SPSS FOR WINDOWS

简明教程

二〇〇二年三月

目　　录

第一章 SPSS 的安装与概貌

第一节 SPSS 的安装

1.1.1 SPSS 简介

SPSS 的全称是：Statistical Program for Social Sciences，即社会科学统计程序。该软件是公认的最优秀的统计分析软件包之一。SPSS 原是为大型计算机开发的，其版本为 SPSSx，80 年代初，微机开始普及以后，它率先推出了微机版本（版本为 SPSS/PC+ x.x），占领了微机市场，大大地扩大了自己的用户量，我国目前正在使用的用户中，绝大部分是使用 3.0—4.0 版本。

80 年代末，Microsoft 发表 Windows 后，SPSS 迅速向 Windows 移植。至 1993 年 6 月，正式推出 SPSS for Windows 6.0 版本。该版本不仅修正了以前版本的错误，改写一些模块使运行速度大大提高。而且根据统计理论与技术的发展，增加了许多新的统计分析方法，使之功能日臻完善。与以往的 SPSS for DOS 版本相比，SPSS for Windows 显得更加直观易用。首先，它采用现今广为流行的电子表格形式作数据管理器，使用户变量命名、定义数据格式、数据输入与修改等过程一气呵成，免除了原 DOS 版本在文本方式下数据录入的诸多不便；其次，采用菜单方式选择统计分析命令，采用对话框方式选择子命令，简明快捷，无需死记大量繁冗的语法语句，这无疑是计算机操作的一次解放；第三，采用对象连接和嵌入技术，使计算结果可方便地被其他软件调用，数据共享，提高工作效率。

作为统计分析工具，理论严谨、内容丰富，数据管理、统计分析、趋势研究、制表绘图、文字处理等功能，几乎无所不包。本使用指导以 SPSS for Windows 6.0 为蓝本，以医学领域的相关资料为例子，简单明了地介绍它的具体使用方法。

1.1.2 SPSS 的安装

SPSS for Windows 6.0 共有 7 个部分，包括：Base、Pro.Stats、Adv.Stats、Tables、Trends、Categeries 和 LISREL。具体内容介绍如下，用户可根据自身需求选择性安装，这样既节省硬盘空间，又方便使用。

Base system（基本统计系统）	
ACF（时间序列研究中的自动相关分析）	97K
Aggregate（数据文件的汇总）	106K
Anova（方差分析）	137K
Autorecode（变量自动赋值处理）	49K
Correlations（相关分析）	73K

Crosstabs（列联表处理）	302K
Curvefit（11 种曲线模型的拟合）	125K
Date（变量定义与数据录入）	155K
Descriptives（均数、标准差等的描述性统计及 Z-分数转换）	79K
Examine（数值分布形式的探究）	290K
Fit（定义程序运行条件）	94K
Flip（数据行列转换）	44K
Frequencies（频数表分析）	121K
Graph（统计图制作）	219K
List（原始数据显示）	52K
Matrix Data（数据的矩阵处理）	81K
Mconvert（矩阵转化）	42K
Means（均数及均数差别的显著性检验）	140K
Mult Response（多变量数据的处理）	90K
Nonpar Corr（非参数资料的相关分析）	80K
Npar Tests（非参数检验）	199K
Oneway（单因素方差分析）	160K
Partial Corr（偏相关分析）	90K
Plot（曲线绘制）	118K
Rank（等级排序、计算正态分数、百分比等分析）	57K
Regression（回归分析）	453K
Report（结果输出）	226K
Sort（数据排序）	43K
SP Chart（高分辨率的统计制图）	94K
Sysfile Info（显示 SPSS 格式的系统文件信息）	35K
TS Plot（时间序列资料的统计制图）	190K
T-Test（t-检验）	77K
基本统计系统共需硬盘空间	4.1 M

Professional Statistics option（专业统计系统）	
Alscal（利用最小二乘法处理多等级测量资料）	404K
Cluster（聚类分析）	166K
Discriminant（判别分析）	435K
Factor（因子分析）	296K
Proximities（资料相似性分析）	117K
Quick Cluster（快速聚类分析）	104K

Reliability（可靠性分析）	164K
2SLS（两级最小二乘法分析）	107K
WLS（加权最小二乘法分析）	94K
专业统计系统共需硬盘空间	1.9 M

Advanced Statistics option（高级统计系统）	
Cox Regression（Cox 回归模型）	374K
Hiloglinear（多因子系统模式的对数线性模型）	155K
Kaplan-Meier（Kaplan-Meier 生存时间模型）	160K
Loglinear（对数线性模型及最优化检验）	207K
Logistic（Logistic 模型）	351K
Manova（协方差分析）	738K
Matrix（高级矩阵转换）	490K
Nonlinear（非线性分析）	147K
Probit（依照所需概率作拟合最优化分析）	134K
Survival（寿命表方式的生存分析）	178K
高级统计系统共需硬盘空间	2.9 M

Tables option（制表系统）	
共需硬盘空间	1.0 M

Trends option（趋势分析系统）	
Arima（Arima 时间序列分析）	332K
Exsmooth（指数平滑拟合）	123K
Model Name（定义程序运行过程需调用的模块）	58K
Season（季节模型）	60K
Spectra（光谱时间序列分析）	138K
X11 Arima（X11 Arima 时间序列分析）	435K
趋势分析系统共需硬盘空间	1.1 M

Categories option（项目分类分析系统）	
本系统只提供键盘录入式的语法命令，共需硬盘空间	0.99 M

LISREL option（线性结构方程式模型分析系统）	
本系统只提供键盘录入式的语法命令，共需硬盘空间	0.64 M

SPSS 的安装步骤：

1、启动 Windows，在程序管理器中选“文件”菜单的“运行”项，弹出“运行”对话框，点击“浏览...”钮，根据安装盘所在的驱动器（A：或 B：或光盘）及其路径，找到 SPSSINST.EXE 文件，点击“确定”钮返回“运行”对话框，再点击“确定”钮，即运行安装程序。

2、安装程序运行后，出现安装选项对话框（如图 1.1 所示）。用户可根据自己的需要选择欲安装的模块：即在所需的模块名前“☐”内点击，使“☐”内出现“☒”表明选中；若再点击使“☒”转为“☐”表明取消选择。选择完毕后点击 OK 钮。

3、指定安装的目标盘和安装文件的路径。

4、输入软件系列号码、用户姓名和单位名称。

5、根据安装过程的提示，依次顺序插换原盘直至安装完成。

最小安装大约需要 15M 硬盘空间（含必需中心系统 14.2M 和求助系统 1.2M），完全安装大约需要 28M 硬盘空间。

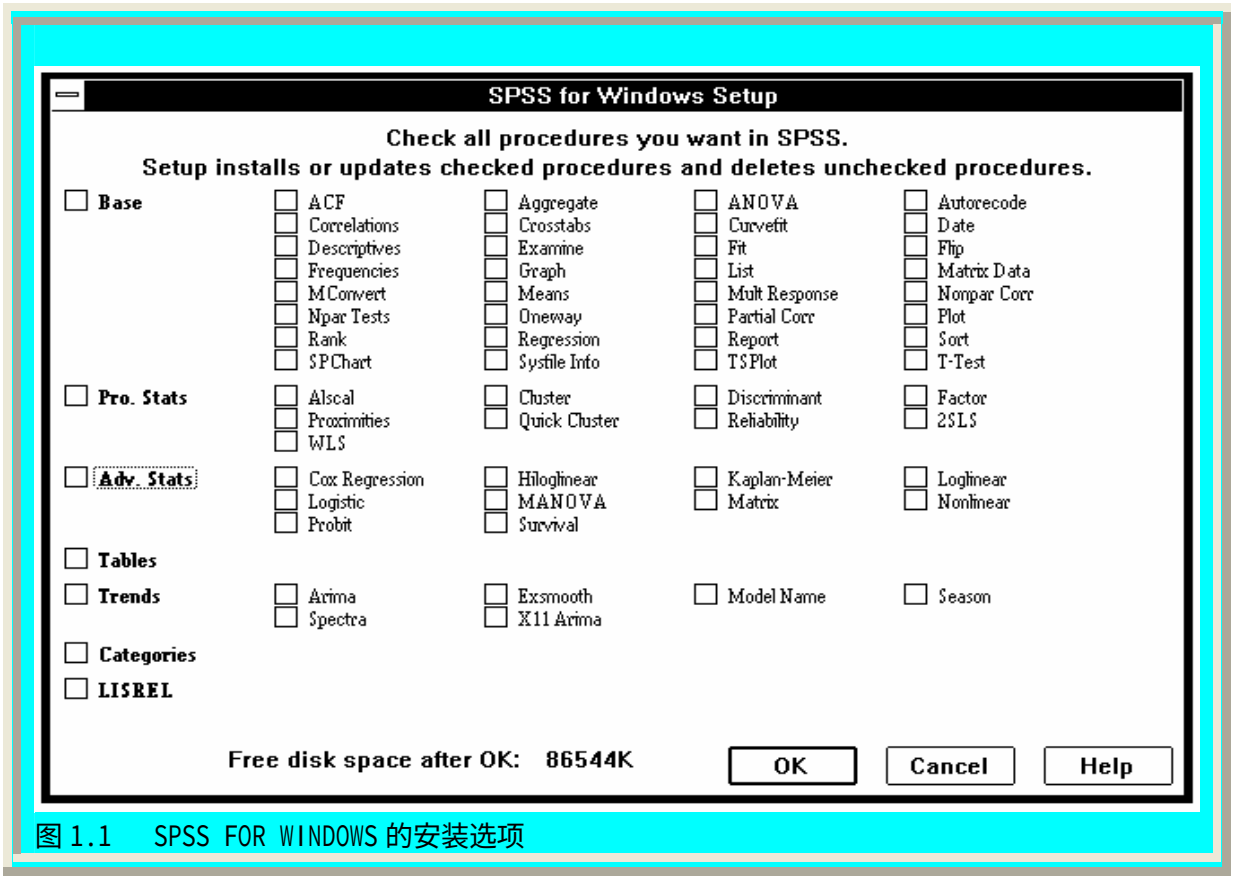


图 1.1　SPSS FOR WINDOWS 的安装选项

第二节 SPSS 的界面初识

1.2.1 SPSS 的启动

在 Windows 的程序管理器中双击 SPSS FOR WINDOWS 图标以打开 SPSS 程序组，选择 SPSS 图标并双击之，即可启动 SPSS。SPSS 启动成功后出现 SPSS 的封面及主窗口，5 秒钟后或点击鼠标左键，封面消失，呈现 SPSS 的预备工作状态（如图 1.2 所示）。

1.2.2 SPSS 的主窗口

SPSS 的主窗口名为 SPSS for Windows，此为窗口的标题栏，当它呈蓝底白字时，表示该窗口为活动窗口，意即用户可对之进行操作。非活动窗口的标题栏呈白底黑字，用户对之不能操作。激活窗口的方法是点击该窗口的标题栏。

标题栏的左侧（即窗口的左上角）为窗口控制钮，点击它选择窗口的还原、移动、大小变换、最小化、最大化、关闭和与其它窗口的切换。标题栏右侧（即窗口右上角）的两个钮：箭头向下的为最小化钮，点击它使窗口缩小为图标（但不是关闭窗口）；箭头向上的为最大化钮，点击它使窗口充满整个屏幕。

该窗口的底部为系统状态栏，显示系统即刻的工作状况，这对用户了解系统情况十分有益。

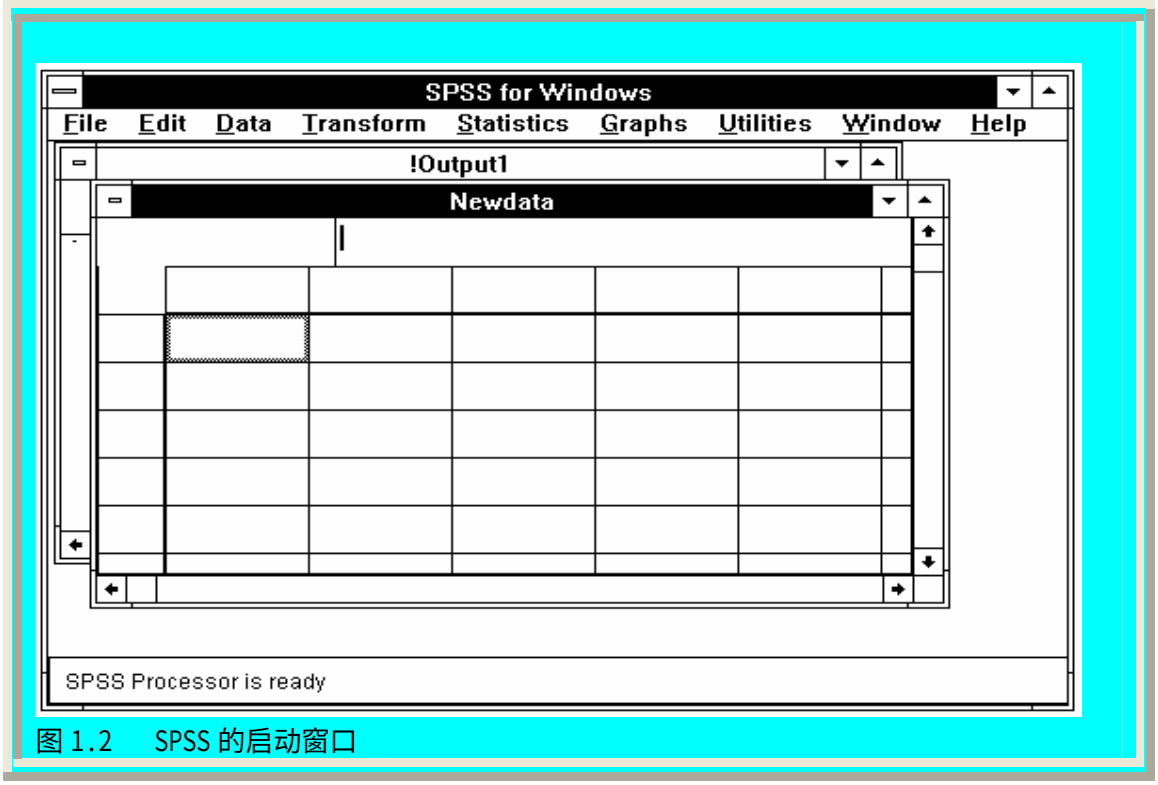


图 1.2　SPSS 的启动窗口

1.2.3 SPSS 的菜单

菜单栏共有 9 个选项：

1、File：文件管理菜单，有关文件的调入、存储、显示和打印等；

2、Edit：编辑菜单，有关文本内容的选择、拷贝、剪贴、寻找和替换等；

3、Data：数据管理菜单，有关数据变量定义、数据格式选定、观察对象的选择、排序、加权、数据文件的转换、连接、汇总等；

4、Transform：数据转换处理菜单，有关数值的计算、重新赋值、缺失值替代等；

5、Statistics：统计菜单，有关一系列统计方法的应用；

6、Graphs：作图菜单，有关统计图的制作；

7、Utilities：用户选项菜单，有关命令解释、字体选择、文件信息、定义输出标题、窗口设计等；

8、Windows：窗口管理菜单，有关窗口的排列、选择、显示等；

9、Help：求助菜单，有关帮助文件的调用、查寻、显示等。

点击菜单选项即可激活菜单，这时弹出下拉式子菜单，用户可根据自己的需求再点击子菜单的选项，完成特定的功能。

1.2.4 SPSS 的其他窗口

在 SPSS 的主窗口中还有两个窗口，一个是数据管理窗口，其标题名称是“Newdata”，且默认为激活状态。数据管理器是一种典型的电子表格形式，用户可通过定义变量名、格式化数据类型后输入原始数值，并可根据需要对数据进行增删、剪贴、修改、存储等操作。

另一个是结果输出窗口，标题名称是“!Output1”，启动时为非活动窗口，只有当完成一项处理后，才在该窗口显示处理过程提示和计算结果。

当进行某项具体的统计方法操作时，可点击对话框的“Paste”钮激活命令编辑窗口，其标题名称是“!Syntax1”，或选 Window 菜单的!Syntax1 项也可激活命令编辑窗口。用户可利用该窗口进行 SPSS 命令的输入、编辑和运行，这对熟悉 DOS 版本的 SPSS 用户是十分方便的。

上述三个窗口在实际操作时，经常因为内容很多，一个窗口中无法看到全部内容。有两种方法可帮助用户看到全部内容：

1、使用窗口的滚动条　　每个窗口的右侧有一个垂直滚动条，用鼠标点击滚动条上下两头的箭号钮或用鼠标按住滚动条中的方块上下拖动，可使窗口中的内容前后翻滚；底边有一个水平滚动条，用鼠标点击滚动条左右两头的箭号钮或用鼠标按住滚动条中的方块左右拖动，可使窗口中的内容左右移动。如此，用户便可看清所有内容。

2、改变窗口的大小　　一般情况下，鼠标指针是一个朝左上方的箭头，当把鼠标指针指向窗口边界时，鼠标指针变成双向箭头形。这时，若按住鼠标左键移动，可改变窗口的大小，同样可看清窗口内容。

1.2.5 SPSS 的退出

完成 SPSS 的统计分析后，退出该系统的方法是：选 File 菜单的 Exit 项，回答系统提出的有关是否需要存储原始数据、计算结果和 SPSS 命令之后，即退到 Windows 的程序管理器中。

1.2.6 SPSS 的求助系统

SPSS 提供了丰富且详尽的在线帮助。主要有下列几种方式：

1、主窗口的 Help 菜单：在软件运行的任何时候，点击 Help 菜单选相关的子菜单，可得到所需的各种帮助。

2、主窗口的 Utilities 菜单：在 Utilities 菜单中，有 Command index...子菜单，它提供有关 SPSS 各项统计分析技术能解决什么问题的信息。

3、各种对话框中的 Help 钮：在具体操作过程中，当弹出某一对话框时，一般总有 Help 钮，点击该钮，用户可得到这一对话框选项内容的详细帮助。

4、结果输出窗口中的 Grossary 钮：当用户在浏览计算结果时，可点击结果输出窗的 Grossary 钮，它显示各种专用统计术语的解释信息以便用户理解。

5、命令编辑窗口中的 Syntax 钮：激活命令编辑窗，可见一 Syntax 钮，点击该钮，可得到与用户正在编辑的命令相关的命令语法提示。

第二章 SPSS 的数据管理

统计分析离不开数据，因此数据管理是 SPSS 的重要组成部分。详细了解 SPSS 的数据管理方法，将有助于用户提高工作效率。SPSS 的数据管理是借助于数据管理窗口和主窗口的 File、Data、Transform 等菜单完成的。

第一节 数据的输入

2.1.1 变量的定义

先激活数据管理窗口，然后选 Data 菜单的 Define Variable...命令项，弹出 Define Variable 对话框（见图 1.1），在 Variable Name:框内输入变量名，如本例为 x1。

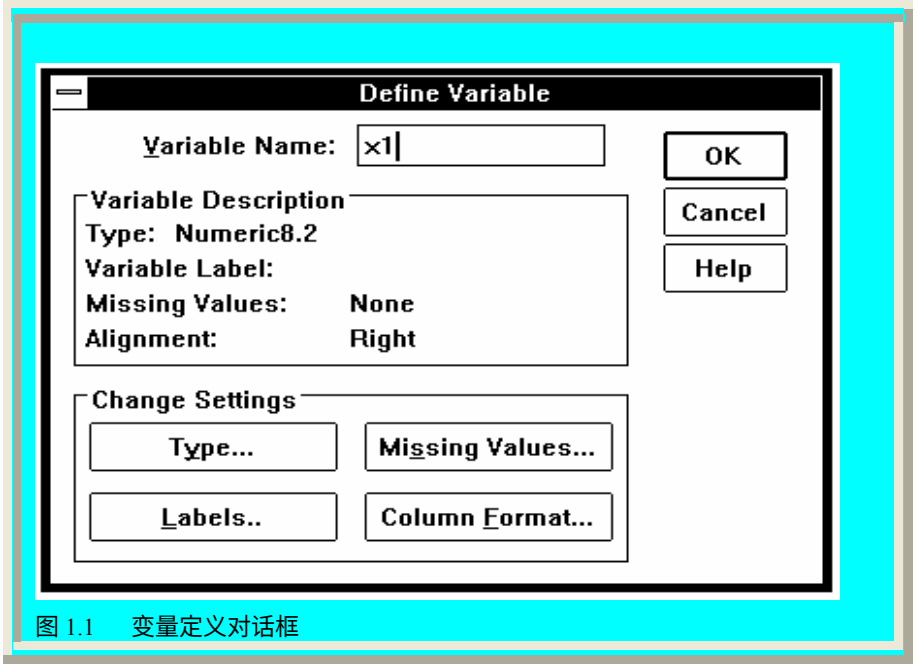


图 1.1 变量定义对话框

2.1.2 数据格式化

在 Define Variable 对话框中点击 Type...钮，弹出 Define Variable Type 对话框（如图 1.2 所示），用户可根据具体资料的属性对数据进行格式化。Define Variable Type 对话框中列出如下 7

种数据类型：

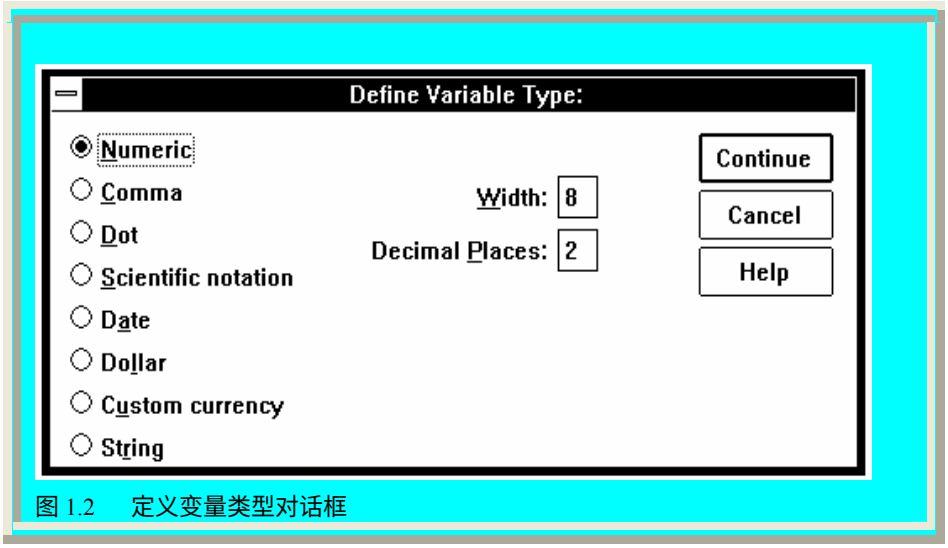


图 1.2　定义变量类型对话框

1、Numeric：数值型，同时定义数值的宽度（Width），即整数部分+小数点+小数部分的位数，默认为 8 位；定义小数位数（Decimal Places），默认为 2 位。

2、Comma：加显逗号的数值型，即整数部分每 3 位数加一逗号，其余定义方式同数值型。

3、Dot：3 位加点数值型，无论数值大小，均以整数形式显示，每 3 位加一小点（但不是小数点），可定义小数位置，但都显示 0，且小数点用逗号表示。如 1.2345 显示为 12.345,00（实际是 12345E-4).

4、Scientific notation:科学记数型,同时定义数值宽度(Width)和小数位数(Decimal Places)，在数据管理窗口中以指数形式显示。如 定义数值宽度为 9，小数位数为 2，则 345.678 显示为 3.46E+02。

5、Date：日期型，用户可从系统提供的日期显示形式中选择自己需要的。如选择 mm/dd/yy 形式，则 1995 年 6 月 25 日显示为 06/25/95。

6、Dollar：货币型，用户可从系统提供的日期显示形式中选择自己需要的，并定义数值宽度和小数位数，显示形式为数值前有 $ 。

7、Custom currency：常用型，显示为整数部分每 3 位加一逗号，用户可定义数值宽度和小数位数。如 12345.678 显示为 12,345.678。

8、String：字符型，用户可定义字符长度（Characters）以便输入字符。

用户选择完毕可点击 Continue 钮返回 Define Variable 对话框。

2.1.3 数据的输入

定义好变量并格式化数据之后，即可向数据管理窗口键入原始数据。数据管理窗口的主要部分就是电子表格，横方向为电子表格的行，其行头以 1、2、3、……表示，即第 1、2、3、……行；纵方向为电子表格的列，其列头以 var00001,var00002,var00003……表示变量名。行列交叉处称为单元格，即保存数据的空格。鼠标一旦移入电子表格内即呈十字形，这时按鼠标左键可激活单元格，被激活的单元格以加粗的边框显示；用户也可以按方向键上下左右移动来激活单元格。单元格被激活后，用户即可向其中输入新数据或修改已有的数据。图 1.3 所示即为一个已输入数据的数据管理窗口。为方便起见，用户亦可省略定义变量和数据格式化两个步骤，一启动 SPSS 即向数据管理窗口

中键入原始数据，这时，变量名默认为 var00001, var00002, var00003……

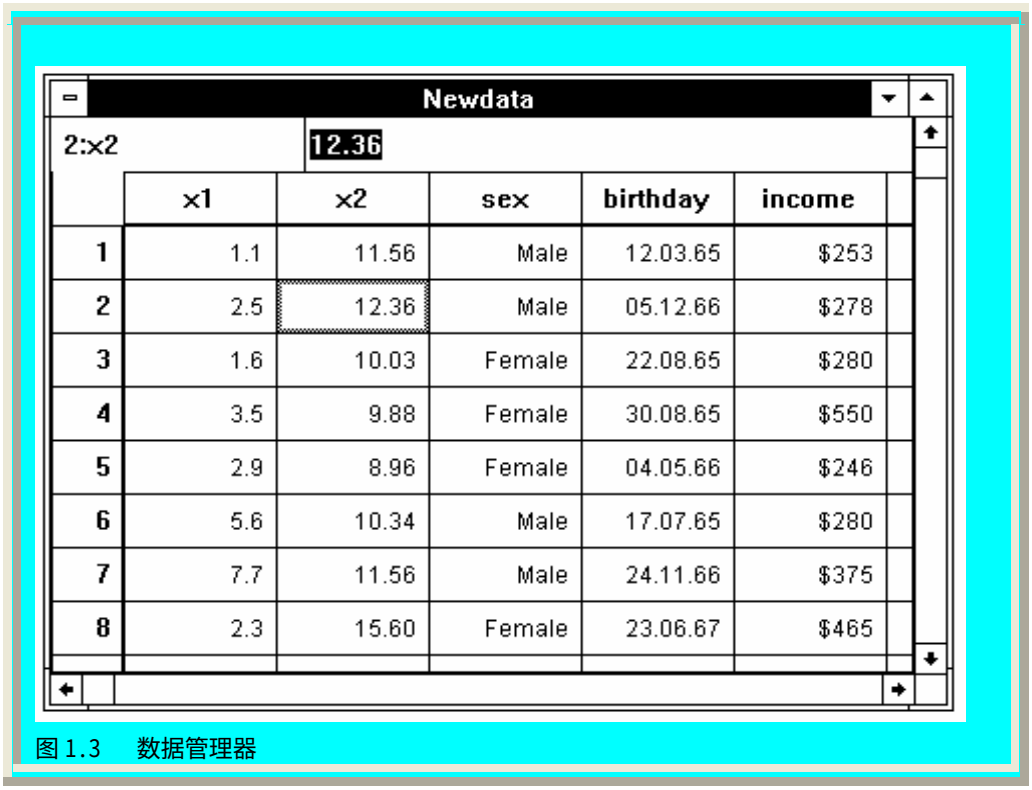


图 1.3　数据管理器

2.1.4 缺失值处理

在实际工作中，因各种原因会出现数值缺失现象，为此，SPSS 提供缺失值处理技术。在 Define Variable 对话框中点击 Missing Value...钮，弹出 Define Missing Values 对话框（图 1.4），用户有 4 个可选项：

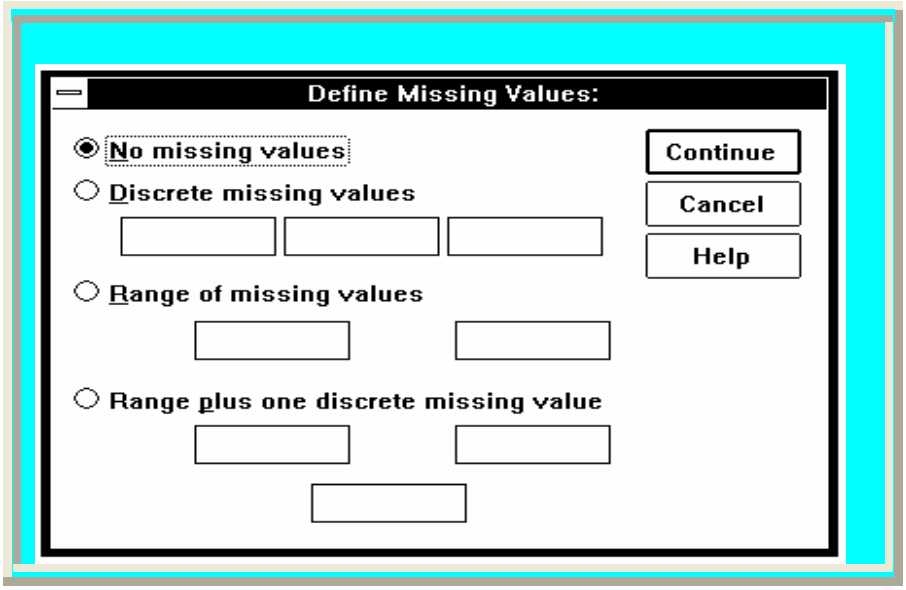


图 1.4　缺失值定义对话框

1、No missing values：没有缺失值；

2、Discrete missing values：可定义 1-3 个。如测量身高（厘米）的资料，可定义 999 为缺失值；性别的资料（男为 1、女为 2），可定义－1 为缺失值；

3、Range of missing values：可定义缺失值的范围。如脉搏资料，可定义 0—9 为缺失值；

4、Range plus one discrete missing value：可定义缺失值的范围，同时定义另外 1 个不是这一范围的缺失值。如定义 0—9 为脉搏的缺失值，同时定义 999 为身高的缺失值。

2.1.5 变量标签

在 Define Variable 对话框中点击 Labels...钮，弹出 Define Labels 对话框（图 1.5），用户可定义变量标签和特定变量值的标签。如定义变量 hb 的标签为“血红蛋白值”，同时定义 12.36 为“正常”，则可在 Define Labels 对话框中的 Variable Label 处输入变量标签名，在 Value Labels 框中的 Value 处指定变量值，在 Value Label 处输入变量值标签，点击 Add 钮表示加入这种标签定义，点击 Change 表示更改原有标签，用户重新定义，点击 Remove 钮表示取消原有标签。

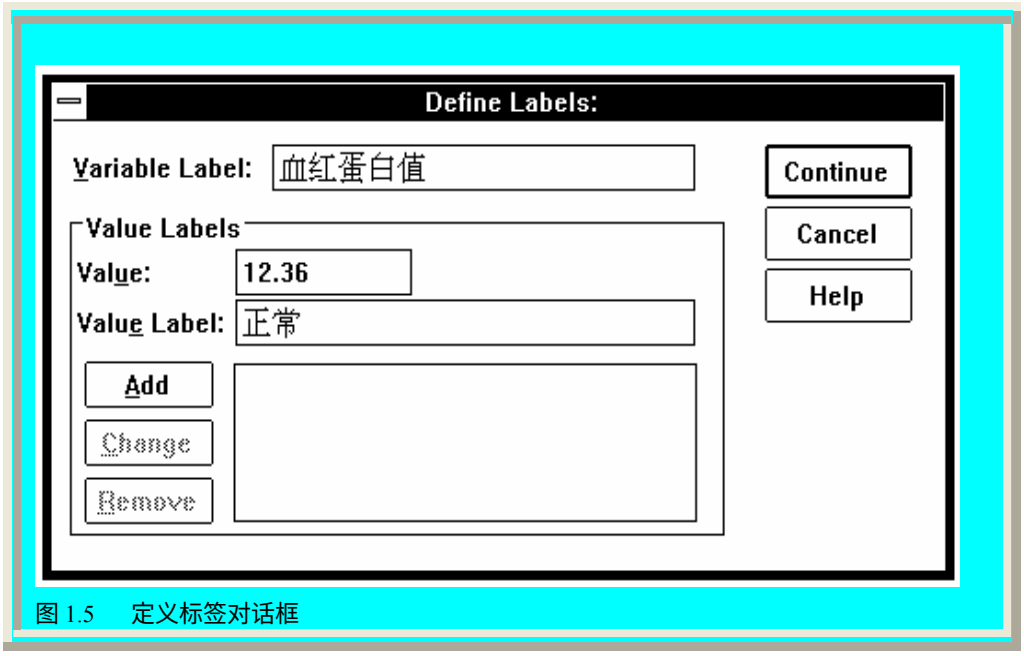


图 1.5　定义标签对话框

2.1.6 数据管理器列宽定义

在 Define Variable 对话框中点击 Column Format...钮，弹出 Define Column Format 对话框（图 1.6），用户可定义数据管理器纵列的宽度，以便显示较长的数值或文字；同时用户还可指定数值或文字在数据管理器单元格中的位置：Left 表示靠左、Center 表示居中、Right 表示靠右（此为默认方式）。

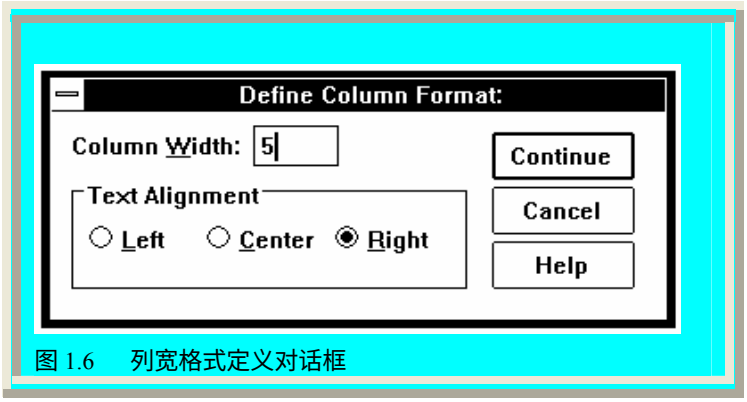


图 1.6　列宽格式定义对话框

第二节 数据的编辑

输入的原始数据，经常在统计分析前或统计分析过程中，需要作一些特殊的处理。为此，系统提供了如下主要方法。

2.2.1 数据的增删

2.2.1.1 增加一个新的变量列

例如要在第 2 列前增加一个新的列，使原来的第 2 列右移变成第 3 列，则可先激活第 2 列的任一单元格，然后选 Data 菜单的 Insert Variable 命令项，系统自动为用户在第 2 列前插入一个新的变量列，原第 2 列自动向右移一列成为第 3 列。

2.2.1.2 增加一个新的观察单位（即增加一个新的行）

例如要在第 6 个观察单位前增加一个观察单位（亦即在第 6 行前增加一行，使原来的第 6 行下移成为第 7 行），则可先激活第 6 行的任一单元格，然后选 Data 菜单的 Insert Case 命令项，系统自动为用户在第 6 行前插入一个新的行，原第 6 行列自动向下移一行成为第 7 行。

2.2.1.3 增加一个新的观察值

例如由于输入错误，造成第 7 个观察单位的第 4 个变量值漏输，结果第 8 个观察单位的第 4 个变量值误为第 7 个观察单位的第 4 个变量值，这样的情形使得数据管理器中的第 4 个变量值从第 7 行起全部上移，而合计例数少一个。于是希望在第 7 行的第 4 列处插入 1 个单元格，原有数据依次下移恢复正常。可先将鼠标指向在第 7 行第 4 列交叉处的单元格，然后按住鼠标左键向下拖动鼠标直至第 4 列从第 7 行起的所有数据被选中（黑底白字），选 Edit 菜单的 Cut 命令项，选中的数据被剪切入剪贴板，再激活第 8 行第 4 列交叉处的单元格，选 Edit 菜单的 Paste 命令项，可将剪贴板中的原第 7 行起的所有数据下移自第 8 行开始，并空出第 7 行第 4 列的单元格以便补入漏输的数值。

2.2.1.4 删除一个行

例如要删除第 9 行（即删除这个观察单位的所有观察值），则可先点击第 9 行的行头，这时整个第 9 行被选中（呈黑底白字状），然后按 Delete 键或选 Edit 菜单的 Clear 命令项，该行即被删除。

2.2.1.5 删除一个变量列

例如要删除第 4 个变量列，则可先点击第 4 列的列头，这时整个第 4 列被选中（呈黑底白字状），然后按 Delete 键或选 Edit 菜单的 Clear 命令项，该列即被删除。

2.2.1.6 删除一个观察值

例如由于输入错误，造成第 6 个观察单位的第 2 个变量值重复输入，结果第 7 个观察单位的第 2 个变量值误为第 6 个观察单位的第 2 个变量值，第 8 个观察单位的第 2 个变量值误为第 7 个观察单位的第 2 个变量值，……，这样的情形使得数据管理器中的第 2 个变量值从第 7 行起全部下移，而合计例数多一个。于是希望将第 7 行第 2 列的单元格删除，原有数据依次上移恢复正常。可先将鼠标指向在第 8 行第 2 列交叉处的单元格，然后按住鼠标左键向下拖动鼠标直至第 2 列从第 8 行起的所有数据被选中（黑底白字），选 Edit 菜单的 Cut 命令项，选中的数据被剪切入剪贴板，再激活第 7 行第 2 列交叉处的单元格，按 Del 键删除该单元格的数值，选 Edit 菜单的 Paste 命令项，可将剪贴板中的原第 8 行起的所有数据上移自第 7 行开始，既填补第 7 行第 2 列的单元格，又恢复原有下移的数值。

2.2.2 数据的整理

2.2.2.1 数据的排序

用户可按要求对数据管理器的数据进行排序。选 Data 菜单的 Sort Cases...命令项，弹出 Sort Cases...对话框（图 1.7），在变量名列框中选 1 个需要按其数值大小排序的变量（用户也可选多个变量，系统将按变量选择的先后逐级依次排序），点击 ➢ 钮使之进入 Sort by 框，然后在 Sort Order 框中确定是按升序（Ascending，从小到大）或降序（Descending，从大到小），点击 OK 钮即可。

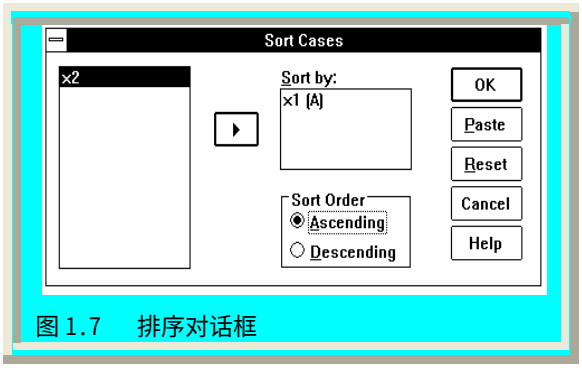


图 1.7　排序对话框

2.2.2.2 数据的行列互换

有时，用户需要将数据管理器中原先按行（列）方向排列的数据转换成按列（行）方向排列的数据，这时可选 Data 菜单的 Transpose...命令项，弹出 Transpose...对话框（图 1.8），在变量名列框中选 1 个或多个需要转换的变量，点击 ➢ 钮使之进入 Variable(s)框，再点击 OK 钮即可。产生的新数据会在第 1 列出现一个 case_lbl 新变量，用于放置原来数值的变量名。若要将数据再转换回原来的排列方式，方法与上述过程相同。

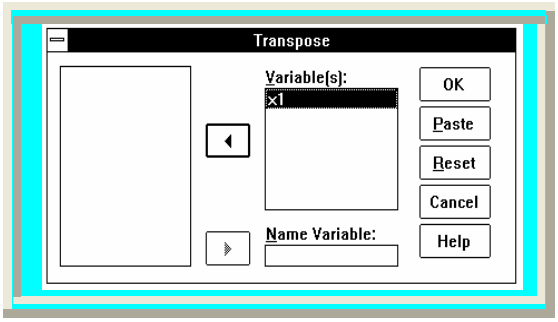


图 1.8　行列互换框

2.2.2.3 数据的分组汇总

用户还可对数据管理器中的数据按指定变量的数值进行归类分组汇总，汇总的形式十分多样。例如，要对下列数据（图 1.9）按变量 group 的大小，把变量 x1 作平均值汇总、把变量 x2 作求和汇总。选 Data 菜单的 Aggregate...命令项，弹出 Aggregate Data 对话框（图 1.10），在变量名列框中选 group 变量，点击➤钮使之进入 Break Variable(s)框，选 x1 变量进入 Aggregate Variable(s)框，因 x1 欲作平均值汇总，故点击 Function...钮弹出 Aggregate Data: Aggregate Function 对话框（图 1.11）选 Mean of values 项点击 Continue 钮返回；选 x2 变量进入 Aggregate Variable(s)框，因 x2 变量欲作求和汇总，故点击 Function...钮选 Sum of values 项点击 Continue 钮返回。再点击 OK 钮即可。结果如图 1.12 所示。

Newdata

1:x2

	group	x1	x2
1	1	3.66	4.00
2	1	3.56	5.00
3	2	2.63	5.00
4	3	6.99	5.00
5	1	5.36	6.00
6	3	4.33	6.00
7	1	3.21	4.00
8	1	3.56	5.00

图 1.9　欲作分组汇总的原始数据

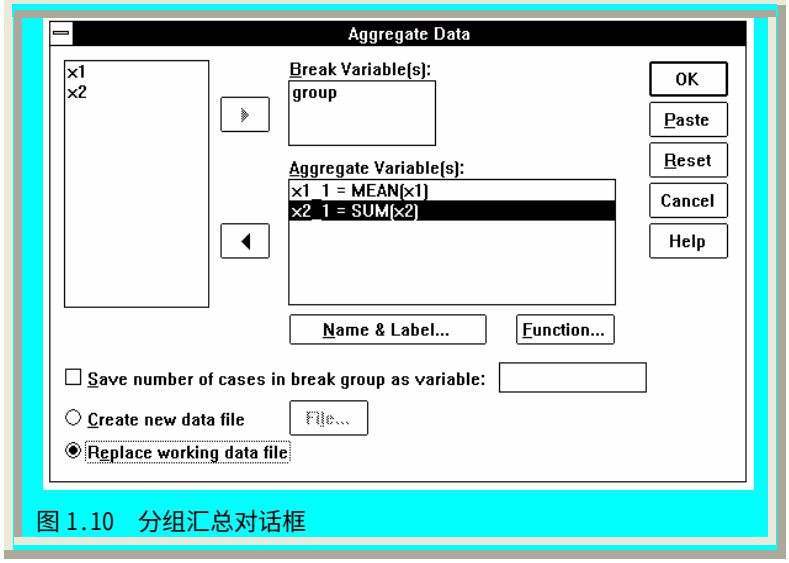


图 1.10　分组汇总对话框

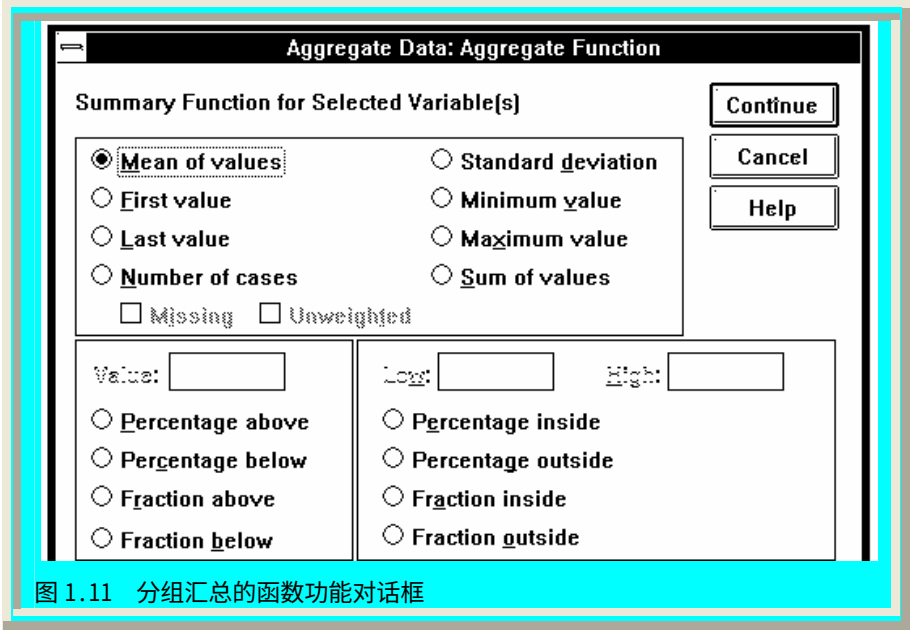


图 1.11　分组汇总的函数功能对话框

图 1.12　分组汇总后的数据

分组汇总提供的函数形式有：

1、Mean of values：求该组的平均值；

2、Standard deviation：求该组的标准差；

3、First value：只保留该组的第 1 个数值；

4、Minimum value：只保留该组的最小值；

5、Last value：只保留该组的最后 1 个数值；

6、Maximum value：只保留该组的最大值；

7、Number of cases：合计该组的观察例数；

8、Sum of values ：求该组所有观察值的和。

9、Percentage above ：先确定 1 个数值，求大于该数值的所有例数占总例数的百分比（0-100%）；

10、Percentage below：先确定 1 个数值，求小于该数值的所有例数占总例数的百分比（0-100%）；

11、Fraction above：先确定 1 个数值，求大于该数值的所有例数占总例数的百分比（0-1）；

12、Fraction below：先确定 1 个数值，求小于该数值的所有例数占总例数的百分比（0-1）；

13、Percentage inside：先确定 1 个下限，再确定 1 个上限，求数值在该区间内的例数占总例数的百分比（0-100%）；

14、Percentage outside：先确定 1 个下限，再确定 1 个上限，求数值在该区间外的例数占总例数的百分比（0-100%）；

15、Fraction inside：先确定 1 个下限，再确定 1 个上限，求数值在该区间内的例数占总例数的百分比（0-1）；

16、Fraction outside：先确定 1 个下限，再确定 1 个上限，求数值在该区间外的例数占总例数的百分比（0-1）。

2.2.2.4 数据的分割

数据也可根据需要，事先按用户的指定作分组（这种分组是系统内定义的，在数据管理器中并不一定明确体现，故亦可称之为分割），此后的所有分析都将按这种分组进行，除非取消数据分割的命令。选 Data 菜单的 Split File...命令项，弹出 Split File 对话框（图 1.13），选 Repeat analysis for each group 表示此后都按指定的分组方式作相同项目的分析，用户可从变量名列框中选 1 个或多个变量点击 ➢ 钮使之进入 Groups Based on 框来作分组的依据。若在数据分割之后要取消这种分组，可选 Analyze all cases 项，则系统恢复如初。

调用 Split File 命令完成定义后，SPSS 将在主窗口的最下面状态行中显示 Split File On 字样；若调用该命令后的数据库被用户存盘，则当这个数据文件再次打开使用时，仍会显示 Split File On 字样，意味着数据分割命令依然有效。

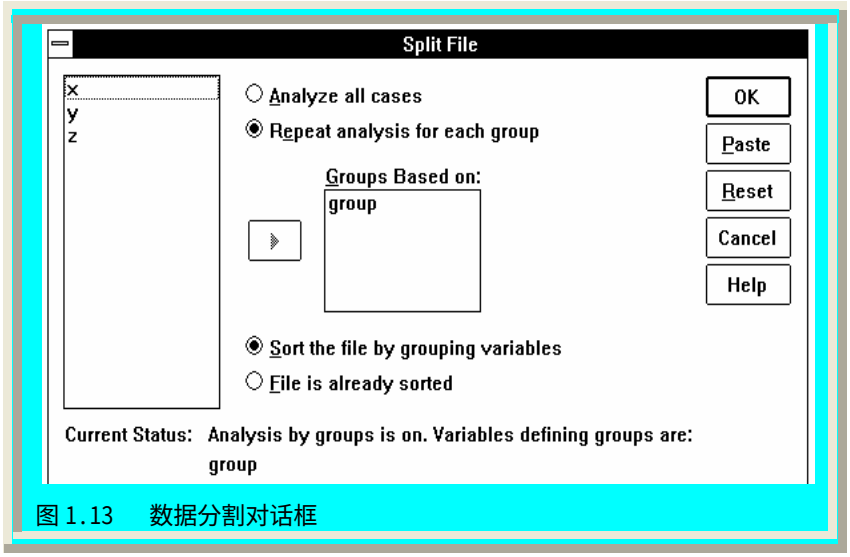


图 1.13　数据分割对话框

2.2.2.5 数据的选择

除按要求作数据分组分别作分析外（但这依然是将所有的资料全部代入分析），还可从所有资料中选择一些数据进行统计分析。选 Data 菜单的 Select Cases...命令项，弹出 Select Cases 对话框（图 1.14），系统提供如下几种选择方法：

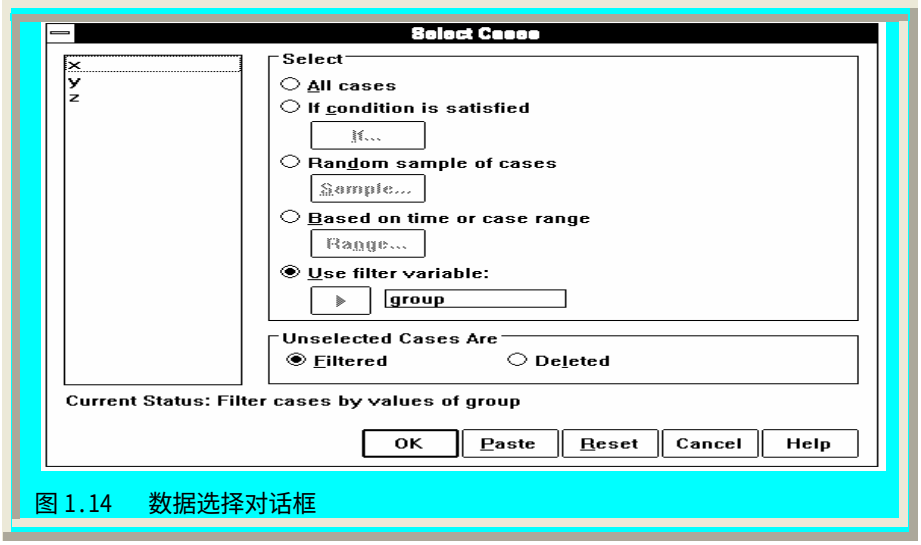


图 1.14　数据选择对话框

1、All cases：表示所有的观察例数都被选择，该选项可用于解除先前的选择；

2、If condition is satisfied：表示按指定条件选择，点击 If...钮，弹出 Select Cases:If 对话框（图 1.15），先选择变量，然后定义条件；

3、Random sample of cases：表示对观察单位进行随机抽样，点击 Sample...钮，弹出 Select Cases:Random Sample 对话框，有两种选择分式，一是大概抽样（Approximately）即键入抽样比例后由系统随机抽取，另一是精确抽样（Exactly）即要求从第几个观察值起抽取多少个；

4、Based on time or case range：表示顺序抽样，点击 Range...钮，弹出 Select Cases:Range 对话框，用户定义从第几个观察值抽到第几个观察值；

5、Use filter variable：表示用指定的变量作过滤，用户先选择 1 个变量，系统自动在数据管理器中将该变量值为 0 的观察单位标上删除标记，系统对有删除标记的观察单位不作分析。若用户在 Select Cases 对话框的 Unselected Cases Are 框中选 Deleted 项，则系统将删除所有被标上删除标记的观察单位。

调用 Select Cases 命令完成定义后，SPSS 将在主窗口的最下面状态行中显示 Filter On 字样；若调用该命令后的数据库被用户存盘，则当这个数据文件再次打开使用时，仍会显示 Filter On 字样，意味着数据选择命令依然有效。

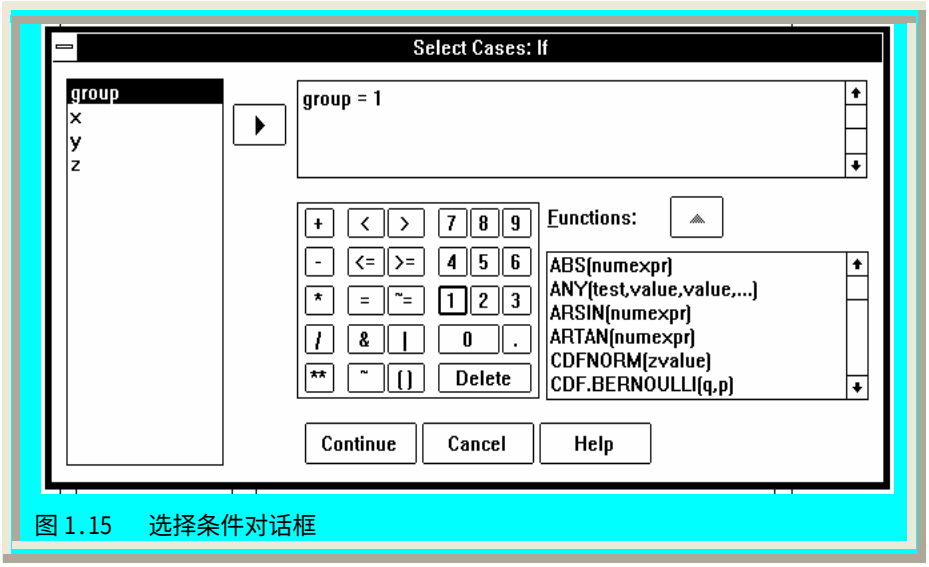


图 1.15　选择条件对话框

2.2.2.6 观察单位的秩次确定

为了解在指定条件下某个或某些变量值的大小顺序，可选 Transform 菜单的 Rank Cases...命令项，弹出 Rank Cases 对话框（图 1.16），从变量名列框中选 1 个或多个变量点击 ➢ 钮使之进入 Variable(s)框作为按该变量值大小排序的依据。若选 1 个或多个变量使之进入 By 框，则系统在排序时将按进入 By 框的变量值分组排序。排序的结果将在数据管理器中新建 1 个变量名为原排序变量前加一特定排序类型字母（如原变量为 x，则普通排序时变量为 rx）的变量用于放置秩次。用户可在 Rank Cases 对话框的 Assign Rank 1 to 框中指定秩次排列方式：Smallest value 表示最小值用 1 标注，之后为 2、3、4.....；Largest value 表示最大值用 1 标注，之后为 2、3、4.....。

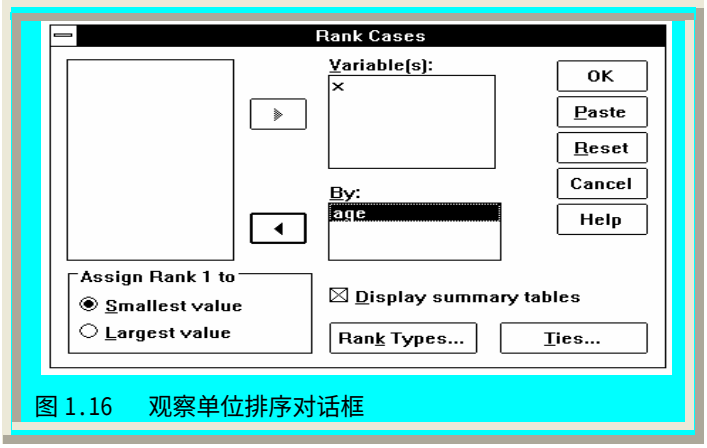


图 1.16　观察单位排序对话框

若点击 Rank Cases 对话框的 Rank Types...钮，可选择排序类型（图 1.17）：

1、Rank：普通秩次，排序类型字母 r；

2、Fractional Rank as percent：累积百分秩次，排序类型字母 p；

3、Savage score：以指数分布为基础的原始分秩次，排序类型字母 s；

4、Sum of case weights：分组例数之和的权重秩次，排序类型字母 n；

5、Fractional Rank：分组例数之和占总例数累积百分比秩次，排序类型字母 r；

6、Ntiles：先给定一个大于 1 的整数，系统按此数范围确定排序的秩次，排序类型字母 n。

图 1.17　观察单位排序类型对话框

2.2.3 数据的算术处理

2.2.3.1 变量的加权

选 Data 菜单的 Weight Cases...命令项，可对指定的数值变量进行加权。在弹出的 Weight Cases 对话框中（图 1.18），Do not weight cases 表示不做加权，这可用于对做过加权的变量取消加权；Weight cases by 表示选择 1 个变量做加权。在加权操作中，系统只对数值变量进行有效加权，即大于 0 的数按变量的实际值加权，0、负数和缺失值加权为 0。

加权操作在χ^2检验中是必不可少的，且一旦该变量做过加权操作，那么，一方面系统自动根据用户对已加权变量值的修改做加权变换，另一方面用户除非取消加权，否则即使改变变量名，系统依然对该变量进行加权操作。

调用 Weight Cases 命令完成定义后，SPSS 将在主窗口的最下面状态行中显示 Weight On 字样；

若调用该命令后的数据库被用户存盘，则当这个数据文件再次打开使用时，仍会显示 Weight On 字样，意味着数据加权命令依然有效。

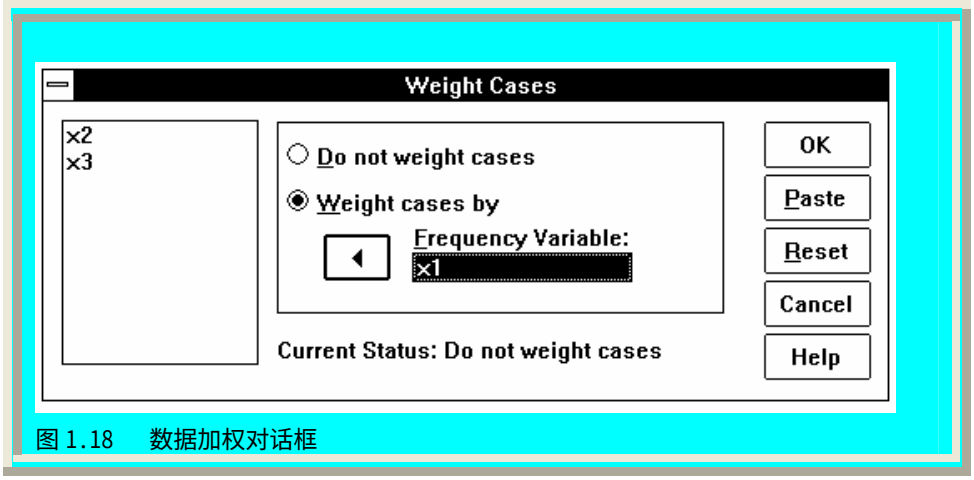


图 1.18　数据加权对话框

2.2.3.2 数据的运算与新变量的生成

选 Transform 菜单的 Compute...命令项，既可对选定的变量进行运算操作，又可通过运算操作让系统生成新的变量。在弹出的 Compute Variable 对话框中(图 1.19)，用户首先在 Target Variable 指定一个变量（可以是数据管理器中已有的变量，也可是用户欲生成的新变量），然后点击 Type&Label...钮确定是数值型变量，还是字符型变量，或加上变量标签。在 Numeric Expression 框中键入运算公式，系统提供计算器和 82 种函数（在 Functions 框内）让用户使用；若点击 If...钮会弹出 Compute Variable:If Cases 对话框（类似于图 1.15 的选择条件对话框），用户可指定符合条件的变量参与运算。

如本例是要求系统生成一个新变量 x1，x1 = x 的绝对值＋y－0.123×z。点击 OK 钮即可。结果在数据管理器中产生一个新变量 x1。

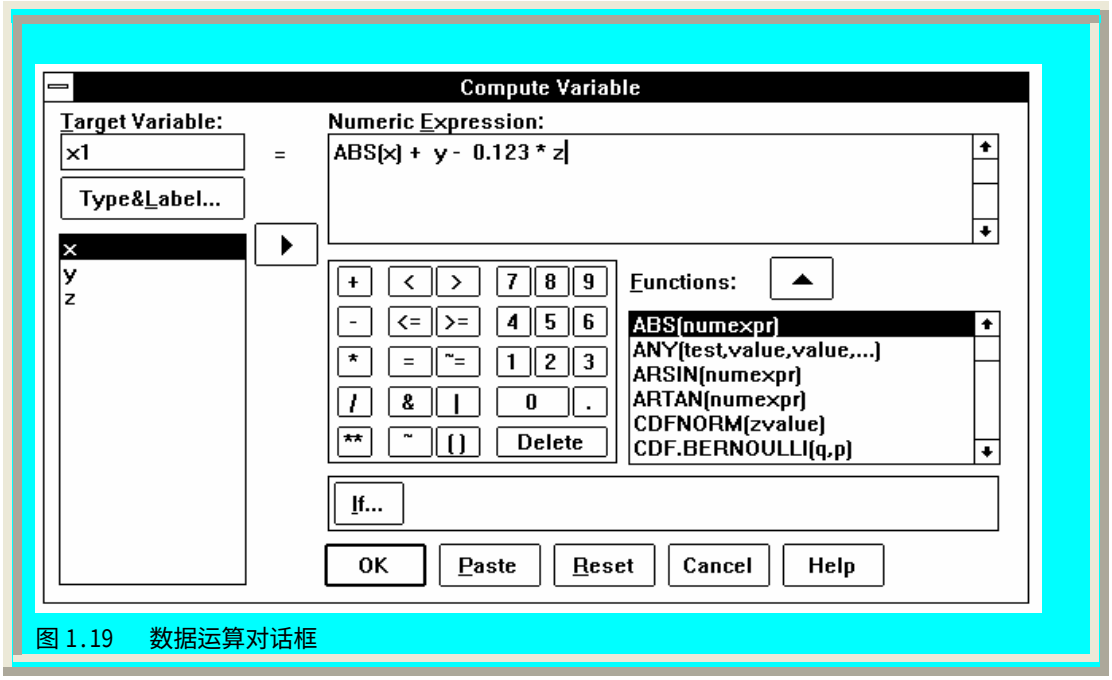


图 1.19　数据运算对话框

2.2.3.3 变量值个数的清点

对于数值型变量，某个或某些值在各观察单位中的出现次数可以作清点。选 Transform 菜单的

Count...命令项，在弹出的 Count Occurrences of Value within Cases 对话框中（图 1.20），先在 Target Variable 指定一个变量（可以是数据管理器中已有的变量，也可是用户欲生成的新变量），然后指定需要清点的变量，即在变量名列中选择 1 个或多个变量点击 ➢ 钮使之进入 Numeric Variable(s)框，再点击 Define Values...钮，弹出 Count Value within Cases:Value to Count 对话框，确定哪些数值作为清点对象。选 Value 表示单一数值为清点对象；选 System-missing 或 System-or user missing 表示系统或用户指定的缺失值为清点对象；选 Range 表示指定数值范围为清点对象。还可点击 If...钮指定条件来确定参与清点的观察单位。

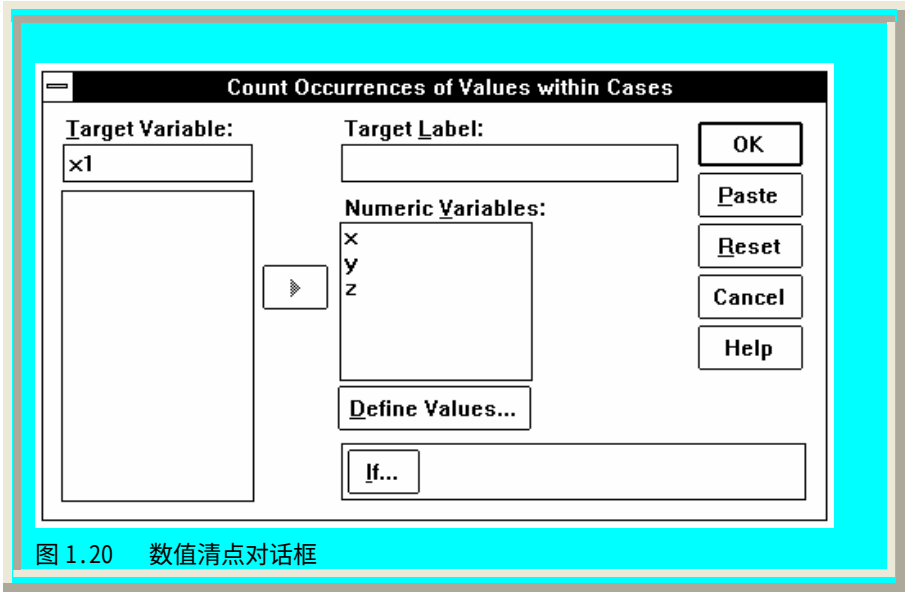


图 1.20　数值清点对话框

2.2.3.4 变量的重新赋值

在数据管理器中，用户可对各单元的数值重新赋予新值。这种操作只适用于数值型变量。选 Transform 菜单的 Recode 命令项，此时有两种选择：一是对变量自身重新赋值（Into Same Variables...），一是对其它变量或新生成的变量进行赋值（Into Different Variables...）。若选第一种赋值方法，在弹出的 Recode into Same Variables 对话框中（图 1.21），先在变量名列中选 1 个或多个变量点击 ➢ 钮使之进入 Numeric Variables 框，然后点击 Old and New Values...钮弹出 Recode into Same Variables:Old and New Value 对话框，用户根据实际情况确定旧值和新值，点击 Continue 钮返回，再点击 OK 钮即可。若选第二种赋值方法，在弹出的 Recode into Different Variables 对话框中（图 1.22），先在变量名列中选 1 个或多个变量点击 ➢ 钮使之进入 Numeric Variable→Output Variable 框，同时在 Output Variable 框确定一赋值变量（可以是数据管理器中已有的变量，也可以是用户要求生成的新变量），然后点击 Old and New Values...钮弹出 Recode into Different Variables:Old and New Value 对话框，用户根据实际情况确定旧值和新值，点击 Continue 钮返回，再点击 OK 钮即可。

在两种赋值情况下，用户均可点击 If...钮指定条件来确定参与清点的观察单位。

与 Compute 方法不同的是：Recode 方法不能进行运算，只能根据指定变量值作数值转换，且这种转换是单一数值的转换。

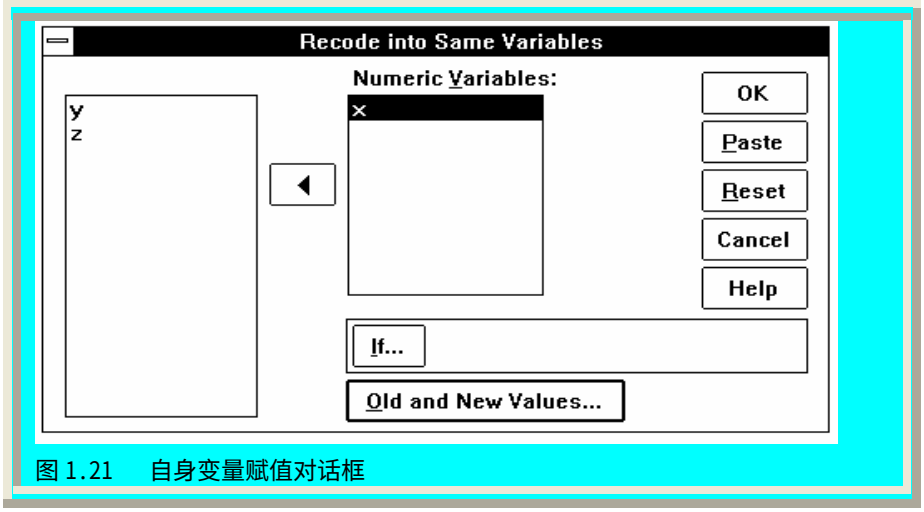


图 1.21　自身变量赋值对话框

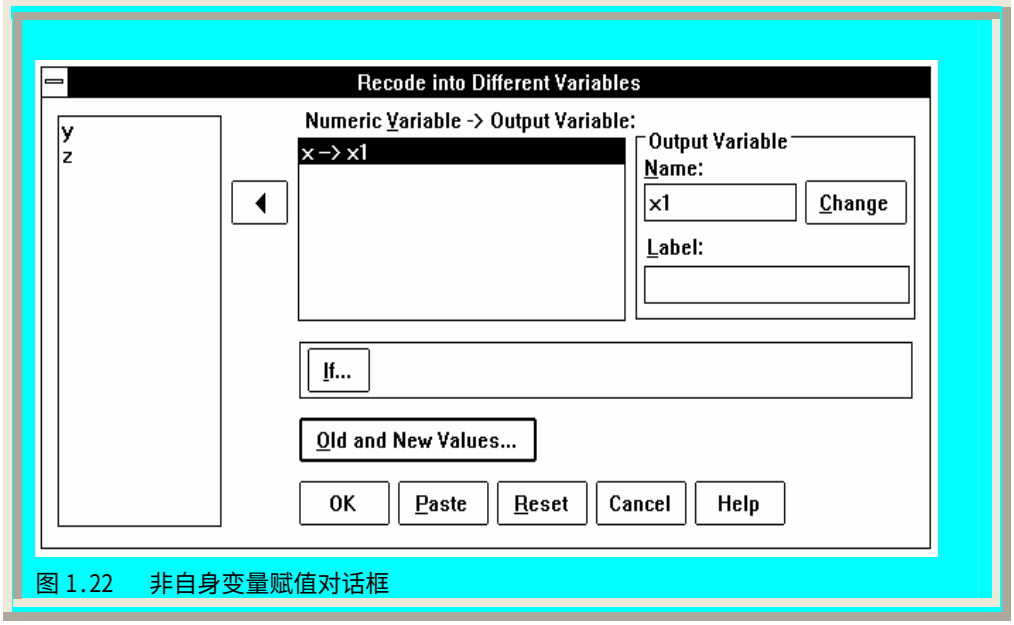


图 1.22　非自身变量赋值对话框

2.2.3.5　缺失值的替代

对于缺失值，可采取多种手段进行科学替代。选 Transform 菜单的 Replace Missing Values... 命令项，在弹出的 Replace Missing Values 对话框中（图 1.23），先在变量名列中选 1 个或多个存在缺失值的变量点击 ➢ 钮使之进入 New Variable(s)框，这时系统自动产生用于替代缺失值的新变量，用户也可在 Name 框处自己定义替代缺失值的新变量名。然后点击 Method 的下箭头选择缺失值的替代方式：

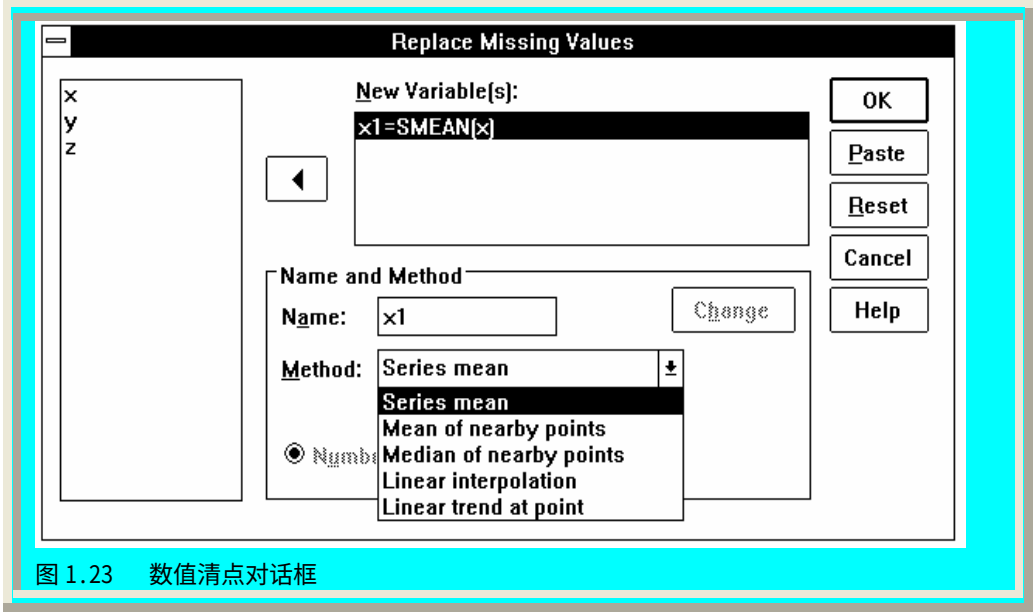


图 1.23　数值清点对话框

1、Series mean：用该变量的所有非缺失值的均数做替代；

2、Mean of nearby points：用缺失值相邻点的非缺失值的均数做替代，取多少个相邻点可任意定义；

3、Median of nearby points：用缺失值相邻点的非缺失值的中位数做替代，取多少个相邻点可任意定义；

4、Linear interpolation：用缺失值相邻两点非缺失值的中点值做替代；

5、Linear trend at point：用线性拟合方式确定替代值。

第三节 数据文件的管理

2.3.1 数据文件的调用

选 File 菜单的 Open 命令项，再选 Data...项，弹出 Open Data File 对话框，用户确定盘符、路径、文件名后点击 OK 钮，即可调入数据文件。

系统支持如下格式的数据文件：

1、SPSS：SPSS for WINDOWS 版本的数据文件，后缀为.sav；

2、SPSS/PC+ ：SPSS for DOS 版本的数据文件，后缀为.sys；

3、SPSS portable：SPSS 的 ASCII 格式的机器码，可用于网络传输，后缀为.por；

4、Excel：微软公司电子表格的数据文件，后缀为.xls；

5、Lotus：莲花公司电子表格的数据文件，后缀为.w*；

6、SYLK：扩展格式电子表格的 ASCII 格式，后缀为.slk；

7、dBASE：数据库的数据文件，后缀为.dbf；

8、Tab-delimited：以空格为分隔的 ASCII 格式的数据文件，后缀为.dat。

2.3.2 数据文件的连接

2.3.2.1 纵向连接——观察单位的追加

利用数据连接功能可以将两个或两个以上的具有相同变量格式的数据文件连在一起。选 Data 菜单的 Merge Files 命令项，选 Add Cases...项，弹出 Add Cases:Read File 对话框（类似于图 1.24），用户确定盘符、路径、文件名后点击 OK 钮，即完成连接。如本例有两个数据文件：data1.sav 和 data3.sav（图 1.25），它们具有共同的变量 name、x1、x2，将之连接后如图 1.26 所示。

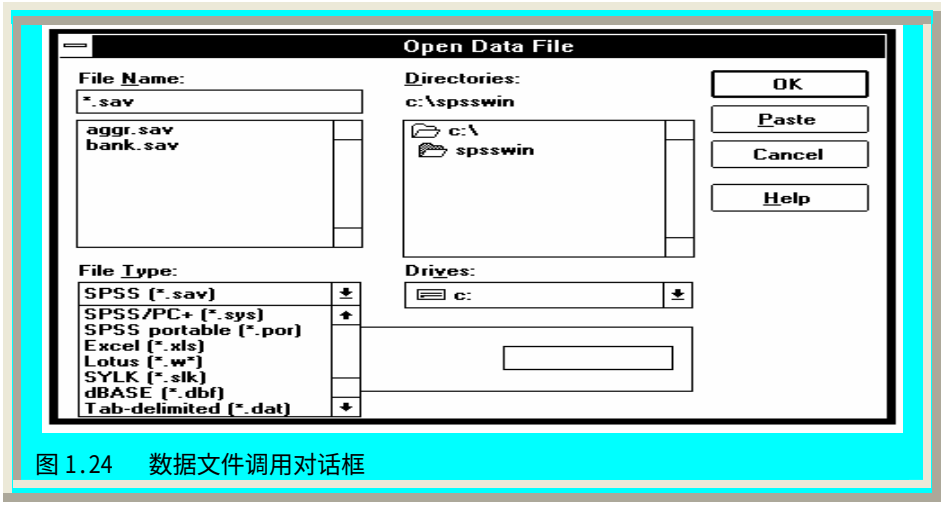


图 1.24　数据文件调用对话框

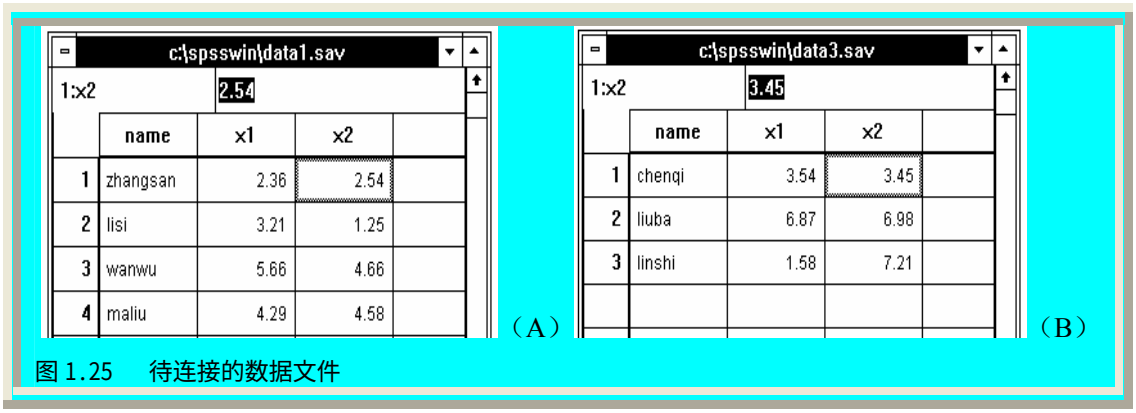


图 1.25　待连接的数据文件

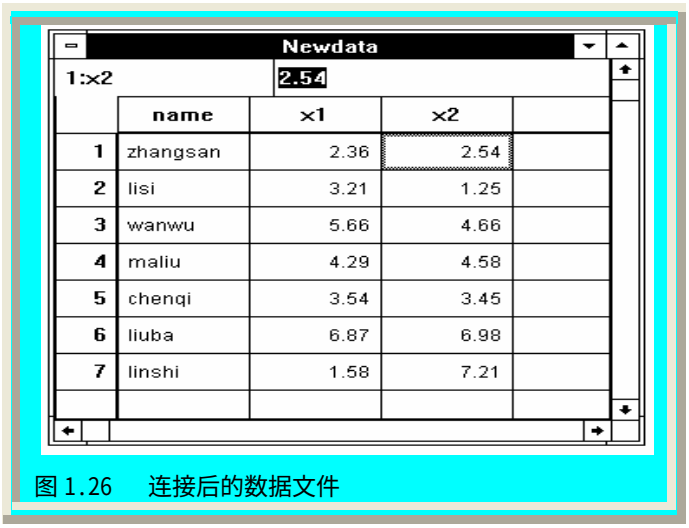


图 1.26　连接后的数据文件

2.3.2.2 横向连接——变量值的合并

利用数据连接功能还可以将两个或两个以上的具有相同观察单位的数据文件连在一起。选 Data 菜单的 Merge Files 命令项，选 Add Variables...项，弹出 Add Variables:Read File 对话框（类似于图 1.24），用户确定盘符、路径、文件名后点击 OK 钮，即完成连接。如本例有两个数据文件：data1.sav 和 data2.sav（图 1.27），它们具有共同的观察单位 zhangsan、lisi、wanwu、maliu，将之连接后如图 1.28 所示。

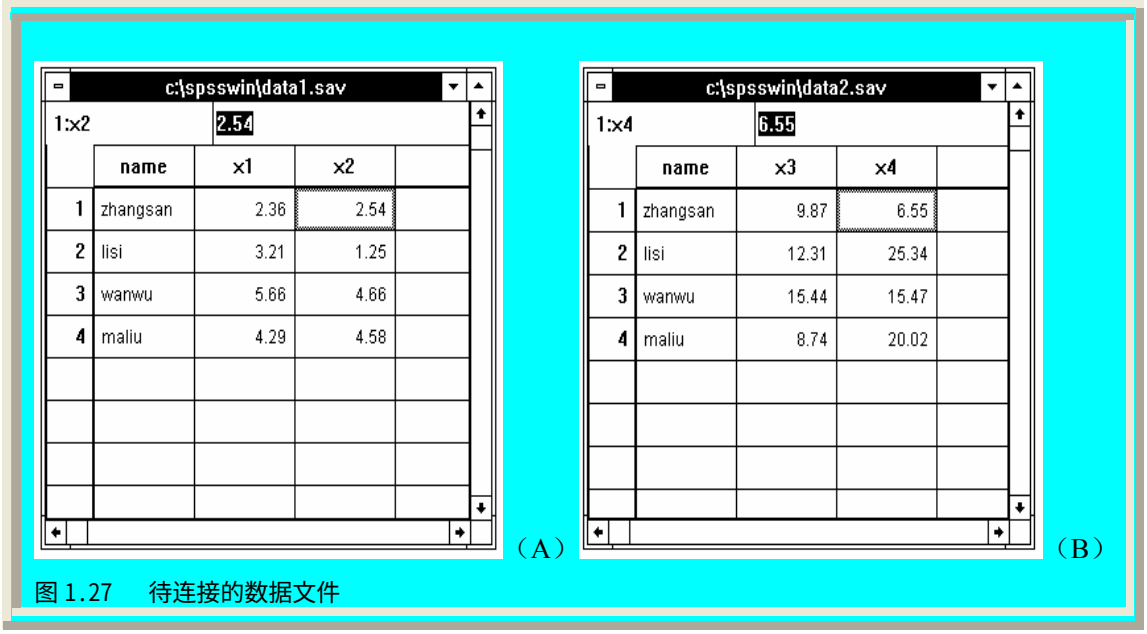


图 1.27　待连接的数据文件

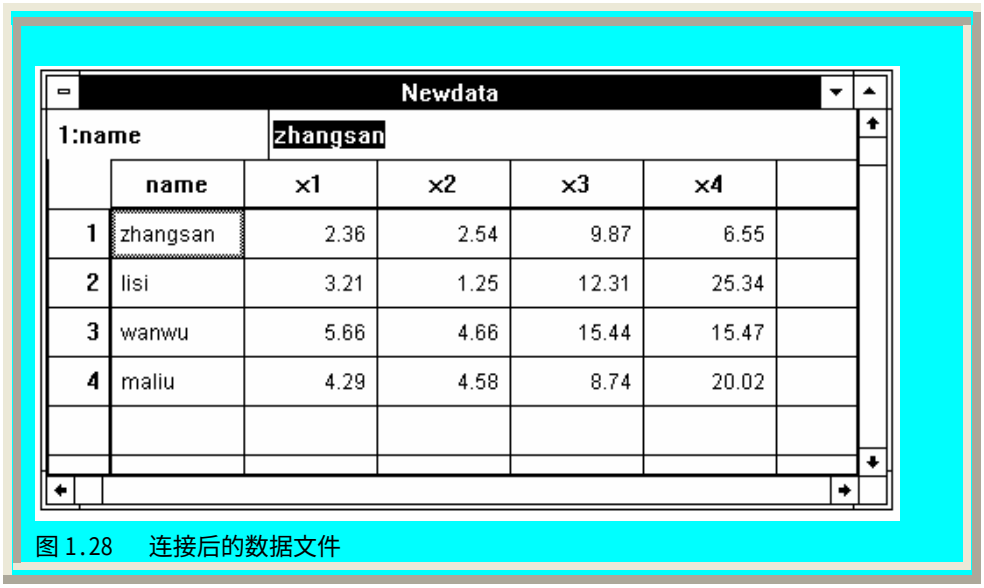


图 1.28　连接后的数据文件

2.3.3 数据文件的保存

输入数据管理器中的数据，无论什么时候（完成统计后或未作任何分析前或数据尚未输完，等），用户均可对之进行保存，以便于再使用（可以用于下次再追加数据、或作其他统计处理、或转成其

他格式的数据文件供别的软件使用，等）都可以将数据文件保存起来。

选 File 菜单的 Save As... 命令项，弹出 Newdata:Save Data As 对话框（图 1.29），用户确定盘符、路径、文件名以及文件格式后点击 OK 钮，即可保存数据文件。

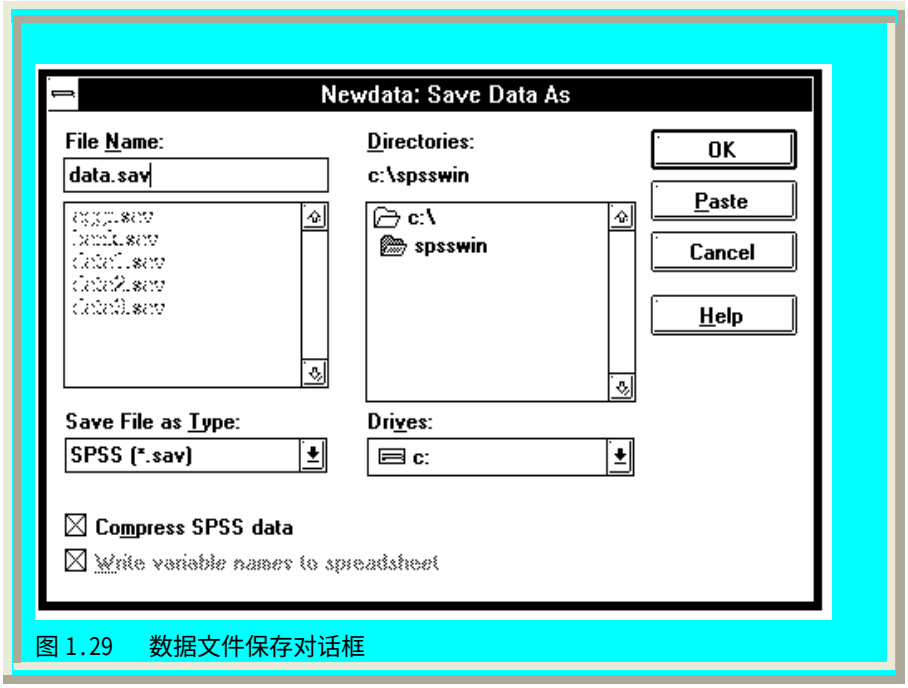


图 1.29　数据文件保存对话框

系统可由用户通过点击 Save File as Type 框的下箭头，选择确定完成下列格式数据文件的存放：

1、SPSS(*.sav)：SPSS for WINDOWS 版本的数据格式；
2、SPSS/PC+(*.sys)：SPSS for DOS 版本的数据格式；
3、SPSS Portable(*.por)：SPSS for WINDOWS 版本的 ASCII 码数据格式；
4、Tab-delimited(*.dat) ：用空格分割的 ASCII 码数据格式；
5、Fixed ASCII(*.dat) ：混合 ASCII 码数据格式；
6、Excel(*.xls) ：Excel 的数据格式；
7、1-2-3 Rel 3.0(*.wk3) ：Lotus 3.0 版本的数据格式；
8、1-2-3 Rel 2.0(*.wk1) ：Lotus 2.0 版本的数据格式；
9、1-2-3 Rel 1.0(*.wks) ：Lotus 1.0 版本的数据格式；
10、SYLK(*.slk) ：扩展方式电子表格的数据格式；
11、dBASE Ⅳ(*.dbf) ：dBASE Ⅳ版本的数据格式；
12、dBASE Ⅲ(*.dbf) ：dBASE Ⅲ版本的数据格式；
13、dBASE Ⅱ(*.dbf) ：dBASE Ⅱ版本的数据格式。

第三章 SPSS 文本文件的编辑

上一章介绍了 SPSS 数据管理窗口的使用方法。在第一章中，我们还提到过 SPSS 的其他窗口，

如结果输出窗口（图 3.1）和命令编辑窗口（图 3.2），这两个窗口是系统用于接收或输出文本的。用户经常在实际工作中需要对之进行必要的编辑。SPSS 的文本编辑是借助于主窗口的 File、Edit 等菜单完成的，本章介绍 SPSS 的文本编辑方法。

图 3.1　结果输出窗口（对变量 x 进行 Descriptives 统计过程分析的结果）

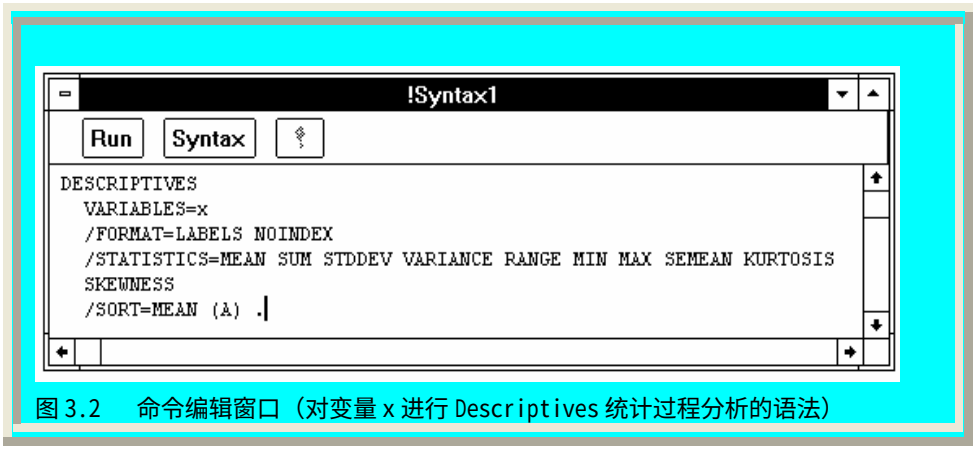


图 3.2　命令编辑窗口（对变量 x 进行 Descriptives 统计过程分析的语法）

第一节 文本文件的管理

3.1.1 文件的生成

SPSS 文本文件主要有两种生成方法：

1、在进行统计分析时，系统会将出错信息、数据转换情况、统计运算的中间环节和最终结果送到结果输出窗口中，这就是结果文本的内容；

2、在调用 Statistics 菜单的统计过程命令项时，会弹出统计过程对话框，这时若点击 Paste 钮就会出现命令编辑窗口，在该窗口中显示了与 SPSS For DOS 相类似的 SPSS 语法命令，这就是命令文本的内容。

无论是结果文本还是命令文本，用户都可以对之进行必要的编辑。

3.1.2 文件的保存

对于出现在结果输出窗口和命令编辑窗口的文本内容，用户可以将之保存起来以便日后查阅。方法是：先激活该窗口（窗口标题栏为蓝底白字时，即为活动窗口），然后选 File 菜单的 Save As... 命令项，弹出 Save As 对话框，用户指定盘符、路径和文件名后点击 OK 钮即可保存文件。

系统对结果文本的文件名默认后缀为.lst，对命令文本的文件名默认后缀为.sps。

3.1.3 文件的调用

对存盘的文本文件，可以在需要时调用它。选 File 菜单的 Open 命令项，再选 SPSS Syntax... 项，弹出 Open SPSS Syntax 对话框，用户指定盘符、路径和文件名后点击 OK 钮即可调用后缀为.lst 的结果文本文件；若选 File 菜单的 Open 命令项，再选 SPSS Output...项，弹出 Open Output 对话框，用户指定盘符、路径和文件名后点击 OK 钮即可调用后缀为.sps 的语法文本文件。

3.1.4 文件的打印

用户还可将文本打印出来以便阅读或保存。先激活需要打印的窗口，然后选 File 菜单的 Print... 命令项，弹出 Print 对话框，用户确定是全部打印（All）还是选择部分打印（Selection），并确定打印份数（Copies）后，点击 OK 钮即可将文本内容送往打印机。

系统在 File 菜单中还提供 Printer Setup...命令项，选择命令项可对打印机类型、纸张尺寸、打印边界、打印输出方向、打印分辩率和打印颜色深浅度进行设定。

第二节 文本文件的编辑

显示在结果输出窗口和命令编辑窗口的文本内容，就象任何显示在文字处理器中的文字内容一样，可以按用户的需要做修改、增删、移动、查找、替换等操作。但 SPSS 毕竟不是专门的文字处理器，而是统计分析软件，因此，其文本编辑的功能相对有限。对其文本内容，尤其是运算结果的有关内容，用户经常需将之体现在专业报告中。如果用户想仅仅依靠 SPSS 有限的文本编辑功能直接就着输出的结果进行文章撰写，那么会发现其排版功能的不足让人捉襟见肘。本节介绍 SPSS 的文本编辑的功能，旨在让用户对输出结果或统计命令作必要的编辑，以便直接打印或通过 WINDOWS 的剪贴板剪切或拷贝后供其他文字处理器（如 Word、Wordperfect 等）使用。

3.2.1 文本中文字的增删与修改

激活结果输出窗口或命令编辑窗口后，用户可使用方向键和 Home、End、PageUp、PageDown 键或直接用鼠标（在文本区内，鼠标呈“I”状）移动和确定光标位置，以便进行文字的增删与修改。其中 ← 键为光标左移；→ 键为光标右移；↑ 键为光标上移；↓ 键为光标下移；Home 键为光标移至行头；End 键为光标移至行尾；Ctrl+Home 键为光标移至篇头；Ctrl+End 键为光标移至篇尾；PageUp 键为上翻一页；PageDown 键为下翻一页。

移动光标至所需位置时，即可进行文字的增删与修改。在默认情况下，编辑处于插入状态，用户在光标位置上击键即可插入文字；若想覆盖原有的文字，可先按 Insert 键关闭插入状态，这时键入的文字将逐一取代光标位置之后的原有文字；若想删除文字，则可使用 Delete 键和 Backpace 键，用 Delete 键可删除光标后面的文字，用 Backpace 键可删除光标前面的文字。

在结果输出窗口中，每隔几行文字，其最左边会显示一个 ⌘ 符号，这是打印分页符（有的是 ◊ 符号，两个 ◊ 之间的内容为一次完整统计过程的结果输出块）。对于一般的打印纸，当保留系统提供的分页符时，会出现每打印十数行就换页的情况，这样十分浪费纸张。故一般需要将分页符删除：即将光标移至分页符后按 Backpace 键即可消除分页符。

必要时，用户可重新对文本加入 ⌘ 符号和 ◊ 符号。选 Edit 菜单的 Add Page Break 命令项可加入 ⌘ 符号；选 Edit 菜单的 Add Output Break 命令项可加入 ◊ 符号。

3.2.2 文本的选择

上面所讲的方法用于少数几个文字的删除是很方便的，但实际工作中需要对几行或数段文字（即文本块）进行删除或移动，这时就需要应用文本选择方法。

将鼠标移至需选择的文本块之首，按住鼠标左键拖动鼠标，直至所需文本块全部选中后放开鼠标左键，被选中的文本块呈黑底白字；若感到拖动鼠标的操作有困难，也可改用键盘选择方式，即先将光标移至需选择的文本块之首，然后按住 Shift 键不放，再同时按方向键移动光标，便可选择所需的文本块。

还可调用 Edit 菜单的 Select 命令项进行文本块选择，它有几个选项：

1、All：窗口里的内容全部选择，可用于结果文本也可用于命令文本；

2、Page：窗口里当前区域内显示的一个页面的内容（即两个分页符之间的内容）被选择，只适用于结果文本；

3、Output Block：窗口里当前区域内显示的一个输出块的内容（即两个 ◊ 符之间的内容）被选择，只适用于结果文本；

4、Command：窗口里当前区域内显示的一个命令段的内容被选择，只适用于命令文本。

3.2.3 文本块的删除、移动与复制

完成文本块的选择之后，就可以进行所需的删除、移动或复制操作了。

1、删除：选好文本块后，按 Del 键或选 Edit 菜单的 Clear 命令项，即可将选好的文本块删除；

2、移动：已有的文本可能需要移到另一处，这时可先选好需要移到别处的文本块，再选 Edit 菜单的 Cut 命令项，将该文本块剪切送入 Windows 的剪贴板中（该文本块从原处消失），然后将光标移到所需的位置，选 Edit 菜单的 Paste 命令项，即完成文本块的移动；

3、复制：已有的文本可能在另一处也需要，这时可先选好该文本块，再选 Edit 菜单的 Copy 命令项，将该文本块拷入 Windows 的剪贴板中（该文本块在原处仍保留），然后将光标移到所需的位置，选 Edit 菜单的 Paste 命令项，即完成文本块的复制。

3.2.4 文本块的打印

被选取的文本块，可直接送打印机输出。选 File 菜单的 Print...命令项，弹出 Print 对话框，

系统默认选 Selection 项，用户确定打印份数后点击 OK 钮即可。

3.2.5 文本中文字的查找

激活结果输出窗口或命令编辑窗口，选 Edit 菜单的 Search For Text...命令项，弹出 Serach For Text 对话框（图 3.3），用户在 Serach for 框中输入需要查找的文字，然后确定是否忽略字母的大小写（Ignore case），点击 Serach Forward 钮可要求系统向后查找，点击 Search Backward 钮可要求系统向前查找。

图 3.3　文本文字查找对话框

3.2.6 文本中文字的替换

激活结果输出窗口或命令编辑窗口，选 Edit 菜单的 Replace Text...命令项，弹出 Replace Text 对话框（图 3.4），用户在 Serach for 框中输入替换前的文字，在 Replace with 框中输入替换后的文字，确定是否忽略字母的大小写（Ignore case），并确定系统的查找方向（向后为 Serach Forward，向前为 Search Backward）。点击 Search 钮，系统找到替换处时会暂停询问用户是否做替换操作，若要点击 Replace then Search 钮，系统替换后继续再查找；若不要可点击 Search 钮再查找或点击 Close 钮结束替换操作。用户在十分肯定的情况下可点击 Replace All 钮，系统将不做任何询问快速自动地全部替换。

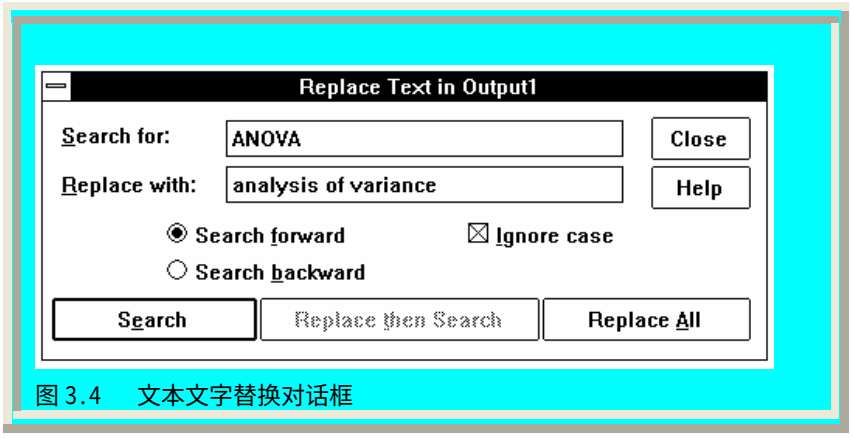


图 3.4　文本文字替换对话框

第四章 摘要性分析

摘要性分析是对原始数据进行描述性分析，这是统计工作的出发点。统计学的一系列基本描述指标，不仅让人了解资料的特征，而且可启发人们对之作进一步的深入分析。通过调用摘要性分析的诸个过程，可完成许多统计学指标，对于计量资料，可完成均数、标准差、标准误等指标的计算；对于计数和一些等级资料，可完成构成比、率等指标的计算和 χ^2 检验。本章将介绍其操作方法。

第一节 Frequencies 过程

4.1.1 主要功能

调用此过程可进行频数分布表的分析。频数分布表是描述性统计中最常用的方法之一，此外还可对数据的分布趋势进行初步分析。

4.1.2 实例操作

[例 4-1]调查 100 名健康女大学生的血清总蛋白含量（g%）如下表，试作频数表分析。

7.43	7.88	6.88	7.80	7.04	8.05	6.97	7.12	7.35	8.05
7.95	7.56	7.50	7.88	7.20	7.20	7.20	7.43	7.12	7.20
7.50	7.35	7.88	7.43	7.58	6.50	7.43	7.12	6.97	6.80
7.35	7.50	7.20	6.43	7.58	8.03	6.97	7.43	7.35	7.35
7.58	7.58	6.88	7.65	7.04	7.12	8.12	7.50	7.04	6.80
7.04	7.20	7.65	7.43	7.65	7.76	6.73	7.20	7.50	7.43
7.35	7.95	7.35	7.47	6.50	7.65	8.16	7.54	7.27	7.27
6.72	7.65	7.27	7.04	7.72	6.88	6.73	6.73	6.73	7.27
7.58	7.35	7.50	7.27	7.35	7.35	7.27	8.16	7.03	7.43
7.35	7.95	7.04	7.65	7.27	7.72	8.43	7.50	7.65	7.04

4.1.2.1 数据准备

激活数据管理窗口，定义血清总蛋白含量的变量名为 X，然后输入血清总蛋白含量的原始数据，结果见图 4.1。

图 4.1　输入血清总蛋白含量值

4.1.2.2 统计分析

激活 Statistics 菜单，选 Summarize 中的 Frequencies...命令项，弹出 Frequencies 对话框（图 4.2）。现欲对血清总蛋白含量值进行频数表分析，故在对话框左侧的变量列表中选 x，点击钮使之进入 Variable(s)框。同时可点击 Format...钮弹出 Frequencies：Format 对话框，在 Order by 栏中有四个选项:Ascending values 为根据数值大小按升序从小到大作频数分布;Descending values 为根据数值大小按降序从大到小作频数分布；Ascending counts 为根据频数多少按升序从少到多作频数分布；Descending counts 为根据频数多少按降序从多到少作频数分布。在 Page Formal 栏中可定义结果输出的格式。本例选 Ascending values 项后点击 Continue 钮返回 Frequencies 对话框。

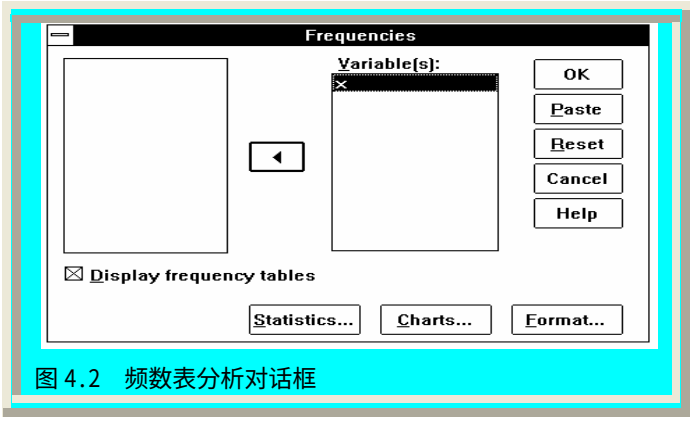


图 4.2　频数表分析对话框

点击 Statistics...钮，弹出 Frequencies:Statistics 对话框（图 4.3），可点击相应项目，要求系统在作频数表分析的基础上，附带作各种统计指标的描述，特别是可进行任何水平的百分位数计算。本例要求计算四分位数(Quartiles)、均数(Mean)、中位数(Median)、众数(Mode)、总和(Sum)、标准差(Std.deviation)、方差(Variance)、全距 (Range)、最小值(Minimum)、最大值(Maximum)、标准误(S.E.mean)、偏度系数（Skewness）和峰度系数(Kurtosis)，选好后点击 Continue 钮返回 Frequencies 对话框。

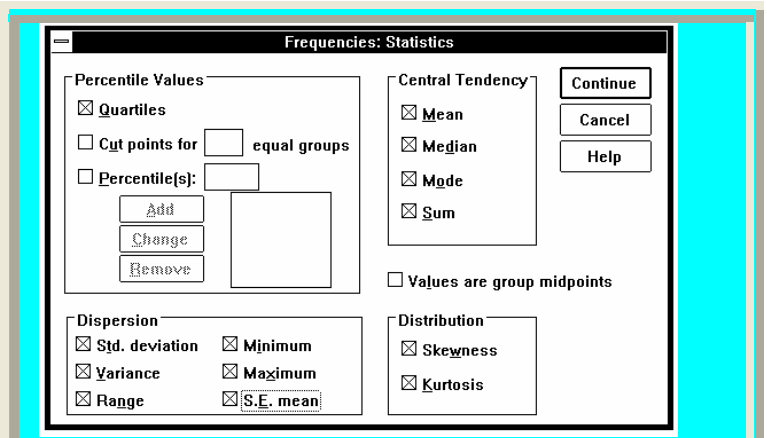


图 4.3　频数表分析的统计指标对话框

点击 Charts...钮，弹出 Frequencies:Charts 对话框，用户可选两种图形，一是直条图（Bar chart），适用于非连续性的变量；另一是直方图（Histogram），适用于连续性的变量。本例要求对变量 x 绘制直方图，故选择 Histogram 项，并要求绘制正态曲线(With normal curve)，点击 Continue 钮返回 Frequencies 对话框，再点击 OK 钮即可。

4.1.2.3 结果解释

在输出结果窗口中将看到如下统计数据：

系统对变量 x 的原始数据作频数分布表，Value 为原始值、Frequency 为频数、Percent 为各组频数占总例数的百分比、Valid percent 为各组频数占总例数的有效百分比、Cum Percent 为各组频数占总例数的累积百分比。

X				Valid	Cum
Value Label	Value	Frequency	Percent	Percent	Percent
	6.43	1	1.0	1.0	1.0
	6.50	2	2.0	2.0	3.0
	6.72	1	1.0	1.0	4.0
	6.73	4	4.0	4.0	8.0
	6.80	2	2.0	2.0	10.0
	6.88	3	3.0	3.0	13.0
	6.97	3	3.0	3.0	16.0
	7.03	1	1.0	1.0	17.0
	7.04	7	7.0	7.0	24.0
	7.12	4	4.0	4.0	28.0
	7.20	7	7.0	7.0	35.0
	7.27	7	7.0	7.0	42.0
	7.35	11	11.0	11.0	53.0
	7.43	8	8.0	8.0	61.0
	7.47	1	1.0	1.0	62.0
	7.50	7	7.0	7.0	69.0
	7.54	1	1.0	1.0	70.0
	7.56	1	1.0	1.0	71.0
	7.58	5	5.0	5.0	76.0

7.65	7	7.0	7.0	83.0
7.72	2	2.0	2.0	85.0
7.76	1	1.0	1.0	86.0
7.80	1	1.0	1.0	87.0
7.88	3	3.0	3.0	90.0
7.95	3	3.0	3.0	93.0
8.03	1	1.0	1.0	94.0
8.05	2	2.0	2.0	96.0
8.12	1	1.0	1.0	97.0
8.16	2	2.0	2.0	99.0
8.43	1	1.0	1.0	100.0
Total	100	100.0	100.0	

接着输出各基本统计指标,其中均数为7.366，标准误为0.039，中位数为7.350，众数为7.350，标准差为0.394，方差为0.155，峰度系数为0.034，峰度系数的标准误为0.478，偏度系数为0.06，偏度系数的标准误为 0.241，全距为 2.000，最小值为 6.430，最大值为 8.430，25%位数为 7.120，50%位数为 7.350，75%位数为 7.580，共 100 个观察值，无缺失值。

Mean	7.366	Std err	.039	Median	7.350
Mode	7.350	Std dev	.394	Variance	.155
Kurtosis	.034	S E Kurt	.478	Skewness	.060
S E Skew	.241	Range	2.000	Minimum	6.430
Maximum	8.430				

Percentile	Value	Percentile	Value	Percentile	Value
25.00	7.120	50.00	7.350	75.00	7.580

Valid cases	100	Missing cases	0

最后系统输出带有正态曲线的直方图（图 4.4），由图中可见，数据基本呈现正态分布形状。

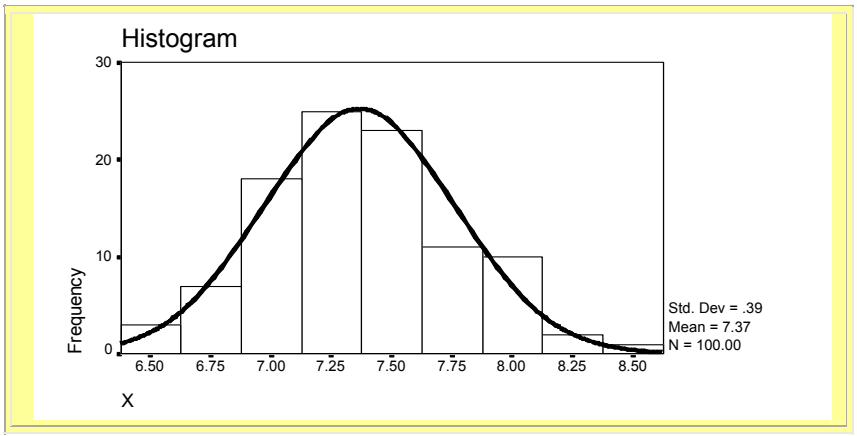


图 4.4　频数分布的直方图

从上述内容可知，系统在未特别指定的情形下，频数分布表是按照原始数值逐一作频数分布的，这与日常需要的等距分组、且组数保持在 8～15 组的要求不符。为此，在调用 Frequencies 过程命令之前，可先对原始数据进行算术处理：已知最小值为 6.430，最大值为 8.430，全距为 2.000，故可要求分成10组，起点为6.4，组距为0.2。选Transform菜单Recode项的Into Different Variable...命令项，在弹出的 Recode Into Different Variable 对话框中选 x 点击钮使之进入 Numeric Variable→Output Variable 框，在 Output Variable 栏的 Name 处输入 x1，点击 Change 钮表示新生成的变量名为 x1。点击 Old and New Values 钮弹出 Recode Into Different Variable:Old and New Values 对话框，在 Old value 栏内选 Range 项，输入第一个分组的数值范围：6.4～6.599，在 New value 栏内输入新值：6.4，点击 Add 钮，依此将各组的范围及对应的新值逐一输入，最后点击 Continue 钮返回 Recode Into Different Variable 对话框，再点击 OK 钮即完成。系统在原数据库中生成一新变量为 x1，这时调用 Frequencies 过程命令将输出等距分组且组数为 10 的频数分布表。

X1

Value Label	Value	Frequency	Percent	Valid Percent	Cum Percent
	6.40	3	3.0	3.0	3.0
	6.60	5	5.0	5.0	8.0
	6.80	8	8.0	8.0	16.0
	7.00	12	12.0	12.0	28.0
	7.20	25	25.0	25.0	53.0
	7.40	23	23.0	23.0	76.0
	7.60	10	10.0	10.0	86.0
	7.80	7	7.0	7.0	93.0
	8.00	6	6.0	6.0	99.0
	8.40	1	1.0	1.0	100.0
	Total	100	100.0	100.0	

Valid cases　100　Missing cases　0

第二节 Descriptives 过程

4.2.1 主要功能

调用此过程可对变量进行描述性统计分析，计算并列出一系列相应的统计指标，且可将原始数据转换成标准 Z 分值并存入数据库，所谓 Z 分值是指某原始数值比其均值高或低多少个标准差单位，高的为正值，低的为负值，相等的为零。

4.2.2 实例操作

[例 4-2]调查 20 名男婴的出生体重（克）资料如下，试作描述性统计。

2770	2915	2795	2995	2860	2970	3087	3126	3125	4654
2272	3503	3418	3921	2669	4218	3707	2310	2573	3881

4.2.2.1 数据准备

激活数据管理窗口，定义男婴出生体重的变量名为 X，然后输入男婴出生体重的原始数据。

4.2.2.2 统计分析

激活 Statistics 菜单选 Summarize 中的 Descriptives...命令项，弹出 Descriptives 对话框（图 4.5）。现欲对男婴出生体重进行描述性分析，故在对话框左侧的变量列表中选 x,点击钮使之进入 Variable(s)框;本例要求将原始数据转换成 z 分值，故选 Save standardized value as variables 项。

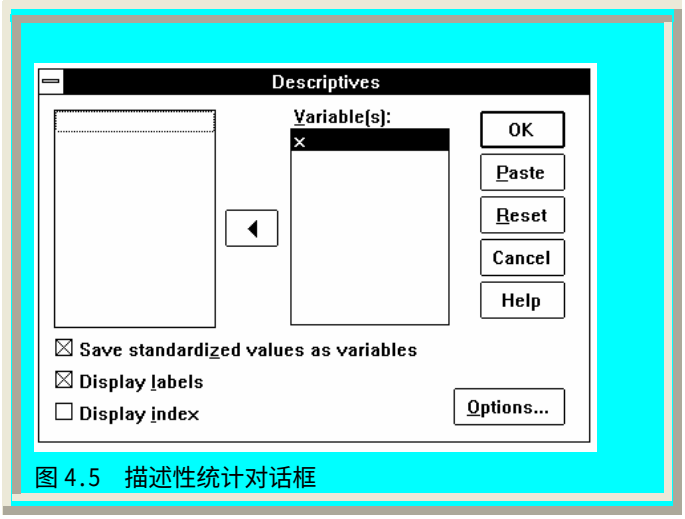


图 4.5　描述性统计对话框

点击 Options... 钮，弹出 Descriptives:Options 对话框（图 4.6）。框中各指标的意义请读者参阅本章第一节。选好项目后点击 Continue 钮返回 Descriptives 对话框，再点击 OK 钮即可。

图 4.6　描述性统计指标对话框

4.2.2.3 结果解释

在结果输出窗口中将看到如下统计数据：均数为 3188.450，标准误为 140.681，标准差为 629.146，方差为 395824.997，峰度系数为 0.118，峰度系数的标准误为 0.992，偏度系数为 0.732，偏度系数的标准误为 0.512，全距为 2382.000，最小值为 2272，最大值为 4654，有效例数为 100，无缺失值。

```
Number of valid observations (listwise) =        20.00
Variable  X

Mean          3188.450                S.E. Mean        140.681
Std Dev        629.146                Variance      395824.997
Kurtosis          .118                S.E. Kurt           .992
Skewness          .732                S.E. Skew           .512
Range         2382.000                Minimum         2272
Maximum       4654                    Sum           63769.000

Valid observations -       20         Missing observations -        0
```

此外，系统以zx为变量名将原始数据转换成标准z分值，存放在原数据库中（图 4.7）。例如，已知均数为 3188.450，标准差为 629.146，故原始值 2770 的Z分值为 $\frac{2770-3188.45}{629.146}$ = - 0.66511；原始值 2770 的Z分值为 $\frac{3881-3188.45}{629.146}$ = 1.10078。新变量具有均值为 0、标准差为 1 的特征，亦即变量的标准化过程。

Newdata

1:x 2770

	x	zx
1	2770	-.66511
2	2915	-.43464
3	2795	-.62537
4	2995	-.30748
5	2860	-.52206
6	2970	-.34722
7	3087	-.16125
8	3126	-.09926
9	3125	-.10085

图 4.7　原始数据及其标准 Z 分值

第三节 Explore 过程

4.3.1 主要功能

调用此过程可对变量进行更为深入详尽的描述性统计分析，故称之为探索性统计。它在一般描述性统计指标的基础上，增加有关数据其他特征的文字与图形描述，显得更加细致与全面，有助于用户思考对数据进行进一步分析的方案。

4.3.2 实例操作

[例 4-3]下表为 30 名 10 岁少儿的身高（cm）资料，试作探索性分析。

编号	身高		编号	身高	
	男孩	女孩		男孩	女孩
1	121.4	133.4	9	128.2	125.4
2	131.5	132.7	10	137.4	137.5
3	132.6	130.1	11	135.5	120.9
4	129.2	136.7	12	129.0	138.8
5	134.1	139.7	13	132.2	138.6
6	135.8	133.0	14	140.9	141.4
7	140.4	140.3	15	129.3	137.5
8	136.0	124.0			

4.3.2.1 数据准备

激活数据管理窗口，定义少儿身高的变量名为 X，然后再定义一个变质为 SEX，用于作性别分组。顺序输入少儿身高的原始数据，在变量 SEX 中，男孩输入 1、女孩输入 2。

4.3.2.2 统计分析

激活 Statistics 菜单选 Summarize 中的 Explore...项，弹出 Explore 对话框（如图 4.8），现欲对少儿身高资料进行分组的探索性分析，故在对话框左侧的变量列表中选 x 点击钮使之进入 Dependent List 框，再选 sex 点击钮使之进入 Factor List 框。

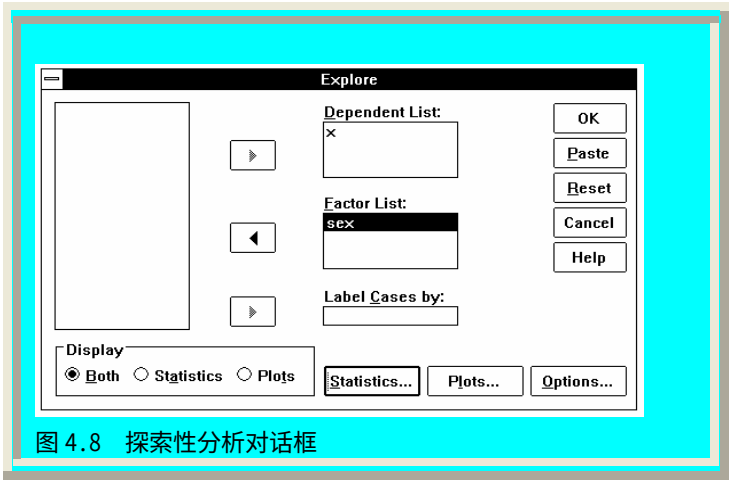


图 4.8　探索性分析对话框

点击 Statistics...钮，弹出 Explore:Statistics 对话框（图 4.9），有如下选项：

1、Descriptives：输出均数、中位数、众数、5%修正均数、标准误、方差、标准差、最小值、最大值、全距、四分位全距、峰度系数、峰度系数的标准误、偏度系数、偏度系数的标准误；

2、M-estimators：作中心趋势的粗略最大似然确定，输出四个不同权重的最大似然确定数；

3、Outliers：输出五个最大值与五个最小值；

4、Percentiles：输出第 5%、10%、25%、50%、75%、90%、95%位数；

5、Grouped Frequency tables：输出分组的例数与数值范围表。

本例全部选择，之后点击 Continue 钮返回 Explore 对话框。

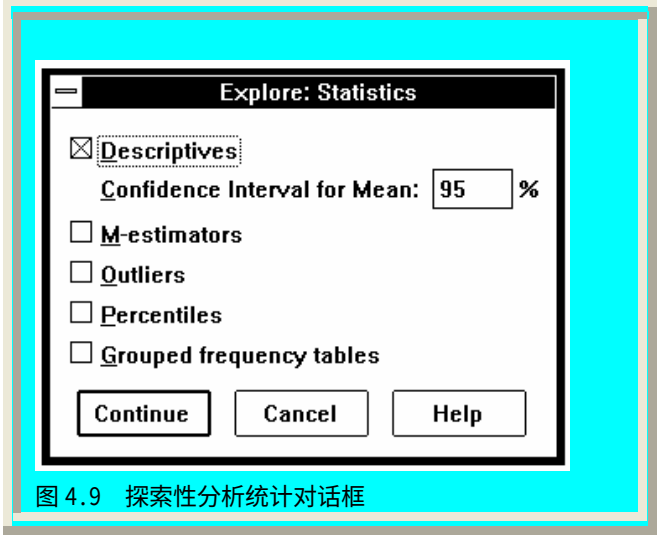


图 4.9　探索性分析统计对话框

点击 Plot...钮弹出 Explore:Plot 对话框(图 4.10),在 Boxplot 栏内选 Factor levels together 项要求按组别进行箱图绘制；在 Descriptive 栏内选 Stem-and-leaf 项要求作茎叶情形描述。之后点击 Continue 钮返回 Explore 对话框，再点击 OK 钮即可。

图 4.10　探索性分析绘图对话框

4.3.2.3 结果解释

在结果输出窗口中将看到如下统计数据：

先输出男孩的数据。共 15 例，无缺失值，其均数为 132.9，中位数为 132.6，5%修正均数为 133.0944，均数的 95%置信区间为 130.0706～135.7294，标准误为 1.3192，方差为 26.1043，标准差为 5.1092,最小值为 121.4,最大值为 140.9,全距为 19.5,四分位全距为 6.8,偏度系数为-0.4239，偏度系数的标准误为 0.5801，峰度系数为 0.4961，峰度系数的标准误为 1.1209。

接着输出四个不同权重下作中心趋势的粗略最大似然确定数，对于伴有长拖尾的对称分布数据或带有个别极端数值的数据，用粗略最大似然确定数替代均数或中位数，结果更准确。系统还进行数据的茎叶情形描述。如系统指出男孩的身高资料中，有一个数值是茎为 12，叶为 1，其实该数值是 121.4；有四个数值是茎为 12，叶为 8999，其实这些数值是 129.2、128.2、190.0、129.3。

再接着输出百分位数：第 5%位数是 121.4，第 10%数是 125.48，第 25%位数是 129.2，第 50%位数是 132.6，第 75%位数是 136，第 90%位数是 140.6。并输出最大五个数和最小五个数：最大五个数是 140.9，140.4，137.4，136.0，135.8；最小五个数是 121.4，128.2，129.0，129.2，129.3。

最后输出频数分布表。

```
 X
 By  SEX              1
 Valid cases:  15.0      Missing cases:  .0       Percent missing:  .0

Mean      132.9000   Std Err     1.3192   Min    121.4000   Skewness    -.4239
Median    132.6000   Variance   26.1043   Max    140.9000   S E Skew     .5801
5%Trim    133.0944   Std Dev     5.1092   Range   19.5000   Kurtosis     .4961
95% CI for Mean (130.0706, 135.7294)      IQR      6.8000   S E Kurt    1.1209

                              M-Estimators
                               ----------
 Huber  ( 1.339)               132.9127   Tukey  ( 4.685)              133.0901
 Hampel ( 1.700, 3.400, 8.500) 133.0153   Andrew ( 1.340 * pi )        133.0904

 Frequency     Stem &  Leaf
     1.00       12 *  1
     4.00       12 .  8999
```

```
     4.00       13 *  1224
     4.00       13 .  5567
     2.00       14 *  00
 Stem width:      10.0
 Each leaf:       1 case(s)

                              Percentiles
                              -----------
Percentiles   5.0000   10.0000   25.0000   50.0000   75.0000   90.0000  95.0000
Haverage    121.4000  125.4800  129.2000  132.6000  136.0000  140.6000
Tukey's Hinges                129.2500  132.6000  135.9000

                             Extreme Values
                             ------- ------
    5    Highest     Case #                5    Lowest      Case #
          140.9      Case: 14                    121.4      Case: 1
          140.4      Case: 7                     128.2      Case: 9
          137.4      Case: 10                    129.0      Case: 12
          136.0      Case: 8                     129.2      Case: 4
          135.8      Case: 6                     129.3      Case: 15

                            Frequency Table
                            --------- -----
                  Bin                                Valid      Cum
                 Center         Freq      Pct        Pct        Pct
                  126.4          5.00    33.33      33.33      33.33
                  136.4         10.00    66.67      66.67     100.00
```

下一部分为系统输出的女孩资料分析结果，其意义同上述。

```
X
 By  SEX              2
 Valid cases:  15.0     Missing cases:  .0     Percent missing:  .0

Mean      134.0000    Std Err     1.6428   Min     120.9000   Skewness    -.8937
Median    136.7000    Variance   40.4829   Max     141.4000   S E Skew     .5801
5% Trim   134.3167    Std Dev     6.3626   Range    20.5000   Kurtosis    -.2747
95% CI for Mean (130.4765, 137.5235)      IQR       8.7000   S E Kurt    1.1209

                              M-Estimators
                              ------------
 Huber   ( 1.339)               135.4183   Tukey   ( 4.685)              136.2104
```

Hampel (1.700, 3.400, 8.500) 135.1852 Andrew (1.340 * pi) 136.2327

Frequency Stem & Leaf
2.00 12 * 04
1.00 12 . 5
4.00 13 * 0233
6.00 13 . 677889
2.00 14 * 01
Stem width: 10.0
Each leaf: 1 case(s)

Percentiles

Percentiles	5.0000	10.0000	25.0000	50.0000	75.0000	90.0000	95.0000
Haverage	120.9000	122.7600	130.1000	136.7000	138.8000	140.7400	
Tukey's Hinges			131.4000	136.7000	138.7000		

Extreme Values

	Highest	Case #		Lowest	Case #
5	141.4	Case: 29	5	120.9	Case: 26
	140.3	Case: 22		124.0	Case: 23
	139.7	Case: 20		125.4	Case: 24
	138.8	Case: 27		130.1	Case: 18
	138.6	Case: 28		132.7	Case: 17

Frequency Table

Bin Center	Freq	Pct	Valid Pct	Cum Pct
125.9	4.00	26.67	26.67	26.67
135.9	10.00	66.67	66.67	93.33
145.9	1.00	6.67	6.67	100.00

此外，按用户要求，系统输出箱图。图中方箱为四分位数，中心粗线为中位数，两端线为最大值与最小值。

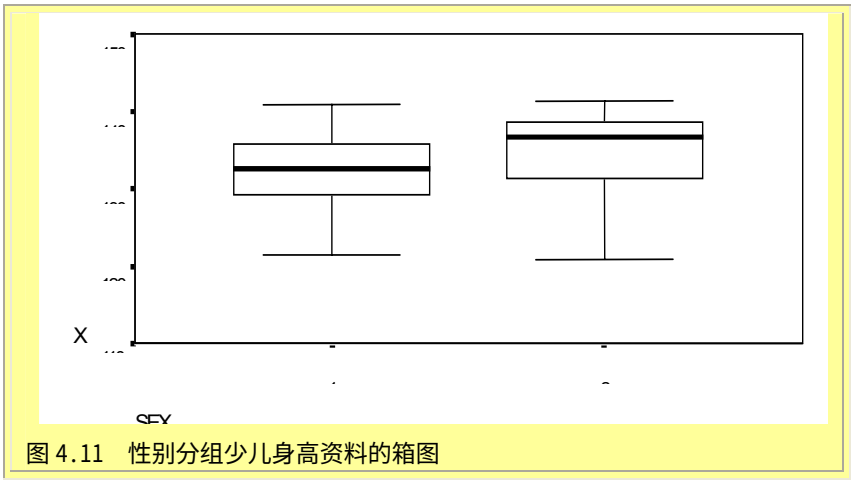


图 4.11　性别分组少儿身高资料的箱图

第四节 Crosstabs 过程

4.4.1 主要功能

调用此过程可进行计数资料和某些等级资料的列联表分析，在分析中，可对二维至n维列联表（RC表）资料进行统计描述和 χ^2 检验，并计算相应的百分数指标。此外，还可计算四格表确切概率（Fisher's Exact Test）且有单双侧（ One-Tail、 Two-Tail），对数似然比检验（Likelihood Ratio）以及线性关系的Mantel-Haenszel χ^2 检验。

4.4.2 实例操作

[例 4-4]用两组大白鼠诱发鼻咽癌的动物实验中，一组单纯用亚硝胺鼻注，另一组附加维生素B_{12}，生癌率如下表，问两组生癌率有无差别？

动物分组	生癌鼠数	未生癌鼠数	合计	生癌率(%)
亚硝胺组	52	19	71	73.2
亚硝胺+B_{12}组	39	3	42	92.9
合计	91	22	113	80.5

4.4.2.1 数据准备

激活数据管理窗口，定义变量名：count 为频数变量（行列对应的频数值），group 为组变量（行），test 为试验结果变量（列）。按顺序输入相应的变量（图 4.12）。

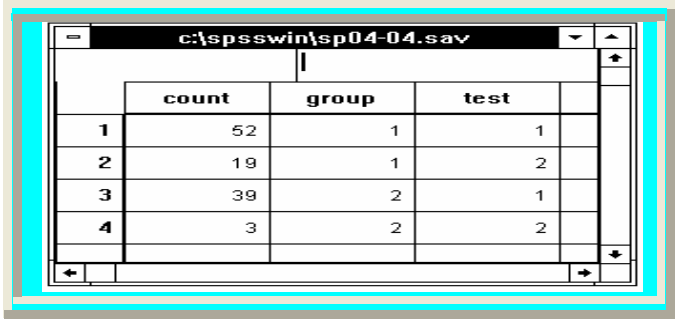


图 4.12　原始数据的输入

4.4.2.2 统计分析

在进行计数资料的分析前，应对频数变量的值进行加权处理。先激活 Data 菜单，选 Weight Cases...项，弹出 Weight Cases 对话框，选 Weight cases by，再选变量 count 点击钮使之进入 Frequence Variable 框中，点击 OK 钮完成加权。

激活Statistics菜单，选Summarize中的Crosstabs...项，弹出Crosstabs对话框（如图 4.13 示）。在Crosstabs对话框中，选group点击钮使之进入Row(s)框，选test点击钮使之进入Column(s)框。点击Statistics...钮，弹出Crosstabs:Statistics对话框（图 4.14），其中Chi-square即为读者所熟悉的 x^2 检验。由于在实际研究中，变量间的依赖强度和特征也是需要考虑的，x^2 值不是列联强度的好的度量，故用户可根据实际需要选择其他相关的指标：

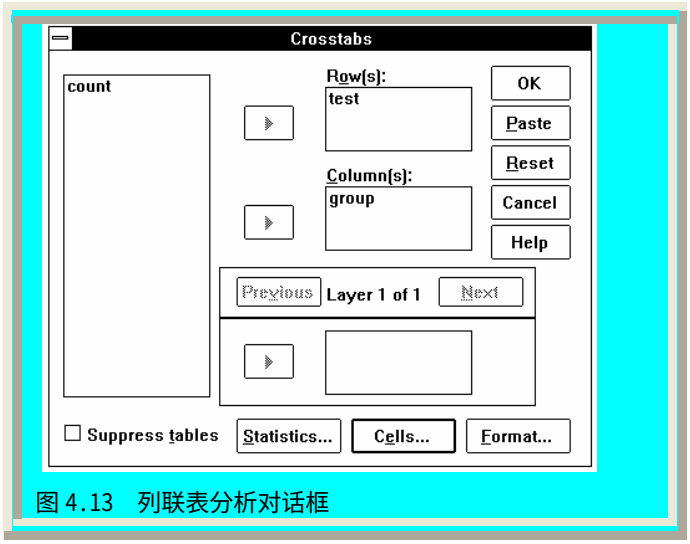


图 4.13　列联表分析对话框

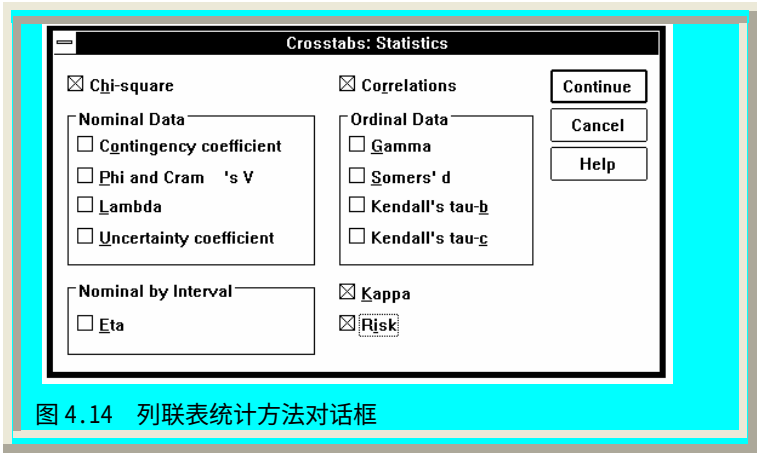


图 4.14　列联表统计方法对话框

1、定距变量的关联指标

Correlations：可作列联表行、列两变量的 Pearson 相关系数或作伴随组秩次的 Spearman 相关系数。

2、定类变量的关联指标

Contingency coefficient：列联系数，其值 = $\sqrt{\frac{x^2}{x^2+N}}$，界于 0～1 之间，其中N为总例数；

Phi and Cramer's V：Ψ系数 = $\sqrt{\frac{\chi^2}{N}}$，用于描述相关程度，在四格表 χ^2 检验中界于-1～1之间，在RC表 χ^2 检验中界于 0～1 之间；Cramer's V = $\sqrt{\frac{\chi^2}{N(k-1)}}$，界于 0～1 之间，其中k为行数和列数较小的实际数；

Lambda：λ 值，在自变量预测中用于反映比例缩减误差，其值为 1 时表明自变量预测应变量好，为 0 时表明自变量预测应变量差；

Uncertainty coefficient：不确定系数，以熵为标准的比例缩减误差，其值接近 1 时表明后一变量的信息很大程度来自前一变量，其值接近 0 时表明后一变量的信息与前一变量无关。

3、定序变量的关联指标

Gamma：γ 值 = $\frac{P-Q}{P+Q}$，P为同序对子数，Q为异序对子数，界于 0～1 之间，所有观察实际数集中于左上角和右下角时，其值为 1；

Somers'D：Somers'D值 = $\frac{P-Q}{P+Q+T_{vd}}$，T_{vd} 为独立变量上不存在同分的偶对中，同序对子数超过异序对子数的比例；

Kendall's tau-b：Kendall τ b = $\frac{P-Q}{\sqrt{(p+Q+T_{v1})(P+Q+T_{v2})}}$，$T_{v1}$ 为在V1 变量上是同序在V2 变量上不是的对子数，T_{v2} 为在V2 变量上是同序在V1 变量上不是的对子数，Kendall τ b 值界于-1～1 之间；

Kendall's tau-c：Kendall τ c = $\frac{2m(P-Q)}{N^2(m+1)}$，m为行数和列数较小的实际数，Kendall τ c 值界于-1～1 之间。

4、其他指标

Kappa：内部一致性系数；

Eta：Eta 值，其平方值可认为是应变量受不同因素影响所致方差的比例；

Risk：相对危险度。

点击 Cells...钮,弹出 Crosstabs:Cells 对话框（图 4.15），用于定义列联表单元格中需要计算的指标。Observed 为实际观察数，Expected 为理论数，Row 为行百分数，Column 为列百分数，Total 为合计百分数，Raw 为实际数与理论数的差值，Standardized 为实际数与理论数的差值除理论数，Adj. Standardized 为由标准误确立的单元格残差。选择后点击 Continue 钮返回 Crosstabs 对话框，再点击 OK 钮即可。

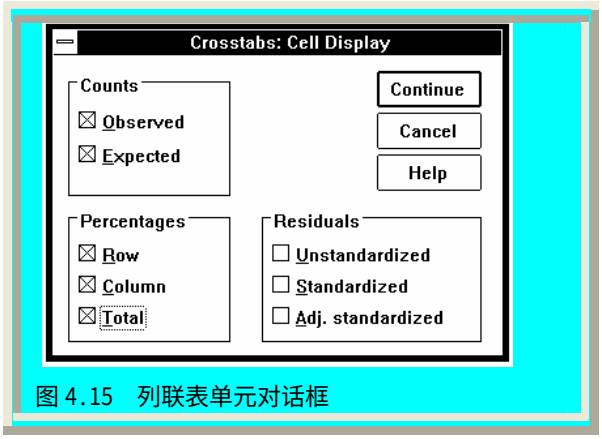


图 4.15　列联表单元对话框

4.4.2.3 结果解释

在结果输出窗中，系统先输出四格表资料，包括实际观察数、理论数、行百分数、列百分数和合计百分数。

TEST by GROUP

		GROUP		
	Count / Exp Val / Row Pct / Col Pct / Tot Pct	1	2	Row Total
TEST	1	52	39	91
		57.2	33.8	80.5%
		57.1%	42.9%	
		73.2%	92.9%	
		46.0%	34.5%	
	2	19	3	22
		13.8	8.2	19.5%
		86.4%	13.6%	
		26.8%	7.1%	
		16.8%	2.7%	
	Column Total	71	42	113
		62.8%	37.2%	100.0%

接着输入有关统计数据，Pearson χ^2值为 6.47766，P值为 0.01092，可认为亚硝胺+B_{12}组的生癌率较高；校正 χ^2值为 5.28685，P值为 0.02149；M-T检验χ^2值为 6.42034，P值为 0.01128；最小理论数为 8.177，故不需作精确概率计算。（如果四格表中有理论频数小于 5 时，Crosstabs命令会自动进行Fisher精确概率计算）。

内部一致性系数为-0.21731，Pearson 相关系数和 Spearman 相关系数均为 0.23943。

第一组对第二组的相对危险性 RR 值为 21%左右（0.21053），即可认为第二组生癌的相对危险性为第一组的 4.75 倍。

Chi-Square	Value	DF	Significance
Pearson	6.47766	1	.01092
Continuity Correction	5.28685	1	.02149
Likelihood Ratio	7.31007	1	.00686
Mantel-Haenszel test for linear association	6.42034	1	.01128

Minimum Expected Frequency - 8.177

Statistic	Value	ASE1	Val/ASE0	Approximate Significance
Kappa	-.21731	.07083	-2.54513	
Pearson's R	-.23943	.07447	-2.59807	.01065 *4
Spearman Correlation	-.23943	.07447	-2.59807	.01065 *4

*4 VAL/ASE0 is a t-value based on a normal approximation, as is the significance

Statistic	Value	95% Confidence Bounds	
Relative Risk Estimate (TEST 1 / TEST 2) :			
case control	.21053	.05816	.76211
cohort (GROUP 1 Risk)	.66165	.51872	.84397
cohort (GROUP 2 Risk)	3.14286	1.06940	9.23654

Number of Missing Observations: 0

第五章 平均水平的比较

在正态或近似正态分布的计量资料中（如临床常见的体温、血压、脉搏、身高、体重等测量值，几乎均为此类资料），经常在使用前一章计量资料描述过程分析后，还要进行组与组之间平均水平的比较。本章将分四节分别介绍这一统计方法：即常用的 t 检验和单因素方差分析。

第一节 Means 过程

5.1.1 主要功能

与第四章中 Descriptives 过程相比，若仅仅计算单一组别的均数和标准差，Means 过程并无特别之处；但若用户要求按指定条件分组计算均数和标准差，如分性别同时分年龄计算各组的均数和标准差，则用 Means 过程更显简单快捷。

5.1.2 实例操作

［例 5 .1］某医师测得如下血红蛋白值（g%），试作基本的描述性统计分析：

对象编号	性别	年龄	血红蛋白值	对象编号	性别	年龄	血红蛋白值
1	女	18	12.83	21	女	16	11.36
2	男	16	15.50	22	男	16	12.78
3	女	18	12.25	23	男	18	15.09
4	女	17	10.06	24	女	18	8.67
5	男	16	10.88	25	女	17	8.56
6	男	18	9.65	26	女	18	12.56
7	女	16	8.36	27	女	17	11.56
8	男	18	11.66	28	男	16	14.67
9	女	18	8.54	29	男	16	7.88
10	女	17	7.78	30	男	18	12.35
11	男	18	13.66	31	男	16	13.65
12	男	18	10.57	32	女	16	9.87
13	男	16	12.56	33	女	18	10.09
14	女	17	9.87	34	女	18	12.55
15	女	17	8.99	35	男	18	16.04
16	女	17	11.35	36	男	18	13.78
17	男	17	14.56	37	男	17	11.67
18	男	16	12.40	38	男	17	10.98
19	女	16	8.05	39	女	16	8.78
20	男	18	14.03	40	男	16	11.35

5.1.2.1 数据准备

激活数据管理窗口，定义变量名：性别为 sex，年龄为 age，血红蛋白值为 hb。按顺序输入数据(sex 变量中，男为 1，女为 2)，结果见图 5.1。

图 5.1　原始数据的输入

5.1.2.2　统计分析

激活 Statistics 菜单选 Compare Means 中的 Means...项，弹出 Means 对话框（如图 5.2 示）。今欲分性别同时分年龄求血红蛋白值的均数和标准差，故在对话框左侧的变量列表中选 hb，点击 ➢ 钮使之进入 Dependent List 框，选 sex 点击 ➢ 钮使之进入 Independent List 框，点击 Next，可选定分组的第二层次（Layer 2 of 2），选 age 点击 ➢ 钮亦使之进入 Independent List 框。点击 Options...可选统计项目：在 Cell Displays 项中，Mean 为均数、Standard deviation 为标准差、Variance 为方差、Count 为观察单位数、Sum 为观察值总和，在 Statistics for First Layer 项中，将为第一层次的分组计算方差分析（ANOVA table and eta）和线性检验（Test of linearity）。选好后点击 Continue 钮返回 Means 对话框，点击 OK 钮即可。

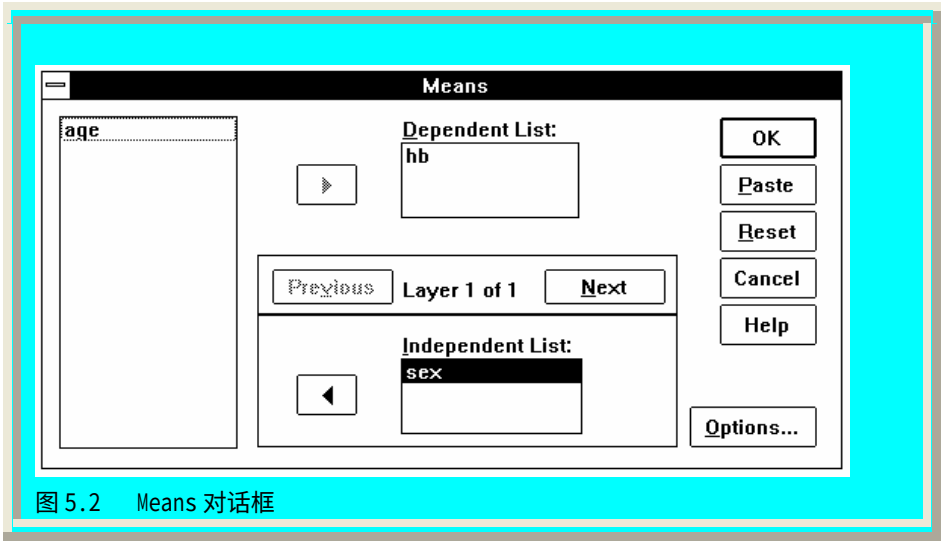


图 5.2　Means 对话框

5.1.2.3　结果解释

在结果输出窗口中将看到如下统计数据：

- - Description of Subpopulations - -
Summaries of HB
By levels of SEX
AGE

Variable	Value	Label	Sum	Mean	Std Dev	Variance	Cases
For Entire Population			457.79	11.4448	2.2690	5.1484	40
SEX	1		265.71	12.6529	2.0531	4.2154	21
AGE	16		111.67	12.4078	2.2455	5.0423	9
AGE	17		37.21	12.4033	1.8993	3.6074	3
AGE	18		116.83	12.9811	2.0933	4.3821	9
SEX	2		192.08	10.1095	1.6989	2.8863	19
AGE	16		46.42	9.2840	1.3494	1.8209	5
AGE	17		68.17	9.7386	1.4036	1.9700	7
AGE	18		77.49	11.0700	1.9158	3.6703	7

Total Cases = 40

For Entire Population 一行表示 40 个观察值合计为 457.79，均数为 11.4448，标准差为 2.2690，方差为 5.1484，例数为 40；接下去各行分别表示先按性别分组（分男性与女性），再按年龄分组（16，17，18 岁三组）的观察值合计、均数、标准差、方差和例数。

若在 Independent List 中未分层次，即 sex 和 age 一起放在 Layer 1 of 1 中，则结果是分别计算男性与女性（不作年龄分组）、16，17，18 岁三组（不作性别分组）的观察值合计、均数、标准差、方差和例数（如下所示）。

- - Description of Subpopulations - -

Summaries of HB
By levels of SEX

Variable	Value	Label	Sum	Mean	Std Dev	Variance	Cases
For Entire Population			457.79	11.4447	2.2690	5.1484	40
SEX	1		265.71	12.6529	2.0531	4.2154	21
SEX	2		192.08	10.1095	1.6989	2.8863	19

Total Cases = 40

Summaries of HB
By levels of AGE

Variable	Value	Label	Sum	Mean	Std Dev	Variance	Cases
For Entire Population			457.79	11.4448	2.2690	5.1484	40
AGE	16		158.09	11.2921	2.4649	6.0759	14
AGE	17		105.38	10.5380	1.9421	3.7719	10

AGE	18	194.32	12.1450	2.1827	4.7640	16
Total Cases = 40						

第二节 Independent-Samples T Test 过程

5.2.1 主要功能

调用此过程可完成两样本均数差别的显著性检验，即通常所说的两组资料的 t 检验。

5.2.2 实例操作

［例 5 .2］分别测得 14 例老年性慢性支气管炎病人及 11 例健康人的尿中 17 酮类固醇排出量（mg/dl）如下，试比较两组均数有无差别。

病　人	2.90	5.41	5.48	4.60	4.03	5.10	4.97	4.24	4.36	2.72	2.37	2.09	7.10	5.92
健康人	5.18	8.79	3.14	6.46	3.72	6.64	5.60	4.57	7.71	4.99	4.01			

5.2.2.1 数据准备

激活数据管理窗口，定义变量名：把实际观察值定义为 x，再定义一个变量 group 来区分病人与健康人。输入原始数据，在变量 group 中，病人输入 1，健康人输入 2。结果如图 5.3 所示。

Newdata

16:group 2

	x	group			
11	2.37	1			
12	2.09	1			
13	7.10	1			
14	5.92	1			
15	5.18	2			
16	8.79	2			
17	3.14	2			
18	6.46	2			
19	3.72	2			

图 5.3　两组资料 t 检验的原始数据

5.2.2.2 统计分析

激活 Statistics 菜单选 Compare Means 中的 Independent-samples T Test...项，弹出 Independent-samples T Test 对话框（如图 5.4 示）。从对话框左侧的变量列表中选 x，点击 ➢ 钮使之进入 Test Variable(s)框，选 group 点击 ➢ 钮使之进入 Grouping Variable 框，点击 Define Groups...钮弹出 Define Groups 定义框，在 Group 1 中输入 1，在 Group 2 中输入 2，点击 Continue 钮，返回 Independent-samples T Test 对话框，点击 OK 钮即完成分析。

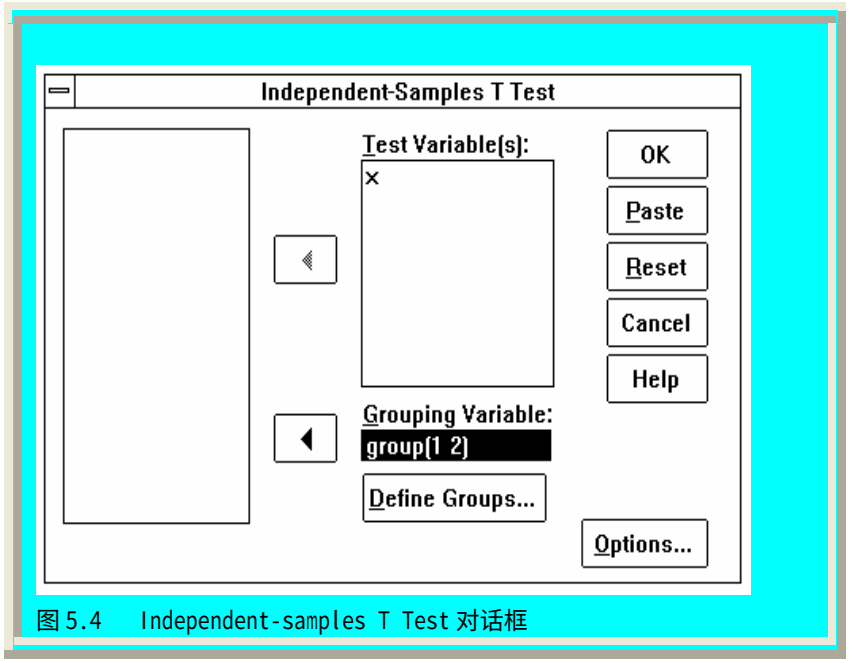


图 5.4 Independent-samples T Test 对话框

5.2.2.3 结果解释

在结果输出窗口中将看到如下统计数据：

```
t-tests for independent samples of  GROUP
                          Number
          Variable       of Cases     Mean        SD       SE of Mean
        -----------------------------------------------------------------
          X
          GROUP 1            14       4.3779      1.450        .387
          GROUP 2            11       5.5282      1.735        .523
        -----------------------------------------------------------------
          Mean Difference = -1.1503
          Levene's Test for Equality of Variances: F= .440   P= .514
```

这一部分显示两组资料的例数（Numbers of cases）、均数（Mean）、标准差（SD）和标准误（SE of Mean），显示两均数差值为 1.1503，经方差齐性检验： F= .440 P= .514，即两方差齐。

```
t-test for Equality of Means                                  95%
  Variances  t-value   df    2-Tail Sig    SE of Diff      CI for Diff
```

```
-----------------------------------------------------------------------
 Equal      -1.81     23          .084        .637        (-2.468, .167)
 Unequal    -1.77     19.47       .093        .651        (-2.513, .213)
-----------------------------------------------------------------------
```

这一部分显示 t 检验的结果，第一行表示方差齐情况下的 t 检验的结果，第二行表示方差不齐情况下的 t 检验的结果。依次显示值（t-value）、自由度（df）、双侧检验概率（2-Tail Sig）、差值的标准误（SE of Diff）及其 95%可信区间（Cl for Diff）。因本例属方差齐性，故采用第一行（即 Equal）结果：t=1.81,P=0.084,差别有显著性意义，即老年性慢性支气管炎病人的尿中 17 酮类固醇排出量低于健康人。

第三节 Paired-Samples T Test 过程

5.3.1 主要功能

调用此过程可完成配对资料的显著性检验，即配对 t 检验。在医学领域中，主要的配对资料包括：同对（年龄、性别、体重、病况等非处理因素相同或相似者）或同一研究对象分别给予两种不同处理的效果比较，以及同一研究对象处理前后的效果比较。前者推断两种效果有无差别，后者推断某种处理是否有效。

5.3.2 实例操作

［例 5 .2］某单位研究饲料中缺乏维生素 E 与肝中维生素 A 含量的关系，将大白鼠按性别、体重等配为 8 对，每对中两只大白鼠分别喂给正常饲料和维生素 E 缺乏饲料，一段时期后将之宰杀，测定其肝中维生素 A 含量(μmol/L)如下，问饲料中缺乏维生素 E 对鼠肝中维生素 A 含量有无影响？

大白鼠对别	肝中维生素 A 含量（μmol/L）	
	正常饲料组	维生素 E 缺乏饲料组
1	37.2	25.7
2	20.9	25.1
3	31.4	18.8
4	41.4	33.5
5	39.8	34.0
6	39.3	28.3
7	36.1	26.2
8	31.9	18.3

5.3.2.1 数据准备

激活数据管理窗口，定义变量名：正常饲料组测定值为 x1，维生素 E 缺乏饲料组测定值为 x2，数据输入后结果如图 5.5 所示。

	x1	x2
1	37.20	25.70
2	20.90	25.10
3	31.40	18.80
4	41.40	33.50
5	39.80	34.00
6	39.30	28.30
7	36.10	26.20
8	31.90	18.30

图 5.5　配对 t 检验的原始数据

5.3.2.2 统计分析

激活 Statistics 菜单选 Compare Means 中的 Paired-samples T Test...项，弹出 Paried-samples T Test 对话框（如图 5.6 示）。从对话框左侧的变量列表中点击 x1，这时在左下方的 Current Selections 框中 Variable 1 处出现 x1，再从变量列表中点击 x2，左下方的 Current Selections 框中 Variable 2 处出现 x2。点击 ➢ 钮使 x1、x2 进入 Variables 框，点击 OK 钮即完成分析。

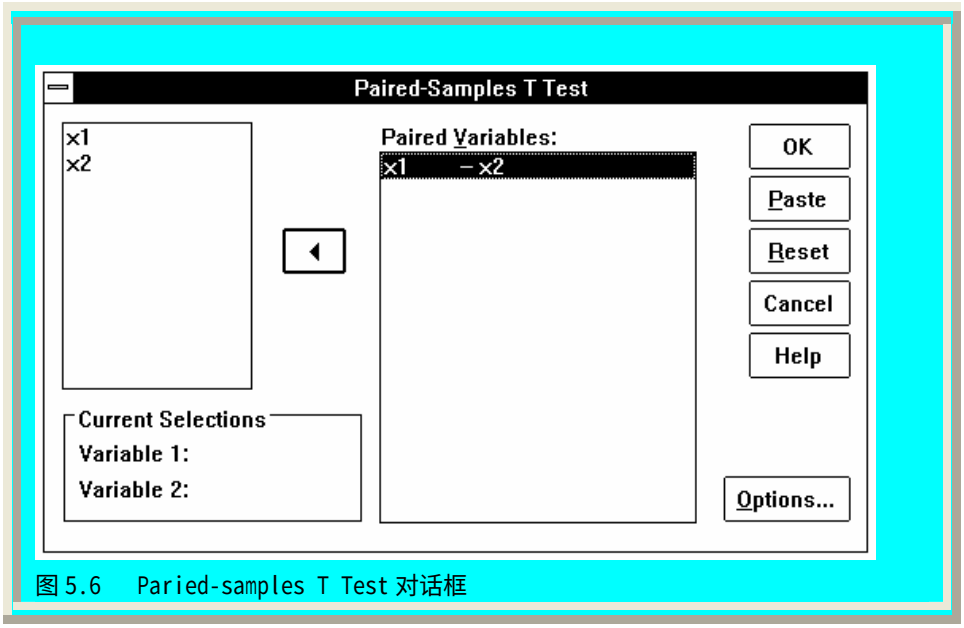


图 5.6　Paried-samples T Test 对话框

5.3.2.3 结果解释

在结果输出窗口中将看到如下统计数据：

- - - t-tests for paired samples - - -

Number of Variable	2-tail pairs	Corr	Sig	Mean	SD	SE of Mean
X1				34.7500	6.649	2.351
	8	.586	.127			
X2				26.2375	5.821	2.058

这段结果显示本例共有 8 对观察值，相关系数(C)为 0.586，相关系数的显著性检验表明 P=0.127；变量 x1 的均数（Mean）、标准差（SD）、标准误（SE of Mean）分别为 34.7500、6.649、2.351，变量 x2 的均数、标准差、标准误分别为 26.2375、5.821、2.058。

Paired Differences					
Mean	SD	SE of Mean	t-value	df	2-tail Sig
8.5125	5.719	2.022	4.21	7	.004
95% CI (3.730, 13.295)					

这段结果显示变量 x1、x2 两两相减的差值均数、标准差、标准误 95%可信区间（95% Cl）分别为 8.5125、5.719、2.022，95%可信区间(95% Cl)为 3.730，13.295。配对检验结果为：t=4.21，P=0.004，差别具高度显著性意义，即饲料中缺乏维生素 E 对鼠肝中维生素 A 含量确有影响。

第四节 One-Way ANOVA 过程

5.4.1 主要功能

在实际研究中，经常需要比较两组以上样本均数的差别，这时不能使用 t 检验方法作两两间的比较（如有人对四组均数的比较，作 6 次两两间的 t 检验），这势必增加两类错误的可能性（如原先 α 定为 0.05，这样作多次的 t 检验将使最终推断时的 α>0.05）。故对于两组以上的均数比较，必须使用方差分析的方法，当然方差分析方法亦适用于两组均数的比较。方差分析可调用此过程可完成。

本过程只能进行单因素方差分析，即完全随机设计资料的方差分析。对于随机区组设计资料方差分析的方法，将在第五章介绍。

5.4.2 实例操作

［例 5 .4］某单位研究两种不同制剂治疗钩虫的效果，用大白鼠作试验。11 只大白鼠随机分配于 3 组：一组为对照组、另外二组分别为使用甲、乙制剂的实验组。试验方法是：用药前每鼠人工感染 500 条钩蚴，感染后第 8 天实验组分别给予甲、乙制剂，对照组不给药，第 10 天全部解剖检查鼠体内活虫数，结果如下，问两制剂是否有效？

对照组	甲制剂组	乙制剂组
279	129	210
334	174	285
303	110	117
338		
298		

5.4.2.1 数据准备

激活数据管理窗口，定义变量名：实际观察值定义为 x，组别用变量 range 表示：其中对照组的值为、甲制剂实验组的值为、乙制剂实验组的值为，输入后的结果如图 5.7 所示。

	x	range			
4	338	1			
5	298	1			
6	129	2			
7	174	2			
8	110	2			
9	210	3			
10	285	3			
11	117	3			

图 5.7　单因素方差分析的原始数据

5.4.2.2 统计分析

激活 Statistics 菜单选 Compare Means 中的 One-Way ANOVA...项，弹出 One-Way ANOVA 对话框（如图 5.8 示）。从对话框左侧的变量列表中选 x，点击 ➢ 钮使之进入 Dependent List 框，选 range 点击 ➢ 钮使之进入 Factor 框，点击 Define Range 钮打开 One-Way ANOVA: Define Range 对话框，因本例为 3 组比较，故在 Minimum 处输入 1，在 Maximum 处输入 3，点击 Continue 钮返回 One-Way ANOVA 对话框。如果欲作多个样本均数间两两比较，可点击该点击对话框的 Post Hoc...钮打开

One-Way ANOVA: Post Hoc Multiple Comparisons 对话框（如图 5.9 所示），这时可见在 Tests 框中有 7 种比较方法供选择：

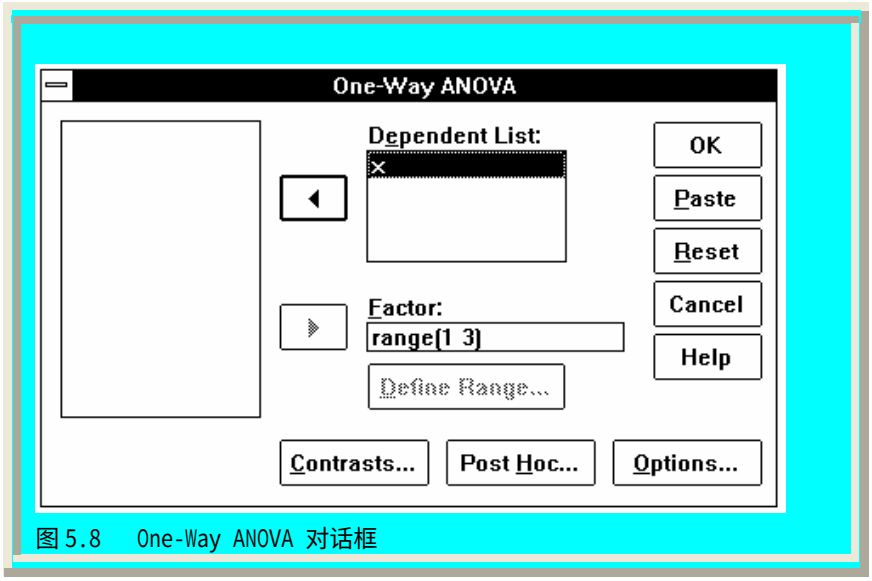


图 5.8　One-Way ANOVA 对话框

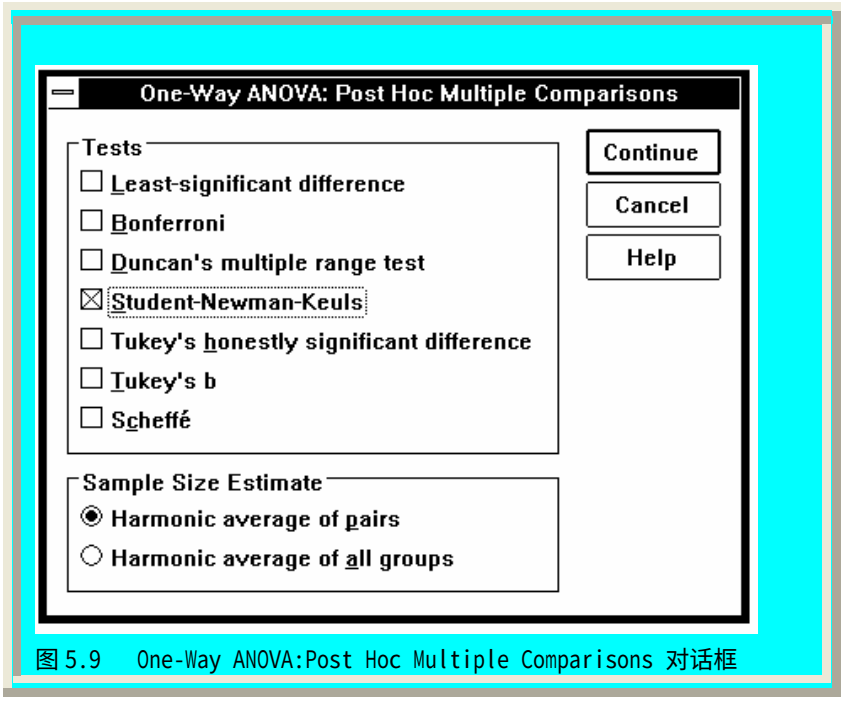


图 5.9　One-Way ANOVA:Post Hoc Multiple Comparisons 对话框

Least-significant difference：最小显著差法。α 可指定 0~1 之间任何显著性水平，默认值为 0.05；

Bonferroni：Bonferroni 修正差别检验法。α 可指定 0~1 之间任何显著性水平，默认值为 0.05；

Duncan’s multiple range test：Duncan 多范围检验。只能指定 α 为 0.05 或 0.01 或 0.1，默认值为 0.05；

Student-Newman-Keuls：Student-Newman-Keuls 检验，简称 N-K 检验,亦即 q 检验。α 只能为 0.05；

Tukey’s honestly significant difference：Tukey 显著性检验。α 只能为 0.05；

Tukey’s b：Tukey 另一种显著性检验。α 只能为 0.05；

Scheffe：Scheffe 差别检验法。α 可指定 0~1 之间任何显著性水平，默认值为 0.05。

本例选用 Student-Newman-Keuls 显著性检验法。在 Sample Size Estimate 框中有 Harmonic average of

pairs 和 Harmonic average of all groups 两选项，前者表示仅采用相互比较两组的调和均数，后者表示采用所有组（含比较的两组和尚未比较的其他组）的调和均数，本例选用前者，点击 Continue 钮返回 One-Way ANOVA 对话框后，再点击 OK 钮即完成分析。

5.4.2.3 结果解释

在结果输出窗口中将看到如下统计数据：

```
- - - - -   O N E W A Y   - - - - -
        Variable    X
     By Variable    RANGE
                                   Analysis of Variance

                                    Sum of            Mean               F        F
            Source           D.F.   Squares           Squares            Ratio    Prob.
Between Groups                 2    59724.3152        29862.1576         12.6804  .0033
Within Groups                  8    18839.8667        2354.9833
Total                         10    78564.1818
```

上述结果显示组间、组内（实际上本例应称之为“剩余”）和合计的自由度（D.F.）、离均差平方和（Sum of Squares,即 SS）、均方（Means Squares,即 SS）、F 值（F Ratio）和 P 值（F Prob.），本例 F=12.6804，P=0.0033，表明甲、乙两种制剂中必有一种制剂治疗钩虫是有效的。

为了解哪一种制剂是有效的，本例采用 SNK 两两比较法，结果如下：

```
- - - - -   O N E W A Y   - - - - -
        Variable    X
     By Variable    RANGE

Multiple Range Tests:   Student-Newman-Keuls test with significance level .050

The difference between two means is significant if
   MEAN(J)-MEAN(I)   >= 34.3146 * RANGE * SQRT(1/N(I) + 1/N(J))
   with the following value(s) for RANGE:

 Step          2         3
 RANGE         3.27      4.04
    (*) Indicates significant differences which are shown in the lower triangle
                                  G G G
                                  r r r
                                  p p p
                                  2 3 1
      Mean         RANGE
    137.6667       Grp 2
```

```
204.0000      Grp 3
310.4000      Grp 1       * *
```

上述结果显示：如果两均数的差值 ≥ 34.3146×RANGE×$\sqrt{\frac{1}{n1}+\frac{1}{n2}}$，则差别有显著性意义。上面已用“*”标出 2、3 两组与 1 组比较均有显著性差异。具体作法是：以甲制剂与对照组的比较为例，均数差值 = 310.4000 - 137.6667 = 172.7333，已知RANGE为 4.04，n1=5，n2=3，按上式求得 101.2418，因 172.7333 > 101.2418，故甲制剂有效；余同。即甲、乙制剂治疗钩虫均有效。因甲制剂与乙制剂比较，均数差值为 66.3333，按上式求得界值为 91.6180，故尚无证据表明甲、乙制剂间效果有差别。

第六章 方差分析

方差分析是 R.A.Fister 发明的，用于两个及两个以上样本均数差别的显著性检验。由于各种因素的影响，研究所得的数据呈现波动状，造成波动的原因可分成两类，一是不可控的随机因素，另一是研究中施加的对结果形成影响的可控因素。方差分析的基本思想是：通过分析研究中不同来源的变异对总变异的贡献大小，从而确定可控因素对研究结果影响力的大小。

方差分析主要用于：1、均数差别的显著性检验，2、分离各有关因素并估计其对总变异的作用，3、分析因素间的交互作用，4、方差齐性检验。

第一节 Simple Factorial 过程

6.1.1 主要功能

调用此过程可对资料进行方差分析或协方差分析。在方差分析中可按用户需要作单因素方差分析（其结果将与第五章第四节相同）或多因素方差分析（包括医学中常用的配伍组方差分析）；当观察因素中存在有很难或无法人为控制的因素时，则可对之加以指定以便进行协方差分析。

6.1.2 实例操作

[例 6-1]下表为运动员与大学生的身高（cm）与肺活量（cm^3）的数据，考虑到身高与肺活量有关，而一般运动员的身高高于大学生，为进一步分析肺活量的差异是否由于体育锻炼所致，试作控制身高变量的协方差分析。

运　动　员	大　学　生

身高	肺活量	身高	肺活量
184.9	4300	168.7	3450
167.9	3850	170.8	4100
171.0	4100	165.0	3800
171.0	4300	169.7	3300
188.0	4800	171.5	3450
179.0	4000	166.5	3250
177.0	5400	165.0	3600
179.5	4000	165.0	3200
187.0	4800	173.0	3950
187.0	4800	169.0	4000
169.0	4500	173.8	4150
188.0	4780	174.0	3450
176.7	3700	170.5	3250
179.0	5250	176.0	4100
183.0	4250	169.5	3650
180.5	4800	176.3	3950
179.0	5000	163.0	3500
178.0	3700	172.5	3900
164.0	3600	177.0	3450
174.0	4050	173.0	3850

6.1.2.1 数据准备

激活数据管理窗口，定义变量名：组变量为 group（运动员=1，大学生=2），身高为 x，肺活量为 y，按顺序输入相应数值，建立数据库，结果见图 6.1。

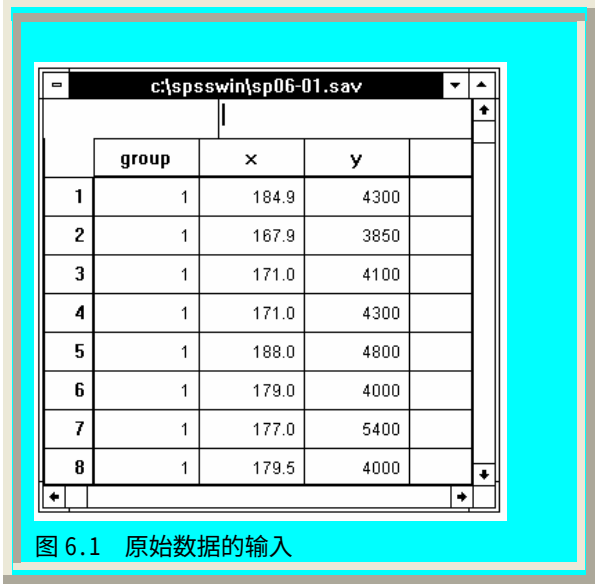


图 6.1 原始数据的输入

6.1.2.2 统计分析

激活 Statistics 菜单选 ANOVA Models 中的 Simple Factorial...项，弹出 Simple Factorial ANOVA 对话框（图 6.2）。在变量列表中选变量 y，点击 ➢ 钮使之进入 Dependent 框；选分组变量 group，点击 ➢ 钮使之进入 Factor(s)框中，并点击 Define Range...钮在弹出的 Simple Factorial ANOVA:Define Range 框中确定分组变量 group 的起止值（1,2）；选协变量 x，点击 ➢ 钮使之进入 Covariate(s)框中。

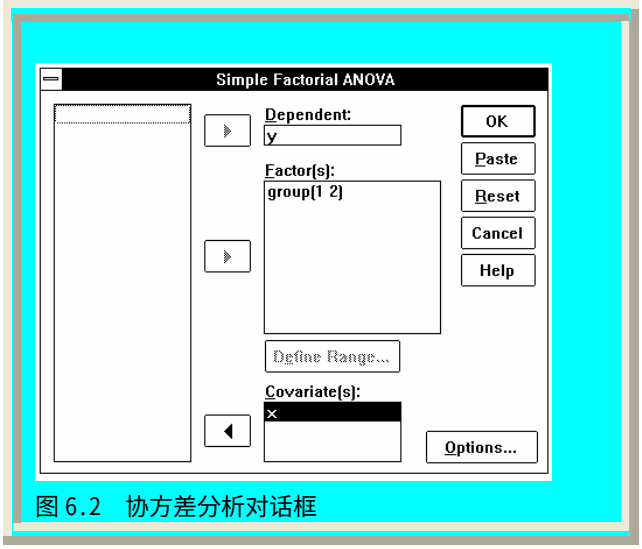


图 6.2　协方差分析对话框

点击 Options...框，弹出 Simple Factorial ANOVA:Options 对话框。系统在协方差分析的方法（Method）上有三种选项：

1、Unique：同时评价所有的效应；

2、Hierarchical：除主效应外，逐一评价各因素的效应；

3、Experimental：评价因素干预之前的主效应。

本例选 Unique 方法，之后点击 Continue 钮返回 Simple Factorial ANOVA 对话框，再点击 OK 钮即可。

6.1.2.3　结果解释

在结果输出窗口中可见如下统计数据：

先输出肺活量总均数和两组的肺活量均数，总均数为 4033.25，运用员组均数为 4399.00，大学生组为 3667.50。

接着协方差分析表明，混杂因素 X（身高）两组间是有差异的（F=10.679，P=0.002），控制其影响后，两组间肺活量的差别依然存在（F=9.220，P=0.004），故可以认为两组间肺活量的均数在消除了身高因素的影响之后仍有差别，运动员的肺活量大于大学生，即体育锻炼会提高肺活量。

最后系统输出公共回归系数，$b_c = 36.002$，该值可用于求修正均数：

$$\overline{Y_i'} = \overline{Y_i} - b_c(\overline{X_i} - \overline{X})$$

本例为 $\overline{Y'}_{运动员} = 4399.00 - 36.002 \times （178.175 - 174.3325）= 4260.6623$

$\overline{Y'}_{大学生} = 3667.50 - 36.002 \times （170.49 - 174.3325）= 3805.8377$

```
Y    by    GROUP
Total Population
```

```
   4033.25
 (       40)

GROUP       1              2
          4399.00     3667.50
        (       20)  (       20)
Y     by     GROUP
        with X
                        UNIQUE sums of squares
                        All effects entered simultaneously
                                        Sum of                  Mean                  Sig
Source of Variation                   Squares       DF          Square        F       of F
Covariates                            1630763        1     1630762.635    10.679    .002
     X                                 1630763          1     1630762.635     10.679    .002
Main Effects                          1407847        1     1407847.095     9.220    .004
     GROUP                             1407847          1     1407847.095     9.220    .004
Explained                             6981685        2     3490842.568    22.860    .000
Residual                              5649992       37      152702.496

Total                                12631678       39      323889.167

40 cases were processed.
0 cases (.0 pct) were missing.
Covariate      Raw Regression Coefficient
X                        36.002
```

第二节 General Factorial 过程

6.2.1 主要功能

调用此过程可对完全随机设计资料、配伍设计资料、析因设计资料、正交设计资料等等进行多因素方差分析或协方差分析。

6.2.2 实例操作

[例 6-2]下表为三因素析因实验的资料，请用方差分析说明不同基础液与不同血清种类对钩端螺旋体的培养计数的影响。

基础液（A）	血清种类（B）			
	兔血清浓度（C）		胎盘血清浓度（C）	
	5%	8%	5%	8%
缓冲液	648 1246 1398 909	1144 1877 1671 1845	830 853 441 1030	578 669 643 1002
蒸馏水	1763 1241 1381 2421	1447 1883 1896 1926	920 709 848 574	933 1024 1092 742
自来水	580 1026 1026 830	1789 1215 1434 1651	1126 1176 1280 1212	685 546 595 566

6.2.2.1 数据准备

激活数据管理窗口，定义变量名：基础液为 base，血清种类为 sero，血清浓度为 pct，钩端螺旋体的培养计数为 X，按顺序输入相应数值，建立数据库。

6.2.2.2 统计分析

激活 Statistics 菜单选 ANOVA Models 中的 General Factorial...项，弹出 General Factorial ANOVA 对话框（图 6.3）。在对话框左侧的变量列表中选变量 x，点击 ➢ 钮使之进入 Dependent Variable 框；选要控制的分组变量 base、sero 和 pct，点 ➢ 钮使之进入 Factor(s)框中，并分别点击 Define Range 钮，在弹出的 General Factorial ANOVA:Define Range 对话框中确定各变量的起止值，本例变量 base 的起止值为 1、3，变量 sero 的起止值为 1、2，变量 pct 的起止值为 1、2。之后点击 OK 钮即可。

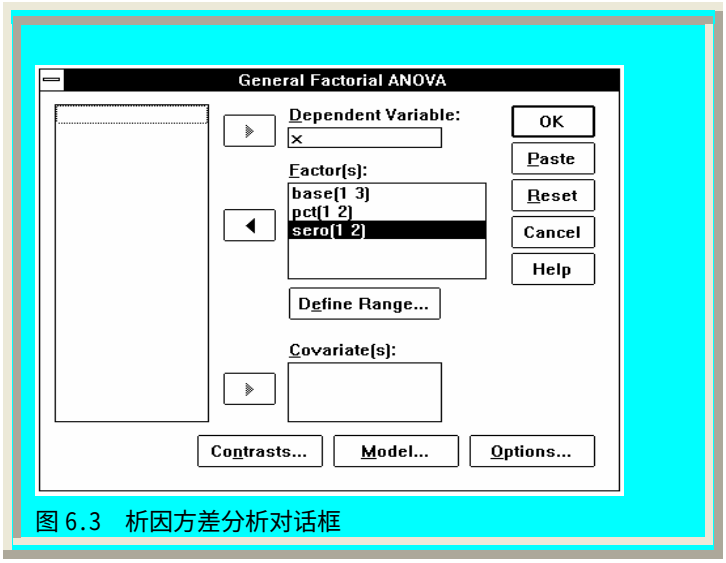


图 6.3 析因方差分析对话框

6.2.2.3 结果解释

在结果输出窗口中，系统显示 48 个观察值进入统计，三个因素按其各自水平共产生 12 种组合。

分析表明，模型总效应的 F 值为 10.55，P 值 < 0.001，说明三因素间存在有交互作用。单因素效应和交互效应导致的组间差别比较结果是：

单因素组间比较：

A：基础液（BASE）

F = 4.98，P = 0.012，说明三种培养基培养钩体的计数有差别；

B：血清种类（SERO）

F = 61.265，P < 0.001，说明两种血清培养钩体的计数有差别；

C：血清浓度（PCT）

F = 3.49，P = 0.070，说明两种血清浓度培养钩体的计数无差别。

两因素构成的一级交互作用：

A×B：基础液（BASE）×血清种类（SERO）

F = 5.16，P = 0.011，交互作用明显；

B×C：血清种类（SERO）×血清浓度（PCT）

F = 15.96，P < 0.001，交互作用明显；

A×C：基础液（BASE）×血清浓度（PCT）

F = 0.78，P = 0.465，交互作用不明显。

三因素构成的二级交互作用：

A×B×C：基础液（BASE）×血清种类（SERO）×血清浓度（PCT）

F = 6.75，P = 0.003，交互作用明显。

```
48 cases accepted.
 0 cases rejected because of out-of-range factor values.
 0 cases rejected because of missing data.
12 non-empty cells.
 1 design will be processed.
 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
 Univariate Homogeneity of Variance Tests
 Variable .. X
        Cochrans C(3,12) =          .34004,       P =    .036 (approx.)
        Bartlett-Box F(11,897) =    1.69822,      P =    .069
 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
* * * * * * A n a l y s i s   o f   V a r i a n c e -- design   1 * * * * * *
 Tests of Significance for X using UNIQUE sums of squares
 Source of Variation          SS              DF          MS            F        Sig of F
 WITHIN+RESIDUAL        2459233.75          36      68312.05
 BASE                    679967.38           2     339983.69         4.98         .012
 PCT                     238713.02           1     238713.02         3.49         .070
 SERO                   4184873.52           1    4184873.5         61.26         .000
 BASE BY PCT             107005.54           2      53502.77          .78         .465
 BASE BY SERO            705473.04           2     352736.52         5.16         .011
 PCT BY SERO            1089922.69           1    1089922.7         15.96         .000
 BASE BY PCT BY SERO     922307.37           2     461153.69         6.75         .003
```

(Model)	7928262.56	11	720751.14	10.55	.000
(Total)	10387496.31	47	221010.56		

R-Squared = .763
Adjusted R-Squared = .691

第三节 Multivarite 过程

6.3.1 主要功能

调用此过程可进行多元方差分析。此外，对于一元设计，如涉及混合模型的设计、分割设计（又称列区设计）、重复测量设计、嵌套设计、因子与协变量交互效应设计等，此过程均能适用。

6.3.2 实例操作

[例 6-3]甲地区为大城市，乙地区为县城，丙地区为农村。某地分别调查了上述三类地区 8 岁男生三项身体生长发育指标：身高、体重和胸围，数据见下表，问：三类地区之间男生三项身体生长发育指标的差异有无显著性？

学生编号	甲地区			乙地区			丙地区		
	身高	体重	胸围	身高	体重	胸围	身高	体重	胸围
1	119.80	22.60	60.50	125.10	23.00	62.00	118.30	20.40	54.40
2	121.70	21.50	55.50	127.00	21.50	59.00	121.30	20.00	54.30
3	121.40	19.10	56.50	125.70	23.40	61.50	121.80	26.60	61.10
4	124.40	21.80	60.50	114.90	17.50	52.50	124.20	22.10	58.60
5	120.00	21.40	57.70	124.90	23.50	58.50	123.50	23.20	60.20
6	117.00	20.10	57.00	117.60	18.90	57.00	123.00	22.90	58.20
7	118.10	18.80	57.10	124.20	20.80	58.50	134.90	32.30	64.80
8	118.80	22.00	61.70	117.90	20.30	61.00	123.70	22.70	59.90
9	124.20	21.30	58.40	120.40	20.00	56.00	105.20	20.20	54.50
10	124.90	24.00	60.80	115.00	19.70	56.50	112.20	20.80	57.50
11	124.70	23.30	60.00	126.20	21.20	56.50	118.60	21.00	57.60
12	123.00	22.50	60.00	125.10	22.10	58.50	112.00	23.20	58.20
13	125.30	22.90	65.20	114.90	19.70	56.00	121.50	24.00	60.30
14	124.20	19.50	53.80	121.50	22.00	57.00	124.50	21.50	55.60
15	127.40	22.90	59.50	114.00	19.00	54.50	119.50	20.50	55.50
16	128.20	22.30	60.00	118.70	19.10	54.50	122.50	23.00	56.70
17	126.10	22.70	57.40	120.60	20.00	55.50	115.50	19.00	54.20
18	128.70	23.50	60.40	122.90	18.50	56.00	122.50	22.50	57.60

19	129.50	24.50	51.00	119.60	19.50	59.50	124.50	25.00	57.90
20	126.90	25.50	61.50	112.30	20.00	58.00	125.00	25.50	60.30
21	126.50	25.00	63.90	121.30	20.00	58.00	117.50	23.00	59.00
22	128.20	26.10	63.00	121.20	21.20	59.00	127.30	22.50	58.90
23	131.40	27.90	63.10	120.20	23.10	59.50	122.30	22.00	58.20
24	130.80	26.80	61.50	120.30	21.00	59.50	121.30	21.00	55.60
25	133.90	27.20	65.80	120.00	22.20	59.50	120.50	22.00	55.10
26	130.40	24.40	62.60	123.30	20.10	56.50	116.00	19.00	53.50
27	131.30	24.40	59.50	122.10	21.00	57.50	120.50	20.00	54.40
28	130.20	23.00	62.60	123.30	21.50	61.00	114.50	19.00	53.40
29	136.00	26.30	60.00	109.90	17.80	56.50	131.00	25.50	58.30
30	141.00	31.90	63.70	125.60	23.30	60.50	122.50	24.50	58.70

6.3.2.1 数据准备

激活数据管理窗口，定义变量名：地区为 G，身高为 X1，体重为 X2，胸围为 X3，按顺序输入相应数值，变量 G 的数值是：甲地区为 1，乙地区为 2，丙地区为 3。

6.3.2.2 统计分析

激活 Statistics 菜单选 ANOVA Models 中的 Multivarite...项，弹出 Multivarite ANOVA 对话框（图 6.8）。首先指定供分析用的变量 x1、x2、x3，故在对话框左侧的变量列表中选变量 x1、x2、x3，点击 ➢ 钮使之进入 Dependent Variable 框；然后选变量 g（分组变量）点击 ➢ 钮使之进入 Factor(s)框中，并点击 Define Range 钮，确定 g 的起始值和终止值。

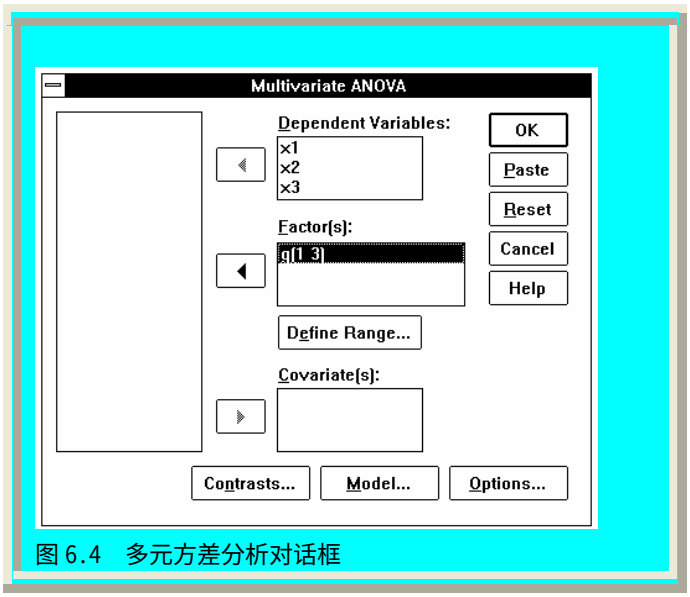


图 6.4 多元方差分析对话框

点击 Options...钮，弹出 Multivarite ANOVA:Options 对话框，选择需要计算的指标。在 Factor(s)栏内选变量 g，点击 ➢ 钮使之进入 Display Means for 框，要求计算平均值指标；在 Matriced Within Cell 栏内选 Correlation、Covariance、SSCP 项，要求计算单元内的相关矩阵、方差协方差矩阵和离均差平方和交叉乘积矩阵；在 Error Matrices 栏内也选上述三项，要求计算误差的相关矩阵、方差协方差矩阵和离均差平方和交叉乘积矩阵；在 Diagnostics 栏内选 Homogeneity test 项，要求作变量的方差齐性检验。之后点击 Continue 钮返回 Multivarite ANOVA 对话框，最后点击 OK 钮即可。

6.3.2.3 结果解释

在结果输出窗口中将看到如下分析结果：

系统首先显示共 90 个观察值进入统计分析，因分组变量 g 为三个地区，故分析的单元数为 3。然后输出 3 个应变量（x1、x2、x3）的方差齐性检验结果，分别输出了 Cochran C 检验值及其显著性水平 P 值、Bartlett-Box F 检验值及其显著性水平 P 值。其中

身高：C = 0.39825，P = 0.540；F = 1.01272，P = 0.363；

体重：C = 0.43787，P = 0.227；F = 4.48624, P = 0.011；

胸围：C = 0.47239, P = 0.089；F = 2.06585, P = 0.127；

可见 3 项指标的方差基本整齐（P 值均大于 0.05）。

```
90 cases accepted.
          0 cases rejected because of out-of-range factor values.
          0 cases rejected because of missing data.
          3 non-empty cells.

          1 design will be processed.

                    CELL NUMBER
                    1       2       3
 Variable
   G                1       2       3

 Univariate Homogeneity of Variance Tests
 Variable .. X1
       Cochrans C(29,3) =    .39825,          P =    .540 (approx.)
       Bartlett-Box F(2,17030) = 1.01272,   P =    .363
 Variable .. X2
       Cochrans C(29,3) =    .43787,          P =    .227 (approx.)
       Bartlett-Box F(2,17030) = 4.48624,   P =    .011
 Variable .. X3
       Cochrans C(29,3) =    .47239,          P =    .089 (approx.)
       Bartlett-Box F(2,17030) = 2.06585,   P =    .127
```

Cochran C检验和Bartlett-Box F检验对考查协方差矩阵的相等性比较方便，但还不够。于是系统接着分别输出了三类地区（即各个单元）各生长发育指标的离均差平方和交叉乘积矩阵和方差协方差矩阵。之后作Box M检验，Box M检验提供矩阵一致性的多元测试，本例Boxs M = 36.93910，在基于方差分析的显著性检验中F = 2.92393；在基于χ^2的显著性检验中χ^2 = 35.09922, 两者P < 0.001，故认为矩阵一致性不佳。

```
Cell Number .. 1
 Sum of Squares and Cross-Products matrix
```

	X1	X2	X3
X1	861.187		
X2	380.137	230.519	
X3	215.937	156.559	314.859

Variance-Covariance matrix

	X1	X2	X3
X1	29.696		
X2	13.108	7.949	
X3	7.446	5.399	10.857

Cell Number .. 1 (Cont.)

Correlation matrix with Standard Deviations on Diagonal

	X1	X2	X3
X1	5.449		
X2	.853	2.819	
X3	.415	.581	3.295

Determinant of Covariance matrix of dependent variables = 444.98354

LOG(Determinant) = 6.09804

Cell Number .. 2

Sum of Squares and Cross-Products matrix

	X1	X2	X3
X1	565.368		
X2	147.222	78.910	
X3	139.430	79.337	147.967

Variance-Covariance matrix

	X1	X2	X3
X1	19.495		
X2	5.077	2.721	
X3	4.808	2.736	5.102

Correlation matrix with Standard Deviations on Diagonal

	X1	X2	X3
X1	4.415		
X2	.697	1.650	
X3	.482	.734	2.259

Determinant of Covariance matrix of dependent variables = 63.90640

LOG(Determinant) = 4.15742

Cell Number .. 3

Sum of Squares and Cross-Products matrix

	X1	X2	X3
X1	944.128		
X2	307.722	217.030	
X3	261.130	186.252	203.702

Variance-Covariance matrix

	X1	X2	X3
X1	32.556		
X2	10.611	7.484	
X3	9.004	6.422	7.024

Correlation matrix with Standard Deviations on Diagonal

	X1	X2	X3
X1	5.706		
X2	.680	2.736	
X3	.595	.886	2.650

Determinant of Covariance matrix of dependent variables = 198.13507
LOG(Determinant) = 5.28895

Pooled within-cells Variance-Covariance matrix

	X1	X2	X3
X1	27.249		
X2	9.599	6.051	
X3	7.086	4.852	7.661

Determinant of pooled Covariance matrix of dependent vars. = 272.06906
LOG(Determinant) = 5.60606

Multivariate test for Homogeneity of Dispersion matrices

Boxs M = 36.93910
F WITH (12,36680) DF = 2.92393, P = .000 (Approx.)
Chi-Square with 12 DF = 35.09922, P = .000 (Approx.)

下面系统输出将三类地区看成一个大样本时的离均差平方和交叉乘积矩阵。如 X1、X2 和 X3 的离均差平方和分别为 662.884、121.562 和 114.902。在此基础上，进行多元差异的检验。通常有四种方法：

1、Pillai轨迹：$V = \sum_{i=1}^{s} \frac{1}{1+\lambda_i}$

2、Wilks λ值：$W=\prod_{i=1}^{s}\frac{1}{1+\lambda_i}$

3、Hotelling轨迹：$T=\sum_{i=1}^{s}\frac{1}{1+\lambda_i}$

4、Roy最大根：$R=\sum_{i=1}^{s}\frac{\lambda_{max}}{1+\lambda_{max}}$

式中 λmax 为最大特征值, λi 为第 i 个特征值,s 为非零特征值个数。根据这些值变换的 F 检验均有显著性（P<0.001），说明三类地区各生长发育指标之间的差别有高度显著性。

这一计算结果对上述三项生长发育指标进行了单因素的方差分析，可见：

X1: SS = 662.88356, F = 12.16335

X2: SS = 121.56200, F = 10.04439

X3: SS = 114.90200, F = 7.49893

差别均有显著性，说明三项生长发育指标各地区间的差别均有显著性。

```
Combined Observed Means for G
 Variable .. X1
                G
                1          WGT.     126.46667
                          UNWGT.     126.46667
                2          WGT.     120.52000
                          UNWGT.     120.52000
                3          WGT.     120.92000
                          UNWGT.     120.92000
 ------------------------------------
 Variable .. X2
                G
                1          WGT.      23.50667
                          UNWGT.      23.50667
                2          WGT.      20.69667
                          UNWGT.      20.69667
                3          WGT.      22.49667
                          UNWGT.      22.49667
 ------------------------------------
 Variable .. X3
                G
                1          WGT.      60.00667
                          UNWGT.      60.00667
                2          WGT.      57.86667
                          UNWGT.      57.86667
                3          WGT.      57.41667
                          UNWGT.      57.41667
```

- -

WITHIN+RESIDUAL Correlations with Std. Devs. on Diagonal

	X1	X2	X3
X1	5.220		
X2	.747	2.460	
X3	.490	.713	2.768

- -

Statistics for WITHIN+RESIDUAL correlations

Log(Determinant) = .00000

Bartlett test of sphericity = . with 3 D. F.

Significance = .

F(max) criterion = 4.50308 with (3,87) D. F.

WITHIN+RESIDUAL Variances and Covariances

	X1	X2	X3
X1	27.249		
X2	9.599	6.051	
X3	7.086	4.852	7.661

- -

WITHIN+RESIDUAL Sum-of-Squares and Cross-Products

	X1	X2	X3
X1	2370.683		
X2	835.081	526.458	
X3	616.497	422.147	666.527

- -

EFFECT .. G

Adjusted Hypothesis Sum-of-Squares and Cross-Products

	X1	X2	X3
X1	662.884		
X2	230.323	121.562	
X3	269.117	78.193	114.902

- -

Multivariate Tests of Significance (S = 2, M = 0, N = 41 1/2)

Test Name	Value	Approx.F	Hypoth. DF	Error DF	Sig. of F
Pillais	.51227	9.87080	6.00	172.00	.000
Hotellings	.70427	9.85978	6.00	168.00	.000
Wilks	.55014	9.86643	6.00	170.00	.000
Roys	.31265				

Note.. F statistic for WILKS' Lambda is exact.

- -

EFFECT .. G (Cont.)

Univariate F-tests with (2,87) D. F.

Variable	Hypoth. SS	Error SS	Hypoth. MS	Error MS	F	Sig. of F
X1	662.88356	2370.68267	331.44178	27.24923	12.16335	.000

X2	121.56200	526.45800	60.78100	6.05124	10.04439	.000
X3	114.90200	666.52700	57.45100	7.66123	7.49893	.001

之后按单元输出各项指标的观察值均数（Obs.Mean）、调整均数（Adj.Mean）、估计均数（Est.Mean）、粗误差（Raw Resid）、标准化误差（Std.Resid）以及不分地区的总均数（Comined Adjusted Means for G）。

Adjusted and Estimated Means

Variable .. X1

CELL	Obs. Mean	Adj. Mean	Est. Mean	Raw Resid.	Std. Resid.
1	126.467	126.467	126.467	.000	.000
2	120.520	120.520	120.520	.000	.000
3	120.920	120.920	120.920	.000	.000

- -

Adjusted and Estimated Means (Cont.)

Variable .. X2

CELL	Obs. Mean	Adj. Mean	Est. Mean	Raw Resid.	Std. Resid.
1	23.507	23.507	23.507	.000	.000
2	20.697	20.697	20.697	.000	.000
3	22.497	22.497	22.497	.000	.000

- -

Adjusted and Estimated Means (Cont.)

Variable .. X3

CELL	Obs. Mean	Adj. Mean	Est. Mean	Raw Resid.	Std. Resid.
1	60.007	60.007	60.007	.000	.000
2	57.867	57.867	57.867	.000	.000
3	57.417	57.417	57.417	.000	.000

- -

Combined Adjusted Means for G

Variable .. X1

G		
1	UNWGT.	126.46667
2	UNWGT.	120.52000
3	UNWGT.	120.92000

- -

Variable .. X2

G		
1	UNWGT.	23.50667
2	UNWGT.	20.69667
3	UNWGT.	22.49667

- -

Variable .. X3

```
G
1        UNWGT.      60.00667
2        UNWGT.      57.86667
3        UNWGT.      57.41667
```

最后，系统输出各变量的离差参数。用户可据此计算预测值，

预测值 Y＝ 总均数 ＋ 该变量离差参数 ＋ 变量间交互效应的离差参数

如本例因无变量间交互效应的离差参数，故甲地区 8 岁男生的身高预测值为 Y ＝ 126.46667 ＋（-1.71555551）= 124.7511145。

上式中 126.46667 可从系统输出的 Combined Adjusted Means for G 一栏中得到，离差参数 -1.71555551 = 0 - 3.83111111 - （-2.1155556），这是因为离差参数的合计总为 0 的缘故。余同，在此不作赘述。

```
Estimates for X1
 --- Individual univariate .9500 confidence intervals
 G
  Parameter       Coeff.     Std. Err.     t-Value      Sig. t    Lower -95%   CL- Upper
          2   3.83111111       .77816      4.92327      .00000      2.28443      5.37780
          3   -2.1155556       .77816     -2.71865      .00791     -3.66224      -.56887
 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
 Estimates for X2
 --- Individual univariate .9500 confidence intervals
 G
  Parameter       Coeff.     Std. Err.     t-Value      Sig. t    Lower -95%   CL- Upper
          2   1.27333333       .36670      3.47237      .00081       .54447      2.00220
          3   -1.5366667       .36670     -4.19048      .00007     -2.26553      -.80780
 - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -
 Estimates for X3
 --- Individual univariate .9500 confidence intervals
 G
  Parameter       Coeff.     Std. Err.     t-Value      Sig. t    Lower -95%   CL- Upper
          2   1.57666667       .41261      3.82117      .00025       .75655      2.39678
          3   -.56333333       .41261     -1.36528      .17568     -1.38345       .25678
```

第七章　相关分析

任何事物的存在都不是孤立的，而是相互联系、相互制约的。在医学领域中，身高与体重、体温与脉搏、年龄与血压等都存在一定的联系。说明客观事物相互间关系的密切程度并用适当的统计指标表示出来，这个过程就是相关分析。

值得注意，事物之间有相关，不一定是因果关系，也可能仅是伴随关系。但如果事物之间有因果关系，则两者必然相关。

SPSS 的相关分析是借助于 Statistics 菜单的 Correlate 选项完成的。

第一节 Bivariate 过程

7.1.1 主要功能

调用此过程可对变量进行相关关系的分析，计算有关的统计指标，以判断变量之间相互关系的密切程度。调用该过程命令时允许同时输入两变量或两个以上变量，但系统输出的是变量间两两相关的相关系数。

7.1.2 实例操作

[例 7-1]某地区 10 名健康儿童头发和全血中的硒含量（1000ppm）如下，试作发硒与血硒的相关分析。

编号	发硒	血硒
1	74	13
2	66	10
3	88	13
4	69	11
5	91	16
6	73	9
7	66	7
8	96	14
9	58	5
10	73	10

7.1.2.1 数据准备

激活数据管理窗口，定义变量名：发硒为 X，血硒为 Y，按顺序输入相应数值，建立数据库（图 7.1）。

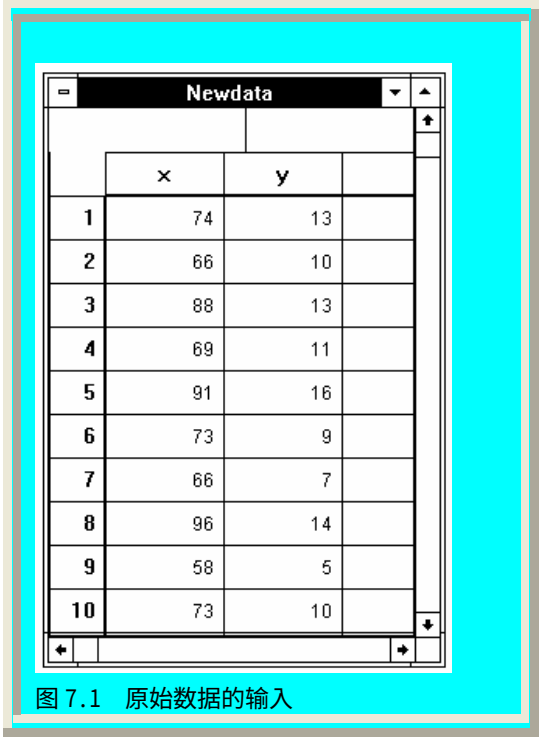


图 7.1　原始数据的输入

7.1.2.2　统计分析

激活 Statistics 菜单选 Correlate 中的 Bivariate...命令项，弹出 Bivariate Correlation 对话框（图 7.2）。在对话框左侧的变量列表中选 x、y，点击 ➢ 钮使之进入 Variables 框；再在 Correlation Coefficients 框中选择相关系数的类型，共有三种：Pearson 为通常所指的相关系数（r），Kendell's tau-b 为非参数资料的相关系数，Spearman 为非正态分布资料的 Pearson 相关系数替代值，本例选用 Pearson 项；在 Test of Significance 框中可选相关系数的单侧（One-tailed）或双侧（Two-tailed）检验，本例选双侧检验。

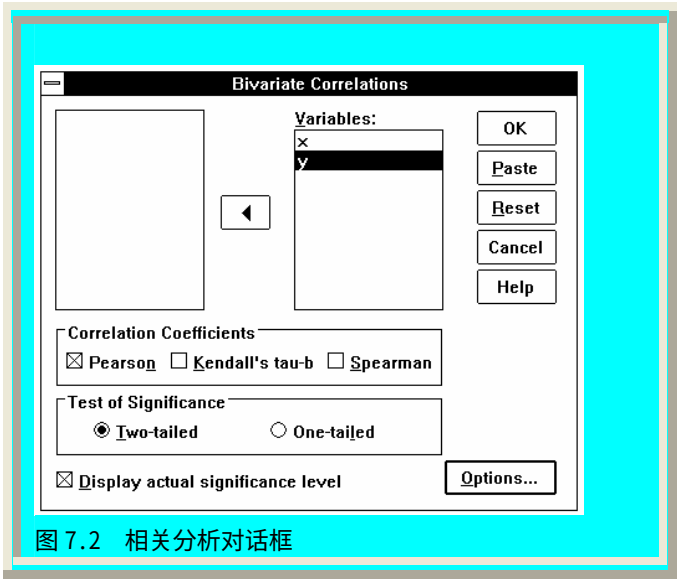


图 7.2　相关分析对话框

点击 Options...钮弹出 Bivariate Correlation:Options 对话框（图 7.3），可选有关统计项目。本例要求输出 X、Y 的均数与标准差以及 XY 交叉乘积的标准差与协方差，故选 Means and standard deviations

和 Cross-product deviations and covariances 项，而后点击 Continue 钮返回 Bivariate Correlation 对话框，再点击 OK 钮即可。

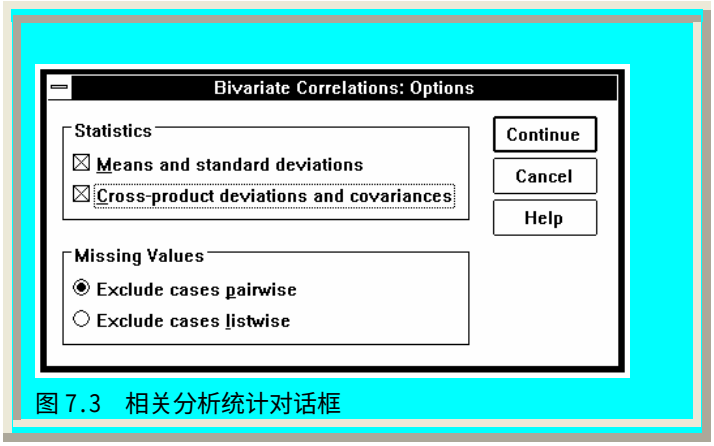


图 7.3　相关分析统计对话框

7.1.2.3　结果解释：

在结果输出窗口中将看到如下统计数据：变量 X、Y 的例数、均数与标准差，变量 X、Y 交叉乘积的例数、标准差与协方差；XY 两两对应的相关系数及其双侧检验的概率，本例 r = 0.8715，P = 0.001。

Variable	Cases	Mean	Std Dev
X	10	75.4000	12.2945
Y	10	10.8000	3.3267

Variables		Cases	Cross-Prod Dev	Variance-Covar
X	Y	10	320.8000	35.6444

	X	Y
X	1.0000	.8715
	(10)	(10)
	P= .	P= .001
Y	.8715	1.0000
	(10)	(10)
	P= .001	P= .

(Coefficient / (Cases) / 2-tailed Significance)
" . " is printed if a coefficient cannot be computed

第二节 Partial 过程

7.2.1 主要功能

调用此过程可对变量进行偏相关分析。在偏相关分析中，系统可按用户的要求对两相关变量之外的某一或某些影响相关的其他变量进行控制，输出控制其他变量影响后的相关系数。

7.2.2 实例操作

[例 7-2]某地 29 名 13 岁男童身高（cm）、体重（kg）和肺活量（ml）的数据如下表, 试对该资料作控制体重影响作用的身高与肺活量相关分析。

编号	身高（cm）	体重(kg)	肺活量(ml)	编号	身高（cm）	体重(kg)	肺活量(ml)
1	135.1	32.0	1750	16	153.0	47.2	1750
2	139.9	30.4	2000	17	147.6	40.5	2000
3	163.6	46.2	2750	18	157.5	43.3	2250
4	146.5	33.5	2500	19	155.1	44.7	2750
5	156.2	37.1	2750	20	160.5	37.5	2000
6	156.4	35.5	2000	21	143.0	31.5	1750
7	167.8	41.5	2750	22	149.4	33.9	2250
8	149.7	31.0	1500	23	160.8	40.4	2750
9	145.0	33.0	2500	24	159.0	38.5	2500
10	148.5	37.2	2250	25	158.2	37.5	2000
11	165.5	49.5	3000	26	150.0	36.0	1750
12	135.0	27.6	1250	27	144.5	34.7	2250
13	153.3	41.0	2750	28	154.6	39.5	2500
14	152.0	32.0	1750	29	156.5	32.0	1750
15	160.5	47.2	2250				

7.2.2.1 数据准备

激活数据管理窗口，定义变量名：身高为 height，体重为 weight，肺活量为 vc，按顺序输入相应数值，建立数据库。

7.2.2.2 统计分析

激活 Statistics 菜单选 Correlate 中的 Partial...命令项，弹出 Partial Correlations 对话框（图 7.4）。现欲在控制体重的影响下对变量身高与肺活量进行偏相关分析，故在对话框左侧的变量列表中选变量 height、vc，点击 ➢ 钮使之进入 Variables 框，选要控制的变量 weight，点击 ➢ 钮使之进入 Controlling for 框中，在 Test of Significance 框中选双侧检验，然后点击 OK 钮即可。

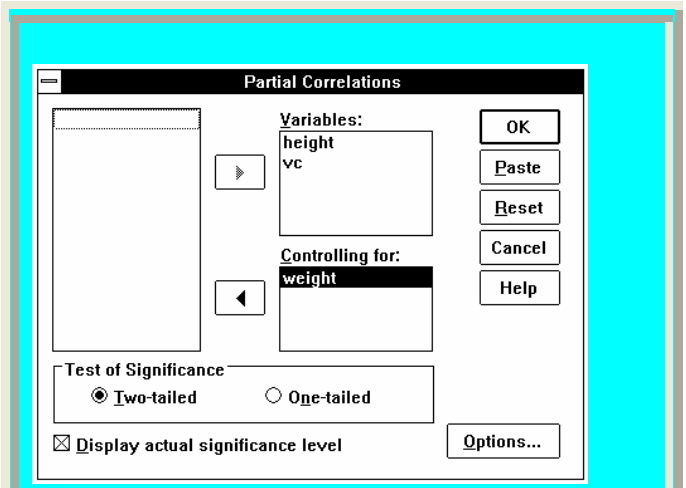


图 7.4　偏相关分析对话框

7.2.2.3　结果解释

在结果输出窗口中将看到如下统计数据：控制体重的影响后，身高与肺活量的相关系数为0.0926，经检验P = 0.639，故身高与肺活量的线性相关不存在。（如果不控制体重的影响，则身高与肺活量的相关系数为0.5884，P为0.001。在有控制的情况下，身高与肺活量的决定系数 $= r^2 = 0.00857$，而无控制的身高与肺活量决定系数 $= r^2 = 0.34621$，可见身高与肺活量的相关有33.764%是由体重协同作用而产生的。）

```
Controlling for..  WEIGHT
                  HEIGHT          VC
HEIGHT             1.0000         .0926
                  (      0)      (     26)
                  P= .           P= .639
VC                  .0926         1.0000
                  (     26)      (      0)
                  P= .639        P= .

(Coefficient / (D.F.) / 2-tailed Significance)

" . " is printed if a coefficient cannot be computed
```

如果控制变量改为身高，则得如下结果：体重与肺活量的相关系数为0.5528，经检验 P = 0.002，故体重与肺活量的线性相关存在。可见，尽管肺活量与身高和体重均有关系，但如果仅仅研究其中一个变量与肺活量的相关关系时，体重的意义会更大。

```
Controlling for..  HEIGHT
                     VC        WEIGHT
VC                   1.0000     .5528
                    (      0)  (     26)
```

```
                    P= .          P= .002
WEIGHT                .5528         1.0000
                    (    26)      (       0)
                    P= .002       P= .

(Coefficient / (D.F.) / 2-tailed Significance)

" . " is printed if a coefficient cannot be computed
```

第三节 Distances 过程

7.3.1 主要功能

调用此过程可对变量内部各观察单位间的数值进行距离相关分析，以考察相互间的接近程度；也可对变量间进行距离相关分析，常用于考察预测值对实际值的拟合程度。

7.3.2 实例操作

[例 7-3]某医师对 10 份标准血红蛋白样品作三次平行检测，结果如下，问检测结果是否一致？

样品号	1	2	3	4	5	6	7	8	9	10
第一次	12.36	12.14	12.31	12.32	12.12	12.28	12.24	12.41	12.33	12.17
第二次	12.40	12.20	12.28	12.25	12.22	12.34	12.31	12.30	12.22	12.24
第三次	12.18	12.22	12.35	12.21	12.10	12.25	12.20	12.46	12.36	12.11

7.3.2.1 数据准备

激活数据管理窗口，定义变量名：第一次测量值为 HB1，第二次测量值为 HB2，第三次测量值为 HB3，输入相应数值即完成。

7.3.2.2 统计分析

激活 Statistics 菜单选 Correlate 中的 Distance...命令项，弹出 Distance 对话框（图 7.5）。在对话框左侧的变量列表中选变量 hb1、hb2、hb3，点击 ➢ 钮使之进入 Variables 框。在 Compute Distances 框中有两个选项，Between cases 表示作变量内部观察值之间的距离相关分析，Between variables 表示作变量之间的距离相关分析，在本例中，因三次平行测量结果分别置于三个变量中，故选择后者。

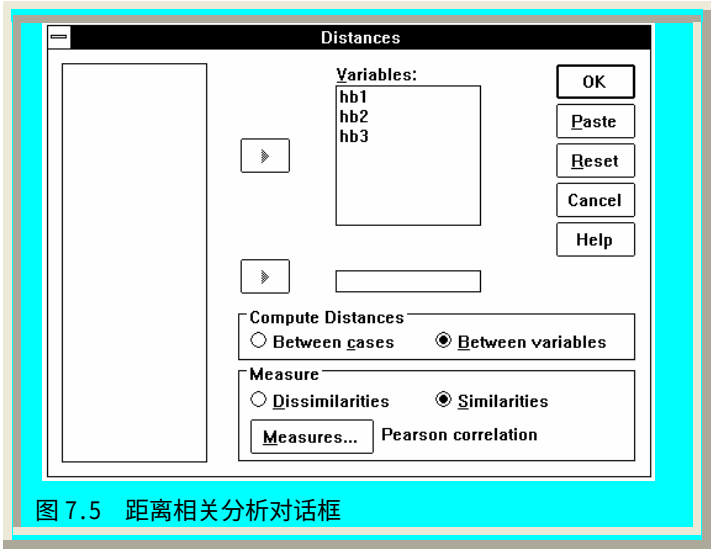


图 7.5　距离相关分析对话框

在 Measure 栏中有两种测距方式：Dissimilarities 为不相似性测距，Similarities 为相似性测距。若选 Dissimilarties 并点击 Measure...钮，弹出 Distance:Dissimilarity Measure 对话框（图 7.6），用户可根据数据特征选用测距方法：

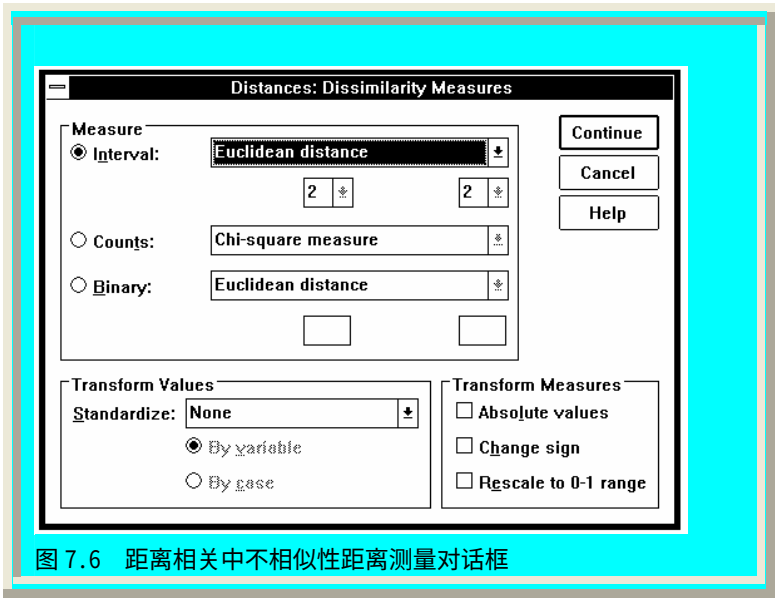


图 7.6　距离相关中不相似性距离测量对话框

1、计量资料

Euclidean distance：以两变量差值平方和的平方根为距离；

Squared Euclidean distance：以两变量差值平方和为距离；

Chebychev：以两变量绝对差值的最大值为距离；

Block：以两变量绝对差值之和为距离；

Minkowski：以两变量绝对差值 p 次幂之和的 p 次根为距离；

Customized：以两变量绝对差值 p 次幂之和的 r 次根为距离。

2、计数资料

Chi-square measure：χ^2值测距；

Phi-square measure：ψ^2值测距，即将χ^2测距值除合计频数的平方根。

3、二分字符变量

Euclidean distance：二分差平方和的平方根，最小为 0，最大无限；

Squared Euclidean distance：二分差平方和，最小为 0，最大无限；

Size difference：最小距离为 0，最大无限；

Pattern difference：从 0 至 1 的无级测距；

Variance：以方差为距，最小为 0，最大无限；

Lance and Williams：Bray-Curtis 非等距系数，界于 0 至 1 之间。

若选 Similarties 并点击 Measure...钮，弹出 Distance: Similarity Measure 对话框（图 7.7），用户可根据数据特征选用测距方法：

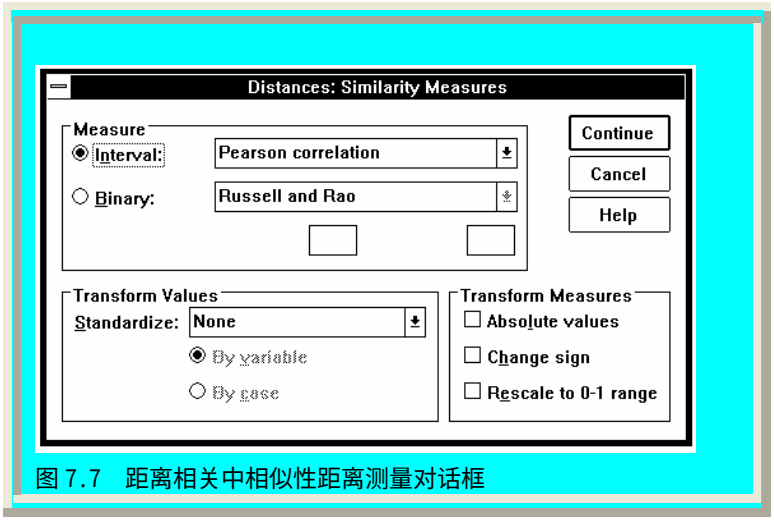


图 7.7　距离相关中相似性距离测量对话框

1、计量资料

Pearson correlation：以 Pearson 相关系数为距离；

Cosine：以变量矢量的余弦值为距离，界于-1 至+1 之间。

2、二分字符变量

Russell and Rao：以二分点乘积为配对系数；

Simple matching：以配对数与总对数的比例为配对系数；

Jaccard：相似比例，分子与分母中的配对数与非配对数给予相同的权重；

Dice：Dice 配对系数，分子与分母中的配对数给予加倍的权重；

Rogers and Tanimoto：Rogers and Tanimoto 配对系数，分母为配对数，分子为非配对数，非配对数给予加倍的权重；

Sokal and Sneath 1：Sokal and Sneath Ⅰ型配对系数，分母为配对数，分子为非配对数，配对数给予加倍的权重；

Sokal and Sneath 2：Sokal and Sneath Ⅱ型配对系数，分子与分母均为非配对数，但分子给予加倍的权重；

Sokal and Sneath 3：Sokal and Sneath Ⅲ型配对系数，分母为配对数，分子为非配对数，分子与分母的权重相同；

Kulczynski 1：Kulczynski Ⅰ型配对系数，分母为总数与配对数之差，分子为非配对数，分子与分母的权重相同；

Kulczynski 2：Kulczynski 平均条件概率；

Sokal and Sneath 4：Sokal and Sneath 条件概率；

Hamann：Hamann 概率；

Lambda：Goodman-Kruskai 相似测量的 λ 值；

Anderberg's D：以一个变量状态预测另一个变量状态；
Yule's Y：Yule 综合系数，属于 2×2 四格表的列联比例函数；
Yule's Q：Goodman-Kruskal γ 值，属于 2×2 四格表的列联比例函数。
3、其他型变量
Ochiai：Ochiai 二分余弦测量；
Sokal and Sneath 5：Sokal and Sneath Ⅴ型相似测量；
Phi 4 point correlation：Pearson 相关系数的平方值；
Dispersion：Dispersion 相似测量。
同时，还可以选择数据转换形式：
None：不作数据转换；
Z-Scores：作标准 Z 分值转换；
Range -1 to 1：作-1 至+1 之间的标准化转换；
Range 0 to 1：作 0 至 1 之间的标准化转换；
Maximum magnitude of 1：作最大量值 1 的标准转换；
Mean of 1：作均数单位转换；
Standard deviation of 1：作标准差单位转换。
本例选 Similarties 项，并以 Pearson correlation 为测量距离。点击 Continue 钮返回 Distance 对话框，再点击 OK 钮即可。

7.3.2.3 结果解释

在结果输出窗口可看到三次测量结果的相关系数矩阵。第一次测量与第二次测量结果的 r = 0.5734，第一次测量与第三次测量结果的 r = 0.7309，第二次测量与第三次测量结果的 r = 0.0878，由此可见，后两次测量的结果一致性较差，这意味着第一次恰好是后两次的“均值”，故对该指标作重复测量意义不大。

Data Information

10 unweighted cases accepted.

0 cases rejected because of missing value.

Correlation measure used.

Correlation Similarity Coefficient Matrix

Variable	HB1	HB2
HB2	.5734	
HB3	.7309	.0878

如果对变量内部各观察值间的一致性进行考核（假定本例 HB1 变量中的数据为对一个标准试样的十次平行测定），那么需在 Distance 对话框中选 Between cases 项，并选 Dissimilarities 项的 Euclidean distance 测距方法，运算结果如下：
在不相似性测量系数矩阵中，最大值为第五个观察值与第八个观察值间的仅为 0.2900，其余的值均较之更小，最小的为第三个观察值与第四个观察值间的仅为 0.0100，可见观察值间的不相似性

差（不相似性系数愈接近 1，不相似性愈好；不相似性系数愈接近 0，不相似性愈差），则意味着测定结果的一致性好。

Data Information

10 unweighted cases accepted.

0 cases rejected because of missing value.

Euclidean measure used.

Euclidean Dissimilarity Coefficient Matrix

	Case 1	Case 2	Case 3	Case 4	Case 5	Case 6	Case 7	Case 8	Case 9
Case 2	.2200								
Case 3	.0500	.1700							
Case 4	.0400	.1800	.0100						
Case 5	.2400	.0200	.1900	.2000					
Case 6	.0800	.1400	.0300	.0400	.1600				
Case 7	.1200	.1000	.0700	.0800	.1200	.0400			
Case 8	.0500	.2700	.1000	.0900	.2900	.1300	.1700		
Case 9	.0300	.1900	.0200	.0100	.2100	.0500	.0900	.0800	
Case10	.1900	.0300	.1400	.1500	.0500	.1100	.0700	.2400	.1600

第八章 回归分析

回归分析是处理两个及两个以上变量间线性依存关系的统计方法。在医学领域中，此类问题很普遍，如人头发中某种金属元素的含量与血液中该元素的含量有关系，人的体表面积与身高、体重有关系；等等。回归分析就是用于说明这种依存变化的数学关系。

第一节 Linear 过程

8.1.1 主要功能

调用此过程可完成二元或多元的线性回归分析。在多元线性回归分析中，用户还可根据需要，选用不同筛选自变量的方法（如：逐步法、向前法、向后法，等）。

8.1.2 实例操作

［例 8.1］某医师测得 10 名 3 岁儿童的身高（cm）、体重（kg）和体表面积（cm^2）资料如下。试用多元回归方法确定以身高、体重为自变量，体表面积为应变量的回归方程。

儿童编号	体表面积（Y）	身高（X_1）	体重（X_2）
1	5.382	88.0	11.0
2	5.299	87.6	11.8
3	5.358	88.5	12.0
4	5.292	89.0	12.3
5	5.602	87.7	13.1
6	6.014	89.5	13.7
7	5.830	88.8	14.4
8	6.102	90.4	14.9
9	6.075	90.6	15.2
10	6.411	91.2	16.0

8.1.2.1 数据准备

激活数据管理窗口，定义变量名：体表面积为 Y，保留 3 位小数；身高、体重分别为 X1、X2，1 位小数。输入原始数据，结果如图 8.1 所示。

Newdata

10:x2 | 16

	y	x1	x2
1	5.382	88.0	11.0
2	5.299	87.6	11.8
3	5.358	88.5	12.0
4	5.292	89.0	12.3
5	5.602	87.7	13.1
6	6.014	89.5	13.7
7	5.830	88.8	14.4
8	6.102	90.4	14.9
9	6.075	90.6	15.2
10	6.411	91.2	16.0

图 8.1　原始数据的输入

8.1.2.2 统计分析

激活 Statistics 菜单选 Regression 中的 Linear...项，弹出 Linear Regression 对话框（如图 8.2 示）。从对话框左侧的变量列表中选 y，点击 ➢ 钮使之进入 Dependent 框，选 x1、x2，点击 ➢ 钮使之进入 Indepentdent(s)框；在 Method 处下拉菜单，共有 5 个选项：Enter（全部入选法）、Stepwise（逐步法）、Remove（强制剔除法）、Backward（向后法）、Forward（向前法）。本例选用 Enter 法。点击 OK 钮即完成分析。

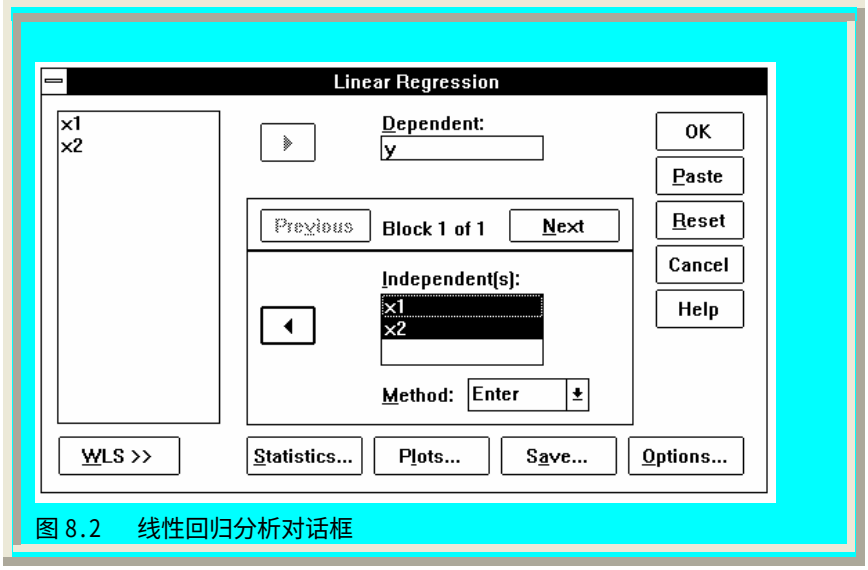


图 8.2　线性回归分析对话框

用户还可点击 Statistics...钮选择是否作变量的描述性统计、回归方程应变量的可信区间估计等分析；点击 Plots...钮选择是否作变量分布图（本例要求对标准化 Y 预测值作变量分布图）；点击 Save...钮选择对回归分析的有关结果是否作保存（本例要求对根据所确定的回归方程求得的未校正 Y 预测值和标准化 Y 预测值作保存）；点击 Options...钮选择变量入选与剔除的 α 、β 值和缺失值的处理方法。

8.1.2.3　结果解释

在结果输出窗口中将看到如下统计数据：

```
* * * *    M U L T I P L E    R E G R E S S I O N    * * * *

Listwise Deletion of Missing Data
Equation Number 1       Dependent Variable..     Y
Block Number   1.   Method:   Enter          X1          X2

Variable(s) Entered on Step Number
     1..       X2
     2..       X1

Multiple R                .94964
R Square                  .90181
Adjusted R Square         .87376
Standard Error            .14335
Analysis of Variance
                          DF          Sum of Squares          Mean Square
Regression                2                1.32104                .66052
Residual                  7                 .14384                .02055
F =          32.14499          Signif F =    .0003
```

---------- Variables in the Equation ----------

Variable	B	SE B	Beta	T	Sig T
X1	.068701	.074768	.215256	.919	.3887
X2	.183756	.056816	.757660	3.234	.0144
(Constant)	-2.856476	6.017776		-.475	.6495

End Block Number 1 All requested variables entered.

结果显示，本例以X1、X2 为自变量，Y为应变量，采用全部入选法建立回归方程。回归方程的复相关系数为 0.94964，决定系数（即r^2）为 0.90181，经方差分析，F=34.14499，P=0.0003，回归方程有效。回归方程为Y=0.0687101X1+0.183756X2-2.856476。

本例要求按所建立的回归方程计算 Y 预测值和标准化 Y 预测值(所谓标准化 Y 预测值是指将根据回归方程求得的 Y 预测值转化成按均数为 0、标准差为 1 的标准正态分布的 Y 值）并将计算结果保存入原数据库。系统将原始的 X1、X2 值代入方程求 Y 值预测值（即库中 pre_1 栏）和标准化 Y 预测值（即库中 zpr_1 栏），详见图 8.3。

Newdata

	y	x1	x2	pre_1	zpr_1
1	5.382	88.0	11.0	5.21050	-1.37292
2	5.299	87.6	11.8	5.33003	-1.06095
3	5.358	88.5	12.0	5.42861	-.80363
4	5.292	89.0	12.3	5.51809	-.57009
5	5.602	87.7	13.1	5.57578	-.41950
6	6.014	89.5	13.7	5.80970	.19105
7	5.830	88.8	14.4	5.89023	.40127
8	6.102	90.4	14.9	6.09203	.92799
9	6.075	90.6	15.2	6.16090	1.10774
10	6.411	91.2	16.0	6.34913	1.59903

图 8.3　计算结果的保存

本例还要求对标准化 Y 预测值作变量分布图，系统将绘制的统计图送向 Chart Carousel 窗口，双击该窗口可见下图显示结果。

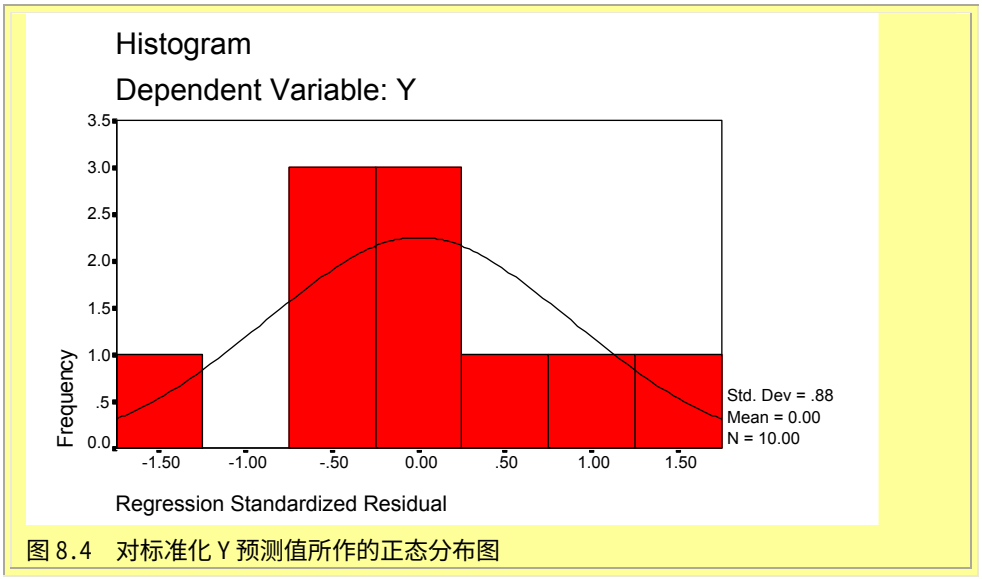


图 8.4　对标准化 Y 预测值所作的正态分布图

第二节 Curve Estimation 过程

8.2.1 主要功能

调用此过程可完成下列有关曲线拟合的功能：

1、Linear：拟合直线方程（实际上与Linear过程的二元直线回归相同，即$Y = b_0 + b_1X$）；

2、Quadratic：拟合二次方程（$Y = b_0 + b_1X + b_2X^2$）；

3、Compound：拟合复合曲线模型（$Y = b_0 \times b_1^X$）；

4、Growth：拟合等比级数曲线模型（$Y = e^{(b0+b1X)}$）；

5、Logarithmic：拟合对数方程（$Y = b_0 + b_1 \ln X$）

6、Cubic：拟合三次方程（$Y = b_0 + b_1X + b_2X^2 + b_3X^3$）；

7、S：拟合S形曲线（$Y = e^{(b0+b1/X)}$）；

8、Exponential：拟合指数方程（$Y = b_0\, e^{b1X}$）；

9、Inverse：数据按$Y = b_0 + b_1/X$进行变换；

10、Power：拟合乘幂曲线模型（$Y = b_0X^{b1}$）；

11、Logistic：拟合Logistic曲线模型（$Y = 1/(1/u + b_0 \times b_1^X)$）。

8.2.2 实例操作

[例 8.2]某地 1963 年调查得儿童年龄（岁）X 与锡克试验阴性率（%）Y 的资料如下，试拟合对数曲线。

年龄（岁） X	锡克试验阴性率（%） Y
1	57.1
2	76.0
3	90.9
4	93.0
5	96.7
6	95.6
7	96.2

8.2.2.1 数据准备

激活数据管理窗口，定义变量名：锡克试验阴性率为 Y，年龄为 X，输入原始数据。

8.2.2.2 统计分析

激活 Statistics 菜单选 Regression 中的 Curve Estimation...项，弹出 Curve Estimation 对话框（如图 8.5 示）。从对话框左侧的变量列表中选 y，点击 ➢ 钮使之进入 Dependent 框，选 x，点击 ➢ 钮使之进入 Indepentdent(s)框；在 Model 框内选择所需的曲线模型，本例选择 Logarithmic 模型（即对数曲线）；选 Plot models 项要求绘制曲线拟合图；点击 Save...钮，弹出 Curve Estimation:Save 对话框，选择 Predicted value 项，要求在原始数据库中保存根据对数方程求出的 Y 预测值，点击 Continue 钮返回 Curve Estimation 对话框，再点击 OK 钮即可。

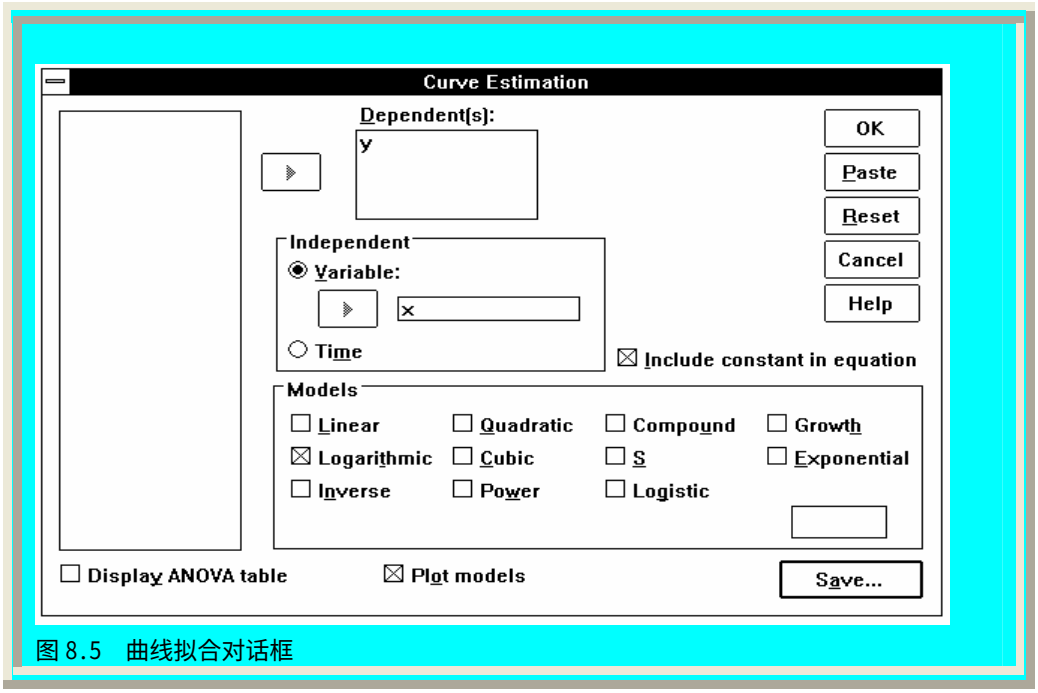


图 8.5 曲线拟合对话框

8.2.2.3 结果解释

在结果输出窗口中将看到如下统计数据：

```
ndependent:   X
   Dependent     Mth      Rsq     d.f.          F          Sigf          b0          b1
```

Y	LOG	.913	5	52.32	.001	61.3259	20.6704

在以X为自变量、Y为应变量，采用对数曲线拟合方法建立的方程，决定系数R^2=0.913（接近于1），作拟合优度检验，方差分析表明：F=52.32，P=0.001，拟合度很好，对数方程为：Y=61.3259+20.6704lnX。

本例要求绘制曲线拟合图，结果如图 8.6 所示。

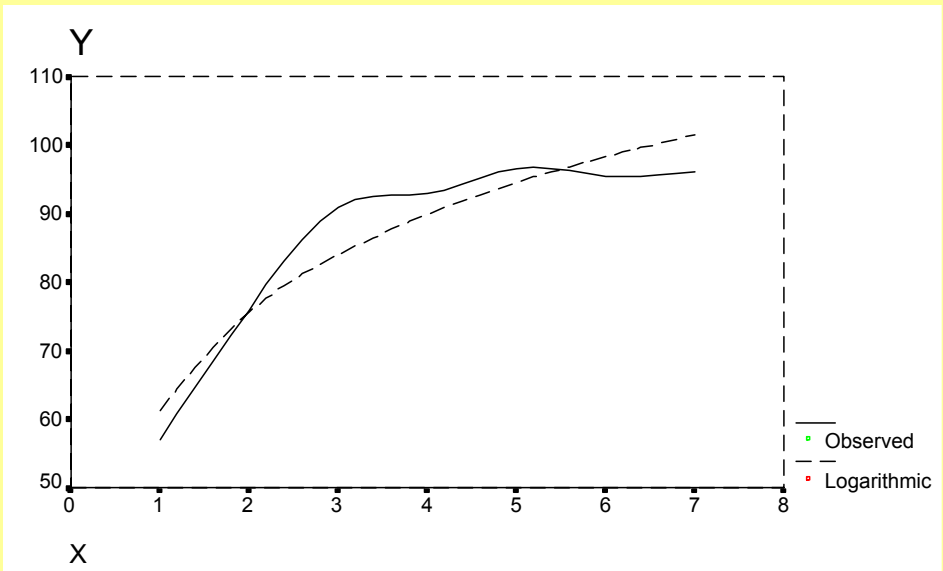


图 8.6　对数曲线拟合情形

根据方程 Y=61.3259+20.6704lnX，将原始数据 X 值代入，求得 Y 预测值（变量名为 fit_1）存入数据库中，参见图 8.7。

Newdata

	x	y	fit_1
1	1	57.1	61.32592
2	2	76.0	75.65356
3	3	90.9	84.03468
4	4	93.0	89.98119
5	5	96.7	94.59366
6	6	95.6	98.36232
7	7	96.2	101.54867

图 8.7　计算结果的保存

第三节 Logistic 过程

8.3.1 主要功能

调用此过程可完成 Logistic 回归的运算。所谓 Logistic 回归，是指应变量为二级计分或二类评定的回归分析，这在医学研究中经常遇到，如：死亡与否（即生、死二类评定）的概率跟病人自身生理状况和所患疾病的严重程度有关；对某种疾病的易感性的概率（患病、不患病二类评定）与个体性别、年龄、免疫水平等有关。此类问题的解决均可借助逻辑回归来完成。

特别指出，本节介绍的 Logistic 过程，应与日常所说的 Logistic 曲线模型（即 S 或倒 S 形曲线）相区别。用户如果要拟合 Logistic 曲线模型，可调用本章第二节 Curve Estimation 过程，系统提供 11 种曲线模型，其中含有 Logistic 曲线模型（参见上节）。

在一般的多元回归中，若以 P（概率）为应变量，则方程为 P=b0+b1X1+b2X2+…+bkXk, 但用该方程计算时，常会出现 P>1 或 P<0 的不合理情形。为此，对 P 作对数单位转换，即 logitP=ln(P/1-P)，于是，可得到 Logistic 回归方程为：

$$P = \frac{e^{b0+b1X1+b2X2+\cdots+bkXk}}{1+e^{b0+b1X1+b2X2+\cdots+bkXk}}$$

8.3.2 实例操作

[例 8.3]某医师研究男性胃癌患者发生术后院内感染的影响因素，资料如下表，请通过 Logistic 回归统计方法对主要影响因素进行分析。

术后感染（有无）Y	年龄（岁）X1	手术创伤程度（5 等级）X2	营养状态（3 等级）X3	术前预防性抗菌（有无）X4	白细胞数（$\times10^9$/L）X5	癌肿病理分度（TNM 得分总和）X6
有	69	4	2	无	5.6	9
有	72	5	3	无	4.4	6
无	57	3	2	无	9.7	4
无	41	1	1	有	11.2	5
无	32	1	1	有	10.4	5
有	65	3	3	有	7.0	5
无	58	3	2	有	3.1	6
有	54	4	2	无	6.6	6
有	55	2	2	有	7.9	7
无	59	1	1	有	6.0	4
无	64	2	2	无	9.1	6
无	36	1	1	有	8.4	8
无	42	3	1	有	5.3	6

无	48	4	2	有	4.6	5
无	50	1	2	有	12.8	4

8.3.2.1 数据准备

激活数据管理窗口，定义变量名：术后感染为 Y（字符变量，有输入 Y、无输入 N），年龄为 X1，手术创伤程度为 X2，营养状态为 X3，术前预防性抗菌为 X4（字符变量，有输入 Y、无输入 N），白细胞数为 X5，癌肿病理分度为 X6。按要求输入原始数据。

8.3.2.2 统计分析

激活 Statistics 菜单选 Regression 中的 Logistic...项，弹出 Logistic Regression 对话框（如图 8.8 示）。从对话框左侧的变量列表中选 y，点击 ➢ 钮使之进入 Dependent 框，选 x1、x2、x3、x4、x5 和 x6，点击 ➢ 钮使之进入 Covariates 框；点击 Method 处的下拉按钮，系统提供 7 种方法：

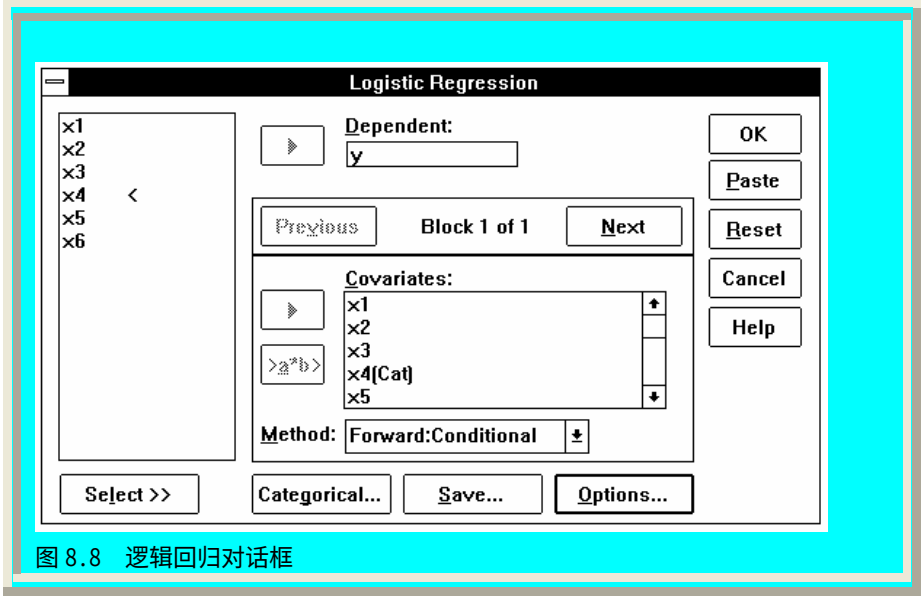


图 8.8 逻辑回归对话框

1、Enter：所有自变量强制进入回归方程；

2、Forward: Conditional：以假定参数为基础作似然比概率检验，向前逐步选择自变量；

3、Forward: LR：以最大局部似然为基础作似然比概率检验，向前逐步选择自变量；

4、Forward: Wald：作 Wald 概率统计法，向前逐步选择自变量；

5、Backward: Conditional：以假定参数为基础作似然比概率检验，向后逐步选择自变量；

6、Backward: LR：以最大局部似然为基础作似然比概率检验，向后逐步选择自变量；

7、Backward: Wald：作 Wald 概率统计法，向后逐步选择自变量。

本例选用 Forward: Conditional 法，以便选择有主要作用的影响因素；点击 Options...钮，弹出 Logistic Regression:Options 对话框，在 Display 框中选取 At last step 项，要求只显示最终计算结果，点击 Continue 钮返回 Logistic Regression 对话框，再点击 OK 钮即可。

8.3.2.3 结果解释

在结果输出窗口中将看到如下统计数据：

Dependent Variable Encoding:

Original Value	Internal Value
y	0
n	1

		Parameter Value	Freq	Coding (1)
X4	n	5		1.000
	y	10		-1.000

系统先对字符变量进行重新赋值，对于应变量 Y，回答是（Y）的赋值为 0，回答否（X）的赋值为 1；对于应变量 X4，回答是（Y）的赋值为-1，回答否（X）的赋值为 1。

Dependent Variable.. Y
Beginning Block Number 0. Initial Log Likelihood Function
-2 Log Likelihood 19.095425
* Constant is included in the model.

Beginning Block Number 1. Method: Forward Stepwise (COND)

Step	Improv. Chi-Sq.	df	sig	Model Chi-Sq.	df	sig	Correct Class %	Variable
1	8.510	1	.004	8.510	1	.004	80.00	IN: X3
2	6.766	1	.009	15.276	2	.000	93.33	IN: X6

No more variables can be deleted or added.

End Block Number 1 PIN = .0500 Limits reached.
Final Equation for Block 1

Estimation terminated at iteration number 12 because
Log Likelihood decreased by less than .01 percent.
-2 Log Likelihood 3.819
Goodness of Fit 3.000

	Chi-Square	df	Significance
Model Chi-Square	15.276	2	.0005
Improvement	6.766	1	.0093

Classification Table for Y

	Predicted y	n	Percent Correct
y			
n			

Observed				
y	y	4	1	80.00%
n	n	0	10	100.00%
			Overall	93.33%

---------- Variables in the Equation ----------

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
X3	-30.5171	298.0526	.0105	1	.9184	.0000	.0000
X6	-10.2797	107.9559	.0091	1	.9241	.0000	.0000
Constant	123.4053	1155.1065	.0114	1	.9149		

结果表明，第一步自变量X3 入选，方程分类能力达 80.00%；第二步自变量X6 入选，方程分类能力达 93.33%（参见结果中的分类分析表）；方程有效性经 χ^2检验， χ^2=15.276，P=0.0005。

Logistic 回归的分类概率方程为：

$$P = \frac{e^{123.4053-30.5171X3-10.2797X6}}{1+e^{123.4053-30.5171X3-10.2797X6}}$$

根据该方程，若一胃癌患者营养状态评分（X3）为 3，癌肿病理分度（X6）为 9，则其$P=4.5\times10^{-27}\approx0$，这意味着术后将发生院内感染；另一胃癌患者营养状态评分（X3）为 1，癌肿病理分度（X6）为 4，则其P=0.98105≈1，这意味着术后将不会发生院内感染。

第四节 Probit 过程

8.4.1 主要功能

调用此过程可完成剂量-效应关系的分析。通过概率单位使剂量-效应的 S 型曲线关系转化成直线，从而利用回归方程推算各效应水平的相应剂量值。

8.4.2 实例操作

[例 8.4]研究抗疟药环氯胍对小白鼠的毒性，试验结果如下表所示。试计算环氯胍的半数致死剂量。

剂量（mg/kg）	动物数	死亡数
12	5	5
9	7	6
7	19	11
6	34	17
5	38	12

4 3	12 5	2 0

8.4.2.1 数据准备

激活数据管理窗口，定义变量名：剂量为 DOSE、试验动物数为 OBSERVE、死亡动物数为 DEATH。然后输入原始数据。

8.4.2.2 统计分析

激活 Statistics 菜单选 Regression 中的 Probit...项，弹出 Probit Analysis 对话框（如图 8.9 示）。从对话框左侧的变量列表中选 death，点击 ➢ 钮使之进入 Response Frequency 框；选 observe，点击 ➢ 钮使之进入 Total Observed 框；选 dose，点击 ➢ 钮使之进入 Covariate(s)框，并下拉 Transform 菜单，选 Log base 10 项（即要求对剂量进行以 10 为底的对数转换）。

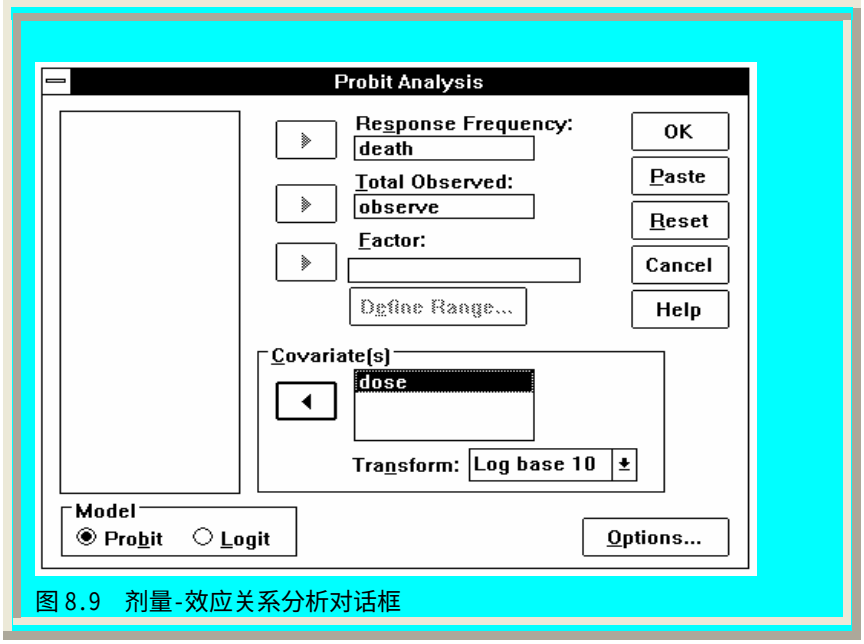


图 8.9 剂量-效应关系分析对话框

系统在 Model 栏中提供两种模型，一是概率单位模型（Probit），另一是比数比自然对数模型（Logit）。本例选用概率单位模型。

点击 Options...钮，弹出 Probit Analysis:Options 对话框，在 Natural Response Rate 栏选 Calculate from data 项，要求计算各剂量组的实际反应率。之后点击 Continue 钮返回 Probit Analysis 对话框，再点击 OK 钮即可。

8.4.2.3 结果解释

在结果输出窗口中将看到如下统计数据：

系统首先显示，共有 7 组原始数据采概率单位模型进行分析。回归方程的各参数在经过 14 次叠代运算后确定，即PROBIT = 5.95215 - 4.66313X 。该方程拟合优度 χ^2检验结果，χ^2 = 0.833，P=0.934，拟合良好。

```
DATA    Information
            7 unweighted cases accepted.
```

0 cases rejected because of missing data.
0 cases are in the control group.
0 cases rejected because LOG-transform can't be done.

MODEL Information

ONLY Normal Sigmoid is requested.

Natural Response rate to be estimated

CONTROL group is not provided.

Parameter estimates converged after 14 iterations.
Optimal solution found.

Parameter Estimates (PROBIT model: (PROBIT(p)) = Intercept + BX):

	Regression Coeff.	Standard Error	Coeff./S.E.
DOSE	5.95215	2.39832	2.48180

Intercept	Standard Error	Intercept/S.E.
-4.66313	2.19942	-2.12017

Estimate of Natural Response Rate = .000000 with S.E. = .26448

Pearson Goodness-of-Fit Chi Square = .833 DF = 4 P = .934

Since Goodness-of-Fit Chi square is NOT significant, no heterogeneity factor is used in the calculation of confidence limits.

Covariance(below) and Correlation(above) Matrices of Parameter Estimates

	DOSE	NAT RESP
DOSE	5.75192	.82927
NAT RESP	.52601	.06995

接着，系统显示剂量对数值（DOSE）、实际观察例数（Number of Subjects）、试验动物反应数（Observed Responses）、预期反应数（Expected Responses）、残差（ Residual）和效应的概率（Prob）。之后，显示各效应概率水平的剂量值及其 95%可信区间值，按本例要求，环氯胍的半数致死剂量（即 Prob = 0.50 时）为 6.07347，其 95%可信区间为 1.86305—7.54282。

Observed and Expected Frequencies

DOSE	Number of Subjects	Observed Responses	Expected Responses	Residual	Prob

1.08	5.0	5.0	4.804	.196	.96082
.95	7.0	6.0	5.917	.083	.84534
.85	19.0	11.0	12.221	-1.221	.64320
.78	34.0	17.0	16.573	.427	.48745
.70	38.0	12.0	11.688	.312	.30757
.60	12.0	2.0	1.682	.318	.14016
.48	5.0	.0	.171	-.171	.03413

Confidence Limits for Effective DOSE

		95% Confidence Limits	
Prob	DOSE	Lower	Upper
.01	2.46942	.02752	4.27407
.02	2.74406	.04534	4.54351
.03	2.93394	.06223	4.72430
.04	3.08539	.07895	4.86574
.05	3.21433	.09580	4.98445
.06	3.32832	.11294	5.08821
.07	3.43158	.13047	5.18134
.08	3.52676	.14845	5.26651
.09	3.61561	.16694	5.34550
.10	3.69937	.18597	5.41954
.15	4.06733	.29060	5.74092
.20	4.38570	.41395	6.01572
.25	4.67862	.56021	6.26792
.30	4.95831	.73436	6.51010
.35	5.23239	.94261	6.75084
.40	5.50646	1.19286	6.99754
.45	5.78528	1.49529	7.25814
.50	6.07347	1.86305	7.54282
.55	6.37600	2.31299	7.86673
.60	6.69886	2.86587	8.25522
.65	7.04974	3.54438	8.75565
.70	7.43943	4.36394	9.46545
.75	7.88416	5.30688	10.59748
.80	8.41075	6.29069	12.60617
.85	9.06910	7.21514	16.40564
.90	9.97116	8.09412	24.20725
.91	10.20216	8.27760	26.73478
.92	10.45919	8.46892	29.82525
.93	10.74928	8.67177	33.68627
.94	11.08278	8.89128	38.64769
.95	11.47580	9.13511	45.27000
.96	11.95538	9.41572	54.59759

.97	12.57252	9.75590	68.85554
.98	13.44250	10.20577	93.92908
.99	14.93751	10.92195	153.73112

最后，系统输出以剂量对数值为自变量 X、以概率单位为应变量 Y 的回归直线散点图，从图中各点的分布状态亦可看出，回归直线的拟合程度是很好的。

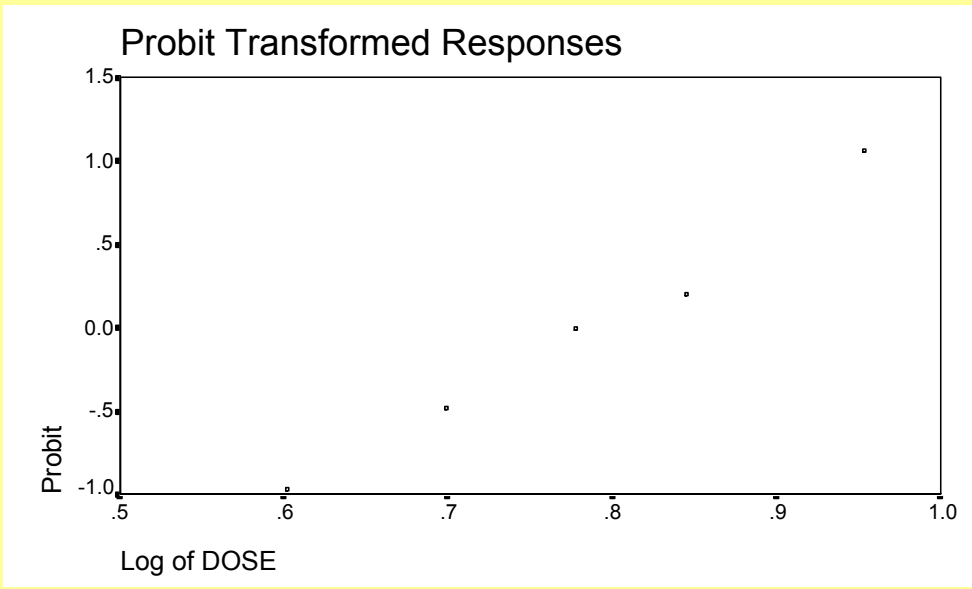


图 8.10　剂量-效应关系回归直线散点图

第五节 Nonlinear 过程

8.5.1 主要功能

调用此过程可完成非线性回归的运算。所谓非线性回归，即为曲线型的回归分析，一些曲线模型我们已在本章第二节中述及。但在医学研究中经，还经常会遇到除本章第二节中述及的曲线模型，对此，SPSS 提供 Nonlinear 过程让用户根据实际需要，建立各种曲线模型以用于研究变量间的相互关系。在医学中，如细菌繁殖与培养时间关系的研究即可借助 Nonlinear 过程完成。

下面一些曲线模型是在论文中较常见的，提供给用户应用时作参考：

模型名称		模型表达式
Asympt. Regression	1	$Y = b1 + b2 \times \exp(b3^{X})$
Asympt. Regression	2	$Y = b1 - (b2 \times (b3^{X}))$
Density		$Y = (b1 + b2 \times X)^{(-1/b3)}$
Gauss		$Y = b1 \times (1 - b3 \times \exp(-b2 \times X^{2}))$
Gompertz		$Y = b1 \times \exp(-b2 \times \exp(-b3 \times X))$
Johnson-Schumacher		$Y = b1 \times \exp(-b2 / (X + b3))$

Log Modified	$Y = (b1 + b3 \times X)^{b2}$
Log-Logistic	$Y = b1 - \ln(1 + b2 \times \exp(-b3 \times X))$
Metcherlich Law of Dim. Ret.	$Y = b1 + b2 \times \exp(-b3 \times X)$
Michaelis Menten	$Y = b1 \times X /(X + b2)$
Morgan-Mercer-Florin	$Y = (b1 \times b2 + b3 \times X^{b4})/(b2 + X^{b4})$
Peal-Reed	$Y = b1 /(1 + b2 \times \exp(-(b3 \times X + b4 \times X^{2} + b5 \times X^{3})))$
Ratio of Cubics	$Y = (b1 + b2 \times X + b3 \times X^{2} + b4 \times X^{3})/(b5 \times X^{3})$
Ratio of Quadratics	$Y = (b1 + b2 \times X + b3 \times X^{2})/(b4 \times X^{2})$
Richards	$Y = b1 /((1 + b3 \times \exp(- b2 \times X))^{(1/ b4)})$
Verhulst	$Y = b1 /(1 + b3 \times \exp(- b2 \times X))$
Von Bertalanffy	$Y = (b1^{(1 - b4)} - b2 \times \exp(-b3 \times X))^{(1/(1 - b4))}$
Weibull	$Y = b1 - b2 \times \exp(- b3 \times X^{b4})$
Yield Density	$Y = (b1 + b2 \times X + b3 \times X^{2})^{(-1)}$

8.5.2 实例操作

[例 8.5]选取某地某年寿命表中 40-80 岁各年龄组的尚存人数资料如下表，请就该资料试拟合Gompertz曲线（$Y = b1 \times b2^{(b3x)}$）。

年龄组（岁）	年龄简化值（X）	尚存人数（Y）
40	0	81277
45	1	79258
50	2	76532
55	3	72850
60	4	67568
65	5	59911
70	6	50800
75	7	39325
80	8	28074

8.5.2.1 数据准备

激活数据管理窗口，定义变量名：年龄简化值为 X，尚存人数为 Y。输入原始数据。

8.5.2.2 统计分析

激活 Statistics 菜单选 Regression 中的 Nonlinear...项，弹出 Nonlinear Regression 对话框（如图 8.11 示）。从对话框左侧的变量列表中选 y，点击 ➢ 钮使之进入 Dependent 框。由于 SPSS 系统尚无法智能地自动拟合用户所需的曲线，故一方面要求用户估计方程中常数项和各系数项进行叠代运算的起始值，另一方面要求用户列出方程模型。对此，可首先点击 Nonlinear Regression 对话框的 Parameters... 钮，弹出 Nonlinear Regression: Parameters 对话框（图 8.12），在 Name 处定义系数名，在 Start Value 处输入起始值（这项工作是十分重要的，否则系统可能无法运算，甚至会因叠代次数过大导致 SPSS 系统的崩溃），本例定义 b1=8500、b2=1、b3=1.5，每定义一个系数，即点击 Add 钮加以确定；若在后面的运算中出错，则还可修改系数项的起始值，修改后点击 Change 钮加以确定；然后点击 Continue 钮返回 Nonlinear Regression 对话框。在 Model Expression 处写出曲线方程表达式，用户可借助系统

提供的数码盘和函数列表写出方程。本例要求计算根据回归方程求出的预测值，可点击 Save 钮，在 Nonlinear Regression:Save New Variables 对话框中选 Predicted value 项。最后点击 OK 钮即可。

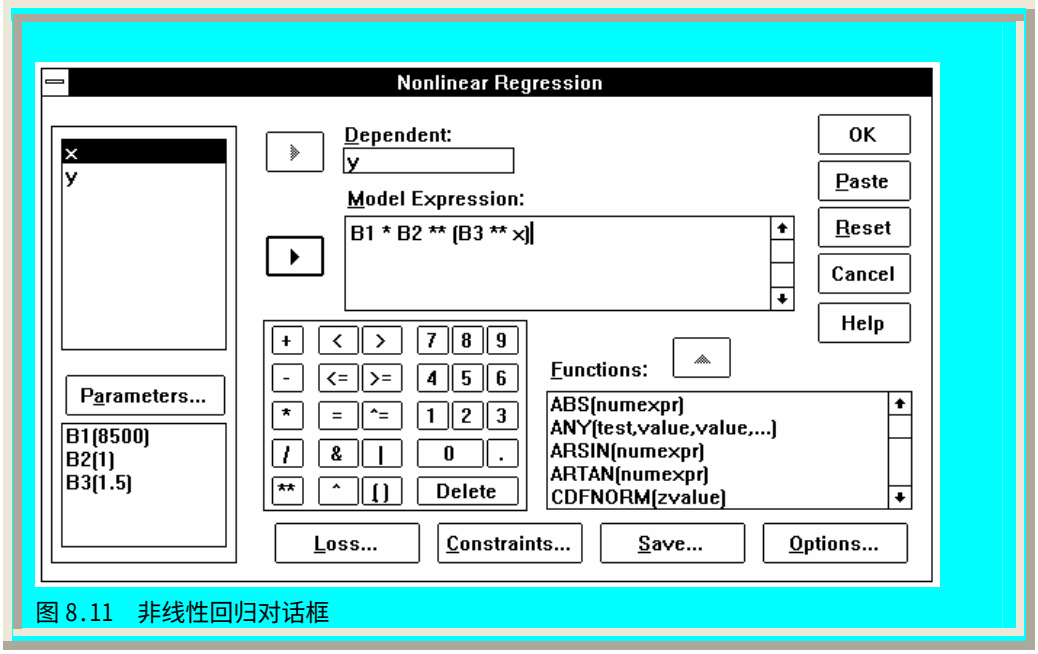


图 8.11　非线性回归对话框

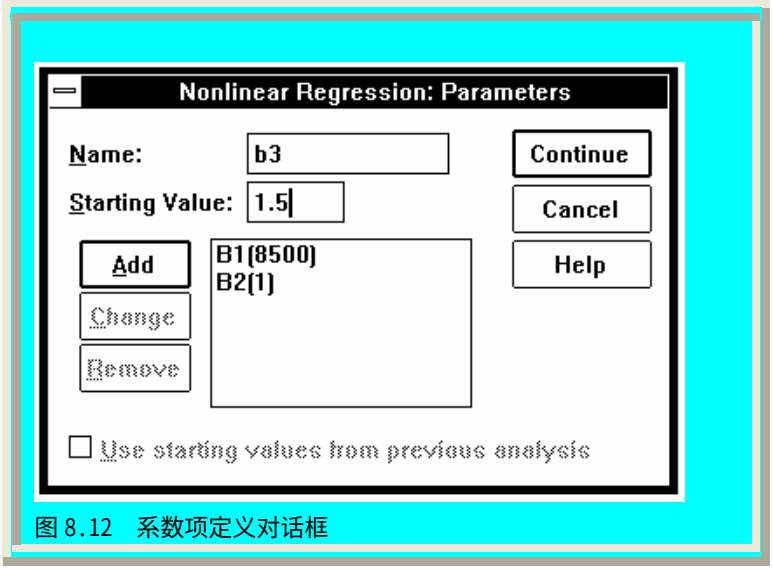


图 8.12　系数项定义对话框

8.5.2.3　结果解释

在结果输出窗口中将看到如下统计数据：

Iteration	Residual SS	B1	B2	B3
1	28327193463	8500.00000	1.00000000	1.50000000
1.1	14333434800	80175.3427	.739240551	1.50000000
2	14333434800	80175.3427	.739240551	1.50000000
2.1	3.8505E+11	194572.013	.006502086	-.21629077
2.2	800135019.6	83185.8046	.842994797	1.19852430
3	800135019.6	83185.8046	.842994797	1.19852430

3.1	12857378788	81201.8322	1.01579267	1.42927791
3.2	550558275.1	85774.2528	.850493197	1.21433127
4	550558275.1	85774.2528	.850493197	1.21433127
4.1	205793117.6	90637.3496	.859429212	1.25276932
5	205793117.6	90637.3496	.859429212	1.25276932
5.1	49937888.65	92251.6832	.905992700	1.33942536
6	49937888.65	92251.6832	.905992700	1.33942536
6.1	438492814.3	83503.5809	.966421043	1.46365602
6.2	14165723.65	91420.4568	.909112694	1.36083115
7	14165723.65	91420.4568	.909112694	1.36083115
7.1	8227661.248	89440.0706	.923463315	1.38898940
8	8227661.248	89440.0706	.923463315	1.38898940
8.1	17416856.86	85916.5498	.948299986	1.45005498
8.2	4600297.866	88467.6768	.930296397	1.40797724
9	4600297.866	88467.6768	.930296397	1.40797724
9.1	2761649.685	86538.9357	.943736707	1.44419408
10	2761649.685	86538.9357	.943736707	1.44419408
10.1	644830.0765	85633.9620	.949714917	1.46896660
11	644830.0765	85633.9620	.949714917	1.46896660
11.1	475140.3684	85680.9561	.949325567	1.46898044
12	475140.3684	85680.9561	.949325567	1.46898044
12.1	475135.4265	85679.2273	.949338713	1.46903683
13	475135.4265	85679.2273	.949338713	1.46903683
13.1	475135.4262	85679.2477	.949338590	1.46903640

Run stopped after 30 model evaluations and 13 derivative evaluations.
Iterations have been stopped because the relative reduction between successive residual sums of squares is at most SSCON = 1.000E-08

Nonlinear Regression Summary Statistics Dependent Variable Y

Source	DF	Sum of Squares	Mean Square
Regression	3	37121583327.6	12373861109.2
Residual	6	475135.42624	79189.23771
Uncorrected Total	9	37122058463.0	
(Corrected Total)	8	2823635793.56	

R squared = 1 - Residual SS / Corrected SS =.99983

			Asymptotic 95 % Confidence Interval	
Parameter	Estimate	Asymptotic Std. Error	Lower	Upper
B1	85679.247671	383.76368720	84740.211757	86618.283585
B2	.949338590	.002336270	.943621944	.955055236

B3	1.469036403	.008908976	1.447236923	1.490835883

Asymptotic Correlation Matrix of the Parameter Estimates

	B1	B2	B3
B1	1.0000	-.9245	-.8880
B2	-.9245	1.0000	.9902
B3	-.8880	.9902	1.0000

经 30 次叠代运算后，相邻两次的方程剩余均方差值不大于规定的 1×10^{-8}，满足要求；回归方程的决定系数$R^2 = 0.99983$，Gompertz曲线方程为：

$Y = 85679.247671\times0.94933859^{(1.469036403^X)}$

本例要求计算预测值，系统将结果存入原始数据库中（图 8.13），系统以 pred_作为预测值的变量名。由结果可见，预测值与实际值十分接近。

Newdata

	x	y	pred_
1	0	81277	81338.62
2	1	79258	79379.15
3	2	76532	76585.92
4	3	72850	72659.82
5	4	67568	67254.30
6	5	59911	60034.14
7	6	50800	50809.25
8	7	39325	39765.18
9	8	28074	27742.08

图 8.13　原始数据及其预测值

第九章 对数线性模型

对数线性模型是用于离散型数据或整理成列联表格式的计数资料的统计分析工具。在对数线性模型中，所有用作的分类的因素均为独立变量，列联表各单元中的例数为应变量。对于列联表资料，通常作 χ^2 检验，但 χ^2 检验无法系统地评价变量间的联系，也无法估计变量间相互作用的大小，而对数线性模型是处理这些问题的最佳方法。

第一节 General 过程

9.1.1 主要功能

调用该过程可对一个或多个二维列联表资料进行非层次对数线性分析。它只能拟合全饱和模型，即分类变量各自效应及其相互间效应均包含在对数线性模型中。

9.1.2 实例操作

[例 9-1]在住院病人中，研究其受教育程度与对保健服务满意程度的关系，资料整理成列联表后如下所示。

对保健服务满意程度(%)	受教育程度		
	高	中	低
满意	65 (91.5)	272 (93.8)	41 (97.6)
不满意	6 (8.5)	18 (6.2)	1 (2.4)

按一般情形作 x^2 检验，结果显示不同受教育程度的住院病人其对保健服务满意程度无差别。但从百分比分析中可见，随受教育程度的提高，满意程度有下降的趋势；且我们还想了解受教育程度与满意程度有无交互作用和交互作用的大小。对此，必须采用对数线性模型加以分析。

9.1.2.1 数据准备

激活数据管理窗口，定义变量名：实际观察频数的变量名为 freq，受教育程度和满意程度作为行、列分类变量（即独立变量），变量名分别为 educ、care。输入原始数据，结果如图 9.1 所示。如同第四章 Crosstab 过程中所述，为使列联表的频数有效，应选 Data 菜单的 Weight Cases...项，弹出 Weight Cases 对话框（图 9.2），激活 Weight cases by 项，从变量列表中选 freq 点击 ➢ 钮使之进入 Frequency Variable 框，点击 OK 钮即可。

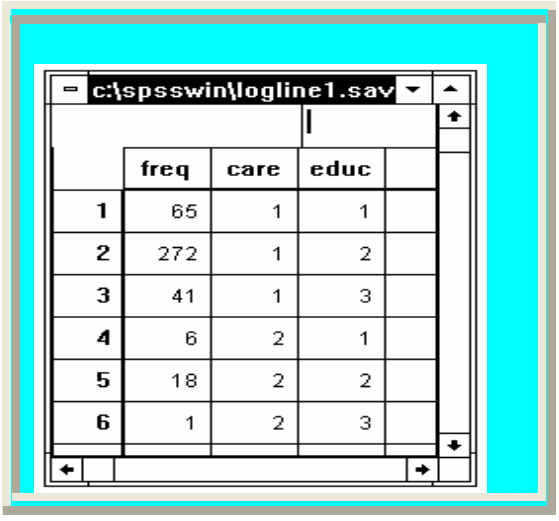


图 9.1　原始数据的输入

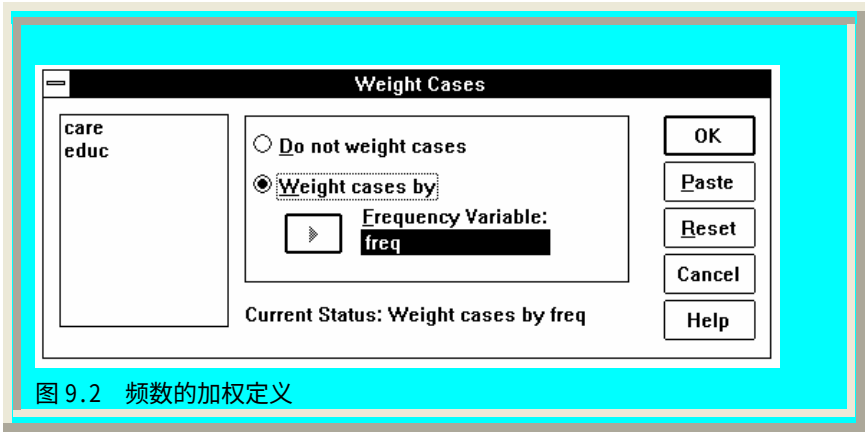


图 9.2　频数的加权定义

9.1.2.2　统计分析

激活 Statistics 菜单选 Loglinear 中的 General...项，弹出 General Loglinear Analysis 对话框（图 9.3）。从对话框左侧的变量列表中选 care，点击 ➢ 钮使之进入 Factor(s)框，点击 Define Range...钮，弹出 General Loglinear Analysis: Define Range 对话框，定义分类变量 care 的范围，本例为 1、2，故可在 Minimum 处键入 1，在 Maximum 处键入 2，点击 Continue 钮返回 General Loglinear Analysis 对话框。同法将变量 educ 选入 Factor(s)框，并定义其范围为 1、3。本例要求计算各分类变量主效应和交互作用的参数估计，故点击 Contrast...钮，弹出 General Loglinear Analysis:Contrasts 对话框，选择 Display parameter estimates 项，点击 Continue 钮返回 General Loglinear Analysis 对话框，最后点击 OK 钮即完成分析。

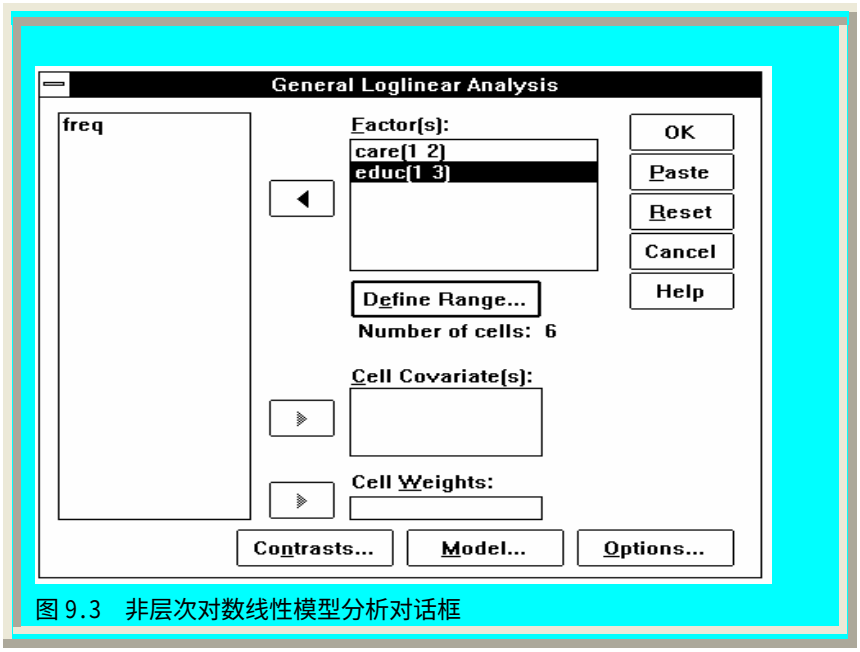


图 9.3　非层次对数线性模型分析对话框

9.1.2.3　结果解释

在结果输出窗口中将看到如下统计数据：

首先显示系统对 403 例资料进行分析，共有二个分类变量：CARE 为 2 水平，EDUC 为 3 水平。分析的效应有三类：满意程度（CARE）、教育程度（EDUC）和两者的交互作用（CARE BY EDUC）。

系统经 2 次叠代后即达到相邻二次估计之差不大于规定的 0.001。

DATA Information
6 unweighted cases accepted.
0 cases rejected because of out-of-range factor values.
0 cases rejected because of missing data.
403 weighted cases will be used in the analysis.

FACTOR Information

Factor	Level	Label
CARE	2	
EDUC	3	

DESIGN Information
1 Design/Model will be processed.
Correspondence Between Effects and Columns of Design/Model 1

Starting Column	Ending Column	Effect Name
1	1	CARE
2	3	EDUC
4	5	CARE BY EDUC

Note: for saturated models .500 has been added to all observed cells.
This value may be changed by using the CRITERIA = DELTA subcommand.

*** ML converged at iteration 2.
Maximum difference between successive iterations = .00000

由于本例对 Model（模型）未作定义，故系统采用默认的全饱和模型，因而期望例数（EXP.count）与实际例数（OBS. count）相同，进而残差（Residual）、标准化残差（Std.Resid）和校正残差（Adj.Resid）均为 0。

Observed, Expected Frequencies and Residuals

Factor	Code	OBS. count & PCT.	EXP. count & PCT.	Residual	Std. Resid.	Adj. Resid.
CARE	1					
EDUC	1	65.50 (16.13)	65.50 (16.13)	.0000	.0000	.0000
EDUC	2	272.50 (67.12)	272.50 (67.12)	.0000	.0000	.0000
EDUC	3	41.50 (10.22)	41.50 (10.22)	.0000	.0000	.0000
CARE	2					
EDUC	1	6.50 (1.60)	6.50 (1.60)	.0000	.0000	.0000
EDUC	2	18.50 (4.56)	18.50 (4.56)	.0000	.0000	.0000
EDUC	3	1.50 (.37)	1.50 (.37)	.0000	.0000	.0000

最后输出参数估计的结果。为了唯一地估计参数，系统强行限定同一分类变量的各水平参数之和为 0，故根据下列结果可推得各参数为：

$\lambda_{满意} = 1.386724028$

$\lambda_{不满意} = -1.386724028$

$\lambda_{高教育程度} = -0.091477207$

$\lambda_{中教育程度} = 1.144301306$

$\lambda_{低教育程度} = -1.052824099$

$\lambda_{满意.高教育程度} = -0.231600045$

$\lambda_{满意.中高教育程度} = -0.041790087$

$\lambda_{满意.低教育程度} = 0.273390132$

$\lambda_{不满意.高教育程度} = 0.231600045$

$\lambda_{不满意.中教育程度} = 0.041790087$

$\lambda_{不满意.低教育程度} = -0.273390132$

λ 值为正，表示正效应；反之为负效应；零为无效应。分析提供的信息是：①对保健服务的满意程度高于不满意程度；②中等教育程度者的满意程度>高等教育程度者的满意程度>低等教育程度者的满意程度；③通过受教育程度与对保健服务满意程度的交互作用研究，结果表明高、中等教育未能增加人们对现有保健服务状况的满意程度。

Estimates for Parameters

CARE

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	1.386724028	.15965	8.68589	1.07381	1.69964

EDUC

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
2	-.091477207	.19895	-.45980	-.48142	.29847
3	1.144301306	.17407	6.57393	.80313	1.48547

CARE BY EDUC

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
4	-.231600045	.19895	-1.16410	-.62154	.15834
5	-.041790087	.17407	-.24008	-.38296	.29938

第二节 Hierarchical 过程

9.2.1 主要功能

调用该过程可对多维列联表资料进行分层对数线性分析。所谓分层即并可根据用户指定的条件，对某一或某些主效应与交互作用进行剔除，从而形成包含特定层次阶项的各种模型。

9.2.2 实例操作

［例 9-2］为了研究 Colles 骨折在不同性别中的年龄分布情况，以说明不同性别者骨折的年龄差异及其年度变化，某地收集了 1978--1981 年的骨折资料，数据见下表。请作对数线性模型的分析。

年龄	1978		1979		1980		1981	
	男	女	男	女	男	女	男	女
0—19	55	17	43	9	89	20	140	41
20--59	165	260	101	233	104	202	137	278
60--89	50	94	29	115	56	95	54	153

9.2.2.1 数据准备

激活数据管理窗口，定义变量名：实际观察频数的变量名为 freq，年份、性别和年龄为分类变量，变量名分别为 year、sex 和 age。输入原始数据，其中年份 1978 至 1981 依次为 1、2、3、4，性别男为 1、女为 2，年龄分组依次为 1、2、3。之后选 Data 菜单的 Weight Cases...项，在 Weight Cases 对话框中激活 Weight cases by 项，从变量列表中选 freq 点击 ➢ 钮使之进入 Frequency Variable 框，点击 OK 钮完成对频数的权重定义。

9.2.2.2 统计分析

激活 Statistics 菜单选 Loglinear 中的 Hierarchical...项，弹出 Hierarchical Loglinear Analysis 对话框（图 9.4）。从对话框左侧的变量列表中选 age，点击 ➢ 钮使之进入 Factor(s)框，点击 Define Range... 钮，弹出 Hierarchical Loglinear Analysis: Define Range 对话框，定义分类变量 age 的范围，在 Minimum 处键入 1，在 Maximum 处键入 9，点击 Continue 钮返回 Hierarchical Loglinear Analysis 对话框。同法将变量 sex 选入 Factor(s)框，定义其范围为 1、2；将变量 year 选入 Factor(s)框，定义其范围为 1、4。

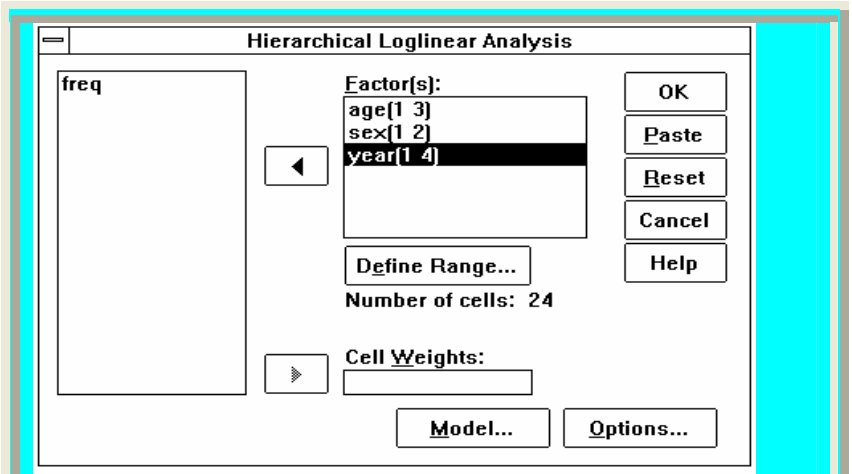


图 9.4　层次对数线性模型分析对话框

为了更好地拟合数据，并尽可能的简单和易于解释，本例选择向后剔除法建立模型，即从所有效应均在模型中开始，然后消除那些不满足保留判据的效应。点击 Model...钮，弹出 Hierarchical Loglinear Analysis: Model 对话框，在 Model Building 栏中选 Use backward elimination 项，点击 Continue 钮返回 Hierarchical Loglinear Analysis 对话框。

本例要求作参数估计，故点击 Options...钮，弹出 Hierarchical Loglinear Analysis: Options 对话框，在 Display for Saturated Model 栏中选 Parameter estimates 项，点击 Continue 钮返回 Hierarchical Loglinear Analysis 对话框，之后点击 OK 钮即完成分析。

9.2.2.3　结果解释

在结果输出窗口中将看到如下统计数据：

首先显示，共有 2540 个观察例数进入分析，其中分类变量 AGE 为 3 水平，SEX 为 2 水平，YEAR 为 4 水平。采用全饱和模型，高阶项为年龄、性别和年份三者的交互作用。（在层次对数线性模型分析中，当指定高阶项时，即意味着包含其所属变量所有可能组合的低阶项；如本例，即包含年龄和性别的交互作用、年龄和年份的交互作用、性别和年份的交互作用、年龄的主效应、性别的主效应、年份的主效应。从最高阶到最低阶共为 3 阶。）

```
DATA     Information
            24 unweighted cases accepted.
             0 cases rejected because of out-of-range factor values.
             3 cases rejected because of missing data.
          2540 weighted cases will be used in the analysis.

FACTOR Information
    Factor        Level       Label
    AGE             3
    SEX             2
    YEAR            4

DESIGN 1 has generating class
     AGE*SEX*YEAR
```

Note: For saturated models .500 has been added to all observed cells.
This value may be changed by using the CRITERIA = DELTA subcommand.

The Iterative Proportional Fit algorithm converged at iteration 1.
The maximum difference between observed and fitted marginal totals is .000
and the convergence criterion is .278

系统以全饱和模型为起始，故显示各变量的实际例数、期望例数、残差和标准化残差，因期望例数与实际例数相同，进而残差、标准化残差均为 0。

Observed, Expected Frequencies and Residuals.

Factor	Code	OBS count	EXP count	Residual	Std Resid
AGE	1				
SEX	1				
YEAR	1	55.5	55.5	.00	.00
YEAR	2	43.5	43.5	.00	.00
YEAR	3	89.5	89.5	.00	.00
YEAR	4	140.5	140.5	.00	.00
SEX	2				
YEAR	1	17.5	17.5	.00	.00
YEAR	2	9.5	9.5	.00	.00
YEAR	3	20.5	20.5	.00	.00
YEAR	4	41.5	41.5	.00	.00
AGE	2				
SEX	1				
YEAR	1	165.5	165.5	.00	.00
YEAR	2	101.5	101.5	.00	.00
YEAR	3	104.5	104.5	.00	.00
YEAR	4	137.5	137.5	.00	.00
SEX	2				
YEAR	1	260.5	260.5	.00	.00
YEAR	2	233.5	233.5	.00	.00
YEAR	3	202.5	202.5	.00	.00
YEAR	4	278.5	278.5	.00	.00
AGE	3				
SEX	1				
YEAR	1	50.5	50.5	.00	.00
YEAR	2	29.5	29.5	.00	.00
YEAR	3	56.5	56.5	.00	.00
YEAR	4	54.5	54.5	.00	.00
SEX	2				

YEAR	1	94.5	94.5	.00	.00
YEAR	2	115.5	115.5	.00	.00
YEAR	3	95.5	95.5	.00	.00
YEAR	4	153.5	153.5	.00	.00

Goodness-of-fit test statistics

Likelihood ratio chi square = .00000 DF = 0 P = 1.000

Pearson chi square = .00000 DF = 0 P = 1.000

下面，系统先显示某一阶及其更高阶交互效应为 0 时的似然比 x^2检验概率值，因K为 3 时的概率值=0.1964>0.05，故认为年龄、性别、年份三者的交互作用为 0，亦即含 1 阶（单一变量主效应）及 2 阶（变量两两交互效应）的模型就能恰当地表述数据。

接着，系统又显示特定阶交互效应为 0 时的似然比 x^2检验概率值，结果表明，单纯含 1 阶（单一变量主效应）或单纯含 2 阶（变量两两交互效应）的模型也能恰当地表述数据。

Tests that K-way and higher order effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
3	6	8.615	.1964	8.547	.2007	4
2	17	404.424	.0000	425.168	.0000	2
1	23	1279.591	.0000	1293.594	.0000	0

Tests that K-way effects are zero.

K	DF	L.R. Chisq	Prob	Pearson Chisq	Prob	Iteration
1	6	875.167	.0000	868.426	.0000	0
2	11	395.809	.0000	416.621	.0000	0
3	6	8.615	.1964	8.547	.2007	0

Note: For saturated models .500 has been added to all observed cells.

This value may be changed by using the CRITERIA = DELTA subcommand.

系统所确定的模型中各参数值如下所示，由于内容较多，各 λ 值如何推算及其所表示的意义，请读者参阅本章第一节。

Estimates for Parameters.

AGE*SEX*YEAR

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	-.1412276052	.08417	-1.67784	-.30621	.02375
2	.1674922915	.10130	1.65335	-.03106	.36605
3	-.0169870288	.07921	-.21447	-.17223	.13826
4	.0577506145	.05557	1.03925	-.05117	.16667
5	-.0069187948	.06504	-.10637	-.13440	.12057

6	-.0817851831	.05570	-1.46819	-.19097	.02740

AGE*SEX

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	.7059980126	.04848	14.56319	.61098	.80102
2	-.2968871102	.03276	-9.06301	-.36109	-.23268

AGE*YEAR

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	-.1762097434	.08417	-2.09344	-.34119	-.01123
2	-.3051792054	.10130	-3.01249	-.50374	-.10662
3	.1339590237	.07921	1.69127	-.02129	.28920
4	.1990874838	.05557	3.58269	.09017	.30800
5	.1982170140	.06504	3.04744	.07073	.32570
6	-.1646071030	.05570	-2.95499	-.27379	-.05543

SEX*YEAR

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	.0471962901	.04918	.95960	-.04920	.14360
2	-.0778801067	.05818	-1.33868	-.19191	.03615
3	.0827715134	.04734	1.74836	-.01002	.17556

AGE

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	-.7212868272	.04848	-14.87857	-.81630	-.62627
2	.7999110228	.03276	24.41872	.73571	.86412

SEX

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	-.0348756276	.02856	-1.22099	-.09086	.02111

YEAR

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	-.0205234390	.04918	-.41728	-.11692	.07588
2	-.3188195595	.05818	-5.48020	-.43285	-.20479
3	-.0126524013	.04734	-.26725	-.10544	.08014

系统开始对全饱和模型进行从高阶到低阶的效应项剔除。第一步，剔除 3 阶交互效应项（AGE*SEX*YEAR）导致 x^2 值为 8.615，概率为 0.1964（不小于默认判据 0.05），故该效应项被剔除。

第二步，剔除 2 阶交互效应项，概率均小于 0.05，故 2 阶交互效应项不能剔除。即本例用 2 阶交互效应项（同时含 1 阶主效应项）描述模型已为最佳。

```
Backward Elimination (p = .050) for DESIGN 1 with generating class
   AGE*SEX*YEAR
 Likelihood ratio chi square =          .00000       DF = 0    P = 1.000

If Deleted Simple Effect is                  DF     L.R. Chisq Change       Prob       Iter
 AGE*SEX*YEAR                                 6                8.615          .1964      4

Step 1
   The best model has generating class
         AGE*SEX
         AGE*YEAR
         SEX*YEAR
   Likelihood ratio chi square =         8.61546       DF = 6    P =    .196

If Deleted Simple Effect is                  DF     L.R. Chisq Change       Prob       Iter
 AGE*SEX                                      2              310.816          .0000      2
 AGE*YEAR                                     6               62.829          .0000      2
 SEX*YEAR                                     3               13.024          .0046      2

Step 2
   The best model has generating class
         AGE*SEX
         AGE*YEAR
         SEX*YEAR
   Likelihood ratio chi square =         8.61546       DF = 6    P =    .196

The final model has generating class
      AGE*SEX
      AGE*YEAR
      SEX*YEAR

The Iterative Proportional Fit algorithm converged at iteration 0.
The maximum difference between observed and fitted marginal totals is        .131
and the convergence criterion is         .278
```

由于剔除了 3 阶交互效应项，故原全饱和模型变为层次模型，因而期望例数改变，期望例数与实际例数不同，进而残差、标准化残差均不为 0。若标准化残差界于-1.96—1.96 范围内，则表示模型是恰当的。从下面的结果可知，本例的标准化残差均在-1.96—1.96 范围内，故层次模型是适合的。

```
Observed, Expected Frequencies and Residuals.
Factor          Code          OBS count    EXP count    Residual    Std Resid
AGE               1
```

SEX	1				
YEAR	1	55.0	59.0	-4.05	-.53
YEAR	2	43.0	39.1	3.88	.62
YEAR	3	89.0	88.3	.69	.07
YEAR	4	140.0	140.5	-.50	-.04
SEX	2				
YEAR	1	17.0	13.0	4.04	1.12
YEAR	2	9.0	12.9	-3.88	-1.08
YEAR	3	20.0	20.7	-.70	-.15
YEAR	4	41.0	40.5	.53	.08
AGE	2				
SEX	1				
YEAR	1	165.0	163.0	1.99	.16
YEAR	2	101.0	97.9	3.07	.31
YEAR	3	104.0	112.6	-8.62	-.81
YEAR	4	137.0	133.5	3.54	.31
SEX	2				
YEAR	1	260.0	262.0	-1.99	-.12
YEAR	2	233.0	236.1	-3.07	-.20
YEAR	3	202.0	193.4	8.62	.62
YEAR	4	278.0	281.6	-3.55	-.21
AGE	3				
SEX	1				
YEAR	1	50.0	47.9	2.06	.30
YEAR	2	29.0	36.0	-6.95	-1.16
YEAR	3	56.0	48.1	7.92	1.14
YEAR	4	54.0	57.0	-3.03	-.40
SEX	2				
YEAR	1	94.0	96.1	-2.05	-.21
YEAR	2	115.0	108.0	6.95	.67
YEAR	3	95.0	102.9	-7.92	-.78
YEAR	4	153.0	150.0	3.02	.25

Goodness-of-fit test statistics

Likelihood ratio chi square = 8.61546 DF = 6 P = .196

Pearson chi square = 8.54688 DF = 6 P = .201

第三节 Logit 过程

9.3.1 主要功能

调用此过程可完成对一个应变量与一个或多个自变量之间对数线性模型的拟合。如果分类变量未区分应变量和自变量，那么应采用本章第一、二节介绍的方法；如果应变量是二分计量，自变量是连续计量，那么应采用 Logistic 回归方法（详见第八章）。

9.3.2 实例操作

[例 9.3]在艾滋病（AIDS）相关的知识、观念、行为研究（KAB Study）中，获得了不同年龄和受教育水平的公众，对预防 AIDS 知识掌握程度的资料，经整理成列联表如下所示。很明显，对预防 AIDS 知识的掌握程度与公众的年龄和受教育水平有关，即若预防 AIDS 知识掌握程度为应变量，则应该受到年龄和受教育水平两个自变量的影响。下面将运用带应变量的对数线性模型进行分析。

受教育水平	年龄	预防 AIDS 知识掌握程度 好	一般	差
高	20- 30- 40- 50-	53 28 31 19	40 21 32 6	2 3 8 11
中	20- 30- 40- 50-	67 71 38 9	103 141 94 66	24 101 87 136
低	20- 30- 40- 50-	2 16 8 3	3 22 98 76	17 19 247 156

9.3.2.1 数据准备

激活数据管理窗口，定义变量名：实际观察频数的变量名为 freq；预防 AIDS 知识掌握程度变量名为 aids，按好、一般、差分别输入 1、2、3；受教育水平变量名为 educ，按高、中、低分别输入 1、2、3；年龄变量名为 age，20-至 50-依次输入 1—4。输入原始数据后选 Data 菜单的 Weight Cases... 项，在 Weight Cases 对话框中激活 Weight cases by 项，从变量列表中选 freq 点击 ➢ 钮使之进入 Frequency Variable 框，点击 OK 钮即可。

9.3.2.2 统计分析

激活 Statistics 菜单选 Loglinear 中的 Logit...项，弹出 Logit Loglinear Analysis 对话框（图 9.5）。

从对话框左侧的变量列表中选 aids，点击 ➢ 钮使之进入 Dependent 框，点击 Define Range...钮，弹出 Logit Loglinear Analysis: Define Range 对话框，定义应变量 aids 的范围，在 Minimum 处键入 1，在 Maximum 处键入 3，点击 Continue 钮返回 Logit Loglinear Analysis 对话框。从对话框左侧的变量列表中选 age，点击 ➢ 钮使之进入 Factor(s)框，点击 Define Range...钮，定义自变量 age 的范围为 1、4；同法将自变量 educ 选入 Factor(s)框，并定义其范围为 1、3。本例要求计算各变量主效应和交互作用的参数估计，故点击 Contrast...钮，弹出 Logit Loglinear Analysis:Contrasts 对话框，选择 Display parameter estimates 项，点击 Continue 钮返回 Logit Loglinear Analysis 对话框，最后点击 OK 钮即完成分析。

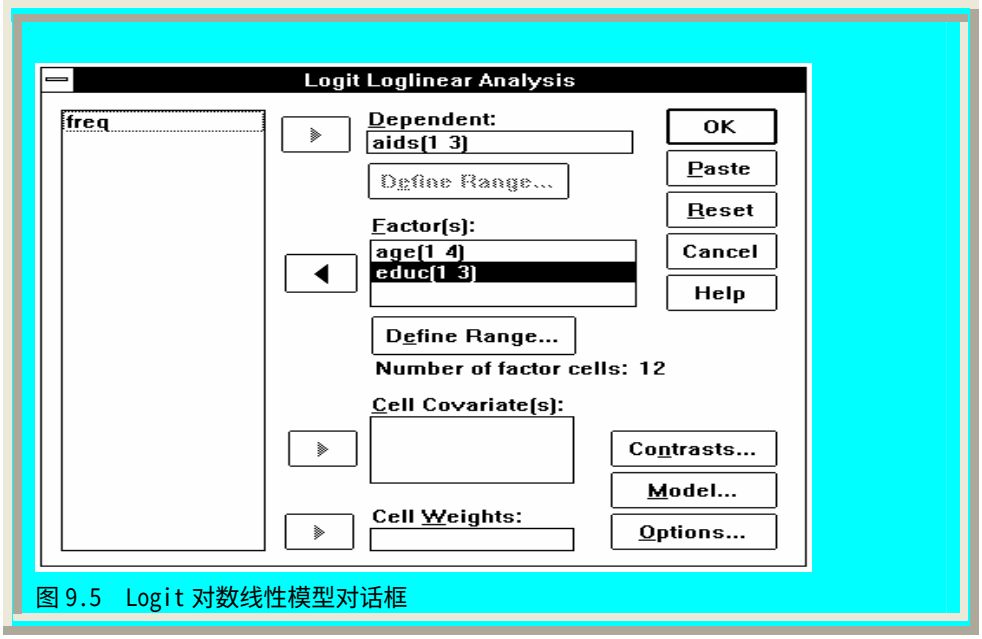


图 9.5　Logit 对数线性模型对话框

9.3.2.3　结果解释

在结果输出窗口中将看到如下统计数据：

系统显示共有 1858 个观察例数进入分析，分析涉及三个变量，其中 AIDS 为 3 水平，AGE 为 4 水平，EDUC 为 3 水平。将产生 3 阶 4 类效应，即：预防 AIDS 知识掌握程度主效应（因 AIDS 被定义为应变量，故不再分析子变量 AGE、EDUC 的主效应），预防 AIDS 知识掌握程度分别与年龄、受教育程度的交互效应，预防 AIDS 知识掌握程度、年龄、受教育程度三者的交互效应。之后系统显示实际例数、期望例数、残差、标准化残差和校正残差。

```
DATA    Information
          36 unweighted cases accepted.
           0 cases rejected because of out-of-range factor values.
           0 cases rejected because of missing data.
        1858 weighted cases will be used in the analysis.

FACTOR Information
    Factor        Level      Label
    AIDS            3
    AGE             4
    EDUC            3
```

DESIGN Information

1 Design/Model will be processed.

Correspondence Between Effects and Columns of Design/Model 1

Starting Column	Ending Column	Effect Name
1	2	AIDS
3	8	AIDS BY AGE
9	12	AIDS BY EDUC
13	24	AIDS BY AGE BY EDUC

Note: for saturated models .500 has been added to all observed cells.
This value may be changed by using the CRITERIA = DELTA subcommand.

*** ML converged at iteration 2.

Maximum difference between successive iterations = .00000.

Observed, Expected Frequencies and Residuals

Factor	Code	OBS. count & PCT.	EXP. count & PCT.	Residual	Std. Resid.	Adj. Resid.
AIDS	1					
AGE	1					
EDUC	1	53.50 (55.44)	53.50 (55.44)	.0000	.0000	.0000
EDUC	2	67.50 (34.53)	67.50 (34.53)	.0000	.0000	.0000
EDUC	3	2.50 (10.64)	2.50 (10.64)	.0000	.0000	.0000
AGE	2					
EDUC	1	28.50 (53.27)	28.50 (53.27)	.0000	.0000	.0000
EDUC	2	71.50 (22.73)	71.50 (22.73)	.0000	.0000	.0000
EDUC	3	16.50 (28.21)	16.50 (28.21)	.0000	.0000	.0000
AGE	3					
EDUC	1	31.50 (43.45)	31.50 (43.45)	.0000	.0000	.0000
EDUC	2	38.50 (17.46)	38.50 (17.46)	.0000	.0000	.0000
EDUC	3	8.50 (2.40)	8.50 (2.40)	.0000	.0000	.0000
AGE	4					
EDUC	1	19.50 (52.00)	19.50 (52.00)	.0000	.0000	.0000
EDUC	2	9.50 (4.47)	9.50 (4.47)	.0000	.0000	.0000
EDUC	3	3.50 (1.48)	3.50 (1.48)	.0000	.0000	.0000
AIDS	2					
AGE	1					
EDUC	1	40.50 (41.97)	40.50 (41.97)	.0000	.0000	.0000
EDUC	2	103.50 (52.94)	103.50 (52.94)	.0000	.0000	.0000
EDUC	3	3.50 (14.89)	3.50 (14.89)	.0000	.0000	.0000
AGE	2					

EDUC	1	21.50 (40.19)	21.50 (40.19)	.0000	.0000	.0000
EDUC	2	141.50 (44.99)	141.50 (44.99)	.0000	.0000	.0000
EDUC	3	22.50 (38.46)	22.50 (38.46)	.0000	.0000	.0000
AGE	3					
EDUC	1	32.50 (44.83)	32.50 (44.83)	.0000	.0000	.0000
EDUC	2	94.50 (42.86)	94.50 (42.86)	.0000	.0000	.0000
EDUC	3	98.50 (27.79)	98.50 (27.79)	.0000	.0000	.0000
AGE	4					
EDUC	1	6.50 (17.33)	6.50 (17.33)	.0000	.0000	.0000
EDUC	2	66.50 (31.29)	66.50 (31.29)	.0000	.0000	.0000
EDUC	3	76.50 (32.35)	76.50 (32.35)	.0000	.0000	.0000
AIDS	3					
AGE	1					
EDUC	1	2.50 (2.59)	2.50 (2.59)	.0000	.0000	.0000
EDUC	2	24.50 (12.53)	24.50 (12.53)	.0000	.0000	.0000
EDUC	3	17.50 (74.47)	17.50 (74.47)	.0000	.0000	.0000
AGE	2					
EDUC	1	3.50 (6.54)	3.50 (6.54)	.0000	.0000	.0000
EDUC	2	101.50 (32.27)	101.50 (32.27)	.0000	.0000	.0000
EDUC	3	19.50 (33.33)	19.50 (33.33)	.0000	.0000	.0000
AGE	3					
EDUC	1	8.50 (11.72)	8.50 (11.72)	.0000	.0000	.0000
EDUC	2	87.50 (39.68)	87.50 (39.68)	.0000	.0000	.0000
EDUC	3	247.50 (69.82)	247.50 (69.82)	.0000	.0000	.0000
AGE	4					
EDUC	1	11.50 (30.67)	11.50 (30.67)	.0000	.0000	.0000
EDUC	2	136.50 (64.24)	136.50 (64.24)	.0000	.0000	.0000
EDUC	3	156.50 (66.17)	156.50 (66.17)	.0000	.0000	.0000

Goodness-of-Fit test statistics

Likelihood Ratio Chi Square =	.00000	DF = 0	P = 1.000
Pearson Chi Square =	.00000	DF = 0	P = 1.000

下一段为拟合优度的检验。系统采用分散相似测量法（Dispersion Similarity Measure），测量值界于-1 至+1 之间，愈靠近| 1 |，拟合优度愈好。本例为 0.145879。

Analysis of Dispersion

Source of Variation	Entropy	Dispersion Concentration	DF
Due to Model	314.875	173.206	
Due to Residual	1642.491	1014.119	
Total	1957.365	1187.325	3750

Measures of Association

Entropy =	.160867
Concentration =	.145879

最后，系统输出对数线性模型的各效应参数值。

由于内容较多，具体推算过程不再赘述（参阅本章第一节）。此处以 AIDS 主效应和 AIDS 与 EDUC 交互效应为例，演示如下：

预防 AIDS 知识掌握程度主效应部分，参数为

λ AIDS-好 = -0.378234829

λ AIDS-一般 = 0.3307195684

λ AIDS-差 = 0- (-0.378234829) - 0.3307195684 = 0.0475152606

这表明公众预防 AIDS 知识掌握程度一般。

预防 AIDS 知识掌握程度与受教育水平交互效应部分，参数为

λ AIDS 好-EDUC 高 = 1.097077448

λ AIDS 好-EDUC 中 = -0.186500026

λ AIDS 好-EDUC 低 = 0 - 1.097077448 - (-0.186500026) = -0.910577422

λ AIDS 一般-EDUC 高 = -0.018774593

λ AIDS 一般-EDUC 中 = 0.0930200827

λ AIDS 一般-EDUC 低 = 0 - (-0.018774593) - 0.0930200827 = 0.0742454897

λ AIDS 差-EDUC 高 = 0 - 1.097077448 - (-0.018774593) = 1.078302855

λ AIDS 差-EDUC 中 = 0 - (-0.186500026) - 0.0930200827 = 0.0934799433

λ AIDS 差-EDUC 低 = 0 - 1.078302855 - 0.0934799433 = -1.1717827983

这表明受教育水平高，预防 AIDS 知识掌握程度好；受教育水平低，预防 AIDS 知识掌握程度一般。为什么不体现受教育水平低，预防 AIDS 知识掌握程度差的信息呢？显然，这还需要结合年龄的因素进行分析。若用户将全部 λ 值都推算出来，其中会得到：

λ AIDS 差-AGE20-EDUC 低 = 0.727782785

λ AIDS 差-AGE30-EDUC 低 = -0.546798785

λ AIDS 差-AGE40-EDUC 低 = 0.181007389

λ AIDS 差-AGE50-EDUC 低 = -0.361991389

其趋势大约是年龄大的、受教育水平低的，预防 AIDS 知识掌握程度就较差。

Estimates for Parameters

AIDS

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
1	-.378234829	.07013	-5.39360	-.51568	-.24079
2	.3307195684	.06115	5.40864	.21087	.45057

AIDS BY AGE

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
3	.5610569048	.14377	3.90234	.27926	.84286
4	.4747703448	.10184	4.66178	.27516	.67438

5	-.317497183	.10192	-3.11520	-.51726	-.11774
6	.0139480312	.13188	.10577	-.24453	.27243
7	.0027846286	.09475	.02939	-.18292	.18849
8	.0999432599	.08316	1.20189	-.06304	.26293

AIDS BY EDUC

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
9	1.097077448	.09401	11.67000	.91282	1.28133
10	-.186500026	.08160	-2.28558	-.34643	-.02657
11	-.018774593	.09212	-.20382	-.19932	.16177
12	.0930200827	.06828	1.36233	-.04081	.22685

AIDS BY AGE BY EDUC

Parameter	Coeff.	Std. Err.	Z-Value	Lower 95 CI	Upper 95 CI
13	-.165975990	.18396	-.90225	-.52653	.19458
14	.1990147721	.15635	1.27292	-.10742	.50545
15	-.400615540	.15180	-2.63910	-.69814	-.10309
16	-.254356695	.11603	-2.19225	-.48177	-.02695
17	.0248775102	.13931	.17858	-.24817	.29793
18	.3092579898	.12025	2.57183	.07357	.54494
19	.5096508480	.17810	2.86156	.16057	.85873
20	.1850931544	.14164	1.30682	-.09251	.46270
21	.1964166680	.15202	1.29206	-.10154	.49437
22	-.088243218	.10424	-.84656	-.29255	.11606
23	.0455872550	.12925	.35269	-.20775	.29893
24	-.198715366	.09651	-2.05897	-.38788	-.00955

第十章 分类分析

人们认识事物时往往先把被认识的对象进行分类，以便寻找其中同与不同的特征，因而分类学是人们认识世界的基础科学。在医学实践中也经常需要做分类的工作，如根据病人的一系列症状、体征和生化检查的结果，判断病人所患疾病的类型；或对一系列检查方法及其结果，将之划分成某几种方法适合用于甲类病的检查，另几种方法适合用于乙类病的检查；等等。统计学中常用的分类统计方法主要是聚类分析与判别分析。

聚类分析是直接比较各事物之间的性质，将性质相近的归为一类，将性质差别较大的归入不同的类。判别分析则先根据已知类别的事物的性质，利用某种技术建立函数式，然后对未知类别的新事物进行判断以将之归入已知的类别中。聚类分析与判别分析有很大的不同，聚类分析事先并不知道对象类别的面貌，甚至连共有几个类别也不确定；判别分析事先已知对象的类别和类别数，它正是从这样的情形下总结出分类方法，用于对新对象的分类。

第一节 K-Means Cluster 过程

10.1.1 主要功能

调用此过程可完成由用户指定类别数的大样本资料的逐步聚类分析。所谓逐步聚类分析就是先把被聚对象进行初始分类，然后逐步调整，得到最终分类。

10.1.2 实例操作

［例 10.1］为研究儿童生长发育的分期，调查 1253 名 1 月至 7 岁儿童的身高（cm）、体重（kg）、胸围（cm）和坐高（cm）资料。资料作如下整理：先把 1 月至 7 岁划成 19 个月份段，分月份算出各指标的平均值，将第 1 月的各指标平均值与出生时的各指标平均值比较，求出月平均增长率（%），然后第 2 月起的各月份指标平均值均与前一月比较，亦求出月平均增长率（%），结果见下表。欲将儿童生长发育分为四期，故指定聚类的类别数为 4，请通过聚类分析确定四个儿童生长发育期的起止区间。

月份	月平均增长率（%）			
	身高	体重	胸围	坐高
1	11.03	50.30	11.81	11.27
2	5.47	19.30	5.20	7.18
3	3.58	9.85	3.14	2.11
4	2.01	4.17	1.47	1.58
6	2.13	5.65	1.04	2.11
8	2.06	1.74	0.17	1.57
10	1.63	2.04	1.04	1.46
12	1.17	1.60	0.89	0.76
15	1.03	2.34	0.53	0.89
18	0.69	1.33	0.48	0.58
24	0.77	1.41	0.52	0.42
30	0.59	1.25	0.30	0.14
36	0.65	1.19	0.49	0.38
42	0.51	0.93	0.16	0.25
48	0.73	1.13	0.35	0.55
54	0.53	0.82	0.16	0.34
60	0.36	0.52	0.19	0.21
66	0.52	1.03	0.30	0.55
72	0.34	0.49	0.18	0.16

10.1.2.1 数据准备

激活数据管理窗口，定义变量名：虽然月份分组不作分析变量，但为了更直观地了解聚类结果，也将之输入数据库，其变量名为 month；身高、体重、胸围和坐高的变量名分别为 x1、x2、x3 和 x4，

输入原始数额。

10.1.2.2 统计分析

激活 Statistics 菜单选 Classify 中的 K-Means Cluster...项，弹出 K-Means Cluster Analysis 对话框（如图 10.1 示）。从对话框左侧的变量列表中选 x1、x2、x3、x4，点击 ➢ 钮使之进入 Variables 框；在 Number of Clusters（即聚类分析的类别数）处输入需要聚合的组数，本例为 4；在聚类方法上有两种：Iterate and classify 指先定初始类别中心点，而后按 K-means 算法作叠代分类，Classify only 指仅按初始类别中心点分类，本例选用前一方法。

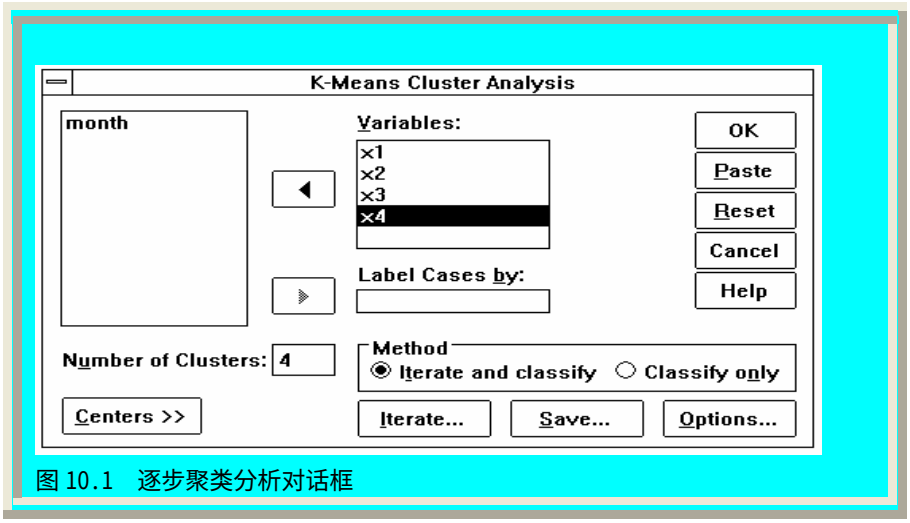


图 10.1 逐步聚类分析对话框

为在原始数据库中逐一显示分类结果，点击 Save...钮弹出 K-Means Cluster:Save New Variables 对话框，选择 Cluster membership 项，点击 Continue 钮返回 K-Means Cluster Analysis 对话框。

本例还要求对聚类结果进行方差分析，故点击 Options...钮弹出 K-Means Cluster:来 Options 对话框，在 Statistics 栏中选择 ANOVA table 项，点击 Continue 钮返回 K-Means Cluster Analysis 对话框，再点击 OK 钮即完成分析。

10.1.2.3 结果解释

在结果输出窗口中将看到如下统计数据：

首先系统根据用户的指定，按 4 类聚合确定初始聚类的各变量中心点，未经 K-means 算法叠代，其类别间距离并非最优；经叠代运算后类别间各变量中心值得到修正。

Initial Cluster Centers.

Cluster	X1	X2	X3	X4
1	11.0300	50.3000	11.8100	11.2700
2	5.4700	19.3000	5.2000	7.1800
3	3.5800	9.8500	3.1400	2.1100
4	.3400	.4900	.1800	.1600

Convergence achieved due to no or small distance change.
The maximum distance by which any center has changed is .0000
Current iteration is 2

Minimum distance between initial centers is 10.5200

Iteration	Change in Cluster Centers			
	1	2	3	4
1	.0000	.0000	2.46E+00	1.27E+00
2	.0000	.0000	.0000	.0000

Case listing of Cluster membership.

Case ID	Cluster	Distance
1	1	.000
2	2	.000
3	3	2.457
4	4	3.219
5	3	2.457
6	4	1.530
7	4	1.346
8	4	.515
9	4	.915
10	4	.266
11	4	.281
12	4	.668
13	4	.467
14	4	.844
15	4	.415
16	4	.873
17	4	1.215
18	4	.619
19	4	1.269

Final Cluster Centers.

Cluster	X1	X2	X3	X4
1	11.0300	50.3000	11.8100	11.2700
2	5.4700	19.3000	5.2000	7.1800
3	2.8550	7.7500	2.0900	2.1100
4	.9060	1.4660	.4820	.6560

之后对聚类结果的类别间距离进行方差分析，方差分析表明，类别间距离差异的概率值均<0.001，即聚类效果好。这样，原有 19 类（即原有的 19 个月份分组）聚合成 4 类，第一类含原有 1 类，第二类含原有 1 类，第三类含原有 2 类，第四类含原有 15 类。具体结果系统以变量名 QCL_1 存于原始数据库中。

Distances between Final Cluster Centers.

Cluster	1	2	3	4
1	.0000			
2	32.4397	.0000		
3	45.3400	13.2521	.0000	
4	52.2325	20.0924	6.9273	.0000

Analysis of Variance.

Variable	Cluster MS	DF	Error MS	DF	F	Prob
X1	37.5806	3	.369	15.0	101.7853	.000
X2	817.1164	3	1.354	15.0	603.2588	.000
X3	45.4089	3	.281	15.0	161.1145	.000
X4	46.0994	3	.235	15.0	195.4933	.000

Number of Cases in each Cluster.

Cluster	unweighted cases	weighted cases
1	1.0	1.0
2	1.0	1.0
3	2.0	2.0
4	15.0	15.0
Missing	0	
Valid cases	19.0	19.0

Variable Saved into Working File.

QCL_1 (Cluster Number)

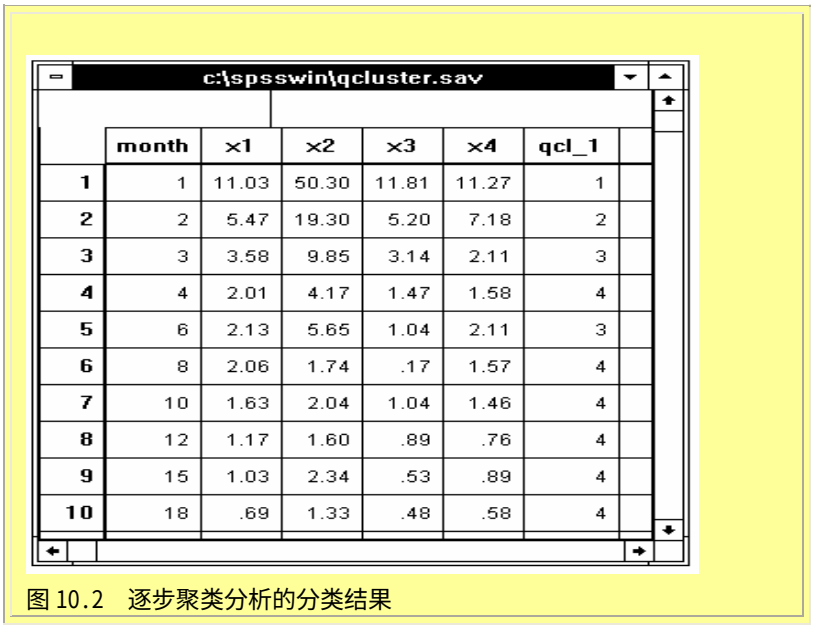
在原始数据库（图 10.2）中，我们可清楚地看到聚类结果；参照专业知识，将儿童生长发育分期定为：

第一期，出生后至满月，增长率最高；

第二期，第 2 个月起至第 3 个月，增长率次之；

第三期，第 3 个月起至第 8 个月，增长率减缓；

第四期，第 8 个月后，增长率显著减缓。



	month	x1	x2	x3	x4	qcl_1
1	1	11.03	50.30	11.81	11.27	1
2	2	5.47	19.30	5.20	7.18	2
3	3	3.58	9.85	3.14	2.11	3
4	4	2.01	4.17	1.47	1.58	4
5	6	2.13	5.65	1.04	2.11	3
6	8	2.06	1.74	.17	1.57	4
7	10	1.63	2.04	1.04	1.46	4
8	12	1.17	1.60	.89	.76	4
9	15	1.03	2.34	.53	.89	4
10	18	.69	1.33	.48	.58	4

图 10.2　逐步聚类分析的分类结果

第二节 Hierarchical Cluster 过程

10.2.1 主要功能

调用此过程可完成系统聚类分析。在系统聚类分析中，用户事先无法确定类别数，系统将所有例数均调入内存，且可执行不同的聚类算法。系统聚类分析有两种形式，一是对研究对象本身进行分类，称为 Q 型举类；另一是对研究对象的观察指标进行分类，称为 R 型聚类。

10.2.2 实例操作

［例 10.2］29 名儿童的血红蛋白（g/100ml）与微量元素（μg/100ml）测定结果如下表。由于微量元素的测定成本高、耗时长，故希望通过聚类分析（即 R 型指标聚类）筛选代表性指标，以便更经济快捷地评价儿童的营养状态。

编号 N0.	钙 X1	镁 X2	铁 X3	锰 X4	铜 X5	血红蛋白 X6
1	54.89	30.86	448.70	0.012	1.010	13.50
2	72.49	42.61	467.30	0.008	1.640	13.00
3	53.81	52.86	425.61	0.004	1.220	13.75
4	64.74	39.18	469.80	0.005	1.220	14.00
5	58.80	37.67	456.55	0.012	1.010	14.25
6	43.67	26.18	395.78	0.001	0.594	12.75
7	54.89	30.86	448.70	0.012	1.010	12.50
8	86.12	43.79	440.13	0.017	1.770	12.25
9	60.35	38.20	394.40	0.001	1.140	12.00
10	54.04	34.23	405.60	0.008	1.300	11.75
11	61.23	37.35	446.00	0.022	1.380	11.50
12	60.17	33.67	383.20	0.001	0.914	11.25

13	69.69	40.01	416.70	0.012	1.350	11.00
14	72.28	40.12	430.80	0.000	1.200	10.75
15	55.13	33.02	445.80	0.012	0.918	10.50
16	70.08	36.81	409.80	0.012	1.190	10.25
17	63.05	35.07	384.10	0.000	0.853	10.00
18	48.75	30.53	342.90	0.018	0.924	9.75
19	52.28	27.14	326.29	0.004	0.817	9.50
20	52.21	36.18	388.54	0.024	1.020	9.25
21	49.71	25.43	331.10	0.012	0.897	9.00
22	61.02	29.27	258.94	0.016	1.190	8.75
23	53.68	28.79	292.80	0.048	1.320	8.50
24	50.22	29.17	292.60	0.006	1.040	8.25
25	65.34	29.99	312.80	0.006	1.030	8.00
26	56.39	29.29	283.00	0.016	1.350	7.80
27	66.12	31.93	344.20	0.000	0.689	7.50
28	73.89	32.94	312.50	0.064	1.150	7.25
29	47.31	28.55	294.70	0.005	0.838	7.00

10.2.2.1 数据准备

激活数据管理窗口，定义变量名：钙、镁、铁、锰、铜和血红蛋白的变量名分别为 x1、x2、x3、x4、x5、x6，之后输入原始数据。

10.2.2.2 统计分析

激活 Statistics 菜单选 Classify 中的 Hierarchical Cluster...项，弹出 Hierarchical Cluster Analysis 对话框（图 10.3）。从对话框左侧的变量列表中选 x1、x2、x3、x4、x5、x6，点击 ➢ 钮使之进入 Variable(s)框；在 Cluster 处选择聚类类型，其中 Cases 表示观察对象聚类，Variables 表示变量聚类，本例选择 Variables。

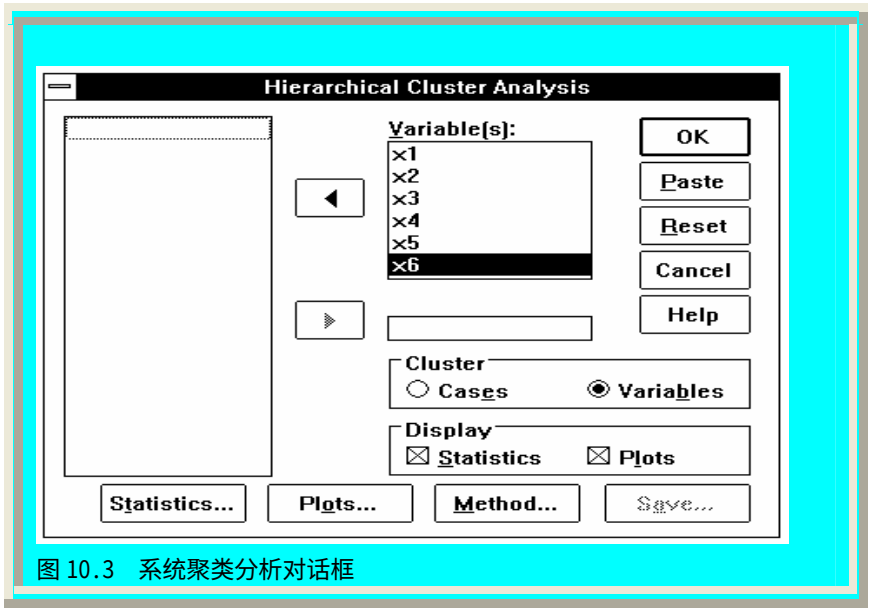


图 10.3 系统聚类分析对话框

点击 Statistics...钮，弹出 Hierarchical Cluster Analysis: Statistics 对话框，选择 Distance matrix，要求显示距离矩阵，点击 Continue 钮返回 Hierarchical Cluster Analysis 对话框（图 10.4）。

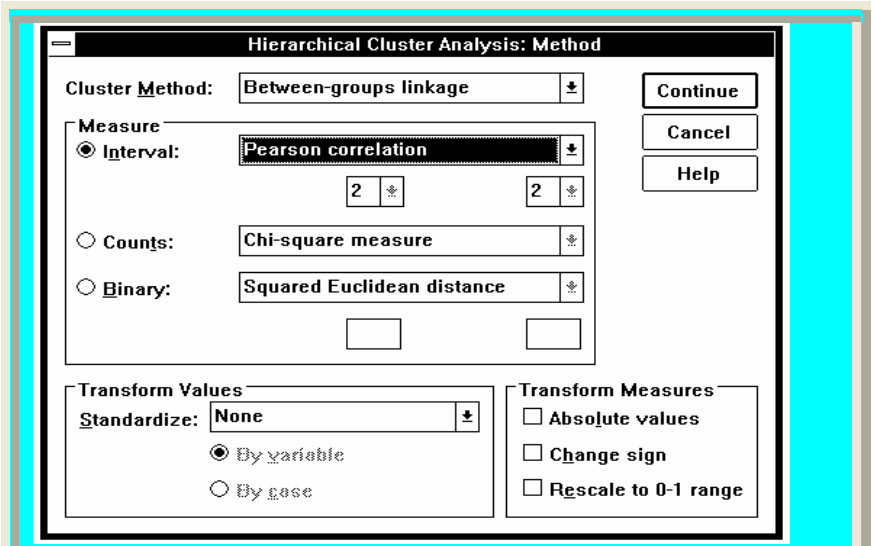


图 10.4　系统聚类方法选择对话框

本例要求系统输出聚类结果的树状关系图，故点击 Plots...钮弹出 Hierarchical Cluster Analysis:Plots 对话框，选择 Dendrogram 项，点击 Continue 钮返回 Hierarchical Cluster Analysis 对话框。

点击 Method...钮弹出 Hierarchical Cluster Analysis:Method 对话框，系统提供 7 种聚类方法供用户选择：

Between-groups linkage：类间平均链锁法；

Within-groups linkage：类内平均链锁法；

Nearest neighbor：最近邻居法；

Furthest neighbor：最远邻居法；

Centroid clustering：重心法，应与欧氏距离平方法一起使用；

Median clustering：中间距离法，应与欧氏距离平方法一起使用；

Ward's method：离差平方和法，应与欧氏距离平方法一起使用。

本例选择类间平均链锁法（系统默认方法）。在选择距离测量技术上，系统提供 8 种形式供用户选择：

Euclidean distance：Euclidean 距离，即两观察单位间的距离为其值差的平方和的平方根，该技术用于 Q 型聚类；

Squared Euclidean distance：Euclidean 距离平方，即两观察单位间的距离为其值差的平方和，该技术用于 Q 型聚类；

Cosine：变量矢量的余弦，这是模型相似性的度量；

Pearson correlation：相关系数距离，适用于 R 型聚类；

Chebychev：Chebychev 距离，即两观察单位间的距离为其任意变量的最大绝对差值，该技术用于 Q 型聚类；

Block：City-Block 或 Manhattan 距离，即两观察单位间的距离为其值差的绝对值和，适用于 Q 型聚类；

Minkowski：距离是一个绝对幂的度量，即变量绝对值的第 p 次幂之和的平方根；p 由用户指定

Customized：距离是一个绝对幂的度量，即变量绝对值的第 p 次幂之和的第 r 次根，p 与 r 由用户指定。

本例选用 Pearson correlation，点击 Continue 钮返回 Hierarchical Cluster Analysis 对话框，再点击 OK 钮即完成分析。

10.2.2.3 结果解释

在结果输出窗口中将看到如下统计数据：

共 29 例样本进入聚类分析，采用相关系数测量技术。先显示各变量间的相关系数，这对于后面选择典型变量是十分有用的。然后显示类间平均链锁法的合并进程，即第一步，X3 与 X6 被合并，它们之间的相关系数最大，为 0.863431；第二步，X1 与 X5 合并，其间相关系数为 0.624839；第三步，X2 与第一步的合并项被合并，它们之间的相关系数为 0.602099；第四步，它们与第二步的合并项再合并，其间相关系数为 0.338335；第五步，与最后一个变量 X4 合并，这个相关系数最小，为 -0.054485。

Data Information

29 unweighted cases accepted.

0 cases rejected because of missing value.

Correlation measure used.

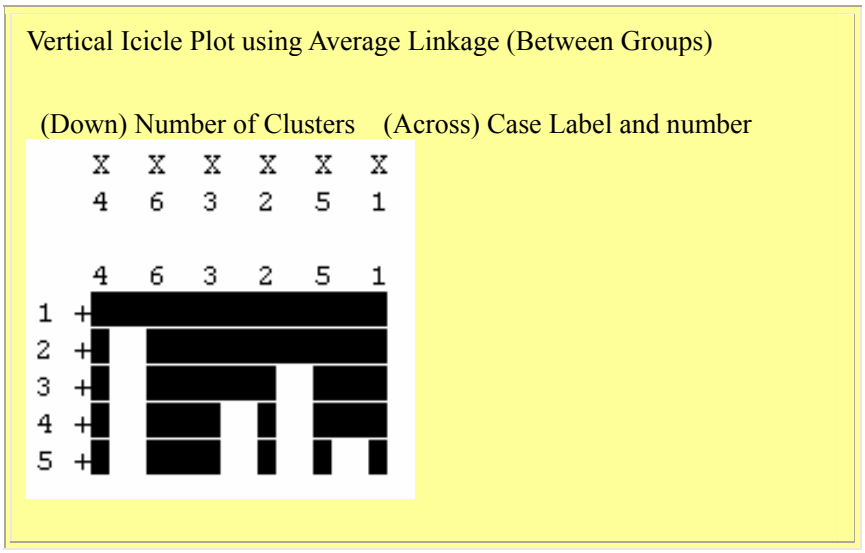
Correlation Similarity Coefficient Matrix

Variable	X1	X2	X3	X4	X5
X2	.5379				
X3	.2995	.6349			
X4	.1480	-.1212	-.2706		
X5	.6248	.5820	.2653	.2939	
X6	.0972	.5693	.8634	-.3226	.2481

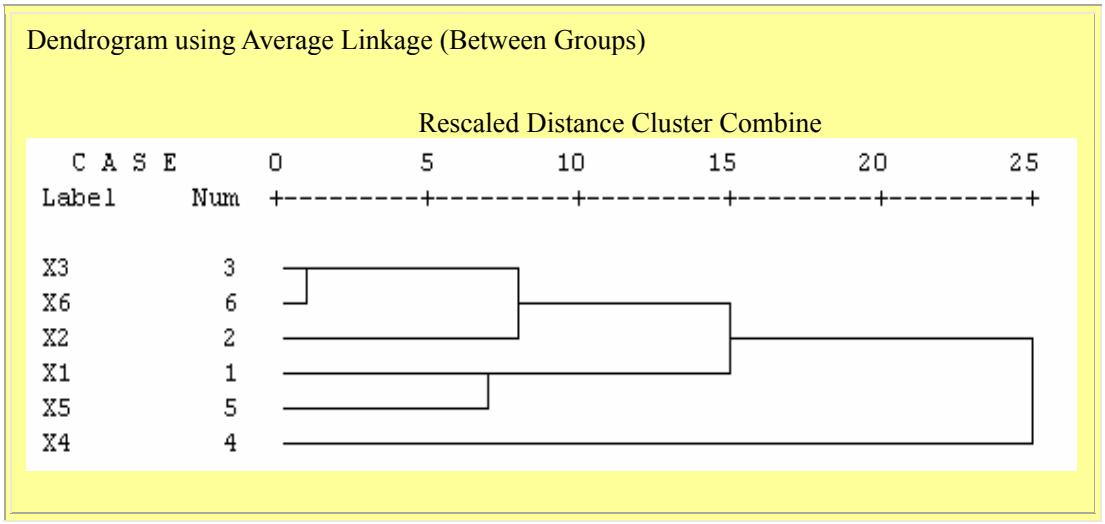
Agglomeration Schedule using Average Linkage (Between Groups)

	Clusters	Combined		Stage Cluster	1st Appears	Next
Stage	Cluster 1	Cluster 2	Coefficient	Cluster 1	Cluster 2	Stage
1	3	6	.863431	0	0	3
2	1	5	.624839	0	0	4
3	2	3	.602099	0	1	4
4	1	2	.338335	2	3	5
5	1	4	-.054485	4	0	0

按类间平均链锁法，变量合并过程的冰柱图如下。先是 X3 与 X6 合并，接着 X1 与 X5 合并，然后 X3、X6 与 X2 合并，接着再与 X1、X5 合并，最后加上 X4，六个变量全部合并。



下面用更为直观的聚类树状关系图表示，即 X1、X2、X3、X5、X6 先聚合后与 X4 再聚合。这表明，在评价儿童营养状态时，可在微量元素钙、镁、铁、铜和血红蛋白 5 个指标中选择一个，再加上微量元素锰即可，其效果与六个指标都用是基本等价的，但更经济更迅速。



微量元素钙、镁、铁、铜和血红蛋白聚合成一类，在这 5 个指标中如何选择一个典型指标呢？先按下式计算类中每一变量与其余变量的相关指数（即相关系数的平方）的均值，而后把该值最大的变量作为典型指标。

$$\overline{R^2} = \frac{\sum r^2}{m-1}$$ （式中m为类中变量个数）

本例相关指数的均值依次为：

$$\overline{R^2_{X1}} = \frac{0.5379^2 + 0.2995^2 + 0.6248^2 + 0.0972^2}{5-1} = 0.1947$$

$$\overline{R^2_{X2}} = \frac{0.5379^2 + 0.6349^2 + 0.5820^2 + 0.5693^2}{5-1} = 0.3388$$

$$\overline{R^2_{X3}} = \frac{0.2995^2 + 0.6349^2 + 0.2653^2 + 0.8634^2}{5-1} = 0.3272$$

$$\overline{R^2_{X5}} = \frac{0.6284^2 + 0.5820^2 + 0.2653^2 + 0.2481^2}{5-1} = 0.2164$$

$$\overline{R^2_{X6}} = \frac{0.0972^2 + 0.5693^2 + 0.8634^2 + 0.2481^2}{5-1} = 0.2851$$

故选择镁（变量 X2）典型指标。

第三节 Discriminant 过程

10.3.1 主要功能

调用此过程可完成判别分析。判别分析目前在医学中得以广泛应用，不仅在于它所建立的判别式可用于临床辅助诊断，而且判别分析可分析出各种因素对特定结果的作用力大小，故亦可用于病因学或疾病预后的推测。

10.3.2 实例操作

［例 10.3］为研究舒张期血压和血浆胆固醇对冠心病的作用，某医师测定了 50-59 岁冠心病人 15 例和正常人 16 例的舒张压和胆固醇指标，结果如下，试作判别分析，建立判别函数以便在临床中用于筛选冠心病人。

编号	冠心病人组		编号	正常人组	
	舒张压 kPa	胆固醇 mmol/L		舒张压 kPa	胆固醇 mmol/L
	x1	x2		x1	x2
1	9.86	5.18	1	10.66	2.07
2	13.33	3.73	2	12.53	4.45
3	14.66	3.89	3	13.33	3.06
4	9.33	7.10	4	9.33	3.94
5	12.80	5.49	5	10.66	4.45
6	10.66	4.09	6	10.66	4.92
7	10.66	4.45	7	9.33	3.68
8	13.33	3.63	8	10.66	2.77
9	13.33	5.96	9	10.66	3.21
10	13.33	5.70	10	10.66	5.02
11	12.00	6.19	11	10.40	3.94
12	14.66	4.01	12	9.33	4.92
13	13.33	4.01	13	10.66	2.69
14	12.80	3.63	14	10.66	2.43
15	13.33	5.96	15	11.20	3.42
			16	9.33	3.63

10.3.2.1 数据准备

激活数据管理窗口，舒张压、胆固醇的变量名分别以 x1、x2 表示，将冠心病人资料和正常人资料合并，一同输入。而后，再定义一变量名为 result，用于区分冠心病人资料和正常人资料，即冠心病人资料的 result 值均为 1，正常人资料的 result 值均为 2。

10.3.2.2 统计分析

激活 Statistics 菜单选 Classify 中的 Discriminant...项，弹出 Discriminant Analysis 对话框（图 10.5）。从对话框左侧的变量列表中选 result，点击 ➢ 钮使之进入 Grouping Variable 框，并点击 Define Range... 钮，在弹出的 Discriminant Analysis:Define Range 对话框中，定义判别原始数据的类别区间，本例为两类，故在 Minimum 处输入 1、在 Maximum 处输入 2，点击 Continue 钮返回 Discriminant Analysis 对话框。再从对话框左侧的变量列表中选 x1、x2，点击 ➢ 钮使之进入 Independents 框，作为判别分析的基础数据变量。

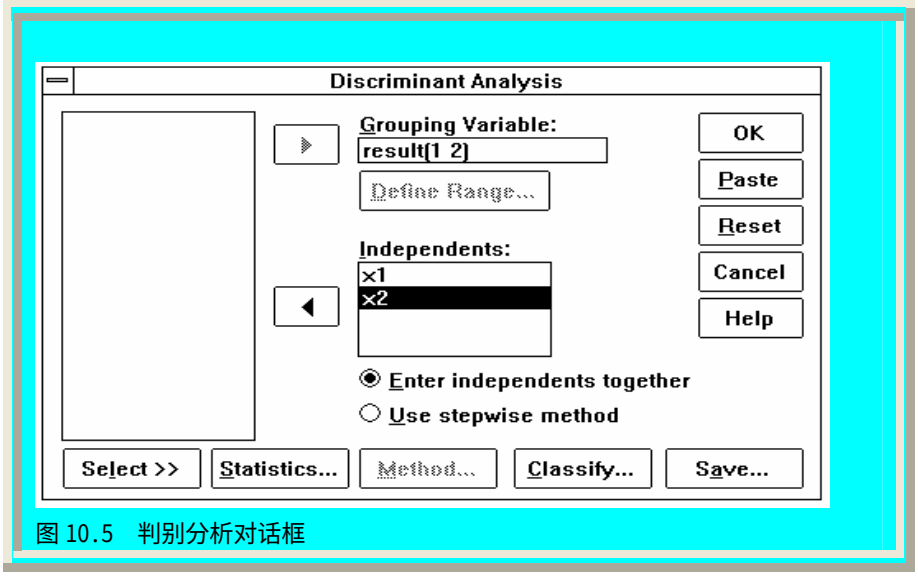


图 10.5 判别分析对话框

系统提供两类判别方式供选择，一是 Enter Independent together，即判别的原始变量全部进入判别方程；另一是 Use stepwise method，即采用逐步的方法选择变量进入方程。对于后者，系统有 5 种逐步选择方式：

Wilks' lambda：按统计量 Wilks λ 最小值选择变量；

Unexplained variance：按所有组方差之和的最小值选择变量；

Mahalanobis' distance：按相邻两组的最大 Mahalanobis 距离选择变量；

Smallest F ratio：按组间最小 F 值比的最大值选择变量；

Rao's V：按统计量 Rao V 最大值选择变量。

本例由于变量数仅为 2 个，倾向让两个变量均进入方程，故选用 Enter Independent together 判别方式。

点击 Statistics...钮，弹出 Discriminant Analysis: Statistics 对话框，在 Descriptive 栏中选 Means 项，要求对各组的各变量作均数与标准差的描述；在 Function Coefficients 栏中选 Unstandardized 项，要求显示判别方程的非标准化系数。之后，点击 Continue 钮返回 Discriminant Analysis 对话框。

点击 Classify...钮，弹出 Discriminant Analysis: Classification 对话框，在 Plot 栏选 Combined groups 项，要求作合并的判别结果分布图；在 Display 栏选 Results for each case 项，要求对原始资料根据建立的判别方程作逐一回代重判别，同时选 Summary table 项，要求对这种回代判别结果进行总结评价。之后，点击 Continue 钮返回 Discriminant Analysis 对话框。

点击 Save...钮，弹出 Discriminant Analysis: Save New Variables 对话框，选 Predicted group

membership 项要求将回代判别的结果存入原始数据库中。点击 Continue 钮返回 Discriminant Analysis 对话框，之后再点击 OK 钮即完成分析。

10.3.2.3 结果解释

在结果输出窗口中将看到如下统计数据：

首先，系统提示将判别回代的结果以变量名 DIS_1 存于原始数据库中。

接着系统显示数据按变量 RESULT 分组，共 31 个样本作为判别基础数据进入分析，其中第一组 15 例，第二组 16 例。同时，分组给出各变量的均数（means）与标准差（standard deviations）。

Following variables will be created upon successful completion of the procedure:

Name	Label
DIS_1	--- Predicted group for analysis 1

On groups defined by RESULT

31 (Unweighted) cases were processed.
0 of these were excluded from the analysis.
31 (Unweighted) cases will be used in the analysis.

Number of cases by group

	Number of cases		
RESULT	Unweighted	Weighted	Label
1	15	15.0	
2	16	16.0	
Total	31	31.0	

Group means

RESULT	X1	X2
1	12.49400	4.86800
2	10.62875	3.66250
Total	11.53129	4.24581

Group standard deviations

RESULT	X1	X2
1	1.64064	1.12948
2	1.09681	.92467
Total	1.65996	1.18231

On groups defined by RESULT
Analysis number 1
Direct method: all variables passing the tolerance test are entered.
Minimum tolerance level..... .00100

Canonical Discriminant Functions

```
        Maximum number of functions.....          1
        Minimum cumulative percent of variance...  100.00
        Maximum significance of Wilks' Lambda....  1.0000

Prior probability for each group is   .50000
```

下面为典型判别方程的方差分析结果，其特征值（Eigenvalue）即组间平方和与组内平方和之比为 1.2392，典型相关系数（Canonical Corr）为 0.7439，Wilks λ 值为 0.446597，经 χ^2 检验，χ^2 为 22.571，P<0.0001。

用户可通过判别方程的标准化系数，确定各变量对结果的作用大小。如本例舒张压（X1）的标准化系数（0.88431）大于胆固醇（X2）的标准化系数（0.82306），因而舒张压对冠心病的影响作用大于胆固醇。考察变量作用大小的另一途径是使用变量与函数间的相关系数，本例显示 X1 的变量与函数间的相关系数为 0.62454，X2 为 0.54396，同样表明舒张压对冠心病的影响作用大于胆固醇。

根据系统显示的非标准化判别方程系数，得到判别方程为：

D = 0.6379195X1 + 0.8001452X2 - 10.7532968

依此方程，病人组的中心得分点为 1.11198，正常人组的中心得分点为-1.04248。本例为二类判别，二类判别以 0 为分界点，若将某人的舒张压和胆固醇值代入判别方程，求出的判别分>0 的为冠心病人，判别分<0 的为正常人。

```
Canonical Discriminant Functions

                    Pct of    Cum   Canonical      After    Wilks'
 Fcn   Eigenvalue   Variance   Pct     Corr          Fcn    Lambda   Chi-square   df    Sig
                                                  :   0    .446597     22.571     2   .0000
  1*      1.2392     100.00  100.00   .7439       :

   * Marks the 1 canonical discriminant functions remaining in the analysis.

Standardized canonical discriminant function coefficients
                  Func  1
X1                 .88431
X2                 .82306

Structure matrix:
Pooled within-groups correlations between discriminating variables
                                          and canonical discriminant functions
(Variables ordered by size of correlation within function)
                  Func  1
X1                 .62454
X2                 .54396

Unstandardized canonical discriminant function coefficients
```

	Func 1
X1	.6379195
X2	.8001452
(Constant)	-10.7532968

Canonical discriminant functions evaluated at group means (group centroids)

Group	Func 1
1	1.11198
2	-1.04248

下面为原始数据逐一回代的判别结果显示。其中病人组有 3 人被错判（编号为 1、6、7，打**者），正常人组有 3 人被错判（编号为 17、18、25，打**者）。接着用分布图的形式显示判别结果，图中 1 代表病人，2 代表正常人，每四个 1 或 2 代表一个人；图中可见，有三个病人跨过 0 界进入负值区，被错判为正常人，也有三个正常人跨过 0 界进入正值区，被错判为病人。最后系统对回代判别的情况作评价，即病人组判别正确率为 80.0%，正常人组为 81.3%，总判别正确率为 80.65%。

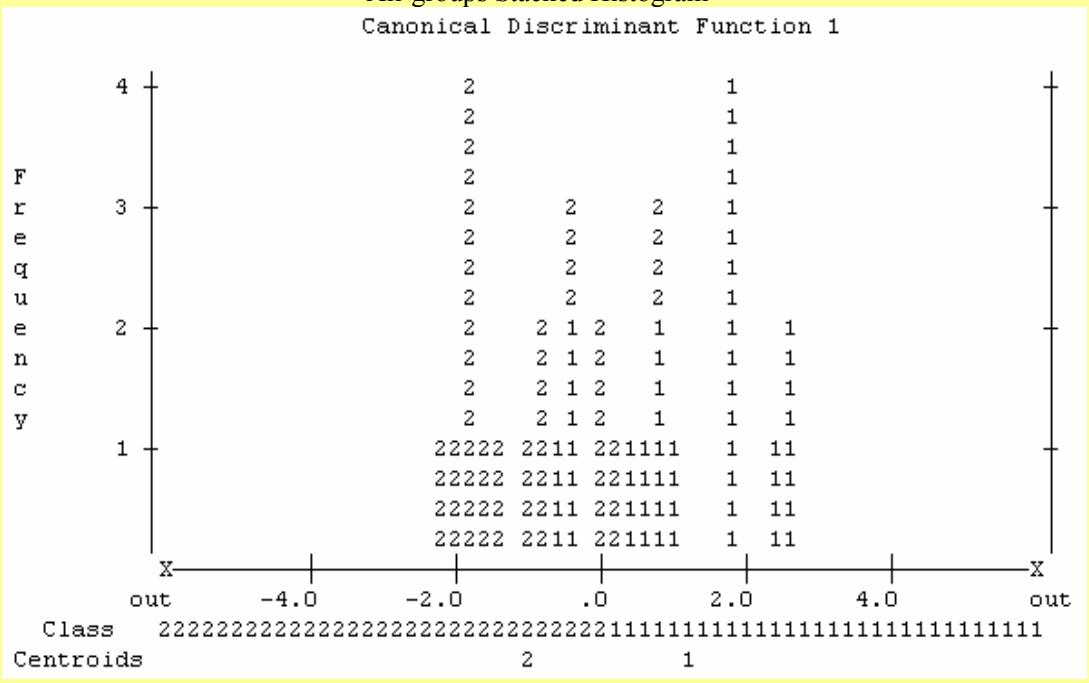
Case	Mis		Actual	Highest Probability			2nd Highest		Discrim
Number	Val	Sel	Group	Group	P(D/G)	P(G/D)	Group	P(G/D)	Scores
1			1 **	2	.4692	.6817	1	.3183	-.3187
2			1	1	.7060	.8188	2	.1812	.7347
3			1	1	.5490	.9737	2	.0263	1.7112
4			1	1	.8162	.8606	2	.1394	.8795
5			1	1	.4884	.9784	2	.0216	1.8049
6			1 **	2	.7174	.8236	1	.1764	-.6805
7			1 **	2	.5157	.7151	1	.2849	-.3924
8			1	1	.6475	.7918	2	.2082	.6547
9			1	1	.1594	.9953	2	.0047	2.5190
10			1	1	.2305	.9926	2	.0074	2.3110
11			1	1	.4577	.9806	2	.0194	1.8546
12			1	1	.4869	.9785	2	.0215	1.8072
13			1	1	.8782	.8798	2	.1202	.9588
14			1	1	.4264	.6473	2	.3527	.3166
15			1	1	.1594	.9953	2	.0047	2.5190
16			2	2	.2097	.9935	1	.0065	-2.2968
17			2 **	1	.7554	.8389	2	.1611	.8005
18			2 **	1	.3611	.5874	2	.4126	.1986
19			2	2	.5442	.9741	1	.0259	-1.6489
20			2	2	.5157	.7151	1	.2849	-.3924
21			2	2	.3048	.5275	1	.4725	-.0164
22			2	2	.4154	.9833	1	.0167	-1.8570
23			2	2	.4876	.9785	1	.0215	-1.7367
24			2	2	.7323	.9551	1	.0449	-1.3846

25	2 **	1	.2945	.5156	2	.4844	.0637
26	2	2	.9393	.8963	1	.1037	-.9664
27	2	2	.8590	.8741	1	.1259	-.8648
28	2	2	.4483	.9812	1	.0188	-1.8007
29	2	2	.3339	.9879	1	.0121	-2.0087
30	2	2	.8647	.8759	1	.1241	-.8721
31	2	2	.3928	.9847	1	.0153	-1.8970

Symbols used in plots

Symbol	Group	Label
1	1	
2	2	

All-groups Stacked Histogram



Classification results -

Actual Group		No. of Cases	Predicted Group Membership 1	2
Group	1	15	12	3
			80.0%	20.0%
Group	2	16	3	13
			18.8%	81.3%

Percent of "grouped" cases correctly classified: 80.65%

Classification processing summary

31 (Unweighted) cases were processed.

0 cases were excluded for missing or out-of-range group codes.
0 cases had at least one missing discriminating variable.
31 (Unweighted) cases were used for printed output.
31 cases were written into the working file.

系统将判别回代的结果以 dis_1 为变量名存入原始数据库中，如下图所示。用户可通过翻动原始数据库详细查阅。

c:\spsswin\discri.sav

	x1	x2	result	dis_1
2	13.33	3.73	1	1
3	14.66	3.89	1	1
4	9.33	7.10	1	1
5	12.80	5.49	1	1
6	10.66	4.09	1	2
7	10.66	4.45	1	2
8	13.33	3.63	1	1
9	13.33	5.96	1	1
10	13.33	5.70	1	1
11	12.00	6.19	1	1

图 10.6　原始数据及判别结果

第十一章　因子分析

11.1　主要功能

多元分析处理的是多指标的问题。由于指标太多，使得分析的复杂性增加。观察指标的增加本来是为了使研究过程趋于完整，但反过来说，为使研究结果清晰明了而一味增加观察指标又让人陷入混乱不清。由于在实际工作中，指标间经常具备一定的相关性，故人们希望用较少的指标代替原来较多的指标，但依然能反映原有的全部信息，于是就产生了主成分分析、对应分析、典型相关分析和因子分析等方法。

调用 Data Reduction 菜单的 Factor 过程命令项，可对多指标或多因素资料进行因子分析。因子分析的基本目的就是用少数几个因子去描述许多指标或因素之间的联系，即将相关比较密切的几个变量归在同一类中，每一类变量就成为一个因子（之所以称其为因子，是因为它是不可观测的，即不是具体的变量，这与上一章的聚类分析不同），以较少的几个因子反映原资料的大部分信息。

11.2 实例操作

[例 11-1]下表资料为 25 名健康人的 7 项生化检验结果,7 项生化检验指标依次命名为 X1 至 X7,请对该资料进行因子分析。

X1	X2	X3	X4	X5	X6	X7
3.76	3.66	0.54	5.28	9.77	13.74	4.78
8.59	4.99	1.34	10.02	7.50	10.16	2.13
6.22	6.14	4.52	9.84	2.17	2.73	1.09
7.57	7.28	7.07	12.66	1.79	2.10	0.82
9.03	7.08	2.59	11.76	4.54	6.22	1.28
5.51	3.98	1.30	6.92	5.33	7.30	2.40
3.27	0.62	0.44	3.36	7.63	8.84	8.39
8.74	7.00	3.31	11.68	3.53	4.76	1.12
9.64	9.49	1.03	13.57	13.13	18.52	2.35
9.73	1.33	1.00	9.87	9.87	11.06	3.70
8.59	2.98	1.17	9.17	7.85	9.91	2.62
7.12	5.49	3.68	9.72	2.64	3.43	1.19
4.69	3.01	2.17	5.98	2.76	3.55	2.01
5.51	1.34	1.27	5.81	4.57	5.38	3.43
1.66	1.61	1.57	2.80	1.78	2.09	3.72
5.90	5.76	1.55	8.84	5.40	7.50	1.97
9.84	9.27	1.51	13.60	9.02	12.67	1.75
8.39	4.92	2.54	10.05	3.96	5.24	1.43
4.94	4.38	1.03	6.68	6.49	9.06	2.81
7.23	2.30	1.77	7.79	4.39	5.37	2.27
9.46	7.31	1.04	12.00	11.58	16.18	2.42
9.55	5.35	4.25	11.74	2.77	3.51	1.05
4.94	4.52	4.50	8.07	1.79	2.10	1.29
8.21	3.08	2.42	9.10	3.75	4.66	1.72
9.41	6.44	5.11	12.50	2.45	3.10	0.91

11.2.1 数据准备

激活数据管理窗口，定义变量名：分别为 X1、X2、X3、X4、X5、X6、X7，按顺序输入相应数值，建立数据库，结果见图 11.1。

	x1	x2	x3	x4	x5	x6	x7
1	3.76	3.66	.54	5.28	9.77	13.74	4.78
2	8.59	4.99	1.34	10.02	7.50	10.16	2.13
3	6.22	6.14	4.52	9.84	2.17	2.73	1.09
4	7.57	7.28	7.07	12.66	1.79	2.10	.82
5	9.03	7.08	2.59	11.76	4.54	6.22	1.28
6	5.51	3.98	1.30	6.92	5.33	7.30	2.40
7	3.27	.62	.44	3.36	7.63	8.84	8.39
8	8.74	7.00	3.31	11.68	3.53	4.76	1.12
9	9.64	9.49	1.03	13.57	13.13	18.52	2.35

图 11.1　原始数据的输入

11.2.2　统计分析

激活 Statistics 菜单选 Data Reduction 的 Factor...命令项，弹出 Factor Analysis 对话框（图 11.2）。在对话框左侧的变量列表中选变量 X1 至 X7，点击 ➢ 钮使之进入 Variables 框。

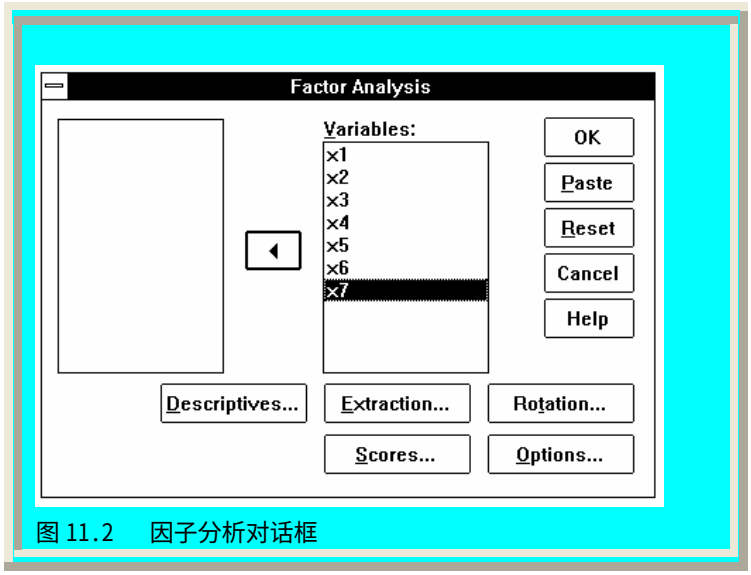


图 11.2　因子分析对话框

点击 Descriptives...钮，弹出 Factor Analysis:Descriptives 对话框（图 11.3），在 Statistics 中选 Univariate descriptives 项要求输出各变量的均数与标准差，在 Correlation Matrix 栏内选 Coefficients 项要求计算相关系数矩阵，并选 KMO and Bartlett's test of sphericity 项，要求对相关系数矩阵进行统计学检验。点击 Continue 钮返回 Factor Analysis 对话框。

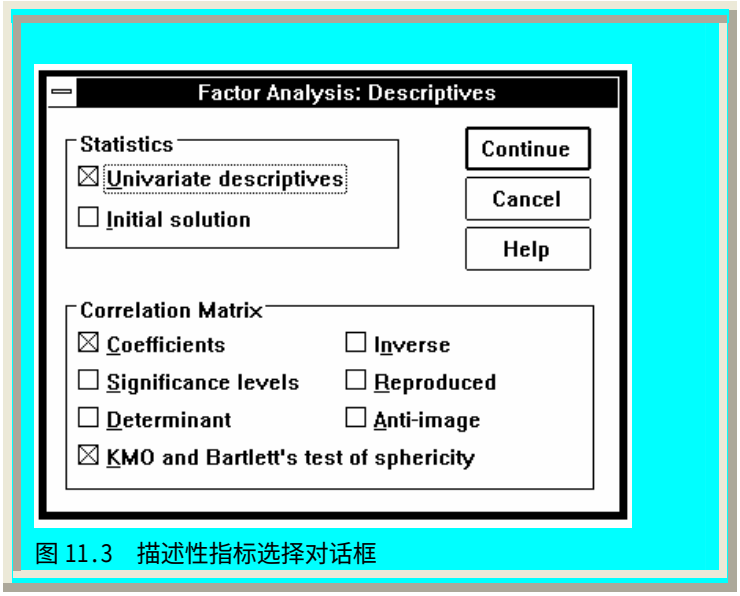


图 11.3　描述性指标选择对话框

点击 Extraction...钮，弹出 Factor Analysis: Extraction 对话框（图 11.4），系统提供如下因子提取方法：

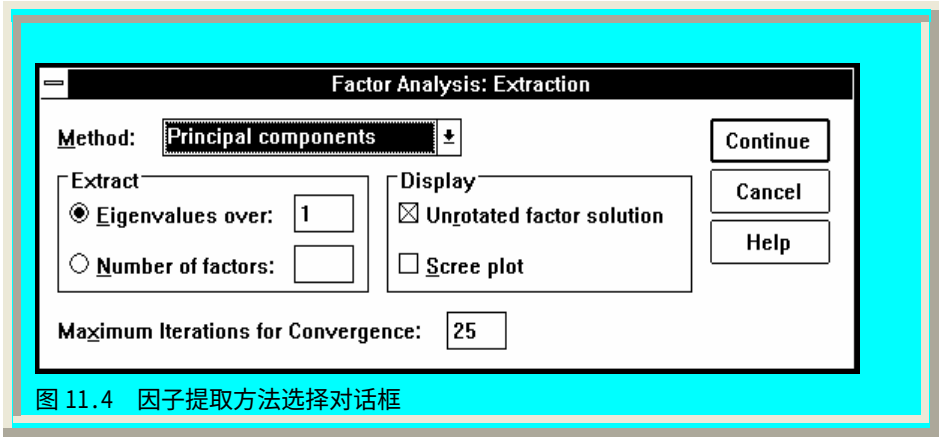


图 11.4　因子提取方法选择对话框

Principal components：主成分分析法；
Unweighted least squares：未加权最小平方法；
Generalized least squares：综合最小平方法；
Maximum likelihood：极大似然估计法；
Principal axis factoring：主轴因子法；
Alpha factoring：α 因子法；
Image factoring：多元回归法。
本例选用 Principal components 方法，之后点击 Continue 钮返回 Factor Analysis 对话框。
点击 Rotation...钮，弹出 Factor Analysis:Rotation 对话框（图 11.5），系统有 5 种因子旋转方法可选：

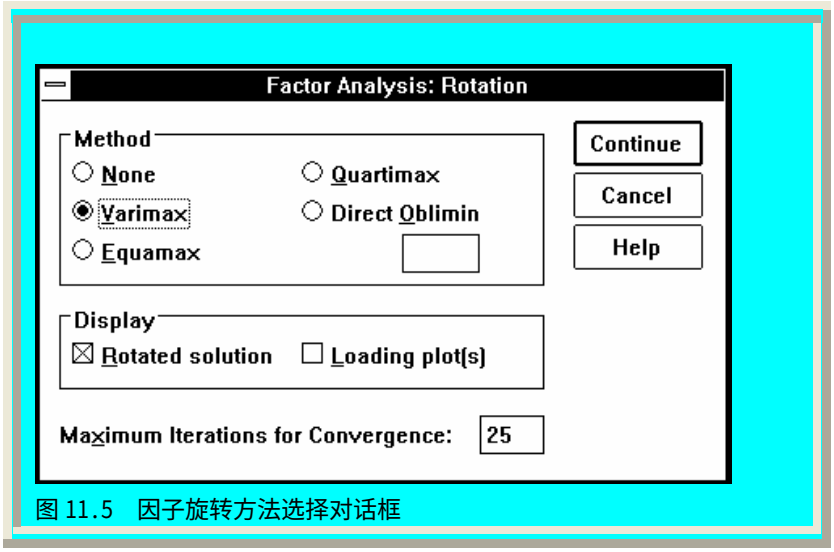


图 11.5　因子旋转方法选择对话框

None：不作因子旋转；

Varimax：正交旋转；

Equamax：全体旋转，对变量和因子均作旋转；

Quartimax：四分旋转，对变量作旋转；

Direct Oblimin：斜交旋转。

旋转的目的是为了获得简单结构，以帮助我们解释因子。本例选正交旋转法，之后点击 Continue 钮返回 Factor Analysis 对话框。

点击 Scores...钮，弹出弹出 Factor Analysis: Scores 对话框（图 11.6），系统提供 3 种估计因子得分系数的方法，本例选 Regression（回归因子得分），之后点击 Continue 钮返回 Factor Analysis 对话框，再点击 OK 钮即完成分析。

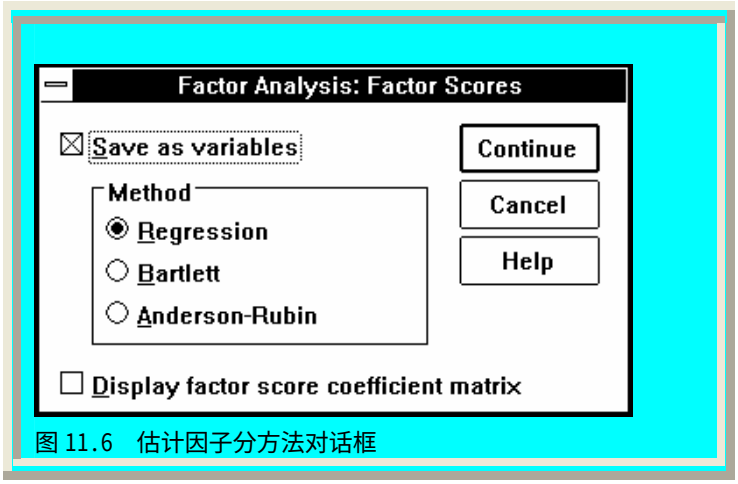


图 11.6　估计因子分方法对话框

11.2.3　结果解释

在输出结果窗口中将看到如下统计数据：

系统首先输出各变量的均数（Mean）与标准差（Std Dev），并显示共有 25 例观察单位进入分析；

接着输出相关系数矩阵（Correlation Matrix），经 Bartlett 检验表明：Bartlett 值 =326.28484，P<0.0001，即相关矩阵不是一个单位矩阵，故考虑进行因子分析。

Kaiser-Meyer-Olkin Measure of Sampling Adequacy 是用于比较观测相关系数值与偏相关系数值的一个指标，其值愈逼近 1，表明对这些变量进行因子分析的效果愈好。今 KMO 值 =0.32122，偏小，意味着因子分析的结果可能不能接受。

```
Analysis number 1   Listwise deletion of cases with missing values

               Mean      Std Dev   Label
X1           7.10000     2.32380
X2           4.77320     2.41779
X3           2.34880     1.66556
X4           9.15240     3.01405
X5           5.45840     3.27344
X6           7.16720     4.55817
X7           2.34600     1.61091

Number of Cases  =      25

Correlation Matrix:
              X1        X2        X3        X4        X5        X6        X7
X1         1.00000
X2          .58026   1.00000
X3          .20113    .36379   1.00000
X4          .90900    .83725    .43611   1.00000
X5          .28347    .16590   -.70423    .16328   1.00000
X6          .28656    .26119   -.68058    .20309    .99020   1.00000
X7         -.53321   -.60846   -.64918   -.67758    .42733    .35732   1.00000

Kaiser-Meyer-Olkin Measure of Sampling Adequacy =  .32122
Bartlett Test of Sphericity = 326.28484, Significance =     .00000
```

使用主成分分析法得到 2 个因子，因子矩阵（Factor Matrix）如下，变量与某一因子的联系系数绝对值越大，则该因子与变量关系越近。如本例变量 X7 与第一因子的值为-0.88644，与第二因子的值为 0.21921，可见其与第一因子更近，与第二因子更远。或者因子矩阵也可以作为因子贡献大小的度量，其绝对值越大，贡献也越大。

在 Final Statistics 一栏中显示各因子解释掉方差的比例，也称变量的共同度（Communality）。共同度从 0 到 1，0 为因子不解释任何方差，1 为所有方差均被因子解释掉。一个因子越大地解释掉变量的方差，说明因子包含原有变量信息的量越多。

```
Extraction   1 for analysis   1, Principal Components Analysis (PC)
PC    extracted   2 factors.
```

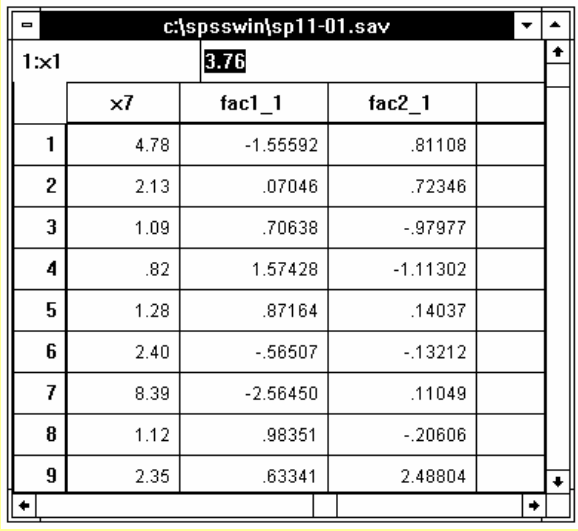
```
Factor Matrix:
              Factor  1     Factor  2
X1              .74646        .48929
X2              .79644        .37219
X3              .70890       -.59727
X4              .91054        .38865
X5             -.23424        .96350
X6             -.17715        .97172
X7             -.88644        .21921

Final Statistics:
Variable     Communality  *  Factor   Eigenvalue   Pct of Var   Cum Pct
                          *
X1               .79660   *     1       3.39518       48.5         48.5
X2               .77284   *     2       2.80632       40.1         88.6
X3               .85927   *
X4               .98014   *
X5               .98320   *
X6               .97561   *
X7               .83384   *
```

下面显示经正交旋转后的因子负荷矩阵（Rotated Factor Matrix）和因子转换矩阵（Factor Transformation Matrix)。旋转的目的是使复杂的矩阵变得简洁，即第一因子替代了 X1、X2、X4、X7 的作用，第二因子替代了 X3、X5、X6 的作用。

```
VARIMAX   rotation   1 for extraction   1 in analysis  1 - Kaiser Normalization.

VARIMAX converged in 3 iterations.

Rotated Factor Matrix:
             Factor  1     Factor  2
X1             .87795        .16064
X2             .87848        .03332
X3             .42098       -.82586
X4             .99001        .00414
X5             .15872        .97878
X6             .21452        .96415
X7            -.73151        .54656

Factor Transformation Matrix:
              Factor  1     Factor  2
Factor  1       .92135       -.38873
```

```
Factor  2        .38873        .92135
```

最后将第一因子的因子分用变量名 fac_1、第二因子的因子分用变量名 fac_2 存入原始数据库中。这些值既可用于模型诊断，又可用于进一步分析。



c:\spsswin\sp11-01.sav

1:x1 3.76

	x7	fac1_1	fac2_1
1	4.78	-1.55592	.81108
2	2.13	.07046	.72346
3	1.09	.70638	-.97977
4	.82	1.57428	-1.11302
5	1.28	.87164	.14037
6	2.40	-.56507	-.13212
7	8.39	-2.56450	.11049
8	1.12	.98351	-.20606
9	2.35	.63341	2.48804

图 11.7　因子分的获得并存盘

第十二章　可靠性分析

12.1　主要功能

在精神卫生与社会医学研究中，经常需要借助量表来了解对象的某一特性。如常用的症状自评量表（SCL-90）即用于评定对象精神病症状的表现形式与强度；又如生活事件量表（LES）即用于对精神刺激进行定性和定量分析。在完成一份量表的编制工作后，或在准备将一份已有的量表作实际应用前，需要对量表的信度进行考核。

量表的使用是为了了解被测对象的某一特征，因而在编制一份量表时，所设立的一系列项目是为了体现量表需要测定的这一特征。如果所设立的测定项目无法获得这一特征，则表示该量表可靠性差，即信度低。所以，研究者有时需要了解量表中各测定项目之间的一致性（同质信度考核），有时需要将量表的测定项目按原编号的奇、偶数分半后，对各自的测定结果进行相关性检验（分半信度考核），等等，这就是量表的可靠性分析，亦即信度研究。

量表的可靠性分析可通过调用 Reliability 过程完成。

12.2 实例操作

［例 12.1］采用家庭环境量表（FES）研究 30 名女医师的家庭特征，测定结果按 10 个分量表的实际得分整理如下。请以此资料对 FES 的信度作评价。

编号 No.	亲密度 FES1	情感表达 FES2	矛盾性 FES3	独立性 FES4	成功性 FES5	知识性 FES6	娱乐性 FES7	道德宗教观 FES8	组织性 FES9	控制性 FES10
1	2	4	5	7	8	6	3	4	5	6
2	4	2	1	4	5	1	1	6	6	2
3	4	3	2	5	5	3	2	5	3	2
4	4	3	3	7	7	2	6	6	4	3
5	4	3	4	5	5	3	4	5	6	3
6	4	2	1	5	3	4	7	4	5	2
7	5	4	6	4	6	3	4	5	6	3
8	3	1	4	7	5	3	5	7	6	3
9	5	2	5	5	7	3	3	6	6	3
10	5	3	2	6	7	3	3	5	4	2
11	3	4	3	4	5	4	3	6	4	2
12	3	3	2	4	4	4	2	4	4	2
13	2	2	0	3	4	3	6	5	7	2
14	2	3	3	4	4	4	4	6	5	3
15	1	3	4	5	7	4	2	6	4	4
16	5	2	3	4	6	4	4	6	4	2
17	5	4	3	5	6	3	3	4	4	2
18	3	3	3	4	5	3	3	5	3	2
19	6	3	6	5	6	3	5	4	4	3
20	5	3	2	4	5	3	6	5	6	3
21	2	3	2	4	4	3	1	4	5	3
22	2	3	2	3	6	5	4	5	5	3
23	4	5	3	5	4	5	5	5	3	4
24	2	2	4	5	6	3	3	5	4	3
25	5	5	3	3	4	5	2	7	7	3
26	4	2	2	5	7	5	3	6	4	2
27	4	2	5	6	4	4	5	3	4	2
28	4	6	5	4	5	4	3	6	5	5
29	5	5	2	4	4	5	6	4	5	2
30	3	3	3	4	4	4	3	7	4	1

12.2.1 数据准备

激活数据管理窗口，定义变量名：亲密度、情感表达、矛盾性、独立性、成功性、知识性、娱乐性、道德宗教观、组织性、控制性等十个分量表的变量名依次是 FES1、FES2、FES3、FES4、FES5、FES6、FES7、FES8、FES9、FES10，输入原始数据。

12.2.2 统计分析

激活 Statistics 菜单选 Scale 中的 Reliability Analysis...项，弹出 Reliability Analysis 对话框（如图 12.1 示）。从对话框左侧的变量列表中选 fes1~fes10 共十个变量，点击 ➢ 钮使之进入 Items 框。点击 Model 处的下拉菜单，系统提供 5 种分析模型：

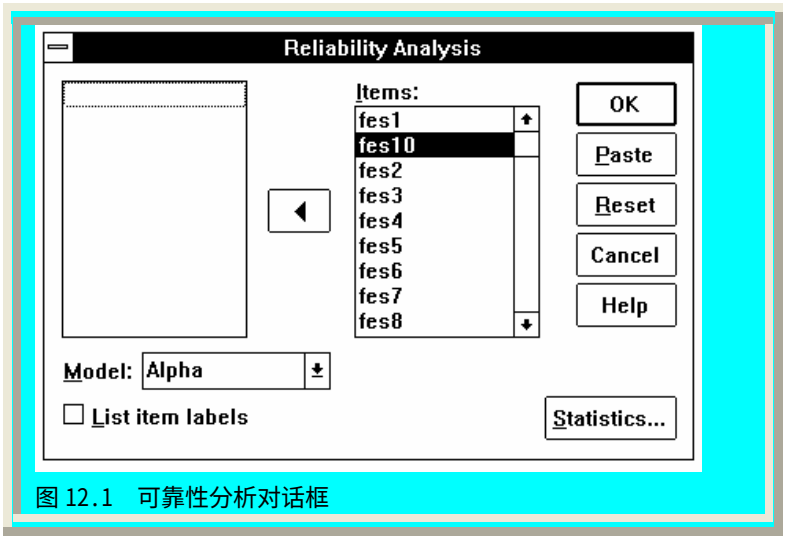


图 12.1　可靠性分析对话框

Alpha：计算信度系数 Cronbach α 值；
Split half：分半信度的分析；
Guttman：真实可靠性的 Guttman 低界；
Parallel：并行模型假定下的极大似然可靠性估计；
Strict parallel：严格并行模型假定下的极大似然可靠性估计。
本例选用 Alpha 模型。
点击 Statistics...钮，弹出 Reliability Analysis: Statistics 对话框（图 12.2），该对话框内含如下选项：

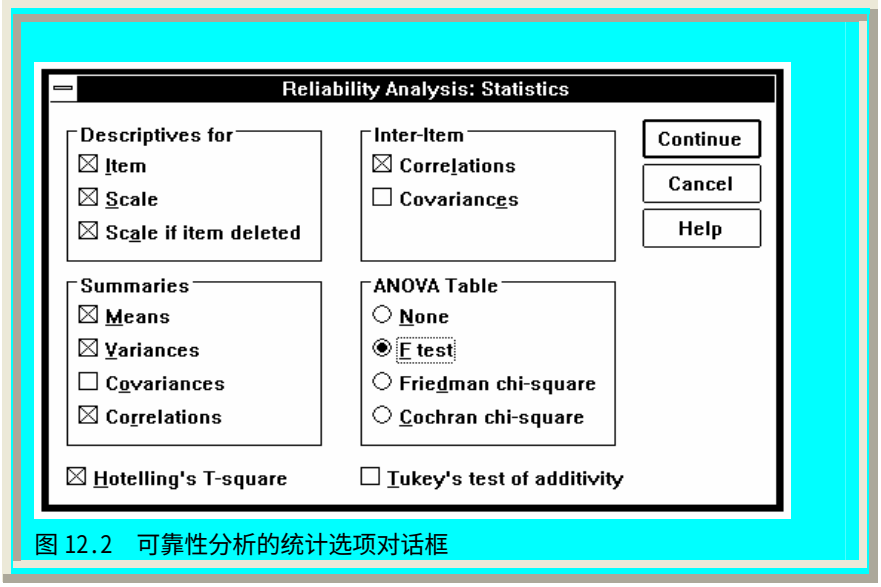


图 12.2　可靠性分析的统计选项对话框

在 Descriptives for 栏中选 Item、Scale、Scale if item deleted 项，以指定对各项目、测定得分情况和项目与量表总体特征关系进行描述性统计；

在 Summaries 处有四个选项：Means、Variances、Covariances 和 Correlations，可分别要求系统计算在 Descriptives for 栏中指定对象的平均数、方差、协方差和相关系数，本例选 Means、Variances 和 Correlations 三项；

在 Inter-Item 处有 Correlations 和 Covariances 两项，前者可计算项目间的两两相关系数，后者可计算项目间的两两协方差值，本例选 Correlations 项；

在 ANOVA Table 处有 None、F test、Friedman chi-square、Cochran chi-square 四个选项，其意义

分别是：不作方差分析、作重复度量的方差分析、计算 Friedman 和 Kendall 谐和系数（适用于等级资料）、计算 Cohran Q 值（适用于所有项目均为二分变量），本例选 F test 项；

此外，还有 Hotelling's T-square 选项，可要求作项目间平均得分的相等性检验；Tukey's test of additivity 选项，可要求作可加性的 Tukey 检验，本例仅选前一项。在完成各选项的选择之后，点击 Continue 钮返回 Reliability Analysis 对话框，再点击 OK 钮即完成分析。

12.2.3 结果解释

在结果输出窗口中将看到如下统计数据：

首先计算各项目在 30 名被试中测定结果的均数与标准差。然后输出项目间的两两相关系数矩阵，从中可见，第三项目（矛盾性）与第十项目（控制性）的相关程度最密切（r = 0.5038）。

		Mean	Std Dev	Cases
1.	FES1	3.6667	1.2685	30.0
2.	FES10	2.7333	1.0148	30.0
3.	FES2	3.1000	1.1250	30.0
4.	FES3	3.1000	1.4704	30.0
5.	FES4	4.6667	1.0933	30.0
6.	FES5	5.2667	1.2576	30.0
7.	FES6	3.6333	1.0334	30.0
8.	FES7	3.7000	1.5570	30.0
9.	FES8	5.2000	1.0306	30.0
10.	FES9	4.7333	1.1121	30.0

Correlation Matrix

	FES1	FES10	FES2	FES3	FES4	FES5	FES6	FES7	FES8	FES9
FES1	1.0000									
FES10	-.2321	1.0000								
FES2	.1933	.4168	1.0000							
FES3	.1849	.5038	.2022	1.0000						
FES4	.0414	.2901	-.2523	.3646	1.0000					
FES5	.0360	.3819	-.0926	.3953	.4681	1.0000				
FES6	-.1754	.3310	.4183	.0930	-.0814	-.0283	1.0000			
FES7	.2444	-.0524	-.1004	-.0467	.1823	-.2043	.0793	1.0000		
FES8	-.0791	-.0132	-.0476	-.0137	-.1224	.1171	-.0907	-.2192	1.0000	
FES9	.0570	.1487	-.0606	-.0675	-.2741	-.1446	-.1180	.1314	.2287	1.0000

N of Cases = 30.0

接着显示对整个量表的统计分析结果。该量表的平均得分为 39.8000，标准差为 4.8309；平均每个项目的得分为 3.9800，极差为 2.5333；各项目的方差平均为 1.4634；项目间的相关系数范围为 -0.2741—0.5038。

之后考查项目与量表得分的关系，方式是：若将某一项目从量表中剔除，则量表的平均得分（Scale Mean if Item Deleted）、方差（Scale Variance if Item Deleted）、每个项目得分与剩余各项目得分间的相关系数（Corrected Item-Total Correlation）、以该项目为自变量所有其他项目为应变量建立回归方

程的R^2值（Squared Multiple Correlation）以及Cronbach α值（Alpha if Item Deleted）会是多少。如本例在每个项目得分与剩余各项目得分间的相关系数中，第十项目（控制性）最大，为 0.5009，表明该项目与其他各项目关系最密切。又如R^2值，第十项目（控制性）最大，为 0.7345，表明其含义有 73.45%可被其他项目解释掉，而第八项目（道德宗教观）最小，为 0.1556，表明其含义仅有 15.56%可被其他项目解释掉。

Statistics for Scale	Mean	Variance	Std Dev	N of Variables
	39.8000	23.3379	4.8309	10

Item Means	Mean	Minimum	Maximum	Range	Max/Min	Variance
	3.9800	2.7333	5.2667	2.5333	1.9268	.8425

Item Variances	Mean	Minimum	Maximum	Range	Max/Min	Variance
	1.4634	1.0299	2.4241	1.3943	2.3538	.2349

Inter-item Correlations	Mean	Minimum	Maximum	Range	Max/Min	Variance
	.0665	-.2741	.5038	.7779	-1.8376	.0450

Item-total Statistics

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Alpha if Item Deleted
FES1	36.1333	20.3954	.1164	.5124	.4065
FES10	37.0667	17.9954	.5009	.7345	.2745
FES2	36.7000	20.4241	.1621	.6318	.3886
FES3	36.7000	15.8724	.4527	.4233	.2410
FES4	35.1333	20.0506	.2136	.5457	.3710
FES5	34.5333	19.1540	.2364	.4650	.3583
FES6	36.1667	21.2471	.1074	.3106	.4067
FES7	36.1000	20.8517	.0044	.3135	.4662
FES8	34.6000	23.0069	-.0739	.1556	.4613
FES9	35.0667	22.3402	-.0227	.4297	.4503

方差分析表明，F = 18.4933，P<0.0001，即该量表的重复度量效果良好。

Analysis of Variance

Source of Variation	Sum of Sq.	DF	Mean Square	F	Prob.
Between People	67.6800	29	2.3338		
Within People	584.2000	270	2.1637		

Between Measures	227.4800	9	25.2756	18.4933	.0000
Residual	356.7200	261	1.3667		
Total	651.8800	299	2.1802		

经Hotelling T^2检验可知，该量表的项目间平均得分的相等性好，即项目具有内在的相关性；在量表的信度检验中，Cronbach α = 0.4144，标准化Cronbach α = 0.4161。Cronbach α 系数的意义是：个体在这一量表的测定得分与如果询问所有可能项目的测定得分的相关系数的平方，即这一量表能获得真分数的能力。本例为 0.4144，意味着对于家庭情况，FES量表尚有 58.56%的内容未曾涉及，故信度不高。

Hotelling's T-Squared = 277.1019	F = 22.2956	Prob. = .0000
Degrees of Freedom:	Numerator = 9	Denominator = 21
Reliability Coefficients	10 items	
Alpha = .4144	Standardized item alpha = .4161	

第十三章 非参数检验

许多统计分析方法的应用对总体有特殊的要求，如 t 检验要求总体符合正态分布，F 检验要求误差呈正态分布且各组方差整齐，等等。这些方法常用来估计或检验总体参数，统称为参数统计。

但许多调查或实验所得的科研数据，其总体分布未知或无法确定，这时做统计分析常常不是针对总体参数，而是针对总体的某些一般性假设（如总体分布），这类方法称非参数统计（Nonparametric tests）。

非参数统计方法简便，适用性强，但检验效率较低，应用时应加以考虑。

第一节 Chi-Square 过程

13.1.1 主要功能

调用此过程可对样本数据的分布进行卡方检验。卡方检验适用于配合度检验，主要用于分析实际频数与某理论频数是否相符。

13.1.2 实例操作

[例 13-1]某地一周内各日死亡数的分布如下表，请检验一周内各日的死亡危险性是否相同？

周　日	死亡数
一	11
二	19
三	17
四	15
五	15
六	16
日	19

13.1.2.1　数据准备

激活数据管理窗口，定义变量名：各周日为 day，死亡数为 death。按顺序输入数据， 结果见图 13.1。激活 Data 菜单选 Weight Cases...命令项，弹出 Weight Cases 对话框（如图 13.2），选 death 点击钮使之进入 Frequency Variable 框，定义死亡数为权数，再点击 OK 钮即可。

图 13.1　数据录入窗口

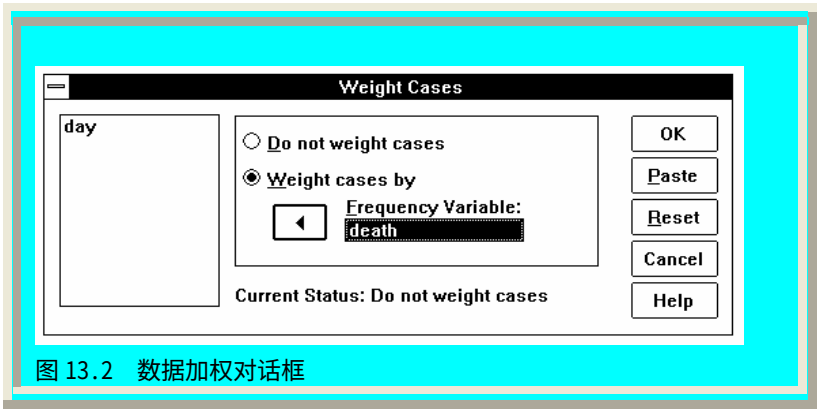


图 13.2　数据加权对话框

13.1.2.2　统计分析

激活 Statistics 菜单选 Nonparametric Tests 中的 Chi-Square...命令项，弹出 Chi-Square Test 对话框（图 13.3）。现欲对一周内各日的死亡数进行分布分析，故在对话框左侧的变量列表中选 day，点击钮使之进入 Test Variable List 框，点击 OK 钮即可。

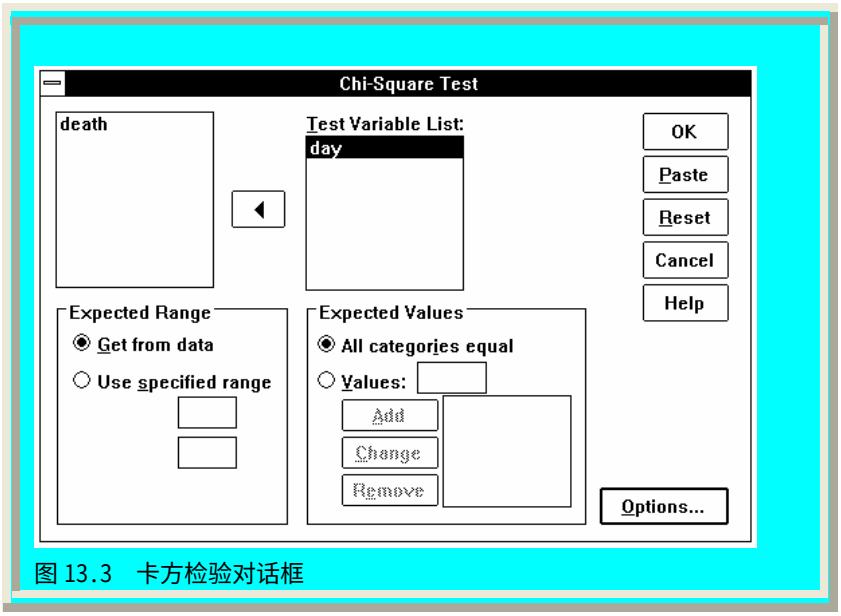


图 13.3　卡方检验对话框

13.1.2.3　结果解释

在结果输出窗口中将看到如下统计数据：

运算结果显示一周内各日死亡的理论数（Expected）为 15.71，即一周内各日死亡均数；还算出实际死亡数与理论死亡数的差值（Residual）；卡方值 $\chi^2 = 3.4000$，自由度数（D.F.）= 6 ，P = 0.7572 ，可认为一周内各日的死亡危险性是相同的。

DAY

Category	Cases Observed	Expected	Residual
1.00	11	15.71	-4.71
2.00	19	15.71	3.29
3.00	17	15.71	1.29
4.00	15	15.71	-.71
5.00	13	15.71	-2.71
6.00	16	15.71	.29
7.00	19	15.71	3.29
Total	110		

Chi-Square	D.F.	Significance
3.4000	6	.7572

第二节 Binomial 过程

13.2.1 主要功能

有些总体只能划分为两类，如医学中的生与死、患病的有与无。从这种二分类总体中抽取的所有可能结果，要么是对立分类中的这一类，要么是另一类，其频数分布称为二项分布。调用 Binomial 过程可对样本资料进行二项分布分析。

13.2.2 实例操作

[例 13-2]某地某一时期内出生 40 名婴儿，其中女性 12 名（定 Sex=0），男性 28 名（定 Sex=1）。问这个地方出生婴儿的性比例与通常的男女性比例（总体概率约为 0.5）是否不同？

13.2.2.1 数据准备

激活数据管理窗口，定义性别变量为 sex。按出生顺序输入数据，男性为 1 ，女性为 0。

13.2.2.2 统计分析

激活 Statistics 菜单选 Nonparametric Tests 中的 Binomial Test...命令项，弹出 Binomial Test 对话框（图 13.4）。在对话框左侧的变量列表中选 sex，点击钮使之进入 Test Variable List 框，在 Test Proportion 框中键入 0.50，再点击 OK 钮即可。

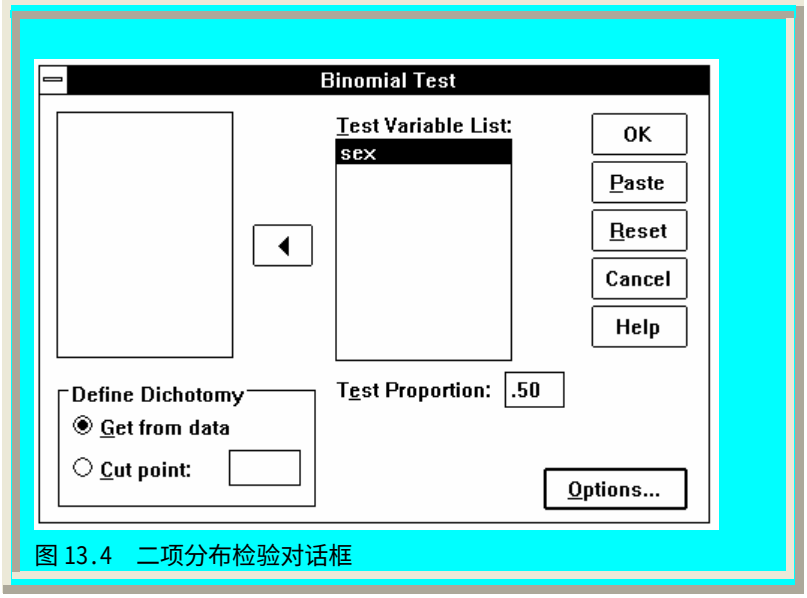


图 13.4　二项分布检验对话框

13.2.2.3　结果解释

在结果输出窗口中将看到如下统计数据：

二项分布检验表明，女婴 12 名，男婴 28 名，观察概率为 0.7000（即男婴占 70%），检验概率为 0.5000，二项分布检验的结果是双侧概率为 0.0177，可认为男女比例的差异有高度显著性，即与通常 0.5 的性比例相比，该地男婴比女婴明显为多。

```
SEX
      Cases
                                Test Prop. =    .5000
         28     = 1.00          Obs. Prop. =    .7000
         12     =  .00
         --                     Z Approximation
         40    Total            2-Tailed  P = .0177
```

第三节 Runs 过程

13.3.1 主要功能

依时间或其他顺序排列的有序数列中，具有相同的事件或符号的连续部分称为一个游程。调用 Runs 过程可进行游程检验，即用于检验序列中事件发生过程的随机性分析。

13.3.2 实例操作

[例 13-3]某村发生一种地方病，其住户沿一条河排列，调查时对发病的住户标记为“1”，对非发病的住户标记为“0”，共 17 户：

```
0  1  1  0  0  0  1  0  0  1  0  0  0  0  1  1  0  0  1  0  0  0  0  1  0  1
```

问病户的分布排列是呈聚集趋势，还是随机分布？

13.3.2.1 数据准备

激活数据管理窗口，定义住户变量为 epi。按住户顺序输入数据，发病的住户为 1，非发病的住户为 0。

13.3.2.2 统计分析

激活 Statistics 菜单选 Nonparametric Tests 中的 Runs Test...项，弹出 Runs Test 对话框（图 13.5）。在对话框左侧的变量列表中选 epi，点击钮使之进入 Test Variable List 框。在临界割点 Cut Point 框中有四个选项：

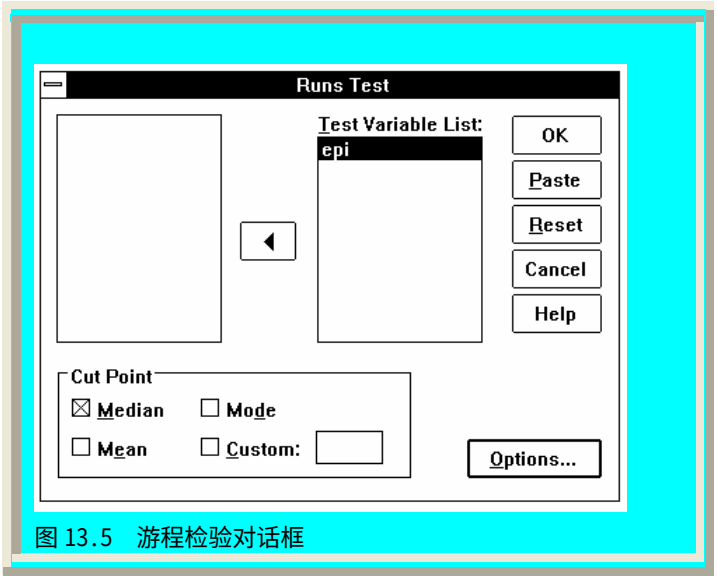


图 13.5　游程检验对话框

1、Median：中位数作临界割点，其值在临界割点之下的为一类，大于或等于临界割点的为另一类；

2、Mode：众数作临界割点，其值在临界割点之下的为一类，大于或等于临界割点的为另一类；

3、Mean：均数作临界割点，其值在临界割点之下的为一类，大于或等于临界割点的为另一类；

4、Custom：用户指定临界割点，其值在临界割点之下的为一类，大于或等于临界割点的为另一类；

本例选 Custom 项，在其方框中键入 1（根据需要选项，本例是 0、1 二分变量，故临界割点值用 1），再点击 OK 钮即可。

13.3.2.3　结果解释

在结果输出窗口中将看到如下统计数据：

检验结果可见本例游程个数为 14，检验临界割点值（Test value） = 1.00，小于 1.00 者有 17 个案例，而大于或等于 1.00 者有 9 个案例。Z = 0.3246，双侧 P = 0.7455。 所以认为此地方病的病户沿河分布的情况无聚集性，而是呈随机分布。

```
EPI
          Runs:   14          Test  value  =  1.00

          Cases:  17   LT 1.00
                   9   GE 1.00           Z =   .3246
                  --
                  26   Total     2-Tailed  P =   .7455
```

第四节 1-Sample K-S 过程

13.4.1 主要功能

调用此过程可对单样本进行 Kolmogorov-Smirnov Z 检验，它将一个变量的实际频数分布与正态分布（Normal）、均匀分布（Uniform）、泊松分布（Poisson）进行比较。

13.4.2 实例操作

[例 13-4]某地正常成年男子 144 人红细胞计数（万/立方毫米）的频数资料如下，问该资料的频数是否呈正态分布？

红细胞计数	人数	红细胞计数	人数
420-	2	540-	24
440-	4	560-	22
460-	7	580-	16
480-	16	600-	2
500-	20	620-	6
520-	25	640-	1

13.4.2.1 数据准备

激活数据管理窗口，定义频数变量名为 f，依次输入人数资料。

13.4.2.2 统计分析

激活 Statistics 菜单选 Nonparametric Tests 中的 1-Sample K-S . . . 命令项，弹出 One-Sample Kolmogorov-Smirnov Test 对话框(图 13.6)。在对话框左侧的变量列表中选 f,点击钮使之进入 Test Variable List 框，在 Test Distribution 框中选 Normal 项，表明与正态分布形式相比较，再点击 OK 钮即可。

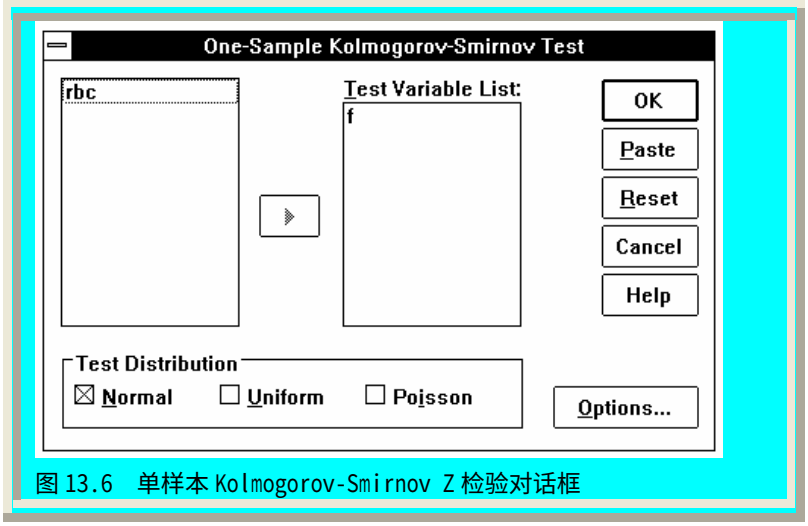


图 13.6　单样本 Kolmogorov-Smirnov Z 检验对话框

13.4.2.3　结果解释

在结果输出窗口中将看到如下统计数据：

K-S 正态性检验的结果显示，Z 值=0.7032，双侧 P 值=0.7060，可认为该地正常成年男子的红细胞计数符合正态分布。

```
F
     Test distribution  -  Normal                    Mean:   12.0000
                                          Standard Deviation:    9.3808
              Cases:  12
              Most extreme differences
        Absolute        Positive       Negative        K-S Z       2-Tailed P
         .20298          .20298        -.16509          .7032          .7060
```

第五节 2 Independent Samples 过程

13.5.1 主要功能

调用此过程可对两个独立样本的均数、中位数、离散趋势、偏度等进行差异比较检验。

13.5.2 实例操作

[例 13-5]调查某厂的铅作业工人 7 人和非铅作业工人 10 人的血铅值（μg / 100g）如下，问两组工人的血铅值有无差别？

非铅作业组	5	5	6	7	9	12	13	15	18	21
铅作业组		17	18	20	25	34	43	44		

13.5.2.1 数据准备

激活数据管理窗口，定义分组变量为 group（非铅作业组为 1，铅作业组为 2），血铅值为 Pb。按顺序输入数据。

13.5.2.2 统计分析

激活 Statistics 菜单选 Nonparametric Tests 中的 2 Independent Samples...命令项，弹出 Two-Independent-Samples-Test 对话框（图 13.7）。在对话框左侧的变量列表中选 Pb,点击钮使之进入 Test Variable List 框;选 group,点击钮使之进入 Grouping Variable 框,点击 Define Groups...钮，在弹出的 Two Independent Samples:Define Groups 对话框内定义 Group 1 为 1，Group 2 为 2，之后点击 Continue 钮返回 Two-Independent-Samples-Test 对话框；在 Test Type 框中有四种检验方法：

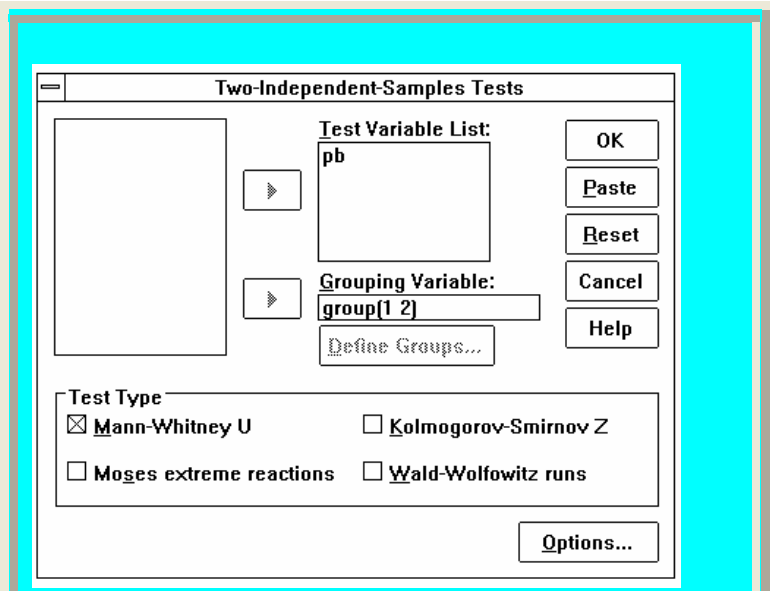


图 13.7　两独立样本检验对话框

Mann-Whitney U：主要用于判别两个独立样本所属的总体是否有相同的分布；

Kolmogorov-Smirnov Z：推测两个样本是否来自具有相同分布的总体；

Moses extreme reactions：检验两个独立样本之观察值的散布范围是否有差异存在，以检验两个样本是否来自具有同一分布的总体；

Wald-Wolfowitz runs：考察两个独立样本是否来自具有相同分布的总体。

本例选 Mann-Whitney U 检验方法，之后点击 OK 钮即可。

13.5.2.3　结果解释

在结果输出窗口中将看到如下统计数据：

结果表明，第 1 组的平均秩次（Mean Rank）为 5.95，第 2 组的平均秩次为 13.36，U = 4.5，W = 93.5，精确双侧概率 P = 0.0012，可认为铅作业组工人的血铅值高于非铅作业组。

```
PB   by   GROUP
      Mean Rank     Cases
```

```
     5.95        10     GROUP = 1
    13.36         7     GROUP = 2
                 --
                 17   Total
                                Exact               Corrected for ties
        U                W       2-Tailed P          Z        2-Tailed P
       4.5              93.5      .0012         -2.9801         .0029
```

第六节 k Independent Samples 过程

13.6.1 主要功能

调用此过程可对多个独立样本进行中位数检验和 Kruskal-Wallis H 检验。

13.6.2 实例操作

[例 13-6]随机抽样得以下三组人的血浆总皮质醇测定值（μg / L），试比较有无差异？

正常人	单纯性肥胖	皮质醇增多症
0.4	0.6	9.8
1.9	1.2	10.2
2.2	2.0	10.6
2.5	2.4	13.0
2.8	3.1	14.0
3.1	4.1	14.8
3.7	5.0	15.6
3.9	5.9	15.6
4.6	7.4	21.6
7.0	13.6	24.0

13.6.2.1 数据准备

激活数据管理窗口，定义分组变量为 group（正常人为 1，单纯性肥胖为 2，皮质醇增多症为 3），总皮质醇测定值为 pzc。按顺序输入数据。

13.6.2.2 统计分析

激活 Statistics 菜单选 Nonparametric Tests 中的 k Independent Samples...项，弹出 Tests for Several Independent Samples 对话框（图 13.8）。在对话框左侧的变量列表中选 pzc，点击钮使之进入 Test Variable List 框。选 group，点击钮使之进入 Grouping Variable 框，点击 Define Range...钮，在弹出的 K Independent Samples:Define Range 对话框内定义 Mininum 为 1，Maxinum

为 2，之后点击 Continue 钮返回 Two-Independent-Samples-Test 对话框。在 Test Type 框中有两个检验方法的选项：Kruskal-Wallis H 为单向方差分析，检验多个样本在中位数上是否有差异，Median 为中位数检验，检验多个样本是否来自具有相同中位数的总体；本例选 Kruskal-Wallis H 项。之后点击 OK 钮即可。

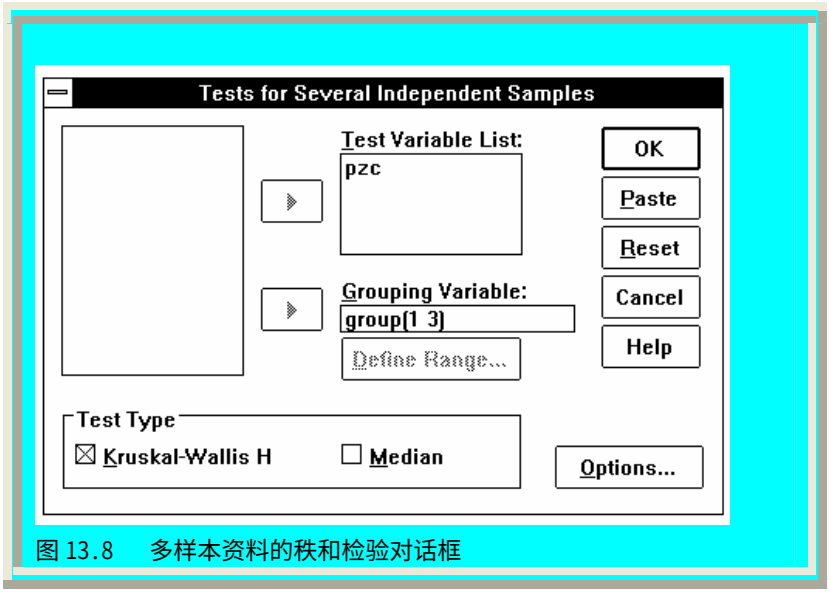


图 13.8　多样本资料的秩和检验对话框

13.6.2.3　结果解释

在结果输出窗口中将看到如下统计数据：

结果表明，1 至 3 组的平均秩次（Mean Rank）分别为 9.65、11.75、25.10，χ^2值（即H值）为 18.1219，P = 0.0001；可认为三组人的血浆总皮质醇测定值有差异，根据本例情况可看出皮质醇增多症组高于其他两组人。

```
PZC   by   GROUP
      Mean Rank     Cases
           9.65        10    GROUP =     1
          11.75        10    GROUP =     2
          25.10        10    GROUP =     3
                       --
                       30    Total
                                          Corrected for ties
 Chi-Square       D.F.     Significance     Chi-Square       D.F.     Significance
    18.1219          2          .0001          18.1300          2          .0001
```

第七节　2 Related Samples 过程

13.7.1　主要功能

调用此过程可对两个相关样本资料（如配对、配伍资料）进行秩和检验。

13.7.2 实例操作

[例 13-7]研究饲料中缺乏 Vit E 对大鼠肝中 Vit A 含量的关系，将大鼠按性别相同、体重相近的原则配成 8 对，并将每对大鼠随机分为 2 组（正常饲料组、Vit E 缺乏饲料组），一定时间后杀死大鼠，测定肝中 Vit A 含量，结果如下表，问：饲料中缺乏 Vit E 对大鼠肝中 Vit A 含量有无影响？

大鼠对别	正常饲料组	Vit E 缺乏饲料组
1	37.2	25.7
2	20.9	25.1
3	31.4	18.8
4	41.4	33.5
5	39.8	34.0
6	39.3	28.3
7	36.1	26.2
8	31.9	18.3

13.7.2.1 数据准备

激活数据管理窗口，定义正常饲料组变量名为 va1，Vit E 缺乏饲料组变量名为 va2，按顺序输入数据。

13.7.2.2 统计分析

激活 Statistics 菜单选 Nonparametric Tests 中 2 Related Samples...项，弹出 Two-Related-Samples Tests 对话框（图 13.9）。在对话框左侧的变量列表中选 va1，在 Current Selections 栏的 Variable 1 处出现 va1，选 va2，在 Current Selections 栏的 Variable 2 处出现 va2，然后点击钮使 va1 -va2（表明是配对变量）进入 Test Pair(s) List 框。在 Test Type 框中有三种检验方法：

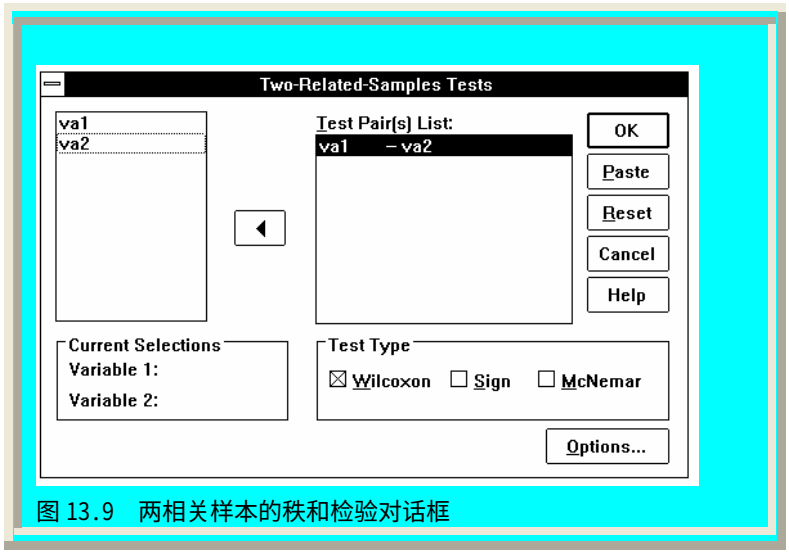


图 13.9 两相关样本的秩和检验对话框

1、Wilcoxon：配对符号等级秩次检验，

2、Sign：符号检验；

3、McNemar：以研究对象作自身对照，检验其“前后”的变化是否显著，该法适用于相关的二分变量数据。

本例选 Wilcoxon 和 Sign 两项。点击 Options...钮，弹出 Two-Related-Samples:Options 对话框，在 Statistics 栏中选 Decriptive 项，要求计算均数、标准差等指标，点击 Continue 钮返回 Two-Related-Samples Tests 对话框，之后点击 OK 钮即可。

13.7.2.3 结果解释

在结果输出窗口中将看到如下统计数据：

首先显示两变量 va1 和 va2 的例数、均数、标准差、最大值和最小值；配对符号秩和检验（Wilcoxon Matched-Pairs Signed-Ranks Test）结果，其平均秩分别为 5.00 和 1.00 ，Z = -2.3805，双侧 P = 0.0173，可认为两组大鼠肝中 Vit A 含量有差别，饲料中缺乏 Vit E 会使大鼠肝中 Vit A 含量降低；但符号检验（Sign Test）的结果，双侧 P = 0.0703，则认为两组大鼠肝中 Vit A 含量无差别。在这种情况下，应取配对符号秩和检验（Wilcoxon）结果，因两法比较之下，配对符号秩和检验较为敏感，效率较高。

```
                N        Mean      Std Dev  Minimum  Maximum
VA1             8    34.75000      6.64852    20.90    41.40
VA2             8    26.23750      5.82064    18.30    34.00

- - - - - Wilcoxon Matched-Pairs Signed-Ranks Test
       VA1
 with  VA2
     Mean Rank    Cases
          5.00        7  - Ranks (VA2 LT VA1)
          1.00        1  + Ranks (VA2 GT VA1)
                      0    Ties  (VA2 EQ VA1)
                     --
                      8    Total
         Z =   -2.3805           2-Tailed P =  .0173
- - - - - Sign Test
      VA1
with  VA2
           Cases
               7  - Diffs (VA2 LT VA1)
               1  + Diffs (VA2 GT VA1)       (Binomial)
               0    Ties                    2-Tailed P =      .0703
              --
               8    Total
```

第八节 K Related Samples 过程

13.8.1 主要功能

调用此过程可对多个相关样本资料（如配伍资料）进行秩和检验。

13.8.2 实例操作

[例 13-8]用某药治疗血吸虫病患者，在治疗前和治疗后一周、二周和四周各测定 7 名患者血清 SGPT 值的变化，以观察该药对肝功能的影响，结果如下表，问：患者四个阶段的血清 SGPT 值有无不同？

患者编号	治疗前	治疗后		
		一周	二周	四周
1	63	188	138	54
2	90	238	220	144
3	54	300	83	92
4	45	140	213	100
5	54	175	150	36
6	72	300	163	90
7	64	207	185	87

13.8.2.1 数据准备

激活数据管理窗口，定义变量名：治疗前为 before、治疗后一周为 w1、二周为 w2、四周为 w4，按顺序输入各组 SGPT 数据。

13.8.2.2 统计分析

激活 Statistics 菜单选 Nonparametric Tests 中的 k Related Samples...命令项，弹出 Tests for Serveral Related Samples 对话框（图 13.10）。在对话框左侧的变量列表中选 before、w1、w2 和 w4，点击钮使 before、w1、w2 和 w4 均进入 Test Variables 框。在 Test Type 框中有三种选项：

1、Friedman：双向方差分析，考察多个相关样本是否来自同一总体；

2、Cochran's Q：作为两相关样本 McNemar 检验的多样本推广，特别适用于定性变量和二分字符变量；

3、Kendall's W：Kendall 和谐系数检验，通过计算 Kendall 和谐系数 W，以检验多个相关样本是否来自同一分布的总体。

本例选 Friedman 和 Kendall's W 两种检验方法，再点击 Statistics...钮，弹出 K Related-Samples:Statistics 对话框，在 Statistics 栏中选 Decriptive 项，要求计算均数、标准差等指标，点击 Continue 钮返回 K Related-Samples Tests 对话框；最后点击 OK 钮即可。

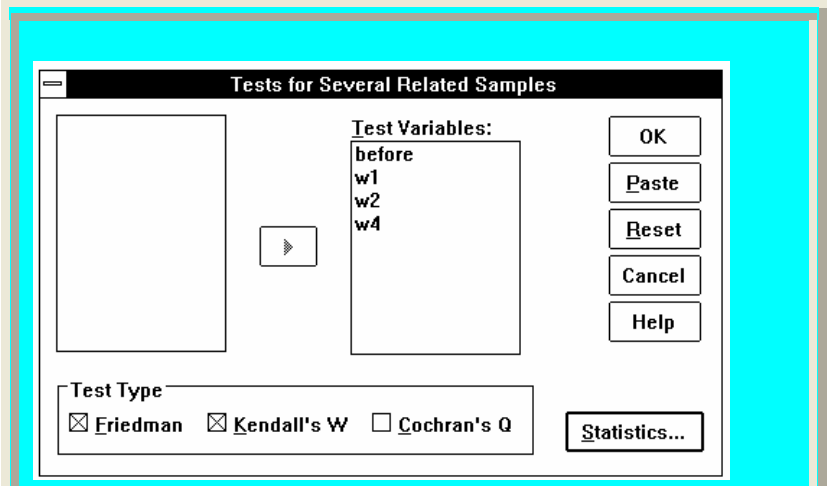


图 13.9　多个相关样本的秩和检验对话框

13.8.2.3　结果解释

在结果输出窗口中将看到如下统计数据：

首先显示的是四个变量 before、w1、w2、w4 的例数、均数、标准差、最大值和最小值。

接着显示检验结果：

Friedman双向方差分析，平均秩次分别 1.29、3.86、3.00 和 1.86 ， x^2= 16.7143，P = 0.0008，可认为患者四个阶段的血清SGPT值有差别。

Kendall和谐系数检验，平均秩次分别 1.29、3.86、3.00 和 1.86 ，和谐系数W = 0.7959， x^2 = 16.7143，P = 0.0008，结论同前。

```
                N          Mean      Std Dev    Minimum    Maximum
BEFORE          7     63.14286     14.70180      45.00      90.00
W1              7    221.14285     61.55331     140.00     300.00
W2              7    164.57143     47.27528      83.00     220.00
W4              7     86.14286     34.48878      36.00     144.00
- - - - - Friedman Two-Way Anova
     Mean Rank   Variable
          1.29   BEFORE
          3.86   W1
          3.00   W2
          1.86   W4
          Cases        Chi-Square          D.F.    Significance
              7          16.7143             3           .0008

- - - - - Kendall Coefficient of Concordance
     Mean Rank   Variable
          1.29   BEFORE
          3.86   W1
          3.00   W2
          1.86   W4
          Cases               W       Chi-Square         D.F.   Significance
```

7	.7959	16.7143	3	.0008

第十四章 生存分析

在临床诊疗工作的评价中，慢性疾病的预后一般不适合用治愈率、病死率等指标来考核，因为其无法在短时间内明确判断预后情况，为此，只能对患者进行长期随访，统计一定时期后的生存或死亡情况以判断诊疗效果。这就是生存分析。

第一节 Life Tables 过程

14.1.1 主要功能

调用此过程时，系统将采用即寿命表分析法，完成对病例随访资料在任意指定时点的生存状况评价。

14.1.2 实例操作

［例 14-1］ 用中药+化疗（中药组，16 例）和单纯化疗（对照组，10 例）两种疗法治疗白血病患者后，随访记录存活情况如下所示，试比较两组的生存率。

中药组		对照组	
随访月数	是否死亡	随访月数	是否死亡
10	否	2	是
2	是	13	否
12	是	7	是
13	否	11	是
18	否	6	否
6	是	1	否
19	是	11	否
26	是	3	否
9	否	17	否
8	是	7	否
6	是		
43	是		
9	是		
4	否		
31	否		
24			

14.1.2.1 数据准备

激活数据管理窗口，定义变量名：随访月数的变量名为 TIME，是否死亡的变量名为 DEATH，分组（即中药组与对照组）的变量名为 GROUP。输入原始数据：随访月数按原数值；是否死亡的，

是为 1，否为 0；分组的，中药组为 1，对照组为 2。

14.1.2.2 统计分析

激活 Statistics 菜单选 Survival 中的 Life Tables...项，弹出 Life Tables 对话框（图 14.1）。从对话框左侧的变量列表中选 time，点击 ➢ 钮使之进入 time 框；在 Display Time Intervals 栏中定义需要显示生存率的时点，本例要求从 0 个月显示至 48 个月，间隔为 2 个月，故在 0 through 框中输入 48，在 by 框中输入 2。选 death，点击 ➢ 钮使之进入 Status 框，点击 Define Event...钮弹出 Life Tables:Define Event for Status Variable 对话框，在 Single value 栏中输入 1，表明 death = 1 为发生死亡事件者；点击 Continue 钮返回 Life Tables 对话框。选 group，点击 ➢ 钮使之进入 Factor 框，点击 Define Range...钮，弹出 Life Tables:Define Range for Factor Variable 对话框，定义分组的范围，在 Mininum 框中输入 1，在 Maxinum 框中输入 2，点击 Continue 钮返回 Life Tables 对话框。

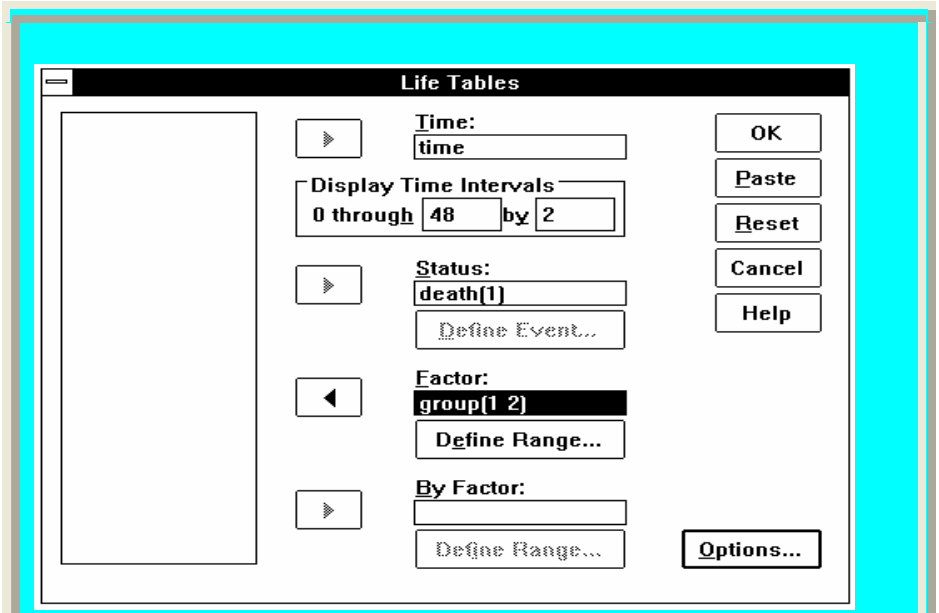


图 14.1 生存资料的寿命表分析对话框

点击 Options...钮弹出 Life Tables: Options 对话框，在 Plot 栏中选 Survival 项，要求绘制生存率曲线图；在 Compare Levels of First Factor 栏中选 Overall 项，要求作组间生存状况的比较。之后点击 Continue 钮返回 Life Tables 对话框，再点击 OK 钮即完成分析。

14.1.2.3 结果解释

在结果输出窗口中将看到如下统计数据：

共有 26 个观察对象进入分析。系统先显示中药组（group = 1）的生存状况寿命表，按用户指定，从 0 月起，隔 2 个月直至 42 个月（原指定从 0—48 个月，但因 42 个月后，生存概率已为 0，故 42 个月后至 48 个月的生存状况不再显示），分别显示进入该时点例数（Number Entrng this Intrvl）、从该时点失去的例数（Number Wdrawn Durong Intrvl）、该时点暴露于死亡危险的例数（Number Exposd to Risk）、该时点死亡的例数（Number of Termnl Events）、该时点死亡概率（Propn Terminating）、该时点生存概率（Propn Surviving）、该时点末生存率（Propn Surv at End）、单位时点的累积概率（Cumul Probability Densty）、该时点风险比例（Hazard Rate）、生存率的标准误（SE of Cumul Surviving）、单位时点累积概率的标准误（SE of Probability Densty）、风险比例的标准误（SE of Hazard Rate）。如本例，用中药+化疗的方式治疗白血病患者，至 8 个月时，死亡率为 17.39%，生存概率为 82.61%，生

存率为 66.38%，风险比例为 9.52%。至 42 个月时，生存概率和生存率均为 0，此时风险比例为 100%。中药组的 50%生存率在 19.44 个月。

对照组同类结果的显示，因在 16 个月时生存概率已为 0，故仅从 0 月起，隔 2 个月至 16 个月止。分析显示，单纯用化疗，白血病患者的半数生存率约在 16 个月多一点，比中药组少三个月。

This subfile contains: 26 observations

Life Table
Survival Variable TIME
for GROUP = 1

Intrvl Start Time	Number Entrng this Intrvl	Number Wdrawn During Intrvl	Number Exposd to Risk	Number of Termnl Events	Propn Termi-nating	Propn Sur-viving	Propn Surv at End	Cumul Proba-bility Densty	Hazard Rate	SE of Cumul Sur-viving	SE of Proba-bility Densty	SE of Hazard Rate
.0	16.0	.0	16.0	.0	.0000	1.0000	1.0000	.0000	.0000	.0000	.0000	.0000
2.0	16.0	.0	16.0	1.0	.0625	.9375	.9375	.0313	.0323	.0605	.0303	.0322
4.0	15.0	1.0	14.5	.0	.0000	1.0000	.9375	.0000	.0000	.0605	.0000	.0000
6.0	14.0	.0	14.0	2.0	.1429	.8571	.8036	.0670	.0769	.1019	.0441	.0542
8.0	12.0	1.0	11.5	2.0	.1739	.8261	.6638	.0699	.0952	.1231	.0458	.0670
10.0	9.0	1.0	8.5	.0	.0000	1.0000	.6638	.0000	.0000	.1231	.0000	.0000
12.0	8.0	1.0	7.5	1.0	.1333	.8667	.5753	.0443	.0714	.1348	.0420	.0712
14.0	6.0	.0	6.0	.0	.0000	1.0000	.5753	.0000	.0000	.1348	.0000	.0000
16.0	6.0	.0	6.0	.0	.0000	1.0000	.5753	.0000	.0000	.1348	.0000	.0000
18.0	6.0	1.0	5.5	1.0	.1818	.8182	.4707	.0523	.1000	.1453	.0489	.0995
20.0	4.0	.0	4.0	.0	.0000	1.0000	.4707	.0000	.0000	.1453	.0000	.0000
22.0	4.0	.0	4.0	.0	.0000	1.0000	.4707	.0000	.0000	.1453	.0000	.0000
24.0	4.0	1.0	3.5	.0	.0000	1.0000	.4707	.0000	.0000	.1453	.0000	.0000
26.0	3.0	1.0	2.5	.0	.0000	1.0000	.4707	.0000	.0000	.1453	.0000	.0000
28.0	2.0	.0	2.0	.0	.0000	1.0000	.4707	.0000	.0000	.1453	.0000	.0000
30.0	2.0	1.0	1.5	.0	.0000	1.0000	.4707	.0000	.0000	.1453	.0000	.0000
32.0	1.0	.0	1.0	.0	.0000	1.0000	.4707	.0000	.0000	.1453	.0000	.0000
34.0	1.0	.0	1.0	.0	.0000	1.0000	.4707	.0000	.0000	.1453	.0000	.0000
36.0	1.0	.0	1.0	.0	.0000	1.0000	.4707	.0000	.0000	.1453	.0000	.0000
38.0	1.0	.0	1.0	.0	.0000	1.0000	.4707	.0000	.0000	.1453	.0000	.0000
40.0	1.0	.0	1.0	.0	.0000	1.0000	.4707	.0000	.0000	.1453	.0000	.0000
42.0	1.0	.0	1.0	1.0	1.0000	.0000	.0000	.2354	1.0000	.0000	.0727	.0000

The median survival time for these data is 19.44

Life Table
Survival Variable TIME
for GROUP = 2

Number Number Number Number					Cumul		SE of	SE of

Intrvl Start Time	Entrng this Intrvl	Wdrawn During Intrvl	Exposd to Risk	of Termnl Events	Propn Termi-nating	Propn Sur-viving	Propn Surv at End	Proba-bility Densty	Hazard Rate	Cumul Sur-viving	Proba-bility Densty	SE of Hazard Rate
.0	10.0	1.0	9.5	.0	.0000	1.0000	1.0000	.0000	.0000	.0000	.0000	.0000
2.0	9.0	1.0	8.5	1.0	.1176	.8824	.8824	.0588	.0625	.1105	.0553	.0624
4.0	7.0	.0	7.0	.0	.0000	1.0000	.8824	.0000	.0000	.1105	.0000	.0000
6.0	7.0	2.0	6.0	1.0	.1667	.8333	.7353	.0735	.0909	.1628	.0678	.0905
8.0	4.0	.0	4.0	.0	.0000	1.0000	.7353	.0000	.0000	.1628	.0000	.0000
10.0	4.0	1.0	3.5	1.0	.2857	.7143	.5252	.1050	.1667	.2122	.0918	.1643
12.0	2.0	1.0	1.5	.0	.0000	1.0000	.5252	.0000	.0000	.2122	.0000	.0000
14.0	1.0	.0	1.0	.0	.0000	1.0000	.5252	.0000	.0000	.2122	.0000	.0000
16.0	1.0	1.0	.5	.0	.0000	1.0000	.5252	.0000	.0000	.2122	.0000	.0000

The median survival time for these data is 16.00+

接着显示两组比较的结果。系统采用 Gehan 比分检验法，得 $u = 0.012$，$P = 0.9113$，即中药组与对照组的生存率无差别。

Comparison of survival experience using the Wilcoxon (Gehan) statistic

Survival Variable TIME

grouped by GROUP

Overall comparison statistic .012 D.F. 1 Prob. .9113

Group	label	Total N	Uncen	Cen	Pct Cen	Mean Score
1		16	8	8	50.00	.1875
2		10	3	7	70.00	-.3000

最后，系统输出生存率曲线图（图 14.2）。从图中可见，对照组（group = 2）在 8 个月前一段时点的生存率均较中药组（group = 1）略低，而 8-12 个月这一段其生存率又较中药组略高，12 个月后再又下降。但在治疗中加用中药，对个别患者而言，20 个月后依然有一定的生存率。

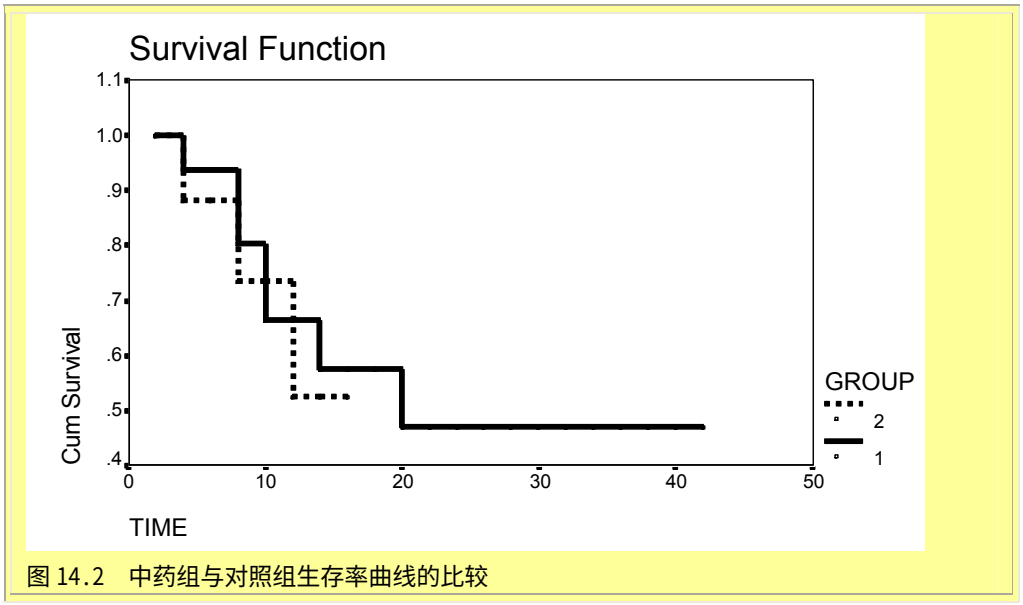


图 14.2　中药组与对照组生存率曲线的比较

第二节 Kaplan-Meier 过程

14.2.1 主要功能

调用此过程，系统将采用 Kaplan-Meier 方法，对病例随访资料进行生存分析，在对应于每一实际观察事件时点上，作生存率的评价。

14.2.2 实例操作

［例 14-2］25 例某癌症病人在不同时期经随机化分配到 A、B 治疗组进行治疗，同时随访观察至 1974 年 5 月 31 日结束，资料整理后如下表，试对其结果进行生存率分析。

病人号	随访天数	是否死亡	治疗方式
1	8	是	A
2	180	是	B
3	632	是	B
4	852	否	A
5	52	是	A
6	2240	是	B
7	220	是	A
8	63	是	A
9	195	是	B
10	76	是	B
11	70	是	B
12	8	是	A
13	13	是	B
14	1990	是	B
15	1976	是	A
16	18	否	B
17	700	否	B
18	1296	是	A

19	1460	是	A
20	210	否	B
21	63	否	A
22	1328	是	A
23	1296	是	B
24	365	是	A
25	23	否	B

14.2.2.1 数据准备

激活数据管理窗口，定义变量名：随访天数为 TIME，是否死亡为 DEATH，治疗方式为 TREAT。变量 TIME 按原数值输入，DEATH 为是的输入 1、否的输入 0，TREAT 为 A 的输入 1、为 B 的输入 2。

14.2.2.2 统计分析

激活 Statistics 菜单选 Survival 中的 Kaplan-Meier...项，弹出 Kaplan-Meier 对话框（图 14.3）。从对话框左侧的变量列表中选 time，点击 ➢ 钮使之进入 time 框；选 death，点击 ➢ 钮使之进入 Status 框，点击 Define Event...钮弹出 Kaplan-Meier:Define Event for Status Variable 对话框，在 Single value 栏中输入 1，表明 death = 1 为发生死亡事件者；点击 Continue 钮返回 Kaplan-Meier 对话框。选 treat，点击 ➢ 钮使之进入 Factor 框。

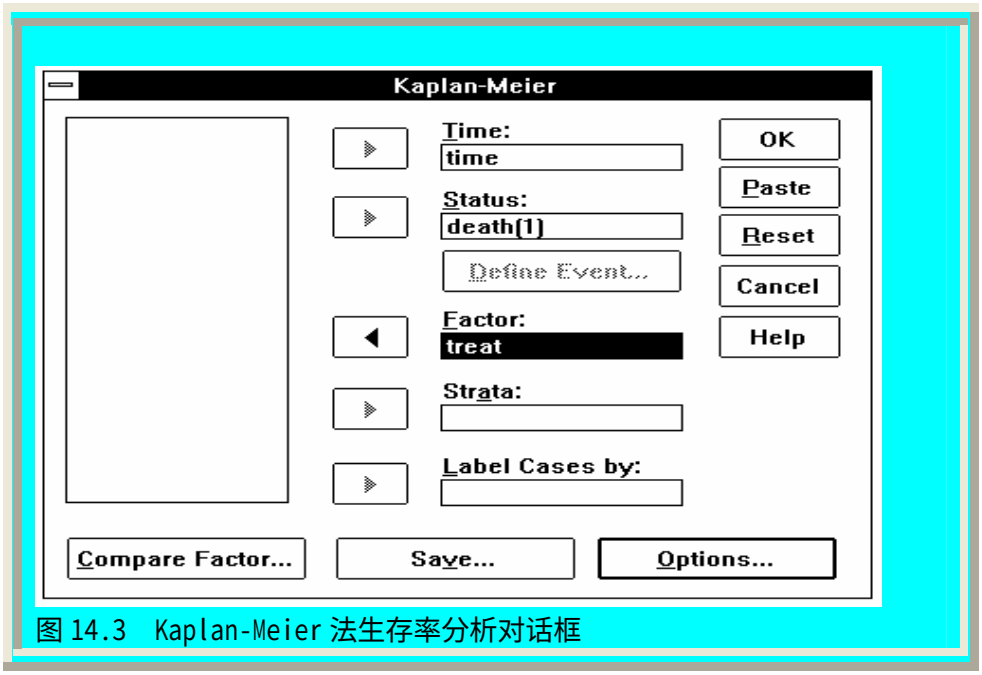


图 14.3 Kaplan-Meier 法生存率分析对话框

点击 Save... 钮弹出 Kaplan-Meier:Save New Variables 对话框，选 Survival 项，要求将各观察样例的生存率存入原始数据库中。点击 Continue 钮返回 Kaplan-Meier 对话框。

点击 Options...钮弹出 Kaplan-Meier: Options 对话框，在 Plot 栏中选 Survival 项，要求绘制生存率曲线图。之后点击 Continue 钮返回 Life Tables 对话框，再点击 OK 钮即完成分析。

14.2.2.3 结果解释

在结果输出窗口中将看到如下统计数据：

先对 A 治疗组资料进行分析。将原资料按生存天数的大小顺次排列，再逐例显示生存状态（Status，即死亡为 1、生存为 2）、生存率（Cumulative Survival）、生存率标准误（Standard Error）、累积死亡例数（Cumulative Event）和尚存活人数（Number Remaining）。如本例，A 组共 12 人，死

亡 6 人，生存 6 人，存活率为 50.00%；平均生存时间为 1023 天，标准误为 276，95%可信区间为 482—1563 天。B 组共 13 人，死亡 12 人，生存 1 人，存活率为 7.69%；平均生存时间为 607 天，标准误为 226，95%可信区间为 163—1051 天。

Factor TREAT = A

Time	Status	Cumulative Survival	Standard Error	Cumulative Events	Number Remaining
8	1			1	11
8	1	.8333	.1076	2	10
52	1	.7500	.1250	3	9
63	1			4	8
63	1	.5833	.1423	5	7
220	1	.5000	.1443	6	6
365	0			6	5
852	0			6	4
1296	0			6	3
1328	0			6	2
1460	0			6	1
1976	0			6	0

Number of Cases: 12 Censored: 6 (50.00%) Events: 6

	Survival Time	Standard Error	95% Confidence Interval
Mean:	1023	276	(482, 1563)
(Limited to	1976)		
Median:	220	.	(., .)

Factor TREAT = B

Time	Status	Cumulative Survival	Standard Error	Cumulative Events	Number Remaining
13	1	.9231	.0739	1	12
18	1	.8462	.1001	2	11
23	1	.7692	.1169	3	10
70	1	.6923	.1280	4	9
76	1	.6154	.1349	5	8
180	1	.5385	.1383	6	7
195	1	.4615	.1383	7	6
210	1	.3846	.1349	8	5
632	1	.3077	.1280	9	4
700	1	.2308	.1169	10	3
1296	1	.1538	.1001	11	2

1990	0			11	1
2240	1	.0000	.0000	12	0

Number of Cases: 13 Censored: 1 (7.69%) Events: 12

	Survival Time	Standard Error	95% Confidence Interval
Mean:	607	226	(163, 1051)
Median:	195	80	(38, 352)

		Total	Number Events	Number Censored	Percent Censored
TREAT	A	12	6	6	50.00
TREAT	B	13	12	1	7.69
Overall		25	18	7	28.00

系统按用户的请求输出生存率曲线图（图 14.4）。从图中可见，生存天数为 200 左右之前，A、B 两组的生存率相近，而后，A 组维持约 50%的生存率，B 组则不断下降。

最后系统将各观察对象的生存率计算结果，逐一送入原始数据库保存（图 14.5），变量名为 sur_1。用户从中可见，如 A 组治疗 8 天死亡者，其 8 天的生存率为 83.333%；又如 B 组治疗 180 天死亡者，其 180 天的生存率为 53.846%。

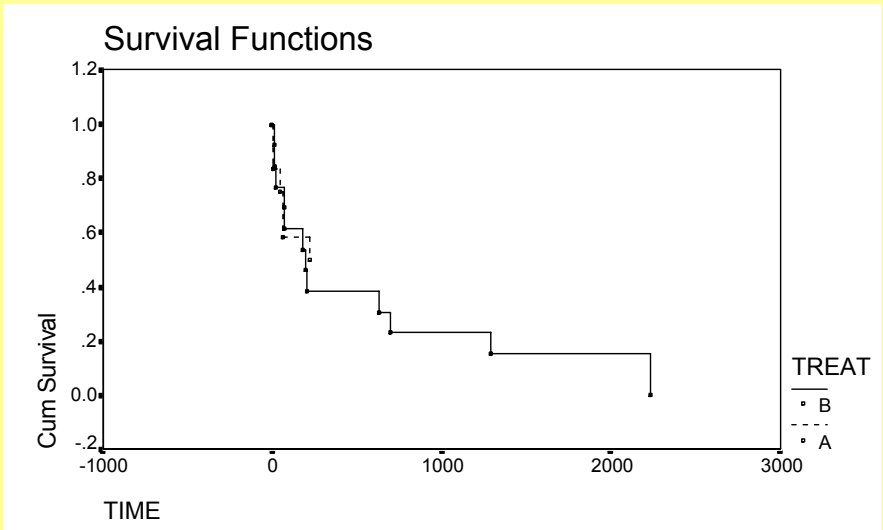
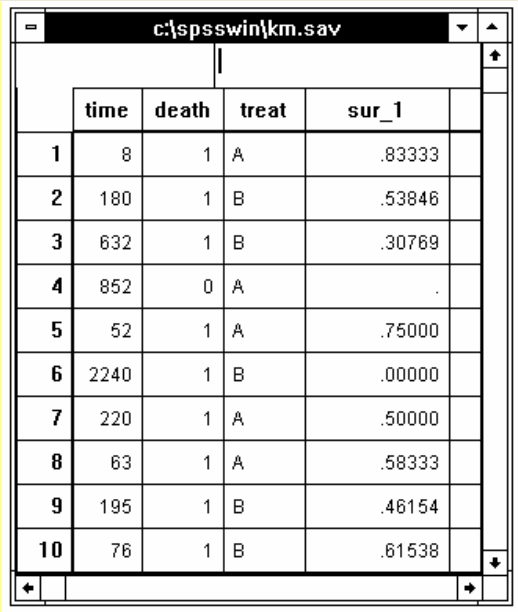


图 14.4 两种治疗方式生存率曲线比较



c:\spsswin\km.sav

	time	death	treat	sur_1
1	8	1	A	.83333
2	180	1	B	.53846
3	632	1	B	.30769
4	852	0	A	.
5	52	1	A	.75000
6	2240	1	B	.00000
7	220	1	A	.50000
8	63	1	A	.58333
9	195	1	B	.46154
10	76	1	B	.61538

图 14.5　生存率分析结果的保存

第三节 Cox Regression 过程

14.3.1 主要功能

调用此过程可完成对病例随访资料中事件发生时点与一系列相关独立变量之间关系的评价，即建立 Cox 回归模型（亦称比例风险模型）。

第一、二节介绍的方法，仅仅是对生存资料作较简单的统计，即描述和分析一个因素（如治疗方式）对生存时间的影响。而在Cox回归模型中，某一时点t，除了有一个本底风险量$h_0(t)$外，第i个影响因素可使该本底风险量$h_0(t)$增至$e^{\beta ixi}$倍而成为$h_0(t)\cdot e^{\beta ixi}$。因此如果有k个因素同时影响生存过程，那么时点t的风险量（常称之为风险函数）表达为：

$$h(t) = h_0(t) \cdot e^{(\beta 1x1+\beta 2x2+...+\beta kxk)}$$

14.3.2 实例操作

[例 14-3]某医师在研究急性白血病患者的生存率时，收集了 33 名患者的资料，按 Ag 阳、阴性分组（Ag 阳性组 17 例，Ag 阴性组 16 例），同时考察白细胞数的影响作用。试据下表资料作 Cox 回归模型的分析。

Ag 阳性组			Ag 阴性组		
生存月数	是否死亡	白细胞数（$\times 10^9$/L）	生存月数	是否死亡	白细胞数（$\times 10^9$/L）
1	是	100.0	2	是	27.0
1	是	100.0	3	是	10.0
4	是	17.0	3	是	28.0

5	是	52.0	3	是	21.0
16	是	6.0	4	否	19.0
22	是	35.0	4	是	26.0
26	否	32.0	4	是	100.0
39	是	5.4	7	是	1.5
56	是	9.4	8	是	31.0
65	否	2.3	16	是	9.0
65	是	100.0	17	是	4.0
100	是	4.3	22	是	5.3
108	是	10.5	30	是	79.0
121	是	10.0	43	是	100.0
134	否	2.6	56	是	4.4
143	是	7.0	65	是	3.0
156	是	0.8			

14.3.2.1 数据准备

激活数据管理窗口，定义变量名：生存月数为 TIME，是否死亡为 DEATH，白细胞数为 WBC，Ag 阳性与否为 AG。TIME 按原数据输入，DEATH 是的输入 1、否的输入 0，WBC 亦按原数据输入，AG 阳性的输入 1、阴性的输入 2。

14.3.2.2 统计分析

激活 Statistics 菜单选 Survival 中的 Cox Regression...项，弹出 Cox Regression 对话框（图 14.6）。从对话框左侧的变量列表中选 time，点击 ➢ 钮使之进入 time 框；选 death，点击 ➢ 钮使之进入 Status 框，点击 Define Event...钮弹出 Cox Regression:Define Event for Status Variable 对话框，在 Single value 栏中输入 1，表明 death = 1 为发生死亡事件者；点击 Continue 钮返回 Cox Regression 对话框。选 wbc 和 ag，点击 ➢ 钮使之进入 Covariates 框。

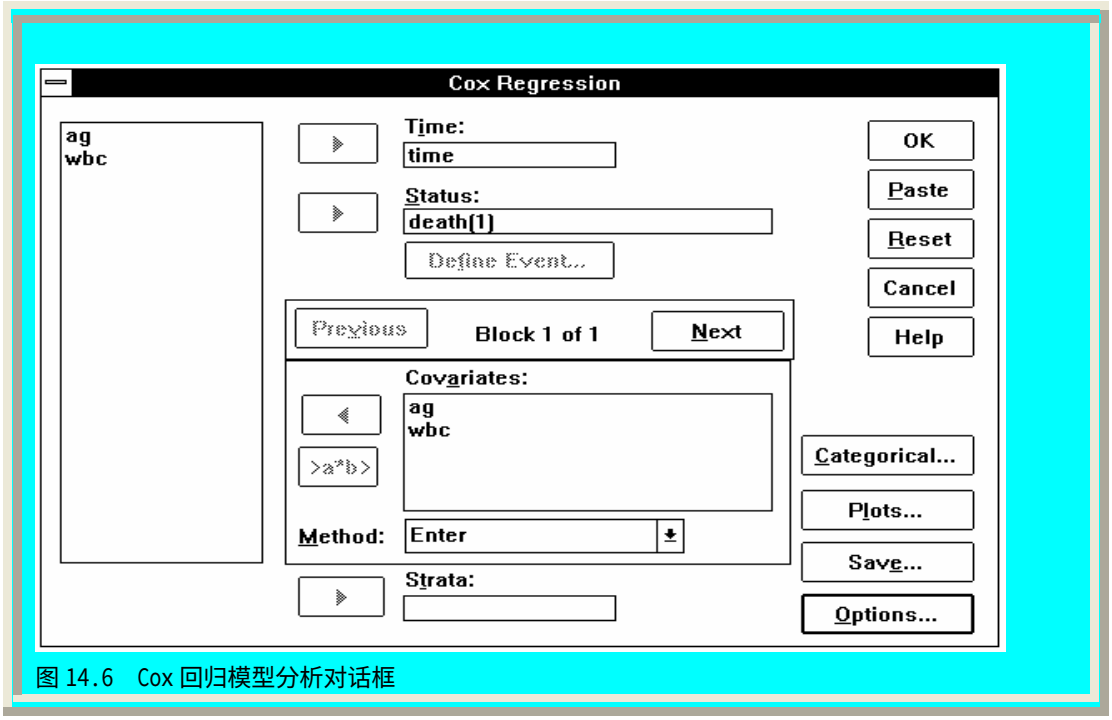


图 14.6 Cox 回归模型分析对话框

在 Method 处有一下拉菜单，系统提供 7 种回归运算方法让用户选择：

1、Enter：所有自变量强制进入回归方程；

2、Forward: Conditional：以假定参数为基础作似然比概率检验，向前逐步选择自变量；

3、Forward: LR：以最大局部似然为基础作似然比概率检验，向前逐步选择自变量；

4、Forward: Wald：作 Wald 概率统计法，向前逐步选择自变量；

5、Backward: Conditional：以假定参数为基础作似然比概率检验，向后逐步选择自变量；

6、Backward: LR：以最大局部似然为基础作似然比概率检验，向后逐步选择自变量；

7、Backward: Wald：作 Wald 概率统计法，向后逐步选择自变量。

本例因自变量较少，故选用 Enter 法。

点击 Plots...钮弹出 Cox Regression:Plots 对话框，在 Polts Type 栏中选 Survival 项，要求绘制生存率曲线图，同时选 Hazard 项，要求绘制风险量变化图。然后点击 Continue 钮返回 Cox Regression 对话框。

点击 Save... 钮弹出 Cox Regression:Save New Variables 对话框，在 Survival 栏中选 Function 项，要求将生存率计算结果存入原数据库；在 Diagnostics 处选 Hazard function 项，要求将风险函数计算结果存入原数据库；再选 X*Beta 项，要求计算各自变量与其系数的乘积并存盘。完成选择后点击 Continue 钮返回 Cox Regression 对话框。

点击 Options...钮弹出 Cox Regression:Options 对话框，在 Model Statistics 栏中选 At last step 项，要求只显示回归方程拟合过程的最终结果；同时选 Display baseline function 项，要求显示各样本的本底风险量。之后点击 Continue 钮返回 Cox Regression 对话框，再点击 OK 钮即完成分析。

14.3.2.3 结果解释

在结果输出窗口中将看到如下统计数据：

系统显示共有 33 例样本进入分析，其中 29 例在观察期内死亡，4 例仍存活，存活率为 12.1%。Cox回归方程拟合结果的 χ^2检验，χ^2值为 11.773，P = 0.0028，表明AG与WBC两自变量对生存状态均有作用。得到风险量增加倍数为$e^{(0.0089\times WBC-1.1219\times AG)}$，其中白细胞数的变量系数为正值，意味着白细胞数愈高，死亡风险愈大；Ag的变量系数为负，意味着Ag阳性者，死亡风险小。

```
33    Total cases read
       0    Cases with missing values
       0    Valid cases with non-positive times
       0    Censored cases before the earliest event in a stratum
       0    Total cases dropped
      33    Cases available for the analysis

Dependent Variable:    TIME

          Events          Censored
            29           4 (12.1%)

Beginning Block Number 0.    Initial Log Likelihood Function

-2 Log Likelihood          153.394
```

Beginning Block Number 1. Method: Enter

Variable(s) Entered at Step Number 1..
AG
WBC

-2 Log Likelihood 142.761

	Chi-Square	df	Sig
Overall (score)	11.773	2	.0028
Change (-2LL) from			
Previous Block	10.633	2	.0049
Previous Step	10.633	2	.0049

---------- Variables in the Equation ----------

Variable	B	S.E.	Wald	df	Sig	R	Exp(B)
AG	-1.1219	.4505	6.2025	1	.0128	-.1655	.3256
WBC	.0089	.0052	2.9703	1	.0848	.0795	1.0089

接着，系统显示各生存时点（亦即各样本）的本底风险量（Cum Hazard）、生存率（Survival）、生存率的标准误（SE）和本底风险量的标准误（SE of Cum hazard）。并提示将在原数据库中产生三个新的变量，即生存率、风险比例和风险量倍数。

	Baseline	---- At mean of covariates ---		
Time	Cum Hazard	Survival	SE	SE of Cum Hazard
1	.0701	.9503	.0338	.0510
2	.1069	.9252	.0415	.0778
3	.2286	.8468	.0574	.1663
4	.3730	.7623	.0691	.2714
5	.4305	.7311	.0738	.3132
7	.4907	.6998	.0776	.3570
8	.5555	.6676	.0809	.4041
16	.6977	.6020	.0858	.5075
17	.7755	.5688	.0877	.5642
22	.9484	.5016	.0895	.6899
30	1.0552	.4641	.0901	.7677
39	1.1770	.4248	.0910	.8562
43	1.3251	.3814	.0894	.9640
56	1.7359	.2829	.0829	1.2628
65	2.3579	.1799	.0679	1.7153
100	2.8882	.1223	.0635	2.1011

108	3.5394	.0762	.0521	2.5748
121	4.3891	.0411	.0369	3.1929
143	6.4697	.0090	.0133	4.7066
156	.	.0000	.	.

Covariate Means

Variable	Mean
AG	.5152
WBC	29.1667

3 New variables have been added:

Name	Label
SUR_1	Survival Function
HAZ_1	Cumulative Hazard Function
XBE_1	X'Beta

从输出的 Cox 模型生存率曲线图（图 14.7）中可见，随时间的延长，患者生存率逐渐下降，接近 160 个月时，生存率几乎为 0。

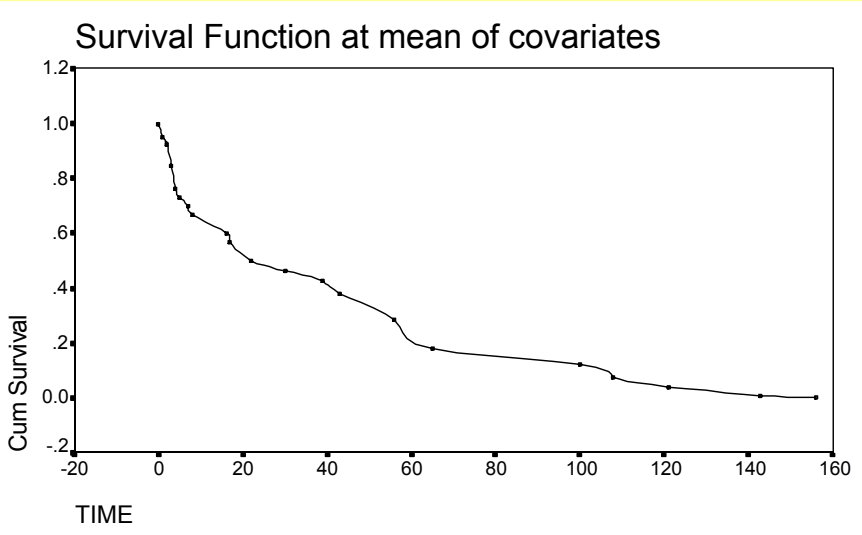


图 14.7　Cox 模型的生存率曲线

下图为 Cox 模型的风险量曲线图，其趋势也十分明显，即随时间的延长，患者在生存上所经历的死亡风险愈来愈大，到 140 个月时，大约是起初（0 个月）的 5 倍。

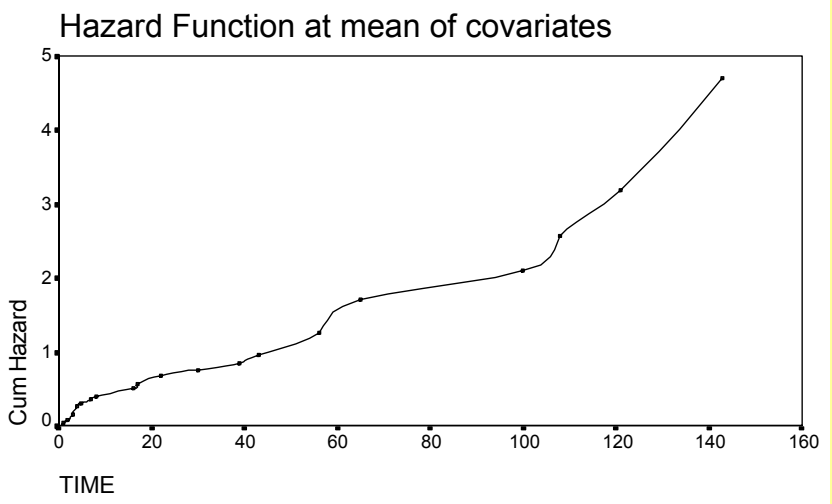


图 14.8　Cox 模型的风险量曲线

系统在原始数据库中将生存率以变量sur_1、风险比例以变量haz_1 和风险量倍数以变量xbe_1 存盘（图 14.9）。用户从中可见，如经治疗后 1 月内死亡、其白细胞数为 100.0×10^9/L、Ag阳性者，生存率为 94.592%、风险比例为 5.560%、风险量倍数为 0.08695；又如经治疗后 26 月内尚存活、其白细胞数为 32.0×10^9/L、Ag阳性者，生存率为 66.318%、风险比例为 41.070%、风险量倍数为-0.51874。

c:\spsswin\cox.sav

	time	death	wbc	ag	sur_1	haz_1	xbe_1
1	1	1	100.0	1	.94592	.05560	.08695
2	1	1	100.0	1	.94592	.05560	.08695
3	4	1	17.0	1	.86820	.14133	-.65235
4	5	1	52.0	1	.80027	.22280	-.34059
5	16	1	6.0	1	.78690	.23966	-.75032
6	22	1	35.0	1	.65585	.42182	-.49202
7	26	0	32.0	1	.66318	.41070	-.51874
8	39	1	5.4	1	.66887	.40217	-.75567
9	56	1	9.4	1	.54083	.61465	-.72004
10	65	0	2.3	1	.45671	.78372	-.78328

图 14.9　Cox 回归模型分析结果的保存

第十五章 统计图的绘制

统计图是用点的位置、线段的升降、直条的长短或面积的大小等来表达资料的内容。它可以把资料所反映的变化趋势、数量多少、分布状态和相互关系等形象直观地表现出来，以便于读者的阅读、比较和分析。

本章将介绍 SPSS 在绘制常用统计图方面的功能。由于计算机绘图具有快速、清晰、规范、可修正以保证准确无误等特点，故在论文、报告等写作中有着十分重要的应用价值。

第一节 直条图

15.1.1 主要功能

调用 Graphs 菜单的 Bar 过程，可绘制直条图。直条图用直条的长短来表示非连续性资料（该资料可以是绝对数，也可以是相对数）的数量大小。

15.1.2 实例操作

[例 15-1]研究血压状态与冠心病各临床型发生情况的关系，分析资料如下所示，试绘制统计图。

血压状态	年龄标化发生率（1/10 万）			
	冠状动脉机能不全	猝死	心绞痛	心肌梗塞
正常	8.90	12.00	34.71	44.00
临界	10.63	18.05	46.18	67.24
异常	19.84	30.55	73.06	116.82

15.1.2.1 数据准备

激活数据管理窗口，定义变量名：年龄标化发生率为 RATE，冠心病临床型为 DISEASE，血压状态为 BP。RATE 按原数据输入，DISEASE 按冠状动脉机能不全=1、猝死=2、心绞痛=3、心肌梗塞=4 输入，BP 按正常=1、临界=2、异常=3 输入。

15.1.2.2 操作步骤

选 Graphs 菜单的 Bar...过程，弹出 Bar Chart 定义选项框（图 15.1）。在定义选项框的下方有一数据类型栏，系统提供 3 种数据类型：

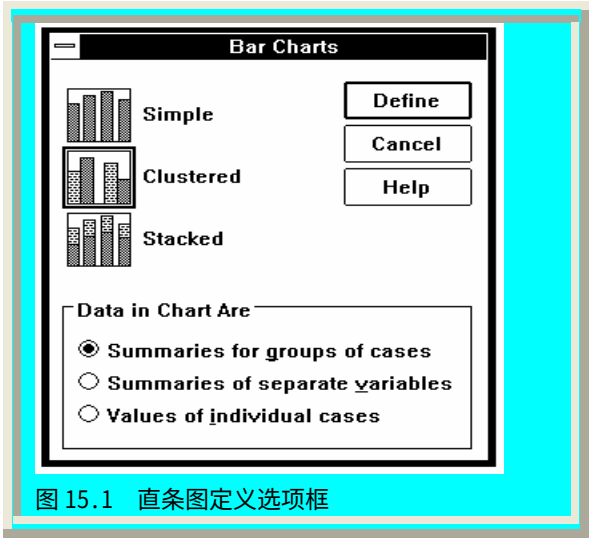


图 15.1　直条图定义选项框

Summaries for groups of cases：以组为单位体现数据；

Summaries of separate variables：以变量为单位体现数据；

Values of individual cases：以观察样例为单位体现数据。

大多数情形下，统计图都是以组为单位的形式来体现数据的。在定义选项框的上方有 3 种直条图可选：Simple 为单一直条图、Clustered 为复式直条图、Stacked 为堆积式直条图，本例选复式直条图。

点击 Define 钮，弹出 Define Clustered Bar:Summaries for Groups of Cases 对话框（图 15.2），在左侧的变量列表中选 rate 点击 ➢ 钮使之进入 Bars Represent 栏的 Other snmmary function 选项的 Variable 框，选 disease 点击 ➢ 钮使之进入 Category Axis 框，选 bp 点击 ➢ 钮使之进入 Define Clusters by 框。

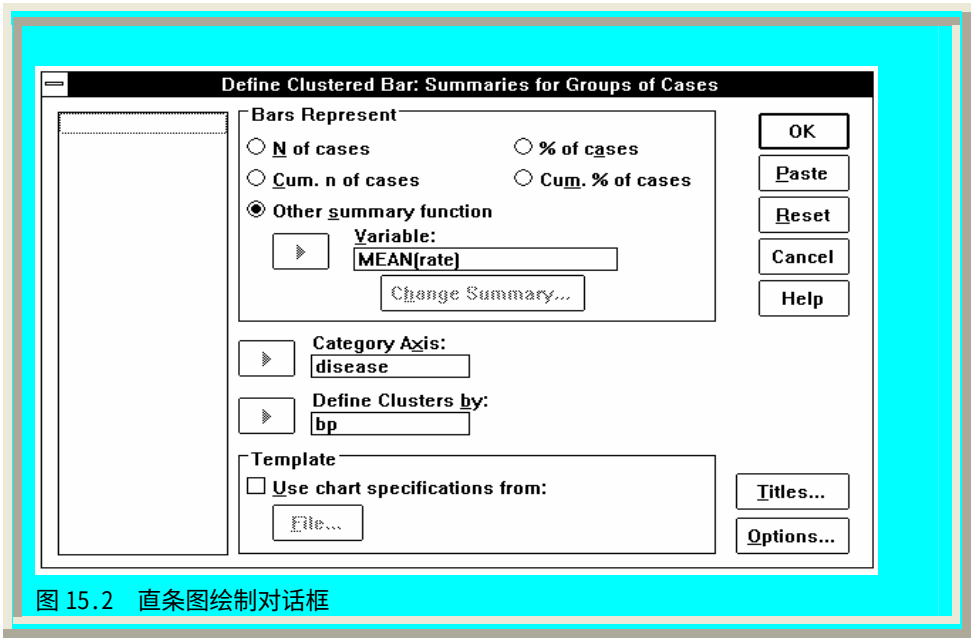


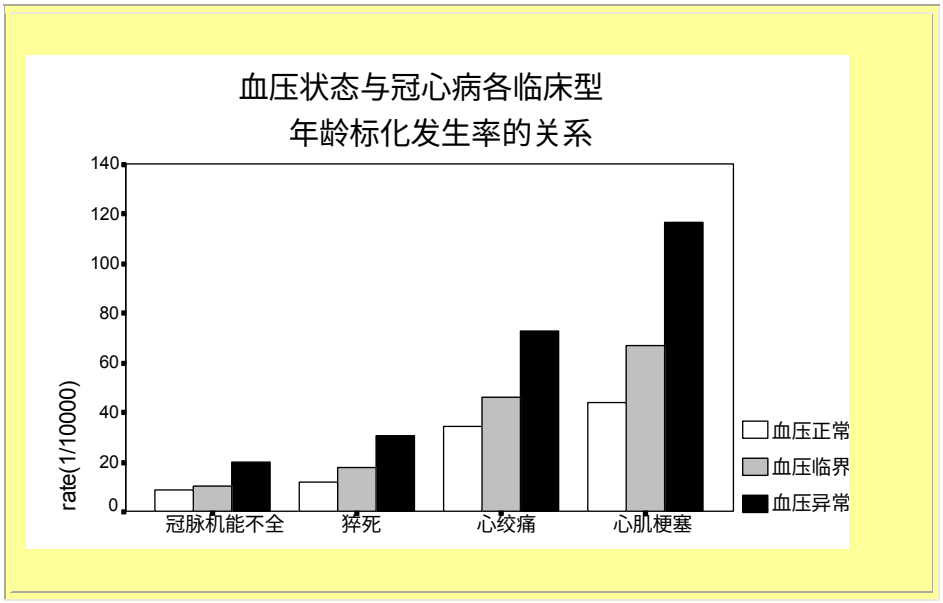
图 15.2　直条图绘制对话框

点击 Titles...钮，弹出 Titles 对话框，在 Title 栏内输入“血压状态与冠心病各临床型年龄标化发生率的关系”，点击 Continue 钮返回 Define Clustered Chart:Summaries for Groups of Cases 对话框，再点击 OK 钮即完成。

系统在统计图编辑窗口中输出直条图。由于在原始数据库中，为了输入的方便，分组采用简单的 1、2、3……等数字表示，故体现在统计图中的分组条目会让读者感到不理解。为此，用户可点击窗口上端工具栏中的 Edit 钮，对统计图进行编辑。用户欲在图中的哪一部位（如：标题、纵横轴的尺度与标目、统计图的色彩或花纹，等等）进行编辑，只须将鼠标箭头指向这一部位并双击鼠标左键，系统即弹出相应的编辑对话框。编辑过程简便易行，用户不妨一试。本章对此内容的介绍从略。

15.1.2.3 结果显示

下图为经编辑（主要是将分组的标目改为中文）后血压状态与冠心病各临床型年龄标化发生率关系的直条图。从图中可见，冠心病各临床型的发生率以冠状动脉机能不全最低、心肌梗塞最高；随血压的升高，疾病发生率升高；异常血压对心肌梗塞发生的影响作用大于其他临床型。



第二节 线图

15.2.1 主要功能

调用 Graphs 菜单的 Line 过程，可绘制线图。线图是用线条的上下波动形式，反映连续性的相对数资料的变化趋势。非连续性的资料一般不用线图表现。

15.2.2 实例操作

[例 15-2]某地调查居民心理问题的存在现状，资料如下表所示，试绘制线图比较不同性别和年龄组的居民心理问题检出情况。

年龄分组	心理问题检出率（%）	
	男性	女性

15-	10.57	19.73
25-	11.57	11.98
35-	9.57	15.50
45-	11.71	13.85
55-	13.51	12.91
65-	15.02	16.77
75-	16.00	21.04

15.2.2.1 数据准备

激活数据管理窗口，定义变量名：心理问题检出率为 RATE，年龄分组为 AGE，性别为 SEX，AGE 与 SEX 可定义为字符变量。RATE 按原数据输入，AGE 按分组情况分别输入 15-、25-、35-、45-、55-、65-、75-，SEX 是男的输入 M、女的输入 F。

15.2.2.2 操作步骤

选 Graphs 菜单的 Line...过程，弹出 Line Chart 定义选项框，有 3 种线图可选：Simple 为单一线图、Multiple 为多条线图、Drop-line 为落点线图，本例选多条线图。

点击 Define 钮，弹出 Define Multiple Line:Summaries for Groups of Cases 对话框（图 15.3），在左侧的变量列表中选 rate 点击 ➢ 钮使之进入 Lines Represent 栏的 Other snmmary function 选项的 Variable 框，选 age 点击 ➢ 钮使之进入 Category Axis 框，选 sex 点击 ➢ 钮使之进入 Define Lines by 框。

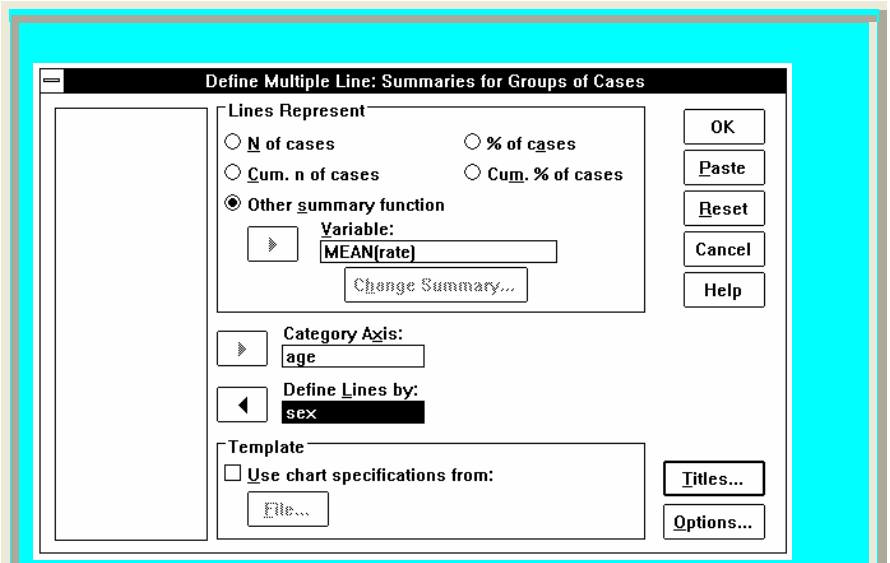
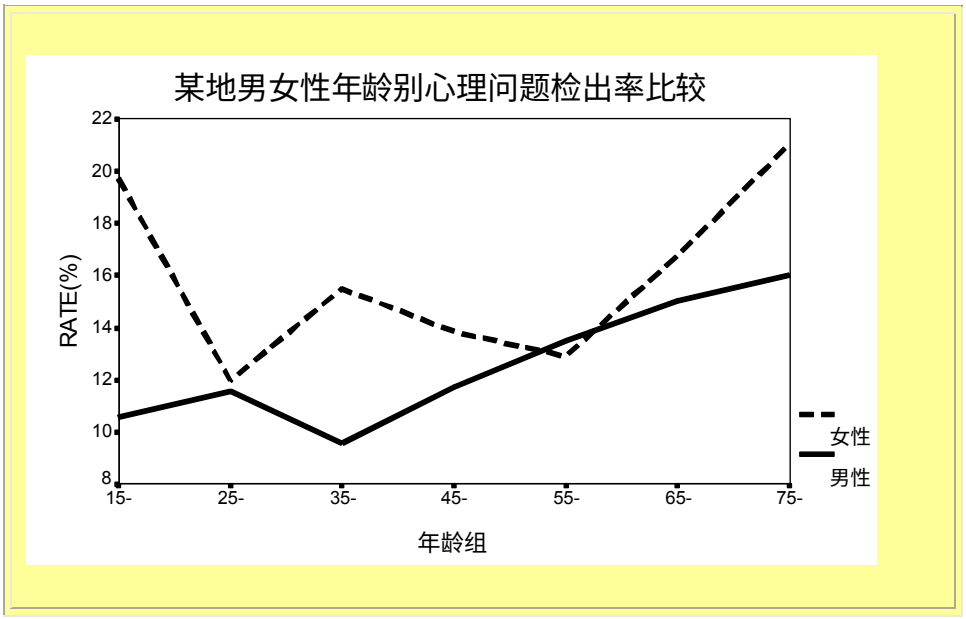


图 15.3 线图绘制对话框

点击 Titles...钮，弹出 Titles 对话框，在 Title 栏内输入“某地男女性年龄别心理问题检出率比较”，点击 Continue 钮返回 Define Multiple Line:Summaries for Groups of Cases 对话框，再点击 OK 钮即完成。

15.2.2.3 结果显示

下图即为系统输出的线图，分析表明，15-岁组和 65-岁以上组的心理问题检出率较其他年龄组为高，女性的心理问题检出率较男性为高。



第三节 区域图

15.3.1 主要功能

调用 Graphs 菜单的 Area 过程，可绘制区域图。实际上区域图是用面积来表现连续性的频数分布资料，面积越大，频数越多，反之亦然。

15.3.2 实例操作

[例 15-3]在某城市抽样研究 20-49 岁已婚育龄妇女的避孕现状，频数分布资料参见下表，试绘制区域图。

年龄分组	避孕现状	
	有	无
20-	63	68
25-	939	184
30-	1860	273
35-	1277	91
40-	1141	173
45-	987	399

15.3.2.1 数据准备

激活数据管理窗口，定义变量名：避孕有无的人数为 NUMBER，年龄分组为 AGE，避孕现状为 CONTRA，AGE 与 CONTRA 可定义为字符变量。NUMBER 按实际人数输入（有无避孕的人数全部输入变量

NUMBER 中），AGE 按分组情况分别输入 20-、25-、30-、35-、40-、45-，CONTRA 有的输入 Y、无的输入 N。

15.3.2.2 操作步骤

选 Graphs 菜单的 Area...过程，弹出 Area Chart 定义选项框，有 2 种线图可选：Simple 为简单区域图、Stacked 为堆积区域图，本例选堆积区域图。

点击 Define 钮，弹出 Define Stacked Area:Summaries for Groups of Cases 对话框（图 15.4），在左侧的变量列表中选 number 点击 ➢ 钮使之进入 Areas Represent 栏的 Other snmmary function 选项的 Variable 框，选 age 点击 ➢ 钮使之进入 Category Axis 框，选 contra 点击 ➢ 钮使之进入 Define Areas by 框。

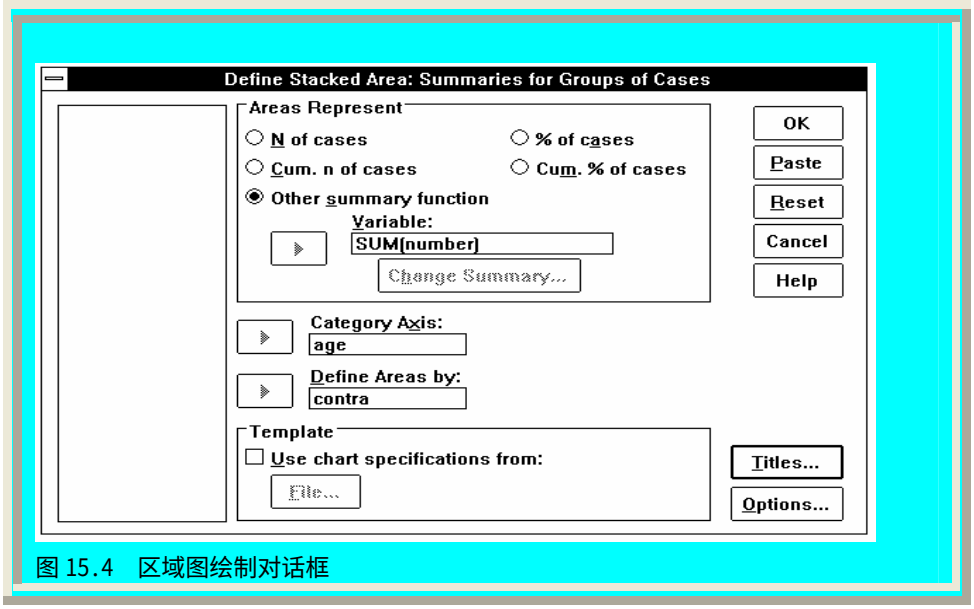
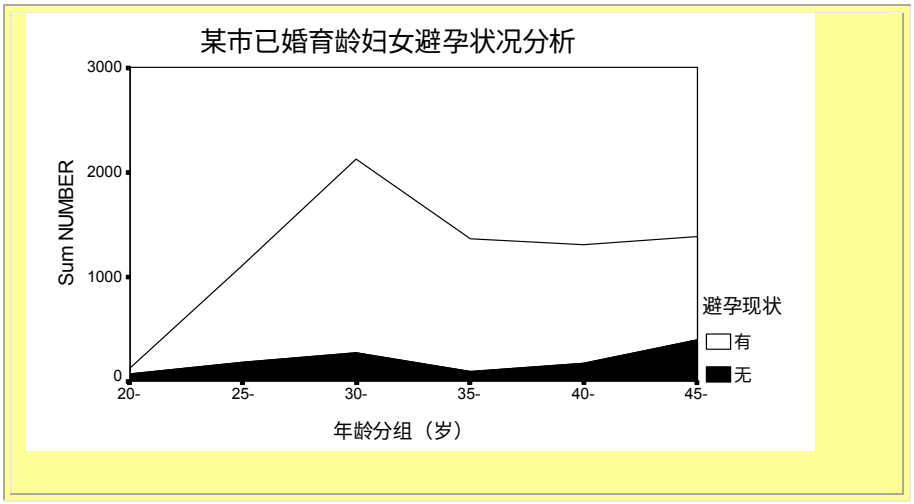


图 15.4 区域图绘制对话框

点击 Titles...钮，弹出 Titles 对话框，在 Title 栏内输入“某市已婚育龄妇女避孕状况分析”，点击 Continue 钮返回 Define Stacked Area:Summaries for Groups of Cases 对话框，再点击 OK 钮即完成。

15.3.2.3 结果显示

下图显示：年轻妇女（25 岁之前）有避孕人数与无避孕人数差不多，25 岁之后，有避孕人数占绝大多数，而 45 岁以后，无避孕人数又开始增加。



第四节 构成图

15.4.1 主要功能

调用 Graphs 菜单的 Pie 过程，可绘制构成图。构成图也称馅饼图，用一个圆来表现百分构成，读者可根据圆中各个扇形面积的大小，判断某一部分在全部中所占比例的多少。

15.4.2 实例操作

[例 15-4]某年某医院用中草药治疗 182 例慢性支气管炎患者，其疗效如下所示，试绘制构成图。

疗效	病例数	百分构成（%）
控制	37	20.3
显效	71	39.0
好转	60	33.0
无效	14	7.7
合计	182	100.0

15.4.2.1 数据准备

激活数据管理窗口，定义变量名：百分构成资料为 DATA，构成部分的名称为 TEXT，TEXT 定义为字符变量。DATA 按实际百分数输入，TEXT 依次输入 1、2、3、4。

15.4.2.2 操作步骤

选 Graphs 菜单的 Pie...过程,弹出 Pie Chart 定义选项框,构成图仅有一种,故直接点击 Define 钮，弹出 Define Pie:Summaries for Groups of Cases 对话框（图 15.5），在左侧的变量列表中选 data 点击 ➢ 钮使之进入 Slices Represent 栏的 Other snmmary function 选项的 Variable 框，选 text 点击 ➢ 钮使之进入 Define Slices by 框。点击 Titles...钮，弹出 Titles 对话框，在 Title

栏内输入“中草药治疗慢性支气管炎效果构成图”，点击 Continue 钮返回 Define Pie:Summaries for Groups of Cases 对话框，再点击 OK 钮即完成。

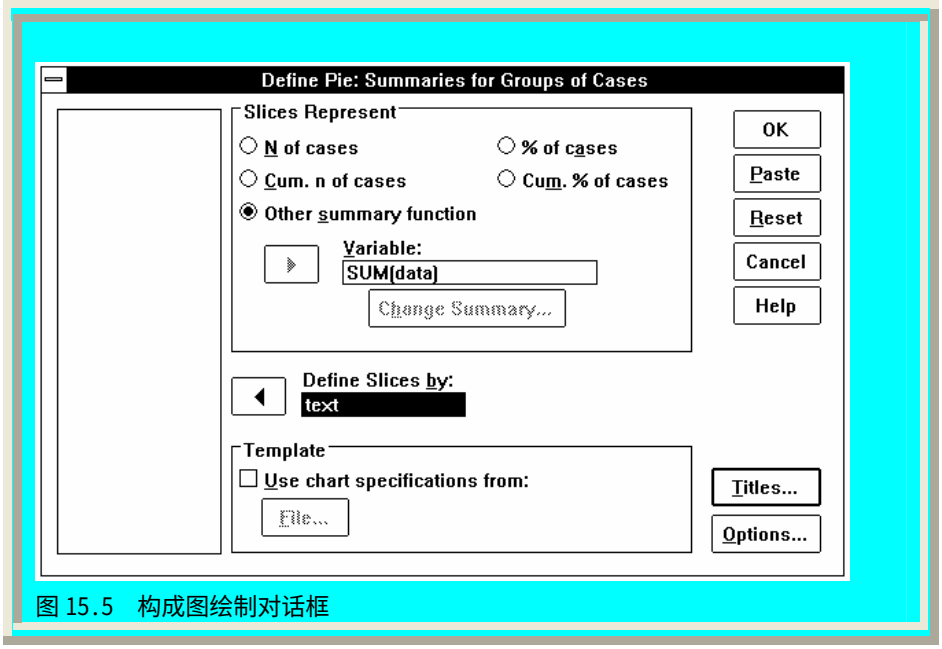
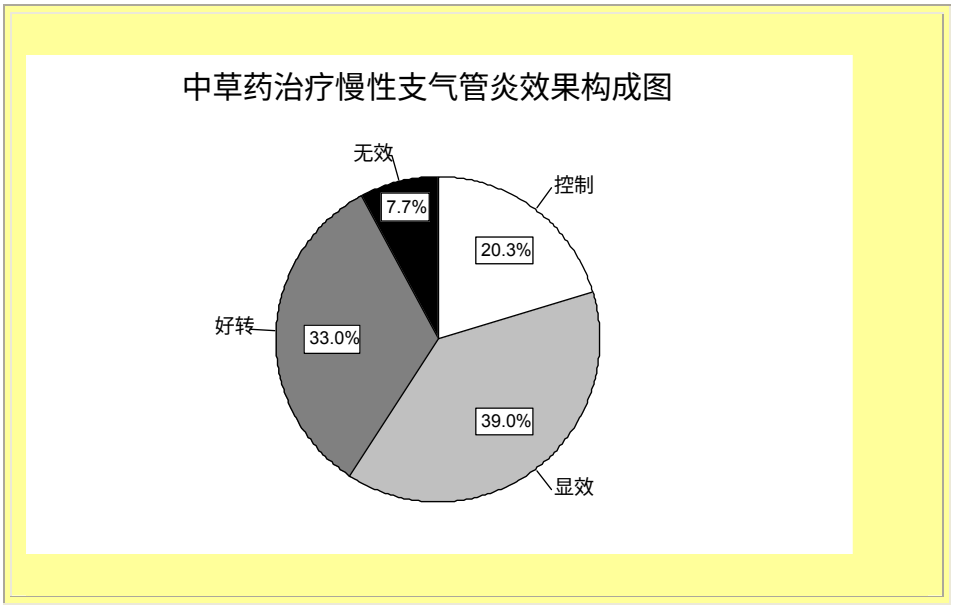


图 15.5　构成图绘制对话框

15.4.2.3　结果显示

下图显示：该中草药效果良好，无效的比例很小。



第五节 高低区域图

15.5.1 主要功能

调用Graphs菜单的High-Low过程，可绘制高低区域图。高低区域图用于表现多种形式的数据区域，如一组测定值的范围（最小值—最大值）、95%可信区间值（低限—高限）、$\overline{X}$ ±1.96·SD（低值—均值—高值）等，形象直观。

15.5.2 实例操作

[例 15-5]为了解水体污染情况，某市测定三种水源中放射性元素锶(^{90}Sr)的含量($10^{-2}Bq \cdot L^{-1}$)，资料如下，试绘制高低区域图。

水源点	范围	均值
自来水	0.65～0.93	0.79
湖　水	1.31～2.11	1.71
水库水	1.01～2.16	1.58

15.5.2.1 数据准备

激活数据管理窗口，定义变量名：数据的变量名为 DATA，将范围的低值与高值以及均值一并输入；设一变量为 CAT，用于定义低值、高值和均值，低值为 1、高值为 2、均值为 3；水源点变量名为 GROUP，依次输入 1、2、3。

15.5.2.2 操作步骤

选 Graphs 菜单的 High-Low...过程，弹出 High-Low Chart 定义选项框，高低区域图有 5 种，即：

Simple High-Low-Close：简单线型高低区域图；

Clustered High-Low-Close：复式线型高低区域图；

Simple Range Bar：简单直条型高低区域图；

Clustered Range Bar：复式直条型高低区域图；

Difference Line：差异线区域图。

本例选用简单线型高低区域图。然后点击 Define 钮，弹出 Define Simple High-Low-Close:Summaries for Groups of Cases 对话框（图 15.6），在左侧的变量列表中选 data 点击 ➢ 钮使之进入 Bars Represent 栏的 Other snmmary function 选项的 Variable 框，选 cat 点击 ➢ 钮使之进入 Category Axis 框，选 group 点击 ➢ 钮使之进入 Define High-Low-Close by 框。点击 Titles...钮，弹出 Titles 对话框，在 Title 栏内输入“某市测定不同水体放射性元素锶的含量比较”，点击 Continue 钮返回 Define Simple High-Low-Close:Summaries for Groups of Cases 对话框，再点击 OK 钮即完成。

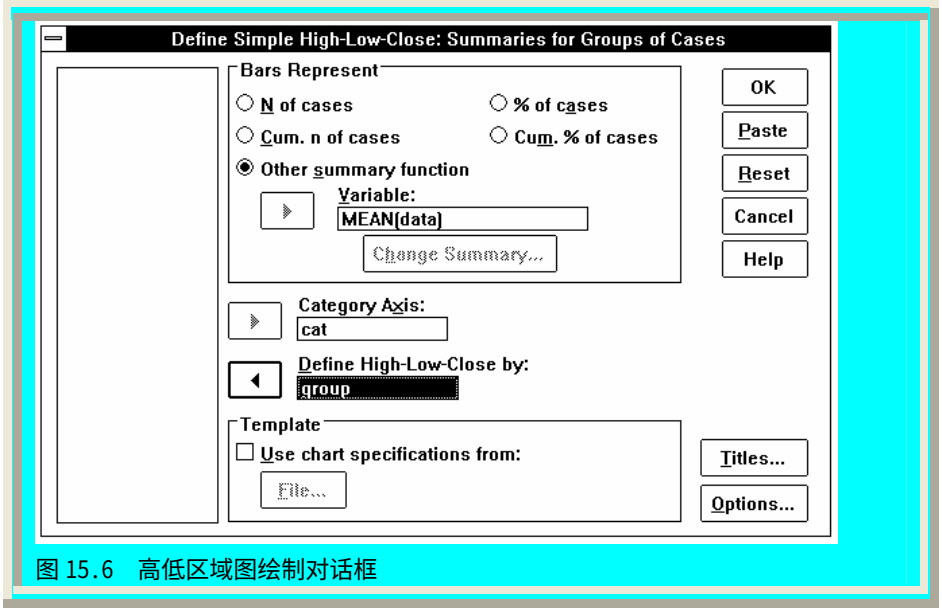
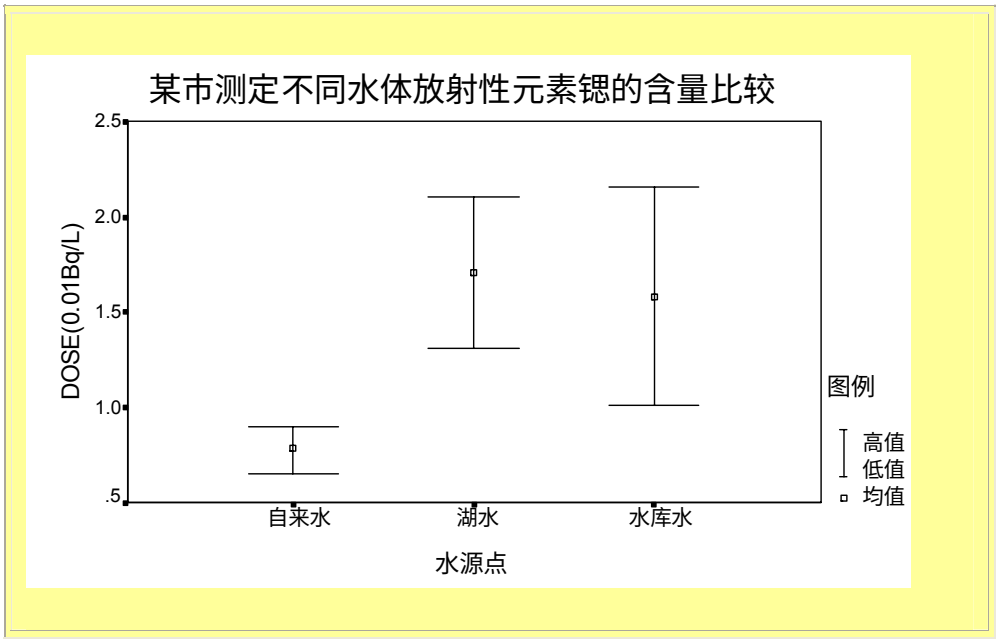


图 15.6　高低区域图绘制对话框

15.5.2.3　结果显示

下图显示放射性元素锶的含量在湖水中最高、在自来水中最低，但水库水中其含量不仅高而且变化幅度最大。



第六节　直条构成线图

15.6.1　主要功能

调用 Graphs 菜单的 Pareto 过程，可绘制直条构成线图（又称佩尔托图）。直条构成线图是直条图与构成图的结合，它用直条的长短表现各组绝对数的多少，同时用线段的逐渐上升趋势表现各组

百分构成比接近 100.00%的过程。

15.6.2 实例操作

[例 15-6]随访 1000 名 20-25 岁的男性一年，分季节考察其感冒发生情况，结果如下，试绘制直条构成线图。

季节	病例数	百分构成（%）
春	443	39.80
夏	104	9.35
秋	379	34.05
冬	187	16.80
合计	1113	100.00

15.6.2.1 数据准备

激活数据管理窗口，定义变量名：各季节病例数的变量名为 DATA，输入具体数字；季节的变量名为 SEASON，依次输入 1、2、3、4。百分构成不必建立变量，也不必输入数据，系统会自动生成。

15.6.2.2 操作步骤

选 Graphs 菜单的 Pareto...过程，弹出 Pareto Chart 定义选项框，有 2 种直条构成线图可选：Simple 为单一直条构成线图，Stacked 为堆积式直条构成线图，本例选用单一直条构成线图。然后点击 Define 钮，弹出 Define Simple Pareto:Summaries for Groups of Cases 对话框（图 15.7），在左侧的变量列表中选 data 点击 ➢ 钮使之进入 Sums of variable 框，选 season 点击 ➢ 钮使之进入 Category Axis 框。点击 Titles...钮，弹出 Titles 对话框，在 Title 栏内输入“1000 名 20-25 岁男性各季节感冒发生人数分析”，点击 Continue 钮返回 Define Simple Pareto:Summaries for Groups of Cases 对话框，再点击 OK 钮即完成。

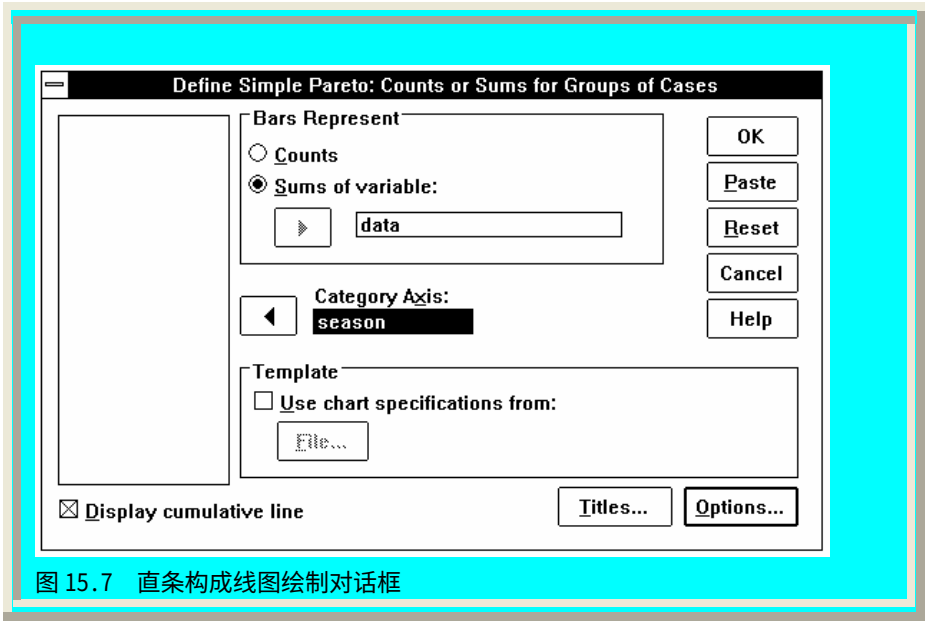
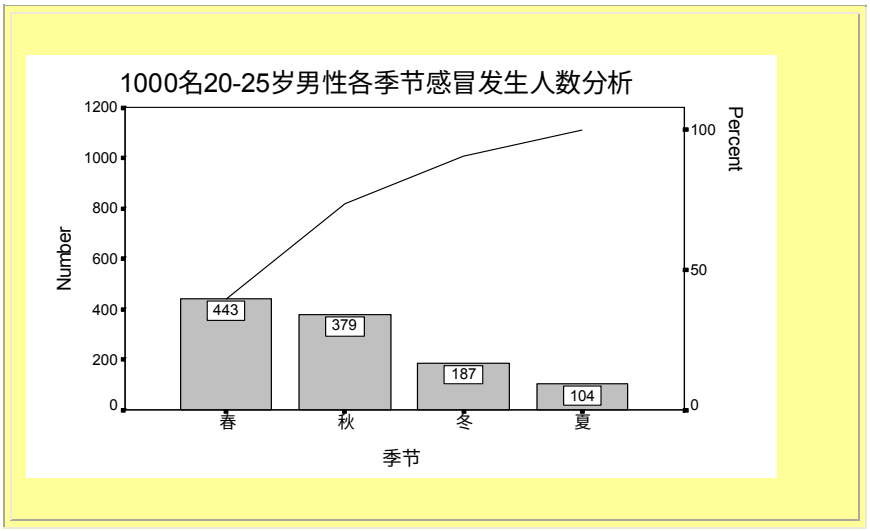


图 15.7 直条构成线图绘制对话框

15.6.2.3 结果显示

下图显示，春秋季的感冒病例比其他季节多，仅春季的病例数已接近全年病例数的一半。夏季感冒病例最少，占全年病例数的比例不到 10%。



第七节 质量控制图

15.7.1 主要功能

调用 Graphs 菜单的 Control 过程，可绘制质量控制图。质量控制图是进行质量控制的常用工具，可提示工作过程中所发生的变化及其趋势，从而提醒人们的警觉与注意，以便分析原因、采取解决对策。

15.7.2 实例操作

[例 15-7]对一种标准试液中某物质含量测平行样 5 次，结果如下，试绘制质量控制图以便对准确度与精确度进行评价。

测定次序	平行样		均数	极差
	第一次	第二次		
1	10.4	10.1	10.25	0.3
2	10.8	11.0	10.90	0.2
3	9.8	10.4	10.10	0.6
4	9.4	11.0	10.20	1.6
5	10.1	11.3	10.70	1.2

15.7.2.1 数据准备

激活数据管理窗口，定义变量名：平行样数据的变量名为 DATA，将测定数据一并输入；设一变

量为 GROUP，用于定义测定次序，依次输入 1、2、3、4、5。均数和极差的数据不必输入，系统会自动生成。

15.7.2.2 操作步骤

选 Graphs 菜单的 Control...过程，弹出 Control Chart 定义选项框，有 5 种质量控制图可选：

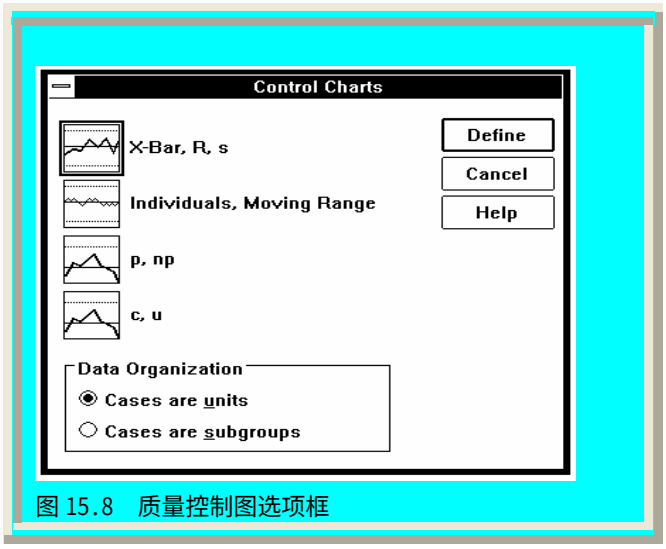


图 15.8 质量控制图选项框

X-Bar, R, s：均数控制图和极差（标准差）控制图。均数控制图又称 $\overline{X}$ 图，用于控制重复测定的准确度；极差控制图又称R图，用于控制例数较少时重复测定的精确度；标准差控制图又称s图，用于控制例数较多时重复测定的精确度。

Individuals, Moving Range：个值控制图。根据容许区间的原理绘制，适用于单个测定值的控制。

p, np：率的控制图。根据率的二项分布原理绘制，适用于率的控制。

c, u：数量控制图。根据组中非一致测定值绘制，各组例数相等时用 u 图，不相等时用 c 图，适用于属性资料的质量控制。

本例选用 X-Bar, R, s。选项框的下方为数据类型选择栏（Data Organization），Cases are units 表示数据文件中各观察样例只是一个值，其分组需要再定义；Cases are subgroups 表示数据文件中各观察样例本身就是一个组。

点击 Define 钮，弹出 X-Bar, R, s:Cases Are Units 对话框（图 15.9），在左侧的变量列表中选 data 点击➢钮使之进入 Process Measurement 框，选 group 点击➢钮使之进入 Subgroups Defined by 框。因本例样品少，故在 Charts 栏中选 X-Bar and range 项，要求输出均数控制图和极差控制图。点击 Titles...钮，弹出 Titles 对话框，在 Title 栏内输入“样品测定的质量控制图”，点击 Continue 钮返回 X-Bar, R, s: Cases Are Units 对话框，再点击 OK 钮即完成。

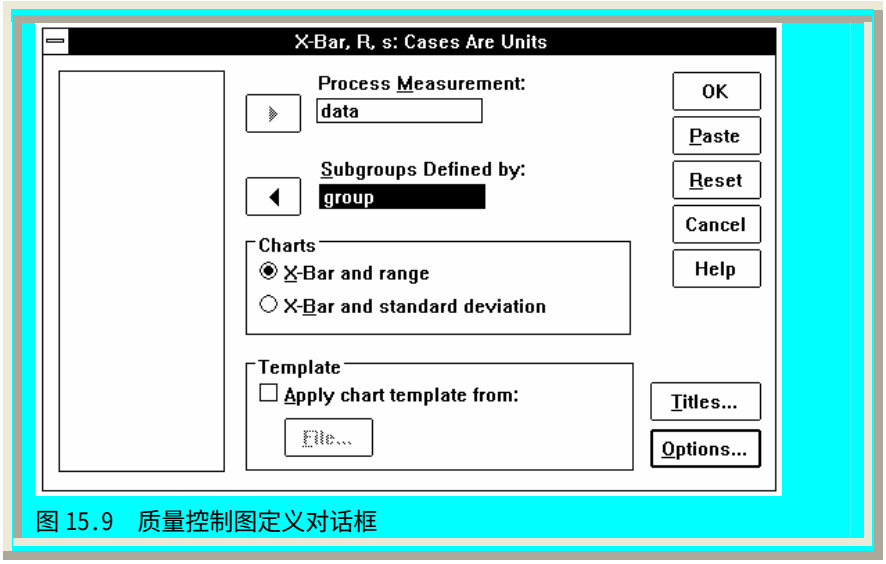
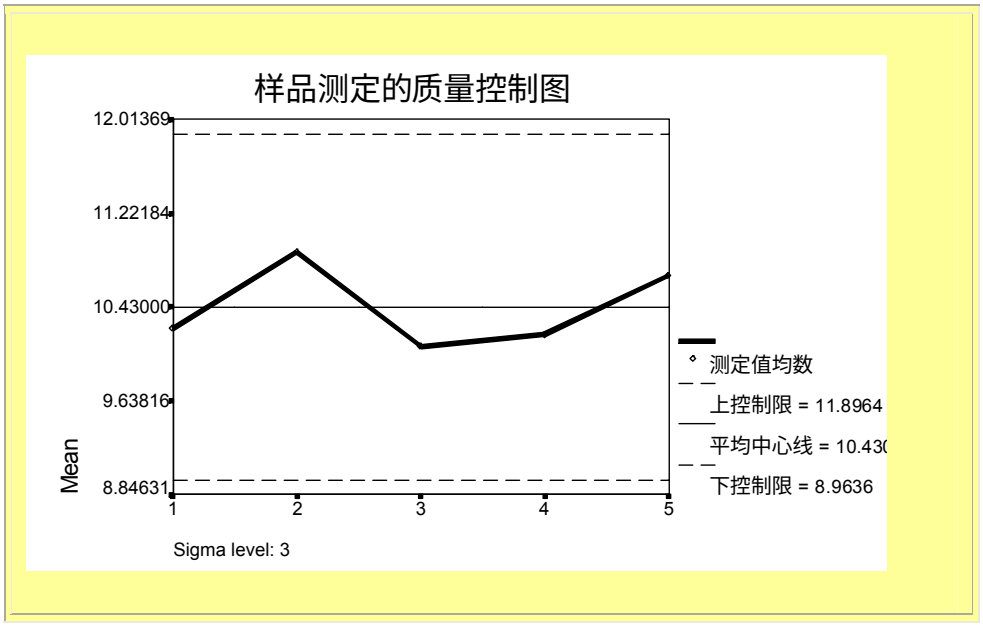
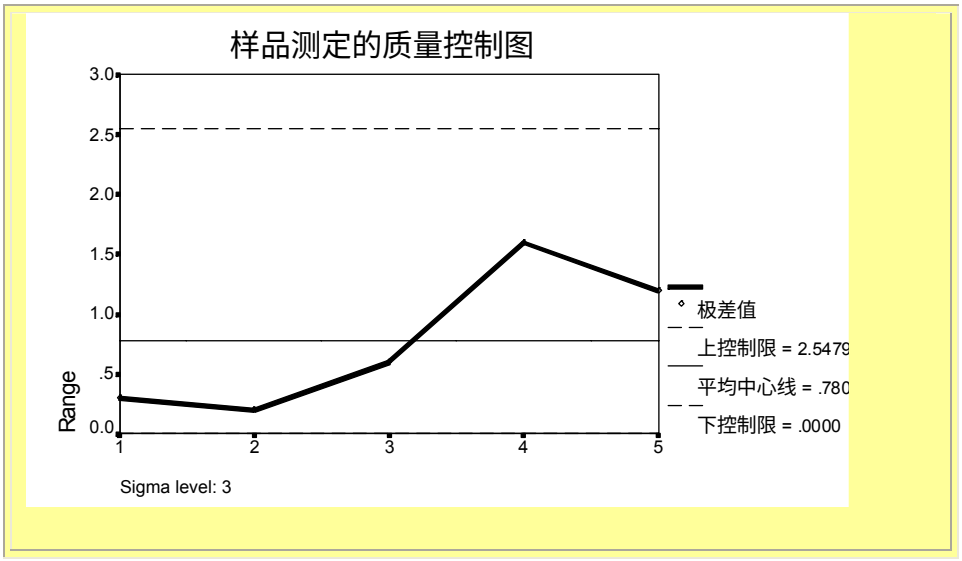


图 15.9　质量控制图定义对话框

15.7.2.3　结果显示

系统输出两张图，第一张为均数控制图，平均中心线的值为五组均数的平均值，其极差均值为五组极差的平均值，并由此计算得到上、下控制限；该图将用于日后测定的准确度检查，测定值在上、下控制限之内的属随机波动，超出上、下控制限的为测定失控。第二张为极差控制图，将用于日后测定的精确度检查；测定值的极差在上、下控制限之内的属随机波动，超出上、下控制限的为测定失控。





第八节 箱图

15.8.1 主要功能

调用 Graphs 菜单的 Boxplot 过程，可绘制箱图。箱图可用于表现观测数据的中位数、四分位数和两头极端值。

15.8.2 实例操作

[例 15-8]研究甲基汞对肝脏脂质过氧化的毒性作用，选用 25 只大白鼠，随机分成五组，按不同剂量染毒一段时期后测定肝脏 LPO 含量（n mol/L），资料如下表，试绘制箱图。

编号	染毒剂量（mg/kg 体重）				
	5.0	10.0	20.0	30.0	40.0
1	184.30	391.50	1025.40	1897.21	1821.33
2	268.20	487.25	1289.24	1705.33	2897.53
3	222.64	345.69	1463.55	1532.46	2001.40
4	127.52	574.12	1168.47	2015.46	2748.97
5	291.50	526.78	1356.70	2100.40	4539.75

15.8.2.1 数据准备

激活数据管理窗口，定义变量名：所测定肝脏 LPO 含量数据的变量名为 DATA，输入原始数据；再设一变量为 GROUP，用于定义不同染毒剂量组，依次输入 1、2、3、4、5。

15.8.2.2 操作步骤

选 Graphs 菜单的 Boxplot...过程，弹出 Boxplot Chart 定义选项框，有 2 种箱图可选：Simple 为简单箱图，Clustered 为复式箱图，本例选用简单箱图。然后点击 Define 钮，弹出 Define Simple Boxplot:Summaries for Groups of Cases 对话框（图 15.10），在左侧的变量列表中选 data 点击 ➢ 钮使之进入 Variable 框，选 group 点击 ➢ 钮使之进入 Category Axis 框。点击 OK 钮即完成。

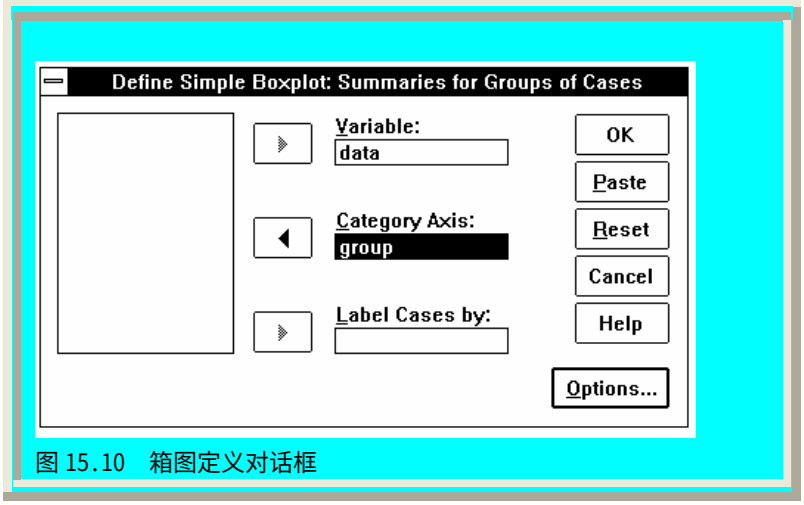
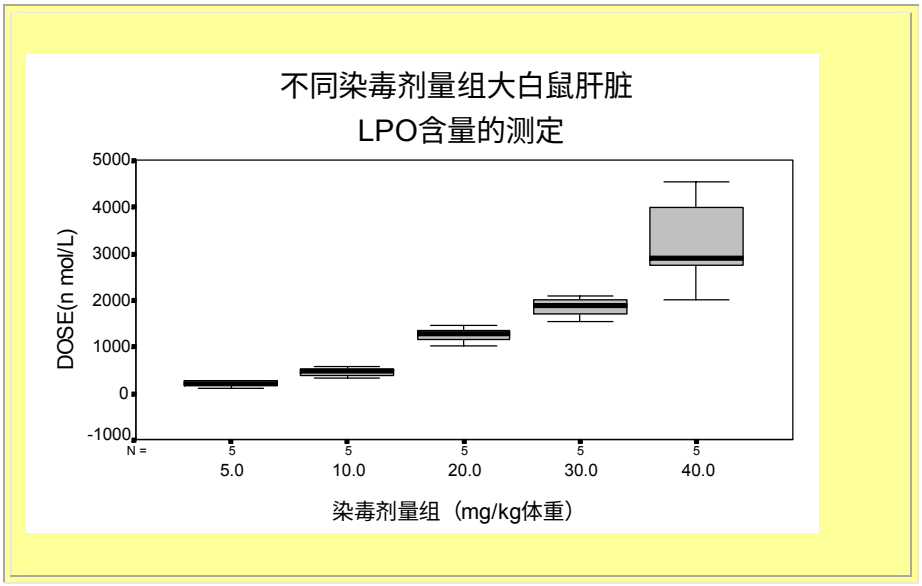


图 15.10　箱图定义对话框

15.8.2.3 结果显示

下图即为箱图，图形的含义是：中间的粗线为中位数，灰色的箱体为四分位（箱体下端为第二十五百分位数、上端为第七十五百分位数），两头伸出的线条表现极端值（下边为最小值、上边为最大值）。从图中可见：随染毒剂量的增加，大白鼠肝脏过氧脂质化的程度更严重，且 LPO 含量的变动范围也随之加大。



第九节 均值相关区间图

15.9.1 主要功能

调用 Graphs 菜单的 Error Bar 过程，可绘制均值相关区间图。正态分布资料的描述性指标：如均值、标准差、标准误，并由此求得的参照值范围、总体均值的可信区间等，都可用均值相关区间图来表现。

15.9.2 实例操作

[例 15-9]食品中微量砷（As）主要采用两种方法测定，一是新银盐法，另一是 DDC-Ag 法。今比较两种方法测定不同浓度 As 标准液（μg/5ml）的光密度值可信区间，试绘制均值相关区间图。

编号	新银盐法					DDC-Ag 法				
	1.0	2.0	3.0	4.0	5.0	1.0	2.0	3.0	4.0	5.0
1	0.150	0.330	0.490	0.690	0.990	0.021	0.065	0.087	0.112	0.169
2	0.130	0.410	0.510	0.620	0.810	0.033	0.059	0.089	0.109	0.148
3	0.140	0.250	0.570	0.730	0.860	0.038	0.048	0.092	0.119	0.134
4	0.180	0.380	0.530	0.780	0.940	0.041	0.069	0.073	0.134	0.175
5	0.190	0.350	0.550	0.810	0.950	0.029	0.057	0.081	0.127	0.162

15.9.2.1 数据准备

激活数据管理窗口，定义变量名：数据的变量名为 DATA，将新银盐法和 DDC-Ag 法的测定所得光密度值一并输入；然后设一变量为 GROUP，用于定义不同标准液浓度组，依次输入 1、2、3、4、5；再设一变量为 CATE，用于定义不同方法组，依次输入 1、2。

15.9.2.2 操作步骤

选 Graphs 菜单的 Error Bar...过程，弹出 Error Bar 定义选项框，均值相关区间图有 2 种，Simple 为简单均值相关区间图，Clustered 为复式均值相关区间图，本例选用复式均值相关区间图。然后点击 Define 钮，弹出 Define Clustered Error Bar:Summaries for Groups of Cases 对话框（图 15.11），在左侧的变量列表中选 data 点击 ➢ 钮使之进入 Variable 框，选 group 点击 ➢ 钮使之进入 Category Axis 框，选 cate 点击 ➢ 钮使之进入 Define Clusters by 框。

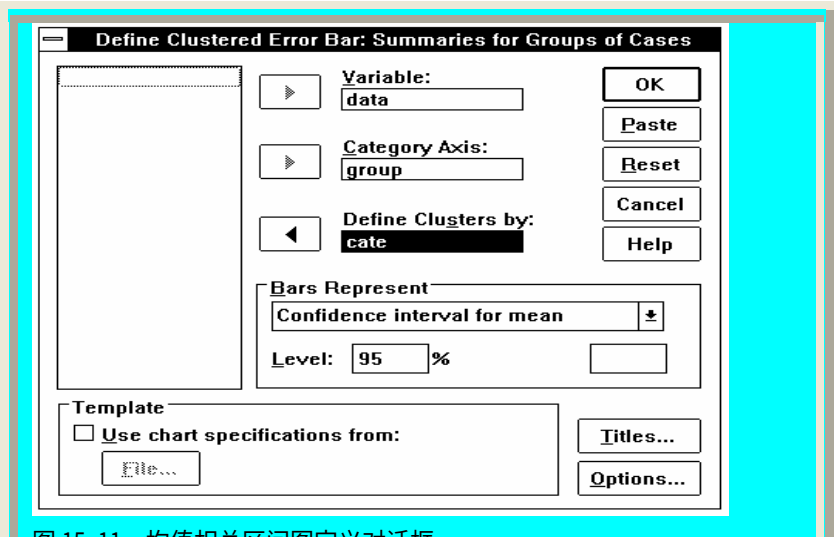


图 15.11　均值相关区间图定义对话框

在 Bar Represent 栏中有一下拉菜单，系统提供 3 种图形表现方式让用户选择：

Confidence interval for mean：绘出总体均值的可信区间，要求输入区间的百分数；

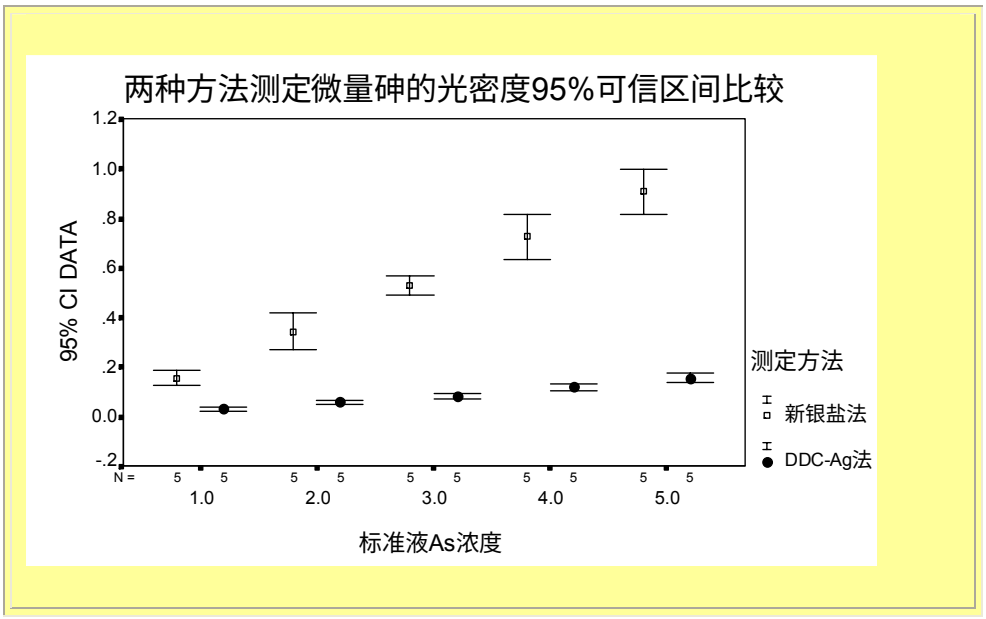
Standard error of mean：绘出均值+/-两倍标准误的区间；

Standard deviation：绘出均值+/-两倍标准差的区间。

本例选用 Confidence interval for mean。之后点击 Titles...钮，弹出 Titles 对话框，在 Title 栏内输入“两种方法测定微量砷的光密度 95%可信区间比较”，点击 Continue 钮返回 Define Clustered Error Bar:Summaries for Groups of Cases 对话框，再点击 OK 钮即完成。

15.9.2.3　结果显示

下图可见，用 DDC-Ag 法测定的光密度精确度高于新银盐法，但其对浓度的区分度低于新银盐法。



第十节 散点图

15.10.1 主要功能

调用 Graphs 菜单的 Scatter 过程，可绘制散点图。散点图用于表现测量数据的原始分布状况，读者可从点的位置判断测量值的高低、大小、变动趋势或变化范围。

15.10.2 实例操作

[例 15-10]研究饮茶对人体血清微量元素（μmol/L）的影响作用，结果如下，试绘制散点图。

编号	多喝茶组			少喝茶组			不喝茶组		
	Zn	Fe	Mn	Zn	Fe	Mn	Zn	Fe	Mn
1	15.87	36.87	0.29	13.25	32.40	0.21	11.40	29.87	0.12
2	16.27	35.90	0.28	12.98	32.65	0.22	10.89	30.14	0.14
3	16.77	37.45	0.24	13.64	33.04	0.19	11.05	30.70	0.11

15.10.2.1 数据准备

激活数据管理窗口，定义变量名：数据的变量名为 DATA，将各组各微量元素的测定值一并输入；设一变量为 CATE1，用于定义喝茶状况组，多喝茶组为 1、少喝茶组为 2、不喝茶组为 3；再设一变量为 CATE2，用于定义微量元素种类，Zn 为 1、Fe 为 2、Mn 为 3。

15.10.2.2 操作步骤

选 Graphs 菜单的 Scatter...过程，弹出 Scatterplot 定义选项框（图 15.12），散点图有 4 种，Simple 为单层散点图，Overlay 为多层散点图，Matrix 为矩阵散点图，3-D 为立体散点图，本例选用单层散点图。然后点击 Define 钮，弹出 Simple Scatterplot 对话框（图 15.13），在左侧的变量列表中选 data 点击 ➢ 钮使之进入 Y Axis 框，选 cate1 点击 ➢ 钮使之进入 X Axis 框（指定饮茶状况为横轴标目），选 cate2 点击 ➢ 钮使之进入 Set Markers by 框（指定微量元素种类为散点标志）。点击 Titles...钮，弹出 Titles 对话框，在 Title 栏内输入“饮茶与人体血清微量元素的关系”，点击 Continue 钮返回 Simple Scatterplot 对话框，再点击 OK 钮即完成。

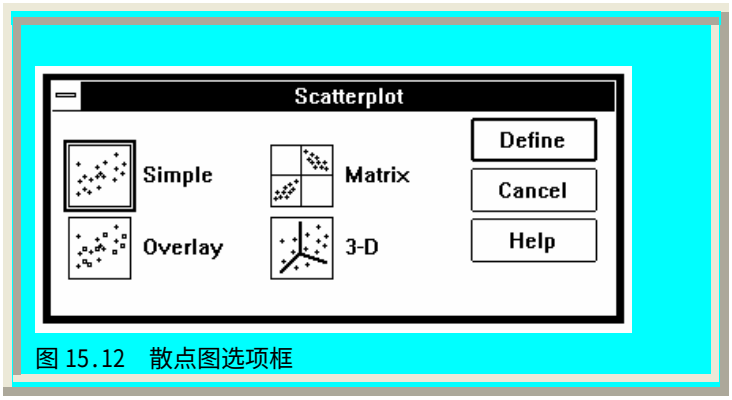


图 15.12 散点图选项框

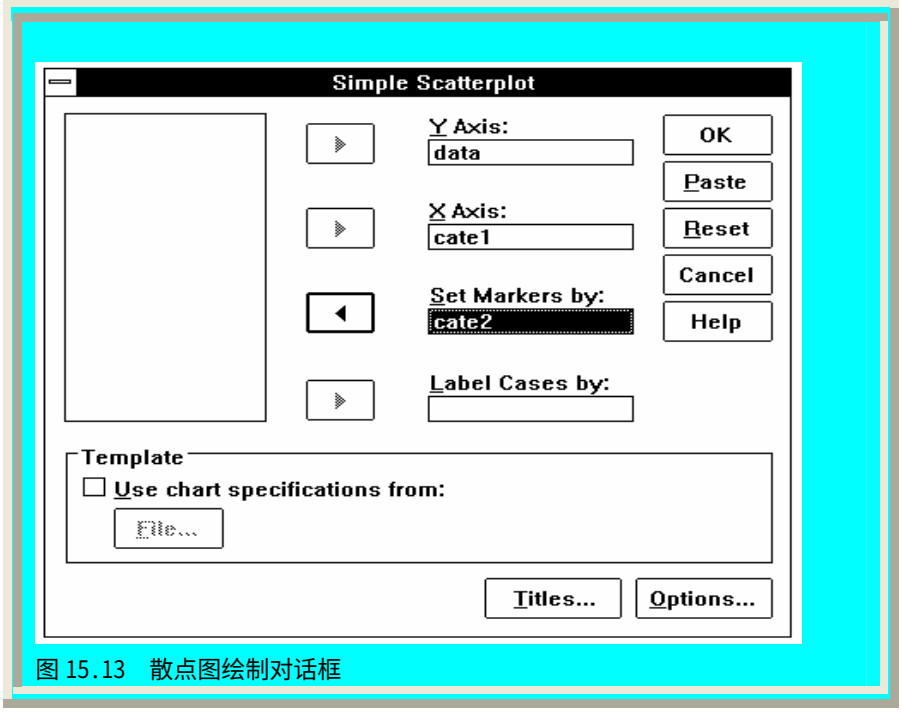
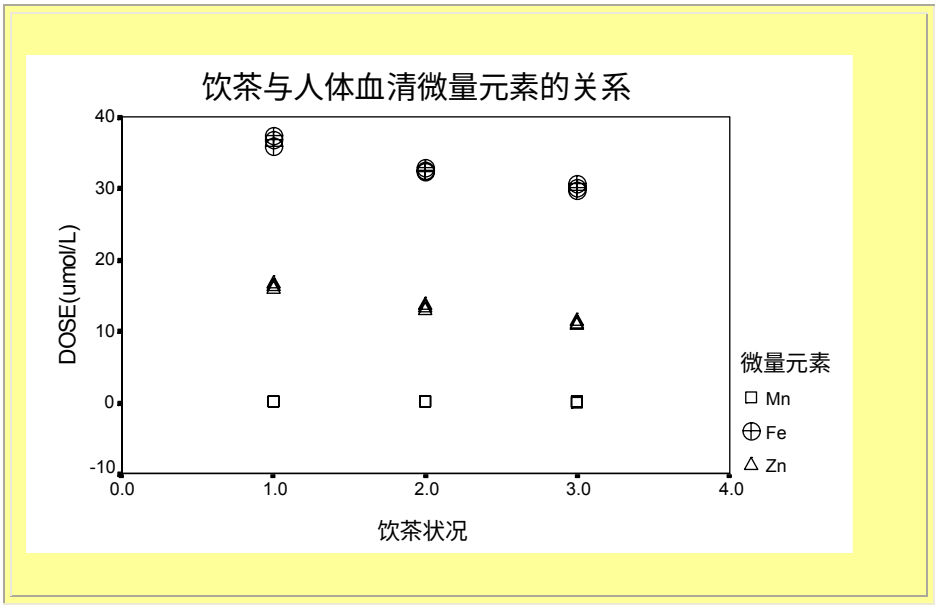


图 15.13　散点图绘制对话框

15.10.2.3　结果显示

下图横轴标目中 1.0 为多喝茶组、2.0 为少喝茶组、3.0 为不喝茶组。图中可见，饮茶对提高人体血清铁（Fe）和锌（Zn）的含量有明显的作用，且饮茶量越大，效果越明显；但对锰（Mn）的作用较小，饮茶量与效果的关系不明显。



第十一节 直方图

15.11.1 主要功能

调用 Graphs 菜单的 Histogram 过程，可绘制直方图。直方图是用直条的长短来表示连续性的绝对数（或称频数）资料的多少，其意义与本章第三节介绍的区域图相似，但区域图能进行多组资料的比较（如堆积式区域图），而直方图不能。

15.11.2 实例操作

[例 15-11]现有某地某年流行性乙型脑炎患者的年龄分布资料如下表，试绘制直方图。

年龄分组	患者人数	每岁患者人数
0-	30	30.0
1-	30	30.0
2-	75	75.0
3-	78	78.0
4-	77	77.0
5-	49	49.0
6-	71	71.0
7-	59	59.0
8-	56	56.0
9-	67	67.0
10-14	143	28.6
15-19	77	15.4
20-24	16	3.2
25-29	10	2.0
30-34	12	2.4
35-44	7	0.7
45-54	3	0.3
55-64	1	0.1

15.11.2.1 数据准备

激活数据管理窗口，定义变量名：频数资料的变量名为 NUMBER，将每岁患者人数资料输入；设一变量为 GROUP，用于定义年龄组，将各年龄分组的下限值输入。为使频数资料在作图中生效，应选 Data 菜单的 Weight Cases...命令项，在弹出的 Weight Cases 对话框中取 Weight cases by 项，并选变量 NUMBER 点击 ➢ 钮使之进入 Frequency Variable 框，点击 OK 钮即可。

15.11.2.2 操作步骤

选 Graphs 菜单的 Histogram...过程，因直方图只有一种类型，故直接弹出 Histogram 对话框

（图 15.14），在左侧的变量列表中选 group 点击 ➢ 钮使之进入 Variable 框；点击 Titles...钮，弹出 Titles 对话框，在 Title 栏内输入“某地某年流行性乙型脑炎患者年龄分布”，点击 Continue 钮返回 Histogram 对话框，再点击 OK 钮即完成。

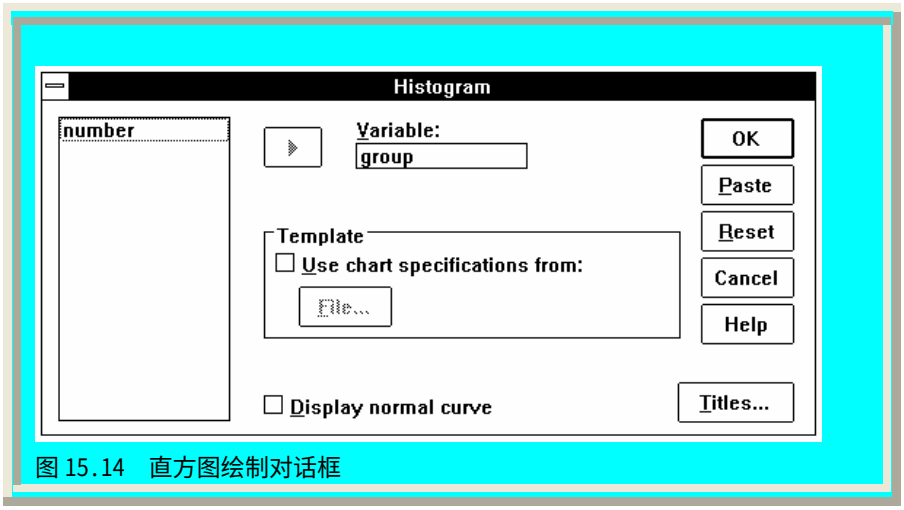


图 15.14　直方图绘制对话框

系统在 Chart Carousel 窗口输出直方图。由于本例资料的分组情形比较细（即每岁一组），而系统只按默认的每 5 岁一组方式输出图形，所以需要安用户的要求对统计图进行编辑。点击 Chart Carousel 窗口上端工具栏中的 Edit 钮，弹出!Chart1 窗口，将鼠标箭头指向图的横轴并双击鼠标左键，弹出 Interval Axis 对话框（图 15.15），在 Axis Title 处将“GROUP”改为中文“年龄分组”；在 Intervals 栏选择 Custom 项，点击 Define...钮，弹出 Interval Axis:Define Custom Interval 对话框（图 15.16），在 Definition 栏的 Interval width 处输入 1，要求按每岁一组的方式作图；在 Range 栏的 Minimum 处输入 0，在 Maximum 处输入 65，要求横轴从 0 岁开始至 65 岁止。点击 Continue 钮返回 Interval Axis 对话框，再点击 OK 钮即可。

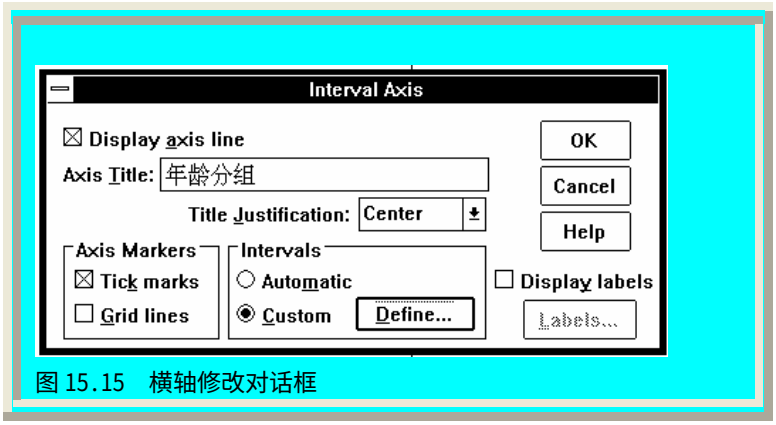


图 15.15　横轴修改对话框

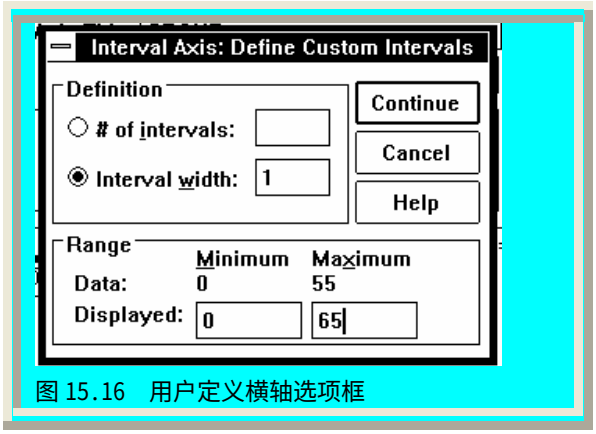
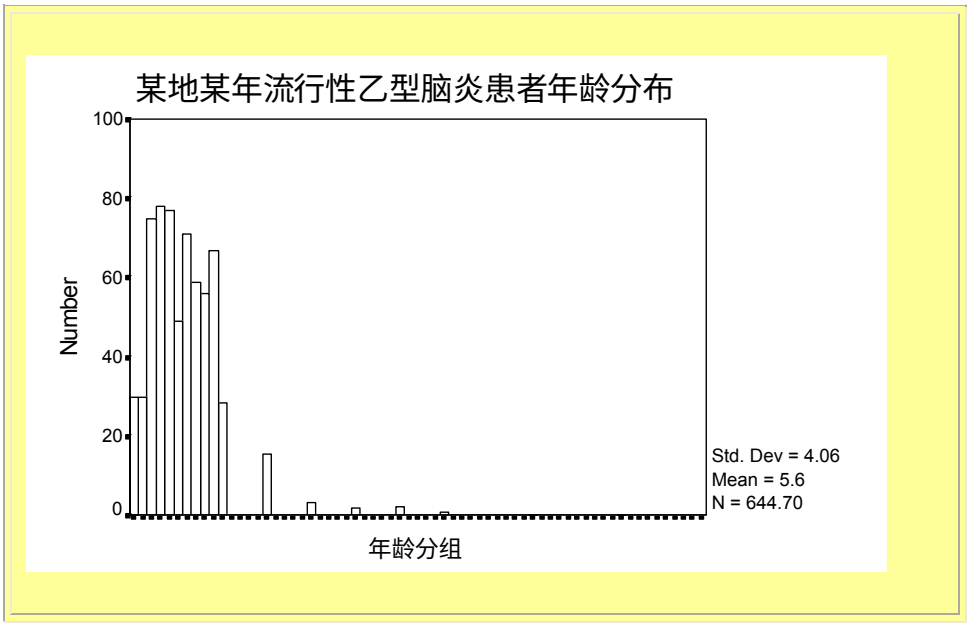


图 15.16　用户定义横轴选项框

15.11.2.3　结果显示

下图即为经用户编辑后的直方图。由图中可见：该地流行性乙型脑炎患者主要集中在 10 岁之前，10 岁之后患者数骤减，尤其是 35 岁之后患者数几乎为 0。



第十二节 正态概率分布图

15.12.1 主要功能

调用 Graphs 菜单的 Normal P-P 过程，可绘制正态概率分布图。如果变量值是正态分布的，则所绘制的正态概率分布图将呈现一条从纵轴零点指定右上角的直线。

15.12.2 实例操作

[例 15-12]某医师测得 30 名健康女大学生血清总蛋白含量（g/L）资料如下，试绘制正态概率分布图。

74.3	78.8	68.8	78.0	70.4	80.5	80.5	69.7	71.2	73.5
79.5	75.6	75.0	78.8	72.0	72.0	72.0	74.3	71.2	72.0
75.9	73.5	78.8	74.3	75.8	65.0	74.3	71.2	69.7	68.0

15.12.2.1 数据准备

激活数据管理窗口，数据的变量名为 DATA，将原始测定值输入。

15.12.2.2 操作步骤

选 Graphs 菜单的 Normal P-P...过程，弹出 Normal P-P Plot 对话框（图 15.17），在左侧的变量列表中选 data 点击 ➢ 钮使之进入 Variable 框。在 Transform 栏中，系统有 4 种数据转换形式供用户选择：

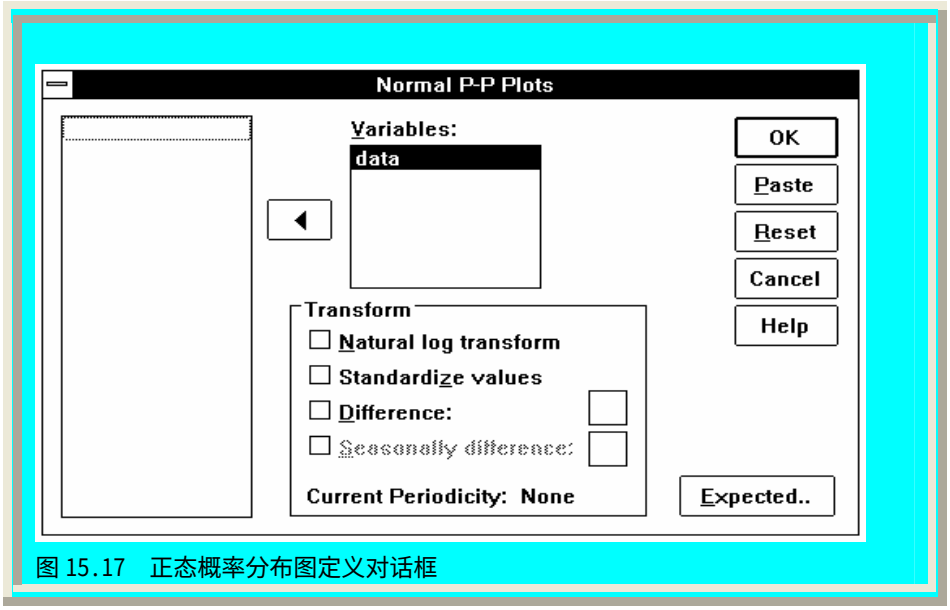


图 15.17 正态概率分布图定义对话框

Natural Log transform：作自然对数的转换；

Standardize values：作标准化值（即 Z 值）的转换；

Difference：使用系列值与 n 个相近观察值的差别值替代原始值；

Seasonally difference：使用系列值与 n 个时期值的差别值替代原始值。

本例不作数据转换。点击 Expected...钮弹出 Normal P-P Plot:Expected 对话框，系统询问用户采用什么方法计算预期正态概率值，共 4 种方法：

Blom：使用公式 $\frac{r-3/8}{n+1/4}$ 推算；

Tukey：使用公式 $\frac{r-1/3}{n+1/3}$ 推算；

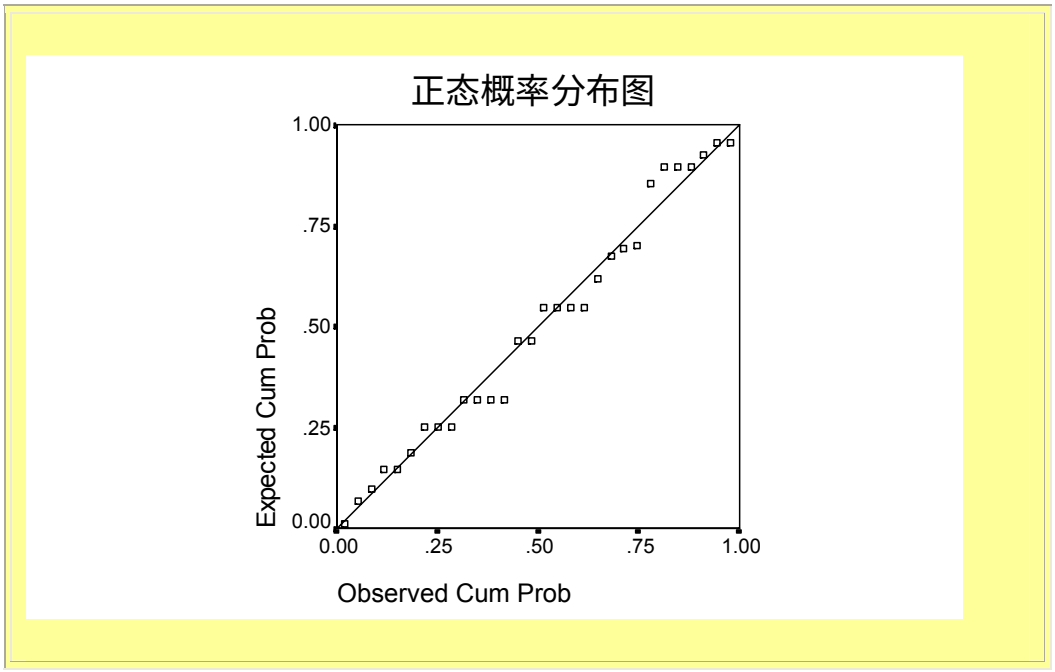
Rankit：使用公式 $\frac{r-1/2}{n}$ 推算；

Van der Waerden：使用公式 $\frac{r}{n+1}$ 推算。

上列各式中，n 为观察单位数，r 为 1～ n 的秩次。本例选 Blom 方法。点击 Continue 钮返回 Normal P-P Plot 对话框，再点击 OK 钮完成操作。

15.12.2.3 结果显示

下图显示观察值紧贴正态概率线分布，由此可知其服从正态分布。



第十三节 正态概率单位分布图

15.13.1 主要功能

调用 Graphs 菜单的 Normal Q-Q 过程，可绘制正态概率单位分布图。绘制正态概率单位分布图的意义与本章第十二节介绍的正态概率分布图一样，只是其纵轴采用概率单位而不是采用概率。

15.13.2 实例操作

[例 15-13]某医师研究不同布氏菌苗免疫后布氏菌素皮肤反应情况，资料如下，试绘制正态概率单位分布图。

A 组菌苗皮肤浸润直径（mm）									
20.0	24.0	25.0	23.0	22.0	20.5	17.5	19.0	25.0	23.0
24.0	24.0	16.5	21.0	15.0	11.5	20.5	15.5	22.5	25.5
B 组菌苗皮肤浸润直径（mm）									
20.0	20.0	22.0	20.0	15.5	18.5	16.5	15.5	18.0	16.5
15.5	20.0	14.0	13.5	13.5	12.5	16.5	14.5	13.0	20.5
20.5	25.0	10.0	20.0	22.5	18.0				

15.13.2.1 数据准备

激活数据管理窗口，A 组菌苗皮肤浸润直径的数据变量名为 X1，将原始测定值输入；B 组菌苗皮肤浸润直径的数据变量名为 X2，也将原始测定值输入。

15.13.2.2 操作步骤

选 Graphs 菜单的 Normal Q-Q...过程，弹出 Normal Q-Q Plot 对话框（图 15.18），在左侧的变量列表中选 data 点击 ➢ 钮使之进入 Variable 框；不作数据转换；选择 Blom 方法推算预期正态概率单位值；点击 OK 钮完成操作。

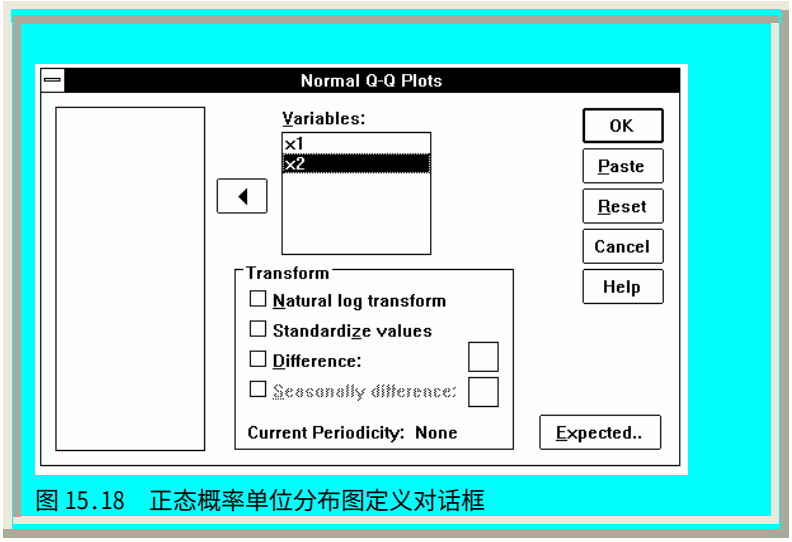
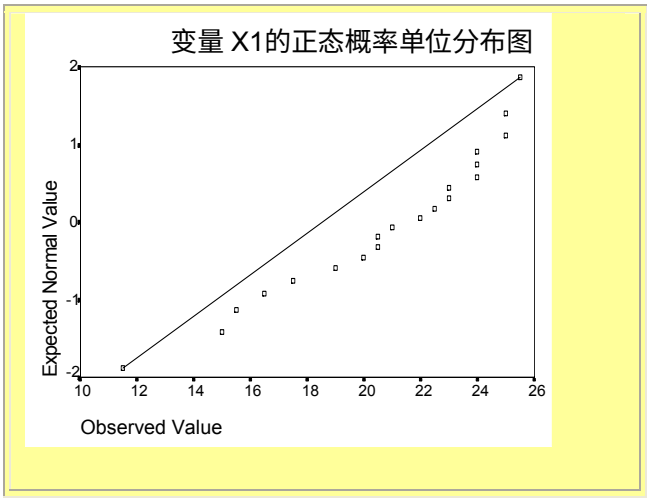
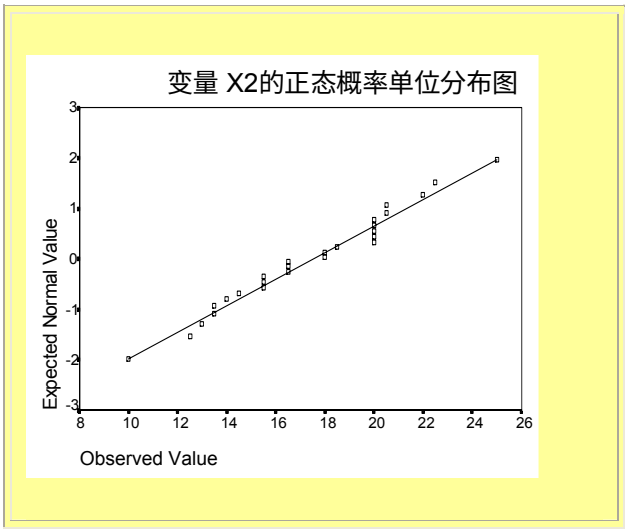


图 15.18 正态概率单位分布图定义对话框

15.13.2.3 结果显示

由于是对两个变量作正态概率单位分布图，故系统输出两张图。比较变量 X1 与 X2 的正态概率单位分布图，可看到变量 X1 的分布形式偏离正态，因其值大多落在 0 概率单位之下，故呈现一定程度的正偏态；而变量 X2 的分布形式接近正态，其值较均匀地紧贴正态概率单位线。


变量 X1的正态概率单位分布图
Expected Normal Value
2
1
0
-1
-2
10
12
14
16
18
20
22
24
26
Observed Value


变量 X2的正态概率单位分布图
Expected Normal Value
Observed Value
3
2
1
0
-1
-2
-3
8
10
12
14
16
18
20
22
24
26

第十四节 普通序列图

15.14.1 主要功能

调用 Graphs 菜单的 Sequence 过程，可绘制普通序列图。普通序列图用于表现一组或几组观察值随另一序列性变量变化的趋势。

15.14.2 实例操作

[例 15-14]研究气相色谱仪采用硅胶采样管测定苯胺的稳定性，资料如下，试绘制普通序列图。

分析时间 (min)	加入值 (μg)	测定值 (μg)
0	100	101.8
1	100	106.0
3	100	104.0
5	100	98.0
7	100	99.0
10	100	96.0

15.14.2.1 数据准备

激活数据管理窗口，定义变量名：加入值数据的变量名为 X1，测定值数据的变量名为 X2，分析时间的变量名为 TIME。将原始数据一一输入即可。

15.14.2.2 操作步骤

选 Graphs 菜单的 Sequence...过程，弹出 Sequence Chart 对话框（图 15.19），在左侧的变量列表中选 x1、x2 点击 ➢ 钮使之进入 Variable 框，选 time 点击 ➢ 钮使之进入 Time Axis Labels 框，点击 OK 钮即完成。

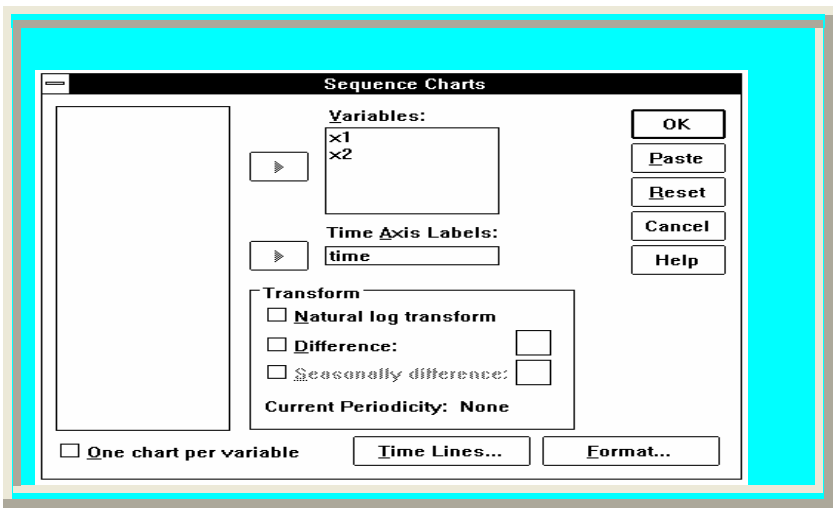
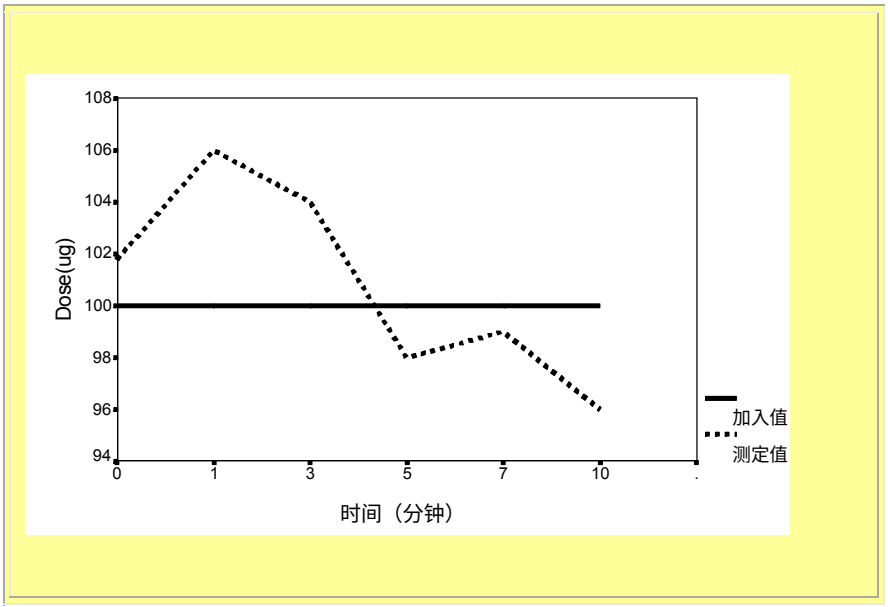


图 15.19　普通序列图定义对话框

15.14.2.3　结果显示

下图为气相色谱仪采用硅胶采样管测定苯胺的稳定性研究图。该图显示：测定值围绕加入值作小幅度（<6%）的波动。测定前期，测定值高于实际加入值；测定后期，测定值低于实际加入值。最佳测定时间为样品加入后 3～5 分钟内。



第十五节 时间序列图

15.15.1 主要功能

调用 Graphs 菜单的 Time Series 过程，可绘制时间序列图。时间序列是指按时间顺序排列的随机变量的一组实测值。分析时间序列图，可以从运动的角度认识事物的本质，如几个时间序列之间的差别、一个较长时间序列的周期性，或对未来的情况进行预测。

15.15.2 实例操作

[例 15-15]连续 4 周（每周 5 个工作日）测定某无菌操作室空气中的细菌含量（$\times 10^3/M^3$），资料如下，试绘制时间序列图，看是否存在周期性变动趋势。

时间	第一周	第二周	第三周	第四周
第 1 天	3.7	3.0	4.4	3.2
第 2 天	4.0	6.6	4.9	4.8
第 3 天	11.5	12.7	10.8	10.9

第 4 天	7.5	5.9	5.6	5.8
第 5 天	3.8	2.0	4.1	6.0

15.15.2.1 数据准备

激活数据管理窗口，定义变量名为 DATA，然后按时间顺序从第一周第 1 天起将观察数据依次输入。

15.15.2.2 操作步骤

在 Graphs 菜单的 Time Series 项中，含两个过程：

Autocorrelations：自相关时间序列图，自相关指相关值来自一组时间序列中前一时间序列与其后序列的对应各观测值的相关；

Cross-correlations：交叉相关时间序列图，交互相关指相关值来自某一时间序列变量与相邻另一个或一些时间序列变量的对应各观测值的相关。

本例只有一个随机变量，故选用 Autocorrelations 过程。在弹出的 Autocorrelations 对话框（图 15.20）中，选左侧变量列表中的 data 点击 ➢ 钮使之进入 Variable 框。在 Display 栏选 Autocorrelations 项，要求仅绘制自动相关的时间序列图。点击 Options... 钮，弹出 Autocorrelations:Options 对话框，在 Maximum Number of Lags 处输入 5，表示时间序列阶段为每 5 天一个周期，点击 Continue 钮返回 Autocorrelations 对话框，再点击 OK 钮即完成。

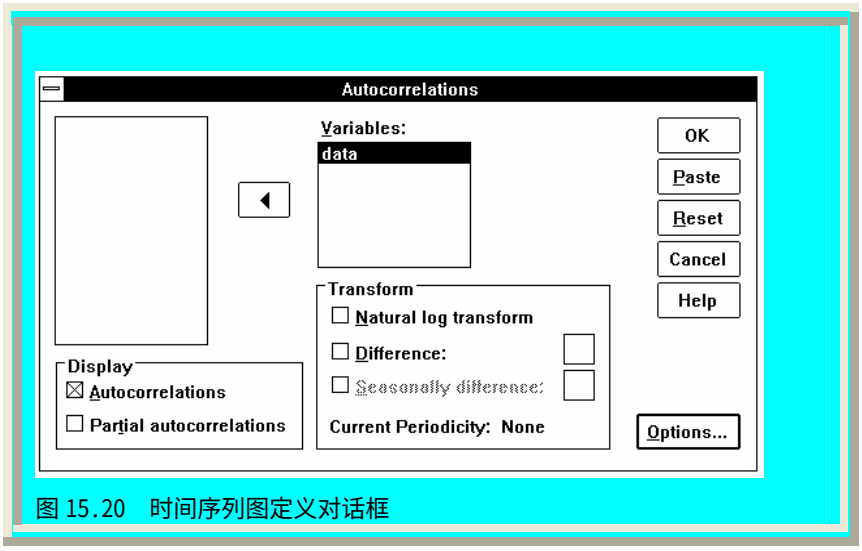
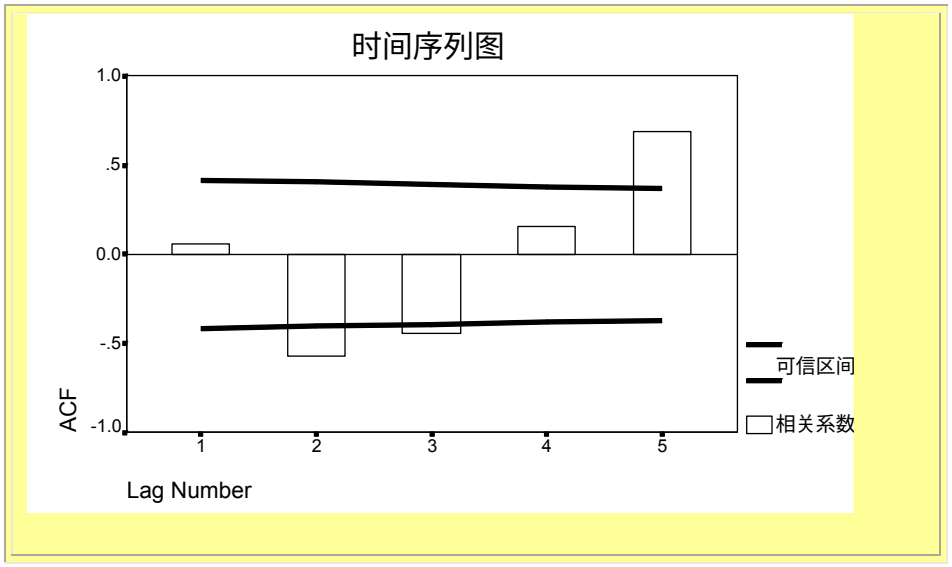


图 15.20 时间序列图定义对话框

15.15.2.3 结果显示

在时间序列图中，用户可根据相关系数的大小来判断序列模型的变动趋势。一般地说，相关系数为 0 或为<0，则前后序列或相邻序列的变动趋势保持原状；当最大的正相关系数出现在最后一个时点之前的任一时点时，表明趋势变动，完整地说是后面的或相邻变量的序列较前面的或相邻前面变量的序列延迟，前面的或相邻前面变量的序列超前的时点即在最大正相关系数所在的时点。

如本例，系统按用户指向，一个时间序列为 5 个时点段，下图显示最大正相关系数位于最后一个时点，故表明前后时间序列稳定，即具有周期性。


时间序列图
1.0
.5
0.0
-.5
-1.0
ACF
1
2
3
4
5
Lag Number
可信区间
相关系数