中国刑事警察学院专著类创新教材

实验设计与优化

SHIYAN SHEJI YU YOUHUA

张振宇 编著

群众出版社

中国刑事警察学院专著类创新教材

实验设计与优化

SHIYAN SHEJI YU YOUHUA

张振宇 编著

群众出版社
·北京·

图书在版编目（CIP）数据

实验设计与优化／张振宇编著. —北京：群众出版社，2016. 3
中国刑事警察学院系列教材
ISBN 978－7－5014－5507－2

Ⅰ. ①实… Ⅱ. ①张… Ⅲ. ①试验设计—高等学校—教材
Ⅳ. ①O212. 6

中国版本图书馆 CIP 数据核字（2016）第 063217 号

实验设计与优化
张振宇　编著

出版发行：群众出版社
地　　址：北京市西城区木樨地南里
邮政编码：100038
印　　刷：北京泰锐印刷有限责任公司

版　　次：2016 年 5 月第 1 版
印　　次：2016 年 5 月第 1 次
印　　张：4.875
开　　本：880 毫米×1230 毫米　1/32
字　　数：130 千字

书　　号：ISBN 978－7－5014－5507－2
定　　价：18. 00 元

网　　址：www. cppsup. com. cn　　www. porclub. com. cn
电子邮箱：zbs@ cppsup. com　　zbs@ cppsu. edu. cn

营销中心电话：010－83903254
读者服务部电话（门市）：010－83903257
警官读者俱乐部电话（网购、邮购）：010－83903253
教材分社电话：010－83903259

前　言

分析化学是化学的一个重要分支，它的基本任务是鉴定物质的化学成分和化学结构以及测定有关成分的含量。一方面，分析化学离不开分析测试，分析测试会产生大量实验数据，而实验数据又都存在误差。如何处理这些实验数据，最大限度地提取其中相关的化学信息，从而得出正确的实验结论，是分析化学工作者必须面对的一个问题。另一方面，所有的分析测试都是在一定的实验条件下进行的，实验条件的选择对于能否得到正确的实验结果尤为重要。如何设计实验过程，优化实验条件，是分析化学工作者要面对的另一个问题。为了解决上述问题，就要用到误差理论和统计学的相关知识。将统计学应用到分析化学领域，便形成了一个交叉学科——化学统计学，进而又发展成化学计量学。

在传统的化学研究中，人们能够使用的化学测试技术有限，对实验数据处理的要求不高，实验数据包含的信息量也较低。电子技术、计算机技术及现代仪器制造技术的发展，使现代分析仪器能提供大量多维数据，因而对实验数据的处理和解析提出了更高的要求，计算机的广泛应用使这些任务的完成成为可能。化学计量学正是以化学实验设计与优化、数据处理、信息提取及结果解释为主要研究内容的一门新兴的化学分支学科。

本书只涉及化学计量学中的部分内容，包括统计学基础知识、假设检验、方差分析、回归分析、正交设计、均匀设计、实验的最优设计等，可作为高等公安院校分析化学专业和刑事科学技术专业研究生相关课程的教材使用。

本教材已在中国刑警学院分析化学专业研究生教学中试用多年。在正式出版之际，承蒙中国刑警学院数学教研室王庆丰教授

审阅，提出了许多宝贵的修改意见，在此谨表谢意。限于编者的学识和能力，书中的错误和不足之处在所难免，诚恳希望读者批评指正。

编著者

2015 年 10 月

目　录

第一章　统计学基本概念

统计学是数学的一个分支学科，在社会科学和自然科学中都有广泛的应用。

化学是一门实验科学。对于一个化学问题，化学工作者首先要做出一个假设，然后确定该做哪些实验来验证假设的有效性。实验产生大量数据，化学工作者要利用这些数据提取尽可能多的信息。最后，化学工作者应用这些信息和化学知识给出关于研究系统的更多的知识。步骤如下：

提出假设 → 设计实验方案 → 分析处理数据 → 得出实验结论

把统计学原理和方法应用到化学实验过程，分析处理实验数据，获取相关化学信息，便形成了化学的一个分支领域——化学统计学。在传统的化学研究中，人们能够使用的化学测试技术有限，对实验数据处理的要求不高，而且实验方法提供数据的能力不强，数据包含的信息量较少。随着电子技术、计算机技术和现代仪器制造技术的不断发展，现代分析仪器能够快速地提供大量多维数据，因而对实验数据的处理、分类、解析和预测提出了更高的要求。将统计学、数学、逻辑学以及计算机科学应用到化学实验过程，又形成了一个比化学统计学涵盖内容更全面、更广泛的化学分支学科——化学计量学。它以研究化学过程中的方法选择、数据处理、信息提取及结果解释为主要内容，具体包括分析采样理论、实验设计与优化、分析信号检测与处理、化学校正技术、化学模式识别、人工神经网络及遗传算法等。

一、实验数据的随机性

人们对客观事物或现象的定性或定量描述被称为数据，它又分为软数据和硬数据。软数据是人们对客观事物或现象的定性描述，如水是“热”的，压力是“大”的等；硬数据是人们对客观事物或现象的定量描述，如水的温度是90℃，气体的压力是101325Pa等。本书中讨论的数据都是硬数据。

实验数据在一定程度上总是存在不确定性，称为实验误差。误差的来源有两个方面：一方面来源于系统误差，另一方面来源于偶然误差。系统误差是实验误差的恒定部分，它使实验数据偏向某一侧，不能通过重复实验来减小或消除，但可通过标准物质对测定进行校正来减小或消除。偶然误差是由实验过程中的随机因素造成的，不能通过用标准物质校正来减小或消除，但可通过增加测定次数来减小或消除。

系统误差和偶然误差总是交织在一起，即使通过某种措施消除了系统误差，偶然误差仍是不可避免的，从而使实验测定的数据成为一个随机变量，而统计学正是研究随机变量分布规律的一门科学。

二、总体和样本

1. 总体

我们把所研究的同质的个体构成的整体称为总体。例如，要研究东北地区成年人血液中红细胞的正常值，那么东北地区全体成年人血液中的红细胞值就是总体。

总体的数量一般是巨大的，有时甚至是无穷的。要研究东北地区成年人血液中红细胞的正常值，我们不可能对东北地区所有成年人都进行血液检测，而只能按照一定的科学方法从中抽取一部分人进行血液检测，然后根据检测结果来推断东北地区成年人血液中红细胞的正常值。

2. 抽样与样本

从总体中抽取部分个体的过程叫作抽样。抽样时从总体中抽取的个体称为样本。抽样时抽取的个体数目称为样本含量。因此，样本含量就是样本中含有的个体数目。

从总体中抽取样本时，必须遵循一定的科学原则。一般来说，一个样本应当具有代表性、可靠性和随机性，这称为样本的“三性”。

代表性是指样本中的每一个个体必须符合总体的规定。这就要求对总体有一个明确的规定，这种规定是根据研究的目的而确定的。例如，要研究东北地区成年人血液中红细胞的正常值，我们可以规定研究对象为体温正常、无急性病、无血液消耗性疾病的东北地区成年人。那么，所有被抽取的对象都必须符合上述规定。

可靠性是指由样本的实验结果对总体所作出的推论要有较大的可信度。由于个体之间存在差异，样本的含量越小，对总体所作出的推断可靠性就越差，但样本含量增大，实验所耗的人力和物力都会加大。这是一对矛盾。

随机性是指总体中每个个体都有相同的概率被抽作样本。必须指出，随机性并不等于随意性。例如，欲将 40 只小鼠随机分为两组，如果闭上眼睛随意抓取 20 只作为第一组，剩余的作为第二组，表面看来是随机的，实际上是不随机的。因为体壮和活泼性强的小鼠难以抓到，故大部分留在了第二组，所以两组的活泼性有较大差异。为了保证抽样的随机性，可把 40 只小鼠随机编上 40 个号码，作出标签，将标签混匀，再随机抽取标签。

三、算术平均值与几何平均值

平均值是描述一组测定值（或叫作观察值）集中位置或平均水平的统计指标，常作为一组数据的代用值用于统计分析中。下面我们用 x_1、x_2、$\cdots x_n$ 表示一组测定值，n 是测定次数（样本

含量）。

1. 算术平均值

将一组数据的各测定值相加，再除以测定次数，就能得到这组数据的算术平均值，用符号$\bar{x}$表示。它是统计学中最基本和最常用的一种平均指标，主要适用于数值型数据。算术平均值的计算公式为：

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1-1)$$

2. 几何平均值

几何平均值是指 n 个测定值连乘积的 n 次方根，用符号 G 表示，计算公式为：

$$G = \sqrt[n]{x_1 x_2 \cdots x_n} \qquad (1-2)$$

为计算方便，常用对数计算几何平均值：

$$\lg G = \frac{1}{n}(\lg x_1 + \lg x_2 + \cdots + \lg x_n) \qquad (1-3)$$

几何平均值常用于反映一组经对数转换后呈对称分布的变量值在数量上的平均水平，如医学研究中的免疫学指标，而化学实验中一般很少用到。

四、偏差

测定值与平均值之间的差值叫作偏差，用 d_i 表示：

$$d_i = x_i - \bar{x} \qquad (1-4)$$

偏差是衡量实验结果精密度的指标。为充分利用每一个测定值，一个很自然的想法是计算各测定值的平均偏差。

1. 平均偏差

平均偏差用符号 $\bar{d}$ 表示，计算公式为：

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} |d_i| = \frac{1}{n}\sum_{i=1}^{n} |x_i - \bar{x}| \qquad (1-5)$$

加入绝对值符号的目的是避免正、负偏差相互抵消。平均偏

差比较直观地反映出一组实验结果的精密度，但由于使用了绝对值符号，不便于作进一步的数学处理。

2. 离均差平方和

为了克服平均偏差使用绝对值不便于进一步运算的缺点，可以不采用取绝对值的方法，而是通过取平方的方法避免正、负偏差相互抵消，为此引入“离均差平方和”的概念。所谓离均差平方和，就是先将各测定值的偏差取平方然后再加和，用符号 SS 表示，计算公式为：

$$
\begin{aligned}
SS &= \sum_{i=1}^{n}(x_i - \bar{x})^2 \\
&= \sum_{i=1}^{n}(x_i^2 - 2x_i\bar{x} + \bar{x}^2) \\
&= \sum_{i=1}^{n}x_i^2 - 2\bar{x}\sum_{i=1}^{n}x_i + n\bar{x}^2 \\
&= \sum_{i=1}^{n}x_i^2 - 2n\bar{x}^2 + n\bar{x}^2 \\
&= \sum_{i=1}^{n}x_i^2 - n\bar{x}^2 \\
&= \sum_{i=1}^{n}x_i^2 - \frac{1}{n}\left(\sum_{i=1}^{n}x_i\right)^2 \qquad (1-6)
\end{aligned}
$$

3. 方差

离均差平方的平均值叫作方差，用 S^2 表示，计算公式为：

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \qquad (1-7)$$

式中，$n-1$ 称为自由度，意为在 n 个离均差平方项中，由于受到样本均值的限制，只有 $n-1$ 个是独立的。

4. 标准偏差

使用方差来表示测定值的精密度，一个不方便之处是方差的单位与测定值的单位不一致，为此使用方差的平方根，称为标准偏差，用 S 表示，计算公式为：

$$S=\sqrt{S^2}=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad (1-8)$$

标准偏差常称为标准差。

5. 相对标准偏差

标准偏差的单位与原始数据相同，因此在比较两组数值相近、单位相同的数据的精密度时，使用标准偏差是方便的。但是，如果两组数据的均值相差较大或单位不同时，就不方便了。为此，引入相对标准偏差（CV），计算公式为：

$$CV=\frac{S}{\bar{x}}\times 100\% \qquad (1-9)$$

相对标准偏差也称为变异系数。

例 1 用气相色谱法（GC）和高效液相色谱法（HPLC）分别测得某人尿液中三唑仑含量（ng/cm^3），列于下表：

实验编号	1	2	3	4	5	均值
GC 法	10.36	10.47	12.45	11.69	12.75	11.54
HPLC 法	9.86	10.43	10.26	9.97	11.33	10.37

求两种方法的平均偏差、离均差平方和、方差、标准偏差和相对标准偏差。

解：

（1）气相色谱法

$$\bar{d}=\frac{1}{n}\sum_{i=1}^{n}|x_i-\bar{x}|$$
$$=\frac{1.18+1.07+0.91+0.15+1.21}{5}$$
$$=0.904(ng/cm^3)$$

$$SS=1.18^2+1.07^2+0.91^2+0.15^2+1.21^2=4.852$$

$$S^2=\frac{4.852}{5-1}=1.213$$

$$S = \sqrt{1.213} = 1.10(ng/cm^3)$$

$$CV = \frac{1.10}{11.54} \times 100\% = 9.54\%$$

(2) 高效液相色谱法

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

$$= \frac{0.51 + 0.06 + 0.11 + 0.4 + 0.96}{5}$$

$$= 0.408(ng/cm^3)$$

$$SS = 0.51^2 + 0.06^2 + 0.11^2 + 0.4^2 + 0.96^2 = 1.357$$

$$S^2 = \frac{1.357}{5-1} = 0.339$$

$$S = \sqrt{0.339} = 0.583(ng/cm^3)$$

$$CV = \frac{0.583}{10.37} \times 100\% = 5.6\%$$

第二章　随机变量的分布

第一节　正态分布

样本中每个个体的测定值（观察值）实际上是一些随机数。随机数的分布通常用其在一定区域出现的频率来表示。

表 2－1 给出了某溶液在某波长处的吸光度的多次测定结果。

表 2－1　某溶液在某波长处的吸光度值

测量序号	测量值 x	测量序号	测量值 x
1	0.3410	8	0.3460
2	0.3350	9	0.3430
3	0.3470	10	0.3420
4	0.3590	11	0.3560
5	0.3530	12	0.3500
6	0.3460	13	0.3630
7	0.3470	14	0.3530

测得的吸光度分布在 0.33～0.37 范围内，若将其划分为 12 个区域，各测量值在每一区域出现的次数和频率见表 2－2，用图表示即为图 2－1。

图 2－1 中纵坐标的意义是：测定值出现在某一区域内的次数在全部测定次数中所占的比例，即单位区域内某一测定值出现的几率（频率），通常称为几率密度。当测定次数非常多（样本含量非常大），区间间隔划分得非常小时，图 2－1 就近似成为一个正态分布。