普通高等教育“十一五”国家级规划教材

主编　喻荣彬

医学研究的数据管理与分析

第2版

人民卫生出版社

普通高等教育“十一五”国家级规划教材

医学研究的数据管理与分析

第 2 版

主　编　喻荣彬（南京医科大学）

副主编　邱洪斌（佳木斯大学）

主　审　汪　宁（中国疾病预防控制中心）

编　委　（以姓氏笔画为序）

王　蓓（东南大学）　　郝元涛（中山大学）

许　锬（苏州大学）　　郝加虎（安徽医科大学）

李淑珍（山西医科大学）　　姚应水（皖南医学院）

张铁军（复旦大学）　　姚振江（广东药学院）

沈　冲（南京医科大学）　　寇长贵（吉林大学）

邱洪斌（佳木斯大学）　　喻荣彬（南京医科大学）

赵　杨（南京医科大学）　　路　滠（南京医科大学）

赵景波（哈尔滨医科大学）　　潘发明（安徽医科大学）

秘　书　路　滠（南京医科大学）

人民卫生出版社

图书在版编目（CIP）数据

医学研究的数据管理与分析/喻荣彬主编. —2 版. —北京：人民卫生出版社，2009.2

ISBN 978-7-117-11169-0

Ⅰ.医… Ⅱ.喻… Ⅲ.医学-数据管理-应用软件-医学院校-教材 Ⅳ.R319

中国版本图书馆 CIP 数据核字（2008）第 213272 号

医学研究的数据管理与分析

第 2 版

主　　编：喻荣彬

出版发行：人民卫生出版社（中继线 010-67616688）

地　　址：北京市丰台区方庄芳群园 3 区 3 号楼

邮　　编：100078

网　　址：http://www.pmph.com

E-mail：pmph@pmph.com

购书热线：010-67605754　010-65264830

印　　刷：三河市富华印刷包装有限公司

经　　销：新华书店

开　　本：787×1092　1/16　　印张：21.25

字　　数：513 千字

版　　次：2003 年 12 月第 1 版　2009 年 2 月第 2 版第 2 次印刷

标准书号：ISBN 978-7-117-11169-0/R·11170

定　　价：38.00 元

前　言

数据管理和统计分析是医学研究工作中的重要步骤，医学研究实施过程可被认为是一个研究设计、数据收集、整理、统计分析和结果解释的连续过程。通常，绝大部分医学研究的数据管理和分析是在计算机上通过运行相应的软件来实现的。目前，国内医学统计学教材多侧重于统计学基本理论，或辅以某个统计分析软件（如 SAS、SPSS、STAT 等）介绍其实际应用，没有真正地把研究设计、数据收集、录入、整理、统计分析和结果解释等作为一个连续的过程来阐述，往往在不同程度上存在理论教学和实际应用脱节。作为普通高等教育国家“十一五”规划教材，本书在教学内容和形式上进行了创新性的探索和尝试，力求有所突破。

本教材旨在将医学统计学、流行病学和计算机软件等相关课程知识有机结合，在论述研究设计、数据资料收集和数据库等相关知识的基础上，介绍目前国际上常用的几种数据管理和统计分析软件的应用，将医学研究的设计、数据收集、录入、整理、统计分析和结果解释等作为一个连续的过程，理论和实际应用紧密结合，系统、全面地进行介绍，有利于综合性地提高学生医学研究设计、实施和评价的能力，尤其是数据处理和统计学分析应用能力。

全书共分十六章。第一章至第六章为总论部分。第一章简要介绍医学研究实施过程中数据管理的主要内容及其和研究设计的关系，重点阐述医学研究数据类型及其统计分析方法选择，以及数据管理、统计分析和结果解释中应遵循的原则。第二章概括介绍现场调查技术，阐述如何通过严格的调查设计和实施获取研究数据，内容包括现场调查设计、质量控制和敏感问题调查技术。调查表是医学研究数据收集的主要工具，第三章系统介绍调查表设计知识，包括调查表的结构和内容、设计基本原则、修改完善和评价等，并提供传染病、慢性非传染病和行为流行病学调查表示例，对医学研究常用的几种量表也作了简单介绍。第四章全面介绍常用流行病学研究设计和实验设计的统计分析思路、方法和指标选择、结果解释和注意事项等。第五章在简介有关数据库知识的基础上，介绍几种常用数据库软件应用、数据文件特点和转换，并通过 StatTransfer 软件系统介绍如何进行数据转换。第六章系统介绍数据处理及其质量控制知识，包括数据的逻辑检查和核对、数据编码和赋值以及缺失值处理等。第七章介绍 EpiData 3.1 软件的应用，重点讲述如何通过计算机录入调查表数据、建立数据文件、进行数据管理、数据转换以及质量控制等。第八章至第十一章详细介绍 SPSS 13.0 软件的应用，重点为数据文件的建立、管理和常用的统计分析过程，包括描述性统计分析、均数比较、无序分类资料的统计分析、非参数检验、重复测量数据的统计分析，相关分析、回归分析（线性回归、曲线拟合、二分类、多分类和多项有序分类变量 Logistic 回归、配对 Logistic 回归、剂量－反应关系分析和非线性回归等），以及生存分析、聚类分析、判别分析、主成分分析和因子分析等。第十二章介绍通过 SPSS 13.0 和 Excel 2003 软件绘制常用的统计图形，重点介绍统计图的修饰和完善技巧。第十三章讲述 EpiCalc2000 软件的应用，介绍医学研究表格数据的统计分析、样本含量估计和随机数字表产生等内容。第十四章在初步介绍循证医学和 Meta 分析知识的基础

上，重点介绍如何通过 Review Manager（RevMan）4.2.8 软件制作、保存 Cochrane 系统评价计划书和论文，并对录入数据进行 Meta 分析。第十五章较详细地介绍 WHO 推荐使用的 Epi Info 3.4.3 软件的应用知识，包括数据录入、常用数据统计分析方法，并介绍通过 Epi Map 软件制作统计地图。第十六章简介 SAS 9.1.3 软件基本知识，重点介绍 SAS 软件的交互式应用。

有关统计学分析的基本原理，书中未做详细介绍。学习本书需要具备初步的医学统计学基础。书中所涉及的统计学分析原理，请参见相关专业书籍。

本书可作为医学院校各专业的本科生、七年制或八年制学生和研究生的必修课或选修课教材，也可供临床医生、卫生防疫医生、卫生管理人员以及其他相关人员使用。

在申报普通高等教育国家“十一五”规划教材和编写过程中，得到了人民卫生出版社、南京医科大学和公共卫生学院的大力支持。中国疾病预防控制中心性病艾滋病预防控制中心汪宁教授审阅了书稿，并提出了许多宝贵的修改意见和建议。

由于编者水平有限，书中难免会存在一些不尽如人意甚至错误之处，诚恳希望读者们批评指正。

喻荣彬　邱洪斌

2008 年 9 月 16 日

目　录

第一章　绪　论

科学研究是通过实验或观察取得信息，并对信息进行处理、分析的过程，目的是为了发现、分析和解决问题。医学科学研究是研究人体正常生理、病理、健康和疾病的科学，主要任务是揭示人体生命本质与疾病发生、发展的现象和机制，认识人和环境的相互关系、健康与疾病相互转化的客观规律，为防治疾病、促进健康提供技术方法和手段。

医学研究的基本程序包括选题（立题）、设计、观察和实验、资料整理和数据统计分析及理性概括等。其中，数据管理和分析贯穿整个医学研究过程。科学合理地进行数据管理和统计分析，对医学研究的顺利实施至关重要。

第一节　研究设计与数据管理分析

研究设计是医学研究工作的起始步骤，也是最重要的环节。首先要提出研究设想，确定要回答或解决的问题，明确研究目的；其次，要根据研究目的确定相应的分析指标，通过调查和实验，收集研究数据。

研究的实施步骤则和设计思路相反。在严格设计的基础上，首先进行观察和实验，收集相关的数据资料，并对数据进行整理和统计分析，结合归纳、演绎和推理，最终验证所提出的假设或回答所要解决的问题。

数据管理和统计分析贯穿于医学研究设计和实施的整个过程中。在研究设计时即应明确所要收集的数据类型、测量方法、统计分析方法和指标。实际上，医学研究实施过程也可以被认为是数据收集（data collection）、数据整理（data processing）、统计分析（statistical analysis）和结果解释（interpretation）的过程。

医学研究设计与数据管理分析步骤的关系如图 1-1-1 所示：

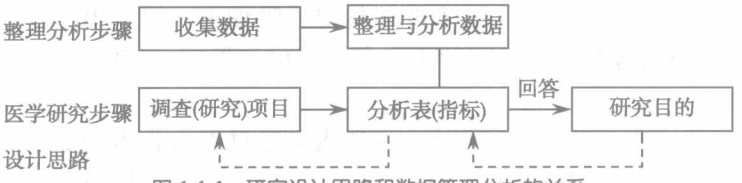

图 1-1-1　研究设计思路和数据管理分析的关系

一、明确研究目的

各项医学研究的目的可能不同，但从统计学角度来说，无非解决下列两个问题：

（1）了解参数（parameter），用以说明总体。如通过抽样调查了解某地学龄儿童的身高和体重水平。

（2）研究变量间的关系，通过确立统计学联系来验证因果联系，如探索暴露和疾病之间的因果关系、评价某种预防措施或药物的防治效果等。

研究目的需要通过具体的指标来阐明。确定研究目的是选定研究指标的依据，而研究指标又是研究目的的具体体现。

二、确定研究对象和观察单位

根据研究目的和指标，确定研究对象和观察单位（observation unit），即划清研究总体的同质范围。如评价某药物对高脂血症的治疗效果，研究总体应包含所有患高脂血症的个体，研究对象应是该总体的一个代表性样本，观察单位则是该样本中每一个患高脂血症的个体。

三、医学研究设计类型的选择

基础医学研究通常涉及各种不同类型的实验设计，如完全随机设计（completely random design，又称单因素设计）、配对设计（paired design）、随机区组设计（randomized block design，也称配伍组设计）、析因试验设计（factorial design）、拉丁方设计（Latin square design）、正交设计（orthogonal design）和序贯设计（sequential design）等。

除了上述实验设计类型外，临床医学和预防医学研究的对象往往是群体（某病患者、某个人群），需借助流行病学研究方法。流行病学研究设计类型包括描述性研究（普查和抽样调查、生态学研究、个案调查和病例分析等）、分析性研究（病例对照研究、队列研究）和实验性研究（临床试验、现场试验和社区干预试验）等。设计类型的选择主要取决于研究目的和客观条件的限制。如要评价某药物的疗效，可采用临床试验研究设计；要探索某罕见病的危险因素，可采用病例对照研究；要了解某病预后的影响因素，则可采用队列研究。

不同类型的研究，数据管理和统计分析的方法和指标选择有所不同，应掌握每种具体方法的应用条件，科学合理地选用。

四、确定研究项目，拟定调查表或原始数据记录表

1. 研究项目的确定　主要取决于研究目的和分析手段。例如，流行病学研究项目通常包括一般情况（如姓名、性别、出生日期、出生地、民族、文化程度、职业等）、研究项目（也称研究变量，包括疾病史、家族遗传史、吸烟史、饮酒史、饮食习惯、体力活动、月经生育史、职业暴露史、体格检查和实验室检测等）。而基础医学研究的变量相对较少、较明确。

2. 调查表和原始数据记录表的设计　调查表是通过把拟收集的数据项目用恰当的措辞构成一系列问题的“答卷”，也称“问卷（questionnaire）”。调查表是调查研究资料收集的最主要工具。调查表如何设计取决于研究目的和分析手段的需要，关键在于保证所获得信息具有全面性、针对性、准确性和可靠性。

五、样本含量的估计

样本含量（sample size）的估计是医学研究设计的一个重要内容。基础医学研究一般采用动物实验，研究条件易于标准化，样本含量相对容易确定。临床和预防医学研究对象通常为人群样本，影响研究结果的因素多而复杂，研究变量变异较大，样本含量的估计更为重要。样本含量的大小至少满足“统计学效率”。样本含量大小主要取决于研究单位的变异大小、两组或多组可能差异的大小、精确性（容许误差）的要求、第一类错误（α）

和第二类错误（β）的设定。不同研究设计可用各自样本含量计算公式来估计，也可采用专门的软件（如 Epi Info、EpiCalc 等）来估算。

六、原始资料的收集

原始资料（raw data）的来源包括常规报表、实验数据和现场调查资料等。收集方式包括直接观察法（体格检查、各种实验室检测分析等）和采访法（访问、调查会及信访、电话访问等）。资料的收集是整个研究工作的中间环节，所收集资料的质量将直接影响研究结果的真实性与结论的正确性。收集原始数据时应严格质量控制措施，避免或减少信息偏倚。

七、医学研究的质量控制

质量控制是决定研究结果科学性的关键。调查研究所获取的数据只有准确地反映客观现实，通过归纳、比较、推理所获的结果才具有科学性，否则就会产生系统误差（systematic error），即偏倚（bias）。偏倚包括选择偏倚（selection bias）、信息偏倚（information bias）和混杂偏倚（confounding bias）三类。医学研究的质量控制就是要控制这三类偏倚对结果科学性的影响。只有通过严格的质量控制措施，才可以保证所获研究资料的准确性、可靠性和完整性。

八、数据资料的管理

数据资料的管理包括录入计算机前的核对、录入时的质量控制和录入后的核对、分组、编码等。

1. 录入前的核对　在调查研究开始时，应采取措施保证原始数据的准确性。通常通过规范的质量控制措施来避免或减少调查研究中的信息偏倚。录入前的核对包括检查调查表中有无漏项、填写错误、及时纠正等内容，录入前核对和纠错有利于数据录入。

2. 录入计算机，建立数据库　可以通过统计分析软件或数据库管理软件录入调查表信息，建立数据库（database）。常用的软件有 Epi Info、EpiData、SPSS、FoxBase、FoxPro、Visual FoxPro（VFP）、Access、Excel 和 Lotus 等。录入软件的选择，取决于数据量的大小（包括记录数、变量数）和对录入效率的要求等。记录数和变量数较大时，建议采用 EpiData 或 Epi Info 软件录入数据。

3. 录入后处理　主要包括逻辑核对、编码、新变量的建立和变量转换等。

（1）逻辑核对（logic checking）：在数据库或统计分析软件中通过排序（sorting）等方法查看极大值或极小值，再重新核对某些极端值，以决定取舍或修正。

（2）数据的编码（coding）和转换（transforming）：有时需要根据连续性资料的值来对个体进行分类，如根据血压值判定是否为高血压患者，或根据既往有无糖尿病病史或口服糖耐量试验（OGTT）血糖值综合判定其是否为糖尿病患者或分型，则需要重新编码。

（3）建立新变量：将数据编码和转换的结果赋值于新变量。如新建立“DM”变量，“1”表示糖尿病患者，“0”表示非糖尿病患者；又如建立“BMI”变量表示体质指数（BMI），根据体重“weight”和身高“height”两个变量值，利用公式对“BMI”赋值。

（4）分类变量转换成哑变量（dummy variable）：对于名义数据（如血型、性格类型等），因为各类别间并不呈等级关系，在进行多因素分析（如多元回归分析、Logistic 回归分析，Cox 回归分析）时，不能使用原始数值，必须进行变量变换，将该变量转换成 $n-1$（水平数－1）个哑变量，再将这些变量纳入多因素模型中。

第二节　数据类型与统计分析方法选择

一、医学研究的数据类型

整理和统计分析资料时一般先区分数据的类型。医学研究的研究数据大体上可分为三种类型：定量数据（quantitative data）、等级数据（ranked data）和名义数据（nominal data）。

1. 定量数据　用定量的方法测量每个观察单位的某项（或几项）指标，所得的数据资料称为定量数据，也称计量资料（measurement data）。每个观察个体的某项指标记以一个数值，如身高、血压、白细胞数等。定量数据又可分为两类：一类是离散型（discrete data）或间断型数据（discontinuous data），它们往往是一种计数，如每名儿童口腔中的龋齿数、一个路段一年内的车祸次数、一个显微镜视野下的阳性细胞数等，这种计数只能是 0 和正整数，不会是负数，也没有小数点；另一类是连续型数据（continuous data），理论上在任何两个数值之间都还有无穷多个数据，如身高在 165.5cm 和 165.6cm 之间理论上存在着无穷多个数据。

2. 等级数据　将观察单位按某种属性的不同程度分组，所得的各组观察单位数为等级资料，又称有序分类数据（ordinal data）或半定量数据（semi-quantitative data）。这类数据一般无单位，但组与组之间有大小之分，或程度差别，而组内不分大小。例如临床疗效痊愈、显效、好转和无效。

3. 名义数据　各类数据之间没有顺序或等级关系。如白细胞分类的中性粒细胞、淋巴细胞、嗜酸性粒细胞、嗜碱性粒细胞等，男性和女性。

等级数据和名义数据也被称为定性数据（qualitative data）、属性资料（attribute data）或计数资料（enumerative data 或 count data），资料中每一观察指标是以其性质为特点的，如血型、性格类型、发病与否、病情轻重等。对计数资料作整理，主要就是清点各种属性的个数，有时还需要对属性本身作归类。有些等级数据或名义数据可以分成两类。如生、死，男、女，阳性、阴性，有效、无效，暴露、不暴露，发病、未发病，患病、未患病等，属两分类数据（dichotomic data）。等级和名义数据也可以是多分类的。

不同类型的数据，在选择研究方法时有所不同。实际工作中，根据统计分析的需要，对这三类数据可进行适当的变换或重新编码（recoding）。

二、数据类型的转换

根据研究目的和统计分析的需要，定量数据和定性数据可以互相转化。例如，血压值为定量数据，但如果将一组 20～40 岁成年人的血压按诊断标准分“正常”与“异常”两组，再统计各组人数，于是血压这一定量数据就转化成为定性数据了。又如年龄资料为定量数据，但可以按 10 岁为一年龄组，将人群年龄分为＜10、10～、20～、30～、40～、

50～、60～、70～等8个年龄组，这样定量数据便转换为等级数据。又如诊断试验中，将某些阳性体征根据确诊患者的概率赋予分数，分数的多少代表确诊概率的大小，这样原来的定性数据就转化为定量数据。

在数据转换过程中，值得注意的是：①定量数据转换为定性数据一般比较简单，但从名义数据、等级数据转换为定量数据，则比较繁琐且损失数据信息。因此，在医学研究中收集数据或计算机储存数据时，应考虑收集定量数据，只有在数据处理时根据需要再转换为等级数据或定性数据。②对两组或多组研究单项的某项指标进行统计学检验时，数据从定量转换为定性或等级数据时，统计学的效率会下降。

定量数据转换为定性或等级数据时，常用的分组切割值（cut-point value）选择方法有：①以正常参考值或临床诊断标准作为分组依据。如空腹血浆血糖值根据临床诊断标准：＜6.1mmol/L（110mg/dL）为“正常血糖＝0”；≥6.1mmol/L（110mg/dL）及＜7.0mmol/L（126mg/dL）为“糖耐量低减（IGT）＝1”；≥7.0mmol/L（126mg/dL）为“糖尿病＝2”。②某些定量指标尚无公认的正常参考值，可根据均数或四分位间距值，将其分为两组或四组。③根据数据的分布特点和研究需要，自行确定，但要能对统计分析结果做出合理的解释。

三、医学研究的数据统计分析方法和指标

医学研究中，首先应考虑研究目的和研究设计，再根据资料的类型和资料的分布情况选择合适的统计分析方法进行数据分析。医学研究数据的统计分析，可以利用统计软件在微机上进行。常用的统计分析软件有SAS、SPSS、STAT、Epi Info、EpiCalc等，其中SAS和SPSS适应于数据库数据的统计分析，EpiCalc适用于表格数据的统计分析，Epi Info则两者均可。统计分析包括统计描述和统计推断。

（一）统计描述

1. 定量数据的描述　定量数据的描述指标包括均数（或几何均数）±标准差（$\bar{x}\pm s$）、中位数（median）、百分位数（percentile）、变异系数（coefficient of variation，*CV*）、极差（range）以及偏度系数（coefficient of skewness）、峰度系数（coefficient of kurtosis）和总体95%可信区间（confidence interval，*CI*）。定量数据的统计描述方法见图1-2-1。

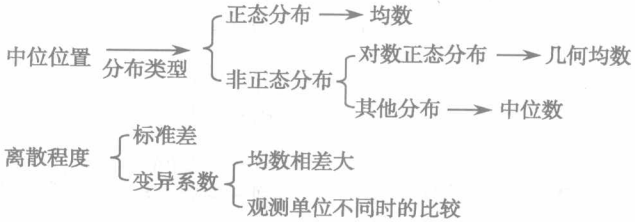

图1-2-1　定量数据统计描述方法小结

2. 定性数据的统计描述　定性数据可通过计算各种相对指标来描述，包括率（rate）、比值（ratio）或构成比（proportion）。如发病率、病死率、N年生存率、治愈率、缓解率、相对危险度（relative risk，*RR*）、比值比（odds ratio，*OR*）、标化死亡比（*SMR*）等。应用过程中，应注意率和比的区别。定性数据的统计描述方法见图1-2-2。

图 1-2-2　定性数据统计描述方法小结

（二）统计推断

1. 假设检验　包括定量数据分布类型的假设检验——正态性检验；定量数据方差的假设检验——方差分析（analysis of variances，ANOVA），包括成组设计多个样本均数的比较、配伍组设计多个样本均数比较、多个样本均数的两两比较、多个实验组和一个对照组均数间的两两比较等；定量数据均数的假设检验——t 检验和 ANOVA；定性数据分布情况或位置的假设检验（χ^2 检验）等。定量数据差别的假设检验方法见图 1-2-3。

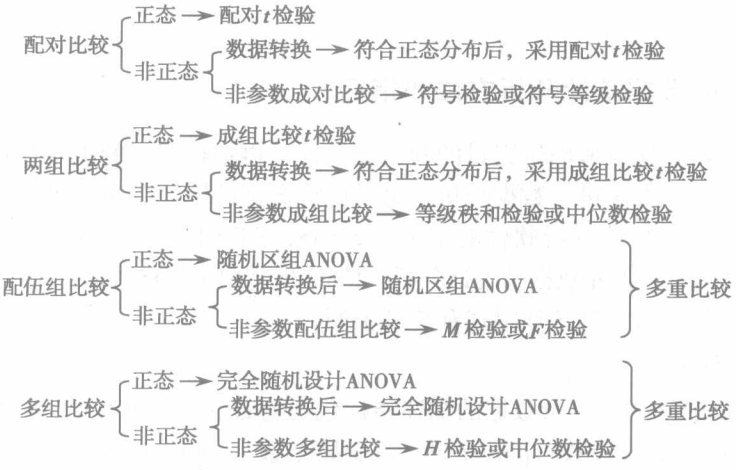

图 1-2-3　定量数据差别的统计意义检验小结

2. 变量之间的关系分析　包括定量数据相关分析（以直线相关为例，用于双变量正态分布资料）、回归分析（包括直线回归、多元线性回归、Logistic 回归和 Cox 回归等分析）和定性数据（R×C 表数据）的关系分析，见图 1-2-4 和图 1-2-5。

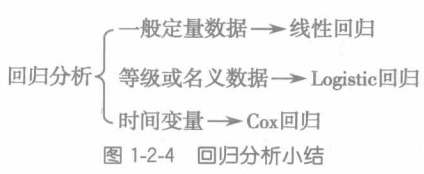

图 1-2-4　回归分析小结

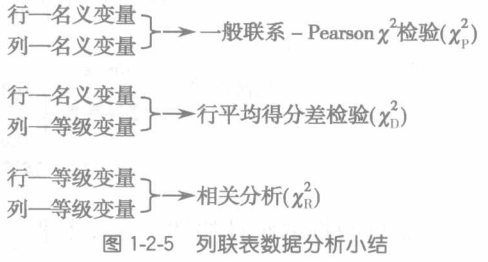

图 1-2-5　列联表数据分析小结

（三）统计分析方法汇总

按应变量和自变量性质归类，相应的统计分析方法见表 1-2-1。

表 1-2-1 统计分析方法汇总表

应变量个数	自变量性质	应变量性质	采用的统计分析方法
1	无自变量（1 个总体）	连续且正态	单样本 t 检验
		有序或连续	单样本中位数检验
		二分类	二项检验
		分类	拟合优度检验
	1 自变量 2 水平（组间独立）	连续且正态	两独立样本 t 检验
		有序或连续	Wilcoxon-Mann Whitney 检验
		分类	χ^2 检验
			Fisher 确切概率检验
	1 个自变量，2 个或以上水平（组间独立）	连续且正态	单因素 ANOVA
		有序或连续	H 检验（Kruskal Wallis 法）
		分类	χ^2 检验
	1 自变量 2 水平（组间相关/配对或配伍）	连续且正态	配对 t 检验
		有序或连续	Wilcoxon 符号秩和检验
		分类	McNemar χ^2 检验
	1 个自变量，2 个或以上水平（组间相关/配对或配伍）	连续且正态	单因素重复测量 ANOVA
		有序或连续	Friedman 检验
		分类	重复测量 Logistic 回归分析
	2 个或以上自变量（组间独立）	连续且正态	ANOVA
		有序或连续	秩变换后 ANOVA
		分类	Logistic 回归分析
	1 个连续性自变量	连续且正态	相关分析
			简单线性回归分析
		有序或连续	非参数相关/Logistic 回归分析
		分类	单因素 Logistic 回归分析
	1 或多个连续性自变量和（或）1 或多个分类自变量	连续且正态	多因素线性回归分析
			协方差分析
		分类	多因素 Logistic 回归分析
			判别分析
2 个或以上	1 个自变量，2 个或以上水平	连续且正态	One-way MANOVA
2 个或以上	2 个或以上自变量	连续且正态	多变量多重线性回归分析
2 组变量	0	连续且正态	典型相关分析
2 个或以上	0	连续且正态	因子分析

（James DL，2006）

第三节　数据管理和分析的原则

一、忠实于原始数据

忠实于原始数据是必须具备的科学精神。科学研究必须遵循客观现实，医学研究的本质即是通过观察、实验，描述或模拟疾病和健康状态的人群现象，或者通过实验动物模型模拟疾病和健康状态的发生、发展，经过科学的归纳、分析和逻辑推理得出普遍性的规律。只有客观地记录、复制原始数据，才能使所获结果接近真实的情况，才能再现出客观规律。通常，医学研究的结果和客观现实之间，总会存在或多或少的不一致，即误差（error），包括随机误差和系统误差。医学研究过程中，应尽量通过科学的设计和严格的质量控制措施，控制系统误差。任何篡改研究数据的行为，不管出于什么目的，都违背科学精神。

二、重视研究数据资料的处理过程

数据处理是统计分析前数据管理中必不可少的步骤，应给予足够的重视。数据资料处理的目的：一是保证被分析数据的正确性，和获得的客观结果尽可能保持一致，控制信息偏倚；二是使原始数据经过编码、转换、重新赋值后符合进一步统计分析的需要。

数据处理过程往往会花费研究者大量的时间，尤其在涉及较大规模的人群调查研究时。由于现今的统计分析软件大都具有较好的功能模块，一旦研究数据处理充分，统计分析过程就会大大简化。

三、选择合适的统计分析方法和指标

统计分析方法的选择主要取决于数据的类型，定量数据、定性数据的统计分析方法各不相同；同时，描述和统计分析方法的选择又取决于数据的分布类型，大多数统计分析方法要求符合正态分布或近似正态分布。

（一）数据转换

选择统计分析方法时，必须遵循科学和客观的原则，只能根据研究数据的类型和分布特点来做出选择，并要求最大限度地利用数据的“统计学信息”。不能满足正态分布的条件时，可以通过适当的数据转换（如对数转换、平方根转换等）以达到要求。避免主观地选择统计分析方法和指标，以迎合自己的需要。

常用的数据转换类型及方法见表 1-3-1。

表 1-3-1　常用的数据转换类型及方法

数据类型	转换方法	举　例
Poisson 分布	平方根转换 $x'=\sqrt{x}$	水中细菌数、单位时间放射性计数等
二项分布	反正弦函数转换 $x'=\arcsin\sqrt{p}$	非传染病患病率、白细胞百分数、淋巴细胞转换率等
标准差与均数呈正比关系	对数转换 $x'=\log x$	发汞含量

（二）正态性检验

流行病学研究数据分析中常用的 t 检验和 ANOVA 是统计学家根据数据为正态分布且各组总体方差相同的条件下推导出来的，因而用以分析的数据应该是正态的而且样本方差间差别无统计意义（方差齐）。正态性及方差齐性检验的方法见表 1-3-2。在 SPSS 软件中，可通过 Nonparametric Tests 过程中的 One-Sample Kolmogorov-Smirnov Test 进行正态性检验。

表 1-3-2 常用正态性及方差齐性检验的方法

检验内容	检验方法
正态性	用直方图或正态概率纸进行观察 用矩法、W 法或 D 法进行统计检验
两组方差齐性	F 检验
多组方差齐性	Bartlett 检验

（金丕焕，2000 年）

一般来说，t 检验和 ANOVA 是比较稳健的（robust）。当有所违反上述前提条件时对结果影响不太大。因而在一般情况下还是可以用的，不必太多顾虑。只有在与这些前提条件要求相距过远时才会有重大影响。

（三）非参数统计法

当 t 检验或 ANOVA 的前提条件不能满足而对数据的总体分布不能确定或没有适当的转换方法时，可以用一种不依赖于某一专门的总体分布因而也与参数无关的方法，称为非参数统计法。非参数统计法往往也适用于等级数据。非参数统计法与参数法在无效假设是正确时，其效率相同。当无效假设不正确而分布为正态时其效率稍差；当分布为非正态时，其效率优于参数法。

相应于 t 检验和 ANOVA，有以下一些非参数统计方法（表 1-3-3）：

表 1-3-3 常用非参数统计方法小结

设计方法	参数统计方法	非参数统计方法
配对比较	配对 t 检验	符号检验*、符号等级检验（Wilcoxon 法）
两组比较	成组比较 t 检验	两样本等级秩和检验（Wilcoxon Mann and Whitney 法）、中位数检验*
配伍组比较	随机区组 ANOVA	M 检验（Friedman 法）
多组比较	完全随机设计 ANOVA	H 检验（Kruskal and Wallis 法）

*：效率较差的方法

非参数方法在配伍组设计或多组比较时也有多重比较的方法可用。具体参见相关统计学书籍。

（四）分析指标的选择

对不同的研究设计类型来说，应选择合适的分析指标。分析指标的选择主要取决于研究的目的、设计的类型和所获数据信息。医学研究常用分析指标包括各种率（如发病率、