中等卫生学校配套教材

卫生统计学学习指导

主编 韩 敏
刘宝俊

民卫生出版社

图书在版编目（CIP）数据

卫生统计学学习指导/韩敏，刘宝俊主编．－北京：人民卫生出版社，1999

ISBN 7-117-03197-2

Ⅰ．卫… Ⅱ．①韩… ②刘… Ⅲ．卫生统计-专业学校-学习参考资料 Ⅳ．R195.1

中国版本图书馆 CIP 数据核字（1999）第 00670 号

卫生统计学学习指导

韩 敏 刘宝俊 主编

人民卫生出版社出版发行

（100078 北京市丰台区方庄芳群园 3 区 3 号楼）

三河市潮河印刷厂印刷

新华书店经销

787×1092 16 开本 7 1/2 印张 162 千字

1999 年 3 月第 1 版 1999 年 3 月第 1 版第 1 次印刷

印数：00 001—6 000

ISBN 7-117-03197-2/R·3198 定价：8.50 元

（凡属质量问题请与本社发行部联系退换）

前　言

为配合第三版规划教材《卫生统计学》的教学，在卫生部教材办公室、人民卫生出版社的大力支持和指导下，由山东省淄博第二卫生学校组织，部分卫生学校参加编写了这本《卫生统计学学习指导》，作为新一版教材的配套教材，供师生教学使用。

本书从章节顺序到具体内容均与教材相对应。本着提高学生统计操作能力的宗旨，确定了每章的学习目标和重点内容，对难以理解的抽象问题，作了较为具体的说明。设置了大量练习题、自测题，在让学生熟悉基本知识的基础上，给学生提供了训练和培养统计操作能力的良好条件。同时根据需要，对部分章节的内容进行了适当的扩充。

本书在编写过程中得到了规划教材主编周士楷老师及有关领导、老师的支持和帮助，在此一并致谢。

由于我们水平有限，书中难免有不妥之处，敬请广大师生指正。

编　者

1998 年 10 月

目　录

第一章 概 述

课 时 目 标

1. 简述卫生统计学的性质、任务和学习目的。
2. 列出卫生统计工作步骤。
3. 解释几个基本概念。
4. 具有实事求是的科学态度。

内 容 提 要

本章主要介绍了卫生统计学的概念、主要内容、应用和发展，以及预防医学专业学习的目的、学习要求；卫生统计工作包括统计全过程设计、搜集资料、整理资料和分析资料四个基本步骤；统计中关于观察单位、变异、变量、变量值、变量类型、总体与样本、概率与频率、误差、参数与统计量和统计推断等几个基本概念。

学 习 指 导

一、卫生统计学的性质、任务和学习目的

卫生统计学是把统计理论与方法应用于居民健康状况、医疗卫生工作实践和医学科学研究的一门学科。

卫生统计学的任务是借助于统计方法，从有限的观察中，从表现为偶然的数据中，把所研究的事物或现象的本质特征、整体情况和与其他事物或现象间的关系一一揭示出来。

学习卫生统计学的目的是掌握卫生统计学的理论、基本知识和基本技能，为学习各门专业课、阅读专业书籍、从事预防医学工作打下必要的统计学基础。学习时应明确目的，注意学习方法，培养科学作风和统计思维能力。

二、卫生统计工作的步骤

统计全过程设计、搜集资料、整理资料和分析资料是卫生统计工作的四个基本步骤，各步骤的有关问题可参见教材内容。注意理解各步骤的基本概念和方法，这些内容将在以后各章中进一步进行描述。

三、统计中的几个基本概念

1. 观察单位 是获得数据的最小单位。观察单位可以是人、标本、家庭、国家等。例如，了解某班学生学习成绩，每一个同学就是一个观察单位；然而了解全校各班学习情况，则每一个班就是一个观察单位，要注意理解掌握观察单位的概念。

2. 变异 是统计工作的前提。统计研究的就是有变异性的事物，没有变异当然就

无所谓统计，然而在任何相同条件下，个体间始终存在着一定的差异，例如，对于同时出生的人来说，假定在相对同等条件下，其身高、体重、心率等生理指标也不尽相同。又如，同样接触了某种传染病的人，有的发病有的不发病，发病的又有轻有重等，这种差异统计学中称为变异，即在同质条件下的观察单位，其同一标志的数量间存在着差别。结合日常工作的实际，进一步理解变异这一概念，对于认识统计的作用具有十分重要的意义。

3. 变量和变量值　变量是观察单位的某项特征。例如儿童的身高、体重、心率，工人的工资收入，学生的学习成绩等都是变量，工人的工种，人口的性别、民族等也是变量。这些变量在观察过程中，每一个观察单位都有一个观察结果，而这些结果可能不尽相同，我们称这些观察结果为变量值。

4. 变量类型　根据观察值的性质不同可以把变量分为数值变量和分类变量。

（1）数值变量（又称定量变量）：是以计量方式所得到的观察结果，一般都带有度量衡单位，如脉搏（次/分）、血压（kPa）等。数值变量可以是连续性变量也可以是不连续性变量，如身高是连续性变量，某日某交通点通过的汽车辆数是不连续性变量。

（2）分类变量（又称定性变量或字符变量）：分类变量的变量值是代表互不相容类别或属性的字符，例如，职称、工种等。分类变量可以是两项分类变量也可以是多项分类变量；可以是有序分类变量也可以是无序分类变量，如工种是无序分类变量，职称是有序分类变量。当分类变量值之间存在程度或等级上的差别，这种程度或等级带有“半定量”的性质，称为有序分类变量或等级变量。如治疗效果分为痊愈、显效、好转、无效等；学习成绩分为优秀、良好、及格、不及格等。

以数值变量值为原始数据的资料称为数值资料；以分类变量为原始数据，清点并汇总具有不同类别变量值的观察单位的个数，编制成分类变量频数表的统计资料称为分类资料，其中以等级变量为原始数据，分类计数其观察单位个数，编制成频数表的资料称为等级资料。正确理解数值资料和分类资料是统计分类的前提条件，因为不同的资料其适用的统计分析方法不同，只有正确地鉴别资料属于何种类型，才能正确使用统计分析的方法。教材有关章节分别介绍了各种不同类型的资料所适用的统计方法。如数值资料进行平均数、标准差、t 检验、F 检验、相关与回归分析等，分类资料进行相对数、χ^2 检验分析等。有时为了分析上的需要也可以把有关资料通过变换改变其资料类型。如数值资料定性化，分类资料定量化等，根据变化后的资料类型选择相应的统计分析方法。

5. 总体与样本

（1）总体：是根据研究目的确定的性质相同的所有观察单位某种变量值的集合体。在理解这一概念时应注意总体是由研究目的确定的，不同的研究目的就会有不同的总体，构成总体的每一个观察单位具有相同的性质，这样其观察值才有统计分析的意义。例如，研究某地区医疗网点工作量时，所有的医疗网点构成统计总体；而研究某医院工作质量时，则这一医院就是一个总体。

（2）样本：是总体内随机抽取的一部分。在理解这一概念时要注意样本是为研究总体时而抽取的。因此为了使样本具有一定的代表性，抽取样本时必须遵循随机化原则从总体中抽取总体单位，构成样本，这样样本才具有代表性。所谓随机化，是指总体内的每一个单位都有均等的机会进入样本。教材所介绍的许多由样本信息推断总体特征的方

法都是建立在随机抽样基础上的，如果样本不是随机抽取的，则有关统计分析的方法就失去了理论依据，因此也就不能用样本指标来推断总体指标了。

6. 概率与频率　概率是事件发生的可能性大小的量度，通常以符号 P 表示，当某实际事件肯定发生时称为必然事件，其概率 $P=1$；当某事件不可能发生时称为不可能事件，其概率 $P=0$；当某事件在一定条件下可能发生也不可能发生时称为随机事件，其概率在 $0<P<1$ 的范围内。统计研究的是随机事件，在大量观察中找出随机事件发生发展变化的规律和趋势是统计研究的目的之一。统计中有一个公认的道理，即“小概率事件在一次观察中，可以认为不会发生”。正确理解这一点具有重要的意义。如在装有 100 个苹果的箱子里，有不到 5 个坏苹果，其坏苹果的概率 $P<0.05$，并且这些坏苹果是均匀的分布在箱子内的，从中“随机抽取一个就是坏苹果”这一事件可以认为是不会发生的，假设一次抽到的就是一个坏苹果时，我们可认为其坏苹果的概率 $P>0.05$，利用这一道理可以帮助我们来解释假设检验中的假设是否成立这一问题。

频率也是某事件出现的可能性大小的量度，只不过概率是对总体而言，频率是对样本而言，在相同的条件下，当 $n\to N$ 时，频率可作为概率的近似值。

7. 误差　误差是指测得值与真值之差或样本指标与总体指标之差，从误差的性质来看，可以分为两大类，即偶然误差和系统误差。

(1) 偶然误差（又称随机误差）：包括抽样误差和随机测量误差。抽样误差是指由于抽样造成样本指标与总体指标之差，这是由于总体内各观察单位存在着个体差异，在这种偶然因素的影响下，不可避免地会出现样本结构不同于总体结构，因而样本指标也就不会等于总体指标，但是当不断增加样本含量时，可以缩小抽样误差。随机测量误差是指由随机测量变异引起的误差，在对总体单位观察的过程中不可避免会产生随机测量误差，但是改善测量手段和条件，随机测量误差可以控制在比较小的范围内（一般是指可以允许的范围内）。

(2) 系统误差：是指由确定的原因引起的观察值与真值之间或样本指标与总体指标之间的偏差。产生系统误差的原因很多，其中常见的是由于观察条件不同引起的偏差，如试验仪器、试剂、操作方法、疗效判断标准不同等，由此而造成的系统误差在观察过程中发生，表现为观察值偏离真值。例如，天平砝码未校正，其真实重量（严格地说应为质量）比其所示值偏小时，测得的物质重量将比物质的真实重量为小。注意正确理解系统误差的性质，系统误差与偶然误差的性质比较见表 1.1。

表 1.1　系统误差与偶然误差的性质比较

误差类型	大小	方向	大小和方向的重现性	产生的原因	可专消除	统计规律性
偶然误差	一般较小	双向	不一定重现	多种影响较小的因素综合的影响	不可避免但可控制	有
系统误差	一般较大	单向	可重现	有少数确定的原因	消除原因即可避免	无

系统误差一般发生在观察过程中，但也可以发生在统计设计和资料分析时，此时从观察单位所获得的观察值可以是正确的，误差表现为结论偏离真实情况。如调查某地青年吸烟情况，以大中专学生为观察单位，调查结果吸烟率必然比真实吸烟率为低。又

如，分析不同职业与高血压患病率的关系时，由于不同职业人群的平均年龄可能差别很大，而年龄不同的人群高血压患病率差别可以很大。因此，必须排除年龄因素的干扰，才能得出正确的结论。统计上把这种干扰因素称之为混杂因素，由于混杂因素的干扰，而造成了统计结论偏离真实情况的现象称为混杂。事物或现象间存在着广泛的、错综复杂的联系，因此当我们分析研究某些事物或现象间的联系时，必须注意排除同时存在着的混杂因素的干扰。如何在统计设计和统计分析时控制和排除混杂因素的干扰，是统计学的重要内容之一。

还有一类误差是由于过失造成的，它也具有系统误差的性质，但它是一种非技术性的、责任性的错误。如读数错、计算错、记录错、录入错等，这类误差不包括在通常所说的系统误差之中。

8．参数和统计量　总体的指标称为参数，样本的指标称为统计量。例如，某中专学校18岁男生身高的总体均数就是一个参数，而该校随机抽取的100名18岁男生，其身高的均数就是一个统计量。统计学约定参数用希腊字母表示，统计量用拉丁字母表示。如 μ 表示总体均数，π 表示总体率，σ 表示总体标准差，ρ 表示总体相关系数等，$\bar{x}$ 表示样本均数，p 表示样本率，s 样本标准差，r 表示样本相关系数等。学习时除了理解基本概念外，还应注意它们代表符号的书写和读音。

9．统计推断　根据样本资料所提供的信息，对总体的特征作出推断，称为统计推断。统计推断包括两个方面：

（1）参数估计　参数估计是根据样本资料所提供的信息，对总体指标的大小或所在范围作出估计。这种估计又分为点估计和区间估计两种。①点估计：是对总体指标作出一个定值的估计，虽然能给人一个明确的数量概念，但这只是一个近似值，常常不能满足实际工作的需要。②区间估计：是估计总体参数所在的范围以及在这个范围内包含总体参数的可能性的大小。

（2）假设检验　首先对总体指标作出一个假设，然后根据样本资料所提供的信息及有关统计量分布理论，对这个假设作出拒绝或不拒绝的判断。这种对于假设的拒绝或不拒绝的判断是具有概率性的，也就是说，这种判断的正确性不是百分之百的，即它是冒着犯有一定概率的错误风险作出判断的。然而这种判断比之于那种说不出判断错误概率的经验判断，要严密得多，可靠得多。因而假设检验作为一种经典的数据处理方法，早就成为自然科学和社会科学研究中的一种通用的方法。

假设检验有许多种，根据其所计算的统计量不同而命名，如 t 检验、U 检验、F 检验、χ^2 检验等。本书将以大量篇幅介绍各种假设检验方法。

四、具有实事求是的科学态度

统计是认识社会的重要手段，也是一种重要的管理工具，为了使我国的统计工作适应社会主义现代化建设事业的需要，国家制定并颁发实施了《中华人民共和国统计法》，其目的是为了科学有效地组织统计工作，为保障统计资料的准确性、及时性提供法律保证，以发挥统计在社会主义现代化建设中的服务和监督作用。我们应当自觉养成实事求是、严肃认真的科学态度，正确对待各种统计结果，避免片面性，把统计分析与专业理论有机地结合起来。

综合练习题

一、填 空 题

1. 卫生统计学是把_____，应用于_____、_____和_____的一门科学。

2. 卫生统计学的主要内容包括_____、_____、_____。

3. 卫生统计工作包括_____、_____、_____、_____四个基本步骤。

4. 观察单位是_____的最基本的、最小的单位。

5. 变量值是指对于_____的某项特征（变量）的_____。

6. 按观察值的性质不同可以把变量分为_____和_____。

7. 总体是根据_____确定的_____的所有观察单位_____的集合。

8. 概率是_____发生的_____的度量，我们经常遇到的事件可分为_____、_____和_____三种类型。

9. 统计上所说的误差，包括_____与_____之差和_____与_____之差。

10. 从误差的性质看，可以把误差分为_____和_____误差。

11. _____称为参数，_____称为统计量。

12. 统计推断包括_____、_____两个方面。

二、是 非 题

1. 学习卫生统计学的目的是为了掌握统计分析的方法。（ ）

2. 学习卫生统计学应培养科学的统计思维能力。（ ）

3. 由现场调查和实验研究搜集的资料称为经常性资料。（ ）

4. 观察单位可以是人、标本、家庭，但不能是国家。（ ）

5. 由于测量手段或条件的波动而造成测量结果的差异是个体变异。（ ）

6. 数值变量都是连续变量。（ ）

7. 分类变量的变量值是代表互不相容类别或属性的字符。（ ）

8. 数值资料是以数值变量为原始数据的统计资料。（ ）

9. 从总体内抽取的一部分称为样本。（ ）

10. 随机事件的概率为 1。（ ）

11. 小概率事件在一次观察中，可以认为不会发生。（ ）

12. 偶然误差就是偶然发生的误差。（ ）

13. 系统误差是由确定性的原因引起的观察值与真值之间和样本指标与总体指标之间的偏差。（ ）

14. 偶然误差是可以避免的，系统误差是不可避免的。（ ）

15. 区间估计是对总体指标作出一个定值的估计。（ ）

三、选 择 题

1. 了解医院住院病人，其观察单位是____。
   A. 一所医院　B. 全部医院　C. 一个科室　D. 某一地区
2. 观察单位间的变异是____
   A. 个体变异　B. 群体变异　C. 随机测量变异　D. 偶然误差
3. 下列属于数值变量的有____
   A. 脉搏　B. 血压　C. 学习成绩　D. 职称
4. 了解某校学生学习成绩，其总体是____
   A. 全体学生　B. 全部学生学习成绩
   C. 每一个学生　D. 每一个学生的学习成绩
5. 某事件发生的概率 $P=0.05$，这一事件是____
   A. 必然事件　B. 不可能事件　C. 随机事件

四、问 答 题

1. 卫生统计学的任务是什么？为什么要学习卫生统计学？
2. 统计资料分哪几种类型？区分统计资料类型的依据是什么？
3. 系统误差与偶然误差的性质如何？各举例说明。
4. 为什么说即使观察值是正确的，系统误差仍然可能发生？
5. 什么是混杂因素？它对统计结论有什么影响？

五、名 词 解 释

1. 数值变量　2. 分类变量　3. 总体　4. 系统误差　5. 统计推断

（山东省淄博第二卫生学校　韩　敏）

第二章　数值资料的统计描述

课 时 目 标

1．列出频数分布表和绘制频数分布图。

2．区分频数分布类型。

3．列举集中趋势指标与离散程度的指标，叙述其应用条件并会计算。

4．描述正态分布曲线的特征及规律。

内 容 提 要

本章主要介绍了数值资料统计描述的一般方法，编制数值资料的频数分布表、频数分布图，集中趋势指标（算术平均数、几何平均数、中位数）与离散程度指标的计算及适用的资料类型，正态分布及其应用。

学 习 指 导

一、数值资料的统计描述

首先应编制频数分布表以了解其分布状况，频数就是观察值的个数。频数分布就是观察值在其所取值的范围内分布的情况。

频数分布表的编制步骤：

1．计算全距　全距＝最大值－最小值

2．确定组段数、组距和组段　组段数一般为10～15个，全距大，观察值个数多可多取些，反之可少取。组段数太多，较繁琐，不易反映分布的特征，组段太少计算误差较大，实际工作中可根据具体情况决定。组距＝全距/组数，但在实际中组距常取整数或比较合适的小数，故此可将计算结果适当调整。第一组组段要包括资料的最小值，一般可从一个较为合适的数值开始，最后一组的组段应包括资料的最大值，同时应封口。

3．列表归组汇总　将各组段列入频数分布表的第一栏，用划记法将各观察值划记到各组段，即频数分布表第二栏，求出各组段频数及总频数列入第三栏。其次，在编制频数分布表的同时，也可绘制频数分布图，以更加直观地了解频数分布情况。频数分布图是表达频数分布的几何图形，常见的频数分布图有两种，即直方图和多边图。频数分布图绘制时应先划一直角坐标系，横轴表示各组段，纵轴表示各组段的频数，然后根据资料的实际数值绘制直方图。

二、频数分布类型

数值资料常见的频数分布类型有三种，如何区分关键是看分布高峰的位置。

1．正态分布型　频数分布的高峰位于中央，图形左右对称。正态分布属于此类型。

2．正偏态分布型　频数分布的高峰偏左，图形左右不对称，即观察值较小的一端集中了较多的频数。

3．负偏态分布型　频数分布的高峰偏右，图形左右不对称，即观察值较大的一端集中了较多的频数。

三、集中趋势指标

集中趋势指标又称平均数，它反映了观察值的集中位置或平均水平，是观察值的典型水平或代表值。常用的集中趋势指标有算术均数（均数）、几何均数和中位数等。计算平均数时，首先应搞清楚它们的应用条件，现把各种平均指标的应用条件归纳如表2.1。

表2.1　各平均指标的应用条件

指　标	适 用 条 件	计算公式
算术平均数	常用于描述对称型分布，尤其是正态	$\overline{X}=\frac{\sum X}{n}$
	分布资料的集中趋势	$\overline{X}=\frac{\sum fX}{\sum f}$
几何均数	常用于描述对数正态分布资料和观察	$G=\lg^{-1}\left(\frac{\sum \lg X}{n}\right)$
	值呈等比数列资料的集中趋势	$G=\lg^{-1}\left(\frac{\sum f\lg X}{\sum f}\right)$
中位数	常用于描述偏态分布资料、一端或两	$M=X_{(n+1)/2}$
	端无界的资料、频数分布类型不清楚	$M=(X_n+X_{n+1})/2$
	的资料的集中趋势	$M=L+\frac{i}{f_m}\left(\frac{n}{2}-\sum f_L\right)$

四、离散程度指标

离散程度指标又称变异程度指标。它反映观察值之间参差不齐的程度。常用的离散程度指标有极差、标准差和变异系数等。现将离散程度指标、计算公式及主要优缺点归纳如表2.2。

表2.2　离散程度指标比较表

指　标	计 算 公 式	主 要 优 缺 点
极　差	$R=X_{max}-X_{min}$	计算简单，易于理解；但只反映了一组观察值的最大值与最小值的差异，不能反映其他观察值间的变异情况
离均差平方和	$SS=\sum(X-\overline{X})^2$	反映了各变量值之间的变异情况，但单位是原观察值单位的平方，不易理解，同时又受观察值个数的影响，不利于比较

续表

指　标	计 算 公 式	主 要 优 缺 点
方差	$s^2 = \sum (X - \bar{X})^2 / n$	反映了各观察值间的变异情况，不受观察值个数的影响，但单位是原观察值单位的平方，不易理解
标准差	$s = \sqrt{\frac{\sum X^2 - (\sum X)^2 / n}{n - 1}}$	反映了各观察值间的变异情况，不受观察值个数的影响，单位与原观察值单位相同，是最常用的离散程度指标之一，但在两组和多组资料比较时，常受到计量单位不同和均数相差较大的影响而不能比较和不便于比较
变异系数	$CV = \frac{s}{\bar{X}} \times 100\%$	两组或多组资料比较变异程度，如均数相差过大或观察值单位不同时用变异系数比较

五、正态分布曲线的特征及规律

正态分布曲线是一条高峰位于中央（即均数所在处）两侧逐渐下降并完全对称，两端永远不与横轴相交的钟型曲线。

正态曲线的特征是整个曲线都在横轴的上方，均数处最高；以均数为中心，左右对称。正态分布曲线有两个重要参数，即 μ 和 σ，μ 决定曲线的位置，σ 决定曲线的形状。

正态分布曲线下面积分布规律：

1．一般正态分布

$\mu \pm \sigma$ 范围内的面积占总面积的 68.27%

$\mu \pm 1.96\sigma$ 范围内的面积占总面积的 95.00%

$\mu \pm 2.58\sigma$ 范围内的面积占总面积的 99.00%

2．标准正态分布（标准正态分布中 $\mu = 0$，$\sigma = 1$）

$-1 \sim 1$（0 ± 1）之间的面积占总面积的 68.27%

$-1.96 \sim 1.96$（0 ± 1.96）之间的面积占总面积的 95.00%

$-2.58 \sim 2.58$（0 ± 2.58）之间的面积占总面积的 99.00%

六、标准差的应用

标准差用来描述观察值间的变异程度（离散程度），用于正态或近似正态分布资料，标准差结合均数描述分布特征。标准差主要用来衡量观察值间的离散（或变异）程度。标准差还可以用于计算变异系数，变异系数又称离散系数，它是标准差对均数的相对百分数，故又有相对标准差之称，以符号 CV 表示，按式（2.1）计算。

$$CV = \frac{s}{\bar{X}} \times 100\% \qquad (2.1)$$

和标准差一样，CV 越小，表示观察值的离散程度越小。当比较不同组观察值的离散程度时，如果不同组的均数相差较大，就不能直接用标准差作为比较指标，而应采用变异系数作为比较指标。当比较不同组观察值的离散程度时，如果被比较的观察值的单