实验设计数据处理与计算机模拟

孙培勤　刘大壮　编著

河南科学技术出版社

内 容 提 要

本书主要涉及实验设计、数据处理和计算机模拟三部分内容；具体章节安排为：误差理论和测定结果表达，统计推断和显著性检验，线性回归，曲线拟合，误差分析和实验设计，单因素及双因素优选法，多因素优选的正交设计法，数学模拟实验，模型判别与序贯实验设计，置信域与统计的实验设计，准确求取反应动力学参数，Monte Carlo 模拟，分形的基础及应用，人工神经网络，Excel、Origin和 Mathcad 软件简介等。

本书适合化学、化工、轻工、材料、环境等相关专业的人员阅读，也可供博士生、工程师和高等院校的教师参考。

图书在版编目（CIP）数据

实验设计数据处理与计算机模拟/孙培勤等编 .—郑州：河南科学技术出版社，2001.7

ISBN 7-5349-2650-5

Ⅰ.实… Ⅱ.孙… Ⅲ.①化学工业-化学实验-设计②化学工业-化学实验-数据处理③化学工业-化学实验-计算机模拟 Ⅳ.TQ016

中国版本图书馆 CIP 数据核字（2001）第 26296 号

责任编辑 韩家显

河南科学技术出版社出版

郑州市经五路 66 号

邮政编码:450002 电话:(0371)5737028

郑州通达印刷厂印刷

全国新华书店发行

开本：850×1168 1/32 印张：7.125 字数：170 千字

2001 年 7 月第 1 版 2001 年 7 月第 1 次印刷

印数:1—3 180

ISBN 7-5349-2650-5/O·9 定价：15.00 元

前　言

化学、化工、轻工、材料、环境等需要实验与观测的多门类学科专业，都有一个共同目标，就是寻找所研究与观测对象的变化规律。通过对规律的研究达到各种实用的目的，如最佳的配方和工艺条件，优异性能的产品，对产品质量、环境质量作出评价等。自然科学和工程技术中所进行的实验，是一种有计划的实践，实验结果会产生大量的观测数据。如何科学地设计实验，对实验所观测的数据进行分析和处理，获得研究观测对象的变化规律，达到各种实用的目的就是本书要讲述的主要内容。所谓规律，一般都具有定性和定量两个方面。本书的目的是为读者提供各种寻找规律的工具，且把重点放在定量的方法上。

作者在多年的教学和科研活动中，在指导硕士生、博士生攻读学位、发表论文的工作中遇到和解决了大量的实验设计和数据处理问题，积累了一些经验。编写本书的意图，是将这些经验介绍给读者，为读者提供一套解决问题的实用的思路和方法，以启发思路，给专科生、本科生、研究生提供最基础的训练和了解新知识、新方法的机会。为了给读者留下深刻、清晰和简洁的解决问题的思路，作者对书的篇幅做了最大程度的压缩，若读者须了解更为深入和具体的内容，可参阅列出的参考文献。本书选择低、中、高三个层次的内容，以满足不同层次读者的需要。第1章至第7章是最基础的部分，包括：误差理论和测定结果表达，统计推断和显著性检验，线性回归，曲线拟合，误差分析和实验

设计，单因素及双因素优选法，多因素优选的正交设计法。第8章至第11章是中级部分，包括：数学模拟实验，模型判别与序贯实验设计，置信域与统计的实验设计，准确求取反应动力学参数。第12章至第14章是提高部分，包括：Monte Carlo模拟，分形的基础及应用，人工神经网络。第15章是常用数据处理软件Excel、Origin和Mathcad软件简介，是学习应用数据处理软件的基础。本书强调实用性、可操作性以及解决问题的思路，简化数学原理的叙述，着重讲清数学公式的具体应用和操作步骤。使读者在学完本书内容之后能独立处理问题，包括常用数据处理软件的使用和进一步在相关理论的基础上建立模型、进行计算机模拟。书中所列举的实例大多是作者处理过的问题。

本书第1章至第3章、第9章至第15章由孙培勤编写，第4章至第8章由刘大壮编写。

本书介绍的部分工作是在河南省模具、材料工程及装备重点学科开放实验室基金资助下完成的。

限于水平，书中可能存在许多不足，甚至错误，欢迎广大读者和同行专家批评指正。

编　者

2001年2月

目　录

第1章

误差理论和测定结果表达

为了定量地研究目标对象，需要采集能反映对象性质的各种观察和测量数据。对于环境保护工作者，有害物质种类和含量是最基础的数据，不论是大气、废水还是废渣中有害物质的含量都要靠测量才能得到。对于化学工艺工作者，温度、压力、浓度、转化率是最基础的数据，也要靠测定才能够得到。对于材料科学工作者，必须了解材料的力学、电学等多方面的性能，这些性能也还是要靠测量才能得到。但是，不论测量工作者如何精心，在采样和分析测试过程中仍不可避免地会产生误差。

各种测定量的大小（真值）是客观存在的，但常常是未知的，只能随着人类认识水平和科学技术水平的提高而逐步逼近于真值。在实际工作中，我们只能用多次测量的平均值代替真值，得到在一定范围内相对准确的结果。要确定这样一个结果，就必须在测定过程中尽量减少误差，并在测量和处理数据中采用数理统计的方法。本章的内容就是在介绍误差理论的基础上，讨论测量结果的正确表达方法及测量值的坏值剔除原则。

1.1 测量误差的分类

误差是测量结果与真值的接近程度，因此，也是测量结果与真值之差。误差按其来源和性质可分为三类。

(1) 系统误差：系统误差是由较确定的原因引起的，对结果的影响较为恒定。如用未经校准的天平称量样品，用未经校核的移液管量取溶液，用未经纯化的试剂进行化学分析等都会产生系统误差。系统误差有一定的方向性，即测量结果总是偏高或偏低，重复测定不能发现和减少系统误差。系统误差，可采用不同的方法校正和消除。

(2) 随机误差：随机误差是由不确定原因引起的。操作者虽然仔细操作，外界条件尽量取得一致，但测得的一系列数据往往仍有差别，且测量值的误差有时正，有时负，有时大，有时小，这是由某些微小的偶然变化因素造成，是不能控制的。如用分析天平称某一试样，多次测量，仍在 0.1mg 上下波动。气流、环境震动、试样暴露在空气中的时间等细微变化都将影响结果。测量次数少，似乎看不出什么规律性，但测量次数多了，就可发现它的统计规律性。增加实验次数可减少随机误差。

(3) 过失误差：过失误差是指一种显然与事实不符的误差，往往由于操作者操作不正确或其他疏忽而引起。例如，器皿不洁净、看错砝码、读错刻度、加错试剂、计算或记录错误等，这些都属于不应有的过失，会对分析结果带来严重影响，必须注意避免，已经发现因上述过失测定的结果，应予剔除。

1.2 随机误差的统计规律性

为了了解随机误差的统计规律，先研究一个实例。

例 1.1 如果让全班同学都从同一瓶溶液中取出样品，各自进行滴定，测出浓度，尽管绝大部分同学测出的数值相差不大，但总不会完全相同。为了研究随机误差的特性，曾对某一吸附残液的 $HgCl_2$ 浓度做了多方面的核对，证明其浓度为 0.804 g/L。这个值，可以作为残液浓度的真值。随后又组织一个班的同学，共进行 120 次滴定，结果见表 1.1。

表 1.1　对 $HgCl_2$ 浓度（g/L）120 次重复测定结果

0.86	0.83	0.77	0.81	0.81	0.80	0.79	0.82
0.82	0.81	0.81	0.87	0.82	0.78	0.80	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	0.79	0.80	0.77
0.76	0.82	0.80	0.82	0.84	0.79	0.90	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	0.80	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
0.80	0.80	0.77	0.81	0.75	0.83	0.90	0.80
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：如果用横坐标表示报告的顺序，纵坐标表示报告结果（浓度），绘出图来（图 1.1），可见多次测量的结果，尽管互不相同，但是，它们都在真值 0.804 附近摆动。说明测量误差主要是随机误差，而不是系统误差。

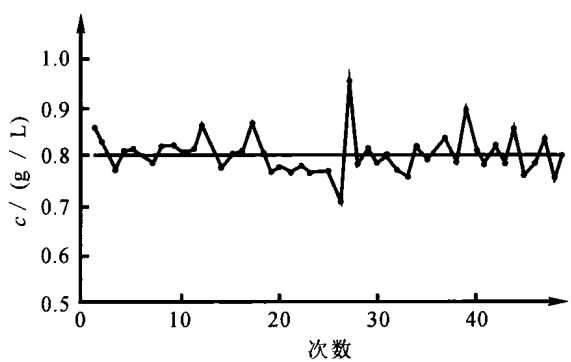

图 1.1　测量值的随机误差

为了更清楚地看出这 120 次结果遵从的分布规律，我们再把

这 120 个数据作成频数分布图或称直方图，确定它的分布密度。为此，先对这 120 个数据由小到大进行分组，为了使每一个数据都能被归并到组内，分组边界值多取一位数字。取起点为 0.635，终点为 0.955，均匀分成 16 组，组距 0.02，结果见表 1.2。

表 1.2　对 120 个实验数据分组结果

分组	频数	分组	频数
0.635～0.655	2	0.795～0.815	24
0.655～0.675	0	0.815～0.835	21
0.675～0.695	0	0.835～0.855	14
0.695～0.715	2	0.855～0.875	6
0.715～0.735	2	0.875～0.895	2
0.735～0.755	8	0.895～0.915	2
0.755～0.775	13	0.915～0.935	0
0.775～0.795	23	0.935～0.955	1

将表 1.2 的数据作出直方图，如图 1.2 所示。

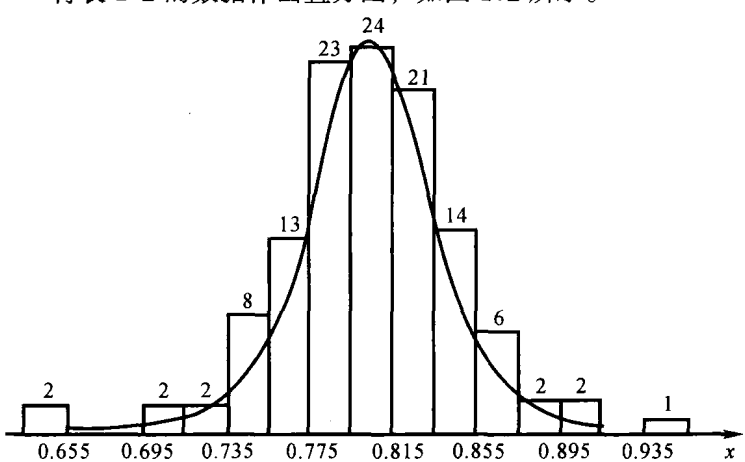

图 1.2　$HgCl_2$ 浓度测量值的直方图

从直方图 1.2 上，我们可以更清楚地看到随机误差的四个特

性：

（1）全部数据中最大值为 0.95，最小值为 0.64。与真值相比，最大负误差约为 −0.16，最大正误差约为 +0.15，120 个数据的误差，都不超过这个界限。误差的这个特性，我们称之为“有界性”。

（2）绝对值小的误差出现的次数多，并集中在中线左右；绝对值大的误差出现的次数少，并分布于左右两侧。这一特性称之为“单峰性”。

（3）绝对值相等的正误差与负误差出现的次数大致相等，这一特性称为“对称性”。

（4）在同一条件下，对同一量进行测量，测量次数增加，随机误差减小。测量次数无限多时，观测值的算术平均值趋于真值，误差平均值的极限为零。这一特性称为随机误差的“补偿性”。

有许多观测对象，它们的真值是无法直接得到的。但是，根据第四条特性，我们可以用全班同学测定的大批数值的算术平均值代替真值。

1.3 正态分布与 t 分布

1.3.1 正态分布

从直方图 1.2 可以看出，测定值的分布曲线大体上是一正态分布。大量的实验证明，随机误差服从正态分布。按照概率论，正态分布密度函数 $p(x)$ 是

$$p(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left[\frac{1}{2\sigma^2}(x-\mu)^2\right] \tag{1.1}$$

式中：参数 μ、σ 是正态分布的数字特征。当 μ 和 σ 值确定之后，$p(x)$ 随 x 的变化曲线就确定了。$p(x)$ 在直角坐标系内的图形，见图 1.3。该线呈钟形，最大点在 $x=\mu$ 处，曲线相对

于直线 $x=\mu$ 对称，在 $x=\mu\pm\sigma$ 处有拐点曲线，以 x 轴为渐进线。

在测量次数很多、测量值误差分布服从正态分布时，其算术平均值 $\bar{x}$ 趋于真值 μ，μ 称为数学期望。σ^2 称为总体方差，σ 称为总体标准差。从图 1.3 可以看出，σ 越小，曲线越陡峭，随机变量 x 离散程度越小。σ 越大，随机变量 x 离散程度越大，曲线分布越宽。相关的计算式如下：

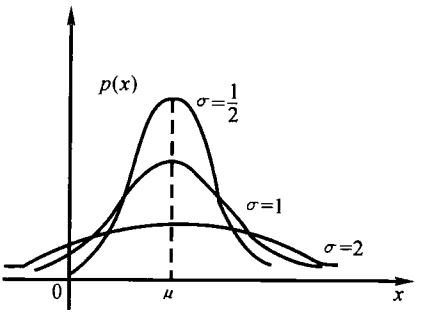

图 1.3　正态分布曲线图

总体平均值：$\bar{x}=\frac{1}{n}\sum_{1}^{n}x_i$　　(1.2)

总体方差：$\sigma^2=\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}$　　(1.3)

总体标准差：$\sigma=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n}}$　　(1.4)

作数学变换，令

$$u=\frac{x-\mu}{\sigma}\tag{1.5}$$

u 仍是正态分布。只是变成了均值为 0、标准差为 1 的标准正态分布。这时，式（1.1）简化为

$$f(u)=\frac{1}{\sqrt{2\pi}}\exp\left(\frac{u^2}{2}\right)\tag{1.6}$$

写出积分式

$$F(u)=\int_{u_1}^{u_2}f(u)\mathrm{d}u\tag{1.7}$$

$F(u)$ 表示 u 值在 u_1 到 u_2 之间出现的概率。若 $u_1=-\infty, u_2=\infty$，按归一化的原则，$F(u)=1$，表示 u 出现在 $\pm\infty$ 区间的概率为 100%。若取 $u_1=-1$，$u_2=+1$，则 $F(u)=68.26\%$。u 在 ± 2 的区间 $F(u)=95.44\%$，$u=\pm 3$，$F(u)=99.74\%$。见图 1.4。

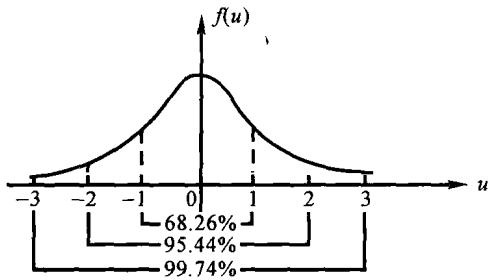

图 1.4　正态分布下的概率

因为，$\mu=\bar{x}$，由式（1.5）得出 $x=\bar{x}\pm u\sigma$。这就是说，测量值 x 出现在 $x=\bar{x}\pm\sigma$ 的区间的概率为 68.26%，不出现在这个范围的概率为 31.74%。同样，在 $\pm 3\sigma$ 的范围内，测量值 x 出现的概率为 99.74%，不出现在这个范围的概率为 0.26%。在统计上，把出现的概率用 $1-\alpha$ 表示，不出现的概率用 α 表示。由于曲线是对称的，左右两边不出现的概率各占 $\alpha/2$。α 也称为检验水平，若 α 取 0.05，就是说，有 95%的把握说明测量值不应该出现在这个区间。每一个 α 值对应一个 $F(u)$ 值，如 $\alpha=0.045\,6$，$u=2$，$\alpha=0.05$，$u=1.96$。

1.3.2　t 分布 小样本

通常在实际的实验室测试工作中，都是小样本实验或小样本监测，测量次数 n 大多小于 30。不符合正态分布适应于大样本的要求。由小样本观测的结果不能代表总体，所以也不能求得总体平均值和总体标准差。这样，以正态分布为基础的统计推断会使实验工作者得出错误的结论，爱尔兰化学家戈塞特（W. S.

Gosstt）首先发现了这个问题。在1908年，他用Student（学生）的笔名发表了一篇论文，题目是“平均值的概率误差”。他一方面从理论考虑，另一方面抽取一些小的随机样本，导出了来自正态分布的小样本平均值的理论分布。这就是在统计检验中应用十分广泛的学生氏 t 分布。

样本方差 s^2 和样本标准差的计算式如下：

样本方差：$$s^2=\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1} \tag{1.8}$$

样本标准差：$$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \tag{1.9}$$

这时，样本均值就是 n 次测量的算术平均值 $\bar{x}$，仍用式（1.2）计算。但是，由于样本较少，不能认为样本均值 $\bar{x}$ 等趋于真值 μ，它与真值可能存在着偏差，偏差值为 $\bar{x}-\mu$。样本方差和样本标准差的计算式（1.8）和（1.9）则与式（1.3）和（1.4）稍有不同，应予注意。在 $n\to\infty$ 时，$s\to\sigma$。

定义统计量：

$$t=\frac{\bar{x}-\mu}{s/\sqrt{n}} \tag{1.10}$$

式（1.10）和式（1.5）对应，只是 x 换为 $\bar{x}$，σ 换为 $s/\sqrt{n}$。

经过一系列的推导，证明 t 分布的密度函数：

$$f(t)=\frac{\Gamma(\frac{n}{2})}{\sqrt{(n-1)\pi}\Gamma(\frac{n-1}{2})}(1+\frac{t^2}{n-1})^{-\frac{n}{2}} \qquad -\infty<t<\infty \tag{1.11}$$

这个随机变量只与样本容量 n 有关。$f=n-1$，称为自由度。当 α 和 f 确定之后，t 分布的临界值可自附表1中查出。不同自由度 f 下的 t 分布曲线，见图1.5。

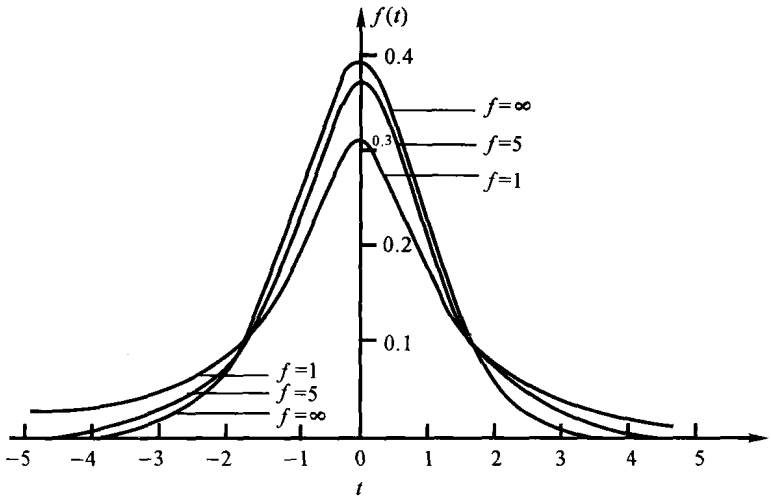

图 1.5　不同自由度下的 t 分布曲线

t 分布的概率密度函数取决于样本的自由度 f 值和 t 值。t 分布是个对称分布，类似于正态分布，对称中线在 $t=0$ 处。曲线的中间比正态分布低，两侧翘得比正态分布略高（图 1.5）。它的形状随自由度而变，当自由度小于 10 时，t 分布曲线与正态分布曲线差别较大。当自由度大于 20 时，t 分布曲线逐渐逼近于正态分布；当自由度趋向无限大时，t 分布曲线就完全成为正态分布曲线。

图 1.5 和图 1.4 是对应的。在标准正态分布中，$\sigma=1$，图中只有一条曲线。但在 t 分布中，$f(t)$ 还与自由度 f 或测量次数 n 有关。随 f 不同，就有不同的曲线，这是二者不同之处。当 f 或 n 确定后，可以作与式（1.7）相似的积分，求出 t 值在指定区间出现的概率。同样，出现的概率用 $1-\alpha$ 表示，不出现的概率用 α 表示。由于曲线是对称的，曲线双方不出现的概率也是各为 $\alpha/2$。