邓 勃 编

数理统计方法在化学分析中的应用

化 学 工 业 出 版 社

数理统计方法
在化学分析中的应用

邓 勃 编

化 学 工 业 出 版 社

本书从应用的角度出发，通过对具体例子的分析，介绍了数理统计方法的原理及其在研究分析测试方法和试验数据处理中的应用。全书共分六章：绪论、分析试验数据统计处理的理论基础、化学分析结果的统计检验、方差分析、试验设计、分析结果的表示方法，书末附有必要的数据表。

本书可供从事分析研究工作的技术人员与大专院校化学分析专业的师生阅读。

数理统计方法
在化学分析中的应用
邓　勃　编
*
化学工业出版社出版
（北京和平里七区十六号楼）
化学工业出版社印刷厂印刷
新华书店北京发行所发行
*
开本787×1092 1/32印张5 3/4字数127千字印数1—13,150
1981年1月北京第1版1981年1月北京第1次印刷
统一书号15063·3246定价0.61元

编 者 的 话

数理统计方法在分析化学中得到了日益广泛的应用，它是分析工作者不可缺少的一个有力工具，越来越受到人们的重视。

最近几年，国内已出版了一些介绍数理统计方法方面的书籍和论文，但专门介绍数理统计方法在分析测试方面的应用的书籍还未见到。国内从事分析测试工作的人员很多，在实际工作中，也经常遇到这样或那样的与试验设计和试验数据处理有关的问题。编者认为，要是能有一本这方面通俗易懂的书，供从事分析测试实际工作的同志在遇到有关问题时参考，对这些同志的工作无疑将会有所帮助。出于这种想法，编者虽然深知自己学识和实际工作经验贫乏，但本着向各位前辈和广大群众学习的精神，还是鼓起勇气作一次尝试，承担编写这本小册子的任务。编者衷心地希望这本小册子能起到“抛砖引玉”的作用，在这本小册子出版之后，能见到更多更好的这方面的书籍问世。

在这本小册子的编写过程中，得到了武汉大学数学系胡迪鹤、北京大学化学系郑用熙、北京化工学院工业分析教研室柯以侃等诸位同志的热情帮助，他们对本书提出了许多宝贵的意见和建议。本书能够出版，也和化学工业出版社的编辑同志，以及编者所在单位同志们的鼓励和大力支持分不开的。编者在此一并向所有这些同志致以衷心的谢意。

鉴于编者的学识和实际工作经验有限，这本小册子无论

在内容选材，还是在编排方式和文字表达方面，难免会存在这样或那样的缺点，甚至错误，敬希各位专家和广大读者批评和指正。

编　者

1978年7月

目　录

第一章 绪 论

1.1 化学分析中应用统计方法的意义

分析测试的目的，是从欲测试研究的对象中抽出一部分样品（统计上称为样本）来进行分析测试，并由样本的分析测试结果来推断研究对象全体（统计上称为总体❶）的某个或某些性质。人们知道，由于实际的分析测试工作不可避免地总要或多或少地受到许多未知因素的影响，因此常常使所得到的测定值参差不齐。但是，这些参差不齐的测定值既然是被测试对象某种性质的客观反映，就不会是“杂乱无章”的，其中必寓有反映这种性质的某种客观规律。现在的问题是，分析工作者如何从这些参差不齐甚至表面上看起来近乎“杂乱无章”的测定值中找出其应有的规律性，进而利用这种规律性来指导以后的实践。要做到这一点，除了必需具备本专业必要的理论知识和有一定的实践经验之外，正确地应用科学的试验数据处理方法亦是不可缺少的，数理统计方法正是帮助分析工作者完成这一任务的有力工具。

例如，它可以帮助决定一组测定值中奇异值的取舍问题。在实际工作中，经常见到这种情况：有些分析人员，特别是一些经验不足的分析人员，对一组测定值中偏差较大者，不加仔细地分析以确定该测定值的出现是否合理，就随

---

❶ 总体是指所研究的对象的全体，其中一个单位称为个体，从总体中抽取出来的部分个体的集合体称为样本。

意地加以取舍，这种做法显然是不正确的。

研究和控制某个或某些因素对测试结果的影响，是试验工作中经常遇到的问题。按照过去惯用的试验方法，是将其它因素固定在某个水平上，然后对这些因素逐个地改变和进行研究，以确定其影响的存在与否及其影响的大小。很显然，当欲研究的因素多，而且因素之间存在交互影响时，这种不顾各因素之间相互影响的孤立的试验方法，不仅使工作量大大地增加，而且由此得出的结论亦未必正确，因为在实际上通常总是多种因素的作用及其相互作用同时存在。只有在同时比较了不同条件下的测试结果之后，才能确定某个或某些因素的影响及各因素之间的相互影响，数理统计方法以严密的形式为进行这种比较研究提供了可能性。

数理统计方法还为合理设计和安排试验；准确、简练、明晰地表达分析测试结果；恰当地估计分析测试结果的可信程度，提供了有益的理论指导，使试验研究工作收到多快好省的效果。

应该强调指出的是，数理统计方法虽然是分析工作者解决面临的日益复杂的任务的一个相当有用的工具，但也仅仅是一个工具，它不能代替必要的严格的试验工作，恰恰相反，它只有在可靠的试验基础上才能发挥其应有的作用。

1.2 化学分析中误差的分类

按照误差的性质和产生的原因，可将误差分为三类：偶然误差，亦称随机误差；系统误差，亦称方法误差或固定误差；过失误差。

偶然误差是指在一组测定中，由于各种因素的偶然变动而引起的单次测定值对多次测定平均值的偏差，它决定试验

测定结果的精密度（通常简称精度）。常用标准差来表征。偶然误差的特点是，当测定次数足够多时，出现数值相等、符号相反的偏差的概率近乎相等，各种大小偏差出现的概率遵循着统计分布规律。引起偶然误差的因素是无法控制的。然而，虽然不能找到适当的因数对偶然误差予以校正，但可以通过增加测定次数在某种程度上将它减小。

系统误差是指在一定试验条件下，由某个或某些恒定因素按照确定的一个方向起作用引起的多次测定平均值对真值的偏差，它决定测定结果的准确度，可用绝对误差或相对误差来表示。系统误差的特点是，引起误差的因素在一定条件下是恒定的，误差的符号偏向同一方向，因此，可以按照它作用的规律对它进行校正或设法消除它。增加试验测定次数，不能使系统误差减小。

过失误差是指一种显然与事实不符的误差，主要是由于分析人员的粗心或疏忽而造成的，没有一定的规律可循，但只要分析人员加强工作责任心，这种误差是完全可以避免的。

表征系统误差的准确度与表征偶然误差的精度是性质迥然不同的两回事，决不可以混淆。事实上，在一组测定中，测定精度好，并不意味着准确度也好，反之，精度不好，就不可能有良好的准确度，因为精度是保证准确度的先决条件。对于一个理想的测定结果，既要求精度好，又要求准确度好。精度与准确度的关系如图1-1所示。

就化学分析而言，它是一个复杂的过程，其测定结果受到多种因素的影响。例如对一个已知含量的标准样品，在同一实验室于短期内进行多次重复测定，不仅由于偶然因素的作用出现单次测定值围绕多次测定平均值在一定范围内波动的情况，而且由于恒定因素的作用，使得多次测定平均值与

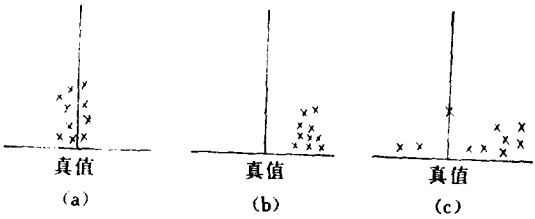

图 1-1 测定精度和准确度相互关系示意图

（a）精度和准确度都好；（b）精度好，准确度不好；（c）精度和准确度均不好

真值之间也存在差值。而当测定是在一个较长时期内进行时，就会产生新的问题，即使在短期内可以认为是不变的因素，现在则可能变为可变因素，比如水和试剂的纯度由于温度和压力的变化、天平砝码的磨损、仪器灵敏度的变化等等，从而使偶然误差增大。如果在不同实验室进行测定，可以想象得到，在一个实验室是不变的因素，而在另一个实验室则完全可能变成可变因素，一般说来，不同实验室对同一样品进行测定，所引进的误差常常是不相同的。鉴于上述这种复杂性，因此，在确定误差性质时，需要特别慎重。

在分析工作中，为了确定一个分析方法的误差大小，常用同类型的但被测组分含量不同的标准样品来进行检验，以确定分析结果与样品真实含量之间彼此符合的程度。对一种标准样品检验如没有系统误差，并不能说明对另一种组成不同的标准样品也无系统误差。人们知道，在重量分析中，由沉淀沾污损失引起的误差，不仅取决于水、试剂的纯度及器皿的表面状况，而且也和被测组分、共存组分的浓度有关。

因此，决不能轻意地用已知校正因数对新的未经检验的样品进行校正。

1.3 精度表示方法

1.3.1 极差

极差是指一组测定值中最大值和最小值之差，表示误差的范围，因此又称为范围误差。

极差虽然反映实际情况的精确性较差，但由于它计算简便，在快速检验中得到广泛的应用。

若所抽取的样本的测定值为x_1、x_2、……，x_n，则极差为

$$R=Max\{x_1, x_2,\cdots\cdots,x_n\}-Min\{x_1,x_2,\cdots\cdots, x_n\}, \tag{1-1}$$

式中$Max\{x_1,x_2,\cdots\cdots,x_n\}$和$Min\{x_1,x_2,\cdots\cdots,x_n\}$分别表示测定值$x_1$、$x_2$、……、$x_n$中最大值和最小值。

1.3.2 算术平均差

算术平均差是单次测定值与多次测定平均值的偏差的绝对值的平均值。

若令各单次测定值为x_1、x_2、……、x_n，测定的算术平均值为$\bar{x}$，测定值与算术平均值的偏差$\delta_i=x_i-\bar{x}$则算术平均差为

$$d=\frac{\sum_{i=1}^{n}|\delta_i|}{n}=\frac{\sum_{i=1}^{n}|x_i-\bar{x}|}{n} \tag{1-2}$$

算术平均差的缺点是，无法表示出各次测定值之间彼此符合的情况，因为有这样一种可能，即在一组测定中偏差彼此接近，而另一组测定中偏差有大有小，然而它们的算术

平均差则完全可能是相同的。例如，甲乙两人对某一样品进行分析，得到的测定值分别为

甲：2.9、2.9、3.0、3.1、3.1

乙：2.8、3.0、3.0、3.0、3.2

甲得到的各次测定值之间彼此符合的程度显然比乙要好一些，但按（1-2）式计算，甲乙两人测定结果的算术平均差却是相同的。

1.3.3 标准离差

标准离差，简称标准差。若令各单次测定值为x_1、x_2、……、x_n，测定的算术平均值为$\bar{x}$，则标准差为

$$S=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \qquad (1-3)$$

标准差是一个表示测定精度的好方法，因为它不仅取决于一组测定中的各个测定值，而且对一组测定中的极值反应比较灵敏。仍以算术平均差一节中提到的例子为例，根据（1-3）式计算，甲乙两人测定结果的标准差分别为0.10和0.14，说明甲的测定结果比乙好，事实上也是如此，然而用算术平均差表示时，则不能区别甲乙两人测定结果的优劣，由此可见，用标准差表示测定精度比用算术平均差表示要更好些。

表示误差时，仅写出误差绝对值大小是不够的，还必须把它与所测定的量联系起来，为此在例行分析中，常用变异系数CV来表示测定的精度

$$CV=\frac{S}{\bar{x}} \qquad (1-4)$$

第二章　分析试验数据统计处理的理论基础

2.1　偶然误差的分布特性

2.1.1　偶然误差的正态分布

在化学分析中，即使是在严格控制的试验条件下，对一个样品进行多次重复测定，由于不可避免的某些偶然因素的作用，各次测定值也并非一定完全相等，而是在一定的范围内波动。表2-1列出了测定某催化剂中碳的数据。这些数据骤然看起来，似乎显得有点“杂乱无章”，然而，如果将这些表面上看来似乎“杂乱无章”的数据进行适当的整理，例如按下列方法进行整理，即将全部测定数据依其大小排列起来，并按一定间隔分成若干组，数出测定值落在每个组的数目（称为频数），于是，便可得到表2-2所示的频数分布表（亦称为分组频数表）。如果以分组为横坐标，相应的频数或相对

表 2-1　某催化剂中碳含量测定值

1.60	1.67	1.67	1.64	1.58	1.64	1.67	1.62	1.57	1.60	1.59	1.64
1.74	1.65	1.64	1.61	1.65	1.69	1.64	1.63	1.65	1.70	1.63	1.62
1.70	1.65	1.68	1.66	1.69	1.70	1.70	1.63	1.67	1.70	1.70	1.63
1.57	1.59	1.62	1.60	1.53	1.56	1.58	1.60	1.58	1.59	1.61	1.62
1.55	1.52	1.49	1.56	1.57	1.61	1.61	1.61	1.50	1.53	1.53	1.59
1.66	1.63	1.54	1.66	1.64	1.64	1.64	1.62	1.62	1.65	1.60	1.63
1.62	1.61	1.65	1.61	1.64	1.63	1.54	1.61	1.60	1.64	1.65	1.59

表 2-2 频数分布表

分　组	频数	相对频数	分　组	频数	相对频数
1.485～1.515	2	0.024	1.635～1.665	20	0.238
1.515～1.545	6	0.071	1.665～1.695	7	0.084
1.545～1.575	6	0.071	1.695～1.725	6	0.071
1.575～1.605	14	0.167	1.725～1.755	1	0.012
1.605～1.635	22	0.262	Σ	84	1.000

频数（频数与样本总数之比）为纵坐标，画成直方图，便可得到如图2-1和图2-2所示的频数分布直方图和相对频数分布直方图。

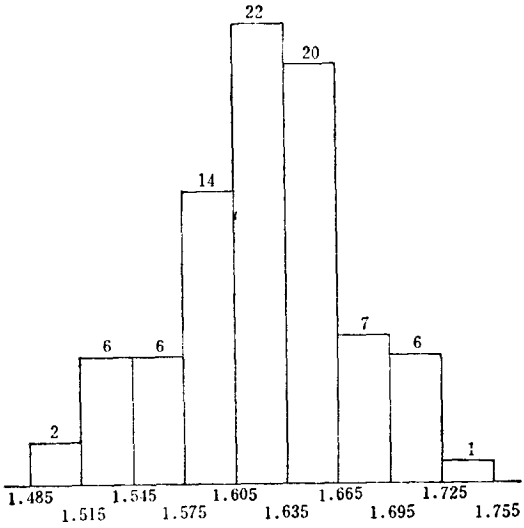

图 2-1 频数分布直方图

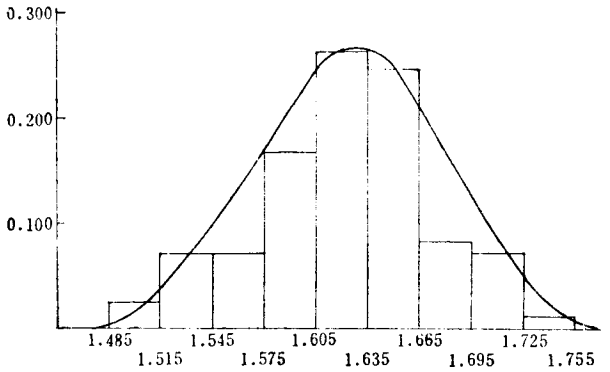

图 2-2　相对频数分布直方图

从经过这样整理后所得到的频数分布表与频数分布直方图以及相对频数分布直方图便可以明显地看到，表 2-1 中的测定数据并不是“杂乱无章”的，而是有其应有的规律性的。虽然不同测定值之间各种大小偏差的出现是彼此独立、互不相关的，但在全部测定数据中，测定值有明显的集中趋势，大多数测定值集中在测定平均值1.620附近，相对于平均值而言，具有各种大小偏差的测定值都有；偏差大小相等、符号相反的测定值出现的次数大致上差不多；偏差小的测定值比偏差较大的测定值出现的次数要多些，偏差大的测定值出现的次数很少。可以想象得到，如果测定数据更多，组分得更细，各组相对频数趋向一个稳定值（这个稳定的比例称为概率），则图2-2相对频数分布直方图逐渐趋于一条曲线，它反映了测定值偶然误差分布的一般状况。当测定值连续变化时，其偶然误差的这种分布特性，在数学上可用一个称之为高斯分布的正态概率密度函数来表示。

$$\varphi(x)=\frac{1}{\sqrt{2\pi}\,\sigma}\,e^{-\frac{(x-\mu)^2}{2\sigma^2}} \qquad (2\text{-}1)$$

式中 x 是从此分布中随机抽取的样本值。μ是相应于正态分布密度曲线最高点的横坐标，称为正态分布的均值，在不存在系统误差的情况下，就是真值，它表示样本值的集中趋势。曲线关于直线$x=\mu$对称。σ是正态分布的标准差，代表从总体均值μ到正态分布曲线两个拐点中任何一个的距离，它表示样本值的离散特性。e＝2.718，是自然对数的底。为简便起见，人们把均值为μ、标准差为σ的正态分布记作$N(\mu,\sigma)$。如果用图形来表示，就得到如图 2-3所示形状的分布曲线，它称为偶然误差正态分布曲线。由图2-3和正态概率密度函数可以看出，集中趋势和离散特性是正态分布的两个基本参数，当给定了均值μ和标准差σ，正态分布就完全被确定了。

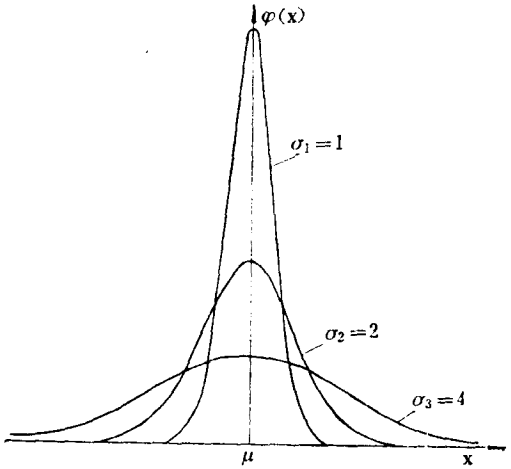

图 2-3　偶然误差正态分布曲线