绪　　论

一、化工原理实验的意义、目的

化工原理是以石油加工和化学工业生产过程中单元操作过程及设备为研究对象，紧密联系生产实际的化学工程学科的主干课程，是石油加工和化学工程专业的一门重要的专业基础课。除课堂教学外，实验教学是学生进一步学习、掌握课堂所学知识以及运用课堂所学知识分析、解决实际问题所必不可少的重要环节，是学生学好化工原理课程的重要组成部分。

化学工程学科的建立和发展证明，实验研究是不可缺少的手段。由于化学工业及石油加工生产过程中的许多问题和现象都是十分复杂的，其影响因素也十分繁多，单靠一些基本假设及理论模型还不能将其圆满解决，而且所建立的数学模型还必须得到实验的验证。因此，实验研究是解决化工生产问题的一条重要途径。化工原理课程中所讲述的许多经验公式无一不是通过实验研究得到的。由此可见，实验工作是化学工程这一学科建立和发展的重要基础。学生上好化工原理实验课对学好化工原理这门课程、正确理解和掌握课堂所学知识具有重要的意义。

通过实验教学可以达到以下目的：

(1) 配合课堂理论教学，在实验的过程中进一步掌握和正确运用所学基本理论和基本知识。

(2)通过实验教学,培养学生发现问题、分析问题和解决问题的实际工作能力 。

(3) 了解实验设备的结构、特点，掌握单元设备的工作原理和常规仪器的使用方法及实验研究的方法,以培养学生的实际动手能力 。

(4) 通过实验数据的分析、处理，编写实验报告，培养学生建立数学模型的能力及组织编写报告的能力。

(5) 通过实验培养学生严谨、求实和科学的工作作风。

二、本课程的内容、特点

本课程包括实验数据处理的基本理论及实验两大部分。在实验数据处理的基本理论中，主要介绍实验数据的误差分析基本原理及实验数据处理的基本方法；在实验部分主要介绍动量传递、热量传递及质量传递过程中所涉及的主要单元过程及设备的工作原理实验和性能测试实验。

本课程的特点之一是注重培养学生的综合能力，尊重学生自主学习的精神，重视学生在学习过程中的主体地位及学生的个性，给学生留有更多的思维发挥的空间，以利于培养学生创新思维能力 。学生通过自行拟定实验方案、制定实验步骤，自行拟定实验数据记录表,以及整理实验数据、得出实验结论、编制实验报告等一系列的过程,培养学生的综合能力。本课程的特点之二是注重理论联系实际，所开设的实验与课堂理论教学紧密配合，所

采用的实验设备大多可用于实际生产过程。如所采用的离心泵、板框式过滤机、换热器、精馏塔、萃取塔均可用于小型的生产过程中。这不仅使实验更具有实际性，而且实验中所遇到的问题和故障也是工业生产中这些设备常出现的，这有利于培养学生理论联系实际发现问题、分析问题、解决问题的能力。

三、对实验的要求

为了使学生充分利用有限的实验时间，在实验课中掌握更多的知识、达到实验的目的，要求学生做到以下几点。

(1) 实验的准备和预习。要完成好每一个实验，就必须做好实验课前的准备工作。在实验前要认真预习所学过的有关知识及实验指导书中的相关内容，并写出预习报告（包括实验目的、实验原理、实验步骤、实验数据记录表及预习中遇到的问题）。进入实验室后结合预习报告，了解和熟悉实验设备，掌握主要设备的流程及工作原理、测试仪器设备的使用方法。

(2) 在实验过程中，要精力集中，认真操作，注意观察实验现象。待过程稳定后测定有关数据。在实验中，要注意分析判断所测数据的可靠性，偏差较大的数据要找出原因并重新测定。

实验完毕后，应整理出实验记录数据，经教师审核后，关闭实验设备结束实验，并使实验设备恢复原状。实验中发现的设备问题要及时向指导教师报告，以便及时解决。

(3) 实验数据的记录。实验数据及现象的记录是处理数据得出正确结论的基础，做好记录是十分重要的。实验数据应记录在预先制定的表格中，记录数据要真实、认真、仔细、整齐、清楚。在测取及记录数据时要做到：

①稳定操作过程。在改变操作条件后，一定要待过程稳定后再读取数据，以反映真实的实验现象。

②记录的数据应是直接读取的数据。所记录下来的数据应是实验仪器及仪表所显示的数据，不要进行换算。如秒表的读数是1 分 50秒,应记为 1′50″而 不要记为 110″。

③根据测量仪器的精度，正确读取有效数字。

④在记录实验数据时，要采取科学的态度，不能随意修改或取舍实验数据，对所有的实验数据要经过误差分析等处理。

(4) 实验数据的处理及实验报告的编写。实验结束后，应及时处理实验数据，得出实验结论，按要求编写实验报告。实验报告是学生对所做实验的总结，通过编写实验报告，培养学生分析问题、解决问题及得出结论的能力。因此，应独立完成。实验报告的内容包括：

①实验项目名称、班级、姓名及同组者姓名、实验日期；

②实验目的；

③实验基本原理；

④实验设备流程简图及主要操作过程；

⑤ 原始实验数据记录表；

⑥ 实验数据的处理过程及实验结论；

⑦ 对实验现象及所存在的问题进行讨论与说明。

第一章　实验数据的误差分析及处理

实践证明，实验误差的估计及误差分析在评判实验结果及设计实验方案方面都有重要的指导意义。下面结合本课程的特点及化工过程中的具体情况，阐述这些内容。希望通过这一环节，掌握有关实验数据的误差分析及处理的基本方法。

第一节　误差的基本概念

一、真值与平均值

1. 真值

真值(又称真实值)是指某物理量客观存在的确定值,它通常是未知的。由于测量时所使用的测量仪器、测量方法以及环境、人的观察力、测量程序等方面的原因，实验误差很难避免，所以真值是无法测得的。根据正负误差出现几率相等的规律，当实验次数无限多时，测量结果的平均值，可以无限逼近于真值。但是，测量次数总是有限的，由此求出的平均值只能近似于真值，称此平均值为最佳值。计算时可将此最佳值作为真值使用，在实际应用过程中，有时也把高一级精度测量仪器的测量值作为真值使用。

2. 平均值

在工程计算中常将测量的平均值作为真值，但是，化工过程中所研究的问题不同，平均值的定义不同。化工中常用的平均值有：

(1) 算术平均值　在工程中算术平均值最常用。设 $x_1, x_2, \cdots, x_i, \cdots, x_n$ 代表各次测量的测量值,其中 n 为测量次数,x_i 为第 i 次的测量值，则算术平均值的表达式为：

$$\bar{x} = \frac{x_1 + x_2 + \cdots\cdots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1-1}$$

用最小二乘法的原理可以证明，在测定中当测量值的误差服从正态分布时，则在同一等级精度的测量中，算术平均值为最佳值或最可信赖值。

(2) 几何平均值　在工程计算中几何平均值也经常用到。其表达式为：

$$\bar{x} = (x_1 \cdot x_2 \cdot \cdots\cdots \cdot x_n)^{1/n} \tag{1-2}$$

以对数形式表示为：

$$\lg \bar{x} = \frac{\sum_{i=1}^{n} \lg x_i}{n} \tag{1-3}$$

当将一组测量值取对数，所得图形的分布呈对称形时，常用几何平均值表示。可以看出，几

何平均值的对数等于这些测量值的对数的算术平均值，几何平均值常小于算术平均值。

(3) 对数平均值　在化学反应过程、三传过程中，许多物理量的变化分布曲线常具有对数特性，此时采用对数平均值才符合实际情况。对数平均值的表达式为：

$$\bar{x} = \frac{x_2 - x_1}{\ln x_2 - \ln x_1} = \frac{x_2 - x_1}{\ln (x_2/x_1)} \tag{1-4}$$

对数平均值总小于算术平均值，当 $x_2 > x_1$,且 $x_2/x_1 < 2$ 时,可以用算术平均值代替对数平均值，所引起的误差不超过 4%,这在工程计算中是允许的 。

(4) 均方根平均值　均方根平均值多用于计算气体的分子平均动能。其表达式为：

$$\bar{x} = \sqrt{\frac{x_1^2 + x_2^2 + \cdots\cdots + x_n^2}{n}} = \sqrt{\frac{\sum_{i=1}^{n} x_i^2}{n}} \tag{1-5}$$

应当指出的是，在化工过程及化工实验研究中，数据的分布大多属于正态分布，所以常采用算术平均值。

二、误差的分类

误差通常是指测量值与真值之差，而偏差是指测量值与平均值之差。当测量次数足够多时，误差与偏差很接近，所以通常将二者混用。根据误差产生的原因及性质，可将误差分为系统误差、偶然误差和过失误差三类。

1. 系统误差

系统误差是由某些固定的原因造成的，它具有单向性，即在相同的条件下进行多次测量时，其正负、大小都有一定的规律性；或者是随着条件的改变而有规律的变化。引起系统误差的原因主要有：

(1) 测量仪器、设备方面的因素。如由仪器设计、制造上存在的某些缺陷，安装不合乎要求或未经核准而使用等引起的误差。

(2) 测量方法方面的影响因素。如由使用近似的测量方法或使用近似的计算公式而引起的误差。

(3) 测量环境方面的因素。如由环境温度、压力、湿度、振动等引起的测量误差。

(4) 测量者的因素。如由测量者读数等某些习惯上的偏向等引起的误差。

(5) 过程滞后的因素。如在动态过程的测量中，由于过程的滞后因素，测量时并未达到平衡或稳定的状态而引起的误差。

尽管系统误差的影响因素很多，但具有一定的规律性，一般情况下只要根据产生误差的原因采取适当的措施进行修正，就可以消除系统误差。

2. 偶然误差

偶然误差又称随机误差，它是由某些意想不到的因素或难以控制的因素引起的。其主要表现为：在相同的条件下进行测量时，其误差值无固定的规律可循。它不同于系统误差，不能从系统中消除。但是，它的出现服从统计规律，测量误差与测定次数有关，随着测量次数的增加，误差有正负抵消的可能。因此，多次测量值的算术平均值将通近于真值。可采用统计概率的方法对偶然误差进行研究。

3. 过失误差

过失误差主要是由测量人员在测量过程中粗心大意或操作不当引起的，它是明显与实际不符的误差。其消除要靠测量人员严肃认真的工作态度和细致的校对工作来避免。对这种误差，可通过某些原则加以判断，在处理数据时进行取舍。

综上所述，系统误差和过失误差是可以消除的。如在使用前应对仪器、设备进行校正，读数时要待过程稳定等。而偶然误差是不易消除的，这种误差是误差理论的主要研究对象。

三、误差的表示方法

前面所述误差的概念，不能说明测量值与真值的近似程度。如工人甲平均每生产 100 个零件有 1 个次品，而工人乙平均每生产 500 个零件有 1 个次品。他们的次品虽然都是 1 个，但显然乙的技术要比甲的高，这就启发我们不但要看次品的数，而且还要注意到产品的次品率。显然甲的次品率是 1%，而乙的次品率是 0.2%。因此，误差有多种表示方法，要依据具体情况使用相应的误差表示方法。

1. 绝对误差

绝对误差是近似值（测量值）与真值之间的差值。

设测量值为 x，真值为 X，绝对误差为 e，则有

$$e = |x - X| \tag{1-6}$$

即

$$x - X = \pm e$$

或

$$x - e \leqslant X \leqslant x + e \tag{1-7}$$

由于在一般情况下真值 X 是未知的，所以误差 e 的绝对值也不能求出，但根据测量或计算的实际情况，可事先估计出误差的绝对值不能超过某一个正数 ε，我们称 ε 叫做误差绝对值的上限或最大误差，又记为 ε_{max}。此时，真值 X 符合：

$$x_1 = \bar{x} + \varepsilon_{max} > X > \bar{x} - \varepsilon_{max} = x_2$$

$$\bar{x} = (x_1 + x_2)/2 \tag{1-8}$$

$$\varepsilon_{max} = (x_1 - x_2)/2 \tag{1-9}$$

式中：x_1——测量的最大值；

x_2—— 测量的最小值；

$\bar{x}$——两次测量值的算术平均值。

也就是说，数 $\bar{x}$ 是误差为 ε_{max} 的数 X 的近似值。

2. 相对误差

由于绝对误差不能全面地反映测量值与真值的近似程度，所以引入相对误差。相对误差的表达式为：

$$e_r = e/X \tag{1-10}$$

式中 e_r——相对误差。

一般情况下，真值是未知的，可以用多次测量的近似值(平均值)来代替。

3. 算术平均误差 δ

算术平均误差的表达式为：

$$\delta = \frac{\sum_{i=1}^{n} |x_i - \bar{x}|}{n} \tag{1-11}$$

式中 δ——算术平均误差；

x_i——第 i 次的测量值；

$\bar{x}$——n 次测量值的平均值(近似值)；

n——测量次数。

式(1-11)中必须取绝对值,否则，$\sum_{i=1}^{n}(x_i - \bar{x}) \equiv 0$。

算术平均误差的缺点是无法表示出各组测量之间彼此符合的情况。因为在一组测量值很接近（各次测量的误差接近）的情况下所得的算术平均误差，可能与另一组测量值中测量误差有大有小所得的算术平均误差相同。

4. 均方根误差 σ

均方根误差又称标准误差，它不仅与一组测量值中的每一个数据有关，而且对一组测量值中较大的误差和较小的误差的敏感性很强。当测量次数为无穷多时，其表达式为：

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - X)^2}{n}} \tag{1-12}$$

当测量次数有限时，真值 X 可用平均值 $\bar{x}$ 代替,此时,均方根误差可用式(1-13)计算：

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \tag{1-13}$$

算术平均值相同的两组测量值，其均方根误差也会不同，它能反映出一组测量值的离散程度。因而这种误差广泛用于化学工程的实验数据处理过程中。

四、精密度、正确度及准确度

在测量时，可以用误差表示数据的可靠性，也可以用精密度（简称为精度）等概念来表示。习惯上，所讲的精密度，通常是指误差。这种误差的来源、性质一般可用以下概念来描述。

1. 精密度

精密度是对某物理量进行几次平行测定的测量值相互接近的程度，即重现性。它反映了偶然误差的影响程度，偶然误差越小则精密度越高。如果纯由偶然误差引起的实验的相对误差为 0.1%,则可认为精密度为 10^{-3}。

2. 正确度

正确度是指在一定的测量条件下，没有偶然误差的影响，测量值与真值的符合程度，是测量中所有系统误差的综合。它反映了所有系统误差对测量值的影响，系统误差愈小则

正确度愈高。如果纯由系统误差引起的实验相对误差为 0.1%，则可认为其正确度为 10^{-3}。

3. 准确度

准确度是指在测量过程中，测量值与真值之间的符合程度，是所有偶然误差及系统误差的综合。它反映了偶然误差及系统误差对测量值的影响程度，准确度越高则表示系统误差及偶然误差越小。也可以说准确度表示的是测量值与真值之间的符合程度。如果由偶然误差及系统误差引起的测量的相对误差为 0.1%，则测量值的准确度为 10^{-3}。

对于实验或测量而言，精密度好，并非表示正确度一定好，反之亦然。但是准确度好则必须是精密度和正确度都好。图 1-1 中，(a)说明测量结果与真值接近，系统误差与偶然误差均小，准确度好；(b)说明精密度高，偶然误差小，但系统误差大；(c)说明偶然误差大，但系统误差较小，即精密度低而正确度较高。

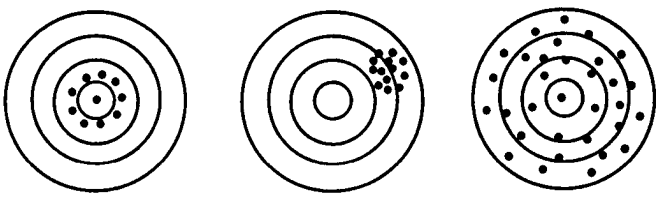

图 1-1　准确度、精密度及正确度示意图

第二节　有效数字及运算法则

一、有效数字

在记录测量数据以及对测量数据进行计算时，确定测量数据及计算结果的有效数字的位数是很重要的。测量值有效数字的位数，应正确反映所使用测量仪器和测量方法所能达到的精度。如一支量程 0～100 ℃的温度计，其最小刻度为 0.1 ℃，当读数为 50.25 ℃时，有效数字是 4 位；若指示液面正好位于 50.2 ℃时，应记为 50.20 ℃，其有效数字也为 4 位。这里所记录的最后一位数字是估计的，称为可疑数字，而前 3 位数字是从刻度上直接读出的，称为可靠数字。可靠数字比有效数字少 1 位。即记录数据时，有效数字应保留 1 位可疑数字。如上面提到的 50.20 ℃，可疑数字表示该位上有 ±1 个单位或下一位有 ±5 个单位的读数误差。

一个数据中，除定位用的“0”外，其他数字都是有效数字(包括 1 至 9 以及它们中间的“0”和四舍五入后保留下来的数字“0”)。也就是说，数字“0”在前面不是有效数字，在后面用于定位的也不是有效数字。例如：长度为 0.002 34 m，前面的 3 个“0”不是有效数字，这与所用的单位有关。若以 mm 为单位，则为 2.34 mm，其有效数字是 3 位。那么长度为 360 000 cm 的有效数字是几位呢？若后面的 3 个“0”是用来定位的，则都不是有效数字，其有效数字为 3 位。为了能明确表示数据的有效数字位数，工程上常用一种科学的记数

法，如上面的读数应写成 3.60×10^{5} cm。这种记数法的特点是小数点前总是一个非“0”的数字，“×”前面的数字都是有效数字。这样，有效数字的位数就一目了然了。如 0.000 356 记为 3.56×10^{-4}，其有效数字为 3 位。

二、运算法则

在实验数据处理过程中，常常会遇到不同精度的数据一同运算，这时需按一定的法则进行运算，这不仅可以保证数据的有效数字位数，而且还可以避免由运算过于繁琐而引起的误差。

1. 四舍六入五留奇

当有效数字位数确定后，其余数字应一律舍去。目前多采用“四舍六入五留奇”或“四舍六入五变偶”规则对数字进行修约。即当末位有效数字之后第一位数字小于 5 时，舍去不计；大于 5 时，有效数字末位加 1；等于 5 时，末位有效数字为奇数时，则末位有效数字加 1 变为偶数，如末位有效数字为偶数则舍去。

如：1.256 76 有 4 位有效数字时记为 1.257；

1.265 56 有 4 位有效数字时记为 1.266；

1.265 56 有 3 位有效数字时记为 1.26。

2. 加减运算法则

在加减运算过程中，所得计算结果的小数点后位数，应与各加减数中小数点后的位数最少的那个数相同。例如：

$$\begin{aligned} &134 + 58.6 + 0.258 + 0.025\ 8 \\ =\ &192.883\ 9 \\ =\ &193(\text{与 } 134 \text{ 小数点后的位数相同}) \end{aligned}$$

又如：

$$\begin{aligned} &13.45 + 1.345 + 0.007\ 345 \\ =\ &14.802\ 345 \\ =\ &14.80(\text{与 } 13.45 \text{ 小数点后的位数相同}) \end{aligned}$$

实际计算时，为了简化起见，可以在进行加减计算之前就将各数据进行修约，舍去没有意义的数字。具体原则是，使加减数据中各数据的小数点后的位数与最少位数者相同。如上面的例子可以作下面的简化运算：

$$\begin{aligned} &134 + 58.6 + 0.258 + 0.025\ 8 \\ =\ &134 + 59 + 0 + 0(\text{与 } 134 \text{ 小数点后的位数相同}) \\ =\ &193 \end{aligned}$$

又如：

$$\begin{aligned} &13.45 + 1.345 + 0.007\ 345 \\ =\ &13.45 + 1.34 + 0.01(\text{与 } 13.45 \text{ 小数点后的位数相同}) \\ =\ &14.80 \end{aligned}$$

3. 乘除运算法则

在乘除法运算中，所得计算结果的有效数字位数，应与各数据中最少的有效数字的位

数相同，而与小数点的位置无关。例如：

$$0.012\,1 \times 25.64 \times 1.057\,82$$
$$= 0.328\,182\,3$$
$$= 0.328$$

此处以有效数字位数最少的 0.012 1 为准。

在计算中，也可以以有效数字位数最少的数据为准，先将各数据的有效数字进行简化，而后进行乘除计算。如上面的例子也可进行以下运算：

$$0.012\,1 \times 25.64 \times 1.057\,82$$
$$= 0.012\,1 \times 25.6 \times 1.06$$
$$= 0.328$$

此处先以有效数字位数最少的 0.012 1 为准，对各个数据进行简化，而后再进行计算。

4. 常数的有效数字

对于常数 g、π、e 及某些因子 1/3、$\sqrt{2}$、$\sqrt{3}$ 等的有效数字，可认为是无限的，需要几位就写几位。

5. 平均值的计算

若对 4 个或超过 4 个数据进行平均值计算时，则平均值的有效数字可增加一位。

6. 精度(或误差)的表示

在表示精度(或误差)时，一般只取 1～2 位有效数字，过多的位数已失去意义。如误差为 0.013 84，可写为 0.014。由于误差是用来表征数据结果的准确程度的，并提供必要的保险，所以适用于在误差值截断后末位进 1，以使误差大一些，而无须考虑通常的“四舍五入”原则。如：

$0.241\,2 \times 10^{-8}$ 可记为 0.25×10^{-8}

当然，这种方法是对最终表达误差而言的。

7. 测量结果及实验数据的表达

在表达测量及实验数据时，其最少位数应与保留的误差的位数对齐并取舍，其取舍应按“四舍六入五留奇”的原则进行。例如：

数据为：1.835 49	误差为 ±0.014	则记为： 1.835
数据为 6.325 0×10^{-6}	误差为 ±0.25×10^{-6}	则记为： 6.32×10^{-6}
数据为 7.385 5×10^{5}	误差为 ±0.048×10^{5}	则记为： 7.386×10^{5}

第三节　误差分析理论基础

为了找出一组含偶然误差（随机误差）的实验测量值与真值之间的关系，从而确定一最佳值，找出评判化工实验质量及水平高低的数量标准和确定可疑数据的取舍等问题，我们将分析各测量值的偶然误差及其频率的分布情况。

一、测量值的最佳值

由式(1-6)可知，$e_i = |x_i - X|$，即 e_i 为每次测量值与真值之差，在此又称其为每次测

量的偶然误差。对 n 次测量的偶然误差求和得：

$$\sum_{i=1}^{n} e_i = \sum_{i=1}^{n} x_i - nX \tag{1-14}$$

则

$$X = \frac{\sum_{i=1}^{n} x_i}{n} - \frac{\sum_{i=1}^{n} e_i}{n} \tag{1-15}$$

根据偶然误差的正态分布特征，当 $n \to \infty$ 时：

$$\frac{\sum_{i=1}^{n} e_i}{n} = 0$$

则

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \to X \tag{1-16}$$

即当测量次数为无限多时，测量值的分布为一正态分布。在一组等精度测量中，算术平均值极接近于真值。而对于有限次数的测量时，算术平均值为真值的近似值，称为最佳值，或称为最可信值。

二、粗差及异常数值的剔除

在实验或测量过程中，由于可能发生的测量者的读数错误、记录错误，或者由于仪器设备的突然波动以及客观条件的突然变化，都会造成异常数据的出现。由异常数据产生的误差不能从实验或测量的客观条件中得到合理的解释，称为粗差。

如果一组测量值中混入含有粗差的异常数据，必然会歪曲测量结果，从而得到与客观实际不相符的结论。因此，从测量值中剔除异常数据是很重要的一个步骤。但是，也有另外一种情况，一组正确的测量数据本身具有较大的分散性，并反映了客观实际情况，如果人为地舍弃了一些主观上认为误差较大而实际上并非真正的异常数据，同样会导致得到虚假的结论。所以异常数据的剔除必须遵循一定的准则，决不能任意地主观确定。鉴别异常数据的方法有两类：一类称为技术判别法，即根据物理的或化学的原因作出明确的技术分析来判别某个误差较大的数据是否确属异常数据；另一类称为统计判别法，即在不能用技术分析来确定测量数据中的异常数据时，往往需要用统计判别的方法进行判别。下面分别介绍几种常用的统计判别方法。

1. 三倍标准误差(3σ)判别法

这是一种最简单的判别方法。若对某一物理量独立测量 n 次，得到 $x_1, x_2, x_3, \cdots, x_n$ 等 n 个测量值。首先，取样本平均值：

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} \tag{1-17}$$

以及每个测量值对样本平均值的误差，即残差：

$$e_i = x_i - \bar{x}, \qquad i = 1,2,3,\cdots,n \tag{1-18}$$

然后由 Bessel 公式计算出测量值的标准差估计值：

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} e_i}{n-1}} \tag{1-19}$$

这一方法的判别准则是：若某个测量值 x_i 的残差 $e_i(1 \leqslant i \leqslant n)$满足

$$|e_i| > 3\sigma \tag{1-20}$$

则 x_i 是含粗差的异常数据，应予以剔除。

2. Grubbs 方法

将 n 个测量值从小到大排列成：$x_1, x_2, x_3, \cdots, x_i, \cdots, x_{n-1}, x_n$。很显然，最有可能为异常数据的应该是 x_1 或 x_n，Grubbs 导出：

$$g_1 = \frac{\bar{x} - x_1}{\sigma} \tag{1-21}$$

及

$$g_n = \frac{x_n - \bar{x}}{\sigma} \tag{1-22}$$

g_1 和 g_n 有相同的概率分布，即：

$$P\{g \geqslant G(n, \alpha)\} = \alpha \tag{1-23}$$

式中 g——为 g_1 或 g_n；

α——风险率；

$G(n,\alpha)$——Grubbs 临界值，见附录七；

n——测量次数。

风险率 α 是按 Grubbs 临界值 $G(n,\alpha)$判断为异常数据但实际上并不是异常数据，从而导致犯错误的概率。

Grubbs 方法剔除异常数值的判据为：对于测量中所得的最大值或最小值 x_k，若其残差 e_k 满足：

$$|e_k| > G(n,\alpha)\sigma \tag{1-24}$$

则 x_k 为异常数据，应予以舍弃。

例 1-1 有一组测量数据如下：

$x_i = 9.65, 9.84, 9.89, 9.93, 9.95, 9.99, 10.02, 10.05, 10.08, 10.12$；$i = 1, 2, \cdots, 10$。取风险率 $\alpha = 0.05$。试剔除该组数据中的异常数据。

解 (1) 计算标准算术平均值：

$$\bar{x} = \frac{\sum_{i=1}^{10} x_i}{n} = 9.95$$

误差为：$\sigma = \sqrt{\sum_{i=1}^{10} e_i/(n-1)} = 0.137$

(2) 设 x_1 为可疑数据，则偏差为：

$$e_1 = \bar{x} - x_1 = 9.95 - 9.65 = 0.30$$

(3) 查附录七得 Grubbs 临界值为：

$$G(10,0.05) = 2.176$$

则

$$G(n,\alpha)\sigma = 2.176 \times 0.137 = 0.298\,1 < |e_1|$$

故应将 x_1 舍弃。

对剔除第一个异常数据后的数据再作一次检查，取相同的风险率 α 判断是否还存在异常数据，直到无异常数据为止。

值得注意的是，在一般情况下 α 值不宜取得太小。这是因为 α 值越小，虽然把不是异常数据的数据判断为异常数据的概率 α 降低了，然而，却把异常数据判定为正常数据的错误概率增大了。

第四节　回归分析

描述一个事物的各个影响因素之间的关系，可分为两种类型。

一种类型是各变量之间存在着确定的关系。如通过电阻为 R 的电路中的电流 I 与电阻两端的电压 U 之间存在着确定的关系：

$$I = \frac{U}{R} \tag{1-25}$$

变量之间这种确定的关系称为函数关系。

另一种类型是各变量之间的关系往往存在着一定的不确定性。如在许多化工实际问题中，就存在着这样的问题。如流体流过管路时摩擦阻力与其影响因素之间的关系；对流传热系数与其影响因素之间的关系等等。造成这种不确定性的原因是过程中存在着暂时还没有认识到的影响因素或在测量中存在着不同程度的误差等。但从统计学的角度看，它们之间又确实存在着一定的关系，这种关系称为统计相关关系。

回归分析的主要任务就是通过大量的实验测量数据找出相关变量之间的定量的数学关系式。另外，还可以对所建立关系式的有效性进行检验，对影响因素进行分析，判断影响因素的重要性。通过实验数据的回归分析，建立数学模型（关系式），是解决影响因素复杂的化工过程问题的重要而有效的方法之一。

一、一元线性回归——直线拟合

在变量之间统计相关的问题中，最简单的情况就是两个变量之间的关系。例如，通过实验测定出变量 y 在一定程度上随另一个变量 x 的变化而变化的关系，通常称 x 为自变量，y 为因变量。如果这两个变量之间具有线性关系，就可以用一元线性回归分析方法找出统计地描述这两个变量之间的定量相关的数学表达式。

1. 直线拟合

直线拟合的出发点是应用最小二乘法原理处理实验数据，最后得到经验方程（公式）常数的最佳估计值。这一方法有两条基本假定：①实验中自变量 x 为给定数据，不带有实验误差，或者误差很小可以忽略不计。而因变量 y 是带有一定误差的实测值。②回归或拟

合得到的最好直线或曲线，与实验点的偏差（残差）的平方和最小。这两条假设适用于各种形式方程的回归。

设有一批实验数据，测量值为 $y_i, x_i (i=1,2,\cdots,n)$。若实验数据符合线性关系，或已知经验方程为直线形式，都可以回归为直线方程，即：

$$y^* = a + bx \tag{1-26}$$

由于实验存在误差，回归方程的计算值 y_i^* 与实验值 y_i 存在一定的残差，设为 e_i，则有：

$$e_i = y_i - y_i^* = y_i - (a + bx_i), \quad (i = 1,2,\cdots,n)$$

根据最小二乘法假设，回归方程的 y_i^* 与实验测定值 y_i 的残差平方和应该最小。即：

$$\sum_{i=1}^{n} e_i^2 = \text{最小}, (\text{以下记} \sum_{i=1}^{n} \text{为} \sum)$$

令：

$$Q = \sum e_i^2 = \sum [y_i - (a + bx_i)]^2 = \text{最小} \tag{1-27}$$

既然 Q 是 a 和 b 的非负二次函数，所以 Q 一定存在着最小值，可通过微分学中的极限原理对 a、b 求偏导数。令偏导数等于零求得 a 及 b。即：

$$\frac{\partial Q}{\partial a} = 0, \qquad \frac{\partial Q}{\partial b} = 0$$

$$\because \quad Q = \sum [y_i^2 - 2y_i a - 2y_i b x_i + a^2 + 2abx_i + b^2 x_i^2]$$

$$\therefore \quad \frac{\partial Q}{\partial a} - 2\sum [y_i - a - bx_i] = 0$$

$$\sum [y_i - a - bx_i] = 0$$

或：

$$\sum y_i - na - b\sum x_i = 0 \tag{1-28}$$

同理：

$$\frac{\partial Q}{\partial b} = -2\sum [y_i x_i - a x_i - b x_i^2] = 0$$

$$\sum (y_i x_i) - a\sum x_i - b\sum x_i^2 = 0 \tag{1-29}$$

整理式(1-28)及式(1-29)得：

$$na + b\sum x_i = \sum y_i \tag{1-30}$$

$$a\sum x_i + b\sum x_i^2 = \sum (x_i y_i) \tag{1-31}$$

以上两式为直线回归得到的正规方程，联立求解可得到 a 和 b 的值，此值即为最小二乘法原理得到的回归方程的最佳估计值。

由式(1-30)和式(1-31)联立方程组求解得：

$$a = \frac{(\sum x_i^2)(\sum y_i) - (\sum x_i)[\sum (x_i y_i)]}{n\sum x_i^2 - (\sum y_i)^2} \tag{1-32}$$

$$b = \frac{n[\sum (x_i y_i)] - (\sum x_i)(\sum y_i)}{n\sum x_i^2 - (\sum x_i)^2} \tag{1-33}$$

为了应用方便，也可以将解的形式表示为其他形式：

$$a=\frac{\sum y_i-b\sum x_i}{n}=\bar{y}-b\bar{x} \tag{1-34}$$

$$b=\frac{\sum[(x_i-\bar{x})(y_i-\bar{y})]}{\sum(x_i-\bar{x})^2} \tag{1-35}$$

式中：$\bar{x}=\frac{\sum x_i}{n}$，$\bar{y}=\frac{\sum y_i}{n}$，$n$ 是实验组数。

在实际应用中，凡可以化为直线方程的函数都可以用以上各式求直线方程的常数。例如幂函数 $y=ax^b$，取对数后可以化为直线方程，利用直线方程的斜率和截距，可求出原函数的系数 a 和指数 b。

2. 相关系数

在前面介绍的求线性回归方程系数的方法中，并没有事先判断两个变量之间是否真正存在着某种程度的线性相关。也就是说，就方法而言，即使是一组毫无关系的数据点，也能得到一条回归直线。因此，在用最小二乘法得到回归方程后，必须判断所得方程有多大意义。或者说，必须判断所关联的两个变量之间，究竟有多大的相关性，相关程度如何。通常，采用相关系数来判别。

在回归方程式 $y_i=a+bx_i$ 中，因为 $a=\bar{y}-b\bar{x}$，所以，实验值与计算值的残差平方和可表示为：

$$Q=\sum[y_i-y_i^*]^2=\sum[y_i-(\bar{y}-b\bar{x}+bx_i)]^2$$

$$=\sum[(y_i-\bar{y})-b(x_i-\bar{x})]^2$$

展开上式，并将式(1-35)变为 $\sum[(x_i-\bar{x})(y_i-\bar{y})]=b\sum(x_i-\bar{x})^2$ 代入得：

$$Q=\sum(y_i-\bar{y})^2-b^2\sum(x_i-\bar{x})^2=\sum[(y_i-\bar{y})^2-b^2(x_i-\bar{x})^2]$$

$$=\sum(y_i-\bar{y})^2\left[1-\frac{b^2\sum(x_i-\bar{x})^2}{\sum(y_i-\bar{y})^2}\right] \tag{1-36}$$

令：

$$r^2=\frac{b^2\sum(x_i-\bar{x})^2}{\sum(y_i-\bar{y})^2}$$

将式(1-35)代入上式，两边开方得：

$$r=\frac{\sum[(x_i-\bar{x})(y_i-\bar{y})]}{\sqrt{\sum(x_i-\bar{x})^2\sum(y_i-\bar{y})^2}} \tag{1-37}$$

这里的 r 称为相关系数，r 的符号取决于 $\sum(x_i-\bar{x})(y_i-\bar{y})$，因此与回归方程的斜率 b 一致，由于残差平方和 Q 总是大于零，由式(1-36)可知，$(1-r^2)\geqslant 1$，所以 $r^2\leqslant 1$，r 的变化范围为 $-1\leqslant r\leqslant 1$。相关系数 r 的意义如下：

(1) 当 $r=\pm 1$ 时，$Q=0$，即 n 组实验数据全部落在直线 $y=a+bx$ 上，见图 1-2(a)、(b)。

(2) $|r|$越接近于 1,则 Q 越小,那么 n 组实验数据越接近直线 $y=a+bx$;$|r|$偏离 1 越大,则实验数据点越偏离直线,见图 1-2(c)、(d)。

(3) 当 $r=0$ 时,则变量间无线性关系。实验点数据分散在直线周围,见图 1-2(e)、(f)。值得注意的是,$r=0$ 只是说明变量 x,y 之间不存在线性关系,但不说明它们之间不存在其他相关关系。

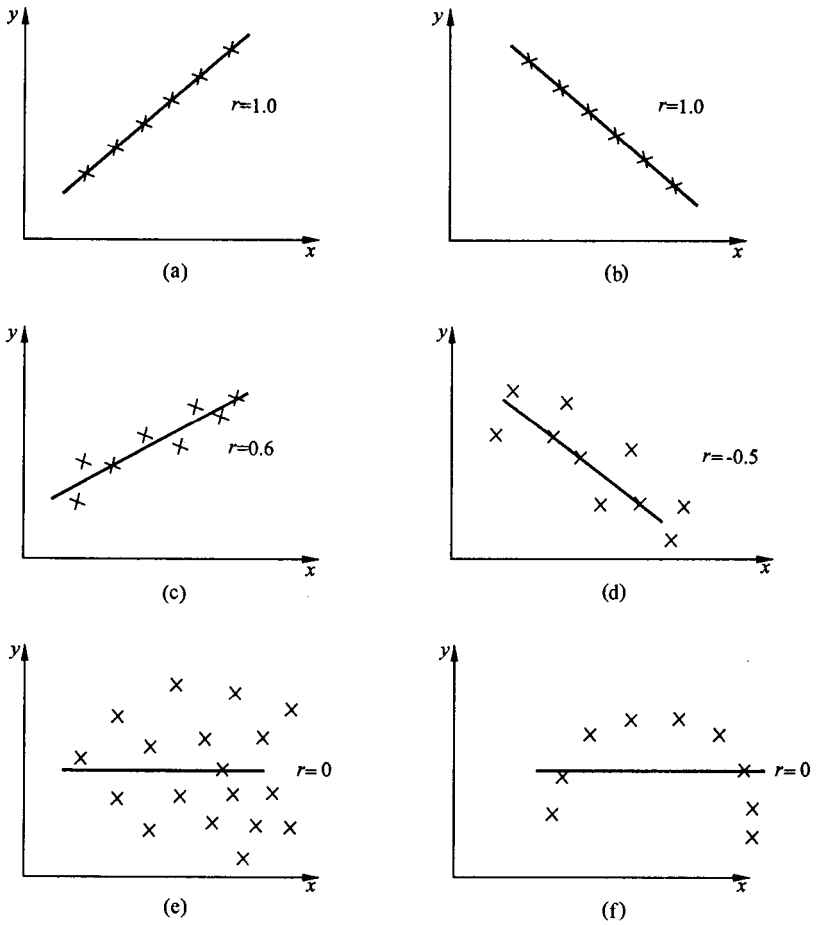

图 1-2 不同相关系数散点示意图

3. 回归方程的精度

回归方程的精度也是描述回归方程 $y=a+bx$ 对诸实验点 (x_i,y_i)的拟合程度,它是以标准残差的量作为衡量标准,其定义为:

$$\sigma_r=\sqrt{\frac{\sum(y_i-\overline{y}_i)^2}{n-q}}=\sqrt{\frac{\sum e_i^2}{n-q}},\quad (n>q) \tag{1-38}$$

式中,n 是实验点数,q 为回归方程中所需要的未知数的总个数。对于含两个变量的线性回归方程,$q=2$。由上式可以看出 σ_r 的值愈小,则回归方程的精度愈高。这一标准残差的概

念同样适用于多项式回归方程精度的检查和量度。

4. 回归方程各常数的标准误差

直线回归方程斜率和截距的标准误差可用式（1-39）和式(1-40)计算：

$$\sigma_a = \sqrt{\frac{n\sum e_i^2}{(n-2)[n\sum x_i^2 - (\sum x_i)^2]}} \tag{1-39}$$

$$\sigma_b = \sqrt{\frac{\sum e_i^2}{n-2}}\sqrt{\frac{\sum x_i^2}{n\sum x_i^2 - (\sum x_i)^2}} \tag{1-40}$$

5. 计算机程序

由已知实验数据 $y_i, x_i (i=1,2,\cdots,n)$，回归直线方程 $y = ay + b$ 的斜率 a 和截距 b。可用前面讲过的方法进行计算，其程序如下。

```
    REM THIS IS LINER REGRESSION COMPUTE PROGRAM
PRINT"LINER REGRESSION"
PRINT
PRINT"NUMBER OF KNOWN POINTS";
INPUT N
J=0
K=0
L=0
M=0
R2=0
FOR I=1 TO N
    PRINT "X,Y OF POINT";I;
    INPUT X,Y
    REM ACCUMULATE INTERMED IATE SUMS
    J=J+X
    K=K+Y
    L=L+X^ 2
    M=M+Y^ 2
    R2=R2+X * Y
NEXT I
REM COMPUE CURVE COEFFIC IENT
B=(N * R2-K * J)/(N * L-J^ 2)
A=(K-B * J)/N
PRINT
PRINT"F(X)=";B;" * X";"+";A
REM COMPUTE REGRESSION ANALYSIS
```

```
J=B*(R2-2J*K/N)
M=M-K^2/N
K=M-J
PRINT
R2=J/M
PRINT"COEFFICIENT OF DETERMINATION
(R^2)=";R2
PRINT"COEFFICIENT OF CORRELATION=";SQR(R2)
PRINT"STANDARD ERROR OF ESTIMATE=";SQR(K/(N-2))
PRINT
END
```

(Compaq 386. IBM PT/XT 及其他兼容机上运行通过)

(1) 符号说明：

N 实验数据组数 n，

X 实验数据 x_i

Y 实验数据 y_i

A 回归直线的斜率 a

B 回归直线的截距 b。

(2)使用说明

输入实验点数（数组数）n,再输入实验数据 y_i, x_i，随后便可以计算并输出所回归直线的截距 a 及斜率 b 以及回归方程的相关系数、标准误差等参数。

二、多元线性回归

在数据处理中，还经常遇到受多个变量影响的实验现象和系统。例如在传热过程中，流体在管内强制对流传热系数经验公式，湍流时为：

$$Nu = ARe^m Pr^n = 0.023Re^{0.8}Pr^n \tag{1-41}$$

层流时为：

$$Nu = ARe^m Gr^k Pr^n = 0.74Re^{0.2}(PrGr)^{0.1}Pr^{0.2} \tag{1-42}$$

都是多变量影响的系统，方程中的系数、指数，可以利用最小二乘法求出（将方程取对数化为多元线性方程）。这类问题就是属于多元线性回归问题，多元线性回归的基本原理与方法同前面讲的一元线性回归都是一样的，下面先以二元线性方程的回归为例，说明推导过程和结果，最后推广至多元线性回归。

1. 二元线性回归

设有一批实验数据，$y_i, x_{1i}, x_{2i}, (i=1,2,\cdots,n)$，回归得到的二元线性方程为：

$$y^* = a_0 + a_1x_1 + a_2x_2 \tag{1-43}$$

将实验值与计算值的残差设为 e_i,即有：

$$e_i = y_i - y_i^* = y_i - (a_0 + a_1x_{1i} + a_2x_{2i}) \tag{1-44}$$

根据最小二乘法原理，回归二元线性方程的计算值 y_i^* 与实验值 y_i 的残差的平方和应最