数据分析与数据挖掘实验指导书

Experiment Manual of Data Analysis and Data Mining

郝文宁　靳大尉　程恺　编著

国防工业出版社
National Defense Industry Press

数据分析与数据挖掘实验指导书

郝文宁　靳大尉　程　恺　编著

国防工业出版社

·北京·

内容简介

本书是数据分析与数据挖掘课程的实验指导书，结合大量实例全面阐述了使用IBM SPSS系列软件进行数据分析与挖掘的原理、方法和步骤。紧密配合理论教学，使学生在有限的实验课时中，加深对所学知识的理解和掌握。

全书分为两个部分，第一部分为数据分析实验，主要介绍如何利用IBM SPSS Statistics软件进行统计分析，具体包括描述性统计、参数检验、非参数检验、相关分析、回归分析和因子分析等七项实验科目，第二部分为数据挖掘实验，主要介绍如何利用IBM SPSS Modeler软件进行数据挖掘，具体包括关联规则挖掘、决策树分类、人工神经网络分类、贝叶斯方法分类和聚类等七项实验科目。

本书可作为数据工程相关专业本科生教材，也可为从事各领域数据分析和数据挖掘的专业人员提供指导和帮助。

图书在版编目(CIP)数据

数据分析与数据挖掘实验指导书/郝文宁，靳大尉，程恺编著. —北京：国防工业出版社，2016.3
ISBN 978-7-118-10797-5

Ⅰ.①数… Ⅱ.①郝… ②靳… ③程… Ⅲ.①统计分析－应用软件 ②数据采集 Ⅳ.①C812 ②TP274

中国版本图书馆CIP数据核字(2016)第036979号

※

国防工业出版社出版发行
(北京市海淀区紫竹院南路23号 邮政编码100048)
三河市腾飞印务有限公司印刷
新华书店经售

*

开本 787×1092 1/16 **印张** 11½ **字数** 255千字
2016年3月第1版第1次印刷 **印数** 1—3500册 **定价** 40.00元

(本书如有印装错误，我社负责调换)

国防书店：(010)88540777 发行邮购：(010)88540776
发行传真：(010)88540755 发行业务：(010)88540717

前　言

数据分析和数据挖掘作为定量研究中的两种主要方法和工具，主要用于揭示所研究问题的数量关系、事物特征、现象规律等，是数据与知识工程领域中最为活跃紧贴前沿的研究方向，学习掌握数据分析与挖掘的相关方法和技术，进而学会“用数据说话，用数据管理，用数据决策和用数据创新”，已经成为当今社会各个行业信息化建设、数据资源开发利用及其业务变革的重要目标。

一般来说，数据分析有广义和狭义两种含义，从狭义角度来理解，数据分析是以统计学为基础，研究探索各类事物的统计特征以及变量或样本之间的量化统计关系，对于样本数据有着比较严格的假设和约束，从 20 世纪 30 年代起，各类数据分析方法就在自然科学、管理科学和社会经济等领域广泛应用。从广义角度来看，数据分析涵盖着数据挖掘的相关理论方法和技术，属于一个典型的交叉学科领域，主要研究如何从数据中提取局部模式或全局模型，形成描述、区分事物的知识和能力，实现关联分析、分类、预测、聚类等典型应用功能，可用于不同类型数据资源的开发利用。从 20 世纪 80 年代起，数据挖掘技术取得突飞猛进的发展，涉及到数据库系统、数据仓库、统计学、机器学习、人工智能、信息科学等多个相关领域的知识。

如上所述，数据分析与数据挖掘在理论方法和技术手段上存在较大差异，因此，本实验指导书在内容编排方面，将二者的实验内容及实验手段进行了区分，主要思路如下。

数据分析实验依托 IBM SPSS Statistics 展开，该软件已有 40 多年的成长历史，是目前应用比较广泛的专业统计软件，支持多国语言交互界面，用户操作清晰友好，帮助文档齐全，能够跨平台运行，具备全面统计分析功能，分析结果简洁直观，因此实验 1 主要用来帮助实验者熟悉 IBM SPSS Statistics 的基本操作。

数据分析实验内容主要围绕描述性统计、推断性统计和多元统计分析三大统计学原理部分展开。

描述性统计是采用概括性数据指标或图表综合说明事物特征、关系和规律的一种方法，主要通过集中趋势、离散趋势、分布形状和相对（分布）位置四大类统计量来描述数据集特征，实验 2 主要用来帮助实验者理解和掌握描述性统计的主要方法。

推断性统计则是通过样本归纳总体的一种统计方法，主要研究解决两大类问题：一类是估计问题，主要解决通过样本估计总体（其分布形式已知）未知参数值及范围的问题；另外一类则是假设检验问题，主要解决通过样本对未知总体的统计特征假设进行合理性、有效性和可靠性判断问题。由于估计问题在实践中独立的应用较少，本书的推断性统计实验侧重于假设检验。在假设检验实验部分，区分为参数检验和非参数检验两大类展开，涵盖了均值差异性、分布类型和分布位置等主要检验方法。实验 3 以 t 检验和方差分析两类数据分析过程为主帮助实验者理解掌握均值差异性检验方法，实验 4 以比率检验、秩

次检验、中心位置检验等数据分析过程为主帮助实验者理解掌握数据分布类型和数据分布位置检验方法。

多元统计分析是基于经典统计学研究多个关联对象或关联指标之间统计规律的一种综合分析方法，本书主要选择相关分析和因子分析作为实验内容，聚类分析、判别分析等内容合并到随后的数据挖掘实验部分。其中，相关分析用于研究原始变量之间的依存关系，分为相关关系（非确定性关系）分析和函数关系（确定性关系）分析两大类展开，实验5以相关系数分析和偏相关分析过程为主帮助实验者理解掌握变量间相关程度及相关方向的统计分析方法，实验6则以一元线性回归和多元线性回归为主帮助实验者理解掌握变量间线性回归模型建立以及线性关系显著性检验的方法。因子分析用于研究原始变量群中共性因子的综合提取，其实质也是一种变量间相关关系研究方法，实验7通过因子分析实验帮助实验者理解掌握如何依据相关性大小对于原始变量分组表示为综合变量的方法。

数据挖掘实验依托IBM SPSS Modeler展开，该软件提供了各种基于机器学习、人工智能和统计学的建模方法。在数据概览、数据预处理、数据挖掘建模、工作流图形界面、工程管理、自动报告生产等方面，均具有非常强大的功能和便捷的使用模式。实验8主要用来帮助实验者熟悉IBM SPSS Modeler的基本操作。

数据挖掘实验内容主要依据待挖掘知识（模式或模型）的类型不同，围绕关联规则挖掘、预测性数据挖掘和描述性数据挖掘三大数据挖掘原理展开。

关联规则挖掘就是通过频繁模式挖掘的方法从数据中发现有价值的描述数据项之间关联规则，关联类型可分为简单关联、时序关联、因果关联。实验9通过基于Apriori的频繁项集挖掘和频繁序列模式挖掘实验帮助实验者理解掌握关联规则、序列模式等不同模式类型的挖掘方法。

预测性数据挖掘是基于数据提供的先验知识通过各类归纳算法构建预测模型，在对象的已知特征值和未知特征值之间建立映射关系，实现预测功能。预测模型可分为分类和回归两大类，其中，回归模型的建立及分析实验在数据分析部分完成。分类模型一般又可分为判别模型和概率模型。实验10选择决策树分类实验帮助实验者理解掌握不同类别的决策区域的确定方法，以及CART、C4.5等决策树分类算法的优劣对比。实验11通过支持向量机分类实验帮助实验者理解掌握当不同类别间决策边界不清晰时，通过判别函数度量类别之间的最大化差异，进而实现分类的方法。实验12通过人工神经网络分类实验帮助实验者理解掌握神经元之间连接权值的迭代优化过程，以及利用多个神经元输出结果的非线性变换输出实现分类的方法。实验13通过贝叶斯分类实验帮助实验者理解掌握基于概率模型实现分类的方法。

描述性数据挖掘是基于数据通过观察实现对数据的高度概括，获取数据的重要特征，主要包括聚类分析、密度估计以及多元统计分析等方法。实验14通过k均值聚类和两步法聚类实验帮助实验者理解掌握基于对象的相似性度量实现组间相似最小化、组内相似最大化的聚类方法。

本实验指导书编写过程中，尽可能做到叙述简明扼要，力求实用好用，但由于数据分析与数据挖掘相关方法和技术手段发展迅速，应用领域广泛，加之我们的水平有限，因此书中一定存在许多不足之处，希望同行和广大读者提出批评建议。

符 号 表

N	样本数量
$T(t)$	t 统计量,对于不同类型的检验,有不同的计算公式
df	自由度
Sig	显著性水平
F	方差齐性检验统计量
W	Wilcoxon 检验中的秩和值
Z	秩和检验统计量
R	多元(复)相关系数
R^2	决定系数
B	回归系数

目　录

实验 1　IBM SPSS Statistics 软件使用基础 …… 1
1.1　实验目的与要求 …… 1
1.2　实验原理 …… 1
1.3　实验内容与步骤 …… 1
1.3.1　安装、启动与退出 …… 1
1.3.2　定义变量 …… 3
1.3.3　数据的输入与保存 …… 6
1.3.4　数据文件的编辑与转换 …… 7
1.4　思考题 …… 12
实验 2　描述性统计 …… 13
2.1　实验目的与要求 …… 13
2.2　实验原理 …… 13
2.3　实验内容与步骤 …… 14
2.3.1　中心、离散趋势描述实验 …… 14
2.3.2　频数分布分析实验 …… 18
2.4　思考题 …… 24
实验 3　参数检验 …… 26
3.1　实验目的与要求 …… 26
3.2　实验原理 …… 26
3.3　实验内容与步骤 …… 28
3.3.1　单样本 t 检验 …… 28
3.3.2　两独立样本 t 检验 …… 29
3.3.3　两配对样本 t 检验 …… 31
3.3.4　单因素完全随机设计的方差分析 SPSS 过程 …… 33
3.3.5　单因素重复测量设计的方差分析 SPSS 过程 …… 37
3.3.6　多因素完全随机设计方差分析的 SPSS 过程 …… 40
3.4　思考题 …… 45
实验 4　非参数检验 …… 46
4.1　实验目的与要求 …… 46
4.2　实验原理 …… 46
4.3　实验内容与步骤 …… 47

4.3.1 单样本二项分布检验的 SPSS 过程 …… 47
4.3.2 相关样本二项分布检验的 SPSS 过程 …… 50
4.3.3 独立样本二项分布检验的 SPSS 过程 …… 53
4.3.4 适合性卡方检验的 SPSS 过程 …… 55
4.3.5 独立性卡方检验的 SPSS 过程 …… 59
4.3.6 符号与符号秩次检验的 SPSS 过程 …… 62
4.3.7 秩和检验(曼 - 惠特尼 U 检验)的 SPSS 过程 …… 63
4.3.8 中位数检验的 SPSS 过程 …… 65
4.4 思考题 …… 67
实验 5 相关分析 …… 69
5.1 实验目的与要求 …… 69
5.2 实验原理 …… 69
5.3 实验内容与步骤 …… 70
5.3.1 二元变量相关分析的 SPSS 过程 …… 70
5.3.2 肯德尔和谐系数计算的 SPSS 过程 …… 72
5.3.3 偏相关分析的 SPSS 过程 …… 75
5.4 思考题 …… 79
实验 6 回归分析 …… 80
6.1 实验目的与要求 …… 80
6.2 实验原理 …… 80
6.3 实验内容与步骤 …… 81
6.3.1 一元线性回归分析的 SPSS 过程 …… 81
6.3.2 多元线性回归分析的 SPSS 过程 …… 84
6.4 思考题 …… 89
实验 7 因子分析 …… 91
7.1 实验目的与要求 …… 91
7.2 实验原理 …… 91
7.3 实验内容与步骤 …… 92
7.3.1 因子分析的 SPSS 过程 …… 92
7.3.2 因素分析结果的读取与解释 …… 97
7.4 思考题 …… 102
实验 8 IBM SPSS Modeler 软件使用基础 …… 103
8.1 实验目的与要求 …… 103
8.2 实验原理 …… 103
8.2.1 IBM SPSS Modeler 简介 …… 103
8.2.2 数据挖掘的 CRISP - DM 模型 …… 103
8.2.3 Modeler 软件使用的技巧 …… 105

8.3 实验内容与步骤 …… 107
8.3.1 Modeler 的启动和界面布局 …… 107
8.3.2 完整建模流程的介绍 …… 110
8.4 思考题 …… 114
实验 9 关联规则挖掘实验 …… 115
9.1 实验目的与要求 …… 115
9.2 实验原理 …… 115
9.2.1 关联规则处理数据的两种形式 …… 115
9.2.2 关联规则相关概念 …… 116
9.3 实验内容与步骤 …… 117
9.3.1 Apriori 算法应用 …… 117
9.3.2 序列关联应用 …… 123
9.4 思考题 …… 127
实验 10 决策树分类实验 …… 128
10.1 实验目的与要求 …… 128
10.2 实验原理 …… 128
10.2.1 决策树分类原理 …… 128
10.2.2 决策树分类常用算法 …… 128
10.3 实验内容与步骤 …… 129
10.3.1 导入数据 …… 129
10.3.2 数据认识与处理 …… 130
10.3.3 建立模型与评估 …… 134
10.4 思考题 …… 136
实验 11 支持向量机 SVM 分类实验 …… 137
11.1 实验目的与要求 …… 137
11.2 实验原理 …… 137
11.3 实验内容与步骤 …… 138
11.3.1 导入数据 …… 138
11.3.2 建立模型 …… 139
11.4 思考题 …… 143
实验 12 人工神经网络分类实验 …… 144
12.1 实验目的与要求 …… 144
12.2 实验原理 …… 144
12.3 实验内容与步骤 …… 145
12.3.1 导入数据 …… 145
12.3.2 模型建立 …… 145
12.4 思考题 …… 151

实验 13　贝叶斯方法分类实验 …… 152
13.1　实验目的与要求 …… 152
13.2　实验原理 …… 152
12.2.1　贝叶斯定理和朴素贝叶斯 …… 152
13.2.2　Modeler 中的贝叶斯分类器 …… 153
13.3　实验内容与步骤 …… 154
13.3.1　数据导入 …… 154
13.3.2　贝叶斯网络建模 …… 155
13.4　思考题 …… 159
实验 14　*K* 均值与二分法聚类实验 …… 160
14.1　实验目的与要求 …… 160
14.2　实验原理 …… 160
14.2.1　聚类分析 …… 160
14.2.2　K－Means 聚类 …… 161
14.2.3　两步聚类 …… 162
14.3　实验内容与步骤 …… 162
14.3.1　*K* 均值聚类 …… 162
14.3.2　两步法类 …… 167
14.4　思考题 …… 171
参考文献 …… 172

实验1　IBM SPSS Statistics 软件使用基础

1.1　实验目的与要求

（1）了解 IBM SPSS Statistics 软件主要功能。

（2）熟悉 IBM SPSS Statistics 软件的安装及相关基本操作操作过程。

（3）掌握建立数据文件以及对数据进行编辑整理的方法。

1.2　实验原理

1. SPSS 简介

统计产品与服务解决方案（Statistical Product and Service Solutions，SPSS）软件，最初全称为社会科学统计软件包（Solutions Statistical Package for the Social Sciences），但是随着SPSS 产品服务领域的扩大和服务深度的增加，SPSS 公司已于 2000 年将其正式更改为“统计产品与服务解决方案”，用于统计学分析运算、数据挖掘、预测分析和决策支持任务的系列软件产品及相关服务的总称 SPSS，其中 SPSS Statistics 软件作为 SPSS 中的重要一员，在各领域的统计分析中有着广泛的应用，有 Windows 和 Mac OS X 等版本，本书主要以 Windows 下 SPSS Statistics 20.0 版本为例进行相关介绍，简称 SPSS。

2. SPSS 主要功能

SPSS 软件的基本功能包括数据管理、统计分析、图表分析、输出管理等。SPSS 采用类似 EXCEL 表格的方式输入与管理数据，数据接口较为通用，能方便的从其他数据库中读入数据，提供了从简单的统计描述到复杂的多因素统计分析方法，统计分析过程包括描述性统计、均值比较、一般线性模型、相关分析、回归分析、对数线性模型、聚类分析、数据简化、生存分析、时间序列分析、多重响应等几大类，每类中又分好几个统计过程。比如回归分析中又分线性回归分析、曲线估计、Logistic 回归、Probit 回归、加权估计、两阶段最小二乘法、非线性回归等多个统计过程，而且每个过程中又允许用户选择不同的方法及参数。SPSS 输出结果十分美观，存储时则是专用的 SPO 格式，可以转存为 HTML 格式和文本格式，具有专门的绘图系统，可以根据数据绘制各种图形。

1.3　实验内容与步骤

1.3.1　安装、启动与退出

Statistics 软件全面支持 Windows 操作系统，其基本操作方式和界面窗口与一般软件相同，操作十分简便。

1. SPSS 20.0 的安装

双击 spss20. exe 安装文件,启动安装程序,如图 1 -1 所示。

图 1 -1　安装启动界面

系统自动进行准备工作后,即可按照提示一步步进行,如图 1 -2 所示。

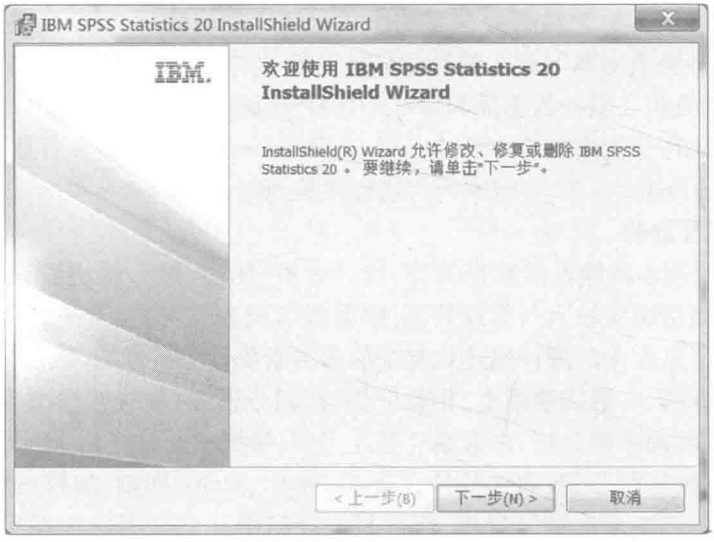

图 1 -2　安装过程界面

每步操作均要认真阅读屏幕显示的信息和提示,安装完成后,系统给出安装报告,可以查看 SPSS 的安装位置、已经安装的 SPSS 组件和统计分析模块。如果对安装的内容有不同的考虑,还可以使用鼠标单击"上一步"返回,或再向前到更前面的步骤,以便改变安装位置和重新选择所需要安装的模块。

2. SPSS 20.0 的启动与退出

SPSS 20.0 的启动和退出方式与 Windows 操作系统下的一般软件完全相同。

1) SPSS 20.0 的启动

安装后双击桌面上的 SPSS Statistics 20.0 图标即可，或者在“开始”菜单中依次选择“程序”→“IBM SPSS Statistics”→“IBM SPSS Statistics 20”命令。启动后会出现启动选项界面如图 1－3 所示，提示 SPSS 20.0 成功启动。

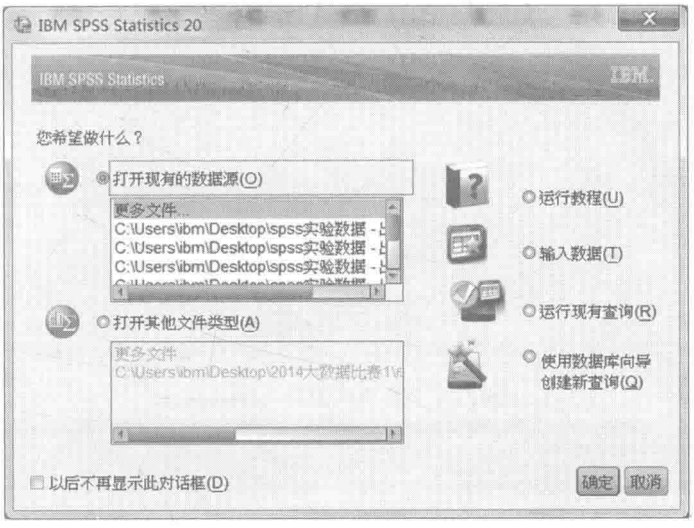

图 1－3　SPSS 启动界面

SPSS 有 6 个启动选项：“运行教程”“输入数据”“运行现有查询”“使用数据库向导创建新查询”“打开现有的数据源”和“打开其他文件类型”。

（1）运行教程：可以浏览运行指导。

（2）输入数据：选择此项，系统将进入数据编辑窗口，用户可以建立新的数据文件或输入数据。

（3）运行现有查询：选择此单选按钮后，系统会让用户选择运行一个查询文件。

（4）使用数据库向导创建新查询：选择此单选按钮后，系统将进入数据库向导，用户可以利用数据库向导导入数据以创建一个新的数据文件。

（5）打开现有的数据源：选择此单选按钮后，系统会让用户选择运行一个 SPSS 数据文件。

（6）打开其他文件类型：选择此单选按钮表示要打开一个其他类型的数据文件。

2）SPSS 20.0 的退出

在菜单栏中选择“文件”→“退出”命令或者单击数据编辑窗口右上角的“关闭”按钮，都可以退出 SPSS。

1.3.2　定义变量

启动 SPSS 后，选择“输入数据”选项，出现如图 1－4 所示的数据编辑窗口。由于目前还没有输入数据，因此显示的是一个空文件。

输入数据前首先要定义变量。定义变量即要定义变量名、变量类型、变量长度（小数位数）、变量标签（或值标签）和变量的格式。

单击数据编辑窗口左下方的“变量视图”标签或双击列的题头（变量），进入如

图 1 -4所示的变量定义窗口,在此窗口中即可定义变量。

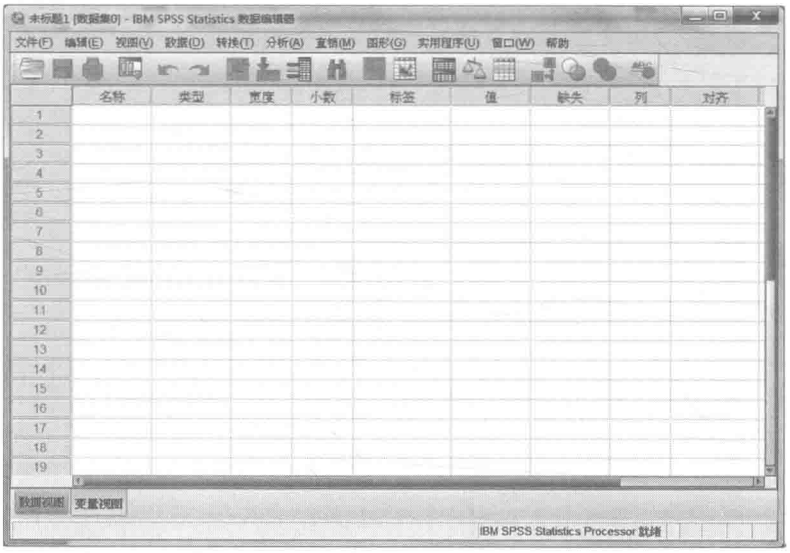

图 1 -4 IBM SPSS Statistics 数据编辑器示意图(变量视图)

1. 变量的定义信息

该窗口的每一行代表一个变量的定义信息,包括名称、类型、宽度、小数、标签、值、缺失、列、对齐、度量标准等。

1) 定义变量名——名称

SPSS 默认的变量为 Var00001、Var00002 等。用户也可以根据自己的需要来命名变量。SPSS 变量的命名和一般的编程语言一样,有一定的命名规则,具体内容如下。

(1) 变量名必须以字母、汉字或字符@ 开头,其他字符可以是任何字母、数字或“_”“@”“#”“$”等符号。

(2) 变量最后一个字符不能是句号。

(3) 变量名总长度不能超过 8 个字符(即 4 个汉字)。

(4) 不能使用空白字符或其他特殊字符(如“!”“?”等)。

(5) 变量命名必须唯一,不能有两个相同的变量名。

(6) 在 SPSS 中不区分大小写,例如,HXH、hxh 或 Hxh 对 SPSS 而言,均为同一变量名称。

(7) SPSS 的句法系统中表达逻辑关系的字符串不能作为变量的名称,如 ALL、AND、WITH、OR 等。

2) 定义变量类型——类型

单击“类型”相应空单元中的按钮,出现如图 1 -5 所示的对话框,在对话框中选择合适的变量类型并单击“确定”按钮,即可定义变量类型。

SPSS 的常用变量类型如下:

(1) 数值:数值型。定义数值的宽度(Width),即“整数部分 + 小数点 + 小数部分”的位数,默认为 8 位;定义小数位数(Decimal Places),默认为 2 位。

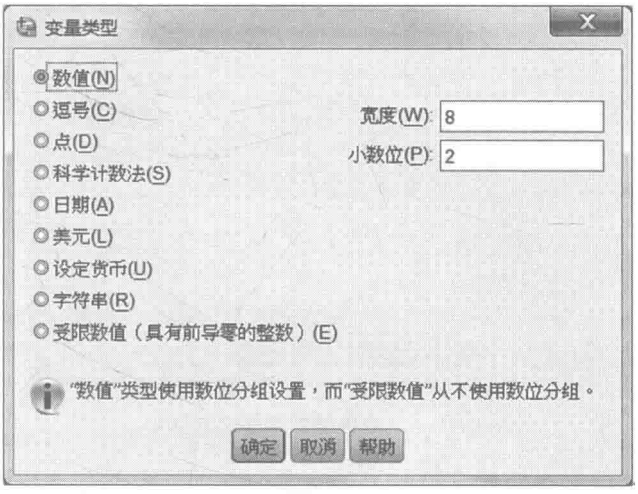

图 1 -5　变量类型定义对话框

（2）逗号：加显逗号的数值型。即整数部分每 3 位数加一逗号，其余定义方式同数值型，也需要定义数值的宽度和小数位数。

（3）点：用户自定义型。如果没有定义，则默认显示为整数部分每 3 位加一逗号。用户可定义数值宽度和小数位数。如 12345.678 显示为 12,345.678。

（4）科学计数法：科学记数型。同时定义数值宽度（width）和小数位数（Decimal），在数据编辑窗口中以指数形式显示。如定义数值宽度为 9，小数位数为 2，345.678 就显示为 3.46E +02。

（5）字符串：字符型。用户可定义字符长度（Characters）以便输入字符。

3）变量长度——宽度

·设置变量的长度，当变量为日期型时无效。

4）变量小数点位数——小数

设置变量的小数点位数，当变量为日期型时无效。

5）变量标签——标签

变量标签是对变量名的进一步说明或注释，变量只能由不超过 8 个字符组成，而 8 个字符经常不足以说清楚变量的含义。而变量标签可长达 120 个字符、可显示大小写，需要时可借此对变量名的含义加以清晰的解释。

6）变量值标签——值

变量值标签是对变量的每一个可能取值的进一步描述。当变量是称名变量或顺序变量时，这是非常有用的。例如，在统计中经常用不同的数字代表被试的性别是男或女；被试的装备是飞机、坦克，还是枪支；被试的教育程度是高中以下、本科、硕士，还是博士等信息。为避免以后对数字所代表的类别发生遗忘，就可以使用变量值标签加以说明和记录。比如用 1 代表“male（男）”、2 代表“female（女）”，其设置方法为：单击“值”相应单元，出现如图 1 -6 所示的对话框；在第一个“值”文本框内输入 1，在第二个“标签”文本框内输入“male”；单击“添加”按钮；再重复这一过程完成变量值 2 的标签，就完成了该变量所有可能取值的标签的添加。

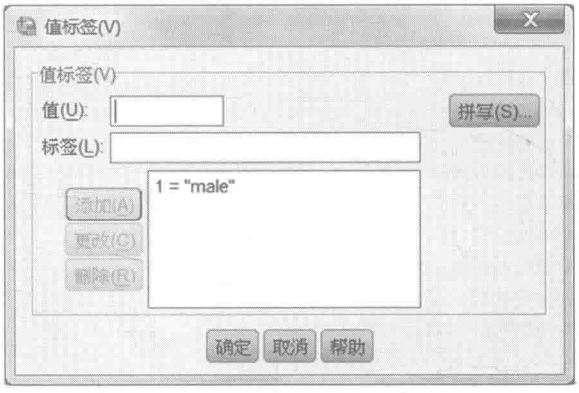

图 1-6　变量值标签定义对话框

7）变量的显示宽度——列

输入变量的显示宽度，默认为 8。

8）变量的测量尺度——度量

变量按测量水平可被划分为称名变量、顺序或等级变量、等距变量和等比变量几种。这里可根据测量量表的不同水平设置对应的变量测量尺度，设置方式为：称名变量选择名义，顺序或等级变量选择序号，等距变量和等比变量均选择度量。

2. 变量定义信息的复制

如果有多个变量的类型相同，可以先定义一个变量，然后把该变量的定义信息复制给其他类型相同的变量。具体操作为：先定义好一个变量，在该变量的行号上单击右键，在弹出的快捷菜单中选择“复制”命令，然后选样其他同类型变量所在行，单击鼠标右键，在弹出的快捷菜单中选择“粘贴”。这样就复制了同样的变量定义信息给一个新的变量，用户再根据需要将自动产生的新变量名改为所要的变量名。

1.3.3　数据的输入与保存

1. 数据输入的一般方法

定义了所有变量后，单击“数据视图”标签，即可在数据视图中输入数据。数据编辑窗口中黑框所在的单元为当前的数据单元，表示用户正在对该数据单元录入数据或正在修改该单元中的数据。因此，在录入数据时，用户应首先将黑框移至想要输入数据的单元格上。

数据录入时可以逐行录入，即完成一个个案行所有变量数值的录入，再转入下一行即下一个个案；也可以逐列录入，即按照变量录入数据，录完一个变量列后再转入到下一个变量列。

2. SPSS 数据文件的保存

在录入数据时，应及时保存数据，防止数据的丢失，以便以后再调用该数据。具体步骤如下。

（1）选择“文件”菜单的“保存”命令，可直接保存为 SPSS 默认的数据文件格式（*.sav）。

（2）选择“文件”菜单的“另存为”命令，弹出“另存为”对话框，根据自己的需要指定

数据文件储存的路径和文件名。

1.3.4 数据文件的编辑与转换

经过变量定义与数据的录入,初期的数据文件即可建成。但在后续的数据分析过程中,常常需要对数据文件进行多方面的修订、编辑与变换。我们选择其中最为常用的操作给予简明的介绍。

1. 数据的编辑

1) 增加和删除一个个案

研究者经常需要在某个个案前面或后面插入新的个案。例如要在第6个观察单位前增加一个观察单位(即在第6行前增加一行,使原来的第6行下移成为第7行),可先激活第6行的任一单元格,然后选样"编辑"菜单中的"插入个案"命令,系统自动在第6行前插入一个新的行,原第6行自动下移一行成为第7行。然后把新增个案的各个变量值输入相应的单元格。

如要删除第9行(即删除这个个案的所有观察值),则可先单击第9行的行头,这时整个第9行被选中(呈黑底白字状),然后按删除键或选择"编辑"菜单中的"清除"命令,该行即被删除。

2) 数据的排序

在数据文件中,可根据一个或多个排序变量的值重排个案的顺序。选择"数据"菜单的"排序个案"命令,弹出对话框,如图1-7所示。

图1-7 根据变量值对个案重新排序对话框

在变量名列表框中选择1个需要按其数值大小排序的变量(也可选多个变量,系统将按变量选择的先后逐级依次排序),单击图中"▶"按钮使之添加到"排序依据"框中,然后在"排列顺序"框中选择是按升序(从小到大)还是降序(从大到小)排列,单击"确定"钮即可。

3) 选择个案子集

在数据统计中可从所有资料中选择部分数据进行统计分析。选择"数据"菜单中的"选择个案"命令,弹出对话框,如图1-8所示。通过单击该对话框上不同的按钮,可以确定用不同的方式对个案进行选择。系统提供的选择方式有五种,但是常用的主要有如下两种。