JMWSTJX

简明卫生统计学

周明河主编

北京大学出版社

简明卫生统计学

主　编

周明河

编　委

李寿鹤　　赵敬忠

吴永贤　　翟淑妙

张二振　　江建明

一九九二年十月十日

北京大学出版社

内　容　简　介

本书是在多年卫生统计在职培训讲稿的基础之上加以修订和扩充而写成的。着意于从实用出发，解决医学领域里的卫生统计问题。每介绍一种方法举一个例题，不作过多阐述。文字简炼概括，但内容比较广泛。为了便于在职培训之用，将基本内容分为预防医学、卫生检验、临床医学三部分。短期培训时可只讲与受训人员专业对应的部分，较长时间的系统培训时可将全书讲完，因此本书可作为各类长短期学习班、专业证书班、成人专科班的卫生统计教材，也可作为各类医务人员在科研工作和日常工作中自学和查询的工具书。

简明卫生统计学

周明河
主　编

*

北京大学出版社出版
（北京大学校内）
北京昌平百善印刷厂电脑排版部排版
北京大学印刷厂印刷
新华书店北京发行所发行　　各地新华书店经售
787×1092毫米　16开本　12 印张　293千字
1991 年 6 月 第1版　1991 年 6 月 第1次印刷
印数：0001－2000 册
ISBN 7－301－01467－8 / R·6
定价：6 元

目　　录

1. 概 述

1.1 卫生统计学的意义和内容

卫生统计也称生物统计，主要用于分析人和生物体的变异规律．卫生统计是数理统计在医疗卫生方面的应用，是数理统计的一个分支。数理统计是以样本为根据，运用数学模型来推断总体的一门科学，运用数理统计，可以研究大量的自然现象和社会现象的规律性。数理统计是一门应用数学，它的理论基础是概率论。概率论是一古老的科学，其原理是18世纪从赌场上发展起来的，当时欧州的赌博是从一付扑克牌中随机地抽样。这就需要科学家解决，如果每次抽10张牌，抽到1、2、3…… 10张黑牌的概率有多大？1713年贝努里通过大量的实践和推导，提供了简单的解决办法，这就是贝努里分布(二项分布)，成为概率论的奠基人。以后贝叶斯，费舍等人又为概率论的发展作出了卓绝的贡献，使之成为一门非常有用的学科。当然，我国古代的孙膑赛马已经具有了概率论的萌芽。

卫生统计的重要性已越来越被广大医疗卫生人员所重视，尤其是恢复职称晋升以后。许多人在工作实践中积累了大量的资料、数据，如何整理分析这些数据，如何从中提炼出规律性的东西和新见解，如何写出一篇象样的论文，都需要卫生统计知识。还有许多人想搞调查研究，从何处下手？如何搞出一份科学合理的设计方案，以保证结果的准确可靠和提高工作效率，也离不开卫生统计。要搞防治效果评价，病因探讨等更离不开卫生统计。甚至看一篇文献资料，没有统计知识也会遇到困难。卫生统计是各类医务人员，尤其是业务骨干的一门不可或少的技术工具。希望大家努力把它学好。

卫生统计的研究对象是有变异的东西，例如同为健康人，即使是同年龄、同性别，他们的身长、体重、血压、脉搏、体温、红细胞、白细胞等数值会各不相同。同为一种病人，病情轻重可以有所不同。对病情相同的人，用同一种疗法进行治疗，疗效也不一样。这些变异规律只有在大数量的基础上才能表现出来，因此，统计的研究对象要有一定数量。也可以说卫生统计的研究对象是有变异的群体。

卫生统计学的内容主要有三个大的方面：

1.1.1 定水平 因为卫生统计的研究内容是有变异的群体，因此需要有一些描述群体水平的典型指标，对计量资料，要计算集中趋势指标如算术均数、几何均数、中位数。还要计算离散趋势指标如全距、标准差、变异系数等。对计数资料要计算相对数如频数指标、构成指标、对比指标等。这些定水平的内容，是各类业务工作中时时处处都用得到的计算方法。

1.1.2 作比较 即作显著性的检验。在日常业务工作中和科研活动中，经常要比较两种事物、两种效果谁好谁坏、谁高谁低。它们之间的差别是固有的，还是由抽样误差造成，这都要通过显著性检验作出回答。作比较是卫生统计中所占比重最大的一部分内容。其方法也非常繁多，不下几十种。常用的方法有 u 检验、t 检验、χ^2 检验。方差分析，二项分布，泊松分布及一系列非参数检验方法，是卫生统计学习的重点内容。

1.1.3 找关系 自然界中的许多事物之间存在着相互依存、相互制约的关系。尤其人的生老病死、健康长寿和很多自然因素、社会因素息息相关。分析它们之间的关系，

能够使人们扬起有利的因素、拟制不利的因素，是增进健康和发展技术的重要前提。尤其对病因不明疾病的研究和探索，更离不开相关分析。分析事物之间有无关系，关系的密切程度和方向的过程叫相关分析；对有关的事物，将其关系用函数式表达出来，从而由一种事物（自变量）去推测另一种事物（因变量）的方法叫回归分析。回归分析可用于疾病的预报和预测。找关系的内容包括直线回归、曲线回归和多因素回归，其计算方法比较复杂，尤其是多因素回归，必需借助计算机程序进行运算。但多元回归的重要性已为人们所认识，新方法新程序正在迅速开发。

1.2 卫生统计工作的步骤和方法

卫生统计工作和一般科研工作相似，可分为四个步骤：设计、搜集资料、整理资料和分析资料。这四个步骤是相互联系不可分割的，任何步骤的缺陷，都会影响统计分析的结果。

1.2.1 设计 即有一个全面的工作计划。任何统计工作，第一步都需要有一个严密的设计，这是最关键的一步。首先要明确研究目的，要对被研究的事物有一定的了解，可根据以往的经验和参考文献，必要时通过预调查，掌握较多的信息。根据研究目的，明确要搜集哪些资料，计算哪些指标。设计完善可行的调查表。还包括对调查对象的选择和样本数量的估算。如何保证资料的准确可靠也是需要考虑的问题，如对照组的设置、干扰因素的消除等。凡此种种，都要经过周密的考虑，作出明确的回答。要结合自身的人力、物力，作出切合实际的科学详细的安排，以期用最少的人力物力达到最好的效果。

1.2.2 搜集资料 搜集资料是统计工作的基础，它的任务是按照设计要求，及时取得完善可能的原始资料。没有完整准确的第一手资料就不可能取得预期的研究结果，甚至作出错误的结论。统计资料的来源可概括为经常性资料和一时性资料两大类。

1.2.2.1 经常性资料 这些资料主要包括医疗卫生工作记录和报告卡片，如门诊住院病历，检查报告单，健康卡片、传染病报告卡片等，这些资料常会出现漏填，重复和项目填写不清等情况。必须使医务人员认识到原始记录的重要性，它是从事医学科学研究和评价医疗预防工作质量的重要资料，要准确、及时、认真填写、妥善保管，以备随时查用。还有一类经常性资料是统计报表，如疫情报表、食物中毒报表、农药中毒报表、卫生监测报表，医院工作报表等，这是根据国家规定的报告制度填报的，它全面经常地提供居民健康状况和医疗卫生工作的主要数字，也给科学研究工作提供了基础资料。

经常性资料的搜集省时、省力。可以用很少的经费，甚至不用经费就能获得重要的资料和结论。但这些资料不是经过专门设计获得的，经常不大合用，资料的准确性也较差。

1.2.2.2 一时性资料 是根据某项工作和研究的需要通过临时设计的调查研究和实验研究方案而取得的资料。这些资料不能从现成的统计报表或医疗单位的原始记录获得，必须进行专门调查或实验才能取得所需要的资料，如肿瘤的病因研究，药物的疗效观察，儿童生长发育研究，分析方法的条件试验等。这些资料比较准确、合用，但要特别注意对照组的设置及混杂因素的消除。还要注意测量中的质量控制及填表中的遗漏和差错。收集一时性资料要设计一套合理的调查表格。

1.2.3 整理资料 上面所搜集到的大量

原始资料，是分散的资料，要了解事物的特征及规律，必须对这些原始资料进行科学的分组归纳，使资料系统化。资料的整理可分为检查、分组，拟整理表和归纳四个步骤。

1.2.3.1 资料的检查 搜集到的资料难免有误，若不进行事先检查核对，就会以误就误，使分析发生困难，甚至误入歧途。检查的内容主要是资料的完整性和正确性，如填表有无遗漏、空项、重复，各项之间有无矛盾，数字有无不合理之处等。发现错漏之处，要及时订正。在专题调查时，要边调查边检查核对，以免时过境迁，使检查订正工作发生困难。

1.2.3.2 分组 资料的分组是将性质相同的事物归纳到一起、以反映事物的内在规律性。分组有质量分组和数量分组。整理资料时要按各种标识尽可能多分组，以免漏掉可能的信息。

1.2.3.3 拟整理表 整理表是用于原始资料归纳的一种过渡性表格，格式要求不严。根据分组情况而设计，主要作用是把各组的频数归纳进来，要有合计。通常把关系密切的项目放在一张表内，可使相互关系表达出来。表的大小，繁简都无一定要求，使划记和计算均数、率、相关系数等比较方便即可。注意不要把整理表放入论文中。

1.2.3.4 归纳 归纳汇总是将原始数据按照作好的整理表分到各组中去。最常用的方法是划记法，方便简便，但容易出错。还有用记录卡片进行分组归纳的。把原始数据归纳到各组是很重要的一步，要随时进行核对以免错漏，注意合计栏的数字是否等于总样本数量。

1.2.4 分析资料 分析资料是揭示事物内在规律的关键步骤，其任务是按照要求，将归纳好的资料进行统计运算，如计算相对数、计算集中趋势指标和离散趋势指标，进行显著性检验和相关回归分析等。为了使结果更可靠和不漏掉有用的信息，经常需要选用多种统计方法，如显著性检验中，可用多种方法加以验证，回归分析中，可选用多种回归模式。近年来还提倡对统计资料的再开发再利用，即对已经用过的资料，再经过进一步分析，往往可以提炼出更多的信息。分析资料要从多方面进行考虑和探索，深挖事物的内在规律，草率和浮浅往往造成很大浪费。通常，分析过程完成，事物的规律和结论就基本明朗。为了更好的进行表述，还需要绘制统计图表。

1.3 卫生统计学的几个基本概念

1.3.1 计量资料和计数资料 卫生统计资料通常分为计量资料和计数资料两大类，还有介于其中的等级资料。因为不同类型的资料要采用不同的分析方法，所以首先要分清资料的类型。

计量资料是用度量衡测量出的资料，度、量、衡分别是测量长度、体积和重量的，随着科学的发展，测量的方法越来越多，范围越来越广。但它们的共同特征是有计量单位，如厘米、千克、毫升、毫克%、毫克/升等，而且能按大小进行分组。通常将一个计量数据叫一个变量值，一群计量数据叫一组变量值。计量资料能够计算平均数。

计数资料是清点个数数出来的资料，只有个数，没有其它计量单位。它是按属性进行分组的，如男女、阴性阳性、病人健康人各多少，不同职业、不同住址的人各多少等。没有大小的概念。这些资料能够计算率。

等级资料既有计数资料的属性，也有量的概念，如尿糖和尿蛋白检验的－、

+、++、+++、++++,可分出阴性阳性,但阳性中又有量的不同.还有象疗效等级,无效、好转、显效、疮瘀等也属於等级资料。

1.3.2 总体和样本 总体是根据研究目的确定的同质的研究对象的全体。医学研究中,很多是无限总体,要直接研究总体情况是不可能的.如研究正常成人的脉搏情况,则总体包括所有的正常成年人,这是个无限的总体。即使对有限总体,由于包含的观察单位过多,直接研究总体耗费人力、物力,也往往是不可能和不必要的。如成批罐头食品的生产,不能把所有罐头都拿出来作化验。

在实际工作中,经常是从研究总体中随机抽取部分观察单位,这些观察单位称为样本,用样本信息来推测总体的特证。对样本的要求一是样本要具有代表性,成为总体具体细微的缩影,而不是总体中的一个特殊部分。样本要具有代表性,必需遵循随机抽样的原则。二是样本要有足够的数量,若样本数量过少,抽样误差必增大,样本就不能很好地代表总体。如何确定样本数量,有一套估算方法。样本数量太大,也不必要,徒增加研究工作的困难。如研究某地区正常成年人的血压水平,可通过各种随机抽样方法,随机抽取200人,作为样本,分别测量其血压值,计算样本均数,用以推测本地区正常成年人血压的总体均数。这种方法称为抽样研究,是颇为常用并且极其重要的研究方法。

1.3.3 变异和抽样误差 世间的一切事物,千差万别、各不相同,事物间的差异称为变异。变异更是生物的重要特征,全世界50多亿人,各有各的面貌,各有各的性格,没有两个完全相同的人。同年龄、同性别的儿童(统计上称为同质观察单位),虽然身长、体重都与年俱增,但胖瘦、高矮各不相同,一些微细指标更是千差万别。同样一组病人,给予相同的治疗,后果也各有参差。实验动物也是如此。生物间的变异是多种因素的综合影响造成的,其中许多因素是未知的、难于控制的。卫生统计学所研究的正是这些有变异的群体,通过对个体变异的研究、透过偶然现象,揭示事物的本质规律。

变异是抽样误差的根源.没有变异也就没有抽样误差。正因为变异是无处不在的,因而抽样误差是不可避免的。在抽样研究中,样本指标与总体指标的相差称抽样误差。例如,某市2万名7岁男孩,假设平均身高为120厘米(总体均数),每次抽样取100名儿童作调查,由于个体变异,每次的样本均数不可能恰好是120厘米,这就产生了抽样误差。抽样误差随变异程度和样本数量而变化,研究对象的变异程度小,样本数量大,可使抽样误差减少。

1.3.4 显著性和显著性检验 两个样本或多个样本进行比较时,都要判断其均数或率有无显著性。所谓显著性,就是判断其差别是否由偶然的抽样误差所致,亦即判断样本所代表的总体之间有无本质差别.若样本间的差异可能是由抽样误差所致,我们就说这种差异无显著性,则不予承认.若样本间的差异不大可能是由抽样误差所致,我们就说这种差别有显著性,而予承认,可进一步分析造成差异的原因等。统计学是用概率说话的,只能推断事物发生的可能性大小,而不能说绝对是或绝对不是。这种可能性称为概率,可通过一系列的方法加以推导。计算这种概率的过程称显著性检验。运算结果若两样本之差由抽样误差所致的概率(可能性)大于5%(即$P>0.05$),即认为这种可能性已经比较大了,其差别可能由抽样误差所致,则判断为差异无显著性。若算出的概率比较小($P \leqslant 0.05$),则认为不大可能由抽样误差所致,判断为差异有显著性。若概率更小($P \leqslant 0.01$)则认为非常不可能由抽样误差所致,判断为差异有非常显著性。显著性检验的方法很多,各种方法都有其适用条件,也有一个

问题可用多种方法运算的。各种方法的计算结果都是得出概率的大小。

显著性是个非常重要的概念，对初学者经常是一阵清楚，一阵糊涂，需要脑子多转几个弯儿。为了帮助大家理解，再举一个浅显的例子。现有两口袋枣儿，可当作两个总体，甲袋好枣儿占50%，乙袋好枣儿占90%。现从总体甲抽取两个样本，每个样本100枚，好枣儿率分别为45%（P_1）和58%（P_2）；再从总体乙抽取两个样本，好枣儿率分别为95%（P_3）和87%（P_4）。这4个率互有差异，但差异来源不同，P_1和P_2，P_3和P_4分别来自同一总体，它们所代表的是同一个总体，它们之间的差异是由抽样误差所致，因此没有显著性。P_1和P_3，P_1和P_4，P_2和P_3，P_2和P_4进行比较时，它们所代表的是性质不同的两个总体，它们之间的差异不是由抽样误差所致，而是其间存在着本质差异，因而认为其差有显著性。

1.4 概率论的基本知识

1.4.1 概率的概念 概率的简单含义可理解为一个事件发生的可能性的大小。在一定条件下进行试验时，所发生的现象叫事件。如果在每次试验中，某事件一定发生，则这一事件叫必然事件，如在一个大气压下，将纯水加热到100℃，则水必定沸腾。水沸腾是事件，每次试验必定发生。必然事件的概率为1。相反地，如果某事件一定不发生，则称不可能事件。如研究让人或动物长生不老，每次试验必定失败。不可能事件的概率等于0。但医疗卫生实践中大多数是随机事件，即在试验中可能发生，也可能不发生的事件。如每个人可能得病，也可能不得病。每个病人，可能好，也可能死。每次试验，可能成功，也可能失败等。随机事件的概率介于0和1之间。

直接估计某一随机事件的概率是非常困难的，甚至是不可能的，仅在比较简单的情况下才可以直接计算随机事件的概率。通常把在很多次试验中随机事件的频率当作概率的近似值。设随机事件A在n次试验中发生了m次，则比值m/n叫做随机事件A的频率。经验证明，当试验次数重复很多次时，随机事件A的频率具有一定的稳定性。就是说，在不同的试验序列中，当试验次数充分大时，随机事件A的频率常在一个确定的数字附近摆动。这个确定的数字正是随机事件A的概率。即随机事件发生的可能性可以用一个数来表示，这个刻画随机事件A在试验中发生的可能性程度的，小于1的正数叫做随机事件A的概率。随机事件的频率可以看作是它的概率的随机表现。

1.4.2 概率的古典定义 在概率论发展的初期是以等可能性的、互不相容的完备群为主要研究对象，只有在这种情况下才可以直接计算随机事件的概率，称古典概率。所谓等可能性是试验中每个基本事件发生的可能性在客观上是完全相等的，如抛硬币正面反面，生孩子的生男生女。所谓互不相容性是任何两个事件在试验中都不可能同时发生，如一次抛硬币，不是出现正面，就是出现反面，不能两面同时出现。如果试验时，若干个随机事件中至少有一个事件发生，则称它们构成完备群。具有这三种性质的事件叫做基本事件。如一次接生，基本事件为（男）、（女）两个。二次接生，按接生顺序，基本事件为（男、男）、（男、女）、（女、男）、（女、女）四个。三次接生，基本事件则有（男、男、男）、（男、男、女）、（男、女、男）、（男、女、女）、（女、男、男）、（女、男、女）、（女、女、男）、（女、女、女）八个。如果试验时某一基本

事件的发生导致随机事件A的发生，则称此基本事件是有利于随机事件A的。

概率的古典定义可叙述如下：设试验的一切可能结果可以表为由 N 个互不相容且等可能性的事件构成的完备群，而其中 M 个事件是有利于随机事件A的，则随机事件A的概率等于有利的基本事件数 M 与基本事件的总数 N 的比值：

$$P(A)=\frac{M}{N}$$

例如三次接生，设随机事件A为“恰有一人为女性”，则有利的基本事件为

（男、男、女）、（男、女、男）、（女、男、男）。

$M=3$，　$N=8$，　$P(A)=3/8=0.375$

还可计算稍繁杂一点的问题。

例1.1：袋内有8个球，其中5个白球3个红球。每次从中任取两球（取后再放回），问两球都是白球的概率 P(A) 是多少？

此例的基本事件总数

$$N=C_8^2=\frac{8!}{2!(8-2)!}=28$$

有利的基本事件数

$$M=C_5^2=\frac{5!}{2!(5-2)!}=10$$

所求的概率

$$P(A)=\frac{M}{N}=\frac{10}{28}=0.375$$

从概率的古典定义可引出概率论的两个基本定理，即概率加法定理与乘法定理。

【概率加法定理】二互不相容事件和的概率，等于这二事件概率的和，即

$$P(A+B)=P(A)+P(B)$$

这一定理也可推广为：有限个互不相容事件和的概率，等于这些事件概率的和。如果事件 A_1，A_2 ……，A_n 构成互不相容的完备群，则这些事件的概率和等于1。尤其仅由两个互不相容事件构成的完备群，称对立事件，对立事件的概率和等于1，即

$$P(A)+P(\bar{A})=1,$$

设

$$P(A)=P,\quad P(\bar{A})=q,\quad P+q=1,$$

这是计数资料运算常常用到的。

例1.2：某工厂在生产中，出现二级品的概率为3%，三级品的概率为1%，其余都是一级品。如二级品、三级品为次品，求出现次品的概率。

这两种次品分别为事件A和事件B，是互不相容事件，次品概率

$$P(A+B)=P(A)+P(B)=3\%+1\%=1/25$$

【概率乘法定理】二独立事件积的概率等于这二事件概率的乘积，即

$$P(AB)=P(A)P(B)$$

也可以把这一概念推广到多个事件中去。即有限个独立事件积的概率等于这些事件概率的积，即

$$P(A_1A_2\cdots A_n)=P(A_1)P(A_2)\cdots P(A_n)$$

所谓独立事件即两事件中任一事件的发生不影响另一事件的概率。

例1.3：一产品从毛坯到成品要经过三道工序，三道工序的次品率分别为1%、5%、10%，问总次品率有多少？

各道工序加工情况互不影响，所以事件 A_1、A_2、A_3 是独立的。合格率

$$P(A)=P(A_1)P(A_2)P(A_3)$$

其中　$P(A_1)=99\%$　　$P(A_2)=95\%$，

$P(A_3)=90\%$

不合格率 $P(\bar{A})=1-P(A)$

即 $P(\bar{A})=1-P(A_1)P(A_2)P(A_3)$

$=1-99\%\times 95\%\times 90\%=15.4\%$。

但在实际工作中不符合古典定义的情况很多，如疾病的治疗效果，血型的分布等不是等可能性的；传染病的互相传染，精神因素对发病的影响等不是独立性的；还有一些试验的基本事件是无限的。这些情况对概率的古典定义都是不适用的。

1.4.3 大数定律　大数定律是概率论的一个中心问题。它的一般意义是：在个别各因素具有个体偏差的总体中，一般的规律性表现在集团中，表现在大量的事实中。也就是

说，如果构成总体的各单位各自具有个体的偶然的差异，则在这样的总体中，总的规律是在大量现象中表现出来的。例如，我们说吸烟是肺癌的重要因素，多吸烟可引起肺癌，是从大量人群的观察中得出的规律，而就单个的个别人来说，不吸烟也可以得肺癌，吸烟也可以不得肺癌。但只要观察数量足够大，这些个体差异就会被消除而表现出事物的内在规律来。还有我们说适当的体育锻炼有助于增强儿童的生长发育，而就个别从不参加锻炼的人来说，也有生长发育很好的，增强表现在大量人群的平均数上。概率论中用来阐明大量随机现象平均结果的稳定性的一系列定理统称大数定律。如果努里定理、泊松定理、切贝雪夫定理等。大数定律是一种表现必然性与偶然性之间的辩证联系的规律，由于大数定律的作用，大量随机因素的总和作用，必然导致某种不依赖于个别随机事件的结果。其中切贝雪夫定律是大数定律的通式，含义为：当试验次数或样本容量（n）充分大时，样本均数或样本率与总体指标之间的差量将很小。在实践中，根据大数定律，在大量情况下取得的事件的样本均数或样本率，只能是该事件出现的概率的近似值，而观察的数量越大，算出的概率越正确。

1.5 学好卫生统计的诀窍

1.5.1 提高认识、树立学好统计的信心 学好卫生统计应以辩证唯物法为指导，以医学科学为基础。有了正确的观点与思想方法才能更好的发挥卫生统计作用。统计无用和统计万能的观点都是错误的。有些人认为凭常识判断，只凭试验结果的表面差别就能得出结论，不需要什么统计处理。这种没有考虑到抽样误差问题的判断往往是错误的。用错误的结论去推广、去指导实践就会失败或造成损失。通过统计处理再得出结论，失误的机会就会大大减少。但统计不是万能的，它只能帮助我们去分析和认识客观规律，它决不能改变事物的本来面目，把原不存在的规律创造出来。有些人甚至只凭主观愿望出发，滥用或乱用统计方法，弄来一些虚假数据去凑合预定的结论，这是非常错误的。统计学需要的是老实、客观和辩证的态度。

统计判断是概率性的，只说可能性大小，不下绝对性结论。说两种事物有差异，允许有5%以下说错的可能。没有显著性差异的事物，当增大样本数量后可能变得有显著性。当不同事物之间作比较时，有时有显著性好，有时无显著性好。有时统计上有显著性，在实际中并无意义。这些都需要我们客观地辩证地看问题，不要臆测，不能武断。

认识到统计学的重要性之后，还有一个如何学好的问题。因为学医的人大多不喜欢数学，不少人觉得卫生统计难学。统计虽属数学范畴，其实卫生统计所用的多是现成公式，我们的任务主要是会套公式，会分析结果，并不要过多地演算。只要有最基本的数学基础如加、减、乘、除、乘方、开方、求对数等就能学会，要树立一定能学会的信心。卫生统计就象一台由数学工作者造就的机器，我们的任务是学会开开关、操作机器、使用机器。至于这台机器的内部构造及维修，一般不用去管。

1.5.2 着重理解卫生统计的基本原理与基本概念至关重要 卫生统计学的公式很多，有的还很繁杂，要记住各种公式，通常不容易作到，也不必要，因为我们可以随时查到公式。关键是哪些资料用哪些公式，计算结果如何解释。这就需要掌握基本概念如计量资料、算术均数、几何均数、标准差、标

准误差、构成指标、频数指标、抽样误差、显著性检验、概率、相关、回归等一系列概念，才能正确地套公式，解释结果。概念清楚了可以长期不忘，公式记往后很快就忘。记住概念是学好卫生统计的重要诀窍。

1.5.3 多用、多练，熟能生巧，细致耐心，不怕麻烦 卫生统计是应用科学，应用科学就得多实践。每学过一个问题、一种方法后一定要亲自练一练、做一做，或结合工作中的实践资料做一做。而且做得次数越多越好，只听一遍课掌握不了方法。再者看业务杂志时，要留心统计处理，看自己能否理解、能否做得出来。一旦自己能解决一定的实际问题，学习的兴趣就会越来越大。

卫生统计毕竟还是一门数学，有较多的数学运算，其中有些还比较麻烦。这就需要细致、耐心，不怕麻烦，不能使计算出错。要静下心来、钻进去、有条不紊、一步一个脚印。天长日久，就会算出兴趣来。如若毛毛草草，定会错误百出，不能使其更好地为业务服务。

2. 卫生防疫常用卫生统计

2.1 集中趋势指标

集中趋势指标用以表示一组变量值的代表性和典型性水平，是对计量资料进行统计分析的基本指标。常用的指标有算术均数、几何均数和中位数三种。集中趋势指标能给人一清楚明确的概念，又便于进行相互比较。是卫生统计学中定水平的基本内容。

2.1.1 算术均数 算术均数也可径称均数。用于变量值呈正态分布或接近正态分布的资料，象生长发育指标、生理生化指标、常量检验指标等基本成正态分布者，算术均数的计算有直接法、加权法和简捷法。

2.1.1.1 直接法 即将全部变量直接相加，再被变量值的个数除。通常用于例数较少的不分组资料，若有计算器也可用于例数很大的资料。可用公式表示为：

$$\bar{X}=\frac{x_1+x_2+x_3+\cdots\cdots x_n}{n}=\frac{\sum x}{n}$$

式中：$\sum$为希腊字母 Sigma，为求和的符号，

$\bar{X}$ 为平均数，x为变量值，

n为变量值的个数。

例 2.1 12名健康成人血沉为：3、9、8、6、5、5、7、3、10、8、4毫米/小时，求平均血沉。

$$\bar{X}=\frac{3+9+8+6+5+5+7+3+10+8+4}{12}$$

$$=\frac{78}{12}=6.5\ (\text{毫米/小时})$$

2.1.1.2 加权法 权即法码、秤锤。当一组变量值相同数值的个数较多时，可将相同数值的个数（频数）乘以该变量值以代替相同变量值逐个相加。此法主要用于分组资料，以每个组段的组中值作为一组数的代表。计算公式可写作：

$$\bar{X}=\frac{\sum fx}{\sum f}$$

式中：f为相同变量值的个数，即每组的频数，x为每组的组中值。

例 2.2 求120名12岁男孩身高均数。

表 2－1 120名12岁男孩身高频数表(加权计算表)

分 组	组中值 x	频数 f	fx
	127	1	127
129－	131	4	624
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－	159	1	159

$\sum f=120$ $\sum fx=17172$

$$\bar{X}=\frac{17172}{120}=143.10\ (\text{cm})$$

分组频数表的编制方法是将一组变量值先找出最大值和最小值。两值相减为全距。通常将全距分为10组左右，即全距/10＝组距。组距要取比较整齐、简单的数值，如0.1,0.2,0.5,1,2,5,10,20等。这样即可定出每组的上、下限，如第1组，下限为125，上限为129。上、下限相加除2即为组中值。最后用划记法将每一变量值编到相应的组中即成频数表。

加权法计算时数值很大，容易出错，所以对大的分组资料多采用简捷法计算。

2.1.1.3 简捷法 编好频数表后，先设一假定均数，假定均数多选频数最多的组中值。然后计算其它各组和假定均数相差的组距数（化减值），其实化减值不用计算，对着假定均数划0，往上各组依次划 −1,−2,−3,……，往下各组依次划1,2,3,……。计算原理是：比假定均数多的组距数与比假定均数少的组距数相互补平后，余（或欠）下的组距数被总频数平均后，乘以组距，加（或减）在假定均数上即为实际均数。计算公式写作：

$$\bar{X} = \bar{X}_0 + \frac{\sum fx}{\sum f} i$$

式中：$\bar{X}_0$为假定均数，　x为化减值，
f为各组频数，　i为组距。

表2−2 120名12岁男孩身高简捷法计算表

分　组	组中值	频数f	化减值x	fx	
125−	127	1	−4	−4	
129−	131	4	−3	−12	
133−	135	9	−2	−18	−62
137−	139	28	−1	−28	
141−	143	35	0	0	
145−	147	27	1	27	
149−	151	11	2	22	
153−	155	4	3	12	+45
157−	159	1	4	4	

$\sum f = 120$　　$\sum fx = 3$

例2.3　仍求平均身高。

表2.2中将频数最多的组中值143作为假定均数，划化减值，各化减值乘以对应的频数（fx）。假设以143作为身高均数，则上边4组共短62个组距，下边4组共长65个组距，长短相补，还富余3个组距。将3个组距变成实际单位（厘米），平均到120人头上即为实际均数。代入公式：

$$\bar{X} = 143 + \frac{3}{120} \times 4 = 143.10 \text{ (cm)}$$

此法计算起来简便、快速，数据小，不易出错。

2.1.2 几何均数 几何均数也称对数均数，用于呈正偏态分布的计量资料。此类资料的各个变量值取对数后基本呈正态分布。常见的有变量值呈倍数关系的资料，如稀释度，抗体效价等。生物材料中微量元素含量，酶活性，食物中毒潜伏期等资料也多呈正偏态分布。几何均数的代号为G，常用的计算方法有直接法、加权法和编码法。

2.1.2.1 直接法 用于不分组的简单资料，先把每一变量值取对数，求和后，被变量值的个数除，最后求反对数。表达公式为：

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

式中：$\sum \lg x$ 为各变量值求对数后的和
$\lg^{-1}$ 为反对数符号。

例2.4 求5人的血清总体平均效价，其效价分别为1:10,1:100,1:1000,1:10000,1:100000。

代入公式：

$$G = \lg^{-1}\left(\frac{\lg 10 + \lg 100 + \lg 1000 + \lg 10000 + \lg 100000}{5}\right)$$

$$= \lg^{-1}(3) = 1000$$

即5人的平均效价为1:1000

2.1.2.2 加权法 用于分组资料，即有较多数值相同的变量值，计算步骤是先将各变量值求对数，再乘以所对应的频数，求和后被总频数除，求反对数。表达公式为：

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

例2.5 某疫苗接种后，求48人的血清平均凝集效价。

计算如下：

表 2－3　平均凝集效价计算表

凝集效价 x	人数 f	$\lg x$	$f\lg x$
20	6	1.3010	7.8060
40	9	1.6021	14.4189
80	21	1.9031	39.9651
160	7	2.2041	15.4287
320	5	2.5051	12.5255

$\sum f \lg x = 90.1442$

代入公式：

$$G = \lg^{-1}\left(\frac{90.1442}{48}\right) = \lg^{-1}1.8780$$

$$= 75.51$$

平均凝集效价为 75.51。

2.1.2.3　编码法　当各变量值都是 2 的整数倍时，用 2 的指数表示（编码），用加数法求出平均数。几何均数即 2 的多少次方。实际等于求以 2 为底的对数。此法运算更为简便，不用查对数表。仍以上例用编码法计算。

表 2－4　平均凝集效价编码法计算表

凝集效价	人数 f	10×2^{x}（编码）	fx
20	6	1	6
40	9	2	18
80	21	3	63
160	7	4	28
320	5	5	25

$\sum fx = 140$

$$\overline{X} = \frac{\sum fx}{\sum f} = \frac{140}{48} = 2.9167$$

$$G = 10\times 2^{2.9167} = 75.51$$

2.1.3 中位数　把一组变量值按大小顺序排列起来，中间位置所对应的数值称中位数，代号 M，即百分位数法的 50% 位数。中位数的应用范围比较广泛，可用于呈极度偏态分布的计量资料，如环境中的细菌数，（空气、土壤、食品、水等），体液中的细胞数等，还可用于分布状态不明或任意分布的资料。计算方法简单，可分简单资料和分组资料。

2.1.3.1 简单资料　把变量值按大小排队，找出中间位置 $\frac{n+1}{2}$，所对的数值即为中位数，几乎不用计算。一组变量值的个数为偶数，需将中间的两个数值相加被 2 除。

例 2.6　10名霉菌性脑膜炎病人脑脊液中细胞数如下，求中位数。

15,28,53,74,96,124,160,286,450,1098。中间位置为 $\frac{10+1}{2}=5.5$，即第 5 个数和第 6 个数之间的一个数值，把两数相加被 2 除。

$$M = \frac{96+124}{2} = 110$$

2.1.3.2 分组资料　用频数分布表进行计算。先求出中间位置，因例数较多，用 $\frac{n}{2}$ 即可。累积各组频数到包含中间位置，包含中间位置的组称中位数组。在中位数组用插入法即可求出中间位置所对应的确切数值，即中位数，也可用公式计算。

例 2.7　146 名食物中毒病人潜伏期分布如下，计算其中位数。

表 2－5　146 名食物中毒病人潜伏期

潜伏期（小时）	例 数	累计频数
0－	17	17
6－	46	63
12－	38 < 10/28	101
18－	32	
24－	6	
30－	1	
36－	4	
42－	2	

先求中间位置：$\frac{146}{2}=73$。累计频数至 101，包含 73，故 12－18 即为中位数组。此组的第 10 人所占的位置正是中间位置，它的潜伏期正是中位数。插入法是假定在 12－18 时区间的 6 小时范围内，均匀排列着 38 人，求第 10 人所对的时数，即

$$12+\frac{6}{36}\times 10 = 13.58(\text{小时})$$

公式为：

$$M=L+\frac{i}{f_m}\left(\frac{n}{2}-c\right)$$

$$=12+\frac{6}{38}\left(\frac{146}{3}-63\right)=13.58\text{(时)}$$

式中：L 为中位数组下限，此例为 12，

i 为组距，此例为 6，

f_m 为中位数组频数，此例为 38，

c 为中位数组前的累计频数，此例为 63。

插入法和公式计算完全一样，按照插入法的思路可以完全不记公式。同理也可求得其它任一百分位数。

2.2 变异指标

自然界中的一切事物，包括所有生物个体之间都存在着一定的矛盾——差异，这种差异称为变异。只用集中趋势指标不能全面地表示一组变量值的特征，还需用一些指标反映其变异即离散程度的大小。常用的变异指标有全距、离均差平方和、方差、标准差，变异系数等。

例 2.8：譬如有甲、乙两群羊，各 5 只，其 kg 数如下。

甲：10、20、30、40、50(kg)

乙：26、28、30、32、34(kg)

这两群羊的均数都是 30 kg，但其变异程度却大不一样，乙群比较匀称，甲群比较离散。

2.2.1 全距 是一组变量值最大值和最小值之差，也称极差。全距是表示变异程度的最简单指标。例 2.8 中，甲群的全距为 $50-10=40$(kg)，乙群为 $34-26=8$(kg)，两者相差很大。但全距只考虑最大值和最小值，其它变量值全不考虑。如果还有一群的体重为 10、30、30、30、50 kg，虽然变异程度减小了，但全距和甲群没有区别。

2.2.2 标准差 标准差的确立有个逐步演化的过程。开始首先想到离均差之和，即 $\sum(x-\bar{x})$，但此值为 0；又想到离均差平方和，即 $\sum(x-\bar{x})^2$，此值能反映变异情况，考虑到全部变量值的大小，但是受变量值个数的影响，个数越多，此值越大；因此又想出方差，即离均差平方和除以例数，

$$\text{方差}=\frac{\sum(x-\bar{x})^2}{2}$$

似较合理，但各变量值的单位取平方，如 kg^2，变得无法理解；最后想出来标准差，即方差的平方根，$s=\sqrt{\frac{\sum(x-\bar{x})^2}{n}}$，使之更为合理，应用也最广。计算公式有分组资料和不分组资料之分。

2.2.2.1 不分组资料 其公式有两种形式

$$S=\sqrt{\frac{\sum(x-\bar{x})^2}{n-1}}$$ 或

$$S=\sqrt{\frac{\sum x^2-\frac{(\sum x)^2}{n}}{n-1}}$$

式中分母为 $n-1$ 是为了使算得的结果更接近总体标准差。

例 2.8 中两群羊的标准差分别为：

$$S_{甲}=\sqrt{\frac{(10-30)^2+(20-30)^2+(30-30)^2+(40-30)^2+(50-30)^2}{5-1}}$$

$$=\sqrt{250}=\pm 15.81\text{(kg)}$$

或

$$S_{甲}=\sqrt{\frac{(10^2+20^2+30^2+40^2+50^2)-\frac{(10+20+30+40+50)^2}{5}}{5-1}}=\sqrt{\frac{5500-\frac{22500}{5}}{4}}=\sqrt{\frac{1000}{4}}=15.81\ (kg)$$

$$S_{乙}=\sqrt{\frac{(26-30)^2+(28-30)^2+(30-30)^2+(32-30)^2+(34-30)^2}{5-1}}=\sqrt{10}=\pm 3.16\ (kg)$$

2.2.2.2 分组资料 分组资料的算术标准差多用简捷法，其公式为：

$$S=\sqrt{\frac{\sum fx^2-\frac{(\sum fx)^2}{\sum f}}{\sum f-1}}\ i$$

在简捷法求均数的计算表上再算一步fx^2，即能求标准差。

例 2.9 求 120 名 12 岁男孩的身高标准差。

表 2-6 120 名 12 岁男孩身高标准差计算表

分 组	组中值	频数 f	化减值 x	fx	fx^2
125-	127	1	-4	-4	16
129-	131	4	-3	-12	36
133-	135	9	-2	-18	36
137-	139	28	-1	-28	28
141-	143	35	0	0	0
145-	147	27	1	27	27
149-	151	11	2	22	44
153-	155	4	3	12	36
157-	159	1	4	4	16
		$\sum f=120$		$\sum fx=3$	$\sum fx^2=239$

将表中的数值代入公式：

$$S=\sqrt{\frac{239-\frac{3^2}{120}}{120-1}}\times 4=\sqrt{\frac{238.925}{119}}\times 4=5.67\ (cm)$$

2.2.2.3 几何标准差 是将各变量值求对数，几何均数也用对数，代入公式运算后再求反对数。算几何标准差常用加权法，其公式为：

$$S=\lg^{-1}\left(\sqrt{\frac{\sum f(\lg x-\lg G)^2}{\sum f-1}}\right)$$

表 2-7 凝集效价几何标准差计算表

凝集效价 x	人数 f	$\lg x$	$f(\lg x-\lg G)^2$
20	6	1.3010	1.9976
40	9	1.6021	0.6851
80	21	1.9031	0.0132
160	7	2.2041	0.7444
320	5	2.5051	1.9963
	$\sum f=48$		$\sum f(\lg x-\lg G)^2=5.4066$

例 2.10 某疫苗接种后，求 48 人血清凝集效价的几何标准差，（$\lg G=1.8780$）