教育部职业教育与成人教育司推荐教材

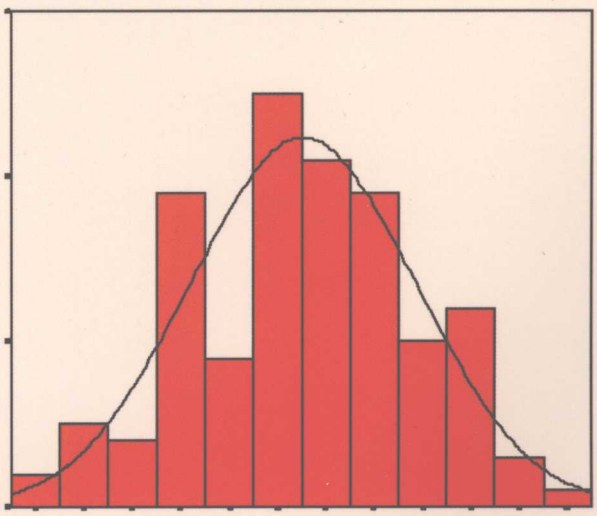
全国卫生职业院校规划教材

供中高职护理、英语护理、助产、检验、药剂、卫生保健、社区医学、药学、眼视光、康复、口腔工艺、影像技术、中医、中西医结合等专业使用

医学统计方法

（第二版）

师明中　封苏琴　主编



科学出版社
www.sciencep.com

教育部职业教育与成人教育司推荐教材
全国卫生职业院校规划教材

供中高职(共用课)护理、英语护理、助产、检验、药剂、卫生保健、社区医学、药学、眼视光、康复、口腔工艺、影像技术、中医、中西医结合等专业使用

医学统计方法

(第二版)

主　编　师明中　封苏琴

副主编　胡晓东　张勤国

编　委　(按汉语拼音排序)

段爱旭　大同大学医学院

封苏琴　常州卫生高等职业技术学校

胡晓东　聊城职业技术学院

黄　涛　井冈山学院医学院

李俊萍　邢台医学高等专科学校

师宏儒　大同大学职业技术学院

师明中　大同大学医学院

张勤国　襄樊职业技术学院

赵　宏　营口市卫生学校

郑　扬　汕头市卫生学校

科学出版社

北京

内 容 简 介

本书是教育部职业教育与成人教育司推荐教材和全国卫生职业院校规划教材之一，针对在医学、卫生科研、医疗及护理工作实践中对医学统计的需求，主要讲述了单变量数值变量资料和分类变量资料以及等级资料的统计分析方法、双变量的直线相关与回归统计分析方法、秩和检验、方差分析以及用于表达统计分析结果的统计表与统计图等。本书在第一版的基础上，对各种资料的统计分析方法做了适当的拓展，考虑到计算机的普及与统计软件的开发，对医学统计方法的上机实习与应用也做了适当的拓展和提高。本书内容结构严谨，版式生动新颖，各章有学习目标、案例分析、链接、小结及目标检测，书后附有医学统计方法教学基本要求、常用附表及目标检测参考答案，便于学习与实习。

本书适合中高职护理、英语护理、助产、检验、药剂、卫生保健、社区医学、药学、眼视光、康复、口腔工艺、影像技术、中医、中西医结合等专业学生使用，也可作为在职岗位培训及执业护士基础理论考试的教材和参考书。

图书在版编目(CIP)数据

医学统计方法 / 师明中，封苏琴主编. —2 版. —北京：科学出版社，2007
教育部职业教育与成人教育司推荐教材·全国卫生职业院校规划教材
ISBN 978-7-03-019829-7

Ⅰ. 医… Ⅱ. ①师…②封… Ⅲ. 医学统计－教材 Ⅳ. R195.1

中国版本图书馆 CIP 数据核字（2007）第 134942 号

责任编辑：郭海燕 / 责任校对：钟 洋
责任印制：刘士平 / 封面设计：黄 超

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
铭浩彩色印装有限公司印刷
科学出版社发行 各地新华书店经销

*

2003 年 8 月第 一 版 开本：850×1168 1/16
2007 年 8 月第 二 版 印张：8
2007 年 8 月第五次印刷 字数：201 000
印数：18 001—23 000

定价：16.00 元

（如有印装质量问题，我社负责调换〈环伟〉）

技能型紧缺人才培养培训教材
全国卫生职业院校规划教材
共用课教材建设指导委员会委员名单

主任委员 刘 晨

委 员（按汉语拼音排序）

陈劲松 四川省卫生学校
陈 均 上海市公共卫生学校
陈 沁 广州医学院护理学院
代凤兰 聊城职业技术学院
丁 玲 沧州医学高等专科学校
封苏琴 常州卫生高等职业技术学校
高健群 宜春职业技术学院
官素琼 玉林市卫生学校
胡希俊 沧州医学高等专科学校
纪 霖 辽源市卫生学校
李长驰 汕头市卫生学校
李怀珍 沧州医学高等专科学校
李 军 山东医学高等专科学校
李晓惠 深圳职业技术学院
李小龙 岳阳职业技术学院
蔺惠芳 中国协和医科大学护理学院
罗志君 四川省卫生学校
牛彦辉 定西市卫生学校
潘道兰 达州职业技术学院
潘凯元 海宁市卫生学校
覃琥云 四川省卫生学校
邱志军 岳阳职业技术学院
任海燕 内蒙古医学院护理学院
石海兰 太原市卫生学校
师明中 大同大学医学院
史学敏 深圳职业技术学院
宋金龙 三峡大学护理学院
孙巧玲 聊城职业技术学院
汪洪杰 安徽医学高等专科学校
王者乐 上海职工医学院
吴丽文 岳阳职业技术学院
肖京华 深圳职业技术学院
徐冬英 广西中医学院护理学院
许练光 玉林市卫生学校
杨玉南 广州医学院护理学院
姚军汉 张掖医学高等专科学校
余剑珍 上海职工医学院
曾志励 广西医科大学护理学院
张金生 聊城职业技术学院
张 宽 嘉应学院医学院
张妙兰 忻州市卫生学校
赵 斌 四川省卫生学校
钟埃莉 成都铁路卫生学校
钟 海 四川省卫生学校
周 琦 广西中医学院护理学院
邹玉莲 岳阳职业技术学院

第一版前言

21世纪是信息技术与生命科学快速发展的时代，医学统计一方面作为两者沟通的桥梁，另一方面也促进了两者的发展与进步。医学信息的正确搜集、整理与分析对临床医学、预防医学、基础医学、医院信息管理与卫生经济的发展都起着重要的促进作用。

本教材为“面向21世纪全国卫生职业教育系列教改教材”的共用课教材，它的使用对象为护理与医学类高、中职各专业的学生。“面向21世纪全国卫生职业教育系列教改教材”编辑委员会提出了教材的编写要求，要坚持“三个基本原则”，即贴近学生、贴近社会、贴近岗位，保证教材的“五个基本特性”，即科学性、思想性、实用性、可读性与创新性，要做到“三个体现”，即要体现社会对卫生职业教育的需求和专业人才能力的培养的要求，体现与学生心理取向和知识、方法、情感前提的有效连接，体现开放发展的观念及其专业思维、行为的方式。

根据上述要求，我们教材的编者认为，一是要保证教材正文内容的系统性，满足学生培养目标的“基本”、“必需”、“够用”等特点，不盲目地求“全”求“大”。二是教材通过“链接”与“接口”的方式使其具有一定的扩展性，考虑到现代教育技术的使用和各学科科研发展的需求，还要兼顾高、中职学生的不同需要，我们适当地拓展了部分内容，比如四格表的精确概率法，χ^2 检验中行×列表的 χ^2 值的分割计算与检验水准的校正；计算机统计软件在医学统计中的应用与实习内容。三是要突出理论与实际结合，以现代教学技术和手段为依托，多种教学方法的综合运用和相互配合，以达到课堂教学的最优化；尽可能深入浅出地对医学统计的基本概念、基本理论、基本知识和基本技能，尤其是对各种统计方法的适用范围、使用方法及统计结果的理解作正确解释，而不拘泥于大量的计算过程；四是要注重学生的能力培养，使学生正确思维、善于学习，具有创新精神，体现“教为主导，学为主体”，在和谐、自然、愉悦、轻松、贴近实际的良好教育气氛中发展学生智能，促进学生个性的健康成长。

限于我们的知识和能力以及时间紧张，本书还可能有许多不足之处，希望广大师生与读者提出批评和建议，我们愿意虚心听取，以便改进。

衷心感谢“面向21世纪全国卫生职业教育系列教改教材”编写委员会与科学出版社在专业理念上与内容上的严谨把关，感谢各参编学校领导的大力支持。

张勤国

2003年5月

第二版前言

《医学统计方法》自2003年出版以来，经过各兄弟院校的教学实践，一致认为它符合中高职教育的培养目标与教学计划，是符合教学规律的，因此本书原则上保留了第一版原教材的基本结构。

为了满足中高职不同院校的教学要求，适用更多的专业，扩大使用范围，本书在第一版的基础上做了适当的调整，主要表现：绪论简明扼要；将数值变量资料（计量资料）和分类变量资料（计数资料）的统计分析都一分为二为“统计描述”与“统计推断”；“方差分析”虽然属于数值变量资料的统计推断，但考虑到它不是中高职教育的教学重点，故仍单独列为一章，供学生了解和参考；考虑到学生的认识过程，在先接触了统计分析过程中的大量统计表及统计图以后，再将“统计表和统计图”单独列为本书中的第9章；对“秩和检验”、“直线相关与回归”做适当的修改后仍单独列章。此外，电子计算机的普及为统计数据的搜集、整理和分析提供了十分便利的条件，因此，第二版仍保留了“医学统计方法上机实习与应用”内容。

另外，为方便查询和自学，本书附有医学统计方法常用值表，它们分别是标准正态曲线下的面积、t 界值、百分率的可信区间、χ^2 界值、T 界值、H 界值、F 界值、q 界值、相关系数 r 界值、等级相关系数 r_s 界值。

本次再版，除汲取了第一版的编写经验外，还引用了有关教材、专著及杂志期刊的部分资料，在此一并致以衷心的感谢。本书在编写过程中，封苏琴、胡晓东、张勤国等兄弟院校的老师提出了许多宝贵的意见和建议，各参编院校领导给予了大力支持，大同大学段爱旭、师宏儒两位老师协助主编做了大量的工作，在此一并表示衷心的感谢。

本书第1章由李俊萍、赵宏、师明中编写，第3章、第4章由段爱旭编写，第2章、第5章由师明中、师宏儒编写，第6章、第7章由封苏琴编写，第8章由黄涛编写，第9章由郑扬编写，第10章由胡晓东编写，全书最后由师明中做统一审定工作。

本次再版，力图提高质量，但限于业务水平和编写时间仓促，难免存在不少缺点和错误，恳请使用本书的广大师生和读者批评指正。

师明中

2007年8月

目　录

第1章 绪 论

学习目标

1. 说出医学统计学的几组基本概念及小概率事件的意义

2. 能够判断资料的性质并说出不同资料的特征

3. 叙述统计工作的基本步骤

第1节 医学统计的意义与基本概念

一、医学统计的意义

医学统计是认识医学现象数量特征的重要工具，是运用概率论与数理统计的基本原理与方法，进行医学科研设计和资料的搜集、整理、分析与推断的过程。

医学研究的主要对象是人群健康状况及影响健康的诸多因素。人群健康与疾病是一种复杂的生物现象和社会现象。生物现象的变异很大，同一性别、同一年龄的人，其各种指标的正常值变异范围很大；各种影响因素也极其复杂，它不仅表现为生物因素方面，也表现在社会心理因素方面。医学统计则可透过偶然现象来探测其规律性。因此，医学统计方法已成为医学科学研究的重要前提和手段。

二、医学统计的基本概念

（一）总体与样本

（1）总体（population）：根据研究目的所确定的同质研究对象所有观察单位某种变量值的集合。

（2）样本（sample）：从总体中随机抽取的一部分观察单位，它是总体中有代表性的一部分。

（二）个体与变异

（1）个体（unit）：组成总体的每个具体观察单位。每项指标的测得值称为观察值（observed value），或者变量值（variable value），通常用英文字母 X 来表示。

（2）变异。同一性质的变量值（即观察值），其大小可能参差不齐，这种变量值之间的差异在统计学上称为变异（variation）。

（三）参数与统计量

（1）参数（parameter）：根据分布特征而计算的总体指标。如总体均数（μ）、总体率（π）、总体标准差（σ）等。

（2）统计量（statistic）：由总体中随机抽取的样本所计算的统计指标。如样本均数（$\bar{x}$）、样本率（p）、样本标准差（s）等。

上述几个统计学概念是密切联系的。例如，要调查某年某地区12岁健康男孩的身高水平，那么该地区同年龄的全部12岁健康男孩的身高值就是一个总体；该地区具体的每一个12岁健康男孩就是个体，其身高值就是观察值或变量值；从该地区随机抽取120名12岁健康男孩进行身高测量，这120名男孩的身高值就是样本；每个12岁健康男孩的身高不尽相同，这种身高值间的差异就是变异。通过计算这120名12岁健康男孩的平均身高（即统计量），就可以运用统计方法估计出该地区该年度全部12岁健康男孩的身高水平（即参数）。

从总体中随机抽样，用样本指标估计总体指标的方法，称为抽样研究方法。在抽样过程中为了避免主观意愿或客观无意识的偏性影响，使样本能够充分反映总体的情况，必须遵循“随机化”和样本含量足够大的原则。

（四）误差

误差（error）指测得值与真值之差，或样本

指标与总体指标之差。误差主要指下列三种：

（1）系统误差（systematic error）：在搜集资料的过程中，由于仪器不准、标准试剂未经校正、医生掌握疗效标准偏高或偏低等原因，可使观察结果呈倾向性的偏大或偏小。系统误差可影响原始资料的准确性，应力求避免。如已发生，则要查明原因，予以校正。

（2）随机测量误差（random measure error）：在搜集资料过程中，即使方法统一，仪器及标准试剂已经校正，但由于偶然因素的影响，造成同一对象多次测定的结果不完全一致，这种误差往往没有固定的倾向，而是有的偏高、有的偏低。随机测量误差是不可避免的，但应努力做到仪器性能及操作方法稳定，使其控制在一定的允许范围内。必要时，可作统计处理。

（3）抽样误差（sampling error）：即使消除了系统误差，并把随机测量误差控制在允许范围内，样本指标与总体指标间仍可能有差异。这是由于个体变异造成的，如居住在同一地区同年龄的12岁健康男孩，他们的身高总是有高有矮，这些个体差异是客观存在、不可避免的。因此，从该地区12岁健康男孩中随机抽取一个120人的样本，如算得他们的平均身高为143.10cm，这个样本指标不一定恰好等于该地区所有12岁健康男孩的真实平均身高，这就是抽样误差。抽样误差有一定的规律性，研究和运用抽样误差的规律，进行调查或实验设计与资料分析，是医学统计的重要内容之一。

（五）概率

1. 事件 物质世界处于普遍联系与相互制约的状态之中。作为这种联系和制约的最简单的情形是观察在某一确定条件之下所发生的现象，这种现象称为事件（event）。在进行观察时，有下列三种情形：

（1）必然事件（certain event）：在一定条件下必然出现的现象。通常用字母 U 或 Ω 表示。例如，在标准大气压下，水加热到100℃时必然沸腾；人在没有氧气的环境中必然要死亡等。

（2）不可能事件（cannot event）：在一定条件下必然不会出现的现象。通常用字母 V 或 Φ 表示。例如，“在标准大气压下，水加热到100℃时不沸腾”、“人能在没有氧气的环境中存活”等，都是不可能事件。

（3）随机事件（random event）：是在一定条件下可能出现，也可能不出现的现象。通常用字母 $A, B, C, \cdots$ 或 $A_1, A_2, A_3, \cdots$ 表示。例如，病者对药物的反应，可能有效，也可能无效；新生儿可能是男婴，也可能是女婴；投掷硬币后可能呈正面，也可能呈背面；小麦播种后，每粒种子可能发芽，也可能不发芽，等等，都是随机事件。

2. 频率（frequency） 对样本而言的，在相同条件下进行 n 次重复试验，事件 A 发生数为 $x(x \leqslant n)$，则 x 与 n 的比 (x/n) 是事件 A 的频率。

3. 概率（probability） 对总体而言的，它是描述某事件发生可能性大小的一个度量。现通过具体例子，以表述其统计定义。历史上，有些人做过成千上万次投掷硬币的试验，其试验的记录见表1-1：

表1-1 投掷硬币的历史试验记录

试验者	投掷次数/n	出现正面朝上的次数/x（即频数）	频率/(x/n)
Demorgan	2048	1061	0.518 1
Buffon	4040	2048	0.506 9
Pearson	12 000	6019	0.501 6
Pearson	24 000	12 012	0.500 5

从表1-1可知，随着投掷次数 n 的增大，频率 (x/n) 愈来愈稳定在0.5左右，因此0.5这个数值反映了投掷硬币出现正面朝上的概率。由此得出概率的统计定义：在多次重复进行同一试验时，随机事件 A 发生的频率所稳定接近的值 P，称为随机事件 A 的概率。

从概率统计定义可得出下列基本性质：①必然事件的概率等于1，即 $P(U$ 或 $\Omega)=1$；②不可能事件的概率等于零，即 $P(V$ 或 $\Phi)=0$；③任何随机事件的概率都在0与1之间，即 $0 \leqslant P(A, B, C, \cdots$ 或 $A_1, A_2, A_3, \cdots) \leqslant 1$。由此可见，某事件发生的概率愈接近于零，表示该事件发生的可能性愈小；概率愈接近于1，表示该事件发生的可能性愈大。习惯上常将 $P \leqslant 0.05$ 的事件称为小概率事件，表示该事件发生的可能性很小。统计分析的结论常以某

事件发生的概率 P 值的大小得出的，如在现代医学文献中经常会看到 $P>0.05$，$0.01<P\leqslant 0.05$，$P\leqslant 0.01$ 等，以表达统计分析结果。

第2节 统计资料的类型

统计资料可分为数值变量资料（计量资料）、分类变量资料（计数资料）和等级资料三种类型。各种资料又可根据需要进行相互转化。不同类型的资料宜采用不同的统计分析方法。

1. 数值变量资料 对每个观察单位用定量方法测定某项指标量的大小，所得的资料称为数值变量资料（计量资料），其一般有度量衡单位。例如，身高（cm）、体重（kg）、脉搏（次/分钟）、血压（kPa 或 mmHg）、白细胞数（个/L）等数值，都属于数值变量资料。这类资料的统计描述有平均指标、变异指标等。统计分析方法有 t 检验、方差分析、相关与回归等。

2. 分类变量资料 将观察单位按某一属性来分类计数的资料，称为分类变量资料（计数资料）。例如，临床治疗的有效与无效；化验结果的阳性与阴性；人群血型的 A，B，O，AB 等都属于分类变量资料。这类资料常用相对数（率、构成比等）作为统计描述指标，用卡方检验等作为假设检验的分析方法。

3. 等级分组资料 将观察单位按某一属性的不同程度分组计数，所得各组的观察单位数称为等级资料。等级资料是介于分类变量资料与数值变量资料之间的一种资料。例如，临床治疗效果按照痊愈、显效、好转、无效、恶化等级分组，然后清点每组病人数；化验结果按照反应程度的 -、±、+、+ +、+ + + 等级分组，然后清点各组病人数等，都属于等级资料。这类资料与分类变量资料不同，属性的分组有程度的差别，各组按大小顺序排列；与数值变量资料也不同，每个观察单位有数量上的差别，但不确切，因而等级资料又称为半计量资料。等级资料的假设检验分析，常用秩和检验等。

根据研究分析的需要，数值变量资料、分类变量资料和等级资料之间可以互相转化。例如，年龄（岁）是一个数值变量资料，如按年龄大小分为成年人与非成年人则转化成分类变量资料；如果再分为婴儿、幼儿、儿童、少年、青年、壮年和老年人，则转化成等级资料。又如，每个人的血红蛋白，原属数值变量资料，若按血红蛋白的正常与异常分为两组，清点各组人数，就成为分类变量资料；若将血红蛋白按含量（g/L）的多少分为五个等级：<60（重度贫血）、60 ~（中度贫血）、90 ~（轻度贫血）、120 ~160（血红蛋白正常）、>160（血红蛋白增高），清点各等级人数，则成为等级资料。

分类变量资料和等级资料转化为数值变量资料是将具有相同属性的事物，按其顺序、轻重、大小、主次标以数码。例如，在多变量分析中，定性指标（计数资料）数量化时，将无病和有病分别取为 0 和 1；或者将上述血红蛋白量的五个等级（等级资料）分别取为 1，2，3，4，5，这时分类变量资料或等级资料则转化为数值变量资料。

第3节 统计工作的基本步骤

统计工作的基本步骤，包括统计设计、搜集资料、整理资料和分析资料四个步骤。这四个步骤是互相联系、不可分割的，任何步骤的缺陷都会影响统计分析的结果。

一、统计设计

在进行医学科学研究之前，首先要明确研究目的，并对工作的全过程作以全面的设想，制定出完整、全面的研究计划。研究者要对被研究的事物有一定的了解，可根据以往工作的实践经验和查阅的文献，并通过预调查或预试验掌握较多的信息后，再制定计划。研究计划应包括研究目的、技术路线、方法、人力、财力、组织等项目。

制定好计划后，则要进一步做好实验设计或调查设计。按研究者是否对观察对象施加干预（即处理因素），科研可分为调查与实验两大类。以人群为观察对象的实验，通常称为试验，如临床试验、现场试验等。而实验设计的基本步骤中确定可比的实验组和对照组是非常重要的。所以一个良好的统计设计，应该做到科学、周密、简明，用尽可能少的人力、物力和时间，获得尽可能多的与研究有关的准确数据。

确定可比的对照组

设立对照组的目的是为了排除与研究无关的外在因素的影响，对照组和实验组应在尽可能相同的条件下进行观察，使结果具有可比性。

实验研究中常用的设立对照的方法：

（1）空白对照。不给任何干预的对照。

（2）实验对照。与实验组操作相同、但与处理效应无关的对照，通常用于有损伤、有刺激的动物实验。如，实验动物注射药物、对照组动物注射无药理作用的生理盐水，目的是使实验组和对照组接受损伤和刺激相同。

（3）安慰剂对照。安慰剂的主要成分是乳糖、淀粉、生理盐水，不含任何有效药物，但其外型、大小、味道与试验药相同。对照组使用安慰剂，目的是保持对照组与试验组患者心理作用对疗效评价影响的一致。

链接

二、搜集资料

搜集资料指根据统计全过程设计所提出的要求，实施有关资料的搜集工作。

（一）资料的来源

医学与卫生统计资料的主要来源，有下列三个方面：

（1）统计报表。医疗卫生工作统计报表是根据国家规定的报告制度，由医疗卫生机构定期逐级上报的。它是提供居民健康状况和医疗卫生机构工作的主要基础资料，为拟定卫生工作计划与措施以及检查和总结工作提供依据，如医院工作报表、居民病伤死亡原因报表、疫情报表等。

（2）医疗卫生工作记录和报告。医院各科门诊病历、住院病历、健康检查记录、各种医疗与检验记录以及传染病报告卡等，都是统计工作的重要原始资料。应注意原始资料有否漏填、重复和项目填写不清等情况。

（3）专题调查或实验。当统计报表、医疗卫生工作记录和报告的资料不能满足研究需要时，可组织专题调查或实验研究，如糖尿病调查、高血压调查、某种药物的疗效观察等。

（二）统计资料的要求

笔记栏

原始资料是统计工作的基本依据，对所需资料应作严格审查，并做到：

（1）资料完整、准确和及时。“完整”是指调查单位数量的完整，调查项目应完整填写，做到记录完整无缺，无重复和遗漏；“准确”指填写的项目准确无误，界限明确，不造成混淆，资料真实可靠；“及时”指资料的时间性，要求按规定时间完成调查登记或填报工作，不能任意拖延时间。

（2）资料有足够的数量。应根据研究目的、资料性质、调查或实验条件等因素，决定资料的数量多少。若要全面摸清情况则需要进行普查，如人口普查、疾病普查等；若调查发病率或死亡率较低的疾病，也要调查较多的数量；当调查对象的个体变异较大时，则应增加调查数量。一般来说，数量多一些较好，但也并非愈多愈好，因为调查数量太多，势必耗费较多的人力、物力和时间。

（3）资料的代表性及可比性。“代表性”指在抽样研究中样本对总体的代表性。只有遵守抽样的随机化原则（即总体中每个对象都有同等的机会被抽取），才能避免主观和其他偏性。

随机抽样方法

（1）单纯随机抽样。在总体中以完全随机的方法抽取一部分个体组成样本的抽样方法。有抽签、摸球、掷币或使用随机数字表等方法。适用于样本含量小的资料。

（2）系统抽样。又称等距抽样或机械抽样，指随机地在所要抽样的名单中每间隔若干个个体抽取一个个体的抽样方法。

（3）分层抽样。按照与研究目的明显有关的因素（或某种特征），将总体分为若干类型或区域，统计上叫“层”，然后从每一层内按比例抽取一定数量的观察单位，将各层的观察单位合起来组成样本。

（4）整群抽样。首先将总体按照某种与研究目的无关的分布特征（如地区范围、不同的团体、病历、格子等）划分为若干个“群”组，每个群包括若干观察单位；然后根据需要随机抽取其中部分“群”，并调查被抽中的各“群”中的全部观察单位。

接链

“可比性”指在进行统计比较时，对比的各组之间，除观察问题或试验因素不同外，其他一切条件都要求尽量一致。例如，临床疗效观察

时，要求病情相同、诊断一致，年龄、性别及护理条件一致等。只有具备了齐同对比条件，才能清楚地说明被观察的因素所起的作用。

三、整理资料

整理资料是把搜集到的原始资料，有目的、有计划地进行科学地加工（如分组或汇总等），使其系统化、条理化，以便进行下一步的统计分析。

（一）原始资料的检查与核对

对于原始统计资料，在进行整理之前必须进行一次系统而认真的检查，以保证资料的可靠性、完整性和正确性。检查核对各个项目是否正确，有无错误、矛盾、重复、遗漏等。一经发现问题，必须及时加以修正、补充。如有不能改正的，则应作重新调查或予以剔除，以免影响统计分析质量。

（二）资料的分组

分组是根据资料的性质或数量特征，把资料进行分组整理，以反映事物的特点。医学科研资料常用的分组方法有下列两类：

（1）质量分组。即按事物的性质或类型分组，也就是以事物的质量标志进行分组。这种方法多适用于分类变量资料及等级资料。例如，病人按性别、职业等分组；疗效按治愈、好转、无效等分组。

（2）数量分组。即按变量值的大小来分组。这种方法多适用于数值变量资料的分组。数量分组的多少决定于资料的性质、数据的多少及分析的目的，以能说明资料的规律性为准。例如，按血压的高低，红细胞数、血红蛋白含量的多少和年龄的大小等分组。

四、分析资料

分析资料，包括统计描述、统计推断，以及解释、分析统计结果。统计描述指用一些统计指标、统计图表等来描述数据的分布特征、变化趋势等。统计推断指用调查、实验取得的样本信息估计总体特征，并对样本统计指标选择适宜的假设检验方法进行检验。最后再根据专业知识解释分析结果，阐明事物的内在联系和规律。具体的统计分析方法，将在以后章节中详细介绍。

案例

1. 某医师研究中药治疗感冒的疗效，在进行简单的设计后，随机抽取60例感冒患者作为研究对象，经用药观察，治愈52人，他认为治愈率为86.7%，该药治疗感冒有效。

思考题 （1）该医师下的结论是否正确？

（2）该医师的试验设计是否合理？

分析 （1）该医师认为该中药治疗感冒有效的结论欠妥。因为即使是不用药的情况下，自身免疫力也是感冒痊愈的一个因素。

（2）该医师的试验设计不合理，应该加设对照组，以消除自身免疫力的影响。

2. 某医师研究中药治疗感冒的疗效，随机抽取120例感冒患者作为研究对象，用随机方法将研究对象分为试验组和对照组（每组各60例），试验组用中药治疗，对照组给予安慰剂，经用药观察，试验组治愈52人，治愈率为86.7%；对照组治愈46人，治愈率为76.7%，该医师认为中药治疗感冒有效。

思考题 该医师的试验设计是否合理？结论是否可靠？

分析 该医师同时设立试验组和对照组，设计较合理。但不能直接凭借治愈率数值大小下结论，因为采取的是抽样研究，不可避免存在抽样误差，要想下正确客观的结论，还需进一步进行统计推断。

医学统计学是认识医学现象数量特征的重要工具，是运用概率论与数理统计的基本原理与方法，进行医学科研设计和资料的搜集、整理、分析与推断的过程，医学统计可透过偶然现象来探测其规律性。因此，医学统计方法已成为医学科学研究的重要前提和手段。

医学统计学的基本概念有总体与样本、同质与变异、抽样误差、概率、小概率事件等。

医学统计工作的步骤包括统计设计、搜集资料、整理资料和分析资料。

小结

目标检测

一、名词解释

总体	样本	变异	抽样误差
随机事件	概率	参数	统计量

二、问答题

1. 概率的基本性质有哪些？
2. 统计资料的类型有哪些？
3. 统计工作的基本步骤是什么？

笔记栏

第2章 数值变量资料的统计描述

学习目标

1. 说出频数表的编制方法，频数分布的两种趋势

2. 学会数值变量资料常用指标（平均水平、离散程度）的计算方法

3. 描述正态分布的基本特征与内容，了解医学参考值范围的估计方法

第1节 数值变量资料的频数表

一、频数表的概念

所谓频数就是观察值的个数。频数分布（frequency distribution）就是观察值在其所取值的范围内，于各组段中分布的情况。所谓频数表（frequency table）指一种统计表同时列出观察值的可能取值及其出现频数。具体做法是，先根据观察值数量大小进行分组，然后求出每组中观察值出现的次数。由于这种资料的表达方式较完整地体现了观察值的分布规律，所以叫作频数分布表。简称频数表。

二、频数表的编制

现举例说明频数表的编制方法。

例2-1 某年某市用随机抽样方法检查了120名12岁健康男孩的身高，检查结果如表2-1，试编制频数表。

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

注：* 最小值；** 最大值

用手工整理资料编制频数表时，通常先设计并编制划记表（表2-2），即先将选定的分组列好，每个组段的起点称下限，终点称上限（上限一般不写出）。然后在原始数据中逐项观察，观察到的数据应当归入哪一组，就在划记表的相应组段内划一道，划满五道成一个“正”字，将全部数据划记完毕后计算各组中“正”字的笔画数目，即得到各组的频数（表2-3）。

具体步骤如下：

（1）计算全距。从表2-1的120个数据中，找出最大值160.9和最小值125.9，两者之差即为全距（range），用R表示。即：

$$R = 160.9 - 125.9 = 35.0(\text{cm})$$

笔记栏

表 2-2　120 名 12 岁健康男孩身高(cm)资料的划记设计

组段(1)	划记(2)	频数/f(3)
125 ~		
129 ~		
133 ~		
137 ~		
141 ~		
145 ~		
149 ~		
153 ~		
157 ~ 161		
合计		

表 2-3　120 名 12 岁健康男孩身高(cm)资料的频数分布

组段(1)	划记(2)	频数/f(3)
125 ~	一	1
129 ~	丅	4
133 ~	正丅	9
137 ~	正正正正正丅	28
141 ~	正正正正正正正	35
145 ~	正正正正正丅	27
149 ~	正正一	11
153 ~	丅	4
157 ~ 161	一	1
合计		120

(2) 确定组段数、组距。根据全距(R)的大小、观察值个数(n),决定组段数(k)。全距大,观察值个数多,则组段数可适当增加。一般取 8 ~ 15 个组段为宜。组段数过多,编制过程和计算较繁,组段数过少,计算误差较大。

根据组段数和全距,决定组距。组距为相邻两组段最小值之差,用 i 表示。组距(i) = 全距(R)/组段数(k)。本例全距为 35cm,组段数拟定为 10 个,则组距为

$$i=\frac{R}{10}=\frac{35}{10}=3.5(\text{cm})$$

一般取靠近的整数作为组距,本例取$i=4$cm。

(3) 划分组段。每个组段应有一个起始值作为组下限(lower limit)和一个终止值作为组上限(higher limit);第一组段应包括最小值,最后组段应包括最大值。为了避免两组段界限互相包含,组段常用各组段的下限及波纹(~)表示。例如,第一组段 125 ~,第二组段 129 ~,第三组段 133 ~,…,第九组段 157 ~ 161。例如,遇到 129,则应划在第二组段中。

(4) 设计划记表。按照已确定的分组数和组距,设计出如表 2-2 形式的划记表。在设计划记表时,还应考虑项目之间的关系,以适合统计分析的要求。

(5) 归纳计数。绘制划记表后,按照不同组段分别将原始资料进行归纳计数,可划"正"字计数。表 2-3 即为上述 120 名 12 岁健康男孩身高(cm)的划记整理结果,即频数表。

如应用电子计算机,可将原始记录表中各项目编成数码后输入计算机,由计算机按事先编好的程序进行整理汇总及统计分析。

频数表的用途

(1) 揭示数值变量资料的分布特征。

(2) 描述数值变量资料分布的集中趋势和离散趋势。

(3) 便于发现某些特大或特小的可疑值。

第2节　集中趋势的指标

从频数表中,可以大致看出频数分布的规律。也可以大致了解频数分布的特征——集中趋势和离散程度,但这种认识是粗略的。如欲准确掌握频数分布的特征,就应作频数分布的定量描述,即计算集中趋势指标和离散程度指标。

集中趋势指标又称平均指标。它反映了观察值的集中位置或平均水平。也可以说,是观察值的典型水平或代表值。集中趋势指标的使用场合不同,计算方法也不一样。常用的集中趋势指标有算术均数(均数)、几何均数、中位数和百分位数等。

一、算术均数

算术均数(arithmetic mean)简称均数(mean),适用于观察值呈正态分布或对称分布的数值变量资料。总体均数用希腊字母 μ 表示,样本均

数用 $\bar{x}$ 表示。常用的计算方法：

（一）直接法

当观察值个数不多时，可直接将各观察值相加后除以观察值的个数，即直接法（direct method）。公式为

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n} = \frac{\sum x}{n} \quad (2.1)$$

式中，$\bar{x}$ 为样本均数，$x_1, x_2, \cdots, x_n$ 为各观察值，$\sum$ 为求和符号，n 为观察值个数。

例 2-2 测定五名健康人第一小时末血沉，分别是6,3,2,9,10(mm)，求均数。

$$\bar{x} = \frac{6+3+2+9+10}{5} = \frac{30}{5} = 6(\text{mm})$$

（二）加权法

当观察值个数较多或观察值为频数表资料时，宜用加权法（weighting method）计算均数。公式为

$$\bar{x} = \frac{f_1x_1 + f_2x_2 + \cdots + f_kx_k}{f_1 + f_2 + \cdots + f_k} = \frac{\sum fx}{\sum f} \quad (2.2)$$

式中，$f_1, f_2, \cdots, f_k$，分别为第一组段至第 k 组段的频数；$x_1, x_2, \cdots, x_k$，分别为第一组段至第 k 组段的组中值；$\sum fx$ 为各组段内组中值与频数乘积的总和；$\sum f = n$，为总频数。

例 2-3 计算表 2-1，某年某市 120 名 12 岁健康男孩身高(cm)的平均数。

首先编制频数表，具体方法前面已述。本例全距为 35cm，组距为 4cm，划记结果（即频数分布）见表 2-3。从表 2-3 看出身高在“125 ~”组段内有 1 人，在“129 ~”组段内有 4 人，…；同一组段内每个人身高是不相等的，可取组中值代表该组段每个人的身高。组中值 =（本组段下限值 + 下一组段下限值）/2，所以第一组段的组中值为（125 + 129）/2 = 127，第二组段的组中值为（129 + 133）/2 = 131，…余类推，见表 2-4 中的第(2)列。

表 2-4 中各组段内第(2)列组中值 x 与第(3)列频数 f 的乘积为第(4)列 fx，将第(4)列各组段的 fx 相加得 $\sum fx$。再将此值除以总频数 $\sum f$ 即得 120 名 12 岁健康男孩的平均身高。本例 $\sum fx = 17\,172$，$\sum f = 120$，将其代入公式(1.2)，得平均数为

$$\bar{x} = \frac{\sum fx}{\sum f} = \frac{17\,172}{120} = 143.10(\text{cm})$$

因为各组段频数起到了“权数”的作用，它“权衡”了各组中值由于频数不同对均数的影响。所以这种计算均数的方法，称为加权法。

表 2-4 120 名 12 岁健康男孩身高(cm)均数的加权法计算

组段 (1)	组中值/x (2)	频数/f (3)	fx (4) = (2) × (3)
125 ~	127	1	127
129 ~	131	4	524
133 ~	135	9	1215
137 ~	139	28	3892
141 ~	143	35	5005
145 ~	147	27	3969
149 ~	151	11	1661
153 ~	155	4	620
157 ~ 161	159	1	159
合计	-	120($\sum f$)	17 172($\sum fx$)

（三）简捷法

当观察值个数较多，同时数值又较大时，可在加权法基础上，进一步简化为简捷法计算均数：

$$\bar{x} = x_0 + \frac{\sum fd}{\sum f} \cdot (i) \quad (2.3)$$

式中，x_0 为假定均数；d 为缩减值；f 为频数；i 为组距。

例 2-4 以简捷法计算例 2-3 的平均数。

计算步骤如下：

（1）列计算表 2-5：表 2-5 中第（1），(2)，(3)列同表 2-4 中的第(1)，(2)，(3) 列。

（2）选“假定均数”（以 x_0 表示）。一般选频数较多，并且位置比较居中的组中值为假定均数(assumed mean)，本例选 143 为假定均数。

（3）求缩减值 d。将各组段的组中值减去假定均数 143，再除以组距 4，求得缩减值 d，见表 2-5 第(4)列。例如，第一组段的缩减值 d_1 为：$d_1 = (127 - 143)/4 = -4$，第二组段的缩减值 d_2 为：$d_2 = (131 - 143)/4 = -3$，…余类推。由于组距相等，d 值是很有规律的。

笔记栏

表 2-5　120 名 12 岁健康男孩身高(cm)均数的简捷法计算

组段 (1)	组中值/x (2)	频数/f (3)	$d=(x-x_0)/i$ (4)	fd (5)=(3)·(4)
125 ~	127	1	-4	-4
129 ~	131	4	-3	-12
133 ~	135	9	-2	-18
137 ~	139	28	-1	-28
141 ~	143(x_0)	35	0	0
145 ~	147	27	1	27
149 ~	151	11	2	22
153 ~	155	4	3	12
157 ~ 161	159	1	4	4
合计	-	120($\sum f$)	-	3($\sum fd$)

假定均数所在组段的缩减值 $d=0$，组中值小于假定均数各组段的 d 值依次为 -1，-2，-3，…，组中值大于假定均数各组段的 d 值依次为 1，2，3，…，所以 d 值可直接写出，不必通过计算。d 是简化后的组中值，故称为缩减值。

(4) 求 $\sum fd$。各组段内第(3)列 f 与第(4)列 d 的乘积为第(5)列的 fd，其合计值即 $\sum fd$，本例 $\sum fd=3$。

(5) 求均数。将表 2-5 有关数值代入公式(2.3)，得：

$$\bar{x}=x_0+\frac{\sum fd}{\sum f}\cdot(i)=143+\frac{3}{120}\times 4$$

$$=143.10(\text{cm})$$

用简捷法算得的均数与加权法结果完全一致，而计算特别简便。

二、几何均数

几何均数(geometric mean)即几何平均数，用 G 表示。适用于观察值呈对数正态分布或观察值为等比数列(如血清抗体滴度)的资料。常用的计算方法有如下几种。

(一) 直接法

当观察值个数不多时，直接将 n 个观察值 $x_1,x_2,\cdots,x_n$ 的乘积开 n 次方，写出公式为

$$G=\sqrt[n]{x_1x_2\cdots x_n} \qquad (2.4)$$

它的对数形式，即计算公式为

$$G=\lg^{-1}\left(\frac{\lg x_1+\lg x_2+\cdots+\lg x_n}{n}\right)$$

即

$$G=\lg^{-1}\left(\frac{\sum \lg x}{n}\right) \qquad (2.5)$$

式中，$\lg^{-1}$ 为求反对数的符号，$\sum \lg x$ 是各观察值的对数值之和，n 是总频数。

例 2-5　5 人的血清滴度分别为 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，代入公式(1.4)，得：

$$G=\sqrt[5]{2\times4\times8\times16\times32}=8$$

或代入公式(1.5)，得：

$$G=\lg^{-1}\left(\frac{\lg2+\lg4+\lg8+\lg16+\lg32}{5}\right)$$

$$=\lg^{-1}(0.903)=8$$

故平均滴度为 1:8。

(二) 加权法

当观察值个数较多或观察值为频数表资料，可用加权法求几何均数，其计算公式：

$$G=\lg^{-1}\left(\frac{\sum f\lg x}{\sum f}\right) \qquad (2.6)$$

式中，$\sum f\lg x$ 为各观察值的对数与相应频数乘积之总和，$\sum f$ 为频数的总和。

例 2-6　某年某市 100 名儿童接种某种疫苗后，测定抗体滴度的资料如表 2-6 第(1)，(2)列所示，求该疫苗的抗体平均滴度。

表 2-6　抗体平均滴度的加权法计算

抗体滴度 (1)	人数/f (2)	滴度倒数/x (3)	$\lg x$ (4)	$f\lg x$ (5)=(2)×(4)
1:2	2	2	0.301 0	0.602 0
1:4	11	4	0.602 1	6.623 1
1:8	18	8	0.903 1	16.255 8
1:16	36	16	1.204 1	43.347 6
1:32	22	32	1.505 1	33.112 2
1:64	8	64	1.806 2	14.449 6
1:128	3	128	2.107 2	6.321 6
合计	100($\sum f$)	-	-	120.711 9($\sum f\lg x$)

引自：师明中．1992．几何均数简捷计算法的两个公式．数理医药学杂志，5(4)：69 ~ 70

将表 2-6 有关数值代入公式(2.6)，得：

$$G=\lg^{-1}\left(\frac{120.711\,9}{100}\right)=\lg^{-1}(1.207\,1)=16.11$$