★ 高等医药院校护理学专业规划教材

医学统计学

MEDICAL STATISTICS

★ 主编　王乐三

中南大学出版社
www.csupress.com.cn

★ 高等医药院校护理学专业规划教材

医学统计学

MEDICAL STATISTICS

★ 主编　王乐三

编者（按姓氏笔画排序）

王一任　王乐三　史静琤

罗建清　胡平成　胡国清

胡　明　曾小敏　谢和宾

童　瑶　虞仁和

Medical Statistics

中南大学出版社

www.csupress.com.cn

图书在版编目(CIP)数据

医学统计学/王乐三主编. —长沙:中南大学出版社,2010
ISBN 978-7-5487-0010-4

Ⅰ.医… Ⅱ.王… Ⅲ.医学统计-高等学校:技术学校-教材 Ⅳ.R195.1

中国版本图书馆 CIP 数据核字(2010)第 047620 号

医学统计学

王乐三 主编

□责任编辑 李 娴
□责任印制 文桂武
□出版发行 中南大学出版社
社址:长沙市麓山南路 邮编:410083
发行科电话:0731-88876770 传真:0731-88710482
□印 装 长沙市宏发印刷厂

□开 本 787×1092 1/16 □印张 18.75 □字数 460 千字
□版 次 2010 年 5 月第 1 版 □2010 年 5 月第 1 次印刷
□书 号 **ISBN 978-7-5487-0010-4**
□定 价 **35.00 元**

前　言

新世纪是生命科学和信息技术快速发展的时代。要在大量的医学信息中获得有价值的结果，需要对信息进行科学的分析。医学统计学是应用统计学的基本原理和方法，研究医学及其相关领域数据信息的收集、整理、分析、表达和解释的一门科学，是医学生的一门重要的基础课程，是医护工作者在从事临床工作和科学研究过程中必须掌握和了解的基本知识。

本书从专业培养目标和教学实际出发，精选“基础理论、基本知识和基本技能”内容，突出学生自学能力、实践能力、学以致用能力、举一反三能力的培养。全书共13章，第一章绪论，第二章至第十一章基本统计方法，第十二章医学人口统计与疾病统计，第十三章SPSS for Windows统计分析。还有统计用表、练习题和英汉名词对照3个附录。涵盖了医学本科生培养目标中需要了解和必须掌握的统计学的基本理论、基本方法和基本技能。

本书的编写得到中南大学出版社、中南大学网络学院和中南大学湘雅公共卫生学院的大力支持。李娴编辑为书稿编辑加工做了大量深入细致的工作。谨在此一并致以崇高敬意和衷心谢意。

限于学识和水平，不妥之处在所难免，诚恳希望读者和医学界同仁不吝指正。

王乐三

2010年4月于长沙

目 录

第一章 绪 论

医学研究的对象主要是人以及与其健康有关的各种影响因素。由于医学研究对象变异现象的客观存在，因此研究中的许多观测结果具有不确定性。例如同性别、同年龄儿童的体重有轻有重；临床上同一药物治疗患同一疾病的患者，其疗效有好有差。统计学(statistics)是研究随机事件数据收集、整理、分析、推断等原理和方法的学科，是了解随机事件偶然现象背后内在规律性的有效手段和工具。医学统计学(medical statistics)是以医学理论为指导，运用数理统计学的原理和方法，研究医学科研中有关数据的收集、整理和分析的应用科学。例如，在制订调查计划或实验设计时如何保证样本的代表性和样本间的可比性；对调查或实验结果如何选用恰当的统计指标进行描述和对总体进行相应的统计推断；对调查或实验结果存在的差异与关联如何选择适当的检验方法进行统计分析；在撰写研究报告时，如何正确表达和解释统计分析结果等。随着医学的发展，作为医学科学研究方法学的医学统计学已逐渐为广大医务工作者和医学科学工作者所认识、所接受，并在基础医学、临床医学、预防医学等各个研究领域广为应用。

国际统计学界通常把生命科学研究、临床医学研究和预防医学研究中的统计学内容统称为生物统计学(biostatistics)。由于各研究领域的侧重点不同，我国统计界通常把生命科学实验研究中的统计学内容称为生物统计学，把基础医学和临床医学研究中的统计学内容称为医学统计学，把预防医学研究中的统计学内容称为卫生统计学。随着医学研究模式的转变，医学领域各个学科相互渗透，所涉及的统计学研究工作已难以区分它们之间的差别。

电子计算机的普及与统计软件(如 SAS、SPSS)的开发，为医学科学研究中的数据信息的储存、整理和分析提供了便利的条件，同时也促进了医学统计学的迅速发展和不断完善。

第一节 统计学中的几个基本概念

一、同质与变异

统计的研究对象是由观察单位(observed unit)构成的群体。统计研究中，给观察单位规定一些相同的因素情况，称为同质(homogeneity)。如研究儿童的生长发育，规定的同性别、同年龄、同地区、同民族、健康的儿童即为同质的儿童。但即使是同质个体，其研究因素也存在差异，称为变异(variation)。如同质的儿童身高有高有矮，体重有重有轻。统计学的任务就是在同质的基础上，对个体变异进行分析研究，揭示由变异所掩盖的同质事物内在的本质和规律。

二、总体与样本

根据研究目的而确定的同质观察单位的全体称为总体(population)，更确切地说，它是同质的所有观察单位某种观察值的集合。例如调查某地 2008 年 7 岁正常女童的身高，则观察

对象是该地2008年全体正常7岁女童，观察单位是每个女童，观察值（变量值）是测得的身高值，该地2008年全体7岁正常女童的身高值就构成一个总体。它的同质基础是同一地区、同一年份、同一年龄的正常女童。这里的总体明确规定了空间、时间、人群范围内有限个观察单位，称为有限总体（finite population）。在另一些情形下，总体的概念是设想的或抽象的，如研究用某药治疗缺铁性贫血的疗效，这里总体的同质基础是缺铁性贫血患者，同时用某药治疗，该总体应包括用该药治疗的所有缺铁性贫血患者的治疗结果，是没有时间和空间范围限制的，因而观察单位数是无限的或不易确定的，称为无限总体（infinite population）。

医学研究中，多数的总体是无限的，即使是有限总体，由于观察单位数太多，耗费很大的人力、物力和财力，因此不可能也不必要对总体进行全面的研究。实际研究中，常常是从总体中随机抽取一部分观察单位组成样本，对样本进行研究，用样本信息来推断总体特征。样本（sample）是从总体中随机抽取的部分观察单位变量值的集合。样本的例数称为样本含量（sample size）。如上例，可从某地2008年7岁正常女童中，随机抽取110名女童，逐个进行身高测量，得到110名女童的身高测量值，组成样本。抽样一定遵循随机的原则，并要有足够的样本含量。应当强调，获取样本仅仅是手段，而通过样本信息来推断总体特征才是研究的目的。

三、参数与统计量

根据总体个体值统计计算出来的描述总体（更确切地说，是指有规律分布的总体）的特征量，称为总体参数（parameter）。总体参数一般用希腊字母表示，如总体均数，总体标准差，总体率等。和总体参数相对应，根据样本个体值统计计算出来的描述样本的特征量，称为样本统计量（statistic）。样本统计量用拉丁字母表示，如样本均数$\bar{X}$，样本标准差S，样本率p等。如研究某年某地50岁以上男子慢性支气管炎的患病情况，该地所有50岁以上男子慢性支气管炎的患病率即为总体参数。若进行抽样研究，用随机的方法从该地抽取一部分50岁以上男子来调查其患病情况，计算的患病率即为统计量。总体参数一般是不知道的，抽样研究的目的就是用样本统计量来推断总体参数，包括区间估计和假设检验。

四、误差

误差（error）是指实测值与真值之差，按其产生原因和性质可分为随机误差（random error）与非随机误差（nonrandom error）两大类，后者又可分为系统误差（systematic error）与非系统误差（nonsystematic error）两类。

1. 随机误差

是一类不恒定的、随机变化的误差，由多种尚无法控制的因素引起。例如，在实验过程中，在同一条件下对同一对象反复进行测量，虽极力控制或消除系统误差后，每次测量结果仍会出现一些随机变化即随机测量误差（random error of measurement）以及在抽样过程中，由于抽样的偶然性而出现的抽样误差（sampling error）。统计分析主要是针对抽样误差。

2. 系统误差

是一类恒定不变或遵循一定变化规律的误差，其产生原因往往是可知的或可能掌握的。可能来自于受试者抽样不均匀，分配不随机；可能来自于不同实验者个人感觉或操作上的差异；可能来自于不标准的仪器，也可能来自于外环境非实验因素的不平衡等。应尽可能设法

预见到各种系统误差的具体来源，力求通过周密的研究设计和严格的技术措施对系统误差加以消除或控制。

3. 非系统误差

亦称为过失误差(gross error)，是由研究者偶然失误而造成的误差。例如，抄错数字，点错小数点，写错单位等，这类误差应当通过认真检查核对予以清除。

五、概率

概率(probability)是描述事件发生可能性大小的一个量值，常用符号 P 表示。概率的取值范围在0～1之间。在一定条件下，肯定发生的事件称为必然事件，肯定不发生的事件称为不可能事件，可能发生也可能不发生的事件称为随机事件。必然事件的概率等于1，不可能事件的概率等于0，随机事件的概率在0与1之间。在实际问题中，当重复观测次数足够大时，可以频率作为概率的估计值。例如用某药治疗某病患者的预后有治愈、好转、无效、死亡四种结果，但对于每个患者治疗后发生哪种结果是不确定的，这里的每一种可能结果都是一个随机事件，如果将结果为"治愈"这个事件记为A，则该患者治愈的概率可记为 $P(A)$，或简记为 P。本例在相同的条件下，经过一定数量患者的治疗，就可得到治愈例数 f 占总病例数 n 的比值，即频率 f/n。当 n 逐渐增大时，这个比值越来越接近一个稳定的数值，即该病治愈的概率 $P(A)$。

统计上一般将 $P\leqslant 0.05$ 或 $P\leqslant 0.01$ 的事件称为小概率事件，表示其发生的可能性很小，可以认为在一次抽样中不会发生。由于存在抽样误差，用样本统计量推断总体参数不可能是肯定推断，只能是概率推断。

第二节 统计资料的类型

确定总体之后，研究者应对每个观察单位的某项特征进行观察或测量，这种特征能表现观察单位的变异性，称为变量(variable)。对变量的观测值称为变量值(value of variable)或观察值(observed value)，由变量值构成资料(data)。例如，以人为观察单位调查某地某年7岁健康儿童的生长发育状况，性别、身高、体重等都可视为变量，性别有男有女，身高可高可矮，体重可轻可重，不同个体不尽相同。变量的观察结果可以是定量的，如身高的厘米数，也可以是定性的，如新生儿属男属女。按变量属定量或定性的类型，医学统计资料一般可分为数值变量资料和分类变量资料两大类，后者又可分为无序分类变量资料和有序分类变量资料。不同类型的资料应采取不同的统计方法分析处理。

一、数值变量资料

数值变量(numerical variable)资料又称定量资料(quantitative data)或计量资料(measurement data)，为观测每个观察单位某项指标的大小而获得的资料。其变量值是定量的，表现为数值大小，一般有度量衡单位。根据其观测值取值是否连续，又可分为连续型(continuous)或离散型(discrete)两类。前者可在实数范围内任意取值，如身高、体重、血压等；后者只取整数值，如某医院每年的病死人数等。

二、无序分类变量资料

无序分类变量(unordered categorical variable)资料又称定性资料(qualitative data)或计数资料(enumeration data),亦称名义变量(nominal variable)资料,为将观察单位按某种属性或类别分组计数,分组汇总各组观察单位数后而得到的资料。其变量值是定性的,表现为互不相容的属性或类别,如试验结果的阳性阴性,家族史的有无等。定性资料分两种情形:

(1)二分类:如检查某小学学生大便中的蛔虫,以每个学生为观察单位,结果可报告为蛔虫卵阴性与阳性两类;如观察某药治疗某病患者的疗效,以每个患者为观察单位,结果可归纳为治愈与未愈两类。两类间相互对立,互不相容。

(2)多分类:如观察某人群的ABO血型分布,以人为观察单位,结果可分为A型、B型、AB型与O型,为互不相容的4个类别。

三、有序分类变量资料

有序分类变量(ordinal categorical variable)资料又称半定量资料(semi-quantitative data)或等级资料(ranked data)。为将观察单位按某种属性的不同程度分成等级后分组计数,分类汇总各组观察单位数后而得到的资料。其变量值具有半定量性质,表现为等级大小或属性程度。如观察某人群某血清反应,以人为观察单位,根据反应强度,结果可分-、±、+、++、+++、++++6级;又如观察用某药治疗某病患者的疗效,以每名患者为观察单位,结果可分为治愈、显效、好转、无效4级。

统计分析方法的选用,是与资料类型密切联系的。在资料分析过程中,根据需要在有关专业理论指导下,各类资料间可以互相转化,以满足不同统计分析方法的要求。例如,以人为观察单位观察某人群脉搏数(次/min),属计量资料;若根据医学专业理论,定义脉搏数在60次/min至100次/min为正常,小于60次/min或大于100次/min为异常,按"正常"与"异常"两种属性分别清点人数,汇总后可转化为计数资料;若进一步定义脉搏数小于60次/min为缓脉,大于100次/min为速脉,按"缓脉"、"正常"与"速脉"3个等级分别清点人数,汇总后可转化为等级资料。

第三节 统计工作的基本步骤

统计工作可分为4个步骤,即统计设计、收集资料、整理资料和分析资料。这4个步骤密切联系,缺一不可,任何一个步骤的缺陷和失误,都会影响统计结果的正确性。

一、统计设计

设计(design)是统计工作的第一步,也是关键的一步,是对统计工作全过程的设想和计划安排。统计设计就是根据研究目的确定研究因素、研究对象和观察指标,并在现有的客观条件下决定用什么方式和方法来获取原始资料,并对原始资料如何进行整理,以及整理后的资料应该计算什么统计指标和统计分析的预期结果如何等进行计划安排,力争以较少的人力、物力和时间取得较好的效果。医学科研设计按是否对研究对象施加处理因素分为调查设计和实验设计。

二、收集资料

收集资料(collection of date)是根据设计的要求，获取准确可靠的原始资料，是统计分析结果可靠的重要保证。没有完整、准确的原始数据，即使有先进的整理和分析方法，也不会产生准确的分析结果。医学统计资料的来源主要有以下3个方面：①统计报表，如法定传染病报表、出生死亡报表、医院工作报表等；②医疗卫生工作记录，如病历、医学检查记录、卫生监测记录等；③专题调查或实验研究。

三、整理资料

整理资料(sorting data)就是将收集到的原始资料进行反复核对和认真检查，纠正错误，分类汇总，使其系统化、条理化，便于进一步的计算和分析。资料整理的过程如下：①审核，即将收集到的原始资料进行认真的检查核对，以保证资料的准确性和完整性；②分组，将完整准确的原始资料按照观察单位的类别或数值大小进行归纳分组；③汇总，即按照设计的要求将分组后的资料汇总整理成统计表。

四、分析资料

分析资料(analysis of data)是根据设计的要求，对整理后的数据进行统计学分析，结合专业知识，作出科学合理的解释。统计分析包括以下两大内容：①统计描述(statistical description)，是利用统计指标、统计表和统计图相结合来描述样本资料的数量特征及分布规律。②统计推断(statistical inference)，是使用样本信息来推断总体特征。统计推断包括区间估计和假设检验。

医学科研一般是抽样研究，得到的是样本统计量，所以对样本分析并不是真正的科研目的。通过样本统计量进行总体参数的估计和假设检验，以达到了解总体的数量特征及其分布规律，才是最终的研究目的。

本课程是为医学生学习专业课程和从事医学领域工作打下必要的统计学基础。学习本课程时应注意：①结合专业，联系实际来理解医学统计学中的基本概念、基本理论和基本方法。②注重统计思维的培养，对书中的统计公式不必深究其数学推导，着重在其意义、用途和应用条件的理解。③熟悉有关统计软件的基本使用方法，增强实际动手能力，提高分析问题和解决问题的能力。

第二章 数值变量资料的统计描述

统计描述是用统计图表和统计指标来描述资料的分布规律及其数量特征。本章主要介绍常用的数值变量资料描述性指标、正态分布及其应用。

第一节 频数分布表

对于一群同质个体的某项定量指标，收集到计量数据之后，欲了解其分布的范围、数据最集中的区间以及分布的形态，可通过编制频数分布表[简称频数表(frequency table)]来实现。频数分布(frequency distribution)通常是针对样本而言。对于连续变量(continuous variable)，频数分布为 n 个变量值在各变量值区间内的变量值个数的分配[(见表2-2第(1)栏和第(2)栏)]。对于离散变量(discrete variable)，频数分布为 n 个变量值在各(或各几个)变量值处的变量值个数的分配[见表2-3第(1)栏和第(2)栏]。现以连续变量为例介绍频数分布表的编制步骤。

例2-1 某年某市120名12岁健康男孩身高资料如表2-1，试编制频数分布表。

表2-1 某年某市120名12岁健康男孩身高(cm)测量资料

142.3	156.6	142.7	145.7	138.2	141.6	142.5	130.5	132.1	135.5
134.5	148.8	134.4	148.8	137.9	151.3	140.8	149.8	143.6	149.0
145.2	141.8	146.8	135.1	150.3	133.1	142.7	143.9	142.4	139.6
151.1	144.0	145.4	146.2	143.3	156.3	141.9	140.7	145.9	144.4
141.2	141.5	148.8	140.1	150.6	139.5	146.4	143.8	150.0	142.1
143.5	139.2	144.7	139.3	141.9	147.8	140.5	138.9	148.9	142.4
134.7	147.3	138.1	140.2	137.4	145.1	145.8	147.9	146.7	143.4
150.8	144.5	137.1	147.1	142.9	134.9	143.6	142.3	143.3	140.2
125.9	132.7	152.9	147.9	141.8	141.4	140.9	141.4	146.7	138.7
160.9	154.2	137.9	139.9	149.7	147.5	136.9	148.1	144.0	137.4
134.7	138.5	138.9	137.7	138.5	139.6	143.5	142.9	146.5	145.4
129.4	142.5	141.2	148.9	154.0	147.7	152.3	146.6	139.2	139.9

1. 计算全距

一组变量值最大值和最小值之差称为全距(range)，亦称极差，常用 R 表示。本例最大值为160.9，最小值为125.9，故 $R=160.9-125.9=35$(cm)。

2. 确定组距

组距(class interval)用 i 表示，组距大小决定于分组多少。变量值在100例左右一般分为8~15组。若变量值较少，组数可相应少些，变量值很多，组数可酌情多些，总之，以能显示变量值的分布规律为宜。组距=全距/组数，本例拟分10组，组距=35/10=3.5，一般取靠近的整数作为组距，本例取 $i=4$ cm。

3. 划分组段

每个组段的起点称组段下限，终点称组段上限。第 1 组段应包括最小变量值，故其下限取小于或等于最小值的较为整齐的数值，本例取小于 125.9 的 125 作为第 1 组段的下限。本例为连续变量，组段应写为上限开口型，如 125 ~，129 ~，133 ~，…。第 2 组段的下限 129 为第 1 组段的上限，第 3 组段的下限 133 为第 2 组段的上限，余类推。最后 1 个组段应包括最大变量值，一般写为上限闭口型，本例最大值为 160.9，最后 1 个组段写为 157 ~ 161。如表 2 -2 第(1)栏，本例共分 9 组，写成 9 个组段。

4. 统计频数

将所有变量值通过划记逐个归入相应组段，如表 2 -2 第(1)栏 125 ~，表示所有身高值等于或大于 125 cm，小于 129 cm，都应归入此组，余仿此。表 2 -2 第(1)、(2)栏即为所需的频数表。

5. 频率与累计频率

频数表中的各组频数之和等于总例数 n，将各组的频数除以 n 所得的比值被称为频率。频率描述了各组频数在全体中所占的比重，各组频率之和应为 100%，见表 2 -2 第(3)栏。累计频数等于该组段及前面各组段的频数之和，累计频率等于累计频数除以总例数，见表 2 -2 第(4)栏和第(5)栏。累计频率描述了累计频数在总例数中所占的比例。图 2 -1 为描述 120 名 12 岁健康男孩身高的频数分布的直方图。

表 2 -2　某年某市 120 名 12 岁健康男孩身高(cm)的频数分布

组段 (1)	频数 (2)	频率(%) (3)	累计频数 (4)	累计频率(%) (5)
125 ~	1	0.83	1	0.83
129 ~	4	3.33	5	4.17
133 ~	10	8.34	15	12.50
137 ~	27	22.50	42	35.00
141 ~	35	29.17	77	64.17
145 ~	27	22.50	104	86.67
149 ~	11	9.17	115	95.83
153 ~	4	3.33	119	99.17
157 ~ 161	1	0.83	120	100.00
合　计	120	100.00	—	—

频数分布表的用途主要有：

1. 揭示资料的分布类型

频数分布的类型可分为对称分布和偏态分布两种。若各组段的频数以频数最多组段为中心左右两侧大体对称，就认为该资料是对称分布，如表 2 -2 及图 2 -1；反之，则认为是偏态分布，如表 2 -4 及图 2 -2、表 2 -5 及图 2 -3。图 2 -2 频数最多组段(21 ~)右侧的组段数多于左侧的组段数，频数向右侧拖尾，称右偏态分布(skewed to the right distribution)，也称正偏态分布(positive skewness distribution)。图 2 -3 频数最多组段(30 ~)左侧的组段数多于右侧的组段数，频数向左侧拖尾，称左偏态分布(skewed to the left distribution)，也称负偏态分布(negative skewness distribution)。

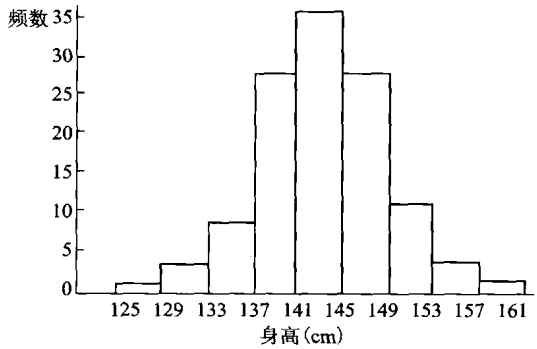

图 2－1　某年某市 120 名 12 岁健康男孩身高的频数分布

表 2－3　某医院 1123 名产后出血孕妇的人流次数分布

人流次数 (1)	产后出血人数 (2)	累计频数 (3)	累计频率(%) (4)
0	402	402	35.80
1	330	732	65.18
2	232	964	85.84
3	118	1082	96.35
4	27	1109	98.75
5	11	1120	99.73
6	3	1123	100.00
合计	1123	—	—

表 2－4　某年某市 110 名正常成年女子的血清转氨酶(mmol/L)含量分布

血清转氨酶含量	人 数
12 ~	2
15 ~	8
18 ~	13
21 ~	22
24 ~	18
27 ~	13
30 ~	11
33 ~	9
36 ~	7
39 ~	4
42 ~45	3

表 2－5　某年某市 100 名正常成年人的血清肌红蛋白(μg/mL)含量分布

血清肌红蛋白含量	人 数
0 ~	2
5 ~	3
10 ~	7
15 ~	9
20 ~	10
25 ~	22
30 ~	23
35 ~	13
40 ~	9
45 ~50	2

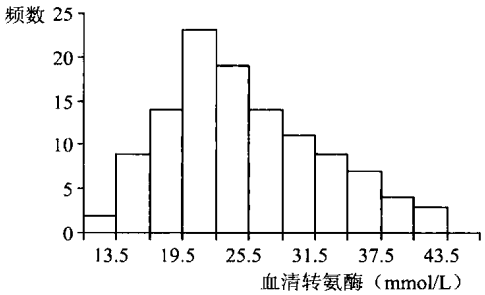

图 2－2　某年某市 110 名正常成年女子血清转氨酶的频数分布

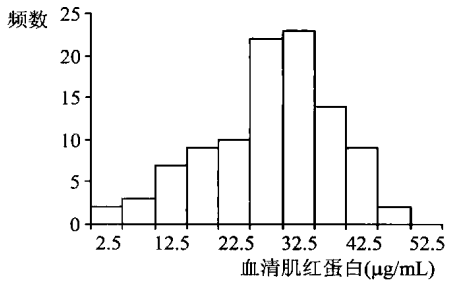

图 2－3　某年某市 100 名正常成年人血清肌红蛋白的频数分布

2. 观察资料的集中趋势和离散趋势

如表 2－2 可见 120 名 12 岁男孩身高位于中央部分“141 ～”组段人数最多，是为集中趋势；从中央部分到两侧频数分布逐渐减少，是为离散趋势。

3. 便于发现某些特大或特小的离群值

若在频数表的两端，连续出现几个组段的频数为 0 后，又出现一些特大或特小值，需要进一步检查和核对这些离群值(outlier)，如有错，应予纠正。

4. 便于进一步计算统计指标和作统计处理

详见下文。

第二节　集中趋势指标

数值变量资料的集中趋势指标是用平均数(average)来描述的，代表一组同质变量值的平均水平。常用的平均数有算术均数、几何均数和中位数。

(一)算术均数

算术均数（arithmetic mean）简称均数，适用于对称分布或近似对称分布的资料。习惯上以希腊字母 μ 表示总体均数(population mean)，以 $\bar{X}$ 表示样本均数(sample mean)。常用计算方法有直接法和加权法。

1. 直接法

计算公式为：

$$\bar{X} = \frac{X_1 + X_2 + \cdots + X_n}{n} = \frac{\sum X}{n} \qquad (2.1)$$

式中，X_1，X_2，…，X_n 为所有变量值，n 为样本含量，$\sum$（希腊字母，读作 sigma）为求和的符号。

例 2－2　现有 11 名 5 岁女孩的身高值(cm)为 112.9、99.5、100.7、101.0、112.1、118.7、107.9、108.1、99.1、104.8、116.5，求其均数。

将身高数值代入公式(2.1)得：

$$\bar{X} = \frac{112.9 + 99.5 \cdots + 116.5}{11} = \frac{1181.3}{11} = 107.39(\text{cm})$$