第二章　数据录入与数据获取

数据是统计研究的基础，没有数据，分析也就无从谈起。在 SPSS 中建立数据文件大致有两种情况：一种是非电子化的原始数据资料，需要直接将调查问卷中的数据录入进 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。

针对上述两种情况，这一讲将主要介绍两个问题，即如何将数据录入进 SPSS 中，以及如何将其它格式的数据读进 SPSS 中。对于第一个问题，根据问题类型的不同，将会介绍开放题、单选题和多选题的录入方式；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。

2.1　CCSS 案例项目背景介绍

为使本书内容更贴近实战，全书将尽量使用中国消费者信心调研项目的数据作为教学案例，通过该项目数据的实际运用对 SPSS 的各项功能进行讲解。本节将首先对该项目的背景做一介绍，以利读者的后续阅读。

2.1.1　项目背景

消费者信心是指消费者根据国家或地区的经济发展形势，对就业、收入、物价、利率等问题的综合判断后得出的一种看法和预期，消费者信心指数则是对消费者整体所表现出来的信心程度及其变动的一种测度。消费者信心指数的概念和方法最早是由美国密歇根大学调查研究中心的乔治 卡通纳在上世纪 40 年代后期提出，随后在美联储的委托之下开展了相应调研直至今日。六十余年的历史已经证明了这一指标体系在预测未来宏观经济走向方面具有不可替代的价值，目前已成为各市场经济国家非常重要的经济风向标之一。

联恒市场研究看到了这一指标体系潜在的市场价值，于 2007 年启动了中国消费者信心调研（CCSS）项目，这一项目是联恒与美国密歇根大学社会研究所消费者信心调查课题组负责人 Richard Curtin 博士共同设计开发完成，整个方法体系与密歇根大学的消费者信心调查基本相同，同时也根据中国的具体国情进行了补充和完善，使之更贴近中国的实际情况。

CCSS 的调查始于 2007 年 4 月，每月在东部与中西部 30 个具有代表性的中国城市中抽取 1,000 个左右的家庭，通过电脑辅助电话访问(CATI)取得，目前已累计了三年多近四

万个样本的历史的数据。为化繁为简，这里我们将只截取北京、上海、广州三个城市在 2007 年 4 月、2007 年 12 月、2008 年 12 月和 2009 年 12 月共 1147 个样本用于随后的讲解，具体数据参见文件 CCSS_Sample.sav。

CCSS 现已成为德意志证券交易所集团旗下产品，本书所涉及的只是完整历史数据库的一小部分，且出于产品保密需要，在数据文件中删除了对指数计算至关重要的权重值，因此分析结果仅用于案例教学，所计算出的指数值会和真实指数值有一定偏差，不代表真实情况。

2.1.2 项目问卷

CCSS 项目的问卷是标准化的，每月固定执行。由于问卷内容较长，我们选择了其中部分题目作为教学案例，具体如下（注意：为了便于讲解，下列题目顺序和内容均进行过调整，并非访问时的原始状况）：

中国消费者信心指数研究问卷

S0　受访者所在城市：

100 北京　200 上海　300 广州

S1　请问您贵姓是？___

S2　记录被访者性别：

1 男性　2 女性

S3　请问您的十足年龄是？___

S4　请问您的学历是？

1 初中/技校或以下　2 高中/中专　3 大专　4 本科　5 硕士或以上

S5　请问您的职业是？

1 企/事业管理人员　2 工人/体力工作者（蓝领）　3 公司普通职员（白领）

4 国家公务员　5 个体经营者/私营业主　6 教师

7 学生　8 专业人士（医生、律师等）　9 无/待/失业、家庭主妇

10 退休 11 其他职业

S7 请问您的婚姻状况是？

1 已婚 2 未婚 3 离异/分居/丧偶

S9 请问您的家庭月收入（包括工资、奖金和各种外快收入）大约在什么范围呢？

1 999元或以下 2 1000-1499元 3 1500-1999元

4 2000-2999元 5 3000-3999元 6 4000-4999元

7 5000-5999元 8 6000-7999元 9 8000-9999元

10 10000-14999元 11 15000-19999元 12 20000-29999元

13 30000以上 98 无收入 99 拒答

C0 请问您的家庭目前有下列还贷支出吗？

C0_1 房贷 1 有 2 无 99 拒答

C0_2 车贷 1 有 2 无 99 拒答

C0_3 其他一般消费还贷 1 有 2 无 99 拒答

O1 请问您家里有家用轿车吗?

1 有 2 没有

A3 首先，请问与一年前相比，您的家庭现在的经济状况怎么样呢？是变好、基本不变还是变差？

1 明显好转 2 略有好转 3 基本不变 4 略有变差 5 明显变差 9 说不清/拒答

A3a 为什么您这样说呢？（最多限选两项）___

0 中性原因 90 不知道/拒答

10 改善：收入相关 110 恶化：收入相关

20 改善：就业状况相关 120 恶化：就业状况相关

30 改善：投资相关 130 恶化：投资相关

40 改善：家庭开支相关　140 恶化：家庭开支相关

50 改善：政策/宏观经济　150 恶化：政策/宏观经济相关

A4　那么与现在相比，您觉得一年以后您的家庭经济状况将会如何变化？

1 明显好转　2 略有好转　3 基本不变　4 略有变差　5 明显变差　9 说不清/拒答

A8　那么与现在相比，您认为一年以后本地区的经济发展状况将会如何？

1 非常好　2 比较好　3 保持现状　4 比较差　5 非常差　9 说不清/拒答

A9　您认为一年之后本地区的就业状况将会如何变化？

1 明显改善　2 略有改善　3 保持现状　4 略有变差　5 明显变差　9 说不清/拒答

A10　那么与现在相比，您认为五年之后，本地区的经济将会出现怎样的变化？

1 明显繁荣　2 略有改善　3 保持现状　4 略有衰退　5 明显衰退　9 说不清/拒答

A16　对于大宗耐用消费品的购买，如家用电器，家用电脑，以及高档家具之类的，您认为当前是购买的好时机吗？

1 很好的时机　2 较好时机　3 很难说，看具体情况而定　4 较差时机　5 很差的时机　9 不知道/拒答

2.2 数据格式概述

2.2.1 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需要遵守相应的格式要求，其基本原则如下：

- ✧ 不同个案（Case）的数据不能在同一条记录中出现，即同一个案的数据应当独占一行。
- ✧ 每一个测量指标/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。

但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“一个个案占一行，一个变量占一列”的原则，这种情况在配对数据和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某个观察指标的不同次测量看成是不同的指标，

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。

2.2.2 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进一步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（Decimals）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十一种变量属性，但这里将重点介绍变量类型和测量尺度这两个属性，对于其它的一些属性，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。

变量的存储类型

SPSS 中的变量有三种基本类型，分别是：数值型、字符串和日期型。根据不同的显示方式，数值型又被细分为了五种（在 20 版中则分为六种），所以 SPSS 中的变量类型共有八种（在 20 版中则为九种）。在变量视图中选择“类型”单元格时，右侧会出现形如…的省略号按钮，单击…会弹出变量类型对话框，如图 2.1 所示。左侧为具体的存储类型，右侧则用于进一步定义变量宽度、小数位数等。

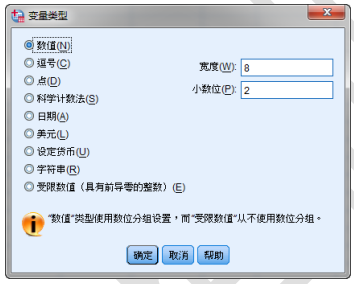

图 2.1 变量类型对话框

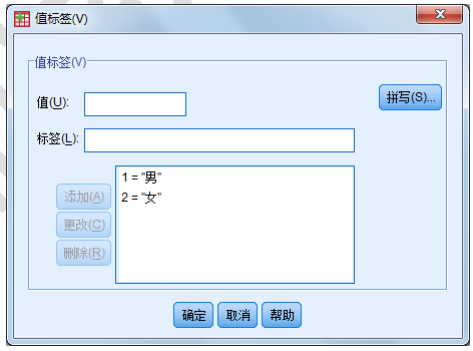

图 2.2 变量值标签对话框

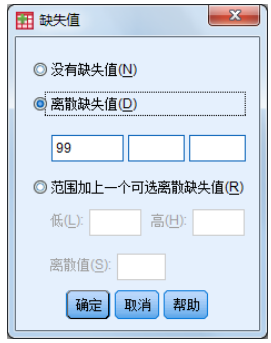

图 2.3 缺失值对话框

1．数值型（Numeric）

在以上三大类变量类型中，数值型是 SPSS 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）、每三位用逗号分隔的逗号数值型（Comma）、每三位用圆点分隔的圆

点数值型（Dot）、科学计数型（Scientific Notation）、显示时带美元符号的美元数值型（Dollar）、用户自定义型（Custom Currency）这六种不同的表示方法。实际上上述方式只有标准数值型最为常用，其余几种方式的详情读者有兴趣的话可以直接查阅软件帮助，这里不再赘述。

2．字符型（String）

字符型也是 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。

3．日期型（Date）

该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 在对话框右侧会以列表框的方式列出各种显示格式以供用户选择。如果此处选择 mm/dd/yy 或类似的两位数年份记录方式，则需要在系统选项的“数据”选项卡中确定具体的世纪范围，目前系统默认为 1941-2040 年区间。

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=86400，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。

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 CCSS 数据中的以下几个变量：

- ✧ 变量 S2“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 S4“学历”：用 1 表示“初中”，2 表示“高中”，3 表示“本科”等，1 和 2 虽然也是符号，但这里有一个顺序之分了，1 就是比 2 的学历低。但是究竟低多少？本科和高中的差距更大，还是高中和初中的差距更大？不知道，各级别之间的差距大小无法衡量，更无法进行比较。
- ✧ 变量 S3“年龄”：20 和 21 就是有区别的，差多少呢？差 1！而且这个差距大小，和 39 与 40 之间的差距是相等的，都是 1，也都等于 50 和 55 之间差距的 1/5！

由上可知，上述三个变量的存储类型同样都是数值型，但数值的具体含义不同，所携带的信息量不同，适用的统计方法也就不同。如果只以存储类型来说明这个变量的属性的话，就不能反映上述区别。为此，就有必要给变量增加测量尺度这一属性。

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四个层次：定类尺度、定序尺度、定距尺度和定比尺度。在这四种测量尺度之间，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。但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。

1．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，不能比较各类之间的大小，所以各类之间没有顺序或等级，如变量 S0“城市”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，如在所有个案中，北京有多少人，占总人数的百分率是多少等。对于 S2“性别”这种两分类变量，一般人们仍然将其归为定类尺度变量。但是两分类变量较为特殊，即使将其归为其他类型，一般也不会影响后续分析。

在 SPSS 中使用度量标准（Measure）属性对变量的测量尺度进行定义，其中定类尺度变量用“名义(N)”来表示。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”。

2．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。CCSS 数据中的变量 S4“学历”就是一个典型的定序变量。

在 SPSS 的度量标准属性框中，定序尺度变量用“序号(O)”来表示。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。

3．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，生活中最典型的定居尺度变量就是温度。

4．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”就真正的表示“没有”。比如温度，0℃只是一个普通的温度（水的冰点），并非没有温度，因此它只是定距变量，而重量则是真正的定比变量，0KG 就意味着没有重量可言。上文中提到的变量 S2“年龄”就是一个典型的定比变量。

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，而定距变量严格来说只可进行加减运算。

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“度量(S)”。

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。

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，比如说，标签（Label）属性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，值（Values）属性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。

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 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”，最后按 OK，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。如果现在就想看显示效果，请切换回数据视图，然后选择菜单视图➔值标签，怎么样，看到了吗？

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。

缺失值

缺失（Missing）属性是一个重要而且容易被忽视的变量属性，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/拒答。相应的代码可能用 9 或者 99 来表示，例如 CCSS 项目中的 S9“家庭月收入”中就是以 99 来表示拒答。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缺失属性框右侧的  ，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“没有缺失值”；第二项指定离散的缺失值（Discrete missing values），最多可以定义 3 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。

角色

该属性是较新的 SPSS 版本中新增的，实际上来源于数据挖掘方法体系的要求，某些对话框支持可用于预先选择分析变量的预定义角色。当打开其中一个对话框时，满足角色要求的变量将自动显示在目标列表中。可用角色包括：

- ✧ 输入：变量将用作输入（例如，预测变量、自变量）。
- ✧ 目标：变量将用作输出或目标（例如，因变量）。
- ✧ 两者：变量将同时用作输入和输出。
- ✧ 无：变量没有角色分配（将不纳入分析）。
- ✧ 分区：变量将用于将数据划分为单独的训练、检验和验证样本。
- ✧ 拆分：该项的存在主要是为了能够和 Clementine（即现在的 IBM SPSS Modeler）相互兼容。具有此角色的变量不会在 SPSS 中用作拆分文件变量。

缺省情况下，SPSS 将为所有变量分配输入角色，需要指出的是，角色分配只影响支持角色分配的对话框。而此类对话框在现有版本的 SPSS 中仅是凤毛麟角，因此一般用户可以直接无视这一属性。

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，这里就不浪费各位的时间了。但是有一点要强调的是，就数据录入这部分内容来说，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到对错，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，所以读者们不可小看这部分工作。

2.3 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如图所示，用户只要按自己的需要在其中定义变量、输入数据，然后保存，就一切 OK 了。

对于这个空数据文件，恐怕我们还要再多说两句，请大家注意窗口左上角的文字是“未标题 1[数据集 0]”，其含义是说该数据暂时未被存储为数据文件，所以没有文件名称（未标题）；但是 SPSS 系统内部在使用该数据文件时，将会按照“数据集 0”这个名称来标识该文件，这就是所谓的工作名称。

2.3.1 操作界面说明

数据窗口是一个典型的 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。请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。

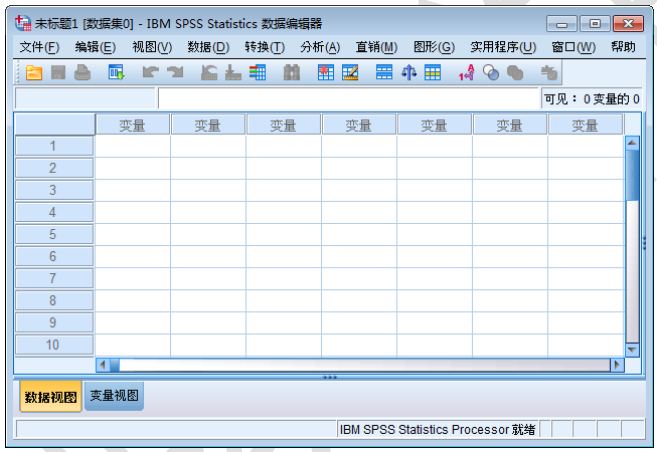

图 2.4 SPSS 的数据编辑窗口

在界面的左下角，可以看到“数据视图”和“变量视图”的标签，现在图中显示的是数据视图，如果点击右边的“变量视图”，则会切换入变量视图。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。

初学者往往会关心 SPSS 的数据容量问题，实际上，作为一个功能完善的统计软件，只要相应机器的内存和硬盘足够大，SPSS 理论上可以加载的变量数和案例数是无限大的，至少笔者亲自处理过的数据文件变量数最多上千，案例数最多达到百万条的级别。而且在此数据量时，SPSS 用现在的主流硬件配置完成常用的统计分析工作耗时也是非常少的。

2.3.2 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，以 CCSS 问卷为例，可以发现在这份问卷中，ID 是数值型开放题，S1“姓名”是字符型开放题，S2“性别”是单选题，C0 系列、A3a 系列均为开放题。下文将分别就这几种类型题目的录入方式加以介绍。

在 SPSS 中定义变量

前边已经说过，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，变量名是变量的惟一标识，前边已经讨论过相关的知识，这里就不再重复，在前三行的名称属性列中直接输入变量名——ID、S1、S2，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。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作修改。
- ✧ 变量 S1 是被访者姓名，应是字符型变量，这里应当将“类型”中的“数值”改成“字符串”，并在必要的时候放大默认的 8 位宽度以满足需要，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宽度，只要最大不超过 256 个字符即可。

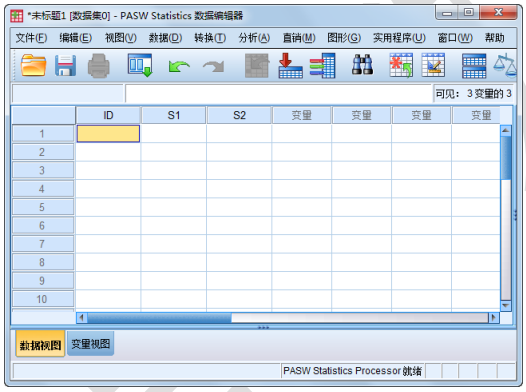

图 2.5 变量定义

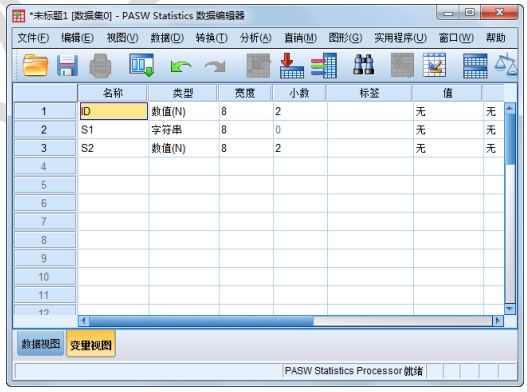

图 2.6 定义好变量的数据编辑窗口

现在切换回数据视图，数据编辑窗口如三张图中的第一张图所示。可见前 4 列的名称均为深色显示，就是刚才定义的内容，表明这四列已经被定义为变量，其余各列的名称仍为灰色的“变量”，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。

开放题的录入

现在开始录入数据，首先来输入变量 ID 的值，请确认一行一列单元格为当前单元格，弃鼠标而用键盘，输入第一个数据 1，此时界面显示如第二张图所示：

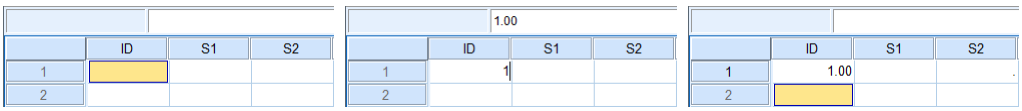

图 2.7 录入数据过程

请注意：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，界面如第三张图所示。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（num）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）因为没有输入过数据，显示为空，一行三列单元格（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。

如果要继续录入，则用类似的输入方式将数据录入完毕即可，但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，辛苦工作的成果付之东流，心痛啊！

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，而且还方便了后边的数据分析工作，何乐而不为呢？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”定义后的界面参见下图：

	名称	类型	宽度	小数	标签	值
1	S2_1	字符串	8	0	性别	无
2	S2_2	字符串	8	0	性别	{1, 男}...
3	S2_3	数值(N)	8	2	性别	{1.00, 男}...

图 2.8　单选题的三种录入方式说明

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

半开放题的录入

半开放题指的是问卷数据中有含“其它，请指出”选项的单选题，此类题目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，“其它，请指出”作为选项中的一个可进行选择；第二个变量将“其它，请指出”的具体内容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。

为使得变量名之间具有一定的逻辑联系，可以考虑第二个变量的名称设置为由第一个变量名称后直接加“a”之类的方式，另外在数据录入完毕后，可能会在数据预处理阶段对第二个变量中的数据进行编码处理，以利后续分析，在 SPSS 中的相关功能可参见第三、四章的相应讲解，此处不赘。

2.3.3　多选题的录入

多选题，又被称为多重响应（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。而多选题，比如 CCSS 项目中的 C0 和 A3a 题目均为多选题，被访者可以选择一个选项，也可以选择两个或者多个。这样一来，由于在多选题中每道题都可能有一个以上的答案，多选题就不能用一个变量来直接编码（否则无法进行分析），而需使用几个变量来进行记录。统计软件中对多选题的常见的方法有两种：多重二分法（Multiple Dichotomy Method）和多重分类法（Multiple Category Method）。下文将进行详细说明。

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，这些变量各自代表对其中一个选项的选择结果，一般均为二分类，而其中必然有一个类别代表选中了这一选项。

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，如 CCSS 项目中的 C0 题目，对应所需选择的三种选项，分别设定了 C0_1、C0_2、C0_3 这三个变量，且均以 1 表示选中，2 表示未选中，如图 2.9 所示，可见图中第二个个案每月有房贷支出，但没有车贷和其他消费还贷支出。而第三个个案则每月只有其他消费还贷支出。

显然，在多重二分法中无论有多少个变量，其变量值标签的定义应该一致，否则将会引起混乱。还有一点要说明的是，C0 题目中我们还增设了代码 99 代表拒答，这主要是根据访问的实际需求增设的，后续分析中可以将 99 和 2 合并成一类，即未选中该选项来进行分析。

c0_1	c0_2	c0_3
2	2	2
1	2	2
2	2	1
1	1	2
2	2	2
2	2	2
2	2	2

图 2.9　多重二分法数据录入格式

a3a_1	a3a_2
120	140
0	140
0	.
130	.
30	.
0	0
90	.

图 2.10　多重分类法的数据格式

多重分类法

多重二分法实际上是多选题的标准数据格式，但这种数据格式有的时候也会给数据录入带来麻烦，以 CCSS 项目中的 A3a 题目为例，每个受访者被限制只能回答最多两个选项，但总选项数量多达 12 个，显然，如果使用多重二分法录入，则大部分数据都需要录入为“未选中”，徒增许多数据录入的工作。对于此类多选题，则使用多重分类法进行记录更为便捷。

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。

多重分类法适合于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。以 A3a 为例，由于限定最多回答两个选项，因此只需要设定 A3a_1 和 A3a_2 两个变量即可，图中可见个案一选择了 120 和 140 两个选项，而个案四只回答了 130 这一个选项，随后的 A3a_2 则为缺失值。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。

设定多选题变量集

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 SPSS 只会默认他们是若干个分散的变量，并不明白它们代表的是一道多选题，只有将其设定为多选题变量集（也称多重响应集），SPSS 才能对其正确识别，从而将多选题的全部变量当作一整道题目来进行分析。

在 SPSS 中提供了专门的菜单用来对付多选题，Tables 模块和多重响应（Multiple Response）菜单都可以用来设定多选题变量集。所不同的是，多重响应菜单中的定义变量集（Define Sets）项定义的多选题变量集信息不能在 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 模块可以保存所定义的信息。所幸的是这两个过程的操作基本相同，现在就以 Define Sets 过程为例来看一下是如何定义多选题集的。

图 2.11　定义多选题变量集

在 SPSS 中选择分析➔多重响应➔定义变量集，打开定义多选题集的对话框，界面如图 2.11。在该对话框中，需要注意的有这样几个地方：

- ✧ 集合中的变量（Variables in Sets）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。
- ✧ 将变量编码为（Variables Are Coded As）单选框组：选择变量的编码方式。在多重二分法方式中，需要在右侧的计数值框中指定是指用哪个数值表示选中。在多重分类法方式中则此时需要设定取值范围，在该范围内的记录值将纳入分析，注意在 Tables 模块中是不需要设定取值范围的。
- ✧ 名称（Name）框：键入多选题变量集的名称，在此定义的变量集名为 C0，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。

所有设定均完成后单击右侧的“添加”钮，相应的多选题变量集设定就会被加入最右侧的“多响应集”列表了。

半开放多选题的处理方式

对于含有“其它，请指出”答案的附加内容的多选题，基本处理思路和半开放单选题非常相似，即首先把其它算作一个答案选项，而用另一个变量来表示其它的内容。在数据录入完毕后再对附加内容根据频次高低进行二次编码，以进行更为深入的分析。

（2.4-2.6 节本样章略）

2.4 外部数据的获取

2.4.1 读取电子表格数据文件

可支持的文件类型

操作实例

2.4.2 读取文本数据文件

2.4.3 用 ODBC 接口读取各种数据库文件

2.5 数据的保存

2.5.1 存为 SAV 格式

2.5.2 存为其他数据格式

2.6 数据编辑窗口常用操作技巧集锦

2.6.1 连续多个相同值的输入

2.6.2 快速定义成批变量

2.6.3 将 EXCEL 或 WORD 中的数据直接导入 SPSS

2.6.4 快速改变变量排列次序

2.6.5 记录的快速定位

2.6.6 利用排序功能快速查找异常值、极端值

2.6.7 利用变量值标签检查录入错误

2.6.8 冻结行或列

2.6.9 快速重复调用对话框

2.6.10 从其它窗口中快速切换回数据窗口

思考与练习

1．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。

2. 在完成上述练习的基础上，请尝试自行在 SPSS 中按照 CCSS 项目的问卷建立相应的数据集结构。