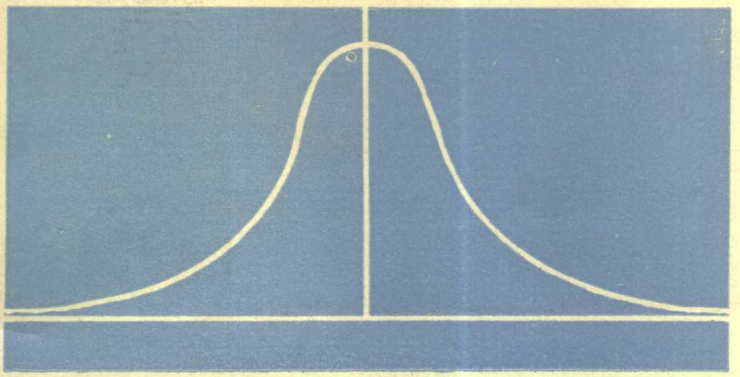
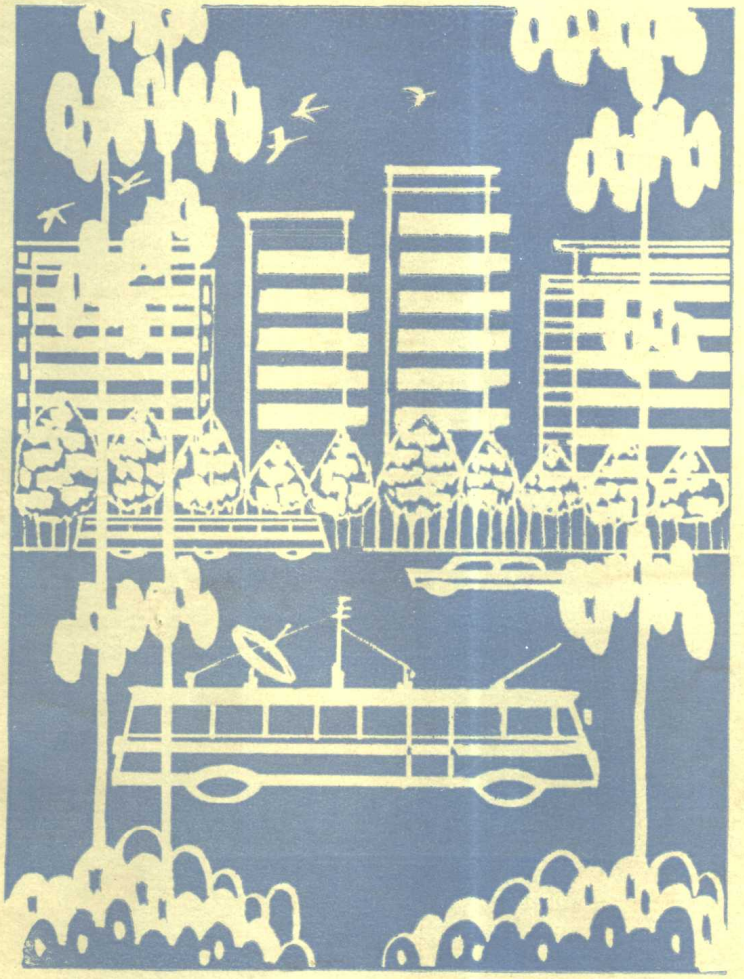


环境监测常用统计方法



高玉堂 主编

原子能出版社

环境监测常用统计方法

高玉堂　主编

原子能出版社

1980

内 容 简 介

本书介绍对环境监测数据常用的统计分析方法，内容包括数据整理、统计图表、样本特征数的计算、假设检验、参数估计和回归相关分析等，也提到了非参数方法、极差和控制图的应用。最后扼要地介绍了放射性测量和制订环境监测方案中的统计学问题。书末附有29种供查阅的统计用表。

本书可供环境保护工作者参考，书中所列的统计公式和附录用表也可供其他领域的工作者查阅。

环境监测常用统计方法

高玉堂　主编

原子能出版社出版

（北京2108信箱）

二龙路印刷厂印刷

新华书店北京发行所发行·新华书店经售

☆

开本 787×1092 1/16·印张 16 3/4·字数 395 千字

1981年 4月第一版·1981年 4月第一次印刷

印数 001—6000·统一书号：15175·195

定价：2.05 元

目　录

第一章　引言 ……（1）
第二章　数据整理和样本特征数计算 ……（3）
一、有效数字和一般运算规则 ……（3）
二、数据整理 ……（5）
三、样本特征数计算 ……（6）
四、本章内容提要 ……（15）
第三章　统计图表 ……（16）
一、统计表 ……（16）
二、统计图 ……（18）
三、本章内容提要 ……（23）
第四章　几种理论分布 ……（24）
一、概率 ……（24）
二、离散型和连续型随机变量分布 ……（25）
三、二项分布和泊松分布 ……（27）
四、正态分布和对数正态分布 ……（30）
五、概率纸的应用 ……（33）
六、几种抽样分布 ……（37）
七、本章内容提要 ……（39）
第五章　假设检验 ……（41）
一、假设检验概述 ……（41）
二、由一个样本检验总体参数的假设 ……（43）
三、由两个样本检验总体参数相等的假设 ……（48）
四、分布假设的检验 ……（57）
五、本章内容提要 ……（63）
第六章　参数估计 ……（68）
一、概述 ……（68）
二、“良好的点估计”具有的性质 ……（68）
三、区间估计 ……（69）
四、本章内容提要 ……（78）
第七章　方差分析 ……（81）
一、概述 ……（81）
二、单因素分组数据的方差分析 ……（82）
三、两因素分组数据的方差分析 ……（93）
四、三因素分组数据的方差分析 ……（107）
五、本章内容提要 ……（112）
第八章　回归和相关 ……（113）

一、直线回归方程 …… (113)
二、直线回归中关于参数的显著性检验和置信区间 …… (115)
三、相关系数 …… (117)
四、曲线回归直线化 …… (120)
五、其他有关问题 …… (123)
六、属性分组资料的关联性 …… (125)
七、本章内容提要 …… (126)
第九章　非参数统计方法 …… (127)
一、概述 …… (127)
二、符号检验 …… (127)
三、符号秩次检验（Wilcoxon 氏配对秩次检验） …… (128)
四、二样本秩和检验（Wilcoxon 氏二样本秩次检验） …… (129)
五、单因素分组数据的秩和检验（Kruskal-Wallis 氏检验） …… (131)
六、二因素分组数据的秩和检验（Friedman 氏检验） …… (132)
七、随机性检验 …… (133)
八、秩次相关 …… (137)
九、本章内容提要 …… (141)
第十章　极差在统计分析中的应用 …… (142)
一、由极差估计标准差 …… (142)
二、用极差推断总体均数置信区间 …… (142)
三、样本均数与总体均数差别显著性检验 …… (143)
四、两个样本均数比较 …… (144)
五、配对数据比较 …… (144)
六、单因素分组数据极差分析（重复数相等） …… (145)
七、两因素分组数据极差分析 …… (146)
八、本章内容提要 …… (149)
第十一章　控制图 …… (150)
一、常用的几种控制图的原理简释和编制方法 …… (150)
二、控制图在环境监测中应用举例 …… (152)
三、本章内容提要 …… (157)
第十二章　放射性测量的统计学问题 …… (159)
一、放射性计数的统计误差 …… (159)
二、样品净计数率统计误差的控制 …… (165)
三、放射性测量装置探测限和测定限的统计学估计 …… (168)
四、优质因子 …… (174)
五、放射性测量装置工作可靠性的统计学检验方法 …… (176)
六、本章内容提要 …… (179)
第十三章　环境监测方案中的统计学问题 …… (180)
一、环境监测方案的一般内容 …… (180)
二、随机抽样方法 …… (181)
三、采样时间和采样频度 …… (184)

四、监测方案中的质量保证 ……(188)
五、本章内容提要 ……(189)
附　录……(191)

第一章 引 言

随着工业生产的发展，环境保护问题越来越为人们所重视。党和国家对环境保护工作极为重视，我国宪法第一章第十一条中明确规定："国家保护环境和自然资源，防治污染和其他公害。"环境监测是环境保护事业的重要组成部分，是做好环境保护工作的先导。通过经常性的收集和分析环境监测数据，可以了解工业企业的"三废"排放对周围环境污染的现状和趋势，并为评价环境质量、制定"三废"治理措施和考核措施的效果、开展环境污染与人群健康关系的研究、制定和修订国家有关卫生标准等，提供必要的基础资料和科学依据。

为发现工业企业对环境的影响，有关部门制定了环境监测方案，有目的、有计划地经常收集大量的环境监测数据，并通过对不同地点、不同时期监测数据的比较，来判定企业的"三废"排放是否造成了环境中污染物水平的改变以及这种变化程度的大小。但是，使监测数据产生差异的原因是众多的，仅凭监测数据表面数值的差别，并不足以判定企业对环境的影响。大家都知道，在实验室里即使严格控制了分析、测量的条件，对同一样品重复分析、测量的结果是互有差异的，这种差异是由分析、测量过程中不可避免的误差造成的。在同一采样点上于同一时间采集多个样品，即使采样、制样和以后的分析、测量条件保持相同，各个样品的分析、测量结果间也有差异，这种差异是由采样、制样和分析、测量过程中为数极多的不能控制的原因造成的。因此，我们平时在不同地点、不同时间上取得的监测数据，其数值中既反映了地点、时间因素（与企业排放量、自然界的气象、水文等条件有关）对监测结果的影响，也还包括了上述仍不能控制的偶然性因素导致的误差在内。由于必然性因素（如企业排放量）的影响总是和偶然性因素（如所拟监测的物质在自然界分布的不均匀性，采样、制样和分析、测量过程中存在的误差）的影响纠缠在一起，在分析和比较环境监测数据时，只有充分认识和掌握偶然性因素所致差异的规律，排除其影响，才能作出有无必然性影响的正确判断。统计方法是以研究偶然现象的规律性的概率论为基础的科学分析方法，因此它是分析环境监测数据不可缺少的重要工具和手段。

在数理统计中，将研究对象的全体（满足指定条件的所有个体或单位的集合）称为总体，而将取自总体的一部分实际观测的个体或单位的集合称为样本。如果样本是用一定的方法按机遇的原则组成的❶，这种样本称为随机样本，样本包含的实测个体或单位的数目则为样本容量。例如，要测量某一时间某一地区内土壤的放射性水平，则该地区内全部应测的土壤构成总体。将全部应测土壤划分成许多相等体积的单位（观测单位），则总体由这所有的观测单位组成。按机遇的原则从中抽取若干个观测单位的土壤样品进行实测，这些实测的土壤样品就是总体的一个随机样本，而抽取的观测单位（土壤样品）的数目就是该样本的容量。显然，这样的总体和样本的概念同样可推广于其他的环境介质，如大气、水、动植物等。

---

❶ 参阅第十三章的随机抽样。

在实际工作中，一般只能对来自总体的样本进行观测，然而观测的结果却总是要推广到样本所属的未知总体，这是一个十分重要的科学推断过程。譬如说，为了解某一采样点上一个时期内某介质中的放射性水平，为此在该时期内采集了该介质的若干样品组成样本；由样本数据得到的平均水平是拟了解的总体放射性水平的估计值，接着的问题是如何通过样本的平均水平去估计未知总体的真实水平。又如为比较同一时期内二个采样点上放射性水平有无差别，我们只能从二个采样点上各自采集一批样品组成二个样本，通过比较二个样本的数据来了解该时期内二个点上总体的真实水平是否有差别。其实，数理统计的中心问题在于如何根据样本探求有关总体的种种知识，以及从样本取得的资料去检验关于总体的种种假设。

统计分析一般包括以下二个步骤：统计叙述和统计推断。统计叙述是指对实际观测的样本数据进行整理和归纳，用少数几个特征数（统计量），有时辅以必要的图表，将收集所得资料的主要特征反映出来，便于理解并作进一步的分析和比较。统计推断则指由实测样本推断所属未知总体的问题。

应用统计方法处理和分析环境监测数据的必要性，目前已为人们所认识，但同时应强调指出统计方法在研究设计中的重要性。如果监测方案制订得不好，所拟分析比较的因素安排不恰当，分析数据常会发生困难，甚至无法进行分析。因样本容量太小，不能达到研究的预期目的，或相反，收集的数据过多，造成人力、物力上的浪费，这些还是目前常遇到的情况。制订监测方案时应用统计方法，可以帮助合理地确定所需的样本容量（从而确定了采样周期），既保证获得为分析问题所必要的数据，又避免无谓的浪费。对以往监测数据的统计分析，有助于修订今后的监测方案，使方案不断完善和符合实际需要，这是统计方法应用于研究设计的一个重要方面。此外，由于放射性测量的结果计数具有统计涨落的性质，在选择和合理分配测量的时间、确定探测仪器和方法的灵敏度以及检验仪器的稳定性等方面，也必然涉及统计学知识的应用。因此，我们应把统计方法不仅看作为处理和分析数据的一种手段，也应将其应用视为研究设计的一个有机组成部分。目前，对环境监测工作中具体的采样和分析、测量方法，已有大量文献作了很详尽的叙述；然而，对制订环境监测方案时的统计学考虑，报道仍少。在本书的最后一章内，仅一般性地提出监测方案中的一些统计学问题，这些问题尚有待进一步的研究和讨论。

随着数理统计方法知识的普及，越来越多的工作单位在环境监测报告中应用统计方法分析数据和解释监测结果，也有更多的从事监测工作的同志要求熟悉和应用统计方法。为此，我们将近几年来各单位在工作中使用的一些统计方法，整理成册，便于从事实际工作的同志在日常工作中查阅使用。但是应该指出，考虑到环境监测数据的复杂性，影响监测结果的因素众多，为深入研究环境变化的规律，本书叙述的一些统计方法远远不能满足研究工作的需要。例如，为监测数据建立符合客观实际的关系模式，应用多元分析、协方差分析、时间序列分析等统计方法分析数据（包括电子计算机在环境监测中的应用），这些还有待探索和研究，需在今后的研究实践中积累这方面成功的经验，使环境监测领域内应用统计方法的内容得到不断的充实和更加丰富，使统计方法更有成效地为环境保护事业服务。

本书在编写过程中承蒙有关单位提供了许多资料和宝贵意见，定稿后，承医学科学院卫生研究所田凤调同志协助审阅，在此一并表示感谢。

第二章　数据整理和样本特征数计算

本章先扼要地介绍有效数字的概念，然后重点地叙述描述样本分布特征的几类常用数值。

一、有效数字和一般运算规则

用最小刻度表示的测量仪器的准确度总是有限的，测量结果存在“观测误差”，数值计算中也存在“舍入”等误差，因此，用数值表示的结果并不是绝对准确的，而是一些近似值。用几位数字来表示测量或计算的结果，是一个应该考虑的问题。

设 a 代表准确值 A 的一个近似值，则

$$\Delta_a = A - a \tag{2-1}$$

表示近似值 a 的误差。如果

$$|A - a| = |\Delta_a| \leqslant \varepsilon_a \tag{2-2}$$

称 ε_a 为近似值 a 的误差限，并将

$$\varepsilon_a / |a| \tag{2-3}$$

称为近似值 a 的相对误差限。（ε_a 一定是正数。）如果误差限 ε_a 已知，准确值 A 可用下面的不等式表示：

$$a - \varepsilon_a \leqslant A \leqslant a + \varepsilon_a \tag{2-4}$$

例如，用最小刻度为 0.1 厘米的尺测量一线段的长度 L，规定取最接近于线段终末端的刻度读数 l 为 L 的近似值（即采用所谓的靠近尺度），这时 l 的误差限为最小刻度单位的一半，即 0.05 厘米。如读得 $l = 6.2$ 厘米，则相对误差限为 $0.05/6.2 \cong 0.008$，准确值 L 在[6.15，6.25]区间内。

如果近似值 a 的误差限是某一位上的半个单位❶，该位到 a 的第一位非零数字共有 n 位，则说 a 有“n 位有效数字”。如上面的 6.2 厘米，误差限为最末位（数字 2）上的半个单位（0.05 厘米），从数字 2 起到前面的第一位非零数字 6 为有效数字，此数值有二位有效数字。又如 1007，100.7，1.007 三个数值，按上述定义均有四位有效数字(1,0，0，7)，但这些数值的误差限是不同的，分别为0.5，0.05，0.0005。在有效数字中，除末位数字是不准确的以外，其余数字都是准确的。在记录测量结果时，一般应只保留一位不准确的数字。从有效数字的位数上，结合所用的量度单位，可以了解测量结果所达的准确程度。

第一位非零数字前若有数字 0 存在，这些 0 不作为有效数字。如 6.2 厘米改用米为单位，写为 0.062 米，这时误差限不变，此数值仍为二位有效数字，前面二个 0 不算有效数

❶当准确值 A 在二个最小刻度读数 a_1（较小值）和 a_2（较大值）之间，若规定恒取 a_1（或恒取 a_2）为 A 的近似值，这时误差限为最小刻度的一个单位。但像本文中那样定义的有效数字，用起来有方便之处，因为一些测量工具的误差限常为最小刻度单位的一半，在计算时由四舍五入产生的误差也是不超过最末位数字的半个单位。

字。在数值中小数点的位置只与所用量度单位的大小有关，并不反映测量结果的准确程度。但如果测量结果的末位数字记为0，此数字0仍为有效数字。例如，将线段的长度记为$l=6.20$厘米，这样的写法表示量尺的最小刻度为0.01厘米，误差限为0.005厘米，准确值L在[6.195，6.205]区间内，因此6.20有三位有效数字。要注意6.2和6.20二种写法的区别，它们所表示的测量结果的准确度是不同的。

为清楚地指明有效数字的位数，也常用10的幂次前面的数字表示有效数字的方法。如称重结果写为1.3公斤，此数值有二位有效数字，表示误差限为0.05公斤，准确值在[1.25，1.35]区间内。若要改写成以克为单位，宜写成1.3×10^3克，这种表示方法既反映了量的大小，又正确表明了有效数字的位数，这时有效数字仍为二位，误差限不变。现若直接写成1300克而不给以任何说明，可以理解成误差限为0.5克，准确值在[1299.5，1300.5]区间内，这时有效数字将被认为是四位，容易造成混淆。

在实际计算中，用四舍五入法以较少位数代替较多位数时，结果的误差限为保留的数值末位上单位的一半。保留n位数字时，凡弃去的第n位后的数字小于第n位上单位的一半，第n位的数字不变；若大于第n位上单位的一半，在第n位数字上加1。如果弃去的数字恰等于第n位上单位的一半，当第n位数字为偶数时其数字不变，若为奇数，则在该数字上加1。例如，$\pi=3.14159265\cdots\cdots$。如拟保留三位数字，取$a=3.14$为$\pi$的近似值，这时

$$|\pi-3.14|=0.00159265\cdots<\frac{1}{2}\times0.01$$

如果保留五位数字，取$a=3.1416$，这时

$$|\pi-3.1416|=0.00000734\cdots<\frac{1}{2}\times0.0001$$

也就是说，用上述四舍五入法时，误差不超过最末位上的半个单位。

近似值进行加减乘除等运算时，可应用下述一些规则：

1．加减计算结果的误差限应与各数中误差限最大者相同。如在小数运算中，加减计算结果保留的小数点后的位数，与各数中小数点后位数最少者相同。在实际计算中，可将各数比小数点后位数最少的数多保留一位小数，计算结果则按上述规则表示。如

$$561.32+491.6+86.954+3.9462$$

$$\approx561.32+491.6+86.95+3.95=1143.82$$

最后用一位小数表示为1143.8。

当二个数值相近的近似值相减时，差数的有效数字位数比原数值大为减少。例如，$12.1471-12.1460=0.0011$，原数值有六位有效数字，前面相同的有效数字在减的过程中消失了，差数只剩下二位有效数字。遇有这种情况，如果可能，预先在原来数值内多保留几位有效数字。有时也可选用其他适当的方法处理。

2．在乘除运算中，较稳妥的规则是：一开始在有效数字位数较多的数值中比位数最少的数值多保留一位有效数字，计算结果的有效数字位数则与位数最少者相同。如计算$\frac{4.892\pi}{6.7}$时，因其中6.7只有二位有效数字，故可作如下的计算：

$$\frac{4.89\times3.14}{6.7}\cong2.29$$

最后的结果保留二位有效数字，即 2.3。

3．近似值的平方或立方运算时，计算结果的有效数字位数与原数值的位数相同。近似值的平方根或立方根的有效数字位数也与原数值相同。

4．准确度相同的多个近似值求平均数时，结果的有效数字位数可增加一位。

5．计算式中常数 π，e 以及 $\sqrt{2}$，1/3 等数值的有效数字位数，在计算时需要几位就写几位。

上面对有效数字和近似值运算只作了一般性叙述，关于这方面的问题还可参阅有关的文献❶。最后应该指出，统计分析中的计算是由一系列连续运算组成，并且都使用计算工具，在计算过程中一般可以多留几位数字，不必在计算的每一步上均拘泥于上述规则，而最终报告结果的有效数字位数则应限制在合理的范围内。

二、数 据 整 理

着手分析数据以前，要对原始数据进行必要的整理。应先逐一检查原始记录是否按规定的要求填写完全、正确。查明有过失错误的数据（如采样、分析、测量过程中操作错误或发生意外污染等），应予舍去。发现有计算或记录错误的数据，应予订正。但不能轻易剔除数值异常的数据，因为这些数据可能反映了企业事故排放对环境的污染或其他因素的影响，应该进一步查明其原因。

为了解监测项目在时间上的动态变化或在不同距离地点上的变化，将数据按出现的时间先后或距离远近依次列出（有时将数据按时间、地点适当归组）进行分析。也常将数据绘制成图（参阅第三章），便于从直观上分析监测项目的变化。

在采样遵从随机原则的条件下，收集所得的一批数据常视作来自一特定“环境总体”的一个随机样本。当样本数据较少时，可直接进行计算和分析。倘若数据的数量众多，通常先将数据按其数值大小分组，数出归入各组的数据数目，编制成样本频数分布。

观测指标（下面称为变量）有离散型和连续型之分。离散型的变量只能取数轴上数个孤立的数值，如放射性测量时给定测量时间内的计数，其结果只能为 0，1，2，…… 等。连续型的变量则能取数轴的某个区间上的一切数值，如测定空气中铀浓度（微克/立方米），测定结果并不表明确切为某一数值，而是位于某个区间上的数值。

整理离散型数据时，将可能出现的变量值（如 0，1，2，……）列出，再将数据归组，数出不同变量值出现的次数，即为频数（见表 4-1）。有时也将变量值分组，如列为 0—4，5—9，10—14，……等，并用各组的算术均数（如2，7，12，……等）代表归入各该组的数据。

整理连续型数据时，分组以区间（称组段）形式出现。组段的宽度称为组距，常用的

---

❶ 高等数学（基础部分），§ 10.1，人民教育出版社，1960。

Пиголев Б.М.: Математическая Обработка Наблюдений, часть I, Гос. Издат. Физ-Мат. Лит., Москва, 1960.

是等距分组。每一组段的起始值称为组段的下限，截止值称为组段的上限。计算时以各组段的组中值（即该组段下限和上限的算术均数）代表归入各该组段的数据。编制连续型数据频数分布的具体步骤如下：

（1）找出数据中的最大值和最小值，二者之差称为极差（或全距）。

（2）考虑分组的组数，一般分为10—15组。如果数据较少，组数可适当减少。

（3）将极差除以所拟分组数，估计出所用的组距。将组距适当调整为较方便的数值，如0.10、0.20等，写出各组段。第一组段下限的末一位数字最好取惯用的0，5等数值，便于以后划记归组。要使第一组段包括最小值在内，末一组段包括最大值在内。组段的写法如表2-1所示，如第一组段“0.50—”表示包括从0.50起不到0.60的变量值。

（4）将数据按所分组段划记，最后得到归入各组段的变量值数目，即频数。

表2-1 某年某地土壤样品中铀含量(微克/克)的频数分布

分　组	划　记	频　数
0.50—	丨	1
0.60—	卌	5
0.70—	卌	5
0.80—	卌卌丨丨	12
0.90—	卌卌卌丨	16
1.00—	卌卌卌卌卌卌丨	32
1.10—	卌卌卌卌卌卌卌卌卌卌	50
1.20—	卌卌卌卌卌卌卌卌卌丨	46
1.30—	卌卌卌卌卌卌卌卌丨丨丨	43
1.40—	卌卌卌卌丨丨丨	24
1.50—	卌卌卌丨丨丨	18
1.60—	卌卌丨丨丨丨	14
1.70—	丨丨	2
1.80—1.90以下	丨	1
合　计		269

为了解变量值出现在各组段内的相对频度，将各组段的频数 f_i 除以总频数 $\sum_i f_i$，其商称为频率。若欲了解变量值小于某一组段下限出现的频数或频率，可将该组段之前各组的频数或频率累加，求得相应的累积频数或累积频率（表2-2）。

离散型和连续型变量频数（或频率）分布的图示，可分别用条图和直方图绘制（见第三章）。

三、样本特征数计算

为描述一批数据具有的某些重要特征，常常要计算几个单一的数值综合地反映这些特征。这些数值中有的表明分布的集中位置，有的表明测定分布的离散程度，还有的表明测定分布的形状（偏度和峰度）。

表 2-2　某年某地土壤样品中铀含量(微克/克)的频率分布和累积频率

分　组	频　数	频　率	累积频数	累积频率
0.50—	1	0.0037	1	0.0037
0.60—	5	0.0186	6	0.0223
0.70—	5	0.0186	11	0.0409
0.80—	12	0.0446	23	0.0855
0.90—	16	0.0595	39	0.1450
1.00—	32	0.1190	71	0.2639
1.10—	50	0.1859	121	0.4498
1.20—	46	0.1710	167	0.6208
1.30—	43	0.1599	210	0.7807
1.40—	24	0.0892	234	0.8699
1.50—	18	0.0669	252	0.9368
1.60—	14	0.0520	266	0.9888
1.70—	2	0.0074	268	0.9963
1.80—1.90以下	1	0.0037	269	1.0000
合　计	269	1.0000	—	—

由样本数据计算得到的这类数值（称为样本特征数），是总体相应的数值（称为总体参数）的估计值。由于抽样的随机性，样本特征数随抽取样本不同本身也是一些变量，而它们所估计的相应的总数参数则是一些常量。

1．未分组数据的计算

(1) 算术均数

设有 n 个变量值 x_1，x_2，……，x_n，则算术均数 $\bar{x}$ 为

$$\bar{x}=\frac{x_1+x_2+\cdots\cdots+x_n}{n}=\sum_{i=1}^{n}x_i/n \qquad (2\text{-}5)$$

式中 Σ 为求和的符号，n 为该样本的容量。

算术均数是最常用的表明数据集中位置的数值，反映数据的平均水平，但易受数据中特大或特小值的影响。对于不对称（偏态）分布的数据，它并不反映数据的典型水平。

(2) 中位数

将变量值由小到大排列以后，居于中间位置的变量值就是中位数 M_e。

当样本容量 n 为奇数时，

$$M_e=\text{第}\frac{n+1}{2}\text{个变量值} \qquad (2\text{-}6)$$

n 为偶数时，

$$M_e=\frac{1}{2}\left[\text{第}\frac{n}{2}\text{个变量值}+\text{第}\left(\frac{n}{2}+1\right)\text{个变量值}\right] \qquad (2\text{-}7)$$

中位数的特点是将全部变量值一分为二，大于或小于中位数的变量值各占一半。它不受特大或特小值的影响。在偏态分布中，它比算术均数更能代表数据的典型水平。

【例 2-1】某年某采样点大气自然沉降物（自然沉降天数为 7 天）中铀含量（微克/平方米·天）的 25 个监测数据为：

0.63, 0.51, 0.67, 0.86, 0.56, 0.56, 0.63, 0.55, 0.74, 1.10, 1.04,
0.34, 0.37, 1.16, 0.27, 0.83, 0.51, 0.30, 0.17, 0.36, 0.18, 0.30,
0.50, 0.98, 0.60

求其算术均数和中位数。

由公式（2-5）得

$$\bar{x}=\frac{0.63+0.51+\cdots\cdots+0.60}{25}=\frac{14.72}{25}=0.589$$

然后将上列数据由小到大重新排列如下：

0.17, 0.18, 0.27, 0.30, 0.30, 0.34, 0.36, 0.37, 0.50, 0.51, 0.51,
0.55, 0.56, 0.56, 0.60, 0.63, 0.63, 0.67, 0.74, 0.83, 0.86, 0.98,
1.04, 1.10, 1.16

样本容量 $n=25$（为奇数），由公式（2-6），$\frac{n+1}{2}=13$，故第 13 个数值 0.56 为所求的中位数。在本例中，算术均数和中位数很接近。

（3）众数

众数是数据中出现频数最多的变量值，用符号 M_o 表示。由大数量分组数据可得到众数的近似值，但它在进一步统计分析中的应用不广。

（4）几何均数

n 个变量值 x_1，x_2，……，x_n 的几何均数 G 等于它们的连乘积的 n 次方根，即

$$G=\left\{\prod_{i=1}^{n}x_i\right\}^{1/n} \tag{2-8}$$

式中 Π 为连乘的符号。

实际上，通常用下式运算：

$$\lg G=\frac{1}{n}\sum_{i=1}^{n}\lg x_i \tag{2-9}$$

查 $\lg G$ 的反对数，得到所求的几何均数。

据报道❶，环境介质中许多物质浓度数据的分布近似呈第四章提到的对数正态分布，这时计算和应用几何均数有着重要的意义。

【例 2-2】某年某采样点测得空气中α放射性强度（$\times 10^{-17}$居里/升）的 29 个监测数据为

37, 43, 28, 30, 12, 11, 4, 13, 9, 11, 9, 18, 5, 18, 9, 19, 19,
16, 7, 11, 28, 25, 69, 9, 23, 41, 63, 4, 53

求其几何均数、中位数和算术均数。

将变量值由小到大重新排列，并各取其对数值（括号内数值）如下：

4(0.6021), 4(0.6021), 5(0.6990), 7(0.8451), 9(0.9542),

❶ 未发表资料。

9(0.9542), 9(0.9542), 9(0.9542), 11(1.0414), 11(1.0414),
11(1.0414), 12(1.0792), 13(1.1139), 16(1.2041), 18(1.2553),
18(1.2553), 19(1.2788), 19(1.2788), 23(1.3617), 25(1.3979),
28(1.4472), 28(1.4472), 30(1.4771), 37(1.5682), 41(1.6128),
43(1.6335), 53(1.7243), 63(1.7993), 69(1.8388)

由公式（2-9）得

$$\lg G = \frac{2\times 0.6021 + 0.6990 + \cdots\cdots + 1.8388}{29} = \frac{35.4627}{29} = 1.2229$$

查反对数，　$G = 16.7$。

在本例中，中位数为第 15 个变量值，即　$M_e = 18$。算术均数为

$$\bar{x} = \frac{2\times 4 + \cdots\cdots + 69}{29} = \frac{644}{29} = 22.2$$

本例的几何均数与中位数甚为接近，而算术均数与中位数之差则较大。因此，本例的几何均数比算术均数更能代表数据的典型水平。这是第四章所述对数正态分布数据的重要特点。

（5）标准差和方差

标准差是最常用来测定分布离散程度的数值，其平方称为方差。

样本方差（用 S^2 表示）的计算公式为

$$S^2 = \sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2/(n-1) \qquad (2\text{-}10)$$

在实际运算中，上式中的分子一般用下式计算：

$$\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2 = \sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2/n \qquad (2\text{-}11)$$

因此，公式（2-10）可写成

$$S^2 = \left[\sum x^2 - \frac{(\sum x)^2}{n}\right]/(n-1) \qquad (2\text{-}12)$$

而样本标准差的计算式则为

$$S = \sqrt{\left[\sum x^2 - \frac{(\sum x)^2}{n}\right]/(n-1)} \qquad (2\text{-}13)$$

标准差恒取正值，它的量纲与原变量相同。标准差和方差在统计分析中的应用极广。

有时要计算样本的几何标准差（用 S_g 表示），其值为变量值取对数后的标准差 $S_{\lg x}$ 的反对数，即

$$S_g = anti\lg S_{\lg x} \qquad (2\text{-}14)$$

式中

$$S_{\lg x} = \sqrt{\left[\sum (\lg x)^2 - \frac{(\sum \lg x)^2}{n}\right]/(n-1)} \qquad (2\text{-}15)$$

几何标准差是无量纲的数值。

【例 2-3】用例 2-1 数据计算标准差

本例 $n = 25$，$\sum x = 14.72$，$\sum x^2 = 10.5726$。由公式（2-13）得

$$S=\sqrt{\left[10.5726-\frac{(14.72)^2}{25}\right]/(25-1)}$$

$$=\sqrt{0.0794}=0.282\ (\text{微克/平方米·天})$$

【例 2-4】用例 2-2 数据求几何标准差

本例 $n=29$，$\sum\lg x=35.4627$，$\sum(\lg x)^2=46.600419$。由公式（2-15）得

$$S_{\lg x}=\sqrt{\left[46.600419-\frac{(35.4627)^2}{29}\right]/(29-1)}$$

$$=\sqrt{0.115528}=0.3399$$

查反对数，$S_g=2.2$。

（6）极差

数据中最大值与最小值之差称为极差（用R表示），

$$R=x_{max}-x_{min} \qquad (2\text{-}16)$$

式中 x_{max} 为最大值，x_{min} 为最小值。

极差也可用来测定数据的离散程度，它计算简便，但易受数据中个别特大或特小数值的影响。在一定条件下，在样本容量较小（$n\leqslant 10$）的数据分析中极差仍有较广的应用，本书第十章和第十一章内将对极差的应用作进一步叙述。

（7）变异系数

为表明分布的相对离散程度，有时计算标准差与算术均数的百分比值这一无量纲的数值，并称之为变异系数（用 C.V.表示）。

$$\text{C.V.}=\frac{S}{\bar{x}}\times 100\ (\%) \qquad (2\text{-}17)$$

如例 2-1 数据的 $\bar{x}=0.589$，$S=0.282$，于是

$$\text{C.V.}=\frac{0.282}{0.589}\times 100\ (\%)=47.9\%$$

在环境监测工作中，为比较不同环境介质或不同核素的监测数据的离散程度，可考虑应用变异系数这一相对量度。

2. 分组数据的计算

（1）算术均数和标准差

对已编制成频数分布的数据计算算术均数和标准差时，用各组段的组中值代表归入各组的变量值（表 2-3），这时计算的公式为

$$\bar{x}=\sum fx/\sum f \qquad (2\text{-}18)$$

$$S=\sqrt{\left[\sum fx^2-\frac{(\sum fx)^2}{\sum f}\right]\Big/(\sum f-1)} \qquad (2\text{-}19)$$

式中 x 为各组段的组中值。

为计算简便起见，一般将组中值 x 变换成新变量 x'，然后再行计算。这时令

$$x' = \frac{x - x_0}{i} \tag{2-20}$$

式中 i 为等距分组的组距，x_0 为任一指定组段的组中值（常取频数较大一组的组中值为 x_0）。实际上不必一一计算各组段的 x'，可将 x_0 所在组的 x' 定为 0，向上各组的 x' 依次（由下而上）写成 -1，-2，……，向下各组的 x' 依次（由上而下）写成 $+1$，$+2$，……就可以了。若某一组段的频数为 0，仍需依次写下去，而不能跳过此组段。

用变量 x' 计算算术均数和标准差的公式为

$$\bar{x} = x_0 + \frac{\sum fx'}{\sum f} \times i \tag{2-21}$$

$$S = i \times \sqrt{\left[\sum fx'^2 - \frac{(\sum fx')^2}{\sum f}\right] \Big/ (\sum f - 1)} \tag{2-22}$$

【例 2-5】用表 2-1 资料计算算术均数、标准差和变异系数。

由表 2-3，$x_0 = 1.15$，$i = 0.10$，$\sum f = 269$，$\sum fx' = 214$，$\sum fx'^2 = 1638$。于是

$$\bar{x} = 1.15 + \frac{214}{269} \times 0.1 = 1.23$$

$$S = 0.1 \times \sqrt{\left[1638 - \frac{214^2}{269}\right] \Big/ (269 - 1)} = 0.23$$

$$C.V. = \frac{0.23}{1.23} \times 100(\%) = 18.7\%$$

表 2-3　用表2-1 资料计算算术均数和标准差

分组(微克/克)	组中值 x	频　数 f	$x' = \frac{x - x_0}{i}$	fx'	fx'^2
0.50—	0.55	1	−6	−6	36
0.60—	0.65	5	−5	−25	125
0.70—	0.75	5	−4	−20	80
0.80—	0.85	12	−3	−36	108
0.90—	0.95	16	−2	−32	64
1.00—	1.05	32	−1	−32	32
1.10—	1.15(x_0)	50	0	0	0
1.20—	1.25	46	1	46	46
1.30—	1.35	43	2	86	172
1.40—	1.45	24	3	72	216
1.50—	1.55	18	4	72	288
1.60—	1.65	14	5	70	350
1.70—	1.75	2	6	12	72
1.80—1.90 以下	1.85	1	7	7	49
合　　计	—	269	—	214	1638