安徽省高等学校“十一五”省级规划教材
医学研究数据管理与统计分析软件

EpiData 软件实用教程

主　编　孙业桓
副主编　张秀军

安徽大学出版社

安徽省高等学校“十一五”省级规划教材
医学研究数据管理与统计分析软件

EpiData 软件实用教程

主　　编　孙业桓

副 主 编　张秀军

编写人员（依姓氏音序排名）

曹红院　付连国　郝加虎　刘国花　孙　良

孙业桓　王　波　吴红燕　吴学森　杨林胜

姚应水　虞　晨　张秀军　张泽坤　张志华

秘　　书　王　波

安徽大学出版社

图书在版编目(CIP)数据

EpiData 软件实用教程/孙业桓主编．—合肥：安徽大学出版社，2009.4
ISBN 978－7－81110－557－5

Ⅰ.E... Ⅱ.孙... Ⅲ.数据管理系统—应用软件，EpiData 3.1—教材 Ⅳ.TP317

中国版本图书馆 CIP 数据核字(2009)第 044070 号

内容简介

EpiData 软件是丹麦的一个非盈利组织编写的专门用于数据录入、分析及数据管理的免费软件。目前已有 EpiData Entry 和 EpiData Analysis 两个模块。EpiData Entry 可用于数据录入与数据管理(最新版本为 V3.1，有汉化版本)；EpiData Analysis(最新版本为 V2.2)用于基本统计描述、分析与制图，目前尚无汉化版。该软件易学实用，在国内疾病预防控制系统、医学院校、医学研究机构已有众多用户。

本书通过图例相结合的方式，突出实用，系统地介绍了 EpiData Entry 和 EpiData Analysis 软件的特点、基本功能和操作方法。全书内容丰富，可读性强，适合于医学(预防医学、预防保健学、临床医学等)、心理学、社会科学、生物学等学科的教学与研究，本书对于涉及统计分析的实际应用部门也有参考作用。

EpiData 软件实用教程

主　编　孙业桓
副主编　张秀军

出版发行　安徽大学出版社
(合肥市肥西路 3 号 邮编 230039)
联系电话　编辑室 0551－5108871
发行部 0551－5108397
网　　址　www.ahupress.com.cn
电子信箱　ahdxchps@mail.hf.ah.cn
责任编辑　徐　建　刘中飞
封面设计　孟献辉

经　销　全国新华书店
印　刷　合肥创新印务有限公司
开　本　787×1092　1/16
印　张　13
字　数　300 千
版　次　2009 年 4 月第 1 版
印　次　2009 年 4 月第 1 次印刷

ISBN 978－7－81110－557－5　　定价 20.00 元

前　言

近十年来，笔者一直为医学相关专业研究生、预防医学专业本科生、临床医学和公共卫生工作者讲授医学数据管理及相关统计分析软件，并为医学科研人员做科研设计咨询工作，从中受到了很多启迪，同时，增强了我们的紧迫感、危机感和使命感。如何对医学研究数据进行高质量的管理呢？这使我们陷入了深深的思考。

目前，计算机技术日新月异，可以进行医学数据管理与统计分析的软件种类繁多，如 SAS、SPSS、Stata、Excel、Access、Epi Info 软件等，但从“学即有用，学即会用，学即实用”，且能满足各专业硕士研究生、本科生及医疗卫生工作者在课题研究阶段顺利进行科研数据管理与统计分析要求的角度考虑，同时也考虑到商品化软件价格昂贵等因素，在目前情况下我们首推免费共享的 EpiData 系列软件。

1997 年以来，我们根据当时实际情况陆续编写了《Epi Info 实用教程》、《Epi Info&SPSS 实用教程》系列讲义，受到了广大研究生及医学研究工作者的热烈欢迎，但随着计算机 Windows 操作系统的推广应用，MSDos 版本下的Epi Info软件已难以适应现阶段的研究生及医疗卫生工作者的需要。虽然美国 CDC 的 Epi Info 工作组将 Dos 版 Epi Info 更新到视窗版 Epi Info，但更新的Epi Info视窗版采用了一种全新的工作策略，数据库采用 Access 数据库格式，操作上远没有 Dos 版易学好用，因此使得原 Epi Info 6.0 用户群纷纷另觅其他软件，由此催生了 EpiData 的开发与发展。

1999 年末，Jens M. Lauritsen、Mark Myatt 和 Michael Bruus 组成研发小组，开发出一个简单、易用、独立的 EpiData 应用程序。该软件不需要任何专门的数据库系统驱动（基于.dll），堪称“绿色环保”，并免费发布。任何人均可从 http://www.epidata.dk/download.php

网站下载使用，并可随时更新升级。目前，已有 EpiData Entry 和 EpiData Analysis 两个模块。EpiData Entry 可用于数据录入与数据管理（最新版本为 V3.1，有汉化版本），功能相当完善；EpiData Analysis（目前版本为 V2.2）用于基本统计描述、分析与制图，目前尚无汉化版本，根据软件开发计划，EpiData 的功能将会逐步完善，读者应经常访问 EpiData 官方网站，以了解软件的升级情况。

EpiData 正因为其有免费共享、易用、实用等优点，在国内疾病预防控制系统、医学院校、医疗卫生研究机构已有众多用户，原先对研究生、本科生开设 Epi Info 课程的院校，现大多已开始开设 EpiData相关课程。因此，根据医学院校研究生、预防医学专业本科生及广大医学工作者的实际需求，笔者开设了 EpiData 软件应用课程，编写了《EpiData Entry & Analysis 实用教程》，在此基础上我们组织安徽医科大学、皖南医学院、蚌埠医学院、安徽理工大学部分相关教师及研究生编写了本教材，申报并获批为安徽省高等学校“十一五”省级规划教材。

全书内容丰富，可读性强，突出实用与可操作性，在各章介绍具体内容的时候，结合示例数据库“安徽省产妇研究数据库”做到了理论与实践紧密结合。经过短期培训学习，可使研究生、相关专业本科生及医学工作者具备基本的医学研究数据管理与统计分析能力。

由于编者水平所限，加之计算机科学技术发展迅速，书中错误和不妥之处在所难免，恳请读者批评指正。

孙业桓　张秀军

2009 年 3 月

目　录

前言 …… I

第一章　学习 EpiData Entry & Analysis 软件背景知识 …… 1

第一节　医学科学研究的基本程序 …… 1

第二节　调查表设计 …… 3

第三节　微型计算机在科研数据管理与分析中的应用基础 …… 9

第四节　常用统计软件包简介 …… 15

第二章　EpiData Entry & Analysis 软件概述 …… 21

第一节　EpiData 的历史 …… 22

第二节　EpiData 与 Epi Info 的兼容性及区别 …… 24

第三节　EpiData 主要功能特点 …… 25

第四节　EpiData 主界面简介 …… 26

第三章　QES 文件的建立(Create Questionnaire File) …… 32

第一节　建立 QES 文件的准备工作 …… 32

第二节　QES 文件的结构 …… 35

第三节　编辑器(Editor) …… 39

第四章　数据库的建立和修改(Make and Revise Data File) …… 51

第一节　生成 REC 文件(Make Data File) …… 51

第二节　修改 REC 文件(Revise Data File) …… 53

第三节　变量重命名(Rename Fields) …… 54

第五章　核查文件(Checks) …… 56

第一节　Check 文件的建立 …… 56

第二节　基本 Check 命令的设置 …… 57

第三节　Check 文件的基本结构 …… 64

第四节　Check 命令详解 …… 69

第六章　数据录入(Enter Data) …… 90

第一节　概述 …… 90

第二节　变量(Field)、记录(Record)、关联数据库(Relate Data)间切换 …… 90

第三节　查找变量(Field)、关联变量(Relate Field)和记录(Record) …… 92

第四节　滤过(Filter) …… 93

第七章 数据库的管理和维护 …… 95
第一节 数据库的追加与合并(Append/Merge Data Files) …… 95
第二节 双录入核查(Double Entry and Validation) …… 98
第三节 一致性逻辑核查(Logical Consistency Check) …… 100
第四节 根据数据库创建 QES 文件(Make QES File from Data File) …… 101
第五节 重新编码数据库(Recode Data File) …… 101
第六节 数据库的清理 …… 102
第七节 压缩数据库(Compress Data File) …… 103
第八节 准备双录入实时校验 …… 103
第九节 创建和释放压缩文件(Archive) …… 103
第十节 数据库相关信息的管理 …… 104
第八章 数据库的导出和导入(Data in/out) …… 109
第一节 数据库的导出(Export Data) …… 109
第二节 数据库的导入(Import Data) …… 113
第九章 数据分析(EpiData Analysis)模块 …… 116
第一节 数据分析(EpiData Analysis)模块的运行 …… 116
第二节 数据统计分析(Data Analysis) …… 118
第十章 利用原始数据制作统计图 …… 128
第一节 统计图制作的原则和要求 …… 128
第二节 EpiData 常用统计图的制作 …… 128
第三节 常用统计图形的选择及注意事项 …… 136
附录一 实用的 Openepi 统计软件 …… 138
附录二 Epi Info 软件 …… 144
参考文献 …… 199

第一章　学习 EpiData Entry & Analysis 软件背景知识

第一节　医学科学研究的基本程序

科学研究(Scientific Research,简称科研)是以科学的观点和方法,对未知事物进行探索、观测和分析,从而发展有关科学知识(理论和技术)的认识活动。科学研究是人类认识活动的一种主要形式,凡是以正确的观点与方法考察人类社会、自然现象与思维规律,并通过理性概括以揭示其本质的认识活动,都可以说是科学研究;凡是创造知识或整理知识并产生新的知识或新的认识的工作,都是科学研究。

医学是研究人类生命过程、寻求防病治病规律、维护人类健康长寿的科学。医学研究的对象是人的生命现象与疾病过程。人类的生命现象与疾病的发生发展过程,几乎包括了地球上物质运动的一切形式。医学研究所采用的方法也极为多样,从简单的肉眼观察到电子显微镜的应用,从试管分析到社会调查,几乎无所不包。医学是一门综合性应用科学,它是在生物学、物理学、化学、生物化学……发展的基础上产生与形成的,因而,相对地说,医学发展比其他科学技术发展较为缓慢,研究的难度也较大。

医学科学研究和其他科学研究一样具有探索性、创新性、复杂性、继承性、积累性,以及个人独立思考与集体性等特征,因此科学合理的科研程序可以有效地指导研究活动,使科研工作符合科学规律,取得科学结果。医学科研的基本程序可概括为选题、科研设计、计划实施、数据管理与分析和总结五个阶段。

一、选题

科研选题是指确定研究具体科学技术问题的酝酿决策过程。科学研究是始于问题而终于问题的具有创造性的活动。关于科学研究问题,爱因斯坦在《物理学的进化》一书中指出:"提出一个问题往往比解决一个问题更重要,因为解决问题也许仅仅是一个数学上或实验上的技能而已。而提出新的问题、新的可能性,从新的角度去看旧的问题,却需要创造性的想象力,而且标志着科学的真正进步。"可见,发现和提出问题是科学研究开好头的关键一步,科学研究的实践证明,选题不仅决定着预期成果的价值,而且关系到研究工作的成败。科研选题要有明确的目的和先进性、创造性、科学性与可行性。要防止随波逐流、盲目跟从,避免低水平的重复性研究。科研选题集中体现了选题者的科学思维、理论认识、实验能力、科研能力以及预期要达到的成果。科研工作者务必以严肃认真的态度对待选题。一个良好的选题应当是"情况明,起点高,新意强,思路好"。为集思广益,对选题应进行开题报告,广泛听取专家与同行的意见。

二、科研设计

科研设计就是针对某项科研课题而制订的总体计划与方案，是研究者验证假设的“蓝图”。通过严谨的科研设计，可以保证科研（实验、调查）结果的准确性、可靠性，可以使科研工作少走弯路，减少不必要的浪费，并可以保证科研结论的可重复性。科研设计的质量关系到科研工作的成败，良好的科研设计是科研工作的先导，是实验数据统计处理的前提，是科研工作实施的路线，是决定研究水平高低和成败的关键。

科研设计包括专业设计与统计学设计两方面。专业设计就是完成课题的专业思路、技术路线与方法的确定，它是科研创新性与学术水平的决定因素；统计学设计就是控制误差、改善实验有效性与资料分析方法的规定，它是保证专业设计的布局合理性和实验结论可信性的要素。专业设计不科学、统计学设计不正确，必然不能得到正确可靠的结论，造成人力、物力、财力的浪费，甚至造成严重的危害。专业设计是为了回答和解决科研所提出的问题，是科研是否有用、是否先进的前提和基础，主要解决研究结果的有用性和独创性；统计学设计是为了减少或控制误差，保证样本的代表性和样本间的可比性，是科研结果可靠性和经济性的保证，主要解决研究结果的可重复性和经济性。两者相辅相成，不可或缺。

三、计划实施

在实施阶段要严格按照科研设计方案进行观察和实验，以获取第一手客观事实资料，收集的资料必须全面、客观，能准确地反映研究对象和事物的本来面目。要注意资料的完整性、可重复性和真实性，切忌主观和片面。充分地占有材料，这是科研的基本要求。材料不会从天上掉下来，而是来自于实践、来自于实际工作。在医学科研中，常用调查、观察和实验的方法收集资料。

对以人群为对象的医学科学研究来说，组织管理工作非常重要，它包括两个方面，一是争取有关机构、部门及人员对研究工作的支持、配合与保证；另一方面是组织、管理参加研究的工作人员。

在良好的组织和管理之下，还需要考虑质量控制。除严格遵守设计相关工作规范外，还应定期检查，严格把住验收关。比如，定期对 10% 的工作进行抽查，评定其好坏，以便随时纠正出现的问题，不合格者返工或废弃重做，阶段性地或最后验收时逐个查看，以保证质量，其目的就是使研究获得高度真实性、完整性、可靠性的资料。

通过调查或检验得到的资料，经过验收后，应该妥善保存。验收与保存都应有一定的交接手续，以明确责任，防止遗失。资料保存时应有恰当的分类，以利于分析时查询使用。已归档的资料不得再涂改或补充，也不得随意复制。应注意资料的保密原则，特别是涉及个人隐私的项目，更不应该让人随便利用。某些标本，例如血清的保存，除了注意防止丢失外，还要防止因保存不善而损坏；即便是文字数据的资料，也应考虑是否用较好的纸张，用不褪色的笔墨书写。

四、数据管理与分析

对观察与实验中所收集到的大量数据资料应当依据科研设计要求进行科学的管理。

由科研实践得来的材料，其数据总是离散的，但它们的分布具有一定的规律性。这种规律性的揭示有赖于统计分析。所谓统计分析，就是按照数理统计方法要求，对收集的资料进行统计描述与统计推断。只有通过统计分析才能排除偶然，发现必然，才能根据局部（样本）结果，引出普遍（总体）结论。所以正确掌握与运用数理统计方法是科研的基本功之一。

五、总结

总结是科研过程的最后一环，就是根据观察事实与统计处理的结果，运用分析、综合、归纳与演绎的方法，把感性材料上升为理性概念。总结、归纳的基本形式是撰写科研论文或撰写结题报告等。

第二节　调查表设计

调查表是指将调查内容和调查项目按提问的逻辑顺序列成的便于调查者收集资料的表格。调查表或称问卷，是调查研究的重要工具，其调查内容的设置及研究项目的安排有着高度的科学性和严谨的逻辑性。调查表设计的质量直接影响到资料收集过程的质量和研究结果的可靠性与准确性。广义上讲，实验记录表格也属调查表范畴。

一、调查表设计原则

（一）结构合理，项目齐全

一份完整的调查表应包括开场白说明语、调查项目及审核记录等项目，其结构安排及篇幅比例均应合理，不可乱编乱排，项目的罗列层次感要强，要有逻辑，表格大小设计应合适、美观。更重要的是调查项目应齐全，不要漏查，但也不要累赘。

（二）语言表述规范、准确

每个调查项目的语言表述应通俗、易懂，易于被调查者接受，忌用专业术语，以免令人费解。如“心悸”是专业术语，可改为“心里发慌”；“蜱”是专业术语，可改为当地人的俗语“牛虱子”等。项目中的数字表达必须准确。如“年龄是多少”应改为“出生年、月、日”，并加“属相”供交叉核对，数量范围的等级、界定同样要明确，不能有遗漏或交叉重叠。

二、调查表的类型

（一）按填写方式分类

1. 询问式调查表

由调查人员向被试者询问而进行的调查。其准确性高，应答率高，但工作量大，组织实施较难。

2. 自填式调查表

由被试者针对各调查项目逐项自行填写而进行的调查。其可靠性高，易组织实施，但应答率较低，容易失访。

在设计询问式和自填式调查表时，一定要注意语气及主语人称的区别，不要“你、我”不分，并根据调查方式及组织实施情况进行设计。

（二）按资料的性质和内容分类

1. 一览表

指一张调查表可同时调查多个调查对象的有关信息。一览表多用于统计报告、医学监测等，很少用于专题调查。

2. 单一表

指一张调查表只调查一个调查对象的信息。

设计何种类型的调查表主要应根据调查内容和调查对象的情况来进行选择。

（三）按调查方式分类

1. 信访调查表

通过邮寄或直接送达被调查对象，由被调查对象自行填写的一种问卷。

2. 电话访问调查表

采用电话询问调查内容以获得研究信息的一种问卷。

3. 面访调查表

由调查员对被调查对象采用面对面方式进行询问，以获取研究信息。

三、调查表设计的步骤

（一）明确研究目的

通过严格操作，使研究目的可以用一系列指标来测量，即具有可操作性。例如，调查某种疾病患者的生命质量，生命质量作为一个不易测量的概念，可以应用生理、心理、社会生活状态等一系列指标，甚至综合指标进行测量。

（二）建立问题库

1. 头脑风暴法

与调查相关的人员组成研究小组，自由发表意见，提出有关指标。然后将指标进行归类、合并、删除等，以形成调查项目问题库。

2. 借用其他问卷的项目

从已有的调查问卷中选用符合本次研究目的的项目，需注意的是，重新组合的调查问卷应进行效度和信度检验。

（三）设计调查表初稿

将零散的问题组装成一份合适的调查表应考虑到各种问题的前后顺序、逻辑结构、对被调查者的心理影响、是否便于回答等多方面因素，尽可能统筹兼顾，形成调查表初稿。

（四）试用和修改

试用的方法包括客观检查法和主观评价法，前者是指用调查表初稿进行预调查，以发

现调查表中的问题；后者是将调查表初稿分送该领域的专家，请他们评论。有条件时最好这两种方法都采用，先用主观评价法进行修改，再用客观检查法进行再修改。

（五）效度和信度检验

通过效度和信度检验来评价调查表的质量。

四、调查表的结构

设计调查表时，首先应该构筑调查表的框架结构，然后再进一步扩充、细化。一份完整的调查表结构除说明语外，通常由三部分组成：识别项目、专题项目、审核项目。

（一）识别项目

识别项目是指对调查对象个体的一般概况的调查，如姓名、性别、出生年月日、籍贯、民族、文化程度、职业、婚姻、工作地址等，必要时还应包括有关背景资料，如家庭成员人数、家庭类型、有关可查资料的编号（病历号、X 片号、档案号等）以及身份证号码和电话号码等，这些项目不是专题调查内容，但可为专题资料分析提供基本信息和参考价值，某些情况下甚至会影响研究的成败。

（二）专题项目（研究项目）

这是针对一项研究时调查表的核心内容，不同的研究目的，其调查表的研究项目不同，研究者应根据研究目的所确定的调查内容逐项设置，供分析研究之用。

（三）审核项目

审核项目是指由调查员或审核人员对调查表各项目调查结果进行审查核对时所需填写的项目，主要用于质量控制。如调查人员姓名、调查日期、核对人姓名、调查表分类编号以及获得受试者有关信息的真实性、可靠性的评价等，目的是提高调查人员的责任感，防止调查表中的错、漏、笔误等，以保证所收集资料的质量。

五、调查表项目的设计方法

（一）项目设计的形式

1. 开放式

即不限制答案的范围，被调查者可根据自己的情况填写答案。对于姓名、出生日期或呈连续性分布的变量可以采用开放式项目设计。如：姓名_____；年龄_____岁。

2. 封闭式

即针对某一问题所有的可能性回答，提出两个或多个固定的答案列到调查表上，由被调查者选择。如：既往糖尿病史（有＝1，无＝0）。

3. 混合式

在一张调查表上同时包括开放式和封闭式问题，绝大部分调查表为混合式。

许多采用封闭式问题的问卷，常常在预调查时先用部分开放式问题，以确定封闭式问

题的答案种类。为了保证封闭式问题包括全部答案，可以在主要答案后加上“其他”之类的答案，以作补充，避免强迫被调查者选择不真实的答案。

(二)项目调查句式

项目调查句式主要有两种：陈述句式和疑问句式，两者没有本质区别，例如：陈述句式：你的性别：男＝1，女＝2；疑问句式：你的性别是什么？男＝1，女＝2。

调查句式主要依据被试者的文化程度高低来进行选择。被试者文化程度低或理解能力差多用一些疑问句式，大多数情况用陈述句式。

(三)项目的数量和顺序设计

为了使项目数量精炼，设计问卷项目时常遵循“五不问”原则：①可问可不问的项目不问；②复杂问题项目不问；③查找资料才能回答的项目不问；④被试者不愿意回答的项目不问；⑤通过其他手段才能解决的项目不问。

1. 项目的数量

一份调查表的项目数量设计，通常情况下以限制被试者 30 分钟内完成为宜。若超过 30 分钟，其调查信度很难达到要求。

2. 项目的排列顺序

当将零散的问题组成一张问卷时，必须考虑各个问题在问卷中的排列顺序。调查表各专题项目的排列一定要顺序合理、逻辑性强，否则会产生系统误差。项目顺序设置应注意以下几点：

(1)先排列容易回答的、无威胁性的问题。如年龄、性别、职业等事实问题宜放在前面。一般情况下，敏感性问题宜放在问卷的后面，以免引起回答人的反感，影响对后面问题的回答。

(2)先排列封闭式问题。开放式问题需要时间考虑，回答不易，如将这类问题放在前面，容易导致被试者拒绝回答，影响应答率或问卷的回收率；对于封闭式问题，每个项目的所有相关子项目应作为一个封闭性整体依次排列，该项目调查完之后再调查下一个项目。

例如：6 你有饮酒习惯吗？(是＝1，否＝0)如否，转向问题 7；如是，继续问答下列问题：

6.1 第一次喝酒年龄_____岁；

6.2 经常喝酒种类等……

即把饮酒这个项目作为一个封闭性整体，有关子项目全部调查完之后再问另一个项目。

(3)问题要按一定的逻辑顺序排列。应考虑到人们的思维方式，按事物的内容和相互关系以及事情发生或发展的先后顺序排列。

内容和性质相同或相近的问题应集中在一起，问完一类问题之后再转向另一类问题，避免跳跃性的提问。对有时间关系的问题，应按顺时针或逆时针方向提问，不要随意更换问题的次序，否则可能扰乱回答者的思维。有时为了防止被调查者的厌倦或不假思索地随便答问，也可随机地使用各类形式的问题和不同的排列次序相结合，增加问卷的多样性。

(4)检查信度的问题须分隔开来。在很多问卷中，研究者有意设置一些高度相关或内容完全相同而形式不同的问题，这些成对出现的问题，目的是检验问卷的信度，但它们不

能排在一起，否则回答者很容易察觉，并使回答无矛盾，达不到检验的目的。

(5)漏斗式的问题顺序技术。使用漏斗技术时，先排列范围广的、普遍的问题，然后漏斗变狭，安排较具体、较特殊的问题。例如调查吸烟，不是先问你吸几包烟，而是先问你是否吸烟。

(四)项目的测量设计

调查表各项目力求量化，不能定量的应定性。定性、定量均涉及测量问题。常见的调查表的项目测量方法有四种：定类测量、定序测量、定距测量、定比测量。

1. 定类测量

主要是对定性项目的测量，如性别、文化、职业、婚姻等。定类测量要求定类要规范、齐全、无交叉。比如，你的职业：工人、农民、干部、知识分子、商业、家务、其他。其中干部和知识分子有交叉，工人和商业也存在交叉，所以，设计时，定类测量常容易出现此类错误，必须避免。

2. 定序测量

定序测量是指按照一定的数量顺序进行的测量。常用的有等序法和非等序法。等序法是指用相等的间距数字顺序赋值测量；非等序法是指按级别差异顺序进行测量。

例如：(1)评委评分结果(等序法)：

评委	甲	乙	丙
A	3	5	…
B	2	1	…
C	3	3	…
D	1	2	…

注：1——最差，…，5——最好。

(2)你的学习成绩(非等序法)：

①好，②一般，③差。

3. 定距测量

定距测量是指按照时间或空间距离的顺序依次测量。如时间距、尺度距等。

例：1990 年你家人口数——

1991 年你家人口数——

1992 年你家人口数——

……

1999 年你家人口数——

注意：定距测量不能有“跳跃”现象，一定要等间距设置，距离的粗细依研究目的而定。

4. 定比测量

定比测量是指以一定的数量比值顺序进行的测量，通常用于有递增或递减变化的调查项目。

(五)项目编写格式的设计

项目的编写格式依据项目的测量方式而定。常见的有二项式、多项式、矩阵式、序列式、填空式、图画式、尺度式、自由式等。举例如下：

(1)二项式：如，是＝1，否＝2；

(2)多项式：如，很满意＝1，较满意＝2，较不满意＝3，不满意＝4，很不满意＝5；

(3)填空式：如，你的年龄28岁；

(4)图画式：如，请画出一个三角形；

(5)矩阵式：如，你对下列问题的满意度如何？

项目	满意	不满意
工作环境	√	
工资收入		√
居住环境	√	

(6)尺度式：如，用药后你的疼痛程度；

0	1	2	3	4	5	6	7	8	9	10
不痛			√							很痛

上述各种编写形式各有利弊，应根据问题的性质和研究目的结合应用。

(六)项目设计的常见错误

1. 双重装填

指一个问题中包括了两个或两个以上的问题，有些应答者可能难以准确而全面地做出回答。例如，你是否患有糖尿病并接受合理治疗？

2. 含糊不清

使用了一些词意含糊不清的词，或使用了一些专业术语、俗语，从而使问题不易为他人理解。有时也可能因为对问题的表述不准确或修饰语过多，从而使问题的意思含糊不清。

3. 诱导性提问

这类提问会人为地增加某些回答的概率，从而产生偏倚(Bias)。因为带有诱导性的提问，容易使无主见的回答者顺着你的意思回答，故提问时应采用中性的提问方式。

4. 抽象的提问

涉及幸福、爱、正义等抽象概念的提问一般较难回答。许多回答者遇到这类提问时，可能发现自己从未思考过此类问题。如果问卷一定要涉及这方面的提问，最好给出一些具体的看法，让回答者仅回答赞成与否。

5. 敏感性问题

有些问题对于回答者是非常敏感的，如未婚先孕、流产、同性恋、吸毒等。这类问题宜慎重设计，否则将因回答者说谎造成偏误。此类问题可采用特殊的敏感问题调查技术。

六、调查表使用的一些注意事项

(1)调查表设计时必须同时编写使用手册或操作指南，实际调查时要严格按照其中的

规定和要求执行。

(2)必须对调查员进行统一的培训和考核。

(3)填写的字迹要工整、清楚,不能缺项。

(4)调查员要签名,并注明调查日期。

第三节　微型计算机在科研数据管理与分析中的应用基础

随着微型计算机应用的迅速普及,这为广大医学科研工作者高效率地处理大量医学科研数据创造了有利的条件。微型计算机不仅使疾病调查数据和实验观察数据能得以长期保存和高效的数据管理,而且更为重要的是,它使这些被保存的原始数据能在将来根据研究需要不断地被重新分类、检索和更新。与任何其他医学领域中的研究方法一样,利用计算机进行科研数据的管理和处理也是当今医学科研工作者必须要了解和掌握的一门知识,这门知识将有助于医学科研工作者更好地利用现有计算机资源来处理科研数据,并且尽量减少或避免数据管理和处理过程中的盲目性,以便达到事半功倍的效果。本节将叙述和讨论涉及疾病调查数据和实验研究数据计算机处理的有关问题及其解决方法。

通常,我们遇到的科研数据可分为频数数据和个案记录(Individual Record)两种类型。频数数据经常来源于已用手工整理过的各种统计报表或某次调查的计算机中间结果。所谓个案记录即为调查表,它记载了调查研究对象与研究有关的全部信息,任何类型调查研究都应该用个案记录方式将调查研究数据保存在计算机数据文件里。这些被保存的个案记录可在将来进一步用计算机软件来进行分组或选择。下面所要讨论的计算机数据处理过程主要是针对个案记录。

一、科研数据的计算机处理过程和方法

有些调查投入大量的人力和物力来收集疾病和实验数据,但却不太注意数据计算机处理过程中的各个环节,以至于工作效率很低,或者输入计算机数据文件中的数据错误百出。对于个案记录,科研数据的计算机处理过程可分为三个阶段,即数据输入准备阶段、数据输入阶段以及数据预处理阶段。经历了这三个阶段以后,疾病调查和实验数据可再被进一步送到专门的统计软件进行分析。

(一)数据输入准备

在正式用计算机输入数据之前,应先进行各项准备工作,例如编写数据编码说明、设计数据文件结构以及培训数据输入员等。充分的准备工作将有助于避免或减少输入错误数据,并且提高数据输入速度。

1. 数据编码

在应用计算机处理科研数据过程中,各原始数据都应转换成计算机能够识别的代码。在调查设计阶段最好就已考虑到数据的编码方案。如果调查表项目很多,还应编写编码说明或手册。编码说明或手册将记录每一变量的名称以及代码数据的实际意义。它类似于电报密码手册,调查者要根据它将调查数据转换成代码数据,输入计算机进行处理和分析,然后再根据它将计算机结果变换成符合常规写法的数字或文字内容。

代码数据应尽可能保留原始调查数据的形态及包含的信息。一般情况下，定量数据可不经转换直接写成代码数据，例如年龄、血压、药物剂量、吸烟支数等。如果人为地将定量数据转变为等级（分组）代码，如先对年龄进行分组，即 1～19＝1，20～39＝2，..，＞80＝5，然后再将这些分组的年龄代码数据 1，2，3，4，5 输入计算机数据文件，这样会造成原有数据信息的丢失，严重时还会导致研究失败。定性数据，例如性别、职业、婚姻状况、疾病类型等必须设立能相应于实际意义的代码。尽管调查者设置的代码可以是任意的，但仍应提倡采用规范方法来进行设置。比如，当一般项目涉及“有、无，正、负，阳性、阴性”时，可用“2”代表“有、正或阳性”，用“1”代表“无、负、阴性”。对阳性事物设立高次代码有助于调查者在分析时获得正的参数估计值。编码方案一旦确立，调查者必须严格遵守。在同一批调查和实验数据里，调查者不应对某些定性数据的阳性事物设立高次代码，而对另外定性数据的阳性事物又改为设立低次代码，这有可能造成代码数据意义的混乱。有序的定性数据代码也应根据原有次序和高低保持有序，例如疾病程度分为轻度、中度和重度，其相应的代码可设为 1、2 和 3。

对于在实际调查和实验过程中出现的缺失数据，应根据缺失数据的不同类型——即失访数据（又称漏失数据）和空白数据，确定编码方法。虽然这两种类型的缺失数据在客观上都造成了调查表项目的空白，但它们的性质及其在分析时的处理方法互不相同，漏失数据是指应该调查而未被调查到的数据，例如调查对象回答不准确、调查员记录不清、项目遗漏或随访过程中调查对象失访等原因都会导致漏失数据的产生。空白数据是在调查对象不具有某些项目的情况时产生的，它属于不必调查的数据，例如未生育者的首次生育年龄或哺乳年龄的调查项目一定呈空白状态。有些调查者将漏失数据和空白数据都视做同一代码，例如“9”或“0”处理，这对统计分析特别是多元统计分析极为不利。如果某项目漏失数据较少，仍可用适当方法加以处理并使其参与分析，但真正的空白数据完全没有必要也不应该这样来做。一般来说，可用适当位数“9”或“×”代码来表示漏失数据，用“0”或“—”代码表示空白数据。如果调查者在编码方案中已考虑用“0”代码表示空白数据，那么最好不要再将其用做表示调查项目的数据代码。

2. 数据文件结构的设计

一个标准的数据文件是由原始数据代码和文件结构两部分组成的。数据文件结构设计是否合理将直接影响数据的储存量、输入速度和误码率，以及在分析数据时的可行性。数据文件结构主要由三部分组成，即变量（调查项目）的名称、类型和宽度。一般来说，一个理想的变量（也称字段）名称应在屏幕显示时能反映代码数据的实际意义，而在分析时它又能做到简明扼要。变量名称太短，会使数据输入员在输入数据时无法正确理解变量内容，以至于使误码输入增加；然而过长的变量名又会减慢数据分析速度，这是因为调查者每次需要输入一大串变量名。

原则上，变量类型应与原始代码数据相一致，例如，可将姓名设计为字符型（文本型）变量，将年龄设计为数字型变量，以及将发病日期设计为日期型变量，等等。一般来说，诸如年龄、血压之类的定量数据应保留其数据的原形态，但对于某些定性数据，例如性别、是否吸烟等，既可将它们设计成字符型变量（男性＝M；女性＝F），也可将它们设计成数字型变量（男性＝1；女性＝2）。另外，假如某些调查项目发生重叠现象，那么可考虑将它们设计成字符型变量，以便能减少变量数，使数据文件结构精简。比方说，在疾病症状调查