化学计量学研究方法

卢小泉 陈 晶 周喜斌 编著

科学出版社

国家科学技术学术著作出版基金资助出版

化学计量学研究方法

卢小泉　陈　晶　周喜斌　编著

科学出版社

北京

内 容 简 介

作为一门正在发展的新兴学科，化学计量学主要应用数学、统计学、计算机科学、化学等学科的理论和方法，研究化学量测理论和方法，设计和选择最优的化学量测方法，并通过对化学数据的解析，最大限度地获取有关物质系统的化学信息。

全书共分 12 章，内容包括：数据预处理、线性回归分析、主成分分析、因子分析、偏最小二乘法、小波分析、模式识别、遗传算法、人工神经网络、支持向量机及定量构效活性关系等。

本书可作为化学及相关学科的研究生和高年级本科生的教材和参考书，也可作为化学、化工、工程等领域的科技工作者和高等学校教师的参考书。

图书在版编目 CIP 数据

化学计量学研究方法/卢小泉，陈晶，周喜斌编著. —北京：科学出版社，2013

ISBN 978-7-03-037271-0

Ⅰ. 化… Ⅱ. ①卢…②陈…③周… Ⅲ. 化学计量学-研究方法 Ⅳ. ①06-04

中国版本图书馆 CIP 数据核字(2013)第 069446 号

责任编辑：张 析 / 责任校对：郑金红

责任印制：钱玉芬 / 封面设计：东方人华

科 学 出 版 社 出版

北京东黄城根北街 16 号

邮政编码：100717

http://www.sciencep.com

双 青 印 刷 厂 印刷

科学出版社发行 各地新华书店经销

*

2013 年 4 月第 一 版 开本：B5(720×1000)

2013 年 4 月第一次印刷 印张：20 1/2

字数：400 000

定价：85.00 元

(如有印装质量问题，我社负责调换)

前　言

化学计量学(Chemometrics)作为化学领域中的一个重要交叉学科已在科研领域显示出了强大的生命力。

化学计量学一词最早由瑞典科学家 S. Wold 提出。它运用数学、统计学、计算机科学以及其他相关学科的理论和方法，优化化学量测过程，并从化学量测数据中最大限度地获取有用的化学信息，可以说是一门化学量测的基础理论与方法学。国际化学计量学学会于 1974 年由美国的 B. R. Kowalski 和瑞典的 S. Wold 发起成立；自 1982 年起分析化学(Anal. Chem.)在基础评论中加入了化学计量学专题；*Journal of Chemical Information and Computer Science*，*Journal of Chemometrics*，*Chemometrics Intelligent Laboratory System* 等相继创刊；*Chemometrics* (Matthias Otto. Wiley, John & Sons, 1999 年)、《化学计量学导论》(俞汝勤. 湖南教育出版社，1991 年)、《化学计量学方法》(许禄，邵学广著. 科学出版社，1995 年第一版，2004 年第二版)等书问世，都大大促进了化学计量学学科的发展。目前，化学计量学研究方法已经在数据处理各个相关科研领域得到了广泛应用。

本课题组自 1994 年开始从事化学计量学相关领域的应用研究工作，取得了一些有价值的创新研究成果。本书除了总结数据预处理和平滑方法、线性回归分析方法、主成分分析、因子分析、偏最小二乘法、小波分析、模式识别、遗传算法、人工神经网络和支持向量机等化学计量学研究方法的最新研究成果外，还详细介绍了本课题组近年来在化学计量学方面的研究成果。全书文字力求简洁，部分章节引入相关实例，以使方法易懂。同时，附录列出了化学计量学的一些常用数据信息，以方便读者查阅。

本书由卢小泉教授做整体安排并编写了第 4、6、7、9、10 章，陈晶博士编写了第 1、2、3、5 章，周喜斌博士编写了第 8、11、12 章。全书最后由卢小泉教授修改定稿。

本书获国家科学技术学术著作出版基金资助，书中相关研究成果得到了国家自然科学基金(编号：20927004，21005063，21175108，21165016)等项目的资助。郭晓斌博士对全书的公式作了修正，研究生马玲玲、厦红、张帆、王莉、马良、张丽萍、袁彩霞、冯严俊等在资料收集过程中做了一定工作，在此表示感谢。

书中虽有探索性内容，但因水平有限，缺点在所难免，敬请学界同仁批评指正。

卢小泉

2012 年 5 月于兰州西北师范大学

目　录

第 1 章　误差及数理统计基础

数理统计是以概率论为基础，从实际观测或实验的资料出发，研究随机现象统计规律的数学分支学科，其中心任务是研究如何利用观测的资料来对随机变量的分布函数和数字特征进行估计、分析与推断。它在农业、国防、医药、生物和化学等各个领域的科学研究和实践工作中有着广泛的应用。在化学学科中，数理统计是解决化学量测、质量控制和数据处理最重要的数学工具。研究如何应用数理统计方法解决化学量测中的实际问题，对实验数据进行统计分析，成为化学计量学的基本任务之一。因本书中许多章节的内容涉及数理统计知识，故在第 1 章作简单介绍，以利于后面章节的讨论。

1.1　统计学中的几个基本概念

误差在测量分析过程中是客观存在的。对测量数据的评价，可靠性的判断，数据规律的探索和指导实践等方面，统计学方法都是分析工作者不可缺少的工具。本章将着重介绍统计学的几个常用的重要概念。

1.1.1　随机变量

随机变量体现的是随机事件的结果，而每一个结果的可能性都与一定的概率相对应。研究一个随机变量不仅要知道其取值范围有多大，还要知道取这些值的概率有多大。随机变量建立了样本空间与实数的联系，其取值范围可以表示所有感兴趣的随机事件，由此计算事件的概率便转化为求它的分布函数。随机变量可分为离散型随机变量、连续型随机变量和混合随机变量。理论上常常通过分布函数来引入随机变量，分布函数只是实直线上的函数，同一个分布函数可以联系不同的概率空间，同一概率空间上的不同的随机变量也可以有相同的分布函数，故常用分布函数引入的是分布函数相同的随机变量。

1.1.2　总体与样本

总体在统计学中是研究对象的全体，其中每个单位称为个体。例如，考察 1000 个保温杯的质量，这 1000 个保温杯的全体就是总体，而其中的每一个保温杯都是个体。总体包含的个体数可以是有限的，也可以是无限的。对每个个体来说，它有各个方面的特性，而人们关心的往往只是它的某个（或者某几个）数量指标及

其在总体中的概率分布。例如，在研究一批电脑组成的总体时，可能关心的是电脑显示器寿命的概率分布情况。由于任何一台电脑的显示器寿命事先不能确定，而每一台电脑显示器都确实有一个寿命值，所以可认为电脑显示器寿命是一个随机变量。由此，对总体的研究也就转化为对表示总体的随机变量的统计规律的研究。即总体就是一个具有确定概率分布的随机变量。为了对总体分布进行研究，从总体中抽取若干个个体加以观察和研究的方法叫做抽样法。供研究用的从总体中随机抽取的若干个体称为样本。样本中包含的个体数目 n 称为样本容量。例如在一定条件下，对某矿石中铁的含量作无限次测定，得到无限多个数据的集合，就是总体；如果只作了 6 次平行测定，得到 6 个数据，这组数据就是该矿石总体的一组随机样品。一次抽取的样本是 n（样本容量）个具体数值，称为样本的一个观察值——样本观察。一般不同的抽样就得到不同的样本观察值。样本所有可能取值的全体称为样本空间。一个样本观察值就是样本空间中的一个点。

分析化学中的样本(sample)是指分析实物，即试样；而在统计学中，是指总体中随机抽取的一组测量值。样本需满足以下条件才能很好地反映总体的特性：

① 代表性：相同条件下进行重复抽样得到的每个样本——随机变量，都具有总体的特性；

② 独立性：独立抽样使各个观察之间无相互影响，即各样本间是相互独立的随机变量。

满足上述条件的样本，为简单随机样本。如无特别说明，本书中的样本都是简单随机样本。

1.1.3 统计量

样本是总体的代表和反映，是对总体进行统计分析和推断的依据。但在处理具体的理论和应用问题时，往往样本包含的总体信息比较分散，并不能直接用于解决研究的问题。针对不同的问题构造样本的不同函数，可以把分散在样本中的关于总体的有用信息集中起来，再利用这些函数去推断总体的性质。统计量是样本中不含未知参数的函数，是部分个体样本经过统计计算得到的量。因为样本是随机变量，统计量作为样本的函数，也是随机变量。常用的重要统计量[1]如下：

1. 均值和总体均值

对某试样进行 n 次测定所得 n 个结果 $x_i(i=1,2,\cdots,n)$ 的均值（即算术平均值）为 $\bar{x}$，表达式为

$$\bar{x}=\frac{\sum_{i=1}^{n}x_i}{n} \tag{1.1}$$

若作无限次测定，所得均值 μ 称为总体的均值(也称期望值)。若无系统误差，μ 为真值。事实上不可能进行无限次测定。平均值 $\bar{x}$ 反映测定量的大小和各个测定结果的集中趋势，可作为真值 μ 的最佳估计值。

2. 标准偏差

(1) 标准偏差

定量分析中两组数据的均值可能相同，需要用标准偏差 s 来衡量分析数据的分散程度。

s 的表达式为

$$s=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\bar{x})^2}{n-1}} \tag{1.2}$$

标准偏差可以表征测定结果对于均值的离散程度，却不能表示这些数据的分布情况。数据的分布情况要用直方图(或频谱图)表示。

无限次测量的总体标准偏差用 σ 表示

$$\sigma=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}} \tag{1.3}$$

由此：s 为 σ 的估计。当 n 趋于无穷大时，s 将趋近于 σ。

(2) 相对标准偏差(变异系数)

$$s_r=s/\bar{x}\times 100\% \tag{1.4}$$

(3) 方差

在统计分析中，常常用到样本方差 s^2：

$$s^2=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2 \tag{1.5}$$

(4) 平均值的标准偏差

将一组独立的重复测定值进行平均时，一部分偶然误差相互抵消，使平均值带有的误差比原测定值要小。平均值的标准偏差 $\sigma_{\bar{x}}$ 又称“标准误差”，与单次测量值的 σ_x 之间的关系为

$$\sigma_{\bar{x}}=\sigma_x/\sqrt{n} \tag{1.6}$$

故平均值的标准偏差 $\sigma_{\bar{x}}$ 服从 $N(\mu,\sigma^2/\sqrt{n})$ 正态分布。

对于有限次的测定，上式可改为

$$s_{\bar{x}}=s_x/\sqrt{n} \tag{1.7}$$

标准偏差和标准误差反映了变量值的离散程度。但标准偏差说明观察个体的离散程度，而标准误差说明样本均数的离散程度。样本误差小，说明样本均数和总体均

数比较接近，用样本均数估计总体均数的可靠性大。为了减小标准误差，须减小标准偏差或适当增加测定次数。

3. 样本极差

一组数据中最大值(maximum)与最小值(minimum)之差为极差。用 R 表示：

$$R = x_{\max} - x_{\min} \tag{1.8}$$

4. 其他统计量

这些样本的统计量也是随机变量，但样本的统计量的分布不同于原来随机变量的分布。

(1) k 阶样本原点矩

$$M_k = \frac{\sum_{i=1}^{n} x_i^k}{n}, k = 1,2,\cdots \tag{1.9}$$

(2) k 阶样本中心矩

$$M'_k = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^k}{n}, k = 2,3,\cdots \tag{1.10}$$

(3) 样本中位数

将一组测量数据从小到大排列起来，当测量值的个数 n 是奇数时，排在正中间的数据为中位数；当 n 为偶数时，中间相邻两个测量值的平均值为中位数。

$$\tilde{x} = \begin{cases} x_{m+1}, n = 2m+1 \\ \frac{1}{2}(x_m + x_{m+1}), n = 2m \end{cases} \tag{1.11}$$

中位数与平均值相比较，其优点是受离群值的影响较小，且当 n 很大时，求中位数相对简单，其缺点是不能充分利用数据。样本的数字特征是在统计分析中最有价值的统计量，最常用的是均值和标准偏差。

例 1.1 测定某试样的含氮量，6 次平行测定的结果为 20.48%、20.55%、20.58%、20.60%、20.53%和 20.50%。计算这组数据的平均值、中位数、极差、标准偏差和相对标准偏差。

解：平均值：$\bar{x} = \dfrac{20.48\% + 20.55\% + 20.58\% + 20.60\% + 20.53\% + 20.50\%}{6} = 20.54\%$

中位数：$\tilde{x} = \dfrac{20.53\% + 20.55\%}{2} = 20.54\%$

极差：$R = x_{\max} - x_{\min} = 20.60\% - 20.48\% = 0.12\%$

标准偏差：$s = \sqrt{\dfrac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}}$

$$=\sqrt{\frac{(0.06\%)^2+(0.01\%)^2+(0.04\%)^2+(0.06\%)^2+(0.01\%)^2+(0.04\%)^2}{6-1}}$$

$$=0.05\%$$

相对标准偏差：$s_r = s/\bar{x} \times 100\% = \frac{0.05\%}{20.54\%} \times 100\% = 0.2\%$

1.2　误　　差

实际工作中，对某样品的分析测试不可能绝对准确，即使用最好的方法和仪器去认真分析，多次分析所得数据之间仍会有或大或小的差别。误差是客观存在的，任何一种定量分析的结果，都必然带有不确定度。因此研究工作中需要将实验所得的各组数据用统计方法进行科学的处理，以便对真值作出相对准确的估计。如下内容是对误差相关知识的简单介绍。

1.2.1　误差的定义和表示

真值是试样中某组分客观存在的真实含量。任何测定都有误差，一般难以获得。一般真值有三类：

① 理论真值：如三角形的三个内角和为 180°。

② 约定真值：如由国际原子量委员会讨论修订的原子质量。

③ 相对真值：如一些标准试样中有关成分的含量，以及由有经验的人员采用公认的可靠方法经过多次试验而得出的结果。

误差是指测量值与真值之间的差值。误差的大小是衡量准确度高低的尺度：误差越小，表示分析结果的准确度越高；反之，误差越大，准确度就越低。

绝对误差是指测定值(x)与真实值(μ)之差，即

$$E = x - \mu \tag{1.12}$$

E 为正时是正误差，测定值大于真值，结果偏高；反之，E 为负时是负误差，测定值小于真值，结果偏低。μ 的不确定性决定了 E 是个理想的值，常用指定值或多次测量的算术平均值(即约定真值)作为 μ 的估计值，得到的是 E 的估计值，它的量纲与测量值和真值相对应。

相对误差是指绝对误差在真实值中所占的比率，通常以百分率(%)表示。

$$E_r = \frac{E}{\mu} \times 100\% = \frac{x-\mu}{\mu} \times 100\% \tag{1.13}$$

真值的不确定性决定了相对误差也是理想值，实际得到的是其估计值，没有量纲。

例 1.2　用重量分析法测定纯 $BaCl_2 \cdot 2H_2O$ 试剂中的 Ba 含量，结果为 56.14%，56.16%，56.17%，56.13%，计算测定结果的绝对误差和相对误差。