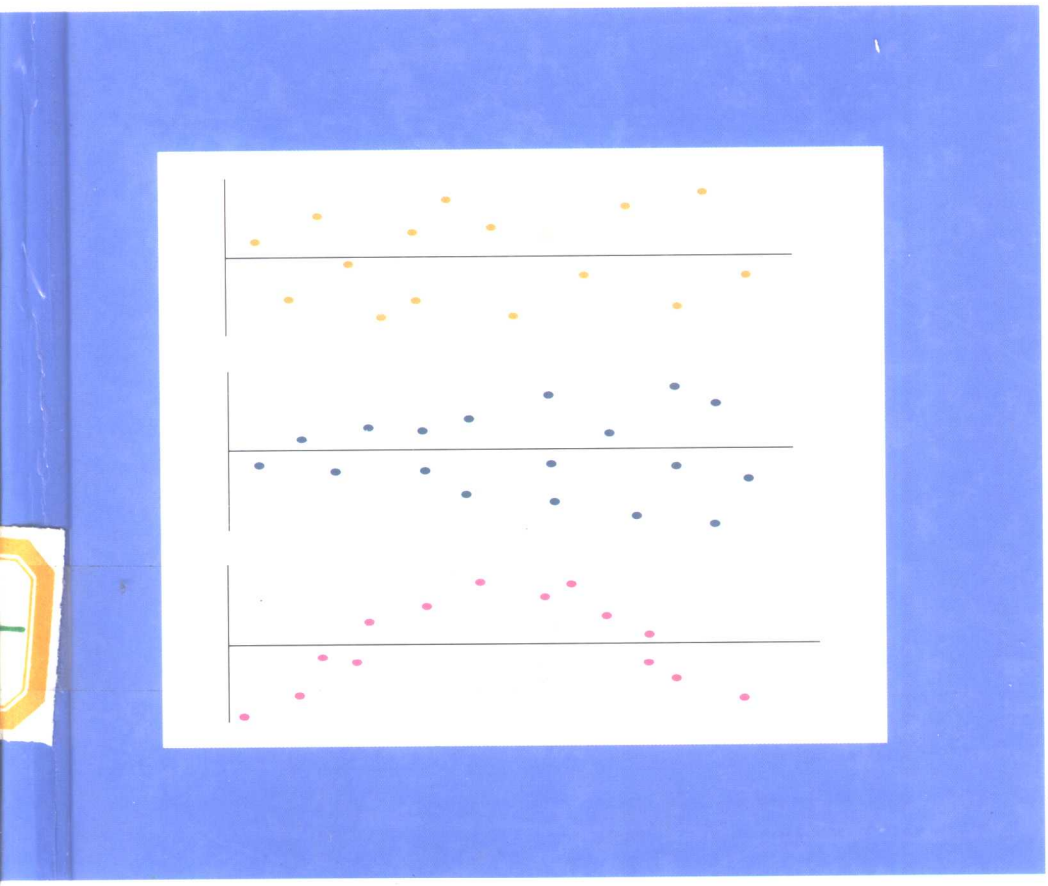
化学统计学

罗　旭　主编



科学出版社

化学统计学

罗　旭　主编

科学出版社
2001

内 容 简 介

本书介绍了化学统计学的初步知识、基础方法、常用方法和几个现代概念。它有以下特点：重视统计学基本原理和基本知识的系统阐述，并给出足够的实例，使读者容易理解和应用；在有关章节用统计学的现代内容补充、更新其古典部分，同时分章论述蒙特卡罗方法、模式识别和人工智能这几个现代统计学概念及其应用；章、节大致按由简到繁，由古典到现代的顺序安排，但各章又有相对的独立性，可根据需要单学。为增强本书的可读性，在初步知识中写了概率知识，在附录中还简要地介绍了矩阵代数初阶和 SAS 系统的知识。

本书可供化学、化工和相关专业的教师、学生和科技工作者使用，其他专业的读者也可参考。

图书在版编目(CIP)数据

化学统计学/罗旭 主编. -北京：科学出版社，2001

ISBN 7-03-009139-6

Ⅰ.化… Ⅱ.罗… Ⅲ.化学-统计学 Ⅳ.O6-04

中国版本图书馆 CIP 数据核字(2001)第 02163 号

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

新蕾印刷厂 印刷

科学出版社发行 各地新华书店经销

*

2001 年 7 月第 一 版 开本：B5(720×1000)

2001 年 7 月第一次印刷 印张：30 3/4

印数：1—3 000 字数：587 000

定价：50.00 元

(如有印装质量问题，我社负责调换〈新欣〉)

《化学统计学》编委会

主　编　　罗　旭

副主编　　徐筱杰　毕开顺

编　者　　（以姓氏笔画为序）

王　玺　毕开顺　乔延江

乔学斌　何春馥　罗　旭

钟大放　侯廷军　赵春杰

徐筱杰　戴荣华

序

为了使科学健康地发展，必须重视科学实验。但在科学实验中观测到的数据不可避免地带有随机误差，增加了对随机事件预测的复杂性。于是，通过对样本数据的搜集、整理、分析和解释以获得关于总体信息和知识的统计学就应运而生，这可上溯至公元前4世纪古希腊的亚里士多德时代。时至能用电子计算机处理或产生数以百万计的数据、新统计方法迭出不穷的当代，统计学的上述内涵定义并无实质性的变化。

《化学统计学》一书，很好地论述了统计学理论，适应了化学学科的特点，并将统计学理论与化学的实际问题紧密地联系起来，使之更容易被理解。此外，它还纳入统计模拟、模式识别、人工智能等技术作为统计学的现代发展，这是本书的主要特色。

该书十分重视阐述的科学性和可读性并强调对统计学基本理论、知识和技能的掌握，必要时补充所需的知识，以大量联系化学实际的例题充分说明，便于读者逐步学习、理解和运用，同时注意更新和创新，兼顾统计学的经典内容和现代发展，并把它们融为一体。该书最后三章介绍了三个现代统计学方法，包括编著者在化学及药物研究领域的应用。在统计蒙特卡罗即蒙特卡罗方法一章中，介绍了他们用伪蒙特卡罗方法制定的计量型制剂均匀度检验新方法。该方法的效率居世界领先地位，已被《中华人民共和国药典》简化后使用。这标志着我国蒙特卡罗方法研究领域已进入其主流，该成果已转化为现实生产力。在模式识别一章中，介绍了他们把模式识别成功地应用于中药质量控制，在继承发扬祖国医药学遗产、实现中医药学现代化所进行的研究工作。在人工智能一章中，介绍了他们提出的基于遗传算法的二维和三维构效关系方法。通过引入遗传算法，很好地解决了二维构效关系模型搭建中分子参数的选取问题以及三维构效关系中活性构象的选择问题。

该书能够满足化学和药学等有关学科的本科生、研究生、教师和科研工作者为提高实验水平学习统计学的需要，对其他学科的读者也有参考价值。《化学统计学》一书的出版，将对上述有关学科的教学及科研起到积极的推动作用。

刘若庄
中国科学院院士
北京师范大学化学系教授
2001年3月25日

前　言

没有只适用于一个学科而不适用于其他学科的统计方法。一个统计方法总能适用于几个相关学科甚至所有学科。但统计方法在应用于某一学科的过程中，总要进行适应而形成一些特点，另一方面该学科的学生和科技工作者在学习统计学时，常由于图书和教学内容理论不能联系实际而感到枯燥和困难。这样，研究数据的收集或产生、描述、分析、综合和解释以获得新化学知识或信息或做出新化学推断的化学统计学(chemstatistics)，就应运而生。化学统计学是合成名词，反映了科学的继承性和交叉学科的发展。

本书共分4篇16章：第一篇3章，论述了化学统计学的初步知识；第二篇4章，介绍了基础统计方法，它们是第三、四篇的基础；第三篇用6章讲述了常用统计方法；第四篇论述了几个现代统计学概念，共3章。本书力求突出以下特点：

(1)重视统计学基本理论和基本知识的阐述，如在第十二章"实验设计与分析之二"中陈述了计算平方和、自由度和期望方差的规则；各章配置了足够的例题，力求理论联系实际，特别是化学实际，使读者容易理解和应用。

(2)兼顾统计学的古典部分及其现代发展。统计学的古典部分也在发展，虽然较慢，但值得重视。方差分析中的显著性过去分为显著和非常显著两个水平，它们是不连续的，现在则由计算机给出实际的显著性水平，称为描述性显著水平；区间估计中的置信区间是众所周知的，但对同样重要的容许区间和预测区间，特别是后者，国内介绍较少；与方差分析同样重要的均值分析也是如此。这类内容，本书都放在第一、二、三篇的有关章、节中陈述。第四篇论述了现代统计学的几个概念，包括蒙特卡罗方法、模式识别和人工智能等3章。蒙特卡罗方法也称统计模拟方法，实际上是由计算机进行的对随机现象的统计模拟，是古典抽样方法的现代发展。其他两章亦有类似之处。这三章分别论述了在该领域概率统计方法的理论基础和应用背景，同时也适当地介绍了编者们独创的领先工作。

(3)章节大致是按由简单到复杂、由古典到现代的顺序安排的，但各章又具有相对的独立性，可以单独阅读。

本书以熟悉微积分的化学和相关学科的读者为主要对象。考虑到高校化学专业教学计划的实际情况，将概率作为统计学的初步知识进行了介绍；此外，由于在第十章回归分析中要用到矩阵和SAS软件系统，本书在附录中加入了"矩阵代数

初阶"和"SAS 系统简介"。希望它们对需要这些知识的读者有所帮助。

本书第十五章由毕开顺执笔，第十六章由徐筱杰执笔，其余部分由罗旭执笔，各位编者对本书都做出了重要贡献。

在编写过程中，本书得到了中国科学院院士陆婉珍同志的支持和胡亚东、薛华等同志的鼓励和帮助，编者向他们表示衷心的感谢。研究生姚美村和李玉娟担任了本书的大部绘图工作，高崇祥同志协助工作，谨致谢意。

由于水平所限，书中的缺点和错误在所难免，殷切希望读者批评指正。

编　者

2000 年 9 月

目　录

第一篇　初步知识

第二篇　基础统计方法

第三篇 基础统计方法

第四篇　几个现代统计学概念

第一篇 初步知识

第一章 绪 论

§1.1 化学统计学的内涵和形成

人类对数据的研究,中外都可追溯到古代。中国大禹治水,依山川性质、人口物产及贡赋多寡分全国为九州。埃及为建筑金字塔对全国人口和财富进行普查。希腊在亚里士多德时代开始以"国事"(matters of state)为研究对象,英文统计学(statistics)一词盖源于此。16 至 17 世纪,统计学开始以概率论为理论基础研究博弈中的概率问题。在此期间,A. de Moivre(1667～1754)发表了一个后世称为正态分布的曲线方程。18 世纪至 19 世纪初叶,P. S. Laplace(1749～1827)和高斯(1777～1855)都使这个极重要分布的研究得到发展。高斯于 19 世纪初叶证明,如果影响一个量的独立随机因素众多而每个因素的影响小,则这个量呈现为概率密度两头小、中间大的分布,并称之为观测误差理论(theory of errors of observation)。19 世纪中叶,A. Quetelet(1796～1874)把统计学应用于包括气象学、人类学、社会学等许多科学领域,并深信政府官员必须接受统计学知识的指导,否则在工作中要发生失误。相关、回归以及相关系数的概念,就是在这个时期形成的。但 20 世纪以前,统计工作都是以收集含有众多个体的大样本和个体遵守正态分布为基础。进入 20 世纪,统计学的一个重要发展是 1908 年 W. S. Gosset(1876～1937)以"学生"(Student)为笔名在 *Biometrica* 上发表的 t 分布,它突破了统计工作必须用大样本方能进行的限制。20 世纪 20 年代,R. A. Fisher 等在小样本理论的基础上建立了显著性检验和方差分析方法。在上述统计理论指导下设计科学实验,称为实验设计。于是统计学发展成一个研究数据的收集或产生、描述、分析、综合和解释,以获得新知识或信息或做出新推断的学科。

20 世纪初叶,统计学适应了一些学科的特点形成一些分支,如生物统计学和经济统计学等。50 年代后,在化学领域出现了一些专著,如 *Statistical Analysis in*

Chemistry and the Chemical Industry(Bennett C A et al. 1954)、*The Application of Mathematical Statistics to Chemical Analysis*(Nalimov V V 1963)、*Statistics for Chemistry*(Youmans H L 1973)、《化学统计学基础》(罗旭 1985)。但直至六七十年代,统计学中早已存在、必须收集大样本才能揭示潜在规律的问题,才在计算机技术和现代数学方法的辅助下迅速得到解决。于是出现了一些新统计方法和相应的软件,如后来也称为计算机统计模拟的蒙特卡罗(Monte Carlo)方法。在这个时期,适应不同学科特点而形成的统计学有了不同的名称,如生物计量学(biometrics, biometry)、经济计量学(econometrics)、技术计量学(technometrics)、化学计量学(chemometrics),同时也出现了一些收载相应论文的期刊,如 *biometrics* (1945), *technometrics* (1959), *Journal of Chemometrics* (1987), *Chemometrics and Intelligent Laboratory Systems* (1987);此外,还出版了不少专著,仅在化学领域就有 *Chemometrics: Theory and Application*(Kowalski B R 1977)、《化学计量学导论》(俞汝勤 1991)、《化学计量学方法》(许禄 1995)等。这一切都酝酿着统计学中一个新兴化学分支学科的形成。

化学计量学、生物计量学、经济计量学分别是由相应的原文直译而来,但它们的内容与计量工作不同,而与统计工作相同或相似。在计算机技术的推动下,上述各学科领域里都提出了一些研究数据的新方法,但其内容并未超出统计学的上述内涵定义,不宜另立学科名称。国外早已有合成名词 biostatistics(生物统计学),近年美国统计学联合会(ASA, American Statistics Association)也用合成名词 chemstatistics(化学统计学)代替 chemometrics 称呼它的一个分部(Brown S D 1995),这些合成名词反映了科学的继承性和形成交叉学科的发展,因而是可取的。这些事实是统计学中这个新兴的化学分支学科宜命名为化学统计学的主要根据。

应该指出,统计方法现已成为科学领域各分支的重要部分。不存在只适用于某一知识领域的特殊统计方法,但有可用于所有知识领域的一般统计方法。为完成某一领域的研究工作而发展起来的统计方法,一定能通过适应,在其他领域找到用场。事实上一些统计方法在几个相关学科的确比在其他学科用得频繁。

§1.2 几个基本的统计学概念

§1.2.1 必然事件与随机事件

人们在实践活动中常遇到必然现象,即在某条件实现后一定发生或一定不发生的事件。例如,"冬去春来","蓝色石蕊试纸遇酸变红","金不生锈"等这样的事件称为必然事件。

但也有在某条件实现后某事件不一定发生的情况。这样的事件称为随机事

件。例如,任意抛掷硬币,落下时国徽不是朝上,就是朝下;对一批针剂进行抽样检查,样本中一支可能是合格的,也可能是不合格的。在前一事件中,掷抛一次的结果是无法预言的,虽然任意一次的结果都是由某种因素所决定,如抛掷的初速、转动的快慢、抛掷角的大小等。但这些因素都不能控制,因而每次抛掷的结果也就无法预言。在后一事件中,一支针剂灌封的质量,包括所含药量、杂质等,受到许多随机因素的影响,不易得到完全控制,因而会是合格的,也会是不合格的,虽然合格的可能性一般要大得多。

§1.2.2 频率与概率

设随机事件 A 在 n 次试验中发生了 n_A 次,则 n_A 称为事件 A 的频数,比值 n_A/n 称为事件 A 的频率或相对频率

$$f(A)=\frac{n_A}{n}=\frac{f_A}{\sum f_i} \tag{1.1}$$

在统计学中也有用 ω 代表频率而把上述 $f(A)$ 写成 $\omega(A)$ 的。但多数文献用 f 代表频率,因为从有无下标和上下文(包括整个公式)不难区别是指频数,还是指频率。

在式(1.1)中 $f(A)$ 代表事件 A 的频率,f_A 代表事件 A 的频数,$\sum f_i$ 代表试验中各类事件的频数之和。显然,任何随机事件的频率都在 0 与 1 之间,即

$$0\leqslant f(A)\leqslant 1 \tag{1.2}$$

对必然事件而言,$n_A=n$,所以频率为 1;对不可能事件而言 $n_A=0$,所以频率为 0。随着某一随机事件重复次数的增多,它的频率就逐渐稳定下来;在重复次数非常多时,频率就逐渐稳定为某一个确定的数值。这个数值就是这一事件出现的概率

$$f(A)=\frac{n_A}{n}\to P(A) \tag{1.3}$$

显然,任何随机事件的概率也是在 0 与 1 之间:

$$0\leqslant P\leqslant 1 \tag{1.4}$$

概率具有非负性;必然事件的概率为 1,不可能事件的概率为 0。这是概率的两个公理。随机事件的频率与事件发生的次数有关,是其概率的随机表现;而随机事件的概率是确定的数值,是随机事件内在规律性的数学量度。表 1.1 描述的是一个抛掷硬币的试验。试验分成 15 组,每组进行 20,200 和 2000 次,国徽朝上的频率依次逼近 0.5000。

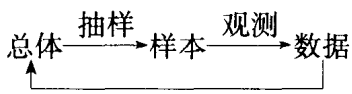
表 1.1 抛掷硬币试验中国徽朝上的频数和频率

	抛掷数					
	20		200		2000	
组	频数	频率	频数	频率	频数	频率
1	14	0.70	104	0.520	1010	0.5050
2	11	0.55	91	0.455	990	0.4950
3	13	0.65	99	0.495	1012	0.5060
4	7	0.35	96	0.480	986	0.4930
5	14	0.70	99	0.495	991	0.4955
6	10	0.50	108	0.540	988	0.4940
7	11	0.55	101	0.505	1004	0.5020
8	6	0.30	101	0.505	1002	0.5010
9	9	0.45	101	0.505	976	0.4880
10	9	0.45	110	0.550	1018	0.5090
11	9	0.45	108	0.540	1021	0.5105
12	6	0.30	103	0.515	1009	0.5045
13	6	0.30	98	0.490	1000	0.5000
14	10	0.50	101	0.505	998	0.4990
15	13	0.65	109	0.545	988	0.4940

随机事件以概率形式表现的规律性称为统计规律性。例如，抛掷硬币落下时国徽朝上的概率是 0.5，就是该随机事件的统计规律性；又如，多次测定一个样品中铁含量 μ 的结果 X 不会是完全相同的，但误差 $\varepsilon = X - \mu$ 具有统计规律性。ε 小于或等于 ΔX 的结果，在全部结果中所占的百分数是相当稳定的，具有一个确定的概率，记作 $P(\varepsilon \leqslant \Delta X)$。

§1.2.3 总体与样本

抛掷硬币国徽朝上的概率为 1/2，是指它的验前概率，即根据硬币的对称性在试验前做出推断。但对多数随机事件出现的概率，在试验前却无法推断。即使两个事件是互不相容的，但却不一定是等可能性的，即概率不一定都是 1/2。由于硬币两面铸造的图形和文字不同等原因，一枚硬币的质量中心与其几何中心可能不重合，于是造成抛掷后国徽朝上、朝下的概率不是 1/2。这样，随机事件的规律性，通常就要通过对它进行非常多次的观测求出验后概率才能发现。然而在实际工作中，只能对随机事件进行次数有限的观测。统计学要解决的正是这个矛盾。统计学的中心任务是统计推断，即通过对事物的局部进行次数有限的观测，获悉统计特性以推断事物整体的统计特性，这个过程可以表示为



研究对象的整体称为统计总体，简称总体。例如，(1)某化肥厂某日生产的全

部硫胺,(2)某制剂厂某年生产的所有维生素 C 针剂,(3)一个化验员为比较两个化验方法的精密度多年进行的化验工作。从总体中随机抽取的一部分个体称为随机样本,简称样本,也叫子样。样本中所含个体的多少称为样本容量或样本大小。例如,表 1.1 中的样本容量依次是 20,200 和 2000。有时允许研究总体中的每一个体,如可以检查所有维生素 C 针剂的澄明度是否合格,即进行全检。但有时不可能研究总体中的每一个体,如一个化肥厂不可能化验其生产的全部硫胺,因为硫胺化验是破坏性的,化验完毕,生产的硫胺也就全部化为乌有。有时总体就是一个样本,例如要研究防锈层金属组成的越王剑。样本可以是事,也可以是物。在样本是物的情况时,常把它叫做样品。

在统计学中,习惯用希腊字母代表总体参数而用英文字母代表样本参数,以示区别。例如,用 μ 代表总体均值而用 $\bar{X}$ 代表样本均值;用 σ^2 代表总体方差而用 S^2 代表样本方差。

§1.2.4 取样的随机性

既然统计学的中心任务是由样本的特征推断总体的特征,样本就必须能够代表总体。这样,统计学对取样自然有一定的要求。这个要求就是随机性,它包括:(1)总体中个体的抽取必须是相互独立的;(2)总体中所有个体被抽取的机会相等。满足以上两个要求的取样,称为简单随机取样(SRS, simple random sampling),这样抽取的样本称为简单随机样本。但由于取样的随机性实际上受到限制,在统计工作中,有时要对取样的随机性进行一定程度的变通。例如,在预测未来事件发生的概率时,统计工作者只能从由全部已发生事件的个体组成的总体中取样,这个总体不可能包括未发生事件的个体,因而取样的随机性受到限制;矿石在地表的分布常是分层的,而各层矿脉有用金属的含量也常不同。这样,为估计其储量所进行的分析工作,就不能采用简单随机取样方法,而应采用分层随机取样方法(stratified random sampling)。这样的变通都是可以理解的。这样的取样仍叫随机取样,但不叫简单随机抽样;这样抽取的样本仍叫随机样本,但不叫简单随机样本。

即使进行随机抽样,根据对样本的观测值推断总体,结论也不可能百分之百正确。但这并不意味着用统计方法所做的结论不可靠。统计方法所做结论的可靠性用概率衡量,称为置信概率(也称置信度或径称信度)。概率小的事件实际上是不可能发生的,可以看成是不可能事件;而概率接近 1 的事件实际上是必然发生的,可以看成是必然事件。概率论中的这个原理对统计推断很重要,称为实际推断原理。

取样不遵行随机性的规定,常是科研工作的结论不正确、成果不能转化为现实生产力的原因。至于在科研工作中对数据采用"合乎我者则取之,不合乎我者则舍

之"的手段,甚至弄虚作假,则已不属于统计学讨论的内容。

§1.3 数据的描述

在§1.2.3中关于"总体与样本"的论述,已涉及数据的收集必须遵行随机性的规定。收集到用各种方法观测的数据样本后,还必须用统计方法加以概括,才能描述或表示总体的特征,并进一步给予正确的解释。表示总体的特征可以用数字,也可用图形。当然,数形合一也是可以做到的。

§1.3.1 量度数据集中位置的统计量

常用两类特征统计量表示数据:一类是量度数据的集中位置;另一类是量度数据的离散程度。有时把两类统计量联合在一起成为一个综合统计量,这更能表示数据的特征。

数据的集中位置可用均值、众数和中位数量度,其中最重要的是算术平均值,简称均值。

1. 均值 均值通常指样本均值,是全部观测值之和除以观测次数所得的商

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \tag{1.5}$$

在不会发生误会的情况下,可把$\sum_{i=1}^{n} X_i$简写为$\sum_i X$或$\sum X$。求均值的目的,是使多个观测值中的正负随机误差相互抵偿而减免;在无系统误差的条件下,观测重复的次数愈多,这种抵偿愈充分,样本均值$\bar{X}$愈接近总体均值μ。总体均值μ也称为观测值X的期望或数学期望E(expectation),记作

$$\mu = E(X) = \sum \text{取值} \cdot \text{概率} = \sum_i X_i P_i \tag{1.6}$$

均值有两个重要性质:(1)观测值与均值之差即偏差的和为零;(2)观测值与均值之差即偏差的平方和最小。这不难用简单的代数运算证明:

$$\begin{aligned}\sum (X - \bar{X}) &= (X_1 - \bar{X}) + (X_2 - \bar{X}) + \cdots + (X_n - X\bar{X}) \\ &= (X_1 + X_2 + \cdots + X_n) - n\bar{X} \\ &= n\bar{X} - n\bar{X} = 0\end{aligned}$$

例 1.1 5个氖灯的寿命分别是1357,1090,1666,1494,1623 h,平均寿命是1446 h,偏差之和为零。

观测值与除均值外的任何值a之差的平方和