中国科学院教材建设专家委员会规划教材
全国高等医药院校规划教材

供药学、药物制剂、临床药学、中药学、制药工程、医药营销等专业使用

案例版™

医药数理统计

主 编 张丕德 马洪林

科学出版社

中国科学院教材建设专家委员会规划教材
全国高等医药院校规划教材
供药学、药物制剂、临床药学、中药学、制药工程、医药营销等专业使用

案例版™

医药数理统计

主　　编　张丕德　马洪林
副 主 编　陈　群　王在翔　陈　征
编　　委　（按姓氏拼音排序）
曹红艳　山西医科大学
陈　群　宁夏医科大学
陈　征　南方医科大学
楚慧珠　广东药科大学
崔　凯　锦州医科大学
金英良　徐州医科大学
李彩艳　包头医学院
李望晨　潍坊医学院
林剑鸣　广州中医药大学
马　勇　包头医学院
马洪林　锦州医科大学
宋桂荣　大连医科大学
王在翔　潍坊医学院
袁　晶　宁夏医科大学
张丕德　广东药科大学
赵　军　湖北医药学院
赵华硕　徐州医科大学
周舒冬　广东药科大学
庄　严　南方医科大学

科学出版社
北　京

郑 重 声 明

为顺应教育部教学改革潮流和改进现有的教学模式，适应目前高等医学院校的教育现状，提高医学教育质量，培养具有创新精神和创新能力的医学人才，科学出版社在充分调研的基础上，引进国外先进的教学模式，独创案例与教学内容相结合的编写形式，组织编写了国内首套引领医学教育发展趋势的案例版教材。案例教学在医学教育中，是培养高素质、创新型和实用型医学人才的有效途径。

图书在版编目（CIP）数据

医药数理统计 / 张丕德，马洪林主编，—北京：科学出版社，2018.1

中国科学院教材建设专家委员会规划教材·全国高等医药院校规划教材

ISBN 978-7-03-052786-8

Ⅰ. ①医… Ⅱ. ①张… ②马… Ⅲ. ①医用数学–数理统计–医学院校–教材 Ⅳ. ①R311

中国版本图书馆 CIP 数据核字（2017）第 102581 号

责任编辑：王 超 胡治国 / 责任校对：郭瑞芝

责任印制：赵 博 / 封面设计：陈 敬

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

北京市密东印刷有限公司 印刷

科学出版社发行 各地新华书店经销

*

2018 年 1 月第 一 版 开本：787×1092 1/16

2018 年 1 月第一次印刷 印张：16 1/2

字数：466 000

定价：55.00 元

（如有印装质量问题，我社负责调换）

前　言

教材建设是高等教育的关键环节，教学改革首先从教材改革做起，为了满足当前国内教育改革的需要，科学出版社率先组织编写“案例版”系列教材，本编委有幸承担《医药数理统计》的编写工作。

本书虽然重在应用，但又有很强的数理逻辑性，对高等数学较薄弱的医药类学生而言，如果一味讲授数学定理、方法，就显得枯燥乏味、抽象难懂；如果只讲应用，完全避开定理、公式的解析和证明，又很难对统计方法有较为深刻的理解。因此，本书编写的思路是：以案例为导引，提出实际问题，叙述相关统计背景，逐步展开具体统计方法的数理描述，深入浅出，解决问题，将实际应用与理论方法紧密结合，便于读者把实际问题与具体方法直观对应，遇到同类问题可立即处理，并达到举一反三的效果。为不冲淡主题，保证内容的完整性，把一些数学证明单独列出，放到每一章的后面，供读者进一步深入了解。由于统计应用中，需要大量的数据分析，适当增加软件实习，有助于提高学生的统计应用能力。同时，建议读者在学习每章后，把本章的主要内容进行总结，便于掌握知识要点。对一些经典的理论和方法，查阅相关的统计历史，有助于提高学习兴趣，也可以得到更多的启发。

本书编写过程中，按照教育部相关要求和各高校教学实际选择内容，注重基础性，突出实用性，追求创新性。在内容编排方面，本书每一章，开头增加学习要求，突出重点，例题与习题的软件操作及运算结果，请按配套的实习指导教材进行。对于不同学校规定的教学时数，如总学时为 36 学时、45 学时、54 学时，可以对全书内容选择教学。

本书是全体编委集体劳动的结晶，考虑篇幅所限，附录只列常用统计表和参考文献两部分。参考文献仅列出一部分经典文献和相似教材，我们尊重所有同类教材与专著的编写工作。尽管编委们来自不同的高校，各自有丰富的教学经验，但编辑过程中难免出现纰漏或不足，欢迎读者批评指正。

编　者

2016 年 11 月

目　录

第一章 描述统计

学习要求

1. 掌握：数据的类型；频数表与频数图的制作；描述分布集中趋势和离散趋势的常用指标与计算。

2. 熟悉：统计表与统计图的制作。

3. 了解：描述分布形状的指标。

概率论（probability）是一门研究客观世界随机现象数量规律的数学分支学科。16 世纪意大利学者开始研究掷骰子等赌博中的一些问题，17 世纪中叶，法国数学家帕斯卡（B.Pascal，1623～1662）、荷兰数学家惠更斯（C.Huygens，1629～1695）基于排列组合的方法，研究了较复杂的赌博问题，解决了“合理分配赌注问题”（即得分问题），开创了概率论研究的新纪元。

对客观世界中随机现象的分析产生了概率论，使概率论成为数学的一个分支的真正奠基人是瑞士数学家伯努利（J.Bernoulli，1700～1782）；而概率论的飞速发展则在 17 世纪微积分学说建立以后。

第二次世界大战军事上的需要以及大工业与管理的复杂化产生了运筹学、系统论、信息论、控制论与数理统计学等学科。

数理统计（mathematical statistics）是一门研究怎样去有效地收集、整理和分析带有随机性的数据，以对所考察的问题作出推断或预测，直至为采取一定的决策和行动提供依据和建议的数学分支学科。

其主要内容包括：①统计设计，即根据研究目的确定研究对象、研究范围和样本获取的方式与大小等，通常分为调查设计与试验设计；②数据搜集，即取得统计数据，是进行统计的基础；③数据整理，即用图表等形式来展示数据特征，使数据更加系统化、条理化，便于进一步统计分析；④数据分析，就是用统计方法来研究数据，是统计学的核心部分；⑤数据解释，即对统计分析结果进行说明和应用。

统计方法的数学理论要用到很多近代数学知识，如函数论、拓扑学、矩阵代数、组合数学等，但关系最密切的是概率论，故可以这样说：概率论是数理统计学的基础，数理统计学是概率论的一种应用。

数理统计学分为：描述统计学，即对随机现象进行观测、试验，以取得有代表性的观测值；推断统计学，对已取得的观测值进行整理、分析，作出推断、决策，从而找出所研究的对象的规律性。本章着重介绍描述统计学。

第一节 数据的基本类型

一、数据的分类

数据（data）也称为资料，是对客观事物的某种属性计量的结果。根据研究目的，对研究对象的某个或某些特征（亦称研究指标）实施观察，这些特征（指标）称为变量（variable），变量的观

测值（即变量值）构成数据或资料。例如，统计某年中国恶性肿瘤的发病情况，可以按照地区、性别、年龄、肿瘤发病率等指标；对药品质量的计量可以得到药品是合格或不合格的数据。由于对事物计量的精确程度不同，得到的数据类型也会有所不同。对数据进行正确的分类、合理的统计是利用统计方法进行分析的基础。

根据不同的理解和原则，统计资料有不同的分类方法。本书将资料分为计量资料（measurement data 或 quantitative data）、计数资料（count data 或 qualitative data）和等级资料（ordinal data）。不同类型的统计资料需要用不同的统计方法进行分析，因此，正确判定变量的类型十分重要。

（一）数据类型及定义

1. 计量数据 是用定量的方法对每一个观察单位的某项指标进行测定所得的资料。计量资料的变量值是定量的，表现为数值大小，一般具有度量衡单位。如表 1-1 中年龄、病程资料等。

表 1-1 某药物治疗成人急性气管炎疗效的观测结果

病例号	年龄/岁	性别	职业	血型	病情	病程/天	临床治疗效果
0001	35	男	工人	A	重	8	有效
0002	42	女	教师	O	轻	3	显效
0003	25	女	学生	B	中	6	无效
0004	33	男	营业员	AB	重	10	有效
0005	27	女	公务员	O	轻	4	治愈
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
0102	45	男	工人	O	中	7	有效

2. 计数数据 是将观察单位按照某种属性分组计数的定性观察结果。计数资料分为二分类资料和无序多分类数据。二分类资料，比如表 1-1 中患者的性别分为男、女等；无序多分类资料，比如表 1-1 中患者的血型分为 O 型、A 型、B 型和 AB 型等。计数数据又称为定类数据（categorical data）或名义数据（nominal data）。

3. 等级数据 是将观察单位按某种属性的不同程度或次序分成等级后分组计数的观察结果，也称为有序多分类数据。例如，表 1-1 中患者的临床治疗效果分为无效、有效、显效、治愈，效果依次递增。临床研究中的等级资料较为多见，如尿蛋白的临床检验结果为-、±、+、++、+++5 个等级，且 5 个等级的尿蛋白量从无到有，从少到多。等级数据又称为定序数据或有序数据。

有时为了研究需要或者是数据分析的需要，需要将资料进行转换。一般情况下将计量资料转化成有序多分类资料或者是二分类资料。例如，血红蛋白水平为计量资料，为了统计某资料中孕妇为正常和贫血的人数，需要将资料按照“联合国儿童基金会推荐的贫血判断标准（2001 年），孕妇血红蛋白水平低于 110g/L 为异常”，将孕妇分为正常和贫血两类，分别统计她们的人数。

（二）变量及其类型

1. 变量（variable） 统计学中称将客观事物的某种属性或标志称为变量，对变量进行观察或测量所得到的值称为观察值（observation）或变量值（variable value）。统计数据就是变量的观察值。

2. 变量的分类 可根据记录变量的数据类别将变量进行相应的分类。记录的数据形式为计量数据、计数数据和等级数据时，相应的变量可分别称为计量变量（measurement variable）或数值变量（numerica variable）、计数变量（count variable）或定类变量（categorical variable）、名义变量（nominal variable）、等级变量（rank variable）或定序变量（ordinal variable）。

在实际中，应用最多的是计量变量，故将它简称为变量。通常所说的变量主要是计量变量，而大多数统计方法所处理的也都是计量变量。

3. 计量变量的分类 根据计量变量的取值可将它分为两种类型：离散型变量（discrete variable）和连续型变量（continuous variable）。离散型变量通常只取整数值。例如，一个月中的门诊人数，某医院一年的新生儿数。连续型变量可以取实数轴上的任何数值，如身高、体重、白细胞数等，它们通常可通过测量得到。

根据计量变量的取值可将它进一步分类，若其可以取值为有限个或无穷可列个，则称其为离散变量；若其可以取值为无限且不可列个，则称其为连续变量。在实际应用时，当离散变量的取值非常多时，也可以近似当作连续变量来处理。

区分数据和变量的类型非常重要，关系到进一步采取何种统计方法进行处理和分析，表 1-2 给出了它们之间的比较。

表 1-2 不同数据类型之间的比较

数据类型	定性数据（品质数据）		定量数据
	计数数据（定类数据）	等级数据（定序数据）	计量数据（数值数据）
表现形式	类别（无序）	类别（有序）	数值
对应变量	计数变量	等级变量	计量变量（离散变量、连续变量）
主要统计方法	计算各组频数，进行列联表分析、χ^2 检验等非参数方法		计算各种统计量，进行参数估计和检验、回归分析、方差分析等参数方法

（三）两类数据的转换

为了数据分析的方便，有时候需要对变量进行转化。但变量转化只能由“高级”向“低级”转化：定量→有序→分类→二值。例如，考试成绩可由百分制转化为五等级制：90～100 分为“优”、80～89 分为“量”、70～79 分为“中”、60～69 分为“及格”、59 分以下为“不及格”，成年男子的血清胆固醇按是否小于 6（mmol/L）划分成血脂正常和异常两类，即将定量数据转化为定性数据。

（四）统计数据的搜集和来源

对统计数据的搜集要做到完整、准确、及时、可靠。医药科学研究的数据主要来源于 3 个方面：①日常工作记录：包括病例、卫生监测记录、药物反应记录等，这些记录还没有经过研究设计，可能会产生不完整和不准确的情况，要注意避免。②统计报表：包括工作报表、统计年鉴、疫情报表等，这些数据的准确性取决于填报人员的业务素养，使用时应对其作出判断。③专题调查和试验：这类数据一般经过严格的设计过程，但也应注意数据搜集过程的质量监控和审核。无论以何种方式收集数据，都要强调它的准确性和完整性。

二、数据的整理

当数据的搜集工作完成后，要对数据资料进行科学的汇总和处理，使其系统化，初步反映出研究总体的特征、规律和趋势。在对数据进行整理时，首先要进行审核筛选，以保证数据的质量，然后确认数据的类型，有针对地进行处理，对定性数据和定量数据分别作出分类整理和分组整理。

（一）定性数据的整理

频数（frequence 或 frequency）是指分布在各组中的数据个数；频率（relative frequence）是指分布在各组中的数据个数占数据总个数的比例值；将各组类别及相应的频数（频率或百分比）用表格形式全部列出就是频数分布表（frequency table），如表 1-3 所示。

表 1-3 2012 年我国各地区三级医院频率分布

地区	三级医院		二级医院		一级医院	
（1）	数量（2）	频率/%（3）	数量（4）	频率/%（5）	数量（6）	频率/%（7）
东部	726	46.6	2029	33.8	1061	37.8
中部	443	28.4	2018	33.7	1052	37.5
西部	389	25.0	1948	32.5	693	24.7
合计	1558	100.0	5995	100.0	2806	100.0

数据来源：中国统计年鉴 2013. 中国统计出版社，33.

表 1-3 中第（1）列“地区”是一个三分类变量，第（3）列是 1558 个三级医院不同地区的频率分布。其中，东部的三级医院频率分布最高，为 46.6%；其次分别为中部、西部。频率分布的特点是，定性变量各类别的频率之和为 100%。

定性数据自身的表现形式就是类别化，进行整理时，只需按不同类别分组，算出各组频数或频率、百分比，列出频数分布表，再用条形图或圆形图等统计图直观地显示其整理结果。如果是定序数据，还可以算出各组累积的频数或频率、百分比。

我们将在本章第三节介绍，在表 1-3 的基础上作出统计图，使其更直观地表示分布状况。

（二）定量数据的整理

案例 1-1

某年抽样调查某地 120 名健康成年女性的红细胞数，如表 1-4 所示。

问题：

（1）案例属于什么类型的资料？

（2）用什么方法进行处理，可以使上述数据更直观、更系统？

表 1-4 某地 120 名健康成年女性的红细胞数 （单位：$\times 10^{12}$/L）

5.12	4.45	4.07	3.58	4.41	4.03	4.22	3.53	4.69	4.62
4.56	3.99	4.28	3.10	4.98	3.37	4.01	3.59	4.20	4.13
4.17	4.47	4.31	4.31	4.04	4.37	4.10	5.05	4.68	4.64
4.64	4.14	5.45	4.00	4.95	4.18	4.44	3.64	3.96	4.12
4.61	3.86	4.63	3.67	4.37	4.40	4.35	4.33	4.20	4.43
4.29	4.16	4.41	4.49	4.29	4.07	4.42	4.31	4.10	4.82
4.34	3.99	4.02	3.32	4.45	4.08	3.23	4.58	3.82	4.16
4.24	4.28	4.41	4.29	4.23	3.22	4.00	4.26	4.36	4.28
3.99	4.24	4.39	4.16	4.31	4.17	4.54	5.07	3.74	4.03
4.42	4.01	4.11	3.12	4.03	4.88	3.87	4.61	4.84	4.36
3.73	4.67	4.68	4.34	3.69	4.47	4.13	4.08	3.44	4.33
4.23	4.28	4.22	4.42	4.13	3.85	4.05	4.18	4.42	4.67

分析讨论：

本案例中，健康成年女性的红细胞数属于定量数据。通过对定量数据的整理了解其分布规律和类型，进而选用合适的统计指标描述其分布的集中趋势、离散趋势。一般主

要按数据的数量标志进行分组、编制频数分布表，必要时采用直方图及频数折线图等统计图形来表示其整理结果，使其频数分布状态更加直观清晰。

下面结合本案例介绍频数分布表的具体编制。

第一步：求极差（全距）。

极差（range，R）又称为全距，为变量的最大值与最小值之差。

$$R=\text{Max}-\text{Min}=5.45-3.10=2.35\ (\times 10^{12}/\text{L})$$

第二步：确定组数、组距和组段。

组数 k 的确定是由数据本身的特征和个数 N 来确定的，实际工作中常常采用等距分组，变量值个数较多时，组段数一般取 10 左右，且可用极差/10 来估计组数。

在本案例中，N=120，则 k=120/10=12，可考虑大致分为 12 组。

组距 d 是指每组上限与下限之差，且将每组的最小值称为该组的下限（lower limit），每组的最大值称为该组的上限（upper limit）。

$$\text{本案例中，组距}=\frac{\text{极差}(R)}{\text{组数}(k)}=\frac{2.35}{12}=0.196\approx 0.2(\times 10^{12}/\text{L})$$

为了计算方便，组段下限一般取较整齐的数值。各组段要连续但不能重叠，除了最后一组，各组段只包含下限值，不包含上限值。但要注意，第一组段应包含最小值，最后一个组段要包含最大值。

第三步：计算频数，编制频数分布表。

对数据进行分组，计算各组频数、频率、累计频数、累计频率，见表 1-5。观察数据某一数值以下（或以上）的频数或频率之和，称为累积频数（cumulative frequence）或累积频率（cumulative frequency）。

对于组距分组还应该注意以下几个问题：

（1）要做到“不重复不遗漏”，当前一组的上限与后一组的下限重叠时，一般规定“组上限不在本组内”，只有最后一组包括上限。例如表 1-5 分组中“3.10～”表示[3.10,3.30]，即上限 3.40 不计入该组。

表 1-5 某地 120 名健康成年女性的红细胞数（$\times 10^{12}$/L）频数表

组段	组中值	频数	频率/%	累计频数	累计频率/%
3.10～	3.20	4	3.33	4	3.33
3.30～	3.40	3	2.50	7	5.83
3.50～	3.60	6	5.00	13	10.83
3.70～	3.80	6	5.00	19	15.83
3.90～	4.00	18	15.00	37	30.83
4.10～	4.20	31	25.83	68	56.67
4.30～	4.40	29	24.17	97	80.83
4.50～	4.60	14	11.67	111	92.50
4.70～	4.80	3	2.50	114	95.00
4.90～	5.00	4	3.33	118	98.33
5.10～	5.20	1	0.83	119	99.17
5.30～5.50	5.40	1	0.83	120	100.00
合计	—	120	100.00	—	—

（2）为了避免遗漏极端值，第一组和最后一组可采用开口组，如“3.10以下”和“5.30以上”的形式。开口组需要确定组距时可以与相邻组等同。

（3）为了满足特定需要，有时也采用不等距分组。例如，对人口年龄的分组，为了兼顾人口生理特点，可分为0～6岁（婴幼儿组）、7～17岁（少儿组）、18～59岁（中青年组）、60岁以上（老年组）的不等距分组。

（4）为了反映一组数据的一般水平，引入组中值（middle point value）的概念，即

$$\text{组中值}=\frac{\text{上限值+下限值}}{2}$$

组中值作为该组数据的代表值，在利用频数分布表进行均值、方差等计算或作频数折线图时将起到重要作用。

第二节　数据的统计描述

通过上节内容的学习，我们了解了数据的分类和数据的整理，通过频数分布表，初步反映数据分布的特征。为了进一步对数据的分布特征和规律进行全面掌握和定量刻画，则需要了解更多的、从不同侧面反映数据分布特征的统计指标即统计量。

本节将从集中趋势、离散趋势两个方面定量描述数据的平均水平和变异程度，并且介绍描述数据分布形态的指标。

一、数据集中趋势的描述

集中趋势（central tendency）指的是一个计量资料的大多数观察值所在的中心位置。常用的描述集中趋势的统计指标主要有算术均数、几何均数和中位数。

（一）算数均数

1. 算数均数（arithmetic mean）　简称为均数（mean），等于一个指标变量所有观察值的和除以观察值的个数。适用于服从对称分布变量的平均水平的描述，这时均数位于分布的中心，能反映一个变量所有观察值的平均水平。

一般地，总体均数用希腊字母 μ 表示，样本均数用 $\overline{X}$ 表示。

2. 均值的计算　可以由原始数据直接计算均值：

设原始数据为 $X_1, X_2, \cdots, X_n$，均值的计算公式：

$$\overline{X}=\frac{X_1+X_2+\cdots+X_n}{n}=\frac{1}{n}\sum_{i=1}^{n}X_i \qquad (1\text{-}1)$$

案例 1-2

某人调查了13名初中女生的身高，其测量值为：146.6，150.2，165.3，159.2，150.6，152.3，156.4，162.3，160.2，155.4，156.2，148.2，150.2，计算这13名初中女生身高的样本均数。

按照公式（1-1），可以得到

$$\overline{X}=(146.6+150.2+165.3+159.2+150.6+152.3+156.4+162.3+160.2+155.4+156.2+148.2+150.2)/13=154.9(\text{cm})$$

也可以利用频数和组中值近似计算频数：

对分组整理的数据，分组数为 k，各组数据出现的频数分别为 f_1，f_2，…，f_k，当然

$\sum_{i=1}^{k} f_i = n$，各组的组中值为 $m_1, m_2, \cdots, m_k$ 时，均值的近似计算公式：

$$\overline{x} \approx \frac{m_1 f_1 + m_2 f_2 + \cdots + m_k f_k}{f_1 + f_2 + \cdots + f_k} = \frac{1}{n}\sum_{i=1}^{k} m_i f_i \qquad (1\text{-}2)$$

由表 1-5 可计算案例 1-1 的算数均数为

$$\overline{X} = \frac{3.20\times 4 + 3.40\times 3 + \cdots + 5.40\times 1}{120} = 4.22(\times 10^{12}/\mathrm{L})$$

上式虽然得到的是成绩均值的近似值，但近似程度已经很好，而且也很大程度地减小了计算量。

3. 均值的性质 均值是一组数据差别相互抵消的结果，所以是数据的重心所在，是进行统计分析和统计推断的基础，具有以下良好的数学性质：

（1）各数据与均值的离差之和为零，即

$$\sum_{i=1}^{n}(x_i - \overline{x}) = 0$$

（2）各数据与均值的离差之平方和为最小值，即

$$\sum_{i=0}^{n}(x_i - \overline{x})^2 \leqslant \sum_{i=1}^{n}(x_i - a)^2 \quad （a 为任意实数）$$

由此可见，均值是误差最小的总体数据的代表值，特别当数据分布为对称或近似对称时，均值是数据集中趋势的最好代表值。但是当数据分布偏斜程度较大时，均值也容易受到极端值的影响，不能很好地反映数据的集中趋势，也可采用截尾均值，即在数据中先去掉若干最小值和最大值，再进行均值计算。这在某些评奖、比赛中较为常见。

（二）几何均数

几何均数（geometric mean，G）是一个变量的所有 n 个观察值乘积的 n 次方根。其计算公式如下：

$$G = \sqrt[n]{X_1 X_2 \cdots X_n} = \lg^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\log X_i\right) \qquad (1\text{-}3)$$

式（1-3）中的 $\log X_i$ 表示对 X_i 求对数，其计算可以采用以 10 为底数（记为 lg），也可采用以自然数 e 为底数（记为 ln）。该公式中的 $\lg^{-1}$ 是取以 10 为底数或以 e 为底数的反对数。注意，观察值中不能包含小于等于零的数据。

案例 1-3

某社区对 8 名儿童进行免疫接种，日后测得其抗体滴度分别为 1∶4，1∶8，1∶16，1∶16，1∶32，1∶64，1∶64，1∶128，求平均滴度。

按照式（1-3），几何均数为

$$G = \sqrt[8]{4\times 8\times 16\times 16\times 32\times 64\times 64\times 128} = 24.68$$

或者

$$G = \lg^{-1}\left(\frac{\lg 4 + \lg 8 + \lg 16 + \lg 16 + \lg 32 + \lg 64 + \lg 64 + \lg 128}{8}\right) = \lg^{-1} 1.39 = 24.68$$

（三）中位数

1. 中位数（median） 是将全部数据按从小到大排序后处于中间位置的数值，记为 M。中位数将全部数据分为两个相等部分，上下各有一半数据值。显然中位数只适用于计量数据和计数数据，不能用于等级数据。

2. 中位数的确定

（1）直接法（基于原始数据）：

将全部数据 $x_1, x_2, \cdots, x_n$，按从小到大的顺序排列后为 $x_{(1)}, x_{(2)}, \cdots, x_{(n)}$（其中 $x_{(1)} \leqslant x_{(2)} \leqslant \cdots \leqslant x_{(n)}$），则中位数为

$$M = \begin{cases} x_{\frac{n+1}{2}}, & \text{当}n\text{为奇数} \\ \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right), & \text{当}n\text{为偶数} \end{cases} \tag{1-4}$$

即中位数位于位置 $\frac{n+1}{2}$，当 n 为奇数时，数据的中间值取作中位数；当 n 为偶数时，排在中间的两个数据的平均值取作中位数。

案例 1-4

某人调查了 13 名初中女生的身高，其测量值为：146.6，150.2，165.3，159.2，150.6，152.3，156.4，162.3，160.2，155.4，156.2，148.2，150.2，计算这 13 名初中女生身高的中位数。

因为 n=13 是奇数，按照从小到大排序后，最中间的数据为第 7 个数据 155.4，所以公式为

$$M = M_{\frac{13+1}{2}} = M_7 = 155.4(\text{cm})$$

另外，低于已分组的数据，可以很容易地确定中位数所在组，即累积频数超过 $\frac{n}{2}$（或累积频率超过 0.5）的最低组。例如，由表 1-5 可知累积频数超过 $\frac{n}{2}=60$ 的最低组为 4.10～组，则中位数在 4.10～组。

（2）频率表法（基于频率表资料）：可以通过计算百分位数来近似计算中位数。百分位数（percentile，P_x）是指将 n 个观察值从小到大依次排序后，对应于 x%位的数值；表示将原始观察值分为两部分，理论上有 x%的观察值小于 P_x，有（100–x）%的观察值大于 P_x，百分位数 P_{50} 就是中位数。

对于频率表资料，百分位数 P_x 的计算公式为

$$P_x = L + \frac{i}{f_x}(n \times x\% - F_L)$$

其中，L 为百分位数所在组段的下限；i 为该组段的组距；f_x 为该组段的频数；n 为总频数；F_L 为小于 L 所在组段的累计频数。

案例 1-5

某医院 2015 年治疗某病的费用（千元），如表 1-6 所示。

表 1-6 某医院 2015 年治疗某丙的费用

组段/千元	频数（f_i）	累计频数	累计频率/%
0～	9	9	2.75
1～	53	62	18.96
2～	68	130	39.76
3～	60	190	58.10
4～	50	240	73.39
5～	39	279	85.32
6～	21	300	91.74
7～	14	314	96.02
8～	5	319	97.55
9～	3	322	98.47
10～	5	327	100.00
合计	327	—	—

$$M = P_{50} = 3 + \frac{1}{60}(327 \times 50\% - 130) = 3.56 \text{（千元）}$$

3. 中位数的特点 中位数是数据的位置平均数，其特点与均值相比较不受极端值影响，特别是在存在开口组时，在描述集中趋势时比均值更为贴切。但其计算功能较差，敏感度也不足。

（四）众数

众数（mode）即数据中出现次数最多的数据值，用 M_0 表示。主要用于描述定性数据的集中趋势，对于定量数据，可能有多个众数或没有众数，意义不大。

众数的特点是较直观、易理解、不受极端值影响，但灵敏度、计算功能和稳定性差，存在着不唯一性，当数据集中趋势不明显或有两个以上分布中心时不宜使用。

二、数据离散趋势的描述

反映数据集中趋势的统计量，体现了数据的中心值、代表值，是对数据的概括性度量，但它对数据一般水平代表性的好坏取决于数据的离散程度，也就是各数据值偏离其中心值的程度。同一总体中不同个体间存在的差异称为变异（variation）。不同的观察指标，其变异是不同的；即使是同一观察指标，在不同总体中，其变异程度也有所不同。

统计学中常用以描述数据离散程度的统计量有极差、四分位数间距、方差、标准差、变异系数等。

（一）极差

极差（range）又称全距，是一组数据最大值与最小值之差，用 R 来表示，即

$$R = \text{Max} - \text{Min} \tag{1-5}$$

样本量接近的同类资料相比较时，极差越大意味着数据越离散，或者说数据间变异越大。

案例 1-6

根据案例 1-4 给出的数据，计算这 13 名初中女生身高的极差。

因为这 13 名初中女生身高的最大值 Max =165.3cm，最小值 Min =146.6cm，由式（1-5）得，$R = \text{Max} - \text{Min} = 165.3 - 146.6 = 18.7(\text{cm})$。

由于极差的计算只利用了两个极端值，而且往往样本量越大，极差越大，所以一般不太直接用极差描述离散趋势。

（二）分位数和四分位数间距

1. 分位数（quantile） 就是将全部数据按从小到大排序后等分，位于等分点上的数据值。统计学中常将数据四等分和百等分，相应产生四分位数和百分位数。

四分位数（quartile）：就是用 3 个点将从小到大排序后的全部数据四等分后在分位点上的数值，也被称作四分位点。

3 个四分位点各自有它们的名称：第一个等分点称为下四分位数（lower quantile），记为 Q_1；第二个等分点就是中位数 M，记为 Q_2；第三个等分点称为上四分位数（upper quantile），记为 Q_3。

四分位数的计算与中位数相似，即先对数据进行排序，再确定其位置，然后确定其数值。

$$Q_1 \text{ 位置} = \frac{1}{4}(n+1)，\ Q_2 \text{ 位置} = \frac{1}{2}(n+1)，\ Q_3 \text{ 位置} = \frac{3}{4}(n+1)$$

当四分位数的位置不是整数时，应该根据其位置按比例分摊两侧数值的差值。

案例 1-7

某医院 2015 年治疗某病的费用（千元），如表 1-6 所示。

Q_1 位置 $= \frac{1}{4}(327+1) = 82$，则 Q_1 在第 3 组段，即 2～，故

$$Q_1 = 2 + \frac{1}{68}(327 \times 25\% - 62) = 2.29$$

Q_3 位置 $= \frac{3}{4}(327+1) = 246$，则 Q_3 在第 6 组段，即 5～，故

$$Q_3 = 5 + \frac{1}{39}(327 \times 75\% - 240) = 5.13$$

事实上，$P_{25} = Q_1; P_{50} = M; P_{75} = Q_3$。

2. 四分位数间距（quartile range） 四分位数间距或四分位差、内距，即上四分位数与下四分位数之差，记为 Q_d。计算公式为

$$Q_d = Q_3 - Q_1 \qquad (1\text{-}6)$$

四分位数间距充分反映了中间 50% 数据的离散程度，其数值的大小，体现出中间的数据的集中与分散。并且不受极端值的影响，大大克服了用极差描述数据离散程度的不足。但它不适用于等级数据。

对于案例 1-7 的数据，四分位数间距为

$$Q_d = Q_3 - Q_1 = 5.13 - 2.29 = 2.84$$

（三）方差和标准差

方差（variance）即各数据观测值与均值的离差的平方的算术平均值，是反映计量数据离散程度的最重要的统计量，方差的算术平方根就是标准差（standard deviation）。根据观察数据的不同，方差又分为总体方差和样本方差。

1. 总体方差和标准差 当观察数据 $x_1, x_2, \cdots, x_n$ 为研究对象的全体数据时，称为总体数据（population data）。此时的方差为总体方差（population variance），记为 σ^2，其计算公式为

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{1-7}$$

总体方差的算术平方根即总体标准差（population standard deviation），用 σ 表示，其计算公式为

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{1-8}$$

为了便于计算，通常采用下列等价的简化公式：

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2, \quad \sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}x_i^2 - \overline{x}^2} \tag{1-9}$$

2. 样本方差和标准差 在实际统计研究中，观察数据一般都是研究对象的部分个体的数据 $x_1, x_2, \cdots, x_n$，称为样本数据（sample data）。此时的方差为样本方差（sample variance），记为 S^2，其计算公式为

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \tag{1-10}$$

样本方差的算术平方根即样本标准差（sample standard deviation），用 S 表示，其计算公式为

$$S = \sqrt{S^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2} \tag{1-11}$$

虽然标准差和方差都能反映每个数据偏离其均值的平均程度，但标准差具有与实际观察值相同的量纲，其意义较方差更明确，故比方差常用。

在案例 1-4 中，n=13，均值 $\overline{x} = 154.9$，故样本方差和标准差分别为

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2 \\ &= \frac{1}{12}[(146.6-154.9)^2 + (150.2-154.9)^2 + \cdots + (150.2-154.9)^2] \\ &\approx 33.08 \end{aligned}$$

$$S = \sqrt{S^2} = \sqrt{33.08} \approx 5.75$$

为了减少计算量，对已分组的数据可以利用各组频数和组中值近似计算样本方差和标准差。

设分组数据组数为 k，各组数据出现的频数分别为 $f_1, f_2, \cdots, f_k$，当然 $\sum_{i=1}^{k} f_i = n$，各组的组中值为 $m_1, m_2, \cdots, m_k$，样本方差和标准差的近似计算公式：

$$S^2 \approx \frac{\sum_{i=1}^{k}(m_i - \overline{x})^2 f_i}{\sum_{i=1}^{k} f_i - 1} = \frac{1}{n-1}\sum_{i=1}^{k}(m_i - \overline{x})^2 f_i \tag{1-12}$$

$$S = \sqrt{S^2} \approx \sqrt{\frac{1}{n-1}\sum_{i=1}^{k}(m_i - \overline{x})^2 f_i} \tag{1-13}$$

例如，根据表 1-3 中的数据，已经求出 $\overline{x} = 76.83$，可以通过表 1-7 计算成绩的样本方差和标准差：