HUAN
JING
JIAN
CE
CHANG
YONG
SHU
LI
TONG
JI
FANG
FA

环境监测常用数理统计方法

四川省环境科学学会编

四川科学技术出版社

环境监测
常用数理统计方法

四川省环境科学学会

四川科学技术出版社

一九八三年·成都

责任编辑：崔泽海、罗孝昌
封面设计：

环境监测常用数理统计方法

四川科学技术出版社出版 （成都盐道街三号）
四川省新华书店发行 资阳县印刷厂印刷

开本850×1168毫米 1/32 印张6.5 字数152千
1983年11月第一版 1983年11月第一次印刷
印数：1—6,900册

书号：13298·6 定价：0.95元

前　言

环境科学是一门综合性的边缘科学。环境监测数理统计是为环境科研、设计方案、生产实践提供可靠依据的重要手段；它为分析资料、处理、判断试验数据，科学地安排试验，制定试样检验方案提供了途径。已日益为国内外科学工作者及生产部门所重视，迅速得到普遍推广和广泛运用。

环境监测数据与其它试验数据比较，它要求监测数据具有更充分的可比性，不但需要对实验室内和实验室间的数据进行比较，还要进行区域性甚至全球性数据间的比较。这就要求为了保证做好环境监测质量工作，各方面都必须提供准确一致的监测数据。如何应用数理统计方法，处理和判断环境监测数据是环境科技工作者十分关注的问题。

我会在有关人员的迫切要求下，聘请成都科技大学讲师黄文辉和四川省环保科研所工程师林兆升担任主讲，于1982年6月份，举办了“环境监测常用数理统计方法”培训班。本书是在培训班讲义的基础上，吸取了各方面的意见进一步修改补充后写成的。书中还结合作者在环境监测的科研实践和教学中的经验，用具体的例子介绍了数理统计方法。每种方法着重介绍直观的原理和最常用的一些数理统计方法，而不作数学上的严格推导，通俗易懂，以便读后能够基本掌握和实际运用，以期能够对有关人员正确处理分析监测数据提供一些帮助。对于想知道数学原理或要解决更深一步问题的读者，书末推荐了有关参考资料。

本书编写前后经历了3—4个月时间。在此，我们谨向编著者

黄文辉讲师、林兆升工程师，参加者吴波同志，审阅者高华寿教授、谢清成副总工程师，以及对本书的编辑出版给予积极支持和鼓励的高步仁、吴家兴、吴鹏鸣、寇洪如、李献文、洪水皆、郗昌杰、刘殿生、任耐安、白大森、陈镇华、黄纪猷、涂金全、丁国斌、林云开、刘香玲、桂先伦、郑宋、刘树刚、张文选、康人桂、周顶昌、李延嗣等同志致以衷心的感谢。

由于作者水平有限，时间匆忙，书中缺点、错误之处，恳望读者批评指正。

四川省环境科学学会

一九八三年三月

说　明

承四川省环境科学学会的委托，编写了这本书，供从事环境监测的同志参考。此书也可供分析研究工作者及高等学校分析化学专业师生参考。

本书主要内容包括：误差及数据处理的理论基础，测定结果的统计检验方法，方差分析、回归分析及正交设计等，并附环境监测数理统计实例。

书的第一、二章由黄文辉同志编写、第三、四、五章和实例由林兆升同志编写，并邀请高华寿教授、谢清成副总工程师进行了审阅。吴波同志参加了部分资料整理及试验数据统计工作。

由于时间匆忙，编者水平有限，内容、错误不妥之处，请读者批评指正。

编著者　1983年3月于成都

目　　录

原
书
缺
页

第一章　误差及数据统计处理的理论基础

误差自始自终存在于一切科学试验的过程之中。如何科学地处理环境监测中得到的大量试验数据，如对数据中离群较远的极值（极大值或极小值）的取舍，估计数据的可靠程度，对影响试验结果因素的分析，对误差进行计算，准确简练地表达分析结果，并给以合理的解释等，都需要使用数理统计方法。也就是要具体应用概率论的一些知识，通过样本分析，了解和判断总体的统计特性。

什么是总体和样本呢？总体（或称母体）指从研究对象得到的所有可能的观测结果。样本（或称子样）指从总体中抽取出来的一部分样品x_1、x_2……x_n的测定值。样本中样品的个数称为样本的大小（或容量），当n＞30时，称为大样本。

必须指出，使用数理统计方法仅仅是分析工作者解决问题的有力工具，它不能代替严格的试验工作；而恰恰相反，它只能在可靠的分析测试基础上，才能发挥其应有的作用。

§1—1　基　本　概　念

误差，根据生产的原因可分为系统误差（可测误差）偶然误差（随机误差）及粗差（过失误差）。

系统误差和偶然误差并没有绝对严格的界限，有时人们对系统误差的复杂规律认识不清时，往往把系统误差当作偶然误差来处理。

系统误差和偶然误差在分析过程中传递方式是不同的。

误差与偏差：误差指分析测定值与真实值之差。真实值是客观存在，一般是不知道的。随着测定技术的不断提高，可使测定值逐步接近真实值。在分析工作中人们常把标准试样中的某成份的含量作为该成份的真实值（至于标准试样的含量如何确定，不在此讨论）或把大样本的平均值作为真实值与测定值比较，可以估计出误差的大小。

偏差指分析测定值与平均值之差，反映数据之间的离散程度。偏差大小既与方法本身的精确与否，也与实验人员的操作水平有关。通常人们以平均值代替真实值计算误差，严格说来应称作偏差。

精密度和准确度：精密度指测得的数据之间重复的程度，反映偶然误差的大小。准确度指测定值与真实值（或多次测定平均值）符合的程度，反映偶然误差和系统误差的大小。评定分析数据的好坏，首先要考虑精密度，其次要考虑准确度。一般来讲，在系统误差已消除的情况下，精密度愈高分析结果愈准确。但若有系统误差存在，则精密度高、准确度不一定高。如图1—1所示。

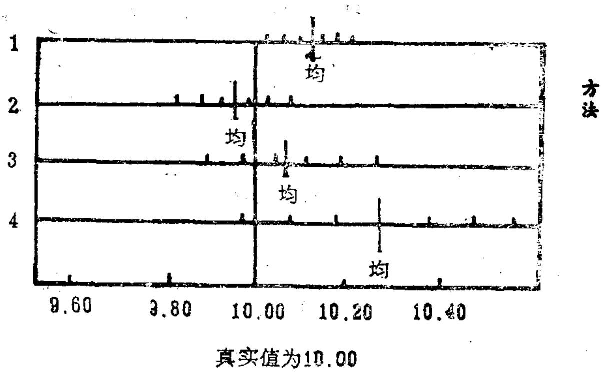

图 1—1 方法的准确度和精密度示意

图中把测定值表示在数轴上，并标出每种方法测得的平均值。

由图1—1可见，方法1的精密度高，说明偶然误差小，但平均值与真实值相差较大，有系统误差存在，故准确度不高。方法2的精密度和准确度都很高，说明偶然误差和系统误差都很小。方法3的精密度很差，偶然误差大，平均值虽然接近真实值，那是由于正负误差正好相互抵消的结果。方法4的精密度和准确度都很差即系统误差和偶然误差都很大。

平均值是最可信赖值：平均值的误差最小，是真实值μ的最好估计值。常以平均值$\bar{x}$作为μ的代表值，这已为人们所公认，称之为平均值公理，可以简单地证明如下：

在相同条件下独立地进行n次测定，得数据x_1，x_2……x_n。如系统误差已消除，则

$$d_1 = x_1 - \mu$$

$$d_2 = x_2 - \mu$$

$$\vdots$$

$$d_n = x_n - \mu \,,$$

$$\frac{1}{n}\sum_{i=1}^{n} d_i = \frac{1}{n}\sum_{i=1}^{n} x_i - \frac{n\mu}{n} = \bar{x} - \mu$$

由于偶然误差具有抵偿性，

故$\frac{1}{n}\sum_{i=1}^{n} d_i \to 0$，　　　　则　$\bar{x} \approx \mu$

平均值的计算方法

(1)算术平均值　这是常用的计算方法

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \text{。}$$

(2)加权平均值　计算用不同方法或不同条件下对同一样本得到的测定值的平均值时，因为方法及条件不同，其数值的精度

与测定次数可能不一致，可靠程度也有差异。要把这些因素反映出来，常对不同的数据给以不同的"权"，即是对一系列不同条件下得到的测定值，用数学的方法对其中好的测定值给予大的信任，在计算平均值时，使好的测定值占有较大的比例。

由数学推导可得出"权"的大小与标准差* S的平方成反比，测定精度高，标准差小，"权"的值就大。所谓加权，就是对精度较高的测定值乘一个较大的系数，对精度较差的测定值乘一个较小的系数，这个系数就称为"权"，一般用ω_i代表第i个测定值的"权"。

$$\omega_i = \frac{1}{S_i^2},$$

按此式求出"权"后，再用下式求出加权平均值

$$\bar{x}_{加} = \frac{\sum_{i=1}^{n} \omega_i x_i}{\sum_{i=1}^{n} \omega_i}$$

例1—1 有五组测定值，精密度不一致，其测定值及计算结果如下：

组测定值 $\bar{x}_i$	标准差 S_i	权 $\omega_i\left(\frac{1}{S_i^2}\right)$	权 ω_i'
14.7	0.22	20.66	1.62
14.1	0.39	6.57	0.52
14.2	0.28	12.76	1
14.6	0.50	4.00	0.31
14.9	0.10	100	7.84

* 标准差S的计算方法见§1—2节。

由于ω_i数值太大，计算加权平均值不方便，可任选一个ω值令其等于1，求出其余的ω与它的比例系数。如令12.76(ω_3)改为1，则 $\omega_1=\frac{20.66}{12.76}\approx1.62$， $\omega_2=\frac{6.57}{12.76}\approx0.52$……如$\omega_i'$所示。

加权平均值 $\bar{\bar{x}}_{加}=\frac{\sum\omega_i\bar{x}_i}{\sum\omega_i}=\frac{166.6}{11.29}=14.75$

这五个平均值中，14.9的精度最高，S值最小，权的值也就最大（7.84），表示对它的信任程度最高。

如按平均值的一般计算法，则总平均值$\bar{\bar{x}}_i=14.50$，与加权平均值相差较大。

如果测量的数据，精度基本一致，而测定次数不同，有时也需用加权平均。

例1—2 测定值

甲组 4.3,4.5,4.7 $\bar{x}_甲=4.5$ $S_甲=0.20$

乙组 4.1,4.4,4.3,4.5,4.6,4.6 $\bar{x}_乙=4.4$ $S_乙=0.195$

求总平均值

解：可以认为是等精度测定，把每组数据的次数作为权ω_i，即权$\omega_甲=3$，$\omega_乙=6$

$$\bar{\bar{x}}_{加}=\frac{\sum\omega_i\bar{x}_i}{\sum\omega_i}=\frac{3\times4.5+6\times4.4}{3+6}=4.43$$

(3) 中位数 在环境监测中，当其得到数据比较分散，其中有少数数据离群较远，而取舍又难以确定时，可用中位数代替平均值。使用中位数并不要求分析数据必须遵循正态分布，比较方便。确定中位数的方法是把n个测定值依大小顺序排列，当n为奇数时，取中间位置的数，即中位数$\bar{x}_m=x_{\frac{n+1}{2}}$，当n为偶数时，

取中间两数的算术平均值，即$\bar{x}_m=(x_{\frac{n}{2}}+x_{\frac{n}{2}+1})/2$。

§1—2 精密度的表示方法

精密度高低的表示方法有多种，现介绍常用的几种：

1．极差

极差是指一组测定值中最大值和最小值之差，也称为范围误差。

以$R=X_{max}-X_{min}$表示。

求极差时，未充分利用所有数据，用它反映精密度的高低是比较粗糙的，有时造成误差较大；但计算简便，在快速检验中常有应用。

2．算术平均偏差

算术平均偏差是每个测定值与平均值之差的绝对值的平均。

设测定值为$x_1, x_2, \cdots\cdots x_n$。算术平均值为$\bar{x}$，测定值与算术平均值的偏差为$d_i=x_i-\bar{x}$。

则算术平均偏差 $\bar{d}=\frac{1}{n}\sum_{i=1}^{n} d_i=\frac{1}{n}\sum_{i=1}^{n}\left|x_i-\bar{x}\right|$。

算术平均偏差用百分数或千分数表示时，称为相对平均偏差

$$\%=\frac{\bar{d}}{\bar{x}}\times 100, \qquad ‰=\frac{\bar{d}}{\bar{x}}\times 1000$$

用算术平均偏差表示各次测定值之间彼此符合的情况，有时产生与实际情况不一致的错误，这是因为有可能在一组测定值中偏差彼此接近，而另一组测定值中偏差有大有小，而$\bar{d}$值完全可能相同。

例1—3 两组测定值，各有5个数据求$\bar{d}$

甲：2.9，2.9，3.0，3.1，3.1　$\bar{d}_{甲}\frac{0.1\times4}{5}=0.08$，

乙：2.8，3.0，3.0，3.0，3.2　$\bar{d}_{乙}\frac{0.2\times2}{5}=0.08$。

$\bar{d}$ 值虽然一样，但甲组数据重复性好一些，精密度高一些，

3．标准偏差（简称标准差）

$$总体标准差\sigma=\sqrt{\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{n}} \qquad (1-1)$$

式中：σ——总体标准差；

μ——总体平均值；

x_i——i次测定值；

n——测定值数目。

在实际工作中都是取少数几个样本进行测定，求不出总体标准差，故式（1—1）在实际运算中使用较少。

样本标准差：样本标准差S的计算方法有多种，在通常情况下采用标准法——贝塞尔（Bessel）公式为好。

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

$$或\quad S=\sqrt{\left[\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x)^2/n\right]/n-1} \qquad (1-3)$$

式（1—2）与式（1—3）是等效式，将式（1—2）平方项展开、整理，可得式（1—3）。

式中（n－1）称为自由度，表示独立偏差的数目，n次测定，测定值与平均值之偏差共有n个，当引入一个平均值后，n次测定中独立偏差数为（n－1）。也有人提出，自由度在这里

的意义表示独立变数（测定次数n）的个数减去计算偏差时所用非独立变数（平均值）的个数，故式中自由度为（n－1）。

标准差用百分数表示时，称为变动系数（或变异系数），用CV表示。

$$CV=\frac{S}{\bar{x}}\times100\% \qquad (1-4)$$

标准差是表示精密度的好方法，原因是使用了平方项，对一组测定中的较大误差和较小误差反映比较灵敏，反映的是有效精密度。仍以例1—3数据来看，分别求出标准差，

$$S_{甲}=\sqrt{\frac{4\times(0.1)^2}{4}}=0.10\ ;\quad S_{乙}=\sqrt{\frac{2\times(0.2)^2}{4}}=0.14\ ,$$

说明甲的精密度比乙的高，这是符合事实的。尽管对同一组数据算出的S比 $\bar{d}$ 稍大，也应该用S来表示精密度。所以近年来化学工作者在计算实验误差时，多采用标准差而不用算术平均偏差。

标准差的计算公式除上述标准法外，还常使用简便的近似计算法，如Dean和Dixon* 提出的适用于少量数据的标准差计算方法，称为极差法（或称瞬时标准差）。此法是把n次测定值的极差乘以与测定次数有关的偏差因子。

即　　$S_{极}$＝极差×偏差因子。　　　　（1—5）

例1—4　有两组数据，求标准差S

甲组数据	计算	乙组数据	计算
17.65	$\bar{x}=17.75$	4.51	$\bar{x}=4.52$
17.83	$R=17.92-17.63$	4.49	$R=0.13$
17.63	$=0.29$	4.59	
17.71	$S_{标}=0.124$	4.53	$S_{标}=0.049$
17.92	$S_{极}=0.29\times0.430$	4.46	$S_{极}=0.13\times0.43$
	$=0.125$		$=0.056$

*注：R.B.Dean and W.J.Dixon.Anal.Chem;23.636(1951)

$S_标$ 是按标准法（Bessel法）计算的标准差值。

$S_极$ 是按极差法计算的标准差值。

极差法计算简便，分析数据不多时可以使用，但不如用Bessel公式计算精确。

表1—1　偏差因子

测定次数	偏差因子	测定次数	偏差因子
2	0.886	9	0.337
3	0.591	10	0.325
4	0.486	11	0.315
5	0.430	12	0.308
6	0.395	13	0.300
7	0.370	14	0.294
8	0.350	15	0.288

4．多个样本测定时标准差的计算

若从同一总体中抽样m个，每个样本重复进行多次测定，每个样本重复分析次数不同，或各样本测定精度不同，这时总体偏准差可按下式计算：

$$S=\sqrt{\frac{\sum_{j=1}^{m}\sum_{i=1}^{n}(x_{ij}-\bar{x})^2}{\sum n_i-m}} \qquad (1-6)$$

例1—5　测定一批鱼体内的含汞量，抽取7条鱼，每条鱼测定n_i次，测定结果如下，计算标准差。