普通高等教育“十一五”国家级规划教材

全国高等学校药学类规划教材

医药统计学

主编　罗　旭　毕开顺

高等教育出版社
Higher Education Press

医药统计学

普通高等教育“十一五”国家级规划教材
全国高等学校药学类规划教材

医药统计学

Yiyao Tongjixue

主　编　罗　旭　毕开顺
副主编　张丕德　赵春杰
编　委　（以姓氏笔画为序）
石　娟　毕开顺　乔延江　刘晓娟
苏薇薇　杜培革　李　海　余露山
张丕德　罗　旭　赵春杰　姚美村
董莉萍

高等教育出版社·北京
HIGHER EDUCATION PRESS　BEIJING

内容简介

本书是普通高等教育“十一五”国家级规划教材。

本书系统地介绍了药学统计学的基本内容。它有以下几个特点：一是理论联系实际、起点低，因而容易学习。本书着重联系药学实际，深入浅出地阐明了统计学的基本知识、基本方法和基本计算技术，并为此设专章补充了概率论方面的基础知识。二是重点突出，设专章论述“数据的误差叠加与处理”。误差叠加规律不仅对理解在统计学理论中占重要地位的中心极限定理是有益的，而且对误差分析的实践起指导作用。对计算机软件和当代统计方法如蒙特卡罗方法等则不作为重点。三是便于学习。各章内容相对独立，可按顺序逐章地学，也可单学几章或一章的全部内容或其基本部分。书中提供了药学领域中与经常面临的问题相对应的实际例子，读者可以按其步骤模仿使用。四是展望未来。把应用当代统计方法和计算机技术于药学领域取得的重要成果作为前瞻性内容，在“后记”中扼要地说明两者的重要性。

本书可供药学专业的本科生、研究生以及药学工作者作为教材或自学教材，也可供其他专业的读者作参考书使用。

图书在版编目(CIP)数据

医药统计学/罗旭，毕开顺主编. —北京：高等教育出版社，2010.1

ISBN 978-7-04-028139-2

Ⅰ.医… Ⅱ.①罗…②毕… Ⅲ.医学统计-统计学-高等学校-教材 Ⅳ.R195.1

中国版本图书馆 CIP 数据核字(2009)第 224068 号

策划编辑 席 雁 **责任编辑** 董达英 **封面设计** 于文燕 **责任绘图** 尹 莉
版式设计 王 莹 **责任校对** 杨凤玲 **责任印制** 尤 静

出版发行	高等教育出版社	**购书热线**	010-58581118
社 址	北京市西城区德外大街 4 号	**咨询电话**	400-810-0598
邮政编码	100120	**网 址**	http://www.hep.edu.cn
总 机	010-58581000		http://www.hep.com.cn
		网上订购	http://www.landraco.com
经 销	蓝色畅想图书发行有限公司		http://www.landraco.com.cn
印 刷	北京京科印刷有限公司	**畅想教育**	http://www.widedu.com
开 本	787×1092 1/16	**版 次**	2010 年 1 月第 1 版
印 张	17.25	**印 次**	2010 年 1 月第 1 次印刷
字 数	410 000	**定 价**	22.40 元

本书如有缺页、倒页、脱页等质量问题，请到所购图书销售部门联系调换。

物料号 28139-00

全国高等学校药学类规划教材

药学概论	主编　叶德泳
天然药物化学	主编　吴继洲(普通高等教育“十一五”国家级规划教材)
生药学	主编　蔡少青
药理学	主编　李元建
药剂学	主编　张志荣(普通高等教育“十一五”国家级规划教材)
药物分析	主编　曾　苏(普通高等教育“十一五”国家级规划教材)
生物药剂学与药物动力学	主编　蒋新国
临床药动学	主编　蒋学华(普通高等教育“十一五”国家级规划教材)
药物设计学(第二版)	主编　仇缀百(普通高等教育“十一五”国家级规划教材)
临床药物治疗学	主编　胡晋红
药事管理学	主编　刘红宁(普通高等教育“十一五”国家级规划教材)
医药统计学	主编　罗　旭　毕开顺(普通高等教育“十一五”国家级规划教材)
药物经济学	主编　胡善联

其他药学类规划教材

药物化学(第二版)	仉文升　李安良(北京市精品教材)
大学化学基础	曹凤岐(普通高等教育“十五”国家级规划教材)
药学实用仪器分析	陈玉英(普通高等教育“十五”国家级规划教材)
药物化学	华维一(普通高等教育“十五”国家级规划教材)
生物技术药物学	吴梧桐(普通高等教育“十五”国家级规划教材)

前　言

本书系统地介绍了药学统计学的基本内容，这是药学在其发展中与统计学结合而形成的一个交叉学科。它包括药学与经典统计方法的结合以及与当代统计方法如蒙特卡罗方法的结合。它的名称体现了科学的继承性和形成交叉学科的发展。药物统计学被规定为药学的一个二级学科[代码350.50，全国普通高等学校科技统计工作(理、工、农、医类)文件，2001]。

但由于历史上的原因，当今多数药学工作者并不把药学统计学视为药学的一部分，而认为它是数学。工作中遇到需用统计方法解决的问题，虽想自学，但拿起概率论与数理统计教材或专著阅读，由于这类书籍并不结合药学实际，且有诸多难点，因此不禁视为畏途。近年，虽有部分药学院校鉴于统计学的重要性而设课，但由于缺乏适当的教材，即使授课教师联系药学实际，也很不充分，还是达不到预期的效果。本书就是适应这个需要而编写的。它立足于科学、面向教育、是一部教材和参考书。

本教材有四个特点。一是理论联系实际，起点低而容易学习。本书着重用药学领域的实例系统而深入浅出地阐明统计学基本知识、基本方法和基本计算技术。假定读者只学过高中数学，当然若学过微积分及概率论有关课程则更好(只有个别章节用到微积分知识)，并设专章论述有关概率论基础方面的内容。二是抓住重点，兼顾其余，但不旁骛。本书设专章论述"数据的误差叠加与处理"，强调系统误差与随机误差叠加方式的不同，推导了计算结果误差与直接观测误差的一般关系。这个关系不仅对误差分析的实践起指导作用，而且对理解在统计学理论中占重要地位的中心极限定理是有益的。对所举例子中涉及的药学专业的基本知识，化学等有关专业的读者不难理解；其他专业的读者可透过其细节，从如何应用的角度学习。鉴于非参数检验在统计学中只占次要地位，本书不设专章讨论，只在假设检验中作为与参数法的对比，介绍均值和标准差的非参数检验。对当代统计方法而言，计算机软件非常重要，在学习本书时有一个袖珍计算器则更方便。但由于本书的性质和篇幅所限，并不把计算机技术作为重点内容与统计学相提并论，从而毕其功于一役，更不把它视为可以完全取代统计学的技术——计算机技术取代的是统计学中烦琐的，乃至非用计算机不可的计算工作。本书对计算机技术在正文中虽有涉及，但仅在附录中给以扼要的陈述。三是长短结合，便于使用。本书大致是按由简单至复杂的顺序编写的，建议按顺序逐章地学。但这样学需要的时间长。由于各章的内容有相对的独立性，读者可以单学几章，或某一章的全部或其基本部分，甚至可以急用先学，从本书大量例题中找到与面临问题相应的例证，按其步骤模仿，先用起来解决问题，再追究其中的道理。这样学需要的时间短。这两种学习方法各有优缺点，可以结合起来。但学习统计学无论是用按部就班的方法，还是用按图索骥的方法，读者都必须付出努力，才能学懂、会用，乃至逐渐融会贯通，有所发现，有所进步。四是把当代统计方法和计算机技术应用在药学领域所取得的重要成果作为前瞻性内容，在"后记"中扼要说明两者的重要性。

参加本教材编写的人员均在药学统计学领域中具有丰富的教学和科研经验，有罗旭、毕开顺(第一、三章，沈阳药科大学)，石娟(第二、四章，西安交通大学)，赵春杰(第五章，沈阳药科大学)，

杜培革、董莉萍(第六章,北华大学),李胜联(第七章,桂林医学院),李海(第八章,四川大学),刘晓娟(第九章,辽宁医学院),姚美村(第十章,中山大学),余露山(第十一章,浙江大学),乔延江(第十二章,北京中医药大学),张丕德(第十三章,广东药学院)。

本教材的编写是以药学专业的读者能学懂、会使用、长学识,不同专业的读者可资借鉴,统计学家予以肯定,达到雅俗共赏而被接受为目的。虽做出很大努力,但缺点和错误在所难免,衷心欢迎广大读者和各界人士批评指正。

沈阳药科大学吴春福校长始终关心本书的出版,何春馥教授通阅了全部书稿并提出了许多宝贵意见,值本书出版之际,一并致谢。

编 者

2009 年 5 月 4 日

目　录

第一章 绪 论

§1.1 统计学简史与定义

人类对数据的研究，中外都可追溯至古代. 16 世纪至 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(1771—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多而每个因素的影响小，则这个量呈现为一种概率密度两头小、中间大的分布，并称之为观测误差理论. 19 世纪中叶，A. Quetelet(1796—1847)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，进而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非常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才能进行的限制. 20 世纪 20 年代，R. A. Fisher(1890—1962)等在小样本理论的基础上建立了显著性检验和方差分析. 在上述统计学理论指导下设计实验，称为实验设计. 于是统计学发展成为一门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科学. 统计学也被定义为“在面临不确定性时做出明智决策的科学”(Brownlee K. B.，1960). 这个定义很简短准确，但失之过于概括而不够具体.

现在，统计方法已成为科学许多分支学科的重要组成部分，如化学统计学、医学统计学、药学统计学和经济统计学. 药学统计学可以定义为研究数据的收集、描述、分析、综合和解释，以在药学领域获得新知识或信息或做出新推断的学科. 必须指出，不存在只适用于一个学科的特殊统计方法，但有可用于所有学科的统计方法. 为完成某一学科研究工作所发展起来的统计方法，一定能通过适应而在其他学科找到用场. 不过在学习、应用统计方法的过程中，通过本学科的具体事例比较容易，常收到事半功倍的效果.

§1.2 统计学的几个基本概念

在分章讨论药物统计学的繁多内容之前，本章要先阐述统计学的几个基本概念.

§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事件，即在某条件实现后一定发生或一定不发生的事件. “蓝色石蕊试纸遇酸变红”、“金不生锈”、“贵重药物用分析天平比用戥子称得准确”等一类事件称为必然事件(certain event). 但也有在一定条件下不一定发生的事件，即事件在未发生之前有两个

或两个以上可能出现的结果，一旦事件发生便只能出现其中的一个结果，这样的事件称为随机事件(accidental event). 例如，随便抛掷硬币，落下时标有面值的一面朝上朝下不一定；对一批针剂进行抽样检查，样本中的一支合格或不合格都有可能. 在前一事件中，每次抛掷硬币的结果，是由抛掷的初速、转动的快慢、抛掷角的大小等因素决定的，但这些因素都无法控制，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，包括含药量、杂质等，受到很多随机因素的影响，不可能得到完全的控制，因而可能合格，也可能不合格，虽然合格的可能性大.

§1.2.2 频率与概率

如果在 n 次试验中事件 A 发生了 n_A 次，则称 n_A 为事件 A 的频数(frequency)，比值 n_A/n 称为 A 在 n 次试验中的频率(relative frequency)：

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

其中 $f_n(A)$ 代表事件 A 在试验 n 次时的频率，f_A 与 $\sum f_i$ 分别代表在该试验中事件 A 的频数与各类事件频数之和. 显然任何事件的频率都在 0 与 1 之间，即

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

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；在重复次数非常多时，频率就逐渐稳定为一个确定的数值. 这个数值就是一个事件出现的概率

$$\lim_{n\to\infty}\frac{n_A}{n}\to P(A), \tag{1.3}$$

这是概率的频率定义. 显然，任何随机事件 E 的概率也在 0 与 1 之间

$$0\leqslant P(E)\leqslant 1. \tag{1.4}$$

概率必须是一个客观、确定的值，否则，以此为基础的概率论学科，包括统计学，将无科学性可言，其中的关键在于从不确定现象中寻找出确定的规律. 表 1.1 是投抛硬币的结果. 在一次实验中，一枚均匀硬币“面值朝上”的事件可能发生，也可能不发生，但在很多次试验中其出现的频率却很稳定，而且试验的次数越多，出现的频率越接近 1/2. 这说明，随机事件发生可能性的大小是事件本身固有的一种客观属性，不依人的主观意志为转移，称为该随机事件的概率.

表 1.1 抛掷硬币实验中面值朝上的频数和频率

组	抛掷数					
	20		200		2000	
	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

续表

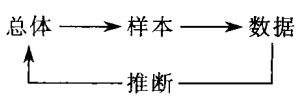
组	抛掷数					
	20		200		2000	
	频数	频率	频数	频率	频数	频率
8	6	0.30	101	0.505	1002	0.5010
9	9	0.45	101	0.505	976	0.4880
10	9	0.45	110	0.550	1018	0.5090
11	9	0.45	108	0.540	1021	0.5105
12	6	0.30	103	0.515	1009	0.5045
13	6	0.30	98	0.490	1000	0.5000
14	10	0.50	101	0.505	998	0.4990
15	13	0.65	109	0.545	988	0.4940

§1.2.3 总体与样本

对一批数据，首先要关注它是否为全部数据，抑或只是更大数据批的一部分. 为避免以偏概全做出错误的结论，在两者之间划出一条明确的界限是极为重要的. 总体(population)是在给定调查和试验中所有个体的总和，具有恒定的特征. 在条件相同的调查和实验中，个体(individual)显示的变异性称为随机变量(random variable)的取值(value). 这种变异性的根源在于随机变量本身，以及试验误差. 这样，总体就可定义为全部且有不同取值的个体. 它们不一定都是不同的，其数目也不一定是有限的. 样本(sample)是总体的一部分. 在特殊情况下，总体就是一个样本，也可以就是一个个体. 但通常人们总希望从样本的信息对总体进行推断. 因此，在一般情况总要定义所考虑的总体，并依据随机性原则从该总体得到一个有代表性的样本. 随机性原则体现了研究随机现象数量规律的概率论对取样工作的约束. 以下还要讨论两个有关的具体问题：

1. 总体与样本的关系

抛掷均匀硬币面值朝上的概率为 1/2，是对它的验前概率，即根据硬币的对称性在试验前进行的推断. 但多数随机事件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. 由于硬币铸造的图形和文字不对称等原因使硬币的质量中心与几何中心不重合，一枚硬币抛掷后落下面值朝上、朝下的概率不恰好都是 1/2 并非不可能. 这样，要寻求随机事件的规律，通常就需要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，限于人力、物力和时间，只能对随机事件进行次数有限的观测. 统计学所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所获知的统计特性，推断该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，某原料药厂某日生产的合成抗菌药磺胺嘧啶、某制剂厂某年生产的静注维生素 C 针剂、一个化验员为比较两个化验方法的精密度多年收集的全

部数据. 从总体随机抽取的一部分个体称为随机样本，简称样本. 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. 有时允许对研究总体中的每一个个体进行检验，如检查所有维生素 C 针剂的澄明度是否合格就是进行全检. 但在绝大多数情况下不能这样做. 例如，一个原料药厂不可能化验其生产的全部磺胺嘧啶，因为这种化验是破坏性的，全部化验完毕，生产的磺胺嘧啶也就全部化为乌有. 样本可以是事，也可以是物. 在样本是物的情况下，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.

2. 取样的随机性

既然统计学的中心任务是由样本特征推断总体特征，样本就必须能代表总体. 这样，统计学对取样就自然有一定的要求. 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. 在预测事件发生的概率时，只能从已发生事件的全部个体组成的总体取样. 这个总体不可能包括未发生事件的个体. 又如，药用植物麻黄在我国新疆、内蒙古自治区等地广泛分布，形成不同群落. 不同麻黄群落的各种手性麻黄碱含量有区别. 对这种情况，若要估计它们各自的总含量，就不可采用简单随机取样方法，而应采用分层随机取样方法. 这样的取样仍叫随机取样，但不叫简单随机取样；这样抽取的样本仍叫随机样本，但不叫简单随机样本.

根据样本的观测值推断总体，结论无论如何也不可能绝对正确. 统计方法所做结论的可靠性称为置信概率(confidence probability)、置信度或信度. 概率接近零的事件即小概率事件在一次试验中实际上是不可能发生的，可以看成是不可能事件(impossible event)，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. 概率论中这个原理对统计推断很重要，称为实际推断原理. 至于概率与 1 接近到何种程度该事件才可看成是必然事件，应视推断错误所产生后果的严重性而定. 后果越严重，概率应与 1 越接近.

取样过程未遵守随机性的规定，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进行无根据的选择，甚至假造，以达到个人不正当的目的，则已不属于统计学讨论的内容.

§1.2.4 观测值的特征——集中位置与离散程度

数据按其分布的特点而被归于正态分布、二项分布、泊松分布、指数分布和其他一些分布，其中多数是连续型的，少数是离散型的，如二项分布. 但数据无论属于哪种类型，都有其集中位置和离散程度两个特征.

1. 描述集中位置的统计量

(1) 均值 均值常指样本均值，是全部观测值之和除以观测次数所得的商：

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

求均值的目的，是要使多个观测值的正负随机误差相互抵偿；在观测过程无系统误差的条件下，观测值重复的次数越多，这种抵偿越充分，均值 $\bar{X}$ 越接近总体均值 μ. 总体均值 μ 也称观测值 X

的(数学)期望(expectation),记作

$$\mu=E(X)\approx\sum\text{取值}\cdot\text{概率}=\sum_i X_iP_i. \tag{1.6}$$

其中 P_i 为取值为 X_i 的概率.

均值有两个重要性质:(1)观测值与均值之差即偏差之和 $\sum_i (X_i-\bar{X})$ 为零;(2)偏差的平方和 $\sum_i (X_i-\bar{X})^2$ 最小. 这都不难用简单的代数运算证明. 偏差的平方和简称平方和或离差平方和(sum of squares),本书用大写斜体的 SS 表示.

例 1.1 T. Higuchi 实验室比较非水和重氮化滴定法测定磺胺乙基噻二唑样品的含量,得到表 1.2 中的数据并算出均值(%). 样品不同(B,C,D),但溶剂、指示剂相同的非水滴定结果的均值为 99.8. 样品相同(A),但溶剂、指示剂不同的非水滴定结果的均值为 100.2. 样品不同(B,C,D)的重氮化滴定结果的均值是 100.1.

表 1.2 非水和重氮化滴定磺胺乙基噻二唑[1]样品的结果

样品	溶剂	指示剂	磺胺乙基噻二唑			
			非水		重氮化	
			X_i	$\bar{X}$	X_i	$\bar{X}$
A	DMF[2]	TB[3]	100.2			
A	DMF	AV	99.8	100.2		
A	ED[2]	AV	100.7			
B	ED	AV	99.9		100.2	
C	ED	AV	99.7	99.8	100.1	100.1
D	ED	AV	99.8		100.1	

1) The Merck Index, 10th ed., 8777

2) DMF -N,N'-二甲基甲酰胺,ED -1,3-乙二胺

3) TB 麝香草酚兰,AV 偶氮紫

(2) 众数 在变量属于离散型的情况,计算得出的均值会是实际上不能取的数值. 这时,知道变量最常取的数值是有用的. 频数最大的观测值叫众数. 众数常用于表示离散型随机变量的集中位置.

例 1.2 一种药物在临床前按规定剂量给 10 组每组 10 名受试者服用. 各组显示有副作用的人数分别为 2,3,1,3,4,4,6,3,3,1. 在这里,每组中有副作用的人数作为随机变量是离散的,但它们在属性分组(有副作用和无副作用的两组)中患者人数的取值上可以根据其大小排列. 有副作用的属性分组的均值为 3 人,即每 10 名受试者中的 3 名有副作用,占 30%. 众数也是 3 人. 离差平方和 20 是最小的平方和,它表明各组有副作用的患者人数的离散程度.

(3) 中位数 均值的另一缺点,是由于受极端值的影响而给出集中位置的虚假印象. 例如,如果将某省自动化程度和管理水平很高的个别大制剂厂和这两方面水平都较低的众多小制剂厂放在一起计算它们的平均年产值,意义就不大. 因为这样的均值反映不了该省制剂厂生产的一般水平. 居民的年收入也不应采取大平均的方法计算,道理相同. 在这种情况下,中位数作为一种衡量集中位置的特征统计量,就具有优越性. 把变量的观测值按大小顺序排列,排在中间位置的一

个观测值叫中位数. 当观测次数 n 为奇数时，排在中间位置的只有一个；当 n 为偶数时，有两个. 在后一情况，中位数取为这两个数的平均值.

中位数还有容易确定的优点. 如果生产氖灯的工厂同时试验 5 个氖灯的寿命，则当第三个氖灯烧毁时，中位数即已确定. 在物理学中测量放射性同位素衰变速度的半衰期和在药理学中观察半数致死量 LD_{50}（使 50%试验动物死亡的剂量）时，都是利用了中位数的这个优点. 如果要计算一种放射性原子蜕变所需时间的均值，则必须等样本中所有原子都蜕变完毕. 这显然是行不通的. LD_{50} 虽不是时间而是剂量，也是一个中位数.

2. 描述数据离散程度的统计量

数据的离散程度可用极差、平均偏差、方差和标准差量度，其中最重要的是方差.

(1) 极差　极差也称范围(range)，在描述一组数据离散程度的多种统计量中是最简单的. 极差是一组观测值中的最大值与最小值之差：

$$R=X_{\max}-X_{\min}. \tag{1.7}$$

表 1.3 中所列的，是服用一种降胆固醇药后血清中胆固醇含量的变化(mg/100 mL). 其极差 $R=55-(-97)=152$(mg/100 mL).

表 1.3　150 名患者服用某种降胆固醇药后血清中胆固醇含量的变化[1)]

−10	11	11	−39	19	−32
4	−15	−18	35	6	20
46	24	−27	−19	5	−60
27	23	−22	−1	12	−27
−13	−39	39	−34	−97	−26
38	14	−47	8	16	−15
−62	12	−53	11	21	−47
−54	−11	−5	0	55	34
−69	−11	−44	20	−50	19
0	−25	−24	−4	14	2
−34	16	−23	−71	−58	9
9	2	−2	−58	13	14
17	−13	−22	−3	−17	1
17	−12	25	−37	−29	−39
−22	0	−22	−63	34	−31
−64	−12	−49	5	−8	33
−50	−7	16	−11	−38	−17
0	−9	−21	1	2	−30
−32	−34	−14	−18	5	6
24	−6	−49	−8	−49	−37
−25	−12	14	10	−41	−66
−31	35	21	−19	−27	17
−6	−17	−6	1	−28	40
−31	17	−54	−27	−16	16
−44	10	−3	−3	5	6
−19	9	−10	−20	−9	−8

1) −为减小，+为增大

极差在控制图中用得较多. 这是因为极差计算简单,从极差的大小便可知道产品质量的波动情况.

简单是极差的优点,也是它的缺点. 顾名思义,极差不考虑绝大多数观测值如何接近,因而不能充分利用数据提供的关于其离散情况的全部信息.

(2) 平均(绝对)偏差 另一个用于量度一组数据离散程度的统计量,与诸观测值关于均值的偏差有关. 已知偏差之和为零,其均值也为零. 这个困难可以通过对偏差取绝对值解决:

$$d=\frac{1}{n}\sum_{i=1}^{n}|X_i-\bar{X}|. \tag{1.8}$$

习惯上,把平均绝对偏差叫平均偏差.

(3) 方差 对偏差取绝对值,似乎已使正负偏差在求和中抵偿的问题得到解决,但由于给以后的计算带来困难而不够成功. 数学在处理模量(绝对量)函数上麻烦. 由于$|X-\bar{X}|$在$X>\bar{X}$时是指$X-X$,而在$X<\bar{X}$时是指$X-X$,就这个模量而言,就必须考虑X在全部取值范围的这两个分组. 两全其美的方法,是以求$\sum_i(X_i-\bar{X})^2$代替求$\sum_i|X_i-\bar{X}|$. 这样,就避免了上述两个麻烦. 偏差平方和的均值称为方差,其表达式为:

$$S^2=\sum_{i=1}^{n}\frac{(X_i-\bar{X})^2}{n-1}. \tag{1.9}$$

这是通过n个观测值求得的样本方差. 观测值X的样本方差也用 var(X)表示(注意 var 通常用正体,它是 variance 的缩写).

重复观测的次数越多,样本方差S^2越接近总体方差σ^2. σ^2也称S^2的数学期望,当变量为有限的离散总体时,也可记作

$$\begin{aligned}\sigma^2&=\sum_{i=1}^{n}(X_i-\mu)^2P_i=E[(X-\mu)^2]\\&=E(X^2)-2\mu E(X)+\mu^2=E(X^2)-\mu^2.\end{aligned} \tag{1.10}$$

式(1.10)的意义,是把X的方差σ^2定义为$(X-\mu)^2$的数学期望或X^2的数学期望与μ^2之差. 这就是方差的词义. 为计算此值,可将$(X-\mu)^2$的诸取值$(X_i-\mu)^2$分别乘以它们的概率$P(X_i)$即P_i后再求和. 式(1.10)与(1.9)的联系是明显的. 随着n的增大,$\bar{X}$逼近μ,同时n与$n-1$的区别趋于消失.

按式(1.5)求得的$\bar{X}$是μ的无偏估计量. 但如果把式(1.9)的分母$n-1$换为n,则求得的S^2不是σ^2的无偏估计量. 用$\bar{X}$估计μ有时估计得高一些,有时低一些,但平均起来既不高,也不低,所以称X为μ的无偏估计量. 在估计总体方差时,自然想用统计量$\frac{1}{n}\sum(X_i-\bar{X})^2$. 但研究结果表明,用它估计的$\sigma^2$平均起来要小一些,是$\frac{n-1}{n}\sigma^2$而不是$\sigma^2$. 这个偏倚可通过用$\frac{n}{n-1}$乘以$\frac{1}{n}\sum(X_i-\bar{X})^2$进行校正,于是得到$\frac{1}{n-1}\sum(X_i-\bar{X})^2$. 这样求出的$S^2$,才是$\sigma^2$的无偏估计量.

式(1.9)中的$n-1$在统计学中叫自由度(degree of freedom),常用希腊字母ν表示,也有用 df 表示的. 自由度的概念可以这样说明:由于$\sum_i(X_i-\bar{X})=0$,前$n-1$个偏差都能自由选择,但第n个$(X_n-\bar{X})$不能,而是固定的: