目 录

第1章 统计分析软件SPSS for Windows概述 …… (1)
§1.1 软件的发展及特点 …… (1)
§1.2 软件基础知识 …… (2)

第2章 SPSS数据文件的建立和管理 …… (9)
§2.1 数据文件的特点与结构 …… (9)
§2.2 定义数据文件的结构 …… (15)
§2.3 数据的录入与保存 …… (18)
§2.4 数据的编辑 …… (20)
§2.5 与其他计算机软件数据共享 …… (23)

第3章 SPSS数据文件的基本加工和处理 …… (28)
§3.1 数据文件的整理 …… (28)
§3.2 数据加工 …… (38)
§3.3 数据分组 …… (50)
§3.4 数据文件的其他处理功能 …… (55)

第4章 SPSS输出窗口的操作和管理 …… (59)
§4.1 索引输出区管理 …… (60)
§4.2 详解输出区管理 …… (63)
§4.3 输出窗口中分析结果的存取 …… (69)

第5章 SPSS基本统计分析 …… (71)
§5.1 案例说明 …… (71)
§5.2 计算基本描述统计量 …… (73)
§5.3 变量的频数统计 …… (79)
§5.4 交叉分组下的频数分析 …… (84)
§5.5 多选项分析 …… (92)
§5.6 基本统计分析的报表制作 …… (105)

第6章 SPSS的参数检验 …… (113)
§6.1 统计推断的基本方法 …… (113)
§6.2 单样本 t 检验 …… (115)
§6.3 两独立样本 t 检验 …… (117)
§6.4 两配对样本 t 检验 …… (122)

第7章 SPSS的非参数检验 …… (127)
§7.1 单样本非参数检验 …… (127)
§7.2 两独立样本非参数检验 …… (138)
§7.3 多独立样本非参数检验 …… (146)
§7.4 两配对样本非参数检验 …… (152)
§7.5 多配对样本非参数检验 …… (158)

第8章 SPSS的方差分析 …… (166)
§8.1 方差分析概述 …… (166)
§8.2 单因素方差分析 …… (167)
§8.3 单因素方差分析案例分析 …… (172)
§8.4 均值比较 …… (175)
§8.5 多因素方差分析和协方差分析 …… (178)
§8.6 多因素方差分析案例分析 …… (183)

第 9 章　SPSS 的相关分析和线性回归分析 …… (185)
§9.1　相关分析 …… (185)
§9.2　相关分析案例分析 …… (191)
§9.3　偏相关分析 …… (192)
§9.4　线性回归分析 …… (195)
§9.5　线性回归分析案例分析 …… (209)
§9.6　曲线估计 …… (217)
§9.7　权重估计 …… (221)

第 10 章　SPSS 的聚类分析 …… (224)
§10.1　聚类分析概述 …… (224)
§10.2　层次聚类分析 …… (225)
§10.3　快速聚类分析 …… (236)
§10.4　案例分析 …… (242)

第 11 章　SPSS 的因子分析 …… (250)
§11.1　什么是因子分析 …… (250)
§11.2　因子分析的数学模型和相关概念 …… (251)
§11.3　因子分析的基本步骤 …… (253)
§11.4　案例分析 …… (259)

参考文献 …… (268)

第 1 章

统计分析软件 SPSS for Windows 概述

§1.1 软件的发展及特点

SPSS 是 Statistical Package for Social Science 的英文缩写，即：社会科学统计分析软件包。它是当今世界上公认和流行的综合统计分析软件包。在我国，SPSS for Windows 以其强大的统计分析功能、方便的用户操作界面、灵活的表格式分析报告及其精美的图形展现，受到了社会各界统计分析人员的喜爱。目前，SPSS 软件使用已经成为许多大专院校统计学专业和财经类、管理类专业本科学生的必修课程。

SPSS 起源于 20 世纪 60 年代的美国斯坦福大学。进入 70 年代，专门研制和经营 SPSS 软件的 SPSS 公司成立。此时的 SPSS 软件是在中小型计算机上运行，其版本统称为 SPSSx。80 年代初，随着微型计算机的出现，SPSS 公司以其敏锐的目光，迅速研制成功了运行在微型计算机 DOS 操作系统上的 SPSS 的第一版、第二版、第三版等，从而使 SPSS 得到了更为广泛的使用，并占领了计算机统计分析软件的大部分市场份额。此时的 SPSS 版本统称为 SPSS/PC+。90 年代，随着微机 Windows 图形操作系统的出现和盛行，SPSS 公司又研制出了 SPSS for

Windows的第五版、第六版。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以及对国际互联网Internet的广泛使用，SPSS for Windows的第七版、第八版、第九版、第十版又相继诞生。目前，SPSS for Windows已经成为世界各地流行使用的统计分析软件，并和权威的SAS等统计分析软件共享市场。

从SPSS/PC+版本到SPSS for Windows版本，SPSS在用户操作和分析结果的展现方面有了非常大的改进。SPSS/PC+版本是通过用户输入SPSS命令程序和参数的方式来完成数据的管理和统计分析工作的，统计分析文字结果和图形结果均以文本字符方式展现。SPSS for Windows的第五版和第六版，在保留以前版本的人工输入命令、参数操作方式的同时，还为用户提供了直观的图形化菜单界面。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点击菜单或按钮并配合简单的对话框输入来实现，从而免去了记忆命令和参数的负担，也不需要任何计算机编程。同时，对于图形分析结果，能够以图形点阵的方式显示，并允许用户对其进行编辑。SPSS for Windows第七版以后的版本，进一步完善了用户的操作界面、增加了用户熟悉的与Windows 95风格一致的工具栏按钮。在分析结果的输出处理方面，改变了原来版本的输出方式，以崭新的树型结构方式管理分析结果，以文字、表格、图形混排的形式展现分析结果，以强大、灵活的编辑功能随心所欲地编辑分析结果，以内容丰富的“统计教练”联机帮助方式帮助用户理解分析结果。本书以SPSS for Windows 8.0版本为讲解内容。

§1.2 软件基础知识

一、SPSS for Windows的安装和启动

SPSS for Windows作为Windows 95/98下的软件产品，具有和其他软件基本相同的安装步骤。基本步骤概括如下：

(1) 启动Windows，将存有SPSS for Windows的光盘放入光盘驱动器。

(2) 运行资源管理器，读取光盘文件。

(3) 双击执行安装程序文件setup。

(4) 根据用户的需要选择所需要安装的程序项。

正确完成安装步骤以后，就可以启动运行SPSS for Windows软件了。

二、SPSS for Windows中的两个基本窗口

1. SPSS的数据编辑窗口

正常启动 SPSS for Windows 后屏幕显示如图 1—1 所示。

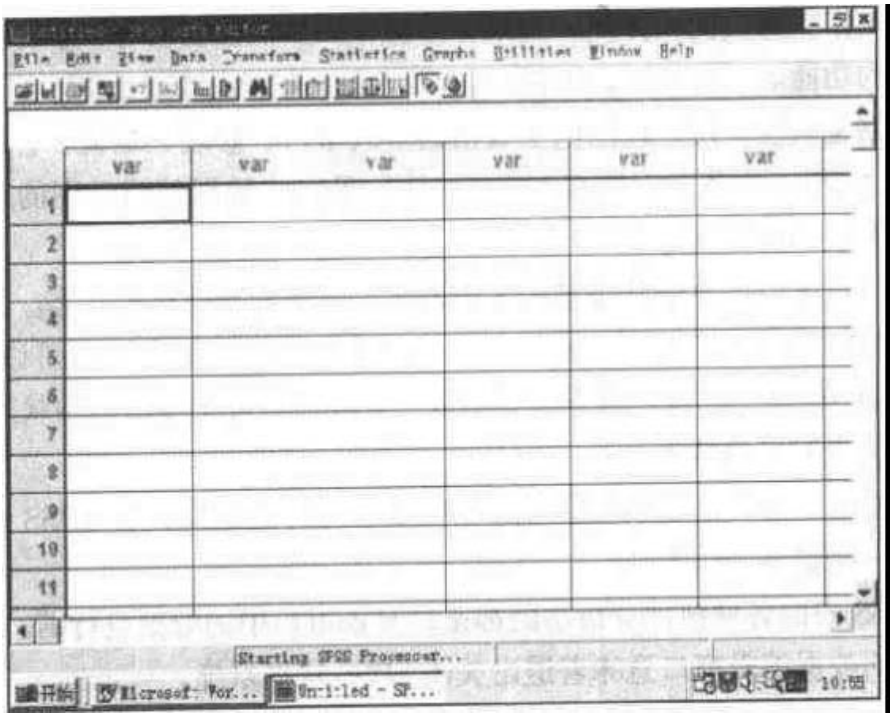


图 1—1　SPSS 数据编辑窗口

SPSS 数据编辑窗口的窗口标题为 data editor，它是 SPSS for Windows 的主程序窗口。数据编辑窗口是对分析对象——SPSS 数据文件进行录入、修改、管理等基本操作的窗口。SPSS 数据文件均以 .sav 作为文件扩展名存储在磁盘上。

数据编辑窗口包括几个部分：窗口主菜单、工具栏、数据编辑区、系统状态显示区。

数据编辑窗口的窗口主菜单基本功能是：

File：SPSS 文件的操作菜单

Edit：SPSS 文件的编辑菜单

Data：SPSS 数据文件的建立和编辑菜单

View：SPSS 用户界面基本状态设置等菜单

Transform：SPSS 数据的基本处理菜单

Statistics：SPSS 的统计分析菜单

Graphs：SPSS 的统计图形菜单

Utilities：SPSS 的相关应用和设置菜单

Windows：SPSS 中各窗口的切换菜单

Help：SPSS 的帮助菜单

以图形按钮组成的工具栏将常用的 SPSS 功能列示出来，这些功能都可以在

相应的功能菜单中找到。用户可以直接点击工具栏上的某个按钮完成其相应的功能，使得操作更加快捷和方便。当鼠标停留在工具栏按钮上时，系统会自动提示相应按钮的功能。

在数据编辑区，用户按照电子表格的形式录入、修改、编辑和管理待分析的数据。通过 View 主菜单下的 Grid Lines 选项可以将数据编辑区中的表格线设置成显示或不显示两种状态。

数据编辑窗口的系统状态显示区用来显示系统的当前运行状态。当系统等待用户操作时，会出现“SPSS processor is ready”的提示信息。也可以通过 View 主菜单下的 Status Bar 选项将系统状态区设置成显示或不显示两种状态。

数据编辑窗口的特点：

- 数据编辑窗口是在系统启动过程中自动打开的窗口，在 SPSS 运行过程中只能打开一个数据编辑窗口。
- SPSS 中的各种统计分析功能都是针对该窗口中的数据进行的。
- 关闭数据编辑窗口意味着退出关闭 SPSS 软件系统。

2. SPSS 的输出窗口

SPSS 输出窗口的窗口名为 Viewer，它是显示和管理 SPSS 统计分析结果、报表及图形的窗口，如图 1—2 所示。出现在该输出窗口中的内容均以 .spo 作为文件扩展名存储在磁盘上。

图 1—2　SPSS 输出窗口

输出窗口包括几个部分:窗口主菜单、工具栏、分析结果显示区、状态显示区。

输出窗口的窗口主菜单的一部分菜单，如：Statistics、Graphs、Utilities、Windows、Help 的功能与数据编辑窗口中的相同，是针对数据编辑窗口中的分析数据的；还有一部分功能菜单，如：View、Insert、Format 是输出窗口的专用菜单；其余的功能菜单，如：File、Edit 则是数据编辑窗口和输出窗口共用的。菜单这样设置的目的是为了方便用户的操作。

需要说明的是，虽然 SPSS 中有些窗口的主菜单名有重名的情况，但是它们下面的子菜单功能却是随不同的当前窗口而不同的，其服务的对象也是不同的。因此，在使用时，用户应注意哪个窗口是当前窗口。

输出窗口的工具栏除了保留了数据编辑窗口中的某些图形按钮外，还增添了自己特有的功能按钮。

分析结果显示区是显示统计分析结果的地方。区域分成左右两个部分。左边以类似于 Windows 资源管理器的树形结构形式显示已有的分析结果的索引；右边显示的是各个分析结果的详细内容。左右两边的内容以两个红色的右箭头为标记一一对应起来。用户可以对该区域中的内容进行增、删、改等编辑管理操作。

输出窗口的状态显示区与数据编辑窗口的状态区的区别是多了一个红色的“!”标记。该标记表示当前的输出窗口为主输出窗口。主输出窗口就是下一次分析结果将输出到的那个窗口。

输出窗口的特点：

● 输出窗口是在用户进行第一次统计分析时被系统自动打开的。

● 在 SPSS 运行过程中用户还可以手工创建若干个新输出窗口。菜单选项为：File －＞ New －＞ Output。手工创建若干个新输出窗口后，用户可以将同一批分析数据的不同分析结果指定输出到不同的输出窗口中，这样就可以将不同输出窗口中的内容以不同的文件名分别存放到磁盘上。

● 不同的输出窗口之间可以相互自由切换，可以利用主菜单中的 Windows 功能菜单实现。它们中只有一个主输出窗口。主输出窗口的状态栏上有一个红色的“!”标记，而其他输出窗口的状态栏上则没有这个标记。如果用户希望将以下的统计分析结果输出到某个输出窗口中，需要按动工具栏上的“!”图形按钮来指定某输出窗口为主输出窗口。

另外，SPSS 还有一个窗口名为 draft Viewer 的输出窗口。它可以通过 File －＞ New－＞draft view 的菜单选项来创建。该窗口功能与上面提到的输出窗口的功能完全相同。但是，出现在 draft viewer 输出窗口中的内容是以.rtf 作为文件扩展名存放到磁盘上的。.rtf 文件是按丰富文本格式（rich text format）

存储的，它可以被 word 等其他软件直接读取和编辑。

三、SPSS for Windows 的退出

SPSS 退出可以通过 File－＞Exit 的菜单选项实现，也可以直接按数据编辑窗口右上角的[x]按钮。在退出 SPSS 之前，一般会向用户提出以下两个问题：

- 是否将 data editor 窗口中的数据存到磁盘上，数据文件扩展名为.sav；
- 是否将 viewer 窗口中的数据存到磁盘上，结果文件扩展名为.spo 或 .rtf。

这时，用户应指定将 SPSS 数据文件或结果文件存放到哪个磁盘上，并输入相应文件名的用户名部分。

四、SPSS for Windows 的运行方式

SPSS 为用户提供了三种基本运行方式，分别为：完全窗口菜单方式、程序运行方式、混合运行方式。

1. 完全窗口菜单方式

完全窗口菜单方式是在 SPSS 运行中，所有分析操作过程都是用户通过点菜单或按钮和输入对话框的方式进行的。如果用户不清楚某些统计选项的含义，可以将鼠标移动到相应的选项上，然后单击鼠标右键。于是，在屏幕上就会出现相应的解释说明，从而有助于用户理解统计选项的意义。

完全窗口菜单方式是一种经常使用的运行方式，用户不需要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。这种工作方式很适用于一般的统计分析人员和 SPSS 的初学者。在以后的章节中，我们主要讲解这种方式。

2. 程序运行方式

程序运行方式是统计分析人员首先根据自己的分析步骤，手工编写相应的 SPSS 命令程序；然后将编写好的程序一次性提交计算机运行。SPSS 软件系统会自动按照程序命令语句的前后顺序自动逐句执行 SPSS 的命令。程序运行方式很适用于熟练的 SPSS 程序员进行特殊分析的需要。在进行大规模的统计分析工作时，程序运行方式能够按照程序自动进行分析，不需要分析过程中的人工干预。

程序运行方式的实现步骤有以下两大步骤：

第一步，打开窗口名为 Syntax Editor 的语句窗口并编写 SPSS 程序。

语句窗口是专门用来供 SPSS 程序员编写 SPSS 程序的窗口。该窗口中的 SPSS 命令程序是按照 SPSS 软件的语法规则编写的，它所分析处理的数据是数

据编辑窗口中的数据。需要说明的是，SPSS for Windows 版本中的 SPSS 命令语法与 DOS 下的 SPSS 各版本的语法是完全兼容的。因此，熟悉 DOS 版本下的 SPSS 用户可以沿袭以前版本的语法来编写 Windows 版本下的 SPSS 程序。语句窗口中的 SPSS 命令程序均以.sps 为文件扩展名存放在磁盘上。

语句窗口有自己的主菜单项、工具栏和系统状态栏，其功能与数据编辑窗口和输出窗口类似。

打开语句窗口的方式可以有两种：

(1) 手工打开

菜单选项：File－＞New－＞Syntax。可以创建一个新的语句窗口。

(2) SPSS 启动时自动打开

菜单选项：Edit－＞Option。如图 1—3 所示。这样，在 SPSS 启动时会自动启动语句窗口。

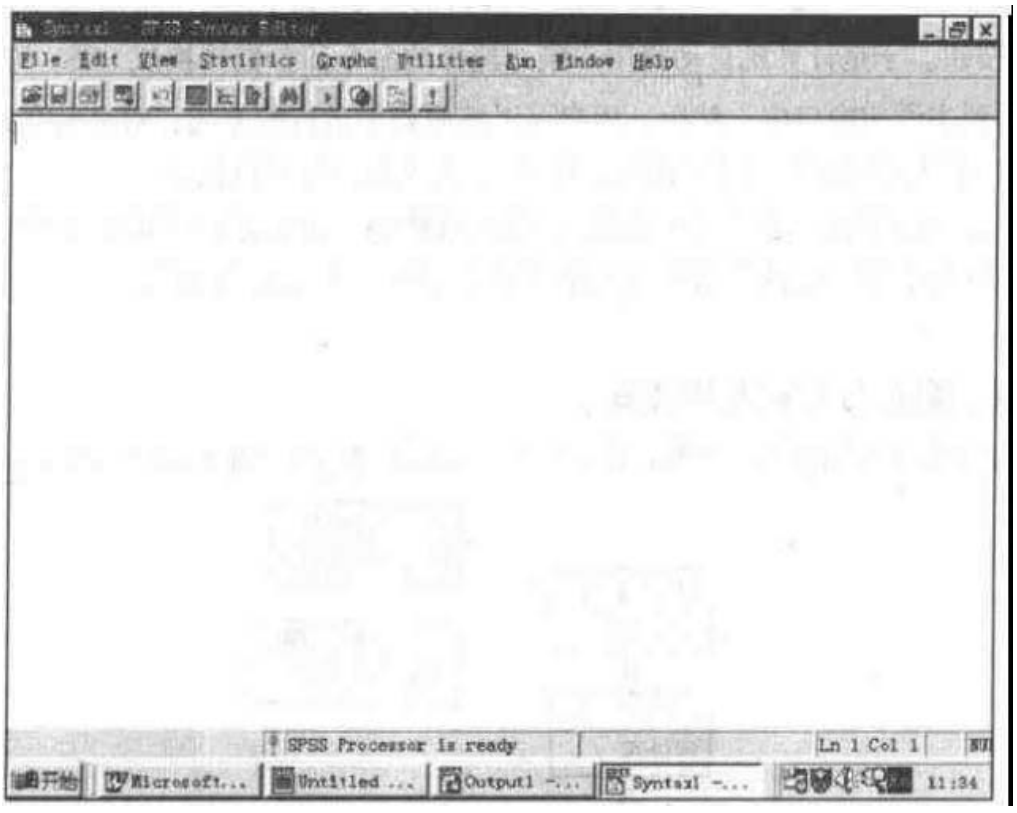


图 1—3　SPSS 语句窗口

语句窗口的特点是：

● 在 SPSS 运行过程中用户可以手工创建若干个新语句窗口。这样就允许用户将对同一批数据的不同分析命令程序分别写在不同的语句窗口中，从而可以将它们分别以不同的文件保存在磁盘上。

● 用户可以在多个语句窗口之间相互切换，同时编辑多个 SPSS 程序。其中在语句窗口的系统状态栏上有红色的“!”的为主语句窗口。

第二步，执行语句窗口中的 SPSS 程序。

选择语句窗口主菜单中的 Run 项。SPSS 提供了四种程序执行的方式：

(1) 执行全部 All：执行当前语句窗口中的所有 SPSS 命令程序。

(2) 执行部分 Selection：仅执行语句窗口中当前已选中的 SPSS 命令程序。命令选中可以通过鼠标在相应 SPSS 命令行上拖动来实现。

(3) 执行当前 Current：仅执行语句窗口中当前光标所在行的 SPSS 命令。

(4) 执行到最后 To end：仅执行语句窗口中当前光标所在行之后的所有 SPSS 命令程序。

3. 混合运行方式

混合运行方式是前两种运行方式的综合。首先按照菜单运行方式选择统计分析的菜单和参数，此时，不马上按 OK 按钮提交系统执行，而是按动屏幕上的 Paste 按钮。于是计算机自动将用户刚才选择的命令和参数转换成 SPSS 的命令，并粘贴到主语句窗口中；然后，用户可以按照程序运行的方式，对显示在语句窗口中的 SPSS 命令进行必要的编辑修改后，再提交给计算机执行。

混合运行方式一般适用于熟练的 SPSS 程序员。可以通过该运行方式手工输入某些分析参数，而这些参数大多是无法通过完全菜单方式给定。

五、数据分析的基本步骤

利用 SPSS 进行数据分析的基本步骤可以简单概括为图 1—4 所示的几步。

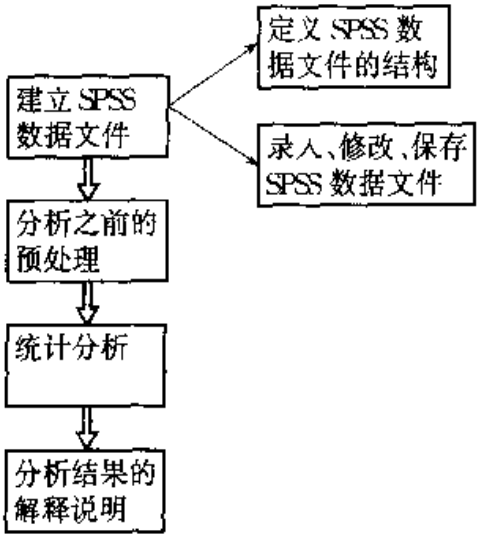


图 1—4

第 2 章

SPSS 数据文件的建立和管理

§2.1 数据文件的特点与结构

一、数据文件的特点

利用 SPSS 进行数据分析的第一步工作就是要建立 SPSS 的数据文件。应首先了解什么是 SPSS 的数据文件。

SPSS 数据文件是一种有结构的数据文件，它包括文件结构和数据两部分。

SPSS 数据文件中的一列数据称为一个变量。每个变量都应有一个名称，即：变量名，变量名是存取 SPSS 每个变量的惟一标识。SPSS 数据文件的结构是对每个变量及相关属性的描述。SPSS 数据文件中的一行数据称为一条个案(Case)。所有个案组成了 SPSS 数据文件的数据部分。

SPSS 数据文件与一般的文本不同点在于：一般的文本文件仅有纯数据部分，而没有关于结构描述的部分。因此，SPSS 数据文件不能像一般的文本文件那样可以直接被大多数编辑软件读取，而只能在 SPSS 软件中打开。

建立 SPSS 数据文件应通过两大步骤实现：

第一步：定义 SPSS 数据文件的结构；

第二步：输入分析数据。

二、数据文件的结构

SPSS 数据文件的结构是对每列变量及其相关属性的描述，它包括：变量名、变量类型、变量宽度、变量名标签、变量值标签、变量列格式、变量缺失值、变量度量方法等信息。

1. 变量名（Variable Name）

变量名是变量存取的惟一标志。在定义 SPSS 数据文件结构时应首先给出每列变量的变量名。对于变量的起名规则是：

● 变量名不能多于 8 个字符。

● 首字符应以英文字母开头，后面可以跟除了“!、?、 *”之外的字母或数字；下划线、圆点不能为变量名的最后一个字母。

● 变量名不区分大小写字母；允许汉字作为变量名，汉字总数不能超过 4 个。

● 变量名不能与 SPSS 内部特有的具有特定含义的保留字符相同，如：ALL、BY、AND、NOT、OR 等。

● SPSS 默认的变量名为以 VAR 三个字母开头，后面跟补足 5 位的数字。如：VAR00001、VAR00012 等。

总之，在变量起名时，为方便记忆，变量名最好与其代表的数据含义相对应。如果变量名不符合 SPSS 的起名规则，系统会自动给出错误提示信息。

2. 变量的类型（Type）和宽度（Width）

SPSS 中的有三种基本的变量类型，它们是：数值型、字符型和日期型。每种变量类型的数据在数据编辑窗口中都有其默认的显示宽度和样式。

（1）数值型

数值型是 SPSS 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、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。数值型有以下六种不同的表示方法。

1）标准数值型（Numeric）

标准数值型是系统默认的数据类型。标准数值型默认的最大显示宽度是 8 位，小数位为 2 位。如果数据是整数，它们的最大显示位是 8 位正整数或 7 位负整数。如：123 456 78、－123 456 7。如果数据是带有小数位的，则总显示宽度的 8 位中包括了 2 位小数，1 位小数点。如：12 345.67、－1 234.56。用户可以根据具体情况对默认显示宽度进行修改。否则，如果数据值总宽度超过了 8 位，

SPSS会在显示宽度为8的总前提下，自动调整数据的显示。如：123 456.78显示为123 456.8,第二位小数四舍五入。又如：1 234 567.89显示为1 234 568,第一位小数四舍五入后截去小数部分。又如：123 456 789显示为 1.23E + 08，－123 456 789显示为－1.2E + 08，即：整数部分大于8位时，以科学记数法显示。

需要说明的是，数据的显示并不影响真正数据的存储，也不影响数据的计算。

2）逗号数值型（Comma）

逗号数值型数据其整数部分从个位开始每3位以一个逗号分隔。它的默认最大显示宽度为8，小数位为2。逗号所占的位数包括在总显示宽度之内。如：1,234.56。如果数据值的总宽度大于默认的8位，SPSS会自动调整，即：首先去掉逗号的显示，然后按标准数值型的显示调整方式调整数据的显示。如：123,456.7显示为123456.7、1234,567.8显示为1234568等。用户在输入逗号数值型数据时，可以不输入逗号，SPSS将自动在相应位置上添加逗号。

3）圆点数值型（Dot）

圆点数值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。它的默认最大显示宽度为8，小数位为2。如：1.234,56。当数据值的总宽度大于默认的8位时，SPSS会按照上述逗号数值型的显示调整规则自动调整数据的显示。用户在输入圆点数值型数据时，可以不输入圆点，SPSS将自动在相应位置上添加圆点。

4）科学记数法型（Scientific Notation）

科学记数法也是数值型数据表示方式的一种。如：120用科学记数法表示为1.2E+02，其中的E表示以10为底，+02表示正的2次方。又如：0.005用科学记数法记为5.0E－03，这里－03表示负的3次方。科学记数法的默认显示总宽度为8，包括整数前的正负符号位、字母E和跟在它后面的正负符号及两位数字。科学记法一般用来表示很大或很小的数据。用户在输入科学记数法数据时，可以按标准数值型方式输入数据，SPSS会自动进行转换。

5）美元数值型（Dollar）

美元数值型主要用来表示货币数据，在数据前附加美元符号$。美元数型的具体显示格式有很多，如：$#、S###、$###，####、$#，###.##等，SPSS会以菜单的方式显示出来供用户选择。

6）用户自定义型（Custom Currency）

用户自定义型是数值型数据表示方式的一种。如果用户认为上述五种数值型

数据的表示方式还不能满足需要时，自己可以再定义不超过4种的表示方式。每种表示方式都应有自己的名称，它们只能是SPSS系统提供的4个名称CCA、CCB、CCC、CCD中的一个。实际上，对CCA、CCB、CCC、CCD，SPSS系统已经预先将它们定义成了逗号数值型，用户可以对它们进行重定义。重定义可以通过Edit－>Options－>Curreny的菜单选项实现，如图2—1所示。

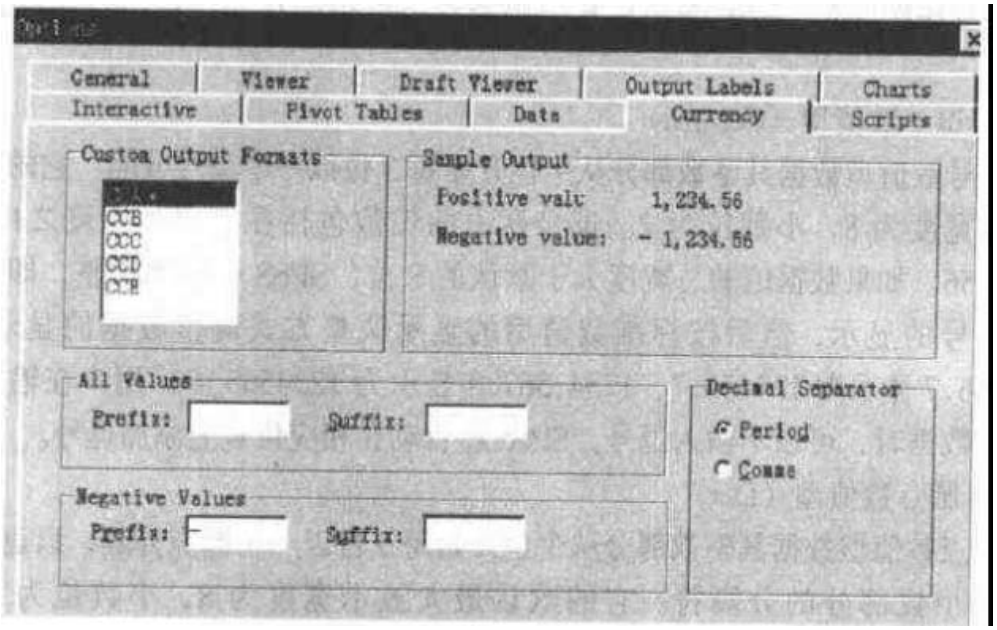


图2—1 SPSS用户自定义数据显示方式窗口

在All Values框中，在Prefix后面输入数值型数据的开始字符，在Suffix后面输入数据的结尾字符。在Negative Values框中，在Prefix后面输入负数值型数据的负号后的开始字符，同样，在Suffix后面输入负数的结尾字符。上述字符SPSS系统都默认是空格。在Decimal Separator框中指定是以圆点（Period）还是以逗号（Comma）作为小数点。

对SPSS预定的CCA或CCB或CCC或CCD进行重新定义后，用户便可以指定某些变量的表示方式是其中的某个了。

（2）字符型（String）

字符型数据类型也是SPSS较常用的数据类型，是由一个字符串组成。如：职工号码、姓名、地点等变量都可定义为字符型数据。字符型数据的默认显示宽度为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数据处理过程中是应用一对双引号括起来的，面在输入数据时不应输入双引号、否则，双引号将会作为字符型数据的一部分。

（3）日期型（Date）

日期型数据用来表示日期或者时间。如：生日、成立日期等变量可以定义为日期型数据。日期型数据的显示格式有很多，例如：dd-mmm-yyyy，dd 表示 2 个字符位的日期，－为数据分隔符，mmm 表示英文月的缩写，yyyy 表示 4 个字符位的年份。如：25－AUG－1999 表示 1999 年 11 月 25 日。又例如：mm/dd/yyyy，mm 表示 2 个字符位的月份，/为数据分隔符，dd 表示 2 个字符位的日期，yyyy 表示 4 个字符位的年份，如：1999 年 11 月 25 日也可以表示为 11/25/1999。SPSS 以菜单的方式将所有的日期显示格式列出来供用户选择。

3．变量名标签（Variable Label）

变量名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。变量名标签可以用中文，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，但我们建议最好给出各变量名的标签。

4．变量值标签（Value Label）

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，假定用 1 表示男，用 2 表示女。如果性别变量没有变量值标签，就很难弄清楚是 1 代表男还是 2 代表男。因此，变量值标签对于顺序水准（如：收入的高、中、低）和名义水准（如：民族、性别）的变量来说是必不可少的，它不但明确了数据的含义，也增强了最后统计分析结果的可读性。变量值标签可以用中文。变量值标签这个属性是可以省略定义的，但我们建议，对顺序水准和名义水准变量最好给出变量值标签。

5．变量列格式（Column Format）

变量列格式包括：变量的列宽度（Column Width）和数据的对齐方式（Text Alignment）。

显示在数据编辑窗口中的数据，其变量列宽度默认值是各变量类型所默认的显示宽度。例如：标准数值型的默认的显示宽度为 8，那么，标准数值型的变量的列宽度就默认为 8。变量的列宽度随变量显示宽度的改变而自动改变。

在某些情况下需要对变量的列宽度进行修改。例如：如果某变量的显示宽度为 4，变量名的字符位宽度为 8，即变量的显示宽度小于变量名宽度。这时，变量名就无法完整地显示在数据编辑窗口中。又例如：在前面讨论变量的类型和宽度时，我们曾说明，如果数据值的总宽度超出了定义的显示宽度，SPSS 将自动调整数据的显示，或截去长出的小数部分，或以科学记数法显示数据，使得数据无法完整地显示在数据编辑窗口中。此时，就可以通过增加变量的列宽度来解决上述问题。

数据在数据编辑窗口中显示时有 3 种对齐方式：左对齐（Left）、右对齐（Right）、中间对齐（Center）。字符型变量默认左对齐显示，数值型变量默认右对齐显示。

6. 变量缺失值（Missing Values）

在统计分析过程中，收集到的数据可能会出现这样的情况。一是，数据中出现明显错误和不合理的数据。例如：某个年龄数据为 213 岁。这显然是不符合实际情况的失真数据；另一种是，有些数据项的数据漏填了。例如：某个职称数据遗漏了。这些数据一般都应当作缺失值数据处理。

缺失值在统计分析中是无法直接采用的。在 SPSS 中处理的方法是：首先，在遗漏的数据处填充入某个特定的标记数据。例如：将遗漏的职称数据以特定的标记数据，如用 999 来填补替代。注意：这里 999 应不是一个合理的职称数据；然后，再告知 SPSS 这个特定的标记数据，如：999、或那些明显的失真数据，如：213 岁，为缺失值数据。这样，在进行统计分析时，SPSS 将区别对待缺失值和正常的分析数据。

SPSS 中缺失值可以有三种指定方法：

● 对一个字符型或数值型变量，缺失值可以是 1 至 3 个特定的值离散值（Discrete missing values）。

● 对一个数值型变量，缺失值可以在 1 个连续的闭区间内（Range of missing values）。

● 对一个数值型变量，缺失值可以在 2 个连续的开间内，并可同时指定一个离散值（Range plus one missing value）。

SPSS 中缺失值有用户自定义缺失值和系统缺失值两大类。上面谈到的缺失值就是用户自定义缺失值。系统缺失值用一个圆点表示。这个圆点不等于 0，也不等于.00，而是系统默认的缺失值，它一般出现在数值型变量数据中。注意：字符型变量中的空格或空不是系统缺失值。

7. 变量度量方法（Measurement）

统计分析数据的度量方法可以大致分为定距度量和非定距度量两类，对应的数据就是定距数据和非定距数据。定距离数据（Scale）是指，如人的身高、体重、血压等的连续型数据，也包括如职工人数、课程门数等的离散化的数据。非定距数据包括：顺序水准（Ordinal）的数据和名义水准（Nominal）的数据。顺序水准的数据，如：高、中、低级职称分别用 1、2、3 等表示、年龄段分别用“老”、“中”、“青”等表示。这里，无论是数值型的 1、2、3 还是字符型的“老”、“中”、“青”，都存在内在固有的大小或高低顺序。名义水准的数据，如：

男、女性别分别用 1、2 表示，民族用“汉”、“回”、“满”等表示。这里，无论是数值型还是字符型都不存在内部固有的大小或高低顺序，而只是一种名义上的指代。

对 SPSS 中的变量也应给出度量方法的描述，默认的度量方法是等距度量。

总之，在描述 SPSS 数据文件结构时，应包括图 2—2 所示的几个方面。

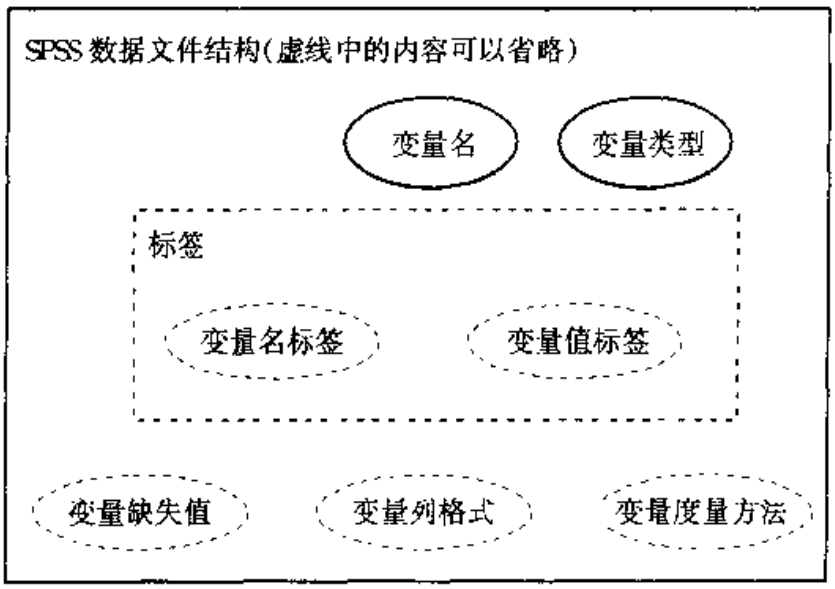


图 2—2　SPSS 数据文件结构

§2.2　定义数据文件的结构

一、手工逐个定义变量

手工逐个定义变量方法是一种常用的定义 SPSS 数据文件结构的方法。它可以通过 3 种操作手段实现。这些手段的最终效果是相同的，用户可以依照自己的操作习惯选其中的任何一种。在以后的操作中也会出现类似的操作方式，我们就不再作详细介绍了。

1. 选菜单方式

步骤：在数据编辑窗口中

(1) 将鼠标定位在数据编辑区的某个变量列上。

(2) 菜单选项：Data －＞ Define Variable，出现图 2—3 所示的窗口。

2. 双击鼠标左键

步骤：在数据编辑窗口中

(1) 将鼠标定位在数据编辑区的某个变量名行上。

(2) 双击鼠标左键，出现图 2—3 所示的窗口。

3. 单击鼠标右键

步骤：在数据编辑窗口中

(1) 将鼠标定位在数据编辑区的某个变量名行上。

(2) 单击鼠标右键，在弹出的菜单中点击 Define Variables 项，出现图 2—3 所示的窗口。

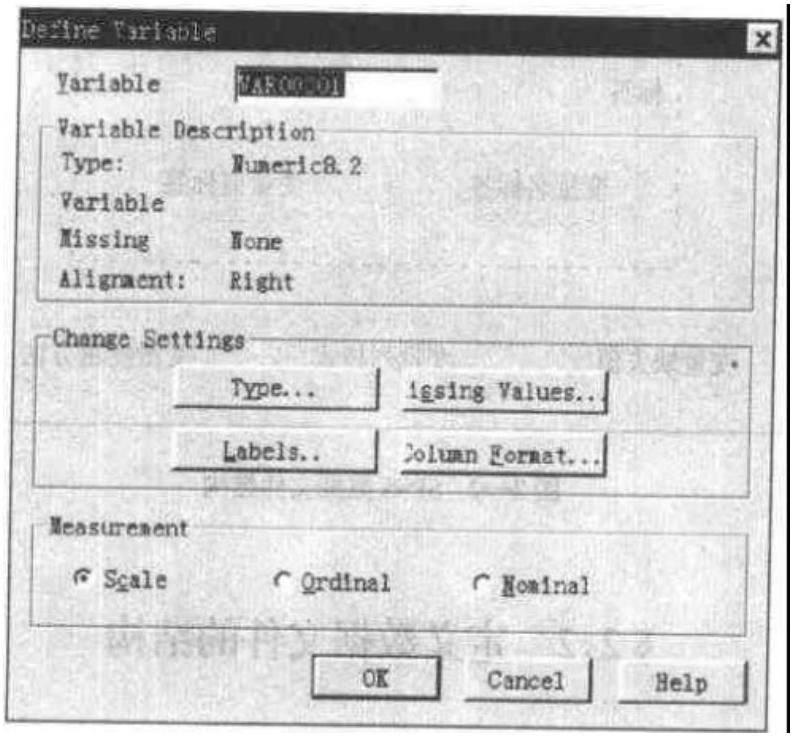


图 2—3　SPSS 数据结构定义窗口

二、利用变量模板定义变量

有时需要定义变量的个数较多，而且这些变量除变量名之外又有很多共同的属性，如：大部分变量的类型都是数值型，显示宽度均为 6，变量值标签也相同。这时，手工逐个定义变量就显得较为烦琐。这时，利用变量模板能够提高变量定义的效率。

变量模板是 SPSS 提供的一种定义变量的方式。它实质就是变量和相应属性的一个“模子”，这个“模子”具有较大的通用性，通过它可以较快捷地“生产”出一批变量。因此，利用变量模板去定义变量，应首先建立一个用户的变量“模

子”，即：用户变量模板，然后才能使用它去“生产”一批变量。

1. 建立用户变量模板

建立用户变量模板主要定义以下几个方面的内容：变量模板名称、变量的类型、变量值标签、变量列格式、变量缺失值。其中，变量模板名称是使用变量模板的惟一标志，其变量属性大都可以被它“生产”出来的变量所继承，否则，建立这个变量模子就没有意义了。

建立用户变量模板的菜单选项是：Date －＞ Templates，然后按 Define 按钮，于是出现如图 2—4 所示的窗口。

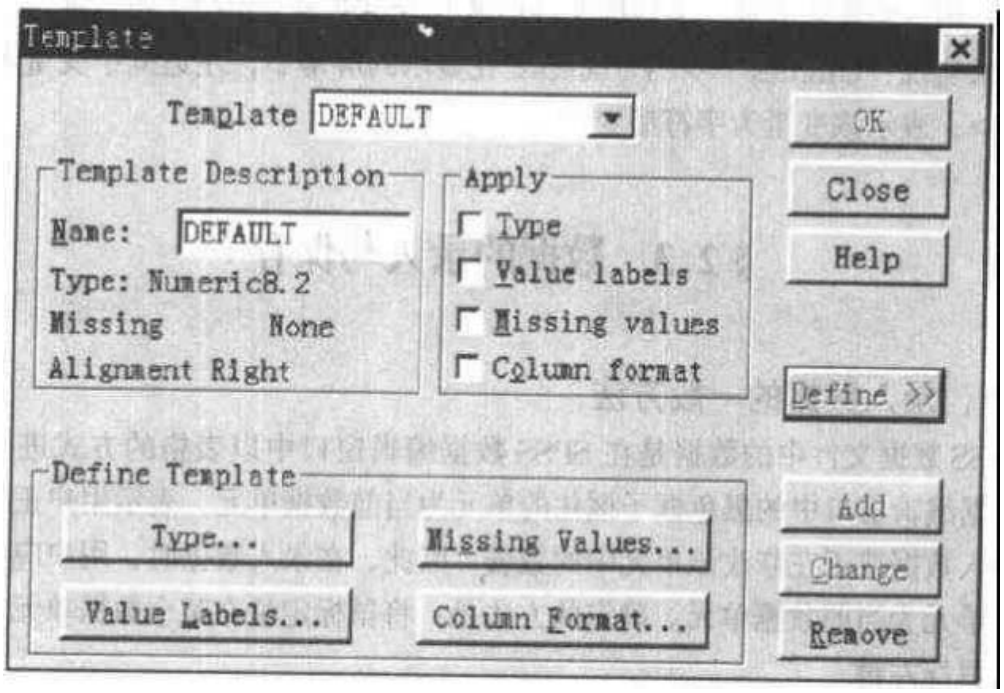


图 2—4　变量模板定义窗口

这里，我们首先假设建立一个名为 XX 的用户变量模板，其类型为标准数值型，显示宽度为 6，并且定义了其变量值标签等。然后，通过按 Add 钮将它加入到系统变量模板集合中。于是，一个用户变量模板便建立完成了。可以通过按 Change 和 Remove 钮实现用户变量模板的修改和将某用户变量模板从系统变量模板集合中删去。

用户变量模板建立后，应指出该用户变量模板中的哪些属性将被继承。其中包括：变量类型、变量值标签、缺式值、变量列格式。

2. 使用用户变量模板

定义完用户变量模板之后便可以通过它来定义变量了。操作步骤是：

(1) 将鼠标定位在数据编辑区的一个列的变量名行上。

(2) 菜单选项：Data－＞Templates，找到相应的用户变量模板名，如：XX，于是，通过用户变量模板就可以“生产”出一个 SPSS 变量了。

需要说明的是：SPSS 系统变量模板中有 4 个系统自定义的变量模板，它们分别名为：DEFAULT、MONTHS、STATES、WEEKDAYS。DEFAULT 是用户在手工定义变量时系统默认的变量模板。其他的变量模板对变量值标签，如：月份、星期和美国的州名等作了详细的定义，用户可以根据情况使用它们，从而快捷地继承其变量值标签。

以上，我们较为详细地说明了如何定义 SPSS 数据文件的结构。结构定义完成之后，用户会希望浏览一下数据文件中所有变量的定义情况。可以通过下列菜单选项来实现：Utilities －＞ Variables。在显示的屏幕中，左边框中变量名后面的字符＞，表示该变量为字符型变量。

§2.3 数据的录入与保存

一、录入数据的一般方法

SPSS 数据文件中的数据是在 SPSS 数据编辑窗口中以表格的方式进行录入的。数据编辑窗口中的黑色框子框住的单元为当前数据单元，表示用户正在对该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，用户应首先确定哪个单元为当前数据单元。确定的方法是：将鼠标定位在某个数据单元上，然后单击鼠标左键。

数据录入时可以逐行录入，即：录入完一个数据后，按 Tab 键，黑框便自动跳到本行的下一个变量列上。也可以逐列录入数据，即：录入完一个数据后，按 Enter 键，黑框便自动跳到本列的下一行上。

二、录入带有变量值标签的数据

在录入带有变量值标签的数据时，我们总希望手工输入的是实际的变量值，而屏幕显示的是与该变量值对应的变量值标签。这样，既方便了数据录入，又有利于数据的检查。其实现的操作步骤是：

(1) 打开显示变量值标签的开关。点菜单选项：View－＞Value Labels。Value Labels 是一个开关选项，如果它前面出现了一个对钩符号，表示开关已经打开，变量值标签将显示在数据编辑窗口中。否则，表示开关尚未打开。

(2) 确定当前数据单元（该列变量是带有变量值标签的）。

(3) 录入相应的变量值后，在数据单元中显示的是其所对应的变量值标签。

如果用户记不住哪个变量值与哪个标签相对应，如：不清楚是1表示男还是2表示男时，可以在确定了当前数据单元之后，单击鼠标右键，在弹出菜单中点击 pick from labels 选项。此时，在当前数据单元的旁边会出现该变量的所有变量值标签，供用户选择。用户只需要双击鼠标左键，选择相应的变量值标签即可

三、SPSS 数据文件的保存

在 SPSS 数据录入过程中，应及时将数据编辑窗口中的数据以文件的形式保存到磁盘上，以防止数据不慎丢失。

1. 数据的保存格式

在保存 SPSS 数据时，可以将它们存储成多种格式的数据文件，以便与其他软件进行数据共享。常见的数据文件格式有：

(1) SPSS 数据文件格式。它是 SPSS 系统默认的数据格式，也是最常用的一种数据格式。在磁盘上以.sav 作为文件扩展名。这种数据格式的优点是：它可以被 SPSS 软件直接读取，更重要的是，它除保存了 SPSS 数据编辑窗口中的变量值数据之外，还将变量的其他属性，如：变量类型、变量名标签、变量值标签、变量列宽度、缺失值等信息也都保存下来。它的缺点是无法直接被其他软件读取。

(2) Excel 工作表格式文件。文件扩展名为.xls。SPSS 数据文件的变量名将默认写在 Excel 工作表的第一行上，用户也可以指定不写。SPSS 中的一个个案为 Excel 中的一行。

(3) Dbase 数据库文件格式文件。文件扩展名为.dbf。SPSS 的变量名和变量类型将自动转成 Dbase 的字段名和字段类型。SPSS 中的一个个案数据为 Dbase 中的一条记录。

(4) 文本格式文件。文件扩展名为.dat。其中包含两种形式：第一种，固定格式(fixed)的文本文件。它依据每个变量的显示宽度，将变量值数据以空格左补足对齐后，一个个案数据为一行；第二种，以 Tab 键作为各数据之间的分隔符，文件的第一行是 SPSS 变量名，从第二行以后是具体数据，一个个案数据为一行。

上述 (2)、(3)、(4) 这些格式的数据文件的优点是：数据可以被相应的软件直接读取。缺点是：仅保存 SPSS 数据编辑窗口中的变量值数据，而不保存变量的其他属性内容，于是那些珍贵的、非常有利于统计分析的必要信息便丢失了。

因此，在实际工作中，如果用户希望将 SPSS 的数据和其他软件共享，可以将 SPSS 数据编辑窗口中的数据保存成两份，一份为 SPSS 数据文件格式，保证

一些重要的分析信息不丢失，另一份为其他软件可以读取的数据文件格式，从而方便数据共享。

2. 数据保存的操作步骤

SPSS 数据文件保存的操作步骤是：

(1) 菜单选项：File －＞ Save data 或者 File －＞ Save as（另存为）。

(2) 给出存放数据文件的目录路径和数据文件的用户名，并根据实际需要，选择数据文件的格式和相应的文件扩展名。

§2.4 数据的编辑

SPSS 数据文件的编辑，主要包括对 SPSS 数据编辑窗口中的分析数据进行必要的增加、删除、修改、复制等操作。

一、打开 SPSS 数据文件

编辑一个 SPSS 数据文件，应该首先利用 SPSS 软件，在数据编辑窗口中建立一个 SPSS 数据文件，或者将已经建立好并存放在磁盘上的 SPSS 数据文件从磁盘外存装入到计算机内存中。装入后的数据会显示在 SPSS 数据编辑窗口之中。

打开 SPSS 数据文件的操作步骤：

(1) 菜单选项：File －＞ Open。

(2) 输入将要打开的 SPSS 数据文件所在的文件目录和用户名，扩展名默认为 .sav。

二、SPSS 数据定位

SPSS 数据定位就是要确定数据编辑窗口中的哪个单元为当前数据单元。定位方式有两种：人工定位和自动定位。

人工定位方式：通过肉眼人工浏览数据，确定当前数据单元。这种方式适用于数据量较少的情况，用户只需通过鼠标，拖动数据编辑窗口右边的滚动钮，或者按键盘上的 Page Up、Page Down 键就可以完成。

当数据量较大时，一般通过自动定位方式进行数据定位，确定当前数据单元。自动定位是计算机按照用户给定的条件自动寻找满足条件的第一个数据单元，并设置它为当前数据单元。SPSS 中用户给定的定位条件可以是：个案号码、

变量值数据。

1. 按个案号码自动定位的操作步骤

(1) 将光标黑框框在某变量列的任何一个个案上。

(2) 菜单选项：Data －＞ Go to case，出现如图 2—5 所示的窗口。

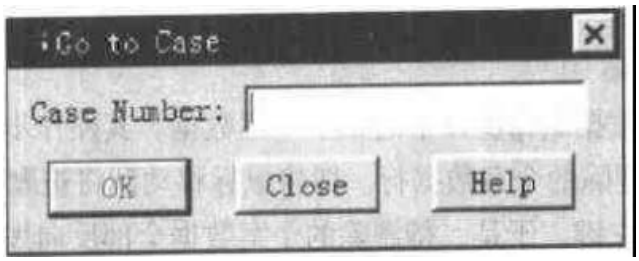


图 2—5　SPSS 按个案号数据自动定位窗口

(3) 输入一个个案号码并确认。计算机自动搜索该号码所对应的个案并确定：当前数据单元的行就是搜索出的个案所在的行，当前数据单元的列就是（1）步中所指定的列。

2. 按变量值数据自动定位的操作步骤

(1) 将光标黑框框在某变量列的任何一个个案上，如年龄。

(2) 选择菜单选项：Edit －＞ Find，出现图 2—6 所示的窗口。

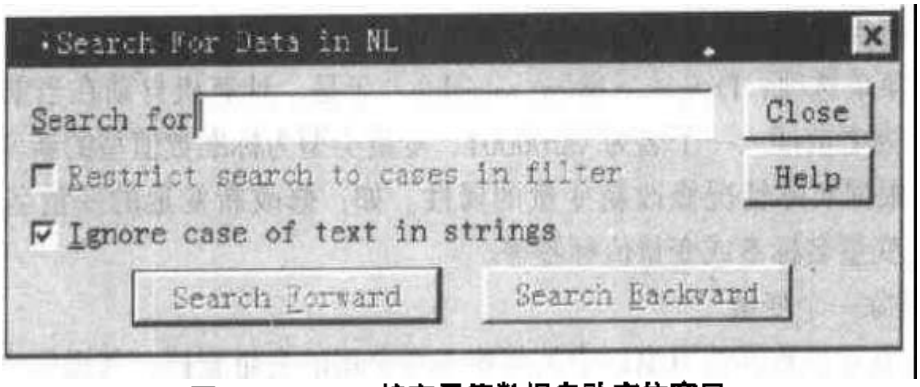


图 2—6　SPSS 按变量值数据自动定位窗口

(3) 输入定位变量值数据，如：32，并确认。计算机便自动对（1）步所指定的列中的数据进行搜索，并找到与变量值数据相等的第一个个案，如：查找到第一个年龄为 32 的个案。于是，当前数据单元的行就是搜索出的个案所在的行，当前数据单元的列就是（1）步中所指定的列。

三、插入和删除一个个案

1. 插入一个个案

数据编辑经常需要在数据编辑窗口的某个个案之前或之后插入一个新个案数

据。其操作步骤是：

(1) 将当前数据单元确定在一个个案数据的任何变量列上。

(2) 菜单选项：Data－＞Insert case。于是，计算机自动在当前数据单元所在行之前插入一个空白行。

(3) 可以在空白行上逐个录入变量值数据。

2. 删除一个个案

有时需要删除数据编辑窗口中的某个个案数据。其操作步骤是：

(1) 选择将被删除的个案数据行。即将鼠标移动到将被删除的个案数据的号码列上，单击鼠标左键。于是，待删除的个案数据全部反向黑色显示。

(2) 单击鼠标右键，从弹出菜单中选择 Cut 选项。于是，待删除的个案数据被删除，其后面的数据自动依次往前提。

四、插入和删除一个变量

1. 插入一个新变量

在数据编辑过程中，有时需要在数据编辑窗口的某个变量之前或之后插入一个新变量。其操作步骤是：

(1) 将当前数据单元确定在某个变量的任何一个个案上。

(2) 菜单选项：Data－＞Insert variable。于是，计算机自动在当前数据单元所在变量列之前插入一个名为 var00001、变量类型为标准数值型的新变量。

(3) 根据实际情况修改新变量的属性。如：修改新变量的变量名和变量类型，加入变量名标签或变量值标签等。

2. 删除一个变量

有时需要将数据编辑窗口中的某些不再有用的变量删掉。其操作步骤是：

(1) 选择将被删除的变量列。即将鼠标移动到将被删除的变量列的变量名行上，单击鼠标左键。于是，待删除变量列上的数据全部反向黑色显示。

(2) 单击鼠标右键，从弹出菜单中选择 Cut 选项。于是，待删除的变量列数据被删除，其右面的变量数据自动依次往左移动。

五、数据单元中数据的移动、复制和删除

在对数据编辑窗口中的数据进行编辑时，有时往往希望对整块数据进行整体操作。如：希望把从某行某列到某行某列围成的整块数据——源数据块，整体地移动、复制到另一个数据单元——目标单元中，或者要将整块的数据清除。在实

现这些功能时，其三大操作步骤是：

（1）定义源数据块。源数据块是将要被移动或复制或清除的对象。定义源数据块的操作方法是：将鼠标移动到某数据块的左上角单元上，然后拖动鼠标至该数据块的右下角单元上。于是，被鼠标拖动过的单元全部反向黑色显示，表明一个数据块已经定义好了。

（2）单击鼠标右键。如果要清除整个数据块的内容，则选择弹出菜单中的Clear选项；如果要复制整个数据块内容，则选择弹出菜单中的Copy选项；如果要移动整个数据块内容，则选择弹出菜单中的Cut选项。

（3）指定目标单元。将鼠标移动到目标单元的左上角单元处，并指定它为当前单元。然后，单击鼠标右键，从弹出菜单中选择Paste选项。于是，源数据块数据就被复制或移动到目标单元中了。

§2.5 与其他计算机软件数据共享

上面，我们主要讲解了建立SPSS数据文件的一般方法，即：利用数据编辑窗口，定义SPSS数据文件结构，输入、修改、删除待分析数据。但是，在实际应用中，我们可能已经将一批待分析的数据存在了其他软件中。例如：我们可能已经利用数据库管理系统将一批数据存储管理在数据库文件中了。如果此时希望利用SPSS软件对这些数据进行统计分析，那么，最简捷的途径就是让SPSS去直接读取它们，也就是说，希望SPSS能够较方便地共享其他软件格式的数据。

一、共享文本格式文件数据

文本文件是计算机中最通用的一种格式的文件，其中的数据均以ASCII码方式存储。在内容上，文本文件中一般没有关于数据属性结构方面的描述信息，仅是纯数据部分。一般的计算机软件都有将自己格式的数据文件转成文本格式文件的功能。因此，如果SPSS能够较方便地读取这种文件，那么，它与其他软件共享数据的能力就无疑增强了许多。庆幸的是，SPSS软件具备了这种能力。

1. 文本文件的格式类型

为方便读取文本文件数据，SPSS规定：它默认的文本文件的扩展名应为.dat。此外，SPSS将文本文件中数据的排列方式也作了规定，即按数据的排列方式将文本文件分成固定格式的文本文件和自由格式的文本文件。

（1）固定格式（Fixed Columns）的文本文件。固定格式的文本文件的特点

是：一个个案数据可以占若干行；每个个案数据的每个变量值的起始列和终止列必须是固定不变的；数据项之间可以有分隔符，也可以没有。分隔符一般是逗号、空格等。

（2）自由格式（Free Field）的文本文件。自由格式的文本文件的特点是：一个个案数据可以占若干行；每个个案数据的每个变量值的起始列和终止列可以不固定；数据项之间必须有一个分隔符。分隔符可以是逗号、空格、Tab 键以及其他的用户自定义分隔符。

2. 读取固定格式的文本文件

SPSS 读取固定格式的文本文件，首先根据用户定义的数据项的起始列和终止列位置，到指定的文本文件列上取到一个数据；然后，将这个数据送入 SPSS 的数据编辑窗口。如此反复不断地从文本文件中读入一行数据，并依次将它们存放在 SPSS 同一行的各个列单元中，各单元列的变量名应是用户事先给定的。于是，固定格式文件中的一条数据就成为 SPSS 中的一个个案数据。这样，反复依次读取数据直到最后一条数据。

SPSS 读取固定格式文本文件的操作步骤是：

（1）菜单选项：File－>Open ASCII data－>Fixed Columns。出现如图 2—7 所示的窗口。

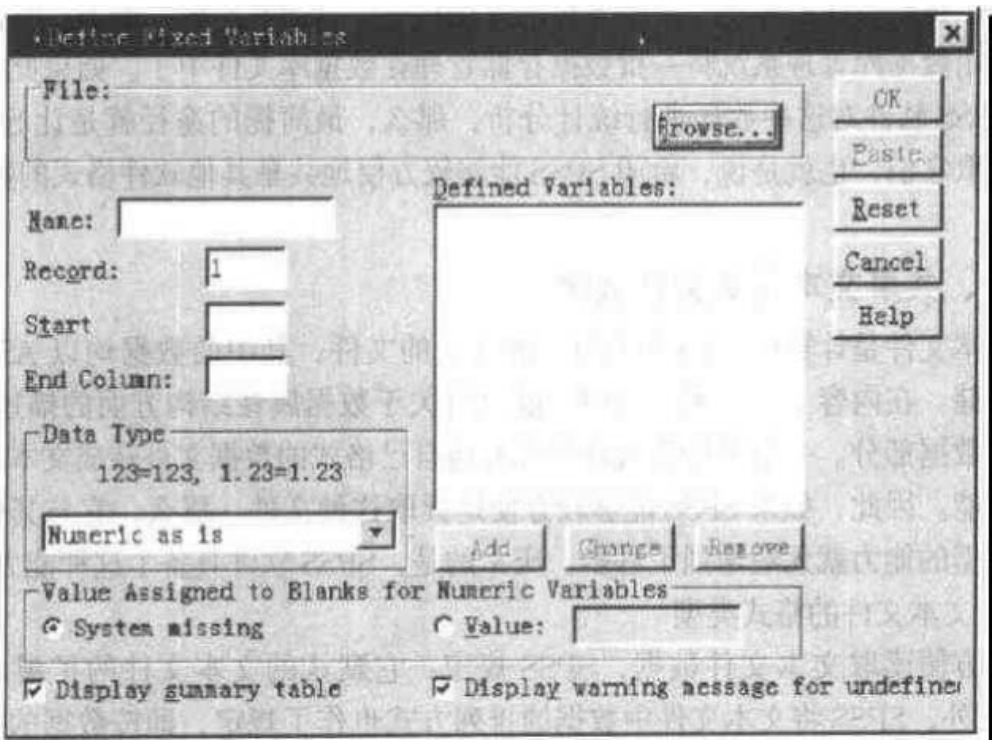


图 2—7　SPSS 读取固定格式的文本文件窗口

(2) 通过 Browse 按钮给出已经存储在磁盘上的待读取的文本文件的名称。

(3) 逐个给出每个数据项的相应参数：name：变量名，表示本数据项在SPSS中将以什么变量名存取；record：指定本数据项在文本文件中的相对行号。在文本文件中，一个个案数据可以分布在多行上，因此，这里要指明本数据项是在第几行上；start：给出在文本文件中本数据项起始列号；end column：给出在文本文件中本数据项终止列号；data type：在下拉框中选定本数据项的数据类型，默认的数据类型为标准数值型；missing value：指明本数据项是否有用户缺失值，缺失值是什么。

(4) 按动Add钮将上述定义信息加到 Define Variables 框中。

(5) 重复步骤 (3)、(4)，直到数据项定义完为止。

如果在定义过程中出现错误，可以通过Change钮实现修改，Remove钮实现删除。

需要注意的是：在读取固定格式的文本文件过程中，用户必须对各列数据项的实际含义以及起止列位置非常清楚。否则，就会出现数据和变量名之间的张冠李戴现象；或者因数据的列位置指定错误而出现不能正确读入数据的情况。

3. 读取自由格式的文本文件

SPSS 读取自由格式的文本文件，是以数据分隔符为界，自左至右地依次顺序读取数据，并将它们按读入的前后次序，依次送入 SPSS 的数据编辑窗口的各列单元中，从而形成 SPSS 数据的一个个案。其中各列变量的变量名应是事先由用户指定的。这样，反复不断读取直到最后一个数据。

SPSS 读取自由格式的文本文件的操作步骤是：

(1) 菜单选项：File－>Open ASCII data－>freefield。出现如图 2—8 所示的窗口。

(2) 通过 Browse 按钮给出已经存储在磁盘上的待读取的文本文件的名称。

(3) 定义在文本文件中，数据项之间是以什么作分隔符的。

(4) 逐个给出每个数据项的相应参数：name：变量名，表示本数据项在SPSS中将以什么变量名存取；data type：指明本数据项的数据类型是标准数值型还是字符型。

(5) 按动Add钮将上述定义信息加到 Define Variables 框中。

(6) 重复步骤 (4)、(5)，直到数据项定义完为止。

如果在定义过程中出现错误，可以通过Change钮实现修改，Remove钮实现删除。

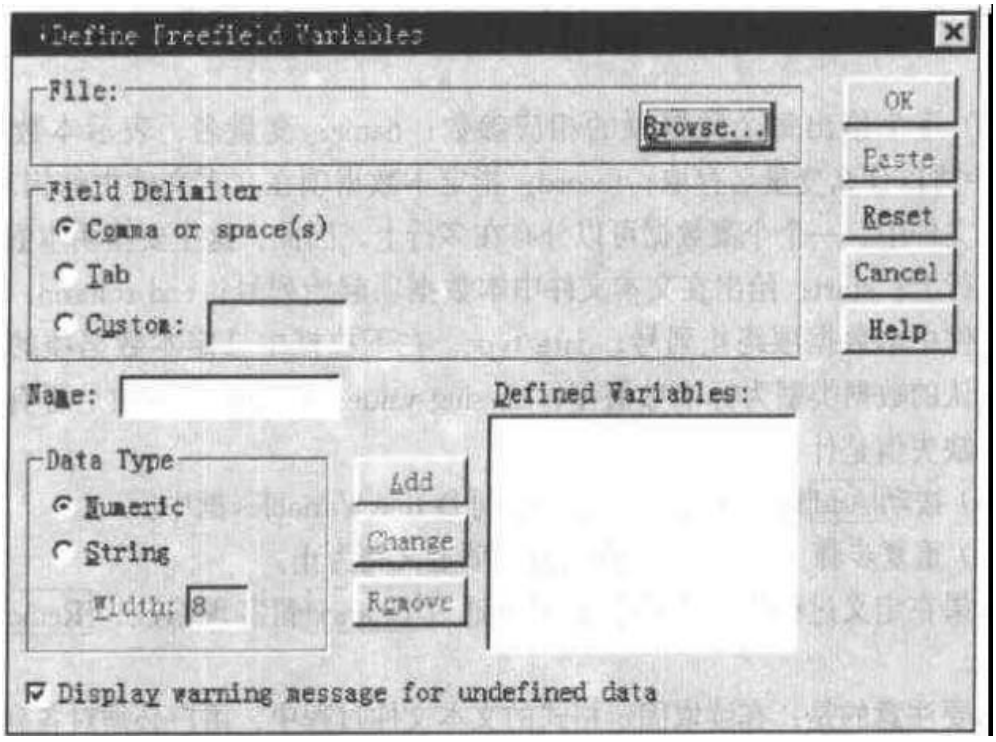


图 2—8　SPSS 读取自由格式的文本文件窗口

需要注意的是：在读取自由格式的文本文件过程中，用户必须要事先了解各列数据项的含义，以免发生数据与变量名不对应的错误。

可见，SPSS 读取自由格式的文件比读取固定格式的文件要简单方便，它不要求用户知道各数据项的起止列位置。因此，在实际工作中，我们往往以自由格式的方式读入数据。但是，在某些情况下，则只能按固定格式的文件读入数据。例如：如果只希望读取文本文件中的某几列数据，而不是全部列数据时，则只能按固定格式读入数据。因为，如果按自由格式读取数据，系统只能按照自左向右的次序，以分隔符作为数据项的分隔标志，依次顺序读入数据。这样，那些本不希望被读入的数据也就被顺序读进来了，必然会出现数据和变量名之间的张冠李戴现象。因此，这时必须按照固定格式的要求，依次定义那些希望被读入的数据项所在的起始列和终止列。于是，系统仅仅到用户指定的相应列位置上去读取数据，而那些没有被指定列位置的数据自然就不会被读入了。

二、共享 dbf 格式文件和 xls 格式文件

Dbase 和 Excel 是计算机中常用的存储、加工、管理数据的软件。Dbase 数据库文件中的数据以及 Excel 工作表中的数据分别以.dbf 和.xls 为扩展名存储在磁盘上。这些格式的数据文件的内容，除了纯数据部分外，还有一些关于数据属性

结构方面的描述信息。

SPSS 提供了与 Dbase 和 Excel 进行数据共享的接口，它能够很方便地将 Dbase 和 Excel 中的数据、共享到自己的数据编辑窗口中。操作步骤是：

(1) 菜单选项：Data -> Open。

(2) 给出待共享的数据文件的名称，并指定文件扩展名。

需要作几点说明：

● 如果共享的是 dbf 数据库文件：数据库文件中的字段名、字段类型自动会转成 SPSS 数据文件中的变量名和变量类型。dbf 中的一条记录为 SPSS 中的一个个案。

● 如果共享的是 xls 工作表文件：一般系统将默认读取 xls 工作表中的全部数据到 SPSS 数据编辑窗口中。但仍然可以允许指定仅读取工作表中某个区域内的数据。区域表示为：以冒号连接两个单元名和地址。如：a5:b10,表示仅读取以 a5 单元为左上角，b10 单元为右下角的矩形区域内的数据到 SPSS 中。工作表上的一行数据为 SPSS 中的一个个案。

● 如果 xls 工作表文件第一行或指定读取区域内的第一行上存储了变量名信息，则应选择 Read variable names 项，即以工作表第一行或指定读取区域内的第一行上的文字信息作为 SPSS 的变量名。否则，不应选择该选项，SPSS 的变量名将自动取名为工作表中的单元名，如：a、b、c 等。

第 3 章

SPSS 数据文件的基本加工和处理

在 SPSS 数据文件建立好之后，一般还需要对它们进行必要的加工处理。如：根据一定条件对原数据文件中的个案数据进行必要筛选抽取；或者，在原有数据的基础之上，计算产生一些含有更丰富信息的新变量数据等。因此，统计分析前的数据加工、处理是统计分析工作中不可缺少的一个关键步骤。

§ 3.1 数据文件的整理

一、个案排序

数据编辑窗口中的个案数据的前后排列次序是随机的，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个案数据。例如：在职工基本情况数据中，希望按工资收入从低到高的顺序，或者按职称从高到低的顺序来浏览数据。观察排序后的个案数据，会方便用户对数据的了解。

SPSS 中的个案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的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。当排序变量只有一个时的排序称为单值排序，即：按照排序变量取值的大小

次序对个案数据重新整理后显示。当排序变量是多个的排序为多重排序。多重排序中的第一个排序变量称为主排序变量，其他排序变量依次称为第二排序变量、第三排序变量等。在多重排序时，个案数据首先按主排序变量值的大小次序排序。然后，对那些具有相等主排序变量值的个案，再对它们按第二排序变量值的大小次序排序。接着，再对那些具有相等第二排序变量值的个案，按第三排序变量值的大小次序排序，依次排序下去。例如：表 3—1 显示的是职工基本情况数据以职称主排序变量降序、工资收入第二排序变量升序的多重排序结果。

表 3—1 **SPSS 个案多重排序举例**

职工号	职　称	工　资
001	高级工程师	1 014.00
011	高级工程师	1 014.00
003	高级工程师	1 044.00
009	工程师	859.00
018	工程师	879.00
020	工程师	879.00
021	工程师	879.00
014	工程师	889.00
002	工程师	984.00
012	工程师	989.00
010	助理工程师	827.00
022	助理工程师	827.00
029	助理工程师	827.00

个案排序的操作步骤：

(1) 菜单选项：Data -> Sort Cases。出现如图 3—1 所示的窗口。

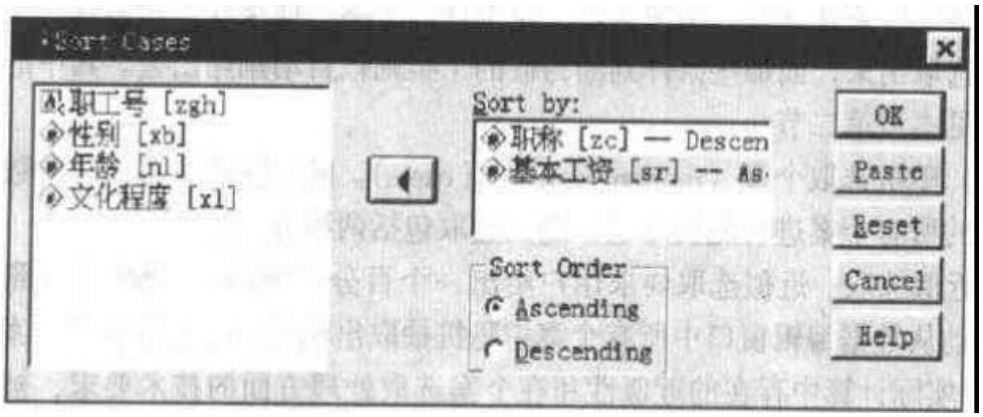


图 3—1　SPSS 个案排序窗口

(2) 指定主排序变量到 Sort by 框中。同时，点 Sort Order 框中的选项指出是按该变量值的升序排还是降序排序。Sort Order 框中的 Ascending 选项表示升序、Descending 表示降序。

(3) 如果是多重排序，在本步还应指定其他排序变量以及相应的排序规则。否则，本步可略。

于是，数据编辑窗口中的个案数据便自动按用户指定的顺序重新显示出来。

需要说明的是：

(1) 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。

(2) 在多重排序中，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。

(3) 排序以后，原来个案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个案标志的变量，如年份等，则应注意保存原数据的排列顺序，以免数据混乱。

二、个案选取

个案选取就是根据不同的需要，从 SPSS 中的个案筛选出希望的个案来。那些未被选取到的个案将不参与以后的所有操作、处理和分析，直到用户再次指定选取它们为止。

1. SPSS 的个案选取方式

在 SPSS 中提供了几种个案选取的方式：

(1) 选择符合一定条件的个案 (if condition is satisfied)。选取符合一定条件的个案要求用户给出一个 SPSS 条件表达式。SPSS 会自动对数据编辑窗口中的所有个案进行条件判断。那些满足给定条件的个案，即条件判断为真的个案，将被自动选取出来，而那些条件判断为假的个案则被自动剔除出去。其中的条件表达式参见本章第二节。

(2) 随机选取个案 (random sample of cases)。随机选取个案就是对数据编辑窗口中的所有个案进行随机筛选。随机选取包括两种方式：

● 近似选取：近似选取要求用户给出一个百分比数值。SPSS 将按照这个比例，自动从数据编辑窗口中所有个案中随机抽取出相应百分比数目的个案。由于 SPSS 在实际计算中存在的近似性和在个案选取处理方面的技术要求，被选取出的个案总数不一定就恰好精确地符合用户指定的百分比数目，会有小的偏差。但这对数据的分析并无大的影响。

● 精确选取：精确选取要求用户给出两个参数。第一个参数是希望选取的个案数，第二个参数是指定在前几个个案中选取。之后，SPSS 会自动在数据编辑窗口的前若干个个案中随机精确地抽出相应个数的个案来。

(3) 选取某一区域内的个案（based on time or case range)。选取某一区域内的个案就是抽取一定范围内的全部个案。它要求用户给出这个范围的上、下界的个案号码。这种个案选取的方式比较适用于时间序列的数据分析。

(4) 使用过滤变量选取个案（use filter variable)。使用过滤变量选取个案要求用户指定一个已存在的变量作为个案选取的标准。那些该变量值取值为 0 或系统缺失值的个案将被自动地剔除掉，而其余的个案将被自动选中。

2. 个案选取的操作步骤

个案选取的操作步骤是：

(1) 菜单选项：Data －＞ Select cases。出现如图 3—2 所示的窗口。

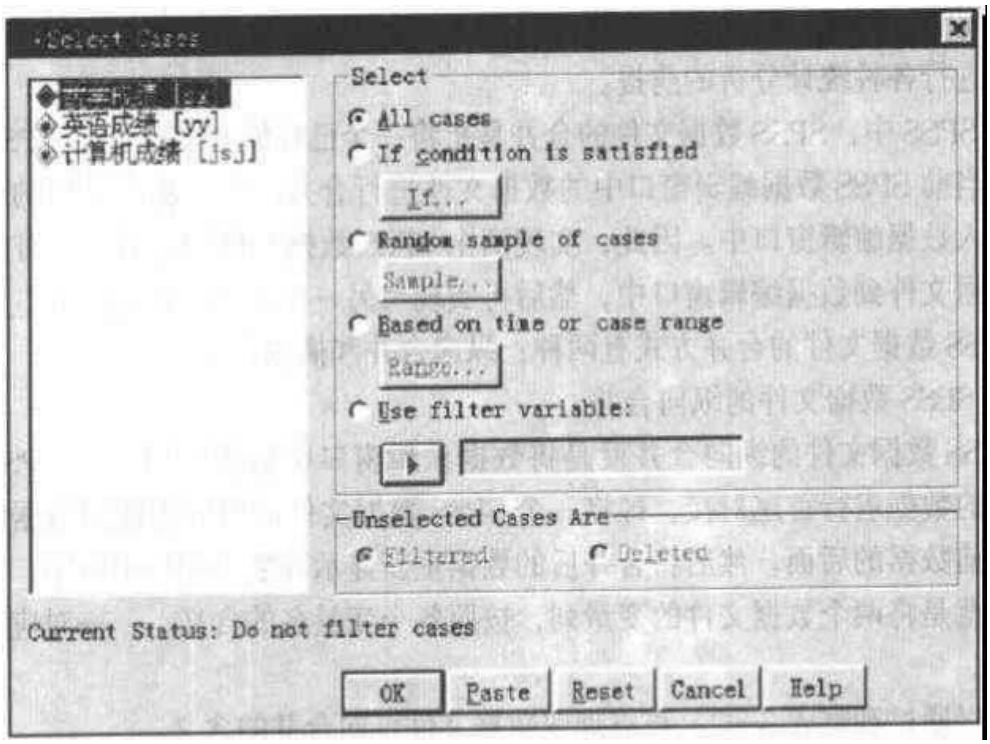


图 3—2　SPSS 个案选取窗口

(2) 指定按上述哪种选取方式进行个案选取。

(3) 指定对未选中个案的处理方式。其中：Filtered 表示在那些未被选中的个案号码上做一个“/”标记；Deleted 表示将那些未被选中的个案从数据编辑窗口中删除掉。一般采取前一个处理方式较好。

需要说明的是：

● 如果按上述方法进行了个案选取，则以后的 SPSS 分析操作仅针对那些被选中的个案进行，直到用户再次利用个案选取操作改变个案的选取为止。

● 采用前三种选取方式进行个案选取之后，SPSS 将在数据编辑窗口中自动生成一个名为 filter _ $ 的新变量。它是 SPSS 操作过程中产生的一个中间变量，取值为 0 或 1。0 表示本个案未被选中，1 表示本个案被选中了。同时，SPSS 自动为该变量添加了变量名标签，使用户能够很直接地了解到本次个案选取的方式。

三、数据文件的合并

进行统计分析的首要任务就是将待分析的数据录入到计算机中。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，然后再分别由几个人去录入，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对这份大数据进行各种统计分析的前提。

在 SPSS 中，SPSS 数据文件的合并是指将一个已存储在磁盘上的 SPSS 数据文件与当前 SPSS 数据编辑窗口中的数据文件进行合并，然后将合并后的数据再重新装入数据编辑窗口中。因此，实现两个 SPSS 数据文件的合并，应首先打开一个数据文件到数据编辑窗口中，然后再实现与另一个数据文件的合并。

SPSS 数据文件的合并方式有两种：纵向合并和横向合并。

1. SPSS 数据文件的纵向合并

SPSS 数据文件的纵向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，然后将合并后的数据重新显示在数据编辑窗口中。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。

可以通过观察表 3—2，直观理解数据文件纵向合并的含义。

表 3—2　　SPSS 数据文件的纵向合并

数据文件 1

职工号	工资收入	职　称
001	1 014.00	高级工程师
002	984.00	工程师
003	1 044.00	高级工程师

数据文件 2

职工号	工资收入	职　称
020	879.00	高级工程师
021	879.00	高级工程师
003	1 044.00	高级工程师

纵向合并后的数据文件

职工号	工资收入	职　称
001	1 014.00	高级工程师
002	984.00	工程师
003	1 044.00	高级工程师
020	879.00	工程师
021	879.00	工程师
003	1 044.00	高级工程师

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，最好起相同的变量名，变量类型也最好相同。这样，将方便 SPSS 对变量的自动对应和匹配。否则，变量名最好不起相同的名称。

SPSS 数据文件纵向合并的操作步骤：

(1) 在数据编辑窗口中打开一个待合并的 SPSS 数据文件。

(2) 菜单选项：Data－＞Merge File－＞Add Cases。于是计算机要求用户输入一个已存在于磁盘上的第二个待合并的 SPSS 数据文件名。输入正确后出现如图 3—3 所示的窗口。

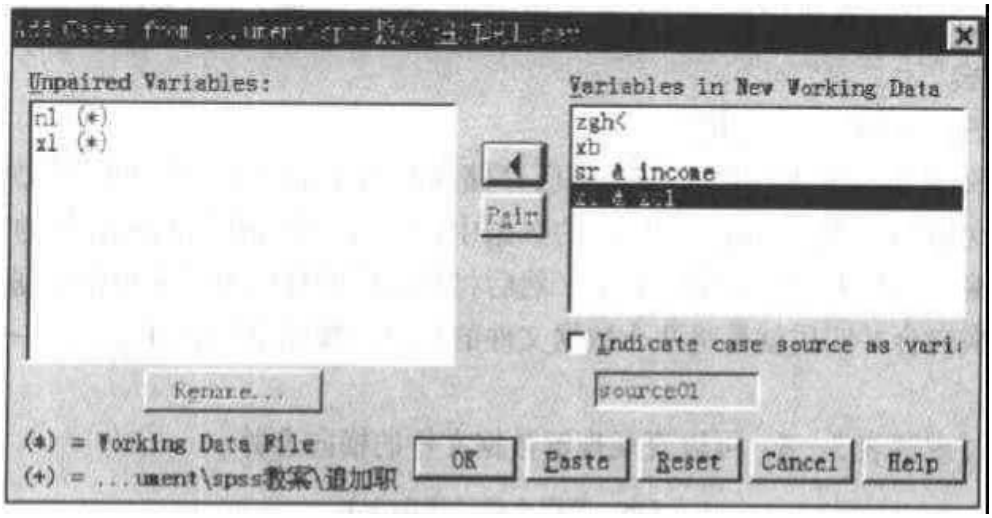


图 3—3　SPSS 数据文件纵向合并窗口

(3) 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 框中。它们被 SPSS 默认：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。用户如果不接受这种默认，可以按[<]钮

将它们剔出到 Unpaired Variables 框中。否则，本步可略。

（4）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（2）中用户指定的磁盘文件中的变量。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，它们被 SPSS 默认：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。用户如果不接受这种默认，可以手工选择两个变量名，按 pair 钮指定配对，表示它们具有相同的数据含义，并被选入 Variables in New Working data 框中。或者，先按 Rename 钮改名后再指定配对。或者，按 > 钮，指定某变量不经任何对应匹配，强行进入 Variables in New Working Data 框中。否则，本步可略。

（5）如果希望在合并后的数据文件中看出哪些个案来自合并前的哪个 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 source1 的变量，取值为 0 或 1。0 表示该个案来自第一个数据文件，1 表示该个案来自第二个数据文件。否则，本步可略。

于是，在数据编辑窗口中会自动显示合并后的数据，用户应根据实际需要将它存储到磁盘的 SPSS 数据文件中。如果其中有缺失数据，则是因为第（4）步中强行指定某变量进入而造成的。

从上述合并步骤可以看出，如果遵循我们前面提到的合并条件，将会大大简化数据文件合并的操作过程。因此，用户在大数据分别录入时就应统筹考虑变量名的取法。

2. SPSS 数据文件的横向合并

SPSS 数据文件的横向合并是将已存储在磁盘上的一个 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。横向合并实质就是将两个数据文件的个案，按照个案的对应，一一进行左右对接。

通过观察表 3—3，可以直观理解数据文件的横向合并。

表 3—3　　SPSS 数据文件的横向合并

数据文件 1

职工号	工资收入	职　称
001	1 014.00	高级工程师
002	984.00	工程师
003	1 044.00	高级工程师

数据文件 2

职工号	奖　金	年　龄
001	100.00	36
002	250.00	50
004	150.00	42

合并后的数据文件

职工号	工资收入	职　称	奖　金	年　龄
001	1 014.00	高级工程师	100.00	36
002	984.00	工程师	250.00	50
003	1 044.00	高级工程师	.	.
004	.	.	150.00	42

实现 SPSS 数据文件的横向合并应遵循三个条件：

第一，两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拼接的依据，称为关键变量。如：职工号、商品代码等。关键变量可以是多个。

第二，两个数据文件都必须事先按关键变量进行升序排序或多重升序排序。

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。

SPSS 数据文件横向合并的操作步骤：

(1) 在数据编辑窗口中打开一个待合并的 SPSS 数据文件。

(2) 菜单选项：Data－>Merge File－>Add Variables。于是计算机要求用户输入一个已存在于磁盘上的第二个待合并的 SPSS 数据文件名。输入正确后出现如图 3—4 所示的窗口。

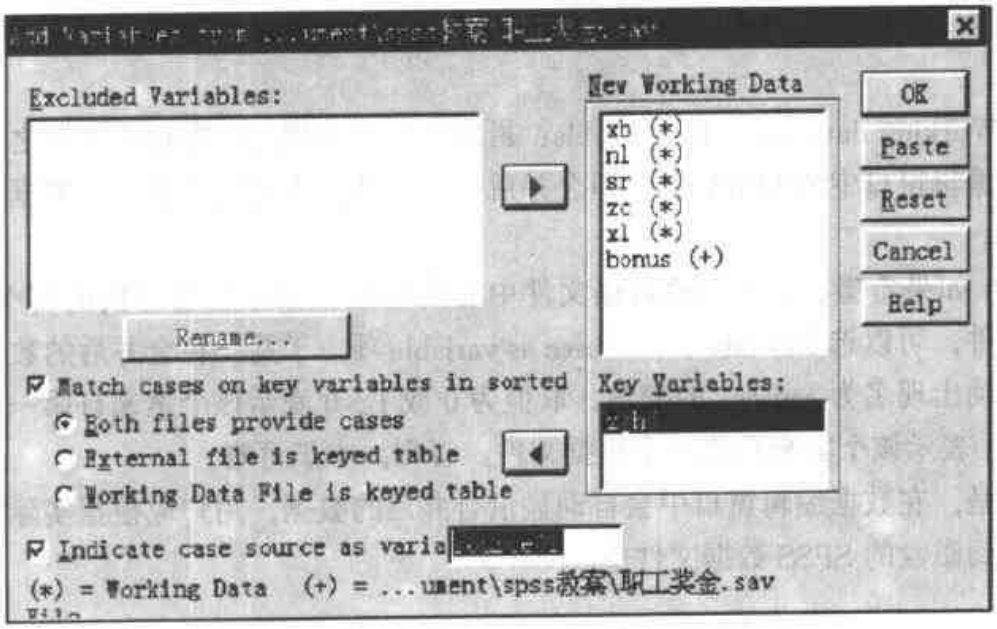


图 3—4　SPSS 数据文件横向合并窗口

(3) 两个待合并数据文件中的所有变量名出现在 New Working Data 框中。变量名后面有 * 或 + 号。* 表示该变量名是数据编辑窗口中的变量，+ 表示该变量名是 (2) 中用户指定的磁盘文件中的变量。它们被 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。用户如果不接受这种默认，可以按 < 钮将它们剔出到 Excluded Variables 框中。或者，剔出后，按 Rename 钮将变量改名，然后按 > 钮将它们从 Excluded Variables 框中重新以新名选回到 New Working Data 框中。否则，本步可略。

(4) 如果两个待合并的数据文件中的个案数据是横向顺序一一对应的，可按 OK 钮完成合并工作。否则，进入第 (5) 步。

(5) 两个待合并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 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。

(6) 指定合并后数据文件中的数据按哪种方式提供。SPSS 有三种数据的提供方式：

● Botb 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个案共同组成合并后的数据文件；

● External file is keyed table：指在数据编辑窗口中的数据基础之上，合并第二个数据文件中的变量数据，即合并后数据文件的个案仅是当前数据编辑窗口中的个案；

● Working data file is keyed table：指在第二个数据文件数据的基础之上，合并数据编辑窗口中的变量数据，即合并后数据文件的个案仅是第二个数据文件中的个案。

(7) 如果希望在合并后的数据文件中看出哪些个案来自合并前的哪个 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 source1 的变量，取值为 0 或 1。0 表示该个案来自第一个数据文件，1 表示该个案来自第二个数据文件。否则，本步可略。

于是，在数据编辑窗口中会自动显示合并后的数据，用户应根据实际需要将它存储到磁盘的 SPSS 数据文件中。

四、数据文件的转置

数据文件的转置就是将数据编辑窗口中数据的行列互换，即将个案转为变

量，将变量转为个案后，重新显示在数据编辑窗口中。

观察表 3—4 数据直观理解数据文件转置的含义。

表 3—4　　SPSS 数据文件的转置

转置前的数据文件

职工号	性　别	年　龄	收　入
001	男	48	1 014.00
002	男	49	984.00
003	女	54	1 044.00
004	女	41	866.00

转置后的数据文件

Case-lbl	V001	V002	V003	V004
职工号	001	002	003	004
性　别	男	男	女	女
年　龄	48	49	54	41
收　入	1 014.00	984.00	1 044.00	866.00

数据转置的操作步骤是：

(1) 菜单选项：Data－＞Transpose。出现如图 3—5 所示的窗口。

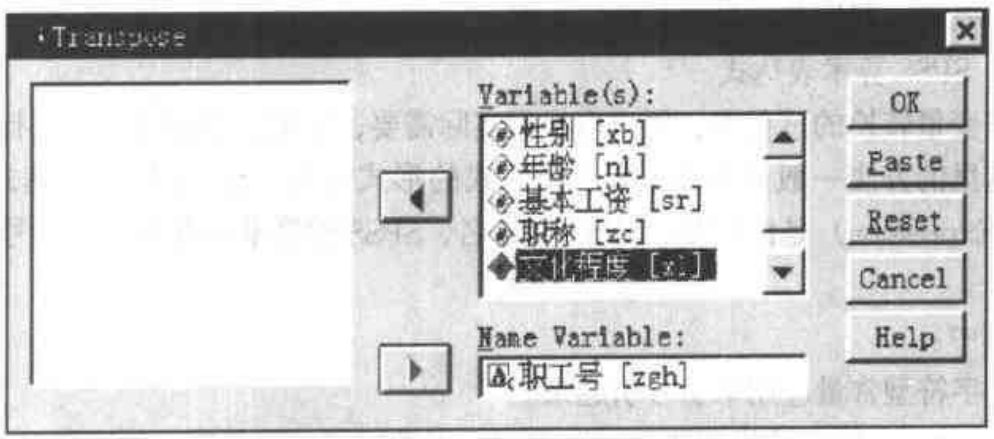


图 3—5　SPSS 数据文件转置窗口

(2) 指定转置的数据文件中保留了原文件中的哪些变量，将它们选入 Variables 框中。

(3) 指定转置后数据文件中各个变量名如何取名。选择一个取值惟一的标记

变量到 Name Variable 框中。如：职工号。于是转置后的数据文件各变量的取名为：以字母 V 开头，加上该标记变量的变量值。如：V001、V002、V003 等。如果略去本步，则新数据各变量的名称默认为 VAR00001、VAR00002、VAR00003 等。

于是，数据编辑窗口中将显示数据文件转置后的数据。

需要说明的是：数据文件转置后，SPSS 会自动产生一个名为 Case-lbl 的新变量，用来存放原数据文件的各变量名。

§3.2 数据加工

一、变量转换

变量转换就是在原有 SPSS 数据文件的基础之上，根据用户给出的 SPSS 算术表达式以及函数，对所有个案或满足 SPSS 条件表达式的个案，计算产生出一系列结果，并将结果存到一个用户指定的 SPSS 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。

例如：收集到学生的数学、英语、计算机三门课程的成绩，现在需要在原有数据的基础之上，计算前 10 名学号的学生的平均成绩。这里，前 10 名学生参与计算，计算公式就是平均成绩的计算公式，计算结果应放在一个新变量中。

在变量转换中，涉及到了 SPSS 算术表达式、SPSS 函数、SPSS 条件表达式等基本概念，我们首先简单讨论这些概念。

1. SPSS 算术表达式

在变量转换的过程中，用户应根据实际需要，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。

其中：

- 字符型常量应用单引号引起来。
- SPSS 变量名是指那些已经存在于数据编辑窗口中的现有的变量名。
- SPSS 的算术运算符主要包括：+（加）、-（减）、*（乘）、/(除)、**(乘方)，它们操作的数据对象应为数值型变量。它们运算的先后次序是：先计算乘方，再计算乘、除，最后计算加、减。在同级运算中，按从左往右的顺序进行计算。如果要改变这种运算次序，可以通过圆括号来实现，即：如果有圆括号应

首先计算圆括号中的式子。例如：计算三门课程的平均成绩的 SPSS 算术表达式为：

（数学 + 英语 + 计算机）/ 3

需要注意的是：

（1）SPSS 算术表达式中的常量及变量的类型应该是一致的。否则，SPSS 无法计算。

（2）在 SPSS 中，算术表达式的计算是针对每个个案进行的。因此，得到的计算结果必然是多个，每个个案都有自己的计算结果。

（3）存放 SPSS 算术表达式结果的变量，其变量类型应与计算结果的数据类型相一致。

2．SPSS 函数

SPSS 函数是由程序员事先编好并存储在 SPSS 软件中的一段计算机程序。这些程序段都有各自的名字，称为函数名。它们能够实现某些特定的计算任务。通过执行这些程序段得到的计算结果称为函数值。用户在使用这些函数时，只需通过书写相应的函数名，并给出一些必要的计算参数，计算机便会自动计算出函数值。

需要注意的是：

● SPSS 中函数的计算是针对每个个案进行的，每个个案都有自己的函数运算结果。

● SPSS 函数值的应存放到一个指定变量中，该变量的变量类型应与函数值的数据类型相一致。

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，有时也可以是多个；有时可以是常量（字符型常量应用单引号引起来），有时也可以是变量名或 SPSS 的算术表达式。

SPSS 函数一般也会与 SPSS 的算术表达式混合出现。例如：((ABS（数学 - MEAN（数学，英语，计算机）） + ABS（英语 - MEAN（数学，英语，计算机）） + ABS（计算机 - MEAN（数学，英语，计算机）))/3。

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。

（1）算术函数。算术函数主要完成一些特定的算术计算功能。函数值为数值型，参数也为数值型。见表 3—5 所示。

表 3—5　　常见的算术函数

函数名	功　能	举　例
ABS（数值表达式）	计算数值表达式的绝对值	ABS（数学－英语）：每个个案的函数值均为：各自数学与英语差的绝对值
SQRT（正数值表达式）	计算正数值表达式的平方根	SQRT（4）：每个个案的函数值均为：2
SIN（弧度单位的角度数）	计算正弦值	SIN（30＊3.14/180）：每个个案的函数值均为：0.5
COS（弧度单位的角度数）	计算余弦值	COS（60＊3.14/180）：每个个案的函数值均为：0.5
EXP（数值表达式）	计算 e 的数值表达式次幂	EXP（5）：每个个案的函数值均为：148.41
LN（数值表达式）	计算以 e 为底的自然对数值	LN（5）：每个个案的函数值均为：1.61
LG10（数值表达式）	计算以 10 为底的对数值	LG10（5）：每个个案的函数值均为：0.7
RND（数值表达式）	得到数值表达式四舍五入后的整数	RDN（2.66）：每个个案的函数值均为：3
TRUNC（数值表达式）	得到截去数值表达式小数部分后的整数	TRUNC（4.7）：每个个案的函数值均为：4
MOD（数值表达式，常数）	得到数值表达式除以常数后的余数	MOD（20，3）：每个个案的函数值均为：2

（2）统计函数。统计函数一般计算基本的描述统计量，函数值为数值型，参数也为数值型。见表 3—6 所示。

表 3—6　　常见的统计函数

函数名	功　能	举　例
MEAN（数值表达式，数值表达式，…）	计算出多个数值表达式的平均值	MEAN（数学，英语，计算机）：对每个个案都算出成绩平均值
SD（数值表达式，数值表达式，…）	计算出多个数值表达式的标准差	SD（数学，英语，计算机）：对每个个案都计算出成绩标准差
VARIANCE（数值表达式，数值表达式，…）	计算出多个数值表达式的方差	VARIANCE（数学，英语，计算机）：对每个个案都计算出各成绩间的方差

续前表

函数名	功　能	举　例
SUM（数值表达式，数值表达式，…）	计算出多个数值表达式的总和	SUM（数学，英语，计算机）：对每个个案都计算出各成绩总和
CFVAR（数值表达式，数值表达式，…）	计算出多个数值表达式的变异系数（标准差/均值）	CFVAR（数学，英语，计算机）：对每个个案都计算出各成绩间的变异系数
MAX（数值表达式，数值表达式，…）	得出多个数值表达式中的最大值	MAX（数学，英语，计算机）：对每个个案都计算出各成绩中的最高分
MIN（数值表达式，数值表达式，…）	得出多个数值表达式中的最小值	MIN（数学，英语，计算机）：对每个个案都计算出各成绩中的最低分

（3）分布函数。分布函数用来随机产生一个服从某种统计分布的数据序列。函数值为数值型，这个序列的数据被随机放置在每个个案的指定变量列中。见表3—7所示。

表 3—7　　常见的分布函数

函数名	功　能	举　例
NORMAL（正数值型数据）	产生一个正态分布的随机序列，其均值为0，标准差为指定的数据	NORMAL(1)：产生一个服从标准正态分布的随机序列
UNIFORM（数值型数据）	产生一个均匀分布的随机序列，其最小取值为0，最大取值为指定的数据	UNIFORM(1)：产生一个服从均匀分布的随机序列，其取值范围是：0至1
RV. 分布名（参数，…）	产生一个服从指定统计分布的随机序列	RV.NORMAL(10,5)：产生一个服从均值为10、标准差为5的正态分布随机序列 RV.T（10）：产生一个有10个自由度的学生分布随机序列
CDFNORM（数值型数据）	计算标准正态分布序列中小于指定数值的累计概率值。对应于PROBIT函数	CDFNORM(0)：得到标准正态分布序列中小于0的累计概率，结果为0.5
PROBIT（概率值） 说明：0<＝概率值＜＝1	计算标准正态分布对应于指定累计概率值的数值	PROBIT(0.5)：得到标准正态分布序列中累计概率为0.5的数值，结果为0
CDF. 分布名（数值，参数，…）	产生一个服从指定统计分布的随机序列，并计算出序列值小于指定数值的累计概率值。对应于IDF函数	CDF.NORMAL(0,0,1)：得到标准正态分布序列中小于等于0的累计概率，结果为0.5

续前表

函数名	功　能	举　例
IDF. 分布名（概率值，参数，…） 说明：0<＝概率值<＝1	产生一个服从指定统计分布的随机序列，并计算出对应于指定累计概率值的数值。对应于 CDF 函数	IDF.NORMAL(0.5,0,1)：得到标准正态分布序列中累计概率为 0.5 的数值，结果为 0

（4）逻辑函数。逻辑函数用来进行逻辑判断。逻辑函数的函数值有两个取值：如果判断结果为真，则函数值为 1；如果判断结果为假，则函数值为 0。见表 3—8 所示。

表 3—8　　常见的逻辑函数

函数名	功　能	举　例
RANGE（参数 1，参数 2，参数 3，…）	判断参数 1 是否在参数 2 至参数 3 范围之间（包括参数 2 和参数 3）	RANGE（数学：80，90）：判断数学成绩是否在 80 至 90 分之间，结果依个案值而定 RANGE（'AA'，'AB'，'AZ'）：判断字符串 AA 是否在字符串 AB 至 AZ 的范围内，结果为 0
ANY（参数 1，参数 2，参数 3，…）	判断参数 1 是否出现在参数 2 或参数 3 中	RANGE（数学：80，90）：判断数学成绩是否为 80 分或 90 分，结果依个案值而定 RANGE（'AA'，'XX'，'DD'）：判断字符串 AA 是否出现在 XX 或 DD 中，结果为 0

（5）字符串函数。字符串函数用来对字符型数据进行处理。字符串函数的参数和函数值有时为字符型，有时也可以是数值型。见表 3—9 所示。

表 3—9　　常见的字符串函数

函数名	功　能	举　例
CONCAT（串 1，串 2，…）	将串 1 和串 2 等首尾相接	CONCAT（'AB'，'CD'）：将字符串 AB 和 CD 首尾相接，结果为 ABCD
INDEX（串 1，串 2） 说明：串 1 应长于串 2	得到串 2 在串 1 中第一次出现的字符位置。如果没出现则结果为 0	INDEX（'ABCDEF'，'CDE'）：找到字符串 CDE 在字符串 ABCDEF 中第一次出现的位置，结果为 3

续前表

函数名	功　能	举　例
RINDEX（串1，串2） 说明：串1应长于串2	得到串2在串1中最后一次出现的字符位置。如果没出现则结果为0	RINDEX（'ABCDCD'，'CD'）：找到字符串CD在字符串ABCDCD中最后一次出现的位置，结果为5
LENGTH（字符串）	得到字符串的总字符数	LENGTH（'ABCD'）：得到字符串ABCD的总字符数，结果为4
LOWER（字符串）	将字符串中的所有字符都转换成小写字符	LOWER（'ABCD'）：将字符串ABCD转换成小写字符，结果为abcd
UPCASE（字符串）	将字符串中的所有字符都转换成大写字符	UPCASE（'abcd'）：将字符串abcd转换成大写字符，结果为ABCD
LPAD（字符串1，数值，字符） 说明：1<=数值<=255	将字符串1左补若干个指定字符后使其字符长度达到指定的数值	LPAD（'AB'，5，'X'）：将字符串AB左补3个字符X，结果为XXXAB
RPAD（字符串1，数值，字符） 说明：1<=数值<=255	将字符串1右补若干个指定字符后使其字符长度达到指定的数值	RPAD（'AB'，5，'X'）：将字符串AB右补3个字符X，结果为ABXXX
LTRIM（字符串）	将字符串前的空格全部删掉	LTRIM（'　　ABC'）：删掉字符串ABC的前两个空格
RTRIM（字符串）	将字符串尾部的空格全部删掉	RTRIM（'ABC　　'）：删掉字符串ABC尾部的两个空格
NUMBER（字符串，格式） 说明：字符串应为数字字符串。格式以字符f开头，见例	将字符串按照格式要求转化成数值。如果字符串不能转换，则结果为系统缺失值	NUMBER（'123456'，f5.2）：将字符串12345转换成总长度为5（不包括小数点）、2位小数的数值型数据，结果为123.45
STRING（数值型数据，格式） 说明：格式以字符f开头，见例	将数值型数据转换成字符型数据	STRING（123.45，f5.1）：将数值型数据123.45转换成总长度为5（不包括小数点）并取1位小数位，结果为123.5
SUBSTR（字符串，数值1，数值2）	取子串，从字符串的数值1位置开始取数值2个字符串	SUBSTR（'ABCDE'，2，3）：取ABCDE第2个字符开始的3个字符，结果为BCD

（6）日期时间函数。日期时间函数主要对日期和时间进行处理。日期时间函数的函数值类型为日期型或数值型。见表3—10所示。

表 3—10　　常见的日期时间函数

函数名	功　能	举　例
DATE.DMY(日,月,年) 说明：函数名中的 D、M、Y 分别表示日期、月份、年份	修改一个已经定义好的日期型变量的值	DATE.DMY（31，12，1999）：将一日期型变量的值修改为 1999 年 12 月 21 日
DATE.QYR（季，年） 说明：函数名中的 Q 表示季度	将季度转化成相应月份后，修改一个已经定义好的日期型变量的值	DATE.QYR（4，1999）：将一日期型变量的值修改为 1999 年 10 月 1 日
DATE.YRDAY（年，天数）	根据给定的年和天数，转化成相应的日期后，修改一个已经定义好的日期型变量的值	DATE.YRDAY（1999，32）：将一日期型变量的值修改为 1999 年 2 月 1 日
XDATE.MDAY（日期型变量）	求出日期型变量值所对应的日期是该月中的第几天	XDATE.MDAY（DATA.DMY（31，12，1999））：结果为 31
XDATE.JDAY（日期型变量）	求出日期型变量值所对应的日期是该年中的第几天	XDATE.JDAY（DATA.DMY（3，02，1999））：结果为 34
XDATE.WEEK（日期型变量）	求出日期型变量值所对应的日期是该年中的第几周	XDATE.WEEK（DATA.DMY（3，02，1999））：结果为 5
TIME.DAYS（天数）	计算给定天数所对应的小时数，结果应赋给一个为时间格式的日期型变量	TIME.DAYS（1.5）：结果为 36:00:00
TIME.HMS（小时，分钟，秒）	修改一个已经定义好的时间格式的日期型变量的值	TIME.HMS（10，10，0）：将一时间格式的日期型变量的值赋成10:10:00
CTIME.DAYS（时间格式的日期型变量）	得到时间格式的日期型变量值所对应的天数	CTIME.DAYS（TIME.HMS（12，0，0））：结果为 0.50

（7）缺失值函数。缺失值函数对每个个案数据判断是否为缺失值。见表 3—11 所示。

表 3—11　　常见的逻辑函数

函数名	功　能	举　例
MISSING（变量名） 说明：该变量必须是数值型变量	对每个个案判断指定变量是否含有系统缺失值或用户缺失值	MISSING（数学）：对每个个案判断数学这个变量是否含有系统缺失值或用户缺失值。1 表示含缺失值，0 表示不含
SYSMIS（变量名） 说明：该变量必须是数值型变量	对每个个案判断指定变量是否含有系统缺失值	MISSING（数学）：对每个个案判断数学这个变量是否含有系统缺失值。1 表示含缺失值，0 表示不含

续前表

函数名	功　能	举　例
NMISS（变量名 1，变量名 2，…）	对每个个案计算在指定变量中有几个变量含有系统缺失值或用户缺失值	NMISS（数学，英语）：对每个个案计算数学和英语这两变量中有几个变量含有缺失值
VALUE（变量名）	忽略用户缺失值，即将用户缺失值看成是普通的数据	VALUE（数学）：忽略数学这个变量中定义的用户缺失值

（8）其他函数。见表 3—12 所示。

表 3—12　　其他函数

函数名	功　能	举　例
LAG（变量名、n）	产生一个新变量，该变量的前 $n-1$ 个数据为系统缺失值，第 n 个以后的数据依次为指定变量的数据值，指定变量的后 $n-1$ 个数据不存储在新变量中。即将指定变量后移 n 期后的结果存入新变量，从而方便时间序列中数据的差分计算	LAG（国民生产总值，1）：对历年的国民生产总值变量数据后移 1 期

3．SPSS 的条件表达式

通过 SPSS 的算术表达式和函数可以对数据编辑窗口中的所有个案计算产生出一系列相应结果。如果仅希望对数据编辑窗口中的部分个案进行计算，则应当利用 SPSS 的条件表达式指定对哪些个案进行计算。

SPSS 的条件表达式就是一个对条件进行判断的式子。其结果有两种取值：如果判断条件成立，则结果为真。反之，如果判断条件不成立，则结果为假。

SPSS 条件表达式中的条件可以分成简单条件和复杂条件，于是条件表达式就有简单条件表达式和复杂条件表达式两类。

（1）简单条件表达式。简单条件表达式就是由 SPSS 的关系运算符、常量、SPSS 变量以及算术表达式和函数组成的式子。其中的关系运算符包括：＞（大于）、＜（小于）、＝（等于）、～＝（不等于）、＞＝（大于等于）、＜＝（小于等于）六种。如：数学＞80，就是一个简单的条件表达式。对于那些数学成绩大于 80 分的个案，条件判断的结果就为真，而对那些数学成绩等于或低于 80 分的个案，条件判断的结果就为假。

（2）复杂条件表达式。复杂条件表达式又称逻辑表达式，它是由 SPSS 的逻辑运算符号、常量、SPSS 变量、圆括号以及简单条件表达式等组成的式子。其中的逻辑运算符包括：& 或 AND（并且）、|或 OR（或者）、～或 NOT（非）三种。NOT 的运算最优先，其次是 AND，最低的是 OR。可以通过使用圆括号来

改变这种运算次序。如：（数学＞＝90）and not（英语＜90），就是一个复杂条件表达式。对于那些数学成绩大于等于 90 分并且英语成绩不低于 90 分的个案，条件判断的结果就为真，而其他的个案，条件判断的结果就为假。

用户根据实际需要构造条件表达式之后，SPSS 会从数据编辑窗口中的所有个案中自动挑出判断结果为真的个案，即挑选出满足一定条件的个案，然后对它们再进行其他的计算处理。

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，系统就会根据用户的要求仅对满足一定条件的个案进行计算处理。

4．变量转换的操作步骤

变量转换的操作步骤是：

（1）菜单选项：Transform－＞Compute。出现如图 3—6 所示的窗口。

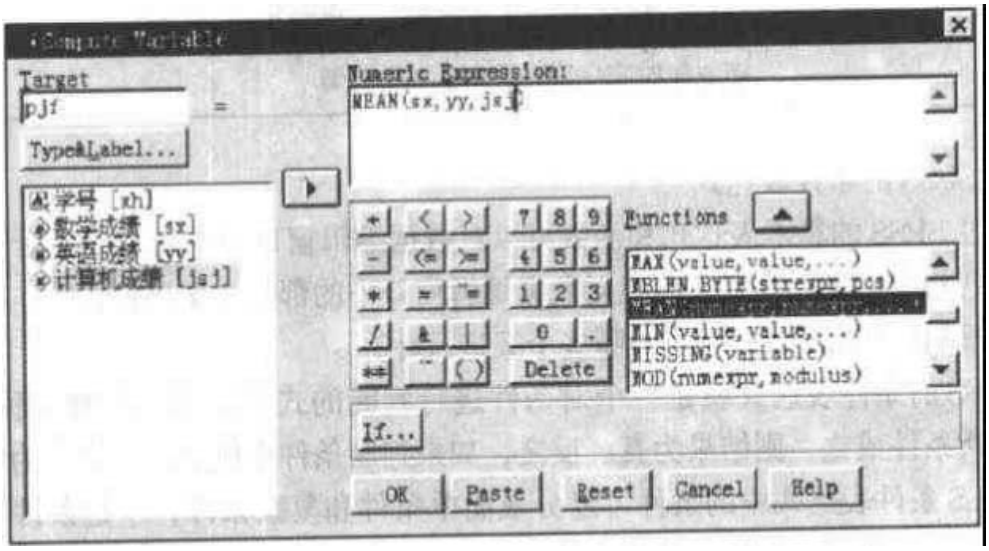


图 3—6　SPSS 变量转换窗口

（2）在 Numeric Expression 框给出 SPSS 算术表达式和函数。可以手工输入，也可以按动屏幕中的按钮以及函数的下拉菜单来完成表达式、函数的输入工作。

（3）在 Target 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，按 Type&Label 钮来修改变量的类型，或对新变量加变量名标签信息。

（4）如果用户仅希望对符合一定条件的个案计算进行变量转换，则按 IF 钮。出现如图 3—7 所示的窗口。点 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的下拉菜单来实现条件表达式的输入工作。否则，本步骤略去。

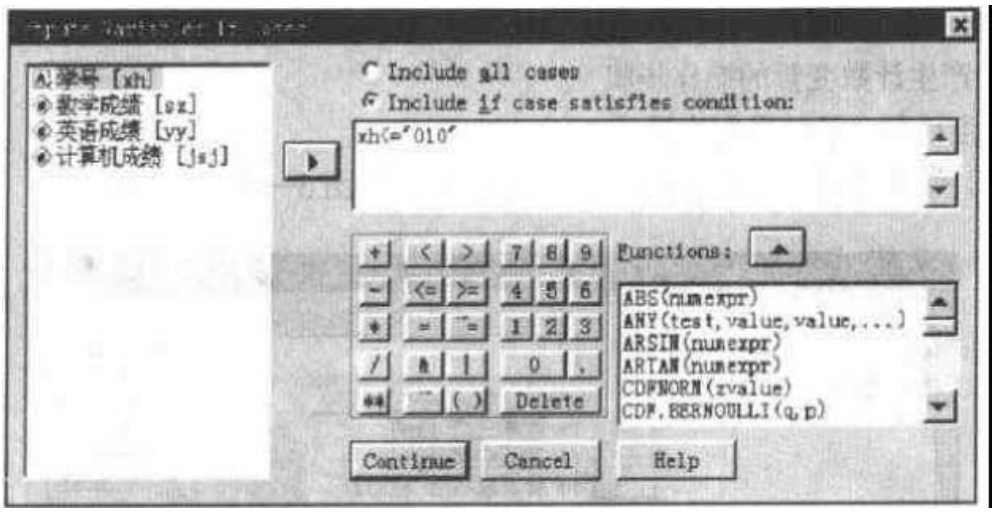


图 3—7　SPSS 条件表达式输入窗口

于是，如果指定存放计算结果的变量为新变量，系统会自动创建它，并放到数据编辑窗口中去。如果指定产生的变量已经存在，则会以计算出的新值覆盖旧值。对那些不满足指定条件的个案，新变量的取值为系统缺失值，已存在变量仍取计算前的旧值。

二、产生计数变量

在实际工作中，往往要做一些计数工作。产生计数变量就是实现计数功能，它对所有个案或满足一定条件的个案，计算若干个变量中有几个变量的值落在指定的区间内，并将计数结果存入一新变量中。

例如：对每个学生统计他们的数学、英语、计算机三门课程成绩中有几门在 90 分以上。这里，所有个案都参与计算，指定的区间在 100 分到 90 分之间。

产生计数变量应首先对区间进行定义。

1．区间的指定

区间的指定是产生计数变量的关键。这里的区间是一个广义的概念，对它的描述可以为：

(1) 一个数据点，包括：一个特定的值（Value）、系统缺失值（System-missing）、系统和用户缺失值（System-missing and User-missing）。

(2) 一个给出了最大值和最小值的闭区间。

(3) 一个没有给出最小值，仅给出最大值的开区间（Lowest through n），即：小于等于 n 的区间。

(4) 一个仅给出最小值，没有给出最大值的开区间（n through highest），即

大于等于 n 的区间。

2. 产生计数变量的操作步骤

产生计数变量的操作步骤是：

(1) 菜单选项：Transform－>Count。出现如图 3—8 所示的窗口。

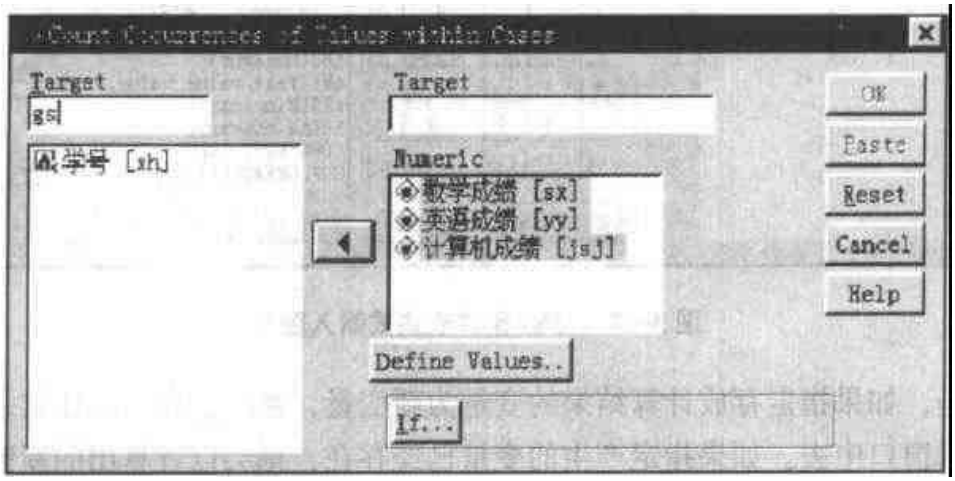


图 3—8　SPSS 产生计数变量窗口

(2) 指定将对哪些变量产生计数变量，将变量名送到 Numeric 框中。

(3) 在 Target 框中输入存放计数结果的变量名。

(4) 按 Define Value 钮来定义区间。出现如图 3—9 所示的窗口。

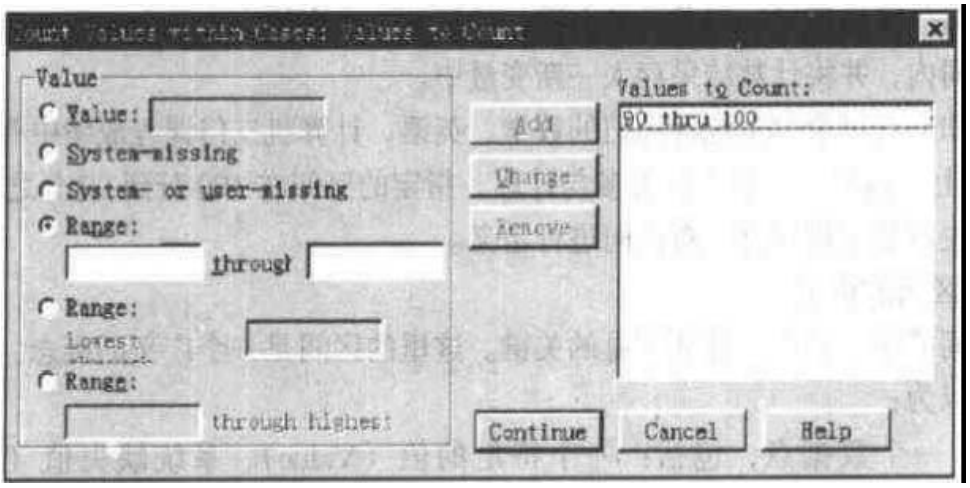


图 3—9　SPSS 区间定义窗口

(5) 如果仅希望对满足一定条件的个案产生计数变量，则按 IF 钮并输入一个条件表达式。否则，本步略去。

三、数据文件的分类汇总

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：对于职工基本情况的数据，现希望了解不同职称职工的平均工资情况。这就需要首先对数据按不同职称分类，然后再分别求出各类职工的工资收入平均值。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，要求用户指定分类变量和汇总变量；然后，计算机自动根据分类变量的若干个不同取值对个案数据分成若干类，并对每类个案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，可以观察表 3—13 数据。

表 3—13　　SPSS 数据文件分类汇总

原始数据文件

职工号	工　资	职　称	职工号	工　资	职　称
001	1 014.00	高级工程师	014	889.00	工程师
011	1 014.00	高级工程师	002	984.00	工程师
003	1 044.00	高级工程师	012	989.00	工程师
009	859.00	工程师	010	827.00	助理工程师
018	879.00	工程师	022	827.00	助理工程师
020	879.00	工程师	029	827.00	助理工程师
021	879.00	工程师			

按职称分类汇总后的数据文件

职称	工资_1
高级工程师	1 024.00
工程师	908.29
助理工程师	827.00

可见，分类汇总产生的 SPSS 数据文件的个案数取决于分类变量的不同类取值个数。这里，分类变量职称有三类取值，则按职称分类汇总后的数据就有三条。

数据文件分类汇总的操作步骤：

(1) 菜单选项：Data －＞ Aggregate。出现如图 3—10 所示的窗口。

(2) 指定分类变量到 Break 框中。

(3) 指定汇总变量到 Aggregate Variable (s) 框中。

(4) 按 Function 钮指定对汇总变量计算哪些描述统计量。SPSS 默认对不同类分别计算汇总变量的均值。

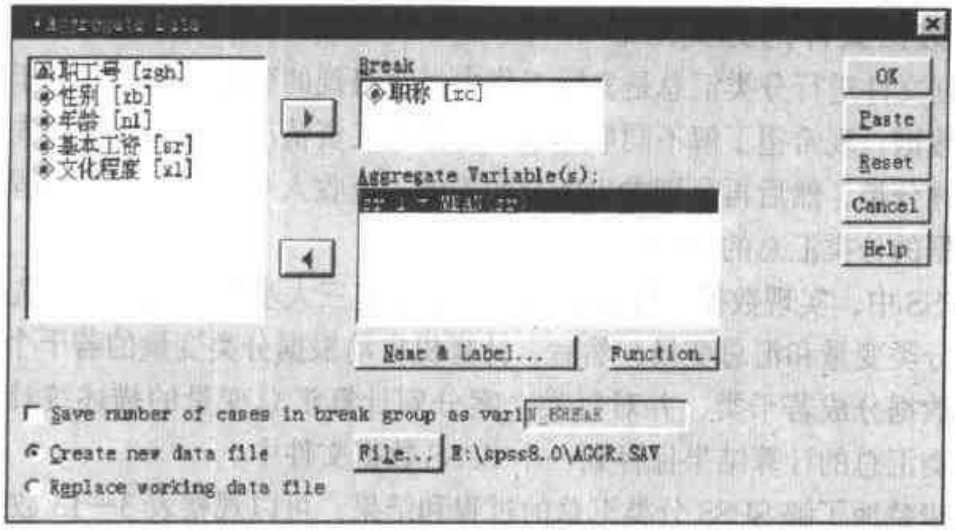


图 3—10　SPSS 数据文件的分类汇总窗口

(5) 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 钮重新指定结果文件名。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。

(6) 按 Name & Label 钮重新指定结果文件中的变量名或加变量名标签。SPSS 默认的结果文件的变量名为原变量名后加_ 1。

(7) 如果希望在结果数据文件中保存各分类组的个案数，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。可以修改该变量名。

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。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。

● 类似于个案排序功能，在多重分类汇总中，多个分类变量的前后指定次序是很关键的，它决定了汇总的先后次序。例如：如果希望了解不同职称中的不同性别的职工工资收入情况，则应指定职称为主分类变量，性别为第二分类变量。

§3.3　数据分组

在实际工作中，人们经常需要对统计数据进行必要的整理，其中统计数据的分组则是统计数据整理的必要环节。SPSS 专门提供了数据分组的功能。

数据分组就是将个案数据概括成若干个组，是对原始数据的进一步概括和总结。

例如：已经收集到了关于职工基本情况的数据。其中：工资收入数据项的内容是具体的实际工资数。如果现在希望了解职工中高、中、低收入的分布情况，显然，直接利用原始数据就太细了。因此，需要首先将工资收入分成高、中、低三个区间；然后，所有个案将根据各自的具体工资值被分派到相应的区间中，形成三个组；最后，再对高、中、低三组中个案的分布情况作统计。

可见，数据分组中的主要步骤是：首先，指定按哪个变量分组，即指定分组变量；其次，定义分组变量的分组区间；最后，指定一个存放分组结果的标志变量，标志每个个案所属组的组号。其中，定义分组区间是最关键的。

SPSS提供了两种数据分组的方式：一种是数据自动分组，另一种是数据手工分组。

一、数据自动分组

SPSS中的数据自动分组本质就是依照分组变量值的大小，将个案分成若干组。存放组号的标志变量记录着每个个案相应分组变量的升序或降序排名的名次，名次相同的个案具有相同的组号。每个个案的标志变量的值标签是相应分组变量的变量值。例如：表3—14所示的就是按升序自动分组的情况。

可见，数据自动分组中，用户只需要指定一个分组变量和一个标志变量，而无需指定分组变量的分组区间。这也正是数据自动分组使用的局限所在。

表3—14　　SPSS自动分组情况

个案号	分组变量	标志变量	标志变量值标签
1	20	2	20
2	18	1	18
3	35	3	35
4	20	2	20
5	89	4	89
6	35	3	35

数据自动分组的操作步骤：

(1) 菜单选项：Transform －> automatic Recode。出现如图3—11所示的窗口。

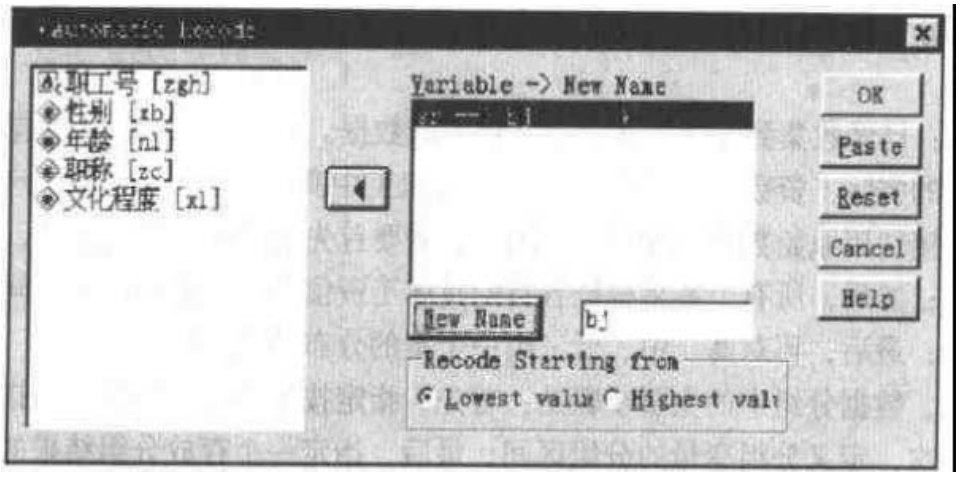


图 3—11　SPSS 数据自动分组窗口

(2) 在 Variable－>New Name 框中指定分组变量。

(3) 在 New Name 框后输入标志变量名，并按 New Name 钮。

(4) 在 Recode Starting from 框中指定组号编制方法：Lowest value 表示组号按分组变量的升序排名编制。Highest value 表示组号按分组变量的降序排名编制。

二、数据手工分组

数据自动分组往往不能满足实际工作中灵活多变的分组需要。SPSS 中的数据手工分组就比较实用。

SPSS 的数据手工分组可以对数据编辑窗口中的所有个案或部分满足条件的个案进行分组。数据手工分组同样需要指定分组变量、标志变量，而关键的一步是定义分组区间。

数据手工分组的区间定义，应给出各分组区间的下界和上界，并指明各区间所对应的分组标志（组号）是什么。

例如：表 3—15 给出了一个手工分组中区间和组号的对应情况。可以指定：35 岁以下为一组，分组标志为 1；36 岁至 54 岁为一组，分组标志为 2；55 岁以上为一组，分组标志为 3。

表 3—15　　SPSS 数据手工分组中区间和组号的对应情况

个案号	年龄为分组变量	标志变量
1	35	1
2	40	2

续前表

个案号	年龄为分组变量	标志变量
3	20	1
4	45	2
5	50	2
6	30	1
7	55	3
8	60	3

本质上讲，SPSS中的分组区间是一个广义的概念，对它的描述可以为：

(1) 一个数据点。包括：一个特定的值（Value）、系统缺失值（System-missing）、系统和用户缺失值（System-missing and user-missing）。

(2) 一个给出了最大值和最小值的闭区间。

(3) 一个没有给出最小值，仅给出最大值的开区间（Lowest through n），即：小于等于 n 的区间。

(4) 一个仅给出最小值，没有给出最大值的开区间（n through highest），即：大于等于 n 的区间。

(5) 其他数据值（All other values）

需要说明的是：

● 如果分组变量为字符型，则它的分组区间只可能是上述的（1）、（5）的情况。

● 在定义分组区间和其对应的分组标志时，区间最好不要交叉重叠。

SPSS中有两种存放分组结果的方式：一种方式是，分组结果存放在原分组变量中，即指定的分组变量就是标志变量。这种方式将分组变量原始数据代之以分组后的组号数据，从而必然造成分组变量原始数据的丢失。第二种方式是，将分组结果存放在一个新的变量中，即指定一个标志变量。可见，第二种方式比第一种方式好一些。

与此对应，SPSS的数据手工分组也有两种操作方法：

1. 分组结果存入原变量

操作步骤是：

(1) 菜单选项：Transform －> Recode －> Into Same Variables。出现如图3—12所示的窗口。

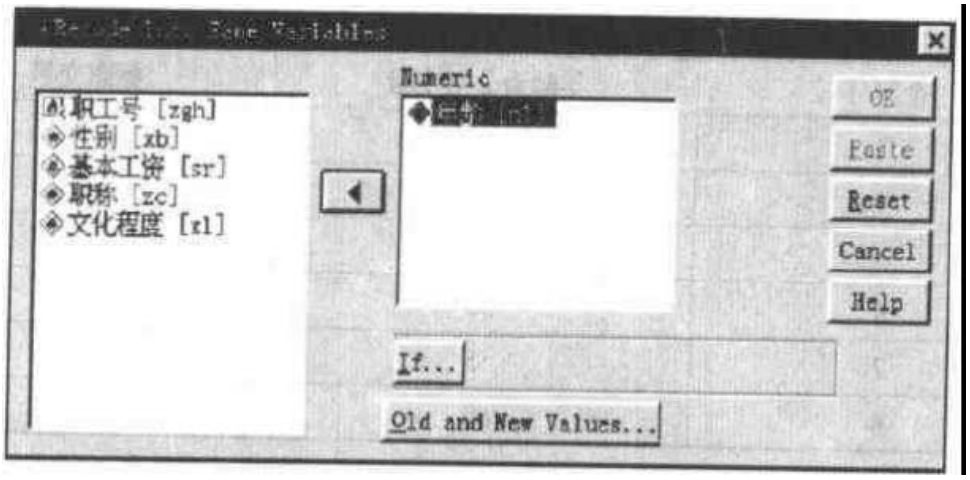


图 3—12　SPSS 数据手工分组窗口（结果存入原变量）

（2）指定分组变量到 Variables 框中。

（3）按 Old and New Values 钮进行分组区间定义。出现如图 3—13 所示的窗口。

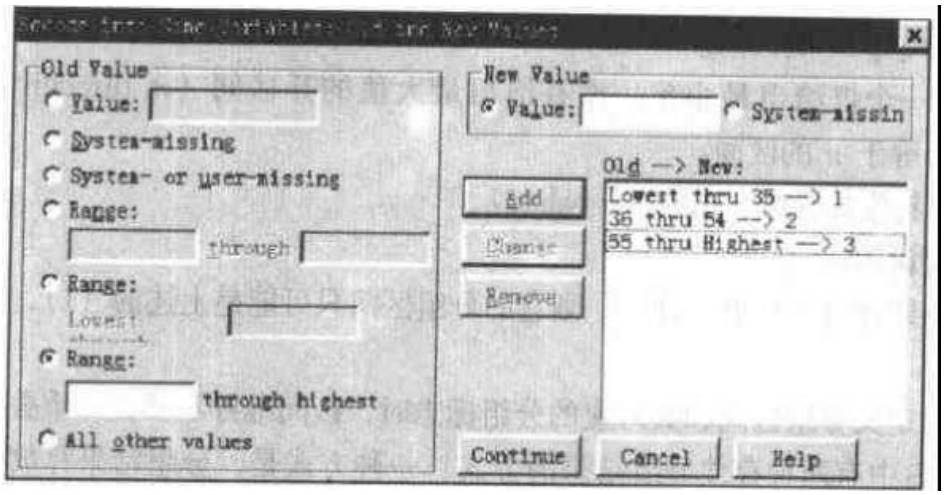


图 3—13　SPSS 数据手工分组的区间定义窗口

（4）定义一个分组区间，在 New Value 框中给出该区间所对应的分组标志是什么，或者指定该区间数据在分组后为系统缺失值。通过按 Add 钮来确认并加到 Old－＞New 框中。Change 钮实现区间定义的修改，Remove 钮实现删除。

（5）如果仅对符合一定条件的个案进行分组，则按 IF 钮并输入条件表达式。否则，本步略去。

于是，数据编辑窗口中的分组变量值被相应分组号替代。

2．分组结果存入新变量

实现步骤：

(1) 菜单选项：Transform －＞ Recode －＞ Into Different Variables。出现如图 3—14 所示的窗口。

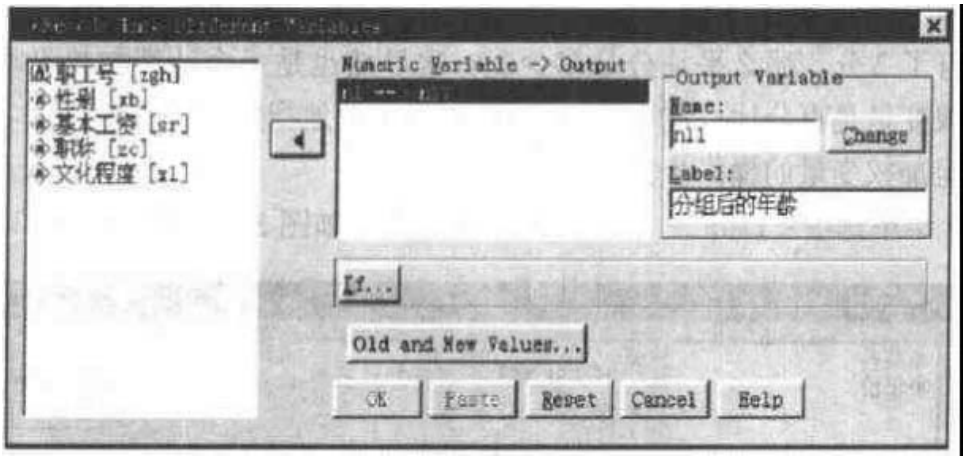


图 3—14　SPSS 数据手工分组窗口（结果存入新变量）

(2) 指定分组变量到 Input Variable －＞Output 框中。

(3) 在 Output Variable 后输入标志变量名，并按 Change 钮确认。可以在 Label 后输入标志变量名标签信息。标志变量名不能与数据编辑窗口中的变量重名。

(4) 按 Old and New Values 钮进行分组区间定义。

(5) 定义一个分组区间，在 New Value 框中给出该区间所对应的分组标志是什么，或者指定该区间数据在分组后为系统缺失值，或者，标志变量仍取分组变量值。通过按 Add 钮来确认并加到 Old －＞New 框中。Change 钮实现区间定义的修改，Remove 钮实现删除。标志变量可以为字符型，其取值可以为 'A'、'B'、'C' 等。在这时，应首先点 Output are string 钮，指定标志变量为字符型。

(6) 如果仅对符合一定条件的个案进行分组，则按 IF 钮并输入条件表达式。否则，本步略去。

于是，数据编辑窗口中将出现一个存放分组结果的新变量。

§3.4　数据文件的其他处理功能

一、指定加权变量

在实际工作中经常会计算加权平均数。例如：希望了解某菜市场某天蔬菜销售的平均价格情况。在这里，如果仅用各种蔬菜销售单价的平均数作为平均价格就不够合理，还应考虑到销售量这个因素对平均价格的影响。因此，以蔬菜的销

售量为权数计算各种蔬菜销售单价的加权平均数，能够较准确地反应平均价格的情况。再例如：在某网站上请足球球迷为国家足球队在亚洲杯上的表现打分。假如有10%的球迷打了5分，25%的球迷打了4分，40%的球迷打了3分，25%的球迷打了2分，那么平均分是多少呢？这显然也是一个计算加权平均数的问题，加权变量是百分比。在SPSS中可以非常方便地指定权数变量。

指定加权变量的操作步骤是：

(1) 菜单选项：Data -> Weight case。出现如图3—15所示的窗口。

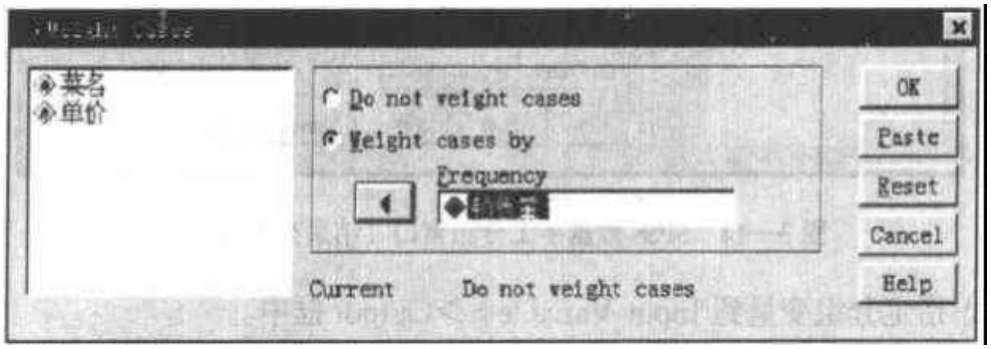


图3—15 SPSS指定加权变量窗口

(2) 点选Weight cases by选项，并指定某变量为加权变量。

于是，在屏幕上会出现相应的信息，提示用户当前的加权变量名。如果取消指定的加权变量应在屏幕上点选Do not weight cases项。

二、变量集的定义和使用

在实际的统计分析中，收集并定义到SPSS中的变量会有几十或上百个。于是，在进行各种数据处理和统计分析时，所有变量都会显示在屏幕上供用户选择。在如此多的变量中反复选择几个变量进行某种处理及分析是很麻烦的。因此，如何减少显示在屏幕上变量名的个数，方便用户对变量的选择，简化操作过程就显得非常必要。SPSS变量集可以简化处理及分析过程中变量选择的操作过程。

1. SPSS变量集的定义

SPSS变量集是一种存放着许多SPSS变量名的集合，它包括系统变量集和用户自定义变量集两大类。

系统变量集是SPSS系统事先已定义好的，它包括两个集合，分别名为：ALLVARIABLES和NEWVARIABLES。其中：ALLVARIABLES变量集中存放了数据编辑窗口中的所有变量名；NEWVARIABLES变量集中存放了数据编辑窗口中所有尚未存盘的新定义的变量名。有时，这两个系统变量集中所包含的变

量名是完全相同的。系统变量集为用户提供了一个完整的变量全体，是 SPSS 系统对变量进行管理的依据。

用户自定义变量集是用户根据实际需要自己定义的，它可以有若干个。一般将待进行相同处理或分析的若干变量名存放在一个用户变量集中。

如果用户仅希望对数据文件中众多变量中的几个变量进行相同的处理及分析，可以先将这几个变量定义为一个用户变量集。然后，再指定使用这个变量集。于是，在数据处理和分析过程中，屏幕上显示出来的变量名就仅是包含在该用户变量集中的变量名，从而大大方便了对变量的选择，可以加快操作速度。

为通过用户变量集来简化变量选择的操作过程，应分别完成两步工作：第一步，定义一个用户变量集。第二步，指定使用该用户变量集。

2. 定义用户变量集

定义用户变量集是使用用户变量集的前提。其本质就是将一些进行相同处理和分析操作的变量名定义在一个集合中。

定义一个用户变量集的操作步骤是：

(1) 菜单选项：Utilities－＞Define Variable Sets。出现如图 3—16 所示的窗口。

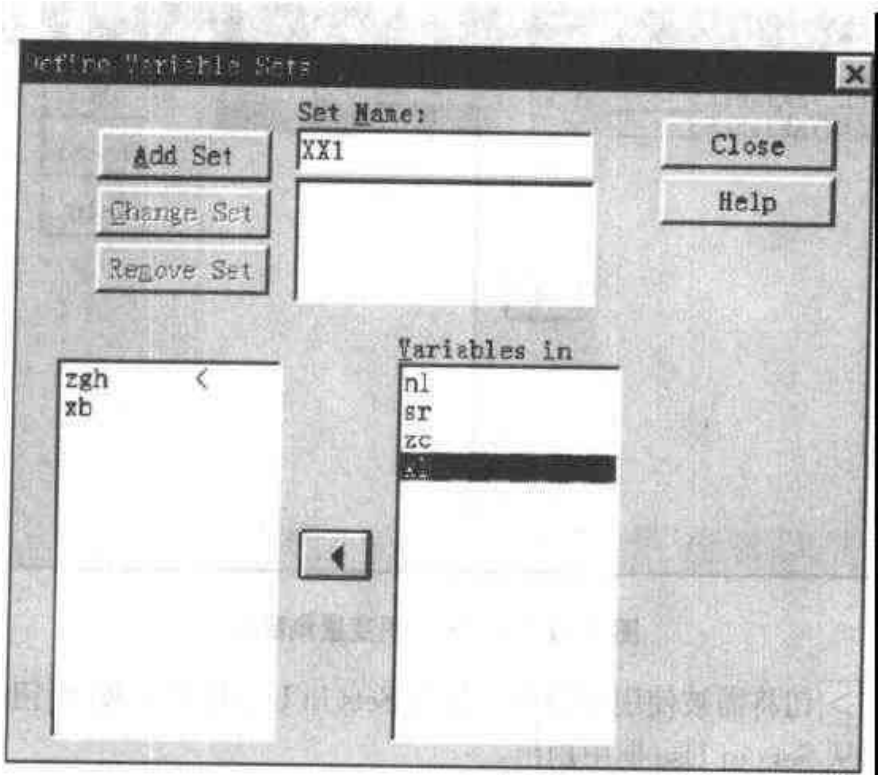


图 3—16　SPSS 定义用户变量集窗口

(2) 在 Set Name 框中输入定义的用户变量集的名称。

(3) 选择若干个变量按[>]钮将它们送到 Variables in 框中，表示这些变量将组成这个用户变量集。

(4) 按[Add Set]钮将上述定义的用户变量集加到 SPSS 变量集中。按[Change Set]钮可实现对已定义的一个用户变量集的修改。按[Remove Set]钮可将一个用户变量集从 SPSS 变量集中删去。

3. 变量集的使用

SPSS 有许多个变量集，用户应告知 SPSS 系统将要使用哪个变量集。所谓使用某个变量集就是指，在进行各种数据处理和分析过程中，屏幕上显示出来的变量仅是这个正在被使用的变量集中的变量。而那些未被使用的变量集中的变量将不被显示出来。

系统默认的正在被使用的变量集是 SPSS 的两个系统变量集，这正是为什么一般屏幕上总显示所有变量的原因。用户可以根据自己的需要，指定使用哪个用户变量集，从而方便操作。

用户变量集使用的操作步骤是：

(1) 菜单选项：Utilities －> Use Sets。出现如图 3—17 所示的窗口。

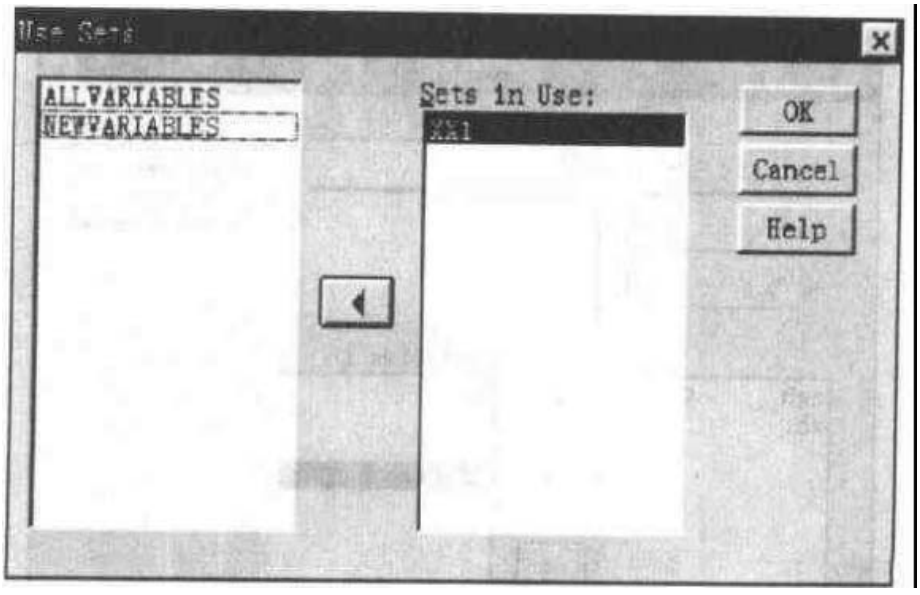


图 3—17 SPSS 使用变量集窗口

(2) 按[>]钮将需被使用的变量集选到 Sets in Use 框中。按[<]钮将不需被使用的变量集从 Sets in Use 框中剔出。

需要说明的是：在 SPSS 运行过程中，应至少有一个变量集正在被使用。否则，SPSS 将无法提供选择变量的依据。因此在 Sets in Use 框中应至少有一个变量集。

第 4 章

SPSS 输出窗口的操作和管理

SPSS 的输出窗口名为 Viewer，它是显示和管理 SPSS 统计分析结果、报表及图形的窗口。为方便用户查看、操作管理统计分析结果，输出窗口中用于显示统计分析结果的区域被分成左右两个部分。左部分以类似于 Windows 资源管理器的树形结构形式显示已有的分析结果的标题和内容索引，称为索引输出区。右部分是各个分析结果的具体详细报告，称为详解输出区。索引输出区中的红色右箭头所指内容与详解输出区中的红色右箭头所指内容一一对应。实质上，索引输出区就是详解输出区的一个视图，它以简洁的方式反映出了详解输出区中分析结果的内容项，并为用户操作和管理它们提供了极为方便的手段和界面。这两个输出区的大小是可以随意改变的，只需将鼠标移动到屏幕中间的分隔栏上，然后左右拖动鼠标即可实现。每个输出区都有各自独立的纵向与横向的屏幕滚动按钮，一个输出区中屏幕内容的滚动不会影响另一个输出区中内容的显示。

对索引输出区和详解输出区的操作管理功能，有一些是完全相同的，也有一些不同。相同的操作功能，用户无论对哪个区中的操作对象进行操作管理，其效果是一样的，并会在两个区中同时体现。而不同的操作功能，则只能针对不同区中的操作对象进行，因此，用户在操作使用时应加以注意。

§4.1 索引输出区管理

一、索引输出区的组成

索引输出区是一棵横向放倒的树。树根名为 Output，各内容项以子树的形式向右下方排列，形成一棵索引树。一个统计分析结果索引就是一棵索引子树，一般由四大子树项组成：

(1) 第一项，索引首部 (Head)：是某棵索引子树的树根，对应到详解输出区中的是某分析报告的整体。索引首部是查找某索引及其所对应详细内容的依据，一般以本次分析功能所对应的菜单项来命名，通过它用户便可以大致了解所作的是哪种分析工作。

(2) 第二项，索引标题 (Title)：位于索引首部的下层，对应到详解输出区中的是某分析报告的标题。索引标题一般也以本次分析功能所对应的菜单项来命名。

(3) 第三项，索引注解 (Notes)：与索引标题在同层上。对应到详解输出区中的是一段 SPSS 的命令和参数。这段 SPSS 命令和参数一般处于不可见的隐藏状态，也可以设置成可见状态。通过索引注解，用户可以方便地知道为完成本次分析功能，SPSS 系统执行的是哪条命令和参数，为进一步掌握 SPSS 命令，手工编写 SPSS 程序解决复杂问题打下了一个良好的基础。

(4) 第四项，多张分析报表的简要说明：与索引标题在同层上。对应到详解输出区中的是多张分析报表。一般一个统计分析的结果会有多张分析报表。

二、索引输出区的操作和管理

索引输出区操作和管理的对象是索引输出区中的内容。但是，有些操作也同时针对详解输出区中的相应内容。也就是说，这些操作在两个区中都可以执行，并且有相同的效果。因此，用户可以依据自己的操作习惯来使用。索引输出区的基本操作有以下几种。

1. 选择索引树、子树、子树项

选择操作是进行其他操作的第一步。选择就是使索引输出区和详解输出区中的相应内容处于选中状态，其表现为：两个区域中的红色右箭头分别指向相应的对象，索引树或相应项呈蓝色显示，详解输出区中的相应内容出现黑色单线

边框。

（1）选择索引树：鼠标点击索引树根 Output 左边的图标，整个索引树都处于选中状态；

（2）选择索引子树：鼠标点击某索引子树根左边的图标，该索引子树处于选中状态；

（3）选择索引子树项：鼠标点击某索引子树项左边的图标，该项内容处于选中状态。

2. 展开（Expand）/折叠（Collapse）索引树

SPSS 默认将索引树和其中的各个子树以展开的形式显示。但在子树较多的情况下，为方便查看和管理，需要将展开的索引树或某些子树以折叠的形式显示。因此，折叠展开的索引树、展开折起的索引树是索引输出区中的常见操作。索引树的展开或折叠将不影响详解输出中的内容。操作步骤是：

（1）选择需要被展开或折叠的索引树或子树。

（2）菜单选项：View－>Expand（Collapse），或者双击树根边的图标。

3. 隐藏（Hide）/显示（Show）索引树

SPSS 默认显示索引子树中除索引注解之外的其他内容。但有时也需要查看索引注解，或要将索引树中的某棵子树或子树项隐藏起来。这时，相应的操作就是隐藏处于可见状态的子树项或显示处于隐藏状态的子树项。隐藏或显示索引树实际上是隐藏或显示详解输出区中的相应内容，而在索引区仅反应在相应索引左边的图标有所变化而已。操作实现步骤是：

（1）选择需要被隐藏或显示的索引子树或子树项。

（2）菜单选项：View－>Hide（Show）。

4. 降低（Demote）/提升（Promote）索引子树

SPSS 默认将所有索引子树以相同的层次关系并排显示在索引树的树根下。有时，需要降低某棵子树的层次，即将某棵子树放到另一棵子树的下层，使其成为另一棵子树的子树，这样它们便有了共同的子树根，从而方便进行管理。或者反之，提升某棵子树。降低或提升索引子树对详解输出区中的内容没有影响。操作步骤是：

（1）选择需要被降低或提升的索引子树。

（2）菜单选项：Edit－>Outline－>Demote（Promote）。

5. 删除索引子树或子树项

在一些统计分析的结果不再有用时，可以将它们删除掉。在删除索引子树或子树项的同时，详解输出中的相应内容自动被删除。如果想恢复删除的内容，可

选主菜单 Edit 中的 Undo 选项。操作步骤是：

（1）选择需要被删除的索引子树或其中的某项。

（2）菜单选项：Edit ＞Delete。或按键盘上的Del键。

6. 重命名子树项

各子树项的名称是 SPSS 自动给出的，如：索引标题名为 Title、索引注解名为 Notes 等。用户可以根据实际需要给它们改名。重命名操作对详解输出区中的内容没有影响。操作步骤是：

（1）选择需要被重命名的子树项。

（2）鼠标点击名称部分，使其处于编辑状态。

（3）输入新名字。

7. 复制索引子树或子树项

用户可以根据实际需要将某些内容复制。在复制索引子树或子树项的同时，详解输出区中的相应内容自动被复制。操作步骤是：

（1）选择需要被复制的子树或子树项。

（2）菜单选项：Edit－＞Copy。或同时按键盘上的Ctrl和C键。

（3）选择目的地上的子树或子树项，复制对象将被复制到它的前面。

（4）菜单选项：Edit－＞Paste After。或同时按键盘上的Ctrl和V键。

8. 移动索引子树或子树项

用户可以根据实际需要调整各个分析内容的位置。在移动索引子树或子树项的同时，详解输出区中的相应内容自动被移动。操作步骤是：

（1）选择需要被移动的子树或子树项。

（2）菜单选项：Edit－＞Cut。或同时按键盘上的Ctrl和X键。

（3）选择目的地上的子树或子树项，移动对象将被移动到它的后面。

（4）菜单选项：Edit－＞Paste After。或同时按键盘上的Ctrl和V键。

9. 插入 SPSS 文本

在实际工作中，用户可能希望在详解输出区中的分析结果中插入一些自己的评述观点。由于 SPSS 不具备类似 Word 等文字处理软件中的文字排版功能，因此，无法在相应位置直接输入文字信息，而只能通过插入文本的操作来实现，即在索引子树中插入 SPSS 文本，在详解输出区中输入评述文字。操作步骤是：

（1）选择相应子树项来确定插入地点。

（2）菜单选项：Insert－＞New Text。于是，在索引子树插入点的后面自动出现 SPSS Text 的字样，在详解输出区的相应位置上出现斜线矩形框。

（3）在斜线矩形框中输入评述文字。

与此类似，用户还可以插入 SPSS 各种图形以及 Word、Excel 等外部对象（Object）。

10. 格式对齐

格式对齐是针对详解输出区中的相应内容的。为使分析结果和报表打印出来后有较好的视觉效果，SPSS 提供了对分析结果的对齐处理。以详解输出区的左右两边界，安排所选中的内容居左对齐（Align Left SPSS 默认）、居中对齐（Center）、居右对齐（Align Right）。操作步骤是：

（1）选择相应子树或子树项。

（2）菜单选项：Format，并选择所需要的对齐方式。

11. 预览和打印

预览和打印功能是预览和打印详解输出区中当前选中的内容项。预览是在计算机屏幕上将打印的效果模拟显示出来，打印则是将相应内容输出到打印机上。操作步骤是：

（1）选择相应子树或子树项。

（2）菜单选项：File－>Print Preview（Print）。

12. 改变索引输出区的显示样式

主要包括：

（1）改变索引树的显示字体。菜单选项：View－>Outline Font。

（2）改变索引树显示的大小。菜单选项：View－>Outline Size，有小（Small）、中（Medium）、大（Large）三种选择项。

§4.2 详解输出区管理

一、分析表格的组成

详解输出区与索引输出区中的内容是相互对应的，一般以分析表格和图形的形式出现。分析表格一般包括以下几个组成部分。

（1）第一部分，表层（Layer）：是表格的一维。表层是 1，表示有一张表，表层是 n，表示有 n 张表。如表 4—1 中的“性别：男职工”。

（2）第二部分，行（Row）标签和行标题：也是表格的一维，说明行数据的含义。如表 4—1 中的“文化程度”。

（3）第三部分，列（Column）标签和列标题：也是表格的一维，说明列数据的含义。如表 4—1 中的“职称”。

（4）第四部分，表标题（Title)。如表 4—1 中的“职称与文化程度分布情况表”。

（5）第五部分，表说明（Caption)：对表作简要说明。如表 4—1 上的“截止到 1999 年底”。

（6）第六部分，表下注（Footnote)：以小写英文字母或数字表示，对表中某些数据内容作必要的说明。如表 4—1 中的“a：不包括离退休人员”。

（7）第七部分，表单元（Cell)：组成一张完整的分析表。

表 4—1　　职称与文化程度分布情况表

性别：男职工

文化程度	职称			合计
	高级工程师[a]	工程师	助理工程师[a]	
本科	1	1	2	4
专科	1	5	1	7
高中	1		3	4
初中			6	6
合计	3	6	12	21

注：截止到 1999 年底

a：不包括离退休人员

二、详解输出区的操作管理

对详解输出区进行操作和管理，应首先明确其中的某张分析表格的当前状态。因为，在不同状态下所能实现的操作管理是不同的。

详解输出区中的分析表格一般有两种状态，它们是：选中状态和编辑状态。选中状态是对分析表格其进行管理的前提，编辑状态是对分析表格进行编辑等操作的必要条件。用户在使用时，应注意观察分析表格的状态，这样才能正确完成操作和管理工作。

使详解输出区中的某张分析表格处于选中状态，操作方法是：鼠标指向详解输出区中的某张分析表格并单击，或按照上面所述的方法，选择索引输出区中的相应索引子树、子树项。于是，相应的表格的四周会出现有黑色单线边框。对处于选中状态下的表格可以进行许多操作和管理，如：表格的隐藏和显示、删除、复制、移动、插入等。具体的操作与上面所讲述的方法是相同的，这里不再重复。下面重点讲述处于编辑状态表格的操作。

1. 使分析表格处于编辑状态

如果要对某张分析表格进行编辑等操作，则需要使其处于编辑状态。操作步骤是：

（1）鼠标指向详解输出区中的某张分析表格。

（2）鼠标双击。于是，相应表格的四周会出现黑色斜线边框，同时输出窗口上的主菜单也发生相应变化。

2. 确定分析表格中各组成部分的状态

一张分析表格是一个整体的操作对象。由于它是由若干部分组成的，因此，在这个整体表格处于编辑状态的情况下，还要将表格继续分解，从而来确定各个组成部分的状态。表组成部分的状态也有选中状态和编辑状态两种，不同的状态所能实现的功能也不一样。

将表格组成部分设[illegible]

(1) 使欲编辑的表格处于编辑状态。

(2) 使该表中的欲修改部分呈编辑状态。

(3) 输入新内容。

需要说明的是：

● 如果用户修改了表格单元中的某个数据，SPSS 不会自动调整修改与之相关的其他数据；

● 对表格中的内容进行编辑以后，SPSS 将自动调整表格中相应部分的行宽度。

5. 调整表格的行列宽度

SPSS 以比较合适的行列宽度显示分析输出表格。如果对表格中的内容进行了编辑，SPSS 将自动调整相应内容所在行的行宽。用户不能手工调整各行的行宽，而可以手工调整各列的列宽度。调整时，可以仅对某列的列宽作调整，也可以将各列的列宽同时调整为相同的宽度。操作步骤是：

(1) 使欲调整的表格处于编辑状态。

(2) 将鼠标移到某列的列表格线上，拖动鼠标调整该列的列宽。

或：

(1) 使欲调整的表格处于编辑状态。

(2) 菜单选项：Format－>Set Date Cell Widths。然后，SPSS 提示用户输入列宽度数据，于是，SPSS 自动将各列的列宽调整到用户指定的宽度。

主菜单 Format 下的 Auto fit 选项也可以实现对某选中表格行列宽度的自动调整。

6. 添加表格组成部分

SPSS 允许用户在一张输出分析表中插入一些内容，其中包括：表标题、表说明、脚注。一般一张表格只能有一个标题、一个表说明和若干条脚注信息。操作步骤是：

(1) 使欲被添加的表格处于编辑状态。

(2) 菜单选项：Insert。选择插入哪个组成部分。

(3) 使添加的部分呈编辑状态。

(4) 输入新内容。

7. 隐藏/显示表格中的表标题、表说明、脚注、部分行列

SPSS 允许用户根据自己的需要隐藏输出分析表格中的某些内容。于是，这些被隐藏的内容不再显示在屏幕上或打印到打印机上。需要说明的是：隐藏仅是不显示，而不是真正删除。因此，用户还可以在必要的时候重新显示它们。操作

步骤是：

（1）使被处理的表格处于编辑状态。

（2）单击鼠标的同时按住 Ctrl 键，使需要被隐藏的行或列处于选中状态。或直接单击鼠标选中需隐藏的标题、说明、脚注。

（3）菜单选项：View－>Hide。于是，SPSS 自动调整表格，被选中的内容不再显示出来。

或：

（1）使被处理的表格处于编辑状态。

（2）菜单选项：View－>Show All。于是，该表格中所有被隐藏的内容将重新显示出来。

8. 表格层、行、列三维的互换

SPSS 可以实现对表格的层、行、列三维的互换显示，体现了 SPSS 在表格处理方面的强大功能。在互换过程中，用户可以将表的行列内容互换、将表层内容移到行列上，或者直接通过 Pivoting Trays 旋转透视窗口，直观地观察并指定表三维间的互换。操作步骤是：

（1）使被透视的表格处于编辑状态。

（2）菜单选项：Pivot－>Transpose Rows and Columns。于是，SPSS 自动将表格的行列内容互换。

或：

（1）使被透视的表格处于编辑状态。

（2）菜单选项：Pivot－>Move Layers to Rows（Columns）。于是，SPSS 自动将表格的层内容放到行或列上。

或：

（1）使被透视的表格处于编辑状态。

（2）菜单选项：Pivot－>Pivoting Trays。于是，出现一个名为 Pivoting Trays 的窗口。在这个窗口中，用图标显示了三维的情况。

（3）鼠标拖动图标到相应的层、行、列位置上。于是，SPSS 自动按照用户的要求重新显示表格。

如果用户希望在屏幕上重新显示原来表格的样式，可以通过主菜单 Pivot 下的 Reset Pivots to Defaults 选项实现。

9. 表格行标签和列标签的隐藏/显示和旋转

表格中的行标签和列标签是 SPSS 自动附加上的。用户可以根据实际需要决定显示或不显示它们，或将它们在表格单元内进行旋转。操作步骤是：

（1）使被处理的表格处于编辑状态。

（2）选中欲隐藏/显示的某行或某列的标签。

（3）菜单选项：View－＞Hide（Show）Dimension Label。于是，相应的行或列标签被隐藏起来或被显示出来。

或：

（1）使被处理的表格处于编辑状态。

（2）菜单选项：Format－＞Rotate Inner Column Labels（Rotate Outer Row Labels)。于是，标签会旋转显示。

需要说明的是：如果在定义SPSS数据结构时没有定义变量值标签，就没有必要也无法进行标签旋转了。

10. 修改表格单元的外观

SPSS按照默认的格式显示表格单元中内容。用户可以根据自己的喜好对它进行修改。包括：更改显示字体、数据显示格式、数据对齐方式等属性。操作步骤是：

（1）使被处理的表格处于编辑状态。

（2）选中欲修改外观的单元。

（3）菜单选项：Format－＞Font。于是，用户可以挑选自己喜欢的字体和字型等。

或：

（1）使被处理的表格处于编辑状态。

（2）选中欲修改外观的单元。

（3）菜单选项：Format－＞Cell Properties。于是，用户可以对数据显示格式、对齐方式、边界、阴影等属性进行选择和定义。

11. 修改表格的外观和样式

如果按照上面所述的方法来逐个修改各表格单元的外观无疑是比较麻烦的，可以通过修改整个表格的外观的方法来实现。同时，用户还可以套用SPSS现成的表格样式。操作步骤是：

（1）使被处理的表格处于编辑状态。

（2）菜单选项：Format－＞Table Propertics。于是，用户可以对表格的行列宽度、脚注字符及位置、单元格式、表格边框等相应项进行定义。

或：

（1）使被处理的表格处于编辑状态。

（2）菜单选项：Format－＞Tablelooks。于是出现多种现存的SPSS表格样

式，用户可以选择其中喜欢的一种。

12．表格书签（Bookmark）

表格书签能够记录表格的不同显示样式。例如：上面所提到的表格的层、行、列之间的各种互换显示。但互换后屏幕只能显示一种样式的表格，如果希望以其他的样式输出表格，还必须再进行互换，这无疑增加了用户的操作。表格书签就能够解决这个问题。它提供了Bookmarks窗口，能够方便地将表格的各种样式以不同书签名的方式记录下来并进行各种管理。操作步骤是：

（1）使将被记录的表格处于编辑状态。

（2）菜单选项：Pivote－＞Bookmarks。于是会出现Bookmarks窗口。在这个窗口中，用户可以为当前屏幕显示的表格样式起一个名字，即书签名，然后将书签名通过Add钮加到书签中。还可以通过Delete钮删掉不再有用的书签名，或通过Rename钮修改已存在的书签名，从而实现对书签的管理。最重要的是，通过这个窗口，用户能够点某书签名，并按Go To钮，从而快捷地在屏幕上看到表格的各种样式。

总之，SPSS对详解输出区中分析表格有较为强大的处理功能，用户在使用过程中应注意摸索学习。

§4.3 输出窗口中分析结果的存取

一、保存输出窗口中的内容

SPSS输出窗口中的内容可以保存到磁盘文件中。正如第1章中提到的，在退出SPSS时，系统会自动提问用户是否保存输出窗口中的内容。输出结果以.spo为文件扩展名存储在磁盘上。

如果在SPSS运行过程中保存输出的分析结果，需要按以下操作步骤实现：

（1）使SPSS的输出窗口为当前窗口，即切换到输出窗口中。

（2）菜单选项：File－＞Save（Save as）。

（3）指定文件夹（路径）并给出存储文件的用户名。注意：此时的文件扩展名应默认为.spo。

于是，SPSS自动将输出窗口中的内容按用户指定的路径和用户名存放到磁盘上。

二、打开存在磁盘上的输出结果文件

正如第1章中讲到的，在SPSS分析过程中，系统会自动创立新的SPSS输出窗口，并将当前的分析结果输出到该窗口中。如果用户希望将已经存在磁盘上的输出结果文件装入SPSS输出窗口，则只能手工打开相应的文件。操作步骤是：

(1) 使SPSS的输出窗口为当前窗口，即切换到输出窗口中。

(2) 菜单选项：File－>Open。

(3) 输入输出结果文件所在的文件夹（路径）和文件用户名。注意：此时的文件扩展名应默认为.spo。

于是，SPSS将自动装入用户指定的输出结果文件到输出窗口中。

第 5 章

SPSS 基本统计分析

基本统计分析是进行其他更深入的统计分析的前提。通过基本统计分析，用户可以对分析数据的总体特征有比较准确的把握，从而选择更为深入的分析方法对分析对象进行研究。

§5.1 案例说明

为清楚地讨论 SPSS 基本统计分析，我们列举两个实际案例数据，并对它们作基本统计分析。

一、案例 1

查阅《中国普通高等院校人文社会科学统计资料汇编》，得到 1999 年全国普通高等学校 22 个学科的人文社会科学活动人员情况数据。如表 5—1 所示。

表 5—1　　1999 年全国普通高等学校人文社会科学活动人员情况表　　（单位：人）

学科	合计	科学家与工程师				技术员	辅助人员
		小计	高级	中级	初级		
管理学	32 099	27 686	7 135	12 479	8 072	2 180	2 233
马克思主义	12 061	11 831	4 214	5 215	2 402	95	135
哲学	4 869	4 789	2 378	1 756	655	18	62
逻辑学	275	273	131	111	31	0	2
宗教学	101	101	76	21	4	0	0
语言学	38 400	37 428	10 636	14 962	11 830	354	618
中国文学	10 846	10 616	4 933	3 773	1 910	92	138
外国文学	4 417	4 264	1 525	1 644	1 095	49	104
艺术学	17 107	16 256	4 603	6 118	5 535	441	410
历史学	7 351	7 197	3 341	2 777	1 079	44	110
考古学	185	175	103	57	15	2	8
经济学	25 588	24 758	8 893	10 092	5 773	424	406
政治学	4 008	3 863	1 374	1 661	828	74	71
法学	8 322	8 107	2 666	3 500	1 941	81	134
社会学	1 813	1 771	693	737	341	11	31
民族学	544	511	253	172	86	3	30
新闻与传播学	2 880	2 683	929	1 190	564	79	118
图书情报文献	31 106	26 173	4 304	13 884	7 985	1 901	3 032
教育学	9 524	8 943	2 783	3 934	2 226	271	310
统计学	1 922	1 828	636	835	357	57	37
心理学	1 638	1 612	597	720	295	12	14
体育学	21 183	20 152	6 247	8 609	5 296	307	724

现在要对上述案例数据作基本统计分析，从而掌握以下几个方面的数据：

（1）全国高等学校人文社科研究中，不同学科的平均人员投人情况。

（2）全国高等学校人文社科研究中，不同各学科的人员投人差异情况。

（3）全国高等学校人文社科研究中，人员投人在不同学科的分布情况。

二、案例 2

为了解某大学某系（非计算机专业）本科学生对该系计算机课程教学的意见，进行随机抽样调查，得到了 390 名学生的调查数据。调查表如表 5—2 所示。由于篇幅所限，具体调查数据略去。

表 5—2　　关于计算机课程教学情况的调查

1. 你的性别是
   (1) 男　(2) 女
2. 你的专业方向是
   (1) 方向 1　(2) 方向 2
3. 你毕业以后的打算是
   (1) 工作　(2) 报考研究生　(3) 出国
4. 你的计算机基础课程的成绩是
   (1) 优　(2) 良　(3) 中　(4) 及格
5. 你的数据库管理系统课程的成绩是
   (1) 优　(2) 良　(3) 中　(4) 及格
6. 你认为本系现在开设的计算机课程能否满足你的专业需要
   (1) 不能　(2) 基本能够　(3) 能够　(4) 太多
7. 你认为现在的上机时间能否满足计算机课程的需要
   (1) 不能　(2) 基本能够　(3) 能够　(4) 太多
8. 在学习计算机课程中，你喜欢下面哪种授课方式
   (1) 一门课程讲几个相关软件，每个软件不必讲得很全面、仔细
   (2) 一门课程只讲一个软件，应讲得很全面、仔细
9. 你认为有必要再增设其他计算机课程吗?
   (1) 没有必要　(2) 两可　(3) 有必要
10. 如果有下列方面的计算机选修课程，你愿意选哪个方面的（请按先后次序选 3 个）
   (1) 计算机理论　(2) 计算机操作系统　(3) 计算机软件开发语言
   (4) 计算机通讯技术　(5) 统计数据库系统　(6) Internet 原理及网页制作
   (7) 数据挖掘技术　(8) 数据仓库
   我的选择是：_____

现在对上面的调查问卷作基本分析，希望掌握以下几个方面的情况：

(1) 学生对计算机课程的设置意见。

(2) 学生对于计算机教学方法的意见。

(3) 学生对于计算机选修课程的意见。

上面两个案例分析所涉及的内容都能够通过基本统计分析实现，在下面的章节中我们将作详细讨论。

§5.2　计算基本描述统计量

得到一批分析数据后，我们要了解这批数据的基本统计特征。

例如：在案例 1 中，希望掌握 1999 年全国人文社科活动人员在各学科上的平均投入情况以及差异情况，就是要利用基本统计分析方法，掌握人力投入的平均值、最大值、最小值以及标准差等基本描述统计数字。可以通过 SPSS 计算基本描述统计量的功能来实现。

我们首先讨论基本统计量的含义，然后再讨论实际计算操作的方法。

一、基本描述统计量

常见的基本描述统计量有以下几个。

1. 均值（Mean）

寻找一个变量所有取值的“中心”往往是我们关注的问题，均值就是这样一种“中心”，它表示了某变量所有取值的集中趋势或平均水平。例如：学生某门课的平均成绩、职工的平均月收入、人文社科人员的平均投入等。

均值的数学定义为：

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \tag{5.1}$$

均值虽然代表了数据的集中趋势，但容易受到极端值的影响。

2. 标准差（Standard Deviation；Std Dev）

找到了数据的“中心”后，单纯以均值来刻画数据并非尽善尽美，还应该考察数据分布的疏密程度，即考察所有数据相对于“中心”分布的疏密程度。如果数据都紧密地集中在“中心”的周围，则说明均值这个“中心”是刻画全部数据的鲜明“代表”。否则，如果数据仅是比较松散地分布在“中心”的周围，那么，仅以均值来说明数据特征就不具有代表性。因此，“中心”和关于“中心”的稀疏程度共同作用，才能使人们对数据有一个比较全面完整地了解。

样本标准差就是表示某变量所有取值的离散程度的统计量。其数学定义为：

$$SD = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2} \tag{5.2}$$

可见，标准差刻画了数据关于均值的平均离散程度。标准差值越大，说明变量值之间的差异越大，距均值这个“中心”的离散趋势越大。

3. 方差（Variance）

方差也是表示变量取值离散程度的统计量。它的数学定义是标准差的平方。

4. 最小值（Minimum）、最大值（Maximum）、总和（Sum）

可以通过掌握变量值的最小值和最大值来了解变量的取值范围，这对总体把握数据是十分有帮助的。总和也是一种把握变量总体取值的手段。

5. 全距（Range）

全距也称为极差，是数据的最大值与最小值之间的绝对差。全距是刻画某变量所有取值离散程度的另一个统计量。在相同样本容量的情况下的两组数据，全距大的一组数据要比全距小的一组数据更分散。当全距非常小时，则意味着数据基本都集中在一起。

6. 峰度（Kurtosis）

当了解了变量取值的“中心”和疏密程度后，还需要了解变量取值的分布形态，从而把握数据整体特征。人们一般对数据的正态分布形态比较熟悉，因此，在刻画一变量取值的分布形态时，通常与正态分布相比较。

分布形态可以从数据分布的陡缓程度方面来描述。峰度就是描述某变量所有取值分布形态陡缓程度的统计量。其数学定义为：

$$\text{Kurtosis}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^4/SD^4-3 \tag{5.3}$$

在 SPSS 中，与正态分布相比，峰度为 0 表示其数据分布与正态分布的陡缓程度相同；峰度大于 0 表示比正态分布高峰要更陡峭，为尖顶峰；峰度小于 0 表示比正态分布的高峰要平坦，为平顶峰。

7. 偏度（Skewness）

偏度也是用来刻画数据分布形态的，它是描述某变量所有取值分布形态的对称性的统计量。其数学定义为：

$$\text{Skewness}=\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\bar{x})^3/SD^3 \tag{5.4}$$

在 SPSS 中，与正态分布相比，偏度为 0 表示其数据分布形态与正态分布偏度相同，左右对称；偏度大于 0 表示正偏差数值较大，为正偏或称右偏，有一条长尾拖在右边；偏度小于 0 表示负偏差数值较大，为负偏或称左偏，有一条长尾拖在左边。偏度的绝对值越大表示数据的分布形态的偏斜程度就越大。

8. 均值标准误差（Standard Error of Mean：S. E. of Mean）

我们知道，样本数据是来自总体的。样本的描述统计量可以反映总体数据的特征，但由于抽样等原因，使得样本数据不一定能够完全准确地反映总体，它可能与总体的真实值之间存在一定的差异。

样本均值作为抽样样本的平均数会与总体均值之间存在差异。进行若干次抽样，会得到若干个不同的样本均值，它们与总体均值之间存在或多或少的差异。均值标准误差就是描述这些样本均值与总体均值之间平均差异程度的统计量。

在 SPSS 中除计算输出均值标准误差外，还可以计算峰度标准误差（S. E. of Kurtosis）、偏度标准误差（S. E. of Skewness）等统计量。

二、计算基本描述统计量的操作步骤

(1) 菜单选项：Statistics－＞Summarize－＞Descriptives. . .。出现如图 5—1 所示的窗口。

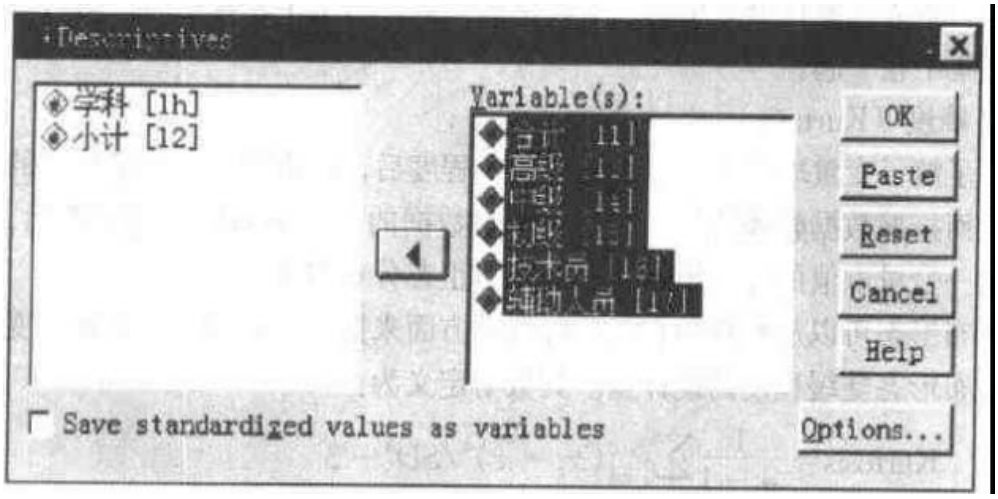


图 5—1　SPSS 计算基本描述统计量窗口

(2) 选择参加计算的数值型变量的变量名到 Variables 框。

于是，在 SPSS 输出窗口会自动显示计算结果。

三、计算描述统计量的其他菜单选项

在进行上述计算过程中，SPSS 还有一些其他选项供用户选择。

1. Save standardized values as variables

表示将指定的变量值进行标准化处理，并将标准化后的结果存入一个新变量中。新变量名以字母“Z“开头，后跟原变量名前 7 个字母。这里，SPSS 进行标准化采用的公式是：

$$z_i = (x_i - \bar{x}) / SD \tag{5.5}$$

标准化后的新变量的均值为 0，标准差为 1。变量值小于 0 的表示该变量值在平均水平以下，大于 0 表示该变量值在平均水平之上。对变量进行标准化的主要目的是为了消除各变量间变量值在数量级上的差异，从而增强数据间的可比性。

2. Option 选项

Option 选项如图 5—2 所示。在这里，用户可以选择计算哪些描述统计量，其中包括：表示集中趋势的统计量、表示离散程度的统计量、表示分布形态的统计量。

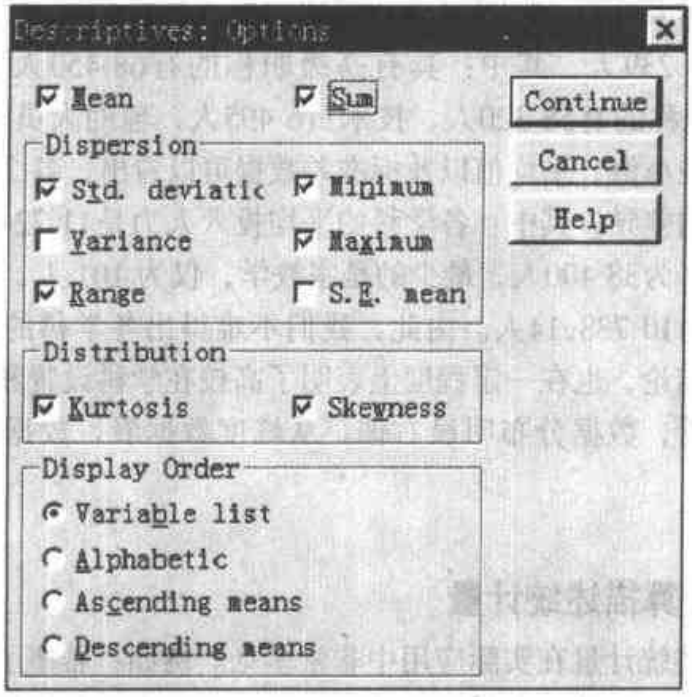


图 5—2　SPSS 计算基本描述统计量的 Option 窗口

同时，用户还可以定义各变量统计量的输出次序（Display Order）。其中：

- Variable list 表示按变量所在数据编辑窗口中从左到右的次序输出；
- Alphabetic：表示按字母顺序输出；
- Ascending means：表示按均值升序输出；
- Descending means：表示按均值降序输出。

四、案例 1 分析说明

对案例 1 数据计算基本描述统计量的分析结果，见输出结果 5—1。

输出结果 5—1　　基本统计分析案例 1 分析结果

	N	Minimum	Maximum	Sum	Mean	Std. Deviation	Skewness	Kurtosis
合计	22	101	38400	236239	10738.14	11724.20	1.170	.223
高级	22	76	10636	68450	3111.36	3000.47	1.081	.558
中级	22	21	14962	94247	4283.95	4734.03	1.197	.235
初级	22	4	11830	58320	2650.91	3299.01	1.483	1.507
技术员	22	0	2180	6495	295.23	583.91	2.784	7.067
辅助人员	22	0	3032	8727	396.68	760.86	2.849	7.837

从上述分析结果可知：1999年全国高等学校人文社会科学人力在22个学科中投入总数为236 239人，其中：具有高级职称的有68 450人，中级职称的有94 247人，初级职称的有58 320人，技术员6 495人，辅助人员8 727人。从人员投入的最大值、最小值、平均值以及标准差数据可以看出，各个学科在人力投入方面存在比较大的差异。其中：各学科的平均投入人力是11 724.2人，投入最多的学科是语言学，为38 400人，最少的是宗教学，仅为101人。同时，投入总人数平均离散程度为10 738.14人。由此，我们不难得出各学科的人力投入总体上存在较大差异的结论，也在一定程度上表明了高校在学科设置和科研方面的倾向性。从偏度数据看，数据分布明显右偏。从峰度数据看，数据分布呈尖顶峰形态。

五、分组计算描述统计量

分组计算描述统计量在实际应用中非常多见。例如：希望了解不同性别职工的工资收入情况等。显然，上述方法计算出的工资收入的基本描述统计量只能反映全体职工的工资收入情况，无法反映不同性别职工的工资收入情况。实际上这是一个分组计算描述统计量的问题。如果按照一般的解决方案实现，应先选择性别为男的职工个案数据来计算描述统计量，然后再选择性别为女的个案数据计算。这种解决方案无疑比较麻烦，因为它需要不断地对样本数据进行筛选。

比较简单的解决方案是：首先利用SPSS的拆分技术对数据文件按性别进行拆分，即分组。这里的拆分变量或称分组变量就是性别；然后再按前面讲到的方法，计算工资收入的基本描述统计量。这样，SPSS将自动分别计算不同性别组的工资收入的基本描述统计量。

数据文件拆分的结果表面上看就是对数据文件按某指定的拆分变量值重新排序。但这里并不仅是作简单排序，重要的是自动进行分组，这将为以后进行的分组计算统计量提供极大的方便。

数据文件拆分的操作步骤是：

(1) 菜单选项：Data－>Split File。于是出现如图5—3所示的窗口。

(2) 选择拆分变量到Groups Based on框中。

(3) 拆分会造成分析结果有两种不同的输出方式：Compare groups表示将各分组计算的结果输出在一张表格中。这种输出方式非常便于不同组之间描述统计量的比较。Organize output by groups表示将各分组计算的结果分别输出在不同的表格中。

(4) 如果数据编辑窗口中的数据文件已经事先按所指定的拆分变量进行了排

序，则可以选择 File is already sorted 项，否则点选 Sort the file by grouping variables 项。前一选项可以提高拆分命令执行的速度，但应注意它的前提要求必须事先满足。

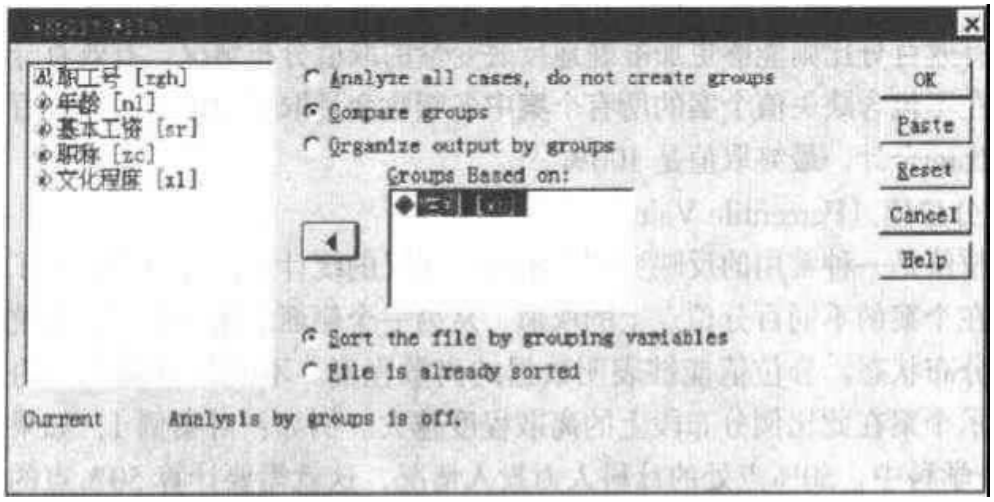


图 5—3　SPSS 数据拆分窗口

需要说明的是：

● 拆分处理对以后的 SPSS 分析操作均起作用，即以后无论进行什么分析都是按拆分变量的不同组分别进行。因此，如果希望恢复到对所有数据的整体分析操作，还需要重新执行拆分功能，并选中 Analyze all cases 项。

● 对数据文件可以进行多重拆分，非常类似于对数据文件的多重排序。多重拆分的次序决定于选择拆分变量的前后次序。

§5.3　变量的频数统计

了解变量的取值分布状况对整体把握数据的特征是非常有利的。

例如：在案例 2 中，希望掌握学生对计算机课程设置的看法和对计算机教学方法的意见。显然，如果计算出以下的频数和百分比数据：多少学生认为目前开设的计算机课程能够满足专业的需要、多少学生认为不能满足、多少学生赞成一门课程讲多个软件、多少学生不赞成．将对分析提供极为重要的数据依据。

变量的频数统计正是实现这种分析的最好手段，它可以使人们非常清楚地了解变量取值的分布状况。

一、变量频数的描述方法

1. 频数、百分比

计算变量的频数（Count）和百分比就是计算变量在各个变量值上取值的个案数、以及占总个案数的比例。如果待分析的个案在某频数统计变量上有缺失值存在，有效百分比则能够更加准确地反映变量的取值分布情况。有效百分比计算的，是在不包含缺失值个案的所有个案中各频数变量取值的比例。累计百分比是各百分比的合计，最终取值是100%。

2. 分位值（Percentile Values）

分位值是一种常用的反映变量取值分布状况的统计量，分位值给出了频数统计变量在个案的不同百分位点上的取值，从另一个侧面清楚刻画了频数统计变量的取值分布状态。分位值能够表明数据的离散程度。不同分位值之间的差距越大，表示个案在此比例分布段上的离散程度越大。例如：对案例1，如果希望了解22个学科中，50%点处的社科人力投入情况，这就需要计算50%点的人力总数分位值。它首先按人力总数的升序对各学科数据重新排序，然后找到50%处的学科，其相应的人力数就是50%点的人力分位值。这里，由于个案数为22，是偶数，那么中间的位置应是（22 + 1）/ 2 = 11.5处。于是，取第11个数据与第12个数据的平均值作为50%分位点的分位值。这个分位值意味着，有50%的学科的人力投入小于该分位值，有另50%的学科的人力投入大于该分位值。

分位点在0至100之间。一般使用比较多的是4分位点（Quartiles），它将100%平均分成4份，于是各分位点是25%、50%、75%。此外，用户还可以自行指定分位点。SPSS中有两种自行定义分位点的方式：一种是：用户直接指定分位点数（Percentile），如：指定30为一个分位点。另一种是：用户指定将100%等分为 n 份（Cut points for n equal groups）。例如：如果给出5，则表示要将100%等分为5份，那么就意味要计算20%、40%、60%、80%分位点的分位值。

3. 统计图形

统计图是一种最为直接的数据刻画方式，它无疑能够非常清晰直观地展示变量的取值状况。常用的统计图形有：条形图（Bar chart）、直方图（Histogram）等。一般来说，条形图适合表示离散型变量，直方图适合表示连续型变量，而不能表示字符型变量。如果给直方图加上正态分布曲线，则非常便于与正态分布形态进行比较。

二、变量频数统计的基本操作步骤

变量频数统计的基本操作步骤是：

（1）菜单选项：Statistics－＞Summarize－＞Frequencies。出现如图 5—4 所示的窗口。

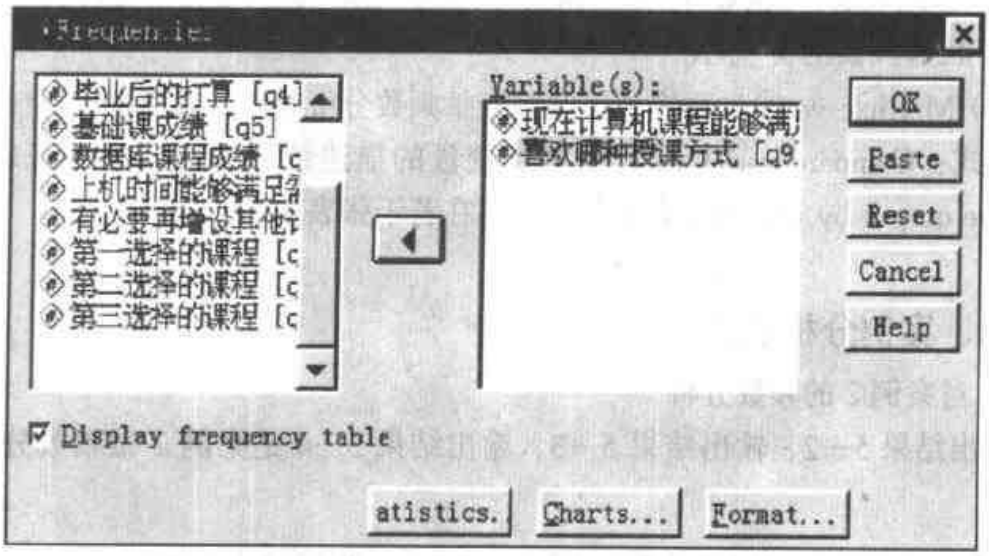


图 5—4　SPSS 频数统计分析窗口

（2）选择需要进行频数分析的频数统计量到 Variable（s）框中。

三、变量频数统计的其他选项

在进行上述计算过程中，SPSS 还有一些其他选项供用户选择。

1. Statistic 选项

该选项用来定义在显示基本频数统计量的同时还需要显示哪些相关的其他统计量。包括：各种分位值、描述集中趋势的基本统计量、描述离散程度的基本统计量、描述分布形态的基本统计量四大部分。这些统计量在以前的章节中已经讲到，这里不再重复。

2. Chart 选项

该选项用来定义在输出频数统计量的同时是否输出统计图形或输出哪种统计图形。同时，可以定义在条形图中纵坐标的含义，它可以是频数（Frequencies）也可以是百分比（Percentage）。

3. Format 选项

Format 选项是用来定义频数表的输出顺序和格式的。

（1）Order by 框定义了频数表输出的顺序。如果同时对多个变量进行频数统计，会产生许多频数表，那么 SPSS 将以什么顺序来依次输出这些频数表呢？Ascending values 表示按变量值升序输出；Descending values 表示按变量值降序输

出；Ascending counts 表示按频数升序输出；Descending counts 表示按频数降序输出。

(2) Suppress tables with more than n categories：表示仅输出频数大于 n 的频数表。SPSS 默认的 n 为 10。

(3) Multiple variable 框定义了在多变量频数分析时所计算出的描述统计量的输出方式：Compare variables 表示多个变量的描述统计量输出在一张表格中；Organize output by variables 表示分别输出在若干张表中。

四、案例分析说明

1. 对案例 2 的频数分析

输出结果 5—2、输出结果 5—3、输出结果 5—4 是案例 2 做频数分析的结果。

输出结果 5—2　　现在开设的计算机课程能否满足专业需要频数分析结果

	Frequency	Percent	Valid Percent	Cumulative Percent
不能	140	35.9	35.9	35.9
基本能够	220	56.4	56.4	92.3
能够	30	7.7	7.7	100.0
Total	390	100.0	100.0	

输出结果 5—2 表明，390 名学生中有 140 人认为目前开设的计算机课程不能满足专业的需要，占总人数的 35.9%，有 220 人认为基本能够满足，占总人数的 56.4%，累计百分比是 92.3%，有 30 人认为能够满足，占总人数的 7.7%。

输出结果 5—3　　计算机教学方法频数分析结果

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	一门课讲几个软件	200	51.3	52.6	52.6
	一门课讲一个软件	180	46.2	47.4	100.0
	Total	380	97.4	100.0	
Missing	System	10	2.6		
Total		390	100.0		

输出结果 5—3 表明，390 名学生中，有 10 人没有对此问题作出回答，SPSS 默认为系统缺失值。有 200 人赞成第一种教学方法，占总人数的 51.3%，有效百分比为 52.6%。有 180 人赞成第二种教学方法，占总人数的 46.2%，有效百分比为 47.4%。可见，学生对两种教学方法的赞成度基本持平，赞成第一种方法比例比第二种略高出 5.2 个百分点。从图 5—5 中也可以直观看出。

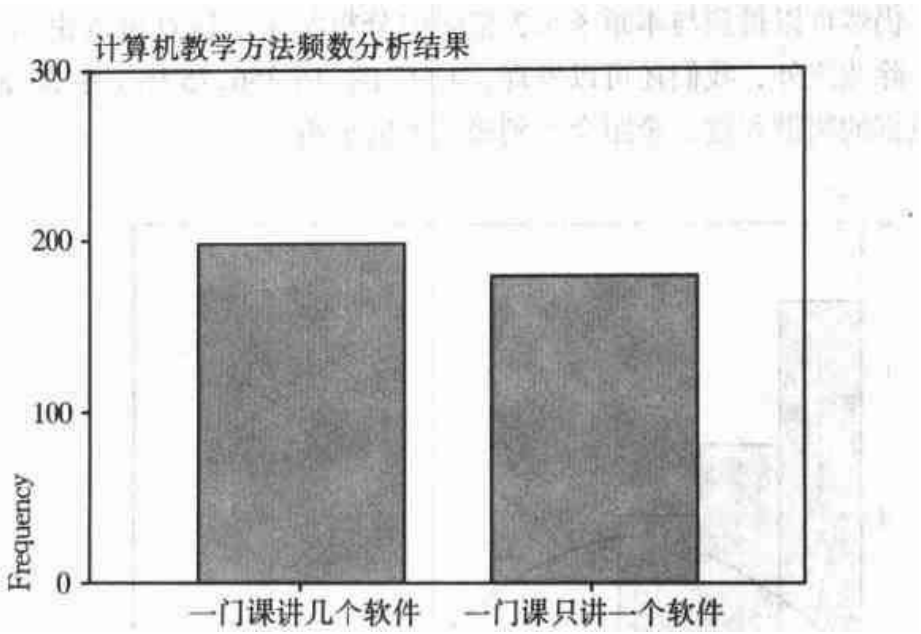


图 5—5　案例 2 中对计算机教学方法的频数分析条图

另外，通过这个例子，我们可以比较清楚地看到，在有缺失值的情况下，有效百分比较百分比能够更加准确地反映实际情况。

2. 对案例 1 的频数分析

输出结果 5—4 是案例 1 中人力投入各分位值的结果。

输出结果 5—4　　案例 1 中频数分析人力总数各分位值

N	Valid	22
	Missing	0
Mean		10738.14
Range		38299
Percentiles	25	1769.25
	50	6110.00
	60	9283.60
	70	12565.60
	75	18126.00

输出结果 5—4 表明：有 1/4 的学科，人力投人总数小于1 769.25人。有一半的学科，人力投人总数小于6 110人；有 3/4 的学科人力投人总数小于18 126人。参照均值10 738.14,发现大概有 65% 的学科人力投人小于平均值。另外，4 分内距为：18 126 − 1 769.25 = 16 356.75，反映了中间 50% 数据大小的差异，是非常大的。由此，我们能够对不同学科人力投人总数的分布情况有比较清晰的认识，仍然可以得到与本章 § 5.2 相同的分析结论。这点可以由图 5—6 中得到验证。除此之外，我们还可以发现，4 分内距 16 356.75 比全距 38 299 更准确地反映数据的离散程度，全距会受到极端值的影响。

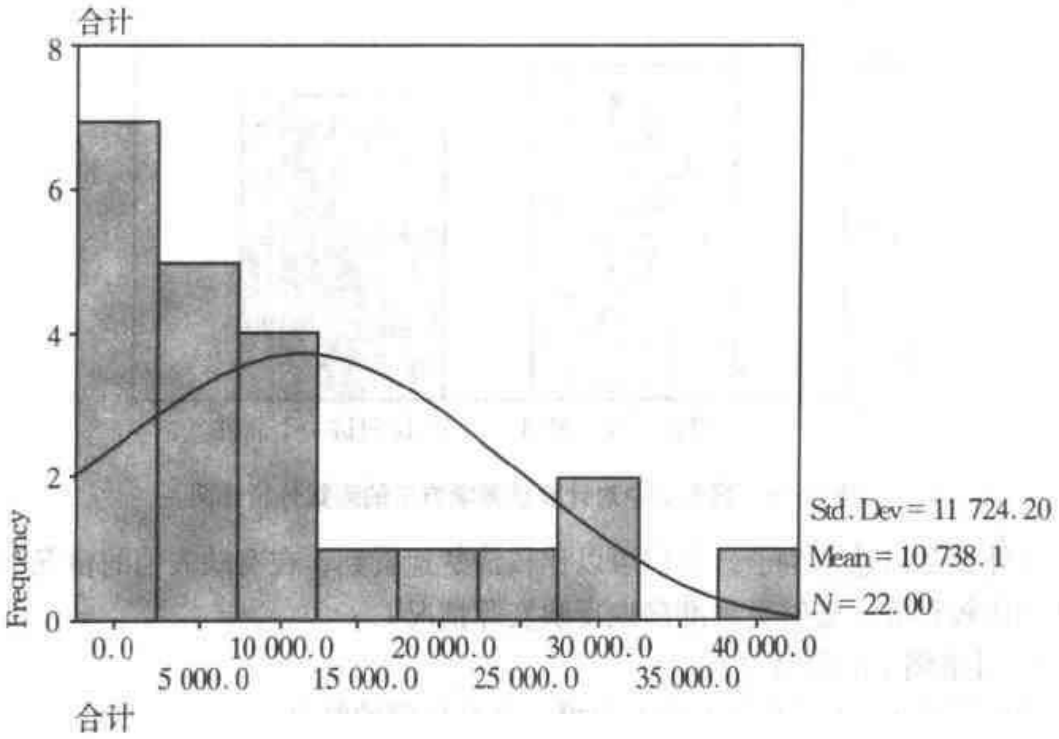


图 5—6　案例 1 中人力总数频数分布图

§ 5.4　交叉分组下的频数分析

一、交叉分组下的频数分析

1. 交叉分组下的频数分析的主要任务

通过频数分析，我们只能够了解单个变量的数据分布情况。但在实际分析中，还需要掌握多个变量在不同取值情况下的数据分布情况，从而进一步深人分析事物（变量）之间的相互影响和关系，这就需要作交叉分组下的频数分析，又

称为交叉列联表分析。

例如：对案例 2 的分析，通过频数分析，我们了解了学生对当前开设的计算机课程满足程度以及计算机教学方法的意见。为了进一步分析，我们希望了解一下，不同专业方向是否会对学生回答这两道题有较大的影响，这里有两个变量参与分析；再进一步，我们希望了解，不同性别、不同专业方向对学生回答这两道题的答案是否产生较大影响，参与分析的变量就有 3 个。

交叉分组下的频数分析的主要任务有两个：第一，根据收集到的样本数据，产生如表 5—3、表 5—4 所示的二维和多维交叉列联表；第二，在交叉列联表的基础之上，分析两两变量之间是否具有独立性或有一定的相关性。

2. 交叉列联表

表 5—3　　两变量交叉列联表示意表

	不能满足	基本满足	满足	合计
专业方向一				
专业方向二				
合计				

在表 5—3 中，专业方向和是否满足专业需要这两变量分别称为交叉列联表的行变量（Row）和列变量（Column）。

表 5—4　　三变量交叉列联表示意表

男学生

	不能满足	基本满足	满足	合计
专业方向一				
专业方向二				
合计				

女学生

	不能满足	基本满足	满足	合计
专业方向一				
专业方向二				
合计				

在表 5—4 中，性别称为交叉列联表的层控制变量（Layer），其他两个变量仍为行变量和列变量。

确定控制变量、行变量和列变量是进行交叉列联表分析的前提。

交叉列联表的单元格中一般有以下几个数据项：

(1) 观察频数（Observed Counts）

(2) 期望频数（Expected Counts）

(3) 行百分比（Row Percentage）

(4) 列百分比（Column Percentage）

(5) 总百分比（Total Percentage）

(6) 剩余（Unstandardized Residuals）

(7) 标准化剩余（Standardized Residuals）

这些数据的具体含义，我们将在案例分析中具体讨论。

3. 行、列变量独立性分析

产生交叉列联表仅是交叉分组下的频数分析的一个任务，另一个主要任务是要对行、列变量作进一步的分析。通过对交叉列联表的频数分布情况的分析，我们希望确定行、列两变量之间有怎样的关系。

通过观察频数，我们可以大致了解个案在相应行、列变量以及控制变量上的取值分布情况。但是行列变量之间是否有联系，联系的紧密程度有多少？在某些极端情况下，我们可以直接通过观察交叉频数表来确定变量之间的相互联系。例如：假如有关于年龄段（老、中、青年人）与工资收人（高、中、低收人）的交叉列联表，如表 5—5 所示。

表 5—5 (a)　　年龄段与工资收入交叉列联表（一）

	低	中	高
青	4	0	0
中	0	5	0
老	0	0	6

表 5—5 (b)　　年龄与工资收入交叉列联表（二）

	低	中	高
青	0	0	5
中	0	6	0
老	4	0	0

在表 5—5（a）中可以看到，所有的频数都出现在主对角线上，说明年龄越小工资收人就越低，年龄越大工资收人就越高，年龄和收人呈完全正相关关系。

在表5—5（b）中可以看到，所有的频数都出现在负对角线上，说明年龄越小工资收入就越高，年龄越大工资收入就越低，年龄和收入呈完全负相关关系。这是一个极端特例的情况，绝大多数情况下，频数的分布会分散在表格的各个方格中。这时，我们就不可能直接通过观察频数分布来确定行、列变量之间的关系以及关系的强弱程度。这时，就需要借助一些度量变量间相关程度的统计量和一些非参数检验的方法。

常用的衡量变量间相关程度的统计量是简单相关系数，但在交叉列联表分析中，由于行列变量往往不是连续等距变量，不符合计算简单相关系数的前提要求。因此，需要根据行列变量的实际取值类型（名义水准、顺序水准）计算其他相关系数，如：Kendall 等级相关系数、Eta 值等。SPSS 提供了若干种适用于不同数据取值范围和条件的相关系数，它们的取值范围有 0 至 1 之间的，也有 -1 至 +1 之间的。对相关系数的检验是必不可少的，SPSS 提供了不同相关系数的不同统计检验方法，这些检验的零假设都是：行、列变量之间彼此独立，不存在显著的相关关系。SPSS 将自动给出各统计检验的统计量和它们对应的概率 P 值。如果 P 值小于等于用户心中的显著性水平 α，则应拒绝零假设，认为行、列变量不独立；否则，不能拒绝零假设。

利用卡方统计量检验行列变量之间是否相关也是经常使用的方法。交叉列联表的卡方检验的零假设是：行、列变量相互独立。计算公式为：

$$\chi^2 = \sum \frac{(f_o - f_e)^2}{f_e} \tag{5.6}$$

其中：f_o 表示实际观察频数，f_e 表示期望频数。由于卡方统计量服从（行数 -1）×（列数 -1）个自由度的卡方分布，SPSS 在自动计算卡方统计量后，还给出它对应的概率 P 值。同样，如果 P 值小于等于用户心中的显著性水平 α，则应拒绝零假设，认为行、列变量不独立；否则，不能拒绝零假设。

需要说明的是：利用卡方检验时应注意满足卡方检验对期望频数的要求，如果交叉列联表中有 20% 以上单元中的期望频数小于 5，则一般不应适用卡方检验方法。这时，可以借助似然率卡方检验、Fisher 精确检验、蒙特卡罗（Monte carlo）随机模拟等方法。关于这些方法具体内容，请参阅有关书目。

二、交叉分组下的频数分析的操作步骤

交叉分组下的频数分析的基本操作步骤是：

（1）菜单选项：Statistics - > Summarize - > Crosstabs，出现如图 5—7 所示的窗口。

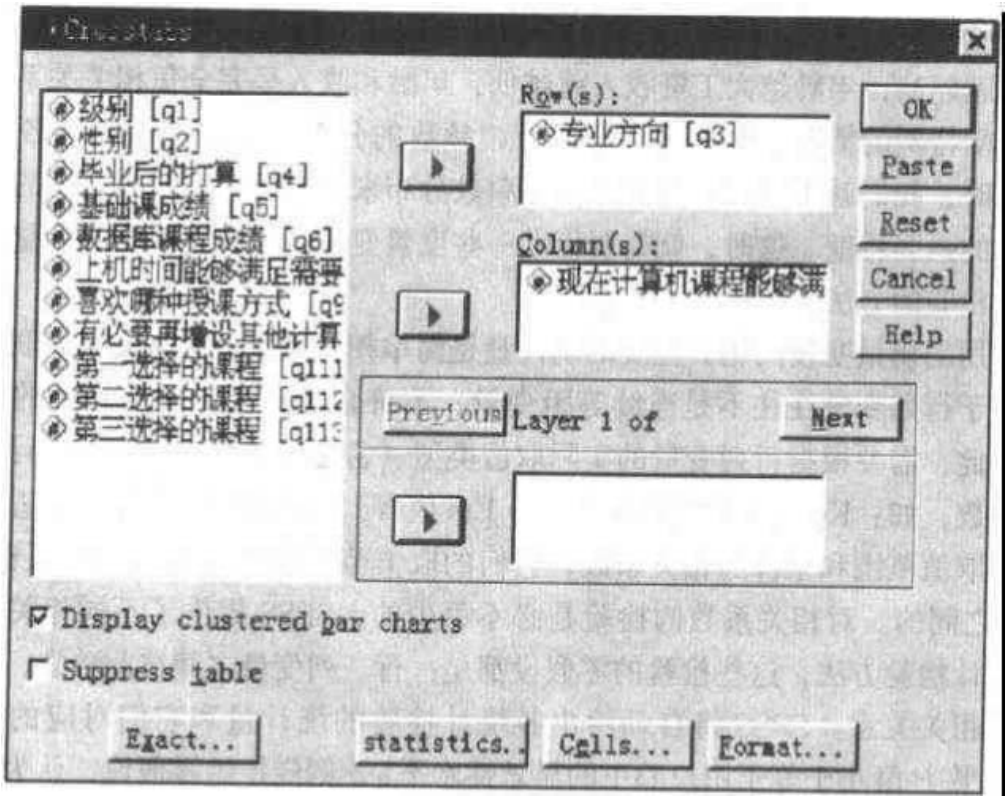


图 5—7　SPSS 交叉分组下的频数分析窗口

(2) 选择一个变量作为行变量到 Row 框中，选择一个变量作为列变量到 Column 框中。

(3) 如果有其他变量参与分析，可以将它们作为层控制变量选定到 Layer 框中。在指定层控制变量时，应根据实际分析的要求通过按动 Previous 或 Next 钮来确定它们的具体层次。例如：假设有 n 个层控制变量且它们都只有 2 个取值水平。如果这 n 个层变量均指定在同一层次上，那么这 n 个变量名会并排出现在同一个 Layer 框中，分析的结果将是 $2n$ 张交叉频数表；如果这 n 个层变量均指定在不同的层次上，那么它们将分别出现在不同的 Layer 框中，分析结果将产生 2 的 n 次方张交叉频数表，它们分别是第一层变量下的第二层变量下的直至第 n 层变量下的不同变量取值的交叉频数表。因此，在指定层变量时，应首先注意它们的层次关系，然后根据实际分析的需要确定控制变量的层次顺序。

(4) Display clustered bar charts 项用来指定是否显示各变量不同交叉取值下关于频数的棒图。

三、交叉分组下的频数分析的其他选项

在进行上述计算过程中，SPSS 还有一些其他选项供用户选择。

1. Cells 选项

SPSS 默认在频数分析表中输出实际的观察频数。但观察频数往往不能确切地反映事物的实质，因此还需要输出其他数据项，如：期望频数、行百分比、列百分比、总百分比等。

2. Format 选项

该选项用来指定交叉频数表的输出顺序。选项 Ascending 表示以行变量取值的升序输出频数表，SPSS 默认此选项；Descending 表示以行变量取值的降序输出频数表。

3. Statistics 选项和 Exact 选项

这两个选项提供了对行、列变量相关性检验的若干种方法，可以根据实际情况选用。

四、案例分析说明

对案例 2 作不同专业方向与对目前开设的计算机课程能否满足专业需要进行分析。

作交叉列联表分析，其中：行变量为专业、列变量为目前开设的计算机课程能否满足专业需要。产生的交叉列联表如输出结果 5—5 所示。

由输出结果 5—5 可了解到以下几种情况。

1. 总体情况

参加调查的共有 390 名学生，实际观察频数中：

● 250 人是第一专业方向，占总人数的 64.1%，140 人是第二个专业方向，占总人数的 35.9%；

● 有 140 人认为目前开设的计算机课程不能满足专业需要，占总人数的 35.9%，220 人认为基本能够，占总人数的 56.4%，30 人认为能够，占总人数的 7.7%。

2. 具体情况

(1) 专业方向一的学生共 250 人，对计算机课程设置的看法：

● 有 70 人认为不能满足，占该专业方向总人数的 28.0%（行百分比），占持“不能满足”意见总人数的 50%（列百分比），占总人数的 17.9%（总百分比）；

● 有 160 人认为基本满足，占该专业方向总人数的 64.0%（行百分比），占持“基本满足”意见总人数的 72.7%（列百分比），占总人数的 41.0%（总百分比）；

● 有 20 人认为能够满足，占该专业方向总人数的 8.0%（行百分比），占持

“能够满足”意见总人数的66.7%（列百分比），占总人数的5.1%（总百分比）；

输出结果 5—5　　案例 2 交叉列联表

			目前开设的计算机课程能否满足专业需要			Total
			不能	基本能够	能够	
专业方向	专业方向一	Count	70	160	20	250
		Expected Count	90.0	141.0	19.0	250.0
		% within 专业方向	28.0%	64.0%	8.0%	100.0%
		% within 能否满足	50.0%	72.7%	66.7%	64.1%
		% of Total	17.9%	41.0%	5.1%	64.1%
		Residual	-20.0	19.0	1.0	
		Std. Residual	-.7	.5	.1	
	专业方向二	Count	70	60	10	140
		Expected Count	50.0	79.0	11.0	140.0
		% within 专业方向	50.0%	42.9%	7.1%	100.0%
		% within 能否满足	50.0%	27.3%	33.3%	35.9%
		% of Total	17.9%	15.4%	2.6%	35.9%
		Residual	20.0	-19.0	-1.0	
		Std. Residual	.9	-.7	-.1	
Total		Count	140	220	30	390
		Expected Count	140.0	220.0	30.0	390.0
		% within 专业方向	35.9%	56.4%	7.7%	100.0%
		% within 能否满足	100.0%	100.0%	100.0%	100.0%
		% of Total	35.9%	56.4%	7.7%	100.0%

（2）专业方向二的学生共140人，对计算机课程设置的看法：

● 有70人认为不能满足，占该专业方向总人数的50.0%（行百分比），占持“不能满足”意见总人数的50.0%（列百分比），占总人数的17.9%（总百分比）；

● 有60人认为基本满足，占该专业方向总人数的42.9%（行百分比），占持“基本满足”意见总人数的27.3%（列百分比），占总人数的15.4%（总百分比）；

● 有10人认为能够满足、占该专业方向总人数的7.1%（行百分比），占持“能够满足”意见总人数的33.3%（列百分比），占总人数的2.6%（总百分比）；

由以上情况可见：

（1）专业方向一中持“基本满足”意见的比例远远高于其他两种意见的比

例，而专业方向二中持该观点的比例与持“不能满足”观点的比例基本持平。这说明，专业方向二的学生比专业方向一的学生更有多学一些计算机课程的愿望。

(2) 持“不能满足”观点的学生比例，专业方向二比专业方向一的高出将近1倍，同样也说明专业方向二的学生对计算机学习的渴望度较高。

可以从图5—8中直观地掌握数据的分布情况。

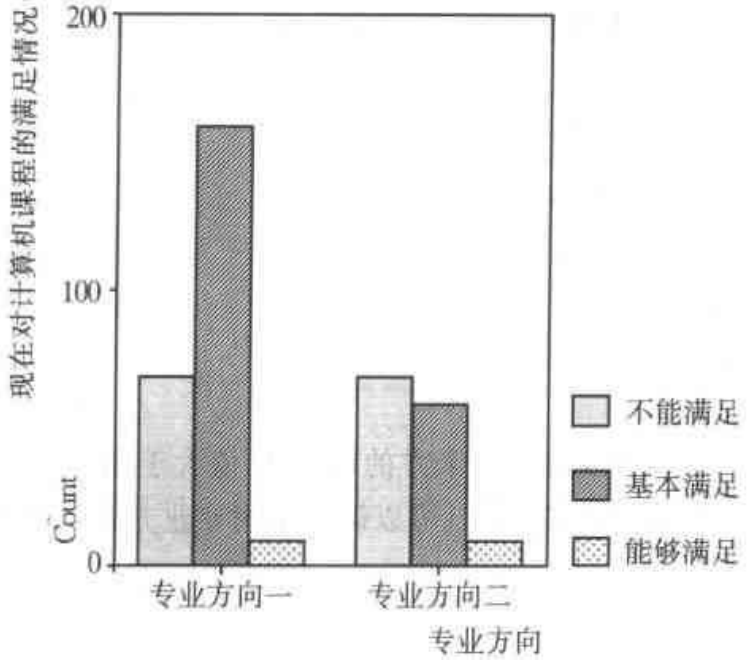


图5—8　案例2交叉分组下的频数分析图形结果

3. 不同的专业方向对回答该问题是否产生显著影响

从总体上讲，持“不能满足”、“基本满足”、“能够满足”观点的学生人数比例为35.9:56.4:7.7。如果仍然按照这个总体比例，专业方向一和专业方向二中持“不能满足”、“基本满足”、“能够满足”观点的学生人数比例仍还应为35.9:56.4:7.7。于是，专业方向一的250名学生中应分别约有90人（250×35.9%）、140人（250×56.4%）、19人（250×7.7%）持三种不同的观点。同理，专业方向二的140名学生中也应分别约有50人（140×35.9%）、79人（140×56.4%）、11人（140×7.7%）持三种不同的观点。这些数据就是各个单元格中的期望频数，实际观察频数与期望频数之差就是剩余，对剩余进行标准化处理便得到标准化剩余。

可见，不同专业方向中的期望频数所构成的分布是和总体分布相一致的分布，期望分布体现出：不同专业方向对该问题的回答并没有产生影响。如果列联表所有单元的实际观察频数整体上与期望频数之间的差距较小，在实际中则说明，不同的专业方向对回答该问题确实没有产生影响，这两个事物（行、列变

量）之间相互独立；相反，如果列联表所有单元的实际观察频数整体上与期望频数之间的差距较大，则说明，不同的专业方向对回答该问题产生了显著影响，这两个事物之间存在明显的相关关系。对于这个差距的度量可以用卡方值描述，从而从卡方分布中确定两事物（行、列变量）是否相互独立。

对案例 2 作卡方检验的结果如输出结果 5—6 所示。

输出结果 5—6　　案例 2 交叉分组下的频数分析卡方检验结果

	Value	Df	Asymp. Sig. (2-sided)
Pearson Chi-Square	1.930	2	.381
Likelihood Ratio	1.911	2	.385
Linear-by-Linear Association	1.282	1	.258
N of Valid Cases	390		

从输出结果 5—6 可以看出，卡方的概率 P 值大于用户心中的显著性水平（0.05 或 0.01），不能拒绝零假设，可以认为不同专业方向对回答该问题没有产生显著影响。

§5.5　多选项分析

一、什么是多选项问题

在问卷调查中，根据实际调查的需要，会要求被调查者从给出的可选择的答案中选择几个答案。

例如：假设设计了关于高考志愿的调查问卷，对考生进行调查，以分析考生的志愿趋向。其中的一道问题是：

高考时你希望报考的学校有：

（1）北京大学，（2）清华大学，（3）中国人民大学，（4）北京理工大学，（5）北京师范大学、（6）北京外国语大学。

请从中选择，学校数不能多于三所。

很显然，不同的考生所选择的答案个数会不同。有的考生可能选择北京大学和北京外国语大学两所大学，即答案是（1）和（6）两个。有的考生可能选择清华大学、北京理工大学和北京师范大学三所大学，即答案是（2）、（4）、（5）三

个，等等。

再例如：在案例 2 中，最后一道题要求学生在列出的计算机选修课程中根据自己的愿望按先后顺序挑选出三门课程。学生对这个问题的答案数明显也不只一个。

对于这种问卷题目，可选择的答案数据都是顺序水准或名义水准的，并且允许选择的答案是多个，有两个或两个以上答案会被同时选中。这种一道问题有两个或两个以上答案同时出现的问题，就是多选项问题。

二、分析多选项问题的一般方案

分析多选项问题的一般做法是：首先将多选项问题分解；然后用前面讲到的频数分析和交叉列联表分析的方法进行分析。

1. 多选项问题的分解

为什么要对多选项问题进行分解？

我们知道，在利用 SPSS 进行问卷调查分析时，对于单选问题，即一道问题只能选择一个答案的问题，一般将一道问题就设为一个 SPSS 的变量，用来存放该问题的一个答案。也就是说，SPSS 中的一个变量对每个问题的答案只能有一个取值。

但是，对于多选项问题，答案个数不只一个，如果也只设计一个 SPSS 变量，就无法同时存放这几个答案。即 SPSS 无法直接处理多选项问题。

一般讲，多选项问题的解决思路是设计一个较好的编码方案，对原始数据进行重新编码。即：将问卷中的一道多选项问题分解，转换成几个子问题，设置几个 SPSS 变量，分别存放描述这些问题的几个可能被选择的答案。

多选项问题的分解可以有两种方法：一种是多选项二分法（Multiple Dichotomies Method），另一种是多选项分类法（Multiple Category Method）。

（1）多选项二分法是将多选项问题中提供的每个可能的答案设为一个 SPSS 变量，每个变量只有 0 或 1 两个取值，分别表示选择了该答案和没有选择该答案。

例如上例中，利用多选项二分法将这一个问题分解成 6 个问题，设置 6 个 SPSS 变量，分别存放是否希望报考北京大学、是否希望报考清华大学、是否希望报考中国人民大学、是否希望报考北京理工大学、是否希望报考北京师范大学、是否希望报考北京外国语大学。其中：用 1 表示希望报考，用 0 表示不希望报考。如表 5—6 所示。

表 5—6　　多选项二分法

设置的变量名	变量名标签（含义）	编码方案（变量取值）	
T1	报考北京大学	0	1
T2	报考清华大学	0	1
T3	报考中国人民大学	0	1
T4	报考北京理工大学	0	1
T5	报考北京师范大学	0	1
T6	报考北京外国语大学	0	1

如果某考生选择了北京大学和北京外国语大学，则 T1 和 T6 的变量值为 1，其余变量值都为 0。如果某考生选择了清华大学、北京理工大学和中国人民大学，则 T2、T3、T4 的变量值为 1，其余变量取值为 0。

同理，我们也可以按照多选项二分法，将案例 2 最后一道题分解成 T1 至 T8 八道问题。

（2）多选项分类法是首先估计多选项问题最多可能同时出现的答案个数。然后，为每个答案设置一个 SPSS 变量，其变量取值为多选项问题中的可选答案。

例如上例中，我们知道考生的答案不超过 3 个，于是，我们就设置 3 个 SPSS 变量，分别表示志愿 A、志愿 B、志愿 C，变量的取值为 1 到 6。如表 5—7 所示。

表 5—7　　多选项分类法

设置的变量名	变量名标签（含义）	编码方案（变量取值）
V1	志愿 A	1：北京大学　2：清华大学　3：中国人民大学 4：北京理工大学　5：北京师范大学 6：北京外国语大学
V2	志愿 B	1：北京大学　2：清华大学　3：中国人民大学 4：北京理工大学　5：北京师范大学 6：北京外国语大学
V3	志愿 C	1：北京大学　2：清华大学　3：中国人民大学 4：北京理工大学　5：北京师范大学 6：北京外国语大学

如果某考生选择了北京大学和北京外国语大学，则 V1 的变量值为 1、V2 的变量值为 6，V3 为缺失数据。如果某考生选择了清华大学、北京理工大学和中国人民大学，则 V1 的变量值为 2，V2 的变量值为 4，V3 的变量值为 3。

同理，我们也可以将案例 2 最后一道题按多选项分类法分解成 V1 至 V3 三

道问题。

需要指出的是：在实际的多选项问题分析中，应根据分析问题的具体情况考虑采用上述哪种分解方案。上例中，如果采用多选项二分法分解问题，对 T1 至 T6 作频数统计，能够很方便地得到某学校的报考情况。但如果采用多选项分类法分解问题，对 V1 至 V3 作频数分析，则只能方便地得到志愿 A、志愿 B、志愿 C 中各学校的报考情况。再如案例 2，要求按照先后顺序填写答案。如果采用多选项二分法进行分解，并对 T1 至 T8 进行汇总，虽然方便地统计出各门课学生希望选修的情况，但却丢失了学生选修倾向的信息。如果采用多选项分解法，并对 V1 至 V3 进行汇总，就可以方便地得到学生的选修倾向数据，但却又无法方便地得到各门课程具体的希望报选情况。

由此可见，对于多选项问题，问题的分解方法、编码的设计方案极为重要。虽然在实际分析中，上面两种分解方法都可以使用，但却需要注意具体的分析问题和分析要求，采用更为合理的分解方法，尽量保证数据信息不丢失，以满足统计分析的需要。

2. 对多选项问题作普通频数分析或交叉列联表分析出现的困难

将多选项问题分解以后，一般可以直接利用频数分析或交叉列联表分析方法对分解后的变量，如：T1 至 T6 或 V1 至 V3 分别进行统计汇总。

但是，稍加注意就会发现，利用普通的频数分析或交叉列联表分析，作某些统计汇总时不方便，它与多选项问题的分解方法有较大关系。例如：关于高考志愿问题，如果采用多选项分类法分解问题，对 V1 至 V3 作频数分析，只能方便地得到志愿 A、志愿 B、志愿 C 中各学校的报考情况。如果希望了解某学校，如中国人民大学在所有考生中的报考人次情况，就还需要作许多手工工作，如：将 V1、V2、V3 中报考中国人民大学的人次作累计，再计算百分比等。否则，就需要重新采用多选项二分法对问题进行分解。但在实际工作中，编码方案修改并非一件容易的事。

那么，是否有一种分析方法，能够比较方便地完成上述手工工作呢？有，这就是多选项分析。

三、多选项分析处理多选项问题

1. 多选项分析的基本思路

对多选项问题采用上面谈到的方法进行频数分析，会比较烦琐麻烦。主要麻烦是需要手工进行一些额外计算。如果问卷中有许多这样的问题，那么可以想像手工工作量的负担。多选项分析则能够很好地解决这个问题。

多选项分析的基本思路是：首先按照多选项二分法或多选项分类法将多选项问题分解成若干子问题，设置若干个 SPSS 变量。然后对这些变量作多选项分析的频数分析或交叉列联表分析。在分析过程中，SPSS 会自动累计个案，将这些变量中有相同取值，即：有相同答案的个案数进行累计，得到的是累计频次。

例如：对高考志愿问题，按多选项分类法对问题作分解。现在希望从收回的 50 份问卷中统计出各学校的报考情况。如果分别对 V1 至 V3 作普通频数分析，假设有表 5—8 所示的统计结果。

首先，对 V1 作普通频数分析。

表 5—8（a）　　多选项问题的普通频数分析结果

变量值标签	变量值	频数	百分比（%）
北京大学	1	45	90.0
中国人民大学	3	5	10.0
合计		50	100.0

从表 5—8（a）的统计结果说明，在志愿 A 中，50 人中有 45 人希望报考北京大学，占总人数的 90%，有 5 人希望报考中国人民大学，占总人数的 10%。

第二，对 V2 作普通频数分析。

表 5—8（b）　　多选项问题的普通频数分析结果

变量值标签	变量值	频数	百分比（%）
清华大学	2	30	60.0
北京师范大学	5	20	40.0
合计		50	100.0

表 5—8（b）的统计结果说明，在志愿 B 中，50 人中有 30 人希望报考清华大学，占总人数的 60%，有 20 人希望报考北京师范大学，占总人数的 40%。

第三，对 V3 作普通频数分析。

表 5—8（c）　　多选项问题的普通频数分析结果

变量值标签	变量值	频数	百分比（%）
北京师范	5	10	20.0
中国人民大学	3	40	80.0
合计		50	100.0

表 5—8（c）的统计结果说明，在志愿 C 中，50 人中有 10 人希望报考北京

师范大学，占总人数的 20%，有 40 人希望报考中国人民大学，占总人数的 80%。

显然，上面三张表的任何一张表都无法说明各学校的报考情况。因为 V1、V2、V3 分别只是代表了考生的某种志愿选择，只有 3 个变量同时考虑才能够全面地体现考生的所有志愿，也才能准确统计出各学校的报考情况。再看表 5—9 所列的情况。

表 5—9　　多选项问题的多选项频数分析结果

变量值标签	变量值	频数	百分比（%）
北京大学	1	45	30.0
清华大学	2	30	20.0
中国人民大学	3	45	30.0
北京师范大学	5	30	20.0
合计		150	100.0

表 5—9 的统计结果是上面三张表综合出来的。这里的频数是人次，百分比也是人次百分比。从中可以知道，在所有 150 个志愿选择中，分别有 30%、20%、30%、20%的志愿选择了北京大学、清华大学、中国人民大学和北京师范大学。这个结果应该是全面的，反映了考生的志愿选择趋向。

再例如：仍然是有关高考的例子。与希望报考的原因作交叉列联：希望报考的原因假设有名气大（1）和其他原因（0）两种。如果用普通交叉列联表分别对 V1、V2、V3 作频数分析，假设有如表 5—10 所示的结果。

先对 V1 作交叉列联。

表 5—10（a）　　多选项问题的普通交叉列联分析结果

变量值标签和 变量值	名气大 (1)	其　他 (0)	Row Total
北京大学 (1)	30	15	45 90.0%
中国人民大学 (3)	3	2	5 10.0%
Column Total (%)	33 66.0	17 34.0	50 100.0%

表 5—10（a）的统计结果表明 50 人中，其中，报考北京大学的 45 人中，因为名气大报考的有 30 人，其他原因报考的有 15 人。报考中国人民大学的 5 人中因为名气大而报考的有 3 人，其他原因报考的有 2 人。因为名气大而选择志愿的有 33 人，占总人数的 66%，由其他原因决定志愿的有 17 人，占总人数的 34%。

再对 V2 作交叉列联。

表 5—10（b）　　多选项问题的普通交叉列联分析结果

变量值标签和 变量值	名气大 (1)	其　他 (0)	Row Total
清华大学 (2)	20	10	30 60.0%
北京师范大学 (5)	15	5	20 40.0%
Column Total (%)	35 70.0	15 30.0	50 100.0%

表 5—10（b）的统计结果表明 50 人中，其中：报考清华大学的 30 人中因为名气大而报考的有 20 人，因其他原因报考的有 10 人。报考北京师范大学的 20 人中因为名气大而报考的有 15 人，其他原因报考的有 5 人。因为名气大而选择志愿的有 35 人，占总人数的 70%，由其他原因决定志愿的有 15 人，占总人数的 30%。

最后对 V3 作交叉列联。

表 5—10（c）　　多选项问题的普通交叉列联分析结果

变量值标签和 变量值	名气大 (1)	其　他 (0)	Row Total
北京师范大学 (5)	8	2	10 20.0%
中国人民大学 (3)	20	20	40 80.0%
Column Total (%)	28 56.0	22 44.0	50 100.0%

表 5—10（c）的统计结果表明 50 人中，其中：报考北京师范大学的 10 人中因为名气大而报考的有 8 人，其他原因报考的有 2 人。报考中国人民大学的 40 人中因为名气大而报考的有 20 人，其他原因报考的有 20 人。因为名气大而选择志愿的有 28 人，占总人数的 56%，由其他原因决定志愿的有 22 人，占总人数的 44%。

上面三张表都无法全面地阐明志愿选择原因和报考学校的关系。再看表 5—11 所列的情况。

表 5—11　　多选项问题的多选项交叉列联表分析结果

变量值标签和 变量值	名气大 (1)	其　他 (0)	Row Total
北京大学 (1)	30	15	45 30.0%
清华大学 (2)	20	10	30 20.0%
中国人民大学 (3)	23	22	45 30.0%
北京师范大学 (5)	23	7	30 40.0%
Column Total (%)	96 64.0	54 36.0	150 100.0%

表 5—11 是由表 5—10 综合出来的，表明 150 人次中，报考北京大学的 45 人次中因为名气大而报考的有 30 人次，其他原因报考的有 15 人次；报考清华大学的 30 人次中因为名气大而报考的有 20 人次，其他原因报考的有 10 人次；报考中国人民大学的 45 人次中因为名气大而报考的有 23 人次，其他原因报考的有 22 人次。报考北京师范大学的 30 人次中因为名气大而报考的有 23 人次，其他原因报考的有 7 人次。因为名气大而选择志愿的有 96 人次，占总人数的 64%，由其他原因决定志愿的有 54 人次，占总人数的 36%。

可见，最后一张表比较准确客观地说明了选择志愿的原因和学校报考情况之间的关系。通过多选项分析中的交叉列联表分析可以方便地得到这个结果。

2. 多选项分析的基本步骤

在 SPSS 中，多选项分析的基本步骤是：

（1）定义多选项分析的多选项变量集，即：将多选项问题按上面谈到的方法分解成多个问题并设置多个变量后，将这些变量定义为一个集合。

（2）进行多选项的频数分析。

（3）进行多选项的交叉列联表分析。

下面对具体的操作步骤分别作说明。

3. 定义多选项变量集

定义多选项变量集的主要目的是为今后的多选项分析中的频数分析和交叉列联表分析作准备的。通过定义多选项变量集，SPSS 今后作多选项的频数分析或交叉列联表分析时就可以确定对哪些变量的相同取值作个案累计。

定义多选项变量集的基本操作步骤是：

（1）菜单选项：Statistics－＞Multiple Response－＞Define Sets。于是出现如图 5—9 所示的窗口。

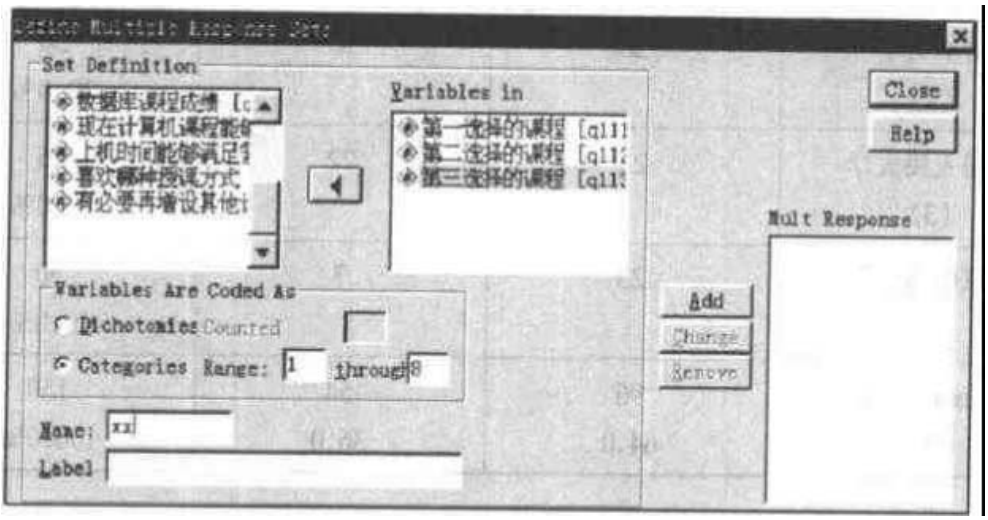


图 5—9　SPSS 多选项变量集定义窗口

（2）从若干个数值型变量中选取进入多选项变量集的变量到 Variables in sets 框中。将来这些变量将参加相同取值的个案累计计算。

（3）定义多选项变量集中的变量是按照哪种方法分解出来的（Variables Are Coded As）。如果是按多选项二分法（Dichotomies Counted）分解，还应输入一个二分值。SPSS 规定：等于该二分值为一组，其余的为另一组。这就表示，在多选项二分法中，变量的取值不一定为 0 或 1。如果是按多选项分类法（Categories）分解的，还应输入这些分解变量取值的最小值和最大值。

（4）为多选项变量集命名。系统会自动在该名字前加字符 $。

（5）按 Add 钮将一个多选项变量集加到 Mult Response 框中。SPSS 可以定义

几个多选项变量集。

4．多选项分析的频数分析

多选项分析的频数分析是对前面定义的多选项变量集中的分解变量作频数统计，它实现分解变量相同取值的个案数累计功能，并计算相应的百分比。

多选项分析的频数分析的基本操作步骤是：

（1）菜单选项：Statistics－＞Multiple Response－＞Frequencies。于是出现如图 5—10 所示的窗口。

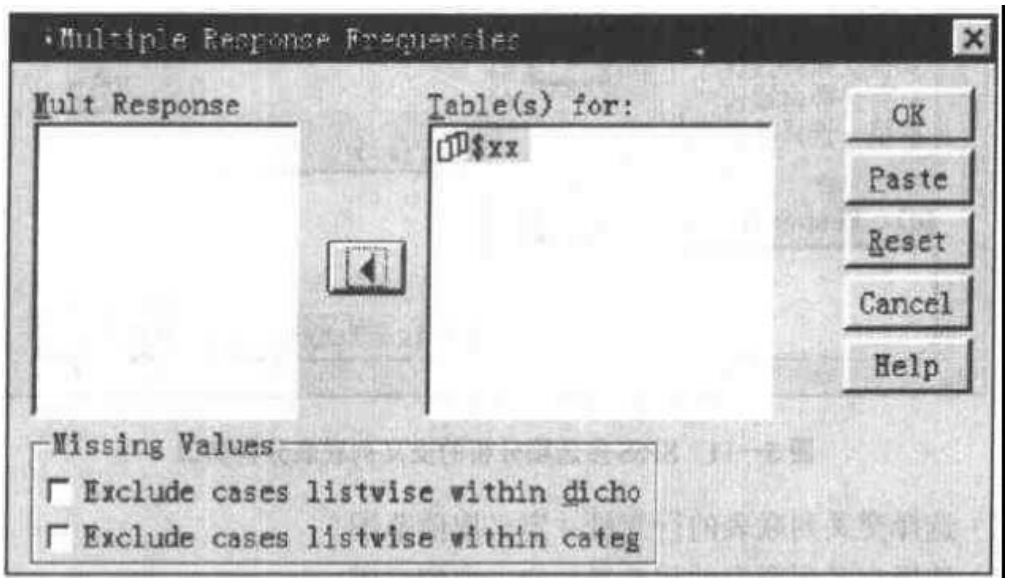


图 5—10　SPSS 多选项的频数分析窗口

（2）选择待分析的多选项变量集到 Tables for 框中。

（3）指定缺失数据处理方法的适用对象：SPSS 规定只要多选项变量集中有一个变量有缺失值就将该个案剔除。Exclude cases listwise within dichotomies 表示缺失值的处理方法适用于采用多选项二分法得到的分解变量的分析。Exclude cases listwise within categories 表示适用于采用多选项分类法得到的分解变量的分析。

通过以上步骤，我们就能够方便地得到类似表 5—9 的综合分析结果。

5．多选项分析的交叉列联表分析

多选项分析的交叉列联表分析是对前面定义的多选项变量集中的各分解变量作交叉列联表分析，也可以与其他变量作交叉列联。它实现分解变量同分组下个案数的累计功能，并计算相应的百分比。

多选项分析的交叉列联表分析的基本操作步骤：

（1）菜单选项：Statistics－＞Multiple Response－＞Crosstabs。于是出现如图5—11所示的窗口。

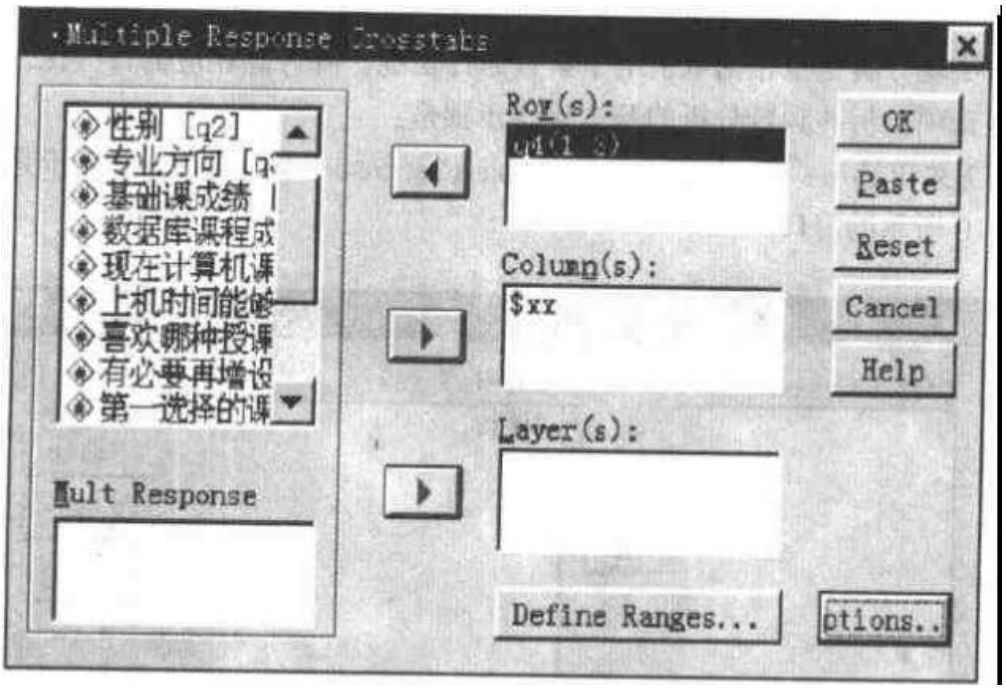


图5—11　SPSS多选项分析的交叉列联表分析窗口

（2）选择交叉列联表的行变量并定义取值范围。

（3）选择交叉列联表的列变量并定义取值范围。

（4）选择交叉列联表的分层变量并定义取值范围。

多选项分析的交叉列联表分析的Option选项：

（1）Cell Percentage：表示显示各单元中的百分比。

（2）Row：表示计算行百分比；Column：表示计算列百分比；Total：表示计算总百分比。

（3）Percentage Based on：指定如何计算百分比。Case：表示分母为个案数；Responses：表示分母为多选项的应答数。

（4）Match Variable Across Response Sets：表示如果交叉列联表的行、列变量均为多选项变量集时，第一个变量集的第一个变量与第二个变量集的第一个变量作交叉分组，第一个变量集的第二个变量与第二个变量集的第二个变量作交叉分组。依次类推。

四、案例分析说明

对案例2的第10题作多选项频数分析，得到输出结果5—7所示的分析

结果。

输出结果 5—7　　案例 2 多选项的频数分析结果

Group $ XX

Category label	Code	Count	Pct of Responses	Pct of Cases
计算机理论	1	30	2.9	8.8
计算机操作系统	2	80	7.8	23.5
计算机软件开发语言	3	150	14.7	44.1
计算机通讯技术	4	210	20.6	61.8
统计数据库系统	5	160	15.7	47.1
Internet 原理及网页制作	6	270	26.5	79.4
数据挖掘技术	7	90	8.8	26.5
数据仓库	8	30	2.9	8.8
Total responses		1020	100.0	300.0

50 missing cases; 340 valid cases

由输出结果 5—7 可见，390 名学生中有 50 人的答案中有缺失值，剔除后有效样本为 340 个，有1 020个答案。其中：学生最希望选修的课程是 Internet 原理及网页制作，占总人次的 26.5%。其次是计算机通讯技术、统计数据库系统、计算机软件开发语言，各占总人次的 20.6%、15.7%、14.7%。

对案例 2 的第 10 题与第 3 题（毕业以后的打算）作交叉列联表分析，得到输出结果 5—8 的分析结果。

输出结果 5—8　　案例 2 多选项分析的交叉列联表分析结果

* * * CROSSTABULATION * * *

$ XX (group)

by Q4　你毕业后的打算

	工作 1	报考研究生 2	出国 3	合计
计算机理论 1	0 0.0 0.0	30 100.0 3.7	0 0.0 0.0	30 2.9

续输出结果

	工作 1	报考研究生 2	出国 3	合计
计算机操作系统 2	30 37.5 20.0	40 50.0 4.9	10 12.5 16.7	80 7.8
计算机软件开发语言 3	20 13.3 13.3	110 73.3 13.6	20 13.3 33.3	150 14.7
计算机通讯技术 4	40 19.0 26.7	150 71.4 18.5	20 9.5 33.3	210 20.6
统计数据库系统 5	10 6.3 6.7	150 93.8 18.5	0 0.0 0.0	160 15.7
Internet 原理及网页制作 6	30 11.1 20.0	230 85.2 28.4	10 3.7 16.7	270 26.5
数据挖掘技术 7	20 22.2 13.3	70 77.8 8.6	0 0.0 0.0	90 8.8
数据仓库 8	0 0.0 0.0	30 100.0 3.7	0 0.0 0.0	30 2.9
Column Total	150 14.7	810 79.4	60 5.9	1020 100

Percents and totals based on respondents 340 valid cases; 50 missing cases

从输出结果 5—8 可见，毕业以后报考研究生的学生比例明显高于其他两种。在学生最倾向选修的 Internet 原理及网页制作课程中，有 85.2% 的人次毕业后将报考研究生，其他的学生比例分别为 11.1% 和 3.7%。

在毕业后准备工作的学生中，有 26.7% 的人次希望选修计算机通讯技术，占的比例是最高的，其次才是 Internet 原理和网页制作；在毕业后准备报考研究生的学生中，有 28.4% 人次希望选修 Internet 原理和网页制作，占的比例也是最高的，其次是计算机通讯技术和统计数据库系统；在准备毕业后出国的学生中，则倾向于计算机软件开发语言和通讯技术，其余为计算机操作系统和 Internet 原

理和网页制作，另对其他课程无人问津。可见，总体上讲，无论学生毕业后有怎样的打算，他们都普遍希望选修 Internet 原理及网页制作和计算机通讯技术课程，这也许体现了当今数字化、网络化社会的普遍现象。

§5.6 基本统计分析的报表制作

基本统计分析的结果可以以统计报表的样式展现。统计报表是统计数据的基本组织形式，也是展现统计分析结果的好方式。因此，制作统计报表在基本统计分析中是极常见的。SPSS 的制作用户统计报表功能允许用户根据自己的需要定义报表的各个组成部分，从而制作出各种用户满意的精美统计报表。

一、用户统计报表概述

1. 用户统计报表的组成

制作统计报表应首先了解表格的各个组成部分。表 5—12 是一张比较典型的统计报表。

表 5—12　　职工工资情况报表

男职工

		文化程度			
		本科	专科	高中	初中
职称	高级工程师				
	工程师				
	技术员				
	无技术职称				

注：截止到本年底

表 5—12 所示的统计表格的基本组成部分是：表格主标题（职工工资情况报表），表格副标题（男职工）：行标题（职称）、列标题（文化程度）、数据单元（基本工资）以及表格左上角部分、表注等部分组成。在 SPSS 中表格的主标题以 Title 指代、表格的左上角部分以 Corner 指代，表格的脚注以 Caption 指代。表格的其他组成部分的指代名称在不同类型的用户报表中略有差别，这将在后面具体讨论。

2. 用户统计报表的种类

SPSS的用户统计分析报表共有四种类型，它们是：基本报表（Basic Ta-

bles)、普通报表（General Tables）、频数报表（Tables of Frequencies）和多选项报表（Multiple response Tables）。我们这里重点讨论基本报表和普通报表。

二、制作用户基本报表

1. 什么是 SPSS 的用户基本报表

SPSS 的用户基本报表样式与表 5—12 所示统计报表样式完全一致。它将报表中的数据单元部分、副标题、行标题、列标题分别用 Summary、Separate、Down、Across 指代。在制作基本报表时，关键要领会 Summary、Separate、Down、Across 的含义。我们看一下 SPSS 对它们是如何具体规定的：

（1）Summary：指代报表中的汇总变量，用来指定对哪个变量进行统计汇总并输出结果。

例如：表 5—12 中的数据单元是就关于基本工资的统计汇总结果，其中的 Summary 就是指代基本工资。

用户基本报表中的汇总变量可以有多个，也可以没有。

（2）Down：指代报表中的行控制变量，用来指定表格的行维由哪个变量控制，也可以理解为：对 Summary 汇总变量在行维上按哪个变量的取值进行分组汇总。

例如：表 5—12 中，行标题是职称，四个不同的职称取值决定了报表由四行组成，意味着将对基本工资这个汇总变量在报表行维上按不同职称作分组汇总。因此，职称是行控制变量。

用户基本报表中行控制变量可以是多个，也可以没有。

（3）Across：指代报表中的列控制变量，用来指定报表的列维由哪个变量控制，也可以理解为：对 Summary 汇总变量在列维上按哪个变量的取值进行分组汇总统计。

例如：表 5—12 中，列标题是文化程度，四个不同的文化程度取值决定了报表由四列组成，意味着将对基本工资这个汇总变量在报表列维上按不同文化程度作分组汇总。因此，文化程度是列控制变量。

用户基本报表中列控制变量可以是多个，也可以没有。

（4）Separate：指代报表中表控制变量。SPSS 可以同时产生若干张表格，这些报表除了表格副标题不同之外，其他部分的含义都是相同的。因此，我们可以将这些表看做是一个整体，其中的每张表就是这个整体中的一张子表。于是，在这个意义上，统计报表就不再是仅有行维和列维的二维表，而是具有表维的三维表了，报表的副标题便成为表维的标志。

SPSS表格中的 Separate 的实质是在指定报表的表维由哪个变量控制，也可以理解为：对 Summary 汇总变量在表维上按哪个变量的取值进行分组汇总。

例如：表 5—12 中仅示意了男职工基本工资情况表，实际还应有一张女职工工资情况的表格。如果将这两张报表作为一个整体来看待，于是在表向上就形成了由“性别”变量控制的第三个维——表维。两个不同的性别取值决定了该表由两张子表组成，意味着将对基本工资这个汇总变量在报表表维上按性别作分组汇总。因此，性别是表控制变量。

用户基本报表中的表控制变量可以有多个，也可以没有。

总之，制作用户基本报表的关键就是指定汇总变量、行控制变量、列控制变量和表控制变量。SPSS 将在报表的行维、列维和表维的各交叉分组下对汇总变量作统计汇总。

2．基本报表的功能

制作基本报表是将 Data editor 窗口中的数据按用户指定的行控制变量、列控制变量以及表控制变量的不同取值分别进行交叉分组后，作各交叉分组下的个案数统计或对所有汇总变量做相同的统计汇总并输出统计结果。

3．制作基本报表的基本操作步骤

制作基本报表的基本操作步骤是：

(1) 菜单选项：Statistics－＞Custom tables－＞Basic tables。于是出现如图 5—12 显示的窗口。

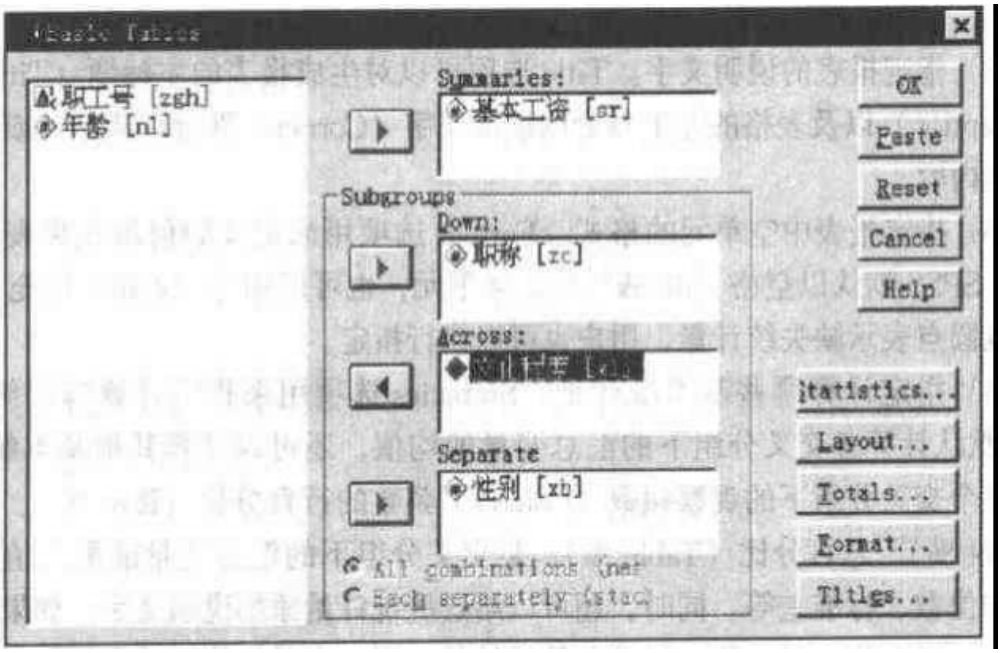


图 5—12　SPSS 制作用户基本报表窗口

(2) 指定汇总变量到 Summaries 框中。如果不指定汇总变量，SPSS 仅作各交叉分组下的个案数统计。

(3) 指定行控制变量到 Subgroups 框中的 Down 框中。如果不指定行控制变量，则汇总变量将不再在报表行维上作交叉分组汇总。

(4) 指定列控制变量到 Subgroups 框中的 Across 框中。如果不指定列控制变量，则汇总变量将不再在报表列维上作交叉分组汇总。

(5) 指定表控制变量到 Subgroups 框中的 Separate 框中。如果不指定表控制变量，则汇总变量将不再在报表表维上作交叉分组汇总。

于是，SPSS 将用户定义的汇总变量在表维、行维、列维上计算各交叉分组下的平均值，并输出相应的统计报表。

4. 制作基本报表的其他选项

如果用户对 SPSS 默认制作出的统计报表中的某些组成部分不满意，可以自己定义。

(1) 指定多个控制变量之间的层次关系。如果用户指定的表控制变量、行控制变量、列控制变量的个数不止一个，SPSS 默认这些变量之间是叠加交叉分组关系，即在原有的交叉分组的基础之上再作进一步的交叉分组，实质上是又产生了新的虚拟维。默认选项是 All Combination (nests)；

如果不希望是这样可以选择 Each Separately (Stack) 选项，表示变量之间是并列存在关系，是在原有交叉分组的基础之上，分别和这些并列变量作交叉分组统计。

(2) 指定报表的说明文字。Title 选项可以对生成报表的主标题 (Title)、表注 (Caption) 以及表格的左上角处的说明文字 (Corner) 进行重新定义或添加新的描述内容。

(3) 指定报表中空单元的格式。Format 选项用来定义如何填充报表中的空单元。SPSS 默认以空格 (Blank) 填充空单元，也可以用零 (Zero) 填充。SPSS 默认以圆点表示缺失统计量，用户也可以自行指定。

(4) 指定计算哪些基本统计量。Statistics 选项用来指定计算哪些统计量。SPSS 默认计算各交叉分组下的汇总变量的均值，还可以计算其他基本统计量。如：某个交叉分组下的观察频数 (Count)、频数的行百分比 (Row %)、列百分比 (Col %)、总百分比 (Table %)、某交叉分组下的汇总变量的最大值、最小值、中位数、标准差等。同时，还可以给这些统计量添加说明文字。如果计算诸如均值、最大值、最小值、标准差等统计量，则一定是计算汇总变量的这些统计量。如果计算的是诸如个案数、个案百分比等统计量，则计算结果与汇总变量没

有关系。

(5) 指定表格的外观。Layout 选项允许用户自行定义汇总变量名的标签、统计量说明文字的显示输出位置等。其中：

● Summary Variable Labels 框定义汇总变量名标签文字（如：基本工资）显示哪个维上。Down the left side 表示在行维上显示；Across the top 表示在列维上显示；In Separate Tables 表示在表维上显示。

● Groups in Summary Variable Dimension 框表示在指定了汇总变量名标签的显示维以后，具体文字的显示位置。All summary variables under each groups 表示在维的每一个分组下都显示一个汇总变量名标签；All groups under each summary variables 表示仅在列维的最上面一行或在行维的最左边显示一个汇总变量名标签。

● Statistics label 框定义计算的统计量的文字说明信息（如：Mean）显示在哪个维上。各选项的含义同上。Labels groups with value labels only 选项表示在统计报表中仅显示表控制变量、行控制变量、列控制变量的具体变量值标签，而不再输出它们的变量名标签信息。

(6) 指定计算汇总结果。Totals 选项用来在统计表中定义输出行、列各个分类的汇总结果和整个表的汇总结果。Totals over each group variable 表示实现前者功能；Table margin totals 表示实现后者功能。

三、制作普通报表

SPSS 的普通报表可以实现基本报表的功能，并且功能更加丰富。用户通过观察基本报表会发现，虽然基本报表中的汇总变量可以是多个，但 SPSS 只能对它们作同一种或几种统计计算。如：对所有汇总变量都计算各交叉分组下的均值、极差。如果希望对不同的汇总变量分别计算不同的统计量，则需要通过制作普通报表来实现。

1. 什么是 SPSS 的用户普通报表

普通报表的样式与表 5—12 所示的典型报表的样式基本类似。报表的行控制变量、列控制变量、表控制变量的指代称谓分别是 Rows、Columns 和 Layers。虽然它们用户基本报表中的称谓相差较大，但含义是相同的。

普通报表中没有专门的汇总变量的英文指代，这点与基本报表不同。如果不指定汇总变量，则 SPSS 只能作各交叉分组下的个案数统计。否则，就需要指定汇总变量。那么 SPSS 如何知道对哪些变量作统计汇总呢？这里，SPSS 规定，可以指定行维或者列维上的某些变量为汇总变量，并且汇总变量只能在同一个维

上，即：如果行（列）维上有汇总变量，则列（行）维上就不能再指定汇总变量，那么，列（行）维上的变量只能作交叉分组用。那么 SPSS 又如何区分行维或列维上哪些变量是用作分组的，哪些变量是要被汇总的呢？这就需要看以下的具体操作步骤了。

2. 普通报表的功能

制作普通报表是将 Data editor 窗口中的数据按用户指定的行控制变量、列控制变量、表控制变量的不同取值分别进行交叉分组，作各交叉分组下的个案数统计或对汇总变量计算基本统计量。这里，普通报表能够实现对行维或列维上指定的各汇总变量计算各自不同的统计量。

3. 制作普通报表的基本操作步骤

（1）菜单选项：Statistics－＞Custom Tables－＞General Tables。于是出现如图 5—13 所示的窗口。

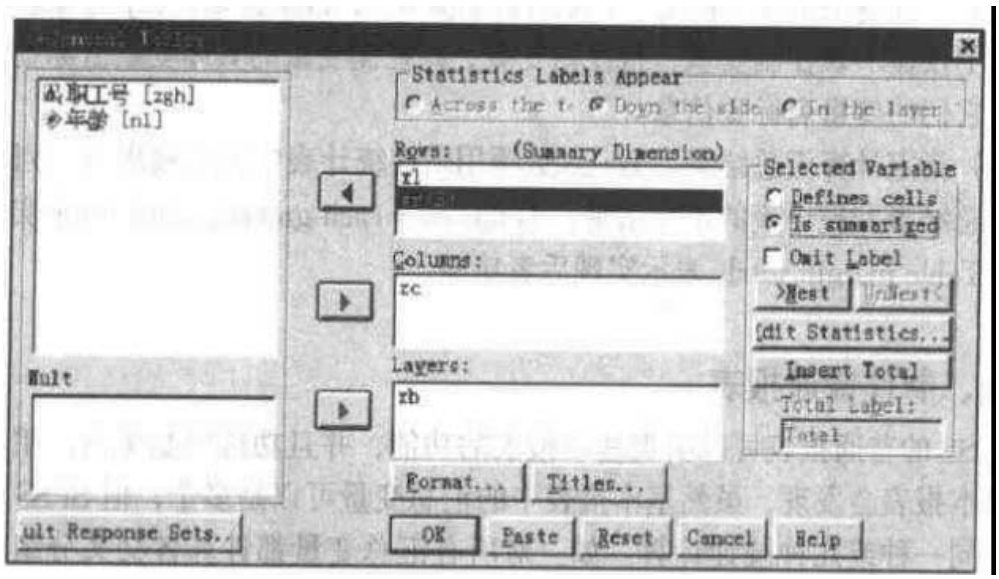


图 5—13　SPSS 制作用户普通报表窗口

（2）指定报表行控制变量到 Rows 框中。行控制变量可以是多个。如果没有行控制变量，则意味着不在行维上作交叉分组。

（3）如果行维中有用户希望汇总的变量，则在鼠标点击选中该变量名后，点中 Selected Variables 框中的 Is summarized 选项。于是，该相应变量名后面出现 (S) 标记，表示以后将对该变量计算交叉分组下的均值。行维中其他非汇总变量，它们在 Selected Variables 框中的对应选项是 Defines cells，将参与各维上的交叉分组。同时，Rows 框中也会出现 Summary Dimension 的标记，表示汇总变量在该维上。否则，略去此步。

(4) 指定报表列控制变量到 Columns 框中。列控制变量可以是多个。如果没有列控制变量，则意味着不在列维上作交叉分组。

(5) 如果行维不是汇总维，则可以在此按 (3) 步中说明的方法指定列维上的某些变量为汇总变量，列维上的其他非汇总变量也将参与各维上的交叉分组。同样，Columns 框中也同样会出现 Summary Dimension 的标记，表示列维是汇总维。否则，略去此步。

于是，SPSS 便可以自动生成表 5—13 所示的一张普通统计分析报表。

表 5—13　　职工工资情况报表

男职工

		职　称			
		高级工程师	工程师	技术员	无技术职称
文化程度	本科	1	1	2	
	专科	1	5	1	
	高中	1		3	
	初中			6	
基本工资		1 024.0	916.0	859.80	

从表 5—13 的统计结果可以看出，在制作这张表时，表控制变量是性别，列控制变量是职称，行控制变量有文化程度和基本工资两个变量，且行维是汇总维。其中的基本工资是汇总变量，而文化程度则分别与性别、职称作交叉分组。汇总结果是各交叉分组下的个案个数以及性别和职称交叉分组下的平均基本工资。可以看出，这里没有给出性别、职称、文化程度三个变量交叉分组下的平均基本工资。如果希望得到这个结果，或对报表的其他部分作修改，请看下面的其他选项。

4. 制作普通报表的其他选项

制作普通报表还有一些其他选项。其中：Format、Title 选项与制作基本报表中的 Format、Title 选项完全相同；Edit Statistics 选项与制作基本报表中的 Statistics 选项也相同，只是在操作上，应首先选中某变量，然后再重新指定对该变量计算哪个基本统计量。具体含义这里不再重复。

(1) 指定变量层次关系。Nest 和 Unnest 选项：前面已经讲到，SPSS 允许表维、行维、列维上有多个控制变量，如表 5—13 中行维上有文化程度和基本工资两个控制变量。如果仅按照前面的基本操作步骤制作报表，从上表得到的结果可以看出，文化程度和基本工资在并列的。如果想得到性别、职称和不同文化程度下的平均基本工资的数据，就需要将基本工资置于文化程度的下层，这时就应使

用 Nest 选项。具体操作是：首先选中基本工资，然后按 Nest 选项按钮。Unnest 选项与 Nest 选项作用相反。因此可见，Nest 选项能够将某维中的若干个控制变量定义为叠加交叉分组关系，Unnest 是定义为并列存在关系。

（2）Omit Label 选项。如果该选项前打上了钩，则表示报表中不输出变量名标签信息。

（3）Inset Total 选项。可以选择某维上的用于交叉分组的控制变量，计算个案合计数。合计项的默认输出标题是 Total，可以重新输入修改。对汇总变量计算个案合计是无意义的。

（4）Statistics Labels Appear 框中的选项。该选项只能用在没有指定汇总变量的情况下。此时，SPSS 默认计算各交叉分组下的个案数，也可以利用 Edit Statistics 选项重新指定计算诸如百分比等统计量。这些统计量的标签可以出现在表维上（In the layer），也可以在行维上（Down the side），或在列维上（Across the top）。

第 6 章

SPSS 的参数检验

§6.1 统计推断的基本方法

一、什么是统计推断

参数检验是统计推断内容的一个组成部分。统计推断是根据已经收集到的样本数据推断样本来自的总体的分布或总体均值、方差等总体统计参数。不直接从总体出发，而根据样本数据推断总体的原因是：一是总体数据无法全部收集到。例如：产品质量的检测问题。如：要检测灯泡的寿命，或要检测某产品所含成分等问题，我们无法对所有产品作实验，而只能从中随机抽取一部分产品（样本）作检测；二是总体数据虽然能够收集到，但如果收集到总体全部数据将要耗费大量的人力和财力。例如：要调查某市小学一年级学生在一学年中的平均作业时间，或调查某市居民 10 月份假期旅游情况等。针对这类大规模的统计调查工作，在实际处理中为节约开销往往也采用抽样技术，并根据样本数据推断总体的情况。

统计推断一般是在两种假设条件下进行的。一种是：假设总体分布已知，根据样本数据推断总体的统计参数，这类统计推断称为参数检验；另一种是：假设

总体分布未知，根据样本数据推断总体的分布或总体的统计参数，这类统计推断称为非参数检验。本章重点讨论参数检验，关于非参数检验将在后面的章节中讨论。

二、统计推断的基本方法

SPSS 中统计推断采用的是假设检验的方法，其基本思路和方法是：

(1) 根据推断检验的目标，对待推断的总体参数或分布作一个基本假设 H_0。

(2) 利用收集到的样本数据和基本假设计算某检验统计量，该统计量服从或近似服从某种统计分布。

(3) 根据该统计量的值得到对应的相伴概率值。某统计量的相伴概率是指该检验统计量在某个特定的极端区域取值在 H_0 成立时的概率。

(4) 作出判断。用户给定一个显著性水平 α。如果相伴概率值小于或等于用户给定的显著性水平 α，则拒绝 H_0。否则，如果相伴概率值大于用户给定的显著性水平 α，则不能拒绝 H_0。

其中，第一步是 SPSS 事先约定好的，第二步和第三步也都是 SPSS 自动完成的。只有第四步需要人工干预，用户只需要根据实际情况，确定一个显著性水平 α，然后与相伴概率值进行比较判断，从而决定是拒绝还是不拒绝零假设。

这里，显著性水平 α 是指当零假设正确时人们却将它拒绝了的概率，即弃真的概率，是拒绝零假设犯错误的概率。我们本着认为零假设是正确的，不应拒绝它的角度出发，认为拒绝零假设一个小概率事件。如果在零假设成立的前提之下，根据样本数据计算出来的相伴概率值小于或等于用户给出的显著性水平，则认为是一个更小的小概率事件。根据小概率事件在一次实验中不可能发生的原则和信念，既然是一个小概率事件，但它居然却在一次实验真的发生了，这时我们就有理由怀疑零假设的正确性，而拒绝零假设成立。从另一个角度讲，既然显著性水平 α 是拒绝零假设犯错误的概率，如果相伴概率值小于或等于 α，则可以认为拒绝零假设犯错误的概率更小，因此应该拒绝零假设。否则，如果相伴概率值大于显著性水平 α，则可以认为拒绝零假设犯错误的概率较大，所以不应该拒绝零假设。

SPSS 参数检验中，最关键一步就是要通过相伴概率值与显著性水平的比较，来决定拒绝还是不能拒绝零假设。

§6.2 单样本 t 检验

一、什么是单样本 t 检验

SPSS 单样本 t 检验是检验某个变量的总体均值与指定的检验值之间是否存在显著差异。这里，前提要求是样本来自的总体应服从正态分布。

例如：通过随机抽样得到若干个周岁儿童身高的样本数据。现在需要通过这些样本数据，分析周岁儿童的平均身高是否为 75 厘米。

再例如：利用第 5 章案例 1 各学科人力投入总数的数据，分析各学科的人力投入平均水平是否与10 200大致相当。

这两个例子都是关于均值检验的。由于身高数据和人力投入数据总体上都可以认为服从正态分布，因此，可以通过单样本 t 检验的方法来实现。

二、单样本 t 检验的基本实现思路

单样本 t 检验的零假设为 H_0：总体均值与指定检验值之间不存在显著差异。采用 t 检验方法，按式（6.1）计算 t 统计量。

$$t=\frac{\bar{D}}{S/\sqrt{n}} \tag{6.1}$$

式中，$\bar{D}$ 是样本均值与检验值的差。因为总体方差未知，用样本方差 S 代替总体方差。n 为样本数。SPSS 将自动计算 t 值，由于该统计量服从 $n-1$ 个自由度的 t 分布，SPSS 将根据 t 分布表给出 t 值对应的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则拒绝 H_0，认为总体均值与检验值之间存在显著差异。相反，相伴概率值大于用户心中的显著性水平 α，则不应拒绝 H_0，可以认为总体均值与检验值之间不存在显著差异。

三、SPSS 单样本 t 检验的操作步骤

（1）菜单选项：Statistics－＞Compare means－＞One－samples t test。于是出现如图 6—1 所示的窗口。

（2）在 Test 框中输入检验值。

SPSS 单样本 t 检验的 Option 选项是用来指定输出内容和关于缺失值的处理方法。其中：

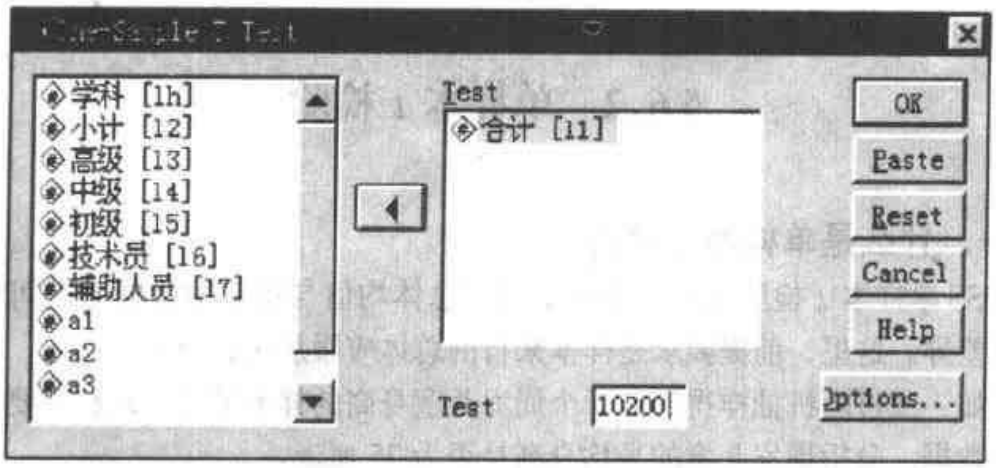


图 6—1　SPSS 单样本 t 检验窗口

● Confidence interval 表示默认输出 95% 的差值置信区间，可以重新修改和指定。

● Missing values 框中，Exclude cases analysis by analysis 表示：当分析计算时涉及到含有缺失值变量，则剔除在该变量上是缺失值的个案；Exclude cases list-wise 表示：剔除所有含缺失值的个案后再进行分析。注意，以后的分析中，关于缺失值的处理办法与此相同，以后就不再重复解释。

四、案例分析说明

利用第 5 章中 1999 年全国普通高等院校人文社科研究数据，分析 1999 年各学科的人力投入平均水平是否与 1998 年的 10200 人大致相当进行单样本 t 检验，结果如输出结果 6—1 所示。

输出结果 6—1　　案例 1 单样本 t 检验分析结果

One-Sample Statistics

	N	Mean	Std. Deviation	Std. Error Mean
合计	22	10738.14	11724.20	2499.61

One-Sample Test

	Test Value = 10200					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
合计	.215	21	.832	538.14	-4660.08	5736.36

由输出结果 6—1 可以看出，1999 年各学科人力投入平均值为10 738.14，标准差为11 724.20，均值标准误差为2 499.61。检验值为10 200，样本均值与检验值的差为 538.14，根据公式计算出的 t 统计量的值为 0.215，得到的相伴概率值为 0.832。差值 95% 的置信区间跨 0，为（−4 660.08，5 736.36），表示 95% 的样本差值将落在这个区间内。如果假设用户心中的显著性水平 α 为 0.05，由于相伴概率值大于 α，则不能拒绝 H_0，可以认为 1999 年各学科人力投入的平均水平与 1998 年相比，没有出现显著变化。

§6.3 两独立样本 t 检验

一、什么是两独立样本 t 检验

两独立样本 t 检验就是根据样本数据对它们来自的两独立总体的均值是否有显著差异进行推断。这个检验的前提要求是：

（1）两样本应是相互独立的，即从一总体中抽取一批样本对从另一总体中抽取一批样本没有任何影响，两组样本的个案数目可以不同，个案顺序可以随意调整。

（2）样本来自的两个总体应服从正态分布。

例如：从北京和上海两个城市，分别随机抽取若干个周岁儿童身高的样本数据，现在需要分析两城市周岁儿童的平均身高是否存在显著差异。

这个问题是一个两独立样本 t 检验的问题，它满足检验的前提要求。首先，北京周岁儿童与上海周岁儿童这两个总体显然是独立的，体现在：分别在两城市抽取随机样本，抽样不会相互影响，且两城市中抽取的样本个数可以是不同的。其次，儿童身高数据可以认为服从正态分布。

再例如：通过对案例 2 的项目分析，研究学生在决定选修哪门课程时，是否受到问卷中提及的其他因素的影响。通过后面对案例分析的具体讨论，可以发现这也是一个两独立样本的 t 检验问题。

二、两独立样本 t 检验的基本实现思路

两独立样本 t 检验的零假设为 H_0：两总体均值之间不存在显著差异。

在具体实现时通过两步完成：第一，利用 F 检验判断两总体的方差是否相等；第二，根据前一步的判断结果决定 t 统计量和自由度的计算公式，进而对 t

检验的结论作出判断。

1. 判断两总体的方差是否相等

判断两总体方差是否相等仍然用假设检验的方法来实现，它的 H_0 假设是：两总体方差无显著差异。SPSS 采用 Levene F 方法进行检验。首先，对来自两个不同总体的样本分别计算样本均值，计算每个样本与本组样本均值的差，并对差取绝对值。于是得到两组绝对值差值序列；然后，利用单因素方差分析方法，判断这两组绝对差值序列之间的均值是否存在显著差异，即判断平均离差是否存在显著差异，从而间接判断两组方差是否存在显著差异。

在检验过程中，SPSS 将自动计算 F 统计量，并依据 F 分布表给出统计量对应的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则拒绝方差分析的零假设，认为这两组绝对差值序列的均值存在显著差异，即：反映离散程度的指标——平均的绝对离差之间存在显著差异，从而必然得到两总体方差不相等的结论；相反，如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝方差分析的零假设，可以认为方差相等。

2. 根据前一步的判断结果决定 t 统计量和自由度的计算公式

(1) 在两总体方差未知且相等的情况下，t 统计量计算公式：

$$t=\frac{\bar{\chi}_1-\bar{\chi}_2}{\sqrt{S_p^2/n_1+S_p^2/n_2}} \tag{6.2}$$

式中：

$$S_p^2=\frac{(n_1-1)\ S_1^2+(n_2-1)\ S_2^2}{n_1+n_2-1} \tag{6.3}$$

这里，t 统计量服从 n_1+n_2-2 个自由度的 t 分布。

(2) 在两总体方差未知且不相等的情况下，t 统计量计算公式：

$$t=\frac{\bar{\chi}_1-\bar{\chi}_2}{\sqrt{S_1^2/n_1+S_2^2/n_2}} \tag{6.4}$$

这里，t 统计量仍服从 t 分布，但自由度采用修正的自由度：

$$f=\frac{(\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2})^2}{\frac{(\frac{S_1^2}{n_1})^2}{n_1}+\frac{(\frac{S_2^2}{n_2})^2}{n_2}} \tag{6.5}$$

通过式 (6.2) 和式 (6.4) t 统计量的计算公式不难看出，如果待检验的两样本均值差异较小，即均值差与零没有显著差异，t 值较小，则说明两样本的均值不存在显著差异；相反，如果待检验的两样本的均值差异较大，即均值差与零

有显著差异，t 值较大，则说明两样本的均值存在显著差异。

那么究竟 t 值达到怎样的数量才可以下结论呢？由于 t 统计量服从 t 分布，SPSS 将自动计算 t 统计量并依据 t 分布表给出相应的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则拒绝 t 检验的零假设，认为两总体均值存在显著差异；相反，如果相伴概率大于用户心中的显著性水平 α，则不能拒绝 t 检验的零假设，可以认为两总体均值不存在显著差异。

在 SPSS 中进行两独立样本 t 检验时，应首先对 F 检验作判断；如果方差相等，看分析结果中 equal 行的 t 检验相伴概率值；如果方差不相等，看 unequal 行的 t 检验相伴概率值，从而判断两独立样本的均值是否存在显著差异。

三、SPSS 两独立样本 t 检验的操作步骤

在利用 SPSS 做两独立样本 t 检验之前，应首先准备好待分析的数据。这里，虽然分析的是两独立样本，但在数据组织时，SPSS 要求两独立样本的数据必须存放在一个 SPSS 变量中，即存放在一个 SPSS 列上。为标识哪些样本是属于哪个总体的，还必须事先准备另一个 SPSS 变量，它存放的是标识数据，可以标识每个样本的所属总体，进而保证 SPSS 能够对样本来自的总体加以区分。

SPSS 两独立样本 t 检验的基本操作步骤是：

(1) 菜单选项：Statistics－＞Compare means－＞Independent-samples t。于是出现如图 6—2 所示的窗口。

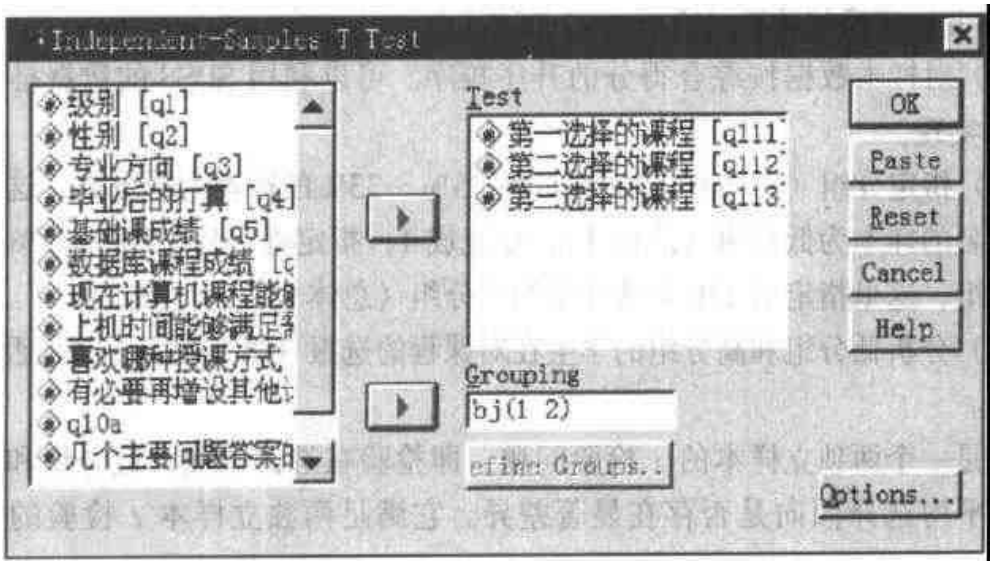


图 6—2　SPSS 两独立样本 t 检验窗口

(2) 选择若干变量作为检验变量到 Test variables 框。

(3) 选择一个变量作为标识变量到 Grouping 框中。该标识变量的两个不同取值可对应两个不同的总体。

(4) 按 Define Groups 钮定义标识变量的标识情况。其中：Use specified values 表示分别输入两个值，每个值代表一个总体；Cut point 表示标识变量为连续变量。输入一个数字，大于等于该值的为一个总体，对应一组样本。小于该值的为另一总体，对应另一组样本。

SPSS 两独立样本 t 检验的 Option 选项含义与单样本 t 检验中的相同。

四、案例分析说明

利用第 5 章中的案例 2 作项目分析，研究学生在决定选修哪门课程时，是否受到问卷中提及的其他因素的影响。分析的主要步骤是：

(1) 对第 9 题“你认为有必要再增设其他计算机课程吗?”重新记分。

由于第 9 题是一个反向记分题，在分析时应重新记分，即将变量值 1 转换成 3、2 仍为 2、3 转换成 1。可以利用 SPSS 的数据分组（Recode）功能实现。

(2) 对第 4、5、6、7、8 题和重新记录分后的第 9 题，计算答案的综合得分。

综合得分可以直接利用这些变量的变量值计算得到。变量值为 1，得分为 1；变量值为 2，得分为 2，依次类推。综合得分为 6 道问题得分的合计。可以利用 SPSS 的变量转换的功能（Compute）实现。

(3) 对样本数据按综合得分的升序排序。可以利用 SPSS 的数据排序功能（Sort）实现。

(4) 指定分组（总体）。指定得分前 25%～33%的样本为低分组，这里指定前 110 名的样本为低分组（总体 1），标记成 1；指定得分后 25%～33%的样本为高分组，这里指定后 110 名的样本为高分组（总体 2），标记成 2。

(5) 分析低分组和高分组的学生在对课程的选报（第 9 题）方面是否存在显著差异。

这是一个两独立样本的 t 检验问题，即检验在选课方面，低分组和高分组学生的平均选择倾向是否存在显著差异。它满足两独立样本 t 检验的前提要求。

这里，仅列出对第一选择课程和第三选择课程的 t 检验结果，如输出结果 6—2 所示。

输出结果 6—2　　案例 2 两独立样本 t 检验的分析结果

第一选择课程的 t 检验：

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality f Means						
	F	Sig.	T	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95 % Confidence Interval of the Difference	
								Lower	Upper
Equal variances assumed	0.091	0.766	0.330	20	0.744	0.272 7	0.822 2	−1.442 4	1.987 8
Equal variances not assumed			0.330	19.913	0.744	0.272 7	0.822 2	−1.442 4	1.987 8

对第三选择课程的 t 检验：

Independent Samples Test

	Levene's Test for Equality of Variances		t-test for Equality f Means						
	F	Sig.	T	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95 % Confidence Interval of the Difference	
								Lower	Upper
Equal variances assumed	0.701	0.414	1.283	17	0.217	1.022 2	0.7170	−0.6594	2.7038
Equal variances not assumed			1.273	16.029	0.221	1.0222	0.8031	−0.68	2.7244

从输出结果 6—2 可以看出：

对第一选修课程作 t 检验：首先，F 值的相伴概率值 0.766 大于显著性水平 0.05，不能拒绝方差相等的零假设，可以认为低分组和高分组的方差无显著差异；然后，看第一行的 T 检验结果。T 统计量的相伴概率值 0.744 大于显著性水平 0.05，不能拒绝 T 检验的零假设，可以认为低分组和高分组的学生在第一选修课程上意见倾向是一致的。另外，从两样本均值差的 95% 置信区间看，区间跨 0，这也是上述结论的一个佐证。

对第三选修课程作 t 检验：首先，F 值的相伴概率值 0.701 大于显著性水平 0.05，不能拒绝方差相等的零假设，可以认为低分组和高分组的方差无显著差异；然后，看第一行的 T 检验结果。T 统计量的相伴概率值 0.217 大于显著性水平 0.05，不能拒绝 T 检验的零假设，可以认为低分组和高分组的学生在第三选修课程上意见倾向是一致的。

由此可见，无论学生倾向选修哪门课程，他们的选择与问卷中提及的其他因素（如：以往的计算机课程成绩、对目前开设的计算机课程是否感到满足、喜欢的授课方式等）没有直接的关系。

另外，在分析结果中，SPSS 还自动输出了两样本均值差值的估计标准误差 (SE of Diff)。在方差相等和不相等时，这个均值估计标准误差的测度方法有所不同。

在方差相等的情况下，估计标准误差的计算方法是：

$$SE = S_p\sqrt{\frac{1}{n_1}+\frac{1}{n_2}} \tag{6.6}$$

在方差不相等的情况下，估计标准误差的计算方法是：

$$SE = \sqrt{\frac{S_1^2}{n_1}+\frac{S_2^2}{n_2}} \tag{6.7}$$

§6.4 两配对样本 t 检验

一、什么是两配对样本 t 检验

两配对样本 t 检验是根据样本数据对样本来自的两配对总体的均值是否有显著差异进行推断。

两配对样本 t 检验的前提要求，一是，两样本应是配对的。配对可以从两个因素考虑，首先，两样本的观察值数目相同；其次，两样本的观察值的顺序不能随意更改。二是，样本来自的两个总体应服从正态分布。

例如：为检验某减肥茶的减肥效果，收集到同一批肥胖人群在服用该减肥茶前后的体重样本数据，检验这批肥胖者在喝减肥茶前后的平均体重是否发生了显著差异。如果体重确实有了显著差异，则说明减肥茶对体重变化产生了明显作用。

这个问题就是一个两配对样本 t 检验问题。首先，两样本是配对的。表现在：收集到的数据都是针对同一批肥胖人群的，喝茶前后两样本中的观察值个数

一定是相同的。每对观察值数据都惟一地对应一个肥胖者，不能随意改变观察值的先后次序。其次，体重数据可以认为是来自正态总体的。

再例如：分析不同的促销形式对商品销售额是否产生显著影响。分别收集某几种商品在不同促销形式下的销售额样本数据，检验不同促销形式下的平均销售额是否有显著差异，从而判定不同促销形式的促销效果。这也是一个配对样本的 t 检验问题，虽然它涉及的样本不只两个，但却仍然是个配对问题。

又再例如：对案例 1 数据，分析 1999 年全国普通高等院校各个不同学科人力投入中，各职称的平均比例结构是否存在显著差异，即分析各学科人力投入中，高级职称的平均比例是否与中级职称的平均比例相同，或是否与初级职称的平均比例相同。这仍然是一个配对样本的 t 检验问题，表现在：首先，一组职称结构的比例数据是惟一对应某一学科的，样本中的观察值次序是不能随意更改的，且各样本中的观察值个数相同；其次，职称结构比例数据可以认为来自正态总体。

二、两配对样本 t 检验的基本实现思路

两配对样本 t 检验的零假设为 H_0，两总体均值之间不存在显著差异。

它首先求出每对观察值（测量值）的差值，得到差值序列；然后对差值求均值；最后检验差值序列的均值，即平均差是否与零有显著差异。如果平均差与零有显著差异，则认为两总体均值存在显著差异；否则，如果平均差与零无显著差异，则认为两总体均值不存在显著差异。它仍然是计算 t 统计量，计算公式与单样本 t 检验中的完全相同。

SPSS 将自动计算 t 值，由于该统计量服从 $n-1$ 个自由度的 t 分布，SPSS 根据 t 分布表给出 t 值对应的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则拒绝 H_0，认为两总体均值存在显著差异。相反，相伴概率值大于用户心中的显著性水平 α，则不应拒绝 H_0，可以认为两总体均值无显著差异。

从配对样本 t 检验的实现思路不难看出，配对样本 t 检验问题是通过转化成单样本 t 检验来实现的，即：检验两配对样本的差值样本的均值是否与零有显著差异。这种解决方案必然要求样本配对，观察值数目相同且次序不可随意更改。

总之，无论单样本 t 检验，还是两独立样本的 t 检验，或是两配对样本的 t 检验，在方法思路上有许多共同之处。在构造 t 统计量时，分子都是均值差，分母都是抽样分布的标准差，只是独立样本 t 检验的抽样分布标准差与配对样本 t 检验的标准差不同。配对样本的 t 检验能够对观察值自身的其他影响因素加以控

制，比独立样本 t 检验更进了一步。

三、SPSS 配对样本 t 检验的基本操作步骤

进行配对样本 t 检验，数据的准备工作比较简单直接，只需将两个配对样本数据分别依次存放在两个 SPSS 变量中即可。

SPSS 两配对样本 t 检验的基本操作步骤是：

(1) 菜单选项：Statistics－＞Compare means－＞Paired－samples t。于是出现如图 6—3 所示的窗口。

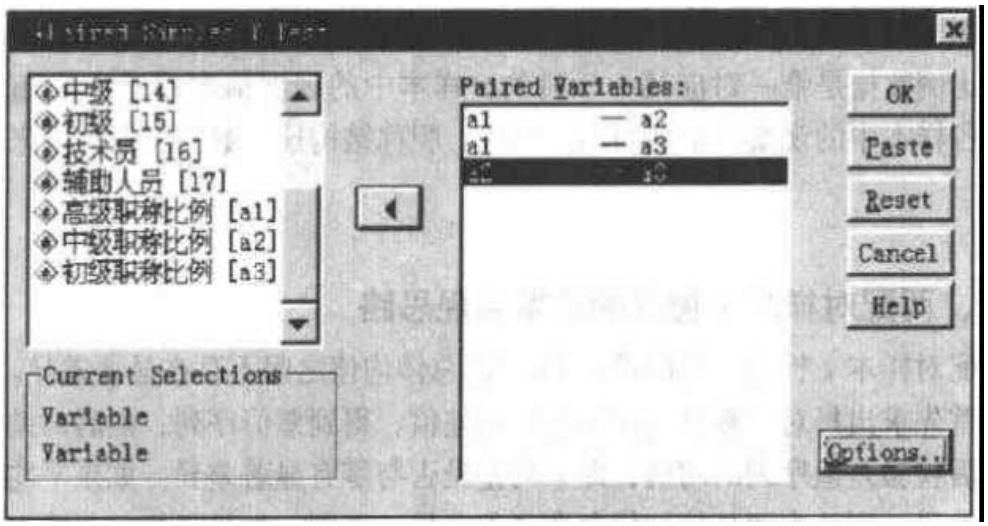


图 6—3　SPSS 两配对样本 t 检验窗口

(2) 选择一对或若干对配对变量作为检测变量到 Paired Variables 框。

SPSS 两配对样本 t 检验的 Option 选项含义与单样本 t 检验中的相同。

四、案例分析说明

对第 5 章案例 1 数据，分析各个学科人力投入中，职称的平均比例结构是否存在显著差异的分析步骤是：

(1) 分别计算高级职称、中级职称和初级职称的比例。可以利用 SPSS 的计算产生变量（Compute）功能实现，并将计算结果分别保存在 3 个 SPSS 变量中。

(2) 检验高级职称的平均比例、中级职称的平均比例、初级职称的平均比例之间是否存在显著差异。

利用 SPSS 两配对样本的 t 检验功能实现。检验结果如输出结果 6—3 所示。

输出结果 6—3（a）　　案例 1 两配对样本 t 检验分析结果

Paired Samples Statistics

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	高级职称比例	.3893	22	.1250	2.665E-02
	中级职称比例	.4031	22	6.304E-02	1.344E-02
Pair 2	高级职称比例	.3893	22	.1250	2.665E-02
	中级职称比例	.2076	22	7.510E-02	1.601E-02
Pair 3	中级职称比例	.4031	22	6.304E-02	1.344E-02
	初级职称比例	.2076	22	7.510E-02	1.601E-02

输出结果 6—3（b）

Paired Samples Correlations

		N	Correlation	Sig.
Pair 1	高级职称比例 & 中级职称比例	22	−.886	.000
Pair 2	高级职称比例 & 初级职称比例	22	−.921	.000
Pair 3	中级职称比例 & 初级职称比例	22	.635	.002

输出结果 6—3（c）

Paired Samples Test

		Paired Differences					t	df	Sig. (2-tailed)
		Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
					Lower	Upper			
Pair 1	高级职称比例 − 中级职称比例	−1.3747E-02	.1832	3.905E-02	−9.4964E-02	6.747E-02	−0.352	21	728
Pair 2	高级职称比例 − 初级职称比例	.1818	.1964	4.186E-02	9.471E-02	.2688	4.342	21	.000
Pair 3	中级职称比例 − 初级职称比例	.1955	6.004E-02	1.280E-02	.1689	.2221	15.275	21	.000

由输出结果 6—3 所列的分析结果可以看出：

职称结构中，中级职称的平均比例最高，其次是高级职称，最低是初级职称。中级职称的平均比例与初级职称的平均比例相差相对最大。这与我国高校人力资源的实际情况是相吻合的。从它们之间的相关系数来看，负相关性较强。两

两配对样本的差值序列的平均值分别为－0.013 8、0.018、0.195 5，计算出的 t 统计量的值分别为－0.352、4.342、15.275，同样表明中级职称的平均比例与初级职称的平均比例相距最大。从各个相伴概率值可得到以下结论：在各学科人力投入中，除高级职称的平均比例与中级职称的平均比例没有显著差异外，其他职称的平均比例之间都存在明显差距，这主要是由于初级职称的平均比例较低造成的。

第 7 章

SPSS 的非参数检验

非参数检验是统计推断的一个重要组成部分，它与参数检验共同构成统计推断的基本内容。参数检验是在假定知道总体分布形式的情况下，对总体分布的某些参数，如：均值、方差等进行推断检验。但是，在现实生活中，由于种种原因，人们往往无法对总体的分布形态作简单的假定，但又希望能从样本数据中获得尽可能多的所需要的信息。非参数检验正是基于这种考虑，在总体分布未知或知道甚少的情况下，利用样本数据对总体的分布形态或分布参数进行推断。

§7.1 单样本非参数检验

一、总体分布的卡方（Chi-square）检验

在得到一批样本数据以后，人们往往希望从中了解样本来自的总体的分布形态是否与已知的分布相吻合。可以通过绘制样本数据直方图的方法来作比较粗略的判断。如果希望进行比较准确的判断，就应该使用非参数检验方法。其中：总体分布的卡方检验就是一种较好的检验方法。

例如：医学家研究心脏病人猝死人数与日期的关系时发现，一周之中，星期

一心脏病人猝死者较多，其他日期则基本相当，比例近似为：2.8:1:1:1:1:1:1。现在需要收集样本数据，推断其总体分布是否与假定的分布相吻合。

1．总体分布的卡方检验的基本思路

总体分布的卡方检验就是根据样本数据推断总体的分布与期望分布或某一理论分布是否有显著差异。它的零假设是 H_0：样本来自的总体其分布形态与期望分布或某一理论分布无显著差异。

总体分布的卡方检验的基本思想是依据这样的理论，即如果从一个随机变量 X 中随机抽取若干个观察样本，这些观察样本落在 X 的 k 个互不相交的子集中的观察频数服从一个多项分布，这个多项分布当 k 趋于无穷时，就近似服从 X 的分布。因此，如果我们假设样本来自的总体服从某个理论分布，那么就需要了解样本数据各子集中的实际观察频数，并依据式（7.1）计算统计量，作出推断。

$$Q = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} \tag{7.1}$$

式中：O_i 表示观察频数，E_i 表示希望频数或理论频数。可见，Q 值越大，说明观察频数与期望频数越不接近；Q 值越小，说明观察频数与期望频数越接近。SPSS 将自动计算 Q 统计量，由于 Q 统计量从 $k-1$ 个自由度的 χ^2 分布，SPSS 会依据 χ^2 分布表，给出 Q 统计量所对应的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则应拒绝 H_0，认为样本来自的总体分布形态与期望分布或某一理论分布存在显著差异；如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，可以认为样本来自的总体分布形态与期望分布或某一理论分布无显著差异。

可见，总体分布的卡方检验是一种吻合性检验，比较适合于一个因素的多项分类的数据分析。

2．SPSS 总体分布的卡方检验的基本操作步骤

SPSS 总体分布的卡方检验对数据的要求是：设计一个 SPSS 变量，其中存放的是收集到的实际样本数据，而并非相应的频数数据。

SPSS 总体分布的卡方检验的基本操作步骤是：

（1）菜单选项：Statistics－>Nonparametric Test－>Chi square。于是出现如图 7—1 所示的窗口。

（2）选定待检验的变量到 Test Variable 框中。

（3）在 Expected Range 框中确定待检验样本的取值范围。其中：Get from data 表示所有观察数据都参与检验；Use Specified Range 表示用户自定义一个取

值范围，只有在这个取值范围内的观察数据才参与检验。

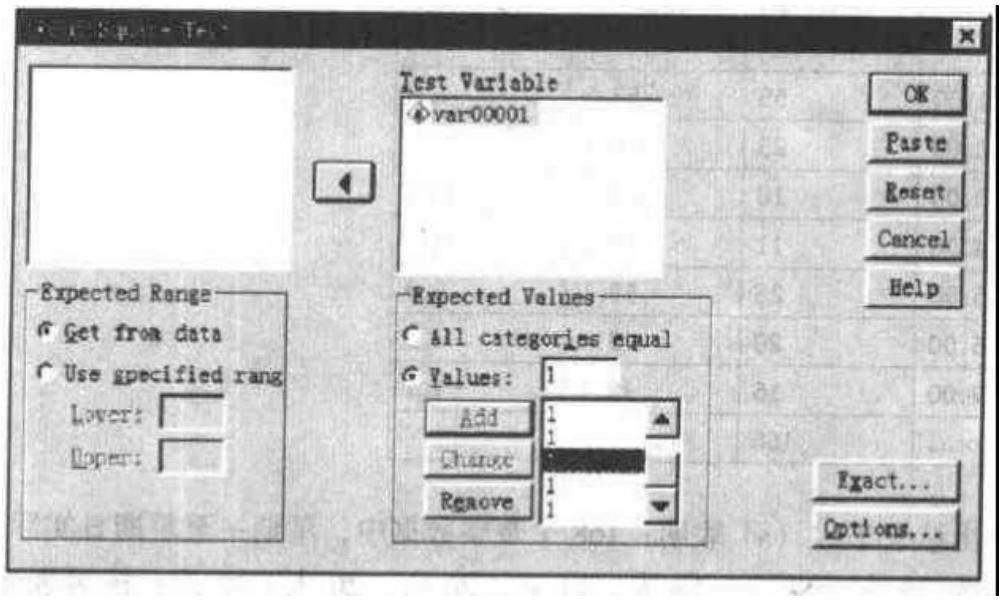


图 7—1　SPSS 总体分布的卡方检验

(4) 在 Expected Values 框中给出期望分布的频数。其中：All categories equal 表示所有分组的期望频数都相同，即期望分布为均匀分布；Value 要求用户输入期望分布的频数值，可以按 Add Change Remove 钮对期望分布的频数作增加、修改和删除。

3. SPSS 总体分布的卡方检验的其他选项

SPSS 总体分布的卡方检验的其他选项有：

(1) Option 选项：Statistics 框用来指定输出附加的分析结果，Missing Values 用来说明对缺失值如何处理。其中，Statistics 框中的 Descriptive 表示：输出待检验变量的描述统计量。Quartiles 表示：输出待检验变量的四分位数；Missing Values 框中的 Exclude cases test by test 表示：只有当检验涉及到含有缺失值的变量时，才暂时剔除在该变量上取缺失值的观察数据；Exclude cases listwise 表示：将所有含有缺失值的观察数据剔除后再作检验。

(2) Exact 选项：表示作精确检验。

4. 举例

为研究上面提到的心脏病人猝死人数与日期的关系，收集到了 168 个观察数据。其中：用 1、2、3、4、5、6、7 表示是星期几死的。现在利用这批样本数据，推断心脏病人猝死人数与日期的关系是否为：2.8∶1∶1∶1∶1∶1∶1。

用总体分布的卡方检验方法，检验结果如输出结果 7—1 所示。

输出结果 7—1（a）　　总体分布的卡方检验举例分析结果（一）

	Observed N	Expected N	Residual
1.00	55	53.5	1.5
2.00	23	19.1	3.9
3.00	18	19.1	-1.1
4.00	11	19.1	-8.1
5.00	26	19.1	6.9
6.00	20	19.1	.9
7.00	15	19.1	-4.1
Total	168		

输出结果 7—1（a）表明：168 个观察数据中，星期一至星期日实际死亡人数分别为 55 人、23 人、18 人、11 人、26 人、20 人、15 人；按照理论分布，168 人在一周各天死亡的期望频数应为 53.5、19.1、19.1、19.1、19.1、19.1、19.1；实际观察频数与期望频数的差分别为 1.5、3.9、-1.1、-8.1、6.9、0.9、-4.1。

输出结果 7—1（b）　　总体分布的卡方检验举例分析结果（二）

Test Statistics

	VAR00001
Chi-Square	7.757
df	6
Asymp. Sig.	.256

输出结果 7—1（b）是计算的卡方统计量，以及对应的相伴概率值。由于相伴概率值大于显著性水平 0.05，因此不能拒绝零假设，可以认为样本来自的总体分布与指定的理论分布无显著差异，即：心脏病人猝死人数与日期的关系基本是2.8:1:1:1:1:1:1。

二、SPSS 的二项分布检验

1. 二项分布

在现实生活中有很多数据的取值是两类的，例如：人群可以分成男性和女性，产品可以分成合格和不合格，学生可以分成三好学生和非三好学生，投掷硬币实验的结果可以分成出现正面和出现反面等。这时，如果某一类情况出现的概率是 P，则另一类情况出现的概率就是 $1-P$. 这种分布称为二项分布。

2．SPSS 二项分布检验的基本思路

SPSS 二项分布检验就是根据收集到的样本数据，推断总体的分布是否与指定的某个二项分布存在显著差异。其零假设是 H_0：样本来自的总体与某个指定的二项分布无显著差异。

SPSS 中的二项分布检验，在样本数小于等于 30 时，按照计算二项分布概率的公式进行计算；在样本数大于 30 时，计算的是 Z 统计量，认为在零假设下，Z 统计量近似服从正态分布。Z 统计量的计算公式为：

$$Z=\frac{K\pm0.5-nP}{\sqrt{nP\ (1-P)}} \tag{7.2}$$

式中：K 是观察值取值为指出检验值的样本个数，P 为检验概率，n 为样本总数。当 K 小于 $n/2$ 时，取加号，K 大于 $n/2$ 时，取减号。

SPSS 自动计算 Z 统计量并给出对应的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则拒绝 H_0，即认为样本来自的总体与某个指定的二项分布有显著差异；如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，即可以认为样本来自的总体与某个指定的二项分布无显著差异。

3．SPSS 二项分布检验的基本操作步骤

SPSS 进行二项分布检验对数据的要求是：设计一个 SPSS 变量，其中存放的是收集到的实际样本数据，而并非相应的频数数据。

SPSS 二项分布检验的基本操作步骤是：

(1) 菜单选项：Statistics－＞Nonparametric Test－＞Binomial。于是出现如图 7—2 所示的窗口。

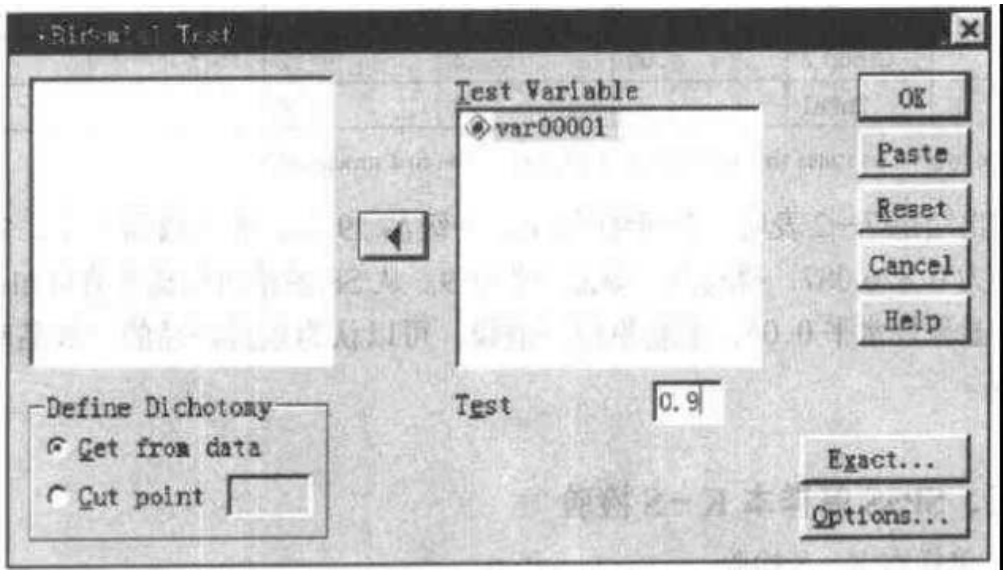


图 7—2　SPSS 二项分布检验窗口

(2) 选定待检验的变量到 Test Variable 框中。

(3) 在 Expected Range 框中确定待检验样本的取值范围。其中：Get from data 适用于检验变量的所有数据都是二值数据的情况；Cut point 表示：如果检验变量的数据不是二值数据，是可以输入一个具体的值，于是小于等于该值的观察值为一类，大于该值的为另一类。

(4) 在 Test 框中输入二项分布的检验概率值。SPSS 将检验第一个观察值所在组的比例是否与输入的检验概率值存在显著差异。在具体操作使用时，应根据观察值数据的具体情况输入相应的检验概率值。

SPSS 二项分布检验的其他选项同前，这里不再重复。

4. 举例

为验证某批产品的一级品率是否达到 90%，现从该批产品中随机抽取 23 个样品进行检测，得到最后的检测结果数据。其中：用 1 表示一级品，0 表示非一级品。

利用二项分布检验方法检验。由于样本数据中第一个观察值为 1（一级品），因此，在检验框中应输入 0.9（检验一级品率是否为 90%）。检验结果见输出结果 7—2 所示。

输出结果 7—2　　二项分布检验举例分析结果

Binomial Test

		Category	N	Observed Prop.	Test Prop.	Exact Sig. (1-tailed)
VAR00001	Group 1	1.00	19	.826087	.9	.193
	Group 2	.00	4	.2		
	Total		23	1.0		

Alternative hypothesis states that the proportion of cases in the first group<0.9

输出结果 7—2 表明：23 个样品中，一级品 19 个，非一级品 4 个，实际一级品率为 0.826 087，检验的一级品率为 0.9。从 SPSS 给出的概率值可知：概率值大于显著性水平 0.05，不能拒绝零假设，可以认为该批产品的一级品率达到了 90%。

三、SPSS 单样本 K－S 检验

1. 单样本 K－S 检验

K－S 检验是以两位苏联数学家柯尔莫哥(Kolmogorov)和斯米诺夫(Smirnov)

命名的，也是一种拟和优度的非参数检验方法。单样本 K－S 检验是利用样本数据推断总体是否服从某一理论分布，适用于探索连续型随机变量的分布形态。

例如：收集一批周岁儿童身高的样本数据，然后利用样本数据推断周岁儿童的身高是否服从正态分布。

2．SPSS 单样本 K－S 检验的基本思路

SPSS 单样本 K－S 检验中的理论分布可以为正态分布、均匀分布、指数分布、泊松分布。其零假设是 H_0：样本来自的总体与指定的理论分布无显著差异。

单样本 K－S 检验的基本思路是：

根据样本数据和用户的指定构造出理论分布，查分布表得到的相应的理论累计概率分布函数 $F_0\ (x)$；利用样本数据计算各样本数据点的累计概率，得到经验累计概率分布函数 $S_0\ (x)$；计算 $S_0\ (x)$ 与 $F_0\ (x)$ 在相同变量值点 x 上的差 $D\ (x)$，得到差值序列 D。单样本 K－S 检验主要对差值 D 序列进行研究。

SPSS 将自动计算 K－S 的 Z 统计量，依据 K－S 分布表（小样本）或正态分布表（大样本）给出对应的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则应拒绝 H_0，认为样本来自的总体与指定的分布有显著差异；如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，可以认为样本来自的总体与指定的分布无显著差异。

例如：收集到 21 个关于周岁儿童身高的样本数据，利用 K－S 方法检验周岁儿童的身高是否服从正态分布。根据上述方法，计算过程如表 7—1 所示。

表 7—1　　单样本 K－S 检验计算方法

序列号	样本数据	Z 化数据	$F_0\ (x)$	$S_0\ (x)$	$D\ (x)$
1	64.00	－1.974 89	0.024 1	0.047 6 (1/21)	0.023 5
2	68.00	－0.969 49	0.166 1	0.190 5 (4/21)	0.024 4
3	68.00	－0.969 49	0.166 1	0.190 5 (4/21)	0.024 4
4	68.00	0.969 49	0.166 1	0.190 5 (4/21)	0.024 4
5	69.00	－0.718 14	0.236 3	0.238 1 (5/21)	0.001 8
6	70.00	－0.466 79	0.320 3	0.381 0 (8/21)	0.060 7
7	70.00	－0.466 79	0.320 3	0.381 0 (8/21)	0.060 7
8	70.00	－0.466 79	0.320 3	0.381 0 (8/21)	0.060 7
9	71.00	－0.215 44	0.414 7	0.619 0 (13/21)	0.204 3
10	71.00	－0.215 44	0.414 7	0.619 0 (13/21)	0.204 3
11	71.00	－0.215 44	0.414 7	0.619 0 (13/21)	0.204 3
12	71.00	－0.215 44	0.414 7	0.619 0 (13/21)	0.204 3
13	71.00	－0.215 44	0.414 7	0.619 0 (13/21)	0.204 3

续前表

序列号	样本数据	Z 化数据	$F_0(x)$	$S_0(x)$	$D(x)$
14	72.00	0.035 91	0.514 3	0.666 7 (14/21)	0.152 4
15	73.00	0.287 26	0.613 0	0.714 3 (15/21)	0.101 3
16	74.00	0.538 61	0.704 9	0.761 9 (16/21)	0.057 0
17	75.00	0.789 96	0.785 2	0.809 5 (17/21)	0.024 3
18	76.00	1.041 31	0.851 1	0.857 1 (18/21)	0.006 0
19	78.00	1.544 01	0.938 7	0.904 8 (19/21)	−0.033 9
20	79.00	1.795 36	0.963 7	0.952 4 (20/21)	−0.011 3
21	80.00	2.046 71	0.979 7	1.000 (21/21)	0.020 3

SPSS 正是针对 $D(x)$ 样本计算 K－S 的 Z 统计量，并给出相应的相伴概率值。

3. SPSS 单样本 K－S 检验的基本操作步骤

将实际样本数据以变量的形式存到 SPSS 变量以后，便可按以下步骤作 SPSS 单样本 K－S 检验。

(1) 菜单选项：Statistics－>Nonparametric Tests－>1－Sample K－S。于是出现如图 7—3 所示的窗口。

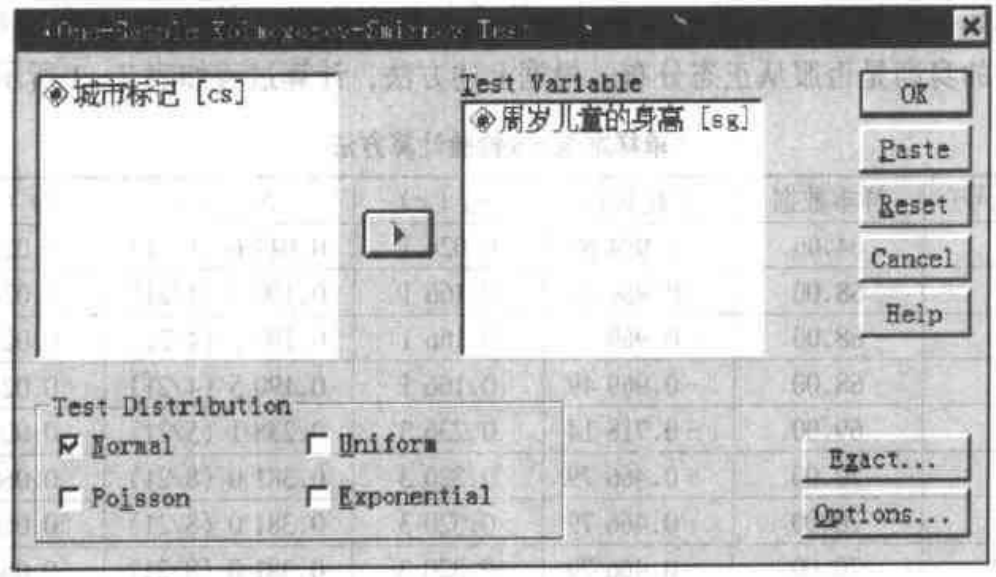


图 7—3 SPSS 单样本 K－S 检验窗口

(2) 选择待检验的变量到 Test variable list 框中。

(3) 在 Test Distribution 框中选择理论分布，其中：Normal 为正态分布、Uniform为均匀分布、Possion 为泊松分布、Exponential 为指数分布。

SPSS 单样本 K－S 检验的其他选项同前，这里不再重复。

4. 举例

针对表 7—1 中周岁儿童的身高数据，利用 K－S 检验，推断周岁儿童的身高是否服从正态分布。分析结果见输出结果 7—3。

输出结果 7—3　　单样本 K－S 检验举例分析结果

One-Sample Kolmogorov-Smirnov Test

	周岁儿童身高
N	21
Normal Parameters Mean	71.8571
Std. Deviation	3.9785
Most Extreme Differences Absolute	.204
Positive	.204
Negative	-.011
Kolmogorov-Smirnov Z	.936
Asymp. Sig. (2-tailed)	.344

输出结果 7—3 表明：样本数据的均值为 71.8574，标准差为 3.9785。$D(x)$ 的最大绝对值为 0.204，$D(x)$的最大正值为 0.204，最小负值为－0.011。根据 $D(x)$计算出的 K－S 的 Z 统计量为 0.936，对应的相伴概率值为 0.344。由于相伴概率值大于显著性水平 0.05，因此，不能拒绝零假设，可以认为周岁儿童的身高服从正态分布。

四、SPSS 单样本变量值随机性检验

1. 单样本变量值随机性检验

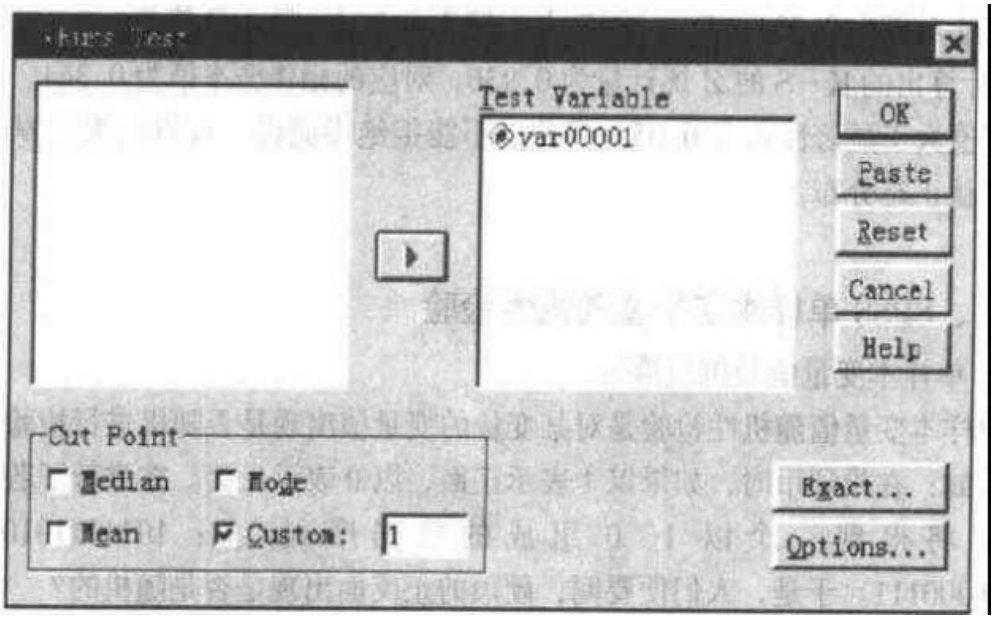
单样本变量值随机性检验是对某变量的变量值出现是否随机进行检验。

例如：在投硬币时，如果以 1 表示正面，以 0 表示反面。在进行了若干次投掷后，将得到一个以 1、0 组成的数据序列，如：1011011010011000101010000111。于是，人们便要问，硬币的正反面出现是否是随机的？

要研究这个问题，就需要对上面的 0、1 序列进行分析，通过验证 0、1 出现是否随机来判定正反面的出现是否随机。

单样本变量值随机性检验正是用来解决这类问题的。

2. 单样本变量值随机性检验的基本思想



(3) 在 Cut point 框中确定计算游程数的分界点。其中：Median 表示以样本中位数为分界点；Mode 表示以样本众数为分界点；Mean 表示以样本均值为分界点；Custom 表示以用户输入的值为分界点。SPSS 将小于该分界点的所有变量值作为一组，将大于或等于该分界点的所有变量值为另一组，然后计算游程数。

SPSS 单样本变量值随机性检验的其他选项同前，这里不再重复。

4. 举例

对上面提到的投掷硬币问题作变量值随机性检验。指定以 1 作为计算游程数的分界点，结果如输出结果 7—4 所示。

输出结果 7—4　　单样本变量值随机性检验举例分析结果

Runs Test

	硬币
Test Value	1
Total Cases	28
Number of Runs	17
Z	.578
Asymp. Sig. (2-tailed)	.563

输出结果 7—4 表明：一共投掷了 28 次，以 1 为分界点计算出的游程数为 17，Z 统计量为 0.578，对应的相伴概率值为 0.563。由于相伴概率值大于显著性水平 0.05，因此，不能拒绝零假设，可以认为硬币的正反面出现是随机的。

五、单样本非参数检验方法的比较

在上面的四种单样本非参数检验中，前三种方法均属于拟和性检验的范畴。因此，当检验样本来自的总体是否服从某种理论分布时，可以选用前三种方法中的一个。但在实际使用中，应注意这些检验方法对数据的要求。

1. 卡方检验一般要求待检验样本有较大的样本容量，比较适合于一个因素的多项分类的数据分析。

2. 二项分布检验只能作二项分布检验。

3. 单样本 K－S 检验比较适合于连续型数据的分析，其检验功效是比较强的。

§7.2 两独立样本非参数检验

两独立样本的非参数检验是在对总体分布不甚了解的情况下，通过分析样本数据，推断样本来自的两个独立总体的分布是否存在显著差异，一般来推断两个独立总体的均值或中位数是否存在显著差异。关于样本是否为独立的，主要看在一个总体中抽取样本对在另一个总体中抽取样本有无影响。如果没有影响，则可以认为这两个总体是独立的。

例如：某工厂用两种不同的工艺生产同一种产品，如果需要检验它们的使用寿命是否存在显著差异，则可以随机抽取样本，测得它们的使用寿命数据，并对其进行两独立样本的非参数检验，从而判定两种工艺生产出的产品，使用寿命分布有无显著差异。

具体数据如下（单位：百小时）：

甲种工艺：675，682，692，679，669，661，693

乙种工艺：662，649，672，663，650，651，646，652

用 1 表示甲种工艺，2 表示乙种工艺。

一、SPSS 两独立样本非参数检验的基本思路

SPSS 提供了四种两独立样本的非参数检验方法。下面分别对它们作简单介绍。

1. 两独立样本的曼－惠特尼 U 检验（Mann-Whitney U）

两独立样本的曼－惠特尼 U 检验的零假设是 H_0：样本来自的两独立总体的均值无显著差异。

曼－惠特尼 U 检验主要是通过对平均秩的研究来实现推断的。秩简单说就是名次。我们可以将数据按照升序进行排序，这时每一个具体数据都会有一个在整个数据中的位置或名次，这就是该数据的秩，数据有多少个，秩便有多少个。

曼－惠特尼 U 检验的基本思路是：首先，将两组样本数据（X_1，X_2，…X_m）和（Y_1，Y_2，…Y_n）混合并按升序排序（m 和 n 分别为两组样本的样本容量），求出每个数据各自的秩 R_i；然后，分别对（X_1，X_2，…X_m）和（Y_1，Y_2，…Y_n）的秩求平均，得到两个平均秩 W_X/m 和 W_Y/n。可以直观理解，如果这两个平均秩 W_X/m 和 W_Y/n 相差甚远，则可以说明一组样本的值普遍偏

小，另一组样本的值普遍偏大。这时，我们就有理由怀疑零假设的成立。

不仅如此，曼－惠特尼 U 检验还计算了（X_1，X_2，…X_m）每个秩优于（Y_1，Y_2，…Y_n）每个秩的个数 U_1，以及（Y_1，Y_2，…Y_n）每个秩优先于（X_1，X_2，…X_m）每个秩的个数 U_2，并对 U_1 和 U_2 进行比较。同样可以直观理解，如果 U_1 和 U_2 相距较远，则有理由怀疑零假设的成立。

为作出精确判断，SPSS将自动计算 Wilcoxon W 统计量和 Mann-Whitcny U 统计量，其中：Wilcoxon W 为：如果 $m<n$，则 Wilcoxon $W=W_Y$；如果 $m>n$，则 Wilcoxon $W=W_X$；如果 $m=n$，则 Wilcoxon W 为第一个观察值所属样本组的 W 值。

Mann－Whiteny U 统计量按式（7.3）计算：

$$U=W-\frac{1}{2}n\ (n+1) \tag{7.3}$$

式中：W 值即为 Wilcoxon W，n 取值为 W 对应组的样本容量。SPSS 计算出 U 值以后依据 Mann－Whitney 分布表给出对应的相伴概率值（Excact Sig）。同时，SPSS 还计算近似服从正态分布的 Z 统计量，计算公式是：

$$Z=\frac{U-\frac{1}{2}mn}{\sqrt{\frac{1}{12}mn\ (m+n+1)}} \tag{7.4}$$

同样 SPSS 也会给出 Z 值对应的相伴概率值（Asymp.Sig）。

因此，在样本个数小于 30 时，应以 U 统计量的相伴概率作为判断标准，在样本个数大于 30 时，属于大样本情况下，应以 Z 统计量的相伴概率作为判断标准。

如果相伴概率值小于或等于用户心中的显著性水平 α，则拒绝 H_0，认为样本来自的两独立总体的均值存在显著差异；如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，可以认为样本来自的两独立总体的均值无显著差异。

例如：对上面提到的关于产品寿命比较的问题，可以用两独立样本的曼－惠特尼 U 检验方法作检验。具体计算过程如表 7—2 所示。

表 7—2　　两独立样本的曼－惠特尼 U 检验计算方法

混合排序后的样本数据	组标记	秩	混合排序后的样本数据	组标记	秩
646.00	2.00	1	663.00	2.00	8
649.00	2.00	2	669.00	1.00	9

续前表

混合排序后的样本数据	组标记	秩	混合排序后的样本数据	各组标记	秩
650.00	2.00	3	672.00	2.00	10
651.00	2.00	4	675.00	1.00	11
652.00	2.00	5	679.00	1.00	12
661.00	1.00	6	682.00	1.00	13
662.00	2.00	7	692.00	1.00	14
			693.00	1.00	15

从表 7—2 可得：第一组样本的样本数为 7，秩总和为 80，平均秩为 11.43；第二组样本的样本数为 8，秩总和为 40，平均秩为 5。W 统计量应为第二组样本的秩总合，依照公式可计算得到 U、Z 统计量分别为 4、 -2.777。

2. 两独立样本的 K－S 检验

K－S 检验不仅能够检验单样本是否服从某一理论分布，还能够对两独立样本的总体分布情况进行比较。其零假设是 H_0：样本来自的两个独立总体的分布无显著差异。

两独立样本的 K－S 检验的基本思想与前面讨论的单样本 K－S 检验的基本思路是一致的，差别在于这里处理的是观察值的秩，而非观察值本身。首先，将两组样本混合并按升序排序；然后，分别计算两组样本秩的累计频数和每个点上的累计频率；最后，将两个累计频率相减，得到差值序列数据。

两独立样本的 K－S 检验仍然关注差值序列。SPSS 将自动计算 K－S Z 统计量，并依据正态分布表给出对应的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则拒绝 H_0，认为样本来自的两独立总体的分布有显著差异；如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，可以认为样本来自的两独立总体分布无显著差异。

例如：对上面提到的关于产品寿命比较的问题，可以用两独立样本的 K－S 检验方法作检验。具体计算过程如表 7—3 所示。

表 7—3　　两独立样本的 K－S 检验计算方法

混合排序后的样本数据	组标记	甲种工艺累计频数	乙种工艺累计频数	甲种工艺累计频率	乙种工艺累计频率	两种工艺累计频率差
646	2.00	0	1	0	0.125	0.125
649	2.00	0	2	0	0.25	0.25
650	2.00	0	3	0	0.375	0.375

续前表

混合排序后的样本数据	组标记	甲种工艺累计频数	乙种工艺累计频数	甲种工艺累计频率	乙种工艺累计频率	两种工艺累计频率差
651	2.00	0	4	0	0.5	0.5
652	2.00	0	5	0	0.625	0.625
661	1.00	1	5	0.143	0.625	0.482
662	2.00	1	6	0.143	0.75	0.607
663	2.00	1	7	0.143	0.875	0.732
669	1.00	2	7	0.286	0.875	0.589
672	2.00	2	8	0.286	1	0.714
675	1.00	3	8	0.429	1	0.571
679	1.00	4	8	0.571	1	0.429
682	1.00	5	8	0.714	1	0.286
692	1.00	6	8	0.857	1	0.143
693	1.00	7	8	1	1	0

从表 7—3 可知，最大的绝对值差 D 为 0.732。

需要说明的是：两独立样本的 K－S 检验适用在大样本情况下。

3. 两独立样本的游程检验（Wald－Wolfwitz Runs）

单样本的游程检验是用来检验变量值的出现是否随机的，两独立变量的游程检验则用来检验样本来自的两独立总体的分布是否存在显著差异。其零假设是 H_0：样本来自的两个总体的分布无显著差异。

两独立样本的游程检验与单样本游程检验的基本思想是相同的，不同点在于如何得到游程数据，这是关键性的步骤。两独立样本的游程检验中，计算游程的方法与观察值的秩有关。首先，将两组样本混合并按升序排序。在数据排序的同时，两组样本的每个观察值对应的样本组标志值序列也随之重新排列；然后，对这个标志值序列按前面讨论的方法求游程。

可以直观理解，如果样本来自的两总体的分布形态存在较大差距，那么计算出的游程数会相对比较小。如果游程数比较大，则应该是由于两样本数据充分混合的结果，那么它们的分布应该不存在显著差异。

为作出精确判断，与单样本游程检验相同，SPSS 也将利用游程数构造计算 Z 统计量，并依据正态分布表给出对应的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则拒绝 H_0，认为样本来自的两独立总体的分布存在显著差异；如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，

可以认为样本来自的两独立总体的分布无显著差异。

例如：对上面提到的关于产品寿命比较的问题，可以用两独立样本的游程检验方法作检验。具体计算过程如表 7—4 所示。

表 7—4　　两独立样本的游程检验计算方法

混合排序后的样本数据	排序后的样本标志值序列	混合排序后的样本数据	排序后的样本标志值序列
646	2	669	1
649	2	672	2
650	2	675	1
651	2	679	1
652	2	682	1
661	1	692	1
662	2	693	1
663	2		

从表 7—4 可知，游程数为 6 个。

需要说明的是：如果两组样本中有相同的观察值，两组样本的前后次序将会影响到游程数的计算。此时，SPSS 会给出相应的提示信息。

4. 两独立样本的极端反应检验（Moses Extreme Reactions）

两独立样本的极端反应检验，也是一种检验样本来自的两总体分布是否存在显著差异的检验。其零假设是 H_0：样本来自的两个总体的分布无显著差异。

它的基本思想是：将一个样本作为控制样本，另一个样本作为实验样本。以控制样本作为对照，检验实验样本是否存在极端反应。如果实验样本不存在极端反应，则认为两独立总体分布无显著差异。相反，如果实验样本存在极端反应，则认为两独立总体分布存在显著差异。首先，将两组样本混合并按升序排序；然后，找出控制样本最低秩和最高秩之间包含的观察值个数，称为跨度（Span）。为控制极端值对分析结果的影响，也可先去掉控制样本两个最极端的观察值后再求跨度，这个跨度称为截头跨度。

两独立样本的极端反应检验注重对跨度和截头跨度的分析。可以直观理解，如果跨度或截头跨度很小，则应是两样本数据无法充分混合后的结果，可以认为实验样本存在极端反应。

SPSS 自动计算跨度和截头跨度后，会依据分布表给出对应的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则拒绝 H_0，认为样本来

自的两独立总体的分布存在显著差异；如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，可以认为样本来自的两独立总体的分布不存在显著差异。

例如：对上面提到的关于产品寿命比较的问题，可以用两独立样本的极端反应检验方法作检验。具体计算过程如表 7—5 所示。假设以甲种工艺生产作为控制样本，乙种工艺生产作为试验样本。

表 7—5　　两独立样本的极端反应检验计算方法

646	649	650	651	652	661	662	663	669	672	675	679	682	692	693
2	2	2	2	2	1	2	2	1	2	1	1	1	1	1

从表 7—5 可知，跨度为 10，截头跨度为 6。

二、SPSS 两独立样本非参数检验的基本操作步骤

首先应明确 SPSS 对两独立样本非参数检验的数据要求：应设计两个 SPSS 变量。其中，一个变量存放两组样本的数据；另一个变量存放样本组标志数据，即每个观察值所对应的样本组号。

SPSS 两独立样本非参数检验的基本操作步骤是：

(1) 菜单选项：Statistics－>Nonparametric Tests－>2 Independent Sample。于是出现如图 7—5 所示的窗口。

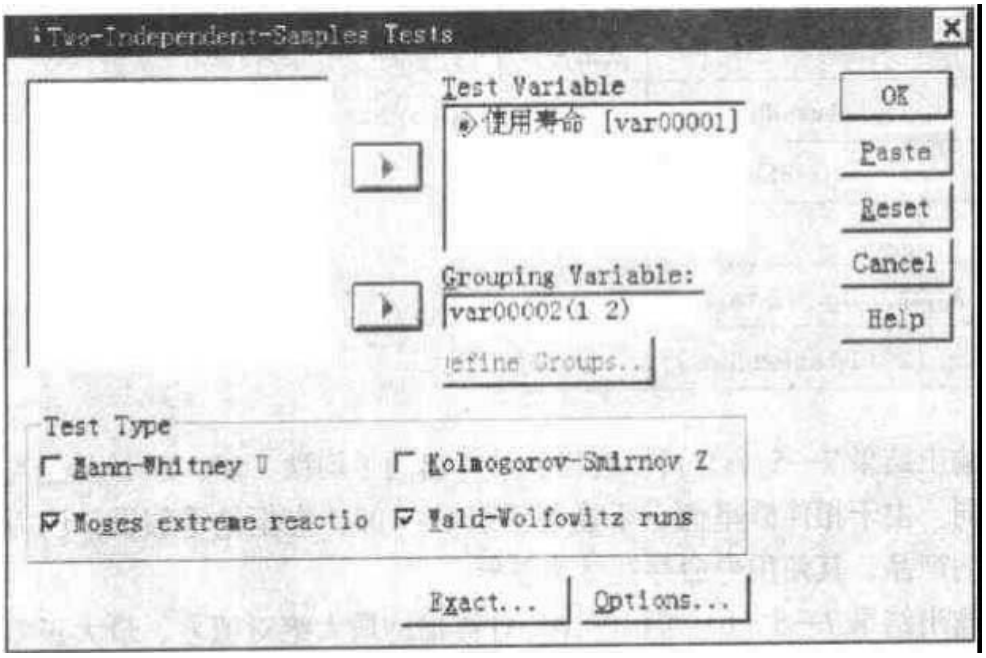


图 7—5　SPSS 两独立样本的非参数检验窗口

(2) 选择待检验的变量到 Test Variable 框中。

(3) 指定存放样本标志值的变量到 Grouping Variable 框，并根据具体情况分别给出样本的标志值。

(4) 选择一种或几种非参数检验方法。

SPSS 两独立样本的非参数检验的其他选项同前，这里不再重复。

三、举例

对上面提到的关于产品寿命比较的问题，用两独立样本的非参数检验方法作检验。检验结果如输出结果 7—5 所示。

输出结果 7—5 (a)　两独立样本的 Mann-Whitney U 检验分析结果

Mann-Whitney Test

Ranks

	样本标志	N	Mean Rank	Sum of Ranks
使用寿命	1.00	7	11.43	80.00
	2.00	8	5.00	40.00
	Total	15		

Test Statistics

	使用寿命
Mann-Whitney U	4.000
Wilcoxon W	40.000
Z	-2.777
Asymp. Sig. (2-tailed)	.005
Exact Sig. [2*(1-tailed Sig.)]	.004

从输出结果 7—5 (a) 可以看出：计算出的平均秩、U、W 统计量与前面讨论的相同。由于相伴概率值小于显著性水平 0.05，应拒绝零假设，认为两种工艺生产的产品，其使用寿命存在显著差异。

从输出结果 7—5 (b) 可以看出：计算出的最大绝对值差、最大正差、最大负差与前面讨论的完全吻合。由于相伴概率值小于显著性水平 0.05，应拒绝零假设，认为两种工艺生产的产品，其使用寿命存在显著差异。

输出结果 7—5（b）　　两独立样本的 K－S 检验分析结果

Tow-Sample Kolmogorov-Smirnov Test

Test Statistics

		使用寿命
Most Extreme Differences	Absolute	.732
	Positive	.732
	Negative	.000
Kolmogorov-Smirnov Z		1.415
Asymp. Sig. (2-tailed)		.037

a Grouping Variable: 样本标志

从输出结果 7—5（c）可以看出：计算出的游程数与前面讨论的一致。但在这个检验中，由于相伴概率值大于显著性水平 0.05，因此不能拒绝零假设，可以认为两种工艺生产的产品，其使用寿命不存在显著差异。

输出结果 7—5（c）　　两独立样本的游程检验分析结果

Wald-Wolfowitz Test

Test Statistics

		Number of Runs	Z	Exact Sig. (1-tailed)
使用寿命	Exact Number of Runs	6	−1.059	.149

从输出结果 7—5（d）可以看出：计算出的跨度和截头跨度与前面讨论的完全吻合。无论是否剔除极端值，相伴概率值均大于显著性水平 0.05，因此，不能拒绝零假设，可以认为两种工艺生产的产品，其使用寿命不存在显著差异。

输出结果 7—5（d）　　两独立样本的极端反应检验分析结果

Moses Test

Test Statistics

		使用寿命
Observed Control Group Span		10
	Sig. (1-tailed)	.084
Trimmed Control Group Span		6
	Sig. (1-tailed)	.100
Outliers Trimmed from each End		1

§7.3 多独立样本非参数检验

多独立样本的非参数检验也是通过分析样本数据，推断样本来自的多个独立总体的分布是否存在显著差异，一般来推断多个独立总体的均值或中位数是否存在显著差异。关于样本是否为独立，同样看在一个总体中抽取样本对在其他总体中抽取样本有无影响。如果没有影响，则可以认为这些总体是独立的。

例如：假设要比较北京、上海、天津、广州四个城市周岁儿童的身高，分析各总体的分布是否存在显著差异。于是在四个城市随机抽取样本，得到以下 20 个数据。

北京：79.00，75.00，78.00，76.00，72.00

上海：72.00，71.00，74.00，74.00，73.00

天津：76.00，78.00，78.00，77.00，75.00

广州：70.00，72.00，71.00，71.00，69.00

用 1 表示北京，2 表示上海，3 表示天津，4 表示广州。

一、SPSS 多独立样本非参数检验的基本思路

SPSS 提供了三种多独立样本的非参数检验方法。下面分别对它们作简单介绍。

1. 多独立样本的中位数检验（Median）

多独立样本的中位数检验通过对多组样本数据的分析推断多个独立总体的分布是否存在显著差异。其零假设是：样本来自的多个独立总体的中位数无显著差异。

它的基本思想是：如果多组独立样本的中位数无显著差异，或者说，如果多组独立样本有共同的中位数，那么，这个中位数应处于每组样本数据的中间位置上。这样，每组样本大于或小于这个共同中位数的数目应大致相同。基本方法是：首先，将多组样本数混合并按升序排序，求出混合样本数据中的中位数，并假设它是一个共同的中位数；然后分别计算每组样本中大于或小于这个共同中位数的样本数，形成表 7—6。

可以直观理解，如果每列中上下单元中的数据大致相同，即：每组样本的大于和小于共同中位数的观察值数目大致相同，则可以认为这多个独立总体的中位数无显著差异。

表 7—6　　多独立样本中位数检验

	1 组样本	2 组样本	……	n 组样本
大于共同中位数的数目				
小于等于共同中位数的数目				

为作出精确判断，SPSS 将依据单元格子中的数据计算 χ^2 统计量，计算公式为：

$$\chi^2 = \sum_{i}^{r}\sum_{j=1}^{k}\frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{7.5}$$

式中：k 和 r 表示表 7—6 中共有 k 列 r 行。O_{ij} 为第 j 列第 i 行单元格中的实际观察频数，E_{ij} 为期望频数。

SPSS 将依据分布表给出 χ^2 对应的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则应拒绝 H_0，认为样本来自的多个独立总体的中位数存在显著差异；如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，可以认为样本来自的多个独立总体的中位数不存在显著差异。

例如：上面四城市周岁儿童身高的问题，利用多独立样本中位数检验作分析。具体的计算过程如表 7—7 所示。

对样本数据混合排序后，共同的中位数为 74。

表 7—7　　多独立样本中位数检验计算方法

	北京	上海	天津	广州	合计
大于 74 的数目	4 (2.25)	0 (2.25)	5 (2.25)	0 (2.25)	9
小于等于 74 的数目	1 (2.75)	5 (2.75)	0 (2.75)	5 (2.75)	11

表 7—7 括号中的数据为期望频数。于是，根据公式可以计算出卡方统计量为 16.768。

应当注意：如果单元格子中的期望频数小于 1 或有 20% 的单元格子中的期望频数小于 5，SPSS 会给出提示。这时，可以考虑采用其他非参数检验方法进行检验。

2. 多独立样本的 K－W 检验

多独立样本的 K－W 检验是 Kruskal-Wallis 检验的缩写，是一种推广的平均秩检验。其零假设是 H_0：样本来自的多个独立总体的分布无显著差异。

多独立样本的 K－W 检验的基本方法是：首先，将多组样本数混合按升序排序，并求出每个观察值的秩；然后，对多组样本的秩分别求平均数。

可以直观理解，如果各组样本的平均秩大致相等，则可以认为多个独立总体的分布无显著差异。如果各样本的平均秩相差甚远，则不能认为多个独立总体的分布无显著差异。

为作出精确判断，将按式（7.6）计算 $K-W$ 统计量：

$$K-W=\frac{12}{N(N+1)}\sum_{i}^{k}n_i(\bar{R}_i-\bar{R})^2 \tag{7.6}$$

式中：k 表示有 k 组样本，每组样本的观察值个数为 n_i 个，$\bar{R}$ 为平均秩。

SPSS 自动计算 $K-W$ 统计量，并依据 Kruskal-Wallis 检验临界值表给出 $K-W$ 统计量的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则应拒绝 H_0，认为样本来自的多个独立总体的分布存在显著差异；如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，可以认为样本来自的多个独立总体的分布不存在显著差异。

例如：上面四城市周岁儿童身高的问题，利用多独立样本的 K－W 检验作分析。具体的计算过程如表 7—8 所示。

表 7—8　　多独立样本的 K－W 检验计算方法

混合排序后的样本数据	样本标志	秩	混合排序后的样本数据	样本标志	秩
69.00	4.00	1	74.00	2.00	10.5
70.00	4.00	2	75.00	1.00	12.5
71.00	2.00	4	75.00	3.00	12.5
71.00	4.00	4	76.00	1.00	14.5
71.00	4.00	4	76.00	3.00	14.5
72.00	1.00	7	77.00	3.00	16
72.00	2.00	7	78.00	1.00	18
72.00	4.00	7	78.00	3.00	18
73.00	2.00	9	78.00	3.00	18
74.00	2.00	10.5	79.00	1.00	20

从表 7—8 可以看出：许多数据的秩是相同的，其原因是由于观察值相同，这种现象称为打结。如：取值为 71 的观察值有 3 个。对于这些打结数据求秩，SPSS 以秩的均值替代。如：取值为 71 的观察值，按照正常的排序顺序，其秩应分别 3、4、5，它们的均值为 4，于是，取值为 71 的各观察值的秩就为 4。基于表 7—8，不难得出北京、上海、天津、广州的平均秩分别为：14.4、8.2、15.8、3.6。

3. 多个独立总体的 Jonkheere－Terpstra 检验

Jonkheer－Terpstr 检验样本来自的多个独立总体的分布是否存在显著差异。其零假设是 H_0：样本来自的多个独立总体的分布无显著差异。

Jonkheer－Terpstr 检验的基本思想方法与两独立样本的曼－惠特尼 U 检验（Mann-Whitney U）比较类似，它也是计算一组样本的观察值小于其他组样本的观察值的个数。如果用 U_{ij} 表示第 i 组样本观察值小于第 j 组样本观察值的个数，那么可以按式（7.7）计算出 $J-T$ 统计量：

$$J-T=\sum_{i<j}U_{ij} \tag{7.7}$$

可见，计算 $J-T$ 统计量时，必然涉及到样本组标志值的大小顺序，这样计算出的 $J-T$ 统计量称为观察的 $J-T$ 统计量（Observed $J-T$）。例如：如果现在有 3 组样本，它们的样本组标志值分别为 1、2、3，那么观察的 $J-T$ 统计量为：1 组样本观察值小于 2 组样本观察值的个数＋1 组样本观察值小于 3 组样本观察值的个数＋2 组样本观察值小于 3 组样本观察值的个数。

SPSS 除了计算观察的 $J-T$ 统计量外，还将计算所有情况下的 $J-T$ 统计量，即：如果仍有 1、2、3 三组样本，除了按照（1，2，3）的顺序计算 $J-T$，还按照（1，3，2）、（2，1，3）、（2，3，1）、（3，1，2）、（3，2，1）的顺序计算所有的 $J-T$ 统计量，并求出这些 $J-T$ 统计量的均值、标准化均值、标准差。

可以直观理解，如果观察的 $J-T$ 统计量远大于或远小于 $J-T$ 均值，那么可以认为，按照样本组标志值的升序，样本数据有明显的上升趋势或下降趋势，从而判定样本来自的多个独立总体的分布存在显著差异。

SPSS 计算出上述统计量后，会依据 Jonkheer－terpstra 检验临界值表给出 $J-T$ 统计量对应的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则应拒绝 H_0，认为样本来自的多个独立总体的分布存在显著差异；如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，认为样本来自的多个独立总体的分布不存在显著差异。

二、SPSS 多独立样本的非参数检验的基本操作步骤

首先应明确 SPSS 对多个独立样本非参数检验的数据要求：应设计两个 SPSS 变量。其中：一个变量存放多组样本的数据；另一个变量存放样本组标志数据，即每个观察值所对应的样本组号。

SPSS 多独立样本的非参数检验的基本操作步骤是：

（1）菜单选项：Statistics－＞Nonparametric Test－＞K Independent Samples，于是出现如图 7—6 所示的窗口。

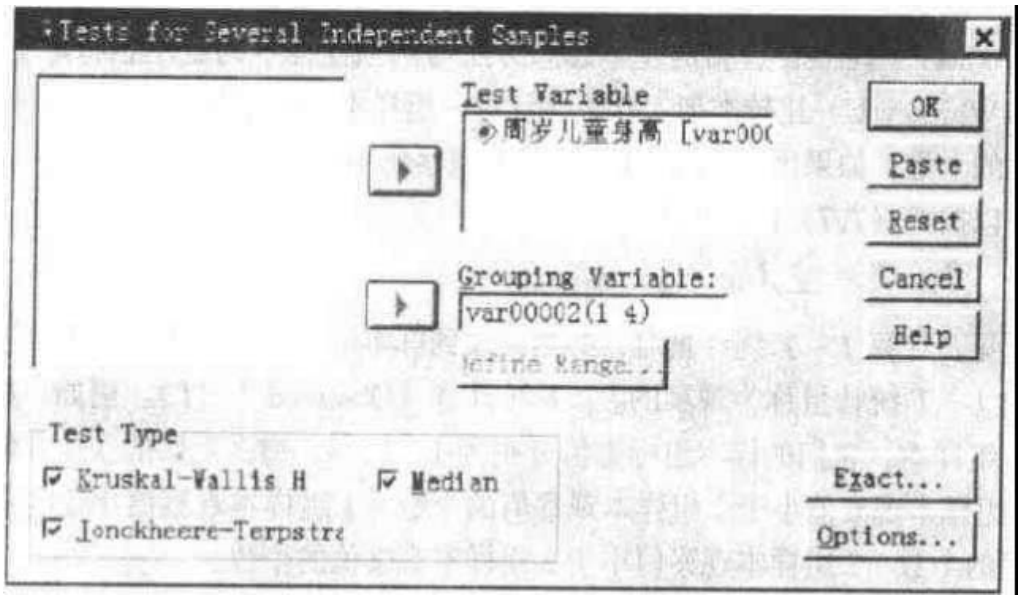


图 7—6　SPSS 多独立样本的非参数检验窗口

（2）选择待检验的变量到 Test Variable 框中。

（3）指定存放样本组标志值的变量到 Grouping Variable 框，并根据具体情况分别给出样本组标志值的最小值和最大值。

（4）选择一种或几种非参数检验方法。

SPSS 两独立样本的非参数检验的其他选项同前，这里不再重复。

三、举例

上面四城市周岁儿童身高的问题，利用多独立样本的非参数检验作分析。分析结果如输出结果 7—6 所示。

输出结果 7—6（a）　　多独立样本的中位数检验举例分析结果

Frequencies

		样本标志			
		1.00	2.00	3.00	4.00
周岁儿童身高	>Median	4	0	5	0
	<= Median	1	5	0	5

Test Statistics

	周岁儿童身高
N	20
Median	74.0000
Chi-Square	16.768
df	3
Asymp. Sig.	.001

从输出结果 7—6（a）可以看出：共同中位数为 74，计算出的 χ^2 统计量为 16.768，其对应的相伴概率值为 0.001，小于显著性水平 0.05，应拒绝零假设，认为四个城市的周岁儿童身高中位数有显著差异。但由于样本数较少，表格单元中的期望频数小于 5，因此，这个结论只能作为参考。

输出结果 7—6（b）　　多独立样本的 K－W 检验举例分析结果

Ranks

	样本标志	N	Mean Rank
周岁儿童身高	1.00	5	14.40
	2.00	5	8.20
	3.00	5	15.80
	4.00	5	3.60
	Total	20	

Test Statistics

	周岁儿童身高
Chi-Square	13.900
df	3
Asymp. Sig.	.003

从输出结果 7—6（b）可以看出：各样本的平均秩与前面讨论的结果是一致的，统计量所对应的相伴概率值小于显著性水平 0.05，因此，应拒绝零假设，认为四个城市周岁儿童的身高分布存在显著差异。

输出结果 7—6（c）　多独立样本的 Jonckheere－Terpstra 检验举例分析结果

Jonckheere－Terpstra Test

	周岁儿童身高
Number of Levels in 样本标志	4
N	20
Observed J-T Statistic	45.500
Mean J-T Statistic	75.000
Std. Deviation of J-T Statistic	14.764
Std. J-T Statistic	-1.998
Asymp. Sig. (2-tailed)	.046

从输出结果 7—6（c）可以看出：由于相伴概率值小于显著性水平 0.05，因此，应拒绝零假设，认为四个城市周岁儿童的身高分布存在显著差异。

§7.4　两配对样本非参数检验

SPSS 两配对样本的非参数检验是在对总体分布不很了解的情况下，对样本来自的两个相关配对总体的分布进行推断。关于两配对总体的特点在 SPSS 参数检验的有关章节曾经讨论过。例如：要检验一种新的训练方法是否对提高跳远运动员的成绩有显著效果，可以收集一批跳远运动员在使用新的训练方法前后的跳远最好成绩，这两组样本便是配对的。再例如：分析不同的广告形式是否对商品的销售额产生显著影响。可以比较几种不同商品在其他条件保持基本稳定的条件下，采用不同的广告形式下的销售额样本数据。这里，不同广告形式下的商品销售额样本数据就是配对的。由此可见，配对样本的观察值个数应是相同的，并且，各个样本观察值的先后次序是不能随意更改的。

一、SPSS 两配对样本非参数检验的基本思路

这里我们讨论三种两配对样本的非参数检验方法。

1．McNemar 变化显著性检验

McNemar 变化显著性检验将研究对象自身作为对照者，检验其“前后”的变化是否显著。其零假设是 H_0：样本来自的两配对总体分布无显著差异。

例如：分析一批学生对统计学在社会生活中的重要性的认识，可以收集到他

们在学习统计学课程之前的看法以及在学完统计学课程以后的看法。假设，我们用 0 表示认为统计学不重要，用 1 表示重要。那么，学习之前和之后的数据就是两组配对样本。

学生对统计学重要性的认识可能会有下列变化：一是，在学习前认为不重要，在学习后认为重要；二是，在学习前认为不重要，在学习后仍认为不重要；三是，在学习前认为重要，在学习后仍认为重要；四是，在学习前认为重要，在学习后却认为不重要。可以把这些数据归纳成表 7—9 的频数表。

表 7—9　　两配对样本的 McNemar 变化显著性检验

学习前	学习后		
	频数	不重要（0）	重要（1）
	不重要（0）	a	b
	重要（1）	c	d

其中：$a+b-c+d$ 为被调查的学生总数。McNemar 变化显著性检验通过对学习前后认识的变化程度，即对上面频数表的分析来推断配对总体的分布。它注重对发生变化的两格中频数 b 和 c 中的研究。

可以直观理解，如果频数 b 和 c 大致相当，即：认为统计学不重要到重要的人数与认为统计学重要到不重要的人数是大致相当，那么，就可以理解为：学习前后学生的认识态度在总体上并没有发生显著的变化。

McNemar 变化显著性检验的基本方法仍采用二项分布检验。它通过对表 7—9中频数 b 和 c 频率的研究，计算二项分布概率值。如果概率值小于或等于用户心中的显著性水平 α，则拒绝 H_0，认为样本来自的两个配对总体的分布有显著差异；如果概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，可以认为样本来自的两个配对总体的分布无显著差异。

需要注意的是：SPSS 中两配对样本的 McNemar 变化显著性检验要求待校验的两组样本的观察值是二值数据，因此，在实际分析中有一定的局限性。

2. 两配对样本的符号（Sign）检验

当我们收集的样本数据不是二值数据时就无法利用 McNemar 变化显著性检验的方法来检验两配对总体的分布是否存在显著差异。这时，可以利用符号检验的方法实现。两配对样本的符号检验的零假设是 H_0：样本来自的两配对总体分布无显著差异。

两配对样本的符号检验的基本思想是：利用正、负符号个数的多少来进行检验。首先，分别将第二组样本的各个观察值减去第一组样本的各个观察值，如果

差值为正，则记为正号，差值为负，记为负号；然后，计算正号的个数 S_+ 和负号的个数 S_-，$S_+ + S_-$ 为总观察值个数 n。

可以直观理解，如果正号个数 S_+ 和负号个数 S_- 大致相当，则可以认为，从总体上讲，第二组样本大于第一组样本的观察值个数，与第二组样本小于第一组样本的观察值个数是大致相当的，这两个配对样本的数据分布差距较小；相反，如果正号个数 S_+ 和负号个数 S_- 相差较多，则可以认为，两个配对样本的数据分布差距较大。

在零假设成立时，两配对样本的符号检验中的正号（S_+）和负号（S_-）的分布是一个概率值为 0.5 的二项分布，因此，现在的任务就是要对差值正负符号序列作单样本二项分布检验。SPSS 将自动根据 S_+、S_- 和 n 计算出实际的概率值。如果概率值小于或等于用户心中的显著性水平 α，则应拒绝 H_0，认为样本来自的两个配对总体的分布有显著差异；如果概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，认为样本来自的两个配对总体的分布无显著差异。

例如：为检验一种新的训练方法是否有助于提高跳远运动员的成绩，收集到了 10 名跳远运动员在使用新的训练方法前后的跳远最好成绩。现在用符号检验方法来作检验，计算过程如表 7—10 所示。

表 7—10　　两配对样本的符号检验计算方法

使用新方法前	使用新方法后	差的符号	使用新方法前	使用新方法后	差的符号
5.74	5.79	+	5.77	5.81	+
6.28	6.12	-	6.41	6.48	+
5.46	5.44	-	6.19	6.32	+
6.03	6.03		5.55	5.64	+
5.39	5.57	+	5.87	5.93	+

可见，正号个数 S_+ 是 7，负号个数 S_- 是 2，不变的个数为 1。

3. 两配对样本的 Wilcoxon 符号平均秩检验

大家可能注意到，在上面的符号检验中，数据信息的利用是不充分的，它只考虑了两配对总体数据变化的性质，即：是变大了还是变小了，但却没有注意其变化的程度，即：大了多少，小了多少。两配对样本的 Wilcoxon 符号平均秩检验在此方面作了改进，其零假设是 H_0：样本来自的两配对总体分布无显著差异。

两配对样本的 Wilcoxon 符号平均秩检验的基本思想是：首先，按照符号检验的方法，分别将第二组样本的各个观察值减去第一组样本的各个观察值，如果差值为正，则记为正号，差值为负，记为负号，并同时保存绝对差值数据；然

后，将绝对差值数据按升序排序，并求出相应的秩；最后，分别计算正号秩总合 W^+、负号秩总合 W^- 以及正号平均秩和负号平均秩。

可以直观理解，如果正号平均秩和负号平均秩大致相当，则可以认为两配对样本数据的正负变化程度基本相当，两配对总体的分布无显著差异。

为作出精确判断，两配对样本的 Wilcoxon 符号平均秩检验将按式（7.8）计算 Z 统计量，它近似服从正态分布：

$$Z=\frac{W-n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}} \tag{7.8}$$

式中，n 为总观察值个数，$W=\min(W^+, W^-)$。SPSS 自动计算 Z 统计量并给出对应的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则应拒绝 H_0，认为样本来自的两个配对总体的分布有显著差异；如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，可以认为样本来自的两个配对总体的分布无显著差异。

例如：仍用上面运动员的例子来作具体说明，计算过程如表 7—11 所示。

表 7—11　　两配对样本的 Wilcoxon 符号平均秩检验计算方法

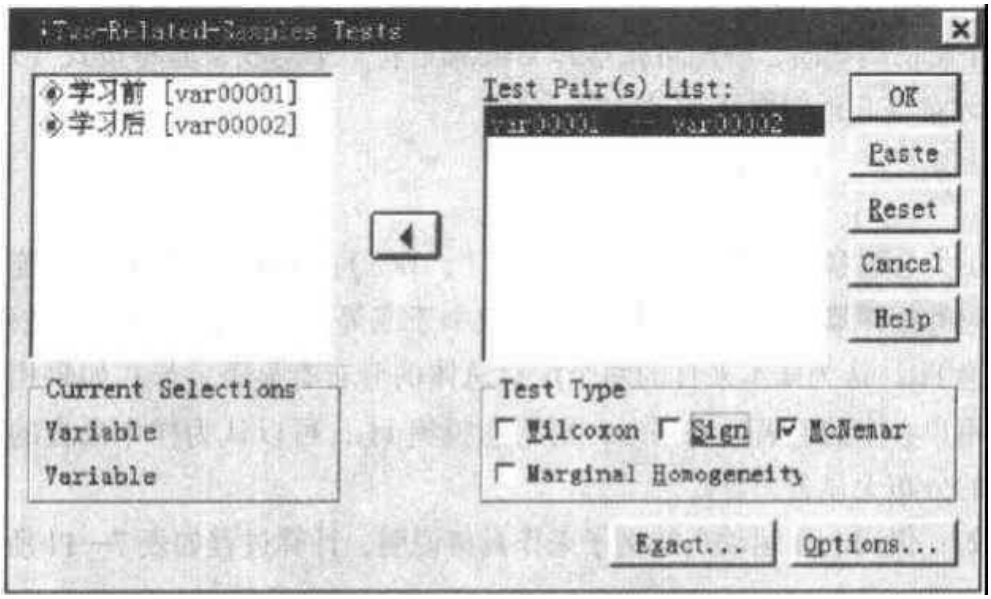
使用新方法前	使用新方法后	绝对差值	秩	差的符号
5.74	5.79	0.05	3	+
6.28	6.12	0.16	8	−
5.46	5.44	0.02	1	−
6.03	6.03	0		
5.39	5.57	0.18	9	+
5.77	5.81	0.04	2	+
6.41	6.48	0.07	5	+
6.19	6.32	0.13	7	+
5.55	5.64	0.09	6	+
5.87	5.93	0.06	4	+

可见，正号秩总和为 36，负号秩总和为 9，正号平均秩和负号平均秩分别为 5.14、4.5。

二、SPSS 两配对样本非参数检验的基本操作步骤

在实际分析操作之前，应先将待检验的两组样本数据分别存放在两个 SPSS

变量中，然后按以下步骤


Two-Related-Samples Tests
学习前 [var00001]
学习后 [var00002]
Test Pair(s) List:
OK
Paste
Reset
Cancel
Help
Current Selections
Variable
Variable
Test Type
Wilcoxon
Sign
McNemar
Marginal Homogeneity
Exact...
Options...

Test Statistics

	学习前与学习后
N	12
Exact Sig. (2-tailed)	.688

从输出结果 7—7 看出：认为统计学不重要变成重要的人数为 4 人，而认为统计学重要变成不重要的人数为 2 人，计算出的二项分布的概率为 0.688。由于这个概率值大于显著性水平 0.05，因此，不能拒绝零假设，认为学习前后学生对统计学的认识没有发生显著变化。

2. 符号检验和 Wilcoxon 符号平均秩检验举例

对上面检验新的训练方法是否有助于提高跳远运动员成绩的例子，进行符号检验和 Wilcoxon 符号平均秩检验。分析结果如输出结果 7—8 所示。

输出结果 7—8（a）　　两配对样本的符号检验举例分析结果

Sign Test

Frequencies

		N
训练后 - 训练前	Negative Differences	2
	Positive Differences	7
	Ties	1
	Total	10

Test Statistics

	训练后 - 训练前
Exact Sig. (2-tailed)	.180

从输出结果 7—8（a）可以看出：10 名运动员中，有 7 人的成绩有所提高，2 人的成绩下降，1 人的成绩仍保持不变。由于 Z 统计量的相伴概率值 0.18 大于显著性水平 0.05，因此，不能拒绝零假设，可以认为新的训练方法没有使跳远运动员的成绩得到显著提高。

输出结果 7—8（b）　两配对样本的 Wilcoxon 符号平均秩检验举例分析结果

Wilcoxon Signed Ranks Test

Ranks

		N	Mean Rank	Sum of Ranks
训练后 - 训练前	Negative Ranks	2[a]	4.50	9
	Positive Ranks	7[b]	5.14	36.00
	Ties	1[c]		
	Total	10		

a　训练后＜训练前

b　训练后＞训练前

c　训练后－训练前

Test Statistics

	训练后 - 训练前
Z	-1.599
Asymp. Sig. (2-tailed)	.110

从输出结果 7—8（b）可以看出：正负平均秩的值和前面讨论的结果是一致的。由于 Z 统计量的相伴概率 0.110，大于显著性水平 0.05，因此，不能拒绝零假设，可以认为新的训练方法没有使跳远运动员的成绩得到显著提高。

§7.5　多配对样本非参数检验

一、SPSS 多配对样本非参数检验的基本思想

多配对样本的非参数检验是对多个相互匹配样本的总体分布是否存在显著差异作统计推断，SPSS 中提供了三种方法，下面分别来讨论。

1. 多配对样本的 Cochran Q 检验

多配对样本的 Cochran Q 检验的零假设是 H_0：样本来自的多配对总体分布无显著差异。

多配对样本的 Cochran Q 检验所处理的样本数据要求是二值的。

例如：根据一项调查，收集到 15 名乘客对三个航空公司服务态度的数据，假定用 1 表示满意，0 表示不满意，样本数据如表 7—12 所示。

表 7—12　　多配对样本的 Cochran Q 检验

乘客号	甲航空公司	乙航空公司	丙航空公司
1	1	1	0
2	1	0	0
3	1	0	0
4	1	0	1
5	1	1	0
6	1	1	0
7	1	0	0
8	1	1	0
9	0	0	0
10	1	0	0
11	1	0	0
12	1	1	1
13	1	1	0
14	1	1	0
15	0	1	0

现在需要根据上面的样本数据推断这三家航空公司服务质量是否有差异。这个问题是一个三配对样本的非参数检验问题，配对的原因是不同的乘客对于满意程度的感受和标准是不尽相同的。如果甲、乙、丙这三组配对样本数据的分布不存在显著差异，则可以认为这三个航空公司的服务质量不相上下。

对配对样本的 Cochran 检验的基本思想是：认为每行中 1 的个数是固定的，也就是说，对每个乘客来说，满意程度的衡量标准是固定的；在零假设成立的条件之下，每列中出现 1 的概率是相等的；这个概率值与每行中 1 的固定个数有关，也就是说，这个概率值与每个乘客的衡量标准有关。依照这种思想构造 Cochran Q 统计量。

$$Q = \frac{k(k-1)\sum_{j=1}^{k}(G_j - \bar{G})^2}{k\sum_{i=1}^{n}L_i - \sum_{i=1}^{n}L_i^2} \tag{7.9}$$

式中：k 为样本数，n 为样本容量；G_j 为第 j 列中取值为真（1）的个数，$\bar{G}$ 为 G_j 的均值；L_i 为第 i 行取值为真（1）的个数。

在观察值个数较大时，Q 统计量近似服从 Chi-Square 分布。SPSS 自动计算 Cochran Q 统计量，并给出对应的相伴概率值。如果相伴概率值小于或等于用户

心中的显著性水平 α，则拒绝 H_0，认为样本来自的多个配对总体的分布有显著差异；如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，可以认为样本来自的多个配对总体的分布无显著差异。

2．多配对样本的 Friedman 检验

多配对样本的 Friedman 检验是利用秩实现对多个配对总体分布进行检验的一种方法，其零假设是 H_0：样本来自的多配对总体分布无显著差异。

多配对样本的 Friedman 检验，要求数据是定距的，使该方法有了较大的适用范围。

例如：对 10 种商品分别采用三种不同的促销形式在三地进行促销，促销 1 个月后，零售额数据如表 7—13 所示。

表 7—13　　多配对样本的 Friedman 检验

商品编号	促销 1	促销 2	促销 3
1	12 866	17 223	9 865
2	4 673	5 894	5 220
3	10 480	14 461	10 072
4	769	1 962	737
5	6 482	13 203	9 423
6	796	742	771
7	843	965	639
8	1 936	1 260	1 793
9	4 694	5 222	4 061
10	635	558	542

现在需要根据上面的样本数据推断这三种促销形式对商品的销售是否产生显著影响。这个问题是一个三配对样本的检验问题，原因是不同商品的销售单价是不同的。分析后，如果发现这三个配对总体的分布不存在显著差异，就可以认为这三种促销形式对商品的销售没有产生显著的差异影响，那么，在以后的促销活动中可以采用其中一种成本较低的形式。

多配对样本的 Friedman 检验的基本方法是：首先，以观察值（行）为单位，将各个样本数据按行升序排序求得各样本数据在各自行中的秩；然后，分别计算各样本的秩总和、平均秩。

可以直观理解，如果多个配对样本的分布存在显著差异，如：第一组样本的

数据普遍偏大，第三组样本的数据普遍偏小，那么在按行排序求秩后，第一组样本的秩总和必然最大，第三组样本的秩总和必然最小，这些秩之间的差异必然较大。因此说、如果各样本的平均秩大体相当，可以认为各总体的分布无显著差异。反之，如果各样本的平均秩相差较大，可以认为各总体的分布存在显著的差异。

为作出精确判断，SPSS按式（7.10）计算Friedman统计量：

$$\chi^2 = \frac{12}{bk(k+1)}\sum_{i=1}^{k}(R_i - \frac{b(k+1)}{2})^2 \qquad (7.10)$$

式中：k 为样本个数，b 为样本观察值数目，R_i 为第 i 组样本的秩总和。

在观察值个数较大时，Frideman统计量服从Chi-Square分布。SPSS自动计算Friedman统计量，并给出对应的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则拒绝 H_0，认为样本来自的多个配对总体的分布有显著差异；如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，可以认为样本来自的多个配对总体的分布无显著差异。

例如：用上述不同促销形式的不同销售额的例子，对Friedman分析方法作具体说明。计算过程如表7—14所示。

表7—14　　多配对样本的Friedman检验计算方法

商品编号	促销1	促销2	促销3	促销1秩	促销2秩	促销3秩
1	12 866	17 223	9 865	2	3	1
2	4 673	5 894	5 220	1	3	2
3	10 480	14 461	10 072	2	3	1
4	769	1 962	737	2	3	1
5	6 482	13 203	9 423	1	3	2
6	796	742	771	3	1	2
7	843	965	639	2	3	1
8	1 936	1 260	1 793	3	1	2
9	4 694	5 222	4 061	2	3	1
10	635	558	542	3	2	1
秩总合				21	25	14
平均秩序				2.1	2.5	1.4

3. 多配对样本的Kendall协同系数检验

多配对样本的Kendall协同系数检验也是一种多配对样本的非参数检验方法，与Friedman检验的方法非常类似，但分析问题的角度有所不同。多配对样

本的 Kendall 协同系数检验主要用在分析评判者的评判标准是否一致公平方面，它将每个评判对象的分数都看作是来自多个配对总体的样本。其零假设是：样本来自的多配对总体分布无显著差异，即评判者的评判标准不一致。

例如：现在有 6 名歌手参加比赛，4 名评委进行评判打分，最后结果如表 7—15 所示。

表 7—15　　多配对样本的 Kendall 协同系数检验

评委/歌手	1 号	2 号	3 号	4 号	5 号	6 号
1	8.75	9.60	9.20	9.65	9.30	9.80
2	8.90	9.55	9.25	9.75	9.45	9.75
3	8.75	9.70	9.25	9.60	9.30	9.70
4	8.80	9.60	9.25	9.75	9.40	9.85

现在需要根据表中的数据推断这 4 名评委的评判标准是否一致。这个问题实质仍是一个多配对样本的检验问题，虽然它研究问题角度，不是推断 6 名歌手的水平是否存在显著差异，而是推断 4 名评委的打分标准是否一致，但采用的基本方法还是 Friedman 检验方法。

可以直观理解，如果利用 Friedman 检验方法检验出 6 名歌手的水平大致相当，即这 6 名歌手的平均秩没有较大差异，那么可以认为这 4 名评委的打分有随意性，评分标准不一致。因为，如果评委的评判标准是一致的，那么，对于某个歌手来说，他将获得一致的打分，即：4 名评委给他的评分的名次或秩应当是完全相同的，这就必然造成各歌手平均秩之间有较大的差异。

除此之外，多配对样本的 Kendall 协同系数检验还按式（7.11）构造了协同系数 W：

$$W = \sum_{i=1}^{n} \frac{(R_i - m(n+1)/2)^2}{(m^2 n(n^2 - 1))/12} \tag{7.11}$$

式中：m 是评判人数，n 是被评判人数，R_i 为第 i 个被评判者的秩总和，W 在 n 较大时近似服从 Chi-Square 分布。

协同系数 W 反映了各行数据之间的相关程度，其取值在 0～1 之间。W 越接近 1 则表示各行数据之间的相关性越强，从另一个角度说明评委的评判标准是一致的。SPSS 将自动计算协同系数，并给出对应的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则拒绝 H_0，认为评判者的评判标准一致；如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，认评判者的评判标准不一致。

二、SPSS 多配对样本的非参数检验的基本操作步骤

应首先准备好待检验的样本数据，每个样本应设计成一个 SPSS 变量，然后按下列步骤进行分析：

（1）菜单选项：Statistics－＞Nonparametric－＞K Paired Sample。于是出现如图 7—8 所示的窗口。

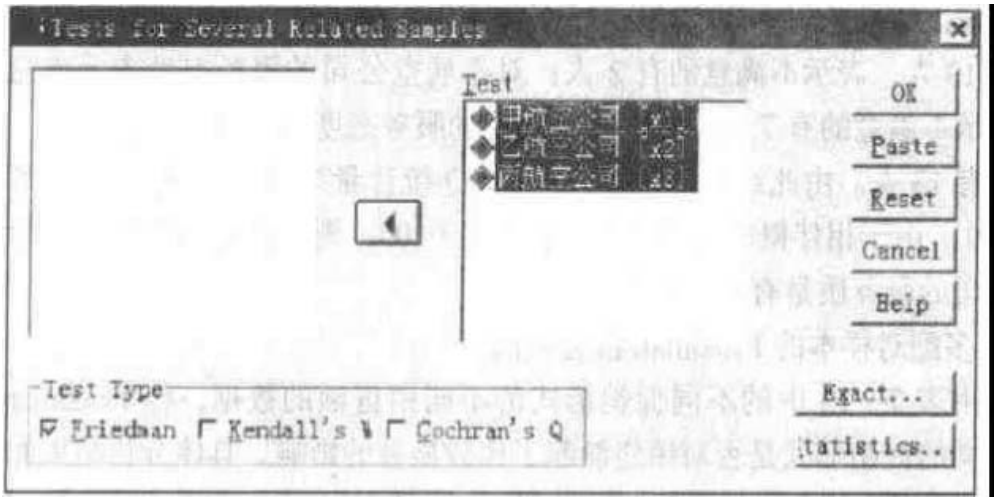


图 7—8　SPSS 多配对样本的非参数检验窗口

（2）选择待检验的若干个配对变量到 Test Variables 框中。

（3）选择一种或几种非参数检验方法。

SPSS 两独立样本的非参数检验的其他选项同前，这里不再重复。

三、举例

1. 多配对样本的 Cochran's Q 检验举例

利用表 7—12 中乘客对不同航空公司服务态度的满意度调查的数据，作 Cochran's Q 检验，分析乘客对三家航空公司服务态度的满意度是否存在显著差异。具体分析结果如输出结果 7—9 所示。

输出结果 7—9　　多配对样本的 Cochran's Q 检验举例分析结果

Cochran's Q Test

Frequencies

	Value	
	0	1
甲航空公司	2	13
乙航空公司	7	8
丙航空公司	13	2

Test Statistics

N	15
Cochran's Q	14.000
df	2
Asymp. Sig.	.001

从输出结果 7—9 可以看出：15 名乘客中，对甲航空公司的服务态度表示满意的有 13 人，表示不满意的有 2 人；对乙航空公司的服务态度表示满意的有 8 人，表示不满意的有 7 人；对丙航空公司的服务态度表示满意的有 2 人，表示不满意的有 13 人。由此计算出的 Cochran's Q 统计量为 14，其对应的相伴概率值为 0.001。由于相伴概率值小于显著性水平 0.05，则应拒绝零假设，认为这三个航空公司的服务质量有较大差距。

2. 多配对样本的 Friedman 检验举例

利用表 7—13 中的不同促销形式的不同销售额的数据，作 Friedman 检验，分析不同的促销形式是否对销售额起了比较显著的影响。具体分析结果如输出结果 7—10 所示。

输出结果 7—10　　多配对样本的 Friedman 检验举例分析结果

Ranks

	Mean Rank
促销形式 1	2.10
促销形式 2	2.50
促销形式 3	1.40

Test Statistics

N	10
Chi-Square	6.200
df	2
Asymp. Sig.	.045

从输出结果 7—10 可以看出：计算出的各种促销形式的平均秩与前面表 7—14 中结果是一致的，由此计算出的 Friedman 统计量为 6.2，其对应的相伴概率值为 0.045。由于相伴概率值小于显著性水平 0.05，因此，应拒绝零假设，认为不同的促销形式对商品的销售额产生了比较显著的影响。

3. 多配对样本的 Kendall 协同系数检验举例

利用表 7—15 中歌手的成绩数据，作 Kendall 检验，分析评委的评判标准是

否一致。具体分析结果如输出结果 7—11 所示。

输出结果 7—11　　多配对样本的 Kendall's *W* 检验举例分析结果

Ranks

	Mean Rank
1 号歌手得分	1.00
2 号歌手得分	4.38
3 号歌手得分	2.00
4 号歌手得分	4.88
5 号歌手得分	3.00
6 号歌手得分	5.75

Test Statistics

N	4
Kendall's W	0.955
Chi-Square	19.094
df	5
Asymp. Sig.	0.002

从输出结果 7—11 可以看出：1 至 6 号歌手的平均秩序分别为：1、4.38、2、4.88、3、5.75，由此计算出的 Frideman 统计量为 19.094，其对应的相伴概率值为 0.002。由于相伴概率值小于显著性水平 0.05，因此应拒绝零假设，认为评委的评分标准是一致的。另外，由于 Kendall 协同系数 W 为 0.955，非常接近于 1、表示各个歌手的分数相关性非常高，也证明了评委的评分标准是一致的。

第 8 章

SPSS 的方差分析

§8.1 方差分析概述

方差分析方法在不同领域的各种分析研究中得到了极为广泛的应用。从方差入手的研究方法有助于我们找到事物的内在规律性。

例如：在农业生产中，我们总是希望在尽量少的投入成本下得到较高的农作物产量。这就需要首先分析农作物的产量究竟受到哪些因素的影响。有许多因素会影响农作物的产量，如：不同品种的种子、不同的施肥量、不同的气候、不同地域，等等，它们都会给农作物的产量带来或多或少的影响。如果我们能够掌握，在这众多的影响因素中，哪些因素对农作物的产量起到了主要的、关键性的影响，我们就可以根据实际情况对这些关键性的因素加以控制。进一步，如果我们掌握了某个关键性的影响因素，如：种子，知道哪种品种的种子更能有效地提高产量，那么我们今后就选用这样的优良品种。同样，对农作物施多少肥料也是一个关键性的影响因素。如果我们能够掌握一个较适宜的施肥量，不仅能够提高产量，也能够降低投入成本。

为实现这样的分析研究，首先，有必要对影响因素加以分析。可以看到，对

农作物的影响因素可以大致分成两大类。其中：第一类因素是人为可以控制的因素，如：种子品种的选定，施肥量的多少等；第二类因素是人为很难控制的因素，如：气候的影响、地域的影响等；然后，针对第一类因素作对比试验。如考察种子因素，可以选用 A、B 两种不同的种子，在其他条件都相近的两批地块中分别种植。考察施肥量因素，可以选用 10 公斤、20 公斤、30 公斤三种不同的施肥量，在其他条件都相近的三批地块中分别施肥。于是，可以得到不同品种的农作物产量数据，或在不同施肥量下的农作物产量的数据。对这些数据进行分析，如果发现 A、B 两种品种的种子所获得的产量相差不多，那么就可以认为 A、B 品种没有对农作物的产量造成显著影响，今后选用哪种品种的种子都可以。同样，如果发现三种不同的施肥量确实给农作物的产量带来了较大的差异，而且，施 20 公斤化肥的地块产量明显高于施 10 公斤化肥的地块产量，但不明显高于施 30 公斤化肥的地块，那么今后就应施 20 公斤肥料，它不仅有助于提高产量，同时能够降低成本。

方差分析方法就是实现上述分析的。在方差分析中，称上述能够人为控制的第一类影响因素为控制因素，或控制变量，称上述人为很难控制的第二类影响因素为随机因素，或随机变量；受控制因素和随机因素影响的事物，诸如：农作物产量，为观察变量；称上述 A、B 两个品种或 10 公斤、20 公斤、30 公斤化肥为控制变量的不同水平。

方差分析就是要分析控制变量的不同水平是否对观察变量产生了显著影响。如果控制变量的不同水平对试验结果产生了显著影响，那么，它和随机变量共同作用必然使得观察变量数据有显著变动；相反，如果控制变量的不同水平对试验结果没有产生显著影响，那么，观察变量数据的变动就不会明显表现出来，它的变动可以归结为受随机变量影响造成的。

根据控制变量个数，可以将方差分析分成单因素方差分析和多因素方差分析。单因素方差分析中的控制变量只有一个，多因素方差分析中的控制变量有多个。

§8.2 单因素方差分析

单因素方差分析就是测试某一个控制变量的不同水平是否给观察变量造成了显著的差异和变动。例如：测试由于施肥量的不同是否给某农作物产量造成显著差异；考察地区的不同是否给妇女的生育率带来显著差异；考察不同学历是否对

工资收入产生显著影响等都是单因素方差分析问题。

一、单因素方差分析的基本思路

单因素方差分析实质也采用了统计推断的方法，其目的是通过实验数据（样本数据）研究观察变量在一个控制变量中的若干个不同水平下，其各个总体在分布上是否存在显著差异。由于方差分析有一个比较严格的前提要求，即不同水平下的各总体应服从方差相同的正态分布，因此方差分析问题就转换成研究不同水平下的各总体的均值有无显著差异的问题。

单因素方差分析的零假设是 H_0：在控制变量的不同水平下，各总体均值无显著差异。即控制变量的不同水平对观察变量没有产生显著的影响。

采用的统计推断方法是计算 F 统计量，进行 F 检验。单因素方差分析将观察变量总的变差平方和，记为 SST，分解为两个部分：一部分是由控制变量引起的变差，记为 SSA；另一部分是由随机变量引起的变差，记为 SSE。于是有：

$$SST = SSA + SSE \tag{8.1}$$

式中：SSA 也称为组间离差平方和（Between Groups），SSE 称为组内离差平方和（Within Groups）。

SSA 的数学定义为：

$$SSA = \sum_{i=1}^{k} n_i (\overline{x}_i - \overline{x})^2 \tag{8.2}$$

式中：k 为水平数，n_i 为第 i 个水平下的样本容量。可见，组间离差平方和是各水平组均值与总均值离差的平方和，反映了控制变量的影响。

SSE 的数学定义为：

$$SSE = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (x_{ij} - \overline{x}_i)^2 \tag{8.3}$$

可见，组内离差平方和是每个数据与本水平组平均值离差的平方和，反映了数据抽样误差的大小程度。

F 统计量为：F =平均组间平方和/平均组内平方和，数学定义为：

$$F = \frac{SSA/(k-1)}{SSE/(n-k)} = \frac{MSA}{MSE} \tag{8.4}$$

式中：$n-k=k(n_i-1)$。可见，如果控制变量的不同水平对观察变量造成了显著影响，那么，观察变量总的变差中，控制变量影响所占的比例必然较大，于是，F 值就比较大；相反，如果控制变量的不同水平没有对观察变量造成显著

影响，那么，观察变量的变差应归结为随机变量造成的，随机变量影响所占的比例必然较大，于是，F 值就比较小。

SPSS 自动计算 F 统计量，由于 F 值服从（$k-1$，$n-k$）个自由度的 F 分布（k 是水平数，n 为个案数），SPSS 依据 F 分布表给出相应的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则拒绝 H_0，认为控制变量的不同水平下各总体均值有显著差异，即控制变量的不同水平给观察变量带来了显著影响；相反，如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，认为控制变量的不同水平下各总体均值没有显著差异，即可以认为控制变量的不同水平没有给观察变量带来显著影响。

二、SPSS 单因素方差分析的基本操作步骤

SPSS 单因素方差分析要求定义两个 SPSS 变量，它们分别存放观察变量数据和控制变量数据。其中，控制变量最好采用顺序水准的数据，不同数值将表示控制变量的不同水平。

SPSS 单因素方差分析的基本操作步骤是：

（1）菜单选项：Statistics－>Compare Means－>One－Way ANOVA。于是出现如图 8—1 所示的窗口。

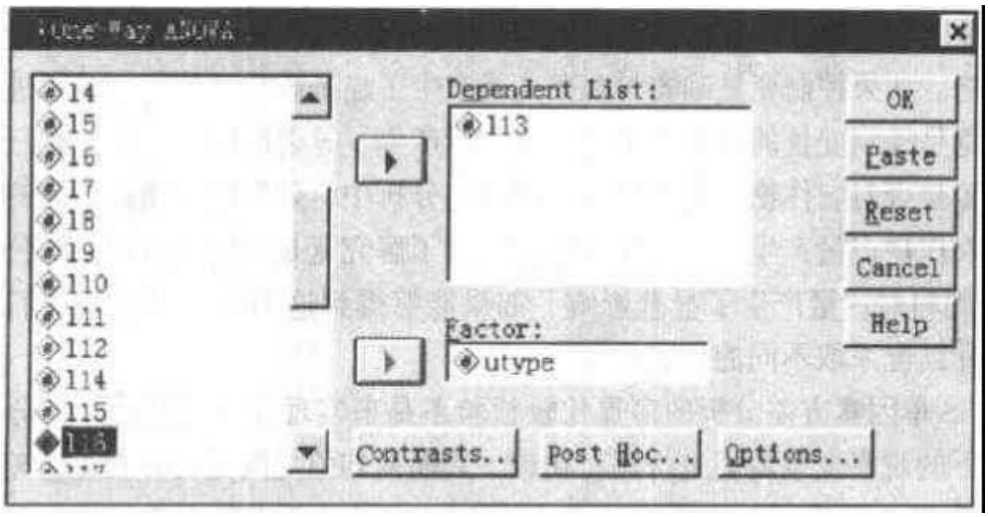


图 8—1　SPSS 单因素方差分析窗口

（2）选择一个或多个变量作为观察变量到 Dependent List 框中。

（3）选择一个变量作为控制变量到 Factor 框。该控制变量有几个不同的取值就表示有几个不同的水平。

三、SPSS 单因素方差分析的进一步分析

上面我们讨论了 SPSS 单因素方差分析的基本分析方法和步骤，下面将对 SPSS 单因素方差分析进一步讨论。

1. SPSS 单因素方差分析的前提检验

由于方差分析的前提是各水平下的总体服从方差相等的正态分布。其中，正态分布的要求不是非常严格，但对于方差相等的要求是比较严格的。因此，我们有必要对方差分析的前提进行检验。

SPSS 单因素方差分析中的方差相等检验采用的方法是 Homogeneity－of－variance 方法。它也是采用统计推断的方法，其零假设是 H_0：各水平下总体方差无显著差异。具体的检验方法与 SPSS 独立样本 t 检验中的方差相等检验方法完全相同，这里不再重复。

SPSS 单因素方差分析的前提检验的操作步骤：按 Option 按钮选项，选定 Homogeneity－of－variance 选项。于是，在方差分析的结果中将出现关于方差是否相等的检验结果和相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则拒绝 H_0，认为各水平下总体方差不等；相反，如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，可以认为各水平下总体方差相等。

2. SPSS 单因素方差分析的多重比较检验

通过上面的检验，我们只能判断控制变量的不同水平是否对观察变量产生了显著影响。如果控制变量确实对观察变量产生了显著影响，我们会希望进一步了解：究竟是控制变量的哪个水平对观察变量产生了显著影响。例如：在三种不同的施肥量是否对农作物产量产生显著影响的分析中，如果已经确定不同的施肥量确实对农作物产量产生影响，便希望进一步了解究竟是 10 公斤或是 20 公斤或是 30 公斤肥料对产量产生了显著影响。如果能够得到这个分析结论，我们就可以根据实际情况采取不同的施肥量方案。

SPSS 单因素方差分析的多重比较检验正是来实现这个功能的。它分别对每个水平下的观察变量均值进行逐对比较，判断两均值之间是否存在显著差异。其零假设是 H_0：相应组的均值之间无显著差异。SPSS 提供的多重比较检验的方法比较多，有些方法适用在各总体方差相等的条件下，有些适用在方差不相等的条件下。它们对均值是否存在显著差异的具体判断标准也不相同。由于涉及的概念较多，这里我们不对每个方法作具体讨论。在实际应用中，用户可以选择其中的几种方法尝试着作比较判断。

SPSS 单因素方差分析的多重比较检验的操作步骤：

按Post Hoc按钮选项，然后选择一种或几种比较分析方法。

于是，在方差分析的结果中将出现关于多重比较判断的检验结果。其中：出现的*号，表示在显著性水平为默认的0.05时，相应两组水平下的观察变量均值存在显著差异。

3．SPSS单因素方差分析的多项式检验和先验对比检验

SPSS单因素方差分析中的多项式检验可以将组间平方和分解成线性、二次、三次或更高次的多项式。于是，在方差分析表中，不仅只计算输出组间平方和，还将计算显示组间平方和的各分解结果以及F统计量和相应的相伴概率值。在分解时，由于各水平中的个案数目可能相同也可能不相同，这样具体的分解计算方法也会不同，因此，SPSS将自动分别按照不同的方法计算组间平方和的分解结果。

（1）SPSS单因素方差分析的多项式检验的操作步骤：

按Contrasts按钮选项，然后在选定Polynomial，并在Degree框中指定是作线性分解，还是二次或三次等分解。

例如：如果指定对组间平方和作线性分解，那么实质就是对观察变量（因变量）与控制变量（自变量）进行一元线性回归分析，计算回归平方和，并对回归方程进行检验，给出F统计量和相应的相伴概率值。如果相伴概率值大于用户心中的显著性水平α，说明控制变量各水平的变化无法反映观察变量的线性变化，也就是认为控制变量的不同水平对观察变量的线性影响是不显著的；相反，如果相伴概率值小于或等于用户心中的显著性水平α，说明观察变量随控制变量各水平的变化而产生了线性变化，也就是认为控制变量的不同水平对观察变量的线性影响显著。同时，还给出方差分析的组间平方和与回归平方和之间的差。

SPSS单因素方差分析中的先验对比检验，要求用户应事先对各水平间观察变量的均值对比有大致的了解和把握，然后再对这种大致把握作出统计检验。这里，它的零假设仍是H_0：相应组的均值之间无显著差异。在检验中，SPSS根据用户凭经验确定的各水平间观察变量均值对比系数，以正负符号确定不同的对比组，然后利用t检验方法判定相应对比组的均值是否为先验指定的对比系数。

（2）SPSS单因素方差分析中的先验对比检验的操作步骤是：

按Contrasts按钮选项，然后在Coefficients框中输入每个水平均值的对比系数值，并要求以正负符号分别代表对比组。

在分析结果中，SPSS将分别计算各总体方差相等和不相等条件下的t统计

量，并给出相应的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平 α，则拒绝 H_0，认为各对比组间的均值比与给定的均值对比系数有显著差异；相反，如果相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，可以认为各对比组间的均值比与给定的均值对比系数间不存在显著差异。

可见，SPSS 单因素方差分析中的多重比较检验和先验对比检验已不再是方差分析本身的问题，涉及到两独立样本 t 检验的问题。

四、SPSS 单因素方差分析的 Option 选项

SPSS 单因素方差分析中 Option 选项主要用来指定输出分析结果以及对缺失值如何处理。其中：

（1）Statistics 框中的 descriptive 表示：输出观察变量在控制变量不同水平下的描述统计量。

（2）Missing Values 框中的 Exclude case analysis by analysis 表示只有当计算涉及的变量含有缺失值时，才暂时剔除那些在该变量上为缺失值的个案；Exclude case listwise 表示将所有含缺失值的个案全部剔除后再分析。

（3）Means plot 表示绘制各水平下观察变量均值的折线图。

§8.3 单因素方差分析案例分析

查阅《中国普通高等院校人文社会科学统计资料汇编》，得到全国重点院校中的 80 所院校的 1999 年人文社科基本情况数据。指标涉及以下几个重要方面：社科活动人员、研究与发展人员、研究与发展经费、研究与发展课题、研究与发展成果、获奖成果。这 80 所院校分布在全国 25 个省市自治区，有综合院校，也有一般的普通院校，办学性质有综合、工科、农业、医药、高等师范、语言、财经、政法等多种类型。由于篇幅所限，具体指标和数据略去。

现在需要利用这些指标数据，分析不同办学类型的院校，它们在研究与发展经费的投入方面是否存在较大的差异。

进行单因素方差分析，指定当年投入的研究与发展经费为观察变量，不同的办学性质为控制变量。分析结果见输出结果 8—1 所示。

输出结果 8—1（a）表明，重点院校中，工科院校最多为 47 所，其次是综合院校。

输出结果 8—1（a）　　案例单因素方差分析结果（一）

Descriptives

L13

	N	Mean	Std. Deviation	Std.Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
综合	12	57639.250	35302.186	10190.863	35209.311	80069.189	10623.0	136181.0
工科	47	9607.362	21294.762	3106.160	3354.985	15859.738	35.0	136378.0
农业	6	5375.833	2977.742	1215.658	2250.885	8500.782	1270.0	8820.0
医药	6	197.667	221.650	90.488	-34.941	430.274	.0	450.0
高等师范	3	36768.000	27963.756	16144.882	-32697.821	106233.821	7887.0	63714.0
语言	2	1120.000	1272.792	900.000	-10315.584	12555.584	220.0	2020.0
财经	2	3040.000	494.975	350.000	-1407.172	7487.172	2690.0	3390.0
政法	2	3269.000	3155.110	2231.000	-25078.543	31616.543	1038.0	5500.0
Total	80	16272.750	28347.199	3169.313	9964.390	22581.110	.0	136378.0

1999 年，这 80 所院校的平均年投入研究与发展经费为 16 272.75 百元，而综合院校和高等师范院校的平均投入经费要明显高于其他类型的院校，医药类院校经费投入最低。同时，也可以看出经费投入在各类学校内部也不平衡。以综合院校为例，最少的年投入经费为 10 623 百元，最高的年投入经费为 136 181 百元，相差 10 倍多。

输出结果 8—1（b）　　案例单因素方差分析结果（二）

Test of Homogeneity of Variances

L13

Levene Statistic	df1	df2	Sig.
2.097	7	72	.055

输出结果 8—1（b）是单因素方差分析的方差是否相等的检验结果。相伴概率值为 0.055，如果显著性水平设为 0.05，可以近似认为不同办学类型下各总体的方差无显著差异，基本满足单因素方差分析中方差相等的要求。

输出结果 8—1（c）是单因素方差分析的主要分析结果。从中可以看出：经费投入的总的离差平方和为 63 481 533 197，其中：由控制变量不同水平造成的组间平方和为 27 293 026 159.732，由随机变量造成的组内平方和为

36 188 507 037.268。组间平方和中能够被控制变量线性解释的平方和(回归平方和)为 4 083 525 927.868,其余不能被线性解释的平方和为 23 209 500 231.864,它与组内平方相加即为回归分析中的剩余平方和。

输出结果 8—1（c）　　案例单因素方差分析结果（三）

ANOVA

L13

			Sum of Squares	df	Mean Square	F	Sig.
Between Groups	(Combined)		27293026159.732	7	3899003737.105	7.757	.000
	Linear Term	Weighted	4083525927.868	1	4083525927.868	8.125	.006
		Deviation	23209500231.864	6	3868250038.644	7.696	.000
Within Groups			36188507037.268	72	502618153.295		
Total			63481533197.000	79			

F 统计量 7.757 是平均的组间平方和除以平均的组内平方和的结果，其对应的相伴概率值为 0.00，小于显著性水平 0.05，因此应拒绝零假设，认为不同办学类型的院校在人文社科研究与发展的年投入经费上存在显著差异；*F* 统计量 8.125 和 7.696 分别是：平均的回归平方和除以平均的组内平方和、不能被控制变量线性解释的平均平方和除以平均的组内平方和的结果，它们对应的相伴概率值分别为 0.006 和 0.000，同样都小于显著性水平 0.05，表明它们比组内差要大的多，也证明了研究与发展年经费投入差异主要是由不同办学类型造成的。

输出结果 8—1（d）是单因素方差分析中用 LSD 方法进行多重对比检验的结果。由于篇幅所限，输出结果 8—1（d）只是其中一部分。第一列和第二列分别表示将控制变量的第一个水平分别与控制变量的其他水平作多重比较检验；第三列是两两水平上观察变量的均值差，数据后面的星号表明该两组间的均值在显著性水平为 0.05 的条件下，存在显著差异；第四列为差值数据的均值标准误差；第五列为 LSD 检验的相伴概率值。如果相伴概率值小于或等于用户心中的显著性水平，则应拒绝 LSD 多重比较检验的零假设，认为相应的两组均值间存在显著差异；其余两列为均值差的 95%的置信区间。

输出结果 8—1（d）　　案例单因素方差分析结果（四）

Post Hoc Tests

Multiple Comparisons

Dependent Variable: L13

LSD

		Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
(I) UTYPE	(J) UTYPE				Lower Bound	Upper Bound
综合	工科	48031.868*	7251.127	.000	33577.030	62486.747
	农业	52263.417*	11209.574	.000	29917.540	74609.293
	医药	57441.583*	11209.574	.000	35095.707	79787.460
	高等师范	20871.250	14471.497	.154	-7977.152	49719.652
	语言	56519.250*	17122.906	.002	22385.360	90653.140
	财经	54599.250*	17122.906	.002	20465.360	88733.140
	政法	54370.250*	17122.906	.002	20236.360	88504.140

* The mean difference is significant at the .05 level.

从输出结果 8—1（d）不难得到以下结论：综合类大学平均研究与发展年投入经费只与高等师范类院校的平均年投入经费没有明显差距，与其他类型的院校间都存在比较显著的差异。同时，从各类院校的平均年投入经费来看，综合类和高等师范类院校明显高于其他类型的院校，这是为什么不同类型的学校经费投入存在显著差异的主要原因。

§8.4　均值比较

一、什么是 SPSS 的均值比较

SPSS 均值比较以计算基本描述统计量为主要内容，后面利用单因素方差分析的方法实现均值的比较分析。其基本思路是：

第一步，用户应首先指定一个或多个变量作为分组变量，并按分组变量的不同取值对个案数据进行分组。如果分组变量有多个，还应指定这些分组变量之间的层次关系。层次关系可以是同层次的或多层次的。同层次意味着将按照各分组

变量的不同取值分别对个案作分组；多层次意味着将首先按第一分组变量分组，然后对各分组下个案再按第二分组变量的不同取值继续分组，这样逐层向下细分下去。如果从单因素方差分析的角度看，这里第一分组变量就是控制变量，各个分组就是控制变量下的各个水平。

第二步，指定一个变量作为汇总变量，并计算该汇总变量在各分组下的基本描述统计量。如果从单因素方差分析的角度看，这里汇总变量就是观察变量。

第三步，对第一分组变量各分组下的汇总变量均值有无显著差异作统计检验。如果从单因素方差分析的角度看，就是对控制变量不同水平下观察变量的均值是否有显著差异进行推断。其零假设是 H_0：第一分组变量各水平下汇总变量均值无显著差异。

二、SPSS 均值比较的基本操作步骤

(1) 菜单选项：Statistics－>Compare Means－>Means。于是出现如图 8—2 所示的窗口。

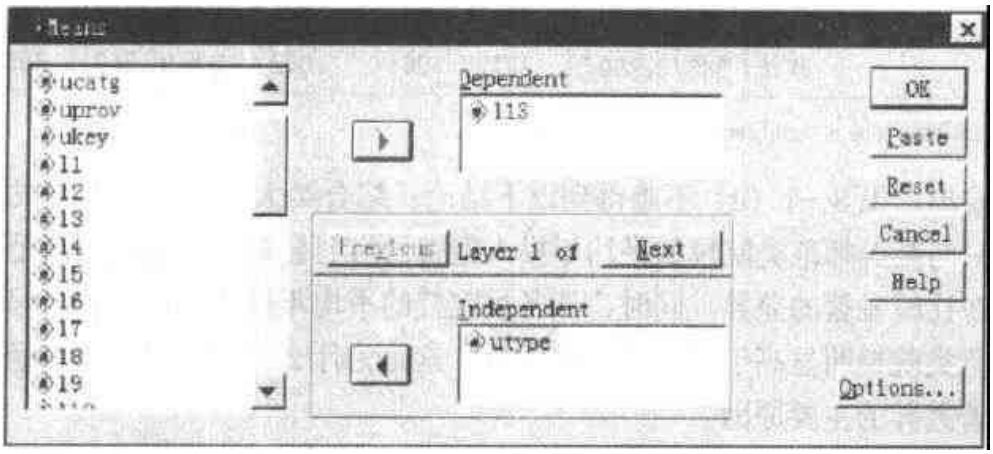


图 8—2　SPSS 的子体均值比较窗口

(2) 指定一个变量为汇总变量到 Dependent list 框中。

(3) 指定一个或多个变量为分组变量到 Independent 框中，并按 Previous 或 Next 按钮指定多个分组变量的层次关系。这里，分组变量指定的先后顺序很关键，它将决定哪个变量是第一分组变量，哪个变量是第二分组变量等。

三、SPSS 均值比较的 Option 选项

在 Option 选项中可以指定计算汇总变量的那些基本统计量。Statistics for

first layer 框中的选项是针对第一分组变量的，具体含义是：

（1）Anova table and eta（η、η^2）：表示输出单因素方差分析表和 eta 值，它们都是关于第一分组变量和汇总变量的。η 值反映了第一分组变量和汇总变量的相关程度，取值在 0 至 1 之间。η^2 是一个比例数，是方差分析表中组间平方和与总平方和的比，表示了第一分组变量（控制变量）对汇总变量（观察变量）变动的解释能力。

（2）Test of linearity：表示作线性度检验。它的限制条件是第一分组变量有三个以上分组水平。该检验的零假设是 H_0：汇总变量是第一分组变量的线性函数。这里，线性度检验实质是对第一分组变量（自变量）和汇总变量（因变量）作一元线性回归分析，给出回归方程显著性检验的 F 统计量值及相应的相伴概率值，并给出判定系数 R^2。与单因素方差分析中多项式检验类似。

四、案例分析说明

对本章案例数据作均值比较分析，其中：第一分组变量为学校类型，汇总变量为研究与发展年投入经费。分析结果如输出结果 8—2 所示。

输出结果 8—2（a）　　案例数据均值比较的分析结果

ANOVA Table

			Sum of Squares	df	Mean Square	F	Sig.
L13 * UTYPE	Between Groups	(Combined)	27293026159.732	7	3899003737.105	7.757	.000
		Linearity	4083525927.868	1	4083525927.868	8.125	.006
		Deviation from Linearity	23209500231.864	6	3868250038.644	7.696	.000
	Within Groups		36188507037.268	72	502618153.295		
	Total		63481533197.000	79			

可见，该分析结果与单因素方差分析中的结果完全一致。

输出结果 8—2（b）中第二列的 R 值是年投入经费与学校类型的简单相关系数，第三列是它们的简单相关系数的平方。从相关系数来看，这两个变量的相关程度比较弱。第四、第五列分别是 η、η^2，表明了经费与学校类型的相关程度、经费投入差异中由学校类型线性解释的比例。从数据来看，学校类型对经费投入

的线性解释能力不高。

输出结果 8—2 (b)　　案例数据子体均值比较的分析结果

Measures of Association

	R	R Squared	Eta	Eta Squared
L13* UTYPE	−.254	.064	.656	.430

§8.5　多因素方差分析和协方差分析

多因素方差分析中的控制变量个数是两个或两个以上，它研究目的是要分析多个控制变量的作用、多个控制变量交互作用、以及其他随机因素作用是否对观察变量的分布产生显著影响。例如：对农作物产量进行分析时，不仅单纯考虑施肥量多少对农作物产量的影响，还要考虑不同品种的种子对产量的影响，同时还要考虑不同施肥量、不同品种的种子的交互作用对产量的影响。

我们用表 8—1 中假设的理想数据来说明交互作用的影响。

表 8—1　　多因素方差分析的举例数据

亩产量	10 公斤化肥	20 公斤化肥	30 公斤化肥
A 品种种子	650、650、650 650、650、650	700、700、700 700、700、700	720、720、720 720、720、720
B 品种种子	750、750、750 750、750、750	950、950、950 950、950、950	712、711、715 709、732、720

从表 8—1 不难看出：如果用 B 种品种的种子，施加 20 公斤的化肥，对提高农作物的产量效果最为明显，而其他品种和施肥量的搭配对产量的影响作用不很显著。

一、多因素方差分析的基本思路

由于多因素方差分析既要分析多个控制变量独立作用对观察变量的影响，又要分析多个控制变量交互作用对观察变量的影响和其他随机因素对观察变量的影响，因此，它将观察变量总的变差平方和分解成由多个控制变量单独作用引起的、由多个控制变量交互影响引起的、由其他随机因素引起的三个部分。以两个

控制变量为例，可以表示成：

$$Q_{总} = Q_{控1} + Q_{控2} + Q_{控1控2} + Q_{随} \tag{8.5}$$

式中：$Q_{控1} + Q_{控2}$称为主效应部分（Main effects），$Q_{控1控2}$称为 N 向交互影响部分（N Way），反映了两个控制变量各水平之间相互搭配时对观察变量的影响。前两部分的和又称为可解释部分（Explained）；$Q_{随}$ 是由其他随机因素引起的部分，也称为剩余部分（Residual）。各部分的计算公式是：

$$Q_{控1} = sl\sum_{i=1}^{r}(\overline{x}_i - \overline{x})^2 \tag{8.6}$$

$$Q_{控2} = rl\sum_{j=1}^{s}(\overline{x}_j - \overline{x})^2 \tag{8.7}$$

$$Q_{控1控2} = l\sum_{i=1}^{r}\sum_{j=1}^{s}(\overline{x}_{ij} - \overline{x}_i - \overline{x}_j + \overline{x})^2 \tag{8.8}$$

$$Q_{随} = \sum_{i=1}^{r}\sum_{j=1}^{s}\sum_{k=1}^{l}(x_{ijk} - \overline{x}_{ij})^2 \tag{8.9}$$

式中：第一个控制变量有 r 个水平，第二个控制变量有 s 个水平，共有 rs 个水平组合，每个水平组合重复 l 次实验，每次实验结果的数据用 x_{ijk}表示。

多因素方差分析仍然采用 F 检验，其零假设是 H_0：多个控制变量的不同交叉水平下，各总体均值无显著差异。F 统计量定义为：

$$F_{控1} = \frac{\dfrac{Q_{控1}}{(r-1)}}{\dfrac{Q_{随}}{rs\ (l-1)}} = \frac{S^2_{控1}}{S^2_{随}} \tag{8.10}$$

$$F_{控2} = \frac{\dfrac{Q_{控2}}{(s-1)}}{\dfrac{Q_{随}}{rs\ (l-1)}} = \frac{S^2_{控2}}{S^2_{随}} \tag{8.11}$$

$$F_{控1控2} = \frac{\dfrac{Q_{控1控2}}{(r-1)\ (s-1)}}{\dfrac{Q_{随}}{rs\ (l-1)}} = \frac{S^2_{控1控2}}{S^2_{随}} \tag{8.12}$$

以上 F 统计量都服从 F 分布。SPSS 将自动计算 F 值，并依据 F 分布表给出相应的相伴概率值。如果 $F_{控1}$的相伴概率值小于或等于用户心中的显著性水平 α，则拒绝 H_0，表示第一个控制变量的不同水平对观察变量均值产生了显著影响；相反，如果 $F_{控1}$的相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，可以认为第一个控制变量的不同水平对观察变量的均值没有产生显著影响。同理，对 $F_{控2}$判断。如果 $F_{控1控2}$的相伴概率值小于或等于用户心中的显著性水

平 α，则拒绝 H_0，表示第一个控制变量和第二控制变量各水平的交互作用对观察变量均值产生了显著影响；相反，如果 $F_{控1控2}$ 的相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，可以认为第一个控制变量和第二控制变量各水平的交互作用对观察变量的均值没有产生显著影响。

二、协方差分析的基本思路

无论单因素方差分析还是多因素方差分析，它们都有一个共同的特点就是控制变量的各个水平是人为可以控制的。但在许多实际问题中，有些因素的不同水平很难人为控制，但它们确确实实对观察变量产生较为显著的影响。在方差分析中，如果忽略这些因素的存在而单纯去分析其他因素对观察变量的影响，往往会夸大或缩小这些因素的影响作用，使得分析结论不准确。

例如：在研究农作物产量问题时，如果仅考察不同施肥量、不同种子对产量的影响，而不去考虑诸如地块对产量的影响．这种分析显然是不全面的。因为有些地块可能有利于农作物的生长，而有些地块可能不利于农作物的生长，即使不同的施肥量、不同的种子对农作物产量没有产生显著影响，用方差分析的方法也很可能得出相反的结论。

再例如：分析不同的饲料对生猪增重是否产生显著差异。如果单纯仅仅分析饲料的作用，而不考虑不同的生猪自身不同的身体条件（如：初始体重不同），那么得出的结论很可能是不准确的，因为体重的变化幅度一定包含了诸如初始体重对体重变化的影响。

因此，为了更加准确地研究控制变量不同水平对观察变量的影响，应尽量排除其他能够排除的因素对分析的影响作用。例如：应尽量排除地块对农作物产量的影响，应尽量排除生猪初始体重对饲养后体重变化幅度的影响等。可以通过设计科学的实验方案，如拉丁方实验方法来实现，也可以利用协方差分析方法来实现。

协方差分析是将那些很难控制的因素作为协变量，在排除协变量影响的条件下，分析控制变量对观察变量的影响，从而更加准确地对控制因素进行评价。

协方差分析的特点体现在：如果对生猪饲养问题用协方差分析方法进行分析，考虑的影响因素应有两类：一类是饲料，它是定性变量，另一类是生猪的初始体重，它是定量变量。方差分析中的影响变量都是定性变量，线性回归分析中的影响变量都是定量变量，而协方差分析中既包含了定性变量，也包含了定量变量，是一种介于方差分析和线性回归分析之间的分析方法。

协方差分析要求协变量应是连续数值型，协变量间没有交换影响，且与控制

变量之间也没有交互影响。

在协方差分析中，将观察变量总的变差平方和分解为由控制变量引起的、由控制变量交互作用引起的、由协变量引起的、由其他随机因素引起的。以单因素协方差分析为例，观察变量总的变差可以分解为：

$$Q_{总} = Q_{协} + Q_{控} + Q_{随} \tag{8.13}$$

也可以理解为：

$$Q_{总} - Q_{协} = Q_{控} + Q_{随} \tag{8.14}$$

即在扣除了协变量对观察变量的影响后分析控制变量对观察变量的影响。具体的扣除方法请参见有关书目。

协方差分析的零假设是 H_0：各交叉分组下各总体均值无显著差异。仍然利用 F 检验方法，F 统计量的计算可以表示为：

$$F_{控} = \frac{S^2_{控}}{S^2_{随}} \tag{8.15}$$

$$F_{协} = \frac{S^2_{协}}{S^2_{随}} \tag{8.16}$$

以上 F 统计量都服从 F 分布。SPSS 将自动计算 F 值，并依据 F 分布表给出相应的相伴概率值。如果 $F_{协}$ 的相伴概率值小于或等于用户心中的显著性水平 α，则应拒绝 H_0，认为协变量对观察变量的影响是显著的；相反，如果 $F_{协}$ 的相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0，可以认为协变量对观察变量的影响是不显著的；如果 $F_{控}$ 的相伴概率值小于或等于用户心中的显著性水平 α，则拒绝 H_0，表示无论协变量是否对观察变量产生显著影响，控制变量的不同水平确实对观察变量的均值产生了显著影响；相反，如果 $F_{控}$ 的相伴概率值大于用户心中的显著性水平 α，则不能拒绝 H_0。如果协变量的相伴概率小于或等于用户心中的显著性水平 α，则可以认为观察变量的变差主要是由有协变量引起的。

三、多因素方差分析和协方差分析的基本操作步骤

多因素方差分析和协方差分析的基本操作步骤是：

（1）菜单选项：Statistics－>General Linear Model－>GLM－General Factorial。于是出现如图 8—3 所示的窗口。

（2）选择观察变量到 Dependent 框中。

（3）选择多个控制变量到 Fixed Factor（s）框中。

（4）如果作协方差分析，选择协变量到 Covariate（s）框中。

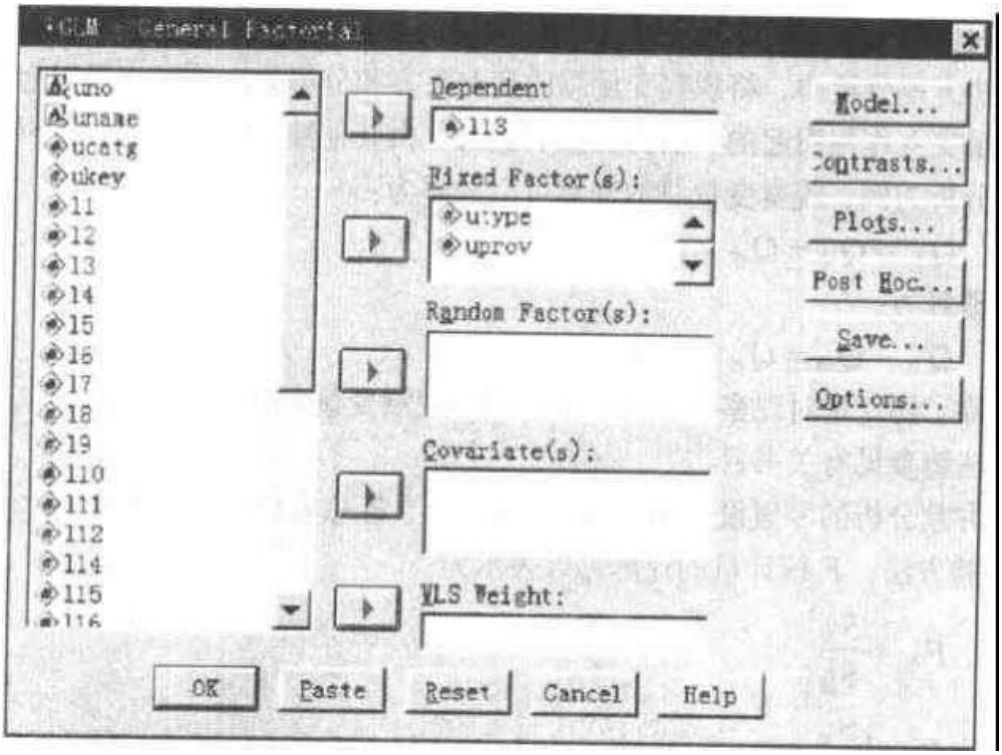


图 8—3　SPSS 多因素方差分析窗口

四、多因素方差分析和协方差分析的其他选项

1. Model 选项

Model 选项允许用户选择建立那种多因素方差分析模型。SPSS 默认建立饱和模型(Full Factorial)。模型中,将观察变量总的变差平方和分解成多个控制变量对观察变量的独立作用部分(主效应部分)、多个控制变量所有交互作用部分(N－Way 影响部分)和随机变量影响部分。例如:如果有 A、B、C 三个控制变量,饱和模型将观察变量总的变差平方和分解成 A、B、C 三个控制变量单独作用部分,A 和 B、A 和 C、B 和 C、A 和 B 和 C 交互影响部分,随机变量部分。其中:A 和 B、A 和 C、B 和 C 的交互作用是二维的,A 和 B 和 C 的交互作用是三维的。

如果在实际分析中,不考虑如此详细,则可以指定按照用户自定义模型(Custom)进行多因素方差分析。例如:仍有 A、B、C 三个控制变量,在用户自定义模型中的 Model 框中选择了 A、B、C、A＊B、A＊C,那么,观察变量总的离差平方和将分解成 A、B、C 三个变量独立作用部分、A 和 B、A 和 C 二维交互影响部分、随机变量影响部分,而 B 和 C、A 和 B 和 C 的交互影响部分一并纳入随机因素影响中。

2. Contrast 选项

如果希望对控制变量各水平上的观察变量的差异进行对比检验,则应该选择

Contrast 选项。其中：

● None：SPSS 默认，不作对比分析。

● Deviation：表示以观察变量的均值为标准，比较各水平上观察变量的均值是否有显著差异。

● Simple：表示以第一水平或最后一个水平上的观察变量均值为标准，比较各水平上观察变量的均值是否有显著差异。

● Difference：表示将各水平上观察变量的均值与其前一个水平上的观察变量均值作比较。

● Helmert：表示将各水平上观察变量的均值与其后一个水平上的观察变量均值作比较。

可见，Contrast 选项允许用户自己定义均值比较的标准。

3．Post Hoc 选项

与单因素方差分析中的功能相同，进行多重比较检验。

4．Plots 选项

以图形的方式展现各控制变量之间是否有交互影响。如果各控制变量之间没有交互作用，则各水平对应的图形近于平行，否则相交。

§8.6 多因素方差分析案例分析

通过两个因素的方差分析，对本章案例，研究不同省市、不同办学类型的院校在研究与发展年投入经费上是否存在显著差异。于是，两个控制变量为省市、办学类型，观察变量为研究与发展年投入经费，建立饱和模型。分析结果如输出结果 8—3 所示。

输出结果 8—3（a）是方差是否相等的结果。可以看出：由于相伴概率值为 0.943，大于显著性水平 0.05，因此，不能拒绝零假设，可以认为两个控制变量各交叉水平下的总体方差相等，符合方差分析的前提要求。

输出结果 8—3（a）　　案例多因素方差分析结果（一）

Levene's Test of Equality of Error Variances

Dependent Variable: 研究与发展当年投入经费

F	Df1	df2	Sig.
.604	43	36	.943

输出结果 8—3（b）是多因素方差分析的主要结果。由于指定建立饱和模型，因此，将观察变量（研究与发展年投入经费）的总变差分解为三大部分：一是各控制变量独立作用造成的变差。不同省市造成的变差平方和为7 465 281 452.404 92，均方为497 685 430.160。不同办学类型造成的变差平方和为23 687 523 390.976 55，均方为3 383 931 912.997，可见，不同办学类型对经费投入的影响要远远大于不同省市的影响。再从它们的 F 值看，分别为0.959和6.518，对应的相伴概率分别为0.514和0.000，进一步证明：不同省市间学校的研究与发展经费没有显著差异，而不同的办学类型的学校，经费差距显著；二是，各控制变量交互作用造成的变差。这里，不同省市和不同办学类型交互作用造成的经费投入变差平方和为8 692 317 245.192 51，均方为413 919 868.819，F 值为0.797，相伴概率为0.705。由于相伴概率大于显著性水平0.05，表明不同省市和不同办学类型的交互作用没有给经费投入造成显著影响。

输出结果 8—3（b）　　案例多因素方差分析结果（二）

Tests of Between－Subjects Effects

Depeodent Variable: 研究与发展当年投入经费

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	44791419800.68810	43	1041660925.597	2.006	.017
Intercept	2913955270.60022	1	2913955270.600	5.613	.023
省市	7465281452.40492	15	497685430.160	0.959	.514
办学类型	23687523390.976 55	7	3383931912.997	6.518	.000
省市＊办学类型	8692317245.19251	21	413919868.819	.797	.705
Error	18690113396.31191	36	519169816.564		
Total	84665724602.00000	80			
Corrected Total	63481533197.00000	79			

第 9 章

SPSS 的相关分析和线性回归分析

§9.1 相关分析

一、什么是相关分析

我们知道在社会生活中存在着许多关系，归结起来大致有两种关系。第一种关系：如销售额与销售量之间的关系，销售额＝销售量×单价。当单价确定后，给出一个确定的销售量，销售额便可以惟一确定。又如：圆的面积和半径之间的关系，圆面积＝3.14×半径×半径。当给定一个半径，便可以惟一确定一个圆面积。可见，这种关系是一种一一对应的确定性关系，即：当一个变量 x 取一定值时，另一变量 y 可以依确定的函数公式取一个确定的值，这种关系称为函数关系。第二种关系：如家庭收入和支出、子女身高和父母身高等。这些事物之间确实存在某种关系，但是这些关系无法像函数关系那样，能够用一个确定的函数公式描述。当一个变量 x 取一定值时，另一变量 y 的值可能有几个。这样，一个变量的值就不能由另一个变量的值惟一确定，这种关系称为统计关系。

事物之间的函数关系比较容易分析和测度。事物之间的统计关系，不像函数关系那样直接，但确实普遍存在，并且有的关系强，有的关系弱，程度各有差

异。如何测度事物间统计关系的强弱程度就是人们关注的问题。

相关分析就是一种测度事物间统计关系强弱的一种手段和工具，旨在衡量事物之间，或称变量之间线性相关程度的强弱。

二、相关分析的基本方法之一：绘制散点图

相关分析的方法较多，比较直接和常用的一种是绘制散点图。通过散点图可以比较直观地看出变量之间的相关程度。

1.SPSS 中散点图的种类

在 SPSS 中提供了四种散点图，分别是简单散点图（Simple）、重叠散点图（Overlay）、矩阵散点图（Matrix）和三维散点图（3－D）。

简单散点图一般是用来显示一对变量之间的散点图。

重叠散点图则是在同一坐标轴上显示多对变量的散点图。

矩阵散点图则是以方矩阵格子的样式，在多个坐标轴上显示多对变量的散点图。其中，行列上的格子数是相等的。正对角线的格子中显示参加绘图的若干个变量的名称。应特别注意每个变量名所在的行和列，它们决定了矩阵散点图各个图形的行列坐标的含义。例如：如果某变量在第三行第三列的格子上，表示第三行格子上的所有图形都以该变量为 y 轴，第三列格子上的所有图形都以该变量为 x 轴。

三维散点图是以立体图的形式展现多对变量的散点图。

2. 绘制散点图的基本操作步骤

（1）菜单选项：Graphs－＞Scatter。于是出现如图 9—1 所示的窗口。

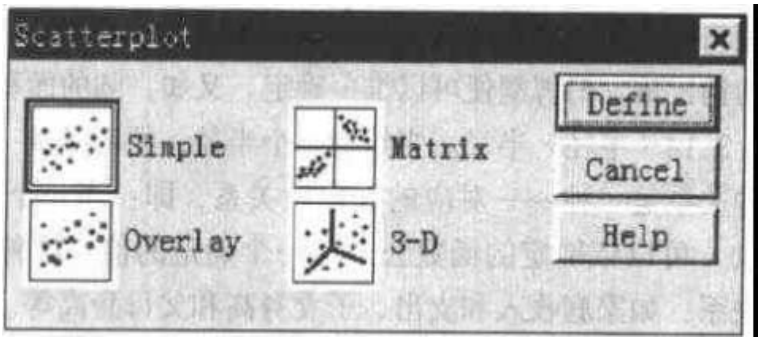


图 9—1　SPSS 制作散点图窗口

（2）选择散点图的类型。

（3）按 Define 钮对不同散点图作具体定义。

不同散点图具体的定义选项稍有差别，我们分别来看。

对于简单散点图主要需要定义：

（1）指定散点图的 y 轴上的变量名到 y Axis 框中。

（2）指定散点图的 x 轴上的变量名到 x Axis 框中。

（3）指定分组变量到 Set Markers 框中。该项可以省略。如果指定了该选项，SPSS 按分组变量的不同取值，将个案分成几组，然后以不同颜色分别绘制散点图。

（4）指定标记变量到 Label Case 框中。该项可以省略。如果指定了该选项，在散点图中的每个点上会标出该个案的相应标记变量的值。

对于重叠散点图主要需要定义：

（1）同时指定两个变量为一对，其中第一个变量的值表示在 y 轴上，另一个表示在 x 轴上。还可以通过按 Swap Pair 按钮调换 y 轴和 x 轴上的变量。因为重叠散点图可以同时展示多对变量，所以这里可以指定多对变量。

（2）指定标记变量到 Label Case 框中。这与制作简单散点图中的相同。

对于矩阵散点图主要需要定义：

（1）定义参加绘图的几个变量到 Matrix 框中。这里应注意：选择变量的先后顺序，决定了矩阵正对角线上变量的排列顺序。

（2）指定分组变量到 Set Markers 框中。这与制作简单散点图中的相同。

（3）指定标记变量到 Label Case 框中。这与制作简单散点图中的相同。

对于三维散点图主要需要定义：

（1）指定散点图的 y 轴上的变量名到 y Axis 框中。

（2）指定散点图的 x 轴上的变量名到 x Axis 框中。

（3）指定散点图的 z 轴上的变量名到 z Axis 框中。

（4）指定分组变量到 Set Markers 框中。这与制作简单散点图中的相同。

（5）指定标记变量到 Label Case 框中。这与制作简单散点图中的相同。

三、相关分析的基本方法之二：计算相关系数

1. 相关系数基本思想

图形虽然能够直观地展现变量之间的相关关系，但不很精确。相关系数能够以数字的方式准确描述变量间的线性相关程度。一般，如果两个变量的相随变动的方向相同，则它们之间的关系称为正线性相关关系；反之，如果两个变量的相随变动的方向相反，则它们之间的关系称为负线性相关关系。

在统计学中，一般用样本相关系数 r 来推断总体相关系数。如果两个变量是正线性相关关系，则 $r>0$；如果两个变量是负线性相关关系，则 $r<0$；如果

两个变量存在一定程度的正线性相关关系，则 $0<r<1$；如果两个变量存在一定程度的负线性相关关系，则 $-1<r<0$。如果两个变量存在完全的正线性相关关系，则 $r=1$；如果两个变量存在完全的负线性相关关系，则 $r=-1$；如果两个变量不存在线性相关关系，则 $r=0$；一般，如果 r 的绝对值大于 0.8，则认为两变量之间具有较强的线性相关关系，如果小于 0.3，则认为两变量之间具有较弱的线性相关关系。

样本相关系数计算出来以后，还有一个非常重要的任务。由于存在样本抽样的随机性、样本数量少等原因，样本相关系数不能直接说明样本来自的两总体是否具有显著相关性，而需要对其进行统计推断。

统计推断，首先总体相关性推断的零假设是 H_0：两总体无显著的线性相关；其次，计算相应的统计量，并得到对应的相伴概率值。如果相伴概率值小于或等于用户指定的显著性水平，则拒绝 H_0，认为两总体存在显著的线性相关关系；如果相伴概率值大于用户指定的显著性水平，则不能拒绝 H_0，可以认为两总体不存在显著的线性相关关系。

对不同类型的变量数据，应采用不同的相关系数来度量，比较常用的主要有 Pearson 简单相关系数、Spearman 和 Kendall tua－b 等级相关系数。

2．Pearson 简单相关系数

Pearson 简单相关系数用来度量定距变量间的线性关系。如：度量国民收入和居民储蓄存款、身高和体重、年龄和收入等变量间的线性相关关系。Pearson 简单相关系数在度量非线性相关关系时，并不是一个好的统计量。它的计算公式为：

$$r=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2\sum_{i=1}^{n}(y_i-\bar{y})^2}} \tag{9.1}$$

SPSS 将自动计算 r 统计量。对 Pearson 简单相关系数的统计检验是计算 t 统计量，公式是：

$$t=\frac{r\sqrt{n-2}}{\sqrt{1-r^2}} \tag{9.2}$$

t 统计量服从 $n-2$ 个自由度的 t 分布。SPSS 会自动计算 t 值，并依据 t 分布表给出其对应的相伴概率值。

3．Spearman 等级相关系数

Spearman 等级相关系数用来度量顺序水准变量间的线性相关关系。它利用

的是非参数检验的分析方法。首先，得到两变量中各数据的排序名次，或称为秩，分别记作 U_i、V_i。然后计算 D_i^2 统计量。计算公式为：

$$\sum_{i=1}^{n} D_i^2 = \sum_{i=1}^{n} (U_i - V_i)^2 \tag{9.3}$$

然后，计算 Spearman 等级相关系数，计算公式为：

$$R = 1 - \frac{6\sum D_i^2}{n(n^2-1)} \tag{9.4}$$

SPSS 将自动计算 R 统计量。可见，Spearman 等级相关系数不是直接通过变量值计算得到的，而利用的是秩。对 Spearman 等级相关系数的统计检验，一般如果个案数 n 小于等于 30，则直接利用 Spearman 等级相关统计量表，SPSS 将自动依据此表给出对应的相伴概率值。如果个案数大于 30，SPSS 将计算 Z 统计量。计算公式是：

$$Z = R\sqrt{n-1} \tag{9.5}$$

Z 统计量近似服从正态分布。SPSS 将依据正态分布表给出对应的相伴概率值。

4. Kendall tua－b 等级相关系数

Kendall tua－b 等级相关也是用来度量顺序水准变量间的线性相关关系的。采用的仍然是非参数检验方法，它利用变量值的秩数据，计算一致对数目（U）和非一致对数目（V）。然后根据下列公式计算 Kendall 相关系数：

$$T = 1 - \frac{4V}{n\ (n-1)} \tag{9.6}$$

SPSS 将自动计算 T 统计量。对 Kendall 等级相关系数的统计检验。一般如果个案数 n 小于等于 30，则直接利用 Kenduall 等级相关统计量表，SPSS 将自动依据此表给出对应的相伴概率值。如果个案数大于 30，SPSS 将计算 Z 统计量。计算公式是：

$$Z = \frac{3T\sqrt{n\ (n-1)}}{\sqrt{2\ (2n+5)}} \tag{9.7}$$

Z 统计量近似服从正态分布。SPSS 将依据正态分布表，给出对应的相伴概率值。

5. 计算各种相关系数的基本操作步骤

计算各种相关系数的基本步骤是：

（1）菜单选项：Statistics－>Correlate－>Bivariate..。于是出现如图 9—2 所示的窗口。

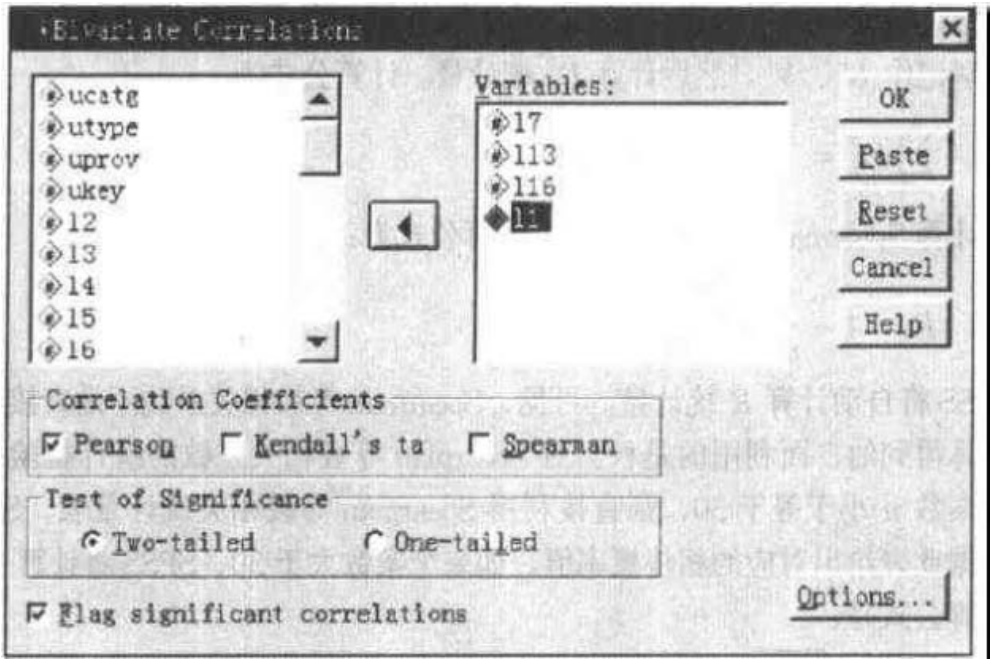


图 9—2　SPSS 相关分析窗口

（2）选择参加计算相关系数的变量到 Variables 框。

（3）在 Correlation Coefficents 框中，选择计算相关系数。

（4）在 Test of Significance 框中，确定在统计检验时作双尾检验（Tow－tailed）还是作单尾检验（One－tailed）。当不明确两变量间的关系是正相关或负相关时，选择作双尾检验。明确两变量间只可能是正相关或负相关时，选择作单尾检验。

（5）Flag of significance 选项，如果选中，表示相关分析结果中，不显示统计检验的相伴概率值，而以星号（＊）表示。一个星号表示：当用户指定的显著性水平为 0.05 时，统计检验的相伴概率值小于等于 0.05，即总体无显著线性相关的可能性小于等于 0.05。两个星表示：当用户指定的显著性水平为 0.01 时，统计检验的相伴概率值小于等于 0.01，即总体无显著线性相关的可能性小于等于 0.01。显然，两个星比一个星的检验更精确。

6. 计算各种相关系数的 Optiont 选项

（1）Statistics 选项：表示当且仅当计算 Pearson 简单相关系数时，选择输出哪些统计量。其中：Means and standard deviations 表示在输出相关系数的同时计算输出各变量的均值和标准差；Cross－product deviations and covariances 表示输出叉积离差和协方差。叉积离差即为 Pearson 简单相关系数公式的分子部分。协方差为：叉积离差/（$n-1$），也反映变量间的相关程度。

(2) Misssing values 选项：对缺失值的处理。Exclude case pairwise，如果正参与计算的两变量中有缺失值，则暂时剔除那些在这两个变量上取缺失值的个案；Exclude cases listwise 为剔除所有具有缺失值的个案后再计算。

§9.2 相关分析案例分析

一、案例说明

查阅《中国普通高等院校人文社会科学统计资料汇编》，得到全国重点院校中的 80 所院校的 1999 年人文社科基本情况数据。指标主要涉及以下几个方面：社科活动人员、研究与发展人员、研究与发展经费、研究与发展课题。由于篇幅所限，具体指标和数据略去。

现对这批数据分析研究与发展课题数与哪些因素有比较密切的关系。

要研究与发展课题与哪些因素有比较密切关系，最简单的方法是绘制散点图，并计算简单相关系数。分析结果如图 9—3 和输出结果 9—1 所示。

从图 9—3 可以看出，研究与发展人员数和研究与发展课题之间存在比较明显的线性相关关系。

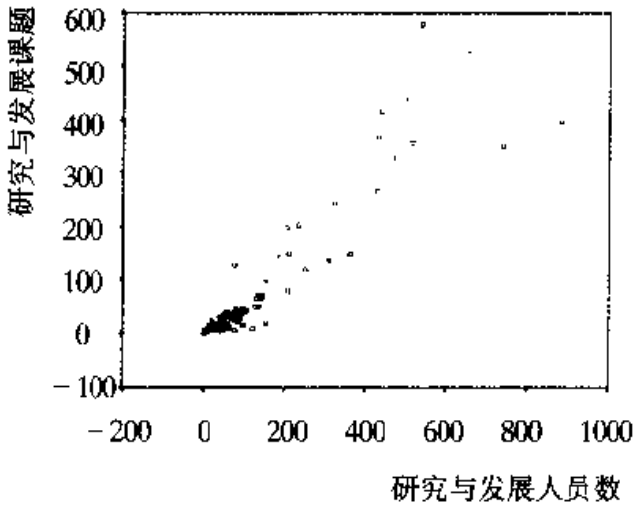


图 9—3 研究与发展人员、研究与发展课题间的散点图

输出结果 9—1 中，每个变量都有三行数据：第一行为该变量与相应列上变量的简单相关系数；第二行为相关系数检验的相伴概率值；第三行为参与计算的观察值个数。

输出结果 9—1　　案例数据简单相关系数矩阵

Correlations

		L1	L7	L13	L16
L1 社科活动人员数	Pearson Correlation	1.000	.889**	.689**	.874**
	Sig. (2-tailed)	.	.000	.000	.000
	N	80	80	80	80
L7 研究与发展人员数	Pearson Correlation	.889**	1.000	.711**	.922**
	Sig. (2-tailed)	.000	.	.000	.000
	N	80	80	80	80
L13 研究与发展年投入经费	Pearson Correlation	.689**	.711**	1.000	.777**
	Sig. (2-tailed)	.000	.000	.	.000
	N	80	80	80	80
L16 研究与发展课题数	Pearson Correlation	.874**	.922**	.777**	1.000
	Sig. (2-tailed)	.000	.000	.000	.
	N	80	80	80	80

** Correlation is significant at the 0.01 level (2-tailed).

从输出结果 9—1 不难看出：研究与发展课题数与研究与发展年投入经费、社科活动人员数、研究与发展人员数的线性相关关系依次增强，简单相关系数依次分别为 0.777、0.874、0.922，且在显著性水平为 0.01 时，都通过统计检验，都呈现出明显的线性相关关系。

§9.3　偏相关分析

一、偏相关分析和偏相关系数

简单相关系数在一些情况下无法较为真实准确地度量事物之间的相关关系，往往有夸大的倾向。例如：在研究商品的需求量和价格、消费者收入之间的关系时会发现，需求量和价格之间的相关关系实际还包含了消费者收入对商品需求量的影响。同时，收入对价格也产生了影响，并通过价格变动传递到对商品需求量的影响中。再如：研究粮食产量与平均气温、月降水量、平均日照时数、月平均湿度之间的关系，产量和平均气温之间的关系中实际还包含了月平均日照时数对产量的影

响。同时，平均日照时数对平均气温也产生了影响等等。在这种情况下，单纯计算简单相关系数，显然不能准确地反映事物之间的相关关系，而需要在剔除其他相关因素的影响的条件下计算相关系数。偏相关分析正是实现这个目的的。

偏相关分析是在控制了其他变量对待分析的两个变量影响的条件下，分析两变量的线性相关关系。在 SPSS 中，那些被控制的因素称为控制变量。

偏相关分析的工具是计算偏相关系数。设有 x_1、x_2 和 y 三个变量。如果计算 x_1 与 y 之间的偏相关系数，则应将 x_2 作为控制因素加以控制。那么，相应的偏相关系数的计算公式是：

$$r_{y1,2}=\frac{r_{y1}-r_{y2}r_{12}}{\sqrt{(1-r_{y2}^2)\ (1-r_{12}^2)}} \tag{9.8}$$

式中：$r_{y1,2}$表示计算 y 和 x_1 的偏相关系数，x_2 为控制变量。r_{y1}表示 y 和 x_1 的简单相关系数。r_{y2}表示 y 和 x_2 的简单相关系数。r_{12}表示 x_1 和 x_2 的简单相关系数。

二、计算偏相关系数

计算偏相关系数的基本操作步骤是：

（1）菜单选项：Statistics－＞Correlate－＞Partial。于是出现如图 9—4 所示的窗口。

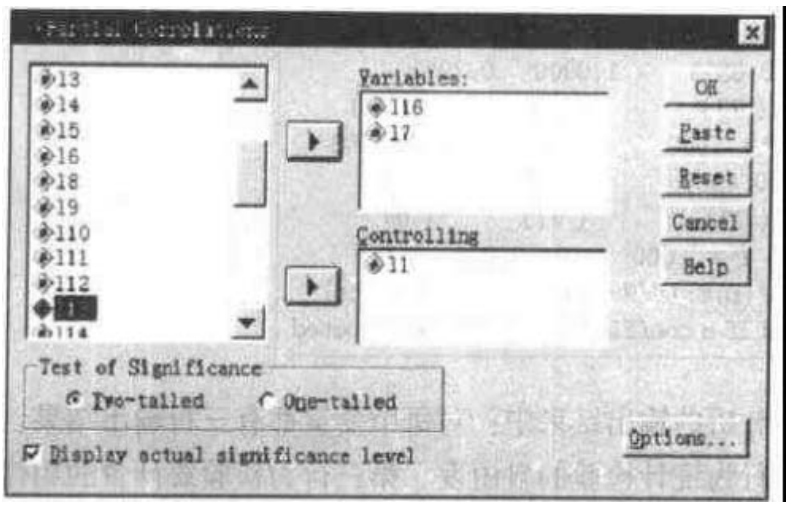


图 9—4　SPSS 计算偏相关系数窗口

（2）选择那些变量参加计算的变量到 Variable 框中。

（3）选择一个变量作控制变量到 Controlling for 框中。

（4）在 Test of Significance 框中，确定在统计检验时作双尾检验（tow－

Tailed）还是作单尾检验（One－tailed）。

（5）Display actual significance level：如果选中该项表示输出结果中显示具体的相伴概率值。否则不显示并以星号作标记。星号的意义与计算相关系数中的相同。

计算偏相关系数的 Option 选项：

其中的大部分选项与计算相关系数中的含义相同。Zero－order Correlations 表示在输出偏相关系数的同时还输出变量间的简单相关系数。

三、案例说明分析

在第二节计算简单相关系数时，我们看到研究与发展课题数与研究与发展人员数有最高的相关关系，那么，这个相关关系中是否包含了社科活动人员数以及研究与发展年投入经费对它们的影响呢？下面计算课题数与研发人员数的偏相关关系，以社科活动人员、年投入经费作为控制变量。结果如输出结果 9—2 所示。

输出结果 9—2　　案例的偏相关分析结果

```
- - PARTIAL CORRELATION COEFFICIENTS - - -
Controlling for..      L1
              L16          L7          L13
L16         1.0000       0.6519      0.4965
             (0)         (77)        ( 77)
            P = .        P = 0.000   P = 0.000
L7           0.6519      1.0000      0.2951
             ( 77)        ( 0)        ( 77)
            P = .000      P = .      P = 0.008
L13          0.4965      0.2951      1.0000
             ( 77)        ( 77)       ( 0)
              P = 0.000   P = 0.008   P = .
(Coefficient /(D.F.) /2 - tailed Significance)
"." is printed if a coefficient cannot be computed
```

在偏相关分析的输出结果中：对每个变量都有三行输出结果。第一行为偏相关系数、第二行为统计检验的自由度、第三行为检验统计量的相伴概率值。

从输出结果 9—2 可知：偏相关系数的计算是在社科活动人员为控制变量的条件下进行的。研究与发展课题数和研究与发展人员数的偏相关系数由原来的简单相关系数 0.922 变为 0.651 9；研究与发展课题数和研究与发展经费之间的偏相关系数由原来的简单相关系数 0.777 变为现在的 0.496 5。由此可见，社科活动人员数对研究与发展人员、研究与发展经费以及课题数的影响。从现实情况分

析，也的确如此。

§9.4 线性回归分析

一、线性回归分析的基本实现思路

线型回归分析是侧重考察变量之间的数量变化规律，并通过一定的数学表达式，即回归方程，来描述这种关系，进而确定一个或几个变量的变化对另一个变量的影响程度，为预测提供科学的数学依据。

一般线性回归分析的基本步骤是：

（1）确定回归方程中的自变量和因变量。

（2）从收集到的样本数据出发确定自变量和因变量之间的数学关系式，即建立回归方程。

（3）对回归方程进行各种统计检验。

（4）利用回归方程进行预测。

利用 SPSS 进行回归分析时，这四个基本步骤中的第一步由用户给定的。第二步和第三步是由 SPSS 自动完成的。第四步的预测工作，用户可以利用第五章中讲到的 Compute 命令，在相应的算术表达式框中输入回归方程公式，SPSS 将依据公式自动计算出预测结果。

二、线性回归方程

线性回归方程从样本数据出发，一般利用最小二乘法，本着回归直线与样本数据点在垂直方向上的偏离程度最低的原则，进行回归方程的参数求解。

根据线性回归方程中自变量的个数，可以将回归方程分成一元线性回归方程和多元线性回归方程。

1. 一元线性回归方程

一元线性回归方程中的自变量只有一个。在现实社会经济现象中，某一事物（因变量）总会受到多方面因素（多个自变量）的影响。一元线性回归分析是在不考虑其他影响因素的条件下，或是在认为其他影响因素确定的情况下，分析某一个因素（自变量）是如何影响另一事物（因变量）的，所进行的分析是比较理想化的。从几何意义上讲，一元回归方程是二维平面上的一条直线。

一元线性回归的经验模型是：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x \tag{9.9}$$

式中：$\hat{\beta}_0$ 表示回归直线在纵轴上的截距。$\hat{\beta}_1$ 是回归系数，表示当自变量 x 变动一个单位所引起的因变量 $\hat{y}$ 的平均变动单位。

2. 多元线性回归方程

多元线性回归方程中的自变量的个数有 2 个以上。

多元线性回归方程的经验模型是：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k \tag{9.10}$$

式中：假设该线性方程中有 k 个自变量。$\hat{\beta}_1$，$\hat{\beta}_2$，…，$\hat{\beta}_k$ 是方程的偏回归系数。$\hat{\beta}_1$ 表示在其他自变量保持不变的情况下，自变量 x_1 变动一个单位所引起的因变量 $\hat{y}$ 的平均变动单位。其他偏回归系数的含义类似。从几何意义上讲，多元回归方程是多维空间上的一个平面。

三、线性回归方程的统计检验

通过样本数据建立一个回归方程后，不能立即就用于对某个实际问题的预测。因为，这个模型的参数是样本数据估计拟合出来的经验方程，还需要对它作各种统计检验。一般经常作以下的统计检验：

1. 回归方程的拟合优度检验

回归方程的拟合优度检验就是要检验样本数据点聚集在回归直线周围的密集程度，从而评价回归方程对样本数据的代表程度。

回归方程的拟合优度检验一般用判定系数 R^2 实现。

我们知道，因变量的各个观察值之间的差异主要是由两个方面原因造成的：一是由于自变量的不同取值造成的；二是由于其他因素造成的。

例如：在对不同人的体重和身高进行研究时发现，人们体重的差异受两类因素的影响。一类：不同的身高会使体重产生差异；另一类：即使身高相同，体重也会因为不同的生活环境、条件等因素的影响而产生差异。将这个问题拿到回归分析中讲就是，因变量（体重）的差异一方面是由自变量（身高）的不同造成的，另一方面是由其他随机因素影响造成的。那么，接下来，人们希望研究体重的差异究竟主要是由哪个方面的影响造成的。

为解决这个问题，一般将因变量总的变差分解成自变量变动造成的变差和其他因素变动造成的变差。数学上可以证明：

因变量总的变差平方和＝由自变量引起的因变量变差的平方和＋由其他随机因素引起的因变量变差的平方和

用数学语言可以表示为：

$$\sum_{i=1}^{n}(y_i-\bar{y})^2=\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2+\sum_{i=1}^{n}(y_i-\hat{y})^2 \tag{9.11}$$

式中：由自变量引起的变差平方和能够由回归方程表示，称为回归平方和，由其他随机因素引起的变差平方和无法用回归方程表示，称为残差平方和或剩余平方和。

判定系数 R^2 定义为：

$$R^2=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}=1-\frac{\sum_{i=1}^{n}(y_i-\hat{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \tag{9.12}$$

可见，判定系数 R^2 等于回归平方和与总平方和之比，体现了回归方程所能解释的因变量变差的比例。于是，$1-R^2$ 则体现了因变量总变差中，回归方程所无法解释的比例。R^2 越接近于 1，则说明回归平方和占了因变量总变差平方和的绝大部分比例，因变量的变差主要由自变量的不同取值造成，回归方程对样本数据点拟合得好；反之，R^2 越接近于 0，则说明回归平方和仅占了因变量总变差平方和的一小部分比例，因变量的变差主要由随机因素造成，回归方程对样本数据点拟合得较差。

判定系数在多元线性模型中是因变量 y 关于自变量 x_i 的复相关系数 R 的平方，实质上测定了因变量 y 与所有自变量全体之间线性相关程度，也测定了实际样本数据与预测数据间的相关程度。因此，从这个意义上讲，判定系数能够比较好地反映回归直线对样本数据的代表程度。

另外，在实际回归分析中可以看到，随着自变量个数的不断增多，它们会给回归模型增添越来越多的新元素，从而必然使得 R^2 不断增加。于是出现的问题就是，当 R^2 增加时，我们无法确切地知道 R^2 的增加究竟是由于 R^2 的数学习性决定的，还是确实是由于引入了好的自变量进入方程而造成的。为了解决这个问题，在作拟合优度的判定时，一般采用修正或调整的 R^2。调整的 R^2 定义为：

$$\overline{R}^2=1-\frac{n-1}{n-k-1}\cdot\frac{SSE}{SST} \tag{9.13}$$

式中：SSE 和 SST 分别表示剩余平方和、因变量的总平方和。调整的 R^2 也等价于：

$$\overline{R}^2=1-\frac{\text{均方误差}}{\text{因变量的样本方差}} \tag{9.14}$$

于是，在某个自变量引入回归方程后，如果该自变量是理想的且对因变量变差的解释说明是有意义的，那么必然使得均方误差减少，从而使调整的 R^2 得到

提高；反之，如果某个自变量对因变量的解释说明没有意义，那么引入它不会造成均方误差减少，从而调整的 R^2 也不会提高。由此可见，调整的 R^2 更能够准确地反映回归方程对样本数据点的拟合程度。它的改进之处在于：调整的 R^2 考虑的是平均的剩余平方和，而不是剩余平方和。因此，一般在线性回归分析中，调整的 R^2 越大越好。

2. 回归方程的显著性检验

回归方程的显著性检验是检验因变量与所有自变量之间的线性关系是否显著，是否可以用线性模型来描述因变量和自变量之间的关系。

回归方程的显著性检验的零假设是 H_0：$\beta_1=\beta_2=\cdots=\beta_k=0$

即检验所有回归系数是否同时与零无显著差异，从总体上对回归方程进行检验，验证因变量和所有自变量之间是否存在显著的线性关系，自变量的变化是否可以反映因变量的线性变化。

回归方程的显著性检验一般采用 F 检验，利用方差分析的方法进行。F 统计量定义成：平均的回归平方和与平均的剩余平方和（均方误差）之比，数学公式为：

$$F=\frac{\sum(\hat{y}_i-\bar{y})^2/k}{\sum(y_i-\hat{y}_i)^2/(n-k-1)} \tag{9.15}$$

式中：n 为样本数，k 为自变量个数。F 服从（k，$n-k-1$）个自由度的 F 分布。

可以直观理解：如果 F 值较大，则说明自变量造成的因变量的变动远远大于随机因素对因变量造成的影响；反之，如果 F 值较小，则说明自变量造成的因变量的变动远远小于随机因素对因变量造成的影响。由于 F 统计量服从 F 分布，因此，SPSS 在自动计算 F 值以后，会依据 F 分布表给出统计量对应的相伴概率值。如果相伴概率值小于或等于用户给定的显著性水平，则拒绝 H_0，认为所有回归系数同时与零有显著差异，自变量与因变量之间存在显著的线性关系，自变量的变化能够反映因变量的线性变化，回归方程显著；相反，如果相伴概率大于用户给定的显著性水平，则不能拒绝 H_0，可以认为所有回归系数同时与零无显著差异，自变量与因变量之间不存在显著的线性关系，自变量的变化无法反映因变量的线性变化，回归方程不显著。

另外，从另一个角度讲，F 统计量也可以反映回归方程的拟合优度。我们将 F 统计量的公式与 R^2 的公式作一结合转换，就有：

$$F=\frac{R^2/k}{(1-R^2)/(n-k-1)} \tag{9.16}$$

可见，如果回归方程的拟合优度高，F 统计量就越显著。F 统计量越显著，回归方程的拟合优度就会越高。

3. 回归系数显著性检验

回归方程显著性检验只能检验所有回归系数是否同时与零有显著性差异。如果检验出所有回归系数不同时为零，仍然不能保证方程中仍存在与零无显著差异的回归系数。也就是说，经过回归方程的显著性检验，仍然不能保证回归方程中不包含不能较好解释说明因变量变化的自变量。因此，可以通过回归系数显著性检验对每个回归系数进行考察。

回归系数显著性检验的零假设是 H_0：$\beta_i=0$。

即：分别检验每个回归系数是否与零无显著差异，从而检验每个自变量对因变量的线性影响是否显著。

回归系数显著性检验一般采用 t 检验的方法，数学公式为：

$$t_i=\frac{\beta_i}{S_{\beta i}} \tag{9.17}$$

式中：$S_{\beta i}$ 是回归系数 β_i 的标准误差（SE.B），它的计算公式是：

$$S_{\beta i}=\sqrt{\frac{S_y^2}{\sum(x_i-\overline{x}_i)^2}} \tag{9.18}$$

式中，S_y 是回归方程标准误差（Standard Error）的估计值，由前面提到的均方误差开方后得到，反映了回归方程无法解释说明样本数据点的程度或对样本数据点的偏差程度。

可见，如果某个自变量 x_i 的回归系数标准误差较大，必然得到一个相对较小的 t 值，表明该自变量 x_i 解释说明因变量变化的能力较差。因此，当某个自变量 x_i 的 t 值小到一定程度时，该自变量 x_i 就不应保留在回归方程中。由于 t 值服从 $n-2$ 个自由度的 t 分布，SPSS 在自动计算每个回归系数的 t 统计量之后，将依据该分布表给出统计量对应的相伴概率值。如果相伴概率值小于或等于用户给定的显著性水平，则应拒绝 H_0，认为该回归系数与零有显著差异，该自变量与因变量之间存在显著的线性关系，它的变动能够较好地解释说明因变量的变动，应保留在回归方程中；反之，如果相伴概率值大于用户给出的显著性水平，则不能拒绝 H_0，可以认为该回归系数与零无显著差异，该自变量与因变量之间不存在显著的线性关系，它的变动无法较好地解释说明因变量的变动，应剔除出回归方程。

在一元线性回归分析中，回归方程的显著性检验可以替代回归系数的显著性检验，并且有这样的关系存在：

$$F = t^2$$

在多元线性回归分析中，这两种统计检验不能互相替代，但又存在一定的联系。我们知道，如果某个自变量 x_i 引入回归方程且通过了相应的回归系数显著性检验，那么必然造成均方误差减小，R^2 有所提高，F 值有所改善，且 F 值的改善量可以用以下公式测定：

$$F_{ch} = \frac{R_{ch}^2 \ (n-k-1)}{1-R^2} \tag{9.19}$$

式中：R_{ch}^2表示 R^2 的改进量。

不难发现：某个自变量 x_i 引入回归方程后，F 统计量的改变量F_{ch}与该自变量的 t 检验统计量之间存在如下的数量关系：$F_{ch} = t_i^2$

因此，从 F_{ch}角度上讲，如果由于某个自变量 x_i 的引入，使得 F_{ch}是显著的(通过观察 F_{ch}的相伴概率值)，那么我们就可以认为该自变量对方程的贡献是显著的，它应保留在回归方程中，从而起到与回归系数 t 检验同等的作用。

由此又涉及到其他两个概念：部分相关系数和偏相关系数。

让我们先来看一下 R_{ch}^2的数学定义：

$$R_{ch}^2 = R^2 - R_i^2 \tag{9.20}$$

式中：R^2 是回归方程当前的判定系数，R_i^2 是如果将某个自变量 x_i 剔除出回归方程后的判定系数。可见，R_{ch}^2是 R^2 的一个变化量，是由于将某个自变量 x_i 引入回归方程造成的，这个变化量的大小在一定程度上取决于自变量 x_i 和因变量的关系密切程度。

不难发现，R_{ch}^2开方后得到的 R_{ch}正是自变量 x_i 与因变量的部分相关系数(part Correlation)。部分相关系数描述了在排除了方程中其他自变量对某个自变量 x_i 的影响之后，因变量和该自变量 x_i 的相关程度。一般认为，如果某自变量 x_i 与因变量的部分相关系数较大，那么它进入回归方程将有助于较大程度地提高 R^2 (R_{ch}^2)。相反，如果某自变量 x_i 与因变量的部分相关系数较小，那么它的进入将对 R^2 (R_{ch}^2) 的提高没有明显作用。

但是，一般我们不能直接依此就下结论说某自变量 x_i 对回归方程是否有重要贡献。原因在于，如果单纯从数量上看 R_{ch}^2，即使一个较小 R_{ch}^2也可能对方程作出了较大贡献。这是因为，有可能因变量的绝大部分信息已被方程中的其他自变量解释完了，而剩余的相对较少信息中的绝大部分信息却仅能被某个自变量 x_i 所解释，那么当方程引入 x_i 后，虽然 R_{ch}^2变化不大，但自变量 x_i 对方程的贡献却仍然是显著的。因此，R_{ch}^2并没有反映出某自变量 x_i 究竟解释了多少方程中

其他自变量所无法解释的因变量信息的比例，但这一点恰恰是评价自变量贡献大小的准确标准。由此可见，部分相关系数不是一个较全面的衡量自变量重要性的统计量，因为它没有排除方程中其他自变量对因变量的影响因素。因此，往往采用以下的一个比例量来衡量自变量的价值：

$$P^2_{\text{partial}}=\frac{R^2-R_i^2}{1-R_i^2} \tag{9.21}$$

式中：分子就是 R^2_{ch}。分母是如果从当前方程中删去某个自变量 x_i 后，回归方程中其他自变量所无法解释的因变量变差的份额。

因此，如果 P_{partial}较大，则说明自变量 x_i 解释了剩余因变量变差中的较大部分份额，它对因变量的解释能力较强。可见，P_{partial}是一个较为理想的衡量自变量价值的指标，而这个指标开方后恰恰为自变量 x_i 与因变量的偏相关系数。

偏相关系数描述了在排除方程中其他自变量对某个自变量 x_i 和因变量的影响之后，因变量和该自变量 x_i 的相关程度，反映了自变量 x_i 与因变量的净相关程度。因此，在决定引入哪个自变量进入回归方程时，应以偏相关系数的大小作为判定依据。

4. 回归方程的残差分析

由于线性回归模型的前提要求残差项应服从方差相等的正态分布，且残差项间互相独立。那么这些前提要求是否成立呢？这就需要通过回归方程的残差分析来进行验证。

对残差序列作处理后，SPSS 可以直接得到标准残差序列。标准化残差便于进行样本奇异值分析，如果残差序列服从正态分布，那么标准化残差序列则应服从标准正态分布。于是就可以依据一些我们较为熟悉的数据作判断。如：大约 99.7% 的标准化残差应落在 −3 至 +3 之间；大约 95% 的标准化残差应落在 −2 至 +2 之间；大约 68% 的标准化残差应落在 −1 至 +1 之间。

回归方程的残差分析主要有：

(1) 残差序列的正态性分析：可以绘制标准化残差序列的带正态曲线的直方图或累计概率图来分析。通过图形的直观观察，可以确定残差序列是否接近正态。

(2) 残差序列的随机性分析：可以绘制残差序列和对应的预测值序列的散点图。如果残差序列是随机的，那么残差序列应与预测值序列无关，残差序列点将随机地分布在经过零的一条直线上下。

(3) 残差序列的独立性分析：分析残差序列是否存在后期值与前期值相关的现象。如果存在相关现象，表示残差序列中还存有一些规律性，回归方程没能较

全面地反映因变量的变化。一般可以利用 $D-W$（Durbin－Watson）检验作残差序列的独立性分析。SPSS 将自动计算 $D-W$ 值。粗略地讲，如果 $D-W=0$，表示残差序列存在完全正自相关；如果 $D-W=4$，表示残差序列存在完全负自相关；如果 $0<D-W<2$，表示残差序列存在某种程度的正自相关；如果 $2<D-W<4$，表示残差序列存在某种程度的负自相关；如果 $D-W=2$，表示残差序列不存在自相关。在实际应用中，如果 $D-W$ 值接近 2 就基本可以认定残差序列具有独立性。

（4）通过对残差序列分析实现样本奇异值（Casewise 或 Outliers）的诊断：奇异值是样本数据中那些远离均值的样本数据点，它们会对回归方程的拟合产生较大偏差影响，应尽量找出它们并加以排除。通过对标准化残差序列的观察，不仅可以知道哪些样本观察值比预测值是大或小，而且还能够知道在绝对值上它比大多数的残差是大还是小。一般认为，如果某样本点对应的标准化残差值超出了 $-3\sim+3$ 的范围，就可以判定该样本数据为奇异值。对奇异值的判定还有许多其他方法。

（5）通过对残差序列分析实现异方差诊断：线性回归模型要求残差序列服从等方差的正态分布，也就是说，在自变量的不同取值点上，残差序列的分布形态不应随自变量或因变量的变化而变化，否则，就认为出现了异方差现象。可以通过绘制学生化残差序列和因变量预测值（或每个自变量）的散点图的方法来诊断异方差现象，或者可以计算学生化残差序列与因变量预测值（或每个自变量）间的相关系数。如果异方差现象确实存在，可以先对因变量实施方差稳定变换后再拟合回归方程。一般，如果残差序列与预测值的平方根成正比例变化，可以对因变量作开方处理；如果残差序列与预测值成比例变化，可以对因变量取对数；如果残差序列预测值的平方成比例的变化，可以对因变量求倒数。另外，对于异方差的情况还可以用加权最小二乘法来代替普通最小二乘法估计回归模型参数。可以根据求式（9.22）的最小的原则进行参数估计：

$$\sum \omega_i \lambda_i^2 = \sum \omega_i \ (y_i - \beta_0 - \beta_1 x_i)^2 \tag{9.22}$$

式中：ω_i 是权数，与残差序列的方差成反比。可见，在加权最小二乘法中如何确定权重变量值是关键。

四、多元线性回归分析中的自变量筛选问题

在多元线性回归分析中，自变量引入的多少是很有讲究的。如果引入的自变量较少，回归方程将无法很好地解释说明因变量的变化。但也并非引入的自变量越多越好，因为这些自变量之间可能存在多重共线性。所谓多重共线性是指自变

量之间存在线性相关关系的现象。自变量间高度的多重共线性会给回归方程带来许多影响。如：会给各个自变量的偏回归系数估计带来困难，偏回归系数的估计方差会随着自变量相关性的增大而不断增大，从而使偏回归系数的置信区间不断增大，偏回归系数估计值的不稳定性不断增强，偏回归系数假设检验的结果不显著等等。因此，有必要对自变量引入回归方程采取一些策略加以控制和筛选。

1. 多元线性回归分析中的自变量筛选方法

多元回归分析中，自变量的筛选一般有三种基本的策略和方法。

（1）向前筛选法（Forward）。该法是变量不断进入回归方程的过程。首先选择与因变量具有最高相关系数的自变量进入方程，并对它进行回归系数显著性检验。然后，在剩余的变量中寻找与因变量偏相关系数最高并通过检验的自变量进入回归方程，并对方程中所有的自变量进行显著性检验。这样一直下去，直到再也没有可进入方程的变量为止。

（2）向后筛选法（Backward）。该法是变量不断剔除回归方程的过程。首先将所有变量全部引入回归方程。然后，进行回归系数显著性检验，在一个或多个 t 检验值不显著的变量中，将 t 值最小的那个变量剔除，然后再重新拟合回归方程，并进行各种检验。如果新方程中所有变量的回归系数的 t 值都是显著的，则变量筛选过程结束。否则，按照上述方法再剔除最不显著的一个自变量，直到再也没有自变量可剔除为止。

（3）逐步筛选法（Stepwise）。该法是向前筛选法和向后筛选法的综合。由于向前筛选法是自变量不断进入回归方程的过程，变量一旦进入回归方程就不会再被剔除出去。但是应注意到，随着自变量的逐个引进，由于自变量之间总存在一定程度的相关性（多重共线性），使得某些已经进入回归方程的自变量的回归系数不再显著，这样造成最终的回归方程可能包含一些不显著的自变量。逐步筛选法是在向前筛选法的基础之上，结合向后筛选法，在每个自变量进入方程后，都判断是否存在应剔除出方程的自变量。如果有则将其剔出。因此，逐步筛选法在选择变量的每一个阶段，都考虑了剔除一个不显著自变量的可能。

2. 自变量间多重共线性的测度

（1）自变量的容忍度。在多元回归分析中，可以用变量的容忍度来度量自变量间多重共线性。某自变量 x_i 的容忍度定义为：$1-R_i^2$，它反映了：如果从当前方程中删去某个自变量 x_i 后，回归方程中其他自变量所无法解释的因变量变差的份额，其取值范围在 0～1 之间。如果某自变量 x_i 对因变量变差有较强的解释贡献能力，即：某自变量 x_i 解释了因变量变差的较大部分份额，那么剔除它之后必然会得到一个相对较大的容忍度。

从另外一个角度讲，R_i^2 也是自变量 x_i 与方程中其他自变量间的复相关系数的平方。如果某自变量 x_i 与方程中的其他自变量有较小的复相关系数，那么其容忍度必然较大。反之，如果某自变量 x_i 与方程中的其他自变量有较大的复相关系数，那么其容忍度必然较小。

可见，如果某自变量 x_i 的容忍度较大，一方面说明该自变量 x_i 对因变量变差的解释贡献能力较强，另一方面也说明该自变量 x_i 与方程中其他自变量的相关性较弱，即：多重共线性较低，应将其引入回归方程。反之，如果某自变量 x_i 的容忍度较小，一方面说明该自变量 x_i 对因变量变差的解释贡献能力较弱，另一方面也说明该自变量 x_i 与方程中其他自变量的相关性较强，即多重共线性较高，可能无需引入回归方程。实际上，这两方面在本质上是一致的。

在 SPSS 中，对自变量间是否存在较强多重共线性，要求不很严格，容忍度太小的自变量也可以进入方程，不过会给出相应的警告信息。

容忍度的倒数称为方差膨胀因子（VIF）。

（2）用特征根刻画自变量的方差。诊断自变量间是否存在较强多重共线性的另一种方法是利用主成分分析方法。其基本思想是：如果自变量间确实存在较强的相关关系，那么它们之间必然存在信息重叠，于是就可以从这些自变量中提取出既能反映自变量信息（方差）、且又相互独立的因素（成分）来。该方法从自变量间的相关系数矩阵出发，计算相关系数矩阵的特征根，得到相应的若干成分。有最大特征根值的成分能够解释说明的自变量的方差比例是最高的，其他特征根随其特征值的减小，对自变量方差的解释能力依次减弱。如果特征根中，有一个特征根的值远远大于其他特征根的值，则说明：这些自变量间一定具有相当多的重叠信息，因为仅一个特征根就能够基本刻画所有自变量的绝大部分信息（方差）。

设方程中所有变量的总方差为 1。每个特征根都能够刻画某自变量信息（方差）的一定比例，所有特征根将刻画某自变量信息（方差）的全部。如果某个特征根既能够刻画某个自变量方差的较大部分比例（如大于 0.7），同时又可以刻画另一个自变量方差的较大部分比例，则表明这两个自变量间存在较强的多重共线性。

通过特征根还可以计算条件指标。

五、SPSS 线性回归分析的基本操作步骤

（1）菜单选项：Statistics－＞Regression－＞Linear。于是出现如图 9—5 所示的窗口。

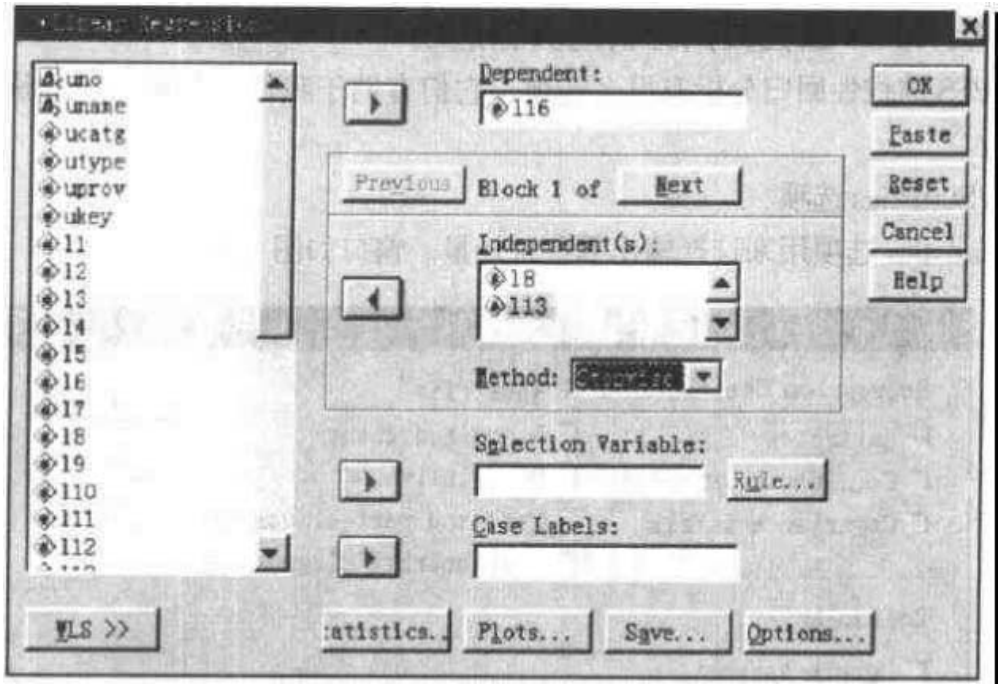


图 9—5　SPSS 线性回归分析窗口

(2) 选择一个变量作为因变量进入 Dependent 框。

(3) 选择一个或多个变量作为自变量进入 Independent 框。

(4) 在 Method 框中选择多元线性回归分析的自变量筛选方法：Enter 表示所选变量全部进入回归方程，它是 SPSS 默认的方法；Remove 表示从回归方程中剔除变量；Stepwise 表示逐步筛选法；Backward 表示向后筛选法；Forward 表示向前筛选法。

(5) 第三步和第四步中选定的自变量及变量筛选方法可以放置在不同的块(Block) 中。一般默认在回归分析中只有一组可进入方程的自变量和相应的筛选方法，但如果希望有多组自变量和对应的多种不同的变量筛选方法，可以按 Next 和 Previous 钮将它们放置在不同的块中。这样，SPSS 将从当前块(默认为 1) 开始，提取自变量组和相应的变量筛选方法对回归方程进行拟合。然后，再自动提取下一块中的自变量组和相应的变量筛选方法，在上个回归方程的基础之上作进一步的拟合工作，直到结束。因此，第四步中的 Remove 方法只可能在第二个以后的块选择。块设置方便了我们作各种探索性的回归分析。

(6) 在 Selection Variable 框中对样本数据进行筛选，即挑选满足一定条件的样本数据进行线性回归分析。

(7) 在 Case Labels 框中指定作图时，以哪个变量作各样本数据点的标志变量。

六、SPSS 线性回归分析的其他选项

SPSS 的线性回归分析有很多选项，它们有助于我们作各种各样的统计检验和分析。

1. Statistics 选项

Statistics 选项用来选择输出哪些统计量。窗口如图 9—6 所示。

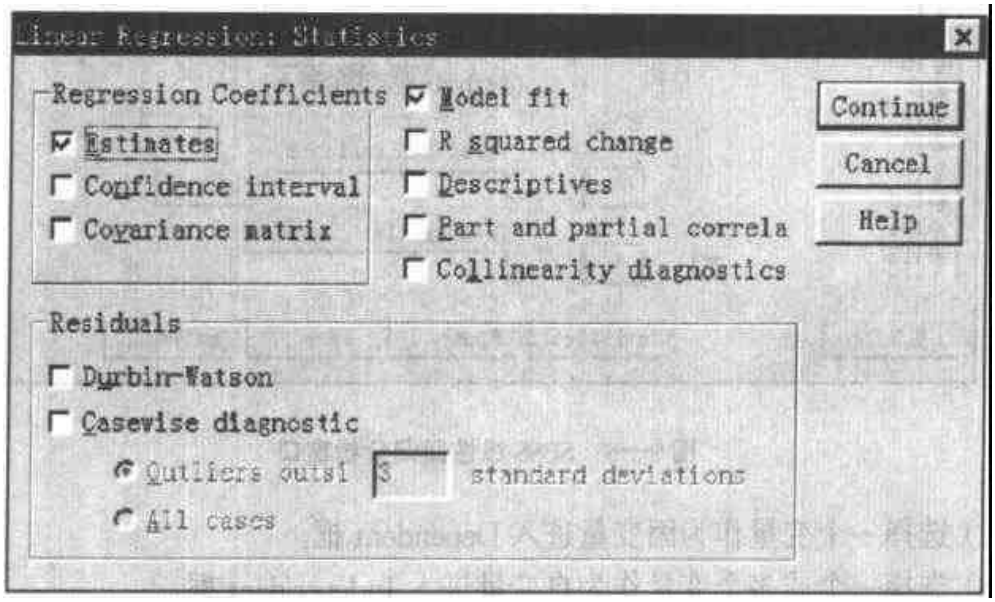


图 9—6　SPSS 线性回归分析的 Statistic 选项窗口

（1）Estimates：SPSS 默认输出项。显示与回归系数相关统计量。如：回归系数、回归系数标准误差、标准化回归系数、t 统计量和相应的相伴概率值、各自变量的容忍度。其中：标准化回归系数有助于判断多元回归方程中各自变量的单位不统一时，哪个自变量对因变量的影响较大。

（2）Confidence Intervals：显示每个非标准化回归系数的 95% 置信区间。

（3）Descriptive：显示自变量和因变量的均值、标准差、相关系数矩阵及单侧检验概率。

（4）Model fit：SPSS 默认输出项。显示判定系数、调整的判定系数、回归方程的标准误差、F 检验的方差分析表。

（5）R square change：显示每个自变量进入方程后对 R^2、F 值产生的变化量。

（6）Part and partial correlation：显示方程中各自变量与因变量之间的简单相关、偏相关系数和部分相关系数。

（7）Covariance matrx：显示方程中各自变量间的相关系数和各自变量间的协方差以及各回归系数的方差。

(8) Collinearity Diagnostics：多重共线性分析，输出各个自变量的容忍度、方差膨胀因子、最小容忍度、特征值、条件指标、方差比例等。

(9) 在 Residual 框中：Durbin－Waston 表示输出 D－W 检验值；Casewise Diagnostic：表示显示标准化残差绝对值大于等于 3（SPSS 默认值）的样本数据点的相关信息，包括：标准化残差、观察值、预测值、残差。另外，还输出最小预测值、最小残差、最大预测值、最大残差、最小标准化预测值、最小标准化残差、最大标准化预测值、最大标准化残差以及关于预测值、残差、标准预测值、标准残差的均值和标准差。

2.Options 选项

Option 选项中有两大部分内容。窗口如图 9—7 所示。

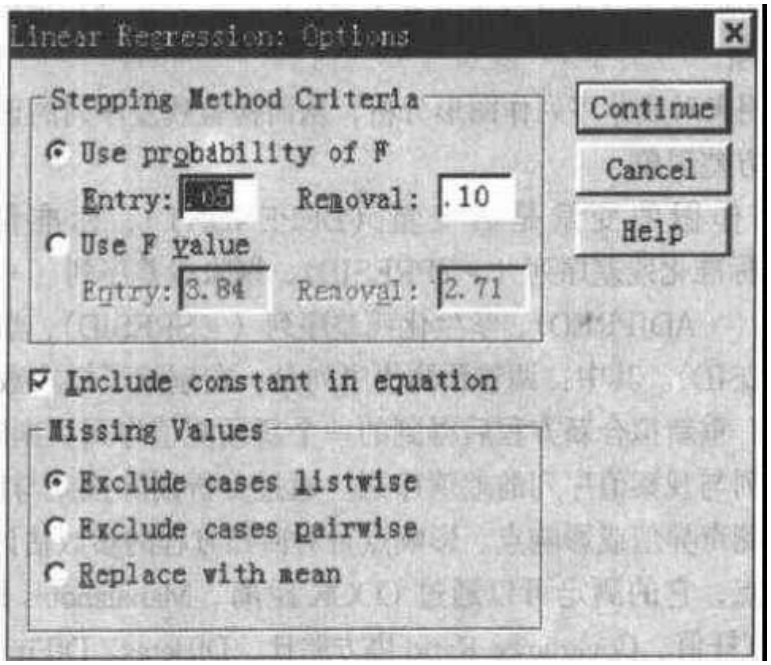


图 9—7　SPSS 线性回归分析的 Option 选项窗口

(1) Stepping Method Criteria 框：与多元线性回归分析中自变量的筛选有关的参数设定。Use probaility of F 项：SPSS 默认，以回归系数显著性检验中的各自变量的 F 统计量的相伴概率值作为自变量是否可以进入方程或是否应剔除出方程的标准。Entry 表示：如果一个自变量的 F 统计量的相伴概率值小于 SPSS 默认的显著性水平 0.05，则应拒绝 H_0，认为该变量对因变量影响是显著的，它应进入回归方程；Remove 表示：如果方程中一个自变量的 F 统计量的相伴概率值水平大于 SPSS 默认的显著性水平 0.10，则不能拒绝 H_0，可以认为该自变量对因变量影响是不显著的，它应剔除出回归方程。在实际应用中，用户可以根据

具体情况修改这两个参数。但应注意：Entry 值应小于 Remove 值。否则，如果 Entry 值应大于 Remove 值，那么自变量一进入方程就会被立即剔除出去；Use F valuse 项表示：以回归系数显著性检验中的各自变量的 F 统计量值作为方程进入会剔除方程的标准。SPSS 默认进入的 F 值为 3.84，剔除的 F 值为 2.27，分别对应的相伴概率值为 0.05 和 0.10。

(2) Included constant in equation：它表示方程中是否有常数项。

(3) Missing Values 框中：Exclude cases listwise 表示删除所有带缺失值的样本数据；Exclude cases pairwise 表示如果计算过程涉及到了某个有缺失值的变量，则暂时删除那些在该变量上是缺失值的个案。如果计算过程没有涉及到有缺失值的变量，那么所有样本数据都参与计算，无论这些样本在其他变量上是否是缺失值；Replace with mean 表示将所有变量的缺失值均以相应变量的均值替代。

3. Plot 选项

Plot 选项用来对残差序列作图形分析，从而检验残差序列的正态性、随机性和是否存在异方差现象。

能够参与作图的变量是因变量（DEPENDNT）、标准化预测值序列（*ZPRED）、标准化残差序列（*ZPRESID）、剔除残差序列（*DRESID）、调整预测值序列（*ADJPRED）、学生化残差序列（*SRESID）、学生化剔除残差序列（*SDRESID）。其中：调整预测值序列是：在剔除了样本数据中的一个奇异值或影响点，重新拟合新方程后得到的一个新预测值序列。剔除残差序列是：调整预测值序列与观察值序列的差值序列。通过分析剔除残差与原残差的差值，可以有助于探测奇异值或影响点。影响点是对回归方程的参数估计结果有较大影响的样本数据点，它的测定可以通过 COOK 距离、Mahalanobis 距离、Leverage values 中心点杠杆值、Covariance Ratio 协方差比、DfBetas、DfFit 等指标来实现，详细内容请参考有关书籍。

这里，SPSS 允许同时作多对变量的简单散点图，可以作标准残差序列的直方图和正态累计概率图。Produce all partial plots 项表示绘制因变量和每个自变量的散点图。

4. WSL 选项

WSL 选项是存在异方差时，利用加权最小二乘法替代普通最小二乘法估计回归模型参数。这时，SPSS 要求选定一个变量作为权重变量。实际工作中，如果无法自行确定权重变量，可以用 SPSS 的权重估计来实现。

5. Save 选项

Save 选项能够将回归分析结果保存到 SPSS 数据编辑窗口的变量中，或某个

SPSS 磁盘数据文件中。其中：

（1）Predicted Values 框中：S.E.of mean prediction 表示保存每个样本所对应的预测均值的标准误差；

（2）Prediction Interval 框中：mean 表示作整体预测时，保存预测均值 95%（默认）的置信区间值；Individual 表示作个体预测时，保存预测值 95%（默认）的置信区间。这两个置信区间是不同的，Individual 选项的置信区间除了 mean 选项中的总体估计预测均值与总体真值间的差距外，还包括个体估计预测值与总体估计预测均值间的差距。

§9.5 线性回归分析案例分析

一、线性回归分析的一般应用

利用第二节案例，对研究与发展课题数作基本预测。在第二节中可以发现课题数与研究发展人员、社科活动人员数投入经费有较强关系，因此，可以在此基础之上，进行多元回归分析，建立回归方程。对案例用逐步筛选法作多元回归分析，分析结果如输出结果 9—3 所示。

输出结果 9—3（a）　　案例多元回归分析结果（一）

Variables Entered /Removed

Model	Variables Entered	Variables Removed	Method
1	研究与发展人员数	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).
2	研究与发展年投入经费	.	Stepwise (Criteria: Probability-of-F-to-enter <= .50, Probability-of-F-to-remove >= .100).
3	具有高级职称的社科活动人员数	.	Stepwise (Criteria: Probability-of-F-to-enter <= .050, Probability-of-F-to-remove >= .100).

输出结果 9—3（a）的含义：本次多元回归分析采用的是逐步筛选法，且回归系数显著性 F 检验的相伴概率值小于 0.05（默认）的自变量引入了回归方程，大于 0.1 的自变量剔除出了回归方程。自变量进入回归方程的次序是：首先，研究与发展人员数进入回归方程，形成模型 1；然后，在模型 1 的基础之上引入第二个自变量研究与发展年投入经费，形成模型 2；最后，在模型 2 的基础之上引

入第三个变量具有高级职称的社科活动人员数，形成最终的回归模型3。

输出结果9—3（b）的含义：第二列是三个回归模型的因变量与所有自变量之间的复相关系数；第三列和第四列分别是用于拟合优度检验的判定系数 R^2 和调整的 R^2；第五列是回归方程的估计标准误差，由均方误差（平均的剩余平方和）开方得到；第六列是自变量逐个引入时引起的 R^2 变化量（R_{ch}）；第七列是自变量逐个引入时引起的回归方程显著性检验 F 值的变化量（F_{ch}）；第十列是 F_{ch} 值的相伴概率值；最后一列是残差独立性检验的 $D-W$ 统计量。

输出结果9—3（b）　　案例多元回归分析结果（二）

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Change Statistics					Durbin-Watson
					R Square Change	F Change	df1	df2	Sig. F Change	
1	.922	.850	.848	51.622	.850	443.377	1	78	.000	1.847
2	.938	.880	.877	46.454	.030	19.323	1	77	.000	
3	.944	.891	.887	44.665	.010	7.289	1	76	.009	

从输出结果9—3（b）可以看出：随着自变量不断引入回归方程，调整的 R^2 在不断提高，回归方程的估计标准误差在不断减小。同时，从 F_{ch} 值的相伴概率值来看，每个都小于显著性水平0.05，说明这些自变量的引入对因变量的解释说明都有比较显著的贡献，它们应保留在回归方程中。

输出结果9—3（c）的含义：第三列的三行分别列出了回归平方和、剩余平方和、因变量的总平方和；第五列是第三列除以第四列自由度的结果，分别是平均的回归平方和、平均的剩余平方和；第六列是用于回归方程显著性检验的 F 统计量，由平均的回归平方和除以平均的剩余平方和得到；第七列是 F 统计量对应的相伴概率值。

输出结果9—3（c）　　案例多元回归分析结果（三）

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1181544.288	1	1181544.288	443.377	.000
	Residual	207860.462	78	2664.878		
	Total	1389404.750	79			

续输出结果

Model		Sum of Squares	df	Mean Square	F	Sig.
2	Regression	1223242.301	2	611621.151	283.426	.000
	Residual	166162.449	77	2157.954		
	Total	1389404.750	79			
3	Regression	1237784.367	3	412594.789	206.814	.000
	Residual	151620.383	76	1995.005		
	Total	1389404.750	79			

从输出结果 9—3（c）可以看出：随着自变量的引入，均方误差在不断减小，说明这些自变量的确为解释因变量作出了贡献，也从另一个角度证明了输出结果 9—3（b）中调整的 R^2 不断升高的原因；另外，这三个模型的 F 检验也证明：所有自变量的回归系数不同时为零，因变量和自变量全体之间确实存在线性关系，可以使用线性模型。

输出结果 9—3（d）的含义：第一、二列为回归分析过程中形成的三个模型的代码和引入自变量的情况，模型 3 是最终的回归模型；第三列是回归模型各自变量的回归系数。依据这个回归系数，案例最终的回归方程是：

研发课题数 = −18.137 + 0.4 研发展人员 + 8.289E−4 经费
+ 0.196 高级职称的活动人员

输出结果 9—3（d）　　案例多元回归分析结果（四）

Coefficients

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
1	(Constant)	−10.670	7.461		−1.430	.157					
	研究与发展人员数	.678	.032	.922	21.057	.000	.922	.922	.922	1.000	1.000
2	(Constant)	−10.524	6.714		−1.568	.121					
	研究与发展人员数	.550	.041	.747	13.336	.000	.922	.835	.526	.495	2.021
	研究与发展年投入经费	1.152E−03	.000	.246	4.396	.000	.777	.448	.173	.495	2.021

续输出结果

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.	Correlations			Collinearity Statistics	
Model		B	Std. Error	Beta			Zero-order	Partial	Part	Tolerance	VIF
3	(Constant)	-18.137	7.044		-2.575	.012					
	研究与发展人员数	.400	.068	.544	5.871	.000	.922	.559	.222	.167	5.973
	研究与发展年投入经费	8.289E-04	.000	.177	2.971	.004	.777	.323	.113	.404	2.477
	具有高级职称的社科人员数	.196	.073	.277	2.700	.009	.910	.296	.102	.137	7.317

第四列是回归系数的标准误差；第五列是标准化回归系数，可以衡量自变量对因变量的贡献程度。依据标准化回归系数，案例最终的标准化回归方程是：

研发课题数=0.544 研发人员+0.177 经费+0.277 高级职称的活动人员

可见，在其他变量保持一定的条下，研究与发展人员增加一个单位，将引起研究与发展课题数平均增加 0.544 个单位；经费增加一个单位，将引起研究与发展课题数平均增加 0.177 个单位；高级职称的活动人员增加一个单位，将引起研究与发展课题数平均增加 0.277 个单位，研发人员数对课题数起了比较主要的作用。

第五、六列是各回归系数显著性检验的 t 统计量和其相应的相伴概率值。从数字上看，这些回归系数的 P 值均小于显著性水平 0.05，因此，应拒绝回归系数显著性检验的零假设，认为这些回归系数与零有显著差异，相应的自变量能够比较好地解释说明因变量的变化，应该保留在回归方程中。

第七、八、九列是关于自变量和因变量相关程度的描述，它们分别是：自变量与因变量的简单相关系数、在控制了回归方程中其他自变量对自变量 x_i 和因变量的影响后，自变量 x_i 与因变量的偏相关系数、在控制了回归方程中其他自变量对自变量 x_i 的影响后，自变量 x_i 与因变量的部分相关系数。可见，首先引入方程的自变量与因变量有最高的偏相关系数。

第十、十一列是自变量共线性检测中的自变量容忍度和方差膨胀因子。仅看模型 3，具有高级职称的社科活动人员具有较小的容忍度，说明它与方程中的其他自变量（主要指研究与发展人员）有一定的相关关系，这符合数据的实际含

义；同样，研究与发展人员也有较小的容忍度，也表明它与方程中其他自变量（主要指具有高级职称的社科活动人员）有一定的相关关系；研究与发展经费的容忍度最高，说明它与方程中其他自变量的相关关系最弱，共线性最小。

输出结果 9—3（e）是回归分析过程中，尚未被引人方程的自变量的情况。从模型 1 看：第三列表示：如果下一步引人社科活动人员，则它的标准化回归系数将是 0.261。同样，如果下一步引人研究与发展经费，则它的标准化回归系数将是 0.246，这个数据与输出结果 9—3（d）中的模型 2 是相吻合的；第四、五列分别是：如果下一步引人某自变量进人回归方程，它的回归系数显著性检验的 t 值和对应的相伴概率值。由于研究与发展经费有最大的 t 值，因此，可以判定下一步它应进人方程，输出结果 9—3（d）的模型 2 证实了这个判断。模型 3 中，由于两个自变量 t 值的相伴概率值都大于显著性水平 0.05，因此可以判定：它们都不能较好地解释说明因变量的变化，最终不能进人方程，输出结果 9—3（d）的模型 3 证实了这个判断。

输出结果 9—3（e）　　案例多元回归分析结果（五）

Excluded Variables

		Beta In	t	Sig.	Partial Correlation	Collinearity Statistics		
Model						Tolerance	VIF	Minimum Tolerance
1	社科活动人员数	.261	2.847	.006	.309	.210	4.767	.210
	具有高级职称的社科活动人员数	0.407	4.193	.000	.431	.168	5.970	.168
	具有高级职称的研究与发展人员	.265	1.271	.207	.143	4.386 E-02	22.800	4.386 E-02
	研究与发展年投入经费	.246	4.396	.000	.448	.495	2.021	.495
2	社科活动人员数	.199	2.346	.022	.260	.203	4.924	.192
	具有高级职称的社科活动人员数	.277	2.700	.009	.296	.137	7.317	.137
	具有高级职称的研究与发展人员	.056	-.272	.786	-.031	3.760 E-02	26.596	3.760 E-02
3	社科活动人员数	.054	.397	.693	.046	7.721 E-02	12.952	5.196 E-02
	具有高级职称的研究与发展人员	-.442	-1.979	.051	.223	2.773 E-02	36.061	2.773 E-02

第六列是在控制了方程中已有自变量的前提下计算出的某自变量与因变量的偏相关系数。从模型1看，由于研究与发展经费具有最高的偏相关系数，因此可以判定下一步它应进入方程，输出结果9—3（d）的模型2证实了这个判断。

第七、八列是共线性检测中的容忍度和方差膨胀因子。从模型1看，研究与发展经费的容忍度最高，表明它与方程中已有自变量的共线性最小，同时它的回归系数显著性检验最显著，因此，可以判定下一步它应进入方程，输出结果9—3（d）的模型2证实了这个判断。从模型2看，虽然社科活动人员具有最高的容忍度，但由于它的回归系数显著性检验没有具有高级职称的社科活动人员显著，因此可以判定下一步具有高级职称的社科活动人员应进入方程，输出结果9—3（d）的模型3证实了这个判断。可见，SPSS判断自变量是否可以进入方程的首要标准是回归系数显著性检验，而容忍度只是一个次要的、不十分严格的标准。

第十列显示的：如果自变量 x_i 进入回归方程后，方程中是所有自变量容忍度的最小值。通过比较最小容忍度和自变量 x_i 的容忍度，可以对目前方程中自变量共线性的大小程度有所把握。

输出结果9—3（f）是利用特征值对方程中自变量共线性的检测结果。仅看模型3，第三列是从自变量相关系数矩阵出发计算出的四个特征值；第四列是条件指标，是最大的特征值除以相应特征值开方后得到的。如：模型3第二行条件指标2.469是3.219/0.528再开方的结果；后面的几列是特征值对各自变量（标准化后）方差贡献度的刻画。如：标准化后，研究与发展人员的方差为1。第一特征值解释了其方差为0.01，第二特征值解释了0.00，第三特征值解释了0.15，第四特征值解释了0.83。同理，标准化后，具有高级职称的社科活动人员的方差为1。第一特征值解释了其方差为0.01，第二特征值解释了0.00，第三特征值解释了0.03，第四特征值解释了0.96。可见，第四特征根既能够解释研究与发展人员的绝大部分方差，同时也可以解释具有高级职称的社科活动人员的绝大部分方差，表明这两个变量具有较强的相关关系。

输出结果9—3（f）　　案例多元回归分析结果（六）

Collinearity Diagnostics

		Eigenvalue	Condition Index	Variance Proportions			
Model	Dimension			(Constant)	研究与发展人员数	研究与发展经费年投入	具有高级职称的社科活动人员数
1	1	1.634	1.000	.18	.18		
	2	.366	2.112	.82	.82		

续输出结果

		Eigenvalue	Condition Index	Variance Proportions			
Model	Dimension			(Constant)	研究与发展人员数	研究与发展年投入经费	具有高级职称的社科活动人员数
2	1	2.292	1.000	.07	.05	.06	
	2	.522	2.095	.77	.02	.21	
	3	.185	3.516	.16	.93	.74	.
3	1	3.219	1.000	.03	.01	.02	.01
	2	.528	2.469	.70	.00	.13	.00
	3	.208	3.938	.17	.15	.75	.03
	4	4.496 E-02	8.462	.10	.83	.09	.96

输出结果 9—3（g）是样本奇异值检验的结果，表明：第 2、36 号样本实际的研究与发展课题数较大，且标准化残差的绝对值已经大于 3，可以认定是奇异值，它们会对回归方程的拟合有一定影响。可以尝试将它们剔除后再建立回归方程。

输出结果 9—3（g）　　案例多元回归分析结果（七）

Casewise Diagnostics

Case Number	Std. Residual	研究与发展课题数
2	-4.625	397.0
36	4.702	577.0

输出结果 9—3（h）是残差分析的结果。第一至第四行数据分别为：根据回归方程计算出的因变量预测值序列、残差序列、标准化预测值序列、标准化残差序列的最小值、最大值、均值、标准差和样本数。

输出结果 9—3（h）　　案例多元回归分析结果（八）

Residuals Statistics

	Minimum	Maximum	Mean	Std.Deviation	N
Predicted Value	-13.430	603.567	88.875	125.173	80
Residual	-206.567	210.022	-4.263E-15	43.809	80
Std.Predicted Value	-.817	4.112	.000	1.000	80
Std.Residual	-4.625	4.702	.000	.981	80

二、线性回归分析在时间序列数据中的应用

线性回归分析不仅可以分析上述截面数据，还可以对时间序列的数据进行分析。表 9—1 是 1960～1986 年美国人均可支配收入和个人消费的有关数据（单位：美元）。利用这些数据，建立消费预测收入的回归方程。

表 9—1　　美国人均可支配收入和个人消费数据

年　份	收　入	消　费	年　份	收　入	消　费
1960	6 036	5 561	1974	8 867	7 826
1961	6 113	5 579	1975	8 944	7 926
1962	6 271	5 729	1976	9 175	8 272
1963	6 378	5 855	1977	9 381	8 551
1964	6 727	6 099	1978	9 735	8 808
1965	7 027	6 362	1979	9 829	8 904
1966	7 280	6 607	1980	9 722	8 783
1967	7 513	6 730	1981	9 769	8 794
1968	7 728	7 003	1982	9 725	8 818
1969	7 891	7 185	1983	9 930	9 139
1970	8 134	7 275	1984	10 419	9 489
1971	8 322	7 409	1985	10 622	9 830
1972	8 562	7 726	1986	10 947	10 142
1973	9 042	7 972			

1. 绘制消费与收入的散点图

利用 SPSS 绘制散点图的功能，绘制消费和收入的散点图如图 9—8 所示。从图 9—8 中可以非常清晰地看出，消费和收入呈明显的线性关系。

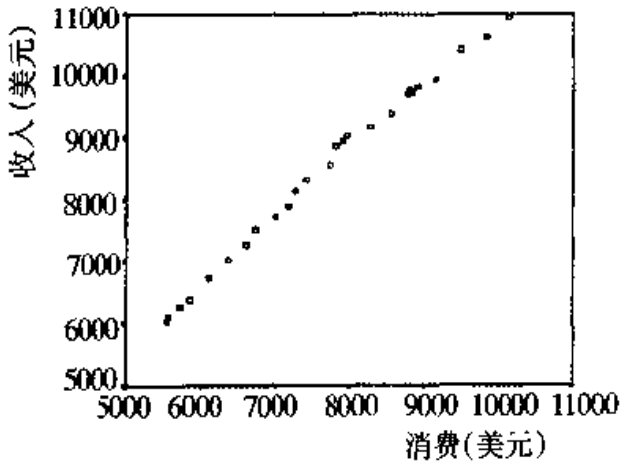


图 9—8　美国 1960—1986 年人均可支配收入和个人消费的散点图

2．建立消费预测收入的回归方程

根据表 9—1 数据，利用 SPSS 进行线性回归分析，结果见输出结果 9—4 所示。

输出结果 9—4　　消费和收入回归分析结果

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate	Durbin Waston
1	.997	.994	.993	118.73	.606

ANOVA

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	55758542.850	1	55758542.850	3955.189	.000
	Residual	352439.224	25	14097.569		
	Total	56110982.074	26			

Coefficients

		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std.Error	Beta		
Model	(Constant)	126.694	135.430		.935	.358
	消费	1.088	.017	.997	62.890	.000

由输出结果 9—4 得到的线性回归方程为：

收入 = 126.694 + 1.088 消费

该回归方程的判定系数为 0.994，拟合优度较高。同时，回归方程的显著性检验和回归系数的显著性检验都非常显著，说明个人消费数据能比较好地解释说明人均可支配收入，方程的质量较好。但 $D-W$ 值为 0.606，说明残差序列存在某种程度的正自相关，这可能主要由于消费和收入之间存在一定的滞后性造成的。

§9.6　曲线估计

一、SPSS 曲线估计的基本思想

在实际的一元回归分析中，一般首先绘制自变量和因变量间的散点图。如果散点图中的数据分布明显呈直线趋势，则可以利用上节讨论的线性回归分析方法估计回归方程。但在很多情况下并非如此，数据在散点图中的分布呈曲线趋势且具有某

种函数的图形特点。这时如果要作回归分析，就需要利用 SPSS 的曲线估计。

SPSS 曲线估计的总体思路是：

首先提供多种可选择的函数模型，用户在不能明确究竟哪种模型更接近样本数据时，可以同时选择几种模型；然后，SPSS 自动完成模型的参数估计，并显示 F 值、对应的相伴概率、R^2 等统计量；最后，用户应选择具有最大 R^2 的模型作为自己的回归模型，并作一些预测。

另外，曲线估计还可以直接用在时间序列的分析中。

SPSS 的曲线估计模型有几种，它们是：线性模型（Linear）、二次项模型（quadratic）、复合模型（Compound）、生长模型（Growth）、对数模型（Logrithmic）、三次项模型（Cubic）、S 模型（S）、指数模型（Exponential）、逆模型（Inverse）、幂模型（Power）、逻辑模型（Logistic），用户可以同时选择几种，然后找出最佳模型。

二、SPSS 曲线估计的基本操作步骤

（1）通过绘制散点图，观察并大致确定几种与样本数据点分布相近的模型。

（2）菜单选项：Statistics－>Regression－>Curve estimation。于是出现如图 9—9 所示的窗口。

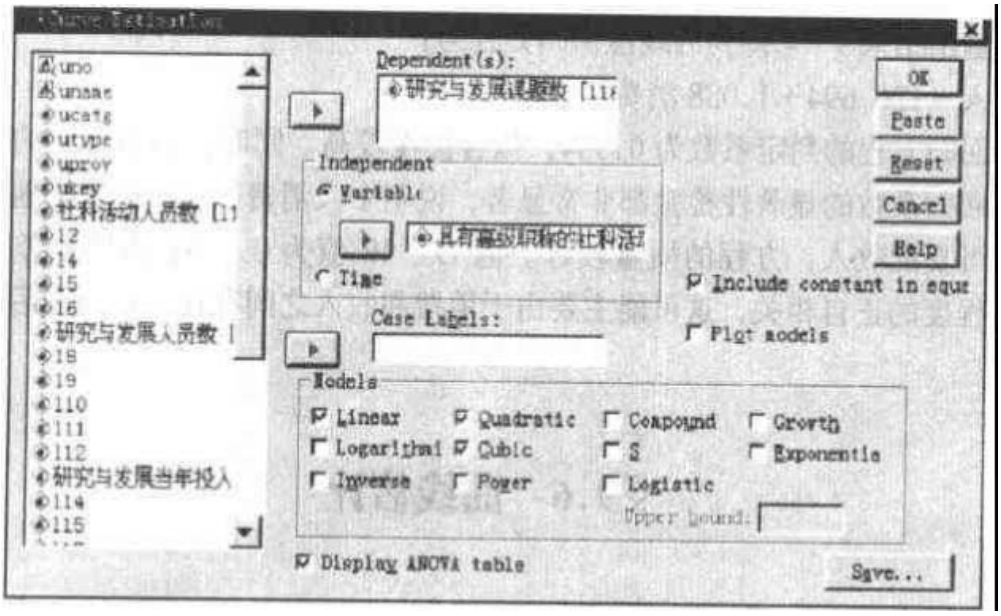


图 9—9　SPSS 曲线估计窗口

（3）选择因变量到 Dependent 框中。

（4）选择一个自变量到 Independent 框。如果选定 Time 参数，表示将以时

间（样本期）作自变量。

（5）选择几种模型。

三、SPSS 曲线估计的其他选项

（1）Display ANOVA table：表示作回归方程显著性检验，并输出相应的方差分析表。

（2）Plot models：表示绘制观察值和预测值的对比图。这样有助于用户直观查看回归线对样本数据点的拟合情况。

（3）Save 选项：表示将分析结果保存到 SPSS 数据编辑窗口中并实现预测。其中：

● Save Variables 框中的 Predicted values 表示保存预测值序列；Residual 表示保存残差值序列；Prediction interval 表示保存预测值 95% 可置信区间的上限和下限值序列；

● Predict case 框表示当以时间作自变量时，可以实现预测。其中：Predict from estimation period through last case 表示计算当前所有样本期内的预测值；Predict through observation 表示计算当前样本期以后若干期的预测值，此时用户应输入一个预测期数。

四、案例分析说明

1. 对高校人文社科统计的课题数与人员数进行曲线估计

对案例数据中的研究与发展课题数、研究与发展人员进行曲线估计。因变量为课题数、自变量为人员数，选择了线性模型、二次曲线、三次曲线三个模型。分析结果如输出结果 9—5 所示。

输出结果 9—5　　案例曲线估计结果

MODEL: MOD-1.

Independent: L3

Dependent	Mth	Rsq	d.f.	F	Sigf	b0	b1	b2	b3
L16	LIN	.827	78	374.03	.000	-25.604	.6447		
L16	QUA	.837	77	197.88	.000	-43.943	.8508	-.0003	
L16	CUB	.887	76	I98.57	.000	11.9463	-.0506	.0025	-.E-06

从输出结果 9—5 中可以看出，各个模型的方程显著性检验都显著，且三次

曲线的拟合优度最高，可以初步选定三次曲线模型。但是通过对图 9—10 的观察发现，如果用三次曲线进行预测，随着具有高级职称的社科活动人员数的增加，研究与发展课题数有下降的趋势。因此，各模型权衡后，决定采用二次曲线，具体模型是：

研究与发展课题数 = −43.943 + 0.8508 具有高级职称的社科活动人员数
−0.0003 具有高级职称的社科活动人员数^2

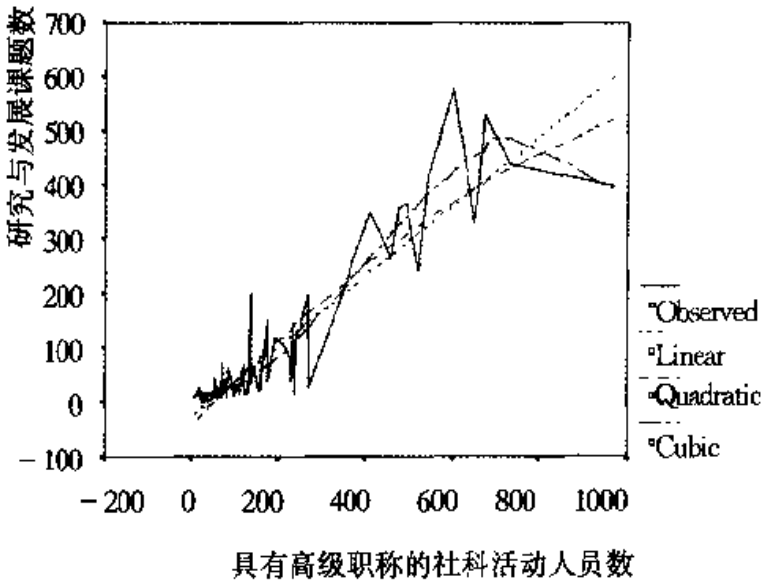


图 9—10　案例曲线估计结果比较

2. 利用线性模型，对表 9—1 中收入数据作外推预测

选择收入为因变量，选择 Time 参数，表示作以年份为单位的时间序列分析。同时，选定预测期数为 28。分析结果和预测值与因变量间的拟合情况图如输出结果 9—6 和图 9—11 所示。

输出结果 9—6　　收入线性回归分析结果

```
MODEL:
Dependent variable.. 收入              Method.. LINEAR
Listwise Deletion of Missing Data
Multiple R                    .98848
R Square                      .97709
Adjusted R Square             .97618
Standard Error              226.74004
        Analysis of Variance:
          DF      Sum of Squares        Mean Square
```

续输出结果

Regression	1	54825705.9	54825705.9		
Residuals	25	1285276.2	51411.0		
F = 1066.41878		Signif F = .0000			
Variables in the Equation					
Variable		B	SE B	Beta	T	Sig T
Time		182.951160	5.602363	.988481	32.656	.0000
(Constant)		5960.498575	89.754454		66.409	.0000

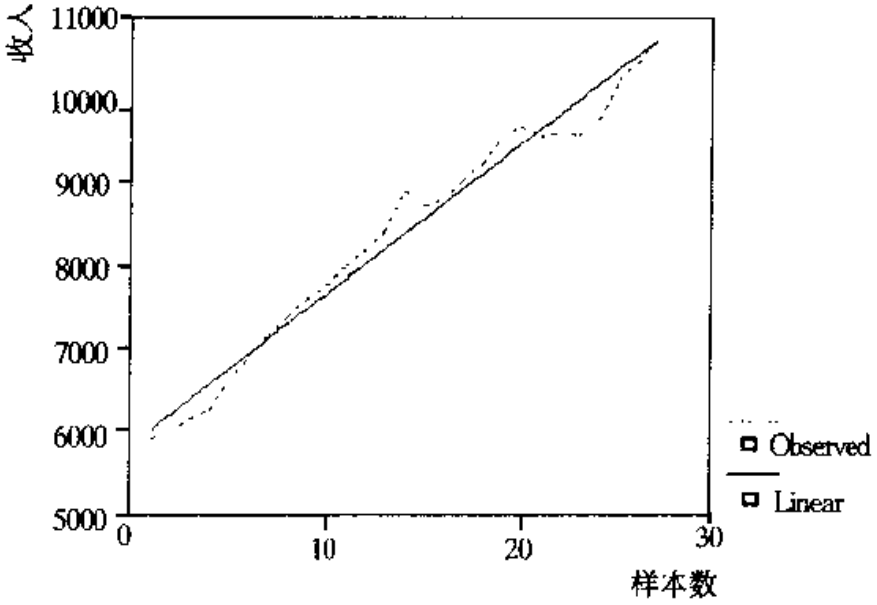


图 9—11　预测值与样本数据的拟合图

于是，建立的回归方程为：

$$收入 = 5\,960.49 + 182.95t$$

预测出的 1987 年人均可支配收入为 11 083 美元。

§9.7　权重估计

一、什么是 SPSS 的权重估计

在上面的章节中讨论的线性回归分析有一个前提要求就是残差序列应是等方差的，但在实际问题中，异方差现象却常常存在。

例如：在分析储蓄和收入的相互关系时，一般在收入水平比价低时，低收入家庭中收入的绝大部分都用于吃的开支，用于储蓄的开支部分极少，而且相当稳

定，使得储蓄的变差比较小。随着收入水平的不断提高，高收入家庭中用于吃的比重会逐渐降低，其他用途的开支所占比例会逐渐增大，在储蓄问题上，即使是同一收入水平的家庭，其储蓄倾向也会极不相同，因而使得储蓄的变差逐渐增大。

又如：分析打字员的打印错误和他的工作时间的长短之间的关系。一般初学打字的人，其出错的可能性是极高的，而且错误的类型也比较分散。随着工作时间的延长，打字技能的不断提高，打印错误出现的概率会不断减少，而且打印错误的类型也比较稳定和集中。

在回归分析中，如果确实存在异方差现象，可以用加权最小二乘法来替代普通最小二乘法。加权最小二乘法的实质是在回归计算过程中给不同的观察值以不同的权数，变差小的观察值赋予的较大的权数，变差大的观察值赋予较小的权数。因此，加权最小二乘法中最为关键的一步是如何确定加权变量的各个权数值。

本节讨论的权重估计就是来解决这个问题的，SPSS 能够自动给出一个加权的范围，并根据对数似然函数值越大越好的原则，给出一个最佳的权数建议值，并依此利用加权最小二乘法建立回归方程。

二、SPSS 权重估计的基本操作步骤

（1）菜单选项：Statistics－＞Regression－＞Weight Estimation。于是出现如图 9—12 所示的窗口。

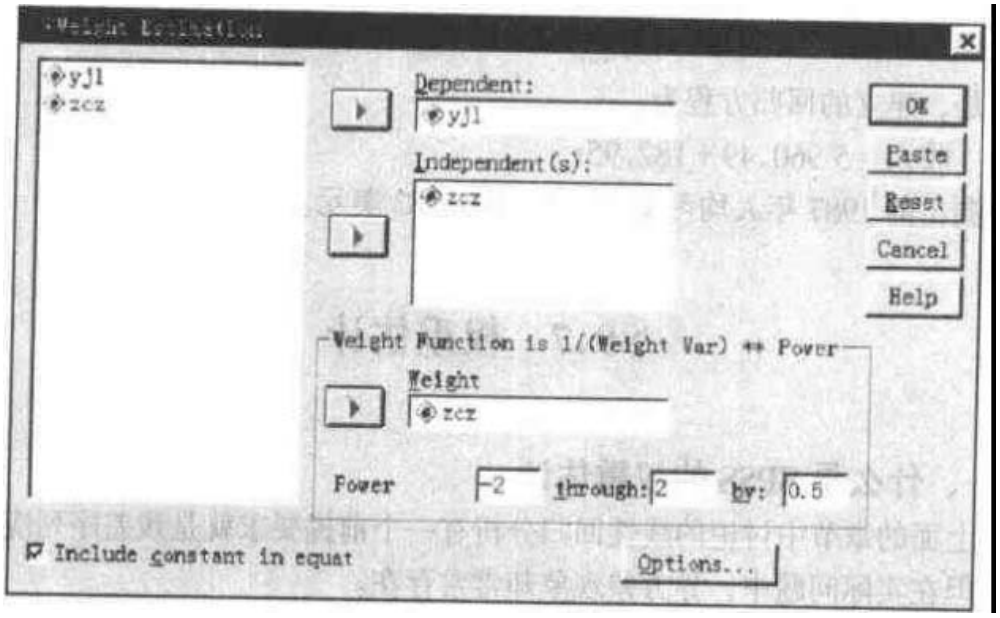


图 9—12　SPSS 权重估计窗口

（2）选择一个变量作为因变量到 Dependent 框中。

（3）选择一个或多个变量作为自变量到 Independent 框中。

（4）选择一个变量作为加权变量到 Weight 框中。SPSS 自动将该变量的 n 次方的倒数作为加权序列值。

（5）在 Power 框中输入计算权数的范围。默认范围 n 是从 -2 至 +2，每次迭代增加 0.5 个步长。用户可以自行调整修改。

于是，SPSS 将根据 n 的范围和增加步长，计算并显示各个对数似然函数值，并自动取使对数似然函数值最大的 n，然后求用户指定的权重变量的 n 次方的倒数序列，并以该序列为权数进行加权回归。

SPSS 权重估计的 Option 选项可以指定是否显示具体的中间计算结果。Display ANOVA and Estimates 框中的 For best power 选项表示只显示最大的 n 和相应的对数似然函数值；For each poere value 选项表示显示每个 n 和相应的对数似然函数值。Save best weight as new variable 表示将计算出的最佳权数序列作为 SPSS 的一个变量存到 SPSS 数据编辑窗口中。

第 10 章

SPSS 的聚类分析

§10.1 聚类分析概述

在社会生活中，我们经常要研究这样一类问题。例如：在学校里，我们会很容易发现这样一种现象，有些学生关系比较密切，会经常在一起，而与另一些学生却很少来往，关系比较疏远。通过对学生家庭情况、成绩、课余爱好数据的分析，会发现学生中确实存在这样一些小群体。在小群体内部，学生们在许多方面有相似性，比如：他们可能有共同的兴趣爱好、有共同的家庭环境等。而在不同的小群体之间，学生们的兴趣爱好和家庭背景则存在比较明显的差异。通过这种分析，我们可以了解学生这个事物的内在结构特征，而首要的工作应是找到学生中的这些小群体。

聚类分析正是统计学中研究这类“物以类聚”问题的一种方法，它属于多元统计分析的范畴。

聚类分析实质是一种建立分类的方法，它能够将一批样本数据（或变量）按照它们在性质上的亲疏程度在没有先验知识的情况下自动进行分类。这里，一个类就是一个具有相似性的个体的集合，不同类之间具有明显的非相似性。在分类

过程中，人们不必事先给出一个分类标准，聚类分析能够从样本数据出发，客观地决定分类标准。

如表 10—1 是 5 名学生的两门课程的成绩。

表 10—1　　聚类分析模拟数据

学号	计算机成绩	英语成绩
001	73	68
002	66	69
003	84	82
004	91	88
005	94	90

现在将这 5 名学生，以两门课程的成绩为依据，将他们分成两类。很明显，001 和 002 号学生应是一类的，003、004、005 号学生是另一类的。进一步、如果将这 5 名学生分成三类，则 001 和 002 号学生是一类的，003 号学生自成一类，004 和 005 号学生是一类。

可见，对学生的分类是在没有指定任何分类标准的前提下进行的，分类的依据完全是从样本数据出发的。001 和 002 号学生之所以在一类中，完全是由于他们的成绩相近，他们之所以没有和 005 号学生分在一类中，也是由于他们的成绩相差较远。

聚类分析正是基于这种出发点，实现样本数据的自动分类的。

SPSS 中提供了两种聚类方法：一种是层次聚类分析法，另一种是快速聚类分析法。

§10.2　层次聚类分析

一、层次聚类法的基本思想

层次聚类的基本思想首先是，在聚类分析的开始，每个样本自成一类；然后，按照某种方法度量所有样本之间的亲疏程度，并把其中最亲密或称最相似的样本首先聚成一小类；接下来，度量剩余的样本和小类间的亲疏程度，并将当前最亲密的样本或小类再聚成一类；再接下来，再度量剩余的样本和小类（或小类和小类）间的亲疏程度，并将当前最亲密的样本或小类再聚成一类；如此反复，直到所有样本聚成一类为止。

由此可见，层次聚类法中，度量数据之间的亲疏程度是极为关键的。那么，如何衡量数据间的亲疏程度呢？这涉及到两个方面的问题，一是如何度量样本数据间的亲疏程度；二是如何度量样本数据与小类之间、小类和小类之间的亲疏程度。

二、样本数据间亲疏程度的度量

样本数据间的亲疏远程度主要是用样本间的距离、或样本间相关系数来度量的。为得到样本间的距离，我们可以将每个样本数据看作是 k 维空间上的一个点。例如：表 10—1 中的 5 个样本就可以看成是 k 为 2 的二维平面上的 5 个点，这个二维平面是由计算机成绩和英语成绩这两个变量构成的。于是，两样本之间的距离就是它们所对应的 k 维空间上的点和点之间的距离，反映了相应样本间的亲疏程度。聚类时，距离较近的样本属于同一类，距离较远的样本属于不同的类。

在 SPSS 中，对不同度量类型的数据采用了不同的测定亲疏程度的统计量。

1. 连续变量的样本距离测度方法

(1) 欧氏距离 (Euclidean Distance)：两样本 (x、y) 之间的距离是各样本每个变量值之差的平方和的平方根 (k 个变量)。

$$EUCLID(x,y)=\sqrt{\sum_{i=1}^{k}(x_i-y_i)^2} \tag{10.1}$$

例如表 10—1 中，001 和 002 号学生的样本欧氏距离是：

$$\sqrt{(73-66)^2+(68-69)^2}$$

(2) 欧氏距离平方 (Squared Euclidean Distance)：两样本 (x、y) 之间的距离是各样本每个变量值之差的平方和 (k 个变量)。

$$SEUCLID(x,y)=\sum_{i=1}^{k}(x_i-y_i)^2 \tag{10.2}$$

例如表 10—1 中，001 和 002 号学生的样本欧氏距离平方和是：

$$(73-66)^2+(68-69)^2$$

(3) Chebychev 距离：两样本 (x、y) 之间的距离是各样本所有变量值之差绝对值中的最大值 (k 个变量)。

$$CHEBYCHEV(x,y)=\max|x_i-y_i| \tag{10.3}$$

例如表 10—1 中，001 和 002 号学生的样本 Chebychev 距离是：

$$\max(|73-66|,|68-69|)$$

(4) Block 距离：两样本 (x、y) 之间的距离是各样本每个变量值之差的绝

对值的总和（k 个变量）。

$$BLOCK(x,y)=\sum_{i=1}^{k}|x_i-y_i| \tag{10.4}$$

例如表 10—1 中，001 和 002 号学生的样本 Block 距离是：

$|73-66|+|68-69|$

（5）Minkowski 距离：两样本（x、y）之间的距离是各样本每个变量值之差的绝对值的 p 次方的总和，再求 p 次方根（k 个变量，p 可以任意指定）。

$$MINKOWSKI(x,y)=\sqrt[p]{\sum_{i=1}^{k}|x_i-y_i|^p} \tag{10.5}$$

例如表 10—1 中，001 和 002 号学生的样本 Minkowski 距离是：

$\sqrt[p]{|73-66|^p+|68-69|^p}$

（6）Customized 距离：用户自定义距离。两样本（x、y）之间的距离是各样本每个变量值之差的绝对值的 p 次方的总和，再求 q 次方根（k 个变量，p、q 可以任意指定）。

$$CUSTOMIZED(x,y)=\sqrt[q]{\sum_{i=1}^{k}|x_i-y_i|^p} \tag{10.6}$$

例如表 10—1 中，001 和 002 号学生的样本 Customized 距离是：

$\sqrt[q]{|73-66|^p+|68-69|^p}$

对表 10—1 中的数据计算欧氏距离的结果如表 10—2 所示。

表 10—2　　表 10—1 数据的各样本欧氏距离矩阵

	Euclidean Distance				
Case	001	002	003	004	005
001		7.071	17.804	26.907	30.414
002	7.071		22.204	31.401	35.000
003	17.804	22.204		9.220	12.806
004	26.907	31.401	9.220		3.606
005	30.414	35.000	12.806	3.606	

从表 10—2 可以知道，004 与 005 号学生的成绩距离最短（3.606），因此，在聚类时它们首先聚成一小类。

连续变量亲疏程度的度量，除了上面的各种距离之外，还可以计算其他统计指标，如：Pearson 简单相关系数、Cosine 相似度等。Cosine 相似度是将样本的

各变量看做是 k 维空间中的向量，然后计算各变量值向量间夹角的余弦（k 个变量）。

$$COSINE(x,y)=\frac{\sum_{i=1}^{k}(x_i y_i)^2}{\sqrt{\left(\sum_{i=1}^{k}x_i^2\right)\left(\sum_{i=1}^{k}y_i^2\right)}} \tag{10.7}$$

2. 顺序或名义变量的样本亲疏测度方法

(1) Chi-square measure：计算 χ^2 统计量。

$$CHISQ(x,y)=\sqrt{\frac{\sum_{i=1}^{k}(x_i-E(x_i))^2}{E(x_i)}+\frac{\sum_{i=1}^{k}(y_i-E(y_i))^2}{E(y_i)}} \tag{10.8}$$

(2) Phi-square measure：计算 φ^2 统计量。

$$PHISQ(x,y)=\sqrt{\frac{\sum_{i=1}^{k}(x_i-E(x_i))^2}{E(x_i)}+\frac{\sum_{i=1}^{k}(y_i-E(y_i))^2}{E(y_i)}}{n} \tag{10.9}$$

对于顺序水准或名义水准的二值变量，也可以通过计算一些有关相似性的统计指标来测定样本间的亲疏程度，这里不对每个指标作具体解释、大家可参考有关书目。

三、样本数据与小类、小类与小类间亲疏程度的度量

在度量了样本数据间的亲疏程度以后，SPSS 会自动将最亲密（距离最小或相关系数最大）的样本首先聚成小类，例如：表 10—2 中，004 和 005 号学生首先聚成一小类。接下来的工作是如何将某个样本与已经形成的小类继续聚集，以及如何将已经形成的小类和小类继续聚集。于是就必然需要度量样本与小类以及小类和小类之间的亲疏远程度。

在 SPSS 中提供了多种度量方法：

1. 最短距离法（Nearest Neighbor）

以当前某个样本与已经形成的小类中的各样本距离中的最小值作为当前样本与该小类之间的距离。

例如：表 10—2 中，样本 001 和（004，005）这个小类的最短距离为 26.907。

2. 最长距离法（Furthest Neighbor）

以当前某个样本与已经形成的小类中的各样本距离中的最大值作为当前样本与该小类之间的距离。

例如：表 10—2 中，样本 001 和（004，005）这个小类的最长距离为 30.414。

3. 组间平均链锁法（Between-groups Linkage）

定义两个小类之间的距离为所有样本对间的平均距离。

例如：样本 1 和样本 2 形成 A 小类，样本 3、4、5 形成 B 小类，那么 A 小类和 B 小类之间的距离就是（样本 1，样本 3）、（样本 1，样本 4）、（样本 1，样本 5）、（样本 2，样本 3）、（样本 2，样本 4）、（样本 2，样本 5）间距离的平均值。可见，组间平均链锁法利用了所有样本对距离的信息，比最短距离法和最长距离法更进了一步。

4. 组内平均链锁法（Within-groups Linkage）

与组间平均链锁法类似，但这里的平均距离是对所有样本对的距离求平均值，包括小类之间的样本对、小类内的样本对。

5. 重心法（Centroid clustering）

将两小类间的距离定义成两小类重心间的距离。每一小类的重心是该类中所有样本在各个变量上的均值所代表的点。

例如：（004、005）小类的重心点为：计算机成绩 92.5，英语成绩 89。

6. 离差平方和法（Ward's Method）

该方法的主要出发点是：在聚类过程中，使小类内各样本的欧氏距离总平方和增加最小的两小类合并为一类。

例如：有 A、B、C 三个小类，如果 A 和 B 合并为一类，那么，（A，B）类中所有样本的欧氏距离总平方和一定小于（A，C）和（B，C）类中所有样本的欧氏距离总平方和。

四、层次聚类分析的几点说明

1.SPSS 中的层次聚类分析有 Q 型聚类和 R 型聚类两种

Q 型聚类就是对样本数据进行聚类，R 型聚类是对变量进行聚类。Q 型聚类是使具有共同特征的样本聚集在一起，下一步可以分别对不同类中的样本作分析；R 型聚类是使具有共同特征的变量聚集在一起，下一步可以从不同类中分别挑选出具有代表性的变量作分析，从而减少分析变量的个数。

2. 应注意聚类分析所选择的变量能否反应所要聚类的特征

聚类分析的目的是找到不同类群体中样本数据的特点，因此在聚类分析选定

变量时，应注意所选择的变量能否反应所要聚类的特征。例如：如果希望对学校规模作聚类分析，但如果分析时没有选定诸如：在校学生人数、教师人数、校园面积、设备、设施情况等变量，那么聚类后的结果就无法反映关于学校规模的特征。

3. 应注意聚类分析的变量数据是否存在数量级上的差异

由于聚类分析是以距离来度量样本亲疏程度的。从各种距离的定义上看，变量不同的数量级别将对距离产生较大的影响。

例如：有表 10—3 三个样本数据。

表 10—3　　聚类分析模拟数据

样本号	社科活动人员数（人）	研究与发展年投入经费（元）	研究与发展课题数（项）
1	410	4 380 000	19
2	336	1 730 000	21
3	490	220 000	8

如果分别以元、万元作为计量单位，这三个样本间的距离如表 10—4 所示。

表 10—4　　聚类分析中不同数量级别对距离的影响

	样本的欧氏距离	
	元	万元
(1, 2)	265 000	81.623
(1, 3)	218 000	193.700
(2, 3)	47 000	254.897

从表 10—4 可以看出：以元为计量单位时，样本 2 和 3 的关系最为密切，其次是 1 和 3，最后是 1 和 2，经费投入起了决定性作用；以万元为计量单位是，样本 1 和 2 的关系最为密切，其次是 1 和 3，最后是 2 和 3，人员数起了决定作用。可见，不同数量级别对距离的影响程度。

为克服上述问题的存在，聚类分析之前，应查看变量数据之间是否存在数量级上的差异。如果数据在数量级上存在差异，应首先进行标准化处理、然后再作聚类分析。

五、SPSS 层次聚类分析的基本操作步骤

（1）菜单选项：Statistics－＞Classify－＞Hierachical Cluster。于是出现如图

10—1 所示的窗口。

(2) 选择参与层次聚类分析的变量到 Variables 框中。

(3) 选择一个字符型变量作为样本的标记变量到 Label Cases 框中。选择标记变量将大大增强聚类分析结果的可读性。

(4) 选择作 R 型聚类还是作 Q 聚类。系统默认作 Q 型聚类。

(5) 按 Method 按钮选项指定距离计算方法。其中：Cluster 下拉框中给出的是计算类间距离的方法，SPSS 默认的是组间平均链锁法；Measure 框中给出的是计算样本距离的方法，Interval 适用于连续型变量，SPSS 默认计算欧氏距离平方；Counts 适用于顺序水准或名义水准的变量，SPSS 默认计算 χ^2 统计量；Binary 适用于二值变量，SPSS 默认计算二元欧氏距离平方。

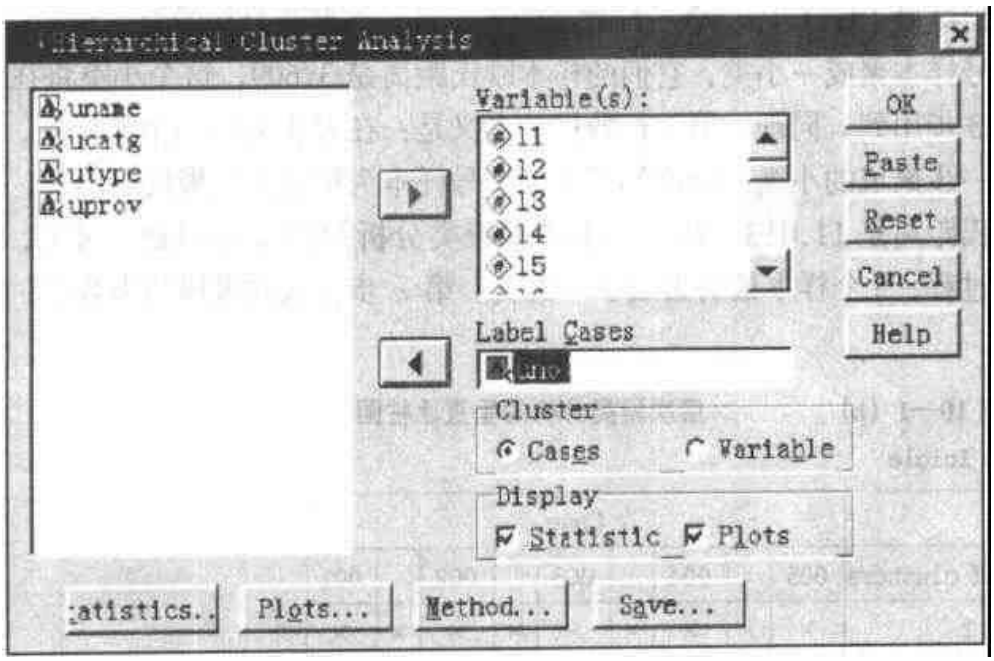


图 10—1 SPSS 层次聚类分析窗口

SPSS 层次聚类分析默认输出的分析结果有凝聚状态表和冰柱图。

输出结果 10—1 (a) 是表 10—1 数据进行层次聚类分析的凝聚状态表。其中，样本距离为欧氏距离，小类间的距离按组间平均链锁法计算。

凝聚状态表的第一列表示聚类分析的第几步；第二列、第三列表示本步聚类中哪两个样本或小类聚成一类；第四列是相应的样本距离或小类距离；第五列、第六列表明本步聚类中，参与聚类的是样本还是小类。0 表示样本，数字 n (非 0) 表示由第 n 步聚类产生的小类参与本步聚类；第七列表示本步聚类的结果将在下面聚类的第几步中用到。

输出结果 10—1（a）　　层次聚类分析的凝聚状态

Agglomeration Schedule

	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
Stage	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	4	5	3.606	0	0	3
2	1	2	7.071	0	0	4
3	3	4	11.013	0	1	4
4	1	3	27.288	2	3	0

输出结果 10—1（a）第一行数据的含义是：在聚类分析的第一步中，4 号样本与 5 号样本聚成一小类，它们的样本欧氏距离是 3.609，这个小类将在聚类分析的第 3 步用到。同理，第三行数据的含义是：在聚类分析的第三步中，3 号样本与第一步聚成的小类（以该小类中的 4 号样本为标记）又聚成一小类，它们的样本欧氏距离是 11.013，这个小类将在聚类分析的第 4 步用到。可见，经过 4 步聚类过程，5 个样本最后聚成了一大类。第 n 步完成时聚成的类数是：总样本数减 n。

输出结果 10—1（b）　　层次聚类分析的垂直冰柱图

Vertical Icicle

	Case								
Number of clusters	005		004		003		002		001
1	×	×	×	×	×	×	×	×	×
2	×	×	×	×	×		×	×	×
3	×	×	×		×		×	×	×
4	×	×	×		×		×		×

冰柱图因其样子非常像冬天房屋顶上垂下的冰柱而得名，它以图形的方式显示层次聚类分析结果。一般从冰柱图的最后一行开始观察，第一列表示类数。

输出结果 10—1（b）中，如果聚成四类，则 005、004 聚成一类，其他样本自成一类；如果聚成三类，则 005、004 为一类，002、001 为一类，003 自成一类；如果聚成两类，则 005、004、003 为一类，002、001 为另一类。

六、SPSS 层次聚类分析的其他选项

SPSS 层次聚类分析还有其他可选项，可以实现其他附加功能。

1. 对不同数量级别的变量作标准化处理

按 Method 按钮选项，Transform Values 框可以指定变量进行标准化处理的方法。SPSS 默认值为 None，表示不作标准化处理。如果希望进行标准化处理，可以根据实际情况指定相应处理方法。其中：

（1）Z scores：表示计算 Z 分数。它将需作标准化处理的变量的每个变量值减去该变量的平均值，然后再除以该变量的标准差。标准化后的变量值平均值为 0，标准差为 1。

（2）Range －1 to 1：表示将需作标准化处理的变量的每个变量值除以该变量的全距。处理以后的变量值的范围在 －1～＋1 之间。变量值中必须含有负数。

（3）Range 0 to 1：表示将需作标准化处理的变量的每个变量值减去该变量的最小值后除以该变量的全距。处理以后的变量值的范围在 0～＋1 之间。

（4）Maximum magnitude of 1：表示将需作标准化处理的变量的每个变量值除以该变量的最大值。处理以后的变量值的最大值为 1。

（5）Mean of 1：表示将需作标准化处理的变量的每个变量值除以该变量的平均值。

（6）Standard deviation of 1：表示将需作标准化处理的变量的每个变量值除以该变量的标准差。

在进行标准化处理时，还应指定标准化处理是针对变量的还是针对样本的。By variables 表示针对变量，适用于 Q 型聚类分析；By cases 表示针对样本，适用于 *R* 型聚类分析。

2. 指定输出分析结果

（1）SPSS 层次聚类分析的图形输出。SPSS 层次聚类的图形结果有两种：一种是输出结果 10—1（b）所示的冰柱图，另一种是树形图（Dendrogram）。

树形图以树的形式展现聚类分析中的每一次类合并情况，SPSS 首先自动将各类之间的距离重新转化到 0～25 之间，然后再将其近似地表示在图上。图 10—2 就是对表 10—1 中的数据绘制聚类分析树形图的结果。

从图 10—2 可以看出，004 和 005 的距离最近，首先合并成一类；其次是 001 和 002，它们合并成一类，之间的距离大于 004 和 005 之间的距离；再次是 003 与（005，004）合并；最后所有样本聚成一类。可见，树形图仅是非常粗略地展现聚类分析的过程，如果样本或小类之间距离差距较小，在图形上就比较难分辨，此时，还应借助于凝聚状态表。

* * * * * * H I E R A R C H I C A L C L U S T E R A N A L Y S I S
*

Dendrogram using Average Linkage (Between Groups)

Rescaled Distance Cluster Combine

CASE		0	5	10	15	20	25
Label	Num	+ - - - - + - - - - + - - - - + - - - - + - - - - -					
004	4	- + - - - - - +					
005	5	- + + - +					
003	3	- - - - - - - + I					
001	1	- - - - - + - +					
002	2	- - - - - +					

图 10—2　层次聚类分析的树形图

绘制图形的具体操作步骤是：

按Plots按钮选项，出现如图 10—3 所示的窗口。选定 Dendrogram 选项，指定输出聚类分析树形图；在 Icicle 框中指定输出冰柱图，并选择冰柱图显示的方向，其中：Vertical 表示纵向显示，Horizontal 表示横向水平显示。另外，根据实际需要，用户还可以指定显示聚类分析全过程的冰柱图，选项是 All clusters；如果仅希望显示聚类分析某一阶段的冰柱图，可以选定 Specified range of clusters 项，并输入从第几类开始显示，到第几类结束显示，中间间隔几类。

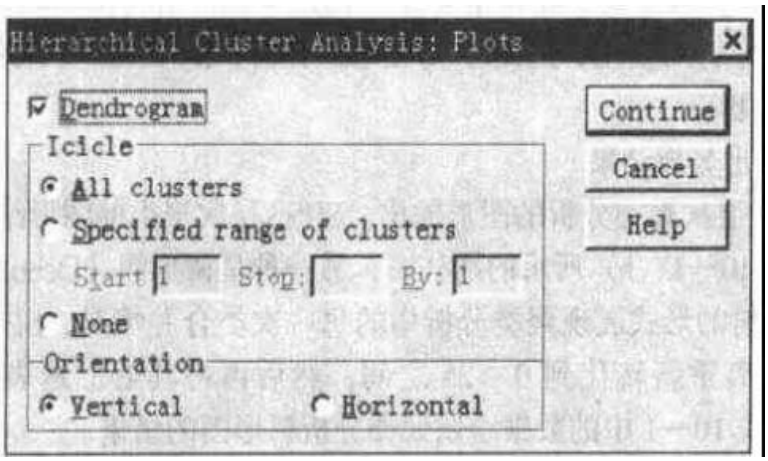


图 10—3　SPSS 层次聚类分析的 Plots 窗口

（2）SPSS 层次聚类分析的所属类成员情况输出。SPSS 层次聚类分析的所属类成员输出能够非常清楚地显示每个样本究竟归属于哪个类。由于在层次聚类分析过程中，SPSS 能够产生任意可能的聚类结果，因此，每个类的类成员情况在

聚类分析的过程中会不断地变化，相应地，同一个样本在聚类分析的过程中的所属情况也在随之发生着变化。例如：某个样本在聚成两类时可能是属于第二类的，但在聚成三类时可能又属于第三类了。所以，如果希望了解每个样本的类归属情况，则应首先确定聚类的类数。输出结果 10—1（c）是类成员情况表。

输出结果 10—1（c）　　层次聚类分析的类成员情况

Cluster Membership

Case	3 Clusters
1:001	1
2:002	1
3:003	2
4:004	3
5:005	3

从输出结果 10—1（c）可知：当聚成三类时，001、002 在一类中，003 在第二类中，004、005 在第三类中。

具体操作步骤是：

按 Statistics 按钮选项，出现如图 10—4 所示的窗口。在 Cluster Membership 框中，可以选择 Single solution 并输入一个具体数值 n（n 应小于总样本数），指定仅显示在聚成 n 类时，所有样本的类属情况，也可以选择 Range of solutions 并输入两个具体数值 n_1、n_2（$n_1 < n_2$），指定显示在聚成 n_1 类到 n_2 类时，所有样本的类归属情况。

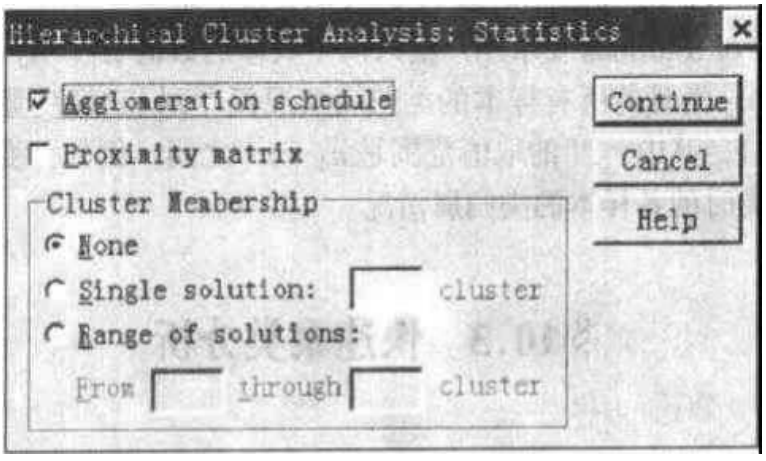


图 10—4　SPSS 层次聚类分析的 Statistics 窗口

另外，如果选择了 Proximity matrix 选项，SPSS 还将显示各样本距离矩阵。通过观察距离矩阵，有助于对 SPSS 层次聚类分析方法的进一步深入理解。

3. 保存 SPSS 层次聚类分析的结果

可以将 SPSS 层次聚类分析的最终结果以变量的形式保存到 SPSS 数据编辑窗口中。

具体操作步骤是：

按 Save 按钮选项，出现如图 10—5 所示的窗口。Cluster Membership 框中：

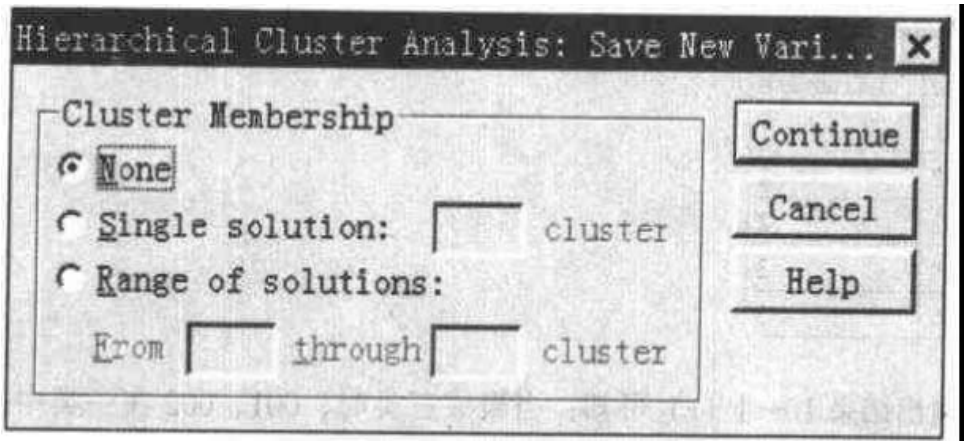


图 10—5 SPSS 层次聚类分析的 Save 窗口

(1) Single solution 要求用户输入一个具体数值 n（n 应小于总样本数），表示将聚成 n 类时所有样本的类归属情况保存到 SPSS 变量中。这个变量的命名为 cluN-M，其中：N 即为 n，是类数。由于不同的距离计算方法会产生不同的聚类分析结果，即使要求聚成 n 类，同一个样本的类归属情况也会因计算方法的不同而不同。在实际分析时，可以对几种方法作尝试，SPSS 会将每次尝试分析的结果都保存下来，M 就表示是第 M 次尝试分析的结果。

(2) Range of solutions 要求用户输入两个具体的数值 n_1、n_2（$n_1 < n_2$），表示将生成 $n_1 \sim n_2$ 类时的所有样本的类归属情况保存到 SPSS 变量中。这些变量的命名为 cluN-1，其中：N 的取值范围是 $n_1 \sim n_2$ 之间。例如：变量 clu4-1 中保存的是聚成 4 类时的各样本的类归属情况。

§10.3 快速聚类分析

从 SPSS 层次聚类分析的过程可以看出，层次聚类分析对计算机的性能要求比较高，在大样本情况下，采用层次聚类分析方法，得到分析结果所需要花费的计算机 CPU 时间比较长，需要占用的计算机内存比较多，易出现死机情况。因此，在大样本情况下，可以采用 SPSS 快速聚类分析方法作聚类分析。SPSS 快

速聚类分析的结果比较简洁易懂，对计算机的性能要求也不是很高，因而受到广泛的使用。

一、SPSS 快速聚类分析的基本思路

SPSS 快速聚类分析仍以距离衡量样本间的亲疏程度，基本思路是：

（1）首先用户应指定希望聚成几类，如：希望聚成 K 类。

（2）确定 K 个类的初始类中心点。初始类中心点可以通过两种方法指定：一种是用户自行指定，指定 K 组数据作为初始类中心点；另一种是有 SPSS 自动指定，系统会根据样本数据的具体情况选择 K 个有代表性的样本数据作为初始类中心点。

（3）计算所有样本数据点到 K 个类中心点的欧氏距离，SPSS 按照距 K 个类中心点距离最短的原则，把所有样本分派到各中心点所在的类中，形成一个新的 K 类，完成一次迭代过程。

（4）重新计算 K 个类的类中心点。计算每类中各个变量的变量值均值，并以均值点作为新的类中心点。

（5）重复（3）和（4）步，直到达到指定的迭代次数或达到终止迭代的判断要求为止。

判断是否结束迭代过程的标准有两个，满足其中一个即可结束快速聚类分析过程。它们是：

- 迭代次数等于指定的迭代次数。系统默认的迭代次数为 10 次。
- 迭代收敛标准：本次迭代产生的新的类中心点距上次迭代后确定的类中心点的最大距离小于 0.02。

从快速聚类的基本思路可以看出，SPSS 快速聚类分析实质就是 K－Mean 聚类。

从聚类分析的结果看，快速聚类与层次聚类法的不同在于：层次聚类可以对不同的聚类类数而产生一系列的聚类解，而快速聚类只能产生单一的聚类解，且需要用户事先指定聚类数目。在初始类中心点的选择上，虽然用户的可选择性比较大，但也应根据实际工作的需要和以往的经验，指定比较合理的初始类中心点，否则，就应增加迭代次数，以保证最终聚类结果的合理性和准确性。

二、SPSS 快速聚类分析的基本操作步骤

（1）菜单选项：Statistics－＞Classify－＞K－Means Cluster。于是出现如图

10—6 所示的窗口。

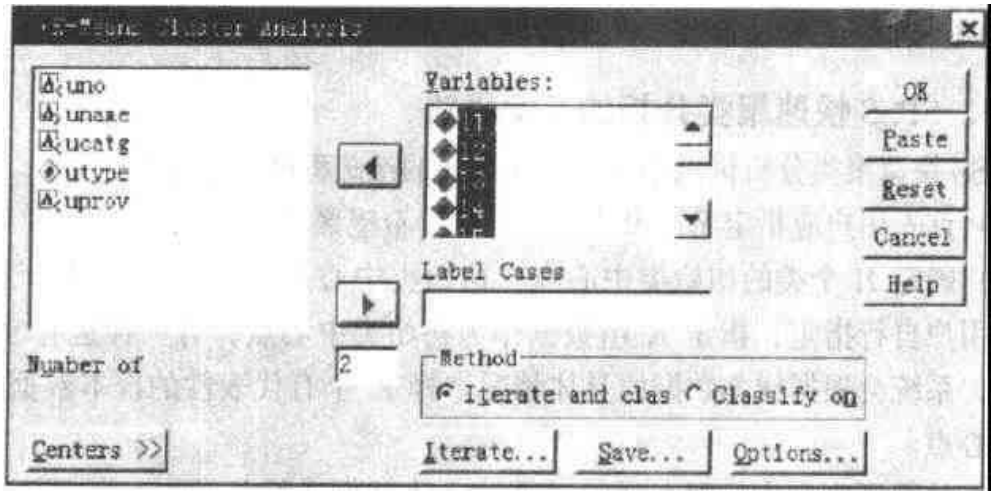


图 10—6　SPSS 快速聚类窗口

（2）选定参加快速聚类分析的变量到 Variables 框中。

（3）选择一个字符型变量作为样本的标记变量到 Label Cases 框中。选择标记变量将大大增强聚类分析结果的可读性。

（4）在 Number of Cluster 框中输入快速聚类的聚类数目，此数目应小于样本总数。

（5）在 Method 框中选择类中心点确定方法。其中：Iterate and classify 是 SPSS 默认的，表示在聚类分析的每一步都重新计算新的类中心点；Classify only 表示聚类分析过程类中心点始终为初始类中心点，仅作一次迭代。

（6）按 Iterate 按钮确定快速聚类的迭代终止条件。其中：Maximum iteration 为最大迭代次数，当达到这个迭代次数时将终止聚类分析过程。SPSS 默认为 10 次；Convergence criterion 为迭代的收敛标准。当本次迭代后形成的若干个类中心点距上次迭代确定的若干个类中心点的最大距离小于指定数据时，终止聚类分析过程。SPSS 默认为 0.02。Use running means：如果选定该项，表示每当一个样本被分配到一类后便立即重新计算新的类中心点。选定该项后，快速聚类的类中心点将与样本进入的前后顺序有关；如果不选定该项表示，完成了所有样本的依次类分配后再计算各类中心，这种方式可以节省运算时间。

SPSS 快速聚类分析默认的输出结果有：初始类中心点情况、迭代步骤、最终各类的样本情况。下面是对表 10—1 数据作快速聚类分析的结果，要求最终聚成两类，系统自行指定初始类中心点。

输出结果 10—2（a）　　快速聚类的初始类中心点

Initial Cluster Centers

	Cluster	
	1	2
计算机成绩	94.00	66.00
英语成绩	90 00	69.00

输出结果 10—2（a）表明：系统自行指定的两个初始类中心点为：（计算机成绩 94，英语成绩 90），（计算机成绩 66，英语成绩 69），分别是 005 号学生和 002 号学生。

输出结果 10—3（b）　　快速聚类的迭代步骤

Iteration History

	Change in Cluster Centers	
Iteration	1	2
1	5.467	3.536
2	.000	.000

输出结果 10—2(b)表明:在迭代过程中,完成第一次迭代后形成的两个新类中心点距初始类中心点的欧氏距离分别为 5.467 和 3.536;第二次迭代后形成的两个新类中心点几乎与上次确定的中心点没有差别。经过两次迭代,快速聚类完成。

输出结果 10—2（c）　　快速聚类的最终结果

Number of Cases in each Cluster

Cluster	1	3.000
	2	2.000
Valid		5.000
Missing		.000

输出结果 10—2（c）表明：当聚成两类时，第一类中有 3 个样本，第二类中有 2 个样本。

三、SPSS 快速聚类分析的其他选项

SPSS 快速聚类分析还有其他选项，用来完成其他附加功能。

1. 保存 SPSS 快速聚类分析的结果

可以将 SPSS 快速聚类分析的结果以变量的形式保存到 SPSS 数据编辑窗

口中。

具体操作步骤是：

按Save按钮，Cluster membership 表示所有样本所属类的类号保存到名为 qcl-1 的 SPSS 变量中；Distance from cluster center 表示将所有样本距所属类中心点的欧氏距离保存到名为 qcl-2 的 SPSS 变量中。

2. 用户指定初始类中心点

K 个类的初始类中心点可以由用户自行指定，它是通过读取另一个 SPSS 数据文件的方式来实现的。这个 SPSS 数据文件应事先存储在磁盘上，其中：各变量的变量名应与当前 SPSS 数据编辑窗口中的变量名完全吻合；应有 K 个样本数据，它们分别为 K 个类的初始类中心。

具体的操作步骤是：

按Center按钮选项，Read initial from 表示某个 SPSS 数据文件中读入初始类中心点；Write final as 表示将快速聚类分析的最终类中心点写入某个 SPSS 数据文件中。

3. SPSS 快速聚类分析的 Option 选项

按Option按钮选项，出现如图 10—7 所示的窗口。可以指定输出其他聚类分析的结果，并选择对缺失数据的处理方法：

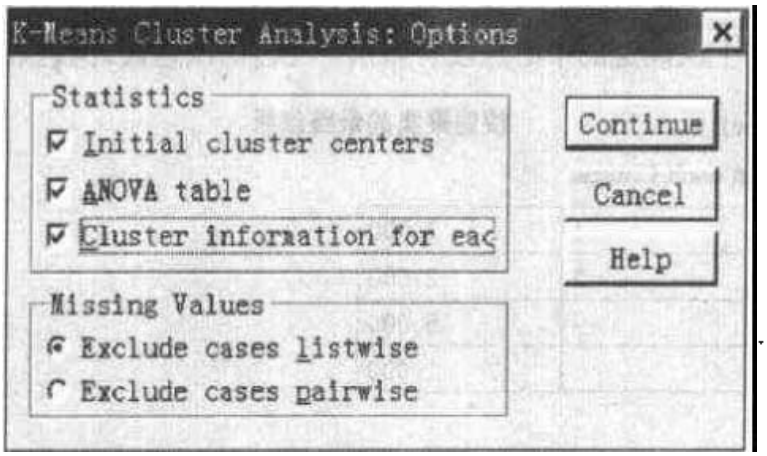


图 10—7　SPSS 快速聚类的 Options 窗口

(1) Statistics 框：选择显示其他分析结果。其中：Initial cluster centers：SPSS 默认项，表示显示有关初始类中心点的数据；ANOVA table：对快速聚类分析产生的类，作单因素方差分析，并输出各个变量的方差分析表。这里的控制变量为存储类号的变量；Cluster information for each case：输出样本的分类信息和它们距所属类中心点的距离。

（2）Missing Values 框：选择缺失数据的处理方法。Exclude case listwise：表示首先将所有带有缺失值的样本全部删除，然后再作聚类分析；Exclude case pairwise：表示只删除那些所有变量均为缺失值的样本。

输出结果 10—2（d）是选择了 Option 选项后附加的分析结果。

输出结果 10—2（d）　　快速聚类的类成员情况

Cluster Membership

Case Number	Cluster	Distance
1	2	3.536
2	2	3.536
3	1	7.341
4	1	1.886
5	1	5.467

输出结果 10—2（d）表明：1、2 号样本在第 2 类中，3、4、5 号样本在第 1 类中。1、2 号样本距本类中心的欧氏距离均为 3.536。3、4、5 号样本距本类中心点的欧氏距离分别 7.341、1.866 和 5.467，4 号样本最接近类中心点，3 号样本距类中心点最远。

输出结果 10—2（e）　　快速聚类的最终类中心点

Final Cluster Centers

	Cluster	
	1	2
计算机成绩	89.67	69.50
英语成绩	86.67	68.50

输出结果 10—2（e）表明：快速聚类最终形成的两个类中心，分别为（计算机成绩 89.67，英语成绩 86.67），（计算机成绩 69.50，英语成绩 68.50）。

输出结果 10—2（f）表明：快速聚类后，样本数据分成两类。第二列是平均的组间平方和，第四列是平均的组内平方和，第六列 F 统计量是第二列除以第四列的结果，最后一列是 F 统计量的相伴概率值。由于相伴概率值均大于显著性水平 0.05，所以不能拒绝单因素方差分析的零假设，可以认为这两类样本在计算机成绩和英语成绩的均值上都没有出现显著差异，聚类效果不很理想，可以再尝试

分成三类的情况。由此也说明，在实际分析中，需要通过反复作快速聚类，来最终确定一个较为合理的聚类数目。或者，可以借助层次聚类分析法，挑选部分样本数据作层次聚类分析，分析的结果可以作为确定快速聚类数目的参考依据。

输出结果 10—2（f）　　快速聚类分析的各类单因素方差分析结果

ANOVA

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
计算机成绩	488.033	1	25.722	3	18.973	.022
英语成绩	396.033	1	11.722	3	33.785	.010

§10.4　案例分析

一、案例说明

查阅《中国普通高等院校人文社会科学统计资料汇编》，得到 1999 年全国 80 所重点院校的人文社会科学研究基本情况数据。这些指标涉及到社科活动人员数、研究与发展人员数、研究与发展经费数、研究与发展课题数、研究与发展成果数、成果获奖数六个方面的指标。

现在希望比较客观地根据上述几个方面的指标，对 80 所重点院校进行分类。这里并没有给定分类的标准，也没有给出应将 80 所重点院校分成几类，而要求比较客观地从数据自身出发进行分类。这时聚类分析便是一个好的解决方案。聚类分析之后，我们还需要研究各类的主要特征，可以通过基本的描述统计分析实现。

二、对案例数据六大方面的 26 个变量分别作层次聚类分析和快速聚类分析

1. 层次聚类分析

层次聚类分析中，样本间的距离计算欧氏距离、小类间距离以组间平均链锁法计算，对样本数据作 Z 化处理。

由于篇幅所限制，输出结果 10—3 仅给出了案例数据作层次聚类分析产生的凝聚状态表的部分结果。在聚类分析的第一步中，第 11 号样本与第 14 号样本聚

成一小类，它们的样本欧氏距离是 0.443，这个小类将在聚类分析的第 3 步用到。同理，在聚类分析的第三步中，第一步聚成的小类(以该小类中的 11 号样本为标记)与第 66 号样本又聚成一小类，它们的样本欧氏距离是 0.548，这个小类将在聚类分析的第 6 步用到。可见，经过 79 步聚类过程，80 个样本最后聚成了一大类。

输出结果 10—3　　案例层次聚类中凝聚状态表的部分输出结果

Agglomeration Schedule

	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
Stage	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	11	14	.443	0	0	3
2	33	77	.515	0	0	10
3	11	66	.548	1	0	6
4	19	30	.587	0	0	12
5	72	73	.628	0	0	21
6	5	11	.704	0	3	17
…	…	…	…	…	…	…
73	8	12	8.146	71	0	74
74	8	65	8.615	73	0	75
75	8	18	8.719	74	72	77
76	1	17	9.252	67	0	78
77	8	36	10.188	75	0	78
78	1	8	10.636	76	77	79
79	1	2	20.930	78	0	0

由于样本较多，其他图形结果略去。

分成五类时：北京外国语大学、北方交通大学、北京理工大学、北京邮电大学、北京科技大学、兰州大学等 66 所大学属第一类；中国人民大学为第二类；北京大学、北京师范大学、南开大学、吉林大学、复旦大学、南京大学等 11 所大学属于三类；清华大学为第四类；华东师范大学为第五类。

对这五类大学的社科活动人员（L1）、研究与发展人员（L7）、研究与发展经费（L13）、研究与发展课题（L16）、研究与发展成果（专著）（L20）、获奖情况（L25）作单因素方差分析，得到输出结果 10—4 所示的结果。

输出结果 10—4　　案例数据聚类后单因素方差分析的结果

ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
L1	Between Groups	7864830.109	4	1966207.527	64.513	.000
	Within Groups	2285833.091	75	30477.775		
	Total	10150663.200	79			
L7	Between Groups	2110970.657	4	527742.664	84.737	.000
	Within Groups	467103.030	75	6228.040		
	Total	2578073.688	79			
L13	Between Groups	49201907113.888	4	12300476778.472	64.110	.000
	Within Groups	14389871204.500	75	191864949.393		
	Total	63591778318.388	79			
L16	Between Groups	1139701.891	4	284925.473	83.453	.000
	Within Groups	256064.909	75	3414.199		
	Total	1395766.800	79			
L20	Between Groups	654631.806	4	163657.951	166.022	.000
	Within Groups	73932.182	75	985.762		
	Total	728563.988	79			
L25	Between Groups	15997.950	4	3999.488	18.362	.000
	Within Groups	16336.000	75	217.813		
	Total	32333.950	79			

输出结果 10—4 表明，这五类院校在社科活动人员、研究与发展人员、研究与发展经费、研究与发展课题、研究与发展成果（专著）、获奖情况方面都存在显著的差异。

从输出结果 10—5 中各变量的平均水平看：

第一类中：社科活动人员平均人数居五类中第五位；研究与发展人员居五类中第四位；研究与发展年平均投入经费居五类中最末位；研究与发展平均课题数居五类中最末位；研究与发展成果（专著）平均数居五类中最末位；获奖成果平均数居五类中第四位。

第二类中：社科活动人员居五类之首位；研究与发展人员居五类之首位；研究与发展年投入经费居五类中第二位；研究与发展课题居五类中第二位；研究与发展成果（专著）居五类中之首位；获奖成果居五类之首位。

输出结果 10—5　　案例数据聚类后各类均值结果

Descriptives

		N	Mean			N	Mean
L1	1	66	333.909	L16	1	66	38.455
	2	1	1889.000		2	1	397.000
	3	11	1034.182		3	11	329.636
	4	1	437.000		4	1	126.000
	5	1	1532.000		5	1	577.000
	Total	80	465.900		Total	80	90.800
L7	1	66	80.848	L20	1	66	22.000
	2	1	883.000		2	1	689.000
	3	11	466.636		3	11	176.727
	4	1	76.000		4	1	93.000
	5	1	537.000		5	1	183.000
	Total	80	149.563		Total	80	54.513
L13	1	66	5903.591	L25	1	66	3.909
	2	1	94372.000		2	1	44.000
	3	11	58199.636		3	11	43.364
	4	1	136378.000		4	1	3.000
	5	1	63714.000		5	1	20.000
	Total	80	16553.713		Total	80	10.025

第三类中：社科活动人员平均人数居五类中第三位；研究与发展人员平均人年数居五类中第三位；研究与发展年平均投入经费居五类中第四位；研究与发展平均课题数居五类中第三位；研究与发展成果（专著）平均数居五类中第三位；获奖成果平均数居五类中第二位。

第四类中：社科活动人员平均人数居五类中第四位；研究与发展人员平均人年数居五类中第五位；研究与发展年平均投入经费居五类中第一位；研究与发展平均课题数居五类中第四位；研究与发展成果（专著）平均数居五类中第四位；获奖成果平均数居五类中第五位。

第五类中：社科活动人员平均人数居五类中第二位；研究与发展人员平均人年数居五类中第二位；研究与发展年平均投入经费居五类中第三位；研究与发展平均课题数居五类中第一位；研究与发展成果（专著）平均数居五类中第二位；获奖成果平均数居五类中第三位。

这五类的各变量的排名次序可以用图 10—8 描述。

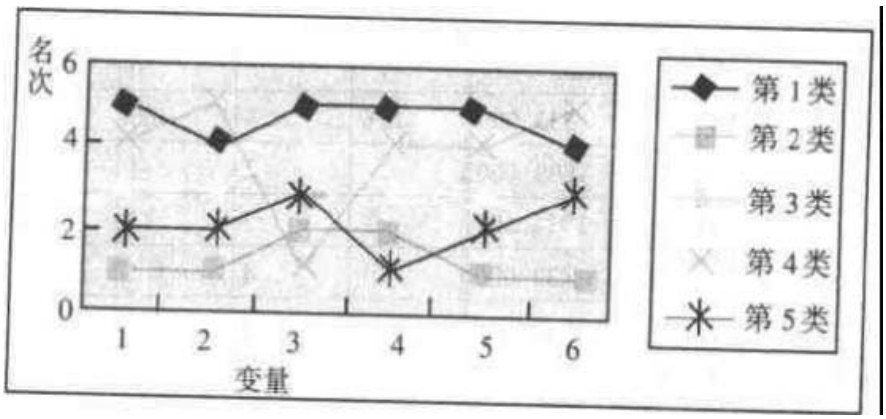


图 10—8　案例数据按层次聚类法聚类后各类的特征

2. 快速聚类分析

指定将案例数据聚成五类，结果与层次聚类法有一些差别。

第一类中少了一所学校；层次聚类分析中的第三、四、五类重新划分成现在的三类，第二类是：武汉大学、吉林大学、北京师范大学等 8 所；第三类是北京大学和清华大学 2 所；第四类是南开大学、复旦大学、南京大学、华东师范大学四所；第五类是中国人民大学，是层次聚类分析中的第二类。

由于篇幅所限，具体的分析结果略去。

从上面对案例数据的分析可以发现，采用不同的聚类方法产生的聚类结果会出现差异，有时差异还比较明显。因此，在实际分析时，应注意多尝试几种方法，综合分析，以最终得到比较科学合理的分析结果。另外，分类数目的确定也是一个需要关注的问题。分类数的确定尚没有统一的规范标准，需要在实践中不断探索，反复分析以确定一个合理的分类数目。一般的原则可以从类间距离入手。假如：从五类聚成四类、四类聚成三类、三类聚成两类、两类聚成一类时，距离变化较大，那么，合理的分类数应在二至五类之间。

三、对 1998 年全国各地区房地产开发建设数据作层次聚类分析

表 10—5 是 1998 年全国各地区房地产开发建设的数据。现根据这些指标，对全国 31 个省市自治区的建设情况作层次聚类分析。指定样本间的距离计算欧氏距离、小类间距离以组间平均链锁法计算，对样本数据作 Z 化处理。

通过观察凝聚状态表，决定聚成 6 类。它们分别是：

第二类：海南

第三类：贵州、甘肃、宁夏、新疆

第四类：西藏

表 10—5　　1998 年全国各地区房地产开发建设数据

地区	施工房屋面积(万平方米)(A)	竣工房屋面积(万平方米)(B)	房屋建筑面积竣工率(%)(C)	竣工房屋价值(万元)(D)	竣工房屋造价(元/平方米)(E)
北京	3 499.06	842.81	24.1	1 927 910	2 287
天津	1379.77	428.74	31.1	633 277	1 477
河北	1 169.47	480.78	41.1	397 584	827
山西	528.37	205.72	38.9	162 817	791
内蒙古	368.46	178.74	48.5	150 360	841
辽宁	2 371.68	1 007.7	42.5	924 671	918
吉林	556.04	281.25	50.6	229 109	815
黑龙江	907.84	471.35	51.9	381 370	809
上海	5 416.1	1 565.34	28.9	3 936 445	2 515
江苏	3 809.78	1 824.62	47.9	1 718 789	942
浙江	3 420.43	1 403.2	41	1 620 532	1 155
安徽	1 228.58	554.16	45.1	352 997	637
福建	2 748.79	578.74	21.1	692 656	1 197
江西	679.84	251.93	37.1	191 222	759
山东	2 396.44	985.06	41.1	781 947	794
河南	1 175.96	392.03	33.3	284 124	725
湖北	1 857.22	735.68	39.6	846 667	1 151
湖南	890.74	328.19	36.8	267 942	816
广东	8 130.79	2 319.09	28.5	3 610 833	1 557
广西	662.74	240.15	36.2	198 957	828
海南	147.64	24.92	16.9	28 756	1 154
重庆	2 058.35	598.63	29.1	510 767	853
四川	2 140.35	756.95	35.4	530 641	701
贵州	478.27	145.91	30.5	107 895	739
云南	767.88	299.7	39	352 179	1 175
西藏	21.36	5.08	23.8	4 718	929
陕西	957.79	262.6	27.4	223 471	851
甘肃	389.74	121.26	31.1	97 658	805
青海	150.13	57.39	38.2	48 124	839
宁夏	207.91	124.6	59.9	84 846	681
新疆	252.61	94.28	37.3	92 663	983

第五类：青海

其余各省市自治区为第一类。

这五类的特征如输出结果 10—6 所示。指标 A、B、C、D、E 分别依次对应表 10—5 的指标，名次为各类相应指标的降序排序结果。

输出结果 10—6　　1998 年各地区房地产开发建设聚类分析结果

类别	指标	N	Mean	名次
1	A	24	2046.7696	1
	B	24	708.0483	1
	C	24	37.3417	3
	D	24	871969.4583	1
	E	24	1059.2083	2
2	A	1	147.6400	4
	B	1	24.9200	4
	C	1	16.9000	5
	D	1	28756.0000	4
	E	1	1154.0000	1
3	A	4	332.1325	2
	B	4	121.5125	2
	C	4	39.7000	1
	D	4	95765.5000	2
	E	4	802.0000	5
4	A	1	21.3600	5
	B	1	5.0800	5
	C	1	23.8000	4
	D	1	4718.0000	5
	E	1	929.0000	3
5	A	1	150.1300	3
	B	1	57.3900	3
	C	1	38.2000	2
	D	1	48124.0000	3
	E	1	839.0000	4

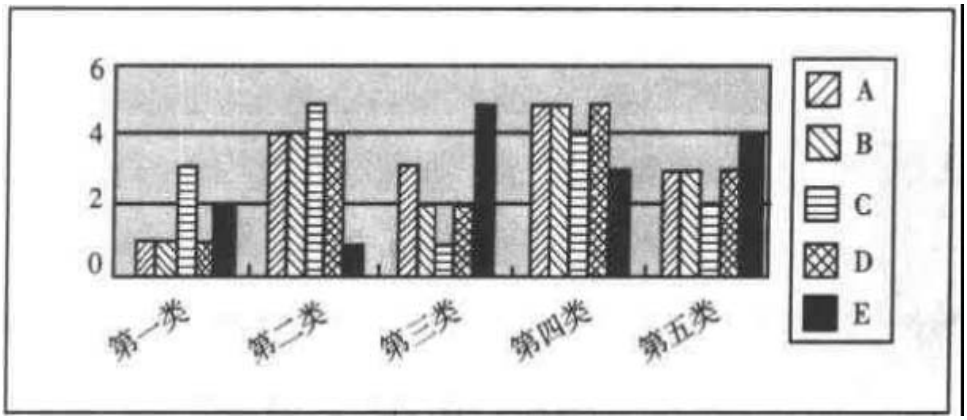


图 10—9　1998 年各地区房地产开发建设聚类分析各类特征

由图 10—9 可以看出：

以北京为代表的第一类的 24 个地区，房屋施工面积、竣工面积最高，但房屋建筑面积竣工率处于中游水平。竣工房屋价值和房屋造价比较高，分别排名第一位和第二位；

海南为第二类地区，它的房屋施工房屋面积、竣工面积均较低，排名第四位，同时房屋建筑面积竣工率最低。竣工房屋价值比较低，但造价最高；

以贵州为代表的第三类地区，它的房屋施工房屋面积、竣工面积均处于中上游水平，但房屋建筑面积竣工率最高。竣工房屋价值比较高，但造价最低；

西藏为第四类地区，它的房屋施工房屋面积、竣工面积最低，竣工率也较低，排名第四位。竣工房屋的价值较低，造价处中游水平；

青海为第五类地区，它的房屋施工房屋面积、竣工面积均处中游水平，竣工率比较高，排名第二位。竣工房屋的价值和造价也处中游。

由此可见，除个别地区之外，全国大部分地区的房地产开发建设水平相差不大。

第 11 章

SPSS 的因子分析

§11.1 什么是因子分析

在实际问题的分析过程中，人们往往希望尽可能多地收集关于分析对象的数据信息，进而能够对它有比较全面、完整的把握和认识。于是，对某个分析对象的描述就会有许多指标。例如：对高等学校科研情况的研究，需要收集诸如：学校教师人数、每年承担多少科研项目、在科研方面投入了多少人年、学校科研项目的经费情况、完成了多少课题项目、获得了多少奖励等几十项指标；对一名学生学习成绩的评价，需要收集诸如：高等数学、高等代数、英语、政治、各门专业课程、选修课程等几十门课程的成绩；即使是做一件上衣，也需要测量诸如身长、袖长、胸围、肩宽、背厚等许多尺寸数据。尽管人们花费了许多力气收集了如此多的指标，但效果如何呢？不难看出，如果收集的变量指标非常多，虽然能够比较全面地对事物有精确的度量，但却给实际的统计分析工作带来了比较大的问题。

首先，收集的变量太多会增加分析中的计算工作量，从而使本来不很复杂的分析工作变得异常烦琐。其次，花费许多人力和财力收集到的变量，由于它们之

间的相关性，造成相当多的信息重叠现象。信息重叠也会给统计分析工作带来麻烦。如：在进行多元回归分析中，自变量高度的多重共线性会给回归方程的参数估计带来许多问题。于是，人们自然想到是否可以通过简单地减少相应指标的办法来解决上述问题，但这必然又造成某些重要信息的丢失。

因子分析正是解决上述问题的一种非常有效的方法。它以最少的信息丢失，将原始的众多指标综合成较少的几个综合指标，这些综合指标称为因子变量。

因子变量的特点是：第一，因子变量的数量远少于原有指标变量的数量，对因子变量的分析能够减少分析中的计算工作量；第二，因子变量并不是原有变量的简单取舍，而是对原始变量的重新组构，它们能够反应原有众多指标的绝大部分信息，不会产生重要信息的丢失问题；第三，因子变量之间有没有线性相关关系，对因子变量的分析能够为研究工作提供较大的便利；第四，因子变量具有命名解释性。因子的命名解释性可以理解为某个因子变量是对某些原始变量的综合，它能够反应这些原始变量的绝大部分信息。例如：对于高等学校科研情况的因子分析，可能得到两个因子，其中一个因子是对诸如科研人力投入、经费投入、申报项目数等指标的综合。另一个因子是对完成课题数目以及获奖等指标的综合。于是，我们说这两个因子命名解释性体现在：其中一个因子是对科研投入情况的反应，可命名为科研投入因子，另一个因子是对科研产出情况的反应，可命名为科研产出因子。因子的命名解释性有助于对因子分析结果的解释评价，具有实际意义。

总之，因子分析是研究如何以最少的信息丢失，将众多原始变量浓缩成少数几个因子变量，以及如何使因子变量具有较强的可解释性的一种多元统计分析方法。

§11.2 因子分析的数学模型和相关概念

在对因子分析的基本方法作详细讨论之前，我们首先对因子分析的数学模型和相关概念进行讨论。

一、因子分析的数学模型和相关概念

因子分析的出发点是用较少的互相独立的因子变量来代替原有变量的绝大部分信息，可以将这一思想用一个数学模型来表示。假设原有变量有 p 个，分别

用 x_1，x_2，x_3、…，x_p 表示，其中 x_i（$i=1$，2，…p）是均值为零、标准差为1的标准化变量，F_1，F_2，F_3、…F_m 分别表示 m 个因子变量，m 应小于 p。于是有：

$$\begin{aligned} x_1 &= a_{11}F_1 + a_{12}F_2 + \cdots + a_{1m}F_m + a_1\varepsilon_1 \\ x_2 &= a_{21}F_1 + a_{22}F_2 + \cdots + a_{2m}F_m + a_2\varepsilon_2 \\ &\cdots \\ x_p &= a_{p1}F_1 + a_{p2}F_2 + \cdots + a_{pm}F_m + a_p\varepsilon_p \end{aligned} \tag{11.1}$$

也可以矩阵的形式表示为：

$$X = AF + a\varepsilon \tag{11.2}$$

在这个数学模型中，F 为因子变量或公共因子，可以把它们理解为在高维空间中的互相垂直的 m 个坐标轴；A 为因子载荷矩阵，a_{ij}称为因子载荷，是第 i 个原有变量在第 j 个因子变量上的负荷。如果把变量 x_i 看成 m 维因子空间中的一个向量，则 a_{ij}表示 x_i 在坐标轴 F_j 上的投影，相当于多元回归分析模型中的标准回归系数；ε 为特殊因子，表示了原有变量不能被公共因子所解释的部分，相当于多元回归分析模型中的残差项。

二、因子分析的几个相关概念

根据因子分析的数学模型，可以引入以下几个基本概念，我们对它们的统计含义稍作解释。

1. 因子载荷

在各公共因子不相关的前提下，因子载荷 a_{ij}的就是第 i 个原有变量与第 j 个公共因子的相关系数，即：表示 x_i 依赖 F_i 的比重，反应了第 i 个原有变量在第 j 个公共因子上的相对重要性。因此，a_{ij}的绝对值越大，则公共因子 F_j 与原有变量 x_i 的关系越强。

2. 变量共同度

变量共同度也称为公共方差。原有变量 x_i 的共同度定义为因子载荷矩阵 A 中第 i 行元素的平方和，即：

$$h_i^2 = \sum_{j=1}^{m} a_{ij}^2 \tag{11.3}$$

在原有变量 x_i 标准化的前提之下，x_i 的方差可以表示成：$h_i^2 + \varepsilon_i^2 = 1$。因此，可以认为，原有变量 x_i 的方差由两个部分组成：第一部分为变量共同度 h_i^2，反应了全部公共因子变量对原有变量 x_i 总方差解释说明的比例，体现了公

共因子对原有变量 x_i 的贡献程度。可见，h_i^2 越接近 1，说明公共因子已经解释说明了原有变量 x_i 的几乎全部信息；第二部分是特殊因子 ε_i^2，它反应了原有变量方差中无法被公共因子刻画的比例。

3. 公共因子 F_j 的方差贡献

公共因子 F_j 的方差贡献定义为因子载荷矩阵 A 中第 j 列各元素的平方和，即：

$$S_j = \sum_{i=1}^{p} a_{ij}^2 \tag{11.4}$$

可见，公共因子 F_j 的方差贡献反映了因子 F_j 对所有原始变量总方差的解释能力，其值越高，说明该因子的重要程度越高。

总之，某变量的共同度刻画了所有公共因子对该变量信息的解释程度，进而可以把握该变量的信息有多少被丢失了。如果大部分变量的共同度都高于 0.8，则说明提取出的公共因子已经基本反映了各原始变量 80% 以上的信息，仅有较少的信息丢失，因子分析效果较好。因此，各个变量的共同度是衡量因子分析效果的一个指标。另外，某公共因子的方差贡献反映了该公共因子对原所有变量信息的解释说明能力，是一个衡量公共因子重要程度的指标。

§11.3 因子分析的基本步骤

因子分析的核心问题有两个：一是如何构造因子变量；二是如何对因子变量进行命名解释。因此，因子分析的基本步骤和解决思路就是围绕这两个核心问题展开的。

因子分析常常有以下四个基本步骤：

(1) 确认待分析的原有若干变量是否适合作因子分析。

(2) 构造因子变量。

(3) 利用旋转方法使因子变量更具有可解释性。

(4) 计算因子变量得分。

下面分别对各步进行讨论。

一、确定原有变量是否适合进行因子分析

因子分析的目的，是从原有众多的变量中综合出少量具有代表意义的因子变量，这必定有一个潜在的前提要求，即原有变量之间应具有较强的相关关系。不

难理解，如果原有变量之间不存在较强的相关关系，那么根本无法从中综合出能够反映某些变量共同特性的几个较少的公共因子变量来。因此，一般在因子分析时，需要对原有变量进行相关分析。最简单的方法是计算变量之间的相关系数矩阵并进行统计检验。如果相关系数矩阵中的大部分相关系数都小于0.3且未通过统计检验，那么，这些变量就不适合作因子分析。除此之外，SPSS还提供了几种帮助判断变量是否适合作因子分析的统计检验方法。

1. 巴特利特球度检验（Bartlett test of sphericity）

巴特利特球度检验是以变量的相关系数矩阵为出发点。它的零假设是 H_0：相关系数矩阵是一个单位阵，即相关系数矩阵对角线上的所有元素都为1，所有非对角线上的元素都为零。巴特利特球度检验的统计量根据相关系数矩阵的行列式计算得到。如果该统计量值比较大，且其对应的相伴概率值小于用户心中的显著性水平，则应拒绝 H_0，认为相关系数矩阵不太可能是单位阵，适合作因子分析；相反，如果该统计量值比较小，且其对应的相伴概率值大于用户心中的显著性水平，则不能拒绝 H_0，可以认为相关系数矩阵可能是单位阵，不适合作因子分析。

2. 反映象相关矩阵（Anti－image correlation matrix）检验

反映象相关矩阵检验以变量的偏相关系数矩阵为出发点，将偏相关系数矩阵的每个元素取反，得到反映象相关阵。不难理解，由于偏相关系数是在控制了其他变量对两变量影响的条件下计算出来的净相关系数，如果变量之间确实存在较强的相互重叠传递影响，也就是说，如果变量中确实能够提取出公共因子，那么控制了这些影响后的偏相关系数必然很小。因此，如果反映象相关矩阵中的有些元素的绝对值比较大，则说明这些变量可能不适合作因子分析。

3. KMO（Kaiser－Meyer－Olkin）检验

KMO 统计量是用于比较变量间简单相关系数和偏相关系数的一个指标，计算公式如下：

$$KMO = \frac{\sum\sum_{i\neq j} r_{ij}^2}{\sum\sum_{i\neq j} r_{ij}^2 + \sum\sum_{i\neq j} p_{ij}^2} \tag{11.5}$$

式中：r_{ij}是变量 i 和变量 j 之间的简单相关系数，p_{ij}是它们之间的偏相关系数。可见，*KMO* 统计量的取值在0和1之间，当所有变量之间的简单相关系数平方和远远大于偏相关系数平方和时，*KMO* 值接近1。*KMO* 值越接近1，则越适合作因子分析，*KMO* 越小，则越不适合作因子分析。Kaiser给出了一个 *KOM* 的度量标准：0.9以上非常适合；0.8适合；0.7一般；0.6不太适合；0.5以下

不适合。另外，在SPSS中还按下列公式计算了另一个与KMO类似的指标，结果显示在反映象相关矩阵的对主角线上。

$$MSA = \frac{\sum_{j\neq i} r_{ij}^2}{\sum_{j\neq i} r_{ij}^2 + \sum_{j\neq i} p_{ij}^2} \tag{11.6}$$

可见，这个指标的求和项中并没有包含所有变量的相关系数或偏相关系数，而只是包含了涉及到的那些变量。

二、确定因子变量和计算因子载荷矩阵

构造因子变量是因子分析的关键步骤之一。因子分析中有许多确定因子变量的方法，如：基于主成分模型的主成分分析法和基于因子分析模型的主轴因子法、极大似然法、最小二乘法等。这里，我们仅对在因子分析中占有主要地位的、使用较为广泛的主成分分析法作简单介绍。

主成分分析法能够为因子分析提供初始解，因子分析也正是在继承了主成分分析结果的基础之上作进一步的分析研究的。

主成分分析法通过坐标变换手段，将原有的 p 个相关变量 x_i 作线性变换，转换成为另一组不相关的变量 y_i，可以表示为：

$$\begin{aligned} y_1 &= u_{11}x_1 + u_{21}x_2 + \cdots + u_{p1}x_p \\ y_2 &= u_{12}x_1 + u_{22}x_2 + \cdots + u_{p2}x_p \\ &\cdots\cdots \\ y_p &= u_{1p}x_1 + u_{2p}x_2 + \cdots + u_{pp}x_p \end{aligned} \tag{11.7}$$

该方程组要求：

$$u_{1k}^2 + u_{2k}^2 + \cdots + u_{pk}^2 = 1 \quad (k = 1, 2, 3, \cdots p)$$

式中，系数 u_{ij} 依照两个原则来确定：

（1）y_i 与 y_j（$i \neq j$，i，$j = 1$，2，3，$\cdots p$）互不相关；

（2）y_1 是 x_1，x_2，x_3，$\cdots$，x_p 的一切线性组合（系数满足上述方程组）中方差最大的；y_2 是与 y_1 不相关的 x_1，x_2，x_3，$\cdots$，x_p 的一切线性组合中方差最大的；y_p 是与 y_1，y_2，y_3，$\cdots y_{p-1}$ 都不相关的 x_1，x_2，x_3，$\cdots$，x_p 的一切线性组合中方差最大的；

我们分别将按照这样的原则确定的变量 y_1、y_2、y_3、$\cdots$、y_p 称为原有变量的第一、第二、第三、$\cdots$、第 p 主成分。其中 y_1 在总方差中占的比例最大，它综合原有变量 x_1，x_2，x_3，$\cdots$，x_p 的能力最强，其余主成分 y_2、y_3、$\cdots$、y_p 在总方差中占的比例依次递减，也就是说，其余主成分 y_2，y_3，$\cdots$，y_p 综合原

有变量的能力依次减弱。在主成分分析的实际应用中，为达到减少变量个数的目的，一般只选取前面几个方差最大的主成分。这样既减少了变量的数目，又能够用较少的主成分反映原有变量的绝大部分信息。

可见，主成分分析中关键的步骤就是如何求出上述方程组中的系数 u_{ij}。通过公式的计算推导可以发现，每个方程中的系数向量 u_i（u_{1i}，u_{2i}，u_{3i}，…，u_{pi}）恰好是原有变量相关系数矩阵 R 的特征值对应的特征向量。因此，为得到主成分，首先应求出 R 的特征值和特征向量。具体的主成分求解过程，大家可以参阅其他书目。这里，我们只讨论主成分分析为因子分析提供初始解的主要步骤（其他步骤略去）。

（1）将原有变量数据作标准化处理；

（2）计算变量间简单相关系数矩阵 R；

（3）求相关系数矩阵 R 的特征值 $\lambda_1 \geqslant \lambda_2 \geqslant \lambda_3 \geqslant \cdots \lambda_p \geqslant 0$ 及对应的单位特征向量 μ_1，μ_2，μ_3，$\cdots \mu_p$。

通过上述几个步骤，我们可以得到 p 个特征值和它们对应的特征向量，这些就是因子分析的初始解。

现在，我们再回到因子分析中来。因子分析利用主成分分析得到的 p 个特征值和对应的特征向量，在此基础之上，按下列方法计算得到因子载荷矩阵：

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \cdots & \cdots & \cdots & \cdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{pmatrix}=\begin{pmatrix} u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} & \cdots & u_{1p}\sqrt{\lambda_p} \\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \cdots & u_{2p}\sqrt{\lambda_p} \\ \cdots & \cdots & \cdots & \cdots \\ u_{p1}\sqrt{\lambda_1} & u_{p2}\sqrt{\lambda_2} & \cdots & u_{pp}\sqrt{\lambda_p} \end{pmatrix} \tag{11.8}$$

由于因子分析的目的是减少变量个数，因此，在计算因子载荷矩阵时，一般不选取所有特征值，而只选取前 m 个特征值和对应的特征向量，得到下面的有 m 个公共变量的因子载荷矩阵：

$$A=\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{pmatrix}=\begin{pmatrix} u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} & \cdots & u_{1m}\sqrt{\lambda_m} \\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \cdots & u_{2m}\sqrt{\lambda_m} \\ \cdots & \cdots & \cdots & \cdots \\ u_{p1}\sqrt{\lambda_1} & u_{p2}\sqrt{\lambda_2} & \cdots & u_{pm}\sqrt{\lambda_m} \end{pmatrix} \tag{11.9}$$

式中：$m<p$。确定 m 可以通过两种方法：

一是，根据特征值 λ_i 确定。一般取大于 1 的特征值；

二是，根据因子的累计方差贡献率确定。

因子的累计方差贡献率的计算方法是：

第一个公共因子的方差贡献率为：

$$C = \lambda_1 / \sum_{i=1}^{p} \lambda_i \tag{11.10}$$

即为：第一个公共因子的方差贡献 S_1/p（有 p 个标准化的原始变量，总方差即为 p）。

第一、第二两公共因子的累计方差贡献率为：

$$C = (\lambda_1 + \lambda_2) / \sum_{i=1}^{p} \lambda_i \tag{11.11}$$

即为：$(S_1 + S_2)/p$

前 m 个公共因子的累计方差贡献率为：

$$C = \sum_{i=1}^{m} \lambda_i / \sum_{i=1}^{p} \lambda_i \tag{11.12}$$

即为：$(S_1 + S_2 + \cdots + S_m)/p$

可以根据实际需要确定累计贡献率的大小，一般累计贡献率应在 80% 以上。

三、因子变量的命名解释

因子变量的命名解释是因子分析的另一个核心问题。对上面计算得到因子载荷矩阵中的 a_{ij} 进行观察，一般会发现这样的现象：a_{ij} 的绝对值可能在某一行的许多列上都有较大的取值，或 a_{ij} 的绝对值可能在某一列的许多行上都有较大的取值。这表明：某个原有变量 x_i 可能同时与几个因子变量都有比较大的相关关系。也就是说，某个原有变量 x_i 的信息需要由若干个因子变量来共同解释；同时，虽然一个因子变量可能能够解释许多变量的信息，但它却只能解释某个变量的一少部分信息，不是任何一个变量的典型代表。这样的情况必然使得某个因子变量的实际含义模糊不清。而实际分析工作中，人们却希望对因子变量的含义有比较清楚的认识。因此，我们希望通过某种手段使每个变量在尽可能少的因子上又有比较高的载荷，即：在理想状态下，让某个变量在某个因子上的载荷趋于 1，而在其他因子上的载荷趋于 0。这样，一个因子变量就能够成为某个变量的典型代表，那么它的实际含义也就清楚了。

实现上述目标的方法是对因子载荷矩阵进行旋转。我们可以从载荷散点图上来直观理解旋转的含义，以因子变量为坐标轴绘制原有变量的散点图。经过坐标旋转后，原有变量点应出现在靠近轴的端点和圆点附近。在轴的端点上变量是只在那个因子上有较高载荷的变量，靠近图的圆点的变量对两个因子都具有小的载

荷。当然，不靠近轴的变量是被两个因子共同解释，旋转后应尽可能少地出现这种情况。

旋转的方法有正交旋转法、斜交旋转法等，其中比较常用的是方差极大法(VariMax)。具体方法可以参阅其他有关书目。

四、计算因子得分

因子得分是因子变量构造的最终体现。在因子分析的实际应用中，因子变量确定以后，对每个样本数据，我们总是希望得到它们在不同因子上的具体数据值，这些数值就是因子的得分。有了因子得分，今后的分析研究就可以不再针对原有变量，而是简化为对它们各因子得分变量的研究，从而达到降维的目的。

那么如何计算因子得分呢？这里的一个基本思想是将因子变量表示为原有变量的线性组合，即通过以下的因子得分函数计算因子得分：

$$F_j = \beta_{j1}x_1 + \beta_{j2}x_2 + \beta_{j3}x_3 + \cdots + \beta_{jp}x_p \quad (j = 1, 2, 3, \cdots, m) \tag{11.13}$$

估计因子得分的方法有很多，如：回归法、Bartlette 法、Anderson－Rubin 法等。具体方法可以参阅其他有关书目。

五、SPSS 因子分析操作的基本操作步骤

(1) 菜单选项：Statistics－>Data Reduction－>Factor。于是出现如图 11—1 所示的窗口。

(2) 选择参与因子分析的变量到 Variables 框。

(3) 可以根据某变量的取值选取部分样本作因子分析。

(4) 按 Extraction 钮选择构造因子变量的方法。SPSS 默认的方法是主成分分析法。其中：

● Analyze 框：选择提取因子变量的依据，可以是相关系数矩阵，也可以是协方差阵；

● Extract 框：指定确定因子个数的标准，可以在 Eigenvaluse over 后输入一个特征值，SPSS 将提取特征值大于该值的因子，SPSS 默认为 1。也可以在 Number of factors 后直接输入提取因子的个数；

● Display 框：选择输出那些与因子提取有关的信息：Unrotated factore solution 表示输出未经旋转的因子载荷矩阵，Scree plot 表示输出因子与其特征值的碎石图。

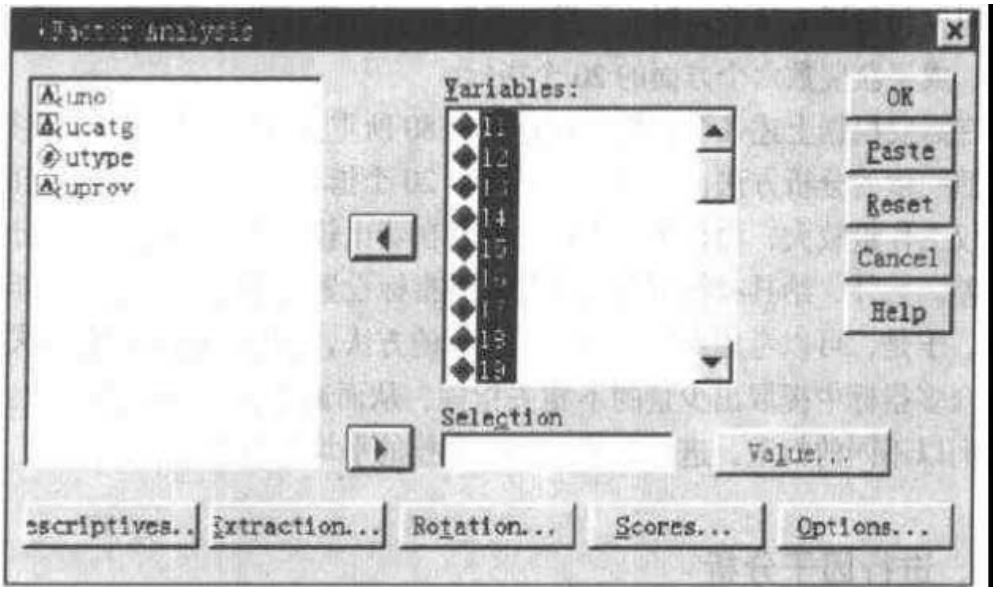


图 11—1　SPSS 因子分析窗口

（5）按Rotation钮选择因子载荷矩阵的旋转方法。SPSS 默认是不进行旋转。一般可以选择 Varimax 选项采用方差极大法旋转。其中：Display 框：指定输出那些与因子载荷矩阵旋转有关的信息，Rotated solution 表示输出旋转后的因子载荷矩阵，Loading plots 表示输出载荷散点图。

（6）按Descriptives钮：在 Statistics 框中选择输出那些相关统计量，其中：

● Univariate descriptives 表示输出各变量的描述统计量，Initial solution 表示输出初始分析结果，是一个中间结果；

● Correlation Matrix 框：提供了几种检验变量是否适合作因子分析的检验方法。

（7）按Scores按钮：选中 Save as variables 项表示：将因子得分保存到一个名为 FACn-m 的 SPSS 变量中，其中：n 是因子变量的名，以数字序号的形式表示；m 表示是第几次作的。选中 Display factor score coefficient matrix 项表示：以矩阵的形式输出因子得分函数。Method 框中提供了估计因子得分的几种方法。

§11.4　案例分析

一、案例说明

查阅《中国普通高等院校人文社会科学统计资料汇编》，得到 1999 年全国 80 所重点院校的人文社会科学研究基本情况数据。这些指标涉及到社科活动人

员数、研究与发展人员数、研究与发展经费数、研究与发展课题数、研究与发展成果数、成果获奖数六个方面的 20 个指标。

现在需要根据上述几个方面的指标，对 80 所重点院校的综合实力进行分析。如果按照一般的分析方法，我们就需要处理 20 个指标，并给它们以不同的权重。这样不仅工作量较大，而且由于指标之间存在着比较高的相关性，会给分析结果带来偏差。另外，给具有较高相关性的众多指标设置权重系数也是一件非常复杂的事情。于是，可以考虑首先采用因子分析的方法，在尽量减少信息丢失的前提下，在众多指标中提取出少量的不相关指标，从而减少分析变量的个数。然后，再定它们以不同的权重，进而计算出各个学校的平均综合实力得分。

二、进行因子分析

选用主成分分析法得到因子分析初始解、方差极大法作因子旋转、回归法计算因子得分。分析结果见输出结果 11—1 所示。

输出结果 11—1 (a)　　案例作因子分析的初始结果

Communalities

	Initial	Extraction
L1	1.000	.912
L2	1.000	.903
L3	1.000	.928
L6	1.000	.880
L7	1.000	.934
L8	1.000	.942
L11	1.000	.860
L12	1.000	.919
L13	1.000	.907
L15	1.000	.836
L16	1.000	.927
L17	1.000	.927
L18	1.000	.890
L19	1.000	.802
L20	1.000	.932
L21	1.000	.850

续输出结果

	Initial	Extraction
L23	1.000	.876
L24	1.000	.865
L25	1.000	.868
L26	1.000	.902

Extraction Method: Principal Component Analysis.

输出结果 11—1（a）的含义：第一列是 20 个原始变量名；第二列是根据因子分析初始解计算出的变量共同度。利用主成分分析方法得到 20 个特征值，它们是因子分析的初始解，利用这 20 个特征值和对应的特征向量计算出因子载荷矩阵。这时，每个原始变量的所有方差都能够被因子变量解释掉，于是，每个原有变量的共同度都是 1；第三列是根据因子分析最终解计算出的变量共同度。根据最终提取的 m 个特征值和对应的特征向量计算出因子载荷矩阵。这时，由于因子变量个数少于原有变量个数，于是，每个变量的共同度必然小于 1。例如：第一行中的 0.912 表示 m 个因子变量总共解释了原变量 L_1 方差的 91.2%。

那么 m 究竟是多少呢？看输出结果11—1(b)。

输出结果 11—1（b）　　案例因子分析的因子提取和旋转结果

Total Variance Explained

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	14.113	70.565	70.567	14.113	70.567	70.567	6.987	34.936	34.936
2	1.464	7.320	77.886	1.464	7.320	77.886	4.777	23.887	58.823
3	1.328	6.638	84.525	1.328	6.638	84.525	3.463	17.316	76.139
4	.955	4.775	89.300	.955	4.775	89.300	2.632	13.161	89.300
5	.559	2.797	92.097						
6	.336	1.678	93.775						
7	.296	1.478	95 253						
8	.235	1.175	96.428						
9	.172	.861	97.290						
10	.166	.829	98.118						
11	.119	.596	98.714						
12	9.000E-02	.450	99.164						
13	5.957E-02	.298	99.462						
14	4.942E-02	.247	99.709						

续输出结果

	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
Component	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
15	2.482E-02	.124	99.833						
16	1.717E-02	8.584E-02	99.919						
17	7.561E-03	3.781E-02	99.957						
18	6.504E-03	3.252E-02	99.990						
19	2.070E-03	1.035E-02	100.000						
20	3.070E-16	1.535E-15	100.000						

Extraction Method: Principal Component Analysis.

输出结果 11—1（b）是案例数据作因子分析后的因子提取和因子旋转结果。各列数据的含义是：

第一列至第四列描述了因子分析的初始解对原有变量总体的刻画情况。第一列是因子分析的 20 个初始解的序号；第二列是因子变量的方差贡献（特征值），它是衡量因子重要程度的指标。例如：第一行的 14.113 表示第一个因子变量刻画了原有变量总方差 20 中的 14.113，它刻画的方差最大，下面各因子刻画的方差依次减少；第三列是各因子变量的方差贡献率，表示该因子刻画的方差占原有变量总方差的比例。例如：第一行的 70.565 即是 14.113/20 的结果；第四列是因子变量的累计方差贡献率，表示前 m 个因子刻画的总方差占原有变量总方差的比例。例如：第二行 77.886 是（14.113 + 1.464）/20 的结果。

第五列至第七列是从初始解中提取了 4 个公共因子后对原变量总体的刻画情况。这是由于分析过程中我们指定了提取 4 个公共因子。各列数据的含义与第二至第四列的含义相同。可见，如果提取 4 个公共因子，那么它们可以描述原变量总方差的 89.3%，大于 80%，可以认为这四个因子基本反映了原变量的绝大部分信息。

第八至十列是旋转后的因子对原变量总体的刻画情况。各列的数据含义同前。

从图 11—2 可以看到：横轴表示公共因子数，纵轴为特征值。当提取 1、2、3 个公共因子时，特征值变化非常明显，当提取 5 个以后的公共因子时，特征值的变化很小，基本趋于平缓。由此说明，提取 3 到 4 个公共因子对原变量信息的刻画有显著的作用，而提取 5 个以上的公共因子对原变量信息的刻画已经无显著贡献，因此，这里提取了 4 个公共因子，基本反映了原变量的绝大部分方差。另外，输出结果 11—1（b）也证明了这点。

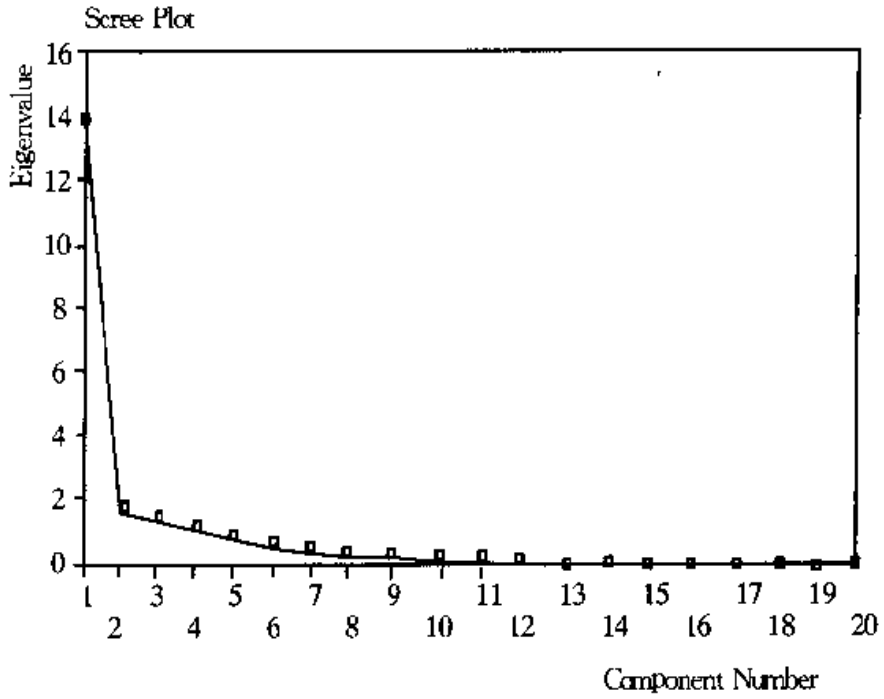


图 11—2　案例公共因子碎石图

输出结果 11—1 (c)　　案例因子分析的因子载荷矩阵

Component Matrix

	Component			
	1	2	3	4
L8	.963	-2.697E-02	-.111	4.645E-02
L3	.960	-2.412E-02	-2.318E-02	-7.090E-02
L7	.954	.128	-5.519E-02	6.296E-02
L2	.939	-5.569E-02	.109	-7.740E-02
L1	.938	-6.107E-02	.146	-8.618E-02
L17	.936	-.201	-8.969E-02	5.649E-02
L12	.930	-.223	-9.069E-03	5.855E-02
L16	.926	-.222	-3.856E-02	-.139
L21	.911	7.962E-02	-.118	-2.179E-02
L20	.900	.337	-4.643E-02	7.689E-02
L18	.864	.115	-.185	-.310
L15	.861	.121	-.198	-.201
L19	.842	.297	-2.939E-02	7.194E-02
L13	.796	4.414E-02	-.303	-.424
L26	.761	-.359	-6.678E-02	.434

续输出结果

	Component			
	1	2	3	4
L25	.701	-.429	-.133	.417
L11	.675	-.139	.615	-8.233E-02
L23	.662	.632	2.860E-02	.194
L24	.589	.559	.284	.353
L6	.461	-.134	.777	-.212

Extraction Method: Principal Component Analysis.

a 4 components extracted.

输出结果 11—1（c）是最终的因子载荷矩阵。由此可以写出如下因子分析的模型：

$$L_8 = 0.963F_1 - 0.0269F_2 - 0.111F_3 + 0.04645F_4$$

$$L_3 = 0.96F_1 - 0.02412F_2 - 0.02318F_3 - 0.0709F_4$$

.....

$$L_6 = 0.461F_1 - 0.134F_2 + 0.777F_3 - 0.212F_4$$

输出结果 11—1（c）中的每个数据表示了相应因子变量对相应原变量的相对重要程度。

输出结果 11—1（d）　　案例作因子分析的因子旋转结果

Rotated Component Matrix

	Component			
	1	2	3	4
L13	.922	.180	.133	8.159E-02
L18	.855	.227	.282	.167
L15	.791	.291	.331	.125
L16	.711	.535	.136	.342
L3	.698	.479	.338	.310
L21	.679	.438	.407	.177
L8	.662	.567	.374	.205
L2	.633	.463	.321	.430
L1	.623	.454	.318	.465
L25	.241	.889	.104	8.956E-02
L26	.248	.882	.200	.150
L17	.619	.656	.232	.247
L12	.580	.654	.226	.321

续输出结果

	Component			
	1	2	3	4
L7	.618	.622	.305	.268
L24	.100	.165	.877	.244
L23	.354	.102	.858	6.350E-02
L20	.599	.335	.659	.161
L19	.553	.322	.604	.167
L6	.122	7.426E-02	.107	.921
L11	.251	.293	.222	.814

Extraction Method: Principal Component Analysis. Rotation Method: Varimax with Kaiser Normalization.

a Rotation converged in 6 iterations.

输出结果 11—1（d）是按照方差极大法对因子载荷矩阵旋转后的结果。从输出结果 11—1（c）中看到，因子变量在许多变量上都有较高了载荷。例如：第一个因子变量，除了变量 L6、L24、L23、L11，几乎在其他所有的变量上都有非常高的载荷数。那么它的实际含义就比较模糊。经过旋转后，第一个因子变量的含义略加清楚，基本反映了研究与发展的人员投入、经费投入和使用情况；第二个因子变量含义比较清楚，基本反映了成果获奖情况；第三个因子变量的实际含义也比较清晰，基本反映了研究与发展的成果应用情况；第四个因子变量反映了辅助人员、人力的投入情况。

输出结果 11—1（e）　　案例作因子分析后因子的协方差阵

Component Score Covariance Matrix

Component	1	2	3	4
1	1.000	-5.820E-16	-1.447E-16	1.773E-16
2	-5.820E-16	1.000	3.016E-16	.000
3	-1.447E-16	3.016E-16	1.000	.000
4	1.773E-16	.000	.000	1.000

Extraction Method: Principal Component Analysis. Rotation Method: Varimax withKaiser Normalization. Component Scores.

输出结果 11—1（c）是因子变量的协方差阵。在前面我们已经谈到，因子变量应是正交、不相关的。从协方差阵上看，其中的数据都非常小，因而可以验证这点。

输出结果 11—1（f）　　案例作因子分析的因子得分函数

Component Score Coefficient Matrix

	Component			
	1	2	3	4
L1	.054	−.013	−.028	.150
L2	.060	−.005	−.025	.122
L3	.098	.002	−.020	.028
L6	−.081	−.153	−.067	.606
L7	.016	.132	−.020	−.014
L8	.048	.089	.021	−.061
L11	−.106	−.042	−.023	.469
L12	.000	.160	−.069	.030
L13	.406	−.225	−.177	−.082
L15	.236	−.122	−.022	−.078
L16	.136	.037	−.160	.061
L17	.026	.160	−.067	−.026
L18	.303	−.191	−.073	−.040
L19	.015	−.035	.214	−.060
L20	.022	−.043	.238	−.077
L21	.097	.002	.046	−.064
L23	−.076	−.101	.452	−.103
L24	−.262	−.002	.510	.037
L25	−.215	.476	−.041	−.123
L26	−.240	.454	.014	−.091

Extraction Method: Principal Component Analysis.

Rotation Method: Varimax with Kaiser Normalization. Component Scores.

输出结果 11—1（f）给出了用回归法计算出的因子得分函数的系数，根据它可以写出以下的因子得分函数：

$$F_1 = 0.054L_1 + 0.06L_2 + 0.098L_3 + \cdots - 0.215L_{25} - 0.24L_{26}$$

$$F_2 = -0.013L_1 - 0.005L_2 + 0.002L_3 + \cdots + 0.476L_{25} + 0.454L_{26}$$

$$F_3 = -0.028L_1 - 0.025L_2 - 0.020L_3 + \cdots - 0.041L_{25} + 0.014L_{26}$$

$$F_4 = 0.15L_1 + 0.122L_2 + 0.028L_3 + \cdots - 0.123L_{25} - 0.091L_{26}$$

SPSS 将根据因子得分函数自动计算各样本的 4 个因子得分。到此，SPSS 因子分析完成。

二、计算综合实力得分

接下来，在因子分析的基础之上，我们对 80 所重点大学进行综合实力分析。

采用的变量不再是原有的 20 个变量，而是经因子分析后得到的 4 个因子变量，利用 4 个因子变量，可以计算出每个学校的综合实力得分。由于 4 个因子变量反映的是人文社会科学研究的不同侧面，因此在计算综合实力时，应给不同的侧面以不同的权数。这些权数一般应由这方面的专家给定，这里，我们仅以这 4 个因子变量的方差贡献率作为权数，于是得到下面的计算公式：

$$综合实力得分 = 0.70567F_1 + 0.0732F_2 + 0.06638F_3 + 0.04775F_4$$

按照这个公式，我们就可以分别计算出 80 所学校的综合实力得分了。

需要说明的是：本案例中，最终得到的因子变量，有的实际含义不很清晰。因此，在利用这些因子变量作进一步分析时，应对这点加以充分考虑。

从最后得分的结果可以看出，综合实力得分较高的是北京大学、中国人民大学等知名的重点大学，这个分析结果与客观实际是吻合的，说明分析方法的运用比较合理。由于篇幅所限，我们略去最后综合得分的具体结果。