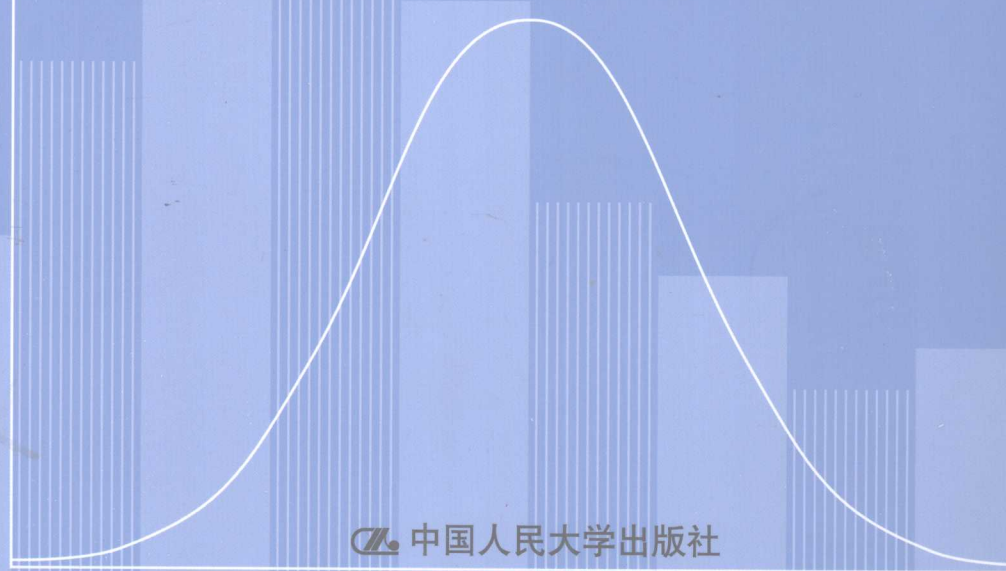
中国人民大学统计咨询研究中心
中国人民大学数据挖掘中心
中国人民大学概率论与数理统计研究所
教育部重点科研基地应用统计科学研究中心
联合推出

数据分析系列教材

统计学
方法与应用

黄向阳　谢邦昌　编著



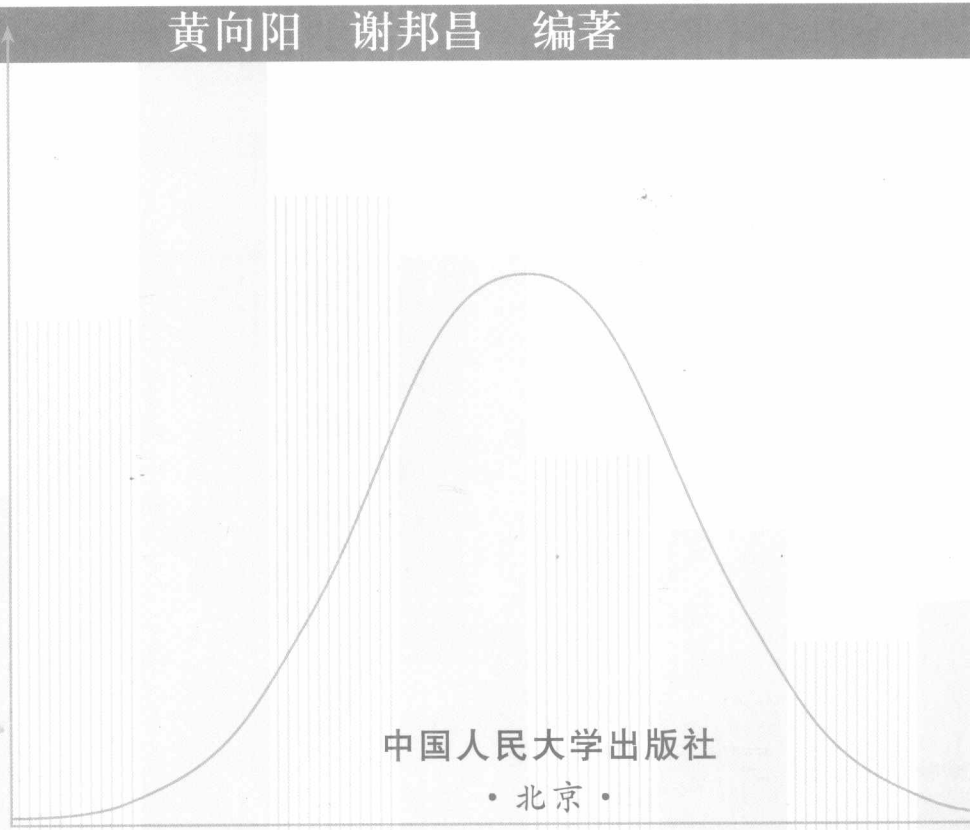
中国人民大学出版社

中国人民大学统计咨询研究中心
中国人民大学数据挖掘中心
中国人民大学概率论与数理统计研究所
教育部重点科研基地应用统计科学研究中心
联合推出

数据分析系列教材

统计学
方法与应用

黄向阳　谢邦昌　编著



中国人民大学出版社
·北京·

图书在版编目（CIP）数据

统计学：方法与应用/黄向阳，谢邦昌编著.
北京：中国人民大学出版社，2009
（数据分析系列教材）
ISBN 978-7-300-10224-5

Ⅰ. 统…
Ⅱ. ①黄…②谢…
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（2009）第 000894 号

数据分析系列教材
统计学：方法与应用
黄向阳　谢邦昌　编著

出版发行	中国人民大学出版社		
社　　址	北京中关村大街 31 号	**邮政编码**	100080
电　　话	010－62511242（总编室）		010－62511398（质管部）
	010－82501766（邮购部）		010－62514148（门市部）
	010－62515195（发行公司）		010－62515275（盗版举报）
网　　址	http://www.crup.com.cn		
	http://www.ttrnet.com(人大教研网)		
经　　销	新华书店		
印　　刷	北京丰印诚印务有限公司		
规　　格	170 mm×228 mm　16 开本	**版　　次**	2009 年 1 月第 1 版
印　　张	13.75 插页 1	**印　　次**	2009 年 1 月第 1 次印刷
字　　数	246 000	**定　　价**	19.00 元

《数据分析系列教材》编委会

出教材学术精品　育人文社科英才

中国人民大学出版社读者信息反馈表

尊敬的读者：

感谢您购买和使用中国人民大学出版社的_____一书，我们希望通过这张小小的反馈卡来获得您更多的建议和意见，以改进我们的工作，加强我们双方的沟通和联系。我们期待着能为更多的读者提供更多的好书。

请您填妥下表后，寄回或传真回复我们，对您的支持我们不胜感激！

1. 您是从何种途径得知本书的：
   ❑书店　❑网上　❑报刊杂志　❑朋友推荐
2. 您为什么决定购买本书：
   ❑工作需要　❑学习参考　❑对本书主题感兴趣
   ❑随便翻翻
3. 您对本书内容的评价是：
   ❑很好　❑好　❑一般　❑差　❑很差
4. 您在阅读本书的过程中有没有发现明显的专业及编校错误，如果有，它们是：_____

   _____

   _____
5. 您对哪些专业的图书信息比较感兴趣：_____

   _____
6. 如果方便，请提供您的个人信息，以便于我们和您联系（您的个人资料我们将严格保密）：

   您供职的单位：_____

   您教授的课程（教师填写）：_____

   您的通信地址：_____

   您的电子邮箱：_____

   请联系我们：

   电话：62515732　62514162　82501704

   传真：62514775

   E-mail：rdcbsjg@crup.com.cn　rdig@rdjg.com.cn

   通讯地址：北京市海淀区中关村大街甲59号文化大厦15层　100872

   中国人民大学出版社工商管理出版分社

总　序

随着社会经济的不断发展、科学技术的不断进步，统计方法越来越成为人们必不可少的工具和手段。在教学过程中，老师们也越来越感觉到运用统计方法解决实际问题的重要性，不少人在探索如何运用统计软件和学习统计方法。谢邦昌教授、黄登源教授在多年的教学中，积累了丰富的经验，他们热情倡议将他们的讲稿提供出来并编写成教材，供更多的人学习和使用。这正与我们的初衷不谋而合。2005 年我们便开始着手这套系列教材的编写，经过不断的讨论、反复的论证，形成了现在的模式。由于有许多研究生的帮忙，再加上几位年轻老师的辛劳，这套书终于问世。

在我们看来，掌握统计方法不仅要在理论上弄明白，更重要的在于能够正确有效地运用这些方法，分析说明实际问题。这套书正是试图利用实际数据，通过统计软件的实际操作，将所能使用的统计方法加以说明，使读者不仅能够了解相应的统计方法，而且能够通过计算机操作学会运用这些方法处理分析实际数据。希望本套书的出版能够为读者提供这样学习的工具。

由于水平有限，书中难免不足之处。恳请读者朋友提出宝贵意见。我们也会遵循这样的思路，在教学以及和读者的交流中不断积累、不断提高、不断完善，奉献给读者更多更好的成果。

感谢为这套书的编写付出汗水的研究生，感谢几位认真用心的年轻老师，感谢中国人民大学出版社的大力支持。为方便读者，书中的所有例题数据，都将放在中国人民大学出版社工商管理分社的网站（www. rdjg. com. cn）上，供读者下载并练习使用。感谢读者，希望能够加强沟通和联系，为提高统计方法实际运用的能力和水平共同努力。

易丹辉

前　言

Conover 在《实用非参数统计》第一章的导言中写道："非参数统计一个诱人的特性是：你并不需要成为一个概率论方面的专家就能理解非参数方法所蕴含的理论。……本章介绍的这些基本概念，所需要的只是耐心、信心和比较好的高中代数知识。"这段话揭示了统计方法大众化时代的典型特征：概率论和数理统计的理论对于实际工作者来说，不再是难以逾越的障碍，在大多数情况下，借助现代计算机的发展，统计方法已经高度工程化，或者说"傻瓜"化了。

而对于应用统计领域的专业人员来说，普及统计方法的重点已经转移到解释统计方法的基本原理和展示统计软件的操作这两个环节上。连接统计方法原理、软件内部过程和最终结果展示的诸多环节都交给统计专家和软件工程师，使用者并不一定需要了解。就像我们对待计算机的态度一样，我们能够大致了解计算机的基本原理，也知道如何通过键盘、界面和鼠标完成任务，但是对技术细节只能有很不充分的理解。统计方法的未来大约也是如此，但这种转变的背后隐藏着滥用统计方法的危险。而统计方法的数学背景和定量分析形象又赋予它很强的说服力和权威性，更容易受到有意或者无意的滥用，所以介绍统计软件操作的时候还要避免过于简化的倾向。

本书的重点是说明如何利用 SPSS 软件完成常见的统计分析，所以它首先是一本操作指南，但是读者还应该留意对统计原理的讲解和对滥用统计方法的提醒。既然是操作指南，读者在使用的时候只需具备基本的统计学知识就可以了。不过，本书第 1 章对所有不熟悉 SPSS 的读者来说都是必读内容，通过这一章的学习，读者可以掌握建立 SPSS 数据集的基本方法，其他章则可以根据分析的需要选择使用。

使用本书的最佳方法是模仿。第一步，建立书中提供的数据集，然后亦步亦趋地实现书中所有的分析功能并解释所得到的结果。第二步，建立自己要分析的数据集，首先实现在第一步中掌握的分析方法的应用并解释所得到的结果是否合理。第三步，探索更多的分析选项，可能得到更好的分析方法，因为越熟悉统计分析软件，它就越能够适合分析需求。为了配合书中的内容，需要一些供练习使用的数据集，在没有数据集的情况下可以使用 SPSS 自带的数据集，本书将在相应章节后列出建议使用的数据集。这些数据集的路径是安装在 SPSS 的子目录下的 \ tutorial \ sample _ files \ ，其中的 . sav 文件可以使用，利用数据集的变量视图（Variable View）就可以了解各个变量的含义。如果安装的时候没有自行指定安装位置，则完整的路径应该是 C: \ Program Files \ SPSS \ Tutorial \ sample _ files。下面简单介绍一下各章的内容和值得关注的重点。

第 1 章介绍了统计分析对数据测量尺度的理解，定类尺度、定序尺度、定距尺度、定比尺度四类尺度和 SPSS 中的变量类型并不相同，注意不要混淆。除了基本操作以外，本章介绍的建立多应答问题交叉表的方法值得特别关注。

第 2 章的内容是均值比较和正态性检验，应该说是比较容易的部分。

第 3 章讲非参数统计方法，读者可以通过操作分析过程，体会一下 Conover 那段话的含义。秩的概念和计算方法是非参数方法的基础，可以花一点时间通过例 3. 1 熟悉秩的计算。此后的内容就是比较程式化的，要点是根据问题的类型选择相应的分析方法。比如两组独立数据集中趋势的比较要使用 Mann-Whitney 检验，成对数据集中趋势的比较要使用符号检验和 Wilcoxon 符号秩检验，等等。判断问题类型和分析方法之间合适与否是应用统计方法的关键所在。

第 4 章介绍相关分析和回归分析。要点是区分变量之间的数量关系和逻辑关系，尤其要注意数量关系导致的伪相关现象，要知道滥用回归分析模型是滥用统计方法的重灾区。从数据分析的一般方法来看，对数据的处理可以是任意的，就是说没有人能够阻止你用 SPSS 建立几个任意变量之间的回归模型，但这个模型的分析结果有无意义就是另外一回事了。逐步回归是建立回归模型的必经阶段，一些起码的模型检验手段，比如偏相关系数、异方差、残差序列相关、多重共线性对于回归分析都是必不可少的。

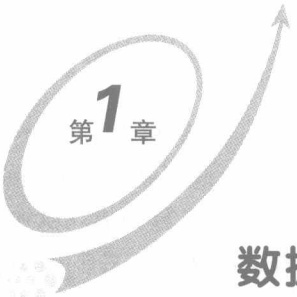
第 5 章介绍方差分析。方差分析在实际使用的时候，也有一个变量选择问题，即考察多少个因素是一个大问题。建议尽量使用多因素方差分析的框架来设计试验。

第 6 章介绍调查数据的信度分析和效度分析。问卷设计和调查实施是非常复杂的领域，最好结合阅读介绍量表设计的书来熟悉软件操作。

本书力图展现统计方法在诸多领域的实际应用，内容比较宽泛，对具体背景和使用条件的交代难免不足，希望读者谅解。中国人民大学统计学院博士研究生邱南南、李扬，硕士研究生程冬旭、刘冬参与了本书的写作，在此谨表谢意。

本书的例题和练习题的数据都放在中国人民大学出版社的网站上，供读者下载和练习使用。

目　录



数据的描述统计分析

数据是对自然和社会现象进行计量所得的结果，是统计分析的基础。在取得数据之后，首先应对数据的类型、平均水平、分散程度、分布形态等做出描述，以反映其基本特征。针对上述问题，本章介绍了数据的测量尺度与类型，通过简单运算概括数据的数字特征，并引入图表对数据进行形象具体的展示。

第1节　测量尺度与数据类型

1.1.1　数据的测量尺度

在收集数据之前，我们首先要对事物和现象进行测量。依据对事物测量的精确水平，可以将测量尺度分为定类尺度、定序尺度、定距尺度、定比尺度四类。这四种测量尺度对事物的测量层次由低到高逐步递进，所包含的信息量也依次增加。我们可以将高层次测量尺度的数据转化成低层次测量尺度的数据，但反之不能进行。

1. 定类尺度

定类尺度也称名义尺度或类别尺度（nominal scale），是最粗略的测量尺度。定类尺度根据某种属性对客观事物进行平行的分类，只能反映事物之间的类别差异，而无法反映各类之间的其他差别。例如，按照性别将人口分为男、女两类，按照婚姻状况将人口分成未婚、已婚和离婚等。我们可以计算每一类别中的个体出现的次数或频率。为分类或处理的方便，可以用不同的数字或编号来表示不同的类别，但这些数字不能区分大小或进行数学运算。定类尺度是对事物最基本的测量，是其他测量尺度的基础。

2. 定序尺度

定序尺度又称为顺序尺度（ordinal scale），它测量事物的顺序或等级差异。定序尺度不仅将事物进行分类，而且确定了这些类别的顺序或优劣。例如，产品按质量等级分为优等品、合格品和残次品等几个等级；考试成绩分为A，B，C，D，E五个等级等。定序尺度不仅测量了类别的差异，而且测量了次序的差异。但定序尺度不能测量类别之间的准确差值，因此，定序尺度的测量结果只能比较优劣，而不能进行数学运算。

3. 定距尺度

定距尺度也称为区间尺度（interval scale），不仅能将事物分类排序，而且可以测量类别或次序之间的准确差异，例如考试成绩用百分制度量，温度用摄氏度或华氏度来度量等。定距尺度的每一间隔都是相等的，可以进行加减运算。但定距尺度没有绝对的原点0，即定距尺度的“0”只是一个相对的水平，而不表示“没有”或“不存在”。例如，一个人的物理成绩为0，只表示在给定的评分标准下他的成绩水平为“0”，并不意味着他没有物理知识。定距尺度中的“0”是人为任意确定的，因而定距尺度的数据不能进行乘除运算。

4. 定比尺度

定比尺度也称为比例尺度（ratio scale），除了具有上述三种测量尺度的全部特性外，还可以计算两个测量值的比值。定比尺度与定距尺度的唯一区别在于，定比尺度中必须有一个绝对固定的、非任意确定的零点，定比尺度的“0”表示“没有”或“不存在”。例如，一个人的年龄不可能小于“0”岁，一个人的收入为“0”表示他没有收入。定比尺度中的“0”具有绝对意义，因而其数值不仅可以比较大小，还可以计算数值之间的比例，加、减、乘、除运算均有意义，在现实生活中，大多数情况下我们都使用定比尺度。

1.1.2　数据的类型

采用不同的测量尺度对事物和现象进行测量会得到不同类型的统计数据，从上述四种测量尺度形成的结果来看，可以将统计数据大体分为属性数据和定量数据两种。谈到统计数据，还涉及变量的概念。变量是说明事物和现象的某种特征的概念，具体表现为变量的取值，也就是统计数据。下面我们将结合变量的概念说明统计数据的两大类型。

1. 属性数据

属性数据也称定性数据或品质数据（qualitative data），说明事物的品质特征，不能用数值表现，由定类尺度和定序尺度测量形成。属性数据反映属性变量，例如性别、产品等级等。

2. 定量数据

定量数据也称数量数据（quantitative data），说明事物的数量特征，能够用数值表现，由定距尺度和定比尺度形成。定量数据反映数量变量，例如温度、年龄、收入等，它们都表现为具体的数值。根据取值情况的不同，定量变量可以分为离散变量和连续变量。离散变量的取值可以一一列举，连续变量可以取连续不断的无穷多个值，不能一一列举。

针对不同类型的数据，我们将采用不同的统计方法进行分析。我们可以计算属性数据的频数或频率，对定量数据则可以采用更为复杂的分析方法。

1.1.3　SPSS 的数据文件

在本书中，我们将结合 SPSS 13.0 进行统计概念和方法的阐述，并结合实例讲述具体的操作步骤和结果分析，以期使读者明了基本的统计知识，并具备一定的实际操作能力。SPSS 是世界上流行的三大统计分析软件之一，在社会科学、自然科学领域的统计分析中发挥了巨大的作用，并以其友好的操作界面、简单易用的分析功能和美观的输出结果赢得了广大用户的欢迎和喜爱。

1. 数据文件的建立

启动 SPSS 程序，出现图 1.1 所示的界面。SPSS 在启动时，已经建立了一个名为“Untitled”的新的数据文件，新的数据文件的建立还可以通过选择菜单“File”→“New”→“Data”实现。SPSS 的数据文件包括数据窗口“Data View”和变量窗口“Variable View”两个标签。“Data View”如图 1.1 所示，每一列代表一个变量，每一行代表一个观测值，可以直接点击数据文件的空白单元格进行数据的录入和编辑。点击“Variable View”，如图 1.2 所示，对变量进

行定义，各个选项的含义依次为：

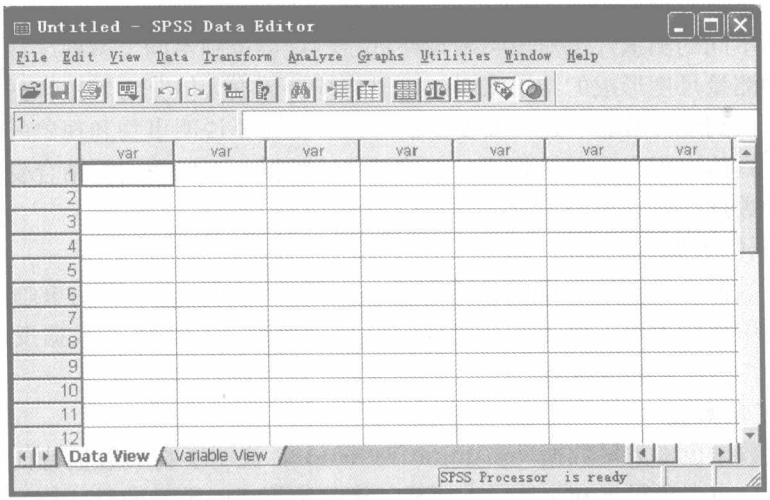


图 1.1 SPSS 的启动

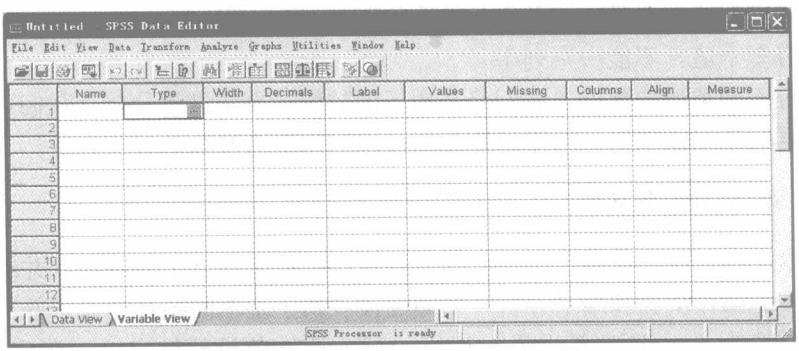


图 1.2 Variable View 标签

Name：变量名。变量名默认为“VAR…”，可自行定义。变量名由不多于 8 个的字母或数字组成，不区分大小写字符。

Type：变量类型。变量有三种基本类型，分别为数值型、字符型和日期型，具体细分为八种，即：Numeric，Comma，Dot，Scientific notation，Date，Dollar，Custom currency，String。单击“Type”，会出现图示的右侧小方框，点击此方框，将出现包含上述变量类型的选项卡（见图 1.3），可从中选择变量类型。

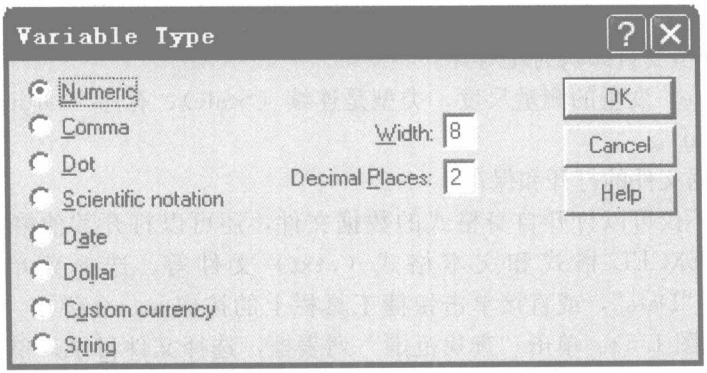


图 1.3　定义变量类型选项卡

Width，Decimals：变量长度和小数位数。

Label：变量标签。当变量名不足以表示变量的含义时，可以在变量标签栏中对变量的具体含义进行详细注释和说明。

Values：变量值标签。变量值标签对变量的可能取值做进一步说明，用于对定类变量的取值进行注释。单击“Values”，会出现类似“Type”的右侧小方框，进而点击此方框，将出现定义变量值标签的选项卡（见图 1.4）。对变量值进行定义之后，可以点击快捷工具栏上的按钮，对定类变量的取值和标签进行切换。

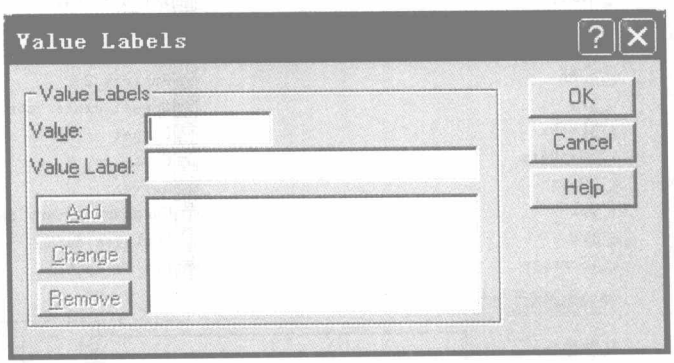


图 1.4　定义变量值标签选项卡

Missing：缺失值的标记。字符型变量默认的缺失值为空格，数值型变量默认的缺失值为零。

Columns：列宽，是指数据窗口中该变量所占的列数。