

## 第 5 章 Excel 在数学建模中的应用

Excel 是 Microsoft Office 套件中的电子表格软件，它的应用很广泛，许多人把它当作一般的制作表格和图表的软件，而不清楚它的强大数据运算能力。其实，Excel 内置了数百个函数供用户调用，还允许用户根据自己的需要随意定义自己的函数，Excel 无需编程就能够实现其他软件需要编程才能完成的复杂计算，能进行各种数据的统计、运算、处理和绘制统计图形，只要善于开发，Excel 一定能够在数学建模中发挥出更大的作用。

### 5.1 Excel 的数据处理功能

在实际问题中，经常会遇到各种数据需要处理，有时数据量比较大，还有各种统计图表需要绘制，如果不熟悉数据处理的方法和统计软件的使用，则可能面对大量数据而束手无策，或者处理起来效率低，耗费大量时间。Excel 擅长数据统计，用它来处理数据能够节省大量时间，提高效率。

Excel 的数据处理功能主要有两大块：

#### 1) 计算功能

Excel 有强大的计算能力，它提供了 300 多个内部函数给用户使用，还允许自定义函数。当大批数据都要用同一个公式计算结果时，只要用鼠标拖动而不需要编程。如果你能熟练应用此项功能，你会感到惊喜，原来 Excel 有如此强大的运算能力！你将获得事半功倍的效果。

#### 2) 数据分析功能

Excel 提供了“数据分析”工具包，内含方差分析、回归分析、协方差和相关系数、傅里叶分析、 $t$  检验等分析工具。

#### 5.1.1 Excel 的函数

Excel 2000 提供了 12 类(类别有常用、财务、日期与时间、数学与三角函数、统计、查找与引用、数据库、文本、逻辑、信息、工程、用户定义)共 300 多个各种内部函数，其中用得比较多的是常用、数学与三角函数以及统计类中的函数。

函数由函数名、参数组成。不同函数对其参数有不同要求，若参数为数值，则可用单元格取代(如 A2，A 代表第 A 列，2 代表第 2 行，A2 代表 A2 格子内的一个数)，有些函数的参数是多个数(数组)，则可用区域取代(如 B2:B9，代表 B 列

从第 2 行到第 9 行的 8 个数, 构成一个数组), 有些函数的参数是矩阵, 则可用矩形区域取代(如 A2 : C4, 其内的数字构成矩阵).

先把光标放在表格的空白格, 点击“插入”→“函数”, 或者点击工具栏中  $f_x$  图标, 弹出插入函数对话框, 如图 5.1.1 所示. 对话框的“选择类别”栏目中显示函数类别, 点击栏目右侧的向下三角形 ▼, 将能看到所有类别.

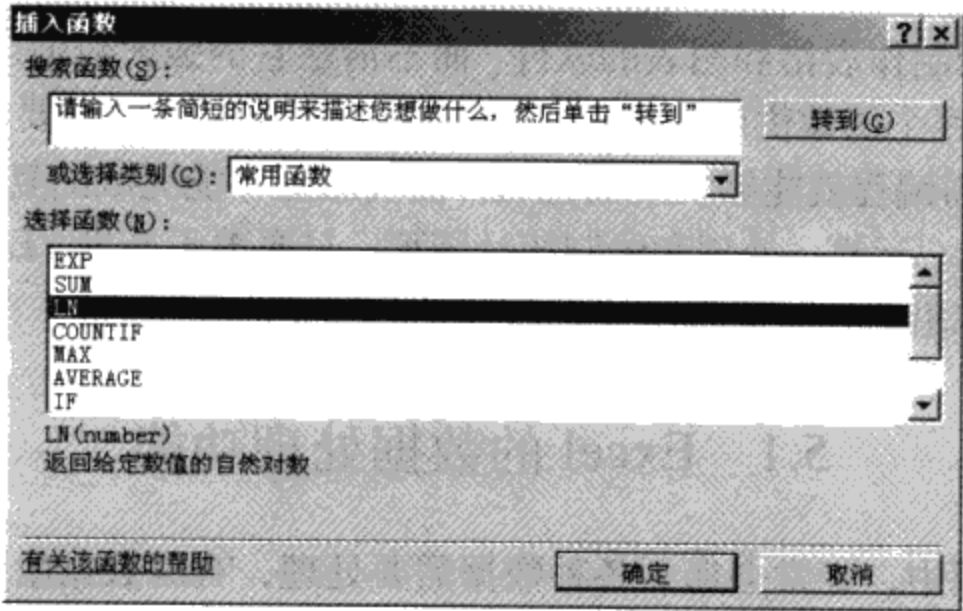


图 5.1.1 插入函数对话框

1. 常用函数

当插入函数对话框的选择类别中显示“常用函数”时, 共有十多个函数供选择, 它们的功能和参数如表 5.1.1 所示.

表 5.1.1 Excel 的常用函数

函 数 名	功 能	参 数
EXP	计算 $e^x$	任意实数
SUM	求和	数组, 如 A2:A10
LN	求自然对数 $\ln x$	正实数
COUNTIF	统计满足某种条件的数据个数	数据区域和条件
AVERAGE	求算术平均值	数组
IF	由条件决定返回值	一个条件, 两个结果
COUNT	统计个数	数组
MAX	求最大值	数组
SIN	正弦	以弧度表示的角度
SUMIF	满足某种条件的所有数据的和	数据区域和条件
HYPERLINK	创建一个快捷方式或链接	路径和文件名、标识符

选择其中一个函数，然后点“确定”，弹出函数参数对话框(图 5.1.2)，列出需要输入的参数，对话框中有必要的文字说明。如果在参数栏目内输入符合要求的参数，则立即显示计算结果，点“确定”，则该结果将被写入表中的当前位置(光标所在的位置)。

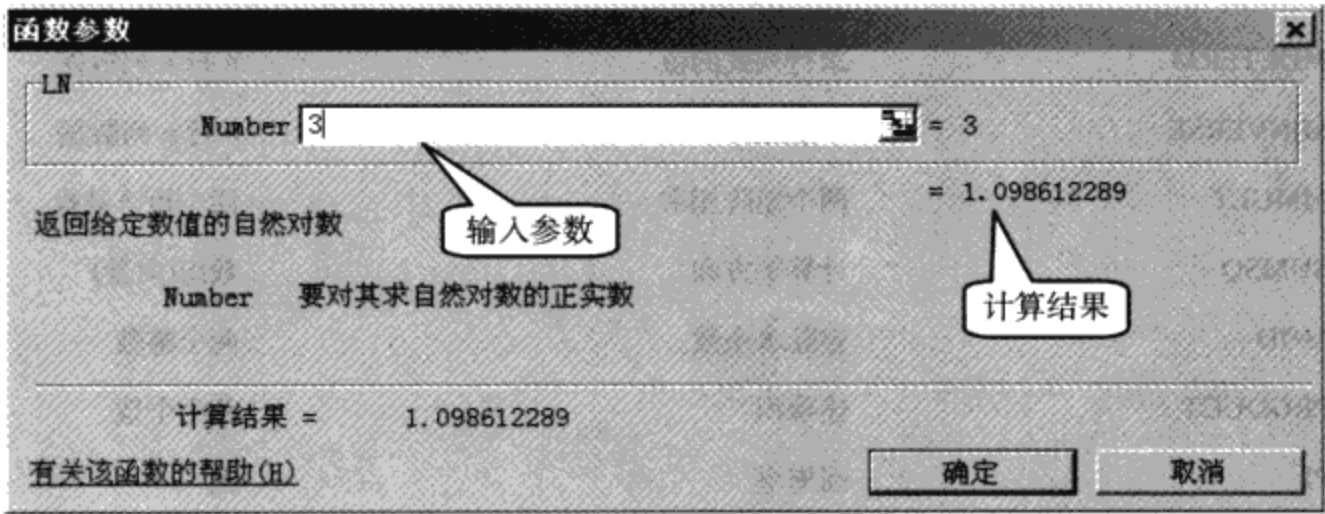


图 5.1.2 函数参数输入对话框

2. 数学与三角函数

数学与三角函数是数值计算时经常用到的函数。在插入函数对话框中选择数学与三角函数，则显示出 58 种函数供选择，其中常用的函数见表 5.1.2。

表 5.1.2 常用的数学与三角函数

函 数 名	功 能	参 数
三角函数 SIN,COS,TAN	求三角函数值	以弧度表示的角度
反三角函数 ASIN,ACOS,ATAN	求反三角函数值	定义域内的数
双曲函数 SINH,COSH,TANH	求双曲函数值	实数
反双曲函数 ASINH,ACOSH,ATANH	求反双曲函数值	定义域内的实数
POWER	$x$ 的 $y$ 次方	两个数 $x$ 和 $y$
EXP	$e^x$	数 $x$ 或单元格
SQRT	$x$ 的平方根	同上
LOG	给定底的对数	真数和底数
LOG10	10 为底的对数	真数或单元格
LN	自然对数	真数或单元格

续表

函 数 名	功 能	参 数
ABS	$x$ 的绝对值	数 $x$ 或单元格
FACT	计算 $n$ 阶乘	整数 $n$
COMBIN	组合数 $C_n^r$	$n$ 和 $r$ 两个整数
MDETERM	求行列式的值	$n$ 行 $n$ 列数据
MINVERSE	求逆矩阵	$n$ 行 $n$ 列数据
MMULT	两个矩阵相乘	两个矩阵数据
SUMSQ	计算平方和	数组(向量)
MOD	整除求余数	两个整数
PRODUCT	连乘积	若干个数
PI	圆周率	无
DEGREES	弧度转换成度	弧度
RADIANS	度转换成弧度	度
LCM	最小公倍数	若干个数
GCD	最大公约数	若干个数
RAND	0-1 之间均匀分布随机数	无
RANDBETWEEN	两个数之间的随机数	两个数
SUMXMY2	两个数组对应数值差的平方和	两个数组
SERIESSUM	求幂级数的和	满足要求的四个数
SIGN	符号函数	实数

还有一些舍入或取整函数没有一一列出, 如 INT, 功能是向下取整.

#### 例 5.1.1 计算 $e^{-2}$ .

**解** 把光标放在表格的空白格, 点击“插入”→“函数”, 或者点击工具栏中  $f_x$  图标, 弹出插入函数对话框, 选择常用类别或数学与三角函数类别中的 EXP, 点击“确定”, 弹出 EXP 函数的对话框, 在 Number 栏目内输入 -2, 点击“确定”, 则光标处的原空白格内显示计算结果 0.13533528...

#### 例 5.1.2 计算 $\sqrt{2} + \ln 3$ 的值.

**解** 点击任一空白单元格, 输入=SQRT(2)+LN(3), 鼠标点击其他地方, 则公式所在的单元格内显示计算结果 2.51282585... 实际上本例的函数 SQRT(2)+LN(3) 是两个内部函数相加, 属于自定义函数.

例 5.1.3 求矩阵  $A = \begin{bmatrix} 1 & 1 & 0 & 1 \\ 1 & 2 & 2 & 2 \\ 2 & -2 & 2 & 1 \\ 3 & -1 & 5 & 3 \end{bmatrix}$  的逆矩阵.

解 求逆矩阵的函数名为 MINVERSE，它的自变量(参数)要求是矩阵，其结果也是矩阵，为此有一些特殊之处. 首先在 4 行 4 列空白区域(如 A1 至 D4)内输入矩阵 A，点击空白格(如 A6)，输入=MINVERSE(A1:D4)，或按照例 5.1.1 中插入函数的方法，选择数学类别中的 MINVERSE，在输入参数栏目内输入 A1:D4，然后点击“确定”，结果只显示了一个数字-4，没有显示完整的逆矩阵，要显示完整逆矩阵的方法是：用鼠标选中 A6 至 D9 一块 4×4 区(恰好能容纳 A 的逆矩阵 16 个数字)，先按 F2 键，再同时按下 Shift+Ctrl+Enter 三个键，则选定的区域内出现逆矩阵的计算结果，如图 5.1.3 所示.

	A	B	C	D
1	1	1	0	1
2	1	2	2	2
3	2	-2	2	1
4	3	-1	5	3
5				
6	-4	7	8	-6
7	-3	5	5	-4
8	-3	4	4	-3
9	8	-12	-13	10

图 5.1.3 求逆矩阵的计算结果

3. 统计函数

在插入函数对话框中选择统计函数，则显示出 76 种函数供选择. 其中常用的统计函数见表 5.1.3.

表 5.1.3 常用的统计函数

函数名	参数	功能(返回值)
AVERAGE	$n$ 个数	求算术平均值
VAR	$n$ 个数	样本方差 $\frac{\sum x_i^2 - n\bar{x}^2}{n-1}$
VARP	$n$ 个数	总体方差 $\frac{\sum x_i^2 - n\bar{x}^2}{n}$
STDEV	$n$ 个数	VAR 的平方根
STDEVP	$n$ 个数	VARP 的平方根

续表

函数名	参 数	功 能(返回值)
DEVSQ	$n$ 个数	$\sum_{i=1}^n (x_i - \bar{x})^2 = \sum x_i^2 - n\bar{x}^2$
AVEDV	$n$ 个数	$\frac{\sum  x - \bar{x} }{n}$
NORMSDIST	数 $x$	标准正态分布的分布函数值
NORMDIST	$x, \mu, \sigma, 1$ 或 $0$	正态分布, 1: 返回分布函数, 0: 返回概率密度
NORMINV	$\alpha, \mu, \sigma$	正态分布概率为 $\alpha$ 时的 $x$ 值
NORMSINV	$\alpha$	标准正态分布由 $\alpha$ 得 $x$
CHIDIST	$x$ , 自由度 $n$	$\chi^2$ 分布 $P\{X > x\}$
CHIINV	$\alpha, n$	$\chi^2$ 分布, 由 $\alpha$ 查 $x$
CHITEST	两组数据	两组数据同分布的概率
POISSON	$k, \lambda, 1$ 或 $0$	0: 返回对应 $k$ 的泊松分布概率 1: 返回累积概率
BINOMDIST	$k, n, p, 1$ 或 $0$	0: 返回二项分布的概率值 $C_n^k p^k q^{n-k}$ 1: 返回累积概率
EXPONDIST	$x, \lambda, 1$ 或 $0$	指数分布, 1: 返回分布函数值, 0: 返回概率密度
TDIST	$x, n, 1$ 或 $0$	$t$ 分布, 1: 返回分布函数, 0: 返回概率密度
TINV	$\alpha, n$	$t$ 分布满足 $P\{T < \alpha\}$ 的 $T$ 值
FDIST	$x$ , 自由度 $n_1, n_2$	$F$ 分布满足 $P\{F < x\}$ 的 $F$ 值
FINV	$\alpha, n_1, n_2$	由 $\alpha$ 查 $F$ 分布临界值
CONFIDENCE	$\alpha, \mu, n$ (数据个数)	总体均值的置信区间(半长度)
COVAR	两组数	协方差
CORREL	两组数	相关系数
FTEST	两组数据	两组数方差相等的概率
CRITBINOM	$n, p, \alpha$	二项分布的临界值(分位数)
SLOPE	两组数	线性回归 $y = a + bx$ 中的 $b$
INTERCEPT	两组数	线性回归 $y = a + bx$ 中的 $a$
LINEST	数组 $y$ , 多维数组 $x$ , 以及逻辑值 $c, s$	多元线性回归 $y = a + \sum b_i x_i$ , $c=0$ 时强制 $a=0$ , $s=1$ 时返回附加回归统计值
LOGEST	同上	指数回归 $y = b \prod_{i=1}^k m_i^{x_i}$ 中的 $b, m_i$
GEOMEAN	$n$ 个数	几何平均数
HARMEAN	$n$ 个数	调和平均数(倒数平均值的倒数)
MIN	$n$ 个数	$n$ 个数中的最小值

以上概率统计函数中,有些函数的名称有一定规律性,凡是后4个字母为DIST的函数,如 NORMDIST、CHIDIST、FDIST、TDIST 等,功能是返回某种分布的分布函数值或概率密度值(如果函数的最后一个参数是逻辑值 1 或 0,则该值为 1 时返回分布函数值或累积概率,为 0 时返回概率密度或分布律的值);如果把前面所说函数名称中的 DIST 改成 INV,如 NORMINV、CHIINV、FINV、TINV 等,它们是对应 DIST 函数的反函数,功能是给定概率反查自变量的值。

#### 4. 自定义函数

能利用现成的库函数当然应当尽量利用,从而省时省力,但实际计算过程中库函数难以完全满足用户的愿望,需要自己定义函数,称为自定义函数。任何一个计算软件,如果不具备允许用户按自己的意愿定义函数的功能,则该计算软件的使用范围很有限。

Excel 允许用户自己定义任意带参数函数,方法是:把光标放在空白处(点击空白格子),先输入一个等号=,然后输入自定义函数的表达式。表达式可以由常量、变量、内部函数和运算符组成,其中运算符包括算术运算符(−,\*,/,^,%,+,-)、比较运算符(=,<,>,<=,>=,<>)和连接符&。举例如下:

例 5.1.4 当  $x=3, 2, 1, 0, -1, -2, -3$  时,计算分段函数  $y = \begin{cases} x \sin x, & x > 0, \\ e^x \cos x, & x \leq 0 \end{cases}$

的值。

**步骤** (1) 选一个空白列(如 D 列)输入自变量  $x$  的值,在第一行(位置编号为 D1,称为一个单元格)输入 3,点击第二行(单元格 D2),输入=D1-1,鼠标点击其他单元格,则 D2 单元格内显示 D1-1 的计算结果 2,再点击 D2 单元格,则它的边框出现加粗的黑色且右下角有一个黑点,如图 5.1.4 所示。用鼠标拉着该黑点向下拖动一直到 D7 单元格,放开鼠标,则 D3 至 D7 单元格内的计算结果依次显示为 1, 0, -1, -1, -3, 如果你继续向下拖动黑点(称为复制公式),则每个单元格的数字是上一行数字依次减 1,即拖动公式(函数所在的单元格)时,如果公式中的自变量是单元格的名称(编号),则随着拖动,公式的自变量作相应变化,计算结果也跟着变化,其变化规律是:纵向拖动(复制)公式时,行号跟着变化而列标不变;横向拖动(复制)公式时,列标跟着变化而行号不变。

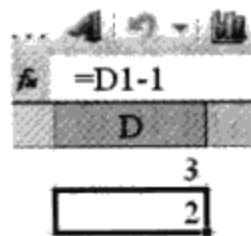


图 5.1.4 边框右下角的黑点

(2) 选另外一个空白列,如标号为 E 的一列,点击第一行 E1 单元格,输入=IF(D1>0,D1\*SIN(D1),EXP(D1)\*COS(D1)),这是分段函数的表达式,注意公式中的自变量为 D1,表示用 D1 中的数字作为函数的自变量。鼠标点击其他单元格,

则 E2 单元格内显示自定义函数的计算结果, 再点击 E2 单元格, 拉着它的边框右下角的黑点, 向下拖动到 E7 放开, 则 E3 至 E7 单元格内依次得到自定义函数当自变量分别为 D3 至 D7 时的值, 如图 5.1.5 所示.

=IF(D4>0,D4*SIN(D4),EXP(D4)*COS(D4))	
D	E
3	0.423360024
2	1.818594854
1	0.841470985
0	1
-1	0.19876611
-2	-0.05631935
-3	-0.049288824

图 5.1.5 自定义函数的计算结果

### 5. 利用自定义函数完成较复杂的计算

从例 5.1.4 我们已经看到, 表达式(自定义函数)可以复制(拖动), 且函数的自变量能够自动改变, 我们利用该功能就能够完成大批量数据计算以及各种较复杂的计算(用其他软件通常需要编程才能进行的计算), 举例如下:

**例 5.1.5** 用迭代法能求非线性方程  $x - \cos x = 0$  的数值解, 迭代公式是  $x_k = \cos(x_{k-1})$ , 取  $x_0 = 1$ , 试用 Excel 计算, 要求精度达到  $10^{-12}$ .

**解** 在某一空白列(如 A 列)的第一个位置(A1)处输入初始值 1, 点击单元格 A2, 输入=COS(A1), 得到计算结果 0.540302306, 然后连续向下拖动黑边框右下角的小黑点, 产生的效果是按迭代公式  $A_k = \cos(A_{k-1})$  不断进行迭代, 放开鼠标就能看见计算结果, 此时单元格内显示的数字格式为小数点后面 9 位, A55 之后的数字不再变化, 说明迭代 55 次之后计算结果的精度达到  $10^{-9}$ . 为了显示小数点后面更多位数, 先选择该列从 A2 开始的单元格, 然后从主菜单选择“格式”—“单元格”, 弹出“单元格格式”对话框, 点数字栏目, 选“数值”, 把小数位数栏目内的数字改为 16, 如图 5.1.6 所示, 点击“确定”, 则数字的显示格式变成小数点后面 16 位.

此时继续向下拖动表达式, 可以观察到迭代结果的精度变化, 如果要求精度达  $10^{-12}$ , 则大约需要迭代 75 次, 结果为 0.739085133215; 如果要求精度达到  $10^{-16}$ , 则大约需要迭代 92 次, 结果为 0.7390851332151610. Excel 的计算精度通常最多能有 16 位有效数字, 继续增加小数点后面位数将无效.

从本例的计算过程我们可以发现, Excel 用于复杂计算有两大优点:

- (1) 不需要编写程序, 这对不熟悉编程, 但急需计算的人员比较实用;
- (2) 显示结果比较直观, 能看见中间结果, 便于数据分析.



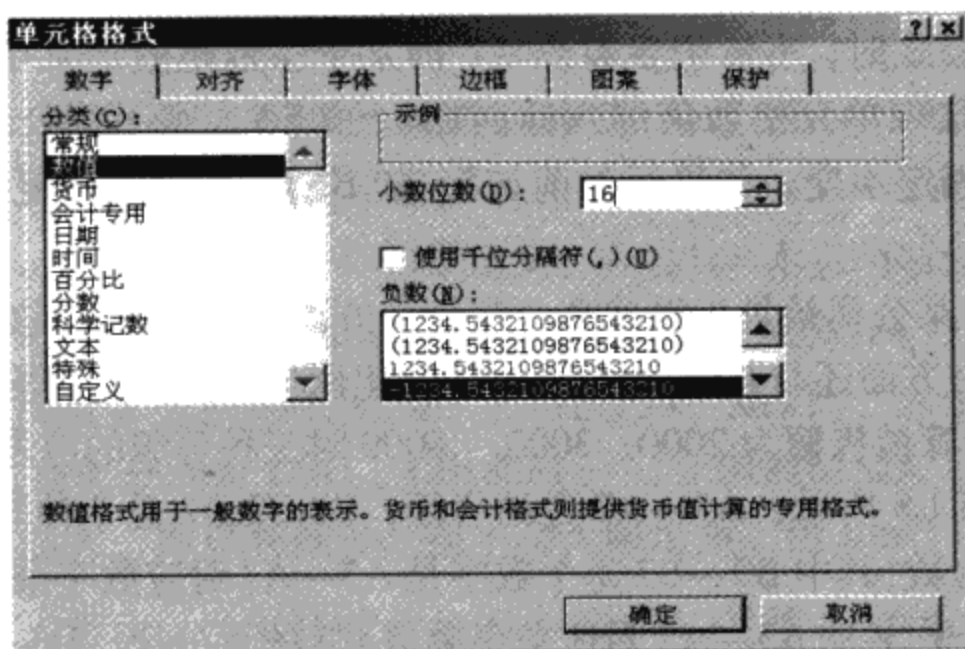


图 5.1.6 设置单元格格式

**例 5.1.6** 利用公式  $\frac{\pi}{2} = 1 + \frac{1}{3} + \frac{1}{3 \cdot 5} + \frac{1}{3 \cdot 5 \cdot 7} + \frac{1}{3 \cdot 5 \cdot 7 \cdot 9} + \dots$  计算  $\pi$  的近似值, 使误差小于  $10^{-14}$ .

**解** 设变量的初始值为  $n=1$ ;  $m=3$ ;  $t=1$ ;  $p=1$ ; 然后在循环中运算:  $n=n+1$ ;  $m=m+2$ ;  $t=t \cdot n/m$ ;  $p=p+t$ ;  $\pi=p \cdot 2$ . 在 Excel 中的计算方法: 在第一行前 4 列依次输入:  $n$ 、 $m$ 、1、1, 在第二行的前 5 列依次输入 1、3、 $=A2/B2 \cdot C1$ 、 $C2+D1$ 、 $D2 \cdot 2$ , 在第三行的前 2 列依次输入  $A2+1$ 、 $B2+2$ , 然后从第一列开始把每一列的公式依次向下拖动, 第 5 列的计算结果就是  $\pi$  的值, 设置第 5 列的数值显示格式为小数点后面 15 位. 可以看出从第 46 行开始, 计算结果稳定在 3.14159265358979, 见图 5.1.7, 此时计算精度已经达到  $10^{-14}$ , 看来这是本例能达到的最高精度.

37	36	73	2.13E-12	1.570796327	3.141592653585650
38	37	75	1.05E-12	1.570796327	3.141592653587750
39	38	77	5.18E-13	1.570796327	3.141592653588780
40	39	79	2.56E-13	1.570796327	3.141592653589290
41	40	81	1.26E-13	1.570796327	3.141592653589550
42	41	83	6.24E-14	1.570796327	3.141592653589670
43	42	85	3.08E-14	1.570796327	3.141592653589730
44	43	87	1.52E-14	1.570796327	3.141592653589760
45	44	89	7.53E-15	1.570796327	3.141592653589780
46	45	91	3.72E-15	1.570796327	3.141592653589790
47	46	93	1.84E-15	1.570796327	3.141592653589790

图 5.1.7 计算  $\pi$  的近似值

**归纳** 如果公式中的自变量是单元格的名称(编号), 则随着表达式的拖动, 公式的自变量作相应变化, 计算结果也跟着变化, 其变化规律是: 纵向拖动(复制)公式时, 行号跟着变化而列标不变; 横向拖动(复制)公式时, 列标跟着变化而行号不变.

如果复制公式时,不希望参数改变,即某个参数是固定某个单元格中的数,为此在公式中代表单元格数值的列标前加\$,如\$A,则不管公式被复制到什么位置上,列标固定不变,如果行号前加\$,如B\$3,则公式被复制时,行号固定不变.

**例 5.1.7** 某公司给员工发奖金,奖金一方面与销售额挂钩(按销售额的一定比例提成),另一方面还与其他指标挂钩(提成比例分为三等:一等 1.5%,二等 1%,三等 0.5%),计算销售额为 2000, 3000,...,6000 时三种等级的应发奖金数.

**解** 如图 5.1.8 所示, A3—A7 是销售额, B2,C2,D2 是 3 种等级, B3—D7 是计算出来的奖金数,其中 B3—B7 每个数字是 A3—A7 对应数字乘以 B2 的百分比,而 C3—C7 每个数字是 A3—A7 对应数字乘以 C2 的百分比, D3—D7 每个数字是 A3—A7 对应数字乘以 D2 的百分比,在 B3 单元格内输入自定义计算公式  $=\$A3*B\$2$ , 公式中 \$A 的作用是不管公式复制到何处,均以 A 列为基数, \$2 的作用是奖金等级始终以第二行的百分比计算. B3 的结果计算出来之后,只需把 B3 单元格右下角的黑点向下并且向右拖动到 D7,则表内所有应发奖金数都能正确计算出来.

**注** B3—D7 的单元格数字显示格式设置为:数值,小数点位数 0.

	A	B	C	D
1		一等	二等	三等
2	销售额	1.50%	1%	0.50%
3	2000	$=\$A3*B\$2$	20	10
4	3000	45	30	15
5	4000	60	40	20
6	5000	75	50	25
7	6000	90	60	30

图 5.1.8 标号加\$则不变

### 例 5.1.8 连续复利问题.

假设银行活期存款年利率为  $r$ (如  $r=1.8\%$ ),若某储户存 10000 元活期存款,那么满一年后,他可以得到利息  $10000r$ ,本息合计  $10000(1+r)$ 元,因为银行允许活期存款随便什么时候支取,如果储户满半年就结算一次,此时的本息合计为  $10000(1+r/2)$ ,把本息取出来以后立即把本息一起再存活期,半年后再次结算,则全年的本息合计为  $10000(1+r/2)^2$ ,因  $(1+r/2)^2 = 1+r+r^2/4 > 1+r$ ,我们发现每半年结算一次的获利比一年结算一次多.试计算每季度、每月、每半月、每天结算一次并立即把本息再存活期情况下的全年获利,你可以发现:活期存款存期越短(即结算越频繁)获利越多!假如活期存款的利息可以按小时,甚至按分钟来结算,那么当储户连续不断地取款再存款,他能依靠这种方式来发大财吗?

解 设  $n$  为一年中的结算次数,  $a_k$  为第  $k$  次结算时的本息,  $a_0=10000$  元是首次存入的本金, 则每次结算时的应得利率为  $r/n$ , 第一次结算时的本息为  $a_1=(1+r/n)a_0$ , 第二次结算时的本息为  $a_2=(1+r/n)a_1=(1+r/n)^2a_0$ , 以此类推可得全年本息为  $a_n=(1+r/n)^na_0$ . 如果定义: 收益比=本息/本金, 则当一年结算一次时年度收益比为  $1+r$ , 而一年结算  $n$  次时年度收益比为  $(1+r/n)^n$ .

当全年结算次数  $n$  分别等于 1、2、4、12、24、365 时, 在 Excel 中计算(图 5.1.9)得到年度收益比分别为 1.018、1.018081、...、1.018162525. 规律是: 结算次数越多, 收益比越大, 即获利越多, 但不会无限增多, 因  $\lim_{n \rightarrow \infty} (1+r/n)^n = e^r$ , 故即使一年中结算无数次, 收益比存在极限, 最多为  $e^{0.018} = 1.018162976$ , 与一年结算一次的收益比 1.018 相比较, 多出来 0.000162976, 在本金 10000 元的情况下, 利息只多出 1.63 元, 所以依靠增加结算次数的方式虽然能增加收益比, 但是发不了大财.

	A	B	C
1	10000		
2	0.018		
3			
4	1	$= (1 + \$A\$2 / A4)^{A4}$	1.0180
5	2	1.018081	10180.81
6	4	1.018121865	10181.21865
7	12	1.018149245	10181.49245
8	24	1.018156107	10181.56107
9	365	1.018162525	10181.62525
10	8760	1.018162958	10181.62958

图 5.1.9 连续复利问题

上述利率的计算方法称为连续复利率, 连续复利率的极限为  $e^r$ , 其中  $r$  为某种基准利率(如活期存款年利率 1.8%), 如果按连续复利率存  $k$  年, 则收益比为  $e^{kr}$ .

5.1.2 Excel 的数据分析功能

Excel 提供了一组称作“数据分析”的统计分析工具包, 内含方差分析、回归分析、协方差和相关系数、傅里叶分析等分析工具, 使用这组分析工具, 可以大大提高工作效率和质量.

在默认安装下, Excel 并不直接提供数据分析工具包, 首次使用时需要进行安装, 方法如下:

- (1) 点击“工具”——“加载宏”;
- (2) 在弹出对话框中列出各种可以加载的项目, 按照需要选择“分析工具库”、

“规划求解”、“与 Access 链接”等等项目，点“确定”；

(3) 如果需要，把 Office 光盘放入光驱，然后按提示进行安装。

安装完成后，“工具”菜单中多出了“数据分析”子菜单，点击它，弹出对话框，显示各种数据分析工具。该工具包含有 19 个工具，大致可分成 5 类，见表 5.1.4。

表 5.1.4 Excel 的数据分析工具

基础分析	检验分析	相关，回归	方差分析	其 他
描述统计	$z$ 检验	协方差	单因素	指数平滑
直方图	$F$ 检验	相关系数	双因素	傅立叶分析
排位	$t$ 检验	回归分析	无重复双因素	随机数发生器
抽样分析				移动平均

下面介绍几种数据分析功能和用法。

### 1. 描述统计

主要统计数据的平均值、中位数、标准差、方差等等统计量，举例说明其用法。

例 5.1.9 某炼钢厂测了 120 炉钢中的 Si 含量，得数据如下：

0.86,0.83,0.77,0.81,0.81,0.8,0.79,0.82,0.82,0.81,0.81,0.87,0.79,0.82,0.78,0.8,0.81,  
0.87,0.81,0.77,0.78,0.77,0.78,0.77,0.77,0.77,0.71,0.95,0.78,0.81,0.8,0.77,0.76,0.82,  
0.8,0.82,0.84,0.79,0.9,0.82,0.79,0.82,0.79,0.86,0.76,0.78,0.83,0.75,0.82,0.78,0.73,  
0.83,0.81,0.81,0.83,0.89,0.81,0.86,0.82,0.82,0.78,0.84,0.84,0.84,0.81,0.81,0.74,  
0.78,0.78,0.8,0.74,0.78,0.75,0.79,0.85,0.75,0.74,0.71,0.88,0.82,0.76,0.85,0.73,0.78,  
0.81,0.79,0.77,0.78,0.81,0.87,0.83,0.65,0.64,0.78,0.75,0.82,0.8,0.8,0.77,0.81,0.75,  
0.83,0.9,0.8,0.85,0.81,0.77,0.78,0.82,0.84,0.85,0.84,0.82,0.85,0.84,0.82,0.85,0.84,  
0.78,0.78.

解 在 Excel 中的 A2—A121 区域内输入 120 个原始数据(或打开数据文件)，然后从菜单上选“工具”—“数据分析”，在弹出对话框中选择“描述统计”，弹出如图 5.1.10 所示描述统计对话框。

在输入区域填入 A1:A121，表示第 A 列第 1 行至第 121 行是需要分析的原始数据，因第一行是表头(标志)，故在对话框的“标志位于第一行”上打上“√”，输出区域定位于 C1，平均数置信度打上“√”，用默认的 95%或根据需要改成其他百分比，点击“确定”，得到分析结果，如图 5.1.11 所示。主要结果有平均值、中位数、标准差、方差、峰度、偏度等，具体含义请参阅概率统计教科书的相关

内容.

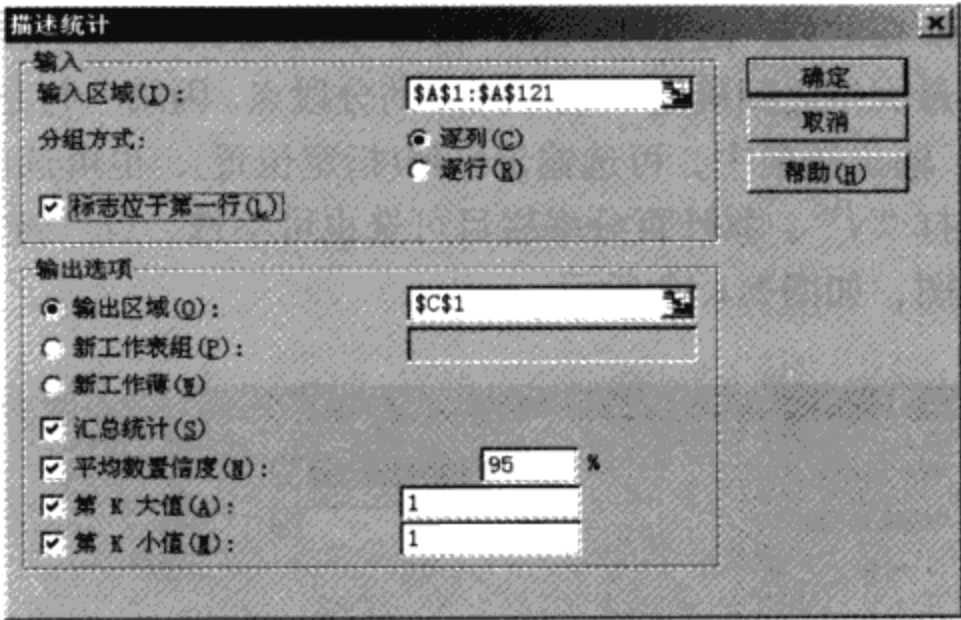


图 5.1.10 描述统计对话框

	A	B	C	D
1	钢的Si含量		钢的Si含量	
2	0.86			
3	0.83	平均		0.8025
4	0.77	标准误差		0.0041069
5	0.81	中位数		0.81
6	0.81	众数		0.81
7	0.8	标准差		0.0449883
8	0.79	方差		0.0020239
9	0.82	峰度		2.195563
10	0.82	偏度		-0.3228812
11	0.81	区域		0.31
12	0.81	最小值		0.64
13	0.87	最大值		0.95
14	0.79	求和		96.3
15	0.82	观测数		120
16	0.78	最大(1)		0.95
17	0.8	最小(1)		0.64
18	0.81	置信度(95.0%)		0.008132

图 5.1.11 描述统计的结果

2. 直方图

直方图是一大批数据的频率分布图，由直方图可以观察和分析数据的概率分布。画直方图的步骤如下：

(1) (在 A 列)输入原始数据，进行描述分析，确定数据的最小值和最大值，把数据所在区域分成若干个小小区间，确定分段点，如例 5.1.9 的数据，最小值为 0.64，最大值为 0.95，假如把该区域分成 16 个小小区间(等间隔，允许不等间隔)，每个小小区间的长度为 0.02，则分段点依次为 0.655,0.675,...,0.955(按由小到大排列)。然后在 B 列输入这些分段点数据。此处分段点比原始数据多一位小数，保证数据不会恰好落在小小区间边界(分段点)上。

(2) 点“工具”——“数据分析”——“直方图”，弹出直方图对话框，如图 5.1.12

所示. 其中“输入区域”是指原始数据所在的区域, 这里是 A1:A121, “接收区域”是指分段点所在的列, 这里填入 B1:B17, 如果空白不填, 则 Excel 会自动在数据的最小值与最大值之间确定一组等间隔的分段点. 因第一行是表头, 故在“标志”上打“√”, 输出选项中, 可选输出区域(指定位置), 也可选新工作表组, 在“图表输出”上打“√”, 累计百分率栏目可选也可不选, 点“确定”. 得到数据统计结果和直方图, 如图 5.1.13 所示.

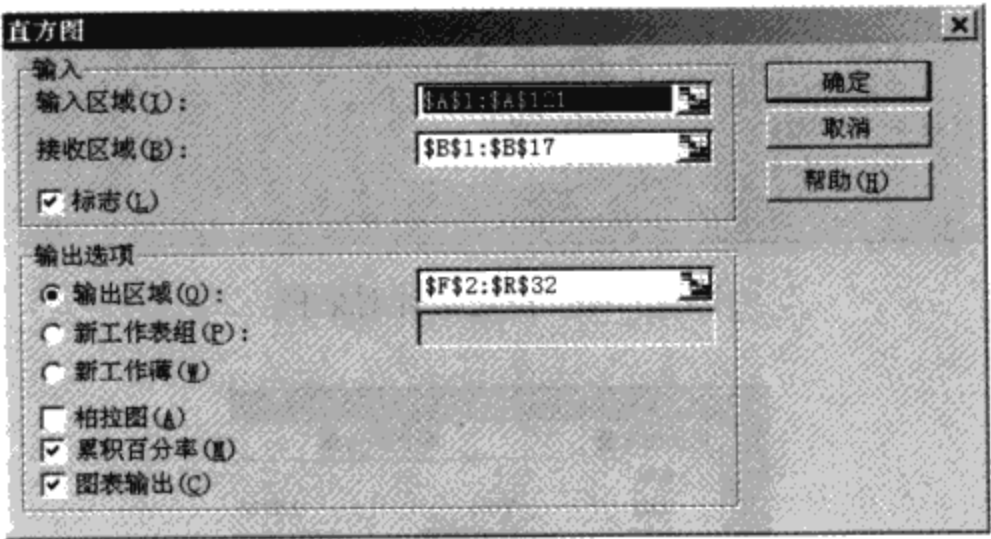


图 5.1.12 直方图对话框

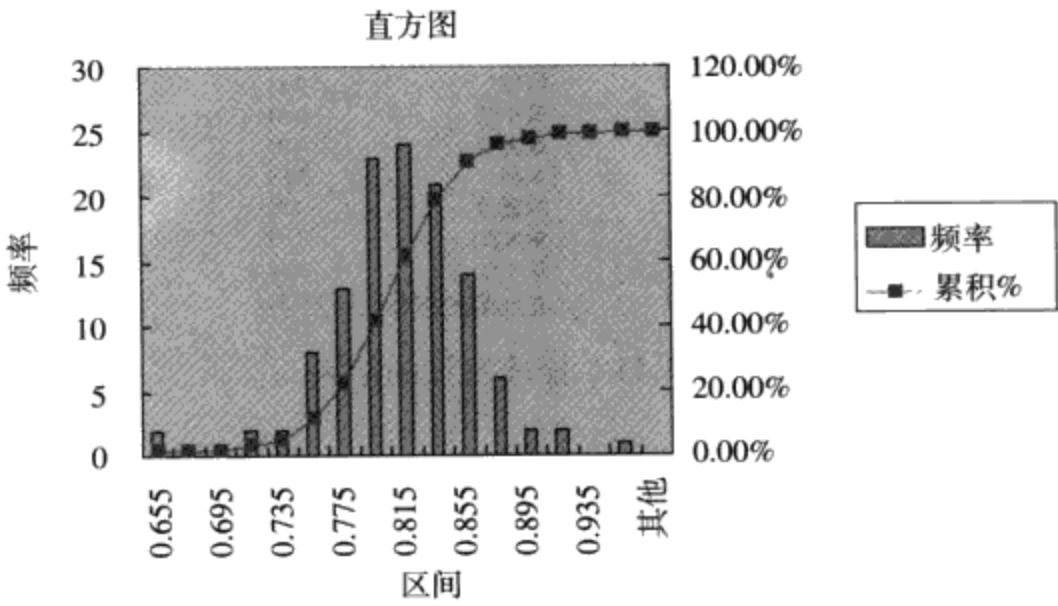


图 5.1.13 直方图

## 5.2 用 Excel 绘制图表

图表是一种直观有效的常用工具, 通过图表, 可以把大量的数据转换成各种格式的直观图形, 便于用户快速地分析数据之间的对比、关联、变化趋势等相互关系. Excel 提供强大的图表绘制功能, 可以非常简便地建立各种统计图表, 如直方图、柱形图、散点图、饼图、条形图、折线图等. 对话框以向导的方式引导用



户使用，既直观又方便。即使初次使用，也能很快掌握。

5.2.1 创建图表的步骤

创建一个图表通常要 5 个主要步骤：


1. 准备数据

数据是图表的依据，要创建图表必须先准备好数据。例如，2004 年数学建模竞赛 A 题(奥运会临时超市网点设计)中的原始数据，经过统计得到如图 5.2.1 所示统计表。

	A	B	C	D	E	F
1	交通工具	人数	消费金额	人数	用餐方式	人数
2	公交(南北)	1774	0-100	2060	中餐	2382
3	公交(东西)	1828	100-200	2629	西餐	556
4	出租车	2010	200-300	4668	商场餐饮	2651
5	私家车	958	300-400	983	合计	10600
6	地铁(东)	2006	400-500	15		
7	地铁(西)	2023	500以上	103		
8	合计	10599	合计	10600		

图 5.2.1 奥运会临时超市网点设计中的统计结果

2. 打开“图表向导”

从菜单选“插入”—“图表”，或者工具栏中的按钮，即可启动“图表向导”，向导中有“标准类型”和“自定义类型”两种类型供选择。

(1) 标准类型。有柱形图、条形图、折线图、饼图、XY 散点图、面积图、圆环图、雷达图、曲面图、气泡图、股价图、圆柱图、圆锥图和棱锥图共 14 类型，每种类型又包含若干个子类，每个子类均用图形表示，如图 5.2.2 所示。你可以选



图 5.2.2 图表向导中的标准类型

择合适的图表类别及子类，然后点击“下一步”。

(2) 自定义类型. 在图表向导中选择自定义类型，出现“内部”和“自定义”两种选择，若选“内部”则出现内置的 20 种图形类型供挑选：彩色堆积图、彩色折线图、带深度的柱形图、对数图、分裂的饼图、管状图、黑白饼图、黑白面积图、黑白折线图、黑白柱形图、蜡笔图、蓝色饼图、两轴线-柱图、两轴折线图、平滑直线图、线柱图、悬浮的条形图、圆锥图、柱状-面积图、自然条形图，如图 5.2.3 所示，选择合适的类型，点击“下一步”，出现“图表源数据”对话框。

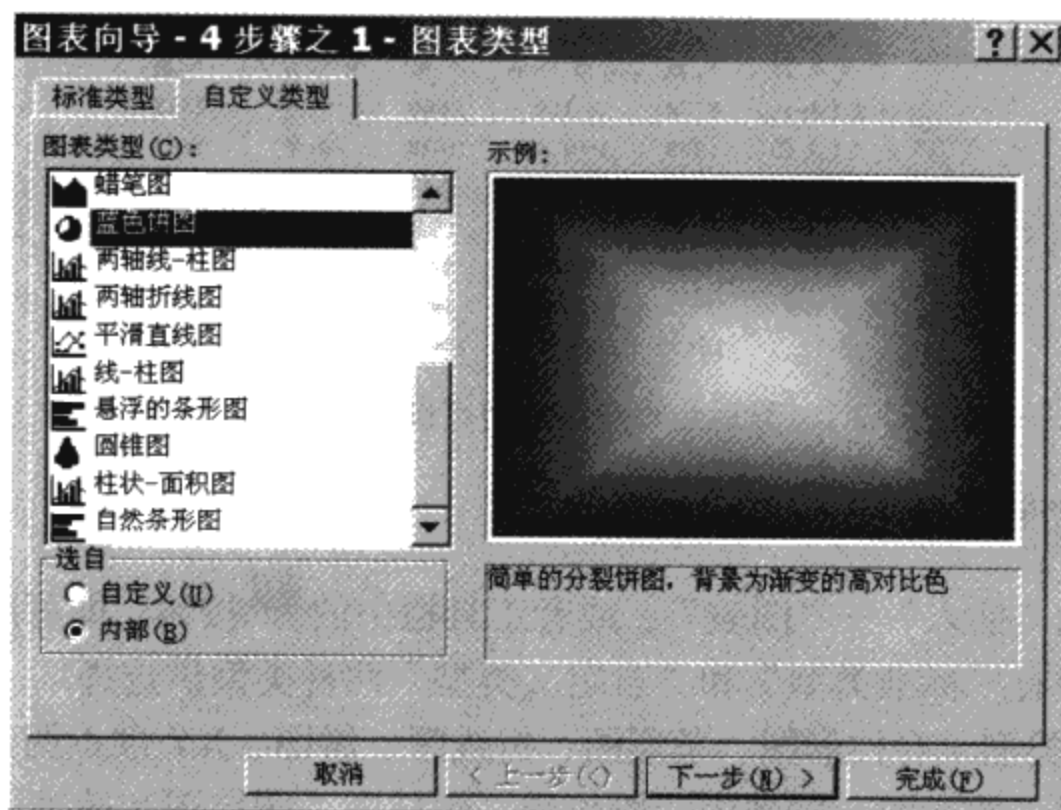


图 5.2.3 图表向导中的内部自定义类型

### 3. 指定数据位置

在“数据区域”栏目内输入数据所在位置，例如输入 A1:B7，该区域的第一行和第一列是表头(文字说明)，数据按列摆放，故对“系列产生在”栏目的两个选项“行”和“列”作出选择“列”(图 5.2.4)，点击“下一步”，出现“图表选项”对话框。

### 4. 设定图表选项

“图表选项”对话框(图 5.2.5)用来设定图表的标题、坐标轴、网格线、图例、数据标志、数据表等项目，具体功能说明如下。





图 5.2.4 确定图表源数据

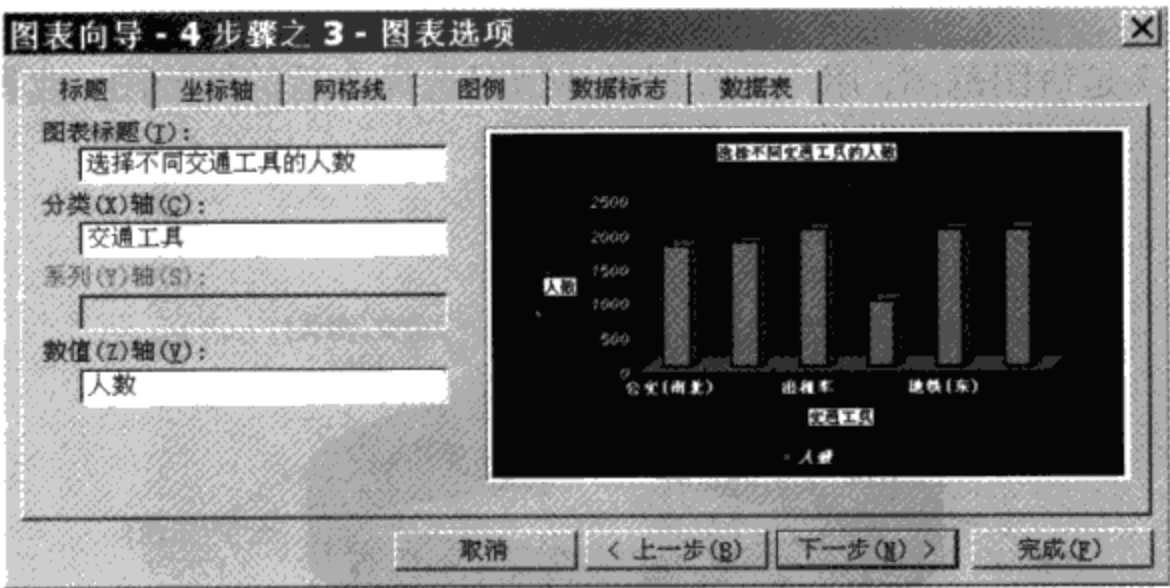


图 5.2.5 图表选项对话框

- (1) 标题选项. 设置图表的标题、坐标轴的文字说明.
- (2) 坐标轴选项. 设置是否显示坐标轴及其刻度.
- (3) 网格线选项. 设定是否显示网格线.
- (4) 图例选项. 设定是否显示图例及其位置.
- (5) 数据标志选项. 设置是否显示数据的名称、数据值等标志.
- (6) 数据表选项. 设置是否显示数据列表.

以上选项的设定有直观图形显示在对话框的右半部分, 所见即所得, 立即能看见效果, 用户可根据需要和爱好决定如何设置.

全部选项设置好以后, 点击“下一步”, 出现图表位置对话框.

5. 设定图表位置

选择“作为新工作表插入”或者“作为其中的对象插入”均可. 点击“完成”，出现“选择不同交通工具人数”的柱形图，如图 5.2.6 所示.

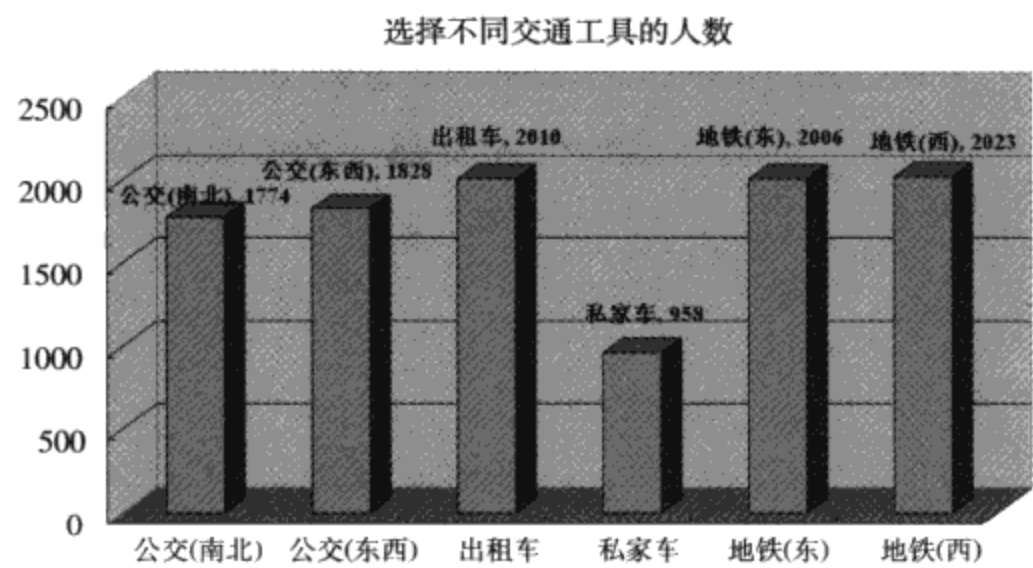


图 5.2.6 用图表向导生成的图表

图 5.2.7 是饼图范例，用户可根据具体情况选择合适的统计图，或者创建自定义类型.

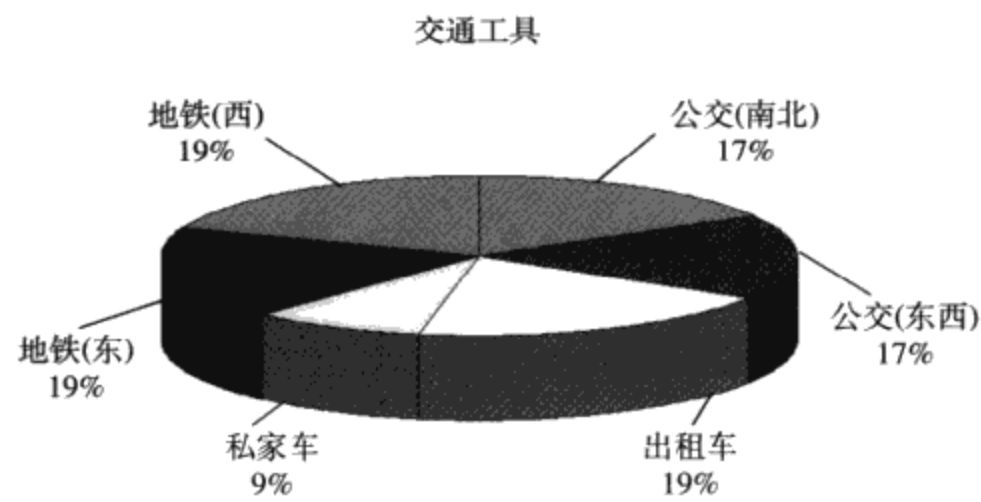


图 5.2.7 饼图范例

5.2.2 编辑和修改图表

在使用图表向导时，可以对图表的标题、坐标轴、网格线、图例、数据标志、数据表等图表组成部分(称为图表的元素或对象)进行设置，但这种设置是粗略的框架，它一般不涉及字体、字型、字号、前景色、背景色、坐标刻度、线条颜色等细节. 图表向导生成的图形通常不够美观，一些细节往往不中意，需要进行编辑、修改、美化和完善.

## 1. 图表的组成

图表由各种元素(部件)组成,以图 5.2.8 的柱形图为例,其组成部分有图表区、绘图区、标题、坐标轴(分类轴和数据轴)、背景墙、网格线、数据标志和数据系列,当鼠标在图上移动时,会弹出相应的元素名称.图 5.2.8 标出了各元素的具体位置,这些元素均可以分别进行设置或修改.

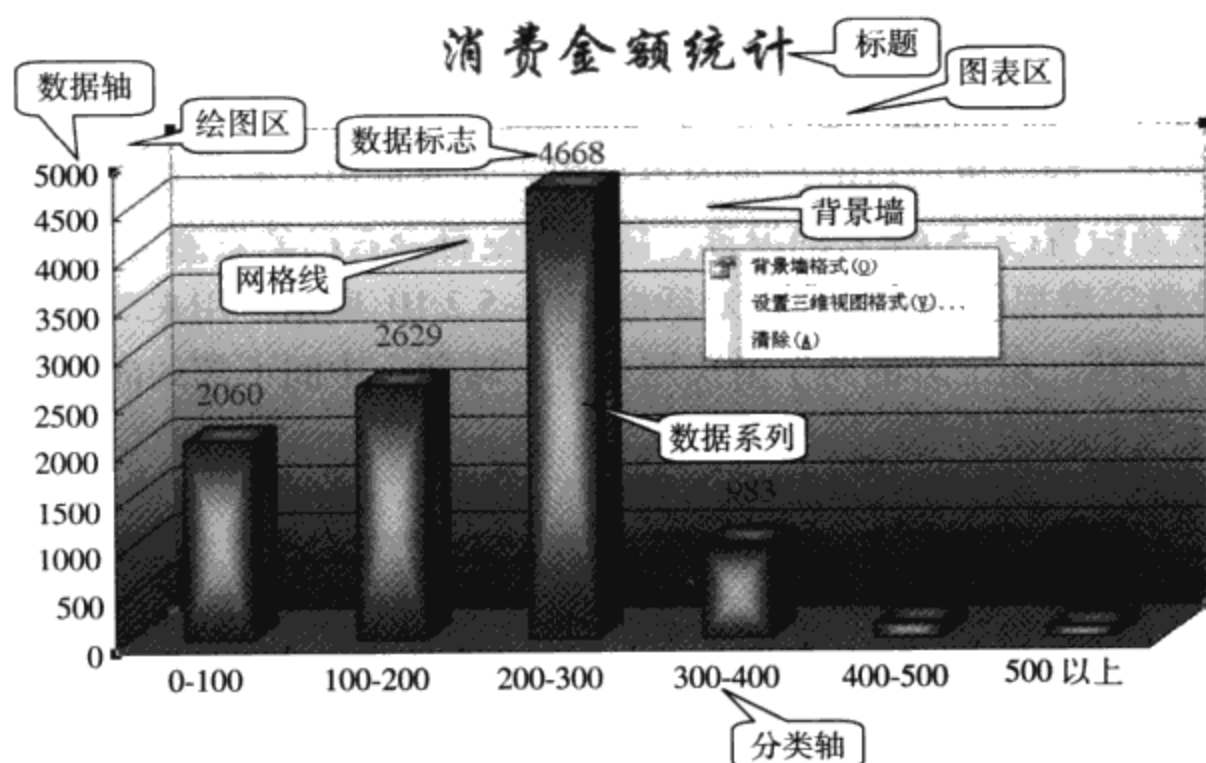


图 5.2.8 用鼠标右键弹出菜单

## 2. 图表的编辑

主菜单的“图表”项目下有“图表类型”、“源数据”、“图表选项”、“位置”、“添加数据”、“设置三维视图”等二级菜单,如图 5.2.9 所示.先选中(点击)某个已经创建的图表,点击主菜单“图表”,在二级菜单中选择其中任一选项,都会弹出一个对话框.二级菜单各项目的主要功能说明如下:

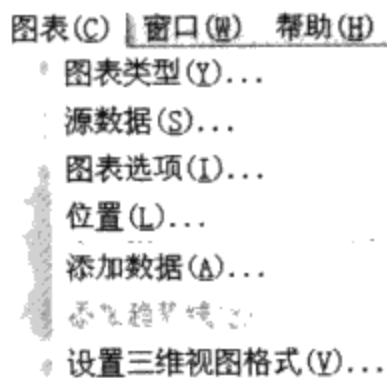


图 5.2.9 图表二级菜单

(1) 图表类型. 功能与图表向导中的“图表类型”对话框(图 5.2.2)类似, 用来更改已创建的当前图表的类型.

(2) 源数据. 用于重新指定源数据的位置, 功能及设置方法与图表向导中的“图表源数据”对话框(图 5.2.4)类似.

(3) 图表选项. 对话框与图表向导中的“图表选项”对话框(图 5.2.5)相似, 用来更改图表的标题、坐标轴、网格线、图例、数据标志、数据表等项目, 具体设置方法与图表向导相同.

(4) 图表位置. 用来更改图表位置, 有“作为新工作表插入”或者“作为其中的对象插入”两种设置供选择.

以上四个二级菜单项目的设置内容与图表向导是相似的, 功能是在图表生成之后用来重新设置(更改)原来的设置, 使图表更符合自己的意愿.

(5) 设置三维视图格式. 弹出对话框如图 5.2.10 所示. 用来对三维视图上下转动和左右旋转, 左上方的两个箭头用来上下转动, 中间下部的向左、向右两个弯箭头用于左右转动图形.

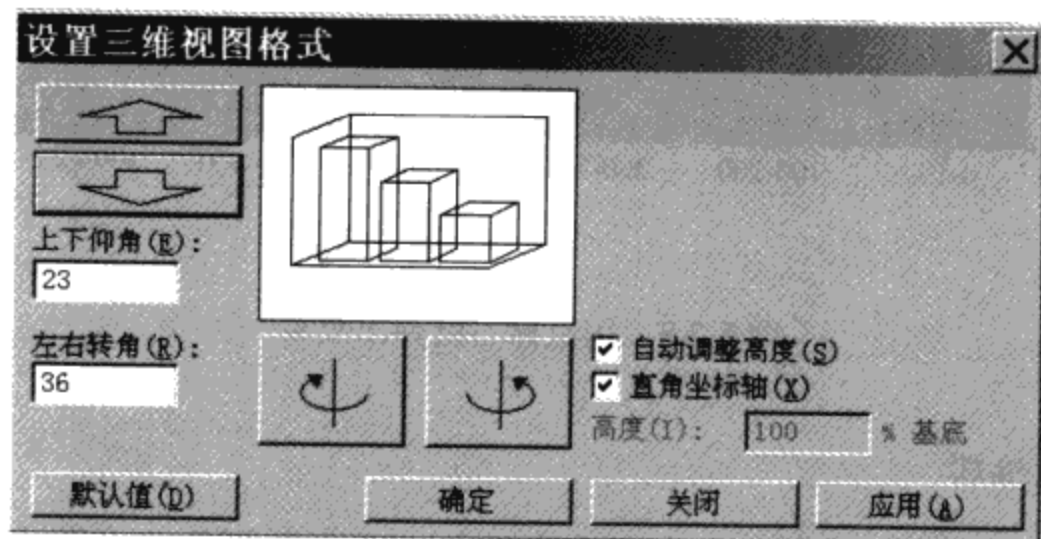


图 5.2.10 设置三维视图格式对话框

### 3. 图表元素的修改(美化)

使用图表向导时无法对字体、字型、字号、前景色、背景色、坐标刻度、线条颜色等细节进行预先设置, 通常采用默认形态, 生成的图形只能算成粗样图表, 元素的一些细节不够美观, 需要进行编辑、修改、美化和完善.

把鼠标移到某个元素上并按右键(称为右键点击, 简称右击), 将弹出一个快捷菜单(图 5.2.8 中显示了右击背景墙以后的弹出菜单), 点击的元素不同, 则弹出菜单的项目也有所不同, 但至少都会包括“××项格式”和“清除”两项. 若选“清除”则从图表中删除该元素, 而“格式”项则用来修改该元素的颜色、图案、线条、字体、字号、格式、刻度等等. 下面举例说明修改方法.

### 1) 标题的修改

有两种办法可以调出“图表标题格式”对话框，一种是鼠标移到标题上然后按左键(称为点击，或称单击)，此时标题四周出现边框(表示标题被选中)，此时可用鼠标拉着标题移动到别处放开(移动定位)，若点击主菜单“格式”，二级菜单第一项即为“图表标题”，点击它，则弹出“图表标题格式”对话框；另一种办法是把鼠标移到标题上并按右键(称为右键点击，简称右击)，出现弹出菜单，它有“图表标题格式”和“清除”两个选项，选择“图表标题格式”，则弹出对话框，如图 5.2.11 所示。

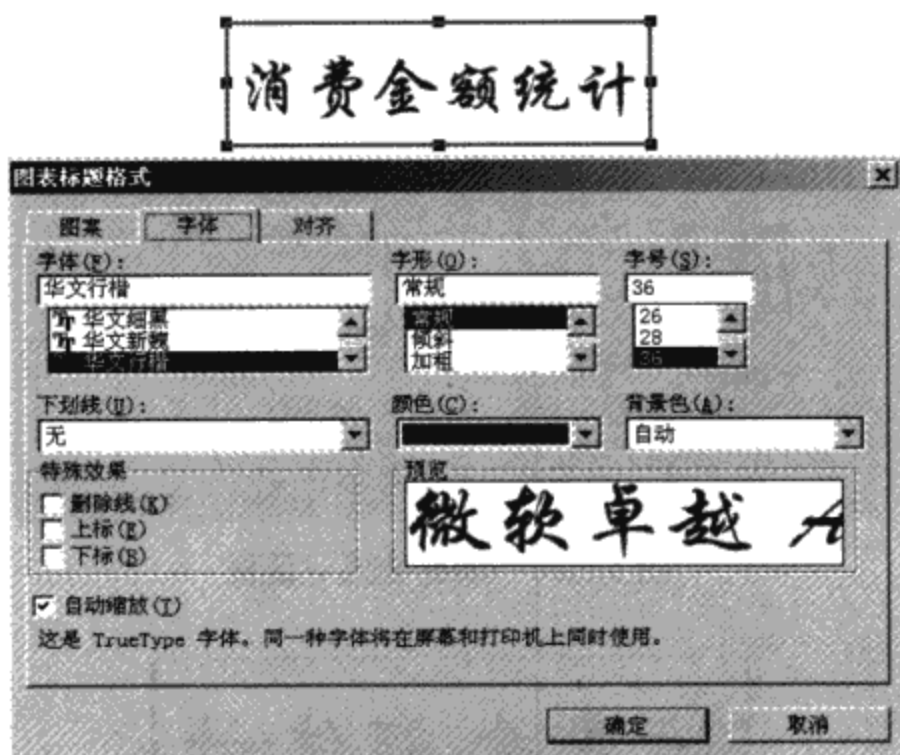


图 5.2.11 图表标题格式对话框

该对话框有“图案”、“总体”、“对齐”三个选项，图中显示“字体”选项，用来设置标题的字体、字形、字号、颜色、特殊效果等。“图案”选项内又有两个选项，一个是“边框”，用于设置边框的颜色、宽度、图案、是否有阴影等；另一个是“区域”，用于设置背景部分的颜色和特殊效果等，点击其中的“填充效果”又会弹出一个对话框，如图 5.2.12 所示。

该对话框用来设置标题背景的填充效果，有“渐变”、“纹理”、“图案”、“图片”四个选项，设置方法与 Office 套件中的其他软件(如 Word, PowerPoint)相类似，你可以根据自己的喜爱进行设置。图 5.2.13 是经过修饰的标题示例。

### 2) 设置数据系列格式

点击图中的柱体，使每个柱体都被选中(各柱体的四角都出现控制标记)，然后点击主菜单的“格式”或者右击柱体，出现二级菜单或者弹出式菜单，都有“数据系列格式”选项，选择它，则弹出数据系列格式对话框，如图 5.2.14 所示。对话框里面有“图案”、“形状”、“数据标志”、“系列次序”、“选项”共 5

个选项. 其中“图案”用于设置柱体表面的颜色和图案效果, 设置方法与标题的背景相同. “形状”选项由于设置柱体的形状, 提供方形柱体、圆柱体、圆锥体、棱锥体、梯形柱体等 6 种形状给用户挑选. “数据标志”的功能和设置方法与图表向导相同(图 5.2.5). “选项”用来设置柱体的间隔、宽度、透视深度等尺寸.

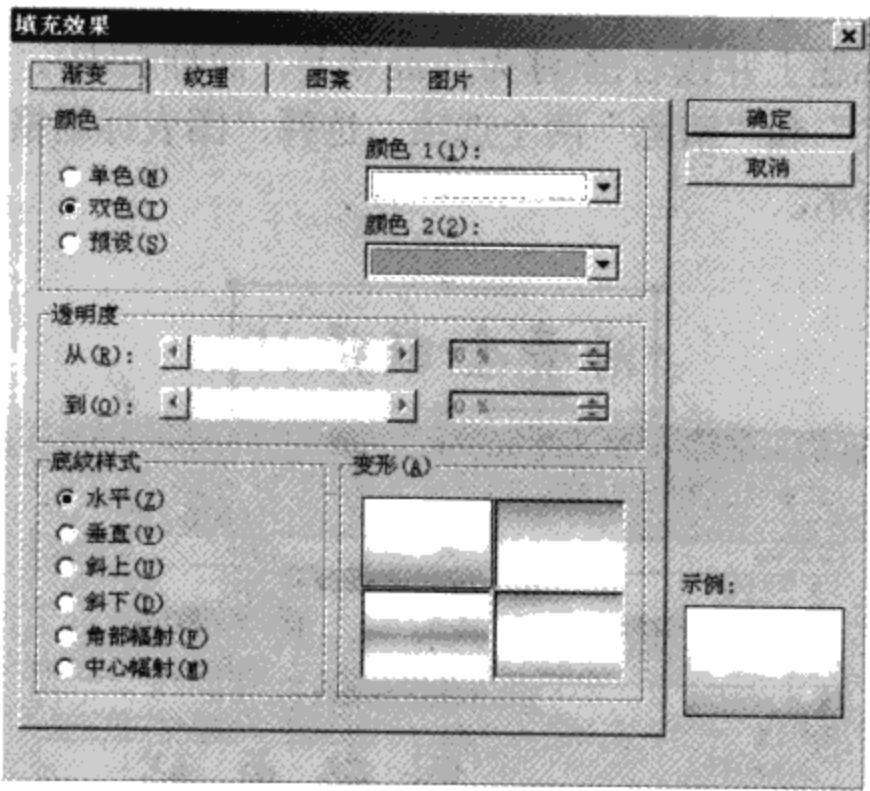


图 5.2.12 填充效果对话框



图 5.2.13 修饰后的标题

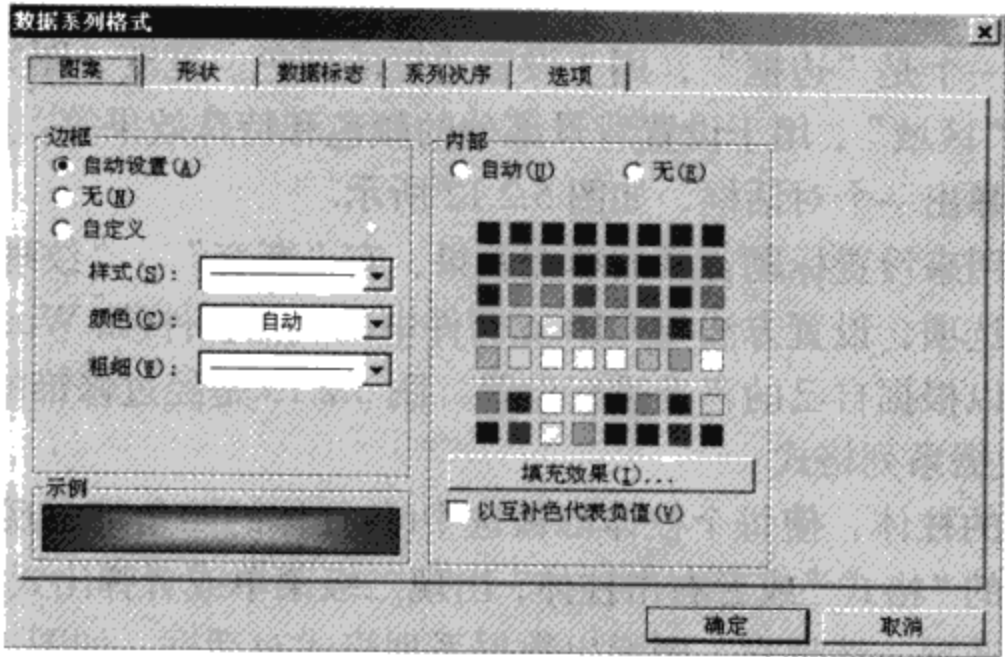


图 5.2.14 数据系列格式对话框



### 3) 修改背景

右击背景区，在弹出菜单中选“背景墙格式”，弹出背景墙格式对话框，见图 5.2.15，点击对话框中的填充效果，又弹出一个对话框(与图 5.2.12 相同)，在颜色栏目里选择“预设”，预设颜色栏目内有多种预设方案供选择，点击栏目右边的▼按钮，列出各种预设方案的名称，你可以从中选择一种试试，如选择“茵茵绿原”，见图 5.2.16。在“底纹样式”项目中选择一种，例如选择“中心辐射”，在“变形”栏目中选择一种，单击“确定”返回上一个对话框，再点击“确定”，即可看到你设置的背景效果。

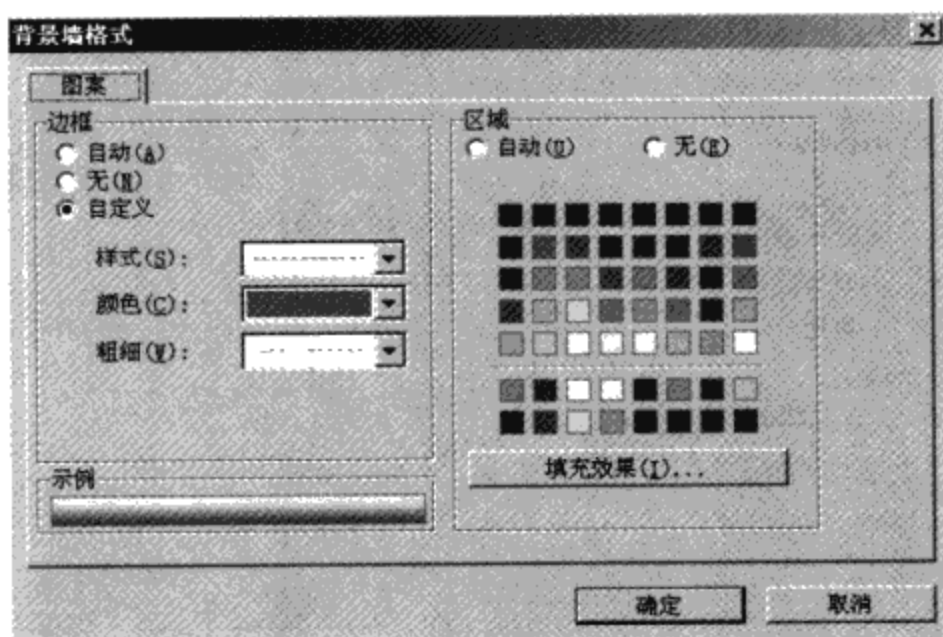


图 5.2.15 背景墙格式对话框

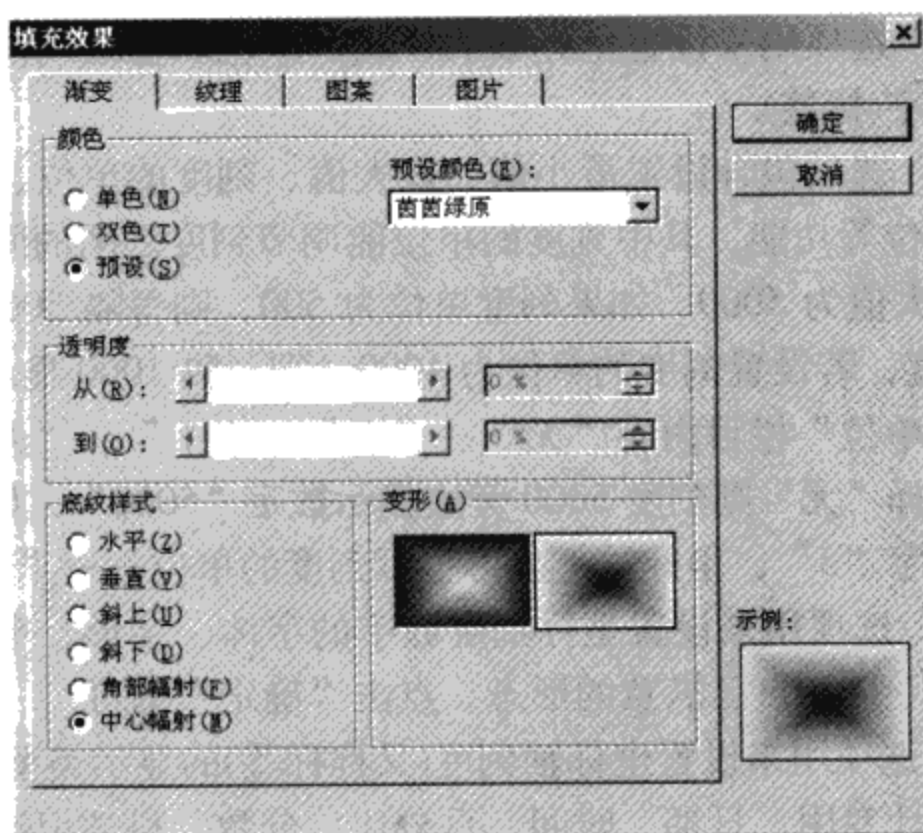


图 5.2.16 填充效果对话框

也可以在填充效果对话框的“颜色”栏目内选择“双色”，颜色 1 栏目选一种浅色(如白色)，颜色 2 栏目可以选青绿色或浅青色等亮一些的冷色调，在“底纹样式”项目中选择“水平”，在“变形”栏目中上白下青样式，点击“确定”即可看到上白下青逐渐过渡的图表背景。

#### 4) 修改坐标轴格式

右击坐标轴区域，弹出“坐标轴格式”对话框，如图 5.2.17 所示。该对话框中有“图案”、“刻度”、“字体”、“数字”、“对齐”5 个选项。分别说明如下：

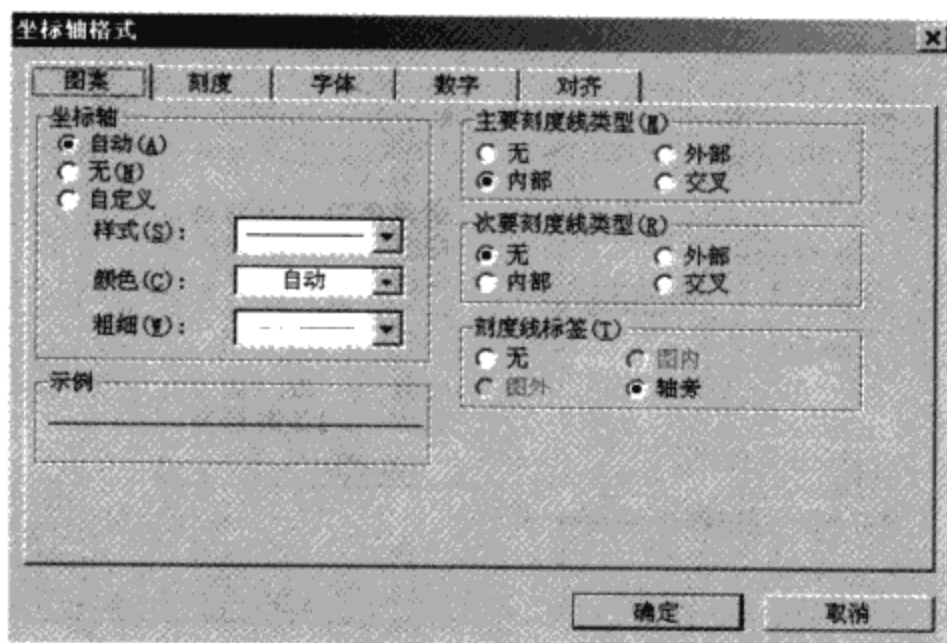


图 5.2.17 坐标轴格式对话框

(1) “图案”选项用来设置坐标轴的样式、颜色和粗细以及刻度线的类型和是否标注坐标数字(标签)。

(2) “刻度”选项包含设置最小值、最大值、刻度的单位、两个坐标轴的交点位置、显示单位等功能，其中刻度的单位能调节刻度之间的间隔，例如刻度的最小值为 0，最大值为 5000，如果刻度单位为 500，则每隔 500 显示坐标刻度，共显示 11 个刻度，若设置刻度的单位为 1000，则每隔 1000 显示刻度，共显示 6 个刻度。“显示单位”栏目内有“无”、“百”、“千”、“万”、“十万”等等选择，如果选择“无”则刻度 5000 旁边显示数字“5000”，如果选择“千”则刻度旁边显示数字“5”，坐标轴上方可显示刻度的单位——“千”字。

(3) “字体”选项用来设置坐标刻度数字的字体、字形、字号、前景色、背景色和特殊效果。选择自己满意的形态，点击“确定”即可。

(4) “数字”选项用来设置坐标轴刻度(又称标签)的显示类型和格式，有常规、数值、货币、会计专用、日期、时间、百分比、分数、科学记数、文本、特殊和自定义等类型供选择。用户可根据实际情况作出选择，如果坐标轴的单位是数字，



则选择“数值”，如图 5.2.18 所示，出现“小数位数”，其内的数字代表小数点后面多少位。如果数字的小数点后面位数太多，标签的数字太长，可以少设几位，如设 0 位，则仅显示整数部分。

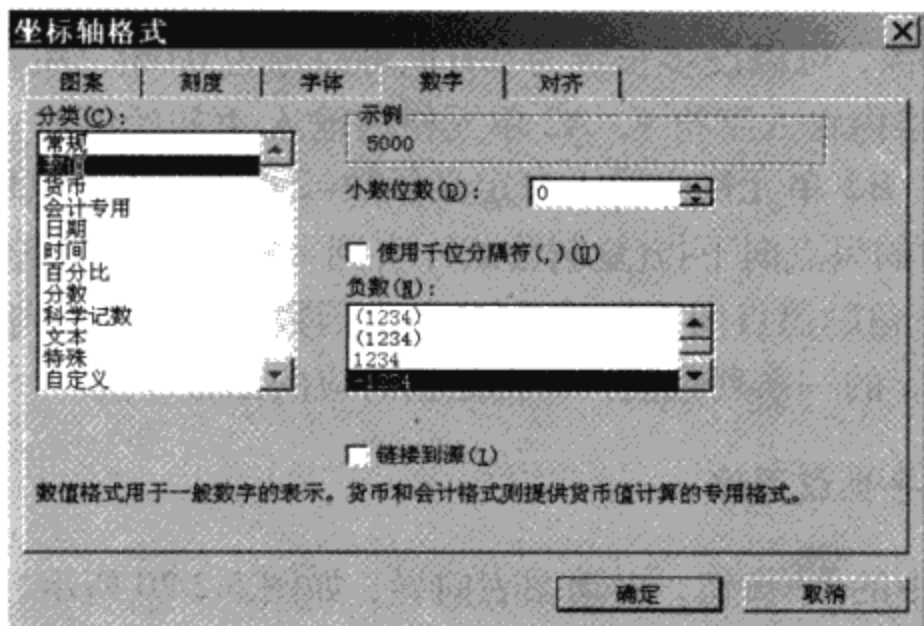


图 5.2.18 坐标轴标签的格式

(5) “对齐”选项设置标签文字的排版方向(角度)，见图 5.2.19。用鼠标可以拉着右边文本的指针转动(图上显示转到了 30 度的位置)，如果点确定，则坐标标签(刻度)是倾斜的。

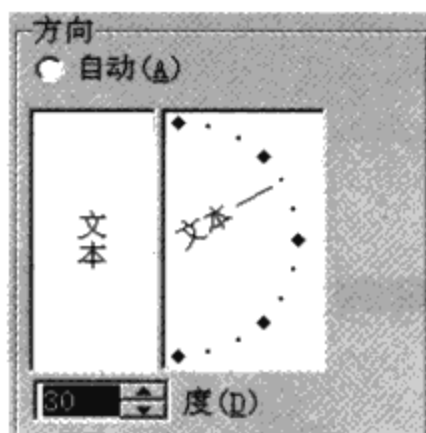


图 5.2.19 对齐选项的设置

其他可以修饰的图表元素还有“数据标志格式”、“网格线格式”、“图表区格式”和“绘图区格式”等等，操作方法与上面介绍的几种大同小异，读者可模仿介绍的内容自己进行摸索，通过实践不难掌握，此处不再一一介绍。

### 5.2.3 绘图实例——用 Excel 绘制任意一元函数的图像

用 Matlab 和 Mathematica 不难画出任意一元函数的图像，但需要编写一小段程序，或者至少要一条语句，用 Excel 也可以绘制任意一元函数的图像，其优点



中的一个到合适的位置放开, 图形区即被放大. 然后做以下工作:

### 1) 修改标题

按照前面介绍的方法, 移到标题的位置, 更改标题文字内容, 启动“图表标题格式”对话框, 设置标题的字体、字形、字号和颜色(前景色和背景色), 直至满意为止.

### 2) 修改坐标轴

按前面的方法调出“坐标轴格式”对话框. 默认  $x$  轴的两边空白区比较多, 本例  $x$  的取值范围是  $-4 \sim 8$ , 但画出来的图中  $x$  的范围是  $-6 \sim 10$ , 两边各增加了 2 个单位的空白区, 似乎空白太多, 可以设置  $x$  轴最小值  $-5$ , 最大值 9, 留一个单位的空白就够了. 默认坐标刻度的数字的字号比较大, 可以设置为 8~10 即可, 选择一种美观的字体. 对刻度间隔(对话框中的“主要刻度单位”)作适当设定, 对坐标轴的线宽和颜色也可进行设置, 点击“确定”生效.

### 3) 修改曲线的颜色

默认曲线的颜色和线宽不一定符合自己的要求, 可根据自己的喜欢作修饰, 右击图中曲线, 调出“数据系列格式”对话框, 选择其中“图案”选项, 其中一个栏目是“线形”, 见图 5.2.21, 选项“自定义”, 点击“颜色”右边的▼, 选择自己想要的颜色. 对“粗细”栏目, 选择中等粗细(比默认线宽粗一些), 点击“确定”.

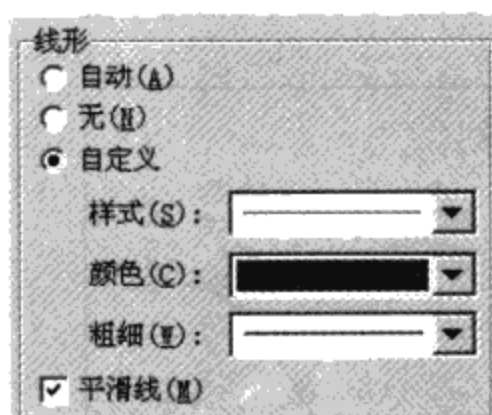


图 5.2.21 线形的设置

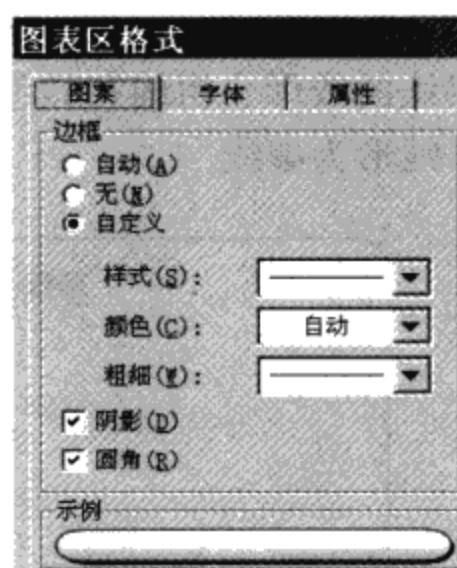


图 5.2.22 设置图表区格式

### 4) 修改图表区格式

调出图表区格式对话框, 选择“图案”选项, 选中“阴影”和“圆角”, 见图 5.2.22. 点击“填充效果”按钮, 弹出“填充效果”对话框(图 5.2.12), 选择双色, 白—青绿从上向下, 水平过渡, 确定. 然后调出“绘图区格式”对话框, 其中“边框”和“区域”都选“无”, 然后点击“确定”.

### 5) 调整坐标轴标记

观察图像,发现坐标轴的标识  $x$  和  $y$  的位置、字体、大小,方向需要调整,字符  $y$  是横躺着的,需要转过来. 右击坐标轴标记字符,调出“坐标轴标题格式”对话框,如图 5.2.23 所示. 设置适当的字体、字号,选择“对齐”选项,显示默认对齐方向是  $90^\circ$ ,用鼠标把文本方向转到  $0^\circ$ ,确定.

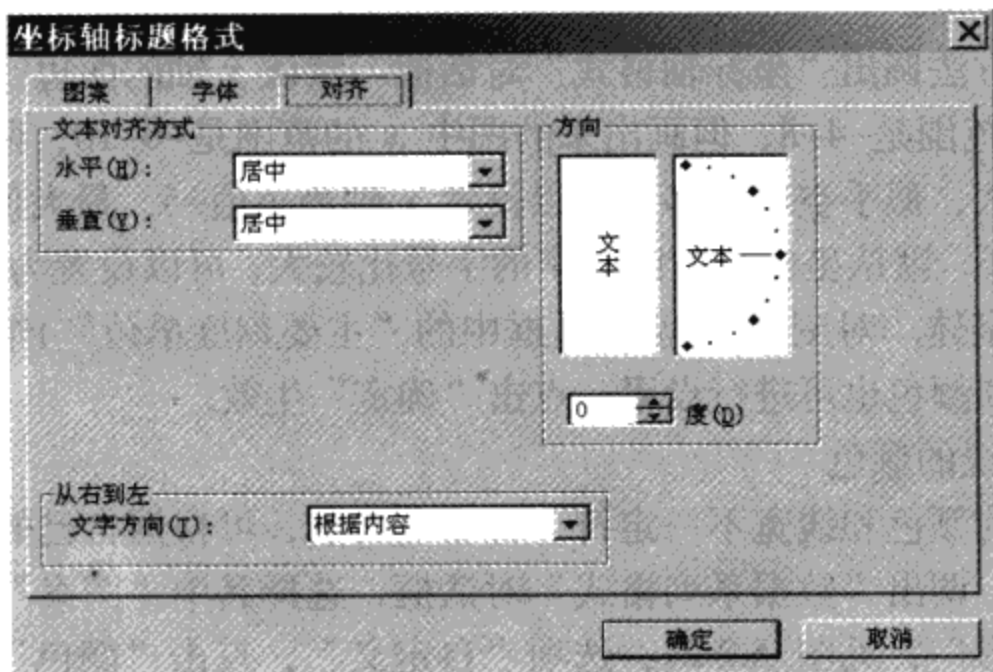


图 5.2.23 设置坐标轴标题

修饰完成以后的图像见图 5.2.24. 图上没有画网格线,如果想加上网格线,可以通过“图表选项”→“网格线”加上网格线,再调出“网格线格式”对话框,设置网格线的线形、粗细和颜色,通常可把网格线的线宽设置细一些,颜色设置浅一些,线形为虚线.

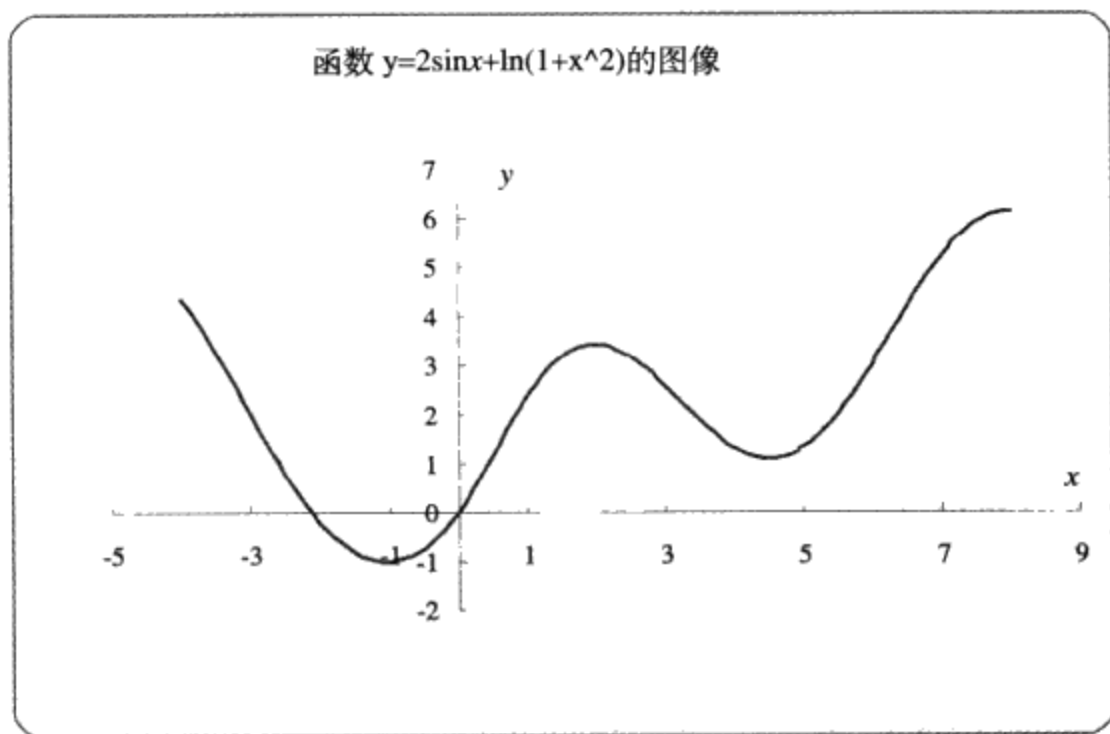


图 5.2.24 用 Excel 画出函数图像

以上介绍的画函数图像的方法虽然是以某个指定函数为特例进行的,但是只要更改 B2 单元格内的自定义函数,就可以生成任意函数的图像,其方法和步骤具有通用性.

## 5.3 总体分布的假设检验

假设有一批数据是某个随机变量的观测结果,怎样根据这批数据来确定该随机变量所服从的分布呢?是正态分布还是其他分布?这就是总体分布的假设检验问题,常用方法是  $\chi^2$  检验法.

### 5.3.1 $\chi^2$ 检验法的基本思路

原假设  $H_0$ : 总体  $X$  的分布函数为  $F(x)$ .

若  $X$  为离散型,则  $H_0$  等价于总体  $X$  的分布律为  $P\{X=t_i\}=p_i$ ; 若  $X$  为连续型,则  $H_0$  等价于总体  $X$  的概率密度为  $f(x)$ . 如果  $F(x)$  中有未知参数,则先用极大似然法作出估计.

#### 1. 分区间(组)

$\chi^2$  检验法通常要求数据的样本容量  $n>50$ , 将实际数据的取值范围分成  $k$  个区间( $k$  按  $n$  的大小来定), 假设各区间表示为  $(t_{i-1}, t_i]$ ,  $i=1, 2, \dots, k$ , 如果原假设成立, 则  $X$  落入区间  $(t_{i-1}, t_i]$  内的概率(理论计算值)可以用下式计算:

$$\hat{p}_i = P\{t_{i-1} < X \leq t_i\} = F(t_i) - F(t_{i-1}), \quad (5.3.1)$$

对离散型  $X$ , 可通过分布律求得该理论值, 对连续型  $X$ , 则  $\hat{p}_i = \int_{t_{i-1}}^{t_i} f(x) dx$ . 理论上  $n$  个观测值中落入区间  $(t_{i-1}, t_i]$  内的个数为  $n\hat{p}_i$ , 用  $m_i$  表示落入区间  $(t_{i-1}, t_i]$  内的样本点(数据)个数.

#### 2. $\chi^2$ 检验法的思路

如果原假设成立, 则  $m_i$  与  $n\hat{p}_i$  比较接近, 因此  $(m_i - n\hat{p}_i)^2$  应当比较小, 令统计量  $\chi^2 = \sum_{i=1}^k \frac{(m_i - n\hat{p}_i)^2}{n\hat{p}_i}$ , 则在原假设成立的情况下, 该统计量的值比较小, 我们可以根据它的大小来决定是接受还是拒绝原假设.

**定理** 若  $n$  充分大, 则当原假设  $H_0$  成立时, 统计量  $\chi^2$  近似服从  $\chi^2(k-r-1)$  分布, 其中  $k$  是小区间(分组)个数,  $r$  是被估参数个数.

3. 作出判别

由具体数据计算出统计量  $\chi^2 = \sum_{i=1}^k \frac{(m_i - n\hat{p}_i)^2}{n\hat{p}_i}$  的值, 如果此值过大就否定原假设, 具体量化: 对给定的  $\alpha$ , 查表得到  $\chi^2_{\alpha}(k-r-1)$ , 若  $\chi^2 > \chi^2_{\alpha}(k-r-1)$ , 则拒绝  $H_0$ , 否则接受  $H_0$ .

5.3.2 方法步骤

下面通过实例介绍  $\chi^2$  检验法的步骤.

例 5.3.1 表 5.3.1 的数据来自于 1999 年数学建模竞赛 A 题——自动化车床管理, 问这批数据服从什么样的分布?

表 5.3.1 100 次刀具故障记录(完成的零件数)


459	362	624	542	509	584	433	748	815	505	612	452	434	982	640
742	565	706	593	680	926	653	164	487	734	608	428	1153	593	844
527	552	513	781	474	388	824	538	862	659	775	859	755	649	697
515	628	954	771	609	402	960	885	610	292	837	473	677	358	638
699	634	555	570	84	416	606	1062	484	120	447	654	564	339	280
246	687	539	790	581	621	724	531	512	577	496	468	499	544	645
764	558	378	765	666	763	217	715	310	851					

2	平均	600
3	标准误差	19.66292
5	中位数	599.5
6	众数	593
7	标准差	196.6292
8	方差	38663.03
9	峰度	0.441398
10	偏度	-0.01117
11	区域	1069
12	最小值	84
13	最大值	1153
14	求和	60000
15	观测数	100
16	最大(1)	1153
17	最小(1)	84
18	置信度(95)	39.01549

图 5.3.1 描述统计结果

解 用  $X$  表示发生刀具故障时已生产的零件数, 则  $X$  是随机变量, 以上 100 次刀具故障记录数据是来之于总体  $X$  的样本, 现在要检验总体  $X$  服从的分布.

1. 数据分析

在 Excel 的 A 列从 A2 开始输入这 100 个数据, B 列 B2—B101 是 A 列数据的复制, 选中 B2-B101 点击工具栏上  图标, 对 B 列数据按由小到大升序排列, 先对数据作描述分析, 从主菜单点击“工具”—“数据分析”—“描述统计”—“确定”, 弹出描述统计对话框(图 5.1.10), 在“输入区域”填入 B1:B101, 在“标志位于第一行”上打上“√”, 输出位置选“新

工作表组”。点击“确定”，得到描述统计的结果如图 5.3.1 所示。由此可知，数据的最小值为 84，最大值为 1153，跨度 1069，平均值 600，标准差为 196.62917。

2. 数据分组

1) 确定分组数  $k$

考虑数据的跨度  $1069 \approx 1100$ ，如果分成 11 组，恰好组距为 100，所以取分组数  $k=11$ ，由此得 11 个区间为： $<150$ ， $(150,250]$ ， $(250,350]$ ， $(350,450]$ ， $(450,550]$ ， $(550,650]$ ， $(650,750]$ ， $(750,850]$ ， $(850,950]$ ， $(950,1050]$ ， $>1050$ 。

2) 统计各组频数，画直方图

统计各区间内的数据个数  $m_i$  为：2,3,4,10,20,24,15,12,5,3,2。各区间的频率  $f_i = m_i / n$  为：0.02,0.03,0.04,0.1,0.2,0.24,0.15,0.12,0.05,0.03,0.02。画出直方图，如图 5.3.2 所示。

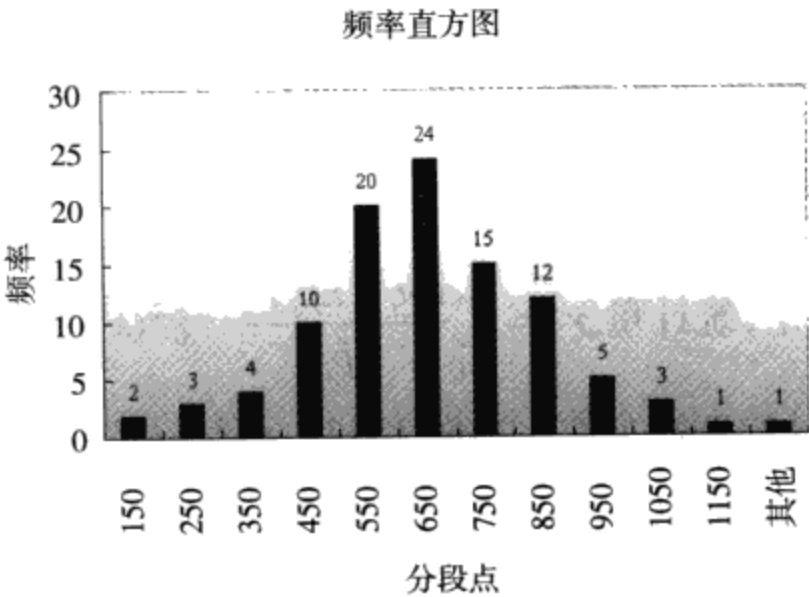


图 5.3.2 频率直方图

从直方图可以看出刀具故障记录数据比较接近正态分布。所以我们先假设  $X$  服从正态分布，其数学期望的无偏估计为  $\mu = \bar{x} = 600$ 。对总体方差的估计有两种

公式，一种是矩法估计  $\hat{\sigma}^2 = \sum_{i=1}^n (X_i - \bar{X})^2 / n$ ，另外一种为样本方差  $\hat{\sigma}^2 = S^2 =$

$\sum_{i=1}^n (X_i - \bar{X})^2 / (n - 1)$ ，标准差(均方差)是方差估计的平方根  $\hat{\sigma} = \sqrt{\hat{\sigma}^2}$ 。这两种估计

可以在 Excel 内调用函数 STDEVP 或 STDEV 来计算，在弹出对话框内输入数据所在的区域 B1:B101，结果分别为 195.64355 和 196.62917。采用其中任意一个都可以。



3. 计算各区间理论频率

假如原假设  $X \sim N(\mu, \sigma^2)$  成立, 则  $X$  落入区间  $(t_{i-1}, t_i]$  内的理论概率为  $\hat{p}_i = P\{t_{i-1} < X \leq t_i\} = F(t_i) - F(t_{i-1})$ , 这可以用 Excel 的函数 NORMDIST 函数来计算, 该函数相当于查正态分布表, 它需要 4 个参数:  $x, \mu, \sigma, 1$ , 其中  $x$  是分布函数  $F(x)$  的自变量,  $\mu$  是数学期望, 此处用估计值  $\mu=600$ ,  $\sigma$  是标准差, 采用矩法估计量  $\sigma=195.6436$ . 求得各区间上的理论概率  $\hat{p}$  之后, 再计算各区间理论频数  $n\hat{p}_i$ .

4. 计算统计量 
$$\chi^2 = \sum_{i=1}^k \frac{(m_i - n\hat{p}_i)^2}{n\hat{p}_i}$$

其中  $m_i$  是  $n$  个样本中落入区间  $(t_{i-1}, t_i]$  内的样本个数. 计算方法是利用自定义带参数计算公式, 公式中的参数相当于自变量. 点击空白列的数据区第二行, 先输入一个等号=(任何公式都以等号开始), 然后输入自定义公式. 假设  $m_i$  数据在 D 列,  $n\hat{p}_i$  数据在 I 列, 则输入=(D2-I2)^2/I2, 该单元格显示计算结果为 0.803032, 然后把该公式向下拖动直到第 12 行, 每一行的结果就自动计算出来了, 用 SUM(J2:J12)可以求出  $\chi^2 = 4.71647$ .

5. 根据  $\chi^2$  值作出判断

按照  $\chi^2$  检验理论, 统计量  $\chi^2 = \sum_{i=1}^k \frac{(m_i - n\hat{p}_i)^2}{n\hat{p}_i} \sim \chi^2(k-r-1)$ , 式中  $k$  是分组(区间)个数,  $r$  是被估计的参数个数, 本例  $k=11, r=2$ , 故  $\chi^2 \sim \chi^2(8)$ . 查  $\chi^2$  分布表得临界值为  $\chi^2_{0.05}(8) = 15.5073$ , 该值也可以通过 Excel 函数 CHINV 得到, 参数  $\alpha=0.05$ , 自由度为 8, CHINV(0.05,8)=15.5073. 由于  $\chi^2 = 4.71647$ , 小于临界值, 故接受原假设: 总体  $X$  服从正态分布. 图 5.3.3 是在 Excel 中进行总体分布假设检验的统计和计算表.

	A	B	C	D	E	F	G	H	I	J
	刀具故障	排序	分组	频数 $m_i$	频率	分段点	F(x)	概率 $p_i$	理论频数 $np_i$	$\chi^2$
1										
2	459	84	<150	2	0.02	150	0.01072	0.01072	1.072126	0.803032
3	362	120	(150,250]	3	0.03	250	0.03681	0.02609	2.608879	0.058637
4	624	164	(250,350]	4	0.04	350	0.10065	0.06384	6.384398	0.890507
5	542	217	(350,450]	10	0.1	450	0.22163	0.12098	12.09759	0.363698
6	409	246	(450,550]	20	0.2	550	0.39914	0.17751	17.75128	0.284866
7	584	280	(550,650]	24	0.24	650	0.60086	0.20171	20.17146	0.726656
8	433	292	(650,750]	15	0.15	750	0.77837	0.17751	17.75128	0.426423
9	748	310	(750,850]	12	0.12	850	0.89935	0.12098	12.09759	0.000787
10	815	339	(850,950]	5	0.05	950	0.96319	0.06384	6.384398	0.300194
11	405	358	(950,1050]	3	0.03	1050	0.98928	0.02609	2.608879	0.058637
12	612	362	>1050	2	0.02	1150	0.99753	0.01072	1.072126	0.803032
13	452	378							卡方值	4.716468
14	434	388							临界值	15.50731

图 5.3.3 用 Excel 作总体分布假设检验



## 6. 计算 $\chi^2$ 值的另一种方法

在 Excel 中还有另一种计算统计量 $\chi^2$ 值的方法: 用函数 CHITEST, 它需要两组参数, 一组是各小区间内样本点的个数, 即图 5.3.3 所示 Excel 表格中 D 列的数据: 2, 3, 4, 10, 20, 24, 15, 12, 5, 3, 2, 另一组是理论频数, 即表中第 I 列的数据. 用 CHITEST 函数, 它需要两个参数: 在 Actual\_range 栏目内输入 D2:D12, 在 Expected\_range 栏目内输入 I2:I12, 得到的结果是 0.90929593(图 5.3.4). Excel 规定的自由度为  $k-1$ , 这里  $k-1=10$ . 以上结果表示的含义是:  $P\{\chi^2(10) > \chi^2\} = 0.90929593$ , 由上式可以求出统计量 $\chi^2$ 的值, 方法是用 CHINV 函数, 它需要两个参数, 在 Probability 栏目内输入刚才的结果 0.90929593(鼠标点一下刚才的结果即可), 在 Deg\_freedom(自由度)栏目内输入 10, 得到结果 4.716468, 这就是统计量 $\chi^2$ 的值, 与前面求出的结果相同.

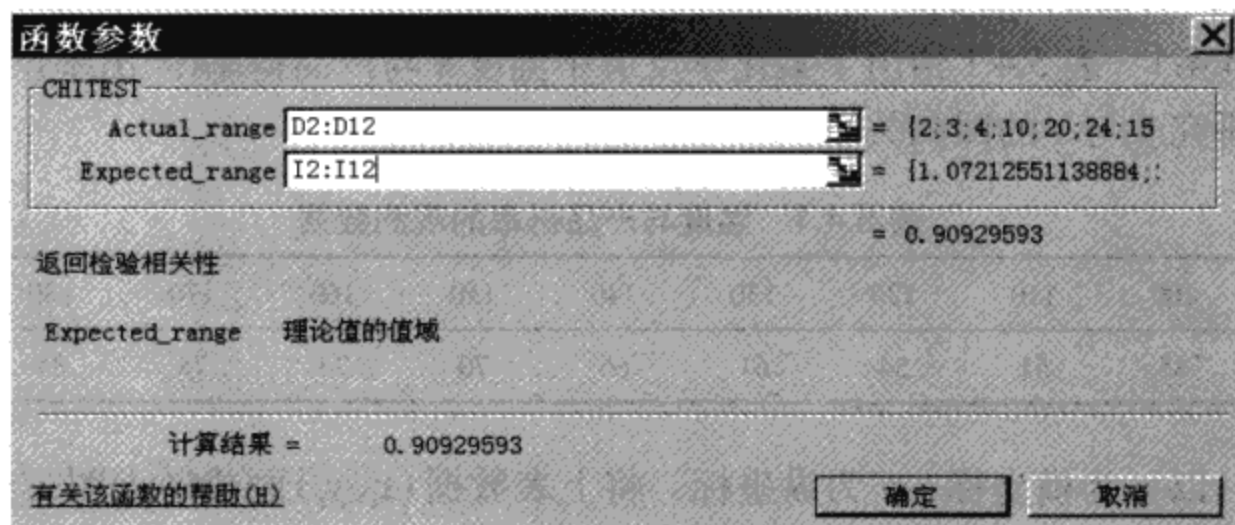


图 5.3.4 调用 Excel 中 CHITEST 函数

**说明** 本例分组(区间)的结果前两个组和最后两个组的数据个数比较少, 可以将前两个区间合并成一个区间 $(-\infty, 250)$ , 将最后两个区间也合并成一个区间 $(950, +\infty)$ . 此时  $k=9$ , 检验结果不变.

## 5.4 回归分析

### 5.4.1 回归分析的概念

#### 1. 回归分析研究的问题

回归分析研究因变量与自变量的相关关系, 因变量是随机变量, 自变量是可以控制或测量的变量(非随机变量, 也称因素变量), 例如, 成人的血压与年龄的关系、商品销售量与价格的关系、农作物的产量与降雨量以及施肥量的关系, 等等.

回归分析通常解决以下问题, 第一, 确定因变量与一个或多个自变量之间的

近似表达式, 称为回归方程或经验公式; 第二, 用求得的回归方程对因变量进行预测或控制; 第三, 进行因素分析, 区别重要因素和次要因素.

## 2. 回归的分类

按自变量的数量来分, 可分成:

- (1) 一元回归: 随机变量  $Y$  与单个自变量  $x$  的相关关系;
- (2) 多元回归: 随机变量  $Y$  与几个自变量  $x_i$  之间的关系.

按回归方程的形式来分, 可分成:

- (1) 线性回归: 回归方程的形式是线性表达式;
- (2) 非线性回归: 回归方程的形式是非线性表达式.

### 5.4.2 一元线性回归

设观测数据为  $(x_i, y_i)$ ,  $i = 1, 2, \dots, n$ . 回归方程的形式为  $y = a + bx$ .

**例 5.4.1** 表 5.4.1 给出了某化学反应中温度  $x$  与产品得率(产出率) $y$  的观测数据, 试研究  $y$  与  $x$  的回归关系.

表 5.4.1 温度与产品得率的观测数据

温度 $x$	100	110	120	130	140	150	160	170	180	190
得率 $y$	45	51	54	61	66	70	74	78	85	89

**解** 以  $x$  为横坐标,  $y$  为纵坐标, 将上表数据  $(x_i, y_i)$  画成散点图, 如图 5.4.1 所示, 从图上可以看出,  $y$  与  $x$  近似存在线性关系.

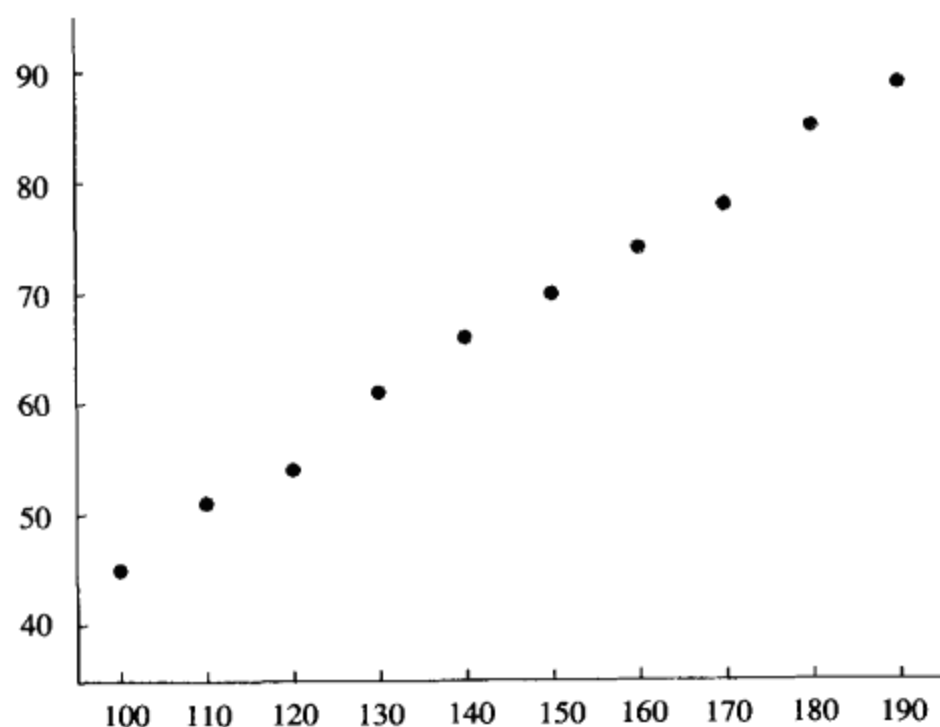


图 5.4.1 温度与得率的散点图

Excel 提供了一组称作“数据分析”的统计分析工具包，内含方差分析、回归分析、协方差和相关系数、傅立叶分析等分析工具，使用这组分析工具，可以大大提高工作效率和质量。

在默认安装下，Excel 并不直接提供数据分析工具包，如果是第一次使用这类工具，则“工具”菜单中没有“数据分析”子菜单，需要按照 5.1 节介绍的方法进行安装。安装完成之后，点击“工具”菜单中的“数据分析”子菜单，弹出对话框，显示各种数据分析工具。

在进行回归分析之前先输入数据，如图 5.4.2 所示。然后点击“工具”→“数据分析”→“回归”→“确定”，弹出“回归分析”对话框，如图 5.4.3 所示。

	A	B
1	温度	得率
2	100	45
3	110	51
4	120	54
5	130	61
6	140	66
7	150	70
8	160	74
9	170	78
10	180	85
11	190	89

图 5.4.2  输入数据

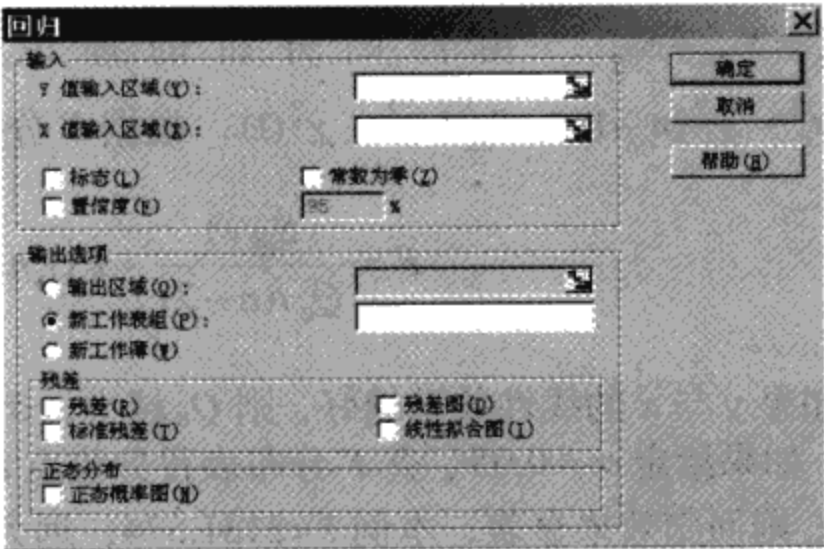


图 5.4.3  回归分析对话框

在 Y 值输入区域填入 B1:B11，表示 B 列第 1 行至第 11 行是因变量 y 的数据，在 X 值输入区域填入 A1:A11，表示 A 列第 1 行至第 11 行是自变量 x 的数据，因第一行是表头(标志)，故在对话框的标志上打上“√”，见图 5.4.3，置信度打上“√”，用默认的 95%或根据需要改成其他百分比。输出选项可选“新工作表组”，残差项目可选，也可不选，正态概率图一般不选。点击“确定”，立即得到回归分析的结果，如图 5.4.4 所示。

回归统计					
Multiple R	0.9981287				
R Square	0.9962609				
Adjusted R	0.9957936				
标准误差	0.9502791				
观测值	10				
方差分析					
	df	SS	MS	F	Significance F
回归分析	1	1924.876	1924.876	2131.574	5.35253E-11
残差	8	7.224242	0.90303		
总计	9	1932.1			
	Coefficients	标准误差	t Stat	P-value	Lower 95%
Intercept	-2.7393939	1.5465	-1.77135	0.11445	-6.305629201
温度	0.4830303	0.010462	46.16897	5.35E-11	0.458904362

图 5.4.4  回归分析的结果

结果中的项目比较多, 其中“回归统计”表中的主要项目解释如下:

(1) Multiple R: 相关系数  $r$ , 其值  $|r| \leq 1$ , 越接近 1 线性关系越显著;

(2) R Square: 相关系数  $r$  的平方, 越接近 1 线性关系越显著;

(3) 标准误差: 均方差的估计值  $\hat{\sigma}$ .

方差分析表的主要项目见表 5.4.2, 其中  $df$  是统计量所服从的  $\chi^2$  分布的自由

度,  $Q_e$  是残差的平方和:  $Q_e = \sum_{i=1}^n (y_i - \hat{a} - \hat{b}x_i)^2$ , 式中  $\hat{a}$  和  $\hat{b}$  是回归系

数.  $\hat{\sigma}^2 = Q_e / (n-2)$  是总体  $X$  的方差  $\sigma^2$  的无偏估计. 若令统计量  $S_{xx} = \sum (x_i - \bar{x})^2$ ,

$S_{yy} = \sum (y_i - \bar{y})^2$ ,  $S_{xy} = \sum (x_i - \bar{x})(y_i - \bar{y})$ , 则  $Q_e = S_{yy} - S_{xx}^2 / S_{xy}$ , 令  $S_{\square} = S_{xx}^2 / S_{xy}$ ,

则  $Q_e = S_{yy} - S_{\square}$ . 由回归分析的理论可知,  $Q_e / \sigma^2 \sim \chi^2(n-2)$ ,

$S_{yy} / \sigma^2 \sim \chi^2(n-1)$ ,  $S_{\square} / \sigma^2 \sim \chi^2(1)$ . 根据  $F$  分布的定义, 统计量

$$F = \frac{S_{\square} / 1}{Q_e / (n-2)} \sim F(1, n-2). \quad (5.4.1)$$

如果  $y$  与  $x$  的线性关系越好, 则  $Q_e$  越小,  $\hat{\sigma}^2$  也越小,  $F$  值越大, 回归效果越显著. 如果给定  $\alpha = 0.05$ , 查  $F$  分布表得分位点  $F_{\alpha}(1, n-2) = F_{0.05}(1, 8) = 5.318$ , 若  $F > F_{\alpha}$  则回归效果显著, 本例  $F=2131.574$ , 远大于分位点  $F_{\alpha}$ , 故回归效果很好.

表 5.4.2 方差分析表的重要项目

	df(自由度)	SS	MS	F 值
回归分析	1	1924.876( $S_{\square}$ )	1924.876( $S_{\square}/1$ )	2131.574
残 差	8	7.22424( $Q_e$ )	0.90303( $Q_e/8$ )	
总 计	9	1932.1( $S_{yy}$ )		

图 5.4.4 中由 Excel 得到的回归系数为  $\hat{a} = -2.7393939$ ,  $\hat{b} = 0.48030303$ , 回归方程为  $y = -2.7393939 + 0.48030303x$ . 回归结果中  $t$  Stat 和温度所对应的栏目内的

数值 46.16879 是统计量  $t$  的值,  $t = \frac{S_{xy}}{\hat{\sigma} \sqrt{S_{xx}}} \sim t(n-2)$ . 如果给定  $\alpha$ , 查  $t$  分布表得

分位点  $t_{\alpha/2}(n-2)$ , 若  $|t| > t_{\alpha/2}(n-2)$  则回归效果显著.

### 5.4.3 多元线性回归

在实际问题中, 大多数情况下随机变量  $y$  往往与多个变量  $x_1, x_2, \dots, x_k$  有关, 这是多元回归问题, 多元线性回归是最基本的类型.

## 1. 多元线性回归模型

假设  $y$  的数学期望  $E(y)$  是  $k$  个自变量:  $x_1, x_2, \dots, x_k$  的线性函数, 写成  $E(y) = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k$ , 于是多元回归模型为

$$y = b_0 + b_1x_1 + b_2x_2 + \dots + b_kx_k + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2). \quad (5.4.2)$$

其中  $b_0, b_1, b_2, \dots, b_k$  是回归系数,  $\sigma^2$  是待估参数.

用最小二乘法求出回归系数的估计值  $\hat{b}_0, \hat{b}_1, \dots, \hat{b}_k$ , 得到多元线性回归方程

$$\hat{y} = \hat{b}_0 + \hat{b}_1x_1 + \hat{b}_2x_2 + \dots + \hat{b}_kx_k, \quad (5.4.3)$$

由该式求出的  $\hat{y}$  称为  $y$  的预报值. 在回归系数求出来以后,  $Q_e$  的值为

$$Q_e = \sum_{i=1}^n (y_i - \hat{y}_i)^2, \text{ 理论上 } Q_e / \sigma^2 \sim \chi^2(n-k-1), \text{ 由此可得 } \hat{\sigma}^2 = \frac{Q_e}{n-k-1} \text{ 是 } \sigma^2 \text{ 的}$$

无偏估计.

## 2. 多元线性假设的显著性检验

与一元线性回归相类似, 多元线性回归的显著性检验等价于检验假设:

原假设  $H_0: b_1 = b_2 = \dots = b_k = 0$  (回归效果差);

对立假设  $H_1$ : 至少有一个  $b_j \neq 0$  (回归效果好).

先考察几个量之间的关系, 已知  $Q_e = \sum_{i=1}^n (y_i - \hat{y}_i)^2$ ,  $S_{yy} = \sum (y_i - \bar{y})^2$ , 令

$$S_{\text{reg}} = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2, \text{ 式中 } \hat{y}_i \text{ 由式(5.4.3)计算得到. 这三个量之间的关系为:}$$

$$Q_e = S_{yy} - S_{\text{reg}}.$$

理论上可以证明  $Q_e / \sigma^2 \sim \chi^2(n-k-1)$ ,  $S_{yy} / \sigma^2 \sim \chi^2(n-1)$ , 由  $\chi^2$  分布的可加性得到  $S_{\text{reg}} / \sigma^2 \sim \chi^2(k)$ . 根据  $F$  分布的定义, 统计量

$$F = \frac{S_{\text{reg}} / k}{Q_e / (n-k-1)} \sim F(k, n-k-1). \quad (5.4.4)$$

若  $F$  值越大, 则回归效果越好. 如果给定  $\alpha$ , 查  $F$  分布表得分位点  $F_\alpha(k, n-k-1)$ , 若  $F > F_\alpha(k, n-k-1)$ , 则拒绝  $H_0$ , 回归效果好; 否则接受原假设  $H_0$ , 回归效果差.

由于  $Q_e = S_{yy} - S_{\text{reg}}$ ,  $S_{yy}$  是定值,  $S_{\text{reg}}$  越大, 则  $Q_e$  越小, 从而  $F$  值越大, 线性

关系越显著，反之亦然.

3. 用 Excel 进行多元线性回归分析

下面通过实例说明用 Excel 进行多元线性回归分析的方法和步骤.

**例 5.4.2** 某种水泥在凝固时放出的热量与水泥中的下列四种化学成分的含量(%)有关  $x_1$ :3Cao. $\text{Al}_2\text{O}_3$ ,  $x_2$ :3Cao. $\text{SiO}_2$ ,  $x_3$ :4Cao.  $\text{Al}_2\text{O}_3$ . $\text{Fe}_2\text{O}_3$ ,  $x_4$ :2Cao. $\text{SiO}_2$ . 数据见表 5.4.3.

表 5.4.3 测试数据

编号	1	2	3	4	5	6	7	8	9	10	11	12	13
$x_1$	7	1	11	11	7	11	3	1	2	21	1	11	10
$x_2$	26	29	56	31	52	55	71	31	54	47	40	66	68
$x_3$	6	15	8	8	6	9	17	22	18	4	23	9	8
$x_4$	60	52	20	47	33	22	6	44	22	26	34	12	12
$y$	78.5	74.3	104.3	87.6	95.9	109.2	102.7	72.5	93.1	115.9	83.8	113.3	109.4

	A	B	C	D	E
1	x1	x2	x3	x4	y
2	7	26	6	60	78.5
3	1	29	15	52	74.3
4	11	56	8	20	104
5	11	31	8	47	87.6
6	7	52	6	33	95.9
7	11	55	9	22	109
8	3	71	17	6	103
9	1	31	22	44	72.5
10	2	54	18	22	93.1
11	21	47	4	26	116
12	1	40	23	34	83.8
13	11	66	9	12	113
14	10	68	8	12	109

图 5.4.5 输入数据

**解** 首先在 Excel 中输入数据,如图 5.4.5 所示,  $x_1$ ,  $x_2$ ,  $x_3$ ,  $x_4$  是自变量,  $y$  是因变量. 回归方程形式为  $y = b_0 + b_1x_1 + b_2x_2 + b_3x_3 + b_4x_4$ .

然后点击“工具”→“数据分析”→“回归”→“确定”,弹出回归对话框,在该对话框的“Y 值输入区域”填入 E1:E14,“X 值输入区域”填入 A1:D14,然后点击“确定”,立即得到回归结果,见图 5.4.6.

对结果的说明:

Multiple: 相关系数  $R$ , 越接近 1 越好;

R Square:  $R$  的平方;

标准误差: 即  $\sigma$  的估计值.

方差分析表中的重要数据:

df 是自由度; SS 是  $S_{\text{总}}$  (2667.899); 残差是  $Q_e$ (47.86364);  $F$  值 111.4792 是判别线性假设是否成立的依据,  $F$  值越大越好. 本例临界值为  $\text{FINV}(0.05,4,8) = 3.8379$ , 由于  $F$  值大于临界值, 所以认为线性回归效果好.

Coefficient 所在的一列表示回归系数, 其中  $b_0 = 62.40537$ ,  $b_1 = 1.551103$ ,  $b_2=0.510168$ ,  $b_3 = 0.101909$ ,  $b_4=-0.14406$ .

	A	B	C	D	E	F	G	H	I
4	Multiple F 0.991149								
5	R Square 0.982376								
6	Adjusted 0.973563								
7	标准误差 2.446008								
8	观测值 13								
9									
10	方差分析								
11		df	SS	MS	F	Significance F	3.8378534		
12	回归分析	4	2667.899	666.9749	111.4792	4.75618E-07			
13	残差	8	47.86364	5.982955					
14	总计	12	2715.763						
15									
16		Coefficient	标准误差	t Stat	P-value	Lower 95%	Upper 95%	下限 95.0%	上限 95.0%
17	Intercept	62.40537	70.07096	0.890602	0.399134	-99.17855226	223.98929	-99.17855	223.989291
18	x1	1.551103	0.74477	2.08266	0.070822	-0.166339744	3.268545	-0.16634	3.26854504
19	x2	0.510168	0.723788	0.704858	0.500901	-1.158890544	2.1792257	-1.158891	2.1792257
20	x3	0.101909	0.754709	0.135031	0.895923	-1.638452774	1.8422716	-1.638453	1.84227158
21	x4	-0.14406	0.709052	-0.20317	0.844071	-1.779138018	1.491016	-1.779138	1.49101596

图 5.4.6 多元线性回归分析的结果

4. 因素主次的判别

在变量  $x_1, x_2, \cdots, x_k$  中，各变量是否都同等重要呢？哪些变量是重要的(影响大)？哪些变量不太重要(影响小)？能否把影响较小的次要变量剔除？剔除之后有什么影响？某个变量  $j$  对应的回归系数  $\hat{b}_j$  越小，则该变量对  $y$  的影响越小，以例 5.4.2 为例， $x_3$  对应的回归系数  $\hat{b}_j = 0.1019$  是绝对值最小者，剔除  $x_3$  试一试，此时还有 3 个自变量，改动数据，其他不变，重新用 Excel 进行回归分析，得到回归方程为：

$$\hat{y} = 71.6483 + 1.45194x_1 + 0.41611x_2 - 0.23654x_4.$$

此时的  $S_{[u]}$  为 2667.79，比原来的  $S_{[u]}=2667.90$  稍微小一点点，其差值记为  $u_j$ ，此值越小，说明变量  $x_j$  的作用越不明显.  $u_3=0.11$ ，说明变量  $x_3$  的作用较小.

下面介绍具体量化方法， $u_j$  小到什么程度可以剔除变量  $x_j$ .

原假设  $H_0: b_j=0$ ，变量  $x_j$  的作用不显著(可以剔除)；

对立假设  $H_1: b_j\neq 0$ ，变量  $x_j$  的作用显著.

理论上  $u_j/\sigma^2 \sim \chi^2(1)$ ，根据  $F$  分布的定义，统计量

$$F_j = \frac{u_j}{Q_e(n-k-1)} \sim F(1,n-k-1). \tag{5.4.5}$$

给定  $\alpha$ ，查  $F$  分布表得分位点  $F_\alpha(1,n-k-1)$ ，若  $F_j > F_\alpha(1,n-k-1)$ ，则拒绝  $H_0$ ，变量  $x_j$  的作用显著；否则接受原假设  $H_0$ ，变量  $x_j$  的作用不显著(可以剔除)， $F_j$  越小，变量  $x_j$  的作用越不显著.

本例  $F_3=0.0184$ , 临界值为  $F_{0.05}(1,8)=5.32$ ,  $F_3 < F_{0.05}$ , 接受  $H_0$ , 即变量  $x_3$  的作用不显著, 可以剔除.

5.4.4 可化为线性的非线性回归

在实际问题中, 线性关系仅是一种最简单、最基本的情况, 更多的是非线性关系. 解决非线性回归的做法可以有两种:

- (1) 通过适当的变换, 化为线性问题;
- (2) 直接用最小二乘法.

常用方法是通过变量代换把非线性关系化成线性关系, 然后用线性回归方法求出回归系数, 再返回原来的函数关系, 得到符合要求的回归方程. 表 5.4.4 列出常见的可化为线性方程的非线性方程.

表 5.4.4 常见的可化为线性方程的非线性方程

非线性方程		变换公式	变换后的线性方程
双曲线 $1/y = a + b/x$		$y^* = 1/y, x^* = 1/x$	$y^* = a + bx^*$
幂函数 $y = cx^b, c > 0, x > 0$		取对数得 $\ln y = \ln c + b \ln x$ 令 $y^* = \ln y, x^* = \ln x, a = \ln c$	$y^* = a + bx^*$
指数函数	$y = ce^{bx}, c > 0$	取对数得 $\ln y = \ln c + bx$ 令 $y^* = \ln y, a = \ln c$	$y^* = a + bx$
	$y = ce^{b/x}, c > 0$	取对数得 $\ln y = \ln c + b/x$ 令 $y^* = \ln y, a = \ln c, x^* = 1/x$	$y^* = a + bx^*$
对数函数 $y = a + b \ln x$		令 $x^* = \ln x$	$y = a + bx^*$
S 型曲线 $y = \frac{1}{a + be^{-x}}$		$y^* = 1/y, x^* = e^{-x}$	$y^* = a + bx^*$
抛物线 $y = b_0 + b_1x + b_2x^2$		$x_1 = x, x_2 = x^2$	$y = b_0 + b_1x_1 + b_2x_2$
多项式 $y = b_0 + b_1x + b_2x^2 + \cdots + b_kx^k$		$x_1 = x, x_2 = x^2, \cdots, x_k = x^k$	$y = b_0 + b_1x_1 + \cdots + b_kx_k$

例 5.4.3 混凝土的抗压强度  $x$  较容易测定, 而抗剪强度  $y$  不易测定, 工程中希望建立一种能由  $x$  推算  $y$  的经验公式. 表 5.4.5 列出了现有 9 对数据.

表 5.4.5 混凝土的抗压强度和抗剪强度数据

x	141	152	168	182	195	204	223	254	277
y	23.1	24.2	27.2	27.8	28.7	31.4	32.5	34.8	36.2

试分别按以下三种形式建立  $y$  对  $x$  的回归方程, 并根据  $F$  值选最优模型.



$$(1) y = a + b\sqrt{x};$$

$$(2) y = a + b\ln x;$$

$$(3) y = cx^b.$$

**解** 对于(1), 令  $x^* = \sqrt{x}$ ; 对于(2), 令  $x^* = \ln x$ ; 对于(3)两边取对数得  $\ln y = \ln c + b\ln x$ , 令  $y^* = \ln y, x^* = \ln x, a = \ln c$ , 三种情况下都有  $y^* = a + bx^*$ .

在 Excel 中输入  $x, y$  原始数据并计算  $\sqrt{x}, \ln x, \ln y$  等数据(图 5.4.7), 调用回归分析工具, 分别得到三种形式下的回归方程.

	A	B	C	D	E
1	x	y	$\sqrt{x}$	$\ln x$	$\ln y$
2	141	23.1	11.87434	4.94876	3.13983
3	152	24.2	12.32883	5.02388	3.18635
4	168	27.2	12.96148	5.12396	3.30322
5	182	27.8	13.49074	5.20401	3.32504
6	195	28.7	13.96424	5.27300	3.35690
7	204	31.4	14.28286	5.31812	3.44681
8	223	32.5	14.93318	5.40717	3.48124
9	254	34.8	15.93738	5.53733	3.54962
10	277	36.2	16.64332	5.62402	3.58906

图 5.4.7 混凝土强度数据

$$(1) \text{ 方程形式 } y = a + b\sqrt{x}.$$

结果:  $a = -9.88055, b = 2.8068, F = 335.61609$ , 相关系数  $R = 0.989732$ , 回归方程为:

$$y = -0.988055 + 2.8068\sqrt{x}.$$

$$(2) \text{ 方程形式 } y = a + b\ln x.$$

结果:  $a = -75.284446, b = 19.87895, F = 451.7927$ , 相关系数  $R = 0.992342$ , 回归方程为  $y = -75.284446 + 19.87895\ln x$ .

$$(3) \text{ 方程形式 } y = cx^b.$$

结果:  $a = -0.20053, b = 0.6781, c = e^a = 0.8183, F = 301.72$ , 相关系数  $R = 0.9886$ , 回归方程为  $y = 0.8183x^{0.6781}$ .

对以上三种形式经验公式的计算结果进行比较, 第二种经验公式的  $F$  值最大为 451.7927, 相关系数也最大, 所以第二种经验公式  $y = -75.284446 + 19.87895\ln x$  是最优模型, 画出散点图和三种经验公式(回归方程)的曲线图形, 见图 5.4.8, 图中以第二种经验公式拟合程度最好.

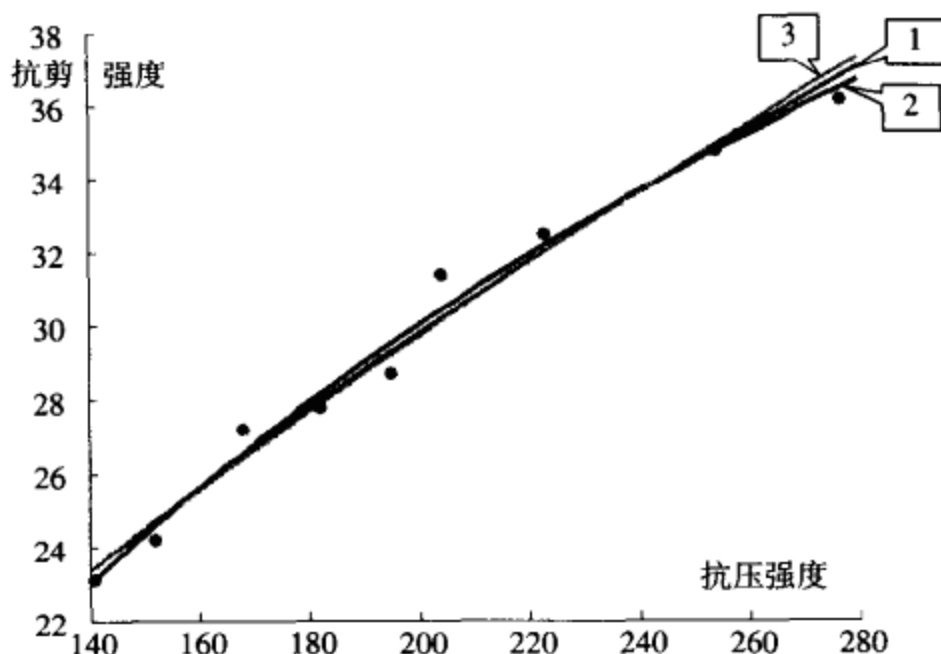


图 5.4.8 混凝土强度的散点图和三种回归曲线

### 习 题 五

1. 用迭代法能求任意正实数  $x$  的平方根, 迭代公式为  $a_n = \frac{1}{2} \left( a_{n-1} + \frac{x}{a_{n-1}} \right)$ , 令  $a_0=1.0$ , 则有  $\lim_{n \rightarrow \infty} a_n = \sqrt{x}$ , 用该方法在 Excel 内计算 5 的平方根, 要求计算结果的精度达到  $10^{-12}$ .
2. 试把方程  $x + 2^x - 4 = 0$  改写成收敛的迭代形式, 用 Excel 求该方程在区间(1, 2)内的根, 要求计算结果的精度达到  $10^{-12}$ .
3. 求方程  $3x - e^x = 0$  在区间(0, 1)内的实数根, 精度达到  $10^{-12}$ .
4. 利用公式  $e^x = 1 + x + \frac{1}{2!}x^2 + \frac{1}{3!}x^3 + \cdots + \frac{1}{n!}x^n + \cdots$ , 在 Excel 内计算  $e$  和  $e^2$  的值, 要求误差小于  $10^{-12}$ .
5. 拉马努金(Ramanujan)在 1916 年提出公式:

$$\frac{1}{\pi} = \frac{2\sqrt{2}}{9801} \sum_{n=1}^{\infty} \frac{(4n)!}{(n!)^4} \cdot \frac{1103 + 26390n}{396^{4n}},$$

试用该式在 Excel 内计算  $\pi$  的近似值, 要求计算结果的误差小于  $10^{-14}$ .

6. 测量了 50 颗滚珠的直径, 得数据如下:

15.8, 15.2, 15.1, 15.9, 14.7, 14.8, 15.5, 15.6, 15.3, 15.1, 15.3, 15, 15.6, 15.7, 14.8, 14.5,  
14.2, 14.9, 14.9, 15.2, 15, 15.3, 15.6, 15.1, 14.9, 14.2, 14.6, 15.8, 15.2, 15.9, 15.2, 15, 14.9,  
14.8, 14.5, 15.1, 15.5, 15.5, 15.1, 15.1, 15, 15.3, 14.7, 14.5, 15.5, 15, 14, 14.7, 14.2, 15.0.

试对以上数据作描述统计, 并画出直方图.

7. 设有如下表格数据:

项 目	应到位数	实际到位数	实际到位率
预 算 拨 款	100	80	
基 金 拨 款	100	70	
资本金拨款	50	20	
设 备 转 帐	50	35	
器 材 转 帐	100	70	
其 他 拨 款	200	150	

试求实际到位率, 并就实际到位数和实际到位率画出三维柱形图和三维饼图, 对图表进行修饰, 使之尽可能美观.

8. 用 Excel 画出函数  $y = \cos x - e^{-x^2/4} \ln(1+x^2)$  在区间  $[1, 6]$  上的图像, 在该区间内有无极小值? 极小值是多少?

9. 用 Excel 画出函数  $f(x) = 10x + e^x - 2$  在区间  $[-2, 4]$  上的图像, 并求解:

(1) 函数  $f(x)$  在该区间内的极大值是多少?

(2) 方程  $f(x) = 0$  在该区间内有几个根? 试把该方程改写成收敛的迭代形式, 取适当的初始值, 求出它的根.

10. 用 Excel 画出函数  $f(x) = e^{-x}(5\sin x + \ln(1+x^2))$  在区间  $[0, 9]$  内的图像, 有几个极小值? 几个极大值? 试用 Excel 求出这些极值点和极值, 并求出方程  $f(x) = 0$  在区间  $[2, 6]$  内的所有实数根.

11. 某罐头食品厂某天检查了某型号罐头 100 个, 得到净重数据(单位 g)如下:

342 341 348 346 343 342 346 341 344 348 346 346 341  
 344 342 344 345 340 344 344 343 344 342 342 343 345  
 339 350 337 345 349 336 348 344 345 332 342 341 350  
 343 347 340 344 353 341 340 353 346 345 346 341 339  
 342 352 342 350 348 344 350 335 340 338 345 345 349  
 336 342 338 343 343 341 347 341 347 344 339 347 358  
 343 347 346 344 345 350 341 338 343 339 343 346 342  
 339 343 350 341 346 341 345 344 342

试画出直方图, 并检验总体是否服从正态分布( $\alpha = 0.05$ ).

12. 测量合金强度  $y(\text{kg/mm}^2)$  与其中的含碳量(%), 得到数据如下表所示, 试求  $y$  对  $x$  的线性回归方程, 并检验回归效果.

$x$	0.1	0.11	0.12	0.13	0.14	0.15	0.16	0.17	0.18	0.20	0.21	0.23
$y$	42.0	41.5	45.0	45.5	45.0	47.5	49.0	55.0	50.0	55.0	55.5	60.5

13. 已知数据如下:

序号	$x_1$	$x_2$	$x_3$	$y$	序号	$x_1$	$x_2$	$x_3$	$y$
1	0.4	53	158	64	10	12.6	58	112	51
2	0.4	23	163	60	11	10.9	37	111	76
3	3.1	19	37	71	12	23.1	46	114	96
4	0.6	34	157	61	13	23.1	50	134	77
5	4.7	24	59	54	14	21.6	44	73	93
6	1.7	65	123	77	15	23.1	56	168	95
7	9.4	44	46	81	16	1.9	36	143	54
8	10.1	31	117	93	17	26.8	58	202	168
9	11.6	29	173	93	18	29.9	51	124	99

试求  $y$  对  $x_1$ ,  $x_2$  和  $x_3$  的线性回归方程并检验回归效果, 能否剔除一个变量?

14. 炼钢厂出钢时所用的盛钢水的钢包, 由于钢水对耐火材料的侵蚀作用, 随着使用次数的增加, 容积不断增大, 实测得到 15 组数据如下:

使用次数 $x$	2	3	4	5	6	7	8	9
增大容积 $y$	6.70	8.20	9.58	9.50	9.70	10.00	9.96	9.99
使用次数 $x$	10	11	12	13	14	15	16	
增大容积 $y$	10.49	10.59	10.60	10.80	10.60	10.90	10.76	

试分别按以下两种形式建立  $y$  对  $x$  的回归方程, 画出散点图和回归曲线, 并根据  $F$  值判断哪一种好.

$$(1) \frac{1}{y} = a + \frac{b}{x};$$

$$(2) y = ce^{b/x}.$$

15. 已知数据如下:

$x$	0	0.5	1	1.5	2	2.5	3	3.5	4	4.5	5	5.5
$y$	3.6	7.5	10.7	12.7	13.1	12.5	11.3	10.7	11.8	15.4	22.2	3.6

试求形式为  $y = a_0 + a_1x^2 + a_2x^3 + a_3 \sin x$  的回归方程并检验回归效果.