高等学校教材

化学实验数据的统计处理与计算

HUAXUE SHIYAN SHUJU DE TONGJI CHULI YU JISUAN

张卫　编著

化学工业出版社

高等学校教材

化学实验数据的统计处理与计算

HUAXUE SHIYAN SHUJU DE TONGJI CHULI YU JISUAN

张卫 编著

化学工业出版社

·北京·

本书是一本介绍化学实验数据的统计与计算的应用型书籍，具有系统性、实用性强的特点，共5章，分别为：实验数据的误差及数理统计基础、化学中常用的数值分析方法、化学实验的最优设计、数学建模及其在化学中的应用、化学中常用软件的应用。

本书可作为高等院校化学、材料、化工、生命、制药等非计算化学专业高年级学生及研究生了解和掌握化学实验数据的统计和计算等的基本原理及其在该领域的应用基础课程的教材和参考书，也可供信息以及相关专业领域的科技工作者参考。

图书在版编目（CIP）数据

化学实验数据的统计处理与计算/张卫编著. —北京：化学工业出版社，2010.2
高等学校教材
ISBN 978-7-122-07434-8

Ⅰ. 化…　Ⅱ. 张…　Ⅲ. ①化学-实验数据-数据处理-高等学校-教材②化学-实验数据-计算-高等学校-教材　Ⅳ. O6-3

中国版本图书馆CIP数据核字（2009）第239613号

责任编辑：宋林青　　文字编辑：林　媛
责任校对：宋　夏　　装帧设计：史利平

出版发行：化学工业出版社（北京市东城区青年湖南街13号　邮政编码100011）
印　　刷：北京市振南印刷有限责任公司
装　　订：三河市宇新装订厂
720mm×1000mm　1/16　印张10¾　字数226千字　2010年2月北京第1版第1次印刷

购书咨询：010-64518888（传真：010-64519686）　售后服务：010-64518899
网　　址：http://www.cip.com.cn
凡购买本书，如有缺损质量问题，本社销售中心负责调换。

定　　价：25.00元

前　言

随着计算机技术的普及与发展以及计算机与其他学科的交叉、渗透，众多高等院校纷纷开设了计算机在化学中应用类课程，并编写了相应的教材。这些教材从一定程度上体现了计算机在化学方面应用类课程的教学改革中取得的成果，但这类课程整体性不强。对于综合性大学的化学及化学近缘专业的数学以及计算机理论基础相对较强的学生来说，建立计算机在化学中应用类的课程中各学科之间理论知识和计算方法的相互联系，并掌握其综合运用的方法显得尤为重要。

为适应综合性大学化学、应用化学专业及环境等化学近缘专业本科生对计算机在化学中应用类课程教学的要求，本书包含了实验数据的误差及数理统计基础、化学中常用的数值分析方法、实验设计与优化方法、数学建模及其在化学中的应用以及化学中常用软件的应用等基本理论，有较为完整、系统的计算机在化学中应用类课程的知识体系，并通过一定学时的上机训练实践，循序渐进地引导学生根据上机实践去理解其基本原理，了解其在化学、化工、环境等领域的应用，全面提高学生灵活运用理论知识的能力。

在教材的写作过程中，本教材注重将繁复难懂的理论以通俗易懂、简单明了的语言加以阐述，便于学生的理解和运用，为他们后续的学习和科学研究实践打下坚实的基础。

本书共分 5 章，由张卫任主编，在编写过程中，上海交通大学化学化工学院的研究生曾圆、万会师等同学承担了大量的文字编辑、排版等工作，在此表示最诚挚的谢意！另外，上海交通大学化学化工学院的孙淮、李江波、李丰、陈虹锦、马荔等教师对本书的编写也给予了较多的建议和帮助，在此亦表示衷心的感谢！同时，本书的出版得益于上海交通大学教务处、化学化工学院的大力支持，并获得教材出版基金资助，另外，在本书的编写过程中，编者参考了大量的 Internet 上的资源及已出版的相关教科书，并引用了其中的一些图表，在此说明并表示衷心的感谢！

由于编者水平和专业知识所限，书中难免会有疏漏之处，真诚地恳请专家、学者以及广大的读者批评指正。

编　者
2009 年 11 月于上海

目 录

第 1 章
实验数据的误差及数理统计基础

1.1 实验误差概述

实验误差是指测定结果与真实结果之间的差值，是客观存在的，根据其性质及产生的原因，可以分为系统误差和偶然误差两类。系统误差是指在测定过程中由于实验方法本身不够完善、仪器缺陷、试剂不纯等原因所造成的误差，对实验结果的影响比较恒定。重复测定不能发现和减小系统误差，只有改变实验条件才能发现系统误差，通常可以利用对照实验、空白实验、仪器校准等办法加以校正。偶然误差是指在测定过程中一系列有关因素微小的随机波动而形成的具有相互抵偿性的误差，是由测定过程中无法避免的偶然因素引起的，有正有负，可大可小，随着测定次数的增加，大量的偶然误差就表现出一定的统计规律性：绝对值相近而符号相反的误差出现的概率相等，绝对值较小的误差比绝对值较大的误差出现的概率大，在一定的测量条件下，偶然误差的绝对值有一定的限值，即超出该限值的误差出现的概率为零，而且对于同一物理量的等精度观测，其偶然误差的算术平均值，随着观测次数 n 的无限增大而趋于零，符合正态分布。因此，增加测定次数，可以减少偶然误差。

表征实验误差的大小通常通过给出实验数据的准确度与精密度加以判别，准确度是指测量值 x 与真实值 μ 的接近程度，分为绝对误差和相对误差，若无法确定真实值，可用测定值偏离测定的平均值的程度来衡量测定结果的好坏，将对同一样品进行多次重复测定时各测定值相互接近的程度称为精密度，其大小用偏差表征，偏差同样可以用绝对偏差和相对偏差来表示。对于一组实验结果，通常还采用平均偏差、样本标准偏差或总体标准偏差等物理量来表征精密度的大小（注：研究对象全体的集合称为总体，样本是指通过观察或实验得到的数据）。精密度是保证准确度的先决条件，精密度差，所得结果不可靠，但高的精密度也不一定能保证高的准确度。

1.1.1 误差与偏差的基本计算公式

实验数据的统计处理中常用的误差及偏差的数学表达式列于表 1.1.1 中。比较表中相应的数学表达式，可以得出：相对误差能更真实地表征测量值与真实值的接近程度，样本标准偏差比平均偏差能更灵敏地反映出大偏差的存在，能较好地反映

测定结果的精密度。因此，实验测定结果的准确度用相对误差来表征更为确切些，精密度则采用样本标准偏差或总体标准偏差来衡量，若作 n 次测定，其均值 $\bar{x}$ 是真实值 μ 的估计，样本标准偏差 s 为总体标准偏差 σ 的估计，当 n 趋于无穷大时，s 趋近于 σ。

表 1.1.1 常用的误差及偏差的数学表达式

误差及偏差名称	定义及数学表达式	变量说明	备　注
绝对误差	测定值与真实值的接近程度，绝对误差 $=x_i-\mu$	x_i 为第 i 次的测定值，μ 为真实值	
相对误差	绝对误差在真实值中所占的百分率，相对误差 $=\dfrac{x_i-\mu}{\mu}\times 100\%$		
绝对偏差	测定结果与平均值之差，绝对偏差 $=x_i-\bar{x}$，$\bar{x}=\dfrac{\sum_{i=1}^{n}x_i}{n}$	$\bar{x}$ 为 n 次测定结果的平均值	$\bar{x}$ 相当于真实值 μ 的估计值
相对偏差	绝对偏差在平均值中所占的百分率，相对偏差 $=\dfrac{x_i-\bar{x}}{\bar{x}}\times 100\%$		
平均偏差	各测定结果的绝对偏差绝对值的平均值，$\bar{d}=\dfrac{\sum_{i=1}^{n}\lvert x_i-\bar{x}\rvert}{n}$		
相对平均偏差	平均偏差在平均值中所占的百分率，相对平均偏差 $=\dfrac{\bar{d}}{\bar{x}}\times 100\%$		
总体标准偏差	$\sigma=\sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}}$	μ 为无限多次测定的平均值，称为总体平均值。即 $\lim\limits_{n\to\infty}\bar{x}=\mu$，若无系统误差，则为该样本的真实值	适用于测定次数趋于无穷大的测定
样本标准偏差	$s=\sqrt{\dfrac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}}$	$n-1$ 为自由度	适用于有限次数的测定，是总体标准偏差 σ 的估计值，s^2 也称方差
相对标准偏差	样本标准偏差在平均值中所占的百分率，$\mathrm{RSD}=\dfrac{s}{\bar{x}}\times 100\%$		$s/\bar{x}$ 的值也称变异系数(CV)
均值标准偏差	$s_{\bar{x}}=\dfrac{s}{\sqrt{n}}$		

1.1.2 误差的分布

虽然样本标准偏差或总体标准偏差可以表征测定结果相对于平均值的离散程度，但却不能指示这些数据的分布情况。通过观察或实验得到的样本总是由某个具体事物产生并反映该事物的特征，而作为总体的该事物本身可视为所有个体的集合，这是一个随机变量，其分布通常用在一定区域内出现的频率来表示，对于理想

的无穷多次随机测定结果来说，通常用正态分布或高斯分布来描述某样本的总体，相应的函数表达式为：

$$y(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{1.1.1}$$

式中，x 为样本测量值；$y(x)$ 为测量值的概率密度函数；μ 为样本的真实值；σ 为总体标准偏差。正态分布具有如下主要特点：

① 数据关于真实值 μ 呈对称性分布，极大值出现在 μ 处，在 $\mu\pm\sigma$ 处各有一个拐点；

② 真实值 μ 的大小不影响曲线形状，曲线形状主要由总体标准偏差 σ 确定，σ 值越大，数据的离散程度越大，如图 1.1.1 所示；

③ 测定数据 x 出现在确定范围内的概率是有规律的，在 $\mu\pm\sigma$ 的范围内，约为总体的 68.3%；在 $\mu\pm2\sigma$ 的范围内约为总体的 95.5%；在 $\mu\pm3\sigma$ 的范围内约为总体的 99.7%。

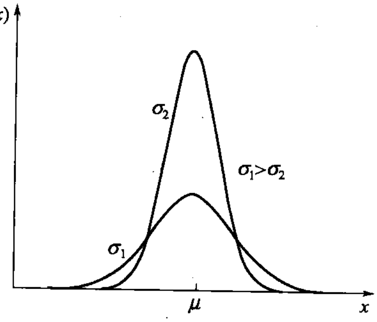

图 1.1.1 真实值 μ 相同，总体标准偏差 σ 不同（$\sigma_1>\sigma_2$）的正态分布图

若令真实值 $\mu=0$，标准偏差 $\sigma=1$，此时的正态分布称为标准正态分布，相应的数学表达式为：

$$y(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}} \tag{1.1.2}$$

对于正态分布，由于其曲线下的面积表示的是测量值出现的概率，因此，若测量数据的分布符合标准正态分布，其概率可以利用 x、μ、σ 值查表获得。

1.2 置信区间及其应用

1.2.1 置信区间的基本概念

根据正态分布的性质，可以定义真实值存在的范围，称为置信区间，置信区间的大小依赖于对该真实值存在范围的指定所具有的确定性，将此确定性称为置信水平。显然，置信水平越高，该确定性所需要的置信区间也就越大。

根据正态分布的特点，对于 n 次采样的均值 $\bar{x}$ 在置信水平为 95%时可能的取值范围是：

$$\mu-1.96(\sigma/\sqrt{n})<\bar{x}<\mu+1.96(\sigma/\sqrt{n}) \tag{1.2.1}$$

式(1.2.1) 可改写为：

$$\bar{x}-1.96(\sigma/\sqrt{n})<\mu<\bar{x}+1.96(\sigma/\sqrt{n}) \tag{1.2.2}$$

式(1.2.2) 给出的是置信水平为 95%时均值 $\bar{x}$ 的置信区间。类似地，可以给出其他

的不同置信水平所对应的置信区间，如99.7%的置信区间由下式给出：

$$\bar{x}-2.97(\sigma/\sqrt{n})<\mu<\bar{x}+2.97(\sigma/\sqrt{n}) \qquad (1.2.3)$$

通常，习惯于采用置信水平为99%的置信区间：

$$\bar{x}-2.58(\sigma/\sqrt{n})<\mu<\bar{x}+2.58(\sigma/\sqrt{n}) \qquad (1.2.4)$$

由于实际上σ的值是未知的，通常当测定次数n足够大（$n>100$）时，在计算中常用标准偏差s代替σ，以确定测量值的置信区间。而当n不够大时，由s代替σ所引进的误差较大，不能忽略，此时测量值的置信区间可用如下方式来计算：

$$\mu=\bar{x}\pm t(s/\sqrt{n}) \qquad (1.2.5)$$

式中，t值依赖于置信水平和自由度的大小。自由度是指计算s过程中独立偏差$(x_i-\bar{x})$所具有的个数，由于对n次测量的样本数据有：$\sum_{i=1}^{n}(x_i-\bar{x})=0$，因此只要有$(n-1)$个$(x_i-\bar{x})$为已知，第$n$个$(x_i-\bar{x})$的值可由上式计算出，故在此种情况下独立偏差$(x_i-\bar{x})$所具有的个数为（$n-1$），即自由度为（$n-1$），以符号$f$表示。

实际上，t值来自于用于统计检验的t分布，其值可以通过查表获得，需要注意的是，置信水平也常用显著性水平来表示，二者关系为：显著性水平＝1－置信水平，如置信水平为95%时，显著性水平为0.05，显著性水平用符号α表示。

对应于t值表数据可以看出，当测量的次数超过100，即自由度超过100后，置信水平为95%（显著性水平0.05）的t值接近于1.96，置信水平为99%（显著性水平0.01）的t值接近于2.58，因此，测定次数$n>100$时，可以用标准偏差s代替σ计算置信区间。

【例1.1】 某实验中采用电化学方法测定尿试样中钠离子的含量，结果（单位：$mmol \cdot dm^{-3}$）分别为：101.5，102.1，98.7，97.6，100.3，99.2，95.4，106.3，试分别计算置信水平95%和99%的置信区间。

解： 自由度$f=8-1=7$，查t值表知，对应于置信水平95%和99%的t值分别为2.365和3.499。8次测量的均值$\bar{x}=100.14 mmol \cdot dm^{-3}$，样本标准偏差$s=3.29$，由式(1.2.5)，得置信水平95%（显著性水平0.05）的置信区间为：

$$\mu=100.14\pm\frac{2.365\times3.29}{\sqrt{8}}=100.14\pm2.75\ (mmol \cdot dm^{-3})$$

置信水平99%（显著性水平0.01）的置信区间为：

$$\mu=100.14\pm\frac{3.499\times3.29}{\sqrt{8}}=100.14\pm4.07\ (mmol \cdot dm^{-3})$$

1.2.2 置信区间的应用

置信区间可以用于检测测量实验中是否有系统误差，当某测量数据的理论值不包含在测量数据在某置信水平所对应的置信区间时，可推断出该测量实验中含有系统误差。

【例1.2】 欲采用分光光度法测定一试样的吸光度，通过标准曲线法确定其浓度，首先需要确定分光光度计是否有仪器误差，已知该试样某浓度标准溶液在

560nm 波长处测定的理论吸光度为 0.580，现在进行 10 次测定，测量均值 $\bar{x}=0.576$，样本标准偏差 $s=0.003$，选取置信水平为 95%（显著性水平 0.05），试通过计算说明该分光光度计是否存在仪器误差。

解：自由度 $f=10-1=9$，查 t 值表知，对应于置信水平 95%的 t 值为 2.262，由式(1.2.5)，得置信水平 95%的置信区间为：

$$\mu=\bar{x}\pm t\frac{s}{\sqrt{n}}$$

$$=0.576\pm\frac{2.262\times0.003}{\sqrt{10}}=0.576\pm0.002$$

由于理论吸光度值 0.580 并不落在所得置信区间范围内，因此，该分光光度计存在仪器误差。

另外，置信区间还可用于试样的测定，如检测某批次食品颗粒的重量及某组成成分的含量，不可能对所有颗粒一一进行称重及组成成分分析，此时，可从大量试样中随机取出部分试样进行称重及成分分析，通过获得的测量数据，计算得到相应的均值和样本标准偏差，从而确定测定量的置信区间。

1.3 误差的传递

1.3.1 偶然误差的传递

若某测量数据 y 是由测定量 a、b、c、d 等组合而成，则偶然误差的传递同其组合方式有关，表 1.3.1 汇总了常见的几种组合方式中偶然误差的传递计算公式。

表 1.3.1 常见的几种测定量组合方式中偶然误差的传递计算公式汇总

组合方式	数学表达式	组合后测量数据 y 的标准偏差计算公式
线性组合	$y=k+k_aa+k_bb+k_cc+\cdots$ 式中，k、k_a、k_b、k_c 等分别为常数	$\sigma_y=\sqrt{(k_a\sigma_a)^2+(k_b\sigma_b)^2+(k_c\sigma_c)^2+\cdots}$ 式中，σ_a、σ_b、σ_c 等分别为测定量 a、b、c 等的总体标准偏差；σ_y 为 y 的总体标准偏差
乘除组合	$y=k\frac{ab}{cd}$，式中 k 为常数	$\frac{\sigma_y}{y}=\sqrt{\left(\frac{\sigma_a}{a}\right)^2+\left(\frac{\sigma_b}{b}\right)^2+\left(\frac{\sigma_c}{c}\right)^2+\left(\frac{\sigma_d}{d}\right)^2}$
乘方组合	$y=b^n$，式中 n 为常数	$\frac{\sigma_y}{y}=\left\lvert\frac{n\sigma_b}{b}\right\rvert$
其他函数	$y=f(b)$	$\sigma_y=\left\lvert\sigma_b\frac{dy}{db}\right\rvert$，式中 $\frac{dy}{db}$ 为该函数 $f(b)$ 关于 b 的一阶导数

从表 1.3.1 各种组合偶然误差的传递计算公式中可以看出，在测定量 a、b、c、d 等的线性组合得到的测量值 y 时，组合 y 的标准偏差大于测定量的标准偏差，但小于各量标准偏差之和。例如利用差量法称取样品时，初始值和终值分别为 22.9634g 和 18.6738g，其标准偏差均为 0.0001×2g，则称取样品的重量及标准偏差分别为：

$$称取样品的重量=22.9634-18.6738=4.2896\ (g)$$

$$标准偏差=\sqrt{(0.0002)^2+(0.0002)^2}=0.000283\ (g)$$

而在乘除组合中，由于采用各测定量 a、b、c、d 等的相对标准偏差的平方来确定组合后测量数据 y 的标准偏差 σ_y，这样，具有最大相对标准偏差的那个分量对于 σ_y 的大小将起主要的决定作用，σ_y 值将略大于该分量。因此，在该种组合的测量中，若希望提高 y 的准确度，应首先设法提高具有最大相对标准偏差的那个测定分量的测试精度。

1.3.2 系统误差的传递

偶然误差的传递同其组合方式有关，类似地，系统误差的传递亦同测定量 a、b、c、d 等组合成测量数据 y 的组合方式有关，表 1.3.2 汇总了常见的几种组合方式中系统误差的传递计算公式。

表 1.3.2 常见的几种测定量组合方式中系统误差的传递计算公式汇总

组合方式	数学表达式	组合后测量数据 y 的标准偏差计算公式
线性组合	$y=k+k_a a+k_b b+k_c c+\cdots$ 式中，k、k_a、k_b、k_c 等分别为常数	$\Delta y=k_a\Delta a+k_b\Delta b+k_c\Delta c+\cdots$，式中 Δa、Δb、Δc 等分别为测定量 a、b、c 等的系统误差 Δy 为 y 中的系统误差
乘除组合	$y=k\dfrac{ab}{cd}$，式中 k 为常数	$\dfrac{\Delta y}{y}=\dfrac{\Delta a}{a}+\dfrac{\Delta b}{b}+\dfrac{\Delta c}{c}+\dfrac{\Delta d}{d}$
乘方组合	$y=b^n$，式中，n 为常数	$\dfrac{\Delta y}{y}=\left\lvert\dfrac{n\Delta b}{b}\right\rvert$
其他函数	$y=f(x)$	$\dfrac{\Delta y}{y}=\left\lvert\Delta x\dfrac{\mathrm{d}y}{\mathrm{d}x}\right\rvert$，式中 $\dfrac{\mathrm{d}y}{\mathrm{d}x}$ 为该函数 $f(x)$ 关于 x 的一阶导数

需要注意的是，在线性组合方式中，由于 y 中的系统误差 Δy 为各测定量 a、b、c、d 等中的系统误差的线性组合，而系统误差有正、有负，因此在某种特殊的情况下，y 中的系统误差 Δy 的值有可能为 0。

1.4 实验数据的统计检验及其应用

实验过程中获得的原始测量数据在实际应用之前，首先需要借助数学方法进行检验，去假存真，使测量数据结果更加准确、可靠。通常采取抽样检验方式对总体的某个或某些特征进行估计，在一定的置信水平（或显著性水平）上，检验其是否符合某种假设的分布规律，若支持该假设时接受它，不支持时则舍弃，这种检验统计量的方法称为统计检验，也称假设检验。统计检验的理论依据是科学实践中广泛采用的小概率原理，即“概率接近零的事件在一次抽样检验中实际上是不可能发生的”。若该事件发生了，则有理由认为原假设是不正确的。

一般情况下，统计检验采用的是概率的反证方法，即先令假设成立，然后依据小概率原理检验结论是否合理，进行反证。其基本步骤为：

① 根据具体问题，提出零假设 H_0 和备择假设 H_1；

② 选取适当的显著性水平，确定拒绝域；

③ 选择合适的检验统计量并确定其分布；

④ 计算样本的统计量值；

⑤ 根据统计量的分布，利用相应的分布表值，通过小概率原理，进行统计推断。

零假设 H_0 的含义是测量数据与真实值相同，需要注意的是，这里二者的相同并不意味着没有偶然误差，而备择假设 H_1 是指测量数据与真实值不一致。若能够证明零假设 H_0 为真，那么备择假设 H_1 就被舍弃，反之，则舍弃零假设 H_0，接受备择假设 H_1。

证明零假设 H_0 是否为真的方法，通常是在假设成立的条件下，采用统计方法计算由偶然误差而引起的测量数据与真实值间差异的概率，概率越小，则零假设 H_0 为真的可能性也越小，当该概率小于选取的显著性水平时，则零假设 H_0 不成立，备择假设 H_1 为真。显然，选取的显著性水平越高，判断的准确度越大。

前已述及，按照概率统计原理，当实验测定次数很多（$n>100$）时，偶然误差遵循正态分布，因此，可以用正态函数来描述其误差分布，并进行假设的检验。但在实际应用中，测量数据的测定次数较少，是小样本实验，不能直接用正态分布函数进行统计分析，只能求得平均值$\bar{x}$及样本标准偏差 s，因此需要引入合适的其他分布函数进行显著性差异的统计检验。小样本统计检验中常用的分布函数主要有 t 分布、F 分布和 χ^2 分布函数，相应的统计检验方法称为 t 检验、F 检验和 χ^2 检验。统计检验的内容主要包括参数估计、离群值检验、平均值检验及方差检验等。

1.4.1 离群值的检验

在对同一实验量进行多次重复测定得到的测量数据中，常常会遇到一组平行测定中有个别数据的精密度不甚高的情况，某一两个测定值比其余的测定值明显偏大或偏小，将这些测定值称为离群值。离群值是测定值随机波动的极度表现，其与平均值之差值是否属于偶然误差是可疑的，需要进行统计检验，判断其波动是否处于合理误差范围内，与其余测定值是否属于同一总体。离群值的取舍会影响测量数据的平均值，尤其当数据少时影响更大。因此在分析计算测量数据前必须对离群值进行合理的取舍，如果统计检验表明离群值确为异常值，才可将其舍弃，若检验表明该离群值不是异常值，即使是极值，也需要将其保留，切不可为了单纯追求实验结果的“一致性”，而将这些数据随意舍弃。当然，对于过失误差，不管其是否为异常值，都应直接舍弃，而不必进行统计检验。常用的离群值的检验方法主要有狄克松（Dixon）检验法及格鲁布斯（Grubbs）检验法。

（1）狄克松（Dixon）检验法

狄克松（Dixon）检验法是 Dixon 于 1951 年提出的，该方法的主要内容为：若将一组测定值从小到大排列得到一个序列，即 x_1，x_2，…，x_n，则可能为异常的离群值必然出现在两端，即 x_1 或 x_n。为检验其是否为异常值，所需要计算的统计量 D 为异常数据与其最邻近数据之间的差同该测定值序列中最大与最小值之差的比值，即

$$D=\frac{x_2-x_1}{x_n-x_1} \tag{1.4.1}$$

或

$$D=\frac{x_n-x_{n-1}}{x_n-x_1} \tag{1.4.2}$$

将计算得到的 D 值同指定显著性水平 α 和测定次数 n 时的临界值 $D_{\alpha,n}$（表 1.4.1）

进行比较，若 $D>D_{\alpha,n}$，则弃去该离群值，否则应予保留。

表 1.4.1　Dixon 单侧检验法的临界值表

测定次数 n	$\alpha=0.1$	$\alpha=0.05$	$\alpha=0.01$	$\alpha=0.005$
3	0.886	0.941	0.988	0.994
4	0.679	0.765	0.889	0.926
5	0.557	0.642	0.780	0.821
6	0.482	0.560	0.698	0.740
7	0.434	0.507	0.637	0.680

【例 1.3】 在一组平行测定中，测得某试样中 Fe 的百分含量（单位：%）分别为 34.38，34.39，34.36，34.40 和 34.44，试分析在指定显著性水平 $\alpha=0.05$ 时，实验数据中有无异常值。

解：首先将数据按递增顺序排列：34.36，34.38，34.39，34.40，34.44，初步判断异常值为 34.44。计算统计量 D：

$$D=\frac{x_n-x_{n-1}}{x_n-x_1}=\frac{34.44-34.40}{34.44-34.36}=0.50$$

由显著性水平 $\alpha=0.05$，查表 1.4.1，$D_{0.05,5}=0.642$，由于 $D<D_{0.05,5}$，34.44 应予保留，此组数据无异常值。

（2）格鲁布斯（Grubbs）检验法

Grubbs 检验法的主要内容为：假设测量数据服从正态分布，将实验测定值从小到大排列：x_1，x_2，…，x_n，离群值在最大值或最小值中，首先计算包括可疑值在内的平均值 $\bar{x}$ 及样本标准偏差 s，为确定异常的离群值，需要计算统计量 G，G 为异常值与 $\bar{x}$ 之差同样本标准偏差 s 的比值，即

$$G=\frac{\bar{x}-x_1}{s} \tag{1.4.3}$$

或

$$G=\frac{x_n-\bar{x}}{s} \tag{1.4.4}$$

根据选定的显著性水平 α 及样本测定次数 n，查 Grubbs 检验法临界值表（表 1.4.2），查出临界值 $G_{\alpha,n}$，将二者进行比较，若 $G>G_{\alpha,n}$，则该离群值为异常值，弃去，否则应予保留。

表 1.4.2　Grubbs 检验法的临界值表

n	α		n	α		n	α	
	0.01	0.05		0.01	0.05		0.01	0.05
3	1.155	1.153	12	2.550	2.285	21	2.912	2.580
4	1.492	1.463	13	2.607	2.331	22	2.939	2.603
5	1.749	1.672	14	2.659	2.371	23	2.963	2.624
6	1.944	1.822	15	2.705	2.409	24	2.987	2.644
7	2.097	1.938	16	2.747	2.443	25	3.009	2.663
8	2.221	2.023	17	2.857	2.475	26	3.029	2.681
9	2.323	2.110	18	2.821	2.504	27	3.049	2.698
10	2.410	2.176	19	2.854	2.532	28	3.068	2.741
11	2.185	2.234	20	2.884	2.557	29	3.085	2.730

【例 1.4】 对某药物中重金属的含量进行测定，12 次测定结果如下（单位：10^{-6} g·dm^{-3}）：0.24，0.26，0.28，0.22，0.23，0.25，0.27，0.31，0.40，0.25，0.27，0.30，试利用 Grubbs 检验法分析在指定显著性水平 $\alpha=0.05$ 时，实验数据中有无异常值。

解：首先将数据按递增顺序排列：0.22，0.23，0.24，0.25，0.25，0.26，0.27，0.27，0.28，0.30，0.31，0.40，初步判断异常值为 0.40。计算平均值$\bar{x}$、样本标准偏差 s 及统计量 G：

$$\bar{x}=0.273,\ s=0.048,\ G=\frac{x_n-\bar{x}}{s}=\frac{0.40-0.273}{0.048}=2.646$$

根据指定的显著性水平 $\alpha=0.05$，查表 1.4.2，$G_{0.05,5}=2.285$，由于 $G>G_{0.05,5}$，数据 0.40 为异常值，应舍弃。

1.4.2 t 检验

t 检验常用于测量数据均值与真实值间的比较以及测量数据均值之间的比较，此外，还可用于检验改变实验条件对结果所产生的影响，本节将简要介绍 t 检验的方法及其应用。

（1）判断测量数据均值$\bar{x}$与真实值间是否有显著性差别

对于实际测量数据来说，总体标准偏差 σ 的值是未知的，1.2.1 节讨论了当测定次数 n 足够大（$n>100$）时，在计算中可以用样本标准偏差 s 代替 σ，而当 n 不够大时，通常采用公式 $\mu=\bar{x}\pm t(s/\sqrt{n})$，来确定测量值的置信区间，因此，可以利用 t 分布的统计量检验小样本的平均值与总体平均值 μ 之间的显著性差异，通常采用单边 t 检验，主要步骤为：

① 提出零假设 H_0 和备择假设 H_1；

② 选取适当的显著性水平，如 $\alpha=0.05$ 或 $\alpha=0.01$；

③ 计算样本的统计量 t 值，由式 $\mu=\bar{x}\pm t(s/\sqrt{n})$推导出 t 值的计算公式：

$$t=\frac{|\bar{x}-\mu|}{s}\sqrt{n} \tag{1.4.5}$$

④ 根据所选显著性水平及自由度 f，查 t 值分布表可得 t 临界值 $t_{\alpha,f}$；

⑤ 将计算出的统计量 t 值同 $t_{\alpha,f}$ 比较，若 $t>t_{\alpha,f}$，则零假设 H_0 不成立，接受备择假设 H_1，否则，接受 H_0。

【例 1.5】 欲采用电化学分析方法测定某糖尿病患者的血糖含量，经 9 次测定，血糖含量的平均结果$\bar{x}$为 7.556mmol·dm^{-3}，样本标准偏差 s 为 0.088，试确定此结果与正常人血糖含量 6.7 mmol·dm^{-3}是否有显著性差别？（指定显著性水平 $\alpha=0.05$）

解：由已知条件，得 $n=9$，自由度 $f=9-1=8$，若指定显著性水平 $\alpha=0.05$，则查 t 分布表，$t_{0.05,8}=2.306$；

零假设 H_0：该患者的血糖含量同正常人无显著性差别；

备择假设 H_1：该患者的血糖含量同正常人有显著性差别。

将各数据代入，计算统计量：$t=\frac{|\bar{x}-\mu|}{s}\sqrt{n}=\frac{|7.556-6.7|}{0.088}\times\sqrt{9}=29.182$

由于计算出的统计量 $t>t_{0.05,8}$，则零假设 H_0 不成立，接受备择假设 H_1，即在该指定的显著性水平 $\alpha=0.05$ 条件下，该患者的血糖含量同正常人有显著性差别。

（2）两组测量数据均值之间的比较

采用两种不同的方法或不同分析人员采用同一方法测定试样，所得到测量数据的均值往往是不同的，用 t 检验可以对两组均值进行检验，比较二者之间是否存在显著性差异。设两组均值分别为 $\bar{x}_1$ 和 $\bar{x}_2$，测定次数分别为 n_1 和 n_2，对应的样本标准偏差分别为 s_1 和 s_2，由于比较的是两个均值，因此，任何一个都不能作为真实值，在这种情况下，假设两组测量数据的总体标准偏差 σ 是相等的，两样本属于同一个总体，两组测量数据的样本标准偏差间的差异是随机的，采用双边 t 检验，主要步骤为：

① 提出零假设 H_0 和备择假设 H_1；

② 选取适当的显著性水平，如 $\alpha=0.05$ 或 $\alpha=0.01$；

③ 计算样本的统计量 t 值，首先计算两组测量数据的合并标准偏差 $\bar{s}$，公式为：

$$\bar{s}=\sqrt{\frac{(n_1-1)s_1{}^2+(n_2-1)s_2{}^2}{n_1+n_2-2}} \tag{1.4.6}$$

t 值的计算公式为：

$$t=\frac{|\overline{x_1}-\overline{x_2}|}{\bar{s}}\sqrt{\frac{n_1 n_2}{n_1+n_2}} \tag{1.4.7}$$

④ 根据所选显著性水平 α 及自由度 $f=n_1+n_2-2$，查 t 值分布表可得 t 临界值 $t_{\alpha,f}$；

⑤ 将计算出的统计量 t 值同 $t_{\alpha,f}$ 比较，若 $t>t_{\alpha,f}$，则零假设 H_0 不成立，接受备择假设 H_1，否则，接受 H_0。

【例 1.6】 采用电化学分析方法测定植物叶片上杀虫剂 DDT 的含量，5 次平行测定未喷洒过杀虫剂的叶片，DDT 的含量均值 $\bar{x}_1$ 为 $0.42\mu g\cdot g^{-1}$，样本标准偏差 $s_1=0.249$。现有一植物叶片样品，5 次平行测定 DDT 的含量均值 $\bar{x}_2$ 为 $0.64\mu g\cdot g^{-1}$，样本标准偏差 $s_2=0.351$，试确定该植物是否喷洒过杀虫剂 DDT？

解：由已知条件，得自由度 $f=n_1+n_2-2=8$，若指定显著性水平 $\alpha=0.05$，则查 t 分布表，$t_{0.05,8}=2.306$；

零假设 H_0：该叶片样品的杀虫剂 DDT 含量同未喷洒过杀虫剂的叶片无显著性差别；

备择假设 H_1：该叶片样品的杀虫剂 DDT 含量同未喷洒过杀虫剂的叶片有显著性差别。

将各数据代入，计算统计量：