面向21世纪课程教材

Textbook Series for 21st Century

医用多元统计方法

张家放 主编

华中科技大学出版社

HUAZHONG UNIVERSITY OF SCIENCE AND TECHNOLOGY PRESS

E-mail: hustpp@wuhan.cngb.com

面向21世纪课程教材

医用多元统计方法

主　编　张家放
副主编　方　亚　董时富
参　编　尹　平　熊光炼　邱晓霞

华中科技大学出版社

图书在版编目(CIP)数据

医用多元统计方法/张家放　主编
武汉:华中科技大学出版社，　2002 年 4 月
ISBN 7-5609-2683-5

Ⅰ. 医…
Ⅱ. ①张… ②方… ③董… ④尹… ⑤熊… ⑥邱…
Ⅲ. 医学统计-多元分析-方法
Ⅳ. R195.1

医用多元统计方法　　主编　张家放

责任编辑:李　德　吴锐涛　周正国　　封面设计:秦　茹
责任校对:陈元玉　　责任监印:张正林

出版发行:华中科技大学出版社
武昌喻家山　邮编:430074　电话:(027)87545012

录　　排:华中科技大学出版社照排室
印　　刷:湖北新华印务有限公司

开本:787×960　1/16　　印张:25.75　　字数:475 000
版次:2002 年 4 月第 1 版　　印次:2002 年 4 月第 1 次印刷　　印数:1—3 000
ISBN 7-5609-2683-5/R・24　　定价:33.50 元

内容提要

本书重点介绍了各种多元统计分析方法的基本原理及其在医学上的应用，这些方法包括多因素的方差分析、多变量方差分析、多元线性回归分析、广义线性模型分析、多元 Logistic 回归分析、Poisson 回归分析、对数线性模型分析、生存分析、主成分分析、聚类分析、判别分析、典型相关分析、路径分析、探索性因子分析、确定性因子分析以及结构方程模型分析等。本书内容充实，使用方便。在每一章里详细论述了一种多元统计分析方法的基本原理和分析过程，介绍了 SAS 程序的使用方法、医学应用实例说明、结果解释、结论分析，章末有作为练习用的实习题目。

本书可以作为医科院校卫生统计学、预防医学、社会医学、卫生管理学等本科生以及研究生的统计专业教材，又可作为医科院校师生、医药卫生科研管理单位的科研工作者、科研管理者进一步提高统计分析水平的参考书。对于各种程度的读者都可以通过学习本书将自己的统计分析水平提高到一个新的层次。

编写说明

这本教材是根据教育部《关于积极推进"高等教育面向21世纪教学内容和课程体系改革计划"实施工作的若干意见》的精神和要求，在配合高等医学统计学教学内容和体制改革的实施中编写出来的，是国家高等医学教育学分制体系、教学内容改革及管理模式研究项目的一个重要科研成果。本教材的主要特点是：

第一，在教材体系上作了较大的更新。目前医学院校本科生所使用的医学统计学教材均是以基本统计分析为重点的，对于实用性很强的多元统计分析，由于计算工具的限制，始终停留在一般性的概括介绍的教学水平上。随着医学统计学的快速发展，这种体系的教材已经不能满足很多专业的需求了。为了适应医学统计学学科的发展趋势，为了提高医学院校本科生、研究生，特别是卫生统计、预防医学、流行病学等专业学生的统计分析能力，本教材作为基本医学统计分析教材的延伸，重点讲解了医用多元统计分析方法的原理、数据处理手段和医学应用。这对提高医学统计学教材的先进性和完善性是有重要意义的。

第二，在教学内容上作了较大的改革。传统的医学统计学教材均是以手工或计算器这种落后的计算工具为数据处理手段的。它最大的弊病就是在教材中不得不尽量缩减对基本统计分析原理的介绍，而将大量的篇幅用在繁杂公式的计算上。多元统计分析是一元统计分析的拓广，它的分析原理、计算方法、计算公式和计算量比一元统计要深得多、难得多、繁杂得多和大得多，如仍用落后的计算工具来处理多元统计分析中的数据，不仅耗时耗力，甚至是根本行不通的。多元统计的应用价值是在计算机发展和普及的今天才得以体现出来的。为此，本教材在教学内容上删减了旧教材中大篇幅的使用计算器计算的内容，增加了使用计算机统计软件解决实际问题的方法，同时补充和完善了重要的多元统计分析方法的基本原理。这样不仅使得学生对医学统计学的科学规律有更清楚的认识，而且使得学生具有真正分析和解决实际问题的能力和计算手段，这对于培养具有国际竞争能力的21世纪医学人才是有极其重要意义的。

第三，在教学方法和手段上有所创新。目前，医科院校的统计课程教学还保留着传统的教学法，即理论教学和实践操作脱节，或者说计算例题和解决实际问题脱节。长期采用这样的教学方法培养出来的学生只能死板地套用书上的公式做最简单的数据分析，而真正解决实际医学问题的能力却很有限。另一方面，这种教学方法使得医学统计的应用价值，特别是多元统计在医学研究上的应用价值不能真正体现出来。本教材在编排上从讲求实效的原则出发，将医学统计内容，特别是将统计分析原理和SAS软件的输出结果有机地结合在一起，将课堂理论教学和计算机实际操作融为一体，使学生能够有机会把各种统计分析方法的理论知识和操作计算机解决实际问题的能力结合起来。这种教学法是统计课程教学的方向，也是统计学教学方法改革的主要任务，是符合人才培养规律和教学规律的。

第四，本教材内容充实，重点突出，实用性强，使用方便。本教材的充实性表现在它系统介绍了诸多的多元统计分析方法及其在医学上的应用，例如各种类型的方差分析、各种类型的回归分析、各种类型的生存分析以及分类分析等等。其中，有的是近十年来才逐渐成熟和被人们接受的新方法，如路径分析、结构方程模型分析等。这些新方法使得这本书的内容更加充实，并使得这本书成为覆盖多元统计分析方法较完整的一本教科书。本科生教学可以把后面几个章节作为选修内容，研究生或有一定工作经验的医学科研工作者应当对这部分内容加以重视。本教材的实用性表现在它针对医学院校学生的特点，对每一种医学多元统计方法的基本内容给予了较全面的介绍，它不仅包含基本原理、基本计算方法等理论知识，还包含统计软件的使用介绍、例题说明以及课外实习题目，适合于各种类型的学生使用。对于教学时数较少或统计基础知识较薄弱的学生可以忽略计算公式等抽象的理论部分，把重点放在对基本原理的理解和使用计算机解决实际问题的方法应用上。对于有较好统计基础知识的学生可以通读每一部分的内容，使得自己的医学统计分析能力达到更高一级水平，并为今后进一步学习和研究打下扎实的理论和实践基础。

本教材第一章主要介绍了医学多元统计学在医学研究中的地位和重要性、结构方程模型分析这个最新方法的主要特点和用途、目前流行的统计分析软件以及统计学的几个重要概念。第二章概括地总结了基本统计分析方法的主要内容。从第三章开始逐个介绍了目前世界各国医学领域

里常用的一些多元统计分析方法，内容包括各种多元统计分析方法的基本原理、计算方法、分析步骤、医学应用、SAS 软件处理数据的程序和使用方法、医学应用举例以及结论分析等。另外，每一章还给出了使用 SAS 软件进行数据分析的练习题，以帮助学生理解和掌握各章的内容。为了不影响教材内容的系统性，每一章所涉及到的 SAS 过程步的基本结构和语句说明或使用方法全部列在附录一里面，共介绍了 23 个基本常用的 SAS 统计分析程序。关于 SAS 的基础知识，可参考有关的专业教材。

本教材第四章多变量方差分析由董时富教授编写，第八章 Poisson 回归分析由尹平教授编写，第十章生存分析由方亚教授编写，第十二章聚类分析由熊光炼老师编写，第十四章典型相关分析由邱晓霞老师编写，其余章节由主编完成。

非常感谢华中科技大学同济医学院各级领导对本教材的大力支持和帮助，感谢公共卫生学院以及流行病与卫生统计学系各位领导、教授对本教材编写的积极协作。

张家放

2002 年 1 月于武汉

目　录

第一章　绪论 …… (1)

第一节　多元统计学在医学中的地位 …… (1)

第二节　结构方程模型分析简介 …… (1)

第三节　多元统计分析软件的介绍 …… (3)

第四节　医学统计学的几个重要概念 …… (3)

第二章　基本统计分析方法 …… (6)

第一节　数据的描述性分析 …… (6)

第二节　总体参数的假设检验 …… (10)

第三节　两个总体参数的比较 …… (13)

第四节　多个总体参数的比较 …… (20)

第五节　简单相关分析 …… (30)

第六节　简单线性回归分析 …… (33)

习题二 …… (37)

第三章　多因子方差分析 …… (39)

第一节　无交互效应二因子方差分析 …… (39)

第二节　有交互效应二因子方差分析 …… (46)

第三节　三因子方差分析 …… (53)

第四节　特殊三因子方差分析 …… (57)

习题三 …… (60)

第四章　多变量方差分析 …… (62)

第一节　多变量方差分析的基本原理 …… (62)

第二节　多变量方差分析的过程 …… (64)

第三节　关于多变量方差分析的多重比较 …… (67)

第四节　多变量均值向量的 Hotelling T^2 检验 …… (67)

第五节　多变量方差分析在医学中的应用 …… (68)

习题四 …… (85)

第五章　多元线性回归分析 …… (88)

第一节　多元线性回归分析的基本原理 …… (89)

第二节　多元线性回归分析的数学模型 …… (90)

第三节　多元线性回归分析的方法步骤 …… (91)
第四节　多元线性回归分析的逐步回归法 …… (97)
第五节　多元相关分析 …… (99)
第六节　多元线性回归分析在医学中的应用 …… (101)
习题五 …… (106)
第六章　广义线性模型分析 …… (109)
第一节　广义线性模型的四种类型离均差平方和 …… (109)
第二节　协方差分析 …… (110)
第三节　广义线性回归分析 …… (116)
习题六 …… (124)
第七章　多元 Logistic 回归分析 …… (127)
第一节　Logistic 回归分析的基本原理 …… (127)
第二节　Logistic 回归分析的数学模型 …… (128)
第三节　Logistic 回归模型的建立和检验 …… (130)
第四节　Logistic 回归模型系数的解释 …… (134)
第五节　配对病例-对照研究的条件 Logistic 回归分析 …… (150)
第六节　Logistic 回归模型在医学中的应用 …… (155)
习题七 …… (155)
第八章　Poisson 回归模型分析 …… (158)
第一节　Poisson 回归模型的参数估计和拟合优度检验 …… (158)
第二节　Poisson 回归模型的相对危险度估计 …… (159)
第三节　Poisson 回归模型分析在医学中的应用 …… (159)
习题八 …… (164)
第九章　对数线性模型分析 …… (165)
第一节　对数线性模型的基本概念 …… (165)
第二节　对数线性模型 …… (167)
第三节　对数线性模型分析的方法步骤 …… (171)
第四节　对数线性模型的选择 …… (175)
第五节　对数线性模型在医学中的应用 …… (175)
习题九 …… (180)
第十章　生存分析 …… (183)
第一节　生存分析的基本概念 …… (183)
第二节　非参数生存分析法 …… (187)
第三节　参数生存分析法 …… (198)

第四节　Cox 比例风险回归模型……（204）
习题十……（212）
第十一章　主成分分析……（215）
第一节　主成分分析的基本原理……（215）
第二节　主成分分析的数学模型……（217）
第三节　主成分分析的方法步骤……（217）
第四节　主成分分析在医学中的应用……（219）
习题十一……（229）
第十二章　聚类分析……（231）
第一节　聚类分析的基本思想……（231）
第二节　聚类分析的统计量……（232）
第三节　聚类分析的方法……（235）
第四节　聚类分析在医学中的应用实例……（237）
习题十二……（250）
第十三章　判别分析……（252）
第一节　判别分析的基本思想……（252）
第二节　Fisher 判别分析法……（253）
第三节　Bayes 判别分析法……（255）
第四节　判别分析在医学中的应用……（260）
习题十三……（274）
第十四章　典型相关分析……（276）
第一节　典型相关分析的基本思想……（276）
第二节　典型相关分析的方法……（277）
第三节　典型相关分析在医学中的应用……（281）
习题十四……（287）
第十五章　路径分析……（289）
第一节　多变量线性回归分析……（289）
第二节　路径分析的数学模型……（291）
第三节　路径分析模型的基本要素……（293）
第四节　路径分析的方法步骤……（294）
第五节　路径分析模型的可鉴别性和自由度……（298）
第六节　直接影响、间接影响和总体影响……（299）
第七节　路径分析在医学中的应用……（301）
习题十五……（308）

第十六章　探索性因子分析 …… (309)
第一节　探索性因子分析和确定性因子分析的区别 …… (309)
第二节　探索性因子分析的基本原理 …… (310)
第三节　探索性因子分析的数学模型 …… (312)
第四节　探索性因子分析的方法步骤 …… (313)
第五节　探索性因子分析在医学中的应用 …… (316)
习题十六 …… (321)
第十七章　确定性因子分析 …… (323)
第一节　确定性因子分析的基本原理 …… (323)
第二节　确定性因子分析的数学模型 …… (325)
第三节　确定性因子分析模型的基本要素 …… (326)
第四节　潜在因子的尺度问题 …… (327)
第五节　确定性因子分析模型的可鉴别性和自由度 …… (328)
第六节　样本导出的与模型隐含的方差协方差矩阵 …… (328)
第七节　确定性因子分析的方法步骤 …… (329)
第八节　确定性因子分析在医学中的应用 …… (331)
习题十七 …… (338)
第十八章　结构方程模型分析 …… (339)
第一节　结构方程模型 …… (339)
第二节　结构方程模型的基本要素 …… (342)
第三节　结构方程模型的可鉴别性和自由度 …… (343)
第四节　直接影响、间接影响和总体影响 …… (344)
第五节　样本导出的与模型隐含的方差协方差矩阵 …… (345)
第六节　结构方程模型分析的方法步骤 …… (347)
第七节　结构方程模型分析在医学中的应用 …… (354)
习题十八 …… (370)
附录一　SAS 过程步使用说明 …… (371)
(一)SAS 过程步常用语句和选择项 …… (371)
(二)主要 SAS 过程步使用说明 …… (372)
附录二　主要参考文献 …… (398)

第一章 绪 论

第一节 多元统计学在医学中的地位

我们知道，统计学是概率论和数理统计的一个应用学科，它的主要目的是通过对随机现象的研究来推测总体的特性或判断间的内在联系规律。它的理论基础已经研究了几百年，但由于受到计算工具的限制，以前只能进行较简单的统计分析，如借助计算器计算均值、方差、标准误、检验总体参数、比较均值，进行单因子方差分析、简单相关分析、简单线性回归分析等基本统计分析。对于应用价值很高的深一层次的多元统计分析，如多因素的方差分析、多变量方差分析、多元线性回归分析、多元相关分析、多元 Logistic 回归分析、对数线性模型分析、生存分析、判别分析、聚类分析、主成分分析、因子分析、典型相关分析、时间序列分析、路径分析以及结构方程模型分析等等，由于它们涉及到的公式很复杂，计算起来很困难，若仅依靠初等的计算工具，几乎是行不通的。幸运的是，近几十年来计算机得到了快速发展，它从根本上解决了多元统计分析中繁杂的计算问题，从而使得多元统计分析这一领域里丰富的理论知识有了真正的应用价值。特别是近二十年来，随着计算机的不断普及，随着统计软件的不断更新换代，多元统计分析方法越来越容易地被人们所接受和掌握，人们的工作效率也随之提高了数百倍数千倍甚至数万倍以上。

医学统计学是统计学的一个应用分支，它的主要任务是解决医学科学研究中出现的各种问题，为更准确地鉴别和诊断各种疾病，更积极有效地预防各种疾病提供科学的理论依据。例如，比较几种不同降血压药物的疗效、分析诸多个危险因素中哪些因素对肺癌的发生起主导作用、研究遗传基因和环境因素对心血管疾病的直接影响和间接影响、根据已掌握的确凿数据建立胃癌术后的预后模型等等。随着多元统计学的发展，医学多元统计学也越来越显示出它在整个医学科学研究中的重要地位，它不仅是医学科学研究中不可缺少的一个重要工具，而且它是促进医学科学发展的一门重要相关学科。

第二节 结构方程模型分析简介

本教材的后四个章节介绍了很实用的医用多元统计分析方法之一：结构方程模型分析。这方面内容对于很多医学院校的学生都很陌生。为了帮助读者提高对结构

方程模型分析的兴趣，这里着重介绍一下它的特点和使用价值。

结构方程模型(Structural Equation Modeling，SEM)分析是一种用来分析多个指标变量之间错综复杂关系结构的多元统计分析方法。近二十年来，SEM 已经作为一个强有力的统计分析工具在各科学研究领域里得到广泛应用。虽然它的历史可以追溯到 20 世纪前半叶，当时 Spearman(1904 年)提出了因子分析(factor analysis)和 Wright (1934 年)引进了路径分析(path analysis)，但是直到 20 世纪 70 年代，这种方法才被 Karl Joreskog 等人完善并逐步开始应用到行为科学等研究领域中来。今天随着计算机技术的快速发展，特别是随着有效的 SEM 计算机程序软件的开发和提高，SEM 已成为一种应用价值很高的多元统计分析方法，并和其它传统的各种多元统计分析方法一样，它的重要性和实用性正被越来越多的人所认识和接受。医学科学以人体为研究对象，从生命的本质到疾病的病因、诊断、治疗以及预防、健康保健等广泛的领域进行探索和研究。作为高级动物的人的复杂性，决定了医学科学研究的复杂性，它不仅仅只限于医学领域本身，而且涉及到生物学、心理学、社会学以及行为科学等范畴。也就是说，任何一个与人的健康或疾病有关的问题，总是与多方面的因素相关联的。例如，一种传染疾病的发生，不仅取决于病源，而且还与个人的基本身体素质、生活卫生习惯、居住环境条件以及家庭遗传基因等等多种因素有关。要认识这些因素之间的相互关联关系，就必须运用多元统计分析方法，而 SEM 可以说是各种多元统计分析方法中更为强有力的一个工具。

在诸多种分析多个变量之间相互依存关系的多元统计分析方法中，典型相关分析可以判别一组变量和另一组变量之间是否关联，但不能确定其因果关系(causal relationship)；各种多元回归分析可以用来确定一组变量对一个变量的直接影响关系，但不能确定其间接影响以及相互因果关系；路径分析可以确定变量之间的直接影响关系以及间接影响关系，但不能确定它们之间的相互因果关系；结构方程模型分析能够确定多个变量之间相互错综复杂的因果关系，包括直接关系和间接关系。特别是结构方程模型分析允许潜在变量(latent variable)存在，并允许可测变量(measured variable)的度量误差存在。这种方法解决了一般回归分析所不能处理的许多难题。例如，允许潜在变量的存在，一方面可以解决某些不易直接测量的理论变量的构造问题，如人的智商、家族的遗传基因等；另一方面可以解决模型中含可测变量过多而导致结论偏差的问题；同时还可以解决自变量之间的内在多重相关性问题。允许度量误差存在的处理方法使得分析结果更接近客观真实，控制或减少人为的误差。另外，在数学方法上，它采用逼近法估计的方差和协方差矩阵使得一般回归分析方法所要求的正态分布数据的假定得以放松。上述这些技巧性的数据处理方法使得 SEM 更具特色，成了很多科研工作者，特别是医学科学研究者感兴趣的学习内容。

得益于计算机统计软件的不断发展和完善，SEM 的实用性越来越强。在国际上，很多社会学、医学、教育学、行为科学等领域的科学工作者正在广泛将之运用于工作

实践当中。目前国际上流行的用于SEM分析的软件有LISREL、EQS和SAS中的CALIS等。由于SEM涉及到协方差矩阵的计算等较深较复杂的数学知识，所以很多医学院校学生感到畏惧，其实大可不必。理解和掌握SEM的理论知识需要较高的数学知识，但学习和掌握SEM的使用方法仅要求有一定的数学和统计基础知识以及一定的计算机软件应用技能就可以了。当然，也不能过于轻视学习SEM，它毕竟不同于回归分析。一般来说，加强对其基本原理的理解是掌握SEM的捷径。如果对基本原理理解不深，可能会造成数据处理不当，从而出现不正确的结论或误导。

目前国内医科院校对多元统计分析方法的理论教学和实践应用正在加强，但对SEM的认识还很薄弱，这部分内容的参考资料也很少，所以在本教材中较详细地介绍SEM的分析原理和方法是很有必要的，特别是对培养我国21世纪高级医科研究人员有着重要的意义。本教材从第十五章起开始进入SEM的学习，它以介绍SEM的基本原理和基本方法为主，对高深的数学理论推导不作要求。通过本课程的学习，初步了解使用SEM的意义，掌握建立SEM的方法以及学习有关应用方面的技巧等。

第三节 多元统计分析软件的介绍

目前，国际上用于多元统计分析的软件很多，其功能各有差异。除了SAS以外，像BMDP，SPSS，STATS，GENSTA，EQS，LISREL等，都是很流行的统计软件。其中，BMDP、SPSS、STATS和GENSTA实用于各种统计分析，EQS和LISREL实用于进行结构方程模型分析。这些软件和SAS软件比较起来，显得欠缺一些。它们最主要的弱点是：第一，建立数据库的功能较差，致使应用起来不如SAS灵活方便；第二，没有计算机母语的使用，使得分析工作很被动，很死板；第三，包括的分析程序不如SAS广泛，精度和深度也不如SAS。目前被世界各国公众，包括国家的政府部门、大专院校、科研单位等看好的专业统计软件就是SAS软件。对于多元统计分析，SAS软件可以说是不可多得的高级计算机专业软件。因此，本教材选用SAS统计软件作为主要计算工具，这样可以促使学生更具国际竞争力。

值得一提的是，如果对路径分析和结构方程模型分析感兴趣的读者，还可以学习使用LISREL软件，它是专业的结构方程模型分析软件，比SAS的CALIS过程步更具灵活性，且输出结果整齐清晰。

第四节 医学统计学的几个重要概念

1. 总体与样本

总体(population)就是被研究对象的全体；样本(sample)就是从总体中抽取的部

分对象。统计分析的原理是由样本推断总体的特性，所以样本应从总体中随机抽取，使之具有代表性。

2. 参数与统计量

由总体中的全部个体表示出来的函数值称为参数(parameter)，它用来描述总体的某个特征。参数一般用希腊字母表示，如总体均值 μ，总体标准差 σ，总体率 π。由样本表示出来的函数值称为统计量(statistic)，它用来估计总体参数或检验总体某种特性的准确性。

3. 变量的分类

统计分析研究的目标是变量。所谓变量指的是某个指标的观察值。根据观察值数据的类型和统计处理的方法，变量可以分为两类：

(1)数值型变量(numerical variable)，也称为连续型变量(continuous)或区间变量(interval)。如身高、体重、舒张压等，其取值允许整数、小数和分数。数值型变量的本质特性是它具有原点和单位。数值型变量给出的信息在统计学上也称为是定量资料(quantitative data)。

(2) 分类型变量(categorical variable)，包括顺序变量(ordinal)和名义变量(nominal)。例如，伤残等级(1、2、3、4)，医院等级(1＝甲、2＝乙、3＝丙、4＝丁)等都属于顺序变量。如职业(1＝工人、2＝农民、3＝学生、4＝干部、5＝其它)，种族(1＝汉、2＝回、3＝其它)等就属于名义变量。顺序变量的数值要用数字表示，且数字具有大小顺序的意义。名义变量的数值可以用字符表示，也可以用数字表示，但都不具有任何大小顺序的意义。不论顺序型分类变量和名义型分类变量用何种数值来表示，它们与数值型变量的根本区别是不具有原点和单位。分类型变量给出的信息在统计学上也称为是定性资料(qualitative data)。

在医学中经常遇到另一类型的数据，它们的取值为整数，称为计数资料。这类资料的处理方法是：如果计数资料的取值范围很大(如白血球数、患病人数)，则将其作为数值型变量处理；如果计数资料的取值范围很小(如每周看医生的次数)，则将其作为分类型变量的顺序变量处理。SAS 处理的数据仅为上述两种类型的数据。

4. 统计分析的任务

统计分析的任务包括数据的描述性分析(description analysis)和数据的统计推断分析(statistical inference analysis)。数据的描述性分析指的是对样本观察值的集中趋势和分散程度等给出直接的数字度量结果，对其准确度不作评估。数据的统计推断分析指的是利用样本观察值对总体的某种特性加以研究，并根据统计学基本原理对其准确度作出尽可能可靠的评估和结论。概括地说，统计推断分析所能解决的主要问题有：

* 总体参数的假设检验；
* 两个或多个总体参数的比较；

* 两个或多个变量之间关联程度的分析；
* 两个或多个变量之间相互依存关系的分析；
* 变量的分类分析；
* 时间序列分析。

5. 统计推断分析的分类

从对数据的要求来说，统计推断分析方法可分为参数分析法(parameter analysis)和非参数分析法(non-parameter analysis)。参数分析法要求数据的分布或渐进分布函数是已知的，且大多数参数分析要求数据服从正态分布。例如，t—检验、方差分析、线性回归分析等都属于参数分析法。非参数分析法对数据的分布没有任何要求，例如，符号检验法、符号秩和检验法、logistic 回归分析等都属于非参数分析法。

由于非参数法分析数据不需考虑数据的分布，因而比较简单。但是对于连续型数值变量，使用这种方法仅利用了数据中很少一部分信息，一般情况下最好避免使用，或者说它通常是在某些参数分析法失效的情况下使用，如数据转换后仍不满足假设条件要求。另外，还有几种分析方法，如 χ^2-检验、Cox 比例风险回归模型分析等对数据的分布没有严格要求，但却能够对参数进行检验，因此有时称之为半非参数分析法。

统计分析是一门分析数据资料的严谨科学，必须遵循它自身的规律才能得到较可靠的结论。具体地说就是，当用参数法分析数据时，一定要考虑数据应当满足的基本要求。如果与假设条件或要求相差甚远，则应当考虑对数据加以转换或用其它方法加以验证，而不能匆匆地分析数据和发表结论。

第二章　基本统计分析方法

在统计学中，把仅涉及到一个或两个变量的统计分析称为基本统计分析，如假设检验、一元方差分析、简单相关分析、简单线性回归分析等。基本统计分析有很广的应用价值，并且它是多元统计分析的理论基础。掌握好基本统计分析的基本理论和SAS软件的使用方法，对进一步学习多元统计有很大的帮助。本章简单地概括医学统计学中几个主要的基本统计分析方法及其SAS过程步的应用说明。

第一节　数据的描述性分析

数据的描述性分析指的是对变量的基本统计量(basic statistic)的分析，它是进行统计推断分析必不可少的先头工作。

一、数值型变量的描述性分析

对于数值型变量，它的基本统计量包括均值(mean)、方差(variance)、标准差(standard deviation)、标准误(standard error)、变异系数(coefficient of variation)、极值(extreme value)、百分位数(percentile)、峰度(kurtosis)、偏度(skewness)、众数(mode)、中位数(median)等等。这些统计量比较全面地概括了数据的集中趋势(central tendency)、离散程度(spread or variation)和分布形态(distribution pattern)。

SAS系统中用来对数值型变量进行描述性分析的过程步有univariate、means、summary、chart和tabulate等过程步，其中常用的是means过程步(附录一A)和univariate过程步(附录一B)。

【例2-1】 利用univariate过程步对表2-1数据中的变量weight作描述性分析。

表2-1　儿童发育指标调查数据

序号 (id)	性别 (sex)	年龄 (age)	身高 (height)	体重 (weight)	健康程度 (health)
1	1	3	0.94	13.59	a
2	1	4	1.02	15.44	b
3	2	7	1.16	19.93	c
4	1	6	1.12	18.09	a
5	1	6	1.14	18.80	a
6	2	7	1.15	18.20	b
7	1	4	1.06	15.30	b