高等学校教材

SPSS统计分析基础教程

主　编　张文彤
　　闫　洁

高等教育出版社

内容简介

SPSS是最为优秀的统计软件之一,深受各行业用户的青睐。为满足广大读者学习统计学入门知识和统计软件入门操作的需求,本书改变了以往SPSS书籍对统计理论和软件操作"两条主线、各自表述"的编写方式,将两者完全融合起来。全书共分15章,以SPSS12.0为准,针对统计初学者和SPSS初级用户的需求,以统计理论为主线,详细介绍了在SPSS中的界面操作、数据管理、统计图表制作、统计描述和常用单因素统计分析方法的原理与实际操作。其内容覆盖了目前国内大部分专业本科统计课程的教学范围,并结合SPSS的强大功能做了很好的扩展。各章后均附有参考文献和思考练习题,涉及统计理论的章节还提供了本章小结。全书内容深入浅出,风格简洁明快,是一本难得的统计理论与SPSS操作相结合的教材。

本书可用作各专业本科生和研究生的统计学教材,也可作为SPSS10～12版的通用入门教材,可供各行业中非统计专业背景的人员以及希望从头学习SPSS软件的人员使用。

图书在版编目(CIP)数据

SPSS统计分析基础教程/张文彤,闫洁主编.—北京:高等教育出版社,2004.9
ISBN 7-04-015855-8

Ⅰ.S... Ⅱ.①张...②闫... Ⅲ.统计分析—软件包,SPSS—高等学校—教材 Ⅳ.C819

中国版本图书馆CIP数据核字(2004)第087691号

策划编辑 耿 芳 责任编辑 欧阳舟 市场策划 韩 飞 封面设计 于文燕
版式设计 张 岚 责任校对 朱惠芳 责任印制

出版发行 高等教育出版社
社 址 北京市西城区德外大街4号
邮政编码 100011
总 机 010-58581000

购书热线 010-64054588
免费咨询 800-810-0598
网 址 http://www.hep.edu.cn
http://www.hep.com.cn

经 销 新华书店北京发行所
印 刷

开 本 787×1092 1/16
印 张 24
字 数 580 000

版 次 年 月第1版
印 次 年 月第 次印刷
定 价 32.00元

物料号:15855-00

SPSS统计分析基础教程参编人员

主　　编　张文彤（复旦大学）
　　闫　洁（博塔（中国）有限公司）

编　　者　（以姓氏笔画为序）
　　王　莉（天津财经大学）
　　邝春伟（华东师范大学）
　　行智国（博塔（中国）有限公司）
　　邹艳辉（博塔（中国）有限公司）
　　赵　杨（南京医科大学）
　　赵新平（复旦大学）
　　高　峻（复旦大学）
　　董　伟（博塔（中国）有限公司）

序 言

知识经济时代,数据成为宝贵的经济资源。在国外,电信、医疗、银行、证券、保险、制造、商业、科研、教育等行业已广泛采用先进的统计分析技术从数据中提取有价值的信息和知识。在国内,随着市场的蓬勃发展,了解成熟的统计分析产品,借鉴成功的统计分析行业应用经验,运用科学的统计分析方法,从数据中总结、归纳有用的知识,并将知识用于市场营销、运营决策和信用风险管理等领域,帮助企、事业单位降低消耗、增加效益,从而提高整体运行效率,已成为各行业中有远见的人士普遍关注的问题。

SPSS软件是全球专业统计分析软件的领导者,一直致力于帮助企事业单位提高科学运用统计分析方法的能力,20世纪 80年代就已经被许多学者引入中国市场。它包含了丰富的统计分析算法,而且在使用层面上更多地考虑了客户在整个统计分析过程中的应用感受,其简洁的界面、完善的数据准备功能和杰出的图表输出能力使得 SPSS软件在全球有超过 25万家的机构用户,并成为国内的主流统计分析软件。随着该产品在各行业应用的深入以及 SPSS培训和认证的广泛开展,目前国内已涌现出一大批应用 SPSS的专家。近两年,国内统计分析市场如火如荼,SPSS在产品技术上也不断推陈出新,继 SPSS英文版在国内成功应用之后,SPSS公司在今年首次推出了简体中文版,该产品更加符合中国人的使用习惯,一经推出便受到各行业人士的喜爱。

为了让中国的 SPSS软件爱好者更好地使用软件,我们在经过两年的筹备后向市场推出了SPSS统计分析大型丛书。该丛书是一套全面了解、认识和应用 SPSS最新统计分析软件、掌握统计分析方法的专业书籍,以统计分析在国内的应用现状为切入点,本着学以致用的原则,在介绍统计理论及 SPSS软件功能模块的同时,更侧重于统计分析在各项工作中的实际应用,引导读者不仅掌握 SPSS软件及技术原理,而且学会运用统计方法解决工作和学习中的实际问题。

该丛书由业内权威专家主笔编写,资料引用详实可靠,实例剖析切中肯綮,不仅融合了行业专家在统计应用领域多年的研究成果,而且还融入了很多 SPSS软件新、老行业用户的实际应用经验。丛书总结了 SPSS软件在各行业的实践应用状况,并综合 SPSS最新行业应用方案,使各行业读者能通过学习提高 SPSS软件的运用能力,解决工作中的实际问题。在丛书编写过程中,SPSS公司和博塔(中国)有限公司(SPSS中国地区分销商)的技术专家还及时提供了国际最新的行业发展信息和 SPSS最新产品和技术信息,并结合 SPSS的全球应用状况提出了宝贵意见。

丛书将分期分批出版相应的分册,其中首批面世的为通用教材《SPSS统计分析基础教程》、《SPSS统计分析高级教程》,均由复旦大学张文彤老师主编,全国多所高校的统计教师和统计专业人士参编。张文彤老师长期以来一直致力于积极推进统计分析工具在国内的普及应用,他在2002年编著的前作《SPSS 11统计分析教程》基础篇和高级篇因内容翔实、风格独特,受到了广大读者的热烈欢迎,并被多所高校列为本科生或研究生教材,其中基础篇一书已通过教育部评审,成为 2003—2004年度教育部研究生工作办公室推荐的“研究生教学用书”。他不仅拥有丰富的教学经验,而且熟谙统计分析产品,本次全新编写的这两册教材分别针对不同读者群,由浅入深、

结合实际应用全面介绍了 SPSS产品和应用。全书实例引用突出,分析讲解透彻,读者可由本书管窥全套丛书“应用为本”的特色。

显然,上述两本书还不能完全覆盖 SPSS的所有应用领域,因此,本套书从现在还有数本分册正在组织编写中,包括全新的行业应用分册。这里我们热忱邀请各行各业的 SPSS资深用户,以及各高校的统计教师加入到本套丛书的编写工作中来,以共同推动我国各行业统计应用水平的迅速提高。

希望本套丛书能够让读者更清晰地了解统计分析,从而进一步促进统计分析在国内的普及。为便于读者交流和使用本套丛书,这里特公布相关网址如下:

SPSS最新版本的全模块试用版下载:www.spssbj.com.cn

丛书相关案例数据下载:www.spssbj.com.cn、www.MedStatStar.com

读者答疑、经验交流:www.spssclub.com、www.StatStar.com

博塔(中国)有限公司 SPSS丛书编委会

前 言

笔者前作《SPSS 11统计分析教程》(基础篇)和《SPSS 11统计分析教程》(高级篇)自2002年中面世以来,因其内容翔实、风格独特,受到了广大读者的热烈欢迎,这从读者用E-mail请教问题的数量即可看出,还有数位读者详细指出了书中的用字错误,可见其阅读的详细程度,在此笔者对读者深表谢意。同时,数所高校均将其作为本科生或者研究生教材,而基础篇一书已通过教育部评审,成为2003—2004年度教育部研究生工作办公室推荐"研究生教学用书",这无疑都是对前作质量的充分肯定。

但是,再优秀的作品也有其生命周期,随着时间的推移,上述著作的不足之处也逐渐显现出来。对于基础篇而言,突出表现为以下两点:

1. 由于在2002年国内尚无系统、完整介绍SPSS统计功能及其操作界面的书籍,前作最终将风格定位在操作字典上,使读者能够全面了解SPSS各方面的分析功能及界面操作方法,应当说前作很好地完成了这一任务。但是现在随着SPSS中文版的正式面市,软件的界面操作已不是主要问题,再来编写这种新华字典式的教材已无必要。读者自然希望能够有一个更好的教材体系出现。

2. 对于统计软件教材而言,其本质应当是统计教材,软件仅仅是实现工具。前作顺应当时的需要,主要满足的是已学习过统计理论知识,但不了解如何使用统计软件来实现的读者群的需求,因此书中并未详细阐述统计基础知识。对于统计初学者而言,需要有一本统计教材与之配合。但是在几年的使用后,现在多所高校均希望直接采用基础篇进行本科生的统计教学,而不需要和其他教材配合使用。在软件工具已经越来越易用的时候,教材的最终归宿应当是以统计知识为主线,这无疑是我们重新投入编写工作的最大动力。

综上,在充分考虑了读者们的意见后,笔者毅然决定推翻原有的框架,完全从头编写基础教程。这本新的基础教程以SPSS 12.0为准,定位为统计软件和统计学入门书籍。他针对统计初学者和SPSS初级用户的需求,以统计理论为主线,严格按照本科生统计学教材方式编写,内容共分三大部分:第一部分的任务是SPSS操作入门,讲解了软件使用和数据管理的操作知识;第二部分以统计理论为主线,详细阐述了如何在SPSS中完成数据的统计描述和参数估计,以及如何使用统计图表来进行数据的完美呈现;第三部分则详细介绍了t检验、方差分析、秩和检验、χ^2检验、相关回归等常用的基本统计分析方法,内容覆盖了目前国内大部分专业统计课程的教学范围,并结合SPSS的强大功能作了很好的扩展。书后的附录根据初学者的特点加以编制,各章后均附有参考文献和思考练习题,第三部分的章节还专门提供了本章小结,更好地满足了本科生教学的需要。另外,书中大部分表为设计表格时自动生成的。因此,大部分表及表题为英文。

除作为教材外,本书还适用于各行业中非统计专业需要使用统计方法的人员,以及希望从头学习SPSS软件的人员。我们希望广大读者能一如既往地踊跃提出自己使用中的宝贵意见和建议,使得本书再版的时候能够更上一层楼,更完美地满足大家的学习和工作需求。

张文彤
2004年7月于复旦公共卫生学院

目 录

第一部分 数据管理与软件入门

第二部分　统计描述与统计图表

第三部分　常用假设检验方法

第一部分
数据管理与软件入门

第1章 数据分析概述与软件入门

1.1 SPSS软件概述

SPSS公司总部位于美国芝加哥,创立于1975年,一直以经营统计软件产品开发为主业。1994—1998年间,SPSS公司得到了很大的发展,陆续购并了SYSTAT公司、BMDP软件公司、Quantum公司、ISL公司等,并将各公司的主打产品收纳SPSS旗下,从而使SPSS公司由原来的单一统计产品开发与销售转向为企业、教育科研及政府机构提供全面信息统计决策支持服务,成为最新的"数据仓库"和"数据挖掘"领域前沿的一家综合统计软件公司。

SPSS软件是SPSS公司赖以起家的产品,目前也仍然是该公司的主打产品之一,目前的最新版本为12.0本书也均以12.0版本为准进行讲解。SPSS的英文名称原为Statistical Package for Social Sciences,意为社会科学统计软件包。后来随着SPSS产品服务领域的扩大和服务深度的增加,SPSS公司已于2002年将英文全称更改为Statistical Product and Service Solutions,意为统计产品与服务解决方案。在近30年的发展中,虽然竞争对手不断出现,但SPSS却始终以其鲜明的特色鼎立于统计学软件之中,现在和SAS(另一种统计分析软件)被并称为当今最权威的两大统计软件。

1.1.1 软件的基本特点

SPSS得到用户广泛欢迎的原因在于SPSS强大的统计分析与数据准备功能,方便的图表展示功能,以及广阔的兼容性、界面的友好性满足了广大用户的需求,深受广大应用统计分析人员的喜爱。

1. 功能强大

⑴ 囊括了各种成熟的统计方法与模型,为统计分析用户提供了全方位的统计学算法,为各种研究提供了相应的统计学方法。如方差分析、回归分析、多元统计分析方法、生存分析方法等,方法体系覆盖全面。

⑵ 提供了各种数据准备与数据整理技术。如利用值标签来快捷地录入数据,从而为数据审核与分析提供了便利条件。生成新的变量,对连续性变量进行离散性转换;将几个小类别合并为一个大类别等。利用SPSS强大的数据整理技术,可使数据结构、内容更易于分析。

⑶ 包括自由灵活的表格功能。特别是在SPSS 11.5版本中新增的自定义表格模块(Custom Table),使得制表变得更加简单和直接。

⑷ 提供了各种常用的统计学图形,如条图、线图、饼图、直方图、散点图等多种图形,并且可

将表格图形直接拷贝到 Word文档、幻灯片中,直接进行结果的展现。

2. 兼容性好

⑴ 在数据方面,不仅可在 SPSS中作数据录入工作,还可将日常工作中常用的 Excel表格数据、文本格式数据导入 SPSS中进行分析,不仅节省了相当大的工作量,并且避免了因拷贝粘贴可能引起的错误。

⑵ 在结果方面,SPSS的表格、图形结果可直接导出为 Word、文本、网页、Excel格式等,也可以将表格、交互式图形作为对象选择性粘贴到 Word、PowerPoint等中,并在其中再利用 SPSS对它们进行编辑。

3. 易用性强

SPSS之所以有广大的用户群,不仅因为它是一种权威的统计学工具,提供了强大的统计功能,也因为它是一种非常简单易用的软件。人机界面的友好、操作的简单,使得各位统计分析人员对它青睐不已。另外,SPSS也向一些高级用户提供了编程功能,使分析工作变得更加节省时间和精力。

1.1.2 SPSS的 Client/Server结构

SPSS软件自 10.0版本以来,已发展为 Client/Server的结构体系。用户可以选择只购买单机版,也可以选择购买服务器和单机版。对于大数据量客户,可以利用 SPSS Server来解决速度慢、网络阻塞等由于数据量大而引起的问题。

在分析中使用 SPSS Server的好处在于:

⑴ 更快的分析速度。由于服务器端往往与数据仓库的物理距离更近,而 SPSS Server也对计算进行了优化,加之应用服务器的硬件配置也远高于单机端(客户端),因此对于进行大数据量分析的客户,SPSS Server可以使速度提高很多。

⑵ 缓解网络阻塞。由于数据不需要全部传送到单机端,所以网络上的数据传输量大大减少,从而缓解了网络阻塞问题。

在使用时调用 Server的具体做法是:在应用服务器端安装 SPSS Server,在单机端安装相同版本的 SPSS Client(参见图 1.1)。在单机端打开 SPSS for Windows时,选择菜单 File→ Switch Server,在如图 1.2(a)所示的对话框中指定要连接的 SPSS Server所在服务器地址,如果是第一次使用,则单击“Add”按钮,出现如图 1.2(b)所示对话框,输入服务器名或 IP地址、端口号,单击“OK”按钮,在服务器列表中出现相应的 Server地址,然后输入用户名、密码、域名,单击“OK”按钮,即可登录到 SPSS Server;此时,在 SPSS for Windows下方的状态栏中,就会显示“SPSS Processor on‘服务器名’:‘端口号’ is ready”,表示连接已经建立。

当然对于数据量不大的客户,只用 SPSS Client就可以了。现在国内绝大多数用户所说的 SPSS,实际上就是指的单机版。

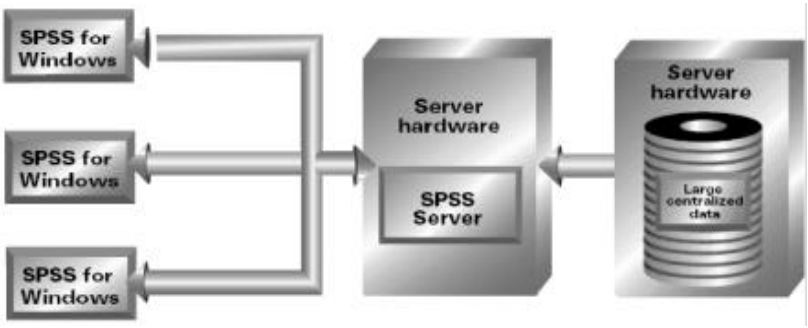
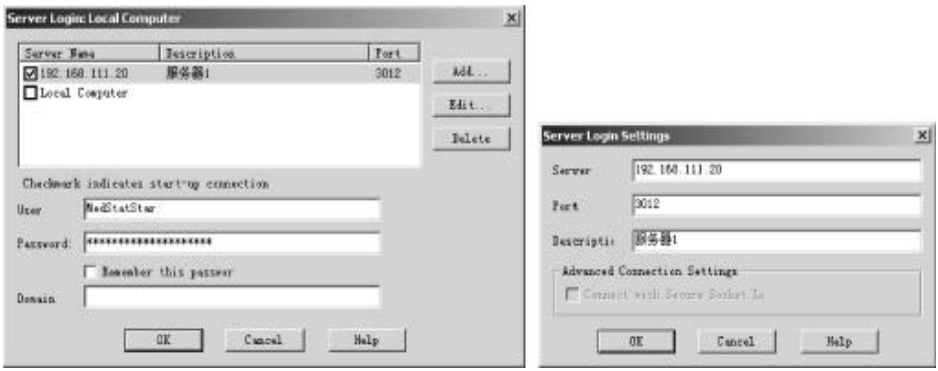


图 1.1 SPSS软件的 Client/Server体系



(a) (b)

图 1.2 SPSS Client调用 SPSS Server的对话框设置

1.1.3 SPSS的模块式结构

无论是 SPSS Client还是 SPSS Server,均是模块式结构,即它把自己的所有功能分放在多个模块上。用户可以根据分析中可能用到的数据处理和统计分析方法,自己选择适当的模块进行购买,而不必花更多的钱购买所有模块。

SPSS 12共由 11个模块构成,它们分别是:SPSS Base, SPSS Advanced, SPSS Categories, SPSS Complex Sample, SPSS Conjoint, SPSS Exact Test, SPSS Maps, SPSS Missing Value Analysis, SPSS Regression, SPSS Tables和 SPSS Trends, 其中 SPSS Base是必需的,因为 SPSS软件的整个框架、基本的数据获取、数据准备等基本功能都被集中在这个模块上,其他模块必须在 SPSS Base搭建的平台上才能工作。其他模块的功能分别如表 1.1所示。

SPSS软件通过其 License来控制模块是否安装。一个模块安装上之后,在 SPSS for Windows的菜单中就会出现相应的菜单项,所以不同客户的 SPSS for Windows的菜单可能有所不同。如果没有购买 SPSS Trends模块,软件中就不会有这样一个菜单:Analyze→ Trends;如果没有购买 SPSS Maps模块,软件中就不会有菜单:Graph→ Maps。

表 1.1 SPSS模块与功能对应表

SPSS附加模块	功 能
SPSS Advanced	一般线性模型、混合线性模型、对数线性模型、生存分析等
SPSS Categories	对应分析、感知图、Proxscal等
SPSS Complex Sample	多阶段复杂抽样技术等
SPSS Conjoint	正交设计、联合分析等,适用于市场研究
SPSS Exact Test	精确 P值计算、随机抽样 P值计算等
SPSS Maps	在地图上展示数据等
SPSS Missing Value Analysis	缺失数据的报告与填补等
SPSS Regression	Logistic回归、非线性回归、Probit回归等
SPSS Tables	交互式创建各种表格 如堆积表、嵌套表、分层表等)
SPSS Trends	Arima模型、指数平滑、自回归等

随着版本的提升,SPSS的各个模块在功能和性能上也会有一定的改进。例如,SPSS Base从11.5版本开始,提供了将结果直接导入 Word、Excel文档的功能,而在 12版本中,变量名也不再有 8字符的位数限制。又比如 SPSS Tables在 11.5版本时发生了重大变化,提供了所见即所得的表格制作功能,详见本书第 7章。SPSS Complex Sample模块则是 12.0版本新增加的内容,详见本丛书的《SPSS与市场研究》中的相关内容。

最后有一点需要澄清:国内许多 SPSS书籍因对 SPSS的功能讲解不全,总是在前言中声明所使用的是 SPSS标准版。实际上 SPSS软件,不存在所谓的标准版和专业版之分,即使安装全部的11个模块,软件也仍然是标准版。这些书籍中所谓的“标准版”,其实质只是 SPSS Base模块的相应功能而已。

1.1.4 SPSS的安装

SPSS的安装非常简单,跟随安装向导即可将 SPSS轻松安装到自己的本机。下面分别简要介绍一下 SPSS Server和 SPSS Client的安装过程。

1. SPSS Server的安装

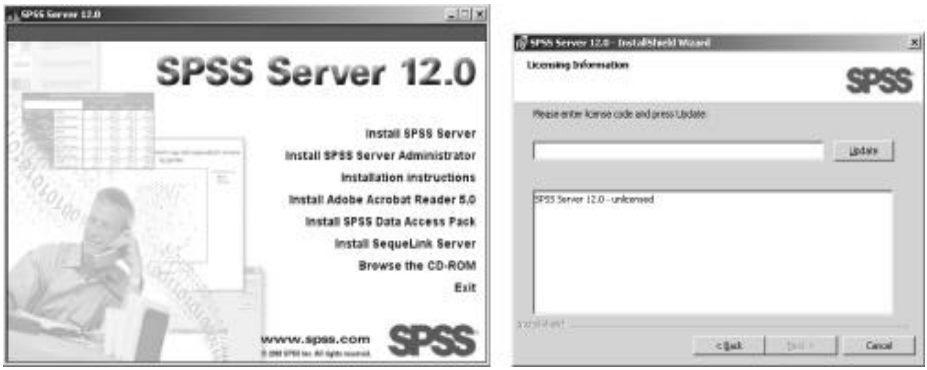
SPSS Server支持的操作平台有 AIX UNIX, HP UNIX, Linux, Windows NT等,根据不同的版本,支持平台略有不同。具体安装步骤如下:

(1) 将 SPSS Server安装光盘插入计算机光驱后,出现如图 1.3 (a)所示的界面。选中“Install SPSS Server”即进入安装向导。

(2) 跟随向导,接受 License协议,并选择安装目录,出现图 1.3 (b)图所示界面。

(3) 键入 SPSS公司提供的 License,单击“Update”按钮,即出现该 License允许安装的模块,单击“Next”按钮。

⑷ 设定该应用服务器的 IP地址和应用端口,再单击“Next”按钮,即开始安装,最后单击“Finish”按钮结束安装。



(a) (b)

图 1.3 SPSS Server安装过程中的几个视图

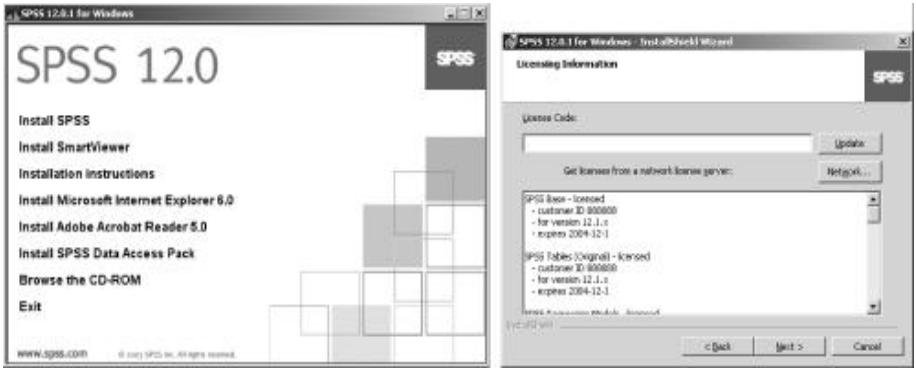
2. SPSS Client的安装

SPSS Client支持的操作平台为 Windows NT4.0,Windows 95、98、2000、XP等。安装要求约100MB~120MB硬盘 视其版本和模块而不同),内存要求为 64MB以上。具体安装步骤如下:

⑴ 将 SPSS Client安装光盘放入计算机光驱后,出现如图 1.4 (a)所示的界面。选中“Install SPSS”即进入安装向导。

⑵ 跟随向导,首先“接受 License协议”,并选择安装目录,在随后的界面中键入名称、公司和 SPSS公司提供的序列号。

⑶ 当要求输入 License时,如图 1.4 (b),键入 SPSS公司提供的 License,单击“Update”按钮,即出现该 License允许安装的模块,单击“Next”按钮,即开始安装,最后单击“Finish”按钮结束安装。



(a) (b)

图 1.4 SPSS Client安装过程中的几个视图

1.2 SPSS操作入门

1.2.1 SPSS软件的启动与退出

在 Windows开始菜单上选择开始→程序→SPSS for Windows→ SPSS for Windows,就启动了SPSS,如图 1.5所示。

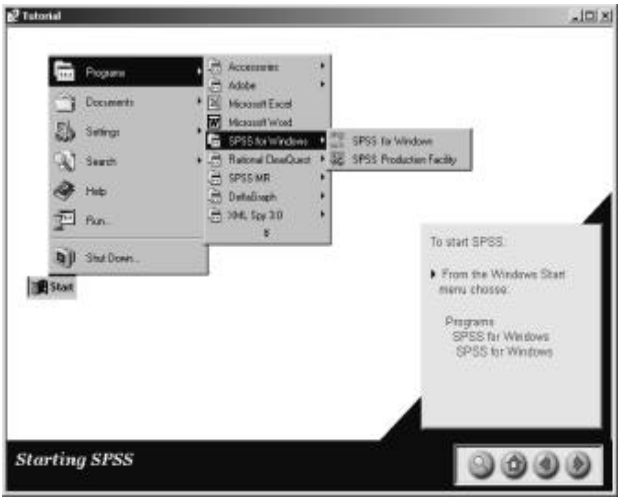


图 1.5 SPSS的启动示意图

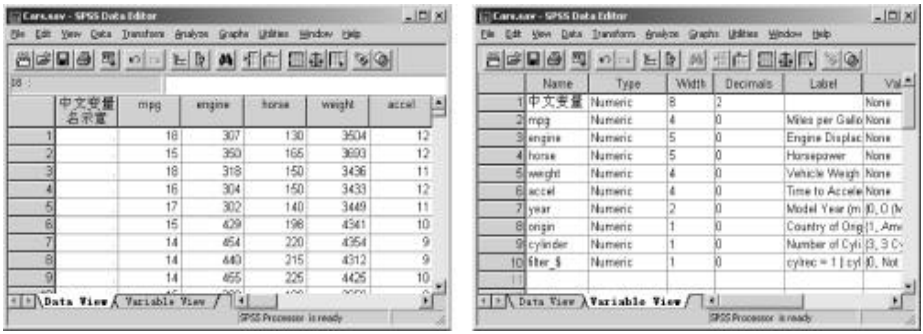
如果要关闭该软件,则选择菜单 File→ Exit;或者直接关闭窗口,即可退出 SPSS。

1.2.2 SPSS的 5个窗口

SPSS运行时使用的窗口种类最多共有 5个:数据编辑窗口、结果管理窗口、语法编辑窗口、脚本窗口和草稿结果窗口。其中数据编辑窗口和结果管理窗口是最常用到的两个窗口。实际上,这 5个窗口分别用于打开 5种格式的 SPSS文件:以“sav”为扩展名的是 SPSS的数据文件;以“sps”为扩展名的是 SPSS的语法文件;以“spo”为扩展名的是 SPSS的结果文件;以“sbs”为扩展名的是 SPSS的脚本文件;以“rtf”为扩展名的是 SPSS的草稿结果文件。

(1) 数据编辑窗口 (SPSS Data Editor):此窗口类似于 Excel窗口,SPSS处理数据的主要工作全在此窗口进行。它分为两个视图:如图 1.6 (a)所示的数据视图用于显示具体的数据,一行代表一个观测个体 (SPSS中称为 Record),一列代表一个属性 (SPSS中称为 Variable);如图 1.6 (b)所示的变量视图则专门显示有关变量的信息:变量名称、变量的类型、变量的格式等,关于变量信

息中各术语的详细解释,可参阅第 2章。

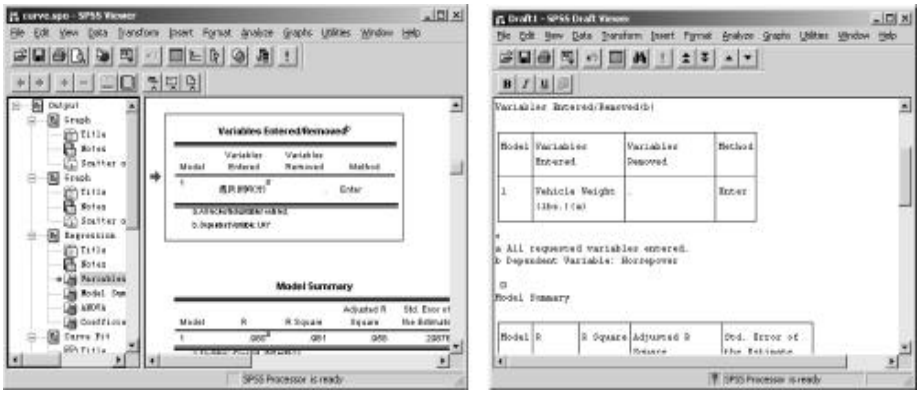


(a) (b)

图 1.6 数据编辑窗口的数据视图和变量视图

⑵ 结果管理窗口 (SPSS Output Viewer):又称结果视图,此窗口用于存放分析结果,如图 1.7(a)所示。类似于 Windows的资源管理器,整个窗口分两个区:左边为目录区,是 SPSS分析结果的一个目录;右边是内容区,是与目录一一对应的内容。

⑶ 草稿结果窗口 (SPSS Draft Viewer):草稿结果是结果的一种简化文本格式。实际上就是 Word所兼容的 rtf超文本格式,因此可以在没有安装 SPSS的 PC机上使用文字编辑软件打开,窗口形式如图 1.7(b)所示。



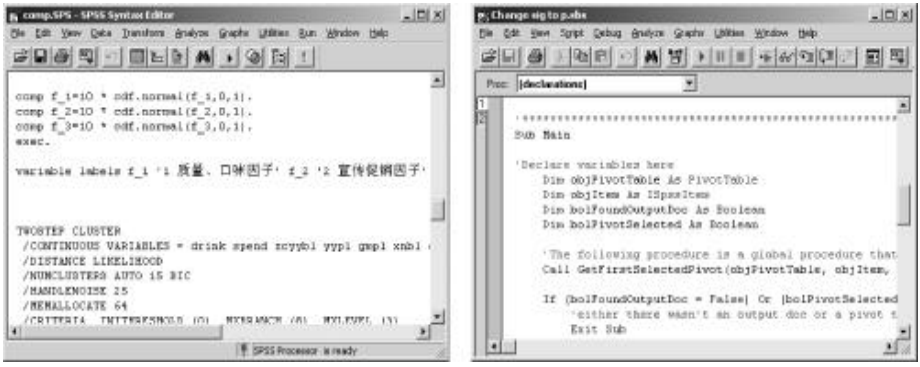
(a) (b)

图 1.7 结果管理窗口和草稿结果窗口

⑷ 语法编辑窗口 (SPSS Syntax Editor):SPSS最大的优势在于其简单易用性,即菜单一对话框式的操作,窗口形式如图 1.8(a)所示。除此之外,SPSS还提供了语法方式或程序方式进行分析。该方法既是对菜单功能的一个补充,也可以使繁琐的工作得到简化,尤其适用于高级分析人员。

⑸ 脚本窗口 (SPSS Script Editor):SPSS脚本是用 Sax Basic语言写的程序,窗口形式如图 1.8(b)所示。脚本可以像 SPSS宏一样构建和运行 SPSS命令,而且脚本可以在命令中利用当前

数据文件的变量信息,还可以对结果进行编辑或者构建一些新的自定义的对话框。脚本可用于使 SPSS内部操作自动化,使结果格式自定义化,实现 SPSS新功能以及将 SPSS与 VB和 VBA兼容应用程序连接。



(a) (b)

图 1.8 语法编辑窗口和脚本编辑窗口

启动 SPSS时,即打开了数据编辑窗口。其他窗口可以通过 File→New/Open→相应的窗口名称而打开。

1.2.3 SPSS的 4种运行方式

SPSS提供了菜单 —对话框方式的操作环境,这是最简单和最常用的运行方法。此外,SPSS还提供了程序运行方法、Include命令方法、Production Facility方法。这几种方法是菜单 —对话框方式的有益补充。下面就以 SPSS自带文件 Employee data. sav中的数据对变量“jobcat”进行频数分析为例说明这 4种运行方法。

1. 菜单对话框方式

首先打开 SPSS软件,然后选择菜单 File→Open File,如图 1.9所示,在 SPSS安装目录下打开数据“Employee data. sav”。

图 1.9 SPSS打开数据示意图

然后,利用菜单 Analyze→ Descriptive Statistics→ Frequencies,如图 1.10所示,选中“ jobcat”,单击“OK”按钮。结果管理窗口会出现如表 1.2所示结果。

图 1.10 利用对话框方式进行频数表分析

表 1.2 Employment Category

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Clerical	363	76.6	76.6	76.6
	Custodial	27	5.7	5.7	82.3
	Manager	84	17.7	17.7	100.0
	Total	474	100.0	100.0	

这里使用到了 SPSS中的对话框,现在以图 1.10所示的对话框为例对其作一简要介绍:

⑴ 列表框:上面共有两个,左边列表为待选变量(包含当前数据文件中的可分析变量或指定变量集)列表,右边列表为分析变量列表。▶为变量移动按钮,用于将变量在“待选”和“分析”列表中移动。

⑵ 5个标准按钮:几乎在任何对话框中均可见到,OK、Cancel、Help的含义非常明确,不再赘述;Reset会将对话框恢复为默认状态;Paste则会将对话框中的选择自动转化为相应的程序语句,详述参见后面相关章节。

⑶ 其他按钮和选项:根据具体功能,不同的对话框还会出现一些特殊的按钮,如本例中最下方有三个按钮,单击“Statistic”按钮会弹出有关“统计量”指定的子对话框,单击“Charts”按钮弹出有关“图形”指定的子对话框,单击“Format”按钮则会弹出有关“表格格式”指定的子对话框。

2. 程序方式

上文中提到对话框中有一个“Paste”按钮,可以将相应的操作转化为所对应的 SPSS程序,事实上,对话框可以被看成是对后台 SPSS程序的打包调用,如果将上文所做的分析使用 SPSS程序方式来分析,则应当在 Syntax编辑窗口中键入以下程序:

```
get file='C:\program files\spss\employee data.sav'.
frequencies variables = jobcat/Order = Analysis.
```

只需要选择菜单 Run→ All,运行该程序也一样会出现相同的分析结果。

对于数据不断更新而分析工作基本相同的分析人员,将常用的分析过程保存为 Syntax文件,在日后,只要在 Run和 All之间轻点鼠标,即可轻松完成繁琐的工作。无疑,这是一个一劳永逸的办法。

3. Include命令方式

当编写 Syntax程序时,如果发现将要编写的程序语句正好是另一个 Syntax文件的内容;或者发现所需的程序语句其实是几个 Syntax文件的总和时,除了可以通过"Copy"、"Paste"的方法来利用原有的资源,生成一个新的 Syntax文件外,还有一种更简单的办法,那就是使用 Include命令。例如,上面的程序如果把它保存为文件:C:\syntaxsample.sps,则以后使用时只需要用下面的一句命令即可等同于上面的整个文件:

```
Include 'C:\syntaxsample.sps'.
```

在 Syntax编辑窗口中键入上面所示的 Include语句,运行后的结果和前面相同。

4. SPSS Production Facility方式

在 Windows的程序菜单中,SPSS菜单组除了有"SPSS forWindows"项之外,还有一个"SPSS Production Facility"。这是 SPSS提供的运行分析的另一种方法,实际上是对 SPSS作了一个简单的开发,让相应的 SPSS程序在系统后台运行,直至运行完毕后才提示用户阅读结束,用户在这期间可同步进行其他工作,从而提高了工作效率。它利用的机制实质上也是 SPSS Syntax,但除此之外,它还可以通过 SPSS宏而更改 SPSS Syntax中的文件名和变量名或其他参数,使得 Syntax的应用更加灵活。

例如现在希望使用这种方式分析上面的问题,则需要利用文件 syntaxsample.sps来进行,打开 SPSS Production Facility,如图 1.11所示,随后的步骤如下:

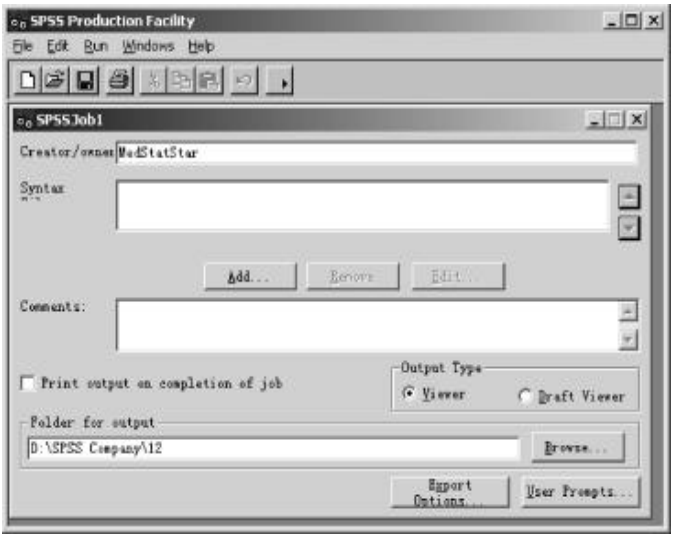


图 1.11 "SPSS Production Facility"的操作界面

⑴ 单击 Syntax框下方的“Add”按钮,到 C盘根目录下打开“syntaxsample.sps”。

⑵ 单击“Edit”按钮,对该程序进行编辑。用@ file代替 C:\program files\spss\Employee data.sav,用@ var代替 jobcat;保存后关闭。

⑶ 单击右下角的“User Prompts”按钮,添加对程序的交互分析界面,如图 1.12所示。

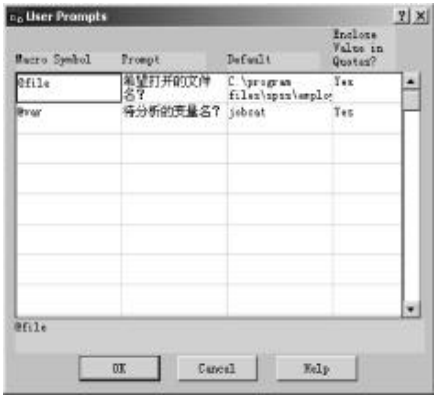


图 1.12 “User Prompts”对话框的设定

⑷ 单击“Browse”按钮指定结果保存路径,单击“ExportOptions”按钮还可以指定结果保存格式。

这样便完成了一个小工程的设定。可以单击 File→ Save,保存该工程为 SPSSJobsample.spp。

下面来运行该工程。单击 Run→ Production Job,即出现如图 1.13所示的对话框。可以按默认的指定去运行该工程,直接单击“OK”按钮,则相应程序会自动转入系统后台运行,运行完毕后会在指定路径下生成结果文件 SPSSJobsample.spo。当然也可以重新指定文件和变量名来运行该工程,这样就可以实现对任何数据中任何变量的频数分析了。

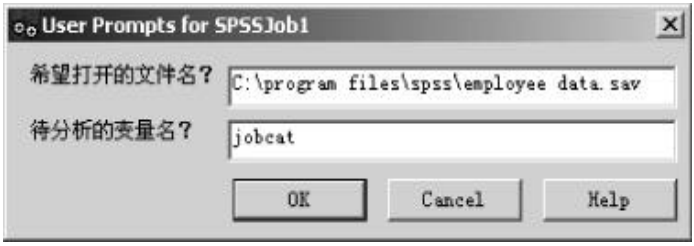


图 1.13 SPSS Production运行时弹出的对话框

1.2.4 SPSS的 4种结果输出

作为功能强大的统计分析工具,为了能够使得分析结果更为美观易读,更好地满足用户的需

求，SPSS一共提供了 4种格式的统计分析结果 表格、文本、标准图和交互图。

1. 表格格式

SPSS可以绘制表格用于表述数据，除此之外，大部分分析结果也都以专用表格的形式展示，如图 1.14所示。这些表可能是二维表，也可能是多维表。二维表、多维表都可以作为“SPSS Pivot Table”对象而粘贴到其他应用程序（如 Word、PowerPoint、Excel）中，并且依然利用 SPSS对这些表格进行编辑。SPSS的制表功能非常强大，能很好地满足用户各种情况下的需求，详见第 6、7两章。

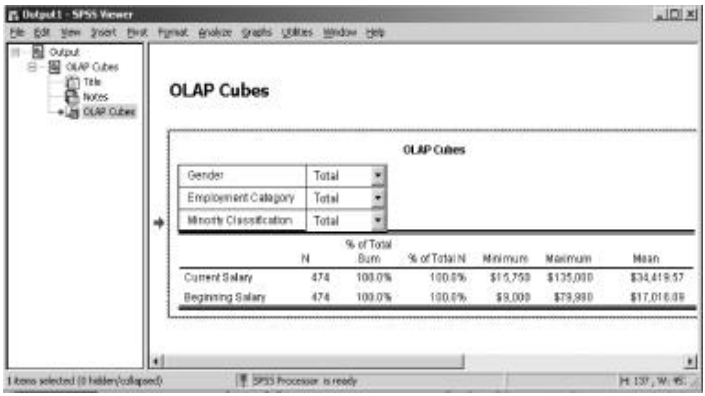


图 1.14 SPSS结果格式之一——表格格式

2. 文本格式

对于一些不便于用表格和图形表达的结果，SPSS提供了文本格式的结果，如图 1.15所示。随着版本的增加，SPSS中的文本输出已经越来越少了，例如在 12版本中，现在只有对数线性模型和 Arima模型进行变量筛选时还使用文本输出。实际上，这里的文本输出并非简单的纯文本，

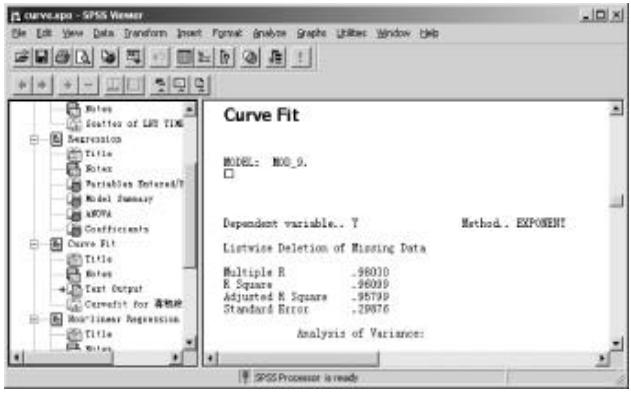


图 1.15 SPSS结果格式之一——文本格式

而是与 Office家族软件完全兼容的 rtf格式,这些文字可以随意进行拷贝粘贴、格式设定等操作。

3. 标准图与交互图

利用图形来展示数据,也是在数据分析中必不可少的。SPSS提供了两种类型的图形。一种是普通图,在 SPSS的手册中称为“标准图”如图 1.16所示;另一种为“交互图”如图 1.17所示。标准图是在 Graphs菜单下直接单击图形生成的,而交互图是在 Graphs→ Interactive下单击图形生成的。与交互图相比,标准图生成速度快,已经可以满足大部分统计绘图的需求,但可编辑能力要弱于交互图;而交互图对系统硬件环境要求更高,但可绘制的图形种类更多,编辑功能更强,尤其值得指出的是,交互图可以生成实时旋转的动态三维图。所以标准图适用于理解数据,而交互图更适合在报告演示中应用。对交互图和标准图的详细介绍参见本书第 8、9章。

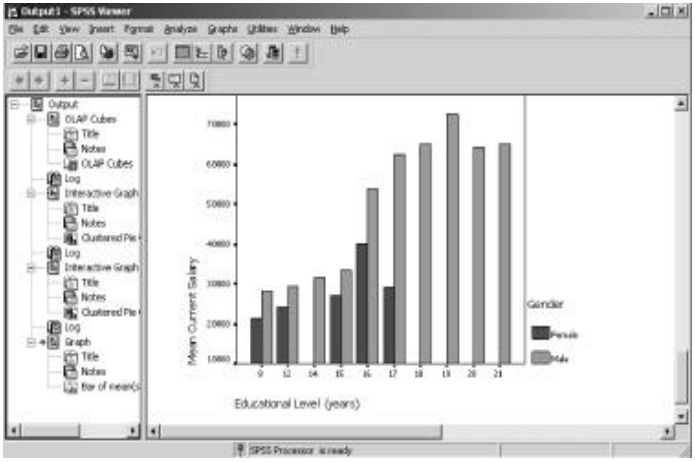


图 1.16 SPSS结果格式之一——标准图格式

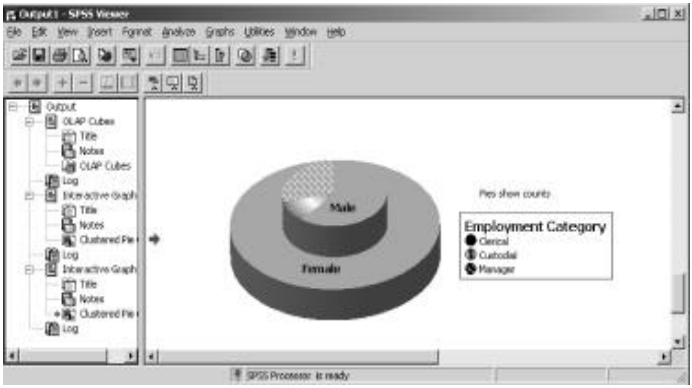


图 1.17 SPSS结果格式之一——交互图格式

4. 结果的保存和导出

SPSS的分析结果可以保存为 SPSS自身的格式：“.spo”格式（从结果编辑窗口的“File”直接点击“Save”即可），但除此之外，还可以使用导出功能存为另外几种常用的格式，具体有以下几种格式可供选择：HTML格式、Word格式、Excel格式和 Text格式。具体操作是：在结果窗口选择菜单 File→ExportOutput，出现如图 1.18所示的对话框。对话框最上方的 Export下拉列表用于选择导出的内容；右下角的 File Type下拉列表则用于选择导出格式（ExportFormat），为上述 4种格式；中部的 File框用于设定导出文件的路径和名称；而左下角的 ExportWhat框组则用于选择希望导出的内容。另外，对于标准图或交互图可以保存为常见的图形格式，如 bmp、jpg等常见格式。只需要在 ExportOutput对话框中选择 Export：Charts only，在 File Type中选择图形格式即可。

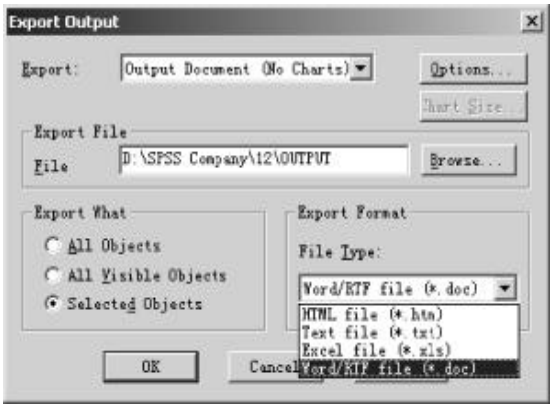


图 1.18 SPSS结果的导出选项

如果只想保存或导出部分结果，只要先选中该结果，然后在图 1.18的 ExportWhat框中选择“Selected Objects”即可。

除了可以保存结果之外，还可以将结果直接通过“Copy”、“Paste”应用到其他软件中。对于SPSS表格、交互图，还可以将它们作为“Object”粘贴到其他应用程序中。这样做有一个好处：粘贴后仍可利用 SPSS提供的功能进行编辑。具体操作是：在“开始”菜单“运行”程序文件：object on.bat（此文件在 SPSS安装目录下）。随后在应用程序中粘贴图表时均使用“选择性粘贴—SPSS Pivot Table控件或 SPSS Interactive Graph控件”即可。

1.2.5 SPSS的帮助系统

SPSS提供了无处不在的“帮助”功能，可以随时随地为不同层次的用户提供帮助。其帮助功能主要包括学习向导、帮助菜单、对话框帮助和语法手册四大类。事实上，国内有相当一部分SPSS教材都是在翻译或引用 SPSS完整而详细的帮助内容，那么绕过这些翻译，直接来见识一下原汁原味的“帮助”功能吧。

1. 学习向导

SPSS为初学者提供了非常完整和系统的自学向导,它相当于一个手把手的教练,浅显易懂地告诉用户各种基本的统计分析问题在 SPSS中是如何实现的。 SPSS中的学习向导有几种,分述如下:

(1) Statistics Coach:对于需要新手紧急完成的一些常用统计分析操作,SPSS提供了统计教练功能,它可以告诉用户为达到分析目的应选择什么统计方法,并一步步地指导用户如何进行统计分析。该模块实际上是一个编译好的交互式网页,使用起来非常舒服。Statistics Coach位于Help菜单中,选择 Help→Statistics Coach即可进入,图 1.19即为统计教练的一个界面。

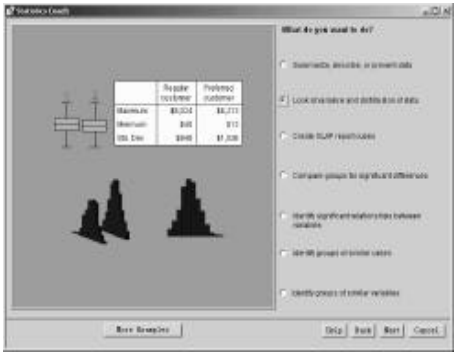


图 1.19 统计教练的界面

(2) Tutorial:同样为初学者提供,是关于某个主题的一步步指导。以示例化、图形化的方式告诉用户如何使用这个软件。初学者可以通过该教程掌握 SPSS的几乎全部常用操作 数据的输入、分析和绘图)。Tutorial模块位于 Help菜单中,选择 Help→Tutorial即可进入,起始界面为一个目录列表,即所有教程内容的索引,用户可在里面选择需要阅读的主题,如果对 SPSS完全不熟悉,则可以从最上面的 Introduction开始,它提供了使用 SPSS的一些最基本的操作教程。图 1.20为 Tutorial在演示如何编辑表格。

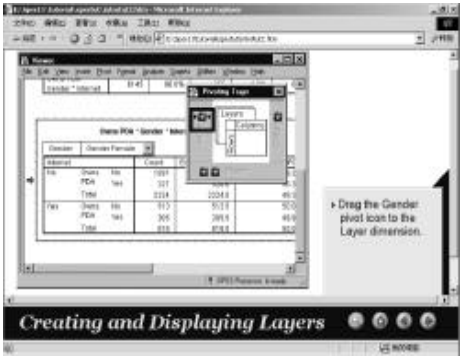


图 1.20 Tutorial在演示如何编辑表格

⑶ Results Coach:是关于结果的解释 (参见图 1.21)。在结果窗口中,只要对相应的输出含义不太清楚,即可选中该输出,并右击鼠标,右键菜单上会有 Results Coach选项,它可以链接到相应的向导界面,详细地对该过程的功能和结果加以讲解。但需要注意的是,对于少数统计上比较复杂,难以解释清楚的方法,SPSS没有提供。

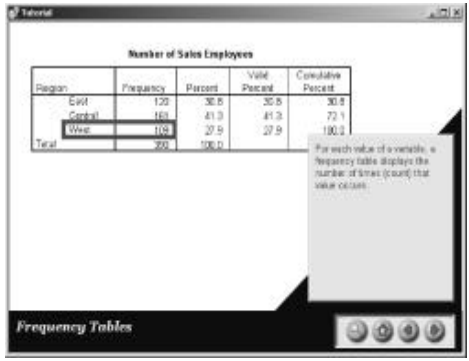


图 1.21 ResultCoach在解释频数表

⑷ Case Studies:上述三个向导多少都有一些入门和救急的味道,对于希望系统学习 SPSS中统计功能的用户而言,就可以使用 Case Studies这一详细的案例向导。用户选择菜单项 Help→ Case Studies即可进入,如图 1.22所示,它为中级用户提供了 SPSS各模块的主要分析方法的基本操作和结果解释。其讲解方式也是示例化、图形化的。只要大家的英文水平和统计功底尚可,实际上可以通过该向导掌握绝大多数的 SPSS基本操作,从而避免了到处寻找一本优秀的SPSS入门教材的痛苦。

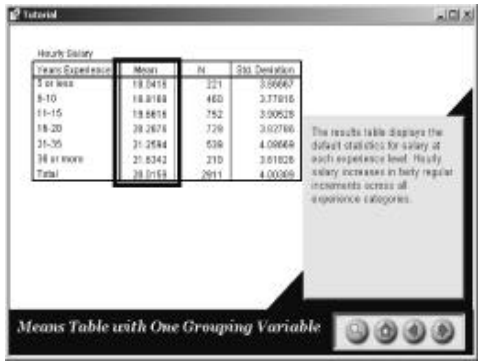


图 1.22 Case Study在演示 Means过程

2. 帮助菜单

SPSS的帮助文件就是一个标准的 Windows帮助文件,在菜单上选择 Help→ Topics进入。在使用上没有太多特殊的地方,主要也是通过目录和索引两种方式查找所需的内容。

⑴ 目录树方式:目录树像一本电子书的目录一样,将所有主题分成了一个树状结构,如图1.23所示。只要循着该目录的各级分支,最终总能找到所需的内容。用户可以在“目录”表中浏览用户手册从而学习SPSS的使用。从左边选择一个主题,如“How to read Excel 5 or later?”,右边内容区即显示此部分内容。

图 1.23 SPSS帮助主题

⑵ 索引方式:目录树的结构比较完整,但使用上要求用户首先要熟悉分类,而且要一层层找下去,如果知道希望查找的关键词,用户就可以在“索引”表中键入关键词,系统会在其左边的索引栏中寻找与键入词完全匹配的内容。如在索引栏中键入“Frequency”,左边的索引栏的第一行即显示“Frequency”,双击并选择其中一个表,即可出现内容。而当关键词不确定时也可以通过“搜索”表查询相关内容。在“搜索”栏中键入待搜索内容,单击“列出主题”,下边即列出包含该搜索内容的所有主题。

3. 对话框帮助

SPSS的界面做得非常友好,对话框界面中到处都是帮助功能。首先,在所有主对话框或子对话框中都会有Help按钮,单击Help后系统会弹出相应的帮助内容,用于解释各个选项、框组的作用是什么。除这种标准的帮助以外,任何时候如果对某个选项的功能不太熟悉,则可以直接在该选项框上方单击鼠标右键,就会立刻弹出相应的解释(见图1.24),注意此处的帮助内容并非Help菜单中相应内容的重复,一般来说要更详细些。

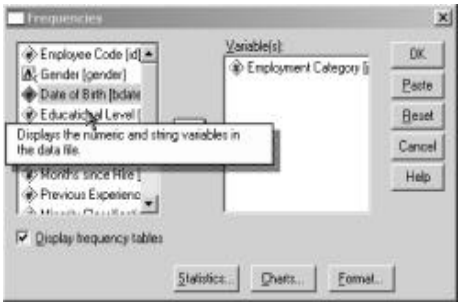


图 1.24 SPSS中的对话框帮助

4. 语法手册

当大家对 SPSS的熟悉达到一定程度时,就会发现许多操作使用对话框来做非常麻烦,甚至无法用对话框来实现。实际上,至少有 20%的高级功能是必须使用程序方式才能实现的,而且使用编程方式来完成相同的工作时,操作效率也要高得多。由于目前国内几乎没有对 SPSS编程加以深入讲解的资料,此时可以直接参考 SPSS附带的语法指南。在 SPSS的安装文件中都附送了所有模块语法指南书的 PDF格式文档,这是 SPSS官方提供的最为权威的使用指导,学会如何使用它,是最有效的学习 SPSS的方法。语法指南的调用非常简单,只要选择 Help→ Command Syntax Reference,就会自动打开相应的 PDF文档。该文档自带一个目录树,通过它就可以查找到希望学习的 SPSS过程名称,从而进行深入的学习。

1.3 数据分析概述

1.3.1 数据分析方法论介绍

任何一个数据分析项目,如果按照整个分析过程的流程结构来看,都可以被分解为大致 7个阶段:计划阶段、数据收集、数据获取、数据准备、数据分析、结果报告和模型发布阶段。下面就来对这 7个阶段做一下简单的探讨。

⑴ 计划阶段。在分析项目的初始阶段,需要花费大量的时间来设计分析计划,以减少盲目分析,避免资源浪费。在该阶段,要对数据分析的各个行动步骤作好规划,主要任务是弄清楚以下几个问题:

① 确定研究问题。从研究分析开始,就确立明确的分析目标是非常重要的。它可以帮助用户合理地计划人员、时间、资源分配,并能指导用户集中精力于研究性分析。

② 建立项目预算。

③ 确定研究范围即确定研究总体和个体。

④ 确定样本的抽取方法。

⑤ 分析评估所需样本量。

⑥ 确定数据收集方式。

⑦ 确定与研究问题相关的数据即确定应该收集个体的哪些数据。

⑧ 确定研究问题的分析方法和分析工具。

⑵ 数据收集阶段。如果手头已经有现成的数据,就可以不必再进行数据收集。数据收集的目标、方式完全取决于在上一步中所制定的计划。数据收集方式有很多种,可以是电话式访问,可以是面谈式收集,也可以是拦截式访问。如果是从头进行数据收集,则应当有一份标准问卷,问题的设计不仅要相关,还要能够从中得出有意义的结论。

⑶ 数据获取阶段。该阶段的目的是将分散的、原始格式各不相同的数据读入分析工具中,

使分析工具可以对数据进行分析。

⑷ 数据准备阶段。该阶段的主要任务是：

① 清理数据以保证数据的准确性。数据准确性是数据分析结果正确的最基本的前提条件。

② 对数据进行必要的转换。如生成一些新的字段以供分析；将连续字段离散化；将字符型字段数值化等。目的是将数据结构转换成合适的形式。

③ 填充缺失数据。对各种缺失字段，利用适当的方法进行填补。

④ 对数据进行合并、汇总等。将数据文件进行合并，将个体数据进行汇总，生成各组数据。

⑸ 数据分析阶段。利用各种数据分析方法对数据进行分析，得出结论。数据分析阶段又可以分为几个部分：

① 预分析：包括概括性统计描述和探索性统计推断两部分，前者是使用统计图和统计表对数据进行更好地理解，而后者则基于对数据的理解开始尝试进行分析，以寻找最终分析模型的雏形。具体使用的方法可以是单因素分析，也可以是简单的多因素分析。

② 精确分析：基于上一步得到的各种信息，开始尝试拟合最佳的统计模型，以寻求对数据中所蕴含信息最完美的解释。完成这一部分的工作往往需要统计知识和专业知识互相补充，而所使用的统计方法一般都是多变量方法，甚至是多元统计分析方法。

⑹ 结果报告阶段。结果报告的目的是将整个数据分析项目的结果以一种非学术化的方式表达出来，使得决策者（报告的阅读者）能够快速理解，并基于此分析结果做出决策。报告可以是文本文档、表格、图形或者是网页。

⑺ 模型发布阶段。结果报告仅仅是对基于历史数据所建立的模型加以阐述，当需要利用该模型进行预测时，具体的做法可以是在分析软件中加以预测，也可以将生成的模型编译成单独运行的控件或程序，将其模型整合到应用平台中去。该阶段的目标是将分析阶段得到的模型、信息和知识带给机构决策者以便他们能为机构做出更好的未来规划。

在大多数分析过程中，不一定会经历所有的这 7步。例如，根据分析的目的，所需的数据仅是日常工作产生的交易数据，那么就不必再经历“数据收集”阶段，而是直接进入“数据获取”阶段。另外，各阶段之间可能交叉进行。例如，有时在对原始数据进行分析之后，即进入了“数据分析”阶段，突然发现其他数据也是分析必需的，所以不得不重新返回“数据收集”阶段。又如，在“数据分析”阶段中发现某个字段因其格式不能参与分析，所以就需要再进行“数据准备”阶段所做的工作。

在一个数据分析项目结束后，可能因该项目中的新发现和对数据的新的理解，从而引发一个新数据分析项目。

1.3.2 SPSS系列产品对数据分析流程的支持

作为一家信息统计决策支持服务的提供商，SPSS公司在以上涉及的各个数据分析阶段均有相应的产品与其对应。例如，在计划阶段可以用 SamplePower来计算样本量，用 SPSS Complex Sample模块设定样本抽取计划，甚至直接抽取样本；数据收集阶段可以用 SPSS Data Entry来进行问卷设计及数据网络录入工作；数据准备阶段可以用 SPSS Base和 Missing Value等对数据进行必要的整理和修补工作；数据分析阶段是 SPSS产品的核心功能，多个 SPSS模块和 SPSS独立

软件为数据分析提供了各种统计分析方法和数据挖掘方法。而 SPSS软件提供的统计图、统计报表功能和结果输出功能则可以很好地支持结果报告阶段的需求。总之,以上提到的各阶段均可以从 SPSS公司的产品线中获得支持。而 SPSS软件本身则作为一个核心平台,在整个数据分析流程中起着关键的作用。

1.3.3 本书内容介绍

本书将以上述数据分析的 7个阶段为主线来组织内容。在本书的第 2章详细介绍了各种格式的数据如何读入 SPSS中,即数据获取阶段的内容;第 3章介绍了数据转换、合并、汇总等各种数据准备问题;第 4、5章介绍了数据分析的最初阶段,即描述性统计分析;第 6章 ~第 9章介绍了各种表格、图形的制作,而表格、图形正是分析报告阶段的必需品;第 10章 ~第 15章讲述数据分析的初级内容,即推断性统计分析的部分方法。更复杂的统计分析方法将在本套丛书的《SPSS统计分析高级教程》中讲解。至于计划阶段、数据收集阶段、结果发布阶段等往往因为会涉及具体的行业应用,不应当是统计教程的讲解内容,所以在基础教程和高级教程中都不会有太多介绍,但将会在本丛书的其他行业应用分册中涉及,感兴趣的读者可参考这些分册中的相关内容。

思考与练习

1. 试检查自己的 SPSS软件共有几个模块,其中包括了哪些功能,并思考平时的统计分析究竟要哪些模块才能够满足需求。

2. 尝试使用本章所介绍的 4种方法来使用 SPSS进行书中例题的分析,并体会这 4种方法各自的优缺点。

参考文献

1 The Basics: SPSS for Windows 10.0. SPSS Inc. Chicago, Illinois, 1999

2 Programming with SPSS Syntax and Macros (v10.0 Revised). SPSS Inc. Chicago, Illinois, 1999

3 张文彤主编. SPSS 11统计分析教程 (基础篇). 北京:北京希望电子出版社, 2002

第2章 数据录入与数据获取

数据是统计研究的基础,如果没有数据,分析也就无从谈起。用于分析的数据资料有两种,一种是原始资料,如调查问卷中的数据需要将它们录入进 SPSS软件,建立数据文件;另一种是已经被录入为其他数据格式的资料,需要将其内容直接读入 SPSS中。

针对上述的两种情况,这一章将主要介绍两个问题,即如何将数据录入进 SPSS中以及如何将其他格式的数据读进 SPSS中。对于第一个问题,根据问题类型的不同,将会从开放题、单选题和多选题的录入方式为例进行介绍;对于第二个问题,则重点介绍如何用 SPSS直接读取 Excel类型和文本格式的数据,以及如何通过 ODBC接口读取数据库文件。良好的开始是成功的一半,录入或者读入数据是数据分析的第一步工作,其重要性是不言而喻的。

2.1 数据格式概述

2.1.1 统计软件中数据的录入格式

统计软件中数据的录入格式和大家平时记录数据用的格式不太相同,SPSS所使用的数据格式也遵守这些基本的格式要求,大致的原则如下:

⑴ 不同观察对象的数据不能在同一条记录中出现,即同一观测数据应当独占一行。

⑵ 每一个测量指标或影响因素只能占据一列的位置,即同一个指标的测量数值都应当录入到同一个变量中去。

有时分析方法会对数据格式有特别的要求,此时可能会违反“一个观测占一行,一个变量占一列”的原则。这种情况在配对数据和重复测量数据中最多见。这是因为根据分析模型的要求,需要将同一个观察对象的某个观察指标的不同次测量看成不同的指标,因此被录入成了不同的变量,这是允许的。但对于统计的初学者而言,最好能够严格遵守以上规则。而且无论表现格式怎样,最终的数据集都应当能够包含原始数据的所有信息。

2.1.2 变量属性介绍

数据录入就是要把每个被访者的每个指标值录入到软件中。在录入数据时,大致可归纳为“数据录入三步曲”:定义各变量名,即给每个指标起个名字;指定每个变量的各种属性,即对每个指标的一些统计特性做出指定;录入数据,即把每个被访者的各指标取值录入为电子格式。因此这里首先介绍一下变量的各种属性问题。

任何一个变量显然都应当有变量名与之对应,但为了进一步满足统计分析的需要,除变量名外,统计软件中还往往对每一个变量定义许多附加的变量属性,如变量类型(Type)、变量宽度(Width)、小数位(Decimal)等。在上一章所讲解的数据管理窗口的变量视图中,可以看到 SPSS会为每一个变量指定 10种变量属性,但这里将重点介绍变量类型和测量尺度这两个属性,对于其他的一些属性,比如变量标签和缺失值等,会给出简单介绍,至于像变量列格式、变量对齐方式这样的属性,不用说,根据字面意思,大家也能理解其内涵。

1. 变量的存储类型

SPSS中,变量有三种基本的类型,分别是:数值型、字符型和日期型。根据不同的显示方式,数值型又被细分为了 5种,所以 SPSS中的变量类型共有 8种。Type项用于设定变量类型,选择 Type单元格时右侧会出现形如…的按钮,单击…会弹出变量类型对话框,如图 2.1所示。

图 2.1 变量类型对话框

在以上三大类变量类型中,数值型是 SPSS中最常用的变量类型。数值型的数据是由 0～9的阿拉伯数字和其他特殊符号,如美元符号、逗号或圆点组成的。如工资、年龄、成绩等变量都可定义为数值型数据。数值型数据根据内容和显示方式的不同,又可分为标准数值型(Numeric)、逗号数值型(Comma)、圆点数值型(Dot)、科学计数法型(Scientific Notation)、美元数值型(Dollar)、用户自定义型(Custom Currency)共 6种不同的表示方法。每种方法的用法根据名称的字面含义也可以猜得出来,这里不再赘述。

字符型数据类型也是 SPSS较常用的数据类型,字符型数据的默认显示宽度为 8个字符位,它区分大小写字母,并且不能进行数学运算。字符型数据在 SPSS的数据处理过程(如在计算生成新变量时)中是用一对引号引起来的。需要注意的是,在输入数据时不应输入引号,否则,双引号将会作为字符型数据的一部分。

日期型数据是用来表示日期或时间的。日期型数据的显示格式有很多,SPSS以菜单的方式列出日期型数据的显示格式以供用户选择。但事实上,SPSS中的日期型变量存储的是该时间与 1582年 10月 14日零点相差的秒数,如 1582年 10月 15日存储的就是 60×60×24=86 400,大家将变量类型变换为数值型就可以看到。但是这里只能存储正数,即 1582年 10月 14日及更早时间在 SPSS中是无效的。日期型数据主要在时间序列分析中比较有用,在较为简单的分析问题中完全可以用普通数值型数据来代替。

2. 变量的测量尺度

如果只使用变量类型,很多时候并不能准确地说明变量的含义和属性。比如说,变量“性别”,用 1代表男,2代表女。在这里,1和 2只是一个符号,没有任何数字意义。2并不比 1大,1也并不比 2小。变量“足球的喜欢程度”,用 1表示“非常喜欢”,2表示“喜欢”,3表示“一般”等,1和 2虽然也是符号,但这里有顺序之分了,1就是比 2喜欢的程度更高。如果以更喜欢为高分,那么 1就比 2大。大多少?不知道,无法衡量。再有一个变量‘薪水’,1和 2就是有区别的,2就是比 1多,多多少?多 1。同样都是 1和 2,都是数值型变量,但是它们的含义不同,适用的统

计方法也不同。如果只以变量类型来说明这个变量的属性,就不能区分出这三个变量的值 1和2彼此的区别。为了区分这三类数字,就有了变量测量尺度这个属性。

在 SPSS中使用 Measure属性对变量的测量尺度进行定义。在统计学中,按照对事物描述的精确程度,将所采用的测量尺度从低级到高级分为 4个层次:定类尺度、定序尺度、定距尺度和定比尺度。

⑴ 定类尺度(Nominal Measurement):定类尺度是对事物的类别或属性的一种测度,按照事物的某种属性对其进行分类或分组。定类变量的特点是其值仅代表了事物的类别和属性,仅能测度类别差异,不能比较各类之间的大小,所以各类之间没有顺序或等级。通常定类尺度的变量又被称为无序分类变量,如性别可取值为“男”、“女”,就是一个定类尺度的变量。对定类尺度的变量只能计算频数和频率,如在所有客户中,男性有多少人,占总人数的百分率是多少。

在 SPSS中,能使用定类尺度的数据可以是数值型,也可以是字符型变量。使用定类变量对事物进行分类时,必须符合穷尽和互斥的原则。穷尽的原则就是指每个个体都必须能归为一个类别,互斥的原则是指每个个体都只能归为一个类别。

⑵ 定序尺度(Ordinal Measurement):定序尺度是对事物之间等级或顺序差别的一种测度,可以比较优劣或排序。定序变量又被称为有序分类变量,它比定类变量的信息量多一些,不仅含有类别的信息,还包含了次序的信息;但是由于定序变量只是测度类别之间的顺序,无法测出类别之间的准确差值,即测量数值不代表绝对的数量大小,所以其计量结果只能排序,不能进行算术运算。定序变量同定类变量一样,其数据可以是数值型,也可以是字符型变量。定序变量除可以计算频率之外,还可以计算累计频率。如足球喜欢程度这一变量的取值有:1——非常喜欢,2——喜欢,3——无所谓,4——不喜欢,5——非常不喜欢,这是一个定序尺度的变量。对它就可以计算累计频数和累计频率。如对“足球喜欢程度”,不仅可以计算喜欢的人数和比例有多少,还可以计算喜欢及非常喜欢的累计人数和比例有多少。

⑶ 定距尺度(Interval Measurement):定距尺度是对事物类别或次序之间间距的测度。定距变量的特点是其不仅能将事物区分为不同类型并进行排序,而且可准确指出类别之间的差距是多少;定距变量通常以自然或物理单位为计量尺度,因此测量结果往往表现为数值,所以计量结果可以进行加减运算。

⑷ 定比尺度(Scale Measurement):定比尺度是能够测算两个测度值之间比值的一种计量尺度,它的测量结果同定距变量一样也表现为数值,如职工月收入,企业销售额等。其与定距变量的差别在于有一固定的绝对“零点”,而定距变量则没有,定距变量中的“0”并不表示“没有”,仅仅是一个测量值,而定比变量中的“0”就真正表示“没有”。比如温度,0℃只是一个普通的温度,并非没有温度,因此它只是定距变量,而体重则是真正的定比变量。定比变量是测量尺度的最高水平,它除了具有其他三种测量尺度的全部特点外,还具有可计算两个测度值之间比值的特点,因此它可进行加、减、乘、除运算,而定距变量只可进行加减运算。

SPSS中默认的变量测量尺度就是定比尺度。但由于后两种测量尺度在绝大多数统计分析中没有本质上的差别,在 SPSS中就将其合并为一类,统称为“Scale”测量。

这三种尺度在许多统计书籍中会有更为通俗的称呼:无序分类变量、有序分类变量和连续性变量。从实用的角度出发,本书将同时采用这两种命名体系。

在这 4种测量尺度之间,按照信息量的高低,可将高层次测量尺度的测量结果转换为低层次

测量尺度的测量结果,但这样会损失一部分信息。不能将低层次的测量尺度转换为高层次测量尺度的结果,这样可能会引入错误的信息。

3. 变量名与变量值标签

除了上边介绍的变量类型和测量尺度外,变量的其他属性是不是就没用了呢?回答当然是否定的。其他的属性仍然很重要,比如,Label项用于定义变量名标签,对变量名的含义进行进一步解释说明,该标签会在结果中输出以方便阅读,增强变量名的可视性和统计分析结果的可读性。另外,Values项也是一个不得不提的选项,用于定义变量值标签(见图 2.2),变量值标签是对变量取值含义的解释说明信息。例如对于性别数据,假设用 1表示男,用 2表示女,如果在录入数据时数据集中没有设定变量值标签,其他人就很难弄清楚是 1表示男还是 2表示男。因此,变量值标签对于定序变量(如职称)和定类变量(如民族、性别)来说,是必不可少的,它不但使定类和定序变量的数据录入变得更加方便,且明确了数据的含义,也同样增强了分析结果的可读性。

图 2.2 变量值标签对话框

变量值标签对话框上部的两个文本框分别为变量值输入框和变量值标签输入框,分别在其中输入“1”和“男”,此时下方的 Add按钮变黑,单击它,该变量值标签就会被加入下方的标签框内。与此类似定义变量值“2”为“女”,最后单击 OK按钮,变量值标签就设置完成。此时做任何分析,在结果中都有相应的标签出现。如果现在就想看效果,切换回 Data View界面,然后选择菜单 View→ Value Labels,就会看到上述结果。

另外,SPSS在 12.0版本以前,对于变量名有一个限制,即要求变量名限制在 8个字符之内。但令人欣喜的是,从 12.0版本开始,此限制已经被取消,变量名最多可以有 64个字符。当然,出于兼容性的考虑,变量名的定义还有一些限制,即不能以数字开头,中间不能有空格,一个数据文件中不能有相同的变量名等。读者只要在使用中尝试即可,不必记那么多规则。

4. 缺失值

Missing项是一个重要而且容易被忽视的选项,它用于定义变量缺失值。SPSS中缺失值有用户自定义缺失值和系统缺失值两大类。对于数值型变量的数据,系统缺失值用一个圆点“.”表示,而字符型变量默认就是空字符串。如果在问卷调查中,有些数据项漏填了,则数据录入时只能跳过,相应的数据单元格就会被系统自动当作缺失值来处理。

另外一类缺失值是用户自定义缺失值,这往往出现在一些设计较严格的大型调查中,在一些题项处会给出一个选项:不知道或拒答。相应的代码可能用 9或者 99来表示。显然,这里的 99不是一个真实的答案,仅仅是缺失值代码,需要告知 SPSS这个特定的标记数据,以在进行统计分析时区别对待缺失值和正常的分析数据。具体做法为单击相应变量 Missing框右侧的省略号,会弹出缺失值对话框如图 2.3所示,利用该对话框,用户可以自定义缺失值。界面上有一列三个单选钮,默认值为最上方的“无自定义缺失值”,第二项指定离散的缺失值(Discrete Missing Values),最多可以定义 3

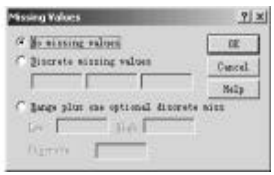


图 2.3 缺失值对话框

个值 最后一项,指定缺失值所在的区间范围,并可同时指定一个离散值。

其他的变量属性,即使不作讲解,大家也可以根据 SPSS界面的提示做出正确的选择,这里就不再详述了。但是有一点要强调的是,就数据录入这部分内容而言,变量属性的设置是最重要的一部分工作,属性的设置不仅涉及对错,而且还有一个设置好坏的问题,属性设置得好,会简化后边的数据分析工作,所以读者不可小看这部分工作。

2.2 数据的直接录入

在 SPSS中,新建一个数据文件非常容易。只要打开 SPSS,系统就已经生成了一个空数据文件,用户只要按自己的需要定义变量、输入数据,然后保存即可。

2.2.1 操作界面说明

初次进入 SPSS系统时会出现一个导航对话框,单击右下方的 Cancel按钮,即可进入 SPSS的主界面,如图 2.4所示。从窗口顶部的“SPSS Data Editor”可以看出,现在所看到的是 SPSS的数据编辑窗口。这个窗口是一个典型的 Windows软件界面,第一次使用 SPSS也会觉得很亲切,从中可以看到菜单栏、工具栏,在 SPSS的工具栏下方的是数据栏,数据栏下方则是数据编辑窗口的主界面。该界面由若干行和列组成,每行对应一条记录,每列对应一个变量。由于现在没有输入任何数据,所以行、列的标号都是灰色的。注意第一行第一列的单元格边框为深色,表明该数据单元格为当前单元格。

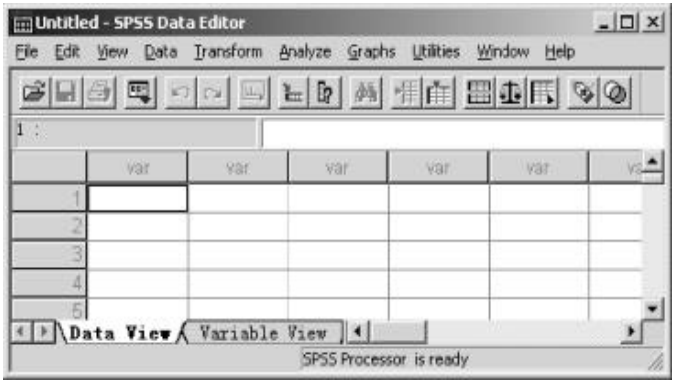


图 2.4 SPSS的数据编辑窗口

在这个界面的左下角,可以看到“Data View”和“Variable View”的标签,现在图中显示的是数据视图,如果点击右边的“Variable View”,就进入变量视图。前面提到的变量属性的设置都在变量视图中进行,而数据的录入工作则应当在数据视图中直接通过键盘完成。

2.2.2 开放题和简单单选题的录入

根据调查问卷中设计问题的类型的不同,定义变量的方式也不同。通常调查问卷中的问题包括单选题、多选题和开放题等几种,所以,下文将分别就这三种类型题目的录入方式加以介绍。为了更好地对此加以说明,这里以这样一份简单的问卷来做例子。

1. 序号:
2. 性别: 1男 2女
3. 姓名:
4. 家庭月收入:. 3 000以下 b. 3000~4 999 c. 5000~6 999 d. 7 000~9 999
   e. 10 000及以上
5. 出生年月日(mm/dd/yyy):
6. 婚姻状况:a 未婚 b. 已婚 c 丧偶 d. 离异
7. 你在选择购物商场的时候,关注以下哪些因素:
   a. 交通条件 b. 促销活动 c. 购物环境 d. 服务质量 e.其他
8. 请问你购物的打折信息主要来自以下哪些渠道 (限选 3项):
   a. 报纸 b. 杂志 c 电视 d. 收音机 e. 网络 f 朋友介绍 g 手机短信
   h. 其他,请指出_____
9. 每天上网的小时数:_____小时

在这份问卷中,包含了开放题、单选题和多选题,其中第 1、3、5、9是开放题,题 1、9是数值型开放题,3是字符型,5是日期型;第 2、4、6题是单选题,第 7、8题是多选题。其中,第 8题有一些特殊,将在后文中说明。下文将分别就这三种类型题目的录入方式加以介绍。

1. 在 SPSS中定义变量

由前文可知,录入数据的第一步是定义变量属性,随后才能进行数据录入。虽然在空白的变量列中直接输入数据,SPSS会自动给该列给定一个变量名,但是这样往往不能完全满足用户的需要,所以还是首先来定义需要使用的变量吧。

定义变量属性,首先要定义变量名,变量名是变量的唯一标识,前边已经讨论过相关的知识,这里不再重复,在前 4行的 Name列中直接输入变量名——“id”、“name”、“born”、“net”,大家同时可以看到 SPSS会在变量类型等列自动填入默认值。

在绝大多数情况下,SPSS给出的默认数据类型和数据精度可以满足需要,如果默认值满足分析的需要,变量定义到此就可以结束了,否则就需要对不满足条件的选项进行进一步的设置。在本例中,变量“id”是被访者的记录号,它的测量尺度应该是定类尺度——“Nominal”。但值得指出的是,因为变量“id”只是方便检查和核对问卷,不参与后边的数据分析工作,所以,要求不严格的情况下,此处的变量类型可采用默认形式不做修改。此外,变量“name”是被访者姓名,应是字符型变量,这里应当将“Type”中的“Numeric”改成“String”。同理,变量“born”代表出生日期,应当更改为日期型数值“Date”。在对变量类型作修改的同时可以看到,变量的其他属性也会自

动进行相应的修改,如图 2.5所示。

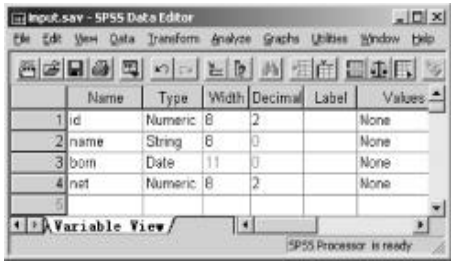


图 2.5 变量定义

引例中的开放题的定义是比较简单的,但是大多时候,开放题的答案可能是一个句子或者一段话,此时要更改该变量的 Width,因为默认的 8个字符的宽度只能存放 4个汉字,要根据该变量可能出现的最大字符长度来确定 Width (最大不超过 256个字符)。

现在切换回数据视图,数据编辑窗口如图 2.6所示。可见前 4列的名称均为深色显示,就是刚才定义的内容,表明这 4列已经被定义为变量,其余各列的名称仍为灰色的“var”,表示尚未使用。同样地,各行的标号也为灰色,表明现在还未输入过数据,即该数据集内没有记录。在变量定义完毕后,就可以向这个文件中录入数据了。

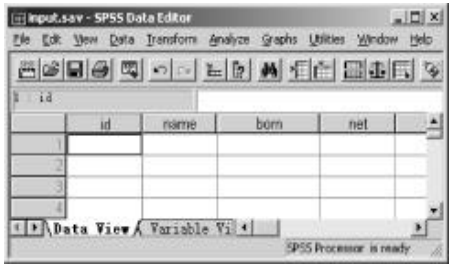


图 2.6 定义好变量的数据编辑窗口

2. 开放题的录入

单选题和开放题的录入方式很相似,在本部分内容中,将首先以问卷中的 1、3、5、9为例来介绍开放题的录入方式,然后说明单选题的录入方式,最后总结二者的区别。

现在开始录入数据,首先来输入变量 id的值,首先确认一行一列单元格为当前单元格,弃鼠标而用键盘,输入数据 1,此时界面显示如图 2.7所示。

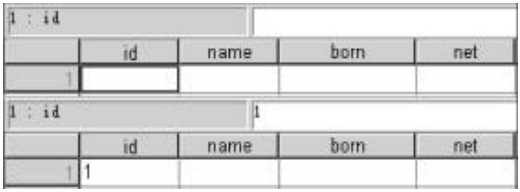


图 2.7 录入数据过程 (一)

注意 在回车之前，输入的数据在数据单元格内左对齐显示，表示该单元格为第一次录入数据，同时数据栏内同步显示出输入的数值。现在回车，界面如图 2.8所示。

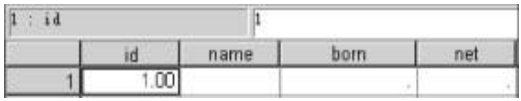


图 2.8 录入数据过程（二）

图 2.8和前面的图形相比，发生了以下变化。首先，当前单元格下移，变成了二行一列单元格，而一行一列单元格的内容则被替换成了 1.00。出现两位小数是因为数值型变量默认为两位小数（由于序号只会是整数，可以将 Decimal设为“0”）；其次，第一行的标号变黑，表明该行已输入了数据；第三，一行二列单元格（字符型变量）因为没有输入数据，显示为空，一行三列和四列单元格（数值型变量）因为没有输入过数据，显示为“.”，这代表该数据为缺失值。用类似的输入方式将数据录入完毕，此时数据编辑窗口如图 2.9所示。

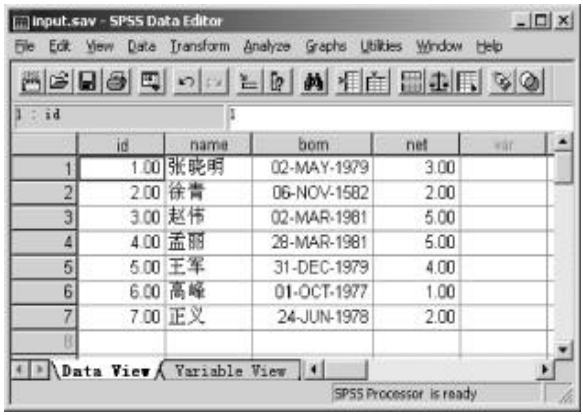


图 2.9 数据录入完毕的窗口

此外，有一点不得不提醒大家，在数据录入过程中，要随时注意保存，如果突然断电或者死机，辛苦工作的成果将付之东流。

3. 单选题的录入

单选题的录入方式与开放题类似，不同的是，单选题中可以定义变量值标签，通过这种方式既可以减少数据录入的工作量，还方便了后边的数据分析工作。具体而言，单选题的录入可以采用字符直接录入、字符代码 +值标签、数值代码 +值标签三种方式。对应这三种录入方式，变量“gender”定义后的界面参见图 2.10。

对于这三种录入方式，原则上都是可以选择的；但是第三种录入方式“数值代码 +值标签”（参见图 2.11）方便了后边的分析工作，推荐读者使用第三种录入方式。

再来看一下对于“收入”的定义，变量“income”为定序型变量，值标签中对变量取值的含义进行了说明，参见图 2.11。

图 2.10 单选题的三种录入方式说明

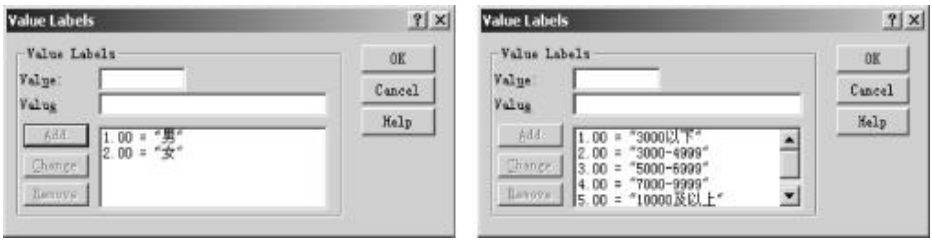


图 2.11 “数值 +值标签”录入方式

如果问卷数据中有含“其他,请指出”选项的单选题,则在录入时可以使用两个变量对其进行定义,在第一个变量中,“其他,请指出”作为选项中的一个可进行选择;第二个变量将“其他,请指出”看作一个独立的开放题,按照开放题的录入方式进行数据录入,将没有选择该选项的被访者作为缺失值处理。

2.2.3 多选题的录入

多选题,又被称为多重应答(Multiple Response),是在社会调查和市场调研中极为常见的一种数据记录类型。通常,问卷中的一个单选题问题对一个被访者只能取一个值。多选题,顾名思义,就是对应一个被访者,一个问题可以取多个值,比如引例中的第 7、8题如下:

> 7. 你在选择购物商场的时候,关注以下哪些因素:
> a. 交通条件 b. 促销活动 c. 购物环境 d. 服务质量 e. 其他
> 8. 请问你购物的打折信息主要来自以下哪些渠道(限选 3项)
> a. 报纸 b. 杂志 c. 电视 d. 收音机 e. 网络 f. 朋友介绍 g. 手机短信
> h. 其他,请指出_____

这是两个典型的多选题,被访者可以选择一个因素,也可以选择两个或者多个,虽然在问卷中这只是一个问题,但实际上答案可以是多个。由于在多选题中每道题都可能有一个以上的答

案,因此多选题不能被直接编码,需要使用几个变量来进行记录。在 SPSS中,常见的方法有两种,多重二分法 (MultipleDichotomyMethod)和多重分类法 (MultipleCategoryMethod)。下文将进行详细说明。

1. 多重二分法

所谓多重二分法,是指在编码的时候,对应每一个选项都要定义一个变量,有几个选项就有几个变量,这些变量均为二分类 (二分类变量是指该变量只有两个取值,此例中这些变量只有两个取值:"选中"与"未选"),它们各自代表对一个选项的选择结果。

在 SPSS中对多选题进行数据录入与单选题的录入程序相同,均是首先在变量视窗进行变量定义,然后直接录入数据,多选题所不同的是变量的定义方式不同,而且,数据录入完毕,在分析之前,还需定义多选题集。

首先来定义变量。每个选项对应一个变量,比如上文的例子,对应第 7题选择商场的 5个因素,定义 5个变量,因为 SPSS12已经取消了对"变量名只能 8个字符长"的限制,所以可以根据自己的习惯和偏好选择是取一个长而详细的变量名,还是取个简短的名字,然后在变量标签中对变量的含义进行说明。很显然本例选择了后者,见图 2.12。

还有一点要说明的是,变量值标签的定义应该一致,即这 5个变量的编码方式应该相同,在这个例子中,"1"和"0"所代表的含义应该一致。比如这个例子,对应选择商场的 5个因素,定义 5个变量,每个变量都是二分类,1代表选择,0代表未选。将数据录入 SPSS中,格式如图 2.12所示。

交通条件	促销活动	购物环境	服务质量	其他
1.00	1.00	.00	.00	.00
.00	.00	1.00	1.00	.00
.00	.00	1.00	1.00	1.00
1.00	1.00	1.00	.00	.00
1.00	.00	1.00	1.00	.00
1.00	1.00	.00	1.00	.00
1.00	1.00	.00	.00	1.00

图 2.12 多重二分法数据录入格式

从图 2.12可以很明显地看出,每个变量都对应一个选项,第一个被访者在这道题的选项中选择了"交通条件"和"促销活动"两项,第二个被访者选择了"购物环境"和"服务质量"两项。那么如果选项过多,比如 20个选项,要求被访者选出最关注的 5个,显然,绝大部分被选中的频率都会较低,使用多重二分法录入,则大部分数据都是 0,不仅增加了数据录入的工作,而且不利于进行分析,这时不适合使用二分法进行数据录入,需使用下文将要介绍的多重分类法。

2. 多重分类法

多重分类法,也是利用多个变量来对一个多选题的答案进行定义,应该用多少个变量,由被访者实际可能给出的最多答案数而定。而且,这些变量须为数值型变量,利用值标签将答案标出,所有变量采用一套值标签。之所以称它为多重分类法,是因为每个变量都是多分类的,每个

变量代表被访者的一次选择。多重分类法适合问题的选项较多的情况,尤其适合于“请在下列选项中选出您最喜欢的几个选项”一类的问题。例如在问卷的第 8题中,研究者希望了解目标人群主要通过什么渠道得到消费信息,在问卷中列出了 8个选项,让被访者从中选择他认为最主要的几个。此时一般都会采用多重分类法的格式来记录数据,如图 2.13所示。图中共有三个变量,均为多分类,各代表被访者的一次选择,即记录的是被选中渠道的代码。注意图中第 6条记录只填入了两个渠道,也就是说该被访者只选出了两种渠道。显然,这种“数据缺失”的现象在多重分类法中其实是一种正常情况。

source1	source2	source3
2.00	4.00	7.00
1.00	2.00	3.00
5.00	6.00	8.00
2.00	6.00	7.00
1.00	4.00	7.00
3.00	7.00	.
2.00	5.00	6.00

图 2.13 多重分类法的数据格式

3. 多选题录入在 SPSS中的实现

在进行多选题录入时,只需要将相应的变量设定好即可进行操作,但是录入完毕后对多选题进行分析前,首先需要定义多选题集,然后才可以把多选题的全部变量当作一道题目来进行分析。在 SPSS中提供了专门的菜单用来对付多选题,Tables模块和 Multiple Response菜单都可以用来设定多选题变量集。所不同的是,Multiple Response菜单中的 Define Sets过程定义多选题变量集的信息不能在 SPSS数据文件中保存,关闭数据文件后相应信息就会丢失,如果再次使用,则必须重新加以定义,而 Tables模块可以保存定义的信息。所幸的是这两个过程的操作是基本相同的,现在就以 Define Sets过程为例来看一下是如何定义多选题集的。在 SPSS中选择 Analyze→Multiple Response→Define Sets,打开定义多选题集的对话框,界面如图 2.14所示。在该对话框中,需要注意以下几点:

图 2.14 定义多选题变量集

(1) Variables in Set框:选入需要加入同一个多选题变量集的变量列表,对于多重二分类法录入的多选题,这些变量必须为二分类,并按照相同的方式来编码(如都用 1代表选中)。对于

多重多分类法录入的多选题,这些变量须为多分类,并共用一套值和值标签。

⑵ Variables Are Coded As单选框组:选择变量的编码方式。Dichotomics即为多重二分法编码方式,counted value是指用哪个数值表示选中。Categories指变量为多重分类法编码方式,此时需要设定取值范围,在该范围内的记录值将纳入分析。

⑶ Name框:键入多选题变量集的名称,在此定义的变量集名为 ques7,当然在 SPSS 12中也可以定义很长的中文变量名。下方的 Label框可以为相应的多选题变量集定义一个名称标签,如同本例中所见。

另外,对于形如问题 8一样的多选题,即含有“其他,请指出”答案的附加内容的问题,也是先把其他算作一个答案选项,而用另一个变量来表示其他的内容。在数据录入完毕后再对附加内容根据频次高低进行编码,以进行更为深入的分析。

2.3 外部数据的获取

对于 SPSS格式的数据,只要点击 File→Open→Data,选择文件路径和文件名打开即可。但如果数据不是 SPSS格式的,是否可以直接读入 SPSS,用 SPSS进行分析呢?回答是肯定的。SPSS可以读入许多非 SPSS默认类型的数据文件,方式主要有三种:直接打开,利用文本向导读入文本数据以及利用数据库 ODBC接口读取数据。对这三种方法,下文将以常见的 Excel格式的数据、文本数据和 Access数据为例,介绍 SPSS获取数据的功能。

2.3.1 电子表格数据如何导入 SPSS中

SPSS中可以直接读入许多常用格式的数据文件,选择菜单 File→Open→Data或直接单击快捷工具栏上的快捷按钮,系统就会弹出 Open File对话框,单击“文件类型”列表框,在里面能看到可以直接打开的数据文件格式,包括如表 2.1所示的 16种类型。

表 2.1 SPSS可以直接打开的数据类型

数据标识	数据类型
SPSS (*.sav)	SPSS数据文件 (6.0版 ~12.0版)
SPSS/PC+ (*.sys)	SPSS 4.0版数据文件
Systat (*.syd)	*.syd格式的 Systat数据文件
Systat (*.sys)	*.sys格式的 Systat数据文件
SPSS Portable (*.por)	SPSS便携格式的数据文件
Excel (*.xls)	Excel数据文件 (5.0版 ~2000版)
Lotus (*.wk*)	Lotus数据文件
SYLK (*.slk)	SYLK数据文件
dBase (*.dbf)	dBase系列数据文件 (dBaseII ~IV)

续表

数据标识	数据类型
SAS Long File Name (*.sas7bdat)	SAS 7～8版长文件名类型数据文件
SAS Short File Name (*.sd7)	SAS 7～8版短文件名类型数据文件
SAS v6 for Windows (*.sd2)	SAS 6版 (for Windows)数据文件
SAS v6 for UNIX (*.ssd01)	SAS 6版 (for UNIX)数据文件
SAS Transport (*.xpt)	SAS便携格式的数据文件
Text (*.txt)	纯文本格式的数据文件
Data (*.dat)	纯文本格式的数据文件

选择所需的文件类型,然后选中需要打开的文件,SPSS就会按照要求打开相应的数据文件,并自动转换为 SPSS格式。

下面以 SPSS自带的文件 demo.xls为例,来看 SPSS如何直接读取这个文件,该文件位于 SPSS目录下的 Tutorial\sample_files子目录中。首先在 Excel中打开 demo.xls,了解一下这个文件的结构,重点需要了解这样几项内容:第一,该文件中包含几个数据表,具体应当打开哪个表;第二,如果不需要该表的所有数据,而只需读入一部分,这时需要了解要读入的数据的精确位置——如单元格 A2:F5。第三,此部分数据的第一行是否是变量名。在这个文件中,很明显可以看出,第一行是变量名,该文件只有一个表,要读取的是该表单中的全部数据。

第一步,在 Open File对话框中,选择路径 (此例中为 SPSS\Tutorial\Sample files),选择文件类型"Excel(*.xls)",文件列表中出现所有的 Excel文件,点击文件 demo.xls。第二步,弹出对话框,如图 2.15所示。Worksheet框中指定哪张表;Range框中指定读取的数据具体位置,用单元格的起 (左上角单元格名称,如 A2)止 (右下角单元格名称,如 F5)位置来表示,中间用冒号":"隔开;"Read variable names from the first row of data?"意为"该单元格范围的第一行是变量名吗?"。指定完毕,点击"OK"按钮之后,数据顺利地读入了 SPSS。

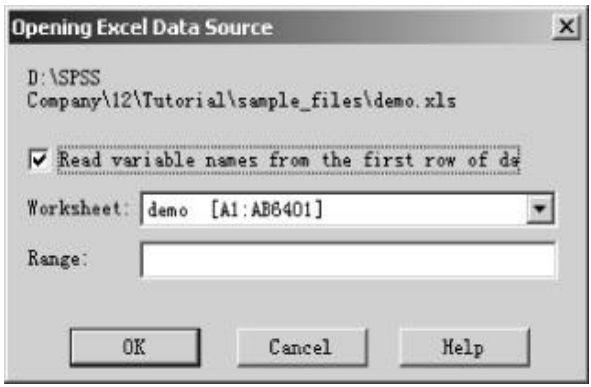


图 2.15 打开 Excel数据文件对话框

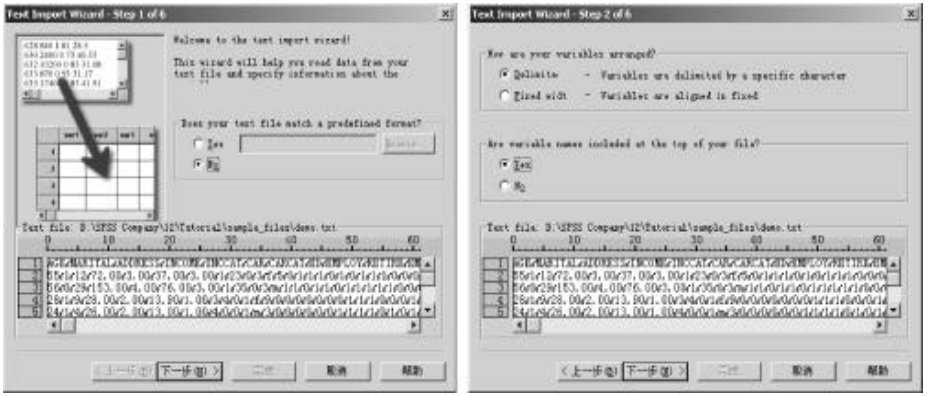
这种直接读取的方法要优于“拷贝 +粘贴”,它不仅可以顺利地进行变量名的转化,最重要的是,它可以直接读取字符型变量,若用“拷贝 +粘贴”的方法,字符型变量就全部变成缺失值了。并且操作简单,不容易出错,就和读取 SPSS自己的文件一样方便。

在上面的例子中只需要读取一个表单的数据,如果需要将两个或者多个 Sheet放在一个数据文件中,是否仍然像读取单个 Sheet文件那样轻松方便呢?回答是肯定的。有两种方式可以实现这一要求,第一种是打开两个 SPSS窗口,分别读取两个 Sheet,然后使用 Merge命令 (详见第 3章)对两个文件进行合并;第二种方式是使用前文的方式,首先读取其中的一个 Sheet,并保存,然后直接从该文件读取另一个 Sheet,实现 SPSS和 Excel的合并。

2.3.2 文本数据如何导入 SPSS中

SPSS可以通过两种菜单操作方式读取文本数据,一种是,选择菜单 File→Read Text Data;另一种是,选择 File→Open→Data,这两种情况是一样的,系统会弹出 Open File对话框,只是前者文件类型自动跳到了 Text(*.txt),后者需要在文件类型下拉菜单中作选择。之所以在菜单上保留“Read Text Data”条目有两个原因:① 读入纯文本的情况非常普遍,放在这里更加醒目;② 为了和 SPSS老版本在菜单上保持兼容。

这里以系统自带的文件“demo.txt”为例来说明如何将文本数据导入 SPSS中。与读取 Excel数据一样,首先打开该数据,观察这个数据的基本结构,如变量间是固定宽度,还是用某种分隔符区分,第一行是否为变量名等。然后关掉这个文本文件,打开 SPSS软件。首先,在 Open File对话框选中相应的文件名并单击“确定”,系统会自动启动文本导入向导对话框如图 2.16 (a)所示,从对话框标题可以看到该向导共分 6步,下面一步步地讲解。



(a) (b)

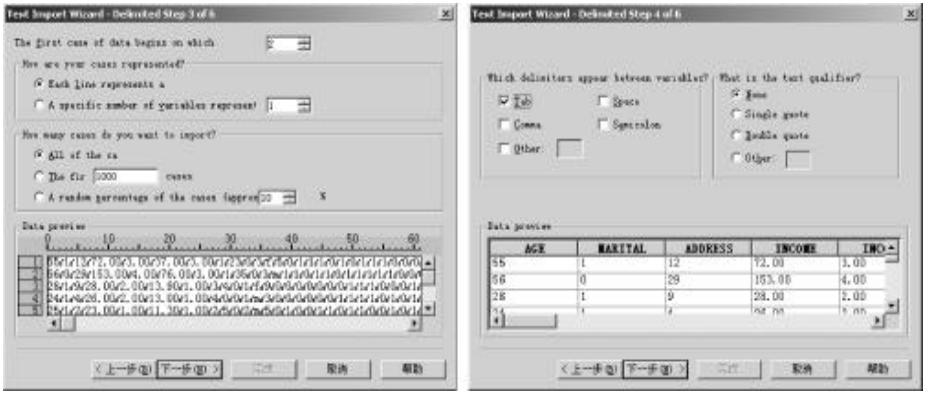
图 2.16 文本导入向导的第一、二个对话框

第 1步 中部为一对单选钮,问题为“你的文本文件和预定义格式相一致吗?”,下方为按预定义格式读入的数据文件的预览。显然,SPSS的预定义格式并没有正确识别该文件。因此选择“No”并单击“下一步”按钮。

第 2步 :最上方的问题是“变量是如何排列的?”,下面的选项分别为 Delimited(用某种字符区分)和 FixedWidth(固定宽度),一般都是 Delimited,该数据也是 ;第二个问题是“变量名包括在文件最前面了吗?”,选“Yes”,然后单击“下一步”按钮,如图 2.16(b)所示。

第 3步 :最上方的句子意为“第一条记录从第几行开始?”,右侧可以输入行数。由于所用数据的第一行为变量名,因此这里输入 2。下面的问题是“你的记录是怎样存储在文件中的?”。可以是“每一行代表一条记录”,或者“每＊＊个变量代表一条记录”,数据一般都是第一种情况。下一个问题是“你想导入多少条记录?”,可以是“所有记录”、“前＊＊条”或“随机导入＊＊%的记录”。一般也选前者,如图 2.17(a)所示。

第 4步 :左上方的问题为“变量间用的是哪种分隔符?”,可选的有 Tab键、空格、逗号、分号或自行定义的其他符号。本数据采用的是 Tab键,可见系统已经自动识别并选择了 Tab键,而下方的数据预览窗口显示出了正确的数据读入情况。右上方的问题意为“数据中采用的是什么文本限定符?”,提供了无、单引号、双引号和自定义 4种选择。如果数据中的字符串变量使用了限定符进行分隔,则需在此处指定,如图 2.17(b)所示。



(a) (b)

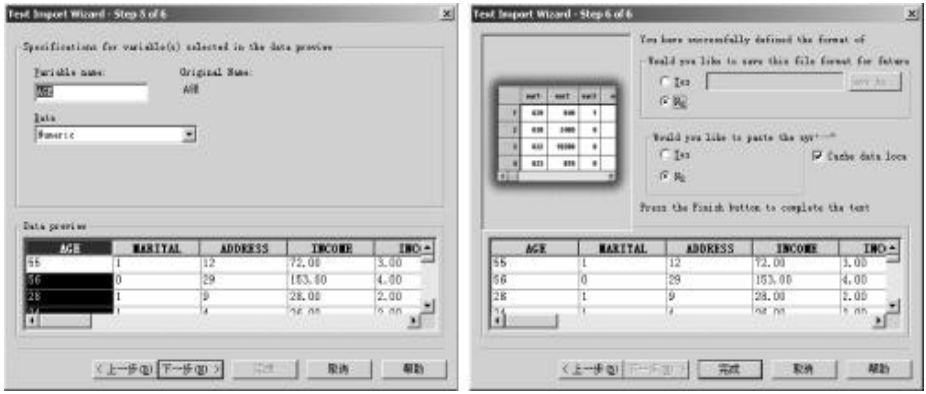
图 2.17 文本导入向导的第三、四个对话框

第 5步 :上方的提示为“定义在数据预览窗口中所选择的变量”。顾名思义,在这个对话框中用户可以在数据预览窗口中选择某一列变量,然后更改其变量名和类型。这里不需要做更改,可以直接单击“下一步”按钮,如图 2.18(a)所示。

第 6步 如图 2.18(b)所示最上面的问题为“你愿意保存这次的文件(读入)格式设置以备下次使用吗?”,第二个问题为“你是否愿意将以上操作粘贴为 SPSS语句吗?”,这里使用默认选项,单击“完成”按钮,可以看到 SPSS成功地读入了该文本数据。

2.3.3 数据库格式数据如何导入 SPSS中

SPSS可以直接读取很多类型的数据文件,对于不能直接打开的数据格式,SPSS提供了利用通用的数据库 ODBC接口读取数据的方法。这里以 SPSS系统自带的文件 demo.mdb为例,来看



(a) (b)

图 2.18 文本导入向导的第五、六个对话框

一下如何使用数据库查询方法读取这个文件。首先,选择菜单 File→ Open Database→ New Query,系统会弹出数据库向导的第一个窗口,其中会列出本机上已安装的所有数据源,如图 2.19 (a) 所示。可见里面列出了需要的 MS Access Database数据源,但该数据源不能直接使用,需要先进行定义:单击下方的 Add Data Source,系统会弹出 ODBC数据管理器窗口如图 2.19 (b) 所示。在用户数据源列表中选中 MS Access Database,单击配置按钮,会弹出该数据源的安装界面,如图 2.20 所示,单击其中的“数据库:选择”按钮,在弹出的文件打开对话框中找到 demo.mdb并单击“确定”按钮,数据源名可以任意指定,此处使用“MS Access Database”,此时安装界面上相应位置就会列出所指向的数据库名。

(a) (b)

图 2.19 向导初始对话框中的数据源列表和系统的 ODBC数据源管理器

单击两次“确定”按钮后回到最初的数据库向导界面,此时即可选中 MS Access Database数据源并单击下一步,系统就会进入向导的第二个窗口,采用拖放式操作将所需变量引入右侧框中,见图 2.21。向导的第 3步 ~第 5步适用于数据的选择性读入、字符值到数值与值标签的转换等操作。第 6步则提供了将生成的 SQL语句保存为文件以供再次使用,将前面的操作粘贴成

图 2.20 MS Access驱动程序安装界面

Syntax语句等功能。如果不需要这些设置,则可在第 2步时直接单击完成,数据就被成功读入了。

图 2.21 数据库向导的第二个对话框

由于 SPSS现在可以直接打开许多常用格式的数据文件,因此数据库查询接口的用处不是很大。但是使用 ODBC接口可以直接和绝大多数流行的数据库进行数据交换,如 SQL Server、DB2、Oracle等,这是直接打开方式无法做到的。其次,在例行工作中,比如每月都要读入相同的数据库,可以将所使用的 SQL语句存储起来,每次只要调用 SQL语句即可。这一方法也可用来解决一些需要对动态数据库进行统计分析的问题。数据仅仅是在需要分析时临时读入,从而可以保证数据始终是最新的。

2.4 数据的保存

数据录入过程中,要随时注意保存,以防出现意外情况,导致信息丢失。SPSS不仅能将数据保存为自己的数据格式(*.sav文件),而且还可以将数据保存为其他类型,如 DBF、FoxPro、Excel、Access等,下边将给出简单介绍。

2.4.1 存为 SPSS格式

无论是数据录入过程还是对数据做了修改,随时保存数据文件是必不可少的工作之一。选择 File→Save,如果数据文件曾经存储过,则系统会自动按原文件名保存数据;否则,就会弹出 Save Data As对话框 (见图 2.22)。此时为所要保存的文件指定文件名和保存的路径就可以了。

另外,有些时候分析者会在分析过程中生成一些临时变量,如果不希望保存全部变量,那么就可以使用 Save Data As对话框中的 Variables按钮来指定需要保存的变量。图 2.23就是在保存文件 input.sav时 Variables子对话框的内容,可见在每个变量的最左侧都有一个复选框,表明它们是否会被保存在文件中。对不需要的变量,单击相应复选框去除选择,则该变量就不会出现在新保存的数据文件中。

图 2.22 Save Data As主对话框

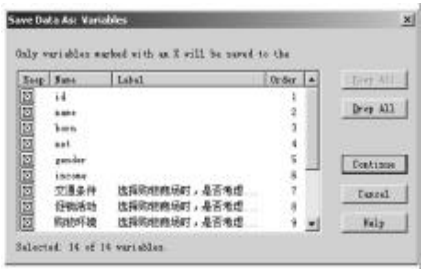


图 2.23 Variables子对话框

2.4.2 存为其他数据格式

SPSS的开放和友好之处不仅在于可以读取非 SPSS类型的数据,而且它还允许将数据存为很多种非 SPSS格式的数据。在 Save Data As对话框中可以看到,最下方有一个“保存”列表框,

单击后可以看到 SPSS能够保存的各种数据类型,有 dbf Excel SAS各版本的各种数据格式、纯文本格式等,用户只需要选择合适的类型,然后确定就可以了。不过,将数据存为 SPSS以外的其他类型,有些设置可能会丢失,如标签和缺失值等。虽然在保存为 SAS等数据格式时 SPSS会提示将标签等另行存储为一个 SAS程序文件,但这样毕竟不太方便,因此除非确实需要和其他软件交换数据,否则在决定保存为其他类型的数据的时候,一定要慎重行事。

思考与练习

针对 SPSS自带文件 demo.xls,进行以下练习:

1. 将该文件读入 SPSS中,仅包含以下变量:年龄、婚姻状况、家庭住址、收入。
2. 对变量 Marital(婚姻状况)设置值标签,1代表已婚,0代表未婚。

参考文献

1 张文彤主编.SPSS 11统计分析教程 (基础篇).北京:北京希望电子出版社,2002
2 SPSS Base 12.0 User's Guide. SPSS Inc. Chicago, Illinois, 2003

第3章 数据管理

不言而喻,一切统计分析都是以数据为基础的,在数据文件建立好之后,还需要对数据进行必要的加工处理。对同一个数据往往要从各种不同的侧面进行研究,采取多种统计方法进行分析,而不同的统计方法对数据文件结构的要求不尽相同,这就需要对数据文件的结构进行重新调整或转换,以便适合于相应的统计方法,这项工作称为数据管理。数据管理直接关系到数据分析的结果,因此是统计分析工作中不可缺少的一个关键步骤。

本章主要介绍 SPSS提供的数据管理方面的一些基本功能。在 SPSS中,数据文件的管理功能基本上都集中在 Data和 Transform菜单上,其中前者主要实现变量级别的数据管理,如计算新变量、变量取值重编码等,而后者的功能主要是实现文件级别的数据管理,如变量排序、文件合并、拆分等,下面将具体介绍这些功能。

3.1 变量级别的数据管理

对变量进行操作的内容主要集中于 Transform菜单 (参见图 3.1),包括新变量的生成、记录的排序、对变量进行计数等。在 12.0版中,SPSS这一菜单的项目可被分为以下几类:

计算新变量:实际上就是指最上面的 Compute过程,这是该菜单中最为常用和重要的过程。

变量转换:包括 Recode, Visual Bander, Count, Rank Cases, Automatic Recode这 5个过程,它们实际上都可以被看成是 Compute过程在某一方面功能的强化和打包,其中第二个过程为 12.0版新增。

专用过程:包括建立时间序列、缺失值替代和设定随机种子三个过程,其中前两个过程实际上专用于时间序列模型,对它们的讲解请参见本丛书中的《统计预测与时间序列模型》一书相关内容。设定随机种子的功能则主要影响伪随机函数的使用,详述见后面相关章节。

图 3.1 Transform菜单

Run Pending Transforms:用于执行编程中被挂起 (Pending)的数据整理操作。属于控制命令,本书对此不作讲解。

3.1.1 计算新变量

计算新变量的功能就是在原有 SPSS数据文件的基础之上,根据用户的要求,使用 SPSS算术表达式及函数,对所有记录或满足 SPSS条件表达式的记录,计算出一个新结果,并将结果存入一

个用户指定的变量中。这个指定的变量可以是一个新变量,也可以是一个已经存在的变量。

1. 常用基本概念

在新变量生成中,涉及了 SPSS算术表达式、SPSS函数、SPSS条件表达式等基本概念,因此首先简单讨论这些概念。

⑴ SPSS算术表达式:在变量转换的过程中,应根据实际需要,指出按照什么方法进行变量转换。这里的方法一般以 SPSS算术表达式的形式给出。SPSS算术表达式 (Numeric Expression)是由常量、SPSS变量名、SPSS的算术运算符、圆括号等组成的式子。

⑵ SPSS函数:SPSS提供了多达 70余种的系统函数。根据函数功能和处理对象的不同,可以将 SPSS函数分成八大类,它们分别是 算术函数、统计函数、分布函数、逻辑函数、字符串函数、日期时间函数、缺失值函数和其他函数。

函数具体的书写形式为 函数名 (参数)。这里,函数名是系统已经规定好的。圆括号中的参数有时是一个,也可以是多个;而参数的类型有时是常量 (字符型常量应用单引号引起来),也可以是变量名或 SPSS的算术表达式。此外,函数中如果有多个参数,各参数之间要用单字符逗号“,”隔开。

SPSS函数一般也会与 SPSS的算术表达式混合出现,用于完成更加复杂的计算。各种函数的释义可参考附录。

⑶ SPSS的条件表达式:通过 SPSS的算术表达式和函数可以对所有记录计算出一个结果,如果仅希望对部分记录进行计算,则应当利用 SPSS的条件表达式指定对哪些记录进行计算。根据实际需要构造出条件表达式之后,SPSS会从所有记录中自动挑选出满足该条件的记录,然后再对它们进行计算处理。

因此,如果用户在给出 SPSS算术表达式和函数的同时,又给出了一个条件表达式,那么,系统就会根据要求仅对满足一定条件的记录进行计算处理。

2. Compute过程的分析实例

了解了 SPSS算术表达式、SPSS函数和 SPSS的条件表达式之后,现在来看看如何通过 Transform命令实现新变量的生成。这里以数据 transform.sav为例,来介绍变量转换的操作步骤。

例 3.1 数据 transform.sav是某年级学生的数学、英语、语文三门课程的成绩,现在需要统计英语成绩在 60分以上的学生的语文和数学的平均成绩。

来看看怎么通过 Compute命令轻松地完成这一任务。选择菜单项 Transform→Compute,出现如图 3.2所示的窗口。该对话框看起来非常复杂,但实际上内容排列很整齐,左上角为需要计算的变量名,右上方的算术表达式 (Numeric Expression)框用于给目标变量赋值,对话框中部是类似计算器的软键盘,可以用鼠标按键输入数字和符号,软键盘右侧为函数窗口,可以在这里找到并使用所需的 SPSS函数。

现在开始具体的设定操作,在 Target Variable框中输入存放计算结果的变量名。该变量可以是一个新变量,也可以是已经存在的变量。新变量的变量类型默认为数值型,用户可以根据需要,点击 Type&Label按钮来修改变量的类型,或对新变量加变量名标签信息。

如果指定存放计算结果的变量为新变量,系统会自动在数据编辑窗口中创建该变量。如果

指定产生的变量名已经存在,则会以计算出的新值覆盖旧值。本例中命名新变量为 score,变量标签和变量类型采用默认,不做更改。

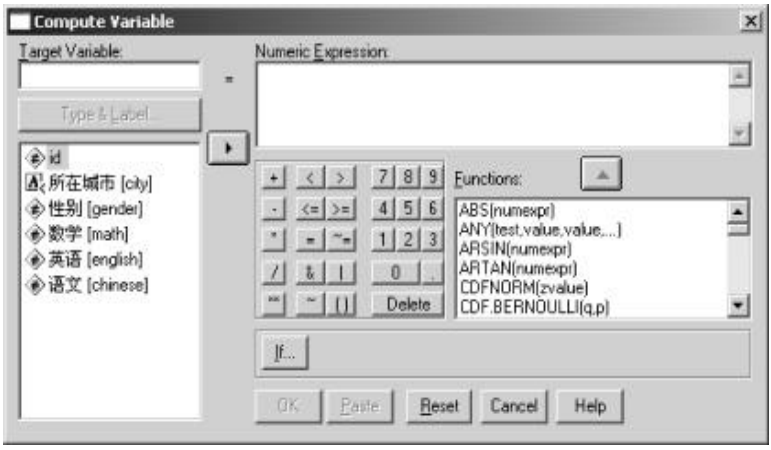


图 3.2 SPSS变量转换窗口

如果要对全部学生计算平均成绩,则直接在主对话框中操作即可,但现在仅希望对符合一定条件的记录进行变量转换,所以按 If钮,出现如图 3.3所示的窗口。点击 Include if case satisfies condition选项,然后通过手工输入或按动屏幕中的按钮和函数下拉菜单来实现条件表达式的输入工作。在本例中,单击“Include if case satisfies condition”选项以后,将左边的变量 english通过黑色的小箭头,使之进入右边的框中。然后利用软键盘输入“english>=60”,这意味着仅对英语成绩在 60分以上的学生进行统计分析。单击“Continue”按钮之后,回到 Transform的主窗口。

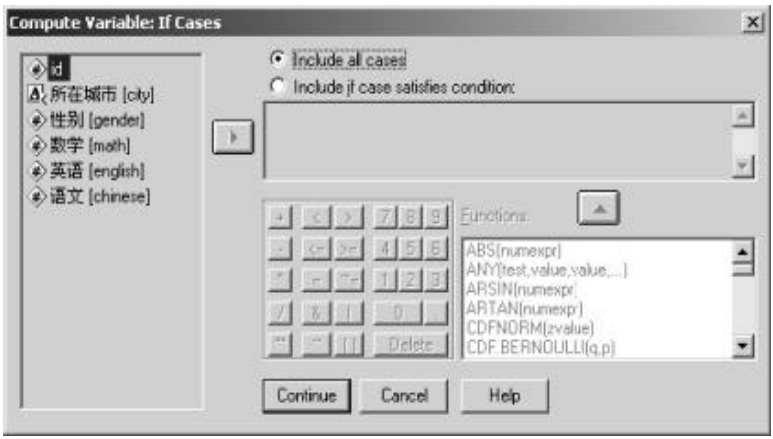


图 3.3 If按钮子对话框

最后,在 Numeric Expression框给出 SPSS算术表达式和函数。可以手工输入,也可以按动数字键盘中的按钮以及函数下拉菜单来完成表达式、函数的输入工作。

在本例中，Numeric Expression框给出了 MEAN (chinese, math)函数表达式，单击 “OK”按钮即可，如图 3.4所示。

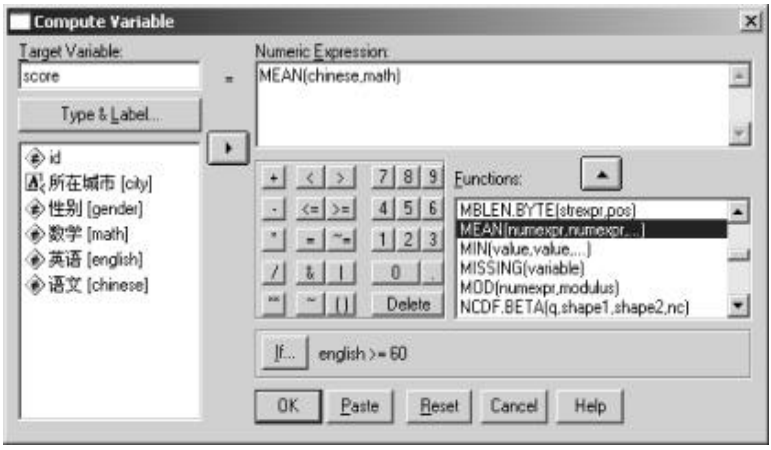


图 3.4 设置完毕的 Transform窗口

如果对全部人群生成一个新变量，但不同人群采用不同的算术表达式，可以通过多次调用 Compute Variable来实现。例如，要根据每个人的调整后工资计算其应缴的 “个人所得税 ”。根据个人所得税法规定 :对于调整后工资额小于 1 200元的人，他应交的个人所得税为 0;对于调整后工资额大于 1 200元且小于 1 700元的人，他应缴的个人所得税为 (调整后工资额 −1 200) × 0.05;对于调整后工资额大于 1 700元且小于 3 200元的人，他应缴的个人所得税为 (调整后工资额 −1 200) ×0.1−25;对于调整后工资额大于 3 200元且小于 6 200元的人，他应缴的个人所得税为 (调整后工资额 −1 200) ×0.15−125;等等。像这样一个问题，依然可以利用 Compute Variable来计算。具体做法是，第一次用 Compute Variable来计算满足条件 “调整后工资额小于 1 200元 ”的人的个人所得税为 “0”;第二次用 Compute Variable来计算满足条件 “调整后工资额大于 1 200元且小于 1 700元 ”的人的个人所得税为 “ (调整后工资额 −1 200) ×0.05”，点击 “OK”后会出现 “Change Existing Variable”，点击 “确定 ”;依次下去即可。这里的操作看似复杂，实际上直接写程序时，代码是很简单的，读者可以利用 Paste按钮粘贴出程序自行练习。

3.1.2 对变量值进行分组合并

数据分析中，将连续变量转换为等级变量，或者将分类变量不同的变量等级进行合并是常见的工作。而 Recode过程可以很好地完成这一类任务。 Recode into Same Variable是对原始变量的取值进行修改 ;而 Recode into Different Variable是根据原始变量的取值生成一个新变量来表示分组情况。但为了保存原始信息的完整性，一般选后者。

1. 对连续变量进行分组

在 SPSS中可以将连续变量转换为离散 (等级或定序)变量，按照某种一一对应的关系生成

新变量值,可以将新值赋给原变量,也可以生成一个新变量。Recode过程和 VisualBander过程都可以完成这一任务,但前者更为简单和常用。现在来看看下边这个例子,SPSS易学易用的特点将会再一次被证明。

例 3.2 在 transform.sav中生成新变量 grade,当英语成绩小于 60时取值为“不及格”,大于等于 60且小于 70为“及格”,大于等于 70且小于 80为“较好”,大于等于 80为“优秀”。

选择菜单 Transform→Record→Into Different Variables,将英语成绩 (English)选入 Input Variable→Output Variable框,此时 Output Variable框变黑,在 Name框键入新变量名 grade并单击“Change”按钮,可见原来的 english->? 变成了 english->grade,如图 3.5所示。

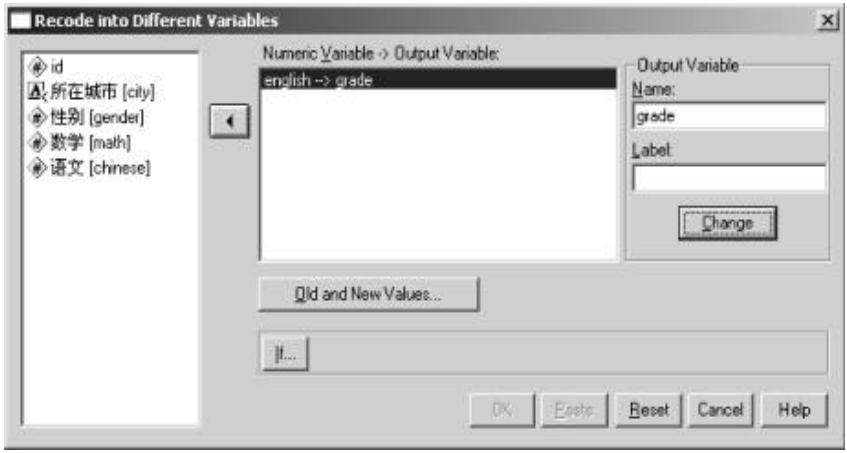


图 3.5 Recode对话框

现在单击“Old and New Values”,系统弹出变量值定义对话框如图 3.6所示。许多东西和前面类似,但要注意所有的范围都是包含了端点的,而前面设定的变换会优于后面的变换,所以为了能得到正确结果,应当将相应界值的变换设定放在最后面。另外,由于这里要生成的变量是字符型变量,需要选择相应的复选框,否则将无法录入变量值。

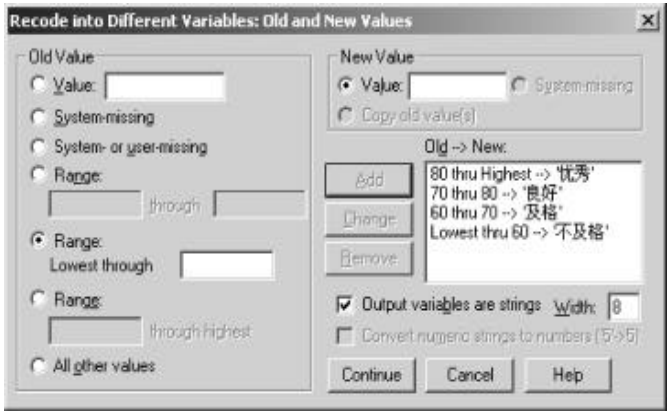


图 3.6 变量值定义对话框

这里的操作比较复杂,因此采用操作表格叙述如下:

Output variables are strings	
Old Value:	Range: 80 through highest New Value: Value:优秀 Add
Old Value:	Range: 70 through 80 New Value: Value:良好 Add
Old Value:	Range: 60 through 70 New Value: Value:及格 Add
Old Value:	Range: Lowest through 60 New Value: Value不及格 Add
Continue	

Recode可以将连续变量转化成数值型或者字符型离散变量,也可将数值型字符变量转化成数值变量,只需选中选项"Convert numeric strings to numbers"即可,轻轻一点,一切尽在掌握中。

2. 分类变量类别的合并

Recode过程也常用于合并某个分类变量的几个水平为一个水平,仍然举个例子来说明问题:将前文在数据 transform.sav中产生的变量 grade中的优秀、良好和及格三个等级合并为一个等级"PASS",将 grade的等级"不及格"转化为"NOPASS"。

界面在前文已经熟悉,现在来看看如何进行相应操作:

Transform→ Record→ Into Different Variables	
Numeric variable:grade	Output variable:Name:grade1: Change
Old and New values:	
Output variables are strings	
Old Value: value不及格	New Value: Value:NOPASS: Add
Old Value: All other values	New Value: Value:PASS: Add
Continue	
OK	

该程序运行之后,就可以看到变量 grade1将变量 grade中前三个水平合并为了一个水平"PASS"。

3.1.3 连续变量的可视化分段

Recode过程提供了精确分组的功能,但是如果希望进行的分组是较有规律的,比如等距分组,或者等样本量分组,使用 Recode过程进行操作就显得非常麻烦,且可视化程度不高,此时可以考虑使用 VisualBander过程进行可视化分段。VisualBander过程是 SPSS 12.0中新增的用于将连续变量进行分段的过程,该过程使用百分位数、标准差范围或者等间距方式将连续变量划分

为若干组段,并采用图形化操作的方式,非常直观好用。

这里仍以数据文件 transform.sav为例,假设现在希望按变量 math将学生分为 5组,60分以下为第一组,60分以上的按照等间距的方式分 4组,则选择菜单 Transform→Visual Bander,首先会弹出变量选择界面,要求选择希望进行分段 (组)的变量,这里选入 math,单击“Continue”后即弹出主界面如图 3.7所示。界面左上角列出的是需要进行分组的变量,选中后则会在右侧以直方图的形式给出变量的分布特征,同时在上方还会指出最大、最小值和缺失值情况。界面最上方的 Current行给出的是原变量的信息,而 Banded行给出的则是生成的分组变量的信息,可以自行定义和更改。

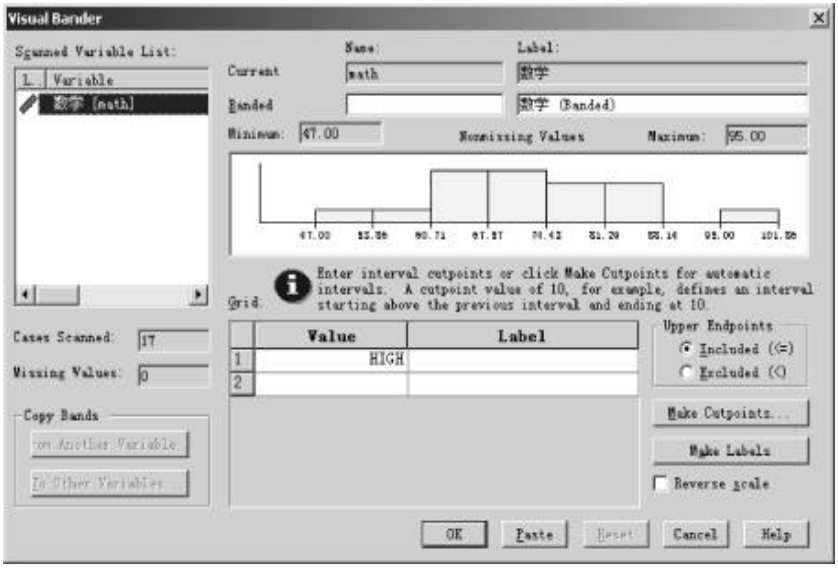
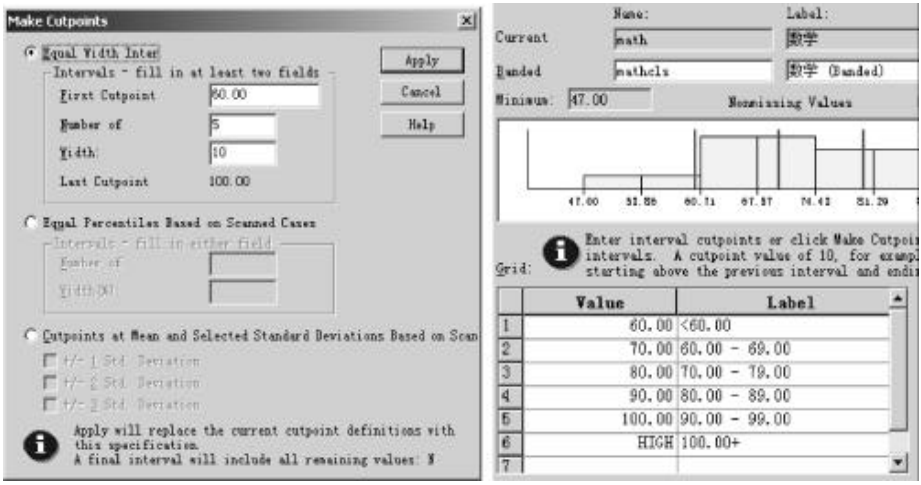


图 3.7 可视化分段对话框

整个界面的中下部均用于定义分组规则,Grid框组用于显示定义好的规则,更改规则可以在该界面上直接进行,但更方便的方式是使用 Make Cutpoints子对话框设定分段规则,用 Make Labels按钮自动填充值标签。以前者为例,它可以选择使用等间距 (Equal Width Intervals)、等比例 (等样本量,Equal Percentiles Based on Scanned Cases)或者按照指定的标准差范围 (Cutpoints at Mean and Selected Standard Deviations Based on Scanned Cases)三种方式进行分段,如图 3.8(a)所示,其中第三种方式显然可以用来在数据分析或质量控制中筛选异常值。本例中为第一种方式,即在对话框中依次定义好分组的起点、组段数或组距,相应的分组定义即可完成。

在单击“Apply”按钮回到主界面后,就会发现变量 math的直方图自动显示出了所定义的分组界限,如图 3.8(b)所示,此时可以通过拖拉分隔线的方式来修改分组界限值。显然,可视化分段过程在操作上要比 Recode过程赋予用户对数据更多的控制能力。本例完整的操作步骤如下:

Transform→ Visual Bander
Variables to Bander:math
Continue
选中 math:
Banded: Name:mathcls
Upper Endpoints: Excluded (<)
Make Cutpoints: First Cutpoint:60 Number of:5 Width:10: Apply
Make Labels
OK



(a) (b)

图 3.8 Make Cutpoints子对话框以及设置完毕的可视化分段对话框

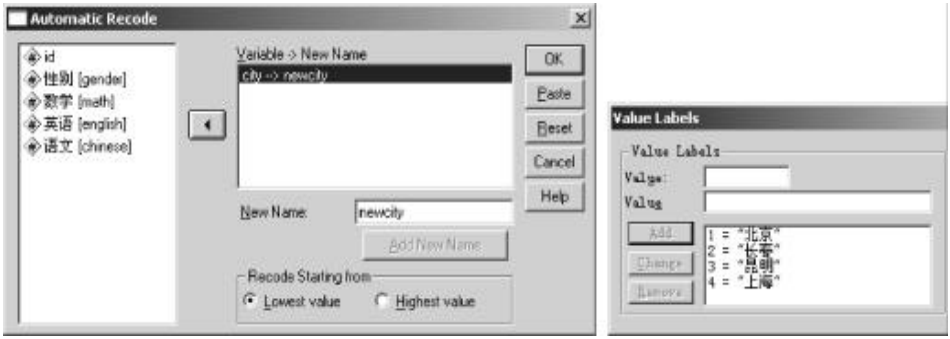
3.1.4 将字符变量转换为数值变量

在数据分析中,将字符变量转换为数值变量是非常实用的一个功能。除了使用 Recode过程手工设定转换规则外,在 SPSS中还可以使用 Automatic Recode过程自动按原变量值的大小或者字母排序生成新变量,而变量值就是原值的大小次序。

例 3.3 在 transform.sav数据中,将字符型变量 city转化为数值变量 newcity。

由于 Automatic Recode过程的操作界面非常简单,这里就不再详述操作过程,直接给出相应的界面和结果如图 3.9所示。

Automatic Recode的排序功能和 Rank Cases类似,所不同的是,Automatic Recode可以用于字符型变量。



(a) (b)

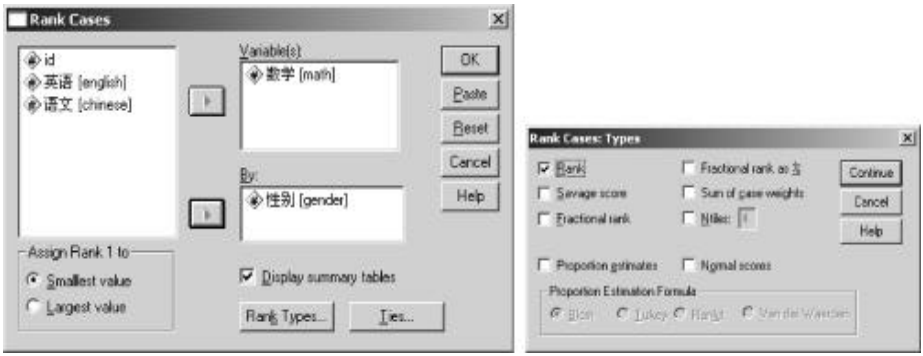
图 3.9 Automatic Recode对话框以及所生成变量 newcity的值标签定义

3.1.5 变量的编秩

所谓编秩,其实就是对记录按照某个变量值的大小来排序。Rank Cases过程就是用来排序的一个专用过程。具体来说,它根据某变量的大小来排出次序(秩次),然后将秩次结果存储到一个新变量中去。这样做有什么用处呢?在许多时候参数检验的条件不被满足,此时不得不使用非参数方法,而稍微复杂些的非参数方法就无法直接用对话框来完成了,需要先计算秩次再进行分析(详见非参数检验一章)。

例 3.4 试根据性别分组计算数学成绩的秩次。

解:选择菜单 Transform→ Rank Cases,弹出 Rank Cases对话框如图 3.10所示。



(a) (b)

图 3.10 Rank Cases对话框

在 Rank Cases对话框中:

(1) Assign Rank1 to单选框组:用于选择将秩次 1赋给最小值或最大值。

(2) ☑ Display summary tables:用于确定是否在结果窗口内输出结果报表。

⑶ Rank Types按钮:用于定义秩次类型,默认为最常用的 Rank(秩分数),另有其他几种选择,单击“More”按钮,还会有更多的设置。由于除了秩分数以外的方法很少被用到,这里不再详述,有兴趣的朋友可参见用户手册。

⑷ Ties按钮:用于定义对相同值观测量的处理方式,可以是取平均秩次、最小秩次、最大秩次或当作一个记录处理,默认值为取平均秩次。

这里将变量 math选入 Variable框,分组变量 gender选入 By框,单击“OK”按钮即可,其他一些设置使用默认。系统会建立一个新变量 Rmath(即原变量名前加 R表示 Rank之意),其取值为 math分组的秩次。

在前面讲解的操作全部结束后,数据集 transform.sav中的数据如图 3.11所示。

	id	city	gender	math	english	chinese	grade	newgrade	mathcls	newcity	Rmath
1	1	昆明	1	67.00	69.00	66.00	及格	PASS	2	3	3.000
2	2	北京	2	78.00	59.00	47.00	不及格	NOPASS	3	1	7.000
3	3	上海	1	56.00	68.00	55.00	及格	PASS	1	4	2.000
4	4	北京	1	78.00	69.00	98.00	及格	PASS	3	1	6.000
5	5	上海	2	69.00	77.00	69.00	良好	PASS	2	4	5.000
6	6	长春	1	87.00	47.00	87.00	不及格	NOPASS	4	2	7.500
7	7	北京	2	69.00	66.00	69.00	及格	PASS	2	1	5.000
8	8	上海	2	88.00	91.00	78.00	优秀	PASS	4	4	8.000
9	9	长春	2	69.00	66.00	69.00	及格	PASS	2	2	5.000
10	10	长春	1	87.00	47.00	87.00	不及格	NOPASS	4	2	7.500
11	11	昆明	1	75.00	66.00	76.00	及格	PASS	3	3	5.000
12	12	长春	1	69.00	60.00	97.00	及格	PASS	2	2	4.000
13	13	北京	1	95.00	79.00	78.00	良好	PASS	5	1	9.000
14	14	上海	2	66.00	69.00	69.00	及格	PASS	2	4	2.000
15	15	上海	1	47.00	97.00	87.00	优秀	PASS	1	4	1.000
16	16	长春	2	66.00	87.00	69.00	优秀	PASS	2	2	2.000
17	17	长春	2	66.00	68.00	55.00	及格	PASS	2	2	2.000

图 3.11 变换后的 transform.sav中的数据

3.1.6 Transform菜单中的其他功能

⑴ Count过程:该过程用于表示某个变量的取值中是否出现某个值,可以是单个数值,也可以指定区间,并且可以仅给出条件,而不必对整个数据集进行操作。该过程的功能可以直接使用 Recode过程来实现。

⑵ Random Number Seed过程:用于设定伪随机函数的随机种子。默认情况下随机种子随着时间在不停改变,这样所计算出的随机数值无法重复,这在临床试验等情况中是不符合要求的。此时可用 Random Number Seed过程人为指定一个种子,以后所有的伪随机函数在计算时都会以该种子开始计算,即结果可重现。但它对真随机函数没有任何影响。

3.2 文件级别的数据管理（一）

Transform菜单提供的数据管理功能虽然很强，但基本上仅限于变量级别，有时候需要对整个数据文件进行加工整理，而不仅仅是对变量进行操作。在 SPSS中，这部分功能主要集中在 Data菜单（参见图 3.12）下。根据各自的功能特点，该菜单中的所有项目可分为以下几类：

图 3.12 Data菜单

① 简单命令：包括插入变量、插入记录和到达某条记录，它们的功能实际上都可以使用鼠标在数据表界面上直接完成，很少会使用菜单来调用，本书不对其进行讲解。

② 常用的简单过程：包括排序、拆分文件、选择记录和加权记录，这几个过程并不复杂，但使用得极为频繁，是大家必须掌握的内容。

③ 变量与数据文件属性向导：是 11.5版新增的两个向导，用于定义数据字典，或者将预定义的数据字典直接引入当前数据文件，对于大型或者连续性的数据分析项目而言，这是一个非常有用的功能。

④ 数据重构向导：用于进行数据转置，或者对重复测量数据进行长型、宽型记录格式间的转换，详述见后面相关章节。

⑤ 文件合并过程：将几个数据文件合并为一个大的 SPSS数据文件，含横向合并和纵向合并两种情况，详述见后面相关章节。

⑥ 正交设计过程：实际上是联合分析模块的一部分，用于生成实施联合分析所需的设计，由于这一分析方法是市场研究中的专用工具，对它的讲解可参见本丛书中的《SPSS与市场研究》一书相关内容。

⑦ 其他过程：包括定义日期变量过程、数据汇总过程和查找重复记录向导。前者用于时间序列数据的分析，将在时间序列一书中讲解，后两个过程将在下文加以讲解，其中查找重复记录向导为 12.0版新增功能。

本节将首先讲解非常重要的几个简单过程，下一节将重点讲述文件级别数据管理中一些比较复杂的功能。

3.2.1 记录排序

数据编辑窗口中记录的前后次序是随机的，由录入时的先后顺序决定。实际工作中，有时用户希望按某种顺序来观察一批数据，例如，在销售报表中，希望按销售额从低到高的顺序，或者按销售时间从早到晚的顺序来浏览数据。观察排序后的记录数据，会方便用户了解数据。

SPSS中的记录排序就是将数据编辑窗口中的数据，按照用户指定的某一个或多个变量值的升序或降序重新排列，这里用户所指定的变量称为排序变量。当对所有记录进行排序时，可按照排序变量取值的大小次序对记录数据重新整理后显示。当对记录进行分组排序时，在每个组内，按照排序变量取值的大小次序对记录数据进行排序。

对于单变量排序，SPSS提供了一种简易操作方法，就是在数据表格的变量名处单击右键，弹出的右键菜单其最后两项就是“Sort Ascending”和“Sort Descending”。但是，对于多变量排序，则需要使用这里讲述的 Sort Cases过程来进行。由于该对话框并不复杂，因此这里不再详细讲解，仅给出一个示意图，如图 3.13所示。

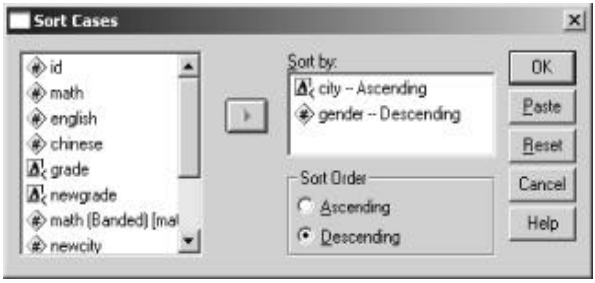


图 3.13 Sort Cases对话框

图 3.13显示的是将数据按照 city升序，gender降序的方式进行排列的操作。其中，比较特殊的是 city和 gender后面分别跟着 Ascending和 Descending，表明前者按升序，后者按降序排列，如果要改变升降序，则选中相应变量，然后直接在 Sort Order单选框组中修改选择即可，同时需要说明以下几点：

⑴ 在多重排序中，指定排序变量名的次序是很关键的，先指定的变量在排序时必然优先于后指定的变量。即记录首先按第一个变量进行排序，对于与第一变量取值相同的记录考虑按第二个变量排序，以此类推。

⑵ 可以指定按某变量值升序排序的同时按另一变量值降序排序，或相反。

⑶ 排序以后，原来记录数据的排列次序将被打乱。因此，在时间序列的数据中，如果数据中没有存放记录标志的变量，如年份等，则应注意保存原数据的排列顺序，以免数据混乱。

3.2.2 记录拆分

用于将数据文件分组进行处理。如果希望分组进行相应的统计分析，或者只分析其中的一部分数据，则可以通过拆分数据集来加以实现。Split File过程用于实现这一功能，其界面非常简单，如图 3.14所示。

这里介绍一下各个对话框元素的用途：

⑴ Analyze all cases：和下面的两个单选框为一组，选中本框不拆分文件。

⑵ Compare groups：按所选变量拆分文件，各组分析结果放在一起便于比较。

⑶ Organize output by groups：按所选变量拆分文件，各组分析结果单独放置。

⑷ Groups Based on框：用于选择拆分数据文件的变量。

⑸ Sort the file by grouping variables：和下面的 File is already sorted为一组，要求拆分时将数据按所用的拆分变量排序。

⑹ File is already sorted：如果数据集很大，而所用的拆分变量已经排过序了，可使用该单

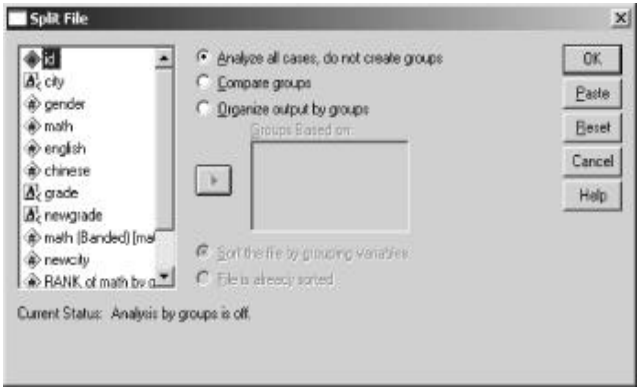


图 3.14 SplitFile过程主对话框

选框以节省运行时间,但实际上较少用到。

当对数据集进行拆分后,可以看到状态栏右侧会出现 SplitOn的提示,表明所做的拆分正在生效,它将在以后的分析中一直有效,而且会被存储在数据集中,直到再次进行设定为止。

3.2.3 记录筛选

很多时候用户不需要分析全部的数据,而是按要求分析其中的一部分,比如只分析职位是经理的人的年薪,或者只对接受教育年限在 12年以上的人进行分析,这时使用 SelectCases过程可以大大简化用户的工作。对话框界面如图 3.15所示。

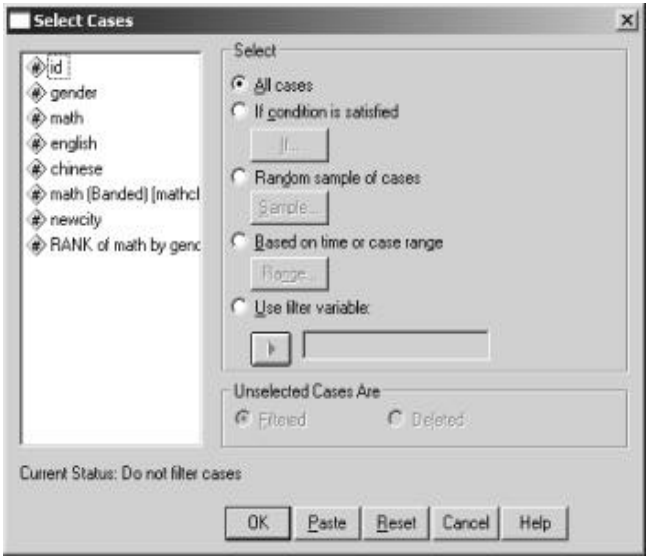


图 3.15 SelectCases过程主对话框

⑴ 主要的对话框元素为 Select单选框组，用于确定选择方式。

All cases:分析所有的记录。

If condition is satisfied:只分析满足条件的记录，单击下方的 If按钮后弹出 If对话框，用于定义筛选条件。

Random sample of cases:从原数据中按某种条件抽样，使用下方的 Sample按钮进行具体设定，可以按百分比抽取记录，或者精确设定从前若干个记录中抽取多少个记录。

Based on time or case range:基于时间或记录序号来选择记录，使用下方的 Range按钮设定记录序号范围。

Use filter variable:使用筛选指示变量来选择记录，必须在下面选入一个筛选指示变量，该变量取值为非 0的记录将被选中，进入以后的分析。

⑵ 最下方的 Unselected Cases Are单选框组用于选择对没有选中的记录的处理方式。

Filtered:表示未被选中的记录只是被隔离，这些记录的记录号上会被加上斜杠以示区别，同时系统会自动产生一个名为 filter_$的筛选指示变量，被选中的记录该变量取值为 1，反之则为 0。

Deleted:未被选中的记录将被删除，一般不要使用，以免误删数据。

当对数据集做出筛选后，可以看到状态栏右侧会出现“Filter On”的提示，表明所做的筛选正在生效，筛选功能将在以后的分析中一直有效，而且会被存储在数据集中，直到再次改变选择条件为止。

3.2.4 记录加权

在默认情况下，每一行就是一条记录，这在多数情况下没有什么问题，但有时却非常麻烦。如图 3.16所示的数据表，如果每一行就是一条记录，则需要输入 121行。这时候，一般使用频数格式录入数据，即相同取值的观测只录入一次，另加一个频数变量用于记录该数值共出现了多少次。这样就需要在分析时用到 WeightCases过程（参见图 3.17）将数据指定为该种格式。该过程的使用极为简单，对话框界面上有两个单选按钮，分别是不按权重记录和按某变量权重记录，如果选择后者，则需要选中一个权重变量。

	gender	group	count
1	1.00	1.00	34.00
2	2.00	1.00	23.00
3	1.00	2.00	45.00
4	2.00	2.00	19.00

图 3.16 频数格式录入数据

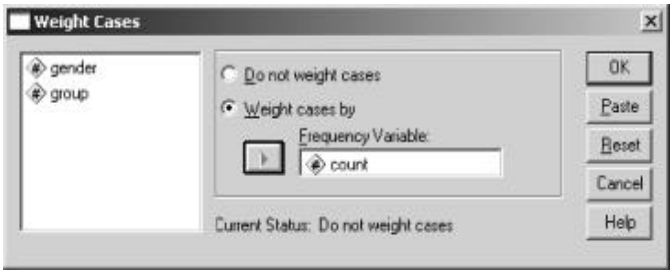


图 3.17 WeightCases过程主对话框

进行权重记录以后，SPSS界面右下角会出现“Weight On”的字样，且可以被存储到数据集中，直到用户取消加权，否则一直按加权对数据进行处理。

3.2.5 数据汇总

所谓分类汇总就是按指定的分类变量对观测值进行分组，对每组记录的各变量求指定的描述统计量，结果可以存入新数据文件，也可以替换当前数据文件。对数据文件进行分类汇总是实际工作中经常遇到的事情。例如，对于学生基本情况的数据，现希望了解不同性别学生的平均分数情况。这就需要首先对数据按不同性别分类，然后再分别求出各类学生的分数平均值。这个过程本质就是一个数据的分类汇总的过程。

在 SPSS中，实现数据文件的分类汇总是经过三大步骤完成的。首先，要指定分类变量（Break Variable(s)）和汇总变量（Aggregate Variable(s)）；然后，SPSS自动根据分类变量的取值将记录数据分成若干类，并对每类记录分别计算汇总变量的描述统计量；最后，将分类汇总的计算结果保存到一个 SPSS数据文件中。

为更清楚地了解 SPSS分类汇总的过程和结果，这里以数据 transform.sav为例来加以演示。

例 3.5 根据数据 transform.sav中学生的性别变量对英语的平均成绩进行汇总。

首先，选择菜单：Data→ Aggregate，出现如图 3.18所示的窗口。然后，指定分类变量到 Break Variable(s)框中，指定汇总变量到 Aggregate Variable(s)框中。使用 Function按钮指定对汇总变量计算哪些描述统计量，此处共提供了 5组函数，分别为常用汇总函数、特定值、记录数、百分比和百分片断（Fraction）。以最常用的第一组为例，可选的函数有均数、中位数、总和、标准差 4种。SPSS默认对各类分别计算汇总变量的均值，见图 3.18。

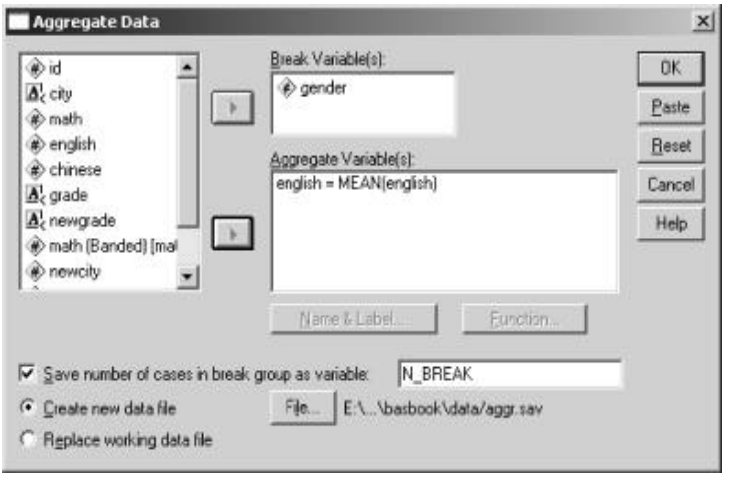


图 3.18 Aggregate过程的主对话框

另外，也可以指定分类汇总的结果保存到何处。有两种选择：第一，Create new data file表示

将结果生成到系统默认的名为 aggr. sav的 SPSS数据文件中，可以按 File按钮重新指定结果文件路径和文件名。第二，Replace working data file:表示以分类汇总后的结果覆盖 SPSS当前数据编辑窗口中的数据。一般采用前一种方式较好。

如果希望在结果数据文件中保存分类组的记录数，则选择 Save number of case in break group as variable选项。于是 SPSS便在结果数据文件中自动生成一个默认名为 N_ BREAK的变量，见图 3.19。

	gender	english	N_BREAK
1	1	66.89	9
2	2	72.88	8

图 3.19 保存分类组的记录数

分类汇总产生的 SPSS数据文件的记录数取决于分类变量的取值个数。这里，分类变量性别有两种取值，则按性别分类汇总后的数据就有两条。

还需要说明的是 分类汇总中的分类变量可以指定多个，称为多重分类汇总。此时汇总数据文件的记录数等于各分类变量类别数的乘积。如分类变量为性别（男、女）和班级（一、二、三），则汇总数据文件中会有 6（2×3）条记录。第一个指定的分类变量为主分类变量，其他的依次为第二、第三分类变量。

3.3 文件级别的数据管理（二）

在上一节中讲解了最为基本和常用的数据管理功能，对于一般的数据分析任务，这些已经足够。但是在较复杂的数据分析项目中，往往会在数据管理中涉及格式化数据、发现重复录入记录，拼接多个数据集和转换存储格式等复杂功能，涉及的数据文件也不止一个，本节的任务就是为大家讲解这些较为复杂的文件级别数据管理功能。

3.3.1 数据字典的定义与应用

在大型的数据分析项目中，数据管理是非常重要的一个环节，为了保证工作质量，数据处理人员往往会事先定义好一个非常详细的数据格式，包括变量格式、变量标签、值标签、缺失值定义等，这被称为数据字典。从 11.5版起，SPSS新增了两个数据管理向导，专门用于定义数据字典，或者将预定义的数据字典直接引入当前数据文件。对于大型或者连续性的数据分析项目而言，这是一个非常有用的功能，可以大大减轻数据处理人员的工作负担。

1. 变量属性定义向导

变量属性定义向导即 Define Variable Properties过程，用于对数据集中已存在的变量进一步定义其属性。具体说来，可以列出所选变量的所有取值；分辨没有值标签的值，并且提供自动给出值标签的功能；可以将另一个变量的属性拷贝到所选的变量，也可以将所选变量的属性拷贝到其他变量。虽然该向导的绝大多数功能都可以在变量视图中实现，但对于复杂的数据管理项目而言，它的可视化能力可以大大提高工作效率，并且对初学者而言，使用该向导进行变量的设置也是非常好的选择。

这里仍以数据集 transform. sav为例对该向导加以说明。假设现在希望对变量 gender进行属

性设定,则选择 Data→Define Variable Properties,此时会弹出预定义对话框,要求选择希望进行设定的变量,可以选择多个,SPSS将会对选入的变量都进行扫描。这里只选入 gender,则进入向导的主界面如图 3.20所示。

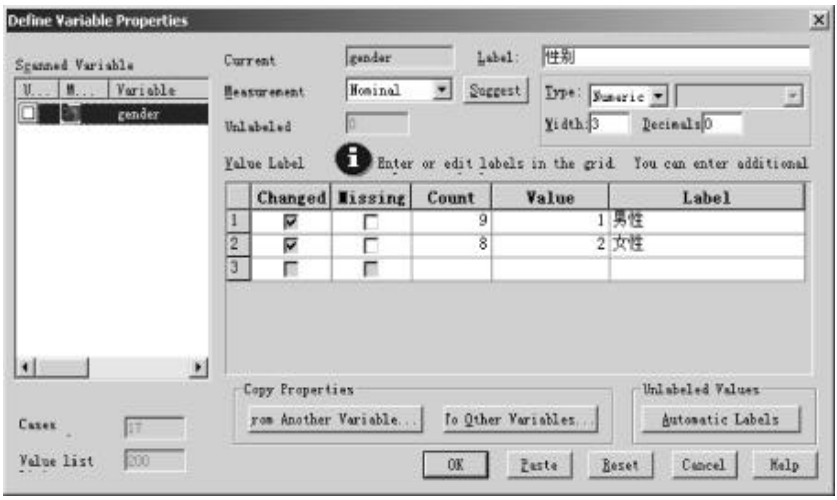


图 3.20 Define Variable Properties对话框

主界面的左侧会列出所有被选择或扫描的变量,选中相应的变量名称,则右侧会显示出相应的设定,并供用户加以更改:上部用于设定测量尺度、存储格式、变量名标签等,如果单击 Suggest按钮,则系统会根据扫描到的数据给出建议的测量尺度;中部的 Value Label网格会列出该变量所有取值的频数、当前值标签和缺失值设定等,这里可以更改标签和缺失值的设定。下部的 Copy Properties按钮组用于将另一个被扫描变量的属性拷贝到所选的变量,也可以将所选变量的属性拷贝到其他被扫描变量,这里由于只选择了一个变量,因此实际上没有用到该按钮组。右下方的 Automatic Labels按钮用于自动生成值标签,实际上就是将所有的变量值均赋给值标签。

如图 3.20所示,此时已经对 gender的属性进行了更改,读者可以看到在这一个界面中就完成了对变量的所有属性定义,而且可以一次性定义多个变量,并且由系统帮助扫描出全部取值范围,这显然要比在变量视图中进行操作要容易得多,可以大大方便数据字典的定义工作。

2. 复制数据文件属性向导

Copy Data Properties过程用于将定义好的数据字典直接应用到当前文件中,在操作时不仅可以将一个外部的数据文件相关属性拷贝到当前数据文件中,还可以进行自定义,只选择某些变量,或者某些属性进行拷贝,这无疑大大提高了连续性项目对原有资源的利用程度。对于一些特殊的文件属性,如多选题变量集、普通变量集、权重变量的设定等,使用该向导进行复制会减少许多重复工作。

例 3.6 将数据集 transform.sav中相关的变量属性作为数据字典应用到另一个数据集 transform2.sav中。

各位读者可以首先分别打开这两个文件,比较一下它们之间的区别,可以发现对于相同的变

量，transform. sav中均设置了标签，且列宽、测量尺度等的设置均不相同。下面开始进行操作，首先打开文件 transform2. sav，然后选择 Data→Copy Data Properties，系统会首先弹出向导的第一个对话框，要求指定希望复制的属性是来自于当前文件，还是另一个外部数据文件，本例中指定为 transform. sav所在位置。单击“下一步”按钮后出现如图 3.21所示的对话框，该界面用于设定希望复制的属性种类，有三种选择，分别为选择同名同类型同长度变量的属性进行复制（Apply properties from selected source file variables to matching working file）、选择一个变量的属性进行复制（Apply properties from a single source variable to selected working file variable）和仅复制文件属性（Apply dataset properties only-no variable selection）如多选题集定义、权重设定等。这里选择第一项，需注意性别变量由于在两个文件中的名称不同，因此未出现在下方的列表中。然后选中源文件变量列表中的全部变量，单击“下一步”按钮，随后的对话框（见图 3.22）会要求用户详细指定希望复制的变量属性，共有 7种之多，并且可以选择是替换原有属性，还是和原属性进行合并。

图 3.21 Copy Data Properties对话框 1

在如图 3.21,图 3.22所示的两个对话框出现时，使用者其实就可以单击“完成”按钮结束向导了，此后出现的界面分别用于选择希望复制的文件属性，以及是否生成相应的 SPSS程序。运行完毕后，大家就会看到，除了未加设定的变量 sex外，其余各变量的属性都套用了 transform. sav中的相应设置。

最后，总结一下如何应用上述两个向导来完成数据管理任务。如果有事先定义的数据字典格式，则可以先生成一个没有记录的空数据文件，将全部的数据字典设定好，将来在数据录入完毕后使用复制文件属性向导套用字典即可 如果没有事先定义的数据字典格式，则可以在录入工

图 3.22 Copy Data Properties对话框 2

作进行了一段时间以后先使用变量属性定义向导完成数据字典的设定工作,然后随着录入工作的进行经常扫描数据的情况,及时更新字典,最后在录入工作完毕后,使用复制文件属性向导应用字典的最终版本。现在大家知道了这两个向导并不是多余的,而是非常重要和实用的。当然,如果数据管理任务不太复杂,则也可以直接在数据字典中录入数据,或者直接在变量视图中修改属性。但是在真正的大型数据管理项目中,单独建立和维护数据字典是非常关键的一环,此时这两个向导的作用就不可忽视了。

3.3.2 查找重复记录

在大型的数据管理或者复杂的数据变换工作中,重复记录的发现是经常需要完成的任务。Identifying Duplicate Cases是 SPSS 12.0新增的功能,通过简单的菜单操作,可以迅速地发现个别变量值重复,或者所有数值完全重复的记录。

下面用数据 company_ rongyu. sav来进行示例。该数据是一份对几个公司的统计表。但由于有的公司提交了数次,因而在这个数据文件中出现了不止一次。在作统计工作时必须把这些重复数据删掉。数据量少时排序后逐个删除当然是没有问题的,但数据量较大时,这将是一个非常庞大的工作。SPSS提供了这种识别重复记录(Identify Duplicate Cases)的程序,下面看一下如何用它来简化工作。选择 Data→ Identify Duplicate Cases,弹出如图 3.23所示的对话框,上方的Define matching cases by框用于选入希望查找重复值的变量(组),这里将企业名称、企业所在地

区两个变量选入；下方的 Sort框组用于设定对于重复的记录按照哪个变量的取值排序，操作方式和 Sort对话框基本相同，此处选入 id整个对话框的下方实际上不会影响重复记录的查找，只是影响相应记录的显示和排列方式，Indicator框组用于设定是将第一个，还是最后一个重复记录设为主记录（相应的，其余记录就成为了“重复”的记录），而 Sequential框用于选择是否要求为重复记录编制流水号。

图 3.23 Identify Duplicate Cases对话框

在操作完毕后，得到的结果如图 3.24所示，可见变量 PrimaryLast等于 0表示相应记录为重复记录，本例中共发现 2、4、7三条重复记录。而重复的记录间又是按照 ID号的大小进行排序，这正是原本所设定的情形。

	id	企业名称	企业所在地区	营业收入	营业成本	Primary Last
1	2	bbbb	上海	1111.64	872.04	0
2	6	bbbb	上海	1111.64	872.04	1
3	4	dddd	北京	10081.3	1785.20	0
4	12	dddd	北京	10081.3	1785.20	1
5	7	ffff	北京	951.65	735.03	0
6	10	ffff	北京	951.85	735.03	1
7	1	aaaa	北京	2730.66	1427.25	1
8	3	cccc	广州	826.64	447.46	1
9	5	eeee	上海	5567.85	1524.12	1
10	8	gggg	广州	4901.66	583.92	1
11	9	hhhh	北京	559.76	14.21	1
12	11	jjjj	上海	1021.16	1516.72	1

图 3.24 操作结束后的数据界面

最后，结果窗口中还会给出本次操作的汇总信息，如表 3.1和表 3.2所示。

表 3.1 Statistics

Indicator of each last matching case as Primary

N	Valid	12
	Missing	0

表 3.2 Indicator of each last matching case as Primary

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	0 Duplicate Case	3	25.0	25.0	25.0
	1 Primary Case	9	75.0	75.0	100.0
	Total	12	100.0	100.0	

3.3.3 数据文件的重新排列与转置

数据文件的重新排列,是数据分析中经常用到的一个功能。数据录入的格式,未必能一步到位地满足用户分析时的要求,很多时候用户要根据分析的要求改变数据的排列格式,Restructure过程是一个图形化界面的数据重构向导,直观地实现了这一功能。

1. 数据的长型与宽型格式

长型格式和宽型格式指的是重复测量数据的两种不同的排列方式,由于重复测量模型可以使用不同的统计模型加以分析,因此,根据模型的要求进行长型格式和宽型格式之间的互转是数据分析中经常要遇到的问题。

这里以 SPSS的自带文件 Anxiety.sav和 Anxiety 2.sav来说明这两种数据排列格式的特点。这两个文件记录的都是 12名精神病患者在接受治疗后的 4个时间点的精神状态评分,其中变量 subject为病人的 id号,score为评分,trial为测量时的时间点编号,anxiety和 tension记录了病人在治疗前有无焦虑和紧张。Anxiety.sav文件是长型格式,以每次测量作为一条记录,用变量 subject和 trial来区分是哪位病人的第几次测量,anxiety和 tension作为携带变量在相同病人的记录中重复出现,这样 12个病人共形成了 48条记录;而 Anxiety 2.sav是宽型格式,每位病人作为一条记录,4次测量分别用 trial1～trial4这 4个变量来分别记录,原先用于区分测量次数的变量 trial不再需要,同一个病人的 subject anxiety和 tension也只出现一次。从图 3.25中可以更清楚地理解这两种数据格式的特点。

事实上,在学习了第 2章后,大家应当能够明白长型格式才是符合统计分析要求的标准记录格式,但是由于重复测量数据会使用特殊的重复测量模型来进行分析,此时就需要将数据变换为宽型格式,该模型的详情参见本丛书的高级教程相关章节。

2. 长型格式转换为宽型格式

现在来看看如何使用 Restructure过程实现数据结构的重建。

例 3.7 将 SPSS自带文件 Anxiety.sav转换为 Anxiety 2.sav的格式。

	subject	anxiety	tension	score	trial
1	1	1	1	18	1
2	1	1	1	14	2
3	1	1	1	12	3
4	1	1	1	6	4
5	2	1	1	19	1
6	2	1	1	12	2
7	2	1	1	8	3
8	2	1	1	4	4
9	3	1	1	14	1
10	3	1	1	10	2
11	3	1	1	6	3
12	3	1	1	2	4

	subject	anxiety	tension	trial1	trial2	trial3	trial4
1	1	1	1	18	14	12	6
2	2	1	1	19	12	8	4
3	3	1	1	14	10	6	2
4	4	1	2	16	12	10	4
5	5	1	2	12	8	6	2
6	6	1	2	18	10	5	1
7	7	2	1	16	10	8	4
8	8	2	1	18	8	4	1
9	9	2	1	16	12	6	2
10	10	2	2	19	16	10	8
11	11	2	2	16	14	10	9
12	12	2	2	16	12	8	8

图 3.25 数据集 Anxiety.sav和 Anxiety2.sav的内容

解：选择 Data→Restructure，系统会弹出 Restructure向导的第一个界面如图 3.26所示，从图中可以看出，在向导中共提供了三种数据重排功能，分别是长型与宽型格式的互换和行列转置。根据要求，在这个例子中要使用的是第二种功能，选择 Restructure selected cases into variables单选框，单击“下一步”按钮后显示向导的第二个界面，见图 3.27。

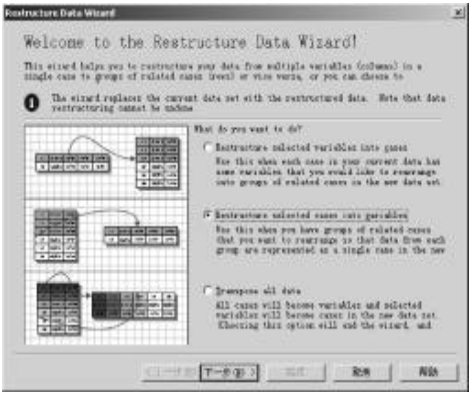


图 3.26 Restructure向导第一步

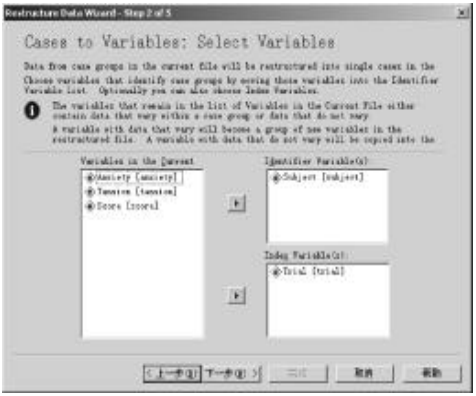


图 3.27 Restructure向导第二步

根据要求可知,用户指定被重复测量个体的 id标识变量和用于反映测量次别的 Index变量,此处分别为 subject和 trial,将它们分别选入 Identifier Variables框和 Index Variables框后单击“下一步”按钮,向导会进一步询问是否根据 id变量和 Index变量对数据进行排序,见图 3.28。

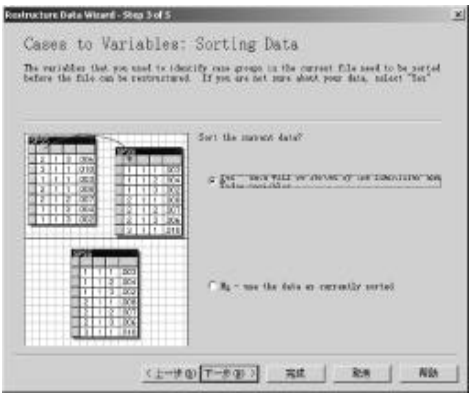


图 3.28 Restructure向导第三步图

系统默认为“Yes”,此时不做更改可以继续单击“下一步”按钮,也可以单击“完成”按钮,系统会自动判断所需的内容。单击“下一步”按钮,看看下边会有些什么内容,从图 3.29中可以看出,这一步是对重新排列以后的数据文件的结构进行设置,给出产生一条新记录的原记录的数目以及选择是否需要标识变量。即使用户对这个界面的功能不了解,根据向导的简短说明,也可以判断出此步骤的意图,这也是 SPSS友好的人机界面的一个展示。在这一步不做更改,单击“下一步”按钮,最后一个对话框用于选择是直接得到结果,还是生成相应的 SPSS程序,默认为前者。直接单击“完成”按钮,就可以得到相应的转换后的数据集,将该结果与数据 Anxiety 2. sav进行比较,可以看出除变量名和标签不同外,两个文件的内容实际上是一致的。另外,也可以看看系统在结果窗口中的汇总输出,如表 3.3和表 3.4所示,这常被用来检查是否操作有误。

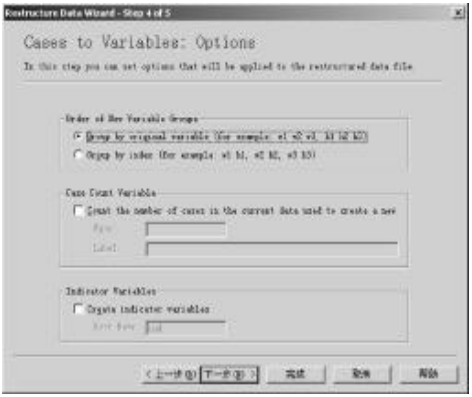


图 3.29 Restructure向导第四步

表 3.3 Generated Variables

Original Variable	Trial	Result	
		Name	Label
Score	1	score.1	score.1: Score
	2	score.2	score.2: Score
	3	score.3	score.3: Score
	4	score.4	score.4: Score

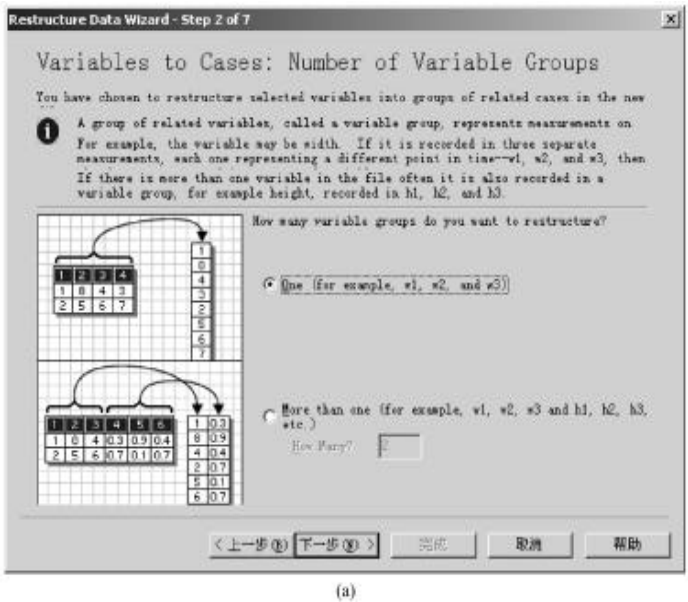
表 3.4 Processing Statistics

Cases In	48
Cases Out	12
Cases In/Cases Out	4.0
Variables In	5
Variables Out	7
Index Values	4

最后还有一个非常有趣的问题：本例中没有说明哪个变量需要转换，但最后程序只将 score 转换为了宽型格式，anxiety和 tension则直接携带了过来，未加转换。这是因为程序会自动扫描需要转换的变量，如果该变量在相同个体内取值均不变，则会被自动携带过来而不加转换，本例中的 anxiety和 tension正属于这种情况。显然，SPSS的这种设计大大方便了用户的使用。

3. 宽型格式转换为长型格式

下面来看看如何将宽型格式的数据转换为长型格式，有了前面的基础，这一部分内容大家应当很容易理解了。假设此处的任务是将 Anxiety 2. sav转换为如 Anxiety. sav的长型格式，则在第一个向导界面上选择第一项，单击“下一步”按钮后弹出界面如图 3.30 (a)所示，询问共有几组重复测量变量需要转换，此处只有一个，单击“下一步”按钮后进入最重要的变量选择界面（参见图 3.30 (b)）：Case Group Identifyication框用于设定重复测量个体的 id标识变量，此处设定为变量 Subject；中部的 Variables to be Transposed框组则用于设定被转换的变量组，首先将变量组名称改为 trial，随后在下方的列表中将 Trial.1～4选入。如果有多组变量需要转换，则依次设定即可；最下方的 Fixed Variable (s)框则用于选入携带变量，此处为 Anxiety和 Tension。



(a)

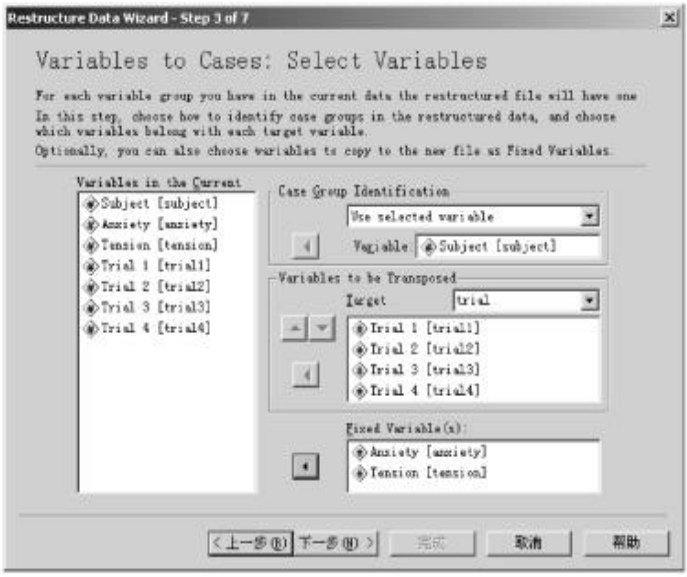
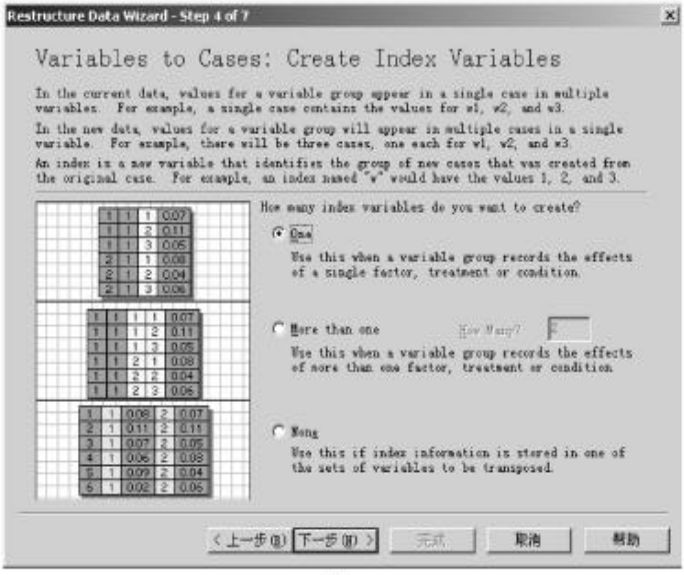
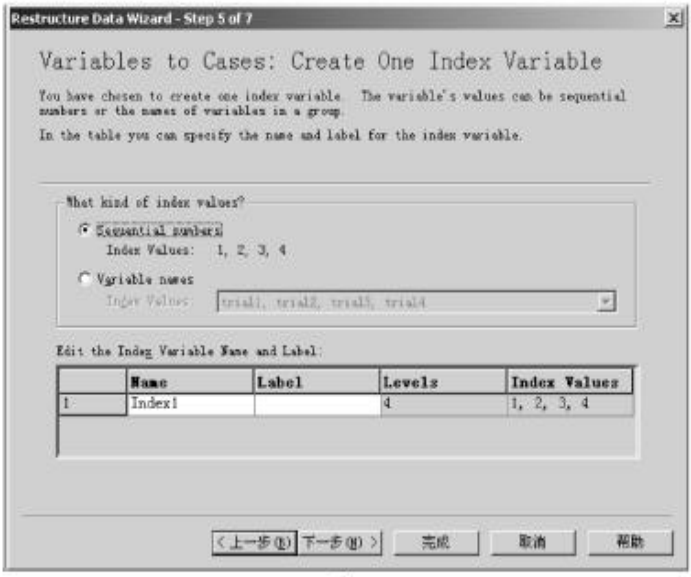


图 3.30 转换向导的第二、三个界面

在正确设定了变量选择界面之后,下面的工作就非常简单了,随后的 Create Index Variables 界面 (参见图 3.31 (a)) 用于设定重复测量指示变量 (如同本例中的变量 trial),而 Create One Index Variable界面 (参见图 3.31 (b)) 则具体设定该变量的数值。实际上现在就可以直接单击“完





(b)

图 3.31 转换向导的第四、五个界面

成”按钮结束本向导了，如果希望更详细地加以设定，则最后还有两个界面用于选择缺失值、未选中变量的处理方式以及是直接执行，还是生成相应的程序。

在本向导全部运行完毕后，数据就会被转换成长型格式，同时结果窗口中会给出操作的汇总表格如表 3.5和表 3.6所示。

表 3.5 Generated Variables

Name	Label
Index1	<none>
trial	Trial 1

表 3.6 Processing Statistics

Variables In	7
Variables Out	5

4. 数据转置

下面看看 Transpose过程，也就是数据重构向导的第三个功能。Transpose过程用于对数据进行行列转置，数据文件的转置就是将数据编辑窗口中数据的行列互换，即将记录转为变量，将变量转为记录后，重新显示在数据编辑窗口中，如图 3.32所示。

	varname	x	group
1	FIRST	.84	1.00
2	SECOND	1.05	1.00

	case_lbl	first	second
1	X	.84	1.05
2	GROUP	1.00	1.00

图 3.32 转置前的数据集和转置后的数据集

Transpose过程的对话框也非常简单（见图 3.33），左侧为候选变量框；右上方为 Variable(s)框，用于选入需要转置的变量，一般应选入除名称变量外的所有其他变量，如果有变量未选入，则转置时会被自动丢弃；右下方为 Name Variable框，用于指定原数据文件中记录转置后变量名的字符变量，但不是必需的，此时系统会将新变量自动按 var001、var002、…的顺序命名。

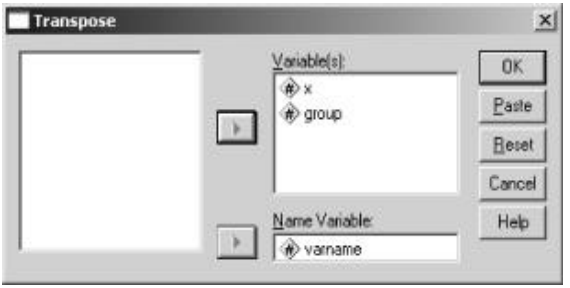


图 3.33 Transpose对话框

对统计分析的初学者而言，可能无法想像这个功能有什么用处。实际上，数据转置主要是用于编程，进行矩阵运算时的矩阵转置操作，对于只需要调用现成的分析程序，不需要自行编写算法的用户而言，转置功能的确没有多少实际用途。

3.3.4 多个数据文件的合并

进行统计分析的第一步工作就是将待分析的数据录入到 SPSS中。在数据量较大时，经常需要把一份大的数据分成几个小部分，然后再分别由不同的录入员进行录入，以缩短数据录入的时间。这样就会出现一份大数据分别存储在几个不同的数据文件中的现象。因此，将这若干个小的数据文件合并成一个大的数据文件是进行数据分析的前提。除此以外，如果数据有多个来源，则可能会使变量分散在几个文件中，需要按照某种规则加以合并后才能进行分析。

SPSS数据文件的合并方式有两种：纵向连接和横向合并，它们分别对应了上述的两种情况。数据集的纵向连接指的是几个数据集中的数据纵向相加，组成一个新的数据集，新数据集中的记录数是原来几个数据集中记录数的总和。横向合并指的是按照记录的次序，或者某个关键变量的数值，将不同数据集中的不同变量合并为一个数据集，新数据集中的变量数是所有原数据集中不重名变量的总和。

在 SPSS中，进行合并的文件必须都存储为 SPSS数据格式。如果是用程序方式，则可以一次实现多个数据文件的合并，但是，如果使用对话框方式，则一次只能进行两个 SPSS数据文件的合并，且其中一个必须是已被打开的当前数据文件。

1. 数据文件的纵向连接

SPSS数据文件的纵向连接或合并就是将数据编辑窗口中的数据与一个 SPSS数据文件中的数据进行首尾对接，即将一个 SPSS数据文件的内容追加到数据编辑窗口中当前数据的后面。纵向合并实质就是将两个数据文件的变量列，按照各个变量名的含义，一一对应进行首尾连接。

实现 SPSS数据文件的纵向合并应遵循两个条件：第一，两个待合并的 SPSS数据文件，其内容合并是有实际意义的；第二，为方便 SPSS数据文件的合并，在不同数据文件中，数据含义相同的列，最好起相同的名字，变量类型和变量长度也要尽量相同。这样，将方便 SPSS对变量的自动对应和匹配。

例 3.8 将数据 transform2.sav中的记录添加到 transform.sav中，注意在 transform2.sav中的变量 sex对应了 transform.sav中的 gender。

首先，在数据编辑窗口中打开数据文件 transform.sav，然后选择菜单 Data→Merge File→Add Cases，并选择待合并的文件 transform2.sav，出现如图 3.34所示的界面。

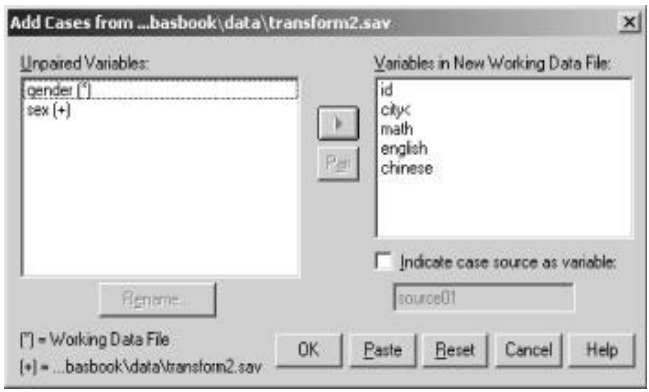


图 3.34 SPSS数据文件纵向合并窗口

在该窗口中，两个待合并的数据文件中共有的变量名会被自动对应匹配，并出现在 Variables in New Working Data File框中。SPSS默认它们具有相同的数据含义，自动成为合并后新数据文件中的变量。如果需要修改默认设置，可以将它们剔除到 Unpaired Variables框中。

在 Unpaired Variables框中，变量名后面有* 或 +号。* 表示该变量名是当前数据编辑窗口中的变量，+表示该变量名是待合并文件中的变量。可见，Unpaired Variables框中的变量名不是待合并的两个文件所共有的，是无法被自动对应匹配的，SPSS默认它们不具有相同的数据含义，不自动成为合并后新数据文件中的变量。同样地，用户可以修改这种默认设置，可以手工选择两个变量名，点击“Pair”按钮强行配对，表示它们具有相同的数据含义，并将其选入 Variables in New Working Data File框中。或者先点击“Rename”按钮改名后再指定配对。当然，也可以指定某变量不经任何对应匹配，强行进入 Variables in New Working Data File框中，但这种方式显然会造成缺失数据。

如果希望在合并后的数据文件中看出哪些记录来自合并前的哪个 SPSS数据文件，可以选 Indicate case source as variable项。于是，在合并后的数据文件中将自动出现名为 source01的变量，取值为 0或 1。0表示该记录来自第一个数据文件，1表示该记录来自第二个数据文件。

2. 数据文件的横向合并

SPSS数据文件的横向合并是将已有的一个 SPSS数据文件中的若干个变量加到当前数据编

辑窗口的数据中,即将一个 SPSS数据文件的内容接到数据编辑窗口中当前数据的右边,然后将合并后的数据重新显示在数据编辑窗口中。横向合并的实质就是将两个数据文件的记录,按照记录对应,一一进行左右对接。

实现 SPSS数据文件的横向合并应遵循三个条件,第一,如果不是按照记录号对应的规则进行合并,则两个数据文件必须至少有一个变量名相同的公共变量,这个变量是两个数据文件横向对应合并的依据,称为关键变量。如学号、贵宾卡号等,关键变量可以是多个;第二,如果是使用关键变量进行合并的对应,则两个数据文件都必须事先按关键变量进行升序排列;第三,为方便 SPSS数据文件的合并,在不同数据文件中,数据含义不相同的列,变量名不应取相同的名称。

例 3.9 将数据 transform3.sav中的变量添加到 transform.sav中。通过这个例子可以直观理解数据文件的横向合并。

首先,在数据编辑窗口中打开数据文件 transform.sav,然后选择菜单 Data→ Merge File→ Add Variables,并选择待合并的文件 transform3.sav,出现如图 3.35所示界面。可以看出,和纵向合并的操作窗口类似,两个待合并数据文件中的所有变量名出现在 New Working Data File框中,外部数据中与当前数据重复的变量,为免于重复而被列入 Excluded Variables(即这些变量是两个文件共有的变量,关键变量的名字一定在这个列表中可以找到)。变量名后面有* 或 +号。* 表示该变量名是当前数据编辑窗口中的变量,+表示该变量为待合并文件中的变量。SPSS默认仍以原变量名取名,成为合并后新数据文件中的变量。同样地,用户也可以做更改。

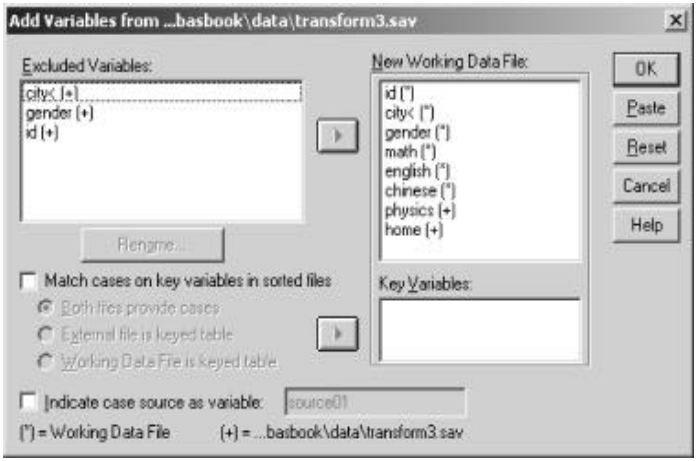


图 3.35 SPSS数据文件的横向合并

如果两个待合并的数据文件中的记录数据是横向顺序一一对应的,可单击“OK”按钮完成合并工作。否则,两个待合并的数据文件中的共有变量名出现在 Excluded Variables框中。点选 Match cases on key variables in sorted files项,并从 Excluded Variables框中选出一个或多个变量作为关键变量送到 Key Variables框中。

关于合并后的数据文件中的数据按哪种方式提供,SPSS有三个选项可供选择:

Both files provide cases:是 SPSS默认的方式,指合并后的数据由原来的两个数据文件共同提供,即由原来两个数据文件中的记录共同组成合并后的数据文件,当两个数据是逐条对应

时,用此选项。

External file is keyed table指在当前已打开数据基础之上,合并第二个数据文件中的变量数据,即合并后数据文件的记录仅包括当前数据编辑窗口中的记录。当外部数据根据关键变量是无重复记录,而当前数据根据关键变量是有重复记录时,用此选项。

Working Data File is keyed table指在第二个数据文件的基础之上,合并数据编辑窗口中的变量数据,即合并后数据文件的记录仅包括第二个数据文件中的记录,当当前数据根据关键变量是无重复记录,而外部数据根据关键变量是有重复记录时,用此选项。

另外,如果希望在合并后的数据文件中看出哪些记录来自合并前的哪个 SPSS数据文件,可以选 Indicate case source as variable项。于是,在合并后的数据文件中将自动出现名为 source01的变量,取值为 0或 1。0表示该记录来自第一个数据文件,1表示该记录来自第二个数据文件。

最后再次提醒大家,使用关键变量进行横向合并前,数据文件必须按照关键变量排序,否则相应的合并操作将会失败。

思考与练习

针对数据 Employee data. sav进行以下练习:

1. 试根据变量 bdate生成一个新变量"年龄"(提示:可以使用函数:XDATE. YEAR())。
2. 试根据 jobcat分组计算 salary的秩次。
3. 试根据雇员的性别变量对 salary的平均值进行汇总。
4. 在 Employee data. sav中生成新变量 grade,当 salary小于 20 000时取值为 d,当取值范围为等于 20 000或 20 000～50 000时为 c,等于 50 000或 50 000～100 000时为 b,大于等于 100 000时为 a。

参考文献

1 张文彤主编. SPSS 11统计分析教程(基础篇).北京:北京希望电子出版社,2002

2 SPSS Base 12 User's Guide. SPSS Inc. Chicago, Illinois, 2003

第二部分
统计描述与统计图表

第4章 连续变量的统计描述与参数估计

统计分析的目的是研究总体特征。但是,由于各种各样的原因,研究者能够得到的往往只能是从总体中随机抽取的一部分观察对象,它们构成了样本。只有通过对样本的研究,才能对总体的实际情况做出可能的推断。因此,在数据收集、整理完毕后,进行深入分析之前,首要的工作就是去了解这个数据的整体情况,通过数据来掌握一定的行业背景,随后才能考虑作深入的推断。

用少量数字(即描述指标)概括大量原始数字,对数据进行描述的统计方法即为描述性统计分析。所谓描述性统计分析,是针对统计学的另一大类——推断性统计分析而言的,后者指从样本信息来回推总体特征。在第二章中介绍了变量按其测量类型可以分为:Nominal变量(即名义型)、Ordinal变量(即定序型)和Scale变量(即定距型)。针对不同测量类型的变量(属性、字段),有不同的描述指标体系和统计图形与之对应。本章将讲述Scale变量,或者说连续变量的统计描述,而下一章将讲述Nominal变量和Ordinal变量以及多选题的统计描述。

4.1 连续变量的统计描述概述

当数据量较少时,如只有5个人的身高,或者7个人的性别资料时,研究者可以通过直接观察原始数据来了解几乎所有的信息。但是,接触到的数据量往往要远大于人脑可以直接处理、记忆的容量。这时就必须借助于各种统计指标来辅助完成对数据的描述工作了。而为了方便统计指标的应用,又以此为基础衍生出了各种描述用工具,最终再使用各种统计软件来加以实现,而SPSS就是最常用的一种。

4.1.1 统计描述中可用的工具

首先,在统计描述中最基本的工具就是列表进行原始数据的频数描述,特别是对于分类数据而言,频数表仍然是现在最常用的描述工具。但是,当数据量较大时,原始频数表显得过于冗长,如果希望深入发掘数据中蕴含的信息,则需要对数据加以浓缩汇总。

⑴ 各种初步汇总描述方法:最直接的汇总描述方法就是将原始数据按照其大小进行分组汇总,计算各组段的频数大小,最终汇总成相应的分组频数表或相应的分组直方图,汇总频数表可以反映出数据的大致趋势。除分段汇总以外,百分位数也能够对数据的分布特征进行刻画,多个百分位数组合起来,也能够反映出数据的分布特征来。但是分组汇总和百分位数对信息的利用仍然比较粗糙,均只能反映比较基础的信息,如果希望对数据的分布特征描述得更为简练,还需要更进一步。

⑵ 各种统计描述指标:这实际上是更复杂的各种描述工具的基础,是针对数据的某种特征

进行精确的数字呈现的一系列指标。对于样本而言,这些统计描述指标也可被称为统计量。常用的统计描述指标在连续变量中有均数、标准差、四分位数间距等,而在分类变量中则有比、率等。

⑶ 统计表:当数据比较复杂,所计算的统计指标较多时,直接观察计算出的数值比较困难,为此人们又会按照一定的排列方式将统计指标组织为一张表格,以方便使用,这就是所谓的统计表。在一张统计表中可以同时呈现多种统计指标,并进行复杂的样本分组、合并计算,因此,统计表是统计描述中常用的工具之一。

⑷ 统计图:统计表虽然能非常精确、详细地对统计指标进行陈列,但是不够直观,如果希望结果更为直观一些,则可以按照统计指标的大小将其绘制为一张图形,这就是所谓的统计图。例如对于连续变量数据,常用直方图、箱图等工具加以展示,而对于分类变量,则常用条图、饼图等加以展示。

显然,统计表和统计图都是建立在各种统计描述指标的基础上的,因此本章和下一章将对统计描述指标体系做详细的讲解,而第 6、7两章将进一步讲解如何利用统计指标制作统计表,第 8、9两章则会讲解统计图的绘制方法。对于在本章和下一章中可能会提前涉及到的统计图形,文中将仅作简单解释,不详细讨论,请大家参阅随后各章的相应内容。

4.1.2 连续变量的统计描述指标体系

图 4.1是对某人群体重分布情况绘制的直方图,这种图形是描述连续性变量最常用的工具,它实际上就是按照数据的大小将数值分成若干个组段,然后计算每个组段内的频数,最终用直条

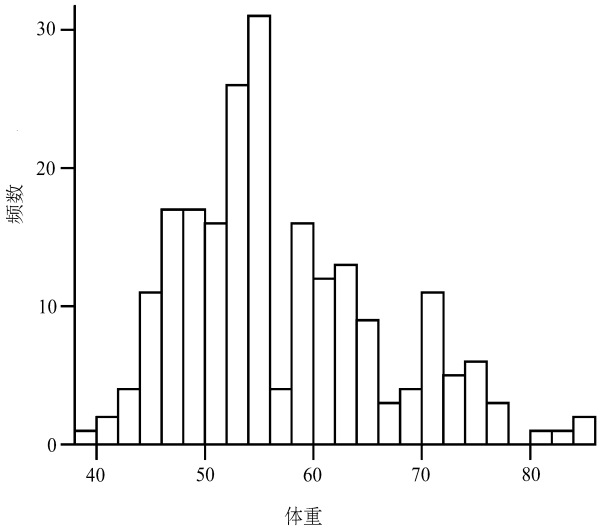


图 4.1 体重的直方图

的高低反映出来,它可以直观地反应数据的分布状况。通过对这张图形的观察,可以发现如果要使用统计指标对该数据加以描述,则主要是表现以下几个趋势:

⑴ 集中趋势（Central Tendency）：该人群的平均体重是多少？这可能是人们希望了解的最基本的汇总信息。人们常说美国人比中国人高，这并不是说美国人比中国人都高，比如姚明就要高于绝大多数美国人，这种说法实际上省略了“平均起来”这个定语。或者说，它实际上是关于数据的“中心位置”的某种表述。在统计学中，相应的用于描述集中趋势，或者说数据分布的中心位置的统计量就被称为位置统计量（Location Statistic）。常用的位置统计量有均数、中位数等，其中均数适用于正态分布和对称分布资料，中位数则适用于所有分布类型的资料，详述后面相关章节。

⑵ 离散趋势（Dispersion Tendency）：显然，仅仅反映数据的集中趋势是远远不够的，图 4.1 中还反映出体重在该人群中的分散状况，最轻的不到 40 kg，而最重的大约在 90 kg上下。应当有某种指标可以反映数据波动范围的大小，这被称为数据的离散趋势。比如人们常说的某国的贫富分化严重，或者某国卫生资源分配的公平性很差，偏远地区还缺医少药的时候，大城市的 CT 等大型医疗设备却大量闲置，占用了大量资源。这些实际上都是在讨论数据的离散趋势，而描述该趋势的统计量就被称为尺度统计量（Scale Statistic）。常用的尺度统计量有标准差、方差、四分位数间距等，其中标准差、方差只适用于正态分布资料，而四分位数间距则适用于各种分布类型的资料。

⑶ 分布特征（Distribution Tendency）：除以上两大基本趋势外，随着对数据特征了解的逐渐深入，研究者常常会提出假设，认为该数据所在的总体应当是服从某种分布的。那么，针对每一种分布类型，都可以由一系列的指标来描述数据偏离分布的程度。例如对于正态分布而言，偏度系数、峰度系数就可以用来反映当前数据偏离正态分布的严重程度。当然，相对而言，这些分布指标使用得较少。

⑷ 其他趋势：统计描述中还会用于许多其他指标，如可同时反映集中趋势和离散趋势的百分位数指标（Percentile），描述数据是呈单峰还是双峰分布，数据的分布是对称的还是偏态的，专门针对存在异常值的数据进行描述的 M统计量（M-Estimators）、极端值（Outlier）列表等，详后。

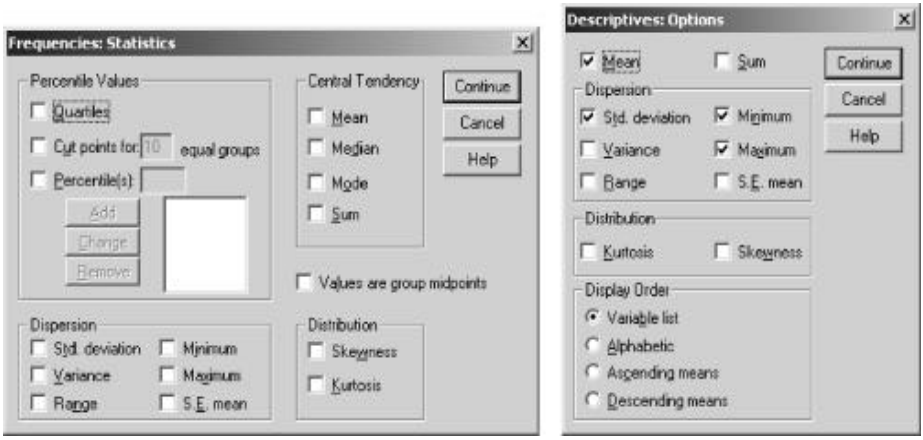
4.1.3 SPSS中的相应功能

SPSS的许多模块均可完成统计描述的任务，除各种用于统计推断的过程会附带进行相关的统计描述外，SPSS还专门提供了几个用于连续变量统计描述的过程，它们均集中在 Descriptive Statistics子菜单中：

⑴ Frequencies过程：其特色是产生原始数据的频数表，并能计算各种百分位数。由图 4.2 (a)可见，它所提供的统计描述功能非常全面，且对话框布置很有规律，基本上按照数据的集中趋势、离散趋势、百分位数和分布指标四大块将各描述指标进行了归类。有了上面的基础，读者使用它应当不存在任何的困难。

除统计指标外，Frequencies过程还可以为数据直接绘制相应的统计图，如用于连续性变量的直方图，用于分类变量的饼图和条图等。

⑵ Descriptives过程：该过程用于进行一般性的统计描述，相对于 Frequencies过程而言，它不能绘制统计图，所能计算的统计量也较少，但使用频率却是最高的。实际上从图 4.2 (b)所示的统计选项可以看出，该过程适用于对服从正态分布的连续性变量进行描述。



(a) (b)

图 4.2 Frequencies过程和 Descriptives过程的统计选项子对话框

(3) Explore过程：顾名思义，该过程用于对连续性资料分布状况不清楚时的探索性分析，它可以计算许多描述统计量，给出各种统计图，并进行简单的参数估计。本章最后的分析实例将以该过程为主加以讲解。

(4) Ratio过程：功能比较特殊，用于对两个连续性变量计算相对比指标，它可以计算出一系列非常专业的相对比描述指标，相对而言使用面比较窄，因此本书将不对它做过多介绍，对此感兴趣的朋友请参见笔者前作《SPSS 11统计分析教程》(基础篇)。

4.2 集中趋势的描述指标

怎样将一个变量的所有个体的值汇总为一个数字，使这个数字代表原数据的中心趋势或平均水平？统计学家提供了多种统计量来代表原始数据的中心趋势，如平均值、中位数和众数等。

4.2.1 算术均数

平均数用于反映一组数值的平均水平，包括算术均数、几何均数、调和均数等，但是以算术均数最为常用，往往也直接将算术均数简称均数。

算术均数 (Arithmetic Mean) 是最常用的描述数据分布的集中趋势的统计量。总体均数 (Population Mean) 用希腊字母 μ表示，样本均数常用 $\bar{X}$表示。

1. 算术平均数的定义和性质

实际上，大家从小学起就已经学习了相关的知识，对一组数据 $X_1, \cdots, X_n$ 而言，其均数的算法为各数据直接相加，再除以总例数 n，即：

$$\bar{x}=\frac{x_1+x_2+\cdots+x_n}{n}=\frac{\sum x_i}{n}$$

显然,有各个变量值与均数离差之和等于零。即:

$$\sum(x_i-\bar{x})=0$$

算术平均数的这条数学性质说明,均数的实质是把总体各单位的差异全部抽象化,采用取长补短的方法把变量值小于平均数的负离差全部用大于平均数的正离差抵消补齐。

除上面的性质外,各个变量值与平均数离差平方之和为最小值。即:

$$\sum(x_i-\bar{x})^2<\sum(x_i-a)^2\quad(a\neq\bar{x})$$

算术平均数的这条数学性质说明,以任意不为平均数的数值为中心计算的离差平方和大于以平均数为中心的离差平方和,因此,算术平均数是误差最小的总体代表值。

2. 均数的意义

任何一个平均数首先是同类现象的平均数,这是平均数的同质性。任何一个平均数总是一个平衡点。在这个平衡点的两边有多有少、有大有小、有高有低、有胖有瘦。而且总是多少相等,大小相同,高低适中,胖瘦相抵。这就是说,用平均数作为观测数据的代表在整体上是没有误差的,而且数学上可以证明,平均数的误差平方和也比其他任何一个数都小。统计学中著名的“最小二乘法”就是根据这个结论建立起来的。但是,由于平均数只是一个平衡点,如果两边加上或去掉相同的砝码,而不管砝码是多少这杆天平总能保持平衡。

平均数最重要的意义在于它高度浓缩了数据,使大量的观测数据转变为一个代表性数值。用平均数作为变量的集中值不仅考虑到变量值的频次、次序,而且还考虑到它的大小。数据资料中任何频次、次序和数值大小的变化,都会引起平均数的改变。因此它是灵敏的,也是对资料所提供信息运用得最为充分的。

但平均数在高度概括观测数据从而使问题简单化的同时,却丢失了某些有用的信息,一方面它把各个观测数据之间的差异性掩盖了起来,另一方面由于平均数对个别极端值反应比较灵敏,因而平均数在某些情况下可能具有一定的欺骗性,这时它就有可能传递不准确的信息。

3. 均数的适用范围

虽然平均数对资料的信息利用最充分,但对严重偏态的分布,会失去它应有的代表性。例如,一个国家会因某些富翁的存在,使平均收入变得很高。假设某单位有 6个人,5个员工,1个经理。员工的月收入分别是:360元、380元、400元、420元、440元,经理的月收入为 40 000元,他们的平均月收入为 7 000元。显然这时用平均数就不能很贴切地反映他们收入的一般水平。所以,平均数的一个主要缺点是容易受极端值的影响。因此,对于偏态的分布,应使用中位数作为集中趋势的统计量。只有单峰和基本对称的分布情况下,使用平均数作为集中趋势描述的统计量才是合理的。由于在统计技术中,发展更多的是平均数,而不是中位数或众数等。因此,应该设法更多地使用平均数,必要时可以考虑对数据进行变量变换,以达到对称分布的要求。

严格地讲平均数只适用于定距变量。但有时对于定序变量,求平均等级也可以使用平均数。对于定类变量,如果人为地把每一类赋予一个数值,如用 1代表男,2代表女,那么男性在总体中

所占的比例,实际就是一种特殊的平均数。

4.2.2 中位数

中位数 (Median)是将总体各单位的标志值按大小顺序排列,处于中间位置的那个标志值。它把全部标志值分成两部分,一半标志值比它小,一半标志值比它大。

1. 中位数的定义

对于未分组的原始资料,首先必须将标志值按大小排序。设排序的结果为:

$$X_1 \leq X_2 \leq X_3 \leq \cdots \leq X_n$$

则中位数就可以按下面的方式确定:

$$M = X_{(n+1)/2}, \quad \text{当 } n\text{为奇数时}$$

$$M = (X_{n/2} + X_{n/2+1})/2, \quad \text{当 } n\text{为偶数时}$$

中位数作为分布数列中处于中等水平的代表值,能够将全部总体单位按标志值的大小等分为两个部分,所以中位数又称为二分位数。

对于按照频数方式分组录入的资料,其中位数的确定方式相对复杂一些,感兴趣的读者可以参看专业统计书籍。

2. 中位数的适用范围

中位数是位置平均数,因此它不受极端值的影响,在具有个别极大或极小标志值的分布数列中,中位数比算术平均数更具有代表性。例如上面员工收入的例子,其中位数就是 410元,显然要比均数更能够代表数据的集中趋势。

中位数适用于任意分布类型的资料,不过,由于中位数只考虑居中位置,其他变量值比中位数大多少或小多少,它是无法反映出来的。所以,用中位数来描述连续变量会损失很多信息。当样本量较小时,中位数会不太稳定,并不是一个好的选择。因此,对于对称分布的资料,分析者往往优先考虑使用均数,仅仅是对均数不能使用的情况下才用中位数加以描述。

中位数对于定序变量、连续变量都可以使用。对于定序变量来说,虽然有众数和中位数两种统计量可供选择,但是,由于众数不考虑变量的次序关系,用众数来描述定序变量会损失很多信息。因此,对于定序变量,应采用中位数来反映更多、更准确的信息。

4.2.3 其他集中趋势描述指标

除上述最常用的两种指标外,在 SPSS中还可以计算一些更为复杂和专业的统计描述指标,这里简介如下:

1. 截尾均数

由于均数较易受极端值的影响,因此可以考虑将数据进行排序后,按照一定比例去掉最两端

的数据,只使用中部的数据来求均数。如果截尾均数和原均数相差不大,则说明数据不存在极端值,或者两侧极端值的影响正好抵消;反之,则说明数据中有极端值,此时截尾均数能更好地反映数据的集中趋势。

常用的截尾均数有 5%截尾均数,即两端各去掉 5%的数据。在 SPSS中 Explore过程可以自动计算 5%截尾均数。

2. 几何均数

几何均数适用于原始数据分布不对称,但经过对数转换后呈对称分布的资料。如医学中的血清滴度资料就常用几何均数描述其分布的集中趋势。样本几何均数常用 G表示,其计算公式是:

$$G = \sqrt[n]{X_1 X_2 \cdots X_n}$$

利用对数的性质,上述公式可表达为:

$$G = \lg^{-1}\frac{\sum \lg X_i}{n}$$

可以发现,几何均数实际上就是对数转换后的数据 lgX的算术均数的反对数。

在 SPSS中,几何均数可以在 Report子菜单中的 4个报表过程中计算输出。

3. 众数 (Mode)

众数指的是样本数据中出现频次最大的那个数字,众数容易理解,也不受极端值影响,但不易确定,且没有太明确的统计特性。

众数适用于任何层次的变量,特别适用于单峰对称的情况,是比较两个分布是否相近首先要考虑的参数。但是,由于众数仅使用了资料中最大频次这一信息,所以它对资料的使用是不完全的,提供的信息有限,用它来反映连续变量会损失很多信息。对于多峰的图形分布,一般也不用它来描述。因此,这里不做详细介绍。

在 SPSS中,众数可以在 Report子菜单和 Tables子菜单的全部报表过程和制表过程中计算输出。

4. 调和均数

调和均数用符号 H表示,现在已经很少使用,它实际上是观察值 X倒数之均数的倒数,常用于完成的工作量相等而所用时间不同的情况,主要用来求平均速度。实际上,中学物理中学习过的并联电路的总电阻就是各分电路电阻的调和均数,各原始数据的大小相差越悬殊,该均数的"调和"作用就越明显。

在 SPSS中,调和均数可以在 Report子菜单中的 4个报表过程中计算输出。

4.3 离散趋势的描述指标

和集中趋势一样,离散趋势也有一系列的描述指标,本节将就一些常用的指标一一加以讲解。

4.3.1 全距

全距 (Range) 又称为极差,是一组数据中最大值 (Maximum) 与最小值 (Minimum) 之差。它是最简单的变异指标:

$$R = X_{max} - X_{min}$$

极差反映的是变量分布的变异范围或离散幅度,在总体中,任何两个标志值之差都不可能超过极差。极差计算简单,含义直观,运用方便。但存在两点不足:一是它仅仅取决于两个极端值的水平,不能反映其间的变量分布情况,提供的信息太少;二是它容易受个别极端值的影响,不符合稳健性的要求。

一般情况下,全距只用于预备性检查,目的是大体上了解数据的分布范围,以便确定随后分析的方法。

4.3.2 方差和标准差

1. 方差 (Variance) 和标准差 (Standard Deviation) 的定义

相对而言,方差和标准差的计算比较复杂,因此这里将从其计算原理开始谈起。首先,对于每个数据而言,其离散程度的大小就是和均数的差值,简称离均差,它可以用来描述个体的变异大小。那么,离均差之和能否表示整个样本的离散程度大小呢?答案是否定的,因为根据均数的性质,所有数据的离均差之和应当正好为 0,这是由于大于均数和小于均数的离均差正好能够完全抵消。为此,可以考虑先将离均差取绝对值,然后再求和,这样就不会出现正负抵消的情况了。显然,离均差绝对值之和可以表示数据离散程度的大小。

但是,使用离均差绝对值之和来表示离散程度仍有不便之处,大家都知道绝对值符号在数学推导中是非常难处理的,该指标很难用来进行后续的统计推断,因此人们又改用将各离均差先平方再求和,这样仍然可以解决符号的问题,同时又可以进行后续的数学推导,该指标被称为离均差平方和 (Sum of Squares of Deviations from Mean, SS)。

离均差平方和在使用上比绝对值要方便一些,但是,它的大小显然是和样本量有关的,观察单位越多,该指标就会越大,因此如果要客观反映变异程度的大小,就应当去除样本量的影响。为此将离均差平方和除以观察例数 N所得,这就是方差:

$$\sigma^2 = \frac{\sum (X_i - \mu)^2}{N}$$

方差相当于平均了每个数据的离均差的平方值,从而克服了离均差平方和受样本含量影响的缺点。故方差可用于不同含量样本数据分布离散程度的比较。方差越大,数据分布离散程度越大。

对于样本数据而言,方差的计算公式有所不同:

$$S^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

其中的 $n-1$被称为自由度（Degree of Freedom），它描述了当样本选定时 n个 X中能自由变动的 X变量值）的个数，由于公式中需要使用均数，这是一个限制条件，因此样本量为 n的样本实际上只有 $n-1$个可以自由取值，最后一个数值可以通过均数算出来。自由度在统计学中也是一个非常重要的概念，后面还会反复遇到。

最后，方差在使用上还有一点小小的不便，就是量纲不合常理，以身高为例，原始数据的量纲为米，则方差的量纲就是其平方，即平方米，这显然很别扭。为此又将方差开平方，这就是所谓的标准差，总体和样本的标准差分别用 σ和 S来表示。标准差度量了偏离平均数的大小，相当于平均偏差，可以直接地、概括地、平均地描述数据变异的大小。对于同性质的数据来说，标准差越小，表明数据的变异程度越小，即数据越整齐，数据的分布范围越集中；标准差越大，表明数据的变异程度越大，即数据越参差不齐，分布越分散。

2. 方差和标准差的适用范围

由于标准差和方差的计算涉及每一个变量值，所以它们反映的信息在离散指标中是最全面、最可靠的变异描述指标。方差还具有可加性，能够参与进一步的统计运算。不过，也正是由于标准差和方差的计算涉及每一个变量值，所以，它们也会受到极端值的影响，当数据中有较明显的极端值时不宜使用。另外，它们在计算中实际上都使用了均数，因此实际上只有均数能反映集中趋势时才能使用方差和标准差来反映离散趋势。因此，实际上方差和标准差的适用范围应当是正态分布。

4.3.3 百分位数、四分位数与四分位数间距

全距的数据最不可靠，因为全距只由数据中的两个极端数据来决定，其余数据均不起作用。为了尽量减少全距缺点，人们又使用了分位差。分位差是对极差指标的一种改进，是从变量数列中剔除了一部分极端值之后重新计算的类似于极差的指标。常用的分位差有四分位差、十分位差以及百分位差。这里以四分位差为例加以说明。

1. 百分位数、四分位数与四分位数间距的定义

百分位数（Percentile）是一种位置指标，用 P_x表示。一个百分位数 P_x将一组观察值分为两部分，理论上有 $x\%$的观察值比它小，$(100-x)\%$的观察值比它大。前面所学习过的中位数实际上就是一个特定的百分位数，即 P_{50}。

除中位数外，常用的百分位数还有四分位数，它实际上是三个数值的总称，分别是 P_{25}、P_{50}和 P_{75}分位数。这三个分位数正好是能够将全部总体单位按标志值的大小等分为四部分的三个数值，符号分别记为 Q_1、Q_2和 Q_3。在许多统计书籍中，也将第一个四分位数 P_{25}称为“下四分位数”，第三个四分位数 P_{75}称为“上四分位数”，分别用符号 Q_L和 Q_U表示。上、下四分位数的差值被称为四分位数间距：

$$Q \cdot R = Q_3 - Q_1$$

显然，P_{25}和 P_{75}这两个分位数间包括了中间 50%的观察值，因此四分位数间距既排除了两段极端值的影响，又能够反映较多数据的离散程度，是当方差、标准差不适用时较好的离散程度描述指标。

同样的道理，还可以计算十分位差、百分位差等。它们的作用都是排除少数极端值对分布变异范围的异常影响。分位的程度越高，分位差所排除的极端值的比例就越小，保留的信息就越多。分位的程度越低，分位差所排除的极端值的比例就越大，保留的信息就越少。实际分析时，需要根据具体情况和要求选择使用。

2. 四分位数与四分位数间距的适用范围

计算四分位差的直接目的是排除部分极端值对变异指标的影响，其计算可以看成是首先从总体分布中剔除最大和最小各 1/4的单位，再对剩下的总体半数单位计算“全距”。因此，四分位数间距可以适用于任意分布类型的资料，它与全距（极差）的区别在于计算范围较窄，反映的是处于分布中间半数单位的变异幅度。

百分位数并非由全部观察值总和计算而来，因此它不如均数和标准差精确，然而中间部分的百分位数因不受极端数据的影响，具有较好的稳定性。但是，靠近两端的百分位数只有在样本含量足够大的时候才比较稳定，如当样本量为 100例时，比 P_{95}大的数值只有 5个，换言之，这 5个数字就决定了 P_{95}的大小。显然，此时 P_{95}是很不稳定的。因此，当样本量较小时，不宜取太接近两端的百分位数。而当样本含量很少时，“百分”位数已名不副实，就更加不用考虑了。

最后需要指出的是，严格地讲百分位数并不应当被仅限于描述离散程度，显然，它也可以对数据的集中趋势等其他特征进行描述，而多个百分位数联合起来，实际上就可以完整地反映整个数据的分布规律。这一点在本章第一节已有所提及，这里再次强调一下。

4.3.4 变异系数

当需要比较两组数据离散程度大小的时候，往往直接使用标准差来进行比较并不合适。这可以被分为两种情况：

⑴ 测量尺度相差太大：例如，希望比较蚂蚁和大象的体重变异，蚂蚁的体重以克计，而大象的体重以吨计，如果直接比较，显然永远都是大象的体重变异更大，但这显然是不合理的，因为体重相差 1 kg对大象的体重而言根本就算不了什么，而蚂蚁则永远也做不到。

⑵ 数据量纲不同：例如希望比较身高和体重的变异程度，两者的量纲分别是 m和 kg，那么，究竟是 1m大，还是 2kg大？根本就没法比较，完全是一笔糊涂账。

在以上情形中，就应当消除测量尺度和量纲的影响，而变异系数（Coefficient of Variation），可简记为 CV就可以做到这一点，它是标准差与其平均数的比率。样本变异系数计算公式为：

$$CV = S/\bar{X}$$

计算出的 CV没有量纲，同时又按照其均数大小进行了标化，这样就可以进行客观的比较。

4.4 连续变量统计描述实例

在系统学习了连续变量的统计描述指标体系后，下面将用一个具体的分析实例来看一下各种描述指标在 SPSS中的实现方法。

4.4.1 数据背景介绍

本例是一次实际调查的部分问卷数据，调查对象为上海部分大专院校的大学生，文件名为 student.sav。主要调查内容和封闭型题目的选项代码如下：性别（1男、2女），出生年、月、日（具体数字），身高（cm），体重（kg），血型（A、AB、B、O），血型代码（1 A、2 AB、3 B、4 O），教育背景（1重点大学本科、2普通大学本科、3大专、4中专/职校），学科（1文史、2理工、3其他），男、女身高级别（1低、2中等、3高，但两者的划分标准不一样），男、女体重级别（1轻、2中等、3重，两者的划分标准不一样）和季度（具体数字）。

需要说明的是，后面的 5个变量：男生身高级别（hm）、女生身高级别（hf）、男生体重级别（wm）、女生体重级别（wf）和季度（quarter），是通过 SPSS的 Recode过程，从前面的相应变量中，经过 Into Different Variables..变换而来。

4.4.2 使用 Explorer过程进行分析

1. 分析操作

这里以 student.sav数据为例，对男性和女性身高数据分别进行描述，具体步骤如下：

Analyze→ Descriptive Statistics→ Explore
Dependent Variables框：height
Factor List框：sex
Plots..：
Descriptive ☑ Histogram
Continue
OK

Explore主对话框如图 4.3所示，Dependent List框用于选入需要分析的变量，下方的 Factor List框用于选入分组变量，从而将希望描述的变量按该因素的取值分组分析，本例中为性别。Explore过程中的 Statistics和 Plots子对话框如图 4.4所示。

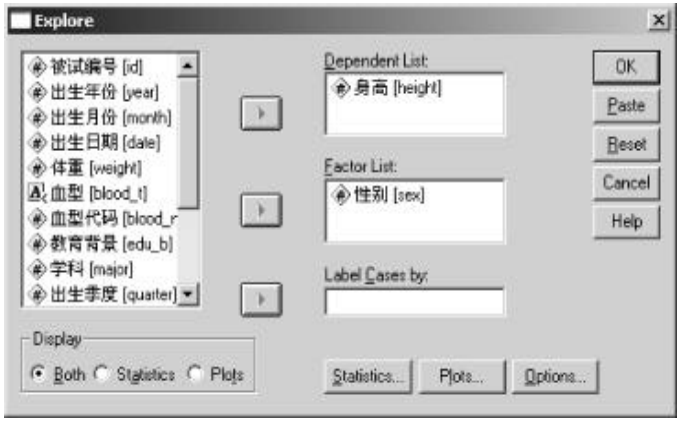
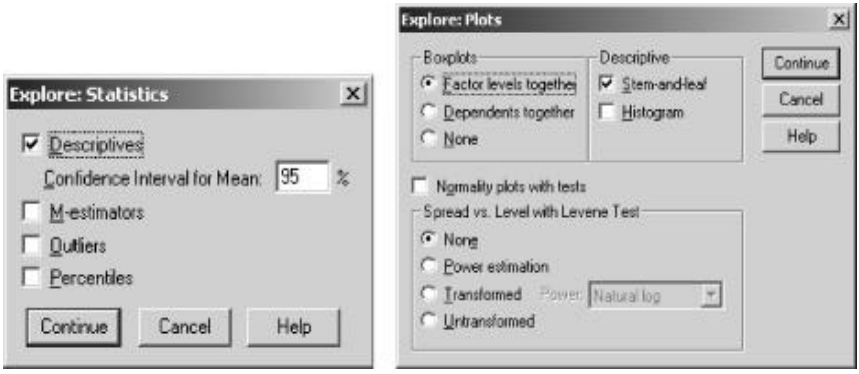


图 4.3 对连续变量进行描述性分析的 Explore过程主对话框



(a) (b)

图 4.4 Explore过程的 Statistics和 Plots子对话框

2. 基本的分析结果

分析结果中首先会给出标题“Explore”,表明随后的输出都属于 Explore过程。

表 4.1 Case Processing Summary

		Cases					
		Valid		Missing		Total	
	性别	N	Percent	N	Percent	N	Percent
身高	男	69	95.8%	3	4.2%	72	100.0%
	女	146	99.3%	1	.7%	147	100.0%

首先是例行的处理记录缺失值情况报告 (见表 4.1),可见对于身高而言,男性、女性两组均存在缺失值,其中男性 3例,女性 1例,最终进入分析的各为 69和 146例有效值。

表 4.2 Descriptives

	性别			Statistic	Std. Error
身高	男	Mean		174.71	.671
		95% Confidence Interval for Mean	Lower Bound	173.37	
			Upper Bound	176.05	
		5% Trimmed Mean		174.70	
		Median		175.00	
		Variance		31.062	
		Std. Deviation		5.573	
		Minimum		159	
		Maximum		188	
		Range		29	
		Interquartile Range		8	
		Skewness		-.034	.289
		Kurtosis		.138	.570

记录汇总报告之后给出的就是身高的统计描述表格,因本例中的结果输出较长,为便于解释,这里仅给出表格上半部男性的分析结果(见表 4.2)。可见 Explore过程的输出结果较多,这里依次解释如下:

⑴ 集中趋势指标:首先可以看到 69名男性学生的平均身高为 174.71 cm (Mean),去掉两侧各 5%的极端值后,截尾均数为 174.70 cm (5% Trimmed Mean),中位数为 175 cm (Median)。对于对称分布,且不存在极端值的数据而言,均数、截尾均数和中位数应当基本相同,显然本例符合这种情况,因此从上述指标及可推测出数据应当是对称分布的。

⑵ 离散趋势指标:身高的方差为 31.062 cm (Variance),其平方根即标准差,大小为 5.573 cm (Std. Deviation)。全部男生中最矮的为 159 cm (Minimum),最高的为 188 cm (Maximum)。两者之差即为全距 29 cm (Range),中间一半的男生的身高差即为四分位数间距 8 cm (Interquartile Range)。

⑶ 分布特征指标:表 4.2最下方还会给出表示数据偏离正态分布程度的偏度系数和峰度系数,及其各自的标准误,关于它们的详细解释,请参阅 4.5节。

⑷ 参数估计:以上结果实际上还会给出总体均数的参数估计结果,可见均数的标准误为 0.671 cm,相应的总体均数 95%可信区间为 173.37~176.05 cm,关于可信区间的详细解释详见 4.5节。

女生身高情况请大家自己分析,这里不再详述。

在统计描述表格之后,Explore过程还会给出身高分性别的茎叶图和箱图,从图形分布上可以看出,分性别的升高基本上呈对称的分布状态。对这两种图形的介绍请读者参见第 8、9两章,这里不再详述。

3. 输出百分位数和极端值列表

除默认的统计量输出外，Explore过程中还可以计算一些更深入的描述统计指标，如选中Statistic子对话框的 Outliers复选框后，即可输出如表 4.3所示的极端值列表。

表 4.3 Extreme Values

	性别			Case Number	Value
身高	男	Highest	1	180	188
			2	5	186
			3	154	186
			4	149	184
			5	150	183
		Lowest	1	20	159
			2	18	165
			3	11	165
			4	7	165
			5	183	167[a]

a. Only a partial list of cases with the value 167 are shown in the table of lower extremes.

这里同样只给出了男性的情况，表格中会输出 5个最大值与 5个最小值以及这些数值所对应的记录号，从两侧极值的大小可见，在最大、最小两个方向上并没有特别明显的异常值，该结果同样支持前面得出的数据分布基本对称的结论。

如果选择 Percentiles复选框，则会输出如表 4.4所示的百分位数表。

表 4.4 Percentiles

			Percentiles						
		性别	5	10	25	50	75	90	95
Weighted Average(Definition 1)	身高	男	165.00	168.00	170.00	175.00	178.00	182.00	185.00
		女	155.00	156.70	159.00	163.00	166.00	170.00	172.00
Tukey's Hinges	身高	男			170.00	175.00	178.00		
		女			159.00	163.00	166.00		

上表会输出第 5%、10%、25%、50%、75%、90%、95%分位数，并分别采用了两种算法，当数据量较大，且基本无重复值时，两法的结果相同，反之，则加权平均法会对数据进行内插，两种方法的结果会略有区别。

4.4.3 使用其他过程进行分析

上面使用 Explore过程对数据进行了分析，下面来演示一下另外两个过程的分析结果。但是，由于另两个过程不能直接对身高进行分组描述，因此这里仅给出不分性别的分析结果，希望

给出分组描述的读者可以先采用第 3章介绍过的 SelectCases过程进行数据拆分。

1. Descriptive过程的结果

该过程的操作非常简单,只需要将希望描述的变量选入即可,本例中身高的分析结果如表 4.5所示。

表 4.5 DescriptiveStatistics

	N	Minimum	Maximum	Mean	Std. Deviation
身高	215	151	188	166.67	7.668
Valid N (listwise)	215				

由于这里的大部分内容都在上一节见过,因此就不再多解释了。

2. Frequencies过程的结果

Frequencies过程默认值给出原始频数表,如果希望得到各种统计量,则需要分析者自行加以指定。例如,在上述的分析中,已经得到了描述集中趋势的均值、中位数等,以及描述离散趋势的方差、标准差、极差等统计量。如果还希望知道身高的具体四分位数及 P_5、P_{95}百分位数是多少,则可以利用 Frequencies过程来得到。具体步骤如下:

Analyze→ Descriptive Statistics→ Frequencies
Variables框 :height
Statistics:
Percentile Value: ☑ Quartiles
Percentile Value: ☑ Percentiles: 5: Add ☑ Percentiles: 95: Add
Continue
OK

表 4.6 Statistics

身高		
N	Valid	215
	Missing	4
Percentiles	5	155.80
	25	160.00
	50	165.00
	75	172.00
	95	180.00

从表 4.6中可知,所有学生身高的四分位数为 160 cm、165 cm和 172 cm。意味着,有 1/4的学生身高矮于 160 cm,1/2的学生身高较 165 cm矮,1/4的学生身高高于 172 cm。另外,90%的

学生身高在 155.8～180 cm之间。

4.5 连续变量的参数估计

通过统计描述,研究者已经可以对样本数据的情况有详细的了解。但是,研究的真正目的是考察样本所代表的总体情况如何,这里必然会涉及到如何将样本信息用来推断总体特征的问题,如总体的集中趋势、离散趋势究竟如何?这种根据样本数据对总体的客观规律性作出合理估计的过程被称为统计推断(Statistical Inference),它又可以被分为参数估计和假设检验两大类,而这里涉及到的用样本信息来推断总体特征的推断就被称为总体的参数估计。本节将介绍如何进行连续变量的参数估计。

4.5.1 正态分布

在进行总体数据的描述时,人们往往会对该总体的分布规律作一定的假定。比如假定身高服从正态分布。这些模型假定基本上是根据经验而得,所以仅仅是对现实世界的一个近似。由于分布是由参数确定的,这样就可以将总体描述的任务归结对几个参数的估计(此即参数估计名称的由来)。而且,如果能确认变量符合或大致符合某种分布的话,就可以选择有针对性的研究方法对该数据进行正确和精确的分析。

常见的连续分布有正态分布、均匀分布、χ^2分布、t分布和 F分布等。这里仅介绍统计学中最为重要的正态分布。正态分布又称高斯分布,虽然当初它是数学家高斯作为描述误差(如测量误差)分布规律的模型提出来的,并将其用于天文研究。但令人惊讶的是,最终这条曲线竟为描述来自不同领域的数据分布规律提供了一个完美的模型。

正态分布是概率统计中最重要的一种分布,其重要性可以从以下两方面来理解:在自然现象和社会现象中,大量的随机变量都服从或近似服从正态分布,如测量的偶然误差、炮弹落点距目标的偏差、一个地区男性成人的身高及体重、海洋波浪的高度、电子管噪声电流、工业产品的尺寸(直径、长度、宽度等)、某地区的每日用水量及用电量等都可看作服从或近似服从正态分布。一般说来,若某一随机变量是受多种相互独立的随机因素的影响,而每一种随机因素所起的作用又是极其微小的,那么该随机变量就近似服从正态分布。正态分布具有许多良好的性质,很多分布可以用正态分布来近似描述,另外一些分布又可以通过正态分布来导出。所以正态分布在理论与实践中都占有重要的地位。

1. 正态分布的定义

若连续性随机变量 X的概率分布密度函数为

$$f(X)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(X-\mu)^2}{2\sigma^2}}$$

其中,μ为平均数,σ^2为方差,则称随机变量 X服从正态分布(Normal Distribution),记为 X～N

(μ, σ^2)。不同的 μ 不同的 σ,对应于不同的正态分布。

图 4.5即为正态分布图,正态分布的密度曲线 (横轴为值,纵轴为频率)是一个对称的钟形曲线 (最高点在均值处)。显然,正态分布是一族分布,其曲线依均值和标准差而略有区别。该连续变量落在某个区间的概率就等于在这个区间上,该曲线下的面积,而曲线下的总面积为100%,代表概率总和为 100%。

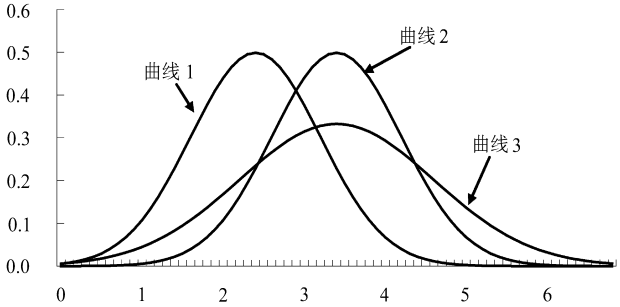


图 4.5 不同均数 μ 不同标准差 σ 的正态分布示意图

2. 正态分布的特征

从正态分布曲线,可以总结出其分布特征如下:

(1) 正态分布曲线是一条对称曲线,关于均数对称,因此均数被称为正态分布的位置参数。

(2) 曲线是单峰,在均值处达到最高点。

(3) 正态分布曲线峰的矮阔与尖峭与标准差有关。标准差越大,个体差异越大,正态曲线也越矮阔;反之,标准差越小,个体差异越小,正态曲线也越尖峭。因此标准差被称为正态分布的尺度参数。

(4) 曲线无论向左或向右延伸,都越来越接近横轴,但不会与横轴相交,以横轴为渐进线。

除此以外,正态曲线下的面积也有一定的分布规律,根据经验法则,有:

(1) 约 68%的个体的取值与平均数的距离在 1个标准差 $(\mu \pm \sigma)$ 之内,或者说一个标准差范围内的曲线下面积为 68%。

(2) 约 95%的个体的取值与平均数的距离在 1.96个标准差 $(\mu \pm 1.96\sigma)$ 之内。

(3) 99%个体的取值与平均数的距离在 2.58个标准差 $(\mu \pm 2.58\sigma)$ 之内。

根据上述规律,可以做出一些相应的总体推断。例如,某单位所有男性员工的平均身高为175 cm,身高的标准差为 5 cm,在身高服从正态分布的前提下,可以得到这样的推断:约 68%的男性员工的身高在 170 cm ~180 cm之间,约 95%的男性员工的身高在 165 cm~185 cm之间。

3. 标准正态分布

统计分析中经常需要求曲线下面积,但这就需要为每个不同的分布单独计算面积分布规律。为了制一张可供不同的 μ σ 共同使用的表,可以考虑引进以下变换:

$$u=\frac{X-\mu}{\sigma}$$

这样做相当于将分布的位置参数移动到 0处,使曲线沿 y轴对称,并且将分布的尺度参数固定为 1。从而将原来的正态分布 $N(\mu,\sigma^2)$,变换成了均数为 0,标准差为 1的正态分布,该分布被称为标准正态分布 (Standard Normal Distribution),而上述变换则被称为标准化变换。在国外,标准正态分布被称为 u分布或者 z分布,因此变换也被称为 u变换或者 z变换。

标准化变换和标准正态分布的意义非常重大,因为这样只需要知道标准正态曲线下面积的分布规律,就可以解决所有正态分布的曲线下面积计算问题了,只需将其进行标准正态变换即可。

在 SPSS中的 Descriptive过程可以将原变量变换为标准正态分布下的得分,只需要选中主对话框左下角的 Save standardized values as variables复选框即可。

4. 偏度和峰度

上文直接引出了正态分布,并指出许多生活中的数据均服从该分布。但是,如果数据实际上不服从该分布,则随后基于正态分布的一切估计和检验都要被推翻。如何来确认这一点呢?对于一个具体的连续变量是否近似于某种类型的分布,通常是通过 P-P概率图及非参数检验法的帮助来鉴别判定的。此处介绍两个有关正态分布的专用统计指标:偏度和峰度。

(1) 偏度 (Skewness):偏度是用来描述变量取值分布形态的统计量,指分布不对称的方向和程度。样本的偏度系数记为 α:

$$\alpha=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^3/s^3$$

式中 s为样本标准差。这是根据矩法 (详见 4.5.2节)测定分布偏度的计算公式。测定分布偏度的其他方法还有分位数法和 Pearson规则等,这里不做介绍,读者可以参考有关专业书籍。偏度是与正态分布相比较而言的统计量。α>0分布为正偏或右偏,即长尾巴在右边,峰尖偏左;α<0分布为负偏或左偏,即长尾巴在左边,峰尖偏右;α=0分布为对称。

需要特别提醒的是,偏态的方向指的应当是长尾的方向,而不是高峰的位置。和左、右偏态的称呼相对应的术语还有正、负偏态,这里的正负是指资料的算术均数与众数之差的符号,对于右偏态分布的资料,此时算术均数大于众数,称之为正偏态;同理称左偏态为负偏态。国内的不少统计书籍对左、右偏态的理解有误,往往正好弄颠倒。

(2) 峰度 (Kurtosis):峰度是用来描述变量取值分布形态陡缓程度的统计量,是指分布图形的尖峭程度或峰凸程度。样本的峰度系数记为 β:

$$\beta=\frac{1}{n}\sum_{i=1}^{n}(x_i-\bar{x})^4/s^4-3$$

同样,式中 S为样本标准差。这也是根据矩法测定分布峰度的计算公式。测定分布峰度的方法还有分位数法 (略)。峰度也是与正态分布相比较而言的统计量。β>0分布为高峰度的,即比正态分布峰要陡峭,峰的形状比较尖;β<0分布为低峰度的,即形状比正态分布的峰要平坦;β=0则分布为正态峰。

Explore过程的结果输出中默认就会给出峰度系数与偏度系数,这在前面的分析实例中已经

见到过了。

4.5.2 参数的点估计

在确定了总体的分布类型后,只需要确定总体分布的几个关键参数,就可以精确的对其中心位置、集中趋势等进行描述。但是总体参数一般都是未知的,需要进行参数估计,也就是要用样本统计量来估计总体参数(及其估计误差)。显然,均数、中位数、标准误等总体参数都可以进行参数估计,但平时遇到的主要是用均数进行参数估计。参数估计分为点估计和区间估计,这里先来讨论前者。

参数的点估计就是选定一个适当的样本统计量作为参数的估计量,并计算出估计值。如选样本均数作为总体均数的估计量,将其大小作为总体均数的点估计值。对于所选统计量是否适于作参数估计量,有无偏性、一致性和有效性三个评选标准。无偏性是指虽然估计量的值不全等于参数,但应当在真实值附近摆动;一致性是指样本量越大,估计值离真实值的差异应当越小;有效性则是指如果有两个统计量都符合上述要求,则应当选取误差更小的一个作为估计值。如前述的均数和中位数,两者在反映正态分布的集中趋势时,在无偏性和一致性方面效果都较好,但中位数的误差更大,所以前面会有应当尽量使用样本均数来反映正态分布集中趋势的结论。

参数点估计时可用的方法有矩法和极大似然估计法两种,这里分别介绍一下。

1. 矩法

矩法的名称比较专业,实际上含义非常简单,它指的是在许多情况下,样本统计量本身往往就是相应的总体参数的最佳估计值,此时就可以直接取相应的样本统计量作为总体参数的点估计值。例如,样本均数、方差、标准差都是相应总体均数、方差、标准差的矩估计量。对于常用的正态分布而言,矩法几乎可以满足全部参数的点估计需求,所以平常教科书上所说的点估计实际上就是用的矩法。

2. 极大似然估计法

极大似然估计法是另一种更好的参数估计方法,其优点在于估计量常能满足一致性、有效性等要求,且具有不变性,不变性是指当原始数据进行某种函数变换后,相应估计量的同一函数变换值仍是新样本的极大似然估计量。

该方法的原理是在已知总体的分布,但未知其参数值时,在待估参数的可能取值范围内进行搜索,使似然函数值(在参数所确定的总体中获得现有样本的概率)最大的那个数值即为极大似然估计值。

因极大似然估计法已超过本书读者需要了解的范畴,这里将不再深入讨论,读者只需要知道还有这样一个点估计的方法即可。

3. 稳健估计值

矩法和极大似然法虽然能够很好的满足点估计的需要,但它们也有很明显的缺陷,就是估计值受异常值的影响十分显著,或因数据分布的偏离而使估计值产生较大变化。在 20世纪 50年

代前后,基于正态分布理论的统计方法的不稳定性引起了统计学家的广泛关注。尤伯(P. J. Huber)于 1964年创立的渐进极小极大理论,以及汉甫(F. R. Hampel)于 20世纪 60年代末提出的崩溃点等概念和有界影响方法最终奠定了稳健统计的理论基础。

稳健统计研究的是具有稳定性的统计方法。即当观测数据符合假定模型,甚至与假定模型有偏离时,性质都较好或至少性质不会很坏的统计方法。而稳健估计指的就是该统计量具有稳健性,当数据存在异常值时受影响较小,而且对大部分的分布而言都很好(当然,这同时意味着它不会对每个分布都是最佳的)。

稳健估计有 M估计、R估计等不同方法,前者是稳健估计常用的方法。M估计最早是由尤伯提出,其实是“极大似然型估计”的简称,即该方法的核心仍然是极大似然估计法,但是在估计时它首先构造一个 Ψ函数,该函数能够减小异常值的影响,而且对所考虑的分布集合中的每个分布都是好的估计量。随后再对 Ψ函数的集中趋势进行参数的极大似然估计,因此相应的估计值受异常值的影响要小得多。

SPSS的 Explore过程能够直接输出 M估计的结果,在 Statistic子对话框中选择 M-Estimator复选框,相应的输出如表 4.7所示。

表 4.7 M-Estimators

	性别	Huber's M-Estimator[a]	Tukey's Biweight[b]	Hampel's M-Estimator[c]	Andrews' Wave[d]
身高	男	174.66	174.74	174.70	174.75
	女	162.80	162.81	162.82	162.81

a. The weighting constant is 1.339.

b. The weighting constant is 4.685.

c. The weighting constants are 1.700, 3.400, and 8.500

d. The weighting constant is 1.340*pi.

表 4.7即为输出的 M估计量的结果,SPSS中输出的 M估计量有 4种,它们分别是 Huber、Andrews、Hampel和 Tukey所提出的,实际上就是所用的 Ψ函数不同。一般而言,Huber法适用于数据接近正态分布的情况,另外三种则适用于数据中有许多异常值的情况。如果 M估计量离平均数和中位数较远,则数据中可能存在异常值。此时,应该用 M估计量替代平均数以反映集中趋势。从输出结果可见,男、女性的 4个 M估计量离均数都很近,这就可以反证数据中应当不存在明显的异常值。

4.5.3 参数的区间估计

显然,仅仅有参数的点估计是不够的,比如打靶,打了 2枪,平均 9环;打了 100枪,平均也是 9环,显然人们更相信后者的确是个好的枪手,而对前者的水平却产生很大的怀疑。这就涉及到了参数的估计值究竟有多大的误差的问题。

1. 标准误

标准误就是用来描述参数估计值可能离真实值究竟有多远的统计量。先考虑这样一种情形:假设现在已知一个正态分布的总体 $N(\mu, \sigma^2)$,从中进行抽样研究,每次抽样的样本量固定为

n,这样对每一个样本均可以计算出其均数 $\bar{X}$。由于这种抽样可以进行无限多次,这些样本均数就会构成一个新的分布总体。统计学家发现,该分布正好就是正态分布 $N(\mu, \sigma^2/n)$。也就是说,样本均数所在分布的中心位置和原数据分布中心位置相同,而其标准差(记为 $\sigma_{\bar{X}}$)则为 $\sigma_{\bar{X}} = \sigma/\sqrt{n}$。为了区分样本所在总体的标准差,通常称样本均数的标准差为样本均数的标准误(简称均数标准误,有的书上也称之为标准误差);而且,即使是从偏态总体随机抽样,当 n足够大时(如 $n>50$),$\bar{X}$也近似正态分布。这一规律就是数理统计中的中心极限定理(Central Limit Theorem)。

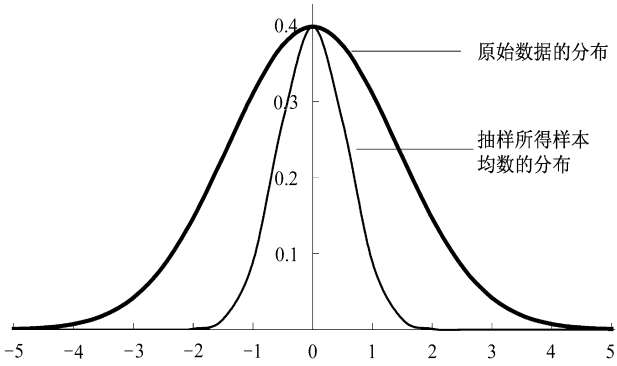


图 4.6 均数的抽样分布示意图

图 4.6就是从均数为 0的一个正态分布总体中进行抽样的示意图,可见样本均数的分布仍然是以 0为均数,但是标准差要比原分布小一些。实际上就是一个倍数关系。

标准误就是一般用来表示参数估计值准确程度的统计量,标准误越大,则说明相应参数的点估计值越不可信。

2. 区间估计的计算

结合样本统计量和标准误可以确定一个具有较大的可信度(如 95%或 99%)包含总体参数的区间,该区间称为总体参数的 $1-\alpha$可信区间或置信区间(Confidence Interval)。

下面来看一下可信区间是如何求取的,显然,由于样本均数 $\bar{X}$的分布规律为正态分布 $N(\mu, \sigma^2/n)$,现在只需要进行如下的标准化变换:

$$U = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

得到的 U将服从标准正态分布 $N(0,1)$。也就是说,若资料服从正态分布 $N(\mu, \sigma^2)$,样本含量为 n的样本均数 $\bar{X}$出现在 $(\mu \pm 1.96\sigma/\sqrt{n})$之中的概率为 0.95,即按照 95%的可信度,应当有:

$$-1.96 < \frac{\bar{X} - \mu}{S/\sqrt{n}} < 1.96$$

对上式进行变换后即得:

$$\bar{X} - 1.96S/\sqrt{n} < \mu < \bar{X} + 1.96S/\sqrt{n}$$

这就是按照 95%可信度计算出的总体均数可信区间。照此类推,对于任意可信度的情况,总体均值 μ的 100(1-α)%可信区间为:

$$\bar{x}-u_{\alpha/2}S/\sqrt{n}<\mu<\bar{x}+u_{\alpha/2}S/\sqrt{n}$$

α值一般取 0.05或 0.01,故 1-α为 0.95或 0.99。上面计算的是双侧可信区间,特殊情况下还会使用单侧的可信区间,这里不再详述。

非常有意思的是,可信度的概念往往会引起误解,它仅仅是大量重复抽样时的一个渐近概念。认为“95%的可信区间包括真实参数值的概率为 0.95”是个错误的理解。这里得到的区间是固定的,而总体参数值也是固定的。因此只有两种可能:包含或者不包含,这当中没有任何概率可言。95%的可信度只是说如果能够大量重复试验的话,则平均下来所计算的每 100个可信区间中,会有大约 95个覆盖真实值。

SPSS的 Explore过程会直接输出标准误和可信区间的大小,例如在上面的例子中男生的身高标准误为 0.671 cm,相应的总体均数 95%可信区间为 173.37~176.05 cm。但是,如果大家直接按照上面的公式利用标准误来计算区间的话,会发现和统计软件的结果略有差异,为什么会这样呢?需要特别指出的是,以上计算公式实际上仅仅适用于大样本,或者已知总体标准差的情形,如果样本量小,且只知道样本标准差,则样本均数所在总体服从的是 t分布,相应的可信区间计算也应当使用 t分布来进行,关于 t分布的知识将在第 11章中继续学习。

思考与练习

1. 请就 student.sav数据,分析学生的体质量分布情况,尝试分性别和合并描述。

2. 使用 Descriptive过程,对 student.sav中的身高和年龄变量进行标准正态变换,对变换后的变量进行统计描述。

参考文献

1 吴喜之主编.统计学基本概念和方法.北京:高等教育出版社,2003
2 杨树勤主编.中国医学百科全书·医学统计学分册.上海:上海科学技术出版社,1982
3 杨树勤主编.卫生统计学.第三版.北京:人民卫生出版社,1995
4 方积乾主编.卫生统计学.第五版.北京:人民卫生出版社,2003
5 张文彤主编.SPSS 11统计分析教程(基础篇).北京:北京希望电子出版社,2002

第5章 类变量的统计描述与参数估计

在第4章中,已经学习了连续变量的统计描述,本章将继续学习分类变量的统计描述及参数估计方法。

首先复习一下分类变量的概念。统计学上把取值范围是有限个值或者是一个数列构成的变量称为离散变量,其中表示分类情况的离散变量又称为分类变量。根据类别的有序性,分类变量又可分为有序分类变量(Ordinal Variable)和无序分类变量(Nominal Variable)两类。但是,这两类变量在统计描述上几乎没有什么差异,因此本章将它们放在一起讲解。

5.1 分类变量的统计描述概述

5.1.1 分类变量的统计描述指标体系

相对于连续变量而言,分类变量的统计描述体系非常简单。由于分类变量不能进行四则运算,因此对变量中包括的几个类型(调查题目中的选项)进行各自频数的统计以及它们在所有类型中所占的比例,就变得非常重要了。

1. 频数分布情况的描述

对于分类变量,首先希望了解各种类别的样本数有多少,除此之外,还会对相对数量比较感兴趣,如每个类别的人数占总人数的比例各为多少。这些信息往往会被整理在同一张频数表中,各个类别的样本数和所占比例分别被称为频数(绝对频数)和百分比(构成比),前者是指本类别出现的次数,百分比则是指本类别出现的次数占总次数的百分比,即本类别出现次数/总次数×100%。如在一项"最受欢迎的软饮料是什么"的调查中,调查者提供了5个答案可供选择:Coke Classic, Diet Coke, Dr. Pepper, Pepsi-Coke, Sprite。50名被调查者都会给出一个答案,统计5种软饮料的每一种在数据集中出现的次数,Coke Classic出现19次,即19人最喜欢Coke Classic;Diet Coke出现8次,即8人最喜欢Diet Coke;5人最喜欢Dr. Pepper;13人最喜欢Pepsi-Coke;5人最喜欢Sprite。这些数字即为每一种饮料的频数。但是,如果不知道总人数为50,或者希望和其他更大/更小人群的调查结果相比较时,就无法确认19这个数字到底有多大,因而又提出了百分比这个概念。如Coke Classic出现的比例为38%(19/50),即38%的人最喜欢Coke Classic;Diet Coke出现的比例为16%(8/50),即16%的人最喜欢Diet Coke;依此类推;10%的人最喜欢Dr. Pepper;26%的人最喜欢Pepsi-Coke;10%的人最喜欢Sprite。这些百分比数字,即为每一种饮料的相对频数(或称百分比)。从38%等这些百分数字,研究者就可以了解到各种饮料为人们所偏

好的程度。

在对有序分类变量进行描述时,除给出分各个类别的频数和百分比外,研究者往往还对累积频数和累积频率感兴趣。累积频数是指本类别及较低类别出现的次数之和,累计百分比则是指本类别及较低类别出现的次数之和占总次数的百分比,即(本类别出现次数+较低类别出现次数)/总次数×100%。比如,在一项员工学历的调查中,希望了解每个员工的文化程度,分别为1——高中及以下,2——大专,3——大学,4——研究生及以上。此时,调查人员不仅希望了解"高中及以下"、"大专"、"大学"、"研究生及以上"各类别员工的人数及比例,还希望了解"大专及以下"、"大学及以下"的人数及所占比例,此时显然就需要使用累积指标了。

当然,出于一些特殊的分析目的,累计频数和累积百分比也可能被用于无序分类变量,如希望知道各少数民族占总人数的比例情况等。但需要注意的是,统计软件一般都只按类别编码从小到大进行频数和百分比的累计,如果编码不符合要求,则研究者只能手工加以统计。

2. 集中趋势的描述

除原始频数外,研究者如果希望哪一个类别的频数最多,还可以使用众数(Mode)来描述它的集中趋势。所谓众数,是指出现次数最多的那个数。显然,众数有时可以多于一个。如果只有一个众数称为单众数,多于一个的称为复众数。在实际工作中,有时利用众数来说明社会经济现象的一般水平。例如,为了说明职工的技术等级、商品销售中卖得最多的服装、鞋的号码等,都可以利用众数来反映其一般水平。但是,众数只反映频数最多的类别的情况,而浪费了所有其他信息,如另一个类别的频数仅少一例,使用众数描述的话就会被完全忽视掉,因此,只有集中趋势显著时,才能用众数作为总体的代表值。实际上,当分类变量的类别数不多时,原始频数表的观察并不复杂,此时众数的使用价值并不高。

可能这里有的朋友会觉得奇怪,为什么本章只提到对分类数据描述其集中趋势,而忽略掉了离散趋势呢?这是因为对于分类数据而言,其数据的离散程度实际上是和集中趋势有关联的,它们往往受相同参数的控制,因此不需要分别描述,对此请参见本章最后一节。

3. 使用相对数进行深入描述

除以上比较简单的频数、比例外,研究者还经常为分类数据计算一些原始频数的相对指标用于统计描述,这些指标被称为相对数,这里简单介绍一下常用的三种相对数:

(1) 比(Ratio):比指的是两个有关指标之比 A/B,用于反映这两个指标在数量/频数上的大小关系。其中 A、B可以是性质相同的两个指标,如两个地区相同时期内交通事故数之比;也可以是性质不相同的两个指标之比,如某地区一周内交通事故数与交通车辆数之比。事实上,比还可以被拓展到连续变量的范畴内,如销售人员属于本月销售额之比等。

(2) 构成比(Proportion):分观察对象为 k个部分($A_1, A_2, \cdots, A_k$),其中某一个/多个部分的例数占总例数的比例称为构成比,它描述某个事物内部各构成部分所占的比重,构成比的计算公式为:

$$\text{构成比} = \frac{\text{某一组成部分的样本数}}{\text{总样本数}}$$

可见构成比的分子必须是分母的一部分,所以其取值为 0~1。实际上,前面提到的百分比

就是一个标准的构成比,而累积频率则是构成比概念的直接延伸。

(3) 率 (Rate):率是一个具有时间概念,或者说具有速度、强度含义的指标,用于说明某个时期内某个事件发生的频率或强度,其计算公式为:

$$某事件的发生率=\frac{观察期内发生某事件的对象数}{该时期开始时的观察对象数}$$

准确的讲,率应当是一个时间点上的强度测量,但这在实际工作中很难做到,因此一般都按一个时段来进行测量。从而它的分子往往是一个时期的累计数。

以上相对数在使用时应当注意适用条件,如样本量较大时相对数才会比较稳定,基数不同的相对数不能直接相加求和等。

5.1.2 分类变量的联合描述

频数表可以描述一个分类变量的数值分布情况,但是研究者往往希望对两个甚至多个分类变量的频数分布进行联合观察,如希望考察一下不同的血型在各民族间的频数分布,甚至于构成比状况如何。此时就需要将这些分类变量的类别交叉起来,分别统计各种类别组合下的频数大小。当一共有两个分类变量时,这种因分类变量的各类别交叉而成的复合频数表被称为行×列表,也称列联表。更多分类变量的交叉表格和两个变量时的交叉表格其实没有本质区别,只是更为复杂而已。在多个分类变量的联合分析中,列联表提供了清楚明白的分析结果,非常直观,容易进行比较。在一般的调查报告中,经常看到作者应用列联表进行变量的交叉分析,它也是调查报告中显示分析结果的主要方式之一。

以二维的 $r\times c$列联表为例。假设有 n个个体根据两个属性 A和 B进行分类。属性 A有 r类:$A_1, A_2, \cdots, A_r$,属性 B有 c类:$B_1, B_2, \cdots, B_c$。n个个体中既属于 A_i类又属于 B_j类的有 n_{ij}个。那么可用如表 5.1所示的一个二维的 $r\times c$列联表表示。

表 5.1 二维的 $r\times c$列联表

	B_1	B_2	…	B_c	合 计
A_1	n_{11}	n_{12}	…	n_{1c}	$n_{1\cdot}$
A_2	n_{21}	n_{22}	…	n_{2c}	$n_{2\cdot}$
…	…	…		…	…
A_r	n_{r1}	n_{r2}	…	n_{rc}	$n_{r\cdot}$
合 计	$n_{\cdot 1}$	$n_{\cdot 2}$	…	$n_{\cdot c}$	n

表 5.1中,除合计栏外的每一个单元格反映了 A B两属性在某种类别交叉下的频数情况,而合计栏则分别反映了 A B两属性各自的类别频数情况,且表格中的数据有如下的换算关系:

$$n_{i\cdot}=\sum_j n_{ij},\ n_{\cdot j}=\sum_i n_{ij},\ n=\sum_i n_{i\cdot}=\sum_j n_{\cdot j}$$

除给出原始频数外,各单元格内还可能给出行百分比、列百分比和总百分比等,分别用于反映该单元格频数占所在行、列、总样本的构成比情况。

5.1.3 SPSS中的相应功能

作为比较基本的功能,SPSS的许多分析过程均可完成分类变量统计描述的任务,但专门用于分类变量统计描述的过程有两个,它们均集中在 Descriptive Statistics子菜单中。

⑴ Frequencies过程:在上一章中已经学习过了,它主要针对单个分类变量输出频数表,从而得到“频数”、“百分比”和“累积百分比”的统计量。除原始频数表外,还可给出描述集中趋势的众数,以及直接绘制用于分类变量的条图和饼图等。

⑵ Crosstabs过程:其强项在于两个或多个分类变量的联合描述,可以产生二维至 n维列联表,并计算相应的行、列、合计百分比和行、列汇总指标等。除强大的描述功能外,该过程也具备了完善的分类资料统计推断功能,详见第 14章。

此外,针对比较特殊的多选题统计描述问题,SPSS也为其提供了专门的模块支持,详见本章第 3节。

5.2 分类变量统计描述实例

这里仍以上一章中使用过的 student.sav为例,来学习一下分类变量的统计描述在 SPSS中的具体实现方法。

5.2.1 使用 Frequencies过程输出频数表

如果研究者希望了解一下共有多少学生,男生和女生各自为多少;各种血型的人数有多少,则可以使用 Frequencies过程输出这两个变量的频数表,具体操作如下:

Analyze→Descriptive Statistics→Frequencies
Variables框:sex, blood_t
OK

相应的分析结果如下:

表 5.2 Statistics

		性别	血型
N	Valid	219	219
	Missing	0	0

首先给出的是统计量列表(见表 5.2),因这里没有选择输出任何统计量,所以只会给出有效样本量。可见一共有 219名学生的数据。这 219名学生的性别和血型数据都是完整的,没有缺失值。

表 5.3 性别

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	男	72	32.9	32.9	32.9
	女	147	67.1	67.1	100.0
	Total	219	100.0	100.0	

表 5.3给出了性别的频数表，Frequency为频数，Percent为各组频数占总例数的百分比（包括缺失记录在内），Valid Percent为各组频数占总例数的有效百分比，Cumulative Percent为各组频数占总例数的累积百分比。可见在 219人中，男性 72人，女性 147人两类人群的累积百分比正好就是 100%。由于不存在缺失值，因此这里的 Percent和 Valid Percent完全相同。

表 5.4 血型

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	A	67	30.6	30.6	30.6
	AB	30	13.7	13.7	44.3
	B	37	16.9	16.9	61.2
	O	85	38.8	38.8	100.0
	Total	219	100.0	100.0	

表 5.4给出的是血型的分析结果，请读者自行分析，这里不再详述。

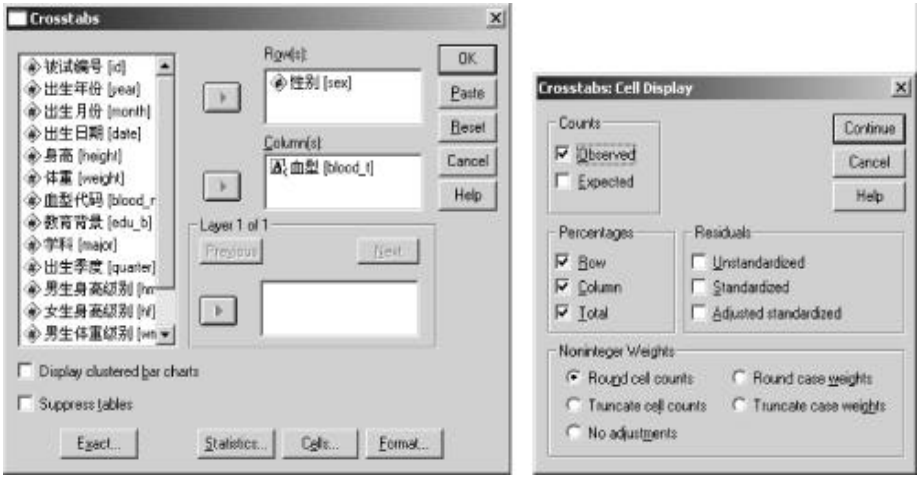
5.2.2 使用 Crosstabs过程输出列联表

如果研究者希望知道性别和血型的交叉频数分布，以及各种百分比的情况，又该如何操作呢？Crosstabs过程可以帮研究者完成这个任务，具体操作如下：

Analyze→ Descriptive Statistics→ Crosstabs
Row (s) 框：sex Column (s) 框：blood_t
Cells：
Percentages: Row, Column, Total
Continue
OK

操作中用到的对话框如图 5.1 (a) 所示，主对话框中的 Rows框、Columns框分别用于选择行×列表中的行、列变量。而下方的 Layer框组则用于选入更多的分类变量，这里被称为层变量（详见第 6章关于表格结构的介绍）。如图 5.1 (b) 所示的 Cell子对话框用于定义列联表单元格中需要显示的指标。这里要求输出三种百分比。

本例相应的输出如下：



(a) (b)

图 5.1 Crosstabs过程的对话框

表 5.5 Case Processing Summary

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
性别 * 血型	219	100.0%	0	.0%	219	100.0%

首先是处理记录缺失值情况报告（见表 5.5），可见 219例均为有效值。

表 5.6 性别 * 血型 Crosstabulation

			血型				
			A	AB	B	O	Total
性别	男	Count	16	8	17	31	72
		% within 性别	22.2%	11.1%	23.6%	43.1%	100.0%
		% within 血型	23.9%	26.7%	45.9%	36.5%	32.9%
		% of Total	7.3%	3.7%	7.8%	14.2%	32.9%
	女	Count	51	22	20	54	147
		% within 性别	34.7%	15.0%	13.6%	36.7%	100.0%
		% within 血型	76.1%	73.3%	54.1%	63.5%	67.1%
		% of Total	23.3%	10.0%	9.1%	24.7%	67.1%
Total		Count	67	30	37	85	219
		% within 性别	30.6%	13.7%	16.9%	38.8%	100.0%
		% within 血型	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	30.6%	13.7%	16.9%	38.8%	100.0%

表 5.6就是性别和血型的交叉表,行变量是性别,列变量是血型,由于系统默认为升序排列(Ascending)。所以 4列血型依次的排列是 A、AB、B和 O。可以看出,在总共 72名男性被调查者中,A型血有 16名,AB型血有 8名,B型血有 17名,O型血有 31名。同样,在总共 147名女性被调查者中,A型血有 51名,AB型血有 22名,B型血有 20名,O型血有 54名。

然而,由于在被调查中男女的数量不同,调查者很难从表 5.6中看出诸如某一个血型男女的比例是不是一样,或有什么差异。不过,每个单元格内已经输出了行百分比、列百分比和合计百分比。这里以 A型血和男性交叉的单元格为例加以说明,该单元格内自上而下依次为:第一个数 16为该单元格的实际频数。第二个数 22.2%为行百分比,它与它右边的男性和 AB、B、O交叉的单元格中的相应百分比 11.1%、23.6%和 43.1%相加正好为 100%;第三个数 23.9%为列百分比,它与它下边的 A型血和女性交叉的单元格中的相应百分比 76.1%相加正好为 100%;第四个数 7.3%为合计百分比,它是该单元格频数 16在所有交叉单元格中所占的总百分比。与其余单元格相应的百分比相加也正好为 100%。

这样就可以进行一些有意义的比较了。比如,在男性被调查者中,A型血的男性占 22.2%。在女性被调查者中,A型血的女性占 34.7%。因此调查者会考虑:是否男性 A型血的人较女性 A型血的人少(假设调查是随机抽样,总体男女数量相同)。同样的道理,男性 AB型血的人较女性 AB型血的人少。男性 B型血的人较女性 B型血的人多。男性 O型血的人较女性 O型血的人多。不过,这样的结果也可能是由于抽样的偶然误差导致的,必须要经过假设检验,才能对以上的猜测加以确定。

5.3 多选题的统计描述

多选题是调查问卷中极为常见的调查题目类型,在第 2章中已对其录入方式进行了讲解,由于它所收集的数据也属于分类数据,因此本章将继续讲解对于这类多选题如何进行描述分析。

5.3.1 多选题的描述指标体系

如何对多选题进行分析呢?当然,可以对每一个单独的题项来进行统计描述,但这样做是不全面的,因为这些变量实际上回答的是一个大问题,将问题割裂开来可能会导致不正确的分析结果,而且无法计算一些汇总指标。在多选题分析中比较特别的描述指标有以下 4个:

⑴ 应答人数:是指选择了本选项的人数,或者说就是原始频数,比如说在 200人中有 178人选择了调理饮食以控制高血压。

⑵ 应答人数百分比(Percent of Cases):选择该项的人占总人数的比例,比如 200个受访者中共有 178人选择了调理饮食以控制高血压,则调理饮食的应答人数百分比为 178/200=89.00%。应答人数百分比可以反映该选项在人群中的受欢迎程度。

⑶ 应答人次:是指选择本选项的人次,一般情况下,应答人次和应答人数是相同的,

但是在有的时候是不同的。例如,您最近买的几管牙膏的品牌各是什么?这种问题,就可能同一个人回答同一个答案多次因为同一个品牌他买了两管。因此,此类多选题就会有可能出现选择某答案的人数不等于选择某答案的次数的情况,因而 Count与 Response就有可能不等。

⑷ 应答次数百分比(Percent of Responses):在做出的所有选择中,选择该项的次数占总次数(总反应数)的比例,比如 200受访者对 4种高血压控制方式分别选择了 178、120、134、160次,则总的应答次数为 178+120+134+160=592人次,而调理饮食的应答次数百分比应为 178/592=30.07%。应答次数百分比可以用于不同选项受欢迎程度的比较。

使用以上几种指标,就可以对多选题进行比较完善的描述了。和录入时相同,SPSS的 Tables模块和 Multiple Response菜单都可以对多选题变量集进行统计描述,但前者生成的是标准的结果表格,可以进行各种复杂编辑,而后者生成的是纯文本表格,功能上也要简单一些。本章将以 Base模块中的 Multiple Response菜单为主加以讲述,Tables模块中的相应功能请参见第 6、7两章。

5.3.2 分析实例

这里使用的是一次市场调查的具体数据 multiplecategory.sav,文件中性别(d1)变量的代码是 1男,2女。其中的第 7题(q7)为多项选择题,具体的题目是:

> q7. 请问促使您买保健品的主要原因是(可多选):
> 1. 广告宣传　　2. 自己需要　　3. 家人需要
> 4. 看望亲友　　5. 朋友推荐　　6. 其他(请注明):

对于多选题的录入和在 SPSS中多选题的定义,在第 2章中已经讲过了,本题是采用多重分类法进行录入,考虑到最多可能答案为 6个,所以共有 6个变量(q7_1～q7_6)。此时应当将这 6个变量定义为一个多选题,该多选题的名称为 q7,标签为“促使购买保健品的主要原因”。

1. 多选题的频数列表

如果希望给出各答案的频数分布情况,则操作步骤如下:

> Analyze→Multiple Response→Frequencies
> Table(s) for框:促使购买保健品的主要原因[$q7]
> OK

所使用的 Multiple Response Frequencies对话框内容非常简单,如图 5.2所示,这里不再详细解释。只是指出下方的 Missing Values复选框组用于选择对缺失值的处理方式,两个复选框分别对应了两种编码的对应方式,不能交错使用。

相应的结果输出如下:

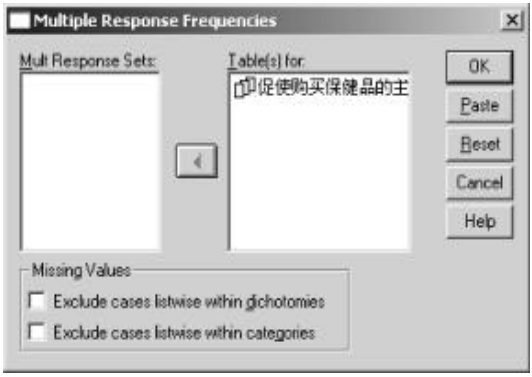


图 5.2 Multiple Response: Frequencies过程的对话框

Group $q7 促使购买保健品的主要原因

Category label	Code	Count	Pct of Responses	Pct of Cases
广告宣传	1	14	2.2	3.1
自己需要	2	299	47.6	66.7
家人需要	3	197	31.4	44.0
看望亲友	4	93	14.8	20.8
朋友推荐	5	17	2.7	3.8
其他	6	8	1.3	1.8
Total responses		628	100.0	140.2

0 missing cases; 448 valid cases

上面的结果提供的信息是：在448个有效的被调查者中，各种原因一共被选择了628次，其中“广告宣传”被选择了14次，“自己需要”被选择了299次，“家人需要”被选择了197次，“看望朋友”被选择了93次，“朋友推荐”被选择了17次，“其他”原因被选择了8次。

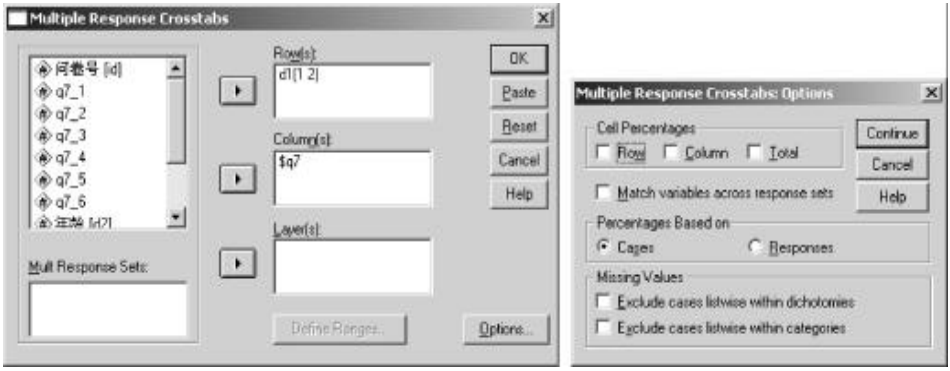
右边的两个百分数是多项选择题比较重要的输出：Pct of Responses计算的是选择次数占总选择次数的比例，比如，这448位被调查者一共进行了628次选择，其中有14人选择了“广告宣传”，该选择次数所占的比例为14/628=2.2%；Pct of Cases计算的则是所有被调查者中选择相应分析方法者占总人数的比例，例如，有14人选择了“广告宣传”，他们占总人数的14/448=3.1%。在调查报告中，研究人员经常使用的是Pct of Cases栏中的百分数。它所表明的意义人们比较容易理解，虽然各个百分数的和大于100%。

2. 多选题的列联表分析

上面直接给出了多选题的频数表，但有的时候还希望能够对不同的人群分别描述，即将多选题变量集和其他分类变量进行交叉描述。如在本例中希望分性别进行考察，则操作如下：

Analyze→Multiple Response→Crosstabs
Row (s)框 :d1
选中 d1:Define Ranges:
Minimum框 :1 Maximum框 :2
Continue
Column (s) 框 :促使购买保健品的主要原因 [$q7]
OK

用到的操作界面如图 5.3 (a)所示,可见多选题的 Crosstabs主对话框和普通 Crosstabs过程的主对话框非常相似,只是下方多了 Define Ranges钮,用于为相应的变量设置取值范围。其 Options子对话框 (见图 5.3 (b))中也可以定义输出行百分比、列百分比和总百分比指标,以及控制缺失值的处理方式。



(a) (b)

图 5.3 Multiple Response: Crosstabs过程的对话框

本例的分析结果如下页框图所示。

交叉表中分性别给出了对各种购买原因的选择情况。在男性被调查者中,购买保健品的原因选择如下:广告宣传 3人,自己需要 133人,家人需要 87人,看望亲友 49人,朋友推荐 6人和其他 3人。同样,在女性被调查者中,购买保健品的原因选择如下:广告宣传 11人,自己需要 166人,家人需要 110人,看望亲友 44人,朋友推荐 11人和其他 5人。在 448位被调查者中,男性 206人,占 46%,女性 242人,占 54%。

由于设置的关系,在系统输出的交叉表的单元格里,只显示了频数的多少,这样看起来比较清楚,美观。但是由于在被调查者中男性与女性的数量不同,仅仅从这个交叉表中的频数中,很难看出性别之间的差异,在一些指标上缺乏可比性。如果在本分析过程 Options的 Cell Percentages复选框组中选择显示变量的行百分比、列百分比和总百分比,就可以更详细的进行性别间的比较了,对此请读者朋友们自行操作,这里不再详述。

* * * CR0SSTABULATION * * *

d1 性别
by $q7(group) 促使购买保健品的主要原因

Count	$q7 广告宣传 1	自己需要 2	家人需要 3	看望亲友 4	朋友推荐 5	其他 6	Row Total
d1							
1 男	3	133	87	49	6	3	206 46.0
2 女	11	166	110	44	11	5	242 54.0
Column Total	14 3.1	299 66.7	197 44.0	93 20.8	17 3.8	8 1.8	448 100.0

Percents and totals based on respondents

448 valid cases;0 missing cases

5.4 分类变量的参数估计

对于分类变量而言,由于只能取若干个离散的值,因此这里在讨论参数估计时往往关心的就是各类别在总体中的比例是多少,或者当从中进行一次抽样时,抽得相应类别的概率是多少,且对于所有可能的类别,其概率之和应当等于 1。

5.4.1 二项分布的参数估计

二项分布是分类变量最为常见的分布类型,下面将讨论一下二项分布的参数估计问题。

1. Bernoulli试验与 Bernoulli试验序列

在许多问题中,人们仅对试验中某事件是否发生感兴趣。例如,掷硬币试验中,关心的是出现正面还是出现反面;产品抽样检查中,注意抽取的产品是好产品还是废品;射击试验中,命中还是不命中;比赛中,胜还是负……在这类问题中,试验的可能结果只有两个,或者事件 A 发生,或者事件 A不发生即 $\overline{A}$发生,这种只有两个可能结果的试验称为贝努利(Bernoulli)试验。

现在开始重复进行 n次独立的贝努利试验。"重复"的意思是指各次试验的条件是相同的,它意味着各次试验中事件 A发生的概率保持不变,设都是 p(从而 $\overline{A}$的概率也保持不变,设都是 q,$q=1-p$);"独立"的意思是指各次试验的结果是相互独立的。这种试验所对应的数学模型称

为贝努利概型,有时为了突出试验次数 n,也称为 n次贝努利概型或 n重贝努利试验。

进行 n次独立重复的贝努利试验,每次试验事件 A发生的概率为 p,若以 ξ表示 n次独立重复的贝努利试验中事件 A发生的次数,那么容易求得 ξ的分布列是

$$P_n(\xi=k)=C_n^k p^k q^{n-k} \quad k=0,1,2,\cdots,n$$

其中:$P(A)=p, P(\bar{A})=q=1-p$

满足以下三个条件的 n次试验构成的序列被称为是 Bernoulli试验序列。

(1) 每次试验结果,只能是两个互斥的结果之一 (A或非 A)。

(2) 每次试验的条件不变。即每次试验中,结果 A发生的概率不变,均为 π。

(3) 各次试验独立。即一次试验出现什么样的结果与前面已出现的结果无关。

2. 二项分布的函数式

一般地,在 Bernoulli试验序列的 n次试验中,事件 A出现的次数 X具有概率

$$P(X=k)=\binom{n}{k}\pi^k(1-\pi)^{n-k} \quad k=0,1,\cdots,n$$

由于 $\binom{n}{k}\pi^k(1-\pi)^{n-k}$是二项式 $[\pi+(1-\pi)]^n$ 展开式中的各项,故称此分布为二项分布。显然,对于不同的 n,不同的 π有不同的二项分布。因此,n、π是二项分布的两个参数。

推而广之,若有一个随机变量 X,它的可能取值是 0,1,…,n且相应的取值概率是

$$P(X=k)=\binom{n}{k}\pi^k(1-\pi)^{n-k}$$

则称此随机变量 X服从以 n、π为参数的二项分布,记为 $X\sim B(n,\pi)$。对于该变量而言,有均数 $\mu_x=n\pi$,方差 $\sigma_x^2=n\pi(1-\pi)$,标准差 $\sigma_x=\sqrt{n\pi(1-\pi)}$。显然,对于样本量 n确定的情形,均数和标准差间存在着明确的换算关系,它们都只受 π的影响,这也是为什么前文不对离散趋势加以描述的理论依据。

3. 二项分布与正态分布的关系

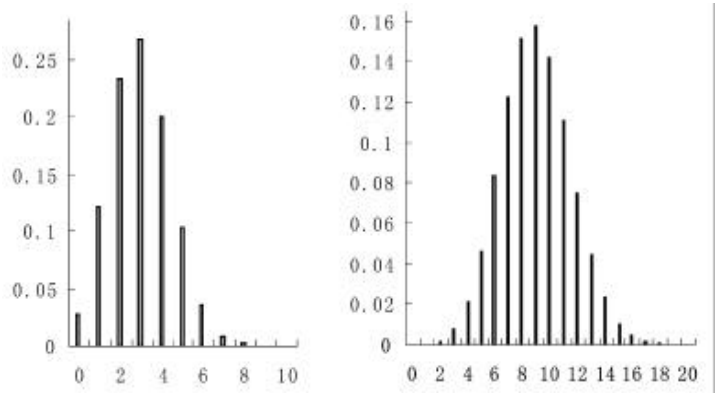
若已知 n与 π,则按上述二项式可计算不同 X取值时的概率,然后以 X为横轴,概率 P为纵轴,可绘制二项分布的图形 (参见图 5.4)。显然,二项分布图的形状取决于 n,π的取值。当 π=0.5时,图形对称;当 π≠0.5时,图形呈偏态,但随 n的增大,图形逐渐对称。

由数理统计学的中心极限定理可得,当 n较大、π不接近 0也不接近 1时 (一般认为这个界限是 $n>40$,且 np和 nq均大于 5),二项分布 $B(n,\pi)$已经非常近似于正态分布 $N(n\pi, n\pi(1-\pi))$。正态分布是许多统计方法的应用基础,二项分布的正态近似拓宽了二项分布的应用范围。

4. 二项分布的参数估计

在实际问题中,对于一个二项分布的总体而言,其试验次数 n是可以人为确定和控制的,因此只需要对参数 π加以估计,就可以明确整个分布的情况。前面已经知道,当 n较大、π也不太极端时,二项分布 $B(n,\pi)$近似正态分布,这样就可以系统的利用正态分布中的相应成果来进行参数估计了。

一般地,从一个阳性率为 π的总体中,随机抽取含量为 n的样本,则样本中的阳性数 X服从



n=10, π=0.3　　n=30, π=0.3

图 5.4 不同参数的二项分布示意图

二项分布 B(n,π),且样本阳性率 p的概率

$$P(X) = \binom{n}{x} \pi^{x} (1-\pi)^{n-x}$$

其中,样本率 p的总体均数 $\mu_p = \pi$,总体标准差(也就是标准误) $\sigma_p = \sigma_X / n$。相应的样本率就是总体均数的点估计值,如果样本足够大,则可以利用正态近似计算可信区间,相应的 100(1−α)%可信区间为:$P \pm 1.96\sqrt{P(1-P)/n}$。

当不满足正态近似的条件时,则可以直接利用二项分布的概率分布规律计算相应的可信区间,此处略。

5.4.2 其他分布类型简介

除二项分布外,在分类资料的描述中偶尔还会遇到一些其他的分布类型,这里简单介绍一下。

1. 多项分布

二项分布用于描述只有两种可能结局事件的概率分布规律,对于有多种可能结果的事件,则需要使用多项分布(Multinomial Distribution)加以描述。比如在掷筛子的时候,每个面都会以一定的概率向上,假定这些概率为 $p_1 \sim p_6$。显然这些概率的和为 1,而人们关心的就是在 n次试验中各种结局分别出现 $k_1 \sim k_6$ 次的概率,且有 $k_1 + k_2 + k_3 + k_4 + k_5 + k_6 = n$。

如果用 $p(m_1, \cdots, m_k)$ 代表多项分布 k种结束在 n次试验中分别出现 $m_1, m_2, \cdots, m_k$ 次的概率,而 $p_1, p_2, \cdots, p_k$ 为一次试验时各种可能结束出现的概率。则应当有:

$$p(m_1, m_2, \cdots, m_k) = \binom{n}{m_1, m_2, \cdots, m_k} p_1^{m_1} p_2^{m_2} \cdots p_k^{m_k}, \sum_{i=1}^{k} m_i = n, \sum_{i=1}^{k} p_i = 1$$

① 本部分内容属拓展知识面,难度较高,跳过此节,不会影响以后各章的阅读。

这里 $\binom{n}{m_1,m_2,\cdots,m_k}=\frac{n!}{m_1!m_2!\cdots m_k!}$为多项式系数，整个多项分布可以用符号 $M(n;p_1,p_2,\cdots,p_k)$来表示。

2. 超几何分布

在质量检查中，往往一次抽取若干物品，每检查一个之后并不放回，这时一个产品不会被重复检查。而如果发现次品数超过标准，就会将整批产品评价为不合格，那么这种结局的概率是多少呢？

如果是“放回式抽样”，也就是每检查一个就把它放回。这样再抽取时，检查过的物品还有可能被抽上。这时每次抽样时得到次品的概率是服从二项分布的，概率等于次品的比例。但是在上述问题中，采用的是“不放回抽样”，此时概率就满足超几何分布（Hypergeometric Distribution）。

显然，超几何分布和排列组合密切相关，仍以质量检查为例，在一批 n个产品中，如果有 m个不合格产品（即有 n－m个合格产品），那么在不放回抽取 t个产品中有 x个不合格产品的概率为：

$$p(x)=\binom{m}{x}\binom{n-m}{t-x}\Big/\binom{n}{t}\qquad x=0,1,\cdots,t$$

3. Poisson分布

Poisson分布也是一种离散随机变量的分布，主要用于描述在单位时间（空间）中某种事件的发生数。如放射性物质在单位时间内的放射次数；在单位容积充分摇匀的水中的细菌数；野外单位空间中的某种昆虫数等。

满足以下三个条件的随机变量服从 Poisson分布：X的取值与观察单位的位置无关，只与观察单位的大小有关；在某个观察单位上 X的取值与前面各观察单位上 X的取值独立（无关）；在充分小的观察单位上 X的取值最多为 1。

X服从以 μ为参数的 Poisson分布可记为 $X\sim P(\mu)$。如果随机变量 X服从 Poisson分布，则 X的取值范围为非负整数，而每种情形下相应取值概率为：

$$P(X=k)=\frac{\mu^k}{k!}e^{-\mu}$$

式中 e为自然对数的底 2.7182；μ是大于 0的常数，被称为 Poisson分布的参数。Poisson分布只有一个参数 μ。这个参数既是 Poisson分布的总体均数，又是分布的总体方差，不同的 μ对应于不同的 Poisson分布。

思考与练习

1. 请就 SPSS自带数据 Employee data.sav，分析员工的性别、受教育程度、少数民族、职位类别的分布情况，并尝试分析这些属性之间的关系以及这些属性和工资之间的关系。

2. 请就 SPSS自带数据 1991 U.S. General Social Survey.sav，分析健康问题（对应的变量为

h1th1～h1th9,为多选题)的分布情况。

参考文献

1 吴喜之主编.统计学基本概念和方法.北京:高等教育出版社,2003
2 杨树勤主编.中国医学百科全书·医学统计学分册.上海:上海科学技术出版社,1982
3 杨树勤主编.卫生统计学.第三版.北京:人民卫生出版社,1995
4 张文彤主编.SPSS 11统计分析教程(基础篇).北京:北京希望电子出版社,2002