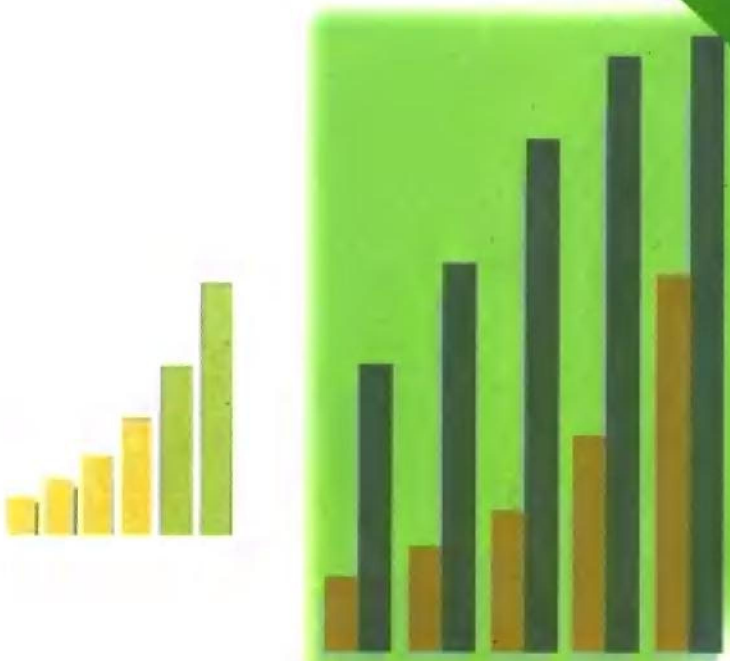

卫生统计学

主编
王 燕
安 琳

北京医科大学出版社

卫生统计学

主　　编　王　燕　安　琳
副 主 编　任正洪

编写人员（按姓氏笔画为序）

王　燕　王晓莉　安　琳　任正洪
朱墨涛　罗树生　高燕秋　康楚云

WEI SHENG TONG JI XUE

图书在版编目（CIP）数据

卫生统计学/王燕，安琳主编．-北京：北京医科大学出版社，1999
ISBN 7-81071-000-1

Ⅰ．卫…　Ⅱ．①王…②安…　Ⅲ．卫生统计　Ⅳ．R195.1

中国版本图书馆 CIP 数据核字（1999）第 37281 号

北京医科大学出版社出版发行
（100083　北京学院路 38 号　北京医科大学院内）
责任编辑：暴海燕
责任校对：齐　欣
责任印制：郭桂兰
山东省莱芜市印刷厂印刷　新华书店经销
*　　*　　*
开本：787×1092　1/16　印张：9　字数：227 千字
1999 年 8 月第 1 版　1999 年 8 月山东第 1 次印刷　印数：1—10000 册
定价：13.50 元

前　言

卫生统计学是一门关于如何运用数理统计学的原理和方法进行医学科研，特别是关于如何收集、整理、分析医学数据的应用科学。卫生统计学是一门基础医学课程，亦可以说是一门基础医学方法学课程。近年来我们欣慰地看到，许多医学工作者，正确应用卫生统计学理论和方法，取得了可喜的成果。但同时需要指出的是，根据我们的经验，卫生统计学是一门较难学习和掌握的课程。它需要学习者花费精力、投入时间、反复实践、不断思考，从而才能够正确掌握灵活应用。

本书的初衷是为北京医科大学护理系成人自学高考的本科学生学习《卫生统计学》而编写的。本书亦可作为医学专业、护理专业大学本科生或专科生及预防医学专业的专科生的自学教材。另外，本书对于已具有一定临床工作经验的医护人员提高卫生统计基础知识亦有一定参考价值。

全书共有十一章。第一章介绍了卫生统计学的一些些本概念；第二、三和四章是关于描述性统计及统计推理；第五、六章讲授计数资料的描述性统计和统计推断。第七章统计图表的制作；第八章是关于秩和检验；第九章是直线相关与回归分析；第十、十一章介绍了基本人口统计与疾病统计指标。

本书的编写人员为中青年教师，在编写此书之前不仅全部从事过预防医学和临床医学专业本科生的教学，而且均教授过各类专升本、大专、进修、夜大等学生的《卫生统计学》教学，积累了一定的教学经验，包括对成人及大专生的教学经验，能体会到初学者，特别是成人学生学习《卫生统计学》的难点。本书以表达正确、思路清晰、重点突出、通俗易懂、便于自学的原则编写，并将教学中积累的体会融入其中。希望该书能成为初学者的良师益友。

北京医科大学出版社为使本书在较短时间内问世鼎力相助，在此表示感谢。

限于编者水平，本书难免有不尽人意之处，敬请读者和专家提出宝贵意见。

编者

一九九九年六月

目　录

第一章 绪 论

第一节 卫生统计学简介

大家都学过哲学，知道客观世界总是处于不断变化之中，只有从变化中去认识它，才能对它获得深刻的了解。从变化的性质来说，变化有量变和质变之分；按变化的现象来分，变化有必然和偶然之别。既然客观世界如此复杂，我们认识它的变化发展规律，就必须用科学的方法和手段去研究和发现隐藏在偶然现象背后的必然规律。**统计学就是这样一门科学，它是研究数据的搜集、整理、分析与推断的科学。**

医学的研究对象主要是人体以及与人体健康有关的各种因素。由于生物现象的变异较大，各种影响因素又错综复杂，故医学也需要运用统计学方法，透过偶然现象来探测其规律性。**卫生统计学是把统计理论方法应用于居民健康状况研究、医疗卫生实践和医学科研的一门应用性学科。**它广泛地应用于基础医学、预防医学、临床医学和卫生事业管理等各个领域。

卫生统计学的研究内容包括三个主要方面：①卫生统计的基本理论和方法，包括医学科研的设计及其数据处理的统计理论和方法。例如医学科研中的现场调查设计和实验设计，科学、准确、及时地搜集与整理统计资料的方法，以及科学地进行统计描述、估计、比较和预测等的方法。②居民健康统计，如出生、疾病及死亡统计等。③卫生资源分布和卫生业务统计，如反映卫生机构、人员与床位等卫生事业基本情况统计；反映医院工作质量的医院统计和防疫统计等。

计算机发展之前，大信息量资料的储存、统计整理及计算分析，是件非常困难的事情，因此统计学尤其是卫生统计学的发展和应用，受到了极大的阻碍。卫生统计学的发展几乎停滞不前，卫生统计学的应用也极有限。近年来，随着计算机科学的飞速发展，计算机的广泛普及，使得大量信息的储存与检索、复杂的数据处理（特别是多因素分析）以及抽样模拟成为可能，大大地促进了卫生统计学的发展，使得卫生统计学的应用越来越广泛，进而极大地促进了医学科研的进步，受到广大医学工作者的关注。

医护人员，在从事临床医护工作的同时，常常需要注意卫生资源的占有和利用情况，工作数量、质量和效果，还需要开展一定的科学研究工作，而所有这些都必须具有一定的卫生统计知识和技能，所以学习卫生统计学是十分必要的。

第二节 统计工作的步骤

上面讲到，对于复杂的客观世界，我们需要用统计学方法，透过现象揭示本质，那么统计工作该怎么做呢?

统计工作和一般科研工作类似，可分为四大步骤，即先要有一个涉及全过程的设计，然后按照设计的要求和规定，搜集资料、整理资料和分析资料。这四大步骤是互相联系的，不

可分割的。任何步骤的不足，都会给统计分析结果带来不利影响。

1. 设计

这是统计工作中最关键的一步。首先要明确研究目的，要对被研究的事物有一定的了解（通过查阅文献、作预试验、预调查等而得到），然后对研究工作的全过程作一个全面的规划。例如明确要搜集哪些资料？人财物是否允许？怎么搜集原始资料？得到资料后如何整理汇总？怎样对汇总资料作进一步的加工？计算哪些统计指标？如何分析这些指标？会得到什么样的结果？所有这些都要经过周密考虑，明确回答，结合实际情况作出科学详尽的计划。

研究设计中，除要规定所搜集的资料类型、搜集资料和分析资料的方法外，还要说明资料质量的控制措施。

研究设计除受研究目的的影响外，还受研究费用及所研究的卫生问题本身等因素的影响。

2. 搜集资料

即按研究设计的规定，及时取得准确完整的原始数据。只有原始数据准确可靠，才能得出可信的结论。医学统计资料主要有三个方面的来源：①统计报表；②医疗卫生工作记录和报告卡片，如病例、出生登记卡和死亡登记卡等；③专题调查和试验。

3. 整理资料

即按研究设计的规定，使原始数据系统化、条理化的过程，以便进一步计算指标，作出分析结论。整理数据的过程一般可分成三步：①核对校正原始数据；②根据研究数据的特征分组；③根据分组情况，将原始数据分别归入相应的组，并做简单的总结。

4. 分析资料

即按研究设计的规定，计算有关指标，进而阐明事物的内在联系和规律。分析资料过程也可分成三步：①计算相关的指标；②绘制适当的统计图表；③进行统计处理，并作出恰当的结论。

第三节　统计资料的类型

统计资料，按照其特点，可分成不同的类型。不同类型的统计资料，其统计处理方法不同。

1. 计量资料

通过度量衡的方法，测定每一个观察单位的某项研究指标的量的大小，而得到的一系列数据资料（即观察值），就称为计量资料。计量资料通常是有单位的，如身高，单位为 cm 或 m；体重，单位为 kg。对于大多数计量资料来说，当观察单位足够多时，这些观察值可以充满某一个数值范围，故属于连续性资料。

2. 计数资料

将全体观察单位，按照某种性质或特征分组，再分别清点各组中观察单位的个数，这样得到的数据资料，就是计数资料。计数资料没有度量衡单位，如果非要加一个单位的话，那就只有“个”或“位”。另外计数资料都是整数，没有小数点，是一种间断性资料。

3. 等级资料

等级资料是介于计量资料和计数资料之间的一种资料，通过半定量方法测量得到。具体来说，将全体观察单位按照某种性质的不同程度分成若干组，再分别清点各组中观察单位的

个数，这样得到的数据资料就称为等级资料。与计量资料相比，等级资料的每一个观察单位没有确切值；与计数资料相比，等级资料的分组有程度的差别，各组按性质的程度大小顺序排列。因此等级资料被称为半定量资料。

需要特别指出的是，对于某一统计资料，其类型不是固定不变的，可以根据统计分析的需要而改变其类型。例如，某患者到医院就诊，当医生怀疑该患者贫血时，会让患者做血红蛋白测定，据此判断该患者是否贫血。这时血红蛋白的测量值，就是计量资料。当医生根据血红蛋白值的大小，诊断该患者是否贫血，这时该资料就变成了计数资料。如果医生要对患者进行治疗，将患者分成轻度、中度、重度或恶性贫血几种类型，这时资料就变成了等级资料。

第四节　统计学中的几个基本概念

1. 变异

性质相同的事物，称为同质的事物。但是即使同质的事物，各个体（也即观察单位）也不是完全相同的。各个体之间的差异，称为变异。例如同年龄同性别儿童的身高，有高有低，不完全相同，这就是个体变异。同质观察单位之间的个体变异，尽管是偶然性（观察单位内外环境的各种因素）的表现，但是事物的重要特征。统计的任务就是要在同质的基础上，通过对个体变异的研究，透过偶然现象，揭示同质事物的本质特征和必然规律。如果事物的各个观察单位之间没有了变异，各个观察单位变成完全一样了，那么统计学就没有存在的必要了。因此变异是统计学存在的基础，统计学就是研究变异的科学。

2. 总体与样本

理论上讲，总体就是根据研究目的而确定的同质研究对象的全体。这是一个非常理论化的概念。实际上，当研究有明确而具体的研究指标时，总体是指性质相同的符合研究要求的所有观察单位的该项变量值的全体。例如研究某地 2000 年正常成人血压的一项研究，其研究对象是该地 2000 年的正常成人，其研究指标是血压值，其研究总体就是该地 2000 年所有正常成人的血压值。当研究没有明确而具体的研究指标时，其研究总体就只能是性质相同的符合研究要求的所有观察单位了。例如，研究某市 1995 年住院分娩妇女的剖腹产情况的一项研究，没有明确而具体的研究指标，这时研究总体就是该市 1995 年住院分娩的产妇。研究总体可以分为有限总体和无限总体。只包括有限数量观察单位的总体，称为有限总体，如研究某校某年级某班同学的体重情况，其总体是该班所有同学的体重，体重数有限，所以其总体属有限总体；而如中国的人口，数量巨大无比，或患有某种疾病的病人数，其数量是无法搞清楚的，针对中国人口或患有某种疾病病人的研究，其研究对象的数量就是巨大无比或不清楚的，其总体就属于无限总体。医学研究的许多研究总体属于无限总体。

在实际工作中，由于研究总体（无论是无限总体，还是有限总体）经常是比较大的，对所有研究对象进行研究是不可能的，一是因为无法找到所有研究对象；二是因为对全体研究对象进行研究，需要花费许多人力、物力、财力和时间；有时研究具有破坏性，如对某种注射剂的检验，检验之后就失去了使用价值，故欲了解该注射剂的质量，不可能对所有该注射剂进行检查，只能检查一部分，所以我们一般不对总体进行研究，而常常通过对样本——从总体中随机抽取的一部分研究对象的研究，用样本的信息来推断总体的特征。为了保证样本研究的准确性，对于总体，样本必须具有足够的代表性。样本要有代表性，挑选样本时必须

遵循随机化原则，就是使总体中每一个个体都有相等的被选机会作为样本。由于样本经常是通过抽样得到的，所以样本研究通常就是抽样研究。此外样本的数量（俗称样本量）要足够大。

常用的选择样本（抽样）方法有概率抽样和非概率抽样。概率抽样就是使总体中每一个个体都有一个已知不为零的被选机会进入样本。概率抽样分为两大类：等概率抽样和不等概率抽样。前者总体中每一个个体被选机会相等，后者每一个个体被选机会不等。等概率抽样又称随机抽样，它能避免抽样过程中人为因素的影响，保证样本的代表性。对于不等概率抽样，可以采用某些加权的方法对不相等的概率作调整。常用的概率抽样方法有：

(1) 简单随机抽样(又称单纯随机抽样)

简单随机抽样是最基本的概率抽样，最直观地体现了抽样的基本原理，它是其它抽样方法的基础。它是对总体的观察单位不进行任何组合，仅按照随机原则直接抽取样本的一种方法。简单随机抽样方法的步骤：第一步，把研究对象总体的各单位编上数字号码；第二步，随机地抽取必要数目的样本。

随机抽取样本的方法有如下三种：

①随机数字表法②抽签法③直接抽选法

本法简单易行，适用于研究对象不多且单纯的调查研究，研究对象多时不适用。

(2) 系统抽样(又称等距抽样或机械抽样)

系统抽样是简单随机抽样的一个变种，其具体做法是：

①将总体的所有个体按顺序排列起来。

②计算抽样间隔 K。抽样间隔是由总体大小和样本大小决定的。

③在头 K 个个体中，用完全随机的方式抽取一个个体，设其所在位置的序号为 k。

④自 k 开始，每隔 K 个个体抽取一个个体，组成样本。

系统抽样简单易行，样本分布更均匀，但不适用于总体排列有规律的情况。

(3) 分层抽样

所谓分层抽样就是先将总体依照一种或几种特征分为几个子总体，每一个子总体称为一层，然后从每一层中随机抽取一个子样本，将它们合在一起，即为总体的样本。

一般来说，用以分层的理想变量是在调查中要加以测量的变量，更多的是选择与调查中欲测量变量高度相关的变量，即将对所要研究的变量有很大影响的因素作为分层变量。

分层抽样时，尽可能使各层内部保持一致，以简化总体的构造，使各层之间有明显的差异性，以便包括总体的各种特性。简而言之，分层抽样的原则是使层间差异大，层内差异小。

(4) 整群抽样(又称聚类抽样)

整群抽样是将总体按照某种标准划分为若干子群体，每一个子群体为一个抽样单位，用随机的方法从中抽取少许子群体，将抽出的子群体中所有个体合起来作为总体的样本。

整群抽样的分层原则与分层抽样不同，它要求群间差异小，群内差异大，能反映总体的变异。

整群抽样的样本单位相对集中，即样本分布比较集中，便于组织调查，节省人力与费用。整群抽样的最大缺点是样本分布不均匀，样本的代表性较差。但由于实施起来比较方便，而且可以节约人财物力，因此，在大规模、大范围的调查中，仍常常采用这种抽样方法。

如果研究对象总体的边界不明确，随机抽样就不能进行。在这种情况下，可采用非随机抽样。非随机抽样是建立在研究人员对总体中某些分子及其个别事件有所了解的基础之上的。同时，在对总体性质有所了解的情况下，非随机地选取若干样本，也属于非随机抽样。依据非随机样本推断总体时，无法确定误差程度，因而获得的推断是不可靠的。定性研究常采用非概率抽样方法。常用的非随机抽样方法有：①偶遇抽样（又称方便抽样）；②主观抽样（又称判断抽样或立意抽样）；③定额抽样（又称配额抽样）；④滚雪球抽样。

3. 抽样误差

我们知道误差可以分为系统误差和随机误差两种。系统误差是指在数据搜集过程中，由于仪器不准确、标准不规范等原因，使得数据倾向性地偏高或偏低。这种误差具有累加性，会因测量次数的增加而增加。随机误差是指在数据搜集过程中，由于一些非人为的偶然因素，使得测量结果或大或小。这种误差，会因测量次数的增加，偏大的测量结果与偏小的测量结果相中和，而变得较小，甚至消除。

抽样误差是指样本指标与总体指标之间的差。只要是抽样研究，就必然存在抽样误差，这有两方面的原因：①研究对象之间存在个体差异即变异。这是抽样误差存在的根本性原因。如果没有变异，各个研究对象都一样，就不可能有抽样误差，也不要进行统计了。②抽样研究只研究一部分研究对象，而不研究全部研究对象，部分不可能等同于全部，也即抽样是抽样误差的直接原因。没有抽样，无论变异存在与否，都不可能产生抽样误差。

抽样误差属于随机误差，但是抽样误差不能被消除，只能被控制。控制抽样误差的方法有①改进抽样方法，使样本更具代表性；②增加样本量，使样本量达到一定水平；（本法有一定的局限性。样本并不是越多越好，存在成本—效益问题。此外增加样本，可能同时增加非抽样误差。）③选择变异程度小的研究指标。

4. 频率与概率

数学上，频率是发生某现象的观察单位数占可能发生该现象的观察单位总数的比值；概率是频率的极限，说明某事件发生的可能性。概率值在 0 到 1 之间。如果某事件的发生概率等于 0，即该事件发生的可能性为 0，则该事件肯定不发生，这类肯定不发生的事件称为不可能事件。若某事件的发生概率等于 1，即该事件发生的可能性为 1（100%），则该事件肯定发生，这类肯定发生的事件称为必然事件。发生概率介于 0 到 1 之间的事件，既可能发生，也可能不发生，这类事件称为随机事件或偶然事件。概率值越大（越接近 1），某事件发生的可能性越大；相反，概率值越小（越接近 0），某事件发生的可能性越小。统计学上认为，某一次抽样研究时不发生小概率事件，而发生大概率事件。当然这不是绝对正确的，因为小概率事件，只是发生的可能性比较小而已，并不是绝对不发生；大概率事件只是发生的可能性比较大而已，并不是一定发生。习惯上常常将概率值小于 0.05 或 0.01 的事件，称为小概率事件；概率值大于 0.05 或 0.01 的事件，称为大概率事件。

习　题

1. 统计工作的基本步骤是什么？
2. 统计资料分成哪几种类型？它们各有哪些特点？这几种类型的统计资料之间有何联系？
3. 什么是总体？什么是样本？样本应具有哪些特点，才能代表总体？
4. 四种基本的概率抽样方法是哪四种？各适用于什么样的研究？
5. 什么是抽样误差？为什么说抽样误差是不可避免的，是必然存在的？如何才能控制或减小抽样误差？

第二章　集中趋势、离散趋势、正态分布及其应用

第一节　计量资料的频数表

为了解计量资料的分布规律，当观察值个数较多时，可将观察值及其相应的频数列表，编制成频数分布表，简称频数表（frequency table）。

一、频数表的编制

例 2.1　1998 年某校 100 名 18 岁健康女大学生的身高（cm）资料如下，试编制频数表。

165.1	165.1	159.5	166.1	162.8	166.2	170.2	165.2	163.0	161.5
169.6	164.0	166.3	167.5	163.6	166.1	158.4	169.0	161.3	162.6
163.0	159.9	168.5	166.0	164.2	156.8	163.0	162.3	165.0	158.3
169.3	160.6	167.3	158.2	161.8	160.6	162.2	162.8	164.2	165.8
160.9	169.1	161.0	161.2	160.7	161.5	162.9	163.4	168.9	170.5
162.0	162.5	163.7	155.6	159.4	161.9	169.7	164.0	162.6	167.5
166.5	171.2	167.1	158.5	163.8	166.6	163.5	164.6	160.4	165.1
165.9	157.1	157.2	154.7	158.0	164.3	162.7	163.1	165.2	164.5
165.4	165.8	163.1	168.2	158.2	167.2	159.9	161.2	164.5	162.8
156.3	168.0	159.0	167.5	173.6	162.0	167.4	170.6	165.0	166.8

编制步骤如下：

1. 求全距　找出观察值中的最大值、最小值，二者的差值即全距（或极差）。本例中最大值为 173.6cm，最小值为 154.7cm，故全距 = 173.6cm - 54.7cm = 18.9cm。

2. 确定组数、组距并写出组段　频数表的组数是根据研究目的和观察例数确定的，一般设 8 ~ 15 个组段，例数较少时组数可相应少些，例数较多时组数可酌情多些。

常取全距的 1/10 取整作为组距，组距为相邻两组段下限值之差，各组段的组距可相等，也可不等。相等组距可用极差/组数来估计。组距相等时，确定组距通常用全距/10（10 为预计分 10 个组段），为了计算方便常取整数。如本例中若取组数为 10，则 18.9/10 = 1.89，取整为 2cm，即组距定为 2cm。

根据组距写出各组段的起点，即下限，各组段的终点，即上限。第一组段要包括最小观察值，最后一个组段要包括最大观察值。一般只写出各个组段的下限，不写出其上限，只有最后一组同时写出上、下限。如表 2.1 第（1）栏："154 ~ "组段，包括了身高在 154 至未

满 156 的观察值。

3. 列表划记。列成表形式，将原始数据采用划记法计数，得到各个组段的观察单位数，这就完成了频数表。见表 2.1。

二、频数分布的两个特征

从频数表可以粗略地看出频数分布的两个重要特征：集中趋势（central tendency）和离散趋势（tendency of dispersion）。如表 2.1，100 名 18 岁健康女大学生的身高分布规律是：身高向中央部分（即 162～组段）集中，且身高在 162～164cm 及接近 162～164cm 者居多，这反映的是集中趋势；从中央到两侧（即由中等身高到较矮或较高）频数分布逐渐减少，这反映的是离散趋势。

三、频数分布的类型

频数分布可分为对称分布和偏态分布两种类型。对称分布是指集中位置在正中，左右两侧频数分布大致对称。如表 2.1，该 100 名 18 岁健康女大学生身高的频数分布是以均数（163.8cm）为中心，低于均数的人数与高于均数的人数大致相等。偏态分布是指集中位置偏向一侧，频数分布不对称，若集中位置偏向数值小的一侧，为正偏态分布，若集中位置偏向数值大的一侧，则为负偏态分布。不同类型的分布，应采用相应的描述性指标和统计分析方法。

表 2.1　1998 年某校 100 名 18 岁健康女大学生身高（cm）的频数分布

身高组段 (1)	划记 (2)	频数 (3)
154～	丅	2
156～	正	4
158～	正正一	11
160～	正正下	13
162～	正正正正丅	22
164～	正正正正	19
166～	正正正	15
168～	正正	9
170～	正	4
172～174	一	1
合计	——	100

四、频数表的用途

1. 频数表可揭示资料的分布特征和分布类型，因而在文献报道中，常将频数表做为陈述资料的形式。

2. 便于进一步计算和分析。

3. 便于发现特大或特小的可疑值。例如，有时在频数表中，连续出现几个组段的频数为 0 后，又出现了一些频数，此时，就要怀疑这些频数对应的特大值或特小值的准确性，应

进一步检查、核对。

第二节　集中位置的描述

对计量资料进行分析时，经常用平均数来反映其集中趋势。平均数（average）是表示一组性质相同的变量值的集中位置或平均水平的指标。常用的平均数有算术均数、几何均数和中位数。

一、算术均数（arithmetic mean）

算术均数简称均数（mean），总体均数用希腊字母 μ 表示，样本均数用 $\overline{X}$ 表示。均数反映一组性质相同的观察值在数量上的平均水平。

1. 均数的计算方法

（1）直接法。即将所有观察值 X_1，X_2，$X_3 \cdots X_n$ 直接相加再除以观察值的个数 n，写成公式为：

$$\overline{X} = \frac{X_1 + X_2 + X_3 \cdots + X_n}{n} = \frac{\sum X}{n} \tag{2.1}$$

式中，$\sum$ 为求和的意思，$\sum X$ 为各观察值的总和。

例 2.1　10 名 18 岁健康女大学生的身高（cm）分别为 165.1，159.5，166.1，162.8，166.2，170.2，165.2，163.0，161.5，169.6，求平均身高。

$$\overline{X} = \frac{165.1 + 159.5 + 166.1 + 162.8 + 166.2 + 170.2 + 165.2 + 163.0 + 161.5 + 169.6}{10}$$

$$= 164.9\ (\text{cm})$$

（2）加权法

当观察值个数较多时，用公式（2.1）计算均数比较麻烦，为计算方便，可先将各观察值分组列成频数表，用加权法求均数。加权法计算公式为：

$$\overline{X} = \frac{f_1 X_1 + f_2 X_2 + f_3 X_3 + \cdots + f_n X_n}{f_1 + f_2 + f_3 + \cdots + f_n} = \frac{\sum fX}{\sum f} \tag{2.2}$$

式中，X_1，X_2，X_3，……，X_n 分别为各组段的组中值，即各组段的下限与相邻较大组段的下限相加除以 2。如“158 -”组段的下限为 158，相邻的下一组为 160 - 组段，其下限为 160，组中值 X_3 =（158 + 160）/2 = 159，余仿此。实际上，公式（2.2）是作为公式（2.1）的近似计算，这里我们用了各组数据的组中值 X 来代表各组数据的观察值。f_1，f_2，f_3，……，f_n 分别为各组段的频数，这里的 f 起了“权重”的作用，它权衡了各组中值由于频数不同对均数的影响，即频数多，权重大，作用也大；频数小，权重小，作用也小，故本法称为加权法。

例 2.2　对表 2.1 资料用加权法求平均身高。

表 2.2　100 名 18 岁健康女大学生身高均数的计算（加权法）

身高组段 (1)	频数 f (2)	组中值 X (3)	fX (4) = (2)·(3)
154 ~	2	155	310
156 ~	4	157	628

身高组段 (1)	频数 f (2)	组中值 X (3)	fX (4) = (2)·(3)
158 ~	11	159	1749
160 ~	13	161	2093
162 ~	22	163	3586
164 ~	19	165	3135
166 ~	15	157	2505
168 ~	9	169	1521
170 ~	4	171	684
172 ~ 174	1	173	173
合计	100 ($\sum f$)		16384 ($\sum fX$)

2. 均数的应用

均数是用来描述一组性质相同的观察值的平均水平的指标。均数适用于对称分布资料，因为这时均数位于分布的中心，最能反映分布的集中趋势。特别是正态分布资料，均数更有其重要作用（见下章）。对于偏态分布资料，均数则不能很好地反映分布的集中趋势，可用几何均数或中位数等描述。

二、几何均数（geometric mean）

某些资料如抗体的滴度、细菌计数等，其频数分布呈明显偏态，资料中的少数数据过分偏大，各观察值之间常呈倍数变化（等比关系），则算术均数对这批资料的集中趋势或平均水平的代表性较差，因为个别大数据使算术均数偏向大的一边。这时，宜用几何均数反映其平均水平。几何均数用 G 表示。

1. 几何均数的计算方法

（1）直接法　即将 n 个观察值（X_1，X_2，X_3，……X_n）的乘积开 n 次方，

写成公式为：

$$G = \sqrt[n]{X_1 \cdot X_2 \cdot X_3 \cdots\cdots X_n} \tag{2.3}$$

写成对数形式为：

$$G = \lg^{-1}\left(\frac{\lg X_1 + \lg X_2 + \cdots + \lg X_n}{n}\right) = \lg^{-1}\left(\frac{\sum \lg X}{n}\right) \tag{2.4}$$

式中：$\sum \lg X$　为变量值取对数后的和，

$\lg^{-1}$为反对数符号。

例 2.3　5 人的血清滴度为 1:2，1:4，1:8，1:16，1:32，求平均滴度。

本例先求平均滴度的倒数：

$$G = \sqrt[5]{2 \times 4 \times 8 \times 16 \times 32} = 8$$

或

$$\lg G = \frac{\lg 2 + \lg 4 + \lg 8 + \lg 16 + \lg 32}{5} = 0.903$$

$$G = \lg^{-1} 0.903 = 8$$

故平均滴度为1:8。

(2) 加权法。当观察值个数 n 较多时，先将观察值分组列成频数表，再用公式 (2.5) 计算：

$$G = \lg^{-1}\left(\frac{f\lg X}{\sum f}\right) \tag{2.5}$$

例2.4 某医院检验一批肝炎患者的相关抗原 (HAA)，滴度分布资料如表2.3第 (1)、(2) 栏，求平均滴度。

表2.3 平均滴度的计算

抗体滴度 (1)	人数，f (2)	滴度倒数，X (3)	$\lg X$ (4)	$f\lg X$ (5) = (2)·(4)
1:1	4	1	0.0000	0.0000
1:2	5	2	0.3010	1.5050
1:4	7	4	0.6021	4.2147
1:8	3	8	0.9031	2.7093
1:16	2	16	1.2041	2.4082
1:32	2	32	1.5051	3.0102
合计	23 ($\sum f$)			13.8474 ($\sum f\lg X$)

$$\lg G = \frac{\sum f\lg X}{\sum f} = \frac{13.8474}{23} = 0.6021$$

$$G = \lg^{-1}0.6021 = 4$$

故平均滴度为1:4。

2. 几何均数的应用

(1) 几何均数常用于呈等比数列的资料，即变量值呈倍数关系，如血清中抗体滴度和血清凝集效价等，尤其适用于对数正态分布的资料。对数正态分布即原始数据呈偏态分布，经对数变换后（用原始数据的对数值 $\lg X$ 代替 X）服从正态分布；

(2) 计算几何均数时应注意：a. 观察值中不能有0；b. 观察值不能同时有正值和负值，若全是负值，计算时可把负号去掉，得出结果后再加上负号。

三、中位数 (median) 及百分位数 (percentile)

中位数（简记为 M）是将数据按大小顺序排队后，位置处在最中间的那个数值。中位数能将数据分成两半，一半数据比中位数大，另一半数据比中位数小。

百分位数也是一种常用来描述计量资料特征的统计指标，用 P_X 表示，指将 n 个观察值从小到大依次排列，再把它分成100等份，对应于 $X\%$ 位的数值即为第 X 百分位数。P_{50} 即中位数（第50百分位数），因此，中位数是一个特定的百分位数。

1. 中位数的计算方法

(1) 直接法 将观察值按大小顺序排列，按式 (2.6) 或式 (2.7) 计算。

n 为奇数时

$$M = X\left(\frac{n+1}{2}\right) \tag{2.6}$$

n 为偶数时

$$M = [X(\tfrac{n}{2}) + X(\tfrac{n}{2}+1)] /2 \tag{2.7}$$

式中，n 为观察值的总个数，$\left(\frac{n+1}{2}\right)$，$\left(\frac{n}{2}\right)$，$\left(\frac{n}{2}+1\right)$为观察值按顺序排列后的位次，$X(\frac{n+1}{2})$，$X(\frac{n}{2})$，$X(\frac{n}{2}+1)$为相应位次上的观察值。

例 2.5　临床观察 7 名某病患者，其潜伏期（天）各为：2，3，3，5，6，9，16，求其中位数。

本例 $n=7$，为奇数，按式（2.6），中位数所在位次为$\frac{7+1}{2}=4$，即第四位数值，即上面由小到大顺序排列的数列中的 4 这个数，

$$M = X(\tfrac{7+1}{2}) = X_4 = 5\text{（天）}$$

例 2.6　若在例 2.5 增加一例，其潜伏期为 20 天，求中位数。

本例 $n=8$，为偶数，按式（2.7），居于中间位置的两个数为 5 和 6，那么中位数就是：

$$M = [X(\tfrac{8}{2}) + X(\tfrac{8}{2}+1)] /2 = [X_4 + X_5] /2 = \frac{5+6}{2} = 5.5\text{（天）}$$

（2）频数表法

步骤为：

a. 列出频数表；

b. 按所分组段，由小到大计算累计频数和累计频率，如表 2.4 第（3）、（4）栏。累计频数即每组例数与其以前各组例数之和，如第一组累计频数为 25，第二组累计频数为 25 + 58 = 83。累计频率即每组频率与其以前各组频率之和。

然后确定第 X 百分位数所在组段。从累计频率一栏可以找出第 X 百分位数所在组段：首先从小到大观察各组段的累计频率，当累计频率刚刚大于或等于 $X\%$ 的累计频率时，这个累计频率所对应的组段为第 X 百分位数所在组段。如由表 2.4 第（4）栏可以看出，中位数位于累计频率为 50.6% 的那个组段，即“12—”组段内；第 95 百分位数位于累计频率为 96.3% 的那个组段，即“48—”组段内。

c. 按式（2.8）、式（2.9）求百分位数 P_X 和中位数 M。

$$P_X = L + \frac{i}{f_X}(n \cdot X\% - \sum f_L) \tag{2.8}$$

式中，P_X——第 X 百分位数；

L——P_X 所在组段的下限；

f_X——P_X 所在组段的频数；

i——该组段的组距；

$\sum f_L$——小于 L 各组段的累计频数。

对于中位数　$f_X = f_{50}$

$$n \cdot X\% = n \cdot 50\% = \frac{n}{2},$$

因此

$$M = P_{50} = L + \frac{i}{f_{50}}(\frac{n}{2} - \sum f_L) \tag{2.9}$$

例 2.7　现有 164 例某种沙门氏菌食物中毒潜伏期的资料（表 2.4），试计算中位数和第 95 百分位数。