## 1.9 数据处理的基本方法

前几节我们从测量误差的概念出发，讨论了测量结果的最佳值和误差的估算.我们进行科学实验的目的是为了找出事物的内在规律，或检验某种理论的正确性，或准备作为以后实验工作的一个依据.因而对实验测量收集的大量数据资料必须进行正确的处理.数据处理是指从获得数据起到得出结论为止的加工过程，包括记录、整理、计算、作图、分析等方面的处理.本节中主要介绍常用的列表法、图示法、图解法和最小二乘法线性拟合等.

### 1.9.1 列表法

在记录和处理数据时，将数据排列成表格形式，既有条不紊，又简明醒目；既有助于表示出物理量之间的对应关系，也有助于检验和发现实验中的问题.列表记录、处理数据是一种良好的科学工作习惯.对初学者来说，要设计出一个项目清楚、行列分明的表格虽不是很难办到的事，但也不是一蹴而就的，需要不断地训练，逐渐形成习惯.

数据在列表处理时，应该遵循下列原则：

（1）各项目（纵或横）均应标明名称及单位，若名称用自定的符号，则需加以说明.

（2）列入表中的数据主要应是原始测量数据，处理过程中的一些重要中间结果也应列入表中.

（3）项目的顺序应充分注意数据间的联系和计算的程序，力求简明、齐全、有条理.

（4）若是函数测量关系的数据表，则应按自变量由小到大或由大到小的顺序排列.

下面以使用螺旋测微器测量钢球直径 $D$ 为例，列表记录和处理数据（表 01-7）.

表 01-7　测钢球直径 $D$*

使用仪器:0~100 mm 一级螺旋测微器,$\Delta_{仪}=\pm0.004$ mm

| 测量次序 | 初读数/mm | 末读数/mm | 直径 $D_i$/mm | $V_i=(D_i-\bar{D})$/mm | $V_i^2/(10^{-8}\text{mm}^2)$ |
|---|---|---|---|---|---|
| 1 | 0.004 | 6.002 | 5.998 | +0.001 3 | 169 |
| 2 | 0.003 | 6.000 | 5.997 | +0.000 3 | 9 |
| 3 | 0.004 | 6.000 | 5.996 | −0.000 7 | 49 |
| 4 | 0.004 | 6.001 | 5.997 | +0.000 3 | 9 |
| 5 | 0.005 | 6.001 | 5.996 | −0.000 7 | 49 |
| 6 | 0.004 | 6.000 | 5.996 | −0.000 7 | 49 |
| 7 | 0.004 | 6.001 | 5.997 | +0.000 3 | 9 |
| 8 | 0.003 | 6.002 | 5.999 | +0.002 3 | 529 |
| 9 | 0.005 | 6.000 | 5.995 | −0.001 7 | 289 |
| 10 | 0.004 | 6.000 | 5.996 | −0.000 7 | 49 |
| 平均 | | | $\bar{D}=5.996\ 7$ mm | $\sum V_i=0$ | $\sum V_i^2=1\ 210\times10^{-8}\text{mm}^2$<br>$S_{\bar{D}}=0.001\ 6$ mm |

* 若用计算器计算 $\bar{D}$,$S_{\bar{D}}$,则后两列可省.

由表 01-5 可查得 $t_{0.683}(n-1)=1.06(n=10)$,A 类不确定度 $S=t_{0.683}(n-1)S_{\bar{D}}=0.000\ 42$ mm,B 类不确定度 $u=0.683\Delta_{仪}=0.002\ 73$ mm.合成不确定度 $U=\sqrt{S^2+u^2}=0.002\ 8$ mm.

最后结果:$D=\bar{D}\pm U=(5.996\ 7\pm0.002\ 8)$mm$(P=0.683)$.

上列表格中数据在计算 $D$ 的平均值时多保留一位.一般处理时中间过程往往多保留一位,以使运算中不至于失之过多.最后仍应按有效数字有关规则取舍.

### 1.9.2　图示法和图解法

#### 1.9.2.1　图示法

物理规律既可以用解析函数关系表示,也可以借助图线表示.工程师和科学家一般对定量的图线最感兴趣,因为定量图线能形象直观地表明两个变量之间的关系.特别是对那些尚未找到适当解析函数表达式的实验结果,可以从图示法所画出的图线中去寻找相应的经验公式.

作图并不复杂,但对许多初学者来说,却并不一定能作得很好.这是由于他们缺乏作图的基本的训练,而且在思想上对作图又不够重视所致.然而

只要认真对待，并遵循作图的一般规则进行一段时间的训练，是能够绘制出相当好的图线的.

制作一幅完整而正确的图，其基本步骤包括：图纸的选择，坐标的分度和标记，标出每个实验点，作出一条与多数实验点基本相符合的图线，以及注解和说明，等等.

（1）图纸的选择　图纸通常有线性直角坐标纸（毫米方格纸）、对数坐标纸、半对数坐标纸、极坐标纸等，应根据具体实验情况选取合适的坐标纸.

因为图线中直线最易绘制，也便于使用，所以在已知函数关系的情况下，作两变量之间的关系图线时，最好通过变量变换将某种函数关系的曲线变换为线性函数关系的直线.

例如：

① $y=a+bx$，$y$ 与 $x$ 为线性函数关系.

② $y=a+b\dfrac{1}{x}$，若令 $u=\dfrac{1}{x}$，则得 $y=a+bu$，$y$ 与 $u$ 为线性函数关系.

③ $y=ax^b$，取对数，则 $\lg y=\lg a+b\lg x$，$\lg y$ 与 $\lg x$ 为线性函数关系.

④ $y=ae^{bx}$，取自然对数，则 $\ln y=\ln a+bx$，$\ln y$ 与 $x$ 为线性函数关系.

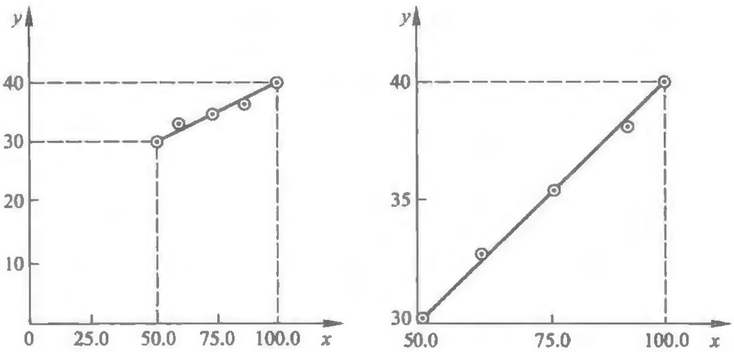
对于①，选用线性直角坐标纸就可得直线；对于②，以 $y$，$u$ 为坐标时，在线性直角坐标纸上也是一条直线；对于③，在选用对数坐标纸后，不必对 $x$，$y$ 作对数计算，就能得一条直线；对于④，则应选择半对数坐标纸作图，才可得到一条直线.如果只有线性直角坐标纸，而要作③、④两类函数关系的直线时，则应将相应的测量值进行对数计算后再作图.

（2）坐标的分度和标记　绘制图线时，应以自变量作横坐标，以应变量作纵坐标，并标明各坐标轴所代表的物理量（可用相应的符号表示）及其单位.

坐标的分度要根据实验数据的有效数字和对结果的要求来确定.原则上，数据中的可靠数字在图中也应是可靠的，而最后一位的存疑数在图中亦是估计的，即不能因作图而引进额外的误差.

在坐标轴上每隔一定间距应均匀地标出分度值，标记所用有效数字位数应与原始数据的有效数字位数相同，单位应与坐标轴的单位一致.坐标的分度应以不用计算便能确定各点的坐标为原则，通常只用 1，2，5 进行分度，避免用 3，7 等进行分度.

坐标分度值不一定从零开始，可以用低于原始数据的某一整数作为坐标分度的起点，用高于测量所得最高值的某一整数作为终点，这样图线就能充满所选用的整个图纸（如图 01-9 所示）.




(a) 不正确　　　(b) 正确

**图 01-9　坐标分度的起点选择**

（3）标点　根据测量数据，用“+”或“⊙”记号标出各数据点在坐标纸上的位置，记号的交叉点或圆心应是测量点的坐标位置，“+”中的横竖线段、“⊙”中的半径表示测量点的误差范围.

欲在同一图纸上画出不同图线，标点应该用不同符号，以便区分，如图 01-10 所示.

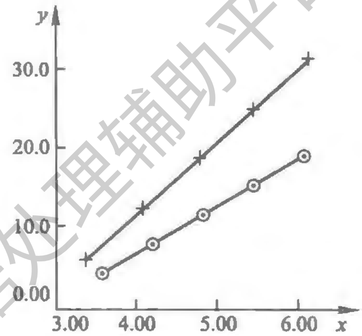
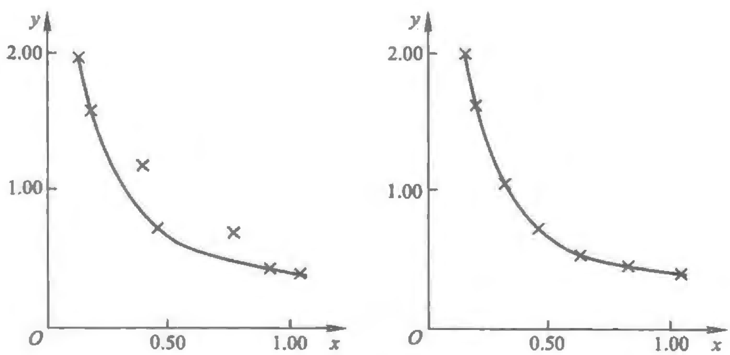



**图 01-10　不同图线的实验点标记**

（4）作一条与标出的实验点基本相符合的图线　连线时必须使用工具（最好用透明的直尺、三角板、曲线板等），所绘的曲线或直线，应光滑匀称，而且要尽可能使所绘的图线通过较多的测量点，但不能连成折线. 对那些严重偏离曲线或直线的个别点，应检查一下标点是否有误，若没有错误，在连线时可舍去不考虑，其他不在图线上的点，应使它们均匀地分布在图线的两侧，如图 01-11 所示.




(a) 不正确　　　(b) 正确

**图 01-11　图线的连接**

对于仪器仪表的校正曲线，连接时应将相邻的两点连成直线，整个校正曲线呈折线形式，如图 01-12 所示.

(5) 图名写在图的正下方.

### 1.9.2.2 图解法

利用已作好的图线，定量地求得待测量或得出经验方程，称为图解法. 尤其当图线为直线时，采用此法更为方便.

直线图解一般是求出斜率和截距，进而得出完整的线性方程，其步骤为：

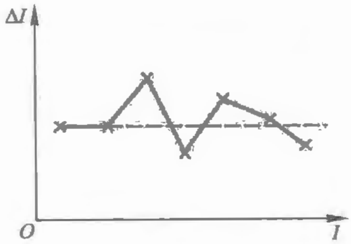



图 01-12 电流表的校正曲线

(1) 选点 为求直线的斜率，通常用两点法，不用一点法，因为直线不一定通过原点. 在直线的两端任取两点 $A(x_1,y_1)$，$B(x_2,y_2)$. 一般不用实验点，而是在直线上选取，并用与实验点不同的记号表示，在记号旁注明其坐标值. 这两点应尽量分开些（如图 01-13 所示）. 如果两点太靠近，计算斜率时会使结果的有效数字减少；但也不能取得超出实验数据范围以外，因为选这样的点无实验依据.

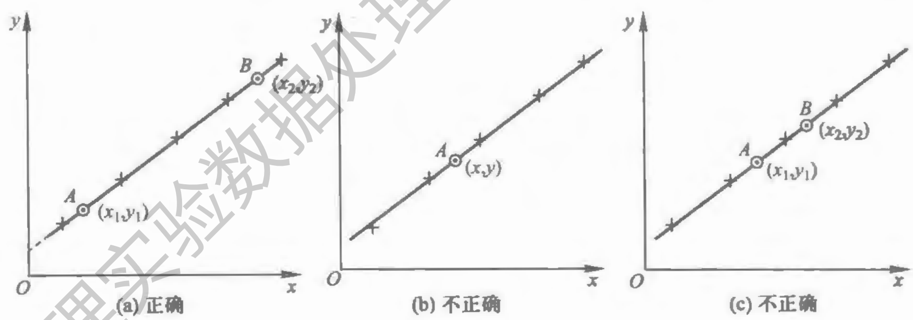



图 01-13 图解法的选点

(2) 求斜率 设直线方程 $y=a+bx$，将两点坐标值代入，可得直线斜率

$$b=\frac{y_2-y_1}{x_2-x_1} \tag{01-23}$$

(3) 求截距 若坐标原点为(0,0)，则可将直线用虚线延长，得到与坐标轴的交点，即可求得截距.

若起点不为零，则可由计算得

$$a=\frac{x_2y_1-x_1y_2}{x_2-x_1} \tag{01-24}$$

下面以测量热敏电阻的阻值随温度变化的关系为例进行图示和图解. 根据理论，热敏电阻的阻值 $R_T$ 与温度 $T$ 的函数关系为

$$R_T=ae^{\frac{b}{T}}$$

式中，$a,b$ 为待定常量，$T$ 为热力学温度，为了能变换成直线形式，将两边取对数得

$$\ln R_T=\ln a+\frac{b}{T}$$

并作变换，令

$$y=\ln R_T,\quad a'=\ln a,\quad x=\frac{1}{T}$$

则得直线方程为 $y=a'+bx$. 实验测量了热敏电阻在不同温度下的阻值后，以变量 $x,y$ 作图. 若 $y-x$ 图线为直线，就证明了 $R_T$ 与 $T$ 的理论关系式是正确的.

实验测量数据和变量变换值列于表 01-8 中. 图 01-14 为 $R_T-T$ 关系曲线；图 01-15 为 $\ln R_T-\frac{1}{T}$ 关系直线.

**表 01-8　热敏电阻阻值与温度的测量数据及其变量变换**

| 序号 | $t$/℃ | $T$/K | $R_T/\Omega$ | $x\left(=\frac{1}{T}\right)/(10^{-3}\mathrm{K}^{-1})$ | $y\left(=\ln\frac{R_T}{\Omega}\right)$ |
|---|---|---|---|---|---|
| 1 | 27.0 | 300.2 | 3 427 | 3.331 | 8.139 |
| 2 | 29.5 | 302.7 | 3 124 | 3.304 | 8.047 |
| 3 | 32.0 | 305.2 | 2 824 | 3.277 | 7.946 |
| 4 | 36.0 | 309.2 | 2 494 | 3.234 | 7.822 |
| 5 | 38.0 | 311.2 | 2 261 | 3.213 | 7.724 |
| 6 | 42.0 | 315.2 | 2 000 | 3.173 | 7.601 |
| 7 | 44.5 | 317.7 | 1 826 | 3.148 | 7.510 |
| 8 | 48.0 | 321.2 | 1 634 | 3.113 | 7.399 |
| 9 | 53.5 | 326.7 | 1 353 | 3.061 | 7.210 |
| 10 | 57.5 | 330.7 | 1 193 | 3.024 | 7.084 |

由 $A(3.050,7.175)$，$B(3.325,8.120)$ 可得

$$b=\frac{\ln(R_2/\Omega)-\ln(R_1/\Omega)}{\left(\frac{1}{T_2}\right)-\left(\frac{1}{T_1}\right)}=\frac{8.120-7.175}{(3.325-3.050)\times10^{-3}}\mathrm{K}=3.44\times10^3\ \mathrm{K}$$

$$a'=\frac{\left(\frac{1}{T_2}\right)\ln(R_1/\Omega)-\left(\frac{1}{T_1}\right)\ln(R_2/\Omega)}{\left(\frac{1}{T_2}\right)-\left(\frac{1}{T_1}\right)}=\frac{(3.325\times7.175-3.050\times8.120)\times10^{-3}}{(3.325-3.050)\times10^{-3}}$$

$$=-3.30$$

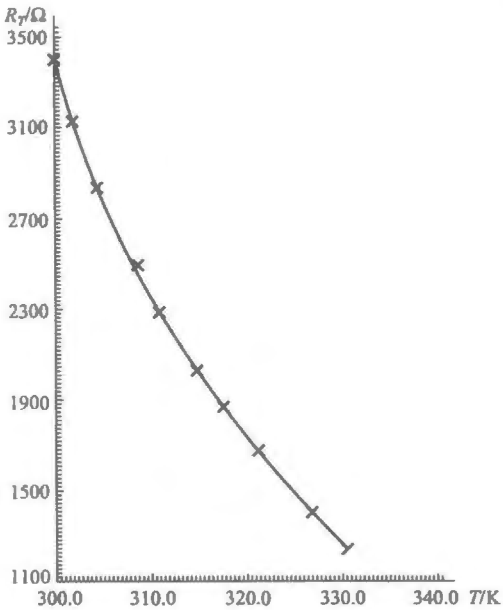



图 01-14　热敏电阻 $R_T-T$ 关系曲线

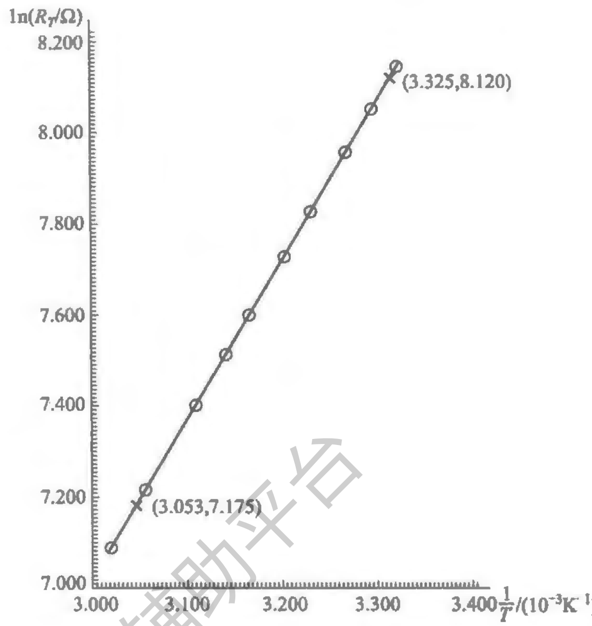



图 01-15　热敏电阻 $\ln R_T-\frac{1}{T}$ 关系曲线

因为　$a'=\ln a$

所以　$a=0.036\ 7\ \Omega$

最后可得该热敏电阻的阻值与温度关系为

$$R_T=0.036\ 7\ \Omega e^{3.44\times10^3 K/T}$$

### 1.9.3　逐差法

逐差法是物理实验中常用的数据处理方法之一.特别是在被测变量之间存在多项式函数关系,自变量等间距变化的实验中,更有其独特的优点.

逐差法就是把实验测量数据进行逐项相减,或者分成高、低两组实行对应项相减.前者可以验证被测量之间的函数关系,后者可以充分利用数据,具有对数据取平均和减少相对误差的效果.

例如:用受力拉伸法测定弹簧劲度系数 $k$,已知在弹性限度范围内,伸长量 $\Delta x$ 与所受拉力 $F$ 之间满足 $F=kx$ 关系.等间距地改变拉力(负荷),将测得的一组数据列于表 01-9 中.

逐项相减得 $\Delta x_i=x_{i+1}-x_i$ 分别为 1.50,1.52,1.48,1.51,1.49,1.50,1.50.可判断出 $\Delta x_i$ 基本相等,验证了 $\Delta x_i$ 与 $F$ 的线性关系.实际上,这一"逐差验证"工作,在实验测量过程中可随即进行,以判别测量是否正确.

表 01-9　弹簧伸长量与所受拉力测量数据

| 次数 | 拉力/($10^{-3}$ N) | 伸长量/($10^{-2}$ m) |
| --- | --- | --- |
| 1 | 0 | 0.00 |
| 2 | 2×9.8 | 1.50 |
| 3 | 4×9.8 | 3.02 |
| 4 | 6×9.8 | 4.50 |
| 5 | 8×9.8 | 6.01 |
| 6 | 10×9.8 | 7.50 |
| 7 | 12×9.8 | 9.00 |
| 8 | 14×9.8 | 10.50 |

但是,如果求弹簧负荷 $2\times9.8\times10^{-3}$ N 的平均伸长量$\overline{\Delta x_i}$,用上述逐项相减再求平均值时,有

$$\overline{\Delta x}=\frac{\sum\limits_{i=1}^{n}\Delta x_i}{n}=\frac{(x_2-x_1)+(x_3-x_2)+(x_4-x_3)+\cdots+(x_7-x_6)+(x_8-x_7)}{7}$$

$$=\frac{x_8-x_1}{7}=\frac{(10.50-0.00)\times10^{-2}}{7}\ \mathrm{m}=1.500\times10^{-2}\ \mathrm{m}$$

中间值全部无用,只有始、末两次测量值起作用,与负荷 $14\times9.8\times10^{-3}$ N 的单次测量等价.

若改用多项间隔逐差,将上述数据分成高组($x_8,x_7,x_6,x_5$)和低组($x_4,x_3,x_2,x_1$),然后对应项相减求平均值,得

$$\overline{\Delta x}=\frac{1}{4}[(x_8-x_4)+(x_7-x_3)+(x_6-x_2)+(x_5-x_1)]$$

于是各个数据全部都用上了.相当于重复测量了 4 次,每次负荷 $8\times9.8\times10^{-3}$ N.这样处理可以充分利用数据,体现出多次测量的优点,减小了测量误差.

### 1.9.4　最小二乘法和线性拟合

用图解法处理数据虽有许多优点,但它是一种粗略的数据处理方法,因为它不是建立在严格的统计理论基础上的数据处理方法,在作图纸上人工拟合直线(或曲线)时有一定的主观随意性.不同的人用同一组测量数据作图,可得出不同的结果.因而人工拟合的直线往往不是最佳的.所以,用图解法处理数据,一般是不求误差的.

由一组实验数据找出一条最佳的拟合直线(或曲线),常用的方法是最小二乘法.所得的变量之间的相关函数关系称为回归方程.所以最小二乘法线性拟合亦称为最小二乘法线性回归.

在这里我们只讨论用最小二乘法进行一元线性拟合问题,有关多元线性拟合与非线性拟合,读者可参阅其他专著.

最小二乘法原理是:若能找到一条最佳的拟合直线,那么这条拟合直线上各相应点的 $y$ 值与测量点的纵坐标之差的平方和在所有拟合直线中应是最小的.

假设所研究的两个变量 $x$ 和 $y$ 间存在线性相关关系,回归方程的形式为

$$y=a+bx \tag{01-25}$$

其图线是一条直线.测得一组数据 $x_i,y_i(i=1,2,\cdots,n)$,现在要解决的问题是:怎样根据这组数据来确定式(01-25)中的系数 $a$ 和 $b$.

我们讨论最简单的情况,即每个测量值都是等精度的,且假定 $x_i,y_i$ 中只有 $y_i$ 有明显的测量随机误差,如果 $x_i,y_i$ 都有误差,只要把相对来说误差较小的变量作为 $x$ 即可.

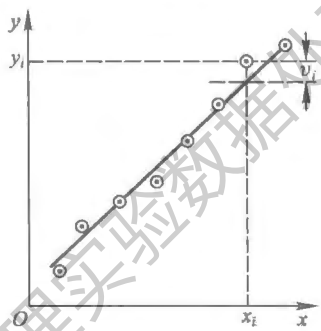


图 01-16　最小二乘法曲线拟合

由于存在误差,实验点是不可能完全落在由式(01-25)拟合的直线上的.对于和某一个 $x_i$ 相对应的 $y_i$ 与直线在 $y$ 方向上的残差为

$$v_i=y_i-y=y_i-a-bx_i \tag{01-26}$$

如图 01-16 所示,按最小二乘法原理应使

$$s=\sum_{i=1}^{n}v_i^2=\sum_{i=1}^{n}(y_i-a-bx_i)^2=s_{\min} \tag{01-27}$$



使其为最小的条件是

$$\frac{\partial s}{\partial a}=0,\quad \frac{\partial s}{\partial b}=0,\quad \frac{\partial^2 s}{\partial a^2}>0,\quad \frac{\partial^2 s}{\partial b^2}>0$$

由一阶微商为零得

$$\begin{aligned}\frac{\partial s}{\partial a}&=\sum 2(y_i-a-bx_i)(-1)=-2\sum(y_i-a-bx_i)=0\\ \frac{\partial s}{\partial b}&=-2\sum(y_i-a-bx_i)x_i=0\end{aligned} \tag{01-28}$$

式(01-28)(亦称正则方程组)可解得

$$a=\frac{\sum x_i\sum y_i-\sum x_i\sum(x_iy_i)}{n\sum x_i^2-(\sum x_i)^2} \tag{01-29}$$

$$b=\frac{n\sum(x_iy_i)-\sum x_i\sum y_i}{n\sum x_i^2-(\sum x_i)^2} \tag{01-30}$$

若令

$$[X]=\frac{\sum x_i}{n}$$

$$[Y]=\frac{\sum y_i}{n}$$

$$[XX]=\frac{\sum x_i^2}{n} \tag{01-31}$$

$$[XY]=\frac{\sum x_iy_i}{n}$$

则 $a,b$ 为

$$a=[Y]-b[X] \tag{01-32}$$

$$b=\frac{[XY]-[X][Y]}{[XX]-[X]^2} \tag{01-33}$$

式(01-27)对 $a,b$ 求二阶微商后，可知 $\frac{\partial^2 s}{\partial a^2}>0$，$\frac{\partial^2 s}{\partial b^2}>0$，这样式(01-32)和式(01-33)给出的 $a,b$ 对应于 $s=\sum v_i^2$ 的极小值，即用最小二乘法对拟合直线所得的两个参量：斜率和截距.于是，就得到了直线的回归方程式(01-25).

在前述假定只有 $y_i$ 有明显随机误差条件下，$a$ 和 $b$ 的标准误差可以用下列两式计算：

$$\sigma_a=\sqrt{\frac{\sum x_i^2}{n\sum x_i^2-(\sum x_i)^2}}\cdot\sigma_y=\sqrt{\frac{[XX]}{n([XX]-[X]^2)}}\cdot\sigma_y \tag{01-34}$$

$$\sigma_b=\sqrt{\frac{n}{n\sum x_i^2-(\sum x_i)^2}}\cdot\sigma_y=\sqrt{\frac{1}{n([XX]-[X]^2)}}\cdot\sigma_y \tag{01-35}$$

式中，$\sigma_y$ 为测量值 $y_i$ 的标准误差，即

$$\sigma_y=\sqrt{\frac{\sum v_i^2}{n-2}}=\sqrt{\frac{\sum(y_i-a-bx_i)^2}{n-2}} \tag{01-36}$$

式中，$n-2$ 是自由度，其意义是：如果只有两个实验点，$n=2$，有两个正规方程就可以解出结果了，而且 $y_i$ 的 $\sigma_y$ 应当为零，所以自由度是 $n-2$.

如果实验是在已知线性函数关系下进行的，那么用上述最小二乘法线性拟合，可得出最佳直线及其截距 $a$，斜率 $b$，从而得出回归方程.如果实验是要通过 $x,y$ 的测量值来寻找经验公式，则还应判断由上述一元线性拟合所找出的线性回归方程是否恰当.这可用下列相关系数 $r$ 来判别：

$$r=\frac{\sum(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum(x_i-\bar{x})^2\cdot\sum(y_i-\bar{y})^2}}=\frac{[XY]-\bar{x}\bar{y}}{\sqrt{([XX]-\bar{x}^2)([YY]-\bar{y}^2)}} \tag{01-37}$$

相关系数 $r$ 的数值大小表示了相关程度的好坏.如图 01-17 所示,若 $r=\pm1$ 表示变量 $x,y$ 完全线性相关,拟合直线通过全部实验点;当 $|r|<1$ 时,实验点的线性不好,$|r|$ 越小线性越差,$r=0$ 表示 $x$ 与 $y$ 完全不相关.

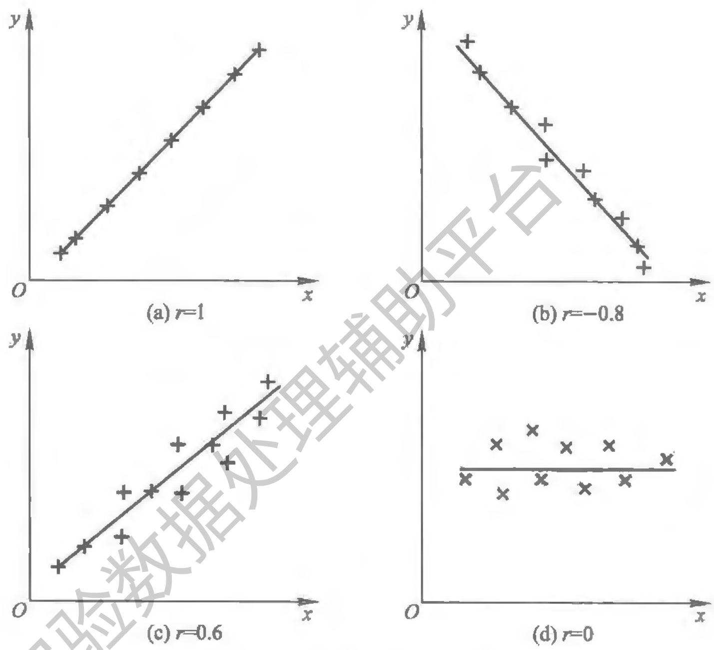



图 01-17　不同相关系数的拟合情况