

普通高中课程标准实验教科书

数学 ③

必修

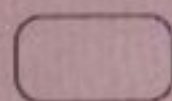
人民教育出版社 课程教材研究所 编著
中学数学课程教材研究开发中心



人民教育出版社

A版

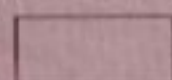
本书部分常用符号



终端框 (起止框)



输入、输出框



处理框 (执行框)



判断框

s

标准差

s^2

方差

$f_s(A)$

事件 A 出现的频率

$P(A)$

事件 A 的概率

目 录

第一章 算法初步	1
1.1 算法与程序框图	2
1.2 基本算法语句	21
1.3 算法案例	34
阅读与思考 割圆术	45
小结	49
复习参考题	50
第二章 统计	53
2.1 随机抽样	54
阅读与思考 一个著名的案例	55
阅读与思考 广告中数据的可靠性	59
阅读与思考 如何得到敏感性问题的诚实反应 ...	62
2.2 用样本估计总体	65
阅读与思考 生产过程中的质量控制图	79



2.3 变量间的相关关系	84
阅读与思考 相关关系的强与弱	92
实习作业	96
小结	98
复习参考题	100

第三章 概率

3.1 随机事件的概率	108
阅读与思考 天气变化的认识过程	122
3.2 古典概型	125
3.3 几何概型	135
阅读与思考 概率与密码	140
小结	143
复习参考题	145

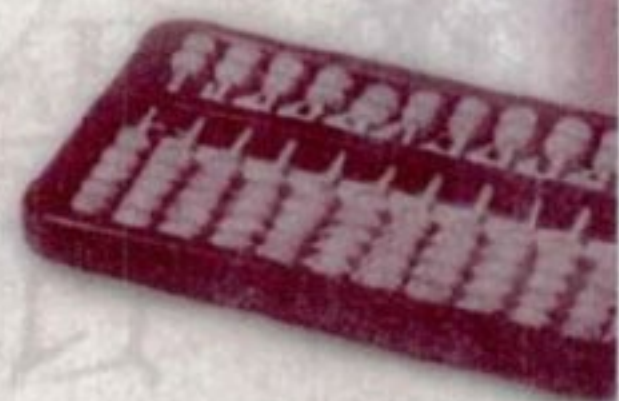


1

又以前式左行齊左式得

得

玉鑑



中国古代数学中蕴涵了丰富的
算法思想，算筹、算盘都是盛行一时的
计算工具。如今，算法已经成为计算机科学
的重要基础，同时计算机又是强大的实现
各种算法的工具。

第一章

算法初步

1.1

算法与程序框图

1.2

基本算法语句

1.3

算法案例

算法不仅是数学及其应用的重要组成部分，也是计算机科学的重要基础。在现代社会里，计算机已经成为人们日常生活和工作不可缺少的工具。听音乐、看电影、玩游戏、打字、画卡通画、处理数据，计算机几乎渗透到了人们生活的各个领域。那么，计算机是怎样工作的呢？要想弄清楚这个问题，算法的学习是一个开始。

从数学发展的历史来看，算法的概念古已有之。比如，在西方数学中很早就有了欧几里得算法，而中国古代数学中蕴涵着更为丰富的算法内容和思想。割圆术、秦九韶算法等等都是很经典的算法。在这一章里，我们将学习算法的概念和程序框图，理解算法的基本结构、基本算法语句，了解一些很有意思的重要算法，体会算法的基本思想，发展有条理的思考与表达的能力，提高逻辑思维能力。

1.1.1 算法的概念

实际上, 算法对我们来说并不陌生.
回顾二元一次方程组

$$\begin{cases} x-2y=-1, & \text{①} \\ 2x+y=1 & \text{②} \end{cases}$$

的求解过程, 我们可以归纳出以下步骤:

第一步, ①+②×2, 得

$$5x=1. \quad \text{③}$$

第二步, 解③, 得 $x=\frac{1}{5}$.

第三步, ②-①×2, 得

$$5y=3. \quad \text{④}$$

第四步, 解④, 得 $y=\frac{3}{5}$.

第五步, 得到方程组的解为 $\begin{cases} x=\frac{1}{5}, \\ y=\frac{3}{5}. \end{cases}$



你能写出求解一般的二元一次方程组的步骤吗?

对于一般的二元一次方程组

$$\begin{cases} a_1x+b_1y=c_1, & \text{⑤} \\ a_2x+b_2y=c_2, & \text{⑥} \end{cases}$$

其中 $a_1b_2-a_2b_1 \neq 0$, 可以写出类似的求解步骤:

第一步, ⑤ $\times b_2$ -⑥ $\times b_1$, 得

$$(a_1b_2-a_2b_1)x=b_2c_1-b_1c_2, \quad (7)$$

第二步, 解⑦, 得 $x=\frac{b_2c_1-b_1c_2}{a_1b_2-a_2b_1}$.

第三步, ⑥ $\times a_1$ -⑤ $\times a_2$, 得

$$(a_1b_2-a_2b_1)y=a_1c_2-a_2c_1, \quad (8)$$

第四步, 解⑧, 得 $y=\frac{a_1c_2-a_2c_1}{a_1b_2-a_2b_1}$.

第五步, 得到方程组的解为
$$\begin{cases} x=\frac{b_2c_1-b_1c_2}{a_1b_2-a_2b_1}, \\ y=\frac{a_1c_2-a_2c_1}{a_1b_2-a_2b_1}. \end{cases}$$

上述步骤构成了解二元一次方程组的一个算法, 我们可以进一步根据这一算法编制计算机程序, 让计算机来解二元一次方程组.

算法① (algorithm) 一词出现于 12 世纪, 指的是用阿拉伯数字进行算术运算的过程. 在数学中, 算法通常是指按照一定规则解决某一类问题的明确和有限的步骤. 现在, 算法通常可以编成计算机程序, 让计算机执行并解决问题.

① 据说英文 algorithm 来源于阿拉伯数学家花拉子米的拉丁译名 Algoritmi.

例 1 (1) 设计一个算法, 判断 7 是否为质数②.

(2) 设计一个算法, 判断 35 是否为质数.

算法分析:

(1) 根据质数的定义, 可以这样判断: 依次用 2~6 除 7, 如果它们中有一个能整除 7, 则 7 不是质数, 否则 7 是质数.

根据以上分析, 可写出如下的算法:

第一步, 用 2 除 7, 得到余数 1. 因为余数不为 0, 所以 2 不能整除 7.

第二步, 用 3 除 7, 得到余数 1. 因为余数不为 0, 所以 3 不能整除 7.

第三步, 用 4 除 7, 得到余数 3. 因为余数不为 0, 所以 4 不能整除 7.

第四步, 用 5 除 7, 得到余数 2. 因为余数不为 0, 所以 5 不能整除 7.

第五步, 用 6 除 7, 得到余数 1. 因为余数不为 0, 所以 6 不能整除 7. 因此, 7 是质数.

② 只能被 1 和自身整除的大于 1 的整数叫质数.

(2) 类似地, 可写出“判断 35 是否为质数”的算法:

第一步, 用 2 除 35, 得到余数 1. 因为余数不为 0, 所以 2 不能整除 35.

第二步, 用 3 除 35, 得到余数 2. 因为余数不为 0, 所以 3 不能整除 35.

第三步, 用 4 除 35, 得到余数 3. 因为余数不为 0, 所以 4 不能整除 35.

第四步, 用 5 除 35, 得到余数 0. 因为余数为 0, 所以 5 能整除 35. 因此, 35 不是质数.



你能写出“判断整数 $n(n>2)$ 是否为质数”的算法吗?

对于任意的整数 $n(n>2)$, 若用 i 表示 $2\sim(n-1)$ 中的任意整数, 则“判断 n 是否为质数”的算法包含下面的重复操作:

用 i 除 n , 得到余数 r . 判断余数 r 是否为 0, 若是, 则 n 不是质数; 否则, 将 i 的值增加 1, 再执行同样的操作.

这个操作一直要进行到 i 的值等于 $(n-1)$ 为止. 因此, “判断 n 是否为质数”的算法可以写成:

第一步, 给定大于 2 的整数 n .

第二步, 令 $i=2$.

第三步, 用 i 除 n , 得到余数 r .

第四步, 判断“ $r=0$ ”是否成立. 若是, 则 n 不是质数, 结束算法; 否则, 将 i 的值增加 1, 仍用 i 表示.

第五步, 判断“ $i>(n-1)$ ”是否成立. 若是, 则 n 是质数, 结束算法; 否则, 返回第三步.

例 2 写出用“二分法”求方程 $x^2-2=0(x>0)$ 的近似解的算法.

算法分析:

令 $f(x)=x^2-2$, 则方程 $x^2-2=0$ 的解就是函数 $f(x)$ 的零点.

“二分法”的基本思想是: 把函数 $f(x)$ 的零点所在的区间 $[a, b]$ (满足 $f(a) \cdot f(b) < 0$) “一分为二”, 得到 $[a, m]$ 和 $[m, b]$. 根据“ $f(a) \cdot f(m) < 0$ ”是否成立, 取出零点所在的区间 $[a, m]$ 或 $[m, b]$, 仍记为 $[a, b]$. 对所得的区间 $[a, b]$ 重复上述步骤, 直到包含零点的区间 $[a, b]$ “足够小”, 则 $[a, b]$ 内的数可以作为方程的近似解.

根据以上分析可以写出如下的算法:

第一步, 令 $f(x)=x^2-2$, 给定精确度 d .

第二步, 确定区间 $[a, b]$, 满足 $f(a) \cdot f(b) < 0$.

第三步, 取区间中点 $m = \frac{a+b}{2}$.

第四步, 若 $f(a) \cdot f(m) < 0$, 则含零点的区间为 $[a, m]$; 否则, 含零点的区间为 $[m, b]$. 将新得到的含零点的区间仍记为 $[a, b]$.

第五步, 判断 $[a, b]$ 的长度是否小于 d 或 $f(m)$ 是否等于 0. 若是, 则 m 是方程的近似解; 否则, 返回第三步.

当 $d=0.005$ 时, 按照以上算法, 可以得到表 1-1 和图 1.1-1.

表 1-1

a	b	$ a-b $
1	2	1
1	1.5	0.5
1.25	1.5	0.25
1.375	1.5	0.125
1.375	1.437 5	0.062 5
1.406 25	1.437 5	0.031 25
1.406 25	1.421 875	0.015 625
1.414 062 5	1.421 875	0.007 812 5
1.414 062 5	1.417 968 75	0.003 906 25

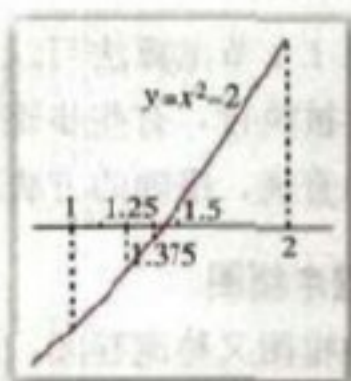


图 1.1-1

于是, 开区间 $(1.414\ 062\ 5, 1.417\ 968\ 75)$ 中的实数都是当精确度为 0.005 时的原方程的近似解.

实际上, 上述步骤也是求 $\sqrt{2}$ 的近似值的一个算法.

计算机解决任何问题都要依赖于算法. 只有将解决问题的过程分解为若干个明确的步骤, 即算法, 并用计算机能够接受的“语言”准确地描述出来, 计算机才能够解决问题.



你能举出更多的算法的例子吗? 与一般的解决问题的过程相比, 你认为算法最重要的特征是什么?

练习

- 任意给定一个正实数, 设计一个算法求以这个数为半径的圆的面积.
- 任意给定一个大于 1 的正整数 n , 设计一个算法求出 n 的所有因数.

1.1.2 程序框图与算法的基本逻辑结构

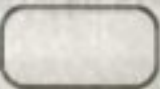
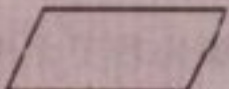
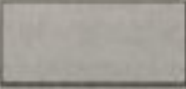

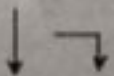

从1.1.1节的算法可以看出,算法步骤有明确的顺序性,而且有些步骤只有在一定条件下才会被执行,有些步骤在一定条件下会被重复执行.因此,我们有必要探究使算法表达得更加直观、准确的方法.

1. 程序框图

程序框图又称流程图,是一种用程序框、流程线及文字说明来表示算法的图形.

在程序框图中,一个或几个程序框的组合表示算法中的一个步骤;带有方向箭头的流程线将程序框连接起来,表示算法步骤的执行顺序.表1-2列出了几个基本的程序框、流程线和它们表示的功能.

表 1-2

图形符号	名称	功能
	终端框(起止框)	表示一个算法的起始和结束
	输入、输出框	表示一个算法输入和输出的信息
	处理框(执行框)	赋值、计算
	判断框	判断某一条件是否成立,成立时在出口处标明“是”或“Y”;不成立时标明“否”或“N”.
	流程线	连接程序框
	连接点	连接程序框图的两部分

例如,1.1.1节中“判断整数 $n(n>2)$ 是否为质数”的算法就可以用下面的程序框图表示.

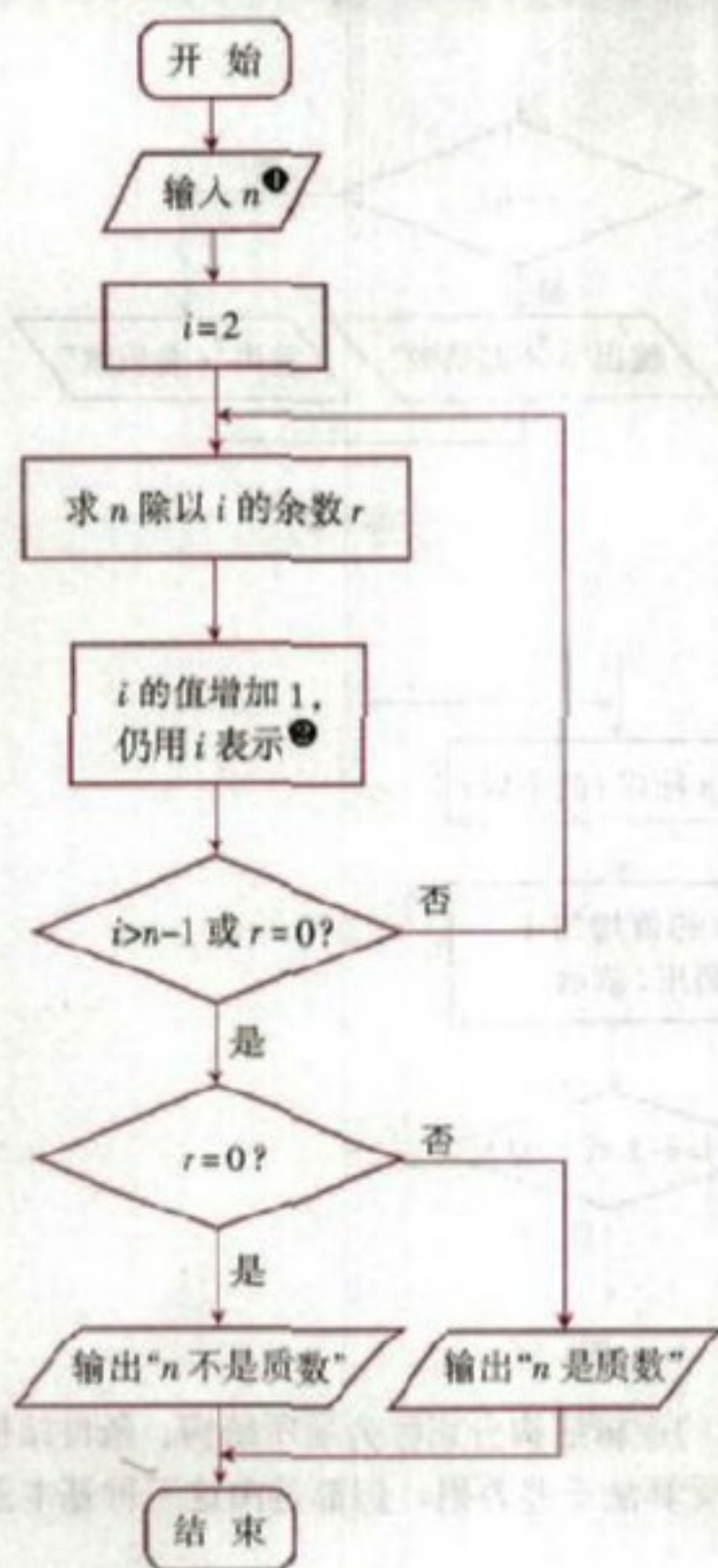


图 1.1-2

① 设 n 是一个大于 2 的整数。

② 一般用 $i = i + 1$ 表示。

程序框图的第一个程序框和最后一个程序框都是终端框，分别表示一个算法的开始和结束。

2. 算法的基本逻辑结构

用程序框图表示算法时，算法的逻辑结构展现得非常清楚，图 1.1-2 的程序框图中包含下面三种逻辑结构：



图 1.1-3

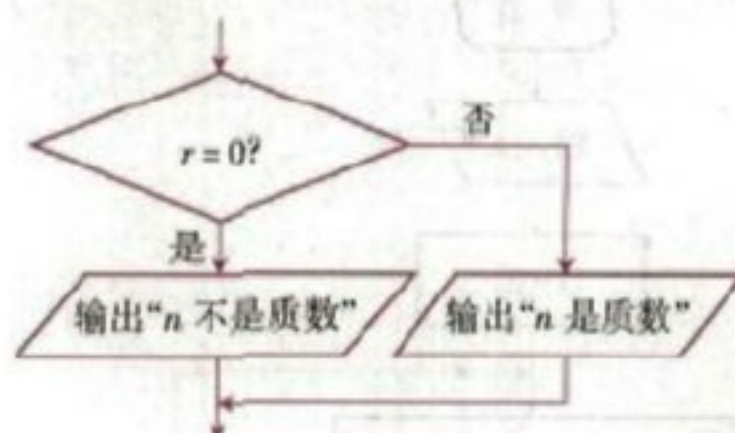


图 1.1-4

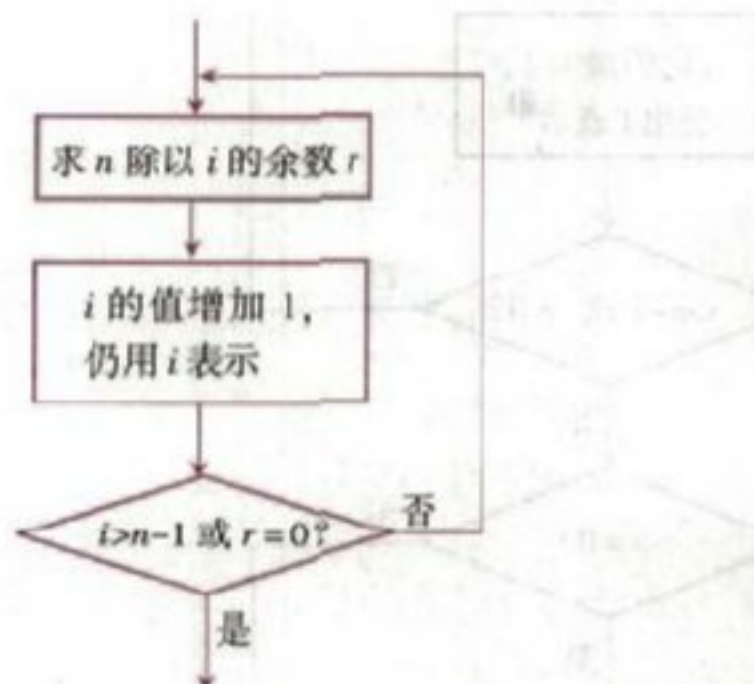


图 1.1-5

图 1.1-3, 图 1.1-4 和图 1.1-5 表示的逻辑结构分别称为**顺序结构**、**条件结构**和**循环结构**, 这是算法的三种基本逻辑结构. 尽管算法千差万别, 但都是由这三种基本逻辑结构构成的.



你能说出这三种基本逻辑结构的特点吗? 条件结构与循环结构有什么区别和联系?

(1) 顺序结构

很明显, 顺序结构是由若干个依次执行的步骤组成的. 这是任何一个算法都离不开的基本结构.

顺序结构可以用程序框图表示为(图 1.1-6):

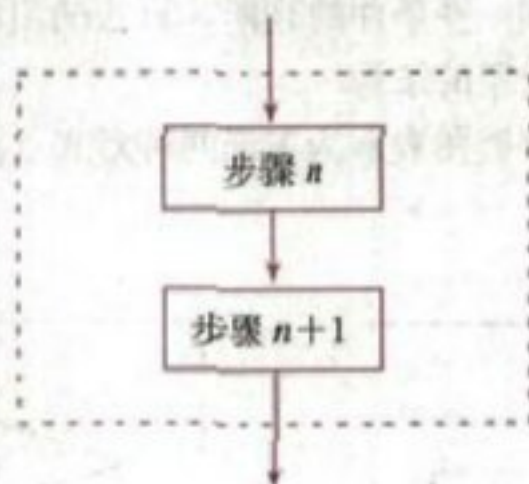


图 1.1-6

例 3 已知一个三角形三条边的边长分别为 a, b, c , 利用海伦—秦九韶公式^①设计一个计算三角形面积的算法, 并画出程序框图表示。

算法分析:

这是一个简单的问题, 只需先算出 p 的值, 再将它代入公式, 最后输出结果。因此只用顺序结构就能表达出算法。

算法步骤如下:

第一步, 输入三角形三条边的边长 a, b, c 。

第二步, 计算 $p = \frac{a+b+c}{2}$ 。

第三步, 计算 $S = \sqrt{p(p-a)(p-b)(p-c)}$ 。

第四步, 输出 S 。

程序框图:

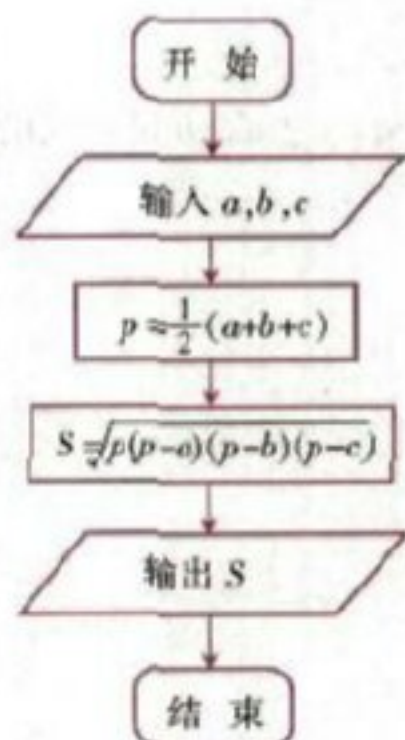


图 1.1-7

① 已知三角形三边边长分别为 a, b, c , 则三角形的面积为 $S = \sqrt{p(p-a)(p-b)(p-c)}$, 其中 $p = \frac{a+b+c}{2}$, 这个公式被称为海伦—秦九韶公式。

(2) 条件结构

在一个算法中,经常会遇到一些条件的判断,算法的流程根据条件是否成立有不同的流向.条件结构就是处理这种过程的结构.

常见的条件结构可以用程序框图表示为下面两种形式(图 1.1-8 和图 1.1-9):

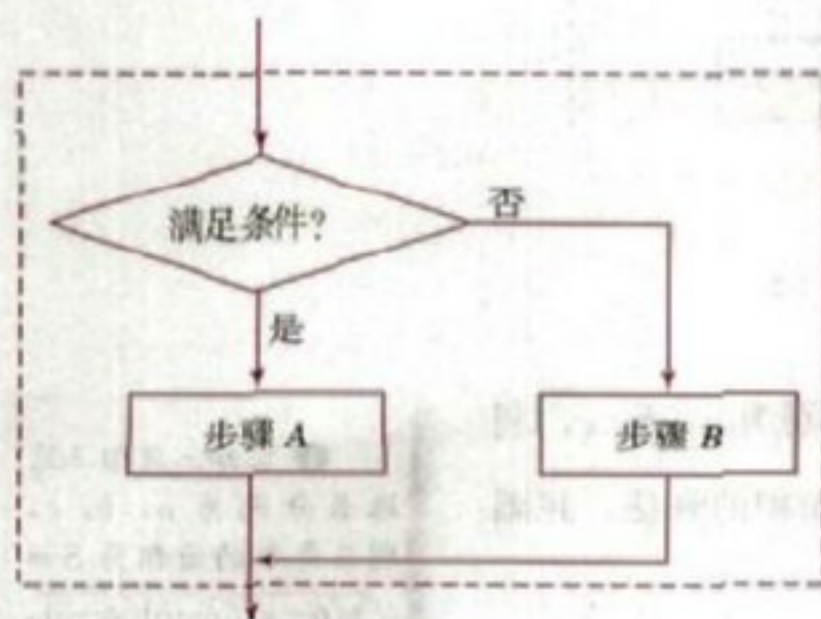


图 1.1-8

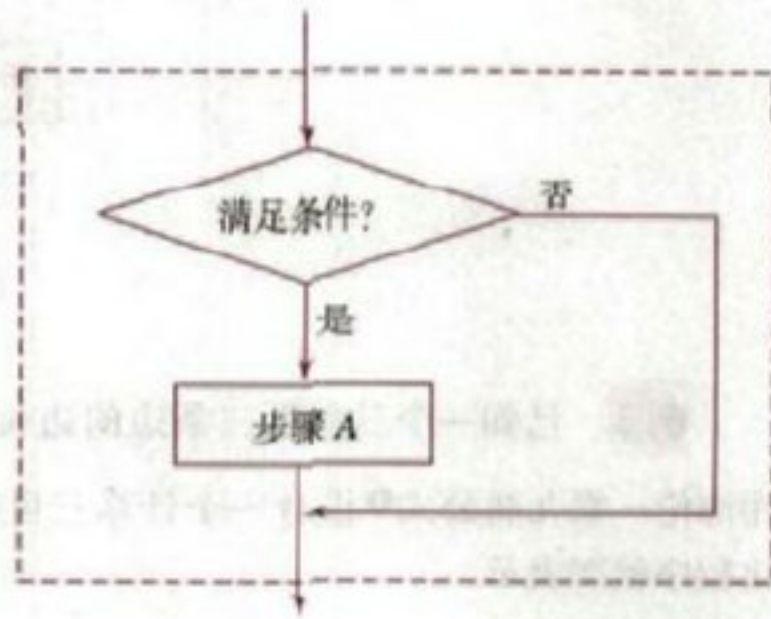


图 1.1-9

例 4 任意给定 3 个正实数,设计一个算法,判断以这 3 个正实数为三条边边长的三角形是否存在,并画出这个算法的程序框图.

算法分析:

判断以 3 个任意给定的正实数为三条边边长的三角形是否存在,只需验证这 3 个数中任意两个数的和是否大于第 3 个数.这个验证需要用到条件结构.

算法步骤如下:

第一步,输入 3 个正实数 a, b, c .

第二步,判断 $a+b>c, b+c>a, c+a>b$ 是否同时成立.若是,则存在这样的三角形;否则,不存在这样的三角形.

程序框图：

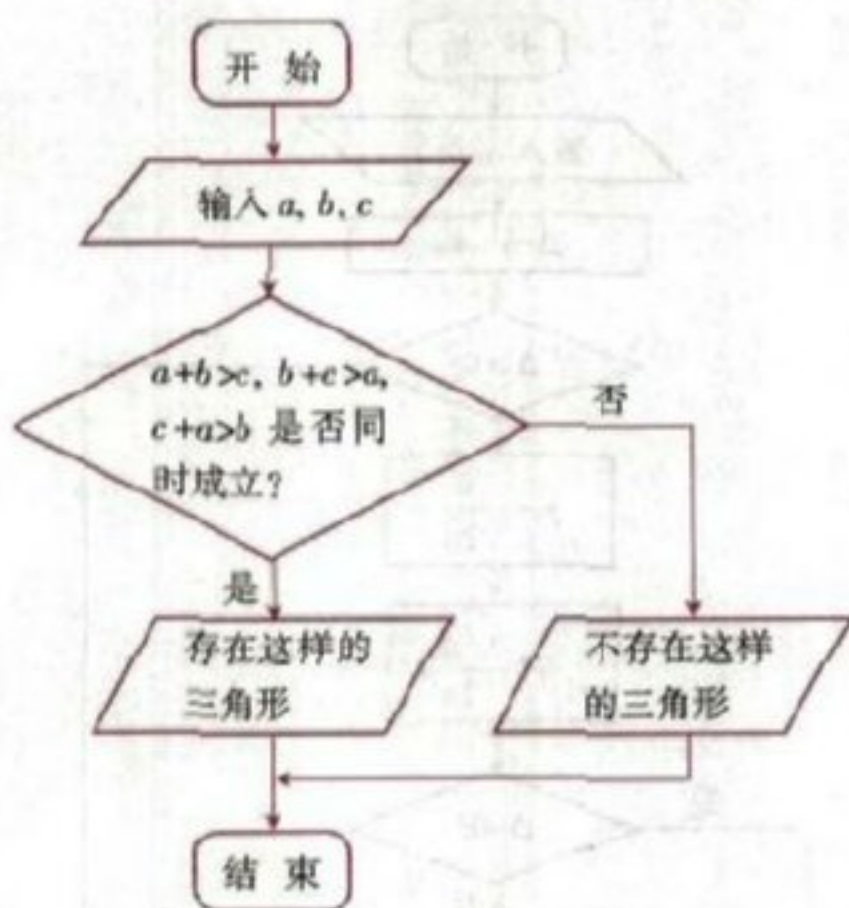


图 1.1-10

例 5 设计一个求解一元二次方程 $ax^2+bx+c=0$ 的算法，并画出程序框图表示。

算法分析：

我们知道，若判别式 $\Delta=b^2-4ac>0$ ，则原方程有两个不相等的实数根 $x_1=\frac{-b+\sqrt{\Delta}}{2a}$ ， $x_2=\frac{-b-\sqrt{\Delta}}{2a}$ ；若 $\Delta=0$ ，则原方程有两个相等的实数根 $x_1=x_2=-\frac{b}{2a}$ ；若 $\Delta<0$ ，则原方程没有实数根。也就是说，在求解方程之前，可以先判断判别式的符号，根据判断的结果执行不同的步骤，这个过程可以用条件结构实现。

又因为方程的两个根有相同的部分，为了避免重复计算，可以在计算 x_1 和 x_2 之前，先计算 $p=-\frac{b}{2a}$ ， $q=\frac{\sqrt{\Delta}}{2a}$ 。

解决这一问题的算法步骤如下：

第一步，输入 3 个系数 a, b, c 。

第二步，计算 $\Delta=b^2-4ac$ 。

第三步，判断 $\Delta\geq 0$ 是否成立。若是，则计算 $p=-\frac{b}{2a}$ ， $q=\frac{\sqrt{\Delta}}{2a}$ ；否则，输出“方程没有实数根”，结束算法。

第四步，判断 $\Delta=0$ 是否成立。若是，则输出 $x_1=x_2=p$ ；否则，计算 $x_1=p+q$ ， $x_2=p-q$ ，并输出 x_1, x_2 。

程序框图：

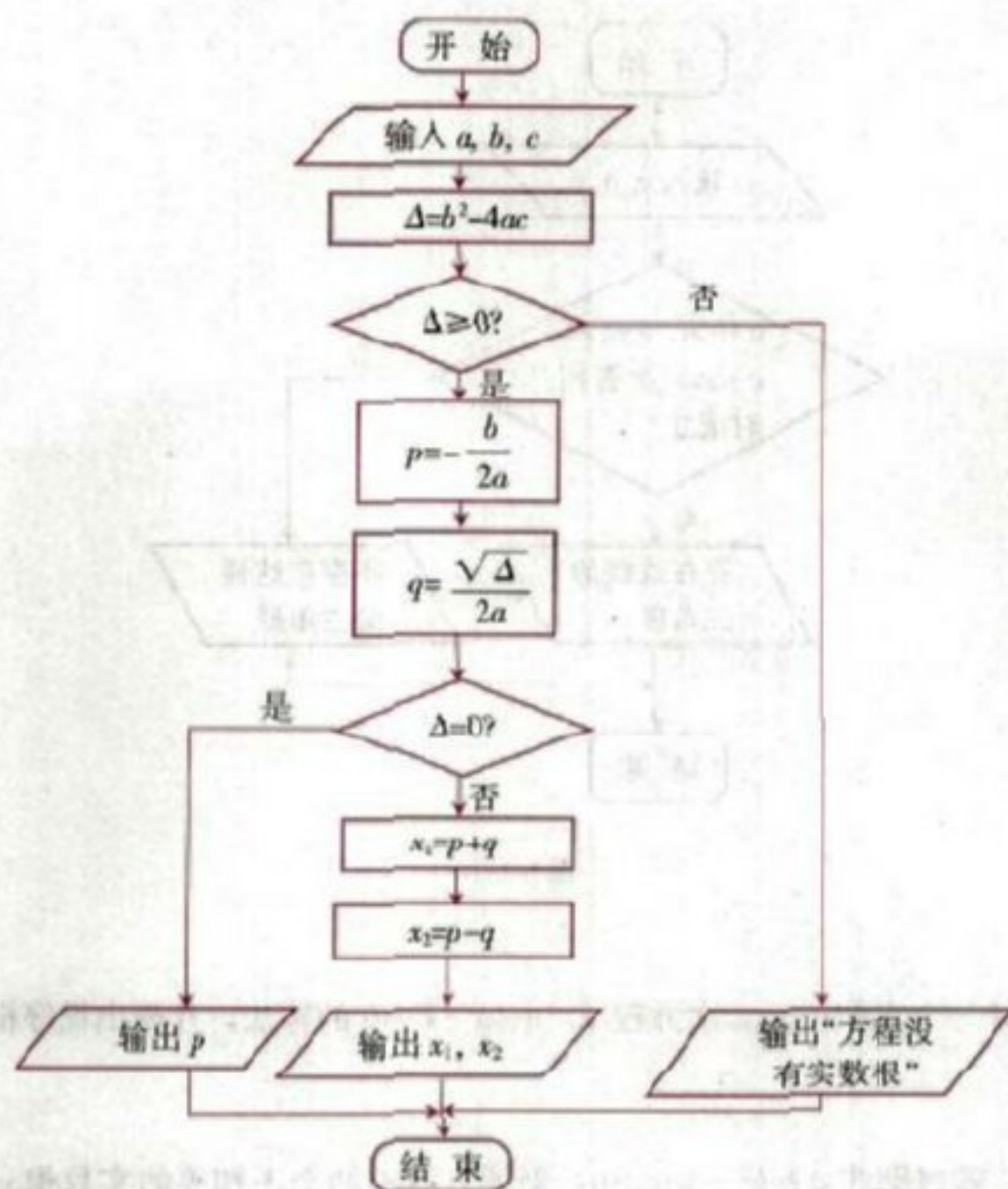


图 1.1-11

(3) 循环结构

在一些算法中，经常会出现从某处开始，按照一定的条件反复执行某些步骤的情况，这就是循环结构。反复执行的步骤称为循环体。

循环结构可以用程序框图表示为（图 1.1-12）：

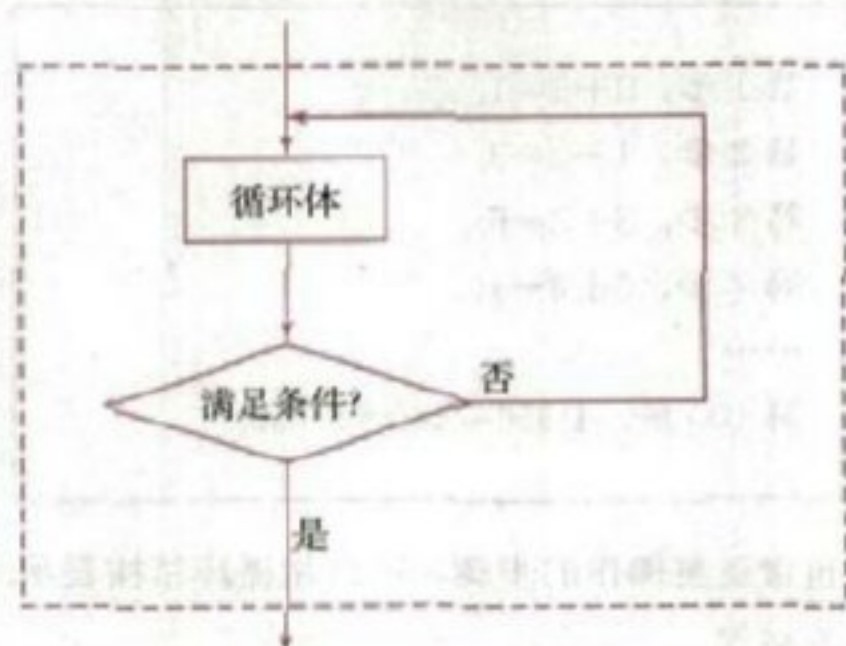


图 1.1-12

这个循环结构有如下特征：在执行了一次循环体后，对条件进行判断，如果条件不满足，就继续执行循环体，直到条件满足时终止循环，因此，这种循环结构称为直到型循环结构。

除直到型循环结构外，图 1.1-13 表示的也是常见的循环结构，它有如下特征：在每次执行循环体前，对条件进行判断，当条件满足时，执行循环体，否则终止循环，因此，这种循环结构称为当型循环结构。

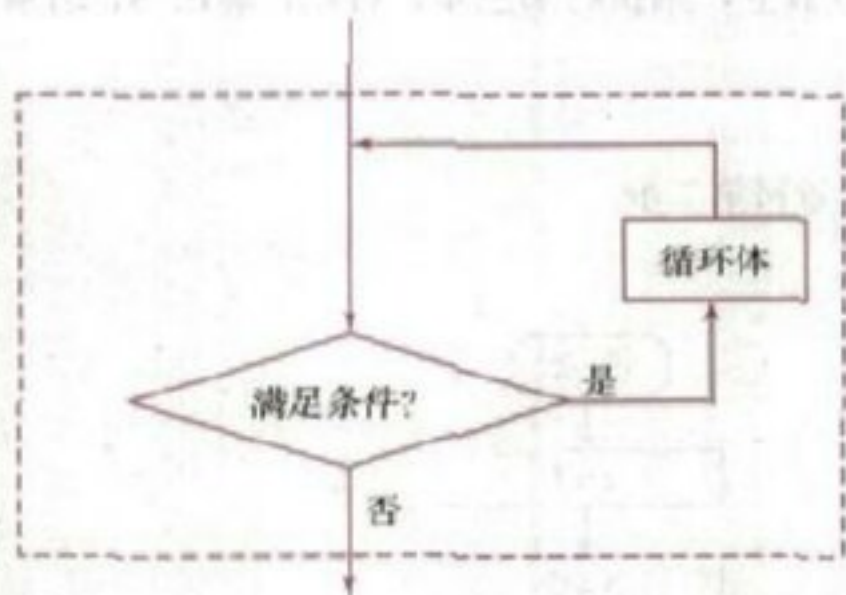


图 1.1-13

从以上两种不同形式的循环结构可以看出，循环结构中一定包含条件结构，用于确定何时终止执行循环体。

例 6 设计一个计算 $1+2+\cdots+100$ 的值的算法，并画出程序框图。

算法分析：

通常，我们按照下列过程计算 $1+2+\cdots+100$ 的值。

第1步, $0+1=1$.
 第2步, $1+2=3$.
 第3步, $3+3=6$.
 第4步, $6+4=10$.

 第100步, $4\,950+100=5\,050$.

显然, 这个过程中包含重复操作的步骤, 可以用循环结构表示. 分析上述计算过程, 可以发现每一步都可以表示为

第 $(i-1)$ 步的结果 $+i$ = 第 i 步的结果.

为了方便、有效地表示上述过程, 我们用一个累加变量 S 来表示每一步的计算结果, 即把 $S+i$ 的结果仍记为 S , 从而把第 i 步表示为

$$S = S + i,$$

其中 S 的初始值为 0, i 依次取 1, 2, ..., 100. 由于 i 同时记录了循环的次数, 所以也称为计数变量.

解决这一问题的算法是:

第一步, 令 $i=1$, $S=0$.

第二步, 若 $i \leq 100$ 成立, 则执行第三步; 否则, 输出 S , 结束算法.

第三步, $S = S + i$.

第四步, $i = i + 1$, 返回第二步.

程序框图:

① 这里的“=”是赋值号, 表示把 $S+i$ 的值仍赋给 S .

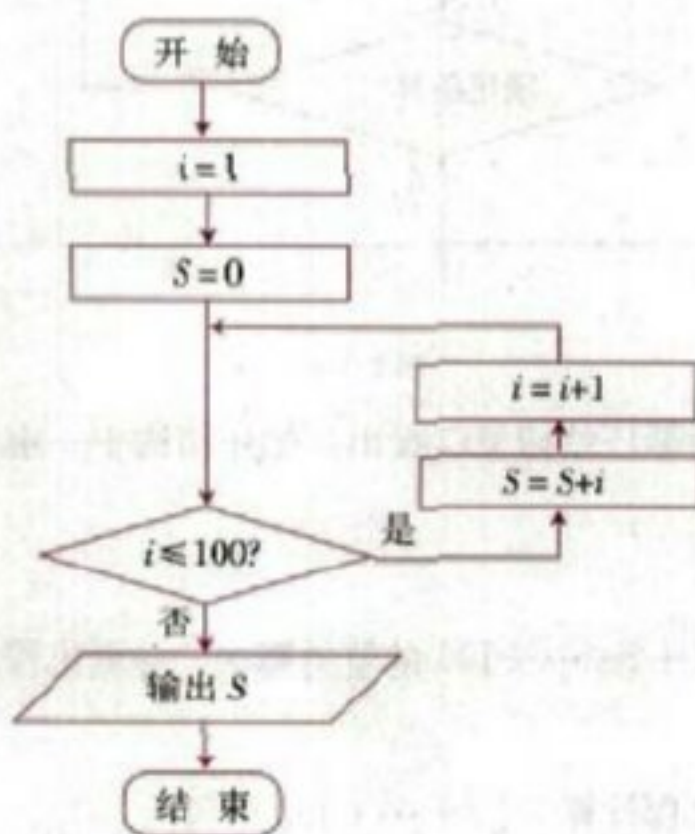


图 1.1-14

上述程序框图用的是当型循环结构, 如果用直到型循环结构表示, 则程序框图为(图 1.1-15);

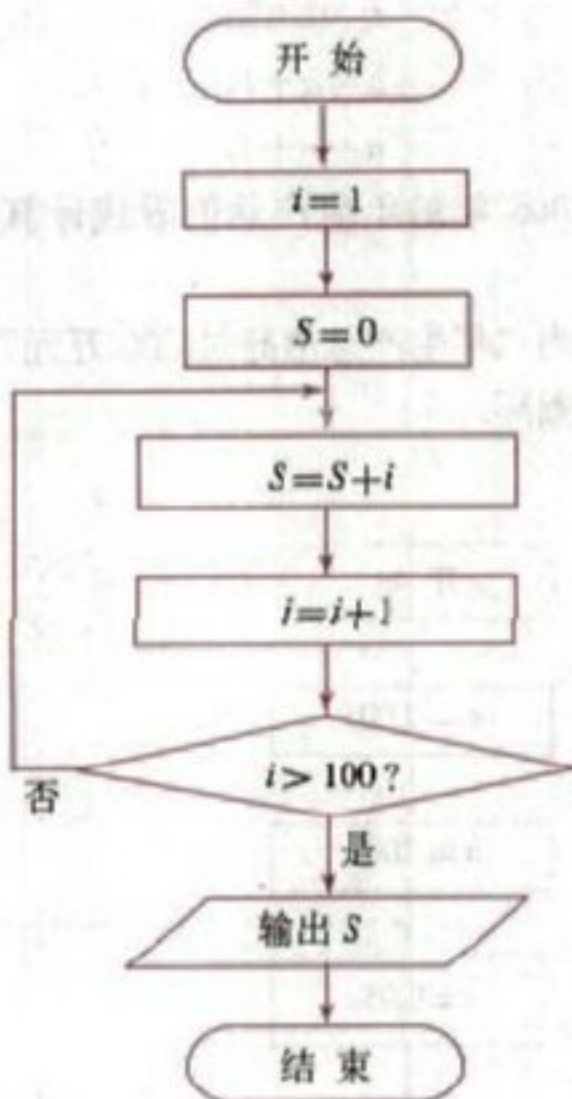


图 1.1-15



如何用自然语言表述图 1.1-15 中的算法? 改进这一算法, 表示输出 $1, 1+2, 1+2+3, \dots, 1+2+3+\dots+(n-1)+n(n \in \mathbf{N}^*)$ 的过程.

例 7 某工厂 2005 年的年生产总值为 200 万元, 技术革新后预计以后每年的年生产总值都比上一年增长 5%. 设计一个程序框图, 输出预计年生产总值超过 300 万元的最早年份.

算法分析:

先写出解决本例的算法步骤:

第一步, 输入 2005 年的年生产总值.

第二步, 计算下一年的年生产总值.

第三步, 判断所得的结果是否大于 300. 若是, 则输出该年的年份; 否则, 返回第二步.

由于“第二步”是重复操作的步骤, 所以本例可以用循环结构来实现. 我们按照“确定循环体”“初始化变量”“设定循环控制条件”的顺序来构造循环结构.

(1) 确定循环体：设 a 为某年的年生产总值， t 为年生产总值的年增长量， n 为年份，则循环体为

$$t=0.05a,$$

$$a=a+t,$$

$$n=n+1.$$

(2) 初始化变量：若将 2005 年的年生产总值看成计算的起始点，则 n 的初始值为 2005， a 的初始值为 200.

(3) 设定循环控制条件：当“年生产总值超过 300 万元”时终止循环，所以可通过判断“ $a>300$ ”是否成立来控制循环.

程序框图：

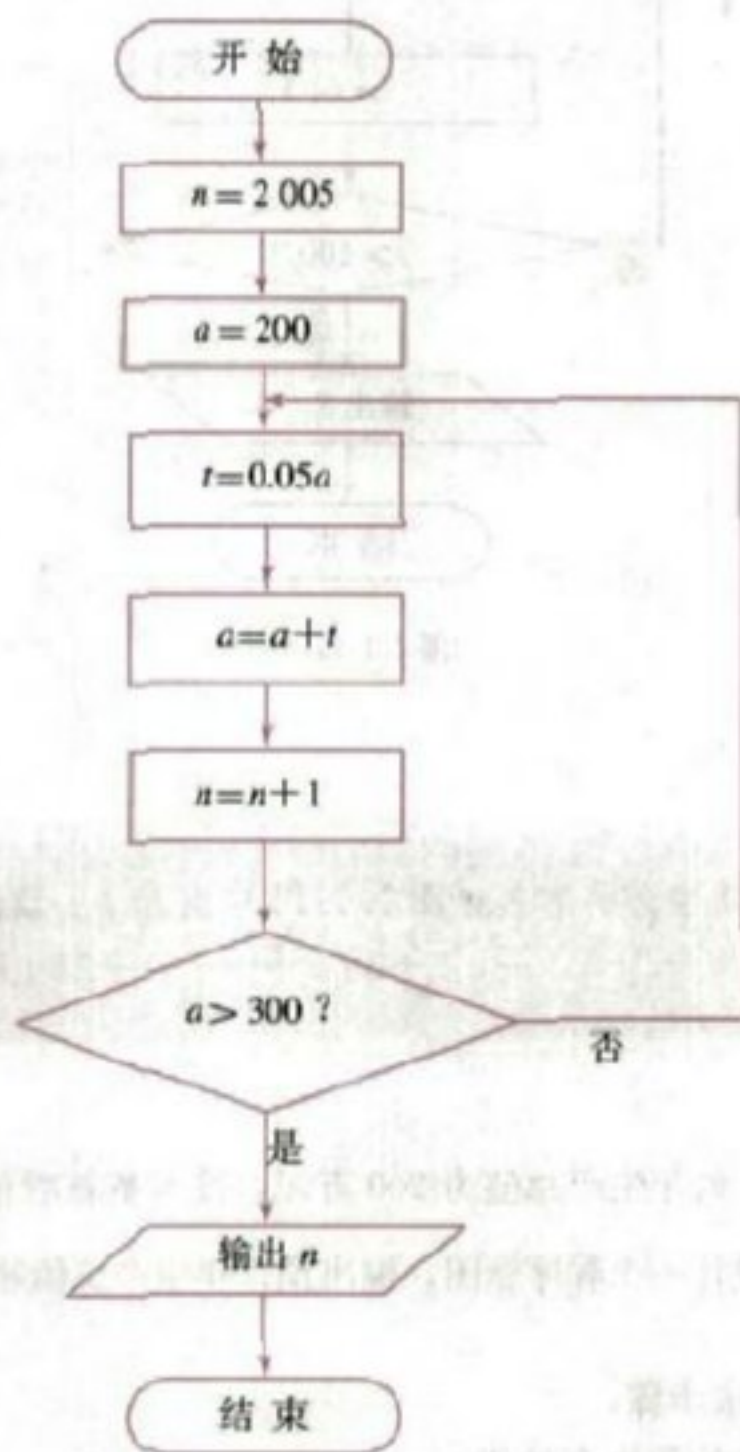


图 1.1-16



图 1.1-16 是包含直到型循环结构的程序框图，你能画出包含当型循环结构的程序框图吗？

3. 程序框图的画法

在用自然语言表述一个算法后，可以画出程序框图，用顺序结构、条件结构和循环结构来表示这个算法。这样表示的算法清楚、简练，便于阅读和交流。

下面，我们根据例 2 的算法步骤，利用三种基本逻辑结构画出程序框图，表示用“二分法”求方程 $x^2 - 2 = 0 (x > 0)$ 的近似解的算法。

(1) 算法步骤中的“第一步”“第二步”和“第三步”可以用顺序结构来表示(图 1.1-17)；

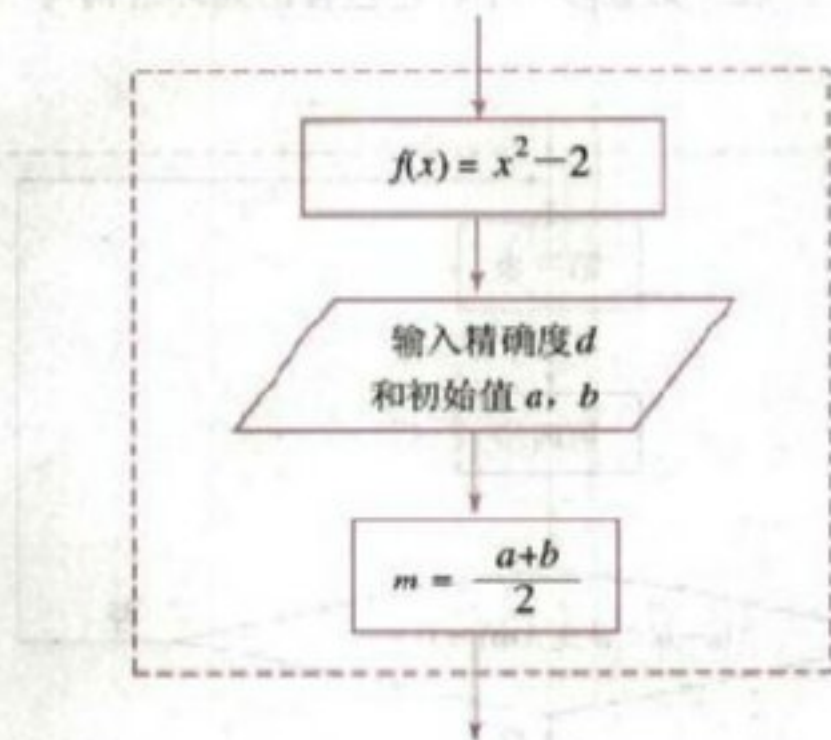


图 1.1-17

(2) 算法步骤中的“第四步”可以用条件结构来表示(图 1.1-18)。在这个条件结构中，“否”分支用“ $a = m$ ”表示含零点的区间为 $[m, b]$ ，并把这个区间仍记成 $[a, b]$ ；“是”分支用“ $b = m$ ”表示含零点的区间为 $[a, m]$ ，同样把这个区间仍记成 $[a, b]$ 。

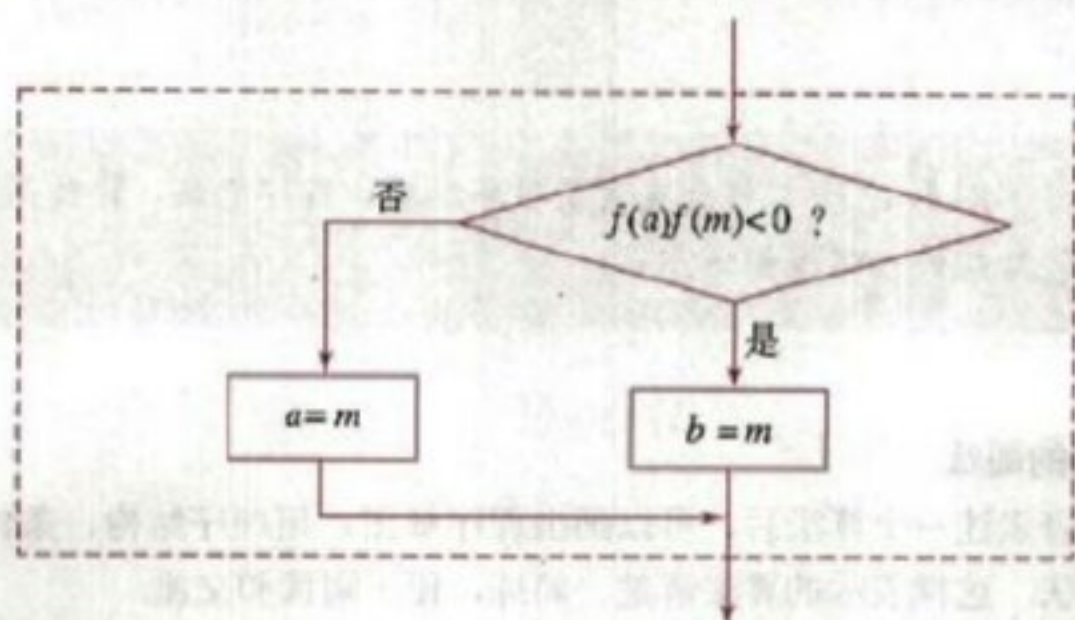


图 1.1-18

(3) 算法步骤中的“第五步”包含一个条件结构，这个条件结构与“第三步”“第四步”构成一个循环结构，循环体由“第三步”和“第四步”组成，终止循环的条件是“ $|a-b| < d$ 或 $f(m) = 0$ ”。在“第五步”中，还包含由循环结构与“输出 m ”组成的顺序结构（图 1.1-19）。

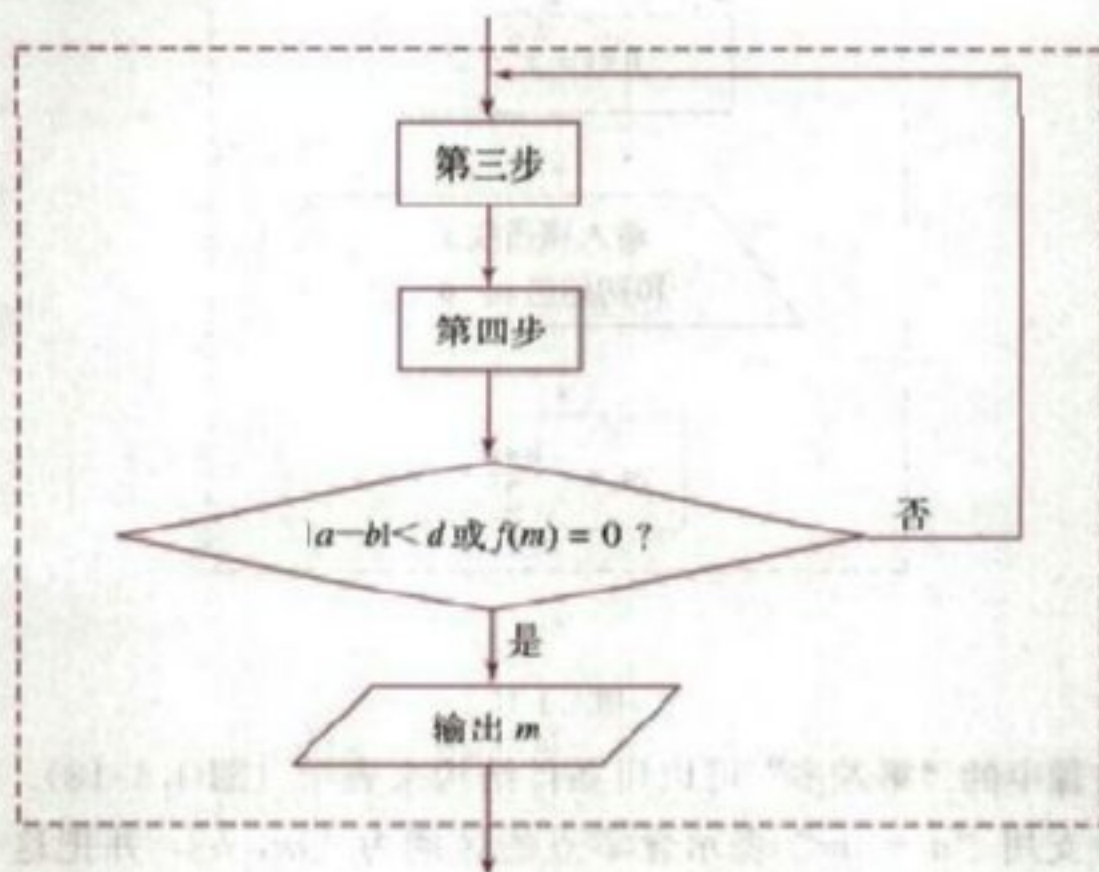


图 1.1-19

(4) 将各步骤的程序框图连接起来，并画出“开始”与“结束”两个终端框，就得到了表示整个算法的程序框图（图 1.1-20）。

从以上过程可以看出，设计一个算法的程序框图通常要经过以下步骤：

第一步，用自然语言表述算法步骤。

第二步，确定每一个算法步骤所包含的逻辑结构，并用相应的程序框图表示，得到该步骤的程序框图。

第三步，将所有步骤的程序框图用流程线连接起来，并加上终端框，得到表示整个算

法的程序框图.

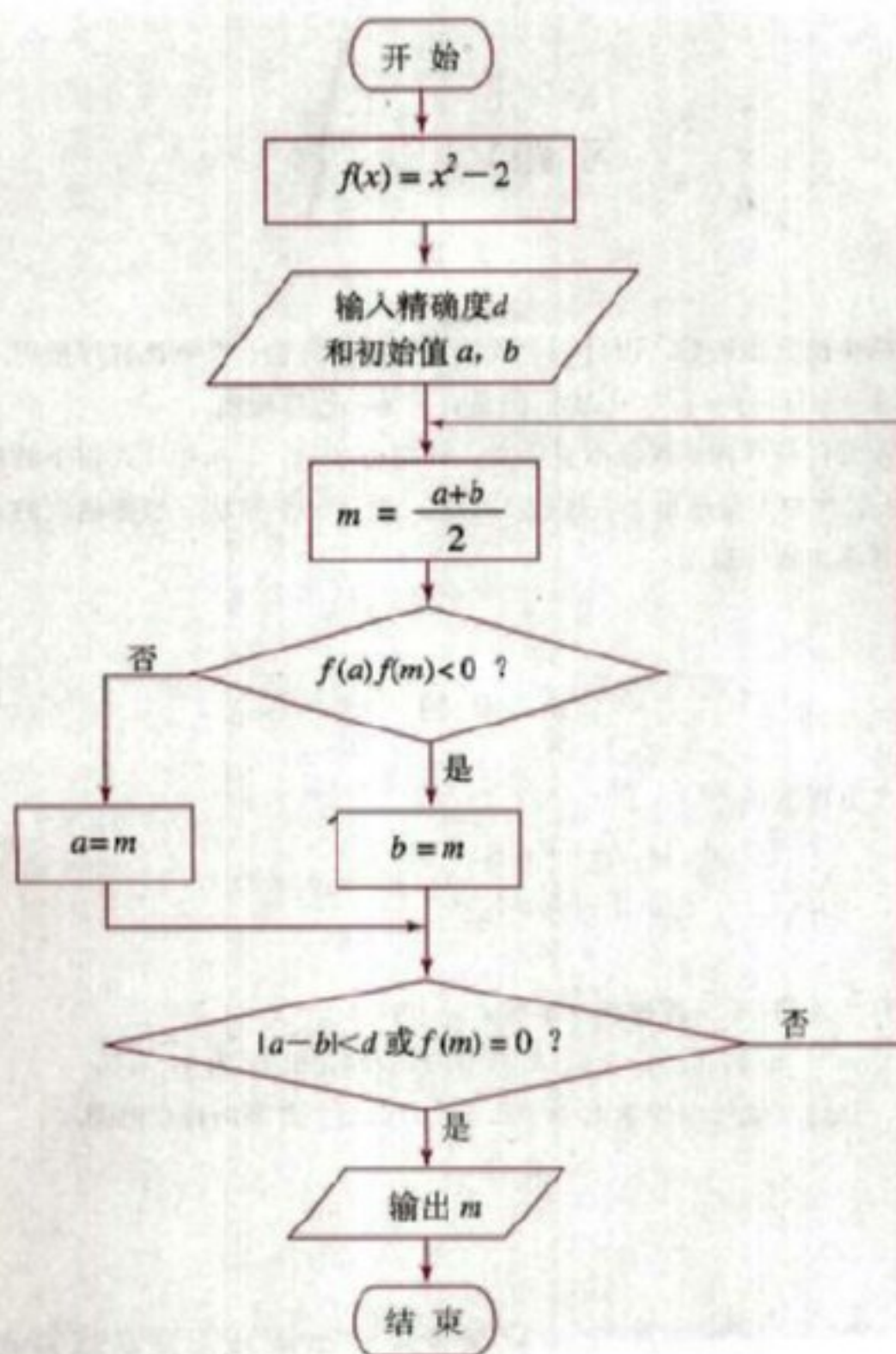


图 1.1-20

练习

设计一个用有理指数逼近无理指数 $5^{\sqrt{2}}$ 的算法，并估计 $5^{\sqrt{2}}$ 的近似值，画出算法的程序框图。

习题1.1

A 组

1. 找一个实际生活中的分段函数, 设计一个求该函数值的算法, 并画出程序框图.
2. 设计一个算法求 $1^2 + 2^2 + \dots + 99^2 + 100^2$ 的值, 并画出程序框图.
3. 某居民区的物业部门每月向居民收取卫生费, 计费方法是: 3 人和 3 人以下的住户, 每户收取 5 元; 超过 3 人的住户, 每超出 1 人加收 1.2 元. 设计一个算法, 根据输入的人数, 计算应收取的卫生费, 并画出程序框图.

B 组

1. 画出求二元一次方程组

$$\begin{cases} a_1x + b_1y = c_1, \\ a_2x + b_2y = c_2 \end{cases} \quad (a_1b_2 - a_2b_1 \neq 0)$$

的解的程序框图.

2. 某高中男子体育小组的 50 m 跑成绩 (单位: s) 为:

6.4, 6.5, 7.0, 6.8, 7.1, 7.3, 6.9, 7.4, 7.5.

设计一个算法, 从这些成绩中搜索出小于 6.8 s 的成绩, 并画出程序框图.

CHAPTER 1

1.2

基本算法语句



计算机完成任何一项任务都需要算法。但是，我们用自然语言或程序框图表示的算法，计算机是无法“理解”的。因此还需要将算法用计算机能够理解的程序设计语言（programming language）表示成计算机程序。

程序设计语言有很多种。为了实现算法的三种基本逻辑结构，各种程序设计语言中都包含下列基本的算法语句，并且形式是类似的。

输入语句 输出语句 赋值语句 条件语句 循环语句

我们使用的语句形式和语法规则与 BASIC^① 语言类似，稍加改造就可以在计算机上运行实现。

^① BASIC 是 Beginner's All-purpose Symbolic Instruction Code（初学者通用符号指令代码）的英文缩写，于 1964 年由美国的两教授设计，具有简单、易学的特点。

1.2.1 输入语句、输出语句和赋值语句

输入语句、输出语句分别与程序框图中的输入、输出框对应，用来输入和输出信息。赋值语句与程序框图中表示赋值的处理框对应，用来给变量赋值。

下面举例说明这几种语句的应用。

例 1 用描点法作函数 $y=x^3+3x^2-24x+30$ 的图象时，需要求出自变量和函数的一组对应值。编写程序，分别计算当 $x=-5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5$ 时的函数值。

算法分析：

根据题意，对于每一个输入的自变量的值，都要输出相应的函数值。写成算法步骤如下：

第一步，输入一个自变量 x 的值。

第二步，计算 $y=x^3+3x^2-24x+30$ 。

第三步, 输出 y .

程序框图:

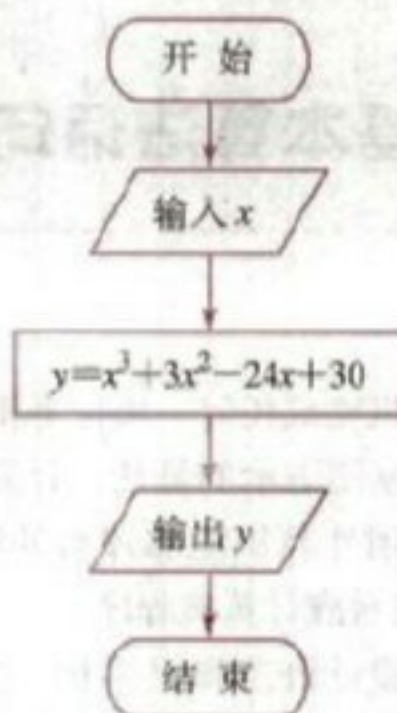


图 1.2-1

显然, 这是一个由顺序结构构成的算法. 按照程序框图中流程线的方向, 依次将程序框中的内容写成相应的算法语句, 就得到了相应的程序.

程序:

```

INPUT "x": x
y=x^3+3*x^2-24*x+30
PRINT y
END
  
```

在这个程序中, 第 1 行中的 INPUT 语句就是**输入语句**. 这个语句的一般格式是

INPUT “提示内容”; 变量

其中, “提示内容”一般是提示用户输入什么样的信息. 每次运行例 1 中的程序时, 依次输入 -5, -4, -3, -2, -1, 0, 1, 2, 3, 4, 5, 计算机每次都把新输入的值赋给变量 “x”, 并按 “x” 新获得的值计算变量 “y” 的值.

例如, 图 1.1-7 中的输入框可以转化为输入语句:

INPUT “a, b, c=”: a, b, c

例 1 中第 3 行的 PRINT 语句是**输出语句**, 它的一般格式是

PRINT “提示内容”; 表达式

我们看到, 用类 BASIC 语言编写的计算机程序是由若干语句行组成的, 计算机按语句行排列的顺序依次执行程序中的语句. 最后一行的 END 语句表示程序到此结束.

PRINT 语句可以在计算机的屏幕上输出常量、变量的值和系统信息。同输入语句一样，这里的表达式前也可以有“提示内容”。例如，图 1.1-7 中的输出框可以转化为输出语句：

```
PRINT "S="; S
```

例 2 编写程序，计算一个学生数学、语文、英语三门课的平均成绩。

算法分析：

先写出解决本例的算法步骤：

第一步，输入该学生数学、语文、英语三门课的成绩 a, b, c 。

第二步，计算 $y = \frac{a+b+c}{3}$ 。

第三步，输出 y 。

程序框图：

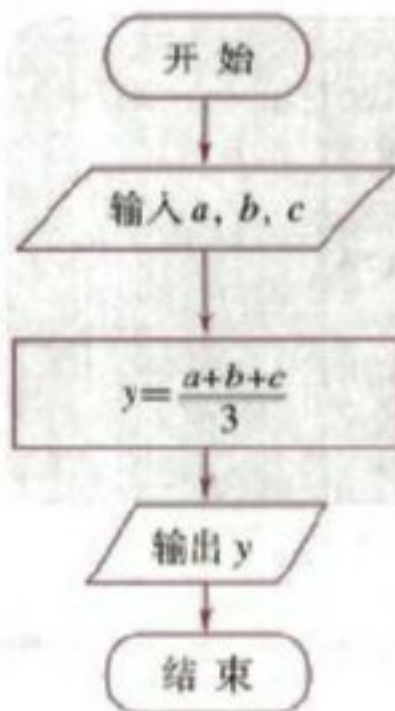


图 1.2-2

由于 PRINT 语句还可以用于输出数值计算的结果，所以这个算法可以写成下列程序。
程序：

```
INPUT "Maths="; a
INPUT "Chinese="; b
INPUT "English="; c
PRINT "The average="; (a+b+c)/3
END
```

除了输入语句，例 1 中第 2 行的**赋值语句**也可以给变量提供初值，它的一般格式是

```
变量 = 表达式
```


顾名思义,赋值语句就是将表达式所代表的值赋给变量.赋值语句中的“=”叫做赋值号,它和数学中的等号不完全一样.计算机执行赋值语句时,先计算“=”右边表达式的值,然后把这个值赋给“=”左边的变量.下面我们来看两个例子.

例 3 给一个变量重复赋值.

程序:

```
A=10
A=A+15
PRINT A
END
```

A 的输出值是多少?

例 4 交换两个变量 A 和 B 的值,并输出交换前后的值.

程序:

```
INPUT A,B
PRINT A,B
x=A
A=B
B=x
PRINT A,B
END
```

程序中的 3 个赋值语句用来交换两个变量的值,变量 x 的作用是什么?

练习

1. 已知华氏温度与摄氏温度的转换公式是:

$$(\text{华氏温度} - 32) \times \frac{5}{9} = \text{摄氏温度}.$$

编写一个程序,输入一个华氏温度,输出其相应的摄氏温度.

2. 编写一个程序,计算两个非 0 实数的加、减、乘、除运算的结果.

(要求输入两个非 0 实数,输出运算结果.)

3. 将图 1.1-7 中的程序框图转化为程序.

4. 春节到了,糖果店的售货员忙极了.请你设计一个程序,帮助售货员算账.已知水果糖每千克 10.4 元,奶糖每千克 15.6 元,果仁巧克力每千克 25.2 元.那么依次购买这三种糖果 a , b , c 千克,应收取多少钱?



(第 4 题)

1.2.2 条件语句

条件语句与程序框图中的条件结构相对应. 图 1.1-9 中的条件结构对应的条件语句是:

```
IF 条件 THEN
    语句体
END IF
```

当计算机执行上述语句时, 首先对 IF 后的条件进行判断, 如果 (IF) 条件符合, 那么 (THEN) 执行语句体, 否则执行 END IF 之后的语句.

图 1.1-8 中的条件结构对应的条件语句是:

```
IF 条件 THEN
    语句体 1
ELSE
    语句体 2
END IF
```

当计算机执行上述语句时, 首先对 IF 后的条件进行判断, 如果 (IF) 条件符合, 那么 (THEN) 执行语句体 1, 否则 (ELSE) 执行语句体 2.

例 5 编写一个程序, 求实数 x 的绝对值.

算法分析:

首先, 我们来设计求实数 x 的绝对值的算法. 因为实数 x 的绝对值为

$$|x| = \begin{cases} x & (x \geq 0), \\ -x & (x < 0), \end{cases}$$

所以算法步骤可以写成:

第一步, 输入一个实数 x .

第二步, 判断 x 的符号. 若 $x \geq 0$, 则输出 x ; 否则, 输出 $-x$.

显然, “第二步” 可以用条件结构来实现.

程序框图：

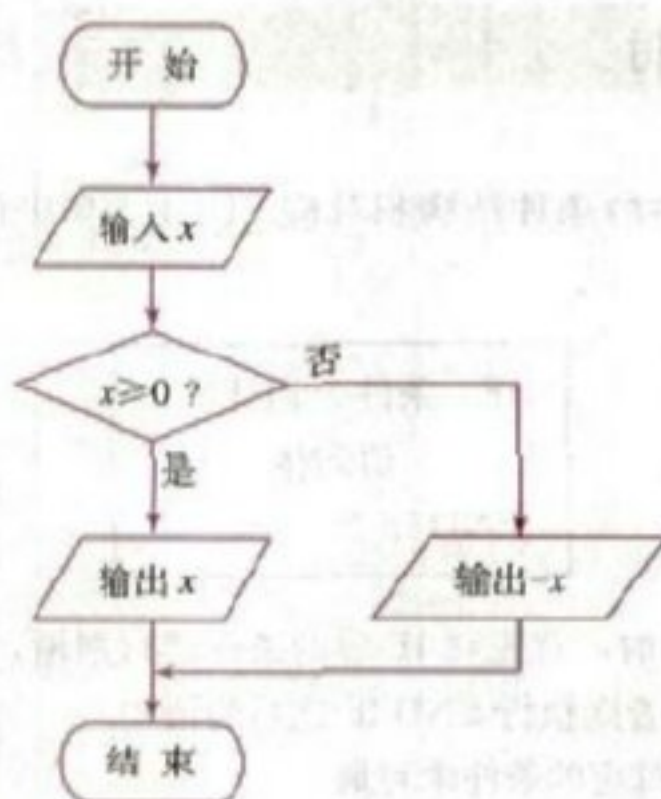


图 1.2-3

程序：

```

INPUT x
IF x ≥ 0 THEN
    PRINT x
ELSE
    PRINT -x
END IF
END
  
```



阅读下面的程序，你能得出什么结论？

```

INPUT x
IF x < 0 THEN
    x = -x
END IF
PRINT x
END
  
```


例 6 把图 1.1-11 中的程序框图转化为程序.

算法分析:

观察图 1.1-11 中的程序框图可以发现, 其中包含两个条件结构, 而且内层的条件结构是外层的条件结构的一个分支. 所以, 可以用 “IF—THEN—ELSE—END IF” 语句来完成转化.

程序:

```

INPUT "a, b, c="; a, b, c
d=b^2-4*a*c
IF d>=0 THEN
    p=-b/(2*a)
    q=SQR①(d)/(2*a)
    IF d=0 THEN
        PRINT "x1=x2="; p
    ELSE
        PRINT "x1, x2="; p+q, p-q
    END IF
ELSE
    PRINT "No real root."
END IF
END
  
```

^①SQR ()

是一个函数, 用来求某个非负数的算术平方根, 即 $\text{SQR}(x)=\sqrt{x}$.

例 7 编写程序, 使任意输入的 3 个整数按从大到小的顺序输出.

算法分析:

用 a, b, c 表示输入的 3 个整数. 为了节约变量, 把它们重新排列后, 仍用 a, b, c 表示, 并使 $a \geq b \geq c$. 具体操作步骤如下:

第一步, 输入 3 个整数 a, b, c .

第二步, 将 a 与 b 比较, 并把小者赋给 b , 大者赋给 a .

第三步, 将 a 与 c 比较, 并把小者赋给 c , 大者赋给 a (此时 a 已是三者中最大的).

第四步, 将 b 与 c 比较, 并把小者赋给 c , 大者赋给 b (此时 a, b, c 已按从大到小的顺序排列好).

第五步, 按顺序输出 a, b, c .

如图 1.2-4 所示, 上述操作步骤可以用程序框图更直观地表达出来.

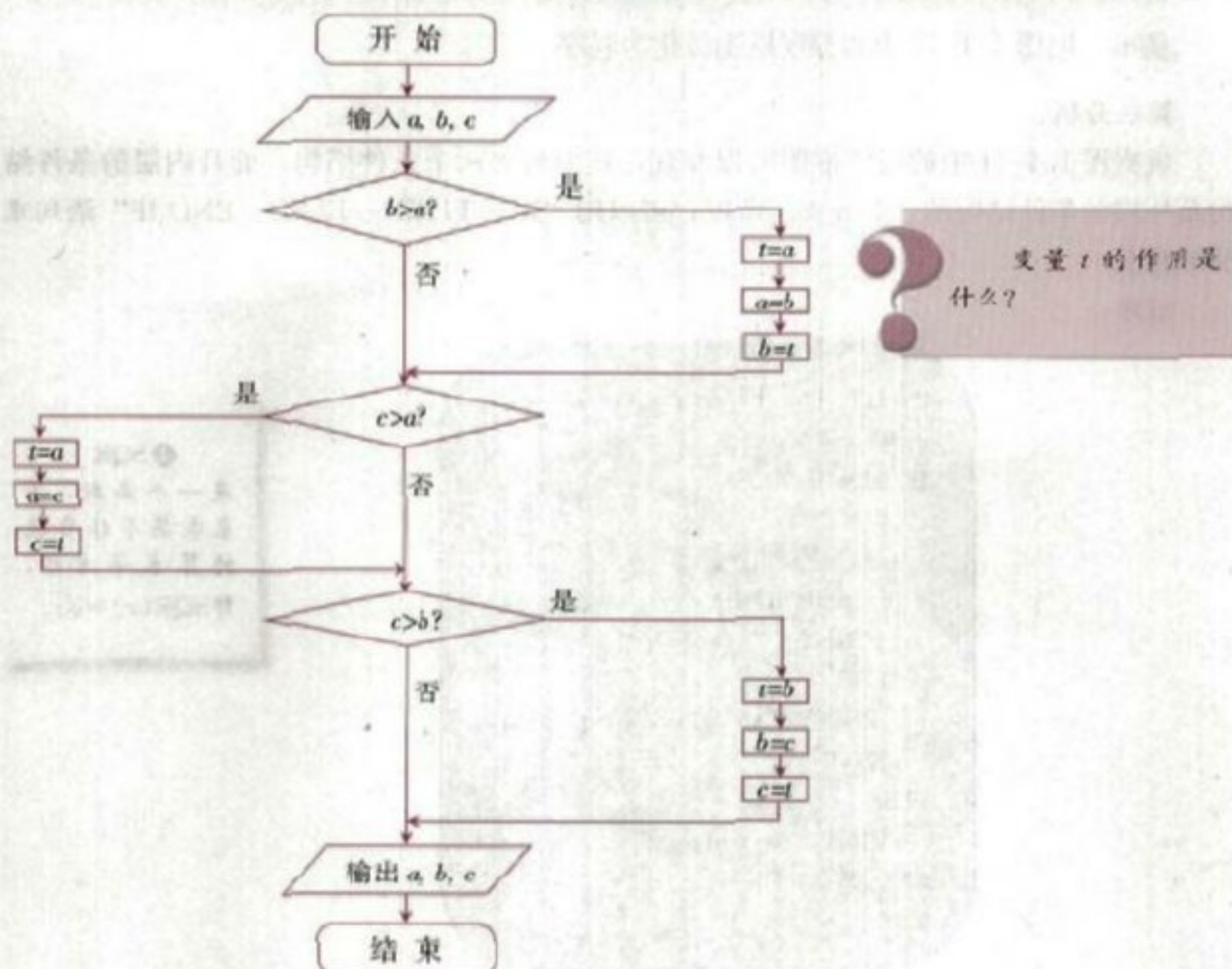


图 1.2-4

根据程序框图，写出相应的计算机程序。

```

INPUT "a, b, c="; a, b, c
IF b > a THEN
    t = a
    a = b
    b = t
END IF
IF c > a THEN
    t = a
    a = c
    c = t
END IF
IF c > b THEN
    t = b
    b = c
    c = t
END IF
PRINT a, b, c
END
  
```


练习

1. 将图 1.1-10 中的程序框图转化为程序。
2. 读程序，说明程序的运行过程。

```

INPUT "Please input an integer: "; x
IF 9 < x AND x < 100 THEN
    a = x \ 10❶
    b = x MOD 10
    x = 10 * b + a
    PRINT x
END IF
END
  
```

❶ 算术运算符 \ 和 MOD 分别用来取商和余数。这里， a 等于 x 除以 10 的商，即把 x 的十位取出来； b 等于 x 除以 10 的余数，即把 x 的个位取出来。

3. 编写一个程序，判断任意输入的整数的奇偶性。
4. 闰年是指能被 4 整除但不能被 100 整除，或者能被 400 整除的年份。编写一个程序，判断输入的年份是否为闰年。

1.2.3 循环语句

循环语句与程序框图中的循环结构相对应。一般程序设计语言中都有直到型 (UNTIL) 和当型 (WHILE) 两种循环语句结构，分别对应于程序框图中的直到型和当型循环结构。

图 1.1-12 中的直到型循环结构对应的 UNTIL 语句是：

```

DO
    循环体❶
LOOP UNTIL 条件
  
```

❶ 这里的循环体是由计算机反复执行的一组语句构成的。

当计算机执行上述语句时，先执行一次 DO 和 UNTIL 之间的循环体，再对 UNTIL 后的条件进行判断。如果条件不符合，继续执行循环体；然后再检查上述条件，如果条件仍不符合，再次执行循环体，直到条件符合时为止。这时，计算机将不执行循环体，直接跳到 UNTIL 语句后，接着执行 UNTIL 语句之后的语句。

下面, 我们根据图 1.1-15 中的程序框图, 用 UNTIL 语句编写计算机程序.

```
i = 1
S = 0
DO
  S = S + i
  i = i + 1
LOOP UNTIL i > 100
PRINT S
END
```

图 1.1-13 中的当型循环结构对应的 WHILE 语句是:

```
WHILE 条件
  循环体
WEND
```

当计算机遇到 WHILE 语句时, 先判断条件的真假, 如果条件符合, 就执行 WHILE 和 WEND 之间的循环体; 然后再检查上述条件, 如果条件仍符合, 再次执行循环体, 这个过程反复进行, 直到某一次条件不符合为止. 这时, 计算机将不执行循环体, 直接跳到 WEND 语句后, 接着执行 WEND 之后的语句.

我们也可以根据图 1.1-14 中的程序框图, 用 WHILE 语句编写计算机程序.

程序:

```
i = 1
S = 0
WHILE i <= 100
  S = S + i
  i = i + 1
WEND
PRINT S
END
```

例 8 修改本节例 1 的程序, 连续输入自变量的 11 个取值, 输出相应的函数值.

算法分析:

与本节例 1 不同的是, 本例要求连续输入自变量的 11 个取值, 并输出相应的函数值. 先写出解决本例的算法步骤:

第一步, 输入自变量 x 的值.

第二步, 计算 $y = x^3 + 3x^2 - 24x + 30$.

第三步, 输出 y .

第四步, 记录输入次数.

第五步, 判断输入的次数是否大于 11. 若是, 则结束算法; 否则, 返回第一步.

显然, 可以用计数变量 n ($1 \leq n \leq 11$) 记录次数, 通过循环结构来实现算法.

程序框图:

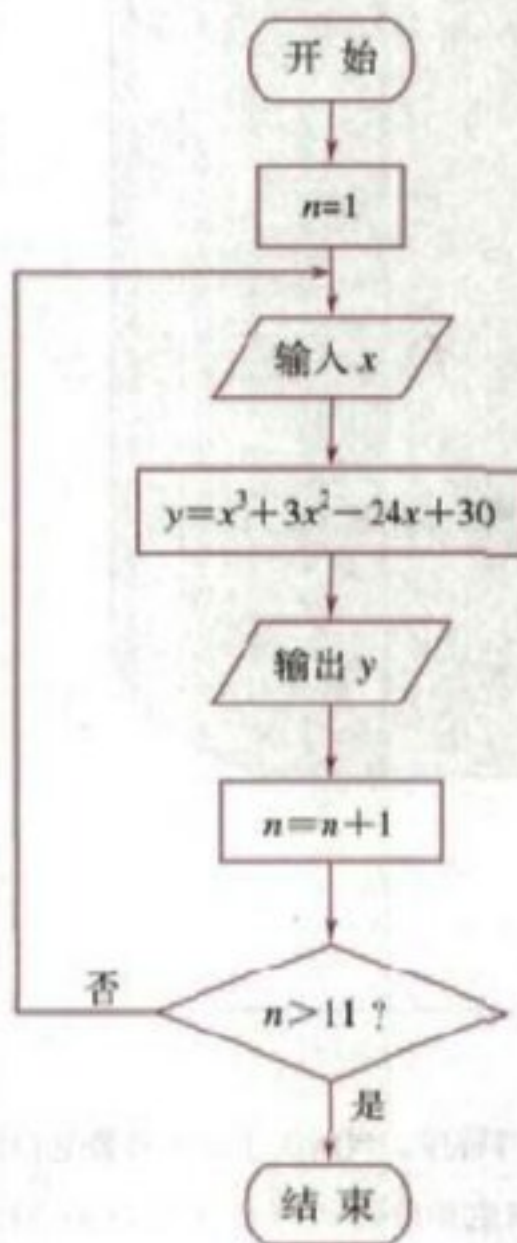


图 1.2-5

程序:

```
n=1
DO
  INPUT x
  y=x^3+3*x^2-24*x+30
  PRINT y
  n=n+1
LOOP UNTIL n>11
END
```


图 1.1-20 中的程序框图包含了顺序结构、条件结构和循环结构. 下面, 我们把这个程序框图转化为相应的程序.

```

INPUT "a, b, d="; a, b, d
DO
  m=(a+b)/2
  g=a^2-2
  f=m^2-2
  IF g*f<0 THEN
    b=m
  ELSE
    a=m
  ENDIF
LOOP UNTIL ABS(a-b)<d OR f=0
PRINT m
END

```

① ABS ()
是一个函数, 用
来求某个数的绝
对值. 即 ABS
 $(x) = |x|$.

练习

1. 根据图 1.1-2 中的程序框图编写程序, 判断大于 2 的整数是否为质数.
2. 编写程序, 输入正整数 n , 计算它的阶乘 $n!$ ($n! = n \times (n-1) \times \cdots \times 3 \times 2 \times 1$).

习题 1.2

A 组

1. 读程序, 写出程序表示的函数.

```

INPUT x
IF x < 0 THEN
    y = -x + 1
ELSE
    IF x = 0 THEN
        y = 0
    ELSE
        y = x + 1
    END IF
END IF
PRINT y
END

```

2. 编写一个程序, 输入梯形的上底、下底和高的值, 计算并输出其面积.
 3. 编写一个程序, 计算下面 n ($n \in \mathbf{N}^*$) 个数的和:

$$2, \frac{3}{2}, \frac{4}{3}, \frac{5}{4}, \dots, \frac{n+1}{n}.$$

B 组

1. 编写一个程序, 求二元一次方程组 $\begin{cases} a_1x + b_1y = c_1, \\ a_2x + b_2y = c_2 \end{cases}$ ($a_1b_2 - a_2b_1 \neq 0$) 的解.
 2. 某牛奶厂 2002 年初有资金 1 000 万元, 由于引进了先进生产设备, 资金年平均增长率可达到 50%. 请你设计一个程序, 计算这家牛奶厂 2008 年底的资金总额.
 3. 编写一个程序, 对于函数

$$y = \begin{cases} x & (x < 1), \\ 2x - 1 & (1 \leq x < 10), \\ 3x - 11 & (x \geq 10), \end{cases}$$

输入 x 的值, 输出相应的函数值.

4. 编写一个程序, 计算 $s = a + aa + aaa + aaaa + \dots + \underbrace{aa \cdots a}_n$ (例如 $2 + 22 + 222 + 2\,222 + 22\,222$, 共有 5 个数相加) 的值, 其中 $a \in \mathbf{N}^*$, 且 $a \leq 9$, 要求输入数字 a 和相加的数的个数 n .

CHAPTER 1

1.3

算 法 案 例

在前面两节中，我们学习了一些简单的算法，如求方程的近似解的二分法、判定质数的算法等，对算法已经有了一个初步的了解。下面，我们将通过几个算法案例，进一步体会算法的思想。

案例1 辗转相除法与更相减损术

在小学，我们学过求两个正整数的最大公约数的方法：先用两个数公有的质因数连续去除，一直除到所得的商是互质数为止，然后把所有的除数连乘起来。

当两个数公有的质因数较大时（如 8 251 与 6 105），使用上述方法求最大公约数就比较困难。下面我们介绍一种古老而有效的算法——**辗转相除法**。这种算法是由欧几里得在公元前 300 年左右首先提出的，因而又叫**欧几里得算法**。

例如，用辗转相除法求 8 251 与 6 105 的最大公约数，我们可以考虑用两数中较大的数除以较小的数，求得商和余数：

$$8\,251 = 6\,105 \times 1 + 2\,146.$$

由此可得，6 105 与 2 146 的公约数也是 8 251 与 6 105 的公约数，反过来，8 251 与 6 105 的公约数也是 6 105 与 2 146 的公约数，所以它们的最大公约数相等。

对 6 105 与 2 146 重复上述步骤：

$$6\,105 = 2\,146 \times 2 + 1\,813.$$

同理，2 146 与 1 813 的最大公约数也是 6 105 与 2 146 的最大公约数。继续重复上述步骤：

$$2\,146 = 1\,813 \times 1 + 333,$$

$$1\,813 = 333 \times 5 + 148,$$

$$333 = 148 \times 2 + 37,$$

$$148 = 37 \times 4.$$

最后的除数 37 是 148 和 37 的最大公约数，也就是 8 251 与 6 105 的最大公约数。

这就是辗转相除法。由除法的性质可以知道，对于任意两个正整数，上述除法步骤总可以在有限步之后完成，从而总可以用辗转相除法求出两个正整数的最大公约数。

例如，求 18 与 30 的最大公约数：

$$\begin{array}{r|rr} 2 & 18 & 30 \\ 3 & 9 & 15 \\ & 3 & 5 \end{array}$$

所以，18 与 30 的最大公约数是 $2 \times 3 = 6$ 。



你能把辗转相除法编成一个计算机程序吗?

算法分析:

从上面的例子可以看出,辗转相除法中包含重复操作的步骤,因此可以用循环结构来构造算法.

算法步骤如下:

第一步,给定两个正整数 m, n .

第二步,计算 m 除以 n 所得的余数 r .

第三步, $m \leftarrow n, n \leftarrow r$.

第四步,若 $r=0$,则 m, n 的最大公约数等于 m ; 否则,返回第二步.

程序框图:

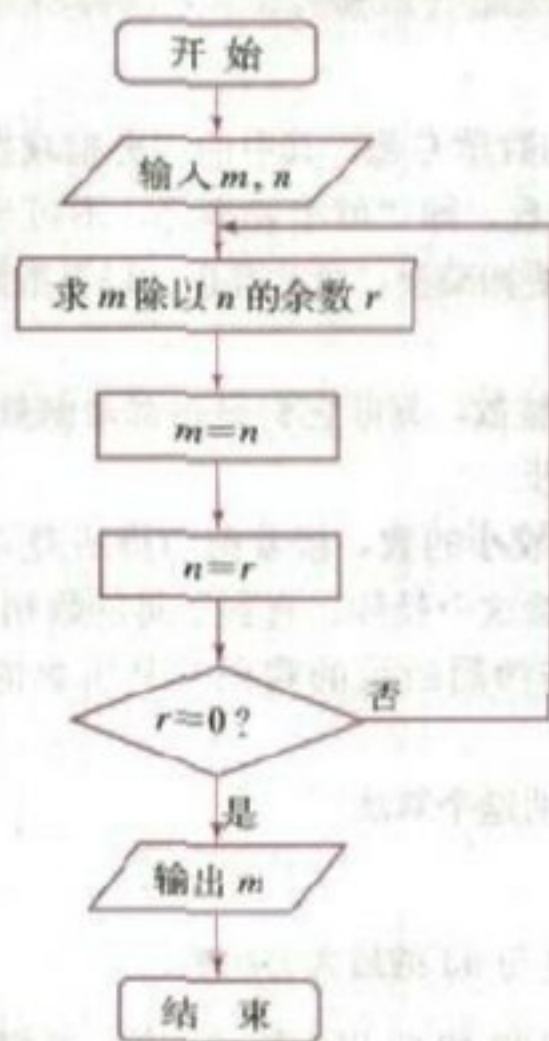


图 1.3-1

程序：

```

INPUT m, n
DO
  r = m MOD n
  m = n
  n = r
LOOP UNTIL r = 0
PRINT m
END

```



你能用当型循环结构构造算法，求两个正整数的最大公约数吗？试写出算法步骤、程序框图和程序。

《九章算术》是中国古代的数学专著，其中的“更相减损术”也可以用来求两个数的最大公约数，即“可半者半之，不可半者，副置分母、子之数，以少减多，更相减损，求其等也，以等数约之。”

翻译为现代语言如下：

第一步，任意给定两个正整数，判断它们是否都是偶数。若是，用2约简；若不是，执行第二步。

第二步，以较大的数减去较小的数，接着把所得的差与较小的数比较，并以大数减小数。继续这个操作，直到所得的数相等为止，则这个数（等数）或这个数与约简的数的乘积就是所求的最大公约数。

下面我们用一个例子来说明这个算法。

例1 用更相减损术求98与63的最大公约数。

解：由于63不是偶数，把98和63以大数减小数，并辗转相减，如图1.3-2所示：



《九章算术》收录了246个数学问题及其解法，分为方田、粟米、衰分、少广、商功、均输、盈不足、方程和勾股九章，算法“更相减损术”包含在方田章中。

$$98-63=35$$

$$63-35=28$$

$$35-28=7$$

$$28-7=21$$

$$21-7=14$$

$$14-7=7$$

图 1.3-2

所以, 98 和 63 的最大公约数等于 7.



把更相减损术与辗转相除法比较, 你有什么发现? 你能根据更相减损术设计程序, 求两个正整数的最大公约数吗?

案例 2 秦九韶算法

怎样求多项式 $f(x)=x^5+x^4+x^3+x^2+x+1$ 当 $x=5$ 时的值呢?

一个自然的做法是把 5 代入多项式 $f(x)$, 计算各项的值, 然后把它们加起来. 这时, 我们一共做了 $1+2+3+4=10$ 次乘法运算, 5 次加法运算.

另一种做法是先计算 x^2 的值, 然后依次计算 $x^2 \cdot x$, $(x^2 \cdot x) \cdot x$, $((x^2 \cdot x) \cdot x) \cdot x$ 的值, 这样每次都可以利用上一次计算的结果. 这时, 我们一共做了 4 次乘法运算, 5 次加法运算.

第二种做法与第一种做法相比, 乘法的运算次数减少了, 因而能够提高运算效率. 对于计算机来说, 做一次乘法运算所用的时间比做一次加法运算要长得多, 所以采用第二种做法, 计算机能更快地得到结果.

有没有更有效的算法呢? 我国南宋时期的数学家秦九韶 (约 1202—1261) 在他的著作《数书九章》中提出了下面的算法.

把一个 n 次多项式 $f(x)=a_n x^n+a_{n-1} x^{n-1}+\cdots+a_1 x+a_0$ 改写成如下形式:

$$\begin{aligned} f(x) &= a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0 \\ &= (a_n x^{n-1} + a_{n-1} x^{n-2} + \cdots + a_1) x + a_0 \\ &= ((a_n x^{n-2} + a_{n-1} x^{n-3} + \cdots + a_2) x + a_1) x + a_0 \\ &= \cdots \\ &= (\cdots ((a_n x + a_{n-1}) x + a_{n-2}) x + \cdots + a_1) x + a_0. \end{aligned}$$

求多项式的值时, 首先计算最内层括号内一次多项式的值, 即

$$v_1 = a_n x + a_{n-1},$$

然后由内向外逐层计算一次多项式的值, 即

$$v_2 = v_1 x + a_{n-2},$$

$$v_3 = v_2 x + a_{n-3},$$

...

$$v_n = v_{n-1} x + a_0,$$

这样, 求 n 次多项式 $f(x)$ 的值就转化为求 n 个一次多项式的值.

上述方法称为**秦九韶算法**. 直到今天, 这种算法仍是多项式求值比较先进的算法.

例2 已知一个5次多项式为

$$f(x) = 5x^5 + 2x^4 + 3.5x^3 - 2.6x^2 + 1.7x - 0.8,$$

用秦九韶算法求这个多项式当 $x=5$ 时的值.

解: 根据秦九韶算法, 把多项式改写成如下形式:

$$f(x) = (((((5x+2)x+3.5)x-2.6)x+1.7)x-0.8.$$

按照从内到外的顺序, 依次计算一次多项式当 $x=5$ 时的值:

$$v_0 = 5;$$

$$v_1 = 5 \times 5 + 2 = 27;$$

$$v_2 = 27 \times 5 + 3.5 = 138.5;$$

$$v_3 = 138.5 \times 5 - 2.6 = 689.9;$$

$$v_4 = 689.9 \times 5 + 1.7 = 3\,451.2;$$

$$v_5 = 3\,451.2 \times 5 - 0.8 = 17\,255.2.$$

所以, 当 $x=5$ 时, 多项式的值等于 17 255.2.



用秦九韶算法求 n 次多项式 $f(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_1 x + a_0$ 当 $x=x_0$ (x_0 是任意实数) 时的值, 需要多少次乘法运算, 多少次加法运算?

计算机的一个很重要的特点就是运算速度快, 但即便如此, 算法好坏的一个重要标志仍然是运算的次数. 如果一个算法从理论上需要超出计算机允许范围内的运算次数, 那么这样的算法就只能是一个理论的算法. 据说国际象棋一盘棋的可能下法有 10^{100} 种, 比整个宇宙中的原子还多. 因此, 用枚举的办法穷尽国际象棋所有可能下法的算法是永远不可能实现的.

算法分析:

观察上述秦九韶算法中的 n 个一次式, 可见 v_k 的计算要用到 v_{k-1} 的值. 若令 $v_0 = a_n$, 我们可以得到下面的公式:

$$\begin{cases} v_0 = a_n, \\ v_k = v_{k-1} x + a_{n-k} \quad (k=1, 2, \cdots, n). \end{cases}$$

这是一个在秦九韶算法中反复执行的步骤, 因此可用循环结构来实现.

算法步骤如下:

第一步, 输入多项式次数 n 、最高次项的系数 a_n 和 x 的值.

第二步, 将 v 的值初始化为 a_n , 将 i 的值初始化为 $n-1$.

第三步, 输入 i 次项的系数 a_i .

第四步, $v = vx + a_i$, $i = i - 1$.

第五步, 判断 i 是否大于或等于 0. 若是, 则返回第三步; 否则, 输出多项式的值 v .

程序框图:

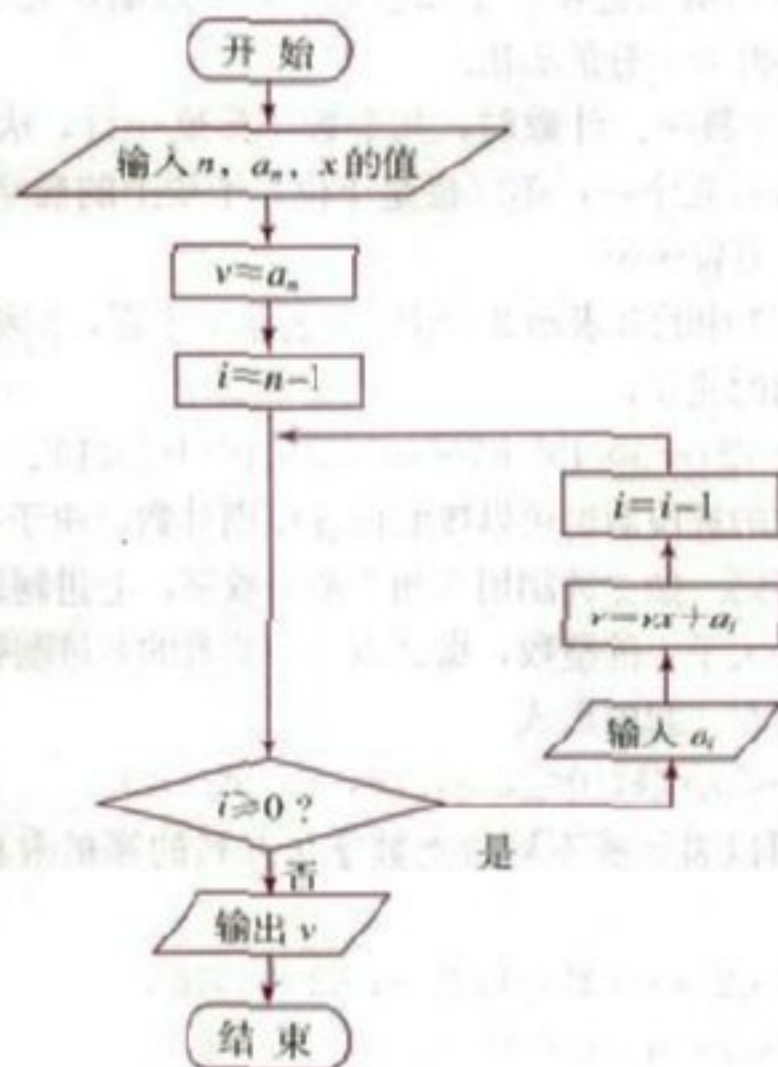


图 1.3-3

程序:

```

INPUT "n="; n
INPUT "a_n="; a
INPUT "x="; x
v=a
i=n-1
WHILE i>=0
    PRINT "i="; i
    INPUT "a_i="; a
    v=v*x+a
    i=i-1
WEND
PRINT v
END
  
```


案例3 进位制

进位制是人们为了计数和运算方便而约定的记数系统,约定满二进一,就是二进制;满十进一,就是十进制;满十二进一,就是十二进制;满六十进一,就是六十进制;等等.也就是说,“满几进一”就是几进制,几进制的**基数**^①就是几.

① 基数都是大于1的整数.

在日常生活中,我们最熟悉、最常用的是十进制,据说这与古人曾以手指计数有关.爱好天文学的古人也曾经采用七进制、十二进制、六十进制,至今我们仍然使用一周七天、一年十二个月、一小时六十分钟的历法.

十进制使用0~9十个数字.计数时,几个数字排成一行,从右起,第一位是个位,个位上的数字是几,就表示几个一;第二位是十位,十位上的数字是几,就表示几个十;接着依次是百位、千位、万位……

例如,十进制数3 721中的3表示3个千,7表示7个百,2表示2个十,1表示1个一.于是,我们得到下面的式子:

$$3\,721 = 3 \times 10^3 + 7 \times 10^2 + 2 \times 10^1 + 1 \times 10^0.$$

与十进制类似,其他的进位制也可以按照位置原则计数.由于每一种进位制的基数不同,所用的数字个数也不同.如二进制用0和1两个数字,七进制用0~6七个数字.

一般地,若 k 是一个大于1的整数,那么以 k 为基数的 k 进制数可以表示为一串数字连写在一起的形式

$$a_n a_{n-1} \cdots a_1 a_0 (k) \quad (0 \leq a_n < k, 0 \leq a_{n-1}, \cdots, a_1, a_0 < k).$$

其他进位制的数也可以表示成不同位上数字与基数的幂的乘积之和的形式,如

$$110\,011_{(2)} = 1 \times 2^5 + 1 \times 2^4 + 0 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0,$$

$$7\,342_{(8)} = 7 \times 8^3 + 3 \times 8^2 + 4 \times 8^1 + 2 \times 8^0.$$

为了区分不同的进位制,常在数的右下角标明基数,如二进制数 $10_{(2)}$,七进制数 $260_{(7)}$,十进制数一般不标注基数.

探究

若 $a_n a_{n-1} \cdots a_1 a_0 (k)$ 表示一个 k 进制数,请你把它写成各位上数字与 k 的幂的乘积之和的形式.



二进制只用0和1两个数字,这正好与电路的通和断两种状态相对应,因此计算机内部都使用二进制.计算机在进行数的运算时,先把接收到的数转化成二进制数进行运算,再把运算结果转化为十进制数输出.

十进制数与其他进位制数之间是怎样转化的呢?下面,我们用例子来说明.

例 3 把二进制数 $110\ 011_{(2)}$ 化为十进制数.

分析: 先把二进制数写成不同位上数字与 2 的幂的乘积之和的形式, 再按照十进制数的运算规则计算出结果.

$$\begin{aligned}\text{解: } 110\ 011_{(2)} &= 1 \times 2^5 + 1 \times 2^4 + 0 \times 2^3 + 0 \times 2^2 + 1 \times 2^1 + 1 \times 2^0 \\ &= 1 \times 32 + 1 \times 16 + 1 \times 2 + 1 \\ &= 51.\end{aligned}$$



如何改进上述算法, 把其他进位制数化为十进制数?

例 4 设计一个算法, 把 k 进制数 a (共有 n 位) 化为十进制数 b .

算法分析:

从例 3 的计算过程可以看出, 计算 k 进制数 a 的右数第 i 位数字 a_i 与 k^{i-1} 的乘积 $a_i \cdot k^{i-1}$, 再将其累加, 这是一个重复操作的步骤, 所以, 可以用循环结构来构造算法.

算法步骤如下:

第一步, 输入 a , k 和 n 的值.

第二步, 将 b 的值初始化为 0, i 的值初始化为 1.

第三步, $b = b + a_i \cdot k^{i-1}$, $i = i + 1$.

第四步, 判断 $i > n$ 是否成立. 若是, 则执行第五步; 否则, 返回第三步.

第五步, 输出 b 的值.

程序框图：

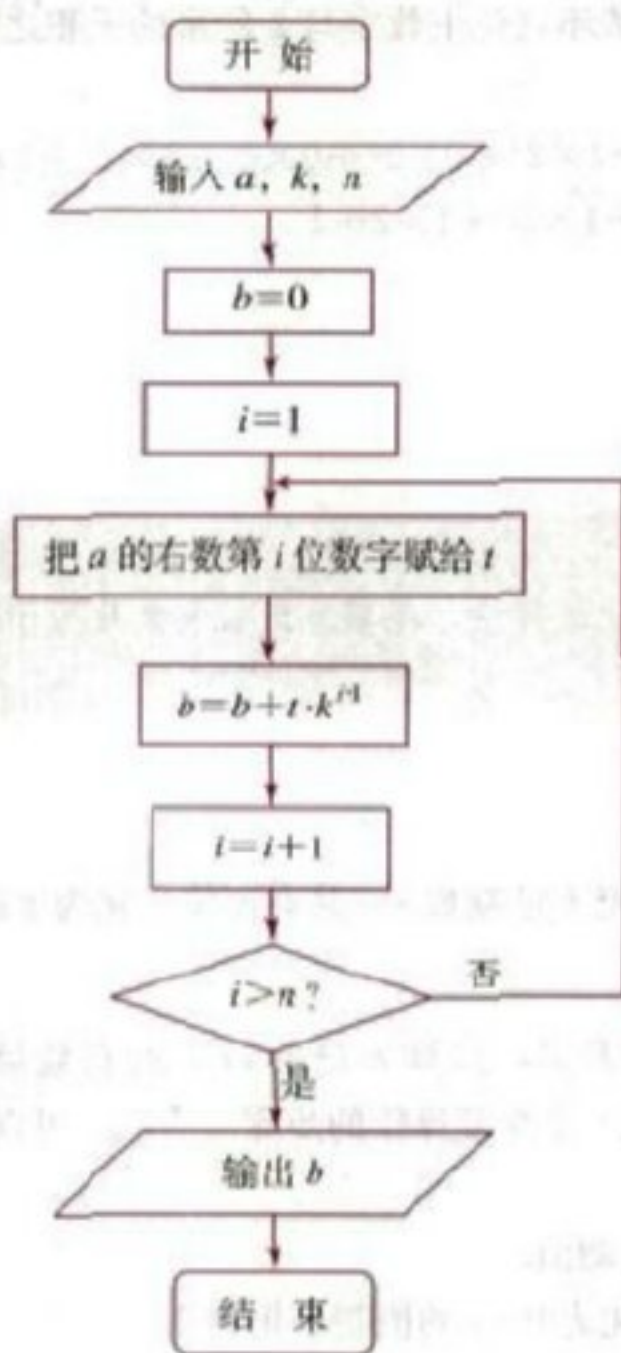


图 1.3-4

程序：

```

INPUT "a, k, n=" :a,k,n
b=0
i=1
t=a MOD 10
DO
  b=b+t*k^(i-1)
  a=a\10
  t=a MOD 10
  i=i+1
LOOP UNTIL i>n
PRINT b
END
  
```


例 5 把 89 化为二进制数.

解: 根据二进制数“满二进一”的原则, 可以用 2 连续去除 89 或所得商, 然后取余数. 具体计算方法如下:

$$\begin{aligned}
 \text{因为} \quad & 89 = 2 \times 44 + 1, \\
 & 44 = 2 \times 22 + 0, \\
 & 22 = 2 \times 11 + 0, \\
 & 11 = 2 \times 5 + 1, \\
 & 5 = 2 \times 2 + 1, \\
 & 2 = 2 \times 1 + 0, \\
 & 1 = 2 \times 0 + 1,
 \end{aligned}$$

所以

$$\begin{aligned}
 89 &= 2 \times (2 \times (2 \times (2 \times (2 \times 2 + 1) + 1) + 0) + 0) + 1 \\
 &= 2 \times (2 \times (2 \times (2 \times (2^2 + 1) + 1) + 0) + 0) + 1 \\
 &= \dots \\
 &= 1 \times 2^6 + 0 \times 2^5 + 1 \times 2^4 + 1 \times 2^3 + 0 \times 2^2 + 0 \times 2^1 + 1 \times 2^0 \\
 &= 1\ 011\ 001_{(2)}.
 \end{aligned}$$

这种算法叫做**除 2 取余法**, 还可以用下面的除法算式表示:

2	89	余数
2	44	1
2	22	0
2	11	0
2	5	1
2	2	1
2	1	0
	0	1

把上式中各步所得的余数从下到上排列, 得到 $89 = 1\ 011\ 001_{(2)}$.

上述方法也可以推广为把十进制数化为 k 进制数的算法, 称为**除 k 取余法**.

例 6 设计一个程序, 实现“除 k 取余法”.

算法分析:

从例 5 的计算过程可以看出如下的规律:

若十进制数 a 除以 k 所得商是 q_0 , 余数是 r_0 , 即 $a = k \cdot q_0 + r_0$, 则 r_0 是 a 的 k 进制数的右数第 1 位数;

若 q_0 除以 k 所得的商是 q_1 , 余数是 r_1 , 即 $q_0 = k \cdot q_1 + r_1$, 则 r_1 是 a 的 k 进制数的右数第 2 位数;

.....

若 q_{n-1} 除以 k 所得的商是 0, 余数是 r_n , 即 $q_{n-1} = r_n$, 则 r_n 是 a 的 k 进制数的左数第 1 位数.

这样，我们可以得到算法步骤如下：

第一步，给定十进制正整数 a 和转化后的数的基数 k 。

第二步，求出 a 除以 k 所得的商 q ，余数 r 。

第三步，把得到的余数依次从右到左排列。

第四步，若 $q \neq 0$ ，则 $a = q$ ，返回第二步；否则，输出全部余数 r 排列得到的 k 进制数。

程序框图：

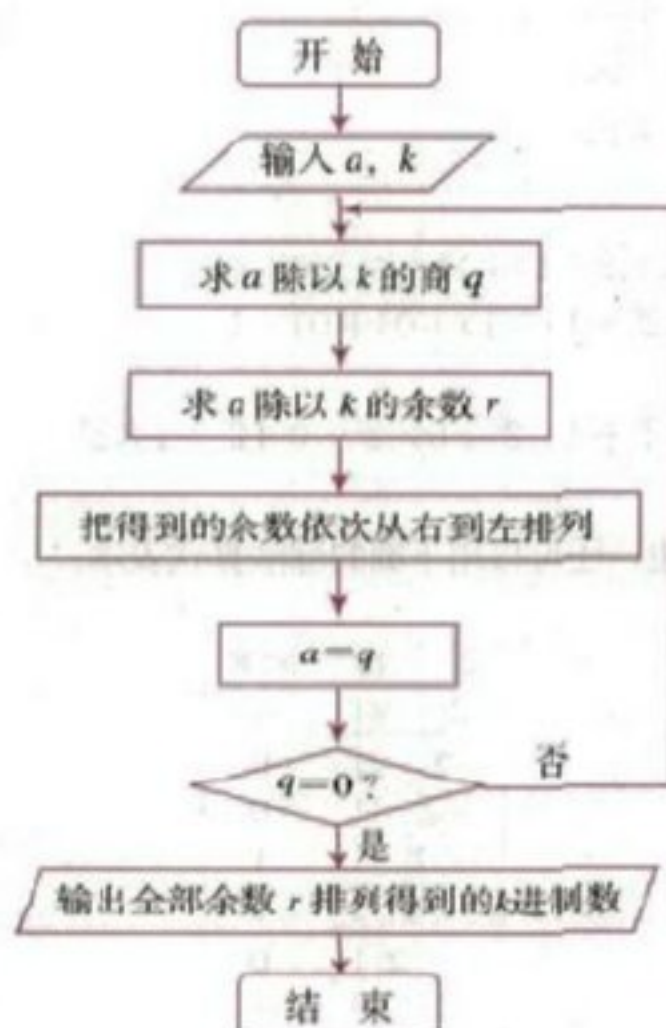


图 1.3-5

程序:

```

INPUT "a, k=" ; a, k
b=0
i=0
DO
    q=a\k
    r=a MOD k
    b=b+r*10^i
    i=i+1
    a=q
LOOP UNTIL q=0
PRINT b
END

```

练习

1. 用辗转相除法求下列两数的最大公约数:

(1) 225, 135;

(2) 98, 196;

(3) 72, 168;

(4) 153, 119.

2. 按照图 1.3-3 中的程序框图给出的步骤, 求

$$f(x) = 0.83x^5 + 0.41x^4 + 0.16x^3 + 0.33x^2 + 0.5x + 1$$

当 $x=5$ 时的值.

3. 用“除 k 取余法”将十进制数 2 008 转化为二进制数和八进制数.



割圆术

“割圆术”是求圆周率的一种算法, 圆周率在解决有关圆和球的计算问题中是非常重要的一个常数. 在古代, 各国数学家都把求出 π 的尽量准确的近似值作为一个重要课题. 历史上对于 π 的研究, 在一定程度上反映了一个时代或地区的数学和计算技术发展的水平.

我国最早采用的 π 值为 3, 即所谓“径一周三”(直径为 1, 周长为 3). 做法是将圆内接正六边形周长作为圆的周长, 从而求出圆周率.

263 年左右, 我国数学家刘徽发现当圆内接正多边形的边数无限增加时, 多边形面积可无限逼近圆面积, 即所谓“割之弥细, 所失弥少, 割之又割, 以至于不可割, 则与圆周

合体而无所失矣”。另一方面，这些圆内接正多边形每边外有一余径，用边长乘以余径，加到正多边形的面积上，则大于圆的面积，这样就可以得到圆面积的上限和下限，于是，刘徽采用了以直代曲、无限趋近、“内外夹逼”的思想，创立了“割圆术”。

“割圆术”的具体操作步骤是这样的：

第一步，从半径为1尺的圆内接正六边形开始，计算它的面积 S_6 。

第二步，逐步加倍圆内接正多边形的边数，分别计算圆内接正十二边形、正二十四边形、正四十八边形……的面积，到一定的边数（设为 $2m$ ）为止，得到一系列递增的数 $S_6, S_{12}, S_{24}, \dots, S_{2m}$ 。

第三步，在第二步中各正 n 边形每边外作一高为余径（如图1中 AB 所示）的矩形，把其面积 $2(S_{2n} - S_n)$ 与相应的正 n 边形的面积 S_n 相加，得 $S_n + 2(S_{2n} - S_n)$ ；这样又得到一系列递增数 $S_{12} + (S_{12} - S_6), S_{24} + (S_{24} - S_{12}), S_{48} + (S_{48} - S_{24}), \dots, S_{2m} + (S_{2m} - S_m)$ 。

第四步，由 $S_{2m} < S_{\pi} < S_{2m} + (S_{2m} - S_m)$ ，估计 S_{π} 的近似值，即圆周率的近似值。

“割圆术”从理论上能够把 π 的值计算到任意精度。刘徽一直计算到192边形，得到了圆周率精确到小数点后两位的近似值3.14，化成分数为 $\frac{157}{50}$ ，这就是著名的“徽率”。我国南北朝时期的数学家祖冲之继承并发展了刘徽的“割圆术”，求得 π 的范围为

$$3.141\ 592\ 6 < \pi < 3.141\ 592\ 7.$$

后人曾推算，若单纯使用“割圆术”，需要计算到圆内接正12 288边形，才能得到这样精确的结果。这不但是当时最精密的圆周率，而且在世界上处于领先地位达1 000多年。

现在，我们可以利用计算机来计算圆周率了。为此，我们先来分析一下圆内接正六边形、正十二边形、正二十四边形……的面积之间的关系，寻求它们的递增规律。

如图1，设圆的半径为1①，弦心距为 h_n ，正 n 边形的边长为 x_n ，面积为 S_n 。由勾股定理，得

$$h_n = \sqrt{1 - \left(\frac{x_n}{2}\right)^2}, \quad x_{2n} = \sqrt{\left(\frac{x_n}{2}\right)^2 + (1 - h_n)^2} \quad (n \geq 6).$$

容易知道 $x_6 = 1$ 。

观察图1，不难发现，正 $2n$ 边形的面积等于正 n 边形的面积加上 n 个等腰三角形的面积，即

$$S_{2n} = S_n + n \cdot \frac{1}{2} \cdot x_n(1 - h_n) \quad (n \geq 6).$$

利用这个递推公式，我们可以得到：



刘徽是我国魏晋时期杰出的数学家，著有《九章算术注》《海岛算经》等。

① 规定圆的半径为1会影响计算结果吗？

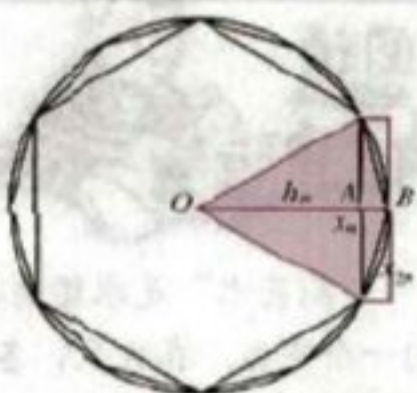


图1

正六边形的面积 $S_6 = 6 \times \frac{\sqrt{3}}{4}$;

正十二边形的面积 $S_{12} = \underline{\hspace{2cm}}$;

正二十四边形的面积 $S_{24} = \underline{\hspace{2cm}}$;

.....

由于圆的半径是1, 所以随着 n 的增大, S_{2n} 的值不断趋近于 π .

我们已经知道, 递推公式可以用循环结构来表达. 因此, 上述步骤可以写成如下的程序.

```
INPUT "n="; n
i=6
x=1
s=6*SQR(3)/4
WHILE i<=n/2
    h=SQR(1-(x/2)^2)
    s=s+i*x*(1-h)/2
    x=SQR((x/2)^2+(1-h)^2)
    i=2*i
WEND
PRINT n, s
END
```

在这个程序中, n 的输入值满足什么条件?

你能进一步完善这个程序, 把“割圆术”编成计算机程序吗?

随着计算机计算速度的高速发展, 到1973年, 人们已把圆周率算到了小数点后100万位, 1989年突破了10亿位大关, 1999年超过了2 061亿位.

现在, 数学家们所关心的问题已不是如何打破纪录, 算出更高精度的 π 值, 而是如何在算法上取得突破, 让计算机更加有效地计算 π 值.

目前, 在几何、微积分和概率领域, 都有求圆周率的近似值的算法. 有兴趣的同学可以查找相关资料, 或在互联网上搜索相关算法.



习题 1.3

A 组

1. 用辗转相除法求下列两数的最大公约数, 并用更相减损术检验你的结果:

(1) 228, 1 995; (2) 5 280, 12 155.

2. 用秦九韶算法求多项式

$$f(x) = 7x^7 + 6x^6 + 5x^5 + 4x^4 + 3x^3 + 2x^2 + x$$

当 $x=3$ 时的值.

3. 完成下列进位制之间的转化:

(1) $10\ 212_{(3)} = \underline{\hspace{2cm}}_{(10)}$; (2) $412_{(3)} = \underline{\hspace{2cm}}_{(7)}$;

(3) $2\ 376_{(8)} = \underline{\hspace{2cm}}_{(10)}$; (4) $119_{(10)} = \underline{\hspace{2cm}}_{(6)}$.

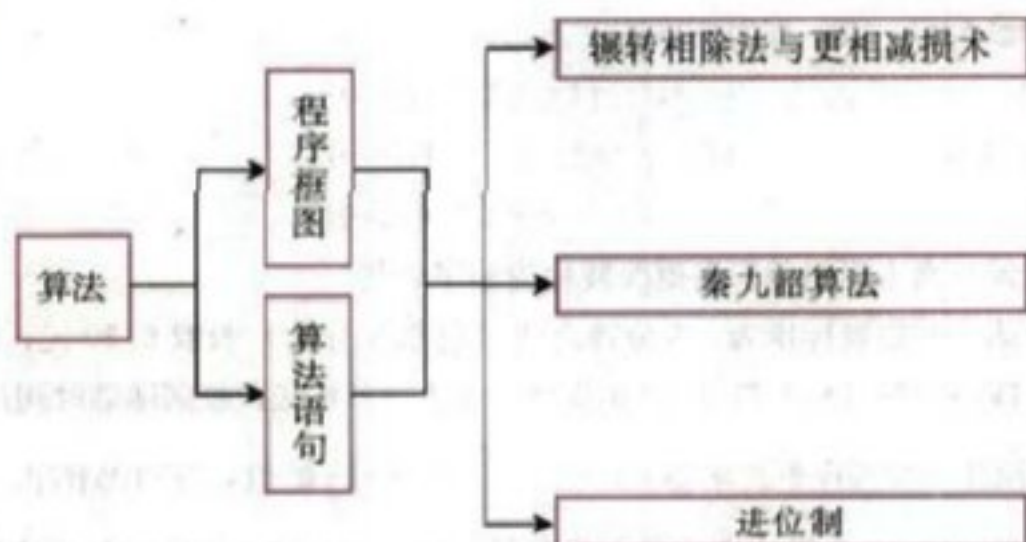
4. 根据阅读与思考“割圆术”中的程序画出程序框图.

B 组

1. 某班有 45 名学生, 设计一个算法, 输入每个学生的数学成绩后, 分别统计在区间 $[0, 60)$ $[60, 80)$ $[80, 100]$ 内的成绩的个数, 用自然语言描述算法步骤, 可用 $a(i)$ 表示第 i 个学生的成绩.
2. 更相减损术、秦九韶算法和割圆术都是中国古代数学中的优秀算法, 查找资料, 搜集其他中国古代数学中的算法.

小结

一、本章知识结构



二、回顾与思考

1. 算法，是我们既熟悉又陌生的，而且非常有用，在计算机科学和数学领域中都占据着重要地位。算法的基本思想在我们的日常生活中是很有用的，学习算法对于发展我们有条理的思考与表达能力，提高我们的逻辑思维能力也是很有帮助的。通过本章的学习，请你说一说算法的涵义是什么，它有什么特点，举出几个蕴涵某种算法的问题，并谈一谈学习算法的体会。

2. 算法的三种基本逻辑结构是什么？你能用相应的程序框图表达吗？

3. 我们可以用自然语言叙述算法，也可以用程序框图表示算法，还可以通过算法语句编写程序使计算机执行算法。自然语言描述的算法步骤、程序框图和程序是不同形式的算法，体现了算法逐渐“精确”的过程。请你说一说完成一个算法的基本步骤。

4. 本章介绍了3个典型的算法案例，想一想它们各蕴涵了怎样的算法，各举一个应用这3个算法的例子。

复习参考题

A 组

1. 画程序框图，对于输入的 x 值，输出相应的 y 值：

$$(1) y = \begin{cases} 0 & (x < 0), \\ 1 & (0 \leq x < 1), \\ x & (x \geq 1); \end{cases}$$

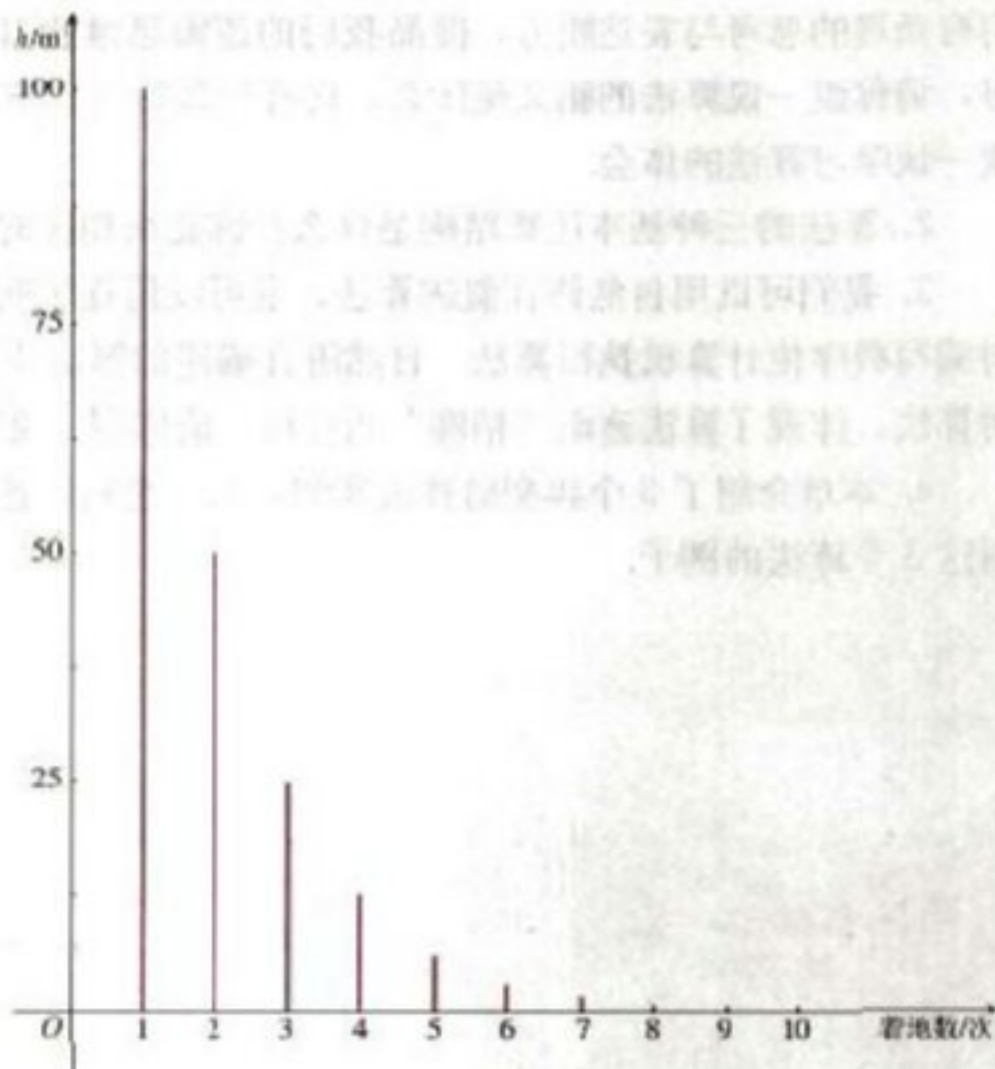
$$(2) y = \begin{cases} (x+2)^2 & (x < 0), \\ 4 & (x = 0), \\ (x-2)^2 & (x > 0). \end{cases}$$

2. 把你画出的求解二元一次方程组的程序框图转化为程序语句.
3. 某市固定电话（市话）的收费标准为：3 分钟之内（包括 3 分钟）收取 0.20 元；超过 3 分钟，每分钟（不足一分钟，按一分钟计算）按 0.10 元收费. 设计一个算法，根据通话时间计算话费.
4. 对任意正整数 n ，设计一个程序框图求 $S = 1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n}$ 的值，并写出程序.
5. 设计两种算法，从输入的 10 个数中选出最大值和最小值，用自然语言描述算法步骤.
6. 一个球从 100 m 高处自由落下，每次着地后又跳回到原高度的一半再落下. 编写程序，求当它第 10 次着地时，

(1) 向下的运动共经过多少米？

(2) 第 10 次着地后反弹多高？

(3) 全程共经过多少米？



(第 6 题)

B 组

1. 编写程序, 将用户输入的正整数转换成相应的星期值输出, 如用户输入 3, 则输出 Wednesday; 用户输入 0, 则输出 Sunday. 如果用户输入的数大于 6, 则用这个数除以 7 所得的余数进行上述操作.
2. 画出程序框图, 用二分法求方程 $1.3x^3 - 26.013x^2 + 0.975x - 19.50975 = 0$ 在 $(20, 21)$ 之间的近似根 (精确度为 0.005).
3. 设计一个算法, 判断一个正的 $n(n \geq 2)$ 位数是不是回文数^❶, 用自然语言描述算法步骤.

❶ 回文数是指从左到右读与从右到左读都是一样的正整数, 如 121, 676, 94 249 等.

第二章

统计

我国是世界上的第13个
贫水国,人均淡水占有量排
列世界第109位。

2.1 随机抽样

2.2 用样本估计总体

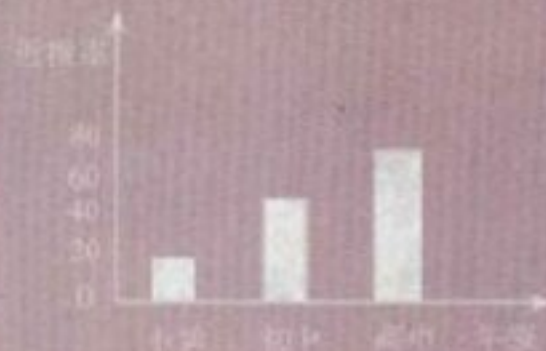
2.3 变量间的相关关系

我们生活在一个数字化时代,时刻都在与数据打交道,例如,产品的合格率、农作物的产量、商品的销售量、当地的气温、自然资源、就业状况、电视台的收视率,你知道这些数据是怎么来的吗?实际上它们是通过调查获得的。怎样调查呢?是对考察对象进行全面调查吗?例如,为了了解一批计算器的使用寿命,我们能将它们逐一测试吗?很明显,这既不可能,也没必要。实践中,由于所考察的总体中的个体数往往很多,而且许多考察带有破坏性,因此,我们通常只考察总体中的一个样本,通过样本来了解总体的情况。进一步,从节约费用的角度考虑,在保证样本估计总体达到一定精度的前提下,样本中包含的个体数越少越好。于是,如何设计抽样方法,使抽取的样本能够真正代表总体,就成为我们要关注的一个关键问题。否则,如果样本的代表性不好,那么对总体的判断就会出现错误。

那么,怎样从总体中抽取样本呢?如何表示样本数据呢?如何从样本数据中提取基本信息(样本分布、样本数字特征等),来推断总体的情况呢?这些正是本章要研究解决的问题。

CHAPTER 2

2.1



学校	近视率	近视人数	近视率
小学	357 000	221 600	258 100
初中	225 200	134 200	112 900
高中	112 000	43 300	6 300

随机抽样

为了回答我们碰到的许多问题，必须收集相关数据。例如食品、饮料中的细菌是否超标，每天城市里的垃圾有多少被回收了，影响学生视力状况的主要原因有哪些，同学们的作息时间是如何安排的，电视台的某个栏目的收视率是多少，某厂产品的合格率是多少……这些问题都需要通过收集数据作出回答。

从节约费用等方面考虑，一般是从总体中收集部分个体的数据来得出结论，也就是要通过样本去推断总体。为此，我们首先必须清楚地知道要收集的数据是什么。例如，在食品质量检验中，为了了解某批袋装牛奶（总体）的细菌超标情况，从中随机地抽取了 n 袋，并测出了每一袋的细菌含量 a_i ($i=1, 2, \dots, n$)，这里， a_i ($i=1, 2, \dots, n$) 就是我们要收集的数据。其次，我们检查样本的目的是为了了解总体的情况。在上述牛奶质量检查中，我们的目的是要了解整批牛奶的细菌含量是否超标，而不是局限在抽查到的那些牛奶的细菌含量是否超标。因此，收集的样本数据应当能够很好地反映总体，这是从样本推断出关于总体的正确结论的前提。再次，我们要知道如何才能收集到高质量的样本数据。我们知道，为了判断一锅汤的味道如何，如果锅里的汤被充分搅拌了，那么我们只需品尝一勺就可以了。同样地，高质量的样本数据来自“搅拌均匀”的总体。如果我们能够设法将总体“搅拌均匀”，那么，从中任意抽取一部分个体的样本，它们含有与总体基本相同的信息。

总之，为了使样本具有好的代表性，设计抽样方法时，最重要的是要将总体“搅拌均匀”，即使每个个体有同样的机会被抽中，下面介绍的抽样方法都是以此作为出发点的。



一个著名的案例

在抽样调查中,样本的选择是至关重要的,样本能否代表总体,直接影响着统计结果的可靠性。下面的故事是一次著名的失败的统计调查,被称作抽样中的泰坦尼克事件,它可以帮助我们理解为什么一个好的样本如此重要。

在1936年美国总统选举前,一份颇有名气的杂志(Literary Digest)的工作人员做了一次民意测验。调查兰顿(A. Landon)(当时任堪萨斯州州长)和罗斯福(F. D. Roosevelt)(当时的总统)中谁将当选下一届总统。为了了解公众意向,调查者通过电话簿和车辆登记簿上的名单给一大批人发了调查表(注意在1936年电话和汽车只有少数富人拥有)。通过分析收回的调查表,显示兰顿非常受欢迎,于是此杂志预测兰顿将在选举中获胜。

实际选举结果正好相反,最后罗斯福在选举中获胜,其数据如下:

候选人	预测结果%	选举结果%
Roosevelt	43	62
Landon	57	38

像本例中这样容易得到的样本称为方便样本。



你认为预测结果出错的原因是什么?

2.1.1

简单随机抽样



假设你作为一名食品卫生工作人员,要对某食品店内的一批小包装饼干进行卫生达标检验,你准备怎样做?

显然,你只能从中抽取一定数量的饼干作为检验的样本。(为什么?)那么,应当怎样获取样本呢?

设计抽样方法时,在考虑样本的代表性的前提下,应当努力使抽样过程简便易行.

得到样本饼干的一个方法是,将这批小包装饼干放入一个不透明的袋子中,搅拌均匀,然后不放回地摸取(这样可以保证每一袋饼干被抽中的机会相等),这样我们就可以得到一个简单随机样本,相应的抽样方法就是简单随机抽样.

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 ($n \leq N$),如果每次抽取时总体内的各个个体被抽到的机会都相等,就把这种抽样方法叫做**简单随机抽样**(simple random sampling).

最常用的简单随机抽样方法有两种——抽签法和随机数法.

(1) 抽签法(抓阄法)

抽签法是大家最熟悉的,也许同学们在做某种游戏,或者选派一部分人参加某项活动时就用过抽签法.例如,高一(2)班有45名学生,现要从中抽出8名学生去参加一个座谈会,每名学生的机会均等.我们可以把45名学生的学号写在小纸片上,揉成小球,放到一个不透明袋子中,充分搅拌后,再从中逐个抽出8个号签,从而抽出8名参加座谈会的学生.

一般地,抽签法就是把总体中的 N 个个体编号,把号码写在号签上,将号签放在一个容器中,搅拌均匀后,每次从中抽取一个号签,连续抽取 n 次,就得到一个容量为 n 的样本.



你认为抽签法有什么优点和缺点?当总体中的个体数很多时,用抽签法方便吗?

抽签法简单易行,当总体中的个体数不多时,使总体处于“搅拌均匀”的状态比较容易,这时,每个个体有均等的机会被抽中,从而能够保证样本的代表性.但是,当总体中的个体数较多时,将总体“搅拌均匀”就比较困难,用抽签法产生的样本代表性差的可能性很大.

(2) 随机数法

随机抽样中,另一个经常被采用的方法是随机数法,即利用随机数表、随机数骰子或计算机产生的随机数进行抽样.这里仅介绍随机数表法.

随机数表由数字0,1,2,...,9组成,并且每个数字在表中各个位置出现的机会都是一样的(见本章附表).

怎样利用随机数表产生样本呢?下面通过例子来说明.

假设我们要考察某公司生产的500克袋装牛奶的质量是否达标,现从800袋牛奶中抽取60袋进行检验.利用随机数表抽取样本时,可以按照下面的步骤进行.

生产实践中,往往是从一大批袋装牛奶中抽样,也就是说总体中的个体数是很大的.你能从这个例子出发说明一下抽样的必要性吗?

第一步, 先将 800 袋牛奶编号, 可以编为 000, 001, ..., 799.

第二步, 在随机数表中任选一个数, 例如选出第 8 行第 7 列的数 7 (为了便于说明, 下面摘取了附表 1 的第 6 行至第 10 行).

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

第三步, 从选定的数 7 开始向右读(读数的方向也可以是向左、向上、向下等), 得到一个三位数 785, 由于 $785 < 799$, 说明号码 785 在总体内, 将它取出; 继续向右读, 得到 916, 由于 $916 > 799$, 将它去掉, 按照这种方法继续向右读, 又取出 567, 199, 507, ..., 依次下去, 直到样本的 60 个号码全部取出, 这样我们就得到一个容量为 60 的样本.

当 $N=100$ 时, 分别以 0, 3, 6 为起点对总体编号, 再利用随机数表抽取 10 个号码. 你能说出从 0 开始对总体编号的好处吗?

练习

1. 请你把抽样调查和普查做一个比较, 并说一说抽样调查的好处和可能出现的问题.
2. 假设要从高一年级全体同学(450人)中随机抽出 50 人参加一项活动, 请分别用抽签法和随机数表法抽出人选, 写出抽取过程.
3. 请举出几个用抽签法或随机数表法抽取样本的实际例子, 你认为抽签法是如何保证样本的代表性的?
4. 你认为用随机数表法抽取样本有什么优点和缺点?

从上所述可知, 简单随机抽样有操作简便易行的优点, 在总体个数不多的情况下是行之有效的. 但是, 如果总体中的个体数很多时, 对个体编号的工作量太大, 即使用随机数法操作也并不方便快捷. 另外, 要想“搅拌均匀”也非常困难, 这就容易导致样本的代表性差. 因此, 为了操作上方便快捷, 在不降低样本的代表性的前提下, 可以采取下面的抽样方法.

2.1.2 系统抽样



某学校为了了解高一年级学生对教师教学的意见,打算从高一年级 500 名学生中抽取 50 名进行调查.除了用简单随机抽样获取样本外,你能否设计其他抽取样本的方法?

我们按照这样的方法来抽样:首先将这 500 名学生从 1 开始进行编号,然后按号码顺序以一定的间隔进行抽取.由于 $\frac{500}{50} = 10$, 这个间隔可以定为 10, 即从号码为 1~10 的第一个间隔中随机地抽取一个号码,假如抽到的是 6 号,然后从第 6 号开始,每隔 10 个号码抽取一个,得到

6, 16, 26, 36, ..., 496.

这样我们就得到一个容量为 50 的样本.这种抽样方法是一种**系统抽样**(systematic sampling).

一般地,假设要从容量为 N 的总体中抽取容量为 n 的样本,我们可以按下列步骤进行系统抽样:

(1) 先将总体的 N 个个体编号.有时可直接利用个体自身所带的号码,如学号、准考证号、门牌号等;

(2) 确定分段间隔 k , 对编号进行分段.当 $\frac{N}{n}$ (n 是样本容量)是整数时,取 $k = \frac{N}{n}$;

(3) 在第 1 段用简单随机抽样确定第一个个体编号 l ($l \leq k$);

(4) 按照一定的规则抽取样本.通常是将 l 加上间隔 k 得到第 2 个个体编号 $(l+k)$, 再加 k 得到第 3 个个体编号 $(l+2k)$, 依次进行下去,直到获取整个样本.

请将这种抽样方法与简单随机抽样做一个比较,你认为这种抽样方法能提高样本的代表性吗?为什么?

如果遇到 $\frac{N}{n}$ 不是整数的情况,可以先从总体中随机地剔除几个个体,使得总体中剩余的个体数能被样本容量整除.

练习

1. 你认为系统抽样有哪些优点和缺点?
2. 设某校共有 118 名教师, 为了支援西部的教育事业, 现要从中随机地抽出 16 名教师组成暑期西部讲师团, 请用系统抽样法选出讲师团成员.
3. 有人说, 我可以借用居民身份证号码 (18 位) 来进行中央电视台春节联欢晚会的收视率调查: 在 1~999 中抽取一个随机数, 比如这个数是 632, 那么身份证后三位数是 632 的观众就是我要调查的对象. 请问, 这样所获得的样本有代表性吗? 为什么?



广告中数据的可靠性

今天已进入数字时代, 各种各样的统计数字和图表充斥着媒体, 由于数字给人的印象直观具体, 所以让数据说话是许多广告的常用手法. 但广告中的数据可靠吗?

在各类广告中, 你会经常遇到由“方便样本 (即样本没有代表性)”所产生的结论. 例如, 某减肥药的广告称, 其减肥的有效率为 75%. 见到这样的广告你会怎么想? 通过学习统计这部分内容, 你会提出下面的问题吗? 这个数据是如何得到的; 该药在多少人身上做过试验, 即样本容量是多少; 样本是如何选取的; 等等. 假设该药仅在 4 个人身上做过试验, 样本容量为 4, 用这样小的样本量来推断总体是不可信的.

“现代研究证明, 99% 以上的人感染有螨虫……”这是一家化妆品公司的广告. 第一次听到此话的人会下意识地摸一下自己的皮肤, 甚至会感觉到有虫在里面蠕动, 恨不得立即弄些药膏抹抹, 广告的威慑作用不言而喻. 但这里 99% 是怎么得到的? 研究共检测了多少人? 这些人是如何挑选的? 如果检测的人都是去医院看皮肤病的人, 这个数据就不适用于一般人群.

某化妆品的广告声称: “它含有某种成分可以彻底地清除脸部皱纹, 只需 10 天, 就能让肌肤得到改善.” 我们看到的数字很精确, 而“能让肌肤得到改善”却是很模糊的. 这样的数字能相信吗? 试验是在什么样的皮肤上做的? 试验的人数是多少?

当我们见到广告中的数据时一定要多提几个问题.



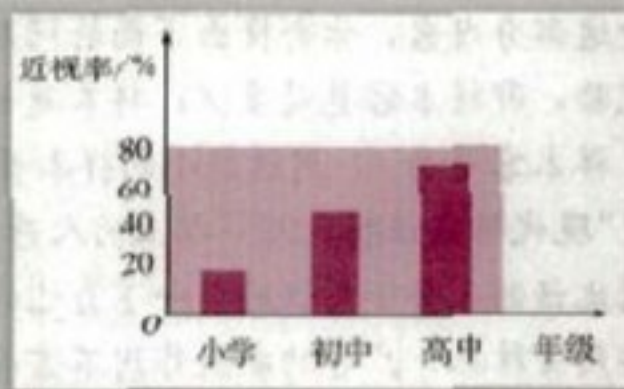
请你从各种媒体中收集一些广告,并用统计的知识分析一下它们所提供的数据和结论的真实性.

我们知道,设计抽样方法时,最核心的问题是要考虑如何使抽取的样本具有好的代表性.为此,在设计抽样方法时,我们应考虑如何利用自己对总体的已有了解.例如,如果要调查某校高一学生的平均身高,由经验可知,男生一般要比女生高,这时就应采用另一种抽样方法——分层抽样.因为用简单抽样方法或系统抽样的方法都有可能产生绝大部分是男生(或女生)或全部都是男生(或女生)的样本,显然,这种样本是不能代表总体的.因此,设计抽样方法时,充分利用事先对总体情况的已有了解是非常重要的.

2.1.3 分层抽样



假设某地区有高中生2 400人,初中生10 900人,小学生11 000人.此地区教育部门为了了解本地区中小学生的近视情况及其形成原因,要从本地区的中小学生中抽取1%的学生进行调查.你认为应当怎样抽取样本?



我们知道,影响学生视力的因素是非常复杂的.例如,不同年龄阶段的学生们的近视情况可能存在明显差异.因此,宜将全体学生分成高中、初中和小学三部分分别抽样.另外,三个部分的学生人数相差较大,因此,为了提高样本的代表性,还应考虑他们在样本中所占比例的大小.

由于样本容量与总体中的个体数的比是1:100,因此,样本中包含的各部分的个体数应该是

$$\frac{2\,400}{100}, \frac{10\,900}{100}, \frac{11\,000}{100}.$$



你认为哪些因素可能影响学生的视力?设计抽样方法时需要考虑这些因素吗?

即抽取 24 名高中生, 109 名初中生和 110 名小学生作为样本.

这样, 如果从学生人数这个角度来看, 按照这种抽样方法所获得的样本结构与这一地区全体中小学生的结构是基本相同的.

一般地, 在抽样时, 将总体分成互不交叉的层, 然后按照一定的比例, 从各层独立地抽取一定数量的个体, 将各层取出的个体合在一起作为样本, 这种抽样方法是一种**分层抽样** (stratified sampling).

从上面的抽样过程可以看出, 分层抽样尽量利用了调查者对调查对象 (总体) 事先所掌握的各种信息, 并充分考虑了保持样本结构与总体结构的一致性, 这对提高样本的代表性是非常重要的. 所以, 分层抽样在实际中有着非常广泛的应用. 通常, 当总体是由差异明显的几个部分组成时, 往往选用分层抽样的方法.

想一想, 为什么要这样取各个学段的个体数?



(1) 简单随机抽样、系统抽样和分层抽样各有其特点和适用范围. 请对这三种抽样方法进行比较, 说说它们各自的优点和缺点.

(2) 某地区中小學生人数的分布情况如下表所示 (单位: 人):

学段	城市	县镇	农村
小学	357 000	221 600	258 100
初中	226 200	134 200	11 290
高中	112 000	43 300	5 300

请根据上述基本数据, 设计一个样本容量为总体中个体数量的千分之一的抽样方案.

在现实生活中, 由于资金、时间有限, 人力、物力不足, 再加上不断变化的环境条件, 做普查往往是不可能的. 因此, 我们一般是把数据的收集限制在总体的一个样本上. 由于总体的复杂性, 在实际抽样中, 为了使样本具有代表性, 通常要同时使用几种抽样方法. 例如, 在上述探究 (2) 中, 我们可以先用分层抽样法确定出此地区城市、县镇、农村的被抽个体数, 再用分层抽样法将城市的被抽个体数分配到小学、初中、高中等不同阶层中去, 县镇、农村的被抽个体数的分配法也一样. 接着, 将城市划分为学生数大致相当的小区, 用简单随机抽样法选取一些小区, 再用简单随机抽样法确定每一小区中的各类学校. 最后, 在选中的学校中用系统抽样法或简单随机抽样法选取学生进行调查.

练习

1. 分别用简单随机抽样、系统抽样和分层抽样的方法, 从全班同学中抽取 10 名同学, 统计他们昨天户外活动的平均时间. 全面调查全班同学昨天户外活动的平均时间, 并与抽样统计的结果进行比较, 你能发现什么问题?
2. 有人说: “如果抽样方法设计得好, 用样本进行视力调查与对 24 300 名学生进行视力普查的结果会差不多, 而且对于教育部门掌握学生视力状况来说, 因为节省了人力、物力和财力, 抽样调查更可取.” 你认为这种说法有道理吗? 为什么?
3. 一般来说, 影响农作物收成的因素有气候、土质、田间管理水平等. 如果你是一个农村调查队成员, 要在麦收季节对你所在地区的小麦进行估产调查, 你将如何设计调查方案?



如何得到敏感性问题的诚实反应

在统计调查中, 问卷的设计是一门很大的学问. 特别是对一些敏感性问题, 例如学生在考试中有无作弊现象, 社会上的偷税漏税等, 更要精心设计问卷, 设法消除被调查者的顾虑, 使他们能够如实回答问题. 否则, 被调查者往往会拒绝回答, 或不提供真实情况. 下面我们用一个例子来说明对敏感性问题的调查方法.

某地区公共卫生部门为了调查本地区中学生的吸烟情况, 对随机抽出的 200 名学生进行了调查. 调查中使用了两个问题.

问题 1: 你的父亲阳历生日日期是不是奇数?

问题 2: 你是否经常吸烟?

调查者设计了一个随机化装置, 这是一个装有大小、形状和质量完全一样的 50 个白球和 50 个红球的袋子. 每个被调查者随机从袋中摸取 1 个球 (摸出的球再放回袋中), 摸到白球的学生如实回答第一个问题, 摸到红球的学生如实回答第二个问题. 回答“是”的人往一个盒子中放一个小石子, 回答“否”的人什么都不要做. 由于问题的答案只有“是”和“否”, 而且回答的是哪个问题也是别人不知道的, 因此被调查者可以毫无顾虑地给出符合实际情况的答案.

请问: 如果在 200 人中, 共有 58 人回答“是”, 你能估计出本地区中学生吸烟人数的百分比吗?

解: 由题意可知, 每个学生从口袋中摸出 1 个白球或红球的概率都是 0.5, 即我们期望大约有 100 人回答了第一个问题, 另 100 人回答了第二个问题. 在摸出白球的情况下, 回答父亲阳历生日日期是奇数的概率是 $\frac{186}{365} \approx 0.51$. 因而在回答第一个问题的 100 人中, 大

约有 51 人回答了“是”，所以我们能推出，在回答第二个问题的 100 人中，大约有 7 人回答了“是”，即估计此地区大约有 7% 的中学生吸烟。

这种方法是不是很巧妙？



在问卷的设计中，不但要考虑“难以启齿”问题本身对调查结果的影响，而且还要考虑其他因素。例如，调查中问题的措辞会对被调查者产生影响，举例来说，“你在多大程度上喜欢吸烟”与“你在多大程度上不喜欢吸烟”两种问法中，前者会比后者给出更为肯定的答案。再如，问题在问卷中的位置也会对调查者产生影响。一般地，比较容易的、不涉及个人的问题应当排在比较靠前的位置，较难的、涉及个人的问题放在后面，等等。

请你设计一个关于青春期问题的调查问卷。

习题 2.1

A 组

1. 在抽样过程中，如果总体中的每个个体都有相等的机会被抽中，那么我们就称这样产生的样本为随机样本。举例说明产生随机样本的困难。
2. 中央电视台希望在春节联欢晚会播出后一周内获得当年春节联欢晚会的收视率。下面是三名同学为电视台设计的调查方案。

同学 A：我把这张《春节联欢晚会收视率调查表》放在互联网上，只要上网登录该网址的人就可以看到这张表，他们填表的信息可以很快地反馈到我的电脑中。这样，我就可以很快统计出收视率了。

同学 B：我给我们居民小区的每一份住户发一个是否在除夕那天晚上看过中央电视台春节联欢的调查表，只要一两天就可以统计出收视率。

同学 C：我在电话号码本上随机地选出一定数量的电话号码，然后逐个给他们打电话，问一下他们是否收看了中央电视台春节联欢晚会，我不出家门就可以统计出中央电视台春节联欢晚会的收视率。

请问：上述三名同学设计的调查方案能够获得比较准确的收视率吗？为什么？

3. 校学生会希望调查有关本学学生活动计划的意见, 你自愿担任调查员, 并打算在学校里抽取 10% 的同学作为样本.
 - (1) 你怎样安排抽样, 以保证样本的代表性?
 - (2) 在抽样中你可能遇到哪些问题?
 - (3) 这些问题可能会影响什么?
 - (4) 你打算怎样解决这些问题?
4. 请用简单随机抽样和系统抽样, 设计一个调查某地区一年内空气质量状况的方案, 哪一个方案更便于实施.
5. 一支田径队有男运动员 56 人, 女运动员 42 人, 用分层抽样的方法从全体运动员中抽出一个容量为 28 的样本.
6. 在一次游戏中, 获奖者可以得到 5 件不同的奖品, 这些奖品要从由 1~50 编号的 50 种不同奖品中随机抽取确定, 用系统抽样的方法为某位获奖者确定 5 件奖品的编号.
7. 设计一个抽样方案, 调查你们学校学生的近视率.

B 组

1. 你可能想了解许多问题, 比如, 全班同学比较喜欢哪门课程, 中学生每月的零花钱平均是多少, 喜欢看《新闻联播》的同学的比例是多少, 中学生每天大约什么时间起床, 每天睡眠的平均时间是多少等. 选一些自己关心的问题, 设计一份调查问卷, 利用抽样的方法调查你们学校的学生的情况, 并解释你所得到的结论.
2. 设计一个抽样方案, 调查中央电视台春节联欢晚会的收视率.

CHAPTER 2.2

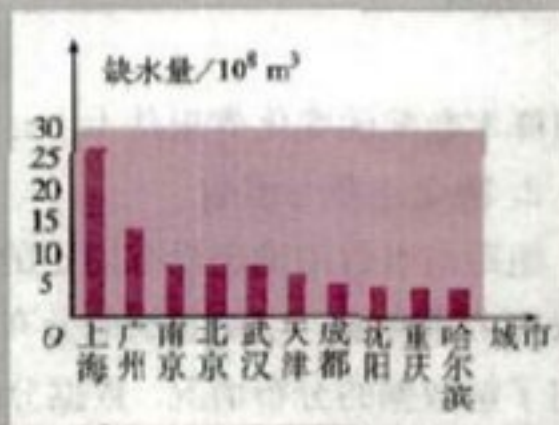
用样本估计总体

前面我们研究了通过抽样来收集数据的方法，了解了提高样本代表性的一些具体方法。数据被收集后，必须从中寻找所包含的信息，以使我们能通过样本估计总体。由于数据多而且杂乱，我们往往无法直接从原始数据中理解它们的含义。因此，必须通过图、表、计算来分析数据，帮助我们找出数据中的规律，使数据所包含的信息转化成直观的容易理解的形式。在此基础上，我们就可以对总体作出相应的估计。这种估计一般分成两种，一种是用样本的频率分布估计总体的分布，另一种是用样本的数字特征（如平均数、标准差等）估计总体的数字特征。

2.2.1 用样本的频率分布估计总体分布

探究

我国是世界上严重缺水的国家之一，城市缺水问题较为突出。某市政府为了节约生活用水，计划在本市试行居民生活用水定额管理，即确定一个居民月用水量标准 a ，用水量不超过 a 的部分按平价收费，超出 a 的部分按议价收费。如果希望大部分居民的日常生活不受影响，那么标准 a 定为多少比较合理呢？你认为，为了较为合理地确定出这个标准，需要做哪些工作？



2000年全国主要城市中
缺水情况排在前10位的城市

很明显，如果标准太高，会影响居民的日常生活；如果标准太低，则不利于节水。为了确定一个较为合理的标准 a ，必须先了解全市居民日常用水量的分布情况，比如月均用水量在哪个范围的居民最多，他们占全市居民的百分比情况等。

由于城市住户较多，通常采用抽样调查的方式，通过分析样本数据来估计全市居民用水量的分布情况。假设通过抽样，我们获得了100位居民某年的月均用水量（单位： t ）：

表 2-1 100 位居民的月均用水量 (单位: t)

3.1	2.5	2.0	2.0	1.5	1.0	1.6	1.8	1.9	1.6
3.4	2.6	2.2	2.2	1.5	1.2	0.2	0.4	0.3	0.4
3.2	2.7	2.3	2.1	1.6	1.2	3.7	1.5	0.5	3.8
3.3	2.8	2.3	2.2	1.7	1.3	3.6	1.7	0.6	4.1
3.2	2.0	2.4	2.3	1.8	1.4	3.5	1.0	0.8	4.3
3.0	2.9	2.4	2.4	1.9	1.3	1.4	1.8	0.7	2.0
2.5	2.8	2.3	2.3	1.8	1.3	1.3	1.6	0.9	2.3
2.6	2.7	2.4	2.1	1.7	1.4	1.2	1.5	0.5	2.4
2.5	2.6	2.3	2.1	1.6	1.0	1.0	1.7	0.8	2.4
2.8	2.5	2.2	2.0	1.5	1.0	1.2	1.8	0.6	2.2

实际抽样时, 样本容量大小应当根据问题的需要来确定, 并不一定样本容量越大越好.

上面这些数字能告诉我们什么呢? 很容易发现的是一个居民月均用水量的最小值是 0.2 t, 最大值是 4.3 t, 其他在 0.2~4.3 t 之间. 除此之外, 很难发现这 100 位居民的用水量的其他信息了. 实际上, 我们很难从随意记录下来的数据中直接看出规律, 为此, 我们需要对统计数据整理与分析.

分析数据的一种基本方法是用图将它们画出来, 或者用紧凑的表格改变数据的排列方式. 作图可以达到两个目的, 一是从数据中提取信息, 二是利用图形传递信息. 表格则是通过改变数据的构成形式, 为我们提供解释数据的新方式.

一幅图胜过一千个字.
看懂图是 21 世纪的成年人
所必须具备的能力.

初中我们曾经学过频数分布图和频数分布表, 这使我们能够清楚地知道数据分布在各个小组的个数. 下面将要学习的频率分布表和频率分布图, 则是从各个小组数据在样本容量中所占比例大小的角度, 来表示数据分布的规律. 它可以使我们看到整个样本数据的**频率分布** (frequency distribution) 情况. 具体的做法如下.

1. 求极差 (即一组数据中最大值与最小值的差)

例如,

$$4.3 - 0.2 = 4.1,$$

说明样本数据的变化范围是 4.1 t.

2. 决定组距与组数

组距与组数的确定没有固定的标准, 常常需要一个尝试和选择的过程. 将数据分组时, 组数应力求合适, 以使数据的分布规律能较清楚地呈现出来. 组数太多或太少, 都会影响我们了解数据的分布情况. 数据分组的组数与样本容量有关, 一般样本容量越大, 所分组数越多. 当样本容量不超过 100 时, 按照数据的多少, 常分成 5~12 组.

为方便起见, 组距的选择应力求“取整”. 在本问题中, 如果取组距为 0.5 (t), 那么

$$\text{组数} = \frac{\text{极差}}{\text{组距}} = \frac{4.1}{0.5} = 8.2,$$

因此可以将数据分为 9 组. 这个组数是较合适的, 于是取组距为 0.5, 组数为 9.

3. 将数据分组

以组距为 0.5 将数据分组时, 可以分成以下 9 组:

$[0, 0.5)$, $[0.5, 1)$, \dots , $[4, 4.5]$.

4. 列频率分布表

计算各小组的频率, 作出下面的**频率分布表**。

表 2-2 100 位居民月均用水量的频率分布表

分组	频数累计	频数	频率
[0, 0.5)	正	4	0.04
[0.5, 1)	正下	8	0.08
[1, 1.5)	正正正	15	0.15
[1.5, 2)	正正正正正	22	0.22
[2, 2.5)	正正正正正正	25	0.25
[2.5, 3)	正正正	14	0.14
[3, 3.5)	正	6	0.06
[3.5, 4)	正	4	0.04
[4, 4.5]	正	2	0.02
合计		100	1.00

表 2-2 的最后一列是各小组的频率, 例如第一小组的频率是:

$$\frac{\text{第一组频数}}{\text{样本容量}} = \frac{4}{100} = 0.04.$$

5. 画频率分布直方图

根据表 2-2 可以得到如图 2.2-1 所示的**频率分布直方图**。

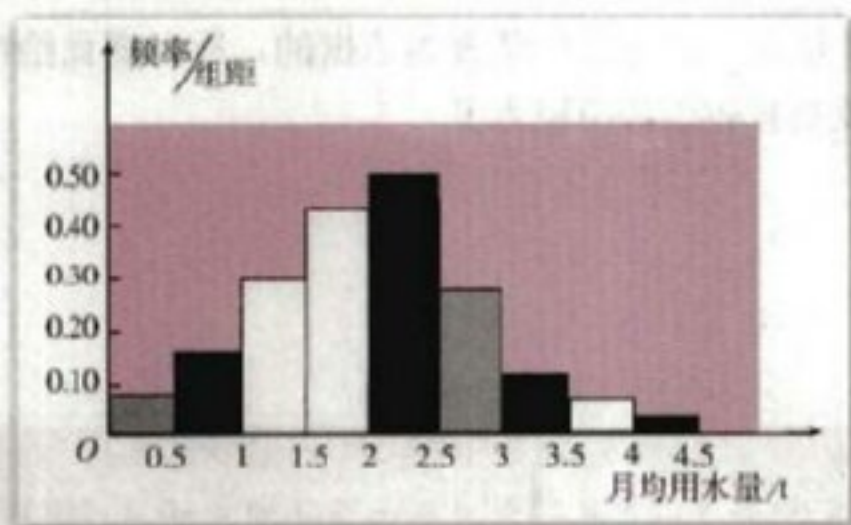


图 2.2-1

图 2.2-1 中, 横轴表示月均用水量, 纵轴表示频率/组距。由于

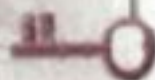
$$\text{小长方形的面积} = \text{组距} \times \frac{\text{频率}}{\text{组距}} = \text{频率},$$

所以各小长方形的面积表示相应各组的频率。这样, 频率分布直方图就以面积的形式反映了数据落在各个小组的频率的大小。

容易知道, 在频率分布直方图中, 各小长方形的面积的总和等于 1。

用计算机中的 Excel 软件很容易作出频率分布直方图, 有条件的同学可以试一试。

图形有“好”与“坏”之分。如果复杂的思想能够在图中清晰、准确、有效地表达出来, 那么就是一幅好图。





探究

同样一组数据,如果组距不同,横轴、纵轴的单位不同,得到的图的形状也会不同,不同的形状给人以不同的印象,这种印象有时会影响我们对总体的判断.分别以0.1和1为组距重新作图,然后谈谈你对图的印象.

表2-2和图2.2-1显示了样本数据落在各个小组的比例大小,从中我们可以看到,月均用水量在区间 $[2, 2.5)$ 内的居民最多,在 $[1.5, 2)$ 内的次之,大部分居民的月均用水量都在 $[1, 3)$ 之间.

直方图能够很容易地表示大量数据,非常直观地表明分布的形状,使我们能够看到在分布表中看不清楚的数据模式.例如,从图2.2-1可以清楚地看到,居民月均用水量的分布是“山峰”状的,而且是“单峰”的,另外还有一定的对称性,这说明,大部分居民的月均用水量集中在一个中间值附近,只有少数居民的月均用水量很多或很少.但是,直方图也丢失了一些信息,例如,原始数据不能在图中表示出来.

根据样本数据的频率分布,我们就可以推测这一城市全体居民月均用水量分布的大致情况.也就是根据样本的频率分布,我们可以大致估计出总体的分布.因为这种估计是以一定的统计调查为依据的,所以据此给市政府提出每位居民月用水量标准的建议,就具有较强的说服力了.

现实中,许多数据的分布都是单峰而且对称的,如身高、体重、考试成绩、农作物产量、某种特定型产品的各种质量指标、股票价格等.请查阅资料做进一步了解.



思考?

如果当地政府希望使85%以上的居民每月的用水量不超出标准,根据频率分布表2-2和频率分布直方图2.2-1,你能对制定月用水量标准提出建议吗?

由表2-2和图2.2-1可以看出,月用水量在3 t以上的居民所占的比例为 $6\% + 4\% + 2\% = 12\%$,即大约有12%的居民月用水量在3 t以上,88%的居民月用水量在3 t以下.因此,居民月用水量标准定为3 t是一个可以考虑的标准.

想一想,你认为3 t这个标准一定能够保证85%以上的居民用水不超标吗?如果不一定,那么哪些环节可能会导致结论的差别?

实际上,这个标准还可能出现偏差,所以,在实践中,对统计结论是需要进行评价的.

类似于频数分布折线图, 连接频率分布直方图中各小长方形上端的中点, 就得到**频率分布折线图** (图 2.2-2).

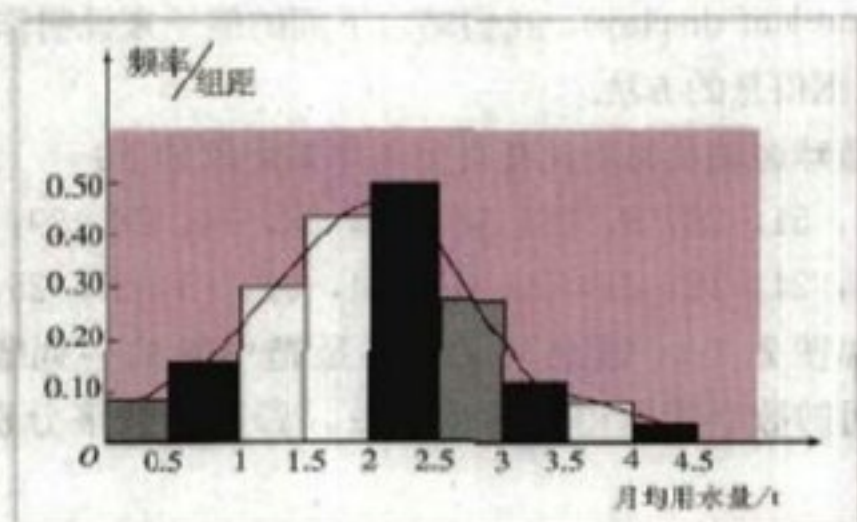


图 2.2-2

一般地, 当总体中的个体数较多时, 抽样时样本容量就不能太小. 例如, 如果要抽样调查一个省乃至全国的居民的月均用水量, 那么样本容量就应比调查一个城市的时候大. 可以想像, 随着样本容量的增加, 作图时所分的组数增加, 组距减小, 相应的频率折线图会越来越接近于一条光滑曲线, 统计中称这条光滑曲线为**总体密度曲线**, 如图 2.2-3 所示. 总体密度曲线反映了总体在各个范围内取值的百分比, 它能给我们提供更加精细的信息. 例如, 图中有阴影部分的面积, 就是总体在区间 (a, b) 内取值的百分比.

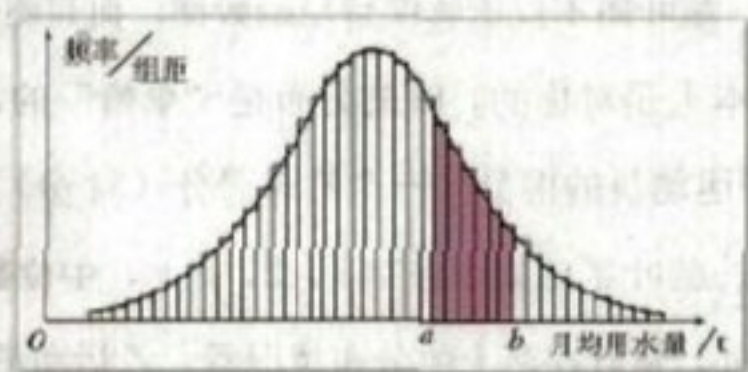


图 2.2-3



可以用样本的频率分布折线图得到准确的总体密度曲线吗?

实际上, 尽管有些总体密度曲线是客观存在的, 但是在实际应用中我们并不知道它的具体表达形式, 需要用样本来估计. 由于样本是随机的, 不同的样本得到的频率分布折线图不同; 即使对于同一样本, 不同的分组情况得到的频率分布折线图也不同. 频率分布折

线图是随着样本的容量和分组情况的变化而变化的,因此不能由样本的频率分布折线图得到准确的总体密度曲线.

除了上面几种图、表能帮助我们理解样本数据外,统计中还有一种被用来表示数据的图叫做**茎叶图**(stem-and-leaf display),我们结合下面的例子来说明作茎叶图的方法,以及从茎叶图中提取样本数据信息的方法.

某赛季甲、乙两名篮球运动员每场比赛得分的原始记录如下:

甲运动员得分:13, 51, 23, 8, 26, 38, 16, 33, 14, 28, 39;

乙运动员得分:49, 24, 12, 31, 50, 31, 44, 36, 15, 37, 25, 36, 39.

用茎叶图表示,如图 2.2-4. 顾名思义,茎是指中间的一列数,叶就是从茎的旁边生长出来的数. 中间的数字表示得分的十位数,旁边的数字分别表示两个人得分的个位数.



图 2.2-4

从图 2.2-4 可以看出,茎叶图不仅能够保留原始数据,而且能够展示数据的分布情况. 比如,乙运动员的得分基本上是对称的,叶的分布是“单峰”的,有 $\frac{10}{13}$ 的叶集中在茎 2, 3, 4 上,中位数是 36;甲运动员的得分除一个特殊得分(51 分)外,也大致对称,叶的分布也是“单峰”的,有 $\frac{9}{11}$ 的叶主要集中在茎 1, 2, 3 上,中位数是 26. 由此可以看出,乙运动员的成绩更好. 另外,从叶在茎上的分布情况看,乙运动员的得分更集中于峰值附近,这说明乙运动员的发挥更稳定.

在样本数据较少时,用茎叶图表示数据的效果较好. 它不但可以保留所有信息,而且可以随时记录,这对数据的记录和表示都能带来方便. 但当样本数据较多时,茎叶图就显得不太方便. 因为每一个数据都要在图中占据一个空间,如果数据很多,枝叶就会很长.

练习

1. 从一种零件中抽取了 80 件, 尺寸数据表示如下 (单位: cm):

362.51×1	362.62×2	362.72×2	362.83×3
362.93×3	363.03×3	363.15×5	363.26×6
363.38×8	363.49×9	363.59×9	363.67×7
363.76×6	363.84×4	363.93×3	364.03×3
364.12×2	364.22×2	364.31×1	364.41×1

这里用 $x \times n$ 表示有 n 件尺寸为 x 的零件, 如 362.51×1 表示有 1 件尺寸为 362.51 cm 的零件.

(1) 作出样本的频率分布表和频率分布直方图;

(2) 在频率分布直方图中画出频率分布折线图.

2. 请班上的每个同学估计一下自己每天的课外学习时间 (单位: min), 然后作出课外学习时间的频率分布直方图. 你认为能否由这个频率分布直方图估计出你们学校的学生课外学习时间的分布情况? 可以用它来估计该地区的学生课外学习时间分布情况吗? 为什么?

3. 下面一组数据是某生产车间 30 名工人某日加工零件的个数, 请设计适当的茎叶图表示这组数据, 并由图出发说明一下这个车间此日的生产情况.

134	112	117	126	128	124	122	116	113	107
116	132	127	128	126	121	120	118	108	110
133	130	124	116	117	123	122	120	112	112

2.2.2

用样本的数字特征估计总体的数字特征

上一节我们学习了用图、表来组织样本数据, 并且学习了如何通过图、表所提供的信息, 用样本的频率分布估计总体的分布. 为了从整体上更好地把握总体的规律, 我们还需要通过样本的数据对总体的数字特征进行研究.

探究

(1) 怎样将各个样本数据汇总为一个数值, 并使它成为样本数据的“中心点”?

(2) 能否用一个数值来描写样本数据的离散程度?

1. 众数、中位数、平均数

初中我们曾经学过众数、中位数、平均数等各种数字特征. 应当说, 这些数字都能够为我们提供关于样本数据的特征信息. 例如, 在上一节调查 100 位居民的月均用水量的问题中, 从这些样本数据的频率分布直方图可以看出, 月均用水量的众数是 2.25 t (最高的矩形的中点) (如图 2.2-5), 它告诉我们, 该市的月均用水量为 2.25 t 的居民数比月均用水量为其他值的居民数多, 但它并没有告诉我们多多少.

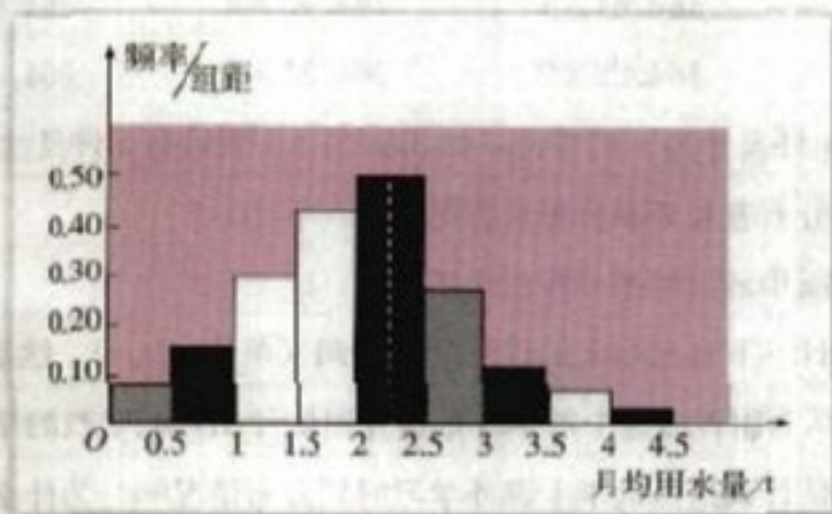


图 2.2-5

那么, 如何从频率分布直方图中估计中位数呢? 在样本中, 有 50% 的个体小于或等于中位数, 也有 50% 的个体大于或等于中位数. 因此, 在频率分布直方图中, 中位数左边和右边的直方图的面积应该相等, 由此可以估计中位数的值. 图 2.2-6 中的虚线代表居民月均用水量的中位数的估计值, 其左边的直方图的面积代表着 50 个单位, 右边的直方图也是 50 个单位. 虚线处的数据值为 2.02.



2.02 这个中位数的估计值, 与样本的中位数值 2.0 不一样, 你能解释其中的原因吗?

同样地, 可以从频率分布直方图中估计平均数. 平均数是频率分布直方图的“重心”, 等于频率分布直方图中每个小矩形的面积乘以小矩形底边中点的横坐标之和. 由图 2.2-6 估计可知, 居民月用水量的平均数是 2.02 t.

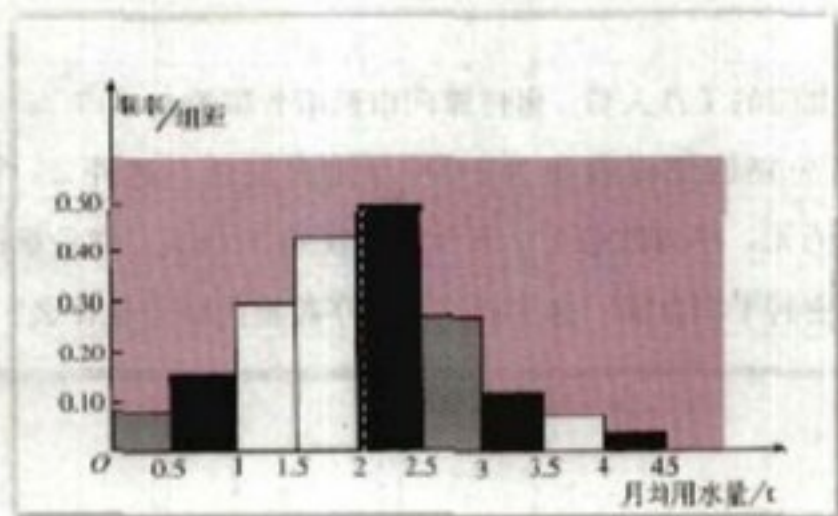


图 2.2-6

图 2.2-6 显示, 大部分居民的月均用水量在中部 (2.02 t 左右), 但也有少数居民的月均用水量特别高. 显然, 对这部分居民的用水作出限制是非常合理的.



样本中位数不受少数几个极端值的影响, 这在某些情况下是一个优点, 但它对极端值的不敏感有时也会成为缺点. 你能举例说明吗?

由于样本平均数与每一个样本数据有关, 所以, 任何一个样本数据的改变都会引起平均数的改变. 这是中位数、众数都不具有的性质. 也正因为这个原因, 与众数、中位数比较起来, 平均数可以反映出更多的关于样本数据全体的信息. 在本例中, 用水量最多的几个居民对平均数影响较大, 这是因为他们的月均用水量与平均数相差太多了.



“用数据说话”, 这是我们经常可以听到的一句话. 但是, 数据有时也会被利用, 从而产生误导. 例如, 一个企业中, 绝大多数是一线工人, 他们的年收入可能是一万元左右, 另有一些经理层次的人, 年收入可以达到几十万元. 这时, 年收入的平均数会比中位数大得多. 尽管这时中位数比平均数更合理些, 但是这个企业的老板到人力市场去招聘工人时, 也许更可能用平均数来回答有关工资待遇方面的提问.

你认为“我们单位的收入水平比别的单位高”这句话应当怎么解释?

练习

假设你是一名交通部门的工作人员. 你打算向市长报告国家对本市 26 个公路项目投资的平均资金数额, 其中一条新公路的建设投资为 2 000 万元人民币, 另外 25 个项目的投资是 20~100 万元. 中位数是 25 万元, 平均数是 100 万元, 众数是 20 万元. 你会选择哪一种数字特征来表示国家对每一个项目投资的平均金额? 你选择这种数字特征的缺点是什么?

2. 标准差

平均数向我们提供了样本数据的重要信息, 但是, 平均数有时也会使我们作出对总体的片面判断. 某地区的统计报表显示, 此地区的年平均家庭收入是 10 万元, 给人的印象是这个地区的家庭收入普遍较高. 但是, 如果这个平均数是从 200 户贫困家庭和 20 户极富有的家庭收入计算出来的, 那么, 它就既不能代表贫困户家庭的年收入, 也不能代表极富有家庭的年收入. 因为这个平均数掩盖了一些极端的情况, 而这些极端情况显然是不能忽视的. 因此, 只有平均数还难以概括样本数据的实际状态.

又如, 有两位射击运动员在一次射击测试中各射靶 10 次, 每次命中的环数如下:

甲 7 8 7 9 5 4 9 10 7 4

乙 9 5 7 8 7 6 8 6 7 7

如果你是教练, 你应当如何对这次射击情况作出评价? 如果这是一次选拔性考核, 你应当如何作出选择?

如果看两人本次射击的平均成绩, 由于

$$\bar{x}_{\text{甲}} = 7, \quad \bar{x}_{\text{乙}} = 7,$$

两人射击的平均成绩是一样的. 那么, 是否两个人的水平就没有什么差异呢?

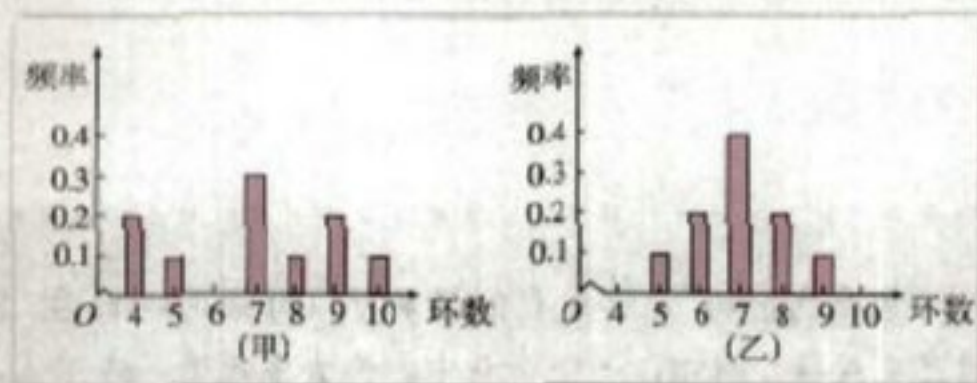


图 2.2-7

由图 2.2-7 来看, 还是有差异的. 例如, 甲成绩比较分散, 乙的成绩相对集中. 因此, 我们还需要从另外的角度来考察这两组数据. 例如, 在作统计图、表时提到过的极差

$$\text{甲的环数极差} = 10 - 4 = 6,$$

$$\text{乙的环数极差} = 9 - 5 = 4,$$

它们在一定程度上表明了样本数据的分散程度, 与平均数一起, 可以给我们许多关于样本数据的信息. 显然, 极差对极端值非常敏感, 注意到这一点, 我们可以得到一种“去掉一个最高分, 去掉一个最低分”的统计策略.

考察样本数据的分散程度的大小,最常用的统计量是**标准差** (standard deviation), 标准差是样本数据到平均数的一种平均距离,一般用 s 表示.

所谓“平均距离”,其含义可作如下理解:

假设样本数据是 x_1, x_2, \dots, x_n , \bar{x} 表示这组数据的平均数, x_i 到 \bar{x} 的距离是

$$|x_i - \bar{x}| \quad (i=1, 2, \dots, n).$$

于是,样本数据 x_1, x_2, \dots, x_n 到 \bar{x} 的“平均距离”是

$$S = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \dots + |x_n - \bar{x}|}{n}.$$

由于上式含有绝对值,运算不太方便,因此,通常改用如下公式来计算标准差

$$s = \sqrt{\frac{1}{n} [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \dots + (x_n - \bar{x})^2]}.$$

考虑一个容量为 2 的样本: $x_1 < x_2$, 其样本的标准差为 $\frac{x_2 - x_1}{2}$.

记 $a = \frac{x_2 - x_1}{2}$, 样本中的个体与平均数之间的距离关系可用图 2.2-8 表示:

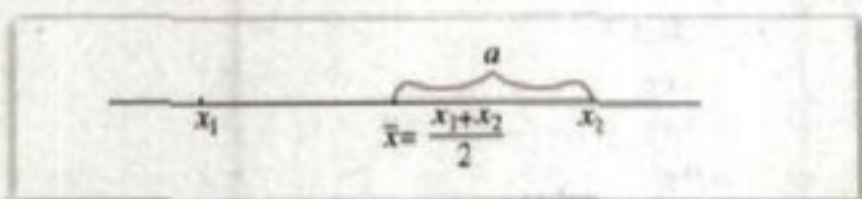


图 2.2-8

显然,标准差越大,则 a 越大,数据的离散程度越大;标准差越小,数据的离散程度越小.

用计算器计算运动员甲的成绩的标准差的过程如下:

MODE 2 (进入统计计算模式)

SHIFT CLR 1 = (清除统计存储器)

7 DT 8 DT 7 DT 9 DT 5 DT

4 DT 9 DT 10 DT 7 DT 4 DT

SHIFT S-VAR 2 = (计算样本标准差)

2

即 $s_{\text{甲}} = 2$.

用类似的方法,可得 $s_{\text{乙}} \approx 1.095$.

由 $s_{\text{甲}} > s_{\text{乙}}$ 可以知道,甲的成绩离散程度大,乙的成绩离散程度小.由此可以估计,乙比甲的射击成绩稳定.

在一般计算器中,均设有计算标准差的按键.在求一组数据的标准差时,只要让计算器处于统计状态,将数据逐个输入,然后按标准差的键,即可立即显示所求标准差的值.

标准差的取值范围是什么? 标准差为 0 的样本数据有什么特点?

不同计算器的参数可能不同,例如有的计算器的统计模式为“MODE 1”,计算样本标准差的参数为 3.

上面两组数据的离散程度与标准差之间的关系可用图2.2-9直观地表示出来.

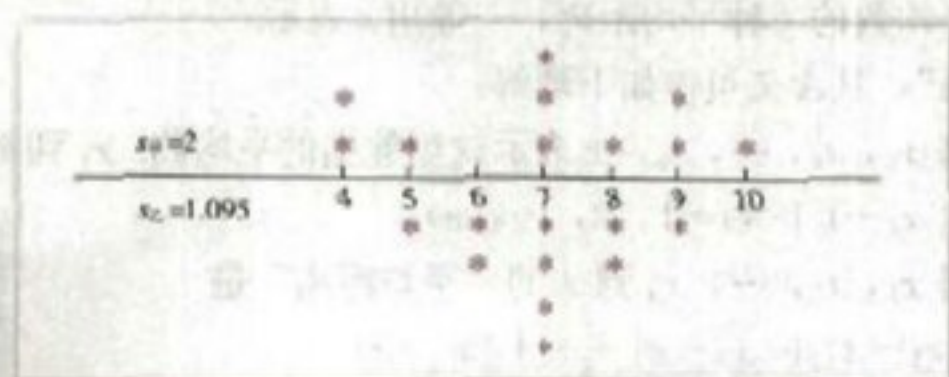


图 2.2-9

例 1 画出下列四组样本数据的条形图, 说明它们的异同点.

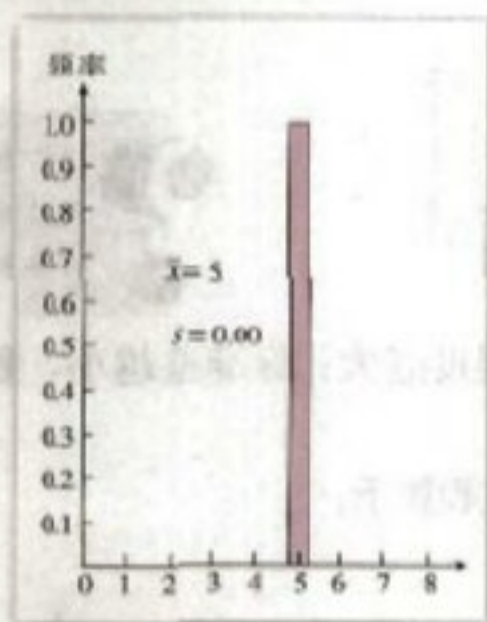
(1) 5, 5, 5, 5, 5, 5, 5, 5, 5;

(2) 4, 4, 4, 5, 5, 5, 6, 6, 6;

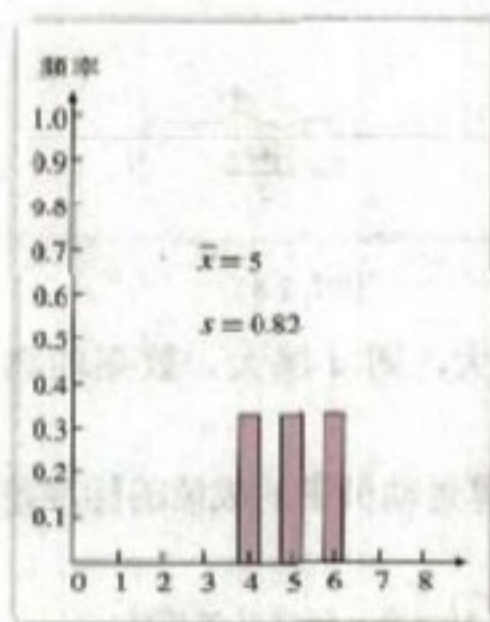
(3) 3, 3, 4, 4, 5, 6, 6, 7, 7;

(4) 2, 2, 2, 2, 5, 8, 8, 8, 8.

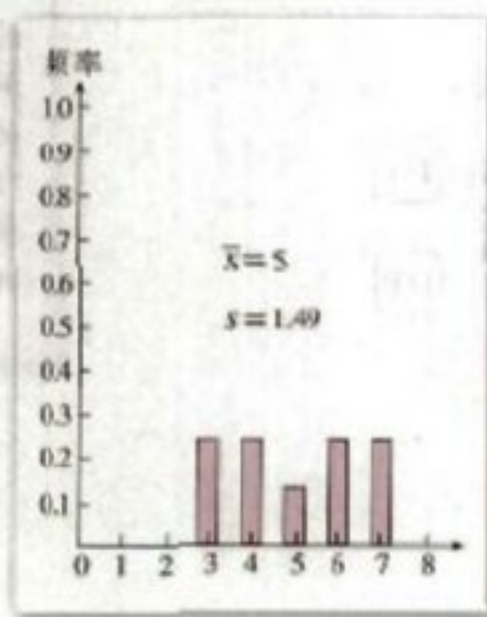
解: 四组样本数据的条形图是:



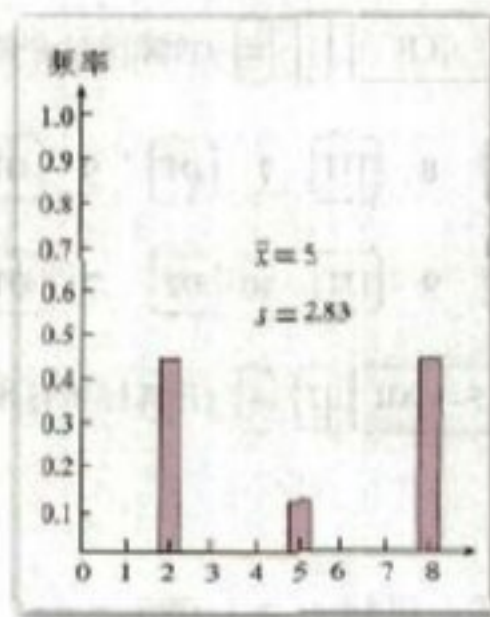
(1)



(2)



(3)



(4)

图 2.2-10

四组数据的平均数都是 5.0, 标准差分别是 0.00, 0.82, 1.49, 2.83. 虽然它们有相同的平均数, 但是它们有不同的标准差, 说明数据的分散程度是不一样的.

标准差还可以用于对样本数据的另外一种解释. 例如, 在关于居民月均用水量的例子中, 平均数 $\bar{x}=1.973$, 标准差 $s=0.868$, 所以

$$\bar{x}+s=2.841, \quad \bar{x}+2s=3.709;$$

$$\bar{x}-s=1.105, \quad \bar{x}-2s=0.237.$$

这 100 个数据中, 在区间 $[\bar{x}-2s, \bar{x}+2s]=[0.237, 3.709]$ 外的只有 4 个, 也就是说, $[\bar{x}-2s, \bar{x}+2s]$ 几乎包含了所有样本数据.

从数学的角度考虑, 人们有时用标准差的平方 s^2 ——**方差**来代替标准差, 作为测量样本数据分散程度的工具:

$$s^2 = \frac{1}{n} [(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2].$$

显然, 在刻画样本数据的分散程度上, 方差与标准差是一样的. 但在解决实际问题时, 一般多采用标准差.

需要指出的是, 现实中的总体所包含的个体数往往是很的, 总体的平均数与标准差是不知道的. 如何求得总体的平均数和标准差呢? 通常的做法是用样本的平均数和标准差去估计总体的平均数与标准差. 这与前面用样本的频率分布来近似地代替总体分布是类似的. 只要样本的代表性好, 这样做就是合理的, 也是可以接受的.

例 2 甲乙两人同时生产内径为 25.40 mm 的一种零件. 为了对两人的生产质量进行评比, 从他们生产的零件中各抽出 20 件, 量得其内径尺寸如下 (单位: mm):

甲

25.46	25.32	25.45	25.39	25.36
25.34	25.42	25.45	25.38	25.42
25.39	25.43	25.39	25.40	25.44
25.40	25.42	25.35	25.41	25.39

乙

25.40	25.43	25.44	25.48	25.48
25.47	25.49	25.49	25.36	25.34
25.33	25.43	25.43	25.32	25.47
25.31	25.32	25.32	25.32	25.48

从生产的零件内径的尺寸看, 谁生产的质量较高?

分析：每一个工人生产的所有零件的内径尺寸组成一个总体. 由于零件的生产标准已经给出 (内径 25.40 mm), 生产质量可以从总体的平均数与标准差两个角度来衡量. 总体的平均数与内径标准尺寸 25.40 mm 的差异大时质量低, 差异小时质量高; 当总体的平均数与标准尺寸很接近时, 总体的标准差小的时候质量高, 标准差大的时候质量低. 这样, 比较两人的生产质量, 只要比较他们所生产的零件内径尺寸所组成的两个总体的平均数与标准差的大小即可. 但是, 这两个总体的平均数与标准差都是不知道的, 根据用样本估计总体的思想, 我们可以通过抽样分别获得相应的样本数据, 然后比较这两个样本的平均数、标准差, 以此作为两个总体之间差异的估计值.

解：用计算器计算可得

$$\bar{x}_1 \approx 25.401, \quad \bar{x}_2 \approx 25.406;$$

$$s_1 \approx 0.037, \quad s_2 \approx 0.068.$$

从样本平均数看, 甲生产的零件内径比乙的更接近内径标准 (25.40 mm), 但是差异很小; 从样本标准差看, 由于 $s_1 < s_2$, 因此甲生产的零件内径比乙的稳定程度高得多. 于是, 可以作出判断, 甲生产的零件的质量比乙的高一些.

从上述例子我们可以看到, 对一名工人生产的零件内径 (总体) 的质量判断, 与所抽取的零件内径 (样本数据) 直接相关. 显然, 我们可以从这名工人生产的零件中获取许多样本 (为什么?). 这样, 尽管总体是同一个, 但由于样本不同, 相应的样本频率分布与平均数、标准差等都会发生改变, 这就会影响到我们对总体情况的估计. 如果样本的代表性差, 那么对总体所作出的估计就会产生偏差; 样本没有代表性时, 对总体作出错误估计的可能性就非常大. 这也正是我们在前面讲随机抽样时反复强调样本代表性的理由. 在实际操作中, 为了减少错误的发生, 条件许可时, 通常采取适当增加样本容量的方法. 当然, 关键还是要改进抽样方法, 提高样本的代表性.

为什么说“两个总体的平均数与标准差都是不知道的”? 25.40 mm 为什么不是它们的平均数?

如果一个总体包含 6 个个体, 现在要从中抽出 3 个作为样本, 请你列出所有可能的样本.

假如以你班全体同学的身高作为总体, 现从中抽出 20 名同学的身高组成样本, 想像一下可能有多少个不同的样本?

练习

1. 农场种植的甲乙两种水稻, 在面积相等的两块稻田中连续 6 年的年平均产量如下(单位: 500 g):

品种	第 1 年	第 2 年	第 3 年	第 4 年	第 5 年	第 6 年
甲	900	920	900	850	910	920
乙	890	960	950	850	860	890

哪种水稻的产量比较稳定?

2. 一个小商店从一家食品有限公司购进 21 袋白糖, 每袋白糖的标准重量是 500 g, 为了了解这些白糖的重量情况, 称出各袋白糖的重量(单位: g) 如下:

486	495	496	498	499	493	493
498	484	497	504	489	495	503
499	503	509	498	487	500	508

求:

(1) 21 袋白糖的平均重量 \bar{x} 是多少? 标准差 s 是多少?

(2) 重量位于 $\bar{x}-s$ 与 $\bar{x}+s$ 之间有多少袋白糖? 所占的百分比是多少?

3. 下列数据是 30 个不同国家中每 100 000 名男性患某种疾病的死亡率:

27.0	23.9	41.6	33.1	40.6	18.8	13.7	28.9	13.2	14.5
27.0	34.8	28.9	3.2	50.1	5.6	8.7	15.2	7.1	5.2
16.5	13.8	19.2	11.2	15.7	10.0	5.6	1.5	33.8	9.2

(1) 作出这些数据分布的频率分布直方图;

(2) 请由这些数据计算平均数、中位数和标准差, 并对它们的含义进行解释.



生产过程中的质量控制图

我们知道, 平均数 μ 表明了总体的重心所在, 标准差 σ 表明了总体的离散程度. 但是, 当我们从样本数据中计算出这两个数值后, 其他信息就丢失了. 所以, 这两个数值并不能刻画总体的全貌. 不过, 现实生活中, 有一些总体(如某地区同龄儿童的身高、体重等)的分布的密度曲线是由它的平均数 μ 与标准差 σ 完全确定的(图 1~图 3), 我们把这种分布记作 $N(\mu, \sigma^2)$, 称为平均数为 μ , 方差为 σ^2 的正态分布.

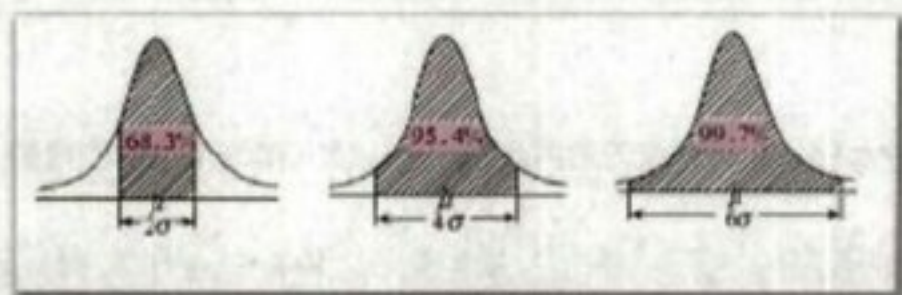


图1

图2

图3

从密度曲线图可以测量出这个总体在 $(\mu - \sigma, \mu + \sigma)$ $(\mu - 2\sigma, \mu + 2\sigma)$ 和 $(\mu - 3\sigma, \mu + 3\sigma)$ 等区间内取值的百分比是:

区 间	取值的百分比
$(\mu - \sigma, \mu + \sigma)$	68.3%
$(\mu - 2\sigma, \mu + 2\sigma)$	95.4%
$(\mu - 3\sigma, \mu + 3\sigma)$	99.7%

上述总体分布在产品质量控制中的应用是非常广泛的. 例如, 工人生产零件时, 零件尺寸一般服从 $N(\mu, \sigma^2)$ 分布. 这样, 零件尺寸在 $(\mu - 3\sigma, \mu + 3\sigma)$ 以外取值的只有 0.3%, 它表明在大量重复试验中, 平均每抽取 1 000 个零件, 属于这个范围以外的尺寸大约有 3 个. 因此在一批产品中随机抽取一个零件, 零件尺寸在 $(\mu - 3\sigma, \mu + 3\sigma)$ 以外是几乎不可能发生的. 一旦这种情况发生, 即零件尺寸 x 满足 $|x - \mu| \geq 3\sigma$, 我们就有理由认为生产中可能出现了异常情况. 比如, 可能原料、机器出了问题, 或工艺规程不完善, 或工人操作时精力不集中等. 这种情况下, 需要停机检查, 找出原因, 使生产过程重新控制在一种正常状态, 从而避免继续生产更多的次品, 以保证产品质量.

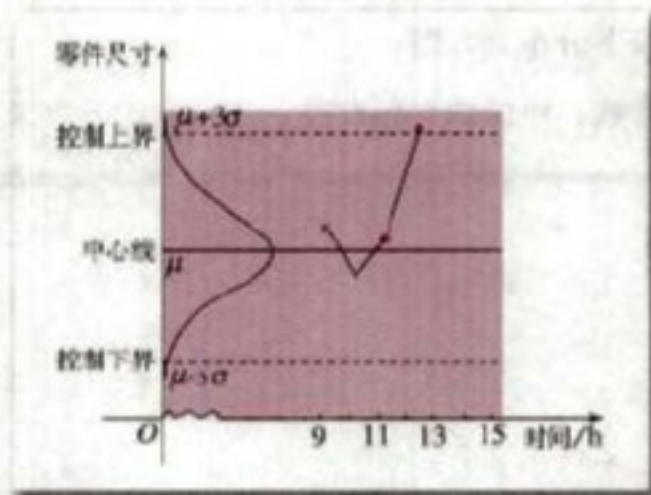


图4

这就是运用统计原理进行产品质量控制的基本思想. 目前, 在生产中广泛运用的质量控制图 (图4), 就是根据上述原理制作的.

图4实际上是将图3旋转 90° 后得到的. 在生产过程中, 从某一时刻起, 每隔一定时间任取一个零件进行检查, 将其尺寸用圆点在图中表示出来, 如果圆点在控制界限以内, 可认为生产情况正常; 如果圆点超出控制界限, 可认为有异常情况发生, 应该停机检查.

至此, 你对标准差的含义是否有了进一步的理解? 请你根据上述阅读材料谈谈你对标准差的认识.

习题 2.2

A 组

1. 有一种鱼的身体吸收汞，汞的含量超过体重的 1.00 ppm（即百万分之一）时就会对人体产生危害。在 30 条鱼的样本中发现的汞含量是：

0.07	0.24	0.95	0.98	1.02	0.98	1.37	1.40	0.39	1.02
1.44	1.58	0.54	1.08	0.61	0.72	1.20	1.14	1.62	1.68
1.85	1.20	0.81	0.82	0.84	1.29	1.25	2.10	0.91	1.31

- (1) 用前两位数作为茎，画出样本数据的茎叶图；
 - (2) 描述一下汞含量的分布特点；
 - (3) 从实际情况看，许多鱼的汞含量超标在于有些鱼在出售之前没有被检查过，每批这种鱼的平均汞含量都比 1.00 ppm 大吗？
 - (4) 求出上述样本数据的平均数和标准差；
 - (5) 有多少条鱼的汞含量在平均数与 2 倍标准差的和（差）的范围内？
2. 在一批棉花中抽测了 60 根棉花的纤维长度，结果如下（单位：mm）：

82	202	352	321	25	293	293	86	28	206
323	355	357	33	325	113	233	294	50	296
115	236	357	326	52	301	140	328	238	358
58	255	143	360	340	302	370	343	260	303
59	146	60	263	170	305	380	346	61	305
175	348	264	383	62	306	195	350	265	385

作出这个样本的频率分布直方图（在对样本数据分组时，可试用几种不同的分组方式，然后从中选择一种较为合适的分组方法）。棉花的纤维长度是棉花质量的重要指标，你能从图中分析出这批棉花的质量状况吗？

3. 以往的招生统计数据显示，某所大学录取的新生高考总分的中位数基本上稳定在 550 分。你的一位校友在今年的高考中得了 520 分，你是立即劝阻他报考这所大学，还是先查阅一下这所大学招生的其他信息？解释一下你的选择。
4. 在去年的足球甲 A 联赛上，一队每场比赛平均失球数是 1.5，全年比赛失球个数的标准差为 1.1；二队每场比赛平均失球数是 2.1，全年失球个数的标准差是 0.4。你认为下列说法中哪一种是正确的，为什么？
 - (1) 平均说来一队比二队防守技术好；
 - (2) 二队比一队技术水平更稳定；
 - (3) 一队有时表现很差，有时表现又非常好；
 - (4) 二队很少不失球。

5. 在一次人才招聘会上, 有一家公司的招聘员告诉你, “我们公司的收入水平很高” “去年, 在 50 名员工中, 最高年收入达到了 100 万, 他们年收入的平均数是 3.5 万”. 如果你希望获得年薪 2.5 万元,
- (1) 你是否能够判断自己可以成为此公司的一名高收入者?
 - (2) 如果招聘员继续告诉你, “员工收入的变化范围是从 0.5 万到 100 万”, 这个信息是否足以使你作出自己是否受聘的决定? 为什么?
 - (3) 如果招聘员继续给你提供了如下信息, 员工收入的中间 50% (即去掉最少的 25% 和最多的 25% 后所剩下的) 的变化范围是 1 万到 3 万, 你又该如何使用这条信息来作出是否受聘的决定?
 - (4) 你能估计出收入的中位数是多少吗? 为什么均值比估计出的中位数高很多?
6. 甲乙两台机床同时生产一种零件, 10 天中, 两台机床每天出的次品数分别是:

甲 0 1 0 2 2 0 3 1 2 4

乙 2 3 1 1 0 2 1 1 0 1

分别计算这两组数据的平均数与标准差, 从计算结果看, 哪台机床的性能较好?

7. 有 20 种不同的零食, 它们的热量含量如下:

110	120	123	165	432	190	174	235	428	318
249	280	162	146	210	120	123	120	150	140

- (1) 以上述 20 个数据组成总体, 求总体平均数与总体标准差.
- (2) 设计恰当的随机抽样方法, 从总体中抽取一个容量为 7 的样本, 求样本的平均数与标准差.
- (3) 利用上面的抽样方法, 再抽取容量为 7 的样本, 计算样本的平均数和标准差, 这个样本的平均数与标准差和 (2) 中的结果一样吗? 为什么?
- (4) 利用 (2) 中的随机抽样方法, 分别从总体中抽取一个容量为 10, 13, 16, 19 的样本, 求样本的平均数与标准差, 分析样本容量与样本平均数和样本标准差对总体的估计效果之间有什么关系.

B 组

1. 在训练运动员的过程中, 需要进行体能测试, 这种测试通常是由专业部门完成的. 下面的结果是由两个权威部门对 10 名游泳运动员进行测试后给出的.

测试	A	B	C	D	E	F	G	H	I	J
T_1	20	23	24	18	17	16	25	24	21	19
T_2	31	39	39	29	28	31	40	30	31	30

已经知道, 对全国样本, 测试 T_1 的平均数为 20, 标准差为 2; 测试 T_2 的平均数是 35, 标准差是 3.

- (1) 上述两个测试哪一个做得更好些?
 - (2) 如果你是教练, 为了增强你的队员的信心, 你应该选择哪个测试?
 - (3) 分值越高, 运动员的运动水平越高, 哪一名运动员最强? 哪一名运动员最弱?
2. 调查本班每位同学的家庭在同一周的用电量, 作出这组数据的频率分布表、频率分布直方图以及频率折线图, 对你所在地区的用电量情况进行估计, 然后在全班进行讨论.

CHAPTER 2

2.3

变量间的相关关系

2.3.1 变量之间的相关关系



在学校里，老师对学生经常这样说：“如果你的数学成绩好，那么你的物理学习就不会有什么大问题。”按照这种说法，似乎学生的物理成绩与数学成绩之间存在着一种相关关系，这种说法有没有根据呢？

凭我们的学习经验可知，物理成绩确实与数学成绩有一定的关系，但除此以外，还存在其他影响物理成绩的因素，例如，是否喜欢物理，用在物理学习上的时间等等。当我们主要考虑数学成绩对物理成绩的影响时，就是要考察这两者之间的相关关系。

我们还可以举出现实生活中存在的许多相关关系的问题，例如：

1. 商品销售收入与广告支出经费之间的关系。商品销售收入与广告支出经费有着密切的联系，但商品销售收入不仅与广告支出多少有关，还与商品质量、居民收入等因素有关。
2. 粮食产量与施肥量之间的关系。在一定范围内，施肥量越大，粮食产量就越高。但是，施肥量并不是决定粮食产量的唯一因素，因为粮食产量还要受到土壤质量、降雨量、田间管理水平等因素的影响。
3. 人体内的脂肪含量与年龄之间的关系。在一定年龄段内，随着年龄的增长，人体内的脂肪含量会增加，但人体内的脂肪含量还与饮食习惯、体育锻炼等有关，可能还与个人的先天体质有关。

应当说，对于上述各种问题中的两个变量之间的相关关系，我们都可以根据自己的生活、学习经验作出相应的判断，因为“经验当中有规律”。但是，不管你的经验多么丰富，如果只凭经验办事，还是很容易出错的。因此，在分析两个变量之间的相关关系时，我们需要一些有说服力的方法。

在寻找变量之间相关关系的过程中，统计同样发挥着非常重要的作用。因为上面提到的这种关系，并不像匀速直线运动中时间与路程的关系那样是完全确定的，而是带有不确定性。这就需要通过收集大量的数据（有时通过调查，有时通过实验），在对数据进行统计分析的基础上，发现其中的规律，才能对它们之间的关系作出判断。

练习

1. 有关法律规定,香烟盒上必须印上“吸烟有害健康”的警示语,吸烟是否一定会引起健康问题?你认为“健康问题不一定是由吸烟引起的,所以可以吸烟”的说法对吗?
2. 某地区的环境条件适合天鹅栖息繁衍,有人经统计发现了一个有趣的现象,如果村庄附近栖息的天鹅多,那么这个村庄的婴儿出生率也高,天鹅少的地方婴儿的出生率低,于是,他就得出一个结论:天鹅能够带来孩子,你认为这样得到的结论可靠吗?如何证明这个结论的可靠性?

2.3.2 两个变量的线性相关

探究

在一次对人体脂肪含量和年龄关系的研究中,研究人员获得了一组样本数据:

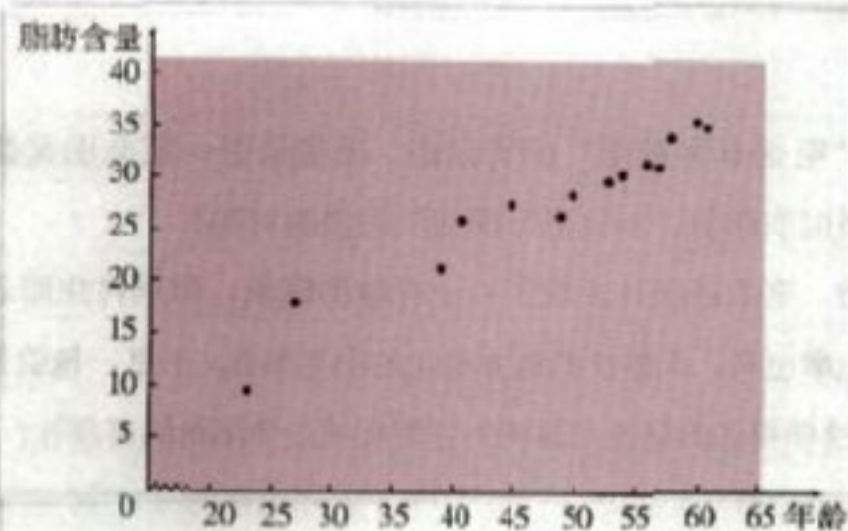
表 2-3 人体的脂肪百分比和年龄

年龄	23	27	39	41	45	49	50
脂肪	9.5	17.8	21.2	25.9	27.5	26.3	28.2
年龄	53	54	56	57	58	60	61
脂肪	29.6	30.2	31.4	30.8	33.5	35.2	34.6

根据上述数据,人体的脂肪含量与年龄之间有这样的关系?

一般地,对于某个人来说,他的体内脂肪不一定随年龄增长而增加或减少,但是如果把很多个体放在一起,这时就可能表现出一定的规律性.各年龄对应的脂肪数据是这个年龄人群脂肪含量的样本平均数.观察表中数据,大体上来看,随着年龄的增加,人体中脂肪的百分比也在增加.为了确定这一关系的细节,我们需要进行数据分析.与以前一样,我们可以作统计图、表.通过作统计图、表,可以使我们对两个变量之间的关系有一个直观上的印象和判断.

下面要作的图叫做**散点图**(scatterplot).对于表 2-3 中的数据,我们假设人的年龄影响体内脂肪含量,于是,按照习惯,以 x 轴表示年龄,以 y 轴表示脂肪含量,得到相应的散点图(图 2.3-1).



计算机可以帮助我们作散点图。实际上，图 2.3-1 就是用计算机作出来的。

图 2.3-1

从散点图可以看出，年龄越大，体内脂肪含量越高。图中点的趋势表明两个变量之间确实存在一定的关系，这个图支持了我们从数据表中得出的结论。

另外，这些点散布的位置也是值得注意的，它们散布在从左下角到右上角的区域。对于两个变量的这种相关关系，我们将它称为**正相关**。还有一些变量，例如汽车的重量和汽车每消耗 1 L 汽油所行驶的平均路程，成**负相关**，汽车越重，每消耗 1 L 汽油所行驶的平均路程就越短，这时的点散布在从左上角到右下角的区域内。



- (1) 两个变量成负相关关系时，散点图有什么特点？
- (2) 你能举出一些生活中的变量成正相关或成负相关的例子吗？

接下来，需要进一步考虑的问题是，当人的年龄增加时，体内脂肪含量到底是以什么方式增加的呢？

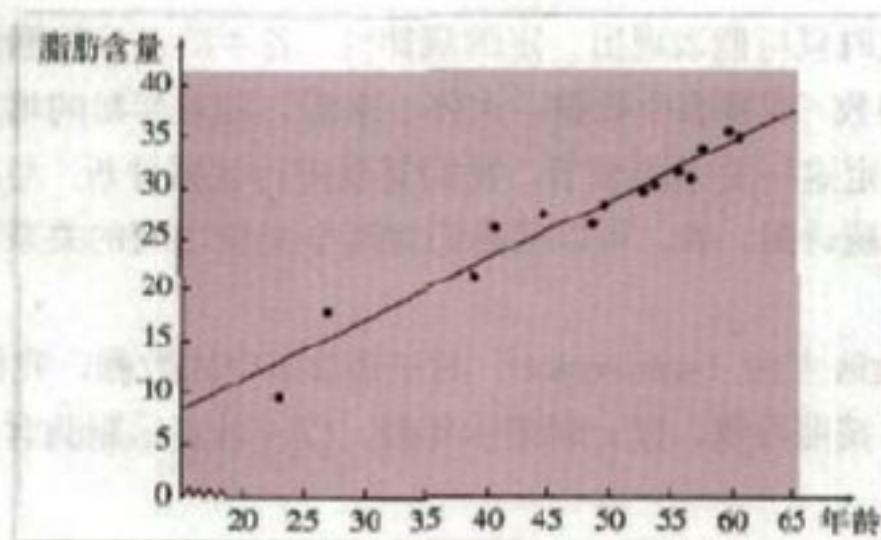


图 2.3-2

从散点图可以看出, 这些点大致分布在通过散点图中心的一条直线附近 (图 2.3-2)。如果散点图中点的分布从整体上看大致在一条直线附近, 我们就称这两个变量之间具有线性相关关系, 这条直线叫做**回归直线** (regression line)。如果能够求出这条回归直线的方程 (简称回归方程), 那么我们就可以比较清楚地了解年龄与体内脂肪含量的相关性。就像平均数可以作为一个变量的数据的代表一样, 这条直线可以作为两个变量具有线性相关关系的代表。

那么, 我们应当如何具体求出这个回归方程呢?

有的同学可能会想, 我可以采用测量的方法, 先画出一条直线, 测量出各点与它的距离, 然后移动直线, 到达一个使距离的和最小的位置, 测量出此时的斜率和截距, 就可得到回归方程了 (图 2.3-3)。但是, 这样做可靠吗?

有的同学可能还会想, 在图中选择这样的两点画直线, 使得直线两侧的点的个数基本相同 (图 2.3-4)。同样地, 这样做能保证各点与此直线在整体上是最接近的吗?

还有的同学会想, 在散点图中多取几组点, 确定出几条直线的方程 (图 2.3-5), 再分别求出各条直线的斜率、截距的平均数, 将这两个平均数当成回归方程的斜率和截距。

同学们不妨去实践一下, 看看这些方法是不是真的可行?

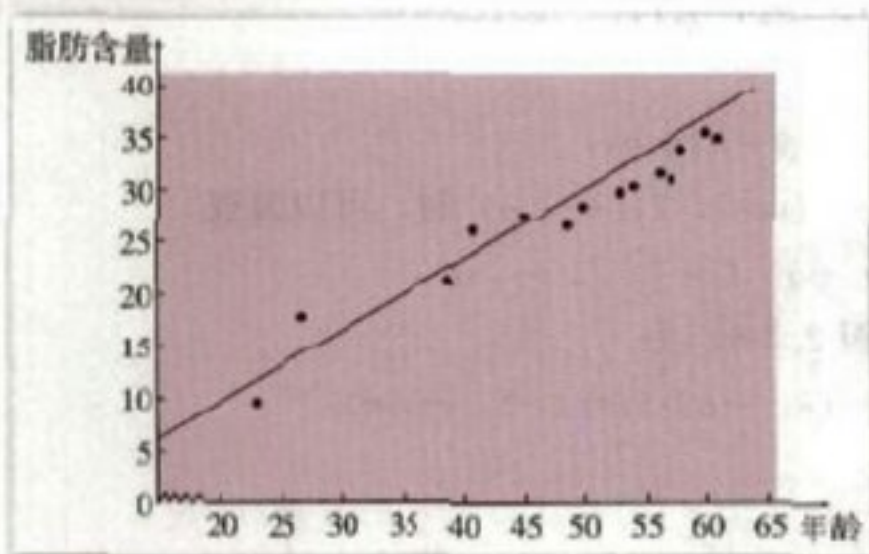


图 2.3-3

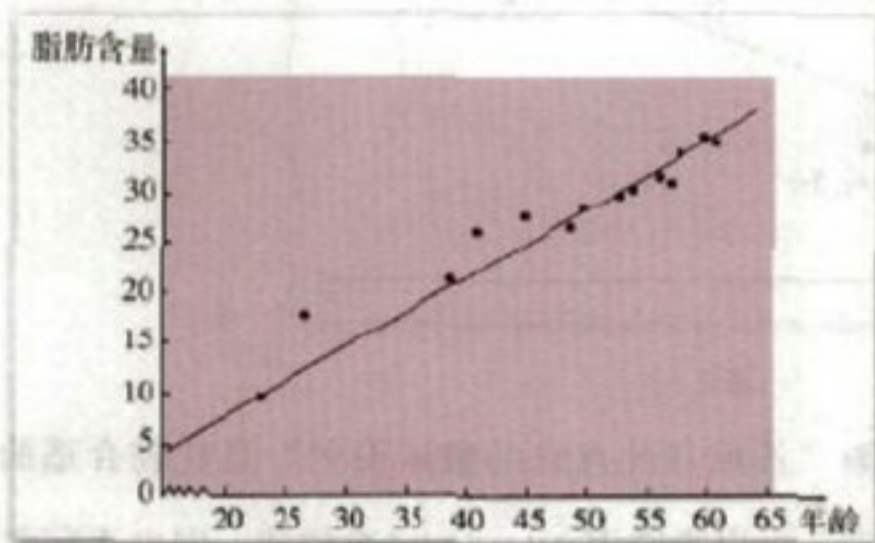


图 2.3-4

回归直线通过样本点的中心, 比照平均数与样本数据之间的关系, 你能说说回归直线与散点图中各点之间的关系吗?

“回归”这个词是由英国著名的统计学家 Francis Galton 提出来的。1889 年, 他在研究祖先与后代身高之间的关系时发现, 身材较高的父母, 他们的孩子也较高, 但这些孩子的平均身高并没有他们父母的平均身高高; 身材较矮的父母, 他们的孩子也较矮, 但这些孩子的平均身高却比他们父母的平均身高高。Galton 把这种后代的身高向中间值靠近的趋势称为“回归现象”。后来, 人们把由一个变量的变化去推测另一个变量的变化的方法称为**回归方法**。

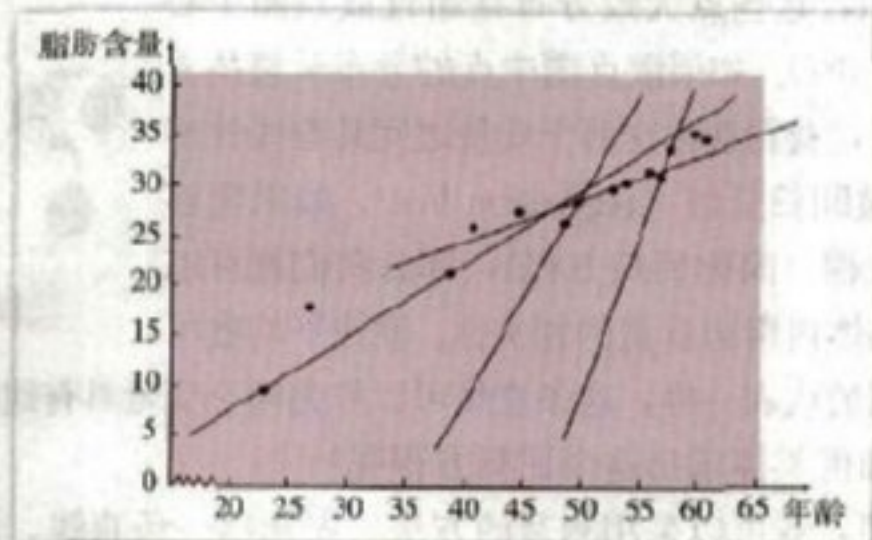


图 2.3-5

上面这些方法虽然有一定的道理,但总让人感到可靠性不强.

实际上,求回归方程的关键是如何用数学的方法来刻画“从整体上看,各点与此直线的距离最小”.

假设我们已经得到两个具有线性相关关系的变量的一组数据

$$(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n).$$

且所求回归方程是

$$\hat{y} = bx + a.$$

其中 a, b 是待定参数. 当变量 x 取 x_i ($i=1, 2, \dots, n$) 时, 可以得到

$$\hat{y}_i = bx_i + a \quad (i=1, 2, \dots, n),$$

它与实际收集到的 y_i 之间的偏差 (图 2.3-6) 是

$$y_i - \hat{y}_i = y_i - (bx_i + a) \quad (i=1, 2, \dots, n).$$

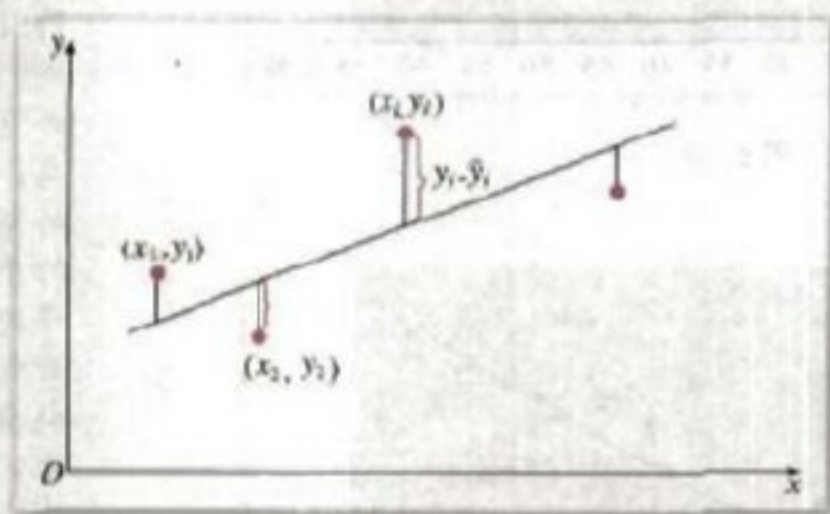


图 2.3-6

这样,用这 n 个偏差的和来刻画“各点与此直线的整体偏差”是比较合适的. 由于 $(y_i - \hat{y}_i)$ 可正可负,为了避免相互抵消,可以考虑用 $\sum_{i=1}^n |y_i - \hat{y}_i|$ 来代替,但由于它含有绝对值,运算不太方便,所以改用



你能解释一下“从整体上看,各点与此直线的距离最小”的含义吗?

$$Q = (y_1 - bx_1 - a)^2 + (y_2 - bx_2 - a)^2 + \cdots + (y_n - bx_n - a)^2 \quad ①$$

来刻画 n 个点与回归直线在整体上的偏差。

这样，问题就归结为：当 a, b 取什么值时 Q 最小，即总体偏差最小，经过数学上求最小值的运算， a, b 的值由下列公式给出：

$$\begin{cases} b = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2} = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}, \\ a = \bar{y} - b\bar{x}. \end{cases} \quad ②$$

其中， b 是回归方程的斜率， a 是截距。

这种通过求①式的最小值而得到回归直线的方法，即求回归直线，使得样本数据的点到它的距离的平方和最小的方法叫做**最小二乘法** (method of least square)。

根据最小二乘法思想和公式②，利用计算器或计算机，可以方便地求出回归方程。

以 Excel 软件为例，用散点图来建立表示人体的脂肪含量与年龄的相关关系的线性回归方程，具体步骤如下：

(1) 在 Excel 中选定表示人体的脂肪含量与年龄的相关关系的散点图 (图 2.3-1)，在菜单中选定“图表”中的“添加趋势线”选项，弹出“添加趋势线”对话框。

(2) 单击“类型”标签，选定“趋势预测/回归分析类型”中的“线性”选项，单击“确定”按钮，得到回归直线。

(3) 双击回归直线，弹出“趋势线格式”对话框，单击“选项”标签，选定“显示公式”，最后单击“确定”按钮，得到回归直线的回归方程 (图 2.3-7)

$$y = 0.577x - 0.448.$$

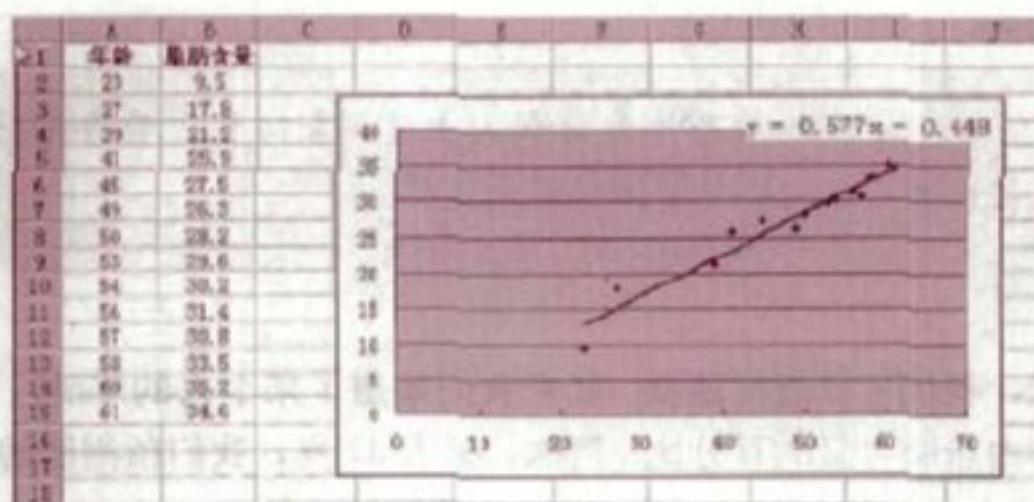


图 2.3-7

用计算器求这个回归方程的过程如下:

MODE 3 1 (进入回归计算模式)

SHIFT CLR 1 = (清除统计存储器)

23	,	9.5	DT	27	,	17.8	DT
39	,	21.2	DT	41	,	25.9	DT
45	,	27.5	DT	49	,	26.3	DT
50	,	28.2	DT	53	,	29.6	DT
54	,	30.2	DT	56	,	31.4	DT
57	,	30.8	DT	58	,	33.5	DT
60	,	35.2	DT	61	,	34.6	DT

SHIFT S-VAR → → 1 = (计算参数 a)

-0.448

SHIFT S-VAR → → 2 = (计算参数 b)

0.577

所以回归方程为 $y = 0.577x - 0.448$.

正像本节开头所说的, 我们从人体脂肪含量与年龄这两个变量的一组随机样本数据中, 找到了它们之间关系的一个规律, 这个规律是由回归直线来反映的.



将表 2-3 中的年龄作为 x 代入上述回归方程, 看看得出的数值与真实数值之间的关系, 从中你体会到什么?

利用回归直线, 我们可以进行预测. 如果我们知道了某个人的年龄, 就可以利用回归方程来预测他的体内脂肪含量的百分比. 例如, 某人 37 岁, 我们预测他的体内脂肪含量在 20.90% ($0.577 \times 37 - 0.448 = 20.90\%$) 附近的可能性比较大. 不过, 我们不能说他的体内脂肪含量一定是 20.90%. 事实上, 这个 20.90% 是对年龄为 37 岁的人群中的大部分人的体内脂肪含量所作出的估计.

例 有一个同学家开了一个小卖部, 他为了研究气温对热饮销售的影响, 经过统

练习

1. 利用本节例题中求出的回归方程, 求当 $x=0$ 时的 y 值, 说明它为什么与实际卖出的热饮杯数不一样.
2. 下表给出了某些地区的鸟的种类数与这些地区的海拔高度, 分析这些数据, 看一看鸟的种类数与海拔高度是否有关.

地区	A	B	C	D	E	F	G	H	I	J	K
种类数	36	30	37	11	11	13	17	13	29	4	15
海拔/m	1 250	1 153	1 067	457	701	731	610	670	1 493	762	549



相关关系的强与弱

我们知道, 两个变量 x 和 y 正(负)相关时, 它们就有相同(反)的变化趋势, 即当 x 由小变大时, 相应的 y 有由小(大)变大(小)的趋势, 因此可以用回归直线来描述这种关系. 与此相关的一个问题是: 如何描述 x 和 y 之间的这种线性关系的强弱? 例如, 物理成绩与数学成绩正相关, 但数学成绩能够在多大程度上决定物理成绩? 这就是相关强弱的问题. 类似的还有吸烟与健康的负相关强度、父母身高与子女升高的正相关强度、农作物的产量与施肥量的正相关强度等.

统计中用相关系数 r 来衡量两个变量之间线性关系的强弱. 若相应于变量 x 的取值 x_i , 变量 y 的观测值为 y_i ($1 \leq i \leq n$), 则两个变量的相关系数的计算公式为

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

不同的相关性可以从散点图上直观地反映出来. 图 1、图 2 反映了变量 x 和 y 之间很强的线性相关关系, 而图 4 中的两个变量的线性相关程度很弱.

对于相关系数 r , 首先值得注意的是它的符号. 当 r 为正时, 表明变量 x 和 y 正相关; 当 r 为负时, 表明变量 x 和 y 负相关. 反映在散点图上, 图 1 中的变量 x 和 y 正相关, 这时的 r 为正, 图 2 中的变量 x 和 y 负相关, 这时的 r 为负.

另一个值得注意的是 r 的大小. 统计学认为, 对于变量 x, y , 如果 $r \in [-1, -0.75]$, 那么负相关很强; 如果 $r \in [0.75, 1]$, 那么正相关很强; 如果 $r \in (-0.75, -0.30]$ 或 $r \in [0.30, 0.75)$, 那么相关性一般; 如果 $r \in [-0.25, 0.25]$, 那么相关性较弱. 反映在散点图上, 图 1 的 $r=0.97$, 这些点有明显的从左下角到右上角沿直线分布趋

计, 得到一个卖出的热饮杯数与当天气温的对比表:

表 2-4

摄氏温度/ $^{\circ}\text{C}$	-5	0	4	7	12	15	19	23	27	31	35
热饮杯数	156	150	132	128	130	116	104	89	93	76	54

- (1) 画出散点图;
- (2) 从散点图中发现气温与热饮销售杯数之间关系的一般规律;
- (3) 求回归方程;
- (4) 如果某天的气温是 2°C , 预测这天卖出的热饮杯数.

解: (1) 散点图如图 2.3-8 所示:

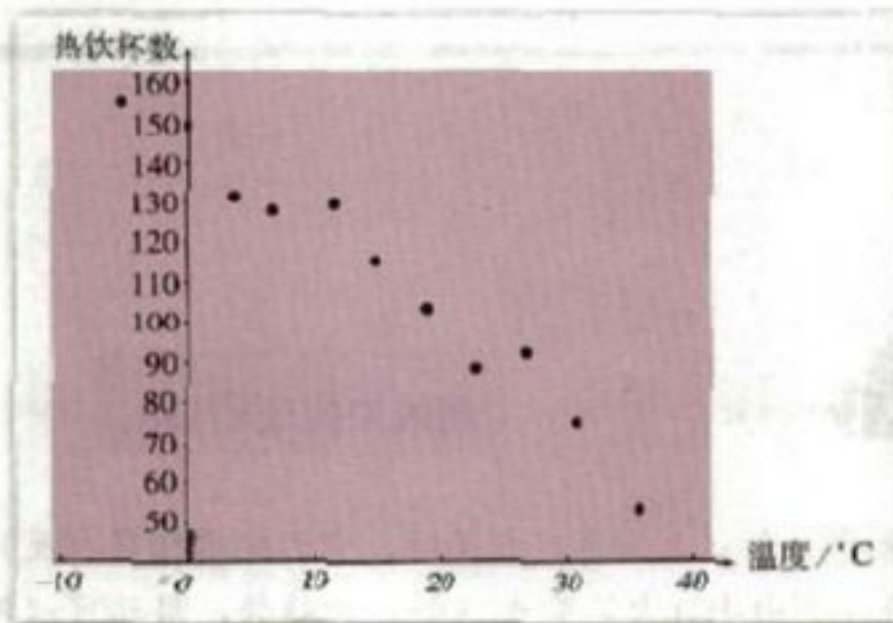


图 2.3-8

(2) 从图 2.3-8 看到, 各点散布在从左上角到右下角的区域里, 因此, 气温与热饮销售杯数之间成负相关, 即气温越高, 卖出去的热饮杯数越少.

(3) 从散点图可以看出, 这些点大致分布在一条直线的附近, 因此, 可用公式②求出回归方程的系数.

利用计算器容易求得回归方程

$$y = -2.352x + 147.767.$$

(4) 当 $x=2$ 时, $y=143.063$. 因此, 某天的气温为 2°C 时, 这天大约可以卖出 143 杯热饮.



气温为 2°C 时, 小卖部一定能够卖出 143 杯左右热饮吗? 为什么?

习题 2.3

A 组

1. “名师出高徒”可以解释为教师的水平越高, 学生的水平也越高. 那么, 教师的水平与学生的水平成什么相关关系? 你能举出更多的描述生活中两个变量的相关关系的成语吗?
2. 有时候, 一些东西吃起来口味越好, 对我们的身体越有害. 下表给出了不同类型的某种食品的数据. 第一列表示此种食品所含热量的百分比, 第二列数据表示由一些美食家以百分制给出的对此种食品口味的评价:

品牌	所含热量的百分比	口味纪录
A	25	89
B	34	89
C	20	80
D	19	78
E	26	75
F	20	71
G	19	65
H	24	62
I	16	60
J	13	52

- (1) 作出这些数据的散点图.
 - (2) 作出回归直线.
 - (3) 关于两个变量之间的关系, 你能得出什么结论?
 - (4) 对于这种食品, 为什么人们更喜欢吃位于回归直线上方的食品而不是下方的?
3. 一个车间为了规定工时定额, 需要确定加工零件所花费的时间, 为此进行了 10 次试验, 收集数据如下:

零件数 x (个)	10	20	30	40	50	60	70	80	90	100
加工时间 y (min)	62	68	75	81	89	95	102	108	115	122

- (1) 画出散点图;
 - (2) 求回归方程;
 - (3) 关于加工零件的个数与加工时间, 你能得出什么结论?
4. 影响消费水平的原因是很多的, 其中重要的一项是工资收入. 研究这两个变量的关系的一个方法是通过随机抽样的方法, 在全国范围内收集被调查者的工资收入和他们的消费状况. 下面的数据来自国家统计局公布的统计年鉴 (2000 年版), 是中国大陆 31 个省、自治区、直辖市的职工平均工资与城镇居民消费水平 (单位: 元).

势, 这时用线性回归模型描述两个变量之间的关系效果很好; 图 2 的 $r=-0.85$, 这些点也有明显的从左上角到右下角沿直线分布趋势, 这时用线性回归模型描述两个变量之间的关系也有好的效果; 图 3 的 $r=0.24$, 这些点的分布几乎没有什么规则, 这时不能用线性回归模型描述这两个变量之间的关系; 图 4 的 $r=-0.05$, 两个变量之间几乎没有什么关系, 这时就更不能用线性回归模型描述两个变量之间的关系。

你能试着对自己身边的某个问题, 确定两个变量, 通过收集数据, 计算相关系数, 然后分析一下能否用线性回归模型来拟合它们之间的关系吗?

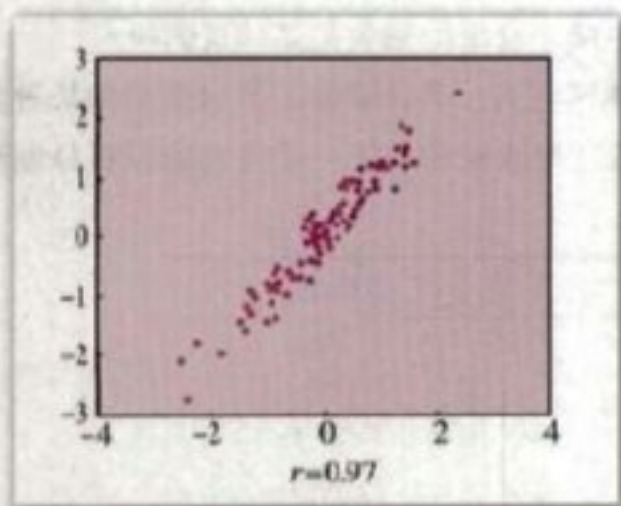


图 1

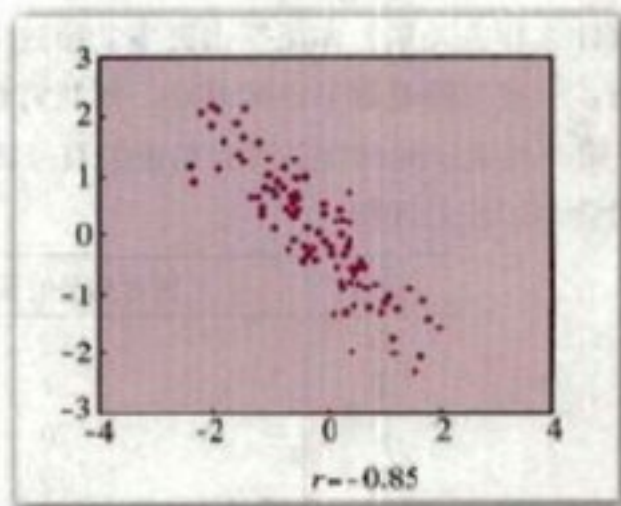


图 2

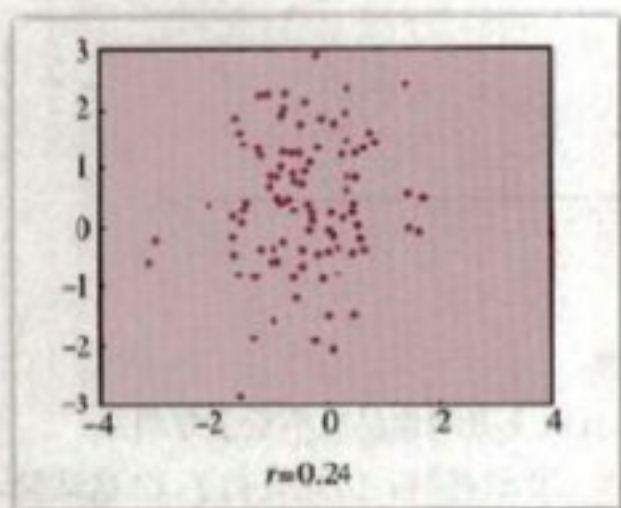


图 3

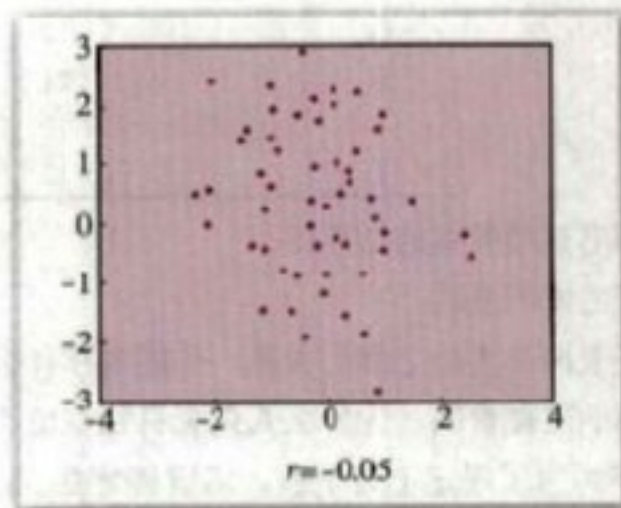


图 4



现实世界的许多问题中都存在相互关联的各种变量,研究这些变量之间的相互关系,能够使我们发现事物发展的一些规律,从而为我们的判断和决策提供依据.下面请同学们根据自己对身边事物的观察和体会,通过查阅资料、讨论等方式,确定要研究的统计问题,然后进行抽样调查,收集数据,并进行整理和分析,最后对问题中的规律作出判断.确定研究问题时,要注意问题的意义.

以下几个问题,供同学们参考.

1. 在校中学生每周使用计算机时间的调查.

(1) 要调查的问题是什么?

(2) 如何设计抽样方案?

(3) 如何分析数据?

(4) 从中能够得出什么规律?

(5) 你能给同学们提出哪些建议?

2. 中学生物理成绩与数学成绩之间的相关关系.

(1) 要研究的问题是什么?

(2) 如何设计抽样方案?

(3) 如何分析数据?

(4) 从中能够得出什么规律?

(5) 你能给同学们提出哪些建议?

下页的“实习报告”供参考.

地区	北京	天津	河北	山西	内蒙古	辽宁	吉林	黑龙江
职工平均工资	14 054	11 056	7 022	6 065	6 347	7 895	7 158	7 094
城镇居民消费水平	7 040	7 346	5 033	3 932	3 765	6 366	5 216	5 632

续表

地区	上海	江苏	浙江	安徽	福建	江西	山东	河南
职工平均工资	16 641	9 171	11 201	6 516	9 490	6 749	7 656	6 494
城镇居民消费水平	11 543	6 239	7 985	4 985	6 255	3 482	6 060	4 214

续表

地区	湖北	湖南	广东	广西	海南	重庆	四川	贵州
职工平均工资	6 991	9 269	12 245	6 776	6 865	7 182	7 249	6 595
城镇居民消费水平	5 295	5 290	8 987	4 987	4 700	6 190	4 876	4 334

续表

地区	云南	西藏	陕西	甘肃	青海	宁夏	新疆
职工平均工资	8 276	12 952	6 931	7 427	9 081	7 392	7 611
城镇居民消费水平	4 933	4 685	4 520	4 615	4 384	3 813	3 988

- (1) 画出散点图；
- (2) 求出回归方程；
- (3) 从回归方程你能得出一些什么结论？

B 组

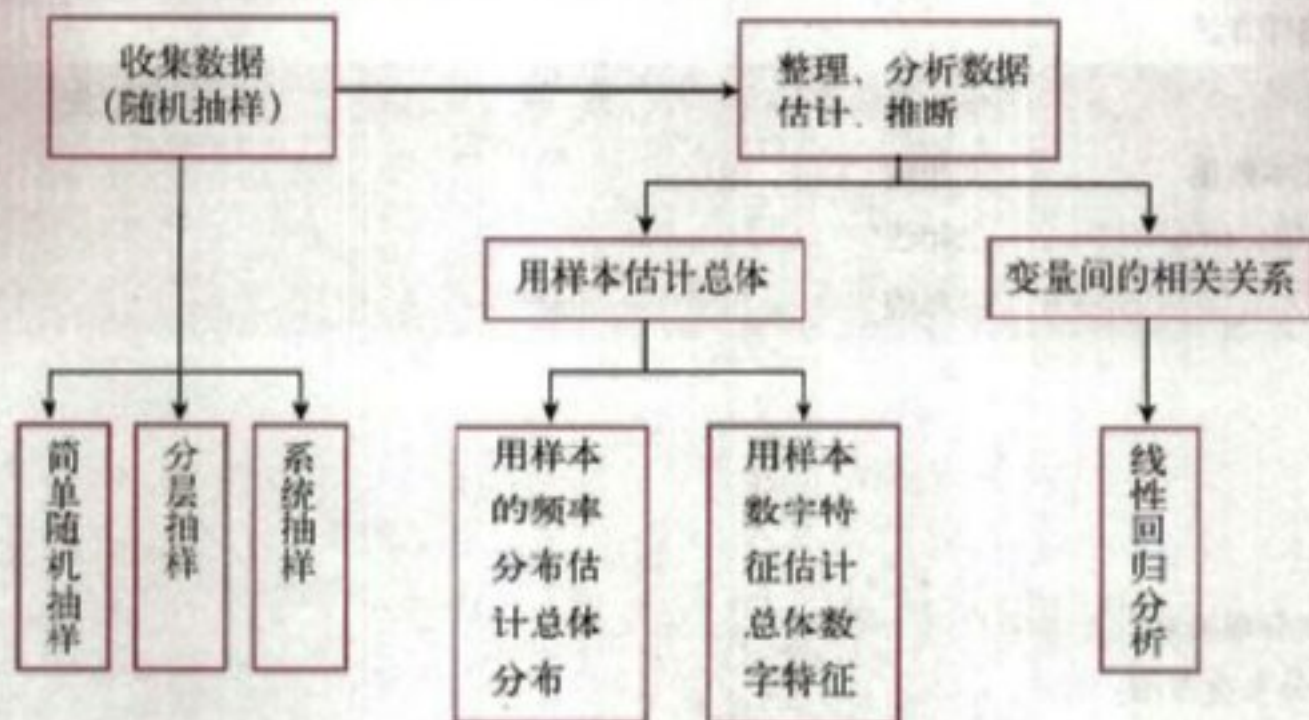
1. 有人收集了 10 年中某城市的居民年收入（即此城市所有居民在一年内的收入的总和）与某种商品的销售额的有关数据：

第 n 年	1	2	3	4	5	6	7	8	9	10
年收入/亿元	32.2	31.1	32.9	35.8	37.1	38.0	39.0	43.0	44.6	46.0
商品销售额/万元	25.0	30.0	34.0	37.0	39.0	41.0	42.0	44.0	48.0	51.0

- (1) 画出散点图；
- (2) 求出回归方程；
- (3) 如果这座城市居民的年收入达到 40 亿元，估计这种商品的销售额是多少。
2. 生活中有许多变量之间的关系是值得我们去研究的。例如，关于我们自己的身体、身高与体重之间是否存在某种相关性呢？请从你自己的班里抽取适当的样本，然后再收集好数据，对它们的相关性进行讨论。

小结

一、本章知识结构



二、回顾与思考

1. 统计与现实生活的联系非常密切.

对于现实中的一些随机现象,通过收集大量数据,再通过一定的统计分析来发现随机现象中的规律性,这就是统计这门学科的主要任务.因此,我们应当逐步学会从现实生活或其他学科中提出有意义的统计问题.

你能从你自己的学习、生活中提出一些统计问题吗?为什么你认为这些问题是有意义的?

2. 抽样调查是收集数据的主要方式.

(1) 在抽取样本的过程中,考虑的最主要的原则是什么?

(2) 本章介绍的三种随机抽样方法,它们有什么联系与区别?它们各自的特点和适用范围是什么?你能给国家统计局的城乡调查队设计一个调查全国公众当前最关心的十大问题的抽样方案吗?

3. 用样本估计总体是统计的基本思想.

一般有用样本的频率分布估计总体分布,以及用样本的特征数估计总体的特征数两类估计.

(1) 现实生活中的许多总体的分布我们并不知道,例如,全国所有高一年级学生的身高作为一个总体,它的分布情况我们是无法准确得知的,那么,通过对全国所有高一

实 习 报 告

年 月 日

题目	本校学生每周使用计算机时间的调查		
抽样方法			
样本数据 (单位: min)	年 级	男 生	女 生
	一年级		
	二年级		
	三年级		
频率分布表和 频率分布直方图			
计算结果	男生 $\bar{x}_1 = \underline{\hspace{2cm}}, s_1 = \underline{\hspace{2cm}};$ 女生 $\bar{x}_2 = \underline{\hspace{2cm}}, s_2 = \underline{\hspace{2cm}};$ 男女生全体 $\bar{x} = \underline{\hspace{2cm}}.$		
结果分析与建议			

复习参考题

A 组

1. 选择题

为了了解某地参加计算机水平测试的 5 000 名学生的成绩，从中抽取了 200 名学生的成绩进行统计分析，在这个问题中，5 000 名学生成绩的全体是（ ）

- (A) 总体. (B) 个体.
(C) 从总体中抽取的一个样本. (D) 样本的容量.

2. 填空题

(1) 在已分组的若干数据中，每组的频数是指_____，每组的频率是指_____.

(2) 一个公司共有 N 名员工，下设一些部门，要采用等比例分层抽样的方法从全体员工中抽取样本容量为 n 的样本，已知某部门有 m 名员工，那么从该部门抽取的员工人数是_____.

3. 在 2002 年春季，一家著名的全国性连锁服装店进行了一项关于当年秋季服装流行色的民意调查，调查者通过向顾客发放饮料，并让顾客通过挑选饮料杯上印着的颜色来对自己喜欢的服装颜色“投票”。根据这次调查结果，在某大城市 A，服装颜色的众数是红色，而当年全国服装协会发布的是咖啡色.

(1) 这个结果是否代表 A 城市的人的想法？

(2) 你认为这两种调查的差异是由什么引起的？

4. 如果调查目的是要确定被调查者的收入水平，请设计一种提问方法.

5. 从一本英语书中随机抽取 100 个句子，数出每个句子中单词数，作出这 100 个数据的频率分布表，由此你可以作出什么估计？

6. 在一个文艺比赛中，12 名专业人士和 12 名观众代表各组成一个评判小组，给参赛选手打分，下面是两个评判组对同一名选手的打分：

小组 A 42 45 48 46 52 47 49 55 42 51 47 45

小组 B 55 36 70 66 75 49 46 68 42 62 58 47

(1) 解释如何衡量每一组成员的相似性.

(2) 对每一组计算这种相似性的度量值，你能据此判断小组 A 与小组 B 哪一个更像是由专业人士组成的吗？

7. 16 种食品所含的热量值如下：

111 123 123 164 430 190 175 236

430 320 250 280 160 150 210 123

(1) 求数据的中位数与平均数；

(2) 用这两种数字特征中的哪一种来描述这个数据集更合适？

8. 改革开放以来,我国高等教育事业有了迅速发展.这里我们得到了某省从1990~2000年18~24岁的青年人每年考入大学的百分比.我们把农村、县镇和城市分开统计.为了便于计算,把1990年编号为0,1991年编号为1……2000年编号为10.如果把每年考入大学的百分比作为因变量,把年份从0到10作为自变量进行回归分析,可得到下面三条回归直线:

$$\text{城市 } \hat{y}=2.84x+9.50;$$

$$\text{县镇 } \hat{y}=2.32x+6.76;$$

$$\text{农村 } \hat{y}=0.42x+1.80.$$

- (1) 在同一个坐标系内作出三条回归直线.
- (2) 对于农村青年来讲,系数等于0.42意味着什么?
- (3) 在这一阶段,三个组哪一个的大学入学率年增长最快?
- (4) 请查阅我国人口分布的有关资料,选择一个在高等教育发展上有代表性的省,以这个省的大学入学率作为样本,说明我国在1991~2000年10年间大学入学率的总体发展情况.



1. 在一家保险公司的董事会上,董事们就我国加入世界贸易组织(WTO)后公司的发展战略问题展开激烈讨论,其中一个议题是如何借鉴国外保险公司先进的管理经验,改进公司的管理模式.会议决定对推销员实行目标管理,即给推销员确定一个具体的销售目标.这个销售目标确定的是否合适,直接影响公司的经济效益.如果目标过高,多数推销员完不成任务,会使推销员失去信心;如果目标定得太低,将不利于挖掘推销员的工作潜力.下面一组数据是部分推销员的月销售额(单位:千元),

19.58	16.11	16.45	20.45	20.24	21.66	22.45	18.22	12.34
19.35	20.55	17.45	18.78	17.96	19.91	18.12	14.65	14.78
16.78	18.78	18.29	18.51	17.86	19.58	19.21	18.55	16.34
15.54	17.55	14.89	18.94	17.43	17.14	18.02	19.98	17.88
17.32	19.35	15.45	19.58	13.45	21.34	14.00	18.42	23.00
17.52	18.51	17.16	24.56	25.14				

请根据这组样本数据提出使75%的职工能够完成销售指标的建议.

2. 想像一下一个人从出生到死亡,在每个生日都测量身高,并作出这些数据散点图,这些点将不会落在一条直线上.但在一段时间内的增长数据有时可以用线性回归来分析.下表是一位母亲给儿子作的成长记录:

年龄/周岁	3	4	5	6	7	8	9	10
身高/cm	90.8	97.6	104.2	110.9	115.4	122.0	128.5	134.2
年龄/周岁	11	12	13	14	15	16		
身高/cm	140.8	147.6	154.2	160.9	167.4	173.0		

- (1) 作出这些数据的散点图.
- (2) 求出这些数据的回归方程.
- (3) 对于这个例子, 你如何解释斜率的含义?
- (4) 用下一年的身高减去当年的身高, 计算每年身高的增长数, 并计算从 3~16 岁身高的年均增长数.
- (5) 解释一下斜率与每年平均增长的身高之间的联系.



美国国家健康与营养调查 (NHANES) 是一项由美国国立卫生研究院 (NIH) 和美国疾病控制中心 (CDC) 联合进行的全国性调查。该调查旨在了解美国成年人的健康状况、营养状况、生活方式以及社会经济因素对健康的影响。调查数据被广泛用于公共卫生研究、政策制定以及临床实践。在 NHANES 中，身高是一个重要的测量指标，用于评估个体的生长发育状况和营养水平。通过长期跟踪调查，研究人员可以观察到身高随时间的变化，从而研究生长激素分泌、营养摄入以及遗传因素对身高发育的影响。

24.11	52.40	18.29	16.0	2.87	1.27	10.41	24.11
25.11	54.40	19.29	17.0	2.90	1.34	11.35	25.11
26.11	56.40	20.29	18.0	2.93	1.41	12.29	26.11
27.11	58.40	21.29	19.0	2.96	1.48	13.23	27.11
28.11	60.40	22.29	20.0	2.99	1.55	14.17	28.11
29.11	62.40	23.29	21.0	3.02	1.62	15.11	29.11

在 NHANES 中，身高数据通常以厘米 (cm) 为单位记录。为了便于分析和比较，数据通常会被转换为英寸 (in) 和磅 (lb) 单位。例如，身高 170 厘米可以转换为 67 英寸，体重 70 公斤可以转换为 154 磅。这种转换有助于研究人员在不同国家和不同研究之间进行数据整合和比较。此外，身高数据还可以用于计算身体质量指数 (BMI)，这是一个常用的指标，用于评估个体的体重是否处于健康范围。通过结合身高和体重数据，研究人员可以更全面地了解个体的健康状况和营养状况。

Year	Age	Height (cm)	Weight (kg)	BMI	Gender	Race
1999	12	150	40	17.8	F	W
2000	13	155	45	18.7	F	W
2001	14	160	50	19.6	F	W
2002	15	165	55	20.5	F	W
2003	16	170	60	21.4	F	W

附表

随 机 数 表

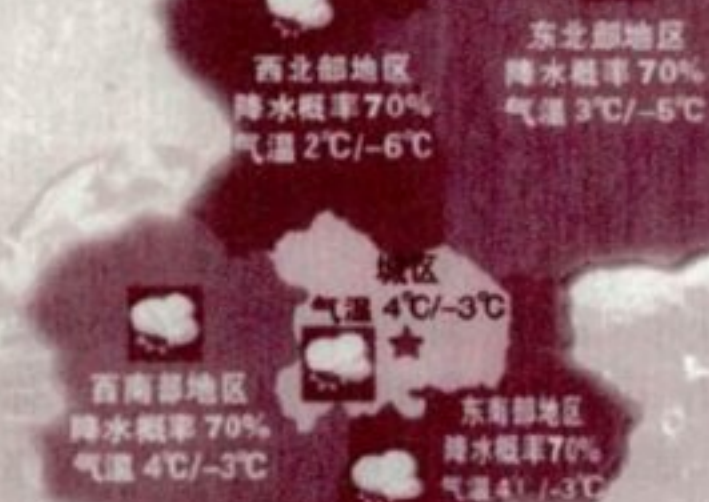
03 47 43 73 86	36 96 47 36 61	46 98 63 71 62	33 26 16 80 45	60 11 14 10 95
97 74 24 67 62	42 81 14 57 20	42 53 32 37 32	27 07 36 07 51	24 51 79 89 73
16 76 62 27 66	56 50 26 71 07	32 90 79 78 53	13 55 38 58 59	88 97 54 14 10
12 56 85 99 26	96 96 68 27 31	05 03 72 93 15	57 12 10 14 21	88 26 49 81 76
55 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 05	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 61 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

第三章

概率



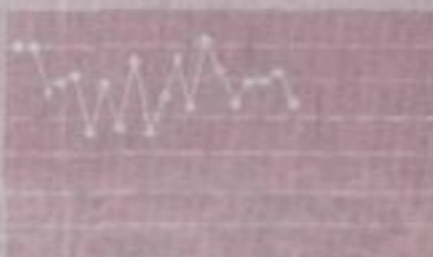
3.1 随机事件的概率

3.2 古典概型

3.3 几何概型

日常生活中,经常会遇到一些无法事先预测结果的事情,它们被称为随机事件.例如,抛掷一枚硬币,它将正面朝上还是反面朝上;明天早上到校的准确时间是几点;购买本期福利彩票是否能够中奖……这些事情的结果都有不确定性,是无法预知的.但当我们把随机的事件放在一起时,它们可能会表现出令人惊奇的规律性.例如,如果你将同样的硬币抛掷100次,尽管事先不能准确预知结果,但由于我们知道正面朝上与反面朝上的可能性各占50%,因此它将差不多50次正面朝上,50次反面朝上.为了研究这种随机事件的规律性,数学中引进了概率.

概率是描述随机事件发生可能性大小的度量,它已经渗透到人们的日常生活中,成为一个常用词汇.概率的准确含义是什么呢?用什么样的方法来计算随机事件的概率?本章我们就来探讨与概率相关的一些基本概念和研究方法.



3.1

随机事件的概率

3.1.1 随机事件的概率

日常生活中，有些问题是很难给予准确无误的回答的。例如，你明天什么时间起床？7：20在某公共汽车站候车的人有多少？12：10在学校食堂用餐的人有多少？你购买的本期福利彩票是否能中奖？等等。显然，这些问题的结果都是不确定的、偶然的，很难给予准确的回答。

客观世界中，有些事情的发生是偶然的，有些事情的发生是必然的，而且偶然与必然之间往往有某种内在联系。例如，北京地区一年四季的变化有着确定的、必然的规律，但北京地区一年里哪一天最热，哪一天最冷，哪一天降雨量最大，哪一天降雪量最大等又是不确定的、偶然的。又如，一方面，某种水稻种子发芽后，在一定的条件（温度、水分、土壤、阳光）下，一定会经历分蘖、生长、颖花、结穗、成熟等过程，这个生长规律是确定的；另一方面，在这个过程中，每一粒发芽种子的分蘖数是多少，结穗率是多少，茎秆高是多少，结穗实粒有多少，不实率是多少，粒重又是多少，这些却都是不确定的。农业生产实践告诉我们，在一定的条件 S （温度、水分、土壤、阳光）下，发芽种子一定会分蘖。像这种在一定的条件 S （温度、水分、土壤、阳光）下，必然会发生的（发芽种子的分蘖）称为必然事件。但是，在一定的条件 S （温度、水分、土壤、阳光）下，一粒发芽种子会分多少蘖，是1支、2支，还是3支……这些又是不确定的，像这种在条件 S 下，不能事先预测结果的事件称为随机事件。另外，“发芽的种子不分蘖”这一事件一定不会发生，像这种在条件 S 下一定不会发生的事件称为不可能事件。

一般地，我们把在条件 S 下，一定会发生的事件，叫做相对于条件 S 的**必然事件**（certain event），简称必然事件；

在条件 S 下，一定不会发生的事件，叫做相对于条件 S 的**不可能事件**（impossible event），简称不可能事件；

必然事件与不可能事件统称为相对于条件 S 的**确定事件**，简称确定事件。

在条件 S 下可能发生也可能不发生的事件，叫做相对于条件 S 的**随机事件**（random event），简称随机事件。

确定事件和随机事件统称为事件，一般用大写字母 A, B, C, \dots 表示。



你能举出一些现实生活中的随机事件、必然事件、不可能事件的实例吗？

对于随机事件，知道它发生的可能性大小是非常重要的。用**概率** (probability) 度量随机事件发生的可能性大小能为我们的决策提供关键性的依据。那么，如何才能获得随机事件发生的概率呢？最直接的方法就是试验（观察）。

下面我们来做一个抛掷一枚硬币的试验，观察它落地时哪一个面朝上。

第一步，全班每人各取一枚同样的硬币，做 10 次掷硬币的试验，每人记录下试验结果，填在下表中：

姓名	试验次数	正面朝上的次数	正面朝上的比例

物体的大小用质量多少、体积大小等来度量，随机事件发生可能性的大小用概率来度量。概率是客观存在的。



与其他同学的试验结果比较，你的结果和他们一致吗？为什么会出现这样的情况？

第二步，每个小组把本组同学的试验结果统计一下，填入下表：

组次	试验总次数	正面朝上的总次数	正面朝上的比例



与其他小组的试验结果比较，各组的结果一致吗？为什么？

第三步，请一个同学把全班同学的试验结果统计一下，填入下表：

班级	试验总次数	正面朝上的总次数	正面朝上的比例

第四步, 请把全班每个同学的试验中正面朝上的次数收集起来, 并用条形图表示.



这个条形图有什么特点?

第五步, 请同学们找出掷硬币时“正面朝上”这个事件发生的规律性.



如果同学们再重复一次上面的试验, 全班的汇总结果还会和这次的汇总结果一致吗? 如果不一致, 你能说出原因吗?

在相同的条件 S 下重复 n 次试验, 观察某一事件 A 是否出现, 称 n 次试验中事件 A 出现的次数 n_A 为事件 A 出现的**频数** (frequency), 称事件 A 出现的比例 $f_n(A) = \frac{n_A}{n}$ 为事件 A 出现的**频率** (relative frequency).



频率的取值范围是什么?

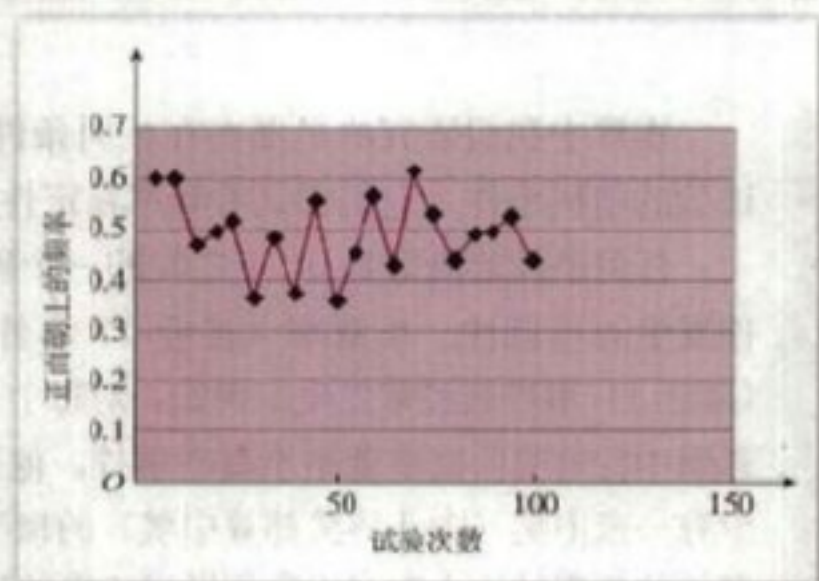
必然事件出现的频率为 1, 不可能事件出现的频率为 0.

可以用计算机模拟掷硬币试验, 以下是一次计算机模拟掷硬币的试验.



表 3-1 计算机模拟掷硬币的试验结果

试验次数	正面朝上的频数	正面朝上的频率
5	4	0.8
10	6	0.6
15	6	0.4
20	14	0.7
25	11	0.44
30	16	0.533 333
35	18	0.514 286
40	20	0.5
45	20	0.444 444
50	20	0.4
55	26	0.472 727
60	31	0.516 667
65	30	0.461 538
70	35	0.5
75	34	0.453 333
80	38	0.475
85	43	0.505 882
90	46	0.511 111
95	56	0.589 474
100	53	0.53



掷硬币的频率图

历史上有人曾经做过大量重复掷硬币的试验, 结果如表 3-2 所示。

表 3-2 历史上一些掷硬币的试验结果

试验次数	正面朝上的频数	正面朝上的频率
2 048	1 061	0.518 1
4 040	2 048	0.506 9
12 000	6 019	0.501 6
24 000	12 012	0.500 5
30 000	14 984	0.499 6
72 088	36 124	0.501 1

我们看到,当试验次数很多时,出现正面的频率值在 0.5 附近摆动.一般来说,随机事件 A 在每次试验中是否发生是不能预知的,但是在大量重复试验后,随着试验次数的增加,事件 A 发生的频率会逐渐稳定在区间 $[0, 1]$ 中的某个常数上.这个常数越接近于 1,表明事件 A 发生的频率越大,频数就越多,也就是它发生的可能性越大;反过来,事件发生的可能性越小,频数就越少,频率就越小,这个常数也就越小.因此,我们可以用这个常数来度量事件 A 发生的可能性的大小.

对于给定的随机事件 A ,由于事件 A 发生的频率 $f_n(A)$ 随着试验次数的增加稳定于概率 $P(A)$,因此可以用频率 $f_n(A)$ 来估计概率 $P(A)$.

这样,抛掷一枚硬币,正面朝上的概率为 0.5,即

$$P(\text{正面朝上}) = 0.5.$$



事件 A 发生的频率 $f_n(A)$ 是不是不变的? 事件 A 的概率 $P(A)$ 是不是不变的? 它们之间有什么区别与联系?



雅各布·贝努利
(Jacob Bernoulli, 1654—1705) 瑞士数学家,被公认为概率理论的先驱.他给出了著名的大数定律.大数定律阐述了随着试验次数的增加,频率稳定在概率附近.

本章中我们研究的是那些在相同条件下可以进行大量重复试验的随机事件,它们都具有频率稳定性.

任何事件的概率是 0~1 之间的一个确定的数,它度量该事件发生的可能性.小概率(接近 0)事件很少发生,而大概率(接近 1)事件则经常发生.例如,对每个人来讲,他买一张体育彩票中特等奖的概率就是小概率事件,他买 10 000 张体育彩票至少有一张中奖(中几等奖都算中奖)的概率是很大的.知道随机事件的概率的大小有利于我们做出正确的决策.

练习

1. 做同时掷两枚硬币的试验, 观察试验结果.

(1) 试验可能出现的结果有几种? 分别把它们表示出来.

(2) 做 100 次试验, 每种结果出现的频数、频率各是多少?

与其他几名同学的试验结果汇总, 你会发现什么? 你能估计每种结果出现的概率吗?

2. 做掷骰子试验, 掷一个骰子 100 次, 并填下表:

	频数	频率
试验的总次数	100	
出现数字 1		
出现数字 5		
出现的数字小于 7		
出现的数字大于 7		
出现的数字为偶数		
出现的数字为奇数		

3. (1) 给出一个概率很小的随机事件的例子;

(2) 给出一个概率很大的随机事件的例子.

3.1.2 概率的意义

1. 概率的正确理解



有人说, 既然抛掷一枚硬币出现正面的概率为 0.5, 那么连续两次抛掷一枚质地均匀的硬币, 一定是一次正面朝上, 一次反面朝上, 你认为这种想法正确吗?

尽管每次抛掷硬币的结果出现正、反的概率都是 0.5, 但连续两次抛掷硬币的结果不一定恰好是正面朝上、反面朝上各一次. 每个同学都连续抛掷两次硬币, 统计全班同学的试验结果, 可以发现有三种可能的结果: “两次正面朝上” “两次反面朝上” “一次正面朝上, 一次反面朝上”. 这正体现了随机事件发生的随机性.



探究

全班同学各取一枚同样的硬币（一角、一元等），连续两次抛掷，观察它落地后的朝向，并记录结果，重复上面的过程10次，将全班同学的试验结果汇总，计算三种结果发生的频率，你有什么发现？

随着试验次数的增加，可以发现，“正面朝上、反面朝上各一次”的频率与“两次均正面朝上”“两次均反面朝上”的频率是不一样的，而且“两次均正面朝上”的频率与“两次均反面朝上”的频率大致相等；“正面朝上、反面朝上各一次”的频率大于“两次均正面朝上”（“两次均反面朝上”）的频率，事实上，“两次均正面朝上”的概率为0.25，“两次均反面朝上”的概率也为0.25，“正面朝上、反面朝上各一次”的概率是0.5。

上述试验告诉我们，随机事件在一次试验中发生与否是随机的，但随机性中含有规律性，认识了这种随机性中的规律性，就能使我们比较准确地预测随机事件发生的可能性。例如，做连续抛掷两枚硬币的试验100次，可以预见：“两个正面朝上”大约出现25次；“两个反面朝上”大约出现25次；“正面朝上、反面朝上各一个”大约出现50次，出现“正面朝上、反面朝上各一个”的机会比出现“两个正面朝上”或“两个反面朝上”的机会大。



思考

如果某种彩票的中奖概率为 $\frac{1}{1000}$ ，那么买1000张这种彩票一定能中奖吗？（假设该彩票有足够多的张数。）

同学们可以把同样大小的9个白色乒乓球和1个黄色乒乓球放在1个袋中，每次摸出1球后再放回袋中，这样摸10次，观察是否一定至少有1次摸到黄球。



有的同学可能认为，中奖概率为 $\frac{1}{1000}$ ，那么买1000张彩票就一定能中奖，但这种想法是不正确的。

实际上，买1000张彩票相当于做1000次试验，因为每次试验的结果都是随机的，所以做1000次的结果也是随机的，这就是说，每张彩票既可能中奖也可能不中奖，因此1000张彩票中可能

没有一张中奖,也可能有一张、两张……中奖.

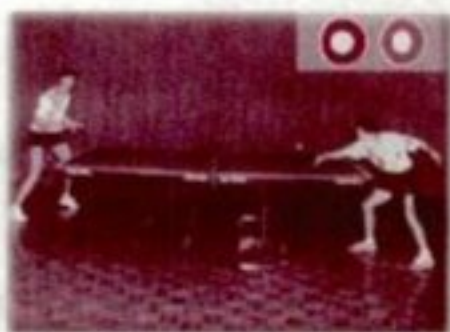
虽然中奖张数是随机的,但这种随机性中具有规律性.随着试验次数的增加,即随着所买彩票张数的增加,其中中奖彩票所占的比例可能越接近于 $\frac{1}{1\,000}$.

实际上,买1 000张彩票中奖的概率约为0.632,没有一张中奖也是有可能的,其概率近似为0.368.

2. 游戏的公平性

在一场乒乓球比赛前,要决定由谁先发球,你注意到裁判是怎样确定发球权的吗?

下面就是常用的一种方法:裁判员拿出一个抽签器,它是一个像大硬币似的均匀塑料圆板,一面是红圈,一面是绿圈,然后随意指定一名运动员,要他猜上抛的抽签器落到球台上时,是红圈那面朝上还是绿圈那面朝上.如果他猜对了,就由他先发球,否则,由另一方先发球.为什么要这样做呢?



这样做体现了公平性,它使得两名运动员的先发球机会是等可能的.用概率的语言描述,就是两个运动员取得发球权的概率都是0.5.这是因为抽签器上抛后,红圈朝上与绿圈朝上的概率都是0.5,因此任何一名运动员猜中的概率都是0.5,也就是每个运动员取得先发球权的概率均为0.5,所以这个规则是公平的.

探究

某中学高一年级有12个班,要从中选2个班代表学校参加某项活动.由于某种原因,一班必须参加,另外再从二至十二班中选1个班.有人提议用如下的方法:掷两个骰子得到的点数和是几,就选几班,你认为这种方法公平吗?

	1点	2点	3点	4点	5点	6点
1点	2	3	4	5	6	7
2点	3	4	5	6	7	8
3点	4	5	6	7	8	9
4点	5	6	7	8	9	10
5点	6	7	8	9	10	11
6点	7	8	9	10	11	12

两个骰子的点数和

3. 决策中的概率思想



如果连续 10 次掷一枚骰子，结果都是出现 1 点，你认为这枚骰子的质地均匀吗？为什么？

利用刚学过的概率知识我们可以进行推断，如果它是均匀的，通过试验和观察，可以发现出现各个面的可能性都应该是 $\frac{1}{6}$ ，连续 10 次出现 1 点的概率约为 0.000 000 016 538，这在一次试验（即连续 10 次投掷一枚骰子）中是几乎不可能发生的^①，而当骰子不均匀时，特别是当 6 点的那面比较重时（例如灌了铅或水银），会使出现 1 点的概率最大，更有可能连续 10 次出现 1 点。

现在我们面临两种可能的决策：一种是这枚骰子的质地均匀，另一种是这枚骰子的质地不均匀，当连续 10 次投掷这枚骰子，结果都是出现 1 点，这时我们更愿意接受第二种情况：这枚骰子靠近 6 点的那面比较重，原因是在第二种假设下，更有可能出现 10 个 1 点。

如果我们面临的是从多个可选答案中挑选正确答案的决策任务，那么“使得样本出现的可能性最大”可以作为决策的准则，例如对上述思考题所作的推断，这种判断问题的方法称为**极大似然法**，极大似然法是统计中重要的统计思想方法之一。



①在一次试验中几乎不可能发生的事件称为**小概率事件**。

4. 天气预报的概率解释



某地气象局预报说，明天本地降水概率为 70%，你认为下面两个解释中哪一个能代表气象局的观点？

(1) 明天本地有 70% 的区域下雨，30% 的区域不下雨；

(2) 明天本地下雨的机会是 70%。

(1) 显然是不正确的, 因为 70% 的概率是说降水的概率, 而不是说 70% 的区域降水. 正确的选择是 (2).

生活中, 我们经常听到这样的议论: “天气预报说昨天降水概率为 90%, 结果连一点雨都没下, 天气预报也太不准确了.” 学了概率后, 你能给出解释吗?

天气预报的“降水”是一个随机事件, “概率为 90%”指明了“降水”这个随机事件发生的概率. 我们知道: 在一次试验中, 概率为 90% 的事件也可能不出现. 因此, “昨天没有下雨”并不能说明“昨天的降水概率为 90%”的天气预报是错误的.

你能根据频率与概率的关系, 通过观察收集数据来判断气象局做出的“降水概率为 90%”预报是否正确吗?

5. 试验与发现

奥地利遗传学家孟德尔 1856 年开始用豌豆作试验, 这个试验大约持续了七八年的时间. 他把黄色和绿色的豌豆杂交, 第一年收获的豌豆都是黄色的. 第二年, 当他把第一年收获的黄色豌豆再种下时, 收获的豌豆既有黄色的又有绿色的. 同样他把圆形和皱皮豌豆杂交, 第一年所收获的都是圆形豌豆, 连一粒皱皮豌豆也没有. 第二年, 当他把这种杂交圆形豌豆再种下时, 得到的却既有圆形豌豆, 又有皱皮豌豆. 类似地, 他把长茎的豌豆与短茎的豌豆杂交, 第一年长出来的都是长茎的豌豆, 另外的那种特征则完全消失了. 当他把这种杂交长茎豌豆再种下时, 得到的却既有长茎豌豆, 又有短茎豌豆. 试验的具体数据如下:

表 3-3 豌豆杂交试验的子二代结果

性状	显性		隐性		显性: 隐性
子叶的颜色	黄色	6 022	绿色	2 001	3.01 : 1
种子的性状	圆形	5 474	皱皮	1 850	2.96 : 1
茎的高度	长茎	787	短茎	277	2.84 : 1

为什么表面完全相同的豌豆会长出这样不同的后代呢? 而且每次试验的结果比例如此稳定? 比例都接近 3 : 1. 孟德尔认为其中一定有某种遗传规律. 经过长期的、坚持不懈的研究, 孟德尔终于找到了这种规律, 这一发现为近代遗传学奠定了基础, 孟德尔本人也成了遗传学的奠基人.

天气预报是气象专家根据观测到的气象资料和专家们的实际经验, 经过分析推断得到的.



孟德尔 (Gregor Mendel, 1822—1884), 奥地利遗传学家, 被公认为传统遗传学之父, 1865 年发现了遗传定律.

6. 遗传机理中的统计规律

孟德尔从豌豆实验中洞察到的遗传规律是一种统计规律,下面给出简单的解释.

纯黄色和纯绿色的豌豆均有两个特征(用符号 YY 代表纯黄色豌豆的两个特征,符号 yy 代表纯绿色豌豆的两个特征):

纯黄色的豌豆 YY

纯绿色的豌豆 yy

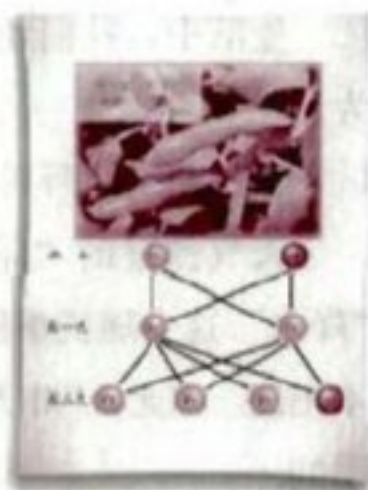
当这两种豌豆杂交时,下一代是从父母辈中各随机地选取一个特征,于是第一年收获的豌豆的特征为:

第一代(第一年收获的豌豆) Yy

当把第一代杂交豌豆再种下时,下一代同样是从父母辈中各随机地选取一个特征,所以第二代的豌豆的特征如下:

第二代(第二年收获的豌豆) YY, Yy, yy

这里对于豌豆的颜色来说, Y 是显性因子, y 是隐性因子. 当显性因子与隐性因子组合时,表现显性因子的特性,即 YY, Yy 都呈黄色;当两个隐性因子组合时才表现隐性因子的特性,即 yy 呈绿色. 由于下一代的两个特征是从父母辈中各随机选取的,因此在第二代中 YY, yy 出现的概率都是 $\frac{1}{4}$, Yy 出现的概率是 $\frac{1}{2}$, 所以黄色豌豆 (YY, Yy): 绿色豌豆 (yy) $\approx 3:1$. 这与连续掷一枚硬币的试验相同,两次均出现反面的概率为 $\frac{1}{4}$, 至少出现一次正面的概率为 $\frac{3}{4}$. 在多次试验中,至少出现一次正面的次数:两次均出现反面的次数 $\approx 3:1$.



练习

1. 在网上或报纸中找出使用概率的例子,并说明这个概率是如何被使用的.
2. 在乒乓球、排球等比赛中,裁判员还用哪些方法决定谁先发球? 这些方法公平吗?
3. “一个骰子掷一次得到 2 的概率是 $\frac{1}{6}$, 这说明一个骰子掷 6 次会出现一次 2”, 这种说法对吗?

说说你的理由.

3.1.3 概率的基本性质

探究

在掷骰子试验中, 可以定义许多事件, 例如:

$C_1 = \{\text{出现 1 点}\}$; $C_2 = \{\text{出现 2 点}\}$; $C_3 = \{\text{出现 3 点}\}$;

$C_4 = \{\text{出现 4 点}\}$; $C_5 = \{\text{出现 5 点}\}$; $C_6 = \{\text{出现 6 点}\}$;

$D_1 = \{\text{出现的点数不大于 1}\}$; $D_2 = \{\text{出现的点数大于 3}\}$; $D_3 = \{\text{出现的点数小于 5}\}$;

$E = \{\text{出现的点数小于 7}\}$; $F = \{\text{出现的点数大于 6}\}$;

$G = \{\text{出现的点数为偶数}\}$; $H = \{\text{出现的点数为奇数}\}$

.....

你能写出这个试验中出现的其他一些事件吗? 类比集合与集合的关系、运算, 你能发现它们之间的关系与运算吗?

1. 事件的关系与运算

(1) 显然, 如果事件 C_1 发生, 则事件 H 一定发生, 这时我们说事件 H 包含事件 C_1 , 记作 $H \supseteq C_1$.

一般地, 对于事件 A 与事件 B , 如果事件 A 发生, 则事件 B 一定发生, 这时称**事件 B 包含事件 A** (或称**事件 A 包含于事件 B**), 记作 $B \supseteq A$ (或 $A \subseteq B$). 与集合类比, 可用图 3.1-1 表示. 不可能事件记作 \emptyset , 任何事件都包含不可能事件.

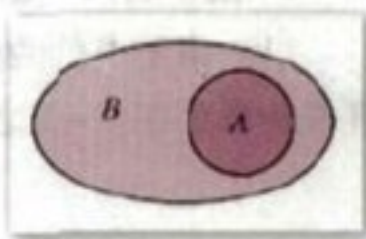


图 3.1-1

(2) 如果事件 C_1 发生, 那么事件 D_1 一定发生, 反过来也对, 这时我们说这两个事件相等, 记作 $C_1 = D_1$.

一般地, 若 $B \supseteq A$, 且 $A \supseteq B$, 那么称事件 A 与事件 B 相等, 记作 $A = B$.

(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**并事件** (或**和事件**), 记作 $A \cup B$ (或 $A + B$).

例如, 在掷骰子的试验中, 事件 $C_1 \cup C_5$ 表示出现 1 点或 5 点这个事件, 即 $C_1 \cup C_5 = \{\text{出现 1 点或 5 点}\}$.

(4) 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**交事件** (或**积事件**), 记作 $A \cap B$ (或 AB), 如图 3.1-2.

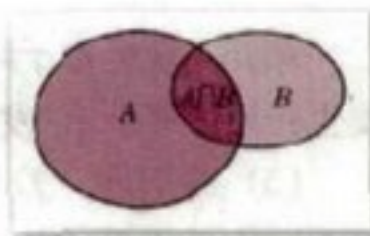


图 3.1-2

例如, 在掷骰子的试验中, $D_2 \cap D_3 = C_4$.

(5) 若 $A \cap B$ 为不可能事件 ($A \cap B = \emptyset$), 那么称**事件 A 与事**

件B互斥，其含义是：事件A与事件B在任何一次试验中不会同时发生，如图3.1-3.

例如，上述试验中的事件 C_1 与事件 C_2 互斥，事件G与事件H互斥.

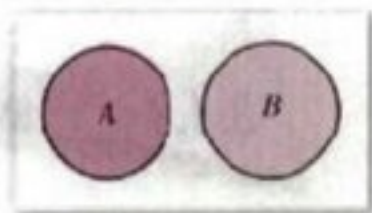


图 3.1-3

(6) 若 $A \cap B$ 为不可能事件， $A \cup B$ 为必然事件，那么称事件A与事件B**互为对立事件**，其含义是：事件A与事件B在任何一次试验中有且仅有一个发生.

例如，在掷骰子试验中， $G \cap H$ 为不可能事件， $G \cup H$ 为必然事件，所以G与H互为对立事件.



事件的关系、运算与集合的关系、运算十分类似，在它们之间可以建立一个对应关系，如事件A与B之并对应于两个集合的并 $A \cup B$ ，事件A与B之交对应于两个集合的交 $A \cap B$ ……因此，可以从集合的观点来看待事件，请同学们找出事件与集合之间的其他对应关系.

2. 概率的几个基本性质

(1) 由于事件的频数总是小于或等于试验的次数，所以频率在0~1之间，从而任何事件的概率在0~1之间，即

$$0 \leq P(A) \leq 1.$$

(2) 在每次试验中，必然事件一定发生，因此它的频率为1，从而必然事件的概率为1. 例如，在掷骰子试验中，由于出现的点数最大是6，因此 $P(E)=1$.

(3) 在每次试验中，不可能事件一定不出现，因此它的频率为0，从而不可能事件的概率为0. 例如，在掷骰子试验中， $P(F)=0$.

(4) 当事件A与事件B互斥时， $A \cup B$ 发生的频数等于A发生的频数与B发生的频数之和，从而 $A \cup B$ 的频率 $f_n(A \cup B) = f_n(A) + f_n(B)$.

由此得到**概率的加法公式**：

$$\begin{aligned} &\text{如果事件A与事件B互斥，则} \\ &P(A \cup B) = P(A) + P(B). \end{aligned}$$

例如，在掷骰子时，由于在一次试验中事件 C_1 与事件 C_2 不会同时发生，因此， $C_1 \cup C_2$ 发生的频数等于 C_1 发生的频数与 C_2 发生的频数之和， $P(C_1 \cup C_2) = P(C_1) + P(C_2)$.

(5) 特别地，若事件B与事件A互为对立事件，则 $A \cup B$ 为必然事件， $P(A \cup B) = 1$. 再由加法公式得 $P(A) = 1 - P(B)$. 例如，在掷骰子试验中，G与H互为对立事件，因此 $P(G) = 1 - P(H)$.

利用上述概率性质, 可以简化概率的计算.

例 如果从不包括大小王的 52 张扑克牌中随机抽取一张, 那么取到红心 (事件 A) 的概率是 $\frac{1}{4}$, 取到方片 (事件 B) 的概率是 $\frac{1}{4}$. 问:

(1) 取到红色牌 (事件 C) 的概率是多少?

(2) 取到黑色牌 (事件 D) 的概率是多少?

解: (1) 因为 $C=A \cup B$, 且 A 与 B 不会同时发生, 所以 A 与 B 是互斥事件. 根据概率的加法公式, 得

$$P(C) = P(A) + P(B) = \frac{1}{2}.$$

(2) C 与 D 也是互斥事件, 又由于 $C \cup D$ 为必然事件, 所以 C 与 D 互为对立事件, 所以

$$P(D) = 1 - P(C) = \frac{1}{2}.$$



练习

- 如果某人在某种比赛 (这种比赛不会出现“和”的情况) 中获胜的概率是 0.3, 那么他输的概率是多少?
- 利用简单随机抽样的方法抽查了某校 200 名学生, 其中戴眼镜的学生有 123 人, 若在这个学校随机调查一名学生, 问他戴眼镜的概率近似值是多少?
- 某工厂为了节约用电, 规定每天的用电量指标为 1 000 千瓦时, 按照上个月的用电记录, 30 天中有 12 天的用电量超过指标, 若第二个月仍没有具体的节电措施, 试求该月的第一天用电量超过指标的概率近似值.
- 一个人打靶时连续射击两次, 事件“至少有一次中靶”的互斥事件是 ()
 (A) 至多有一次中靶. (B) 两次都中靶.
 (C) 只有一次中靶. (D) 两次都不中靶.
- 把红、蓝、黑、白 4 张纸牌随机分给甲、乙、丙、丁 4 个人, 每人分得一张, 事件“甲分得红牌”与事件“乙分得红牌”是 ()
 (A) 对立事件. (B) 互斥但不对立事件.
 (C) 不可能事件. (D) 以上都不对.



天气变化的认识过程

在人类与大自然的较量中,经常面对影响人类生存、反复无常的天气变化,人们对这种随机现象的认识,经历了神化、经验预报、利用现代科学技术进行预报的阶段.



古代,人类对于支配自己的大自然没有科学的认识,认为神灵主宰着天气的变化.中国民间认为,玉皇、雷公、风伯、雨师、龙王等神灵主司刮风下雨、雷鸣电闪.为求风调雨顺,我们的祖先想尽各种办法求助于神灵.现在看来,求雨活动是无知和迷信的,但可以理解的是,这是人们在对大自然束手无策时所做的事情.人类不甘于受自然支配的本能是人类认识与改造世界的动力.

人类对天气变化经历了漫长的认识过程,积累了丰富的气象经验,这些经验的获得实际上有意无意地应用了概率与统计方面的知识.千百年来,我国劳动人民在生产实践中根据云的形状、走向、速度、厚度、颜色等的变化,总结了丰富的“看云识天气”的经验,并将这些经验编成谚语,如“天上钩钩云,地上雨淋淋”“炮台云,雨淋淋”“云交云,雨淋淋”……三国时期,诸葛亮曾经运用自身丰富的气象观测经验,提前三天准确地预报出一场大雾,并在大雾的掩护下,不费吹灰之力,演出了一场“草船借箭”的好戏,令世人惊叹.

近代天气预报已有100多年历史,它建立在对遍布世界各地的气象观测台的观测数据的统计与分析的基础上.自从有电报和互联网后,各地同时观测的气象资料能及时集中到各国的气象中心.根据这些天气资料,利用动力学方程可以确定未来一段时间内的气压、温度、湿度、风力等分布的情况,从而对天气进行预报,这称作动力学预报.但是,影响天气变化的因素多种多样,目前的动力学方程只考虑了一些主要因素,因此做出的天气预报还有较大的误差,不能直接用于预报.

统计预报以概率论为基础,其基本思路是:将预报量 P 同其他一些气象要素 (X_1, \dots, X_n) 进行统计分析,建立起回归方程,利用统计决策作出预报.将动力学预报与统计预报相结合,可以提高预报效果.

随着人们对气象研究的深入,人们对影响天气变化的因素的认识会越来越深入,使天气预报的准确率越来越高.

习题 3.1

A 组

- 若 A, B 为互斥事件, 则 ()
 (A) $P(A) + P(B) < 1$, (B) $P(A) + P(B) > 1$.
 (C) $P(A) + P(B) = 1$, (D) $P(A) + P(B) \leq 1$.
- 在一个实验中, 一种血清被注射到 500 只豚鼠体内. 最初, 这些豚鼠中 150 只有圆形细胞, 250 只有椭圆形细胞, 100 只有不规则形状细胞. 被注射这种血清之后, 没有一个具有圆形细胞的豚鼠被感染, 50 个具有椭圆形细胞的豚鼠被感染, 具有不规则形状细胞的豚鼠全部被感染. 根据试验结果, 估计具有下列类型的细胞的豚鼠被这种血清感染的概率:
 (1) 圆形细胞;
 (2) 椭圆形细胞;
 (3) 不规则形状细胞.
- 李老师在某大学连续 3 年主讲经济学院的高等数学, 下表是李老师这门课 3 年来的考试成绩分布:

成 绩	人 数
90 分以上	43
80~89 分	182
70~79 分	260
60~69 分	90
50~59 分	62
50 分以下	8

经济学院一年级的学生王小慧下学期将修李老师的高等数学课, 用已有的信息估计她得以下分数的概率:

- (1) 90 分以上; (2) 60~69 分; (3) 60 分以上.
- 从一本英文 (小说类) 书里随机翻一页, 然后数出在这一页里元音字母的数目和字母的总数, 把数据填入下表:

元音字母	频数	频率
A		
E		
I		
O		
U		
本页字母总数		

利用这个频率预测在另外一页中元音字母 “E” 的数目, 然后数出这页中元音字母 “E” 的数目, 你的预测与实际数目接近吗?

5. 某人捡到不规则形状的五面体石块, 他在每个面上作了记号, 投掷了 100 次, 并且记录了每个面落在桌面上的次数 (如下表). 如果再投掷一次, 请估计石块的第 4 面落在桌面上的概率是多少?

石块的面	1	2	3	4	5
频数	32	18	15	13	22

6. 在一个袋子中放 9 个白球, 1 个红球, 摇匀后随机摸球:

- (1) 每次摸出球后记下球的颜色然后放回袋中;
(2) 每次摸出球后不放回袋中.

在两种情况下分别做 10 次试验, 求每种情况下第 4 次摸到红球的频率. 两个频率相差得远吗? 两个事件的概率一样吗? 第 4 次摸到红球的频率与第 1 次摸到红球的频率相差得远吗? 请说明原因.

B 组

1. 下列说法正确的是 ()

- (A) 事件 A, B 中至少有一个发生的概率一定比 A, B 中恰有一个发生的概率大.
(B) 事件 A, B 同时发生的概率一定比 A, B 中恰有一个发生的概率小.
(C) 互斥事件一定是对立事件, 对立事件不一定是互斥事件.
(D) 互斥事件不一定是对立事件, 对立事件一定是互斥事件.

2. 若 $P(A \cup B) = P(A) + P(B) = 1$, 则事件 A 与 B 的关系是 ()

- (A) 互斥不对立. (B) 对立不互斥.
(C) 互斥且对立. (D) 以上答案都不对.

3. 统计全班同学的生日, 将数据填入下表:

月份	频数	频率
一月		
二月		
三月		
四月		
五月		
六月		
七月		
八月		
九月		
十月		
十一月		
十二月		
合计		

- (1) 全班同学的生日在每个月的频数一样吗?
(2) 收集全年级同学生日的数据, 你能得到一个人在每个月出生是等可能的结论吗?

CHAPTER 3.2

古典概型



通过试验和观察的方法，我们可以得到一些事件的概率估计，但这种方法耗时多，而且得到的仅是概率的近似值。在一些特殊的情况下，我们可以构造出计算事件概率的通用方法。

3.2.1 古典概型

我们再来分析事件的构成。考察两个试验：

- (1) 掷一枚质地均匀的硬币的试验；
- (2) 掷一枚质地均匀的骰子的试验。

在试验(1)中，结果只有两个，即“正面朝上”或“反面朝上”，它们都是随机事件；在试验(2)中，所有可能的试验结果只有6个，即出现“1点”“2点”“3点”“4点”“5点”和“6点”，它们也都是随机事件。我们把这类随机事件称为**基本事件**(elementary event)。

基本事件有如下特点：

- (1) 任何两个基本事件是互斥的；
- (2) 任何事件(除不可能事件)都可以表示成基本事件的和。

在掷硬币试验中，必然事件由基本事件“正面朝上”和“反面朝上”组成；在掷骰子试验中，随机事件“出现偶数点”可以由基本事件“2点”“4点”和“6点”共同组成。

例1 从字母 a, b, c, d 中任意取出两个不同字母的试验中，有哪些基本事件？

分析：为了得到基本事件，我们可以按照某种顺序，把所有可能的结果都列出来。

解：所求的基本事件共有6个：

$$A=\{a, b\}, B=\{a, c\}, C=\{a, d\},$$

$$D=\{b, c\}, E=\{b, d\},$$

$$F=\{c, d\}.$$

上述试验和例1的共同特点是:

- (1) 试验中所有可能出现的基本事件只有有限个;
- (2) 每个基本事件出现的可能性相等.

我们将具有这两个特点的概率模型称为**古典概率模型** (classical models of probability), 简称古典概型.



在古典概型下, 基本事件出现的概率是多少? 随机事件出现的概率如何计算?

对于掷均匀硬币试验, 出现正面朝上的概率与反面朝上的概率相等, 即

$$P(\text{“正面朝上”}) = P(\text{“反面朝上”}).$$

由概率的加法公式, 得

$$P(\text{“正面朝上”}) + P(\text{“反面朝上”}) = P(\text{必然事件}) = 1.$$

因此

$$P(\text{“正面朝上”}) = P(\text{“反面朝上”}) = \frac{1}{2}.$$

对于掷质地均匀的骰子试验, 出现各个点的概率相等, 即

$$\begin{aligned} P(\text{“1点”}) &= P(\text{“2点”}) = P(\text{“3点”}) \\ &= P(\text{“4点”}) = P(\text{“5点”}) \\ &= P(\text{“6点”}). \end{aligned}$$

反复利用概率的加法公式, 我们有

$$\begin{aligned} &P(\text{“1点”}) + P(\text{“2点”}) + P(\text{“3点”}) + P(\text{“4点”}) \\ &\quad + P(\text{“5点”}) + P(\text{“6点”}) \\ &= P(\text{必然事件}) = 1. \end{aligned}$$

所以

$$\begin{aligned} P(\text{“1点”}) &= P(\text{“2点”}) = P(\text{“3点”}) \\ &= P(\text{“4点”}) = P(\text{“5点”}) \\ &= P(\text{“6点”}) \\ &= \frac{1}{6}. \end{aligned}$$

进一步地, 利用加法公式还可以计算这个试验中任何一个事件的概率, 例如,

$$\begin{aligned} P(\text{“出现偶数点”}) &= P(\text{“2点”}) + P(\text{“4点”}) + P(\text{“6点”}) \\ &= \frac{1}{6} + \frac{1}{6} + \frac{1}{6} \\ &= \frac{1}{2}. \end{aligned}$$

即

$$P(\text{“出现偶数点”}) = \frac{\text{“出现偶数点”所包含的基本事件的个数}}{\text{基本事件的总数}}$$

对于古典概型，任何事件的概率为

$$P(A) = \frac{A \text{ 包含的基本事件的个数}}{\text{基本事件的总数}}$$

例 2 单选题是标准化考试中常用的题型，一般是从

A, B, C, D 四个选项中选择一正确答案. 如果考生掌握了考查的内容，他可以选择唯一正确的答案. 假设考生不会做，他随机地选择一个答案，问他答对的概率是多少？

解：这是一个古典概型，因为试验的可能结果只有 4 个：选择 A、选择 B、选择 C、选择 D，即基本事件共有 4 个，考生随机地选择一个答案是指选择 A, B, C, D 的可能性是相等的. 由古典概型的概率计算公式得

$$P(\text{“答对”}) = \frac{\text{“答对”所包含的基本事件的个数}}{4} = \frac{1}{4} = 0.25.$$

假设有 20 道单选题，如果有一个考生答对了 17 道题，他是随机选择的可能性大，还是他掌握了一定的知识的可能性大？

探究

在标准化的考试中既有单选题又有多选题，多选题是从 A, B, C, D 四个选项中选出所有正确的答案，同学们可能有一种感觉，如果不知道正确答案，多选题更难猜对，这是为什么？

例 3 同时掷两个骰子，计算：

- (1) 一共有多少种不同的结果？
- (2) 其中向上的点数之和是 5 的结果有多少种？
- (3) 向上的点数之和是 5 的概率是多少？

解：(1) 掷一个骰子的结果有 6 种. 我们把两个骰子标上记号 1, 2 以便区分，由于 1 号骰子的每一个结果都可与 2 号骰子的任意一个结果配对，组成同时掷两个骰子的一个结果，因此同时掷两个骰子的结果共有 36 种.

(2) 在上面的所有结果中，向上的点数之和为 5 的结果有

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

其中第一个数表示 1 号骰子的结果，第二个数表示 2 号骰子的结果.

(3) 由于所有 36 种结果是等可能的, 其中向上点数之和为 5 的结果 (记为事件 A) 有 4 种, 因此, 由古典概型的概率计算公式可得

$$P(A) = \frac{4}{36} = \frac{1}{9}.$$

你能列出这 36 个结果吗?

思考?

为什么要把两个骰子标上记号? 如果不标记号会出现什么情况? 你能解释其中的原因吗?

如果不标上记号, 类似于 (1, 2) 和 (2, 1) 的结果将没有区别. 这时, 所有可能的结果将是 (1, 1) (1, 2) (1, 3) (1, 4) (1, 5) (1, 6) (2, 2) (2, 3) (2, 4) (2, 5) (2, 6) (3, 3) (3, 4) (3, 5) (3, 6) (4, 4) (4, 5) (4, 6) (5, 5) (5, 6) (6, 6) 共有 21 种, 和是 5 的结果有 2 个, 它们是 (1, 4) (2, 3), 所求概率为

$$P(A) = \frac{2}{21}.$$

两个答案都是利用古典概型的概率计算公式得到的. 为什么会出现不同结果呢? 这就需要考察两种解法是否满足古典概型的要求. 可以发现, 第一种解法中给出的基本事件是等可能发生的, 但第二种解法中构造的 21 个基本事件不是等可能发生的.

你能说明第二种解法中的基本事件不是等可能发生的原因吗?

由此我们看到, 用古典概型计算概率时, 一定要验证所构造的基本事件是否满足古典概型的第二个条件 (每个结果出现是等可能的), 否则计算出的概率将是错误的.

例 4 假设储蓄卡的密码由 4 个数字组成, 每个数字可以是 0, 1, 2, ..., 9 十个数字中的任意一个. 假设一个人完全忘记了自己的储蓄卡密码, 问他到自动取款机上随机试一次密码就能取到钱的概率是多少?

解: 一个密码相当于一个基本事件, 总共有 10 000 个基本事件, 它们分别是 0000, 0001, 0002, ..., 9998, 9999. 随机地试密码, 相当于试到任何一个密码的可能性都是相等的, 所以这是一个古典概型. 事件“试一次密码就能取到钱”由 1 个基本事件构成, 即由正确的密码构成. 所以



$$P(\text{“试一次密码就能取到钱”}) = \frac{1}{10\,000}.$$

发生概率为 $\frac{1}{10\,000}$ 的事件是小概率事件，通常我们认为这样的事件在一次试验中是几乎不可能发生的，也就是通过随机试验的方法取到储蓄卡中的钱的概率是很小的。但我们也知道，如果试验很多次，比如 100 000 次，那么这个小概率事件是可能发生的。所以，为了安全，自动取款机一般允许取款人最多试 3 次密码，如果第 4 次键入的号码仍是错误的，那么取款机将“没收”储蓄卡。另外，为了使通过随机试验的方法取到储蓄卡中的钱的概率更小，现在储蓄卡可以使用 6 位数字作密码。

人们为了方便记忆，通常用自己的生日作为储蓄卡的密码。当钱包里既有身份证又有储蓄卡时，密码泄密的概率很大，因此用身份证上的号码作密码是不安全的。

例 5 某种饮料每箱装 6 听，如果其中有 2 听不合格，问质检人员从中随机抽出 2 听，检测出不合格产品的概率有多大？



解：我们把每听饮料标上号码，合格的 4 听分别记作：1, 2, 3, 4，不合格的 2 听分别记作 a, b ，只要检测的 2 听中有 1 听不合格，就表示查出了不合格产品。

依次不放回从箱中取出 2 听饮料，得到的两个标记分别记为 x 和 y ，则 (x, y) 表示一次抽取的结果，即基本事件。由于是随机抽取，所以抽取到任何基本事件的概率相等。用 A 表示“抽出的 2 听饮料中有不合格产品”， A_1 表示“仅第一次抽出的是不合格产品”， A_2 表示“仅第二次抽出的是不合格产品”， A_{12} 表示“两次抽出的都是不合格产品”，则 A_1, A_2 和 A_{12} 是互不相容的事件，且

$$A = A_1 \cup A_2 \cup A_{12},$$

从而

$$P(A) = P(A_1) + P(A_2) + P(A_{12}).$$

因为 A_1 中的基本事件的个数为 8， A_2 中的基本事件的个数为 8， A_{12} 中的基本事件的个数为 2，全部基本事件的总数为 30，所以

$$P(A) = \frac{8}{30} + \frac{8}{30} + \frac{2}{30} = 0.6.$$

你能列出 30 种基本事件和事件 A 包含的基本事件吗？



随着检测听数的增加,查出不合格产品的概率怎样变化?为什么质检人员一般都采用抽查的方法而不采用逐个检查的方法?

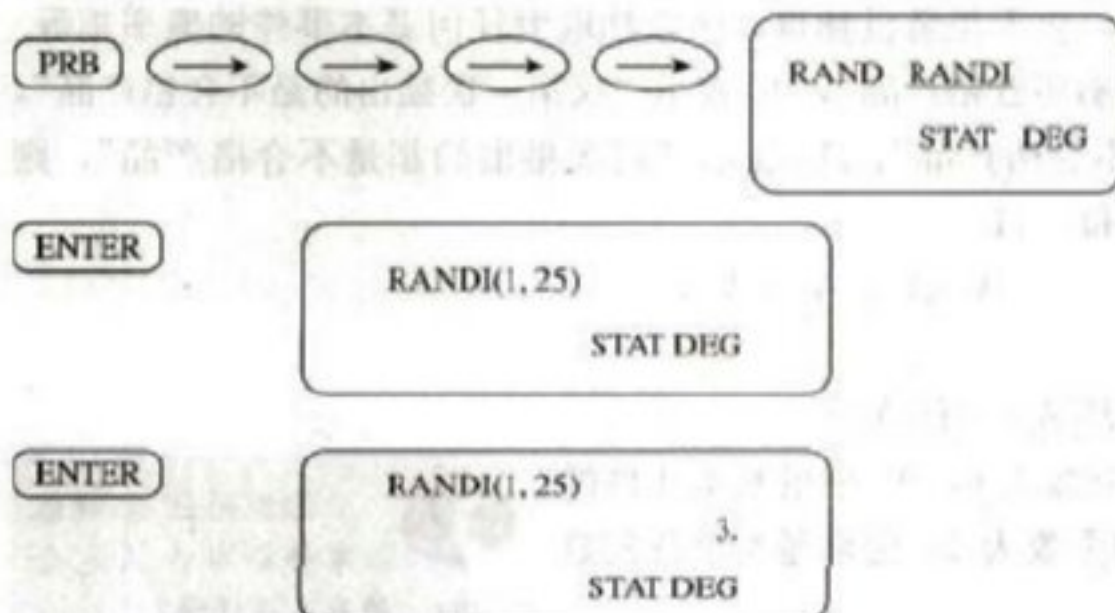
练习

1. 在 20 瓶饮料中,有 2 瓶已过了保质期,从中任取 1 瓶,取到已过保质期的饮料的概率是多少?
2. 在夏令营的 7 名成员中,有 3 名同学已去过北京,从这 7 名同学中任选 2 名同学,选出的这 2 名同学恰是已去过北京的概率是多少?
3. 5 本不同的语文书,4 本不同的数学书,从中任意取出 2 本,取出的书恰好都是数学书的概率为多少?

3.2.2 (整数值)随机数(random numbers)的产生

在第一节中,同学们做了大量重复的试验.有的同学可能觉得这样做试验花费的时间太多了,有没有其他方法可以代替试验呢?

下面我们介绍一种如何用计算器产生指定的两个整数之间的取整数值的随机数.例如,要产生 1~25 之间的取整数值的随机数,按键过程如下:

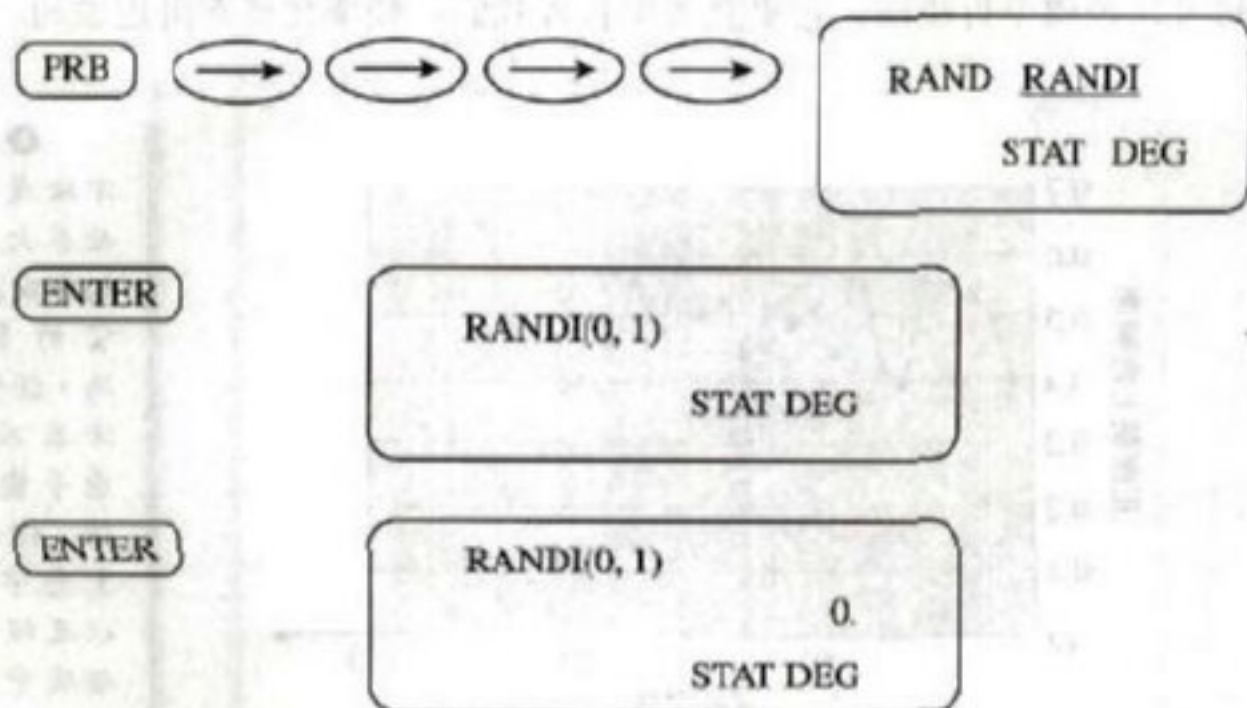


以后反复按 **ENTER** 键,就可以不断产生你需要的随机数.

随机数与伪随机数

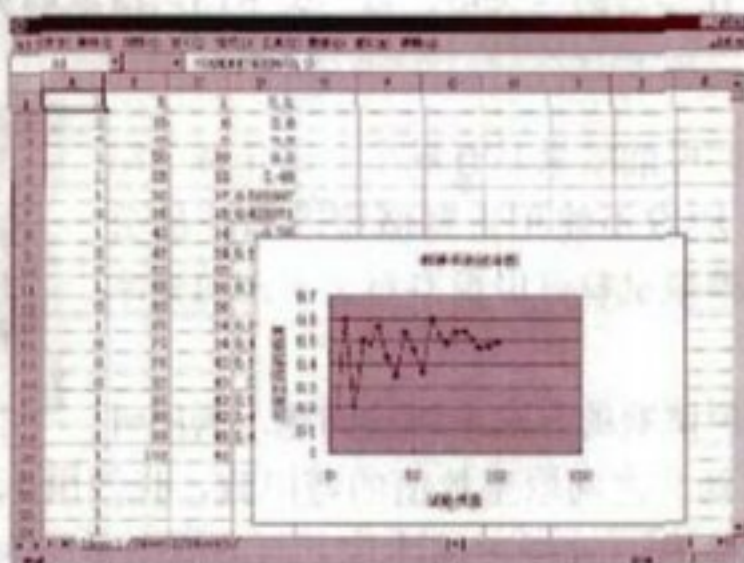
例如我们要产生 1~25 之间的随机整数,我们把 25 个大小形状相同的小球分别标上 1, 2, 3, ..., 24, 25, 放入一个袋中,把它们充分搅拌,然后从中摸出一个,这个球上的数就称为**随机数**.计算机或计算器产生的随机数是依照确定算法产生的数,具有周期性(周期很长),它们具有类似随机数的性质,因此,计算机或计算器产生的并不是真正的随机数,我们称它们为**伪随机数**.

同样地，我们可以用 0 表示反面朝上，1 表示正面朝上，利用计算器不断地产生 0，1 两个随机数，以代替掷硬币的试验。按键过程如下：



我们也可以用计算机产生随机数，而且可以直接统计出频数和频率。下面以掷硬币为例给出计算机产生随机数的方法。

每个具有统计功能的软件都有随机函数。以 Excel 软件为例，打开 Excel 软件，执行下面的步骤：



1. 选定 A1 格，键入 “=RANDBETWEEN (0, 1)①”，按 Enter 键，则在此格中的数是随机产生的 0 或 1。

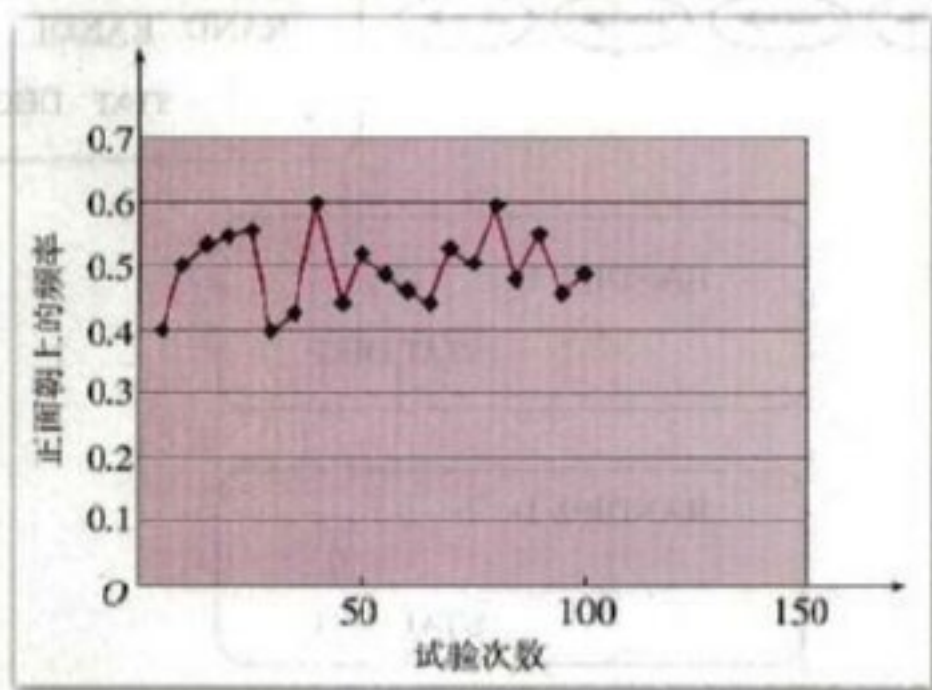
2. 选定 A1 格，按 Ctrl+C 快捷键，然后选定要随机产生 0，1 的格，比如 A2 至 A100，按 Ctrl+V 快捷键，则在 A2 至 A100 的数均为随机产生的 0 或 1，这样我们很快就得到了 100 个随机产生的 0，1，相当于做了 100 次随机试验。

3. 选定 C1 格，键入频数函数 “=FREQUENCY (A1 : A100, 0.5)”，按 Enter 键，则此格中的数是统计 A1 至 A100 中，比 0.5 小的数的个数，即 0 出现的频数，也就是反面朝上的频数。

① 随机函数
RANDBETWEEN
(a, b) 产生从整
数 a 到整数 b 的取
整数值随机数。

4. 选定 D1 格, 键入 “=1-C1/100”, 按 Enter 键, 在此格中的数是这 100 次试验中出现 1 的频率, 即正面朝上的频率.

同时可以画频率折线图, 它更直观地告诉我们: 频率在概率附近波动.



① 蒙特卡罗方法是在第二次世界大战期间兴起和发展起来的, 它的奠基人是冯·诺伊曼, 该方法在应用物理、原子能、固体物理、化学、生物、生态学、社会学以及经济行为等领域中都得到了广泛的应用.

上面我们用计算机或计算器模拟了掷硬币的试验, 我们称用计算机或计算器模拟试验的方法为随机模拟方法或蒙特卡罗 (Monte Carlo) 方法①.

例 6 天气预报说, 在今后的三天中, 每一天下雨的概率均为 40%. 这三天中恰有两天下雨的概率大概是多少?

分析: 这里试验出现的可能结果是有限个, 但是每个结果的出现不是等可能的, 所以不能用古典概型求概率的公式. 用计算器或计算机做模拟试验可以模拟每天下雨的概率是 40%.

解: 我们通过设计模拟试验的方法来解决问题. 利用计算器或计算机可以产生 0 到 9 之间取整数值的随机数, 我们用 1, 2, 3, 4 表示下雨, 用 5, 6, 7, 8, 9, 0 表示不下雨, 这样可以体现下雨的概率是 40%. 因为是 3 天, 所以每三个随机数作为一组. 例如, 产生 20 组随机数

907 966 191 925 271 932 812 458 569 683
431 257 393 027 556 488 730 113 537 989

就相当于做了 20 次试验. 在这组数中, 如果恰有两个数在 1, 2, 3, 4 中, 则表示恰有两天下雨, 它们分别是 191, 271, 932, 812, 393, 即共有 5 个数. 我们得到三天中恰有两天下雨的概率近似为 $\frac{5}{20} = 25\%$.

这里我们用随机模拟的方法得到的仅是 20 次试验中恰有两天下雨的频率或概率的近似值, 而不是概率.

通过例 6, 你能体会到随机模拟的好处吗?

练习

- 将一枚质地均匀的硬币连掷三次，出现“2个正面朝上、1个反面朝上”和“1个正面朝上、2个反面朝上”的概率各是多少？并用随机模拟的方法做100次试验，计算各自的频数。
- 从52张扑克牌（没有大小王）中随机地抽一张牌，这张牌出现下列情形的概率：
 - 是7；
 - 不是7；
 - 是方片；
 - 是J或Q或K；
 - 既是红心又是草花；
 - 比6大比9小；
 - 是红色；
 - 是红色或黑色。
 请设计一种用计算机或计算器模拟上面摸牌试验的方法。
- 盒中仅有4个白球和5个黑球，从中任意取出一个球。
 - “取出的球是黄球”是什么事件？它的概率是多少？
 - “取出的球是白球”是什么事件？它的概率是多少？
 - “取出的球是白球或是黑球”是什么事件？它的概率是多少？
 - 设计一个用计算机或计算器模拟上面取球的试验。
- 掷两粒骰子，计算出现点数总和为7的概率；
 - 利用随机模拟的方法，试验200次，计算出现点数总和为7的频率；
 - 所得频率与概率相差大吗？为什么会有这种差异？

习题3.2

A组

- 下面有三个游戏规则，袋子中分别装有球，从袋中无放回地取球，分别计算甲获胜的概率，哪个游戏是公平的？

游戏1	游戏2	游戏3
1个红球和1个白球	2个红球和2个白球	3个红球和1个白球
取1个球	取1个球，再取1个球	取1个球，再取1个球
取出的球是红球→甲胜	取出的两个球同色→甲胜	取出的两个球同色→甲胜
取出的球是白球→乙胜	取出的两个球不同色→乙胜	取出的两个球不同色→乙胜

2. 某城市的电话号码是 8 位数, 如果从电话号码本中任指一个电话号码, 求:

- (1) 头两位数码都是 8 的概率;
- (2) 头两位数码至少有一个不超过 8 的概率;
- (3) 头两位数码不相同的概率.

3. 某班主任对全班 50 名学生进行了作业量多少的调查, 数据如下表:

	认为作业多	认为作业不多	总数
喜欢电脑游戏	18	9	27
不喜欢电脑游戏	8	15	23
列总数	26	24	50

如果校长随机地问这个班的一名学生, 下面事件发生的概率是多少?

- (1) 认为作业多;
 - (2) 喜欢电脑游戏并认为作业不多.
4. A, B, C, D 4 名学生按任意次序站成一排, 试求下列事件的概率:
- (1) A 在边上;
 - (2) A 和 B 都在边上;
 - (3) A 或 B 在边上;
 - (4) A 和 B 都不在边上.
5. 一个盒子里装有标号为 1, 2, ..., 5 的 5 张标签, 随机地选取两张标签, 根据下列条件求两张标签上的数字为相邻整数的概率:
- (1) 标签的选取是无放回的;
 - (2) 标签的选取是有放回的.
6. 在一个盒中装有 6 枚圆珠笔, 其中 3 枝一等品, 2 枝二等品和 1 枝三等品, 从中任取 3 枝, 问下列事件的概率有多大?
- (1) 恰有一枝一等品;
 - (2) 恰有两枝一等品;
 - (3) 没有三等品.

B 组

1. 某人有 4 把钥匙, 其中 2 把能打开门. 现随机地取 1 把钥匙试着开门, 不能开门的就扔掉, 问第二次才能打开门的概率是多少? 如果试过的钥匙不扔掉, 这个概率又是多少?
2. 假设有 5 个条件很类似的女孩, 把她们分别记为 A, C, J, K, S. 她们应聘秘书工作, 但只有 3 个秘书职位, 因此 5 人中仅有三人被录用. 如果 5 个人被录用的机会相等, 分别计算下列事件的概率:
 - (1) 女孩 K 得到一个职位;
 - (2) 女孩 K 和 S 各自得到一个职位;
 - (3) 女孩 K 或 S 得到一个职位.
3. 假设每个人在任何一个月出生是等可能的, 利用随机模拟的方法, 估计在一个有 10 个人的集体中至少有两个人的生日在同一个月的概率?

我们已经学习了两种方法计算随机事件发生的概率，一是通过做试验或者用计算机模拟试验等方法得到事件发生的频率，以此来近似估计概率；二是用古典概型的公式来计算事件发生的概率。在现实生活中，常常会遇到试验的所有可能结果是无穷多的情况，这时就不能用古典概型来计算事件发生的概率了。

在特定情形下，我们可以用几何概型来计算事件发生的概率。

3.3.1 几何概型

在概率论发展的早期，人们就已经注意到只考虑那种仅有有限个等可能结果的随机试验是不够的，还必须考虑有无限多个试验结果的情况。例如一个人到单位的时间可能是8:00~9:00之间的任何一个时刻；往一个方格中投一个石子，石子可能落在方格中的任何一点上……这些试验可能出现的结果都是无限多个。下面我们通过几个例子来说明相应概率的求法。

问题 图 3.3-1 中有两个转盘，甲乙两人玩转盘游戏，规定当指针指向 B 区域时，甲获胜，否则乙获胜。在两种情况下分别求甲获胜的概率是多少？

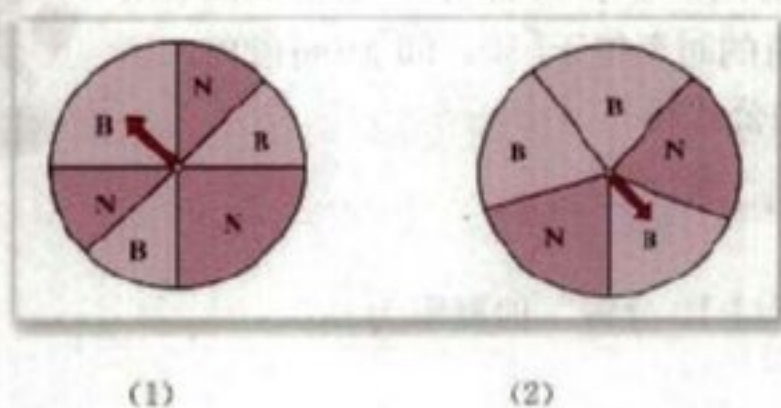


图 3.3-1

显然，以转盘 (1) 为游戏工具时，甲获胜的概率为 $\frac{1}{2}$ ；以转盘 (2) 为游戏工具时，甲获胜的概率为 $\frac{3}{5}$ 。事实上，甲获胜的概率与字母 B 所在扇形区域的圆弧的长度

有关,而与字母B所在区域的位置无关,只要字母B所在扇形区域的圆弧的长度不变,不管这些区域是相邻,还是不相邻,甲获胜的概率是不变的.

如果每个事件发生的概率只与构成该事件区域的长度(面积或体积)成比例,则称这样的概率模型为**几何概率模型**(geometric models of probability),简称为几何概型.

在几何概型中,事件A的概率的计算公式如下:

$$P(A) = \frac{\text{构成事件A的区域长度(面积或体积)}}{\text{试验的全部结果所构成的区域长度(面积或体积)}}$$

因此,如果把图3.3-1中的圆周的长度设为1,则以转盘(1)为游戏工具时,

$$P(\text{“甲获胜”}) = \frac{\frac{1}{2}}{1} = \frac{1}{2};$$

以转盘(2)为游戏工具时,

$$P(\text{“甲获胜”}) = \frac{\frac{3}{5}}{1} = \frac{3}{5}.$$

例1 某人午觉醒来,发现表停了,他打开收音机,想听电

台报时,求他等待的时间不多于10分钟的概率.

分析: 假设他在0~60分钟之间任何一个时刻打开收音机是等可能的,但0~60之间有无穷个时刻,不能用古典概型的公式计算随机事件发生的概率.我们可以通过随机模拟的方法得到随机事件发生的概率的近似值,也可以通过几何概型的求概率公式得到事件发生的概率.因为电台每隔1小时报时一次,他在0~60之间任何一个时刻打开收音机是等可能的,所以他在哪个时间段打开收音机的概率只与该时间段的长度有关,而与该时间段的位置无关,这符合几何概型的条件.

解: 设 $A = \{\text{等待的时间不多于10分钟}\}$. 我们所关心的事件A恰好是打开收音机的时刻位于 $[50, 60]$ 时间段内,因此由几何概型的求概率的公式得

$$P(A) = \frac{60-50}{60} = \frac{1}{6}.$$

即“等待报时的时间不超过10分钟”的概率为 $\frac{1}{6}$.

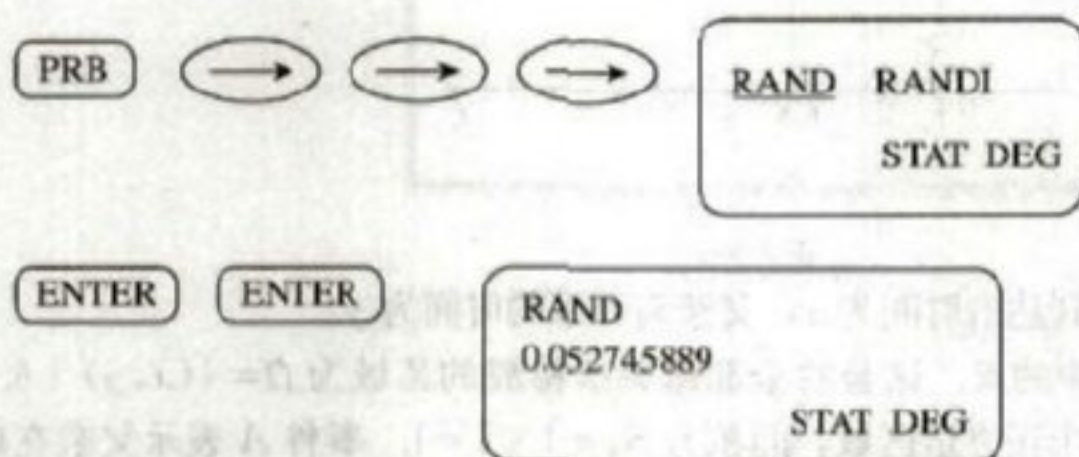
在本例中,打开收音机的时刻X是随机的,可以是0~60之间的任何一刻,并且是等可能的.我们称X服从 $[0, 60]$ 上的均匀分布,X为 $[0, 60]$ 上的均匀随机数.



你能用圆盘等设计一种方法模拟试验吗?

3.3.2 均匀随机数的产生

我们常用的是 $[0, 1]$ 上的均匀随机数，可以利用计算器来产生 $0 \sim 1$ 之间的均匀随机数（实数），方法如下：



注意：每次
结果会有不同。

试验的结果是区间 $[0, 1]$ 内的任何一个实数，而且出现任何一个实数是等可能的，因此，就可以用上面的方法产生的 $0 \sim 1$ 之间的均匀随机数进行随机模拟。



如果试验的结果是区间 $[a, b]$ 上的任何一点，而且是等可能的，如何产生 $[a, b]$ 之间的均匀随机数？



例 2 假设你家订了一份报纸，送报人可能在早上6:30~7:30之间把报纸送到你家，你父亲离开家去工作的时间在早上7:00~8:00之间，问你父亲在离开家前能得到报纸（称为事件A）的概率是多少？

分析：我们有两种方法计算该事件的概率：

- (1) 利用几何概型的公式；
- (2) 用随机模拟的方法。

解：方法一

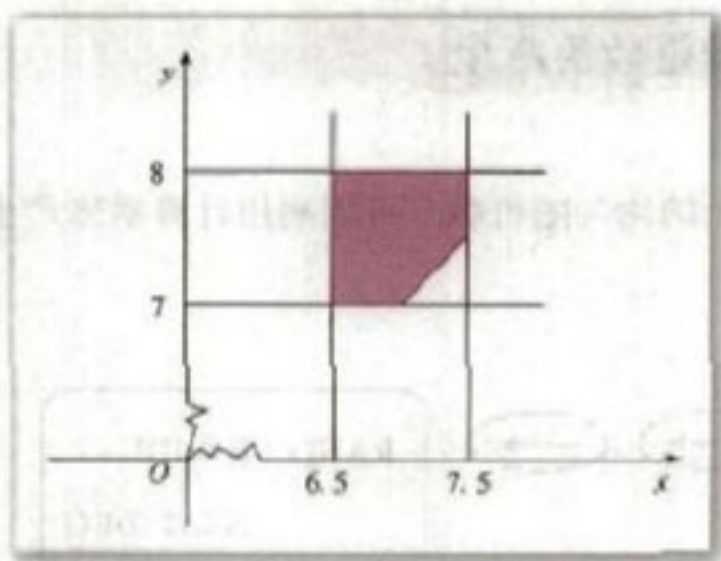


图 3.3-2

如图 3.3-2, 设送报人到达的时间为 x , 父亲离开家的时间为 y .

(x, y) 可以看成平面中的点. 试验的全部结果所构成的区域为 $\Omega = \{(x, y) \mid 6.5 \leq x \leq 7.5, 7 \leq y \leq 8\}$, 这是一个正方形区域, 面积为 $S_{\Omega} = 1 \times 1 = 1$. 事件 A 表示父亲在离开家前能得到报纸, 所构成的区域为 $A = \{(x, y) \mid y \geq x, 6.5 \leq x \leq 7.5, 7 \leq y \leq 8\}$, 即图中的阴影部分, 面积为 $S_A = 1 - \frac{1}{2} \times \frac{1}{2} \times \frac{1}{2} = \frac{7}{8}$. 这是一个几何概型, 所以

$$P(A) = \frac{S_A}{S_{\Omega}} = \frac{7}{8}.$$

思考?

你能设计一种随机模拟的方法, 近似计算上面事件 A 发生的概率吗?
(包括手工的方法或用计算器、计算机的方法.)

方法二 (随机模拟的方法) 我们可以做两个带有指针 (分针) 的圆盘, 标上时间, 分别旋转两个圆盘, 记下父亲在离开家前能得到报纸的次数, 则

$$P(A) = \frac{\text{父亲在离家前能得到报纸的次数}}{\text{试验的总次数}}.$$



我们也可以用计算机产生随机数模拟试验. X 是 $0 \sim 1$ 之间的均匀随机数, Y 也是 $0 \sim 1$ 之间的均匀随机数. 如果 $Y + 7 > X + 6.5$, 即 $Y > X - 0.5$, 那么父亲在离开家前能得到报纸. 下面是我们在计算机上做的 50 次试验, 得到的结果是 $P(A) = 0.88$.

在 Excel 中产生 $[0, 1]$ 区间上均匀随机数的函数为 "rand ()".

例 3 在图 3.3-3 的正方形中随机撒一把豆子，用随机模拟的方法估计圆周率的值。

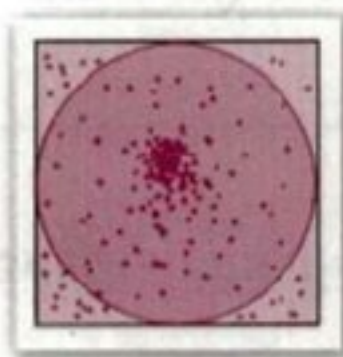


图 3.3-3

解：随机撒一把豆子，每个豆子落在正方形内任何一点是等可能的，落在每个区域的豆子数与这个区域的面积近似成正比，即

$$\frac{\text{圆的面积}}{\text{正方形的面积}} \approx \frac{\text{落在圆中的豆子数}}{\text{落在正方形中的豆子数}}$$

假设正方形的边长为 2，则

$$\frac{\text{圆的面积}}{\text{正方形的面积}} = \frac{\pi}{2 \times 2} = \frac{\pi}{4}$$

由于落在每个区域的豆子数是可以数出来的，所以

$$\pi \approx \frac{\text{落在圆中的豆子数}}{\text{落在正方形中的豆子数}} \times 4,$$

这样就得到了 π 的近似值。

同学们可以自己做一个仪器，具体实践一下上述试验。

另外，我们可以用计算器或计算机模拟上述过程，步骤如下：

(1) 产生两组 $0 \sim 1$ 之间的均匀随机数， $a_1 = \text{RAND}$ ， $b_1 = \text{RAND}$ ；

(2) 经平移和伸缩变换， $a = (a_1 - 0.5) \times 2$ ， $b = (b_1 - 0.5) \times 2$ ；

(3) 数出落在圆内 $x^2 + y^2 < 1$ 的点 (a, b) 的个数 N_1 ，计算 $\pi = \frac{4N_1}{N}$ (N 代表落在正方形中的点 (a, b) 的个数)。

可以发现，随着试验次数的增加，得到的 π 的近似值的精度会越来越高。

本例启发我们，利用几何概型，并通过随机模拟方法可以近似计算不规则图形的面积。

例4 利用随机模拟方法计算图 3.3-4 中阴影部分 ($y=1$ 和 $y=x^2$ 所围成的部分) 的面积.

分析: 在坐标系中画出矩形 ($x=1$, $x=-1$, $y=1$ 和 $y=0$ 所围成的部分), 用随机模拟的方法可以得到它的面积的近似值.

解: (1) 利用计算器或计算机产生两组 $0\sim 1$ 区间的均匀随机数, $a_1 = \text{RAND}$, $b = \text{RAND}$;

(2) 进行平移和伸缩变换, $a = (a_1 - 0.5) * 2$;

(3) 数出落在阴影内的样本点数 N_1 , 用几何概型公式计算阴影部分的面积.

例如做 1 000 次试验, 即 $N=1\ 000$, 模拟得到 $N_1=698$, 所以

$$S \approx \frac{2N_1}{N} = 1.396.$$

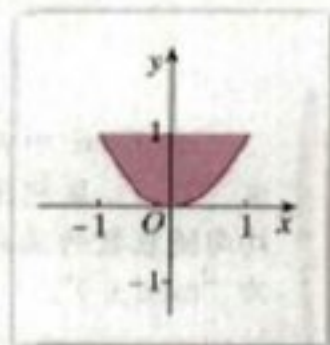
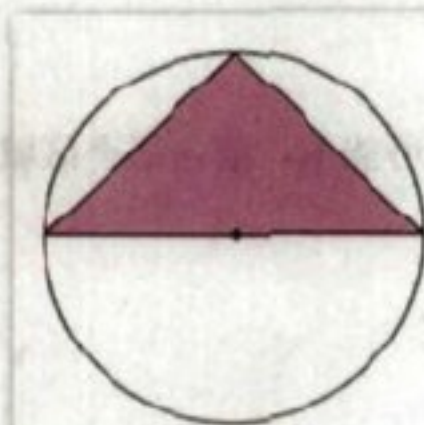


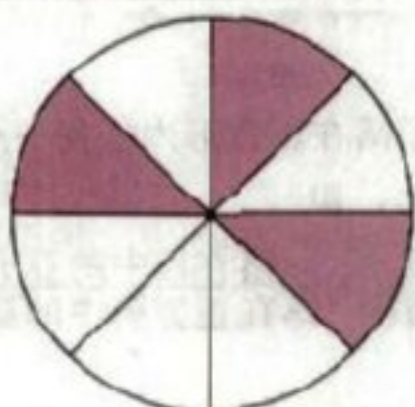
图 3.3-4

练习

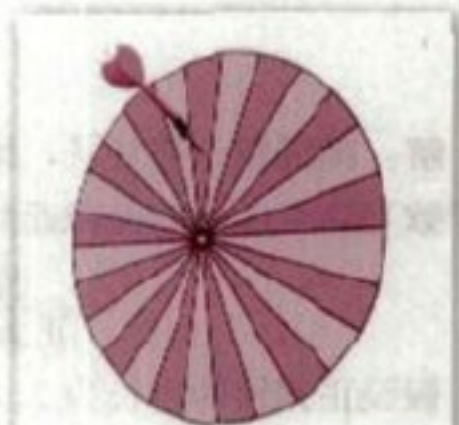
1. 如图, 假设你在每个图形上随机撒一粒黄豆, 分别计算它落到阴影部分的概率.



(第1题)



(第2题)



2. 如图, 如果你向靶子上射 200 镖, 你期望多少个镖落在红色区域 (颜色较深的区域).



概率与密码

从古到今, 在军事、政治、经济等方面, 文件的保密性很重要. 如果文件泄密, 那么可能会导致战役的失败、经济上的重大损失, 甚至会导致国家的灭亡. 为了保证安全, 保密文件的传送经常用“密文”的方式进行.

密文的设计通常利用密码转换. 以传送命令: “We will start the fight at eleven o'clock on Wednesday” 为例, 显然, 在传送过程中应当做到: 即使敌方得到了这个命令

也不知道其含义。最早的加密方法是伟大的罗马军事家和政治家凯撒 (Gaius Julius Caesar, 约前 100—前 44) 发明的。他设计了把密文中的每个字母用按字母次序后移三位的字母代替的方法。用此方法编译上面的命令, 得到 “Zh zcoo vwduw wkh ilkw dw hohy-hq r'fojkw rq Zhgqhvgdb”。如果不知道替换规则, 很难理解其中的含义。后来有人使用把 26 个字母分别对应 1~26 个自然数或其他代码等方法传送密文。只要传送一方和接受一方均知道这个对应表即可。

这种方法使用了很长一段时间后, 有人掌握了破译的方法。你知道是如何破译的吗?

用我们掌握的概率知识, 就可以破译这个密码。经过研究, 人们发现, 书面语言中字母以基本固定的频率出现, 如下表所示。

字母	A	B	C	D	E	F	G	H	I
频率	0.081 6	0.015 5	0.022 3	0.046 3	0.123 1	0.023 7	0.019 8	0.067 1	0.066 9
字母	J	K	L	M	N	O	P	Q	R
频率	0.000 8	0.006 8	0.035 4	0.027 3	0.067 3	0.079 5	0.015 6	0.000 6	0.055 5
字母	S	T	U	V	W	X	Y	Z	
频率	0.057 8	0.097 7	0.028 1	0.011 2	0.027 8	0.001 4	0.020 6	0.000 4	

从表中可以看到, 不同字母出现的频率不同, 这是书面语言的一个重要特征。在通常的文章中, 字母 “e” 平均出现的比例占有所有字母的 12% 左右, “t” 占 9.7% 左右, 而 “j” 的出现远小于 1%。如果掌握了这个规律, 再用上面的方法加密, 通过对用密码写的密文中字母的频率分析, 就比较容易破译出密文。出现频率最高的字母, 无论你在编译中使用什么字母, 它一般都表示 “e”, 出现频率次高的字母大概是 “t”, 等等。

上面编译密码的方法的共同特点是一个字母对应另一个确定的字母。当收到的只是短短的一句话时, 要找出这种对应关系是比较困难的。但如果文件比较大, 或者经常收到一个地方的密文, 经过一段时间的积累, 就可以利用对字母的频率分析, 得到字母与密码的对应关系, 这样编译的密码就容易被破译了。

为了使密码设计得更难破译, 人们发明了许多反破译的方法。利用随机序列就是一种极为重要的方法, 其原理是: 利用取值于 1~26 之间的整数值随机数序列, 使每个字母出现在密码中的概率都相等。一种理论上不可破译的密码是 “(用后即销毁的) 一次密码本”。在实际应用中, 这种密码本是伪随机序列, 序列中的每一个数都是 1 到 26 之间的整数。例如, 若组成这个密码本的伪随机序列为: 12, 16, 5, 7, 21, 19, 15, 13, 4, 14, 11, 10, 16, 24, 18, 15, 19, 11, 5, …, 要发送的命令是 “We will start the fight at eleven o'clock on Wednesday”, 那么在 “We” 这个词中, W 对应于伪随机数 12, 就按字母顺序用 W 后面的第 12 个字母 I 表示 W, e 对应于伪随机数 16, 就用 e 后面第 16 个字母 u 表示 e, “will” 编译的过程为 $w+5 \rightarrow b$, $i+7 \rightarrow p$, $l+21 \rightarrow g$, $l+19 \rightarrow e$ 等等。全句的密文为 “lu bpge hgefe dxb …”。这样一来, 对方再想通过分析每个字母出现的频率来破译密码就不可能了, 因为在密文中每个字母出现的频率几乎相等。

密码虽然神秘, 但只要掌握一些概率的知识, 我们就能编译它。概率的应用是不是很奇妙?

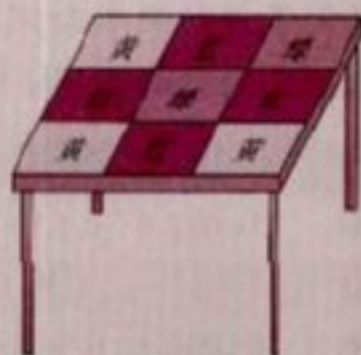
你还能发现概率在其他方面的应用吗?

习题 3.3

A 组

1. 一张方桌的图案如图所示. 将一颗豆子随机地扔到桌面上, 假设豆子不落在线上, 求下列事件的概率:

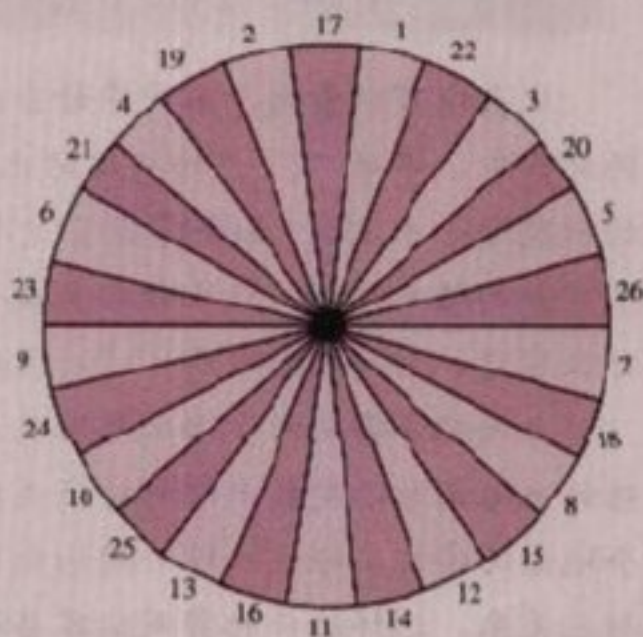
- (1) 豆子落在红色区域;
- (2) 豆子落在黄色区域;
- (3) 豆子落在绿色区域;
- (4) 豆子落在红色或绿色区域;
- (5) 豆子落在黄色或绿色区域.



(第 1 题)

2. 一个靶子如图所示. 随机地掷一个飞镖扎在靶子上, 假设飞镖既不会落在黑色靶心, 也不会落在两种颜色之间, 求飞镖落在下列区域的概率:

- (1) 编号为 25 的区域;
- (2) 绿色区域 (颜色较浅的区域);
- (3) 编号不小于 24 的区域;
- (4) 编号在 6 号到 9 号之间的区域 (按逆时针方向);
- (5) 编号为奇数的区域;
- (6) 红色 (颜色较深) 的编号为奇数的区域.



(第 2 题)

3. 一个路口的红绿灯, 红灯的时间为 30 秒, 黄灯的时间为 5 秒, 绿灯的时间为 40 秒. 当你到达路口时, 看见下列三种情况的概率各是多少?

- (1) 红灯; (2) 黄灯; (3) 不是红灯.

B 组

甲、乙两艘轮船都要在某个泊位停靠 6 小时, 假定它们在一昼夜的时间段中随机地到达, 试求这两艘船中至少有一艘在停靠泊位时必须等待的概率.

小结

一、本章知识结构



二、回顾与思考

1. 随机事件的概率：随机事件在一次试验中是否发生是不确定的，但在大量重复试验中，随机事件的发生是有规律的，概率就是要寻找这种规律性。

你能举几个在日常生活中利用概率的例子吗？

2. 随机现象的产生：在现实中，很多结果的出现受众多随机因素的影响，由于对这些因素难以掌握或缺乏了解，因此在试验前我们不能确定会出现哪个结果，这样就产生了随机现象。你能举出随机现象的例子吗？你会用什么方法了解这个随机现象？

3. 频率与概率的关系与区别：频率是概率的近似值，随着试验次数的增加，频率会越来越接近概率。频率本身也是随机的，两次做同样的试验，会得到不同的结果；而概率是一个确定的数，与每次试验无关。

(1) 试验 100 次得到的频率一定比试验 50 次得到频率更接近概率吗？

(2) 你有办法了解你得到的频率是否接近概率吗？

4. 利用古典概型与几何概型可以求一些随机事件的概率。

(1) 古典概型有哪些特征？

(2) 几何概型有哪些特征？

(3) 古典概型与几何概型的区别是什么？

复习参考题

A 组

1. 甲、乙两人下棋，两人下成和棋的概率是 $\frac{1}{2}$ ，乙获胜的概率是 $\frac{1}{3}$ ，则乙不输的概率是_____，甲获胜的概率是_____，甲不输的概率是_____.
2. 某个制药厂正在测试一种减肥新药的疗效，有 500 名志愿者服用此药，结果如下：

体重变化	体重减轻	体重不变	体重增加
人数	274	93	133

如果另有一人服用此药，估计下列事件发生的概率：

- (1) 此人的体重减轻；
 - (2) 此人的体重不变；
 - (3) 此人的体重增加.
3. 将一枚质地均匀的硬币连续投掷 4 次，出现“2 次正面朝上，2 次反面朝上”和“3 次正面朝上，1 次反面朝上”的概率各是多少？
 4. 某校有教职工 130 人，对他们进行年龄状况和教育程度的调查，其结果如下：

	本科	研究生	合计
35 岁以下	50	35	85
35~50 岁	20	13	33
50 岁以上	10	2	12

随机地抽取一人，求下列事件的概率：

- (1) 具有本科学历；
 - (2) 35 岁以下具有研究生学历；
 - (3) 50 岁以上.
5. 甲袋中有 1 只白球、2 只红球、3 只黑球，乙袋中有 2 只白球、3 只红球、1 只黑球，现从两袋中各取一球，求两球颜色相同的概率.
 6. 有 2 个人在一座 7 层大楼的底层进入电梯，假设每一个人自第二层开始在每一层离开电梯是等可能的，求 2 个人在不同层离开的概率.



(第 2 题)

B 组

1. 掷一枚均匀的硬币 4 次, 求出现正面的次数多于反面次数的概率.
2. 某小组有 3 名男生和 2 名女生, 从中任选 2 名学生参加演讲比赛, 判断下列各对事件是否为互斥事件, 并说明理由.
 - (1) 恰有 1 名男生和恰有 2 名男生;
 - (2) 至少有 1 名男生和至少有 1 名女生;
 - (3) 至少有 1 名男生和全是男生;
 - (4) 至少有 1 名男生和全是女生.
3. 柜子里有 3 双不同的鞋, 随机地取出 2 只, 试求下列事件的概率, 并说明它们的关系:
 - (1) 取出的鞋不成对;
 - (2) 取出的鞋都是左脚的;
 - (3) 取出的鞋都是同一只脚的;
 - (4) 取出的鞋一只是左脚的, 一只是右脚的, 但它们不成对.
4. 利用随机模拟的方法近似计算图形的面积: $y=x^2+1$ 与 $y=6$ 所围区域的面积.

