

因此，“百年一遇的洪水”并不是说一百年这种洪水就一定会出现一次，它实质上是一种与概率有关的描述. 一百年内，这样的洪水可能没有出现，也可能出现多次. 从长期的角度来看，它表示这样的洪水平均一百年出现一次.

讨论题

走访有关水文专家或查阅网络资料，然后讨论年最大洪峰流量累积频率曲线对报告洪水和枯水的意义.

大数据时代

为什么电脑、手机使用一段时间后运行会越来越慢？这是因为里面存储的数据越来越多．随着社会的发展，数据增长的速度越来越快，导致存储数据的空间需求越来越大．大数据已经成为一个大家熟知的热词．

那么，什么是大数据呢？它与我们传统的数据有什么区别？通常来说，大数据(big data)是具有大规模、分布式、多样性和时效性的数据，是无法在一定时间范围内用常规软件工具进行捕捉、管理和处理的数据集合．大数据分析需要新的架构和分析方法，才能有效地发现其中的有用信息．在维克托·迈尔—舍恩伯格及肯尼斯·库克耶编写的《大数据时代》一书中，大数据指不用抽样调查这样的捷径，而是采用所有数据进行分析处理．大数据的获取和分析必须要与相应的实际应用相结合．

大数据有如下几个基本特点：

1．大体量(Volume)：即数据总量非常大，可从数百 TB 到数十数百 PB、甚至 EB 的规模．(数据的计数单位最小的基本单位是 bit，按顺序依次为：bit、Byte、KB、MB、GB、TB、PB、EB、ZB、YB、BB、NB、DB．其中 1 Byte＝8 bit，之后都是 1024 进制．)

2．多样性(Variety)：即包括各种格式和形态的数据，如数值类型、文本类型、图片类型、视频类型、网页浏览数据等等．

3．时效性(Velocity)：即数据的增长和转移速度非常快，因此很多大数据分析问题需要在一定的时间限度下得到及时处理．

通常情况下，大数据中会包含大量杂乱的数据，大数据分析需要从大量各种类型的数据中快速地提取有价值的信息，因此，大数据分析可以带来巨大的商业价值或者管理效益．

下面我们给出几个大数据收集和分析应用的实例．

一、居民身份证

居民身份证是公民进行社会活动、维护社会秩序、保障公

民合法权益、证明公民身份的法定证件，它由公安部门核发和管理. 居民身份证中包含有各种类型的数据信息：文本数据、数值数据、图像数据，还有指纹数据等等. 我国人口众多，幅员辽阔，居民身份证管理系统包含的数据量非常大. 随着社会经济的增长和人员流动的日趋频繁，居民身份证的使用范围在不断扩大，相应的数据更新量也非常大，对数据的存储和管理要求变高. 居民身份证的管理和应用，需要大数据的分析作为支撑. 大数据分析工具的引入，有助于提高公安部门的管理和服务水平，能够方便公民生活，促进社会治安管理. 2016 年开始我国可以异地办理身份证，这就是公安部门引入大数据分析的优势开展的便民服务措施.

二、智能交通管控指挥平台

智能交通管控平台包含了许多数据，如：各种机动车的信息、持有驾驶证的人员信息，分布于各交通节点的监控摄像头实时提供的交通视频数据和抓拍的图片数据等. 基于大数据分析工具建立的智能交通管控平台不仅能及时发现和处理各种交通突发事件，还能提前对交通管理中可能出现的问题及时做出预警. 它的运行有利于提高公安交通管理工作的效率，减轻一线警务人员的工作压力，对加强和保障道路交通的安全、有序和畅通，减少道路交通违法和事故的发生有较好的促进作用.

讨论题

1. 利用搜索引擎搜索“中国统计年鉴”，查询并下载我国往年的各项经济指标数据、人口统计数据.
2. 身份证异地办理已经成为现实，一个基于大数据应用下的便民服务时代正逐步融入我们的生活. 请上网查询更多大数据的应用案例.

数 学 实 验

——中学生阅读课外读物每周所花时间的调查分析

某高中共有16个班，其中高一年级6个班，高二年级5个班，高三年级5个班，每班的人数均在50人左右，各班男、女生基本各占一半. 现在要调查该校学生每周用于阅读课外读物(包括报刊、科普读物、小说等，不包括看电视、上网)的时间，并对调查的结果进行分析.

具体要求：

(1) 为了使得到的数据更加可靠，避免时间过长导致调查数据不准确，要求在所定抽样“周”后的两天内完成抽样调查工作.

(2) 分别对男、女学生各抽取一定容量的样本，样本容量可在40～50之间选择.

(3) 写出实验报告，其中含：

1) 全部样本数据；

2) 各年级男生的样本平均值，各年级女生的样本平均值；

3) 全校男生的样本平均值，全校女生的样本平均值，全校学生的样本平均值；

4) 对计算结果进行简单的统计分析.

实验步骤

1. 准备工作：

(1) 时间安排：(略).

(2) 确定抽样方案：

由于各年级的学生阅读课外读物的时间存在一定的差异，我们应该采用分层抽样的方法．又由于各班的学生人数基本相同，且各班男、女生人数各占一半，分层抽样时以班级为层，从每班中随机抽取男、女生各 3 人，这样分别对男、女生各获得一组容量为 48 的样本，符合对样本容量的要求．各班抽取时，将男、女生分别编号，采用随机数表法．

(3) 确定调查内容：调查抽取出的同学每周阅读课外读物所花的时间(单位：分钟)．

2. 具体实施：

(1) 按选定的时间、已确定的抽样方案确定调查对象，制定出调查表，然后着手调查，做好记录；

(2) 按要求计算出平均数并进行统计分析；

(3) 撰写实验报告．

实验报告

年　月　日

<table>
<tr><td>题　目</td><td colspan="9">×××高中学生周阅读课外读物时间</td></tr>
<tr><td>对抽取样本的要求</td><td colspan="9">1. 阅读课外读物时间，指一周中(包括双休日)阅读报刊、科普读物、小说等时间的总和(看电视、上网不计在内)．
2. 在所确定的“周”后两天内完成抽样调查工作．
3. 男女学生的两个样本容量相同，并在 40～50 之间．</td></tr>
<tr><td>抽样方法和样本容量</td><td colspan="9">采用分层抽样的方法．以班为层，从每班中抽取男、女学生各 3 人，两个样本的容量均为 48；各班抽取时，将男、女生分别编号，采用随机数表法．</td></tr>
<tr><td rowspan="14">样本数据
(单位：分钟)</td><td></td><td colspan="4">男　生</td><td colspan="4">女　生</td></tr>
<tr><td rowspan="5">高一年级</td><td>380</td><td>500</td><td>245</td><td>145</td><td>230</td><td>600</td><td>460</td><td>110</td></tr>
<tr><td>450</td><td>480</td><td>620</td><td>420</td><td>420</td><td>580</td><td>105</td><td>400</td></tr>
<tr><td>280</td><td>660</td><td>550</td><td>350</td><td>380</td><td>420</td><td>180</td><td>500</td></tr>
<tr><td>500</td><td>330</td><td>600</td><td>180</td><td>140</td><td>450</td><td>600</td><td>400</td></tr>
<tr><td>520</td><td>520</td><td></td><td></td><td>125</td><td>540</td><td></td><td></td></tr>
<tr><td rowspan="5">高二年级</td><td>420</td><td>580</td><td>510</td><td>175</td><td>280</td><td>380</td><td>530</td><td>190</td></tr>
<tr><td>630</td><td>400</td><td>150</td><td>450</td><td>570</td><td>300</td><td>220</td><td>320</td></tr>
<tr><td>360</td><td>450</td><td>500</td><td>420</td><td>250</td><td>300</td><td>350</td><td>400</td></tr>
<tr><td>150</td><td>580</td><td>400</td><td></td><td>360</td><td>130</td><td>450</td><td></td></tr>
<tr><td rowspan="4">高三年级</td><td>380</td><td>420</td><td>235</td><td>150</td><td>200</td><td>460</td><td>160</td><td>400</td></tr>
<tr><td>400</td><td>470</td><td>330</td><td>200</td><td>430</td><td>300</td><td>210</td><td>100</td></tr>
<tr><td>280</td><td>300</td><td>410</td><td>180</td><td>130</td><td>250</td><td>170</td><td>320</td></tr>
<tr><td>210</td><td>230</td><td>400</td><td></td><td>100</td><td>270</td><td>80</td><td></td></tr>
</table>

续表

题　目	×××高中学生周阅读课外读物时间		
计算结果		男　生	女　生
	高一年级	$\bar{x}_{11}\approx429$	$\bar{x}_{21}\approx369$
	高二年级	$\bar{x}_{12}\approx412$	$\bar{x}_{22}\approx335$
	高三年级	$\bar{x}_{13}\approx306$	$\bar{x}_{23}\approx239$
	全校男生：$\bar{x}_1\approx385$； 全校女生：$\bar{x}_2\approx318$； 全校学生：$\bar{x}\approx352$.		
统计分析	由计算结果可以估计，就全校范围来讲，每周阅读课外读物的时间： (1) 男生比女生花的时间多； (2) 低年级学生比高年级学生花的时间多； (3) 全校学生平均时间约为 352 分钟.		

思考题

1. 如何从各年级男、女学生阅读时间的平均数直接得出 $\bar{x}_1$，$\bar{x}_2$？如何由 $\bar{x}_1$，$\bar{x}_2$ 直接算出全校学生的平均数？
2. 根据上面的样本数据，还能得出什么结论？比如各年级男、女学生阅读时间的分散程度如何，全校男、女学生课外阅读时间的分散程度如何等等.

复习题

A 组

1. 从某地参加计算机水平测试的 5 000 名学生的成绩中，抽取 200 名学生的成绩进行统计分析. 请指出这个问题中的总体、个体、样本、样本容量.
2. 根据本章“课题学习”中给出的抽样数据，
   (1) 分别给出全校男生、女生阅读课外读物所花时间的分组数据统计表，绘制频率直方图；
   (2) 分别给出高一年级、高二年级、高三年级学生阅读课外读物所花时间的分组数据统计表，绘制频率直方图.
3. 某车间生产一种滚珠，检验人员随机抽取了 50 个产品，测得它们的直径如下(单位：mm)：

15.0	15.8	15.2	15.1	15.9	14.7	14.8	15.5	15.6	15.3
15.1	15.3	15.0	15.6	15.7	15.8	14.5	14.2	14.9	14.9
15.0	15.3	15.6	15.1	14.9	14.2	14.6	15.8	15.2	15.9
15.2	15.2	15.0	14.9	14.8	14.5	15.1	15.5	15.5	15.1
15.1	15.0	15.3	14.7	14.5	15.5	15.0	14.7	14.6	14.2

（1）取 $a=14.0$，$b=16.1$，$m=7$，列出分组数据统计表，绘出频率直方图；

（2）在分组数据下计算样本平均数、样本方差和标准差；

（3）估计滚珠直径在 14.5 mm～15.5 mm 之间的概率；

（4）在分组数据下，求样本中位数、75%分位数、20%分位数.

4. 对第 3 题中的数据，随机抽取容量为 10 的两个样本，计算样本平均值和样本标准差，体会它们的不同.

5. 同一型号的导体其电阻可能不同，下表是随机取出一容量为 20 的样本，测得它们的电阻(单位：Ω)如下：

9.8	14.5	13.7	7.6	10.5	9.3	11.1	10.1	12.7	9.9
10.4	8.3	11.5	10.0	9.1	13.8	12.9	10.6	8.9	9.5

（1）在原始数据下，求样本平均数、样本中位数、25%样本分位数、75%样本分位数；

（2）取 $a=7.5$，$b=15.0$，$m=5$，给出分组数据统计表，绘制频率直方图；

（3）在分组数据下，求样本平均数、样本中位数、25%样本分位数、75%样本分位数.

6. 下表给出了两种型号的计算器充电以后所能使用的时间(单位：h)：

型号 A	5.5	5.6	6.3	4.6	5.3	5.0	6.2	5.8	5.1	5.2	5.9	
型号 B	3.8	4.3	4.2	4.0	4.9	4.5	5.2	4.8	4.5	3.9	4.6	3.7

（1）分别计算两种型号计算器充电后使用时间的样本平均数、25%样本分位数、样本中位数、75%样本分位数和样本标准差；

（2）根据用样本估计总体的思想，对于这两种计算器充电后使用的时间，你能初步得出什么结论?

B 组

1. 统计班里部分同学某次考试中数学成绩 x 和物理成绩 y 的数据，分别计算样本平均数和标准差.

2. 在生产过程中，测得维尼纶的纤度(表示纤维粗细的一种量)有如下的 100 个数据：

1.36	1.49	1.43	1.41	1.37	1.40	1.32	1.42	1.47	1.39
1.41	1.36	1.40	1.34	1.42	1.42	1.45	1.35	1.42	1.39
1.44	1.42	1.39	1.42	1.42	1.30	1.34	1.42	1.37	1.36
1.37	1.34	1.37	1.37	1.44	1.45	1.32	1.48	1.40	1.45
1.39	1.46	1.39	1.53	1.36	1.48	1.40	1.39	1.38	1.40
1.36	1.45	1.50	1.43	1.38	1.43	1.41	1.48	1.39	1.45
1.37	1.37	1.39	1.45	1.31	1.41	1.44	1.44	1.42	1.47
1.35	1.36	1.39	1.40	1.38	1.35	1.42	1.43	1.42	1.42
1.42	1.40	1.41	1.37	1.46	1.36	1.37	1.27	1.37	1.38
1.42	1.34	1.43	1.42	1.41	1.41	1.44	1.48	1.55	1.37

（1）列出样本分组数据统计表；

（2）画出分组数据频率直方图；

（3）利用得出的分组数据频率直方图，任取两个端点值，估计纤度在这两个端点值之间的概率约是多少?

（4）在分组数据下，试求样本平均数、样本中位数、75%样本分位数、95%样本分位数.

思考与实践

1. 请设计一个对你所在社区居民健康状况进行调查的方案，然后就你的方案与同学进行交流，说明你的调查方案的合理性.
2. 在 2.3.3 节例 2 中，我们得到分析结论：根据专家意见，公司应保留专利，自己生产销售. 请思考下列问题：
   （1）这样做，公司一定会获得比卖掉专利更多的收益吗？
   （2）公司获得比卖掉专利更多收益的可能性估计有多大？
3. 对市场上某种生活必需品(如大米、小白菜、猪肉等)的销售价格进行抽样调查. 要求：
   （1）设计抽样方案；
   （2）对样本数据进行整理，列出分组数据统计表，画频率直方图；
   （3）计算样本平均值和样本标准差；
   （4）谈谈你对分析结果的看法.

随机数表

03 47 43 73 86	36 96 47 36 61	46 98 63 71 62	33 26 16 80 45	60 11 14 10 95
97 74 24 67 62	42 81 14 57 20	42 53 32 37 32	27 07 36 07 51	24 51 79 89 73
16 76 62 27 66	56 50 26 71 07	32 90 79 78 53	13 55 38 58 59	88 97 54 14 10
12 56 85 99 26	96 96 68 27 31	05 03 72 93 15	57 12 10 14 21	88 26 49 81 76
55 59 56 35 64	38 54 82 46 22	31 62 43 09 90	06 18 44 32 53	23 83 01 30 30
16 22 77 94 39	49 54 43 54 82	17 37 93 23 78	87 35 20 96 43	84 26 34 91 64
84 42 17 53 31	57 24 55 06 88	77 04 74 47 67	21 76 33 50 25	83 92 12 06 76
63 01 63 78 59	16 95 55 67 19	98 10 50 71 75	12 86 73 58 07	44 39 52 38 79
33 21 12 34 29	78 64 56 07 82	52 42 07 44 38	15 51 00 13 42	99 66 02 79 54
57 60 86 32 44	09 47 27 96 54	49 17 46 09 62	90 52 84 77 27	08 02 73 43 28
18 18 07 92 45	44 17 16 58 09	79 83 86 19 62	06 76 50 03 10	55 23 64 05 05
26 62 38 97 75	84 16 07 44 99	83 11 46 32 24	20 14 85 88 45	10 93 72 88 71
23 42 40 64 74	82 97 77 77 81	07 45 32 14 08	32 98 94 07 72	93 85 79 10 75
52 36 28 19 95	50 92 26 11 97	00 56 76 31 38	80 22 02 53 53	86 60 42 04 53
37 85 94 35 12	83 39 50 08 30	42 34 07 96 88	54 42 06 87 98	35 85 29 48 39
70 29 17 12 13	40 33 20 38 26	13 89 51 03 74	17 76 37 13 04	07 74 21 19 30
56 62 18 37 35	96 83 50 87 75	97 12 55 93 47	70 33 24 03 54	97 77 46 44 80
99 49 57 22 77	88 42 95 45 72	16 64 36 16 00	04 43 18 66 79	94 77 24 21 90
16 08 15 04 72	33 27 14 34 09	45 59 34 68 49	12 72 07 34 45	99 27 72 95 14
31 16 93 32 43	50 27 89 87 19	20 15 37 00 49	52 85 66 60 44	38 68 88 11 80
68 34 30 13 70	55 74 30 77 40	44 22 78 84 26	04 33 46 09 52	68 07 97 06 57
74 57 25 65 76	59 29 97 68 60	71 91 38 67 54	13 58 18 24 76	15 54 55 95 52
27 42 37 86 53	48 55 90 65 72	96 57 69 36 10	96 46 92 42 45	97 60 49 04 91
00 39 68 29 61	66 37 32 20 30	77 84 57 03 29	10 45 65 04 26	11 04 96 67 24
29 94 98 94 24	68 49 69 10 82	53 75 91 93 30	34 25 20 57 27	40 48 73 51 92
16 90 82 66 59	83 62 64 11 12	67 19 00 71 74	60 47 21 29 68	02 02 37 03 31
11 27 94 75 06	06 09 19 74 66	02 94 37 34 02	76 70 90 30 86	38 45 94 30 38
35 24 10 16 20	33 32 51 26 38	79 78 45 04 91	16 92 53 56 16	02 75 50 95 98
38 23 16 86 38	42 38 97 01 50	87 75 66 81 41	40 01 74 91 62	48 51 84 08 32
31 96 25 91 47	96 44 33 49 13	34 86 82 53 91	00 52 43 48 85	27 55 26 89 62
66 67 40 67 14	64 05 71 95 86	11 05 65 09 68	76 83 20 37 90	57 16 00 11 66
14 90 84 45 11	75 73 88 05 90	52 27 41 14 86	22 98 12 22 08	07 52 74 95 80
68 05 51 18 00	33 96 02 75 19	07 60 62 93 55	59 33 82 43 90	49 37 38 44 59
20 46 78 73 90	97 51 40 14 02	04 02 33 31 08	39 54 16 49 36	47 95 93 13 30
64 19 58 97 79	15 06 15 93 20	01 90 10 75 06	40 78 78 89 62	02 67 74 17 33

05 26 93 70 60 22 35 85 15 13 92 03 51 59 77 59 56 78 06 83 52 91 05 70 74
07 97 10 88 23 09 98 42 99 64 61 71 62 99 15 06 51 29 16 93 58 05 77 09 51
68 71 86 85 85 54 87 66 47 54 73 32 08 11 12 44 95 92 63 16 29 56 24 29 48
26 99 61 65 53 58 37 78 80 70 42 10 50 67 42 32 17 55 85 74 94 44 67 16 94
14 65 52 68 75 87 59 36 22 41 26 78 63 06 55 13 08 27 01 50 15 29 39 39 43

17 53 77 58 71 71 41 61 50 72 12 41 94 96 26 44 95 27 36 99 02 96 74 30 83
90 26 59 21 19 23 52 23 33 12 96 93 02 18 39 07 02 18 36 07 25 99 32 70 23
41 23 52 55 99 31 04 49 69 96 10 47 48 45 88 13 41 43 89 20 97 17 14 49 17
60 20 50 81 69 31 99 73 68 68 35 81 33 03 76 24 30 12 48 60 18 99 10 72 34
91 25 38 05 90 94 58 28 41 36 45 37 59 03 09 90 35 57 29 12 82 62 54 65 60

34 50 57 74 37 98 80 33 00 91 09 77 93 19 82 74 94 80 04 04 45 07 31 66 49
85 22 04 39 43 73 81 53 94 79 33 62 46 86 28 08 31 54 46 31 53 94 13 38 47
09 79 13 77 48 73 82 97 22 21 05 03 27 24 83 72 89 44 05 60 35 80 39 94 88
88 75 80 18 14 22 95 75 42 49 39 32 83 22 49 02 48 07 70 37 16 04 61 67 87
90 96 23 70 00 39 00 03 06 90 55 85 78 38 36 94 37 30 69 32 90 89 00 76 33

53 74 23 99 67 61 32 28 69 84 94 62 67 86 24 98 33 41 19 95 47 53 53 38 09
63 38 06 86 54 99 00 65 26 94 02 82 90 23 07 79 62 67 80 60 75 91 12 81 19
35 30 58 21 46 06 72 17 10 94 25 21 31 75 96 49 28 24 00 49 55 65 79 78 07
63 43 36 82 69 65 51 18 37 88 61 38 44 12 45 32 92 85 88 65 54 34 81 85 35
98 25 37 55 26 01 91 82 81 46 74 71 12 94 97 24 02 71 37 07 03 92 18 66 75

02 63 21 17 69 71 50 80 89 56 38 15 70 11 48 43 40 45 86 98 00 83 26 91 03
64 55 22 21 82 43 22 28 06 00 61 54 13 43 91 82 78 12 23 29 06 66 24 12 27
85 07 26 13 89 01 10 07 82 04 59 63 69 36 03 69 11 15 83 80 13 29 54 19 28
58 54 16 24 15 51 54 44 82 00 62 61 65 04 69 38 18 65 18 97 85 72 13 49 21
34 85 27 84 87 61 48 64 56 26 90 18 48 13 26 37 70 15 42 57 65 65 80 39 07

03 92 18 27 46 57 99 16 96 56 30 33 72 85 22 84 64 38 56 98 99 01 30 93 64
62 93 30 27 59 37 75 41 66 48 86 97 80 61 45 23 53 04 01 63 45 76 08 64 27
08 45 93 15 22 60 21 75 46 91 93 77 27 85 42 28 88 61 08 84 69 62 08 42 78
07 08 55 18 40 45 44 75 13 90 24 94 96 61 02 57 55 66 83 15 73 42 37 11 61
01 85 89 95 66 51 10 19 34 88 15 84 97 19 75 12 76 39 43 78 64 63 91 08 25

72 84 71 14 85 19 11 58 49 26 50 11 17 17 76 86 81 57 20 18 95 60 78 46 75
88 78 28 16 84 13 52 58 94 53 75 45 69 80 96 73 89 65 70 31 99 17 48 48 76
45 17 75 65 57 28 40 19 72 12 25 12 74 75 67 60 40 60 81 19 24 62 01 61 16
96 76 28 12 54 22 01 11 94 25 71 96 16 16 88 68 64 36 74 45 19 59 50 88 92
43 31 67 72 30 24 02 94 08 63 88 32 36 66 02 69 36 88 25 39 48 08 45 15 22

50 44 66 44 21 66 06 58 05 62 68 15 54 35 02 42 35 48 96 32 14 52 41 52 48
22 66 22 15 86 26 63 75 41 99 58 42 36 72 24 58 37 52 18 51 03 37 18 39 11
96 24 40 14 51 28 22 30 88 57 95 67 47 29 88 94 69 40 06 07 18 16 36 78 86
31 73 91 61 19 60 20 72 98 48 98 57 07 28 69 65 95 39 69 58 56 80 30 19 44
78 60 73 99 84 43 89 94 36 45 56 69 47 07 41 90 22 91 07 12 78 35 34 08 72

84 37 90 61 56	70 10 23 98 05	85 11 34 76 60	76 48 45 34 60	01 64 18 39 96
36 67 10 08 23	98 93 35 08 86	99 29 76 29 81	88 34 91 58 93	63 14 52 32 52
07 28 59 07 48	89 64 58 89 75	83 85 62 27 89	30 14 78 56 27	86 63 59 80 02
10 15 83 87 60	79 24 31 66 56	21 48 24 06 93	91 98 94 05 49	01 47 59 38 00
55 19 68 97 65	03 73 52 16 56	00 58 55 90 27	33 42 29 38 87	22 13 88 83 34
53 81 29 13 39	35 01 20 71 34	62 33 74 82 14	53 73 19 09 03	56 54 29 56 93
51 86 32 68 92	33 98 74 66 99	40 14 71 94 58	45 94 19 33 81	14 44 99 81 07
35 91 70 29 13	80 03 54 07 27	96 94 78 32 66	50 95 52 74 33	13 80 55 62 54
37 71 67 95 13	20 02 44 95 94	64 85 04 05 72	01 32 90 76 14	53 89 74 60 41
93 66 13 83 27	92 79 64 64 72	28 54 96 53 84	48 14 52 98 94	56 07 93 39 30
02 96 08 45 65	13 05 00 41 84	93 07 54 72 59	21 45 57 09 77	19 48 56 27 44
49 83 43 48 35	82 88 33 69 96	72 36 04 19 76	47 45 15 18 60	82 11 08 95 97
84 60 71 62 46	40 80 81 30 37	34 39 23 05 33	25 15 35 71 30	88 12 57 21 77
18 17 30 88 71	44 91 14 88 47	89 23 30 63 15	56 34 20 47 89	99 82 93 24 93
79 69 10 61 78	71 32 76 95 62	87 00 22 58 40	92 54 01 75 25	43 11 71 99 31
75 93 36 57 83	56 20 14 82 11	74 21 97 90 65	98 42 68 63 86	74 54 13 26 94
38 30 92 29 03	06 23 81 39 38	62 25 06 84 63	61 29 08 93 67	04 32 92 08 09
51 29 50 10 34	31 57 75 95 80	51 97 02 74 77	76 15 48 49 44	18 55 63 77 09
21 31 38 86 24	37 79 81 53 74	73 24 16 10 33	52 83 90 94 76	70 47 14 54 36
29 01 23 87 83	58 02 39 37 67	42 10 14 20 92	16 55 23 42 45	54 96 09 11 06
95 33 95 22 00	18 74 72 00 18	38 79 58 69 32	81 76 80 26 92	82 80 84 25 39
90 84 60 79 80	24 36 59 87 38	82 07 53 89 35	96 35 23 79 18	05 98 90 07 35
46 40 62 98 80	54 97 20 56 95	15 74 80 08 32	16 46 70 50 80	67 72 16 42 79
20 31 89 03 43	38 46 82 68 72	32 14 82 99 70	80 60 47 18 97	63 49 30 21 30
71 59 73 05 50	08 22 23 71 77	91 01 93 20 49	82 96 59 26 94	66 39 67 98 60

后　记

为了全面贯彻党的教育方针，适应时代发展的需要，为学生的终身发展奠定基础，依据《普通高中数学课程标准（2017 年版）》，我们组织专家学者编写了这套普通高中数学教科书.

在本套教科书的编写过程中，我们得到了许多数学教育界前辈、数学课程专家、数学教育理论工作者、中学数学教研员和教师的大力支持和热情帮助，我们对他们的辛勤付出表示衷心的感谢. 我们还要特别感谢华中师范大学数学与统计学学院对本套教科书编写工作的高度重视和大力支持.

本套教科书是全体编写人员集体智慧的结晶. 除已列出的主要编写者外，参加本册教科书编写讨论的还有：罗国彬、邓勤涛、刘运新、岑爱国、高云、李文溢、覃红、孔凡祥、乔安国、徐新斌、张琴、田杰、贾培等.

我们还要感谢使用本套教科书的师生们，期待你们在使用本套教科书的过程中，及时把意见和建议反馈给我们，以便我们进一步修改完善.

责任编辑 田 杰 张 琴
封面设计 牛 红 刘静文

普通高中教科书 数学 必修 第四册

出 版 湖北教育出版社 430070 武汉市雄楚大街 268 号
经 销 新 华 书 店
网 址 http://www.hbedup.com
印 刷 武汉中远印务有限公司
开 本 890mm×1240mm 1/16
印 张 5.5
字 数 100 千字
版 次 2019 年 11 月第 1 版
印 次 2019 年 11 月第 1 次印刷
书 号 ISBN 978-7-5564-3143-4
定 价 5.50 元