

大自然这本书是用数学语言写成的。

——伽利略

一种科学只有在成功地运用数学时，才算达到完善的地步。

——马克思

致 同 学

亲爱的同学，你感到高中阶段的学习生活有趣吗？

我们知道，数学与生活紧密相连。数学可以帮助我们认识世界，改造世界，创造新的生活。数学是高中阶段的重要学科，不仅是学习物理、化学等学科的基础，而且对我们的终身发展有较大的影响。

面对实际问题，我们要认真观察、实验、归纳，大胆提出猜想。为了证实或推翻提出的猜想，我们要通过分析，概括、抽象出数学概念，通过探究、推理，建立数学理论。我们要积极地运用这些理论去解决问题。在探究与应用过程中，我们的思维水平会不断提高，我们的创造能力会得到发展。在数学学习过程中，我们将快乐地成长。

考虑到广大同学的不同需要，本书提供了较大的选择空间。

书中的引言、正文、练习、习题中的“感受·理解”部分、阅读、回顾等内容构成一个完整的体系。它体现了教材的基本要求，是所有学生应当掌握的内容。相信你一定能学好这部分内容。

本书还设计了一些具有挑战性的内容，包括思考、探究、链接，以及习题中的“思考·运用”、“探究·拓展”等，以激发你探索数学的兴趣。在掌握基本内容之后，选择其中一些内容作思考与探究，你会更加喜欢数学。

目 录

第 1 章 算法初步

1.1 算法的含义	5
1.2 流程图	7
1.3 基本算法语句	17
1.4 算法案例	26

第 2 章 统计

2.1 抽样方法	43
2.2 总体分布的估计	53
2.3 总体特征数的估计	65
2.4 线性回归方程	74

第 3 章 概率

3.1 随机事件及其概率	93
3.2 古典概型	100
3.3 几何概型	106
3.4 互斥事件	112

附 录

附录 1 随机数表(部分)	124
附录 2 本章测试答案与提示	125

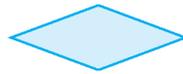
本书部分常用符号



输入、输出框



处理框



判断框



起止框

$$p \leftarrow q$$

将 q 的值赋给 p

$$\bar{x}$$

样本平均数

$$\sum_{i=1}^n a_i$$

n 个实数 a_1, a_2, \dots, a_n 的和

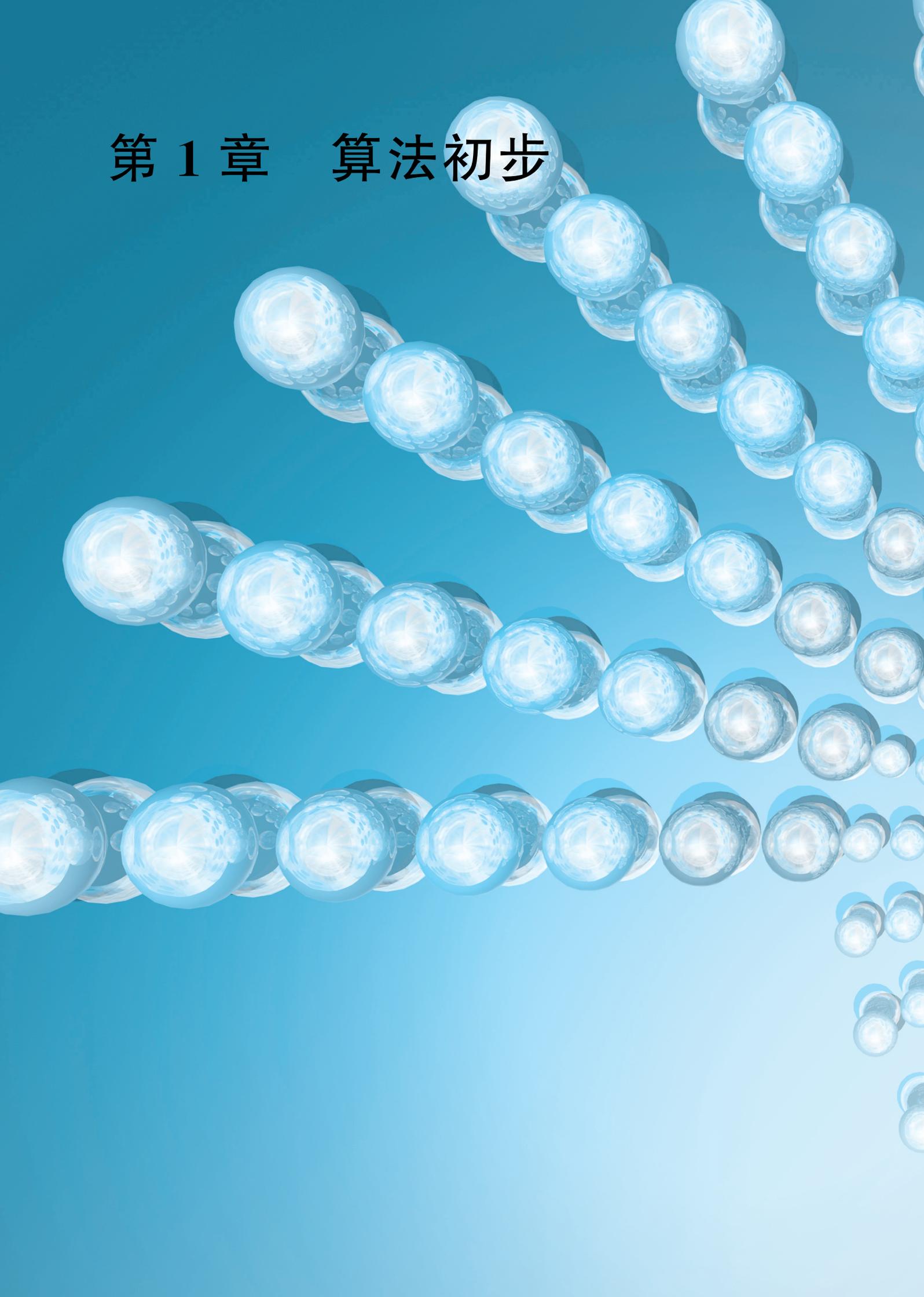
$$P(A)$$

事件 A 的概率

$$\bar{A}$$

事件 A 的对立事件

第 1 章 算法初步





[-]	[book icon]	算法初步
[+]	[folder icon]	算法的含义
[-]	[folder icon]	流程图
[+]	[folder icon]	顺序结构
[+]	[folder icon]	选择结构
[+]	[folder icon]	循环结构
[-]	[folder icon]	基本算法语句
[+]	[folder icon]	赋值语句
[+]	[folder icon]	输入、输出语句
[+]	[folder icon]	条件语句
[+]	[folder icon]	循环语句
[+]	[folder icon]	算法案例

对世界各式各样的观察方式中,……最有趣的方式之一,是被设想为由模式组成的那种方式.

——维 纳

假如你的朋友不会发送电子邮件,你能教会他吗?发送电子邮件的方法很多,下面就是一种操作步骤:

- 第一步 打开电子信箱;
- 第二步 单击“写邮件”按钮;
- 第三步 输入发送地址;
- 第四步 输入主题;
- 第五步 输入信件内容;
- 第六步 单击“发送邮件”按钮.

我们做任何一件事情,都是在一定的条件下按某种顺序执行的一系列操作,解决数学问题也常常如此.例如,用加减消元法解二元一次方程组,就可以按照某一程序进行操作;用配方法解一元二次方程,也是按一定程序操作的.

将上述程序转换成计算机能识别的语言后,就能借助计算机极大地提高解决问题的速度.因此,探索解决问题的统一程序的思想是十分重要的.对一类问题的机械的、统一的求解方法就是算法.

从古代的“百鸡问题”到现代机器证明数学定理的“吴方法”,从二元一次方程组的消元解法到计算机动画的设计,从猜数游戏到集成电路的布线安排,……它们中都蕴含了丰富的算法思想.

面对一个需要解决的问题,

- 如何设计解决问题的操作序列?
- 怎样用数学语言描述上述操作序列?

1.1

算法的含义

电视娱乐节目中,有一种有趣的“猜数”游戏:竞猜者如在规定的时间内猜出某种商品的价格(或重量等),就可获得该件商品.

● 现有一商品,价格在 0~8 000 元之间,采取怎样的策略才能在较短的时间内猜出正确的答案呢?

解决这个问题有多种途径,其中一种较好的方法是:

第一步 报“4 000”;

第二步 若主持人说“高了”(说明答数在 0~4 000 之间),就报“2 000”,否则(答数在 4 000~8 000 之间)报“6 000”;

第三步 重复第二步的报数方法,直至得到正确答案.

以上过程实际上是按一种机械的程序进行的一系列操作.

一般而言,对一类问题的机械的、统一的求解方法称为

算法(algorithm).



阿尔·花拉子米 (al-Khowārizmī, 约 780~约 850),主要著作有《代数学》,被西方人称为“代数学之父”.英文 algorithm 一词据说是源于这位数学家的名字.

例 1 给出求 $1+2+3+4+5$ 的一个算法.

算法 1 按照逐一相加的方法进行.

第一步 计算 $1+2$,得到 3;

第二步 将第一步中的运算结果 3 与 3 相加,得到 6;

第三步 将第二步中的运算结果 6 与 4 相加,得到 10;

第四步 将第三步中的运算结果 10 与 5 相加,得到 15.

算法 2 可以运用公式 $1+2+3+\cdots+n = \frac{n(n+1)}{2}$ 直接计算.

第一步 取 $n = 5$;

第二步 计算 $\frac{n(n+1)}{2}$;

第三步 输出运算结果.

例 2 给出求解方程组

$$\begin{cases} 2x + y = 7, & \text{①} \\ 4x + 5y = 11 & \text{②} \end{cases}$$

的一个算法.

解 我们用消元法求解这个方程组,步骤是:

第一步 方程①不动,将方程②中 x 的系数除以方程①中 x 的

系数,得到乘数

$$m = \frac{4}{2} = 2;$$

第二步 方程②减去 m 乘以方程①的积,消去方程②中的 x 项,得到

$$\begin{cases} 2x + y = 7, \\ 3y = -3; \end{cases}$$

第三步 将上面的方程组自下而上回代求解,得到

$$y = -1, x = 4.$$

所以原方程组的解为

$$\begin{cases} x = 4, \\ y = -1. \end{cases}$$

这种消元回代的算法适用于一般线性方程组的求解.

找到了某种算法,是指使用一系列运算规则能在有限步骤内求解某类问题,其中的每条规则必须是明确定义的、可行的.

算法从初始步骤开始,每一个步骤只能有一个确定的后继步骤,从而组成一个步骤序列,序列的终止表示问题得到解答或指出问题没有解答.

我们过去学过的许多数学公式都是算法,加、减、乘、除运算法则以及多项式的运算法则也是算法.

练习

1. 写出解方程 $2x + 3 = 0$ 的一个算法.
2. 写出求 $1 \times 3 \times 5 \times 7$ 的一个算法.
3. 已知平面直角坐标系中的两点 $A(-1, 0)$, $B(3, 2)$,写出求直线 AB 的方程的一个算法.
4. 写出求 $1 + 2 + 3 + \cdots + 100$ 的一个算法.

1.2

流程图

为了将设计好的算法清晰直观地描述出来,通常采用画流程图的方法来表示.那么,

- 流程图是怎样构成的?如何用流程图描述基本的算法结构?

例如,计算 $1 + 2 + 3 + 4 + 5$ 的算法还可以用下面的流程图来表示(图 1-2-1).

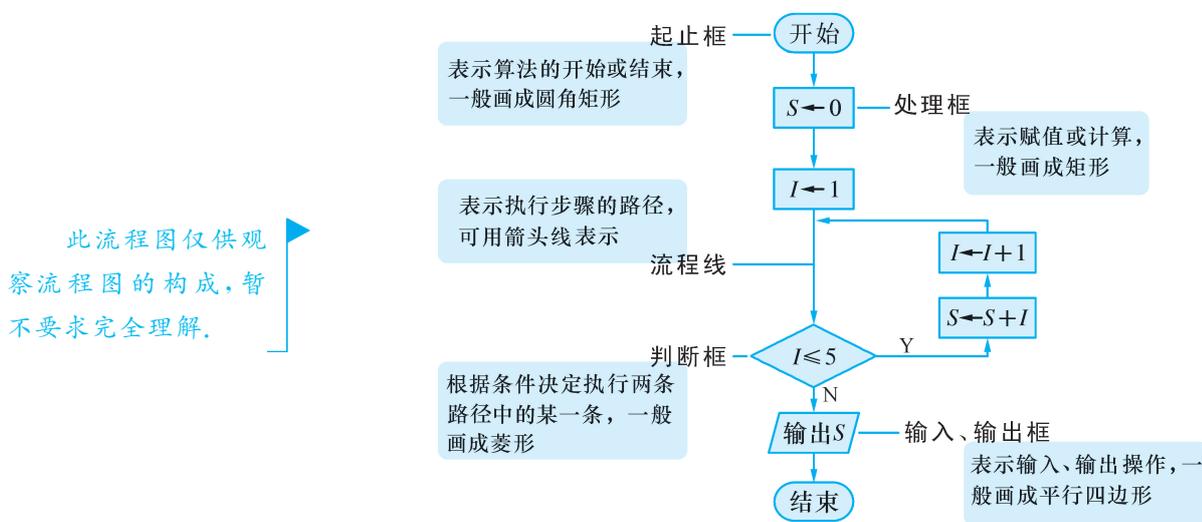


图 1-2-1

流程图(flow chart)是由一些图框和流程线组成的,其中图框表示各种操作的类型,图框中的文字和符号表示操作的内容,流程线表示操作的先后次序.

从流程图 1-2-1 中可以看出,该算法步骤中,有些是按顺序执行,有些需选择执行,而另外一些则需循环执行.

事实上,算法都可以由顺序结构、选择结构和循环结构这 3 块“积木”通过组合和嵌套表达出来.流程图可以帮助我们更方便直观地表示这 3 种基本的算法结构.

1.2.1 顺序结构

这里 S1 代表步骤 1, S2 代表步骤 2, 依此类推. S 是 step (步) 的第一个字母.

作 $\triangle ABC$ 的外接圆可按如下的算法进行:

- S1 作 AB 的垂直平分线 l_1 ;
- S2 作 BC 的垂直平分线 l_2 ;
- S3 以 l_1 与 l_2 的交点 M 为圆心, 以 MA 为半径作圆, $\odot M$ 即为 $\triangle ABC$ 的外接圆.

● 上述算法过程具有怎样的特点?

以上算法通过依次执行 3 个步骤, 完成了作外接圆这一过程(图 1-2-2). 像这种依次进行多个处理的结构称为**顺序结构**(sequence structure).

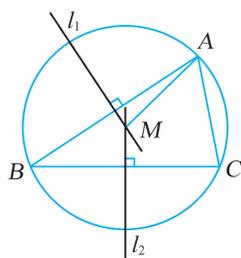


图 1-2-2

如图 1-2-3 所示, 虚线框内是一个顺序结构, 其中 A 和 B 两个框是依次执行的. 顺序结构是一种最简单、最基本的结构.

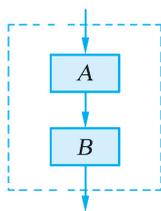


图 1-2-3

例 1 已知两个单元分别存放了变量 x 和 y 的值, 试设计交换这两个变量的值的一个算法, 并画出流程图.

解 为了交换两个变量的值, 可以增加一个单元存放中间变量 p . 其算法如下:

将 x 的值赋给变量 p , 记作“ $p \leftarrow x$ ”.

- S1 $p \leftarrow x$; {先将 x 的值赋给变量 p , 这时存放变量 x 的单元可作他用}
- S2 $x \leftarrow y$; {再将 y 的值赋给变量 x , 这时存放变量 y 的单元可作他用}
- S3 $y \leftarrow p$. {最后将 p 的值赋给变量 y , 两个变量 x 和 y 的值便完成了交换}

流程图如图 1-2-4 所示.

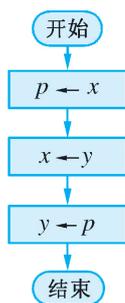


图 1-2-4

例 2 半径为 r 的圆的面积计算公式为

$$S = \pi r^2,$$

当 $r = 10$ 时, 写出计算圆面积的算法, 并画出流程图.

解 算法如下:

S1 $r \leftarrow 10;$

{把 10 赋给变量 r }

S2 $S \leftarrow \pi r^2;$

{用公式计算圆的面积}

S3 输出 S .

{输出圆的面积}

流程图如图 1-2-5 所示.

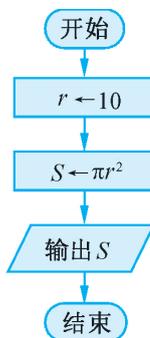


图 1-2-5

练习

1. 写出作 $\triangle ABC$ 的内切圆的一个算法.

2. 写出解方程组 $\begin{cases} x + y = 3, \\ y + z = 5, \\ z + x = 4 \end{cases}$ 的一个算法, 并画出流程图.

1.2.2 选择结构

某铁路客运部门规定甲、乙两地之间旅客托运行李的费用 c (单位: 元) 与行李的重量 w (单位: kg) 之间的关系为

$$c = \begin{cases} 0.53 \times w, & w \leq 50, \\ 50 \times 0.53 + (w - 50) \times 0.85, & w > 50. \end{cases}$$

● 计算费用 c 的算法可以用怎样的算法结构来表示?

为了计算行李的托运费, 应先判断行李的重量是否大于 50 kg, 然后再选用相应的公式进行计算. 其算法如下:

- S1 输入行李的重量 w ;
 S2 如果 $w \leq 50$, 那么 $c \leftarrow 0.53 \times w$,
 否则 $c \leftarrow 50 \times 0.53 + (w - 50) \times 0.85$;
 S3 输出行李重量 w 和托运费 c .

流程图如图 1-2-6 所示.

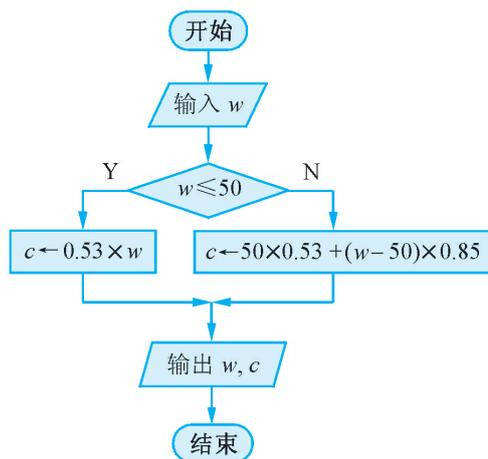


图 1-2-6

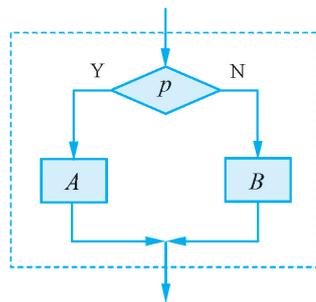


图 1-2-7

在上述算法过程中, S2 进行了判断, 像这种先根据条件作出判断, 再决定执行哪一种操作的结构称为**选择结构**(selection structure)(或称为“分支结构”). 如图 1-2-7 所示, 虚线框内是一个选择结构, 它包含一个判断框, 当条件 p 成立(或称为“真”)时执行 A , 否则执行 B .

思考

在流程图 1-2-1 所表示的算法中, 哪一步进行了判断?

例 3 设计求解一元二次方程

$$ax^2 + bx + c = 0 \quad (a \neq 0)$$

的一个算法, 并画出流程图.

分析 因为一元二次方程未必总有实数根, 所以求解时先计算判别

式 $\Delta = b^2 - 4ac$ ，然后比较判别式与 0 的大小，再决定能否用求根公式进行求解。因此，在算法中要使用选择结构。

解 算法如下：

S1 输入 a, b, c ；

S2 $\Delta \leftarrow b^2 - 4ac$ ；

S3 如果 $\Delta < 0$ ，那么输出“方程无实数根”，否则

$$x_1 \leftarrow \frac{-b + \sqrt{\Delta}}{2a}, x_2 \leftarrow \frac{-b - \sqrt{\Delta}}{2a}, \text{输出 } x_1, x_2.$$

流程图如图 1-2-8 所示。

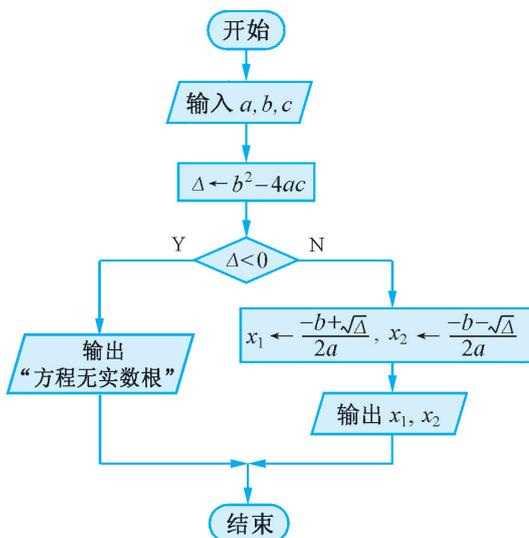
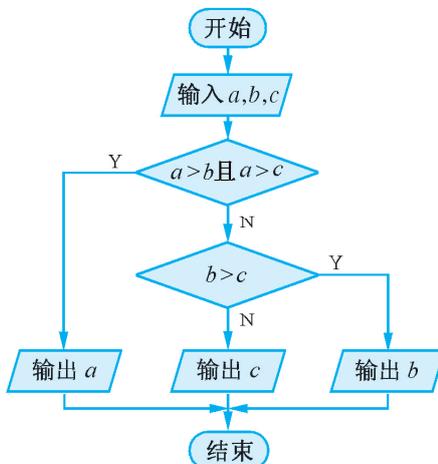


图 1-2-8

练习

1. 如果考生的成绩大于或等于 60 分，那么输出“及格”，否则输出“不及格”。用流程图表示这一算法过程。
2. 下面的流程图表示了一个什么样的算法？



(第 2 题)

3. 写出解方程 $ax + b = 0$ (a, b 为常数) 的一个算法，并画出流程图。

1.2.3 循环结构

在学校的长跑测试中,你每跑 1 圈,会想是否跑完了全程.如果没有跑完全程,那么又会想离终点还有多远.

● 用怎样的算法结构表示这个过程?

以万米长跑(每圈 400 m)为例,我们分步描述上述过程:

S1 起跑;

S2 如果未跑到 10 000 m,那么转 S3,否则转 S4;

S3 跑 1 圈,转 S2;

S4 结束.

上述算法可用图 1-2-9 所示的流程图来表示.

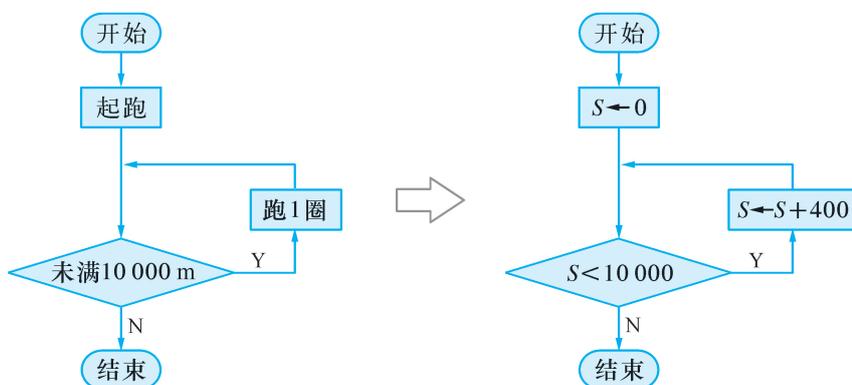


图 1-2-9

在算法中,像这种需要重复执行同一操作的结构称为**循环结构**(cycle structure).

图 1-2-10 是一种常见的循环结构:先判断所给条件 p 是否成立,若 p 成立,则执行 A ,再判断条件 p 是否成立;若 p 仍成立,则又执行 A .如此反复,直到某一次条件 p 不成立时为止.这样的循环结构称为当型循环.

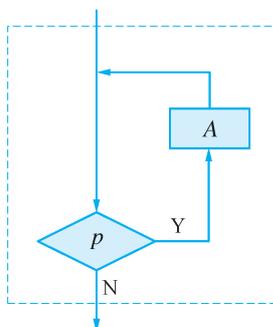


图 1-2-10

对于万米长跑,如果我们先跑 1 圈后再判断,那么还可以这样描述算法过程:

- S1 起跑;
- S2 跑 1 圈;
- S3 如果跑到 10 000 m,那么转 S4,否则转 S2;
- S4 结束.

流程图如图 1-2-11 所示.

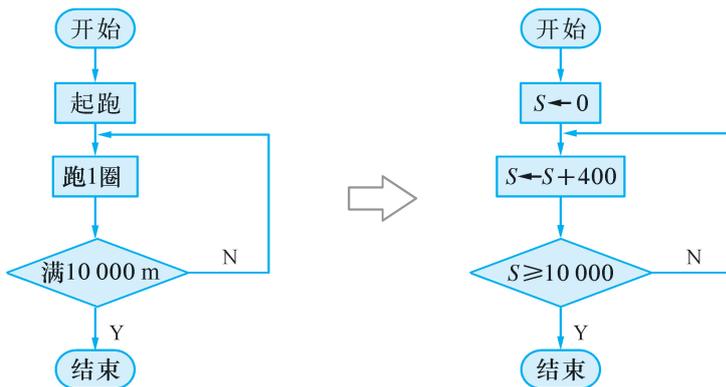


图 1-2-11

上面这种循环结构称为直到型循环(图 1-2-12):先执行 A,再判断所给条件 p 是否成立,若 p 不成立,则再执行 A. 如此反复,直到 p 成立,该循环过程结束.

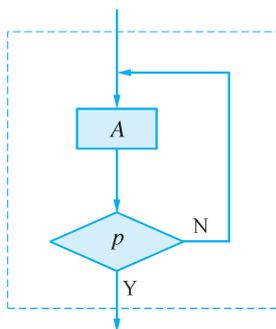


图 1-2-12

思考

在流程图 1-2-1 中,哪些步骤构成了循环结构?

例 4 写出求 $1 \times 2 \times 3 \times 4 \times 5$ 的一个算法,并画出流程图.

分析 我们用变量 T 存放乘积的结果,变量 I 作为计数变量.每循环一次, I 的值增加 1.

- 解** S1 $T \leftarrow 1;$ {使 $T = 1$ }
 S2 $I \leftarrow 2;$ {使 $I = 2$ }

对算法做少许改动,求 $1 \times 3 \times 5 \times 7 \times 9$ 的值.

- S3 如果 $I \leq 5$, 那么转 S4, 否则转 S6; {当 $I \leq 5$ 时循环}
- S4 $T \leftarrow T \times I$; {求 $T \times I$, 其积仍存放在变量 T 中}
- S5 $I \leftarrow I + 1$, 转 S3; {使 I 的值增加 1, 并转到 S3}
- S6 输出 T . {输出结果}

流程图如图 1-2-13 所示.

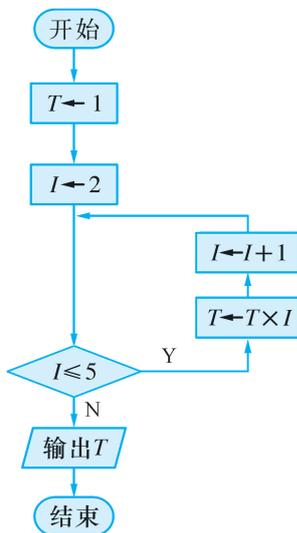


图 1-2-13

思考

在流程图 1-2-13 中,使用了哪一种循环结构? 如何用另一种循环结构来描述同样的问题?

例 5 设计一个计算 10 个数的平均数的算法,并画出流程图.

分析 我们用一个循环依次输入 10 个数,再用一个变量存放数的累加和.在求出 10 个数的总和后,除以 10,就得到这 10 个数的平均数.

- 解**
- S1 $S \leftarrow 0$; {使 $S = 0$ }
 - S2 $I \leftarrow 1$; {使 $I = 1$ }
 - S3 如果 $I \leq 10$, 那么转 S4, 否则转 S7; {当 $I \leq 10$ 时循环}
 - S4 输入 G ; {输入一个数}
 - S5 $S \leftarrow S + G$; {求 $S + G$, 其和仍存放在 S 中}
 - S6 $I \leftarrow I + 1$, 转 S3; {使 I 的值增加 1, 并转到 S3}
 - S7 $A \leftarrow S/10$; {将平均数 $S/10$ 存放在 A 中}
 - S8 输出 A . {输出平均数}

流程图如图 1-2-14 所示.

将 0 赋给 S ,是为这些数的和建立存放空间.

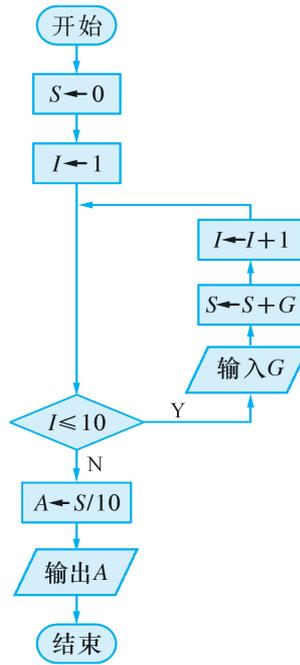
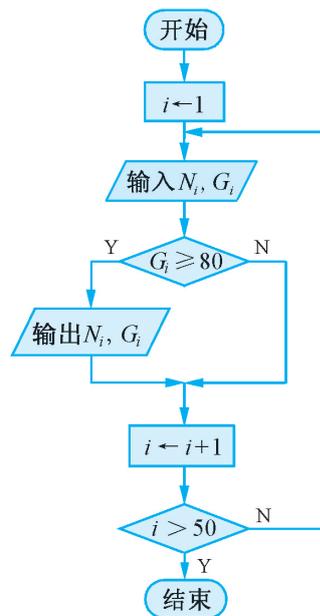


图 1-2-14

由上面的例题可以看到,利用顺序结构、选择结构和循环结构这 3 种基本结构描述的算法,结构清晰,容易阅读、理解和修改.

练习

1. 写出计算 $2 + 4 + 6 + \dots + 100$ 的一个算法,并画出流程图.
2. 用 N_i 代表第 i 个学生的学号, G_i 代表第 i 个学生的成绩 ($i = 1, 2, \dots, 50$), 下面的流程图表示了一个什么样的算法?



(第 2 题)

习题 1.1

感受·理解

1. 三角形面积的计算公式为 $S = \frac{1}{2}ah$ (其中 a 为某边长, h 为该边上的高), 当 $a = 7.85$, $h = 14.29$ 时, 写出计算三角形面积的一个算法, 并画出流程图.
2. 火车站对乘客在一定时段内退票要收取一定的费用, 收费的办法是: 按票价每 10 元 (不足 10 元按 10 元计算) 核收 2 元, 票价在 2 元以下的不退. 试写出将票价为 x (单位: 元) 的车票退掉后, 应返还的金额 y (单位: 元) 的一个算法, 并画出流程图.
3. 画出解方程组 $\begin{cases} 2x - y = 1, \\ 4x + 3y = 7 \end{cases}$ 的一个算法的流程图.
4. 画出求两个正整数 a 与 b 相除所得商 q 及余数 r 的一个算法的流程图.
5. 写出求 3 个实数中的最小数的一个算法, 并画出流程图.
6. 写出解不等式 $ax + b > 0$ ($a \neq 0$) 的一个算法, 并画出流程图.

思考·运用

7. 写出在数 3, 5, 8, 9, 12, 15, 35, 7, 18, 52 中搜索数 18 的一个算法, 并画出流程图.
8. 我国南宋时期的数学家秦九韶发现了求三角形面积的“三斜求积”公式

$$S_{\triangle ABC} = \sqrt{\frac{1}{4} \left[c^2 a^2 - \left(\frac{c^2 + a^2 - b^2}{2} \right)^2 \right]}$$

它与古希腊数学家海伦给出的三角形面积公式 $S_{\triangle ABC} = \sqrt{p(p-a)(p-b)(p-c)}$

(其中 $p = \frac{1}{2}(a+b+c)$) 是一致的. 试根据海伦公式, 画出已知 3 条线段 a, b, c , 求以这 3 条线段为边的三角形面积的一个算法的流程图. (提示: 输入 3 个数后应先判断以它们为三边长能否构成三角形)

海伦 (Heron, 生卒年不详), 古希腊数学家, 约公元 62 年前后活跃于亚历山大.

探究·拓展

9. 写出求 $\frac{1}{2 + \frac{1}{2 + \frac{1}{\ddots + \frac{1}{2}}}}$ (共有 n 个 2) 的值的一个算法, 并画出流程图.

1.3

基本算法语句

- 算法是一种数学语言,如何用更简洁的语句表述算法语言呢?

本节主要通过伪代码学习基本的算法语句.

伪代码(pseudo code)是介于自然语言和计算机语言之间的文字和符号,是表达算法的简单而实用的好方法.下面,我们在伪代码中将使用 VB(Visual Basic)语言的关键词.

1.3.1 赋值语句

在伪代码中,赋值语句(assignment statement)用符号“ \leftarrow ”表示,“ $x \leftarrow y$ ”表示将 y 的值赋给 x ,其中 x 是一个变量, y 是一个与 x 同类型的变量或表达式.

例 1 用伪代码写出求 $x = 23$ 时多项式 $7x^3 + 3x^2 - 5x + 11$ 的值的一个算法.

算法 1

$x \leftarrow 23;$
 $p \leftarrow 7x^3 + 3x^2 - 5x + 11.$

算法 2

$x \leftarrow 23;$
 $p \leftarrow ((7x + 3)x - 5)x + 11.$

上述两种算法,算法 1 要做 6 次乘法,算法 2 只需做 3 次乘法.由此可见,算法的好坏会影响运算速度.

例 1 中算法 2 称为秦九韶算法,其算法特点是:通过一次式的反复计算,逐步得出高次多项式的值;对于一个 n 次多项式,至多做 n 次乘法和 n 次加法.

1.3.2 输入、输出语句

“鸡兔同笼”是我国隋朝时期的数学著作《孙子算经》中的一个有趣且具有深远影响的问题:“今有雉兔同笼,上有三十五头,下有九十四足.问雉兔各几何.”用方程组的思想不难解决这一问题.设有 x 只鸡, y 只兔,则

$$\begin{cases} x + y = 35, \\ 2x + 4y = 94. \end{cases}$$



孙子算经》(局部)

下面我们设计解二元一次方程组 $\begin{cases} a_1x + b_1y = c_1, \\ a_2x + b_2y = c_2 \end{cases}$ ($a_1b_2 - a_2b_1 \neq 0$) 的一个算法:用消元法解得

$$\begin{cases} x = \frac{b_2c_1 - b_1c_2}{a_1b_2 - a_2b_1}, \\ y = \frac{a_1c_2 - a_2c_1}{a_1b_2 - a_2b_1}, \end{cases}$$

因此,只要输入相应的未知数的系数和常数项,就能计算出方程组的解,即可以输出 x, y 的值.

我们用输入语句(input statement)“Read a, b ”表示输入的数据依次送给 a, b ,用输出语句(output statement)“Print x ”表示输出运算结果 x .这样,上述解二元一次方程组的算法流程图与相应的伪代码就可以表示为图 1-3-1.

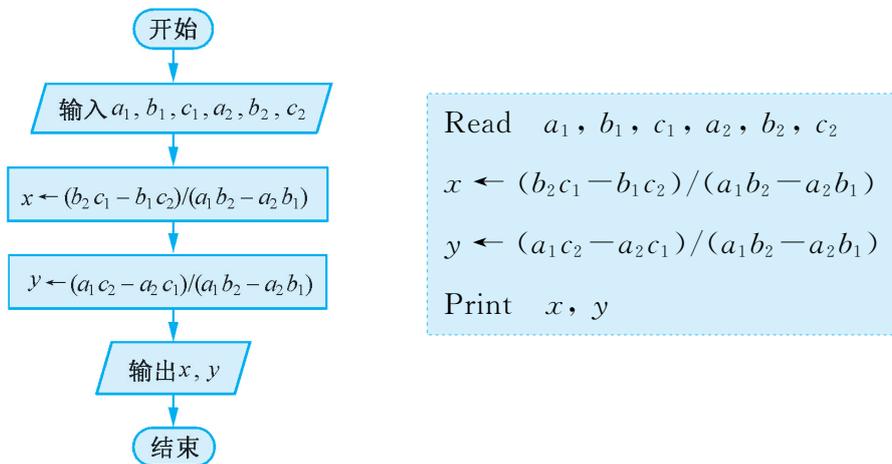


图 1-3-1

当输入 $a_1, b_1, c_1, a_2, b_2, c_2$ 分别为 1, 1, 35, 2, 4, 94 时,输出的 x, y 的值分别为 23, 12. 即“鸡兔同笼”问题的答案是 23 只鸡和 12 只兔.

练习

1. 已知一个正三棱柱的底面边长为 2, 高为 3, 用输入、输出语句和赋值语句表示计算这个正三棱柱的体积的一个算法.
2. 已知三角形的三边长分别为 a, b, c , 借助三角形的面积公式

$$S = \sqrt{p(p-a)(p-b)(p-c)} \quad \left(\text{其中 } p = \frac{1}{2}(a+b+c) \right),$$

用输入、输出语句和赋值语句表示计算三角形面积的一个算法.

3. 某市 2006 年 1~12 月的产值(单位: 亿元)分别为 3.8, 4.2, 5.3, 6.1, 5.6, 4.8, 7.3, 4.5, 6.4, 5.8, 4.7, 6.5, 该市要统计每季度的月平均产值及 2006 年的月平均产值, 试分别用输入、输出语句和赋值语句表示计算上述各个平均值的一个算法.

1.3.3 条件语句

某居民区的物管部门每月按以下方法收取卫生费：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。

● 如何设计算法，根据输入的人数计算应收取的卫生费？

我们令 c (单位：元) 表示应收取的费用， n 表示某户人家的人口数，则有

$$c = \begin{cases} 5, & 0 < n \leq 3, \\ 5 + 1.2(n - 3), & n > 3. \end{cases}$$

解决这一问题的算法步骤如下：

- S1 输入 n ；
- S2 如果 $n \leq 3$ ，那么 $c \leftarrow 5$ ，否则 $c \leftarrow 5 + 1.2(n - 3)$ ；
- S3 输出 c 。

流程图如图 1-3-2 所示，从中可以看出这是一个选择结构。在执行此算法时，要根据一定的条件选择流程线的方向。

我们可以运用 **条件语句** (conditional statement) 来实现上述过程。条件语句的一般形式是

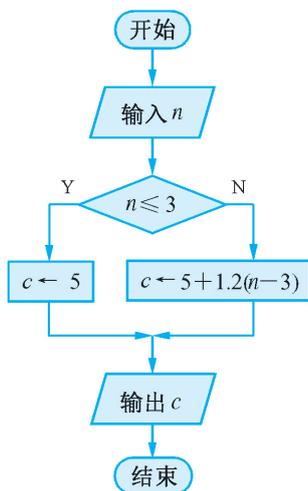


图 1-3-2

```

If A Then
    B
Else
    C
End If
  
```

其中 A 表示判断的条件， B 表示满足条件时执行的操作内容， C 表示不满足条件时执行的操作内容，End If 表示条件语句结束。

上面的算法过程用条件语句可以表示为

```

Read n

If n ≤ 3 Then
    c ← 5
Else
    c ← 5 + 1.2(n - 3)
End If

Print c
  
```

我们把步骤“ $c \leftarrow 5$ ”称为“Then”分支，步骤“ $c \leftarrow 5 + 1.2(n - 3)$ ”称为“Else”分支。为了醒目和便于阅读，这些分支一般缩进书写。

例 2 儿童乘坐火车时,若身高不超过 1.1 m,则无须购票;若身高超过 1.1 m 但不超过 1.4 m,可买半票;若超过 1.4 m,应买全票.试设计一个购票的算法,写出伪代码,并画出流程图.

解 上述购票的算法步骤如下:

S1 测量儿童身高 h ;

S2 如果 $h \leq 1.1$, 那么免费乘车;否则,如果 $h \leq 1.4$, 那么购买半票乘车;否则,购买全票乘车.

用条件语句表示如下:

```

Read  $h$ 
If  $h \leq 1.1$  Then
    Print 免费乘车
Else
    If  $h \leq 1.4$  Then
        Print 半票乘车
    Else
        Print 全票乘车
    End If
End If
    
```

流程图如图 1-3-3 所示.

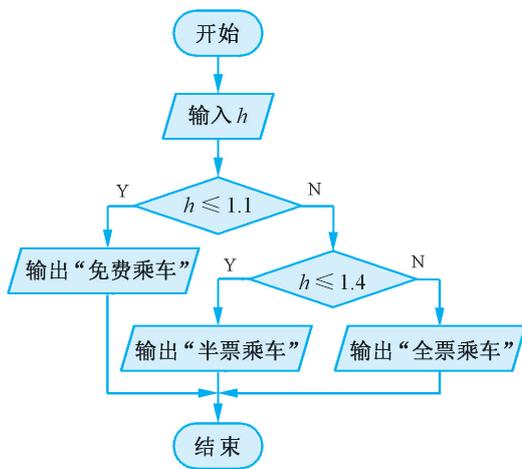


图 1-3-3

由例 2 可知,条件语句“**If-Then-Else**”可以嵌套.

思考

条件语句也可以没有“**Else**”分支.你能举一个例子说明吗?

例 3 已知函数

$$y = \begin{cases} 1, & x > 0, \\ 0, & x = 0, \\ -1, & x < 0, \end{cases}$$

试用伪代码写出根据输入的 x 的值计算 y 值的一个算法,并画出流程图.

解 用条件语句可以方便地表示这类分段函数：

```

Read x
If x > 0 Then
    y ← 1
Else
    If x = 0 Then
        y ← 0
    Else
        y ← -1
    End If
End If
Print y
    
```

流程图如图 1-3-4 所示。

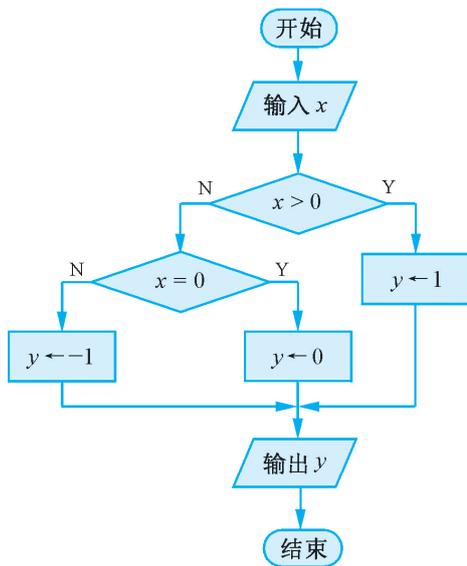


图 1-3-4

本例中的分段函数称为“符号函数”，即通过函数值(输出值)是 1 还是 -1 来判断输入的值是正数还是负数。

练习

1. 用条件语句表示：输入两个数，输出较大的数。
2. 已知函数 $y = \begin{cases} x, & x \geq 0, \\ -x, & x < 0, \end{cases}$ 试用伪代码写出根据输入的 x 的值计算 y 值的一个算法。
3. 办理个人异地汇款(不超过 100 万)时,银行要收取一定的手续费: 汇款额不超过 100 元时收取 1 元手续费,超过 100 元但不超过 5 000 元时按汇款额的 1% 收取手续费,超过 5 000 元时一律收取 50 元手续费. 试用条件语句描述汇款额为 x (单位: 元)时银行收取的手续费 y (单位: 元)的一个算法,并画出流程图。

1.3.4 循环语句

● 设计计算

$$1 \times 3 \times 5 \times 7 \times \cdots \times 99$$

的一个算法.

我们用变量 T 存放乘积的结果, 变量 I 作为计数变量. 每循环 1 次, 将乘积 $T \times I$ 的结果仍存放在变量 T 中, 同时使 I 的值增加 2. 算法如下:

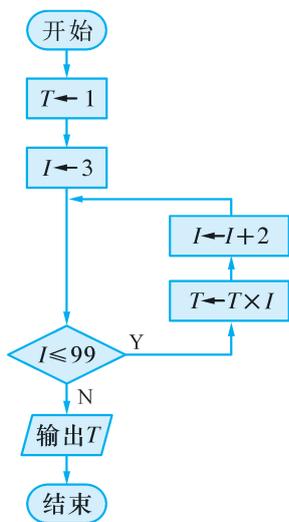


图 1-3-5

- S1 $T \leftarrow 1;$
- S2 $I \leftarrow 3;$
- S3 如果 $I \leq 99$, 那么转 S4, 否则转 S6;
- S4 $T \leftarrow T \times I;$
- S5 $I \leftarrow I + 2$, 转 S3;
- S6 输出 T .

流程图如图 1-3-5 所示. 从流程图可以看出这是一个循环结构, 我们可以运用循环语句 (cycle statement) 来实现上述过程.

当型循环可以用下面的语句形式来描述:

```

While p
    循环体
End While
  
```

它表示当所给条件 p 成立时, 执行循环体部分, 然后再判断条件 p 是否成立. 如果 p 仍然成立, 那么再次执行循环体. 如此反复, 直到某一次条件 p 不成立时退出循环.

上述算法用当型语句“While ... End While”表示如下:

```

T ← 1
I ← 3
  
```

```

While I ≤ 99
    T ← T × I
    I ← I + 2
End While
  
```

```
Print T
```

当型语句的特点是先判断, 后执行.

前面的算法也可以改写为直到型循环:

- S1 $T \leftarrow 1;$
- S2 $I \leftarrow 3;$

当型循环与直到型循环可以相互转换, 一般采用当型语句描述循环结构.

S3 $T \leftarrow T \times I;$
 S4 $I \leftarrow I + 2;$
 S5 如果 $I > 99$, 那么转 S6, 否则转 S3;
 S6 输出 T .

直到型循环可以用下面的语句形式来描述:

```
Do
    循环体
Until  $p$ 
End Do
```

它表示先执行循环体部分, 然后再判断所给条件 p 是否成立. 如果 p 不成立, 那么再次执行循环体部分. 如此反复, 直到所给条件 p 成立时退出循环.

上面的算法用直到型语句“Do ... End Do”表示如下:

$T \leftarrow 1$
 $I \leftarrow 3$

```
Do
     $T \leftarrow T \times I$ 
     $I \leftarrow I + 2$ 
Until  $I > 99$ 
End Do
```

Print T

直到型语句的特点是先执行, 后判断.

如果循环结构中的循环次数已知, 那么还可以采用“For”语句来描述. “For”语句的一般形式为:

For 循环是当型循环, 即当循环变量 I 满足“初值” $\leq I \leq$ “终值”时, 就执行循环.

```
For  $I$  From “初值” To “终值” Step “步长”
    循环体
End For
```

例如, 前面的问题用 For 语句可以表示为:

$T \leftarrow 1$

```
For  $I$  From 3 To 99 Step 2
     $T \leftarrow T \times I$ 
End For
```

Print T

在上面的“For”语句中, 如果省略“Step 2”, 那么重复循环时, I 的值每次增加 1.

例 4 抛掷一枚硬币时,既可能出现正面,也可能出现反面,预先作出确定的判断是不可能的.但是假如硬币质量均匀,那么当抛掷次数很多时,出现正面的频率应接近于 50%.试设计一个循环语句模拟抛掷硬币的过程,并计算抛掷中出现正面的频率.

分析 抛掷硬币的过程实际上是一个不断重复地做同一件事情的过程,利用循环语句,我们很容易在计算机上模拟这一过程.

在程序语言中,有一个随机函数“Rnd”,它能产生 0 与 1 之间的随机数.这样,我们可以用大于 0.5 的随机数表示出现正面,不大于 0.5 的随机数表示出现反面.

解 本题算法的伪代码如下:

```
s ← 0
Read n
For i From 1 To n
    If Rnd > 0.5 Then s ← s + 1
End For
Print 出现正面的频率为  $\frac{s}{n}$ 
```

单行条件语句不需要结束标志“End If”.

练习

1. 在第 1.2 节开头,我们用流程图 1-2-1 给出了计算 $1+2+3+4+5$ 的一个算法,试用循环语句描述这一算法过程.
2. 2000 年我国人口数约为 13 亿.如果每年的人口自然增长率为 15%,那么多少年后我国人口将达到或超过 15 亿?这个问题可以通过循环方式计算完成,即每一次在原有的基础上增加 15%,直到达到或超过 15 亿,再记下循环次数.试用循环语句表示这一过程.
3. 1, 1, 2, 3, 5, 8, 13, … 这一列数的规律是:第 1、第 2 个数是 1,从第 3 个数起,每个数是其前面 2 个数的和.试用循环语句描述计算这列数中第 20 个数的一个算法.

EXCEL

VBA(Visual Basic for Application)是 Excel 自带的一种程序设计语言,它具有一般程序设计语言所具有的功能,可由手工写入或宏记录器两种方式生成.

使用 VBA 宏记录器无须亲自写 VBA 的代码,在计算机内会自动生成 VBA 的代码.你只要打开宏记录器,做 1 次你所需要的操作,例如画 1 个经常要用的表格,宏记录器会用代码记录下你的每一步操作,操作完成后,保存为一个叫宏的文件.下次再做同样的事,你只要执行该文件,就可以自动画出已设计好的表格.当然,如果没有相关记录,就要靠人工编写 VBA 程序来弥补.

如图 1-3-6 所示,在 Excel 工作表中,选择“工具/宏/Visual Basic 编辑器”.在 VB 编辑器窗口中选择“工具/宏”,在弹出的对话框中,在“宏名称”栏内输入宏的名称,如“抛掷硬币”,单击“创建”,出现宏主体语句 Sub 和 End Sub,输入你的程序后按 F5 即可运行.如不满意,可随时修改.

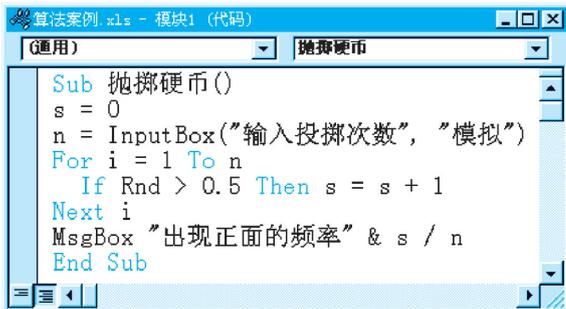


图 1-3-6

当抛掷次数为 10 000 时,可得出出现正面的频率为 0.494 4(你的模拟结果可能与此不同)。

习题 1.2

感受·理解

```

a ← 3
b ← -5
c ← 6
a ← b
b ← c
Print a, b, c
    
```

(第 1 题)

1. 在如图所示的算法中,最后输出的 a, b, c 各是多少?
2. 用算法语句给出用公式法求解方程 $x^2 - 3x - 4 = 0$ 的一个算法.
3. 输入 3 个数 a, b, c ,如果这 3 个数能作为一个三角形的三边长,那么输出 $\frac{1}{2}(a + b + c)$ 的值,否则提示重新输入.试用算法语句表示上述过程.
4. 某班有 50 名学生,现将他们某次测验的成绩分为 3 个等级:不低于 80 分为 A,低于 60 分为 C,其余为 B.试用算法语句表示输出每个学生相应的成绩等级的一个算法.
5. 写出求所有立方小于 1 000 的正整数的一个算法,并画出流程图,写出伪代码.
6. 设计计算 $1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{100}$ 的一个算法,并画出流程图,写出伪代码.

思考·运用

7. 用循环语句描述求 $1 - \frac{1}{2} + \frac{1}{3} - \frac{1}{4} + \dots + \frac{1}{99} - \frac{1}{100}$ 的一个算法.
8. 青年歌手大奖赛有 10 名选手参加,并请了 12 名评委.为了减少极端分数的影响,通常去掉一个最高分和一个最低分后再求平均分.请用算法语句表示:输入 12 名评委给某歌手所打的分数 $a_i (i = 1, 2, \dots, 12)$,用函数 $\text{Max}(a_1, a_2, \dots, a_{12})$ 和 $\text{Min}(a_1, a_2, \dots, a_{12})$ 分别求出 $a_i (i = 1, 2, \dots, 12)$ 中的最大值和最小值,最后输出该歌手的成绩.

探究·拓展

9. 通过计算机验证:存在自然数 n ,使 $1 + \frac{1}{2} + \frac{1}{3} + \dots + \frac{1}{n} > 10$.

本节我们选取了若干引人入胜的数学问题,通过对这些数学问题算法的探讨,可以进一步体会算法的思想,提高逻辑思维能力和算法设计水平.

案例 1 设计解决“韩信点兵—孙子问题”的算法.

韩信是秦末汉初的著名军事家.据说有一次汉高祖刘邦在卫士的簇拥下来到练兵场,刘邦问韩信有什么办法,不要逐个报数,就能知道场上士兵的人数.

韩信先令士兵排成 3 列纵队进行操练,结果有 2 人多余;接着他下令将队形改为 5 列纵队,这一改,又多出 3 人;随后他又下令改为 7 列纵队,这一次又剩下 2 人无法成整列.

在场的人都哈哈大笑,以为韩信无法清点出准确人数,不料笑声刚落,韩信便高声报告共有士兵 2 333 人.

众人听了一愣,不知韩信用什么办法这么快就能得出正确结果.

当然,韩信当时是否真是这样做的,已无从查考,但这个故事却引出一个著名的数学问题,即闻名世界的“孙子问题”.

这种神机妙算,最早出现在我国《算经十书》之一的《孙子算经》中.原文是:“今有物不知其数,三三数之剩二,五五数之剩三,七七数之剩二.问物几何?答曰:二十三.”

自从《孙子算经》中提出这个“物不知数”问题之后,它便引起了人们很大的兴趣.南宋数学家秦九韶对此加以推广,又发现了一种新的解法,叫“大衍求一术”.这种解法后来传入欧洲,欧洲学者发现此解法和高斯的解法本质上是一致的,但比高斯早了 500 余年.

所以,人们将这种问题的通用解法称为“孙子剩余定理”或“中国剩余定理”(Chinese remainder theorem).这个定理在近代数学和电子计算机程序设计中有广泛的应用.

■ 算法设计思想

“孙子问题”相当于求关于 x, y, z 的以下不定方程组的正整数解:

$$\begin{cases} m = 3x + 2, \\ m = 5y + 3, \\ m = 7z + 2. \end{cases}$$



秦九韶(1202~1261),四川安岳人.1247年写成名著《数书九章》,其中最重要的成就之一就是“大衍求一术”,即一次同余式组的解法.有兴趣的同学可上网查阅相关介绍.

Mod(a, b)表示
 a 除以 b 所得的余数,
称 b 为模.

设所求的数为 m , 根据题意, m 应同时满足下列 3 个条件:

- (1) m 被 3 除后余 2, 即 $\text{Mod}(m, 3) = 2$;
- (2) m 被 5 除后余 3, 即 $\text{Mod}(m, 5) = 3$;
- (3) m 被 7 除后余 2, 即 $\text{Mod}(m, 7) = 2$.

首先, 从 $m = 2$ 开始检验条件, 若 3 个条件中有任何 1 个不满足, 则 m 递增 1. 当 m 同时满足 3 个条件时, 输出 m .

流程图与伪代码

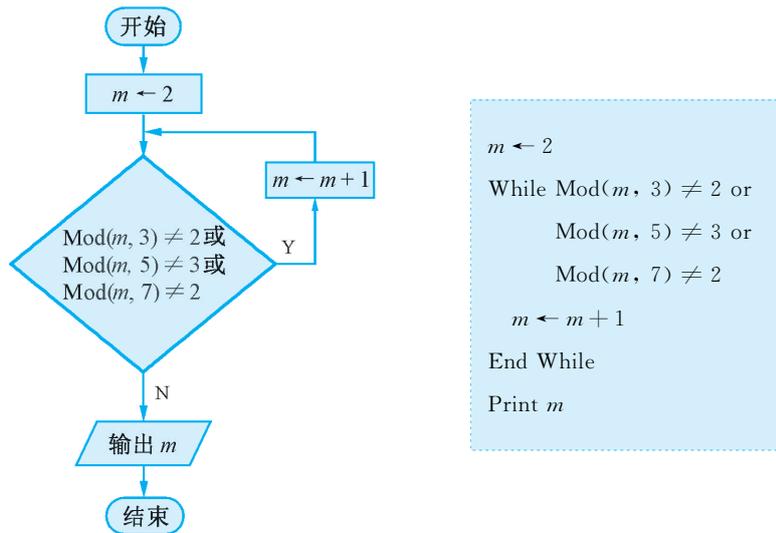


图 1-4-1

EXCEL

如图 1-4-2, 在 VB 编辑器窗口中创建名称为“孙子问题”的宏, 输入如下程序后按 F5 运行, 可以得到不定方程的一个解为 23.

符号“<>”表示
不等于.

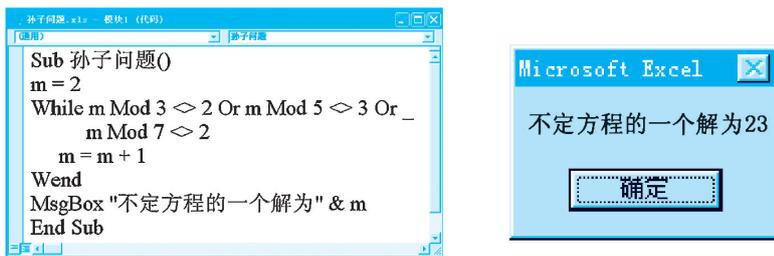


图 1-4-2

案例 2 写出求两个正整数 $a, b (a > b)$ 的最大公约数的一个算法.

公元前 3 世纪, 欧几里得在《原本》第七篇中介绍了求两个正整数 $a, b (a > b)$ 的最大公约数的方法, 即求出一列数:

$$a, b, r_1, r_2, \dots, r_{n-1}, r_n, 0.$$

这列数从第三项开始, 每一项都是前两项相除所得的余数, 余数



欧几里得(Euclid, 约公元前 330~公元前 275), 古希腊数学家, 以其所著的《原本》闻名于世。

为 0 的前一项 r_n , 即是 a 和 b 的最大公约数(greatest common factor). 这种方法称为“欧几里得辗转相除法”(Euclid algorithm).

例如, 求 $a = 204$ 与 $b = 85$ 的最大公约数的步骤为:

$$204 \div 85, \text{余数 } r_1 \text{ 为 } 34, \text{所以 } 204 = 85 \times 2 + 34;$$

$$85 \div 34, \text{余数 } r_2 \text{ 为 } 17, \text{所以 } 85 = 34 \times 2 + 17;$$

$$34 \div 17, \text{余数为 } 0, \text{所以 } 34 = 17 \times 2.$$

因此, 204 与 85 的最大公约数是 $r_2 = 17$.

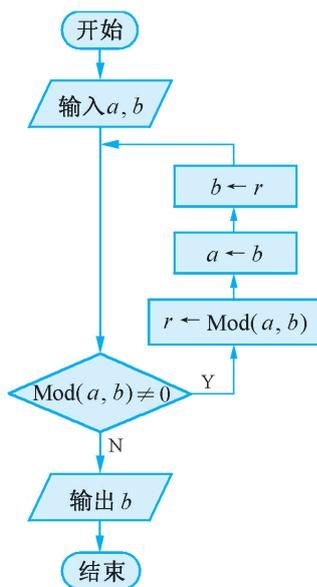
■ 算法设计思想

欧几里得辗转相除法求两个正整数 a, b 的最大公约数的步骤是: 计算出 $a \div b$ 的余数 r , 若 $r = 0$, 则 b 即为 a, b 的最大公约数; 若 $r \neq 0$, 则把前面的除数 b 作为新的被除数, 把余数 r 作为新的除数, 继续运算, 直到余数为 0, 此时的除数即为 a, b 的最大公约数.

求 $a, b(a > b)$ 的最大公约数的算法为:

- S1 输入两个正整数 a, b ;
- S2 如果 $\text{Mod}(a, b) \neq 0$, 那么转 S3, 否则转 S6;
- S3 $r \leftarrow \text{Mod}(a, b)$;
- S4 $a \leftarrow b$;
- S5 $b \leftarrow r$, 转 S2;
- S6 输出 b .

■ 流程图与伪代码



```

Read a, b
While Mod(a, b) ≠ 0
    r ← Mod(a, b)
    a ← b
    b ← r
End While
Print b
    
```

图 1-4-3

EXCEL

创建名称为“最大公约数”的宏，输入程序后按 F5 运行，先后输入两个自然数 204 和 85，可以得到这两个数的最大公约数为 17(如图 1-4-4)。



```

Sub 最大公约数()
a = InputBox("输入第一个自然数")
b = InputBox("输入第二个自然数")
While a Mod b <> 0
    r = a Mod b
    a = b
    b = r
Wend
MsgBox ("最大公约数为" & b)
End Sub
    
```

图 1-4-4

阅 读

辗转相除与更相减损

我国的《九章算术》卷一方田章中的“更相减损术”，与辗转相除法非常相似。书中说：“约分术曰：可半者半之，不可半者付置分母分子之数，以少减多，更相减损，求其等也。以等数约之。”意思是：如果分母、分子都是偶数(可半者)，那么先除以 2；如果不全是偶数，便将分子分母互减，以少减多，直到得出最大公约数为止；用最大公约数去约分子分母，便可使分数最简。

例如，求 49 与 91 的最大公约数，用更相减损术的步骤为：



陕西旬阳出土的唐代象牙算筹

第一步	将 49, 91 上下对齐；	<table border="1" style="border-collapse: collapse; text-align: center;"> <tr><td>49</td><td>≡≡≡</td><td>≡≡≡</td></tr> <tr><td>91</td><td>≡≡≡</td><td> </td></tr> </table>	49	≡≡≡	≡≡≡	91	≡≡≡	
49	≡≡≡	≡≡≡						
91	≡≡≡							
第二步	从 91 中减去 49，余 42；	<table border="1" style="border-collapse: collapse; text-align: center;"> <tr><td>49</td><td>≡≡≡</td><td>≡≡≡</td></tr> <tr><td>42</td><td>≡≡≡</td><td>≡≡</td></tr> </table>	49	≡≡≡	≡≡≡	42	≡≡≡	≡≡
49	≡≡≡	≡≡≡						
42	≡≡≡	≡≡						
第三步	从 49 中减去 42，余 7；	<table border="1" style="border-collapse: collapse; text-align: center;"> <tr><td>7</td><td></td><td>≡≡</td></tr> <tr><td>42</td><td>≡≡≡</td><td>≡≡</td></tr> </table>	7		≡≡	42	≡≡≡	≡≡
7		≡≡						
42	≡≡≡	≡≡						
第四步	从 42 中五次减去 7，出现上下两个相等的数 7。	<table border="1" style="border-collapse: collapse; text-align: center;"> <tr><td>7</td><td></td><td>≡≡</td></tr> <tr><td>7</td><td></td><td>≡≡</td></tr> </table>	7		≡≡	7		≡≡
7		≡≡						
7		≡≡						

于是，这个 7 就是“等数”，也就是最大公约数。由于我国古代用的是筹算，数目都是用算筹排成的，相减时只需取走或重新摆一下算筹即可。因此上述运算过程虽然是 4 步，摆出了 4 个筹式，但实际操作时只需在一个筹式上连续进行就可以了。

案例 3 写出用区间二分法求方程 $x^3 - x - 1 = 0$ 在区间 $[1, 1.5]$ 上的一个近似解(误差不超过 0.001)的一个算法.

算法设计思想

如图 1-4-5,如果估计出方程 $f(x) = 0$ 在某区间 $[a, b]$ 上有一个根 x^* ,就能用二分搜索求得符合误差限制 c 的近似解.算法步骤可以表示如下:

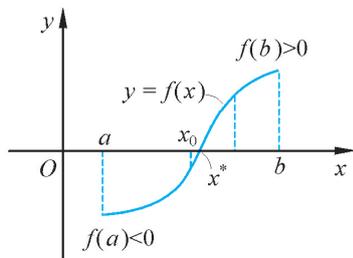


图 1-4-5

- S1 取 $[a, b]$ 的中点 $x_0 = \frac{1}{2}(a + b)$, 将区间一分为二;
- S2 若 $f(x_0) = 0$, 则 x_0 就是方程的根, 否则判断根 x^* 在 x_0 的左侧还是右侧:
 - 若 $f(a)f(x_0) > 0$, 则 $x^* \in (x_0, b)$, 以 x_0 代替 a ;
 - 若 $f(a)f(x_0) < 0$, 则 $x^* \in (a, x_0)$, 以 x_0 代替 b ;
- S3 若 $|a - b| < c$, 计算终止, 此时 $x^* \approx x_0$, 否则转 S1.

流程图与伪代码

“Exit Do”表示退出循环.

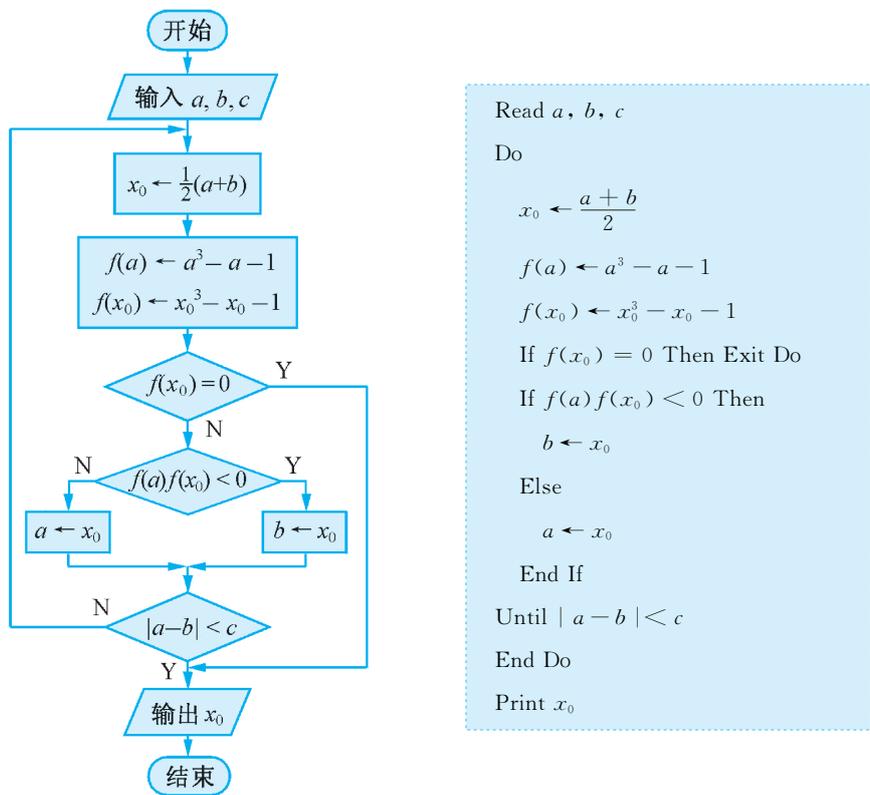


图 1-4-6

EXCEL

创建名称为“二分法”的宏,输入程序后按 F5 运行,依次输入区间左、右端点值 1, 1.5,再输入误差限制 0.001,得方程的近似解为 1.325 195 312 5,如图 1-4-7 所示.

函数 Val() 的作用是将字符串转换为数值,可以提高运算的正确性和精度.

```
Sub 二分法()
a = Val(InputBox("输入区间左端点值"))
b = Val(InputBox("输入区间右端点值"))
c = Val(InputBox("输入误差限制"))
Do
x0 = (a + b) / 2
f1 = a ^ 3 - a - 1
f2 = x0 ^ 3 - x0 - 1
If f2 = 0 Then Exit Do
If f1 * f2 < 0 Then
b = x0
Else
a = x0
End If
Loop Until Abs(a - b) < c
MsgBox "方程的近似解为" & x0
End Sub
```



图 1-4-7

练习

这里 Int(x) 表示不超过 x 的最大整数.

1. 下面一段伪代码的目的是().

```
Read m, n
While  $\frac{m}{n} \neq \text{Int}\left(\frac{m}{n}\right)$ 
 $c \leftarrow m - n \times \text{Int}\left(\frac{m}{n}\right)$ 
 $m \leftarrow n$ 
 $n \leftarrow c$ 
End While
Print n
```

(第 1 题)

- A. 求 m, n 的最小公倍数
 - B. 求 m, n 的最大公约数
 - C. 求 m 被 n 整除的商
 - D. 求 n 除以 m 的余数
2. 在直角坐标系中作出函数 $y = 2^x$ 和 $y = 4 - x$ 的图象,根据图象判断方程 $2^x = 4 - x$ 的解的范围,再用二分法求这个方程的近似解(误差不超过 0.001),写出这个算法的伪代码,并画出流程图.(有兴趣的同学可以上机操作)

习题 1.3

感受·理解

1. 已知一种放射性物质不断发生衰变,每经过 1 年剩余的质量约为原来的 84%. 那么,约经过多少年,剩余的质量是原来的一半? 试写出运用二分法计算这个近似解的伪代码.
2. 设计一个计算两个正整数 a, b 的最小公倍数的算法.

思考·运用

3. 我国古代劳动人民对不定方程的研究作出过重要贡献,其中《张丘建算经》中的“百鸡问题”就是一个很有影响的不定方程问题:今有鸡翁一值钱五,鸡母一值钱三,鸡雏三值钱一. 凡百钱买百只,问鸡翁、母、雏各几何.

其意思是: 1 只公鸡的价格是 5 钱, 1 只母鸡的价格是 3 钱, 3 只小鸡的价格是 1 钱. 想用 100 钱买 100 只鸡, 问公鸡、母鸡、小鸡各可买几只. 设 x, y, z 分别代表公鸡、母鸡、小鸡的只数, 我们可以大致确定 x, y, z 的取值范围: 若 100 钱全买公鸡, 则最多可买 20 只, 即 x 的取值范围是 $0 \sim 20$; 若 100 钱全买母鸡, 则最多可买 33 只, 即 y 的取值范围是 $0 \sim 33$; 当 x, y 在各自的取值范围内确定后, 小鸡的只数 $z = 100 - x - y$ 也就确定了.

根据上述算法思想, 画出求解的流程图, 并写出相应的伪代码.

4. 设计求解不定方程 $x_1 + 2x_2 + 3x_3 + 4x_4 = 0$ ($x_1, x_2, x_3, x_4 \in \{-1, 1\}$) 的一个算法. (提示: 可用循环语句或条件语句)

阅 读



莱布尼茨 (G. W. Leibniz, 1646~1716), 德国自然科学家、哲学家、数学家.

二进制数·计算机

二进制记数法的思想源远流长, 我国古代很早就有研究, 在《易经》上就讲到两仪, 即一黑一白阴阳互补的两条鱼. 以后, 在两仪之上形成了八卦. 《易经》中关于两仪及演变的叙述可以看成是二进制应用的萌芽.

德国数学家莱布尼茨 1679 年撰写的《二进制算术》, 使他成为二进位制数的发明人. 二进制在现代被应用于计算机设计, 但莱布尼茨本人并没有将它用到自己的计算机上. 莱布尼茨后来发现他的二进制可以给中国古老的六十四卦易图一个很好的数学解释, 他是通过他的朋友、法国传教士白晋 (F. J. Bouvet) 得到六十四卦易图的. 莱布尼茨高兴地说: “可以让我加入中国籍了吧!”



莱布尼茨发明的计算机



冯·诺伊曼(John von Neumann, 1903~1957), 美籍匈牙利数学家, 在计算机科学、计算机技术和数值分析等方面作了开拓性的工作。

莱布尼茨于1661年进入莱比锡大学学习,除了学习法律以外,还刻苦研究哲学和数学.他与牛顿几乎同时创立了微积分;在帕斯卡(Blaise Pascal, 1623~1662)加法机(加减法)的基础上,他还研制成功能够进行加、减、乘、除和开平方等运算的机械齿轮计算机,并于1673年在英国伦敦皇家学会上作了表演.

1946年,世界上第一台电子计算机ENIAC(埃尼阿克)诞生,这是科学技术发展史上一座新的里程碑.但是它还不完善.计算机之父冯·诺伊曼积极参与和研究之后,很快提出了改进意见.其中主要的两条对后来计算机科学技术的发展产生了深远影响:第一,用二进制替代原来的十进制(decimal system),这样大大减少了元器件数量,提高了运算速度;第二,存储程序,就是把程序像数据一样放在计算机内部的存储器中,这也就是后人所说的冯·诺伊曼计算机体系结构.此后,电子计算机在短短50多年的时间里得到了飞速发展,成为信息时代的骄子.

计算机为什么要采用二进制呢?

第一,二进制只有0和1两个数字,要得到表示两种不同稳定状态的电子器件很容易,而且制造简单,可靠性高.例如,电位的高与低,电容的充电与放电,晶体管的导通与截止,等等.

第二,在各种记数法中,二进制运算规则简单,有布尔逻辑代数作理论依据,简单的运算规则使得机器内部的操作也变得简单.

二进制加法法则只有4条:

$$\begin{aligned} 0+0 &= 0, 0+1 = 1, \\ 1+0 &= 1, 1+1 = 10, \end{aligned}$$

而十进制加法法则从 $0+0=0$ 到 $9+9=18$,有100条.

二进制的乘法法则也很简单:

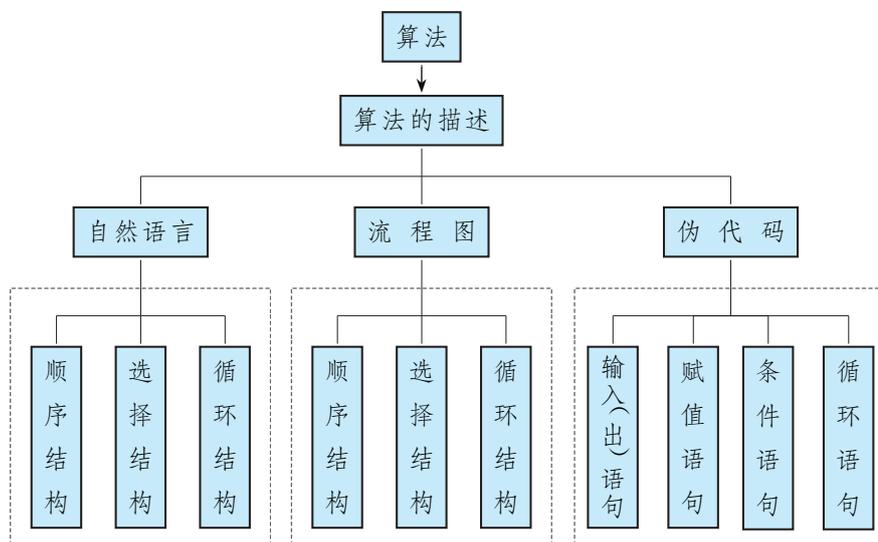
$$\begin{aligned} 0 \times 0 &= 0, 0 \times 1 = 0, \\ 1 \times 0 &= 0, 1 \times 1 = 1, \end{aligned}$$

而十进制的乘法法则要由一张“九九表”来规定,比较复杂.

本章回顾

本章概览

本章通过实例介绍了算法的含义,重点研究了在解决问题的过程中如何设计算法,如何根据算法画出流程图,并在此基础上逐步学会运用基本算法语句来表示算法过程。



算法的基本思想就是探求解决问题的一般性方法,并将解决问题的步骤用具体化、程序化的语言加以表述。

描述算法的方式经历了从简单的自然语言向高级的计算机程序语言的发展过程。自然语言通俗易懂,直接明了;流程图直观形象,能体现算法过程的结构特征;伪代码将流程图的各结构用接近计算机程序语言的算法语句进行表述,为编制计算机程序提供了便利。

内容提要

1. 算法含义

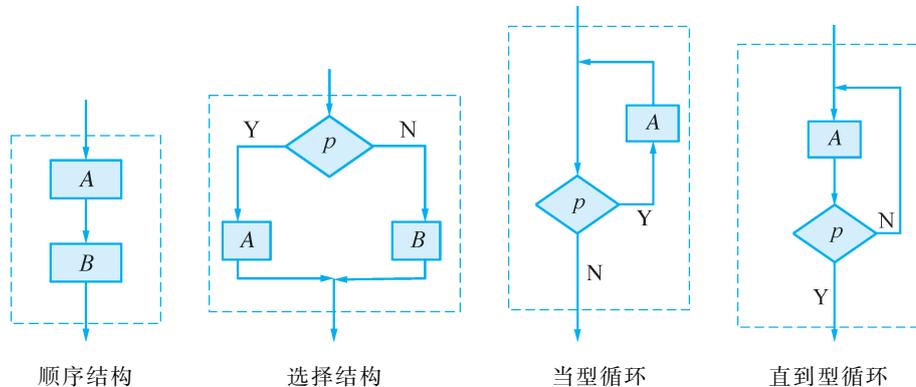
对一类问题的机械的、统一的求解方法称为算法。找到了某种算法,是指使用一系列运算规则能在有限步骤内求解某问题,其中的每条规则必须是明确定义的、可行的。

2. 流程图

是由一些图框和流程线组成的,其中图框表示各种操作的类型,图

框中的文字和符号表示操作的内容,流程线表示操作的先后次序。

- (1) 顺序结构:依次进行多个处理的结构。
- (2) 选择结构:先根据条件作出判断,再决定执行哪一种操作的结构。
- (3) 循环结构:需要重复执行同一操作的结构。



3. 基本算法语句

伪代码:介于自然语言和计算机语言之间的文字和符号。

- (1) 赋值语句:用符号 \leftarrow 表示,如“ $x \leftarrow y$ ”表示将 y 的值赋给 x ,其中 x 是一个变量, y 是一个与 x 同类型的变量或表达式。
- (2) 输入、输出语句
 输入语句:“Read a, b ”表示输入的数据依次送给 a, b ;
 输出语句:“Print x ”表示输出运算结果 x 。(支持多个输入和输出,但是中间要用逗号隔开)
- (3) 条件语句:

```

If A Then
    B
Else
    C
End If
    
```

- (4) 循环语句:

	While 循环	For 循环
对应当型循环	While p 循环体 End While	For I From “初值” To “终值” Step “步长” 循环体 End For
对应直到型循环	Do 循环	
	Do 循环体 Until p End Do	

复习题

感受·理解

1. 已知 $\odot O$, 写出求作 $\odot O$ 的圆心的一个算法.
2. 已知 $a, b \in \mathbf{N}^*$, 且 $a + b = 10$, 设计一个算法, 求出 ab 的最大值.
3. 已知函数 $f(x) = \frac{1}{x}$, 设计一个算法, 当 x 分别取 1.1, 1.01, 1.001, 1.000 1, 1.000 01 时, 计算 $A = \frac{f(x) - f(1)}{x - 1}$ 的值.
4. 先用不同的算法计算 $\frac{1}{1 \times 2} + \frac{1}{2 \times 3} + \frac{1}{3 \times 4} + \cdots + \frac{1}{99 \times 100}$, 再比较其优劣.
5. 已知 $\triangle ABC$ 中, $AB = c, BC = a, \angle B = \alpha$, 试写出求作 $\triangle ABC$ 的一个算法.
6. 用条件语句表示: 输入 x 的值, 通过下列表达式计算 y 的值:

```

I ← 1
While I < 8
    I ← I + 2
    S ← 2I + 3
End While
Print S
    
```

(第 7 题)

$$y = \begin{cases} -2x - 4, & x \in (-\infty, -2], \\ \sqrt{x + 2}, & x \in (-2, 2), \\ 2^{x-1}, & x \in [2, +\infty). \end{cases}$$

7. 根据如图所示的伪代码, 可知输出的结果 S 为().
A. 17 B. 19 C. 21 D. 23
8. 写出求 a_1, a_2, \dots, a_{100} 中最小数的一个算法.

思考·运用

9. 判断某年份是否为闰年, 要看此年份数能否被 4 整除. 若不能被 4 整除, 则是平年, 2 月是 28 天; 若能被 4 整除, 但不能被 100 整除, 则为闰年, 2 月是 29 天; 若能被 4 整除, 又能被 100 整除, 还要看能否被 400 整除, 若能则为闰年, 否则为平年. 画出上述算法的流程图, 并写出伪代码.
10. 函数 $y = x^2$ 与 $y = 2^x$ 的图象有 3 个交点 $(x_1, y_1), (2, 4), (4, 16)$, 其中 $-1 < x_1 < 0$. 试写出运用二分法求 x_1 的近似值的伪代码(误差不超过 0.01).
11. 收集身边的实际问题, 设计相应的算法, 并与同学交流.

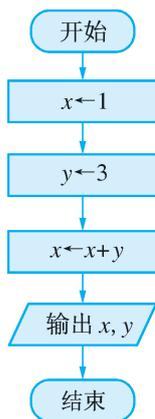
探究·拓展

12. 适合方程 $a^2 + b^2 = c^2$ 的一组正整数称为勾股数或商高数. 试设计一个满足 $a \leq 30, b \leq 40, c \leq 50$ 的勾股数的算法(给出算法步骤, 画出流程图).

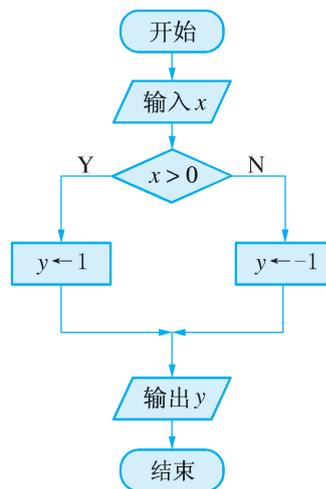
本章测试

一、填空题

1. 如图所示的流程图中,输出的结果是: $x = \underline{\hspace{2cm}}$, $y = \underline{\hspace{2cm}}$.



(第 1 题)



(第 2 题)

2. 如图所示的流程图表示的分段函数的解析式为_____.
3. 根据如图所示的伪代码,当输入的 a, b 分别为 2, 3 时,最后输出的 m 的值是_____.

```

Read a, b
If a > b Then
    m ← a
Else
    m ← b
End If
Print m
  
```

(第 3 题)

```

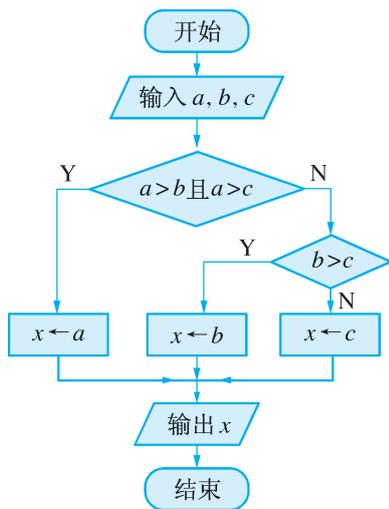
S ← 1
For I From 1 To 5 Step 2
    S ← S + I
End For
Print S
  
```

(第 4 题)

4. 执行如图所示的伪代码,输出的结果是_____.
5. 执行如图所示的伪代码,输出 S 的值为_____.
6. 如图给出的是计算 $1 + \frac{1}{2} + \frac{1}{3} + \frac{1}{4} + \dots + \frac{1}{10}$ 的值的一个流程图,其中判断框内可以填入的条件为_____.

三、解答题

11. 在如图所示的流程图中,求输入的 a, b, c 分别是 1, 2, 3 时输出的 x 的值.



(第 11 题)

12. 已知函数 $y = \begin{cases} x^2 - 1, & x < 2, \\ -x^2 + 1, & x \geq 2, \end{cases}$ 试写出“对确定的自变量 x 求对应的函数值 y ”的一个算法的伪代码.

13. 根据如图所示的伪代码,画出算法流程图.

```

S ← 0
i ← 1
While i ≤ 5
    S ← S + i
    i ← i + 1
End While
Print S
  
```

(第 13 题)

14. 已知函数 $f(x) = \frac{x}{1+x}$, 设 $f_1(x) = f(x)$, $f_2(x) = f(f_1(x))$, ..., $f_n(x) = f(f_{n-1}(x))$, 其中 n 为正整数. 设计求 $f_{100}(1)$ 的一个算法, 分别用流程图和伪代码表示.

15. 为了加强居民的节水意识, 某市制定了以下收费标准: 每户每月用水未超过 7 m^3 时, 按 1.0 元/m^3 收费, 并合计加收 0.2 元 的城市污水处理费; 若超过 7 m^3 , 7 m^3 以内的按 1.0 元/m^3 收费, 超过 7 m^3 的部分按 1.5 元/m^3 收费, 并合计加收 0.4 元 的城市污水处理费. 画出由输入用水量求应收水费的流程图.

第 2 章 统 计



[-] 统计

[-] 抽样方法

[-] 简单随机抽样

[+] 抽签法

[+] 随机数表法

[+] 系统抽样

[+] 分层抽样

[-] 总体分布的估计

[+] 频率分布表

[+] 频率分布直方图与折线图

[+] 茎叶图

[-] 总体特征数的估计

[+] 平均数及其估计

[+] 方差与标准差

[+] 线性回归方程

只有将数学应用于社会科学的研究之后,才能使得文明社会的发展成为可控制的现实.

——怀 特

灯泡厂要了解生产的灯泡的使用寿命,需要将所有灯泡逐一测试吗?

保险公司为对人寿保险制定适当的赔偿标准,需要了解人口的平均寿命,怎样获得相关数据?

国际奥委会 2003 年 6 月 29 日决定,2008 年北京奥运会的举办日期将比原定日期推迟两周,改在 8 月 8 日至 8 月 24 日举行.原因是 7 月末 8 月初北京地区的气温高于 8 月中下旬.这一结论是如何得到的呢?

.....

现实生活中存在大量类似的问题,如国家经济运行状况、人口普查、市场调查及天气预报等.

- 如何科学、合理地收集数据?
- 怎样分析和研究数据,对一般情况作出估计?

2.1

抽样方法

我们已经学习过一些统计知识,了解了统计的基本思想方法是用样本估计总体,即当总体容量很大或检测过程具有一定的破坏性时,不直接去研究总体,而是通过从总体中抽取一个样本,根据样本的情况去估计总体的相应情况.那么,

- 如何科学地进行抽样呢?

2.1.1 简单随机抽样

● 为了了解高一(1)班 50 名学生的视力状况,从中抽取 10 名学生进行检查.如何抽取呢?

1. 抽签法

通常使用抽签法,方法是:将 50 名学生从 1 到 50 进行编号,再制作 1~50 的 50 个号签,把 50 个号签集中在一起并充分搅匀,最后随机地从中抽出 10 个号签.对编号与抽中的号签的号码相一致的学生进行视力检查.

一般地,用抽签法从个体数为 N 的总体中抽取一个容量为 k 的样本的步骤为:

- (1) 将总体中的 N 个个体编号;
- (2) 将这 N 个号码写在形状、大小相同的号签上;
- (3) 将号签放在同一箱中,并搅拌均匀;
- (4) 从箱中每次抽出 1 个号签,连续抽取 k 次;
- (5) 将总体中与抽到的号签的编号一致的 k 个个体取出.

在学习了求概率的方法后,可以说明用抽签法能使每个个体被选中的机会相等.

这样就得到一个容量为 k 的样本.对个体编号时,可以利用已有的编号,如从全班学生中抽取样本时,利用学生的学号作为编号;对某场电影的观众进行抽样调查时,利用观众的座位号作为编号等.

抽签法简单易行,适用于总体中个体数不多的情形.

2. 随机数表法

用抽签法抽取样本时,编号的过程有时可以省略(如用已有的编号),但制签的过程就难以省去了,而且制签也比较麻烦.如何简化制签的过程呢?

链接

同桌的两位同学
相互协作,编制一张随
机数表.

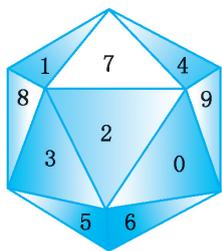


图 2-1-1

随机数表的制作

随机数表是人们根据需要编制出来的,由 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 这 10 个数字组成,表中每一个数都是用随机方法产生的.随机数的产生方法主要有抽签法、抛掷骰子法和计算机生成法.

(1) 抽签法:用 0, 1, 2, 3, 4, 5, 6, 7, 8, 9 这 10 个数字做 10 个签,放入一个箱中并搅拌均匀,再从箱中每次抽取一个签并记下签的数码,再放回箱中,如此重复进行下去即可得到一张随机数表.

如果需要两位数表,那么将所得的各个数码按顺序两两连在一起.类似地,如果需要三位数表,那么就三三连在一起,如 012, 321, 249, 460, 634, 105, ….

(2) 抛掷骰子法:如图 2-1-1,在一个正二十面体的各面写上 0~9 这 10 个数字(相对的两个面上的数字相同),这样就得到一个产生 0~9 的随机数的骰子.不断抛掷这个骰子,并逐一记下朝上一面(与地面或桌面平行)的数字,就能按顺序排成一个随机数表.

(3) 计算机生成法:利用随机函数或随机数发生器让计算机自动生成随机数表.

EXCEL & CALCULATOR

用 Excel 可以方便地产生所需的随机数,下面介绍两种常用的方法.

(1) 利用随机函数生成随机数.在单元格 C1 内输入随机函数“=RAND()”,就能得到一个 0 与 1 之间的随机数,拖曳 C1 的填充柄,便可产生不同的随机数(图 2-1-2 中第 C 列).Excel 有 65 536 行,利用 Excel 的自动填充功能,能快速生成大量的随机数.

如果要生成从 0 到 99 的随机数,且随机数为整数,那么可在单元格内输入“=INT(100 * RAND())”(图 2-1-2 中第 D 列).

(2) 利用随机数发生器.依次选择“工具/数据分析/随机数发生器”后,弹出如图 2-1-3 所示的对话框.它能根据你的要求产生所需的随机数.图 2-1-3 中的设置,表示在第 A 列生成 20 个 0 与 1 之间的随机数.

行	C	D
1	=RAND()	=INT(100 * RAND())
2	0.822902	79
3	0.984081	90
4	0.872911	97
5	0.074953	72
6	0.325	0
7	0.033718	89
8	0.351797	70
9	0.20506	25
10	0.10979	7
11	0.457856	68
12	0.811942	19
13	0.295412	80
14	0.452278	30
15	0.038423	54

填充柄

图 2-1-2

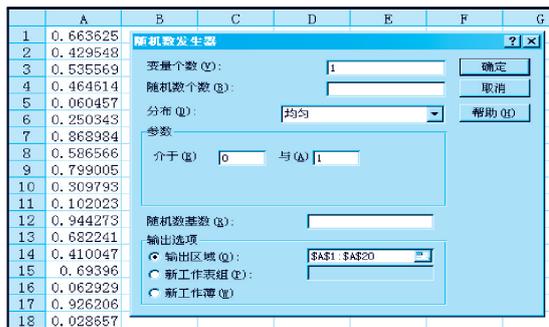


图 2-1-3

练习

1. 一个学生在一次知识竞赛中要回答的 8 道题是这样产生的：从 15 道历史题中随机抽出 3 道，从 20 道地理题中随机抽出 3 道，从 12 道生物题中随机抽出 2 道. 试用抽签法确定这个学生所要回答的 8 道题的序号(历史题编号分别为 1, 2, \dots , 15, 地理题编号分别为 16, 17, \dots , 35, 生物题编号分别为 36, 37, \dots , 47).
2. 从 100 件电子产品中抽取一个容量为 25 的样本进行检测, 试用随机数表法抽取样本.
3. 假设一个总体有 5 个元素, 分别记为 a, b, c, d, e , 采用逐个不放回抽取样本的方法, 从中抽取一个容量为 2 的样本, 这样的样本共有多少个? 写出全部可能的样本.
4. 某学校高一年级共有 200 名学生, 为了了解这些学生的身高状况, 试用简单随机抽样从中抽取一个容量为 15 的样本.

2.1.2 系统抽样

● 某校高一年级共有 20 个班, 每班有 50 名学生. 为了了解高一学生的视力状况, 从这 1 000 人中抽取一个容量为 100 的样本进行检查, 应该怎样抽样?

通常先将各班学生平均分成 5 组, 再在第一组(1~10 号学生)中用抽签法抽取一个, 然后按照“逐次加 10(每组中个体数)”的规则分别确定学号为 11~20、21~30、31~40、41~50 的另外 4 组中的学生代表.

将总体平均分成几个部分, 然后按照一定的规则, 从每个部分中抽取一个个体作为样本, 这样的抽样方法称为 **系统抽样**(systematic sampling).

假设要从容量为 N 的总体中抽取容量为 n 的样本, 系统抽样的步骤为:

- (1) 采用随机的方式将总体中的 N 个个体编号;
- (2) 将编号按间隔 k 分段, 当 $\frac{N}{n}$ 是整数时, 取 $k = \frac{N}{n}$; 当 $\frac{N}{n}$ 不是整数时, 从总体中剔除一些个体, 使剩下的总体中个体的个数 N' 能被 n 整除, 这时取 $k = \frac{N'}{n}$, 并将剩下的总体重新编号;
- (3) 在第一段中用简单随机抽样确定起始的个体编号 l ;
- (4) 按照一定的规则抽取样本, 通常将编号为 $l, l+k, l+2k, \dots, l+(n-1)k$ 的个体抽出.

系统抽样也可称为“等距抽样”。

例 1 某单位在岗职工共 624 人,为了调查工人用于上班途中的时间,决定抽取 10% 的工人进行调查.试采用系统抽样方法抽取所需的样本.

分析 因为 624 的 10% 约为 62,624 不能被 62 整除,所以为了保证“等距”分段,应先剔除 4 人.

解 第一步 将 624 名职工用随机方式进行编号;

第二步 从总体中剔除 4 人(剔除方法可用随机数表法),将剩下的 620 名职工重新编号(分别为 000, 001, 002, \dots , 619),并等距分成 62 段;

第三步 在第一段 000, 001, 002, \dots , 009 这 10 个编号中用简单随机抽样确定起始号码 l ;

第四步 将编号为 $l, l+10, l+20, \dots, l+610$ 的个体抽出,组成样本.

练习

- 为了了解参加一次知识竞赛的 1 252 名学生的成绩,如果决定采用系统抽样的方法抽取 1 个容量为 50 的样本,那么应从总体中随机剔除个体的数目是().
A. 2 B. 3 C. 4 D. 5
- 从 2 000 个编号中抽取 1 个容量为 20 的样本,若采用系统抽样的方法,则抽样的间隔为().
A. 200 B. 150 C. 100 D. 80
- 要从 1 003 名学生中抽取 1 个容量为 20 的样本,试叙述系统抽样的步骤.
- 试用系统抽样的方法从你校学生中抽取 1 个适当的样本,再对抽出学生的两臂平展的长度及身高进行测量,并分别计算两组数据的平均数.

2.1.3 分层抽样

● 某校高一、高二和高三年级分别有学生 1 000 名、800 名和 700 名,为了了解全校学生的视力情况,从中抽取容量为 100 的样本,怎样抽样较为合理?

由于不同年级的学生视力状况有一定的差异,不能在 2 500 名学生中随机抽取 100 名学生,也不宜在 3 个年级中平均抽取.为准确反映客观实际,不仅要使每个个体被抽到的机会相等,而且要注意总体中个体的层次性.

一个有效的办法是,使抽取的样本中各年级学生所占的比与实际人数占总体人数的比基本相同.

据此,应抽取高一学生 $100 \times \frac{1\,000}{2\,500} = 40$ 名,高二学生 $100 \times \frac{800}{2\,500} = 32$ 名,高三学生 $100 \times \frac{700}{2\,500} = 28$ 名.

一般地,当总体由差异明显的几个部分组成时,为了使样本更客观地反映总体情况,我们常常将总体中的个体按不同的特点分成层次比较分明的几个部分,然后按各个部分在总体中所占的比实施抽样,这种抽样方法叫**分层抽样**(stratified sampling),所分成的各个部分称为“层”.

分层抽样的步骤是:

- (1) 将总体按一定标准分层;
- (2) 计算各层的个体数与总体的个体数的比;
- (3) 按各层个体数占总体的个体数的比确定各层应抽取的样本容量;
- (4) 在每一层进行抽样(可用简单随机抽样或系统抽样).

若按比例计算所得的个体数不是整数,可作适当的近似处理.

例 2 某电视台在因特网上就观众对某一节目的喜爱程度进行调查,参加调查的总人数为 12 000 人,其中持各种态度的人数如表 2-1-1 所示.电视台为进一步了解观众的具体想法和意见,打算从中抽取 60 人进行更为详细的调查,应怎样进行抽样?

表 2-1-1

很喜爱	喜 爱	一 般	不喜爱
2 435	4 567	3 926	1 072

分析 因为总体中人数较多,所以不宜采用简单随机抽样.又因为持不同态度的人数差异较大,所以也不宜用系统抽样,而以分层抽样为妥.

解 可用分层抽样,其总体容量为 12 000.

“很喜爱”占 $\frac{2\,435}{12\,000}$,应取 $60 \times \frac{2\,435}{12\,000} \approx 12$ 人;

“喜爱”占 $\frac{4\,567}{12\,000}$,应取 $60 \times \frac{4\,567}{12\,000} \approx 23$ 人;

“一般”占 $\frac{3\,926}{12\,000}$,应取 $60 \times \frac{3\,926}{12\,000} \approx 20$ 人;

“不喜爱”占 $\frac{1\,072}{12\,000}$,应取 $60 \times \frac{1\,072}{12\,000} \approx 5$ 人.

因此,采用分层抽样的方法在“很喜爱”、“喜爱”、“一般”和“不喜爱”的 2 435 人、4 567 人、3 926 人和 1 072 人中分别抽取 12 人、23 人、20 人和 5 人.

练习

1. 在某年有奖明信片销售活动中,规定每 100 万张为 1 个开奖组,通过随机抽取的方式确定号码的后 4 位数是 2 709 的为三等奖.这样确定获奖号码的抽样方法是_____.
2. 某公司生产 3 种型号的轿车,产量分别为 1 200 辆、6 000 辆和 2 000 辆.为检验该公司的产品质量,现用分层抽样的方法抽取 46 辆进行检验,这 3 种型号的轿车应分别抽取_____辆、_____辆和_____辆.
3. 某工厂生产 A, B, C 3 种不同型号的产品,产量之比为 2 : 3 : 5.现用分层抽样的方法抽取 1 个容量为 n 的样本,若样本中 A 种型号的产品有 16 件,则样本容量 $n =$ _____.
4. 某所学校有小学部、初中部和高中部,在校小学生、初中生和高中生人数之比为 5 : 2 : 3,且已知初中生有 800 人.现要从这所学校中抽取 1 个容量为 80 的样本以了解他们对某一问题的看法,应采用什么抽样方法?从小学部、初中部及高中部各抽取多少名学生?总体上看,平均多少名学生中抽取到 1 名学生?
5. 将你所在班级的同学按性别分成两组,分别编号,制成号签,分别放在两个箱子里搅拌均匀,然后按男女生之比各抽出若干个号签,组成两个样本,就他们对某一问题的看法进行调查,以比较男女同学对该问题看法的差异.

以上我们学习了 3 种抽样方法,这些抽样方法的特点及适用范围可归纳如下:

表 2-1-2

类别	特 点	相互联系	适用范围	共同点
简单随机抽样	从总体中逐个抽取		总体中的个体数较少	抽样过程中每个个体被抽到的可能性相同
系统抽样	将总体平均分成几部分,按一定的规则分别在各部分中抽取	在起始部分抽样时,采用简单随机抽样	总体中的个体数较多	
分层抽样	将总体分成几层,按各层个体数之比抽取	各层抽样时采用简单随机抽样或系统抽样	总体由差异明显的几部分组成	

例 3 下列问题中,采用怎样的抽样方法较为合理?

(1) 从 10 台冰箱中抽取 3 台进行质量检查;

(2) 某电影院有 32 排座位,每排有 40 个座位,座位号为 1~40. 有一次报告会坐满了听众,报告会结束以后为听取意见,需留下 32 名听众进行座谈;

(3) 某学校有 160 名教职工,其中教师 120 名,行政人员 16 名,后勤人员 24 名. 为了了解教职工对学校在校务公开方面的意见,拟抽取一个容量为 20 的样本.

分析 (1) 总体容量比较小,用抽签法或随机数表法都很方便.

(2) 总体容量比较大,用抽签法或随机数表法比较麻烦. 由于人员没有明显差异,且刚好 32 排,每排人数相同,可用系统抽样.

(3) 由于学校各类人员对这一问题的看法可能差异较大,故应采用分层抽样.

解 (1) 用抽签法或随机数表法.

(2) 用系统抽样. 将每排的 40 个人组成一组,共 32 组,从第 1 排至第 32 排分别为第 1~32 组. 先在第 1 排用简单随机抽样抽出 1 名听众,再将其他各排与此听众座位号相同的听众全部取出.

(3) 用分层抽样. 总体容量为 160,故样本中教师人数应为 $20 \times \frac{120}{160} = 15$ 名,行政人员人数应为 $20 \times \frac{16}{160} = 2$ 名,后勤人员人数应为 $20 \times \frac{24}{160} = 3$ 名.

思 考

某城市的两所中学分别对自己学校 12~14 岁学生的身高进行了抽样统计,发现这两所学校 12~14 岁学生的平均身高竟相差 19 cm,这可能吗? 他们在抽样的过程中可能出现了哪些问题?

练习

- 某商场想通过检查发票及销售记录的 2% 来快速估计每月的销售总额, 采取如下方法: 从某本 50 张的发票存根中随机抽取 1 张, 如 15 号, 然后按顺序往后将 65 号、115 号、165 号……发票上的销售额组成 1 个调查样本. 这种抽取样本的方法是().
 - 抽签法
 - 系统抽样
 - 分层抽样
 - 随机数表法
- 某科研机构由科技人员、行政人员和后勤职工 3 种不同类型的人员组成, 现要抽取 1 个容量为 45 的样本进行调查. 已知科技人员共有 60 人, 抽入样本的有 20 人, 且行政人员与后勤职工人数之比为 2 : 3, 那么此机构的总人数、行政人员、后勤职工人数分别为多少?
- 一个单位有职工 160 人, 其中业务人员 96 人, 管理人员 40 人, 后勤人员 24 人. 为了了解职工的某种情况, 要从中抽取 1 个容量为 20 的样本, 按下述 3 种方法抽取:
 - 将 160 人按 1~160 编号, 用白纸做成有 1~160 号的签放入箱内搅匀, 然后从中抽取 20 个签, 与签号相同的 20 个人被选出;
 - 将 160 人按 1~160 编号, 按编号顺序分成 20 组, 每组 8 人, 号码分别为 1~8, 9~16, ..., 153~160, 先从第 1 组中用抽签法抽出 k ($0 < k < 9$) 号, 再抽取其余组的 $(k+8n)$ ($n=1, 2, \dots, 19$) 号, 如此抽取 20 人;
 - 按 $20 : 160 = 1 : 8$ 的比例, 从业务人员中抽取 12 人, 从管理人员中抽取 5 人, 从后勤人员中抽取 3 人, 都用随机数表法从各类人员中抽取所需的人数, 他们合在一起恰好抽取 20 人.
 上述 3 种抽样方法中, 按简单随机抽样、分层抽样、系统抽样的顺序是().
 - ①、②、③
 - ②、①、③
 - ①、③、②
 - ③、①、②
- 下列抽样中, 不是系统抽样的是().
 - 从号码为 1~15 的 15 个球中任选 3 个作为样本, 先在 1~5 号球中用抽签法抽出 l 号, 再将号码为 $l+5, l+10$ 的球也抽出
 - 某工厂生产产品, 用传送带将产品送入包装车间的过程中, 检验人员从传送带上每隔 5 min 抽 1 件产品进行检验
 - 某项市场调查, 规定在商店门口随机地询问 1 个人, 直至达到事先规定的调查人数为止
 - 电影院调查观众的某一指标, 邀请每排(每排人数相等)座位号为 14 的观众留下来座谈

习题 2.1

感受·理解

1. 为了了解某市 800 家企业的管理情况,拟抽取 40 家企业作为样本. 这 800 家企业中有中外合资企业 160 家, 私营企业 320 家, 国有企业 240 家, 其他性质的企业 80 家. 如何抽样较合理?
2. 用系统抽样方法从 400 名学生中抽取容量为 20 的样本, 将 400 名学生随机地编号为 1~400, 按编号顺序平均分为 20 个组(1~20 号, 21~40 号, …… 381~400 号). 若第 1 组中用抽签的方法确定抽出的号码为 11, 则第 3 组抽取的号码为_____.
3. 用分层抽样的方法从某校学生中抽取 1 个容量为 45 的样本, 其中高一一年级抽 20 人, 高三年级抽 10 人. 已知该校高二年级共有学生 300 人, 求该校学生总数.
4. 某市的 4 个区共有 20 000 名学生, 且 4 个区的学生人数之比为 3 : 2.8 : 2.2 : 2. 如果要用分层抽样的方法从所有学生中抽取 1 个容量为 200 的样本, 那么在这 4 个区中分别应抽取多少名学生?
5. 用适当的方法对你校高一学生的体重和身高进行抽样调查, 并将数据收集整理, 以备进一步分析.
6. 用系统抽样的方法, 对你校高一到高三学生的身高和心率进行抽样调查, 并制成表格.
7. 某总体的容量为 100, 其中带有标记的有 60 个. 现用简单随机抽样的方法从中抽出 1 个容量为 20 的样本, 试估计抽取到带有标记的个体个数.

思考·运用

8. 用适当的方法, 对你所在学校的学生进行抽样调查, 将其父亲、母亲的年龄收集整理, 并用表格表示出来.
9. 有关方面就男女职工是否同龄退休广泛征求意见, 请对你班学生进行抽样调查, 了解他们对这一问题的看法.
10. 某校高一年级 500 名学生中, 血型为 O 型的有 200 人, A 型的有 125 人, B 型的有 125 人, AB 型的有 50 人. 为了研究血型与色弱之间的关系, 要从中抽取 1 个容量为 40 的样本, 应如何抽样? 写出 AB 血型样本的抽样过程.
11. 举例说明各种抽样方法在实际生活中的应用.

探究·拓展

12. 请你就中学生普遍关心的某一问题, 通过网上调查和对本校学生进行抽样调查(有条件的话可以扩大调查范围), 了解学生对此问题的看法, 并对两种调查方法所得结果进行分析比较.

2.2

总体分布的估计

为了了解7月25日至8月24日北京地区的气温分布状况,可以对北京往年这段时间的日最高气温进行抽样,并对得到的数据进行分析.我们随机抽取近年来北京地区7月25日至8月24日的日最高气温,得到如下两个样本(单位: $^{\circ}\text{C}$):

表 2-2-1

7月25日至 8月10日	41.9	37.5	35.7	35.4	37.2	38.1	34.7	33.7	33.3
	32.5	34.6	33.0	30.8	31.0	28.6	31.5	28.8	
8月8日至 8月24日	28.6	31.5	28.8	33.2	32.5	30.3	30.2	29.8	33.1
	32.8	29.4	25.6	24.7	30.0	30.1	29.5	30.3	

● 怎样通过上表中的数据,分析比较两时间段内的高温天气(日最高气温 $\geq 33^{\circ}\text{C}$)状况呢?

2.2.1 频率分布表

上面两个样本中的高温天气的频率可以用下表表示:

表 2-2-2

时 间	总天数	高温天气频数	频 率
7月25日至8月10日	17	11	0.647
8月8日至8月24日	17	2	0.118

由此表可以发现,近年来,北京地区7月25日至8月10日的高温天气的频率明显高于8月8日至8月24日的频率.

上例说明,当总体很大或不便于获得时,可以用样本的频率分布估计总体的频率分布.我们把反映总体频率分布的表格称为**频率分布表**(frequency distribution table).

初中数学已经介绍了历史上所做的抛掷硬币的试验,并获得了正面向上或反面向上的频率分布表.这类试验只有两种结果,比较简单.下面研究较为复杂的频率分布表的制作方法.

例 1 从某校高一年级的 1 002 名新生中用系统抽样的方法抽取 1 个容量为 100 的身高样本,数据(单位: cm)如下. 试作出该样本的频率分布表.

表 2-2-3

168	165	171	167	170	165	170	152	175	174
165	170	168	169	171	166	164	155	164	158
170	155	166	158	155	160	160	164	156	162
160	170	168	164	174	171	165	179	163	172
180	174	173	159	163	172	167	160	164	169
151	168	158	168	176	155	165	165	169	162
177	158	175	165	169	151	163	166	163	167
178	165	158	170	169	159	155	163	153	155
167	163	164	158	168	167	161	162	167	168
161	165	174	156	167	166	162	161	164	166

分析 该组数据中最小值为 151,最大值为 180,它们相差 29,可取区间 $[150.5, 180.5]$,并将此区间分成 10 个小区间,每个小区间长度为 3,再统计出每个区间内的频数并计算相应的频率. 我们将整个取值区间的长度称为**全距**,分成的区间的长度称为**组距**.

解 (1) 在全部数据中找出最大值 180 和最小值 151,两者之差为 29,确定全距为 30,决定以组距 3 将区间 $[150.5, 180.5]$ 分成 10 组;

(2) 从第 1 组 $[150.5, 153.5)$ 开始,分别统计各组中的频数,再计算各组的频率,并将结果填入下表:

表 2-2-4

分 组	频数累计	频 数	频 率
$[150.5, 153.5)$	4	4	0.04
$[153.5, 156.5)$	12	8	0.08
$[156.5, 159.5)$	20	8	0.08
$[159.5, 162.5)$	31	11	0.11
$[162.5, 165.5)$	53	22	0.22
$[165.5, 168.5)$	72	19	0.19
$[168.5, 171.5)$	86	14	0.14
$[171.5, 174.5)$	93	7	0.07
$[174.5, 177.5)$	97	4	0.04
$[177.5, 180.5]$	100	3	0.03
合 计		100	1

除最后一个区间是闭区间外,其他区间均为左闭右开区间. 称区间的左端点为下组限,右端点为上组限. 此处采用下组限在内而上组限不在内的分组方法,当然也可采用下组限不在内而上组限在內的分组方法.

这张表给出了该身高样本处于各个区间内的人数和频率,由此可以估计该校高一学生的身高的分布状况。

一般地,制作频率分布表的步骤如下:

(1) 求全距,决定组数和组距,组距 = $\frac{\text{全距}}{\text{组数}}$;

(2) 分组,通常对组内数值所在区间取左闭右开区间,最后一组取闭区间;

(3) 登记频数,计算频率,列出频率分布表。

如果取全距时不利于分组(如不能被组数整除),那么可以适当增大全距,如在左、右两端各增加适当范围(尽量使两端增加的量相同)。

思考

在编制频率分布表时,分的组数过少或过多各有何利弊?

EXCEL

在例 1 中,用 Excel 制作频率分布表的操作步骤如下:

(1) 将数据输入 Excel 工作表中(数据区 A1: J10),在空白单元格插入函数“=MAX(A1: J10)”和“=MIN(A1: J10)”可求出数据区域内的最大、最小值分别为 180, 151(图 2-2-1);

=MAX(A1:J10)		
I	J	K
175	174	180
164	158	151
156	162	
163	172	=Min(A1:J10)
164	169	
169	162	
163	167	
153	155	
167	168	
164	166	

图 2-2-1

M2	=COUNTIF(A1:J10,"<153.5")			
	L	M	N	O
1	分组	频数累计	频数	频率
2	[150.5, 153.5)	4	4	=N2/100
3	[153.5, 156.5)	12	8	=M3-M2
4	[156.5, 159.5)	20	8	0.08
5	[159.5, 162.5)	31	11	0.11
6	[162.5, 165.5)	53	22	0.22
7	[165.5, 168.5)	72	19	0.19
8	[168.5, 171.5)	86	14	0.14
9	[171.5, 174.5)	93	7	0.07
10	[174.5, 177.5)	97	4	0.04
11	[177.5, 180.5]	100	3	0.03

图 2-2-2

COUNTIF()为条件计数函数,前一项为计数对象所在的区域,后一项为计数的条件,结果为符合该条件的数据的个数。

(2) 确定全距为 30,组距为 3,将区间[150.5, 180.5]分成 10 组.用函数 COUNTIF()计算频数累计,如计算身高小于 165.5 的人数,只需输入“=COUNTIF(A1: J10, "<165.5)”;

(3) 计算频数和频率.用该组的累计频数减去上一个组的累计频数即得该组的频数(在单元格 N2 中输入 4,在单元格 N3 中输入“=M3-M2”,再双击 N3 的填充柄即完成频数的计算).将频数除以 100 即得频率(在单元格 O2 中输入“N2/100”,再双击 N2 的填充柄便完成全部计算),如图 2-2-2 所示。

练习

1. 下面是不同厂家生产的手提式电脑的重量(单位: kg)数据,试选取适当的组距,列出其频率分布表。

1.9 2.0 2.1 2.4 2.4 2.8 3.2 2.3 1.5 2.6
2.6 1.9 2.4 2.2 1.6 1.7 1.7 1.8 1.8 3.0

2. 一个容量为 20 的数据样本,已知分组与频数如下: [10, 20), 2 个; [20, 30), 3 个; [30, 40), 4 个; [40, 50), 5 个; [50, 60), 4 个; [60, 70], 2 个. 那么,样

本数据在区间 $(-\infty, 50)$ 上的可能性为()。

- A. 5% B. 25% C. 50% D. 70%

3. 容量为 100 的样本数据,按照从小到大的顺序分为 8 组,如下表所示. 请计算第 3 组的频数和频率.

组 数	1	2	3	4	5	6	7	8
频 数	10	13	x	14	15	13	12	9

“组中值”为各个组区间的中点的数值.

4. 在一本书中,分组统计 100 个句子中的字数,得出下列结果: 字数 1~5 个的 15 句,字数 6~10 个的 27 句,字数 11~15 个的 32 句,字数 16~20 个的 15 句,字数 21~25 个的 8 句,字数 26~30 个的 3 句. 请作出字数的频率分布表,并利用组中值对该书中平均每个句子包含的字数作出估计.

自己找一本书,对书中句子所含字数进行抽样估计.

2.2.2 频率分布直方图与折线图

我们还学过一种更为直观地体现数据分布规律的方法——绘制**频数条形图**或**频率直方图**.

- 例 2** 下表是某学校 1 个星期中收交来的失物件数,请将 5 天中收交来的失物数用条形图表示.

表 2-2-5

星 期	一	二	三	四	五
件 数	6	2	3	5	1
累 计	6	8	11	16	17

解 我们用 Excel 作条形图:

- (1) 在 Excel 工作表中输入数据,光标停留在数据区中;
- (2) 选择“插入/图表”,在弹出的对话框中点击“柱形图”(bar graph);
- (3) 点击“完成”(如图 2-2-3).

试作出第 2.2 节开头提出的北京地区 7 月 25 日至 8 月 10 日与 8 月 8 日至 8 月 24 日两时间段的高温天数的频数条形图.

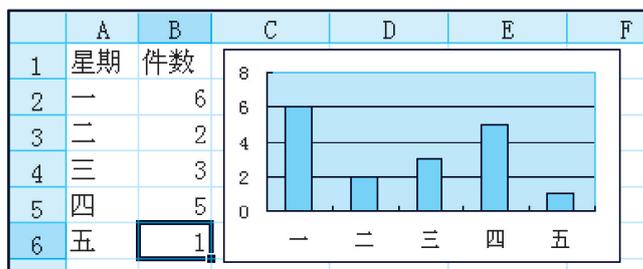


图 2-2-3

我们可以利用直方图反映样本的频率分布规律,这样的直方图称为**频率分布直方图**,简称**频率直方图**.下面用例 1 中的数据加以说明.

例 3 作出例 1 中数据的频率分布直方图.

解 (1) 先制作频率分布表, 然后作直角坐标系, 以横轴表示身高, 纵轴表示 $\frac{\text{频率}}{\text{组距}}$;

(2) 在横轴上标上表示 150.5, 153.5, 156.5, \dots , 180.5 的点 (为方便起见, 起始点 150.5 可适当前移);

(3) 在上面标出的各点中, 分别以连结相邻两点的线段为底作矩形, 高等于该组的 $\frac{\text{频率}}{\text{组距}}$.

至此, 就得到了这组数据的频率分布直方图, 如图 2-2-4 所示.

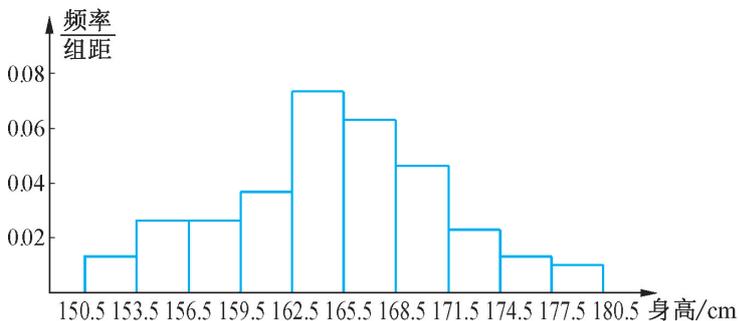


图 2-2-4

一般地, 作频率分布直方图的方法为:

把横轴分成若干段, 每一线段对应 1 个组的组距, 然后以此线段为底作矩形, 它的高等于该组的 $\frac{\text{频率}}{\text{组距}}$, 这样得出一系列的矩形, 每个矩形的面积恰好是该组的频率. 这些矩形就构成了频率分布直方图.

频率直方图比频率分布表更直观、形象地反映了样本的分布规律, 如在 164 附近达到“峰值”, 并具有一定的对称性, 这说明该批学生的身高在 164 cm 附近较为集中. 另外还可看出, 特别高和特别矮的学生都很少.

如果将频率分布直方图中各个相邻的矩形的上底边的中点顺次连结起来, 那么就得到**频率分布折线图**, 简称频率折线图.

例 3 的频率直方图 2-2-4 中, 取各相邻矩形的上底边中点并顺次连结, 再将矩形的边擦去, 就得到如图 2-2-5 所示的频率折线图.

取值区间两端点须分别向外延伸半个组距, 并取此组距上在 x 轴上的点与折线的首、尾分别相连.

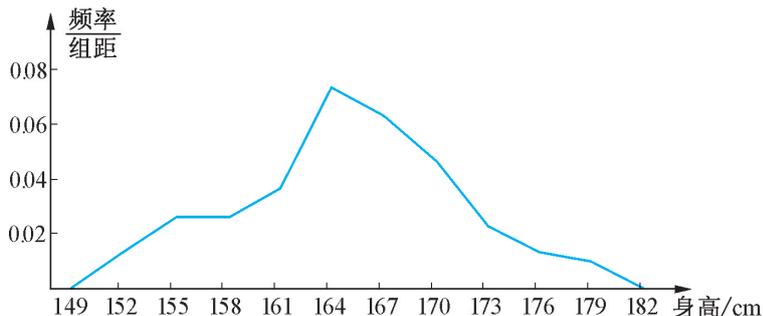


图 2-2-5

思考

不用频率分布直方图,如何直接作频率折线图?比较几种表示频率分布的方法,看看各有哪些优点和不足.

频率折线图的优点是它反映了数据的变化趋势.如果将样本容量取得足够大,分组的组距取得足够小,那么相应的频率折线图将趋于一条光滑曲线,我们称这条光滑曲线为总体分布的**密度曲线**.

如例 3 的频率分布密度曲线可以近似地表示为图 2-2-6.

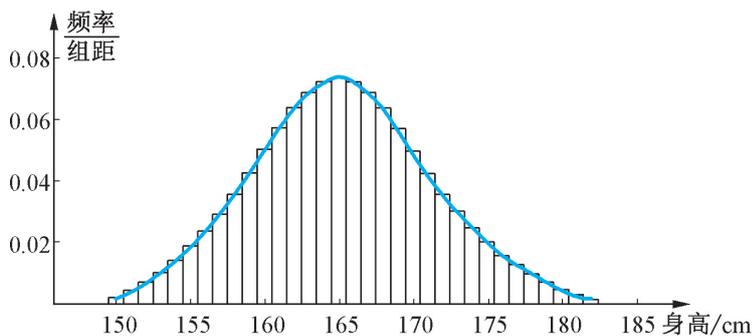


图 2-2-6

例 4 为了了解一大片经济林的生长情况,随机测量其中 100 株树木的底部周长,得到如表 2-2-6 所示的数据(单位: cm):

表 2-2-6

135	98	102	110	99	121	110	96	100	103
125	97	117	113	110	92	102	109	104	112
109	124	87	131	97	102	123	104	104	128
105	123	111	103	105	92	114	108	104	102
129	126	97	100	115	111	106	117	104	109
111	89	110	121	80	120	121	104	108	118
129	99	90	99	121	123	107	111	91	100
99	101	116	97	102	108	101	95	107	101
102	108	117	99	118	106	119	97	126	108
123	119	98	121	101	113	102	103	104	108

- (1) 编制频率分布表;
- (2) 绘制频率分布直方图;
- (3) 估计该片经济林中底部周长小于 100 cm 的树木约占多少,底部周长不小于 120 cm 的树木约占多少.

解 (1) 从表中可以看出,这组数据的最大值为 135,最小值为 80,故全距为 55,可将其分为 11 组,组距为 5.

从第1组[80, 85)开始,将各组的频数、频率和 $\frac{\text{频率}}{\text{组距}}$ 填入下表中.

表 2-2-7

分 组	频 数	频 率	$\frac{\text{频率}}{\text{组距}}$
[80, 85)	1	0.01	0.002
[85, 90)	2	0.02	0.004
[90, 95)	4	0.04	0.008
[95, 100)	14	0.14	0.028
[100, 105)	24	0.24	0.048
[105, 110)	15	0.15	0.030
[110, 115)	12	0.12	0.024
[115, 120)	9	0.09	0.018
[120, 125)	11	0.11	0.022
[125, 130)	6	0.06	0.012
[130, 135]	2	0.02	0.004
合 计	100	1	0.2

(2) 这组数据的频率直方图如图 2-2-7 所示.

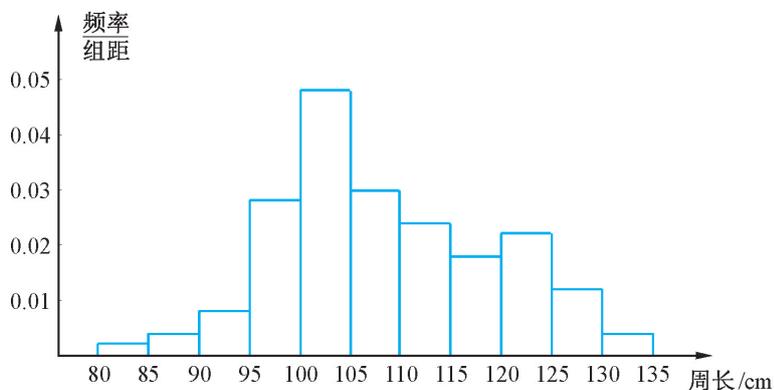


图 2-2-7

(3) 从频率分布表可以看出,该样本中小于 100 的频率为 $0.01 + 0.02 + 0.04 + 0.14 = 0.21$, 不小于 120 的频率为 $0.11 + 0.06 + 0.02 = 0.19$, 故可估计该片经济林中底部周长小于 100 cm 的树木约占 21%, 底部周长不小于 120 cm 的树木约占 19%.

练 习

- 作出习题 2.1 第 5 题和第 6 题中数据的频率直方图和频率折线图.
- 一组数据的频率分布直方图中,所有小长方形的面积总和为_____.
- 为了了解一批灯泡(共 10 000 只)的使用寿命,从中抽取了 100 只进行测试,其使用寿命(单位: h)如下表:

使用寿命	[500, 600)	[600, 700)	[700, 800)	[800, 900)	[900, 1 000)
只 数	1	4	8	15	20
使用寿命	[1 000, 1 100)	[1 100, 1 200)	[1 200, 1 300)	[1 300, 1 400)	[1 400, 1 500]
只 数	24	18	7	2	1

- (1) 制作频率分布表；
- (2) 绘制频率分布直方图；
- (3) 根据样本的频率分布,估计使用寿命不低于 1 000 h 的灯泡约有多少只.

2.2.3 茎叶图

某篮球运动员在某赛季各场比赛的得分情况如下：

12, 15, 24, 25, 31, 31, 36, 36, 37, 39, 44, 49, 50.

如何分析该运动员的整体水平及发挥的稳定程度？

除了上面介绍的几种图、表能帮助我们分析数据外,统计中还有一种用来表示数据的**茎叶图**(stem and leaf display).

在上述问题中,将所有两位数的十位数字作为“茎”,个位数字作为“叶”,茎相同者共用一个茎,茎按从小到大的顺序从上向下列出,共茎的叶一般按从小到大(或从大到小)的顺序同行列出,这样就得到该运动员得分的茎叶图(图 2-2-8).

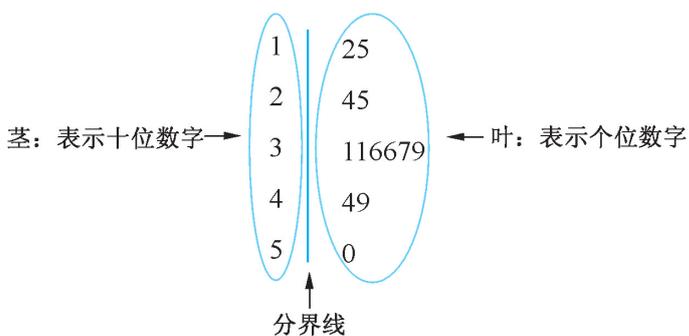


图 2-2-8

图 2-2-8 中第一行分界线左侧的“1”表示十位数字,右侧的“2”和“5”表示个位数字,这一行说明该运动员的得分为 12 分和 15 分.同理,第二行说明得分为 24 分和 25 分,第三行说明有两个 31 分,两个 36 分,一个 37 分,一个 39 分,依此类推.

从这张图可以直观地看出,该运动员平均得分大约在 20~40 之间,且分布较对称,集中程度高,说明其发挥比较稳定.

例 5 甲、乙两名篮球运动员上个赛季每场比赛的得分如下,试比较这两名运动员的得分水平.

甲 12, 15, 24, 25, 31, 31, 36, 36, 37, 39, 44, 49, 50.

乙 8, 13, 14, 16, 23, 26, 28, 33, 38, 39, 51.

解 画出两人得分的茎叶图,为便于对比分析,可将茎放在中间共用,叶分居左、右两侧(图 2-2-9):

左侧的叶按从大到小的顺序写,右侧的叶按从小到大的顺序写.相同的得分要重复记录,不能遗漏.

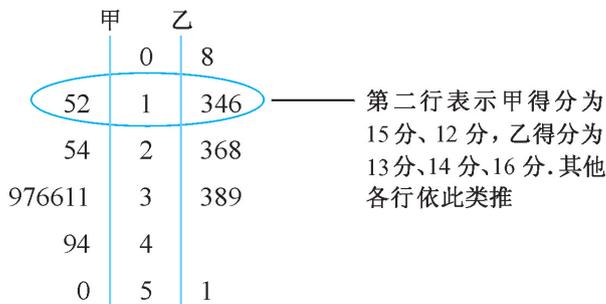


图 2-2-9

从这个茎叶图可以看出,甲运动员的得分大致对称,平均得分大约是 30 多分.乙运动员的得分除一个 51 分外,也大致对称,平均得分大约是 20 多分.因此甲运动员发挥比较稳定,总体得分情况比乙好.

例 5 说明,茎叶图既可以用于分析单组数据,也可以用于对两组数据进行比较分析.

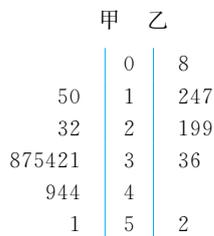
用茎叶图刻画数据有两个优点:一是所有的信息都可以从这张茎叶图中得到,二是茎叶图便于记录和表示.但当样本数据很多时,茎叶图的效果就不是很好了.

练习

- 2003~2004 赛季,某球员在 NBA 一些场次的比赛中所得篮板球数为 16, 6, 3, 5, 12, 8, 13, 6, 10, 3, 19, 14, 9, 7, 10, 10, 9, 11, 6, 11, 12, 14, 8, 6, 10, 5, 10, 11, 13, 9, 10, 10, 7, 6, 11, 12, 17, 4, 12, 8, 10, 12, 9, 15, 15, 12, 13, 18, 8, 16, 请制作这些数据的茎叶图.
- 已知如图是 9 位评委给某作品打出的分数的茎叶图,那么 9 位评委打出的分数平均分为 _____.



(第 2 题)



(第 3 题)

- 如图是甲、乙两名运动员某赛季一些场次得分的茎叶图.
 - 甲、乙两名运动员的最高得分各是多少?
 - 哪名运动员的成绩好一些?

习题 2.2

感受·理解

- 某射手在同一条件下射击 30 次, 结果为: 6 环及 6 环以下 2 次, 7 环 6 次, 8 环 7 次, 9 环 10 次, 10 环 5 次.
 - 列出频率分布表;
 - 估计射手击中 7~9 环的可能性.
- 从大量棉花中抽取 50 根棉花纤维, 纤维长度(单位: mm)的数据分组及各组的频数为: $[22.5, 25.5)$, 3; $[25.5, 28.5)$, 8; $[28.5, 31.5)$, 9; $[31.5, 34.5)$, 11; $[34.5, 37.5)$, 10; $[37.5, 40.5)$, 5; $[40.5, 43.5]$, 4.
 - 列出样本的频率分布表;
 - 画出频率分布直方图;
 - 估计纤维长度小于 36 的百分比.
- 为了检测某种产品的质量, 抽取了 1 个容量为 100 的样本, 数据的分组及各组频数如下表:

分 组	频 数	频 率
$[10.75, 10.85)$	3	
$[10.85, 10.95)$	9	
$[10.95, 11.05)$	13	
$[11.05, 11.15)$	16	
$[11.15, 11.25)$	26	
$[11.25, 11.35)$	20	
$[11.35, 11.45)$	7	
$[11.45, 11.55)$	4	
$[11.55, 11.65]$	2	
合 计	100	

- 完成上面的频率分布表;
 - 画出频率分布直方图;
 - 估计数据落在 $[10.95, 11.35)$ 范围内的可能性.
- 从规定尺寸为 25.40 mm 的一堆产品中任取 100 件, 测得它们的实际尺寸(单位: mm)如下, 试以 0.030 为组距, 列出样本频率分布表, 作出频率分布直方图、折线图.

25.39 25.36 25.34 25.42 25.45 25.38 25.39 25.42 25.47 25.35
 25.41 25.43 25.44 25.48 25.45 25.43 25.46 25.40 25.51 25.45
 25.40 25.39 25.41 25.36 25.38 25.31 25.56 25.43 25.40 25.38
 25.37 25.44 25.33 25.46 25.40 25.49 25.34 25.42 25.50 25.37
 25.35 25.32 25.45 25.40 25.27 25.43 25.54 25.39 25.45 25.43

25.40 25.43 25.44 25.41 25.53 25.37 25.38 25.24 25.44 25.40
 25.36 25.42 25.39 25.46 25.38 25.35 25.31 25.34 25.40 25.36
 25.41 25.32 25.38 25.42 25.40 25.33 25.37 25.41 25.49 25.35
 25.47 25.34 25.30 25.39 25.36 25.46 25.29 25.40 25.37 25.33
 25.40 25.35 25.41 25.37 25.47 25.39 25.42 25.47 25.38 25.39

5. 下面是某市 2005 年 9 月 26 日和 9 月 29 日市区出现堵车的时刻,试列出这两天的堵车时刻的频率分布表和频率分布直方图,并分析该市每天大约在什么时间段是行车高峰期.

9 月 26 日	8:01	8:02	9:30	9:31	9:51	10:24	10:51	
	11:21	15:52	16:30	17:29	17:30	18:04	18:22	
9 月 29 日	8:29	8:32	8:33	9:29	9:58	10:14	10:33	11:43
	14:00	16:08	16:29	16:54	16:55	17:05	18:08	18:09

6. 选一篇英语短文,从它的第一个单词起直到最后一个单词结束,数出各个单词所含字母的个数,并就字母个数列出频率分布表,画出频率分布直方图;也可从电脑的文档里或网上找一篇短文,用计算机查找的方法进行统计.
7. 试用茎叶图表示表 2-2-3 中前 5 行的数据,其中茎用两位数来表示.

思考·运用

8. 为了检测某种产品的质量,抽取了 1 个容量为 40 的样本,检测结果为一等品 8 件,二等品 18 件,三等品 12 件,次品 2 件.
- 列出样本的频率分布表;
 - 画出频数条形图;
 - 估计这种产品为二等品或三等品的百分率.

9. 在某电脑杂志的一篇文章中,每个句子所含字的个数如下:

10 28 31 17 23 27 18 15 26 24 20 19
 36 27 14 25 15 22 11 21 24 27 17 29

在某报纸的一篇文章中,每个句子所含字的个数如下:

27 39 33 24 28 19 32 41 33 27 35 12
 36 41 27 13 22 23 18 46 32 22 18 32

- 将这两组数据用茎叶图表示;
- 将这两组数据进行比较分析,你能得到什么结论?

探究·拓展

10. 操作 1: 将 1 000 粒黑芝麻与 1 000 粒白芝麻放入一个容器中,并搅拌均匀,再用小杯从容器中取出一杯芝麻,计算黑芝麻的频率. 操作 2: 将 1 500 粒黑芝麻与 500 粒白芝麻放入一个容器中,并搅拌均匀,再用小杯从容器中取出一杯芝麻,计算黑芝麻的频率.

通过两次操作,你是否有所发现? 若有一袋芝麻,由黑、白两种芝麻混合而成,你用什么方法估计其中黑芝麻所占的百分比?

阅 读



拉普拉斯(P. S. Laplace, 1749~1827), 法国著名的数学家和天文学家, 分析概率论的创始人, 天体力学的主要奠基人, 天体演化学的创立者之一。

男女出生性别比

1814年, 法国著名的数学家拉普拉斯在他的著作《概率的哲学讨论》中, 记录了一项有趣的统计结果. 他根据伦敦、彼得堡、柏林和全法国的统计资料, 得出几乎完全一致的男婴出生数与女婴出生数的比值约为 22:21, 即在全体出生婴儿中, 男婴约占 51.2%, 女婴约占 48.8%, 或者说男女性别比约为 104.9%.

之后, 经过大量统计资料的验证, 全世界各种人种的婴儿性别比都遵循这一规律, 如果与此标准差异较大, 那么就一定存在某种人为的因素. 如拉普拉斯统计了法国巴黎 1745~1784 年 40 年间的男婴出生率为 50.06%, 与男婴标准出生率相差 1.14%. 经过调查, 发现当时的巴黎人“重女轻男”, 有抛弃男婴的陋俗, 以致歪曲了出生率的真相. 日本的统计也基本符合这一标准, 仅有几个例外. 如明治 38 年男女性别比为 102.7%, 明治 39 年 108.7%, 明治 40 年 102.7%, 原因是明治 39 年是“丙午年”, 由于迷信, 这一年生的女婴, 分别假报到明治 38 年和明治 40 年. 这种现象在现在的日本仍然存在, 如在昭和 41 年还出现过.

我国 1953 年和 1964 年人口普查时统计的男女性别比分别为 107.6% 和 105.5%, 基本符合国际公认的 105%~107% 的标准. 可是, 从 1982 年第三次全国人口普查起, 我国婴儿出生的男女性别比开始出现较为严重的失调: 1982 年为 108.5%, 到 1987 年 1% 人口抽样调查时已达 110.9%, 1990 年第四次人口普查为 111.3%, 1995 年 1% 人口抽样调查时高达 115.6%, 到 2000 年第五次人口普查时已是 116.9%. 男女出生性别比的失调已经引起人们的关注.

2.3

总体特征数的估计

根据第 2.2 节开头的数据,还可以求出北京地区近年来 7 月 25 日至 8 月 10 日的日最高气温的样本平均值为 34.02,我们可将其作为北京地区近年来 7 月 25 日至 8 月 10 日的日最高气温平均值的估计.

在数学中,通常把能反映总体某种特征的量称为总体特征数.

● 怎样通过抽样的方法,用样本的特征数估计总体的特征数呢?

2.3.1 平均数及其估计

某校高一(1)班同学在老师的布置下,用单摆进行测试,以检验重力加速度.全班同学两人一组,在相同条件下进行测试,得到下列实验数据(单位: m/s^2):

9.62 9.54 9.78 9.94 10.01 9.66 9.88 9.68 10.32
9.76 9.45 9.99 9.81 9.56 9.78 9.72 9.93 9.94
9.65 9.79 9.42 9.68 9.70 9.84 9.90

● 怎样利用这些数据对重力加速度进行估计?

n 个实数 $a_1, a_2, a_3, \dots, a_n$ 的和简记为 $\sum_{i=1}^n a_i$, “ \sum ” 读作 sigma [ˈsɪgmə].

我们常用算术平均数 $\frac{1}{n} \sum_{i=1}^n a_i$ [其中 $a_i (i = 1, 2, \dots, n)$ 为 n 个实验数据] 作为重力加速度的“最理想”的近似值,它的依据是什么呢?

处理实验数据的原则是使这个近似值与实验数据越接近越好. 设这个近似值为 x ,它与 n 个实验数据 $a_i (i = 1, 2, \dots, n)$ 的离差分别为 $x - a_1, x - a_2, x - a_3, \dots, x - a_n$. 由于上述离差有正有负,故不宜直接相加. 可以考虑离差的平方和,即

$$(x - a_1)^2 + (x - a_2)^2 + \dots + (x - a_n)^2.$$

因为

$$\begin{aligned} & (x - a_1)^2 + (x - a_2)^2 + \dots + (x - a_n)^2 \\ &= nx^2 - 2(a_1 + a_2 + \dots + a_n)x + a_1^2 + a_2^2 + \dots + a_n^2, \end{aligned}$$

所以当 $x = \frac{a_1 + a_2 + \dots + a_n}{n}$ 时,离差的平方和最小,故可用

$\frac{a_1 + a_2 + \dots + a_n}{n}$ 作为表示这个量的理想近似值,称为这 n 个数据 a_1, a_2, \dots, a_n 的**平均数**(average)或**均值**(mean),一般记为

\bar{a} 读作“ a 平均”.

$$\bar{a} = \frac{a_1 + a_2 + \dots + a_n}{n}$$

这样,我们可以用计算器求得,由高一(1)班学生的实验数据估计的重力加速度的最佳近似值为 9.774 m/s^2 .

例 1 某校高一年级的甲、乙两个班级(均为 50 人)的语文测试成绩(总分: 150 分)如下,试确定这次考试中,哪个班的语文成绩更好一些.

甲班

112	86	106	84	100	105	98	102	94	107
87	112	94	94	99	90	120	98	95	119
108	100	96	115	111	104	95	108	111	105
104	107	119	107	93	102	98	112	112	99
92	102	93	84	94	94	100	90	84	114

乙班

116	95	109	96	106	98	108	99	110	103
94	98	105	101	115	104	112	101	113	96
108	100	110	98	107	87	108	106	103	97
107	106	111	121	97	107	114	122	101	107
107	111	114	106	104	104	95	111	111	110

分析 我们可用一组数据的平均数衡量这组数据的集中水平,因此,分别求出甲、乙两个班级的平均分即可.

解 用计算器分别求出甲班的平均分为 101.1,乙班的平均分约为 105.4,故这次考试乙班成绩要好于甲班.

EXCEL & CALCULATOR

Excel 中函数“**AVERAGE()**”可以直接用于计算给定数据的平均数.如将乙班的成绩输入工作表中 A1: J5 区域后,在某空白单元格中输入“**=AVERAGE(A1: J5)**”,即得乙班的平均分为 105.38(如图 2-3-1).

K1	=AVERAGE(A1: J5)										
	A	B	C	D	E	F	G	H	I	J	K
1	116	95	109	96	106	98	108	99	110	103	105.38
2	94	98	105	101	115	104	112	101	113	96	
3	108	100	110	98	107	87	108	106	103	97	
4	107	106	111	121	97	107	114	122	101	107	
5	107	111	114	106	104	104	95	111	111	110	

图 2-3-1

用计算器计算平均数的方法：

(1) 选择计算模式，按 $\boxed{\text{MODE}}$ $\boxed{2}$ 键，选择“统计”模式；

(2) 输入数据，按 112 $\boxed{\text{DT}}$ 即可输入数据 112，仿此将甲班的语文测试成绩全部输入（按 $\boxed{\text{DT}}$ $\boxed{\text{DT}}$ 键能输入同样的数据两次，若输入 5 个相同的数据 20，则可按 20 $\boxed{\text{SHIFT}}$ $\boxed{;}$ 5 $\boxed{\text{DT}}$ ）；

(3) 求平均数，按 $\boxed{\text{SHIFT}}$ $\boxed{\text{S-VAR}}$ $\boxed{1}$ $\boxed{=}$ 键，即得甲班的平均分为 101.1.

输入数据之前，
按 $\boxed{\text{SHIFT}}$ $\boxed{\text{CLR}}$ $\boxed{1}$
 $\boxed{=}$ 键清除统计存储器。

例 2 下面是某校学生日睡眠时间(单位：h)的抽样频率分布表，试估计该校学生的日平均睡眠时间.

表 2-3-1

睡眠时间	人数	频率
[6, 6.5)	5	0.05
[6.5, 7)	17	0.17
[7, 7.5)	33	0.33
[7.5, 8)	37	0.37
[8, 8.5)	6	0.06
[8.5, 9]	2	0.02
合计	100	1

分析 要确定这 100 名学生的平均睡眠时间，就必须计算其总睡眠时间. 由于每组中的个体睡眠时间只是一个范围，所以可用各组区间的组中值近似地表示.

解法 1 总睡眠时间约为

$$\begin{aligned} & 6.25 \times 5 + 6.75 \times 17 + 7.25 \times 33 + 7.75 \times 37 + \\ & 8.25 \times 6 + 8.75 \times 2 \\ & = 739(\text{h}). \end{aligned}$$

故平均睡眠时间约为 7.39 h.

解法 2 求组中值与对应频率之积的和.

$$\begin{aligned} & 6.25 \times 0.05 + 6.75 \times 0.17 + 7.25 \times 0.33 + 7.75 \times \\ & 0.37 + 8.25 \times 0.06 + 8.75 \times 0.02 \\ & = 7.39(\text{h}). \end{aligned}$$

答 估计该校学生的日平均睡眠时间约为 7.39 h.

是否还有其他的
估算方法？

一般地,若取值为 x_1, x_2, \dots, x_n 的频率分别为 p_1, p_2, \dots, p_n , 则其平均数为 $x_1 p_1 + x_2 p_2 + \dots + x_n p_n$.

例 3 某单位年收入(单位:元)在 10 000 ~ 15 000, 15 000 ~ 20 000, 20 000 ~ 25 000, 25 000 ~ 30 000, 30 000 ~ 35 000, 35 000 ~ 40 000 及 40 000 ~ 50 000 之间的职工所占的百分比分别为 10%, 15%, 20%, 25%, 15%, 10% 和 5%, 试估计该单位职工的平均年收入.

分析 上述所占的比就是各组的频率.

解 估计该单位职工的平均年收入为

$$\begin{aligned} & 12\,500 \times 10\% + 17\,500 \times 15\% + 22\,500 \times 20\% + 27\,500 \times \\ & 25\% + 32\,500 \times 15\% + 37\,500 \times 10\% + 45\,000 \times 5\% \\ & = 26\,125(\text{元}). \end{aligned}$$

答 估计该单位人均年收入约为 26 125 元.

练习

- 根据第 2.1.1 节练习第 4 题所得到的数据,估计该校高一学生的身高的平均值.
- 对第 2.2.2 节练习第 3 题,试估计这批灯泡的平均使用寿命.
- 从某校全体高考考生中任意抽取 20 名考生,其数学成绩(总分:150 分)分别为 102, 105, 131, 95, 83, 121, 140, 100, 97, 96, 95, 121, 124, 135, 106, 109, 110, 101, 98, 97, 试估计该校全体考生的数学平均成绩.
- 下表是一个容量为 20 的样本数据分组后的频数分布.若利用组中值近似计算本组数据的平均数 \bar{a} , 则 $\bar{a} =$ _____.

数据	[12.5, 15.5)	[15.5, 18.5)	[18.5, 21.5)	[21.5, 24.5]
频数	3	3	6	8

- 一份共 3 道题的测试卷,每道题 1 分.全班得 3 分、2 分、1 分和 0 分的学生所占比例分别为 30%, 50%, 10% 和 10%.
 - 若全班共 10 人,则平均分是多少?
 - 若全班共 20 人,则平均分是多少?
 - 如果该班人数未知,那么能求出该班的平均分吗?

2.3.2 方差与标准差

有甲、乙两种钢筋,现从中各抽取一个样本(如表 2-3-2)检查它们的抗拉强度(单位: kg/mm^2),通过计算发现,两个样本的平均数均为 125.

表 2-3-2

甲	110	120	130	125	120	125	135	125	135	125
乙	115	100	125	130	115	125	125	145	125	145

● 哪种钢筋的质量较好?

将甲、乙两个样本数据分别标在数轴上,如图 2-3-2 所示.

这样的图,也称为点线图(lineplot).

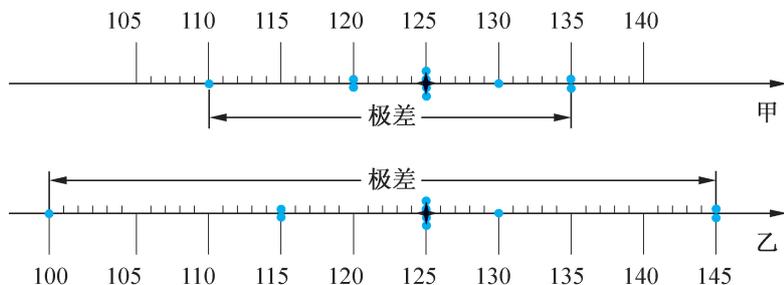


图 2-3-2

从图 2-3-2 中可以看出,乙样本的最小值 100 低于甲样本的最小值 110,乙样本的最大值 145 高于甲样本的最大值 135,这说明乙种钢筋没有甲种钢筋的抗拉强度稳定.

我们把一组数据的最大值与最小值的差称为**极差**(range).从图 2-3-2 中可看出,乙的极差较大,数据点较分散;甲的极差小,数据点较集中.这说明甲比乙稳定.运用极差对两组数据进行比较,操作简单方便,但当两组数据的集中程度差异不大时,就不容易得出结论.

我们还可以考虑每一抗拉强度与平均抗拉强度的离差.结合上节有关离差的讨论,各次抗拉强度与平均抗拉强度的离差的平方和越小,稳定性就越高.由于两组数据的容量可能不同,因此应将上述平方和除以数据的个数,我们把由此所得的值称为这组数据的**方差**(variance).

标准差是样本数据到平均数的一种平均距离.

因为方差与原始数据的单位不同,且平方后可能夸大了离差的程度,所以我们将方差的算术平方根称为这组数据的**标准差**(standard deviation).标准差也可以刻画一组数据的稳定程度.

一般地,

设一组样本数据 x_1, x_2, \dots, x_n , 其平均数为 \bar{x} , 则称

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

为这个样本的**方差**, 其算术平方根 $s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$ 为样本的**标准差**, 分别简称**样本方差**、**样本标准差**.

根据上述方差计算公式可以算得甲、乙两个样本的方差分别为 50 和 165, 故可以认为甲种钢筋的质量好于乙种钢筋.

例 4 甲、乙两种水稻试验品种连续 5 年的平均单位面积产量(单位: t/hm^2)如下, 试根据这组数据估计哪一种水稻品种的产量比较稳定.

表 2-3-3

品 种	第 1 年	第 2 年	第 3 年	第 4 年	第 5 年
甲	9.8	9.9	10.1	10	10.2
乙	9.4	10.3	10.8	9.7	9.8

解 甲品种的样本平均数为 10, 样本方差为

$$\begin{aligned} & [(9.8 - 10)^2 + (9.9 - 10)^2 + (10.1 - 10)^2 + \\ & (10 - 10)^2 + (10.2 - 10)^2] \div 5 \\ & = 0.02. \end{aligned}$$

乙品种的样本平均数也为 10, 样本方差为

$$\begin{aligned} & [(9.4 - 10)^2 + (10.3 - 10)^2 + (10.8 - 10)^2 + \\ & (9.7 - 10)^2 + (9.8 - 10)^2] \div 5 \\ & = 0.244. \end{aligned}$$

因为 $0.244 > 0.02$, 所以, 由这组数据可以认为甲种水稻的产量比较稳定.

例 5 为了保护学生的视力, 教室内的日光灯在使用一段时间后必须更换. 已知某校使用的 100 只日光灯在必须换掉前的使用天数如下, 试估计这种日光灯的平均使用寿命和标准差.

表 2-3-4

使用天数	151~ 180	181~ 210	211~ 240	241~ 270	271~ 300	301~ 330	331~ 360	361~ 390
日光灯数	1	11	18	20	25	16	7	2

分析 用每一区间内的组中值作为相应日光灯的使用寿命,再求平均使用寿命.

解 各区间的组中值分别为 165.5, 195.5, 225.5, 255.5, 285.5, 315.5, 345.5, 375.5, 由此算得平均数约为

$$\begin{aligned} & 165.5 \times 1\% + 195.5 \times 11\% + 225.5 \times 18\% + 255.5 \times 20\% + \\ & 285.5 \times 25\% + 315.5 \times 16\% + 345.5 \times 7\% + 375.5 \times 2\% \\ & = 268.4 \approx 268(\text{天}). \end{aligned}$$

这些组中值的方差为

$$\begin{aligned} & \frac{1}{100} \times [1 \times (165.5 - 268.4)^2 + 11 \times (195.5 - 268.4)^2 + 18 \times \\ & (225.5 - 268.4)^2 + 20 \times (255.5 - 268.4)^2 + 25 \times (285.5 - 268.4)^2 + \\ & 16 \times (315.5 - 268.4)^2 + 7 \times (345.5 - 268.4)^2 + 2 \times (375.5 - 268.4)^2] \\ & = 2\,128.60(\text{天}^2), \end{aligned}$$

故所求的标准差约为 $\sqrt{2\,128.6} \approx 46(\text{天})$.

答 估计这种日光灯的平均使用寿命约为 268 天,标准差约为 46 天.

从例 5 中可以看出,样本数据中在 $[268 - 46 \times 2, 268 + 46 \times 2]$ 外的只有 3 个,也就是说,区间 $[\bar{x} - 2s, \bar{x} + 2s]$ 几乎包含了所有数据.

EXCEL & CALCULATOR

在 Excel 中,可分别用函数“VARP()”和“STDEVP()”计算方差和标准差.

练习

1. 已知一组数据 1, 3, 2, 5, 4, 那么这组数据的标准差为 _____.
2. 从两个班级各抽 5 名学生测量身高(单位: cm),甲班的数据为 160, 162, 159, 160, 159,乙班的数据为 180, 160, 150, 150, 160. 试估计哪个班级学生身高的波动小.
3. 若 k_1, k_2, \dots, k_8 的方差为 3, 则 $2(k_1 - 3), 2(k_2 - 3), \dots, 2(k_8 - 3)$ 的方差为 _____.
4. 第 2.2.3 节练习第 3 题中,哪名运动员的成绩更稳定?
5. 利用计算器计算下列两组数据的平均数和标准差.

甲	9.9	10.3	9.8	10.1	10.4	10.0	9.8	9.7
乙	10.2	10.0	9.5	10.3	10.5	9.6	9.8	10.1

习题 2.3

感受·理解

1. 已知某地连续 10 天的最低气温(单位: $^{\circ}\text{C}$)依次是 x_1, x_2, x_3, x_4, x_5 和 $x_1 + 1, x_2 + 2, x_3 + 3, x_4 + 4, x_5 + 5$, 若前 5 天的平均最低气温为 7°C , 求后 5 天的平均最低气温.
2. 某工厂一个月(30 天)中的日产值如下: 有 2 天的产值是 5.1 万元, 有 3 天的产值是 5.2 万元, 有 6 天的产值是 5.3 万元, 有 8 天的产值是 5.4 万元, 有 7 天的产值是 5.5 万元, 有 3 天的产值是 5.6 万元, 有 1 天的产值是 5.7 万元. 试计算该厂这个月的平均日产值.
3. 为了考察某种大麦穗长的分布情况, 在一块试验田里抽取了 100 穗, 量得它们的长度(单位: cm)如下, 请列出频率分布表, 并估计该试验田里麦穗的平均长度.

6.5 6.4 6.7 5.8 5.9 5.9 5.2 4.0 5.4 5.6
 5.8 5.5 6.0 6.5 5.1 6.5 5.3 5.9 5.5 5.8
 6.2 5.4 5.0 5.0 6.8 6.0 5.0 5.7 6.0 5.5
 6.8 6.0 6.3 5.5 5.0 6.8 6.6 6.0 7.0 6.4
 6.4 5.8 5.9 5.7 6.8 6.6 6.0 6.4 5.7 7.4
 6.0 5.4 6.5 6.0 6.8 5.8 6.3 6.0 6.3 5.6
 5.3 6.4 5.7 6.7 6.2 5.6 6.0 6.7 6.7 6.0
 5.5 6.2 6.1 5.3 6.2 6.8 6.6 4.7 5.7 5.7
 5.8 5.3 7.0 6.0 6.0 5.9 5.4 6.0 5.2 6.0
 6.3 5.7 6.8 6.1 4.5 5.6 6.3 6.0 5.8 6.3

4. 两台机床同时生产一种零件, 在 10 天中, 两台机床每天的次品数如下:

甲	1	0	2	0	2	3	0	4	1	2
乙	1	3	2	1	0	2	1	1	0	1

- (1) 哪台机床次品数的平均数较小?
 - (2) 哪台机床生产状况比较稳定?
5. 甲、乙两台半自动车床加工同一型号的产品, 各生产 1 000 只产品中次品数分别用 x 和 y 表示. 经过一段时间的观察, 发现 x 和 y 的频率分布如下表, 问: 哪一台车床的产品质量较好?

x	0	1	2	3
P	0.7	0.1	0.1	0.1
y	0	1	2	3
P	0.5	0.3	0.2	0

思考·运用

6. 某制造商生产长度为 6 cm 的金属棒, 抽样检查 40 根, 测得每根长度(单位: cm, 保留两位小数)如下:

6.02 6.01 5.94 5.94 5.97 5.96 5.98 6.01 5.98 5.99
 6.00 6.03 5.99 5.97 5.98 6.00 6.03 5.95 6.00 6.00
 5.92 5.93 5.99 5.99 6.00 5.95 6.00 5.97 5.96 5.97
 6.03 6.01 5.98 5.99 6.04 6.00 6.02 5.97 5.96 5.98

- (1) 计算上述样本中金属棒的平均长度;
 (2) 画出频率分布直方图;
 (3) 如果允许制造商生产这种金属棒与 6 cm 的标准有 0.2% 的离差, 那么抽样检查中合格的金属棒有几根? 合格率是多少?

7. 下面是 60 名男生每分钟脉搏数:

72 70 66 74 81 70 74 53 57 62 58 92
 72 67 62 91 73 64 65 80 78 67 75 80
 83 61 72 72 69 70 76 74 65 84 79 80
 76 72 68 65 82 79 71 86 77 69 72 56
 70 62 76 56 86 63 73 70 75 73 89 64

- (1) 画出上述数据的频率分布直方图;
 (2) 根据直方图的各组中值估计总体平均数, 并将所得结果与实际的总体平均数相比较, 计算误差.

探究·拓展

8. 一位研究化肥的科学家将一片土地划分为 100 个 50 m^2 的小块, 并在 50 个小块上施用新化肥, 留下 50 个条件大体相当的小块不施新化肥. 施用新化肥的 50 小块土地的小麦产量(单位: kg)如下:

15 29 22 15 3 30 22 16 5 2
 22 13 20 25 42 25 20 38 12 29
 14 21 26 13 21 27 13 21 11 18
 10 18 24 24 36 34 23 18 10 9
 17 23 33 8 16 23 31 16 23 40

没有施用新化肥的 50 小块土地的小麦产量(单位: kg)如下:

23 16 16 17 22 3 10 10 8 14
 16 5 24 16 32 23 15 18 9 21
 4 24 5 24 15 2 15 25 17 29
 33 39 16 17 2 15 17 17 26 13
 26 11 18 19 12 20 27 12 28 22

你认为新化肥的研制已经取得成功了吗?

2.4

线性回归方程

在实际问题中,变量之间的常见关系有如下两类:

一类是确定性函数关系,变量之间的关系可以用函数表示.例如,圆的面积 S 与半径 r 之间就是确定性函数关系,可以用函数 $S = \pi r^2$ 表示.

一类是相关关系,变量之间有一定的联系,但不能完全用函数来表达.例如,人的体重 y 与身高 x 有关.一般来说,身高越高,体重越重,但不能用一个函数来严格地表示身高与体重之间的关系.

● 用怎样的数学模型刻画两个变量之间的相关关系?

某小卖部为了了解热茶销量与气温之间的关系,随机统计并制作了某 6 天卖出热茶的杯数与当天气温的对照表:

表 2-4-1

气温/ $^{\circ}\text{C}$	26	18	13	10	4	-1
杯数	20	24	34	38	50	64

如果某天的气温是 -5°C ,那么你能根据这些数据预测这天小卖部卖出热茶的杯数吗?

为了了解热茶销量与气温的大致关系,我们以横坐标 x 表示气温、纵坐标 y 表示热茶销量建立平面直角坐标系,将表中数据构成的 6 个数对所表示的点在坐标系内标出,得到图 2-4-1.今后我们称这样的图为**散点图**(scatterplot).

从图 2-4-1 可以看出,这些点散布在一条直线的附近,故可以用一个线性函数近似地表示热茶销量与气温之间的关系.

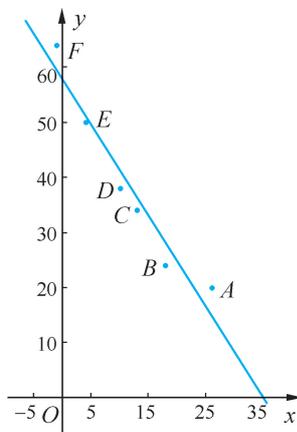


图 2-4-1

\hat{y} 读作“y 估计”。

用方程为 $\hat{y} = bx + a$ 的直线拟合散点图中的点,应使得该直线与散点图中的点最接近.那么,怎样衡量直线 $\hat{y} = bx + a$ 与图中 6 个点的接近程度呢?

我们将表中给出的自变量 x 的 6 个值代入直线方程,得到相应的 6 个 \hat{y} 值:

$$26b + a, 18b + a, 13b + a, 10b + a, 4b + a, -b + a.$$

这 6 个值与表中相应的实际值应该越接近越好.所以,我们用类似于第 2.3.1 节中估计总体平均数的思想,考虑离差的平方和

$$\begin{aligned} Q(a, b) = & (26b + a - 20)^2 + (18b + a - 24)^2 + \\ & (13b + a - 34)^2 + (10b + a - 38)^2 + \\ & (4b + a - 50)^2 + (-b + a - 64)^2, \end{aligned}$$

它是直线 $\hat{y} = bx + a$ 与各散点在垂直方向(纵轴方向)上的距离的平方和,可以用来衡量直线 $\hat{y} = bx + a$ 与图中 6 个点的接近程度.所以,设法取 a, b 的值,使 $Q(a, b)$ 达到最小值.这种方法叫做**最小平方方法**(method of least square).

“最小平方方法”又称“最小二乘法”。

接下来,需要计算使 $Q(a, b)$ 取得最小值的 a 和 b ,我们可以用公式

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

这里 x_i, y_i 是观察数据, $i = 1, 2, \dots, n$, 其计算过程如下:

$$x_1 = 26, y_1 = 20, x_2 = 18, y_2 = 24, x_3 = 13, y_3 = 34,$$

$$x_4 = 10, y_4 = 38, x_5 = 4, y_5 = 50, x_6 = -1, y_6 = 64,$$

$$6 \sum_{i=1}^6 x_i^2 = 6 \times [26^2 + 18^2 + 13^2 + 10^2 + 4^2 + (-1)^2] = 7\,716,$$

$$\left(\sum_{i=1}^6 x_i\right)^2 = (26 + 18 + 13 + 10 + 4 - 1)^2 = 70^2 = 4\,900,$$

$$\sum_{i=1}^6 x_i = 70, \sum_{i=1}^6 y_i = 20 + 24 + 34 + 38 + 50 + 64 = 230,$$

$$\begin{aligned} 6 \sum_{i=1}^6 x_i y_i = & 6 \times [26 \times 20 + 18 \times 24 + 13 \times 34 + 10 \times 38 + \\ & 4 \times 50 + (-1) \times 64] = 11\,460, \end{aligned}$$

$$\bar{x} = \frac{70}{6}, \bar{y} = \frac{230}{6},$$

因此

$$b = \frac{11\,460 - 70 \times 230}{7\,716 - 4\,900} \approx -1.6477,$$

$$a = \frac{230}{6} - \frac{11\,460 - 70 \times 230}{7\,716 - 4\,900} \times \frac{70}{6} \approx 57.5568,$$

从而所求直线方程为

$$\hat{y} = -1.6477x + 57.5568.$$

当 $x = -5$ 时, $\hat{y} \approx 66$. 故当气温为 -5°C 时, 热茶销量约 66 杯.

像这样能用直线方程 $\hat{y} = bx + a$ 近似表示的相关关系叫做 **线性相关关系** (linear correlation).

一般地, 设有 n 对观察数据如下:

表 2-4-2

x	x_1	x_2	x_3	\cdots	x_n
y	y_1	y_2	y_3	\cdots	y_n

当 a, b 使

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

取得最小值时, 就称方程 $\hat{y} = bx + a$ 为拟合这 n 对数据的 **线性回归方程** (linear regression equation), 该方程所表示的直线称为回归直线.

线性回归方程 $\hat{y} = bx + a$ 中的系数 a, b 可以用公式 (*) 计算, 也可以用下面的公式 (***) 计算.

公式 (*), (**)
的推导比较复杂, 这
里不作要求.

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

例 1 下表为某地近几年机动车辆数与交通事故数的统计资料, 请判断机动车辆数与交通事故数之间是否具有线性相关关系. 如果具有线性相关关系, 求出线性回归方程; 如果不具有线性相关关系, 说明理由.

表 2-4-3

机动车辆数 $x/10^3$ 辆	95	110	112	120	129	135	150	180
交通事故数 $y/10^3$ 件	6.2	7.5	7.7	8.5	8.7	9.8	10.2	13

解 在直角坐标系中画出数据的散点图, 直观判断散点在一条直线附近, 故具有线性相关关系. 计算相应的数据之和:

用计算器计算 $\sum_{i=1}^n x_i$, $\sum_{i=1}^n x_i^2$ 可一次完成:先用“M+”键输入 x_i ($i = 1, 2, 3, \dots, n$),再分别按 $\sum x$, $\sum x^2$ 键可得. 求 $\sum_{i=1}^n x_i y_i$ 时,每次将两数之积输入“M+”,最后按 \sum 键即可,不必先算出 $x_i y_i$ 的值.

思考

$$\sum_{i=1}^8 x_i = 1031, \quad \sum_{i=1}^8 y_i = 71.6,$$

$$\sum_{i=1}^8 x_i^2 = 137835, \quad \sum_{i=1}^8 x_i y_i = 9611.7,$$

将它们代入(*)式计算得

$$b \approx 0.0774, \quad a = -1.0241,$$

所以,所求线性回归方程为

$$\hat{y} = 0.0774x - 1.0241.$$

图 2-4-2 是 1991~2000 年北京地区年平均气温(单位: $^{\circ}\text{C}$)与年降雨量(单位: mm)的散点图. 此时求得的回归直线方程有意义吗?

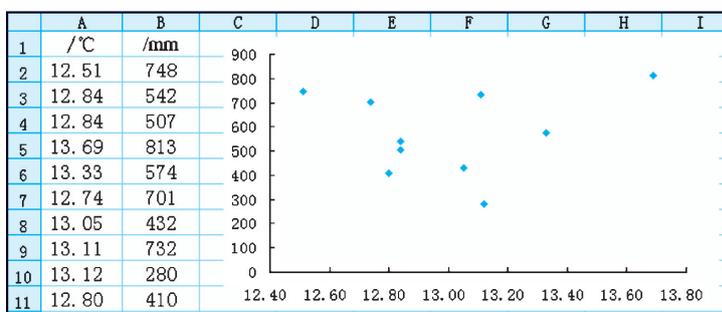


图 2-4-2

一般地,用回归直线进行数据拟合的一般步骤为:

- (1) 作出散点图,判断散点是否在一条直线附近;
- (2) 如果散点在一条直线附近,那么用公式(*)或(**)求出 a, b ,并写出线性回归方程.

EXCEL & CALCULATOR

用 Excel 和计算器进行数据拟合,请参阅《数学 1(必修)》中的相关内容.

求线性回归方程可以借助计算机来完成. 用 Excel 进行数据拟合,得到图 2-4-3.

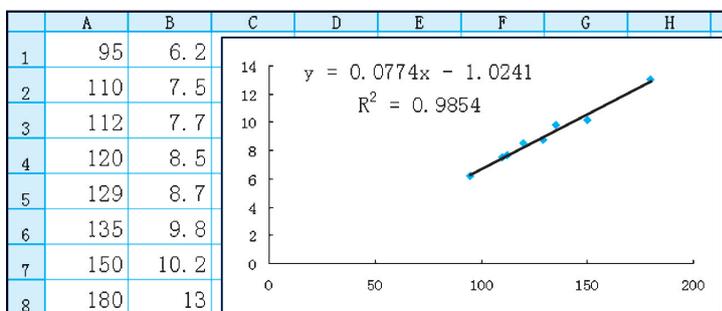


图 2-4-3

图中得到的直线即为回归直线,显示的 R^2 的值越接近 1,其拟合效果越好.

练习

1. 下面是我国居民生活污水排放量的一组数据(单位: 10^8 t), 试分别估计 1996 年和 2004 年我国居民生活污水排放量.

年 份	1995	1996	1997	1998
排放量	151		189.1	194.8
年 份	1999	2000	2001	2002
排放量	203.8	220.9	227.7	232.3

2. 下表是某地今年的房屋销售价格 y (单位: 万元)与房屋的面积 x (单位: m^2) 的数据. 经过研究, 发现 y 与 x 之间具有线性相关关系. 写出这两个变量间的线性回归方程.

面积/ m^2	88	92	102	112	106
价格/万元	112	137	148	153	150

3. 一个工厂在某年里每月产品的总成本 y (单位: 万元)与月产量 x (单位: 万件)之间有如下的一组数据:

x	1.08	1.12	1.19	1.28	1.36	1.48	1.59	1.68	1.80	1.87	1.98	2.07
y	2.25	2.37	2.40	2.55	2.64	2.75	2.92	3.03	3.14	3.26	3.36	3.50

- (1) 画出散点图;
 (2) 求出线性回归方程.

习题 2.4

感受·理解

1. 某小吃店的日盈利 y (单位: 百元) 与当天平均气温 x (单位: $^{\circ}\text{C}$) 之间有如下数据:

$x/^{\circ}\text{C}$	-2	-1	0	1	2
$y/\text{百元}$	5	4	2	2	1

甲、乙、丙 3 位同学对上述数据进行了分析, 发现 y 与 x 之间具有线性相关关系, 他们通过计算分别得到 3 个线性回归方程: ① $\hat{y} = -x + 2.8$; ② $\hat{y} = -x + 3$; ③ $\hat{y} = -1.2x + 2.6$. 其中正确的序号是_____.

2. 某种产品的广告费支出 x 与销售额 y 之间有如下对应数据:

$x/10^6$ 元	2	4	5	6	8
$y/10^6$ 元	30	40	60	50	70

- (1) 画出散点图;
(2) 求出线性回归方程.

3. 每立方米混凝土的水泥用量 x (单位: kg) 与 28 天后混凝土的抗压强度 (单位: kg/cm^2) 之间有如下对应数据:

x	150	160	170	180	190	200
y	56.9	58.3	61.1	64.6	68.1	71.3
x	210	220	230	240	250	260
y	74.1	77.4	80.2	82.6	86.4	89.7

- (1) 画出散点图;
(2) 求出线性回归方程.

4. 在国民经济中, 生产和货运之间有着密切的关系. 下面列出 1991~2000 年某地区货运量与工业总产值的统计资料:

年 份	1991	1992	1993	1994	1995
总产值 $x/10^9$ 元	2.8	2.9	3.2	3.2	3.4
货运量 $y/10^4$ t	25	27	29	32	34
年 份	1996	1997	1998	1999	2000
总产值 $x/10^9$ 元	3.2	3.3	3.7	3.9	4.2
货运量 $y/10^4$ t	36	35	39	42	45

- (1) 画出散点图;
(2) 求出变量 x , y 之间的线性回归方程.

思考·运用

5. 试根据第 2.1.2 节练习第 4 题所得数据, 分析你班学生的两臂平展的长度与身高是否具有线性相关关系. 若存在线性相关关系, 求出线性回归方程; 若不存在线性相关关系, 请说明理由.

相关系数

观察如图 2-4-4 所示的第 2.4 节节首问题和例 1 中数据的散点图：

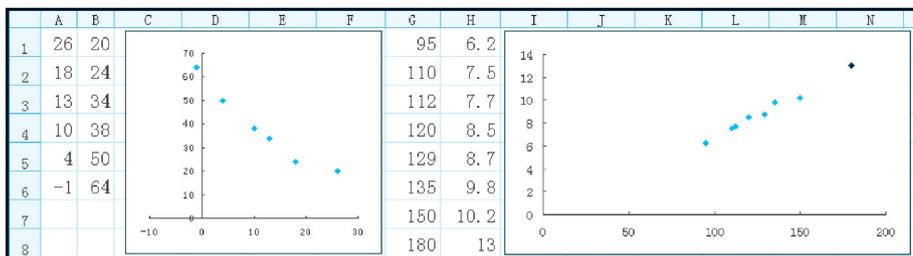


图 2-4-4

它们的线性回归方程分别为

$$\hat{y} = -1.6477x + 57.5568,$$

$$\hat{y} = 0.0774x - 1.0241,$$

两者有何异同？

前者随着自变量的增大，函数值在减小，对应的一次项系数为负数；后者随着自变量的增大，函数值在增大，对应的一次项系数为正数。

一般地，在线性回归方程 $\hat{y} = bx + a$ 中，当一次项系数 b 为正数时，就称这两个变量正相关；当 b 为负数时，就称这两个变量负相关。

另一方面，两个散点图中的点与回归直线的偏离程度也不相同，说明两个变量之间的线性相关程度也有区别。对于变量 y 与 x 的一组观测值来说，我们常用 r 来描述线性相关的程度：

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

即

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

这个数值 r 称为 y 与 x 的样本相关系数，简称**相关系数**。

当 $r > 0$ 时， y 与 x 正相关；当 $r < 0$ 时， y 与 x 负相关。可以证明， $|r| \leq 1$ 。 $|r|$ 越接近 1，线性相关程度越高； $|r|$ 越接近 0，线性相关程度越低。

实习作业

1. 对你校高一年级 15 周岁的男生(或女生)身高进行调查,把抽取的 60 名男生(或女生)测得的身高作为样本,将所得数据进行分析处理,完成相应的实习报告.

2. 自己(或分组)选择适当的课题,进行统计研究,并写出报告.

下列课题供参考:

(1) 在本校范围内就学生的视力情况进行抽样调查,对数据进行统计分析,写出实习报告,并就提高学生视力提出建设性的意见.

(2) 在本校范围内就学生的开支情况(学费、生活费、资料费、交通费、活动费、零花钱等)进行抽样调查,对数据进行统计分析,写出实习报告,并就学生开支情况提出你的看法.

(3) 对你校学生分年龄、性别进行身高抽样调查,就男女生身高及其增长速度等方面进行统计分析,设计调查表,并写出实习报告.

附：实习报告表

实习报告

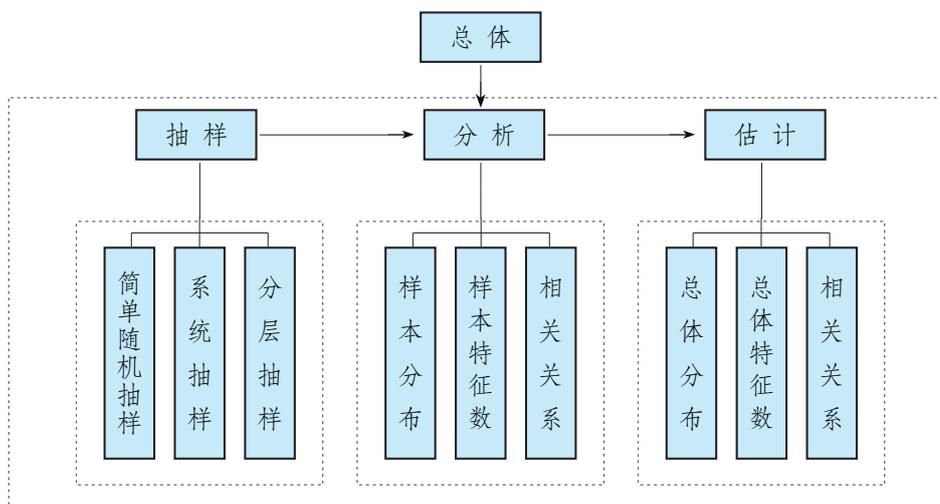
_____年_____月_____日

题目	
基本要求	
抽样说明	
样本数据	
频率分布表	
频率分布直方图	
样本平均数	
结果分析	
负责人及参加人员	
指导教师审核意见	
备注	

本章回顾

本章概览

本章介绍了从总体中抽取样本的常用方法,并通过实例,研究了如何利用样本对总体的分布规律、集中水平、稳定程度及相关关系等特性进行估计和预测。



当总体容量很大或检测具有一定的破坏性时,可以从总体中抽取适当的样本,通过对样本的分析、研究,得到对总体的估计,这就是统计分析的基本过程.而用样本估计总体就是统计思想的本质。

要准确估计总体,必须合理地选择样本,我们学习的是最常用的3种抽样方法.获取样本数据后,将其用频率分布表、频率直方图、频率折线图或茎叶图表示后,蕴含于数据之中的规律就会得到直观的揭示.用样本平均数可以对总体水平作出估计,用样本极差、方差(标准差)可以估计总体的稳定程度。

对两个变量的样本数据进行相关性分析,可以发现存在于现实世界中的回归现象.用最小平方法研究回归现象,得到的线性回归方程可用于估计和预测,为决策提供依据。

总之,统计的基本思想是从样本数据中发现统计规律,实现对总体的估计。

内容提要

1. 抽样方法:简单随机抽样、系统抽样、分层抽样
3种抽样方法的区别和联系如下表所示:

	特 点	相互联系	适用范围	共同点
简单随机 抽样	从总体中逐个 抽取		总体中的 个体数较少	抽样过程 中每个个体 被抽到的可 能性相同
系统抽样	将总体分成几部 分,按一定的规则 在各部分中抽取	在起始部分抽 样时,采用简单 随机抽样	总体中的 个体数较多	
分层抽样	将总体分成几 层,按各层个体数 之比抽取	各层抽样时, 采用简单随机抽 样或系统抽样	总体由差 异明显的几 部分组成	

2. 频率分布表

制作频率分布表的步骤:

- (1) 求全距,决定组数和组距,组数 = $\frac{\text{全距}}{\text{组距}}$; (2) 分组,通常对组内数值所在区间取左闭右开区间,最后一组取闭区间; (3) 登记频数,计算频率,列出频率分布表.

3. 频率分布直方图与折线图

制作频率分布直方图的步骤:

- (1) 先制作频率分布表,然后作直角坐标系; (2) 把横轴分成若干段,每一线段对应 1 个组的组距; (3) 以此线段为底作矩形,它的高等于该组的 $\frac{\text{频率}}{\text{组距}}$,这样得出一系列的矩形,每个矩形的面积恰好是该组的频率. 这些矩形就构成了频率分布直方图.

如果将频率分布直方图中各相邻的矩形的上底边的中点顺次连结起来,那么就得到频率分布折线图,简称频率折线图.

4. 平均数、方差和标准差

$$(1) \text{ 平均数: } \bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}.$$

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

$$(3) \text{ 标准差: } s = \sqrt{s^2} = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}}.$$

5. 线性回归方程

$$(1) \text{ 线性回归方程: } \hat{y} = bx + a.$$

$$(2) \text{ 回归系数: } b = \frac{\sum_{i=1}^n x_i y_i - n\bar{x}\bar{y}}{\sum_{i=1}^n x_i^2 - n\bar{x}^2}, a = \bar{y} - b\bar{x}.$$

$$(3) \text{ 回归直线 } \hat{y} = bx + a \text{ 过定点 } (\bar{x}, \bar{y}).$$

复习题

感受·理解

1. 若将容量为 100 的样本数据分为如下 8 组, 则第 3 组的频率为().
A. 0.14 B. 0.03 C. 0.07 D. 0.21

组号	1	2	3	4	5	6	7	8
频数	10	13		14	15	13	12	9

2. 调查某班学生的平均身高, 从 50 名学生中抽取 $\frac{1}{10}$, 应如何抽样? 如果知道男女生(男生 30 人, 女生 20 人)的身高显著不同, 又如何抽样?
3. 某工厂生产甲、乙、丙、丁 4 类产品共计 3 000 件. 已知甲、乙、丙、丁 4 类产品数量之比为 1 : 2 : 4 : 8. 现要用分层抽样的方法从中抽取 150 件进行质量检测, 求甲类产品抽取的件数.
4. 有 140 位选手参加高尔夫球赛, 他们的成绩统计如下:

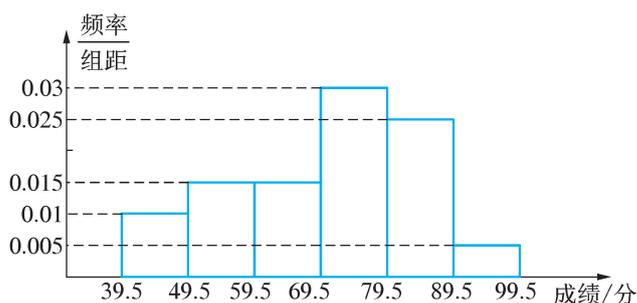
杆数	65	66	67	68	69	70	71	72	73	74	75
选手数	1	2	5	3	8	17	20	31	22	21	10

- (1) 列出频率分布表;
(2) 画出频数条形图.
5. 对你班同学进行抽样调查, 了解零花钱的使用情况.
6. 200 名学生参与研究性学习, 每人仅参加 1 个课题组. 其中参加文学类的有 33 人, 参加理化类的有 30 人, 参加数学类的有 62 人, 参加社会科学类的有 47 人, 参加信息类的有 28 人.
(1) 列出学生参加各类课题组的频率分布表;
(2) 画出频数条形图.
7. 对某种电子元件进行寿命追踪调查, 情况如下表:

寿命/h	100~200	200~300	300~400	400~500	500~600
个数	20	30	80	40	30

- (1) 列出频率分布表;
(2) 画出频率分布直方图;
(3) 估计该电子元件寿命在 100~400 h 内的百分比;
(4) 估计该电子元件寿命在 400 h 以上的百分比.
8. 从参加环保知识竞赛的学生中抽出 60 名学生, 将其成绩(均为整数)整理后画出的频率分布直方图如图所示. 观察图形, 回答下列问题:
(1) $[79.5, 89.5)$ 这一组的频数和频率分别为多少?

(2) 估计该次环保知识竞赛的及格率(60 分及以上为及格).



(第 8 题)

9. 甲、乙两名学生某门课程的 5 次测试成绩依次分别为 60, 80, 70, 90, 70 和 80, 65, 70, 80, 75, 因为 _____, 所以学生 _____ 成绩更稳定.

10. 下面是从某校高一学生中抽取的 20 名学生的学习用书的重量(单位: kg):

8.4 10.1 6.3 7.1 6.2 6.5 7.6 8.0 8.5 6.4
10.3 8.8 5.2 4.6 7.8 3.9 4.8 7.2 8.0 6.8

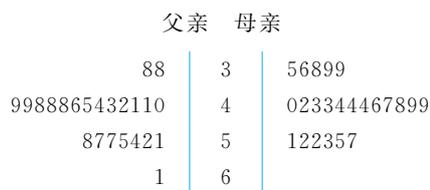
(1) 作出频率分布直方图;

(2) 利用频率分布直方图的组中值对总体平均数及方差进行估计.

11. 下面是南京市与哈尔滨市 2001 年 12 个月的月平均气温(单位: $^{\circ}\text{C}$), 试分析这两个城市的月平均气温是否具有线性相关关系. 若具有, 求出线性回归方程; 若不具有, 说明理由.

月 份	1	2	3	4	5	6
南京月平均气温	2	3.8	8.4	14.8	19.9	24.5
哈尔滨月平均气温	-19.4	-15.4	-4.8	6	14.3	20
月 份	7	8	9	10	11	12
南京月平均气温	28	27.8	22.7	16.9	10.5	4.4
哈尔滨月平均气温	22.8	21.1	14.4	5.6	-5.7	-15.6

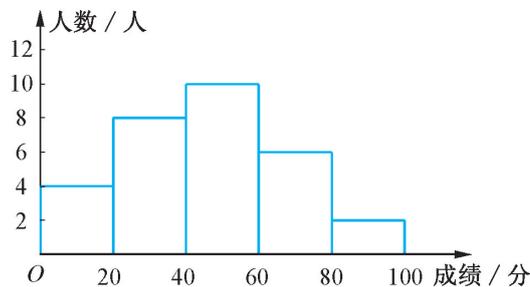
12. 下面是某班学生父母年龄的茎叶图, 试比较该班学生的父亲与母亲的平均年龄.



(第 12 题)

思考·运用

13. 抽样统计你校学生的身高与心率,并用散点图表示数据,判断两者是否线性相关.若是,求出线性回归方程;若不是,请说明理由.
14. 下面是一次考试结果的频数分布直方图,请据此估计这次考试的平均分.



(第14题)

15. 请你收集有关数据,估算我国2006年年龄为18岁的人口数.

探究·拓展

16. 青年歌手大奖赛有10名选手参加,12名评委给出的评判分数如下表:

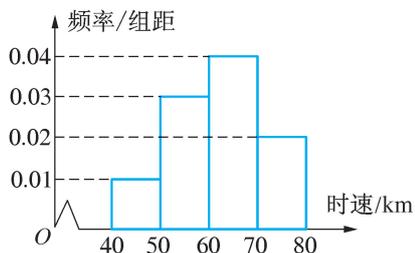
歌 手	1	2	3	4	5	6	7	8	9	10
评委1	9.07	8.98	8.80	8.80	9.00	8.81	9.24	8.91	8.98	9.05
评委2	9.12	8.92	8.83	8.86	9.05	8.85	9.28	9.02	9.06	9.10
评委3	9.15	8.95	8.91	8.90	9.12	9.00	9.28	9.05	9.10	9.12
评委4	9.15	9.00	9.05	8.93	9.15	9.00	9.29	9.15	9.10	9.18
评委5	9.15	9.01	9.10	8.93	9.15	9.04	9.30	9.16	9.10	9.20
评委6	9.18	9.02	9.17	9.03	9.15	9.15	9.32	9.20	9.10	9.20
评委7	9.21	9.15	9.20	9.05	9.16	9.15	9.35	9.21	9.14	9.20
评委8	9.21	9.20	9.20	9.05	9.17	9.18	9.35	9.21	9.17	9.24
评委9	9.24	9.20	9.24	9.15	9.18	9.19	9.38	9.21	9.17	9.29
评委10	9.26	9.23	9.28	9.17	9.18	9.24	9.40	9.28	9.22	9.30
评委11	9.30	9.30	9.29	9.21	9.24	9.25	9.45	9.29	9.29	9.30
评委12	9.35	9.32	9.31	9.21	9.31	9.32	9.45	9.30	9.40	9.35

- (1) 试确定歌手的名次;
- (2) 如何对评委的评判水平进行评价,以便确定下次聘请的10名评委?

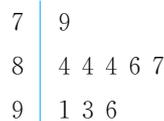
本章测试

一、填空题

1. 已知样本数据 7, 10, 14, 8, 7, 12, 11, 10, 8, 10, 13, 10, 8, 11, 8, 9, 12, 9, 13, 12, 那么这组数据落在区间 $[8.5, 11.5)$ 内的频率为_____.
2. 若 200 辆汽车通过某一段公路时的时速频率分布直方图如图所示, 则时速在区间 $[40, 50)$ 内的汽车大约有_____辆.



(第 2 题)



(第 3 题)

3. 如图是某次青年歌手大奖赛上 9 位评委给某位选手打分的茎叶图, 这组数据的众数是_____.
4. 某校共有教师 200 人, 男学生 1 200 人, 女学生 1 000 人, 现用分层抽样的方法从所有师生中抽取 1 个容量为 n 的样本, 已知从女学生中抽取的人数为 80 人, 那么 $n =$ _____.
5. 在某年的足球联赛中, 甲球队每场比赛平均失球数是 1.8, 全年比赛失球个数的标准差为 1.1; 乙球队每场比赛平均失球数是 1.5, 全年失球个数的标准差是 0.6. 有下列说法: ① 平均说来甲球队的成绩比乙球队的成绩好; ② 乙球队比甲球队防守状况更稳定. 其中, 正确的有_____.
6. 已知某班级 20 名男生俯卧撑的测试成绩统计如下, 那么 20 名男生俯卧撑的平均成绩为_____.

成绩	20	18	15	12	10	6	3
人数	1	2	4	3	6	3	1

二、选择题

7. 为了解 1 200 名学生的课外阅读情况, 打算从中抽取 1 个容量为 30 的样本. 若考虑采用系统抽样, 则分段的间隔为().
A. 12 B. 20 C. 30 D. 40
8. 在某频率分布直方图中, 从左到右共有 11 个小矩形, 若居中的那个小矩形的面积等于其他 10 个小矩形的面积和的 $\frac{1}{4}$, 且样本容量为 160, 则居中的那组数据的频数为().
A. 32 B. 0.2 C. 40 D. 0.25

9. 某设备的使用年限 x (单位: 年) 与所支出的维修费用 y (单位: 万元) 如下表所示. 已知 y 与 x 具有线性相关关系, 且线性回归方程为 $\hat{y} = mx + 1$, 那么实数 m 的值为().

A. 6 B. 4 C. 2 D. 1

x	2	3	4	5	6
y	2	4	6	6	7

10. 已知数据 x_1, x_2, \dots, x_{10} 的均值为 2, 方差为 3, 那么数据 $2x_1 + 3, 2x_2 + 3, \dots, 2x_{10} + 3$ 的均值和方差分别为().

A. 2, 3 B. 7, 6 C. 7, 12 D. 4, 12

三、解答题

11. 要从 10 部手机中抽取 4 部进行质量检验, 请写出用抽签法抽样的过程.
12. 某网站调查网民对当前网页的满意程度, 在登录的所有网民中, 收回有效帖子共 50 000 份, 其中持各种态度的份数如下表所示.

很满意	满意	一般	不满意
10 800	12 400	15 600	11 200

为了了解网民的具体想法和意见, 以便决定如何完善网页设计与维护, 打算从中抽选 500 份进行统计分析. 问: 为使样本更具有代表性, 每类中各应抽样多少份?

13. 有一个容量为 100 的样本, 数据分组及各组的频数如下: $[12.5, 15.5)$, 6; $[15.5, 18.5)$, 16; $[18.5, 21.5)$, 18; $[21.5, 24.5)$, 22; $[24.5, 27.5)$, 20; $[27.5, 30.5)$, 10; $[30.5, 33.5]$, 8.

(1) 列出样本的频率分布表;

(2) 估计总体中在 $[21.5, 30.5)$ 的数据约占的百分比.

14. 对划艇运动员甲、乙两人在相同的条件下进行了 6 次测试, 测得他们最大速度的数据如下. 试根据这些数据, 比较他们的运动水平.

甲	27	38	30	37	35	31
乙	33	29	38	34	28	36

15. 在某种产品表面进行腐蚀刻线试验, 得到腐蚀深度 y (单位: mm) 与腐蚀时间 x (单位: s) 之间的一组观察值如下表:

x/s	5	15	30	50	70
y/mm	6	10	16	19	24

(1) 画出表中数据的散点图;

(2) 求线性回归方程;

(3) 预测腐蚀时间为 60 s 时的腐蚀深度.