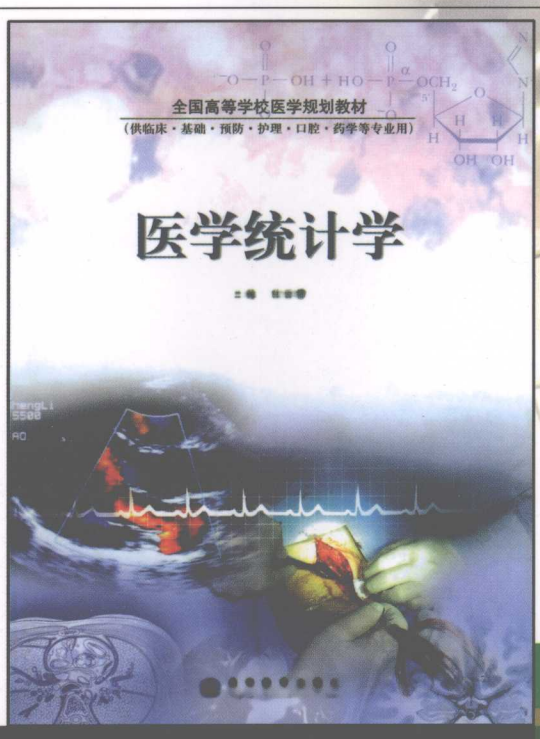
成功笔记系列丛书

医学统计学成功笔记

——含SPSS 13.0实习指导

陈卫中◎主编　成功笔记系列丛书编写委员会◎编



NOTES TO SUCCESS

哈尔滨工程大学出版社

成功笔记系列丛书

医学统计学成功笔记

——含 SPSS 13.0 实习指导

（配倪宗瓒第 1 版教材·高教版）

陈卫中　主编
《成功笔记系列丛书》编写委员会　编

哈尔滨工程大学出版社

内容简介

本书是配合倪宗瓒主编的《医学统计学》一书而编写的辅导书，并包含SPSS 13.0软件学习指导。本书共分为两大部分，二十五章，分别介绍了医学统计学的基本概念、基本方法和基本技能，以及各种统计方法的应用条件、优缺点，并更加注重了各种基本方法的适用条件、适用范围和基本思想的总结，对学生应试以及处理实际问题都有一定的参考价值。

本书适用于各医学专业以及预防医学、卫生事业管理等专业的本科生和研究生使用。

图书在版编目(CIP)数据

医学统计学成功笔记：含SPSS13.0实习指导/陈卫中主编.—哈尔滨：哈尔滨工程大学出版社，2008.1
（成功笔记系列丛书）
ISBN 978-7-81133-075-5

Ⅰ.医… Ⅱ.陈… Ⅲ.医学统计-高等学校-教学参考资料 Ⅳ.R195.1

中国版本图书馆CIP数据核字(2007)第202977号

出版发行 哈尔滨工程大学出版社
社　　址 哈尔滨市南岗区东大直街124号
邮政编码 150001
发行电话 0451-82519328
传　　真 0451-82519699
经　　销 新华书店
印　　刷 黑龙江省教育厅印刷厂
开　　本 787mm×960mm 1/16
印　　张 15.25
字　　数 230千字
版　　次 2008年1月第1版
印　　次 2008年1月第1次印刷
定　　价 25.00元
http://press.hrbeu.edu.cn
E-mail:heupress@hrbeu.edu.cn

成功笔记系列丛书编委会

前言

Preface

医学统计学作为统计学在医学领域中的应用性学科，是理论与实践相结合的典范，是进行医学科研必不可少的工具，是医学生必须掌握的基础课程之一。但是由于其思维方式更加注重抽象思维和推理能力运用，而大多数医学课程主要是对形象思维和识记能力的训练，学生对统计学普遍存在一种恐惧心理，大多数都反映学不会统计，学不通统计。编者作为一个研究统计学多年的人，非常想为普及统计学，使大家理解统计学，发现统计学带给大家的快乐多做点什么，恰逢本丛书筹稿出版，故有了这么一个机会，和大家交流一下有关统计学的学习和掌握的经验。

在学习医学统计学的过程中，关键在于掌握医学统计学的基本概念，基本原理、基本方法和基本技能，以及各种统计方法的应用条件、优缺点，最终的目的是解决医学中的实际问题，因此有的时候我们完全可以忽略那些繁杂的公式，毕竟有现成的软件 SPSS 帮我们完成计算的过程。我们的任务不过是在根据条件，选准方法的基础上点击鼠标而已。因此本书在成书过程中更加注重了各种基本方法的适用条件、适用范围和基本思想的总结，应该说是教材的有力补充和简化，对大家应试以及处理实际问题都有一定的参考价值。

本书由陈卫中老师主编，张丽蓉老师为本书的编校工作做了很多的工作，哈尔滨工程大学出版社给予了大力的帮助。在成书过程中还得到学校领导和教研室老师的大力支持，在此一并致谢。

本书在编写过程中，力求准确并通俗易懂，但限于本人的水平、表达能力、时间有限，谬误在所难免，望各位专家、老师、同学不吝赐教，以期更加完善。

编者

2007 年 10 月

Contents 目录

第一部分　医学统计学基本理论

第二部分 SPSS 13.0 实习指导

第一部分
医学统计学基本理论

第1章 绪 论

1.1 医学统计学的定义

医学统计学是应用概率论和数理统计的基本原理和方法，研究医学领域中数据的收集、整理和分析的一门科学。

1.2 医学统计学的主要内容

1.统计研究设计
2.常用的基本统计方法
3.临床医学中常用的统计方法
4.常用医学人口和疾病统计指标及其应用，寿命表的编制原理及其在医学上的应用
5.多因素分析的统计分析方法

1.3 统计工作的步骤

任何统计工作或统计研究都可以大致分为设计(design)、收集资料(collection of data)和整理资料(sorting data)和分析资料(analysis of data)四个步骤。

设计是研究的开始，贯穿研究的始终，包括专业设计和统计设计。专业设计是基础，回答要做什么的问题；统计设计围绕专业设计进行，使研究少走弯路，是保证

结论具有可推性的重要手段。

收集资料，即研究阶段，是通过试验或调查得到数据的过程。

整理资料是资料收集完毕后，净化原始数据，进一步保证数据准确无误，并使所得资料系统化、条理化，为资料的分析奠定基础。整理资料的过程大致包含以下三个方面：

(1)数据库的建立和数据的录入；

(2)数据的逻辑查错；

(3)数据的预处理，它是对原始数据进行综合、归纳而形成一些综合性指标，或者根据需要对数据作进一步的分组整理，对数据库中的缺失值进行一定的补充处理等过程，比如将年龄按照专业意义分成老、中、青三个层次等。

分析资料也就是根据研究目的计算有关指标，反映数据的综合特征，阐明事物的内在联系和规律，从而实现研究目的，得到研究结论。统计分析包括以下两个方面：

(1)统计描述(descriptive statistics)，指用统计指标、统计表、统计图等方法对资料的数量特征及其分布规律进行综合分析和描述；

(2)统计推断(inferential statistics)，指用样本信息推断总体特征的过程，包括参数估计和统计推断两部分。

1.4 统计学的几个基本概念

1.总体(population)

总体是指所有的同质观察单位的集合，也就是所有观察单位(个体)的变量值的集合。有些总体在规定的时间和空间范围内是可数的，称为有限总体(finite population)。但有些时候总体是抽象的，比如研究某药治疗高血压的疗效，这里，总体的同质基础是确诊高血压患者同时用某药治疗，此时的总体应包括现在以及将来所有用该药治疗的患者，是没有时间和空间范围限制的，因而观察单位数无限，称为无限总体(infinite population)。

(1)目标总体(target population)是指打算将试验结果外推的总体人群。

(2)实际总体(actual population)是指在试验中符合条件的人群，是目标总体中特殊的一部分，就是在试验中规定的纳入或排除标准内的人群。

2.样本(sample)

样本就是从总体中随机抽取(即要求总体中的每个个体都有相同的机会进入样本，而不受人为控制)的一部分观察单位。样本观察的目的在于用试验得到的样本信息来推断总体的特征，因此要求样本必须有足够的代表性和可靠性，除了样本

要随机抽取外，还必须有一定的样本含量(sample size)。

3.变量(variable)

变量是指被观察单位的特征。变量值即为该特征的观察结果，如身高是观察对象的特征，在统计学上称为变量，某观察者的身高为 174 cm，则 174 cm 为变量值。所有变量值的集合，称为资料(data)。

4.资料的类型

(1)数值变量，也叫定量变量、连续变量，其变量值以数值和单位的形式表达，可以是某实数范围区间的任意值。根据需要，数值变量可转化为有序分类变量，即等级化数据，类似于频数表的制作，如年龄可以根据专业意义划分为老、中、青三个等级，称为等级变量。但这个过程必然伴随信息的损失。

(2)分类变量，也叫计数资料，变量值是定性的，变量的特征表现为互不相容的两个或多个类别或属性。根据类别的多少，变量可以分为两分类和多分类变量，其中多分类变量又可根据各类别之间有无程度、等级的差别，分为有序多分类变量和无序多分类变量。

(3)等级资料，实际上就是有序多分类变量的另一种叫法。

5.概率

说明某事件发生的可能性大小，其大小在 0 到 1 之间。在许多情况下，事件发生的概率是未知的，但频率是可以通过试验得到的，频率 f = 事件出现的次数 m/总观察次数 n，其围绕概率上下波动。当观察或试验的次数极多时，频率逐渐逼近于概率，因此可以用频率来估计说明事件发生的可能性大小。

小概率事件，是指某事件发生的可能性很小，在单次的试验或观察中可以认为不发生，习惯上规定小概率事件的发生概率的界值为 0.05 或 0.01。

统计学中的统计推断都是在一定概率标准的基础上进行的。

6.参数的统计量

总体的统计指标，称为参数，它是一个定值，通常用希腊字母表示，如总体均数 μ、总体标准差 σ、总体率 π；而通过抽样调查得到的对应样本的统计指标称为统计量，通常用拉丁字母表示，如样本均数 $\bar{X}$、样本标准差 s、样本率 p 是变化的，可能每次和每次的值都不相同，但它不会离开总体参数太远，总是围绕总体参数上下波动。

第2章 计量资料的统计描述

计量资料的统计描述就是用统计图(表)、统计指标来描述资料的分布规律及其数量特征。计量资料根据其变量取值的特点分为两部分,一是集中趋势,说明一组数值总的集中位置或平均水平,可以作为总体的一个代表值,以给人一个明晰的印象,同时便于进行事物间的比较;二是离散趋势,说明数据的变异性大小、集中趋势指标的代表性等。

2.1 频数分布表和频数分布图

当观察单位较多时,为了了解变量的分布规律,通过资料的整理编制的频数分布表反映各变量取值及相应频数间的关系,它也是选择合适的统计指标和统计方法,进行统计分析的基础。

2.1.1 频数表和分布图的制作过程

1.求全距。

2.确定组数的组距。

3.划分组段。

4.划记求频数。

5.制作直方图。

2.1.2 频数表的主要作用

1.数据的分布范围。

2.数据的集中组段。

3.数据的分布形式。判断数据的分布形式是否对称,是一个近似正态分布它还是偏态分布。它为进一步选择统计指标和分析方法奠定基础。

4.发现特大、特小可疑值,以进一步净化数据。

5.可以使用组中值来代表该组段的所有值,从而简化了手工计算,但在计算机时代,这已经不能算作一个优势了。

2.2 集中趋势的描述

1.算术均数(arithmetic mean)

(1)总体均数用 μ 表示,样本均数用 $\bar{X}$ 表示。

(2)算术均数适用于正态分布资料或近似正态分布资料的描述。

(3)算术均数不适用于开口资料集中趋势的描述。

(4)计算方法有直接求算术均数法和用频数表求加权算术均数法。

(5)算术均数较稳定,同时综合了资料中的全部信息。

2.几何均数(geometric mean)

适用条件:几何均数达用于非正态分布;等级或近似等级资料;对数正态分布资料,如血清学试验中的滴度资料等。

如果资料中有负数或零时,应根据实际情况去掉负号,或加上一个正数使全部资料正值化后再求几何均数。

3.中位数(median)

(1)中位数是一个界值,是在所有资料中从大到小排序后位置居中的数值。

(2)中位数适用于任何分布,不过主要用于以下方面:

①分布不明或偏态分布资料;

②在频数表中一端或两端无确切值的资料;

③在正态分布中,和均数一致。

(3)计算方法有直接法和频数表法。

(4)中位数不受极端值的影响,但未考虑资料中的全部信息,它最多只与两个值有关系。

4.众数(mode)

(1)众数是在一组资料中出现频率最高的值,或者是在频数表中出现频率最高的组段。

(2)众数的概念容易理解,但当出现频率相等的组段或数值不是一个时,难以确定。同时该指标也没有充分利用样本的全部信息。

2.3 离散趋势的描述

1.极差(range)

极差也称为全距,是一组资料中最大值与最小值的差。

(1)可用来描述任何分布的资料。

(2)便于理解和计算。

(3)未充分利用样本的全部信息。

(4)受极端值的影响大,稳定性差。

(5)样本的一端或两端无确切值时,不能得到极差。

(6)样本含量相差悬殊时,较难说明样本间变异大小的比较。

2.百分位数(percentile)

百分位数是一个界值,表示资料从小到大排序后,处于资料某一位置的值常常是几个百分位数结合使用的,它较极差的稳定性要强,但也没有充分利用样本的全部信息。

3.四分位数间距(inter-quartile range)

四分位数间距是将全部资料从小到大排序后,中间50%值的极差,或者是 $P_{75}-P_{25}$ 的值,它常用于描述偏态分布资料的变异,和中位数联合使用分别描述资料的集中趋势和离散趋势,其特点与百分位数相同。

4.离均差平方和(sum of squares)

(1)用于描述正态分布或近似正态分布资料的离散性。

(2)充分利用了样本或总体的全部信息。

(3)具有可相加性和可分解性。

(4)受到样本含量的影响。

(5)量纲和原始资料不一致。

5.方差(mean of square)

方差消除了离均差平方和受样本量影响的缺点,较好地对正态或近似正态分布资料的变异性进行了描述,但依然存在量纲与原始资料不一致的问题。

6.标准差(standard deviation)

标准差由均方开方得到。

(1)用于描述正态或近似正态资料的变异程度,说明均数的代表性强弱。标准差与均数结合起来可以较好地对资料进行统计描述,文献中常见的 $\bar{X} \pm S$ 就是这个意思(这里的"±"不是区间的意思,而只是连接符号)。

(2)可以用于计算标准误。

(3)可以用于计算变异系数。

(4)结合均数和正态分布的规律估计参考值范围。

7.变异系数(*CV*,coefficient of variation)

$$CV = \frac{S}{\bar{X}} \times 100\%$$

变异系数主要用于对量纲不一致或均数相差悬殊的资料间的变异的比较。

对于样本来说，主要统计描述指标适用资料总结如表 2.1 所示。

表 2.1 统计指标的比较

	指标	资料类型		
		正态或近似正态	偏态	等级或对数正态
集中趋势	算术均数	√√		
	几何均数			√√
	中位数	√	√√	
	众数	√	√	√
离散趋势	极差	√	√	√
	四分位数间距	√	√√	√
	均方	√√		√(数据转换后)
	标准差	√√		√(数据转换后)

2.4 正态分布及其应用

正态分布也叫做高斯分布，是一种最常见、最常用的分布，医学中的许多现象都符合正态分布，同时当试验的次数足够多(样本量足够大)时，许多非正态分布也近似地服从正态分布，它也是许多检验方法的基础。

2.4.1 正态分布的图形(如图 2.1 所示)

1. 正态分布曲线是以总体均数为中心，两边对称，中间高、两边低的倒钟形曲线。

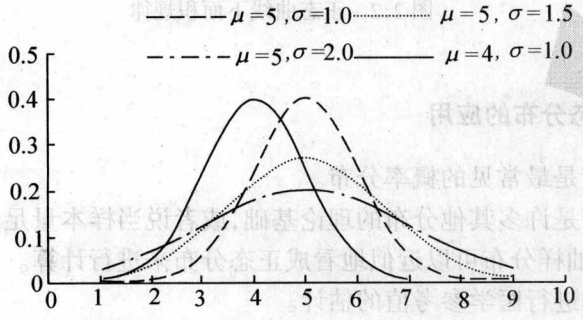


图 2.1 不同 μ 或 σ 的正态曲线