

供基础、临床、预防、医药、口腔等专业用

医学统计学

孙玉文 主编

东北大学出版社

编写人员名单

- 主 编** 孙玉文
- 副主编** (以姓氏笔划为序)
- 于苏荣 (吉林医学院)
 - 于秋英 (沈阳医学院)
 - 石宝利 (中国医科大学)
 - 朱隆高 (吉林医学院)
 - 杜生福 (内蒙古医学院)
 - 李晓霞 (牡丹江医学院)
 - 杨俊英 (河北医学院)
- 编 委** (以编写字数多少为序)
- 孙玉文 (沈阳医学院)
 - 于秋英 (沈阳医学院)
 - 杨俊英 (河北医学院)
 - 潘秀丹 (沈阳医学院)
 - 朱隆高 (吉林医学院)
 - 于苏荣 (吉林医学院)
 - 张宝珍 (沈阳医学院)
 - 宋建荣 (衡阳医学院)
 - 杜生福 (内蒙古医学院)
 - 李晓霞 (牡丹江医学院)
 - 白淑敏 (沈阳医学院)
 - 石宝利 (中国医科大学)
 - 罗凤基 (河北医学院)
 - 李宏羊 (沈阳医学院)

前 言

随着医学科学的飞速发展及教学改革的要求，医学统计学必须适应当前的多种层次（临床专业、预防专业、药学专业等医药学本科生及研究生）的教学以及医学科研工作的需要，并应拓宽其适用范围和加强对学生的三基技能培养。

本书是在微积分学的起点上，以国内外有关的概率统计的一些基本概念、理论、方法作为这本书的主要内容。由于增加了概率论初步，从而解决了微积分学与医学统计方法的衔接问题。该书还介绍了非参数检验的概率原理；回归的各种可信限的确定以及按着计算机软件程序的要求写了多元统计方法。

本书是由沈阳医学院等七所院校部分同仁在总结多年教学科研的基础上合编而成的。在编写过程中得到了主编单位（沈阳医学院）各级领导和同志们的大力支持，在此一并致谢。

鉴于我们水平有限，时间仓促，错误难免，敬请各界同仁批评指正。

编 者

一九九五·四·四

目 录

绪 论	(1)
1 数据的初步处理	(2)
1.1 资料收集	(2)
1.2 资料整理	(3)
1.3 数值变量资料的数字特征	(5)
1.4 分类变量资料的数字特征	(17)
2 概率论初步	(27)
2.1 概 率	(27)
2.2 随机变量及其概率分布	(37)
2.3 数学期望	(44)
2.4 大数定律及中心极限定理	(54)
2.5 几种离散概率分布	(56)
2.6 几种连续分布	(61)
2.7 抽样分布	(70)
3 参数估计	(74)
3.1 无偏估计	(74)
3.2 参数的区间估计	(75)
4 数值变量资料的假设检验	(81)
4.1 假设检验的概念	(81)
4.2 样本均数与总体均数的比较	(85)
4.3 两样本均数的比较	(90)
4.4 样本方差的比较	(94)
4.5 多个样本均数的比较	(98)
4.6 正态性 D 检验	(114)
4.7 变量变换	(118)
4.8 假设检验应注意的问题	(121)

5 分类变量资料的假设检验	(122)
5.1 样本率与总体率的比较	(122)
5.2 两样本率(或构成比)的比较	(123)
5.3 多个或多列样本率(或构成比)比较的 χ^2 检验	(127)
5.4 配对分类变量资料 χ^2 检验	(129)
5.5 χ^2 检验注意事项	(130)
6 线性回归与相关	(131)
6.1 线性回归	(131)
6.2 简单线性回归	(132)
6.3 最小平方估计量的性质	(133)
6.4 置信限和显著性检验	(137)
6.5 简单线性回归的方差分析	(140)
6.6 相关	(144)
6.7 等级相关	(147)
7 非参数检验	(150)
7.1 非参数检验的概念	(150)
7.2 配对比较的符号秩和检验	(150)
7.3 两个样本比较的秩和检验	(152)
7.4 多个样本比较的秩和检验	(155)
7.5 配伍组设计的多个样本比较秩和检验	(158)
8 多元统计分析	(160)
8.1 向量和矩阵	(160)
8.2 一些特殊定理	(168)
8.3 多元正态分布	(169)
8.4 多元线性回归	(172)
8.5 逐步回归	(181)
8.6 距离判别	(190)
8.7 贝叶斯判别	(195)
8.8 逐步判别	(200)
9 统计表与统计图	(208)
9.1 统计表	(208)
9.2 统计图	(211)

10 调查设计	(224)
10.1 调查设计的基本原则	(224)
10.2 制定调查设计的一般步骤和内容	(224)
10.3 四种基本抽样方法	(228)
10.4 样本例数的估计方法	(230)
10.5 非抽样误差的控制	(232)
11 实验设计	(233)
11.1 实验设计的概述	(233)
11.2 实验设计的三要素	(234)
11.3 实验设计的基本原理和原则	(236)
11.4 常用的实验设计方法	(242)
12 人口统计	(247)
12.1 人口数与人口构成	(247)
12.2 人口估计	(250)
12.3 人口预测	(252)
12.4 人口自然变动指标	(254)
12.5 寿命表	(260)
13 疾病统计	(270)
13.1 疾病统计对象和观察单位	(270)
13.2 疾病频度指标	(271)
13.3 疾病严重程度指标	(273)
13.4 医院质量管理指标	(275)
13.5 疾病分类	(280)
附录一 统计用表	(284)
附录二 习题	(317)
习题1 数据初步处理	(317)
习题2 概率论初步	(320)
习题3 参数估计	(323)
习题4 数值变量资料的假设检验	(324)
习题5 分类变量资料的假设检验	(329)
习题6 线性回归和相关	(331)
习题7 非参数检验	(332)

习题 8 多元统计分析	(333)
习题 9 统计表与统计图	(335)
习题 10 调查设计	(337)
习题 11 实验设计	(337)
习题 12 人口统计	(338)
习题 13 疾病统计	(338)
附录三 英汉卫生统计学词汇	(340)
参考文献	(345)

绪 论

医学统计学是应用概率论和数理统计的基本原理和方法,结合医学实际,研究资料和信息搜集、整理、分析的一门学科。近代医学发展十分迅速,许多新的问题需要人们去研究解决,认识其内在的本质规律。医学统计学正是一门帮助人们透过许多偶然现象,去分析和判断事物的内在规律的科学。电子计算机的发展和普及应用,为大量的信息储存与检索,复杂的数据处理,特别是多因素分析,以及抽样模拟等提供了有利的条件。许多供医学统计设计和整理分析专用的统计程序,既利于医务工作者应用医学统计方法解决医学中的实际问题,又增加了应用一些复杂的统计分析方法进行医学科学研究的可行性。因此医学统计学已成为促进医学发展的一门重要应用科学,是医学科研工作者分析和解决问题的重要手段。

医学统计学的主要内容是研究医学统计设计、数理统计方法在医学科学中的应用。根据目前医学生的现状,本书主要介绍以下内容:

(1) 概率论的基本原理。主要介绍概率初步知识。

(2) 医学统计研究设计。进行医学科研设计时,除应用必要的专业知识外,必须应用医学统计设计的基本原理进行周密的考虑,采取必要的有效措施以保证研究的结果能够回答研究假说中提出的问题,从而使用较少的人力、物力和时间以取得较好的效果。设计是后续步骤的依据,是关键的一环。

(3) 常用的基本统计方法。包括:1)定理和定性资料的统计描述和总体指标的估计;2)假设检验:如 t 检验、 u 检验、方差分析、 χ^2 检验、秩和检验等;3)直线相关回归。

(4) 健康统计。包括医学人口统计、疾病统计。

(5) 多元统计分析方法。包括多元线性回归、逐步回归、逐步判别等。

统计学是统计工作实践的经验总结,但它又对统计工作的全过程起指导作用,这个全过程可分为四个步骤:设计、搜集、整理和分析资料。四个步骤互相联系,缺一不可。其中设计是整个统计研究的基础,在设计时应当对后三个步骤进行周密的考虑,并在整个研究中自始至终认真贯彻执行。

学习医学统计学应注意的问题:

(1) 应着重理解各种统计方法的基本原理和基本概念、基本理论,掌握适用范围和注意事项;在学习过程中注意联系实际,结合专业。例如,应多联系医学文献和医学科研工作,评价其统计设计和分析的优缺点。

(2) 培养科学的统计思维方法。例如,关于统计工作步骤间的内在联系;关于生物个体变异的客观存在,抽样误差不可避免,因而在进行样本指标的比较时,不能仅从数字表面大小看问题的思想;统计检验的基本思想;关于统计结论具有概率意义的思想等。

书中带星号部分,医疗等专业学生可不学;学过概率论的本科以上的学生可跳过概率初步这一章不学或只做浏览,选学其他章节。

1 数据的初步处理

一个调查设计、试验设计都要经周密思考之后,制定设计计划。有目的收集资料、合理地整理资料,使之系统化,然后对数据进行分析,作统计描述与统计推断。因此,周密设计、收集资料、整理资料、分析资料是统计工作的四个基本步骤。

数据的初步处理是指收集资料、整理资料、分析资料,三者是数据初步处理的基本步骤是密切相联不可分割、一环扣一环,其中一个环节出现缺陷,都将影响研究结果的正确性,因此必须是准确无误进行统计工作。

1.1 资料收集

资料收集就是根据研究的目的,进行周密的调查设计、试验设计或实验设计之后,有目的地收集准确、完整的原始资料,这是数据处理的前提与基础。统计资料收集的质量对资料整理和资料分析起着决定性作用,如果收集到的原始资料不准确或有遗漏,无论怎样精心进行资料整理和分析都不可能弥补这个缺陷,甚至会使收集到的资料完全失去应用价值。

1.1.1 资料的特性

收集的资料质量如何,直接关系到资料的整理与分析。因此,一个合格的统计资料应具备的特性是:

(1) 完整性

所说完整,是无遗漏。设计调查项目与预期分析指标相吻合,收集资料时认真判定与填写,无残缺。

(2) 准确性

所说准确,是不含糊,无差错,体现实事求是精神。

(3) 及时性

是指有时间紧迫感,不拖拉,当研究目的明确,周密设计后,要立刻搜集资料,其后迅速进行资料的整理与分析,争取以最快速度、分秒必争将研究结果公布于众(以论文、调查报告或成果形式)。可以使资料有时代意义。

1.1.2 统计资料的来源

原始的统计资料收集途径有两个方面,一种是经常性资料——日常工作记录和统计报表;另一种是一时性资料——由专题现场调查、试验研究或实验研究获得的资料。

(1) 经常性资料

1) 日常医疗卫生工作原始记录

种类,一是医疗工作原始记录。如门诊病志、住院病历、检查记录、传染病与职业病报告卡片、出生登记表、死亡登记表、病残记录、随访追踪调查表等。二是卫生工作原始记录。儿童生长发育卡片、各种预防接种登记卡,儿童智力测定卡、疫情报告登记卡等。三是计划生育与妇幼保健原始记录。婚姻登记表、计划生育节育措施登记表、孕产妇保健卡、围产儿登记卡、育龄妇

女登记表等。

以上这些原始资料都有定期的统计报表,根据这些资料可以进行某地、某个时期某些病的发病和死亡情况分析,按民族、婚姻、年龄、性别、职业、时间、地区等分析其三区分布及长期趋势。

日常工作原始记录资料的优点:一是能连续、及时记录原始材料;二是业务管理与科学研究的重要资料,掌握某病流行或其动态变化;三是可以做为评价防治工作的质量和效果的依据。

对这些原始工作记录,必须认真填写,注意积累和保存,并予以充分利用。

原始记录资料的缺点:一是登记时往往会有漏填、重复或有项目不清的情况;二是很难用于评定治疗效果,比如,病历是医院主要工作的日常记录,是临床治疗工作原始记录、经验的积累。由于事先无计划性,又无有对照,故不能做比较与评定;三是无法估算有病(发病率或患病率)的频率或死亡的频率,不能推论总体指标,因为住院病人是人群中特殊部分,是患者中的一部分,且住院病人是有选择性的,无病的人的是不去看病的,没有分母数量(年平均人口数或受检人数),因此无法计算出频率指标。

2) 统计报表

根据国家卫生部规定的各种报告制定统一报表,由医疗卫生机构定期逐级上报,统一的统计报表必须经过有关职能部门审批,内容要有的放矢,简明扼要,切实可行。

统计报表的种类有:①医院工作报表,如各科门诊量、住院病人数、治疗结果等;②预防接种报表,反映预防接种数量与完成情况;③计划生育报表,反映计划生育工作质量等;④疫情报表,有传染病、非传染性疾病,可反映疾病的长期趋势和疾病防治效果;⑤人口统计与疾病统计报表等。

统计报表资料的作用是:①能帮助分析研究一个地区或一个单位对疾病防治工作的组织管理和防治效果;②统计报表不但要上报上级机关,而且填报单位应充分利用统计报表所提供的现实情况,进行分析,提出问题,改进工作;③有较全面的居民健康状况以及医疗卫生机构的主要数据,可为总结、检查、和制订卫生工作计划提供重要数字依据;④是医学研究的资料来源。

(2) 一时性资料

在医疗卫生工作中,常常为了摸清疾病情况,分析疾病规律,探索致病因素,了解并掌握环境污染对居民健康的影响,以及评价防治效果,研究儿童青少年生长发育指标等等。需要进行专题现场调查、试验研究或实验研究。对于这类资料的收集是经立题后,进行一段时间的工作获取资料,故称之为“一时性资料”。

1.2 资料整理

收集来的资料,只能表明观察单位各自的特性,是一些零散的资料,必须要进行资料整理,使之系统化,条理化,为下一步的资料分析(analysis of data)作准备。资料整理的目的,是将原始调查表的项目,按质量特征和数量特征归纳分组。

统计资料的整理一般包括对原始资料进行核查和设计分组,现分述如下。

1.2.1 对原始资料进行核查

在资料收集的过程中,随时都要对原始资料进行修正、补充及合理剔除等核对检查,以利于及时发现问题及时解决。为慎重起见,在资料整理之前,还必须再进行一次系统地、认真地核查,以确保资料的可信性。可从两个方面进行核查:

(1) 逻辑检查

检查各个项目之间有无判断错误和不合逻辑的地方。如果女性疾病的调查,在性别上误写作“男”性;又如,是对老年人进行疾病调查,年龄写成小年龄者;诊断为风湿性心脏病患者,可是在心脏杂音听诊项目上无有标记等。

(2) 计算检查

主要针对统计报表中的合计栏与实际填报数据是否相符,有无差错,进行核算。

经过核查后,能补救、修正的尽量改正,否则予以剔除,以保证资料分析质量。

1.2.2 设计分组

任何一事物都具有两种标识:品质标识和数量标识。按着资料的两种特征,分别进行归纳若干组,以反映事物的一般特性,叫资料分组(data classification)。分组的原则是把性质不相同的分开。医学科研资料就以这个原则进行分组。

(1) 类型分组

按事物的性质和类别分组,如将观察单位按性别、疾病别等分组。

(2) 数量分组

按事物数量标志,如将观察单位按年龄、体重、血压等分组。数量分组的多少,以能说明资料的规律性为原则。数量分组组限要清楚,上下限不能相互包含,也不能留有空隙。如年龄分组正确写法:0~、5~、...(以5岁为一组)。其中0~,指从0岁起至不满5岁,余类推。本节着重介绍数量分组的方法和步骤。

例 1.1 沈阳市 1994 年测定某小学 104 名正常 9 岁女孩身高(cm)值,见表 1-1。

140.8	136.1	139.3	131.4	129.0	127.5	145.1	141.6
144.6	135.5	136.0	135.5	138.7	146.7	136.0	135.5
131.5	136.6	137.0	136.5	139.0	134.4	135.0	147.1
141.0	122.3	132.3	136.2	137.1	139.3	139.0	140.5
136.0	147.5	140.5	132.2	139.7	127.5	127.0	124.5
134.8	128.1	132.3	134.2	125.0	128.3	134.0	130.7
133.0	138.5	143.8	131.8	132.4	124.8	129.6	126.8
135.8	131.0	127.0	119.8	132.5	132.6	135.0	140.2
137.4	137.9	132.3	128.1	139.6	127.3	141.3	139.3
143.5	142.2	141.2	131.5	144.1	133.3	136.8	146.1
154.8	141.3	126.0	128.3	144.1	127.5	142.4	138.0
132.7	128.4	139.0	138.3	133.5	136.2	136.2	134.1
142.1	135.8	134.1	126.7	141.5	142.0	134.0	129.4

数量分组的组数一般为 8~15 组为宜。在 100 例以上不足 200 例的组距最好为 1/10 全距。分组步骤如下:

1) 求全距(Range)

从表 1-1 中找出最大值与最小值之差即为全距。本例全距(R)= $154.8\text{cm} - 119.8\text{cm} = 35.0\text{cm}$

2) 定组距(class interval)

组距是指相邻两组间距离 本例组距(i)= $R/10 = 35.0\text{cm}/10 = 3.5\text{cm} \approx 4.0\text{cm}$ (取整数便于分组)。

3) 定组数

一般用全距除以组距估计组数,本例组数= $R/i = 35.0\text{cm}/4.0\text{cm} = 8.75 \approx 9$ (组)

4) 定组限(class limit)

定组限的原则是第一组要包进最小值,最末组要包进最大值。每组的最大值称为组的上限(upper limit),每组的最小值称为组的下限(lower limit)。本例第一组下限值可取 119.0cm,上限值可近似取作为 123.0cm,以组距 4.0cm 进行确定下组下限值,直至将最大变量值 154.8cm 包进去为止,共可分出 9 组,最末组组限为 151.0~155.0cm(最末组可写出近似上限值 155.0cm)。

5) 列整理表(sorting table)

是把原始资料按上述规定的组段顺序整理归组的表格,将表 1-1 的数据归纳到各组段中去制成表 1-2,把有联系的项目编制在一个表里,可看出它们之间相互联系的规律性。因此在编制整理表时,有关项目要有次序,合理安排,清楚醒目。划记法适用于一览表收集的资料归组,简便易行,但易出差错,核查不便。

表 1-2 1994 年沈阳市某小学 104 名正常 9 岁女孩身高值整理表

组段(cm) (1)	划 记 (2)	频数 (3)
119~	┆	2
123~	正 一	6
127~	正 正 正	15
131~	正 正 正 正 下	23
135~	正 正 正 正 正	25
139~	正 正 正 正 下	22
143~	正 下	8
147~	┆	2
151~155	—	1
合 计		104

若收集的资料是单一表的卡片,可按上表组段划定方位,把卡片分别归入相应组内,尔后清点各组卡片数,即得各组例数(频数)。分卡法简单易行,便于核对,不易出差错,但费用要高于一览表。

6) 编制统计分析表(请见 9 章内容)

7) 资料的统计分析(请见 1.3,1.4,2,8)

1.3 数值变量资料的数字特征

数值变量资料(numerical variable data)是用定量方法测定某项指标的变量值,如身高(cm)、体重(kg)、血红蛋白(g/l)等均属数值变量。在对资料进行分析时,先要进行统计描述。

§ 1.2 资料整理中列出的整理表 1-2 基础上编制频数表(frequency table),表中(1)、(3)栏为所需的频数表 频数分布表(Frequency distribution table)。

1.3.1 频数分布表的作用

从频数分布表可显示出两个重要特征:集中趋势(central tendency)和离散趋势(tendency of dispersion)。从表 1-2 中可看出 104 名 9 岁女孩的身高不相同,但有一定分布规律:身高向中央位置集中,以中等身高者居多,称做集中趋势;从中央位置到两侧频数分布逐渐减少,称做离散趋势。从频数表可揭示出资料的分布特征和分布类型。计算出集中趋势和离散趋势的指标,进行统计描述。

1.3.2 频数分布图与频率分布图

将表 1-2 的整理表绘制成频数表与频率表,见表 1-3,将其第(2)栏、第(4)栏数值分别绘制频数分布直方图与累积频率图,见图 1-1 与图 1-2。

表 1-3 1994 年沈阳市某小学 104 名 9 岁正常女孩身高频数、频率表

组段(cm)	频数 f (人)	累积频数	累积频率(%)
(1)	(2)	(3)	(4)
119~	2	2	1.92
123~	6	8	7.69
127~	15	23	22.12
131~	23	46	44.23
135~	25	71	68.27
139~	22	93	89.42
143~	8	101	97.12
147~	2	103	99.04
151~155	1	104	100.00
合计	104		

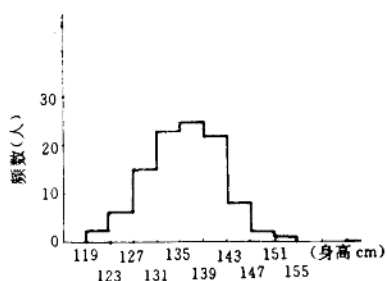


图 1-1 1994 年沈阳市某小学 104 名女孩身高频数分布直方图

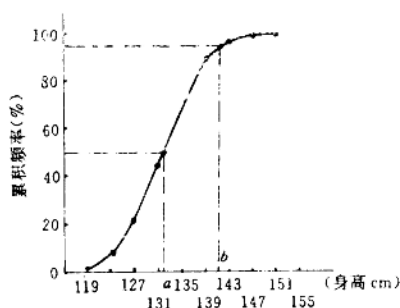


图 1-2 1994 年沈阳市某小学 104 名女孩身高累积频率图

图 1-1 是一个位置集中,左右两侧频数分布基本对称的近似正态分布的直方图,与表 1-2 频数分布相一致。若频数分布集中位置偏向一侧,左右两侧频数分布不对称谓之偏态分布。若集中位置偏向数值小的一侧,称作正偏态分布;若集中位置偏向数值大的一侧,称作负偏态分布。由图 1-2 的累积频率图上可以直接读出横轴上某一组段的累积频率,还能读出任意区间上累积频率的近似值,如 a 点与 b 点。

累积频率图的绘制:先在坐标平面上标出 $(t_0, 0)$, (t_1, F_1) , \dots , (t_{m-1}, F_{m-1}) , $(t_m, 1)$ 诸点,然后再把这些点用线段连接起来。把这样得到的折线段叫做累积频率图,见图 1-3。若记此线段所表示的函数为 $F(x)$, $x \in [t_0, t_m]$, 则 $F(x)$ 在区间 $[t_0, t_m]$ 上是递增的,且在 t_0 处取值为 0,在 t_m 处取值为 1。函数 $F(x)$ 的概率意义是:当 $x = t_i$ 时, $F(x)$ 就是 $[t_0, t_i]$ 上的累积频率;当 $x \neq t_i$ ($i = 1, 2, \dots, m$) 时, $F(x)$ 近似地表示 $(t_0, x]$ 上的累积频率。故从累积频率图上不但能直接读出形如 $(t_0, t_i]$ 的区间上的累积频率,还能读出任意区间 $(t_0, x]$ ($x \neq t_i, i = 1, 2, \dots, m$) 上累积频率的近似值;且可看出,对于任意区间 $(a, b]$ ($t_0 \leq a < b \leq t_m$) 上的频率,记作 $f(a, b]$, 公式如下:

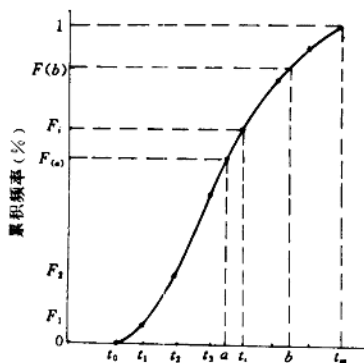


图 1-3 累积频率图

$$f(a, b] \approx F(b) - F(a) \quad (1.1)$$

$$f(t_i, t_{i+1}] \approx F_{(i+1)} - F_{(i)} = f_{(i-1)} \quad (1.2)$$

应当指出,当 a 与 b 或 t_i 与 t_{i-1} 均为等分点时,上式中的近似号可改为等号。

1.3.3 集中趋势的描述

集中趋势是反映数值变量资料的平均水平,通常用平均数指标来表示。

平均数(average)是统计中应用最广泛、最重要的一个指标体系。常用来描述一组同质变量值的集中位置的特征值,集中反映这组变量的典型值,具有代表一组变量值的意义。根据资料性质可有算术平均数、几何平均数、中位数、众数、调和平均数等。常用的平均数为前三种,现分述如下:

(1) 均数(mean)

是算术平均数(arithmetic mean)的简称,反映一组同质观测值在数量上的平均水平。希腊字母 μ 表示总体均数, \bar{X} 表示样本均数符号。均数适用于变量值分布比较集中对称的数值变量资料;或呈正态或近似正态分布的数值变量资料。

1) 均数的计算方法

i) 直接计算法。当变量值例数不多时,直接将各变量值 X_1, X_2, \dots, X_n 相加再除以变量值的例数 n 即是。计算公式:

$$\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} = \frac{\sum_{i=1}^n X_i}{n} \quad (1.3)$$

例 1.2 10 名 9 岁女孩体重(kg)分别: 34.0, 29.0, 30.0, 27.0, 25.0, 25.5, 34.0, 42.0,

38.0+30.0, 求其 10 名女孩平均体重, 依公式(1.3)计算。

$$\bar{X} = \frac{34.0+29.0+30.3+27.0+25.0+25.5+34.0+42.0+38.0+30.0}{10} = 31.5(\text{kg})$$

这 10 名 9 岁女孩平均体重为 31.5kg。

ii) 频数表法。当变量值例数较多时($n \geq 100$), 利用 § 1.2 知识把变量值编制成频数表再来计算

加权法(weighting method) 当资料中相同变量值的个数较多时, 可将相同变量值的个数, 即频数(f), 乘以该变量值组段的组中值(X), 以替代相同变量值逐个相加。用表 1-1 资料制成表 1-4 的频数表资料, (1) 栏为身高分组组段, 组段只表示该组变量值的大小范围, 故计算时必须求出各组的平均值(代表值)即为组中值(X), 组中值 $X = (\text{本组段下限} + \text{本组段上限} (\text{近似为下组段下限})) / 2$ 例如, 第一组组中值(class midvalue) $= (119\text{cm} + 123\text{cm}) / 2 = 121\text{cm}$ 。余类推, 即得表 1-4 中的第(3)栏。

表 1-4 104 名 9 岁女孩身高资料加权法计算表

组段(cm)	频数 f	组中值 X	fX	fX^2
(1)	(2)	(3)	(4) = (2) · (3)	(5) = (2) · (4)
119~	2	121	242	29182
123~	6	125	750	93750
127~	15	129	1935	249615
131~	23	133	3059	406847
135~	25	137	3425	469225
139~	22	141	3102	437382
143~	8	145	1160	168200
147~	2	149	298	44402
151~155	1	153	153	23409
合 计	104	—	14124	1922112

把各组段的组中值(X)与其相应组段的频数(f)相乘得第(4)栏 fX , 相加得 $\sum fX$, 最后除以总频数 $\sum f$ (即 n)。计算公式:

$$\bar{X} = \frac{\sum_{j=1}^m X_j f_j}{\sum_{j=1}^m f_j} \quad (1.4)$$

式中 m 为分組组数, X_1, X_2, \dots, X_m 分别为各组段的组中值, f_1, f_2, \dots, f_m 分别为各组段的频数。频数 f 起到了“权数”作用, 因它权衡了各组中值由于频数不同对均数所产生的影响。频数多, 权衡均数作用大; 反之亦然。故称之为加权法。

现将例 1.1 的表 1-4 资料用加权法按公式(1.4)求 9 岁女孩平均身高。

$$\bar{X} = \frac{121 \times 2 + 125 \times 6 + \dots + 153 \times 1}{1 + 6 + \dots + 1} = \frac{14124}{104} = 135.81(\text{cm})$$

这 104 名 9 岁女孩平均身高为 135.81cm。

从上述可知加权法计算其数值较大, 可进一步简化。自从使用计算器后, 已勿需用简捷法计算

算术均数了。

2) 均数的两个重要特性

i) 即各离均差(即各观察值 X 与均数 \bar{X} 之差)的总和等于零。

$$\sum_{i=1}^n (X_i - \bar{X}) = 0$$

ii) 离均差的平方和小于各观察值 X 与任何数 Z (而 $Z \neq \bar{X}$)之差的平方和。

$$\sum_{i=1}^n (X_i - \bar{X})^2 < \sum_{i=1}^n (X_i - Z)^2 \quad (1.5)$$

以上两特性可以表明,均数是一组观察值最理想的代表值。

均数能反映全部观察值的平均水平,因而应用广泛。当资料集中对称分布时,均数最能反映分布的平均水平,位于分布的中心。当资料呈正态分布时,其均数更有其重要作用。

(2) 几何均数(geometric mean)简记为 G

为同质 n 个变量值相乘积开 n 次方所得的根 $G = \sqrt[n]{X_1 \cdot X_2 \cdot \dots \cdot X_n}$ 用 G 表示。几何均数适用于变量值彼此相差较大,甚至成倍数关系,如抗体效价,人口的几何增长;或有些数据可转换成对数正态分布资料(如偏态分布资料),可用几何均数表示其平均水平。计算几何均数的变量值不能有 0;不能同时有正值和负值;若全为负值,计算时可把负号去掉,得出结果后,把负号加上。

1) 几何均数的计算方法

i) 直接算法。当变量值例数不多时,写成对数形式:

$$G = \lg^{-1} \left(\frac{\lg X_1 + \lg X_2 + \dots + \lg X_n}{n} \right) = \lg^{-1} \left(\frac{\sum_{i=1}^n \lg X_i}{n} \right) \quad (1.6)$$

例 1.3 有 5 人,其血清滴度分别为 1:4, 1:8, 1:16, 1:32, 1:64。求其平均滴度。

本例先求滴度的倒数,然后用几何均数计算。依公式(1.6)计算如下:

$$G = \lg^{-1} \left(\frac{\lg 4 + \lg 8 + \lg 16 + \lg 32 + \lg 64}{5} \right) = \lg^{-1} 1.204119983 = 16$$

故这 5 人血清平均滴度为 1:16。

ii) 频数表加权法计算几何均数。当变量值的个数较多时,可先编制成频数表后再计算几何均数。

例 1.4 某菌苗接种后二周,受试者血清凝集效价资料见表 1-5 的第(1)、(2)栏,计算其平均凝集效价。计算公式:

$$G = \lg^{-1} \left(\frac{f_1 \lg X_1 + f_2 \lg X_2 + \dots + f_m \lg X_m}{f_1 + f_2 + \dots + f_m} \right) = \lg^{-1} \left(\frac{\sum_{j=1}^m f_j \lg X_j}{\sum_{j=1}^m f_j} \right) \quad (1.7)$$

式中 f_1, f_2, \dots, f_m 为各组的频数, m 为组数, $\sum f$ 为总例数, $\sum f \lg X$ 为各组频数 f 与相应变量值的对数值相乘积之总和, $\lg X$ 为变量值的对数值, \sum 为总和符号, \lg^{-1} 为求真数符号。

将表 1-5 有关数据代入公式(1.7)即是

$$G = \lg^{-1} \left(\frac{96.3587}{50} \right) = \lg^{-1} 1.927174 = 84.56$$

故这 50 人血清凝集平均效价为 $1:84.6$ 。

表 1-5 50 人接种某菌苗 2 周后血凝效价

凝集效价倒数 X	人数 f	$\lg X$	$f \lg X$
(1)	(2)	(3)	(4) = (2) · (3)
20	5	1.3010	6.5050
40	10	1.6021	16.0210
80	20	1.9031	38.0620
160	8	2.2041	17.6328
320	5	2.5051	12.5255
640	2	2.8062	5.6124
合 计	50	—	96.3587

(3) 中位数和百分位数

中位数(median, 简记 M), 是将一组变量值按大小顺序排列, 位次居中的变量值即是中位数。在它的上下各有相等的变量值的个数。中位数适用于偏态分布资料; 一组变量值大部分集中在一侧, 少数分散在另一侧, 或资料的分布情况不清楚, 或数据一端(或两端)无界限时(如观测值记录有大于或小于多少, 无确切值), 宜用中位数表示其集中趋势。如传染病的平均潜伏期等。

百分位数(percentile), 也是一种位置指标, 以 P_x 表示。把一组资料的变量值从小到大排列, 分为 100 等份, 与 $x\%$ 相对应的数值, 即为第 $x\%$ 位数, 第百分之五十位数即为中位数, 中位数是一个特定的百分位数。一个百分位数 P_x , 把总体或样本的全部变量值分为两部分, 理论上 $x\%$ 的变量值比它小, 有 $(100-x)\%$ 的变量值比它大, 故百分位数是一个界值, 也是分布数列的百等份分割值。

1) 中位数和百分位数的计算方法

i) 直接算法。变量值例数不多时:

① 变量值个数为奇数 位次居中的变量值即是中位数

$$M = X_{((n+1)/2)} \quad (1.8)$$

② 变量值个数为偶数

$$M = [X_{(n/2)} + X_{(n/2+1)}] / 2 \quad (1.9)$$

式中 M 为中位数符号, n 为变量值的总例数, $(n+1)/2$ 、 $(n/2)$ 、 $(n/2+1)$ 为有序数列中变量值的位次, $X_{(n+1)/2}$ 、 $X_{(n/2)}$ 、 $X_{(n/2+1)}$ 为相应位次上的变量值。

例 1.5 某传染病的 7 例患者其潜伏期(天)分别是 2, 3, 3, 3, 4, 5, 6。求其平均潜伏期。

本例 $n=7$, 为奇数 按式(1.8)计算 M 。

$$M = X_{((7+1)/2)} = X_4 = 3(\text{天})$$

例 1.6 9 岁女孩 8 人, 其体重(kg)分别为 22.9, 23.0, 24.2, 26.1, 26.9, 27.5, 27.5, 41.1。求 M 值(即求 8 名女孩平均体重值)。

本例 $n=8$, 为偶数 按式(1.9)计算 M 。

$$M = [X_{(8/2)} + X_{(8/2+1)}] / 2 = [X_4 + X_5] / 2 = [26.1 + 26.9] / 2 = 26.5(\text{kg})$$

ii) 频数表法计算中位数与百分位数。当变量值例数较多时, 先将资料编制成频数表后, 再进行中位数或百分位数的计算。