第一篇

SPSS 应用

第一章　SPSS 文件建立与编辑

学习目标：

1. 明确 SPSS 软件的用途，了解 SPSS 的基本窗口和菜单，熟练掌握 SPSS 的基本操作。

2. 熟练掌握建立 SPSS 数据文件及管理 SPSS 数据的基本操作方法。

3. 熟练掌握 SPSS 进行数据合并、数据排序、数据选取、变量计算等操作方法。

SPSS 是世界著名的统计分析软件之一，是 Statistical Package for Social Science 的英文缩写，中文译为“社会科学统计软件包”，也是 Statistical Product and Service Solutions 的英文缩写，即统计产品与服务解决方案。20 世纪 60 年代末，美国斯坦福大学的三位研究生研制开发了最早的统计分析软件 SPSS，并于 1975 年在芝加哥成立了专门研发和经营 SPSS 软件的 SPSS 公司。1984 年，SPSS 公司推出了运行在 DOS 操作系统上的 SPSS 计算机版第 1 版，随后又相继推出了第 2 版、第 3 版等。20 世纪 90 年代初，随着计算机 Windows 图形操作系统的出现和盛行，SPSS 公司又研制出了以 Windows 为运行平台的 SPSS 第 5 版、第 6 版。20 世纪 90 年代中后期，为适应用户在 Windows 95 操作系统环境下工作的习惯，并迎合 Internet 的广泛使用，SPSS 第 7 版～第 17 版又相继诞生。目前，SPSS 在全球约有几十万家产品用户，分布于通信、医疗、银行、证券、保险、制造、商业、市场研究、科研教育等多个行业和领域，已成为世界上最流行、应用最广泛的专业统计分析软件之一。

2009 年，IBM 公司斥资 12 亿美元收购了 SPSS 软件公司。SPSS 第 18 版和第 19 版更新命名为 PASW（Predictive Analytics Software，预测分析软件）Statistics。其第 20 版～第 29 版命名为 IBM SPSS Statistics。

从 SPSS/PC+ 版本到目前的最新版本，SPSS 在用户操作和分析结果的展现方面做了非常大的改进。用户的数据管理和统计分析工作可以非常方便地通过鼠标单击菜单或按钮并配合简单的窗口输入来实现，免去了使用者记忆命令和参数的负担，也不需要任何计算机编程。SPSS 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 统计分析过程包括描述统计、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、因子分析、生存分析、时间序列分析、多重响应等，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 也有专门的绘图系统，可以根据数据绘制各种图形。

除了 SPSS 软件之外，这里简要介绍一下其他几种统计软件。

（1）**SAS**。SAS 是功能齐全的软件，价格不菲，许多公司特别是美国制药公司偏爱使用。尽管它现在已经尽量“傻瓜化”，但相关人员仍然需要经过一定的培训才可以使用，使用者也可以对它编程，不过对于基本统计课程使用起来不太方便。

（2）**Stata**。Stata 是用于 Windows、Macintosh 及 Unix 计算机系统下的一种功能完全的统计软件包。它的特点包括易操作、速度快，其具有一整套预先编好的分析与数据管理功能，同时也允许用户根据需要来创建自己的程序、添加更多的功能。Stata 的大部分操作既可以通过下拉菜单系统来完成，也可以更直接地通过键入命令来完成。初学者可以在菜单的帮助下学习使用 Stata，任何人在应用自己所不熟悉的程序时都可以由此获得帮助。Stata 的命令有很强的一致性和直观意义，可以使有经验的用户更为高效地工作，这一特点还使得对更复杂或需要多次重复的任务进行编程变得十分容易。Stata 是应用经济学研究者的一个有力工具，不论所探索的数据是时间序列、面板数据还是横截面数据，Stata 都能帮助人们更容易且更有效地进行分析研究。

（3）**R**。R 是基于 R 语言的一款统计软件。R 语言是一种统计计算语言，是贝尔实验室开发的 S 语言的一种实践，具有许多优点。比如：R 是免费的；R 的更新速度快，可以包含很多最新方法的实现方案，而其他软件的更新则需要比较长的时间；R 可以提供丰富的数据分析技术，功能十分强大；R 的绘图功能强大，可以按照需要画出图形，对数据进行可视化分析。

（4）**Eviews**。Eviews 是一种处理回归和时间序列等问题很方便的经济计量学软件，能够处理以时间序列为主的多种类型数据，进行包括描述统计、回归分析、传统时间序列分析等基本数据分析，以及建立条件异方差、向量自回归等复杂的计量经济模型。

（5）**Excel**。Excel 作为微软公司推出的 Office 软件中的主要成员，因其具有电子表格管理、数据清单管理、商业统计图表处理及数据分析与决策功能，已受到广大用户的青睐，并逐步成为政府机构和企事业单位进行数据处理分析的主要工具。Excel 严格来说并不是统计软件，但有一定的统计计算功能。而且凡是装有 Microsoft Office 的计算机，基本上都有 Excel 。但要注意，有时在安装 Office 时没有安装数据分析的功能，如需使用则必须安装该功能后才能进行数据分析。画图功能是 Excel 默认具备的功能，该功能在对数据的描述统计中非常有用。对于简单分析，如相关分析、一元线性回归分析、方差分析，Excel 还算方便，但随着问题的深入，Excel 就不那么方便了，需要使用宏命令来编程，这时没有相应的简单选项。因此，多数专门的统计推断问题还是需要专门的统计软件来处理。

当然，还有很多其他的软件，此处不再一一列出。对于经济管理类专业的学生而言，如果没有特别的数据分析要求，能够熟练掌握 Excel 和 SPSS 一般就可以达到对数据进行分析的目的。而且，只要学会使用一种软件，使用其他的软件也不会困难，通过阅读帮助和说明即可掌握使用方法。

最后，需要说明的是，学习软件的最好方式是在使用中学习。

1.1　变量和数据

统计学是一门分析数据的科学，因此，在学习数据分析之前，我们有必要对变量和数据有一个全貌的认识。

1.1.1 变量和变量值

变量（variable）是说明现象某种特征的概念。“变量”这个名称来源于某个特征在总体中的所有个体上是变化的这一事实。例如，就业人员的年龄、性别和受教育年限。这里的“年龄”“性别”“受教育年限”就是变量，对于每一位就业人员来说，他们的“年龄”“性别”“受教育年限”都是不同的。

变量的具体取值称为**变量值**。比如，某个人的年龄是 35 岁，性别为女，受教育年限是 12 年，这里的“35 岁”“女”“12 年”就是变量值。对于一个抽取的样本（样本容量为 500）来说，如果我们研究“年龄”这个变量，那么我们所关注的是“一个”变量，但是有“500 个”变量值。

根据变量值的不同，可以将变量分为以下几种类型：

1. **定类型变量**

定类型变量是说明事物类别的一个名称。例如，“性别”就是一个定类型变量，其变量值为“男”或“女”。类似的，“籍贯”“政治面貌”“行业”等都是定类型变量。

2. **定序型变量**

定序型变量是说明事物有序类别的一个名称。例如，“产品等级”就是一个定序型变量，其变量值可以为“一等品”“二等品”“三等品”等；受教育程度也是一个定序型变量，其变量值可以为“小学”“初中”“高中”“大学”等。

3. **数值型变量**

数值型变量是说明事物数字特征的一个名称，例如，“国内生产总值”“产品产量”“年龄”等。根据取值不同，数值型变量又可以分为**离散型变量**和**连续型变量**。

上述的定类型变量和定序型变量属于**定性变量**，数值型变量属于**定量变量**。如果要研究定性变量之间的关系，可采用列联分析、卡方独立性检验等统计分析方法；如果要研究定量变量之间的关系，可采用相关分析、回归分析等统计分析方法；如果要研究定性变量与定量变量之间的关系，可采用方差分析、Logistic 回归、判别分析等统计分析方法。

1.1.2 数据的类型

统计数据是对现象进行测量的结果，上述变量值被统称为**数据**（data）。数据是统计分析的基础，数据质量的优劣直接影响着统计分析结果的准确性。获取高质量的数据是一项系统工程，其中涉及很多问题。本节主要介绍数据的类型及数据获取方法。

1. **分类数据、顺序数据、数值型数据**

按照所采用的计量尺度不同[①]，可以将统计数据分为分类数据、顺序数据和数值型数据。

① 数据的测量尺度有四种。①分类尺度（nominal scale）：按照事物的某种属性对其进行平行的分类，数据表现为类别，没有序次关系，是数据的最低级；②顺序尺度（ordinal scale）：是对事物类别顺序的测度，数据表现为有序的类别，只能比较大小，不能进行加减运算，更不能做乘除运算；③间隔尺度（interval scale）：是对事物类别或次序之间间距的测度，没有绝对零点，数据表现为数字，数据中的 0 是人为设定的，只能做加减运算，不能做乘除运算；④比率尺度（ratio scale）：是对事物类别或次序之间间距的测试，有绝对零点，数据表现为数字，是数据最高级的测度等级，可以做加减乘除运算，以及基于加减乘除的运算。

分类数据是定类型变量的取值。它是只能归于某一类别的非数字型数据，它是对事物进行分类的结果，数据表现为类别，是用文字来表述的。

顺序数据是定序型变量的取值。它是只能归于某一有序类别的非数字型数据。顺序数据虽然也是类别，但这些类别是有序的。例如，考试成绩可以分为优、良、中、及格、不及格。

数值型数据是数值型变量的取值。它是按数字尺度测量的观察值，其结果表现为具体的数值。

分类数据和顺序数据说明的是事物的品质特征，通常是用文字来表述的，其结果均表现为类别，因而也可统称为**定性数据**；数值型数据说明的是现象的数量特征，通常是用数值来表现的，因此也可被称为**定量数据**。

2. **观测数据和实验数据**

按照统计数据的收集方法，可以将其分为观测数据和实验数据。**观测数据**是通过调查或观测而收集到的数据，这类数据是在没有对事物人为控制的条件下得到的，有关社会经济现象的统计数据几乎都是观测数据。**实验数据**则是在实验中控制实验对象而收集到的数据。比如，对一种新药疗效的实验数据，对一种新的农作物品种的实验数据。自然科学领域的大多数数据都是实验数据。

3. **截面数据和时间序列数据**

按照被描述的现象与时间的关系，可以将统计数据分为截面数据和时间序列数据。**截面数据**是在相同或近似相同的同一时点上搜集的数据，这类数据通常是在不同的空间上获得的，用于描述现象在某一时刻的变化情况。例如：同一年份各地区的国内生产总值；第七次人口普查中我国各地区的人口数量；等等。**时间序列数据**是在不同时间收集到的数据，这类数据是按时间顺序收集到的，用于描述现象随时间变化的情况。例如，我国1978—2022 年每年的粮食产量、发电量等。研究宏观经济问题，相关时间序列数据来自国家统计局或一些专业部委的统计年鉴。如果研究微观经济现象，如研究某企业的产值与能耗，那么数据就要在这个企业的相关部门获取。

在分析不同类型的数据时，一定要结合数据特征采用合适的方法。比如：对分类数据，通常计算出各组的频数或频率，进行列联表分析和 χ^2 检验等；对顺序数据，可以计算其中位数和四分位差，计算等级相关系数等；对数值型数据，可以用更多的统计方法进行分析，如计算各种统计量，进行参数估计和假设检验等。

1.1.3　数据的获取

1. **间接获取**

对于大多数使用者来说，亲自去做调查往往是不可能也是不必要的。我们大多会使用有关统计部门和机构发布的统计资料，或者其他机构调查、试验得到的数据，然后将所获取的数据按照自己的需要进行加工、整理，使之成为进行统计分析可以使用的数据。这些数据被称为间接数据或第二手数据。相对来说，这些第二手数据的收集比较容易，收集数据的成本低、花费的时间短。第二手数据的作用也非常广泛，除了可以用于分析所要研究的问题之外，这些资料还可以提供研究问题的背景信息，研究者首先进行探索性分析，回答和检验某些疑

问和假设，寻找研究问题的思路和途径，从而可以更好地定义问题。因此，收集第二手数据是研究者首先要考虑并采用的。分析问题也应该首先从对第二手数据的分析开始。

但是，第二手数据也有很大的局限性，研究者在使用第二手数据时要保持谨慎的态度。因为第二手数据并不是为研究者研究特定的问题而量身定做的，它在解决所研究的问题方面可能有欠缺，如数据的相关性不强、口径不一致、数据时效性不强等。因此，在使用第二手数据前，对第二手数据进行评估是必要的。

对第二手数据进行评估需要考虑下面一些内容。

（1）数据是谁收集的？这主要是考虑数据收集者的实力和社会信誉度。例如，对于全国性的宏观经济数据，与某个专业性的调查机构相比，政府有关部门公布的数据可信度更高。

（2）数据是为什么目的而收集的？一般来说，为了某个群体的得利而收集的数据是值得怀疑的，这样的数据带有某种倾向性。在问题分析中，研究人员一般需要使用权威机构发布的数据。

（3）数据是怎样收集的？收集数据的方法有多种，采用不同方法收集到的数据，其解释力和说服力都是不同的。有些数据是任意抽选的，这样的数据解释力就差，而通过概率手段进行抽样得到的数据解释力就强。如果不了解收集数据所采用的方法，就很难对数据质量做出客观的评价。

（4）数据是何时收集的？数据具有时效性。过时的数据，其说服力和解释力自然会打折扣，因为时代变化很大，过去的数据往往不能准确地描述现在的情况。

使用第二手数据，要注意数据的定义、统计口径和计算方法，避免数据的错用、误用和滥用。在引用第二手数据时，应注明数据的来源，以示对他人劳动成果的尊重并方便对数据的质量进行评估。

有关间接获取的宏观数据和微观数据，本书第十章进行了专门的介绍。

2. 直接获取

虽然第二手数据具有收集方便、耗时少、成本低等优点，但对一个特定的问题而言，第二手资料的主要缺陷可能在于数据的相关性太低。若仅仅依靠第二手资料还不能回答研究所提出的问题，就需要获得第一手数据。

数据的直接来源主要有两种渠道：一是调查或观察；二是试验。

调查是获得社会经济数据的最主要手段，也是很多领域的专家分析研究社会问题的重要基础。调查数据包括国家机关统计部门完成统计调查所获得的数据，也包括企业、机构、个人为特定的需要所完成的调查所获得的数据。数据通常取自有限总体，即总体所包含的个体单位有限。如果调查针对总体中的所有个体单位进行，就把这种调查称为**普查**。普查数据具有信息全面、完整的特点，但是，当总体较大时，进行普查将是一项很大的工程，耗时、费力、成本高，因此普查不可能经常进行。此时，需要把数据收集限制在总体的一个样本上，样本是总体中的一个被选中的部分。选择样本最科学的方法是按照随机原则进行**抽样调查**，产生一个随机样本。调查中常用的概率抽样方法有：简单随机抽样、分层抽样、整群抽样、系统抽样。

试验大多是对自然现象而言。试验数据是在试验中控制试验对象而搜集到的数据。试

验是检验变量间因果关系的一种方法。在试验中，研究人员要控制某一情形的所有相关方面，操纵少数感兴趣的变量，然后观察试验结果。心理学、教育学、社会学、经济学、管理学中有许多使用试验方法获得研究数据的案例。

1.2　SPSS 的使用基础

1.2.1　SPSS 的基本窗口

SPSS 软件运行时有多个窗口，各窗口有各自的作用。使用者要快速入门，只需要熟悉两个基本窗口即可，它们是数据编辑窗口和结果输出窗口。

1. SPSS 数据编辑窗口

打开 SPSS 以后，你会看到如图 1-1 所示的界面，这就是数据编辑窗口。

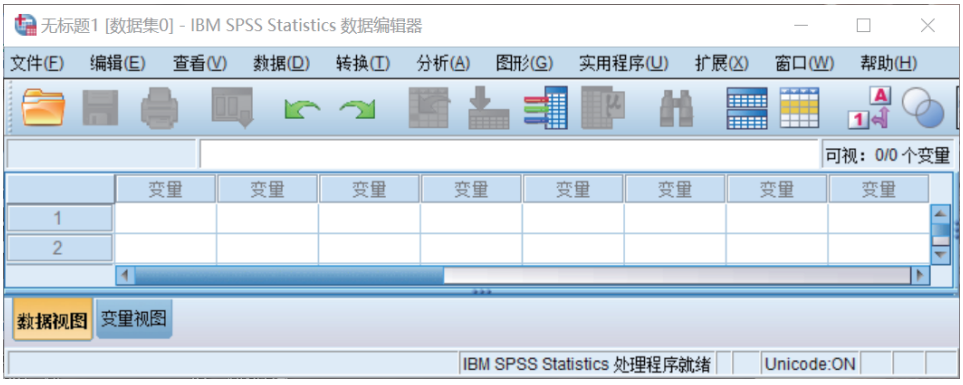


图 1-1　SPSS 数据编辑窗口

这个界面显示的是一个空数据表，其中列代表变量，行代表观测值。界面的上方是 SPSS 的主菜单栏，将数据录入数据表中后，便可以通过选择主菜单中相应的选项（表 1-1），进行数据分析。

表 1-1　SPSS 窗口主菜单及其功能

菜 单 名	功　能	解　释
文件（F）	文件操作	对 SPSS 相关文件进行基本管理（如新建、打开、保存、打印）
编辑（E）	数据编辑	对数据编辑器窗口中的数据进行基本编辑（如撤销 / 恢复、剪切、复制、粘贴），并实现数据查找、软件参数设置等功能
查看（V）	窗口外观状态管理	对 SPSS 窗口外观等进行设置（如状态栏、表格线、变量值标签等是否显示、字体设置等）
数据（D）	数据的操作和管理	对数据编辑器窗口中的数据进行加工整理（如数据的排序、转置、抽样、分类汇总、加权等）
转换（T）	数据基本处理	对数据编辑器窗口中的数据进行基本处理（如生成新变量、计数、分组等）

续表

菜 单 名	功　能	解　释
分析（A）	统计分析	对数据编辑器窗口中的数据进行统计分析和建模（如基本统计分析、均值比较、相关分析、回归分析、非参数检验等）
图形（G）	制作统计图形	对数据编辑器窗口中的数据生成各种统计图形（如条形图、直方图、饼图、折线图、散点图等）
实用程序（U）	实用程序	SPSS 其他辅助管理（如显示变量信息、定义变量集等）
窗口（W）	窗口管理	对 SPSS 的多个窗口进行管理（如窗口切换、最小化窗口等）
帮助（H）	帮助	实现 SPSS 的联机帮助（如语句检索、统计辅导等）

SPSS 数据编辑窗口（窗口标题为 IBM SPSS Statistics 数据编辑器）是 SPSS 的主程序窗口，该窗口的主要功能是定义 SPSS 数据的结构、录入编辑和管理待分析的数据。其中，窗口左下角的【数据视图】用于显示 SPSS 数据的内容，【变量视图】用于显示 SPSS 数据的结构，即对变量进行定义。SPSS 的所有统计分析功能都是针对该窗口中的数据。这些数据通常以 SPSS 数据文件的形式保存在计算机磁盘上，其文件扩展名为 .sav。.sav 文件格式是 SPSS 独有的，一般无法通过 Word、Excel 等其他软件打开。

2. SPSS **结果输出窗口**

结果输出窗口（窗口标题为 IBM SPSS Statistics 查看器）是 SPSS 的另一个主要窗口，该窗口的主要功能是显示管理 SPSS 统计分析结果、报表及图形，如图 1-2 所示。

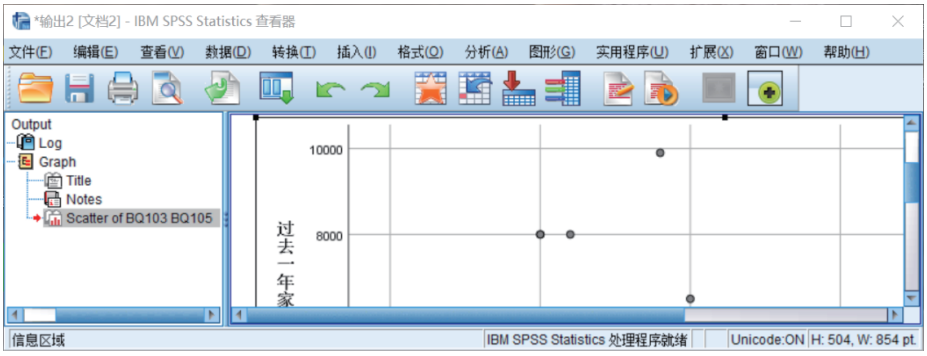


图 1-2　SPSS 结果输出窗口

注：本图是做了一个散点图的输出窗口。

SPSS 统计分析的所有输出结果都显示在该窗口中。输出结果通常以 SPSS 输出文件的形式保存在计算机磁盘上，其文件扩展名为 .spv。.spv 文件格式是 SPSS 独有的，也无法通过 Word、Excel 等其他软件打开。

1.2.2　SPSS 的基本运行方式

SPSS 为用户提供了三种基本运行方式，它们是完全窗口菜单方式、程序运行方式、混合运行方式。这三种运行方式分别适合于不同的用户和不同的统计分析要求。

1. **完全窗口菜单方式**

完全窗口菜单方式是在使用 SPSS 的过程中，所有的分析操作都通过菜单、按钮、输

入对话框等方式来完成。它是一种最常见和普遍的使用方式，其最大的优点是简洁和直观。用户不需要了解任何计算机编程的概念，只要熟悉 Windows 的基本操作并懂得相应的统计知识，就可以非常方便地完成统计分析工作。在操作中，数据编辑器窗口中所有待分析的变量通常显示在窗口左边的列表框中，用户通过鼠标和窗口中间的按钮将本次需要分析的变量选到右边的列表框中。图 1-3 是使用完全窗口菜单方式编辑直方图的对话框。

图 1-3　SPSS 变量选择操作方式

完全窗口菜单方式适合一般的统计分析人员和 SPSS 的初学者。对于一般的数据分析来说，完全窗口菜单方式基本可以满足操作要求。本书在后续软件操作过程中均以该方式为主。

2. 程序运行方式

程序运行方式是在使用 SPSS 的过程中，统计分析人员首先根据自己的分析需要，将数据分析的步骤手工编写成 SPSS 命令程序，然后将编写好的程序一次性提交给计算机执行。SPSS 会按照程序命令语句的前后顺序自动逐句执行相应的命令，并最终给出统计分析结果。

程序运行方式适用于大规模的统计分析工作。它能够依照程序自动进行多步骤的复杂数据分析，分析过程中无须人工干预。这样，即使分析计算的时间较长、分析步骤较多，也能够自动完成，无须人工等待。这种方式需要做两项工作：一是编写 SPSS 程序，二是提交并运行 SPSS 程序。

3. 混合运行方式

混合运行方式是在使用菜单的同时编辑 SPSS 程序，是完全窗口菜单方式和程序运行方式的综合。为实现混合运行方式，用户应首先按照菜单运行方式选择统计分析的菜单和选项，但并不马上单击【确定】按钮提交执行，而是单击【粘贴（P）】按钮。于是，SPSS 将自动把用户所选择的菜单和选项转换成 SPSS 的命令程序，并粘贴到当前语法窗口中。然后，用户可以按照程序运行的方式，对在语法窗口中生成的 SPSS 命令进行必要

的编辑修改，然后再一次性提交给计算机执行。

可见，混合运行方式弥补了完全窗口菜单方式中每步分析操作都要人工干预的不足，同时摆脱了程序运行方式中必须熟记 SPSS 命令和参数的制约，因此是一种较为灵活且实用的操作方式。另外，对于熟练的 SPSS 程序员，可以借助该方式在程序中添加窗口菜单和选项中没有提供的参数。

以上三种使用方式各有千秋，实际使用中应根据应用分析的需要和使用者掌握 SPSS 的程度进行合理的选择。

1.3 SPSS 数据文件的建立和管理

建立 SPSS 数据文件是利用 SPSS 软件进行数据分析的首要工作。没有完整且高质量的数据，也就没有值得依赖的数据分析结论。因此，学会 SPSS 数据文件的建立对于初学者来说也是一件非常重要的事情。

SPSS 数据文件由数据的结构和内容两部分组成。其中，数据的结构记录了数据类型、取值说明、数据缺失情况等必要信息，数据的内容是那些待分析的具体数据。基于此，建立 SPSS 数据文件时应完成两项任务：第一，描述 SPSS 数据的结构；第二，录入编辑 SPSS 的数据内容。这两部分工作分别在 SPSS 数据编辑窗口的【变量视图】和【数据视图】中完成。

1.3.1 SPSS 数据的结构

在输入数据前，通常要先描述数据结构，选择【变量视图】标签，出现的 SPSS 界面如图 1-4 所示。其中，各项内容依次为名称、类型、宽度、小数位数、标签、值、缺失、列、对齐、测量、角色。

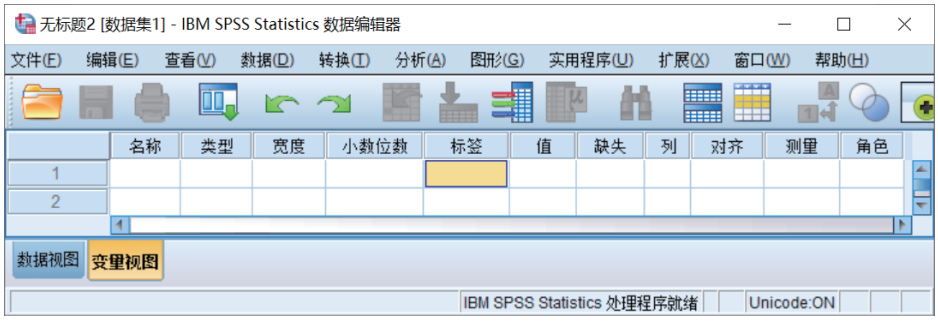


图 1-4 【变量视图】窗口

1. 名称

【名称】是变量访问和分析的唯一标识。在定义 SPSS 数据结构时应首先给出每个变量的名称。在数据编辑窗口中，变量名会显示在数据视图中列标题的位置上。为了方便记忆，变量名最好与其代表的数据含义相对应，变量名可以用英文也可以用汉字，但是它不能与 SPSS 内部特有的具有特定含义的保留字相同，如 ALL、BY、AND、NOT、OR 等，

SPSS 默认的变量名以“变量”命名。

2. **类型、宽度、小数位数**

数据类型【类型】指每个变量的取值类型，相应类型会有默认的列宽【宽度】和小数位宽【小数位数】（图 1-5）。SPSS 中有三种基本数据类型，分别为数值型、字符型和日期型。

图 1-5　定义数据类型

数值型是 SPSS 最常用的数据类型，SPSS 中数值型有五种不同的表示方法。

（1）标准型:【数字（N）】。这是 SPSS 默认的数值类型，默认的列宽为 8，包括正负符号位、小数点和小数位，小数位宽默认为 2。如果数据的实际宽度大于 8，SPSS 将自动按科学计数法显示。

（2）逗号型:【逗号（C）】。逗号型数据其整数部分从个位开始每 3 位以一个逗号分隔，默认的列宽为 8，小数位宽为 2，逗号所占的位数包括在内，如“1,234.5”。用户在输入逗号型数据时可以不输入逗号，SPSS 将自动在相应位置上添加逗号。

（3）圆点型:【点（D）】。圆点型数据其整数部分从个位开始每 3 位以一个圆点分隔，以逗号作为整数和小数部分的分隔符。它默认的列宽为 8，小数位宽为 2，如“1.234,56”。用户在输入圆点型数据时可以不输入圆点，SPSS 将自动在相应位置上添加圆点。

（4）科学计数法型:【科学记数法（S）】。这也是一种常见的数值型数据的表示方式。科学计数法的默认列宽为 8，包括正负符号位、字母 E 和跟在其后的正负符号及两位幂次数字。科学计数法一般用来表示很大或很小的数据。用户在输入科学计数法数据时可以按数值方式输入数据，SPSS 会自动进行转换。

（5）美元符号型:【美元（L）】。美元符号型主要用来表示货币数据，它在数据前附加美元符号。美元型数据的显示格式有很多，SPSS 会以菜单方式将其显示出来供用户选择。用户在输入美元型数据时可以不输入美元符号，SPSS 将自动在相应位置上添加美元符号。

字符型:【字符串（R）】。字符型也是 SPSS 较常用的数据类型，它由一串字符组成。如职工号码、姓名、地址等变量都可定义为字符型数据。字符型数据的默认列宽为 8 个字符位，它不能够进行算术运算，并区分大小写字母。字符型数据在 SPSS 命令处理过程中应用一对双引号引起来，但在输入数据时不用双引号，否则，双引号将会作为字符型数据的一部分。

日期型:【日期（A）】。日期型用来表示日期或时间数据，如生日、成立日期等变量可以定义为日期型。日期型数据的显示格式有很多，如，“dd-mm-yyyy”“mm/dd/yyyy”等。

定义变量时，在 SPSS 数据编辑器窗口的变量视图中的【类型】列下相应的位置单击鼠标，并根据实际数据在弹出对话框中选择相应的数据类型（图 1-5）。

3. **标签**

变量名标签【标签】是对变量名含义的进一步解释说明，它可增强变量名的可视性和统计分析结果的可读性。变量名标签可用中文表示，总长度可达 120 个字符，但在统计分析结果的显示中，一般不可能显示如此长的变量名标签信息。变量名标签这个属性是可以省略定义的，但建议最好给出变量名的标签。不过，如果变量名已经是中文，变量名标签可以省略。在 SPSS 数据编辑器窗口的变量视图中，在【标签】列下相应行的位置输入变量名标签即可。

4. **值**

变量值标签【值】是对变量取值含义的解释说明信息，对于**定类型**（如性别、民族）和**定序型**（如收入的高、中、低）变量尤其重要。例如，对于性别变量，假设用数值 1 表示“男”，用数值 2 表示“女”。那么，人们看到的数据就仅仅是 1 和 2 这样的符号，通常很难弄清楚 1 是代表男还是女。但如果为性别变量附加变量值标签，并给出 1 和 2 的实际指代，就会使数据含义非常清楚，它增强了最后统计分析结果的可读性。变量值标签这个属性是可以省略定义的，但建议最好给出定序或定类变量的变量值标签。

在 SPSS 数据编辑窗口的变量视图中，在【值】列下相应行的位置单击鼠标，可以在弹出对话框中指定变量值标签，如图 1-6 所示。

图 1-6　定义变量值标签

5. **缺失**

缺失数据【缺失】的处理是数据分析准备过程中的一个非常重要的环节。

数据中明显错误或明显不合理的数据，以及漏填的数据都可看作缺失数据。例如，在某项客户满意度的问卷调查数据中，某个被调查者的年龄是 213 岁。这个数据显然是一个不符合实际情况的失真数据。又如，在某项客户满意度的问卷调查数据中，某个被调查者

的年收入没有填，是空缺的。

在利用 SPSS 进行分析时，如果不进行特别说明，SPSS 将会把上述明显错误的数据或空缺数据当作正常且合理的数据进行分析，这必然会影响分析的结果。因此，如果数据中存在缺失数据，分析时通常不能直接采纳，要进行说明。

SPSS 中说明缺失数据的基本方法是指定用户缺失值。首先，在空缺数据处填入某个特定的标记数据。例如，将空缺的年收入数据用特定的标记数据（如 99 999 999）来替代；然后，再指明这个特定的标记数据，以及那些明显的失真数据等为用户缺乏值。这样，在分析时，SPSS 就能够将这些用户缺失值与正常的数据区分开来，并依据用户的处理策略对其进行处理或分析。

对于字符型或数值型变量，用户缺失值可以是 1~3 个特定的离散值【离散缺失值（D）】；对于数值型变量，用户缺失值还可以在一个连续的闭区间内并同时附加一个区间以外的离散值【范围加上一个可选的离散缺失值（R）】。

在 SPSS 数据编辑器窗口的变量视图中，在【缺失】列下相应行的位置单击鼠标，可以根据实际数据在弹出对话框中指定用户缺失值，如图 1-7 所示。

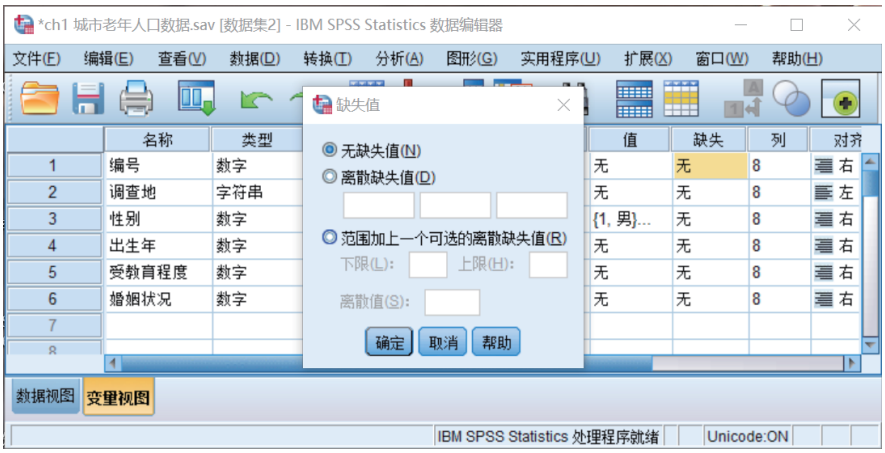


图 1-7　定义用户缺失值

除用户缺失值外，SPSS 还有一类默认的缺失值，被称为系统缺失值。系统缺失值用一个圆点表示，它不等于 0 或 0.00，其通常出现在数值型变量数据中。但是，字符型变量中的空格不是系统缺失值。

如果数据中存在大量缺失值，会对分析产生重大影响。例如，大量的缺失数据会使分析结果出现系统性偏差，会因缺少充分可利用的数据而造成统计计算精度大幅下降，会由于某些模型无法处理缺失数据而限制该模型的应用等。因此，在数据分析前通常需要对缺失数据进行必要的处理。统计学对缺失数据的处理方法有许多，如利用 EM 法（Expectation Maximization）或回归法进行插值估计等，SPSS 也提供了分析缺失数据的专门模块。

6. **测量**

统计学依据数据的计量尺度【测量】将数据划分为三大类：定距型数据、定序型数据和定类型数据。

定距型数据【标度】通常指如身高、体重、收入等连续数值型数据，也包括人数、商品件数等离散型数据。**定序型数据**【有序】具有内在固有大小或高低顺序，一般可以用数值或字符表示。例如：文化程度变量可以有“未受教育”“小学”“初中”“高中”“大学”“研究生”六种情况，可分别用 1、2、3、4、5、6 表示；年龄段变量可以有“老”“中”“青”三个取值，分别用 A、B、C 表示等。这里，无论是数值的 1、2、3，还是字符的 A、B、C，都有固有大小或高低顺序，但数据之间却是不等距的，因为老年与中年、中年与青年之间的差距是不相等的。**定类型数据**【名义】指没有内在固有大小或高低顺序，一般以数值或字符表示的分类数据。例如：性别中的“男”和“女”，可以分别用 1 和 2 表示；民族变量中的各个民族，可以分别用“汉”“回”“彝”等字符表示。这些数值或字符都不存在内在固有的大小或高低顺序，而只是一种分类名义上的指代。在 SPSS 中可根据变量的具体含义指定其计量尺度属于上述哪种类型。

在 SPSS 数据编辑窗口的变量视图中，在【测量】下相应行的位置单击鼠标，可以在弹出的对话框中根据实际数据指定变量的计量尺度，如图 1-8 所示。

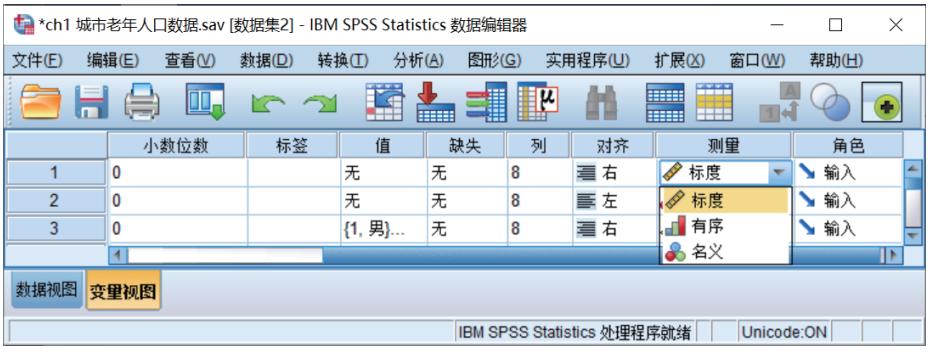


图 1-8　定义测量

图 1-9 是根据一份调查问卷定义好的 SPSS 数据结构。可以看出，各变量用代码表示，在对应的变量名标签【标签】中进行了说明，对应的变量值标签【值】也对性别、民族、受教育程度等定类型数据和定序型数据进行了说明。

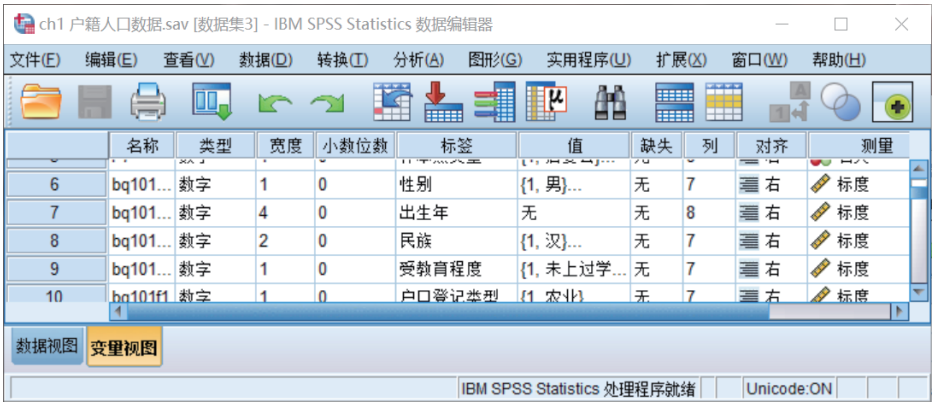


图 1-9　SPSS 数据结构示例

1.3.2　SPSS 数据的录入

在定义好 SPSS 数据结构之后，就可以录入数据了。SPSS 数据的录入操作在数据编辑窗口中的【数据视图】（图 1-1）中实现。其操作方法与 Excel 基本类似，可以通过直接向数据表中录入数据来生成 SPSS 格式的数据集。在实际应用中，可能还会对数据结构进行修改，边录入、边分析、边修改数据结构的情况也是较为常见的。

1. SPSS 数据的基本组织方式

如果待分析的数据是一些原始的调查问卷数据，或是一些基本的统计指标，这些数据就可按原始数据的方式组织。在原始数据的组织方式中，数据编辑器窗口中的一行称为一个“个案”（case）或“观测”，所有个案组成完整的 SPSS 数据。数据编辑器窗口中的一列称为一个“变量”。每个变量都有一个名字，称为变量名，是访问和分析 SPSS 变量的唯一标识。为了有一个更为直观的印象，下面以一份调查问卷数据的录入为例进行说明。

【例 1-1】为了解统计学课程的学习情况，我们进行了为期 8 年（2014.9—2021.7）的问卷调查，调查于每学期结束时进行，对象为修完统计学课程的本科生，问卷主要涉及学生基本信息、学生对统计学课程的认知情况、学生的课程学习情况、学生对课程学习的期待等方面，具体数据见文件“ch1 统计学课程问卷调查”。

例 1-1 的数据就是一份原始数据。在 SPSS 数据编辑器窗口中，一行存储一份问卷数据，是一个个案。对于本例，共有 1 001 人参与了问卷调查，在 SPSS 文件中就有 1 001 行数据，1 001 个个案。SPSS 中的一列通常对应一个问卷问题，是一个变量，每个变量都有变量名，变量名可以与问卷题目相对应。

图 1-10 是该份调查数据在 SPSS 数据编辑器窗口中数据视图的组织方式。

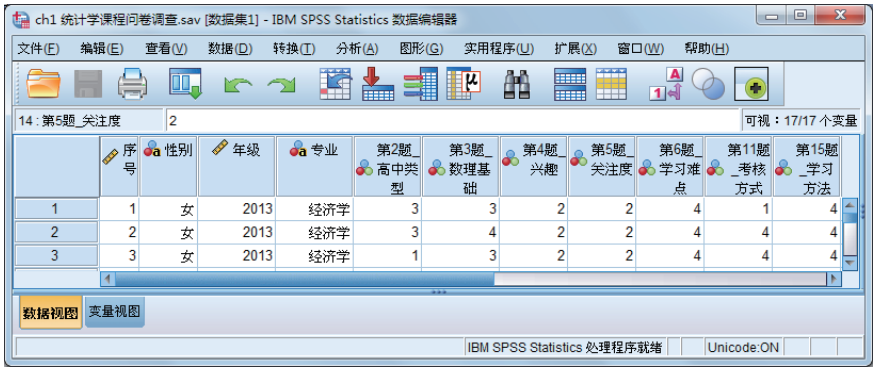


图 1-10　例 1-1 的数据组织方式

在数据录入中，也存在将经过分组汇总后的计数数据录入的情况（图 1-11），此时，数据编辑窗口中的一行是变量的一个分组，所有行包括了该变量的所有分组情况，数据编辑器窗口中的一列仍为一个变量，代表某个问题（或某个方面的特征）及相应的计数结果。

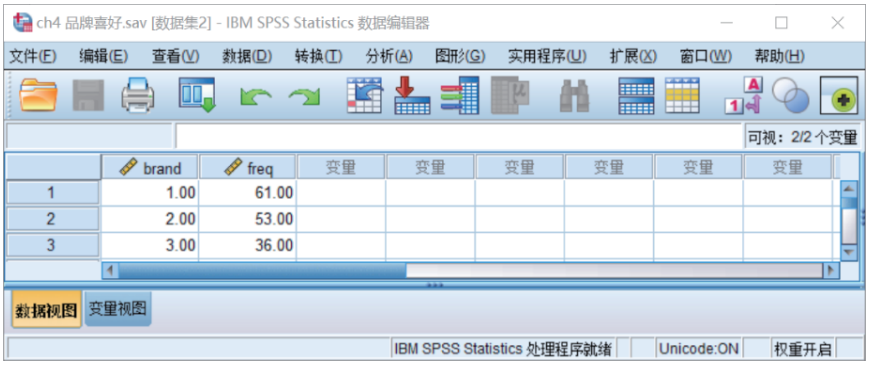


图 1-11　计数数据组织方式示例

选择什么样的数据组织方式主要取决于收集到的数据及所要进行的分析。

2. **从其他文件中导入数据**

在现实应用中，可能会将一批待分析的数据保存在其他软件中，如果希望用 SPSS 对这些数据进行统计分析，就需要将这些数据转换到 SPSS 中。SPSS 能够直接将它们读入数据编辑窗口，用户可再将其保存为 SPSS 格式文件。因此，读取其他格式的文件并将其转换为 SPSS 格式数据，是另外一种建立 SPSS 数据文件的方法。

SPSS 能够直接打开各种类型的数据文件，常见的格式有：

- Excel 格式文件，扩展名为 .xls；
- dBase 系列数据文件，扩展名为 .dbf；
- SAS 格式文件，扩展名为 .sas7bdat。

相应基本操作步骤如下。

第 1 步：选择【文件（F）】→【打开（O）】→【数据（D）】选项，在【查找位置】下拉列表框内查找需要文件所在的文件夹。

第 2 步：在【文件类型】下拉列表框中选择数据文件的类型，这处选 .xls 类型，之后会出现如图 1-12 的对话框。

图 1-12　将 Excel 数据导入 SPSS 文件

第 3 步：选择所需要的数据文件，然后单击【打开（O）】按钮，在【工作表（K）】中选择已准备好的数据表格，单击【确定】按钮。

注意：在使用 SPSS 打开 Excel 文件之前，需要先将所有的 Excel 文件关闭，否则，在打开过程中会出现错误；此外，因为每个 Excel 文件一般都包括多张表格，在选择时要注意打开所需要的文件，最好的方式是在 Excel 文件中对每个表格进行命名，这样，可读性会很强。

1.3.3　SPSS 数据文件的合并

当数据量较少时，一般可以按照上述方式建立 SPSS 数据文件。如果数据量较大，通常会把一份大的数据分成几个小的部分，由几个录入员分别录入，以加快数据录入速度，缩短录入时间。但会出现问题：一份完整的数据分别存储在了几个小的 SPSS 文件中，因此，如果要分析这些数据，就需要首先将若干个小的数据文件合并。SPSS 提供了两种合并数据文件的方式，分别是纵向合并和横向合并。

1. **纵向合并数据文件**

纵向合并数据文件就是将当前数据编辑器窗口中的数据与另一个 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当前数据编辑器窗口中数据的后面，依据两份数据文件中的变量名进行数据对接。

【例 1-2】有两份关于城市老年人口情况的抽样调查数据，分别如表 1-2 和表 1-3 所示。数据文件名分别为“ch1 城市老年人口数据”和“ch1 追加数据”，两份数据文件中的数据不同，且同一数据的变量名也不完全一致。现需要将这两份数据合并到一起。

表 1-2　城市老年人口数据

编　　号	调 查 地	性　　别	出 生 年 份	受教育程度	婚 姻 状 况
1	辽宁省	1	1952	3	1
2	江苏省	2	1946	3	5
3	辽宁省	2	1952	2	5
4	安徽省	2	1954	1	2
5	浙江省	1	1951	2	1
6	安徽省	2	1953	2	2
7	安徽省	2	1954	1	5
8	安徽省	2	1954	2	2
…	…	…	…	…	…
15	浙江省	1	1952	3	2

表 1-3　追加数据

编　　号	调 查 地	性　　别	出 生 年 份	受教育程度	家庭月支出 / 元
16	安徽省	2	1953	2	500
17	安徽省	2	1954	3	1 000
18	江苏省	2	1955	3	600
19	安徽省	1	1955	2	500
20	辽宁省	2	1947	2	500
14	辽宁省	2	1951	2	1 000

本例的合并实质是一个纵向合并，基本操作步骤如下。

第 1 步：在数据编辑器窗口中打开需要合并的 SPSS 数据文件：“城市老年人口数据”，选择菜单【数据（D）】→【合并文件（G）】→【添加个案（C）...】选项，会出现如图 1-13 所示的对话框。

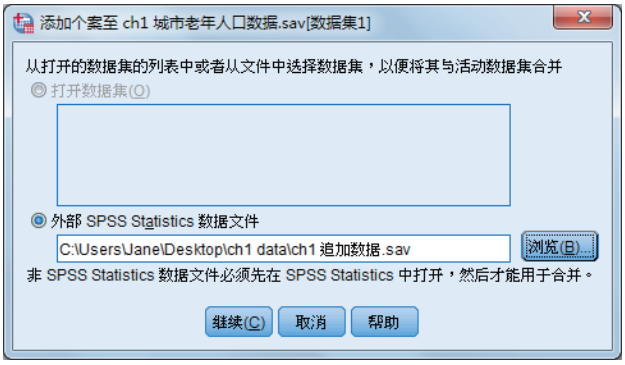


图 1-13　选择合并文件对话框

第 2 步：选择需要合并的另一个数据文件“追加数据”，单击【继续（C）】按钮，出现如图 1-14 所示的对话框。对话框中有两个文本框【非成对变量（U）】和【新的活动数据集中的变量（V）】。

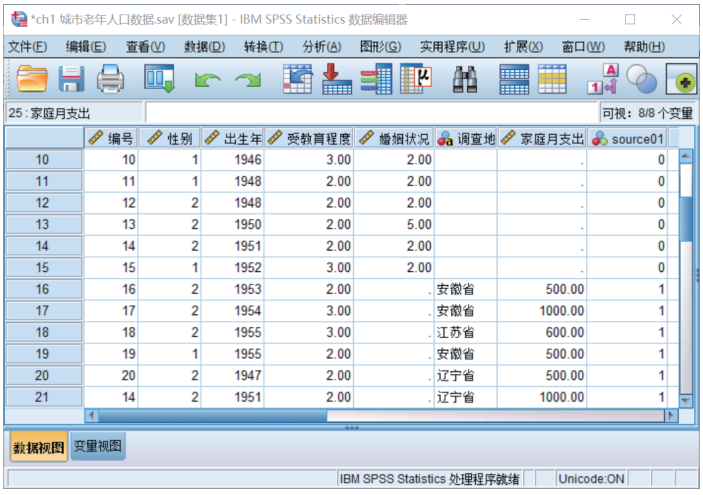
图 1-14　纵向合并数据文件对话框

【新的活动数据集中的变量（V）】文本框中显示的是两个数据文件中具有同名的变量，SPSS 默认它们有相同的数据含义，并将它们作为合并后数据文件中的变量。如果不接受这种默认，可以单击按钮将它们剔出到【非成对变量（U）】文本框中。

【非成对变量（U）】文本框中显示的两个变量“婚姻状况（*）”和“家庭月支出（+）”是两个文件中的不同变量名，带“（*）”的表示该变量是当前数据编辑器窗口中的变量，带“（+）”的表示该变量是“追加数据”文件中的变量。SPSS 默认这些变量的含义不同，且不放入合并后的新文件中。如果不接受这种默认，可选择这两个变量并将它们放入右边的【新的活动数据集中的变量（V）】文本框中。本例中，将这两个变量都放入右侧文本框中。

第 3 步：如果希望在合并后的数据文件中看出个案来自合并前的哪个 SPSS 数据文件，可以在图 1-14 所示的文本框中选中【指示个案源变量（I）】复选框。于是合并后的数据文件中将自动生成一个名为“source01”、取值为 0 或 1 的变量。0 表示个案来自当前数据编辑器窗口中的数据文件，1 表示来自其他数据文件。

第 4 步：单击【确定】按钮，部分结果如图 1-15 所示。



	编号	性别	出生年	受教育程度	婚姻状况	调查地	家庭月支出	source01
10	10	1	1946	3.00	2.00		.	0
11	11	1	1948	2.00	2.00		.	0
12	12	2	1948	2.00	2.00		.	0
13	13	2	1950	2.00	5.00		.	0
14	14	2	1951	2.00	2.00		.	0
15	15	1	1952	3.00	2.00		.	0
16	16	2	1953	2.00	.	安徽省	500.00	1
17	17	2	1954	3.00	.	安徽省	1000.00	1
18	18	2	1955	3.00	.	江苏省	600.00	1
19	19	1	1955	2.00	.	安徽省	500.00	1
20	20	2	1947	2.00	.	辽宁省	500.00	1
21	14	2	1951	2.00	.	辽宁省	1000.00	1

图 1-15　纵向合并后的数据文件

从合并后的数据文件中可以看到，两个数据文件中的重复个案会重复显示，如例中编号为 14 的样本。

通过上述操作，可以看出，为了方便 SPSS 数据文件的纵向合并，在不同数据文件中数据含义相同的数据项最好取相同的变量名，且数据类型也最好相同，这样将大大简化操作过程，有利于 SPSS 对变量的自动匹配。含义不同的数据项其变量名最好不要相同，否则会给数据合并过程带来许多麻烦。

2. 横向合并数据文件

横向合并数据文件就是将数据编辑器窗口中的数据与另一个 SPSS 数据文件中的数据进行左右对接，即将一个 SPSS 数据文件的内容拼到数据编辑器窗口中当前数据的右边，依据两个数据文件中的个案进行数据对接。

【例 1-3】有两份关于城市老年人口情况的抽样调查数据，分别如表 1-2、表 1-4 所示。文件名分别为“ch1 城市老年人口数据”和“ch1 追加数据－食品支出”，两份数据文件中的编号相同，第二份数据中只有部分样本的食品支出金额，现需要将这两份数据文件合并。

表 1-4　追加数据 - 食品支出

编　　号	家庭平均每月食品支出 / 元
1	80
2	100
3	150
4	150
5	200
6	200
7	200
8	200
9	300
10	300
70	500

本例是一个横向合并，基本操作步骤如下。

第 1 步：在数据编辑器窗口中打开需要合并的 SPSS 数据文件：“城市老年人口数据”，选择菜单【数据（D）】→【合并文件（G）】→【添加变量（V）...】选项，会出现与图 1-13 类似的对话框。

第 2 步：指定需要进行合并处理的 SPSS 数据文件名，如本例的“追加数据－食品支出”文件。单击【打开（O）】→【继续（C）】按钮，随后将显示如图 1-16 所示的对话框。在【合并方法】标签中默认选择【基于键值的一对一合并（N）】单选按钮，【键变量（K）】文本框中会自动将变量“编号”选入。

图 1-16　横向合并数据文件对话框（1）

第 3 步：选择【变量】标签，将显示如图 1-17 所示的选项卡。选项卡中有两个文本框：【排除的变量（E）】和【包含的变量（I）】。在图 1-17 中，两个待合并数据文件中除“编号”之外的所有变量名均显示在【包含的变量（I）】文本框中，SPSS 默认这些变量均以原有变量名进入合并后的数据文件中。其中，变量名后的“（*）”表示该变量是当前数据编辑器窗口中的变量，“（+）”表示该变量是被合并文件中的变量。用户如果不接受这种默认，可以选中该变量并按按钮将它们剔出到【排除的变量（E）】文本框中；或者剔出后单击【重命名（A）...】按钮将变量改名，然后再按按钮将它们从【排除的变量（E）】文本框中重新以新变量名选回到【包含的变量（I）】文本框中。

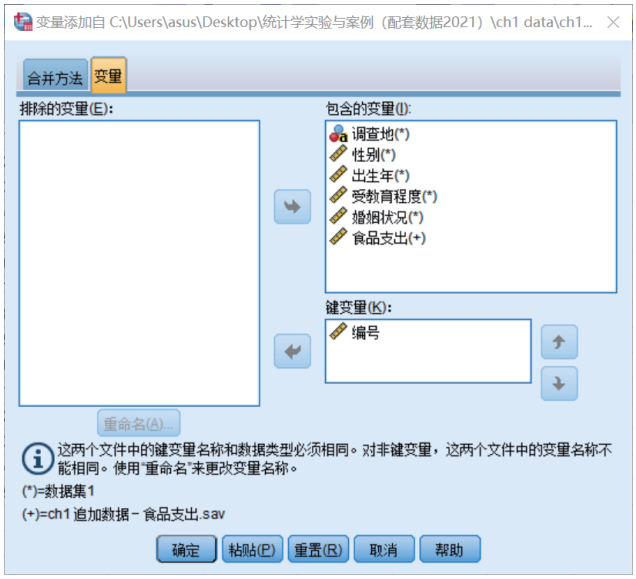


图 1-17　横向合并数据文件对话框（2）

第 4 步：单击【确定】按钮，数据编辑器窗口中会自动显示合并后的数据，用户可根据实际需要将它保存下来。

横向合并数据文件时，通常要注意以下三个问题：

第一，两个数据文件必须至少有一个名称相同的变量，该变量是两个数据文件横向拼接的依据，称为关键变量，如例 1-3 中的“编号”；

第二，两个数据文件都必须事先按关键变量值的升序排序；

第三，为方便 SPSS 数据文件的横向合并，不同数据文件中数据含义不同的数据项，变量名不应相同。

1.4　SPSS 数据的预处理

在数据文件建立好后，通常还需要对待分析的数据进行必要的预加工处理，这是数据分析过程中不可缺少的一个关键环节。而且随着数据分析的不断深入，对数据的加工处理还会多次反复，实现数据加工和数据分析的螺旋上升。

数据的预加工处理是服务于数据分析和建模的，需要解决的问题有很多。例如，缺失值和异常数据的处理、数据的转换、数据排序等。

SPSS 提供了一些专门的功能辅助用户实现数据的预加工处理工作，通过预处理还可以使用户对数据的总体分布有所了解。

1.4.1 数据的基本处理

1. 排序

排序是数据预处理中的常用方法之一，通过关键变量的排序，研究者可以对数据的特征有一个大致的了解。

【例 1-4】在统计学课程学习情况问卷调查中，为了将学生的课程学习情况及考试成绩结合起来进行分析，问卷采用记名方式（后 82 份问卷未记名，未加入成绩变量），在学生考试结束之后，再加入成绩变量。试以学习兴趣为主排序变量的降序，考试成绩为第二排序变量的升序进行多重排序。具体数据见文件“ch1 统计学课程问卷调查”。

SPSS 数据排序的基本操作步骤如下。

第 1 步：选择【数据（D）】→【个案排序（O）】选项。

第 2 步：指定主排序变量到【排序依据（S）】文本框中，并选择【排列顺序】文本框中的选项指出该变量按升序还是降序排序。

第 3 步：如果是多重排序，还要依次指定第二、第三排序变量及相应的排序规则。否则，本步可略。本例为多重排序，对话框如图 1-18 所示。

图 1-18　数据排序对话框

第 4 步：单击【确定】按钮，数据编辑窗口中的数据便自动按用户指定的顺序重新排列并显示。

注意：

第一，数据排序是整行数据排序，而不是只对某列变量排序。

第二，多重排序中指定排序变量的次序很关键。排序时先指定的变量优于后指定变量。多重排序可以在按某个变量升序（或降序）排序的同时再按其他变量值升序（或降序）排序。

第三，数据排序以后，原有数据的排列次序必然被打乱。因此，在时间序列数据中如果没有标示时间的变量（如年份、月份、季度等），则应注意保留数据的原始排列顺序，以免发生混乱。

2. **查找重复个案**

通常分析数据中不应出现与关键变量相同的个案。例如，在 1.2.3 节数据纵向合并后，14 号的数据出现了两次（关键变量是“编号”），这显然是不合理的。导致出现重复个案的主要原因可能是数据录入时的疏忽或缺乏必要的数据编码等。当数据量较大时，自动查找其中的重复个案是必要的。

SPSS 自动查找重复个案的主要方法是排序。它首先按照用户指定的关键变量对所有个案排序，之后，关键变量值相同的个案会被排在一起，便于识别。

【例 1-5】使用 1.2.3 节纵向合并后的数据进行重复个案查找。

第 1 步：选择【数据（D）】→【标识重复个案（U）】选项。

第 2 步：指定关键变量到【定义匹配个案的依据（D）】文本框中，这里指定为“编号”。指定对重复个案的排序变量到【匹配组内的排序依据（O）】文本框中，这里指定为“出生年（A）”，且默认对重复个案按升序排序，如图 1-19 所示。

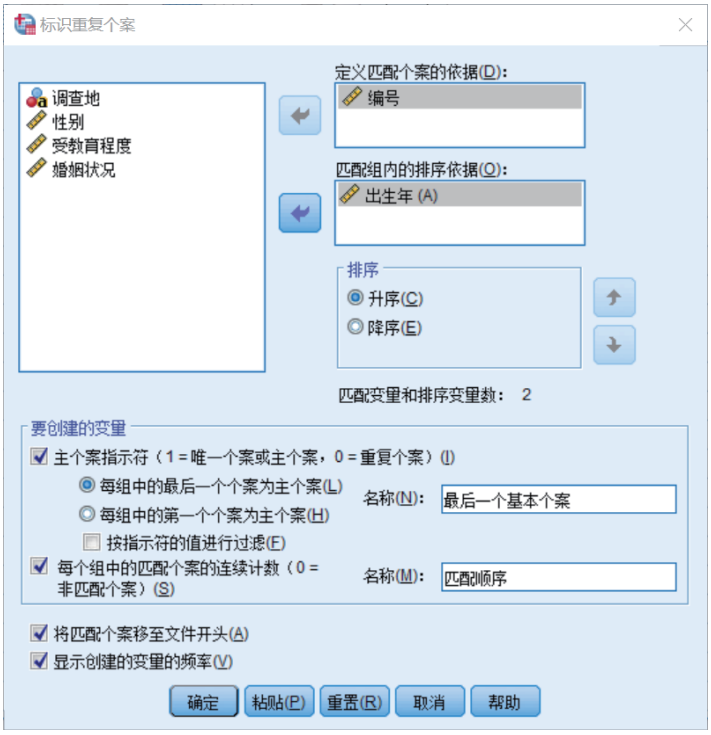


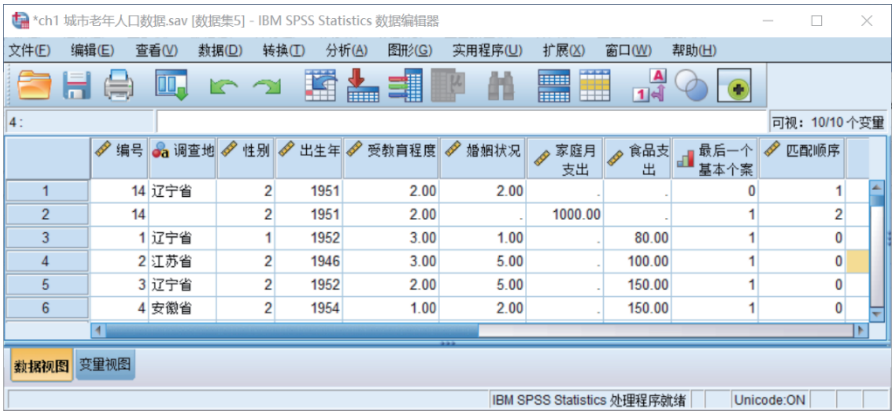
图 1-19　查找重复个案对话框

图中的【将匹配个案移至文件开头（A）】是 SPSS 的默认选项，表示将数据中的重复个案放置在文件开头。

SPSS 默认生成标识重复个案的变量。该变量名默认为“最后一个基本个案”。若选中【主个案指示符（1= 唯一个案或主个案，0= 重复个案）(I)】复选框，表示用 0 标识的个案为重复个案，1 为不重复个案（SPSS 称其为主个案）；单击【每组中的最后一个个案为主个案（L）】单选按钮表示对于具有相同关键变量值的重复个案指定变量的升序或降序排序后，排在最后的个案为主个案。若希望排在最前的个案为主个案，应单击【每组中的第一个个案为主个案（H）】单选按钮。

第 3 步：选中【每个组中的匹配个案的连续计数（0= 非匹配个案）(S)】复选框表示生成一个名为“匹配顺序”的变量，变量取 0 表示该个案为非重复个案，取 1，2，3…表示为第 1 个，第 2 个，第 3 个…重复个案。

第 4 步：单击【确定】按钮，查找重复个案的结果，如图 1-20 所示。



	编号	调查地	性别	出生年	受教育程度	婚姻状况	家庭月支出	食品支出	最后一个基本个案	匹配顺序
1	14	辽宁省	2	1951	2.00	2.00	.	.	0	1
2	14		2	1951	2.00	.	1000.00	.	1	2
3	1	辽宁省	1	1952	3.00	1.00	.	80.00	1	0
4	2	江苏省	2	1946	3.00	5.00	.	100.00	1	0
5	3	辽宁省	2	1952	2.00	5.00	.	150.00	1	0
6	4	安徽省	2	1954	1.00	2.00	.	150.00	1	0

图 1-20　查找重复个案结果

1.4.2　数据选取

数据选取在数据分析过程中很普遍，尤其是在大规模调查数据的分析中，根据研究目的选择合适的数据进行分析是非常普遍的，其目的在于服务于以后的数据分析。

1. **提高数据分析效率**

如果数据量较大，则会在一定程度上影响计算和建模的效率，因此，通常可以依据一定的抽样方法从总体中抽取少量样本，后面的分析只针对样本进行，这样会大大提高分析的效率。当然，抽取出的样本应具有总体代表性，否则分析的结论可能会有偏差。对于这个问题，统计学做了专门研究，一般可通过抽样方法来解决。

2. **检验模型**

在数据分析中，所建的模型是否能够较为完整准确地反映数据的特征，是否能够用于以后的数据预测，这些问题都是研究者极为关心的。为了验证模型，一般可依据一定的抽样方法只选择部分样本参与数据建模，剩余的数据用于模型检验。

SPSS 提供了几种数据选取方法，在此，仅介绍按指定条件选取和随机选取两种常用方法。

【例 1-6】按指定条件选取。文件“ch1 户籍人口数据”为来自北京、上海、大连、无锡、杭州、合肥、广州、贵阳 8 个城市的调查数据，要求选出来自上海的全部样本，并选

出年龄在 60 岁以上（含 60 岁）的老年人口。

第 1 步：选择【数据（D）】→【选择个案（S）】选项，如图 1-21 所示。

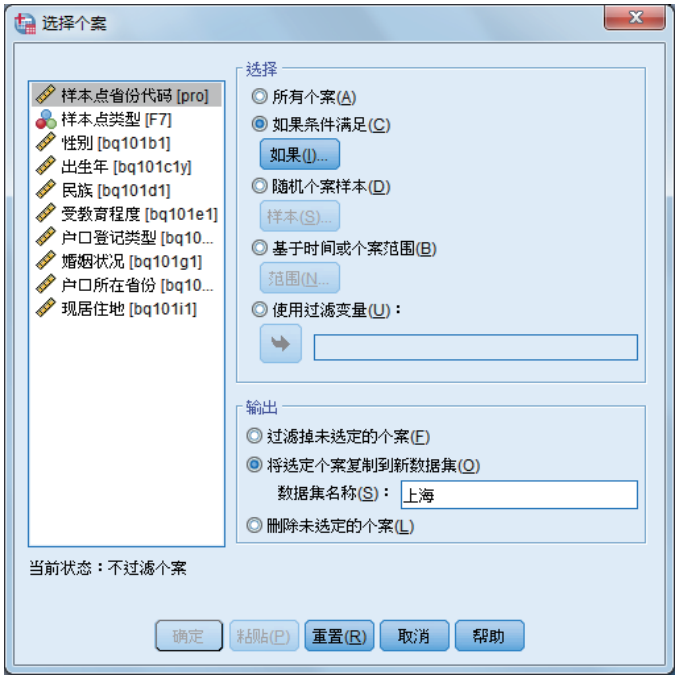


图 1-21　数据选取对话框

第 2 步：在【选择】选项区域中指定选取方法。【所有个案（A）】表示全部选中，本例单击【如果条件满足（C）】单选按钮，单击【如果（I）...】按钮，选择【样本点省份代码 [pro]】选项，在弹出的文本框中输入“pro=31”，如图 1-22 所示。

图 1-22　数据选取条件对话框（1）

第 3 步：在图 1-21 的【输出】选项区域指定输出方式。其中，【过滤掉未选定的个案（F）】表示在未被选中的个案号码上打一个“／”标记，表示暂时筛掉；【将选定个案复制到新数据集（O）】表示将选出的个案复制到一个新的数据编辑器窗口中，此时，应该在下面的【数据集名称（S）】文本框中输入数据集名称，本例选择此种处理方式（图 1-21）；【删除未选定的个案（L）】表示将未被选中的个案从数据编辑器窗口中删除。

第 4 步：单击【确定】按钮，系统会自动生成一个新的 SPSS 数据文件，部分样本数据如图 1-23 所示。

	pro	F1	F2	F3	F7	bq101b1	bq101c1y	bq101d1	bq101e1
1	31	上海市	市辖区	黄浦区	1	2	1924	1	5
2	31	上海市	市辖区	黄浦区	1	2	1951	1	4
3	31	上海市	市辖区	黄浦区	1	2	1955	1	4
4	31	上海市	市辖区	黄浦区	1	1	1920	1	4
5	31	上海市	市辖区	黄浦区	1	2	1938	1	3
6	31	上海市	市辖区	黄浦区	1	1	1951	1	3

图 1-23　数据选取产生的新数据文件

第 5 步：由于本次调查于 2015 年进行，要选出年龄在 60 岁以上（含 60 岁）的老年人口，则在新生成的文件中采用相同的方法，单击【如果（I）...】按钮，选择【出生年 [bq101c1y]】选项，在文本框中将变量出生年设定条件＜=1955，如图 1-24 所示。

图 1-24　数据选取条件对话框（2）

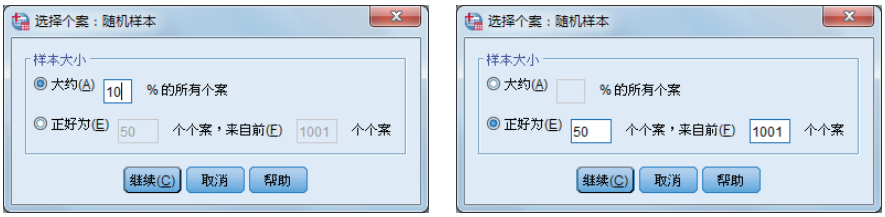
第 6 步：单击【确定】按钮，系统会自动生成另一个新的 SPSS 数据文件。

【例 1-7】随机选取。在“ch1 统计学课程问卷调查”数据文件中，先随机抽取 10% 的样本，再随机抽取 50 个样本。

在本例中，随机抽取 10% 的样本是近似选取。SPSS 将按照这个比例自动从数据编辑器窗口中随机抽取出相应百分比数目的个案。由于 SPSS 在样本选取方面的技术特点，抽取出的个案总数不一定恰好精确等于用户指定的百分比数目，会有小的偏差，因而称为近似选取。这种样本量上的偏差通常不会对数据分析产生重大影响。SPSS 的基本操作步骤如下。

第 1 步：选择【数据（D）】→【选择个案（S）】选项，同上例中图 1-21 所示。

第 2 步：在【选择】选项区域中单击【随机个案样本（D）】单选按钮，再单击【样本（S）...】按钮，出现如图 1-25 所示对话框。



（a）　　（b）

图 1-25　随机选取数据对话框

第 3 步：若随机抽取 10% 的样本，则在对应【大约（A）】后文本框中输入 10，若要随机抽取 50 个样本，则在对应【正好为（E）】后文本框中输入 50，1001（样本容量），如图 1-25 所示。

第 4 步：单击【继续（C）】按钮回到图 1-21 所示窗口，在【输出】选项区域中单击【过滤掉未选定的个案（F）】单选按钮，SPSS 会在数据编辑窗口自动生成一个名为“filter_$”的新变量，取值为 1 或 0，1 表示本条个案被选中，0 表示未被选中。

第 5 步：单击【确定】按钮。随机抽取 10% 样本的输出结果如图 1-26 所示，随机抽取 50 个样本的输出结果如图 1-27 所示。

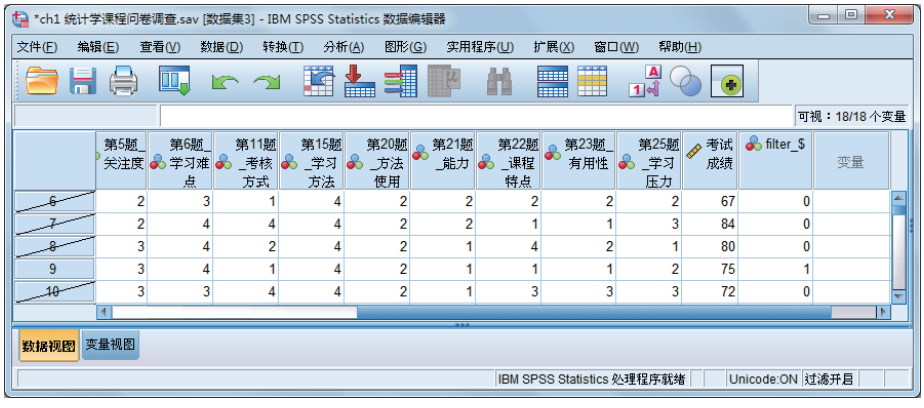


图 1-26　数据选取输出窗口（10% 样本）

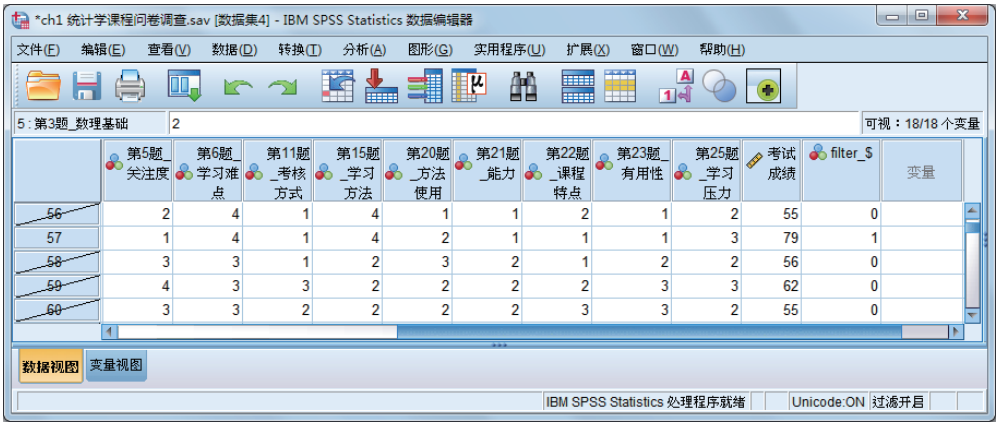


图 1-27　数据选取输出窗口（50 个样本）

1.4.3　变量计算

变量计算是数据分析过程中应用最广泛也是最重要的一环，通过变量计算可以处理许多问题。SPSS 变量计算是在原有数据基础上，根据用户给出的 SPSS 算术表达式及函数对所有个案或满足条件的部分个案计算，产生一系列新变量。

需要**注意**的是：

第一，SPSS 变量计算是针对所有个案（或指定的部分个案），每条个案（或指定的部分个案）都有自己的计算结果；

第二，变量计算的结果应保存到一个指定的变量中，该变量的数据类型应与计算结果的数据类型相一致。

【例 1-8】根据“ch1 城市老年人口数据”中的出生年，计算男性的年龄。

SPSS 的基本操作步骤如下。

第 1 步：选择【转换（T）】→【计算变量（C）】选项。

第 2 步：在【目标变量（T）】文本框中定义新计算的变量名，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，用户可以根据需要单击【类型和标签（L）...】按钮进行修改，还可以给新变量加变量名标签。这里，输入新变量“年龄”。

第 3 步：在【数字表达式（E）】文本框中给出 SPSS 算术表达式和函数。可以手工输入，也可以单击文本框下的按钮完成算术表达式和函数的输入工作。SPSS 将所有函数划分成若干类别，显示在对话框右侧中间，各类别所包含的函数名列在右侧下方。鼠标选中一个函数后，该函数的说明信息会在对话框中间下方的位置显示。这里，直接输入函数表达式（图 1-28）。

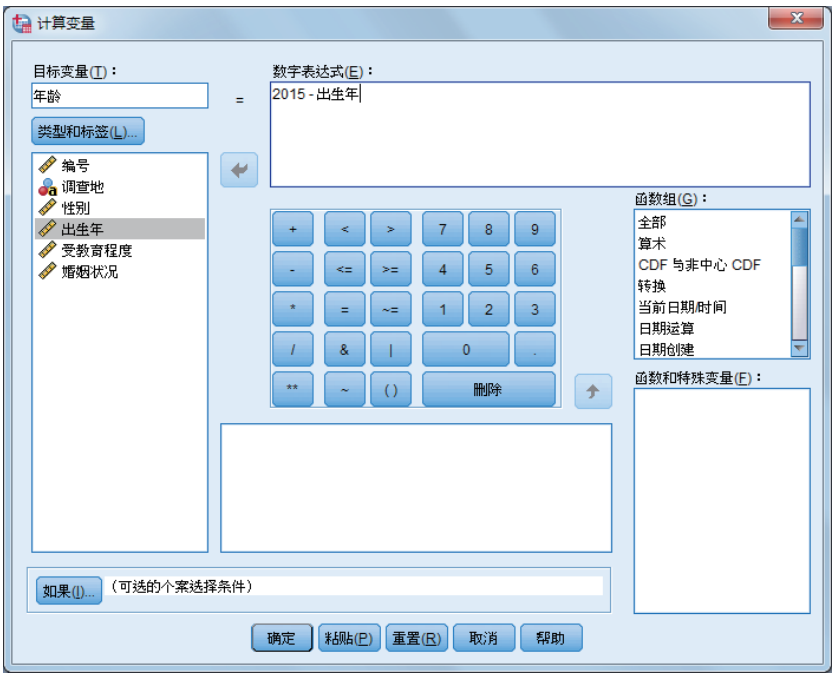


图 1-28 【计算变量】对话框

第 4 步：如果仅希望对符合一定条件的个案计算产生变量，则单击【如果（I）...】按钮，出现如图 1-29 所示对话框。单击【在个案满足条件时包括（F）】单选按钮，然后在文本框中输入条件表达式。否则，本步略去。本例需要计算的是男性的年龄，因此，在文本框中输入“性别 =1”。

图 1-29　条件表达式输入对话框

第 5 步：单击【继续（C）】按钮回到主对话框。

第 6 步：单击【确定】按钮，计算结果如图 1-30 所示。

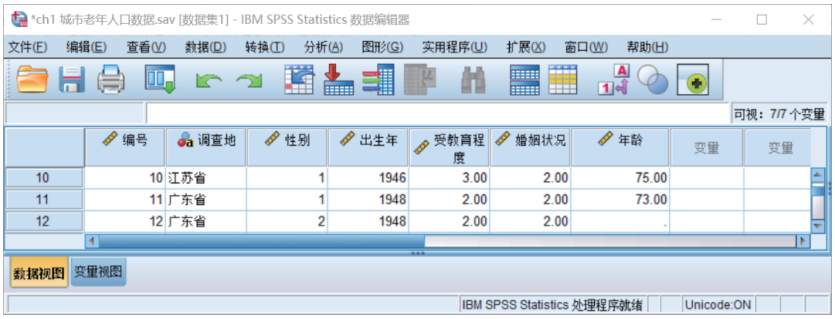


图 1-30　数据计算输出窗口

练　习　题

● 概念辨析

1. 指出下面的变量中哪一个属于无序类别变量。（　　）

A. 年龄　　B. 工资

C. 汽车产量　　D. 购买商品时的支付方式

2. 对高中生的一项抽样调查表明，85% 的高中生愿意接受大学教育。这一叙述是（　　）的结果。

A. 定性变量　　B. 试验　　C. 描述统计　　D. 推断统计

3. 指出下面的变量中哪一个属于定序型变量。（　　）

A. 企业的收入　　B. 员工的工资

C. 员工对企业某项改革措施的态度　　D. 汽车产量

4. 指出下面的变量中哪一个属于数值型变量。（　　）

A. 生活费支出　　B. 产品的等级

C. 企业类型　　D. 员工对企业某项改革措施的态度

5. 一家研究机构从 IT 从业者中随机抽取 500 人作为样本进行调查，其中 60% 的人回答他们的月收入在 5 000 元以上，50% 的人回答他们的消费支付方式是用信用卡。这里的 500 人是（　　）。

A. 总体　　B. 样本　　C. 变量　　D. 统计量

6. 下列不属于描述统计问题的是（　　）。

A. 根据样本信息对总体进行的推断　　B. 了解数据分布的特征

C. 分析感兴趣的总体特征　　D. 利用图表等对数据进行汇总和分析

7. 从含有 N 个元素的总体中，抽取 n 个元素作为样本，使得总体中的每一个元素都有相同的机会（概率）被抽中，这样的抽样方式被称为（　　）。

A. 简单随机抽样　　B. 分层抽样　　C. 系统抽样　　D. 整群抽样

8. 为了解某学校学生的购书费用支出，从男生中抽取 60 名学生调查，从女生中抽取 40 名学生调查，这种调查方法是（　　）。

A. 简单随机抽样　　B. 系统抽样　　C. 分层抽样　　D. 整群抽样

9. 指出下面的变量中哪一个属于定类型变量。（　　）

A. 考试成绩　　B. 民族

C. 受教育程度　　D. 身高

10. 为了解某学院学生的生活费用支出，从全院 30 个班级中抽取 6 个班级学生调查，这种调查方法是（　　）。

A. 简单随机抽样　　B. 系统抽样

C. 分层抽样　　D. 整群抽样

● 上机练习

1. 数据录入

根据自己的个人信息，按照例 1-1 中问卷的部分内容录入数据，要求定义变量名、变量类型、小数位宽、变量名标签（如有必要）、变量值标签和计量尺度。其中：性别按 0/1 变量定义，6~10 题按选项定义变量值标签，注意第 10 题多项选择题的数据录入方法。

问卷内容如下。

第一部分：个人基本信息

（1）年级：_____

（2）专业：_____

（3）性别：_____

（4）你在哪个省（自治区、直辖市）参加高考：_____

（5）你高中学的是：_____　①文科　②理科

第二部分：课程认识及学习情况

（6）你认为你的数理基础：_____

①很好　②好　③一般　④较差

（7）你对统计学的兴趣：_____

①很有兴趣　②有兴趣　③一般　④没兴趣

（8）你认为学习统计学的难点在于：_____

①概念　②原理　③公式和计算　④统计思想

（9）你认为想要学习这门课程，关键靠什么：_____

①逻辑思维能力　②综合能力　③计算能力　④记忆能力

（10）通过统计学课程的学习，你认为自己掌握了哪些能力（可多选）：_____

①统计调查能力　②搜集数据能力　③数据整理能力　④数据分析能力

（注意：多选题中有几个选项就要设置几个 0/1 变量）

2. 定义变量

试录入表 1-5 中的数据文件，并按要求进行变量定义。具体要求如下。

表 1-5　学生信息

学　号	姓　名	性别	生　　日	身高 /cm	体重 /kg	英语成绩	数学成绩	生活费 / 元
200201	刘一迪	男	1982.01.12	156.42	47.54	75	79	345.00
200202	许兆辉	男	1982.06.05	155.73	37.83	78	76	435.00
200203	王鸿屿	男	1982.05.17	144.6	38.66	65	88	643.50
200204	江飞	男	1982.08.31	161.5	41.68	11	82	235.50
200205	袁翼鹏	男	1982.09.17	161.3	43.36	82	77	867.00
200206	段燕	女	1982.12.21	158	47.35	81	74	
200207	安剑萍	女	1982.10.18	161.5	47.44	77	69	1233.00

（1）变量名同表格名，以“/”后的内容作为变量标签。对性别设值标签“男 =0；女 =1”。

（2）正确设定变量类型。将学号设为数值型；日期型统一用“mm/dd/yyyy”型；生活费用货币型。

（3）变量值宽统一为 10，身高与体重、生活费的小数位为 2，其余为 0。

3. 数据处理

设计一个研究问题，选择相应变量，再采用合适的方法，对“上机作业 1.3- 合肥市调查数据”中的数据进行处理。