

# 前言

在当今知识爆炸的年代，人们日常生活的每一个方面几乎都会产生各种各样的数据，同时也离不开数据，这些数据的类型各异，其表现形式纷繁芜杂。随着数据分析工作的作用凸显，如何对现有数据进行整理、加工、处理和分析，以期得到所谓的结论，作为我们进行决策的依据？如何利用现有数据对将来可能出现的数据结果或结论进行预测？不管是针对企事业单位的管理者或决策者还是从事具体数据分析的工作人员而言，都需要进行合理数据分析流程的规划，区分数据类型，利用适合的数据分析方法，使用方便、快捷、可靠的统计软件作为工具，对特定数据进行分析与预测，从而洞察市场动向、观测人心所在、把握商机，从而提升所在单位的竞争力。

具有深厚数学背景的数据分析方法往往会成为人们继续深入学习的“拦路虎”因此，本书就是要降低学习难度，通过笔者积累的大量真实案例和数据，以文字阐述而不是复杂公式推导的形式深入浅出地剖析统计分析方法的基本原理和步骤，重点在于理清数据分析的基本思路，得到恰当的分析结果。

本书通过菜单操作和编程两种途径，用大量的图形展示每一步操作的细节，一步一步地带领读者走入统计数据分析的美妙世界。

为与国际接轨，本书采用 SAS 9.1.3 Service Package 4 的英文版本进行讲解，希望读者边操作边学习。

为了提高学习效率，本书还附送随书案例的全部数据以及利用 SAS 系统进行统计分析的详细操作视频。

## 主要内容

本书全面、细致地讲解了 SAS 系统进行数据预处理和数据分析的全过程，全书分为 7 个部分共 11 章。

第 1 部分——数据预处理。该部分内容包括第 1 章，主要结合笔者实际工作经验对数据分析之前的准备工作和处理流程进行系统讲解。同时，对于 SAS 系统环境和界面的基本元素和操作方式进行了介绍，详细阐释了 SAS 编程语言的基本结构，并明确了 SAS 系统分析的对象是数据库当中的数据集。本部分内容是掌握 SAS 系统的基本知识，也是读者必须学习和掌握的内容。

第 2 部分——数据的描述。该部分内容主要为第 2 章，主要讲解如何通过 SAS 系统绘制常见的统计图形和统计表格来描述数据的概貌，并在此基础上计算反映数据集中趋势、离散程度和分布状况的统计量来进行简单的描述统计分析。该部分内容是数据分析工作的基础和前提，同时也是各种高级统计方法的数据描述基础。

第 3 部分——统计推断。该部分包含第 3~5 章，内容涵盖了简单统计推断、方差分析和非参数检验。本部分主要讲解如何利用已经搜集到的一个或多个样本数据，根据样本数据的特征，在总体分布形式已知或未知的情况下，通过特定的方法，推断总体参数或对总体参

数进行判断的基本思路和详细分析流程。

第 4 部分——相关与回归分析。该部分内容主要是第 6 章，主要研究两个或多个变量之间的相互依存或影响的统计关系，具体内容包括两个变量之间的相关分析、两组变量之间的典型相关分析以及经典回归分析。本章是进行数据统计建模的基础，也是数据分析工作的核心内容之一。

第 5 部分——多元统计分析。该部分内容包括第 7~9 章，具体内容涵盖了因子分析、主成分分析、聚类分析、判别分析、列联表分析和对应分析。本部分内容对多元统计分析中各种常用的方法和原理进行了具体阐述，是对数据分析工作的进一步深入介绍。

第 6 部分——微观计量分析。该部分内容主要是第 10 章，讲解微观计量分析中的常见离散因变量模型，具体内容包括概率线性模型、二元选择模型、多重选择模型及计数模型。这部分内容也是日常问卷数据分析的重点。

第 7 部分——时间序列分析。该部分内容主要为第 11 章，着重讲解时间序列的平稳性以及利用 Box-Jenkins 法进行时间序列建模的模型识别、估计及预测等问题。

## 本书特色

- ▶ 本书是笔者多年一线数据分析工作的经验总结和倾情奉献，所有分析案例均来源于作者第一手的调查数据。
- ▶ 使用 SAS 9.1.3 Service Pack 4 最新版本作为分析工具，其程序和菜单操作过程同样适用于较低版本。
- ▶ 数据分析方法齐全，涵盖了描述统计分析、统计推断、多元统计分析、微观计量分析、时间序列分析等常用方法。
- ▶ 以说理而非公式推导的方式进行原理和理论讲解。
- ▶ 全程真实数据案例引导学习，数据分析报告顺理成章、水到渠成。
- ▶ 数据文件+录像示例，可对照录像进行手把手的实践，学习效率高。

## 本书约定

本书的插图和运行结果可能会与读者实例环境中的操作界面或结果略有差别，这可能是由于操作系统平台、SAS 版本不同而导致的，在此特别说明，一切以实际环境为准。

## 致谢

本书由首都经济贸易大学统计学院阮敬老师编著，成都易为科技有限责任公司审校，参与编辑、排版、校对的同志有：王斌、黄中林、张强林、王晓、夏惠军、余松、江广顺、姚新军。在本书编写过程中，还得到了首都经贸大学统计学院纪宏教授、台湾辅仁大学谢邦昌教授、清华大学朱世武教授的大力支持和帮助，我的师妹刘欢同学提供了部分章节的素材。在此要特别感谢家人对我创作的大力支持，没有他们的支持，本书不可能这么快速地和读者见面。

由于时间有限，书中不足之处在所难免，恳请读者批评指正（电子函件：book\_better@sina.com）。

作者

2009.1

# 目 录

第 1 章	数据预处理 .....	1
1.1	SAS 环境与操作界面 .....	1
1.2	SAS 编程基础 .....	3
1.2.1	SAS 编程语言的基本结构 .....	3
1.2.2	SAS 结构化编程语句 .....	5
1.3	SAS 的数据处理对象 .....	8
1.3.1	SAS 数据库和 SAS 数据集 .....	8
1.3.2	SAS 系统的外部数据文件 .....	15
1.4	数据预处理原理和基本方法 .....	17
1.4.1	数据整理 .....	19
1.4.2	数据分拆与合并 .....	22
1.4.3	数据清洗 .....	24
1.4.4	数据变换 .....	27
1.5	本章小结 .....	30
第 2 章	数据的描述 .....	31
2.1	统计图 .....	31
2.1.1	直方图 .....	31
2.1.2	条形图 .....	34
2.1.3	线图 .....	36
2.1.4	散点图 .....	37
2.1.5	饼图 .....	39
2.1.6	盒式图 .....	40
2.1.7	茎叶图 .....	41
2.2	统计量 .....	42
2.2.1	集中趋势 .....	42
2.2.2	离散程度 .....	45
2.2.3	分布形状 .....	48
2.2.4	利用菜单和程序进行详细的描述统计分析 .....	50
2.3	统计表 .....	55
2.3.1	统计表的基本要素 .....	56
2.3.2	用 TABULATE 过程绘制统计表 .....	56
2.4	数据分布 .....	58
2.4.1	总体分布 .....	59

2.4.2	样本分布	59
2.4.3	抽样分布	59
2.5	本章小结	61
第 3 章	简单统计推断	62
3.1	简单统计推断的基本原理	62
3.1.1	参数估计	63
3.1.2	假设检验	64
3.2	单总体参数的估计及假设检验	68
3.2.1	单总体的参数估计	68
3.2.2	单总体参数的假设检验	71
3.3	两总体参数的估计及假设检验	82
3.3.1	独立样本的参数估计和检验	82
3.3.2	成对样本的参数估计和检验	91
3.4	本章小结	95
第 4 章	方差分析	96
4.1	方差分析的基本原理	96
4.2	单因素方差分析	99
4.2.1	单因素方差分析与方差同质性检验	99
4.2.2	方差分析的多重比较	102
4.2.3	方差分析模型的参数估计和预测	104
4.3	多因素方差分析	109
4.3.1	只考虑主效应的多因素方差分析	110
4.3.2	存在交互效应的多因素方差分析	116
4.4	协方差分析	118
4.5	本章小结	122
第 5 章	非参数检验	123
5.1	非参数检验的基本问题	123
5.2	单样本非参数检验	124
5.2.1	单样本均值的 Wilcoxon 符号秩检验	124
5.2.2	单样本的 Kolmogorov-Smirnov 检验	125
5.3	两个样本的非参数检验	128
5.3.1	两个独立样本中位数比较的 Wilcoxon 秩和检验	128
5.3.2	两个独立样本分布的 Kolmogorov-Smirnov 检验	132
5.3.3	成对样本中位数的 Wilcoxon 符号秩检验	134
5.4	多个样本的非参数检验	136
5.4.1	多个独立样本位置的 Kruskal-Wallis 检验	136
5.4.2	多个独立样本位置的 Jonckheere-Terpstra 检验	138
5.4.3	多个独立样本中位数的 Brown-Mood 检验	139
5.5	本章小结	139
第 6 章	相关与回归分析	141

6.1	相关分析	141
6.1.1	简单相关分析	142
6.1.2	偏相关分析	146
6.1.3	等级相关分析	147
6.2	典型相关分析	151
6.2.1	典型相关分析基本原理	151
6.2.2	典型相关系数的显著性检验	155
6.2.3	典型相关的冗余分析	156
6.3	线性回归分析	158
6.3.1	回归分析的基本原理	158
6.3.2	一元线性回归分析	161
6.3.3	多元线性回归分析	168
6.4	定性自变量回归分析	172
6.4.1	虚拟变量的设定	172
6.4.2	含有虚拟变量的回归分析	173
6.5	本章小结	174
第 7 章	因子分析	175
7.1	数据降维	175
7.1.1	数据降维的基本问题	175
7.1.2	数据降维的基本原理	176
7.2	主成分分析	177
7.2.1	主成分分析的基本概念与原理	177
7.2.2	主成分分析的基本步骤和过程	178
7.3	因子分析	184
7.3.1	因子分析的基本原理	185
7.3.2	因子分析的基本步骤和过程	186
7.4	本章小结	194
第 8 章	聚类分析与判别分析	195
8.1	聚类分析的基本原理	195
8.1.1	分类的基本原则	195
8.1.2	单一指标的系统聚类过程	197
8.1.3	多指标的系统聚类过程	198
8.2	聚类分析的步骤和过程	202
8.2.1	系统聚类	202
8.2.2	快速聚类	210
8.2.3	变量聚类	212
8.3	判别分析的基本原理	215
8.4	判别分析的步骤和过程	216
8.4.1	距离判别	217
8.4.2	Bayes 判别	217
8.4.3	Fisher 判别	224


8.4.4	逐步判别	227
8.5	本章小结	231
第 9 章	列联分析与对应分析	233
9.1	列联分析	233
9.1.1	列联表	233
9.1.2	列联表的分布	237
9.1.3	$\chi^2$ 分布与 $\chi^2$ 检验	238
9.1.4	列联表中的关联度分析	240
9.1.5	$\chi^2$ 分布的期望值准则	241
9.2	对应分析	242
9.2.1	对应分析的基本思想	242
9.2.2	对应分析的步骤和过程	243
9.3	本章小结	249
第 10 章	离散因变量模型	250
10.1	线性概率模型	250
10.2	二元选择模型	251
10.2.1	线性概率模型的缺陷与改进	252
10.2.2	二元选择模型的基本原理	252
10.2.3	BINARY PROBIT 模型	253
10.2.4	BINARY LOGIT 模型	263
10.3	多重选择模型	269
10.3.1	多重选择模型的基本原理	269
10.3.2	ORDINAL PROBIT 模型	271
10.3.3	ORDINAL LOGIT 模型	276
10.3.4	MULTINOMIAL LOGIT 模型	279
10.4	计数模型	280
10.4.1	POISSON 回归模型的基本原理	281
10.4.2	POISSON 回归模型的分析过程和步骤	281
10.5	本章小结	285
第 11 章	时间序列分析	286
11.1	时间序列的基本问题	286
11.1.1	时间序列的组成部分	286
11.1.2	时间序列的平稳性	288
11.2	ARIMA 模型的分析过程	291
11.2.1	ARIMA 模型	291
11.2.2	ARMA 模型的识别、估计与预测	292
11.2.3	利用 SAS 时间序列预测系统进行菜单操作	301
11.3	本章小结	308

# 第 1 章

## 数 据 预 处 理

数据预处理是统计分析过程中不可缺少的重要内容，是定量分析研究的基础，也是进行社会科学研究的基本前提。本章将结合实际定量分析工作中的常见问题，从 SAS 系统的基本界面、操作方式及编程基础等方面对数据预处理的方法进行详细介绍，并通过 SAS 软件的实际操作和真实数据展现一个完整的数据预处理过程。

### 1.1 SAS 环境与操作界面

SAS 系统安装完毕之后，进入“开始”菜单，选择“The SAS System 9.1.3”，或者双击  图标，进入 SAS 系统的主界面，如图 1-1 所示。

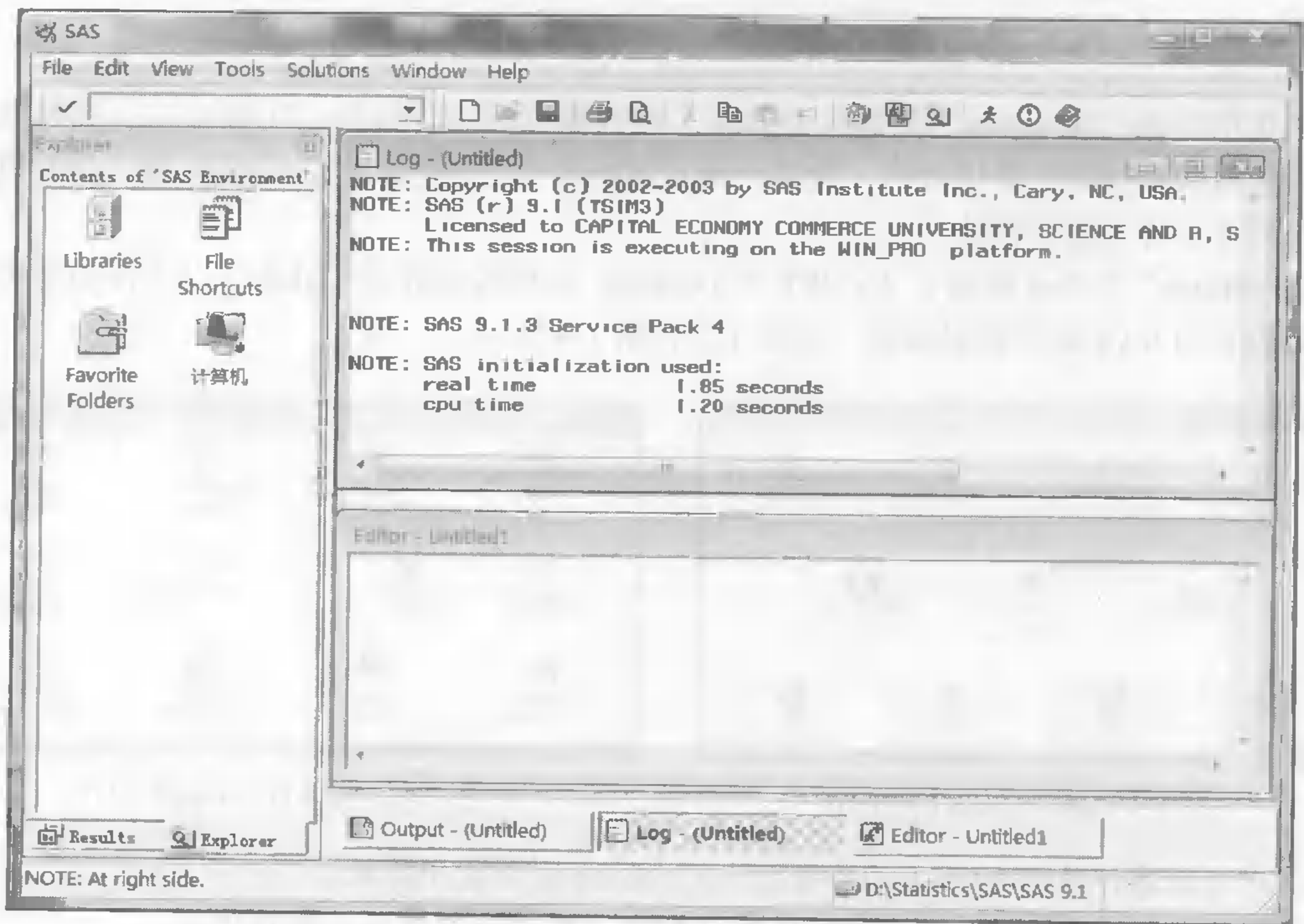


图 1-1 SAS 9.1.3 主界面

在 SAS 主界面中，处于最上面的系统菜单可以实现 SAS 文件和数据库的操作、编辑、视图、分析、作图等一系列功能。系统菜单会因所调用的 SAS 模块不同而有所不同。

- “File” 菜单主要实现 SAS 系统文件的基本操作，包括新建、打开、存储、打印文件，以及导入和导出外部数据文件、利用 E-mail 传送文件等。

- “Edit” 菜单主要实现对文本进行选择、查找、替换、复制、剪切、粘贴等编辑功能。
- “View” 菜单的功能主要是进行窗口切换或打开对应的窗口。
- “Tools” 菜单可以打开数据库查询与管理器，以及表格、图形、报告、文本编辑器，并且可以对 SAS 系统的全局参数、界面、字体、颜色等方案进行调整。
- “Solutions” 菜单是利用 SAS 进行数据分析、程序开发的最主要的菜单，如图 1-2 所示。

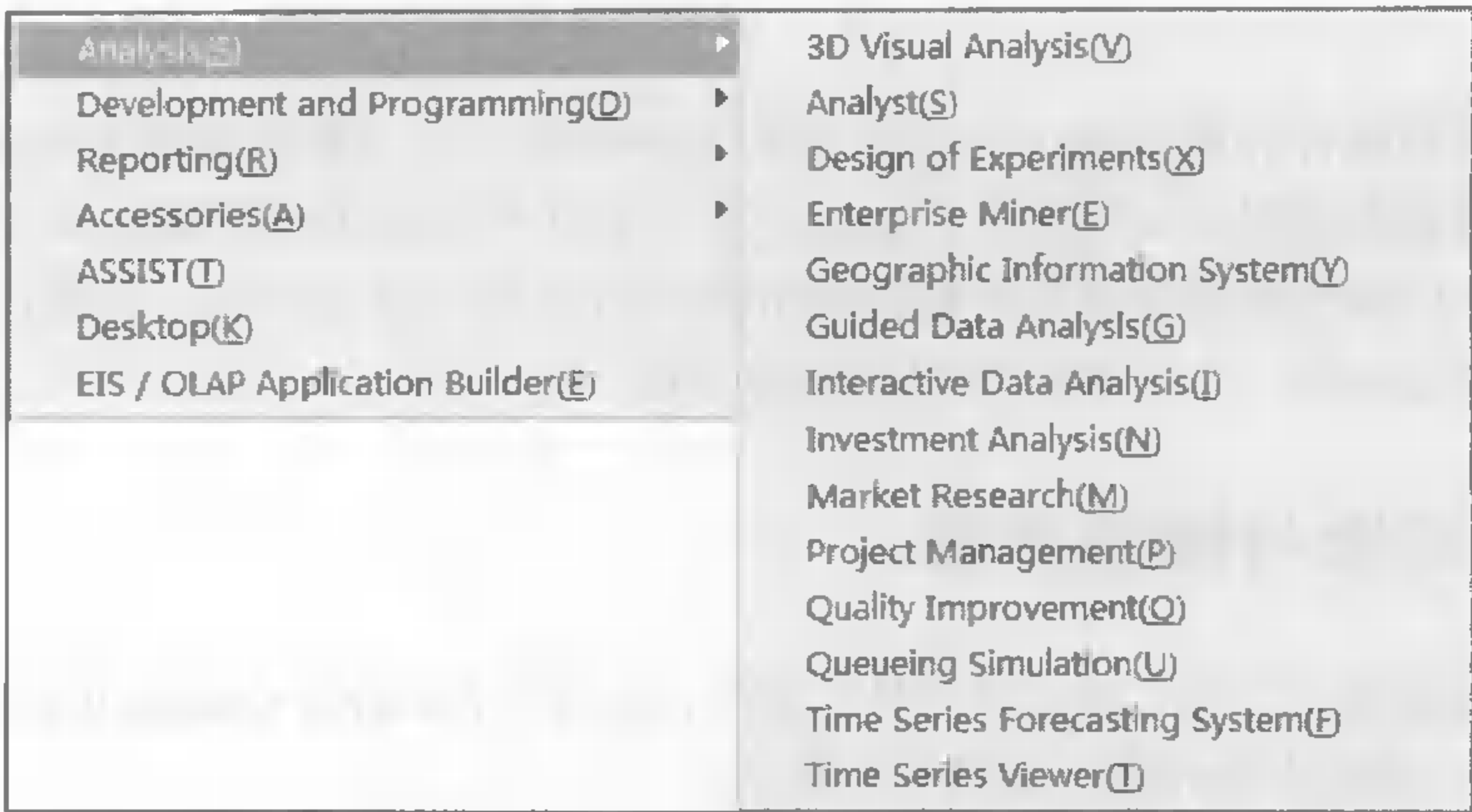


图 1-2 SAS 系统的“Solutions”菜单

该菜单下的“Analysis”二级菜单涵盖了绝大多数的统计分析、实验设计、数据挖掘、地理信息系统、探索性数据分析、投资分析、市场研究、项目管理、质量控制、队列模拟、时间序列分析等功能和模块。

“Solutions”菜单还提供了 ASSIST 和 Desktop 两种图形界面（GUI），以方便初学者在较短时间内利用 SAS 进行数据处理，如图 1-3 和图 1-4 所示。




图 1-3 SAS/ASSIST 的 GUI



图 1-4 SAS/Desktop 的 GUI

用户可以通过单击 GUI 上的图标进入对应的分析功能和模块。

- “Window” 系统菜单主要实现窗口切换、窗口排列等窗口操作功能。
- “Help” 菜单是 SAS 系统的强大帮助系统，提供了从菜单操作到编程语言一系列的帮助功能，同时在任意对 SAS 系统进行分析的过程中，只要单击工具栏上的  图标，便可以快速地转到该过程对应的帮助信息上。

系统菜单下面是 SAS 工具栏。与大多数软件一样，该工具栏提供了能够实现系统菜单中的常用功能的快捷方式，并且以 Windows 操作系统中的常用图标表示，如图 1-5 所示。

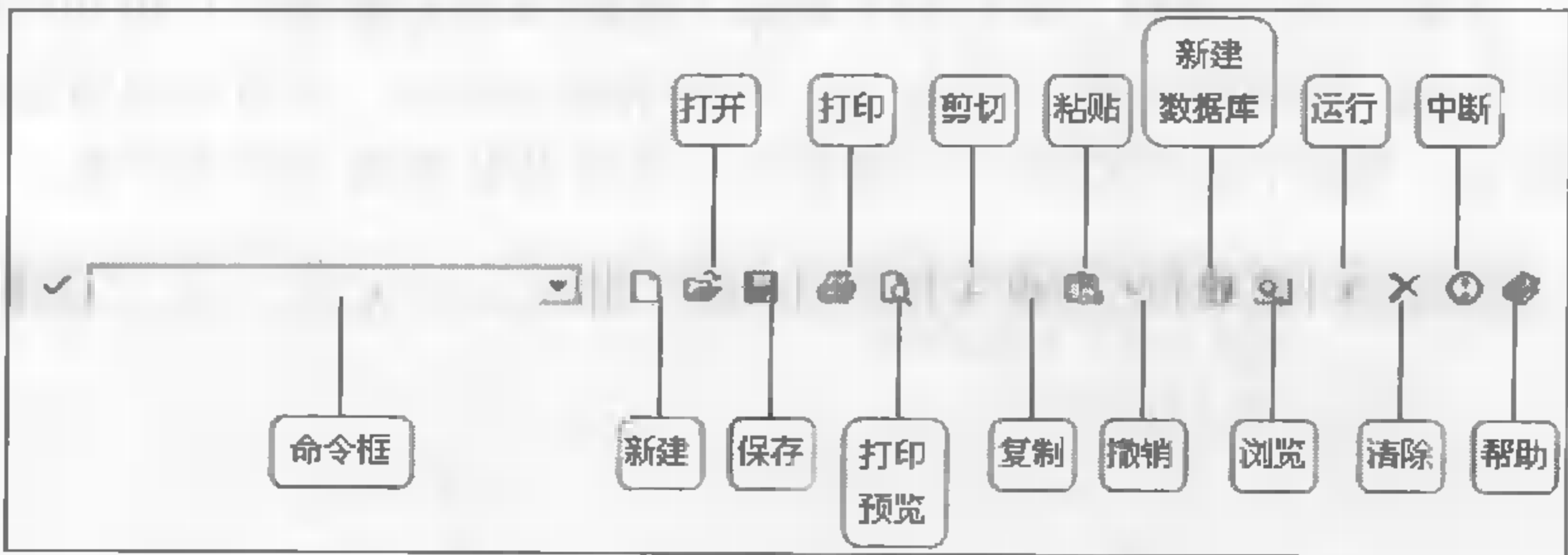


图 1-5 SAS 系统的工具栏

SAS 工具栏的特殊之处在于，工具栏的左边有一个文本输入框（称为文本框），用户可以在该框中输入命令或者调用 SAS 系统的各种模块，故称之为命令框。单击命令框前的✓按钮即可向 SAS 系统提交命令，其功能等同于按键盘上的 Enter（回车）键；单击命令框后的▾按钮，则可从中选择以前使用 SAS 系统时输入的历史命令。

处于主界面最左边的是“Results”（结果）窗口和“Explorer”（浏览）窗口，这两个窗口处于同一界面中，可以通过单击最下边的对应按钮进行切换。

- “Explorer”窗口类似于 Windows 系统下的资源管理器，可以实现对 SAS 数据库、数据文件、程序及其他文件的查看、打开、新建、删除等基本操作，甚至可以直接运行 SAS 系统之外的其他软件。
- “Results”窗口可以以目录或条目的形式提供分析结果的浏览功能。

处于主界面右边的分别是“Output”、“Log”和“Editor”窗口，同样可通过单击最下边的对应按钮进行切换。

- “Output”窗口主要呈现利用 SAS 系统进行分析的结果，按 F7 键可切换至该窗口。
- “Log”窗口是日志窗口，按 F6 键可以切换至该窗口。用户对 SAS 的每一步操作都会以日志的形式列示在“Log”窗口当中（如每次打开 SAS 系统时，该窗口中默认显示一些授权文件的信息）。同时，在进行数据分析的过程中，一些出错信息和程序执行的反馈信息也会出现在该窗口中。
- “Editor”窗口是一个文本编辑器，按 F5 键可以切换至该窗口。用户可以在该窗口中进行编程或输入文字信息（在“Log”和“Output”窗口中，用户不能自行写入信息）。对于这 3 个窗口中的信息，用户可以通过系统菜单或工具栏上的按钮进行复制、新建、存储、打印等操作。

可以根据用户的需要用两种方式关闭和打开这些默认的窗口。一种方式是在命令框中输入对应窗口的名字，另一种方式是在“View”系统菜单中单击对应的窗口名字即可。

## 1.2 SAS 编程基础

SAS 系统不仅可以使使用菜单方式进行数据操作和统计分析，而且还具备强大的编程语言功能，供用户灵活调用各种模块以及调整各种分析参数。

### 1.2.1 SAS 编程语言的基本结构

在 SAS 系统中，可以利用“Editor”或“Program Editor”窗口书写程序。SAS 编程语言

结构比较简单，主要由两个步骤，即 DATA Step（简称 DATA/数据步）和 PROC Step（简称 PROC/过程步）组成。程序中的每一行以“;”号表示输入结束，其语句的语法与常见的高级语言语法大体相似，同样包括关键字、运算符、函数及其参数等基本要素。

```
title "...";           /*设置标题*/
libname ...;           /*定义永久数据库*/
data ...;               /*DATA 步*/
    ...;
    ...;
run;
proc ...;               /*PROC 步*/
run;
```

在 SAS 语言中，通常利用“/\*.....\*/”表示对程序的注释，DATA 步和 PROC 步之间用“run;”或者直接用“;”隔开。通常，一些全局变量的设置语句可以放在 DATA 步之前，如为输入表格或图形定义表头或标题、建立 SAS 数据库等。

## 1. DATA 步（数据步）

DATA 步是 SAS 进行数据管理和操作的基本步骤，其主要功能包括：建立 SAS 数据集，导入外部程序数据文件，分割、合并、修改、更新现有的 SAS 数据集，分析、呈现和管理数据，利用数据集中已有的数据计算或生成新变量等。DATA 步中常用的 SAS 编程语句如下。

- **infile 语句**：从外部文件获取数据。如要使用该语句，必须把其放在其他 data 语句之前，其主要语法如下所示。

```
infile “外部数据路径及文件名” <选项>;
```

如从 D:\student.txt 文件中获取数据，可利用以下 infile 语句。

```
infile “D:\student.txt”;
```

- **input 语句**：指定读入数据的格式以及为读入的数据指定变量名及格式，其语法如下所示。

```
input <变量名 1 变量名 2 ... 变量名 n> <选项>;
```

如从上例读入的外部数据中读入两个变量的数据，并分别命名为 height、weight。

```
input height weight;
```

- **cards 语句**：用于在 SAS 系统中直接输入数据，表明所列示数据的开始。对于 DATA 步编程的详细内容，将在下节详细讲解。

## 2. PROC 步（过程步）

SAS 系统的过程步可引用现有的程序或过程进行相应的数据处理和分析活动。其主要语法如下所示。

```
proc 过程名 <data=数据库名.数据集名> <选项>;
    <var <变量名 1 变量名 2 ... 变量名 n>>;
    <where <条件或表达式>>;
    <by <变量名 1 变量名 2 ... 变量名 n>>;
```

run;

其中的语句功能如下。

- data 语句表示该 PROC 步所处理的数据集。
- var 语句表示处理数据集中的特定变量。对于没有列示的变量，系统则不予处理。
- where 语句表示指定系统处理符合一定条件或表达式的样本。
- by 语句表示指定系统按照所列示的变量进行分组处理。但是要注意的是，使用该语句时，必须先对该语句中指定的分类变量进行排序。

表 1-1 列出了最常用的几种 PROC 步过程。

表 1-1 常见 PROC 步过程的作用及功能

过 程 名	作 用	输 出 结 果
PRINT	显示数据集的变量名及变量值	变量和变量值
SORT	对指定变量进行排序	可对指定变量进行升序、降序排列
MEANS	对数值型变量进行描述统计分析	均值、标准差、极值等统计量
UNIVARIATE	对数值型变量进行描述统计分析	常见统计量、t 检验、分位数、极端值等
FREQ	对定序变量进行描述统计分析	频数、频率、累计频数、累计频率等
CHART	对指定变量绘制文本形式的图形	饼图、横向/纵向直方图、星形图
GCHART	在“Graph”窗口中对指定变量绘制图形	饼图、横向/纵向直方图、星形图

以上输出结果均可在“Output”窗口中显示，并且在“Results”窗口中列示对应结果的标签。双击标签，便会在“Output”窗口中显示对应的具体内容。

PROC 步所能调用的程序或模块非常多。本书出于应用的目的，在后续各章节中就对应的过程或模块进行详细讲解。

3. SAS 编程语言的表达式

SAS 的常用表达式主要有运算表达式和逻辑表达式。SAS 的运算符主要有+(加)、-(减)、\*(乘)、/(除)、\*\* (乘方) 等；SAS 的逻辑符号主要有=(等于)、<(小于)、>(大于)、<=(小于等于)、>=(大于等于)、<>(不等于)、and (和)、or (或)、xor (异或)，这些符号也可以用英文字母等价表示，如表 1-2 所示。

表 1-2 SAS 逻辑运算符的表达方式

符 号	=	<	>	<=	>=	<>
英 文	eq	lt	gt	le	gt	ne

1.2.2 SAS 结构化编程语句

SAS 的结构化编程语句主要有顺序语句、条件语句和循环语句。这 3 种基本形式的语句均可在 DATA 步和 PROC 步中使用。顺序语句是最为常见的语句形式，系统按照语句自身顺序进行解释形式的执行。下面简要介绍条件语句和循环语句的基本使用方法。

1. 条件语句

条件语句能够使程序按照一定的表达式或条件实现不同的操作或执行顺序跳转的功能，其语法如下所示。

```
if 条件或表达式 then
    ...;                                /*当条件或表达式满足时执行的程序*/
else
    ...;                                /*当条件或表达式不满足时执行的程序*/
```

此外，条件语句还可以进行嵌套。



**例 1-1** 比较 x 和 y 两个变量的大小。如果  $x > y$ ，则输出 “ $x > y$ ”；如果  $x < y$ ，则输出 “ $x < y$ ”；如果  $x = y$ ，则输出 “ $x = y$ ”。

现假定 x 赋值为 10，y 赋值为 20，在 Editor 窗口中输入以下程序。

```
data;
x=10;y=20;
if x>y then
    put "x>y";          /*在“Log”窗口中显示引号内的字符*/
else
    if x<y then
        put "x<y";
    else
        put "x=y";
run;
```

单击工具栏上的 （运行）按钮，则在 “Log” 窗口中输出结果 “ $x < y$ ”。

2. 循环语句

循环语句可以使 SAS 系统循环执行一定的程序，主要有计数（DO）循环、当（WHILE）循环、直到循环（UNTIL）3 种形式。

（1）计数循环。

计数循环的主要表达方式如下所示：

```
do 计数变量=初始值 to 终止值 by 步长;
    ...;
end;
```

如果 by 步长省略，则表示计数变量按照默认步长 1 计数。步长也可以是负数，此时计数变量初始值需大于终止值。



**例 1-2** 计算 1~100 之内的所有奇数自然数之和。具体程序如下所示。

方法一：

```
data;
    y=0;
do x=1 to 99 by 2;
```

```

        y=y+x;
    end;
    put "y=" y;
run;


```

方法二:

```

data;
    y=0;
do x=99 to 1 by -2;
    y=y+x;
end;
put "y=" y;
run;

```

单击工具栏上的按钮，则上述两种方法均会在“Log”窗口中显示结果“y=2 500”。

(2) 当循环。

当循环的主要表达式如下所示:

```

do while (继续循环条件表达式);
    ...;
end;

```

该语句执行时会首先判断条件表达式是否成立。如果成立，则系统执行 DO WHILE 中的语句，遇到 END 时返回条件表达式的判断。如此重复，直到条件表达式不能够满足为止。

如在例 1-2 中，同样也可以利用当循环进行计算，程序如下所示:

```

data;
    x=1;
    y=0;
do while (x<100);
    y=y+x;
    x=x+2;
end;
put "y=" y;
run;

```

程序运行结果同样显示“y=2 500”。

(3) 直到循环。

直到循环的主要表达式如下所示:

```

do until (退出循环条件表达式);
    ...;
end;

```

该语句会首先执行循环语句内部的程序，然后判断条件是否成立。如果成立，则退出循环过程，否则继续执行循环语句内部的程序。

仍旧以例 1-2 为例，采用直到循环的程序如下所示:

```

data;
    x=1; y=0;
do until (x>100);

```

```

        y=y+x;  x=x+2;
    end;
    put "y=" y;
run;

```

循环语句均可以使用条件语句跳出循环过程。仍以例 1-2 为例，现计算 50 以内的奇数自然数之和。除了修改以上各种循环语句的条件之外，还可以在不修改循环条件的前提下进行计算，具体程序如下所示：

```

data;
    x=1;
    y=0;
    do until (x>100);
        y=y+x;
        x=x+2;
        if x>50 then leave;
    end;
    put "y=" y;
run;

```

## 1.3 SAS 的数据处理对象

要进行数据预处理，并进一步进行统计分析，必须先弄清利用 SAS 软件包进行统计分析的对象是什么。本节将从两个方面对该问题进行阐述。

### 1.3.1 SAS 数据库和 SAS 数据集

SAS 的数据对象是存在于 SAS 数据库中的数据集。具体而言，数据集是 SAS 的数据处理对象，同时也是数据分析的基础。

#### 1. SAS 数据库

SAS 数据库具体是指存放 SAS 数据文件（即数据集）的文件夹，它与计算机存储器中的某一个具体文件夹相对应。

##### (1) SAS 数据库的分类。

为了让 SAS 系统识别文件夹所对应的数据库，要为每一个数据库指定一个库标记（库名）以识别该库。库标记是逻辑存在的，只存在于 SAS 系统中，同一个文件夹可以对应不同的逻辑库标记，就像一个人可以有很多个称呼或绰号一样。根据数据处理的时效性不同，SAS 数据库又可以分为临时库和永久库。

- 临时库：只有一个，名为 WORK。在每次启动 SAS 时由系统自动生成。关闭 SAS 时，该数据库中的所有数据文件自动被清除。
- 永久库：可有多。用户可以自己指定永久库的库标记，库中的所有数据文件永久保留。但库标记是临时的，每次启动 SAS 系统都要重新指定。

在每次启动 SAS 时，系统都会根据用户安装 SAS 时授权文件的模块自动指定若干个库标记，其中有 3 个库标记是不可缺少的。

- **SASUSER:** 永久库。该库中的数据文件可以被永久保存起来,以便下次系统启动时使用,其中存储了 SAS 用户信息。具体对应于安装 SAS 系统时指定路径下的“\My SAS Files\9.1”文件夹。



本书所用例子的数据文件都存储在该数据库中,读者可把光盘中的数据复制到对应的文件夹中。

- **SASHELP:** 永久库。该库中的数据文件可以被永久保存起来,以便下次系统启动时使用。该数据库包含了控制各种 SAS 运算进程的信息,具体对应于安装 SAS 系统时的根目录。
- **WORK:** 临时库。该库中的数据文件可以被暂时保存起来,关闭 SAS 系统时,该库内的数据文件将被清除。

在 SAS 中,引用临时库中的数据文件可以省略库名,即它被认为是缺省的数据库。如果要引用永久库中的数据文件,则应当指明其对应的永久库名。

## (2) SAS 永久数据库的建立。

为了存储数据文件和进行数据分析,首先须为系统指定数据库。用户根据存储数据的文件夹指定的数据库都是永久数据库。永久数据库的指定可以用菜单和编程两种方式实现。

### ① 通过菜单操作方式制定永久数据库。

单击工具栏上的按钮,或者在 SAS 系统工具栏上的命令框中输入“DMLIBASSIGN”命令,亦或者在“Explorer”窗口的“Libraries”图标上单击鼠标右键,再选择“New”菜单项,均会弹出图 1-6 所示对话框。

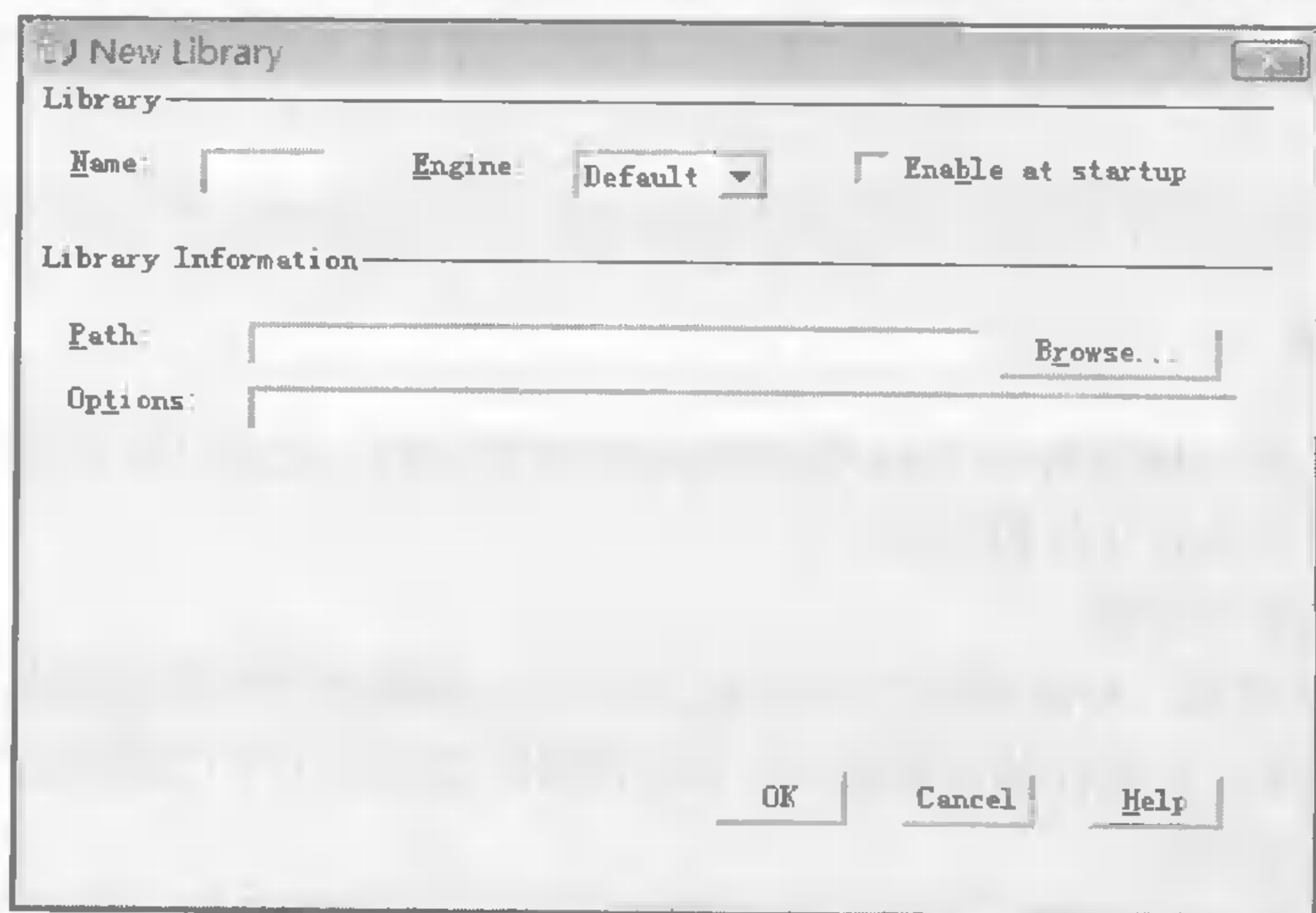


图 1-6 “New Library”对话框

该对话框可以实现指定数据库属性的全部功能。具体如下。

- 在“Library”分栏下,可以在“Name”文本输入框中输入数据库的库标记(库名),“Engine”下拉列表用于指定包括 SPSS、Excel、Access、Sybase 等常用软件的 27 种数据引擎版本,日常使用时选择默认的 Default 选项即可。此外,如果勾选“Enable at startup”复选框,可以让 SAS 系统在每次启动时,自动打开“Name”文本输入框所指定的数据库。

“Library Information”分栏主要用于指定数据库与文件夹路径的对应关系。在“Path”文本输入框中，可以输入将要建立的 SAS 数据库对应的文件夹所处的绝对路径。同时，也可以利用该文本输入框右边的“Browse”按钮调用系统资源管理器找到文件夹所处的位置。



### 例 1-3

建立一个名为“test”的永久数据库，该数据库对应的文件夹绝对路径为“D:\Statistics\SAS\SAS 9.1”。

根据上述菜单操作方法，打开“New Library”对话框，在“Name”文本输入框中输入“test”作为该永久数据库的库标记，在“Path”文本输入框中输入其对应文件夹路径“D:\Statistics\SAS\SAS 9.1”或者单击“Browse”按钮找到该文件夹对应的位置，然后单击“OK”按钮即可。建立后的数据库会在“Explorer”窗口中的“Libraries”中显示出来。

#### ② 利用编程方式建立永久数据库。

利用程序建立永久数据库时要用到 LIBNAME 函数，其语法如下。

LIBNAME 库标记<,路径<,数据引擎<,数据引擎的选项>>>

该函数的参数分别对应于图 1-6 中“Library”下的各项内容。如建立例 1-3 的数据库，可以在“Editor”窗口中输入以下程序。

```
libname test "D:\Statistics\SAS\SAS 9.1";
```

单击工具栏上的  按钮提交系统运行，便可建立一个名为“test”的永久数据库。

#### (3) 清除指定的数据库。

清除指定数据库的操作只是 SAS 系统中的逻辑步骤。当清除一个具体的数据库时，只是清除其在 SAS 系统中先前指定的逻辑库名，并不会清除计算机存储器上对应的物理文件夹。该操作非常简单，仍旧使用 LIBNAME 函数。如清除例 1-3 所指定的数据库，输入并运行以下程序即可。

```
libname test;
```

## 2. SAS 数据集

SAS 数据集具体是指存放在 SAS 数据库中的数据文件，是用 SAS 进行数据分析的基本对象，它与某一个具体的文件相对应。

#### (1) SAS 数据集的分类。

与 SAS 数据库类似，SAS 数据集也可以分为永久数据集和临时数据集。

● 临时数据集：存放在临时数据库（即 WORK 数据库）中的数据文件。关闭 SAS 时，临时数据集会自动被清除。

● 永久数据集：存放在永久数据库的数据文件。关闭 SAS 时，永久数据集仍然可存放在存储器中。

#### (2) SAS 数据集的调用。

每一个数据集都有一个二级名字。第一级是库标记，第二级是数据集名，中间用“.”隔开。在 SAS 系统中，通过指定两级名来识别数据文件。数据文件二级名的一般形式为：库标记.数据集名。

调用永久数据库中的数据集时，应当指定该数据集对应的库标记，而调用临时数据库中的数据集时，则可以省略库标记，直接引用即可。

例如，可以这样引用 Sample 永久库中名为“ABC”的数据集：Sample.ABC；而可以这样引用 Work 临时库中名为“ABC”的数据集：Work.ABC 或 ABC。

### (3) SAS 数据库和数据集的区别和联系。

SAS 数据库对应于文件夹，并且同一个文件夹可以使用不同的库标记（库名）。

如 test 数据库对应“D:\Statistics\SAS\SAS 9.1\”，exam 数据库也可以对应此文件夹。

每个数据集实际上是在存储器的相应文件夹内产生一个名为数据集名、扩展名为.sas7bdat 的文件（SAS V9 下的扩展名命名方式）。图 1-7 所示形象地描述了二者之间的关系。

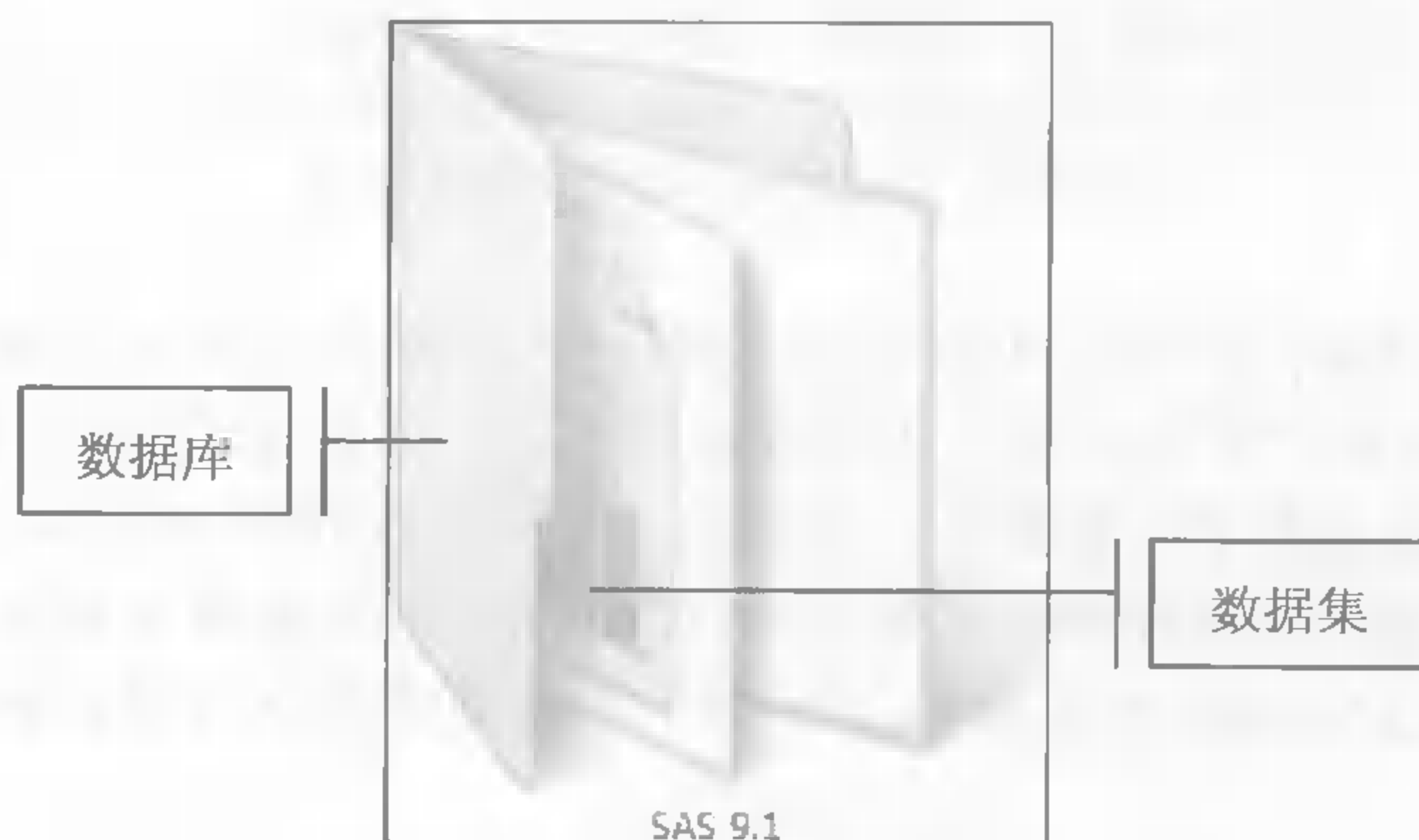


图 1-7 SAS 数据库和数据集之间的关系

之所以要引入数据库和数据集的概念，是出于以下两个方面的考虑。

- 可以使 SAS 系统引用 SAS 数据集和其他 SAS 文件时，只需引用逻辑库名和文件名，不必引用可能很长的路径，从而使引用变得十分简洁。
- 便于 SAS 系统在不同的操作系统下进行移植。因为用逻辑库名和文件名引用 SAS 文件的做法对所有操作系统下的 SAS 是相同的，只是 SAS 逻辑库的物理位置描述随操作系统的不同而有所不同。

### (4) 建立 SAS 数据集。

在建立 SAS 数据集之前，应当先指定存放数据集的数据库。SAS 数据集的建立方法有很多，而且在许多分析模块下均可建立 SAS 数据集。根据使用经验，仍旧可总结为菜单和编程两种方式。此外，为了提高 SAS 系统处理数据的通用性，SAS 还提供了不同格式的数据导入功能。

SAS 数据集的内容包含变量及其对应的数据，即样本观测值。

#### ① SAS 变量的类型。

SAS 变量的基本类型主要有数值型和字符型两种。整数、实数、浮点数、科学计数、表达日期或时间的变量都被存储为数值型数据，默认长度为 8 字节。汉字、字母、符号等存储为字符型数据，默认长度也是 8 字节。对于实际数据中遇到的缺失值，SAS 系统通常用“.”表示。

#### ② 利用 SAS/Insight 模块建立数据集。

**STEP 1** SAS/Insight 是进行探索性数据分析的主要模块。在 SAS 工具栏上的命令框中输入“insight”，或者选择系统菜单“Solutions → Analysis → Interactive Data Analysis”，进入 SAS/Insight 模块的数据集对话框，如图 1-8 所示。

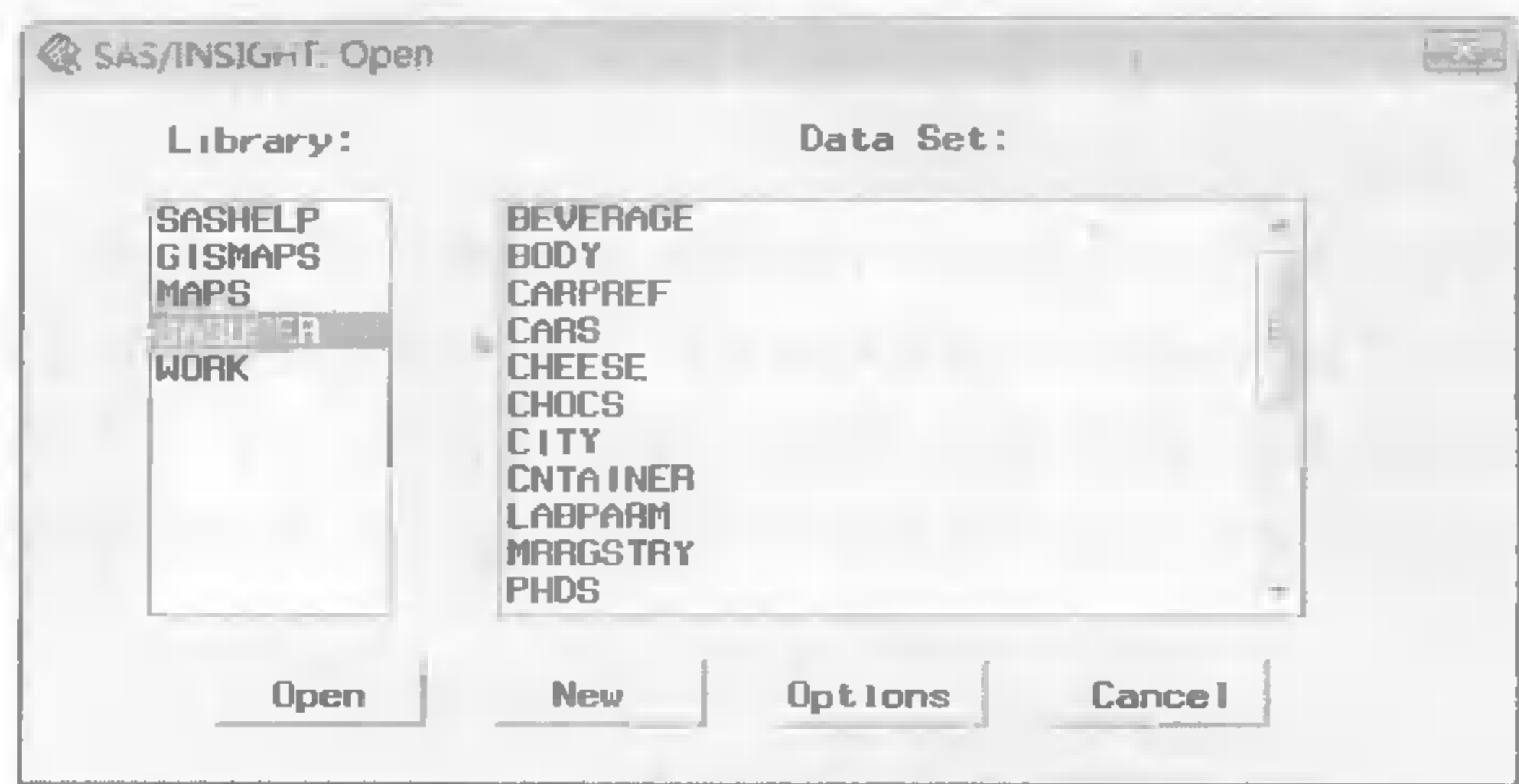


图 1-8 SAS/Insight 的数据集对话框

**STEP 2** 该对话框的左部分“Library”列示的是目前 SAS 系统中指定的数据库，右部分“Data Set”列示的是对应数据库中的所有数据集。在此界面下，通过单击下面的“Open”按钮可以打开对应数据库下的数据集；而单击“New”按钮则可在 Work 临时数据库中建立新的数据集。系统默认把该数据集命名为 Work.A。如果想把该数据集转换为永久数据集，则需在进入 SAS/Insight 界面后，选择“File”系统菜单把该临时数据集保存在永久数据库中。



例 1-4


现有 10 名学生的数学和语文两门功课的期末考试成绩如表 1-3 所示。试用 SAS/Insight 模块在 SASUSER 永久库中建立名为“score”的 SAS 数据集，并且该数据集包含“ID”（学号）、“Literature”（语文）、“Math”（数学）3 个变量和 10 个观测值。

表 1-3 10 名学生的期末考试成绩

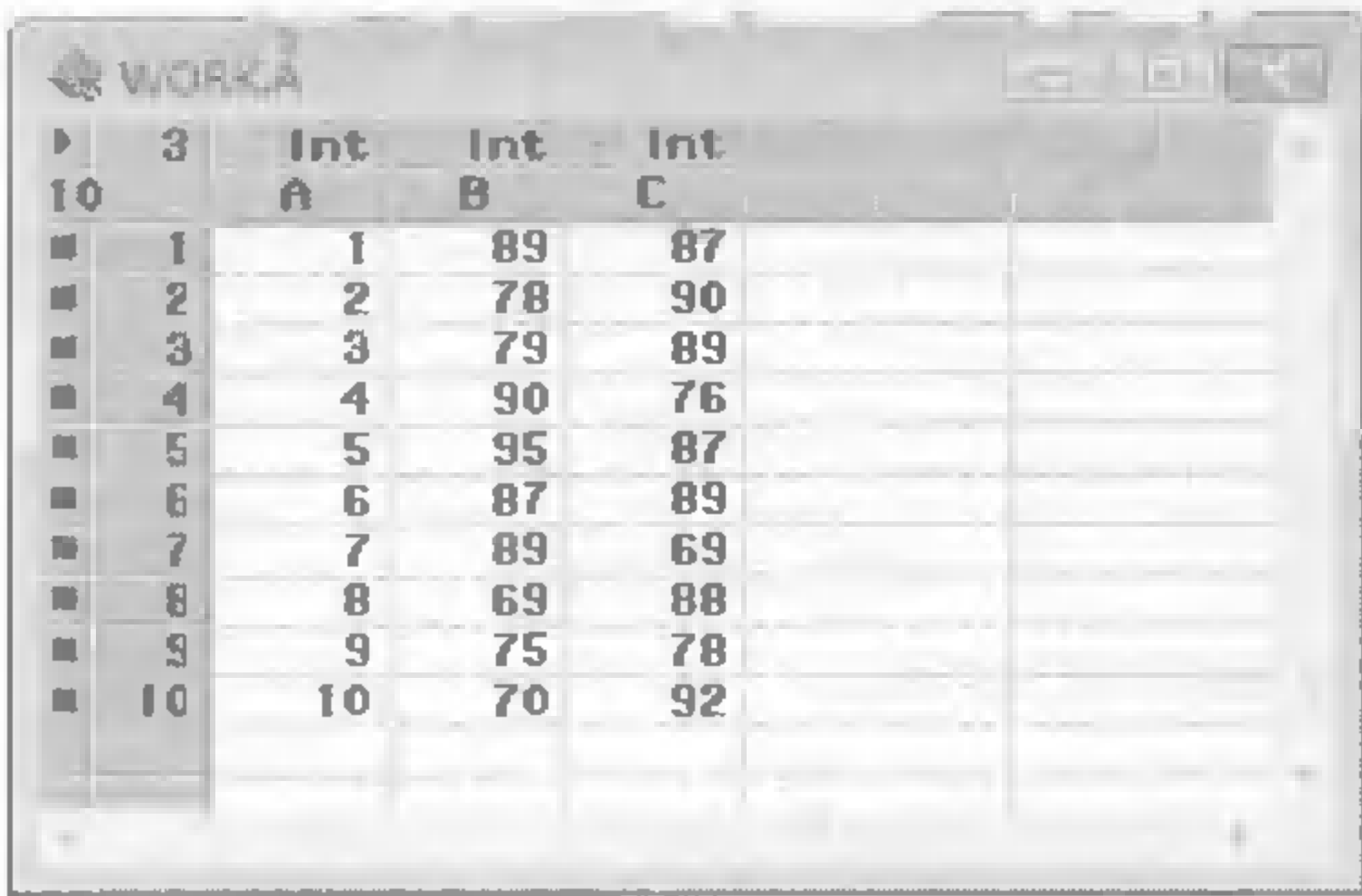
学 号		1	2	3	4	5	6	7	8	9	10
成 绩	语文	89	78	79	90	95	87	89	69	75	70
	数学	87	90	89	76	87	89	69	88	78	92

**STEP 3** 在图 1-8 所示的 SAS/Insight 数据集对话框中，单击“New”按钮，弹出新建临时数据集 Work.A，即 SAS/Insight 的主界面，如图 1-9 所示。

SAS/Insight 的主界面类似于 Excel 电子表格，列代表字段或变量，行代表记录或样本序号，可以直接在空白的表格区域里输入数据。在本例中，把例 1-4 中 3 个变量的数据输入到表格当中，每输入一行或一个样本数据，系统会在最左边自动为该行标上对应的行号；每输入一列或一个变量数据，系统会自动按照英文字母的顺序（A、B、C、D……）将对应的列编号作为缺省变量名。

**STEP 4** 双击对应的缺省变量，或者单击左上角的  按钮，亦或者在表格空白处的任意地方单击鼠标右键，调出 SAS/Insight 数据操作菜单，如图 1-10 所示。

**STEP 5** 然后选择“Define Variables”，弹出设置变量属性的对话框，如图 1-11 所示。



	3	Int	Int	Int
10	A	B	C	
1	1	89	87	
2	2	78	90	
3	3	79	89	
4	4	90	76	
5	5	95	87	
6	6	87	89	
7	7	89	69	
8	8	69	88	
9	9	75	78	
10	10	70	92	

图 1-9 SAS/Insight 数据集主界面

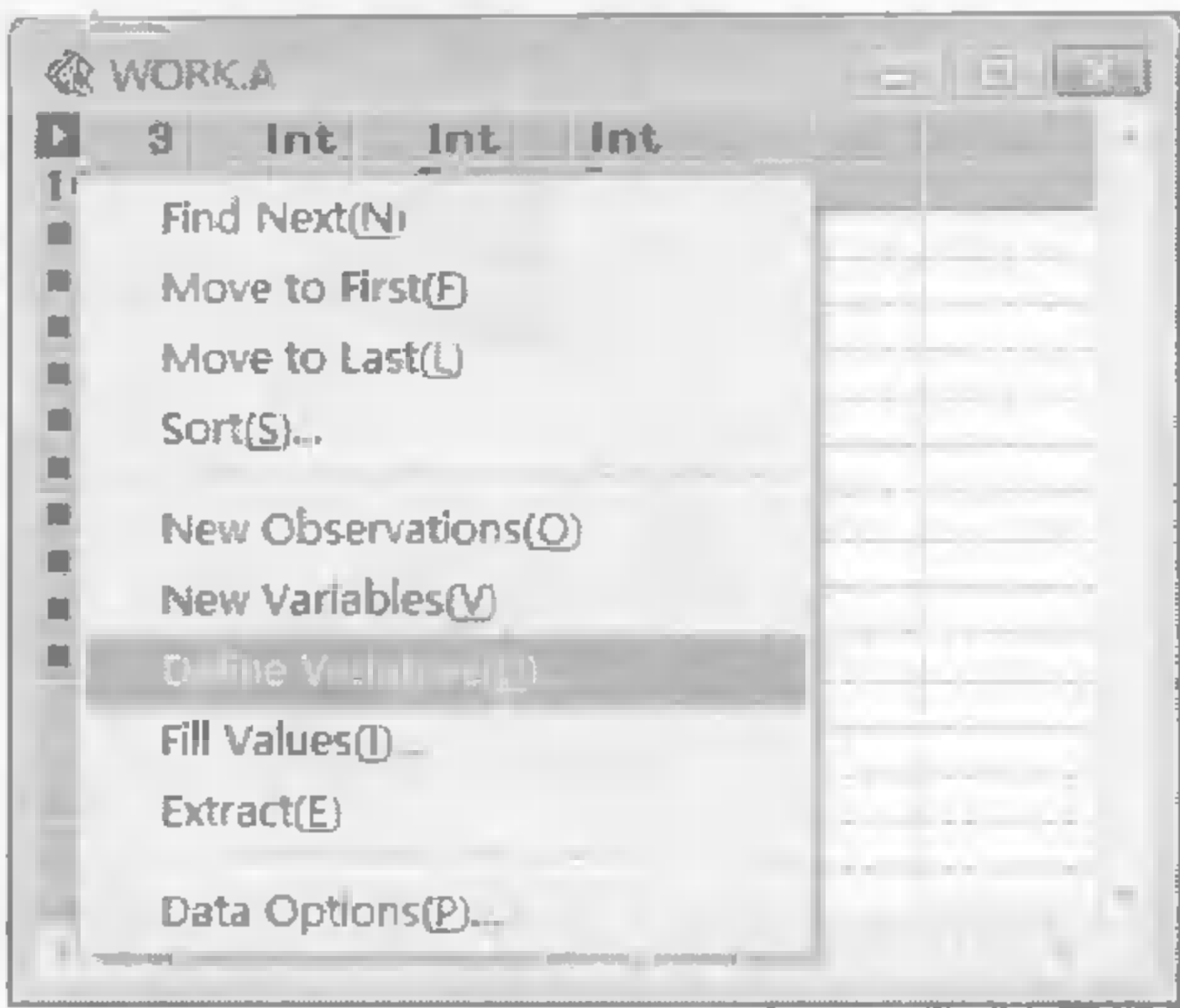


图 1-10 SAS/Insight 数据操作菜单

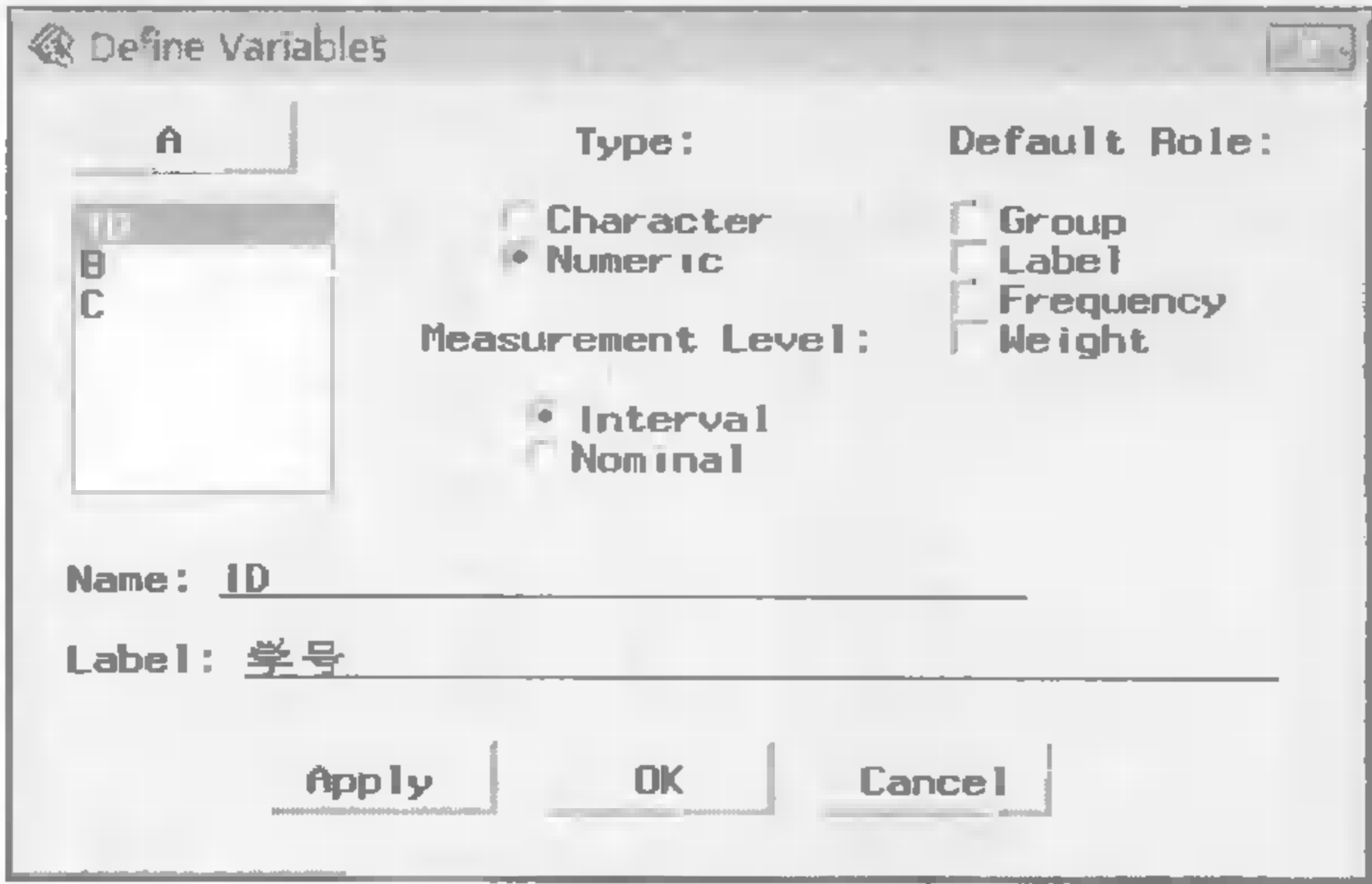


图 1-11 变量属性定义对话框

在变量属性定义对话框中，左上角的“A”按钮表示目前操作的数据集（默认为临时数据库中名为“A”的数据集），该按钮下的空白处列示了该数据集中所包含的所有变量的名字。在“Name”文本输入区中可以输入选中变量的名字，而在“Label”文本输入区可以输入该变量的标签。“Type”下有可供选择的两种变量类型的单选框，“Character”为字符型，“Numeric”为数值型。“Measurement Level”下的“Interval”单选框表示间隔测度水平，“Nominal”表示名义测度水平。“Default Role”项下的复选框可以指定变量在利用 SAS 系统进行分析或绘图过程中的作用，其中“Group”表示该变量可作为分组变量使用，“Label”表示该变量在分析过程中作为标签使用，“Frequency”表示该变量作为其他变量时每个观测值出现的频数，“Weight”表示该变量作为每个观测值的权重。

**STEP 6** 在本例中，在图 1-11 的对话框中依次选中 A 数据集中的 A、B、C 3 个变量，分别在“Name”文本输入区中输入“ID”、“Literature”、“Math”，把默认的变量名修改为例 1-4 中的变量名；在 Label 文本输入区中输入“学号”、“语文”、“数学”作为上述 3 个变量的标签。其他设置保持为默认选项。然后单击“OK”按钮返回 SAS/Insight 主界面。接着选择系统菜单“File→Save→Data”，弹出数据存储对话框，如图 1-12 所示。

**STEP 7** 选中 SASUSER 数据库，在“Data Set”文本输入区中输入“score”，然后单击“OK”按钮，即可在 SASUSER 永久库中建立一个名为“score”的数据集。



图 1-12 SAS/Insight 数据存储对话框

③ 利用 Analyst（分析员）建立数据集。

**STEP 1** Analyst（分析员）是 SAS 系统中的主要分析工具，涵盖了从描述统计、统计推断反多元统计分析等常见内容。在 SAS 系统工具栏上的命令输入框中输入“analyst”，或者选择系统菜单“Solutions→Analysis→Analyst”，打开分析员工具窗口，如图 1-13 所示。

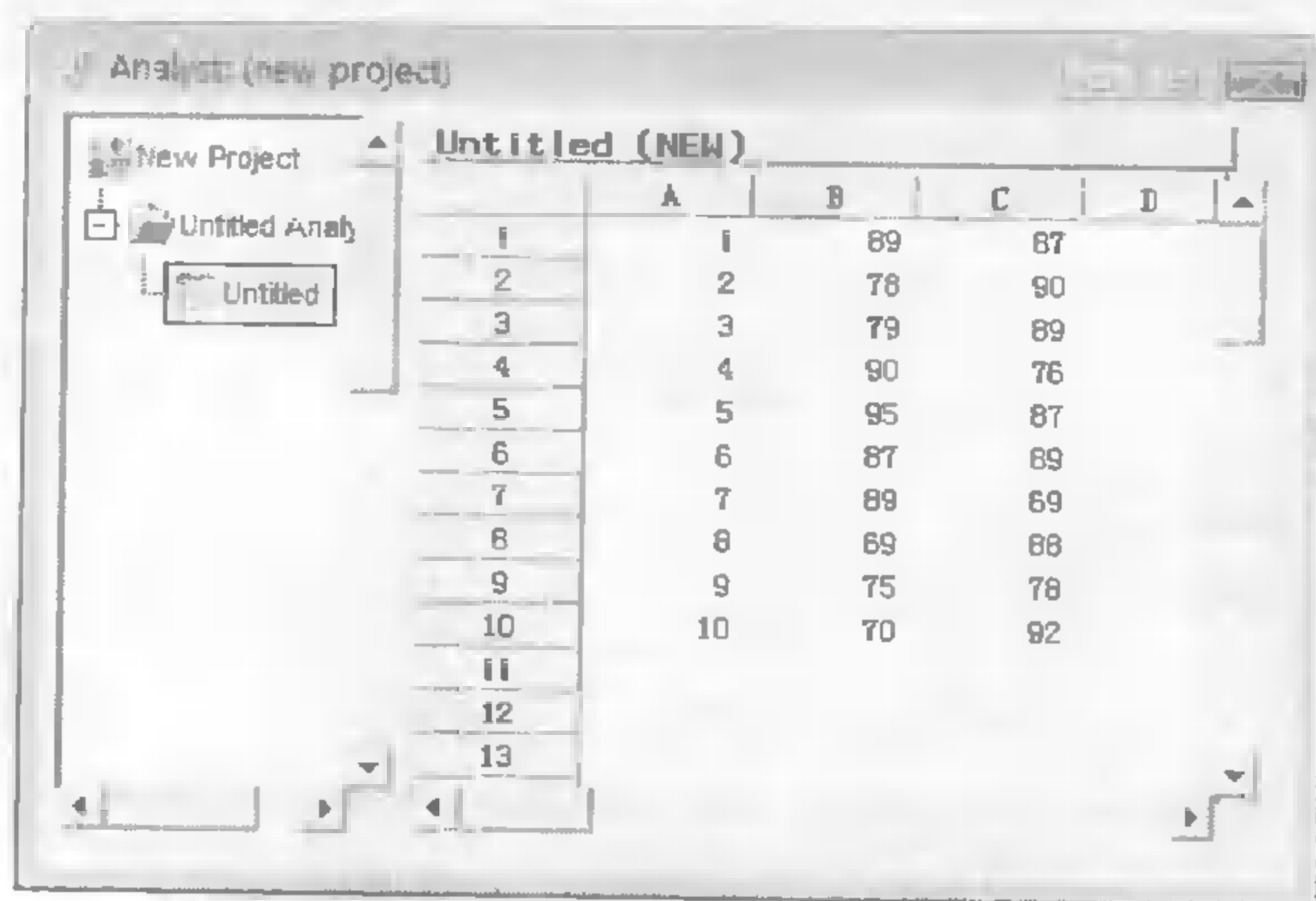


图 1-13 SAS/Analyst 分析员模块

分析员窗口的右半部分类似 Excel 表格，可以建立数据集。

**STEP 2** 以例 1-4 为例，用在 SAS/Insight 中输入数据的方式在表格的空白处输入对应的数据。接下来，只要用鼠标左键双击对应的变量名，便可以对变量名进行修改。如把“A”列修改为“ID”，只需双击“A”，然后输入“ID”即可把原来名为“A”的变量更名为“ID”。把所有的变量都设定好之后，选择系统菜单“File→Save”，在弹出的对话框中把新建的、名为“score”的数据集存储在 SASUSER 数据库中。

④ 利用 SAS 编程建立数据集。

在已有数据库中建立数据集时，主要使用 SAS 语言中的 DATA 步中的 INPUT 函数和 CARDS 选项。DATA 步用于指定数据集的名字，INPUT 函数的主要作用是指定变量及变量属性，并为对应的变量输入指定输入方式，其主要语法如下。

```
INPUT <变量名 1 变量名 2 .....变量名 n> <@@>;
```

CARD 选项则列示了所有变量对应的数据，可以为 INPUT 指定的变量读入数据，一直读到“;”为止。下面仍以例 1-4 为例，在 SASUSER 中建立名为“score”的数据集。

```
data SASUSER.score;                                /*在 SASUSER 数据库中建立名为“score”的数据集*/
  input id literature math @@;                       /*定义数据集中的变量和数据读入方式*/
```

```

label id="学号" literature="语文" math="数学"; /*定义数据集中变量的标签*/
cards;
  189  87  2   78  90  3   79  89  4   90  76  5   95  87
  687  89  7   89  69  8   69  88  9   75  78 10   70  92
;

```

例中 LABEL 语句可以为变量名设置标签。在利用程序输入数据以建立数据集的时候，应当注意数据输入的规则。CARDS 后面紧跟数据列表，并以“;”表示数据输入结束。

INPUT 语句中的“@@”表示按照 INPUT 定义的变量顺序依次连续读入数据，无论数据分为多少行，遇到“;”时则停止数据读入。如在上例中，系统一直按照“id→literature→math”的顺序读入 CARDS 的数据。首先读入 1、89、87，接下来读入 2、78、90，以此类推，读到“;”时停止数据输入。

如果没有“@@”符号，则表示系统按照行读入数据。如在上例中，系统首先按照 input 指定的顺序读入第 1 行中的 1、89、87，然后跳至第 2 行，读入 6、87、89，接着跳至第 3 行。由于第 3 行是“;”，故停止读入数据。因此，最后该数据集只有两行数据。

此外，还可以用 INPUT 语句读入字符型变量，这时需要在 INPUT 语句中指定的字符型变量名后加上“\$”符号。如在例 1-4 中，增加字符型变量“name”，其标签为“姓名”。

```

data SASUSER.score;
  input id name$ literature math @@;
  label id="学号" name="姓名" literature="语文" math="数学";
  cards;
    1 张三 89 87 2 李四 78 90 3 王五 79 89
    4 赵二 90 76 5 孙玉 95 87 6 张一 87 89
    7 李刚 89 69 8 黄源 69 88 9 陈强 75 78
    10 钱刚 70 92
  ;

```

此外，在已有数据集的情况下，也可以利用 SET 语句对数据集进行复制。如根据 SASUSER.score 数据集复制新的 Work.score 数据集，程序如下：

```

data score;
  set SASUSER.score;
run;

```

### 1.3.2 SAS 系统的外部数据文件

除了 1.3.1 小节介绍的在 SAS 系统中建立数据集的方法之外，对于在其他非 SAS 系统环境（如 Excel、Access、Lotus 等）中建立的数据文件，SAS 可以通过数据导入功能把其转换为 SAS 数据集。

#### 1. 利用 SAS/Import Data 菜单进行数据导入

SAS 系统的外部数据文件导入主要利用系统菜单中的“File→Import Data”来实现，如图 1-14 所示。

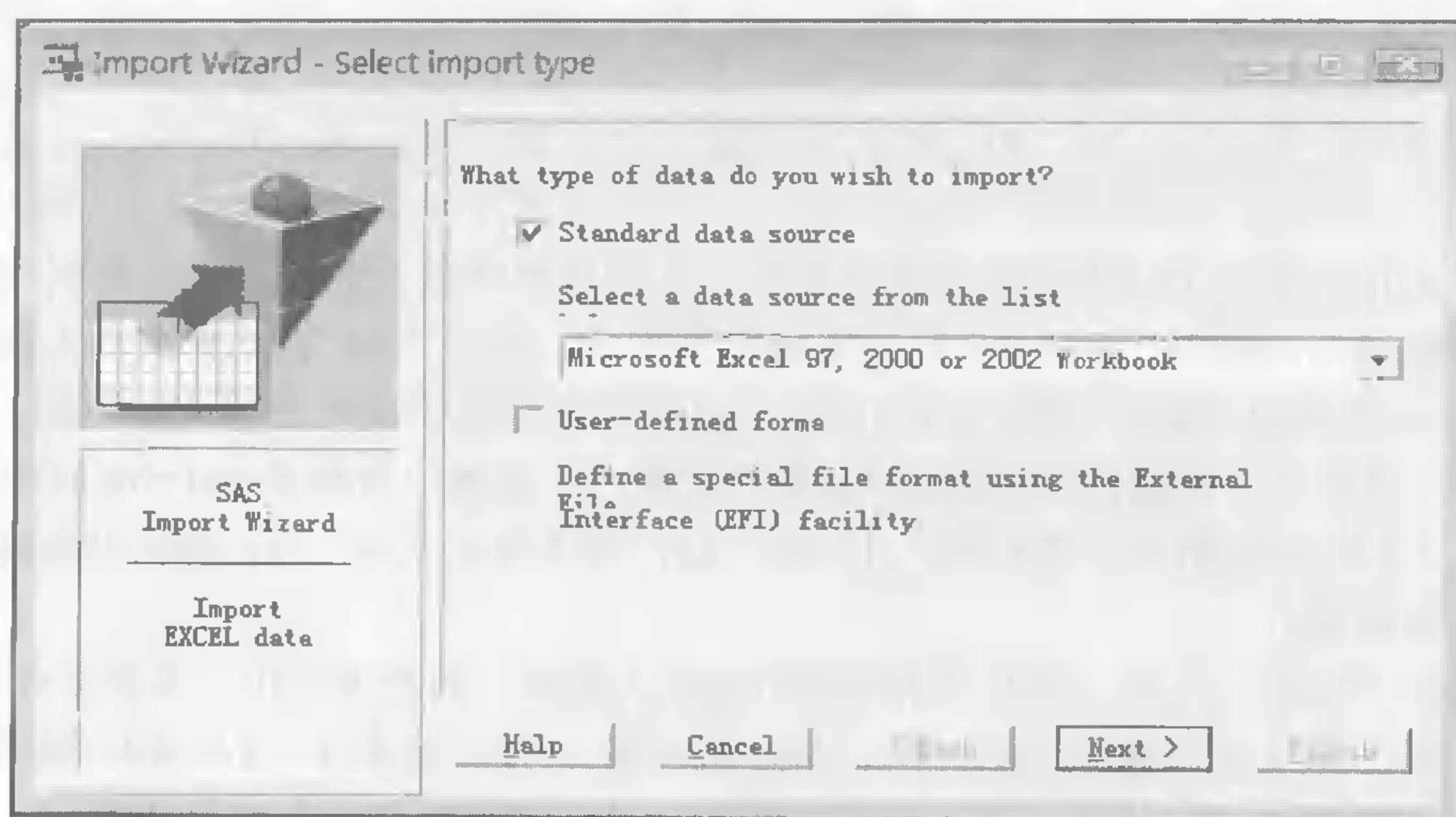


图 1-14 SAS 系统的数据导入对话框

SAS 数据导入程序，提供了两种方式的数据导入功能：一种是标准数据来源的导入，可以通过在该对话框中的“Select a data source from the list”下拉选单中选择对应的外部数据源文件实现；另一种是使用事先设定的数据格式进行数据导入，这需要事先编制好数据格式文件。在一般情况下，用户选择第一种标准数据格式即可。标准数据格式提供 Excel、Access（含 2 000 以上版本和 97 版本两种数据格式）、以逗号分隔的 CSV 格式、以 Tab 制表符分割的纯文本格式、带格式文本文件、DBF 数据库文件、JMP 数据库文件、Lotus 1-2-3 工作簿文件等诸多常用格式。

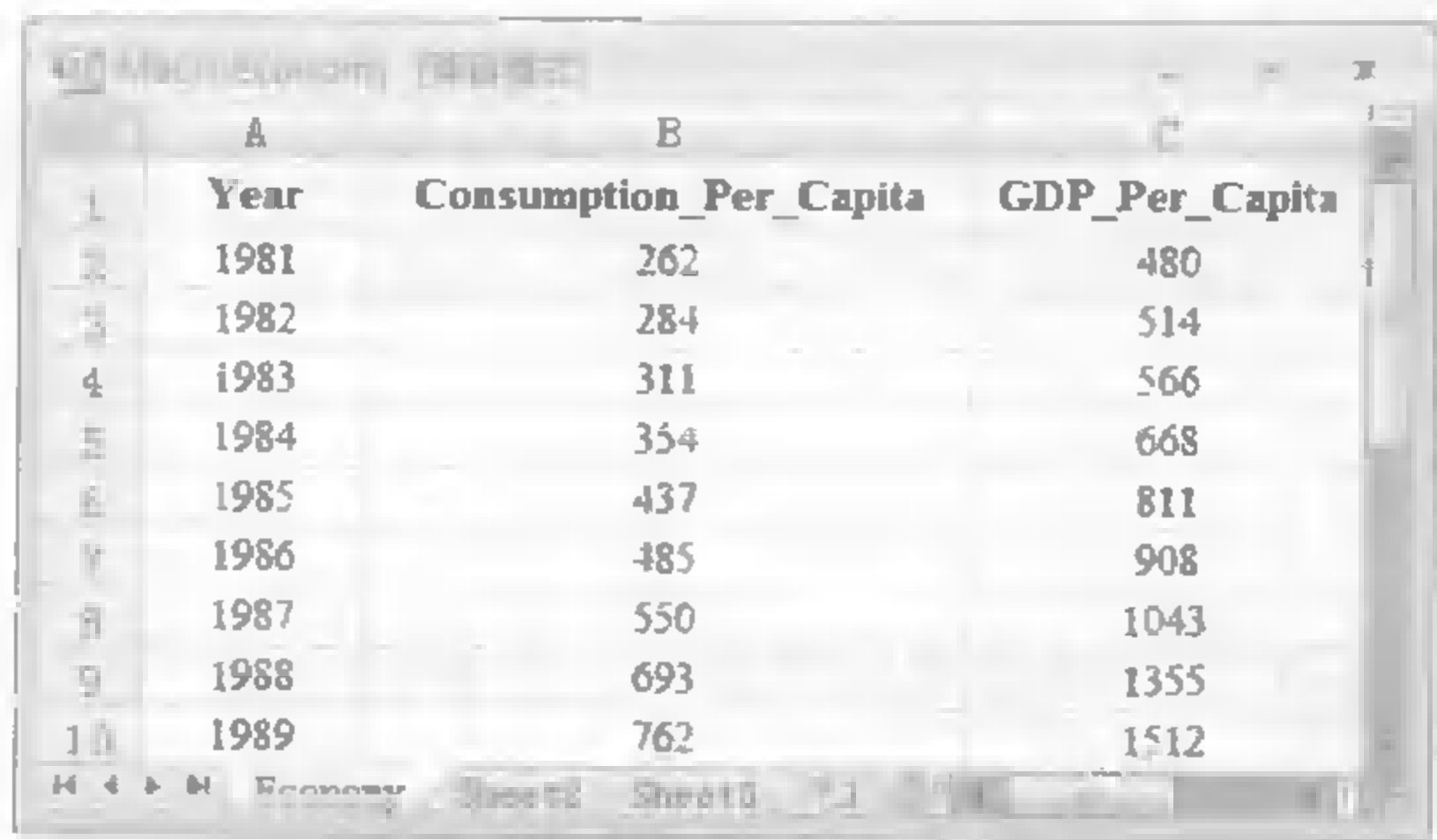
在数据导入对话框的左边，系统以图形和文字说明的形式标注当前数据导入操作步骤和过程的状态。



### 例 1-5

现有我国 20 世纪 80 年代的人均居民消费和人均 GDP 的数据，统计分析人员利用 Excel 2003 进行数据录入，建立了名为“MacroEconomy.xls”的 Excel 工作簿。该工作簿中一共有 3 个工作表，分别是 Economy、Sheet2、Sheet3，其中所有的数据存储在 Economy 工作表中，如图 1-15 所示。要求把该文件中的变量及其对应的数据导入至 SAS 系统的 SASUSER 数据库中，并建立名为“PerCapitaData”的数据集。

**STEP 1** 选择“File→Import Data”，在图 1-14 所示的下拉选单中选中第一项“Microsoft Excel 97, 2000 or 2002 Workbook”，单击“Next”按钮。在弹出的对话框中输入存放 MacroEconomy.xls 工作簿的路径（由于 SAS 系统对中文支持不完善，因此最好把数据文件放在英文路径下），单击“OK”按钮，弹出工作表选择对话框。然后在下拉选单中选择想要导入数据 Economy 的工作表，同时还可以单击“Options”按钮，设置导入数据时的一些规则。在本例中，按照默认的规则，即把 Excel 工作表中的第一行数据当作 SAS 数据集的变量名，进行数据导入。然后单击“Next”按钮，弹出 SAS 数据库和数据集选择对话框，如图 1-16 所示。



	A	B	C
1	Year	Consumption Per Capita	GDP Per Capita
2	1981	262	480
3	1982	284	514
4	1983	311	566
5	1984	354	668
6	1985	437	811
7	1986	485	908
8	1987	550	1043
9	1988	693	1355
10	1989	762	1512

图 1-15 我国 20 世纪 80 年代的宏观经济数据

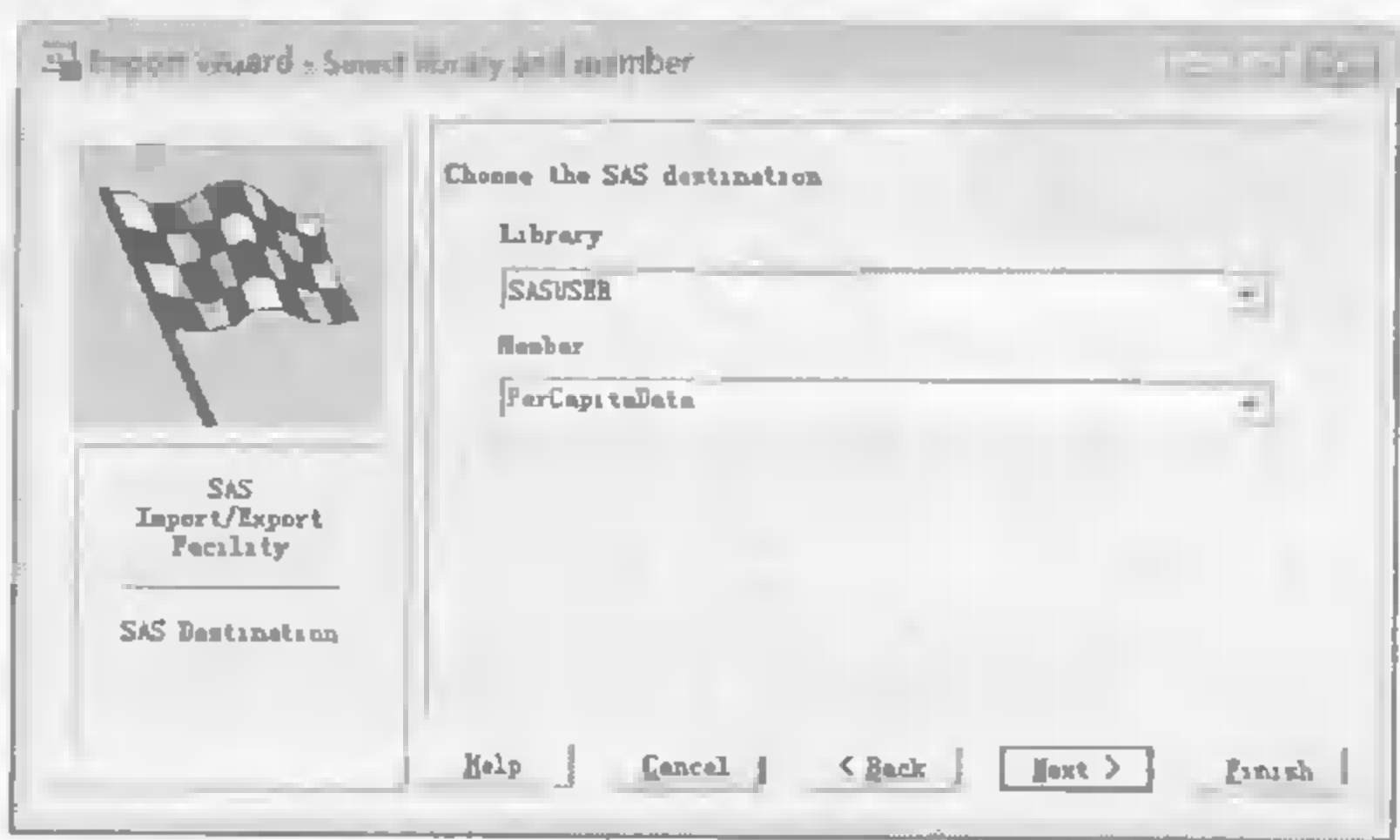


图 1-16 SAS 数据导入功能中的数据  
库和数据集选择对话框

**STEP 2)** 在“Library”下拉选单中选中 SASUSER 数据库，在“Member”文本输入框中输入数据集的名字“PerCapitaData”，单击“Next”按钮。或者直接单击“Finish”按钮，SAS 数据导入对话框自动关闭，数据导入过程完成。回到“Explorer”窗口，可以查看刚导入的 SASUSER 数据库中 PerCapitaData 数据集的具体内容。

利用 SAS 数据导入功能导入其他格式的数据的过程类似于例 1-5，这里不予赘述。

2. 利用 SAS 程序进行数据导入

可用 IMPORT 语句导入外部数据。仍以例 1-5 为例，利用 IMPORT 语句向 SASUSER 数据库中导入“D:\”路径下 MacroEconomy.xls 工作簿中的 Economy 工作表的数据，并命名为“PerCapitaData”数据集，程序如下：

```
proc import datafile="D:\MacroEconomy.xls" /*指定外部数据文件路径及文件名*/
out=sasuser.PerCapitaData; /*指定导入的数据库和数据集名*/
sheet='Economy'; /*指定导入外部 Excel 文件中的工作表*/
run;
```

1.4 数据预处理原理和基本方法

在数据分析过程中，获得进行统计分析和建模的对象（即数据）的过程也是必不可少的重要环节。除前几节讲述的数据库和数据集的建立及其基本操作方法之外，数据预处理的过程还包括数据整理、数据合并及分拆、数据清洗、数据变换等内容。

在数据预处理过程中，通常根据其自身特点把数据划分为脏数据和净数据。从广义上看，脏数据是指没有进行过数据预处理而直接收集到的、处于原始状态的数据，净数据则是指经过一定的选取、清洗、变换等数据预处理之后可以直接作为统计分析对象的数据。脏数据是数据预处理的基本对象，而净数据是数据预处理的目標和结果。

从狭义上看，这两种数据的区别主要是是不是符合研究要求，以及是否能够对其直接进行相应的数据分析。脏数据依照不同的分析目的具有不同的定义，如在常见的数据挖掘工作中，脏数据是指不完整、含噪声、不一致的数据；而在问卷分析中，脏数据则是指不符合问

卷要求的数据。在现实生活中，有许多脏数据的例子。



例 1-6

某咨询公司受某品牌汽车的委托，对该品牌汽车的满意度状况进行了调查。其中对购买了该品牌汽车的消费者有以下几个典型问题。

- A1. 您是否拥有某品牌的汽车？  
1. 是                      2. 否（停止问卷调查）

Q1. 您对某品牌汽车的总体满意程度如何？请打分（满意程度越高，得分越高，反之得分越低）。

1	2	3	4	5	6	7	8	9	10
---	---	---	---	---	---	---	---	---	----

- B1. 您去年的平均月收入是多少？请选择。  
1. 3 000 元以下      2. 3 000～5 000 元      3. 5 000～8 000 元      4. 8 000 元以上  
B2. 您家庭去年的平均月收入是多少？请选择。  
1. 3 000 元以下      2. 3 000～5 000 元      3. 5 000～8 000 元      4. 8 000 元以上

数据录入人员对 10 份该问卷进行了数据录入，录入结果如表 1-4 所示。

表 1-4                      某品牌汽车满意度调查结果

ID	Q1	B1	B2
1	7	3	3
2	8	2	3
3	10	3	5
4	9	4	4
5	10	4	3
6	6	3	3
7	.	2	4
8	7	4	4
9	5	3	4
10	11	2	4

利用编程方式（详见 1.3.1 小节）把例 1-6 中的数据存储在 SASUSER 数据库中的 Car 数据集中。

```
data SASUSER.Car;  
  input ID Q1 B1 B2;  
  cards;  
1 7 3 3  
2 8 2 3  
3 10 3 5  
4 9 4 4  
5 10 4 3  
6 6 3 3  
7 . 2 4  
8 7 4 4
```

```
9 5 3 4
10 11 2 4
run;
```

本节将以该数据集为例，从以下几个方面详细讲解数据预处理的主要内容。

1.4.1 数据整理

对于收集到的原始数据，在进行分析研究之前，应当对其进行整理，如对变量进行排序、调整变量在数据集中的顺序、改变变量数值的显示方式、增加/删除变量或样本等。

SAS 系统进行数据整理的主要功能是利用 SAS/Insight 模块实现的。

1. 对变量排序

**STEP 1)** 进入 SAS/Insight，打开 SASUSER.Car 数据集，在数据区域单击鼠标右键，或者单击最左上角的  按钮，弹出数据操作菜单，如图 1-10 所示。选择“Sort”，弹出变量排序对话框，如图 1-17 所示。

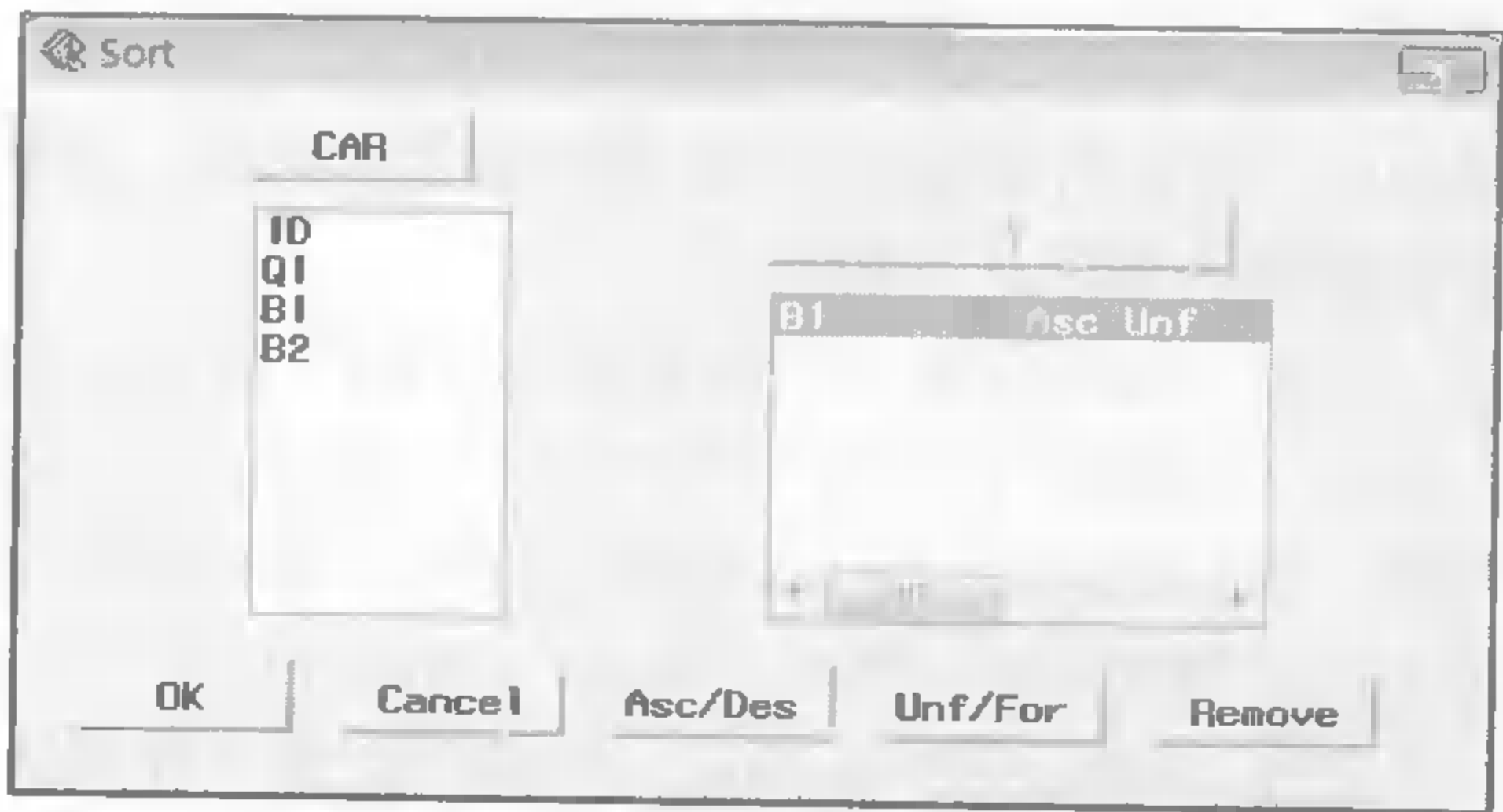


图 1-17 变量排序对话框

**STEP 2)** 在“CAR”按钮下面的数据集中，选择想要进行排序的变量。如对“B1”变量进行降序排序，则可选中“B1”变量，然后单击右边的“Y”按钮。这时，“B1”会被自动放置在“Y”按钮下的空白区域中。然后在该区域选中“B1”变量，单击“Asc/Des”按钮（“Asc”表示升序，“Des”表示降序，默认是升序排列）。以此类推，可以把若干个变量同时选中以放置在该空白区域中，并分别指定排序方式。单击“Remove”按钮可以移除不需要排序的变量。然后单击“OK”按钮，即可在 SAS/Insight 主界面中看到经过排序后的数据。同样也可以用编程方式进行排序。

```
proc sort data=SASUSER.Car;          /*使用 sort 过程对 SASUSER.Car 数据集进行排序*/
  by descending B1;                  /*按照 B1 变量进行降序排列*/
proc print;                          /*在“Output”窗口中显示 SASUSER.Car 数据集*/
run;
```

2. 调整变量在数据集中的顺序

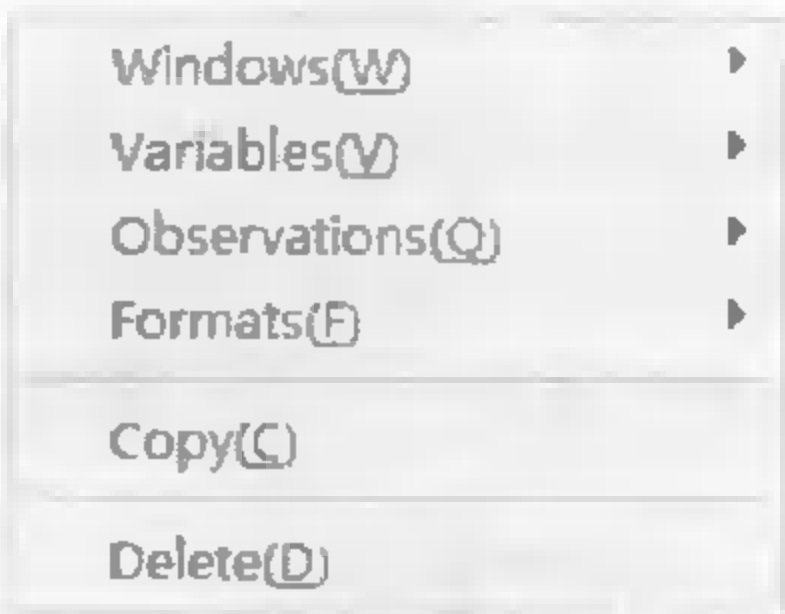
有时候需要调整变量在数据集中的顺序。如把 SASUSER.Car 中的“Q1”作为第 1 个变量，以加强数据分析人员对汽车满意度打分的重视程度，并且把问卷编号变量“ID”作为最后 1 个变量。

进入 SAS/Insight，打开数据操作菜单，首先选择“Move to First”，在弹出的变量移动对话框中选中将要放在第 1 位置的“Q1”变量，单击“OK”按钮；或者直接在数据表上选中“Q1”的变量名，再单击数据操作菜单中的“Move to First”，便会看到“Q1”已经出现在第

1 个变量的位置，而其他变量的相对位置保持不变。然后继续从数据操作菜单中选择“Move to Last”在弹出的变量移动对话框中选中要放在末尾位置的“ID”变量，单击“OK”按钮，“ID”变量便会出现最末位置，而其他变量的相对位置保持不变。

3. 调整变量值格式

变量值的表现可以有多种形式，如数值型数据定义为总长度 10、小数位数为 2 的显示格式，日期型数据可以显示为年/月/日或日/月/年等多种形式。在 SAS/Insight 中，可以对变量值的形式进行多种定义，主要使用 SAS/Insight 中的“Edit”菜单下的“Formats”二级菜单，如图 1-18 所示。



“Formats”二级菜单中提供了常用的 8 种数值型数据显示格式的快捷方式，同时也可以可以在“Other”选项中自定义各变量的显示形式。在自定义显示形式过程中，系统会预览数据格式。

4. 设定变量值的标签

在建立数据集的过程中，可以用 label 语句设置变量名的标签。在某些情况下，不光变量名具有标签，其对应变量的值也可以具有标签。

如例 1-6 中的“B1”小题，如果设置一个变量名为“B1”以表示该道题的选择情况，则“B1”对应的值可能有 4 个，分别是“1”、“2”、“3”、“4”。通常，在 SAS 系统中输入数据时，为了数据录入、运算方便，往往把类似“B1”这种选择性的题目用数字代替选项进行输入。这些输入的数字与问卷上的选项相对应，如“1”表示“3 000 元以下”、“2”表示“3 000~5 000 元”等。那么一旦离开了问卷，谁也不知道“B1”变量中的数字代表什么意思，这时可以为变量值挂上标签，即在系统中指定“1”代表“3 000 元以下”、“2”代表“3 000~5 000 元”等。

设定变量值标签可利用 FORMAT 过程实现，具体程序如下。

```
proc format;
  value B1_Fmt 1='3 000 元以下'      /*指定一个变量值标签对应关系，并命名为“B1_Fmt”*/
               2='3 000~5 000 元'    /*等号左边为变量值，等号右边为变量值的标签*/
               3='5 000~8 000 元'
               4='8 000 元以上';
run;
```

FORMAT 过程只是设定了变量值与其标签的对应关系，该对应关系用用户自己指定的名字表示。在对某个变量进行分析时，必须在对应的分析过程中用 FORMAT 语句引用变量值标签对应关系，才能达到目的，如以下程序。

```
data format_demo;
  input B1@@;
  cards;
    1 2 3 4 4 4 3 2 3 1 1 3 4 2
  :
run;
proc print data=format_demo;
run;
proc print data=format_demo;
  format B1 B1_Fmt.; /*引用 B1_Fmt 变量值标签对应关系，注意变量值标签名称后要加上“.”*/
run;
```

读者可以自行比较两个 PRINT 过程的输出结果。

5. 增加/删除变量或观测值

在 SAS/Insight 窗口中增加变量，只需在对应的位置上输入数据即可，具体方法详见 1.3.1 小节。“Edit” 菜单可以实现删除变量或观测值的功能，如图 1-18 所示。只要单击选中想要删除的变量名字或观测值对应的样本号，然后选择 “Edit→Delete”，系统便会自动删除对应的内容。

删除变量通常也可用编程方式实现。如在例 1-6 中，需要删除问卷编号 “ID” 变量，可以采用以下程序。


```
data Car;                                /*指定名为 Car 的临时数据集*/
  set SASUSER.Car;                       /*从 SASUSER.Car 数据集中读入数据*/
  drop ID;                               /*在 Car 临时数据集中删除 ID 变量*/
run;
```

上面使用 DROP 语句删除指定数据库中的变量，同样也可以用 KEEP 语句达到删除变量的目的，如下所示。

```
data Car;
  set SASUSER.Car;
  keep Q1 B1 B2;                         /*在新临时数据集 Car 中保留 Q1、B1 和 B2 变量*/
run;
```

6. 生成数据子集

生成原始数据集的子集主要是指在原始数据集的基础上，从中选取部分变量或部分观测值以组合成新的数据集，而原始数据保持不变。可以通过 SAS/Insight 中的 “Extract” 功能实现。

如从例 1-6 的 SASUSER.Car 数据集中，选取变量 “Q1”、“B2” 及其对应的第 4、7、9、10 个观测值以建立数据子集。按住键盘上的 Ctrl 键，并用鼠标选中 “Q1”、“B2” 的变量名以及编号为 4、7、9、10 的观测值，然后单击最左上角的  按钮或在数据区域中单击鼠标右键以弹出菜单，如图 1-19 所示。

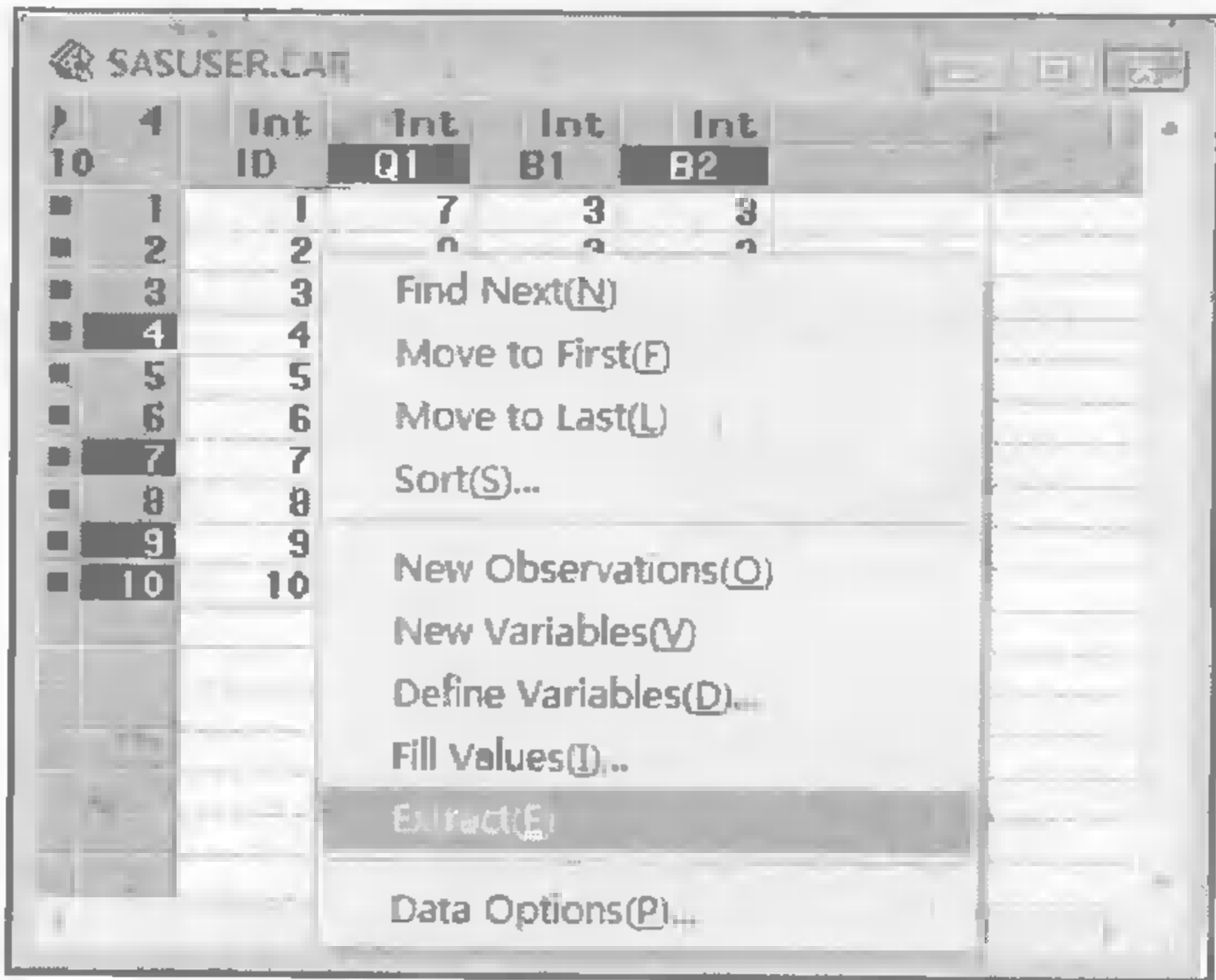


图 1-19 “Extract” 提取数据子集功能

选择 “Extract” 功能，则系统自动弹出一个 SAS/Insight 数据窗口，并自动把该数据子集命名为 SASUSER.Car1。

1.4.2 数据分拆与合并

数据分拆与合并是对整个数据的变量及其观测值进行操作，它是常用的数据预处理过程。

1. 数据分拆

在数据预处理时，有时需要将数据集按照一定的规则拆分为若干个数据集。如在例 1-6 中，需要分别考察低收入和高收入人群对汽车的满意程度，因此必须把总体数据按照个人收入变量，即“B1”变量分拆至 Car\_Low 和 Car\_High 两个数据集以分别进行考察。数据分拆可以用 Output 语句结合 Select 语句实现。

```
data SASUSER.Car_Low SASUSER.Car_High;
  set SASUSER.Car;
  select;
    when (B1<=2) output SASUSER.Car_Low;      /*把低收入的观测值放在指定数据集中*/
    otherwise output SASUSER.Car_High;         /*把其他收入的观测值放在指定数据集中*/
  end;
run;
```

2. 数据合并

有时需要把若干个数据集的信息合并起来以综合考察，如在一次调查活动的数据录入工作中，不同的录入人员分批次录入了多个数据文件，研究人员需要把这些文件进行合并，从而反映整体信息。又如学生考试成绩，可把不同科目的分数合并，并反映在一个数据文件中。根据不同的实际情况，数据合并又可以分为纵向合并与横向合并。

(1) 纵向合并。

纵向合并是指把若干个数据集的观测值按相同的变量进行数据追加。在通常情况下，要求参加合并的各个数据集的结构要相同，如上述反映低收入群体满意度数据集 SASUSER.Car\_Low 和反映高收入群体的满意度数据集 SASUSER.Car\_High 具有相同的变量名和数据结构，故可进行纵向合并，并将合并后的数据集命名为 SASUSER.Car\_Total。在 SAS/Analyst 中，可以进行数据的纵向合并。

**STEP 1** 进入 SAS/Analyst，在其界面下的任意地方单击鼠标右键，可弹出 SAS/Analyst 的系统菜单(弹出的菜单与进入 Analyst 之后 SAS 主界面的菜单一样)，选择“Data→Combine Tables→ Concatenate By Rows”，弹出数据纵向合并对话框，如图 1-20 所示。

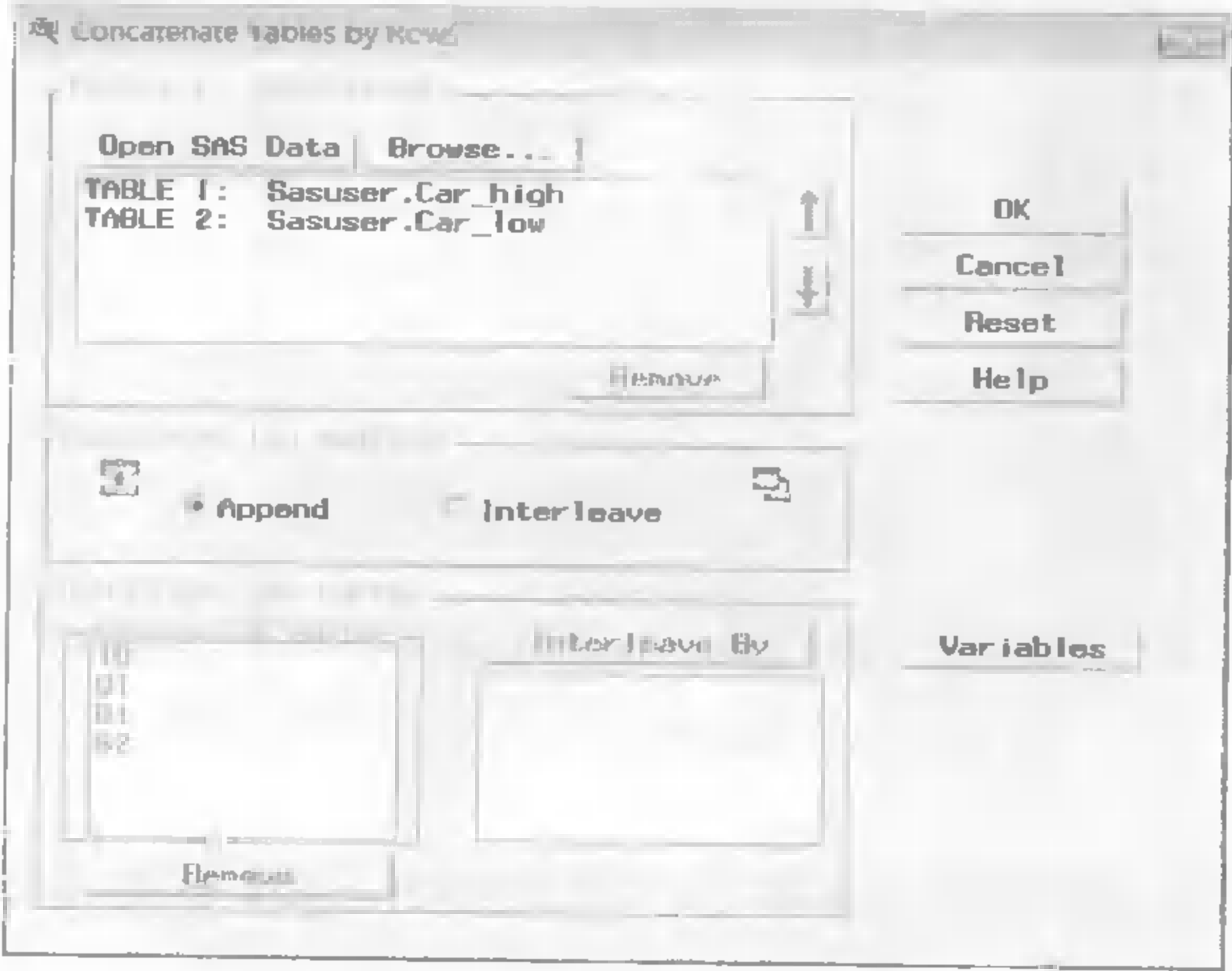


图 1-20 数据纵向合并对话框

**STEP 2** 在该对话框中，依次单击“Open SAS Data”按钮把对应的 SASUSER.Car\_high 和 SASUSER.Car\_low 数据集添加到列表框中，系统会自动在左下方的“Common variables”列表框中显示各数据集共有的变量名，然后单击“OK”按钮即可完成数据集纵向合并，再选择“File→Save As By SAS Name”进行数据集保存。数据集纵向合并也可用以下程序实现。

```
data SASUSER.Car_Total;  
  set SASUSER.Car_Low SASUSER.Car_High; /*SET 语句后列示参加纵向合并的数据集*/  
run;
```

(2) 横向合并。

横向合并是指把若干个数据集的变量按照一定的关键变量进行变量追加。如把学生的各门课程期末考试成绩进行汇总。学籍档案中有一个名为 SASUSER.Student\_Profile 的数据集，存储了学生的学号、姓名等信息，现有另一个名为 SASUSER.Student\_Score 的数据集存储了学生的考试成绩，为了综合考察学生学习成绩的总体情况及其影响因素，需要把这两个数据集按照学号（“ID”）变量进行横向合并，如图 1-21 所示。

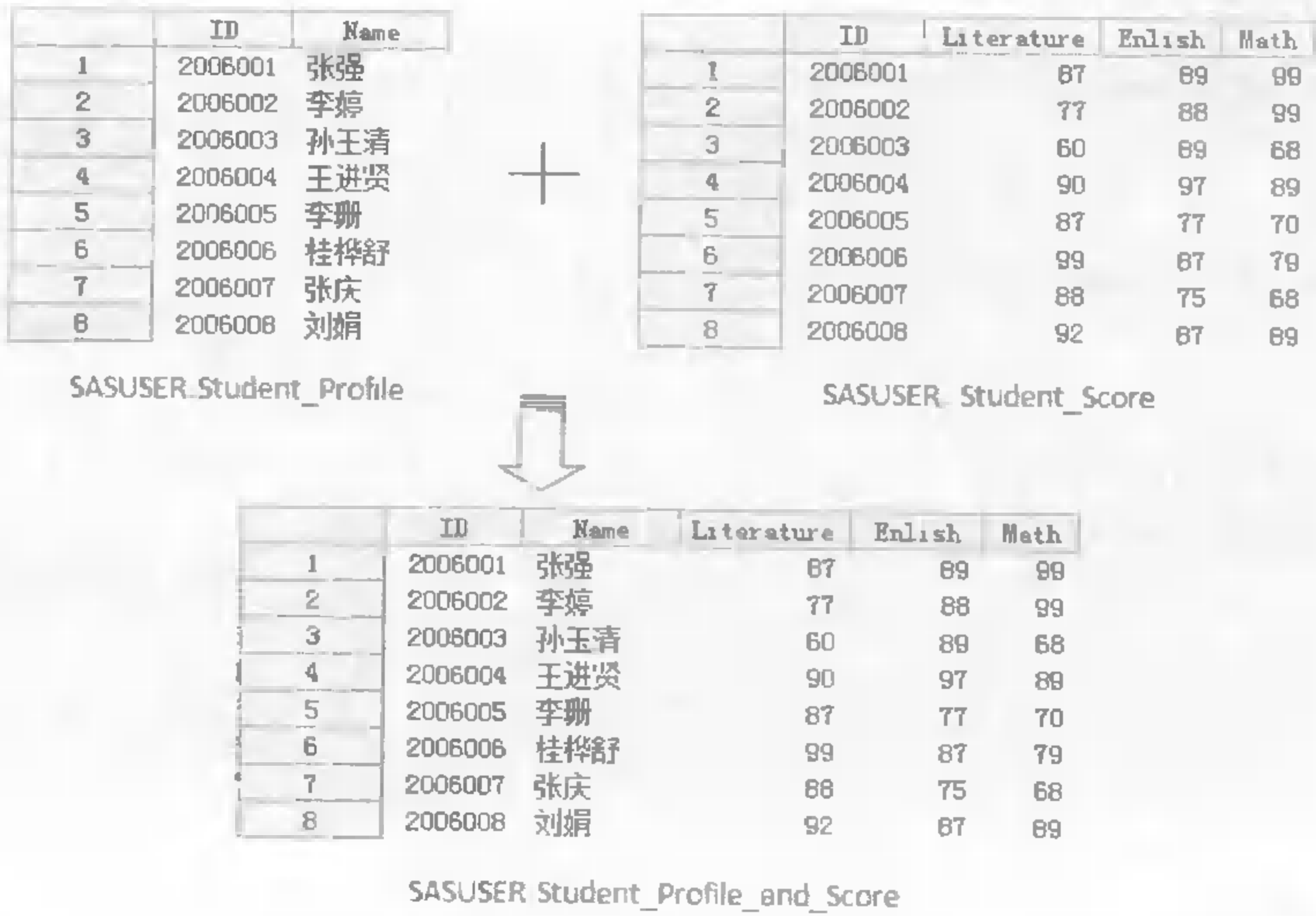


图 1-21 数据横向合并过程

为了把参加合并的各数据集的数据正确对应上，在数据集横向合并过程中，应当首先指定参加合并的数据集中共有的变量，并把该变量作为关键字，再把各数据集按照该关键字进行排序，然后再按照关键字把数据集合并起来。

**STEP 1)** 进入 SAS/Analyst，选择“Data→Combine Tables→Merge By Columns”，弹出数据横向合并对话框，如图 1-22 所示。

**STEP 2)** 在该对话框的“Table 1”和“Table 2”中选择 按钮，分别把 SASUSER.Student\_Profile 和 SASUSER.Student\_Score 两个数据集选中（如果有多个数据集参加合并，则可以单击右边的“More”按钮进行添加）。在“Combined table will keep”栏下选择“All rows”单选框，然后从“Common variables”中选中“ID”变量，单击右边的“Merge By”按钮，把“ID”变量放置在“Merge By”按钮下的列表框中，再单击“OK”按钮，系统自动弹出合并后的数据。最后选择系统菜单“File→Save As By SAS Name”进行数据集保存。

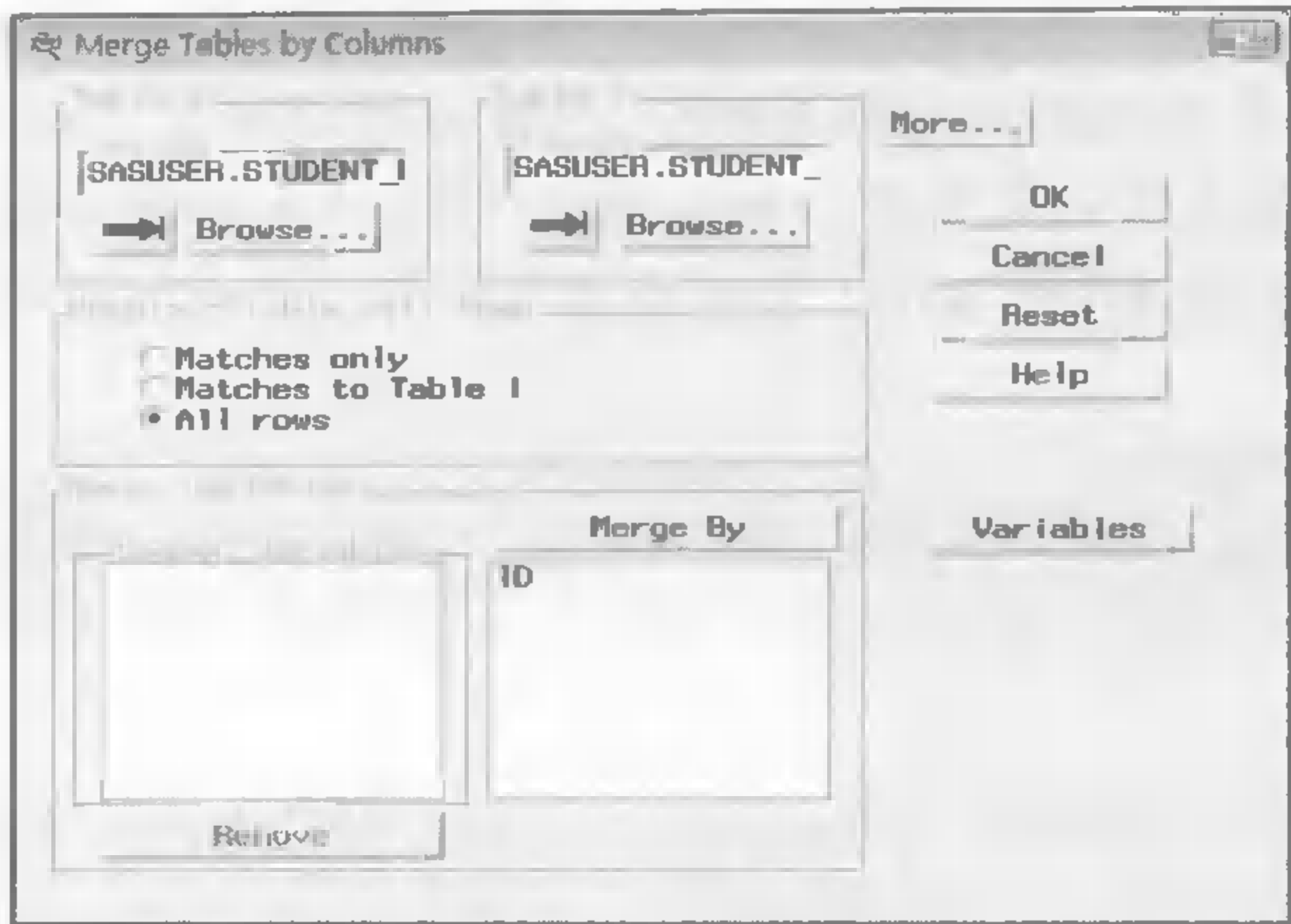


图 1-22 数据横向合并对话框

纵向合并程序主要利用 MERGE 语句来进行。注意在进行合并之前，应当按照合并关键字进行排序。本例程序如下：

```
proc sort=SASUSER.Student_Profile;
  by ID;                                /*按 ID 变量排序*/
run;
proc sort=SASUSER.Student_Score;
  by ID;                                /*按 ID 变量排序*/
run;
data SASUSER.Student_Total;
  merge SASUSER.Student_Profile SASUSER.Student_Score;
  by ID;                                /*把 ID 变量作为合并关键字*/
run;
proc print;
run;
```

1.4.3 数据清洗

数据清洗往往是数据分析工作中最为重要且最容易被忽视的一个基础性内容。数据清洗工作的主要内容是把数据集中的脏数据找出来并修正，以尽量降低非法数据对分析结果的影响。

1. 数据查错

在对例 1-6 数据集 SASUSER.Car 进行审核的过程中很容易发现，“Q1”变量的第 10 个样本数值为 11。而根据问卷中的要求，“Q1”变量的取值范围是从 1 到 10，显然“Q1”的第 10 个样本超出了该范围，不符合调查研究的要求。同样，对于变量“B2”而言，其第 3 个样本的数值为 5，同样也超过了备选答案的编号。类似于这种需要把不符合要求的数据从数据集中找出来并进行修改的过程被称为数据查错。

**STEP 1** 利用菜单方式进行数据查错的方法如下。

进入 SAS/Insight，打开 SASUSER.Car，选择“Edit→Observations→Find，弹出数据查找

对话框，如图 1-23 所示。

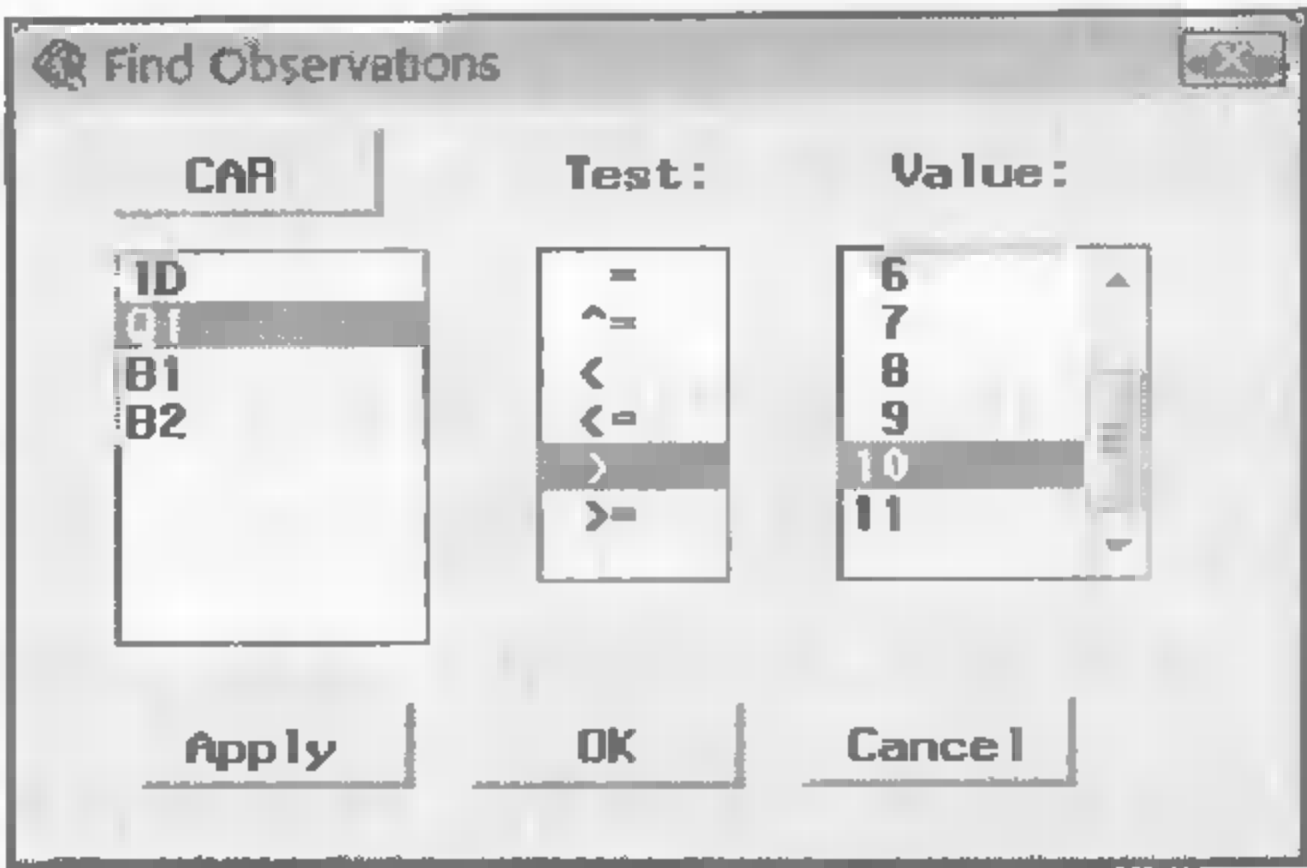


图 1-23 SAS/Insight 的 Find 对话框

**STEP 2** 在该对话框中“CAR”按钮下选择“Q1”，在“Test”列表框中选择“>”符号，在“Vaule”列表框中选择“10”，则表示从数据集中找出“Q1”变量的观测值大于 10 的样本。单击“OK”按钮后，满足该指定条件的观测值会在 SAS/Insight 中标示出来。如本例中符合该条件的观测值是第 10 个，即在系统中标示为 10。对于“B2”变量也可以进行同样操作，以找到对应的出错样本观测值，并直接对其进行修改。

在 SAS 系统中，采用菜单方式进行数据查找虽然直观，但是在样本量比较大的时候，由于数据显示的原因会变得非常不方便，而且菜单操作方式在查找的同时不能进行数据修正。因此，通常也可用编程方式实现查找功能，并实时更新错误的数据。

查找数据时主要使用结构化查询语句（SQL），具体程序如下。

```
proc sql;                                /*调用 SQL 语言*/
  select id, q1                          /*从 SASUSER.Car 数据集中选中 ID、Q1 变量*/
  from SASUSER.Car
  where q1>10;                          /*指定查找条件*/
run;
```

运行程序之后，在“Output”窗口中显示如下。

The SAS System		22:33 Wednesday, March 24, 2 008	1
ID	Q1		
10	11		

该输出结果表明“ID”为 10 的样本对应变量的“Q1”的观测值为 11，即找出的出错样本。同理，对于变量“B2”的查错程序如下。

```
proc sql;
  select id,b2
  from SASUSER.car
  where b2>4;
run;
```

程序提交运行后，显示“ID”为 3 的样本对应变量的“B2”的观测值为 5，即找出的出错样本。

## 2. 检查逻辑关系

除了能够比较直观地将脏数据呈现出来之外，有时脏数据还暗含在数据的逻辑关系当中。

如在例 1-6 中，调查问卷中的“B1”和“B2”问题是有一定逻辑关系的，即被调查者的个人收入不能够大于其家庭收入。这种问题常常被用在实际的问卷调查中，以发挥鉴别问卷主观信度的重要作用。因此，在数据录入过程中，这两个变量之间暗含的逻辑关系便是“B1” $\leq$ “B2”。按照该逻辑关系检查所录入的数据，可以很容易发现第 5 个样本在该两个变量的逻辑关系中是错误的，即“B1” $>$ “B2”。在 SAS 中，可以很方便地通过菜单方式实现变量之间逻辑关系的判断，并找出对应变量的逻辑关系的样本。

**STEP 1** 在“Explorer”窗口中找到例 1-6 中的 SASUSER.Car 数据集，双击鼠标左键，弹出该数据集的 ViewTable 窗口。然后选择系统菜单“Data→Where”，弹出 Where 对话框，如图 1-24 所示。

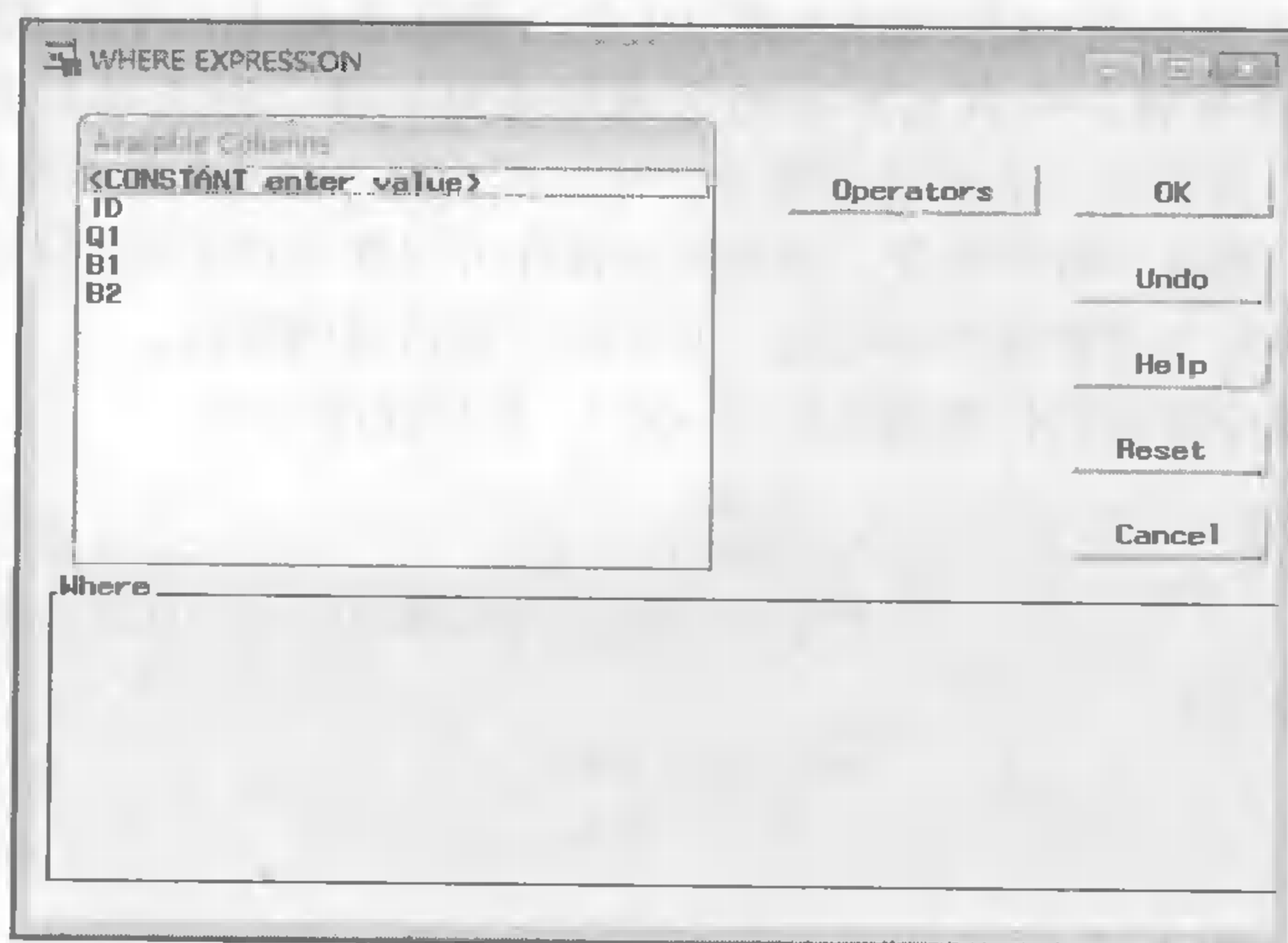



图 1-24 Where Expression 窗口

**STEP 2** 在“Available Columns”列表中框选中“B1”，然后单击“Operators”按钮，选择“GT”（即大于）。接着在“Available Columns”中选中“B2”，系统会自动在对话框下方的“Where”区域中显示“B1 GT B2”表达式。单击“OK”按钮，系统自动返回 ViewTable 窗口，这时该窗口自动显示根据用户设置的逻辑关系找出的样本，如在本例中出现第 5 号样本，其他不符合逻辑关系的样本则自动被隐藏。如果想对数据进行修改，须选择系统菜单“Edit→Edit Mode”或者单击工具栏上的按钮才能进行修改。选择系统菜单“Data→Where Clear”则清除逻辑关系查找，返回数据集中的所有数据。同样，可以使用程序进行逻辑关系查找，如下所示。

```
proc sql;
  select id,b1,b2
  from SASUSER.car
  where b1>b2;
run;
```

程序提交运行后，“ID”显示为5的样本对应变量“B1”和“B2”的观测值分别为4和3，即找出的、逻辑关系出错的样本。

### 3. 数据修正

对于出错样本，可以利用 SAS/Insight、SAS/Analyst 和 ViewTable 窗口进行修改。

为提高数据查错和更正数据的效率，通常利用 UPDATE 语句对数据集进行数据修正。例如，经过认真核对编号为10和3的问卷内容，发现编号为10的问卷在“Q1”这道题上的所选项目应该是“1”，编号为3的问卷在“B2”这道题上的选择应该是“2”，编号为5的问卷在“B1”这道题上选择的应该是“2”。可以利用以下程序进行数据修正。

```
data SASUSER.Car_Upd;          /*建立存储更新信息的数据文件，无须更新的数据用缺失值表示*/
    input id q1 b1 b2;
    cards;
        3 . 2
        5 . 2
        10 1 .
    ;
proc sort SASUSER.Car_Upd;
    by id;
run;
proc sort SASUSER.Car;
    by id;
run;
data SASUSER.Car_Renew; /*使更新后的数据存储在 SASUSER.Car_Renew 数据库中*/
    update SASUSER.Car SASUSER.Car_Upd;
    by id;
run;
proc print data=SASUSER.Car_Renew;
run;
```

在进行数据更新时要注意，应当先分别对被更新的数据集和更新数据集按照关键变量进行排序。

#### 1.4.4 数据变换

数据变换主要是指变换数据的类型、表达方式以及根据函数运算变换数据的数值，也是常见的数据预处理方法之一。数据类型变换主要是变换数据的属性，如把数值型数据变换为日期型或字符型，或把日期型数据转换为数值型数据。该部分内容在定义变量属性的过程中已经涉及，本小节不再赘述。

##### 1. 数据函数变换

数据函数变换是指利用函数对数据进行运算，把原数据变换成函数运算的结果或将原数据按照函数关系生成新变量。在 SAS 系统中，可以通过 SAS/Analyst 模块实现数据函数。

**STEP 1)** 仍以例 1-6 中的 SASUSER.Car 为例，在 SAS/Analyst 中，选择系统菜单“File → Open By SAS Name”，打开 SASUSER 中的 Car 数据集。然后选择系统菜单“Edit → Mode

→Edit”，打开数据编辑模式。选中要进行函数变换的变量，选择系统菜单“Data→Transform”，打开数据变换菜单，如图 1-25 所示。

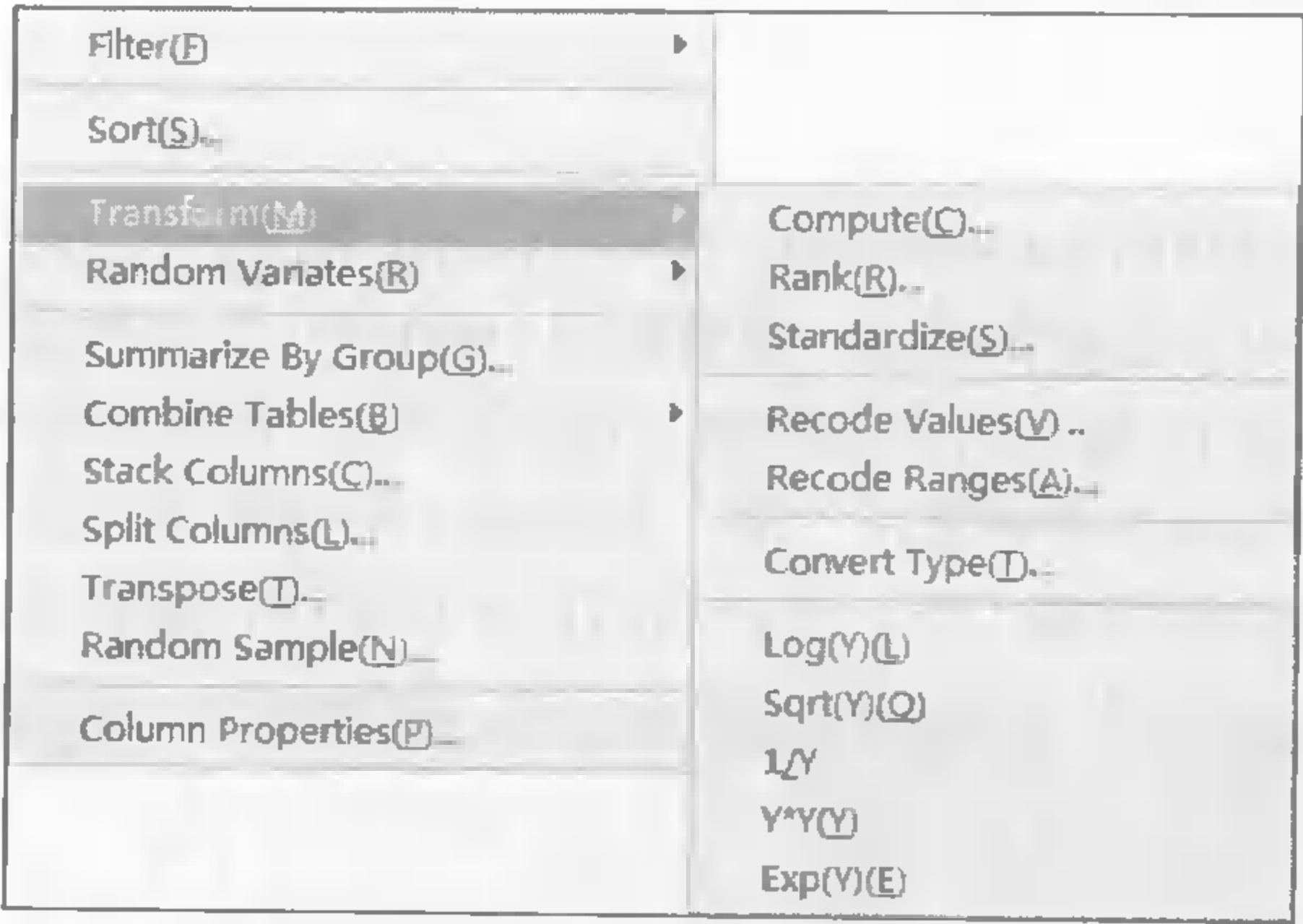


图 1-25 SAS/Insight 的数据变换菜单

**STEP 2** 在“Transform”菜单中，用户可以通过“Compute”功能进行函数变换。同时，该菜单的二级菜单中也列示了常用的变换函数，直接单击对应的函数名便可在数据集中生成新的变量。选择“Convert Type”则可以实现变量在数值型和字符型数据之间进行属性转变。单击“Compute”，打开函数变换对话框，如图 1-26 所示。

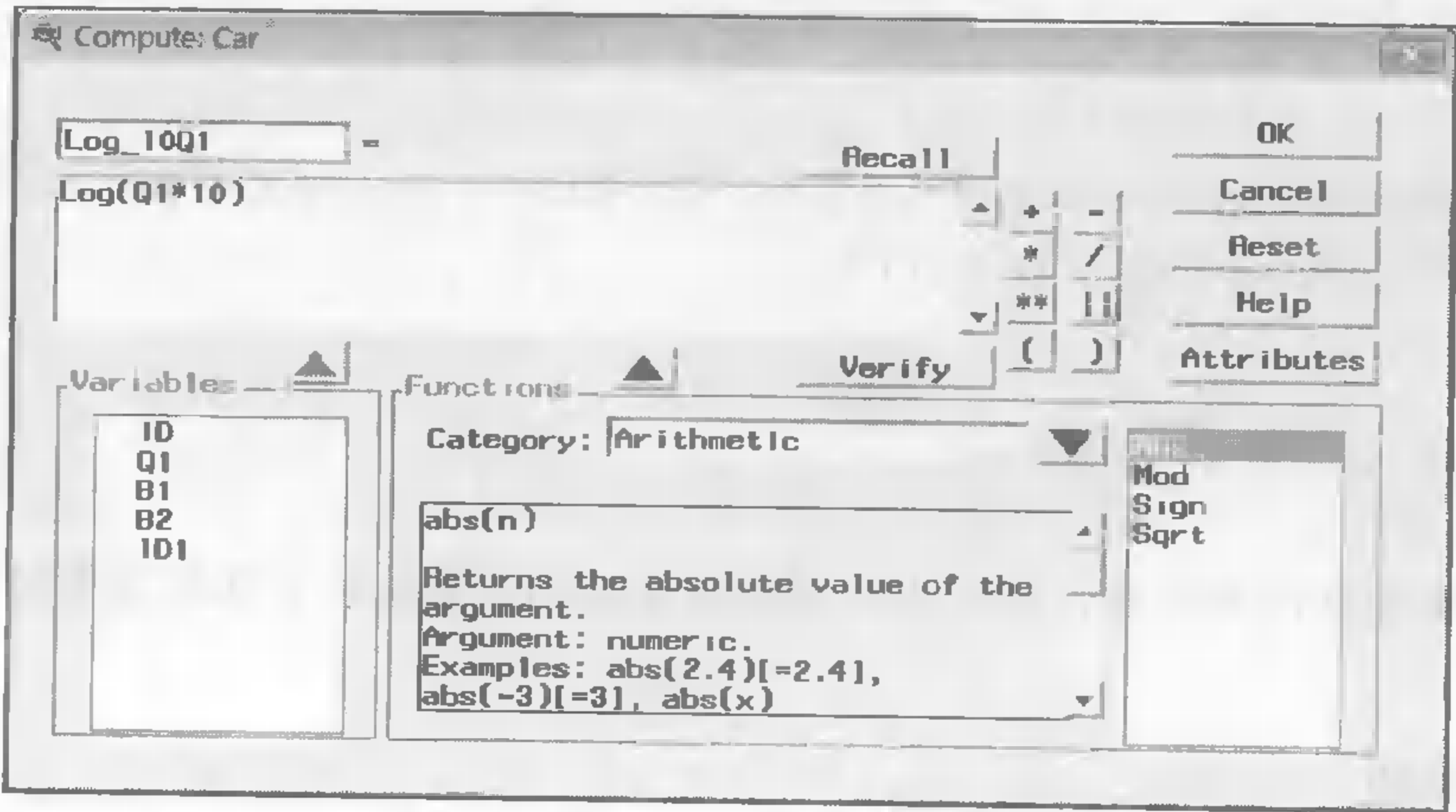


图 1-26 “Compute”对话框

如把例 1-6 中的“Q1”变量进行变换，求“Q1”数值扩大 10 倍之后的对数，用“Log\_10Q1”变量名表示。在该对话框中的最左上方的文本输入框中输入变换后的新变量的名字，在中间的文本输入框中输入变量函数表达式，如图 1-26 所示。除此之外，还可以利用 SAS 系统默认的函数进行数据变换，其“Category”下拉列表框提供了算术、字符、日期、数学、概率等 12 大类常用函数，用户可根据需要选择具体的函数表达式。当选中具体函数时，“Category”列表框下的文本输入框会自动显示该函数的功能及用法。

**STEP 3** 在文本输入框中输入表达式之后，可单击“Verify”按钮检查函数形式和参数设置是否正确。然后单击“OK”按钮，就会在 SAS/Analyst 窗口中按照函数运算结果生成新的变量。

## 2. 数据标准化

数据标准化的主要作用是使各种具有不同计量单位的变量可以进行对比分析,或者出于研究的需要,把各变量转化为均值、方差相同的新变量。标准化也称之为同量纲化。实现数据标准化的方法有很多,SAS 系统默认的方法是 Z-Score(Z 得分)法。如把变量  $X$  进行 Z-Score 变化,具体公式为:  $Z = \frac{X - \mu_X}{\sigma_X}$ 。其中,  $\mu_X$  和  $\sigma_X$  分别表示  $X$  的均值和标准差。

按照如上公式,可把变量  $X$  标准化为均值为 0、标准差或方差为 1 的数列,这也是最为常见的标准化方法。

**STEP 1)** 主要利用 SAS/Analyst 系统菜单“Data→Transform→Standardize”来实现数据标准化。如对 SASUSER.Car 中的满意度得分变量“Q1”进行标准化,使得标准化后的满意度得分均值为 0、标准差为 1。选择上述系统菜单打开“Standardize”对话框,如图 1-27 所示。

**STEP 2)** 在“Standardize”对话框中,选中“Q1”,单击“Standardize”按钮,然后在“Standardize to”分栏下的“Mean”文本输入框中输入均值“0”(系统默认数值即为 0),在“Standard deviation”文本输入框中输入标准差“1”(系统默认数值为 1)。单击“OK”按钮,系统自动生成一个名为“Q1\_stnd”的标准化变量,该变量的均值为 0,标准差为 1。

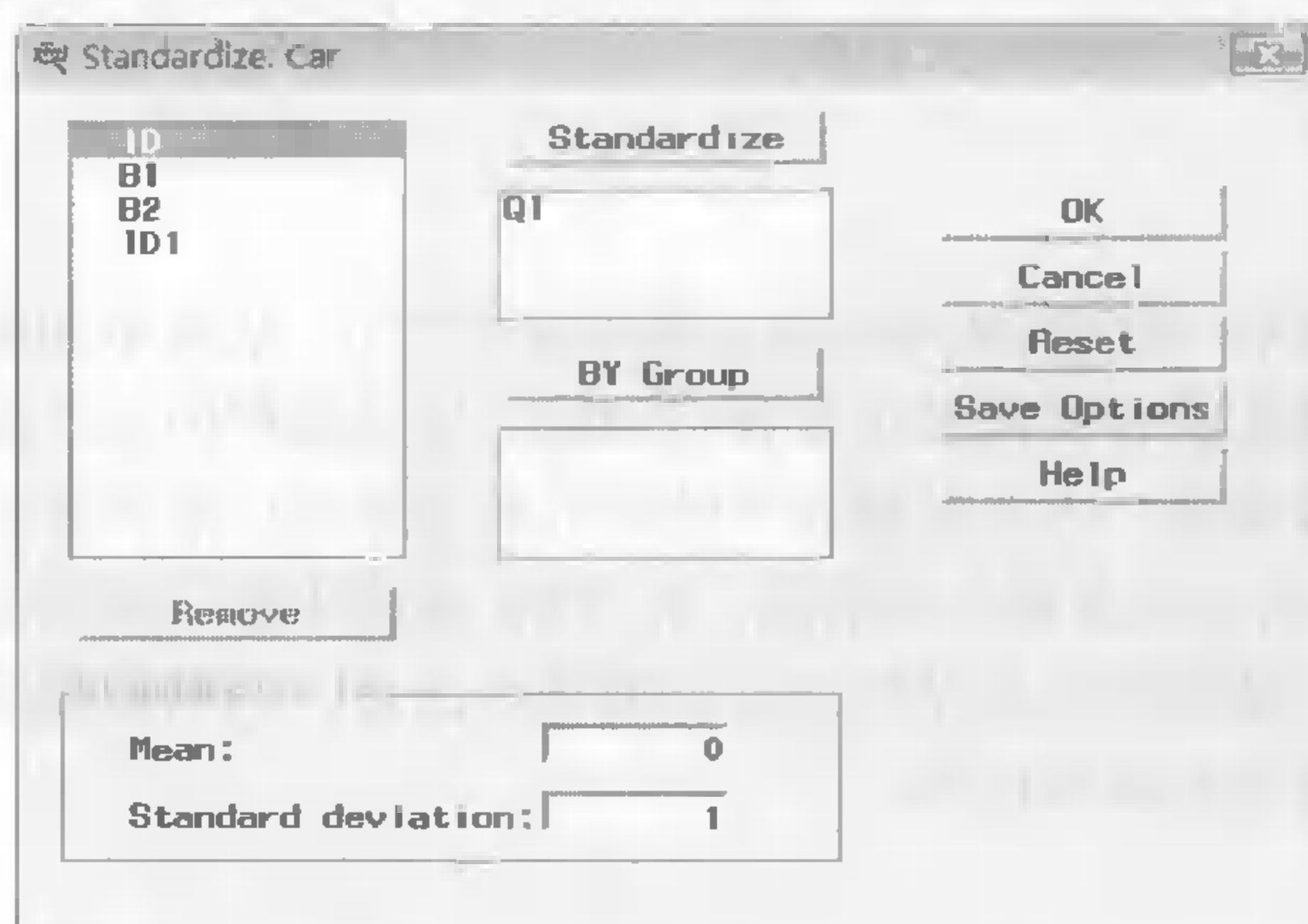


图 1-27 Standardize 对话框

## 3. 数据尺度变换

数据尺度变换主要是指对数据能够反映的取值范围进行调整。如将某个变量的区间范围扩大或缩小,即将原数据区间映射至新的区间范围。这种变换在实际的量表调查问卷中比较常用,如例 1-6 中的满意度“Q1”变量,其原区间范围为[1, 10],而通常所说的满意度指数的取值范围为[0, 100]。这时在计算满意度过程中,需要把“Q1”的尺度区间进行映射,映射后各变量数值的相对位置及数据关系与原数据相同。

设变量  $X$  的原区间为  $[min, max]$ , 新区间为  $[min\_new, max\_new]$ , 则变量  $X$  尺度变换的区间映射可根据以下公式进行。

$$X_{new} = \frac{X - min}{max - min} (max\_new - min\_new) + min\_new$$

如把“Q1”原数据的十分制变换为百分制,即将原区间[1,10]变换为[0,100],则可按以

下公式进行变换。

$$X_{100} = \frac{X_{10} - 1}{9} \times 100$$

尺度变换实际上是函数变换的一种特例，仍然可以通过“Compute”对话框实现，只要把上述变换公式输入至文本输入框中即可。

利用程序中的函数语句，同样可以实现上述数据变化的全部功能。本小节的示例程序如下。

```
data SASUSER.Car;
  set SASUSER.Car;
  Log_10Q1=log(Q1*10);          /*将 10 倍 Q1 求对数得到的数值命名为 Log_10Q1*/
  Interval_Q1=100*(Q1-1)/9;     /*将 Q1 区间转换为[0,100]并命名为 Interval_Q1*/
  Stnd_Q1=Q1;                   /*增加一个变量，用以存储 Q1 的标准化数值*/
run;
proc standard data=SASUSER.Car out=SASUSER.Car /*标准化变换的 standard 过程*/
  mean=0 std=1;                  /*指定标准化过程中的均值为 0、标准差为 1*/
  var Stnd_Q1;                   /*指定要进行标准化的变量*/
run;
proc print data=SASUSER.Car;
run;
```

## 1.5 本章小结

本章主要介绍了 SAS 系统的基本环境及基本操作方式，从菜单和编程两个方面详细讲解了以下内容：SAS 的基本操作界面及主要窗口功能，以及编程的基本语法和构成；SAS 的分析对象是数据库中的数据，永久数据库中的永久数据集可以永久保存；临时数据库中的临时数据集在关闭 SAS 系统时会被自动清除；在 SAS 系统中建立数据集可以通过菜单、编程和导入 3 种方式进行；数据预处理是统计分析的基本前提，包括数据整理、数据库合并及分拆、数据清洗、数据变换等具体内容。

## 第 2 章

# 数据的描述

人们每天都生活在数字的海洋中，如薪水、奖金、股票指数、基金净值、银行利率、汇率、CPI（消费价格指数）、中奖号码等，这些数字使人眼花缭乱；同时，人们也生活在数据的周围，如教育程度、职称、产品等级、政治观点等，这些非数字的数据也会给人们的生活带来巨大的影响。面对这些复杂且交织的数据，没有人能够记住它们的全部信息，但是人们能够通过一定的手段将清数据，把看似错综复杂的数据还原或描述或其本来面貌，并对大量数据进行概括和描述性的分析，使得人们可以快速理解并把数据应用到实际工作中。

本章主要讲述如何利用 SAS 系统通过常见的图形、表格和一些简单的指标来描述各类常见数据，从而把数据的特征及其内在结构直观明了地呈现出来。

### 2.1 统计图

统计图形是最简单的一种数据分析工具，广泛地显现于电视、广告、平面媒体、网络等媒介中。由于数据信息和数据结构可以通过图形方式直观显示，因此人们可以很方便地通过阅读统计图形得到数据结论。

在 SAS 系统中，同样可以通过菜单操作和编程方式实现统计图形的绘制。根据图形自身的属性和作用不同，利用菜单绘图的操作方式分散于各个具体的分析模块当中，而且不同的分析模块可以绘制同样类型的图形；而利用编程方式实现图形绘制也有低分辨率和高分辨率两种主要形式。在 SAS 系统中绘制的各种图形都可以通过 SAS 提供的图形工具进行大小、颜色、刻度等细节的调整。

#### 2.1.1 直方图

直方图是根据变量的取值来显示其频数（次数）分布情况的图形。它的横轴代表数据分组，纵轴可用频数或百分比（频率）表示，这样组别与其相应的频数就形成了一个矩形。在通常情况下，横轴和纵轴也可以互换。

对于等距分组的数据，矩形的高度即可直接代表频数的分布；而对于不等距分组的数据，则需要用矩形面积来表示各组的频数分布特征。



#### 例 2-1

考察 OECD 国家温室气体的排放状况，用以制定温室气体排放国际公约。2002 年对该组织各国进行的调查数据（GAS.sas7bdat）如表 2-1 所示。

从表 2-1 的数据中很难读出温室气体排放状况的基本信息，此时可用直方图进行描述。

表 2-1 OECD 主要国家温室气体排放情况（单位：10 万吨）

国家/地区	二氧化碳	甲烷	氧化亚氮	国家/地区	二氧化碳	甲烷	氧化亚氮
	CO <sub>2</sub>	CH <sub>4</sub>	N <sub>2</sub> O		CO <sub>2</sub>	CH <sub>4</sub>	N <sub>2</sub> O
加拿大	575.86	93.97	52.91	冰岛	2.24	0.53	0.3
日本	1247.61	19.53	35.39	爱尔兰	45.81	12.8	9.74
韩国	465.87	33.85	.	意大利	468.96	34.34	42.2
澳大利亚	358.46	124.29	35.3	卢森堡	10.22	0.47	0.1
新西兰	33.77	27.56	13.16	荷兰	176.65	18.71	15.28
奥地利	69.67	7.46	5.75	挪威	40.95	6.88	5.81
比利时	126.58	9.13	12.89	波兰	308.28	37.79	22.63
捷克	123.05	10.37	8.15	葡萄牙	67.46	8.37	6.1
丹麦	54.16	5.63	7.98	西班牙	325.45	41.14	28.76
芬兰	69.5	5.11	6.82	瑞典	54.75	5.68	8.39
法国	406.04	61.76	72.48	瑞士	43.74	4.26	3.56
德国	864.12	81.45	55.83	土耳其	227.36	18.99	5.44
希腊	105.5	11.44	13.96	英国	537.38	44.07	41.02
匈牙利	57.21	9.78	10.42	斯洛伐克	42.48	4.72	3.83

直方图在 SAS/Analyst、SAS/Insight、SAS/Assist 等模块中均可绘制，本书主要以 SAS/Insight 为主讲解图形的绘制。

**STEP 1** 进入 SAS/Insight，打开例 2-1 的数据集，选择系统菜单 “Analyze→Histogram/Bar Chart”，弹出直方图/条形图对话框，如图 2-1 所示。

**STEP 2** 在该对话框中选择需绘制直方图的变量，如本例中选择 “CO2”（可以同时选中多个变量进行绘图），然后单击 “Y” 按钮。此外，还可以选择 “Group” 按钮绘制按分组变量进行分组的直方图。单击 Output 按钮可以弹出 “Output” 窗口，在此可以设置输出图形中变量的名字或标签、显示坐标轴等。“Method” 按钮可用于设置名义变量异常值的比例。单击 “OK” 按钮后，系统自动弹出 SAS/Insight 的结果窗口，如图 2-2 所示。

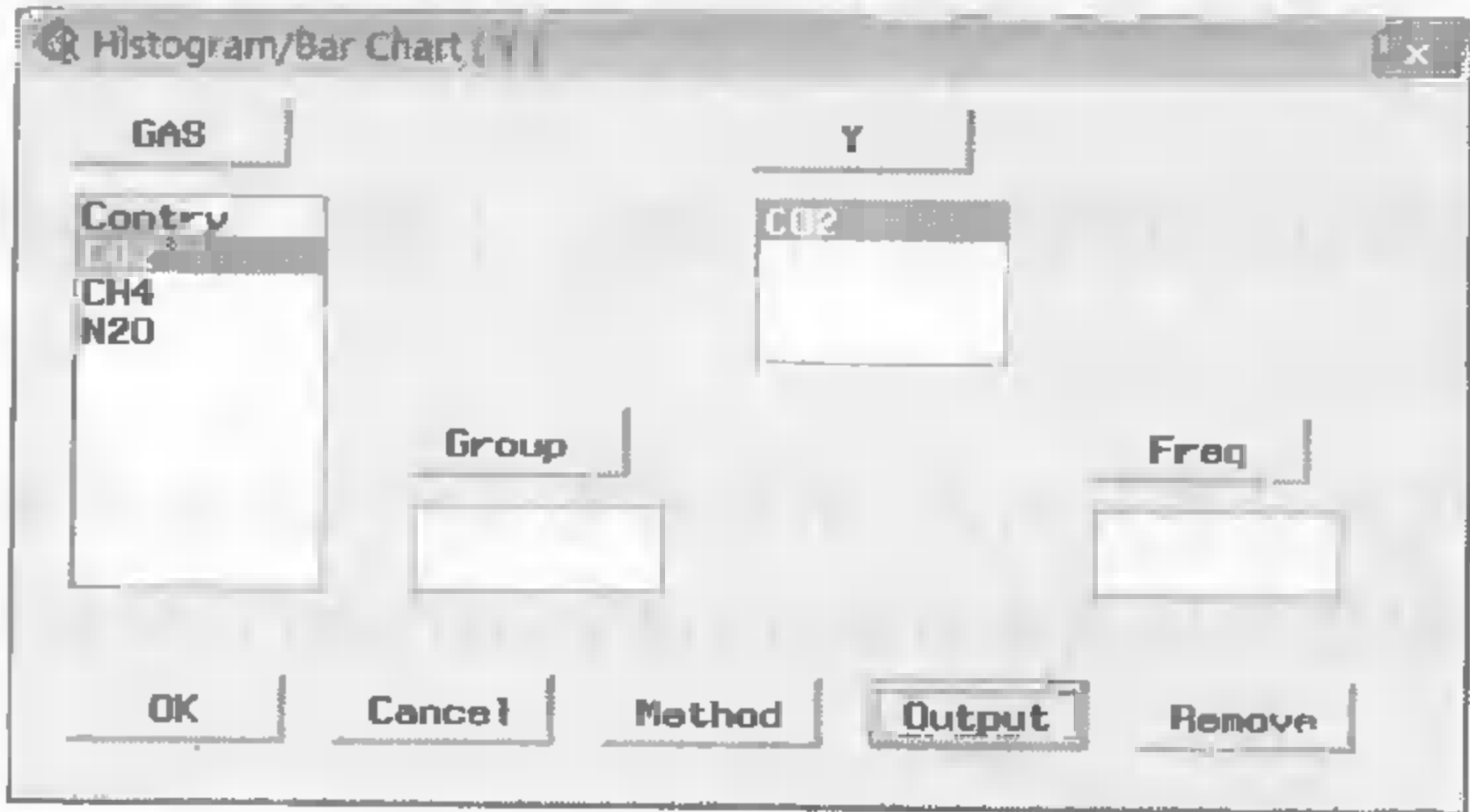


图 2-1 直方图/条形图绘制对话框

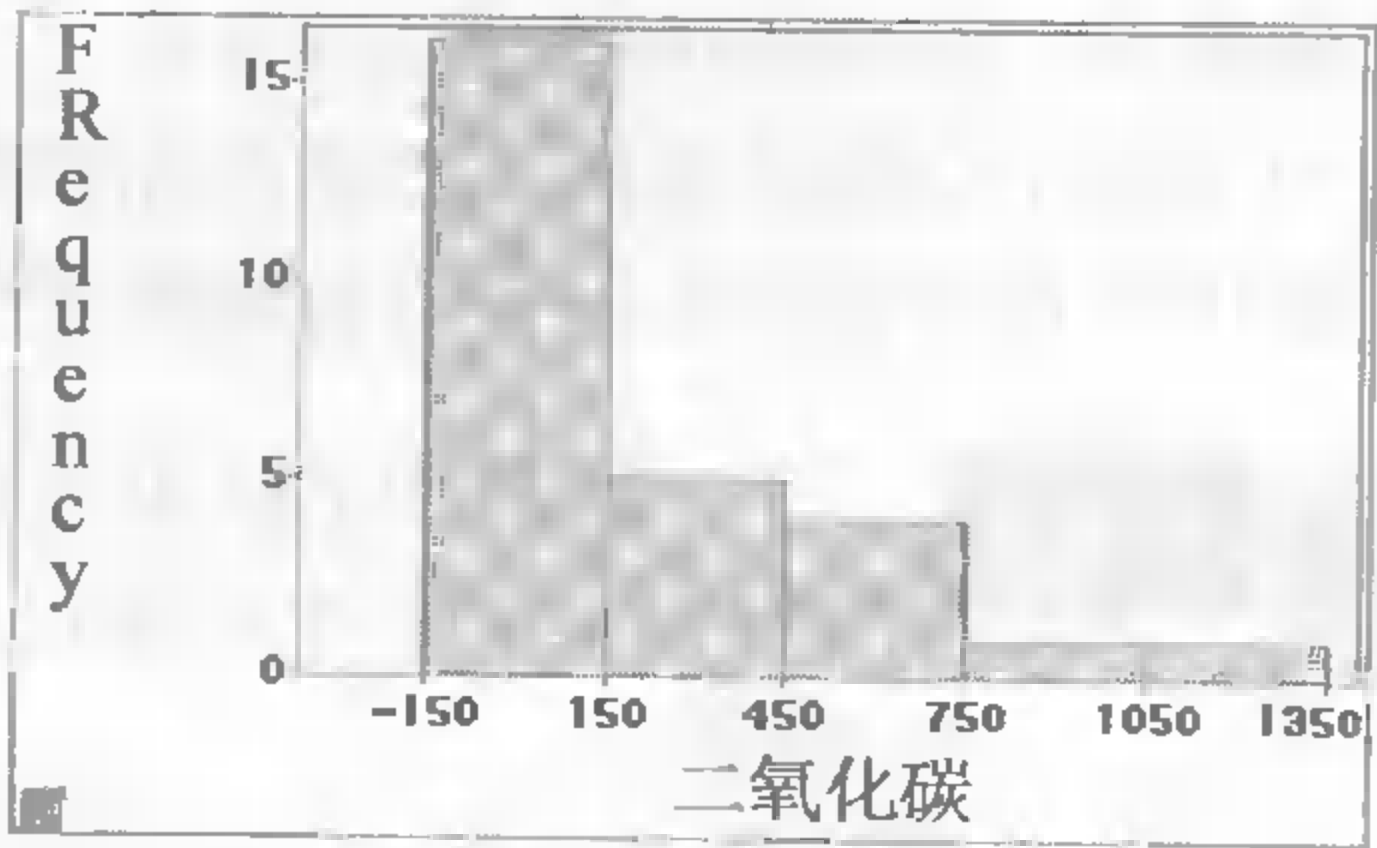



图 2-2 直方图

在图 2-2 中，自动生成的直方图反映了大部分国家或地区的二氧化碳排放量都是比较低的，即处于-150~150 区间范围之内。但是气体排放量指标应当是没有负值的，即最低排放量应当为零，所以应当对图 2-2 的刻度进行适当调整。

**STEP 3** 单击图 2-2 左下方的  按钮，弹出图 2-3 所示的调整选项菜单。其中，各选项功能如下。

- Ticks: 对图形坐标轴刻度的最大值、最小值、刻度间隔等进行调整。
- Axes: 隐藏或显示图形的坐标轴。
- Observations: 隐藏或显示观测值。当观测值被隐藏时，直方矩形在图形中不会出现。
- Values: 隐藏或显示每个直方矩形对应的频数或次数。
- Reference Lines: 在直方图中绘制测量矩形高度的参考线。

**STEP 4** 单击“Ticks”对刻度进行调整，如图 2-4 所示。

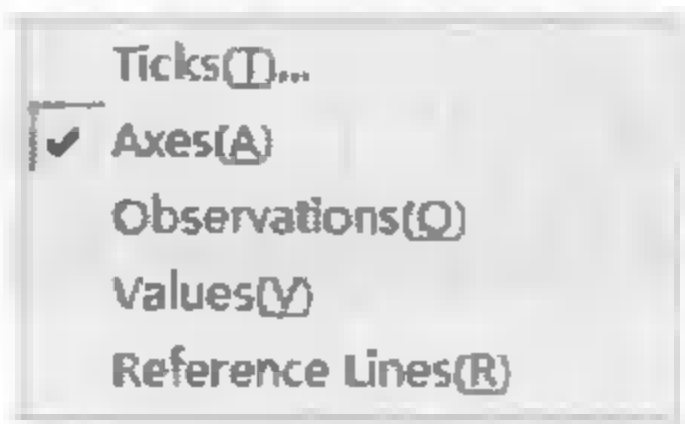


图 2-3 直方图/条形图图形选项菜单

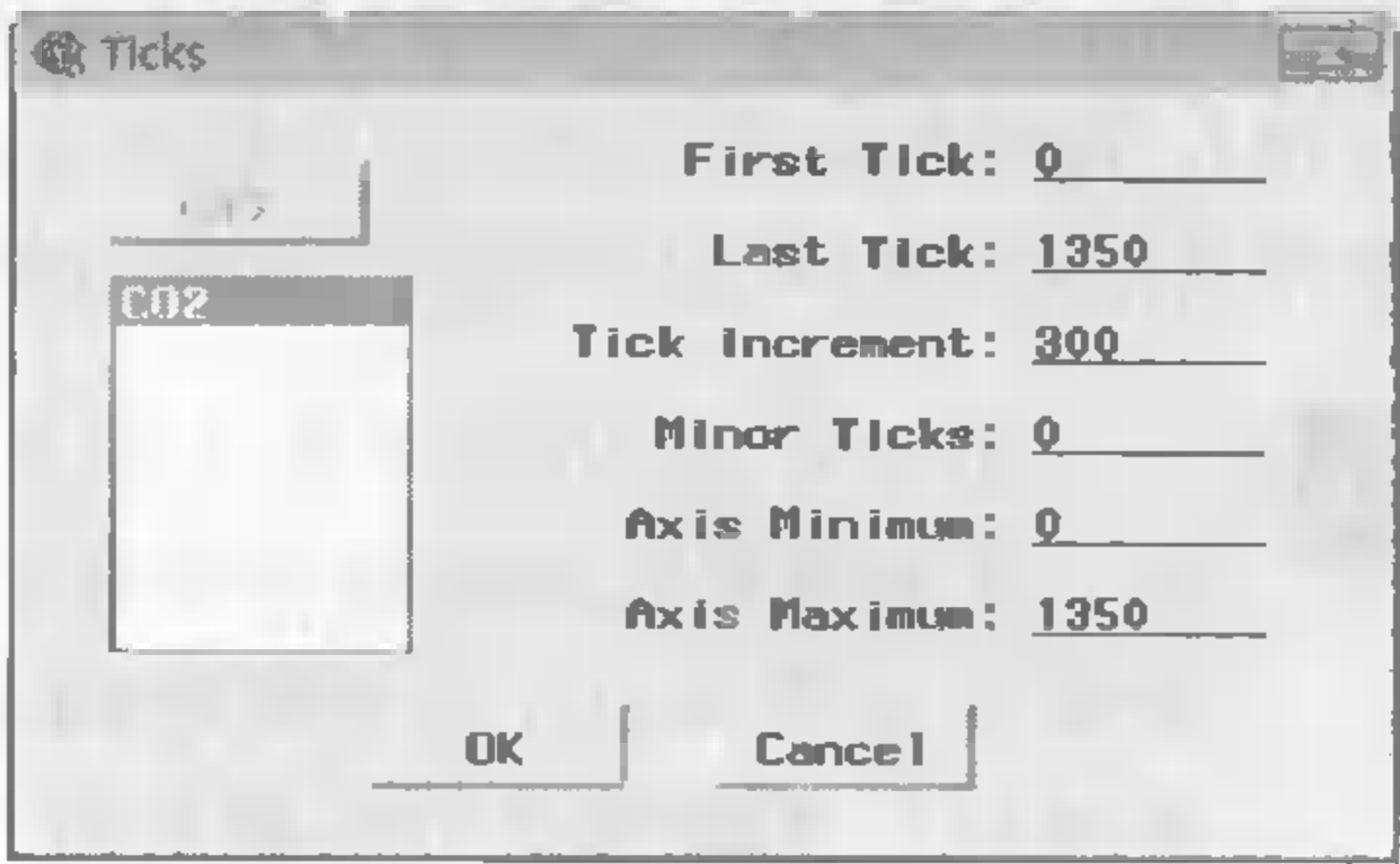


图 2-4 直方图刻度调整对话框

在刻度对话框中，可以设定以下选项。

- First Tick: 初始刻度值。
- Last Tick: 最后一个刻度值。
- Tick Increment: 刻度值的主间隔。
- Minor Ticks: 刻度值的次间隔。
- Axis Minimum: X 轴的最小值。
- Axis Maximum: X 轴的最大值。

**STEP 5** 将“First Tick”中的“-150”修改为“0”，同时将“Axis Minimum”也修改为“0”，单击“OK”按钮，便可得到以下符合实际情况的直方图，如图 2-5 所示。

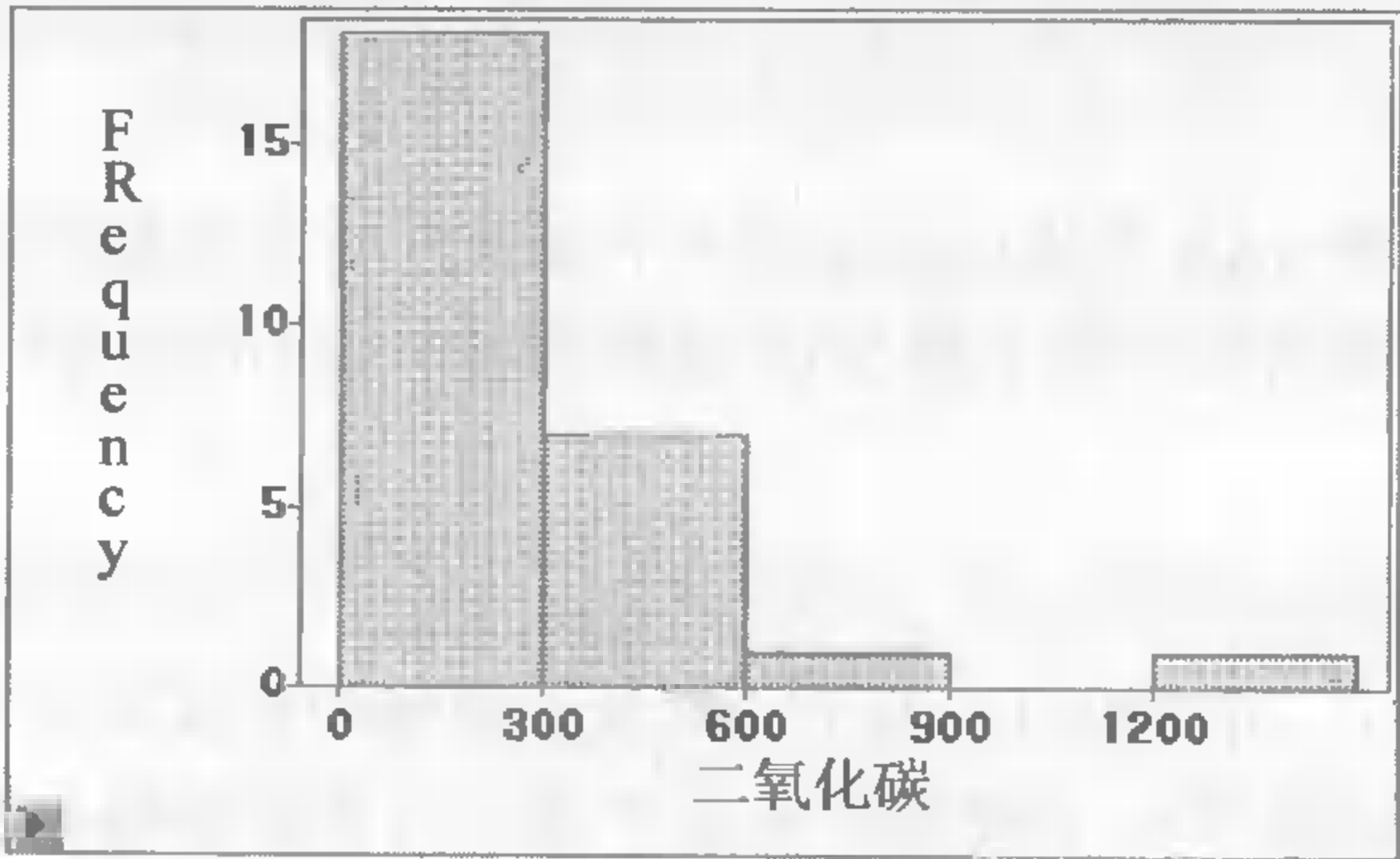


图 2-5 修正的直方图

在直方图的矩形上双击鼠标左键，可以弹出处于该组范围内的具体样本的详细情况。利用 SAS 程序也可以绘制直方图，程序如下。

```
proc capability data=SASUSER.Gas;                                /*调用 capability 模块和 Gas 数据集*/
    histogram CO2 / normal    /*用 CO2 变量绘制直方图，normal 选项表示在直方图中加入正态曲线*/
    midpoin =150 450 750 1050 1350    /*指定直方图每个矩形的中点刻度*/
    ctext=blue;                /*指定直方图的颜色*/
run;
```

在绘制直方图时还应当注意，直方图适用于定量的连续数据，对于定性数据或离散的定量数据不太适用。

2.1.2 条形图

条形图是描述已汇总为频数、相对频数或百分比频数分布的定性数据的图形。通常将图的横轴指定为数据的分组标志，而将纵轴指定为频数、相对频数或百分比频数的刻度（横轴和纵轴也可互换）；每组标志都用相同宽度的条形表示，条形的长度等于观测数值的大小。在绘图时，通常将条形分开以突出每组数据的独立性。



**例 2-2** 为了考察 2008 年第 1 季度北京居民对当前经济总体状况的信心状况，首都经贸大学统计学院对全市 600 户居民进行了入户访问调查，得到了有效样本 595 份（详见 CCI.sas7bdat）。利用条形图对各种不同信心状况的分布情况进行概括分析。

SAS/Insight 中的条形图和直方图是同一个图形，按照绘制直方图的方法可以直接绘制出垂直的条形图。在默认情况下，当使用 SAS/Insight 的绘图功能对定性变量处理时，该功能会自动把 4% 的数据处理为“Other”类型的数据。因此，为了得到完整的条形图，应当在图 2-1 所示的直方图/条形图对话框中单击“Method”按钮，把“Thresholds”（阈值）修改为“0”即可。

利用程序绘制条形图比较灵活，SAS 系统提供了许多可供调整的参数和选项，具体程序如下。

```
proc gchart data=Sasuser.CCI;                                /*利用 gchart 模块绘制高分辨率图形*/
    vbar CCI          /*绘制变量 CCI 的垂直条形图。如绘制水平条形图，则用 hbar 参数*/
    /description="Verical Bar Chart of CCI"    /*命名绘制的图形*/
    type=FREQ;        /*绘制表示频数的条形图*/
run;
```

利用上述程序可以使 SAS 系统自动绘制水平或垂直放置的条形图，但是如果要根据用户自身需要定制条形图，则需要详细了解 SAS 编程语言中的 GCHART 过程。该过程的语法及含义如下。

```
proc gchart <选项>;
```

在 GCHART 过程中，可以用以下两个参数改变图形的形状和位置。

- vbar/hbar 变量表<选项>; /\*根据变量表中指定的变量绘制垂直/水平的条形图\*/
- vbar3d/hbar3d 变量表<选项>; /\*根据变量表中指定的变量绘制立体的条形图\*/

常用的选项语句如下。

- `discrete`: 把数值型变量处理为离散变量。
- `midpoin =`: 设定主轴（即垂直条形图的横轴或水平条形图的纵轴）上的组中值或分组标记。
- `sumvar=`: 设定被求和与求平均数的变量。
- `type=`: 设定条形长度或高度表示的统计量，在不使用 `sumvar` 选项时，为 `FREQ` 和 `PERCENT`，即频数和频率；使用 `sumvar` 时，为 `SUM`（默认）和 `MEAN`，即合计和均值。
- `ref=`: 根据指定数值绘制条形图中的参考线。
- `width=`: 设定条形图的宽度。
- `group=`: 根据指定的变量分组显示图形。
- `subgroup=`: 根据指定的变量绘制堆积条形图。

仍然以例 2-2 为例，假设这次调查的资料中有一个“性别”分类变量，以此了解男性和女性受访市民对本市经济状况的看法是否存在差别。因此，可以把男性和女性的选择情况用条形图的形式呈现出来。为了更好地对比各种看法的性别差异，可以把该问题在性别变量上进行堆积，即把条形图中的每一个条形都分为男性和女性，每一个条型的高度之和等于男性高度加上女性高度。具体程序如下：

```
proc gchart data=Sasuser.CCI;  
  vbar3d CCI  
  /subgroup=gender  
;  
run;
```

运行结果如图 2-6 所示。

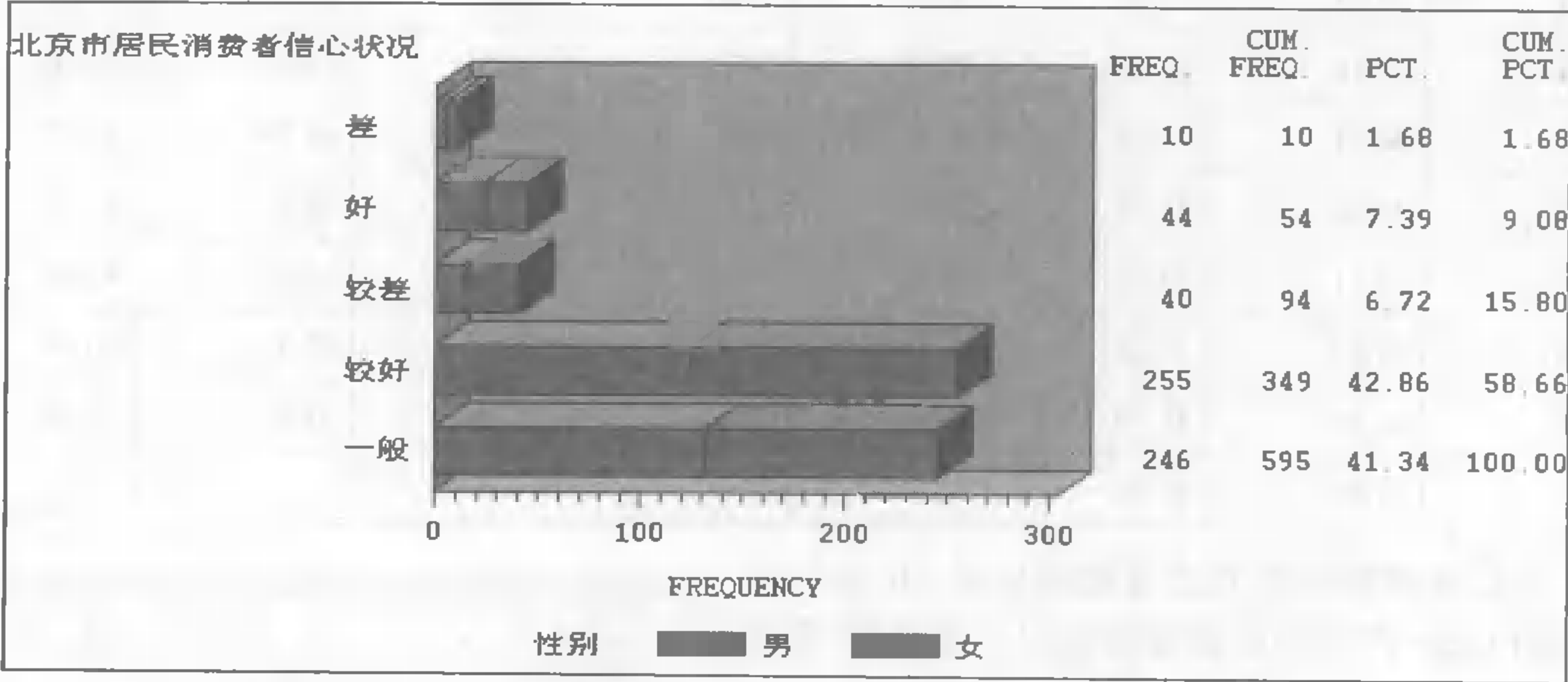


图 2-6 3D 堆积条形图

从图 2-6 中可以明显看出，性别对消费者信心状况大体上没有明显影响，但是女性评价为“较差”的人数明显要比男性多。

这里要注意，如果使用的是“GROUP”选项关键字，则会在坐标轴中出现两个条形图，左边为男性的 5 种选择状况，右边为女性的 5 种选择情况。

2.1.3 线图

线图是由折线或曲线构成的图形。线图在生活中很常见，如股票的 K 线图、价格走势图等。线图一般由两个变量绘制。一个变量作为分析的变量，即线图中线所代表的含义；另一个变量往往是定性变量或时间变量，作为分类变量或参照变量，用以考察分析变量的变动状况。此外，借助线图也可以同时考察多个变量的变动状况，并从中找出数据之间的关系。



例 2-3

为了考察改革开放以来我国的产业结构状况，这里收集了 1978~2006 年三大产业的 GDP 数据（详见 GDP123.sas7bdatt）以进行分析，如表 2-2 所示。

表 2-2 1978~2006 年三大产业 GDP 构成

Year 年份	GDP_1 第一产业 GDP 比重	GDP_2 第二产业 GDP 比重	GDP_3 第三产业 GDP 比重	Year 年份	GDP_1 第一产业 GDP 比重	GDP_2 第二产业 GDP 比重	GDP_3 第三产业 GDP 比重
1978	28.19	47.88	23.94	1993	19.71	46.57	33.72
1979	31.27	47.10	21.63	1994	19.76	46.57	33.57
1980	30.17	48.22	21.60	1995	19.86	47.18	32.86
1981	31.88	46.11	22.01	1996	19.69	47.54	32.77
1982	33.39	44.77	21.85	1997	18.29	47.54	34.17
1983	33.18	44.38	22.44	1998	17.56	46.21	36.23
1984	32.13	43.09	24.78	1999	16.47	45.76	37.67
1985	28.44	42.89	28.67	2000	15.06	45.92	39.02
1986	27.15	43.72	29.14	2001	14.39	45.05	40.46
1987	26.81	43.55	29.64	2002	13.74	44.79	41.47
1988	25.70	43.79	30.51	2003	12.80	45.97	41.23
1989	25.11	42.83	32.06	2004	13.39	46.23	40.38
1990	27.12	41.34	31.55	2005	12.55	47.51	40.04
1991	24.53	41.79	33.69	2006	11.73	48.92	39.36
1992	21.79	43.44	34.76				

可以通过线图的方式直观地呈现 20 多年来我国的 GDP 产业构成和发展状况。在 SAS 系统中，可以通过 SAS/Insight 绘制线图。

STEP 1) 进入 SAS/Insight, 打开对应的数据集 (GDP123.sas7bdatt)。选择系统菜单 “Analyze→Line Plot”，弹出线图对话框，如图 2-7 所示。

STEP 2) 在线图对话框中，“Y” 按钮下的变量即是要进行绘制线图的变量，“X” 按钮下的变

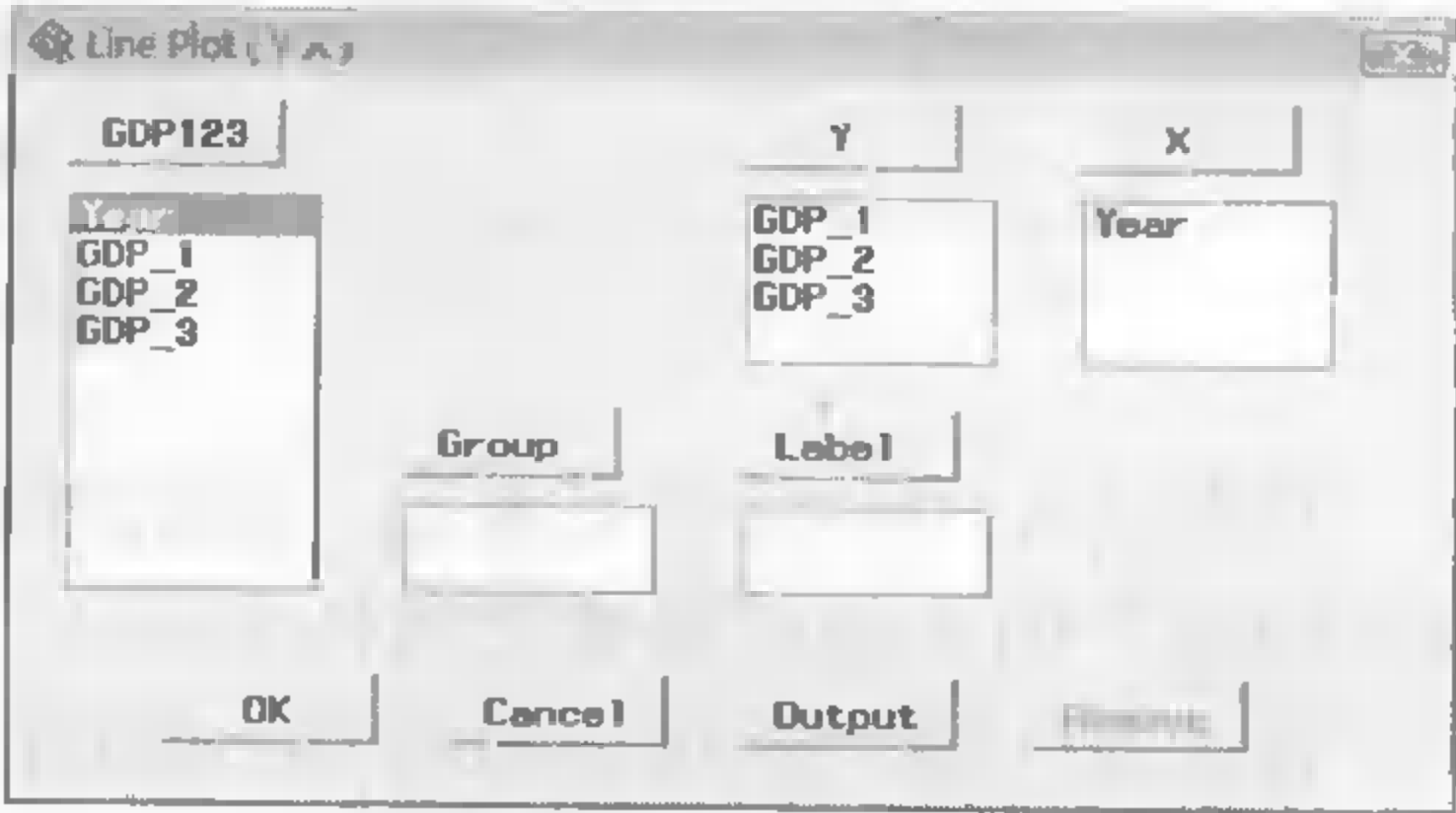


图 2-7 线图对话框

量便是时间变量或定性变量。把年份变量“Year”选中并故置到“X”按钮下的列表框中，把“GDP\_1”、“GDP\_2”、“GDP\_3”同时放在“Y”按钮下的列表框中，然后单击“Output”按钮，选中“Variables”分栏下的“Labels”，单击“OK”按钮，便可在结果窗口中绘制出1978~2006年我国“三大产业”结构的线图，如图2-8所示。

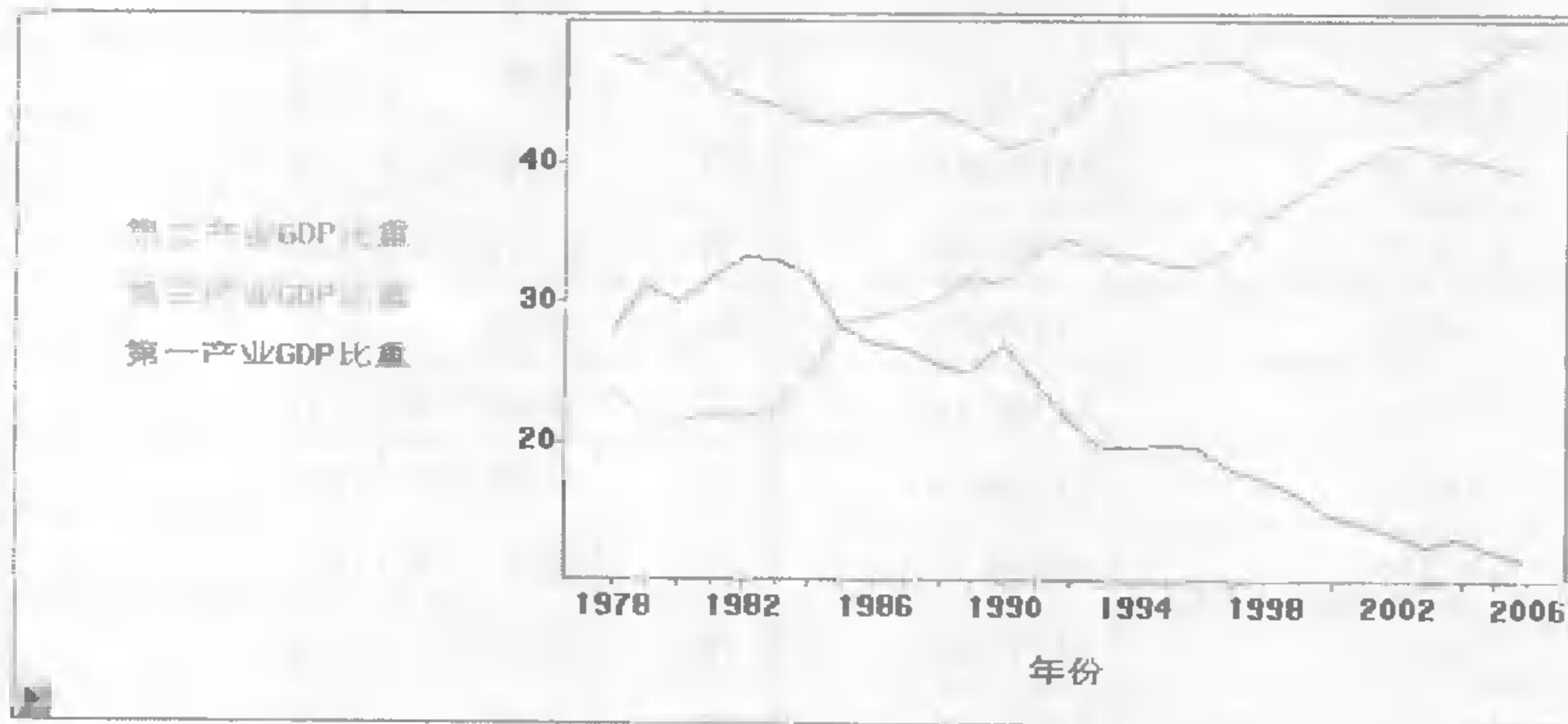


图 2-8 改革开放以来我国三大产业结构变动状况

**STEP 3** 单击左边的变量标签，右边图形对应的变量线图即会以高亮加粗的形式显示。其他的图形调整选项的设置类似于直方图的调整方式。

利用 SAS 程序也可绘制线图，具体程序如下：

```
symbol i=join;                                /*设定两点之间是连续的，否则不能连成线*/
axis1 label=('年份') order=(1978 to 2006 by 1); /*设定一个名为 axis1 的坐标轴属性，其标签为年份，
刻度为 1978 到 2006，每两个刻度之间间隔为 1*/
axis2 label=('各产业 GDP 构成');              /*设定一个名为 axis2 的坐标轴属性*/
axis3 label=('第一产业 GDP 构成');            /*设定一个名为 axis2 的坐标轴属性*/
proc gplot data=Sasuser.GDP123;               /*使用 gplot 语句绘制线图*/
  plot GDP_1 * Year /haxis=axis1 vaxis=axis3;  /*绘制一个只含有第一产业在不同年份变动的线图，横
轴的属性使用上面定义的 axis1 属性，纵轴使用 axis3
属性*/
run;
proc gplot data=Sasuser.GDP123;
  /*把三个产业在不同年份变动的线图进行叠加（使用 overlay 参数），并使用 axis1 和 axis2 分别作为
横、纵轴属性*/
  plot GDP_1 * Year GDP_2 * Year GDP_3 * Year /overlay haxis=axis1 vaxis=axis2;
run;
```

#### 2.1.4 散点图

散点图是由坐标轴上的一系列散点构成的图形，通常用来表示两个变量之间的关系。当坐标轴中的散点多得能够连成线的时候，便成为了线图。因此，线图是散点图的特例。



##### 例 2-4

有关机构对全球 35 个国家的经济发展状况和市场调查研究行业的发展状况进行了研究（数据详见 Survey.sas7bdat），试图从中找出市场调查行业与经济增长之间的相互关系，如表 2-3 所示。

表 2-3 35 个国家的市场调查行业和经济发展状况（单位：美元）

Number 编号	Country 国家	Survey_Percapita 人均调查费用	GDP_Percapita 人均 GDP	Number 编号	Country 国家	Survey_Percapita 人均调查费用	GDP_Percapita 人均 GDP
1	英国	27.65	24122.66	19	巴西	1.2	3201.22
2	美国	21.68	36467.79	20	多米尼加	0.96	2048.19
3	法国	16.21	21759.73	21	秘鲁	0.75	2039.22
4	挪威	15.45	36136.36	22	斯洛伐克	0.74	3333.33
5	丹麦	14.53	30566.04	23	萨尔瓦多	0.65	3064.52
6	澳大利亚	14.37	13315.79	24	泰国	0.5	1860.84
7	比利时	11.18	22941.18	25	保加利亚	0.37	7073.17
8	奥地利	9.63	23536.59	26	菲律宾	0.29	1002.67
9	新加坡	8.21	22564.1	27	斯里兰卡	0.27	752.69
10	西班牙	6.93	13071.07	28	洪都拉斯	0.16	952.38
11	葡萄牙	4.6	9 800	29	中国	0.14	852.54
12	斯洛文尼亚	3	9 000	30	巴基斯坦	0.04	423.79
13	捷克	2.91	4563.11	31	韩国	2.26	9125.8
14	塞浦路斯	2.5	10000	32	瑞士	16.62	33943.66
15	南非	2.34	2923.43	33	加拿大	14.23	22852.46
16	墨西哥	2.19	4979.47	34	智利	2.8	4 200
17	哥斯达黎加	1.94	3888.89	35	俄罗斯	0.29	1100.54
18	巴拿马	1.79	2 500				

STEP 1) 在 SAS 系统中，仍然可以通过 SAS/Insight 绘制散点图。进入 SAS/Insight，打开对应的数据集（Survey.sas7bdat）。选择系统菜单“Analyze→Scatter Plot”，弹出散点图对话框，如图 2-9 所示。

STEP 2) 对于散点图的横、纵坐标没有特别要求，可以把两个变量分别放置在任意一个轴上，其表示的意义不会发生变化。如在本例中，把人均调查费用变量“Survey\_Percapita”放在“Y”轴上，而把人均 GDP 变量“GDP\_Percapita”放在“X”轴上，单击“OK”按钮，便可弹出散点图，如图 2-10 所示。

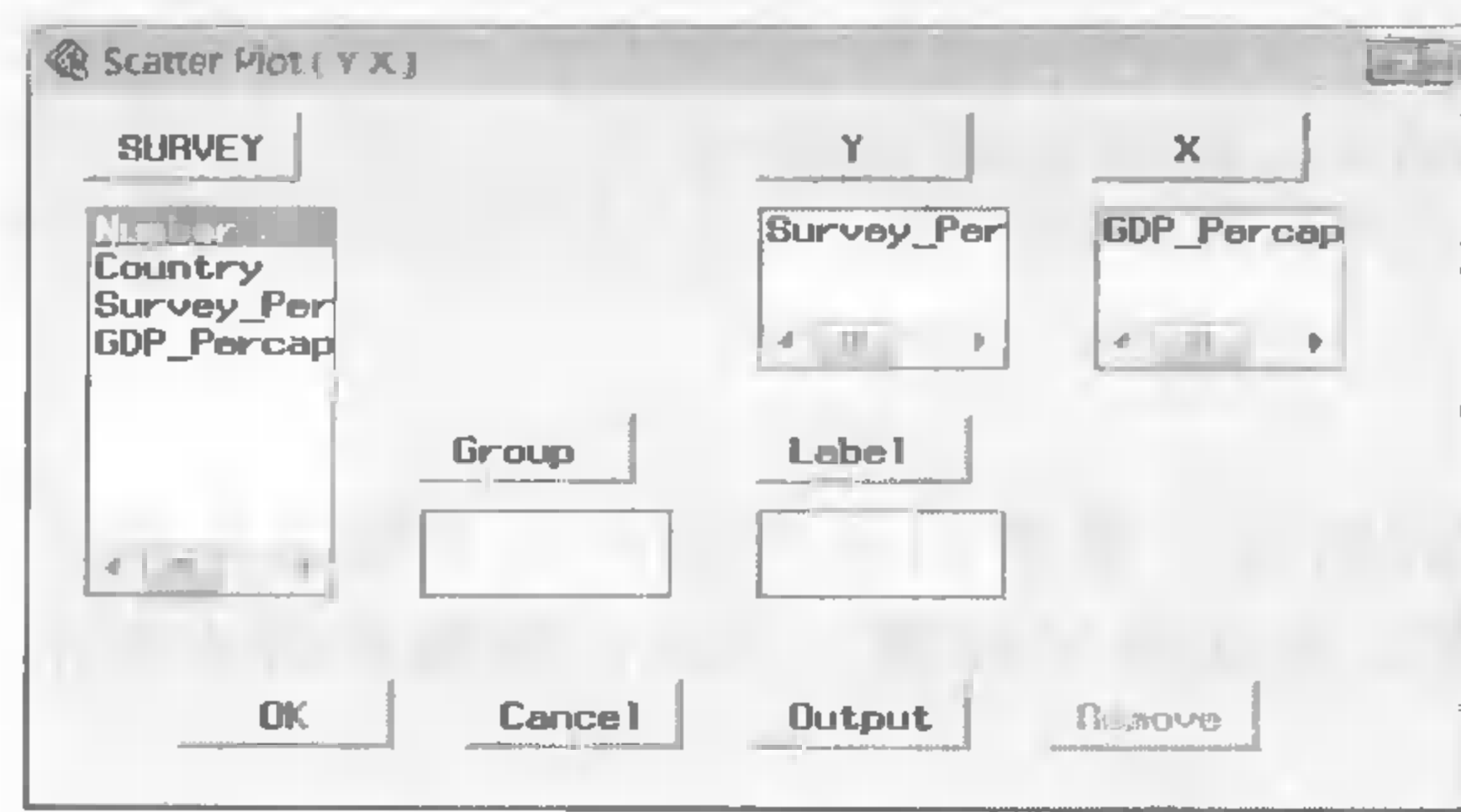


图 2-9 散点图对话框

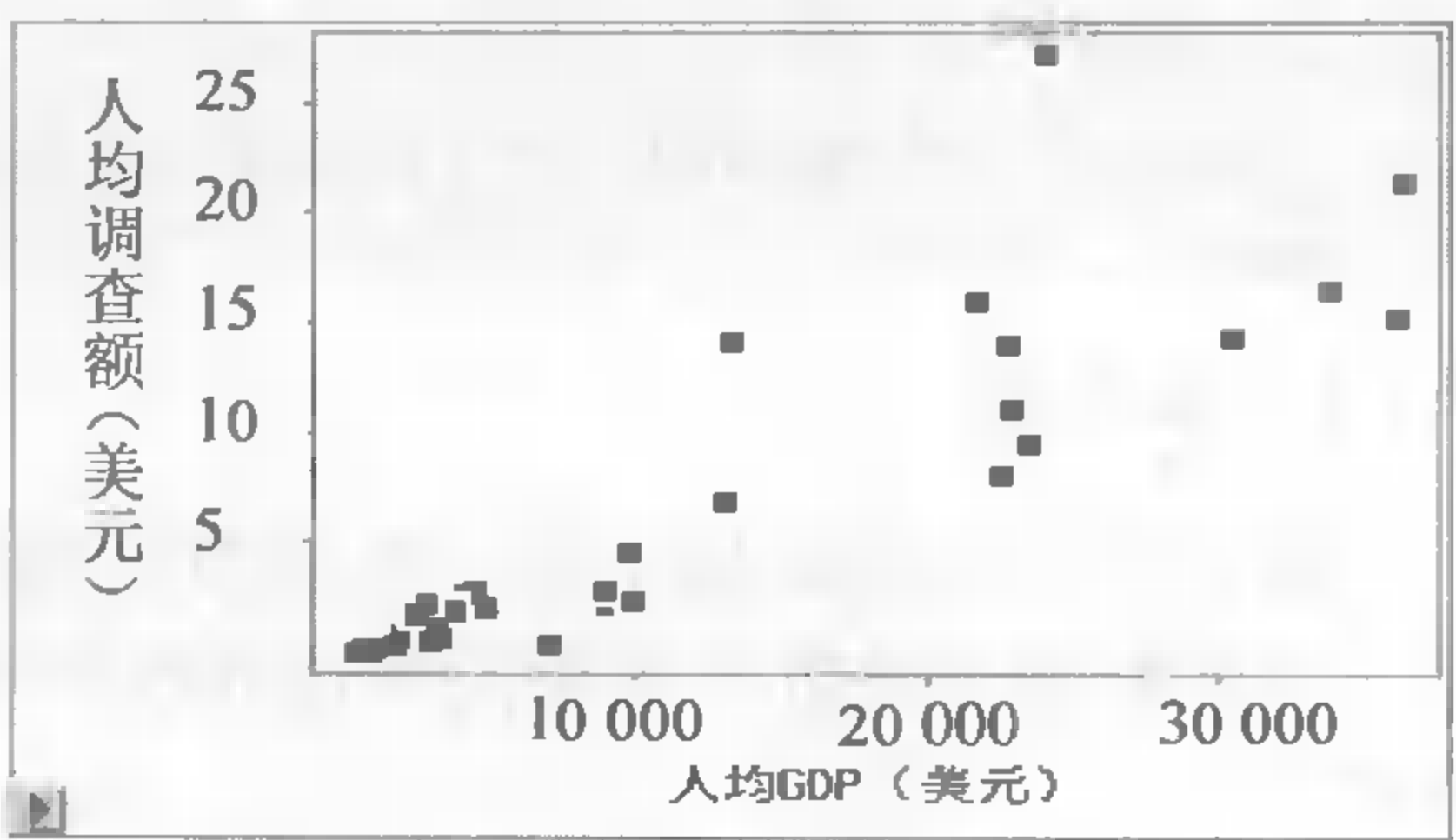


图 2-10 35 个国家人均调查费用和人均 GDP 的散点图

从图 2-10 所描绘的散点图来看，在世界范围内，随着人均 GDP 的增加，人均调查费用/

额度也会增加，二者之间存在正向相关关系。除此之外，大部分国家的人均 GDP 低于 10 000 美元，其人均调查额度也低于 5 美元。

可以单击散点图左下角的三角按钮对图 2-10 进行调整，方法同直方图/条形图，此处不再赘述。同样，利用 SAS 程序也可以很方便地实现散点图的绘制，程序如下。

```
symbol v=square; /*指定散点的符号，如 square 为矩形，dot 为点，triangle 为三角。缺省该行语句系统默认为“+”号*/
proc gplot data=Sasuser.Survey;
  plot survey_percapita * GDP_percapita;
run;
```

2.1.5 饼图

饼图是一种描述定性数据的相对频数和百分比频数分布的图形，通常以圆饼或椭圆饼的形式出现。饼图的整个圆即代表一个总体的全部数据，圆中的一个扇形表示总体的一个类别，其面积大小由相应部分占总体的比例决定，且各部分比例的总和必为 100%。在统计分析中，它主要用来研究结构性问题，如股权结构、投资结构等。

下面用例 2-2 中的北京市消费者信心指数数据 (CCI.sas7bdat) 来考察各种评价的群体所占的比例。

- STEP 1 选择系统菜单 “SAS/Analyst→Graphs→Pie Chart” 弹出饼图对话框，如图 2-11 所示。
- STEP 2 选中 “CCI”，单击 “Chart” 按钮，把该变量设置为分析变量。然后在 “Pie type” 分栏下选择平面饼图 “2-D” 或立体饼图 “3-D” 单选框，选中后单击 “Options” 按钮，可以对饼图进行微调。

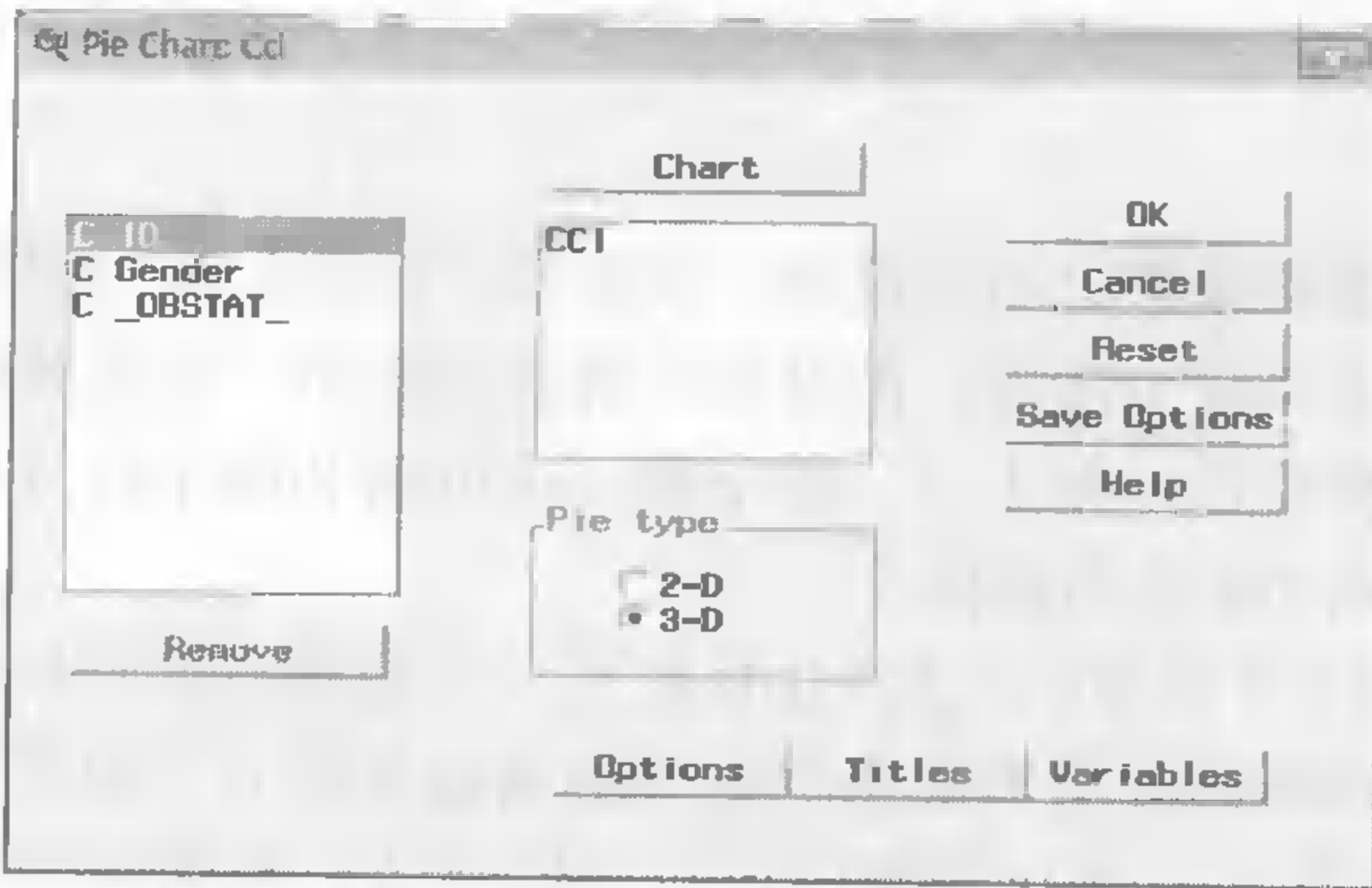


图 2-11 饼图对话框

STEP 3 通常进行微调的是 “Slice Values” 选项卡。在其中的 “Statistic to chart” 分栏下，选择频数 “Frequency” 或百分比 “Percent” 单选框，选中后单击 “OK” 按钮即可返回饼图对话框。在该对话框中，单击 “OK” 按钮，系统会根据用户设置自动绘制饼图。

同样，可以利用 SAS 语言的 GCHART 过程绘制高精度的饼图，程序如下：

```
title '平面 2D 饼图';
proc gchart data=Sasuser.CCI;
  pie cci;
run;
title '立体 3D 饼图';
```

```
proc gchart data=Sasuser.CCI;
  pie3d cci / percent=arrow;      /*为 3D 饼图的扇形加上其表示的百分比*/
run;
title '圆环图';
proc gchart data=Sasuser.CCI;
  donut cci / subgroup=gender;    /*用性别变量将饼图分为两个圆环*/
run;
```

在该段程序中，利用 GCHART 过程生成了一个圆环图，用来对比不同性别的消费者的信心状况，如图 2-12 所示。

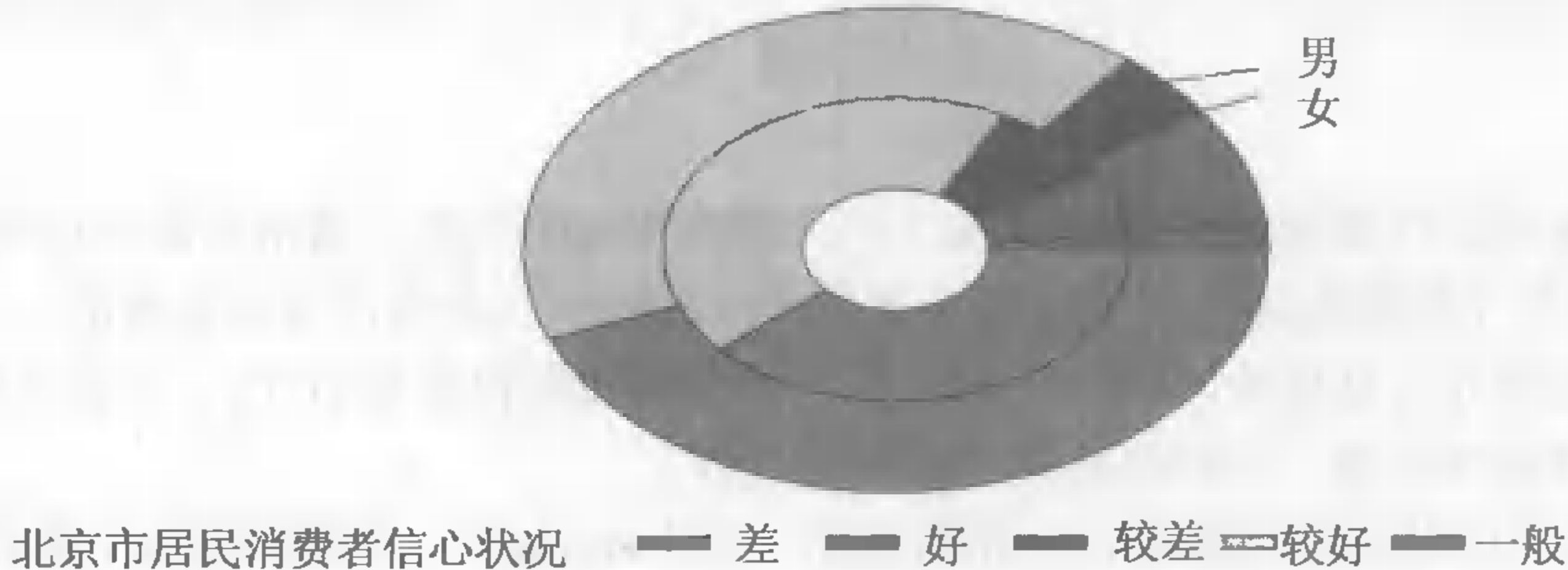


图 2-12 圆环图

在该圆环图中，各部分特征的比例大小并不由其圆环的大小决定，而是由各部分圆环所占的角度，即圆环段占整个自身圆环的比重决定的。由图 2-12 可以看出，评价较差和一般的性别差异比较大，女性评价较差的比例比男性大，而男性评价一般的比例比女性明显要大。

2.1.6 盒式图

盒式图是一个用来描述数据分布状况的、类似盒子的图形，有时也叫盒形图或箱线图。这种图形在一般的媒体上比较少出现，但是却是统计分析中一个重要的描述性分析工具。盒式图可显示数据的 5 个特征值：最大值、最小值、中位数和两个四分位数。为了正确理解盒式图，还是先回顾一下例 2-4 中的数据。

例 2-4 收集了全球 35 个国家的人均 GDP 数据，而这些数据则代表了这些国家的经济发展水平。那么这 35 个国家的经济发展水平是否均衡或有较大差距呢？亦即这些国家中人均 GDP 的分布状况究竟如何呢？是比较集中还是比较分散？这些问题可以通过盒式图感性地进行描述，如图 2-13 所示。

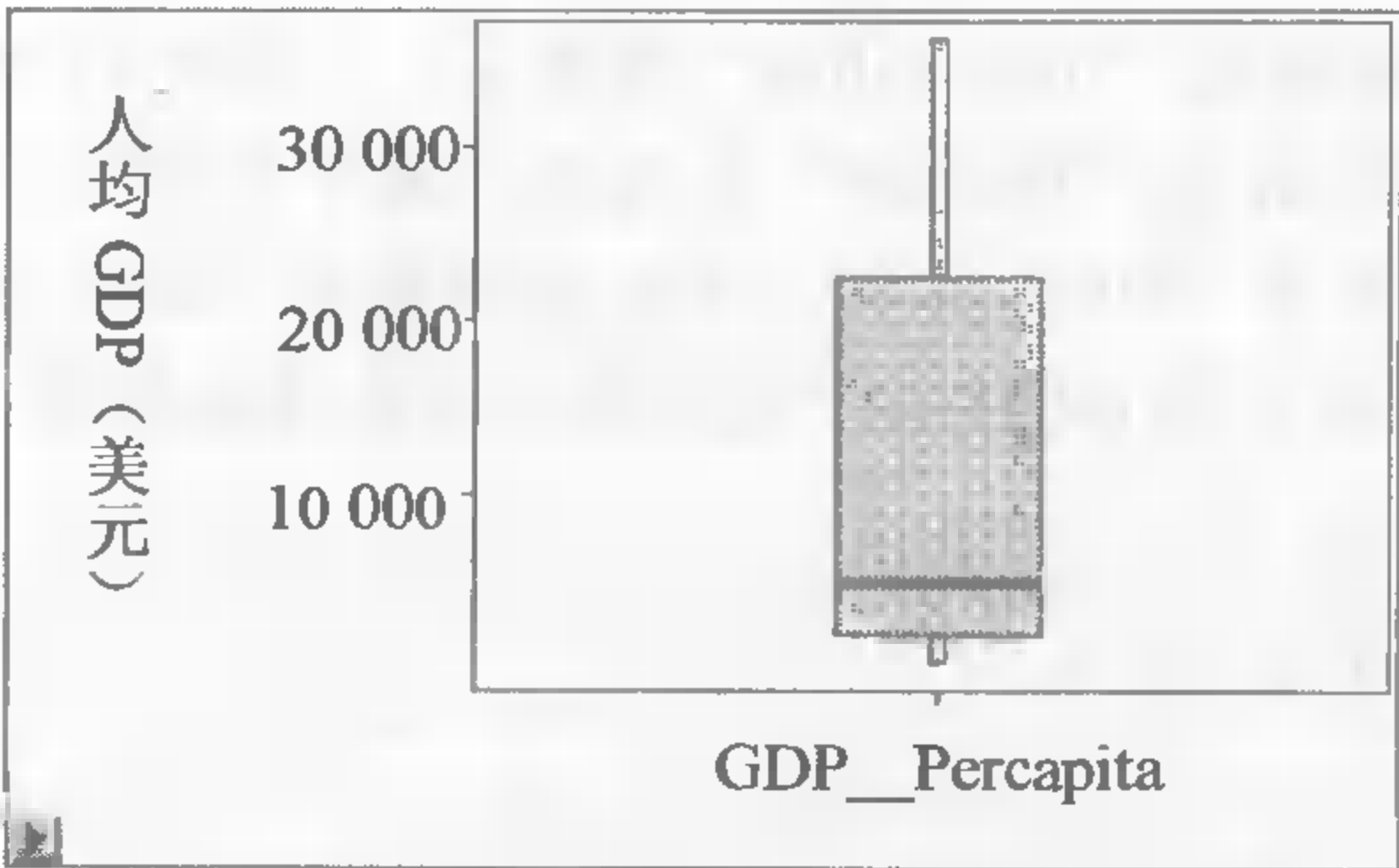


图 2-13 盒式图

下面来看一下盒式图的构成及各部分的含义。

盒式图由一个盒子和两根线构成。盒子的中间线代表数据的中位数，它是数据中占据中间位置的数值，即数据中有一半在该值之上，另一半在其之下；盒子的上下两条边代表上下四分位数，即数据中有四分之一的数值大于上四分位数（在盒子之上），另外有四分之一的数值小于下四分位数（在盒子之下），而盒子的高度则被称为四分位距。因此可知，有一半的数值在中间盒子之内，另一半则分布在盒子的上下两边。盒子上下两端的纵向线段表明箱子外面点的分布。纵向线段的上下两个端点分别表示数据的最大值和最小值。

既然盒子占据了所有数据的一半，也就包含了整个数据的大部分信息，因此主要看盒子的高度。盒子越高，表明大部分数据的变动范围越大，其分散程度即差异也越大；盒子越低，表明大部分数据的变动范围越小，其集中程度也越大。

从图 2-13 来看，盒子比较高，因此可以说这 35 个国家的人均 GDP 还是差异较大的，即它们之间的经济发展水平存在比较大的差异。

选择“SAS/Insight→Analyze→Box Plot/Mosaic Plot”，弹出盒式图对话框。只需把需要绘制盒式图的变量放置在“Y”按钮下的文本框，然后单击“OK”就可以得到图 2-13 所示的盒式图。如果用户选择的变量是离散变量或定性变量，则显示结果是 Mosaic（马赛克）图。

利用 SAS 程序绘制盒式图，在单个变量时比较麻烦，在有分类变量时非常简单。因为盒式图往往也应用于不同总体之间的数据对比，所以 SAS 程序中绘制盒式图的 BOXPLOT 过程要求数据集中有分类变量。具体程序如下：

```
proc sql;
  alter table Sasuser.Survey
  add _dummy_ num label='虚拟分类变量'; /*在原数据集中增加一个名为“_dummy_”的分类变量*/
proc sql;
  update Sasuser.Survey set _dummy_=0; /*为_dummy_变量赋值。可以赋任意值*/
proc boxplot data=Sasuser.Survey; /*调用 BOXPLOT 过程绘制盒式图*/
  plot GDP_Percapita * _dummy_; /*指定分析变量和分类变量*/
run;
```

### 2.1.7 茎叶图

茎叶图由“茎”和“叶”两部分组成，它是由数字形成的。因此，茎叶图不是严格意义上的图形。茎叶图类似于直方图，但与直方图相比，它不仅能够显示数据分布的特征，同时还保留了原始数据的信息。



#### 例 2-5

有关机构每年都对外公布中国大学的排名情况及其综合得分。笔者根据网上最新公布的中国大学排行榜的有关数据（详见 College.sas7bdat）绘制茎叶图进行分析。

在 SAS 系统中，通常使用 SAS 语言中的 UNIVARIATE 过程进行茎叶图的绘制。具体程序如下：

```
proc univariate plot data=Sasuser.College;
  var Score;
run;
```

在 UNIVARIATE 过程的选项中加入 PLOT 关键字, 便可绘制茎叶图、盒式图和正态概率图 3 种图形。运行结果如图 2-14 所示 (节选)。

在图 2-14 中，“Stem”列数是茎，“Leaf”列数是叶（都是个位数字），最右边“#”对应的数字表示频数。图的最下方表明茎的单位是 10，每一个具体的原始数值为：茎单位 $\times$ 茎+叶。如第一行的茎是 10，叶是 0，且只有一个数，则该行的原始数值是： $10 \times 10 + 0 = 100$ 。再看第 7 行，茎是 7，叶分别是 1 和 3（注意不是 13），一共有两个数值，分别是： $10 \times 7 + 1 = 71$  和  $10 \times 7 + 3 = 73$ 。依此类推，可以明显地从图 2-14 中看出，得分为 30~34 分（即倒数第二行）的大学最多，一共有 35 所大学。从整张图形来看，大部分学校的得分

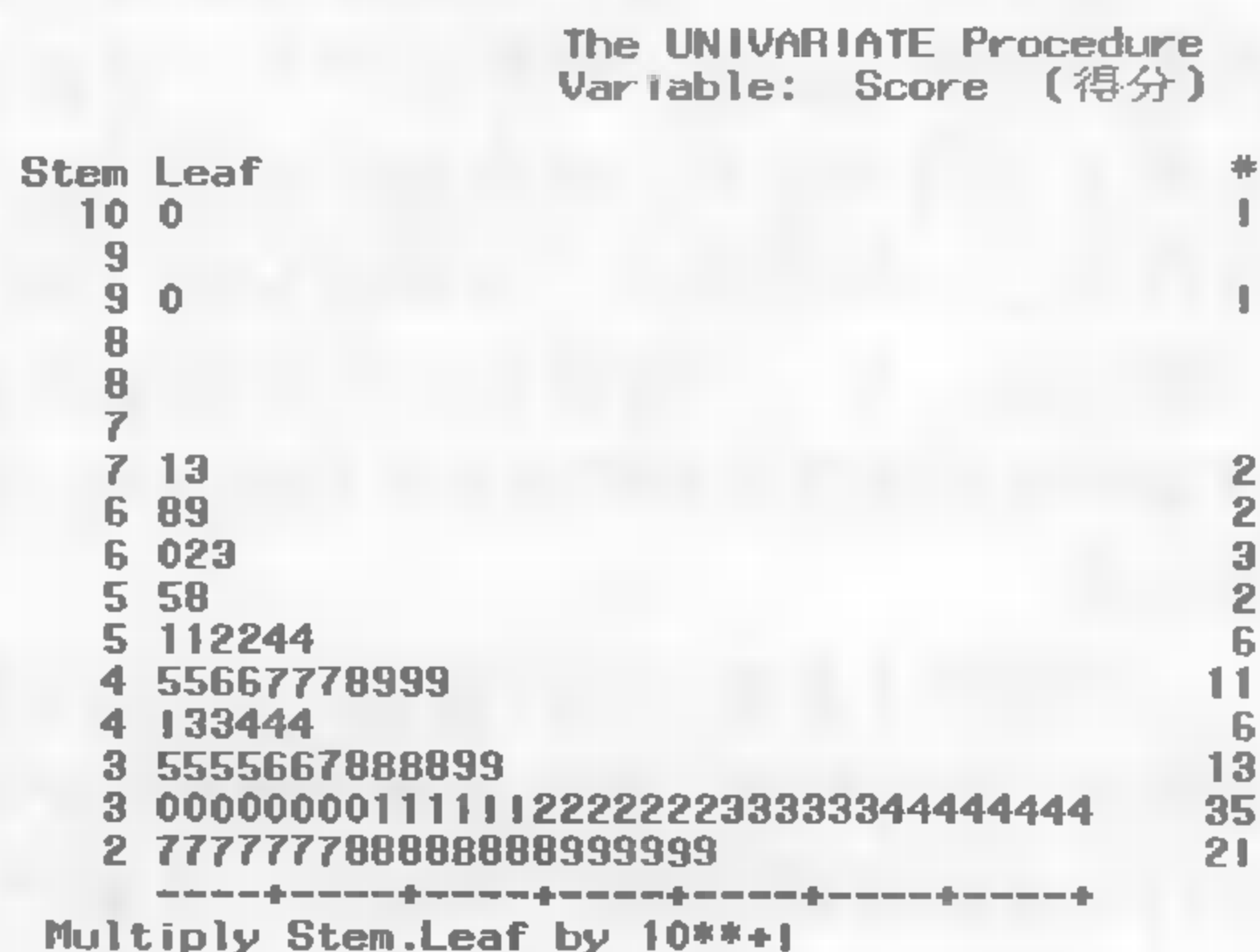


图 2-14 茎叶图

## 2.2 统计量

数据可以用图形来直观呈现，但常常需要增加一些汇总的数据信息进行辅助性的抽象和概括。这些抽象和概括的数据是通过收集而得的原始数据进行归纳总结的，能够用较少的变量来代表全体数据的信息。同时，这些概括性的变量又是能够从收集的样本数据中直接计算出来的，能够在一定的程度上反映总体的特征，因此称之为样本统计量，简称统计量。

统计量是从样本数据中计算出来的，同一总体可以用不同的方式得到不同的样本数据，因此根据不同的样本计算出来的统计量的值就有可能不同。所以，统计量具有不确定性，同时也是不惟一的，但是是已知的。

样本数据的统计量可以从集中趋势、离散程度和分布形状等几个方面进行测量。

### 2.2.1 集中趋势

集中趋势用于描述一组数据的集中位置或平均水平，它代表了一组数据的典型水平，反映了一组数据的中心点位置，具体有以下几种。

## 1. 均值 (Mean)

均值也叫做平均数，SAS 系统中的均值主要指的是简单算术平均数，即原始数据之和除以原始数据的个数： $\bar{x} = \frac{\sum_{i=1}^n x_i}{n}$ 。其中  $x_i$  表示变量的观测值，即原始数据的各个值， $\bar{x}$  表示均值。

均值是最为常见的统计量之一，在日常生活中非常容易接触到，如某人 2007 年的平均工资、我国人口的平均年龄、医院平均每万人可用的床位数、学生考试成绩的平均分等。



例 2-6

10 名学生的统计学期末考试成绩如表 2-4 所示，计算其平均分。

表 2-4 10 名学生的统计学期末考试成绩

学 号	1	2	3	4	5	6	7	8	9	10
成 绩	89	90	78	98	87	76	69	90	92	88

经过计算，这 10 名学生的统计学期末考试平均成绩为  $857/10 = 85.7$ （分）。

均值在统计学中具有极其重要的地位，一般用于寻找数值型数据的中心值，而不适用于分类数据和顺序数据集集中趋势的测度。此外，均值很容易受到极端值的影响。

利用 SAS 编程的 mean 函数可以实现均值的计算，具体程序如下。

```
data;  
x=mean(89,90,78,98,87,76,69,90,92,88);          /*调用 mean 函数*/  
put '均值=' x;                                     /*在“Output”窗口中输出均值计算结果*/  
run;
```

在 SAS 系统中，几乎各个分析模块都能够计算均值，故本节把该部分的软件操作过程放在最后以集中进行讲述。

SAS 系统在 SAS/Insight 中提供以下两种可供选择的均值进行计算。

(1) 截尾均值 (Trimmed Mean)。

计算原始数据中去掉最大 N 个和最小 N 个（或百分之 N）值后的平均值。其中的 N 可以指定为 1、2、3，这是变量中心位置的一种稳健（鲁棒性）估计，但估计量本身不再服从正态分布。

这种均值的计算方法在现实生活也非常实用。去掉头尾若干个最大及最小的数据，有利于克服极端值对数据分析的影响。如在电视歌手大赛中经常会看到，在对歌手进行打分时，主持人唱分时会去掉一个最高分和一个最低分。此处去掉最高分和最低分，实质上相当于求 SAS 系统中当 N = 1 时的截尾均值。

如计算例 2-6 中的截尾均值，取 N = 1，则表示去掉最高成绩 98 分，去掉最低成绩 69 分，还剩下样本容量为 8 的 8 名学生成绩。计算这剩余的 8 名学生成绩平均分为  $690/8 = 86.25$ （分），即为 N = 2 时的截尾均值。

(2) 缩尾均值 (Winsorized Mean)。

把原始数据中最小的 N 个值用第 N+1 小的那个数值替换，同时也把最大的 N 个值用第 N+1 大的那个数值替换，然后计算均值。它也是一种稳健的均值估计。

如计算例 2-6 中的缩尾均值，取 N = 2，则表示把最小的 2 个数用第 3 小的那个数值即 78 进行替换，把最大的 2 个数用第 3 大的那个数值即 90 进行替换，从而得到的新数列为：78、78、78、87、88、89、90、90、90、90，计算该数列的平均数为 85.8（分），即 N = 2 时的缩尾均值。

在 SAS 编程中，可以通过 UNIVARIATE 过程计算上述两种均值，具体程序如下。

```
data null;                                          /*建立一个 Work.null 临时数据集*/  
input score@@;  
cards;  
89 90 78 98 87 76 69 90 92 88
```

```
proc univariate data=null trimmed=2 winsorized=2; /*调用 UNIVARIATE 过程，“trimmed=”指定截尾个数，“winsorized=”指定缩尾个数*/
var score;
run;
```

UNIVARIATE 过程的功能十分强大，本节末尾将详细讲解该过程的具体用法，并给出分析的实际示例。

2. 中位数 (Median)

所谓中位数，就是把所有数据按照一定的顺序（通常按数值大小）进行排序后数据最中间位置对应的那个数值。如果数据个数是奇数，中位数是处在正中心位置的数值；如果数据个数是偶数，中位数则是处在正中心位置的两项数据的平均数。

如在例 2-6 中，学生成绩的中位数对应的是学生成绩排名第 5 名和第 6 名中间的位置，则中位数应该为 $(88+89)/2 = 88.5$ (分)。说明有一半学生的成绩在 88.5 分之上，另一半学生的成绩在 88.5 分之下。

中位数的计算较为简便。只需将全部数据排序，然后用找中点的方法就可得到中位数。而且，它对极端值不敏感，因此也称之为位置平均数。

利用 SAS 程序的 MEDIAN 函数可以计算中位数，具体程序如下。

```
data;
x=median(89,90,78,98,87,76,69,90,92,88); /*调用 median 函数计算中位数*/
put '中位数=' x; /*在“Output”窗口中输出中位数计算结果*/
run;
```

3. 分位数 (Quantile)

除中位数把所有数据等分成两部分外，与其类似的还有四分位数 (Quartile)、十分位数 (Decile) 和百分位数 (Percentile) 等，即分别用 3 个点、9 个点和 99 个点将数据 4 等分、10 等分和 100 等分后各分位点位置上的数值。中位数实际上是第 50 个百分位数、第 2 个四分位数或 5 个十分位数。

在实际应用中，四分位数应用最为广泛。通过 3 个四分位点把原始数据等分为四份，每份的数据量占总数据量的 25%或 1/4。其中，处于 25%位置对应的数值叫做下四分位数，在 SAS 系统中记作“Q1”；处于 75%位置对应的数值称之为上四分位数，在 SAS 系统中记作“Q3”。

在计算分位数时，也需要对原始数据进行排序，分位数只受位置影响，不受极端值影响。

在 SAS 程序中，可以调用 MEANS 过程进行分位数的计算，如计算例 2-6 中的学生成绩的分位数，具体程序如下：

```
data null; /*建立一个名为 Work.null 的临时数据集*/
input score@@;
cards;
89 90 78 98 87 76 69 90 92 88
;
proc means data=null q3 q1; /*调用 MEANS 过程，其中 q3 表示上四分位数，q1 表示下四分位数*/
```

```
var score;  
run;
```

MEANS 过程的功能十分强大，本节末尾将详细讲解该过程的具体用法，并给出分析的实际示例。

4. 众数 (Mode)

众数是指数据中出现次数最多的数值。它既可应用于定量数据，也可应用于定性数据，是一种比较重要的集中趋势测度指标。



例 2-7

某国家机关通过公务员考试招收工作人员，具体参考人员得分情况如表 2-5（数据详见 Gwy.sas7bdat），试指出这些参考人员考试得分的众数。

表 2-5 16 名参考人员的公务员考试笔试成绩

应聘人员	人员 1	人员 2	人员 3	人员 4	人员 5	人员 6	人员 7	人员 8
笔试成绩	190	188	188	185	183	183	180	180
应聘人员	人员 9	人员 10	人员 11	人员 12	人员 13	人员 14	人员 15	人员 16
笔试成绩	180	180	178	177	175	175	174	173

在这 16 名参考人员的笔试成绩中，173、174、177、178、185、190 各出现 1 次，175、183、188 各出现 2 次，180 出现 4 次。因此，这些参考人员笔试成绩的众数为 180。

有时众数不一定像例 2-7 这样只有一个，也可能出现多个众数，即出现次数最多的数值不止一个。如某单位中有 50 个同事，其中叫“张三”的人有 2 个，叫“王五”的人也有 2 个，除此之外没有同名同姓的人，则“张三”和“王五”都是姓名这个变量的众数。当所有数值都只出现 1 次时，这种情况被称为无众数。众数不受位置的影响，也不受极端值的影响。

在 SAS 编程中，可以调用 UNIVARIATE 过程实现众数计算，具体程序如下。

```
data null;  
  input score@@;  
  cards;  
    190  188  188  185  183  183  180  180  
    180  180  178  177  175  175  174  173  
  ;  
proc univariate data=null modes;      /*调用 Univariate 过程的 Modes 选项计算众数*/  
  var score;  
run;
```

2.2.2 离散程度

集中趋势概括数据可以使人们在大体上对数据产生初步的印象，但是在这些指标对数据进行高度抽象的同时，也忽略了一些必要的信息，使得人们在某些情况下只能看到数据呈现出来的假象，而不能读懂其真正的内在涵义。如一个人脚下泡着开水，头上顶着冰块，就冷暖的集中趋势均值而言，冷暖相互抵消并有向平均温度靠近的趋势，他理应感到很舒服。

这是相当滑稽的事情。造成这种荒唐事情的一个主要原因是：没有考虑集中趋势以外的其他数据信息，犯了以偏概全的错误。因此，不光数据集中程度是考虑问题时应注意的一个方面，数据的离散程度也是应当考虑的重要问题。

离散程度的测度指标很多，常见的有级差、平均差、四分位差、异众比率、方差、标准差、标准误差及离散系数等。本小节只介绍 SAS 系统中的常见离散程度测度指标。

### 1. 级差 (Range)

级差也叫全距，是数据中最大值减去最小值所得的差。如计算例 2-6 数据的级差为  $98 - 69 = 29$  (分)。

级差极易受极端值的影响。由于它抛弃了几乎全部的数据信息，中间部分的数据信息无法得到反映，因此不能准确地描述数据的分散状况。

利用 SAS 编程语言的 RANGE 函数可以直接计算级差，具体程序如下。

```
data;
x=range(89,90,78,98,87,76,69,90,92,88);    /*调用 RANGE 级差函数*/
put '级差=' x;                               /*在“Output”窗口中显示级差计算结果*/
run;
```

### 2. 四分位差 (Qrange)

四分位差，故名思义是四分位数之间的差。具体而言，它是指第 3 个四分位数减去第 1 个四分位数得到的差，即上四分位数与下四分位数之间的差值。它反映了中间 1/2 数据的分散情况，其值越小，说明中间的数据越集中。四分位差不受极端值的影响，它在一定程度上也说明了中位数对一组数据的代表程度。

利用 SAS 编程语言的 MEANS 过程可以计算四分位差，如计算例 2-7 中的公务员考试得分的四分位差，具体程序如下。

```
data null;
input score@@;
cards;
190 188 188 185 183 183 180 180
180 180 178 177 175 175 174 173
;
proc means data=null qrange;    /*调用 MEANS 过程的 QRANG 选项，计算四分位差*/
var score;
run;
```

对于 MEANS 过程，本节最后部分将详细介绍。

### 3. 方差 (Variance) 和标准差 (Standard Deviation)

方差是每一个数的原始数值与均值的差（即离差）求平方，然后求这些平方的和，再用此平方和除以数据的个数得到的数值。一言以蔽之，方差即离差平方和的平均数，即：

$$\sigma^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n}。$$

方差是用全体数据进行计算的，因此它反映了所有数据相对于数据中心发散的平均程度，是统计分析中最为重要的统计量之一。方差越大，说明数据离散程度越高。

方差的算术平方根，即标准差，其含义与方差相差不大。

如计算例 2-6 数据中的方差和标准差，计算过程如表 2-6 所示。

表 2-6 10 名学生的统计学期末考试成绩方差和标准差的计算过程

学 号	1	2	3	4	5	6	7	8	9	10
成 绩	89	90	78	98	87	76	69	90	92	88
平 均 成 绩	85.7									
成绩-均值	3.3	4.3	-7.7	12.3	1.3	-9.7	-16.7	4.3	6.3	2.3
(成绩-均值) <sup>2</sup>	10.9	18.5	59.3	151.3	1.7	94.1	278.9	18.5	39.7	5.3
方差 = $\sum (\text{成绩} - \text{均值})^2 / n$	67.81									
标 准 差	8.23									

在 SAS 系统中，系统计算的是样本方差，即样本量在原始数据样本量基础上减 1（即少了一个样本，样本量即公式中的分母，为  $n-1$ ），也称之为样本修正方差。

在 SAS 编程语言中，可以直接利用 VAR 函数和 STD 函数直接计算方差和标准差，具体程序如下：

```
data;
x=var(89,90,78,98,87,76,69,90,92,88);      /*调用 VAR 函数计算样本方差*/
y=std(89,90,78,98,87,76,69,90,92,88);      /*调用 STD 函数计算样本标准差*/
put '方差=' x '标准差=' y;
run;
```

4. 标准误差（Standard Error）

标准误差是样本均值的标准差。在进行数据的抽样调查中，由于随机性的存在，采用同一种抽样方法在不同的时间、地点、环境等条件下进行多次抽样，可能得到多个不同的样本数据；同一种抽样方法在相同的环境下由不同的人员实施，也可能得到多个不同的样本数据。把所有类似情况下获得的所有可能样本找出来，如在总体容量为  $N$  的总体中，随机抽取样本容量为  $n$  的样本。如考虑排列顺序，一共可能获得  $N^n$  个样本；如不考虑排列情况，则可获得  $C_N^n$  个样本。计算每个样本的均值，可以得到由若干个均值组成的数据。然后计算这些数据的标准差，即得标准误差。

标准误差不是观测值的实际误差，也不是误差范围，它只是对一组观测数据可靠性的估计。标准误差小，则观测的可靠性大，反之则不大可靠。世界上多数国家的物理实验和正式的科学实验报告都是用标准误差评价数据的测量精度的。

在 SAS 编程语言中，可以直接利用 STDERR 函数计算标准误差，具体程序如下。

```
data;
x=stderr(89,90,78,98,87,76,69,90,92,88);    /*调用 STDERR 函数计算标准误差*/
put '标准误差=' x;
run;
```

5. 变异系数 (Coefficient of Variation)

变异系数是衡量相对离散程度的一个重要指标。之前介绍的测度指标都是从一个单独数值来反映数据离散程度的，对于平均水平不同或计量单位不同的不同变量而言，则不能采用这几种方法来直接比较其离散程度。为了消除这些因素对离散程度的测度影响，就需要用一个相对指标来对不同总体或样本数据的离散程度进行比较。

变异系数也叫做标准差系数或离散系数，具体是指一组数据的标准差与其相应的均值之比，即： $CV = \frac{\sigma}{\bar{x}}$ 。变异系数越小，说明数据的离散程度也越小。



例 2-8

某学校新生入学体检，经过抽样，考察 20 名新同学的身高和体重差异状况。详细数据如表 2-7 所示，试分析身高和体重的差异。

表 2-7 20 名新生的身高和体重

序号	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
身高	157	170	161	184	184	168	166	158	174	166	173	189	188	163	161	189	183	186	188	155
体重	61	71	74	59	71	55	73	66	50	57	48	68	73	52	60	56	53	67	73	64

身高和体重的差异如何进行对比？直接采用标准差是无法进行对比的，因为标准差是带有单位的，身高的单位是厘米，体重的单位是千克，不同单位的统计量对比起来是没有任何意义的。为此，采用变异系数来消除量纲即计量单位的影响，并进行对比。

在 SAS 语言中，可以直接利用 CV 函数计算数据的变异系数，具体程序如下：

```
data;
  height=cv(157,170,161,184,184,168,166,158,174,166,173,189,188,163,161,189,183,186,188,155);
  weight=cv(61,71,74,59,71,55,73,66,50,57,48,68,73,52,60,56,53,67,73,64);
  put "身高的变异系数=" height "体重的变异系数=" weight;
  if height>weight then
    put "身高比体重的差异大";
  else
    if height<weight then
      put "身高比体重的差异小";
    else
      put "身高和体重的差异相当";
run;
```

在 SAS 系统中，CV 函数计算出来的变异系数的值是数据标准差与均值之比的 100 倍。

2.2.3 分布形状

在对数据进行概括性的分析时，考察集中趋势和离散程度是两个重要方面，但并非仅此而已。就像对一个人进行评价一样，不仅要考察人的高矮情况，也要考察胖瘦情况，更要看看一个人是否站有站相、坐有坐相。对于数据的分布状况，也应当进行概括性的分析，才能掌握数据的全貌。

数据分布的测度主要考察数据分布的偏斜程度、扁平程度，以及数据分布是否对称，其指标主要有偏度和峰度两类。

### 1. 偏度 (Skewness)

偏度是对数据分布对称性的测度。偏度的计算方法有很多，通常采用三阶中心矩的计算方法，其主要考察离差三次方之和与标准差的三次方的比例，即： $SK = \frac{n \sum (x_i - \bar{x})^3}{(n-1)(n-2)s^3}$ 。其中  $s$  表示样本标准差。

如果数据是对称的，则偏度等于 0；如果偏度明显不等于 0，则表明数据分布是非对称的，具体地说，偏度大于 0 时，均值右边的数据更为分散，表明数据右偏；偏度小于 0 时，均值左边的数据更为分散，表明数据左偏。偏度的数值越大，表明数据偏斜的程度就越大。具体描述如图 2-15 所示。

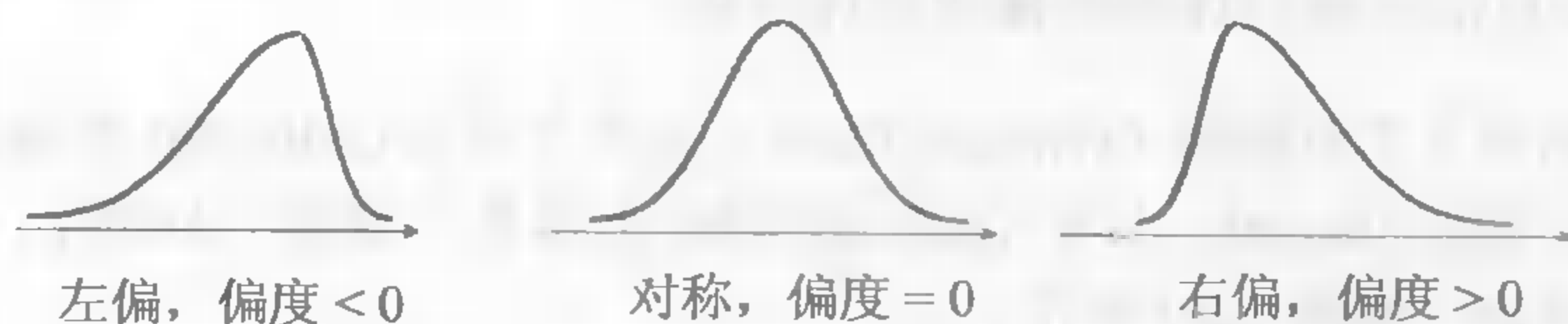


图 2-15 偏度与数据分布

在 SAS 编程语言中，可以直接利用 SKEWNESS 函数计算偏度。以例 2-7 中的公务员考试成绩为例，具体程序如下。

```
data;
x=skewness(190,188,188,185,183,183,180,180,180,180,178,177,175,175,174,173);
put '偏度=' x;
run;
```

### 2. 峰度 (Kurtosis)

峰度是用来反映数据分布曲线顶端陡峭或扁平程度的指标。这里所说的陡峭或扁平是针对标准正态分布而言的。峰度通常用四阶中心矩进行计算，考察四阶矩与标准差四次方之间的比例关系，即： $K = \frac{n(n+1) \sum (x_i - \bar{x})^4 - 3 \left[ \sum (x_i - \bar{x})^2 \right]^2 (n-1)}{(n-1)(n-2)(n-3)s^4}$ 。

SAS 系统中的计算公式是四阶中心矩与标准差四次方的比值减去 3 后的值(即 SAS 计算出来的峰度 =  $K - 3$ )。

如果数据服从标准正态分布，则峰度的值等于 0。如果峰度明显不等于 0，则表示数据分布比标准正态分布更陡峭或更扁平。具体地说，如峰度大于 0，说明它是比正态分布要陡峭；如峰度小于 0，则说明数据分布比正态分布平坦。具体描述如图 2-16 所示。

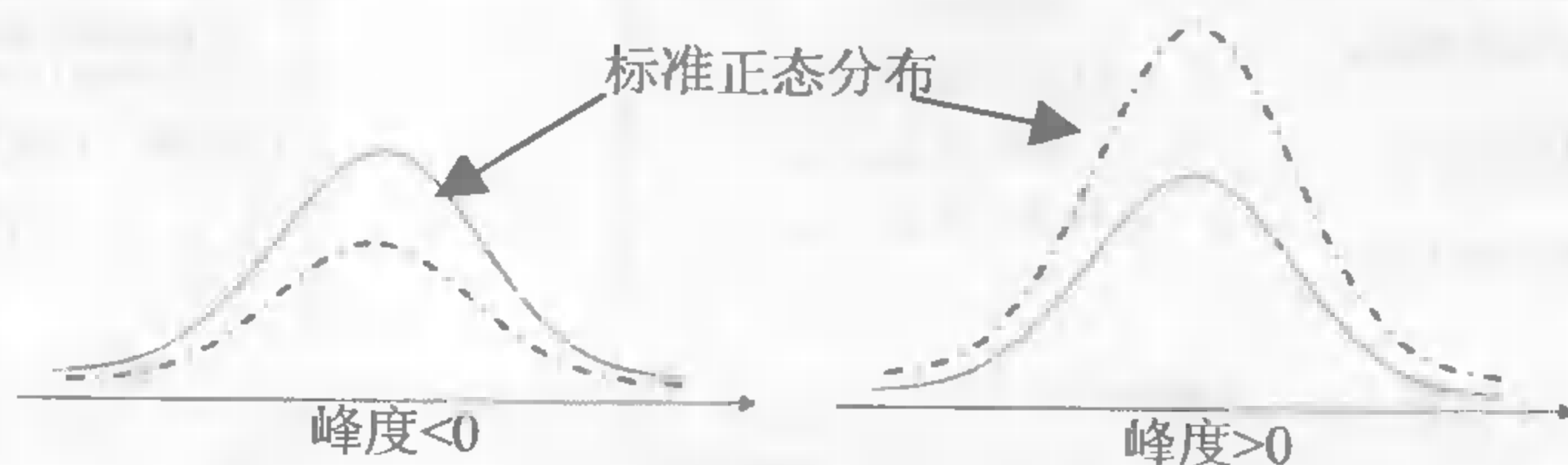


图 2-16 峰度与数据分布

在 SAS 编程语言中，可以直接利用 KURTOSIS 函数计算峰度。以例 2-7 中的公务员考试成绩为例，具体程序如下。

```
data;
x=kurtosis (190,188,188,185,183,183,180,180,180,180,178,177,175,175,174,173);
put '峰度=' x;
run;
```

## 2.2.4 利用菜单和程序进行详细的描述统计分析

在 SAS 系统中，除了可以使用上述的函数形式直接单独计算集中趋势、离散程度和分布状况的测度指标之外，还可以用菜单操作实现。

### 1. 利用 SAS/Insight 进行详细的描述统计分析

本小节主要以例 2-7 的数据（Gwy.sas7bdat）为例介绍 SAS/Insight 的描述统计分析。

**STEP 1** 进入 SAS/Insight，打开 Gwy.sas7bdat 数据集，选择“Analyze→Distribution”，弹出数据分布对话框，如图 2-17 所示。

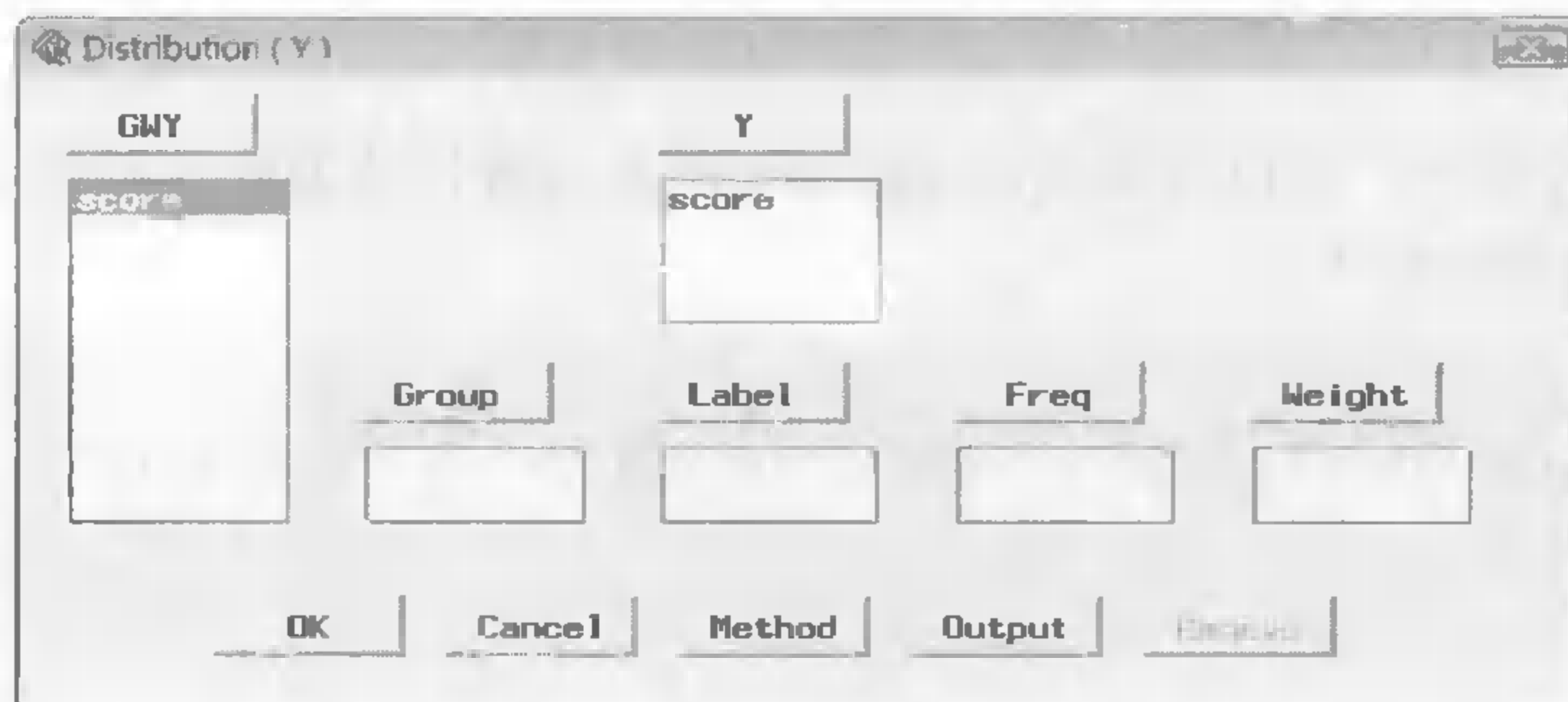


图 2-17 “Distribution”对话框

**STEP 2** 在对话框左边的变量列表中选中变量名“score”，然后单击“Y”按钮，表示将要 score 变量进行描述统计分析。单击“Output”按钮，弹出输出结果对话框，如图 2-18 所示。

**STEP 3** 在“Output”对话框中，可以对输出的统计量进行调整，只需在“Descriptive Statistics”分栏下选中对应的复选框即可。单击“Trimmed/Winsorized Means”按钮，弹出截尾/缩尾均值的设置窗口，如图 2-19 所示。

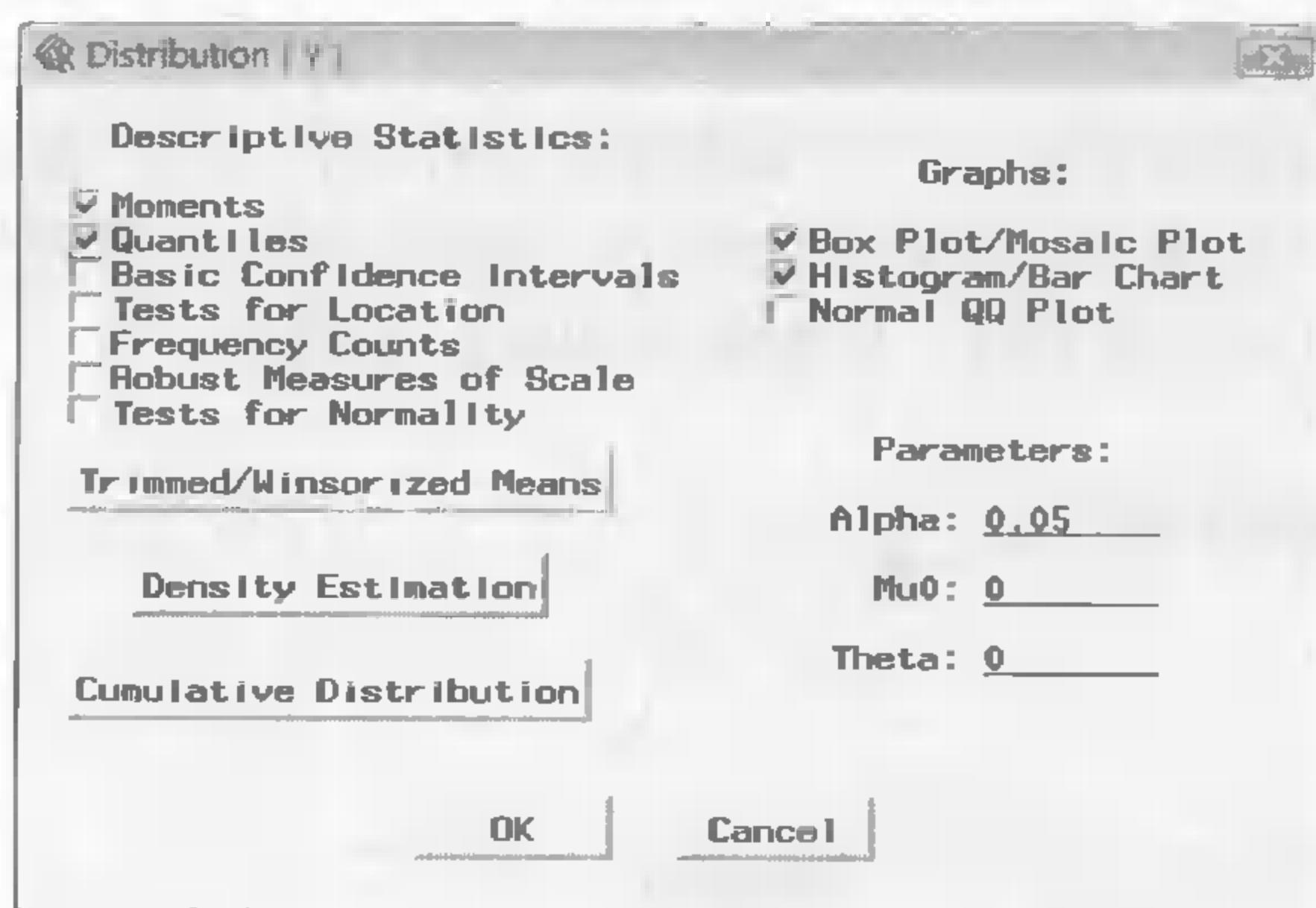


图 2-18 Distribution 的“Output”对话框

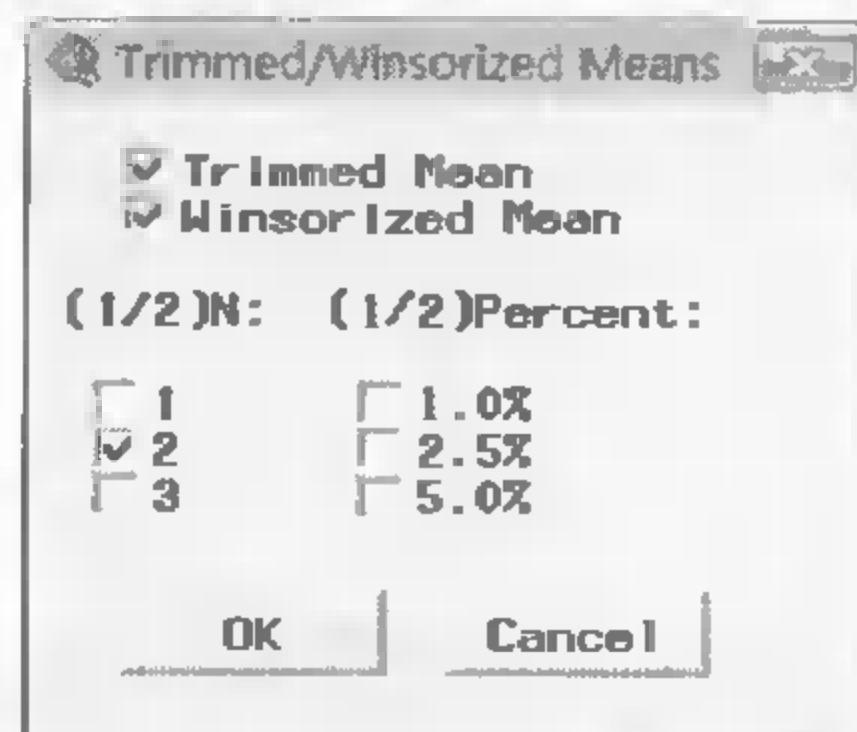


图 2-19 截尾/缩尾均值设定对话框

SAS 系统在默认的情况下是不给出截尾和缩尾均值的，如需计算这两个均值，应当在图 2-19 所示的对话框中选中“Trimmed Mean”和“Winsorized Mean”两个复选框，并且设定截尾或缩尾的样本量或百分比。如本例选择  $N=2$ ，只需在“(1/2) N”栏下选中“2”即可。

把所有的调整选项设置好后，单击“OK”按钮返回“Distribution”对话框，在该对话框中，单击“OK”按钮即可看到输出统计量结果。SAS/Insight 的输出结果包括盒式图和直方图，还有一些表格。

(1) Momen 表格：矩统计量表。本节所述的统计量均可被称为矩统计量。本例具体结果如图 2-20 所示。

该表格所表示的统计量如下。

- N：样本量，即参与计算的数据的个数。
- Mean：均值，即简单算术平均数。
- Std Dev：标注差。
- Skewness：偏度。
- USS：所有数据数值的平方和。
- CV：变异系数（注：在 SAS 系统中，CV 等于标准差与均值之比的 100 倍）。
- Sum Wg：权数之和。在数据没有加权的情况下，其等于样本量。
- Sum：所有数据的数值之和。
- Variance：方差。
- Kurosis：峰度。
- CSS：所有数据的离差平方和。
- Std Mean：标准误差。

(2) Quantiles 表格：分位数表。该表列示了典型的百分位数、四分位数及一些常见的位置平均数，如中位数、众数等。本例输出结果如图 2-21 所示。

Moments			
N	16.0000	Sum Wgts	16.0000
Mean	180.5625	Sum	2889.0000
Std Dev	5.2532	Variance	27.5958
Skewness	0.3456	Kurtosis	-0.8624
USS	522059.000	CSS	413.9375
CV	2.9093	Std Mean	1.3133

图 2-20 矩统计量表

Quantiles			
100% Max	190.0000	99.0%	190.0000
75% Q3	184.0000	97.5%	190.0000
50% Med	180.0000	95.0%	190.0000
25% Q1	176.0000	90.0%	188.0000
0% Min	173.0000	10.0%	174.0000
Range	17.0000	5.0%	173.0000
Q3-Q1	8.0000	2.5%	173.0000
Mode	180.0000	1.0%	173.0000

图 2-21 分位数表

该表格所表示的统计量如下。

- 100% Max：最大值。
- 75% Q3：上四分位数。
- 50% Med：中位数。
- 25% Q1：下四分位数。
- 0% Min：最小值。
- Range：级差，即全距。
- Q3-Q1：四分位差，即上四分位数与下四分位数之差。
- Mode：众数。

如果选择输出截尾或缩尾均值，则系统还会给出截尾或缩尾均值的具体数值，并给出其在 95%置信度下的置信区间的上下限，并进行假设检验。

2. 利用 SAS 编程语言的过程进行详细的描述统计分析

可以用 SAS 编程语言中的 FREQ、MEANS、UNIVARIATE 过程进行描述统计分析。

(1) FREQ 过程。

FREQ 过程主要用于计算数据的频数和一些用于检验的统计量。其主要语法如下。

```
proc freq <选项>;
  by 变量/变量列表;
  exact 统计选项</计算选项>;
  output <out=输出数据集> <输出数据集所包含的统计量> 选项;
  tables 变量名列表</选项>;
  test 选项;
  weight 变量 </选项>;
```

上述关键字的具体解释如下。

- by: 指定输出结果的分组变量，所有结果均按照指定变量的不同值分别进行计算。
- exact: 对特定统计量进行精确检验。
- output: 产生含有特定统计量的新数据集，即把输出结果存放至数据集中。
- tables: 生成多个变量的交叉分析表格，并对相关性进行测度和检验。
- test: 对交叉表格一致性和相关性进行近似检验。
- weight: 指定作为权数的变量。

为了详细说明该语句的使用方法，下面采用例 2-2 的数据 (CCI.sas7bdat) 进行讲解。根据例 2-2 的数据，考察不同性别的北京市民对经济发展状况的评价情况，使用 FREQ 过程的程序如下。

```
proc freq data=Sasuser.CCI; /*调用 FREQ 过程对 Sasuser.CCI 数据集进行分析*/
  table gender CCI; /*分别列出变量 Gender 和 CCI 的频数分布表*/
  table gender * CCI; /*列出变量 Gender 和 CCI 的交叉频数分布表*/
run;
```

使用 FREQ 过程的 TABLE 语句时，应当注意变量的表示方式。如在上述程序的第 2 行中，gender 和 CCI 变量中间用空格分开，表示分别列出 gender 变量的频数表和 CCI 变量的频数表，如图 2-22 所示。

The FREQ Procedure				
性别				
Gender	Frequency	Percent	Cumulative Frequency	Cumulative Percent
男	300	50.42	300	50.42
女	295	49.58	595	100.00
北京市居民消费者信心状况				
CCI	Frequency	Percent	Cumulative Frequency	Cumulative Percent
差	10	1.68	10	1.68
好	44	7.39	54	9.08
较差	40	6.72	94	15.80
较好	255	42.86	349	58.66
一般	246	41.34	595	100.00

图 2-22 频数表



- types: 指定 class 变量的组合形式。
- var: 指定需要进行描述统计分析的变量和次序。
- ways: 指定单一 class 变量组合方式数。
- weight: 指定作为权数的变量。

在 MENAS 过程的选项中，可以设置需要进行计算的统计量，其关键词如表 2-8 所示。

表 2-8 MEANS 过程的统计量关键字

描述统计量		分位数统计量	
CLM	置信区间	MEDIAN P50	中位数
CSS	离差平方和	P1	1%分位数
CV	变异系数	P5	5%分位数
KURTOSIS KURT	峰度	P10	10%分位数
LCLM	置信区间下界	Q1 P25	下四分位数
MAX	最大值	Q3 P75	上四分位数
MEAN	均值	P90	90%分位数
MIN	最小值	P95	95%分位数
N	样本量	P99	99%分位数
NMISS	缺失样本量	QRANGE	四分位差
RANGE	级差	假设检验统计量	
SKEWNESS SKEW	偏度	PROBT	T 统计量的 p 值（双尾）
STDDEV STD	标准差	T	T 统计量的值
STDERR	标准误差		
SUM	和		
SUMWGT	加权和		
UCLM	置信区间上界		
USS	平方和		
VAR	方差		

在默认的情况下，MEANS 过程自动计算样本量、均值、标准差、最小值和最大值等五个统计量。

以例 2-7 为例，对公务员成绩（Gwy.sas7bdat）进行描述统计分析，具体程序如下。

```
proc means data=Sasuser.Gwy qrange sum mean n uss std range cv; /*列示出需要计算的统计量*/
  var score;
  output out=Sasuser.Gwy_Output; /*把计算结果保存在 Sasuser.Gwy 数据集中*/
run;
```

(3) UNIVARIATE 过程。

UNIVARIATE 过程能够进行比较复杂的描述统计分析，除了能够实现 MEANS 的功能之外，还可以绘制数据的分布图、计算变量的频数表以及进行假设检验，其主要语法如下。

```
proc univariate <选项>;
  by 变量;
  class 变量 <选项> <变量 <选项>> </keylevel= 数值| ( 数值 1 数值 2)>;
  freq 变量;
  histogram < 变量> </选项>;
  id 变量;
  inset 关键词列表 </选项>;
  output <out=新数据集> <关键字 1=新变量名...关键字 K=新变量名><百分位数选项>;
  probplot <变量> </选项>;
  qqplot <变量> </选项>;
  var 变量;
  weight 变量;
```

上述关键字的具体解释如下。

- by: 指定进行分组计算的变量。
- class: 指定观测组并计算其统计量。
- freq: 指定作为频数处理的变量。
- histogram: 生成高精度直方图。
- id: 指定识别极端值的变量。
- inset: 在图形中插入统计表。
- output: 把结果存储在新的数据集中。
- probplot: 生成高精度的概率图。
- qqplot: 生成高精度的 Q-Q 图。
- var: 指定进行统计分析的变量。
- weight: 指定作为权数的变量。

在 UNIVARIATE 过程的选项中, plot、freq 选项比较常用。其中, plot 表示利用 UNIVARIATE 过程绘制茎叶图、盒式图和正态概率图, freq 表示生成由变量值、频数、频率、累积频数构成的频数分布表。以例 2-7 为例,对公务员成绩(Gwy.sas7bdat)进行 UNIVARIATE 描述分析,具体程序如下。

```
proc univariate data=Sasuser.Gwy plot freq; /*绘制图形,并计算频数分布表*/
  var score;
  output out=Sasuser.Gwy_Output1 qrange=qr mean=average q3=quateile_3; /*将结果保存在新数据集*/
run;
```

程序中的 output 语句可以用于指定并保存 UNIVARIATE 过程计算的统计量,其能够计算的统计量不仅包括表 2-8 所示的所有统计量,还包括所谓的稳健估计量和假设检验统计量。

## 2.3 统计表

在上一节中,介绍了数据可以以图形的形式呈现和展示。但是当需要对数据进行深入剖析以发现数据的内在联系时,光靠图形展示数据是不够的,因为从图形中能够获得信息量有限。

假设读者作为一个上市公司的财务总监,需要公布公司的财务状况,如果仅作为公司业

绩的展示和宣传，用形形色色的图可以很快地吸引别人的眼球，从而达到广而告之的目的。但是如果要进行正式的财务状况汇报，如公布年报，投资者需要通过年报完全掌握有关公司的详细信息，这时图形是远远不够的，而需要使用统计表格来呈现数据，从而把数据的原始信息和数量关系用二维的形式完全地展现出来，以方便投资者进行进一步的研究。

2.3.1 统计表的基本要素

统计表实质上就是一张二维表格，它有其自身的特点和构成要素。首先来看个例子。



例 2-9

某电脑销售公司对其 2008 年第 1 季度的笔记本电脑销售情况进行了详细记录（数据见 Laptop.sas7bdat）。经汇总得到以下销售情况的简要汇报，如表 2-9 所示。

表 2-9 笔记本销售情况一览表（单位：台）

地 区	品 牌					合 计
	戴尔	华硕	惠普	联想	索尼	
东 部 地 区	18	23	14	27	17	99
西 部 地 区	13	30	23	22	12	100
中 部 地 区	32	15	18	28	7	100
合 计	63	68	55	77	36	299

形如表 2-9 的表格通常被称为统计表格。统计表格的最大特点是表格左右两端是开口的，即左右两端没有竖线。



SAS 系统中绘制的统计表格都是封口的。

统计表通常由行和列组成，分别代表行维和列维，亦即二维表格的由来。如表 2-9 中，“地区”是行，地区变量的 3 个值把表格分成了 3 行；“品牌”是列，其 5 个值把表格分成了 5 列。统计表中往往还有汇总的合计项，又可以分成行合计与列合计，分别统计每行或每列分析变量的汇总情况。如在表 2-9 中，可清晰地看到该公司第 1 季度共销售了 299 台笔记本电脑，其中东部地区销售 99 台，西部地区销售 100 台，中部地区销售 100 台。各个品牌的销售汇总情况也可以从列合计中看出来，如联想笔记本电脑在 3 个地区共销售了 77 台，而索尼只售出了 36 台。

当有多个这样的销售公司时，每个公司都有一个这样的表格，把这些表格排列起来，就形成了一系列的销售报表。表格与表格之间的维度叫做页维，如某表格页维为 3，表明具有 3 个同样结构的表格。

在 SAS 系统中，必须先把上述的页维度、行维度、列维度等概念搞清楚，否则进行分析时很容易出错。

2.3.2 用 TABULATE 过程绘制统计表

在 SAS 系统中，利用 TABULATE 过程进行表格绘制十分方便，本节将详细介绍该语句在实际工作中的使用方法。

TABULATE 过程的主要语法如下：

```
proc tabulate <选项>;
  by <descending> 变量 1<...<descending> 变量 n> <no orted>;
  class 变量 </选项>;
  classlev 变量/style=<格式名称| parent> <[格式属性] >;
  freq 变量;
  keylabel 关键字 1='关键字描述 1' <...关键字 n='关键字描述 n'>;
  keyword 关键字/style=<格式名称 | parent> <[格式属性] >;
  table <<页维度,> 行维度,> 列维度</表格属性>;
  var 变量</选项>;
  weight 变量;
```

各语句的功能如下：

- by: 按指定变量分表制表。
- class: 指定分类变量。
- classlev: 指定分类变量标题的格式。
- freq: 指定作为表示频数的变量。
- keylabel: 指定关键词的标签。
- keyword: 指定关键词的格式。
- table: 指定绘制表格的布局及形式。
- var: 指定表格中的分析变量。
- weight: 指定作为权重的变量。

在 TABULATE 过程的众多语句中，table 是最重要的语句，所有由 TABULATE 过程绘制的表格，都应通过 table 语句对表格的布局进行设计。table 语句的主要功能如下。

1. 设计表格布局

表格的 3 个维度（可以省略）之间以“，”隔开，如果 tABLE 之后只有 1 个变量表达式，则表示列维，如下所示。

```
tABLE 地区, 品牌, 销售方式
```

该语句表示表格以地区的形式分开，每个地区对应一张表格，每张表格的行表示品牌，列表示销售方式。

如果需要考察交叉变量的情况，则相交叉的变量之间用“\*”连接，如下所示。

```
tABLE 地区*品牌
```

该语句表示表格只有列维度，且列维度表示地区和品牌两个变量的交叉情况。变量之间的交叉可以进行嵌套，如下所示。

```
tABLE (A B)*C
```

该语句相当于 A\*C 和 B\*C 两种变量交叉情况。

2. 计算统计量

在 tABLE 语句中，可设置表格中出现的统计量。统计量的关键字可以从表 2-8 中获得。如按照地区计算各品牌的销售均值，语句如下。

```
tABLE 地区*MEAN*品牌
```

此外，在 `TABLE` 语句中还可以用关键字 `all` 指定表格按照行或按照列进行汇总，如下所示。

```
TABLE 地区 all,品牌
```

该语句表示按照地区进行行汇总。  
在使用 `table` 语句时，`table` 语句中的变量必须通过 `class` 语句声明为分类变量。  
以例 2-9 为例，使用 `TABULATE` 过程绘制表 2-9 所示的统计表格，具体程序如下。

```
proc tabulate data=Sasuser.Laptop;
  class district brand;          /*指定 district 和 brand 变量为分类变量*/
  table district all,brand all;  /*指定 district 为行变量，指定 brand 为列变量，行列都进行合计*/
  keylabel N='销售量' all='合计'; /*把次数关键字 N 命名为“销售量”，all 关键字命名为“合计”*/
run;
```

接下来，利用例 2-9 的数据绘制更为复杂的表格，如表 2-10 所示。

表 2-10 笔记本销售汇总表

地区	品 牌														
	戴 尔			华 硕			惠 普			联 想			索 尼		
	销售量 (单位:台)	销售 额(单位:元)	平均价 格(单位:元)	销售量 (单位:台)	销售 额(单位:元)	平均价 格(单位:元)	销售量 (单位:台)	销售 额(单位:元)	平均价 格(单位:元)	销售量 (单位:台)	销售 额(单位:元)	平均价 格(单位:元)	销售量 (单位:台)	销售 额(单位:元)	平均价 格(单位:元)
东部	18	227 514	12 640	23	278 819	12 123	14	180 739	12 910	27	353 638	13 098	17	218 854	12 874
西部	13	159 903	12 300	30	382 401	12 747	23	302 619	13 157	22	285 058	12 957	12	124 051	10 338
中部	32	403 971	12 624	15	158 957	10 597	18	220 858	12 270	28	386 131	13 790	7	83 149	11 878
合计	63	791 388	12 562	68	820 176	12 061	55	704 216	12 804	77	1 024 827	13 309	36	426 054	11 835

表 2-10 中列示了销售量、销售额及平均售价的详细情况，并分别对品牌进行了汇总。从该表格中可以清晰地看到不同品牌电脑在不同地区的销售明细情况。结合表 2-9 和表 2-10 所示，可以得出统计分析结论：东、中、西部 3 大区域的销售量情况相差不大，说明销售公司的区域平衡策略是成功的；但是各个品牌的电脑销售情况差异较大。其中，联想笔记本的销售情况最好，由于该品牌平均价格比较高，因此销售总额处于领先地位；而索尼笔记本的销售情况最差，其平均价格是最低的，而且由于销量偏低，从而造成其销售总额大大低于其他品牌。

利用 SAS 的 `TABULATE` 过程可以绘制表 2-10 所示的统计表格，具体程序如下。

```
proc tabulate data=Sasuser.Laptop;
  class district brand;
  var price;
  table district all,brand*(n price price*mean);
  keylabel N='销售量' all='合计' sum='销售额' mean='平均价格';
run;
```

2.4 数据分布

数据分布也是描述数据的一种形象方式。可以通过数据分布考察数据中各个数值出现的次数，

而且数据分布描绘了数据出现次数的变动状况。同时，数据分布通常是针对随机变量而言的，它也是进行统计推断的基础之一，绝大多数统计推断的结论都是从数据分布开始展开论述的。

数据分布可以细分为离散变量的分布和连续变量的分布，其中连续变量的分布是统计推断的数学基础。根据分布的表现形式，数据分布又可以分为正态分布、 $t$ 分布、 $F$ 分布、卡方分布等许多类型。本章为简单实用起见，同时也为下一章介绍统计推断的内容打下基础，只简单介绍按研究对象进行划分的数据分布，即总体分布、样本分布和抽样分布。

### 2.4.1 总体分布

总体是一个具有确切分布的随机变量，总体分布就是所有数据的分布。总体分布具体是指总体的所有变量值的分布状况，即总体变量值分布状况的一种概括。如研究新生儿的性别，收集所有新生儿的性别并考察他们的性别，男、女二者必居其一，故属于两点分布。又如全世界所有人的身高服从正态分布等，如图 2-24 所示。

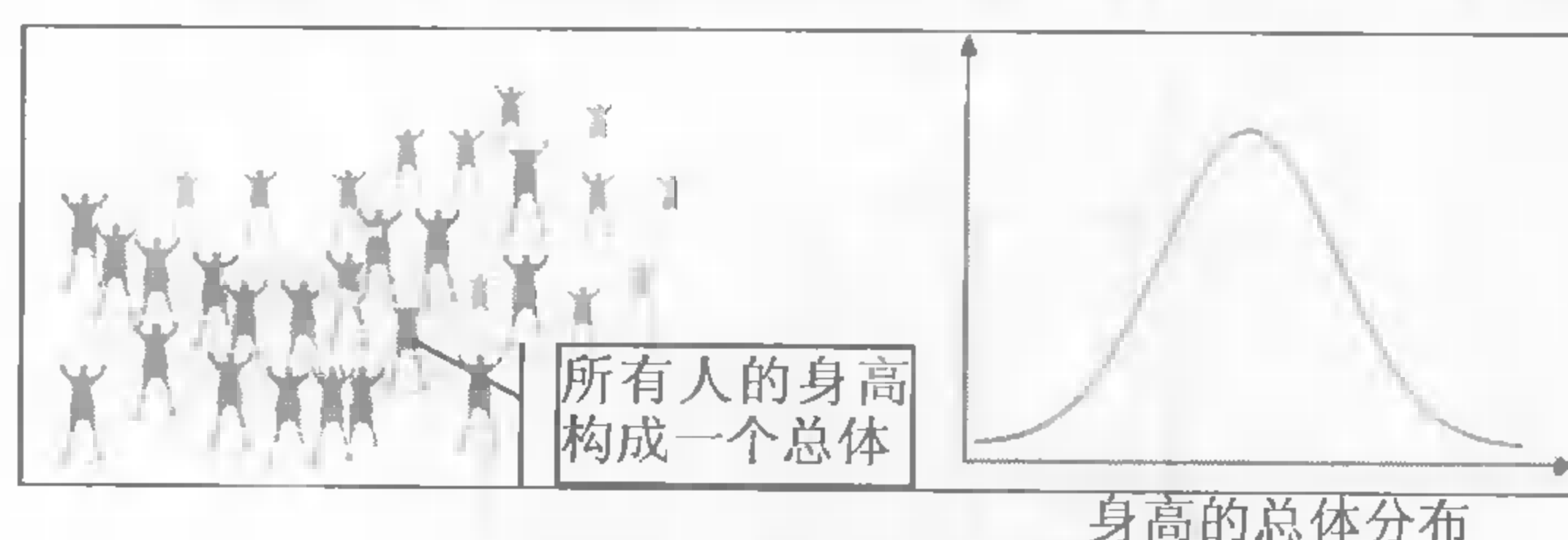


图 2-24 总体与总体分布

在现实生活中，总体的每一个观测值几乎不可能都能够获得，因此才有必要对总体的特征进行推断。所以，总体分布往往是未知的，总体的特征往往也是未知的。在通常的情况下，假定总体服从一个特定的分布，在该假定下进行统计分析。总体的特征也叫做总体参数，由于不能够完全获得总体数据，所以总体参数往往是未知的，但是是唯一确定存在的。这是因为总体一旦确定，总体参数自然而然就确定了。

在 SAS 系统中，不能够直接计算总体参数，总体的特征只能通过样本数据进行推断。

### 2.4.2 样本分布

总体数据既然很难获得，那么人们可从总体中抽取出若干部分的个体进行调查，以进行研究。从总体中抽取一个容量为  $n$  的样本，那么这些样本观测值是有差异的，并形成一個样本分布（或子样分布），也就是样本中各观察值的分布。其中  $n$  叫做样本容量，简称样本量。

样本总是在一定总体中抽取的，其中包含总体的一些信息，所以也被称为经验分布。随着样本量的增大，样本的分布会逐渐接近于总体分布。

样本数据很容易获得，但是从同一个总体中可以抽取出若干个不同的样本，因此样本之间还是有差异的。这种差异是随机的，可以通过标准误差来衡量。在通常情况下，可以通过样本统计量和样本分布来对总体进行推断。

在 SAS 系统中，只能够根据所获得的数据计算样本统计量，并描绘样本分布的状况。

### 2.4.3 抽样分布

抽样分布是指样本统计量的分布。

在进行数据的抽样调查中，由于随机性的存在，同一种抽样方法在不同的时间、地点、

环境等条件下进行多次抽样，可能得到多个不同的样本数据；同一种抽样方法在相同的环境下由不同的人员实施，也可能得到多个不同的样本数据。把所有类似情况下获得的所有可能样本找出来，如在总体容量为  $N$  的总体中，随机抽取样本容量为  $n$  的样本，如考虑排列顺序，一共可能获得  $N^n$  个样本；如不考虑排列情况，则可获得  $C_N^n$  个样本。

计算每个样本的统计量，可以得到一些列的统计量数据，所有这些数据形成的分布就是抽样分布。具体地说，样本均值的分布、样本标准差的分布、样本方差的分布、样本比例的分布等都被叫做抽样分布。

抽样分布提供了样本统计量长远而稳定的信息，是进行推断的理论基础，也是抽样推断科学性的重要依据。但是在实现生活中，不能把所有的样本都抽取出来，因此抽样分布又是一种理论上的分布。为了理解抽样分布，下面来看一个简单的例子。



例 2-10

设一个总体，含有 4 个个体，即总体单位数  $N=4$ 。4 个个体分别为  $x_1=2$ 、 $x_2=4$ 、 $x_3=6$ 、 $x_4=8$ 。总体的均值、方差及分布如图 2-25 所示。

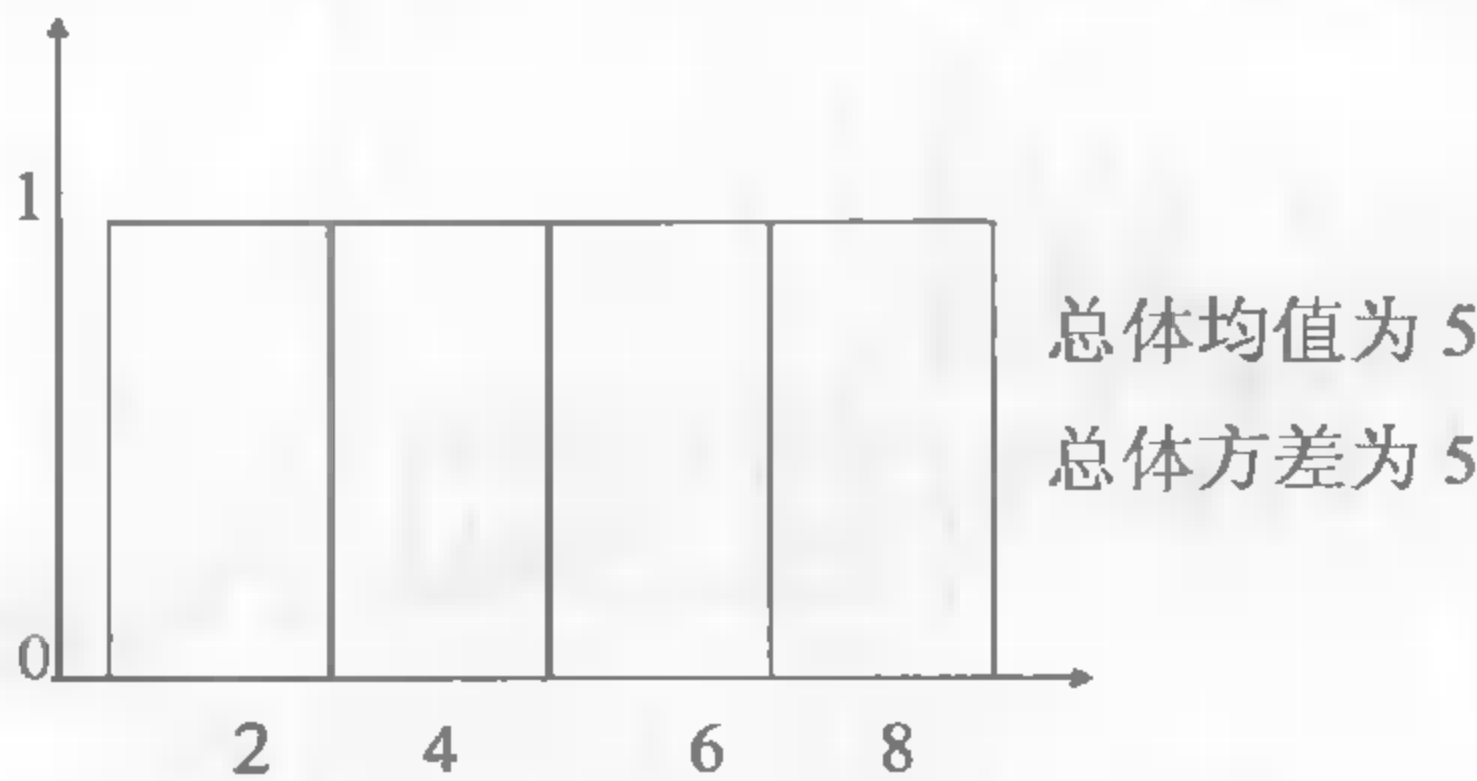


图 2-25 一个总体分布的例子

现按照随机原则，从总体中抽取样本容量  $n=2$  的简单随机样本，在重复抽样条件下，共有  $4^2=16$  个样本。所有样本的结果如表 2-11 所示。

表 2-11 一个总体的所有随机样本

第一个样本的观测值	第二个样本的观测值			
	2	4	6	8
2	2, 2	2, 4	2, 6	2, 8
4	4, 2	4, 4	4, 6	4, 8
6	6, 2	6, 4	6, 6	6, 8
8	8, 2	8, 4	8, 6	8, 8

表 2-11 中所列的 16 个样本已经穷尽了所有可能从 4 个个体组成的总体中抽取 2 个个体的情况。根据表 2-11 计算每个样本的统计量（通常计算均值），结果如表 2-12 所示。

表 2-12 所有随机样本的统计量（均值）

第一个样本的观测值	第二个样本的观测值			
	2	4	6	8
2	2	3	4	5
4	3	4	5	6
6	4	5	6	7
8	5	6	7	8

于是，表 2-12 中所有数据的分布情况构成了均值的抽样分布。把这些均值作为数据，绘制抽样分布图（或直方图），如图 2-26 所示。

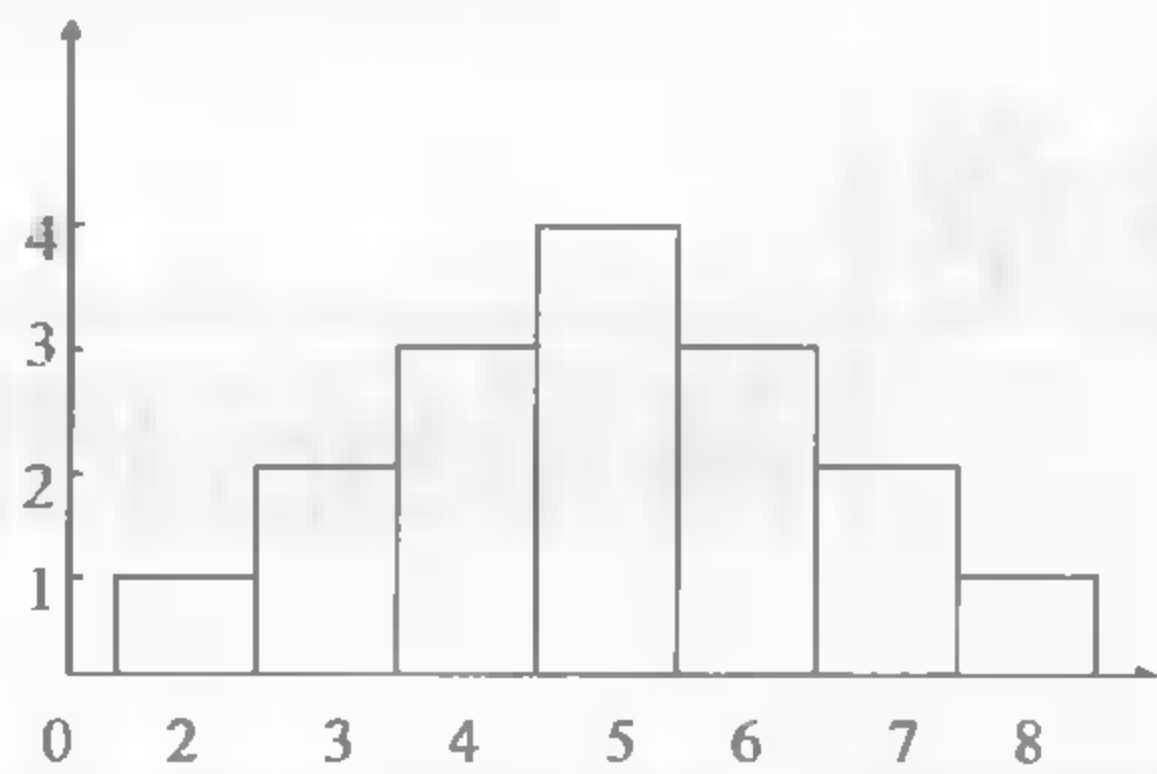


图 2-26 一个抽样分布的例子

可以利用 SAS 程序计算抽样分布的均值和方差，具体程序如下。

```
data;  
  x=mean(2,3,4,5,3,4,5,6,4,5,6,7,5,6,7,8);  
  y=css(2,3,4,5,3,4,5,6,4,5,6,7,5,6,7,8);  
  z=y/16;      /*SAS 中不能直接计算总体方差，故采用总体方差的公式直接计算。在本例中，n=16*/  
  var=5;      /*总体方差的值*/  
  n=var/z;  
  if x=5 then  
    put "抽样分布的均值等于总体均值";  
  else put "抽样分布的均值不等于总体均值";  
  put "抽样分布的方差等于 1/" n "倍总体方差";  
run;
```

由该程序的运行结果可以看出，抽样分布的均值与总体均值相等，抽样分布的方差是总体方差的  $1/n$ （本例  $n=2$ ）倍。

该结论不是本例特有的结论，而是可以根据抽样分布基本特征推导出的中心极限定理，即当样本量  $n$  充分大（ $n \geq 30$ ）时，样本均值的分布服从以总体均值为均值、以总体方差的  $1/n$  倍为方差的正态分布。所以说抽样分布是一种稳定的分布，可以作为统计推断的理论基础。对于该部分理论，第 3 章将详细讲解。

## 2.5 本章小结

本章介绍了利用 SAS 系统进行描述统计分析的步骤和方法，主要内容简要回顾如下：讲述了如何绘制和阅读常用的直方图、条形图、盒式图、茎叶图、饼图等统计图形；详细介绍了常用的描述统计量，集中趋势、离散程度和分布形态等是进行描述的主要方面，可以通过 SAS 系统计算样本均值、中位数、分位数、众数、方差、标准差、变异系数、偏度、峰度等统计量以对数据进行描述；统计表格也是描述数据的一种重要手段，详细介绍了利用 TABULATE 过程制表的方法；数据的分布也是数据描述的一个重要方面，主要介绍了总体分布、样本分布、抽样分布及它们之间的内在关系，为第 3 章介绍统计推断打下了基础。

## 第 3 章

# 简单统计推断

现实世界都是未知的，如果把这个世界作为一个总体，谁也无法准确地知晓这个总体的特征，即准确测算这个总体的参数。但是作为在现实世界中生活的人，可以通过大量的经验和观察推测未知世界的特征，通过收集能够掌握的数据和信息对这个总体进行推测性的描述。当然，这种推测是有概率保证的。统计学家从各种纷繁复杂的现象中，通过数据的收集和分析，利用一定的统计推断算法和统计分析工具，为人们打开未知世界的大门提供了钥匙和捷径。

因此，本章将对简单统计推断的基本原理和内容进行详细讲解，并利用 SAS 系统实现推断的过程。

### 3.1 简单统计推断的基本原理

在第 2 章中介绍描述统计分析时，研究了总体分布、样本分布和抽样分布 3 种分析的特征及其内在联系。统计推断的内容便是基于这 3 者之间的关系以抽样分布为核心展开的。

所谓统计推断，就是在总体中按随机原则抽取一部分单位作为样本，根据样本数据归纳或推断总体数量特征的一种统计方法。基本原理是抽样推断中的大数定律和中心极限定理。

在对总体进行抽样的过程中，由于随机原则的存在，样本和总体总是有差异的，因此总是不可避免地存在抽样误差。

抽样误差具体是指所有可能出现的样本指标的标准差，也可以理解为所有样本指标和总体指标之间的平均离差。当取样本均值作为样本指标时，抽样误差就是第 2 章中所提到的标准误差。

由于误差的存在，在对样本数据进行分析的过程中，得出的结论只能是不确定的结论。这种不确定性并不代表不正确，而是代表在一定的误差容许范围内，结论的正确性是有概率保证的。因此，统计学的结论往往是不确定的。

在对例 2-10 的分析中，得到了一个重要的结论，即抽样分布的均值等于总体均值，抽样分布的方差等于总体方差的  $1/n$  倍（ $n$  为样本量）。在日常生活中，由于条件有限，无法得到所有可能抽样的样本，因此上述情况是对理想状况的一个描述，在样本量充分大的情况下总是成立的，人们称之为中心极限定理，即：

当样本量  $n$  充分大（ $n \geq 30$ ）时，样本均值的分布服从以总体均值为均值、以总体方差的  $1/n$  倍为方差的正态分布。

其中，样本量  $n \geq 30$  是一个经验法则，在日常统计分析过程中，可以据此判断结论的可靠性。

抽样推断的内容涵盖较广，大体上可分为对总体参数的推断和非参数推断。

总体参数的推断主要是指根据抽样分布对总体的特征进行估计和检验，因此参数推断应当事先知道总体的分布状况。非参数推断是在未知总体分布条件下对总体的分布情况进行推断。此外，统计推断既可对一个总体进行，也可以对多个总体进行。

本章阐述的简单统计推断的内容可总结为参数估计和假设检验两个基本问题。

### 3.1.1 参数估计

总体的特征可称为参数。通常，人们很难获得总体的全部数据，总体往往是未知的，因而总体特征即总体参数也往往是未知的。但是总体一旦确定下来，总体参数就确定了，总体参数的数值也不会改变，因此总体参数是确定的。此外，一个研究对象往往对应一个总体，总体参数也是唯一的。

人们总可通过一些手段和方式收集总体的样本数据，利用样本数据计算出样本统计量，根据统计推断的基本原则，在一定的概率或置信度下对总体参数进行推断。而样本统计量来自于样本数据，样本数据相对于总体数据而言具有多样性和不确定性，但是是已知的。

统计推断的参数估计过程便是利用已知、不确定、不唯一的样本统计量去推断未知、确定、唯一的总体参数的过程。此过程可以通过利用样本数据对总体特征进行估计的方法实现。如用样本均值估计总体均值、样本方差估计总体方差、样本比例估计总体比例等。

由于估计值和真实值之间存在着一定的误差，所以参数估计过程一般是在一定的概率或置信度下做出的。这个概率保证通常被称为置信度。

参数估计根据是否有置信度作为保证，可以分为点估计和区间估计两种形式。

#### 1. 点估计

点估计就是直接用实际抽样的样本数据得到的样本统计量的值作为总体参数的估计值，如直接用样本均值作为总体均值的估计值。用于点估计的估计量是一种不考虑抽样误差问题的估计方法，可以利用矩估计、极大似然估计等方法进行测算。

点估计的方法比较简单，但是由于其不考虑抽样误差，因此可靠性比较差，在大样本的情况下比较常用。在实际应用中，点估计应当满足无偏、有效、一致性的要求。

如想知道某学校所有学生的统计学期末考试成绩的平均分和标准差，用点估计的方法可以得到总体平均分和总体标准差的估计值。抽查 100 名同学该门课程的成绩作为样本，分别计算样本平均分和标准差，得到 85 分和 8.9 分，则此 85 分和 8.9 分就是总体平均分和总体标准差的点估计值。

#### 2. 区间估计

为了提高估计的精度，往往在估计时给出一个可靠程度，即置信度。根据置信度的要求，利用随机抽取的样本的统计量来确定总体参数估计值的估计上限和估计下限，即根据样本指标确定置信区间的方法被叫做区间估计。

区间估计，即在一定的置信度下给出总体参数估计值的一个估计区间，其中的置信度是根据抽样误差设置的一种概率保证。置信区间也可以理解为在置信度条件下，根据样本统计量和抽样误差推断总体参数的可能范围。

如在 95% 的置信度下，得出某大学全体同学统计学期末考试成绩的平均分的置信区间为

[82.5, 88.7]。这个估计结果应当这样理解：在重复构造的 100 个总体参数的置信区间中，有 95 个区间都包含了总体参数的真实值，它们的上限是 88.7，下限是 82.5，即有 95% 的样本均值所构成的区间中会包括总体均值。更加通俗地说，是指在 100 次抽样中，大概有 95 次所得到的区间包含总体均值的真实值。而不能理解为这个 [82.5, 88.7] 的区间以 95% 的概率或可能性包含总体平均分的真实值。区间估计的示意图如图 3-1 所示。

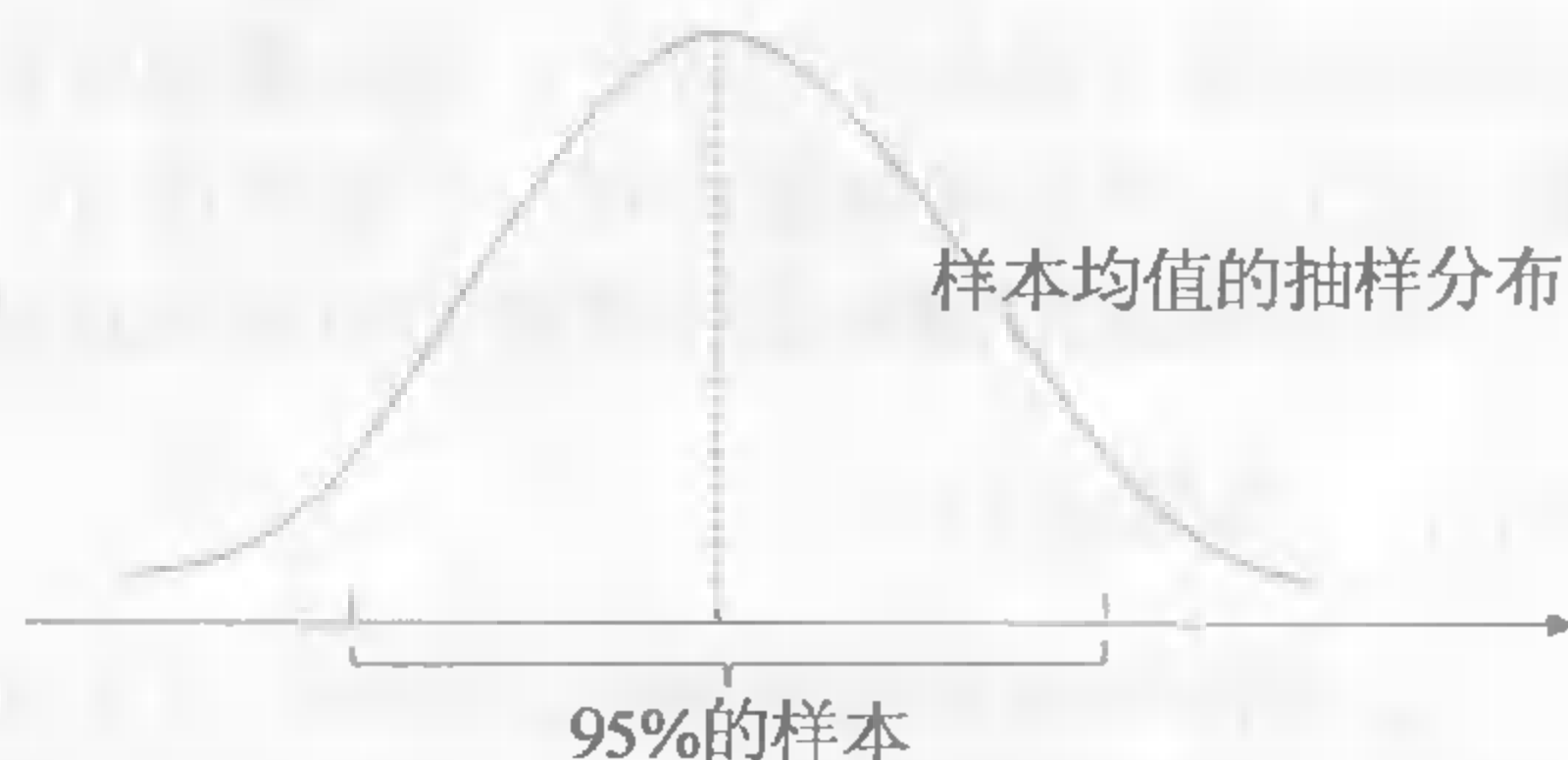


图 3-1 区间估计的一个示意图

置信度和置信区间是相辅相成的。在抽样误差和样本量保持不变的条件下，置信度越大，置信区间也越大；置信度越小，置信区间也越小。常用的置信度为 95%、99%、90% 等。在 SAS 系统中，可对总体均值、总体方差等参数进行区间估计。

### 3.1.2 假设检验

假设检验是统计推断中最为核心的问题之一，主要研究在一定的情况下，总体是否具备某种（些）指定的特征。

#### 1. 假设检验的基本思想

有一个人说：“我是一个从来不做坏事的好人”，那么如何去对这个人所说的话进行检验呢？应当有一个严格的标准，根据好人的严格定义和这个人说的话，好人就是做好事的人，一个人是好人的话就不会做坏事。

在通常情况下，至少有两种方法对“我是一个从来不做坏事的好人”这个结论进行检验。

第一种方法是从这个人出生那一刻起，到他死的那一刻为止，每时每刻、每时每地地观测这个人是否做好事。如果确实做好事的话，就可以肯定这个人是个好人了。抛开观测的实施者是否能够活得比被观测者长的问题，这种方法十分耗时耗力，在实现生活中几乎没有可能实施。为了一个确定的结论付出了巨大的代价，但是这种巨大代价所得到的结论是十分确定无疑的。

第二种方法相对简单。既然肯定一个人这么难，那么否定一个人就显得比较容易了。既然这个人声称是“从来不做坏事的好人”，那么相对于他自身而言，他干坏事的概率或可能性就非常小，换句话说，他干坏事被人发现的几率应当非常小，或者说他即使干了坏事，被人发现的概率也是非常小的。否则的话，也就不会这么理直气壮地发布豪言壮语了。于是，可收集证据以证明这个结论。只要发现他干了坏事，而且有人能够证明他干了坏事，那么他就不是“从来不干坏事的好人”了。因为按照他的假设前提，他是不会干坏事的或者说干坏事的几率非常小，几乎不可能发现他干坏事。但是只要有一个人发现他干过坏事，概率这么小的事情在一次观测中都能发生并被人发现，说明事情的假设是不可靠的，也就可以否定他的说法了。当然，如果已经访问了大部分的人，他们都没有发现他干过坏事，这时只能说没有足够的理由证明他干过坏事，即没有足够的理由否定他说的这句话，因为不可能把在所有时间、地点接触过他的人都访问到。

因此，第二种检验方法的结论是不确定的，得到的结论是有犯错误的概率的，这个概率的最大值就是相对于假设的那件不可能发生的事情在一次观测中发生的概率。

相对于假设而言的、在一次观测或试验中几乎不可能发生的事情，人们称之为小概率事

件。小概率事件在一次试验中发生的概率被称为显著性水平。

假设检验的基本原理就是小概率事件原理,即观测小概率事件在假设成立的情况下是否发生。如果在一次试验中,小概率事件竟然发生了,说明假设在一定的显著性水平下不可靠或不成立;如果在一次试验中,小概率事件没有发生,只能说明没有足够理由相信假设是错误的,但是不能说明假设是正确的,因为无法收集所有的证据证明它是正确的。

从上面的分析过程中可以看到,假设检验的结论是在一定的显著性水平下得出的。因此,当观测事件并下结论时,有可能犯错误。在假设检验过程中,无法保证不犯错误。这些错误归纳起来主要有两类。

- 第 I 类错误:当假设为真时,却否定它而犯的误差,即拒绝正确假设的误差,也叫弃真误差。犯第 I 类错误的概率记为  $\alpha$ ,所以通常也叫做  $\alpha$  误差,  $\alpha = 1 - \text{置信度}$ 。
- 第 II 类错误:当假设为假时,却肯定它而犯的误差,即接受错误假设的误差,也叫纳伪误差。犯第 II 类错误的概率记为  $\beta$ ,所以通常也叫做  $\beta$  误差。

这两类误差在其他条件不变的情况下是相反的,即  $\alpha$  增大时,  $\beta$  就减小;  $\alpha$  减小时,  $\beta$  就增大。要想同时减小两类误差,只能增加样本量。

$\alpha$  误差容易受分析人员的控制,因此,在假设检验中,通常先控制第 I 类误差发生的概率  $\alpha$ ,具体表现为:在做假设检验之前先指定一个  $\alpha$  的具体数值,通常取 0.05,也可以取 0.1、0.001 等表示较小的常用数值。

除了指定理论上的显著性水平之外,在 SAS 系统中,系统可以根据样本分布和样本数据自动计算出一个实际的显著性水平,通常称之为  $P$  值。 $P$  值也具体指在检验过程中实际犯第 I 类误差的概率。

当  $P$  值比  $\alpha$  小时,说明实际计算的显著性水平比理论的显著性水平更小,小概率事件在一次试验中发生的几率更小(比理论设定的概率还小)。此时在  $P$  值的显著性水平条件下,如果能够观测到小概率事件发生,则说明假设更加不可靠,可以对假设做出否定的判断;但是当  $P$  值大于  $\alpha$  时,在  $P$  值的显著性水平条件下,如果能够观测到小概率事件发生,说明假设可能没有任何问题,因为观测一个概率比较大的事件,其发生的可能性本来就比较大,故不能对假设做出否定的判断。因此,在 SAS 系统中进行假设检验,往往是从  $P$  值入手进行判定的。

## 2. SAS 系统中的假设检验基本步骤

假设检验的基本原理描述了检验的步骤。统计软件的检验过程与传统手工计算的检验过程有所区别,本部分主要介绍 SAS 系统中的假设检验基本步骤。

(1) 提出假设。没有假设,就没有检验的对象。假设是对总体特征的一个特定描述。

假设可以分为原假设和备择假设。原假设又被称为零假设,通常把想要搜集证据来否定的结论作为原假设,用  $H_0$  表示。而备择假设又被称为研究假设,通常把想要搜集证据来肯定或支持的结论作为备择假设,用  $H_1$  表示。

原假设和备择假设通常是对立、互斥的。原假设中的表达式通常包含“=”、“ $\geq$ ”、“ $\leq$ ”等含有等号的符号,而备择假设中的表达式通常包含“ $\neq$ ”、“ $<$ ”、“ $>$ ”等含有不等号的符号。

当备择假设含有“ $\neq$ ”时,称之为双测或双尾检验;当备择假设含有“ $<$ ”或“ $>$ ”时,称之为单侧或单尾检验。

例如,某灯泡厂生产一批灯泡,通过抽样调查得到 100 只该批灯泡的平均寿命是 6 000 小时,能否认为这批灯泡的总体平均寿命( $\mu$ )就是 6 000 小时呢?这需要进行假设检验。根

据这个问题提出的假设如下。

$H_0: \mu = 6\,000; H_1: \mu \neq 6\,000$

在 SAS 系统中，用 “Null” 表示原假设，用 “Alternate” 表示备择假设。

(2) 确定理论的显著性水平  $\alpha$ ，通常取 0.05，也可以取 0.1、0.001 等常用数值。

(3) 根据已知条件和总体分布状况，在原假设成立的情况下，选择计算用于检验的统计量。统计量的计算因检验对象而异，通常对总体均值或总体比例进行检验的统计量计算公式是：检验统计量=(点估计值-原假设成立时的参数值)/点估计值的标准误差。

(4) 将 SAS 计算出来的统计量的值对应的  $P$  值与理论的  $\alpha$  值对比。

在传统的手工检验过程中，这一步骤通常是根据  $\alpha$  的大小去查统计量对应的分布表，得到所谓的临界值 (SAS 中用 “Prob” 表示)，然后用计算出来的统计量值与临界值对比。如果统计量值在临界值之外，表示拒绝原假设，否则表示没有充分理由拒绝原假设。

但是在 SAS 软件中，不可能给用户一张分布表以查找其对应的值，SAS 系统给出的  $P$  值等同于 “临界值>统计量值” 的概率。可以依据以下判定法则对假设检验下结论。

如果  $P \leq \alpha$ ，说明在显著性水平  $\alpha$  条件下，原假设不成立，拒绝原假设，选择备择假设；如果  $P > \alpha$ ，说明在显著性水平  $\alpha$  条件下，没有充分证据表明应当拒绝原假设。

3. 假设检验中总体的几种不同情况

在进行统计推断之前，应当先搞清楚总体的分布情况，因为根据不同的总体分布情况计算的统计量形式不同。检验所用的统计量的形式和步骤取决于所抽取样本的样本量大小。无论大样本还是小样本，均假定其总体服从正态分布。

(1) 大样本情形下的检验方法。

大样本就是指样本量  $n \geq 30$  的样本。以总体均值的假设检验为例，在大样本的情况下，根据中心极限定理，均值的抽样分布服从正态分布，所以可以使用正态统计量 (即  $Z$  统计量) 进行假设检验。

(2) 小样本情形下的检验方法。

所谓小样本，即样本量  $n < 30$  的样本。对于小样本的情况，又可以细分两种情况。

- 当总体方差已知时，仍然使用正态统计量 ( $Z$  统计量)。
- 当总体方差未知时，则使用  $t$  统计量。 $t$  统计量服从自由度为  $(n-1)$  的学生分布 (即  $t$  分布)。

上述两种典型情况同样也适用于区间估计。

SAS 系统可以对总体均值、总体比例、总体方差进行假设检验。所有的检验过程均可以通过 “SAS/Analyst→Statistics→Hypothesis Tests” 菜单实现，如图 3-2 所示。

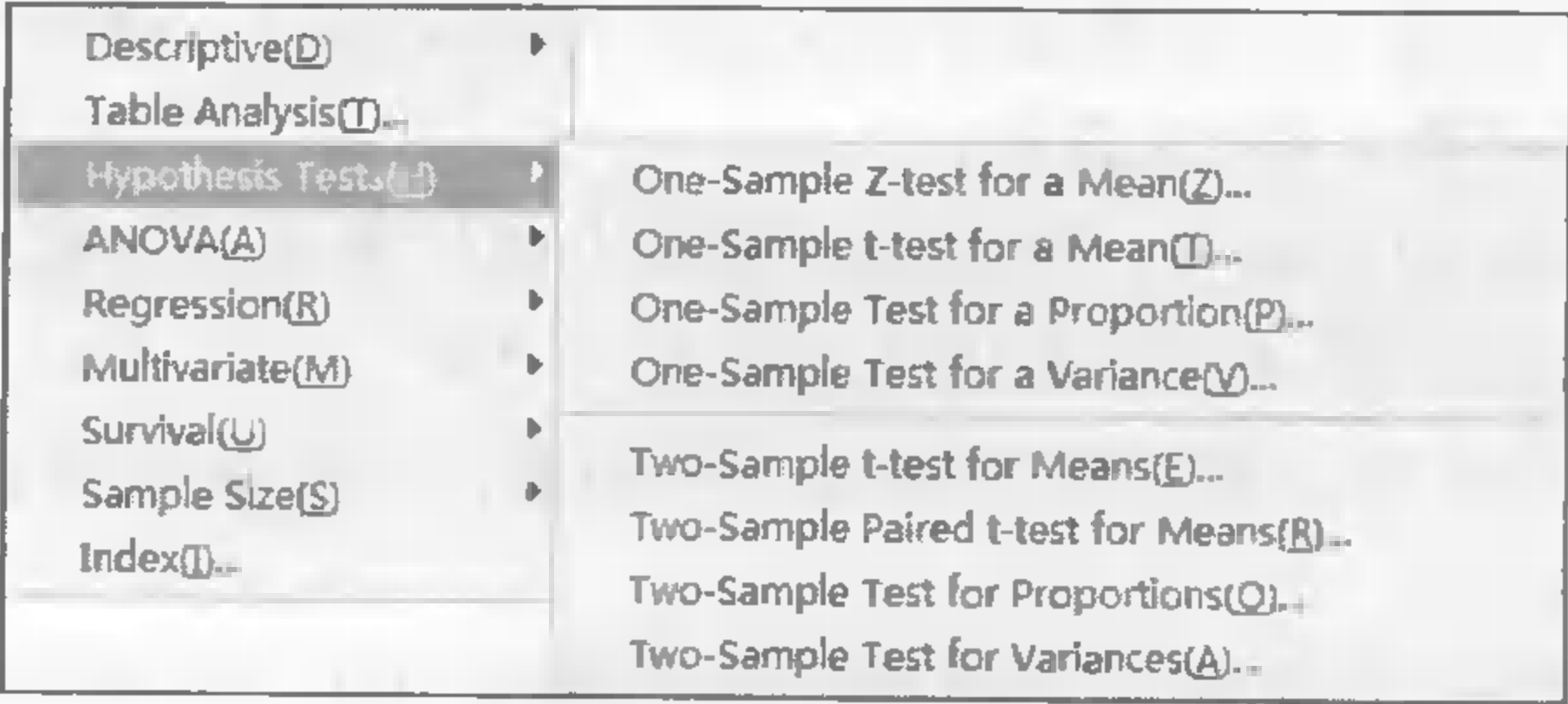


图 3-2 SAS/Analyst 的假设检验功能

图 3-2 所示的 SAS/Analyst 的主要假设检验功能如下。

- One-Sample Z-test for a Mean: 单总体均值的正态检验（大样本条件下，SAS 系统要求给定总体方差）。
- One-Sample t-test for a Mean: 单总体均值的  $t$  检验（小样本条件下）。
- One-Sample Test for a Proportion: 单总体比例的正态检验（比例的检验通常都在大样本下进行）。
- One-Sample Test for a Variance: 单总体方差的卡方（ $\chi^2$ ）检验。
- Two-Sample t-test for Means: 两总体均值之差（独立样本）的  $t$  检验。
- Two-Sample Paired t-test for Means: 两总体均值之差（成对、匹配样本）的  $t$  检验。
- Two-Sample Test for Proportions: 两总体比例之差的正态检验。
- Two-Sample Test for Variances: 两总体方差之比的  $F$  检验。

表 3-1 列出了 SAS 系统在各种情况下所用的检验统计量，以供分析参考。

表 3-1 常用统计推断的标准化统计量及其分布

菜 单	统计推断所用的标准化统计量	统计量的分布
One-Sample Z-test for a Mean	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	标准正态分布
One-Sample t-test for a Mean	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	自由度 $(n-1)$ 的 $t$ 分布
One-Sample Test for a Proportion	$z = \frac{p - \pi_0}{\sqrt{\frac{\pi_0(1 - \pi_0)}{n}}}$	标准正态分布
One-Sample Test for a Variance	$\chi^2 = (n-1)s^2/\sigma_0^2$	自由度 $(n-1)$ 的 $\chi^2$ 分布
Two-Sample t-test for Means	$t = \frac{(\bar{x}_1 - \bar{x}_2) - (\mu_1 - \mu_2)}{\sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2} \left( \frac{1}{n_1} + \frac{1}{n_2} \right)}}$	自由度 $(n_1 + n_2 - 2)$ 的 $t$ 分布
Two-Sample Paired t-test for Means	$t = \frac{\sum_{i=1}^n d_i / n_d - d_0}{\sqrt{\frac{\sum_{i=1}^n (d_i - \sum_{i=1}^n d_i / n_d)^2}{n_d - 1}}} / \sqrt{n_d}$	自由度 $(n-1)$ 的 $t$ 分布
Two-Sample Test for Proportions	$z = \frac{(p_1 - p_2) - d_0}{\sqrt{p_1(1 - p_1)/n_1 + p_2(1 - p_2)/n_2}}$	标准正态分布
Two-Sample Test for Variances	$F = s_1^2/s_2^2$	第一自由度 $(n_1-1)$ 、第二自由度 $(n_2-1)$ 的 $F$ 分布

其中， $\sigma$ 表示总体标准差， $n$  表示样本容量， $n_d$  表示成对样本的个数， $\bar{x}$ 、 $p$ 、 $s^2$  分别表示样本均值、样本比例和样本方差， $d_i$  表示成对样本的两组变量值之差， $\mu_0$ 、 $\pi_0$ 、 $\sigma_0^2$  分别表示原假设成立时的总体均值、总体比例和总体方差。

### 3.2 单总体参数的估计及假设检验

单总体参数估计和假设检验比较简单。点估计实际上就是计算统计量的问题，第 2 章已经对统计量的计算进行了详细讲解。如要进行点估计，只需把计算的统计量作为总体参数的描述即可。本节重点讲述总体参数的估计和假设检验。

#### 3.2.1 单总体的参数估计

单总体点估计在 SAS 系统中实际上就是计算对应参数的样本统计量，详见第 2 章中描述统计分析的相关内容，本小节不再赘述。

区间估计，顾名思义是在估计总体特征时给出一个置信区间。置信区间在其他条件不变的情况下，取决于置信度的水平，而在 3.1.2 小节的分析中，可以得出显著性水平 $\alpha=1-\text{置信度}$ 的结论。在 SAS 系统中，置信区间的置信度是由 $\alpha$ 控制的，因此只要给定系统一个确定的、理论显著性水平 $\alpha$ 的值，就可以估计出置信区间。

在实际应用中，SAS 可以对单总体的均值、方差、标准差、比例进行参数估计。

##### 1. 单总体均值的参数估计

对单总体的均值进行区间估计，主要是指在一定的置信度下对总体均值进行估计，被广泛应用于社会经济领域。



**例 3-1**

某糖果生产商新开发了一种新型的饼干。如果饼干水分超标，就容易促使细菌繁殖，油脂发生氧化，从而严重缩短产品的实际保质期。因此，国家对饼干中的水分含量有严格限定，即水分含量不得超过 4.0%。为了检测水分含量，有关工作人员随机抽取该生产商生产的 50 块规格为 100 克/块的饼干，进行了水分含量测试，具体测试数据（详见 Moisture.sas7bdat）如表 3-2 所示。为了达到良好的测试效果，需要对该厂生产的所有该型号饼干的水分含量进行置信度为 95%的区间估计。

表 3-2 糖果水分（Moisture）测试结果（单位：g）

4.50	3.56	4.09	3.64	3.07
3.50	3.89	4.75	4.11	3.68
3.55	3.86	4.49	4.15	3.97
4.03	3.34	4.51	4.57	3.45
3.19	3.93	3.72	3.57	4.04
3.95	4.43	3.49	4.13	4.26
3.66	4.62	4.36	3.86	3.85
4.18	3.74	4.42	4.43	3.72
4.32	3.52	4.07	3.48	4.28
4.87	3.79	4.18	3.48	4.43

**STEP 1)** 进入 SAS/Insight，打开 Mosisture.sas7bdat 数据集。

**STEP 2** 单击 Analyze→Distribution，弹出“Distribution”对话框，可以进行区间估计选择。在该对话框中，选择“Moisture”变量，单击“Y”按钮，把该变量设置为分析变量。然后单击对话框下部的“Output”按钮，弹出输出结果控制对话框，如图 3-3 所示。

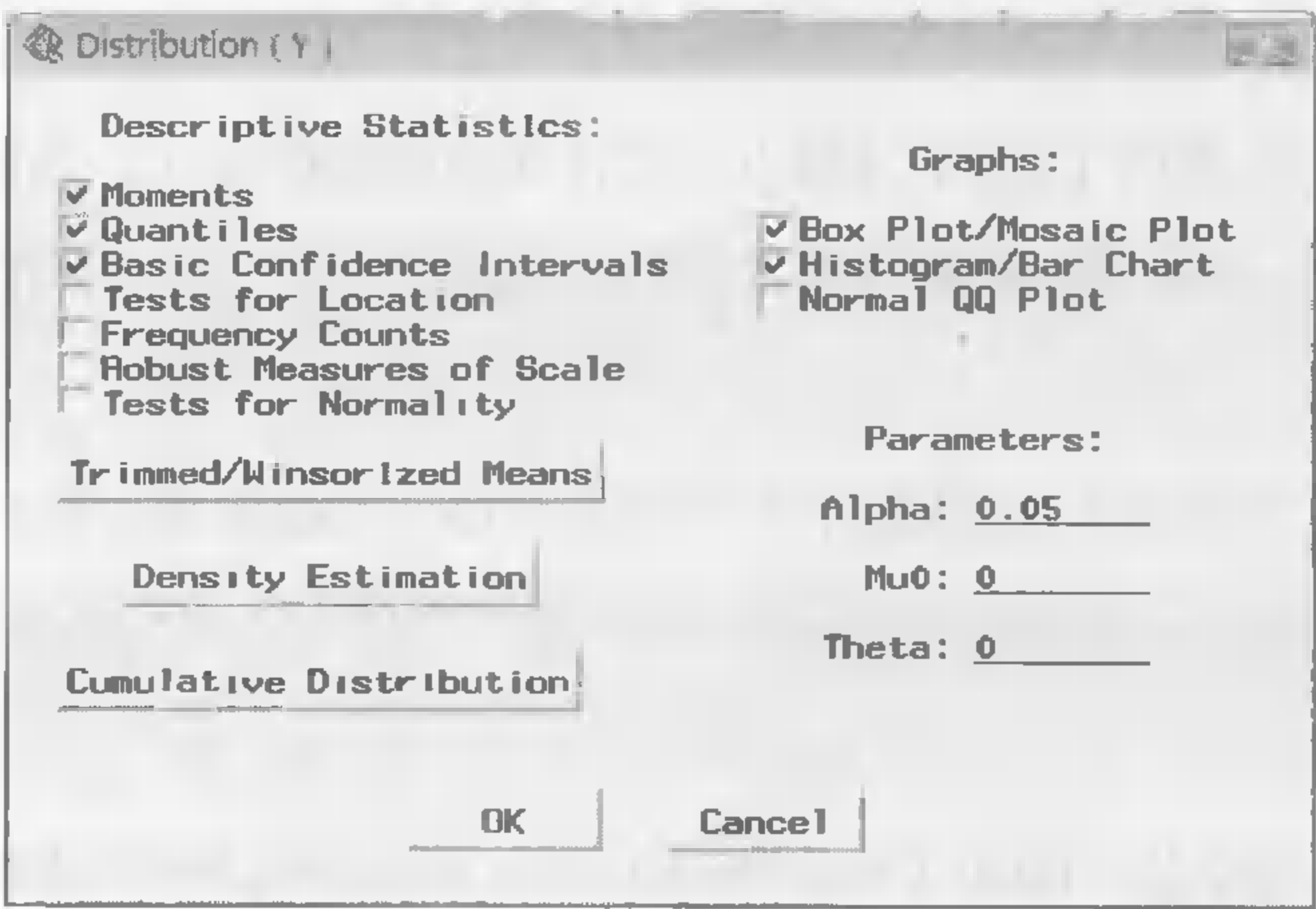


图 3-3 Distribution 输出结果设置对话框

**STEP 3** 在“Descriptive Statistics”分栏下，选中“Basic Confidence Intervals”复选框，即可按照默认的设置进行区间估计。在“Parameters”分栏下，可以设置区间估计的置信度。置信度是由理论显著性水平 $\alpha$ 控制的，在“Alpha”文本输入框中输入指定的 $\alpha$ 值即可，本例中置信度为 95%，因此输入“ $\alpha=0.05$ ”。如果想以 99%的置信度进行区间估计，只需把“Alpha”的值改成“0.01”即可。显著性水平设置好后，单击“OK”按钮返回“Distribution”对话框。在该对话框中，单击“OK”按钮，区间估计的结果便出来了，如图 3-4 所示。

在图 3-4 中，“95% Confidence Intervals”表示 95%置信度下的置信区间，一共可以对 3 个总体参数进行区间估计，分别是均值(Mean)、标准差(Std Dev)、方差(Variance)。“Estimate”列示对应总体参数的点估计值，“LCL”则表示置信下限，“UCL”表示置信上限。如在本例中，饼干水分含量总体均值的点估计是 3.9736，95%的置信区间是[3.8531, 4.0941]。

此外，在 SAS/Insight 的结果输出窗口中，也可以直接进行不同置信度的区间估计。切换至刚才得到的“Distribution”结果窗口，然后单击系统菜单上的“Table”，弹出“Table”菜单，如图 3-5 所示。

95% Confidence Intervals			
Parameter	Estimate	LCL	UCL
Mean	3.9736	3.8531	4.0941
Std Dev	0.4239	0.3541	0.5282
Variance	0.1797	0.1254	0.2790

图 3-4 单总体的区间估计的结果

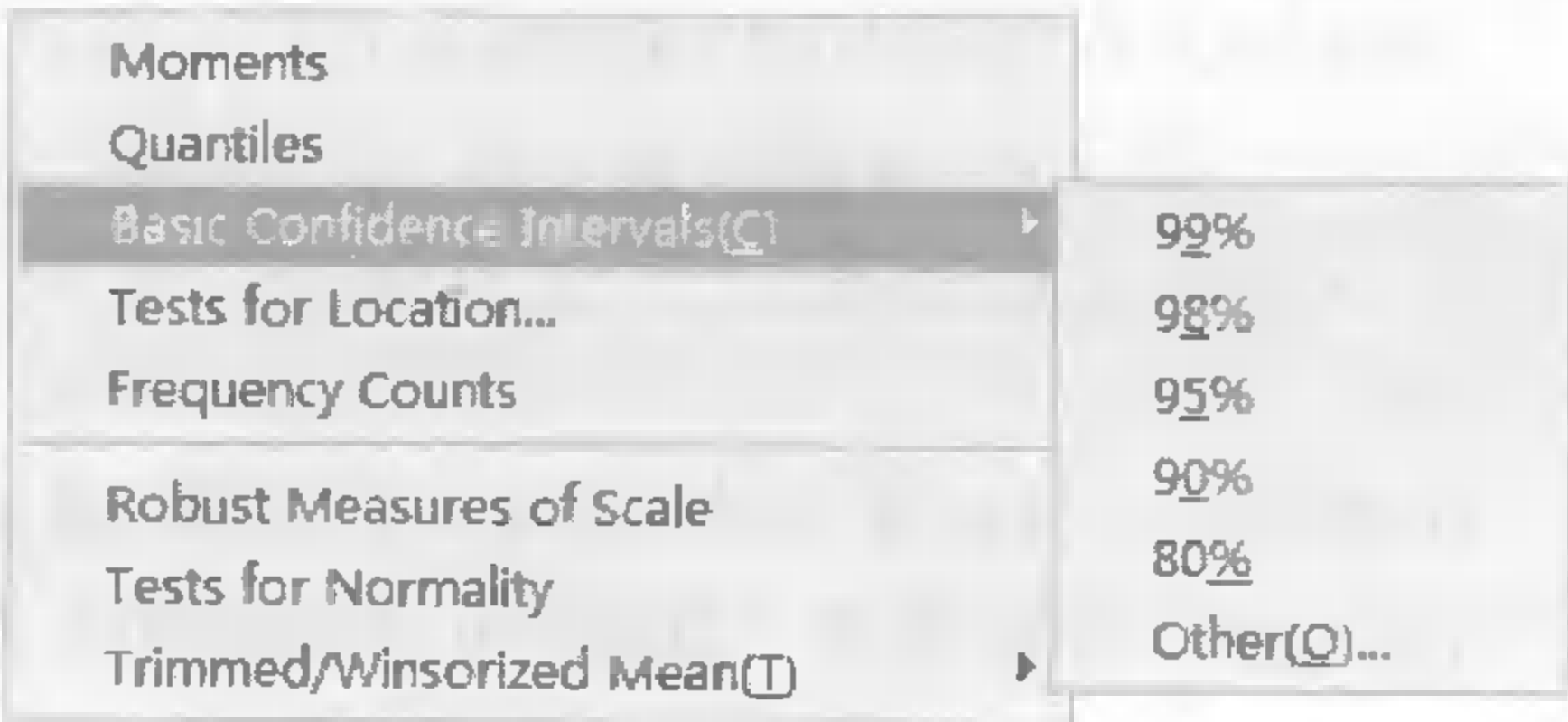


图 3-5 “Distribution”的“Table”菜单

在“Table”菜单中，用户直接单击对应的置信度，就可以在结果窗口中看到区间估计在该置信度下的区间估计结果。此外，还可以通过该菜单下的“Other”选项指定置信度。

利用 SAS 程序的 MEANS 过程可以直接估计均值的置信区间，具体程序如下。

```
proc means data=Sasuser.Moisture
  clm  alpha=0.05;      /*关键字 clm 用于计算置信区间，关键字 “alpha=” 用于指定显著性水平*/
  var  moisture;
run;
```

同样，利用 SAS 程序的 TTEST 过程也可估计均值的置信区间，具体程序如下。

```
proc ttest data=Sasuser.Moisture alpha=0.05; /*关键字 “alpha=” 用于指定显著性水平*/
  var moisture;
run;
```

此外，利用 SAS 程序的 UNIVARIATE 过程也可估计均值的置信区间，具体程序如下。

```
proc univariate data=Sasuser.Moisture cibasic(alpha=0.05); /*关键字 “cibasic(alpha=)” 指定显著性水平*/
  var moisture;
run;
```

在该程序运行结果中的“Basic Confidence Limits Assuming Normality”表格中可看到均值“Mean”的置信区间。

## 2. 单总体方差、标准差的区间估计

单总体方差、标准差的区间估计原理和方法与单总体均值的估计原理方法相同，唯一的区别在于所用的统计量服从卡方分布。在 SAS 系统中，对总体方差、标准差的具体操作步骤与总体均值参数估计的操作步骤相同。

如以例 3-1 为例，要求在给定的置信度下，估计水分含量的总体方差和标准差的区间估计，按照图 3-3 所示的操作步骤，同样可得到图 3-4 所示的结果，即总体方差的点估计值为 0.1797，95%置信度下的区间估计为[0.1254, 0.2790]；总体标准差的点估计值为 0.4239，95%置信度下的置信区间为[0.3541, 0.5282]。同样利用与均值区间估计相同的程序可得到总体方差、标准差的置信区间，程序如下。

利用 SAS 程序的 TTEST 过程估计总体方差、标准差的置信区间。

```
proc ttest data=Sasuser.Moisture alpha=0.05; /*关键字 “alpha=” 用于指定显著性水平*/
  var moisture;
run;
```

利用 SAS 程序的 UNIVARIATE 过程，估计总体方差、标准差的置信区间，具体程序为：

```
proc univariate data=Sasuser.Moisture cibasic(alpha=0.05); /*关键字 “cibasic(alpha=)” 指定显著性水平*/
  var moisture;
run;
```

在该程序运行结果中的“Basic Confidence Limits Assuming Normality”表格中可看到方差“Variance”和标准差“Standard Deviation”的置信区间。

## 3. 单总体比例的参数估计

这里所说的比例是指在反映总体某种特征的变量只有两种属性的情况下，其中某种属性占有所有属性的比重或百分比。如全社会男性人口的比例，反映全社会人口性别特征的“性别”

变量只有两种可能：“男”或“女”。如果要研究总人口中的男性比例，可以根据抽样调查的数据结果进行总体男性人口比例的参数估计。又如收集人们对某项政策实施的“支持”和“反对”看法的样本数据，可以对全社会关于该项政策实施的态度进行参数估计。

考察比例的样本数据通常为大样本，可利用大样本条件下的统计量进行统计推断。



例 3-2

某厂家生产了一批产品。为了检验该批次产品的合格率，从这批产品中随机抽取了 100 个产品进行检验（详见 Quality.sas7bdat），得到的检验结果为：合格品 96 个，不合格品 4 个。试以 99% 的置信度估计该厂家该批次产品合格率的置信区间。

Quality.sas7bdat 数据集中的变量名为“Product”有两个属性，即两个值，分别是“合格”和“不合格”。

SAS 过程处理这种定性变量比较麻烦。因此，对于总体比例的参数估计，可以通过 SAS 编程语言中的 SURVEYMEANS 过程进行计算，具体程序如下。

```
proc surveymeans data=Sasuser.Quality
  mean clm alpha=0.01; /*关键字 mean 表示计算两种属性的比例，关键字 clm 表示估计变量两种属性的置信区间，关键字“alpha=”表示置信度*/
  var product;
run;
```

运行 SAS 程序，得到图 3-6 所示的结果。

Statistics					
Variable	Level	Label	Mean	Std Error of Mean	99% CL for Mean
Product	不合格	产品质量	0.040000	0.019695	0.00000000 0.09172611
	合格	产品质量	0.960000	0.019695	0.90827389 1.00000000

图 3-6 总体比例的参数估计结果

在图 3-6 所示的输出结果中，“Product”变量的两个属性即“不合格”与“合格”的均值（Mean）为 0.04 和 0.96，分别表示产品的不合格率与产品的合格率，为这两个属性值的点估计值。在图 3-6 中，很容易找到 99% 置信度下该批次全体产品合格率的置信区间（99% CL for Mean）为[0.90827389, 1]。

3.2.2 单总体参数的假设检验

单总体参数的假设检验问题是现实生活中十分常见的问题。如例 3-1 中，利用随机抽样得到的样本，根据计算出来的统计量的值能否对产品质量进行判定呢？答案是肯定的，完全可以按照假设检验的基本原理和步骤对实际问题进行检验。

1. 总体均值的假设检验

(1) 大样本和总体方差已知，大样本和总体方差未知或小样本和总体方差已知。

在大样本（样本量≥30）情形下，通常使用正态统计量（Z 统计量）进行假设检验。如果总体方差未知，则用样本方差代替。在小样本（样本量<30）情形下，如果总体方差已知，也可使用正态统计量（Z 统计量）进行假设检验。



例 3-3

饼干水分超标容易促使细菌繁殖，油脂发生氧化，从而严重缩短产品的实际保质期，国家对饼干中的水分含量有严格限定，即水分含量不得超过 4.0%。为了检测水分含量，有关工作人员随机抽取某生产商生产的 50 块某批次规格为 100 克/块的饼干，进行了水分含量测试，具体测试数据（详见 Moisture. sas7bdat）如表 3-2 所示。经过测试，抽样得到的水分含量样本均值为 3.97 克，能否认为该生产厂商该批次的饼干符合国家要求？（已知总体方差为 0.2，设显著性水平  $\alpha=0.05$ ）

该例是例 3-1 的延伸，要研究的是某一批次所有饼干水分含量这个总体的情况。国家规定的水分含量是不得超过 4%，该批次产品的规格是 100 克/块，因此对于整批产品而言，如果是合格产品的话，其总体的平均含水量不应超过  $100 \times 4\% = 4$  克。而得到的样本平均含水量为 3.97 克，貌似小于国家标准，那么是否能直接认为该批饼干就是符合国家标准呢？这个问题需要用统计学的方法进行检验。

因此，根据样本数据以及出于研究目的，设总体水分含量的均值为  $\mu$ ，提出该问题的原假设和备择假设如下。

$H_0: \mu \leq 4; H_1: \mu > 4$

接下来，可以利用 SAS 系统进行假设检验。

**STEP 1)** 在 SAS 系统中，利用 “SAS/Analyst→Statistics→Hypothesis Tests” 进行假设检验。进入 SAS/Analyze 后，打开 Mosisture.sas7bdat 数据集。

**STEP 2)** 本例中的样本量为 50，属于大样本，因此在 Hypothesis Tests 的二级菜单（如图 3-2 所示）中选择 “One-Sample Z-test for a Mean”，弹出单总体均值正态检验对话框，如图 3-7 所示。

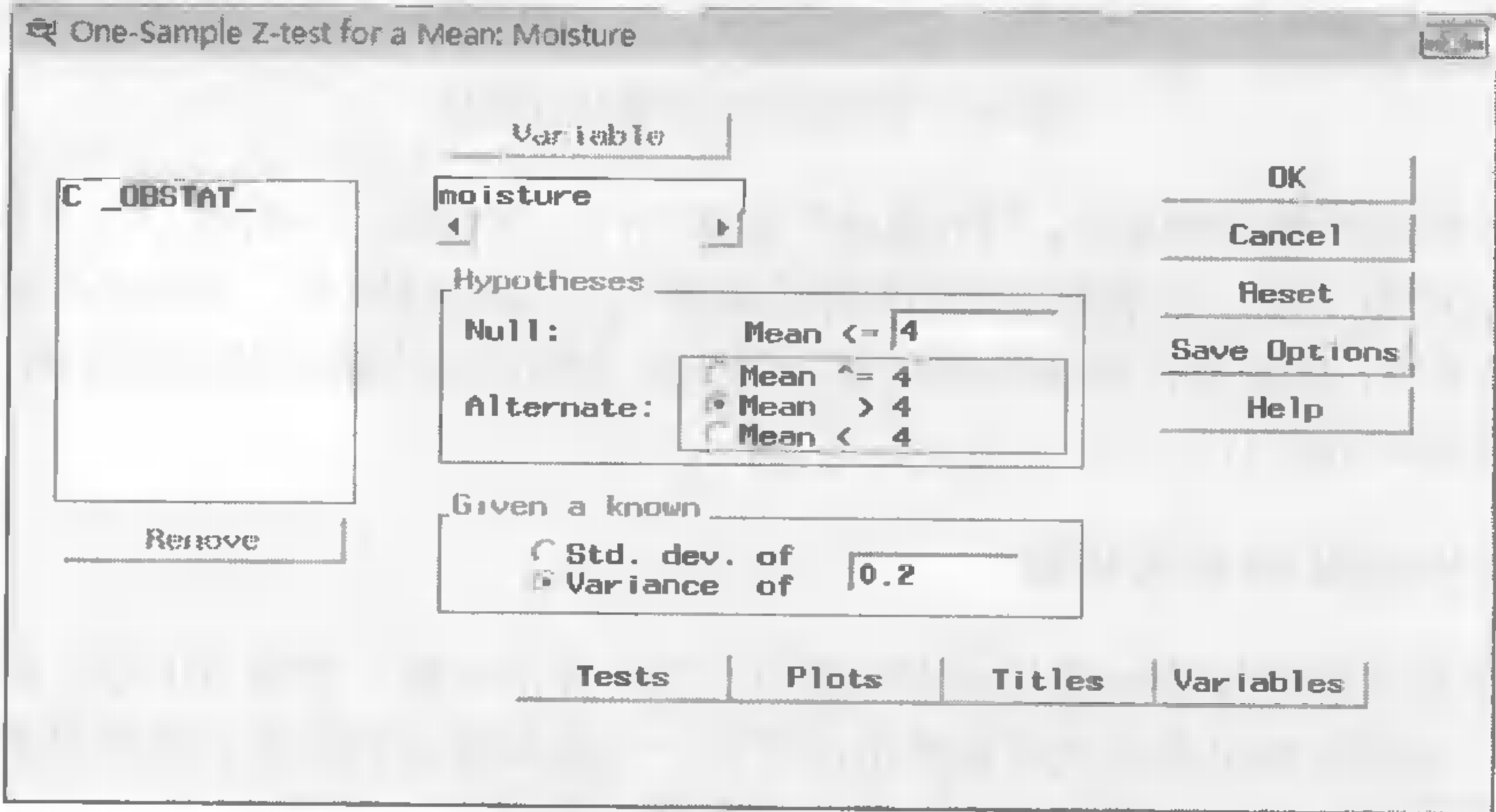


图 3-7 单总体均值正态（Z）检验对话框

**STEP 3)** 图 3-7 所示对话框中上部区域列示了数据集中的所有变量。选中 “moisture” 变量，单击 “Variable” 按钮，把其指定为分析变量。在 “Hypotheses” 分栏下，可以对原假设和备择假设进行设定，“Null” 表示原假设，“Alternate” 表示备择假设。本例中将要总体均值 “4” 进行检验，因此，先在原假设 “Null” 的文本输入框中输入 “4”，并根据本例的备择假设在 “Alternate” 的单选框中选择 “Mean>4”，则在 “Null:” 区域系统自动变为 “Mean<=4”，

表示本例所设定的原假设和备择假设。

**STEP 4** 在“Given a know”的文本输入框中，可以选择输入给定的总体标准差“Std. Dev. of”或给定的总体方差“Variance of”。在本例中，已经给定了总体方差 0.2，所以选择“Variance of”单选框，输入“0.2”。如果本例没有给出总体方差，即总体方差未知，则可以用样本方差代替，本例的样本方差计算结果为 0.1764，即把“0.1764”填入“Variance of”的文本输入框中。

**STEP 5** 此外在进行假设检验的同时，也可以进行参数估计。单击“Tests”按钮，弹出“Tests”对话框，如图 3-8 所示。

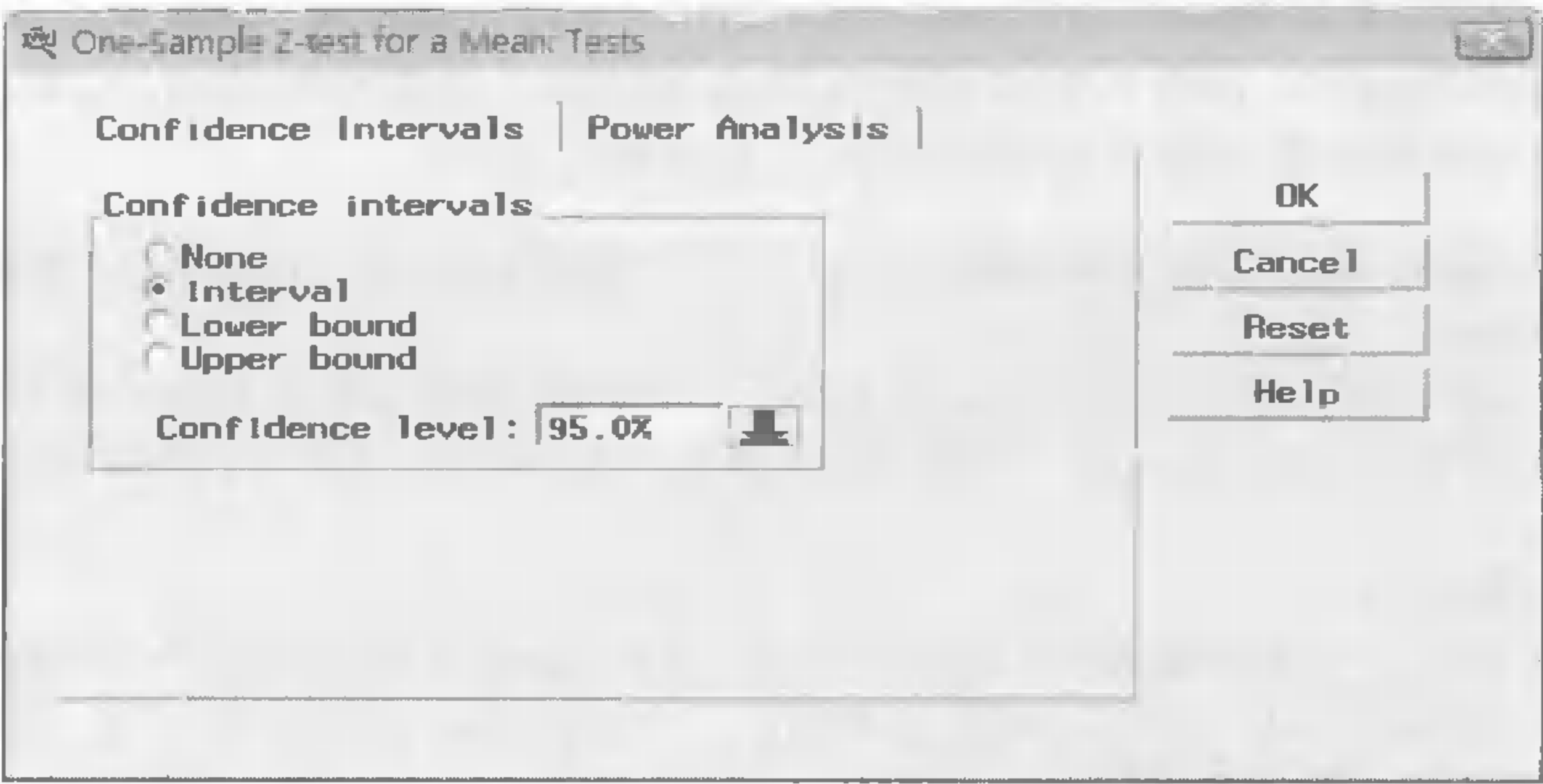


图 3-8 单总体均值假设检验中的“Tests”对话框

**STEP 6** 选择“Confidence Intervals”选项卡，在“Confidence Intervals”中，可以选择不同的单选框以输出不同的区间估计形式。系统默认选择不输出区间估计的结果，即“None”单选框。“Interval”表示输出具体的估计置信区间，“Lower bound”表示只输出置信区间的下限，“Upper bound”表示只输出置信区间的上限。同时可以在“Confidence level”文本输入框中设定区间估计的置信度。如本例选择输出 95%置信度下的置信区间，即选中“Interval”单选框，在“Confidence level”文本输入框中输入 95%。

**STEP 7** 单击“OK”按钮返回假设检验对话框。在该对话框中，单击“OK”按钮，可以得到本例的假设检验输出结果，如图 3-9 所示。

Sample Statistics for moisture			
N	Mean	Std. Dev.	Std. Error
50	3.97	0.42	0.06
Hypothesis Test			
Null hypothesis:		Mean of moisture $\leq 4$	
Alternative:		Mean of moisture $> 4$	
With a specified known variance of 0.2			
Z Statistic		Prob > Z	
-0.417		0.6618	
95% Confidence Interval for the Mean			
Lower Limit		Upper Limit	
3.85		4.10	

图 3-9 大样本的单总体均值假设检验结果

图 3-9 中的头三行列示了样本数据的一些基本特征。“Hypothesis Test”下列示了假设检验的过程和结果。“Null hypothesis”表示原假设是总体均值 $\mu \leq 4$ ，“Alternative”表示备择假设是总体均值 $\mu > 4$ 。因为进行的是正态检验，所以系统输出的正态统计量即 Z Statistic 的值是-0.417。该值是根据样本数据以及均值抽样分布服从正态分布的情况下计算出来。它对应的 P 值即临界值“Prob”大于 Z 值的概率为 0.6618。这个数值表明，在给定的理论显著性水平 $\alpha = 0.05$ 的条件下，P 值为 0.6618，远远大于 0.05，因此没有充分的证据表明应当拒绝原假设，即没有理由认为该生产厂商生产的该批次饼干是不合格的。此外，区间估计的结果与 SAS/Insight 中的估计结果相同，如图 3-4 所示。

在 SAS 编程语言中，没有可供直接进行单总体参数正态检验的过程，但是可以根据样本数据和正态分布函数计算 Z 统计量值与 P 值，具体程序如下。

```
proc means noprint data=Sasuser.Moisture;           /*利用 MEANS 过程得到样本量和样本均值*/
  var moisture;
  output out=work.temp n=_nobs_ mean=_mean_;      /*把得到样本量和样本均值存放在一个名为
“temp”的临时数据集中，其中“_nobs_”变量表示样本量，“_mean_”变量表示样本均值*/
run;
data work.temp1;
  set work.temp;  /*根据 temp 数据集复制一个新的、名为“temp1”的数据集，用来计算统计量和 P 值*/
  Z_Value = (_mean_ - 4) / (sqrt(0.2)/sqrt(_nobs_)); /*根据已知的总体方差“0.2”和原假设成立情况
下的总体均值“4”，计算用于检验的 Z 统计量的值*/
  P_Value = 1-probnorm(Z_Value);                  /*根据 Z 统计量和正态分布函数计算 P 值*/
  put 'Hypothesis Test:';                          /*在“Log”窗口中输出假设检验的结果*/
  put 'Null hypothesis:' 'Mean of moisture <= 4';
  put 'Alternative:' 'Mean of moisture > 4';
  put 'With a specified known variance of 0.2';
  put 'Z Statistic:' Z_Value 'Prob > Z:' P_Value;
run;
```

 例 3-4

一种零件的生产标准是：直径为 10cm，标准容许的直径方差为 0.5。为对生产过程进行控制，质量监测人员定期对一台加工机床进行检查，以确定这台机床生产的零件是否符合标准要求。如果零件的平均直径大于或小于 10cm，则表明生产过程不正常，必须进行调整。现经过随机测试，得到该机床生产的 20 个同样零件的直径（变量名为 Diameter）测试结果（详见 Ware.sas7bdat）如表 3-3 所示。试对该台机床的生产过程是否正常做出判定，设显著性水平 $\alpha = 0.01$ 。

表 3-3 20 个零件的直径

9.72	9.93	9.35	9.87
8.21	9.03	9.96	9.73
8.16	10.93	8.72	8.77
9.66	9.08	9.52	11.00
8.99	10.48	10.75	8.95

在本例中，如果该台机床的生产过程正常的情况下，其生产的所有零件在容许方差为 0.5

的情形下的平均直径 $\mu$ 应当是 10cm。据此，可以提出该问题的原假设和备择假设如下。

$$H_0: \mu = 10; H_1: \mu \neq 10$$

原假设表示该台机床生产过程正常，备择假设表示该台机床生产过程不正常。

**STEP 1)** 进入 SAS/Analyst 后，打开 Ware.sas7bdat 数据集，然后选择“Statistics → Hypothesis Tests”进行假设检验。

**STEP 2)** 本例中的样本量为 20，属于小样本，但标准容许的直径方差为 0.5，即总体方差已知，故仍可以利用 SAS 系统进行正态假设检验。因此，在 Hypothesis Tests 的二级菜单（如图 3-2 所示）中选择“One-Sample Z-test for a Mean”，弹出图 3-7 所示的单总体均值正态检验对话框。

**STEP 3)** 选中“Diameter”变量，单击“Variable”按钮，把其指定为分析变量。在原假设“Null”的文本输入框中输入“10”，并根据本例的备择假设在“Alternate”的单选框中选择“Mean<sup>^</sup> = 10”，表示本例所设定的原假设和备择假设。

**STEP 4)** 在“Given a know”的文本输入框中，输入给定的总体方差“Variance of”为 0.5，单击“OK”按钮即可得到假设检验的结果，（如需进行参数估计，同样可以单击“Tests”按钮，如例 3-3 一样进行设定。）如图 3-10 所示。

Sample Statistics for diameter			
N	Mean	Std. Dev.	Std. Error
20	9.54	0.83	0.18
Hypothesis Test			
Null hypothesis: Mean of diameter = 10			
Alternative: Mean of diameter <sup>^</sup> = 10			
With a specified known variance of 0.5			
Z Statistic		Prob > Z	
-2.912		0.0036	

图 3-10 小样本总体方差已知的单总体均值假设检验结果

在图 3-10 所示的结果中，得到该问题检验结果的  $P$  值（即“Prob>Z”）为 0.0036， $P$  值远远小于  $\alpha$  (0.01)。故根据假设检验的原理可知，在  $\alpha = 0.01$  的显著性水平下，拒绝原假设，即该台机床的生产过程不正常。

该例的具体程序如下。

```
proc means noprint data=Sasuser.Ware; /*利用 MEANS 过程得到样本量和样本均值*/
var diameter;
output out=work.temp n=_nobs_ mean=_mean_; /*把得到样本量和样本均值存放在一个名为
“temp”的临时数据集中，其中“_nobs_”变量表示样本量，“_mean_”变量表示样本均值*/
run;
data work.temp1;
set work.temp; /*根据 temp 数据集复制一个新的、名为“temp1”的数据集，用来计算统计量和 P 值*/
Z_Value = (_mean_ - 10) / (sqrt(0.5)/sqrt(_nobs_)); /*根据已知的总体方差“0.5”和原假设成立情况
下的总体均值“10”，计算用于检验的 Z 统计量的值*/
P_Value = (1-probnorm(abs(Z_Value)))*2; /*根据 Z 统计量和正态分布函数计算 P 值*/
put 'Hypothesis Test:'; /*在“Log”窗口中输出假设检验的结果*/
put 'Null hypothesis:' 'Mean of moisture = 10';
```

```
put 'Alternative:' 'Mean of moisture ^= 10%';
put 'With a specified known variance of 0.5%';
put 'Z Statistic:' Z_Value 'Prob > Z:' P_Value;
run;
```

例 3-4 与例 3-3 的  $P$  值计算公式不同。例 3-3 中的备择假设是“ $\mu > 4$ ”，表示单侧检验，备择假设有“ $>$ ”或“ $<$ ”符号，据此计算出来的  $P$  值被称为单侧  $P$  值；而例 3-4 中的备择假设是“ $\neq$ ”符号的，据此计算出来的  $P$  值被称为双侧  $P$  值。双侧  $P$  值和单侧  $P$  值可以通过公式进行互相推算。

(2) 小样本、总体方差未知。

在小样本（样本量 $<30$ ）情形下，如果总体方差未知，通常使用  $t$  统计量进行假设检验。



例 3-5

某省移动通信公司对其用户进行满意度评估。公司经理根据近期业务发展状况、消费者情况和专家意见，判定该省公司管辖范围内的数据业务用户满意度的评估值应该超过 82 分。为了验证该评估值，委托市场研究公司对该省数据业务用户进行了小规模调查，得到表 3-4 所示的 25 个用户评价满意度（变量名为 Csi\_data）的得分（数据详见 Mobile.sas7bdat）。试在显著性水平  $\alpha = 0.05$  条件下，对该公司数据业务评估值进行检验。

表 3-4 某省移动通信公司数据业务用户满意度得分

76	85	73	77	77
84	72	88	89	83
86	74	72	74	85
90	85	90	78	83
84	83	83	84	75

本例的样本量为 25，没有给定总体方差，因此可以使用  $t$  统计量进行假设检验。

根据样本计算出的满意度平均分为 81.2 分，小于公司经理的评估值 82 分。那么这 0.8 分的差距究竟是由调查中的随机因素造成的，还是总体满意度平均值就是小于 82 分呢？此外，市场研究公司往往是基于客户委托的目的和意愿进行研究的，因此对于这个问题，可对研究目的即总体满意度  $\mu$  大于 82 分进行验证，提出原假设和备择假设如下。

$H_0: \mu \leq 82; H_1: \mu > 82$

原假设表示总体满意度评价不超过 82 分，备择假设表示支持公司经理的结论，即总体满意度超过 82 分。

**STEP 1)** 进入 SAS/Analyst，打开 Mobile.sas7bdat 数据集，利用“Statistics→Hypothesis Tests”进行假设检验。

**STEP 2)** 本例中的样本量为 25，属于小样本，总体方差未知，可利用 SAS 系统进行  $t$  检验。因此，在 Hypothesis Tests 的二级菜单(如图 3-2 所示)中选择“One-Sample t-test for a Mean”，弹出单总体均值  $t$  检验对话框，如图 3-11 所示。

**STEP 3)** 选中“csi\_data”变量，单击“Variable”按钮，把其指定为分析变量。在“Hypotheses”分栏下的原假设“Null”的文本输入框中输入“82”，并根据本例的备择假设在“Alternate”的单选框中选择“Mean>82”，表示本例所设定的原假设和备择假设。

同样，单击“Tests”按钮可对变量进行区间估计。单击“OK”按钮便可得到图 3-12 所示的检验结果。

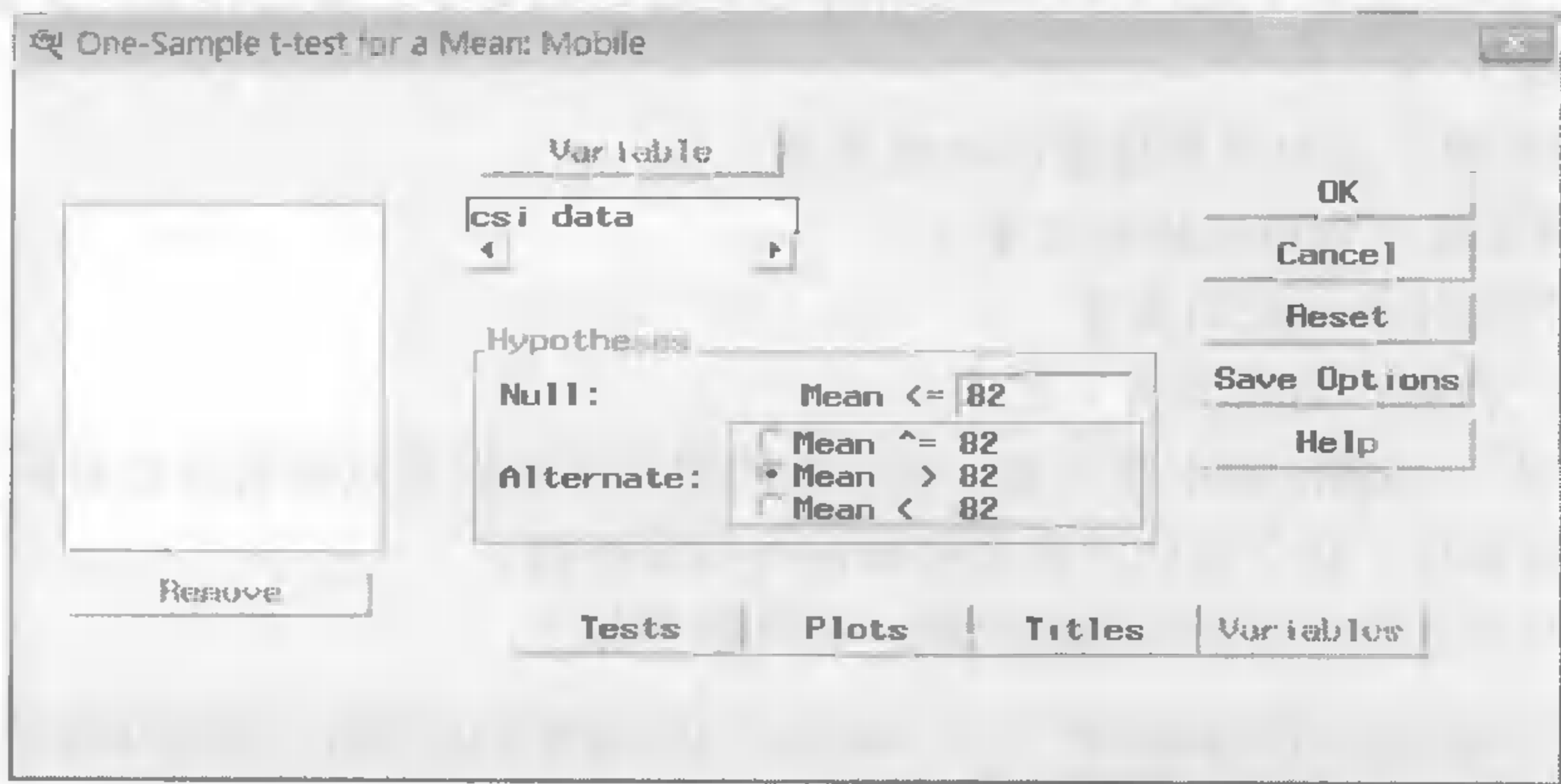


图 3-11 小样本的总体均值  $t$  检验对话框

Sample Statistics for csi_data			
N	Mean	Std. Dev.	Std. Error
25	81.20	5.83	1.17
Hypothesis Test			
Null hypothesis: Mean of csi_data <= 82			
Alternative: Mean of csi_data > 82			
t Statistic	Df	Prob > t	
-0.686	24	0.7504	

图 3-12 小样本总体均值  $t$  检验结果

在检验结果中，可以看到  $t$  统计量的值为-0.686，其对应的自由度（Df）为 24， $P$  值等于 0.7504，远远大于  $\alpha$  (0.05)。因此，在显著性水平  $\alpha = 0.05$  的条件下，没有充分证据表明应当拒绝原假设，即没有理由认为该省移动公司数据业务的用户总体评价大于 82 分，故该公司经理的评估值有问题。

利用 SAS 语言中的 TTEST 过程可以对上述问题进行  $t$  检验，TTEST 过程的具体语法如下。

```
proc ttest <选项>;
  class 变量;
  paired 变量;
  by 变量;
  var 变量;
  freq 变量;
  weight 变量;
```

其中，TTEST 可以对以下与假设检验有关的常用选项进行调整。

- Alpha =  $p$ : 指定用于区间估计的置信度对应的显著性水平，如 Alpha = 0.01 表示以 99% 置信度进行区间估计。在默认情况下， $p = 0.05$ 。
- $H_0 = m$ : 指定原假设成立条件下的总体均值。

各语句的功能如下。

- **class**: 指定对两个独立样本的总体参数检验时的分类变量，单总体检验可以忽略分类变量。
- **paired**: 指定对两个成对样本的总体参数检验时的变量，成对样本的变量之间用 “\*” 号隔开。
- **by**: 指定用于分别进行检验的分组变量。
- **var**: 指定进行分析检验的变量。
- **freq**: 指定作为频数的变量。
- **weight**: 指定作为权数的变量。

TTEST 过程可以输出样本统计量、统计量的置信区间以及双侧假设检验的  $t$  统计量值及  $P$  值，也可以对来自于两个总体的样本数据进行假设检验。

例 3-5 是对单总体均值进行假设检验，具体程序如下：

```
proc ttest data=Sasuser.Mobile h0=82; /* “h0=82” 表示原假设成立条件下的总体均值为 82*/
var csi_data;
run;
```

运行得到图 3-13 所示的结果。

TTEST 过程的假设检验结果得出的  $P$  值是双侧检验的  $P$  值，即备择假设取 “ $\neq$ ” 号时计算出来的  $P$  值，可以直接用该数值与理论显著性水平  $\alpha$  对比；如果备择假设取 “ $<$ ” 或 “ $>$ ” 符号，则单侧  $P$  值应当按照以下原则计算。

T-Tests			
Variable	DF	t Value	Pr >  t
csi_data	24	-0.69	0.4993

图 3-13 Ttest 过程的假设检验结果

- 如果备择假设取 “ $<$ ” 符号，则
- 当  $t \geq 0$  时，进行判定的单侧  $P$  值为  $1-(Pr>|t|)/2$ 。
  - 当  $t < 0$  时，进行判定的单侧  $P$  值为  $(Pr>|t|)/2$ 。

- 如果备择假设取 “ $>$ ” 符号，则
- 当  $t \geq 0$  时，进行判定的单侧  $P$  值为  $(Pr>|t|)/2$ 。
  - 当  $t < 0$  时，进行判定的单侧  $P$  值为  $1-(Pr>|t|)/2$ 。

在例 3-5 中，备择假设是 “ $H_1: \mu > 82$ ”， $t$  统计量的值为 -0.69，故进行判定的单侧  $P$  值为  $1-0.4993/2 = 0.75035$ ，与图 3-12 所示结果一致。

此外，利用 UNIVARIATE 过程也可以得到上述假设检验的过程和结果，具体程序如下。

```
proc univariate data=Sasuser.Mobile mu0=82; /*mu0 指定原假设成立条件下的总体均值为 82*/
var csi_data;
run;
```

UNIVARIATE 过程的结果同样输出双侧检验的 “ $Pr>|t|$ ” 值。在单侧检验情况下， $P$  值的计算方法与 TTEST 过程的计算方法一致。

2. 总体比例的假设检验

总体比例的假设检验是指根据样本数据对总体具备某种属性的个体总数占全体属性总数的比例提出假设并进行检验的过程，通常在大样本的条件下进行。本部分内容以例 3-2 的数据为基础进行假设检验。



例 3-6

某厂家生产了一批产品，根据相关标准规定，该种产品的合格率应当大于 97%。为了检验该批次产品的合格率，从其中随机抽取了 100 个产品进行检验（数据详见 Quality.sas7bdat），得到的检验结果为：合格品 96 个，不合格品 4 个。能否据此认为该批次产品不合格？设显著性水平  $\alpha=0.05$ 。

例 3-6 是例 3-2 的延伸，主要根据样本产品的合格情况来检验总体产品的合格情况。根据题意，从样本中得到的产品合格率为 96%，看起来低于标准规定的合格率 97%。因此，应当考察这 1% 的差距究竟是由随机因素引起的，还是产品确实与标准的规定存在差距。提出原假设和备择假设为如下。

$H_0: p \leq 0.97; H_1: p > 0.97$

原假设表示产品合格率不超过 97%，即本批次产品不合格；备择假设表示产品合格率大于 97%，即本批次产品合格。

**STEP 1** 进入 SAS/Analyst，打开 Quality.sas7bdat 数据集，利用 “Statistics → Hypothesis Tests” 进行假设检验。由于是对总体比例进行假设检验，因此在 “Hypothesis Tests” 二级菜单中选择 “One-Sample Test for a Proportion”，弹出总体比例的假设检验对话框，如图 3-14 所示。

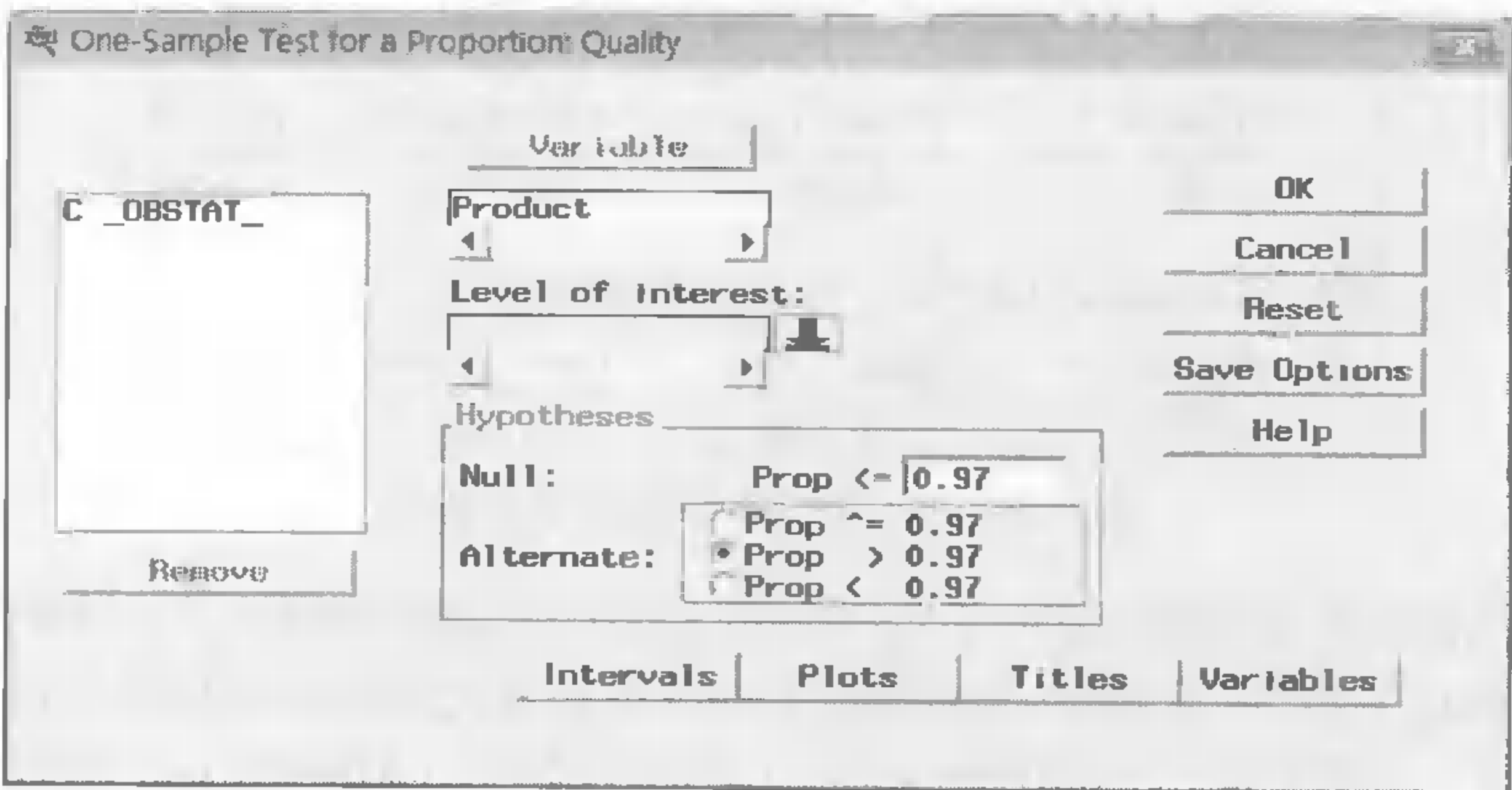


图 3-14 单总体比例的假设检验对话框

**STEP 2** 在图 3-14 所示对话框的中部选中 “Product” 变量，单击 “Variable” 按钮，将其指定为分析变量。在 “Level of interest” 的文本输入框中，可以指定要研究总体的某种属性，并单击 按钮进行设定。因为要对产品合格率进行检验，单击 按钮后，选择 “合格” 属性。

**STEP 3** 在 “Hypotheses” 分栏下的原假设 “Null” 的文本输入框中输入 “0.97”，并根据例 3-6 的备择假设在 “Alternate” 的单选框中选择 “Prop>0.97”，表示本例所设定的原假设和备择假设。同样，单击 “Intervals” 按钮可对变量进行区间估计，如图 3-15 所示。

**STEP 4** 图 3-15 所示对话框类似于图 3-8 所示的总体均值区间估计对话框。在 “Confidence intervals” 分栏下选择 “Interval” 单选框可以得到总体比例的置信区间，选择 “Lower bound” 单选框则可得到总体比例置信区间的下界，选择 “Upper bound” 则可得到总体比例置信区间的下界。在 “Confidence level” 文本输入框中可以选择或输入区间估计的置

信度。系统默认不输出区间估计的置信区间。在本例中，选中“Interval”单选框和 95%置信度进行区间估计。单击“OK”按钮返回图 3-14 所示的对话框。在该对话框中，单击“OK”按钮便可得到总体比例假设检验和区间估计的结果，如图 3-16 所示。

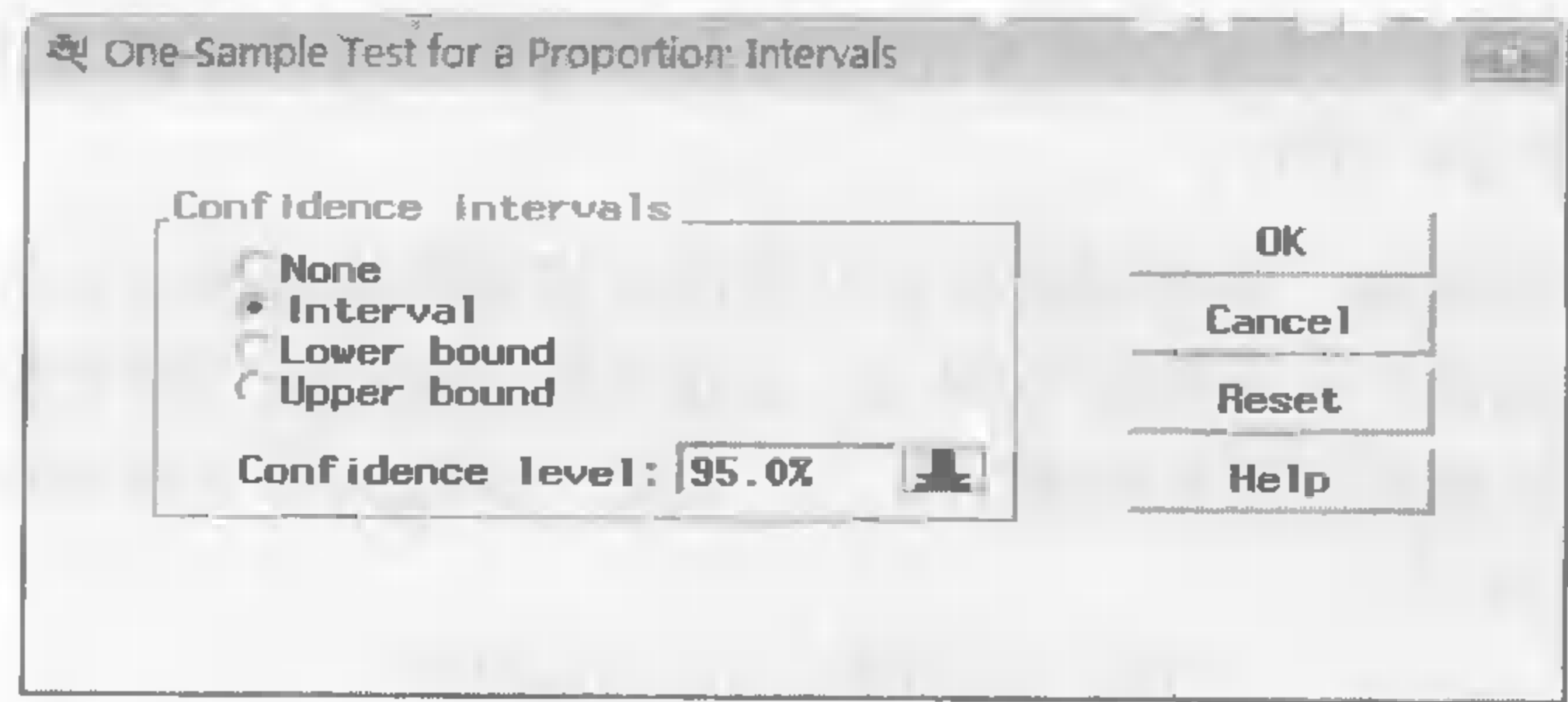


图 3-15 总体比例的区间估计对话框

Sample Statistics			
Product	Frequency		
不合格	4		
合格	96		
<hr/>			
Total	100		
Hypothesis Test			
Null Hypothesis:	Proportion $\leq 0.97$		
Alternative:	Proportion $> 0.97$		
Product	Proportion	Z Statistic	Pr > Z
合格	0.9600	-0.59	0.7211
95% Confidence Interval for the Proportion			
Product	Lower limit	Upper limit	
合格	0.922	0.998	

图 3-16 单总体比例假设检验的结果

在单总体比例的假设检验结果中，首先列示的是样本属性的情况，如本例中不合格产品的频数（Frequency）为 4，合格产品的频数为 96。单总体比例的假设检验的过程与均值的假设检验过程一样，对原假设（Null Hypothesis）和备择假设（Alternative）进行了详细描述，在该过程中，可以看到合格产品率的检验问题，即“合格”属性对应的比例为 0.9600，对应的 Z 统计量为-0.59，P 值为 0.7211。因为 P 值远远大于  $\alpha$  (0.05)，所以没有充分证据表明应该拒绝原假设，即没有充分理由否定产品合格率不超过 97%。

3. 总体方差的假设检验

总体方差的假设检验是指根据样本数据对总体方差提出假设并进行检验的过程。通常可以利用卡方 ( $\chi^2$ ) 统计量进行检验。



例 3-7

学校对学生统计学期末考试成绩进行抽查，以考查本门课程教学质量是否稳定。按照以往教学经验，所有该门课程成绩的标准差应当保持在 12 分。如果某次考试的标准差显著偏离 12 分，则表示教学质量可能存在问题。试用随机抽取的 50 名学生的成绩（如表 3-5 所示，具体数据见 Score\_Variance.sas7bdat）进行教学质量判定。设显著性水平  $\alpha = 0.05$ 。

表 3-5 50 名学生的期末统计学考试成绩

85	87	92	79	95	98	94	85	61	96
85	80	55	76	57	70	80	92	60	93
65	65	89	57	97	75	77	56	68	67
62	99	72	77	59	90	57	96	62	92
63	99	86	98	98	72	55	75	68	79

本例是利用 50 名学生的样本成绩对总体方差进行假设检验。如教学质量稳定，表示总体方差  $\sigma^2$  应当等于  $12^2$ ，即 144，据此可提出该问题的原假设和备择假设。

$H_0: \sigma^2 = 144; H_1: \sigma^2 \neq 144$

原假设表示总体方差等于 144，即本学期课程教学质量稳定；备择假设表示总体方差不等于 144，即本学期课程教学质量不稳定。

**STEP 1)** 进入 SAS/Analyst，打开 Score\_Variance.sas7bdat 数据集，利用 “Statistics → Hypothesis Tests” 进行假设检验。由于是对总体方差进行假设检验，因此在 “Hypothesis Tests” 二级菜单中选择 “One-Sample Test for a Variance”，弹出总体方差的假设检验对话框，如图 3-17 所示。

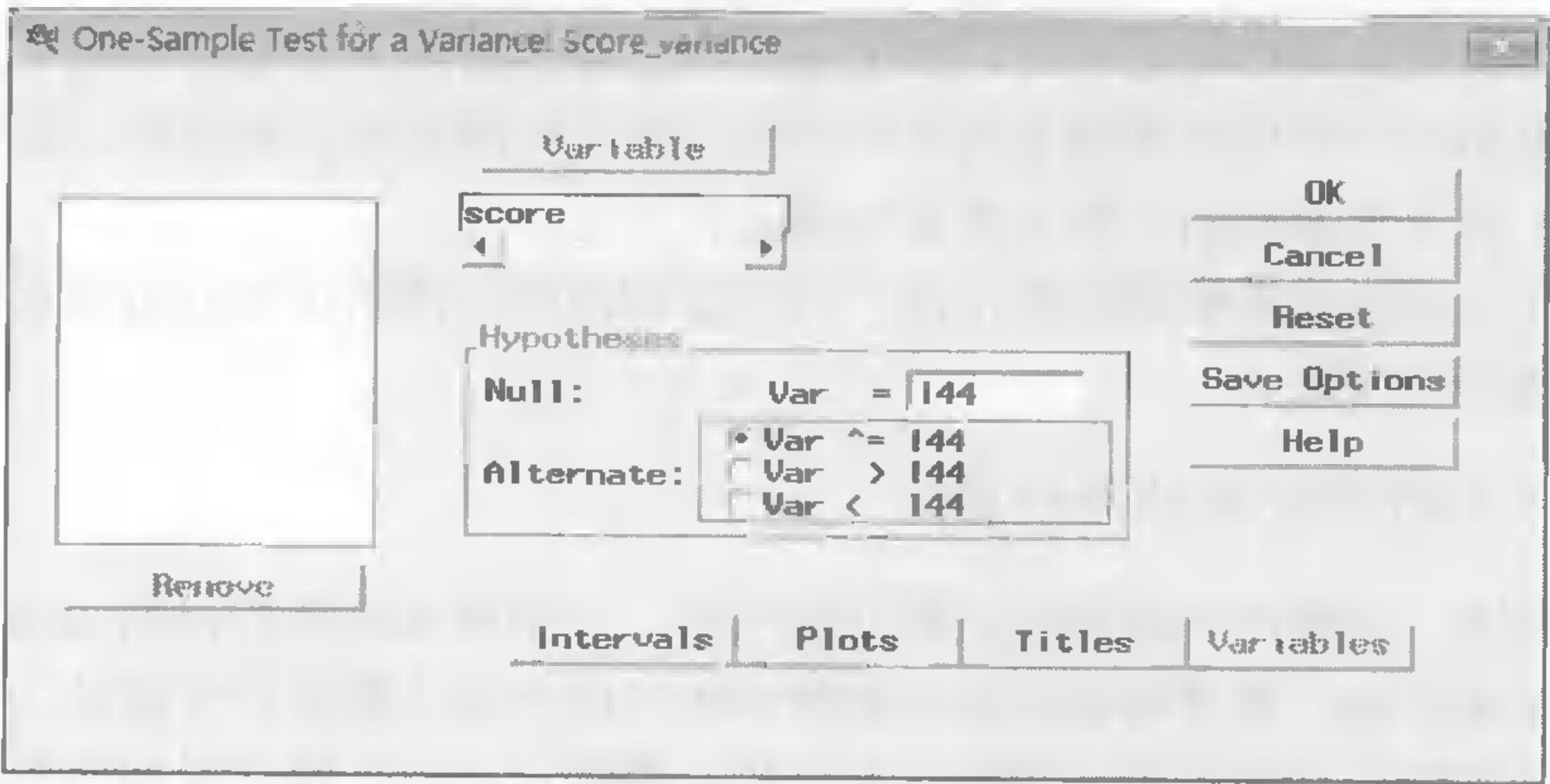


图 3-17 单总体方差的假设检验对话框

**STEP 2)** 在图 3-17 所示对话框的中部选中 “score” 变量，单击 “Variable” 按钮，将其指定为分析变量。在 “Hypotheses” 分栏下的原假设 “Null” 的文本输入框中输入 “144”，并根据例 3-7 的备择假设在 “Alternate” 的单选框中选择 “Var ^= 144”，表示本例所设定的原假设和备择假设。同样，单击 “Intervals” 按钮可弹出与图 3-15 类似的对话框，在该对话框中可对变量进行区间估计。单击 “OK” 按钮，可以得到总体方差的假设检验结果，如图 3-18 所示。

在单总体方差的假设检验结果中，同样首先列示的是样本的情况，如本例中样本方差为 212.46。其假设检验的过程与均值、比例的假设检验过程一样，对原假设 (Null Hypothesis) 和备择假设 (Alternative) 进行了详细描述。在该过程中，可以看到在原假设成立的条件下计算的  $\chi^2$  (Chi-square) 统计量为 72.295，对应的  $P$  值为 0.0337。因为  $P$  值小于  $\alpha$  (0.05)，所以在显著性水平  $\alpha = 0.05$  的情况下，拒绝原假设，即本学期统计学课程的教学质量不稳定。

如果本例给定的显著性水平  $\alpha = 0.01$ ，则  $P$  值大于  $\alpha$ ，因此在显著性水平  $\alpha = 0.01$  的情况下，没有充分理由可以拒绝本学期统计学课程教学质量稳定的结论。

Sample Statistics for score			
N	Mean	Std. Dev.	Variance
50	77.9	14.576	212.46
Hypothesis Test			
Null hypothesis:		Variance of score = 144	
Alternative:		Variance of score ^ = 144	
Chi-square		Df	Prob
72.295		49	0.0337
95% Confidence Interval for the Variance			
Lower Limit		Upper Limit	
148.25		329.92	

图 3-18 单总体方差的假设检验结果

从本例可以看出，做出是否拒绝原假设的结论，与理论显著性水平的设定有关。这也是统计结论不确定性的一个具体表现。

### 3.3 两总体参数的估计及假设检验

参数估计和假设检验的问题也可以扩展至两个总体的情形，主要考察两个总体的参数是否有差异。如两所高等学校高考招生的平均录取分数及标准差是否有差异、新开发的减肥良药是否有疗效、两个国家的新生婴儿男女比例是否有差异等。

根据来自于总体样本数据的性质不同，两个总体的统计推断可细分为独立样本的统计推断及成对样本的统计推断。

#### 3.3.1 独立样本的参数估计和检验

所谓独立样本，即两个样本数据是相互独立的，一个样本数据特征的变动不会影响另一个样本数据特征的变动。如考察由两种不同技术生产的产品产量是否有差异，对从不同技术生产的产品批次中随机抽取的若干样本进行分析。通常，对两个独立样本的参数估计和假设检验不要求两个样本的样本量相等。

##### 1. 独立样本均值之差的参数估计和假设检验

独立样本均值之差的参数估计主要考察两个总体均值的差异程度或置信区间，而假设检验主要考察两个总体的均值是否有差异或检验其差异的具体数值。一般假定两个总体均服从正态分布，在 SAS 系统中使用 *t* 统计量进行检验。



**例 3-8**

某笔记本电脑电池制造商经过研发与创新，开发出两种新的生产工艺，使得使用该品牌电池的笔记本电脑的续航时间有所改进。为了检验这两种新生产工艺对电池续航能力是否有明显的影响，技术人员随机抽取了利用这两种新工艺生产的两个批次的 4 芯电池，考察其在同一型号笔记本电脑上的续航时间（数据详见 Battery.sas7bdat），如表 3-6 所示。设显著性水平  $\alpha = 0.01$ ，检验这两种新工艺对电池续航时间的影响是否有显著差异。如果存在显著差异，在给定的显著性水平下计算差异的置信区间。

表 3-6 由两种新工艺生产的电池在笔记本电脑上的续航时间（单位：小时）

技术类型 Tech	续航时间 Endurance	技术类型 Tech	续航时间 Endurance	技术类型 Tech	续航时间 Endurance	技术类型 Tech	续航时间 Endurance
Type A	4.1	Type A	3.4	Type B	3.9	Type B	4.8
Type A	3.7	Type A	3.9	Type B	4.2	Type B	3.8
Type A	3.5	Type A	3.9	Type B	3.9	Type B	3.8
Type A	3.9	Type A	3.3	Type B	4.5	Type B	3.4
Type A	4.1	Type A	3.4	Type B	4.0	Type B	3.8
Type A	3.5	Type A	4.0	Type B	4.4	Type B	4.1
Type A	3.5	Type A	3.9	Type B	4.1	Type B	4.7
Type A	3.6	Type A	3.7	Type B	3.3	Type B	4.3
Type A	4.1	Type A	3.4	Type B	4.0	Type B	3.3
Type A	4.0	Type A	3.5	Type B	3.4	Type B	3.9
Type A	4.1	Type A	3.6	Type B	4.5	Type B	3.8
Type A	4.0	Type A	3.9	Type B	3.8	Type B	3.8
Type A	3.7	Type A	3.3	Type B	4.7	Type B	4.6
Type A	4.3	Type A	4.0	Type B	3.6	Type B	4.1
Type A	4.2	Type A	3.5	Type B	4.3	Type B	3.4
Type A	3.2	Type A	3.4	Type B	3.5	Type B	3.8
Type A	3.8	Type A	3.6	Type B	3.9	Type B	4.2
Type A	3.4			Type B	3.8		

来自两总体的独立样本数据参数估计和假设检验的问题，在 SAS 系统中对数据格式有一定的要求，即在数据集中应当有用于区分来自于不同总体样本数据的分类变量。如本例 Battery.sas7bdat 数据集中有两个变量，其中的“Endurance”变量为随机抽取的样本数据观测值。分类变量既可以是数值型变量也可以是字符型变量，本例中的“Tech”变量是字符型变量，表示生产工艺的类型，其有两个数值：“Type A”及“Type B”分别表示具体的生产工艺。

设 $\mu_1$ 和 $\mu_2$ 分别表示两种工艺生产的所有电池的总体平均续航时间，如果由两种工艺生产的电池的续航时间没有差异，那么 $\mu_1 = \mu_2$ ，即 $\mu_1 - \mu_2 = 0$ ；如果存在差异，则 $\mu_1 - \mu_2 \neq 0$ 。据此可以根据本例提出原假设和备择假设。

$H_0: \mu_1 - \mu_2 = 0, H_1: \mu_1 - \mu_2 \neq 0$

原假设表示由两种工艺生产的电池的续航时间无差异，而备择假设表示由两种工艺生产的电池的续航时间有差异。

**STEP 1** 进入 SAS/Analyst，打开 Battery.sas7bdat 数据集，利用“Statistics→Hypothesis Tests”进行假设检验。由于是对来自于两总体的独立样本数据进行假设检验，因此在“Hypothesis Tests”二级菜单中选择“Two-Sample t-test for Means”，弹出独立样本均值之差的假设检验对话框，如图 3-19 所示。

**STEP 2** 在图 3-19 所示的对话框中，需要指定分类变量和用于分析的变量。因为本例以“Tech”变量作为分类变量考察电池续航时间，所以首先应当在左上方的“Groups are in”分栏下选择“One variable”单选框，表示用一个分类变量来区分两个样本数据。

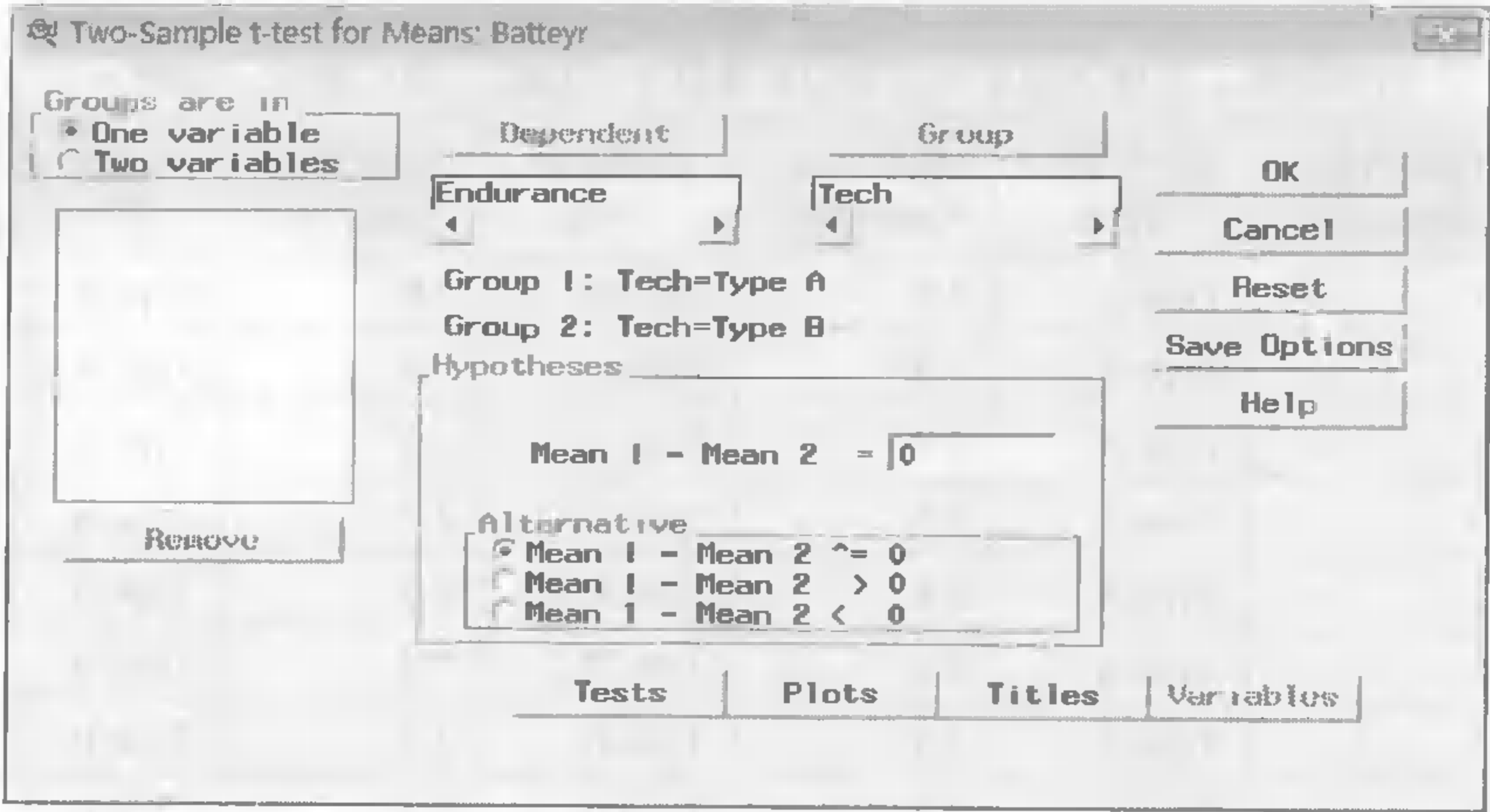


图 3-19 独立样本均值之差假设检验对话框

**STEP 3** 在对话框中部的变量选择区域中，选中“Endurance”变量，单击“Dependent”按钮，将其设置为因变量，即本例分析的对象；选中“Tech”变量，单击“Group”按钮，将其设置为分类变量。这时在“Dependent”按钮和“Group”按钮下的分类状态显示栏中，系统会自动根据选中的 Group 分类变量的具体观测值显示两个分类的具体数值。在本例中，在“Group 1”中显示“Tech = Type A”，在“Group 2”中显示“Tech = Type B”，分别表示由两种生产工艺生产的电池的续航时间。

**STEP 4** 在“Hypotheses”分栏下同样可以指定原假设和备择假设，在“Mean1-Mean2 =”文本输入框中输入“0”（在默认情况下，系统自动指定为“0”），表示原假设所表示的“Group1”的均值与“Group 2”的均值之差为 0，即“Group1”与“Group 2”所代表的总体均值无差异。然后在“Alternative”中选择备择假设“Mean 1-Mean 2 ^= 0”单选框。

**STEP 5** 单击“Tests”按钮，选择“Confidence Intervals”选项卡中的“Interval”单选框，在“Confidence Level”文本输入框中输入“99%”（本例给定的显著性水平  $\alpha = 0.01$ ），即可对两总体均值之差进行区间估计，设置好置信度和置信区间之后，单击“OK”按钮返回假设检验对话框。在该对话框中，单击“OK”按钮便可得到图 3-20 所示的区间估计和假设检验结果。

Sample Statistics				
Group	N	Mean	Std. Dev.	Std. Error
Type A	35	3.725714	0.2994	0.0506
Type B	35	3.982857	0.4112	0.0695
Hypothesis Test				
Null hypothesis:		Mean 1 - Mean 2 = 0		
Alternative:		Mean 1 - Mean 2 ^= 0		
If Variances Are		t statistic	Df	Pr > t
Equal		-2.991	68	0.0039
Not Equal		-2.991	62.13	0.0040
99% Confidence interval for the Difference between Two Means				
		Lower Limit	Upper Limit	
		-0.48	-0.03	

图 3-20 独立样本均值之差的区间估计和假设检验结果

在上述结果中，首先显示的是两个独立样本的常用样本量、样本均值、样本标准差和标准误差等统计量信息，然后给出的是假设检验的过程和结果。独立样本均值之差的假设检验结果不同于单总体均值的假设检验结果。在该结果中，在 “If Variances Are” 中分别给出了两总体方差相等 (“Equal”) 和不相等 (“Not Equal”) 两种情况下的检验结果，可以根据总体方差的具体假定进行判断。如本例中，无论是在总体方差相等还是不相等的情况下，根据样本数据计算出来的  $P$  值都远远小于给定的理论显著性水平  $\alpha$  (0.01)。因此，可以得到以下结论：在显著性水平  $\alpha = 0.01$  的情况下，拒绝原假设，即由两种生产工艺生产的电池的续航时间存在显著差异。这种差异在 99%置信度下的置信区间为  $[-0.48, -0.03]$ 。

利用 SAS 语言的 TTEST 过程可以对独立样本均值进行区间估计和假设检验，具体程序如下。

```
proc ttest data=Sasuser.Battery alpha=0.01; /*调用 TTEST 过程, 并指定区间估计的显著性水平  $\alpha=0.01$ */
  class Tech; /*指定用于区分两个样本数据的分类变量*/
  var Endurance; /*指定分析变量*/
run;
```

运行程序后，得到图 3-21 所示的结果。

The TTEST Procedure										
Statistics										
Variable	Tech	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err	Minimum
Endurance	Type A	35	3.5877	3.7257	3.8638	0.2273	0.2994	0.4297	0.0506	3.2
Endurance	Type B	35	3.7932	3.9829	4.1725	0.3123	0.4112	0.5903	0.0695	3.3
Endurance	Diff (1-2)		-0.485	-0.257	-0.029	0.294	0.3597	0.4592	0.086	
T-Tests										
Variable	Method	Variances	DF	t Value	Pr >  t					
Endurance	Pooled	Equal	68	-2.99	0.0039					
Endurance	Satterthwaite	Unequal	62.1	-2.99	0.0040					
Equality of Variances										
Variable	Method	Num DF	Den DF	F Value	Pr > F					
Endurance	Folded F	34	34	1.89	0.0683					

图 3-21 TTEST 过程进行独立样本均值之差的假设检验结果

利用 TTEST 过程与利用 SAS/Analyst 进行的假设检验结果类似，但是 TTEST 结果呈现了更多信息。首先分析的是两个样本 (Type A 和 Type B) 各自的样本统计量，以及每一个样本的样本均值、样本标准差在给定置信度下的区间估计结果。

TTEST 过程的假设检验结果更为详细，不仅给出了总体方差相等和不相等两种情况下的假设检验结果，还给出了总体方差是否相等的假设检验结果。图 3-21 中的 “Equality of Variances” 信息表示检验总体方差是否相等的检验过程，其原假设是总体方差相等。在原假设成立的情况下，计算出来的  $P$  值 (“Pr>F”) 为 0.068 3，在给定显著性水平  $\alpha = 0.01$  的条件下， $P$  值大于  $\alpha$ 。其检验结果表明，没有充分理由拒绝总体方差相等的假设，即在  $\alpha = 0.01$  条件下，可以认为两个总体方差相等。

根据在  $\alpha = 0.01$  条件下两个总体方差相等的结论，在 “T-Tests” 的结果中，第 3 列的 “Variances” 的 “Equal” 行表示总体方差相等条件下的检验过程，对应的  $P$  值 (“Pr>|t|”) 为

0.0039, 远远小于 0.01, 因此可以在  $\alpha=0.01$  条件下拒绝原假设, 即拒绝由两种工艺生产的电池的续航时间相等的假设, 认为由两种工艺生产的电池的续航时间有差异。



例 3-9

续例 3-8。技术人员经过样本数据的观测, 发现由 Type B 工艺生产的电池的续航时间比由 Type A 工艺生产的电池的续航时间要长。经过长时间的实验, 技术人员估计 Type B 工艺的续航时间要比 Type A 工艺的续航时间长 0.1 个小时。试问能够在显著性水平  $\alpha=0.05$  的条件下对技术人员的估计进行检验吗? (数据仍见 Battery.sas7bdat。)

根据技术人员的推测和研究假设, 可以提出该问题的原假设和备择假设。

$H_0: \mu_2 - \mu_1 \leq 0.1, H_1: \mu_2 - \mu_1 > 0.1$  或者  $H_0: \mu_1 - \mu_2 \geq -0.1, H_1: \mu_1 - \mu_2 < -0.1$

原假设表示 Type B 工艺的续航时间与 Type A 工艺的续航时间之差不超过 0.1 个小时, 而备择假设表示 Type B 工艺的续航时间要比 Type A 工艺的时间长 0.1 个小时。

**STEP 1** 进入 SAS/Analyst, 打开 Battery.sas7bdat 数据集, 利用 “Statistics → Hypothesis Tests” 进行假设检验。同样由于是对来自于两总体的独立样本数据进行假设检验, 因此在 “Hypothesis Tests” 二级菜单中选择 “Two-Sample t-test for Means”, 弹出图 3-19 所示的独立样本均值之差的假设检验对话框。

**STEP 2** 在该对话框中, 仍然按照例 3-8 中的步骤指定分类变量和分析变量。与例 3-8 不同的是, 在 “Hypotheses” 的 “Mean 1-Mean 2 =” 文本输入框中输入 “-0.1”, 在 “Alternative” 分栏下选择 “Mean 1-Mean 2 < -0.1”。在此过程中, 一定要注意 “Mean 1” 和 “Mean 2” 分别代表哪一个工艺。单击 “OK” 按钮便可得到假设检验的结果。

在分析结果中, 可以看到无论是总体方差相等还是不相等, 其检验的  $P$  值都小于  $\alpha(0.05)$ , 表示拒绝原假设, 即在  $\alpha=0.05$  条件下, 可以认为由 Type B 工艺生产的电池的续航时间比由 Type A 工艺生产的电池的续航时间长 0.1 个小时。

对于此问题, 同样可以利用 SAS 语言中的 TTEST 过程进行假设检验, 具体程序如下。

```
proc ttest data=Sasuser.Battery h0=-0.1; /*调用 TTEST 过程, 并设置双侧检验的原假设 h0=-0.1 */
  class Tech;
  var Endurance;
run;
```

运行结果如图 3-22 所示。

T-Tests					
Variable	Method	Variances	DF	t Value	Pr >  t
Endurance	Pooled	Equal	68	-1.83	0.0720
	Satterthwaite	Unequal	62.1	-1.83	0.0724
Equality of Variances					
Variable	Method	Num DF	Den DF	F Value	Pr > F
Endurance	Folded F	34	34	1.89	0.0683

图 3-22 TTEST 过程进行独立样本均值之差的单侧假设检验结果


在对单总体均值的  $t$  检验过程中, 已经知道 TTEST 过程只能输出双侧检验 (即备择假设是 “ $\neq$ ” 符号) 的  $P$  值。如想得到单侧检验结果, 必须按照一定原则计算 (详见 3.2.2 节)。

首先考察总体方差是否相等,“Equality of Variance”中的  $P$  值 (“Pr>F”) 为 0.068 3, 大于显著性水平  $\alpha(0.05)$ , 可以认为总体方差是相等的。因此在 “T-Tests” 的结果中, 应该分析 “Variances” 值为 “Equal” 的行。

本例为单侧检验的过程, 所以单侧检验的  $P$  值为:  $0.0720/2 = 0.036$ , 小于  $\alpha(0.05)$ 。因此, 在显著性水平  $\alpha = 0.05$  条件下, 可以得到与 SAS/Analyst 分析相同的结论, 即认为由 Type B 工艺生产的电池的续航时间比由 Type A 工艺生产的电池的续航时间长 0.1 个小时。

2. 独立样本比例之差的参数估计和假设检验

独立样本比例之差的参数估计主要考察两个总体某种属性比例的差异程度或置信区间而假设检验主要考察两个总体比例是否有差异或检验其差异的具体数值。这里所谓的比例仍然是指总体只具备两种属性, 其中具备某种属性的个体数目占总体数目的比重, 即假定两个总体都服从二项分布, 通常用  $Z$  统计量进行检验。



**例 3-10**

某出版集团为了对旗下两本时尚杂志进行精确的市场定位, 分别对两本杂志的读者性别进行了随机的抽样调查, 调查结果 (数据详见 Magazine.sas7bdat) 如表 3-7 所示。试在显著性水平  $\alpha = 0.01$  条件下分析两本杂志的读者性别是否有显著差异。

表 3-7 两本杂志的读者性别的抽样调查结果

杂志名称 Name	性别 Gender	杂志名称 Name	性别 Gender	杂志名称 Name	性别 Gender	杂志名称 Name	性别 Gender
Fashion	Male	Fashion	Female	Cosmetic	Male	Cosmetic	Male
Fashion	Female	Fashion	Male	Cosmetic	Female	Cosmetic	Female
Fashion	Male	Fashion	Female	Cosmetic	Female	Cosmetic	Female
Fashion	Male	Fashion	Male	Cosmetic	Male	Cosmetic	Female
Fashion	Male	Fashion	Female	Cosmetic	Female	Cosmetic	Male
Fashion	Male	Fashion	Male	Cosmetic	Female	Cosmetic	Female
Fashion	Male	Fashion	Female	Cosmetic	Female	Cosmetic	Male
Fashion	Female	Fashion	Female	Cosmetic	Female	Cosmetic	Female
Fashion	Female	Fashion	Female	Cosmetic	Male	Cosmetic	Female
Fashion	Male	Fashion	Female	Cosmetic	Female	Cosmetic	Male
Fashion	Male	Fashion	Male	Cosmetic	Female	Cosmetic	Female
Fashion	Female	Fashion	Female	Cosmetic	Female	Cosmetic	Female
Fashion	Male	Fashion	Male	Cosmetic	Female	Cosmetic	Female
Fashion	Male	Fashion	Female	Cosmetic	Male	Cosmetic	Male
Fashion	Male	Cosmetic	Female	Cosmetic	Female	Cosmetic	Female
Fashion	Female	Cosmetic	Female	Cosmetic	Female	Cosmetic	Female
Fashion	Male	Cosmetic	Female	Cosmetic	Female	Cosmetic	Male
Fashion	Male	Cosmetic	Female	Cosmetic	Male	Cosmetic	Female
Fashion	Female	Cosmetic	Female	Cosmetic	Female	Cosmetic	Female
Fashion	Female	Cosmetic	Female	Cosmetic	Female	Cosmetic	Female

与独立样本均值之差检验一样，独立样本比例之差的参数估计和假设检验也要求在数据集中应当有用于区分来自于不同总体样本数据的分类变量。本例 Magazine.sas7bdat 数据集中有两个变量，其中的“Gender”变量为随机抽取的样本数据观测值，分类变量“Name”变量是字符型变量，表示两本杂志的名称，其有 2 个数值：“Fashion”及“Cosmetic”。

在本例中，既可以考察男性读者的比例，也可以考察女性读者的比例。在实际生活中，时尚杂志往往受女性读者青睐，所以本例考察女性读者的比例。

设  $p_1$  和  $p_2$  分别表示两本杂志读者的总体女性比例，如果两本杂志读者性别比例没有差异，那么  $p_1 = p_2$ ，即  $p_1 - p_2 = 0$ ；如果存在差异，则  $p_1 - p_2 \neq 0$ 。据此可以根据本例提出原假设和备择假设。

$$H_0: p_1 - p_2 = 0, H_1: p_1 - p_2 \neq 0$$

原假设表示两本杂志的女性读者比例无差异，备择假设表示两本杂志的女性读者比例有差异。

**STEP 1** 进入 SAS/Analyst，打开 Magazine.sas7bdat 数据集，利用“Statistics→Hypothesis Tests”进行假设检验。由于是对来自于两总体的独立样本数据进行假设检验，因此在“Hypothesis Tests”二级菜单中选择“Two-Sample Test for Proportions”，弹出独立样本比例之差的假设检验对话框，如图 3-23 所示。

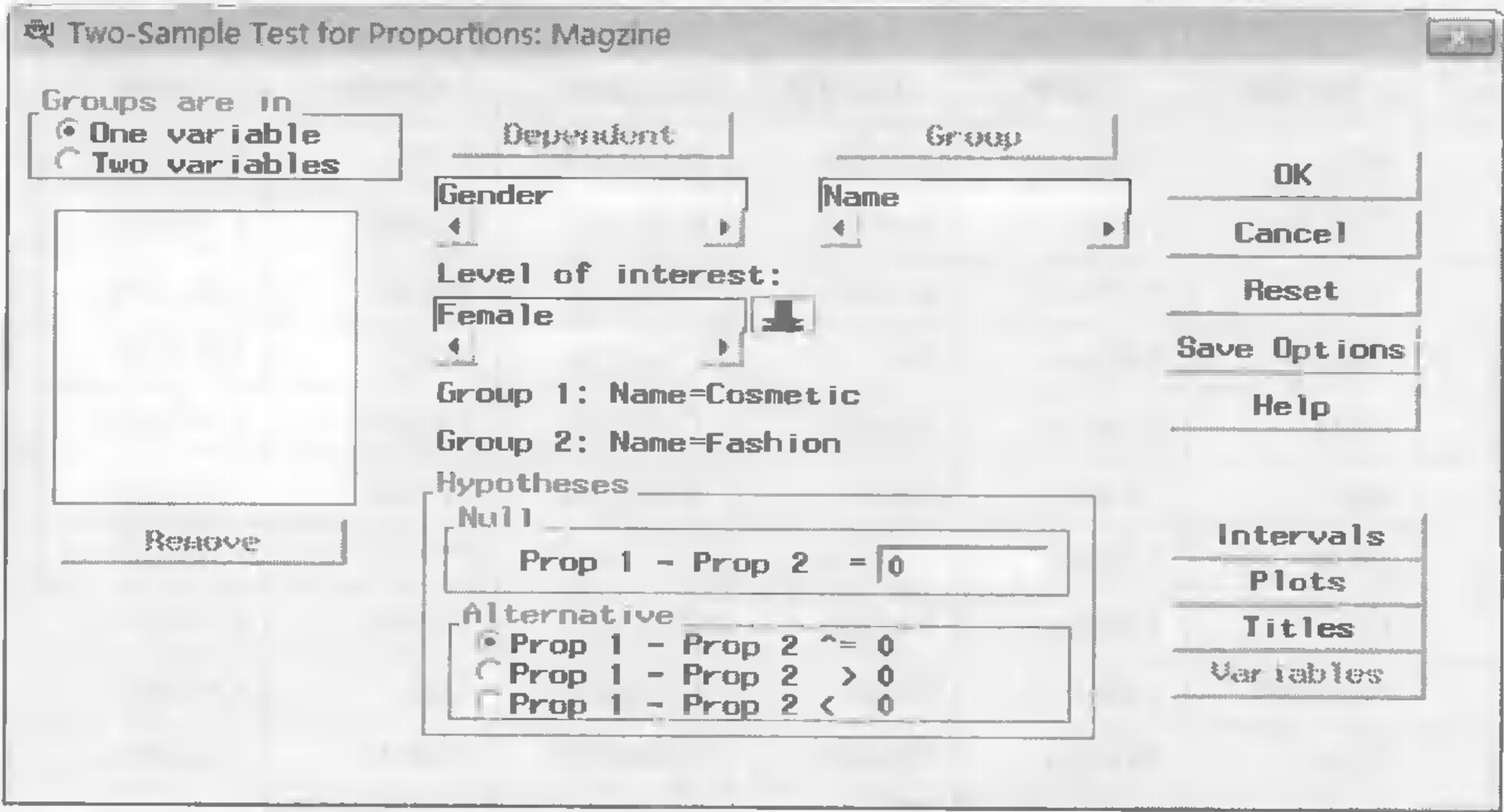


图 3-23 独立样本比例之差的假设检验对话框

**STEP 2** 由于本例使用一个变量对样本来自于不同总体进行区分，因此在对话框左上角的“Groups are in”分栏下仍然选择“One variable”单选框。在变量选择区域中，选中“Gender”变量，单击“Dependent”按钮，将其指定为分析变量；选中“Name”变量，单击“Group”按钮，指定其为分类变量。此时，系统自动指定“Group 1”为“Name=Cosmetic”、“Group 2”为“Name=Fashion”，即在假设中，“Prop 1”代表“Group 1”的总体比例，“Prop 2”代表“Group 2”的总体比例。设置好变量的角色之后，指定想分析的属性值。在“Level of interest”的文本输入框中可以指定要研究的总体的某种属性，可以单击 按钮进行设定。因为要对女性读者比例进行检验，因此单击 按钮后，选

择“Female”属性。然后根据研究需要在“Hypotheses”分栏下设置原假设“ $\text{Prop 1}-\text{Prop 2}=0$ ”和备择假设“ $\text{Prop 1}-\text{Prop 2}^{\wedge}=0$ ”。

**STEP 3** 单击对话框右边的“Intervals”按钮,在弹出的“Interval”对话框中选中“Intervals”单选框,在“Confidence level”文本输入框中输入“99%”(本例给定的显著性水平 $\alpha=0.01$ ),即可对两总体比例之差进行区间估计。设置好置信度和置信区间之后,单击“OK”按钮返回假设检验对话框。在该对话框单击“OK”按钮便可得到图 3-24 所示的区间估计和假设检验结果。

Sample Statistics				
- Frequencies of Gender for Name -				
Value	Cosmetic	Fashion		
Female	35	16		
Male	11	18		
Hypothesis Test				
Null hypothesis:				
Proportion of Gender(Name=Cosmetic) - Proportion of Gender(Name=Fashion) = 0				
Alternative:				
Proportion of Gender(Name=Cosmetic) - Proportion of Gender(Name=Fashion) ^= 0				
- Proportions of Gender for Name -				
Value	Cosmetic	Fashion	Z	Prob > Z
Female	0.7609	0.4706	2.67	0.0076
99% Confidence Interval for the Difference in Two Proportions				
Value	Lower Limit	Upper Limit		
Male	0.017	0.564		

图 3-24 独立样本比例之差的区间估计和假设检验结果

在输出结果中,可看到两本杂志的男性读者与女性读者的样本统计数目。由于假设检验的  $P$  值(“Prob>Z”)为 0.0076 远远小于 $\alpha$  (0.01),因此应当拒绝原假设,即认为两本杂志女性读者的总体比例是有差异的。相应地,两本杂志男性比例之差在 99%置信度下的置信区间为[0.017, 0.564],对应的女性比例之差在 99%置信度下的置信区间为[1-0.564, 1-0.017],即 [0.436, 0.983]。

3. 独立样本方差之比的参数估计和假设检验

对于独立样本方差是否存在差异的考察,往往是从两个总体方差的比值入手的。在通常情况下,假定两个总体服从正态分布,用  $F$  统计量进行检验。



例 3-11

为了对比考察两所大学(A大学和B大学)新生的生源质量,分别从这两所大学新录取的学生中随机抽取 40 人,观测其高考得分(数据详见 Highschool.sas7bdat),如表 3-8 所示。试在显著性水平 $\alpha=0.05$ 条件下,考核两所大学新生高考成绩的分差是否有差异。

本例 Highschool.sas7bdat 数据集中有两个变量,其中的“Score”变量为随机抽取的高考得分样本数据观测值,分类变量“College”是字符型变量,表示两所学校的名称,其有 2 个数值:“A”和“B”。

表 3-8 两所大学的新生高考成绩

学校 College	得分 Score	学校 College	得分 Score	学校 College	得分 Score	学校 College	得分 Score
A	646	A	676	B	648	B	659
A	604	A	609	B	648	B	594
A	617	A	626	B	577	B	590
A	608	A	651	B	677	B	625
A	633	A	591	B	657	B	672
A	609	A	660	B	589	B	655
A	605	A	655	B	568	B	592
A	671	A	654	B	664	B	610
A	587	A	602	B	673	B	575
A	643	A	685	B	669	B	682
A	636	A	677	B	673	B	561
A	635	A	596	B	599	B	660
A	638	A	665	B	694	B	686
A	676	A	606	B	580	B	564
A	599	A	678	B	691	B	684
A	658	A	623	B	568	B	696
A	616	A	686	B	638	B	689
A	646	A	672	B	604	B	588
A	598	A	661	B	644	B	669
A	589	A	638	B	579	B	613

设  $\sigma_1^2$  和  $\sigma_2^2$  分别表示两所学校全体新生高考分数的方差，如果两校全体新生成绩的方差没有差异，那么  $\sigma_1^2 = \sigma_2^2$ ，即  $\sigma_1^2 / \sigma_2^2 = 1$ ；如果存在差异，则  $\sigma_1^2 / \sigma_2^2 \neq 1$ 。据此可以根据本例提出原假设和备择假设。

$$H_0 : \sigma_1^2 / \sigma_2^2 = 1, H_1 : \sigma_1^2 / \sigma_2^2 \neq 1$$

原假设表示两校新生录取分数方差无差异；而备择假设表示两校新生录取分数方差存在显著差异。

**STEP 1)** 进入 SAS/Analyst，打开 Highschool.sas7bdat 数据集，利用 “Statistics → Hypothesis Tests” 进行假设检验。由于是对来自于两总体的独立样本数据进行假设检验，因此在 Hypothesis Tests 二级菜单中选择 “Two-Sample Test for Variances”，弹出独立样本方差之比的假设检验对话框，如图 3-25 所示。

**STEP 2)** 由于本例使用一个变量对样本来自于不同总体进行区分，因此在对话框左上角的 “Groups are in” 分栏下仍然选择 “One variable” 单选框。在变量选择区域中，选中 “Score” 变量，单击 “Dependent” 按钮，把其指定为分析变量；选中 “College” 变量，单击 “Group” 按钮，指定其为分类变量。此时，系统自动指定 “Group 1” 为 “College=A”，“Group 2” 为 “College=B”，即在假设中，“Variance 1” 代表 “Group 1” 的总体方差，“Variance 2” 代表 “Group 2” 的总体方差。然后根据研究需要在 “Hypotheses” 分栏下设置好备择假设 “Variance 1 / Variance 2 ^ = 1”。

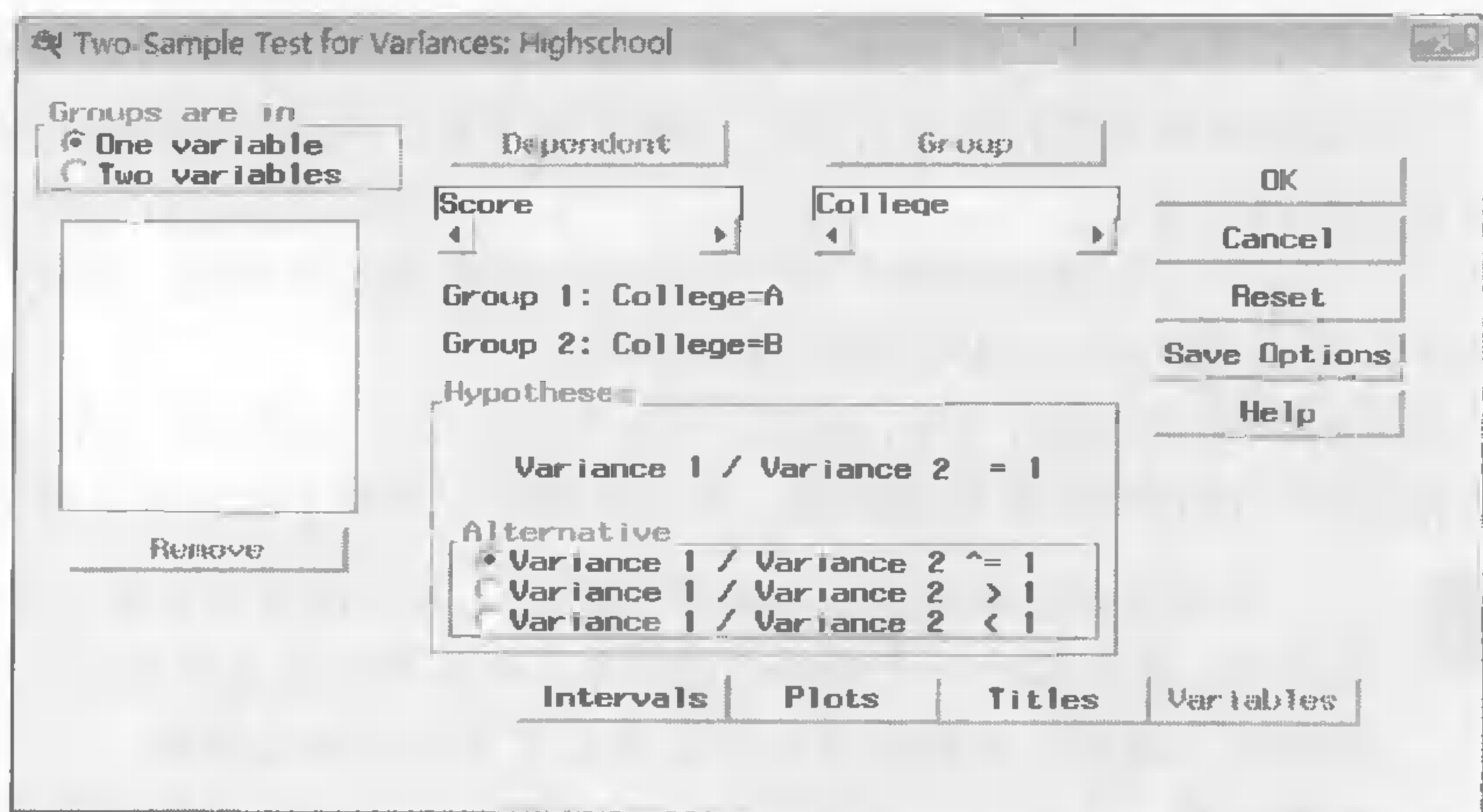


图 3-25 独立样本方差之比的假设检验对话框

**STEP 3** 单击对话框下部的“Intervals”按钮，在弹出的“Interval”对话框中选中“Intervals”单选框，在“Confidence level”文本输入框中输入“95%”（本例给定的显著性水平 $\alpha=0.05$ ），即可对两总体方差之比进行区间估计。设置好置信度和置信区间之后，单击“OK”按钮返回假设检验对话框。在该对话框中，单击“OK”按钮便可得到图 3-26 所示的区间估计和假设检验结果。

Sample Statistics				
College Group	N	Mean	Std. Dev.	Variance
A	40	635.625	30.435	926.2917
B	40	632.6	44.689	1997.118

Hypothesis Test			
Null hypothesis:	Variance 1 / Variance 2 = 1		
Alternative:	Variance 1 / Variance 2 $\neq$ 1		
- Degrees of Freedom -			
F	Numer.	Denom.	Pr > F
0.46	39	39	0.0185

95% Confidence Interval of the Ratio of Two Variances	
Lower Limit	Upper Limit
0.2453	0.8769

图 3-26 独立样本方差之比的区间估计和假设检验结果

从 SAS/Analyst 输出结果中可以得出，该检验问题的  $P$  值（“Pr>F”）为 0.0185，小于 $\alpha$  (0.05)。因此，可以在显著性水平 $\alpha=0.05$ 条件下，拒绝原假设，即两所大学新生的录取成绩方差有显著差异，其方差之比的 95%置信区间为[0.2453, 0.8769]。

3.3.2 成对样本的参数估计和检验

有时来自于两个总体的样本并不是独立的。如对药物的临床疗效进行检验的问题，参与试验的病人吃药前与吃药后的两组考核指标不是独立的，因为这两组数据是来自于相同观测者在不同时期或不同情况下的观测值，这些观测值的具体数值与参与试验的观测者自身素质和药物疗效均有关系。

类似于这样的非独立的两组样本数据通常被称为成对样本或匹配样本。成对样本数据主

要用于对两个总体均值之差进行统计推断。成对样本一般不能使用独立样本参数估计和假设检验的方法，往往要相对样本数据进行处理。处理的理论基础便是组成成对样本的不同个体之间的观测值是相对独立的。如病人 A 吃药前后的考核指标数据与病人 B 吃药前后考核指标数据是独立的。基于此，可以首先把两个样本中配对的观测值逐个相减，形成一个由独立观测值组成的样本，然后用单样本检验方法进行统计推断。

成对样本的数据预处理过程在 SAS 系统中可以自动完成。因此，在 SAS 系统中，用户只要在系统中指定代表成对样本数据的变量，便可直接进行类似于独立样本的分析。



例 3-12

为考察北京市居民生活的幸福程度，首都经贸大学统计学院每年对固定样本（该样本为一随机抽样的样本，每年都按最初抽取的样本进行调查）进行入户调查，并向社会公开发布北京市居民幸福指数。随着社会经济的快速发展，居民收入不断提升，生活水平也相应地不断提高，幸福指数是否会得到提升呢？（设显著性水平  $\alpha=0.05$ 。）对于这个问题，笔者收集了 2006 年、2007 年的调查数据（每年 200 个样本）以进行分析（如表 3-9 所示，数据详见 Happiness.sas7bdat）。

表 3-9 2006 年、2007 年北京市居民幸福指数

06 指数	07 指数	06 指数	07 指数	06 指数	07 指数	06 指数	07 指数	06 指数	07 指数
Happy_06	Happy_07	Happy_06	Happy_07	Happy_06	Happy_07	Happy_06	Happy_07	Happy_06	Happy_07
68.89	75.53	71.37	63.00	65.05	83.81	66.62	69.49	71.53	76.32
80.83	65.55	68.15	54.67	68.77	77.64	67.03	78.52	83.57	71.68
78.00	61.80	71.07	75.77	84.50	71.94	66.87	77.44	82.97	84.44
69.27	68.24	65.11	76.25	77.34	73.86	81.44	73.10	74.85	74.07
74.23	67.02	79.03	63.21	77.95	73.10	72.73	83.72	65.34	81.28
77.41	68.53	75.89	81.83	83.42	77.64	82.67	73.64	69.04	78.24
69.26	60.00	72.71	56.89	75.23	80.92	69.85	84.43	69.55	83.83
77.86	62.13	75.74	72.64	76.25	81.33	75.96	83.50	75.92	68.66
80.66	76.74	75.98	80.42	76.54	73.65	66.48	84.43	65.08	68.66
76.66	53.24	83.26	65.43	68.61	72.25	72.08	67.10	77.19	68.16
82.20	55.72	68.96	65.79	74.90	69.23	74.88	68.94	57.87	76.24
74.15	65.00	81.38	80.04	69.77	57.71	75.01	75.70	74.16	66.36
65.87	54.76	72.02	75.12	72.85	65.67	54.35	77.15	59.98	75.34
71.24	77.13	69.78	76.22	72.12	61.56	60.93	74.99	66.94	80.87
81.27	67.13	77.10	68.62	77.34	74.75	65.51	65.76	54.65	75.12
78.21	69.52	82.30	66.32	65.56	72.12	77.14	77.93	70.99	75.84
71.88	73.48	79.77	69.36	72.85	68.56	75.05	71.58	78.92	77.92
67.44	58.90	76.66	61.78	68.20	63.23	83.69	84.63	65.98	79.91
69.18	56.76	73.48	71.89	81.52	58.79	81.83	67.17	61.40	73.96
83.90	67.85	77.87	80.17	79.56	65.12	73.16	76.37	67.72	75.48
68.08	78.81	80.77	72.68	72.08	83.17	65.56	69.11	82.58	78.57

续表

06 指数	07 指数	06 指数	07 指数	06 指数	07 指数	06 指数	07 指数	06 指数	07 指数
Happy_06	Happy_07	Happy_06	Happy_07	Happy_06	Happy_07	Happy_06	Happy_07	Happy_06	Happy_07
74.50	67.32	75.59	69.60	66.08	82.24	71.13	72.47	77.35	69.94
83.32	66.01	79.13	74.19	74.41	69.78	66.58	74.80	65.53	72.45
69.73	72.20	68.55	72.57	82.97	66.77	75.87	72.01	73.80	67.69
81.34	70.54	67.28	83.53	71.56	84.16	77.10	69.04	68.31	73.97
68.68	68.37	79.92	78.30	84.55	65.48	70.62	76.23	63.19	70.02
71.66	67.32	81.06	84.22	71.13	69.80	82.48	77.32	61.06	68.23
78.02	70.86	72.66	73.09	84.74	78.29	73.19	84.52	57.77	68.15
71.49	66.03	71.60	82.10	76.23	84.24	75.32	75.63	77.51	69.99
77.01	69.52	72.74	78.71	76.92	82.90	84.40	73.84	69.02	78.86
73.48	68.29	82.05	61.38	72.25	73.44	68.78	73.56	63.64	65.00
71.64	79.16	71.83	71.68	67.40	76.31	75.30	72.39	81.27	68.95
65.09	61.71	75.30	79.20	83.29	65.41	83.13	69.61	61.16	82.81
84.93	76.46	79.76	74.10	65.43	82.16	71.04	65.33	65.56	72.89
75.25	73.90	81.74	65.98	73.22	83.61	82.08	72.07	59.33	73.38
78.06	60.29	68.07	79.86	77.81	74.77	82.54	82.31	66.51	69.39
70.08	65.96	83.43	66.68	69.70	81.51	65.90	84.62	52.23	67.81
71.28	79.35	73.53	66.53	72.32	65.48	84.32	80.15	68.96	82.03
81.57	68.28	71.28	68.28	71.02	65.23	78.51	70.13	82.29	81.50
65.43	65.00	77.52	67.97	77.69	78.18	76.53	73.49	77.31	76.72

来自于两总体的成对样本参数估计和假设检验的问题要求成对样本的样本量相同。在本例中，“Happy\_06”表示 2006 年 200 户居民的幸福指数，“Happy\_07”表示这 200 户居民 2007 年的幸福指数。

例 3-12 要分析的假设是 2007 年的北京市居民幸福指数是否在 2006 年的基础之上得到了提升，即利用所收集到的样本数据对 200 户北京市居民的幸福程度进行统计推断。由于该项调查样本在各年中是固定的，即样本数据是通过连续观测（即一点多测）得来的数据，所以可认为这些样本数据观测值是成对样本。

设 2006 年幸福指数的总体平均值为 $\mu_1$ ，2007 年幸福指数总体平均值为 $\mu_2$ ，依据例 3-12 的研究假设，可以提出原假设和备择假设如下。

$$H_0: \mu_1 - \mu_2 \geq 0, H_1: \mu_1 - \mu_2 < 0$$

原假设表示随着经济的发展，居民幸福指数没有得到提升；而备择假设表示随着经济的发展，居民幸福指数没有得到提升。

**STEP 1** 进入 SAS/Analyst，打开 Happiness.sas7bdat 数据集，利用“Statistics→Hypothesis Tests”进行假设检验。由于是对来自于两总体的成对样本数据进行假设检验，因此在 Hypothesis Tests 二级菜单中选择“Two-Sample Paired t-test for Means”，弹出成对样本均值之差的假设检验对话框，如图 3-27 所示。

**STEP 2** 在对话框中部的变量选择区域中，选中“Happy\_06”变量，单击“Group 1”按钮，把其指定为第一组数据；选中“Happy\_07”变量，单击“Group 2”按钮，把其指定为第 2 组数据。然后在“Hypotheses”分栏下设置好本例中的原假设和备择假设。

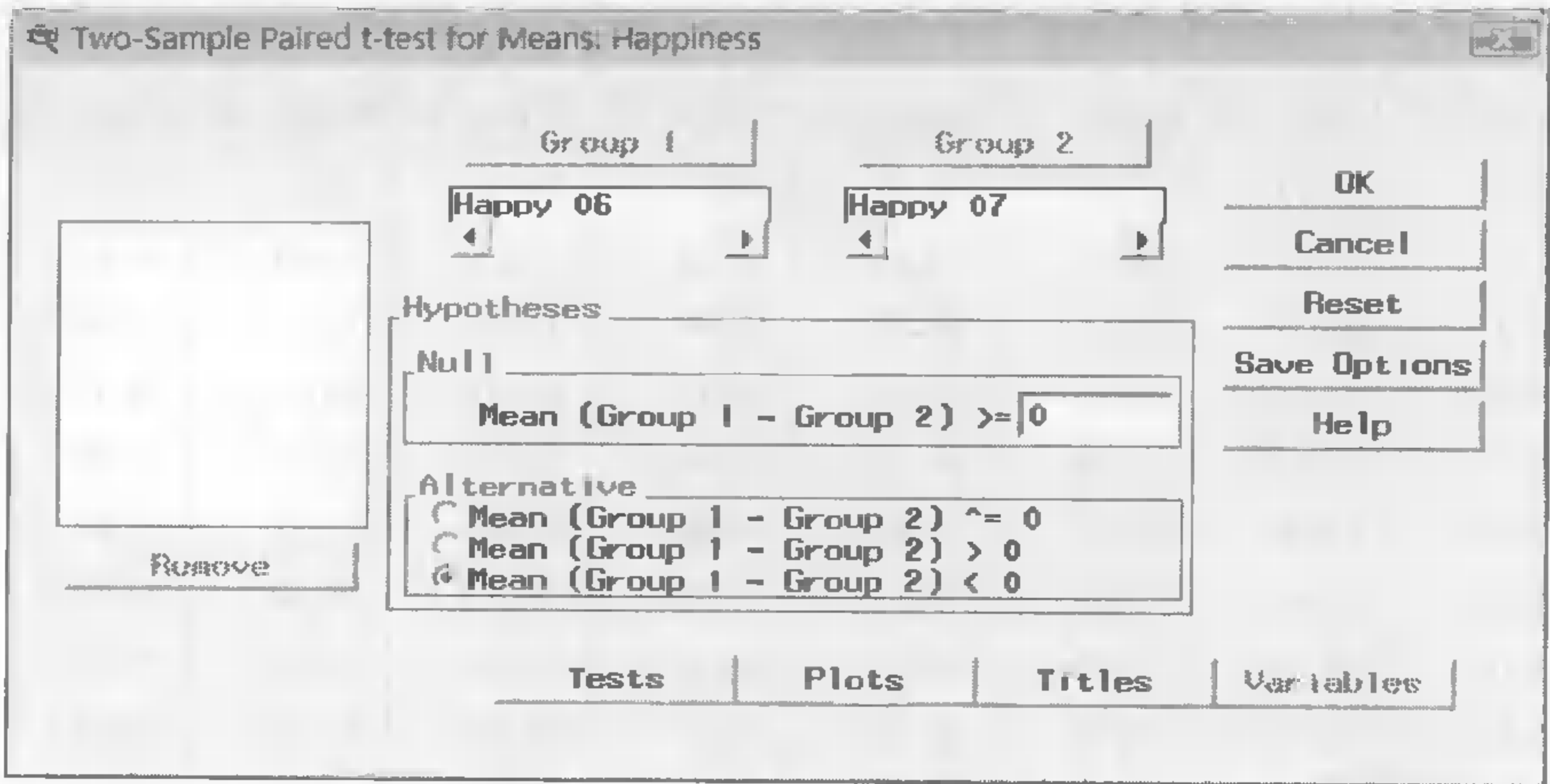


图 3-27 成对样本均值之差的假设检验对话框

**STEP 3)** 单击“Tests”按钮，选择“Confidence Intervals”选项卡中的“Interval”单选框，在“Confidence Level”文本输入框中输入“95%”（本例给定的显著性水平 $\alpha=0.05$ ），即可对两总体均值之差进行区间估计。设置好置信度和置信区间之后，单击“OK”按钮返回假设检验对话框。在该对话框中，单击“OK”按钮便可得到图 3-28 所示的区间估计和假设检验结果。

Sample Statistics				
Group	N	Mean	Std. Dev.	Std. Error
Happy_06	200	73.45655	6.7371	0.4764
Happy_07	200	72.4362	7.0878	0.5012

Hypothesis Test		
Null hypothesis:	Mean of (Happy_06 - Happy_07) => 0	
Alternative:	Mean of (Happy_06 - Happy_07) < 0	
t Statistic	Df	Prob > t
1.424	199	0.9220

95% Confidence Interval for the Difference between Two Paired Means	
Lower Limit	Upper Limit
-0.99	2.43

图 3-28 成对样本均值之差的区间估计和假设检验结果

在输出结果中，可以看到该问题的  $P$  值（“Prob>t”）为 0.9220 远远小于 $\alpha(0.05)$ ，因此没有充分理由拒绝原假设，即在显著性水平 $\alpha=0.05$ 的条件下，2006 年的总体幸福指数要大于 2007 年的总体幸福指数。

对于该问题，同样可以利用 SAS 语言中的 TTEST 过程进行编程分析，具体程序如下。

```
proc ttest data=Sasuser.Happiness h0=0 alpha=0.05; /*调用 TTEST 过程，指定原假设均值之差为“0”，
显著性水平 $\alpha=0.05$  用于设置区间估计的置信度*/
  paired Happy_06*Happy_07; /*指定构成成对样本的两个变量，配对的变量之间用“*”号连接*/
run;
```

运行程序后得到图 3-29 所示的结果。

The TTEST Procedure										
Statistics										
Difference	N	Lower CL Mean	Mean	Upper CL Mean	Lower CL Std Dev	Std Dev	Upper CL Std Dev	Std Err	Minimum	Maximum
Happy_06 - Happy_07	200	-0.393	1.0204	2.4335	9.2289	10.134	11.238	0.7166	-22.8	26.48
T-Tests										
Difference			DF	t Value	Pr >  t					
Happy_06 - Happy_07			199	1.42	0.1561					

图 3-29 成对样本均值之差的 TTEST 检验过程

在图 3-29 所示的输出结果中，系统首先给出了成对样本数值之差的样本统计量及给定显著性水平的均值、标准差的区间估计结果。“T-Tests”则给出了用于假设检验判定的双侧  $P$  值 (“Pr>|t|”) 为 0.1561，由于本例进行的是单侧检验，因此单侧  $P$  值为  $1-0.1561/2 = 0.92195$ ，与用 SAS/Analyst 得到的检验结果的  $P$  值一样，故可得到相同的分析结论。

### 3.4 本章小结

本章主要介绍了简单统计推断的基本内容，简要回顾如下：统计推断内容是建立在总体分布已知或假定已知的基础之上的，可对总体参数进行估计和假设检验；可对单个总体及两总体的均值、方差、比例等进行参数估计和假设检验；在 SAS 系统中，进行假设检验的判断依据与手工检验不同，主要利用检验的  $P$  值进行判定；简单统计推断在 SAS 系统中可用 SAS/Analyst 和 UNIVARIATE、SURVEYMEANS、TTEST 等过程实现；假设检验不仅可对总体参数进行检验，还可以针对模型参数估计的结果进行检验，在后续章节的分析中仍然适用。

## 第4章

# 方差分析

在第3章中,研究了单个和两个总体的参数估计和假设检验问题。但是在实际生活中,往往会遇到对多个总体进行统计推断的问题。如考察某集团公司旗下5个品牌服装的销量是否有差异,从各大商场中收集这5个品牌服装的销售数据以对其总体销量进行分析,其中涉及到对5个总体(把每个品牌服装的销量看作一个总体)参数的假设检验问题。又如在科学实验中常常要探讨不同实验条件或处理方法对实验结果的影响,通常是比较不同实验条件下样本均值间的差异,这是由不同实验条件或处理方法决定的不同总体间的假设检验问题。

方差分析(Analysis of Variance, ANOVA)是利用样本数据检验两个或两个以上的总体均值间是否有差异的一种方法。在研究一个变量时,它能够解决多个总体的均值是否相等的检验问题;在研究多个变量对不同总体的影响时,它也是分析各个自变量对因变量影响的方法。

方差分析不仅广泛应用于社会经济领域,在其他领域往往与实验设计相结合,研究不同因素对研究对象的影响效果,如在医学领域研究几种药物对某种疾病的疗效,在农业领域研究土壤、肥料、气候、日照时间等因素对某种农作物产量的影响,不同饲料对牲畜重量增长的效果等。

### 4.1 方差分析的基本原理

方差分析主要是通过方差比较的方式来对不同总体参数进行假设检验。1928年, Fisher提出一种比较方差的方法,被命名为 $F$ 检验法。 $F$ 作为统计量的表达式如下所示。

$$F = \frac{S_1^2}{S_2^2} = \frac{\Sigma(x_{1i} - \bar{x}_1)^2 / v_1}{\Sigma(x_{2i} - \bar{x}_2)^2 / v_2} = \frac{\text{总体1的方差}}{\text{总体2的方差}}$$

其中 $v_1$ 、 $v_2$ 分别表示总体1和总体2的自由度。

方差是衡量一个总体或样本数据离散程度的重要指标,代表了其所反映数据的差异程度,同时也包含了数据变动的信息。当 $F=1$ 时,表示两个总体方差相等,即没有差异;当 $F \approx 1$ 时,表示两个总体没有显著差异;当 $F \neq 1$ 时,表示两个总体有显著差异。

因此,进行方差分析的关键是找出能够代表 $F$ 表达式的分子分母中的方差测度指标以进行对比分析。为此,可以先从以下例子进行分析。



#### 例 4-1

某市场研究公司受数码相机制造商委托,对市场上销售的消费级数码相机销量进行研究。假定诸如液晶显示屏尺寸、光学变焦倍数、品牌号召

力等影响销量的因素全部相同，现考察数码相机成像元器件像素数是否会对产品销量产生显著的影响。研究人员从地理位置相似、经营规模相当、人气基本无差异的 8 家电脑器件卖场收集了某天的销售数据（详见 DC\_Sale.sas7bdat），如表 4-1 所示。试问成像元器件像素数是否会对数码相机销量产生影响？设显著性水平  $\alpha = 0.05$ 。

表 4-1 8 家电脑卖场的数码相机销售数据（单位：台）

卖场编号 像素分类	卖场 1	卖场 2	卖场 3	卖场 4	卖场 5	卖场 6	卖场 7	卖场 8
500 万像素以下	70	67	82	87	80	80	87	96
500~600 万像素	101	76	97	88	92	99	123	90
600~800 万像素	114	96	128	103	107	91	99	119
800~1 000 万像素	120	98	132	128	132	132	131	119
1 000 万像素以上	132	102	123	119	123	135	126	117

在方差分析中，通常把影响因变量的、可控制的定性变量或离散型变量称为“因素”；而把各个因素具有的表现称为“水平”。如本例中，像素数便是影响因变量的一个因素，其具有 5 个水平。

而影响因变量的定量变量或连续型变量被称为“协变量”，如销售人员奖金对销售量的影响，奖金可作为影响销售量的一个协变量。

在其他条件相同的情况下，本例所要解决的问题就归结为一个多总体的检验问题，即检验成像元器件的像素数对销售量是否有影响。把每一类不同像素的数码相机的总销量分别看成是不同的总体，该问题便转化为以下的假设检验问题。

$$H_0: \mu_1 = \mu_2 = \mu_3 = \mu_4$$

$$H_1: \mu_1、\mu_2、\mu_3、\mu_4 \text{ 不完全相等。}$$

从方差分析目的来看，该问题是要检验 5 类像素数的数码相机销售均值是否相等。可以使用上述方差比较方法来判断。

因此，对于多总体均值比较的方差分析问题，实际上是一个假设检验的问题，其研究本质是比较在因素不同水平下的因变量总体均值是否相等。可建立以下线性模型进行分析。

$$x_{ij} = \bar{x}_i + \varepsilon_{ij}$$

$x_{ij}$  表示作为影响因素的第  $i$  个水平下因变量的第  $j$  个观测值，在本例中即第  $i$  种像素数数码相机在第  $j$  个卖场的销售数据的观测值； $\bar{x}_i$  表示第  $i$  个水平下因变量的均值，在本例中即第  $i$  种像素数的数码相机销售量的均值； $\varepsilon_{ij}$  表示第  $i$  个水平下的因变量第  $j$  个观测值与该水平下因变量均值之间的残差，也被称为随机扰动项，服从均值为零的一个正态分布。在实际问题中，残差表示除所考虑因素之外的其他因素或不可观测的随机因素（如天气、政策变动、不可抗力等）的影响。

对于上述模型可以找出造成因变量差异的各种不同来源，并根据方差比较的方法，对这些不同来源的差异进行分析，找出对因变量影响较大的因素及其水平。然后根据合适的参数估计方法对该线性模型进行估计，得出因素水平变动对因变量变动的具体影响。

首先，把抽样得到的销售数据按不同像素数分为 5 个组，并分别计算各组内部的方差以及组与组之间的方差，如表 4-2 所示。

表 4-2 方差分析各种方差的计算过程

像素数	组内平均数 $\bar{x}_i$	组内离差平方和 (SSA) $\sum (x_{ij} - \bar{x}_i)^2$	组间离差平方和 (SSE) $\sum (\bar{x}_i - \bar{\bar{x}})^2$
500 万像素以下	81.13	698.00	10 472.85
500~600 万像素	95.75	1 375.25	
600~800 万像素	107.13	1 198.00	
800~1 000 万像素	124.00	1 098.00	
1 000 万像素以上	122.13	843.00	
合计	—	4 682.125	
总离差平方和 (SST)		$\sum (x_{ij} - \bar{\bar{x}})^2 = 15 154.975$	

$\bar{\bar{x}}$  表示所有样本数据计算的均值。在表 4-2 所示的计算过程中，可以很容易发现总离差平方和等于组内离差平方和与组间离差平方和之和。

由于离差平方和能够在一定的程度上反映数据的差异，所以总体差异的产生来自两个方面：一方面是由不同像素数的差异（即组间差异）造成的，即不同像素数对销售量产生了影响；另一方面由于抽选样本的随机性而产生的差异，即各不同像素数内部的随机误差，如相同像素数的数码相机在不同卖场的销售量也不同。

上述两个方面产生的差异可以用两个方差来衡量。

- 组间方差：也称为水平之间方差，即组间离差平方和除以自由度  $(r-1)$ ，其中  $r$  为组数或水平数。它既包括系统性因素，也包括随机性因素。
- 组内方差：也称为水平内部方差，即组内离差平方和除以自由度  $(n-r)$ ，其中  $n$  为样本容量总数，仅包括随机性因素。

如果不同的水平（像素数）对因变量没有影响，那么在水平之间的方差中，则仅仅有随机因素的差异，而没有系统性差异，它与水平内部方差应该近似，即：

$$\frac{\text{水平之间（组间）的方差}}{\text{水平内部（组内）方差}} \approx 1$$

两个方差的比值接近于 1。

反之，不同水平对结果有显著影响，水平之间的方差就会大于水平内的方差。当这个比值达到某个程度，或者说达到某临界点时，就可做出不同的水平之间存在着显著差异的判断。

因此，方差分析就是通过不同方差的比较做出拒绝原假设或不能拒绝原假设的判断。组间方差和组内方差之比是一个统计量，该统计量服从第一自由度为  $r-1$ 、第二自由度为  $n-r$  的  $F$  分布。

在对数据进行方差分析时，应该满足以下两个前提条件。

- 应将各组观察数据看作是来自于正态分布总体的随机样本。该条件通常较容易满足。
- 各组观察数据是从具有相同方差的、相互独立的总体中抽取得到的,即具有同方差性。该条件通常要求在做方差分析之前应当进行同方差性检验。

根据所研究的因变量数目不同,方差分析可以分为一元方差分析和多元方差分析。在分析一个因变量时,根据影响因素数目不同,又可以进一步细分为单因素方差分析和多因素方差分析。如果影响因素中具有协变量,则称之为协方差分析。含有协变量的一元多因素方差分析在实际应用中非常常见。

在方差分析的过程中,不仅可以对因素是否对因变量产生显著影响进行检验,还可以通过对因素的多重比较来分析因素的哪个或哪些水平对因变量的影响最显著,也可以通过广义线性模型估计因素水平对因变量的具体影响。

## 4.2 单因素方差分析

单因素方差分析(One-Way ANOVA)主要研究单个因素对因变量的影响。通过因素的不同水平对因变量进行分组,计算组间和组内方差,利用方差比较的方法对各分组所形成的总体进行均值比较,从而对各总体均值相等的原假设进行检验。

### 4.2.1 单因素方差分析与方差同质性检验

本小节以例 4-1 为例进行单因素方差分析。在数据集 DC\_Sale.sas7bdat 中,因变量是“Sale”。在数码相机的生产过程中,可以对成像元器件的像素数进行人为调整,因此可以控制的定性变量便是像素数“Pixel”。其对因变量产生了影响,通常把它认为是影响因变量变动的一个因素。该因素有“500 万像素以下”、“500~600 万像素”、“600~800 万像素”、“800~1 000 万像素”、“1 000 万像素以上”5 个水平。

**STEP 1** 进入 SAS/Analyst, 打开 DC\_Sale.sas7bdat 数据集, 选择菜单“Statistics→ANOVA→One-Way ANOVA”, 打开图 4-1 所示的对话框。

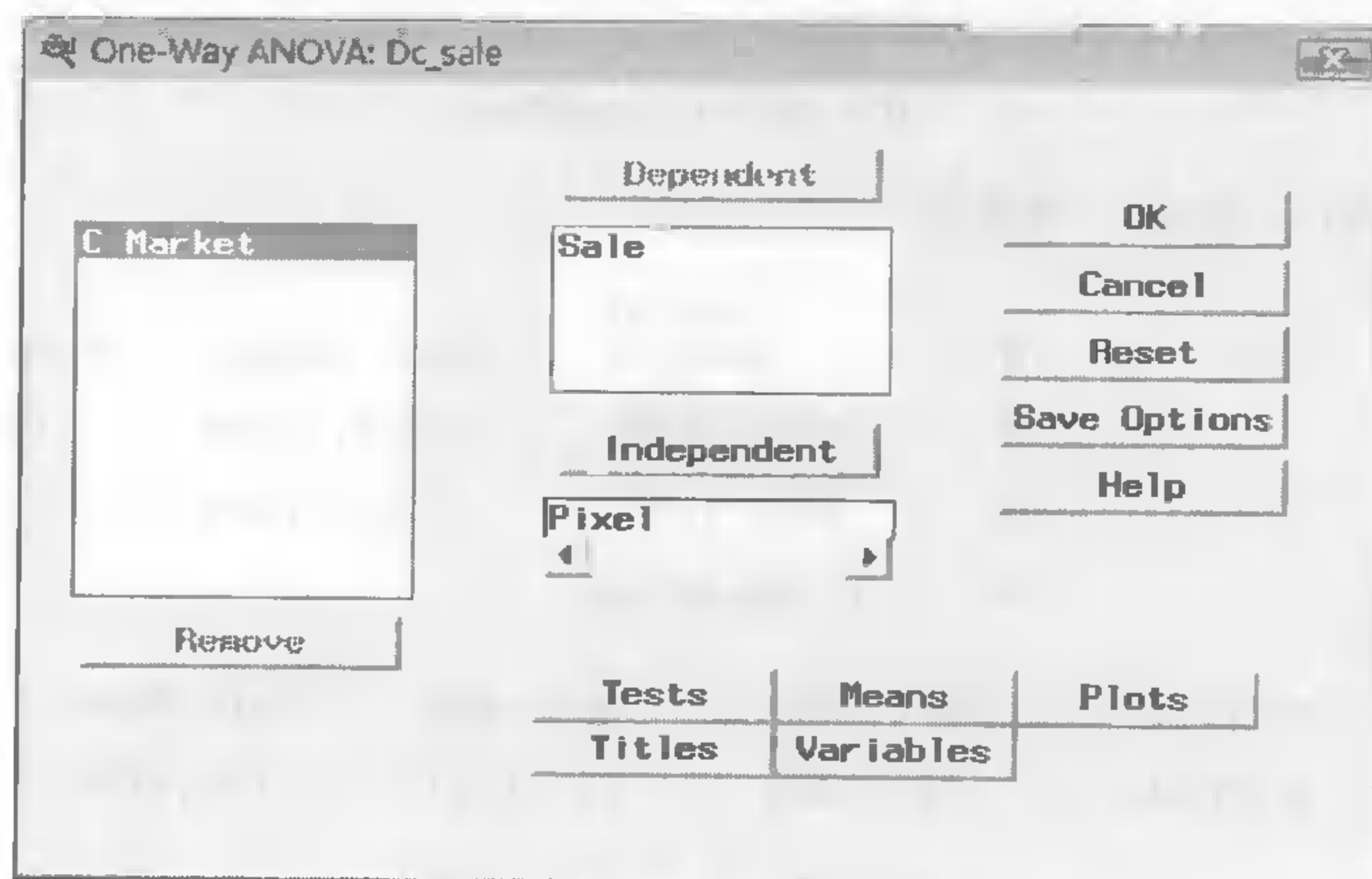


图 4-1 单因素方差分析对话框

**STEP 2** 在“One-Way ANOVA”对话框中部的变量选择区域中选中“Sale”变量, 单击

“Dependent” 按钮，将其指定为因变量；选中 “Pixel” 变量，单击 “Independent” 按钮，将其指定为因素。

**STEP 3** 单击 “Tests” 按钮，弹出图 4-2 所示的对话框。

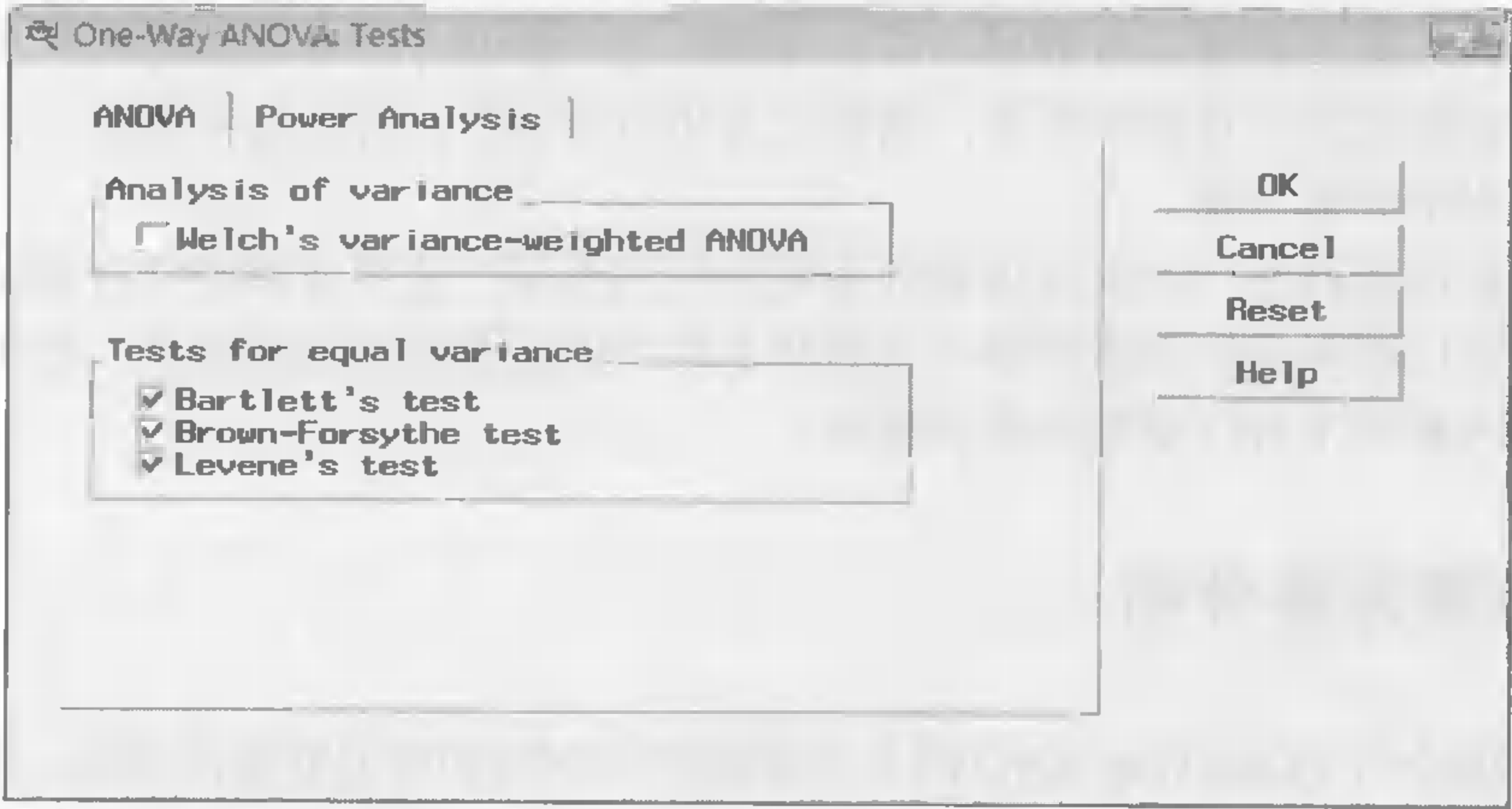


图 4-2 单因素方差分析的检验对话框

在该对话框中的 “Tests for equal variance” 分栏下，提供了 Bartlett’s 检验、Brown-Forsythe 检验和 Levene’s 检验等 3 种同方差性检验的方法，其中对于一元方差分析，常用 Levene’s t 检验方法，而多元方差分析大多使用 Bartlett’s 球形检验法。

3 种检验方法的原假设都是由因素水平所区分的各总体的方差相等。

$$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_r^2$$

其中， $r$  为分组数目或因素的水平个数。

**STEP 4** 为了对比各种检验方法对于本例结果的稳健性，在本例中，把 3 种方法前的复选框都选中，单击 “OK” 按钮后返回 “One-Way ANOVA” 对话框。在该对话框中，单击 “OK” 按钮，可得到图 4-3 所示的方差分析结果和图 4-4 所示的同方差性检验结果。

The ANOVA Procedure					
Dependent Variable: Sale    销售量					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	4	10472.85000	2618.21250	19.57	<.0001
Error	35	4682.12500	133.77500		
Corrected Total	39	15154.97500			
	R-Square	Coeff Var	Root MSE	Sale Mean	
	0.691050	10.90886	11.56611	106.0250	
Source	DF	Anova SS	Mean Square	F Value	Pr > F
Pixel	4	10472.85000	2618.21250	19.57	<.0001

图 4-3 单因素方差分析输出结果

The ANOVA Procedure					
Levene's Test for Homogeneity of Sale Variance ANOVA of Squared Deviations from Group Means					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Pixel	4	36442.6	9110.7	0.26	0.8988
Error	35	1205713	34448.9		
Brown and Forsythe's Test for Homogeneity of Sale Variance ANOVA of Absolute Deviations from Group Medians					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Pixel	4	60.7500	15.1875	0.23	0.9177
Error	35	2277.6	65.0750		
Bartlett's Test for Homogeneity of Sale Variance					
Source	DF	Chi-Square	Pr > ChiSq		
Pixel	4	1.1577	0.8850		
The ANOVA Procedure					
Level of Pixel	N	Mean	Std Dev		
1000万像素以上	8	122.125000	10.1480118		
500-600万像素	8	95.750000	13.5198267		
500万像素以下	8	81.125000	9.3874917		
600-800万像素	8	107.125000	12.4835606		
800-1000万像素	8	124.000000	11.7958831		

图 4-4 方差同质性检验结果

在图 4-3 所示的 “The ANOVA Procedure” 表格中，“Source” 列表示方差分析过程中方差的来源。

- Model: 组间差异。
- Error: 组内差异。
- Corrected Total: 总体差异，为组间差异与组内差异之和。
- DF: 表示各种差异的自由度。
- Sum of Squares: 表示各类差异所代表的离差平方和。
- Mean Square: 表示各类离差平方和除以其对应自由度得到的方差。
- F valae: 由组间方差与组内方差之比计算的 F 统计量值。
- Pr>F: F 统计量对应的 P 值。

由图 4-3 可知，本例中组间离差平方和为 10 472.85，组内离差平方和为 4 682.125，总离差平方和为 15 154.975，对应的组间方差为： $2\,618.2125 = 10\,472.85 / 4$ ，组内方差为： $133.775 = 4\,682.125 / 35$ 。

用于判定组内方差和组间方差是否存在差异的 F 值为： $19.57 = 2\,618.2125 / 133.775$ ，用于该差异是否显著的假设检验的 P 值（“Pr>F”）小于 0.0001，故远远小于  $\alpha(0.05)$ 。

根据方差分析的 P 值可以得到结论：在显著性水平  $\alpha = 0.05$  的条件下，可以拒绝各总体均值相等的原假设，接受备择假设，即数码相机成像元器件的像素数会对数码相机的销售量产生显著影响。

4.1 节提到过单因素方差分析的两个基本条件，其中方差同质性（即各组总体方差相等）条件是可根据样本数据进行检验的条件。在图 4-4 中，分别列示了 3 种检验方法在方差同质

性原假设成立的情况下被用于检验统计量和  $P$  值的情况。

在本例中，方差同质性的 Levene’s 检验  $P$  值 (“Pr>F”) 为 0.8988，Brown-Forsythe 检验  $P$  值 (“Pr>F”) 为 0.9177，Bartlett’s 检验  $P$  值 (Pr>ChiSq) 为 0.8850。由 3 种检验方法计算出的  $P$  值均远远大于显著性水平  $\alpha(0.05)$ ，因此在显著性水平  $\alpha = 0.05$  的条件下，没有充分理由拒绝方差相等的原假设，即可认为本例各总体的方差具有同质性，符合单因素方差分析的前提条件，前面做出的像素数目对销售量影响显著的结论是可靠的。

此外，在方差同质性检验的结果中，SAS 系统会自动计算出各分组因变量的样本量、样本均值和样本标准差。

至此，通过方差分析的方法得到了例 4-1 的结论，但是得到的结论只限于像素数目对销售量有显著影响这么一个表面认识上。

如果进一步研究究竟是因素的哪一个水平对观察变量产生了显著影响，即具体哪种像素数目对销售量有显著影响？这就是单因素方差分析的均值多重比较检验。

4.2.2 方差分析的多重比较

方差分析的多重比较在 SAS/Analyst 中在单因素方差分析对话框中可以实现。打开图 4-1 所示的单因素方差分析对话框，单击 “Means” 按钮，弹出图 4-5 所示的方差分析均值比较检验的对话框。

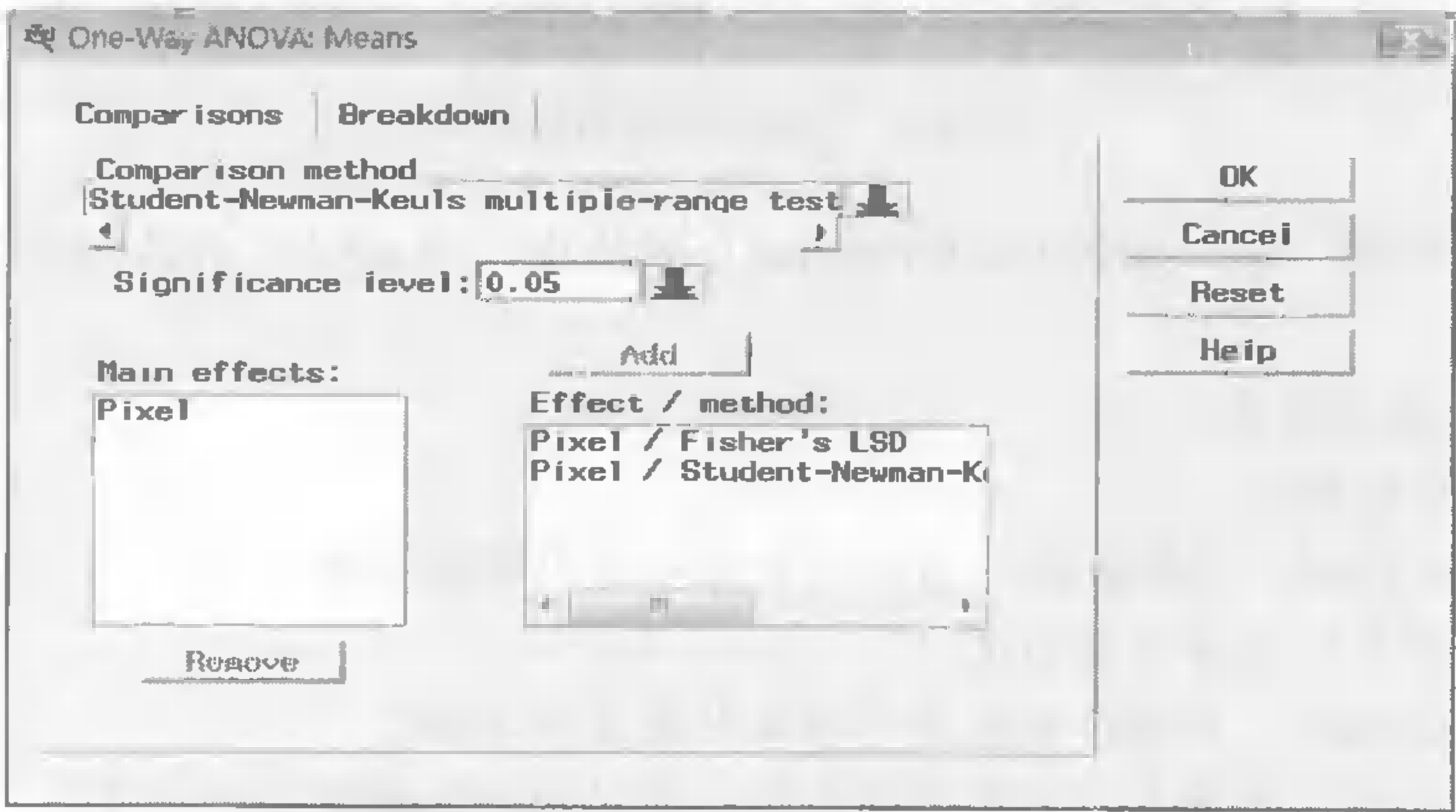



图 4-5 单因素方差分析中的 “Means” 对话框

在 “Means” 对话框中，可通过 “Comparisons” 选项卡指定多重比较的方法、显著性水平以及需要进行多重比较的因素，也可通过 “Breakdown” 选项卡输出数值型因变量的常用统计量，如均值、方差、标准差、最大值、缺失值样本量等。

选择 “Comparisons” 选项卡，单击 “Comparison method” 下拉列表框的最右边的  按钮，弹出可供选择的各种多重比较方法，如图 4-6 所示。

SAS/Analyst 中一共提供了 10 种典型的均值多重比较方法，可以选择其中的任意一种或多种方法进行分析。在实际应用中，比较常用的是 Fisher 的 LSD 均值比较检验方法。本例选择 “Fisher’s LSD” 方法和 “Student-Newman-Keuls multiple-range test”

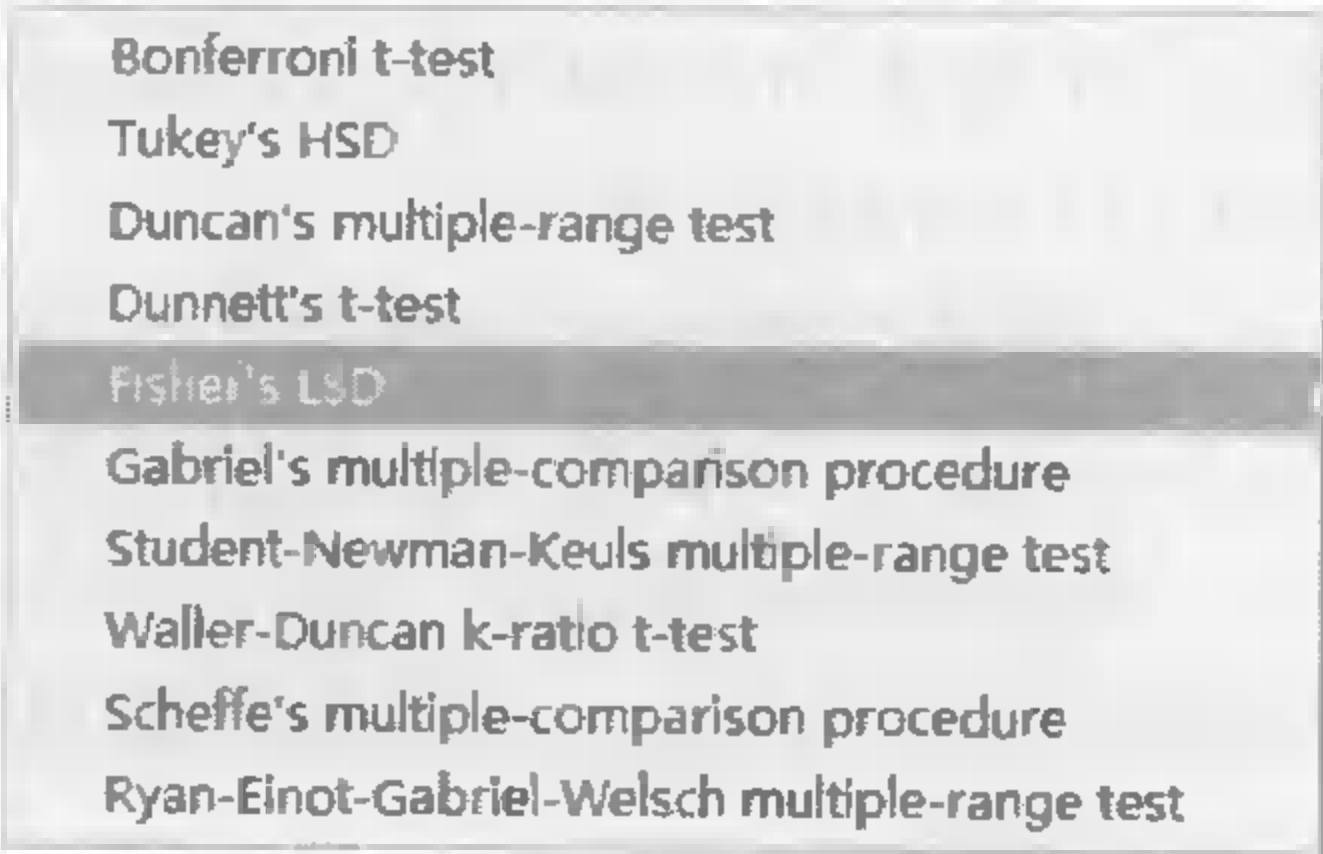


图 4-6 方差分析多重比较的各种方法

方法。

**STEP 1** 选择“Fisher’s LSD”方法，在“Significance level”的下拉列表框中填入或选择理论的显著性水平。本例根据给定的显著性水平选择 0.05。在“Main effects”区域中，选择“Pixel”，然后单击右边“Effect/method”区域上方的“Add”按钮，把“Pixel”变量指定为多重比较检验的因素以进行分析。

**STEP 2** 然后选择“Student-Newman-Keuls multiple-range test”方法，按照同样的方法设定好显著性水平和分析变量。

**STEP 3** 指定好检验方法、显著性水平和因素之后，单击“OK”按钮返回单因素方差分析对话框。在该对话框中，单击“OK”按钮。在 SAS 的输出结果中，除了包含图 4-3 和图 4-4 所示的结果之外，还包含图 4-7 所示的输出结果（节选）。

The ANOVA Procedure			
t Tests (LSD) for Sale			
Means with the same letter are not significantly different.			
t Grouping	Mean	N	Pixel
A	124.000	8	800-1000万像素
A			
A	122.125	8	1000万像素以上
B	107.125	8	600-800万像素
B			
B	95.750	8	500-600万像素
C	81.125	8	500万像素以下

The ANOVA Procedure			
Student-Newman-Keuls Test for Sale			
Means with the same letter are not significantly different.			
SNK Grouping	Mean	N	Pixel
A	124.000	8	800-1000万像素
A			
A	122.125	8	1000万像素以上
B	107.125	8	600-800万像素
B			
B	95.750	8	500-600万像素
C	81.125	8	500万像素以下

图 4-7 方差分析多重比较检验的结果

方差分析的多重比较结果在 SAS 系统中通常以变量分组的形式给出，系统自动给出诸如“A”、“B”、“C”标注的因素水平分组（Grouping）。处于同一分组之内的因素水平之间没有显著差异，处于不同分组之间的因素有明显差异。如在 LSD 检验中，系统会自动把像素水平分成 3 组，分别是“A”、“B”、“C”组。

A 组中有两个水平，分别是“800~1 000 万像素”和“1 000 万像素以上”，因为 A 组中的数码相机像素水平比较高，所以可以将其命名为“高像素组”；B 组中有 2 个水平分别是“600~800 万像素”和“500~600 万像素”，可以将其命名为“中像素组”；C 组中只有一个水平，即“500 万像素以下”，可以将其命名为“低像素组”。

A 组和 B 组内部的两个水平之间没有显著差异，即 A 组中的“800~1 000 万像素”和“1 000 万像素以上”之间对数码相机的销售量没有显著影响；B 组中的“600~800 万像素”和“500~600 万像素”之间对数码相机的销量没有显著影响；A、B、C 组之间对销量影响

显著。由于 C 组中只有一个水平，这个水平与 A 组和 B 组内的水平都对销量有显著的影响。因此，可以认为“500 万像素以下”这个水平对数码相机销售量影响最为显著。

观察数码相机市场的现状可以发现，500 万像素以下的数码相机由于其技术比较落后，消费者需求量不大，与中高像素的数码相机相比，销售量明显萎缩；消费者对于像素数量的要求不同，对销售量也产生了显著的影响，像素高的相机明显比低像素相机的销量大。

本例数据对于“Student-Newman-Keuls multiple-range test”检验方法的结果和分析过程与 LSD 方法一致。

至此，已经通过单因素方差分析的多重比较检验得知对销量影响较大的因素水平，但是这些水平究竟是怎样具体影响因变量，即当从一个水平变为另一个水平时，因变量会增加或降低多少个单位？这需要利用广义线性模型（GLM, Generalized Linear Modeling）对方差分析模型进行参数估计。

4.2.3 方差分析模型的参数估计和预测

在 4.1 节中的原理性解释中可看到，单因素方差分析实际上是对一个线性模型(Linear Model)进行分析。故在 SAS 系统中，除应用系统菜单“SAS/Analyst→Statistics→ANOVA→One-Way ANOVA”进行分析之外，还可利用系统菜单“SAS/Analyst→Statistics→ANOVA→Linear Models”进行分析。

利用“Linear Models”菜单进行方差分析的好处在于：除提供上述完整的方差分析过程之外，还可以对 4.1 节中的线性模型进行参数估计和假设检验，而且根据参数估计的结果可以得出因素水平之间的变动状况对因变量产生的具体影响。此外，利用该菜单还可以进行多因素方差分析、协方差分析和多元方差分析。

**STEP 1** 进入 SAS/Analyst，打开 DC\_Sale.sas7bdat 数据集，选择系统菜单“Statistics→ANOVA→Linear Models”，弹出方差分析线性模型对话框，如图 4-8 所示。

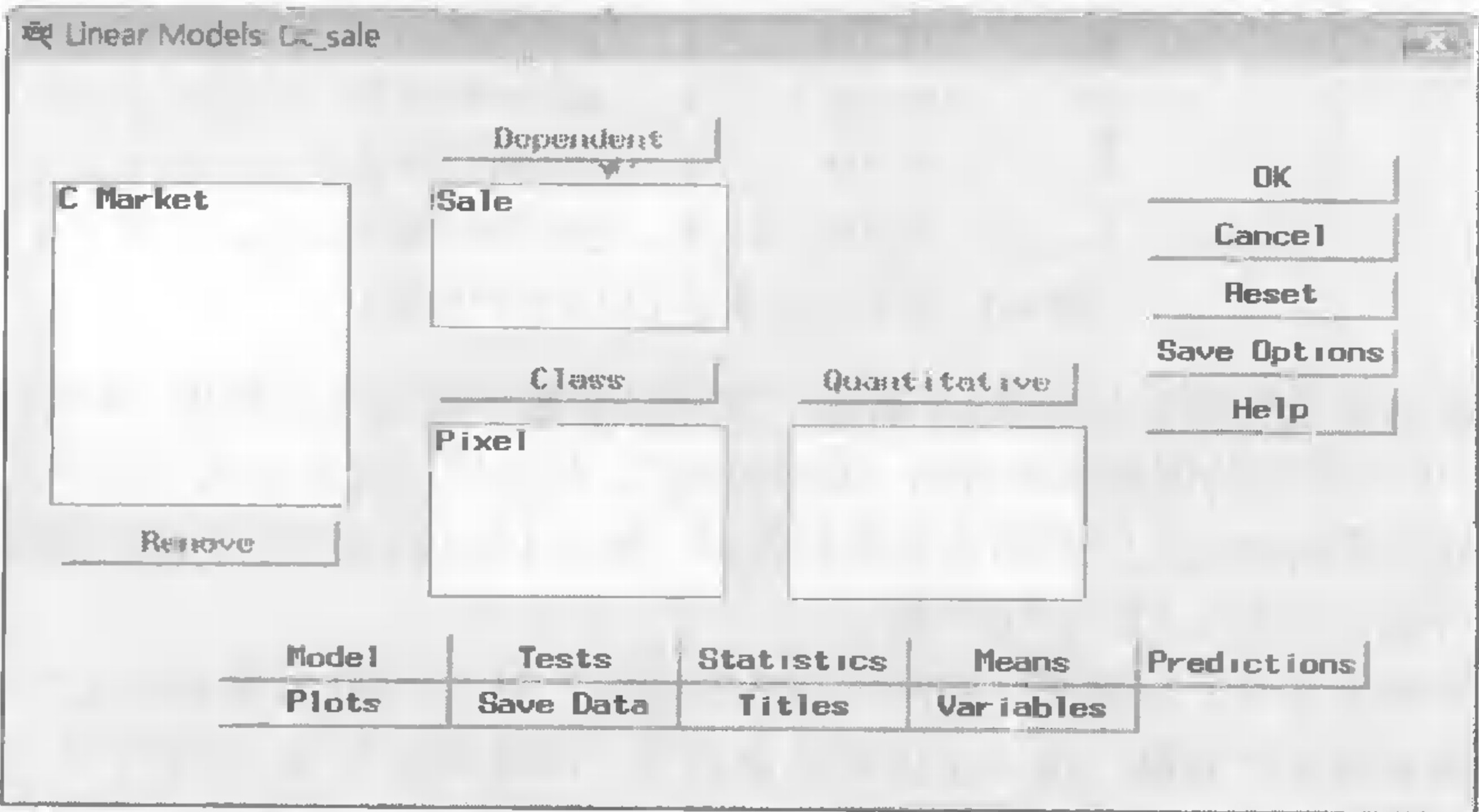


图 4-8 “Linear Models”对话框

**STEP 2** 选中“Sale”变量，单击“Dependent”按钮，将其指定为因变量；选中“Pixel”变量，单击“Class”按钮，将其指定为因素。单击“Means”按钮可以设置方差分析的多重比较检验，设置方法如图 4-5 所示。单击“Statistics”按钮，弹出图 4-9 所示的统计量对话框。

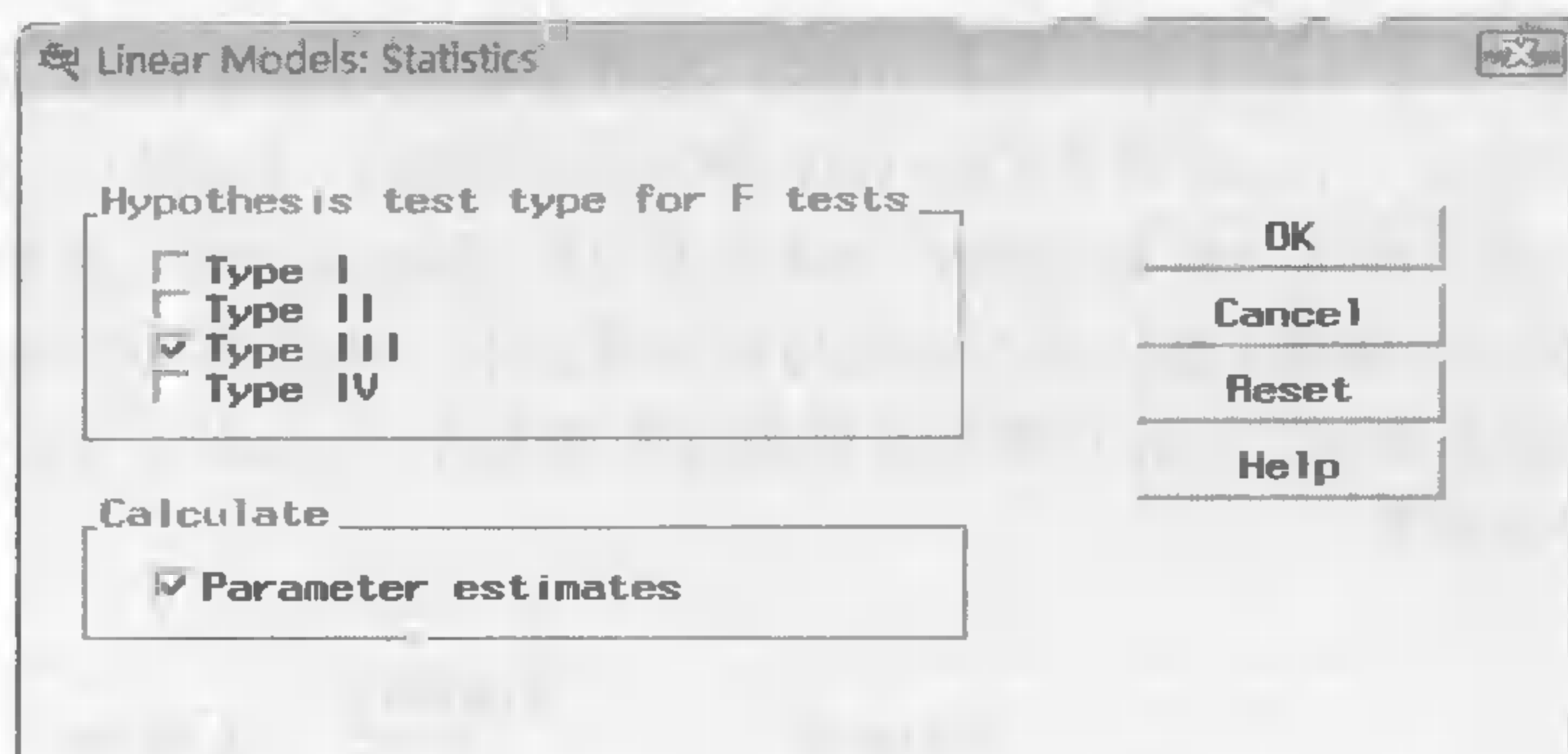


图 4-9 “Statistics”对话框

**STEP 3** 在方差分析中，对 F 统计量通常做“Type III”型检验，在“Hypothesis is test type for F tests”栏下选择默认的“Type III”复选框即可。本例要对模型进行参数估计，因此在“Calculate”栏下选中“Parameter estimates”，单击“OK”按钮返回图 4-8 所示对话框。

**STEP 4** 此外，在图 4-8 所示的对话框中单击“Save Data”按钮可以根据样本数据计算样本统计量并将其存储在数据集中，单击“Plot”按钮可以绘制因变量与因素之间的点线图，单击“Predictions”可以对方差分析模型进行预测。单击“Predictions”按钮，弹出图 4-10 所示的预测对话框。

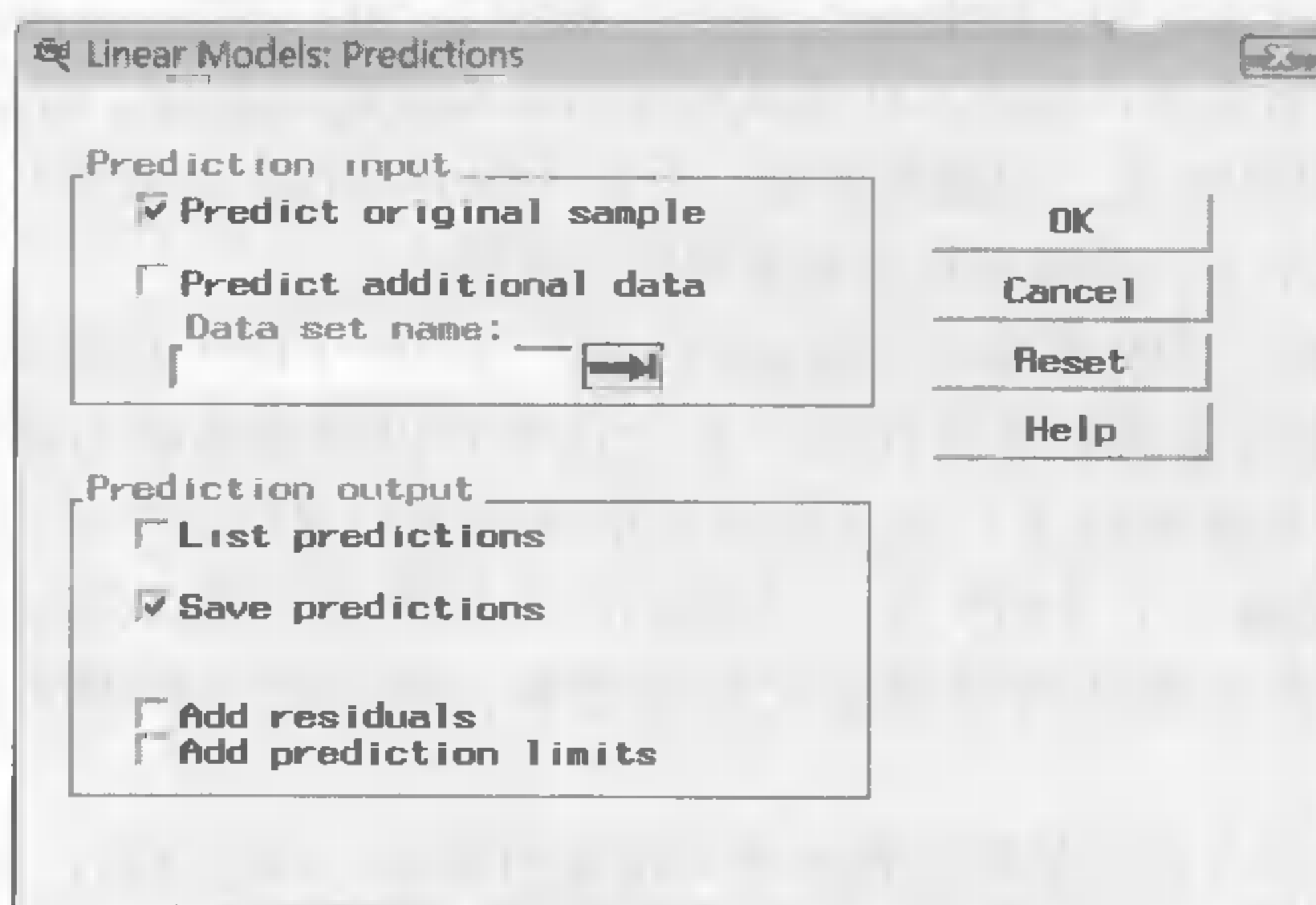
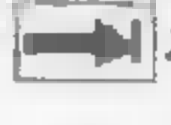


图 4-10 方差分析预测对话框

**STEP 5** 在该对话框中，可以通过“Prediction input”指定被用于预测的数据来源。如果想验证模型的预测效果，用预测值与实际观测值进行对比分析，则可以选择“Predict original sample”复选框进行预测；如果想利用方差分析模型对新的数据进行预测，则可以选择“Predict additional data”复选框，并在该复选框下的“Data set name”的文本输入框中输入新 SAS 数据集的路径和数据集名，或单击  按钮，浏览指定具体数据集即可。在“Prediction output”分栏下，可以选择预测结果的显示方式，“List predictions”复选框表示在 SAS/Analyst 的结果输出窗口中显示预测结果；“Save predictions”复选框表示在所用数据集中存储预测的结果。

**STEP 6** 此外，还可以利用“Add residuals”复选框计算模型的残差项，利用“Add

predictions limits”指定预测值所处的区间。本例选择利用方差分析的原始数据进行预测，并把预测结果存储于原始数据集当中。单击“OK”按钮返回 Linear Models 对话框。在该对话框中，单击“OK”按钮，可以得到用 One-Way ANOVA 得到的、相同的方差分析与多重比较检验的输出结果，模型预测的结果会被自动保存在 DC\_Sale.sas7bdat 数据集中。可以返回 SAS/Analyst 主界面进行查看（在主界面左边的结果索引中，双击“Predictions Table”便可弹出含有预测值的数据集数据，系统自动命名预测值变量名为“\_pred”）。此外，还可以得到图 4-11 所示的参数估计结果。

Parameter		Estimate	Standard Error	t Value	Pr >  t
Intercept		124.0000000 B	4.08923893	30.32	<.0001
Pixel	1000万像素以上	-1.8750000 B	5.78305715	-0.32	0.7477
Pixel	500-600万像素	-28.2500000 B	5.78305715	-4.88	<.0001
Pixel	500万像素以下	-42.8750000 B	5.78305715	-7.41	<.0001
Pixel	600-800万像素	-16.8750000 B	5.78305715	-2.92	0.0061
Pixel	800-1000万像素	0.0000000 B	.	.	.

图 4-11 方差分析的参数估计结果（含截距项）

在进行参数估计结果的分析之前，必须先弄清楚图 4-11 所示结果的含义。

图 4-11 中的“Parameter”列给出了截距项和因素 5 个水平的参数，“Estimate”列则给出了 GLM 方法估计的各个具体参数估计值，其余各列分别给出了估计的标准误差和参数显著性检验的统计量值及其对应的  $P$  值（“Pr>|t|”）。

首先找到参数估计值，即“Estimate”列对应数值为“0”的行，表示水平为“800~1 000 万像素”的参数估计值为 0，这并不代表该水平对因变量毫无影响，而是代表模型中截距项（即“Intercept”）的具体含义，即截距项表示水平“800~1 000 万像素”对因变量（销售量）的影响。在该像素水平下，数码相机的销售量为 124 台。

而其他水平对因变量的影响均以截距项（即“800~1 000 万像素”相机的销量）为基准来衡量，其对应的参数估计值代表了各个水平对因变量影响与截距项对因变量影响的差距。如“1 000 万像素以上”水平的参数估计值为-1.875，表示与截距项相比，该水平下的数码相机销量减少了 1.875 台，其对应的  $P$  值为 0.7477，远远小于  $\alpha(0.05)$ ，表示这两个像素水平之间对数码相机销量无显著影响，该水平下的数码相机销量具体为： $124-1.875 = 122.125$  台。

可以用同样的方法分析其他的因素水平对销量的影响。通过分析，发现“500 万像素”水平与截距项相比，对销量的影响最大，这与前面分析的多重比较检验的分类结果一致。与截距项基准相比，该水平下的销量降低了 42.875 台，其对应的  $P$  值小于 0.001，远远小于  $\alpha(0.05)$ ，具有非常显著的差异，在该水平下的销量为： $124-42.875 = 81.125$  台。

在 SAS 系统中的默认情况下，给出图 4-11 的方差分析模型参数估计的相对比较结果，对于各种不同因素水平，对因变量影响的差异分析非常直观明了。但是如果根据这个结果来考察因素，对因变量影响的绝对数值还需用户自行计算，并根据估计参数数值为“0”的水平确定比较基准，然后计算具体因素水平对因变量的绝对影响，在因素水平非常多的情况下显得十分麻烦。为了避免用户进行繁琐计算，SAS 系统在进行参数估计时提供了可供选择的结果输出方式。

**STEP 1** 在图 4-8 所示的对话框中，单击“Model”按钮，弹出方差分析模型设置对话框，如图 4-12 所示。

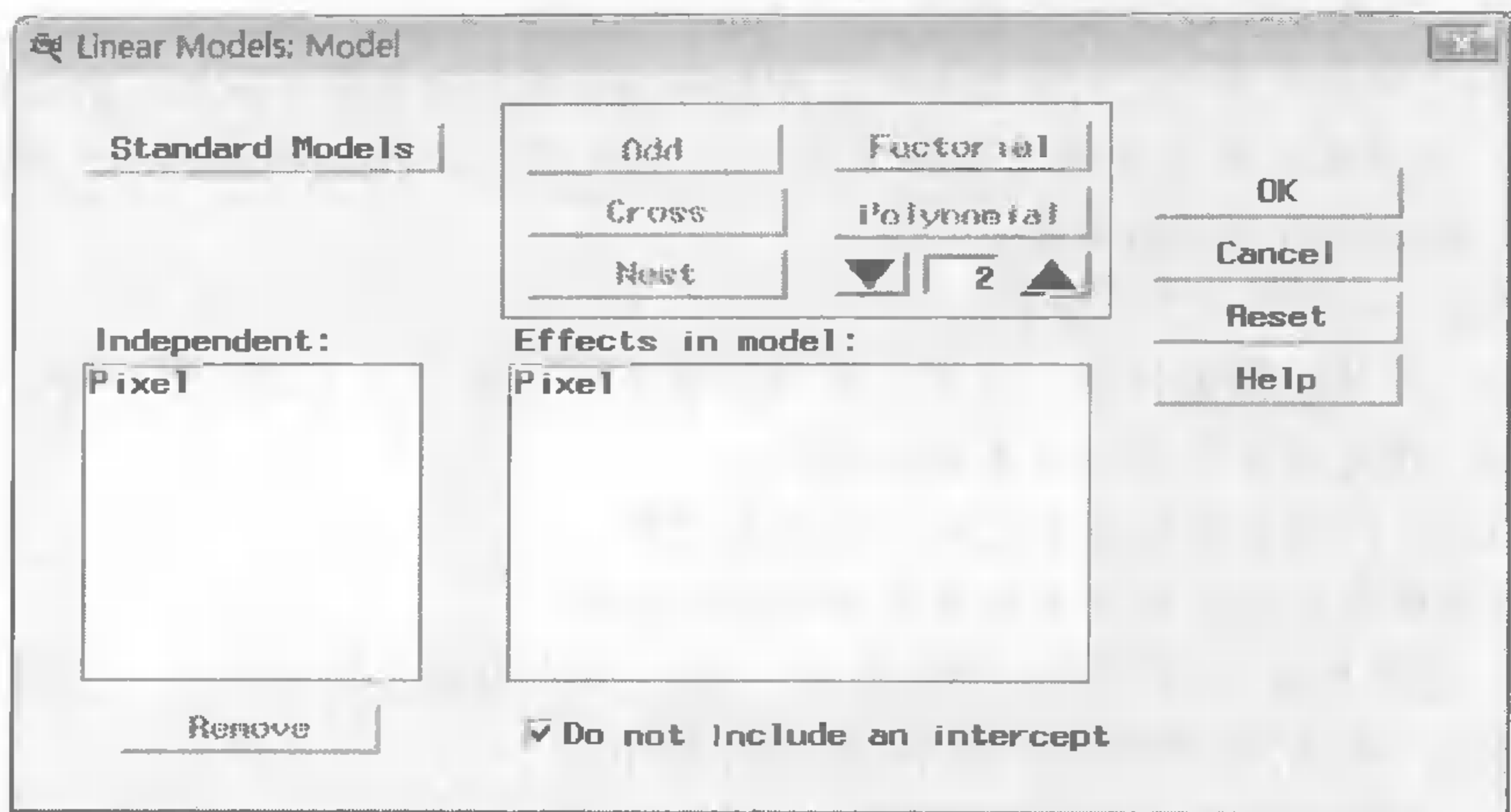


图 4-12 方差分析模型设置对话框

**STEP 2** 在图 4-12 中，“Do not include an intercept”复选框可以指定模型输出的结果形式，默认不选中该复选框，即模型中包含截距项，输出图 4-11 所示的参数估计结果。如果要输出各种因素水平对因变量的绝对影响，则必须选中该复选框。单击“OK”按钮返回图 4-8 所示的对话框。在该对话框中，单击“OK”按钮，便可得到不包含截距项的参数估计结果，如图 4-13 所示。

Parameter		Estimate	Standard Error	t Value	Pr >  t
Pixel	1000万像素以上	122.1250000	4.08923893	29.86	<.0001
Pixel	500-600万像素	95.7500000	4.08923893	23.42	<.0001
Pixel	500万像素以下	81.1250000	4.08923893	19.84	<.0001
Pixel	600-800万像素	107.1250000	4.08923893	26.20	<.0001
Pixel	800-1000万像素	124.0000000	4.08923893	30.32	<.0001

图 4-13 方差分析的参数估计结果（不含截距项）

不含截距项的参数估计结果代表了因素水平对因变量的绝对影响，其具体影响数值与含有截距项用户自行计算出的绝对数值一致。

从上述分析结果来看，高像素（800 万以上）数码相机的销售量比较大，而中低像素（500～800 万）数码相机的销售量一般，低像素（500 万以下）数码相机的销售量最小。

可以用 SAS 编程中的 ANOVA 过程进行单因素方差分析。ANOVA 过程的具体语法如下：

```
proc anova <选项>;
  class 变量 </选项>;
  model 因变量=因素或协变量 </选项>;
  absorb 变量;
  by 变量;
  freq 变量;
  manova <检验选项> </选项>;
  means 因素或协变量 </选项>;
  repeated 因素详单 </选项>;
  test <h=效应> e=效应;
```

具体语句功能解释如下：

- absorb: 在模型中合并分类效应。
- by: 指定分组变量以进行分组分析, 即按照 BY 指定的变量对各组分别进行方差分析。
- class: 指定定性因素变量或分类变量。CLASS 语句必须放在 MEANS 语句之前。
- freq: 指定作为频数的变量。
- manova: 执行多元方差分析。
- means: 计算比较统计量。通常可进行均值多重比较、方差同质性检验。
- model: 指定用于拟合的方差分析模型。
- repeated: 执行多重测度多元和一元方差分析。
- test: 使用指定效应的平方和残差进行假设检验。

ANOVA 过程不能对模型进行参数估计。在一元单因素方差分析中, 该过程主要用到 class、MODEL、MEANS 等语句。例 4-1 的程序如下。

```
proc anova data=Sasuser.DC_Sale; /*调用 ANOVA 过程, 并使用 Sasuser.DC_Sale 数据集进行分析*/
  class pixel; /*指定变量 "pixel" 作为分类的因素变量*/
  model sale=pixel; /*指定变量 "sale" 作为因变量、"pixel" 作为影响因素*/
  means pixel /hovtest=levене lsd; /*关键字 "hovtest=" 用于指定方差同质性的检验方法, 其取值为
  "levене", 表示使用 "Levene's test" 检验法; 关键字 "lsd" 表示按照分类变量 "pixel" 进行均值比较, 方
  法为 LSD*/
run;
```

运行程序后, 可以得到图 4-3、图 4-4 和图 4-7 所示的类似结果。

此外, 还可以用 GLM 过程进行单方差分析, 并且得到参数估计的结果和预测值。

GLM 过程在 SAS 编程中应用非常广泛, 可被应用于回归分析、方差分析、偏相关分析、也可以被用于对定性变量进行分析。其工作原理是用最小二乘法 (Least Square) 对线性模型进行参数估计。其具体语法如下:

```
proc glm <选项>;
class variables </选项>;
model 因变量=自变量 </选项>;
absorb 变量;
by 变量;
freq 变量;
id 变量;
weight 变量;
contrast '标签' 效应值 <... 效应值> </选项>;
estimate '标签' 效应值 <... 效应值> </选项>;
lsmeans 效应 </选项>;
manova <检验选项> </详细选项>;
means 效应 </选项>;
output <out=输出数据集>
  关键字=数据集中的变量名 <... 关键字=数据集中的变量名> </选项>;
random 效应 </选项>;
repeated 因素详单 </选项>;
test <h=效应> e=效应 </效应>;
```

具体语句功能解释如下:

- absorb: 在模型中合并分类效应。
- by: 指定分组变量以进行分组分析, 即按照 by 指定的变量对各组分别进行方差分析。
- class: 指定定性因素变量或分类变量。class 语句必须放在 means 语句之前。
- contrast: 构造和检验线性模型的参数。
- estimate: 对线性模型进行参数估计。
- freq: 指定作为频数的变量。
- id: 在输出结果中指定观测变量。
- lsmeans: 计算最小二乘法(边际)均值。
- manova: 执行多元方差分析。
- means: 计算比较统计量。通常可进行均值多重比较、方差同质性检验。
- mode: 指定用于拟合的方差分析模型。
- output: 指定输出结果存储于一个数据集中。
- random: 指定模型中存在随机效应并计算期望均方。
- repeated: 执行多重测度多元和一元方差分析。
- test: 使用指定效应的平方和残差进行假设检验。
- weight: 指定作为权数的变量。

GLM 过程在一元单因素方差分析中, 主要用到 class、model、means、estimate 等语句。

例 4-1 的程序如下。

```
proc glm data=Sasuser.Dc_sale;      /*调用 GLM 过程, 并打开 Sasuser.Dc_sale 数据集*/
  class pixel;                      /*指定变量“pixel”作为因素*/
  model sale = pixel / solution;    /*指定变量“sale”作为因变量、“pixel”作为影响因素, 并估计
方差分析模型的参数*/
  means pixel / hovtest=levene LSD; /*关键字“hovtest=”指定方差同质性的检验方法, 其取值为
“levene”表示使用“Levene’s test”检验法; 关键字“lsd”表示用 LSD 法按照分类变量“pixel”进行均值
比较*/
run;
```

如果需要估计参数的绝对数值(程序默认模型包含截距项), 即估计不含截距项模型的参数, 可以在上述程序的 model 语句中加上“noint”关键字。

```
model sale = pixel / noint solution;
```

运行程序之后, 在“Output”窗口中可得到用 ANOVA 过程和 SAS/Analyst 分析得到的同样结果。其结果的索引保存在“Results”窗口中, 可以用鼠标双击对应的表格名称, 在“Output”窗口中打开对应的详细结果。

### 4.3 多因素方差分析

当有两个或者两个以上的因素对因变量产生影响时, 可以用多因素方差分析的方法来进行分析。多因素方差分析的原理与单因素方差分析基本一致, 也是利用方差比较的方法, 通过假设检验的过程来判断多个因素是否对因变量产生显著性影响。

在多因素方差分析中, 由于影响因变量的因素有多个, 其中某些因素除了自身对因变量产生影响之外, 它们之间也有可能共同对因变量产生影响。在通常情况下, 把因素单独对

因变量产生的影响称为“主效应”：把因素之间共同对因变量产生的影响或因素某些水平同时出现时除了主效应之外的附加影响称为“交互效应”。

以例 4-1 来说，如果同时考虑成像元器件的像素数和镜头的光学变焦倍数两个因素对销售量的影响，当只考虑主效应时，假定 800~1 000 万像素的相机的销售量可以比其他相机的销售量多 20 台，而光学变焦为 10 倍的相机的销售量可以比其他相机的销售量多 15 台。当因素没有交互效应时，则像素数为 800~1 000 万且光学变焦为 10 倍的数码相机可以多卖出 35(20+15)台。但如果因素存在交互效应，那么同时考虑像素数为 800~1 000 万、光学变焦为 10 倍两个因素，则会产生交互效应，此时多出的销售量就不一定是 35 台了。

因此，多因素方差分析不仅要考虑每个因素的主效应，往往还要考虑因素之间的交互效应。多因素方差分析往往假定因素与因变量之间的关系是线性关系。因此，从这个方面来说，方差分析的模型也是下面线性模型的延续。

因变量 = 因素 1 主效应 + 因素 2 主效应 + ... + 因素 *n* 主效应 + 因素交互效应 1 + 因素交互效应 2 + ... + 因素交互效应 *m* + 随机误差，所以多因素方差分析往往选用广义线性模型 (General Linear Model) 进行参数估计。

4.3.1 只考虑主效应的多因素方差分析

在只考虑主效应的多因素方差分析模型中，只有因素自身对因变量的独立影响，不含任何交互作用对因变量的影响。



例 4-2

开发商开发某个住宅楼盘时，应当根据购房者住房面积的实际需求来设计和建造房屋。而影响购房者房屋使用面积需求的原因有很多。经过前期市场调研和分析，发现学历、购房者所在单位类型、收入水平和户型等几个因素对购房面积需求影响比较显著。现根据某开发商的委托，对上述因素影响购房面积的情况，随机抽取了 472 个样本进行抽样调查（数据详见 House.sas7bdat），简要情况如表 4-3 所示。试根据样本数据分析各种因素对购房面积的影响。设显著性水平  $\alpha = 0.05$ 。

表 4-3 购房面积及其影响因素简况

因 变 量	因 素 名 称	变 量 名	水 平
购房使用面积 (space)	学历	education	4 个水平：初中及以下、高中（中专）、大学（专、本科）、研究生及以上
	单位	unit	6 个水平：大专院校科研单位、国营企业、私营企业、行政事业单位、其他、失业
	年收入	income	5 个水平：10 000 元以下、10 000~25 000 元、25 000~50 000 元、50 000~75 000 元、75 000 元以上
	户型	type	11 个水平：两室一厅、两室两厅、三室一厅、三室两厅、三室三厅、四室两厅单卫、四室二厅双卫、四室三厅单卫、更大户型

首先考虑各个因素对因变量的影响只有主效应而没有交互效应的情况。

**STEP 1** 进入 SAS/Analyst，打开 House.sas7bdat 数据集，选择系统菜单 “Statistics → ANOVA → Linear Model”，弹出图 4-14 所示的对话框。

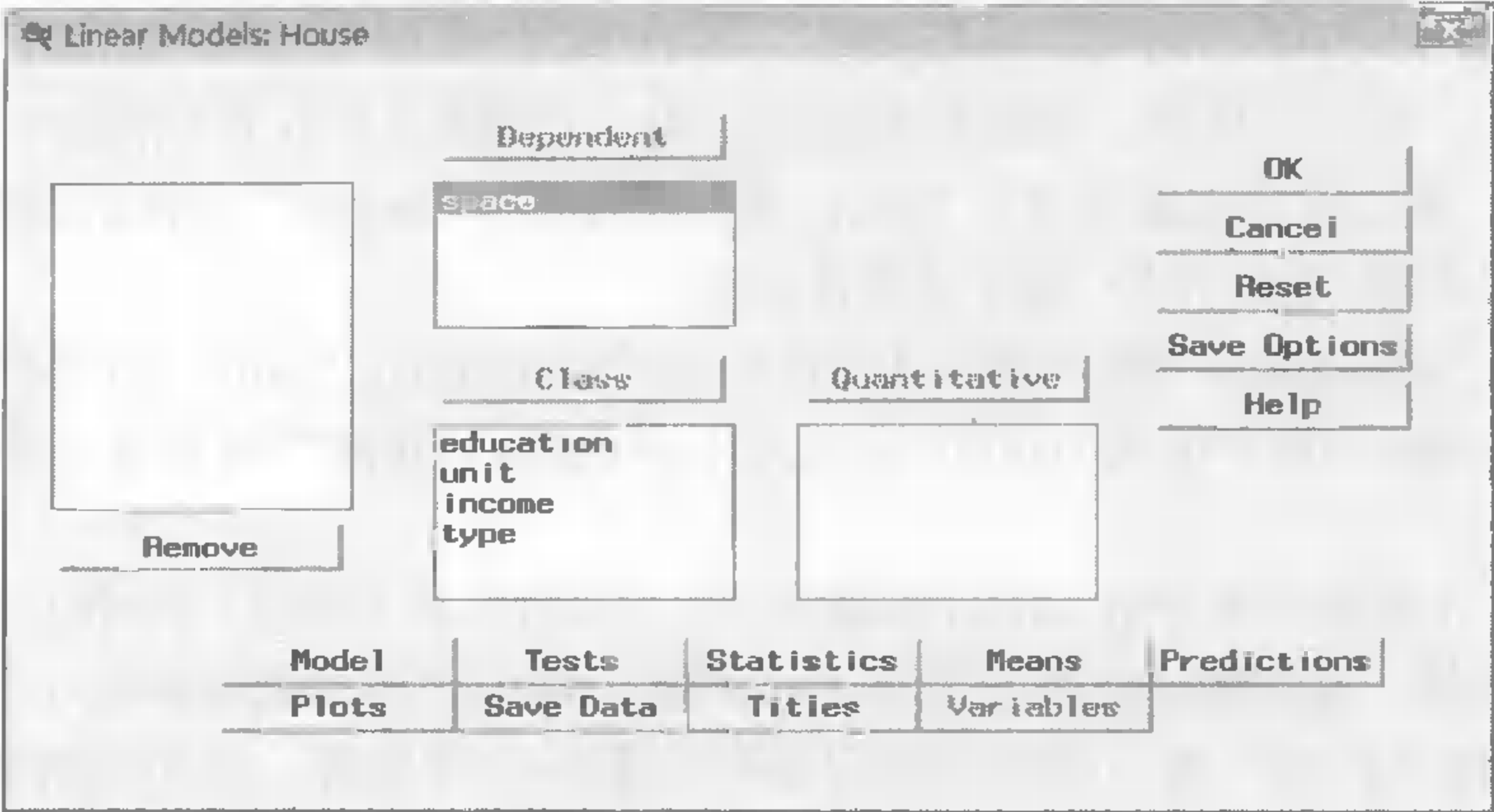


图 4-14 多因素方差分析（“Linear Models”）对话框

**STEP 2)** 在该对话框中部的变量选择区域中，选中“space”变量，单击“Dependent”按钮将其指定为分析的因变量。同时选中“education”、“unit”、“income”和“type”变量，单击“Class”把它们指定为因素变量。图 4-14 所示对话框默认是指考虑因素对因变量主效应的方差分析。单击“Means”按钮弹出类似图 4-5 所示的对话框，设置利用 10 种方法对指定因素进行均值多重比较。单击“OK”按钮返回图 4-14 所示的多因素方差分析对话框。在该对话框中单击“OK”按钮，可以得到方差分析的结果，如图 4-15 所示。

The GLM Procedure						
Class Level Information						
Class	Levels	Values				
education	4	初中及以下 大学（专、本科） 高中（中专） 研究生及以上				
unit	6	大专院校科研单位 国营企业 其它 失业 私营企业 行政事业单位				
income	5	10000~25000元 10000元以下 25000~50000元 50000~75000元 75000元以上				
type	11	更大户型 两室两厅 两室一厅 三室两厅 三室三厅 三室一厅 四室二厅双卫 四室两厅单卫 四室三厅单卫 四室三厅双卫 一室一厅				
		Number of Observations Read		472		
		Number of Observations Used		472		
The GLM Procedure						
Dependent Variable: space		购房面积				
Source	DF	Squares	Mean Square	F Value	Pr > F	
Model	22	32783.3469	1490.1521	4.66	<.0001	
Error	449	143477.4602	319.5489			
Corrected Total	471	176260.8070				
		R-Square	Coeff Var	Root MSE	space Mean	
		0.185993	25.02275	17.87593	71.43873	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
education	3	1519.40199	506.46733	1.58	0.1923	
unit	5	886.37145	177.27429	0.55	0.7347	
income	4	10545.81603	2636.45401	8.25	<.0001	
type	10	9604.42884	960.44288	3.01	0.0011	

图 4-15 多因素方差分析的结果

多因素方差分析的结果首先给出的是用户指定因素的一些基本描述，具体描述了各个因素的水平数量及具体水平，以及进行分析的样本量。然后在“**The GLM Procedure**”表格中首

先列示了模型整体上方差之比的显著性检验，其  $P$  值 (“Pr>F”) 为 “<.0001”，表示其  $P$  值远远小于 0.0001，即非常显著。但是模型拟合优度（详见 6.4.3 小节）的  $R^2$ ，即 “R-Square” 值为 0.185993，表示拟合程度不高；其次，该表格中的 “Source” 列表示因素对因变量主效应，可以通过其  $P$  值 (“Pr>F”) 进行显著性检验。

在本例中，“education” 的  $P$  值为 0.1923，大于  $\alpha(0.05)$ ；“unit” 的  $P$  值为 0.7347，大于  $\alpha(0.05)$ ；“income” 的  $P$  值为 0.0001 远远大于  $\alpha(0.05)$ ；“type” 的  $P$  值为 0.0011 远远小于  $\alpha(0.05)$ 。

由此可知，在显著性水平  $\alpha = 0.05$  的条件下，“学历” 和 “单位” 两个因素对购房面积因变量的影响不显著，说明购房面积的大小受不同学历和工作单位的影响较小。

“年收入” 和 “户型” 两个因素对购房面积的影响非常显著，说明消费者在考虑所购房屋面积时，主要考虑自身收入的实际情况和开发商户型设计两个重要因素。因此，开发商在建造房屋时应当根据消费者的实际购买能力设计户型合理的房子，才能够得到更多的收益。

那么，既然能够得出影响购房面积的两个重要影响因素，具体应当根据何种收入开发何种户型的房屋，才能够对消费者购房面积产生显著影响呢？换句话说，因素的哪个（些）水平对因变量的影响最大呢？首先，可以通过对因素多重比较检验进行分析。

**STEP 3** 在图 4-14 所示的对话框中，单击 “Means” 按钮弹出多重检验对话框。对 “income” 和 “type” 变量利用 Fisher’s LSD 方法进行多重比较检验，其设置过程和图 4-5 相同。返回图 4-14 之后单击 “OK” 按钮，除可以得到图 4-15 所示的结果外，还可以得到图 4-16 所示的多重比较检验结果。

The GLM Procedure					
t Tests (LSD) for space					
Comparisons significant at the 0.05 level are indicated by ***.					
income Comparison		Difference Between Means	95% Confidence Limits		
75000元以上	- 50000~75000元	19.151	4.997	33.306	***
75000元以上	- 25000~50000元	23.660	12.517	34.804	***
75000元以上	- 10000~25000元	31.278	20.443	42.112	***
75000元以上	- 10000元以下	35.884	24.756	47.012	***
50000~75000元	- 75000元以上	-19.151	-33.306	-4.997	***
50000~75000元	- 25000~50000元	4.509	-5.498	14.516	
50000~75000元	- 10000~25000元	12.126	2.465	21.788	***
50000~75000元	- 10000元以下	16.733	6.743	26.722	***
25000~50000元	- 75000元以上	-23.660	-34.804	-12.517	***
25000~50000元	- 50000~75000元	-4.509	-14.516	5.498	
25000~50000元	- 10000~25000元	7.618	3.474	11.761	***
25000~50000元	- 10000元以下	12.224	7.363	17.084	***
10000~25000元	- 75000元以上	-31.278	-42.112	-20.443	***
10000~25000元	- 50000~75000元	-12.126	-21.788	-2.465	***
10000~25000元	- 25000~50000元	-7.618	-11.761	-3.474	***
10000~25000元	- 10000元以下	4.606	0.504	8.709	***
10000元以下	- 75000元以上	-35.884	-47.012	-24.756	***
10000元以下	- 50000~75000元	-16.733	-26.722	-6.743	***
10000元以下	- 25000~50000元	-12.224	-17.084	-7.363	***
10000元以下	- 10000~25000元	-4.606	-8.709	-0.504	***
Comparisons significant at the 0.05 level are indicated by ***.					
type Comparison		Difference Between Means	95% Confidence Limits		
更大户型	- 一室一厅	6.667	-13.616	26.949	
更大户型	- 四室两厅单卫	9.938	-3.882	23.759	
更大户型	- 四室二厅双卫	13.468	1.410	25.527	***
更大户型	- 三室两厅	21.986	11.450	32.523	***
更大户型	- 两室两厅	26.747	15.622	37.872	***
更大户型	- 三室一厅	27.135	16.482	37.788	***
更大户型	- 两室一厅	28.203	17.318	39.087	***

图 4-16 多因素方差分析的多重比较检验结果

在多重比较检验结果中，系统会自动按照因素的水平进行两两配对的比较。如果配对的两种水平对因变量的影响显著（默认的显著性水平为 0.05），则会自动在该配对行的最后一列标注“\*\*\*”号。本例由于因素的水平太多，水平之间配对的多重比较数目过多，对于图 4-15 所示的输出结果，本书进行了适当删节。

对于“income”（收入）因素而言，“75 000 元以上”水平配对的多重比较检验结果有 4 个显著（即对“\*\*\*”号进行计数），“50 000~75 000 元”水平的多重比较检验有 3 个显著，“25 000~50 000 元”水平的多重比较检验结果有 3 个显著，“10 000~25 000 元”和“10 000 元以下”两个水平的多重比较检验结果各有 4 个显著。因此，在显著性水平  $\alpha = 0.05$  的条件下，可以认为低收入群体和高收入群体对购房面积最敏感（影响最大），而中收入群体同样对购房面敏感，但其敏感性不如其他两个收入群体。

而对于“type”（户型）因素而言，同样可以通过上述分析方法找出最有影响的水平。通过分析，可以从 SAS 输出结果中统计出各种水平多重比较显著的个数，如表 4-4 所示。

表 4-4                      户型因素的各种水平多重比较检验结果分析

户    型	多重比较显著的数目
更大户型	8
两室一厅	5
三室两厅	5
三室一厅	5
四室两厅双卫	5
四室两厅单卫	5
两室两厅	4
一室一厅	3
四室三厅双卫	2
三室三厅	1
四室三厅单卫	1

从表 4-4 所示的结果中可以看到，“更大户型”水平显著的数目最多，这与实际情况相吻合。在实际情况中，人们总是希望自己的房屋户型面积越大越好，房间越多越好。此外，从该表中还可以发现，消费者对于实用性比较强的户型，如“两室一厅”、“三室一厅”、“三室两厅”、“四室两厅双卫”等户型的面积比较敏感，其多重比较显著的数目也比较多，所以可以认为这些户型对购房面积的影响比较大。而消费者对“四室三厅双卫”、“三室三厅”、“四室三厅单卫”等实用型稍差的户型的购买欲望不强烈，它们对购房面积的影响不大。

因此，在分析各种因素主效应的基础上，可以为开发商针对消费者购房面积问题做出以下结论：收入和户型对消费者在购房面积决策上的影响非常显著，而学历和工作单位对该决策影响不显著；其中，低收入群体和高收入群体对购房面积的要求比较敏感，实用型的户型对面积影响较大。因此，开发商在单独考虑这些因素主效应的情况下，应当优先考虑低收入和高收入人群的需求，并且为消费者有针对性地设计实用型强的户型。同时，在设计这些户型的同时，必须考虑其面积因素的影响。

那么，究竟这些影响显著的因素具体会对因变量产生什么样的影响呢？如当人们收入提高时，其购房面积会增加或降低多少平米呢？对于这种问题，可以利用方差分析模型的参数估计来解决。

**STEP 1** 在图 4-14 所示的“Linear Models”对话框中，设定好因素和因变量之后，单击“Statistics”按钮，在弹出的对话框中的“Calculate”分栏下选中“Parameter estimates”复选框，单击“OK”按钮返回“Linear Models”对话框。

**STEP 2** 此外，在“Linear Models”对话框中单击“Model”按钮，可以设置参数估计结果的形式，即可选择含有截距项和不含有截距项两种形式（设置方法同图 4-12 所示的单因素方差分析模型，本例选择含截距项）。

**STEP 3** 单击“Linear Models”对话框中的“OK”按钮，除了得到上述分析的结果之外，还会得到模型的参数估计结果，如图 4-17 所示。

Parameter	Estimate	Standard Error	t Value	Pr >  t
Intercept	111.5956848 B	12.56667011	8.88	<.0001
education 初中及以下	2.4257851 B	6.31986942	0.38	0.7013
education 大学（专、本科）	6.4418797 B	5.89792874	1.10	0.2704
education 高中（中专）	2.6115424 B	6.01678380	0.43	0.6645
education 研究生及以上	0.0000000 B	.	.	.
unit 大专院校科研单位	-4.8597270 B	4.00053515	-1.21	0.2251
unit 国营企业	-0.5103139 B	2.44372408	-0.21	0.8347
unit 其它	-2.3339739 B	3.78458994	-0.62	0.5377
unit 失业	2.1847887 B	4.45059890	0.49	0.6237
unit 私营企业	0.6542611 B	2.64989158	0.25	0.8051
unit 行政事业单位	0.0000000 B	.	.	.
income 10000~25000元	-22.3430449 B	5.91407282	-3.78	0.0002
income 10000元以下	-26.4484857 B	6.13002570	-4.31	<.0001
income 25000~50000元	-16.1755749 B	6.01777788	-2.69	0.0075
income 50000~75000元	-9.9957099 B	7.53338728	-1.33	0.1852
income 75000元以上	0.0000000 B	.	.	.
type 更大户型	-7.3115798 B	10.73224380	-0.68	0.4961
type 两室两厅	-25.7535552 B	9.48344103	-2.72	0.0069
type 两室一厅	-24.8674496 B	9.30453816	-2.67	0.0078
type 三室两厅	-22.5019034 B	9.35012245	-2.41	0.0165
type 三室三厅	-30.8800180 B	15.84689780	-1.95	0.0520
type 三室一厅	-25.3260163 B	9.32317712	-2.72	0.0069
type 四室二厅双卫	-16.5104865 B	9.87527167	-1.67	0.0952
type 四室两厅单卫	-10.2563771 B	10.40430241	-0.99	0.3248
type 四室三厅单卫	-32.3929129 B	20.96612086	-1.55	0.1230
type 四室三厅双卫	-30.2940368 B	12.27029426	-2.47	0.0139
type 一室一厅	0.0000000 B	.	.	.

图 4-17 多因素方差分析模型的参数估计结果（包含截距项）

要分析该模型参数的具体含义，必须首先找出其截距项所代表的意思。在单因素方差分析中，截距项代表的是参数估计值为“0”的水平对因变量的影响。同理，在多因素方差分析中，截距项代表的是所有参数估计值为“0”的水平对因变量的影响。在本例中，参数估计值为“0”的各因素水平有“研究生及以上”、“行政事业单位”、“75 000 元以上”和“一室一厅”，所以截距项“Intercept”表示学历为研究生及以上，在行政事业单位工作、年收入 75 000 元以上且购房意向户型为一室一厅的购房者的房屋面积需求，具体需求面积为 111.596 平米。其他水平参数估计值表示其对因变量的影响与该截距项基准对比的相对差距，该差距显著性对应的 P 值会自动计算。如要考察学历为大学、在私营企业工作、年收入为 50 000~75 000 元且购买意向户型为三室两厅的购房者的房屋面积需求为： $111.596 + 6.442 + 0.654 - 9.996 - 22.502 = 86.194$ （平米）。

上述分析过程同样可以用 SAS 编程语言的 ANOVA、GLM 等过程实现。  
使用 ANOVA 过程进行多因素方差分析不能够估计模型的具体参数，在按分组变量对因变量进行分组的各组样本量不相等的情况下不适用。

如果在各组样本量不相等的情况下利用 ANOVA 过程进行分析,系统会自动在输出结果中提示建议采用 GLM 过程进行分析。

本例采用 ANOVA 过程的具体程序如下。

```
proc anova data=Sasuser.House;      /*调用 ANOVA 过程, 并使用 Sasuser.House 数据集进行分析*/
  class education unit income type; /*指定变量“education”、“unit”、“income”、“type”为因素变量*/
  model space=education unit income type; /*指定模型, “space”作为因变量, 受所有因素变量影响*/
  means income type /lsd; /*对“income”、“type”变量进行多重比较, 关键字“lsd”表示采用 LSD 法*/
run;
```

运行程序后,可以得到上述分析过程的主要结果,同时在“Log”窗口中出现红色提示信息“WARNING: PROC ANOVA has determined that the number of observations in each cell is not equal. PROC GLM may be more appropriate.”。

该提示信息表示 ANOVA 过程已确认各组样本量不相等,建议采用 GLM 过程进行分析。本例采用 GLM 过程进行分析的具体程序如下。

```
proc glm data=Sasuser.House;      /*调用 GLM 过程, 并使用 Sasuser.House 数据集进行分析*/
/*指定变量“education”、“unit”、“income”、“type”作为因素变量*/
  class education unit income type;
/*指定分析模型, “space”作为因变量, 受所有因素变量影响, 关键字“ss3”表示对方差分析的 F 统计量做 III 型检验, 关键字“solution”表示对模型进行参数估计*/
  model space=education unit income type /ss3 solution; means income type /lsd; /*对“income”、“type”变量进行多重比较, 关键字“lsd”表示使用 LSD 法*/
run;
```

运行程序之后,可以得到与 Analyst 菜单分析过程一致的结果。

如果要精确分析影响显著的因素效应,可以在原方差分析模型中剔除影响不显著的因素。如在本例中,根据图 4-15 的结果,可知“education”、“unit”变量对因变量影响不显著,则可以考虑从模型中剔除这两个变量,只保留“income”和“type”变量进行分析,(在 SAS/Analyst 中重复图 4-14 所示的变量设定步骤,不选中这两个变量进行分析即可。)则采用 GLM 过程的程序如下。

```
proc glm data=Sasuser.House;
  class income type;
  model space=income type /ss3 solution; /*指定分析模型, “space”作为因变量, 受“income”和“type”*/因素变量影响*/
run;
```

运行程序之后,可得到图 4-18 所示分析结果。

图 4-18 所示结果的分析过程类似于剔除变量之前的分析,其分析结论从统计意义上来说更加精确。因为其分析模型已经排除了那些不显著因素的影响,只分析影响显著的因素效应,因此对于企业精确营销方案的制定有极大的帮助。

在实际生活中,人们在考虑购房面积的同时,往往会根据各种因素的不同情况进行综合考虑。如消费者在买房时不仅要结合自身收入,还要同时考虑开发商所提供住房的户型,而不是单独把其中某项因素拿出来考虑。这涉及到因素之间的交互效应,因此要用到交互效应的多因素方差分析进行具体分析。

The GLM Procedure						
Dependent Variable: space		购房面积				
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	14	30208.5928	2157.7566	6.75	<.0001	
Error	457	146052.2143	319.5891			
Corrected Total	471	176260.8070				
		R-Square	Coeff Var	Root MSE	space Mean	
		0.171386	25.02432	17.87705	71.43873	
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
income	4	12033.24256	3008.31064	9.41	<.0001	
type	10	10553.48523	1055.34852	3.30	0.0004	
Parameter	Estimate		Standard Error	t Value	Pr >  t	
Intercept	115.4981492 B		10.62821417	10.87	<.0001	
income	10000~25000元		-25.1445312 B	5.75295051	-4.37	<.0001
income	10000元以下		-29.3517671 B	5.93716592	-4.94	<.0001
income	25000~50000元		-19.3916752 B	5.81939064	-3.33	0.0009
income	50000~75000元		-13.2227134 B	7.38961605	-1.79	0.0742
income	75000元以上		0.0000000 B	.	.	.
type	更大户型		-4.5995121 B	10.50945987	-0.44	0.6618
type	两室两厅		-22.5007326 B	9.25376124	-2.43	0.0154
type	两室一厅		-23.0807700 B	9.16955590	-2.52	0.0122
type	三室两厅		-19.4781783 B	9.09330556	-2.14	0.0327
type	三室三厅		-28.6264280 B	15.52001645	-1.84	0.0658
type	三室一厅		-23.0634023 B	9.10996382	-2.53	0.0117
type	四室二厅双卫		-12.6134055 B	9.59019536	-1.32	0.1891
type	四室两厅单卫		-7.2395226 B	10.18638036	-0.71	0.4776
type	四室三厅单卫		-38.3536179 B	20.01494163	-1.92	0.0560
type	四室三厅双卫		-26.4947604 B	12.04838864	-2.20	0.0284
type	一室一厅		0.0000000 B	.	.	.

图 4-18 剔除不显著因素之后的方差分析模型

4.3.2 存在交互效应的多因素方差分析

存在交互效应的方差分析不仅只考察因素之间的交互效应，同时也考察各因素的主效应。因为在分析模型中，如果没有主效应的存在，就不会有交互效应。本小节仍以例 4-2 为例进行考察。

如果在实际分析中，不知道因素之间是否存在交互效应，可以首先考虑利用方差分析的全模型进行分析。方差分析的全模型是指在模型中考虑所有因素的主效应及所有因素之间的交互效应。在 SAS 系统中，因素之间的交互作用用 “\*” 表示，如因素 A 和因素 B 之间的交互作用在 SAS 中表示为 “A\*B”。交互作用既可以考虑两个因素之间的交互，也可以考虑多个因素之间的交互。在实际应用中，通常只考虑两个因素之间的交互，多个因素之间的交互效应的分析方法类似于两个因素的交互效应分析。

对于例 4-2，首先考虑全模型，然后根据全模型中各种效应的显著性进行因素分析。

**STEP 1** 进入 SAS/Analyst，打开 House.sas7bdat 数据集，选择菜单 “Statistics→NOVA→Linear Models”，弹出图 4-14 所示的对话框。按照图 4-14 所示的设置方法设置好因变量和因素后，单击 “Model” 按钮，弹出模型设置对话框，如图 4-19 所示。

**STEP 2** 方差分析模型默认考虑全部因素的主效应，如考虑全模型，加入因素之间的交互效应，则需单击该对话框左上角的 “Standard Models” 按钮，在其弹出菜单中选择 “Effect up to 2-way interactions”，即考虑因素二维交互效应，系统会自动在对话框中下部的 “Effects in model” 列表框中列出所有变量之间的二维交互效应，交互的因素之间用 “\*” 号连接。

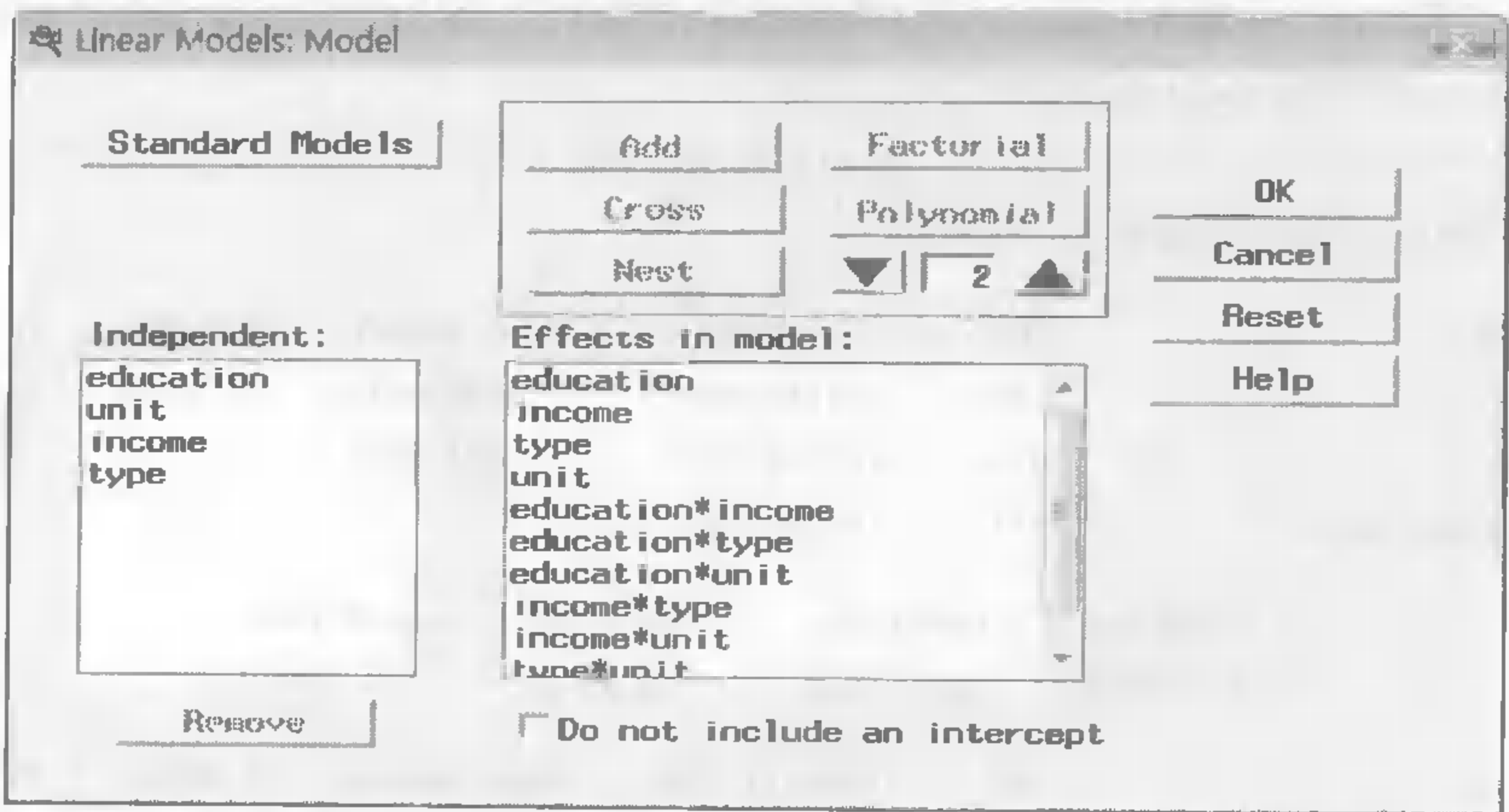


图 4-19 Linear Models 的模型设置对话框

**STEP 3)** 然后单击“OK”按钮返回“Linear Models”对话框。在该对话框中，单击“OK”按钮，得到分析结果。其中，模型各种效应对因变量影响的显著性检验结果如图 4-20 所示。

	R-Square	Coeff Var	Root MSE	space Mean		
	0.435932	23.55897	16.83023	71.43873		
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
education	3	1089.75628	363.25209	1.28	0.2802	
income	4	6550.95541	1637.73885	5.78	0.0002	
type	10	6019.59246	601.95925	2.13	0.0220	
unit	5	1793.97755	358.79551	1.27	0.2778	
education*income	7	1546.03907	220.86272	0.78	0.6047	
education*type	16	4575.76660	285.98541	1.01	0.4456	
education*unit	11	1972.68739	179.33522	0.63	0.8004	
income*type	19	17518.66472	922.03499	3.26	<.0001	
unit*income	14	6986.55816	499.03987	1.76	0.0428	
unit*type	25	6136.65637	245.46625	0.87	0.6528	

图 4-20 方差分析全模型分析结果

在该全模型中，在显著性水平  $\alpha = 0.05$  的条件下，根据各种主效应和交互效应的  $P$  值（“Pr>F”），可知收入、户型两个因素的主效应是显著的。除此之外，收入与户型的交互效应、单位与收入的交互效应均显著。可以从模型当中将其他不显著的因素剔除出去。但是剔除效应不显著的因素时，应当把主效应和交互效应结合考虑。可以考虑以下因素剔除原则：

- 剔除交互效应不显著的交互因素。
- 剔除交互效应不显著且对应主效应也不显著的因素。

模型中的显著因素往往要经过几次剔除过程才能够被筛选出来。

按照上述原则，对模型的因素进行筛选。首先观察交互效应，在给定显著性水平下，教育与收入、教育与户型、教育与单位、单位与户型 4 个交互效应均不显著，所以在模型当中应当予以剔除。收入与户型、单位与收入的交互效应均显著，构成其交互作用的收入、户型两个因素主效应也显著，所以它们的主效应和交互效应都应当被保留在模型当中。而构成单位与收入交互效应的单位因素，虽然其主效应不显著，但是其交互效应显著，因此也应当被保留在模型当中。至此，模型中影响显著的效应有收入、户型、单位主效应以及收入与户型、单位与收入的交互效应。

重新选择“Statistics→ANOVA→Linear Models”，单击“Model”按钮打开图 4-19 所示的模型设置对话框，在“Effects in model”列表框中列示的所有效应中选中应当剔除的效应，

单击左下角的“Remove”按钮即可完成剔除变量的操作。返回“Linear Models”对话框，单击“OK”按钮，得到图 4-21 所示的分析结果。

The GLM Procedure					
Dependent Variable: space		购房面积			
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	56	55176.3061	985.2912	3.38	<.0001
Error	415	121084.5009	291.7699		
Corrected Total	471	176260.8070			
	R-Square	Coeff Var	Root MSE	space Mean	
	0.313038	23.91038	17.08127	71.43873	
Source	DF	Type III SS	Mean Square	F Value	Pr > F
income	4	6622.75922	1655.68981	5.67	0.0002
type	10	8573.40431	857.34043	2.94	0.0014
unit	5	3323.52986	664.70597	2.28	0.0461
income*type	21	16860.64072	802.88765	2.75	<.0001
unit*income	16	6277.06541	392.31659	1.34	0.1663

图 4-21 剔除不显著效应的模型分析结果

然后在图 4-21 所示的结果中分析因素效应的显著性。从该结果可以看出，“unit”即单位因素的主效应虽然显著，但是其与  $\alpha(0.05)$  非常接近，为了模型的精确性，还是考虑把其剔除。经过上一轮因素剔除之后，“unit”与“income”，即单位与收入因素的交互效应不显著，这是由于在全模型中，“unit”变量本身非常不显著，而且此交互效应可能还受到其他因素的影响。因此，在方差分析模型中，把“unit”及其对应的交互效剔除。重复上述效应剔除过程，可以得到图 4-22 所示的结果。

The GLM Procedure					
Dependent Variable: space		购房面积			
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	35	47426.6743	1355.0478	4.59	<.0001
Error	436	128834.1327	295.4911		
Corrected Total	471	176260.8070			
	R-Square	Coeff Var	Root MSE	space Mean	
	0.269071	24.06238	17.18986	71.43873	
Source	DF	Type III SS	Mean Square	F Value	Pr > F
income	4	7716.07240	1929.01810	6.53	<.0001
type	10	8205.30089	820.53009	2.78	0.0025
income*type	21	17218.08151	819.90864	2.77	<.0001

图 4-22 再次剔除不显著效应的模型分析结果

经过 2 次因素效应的剔除过程之后，最终可以把影响效应显著的因素及其交互作用筛选出来。接下来就可以按照多因素方差分析的过程进行分析了。对于含有交互效应的方差分析模型，其分析过程和步骤类似于只含有主效应的方差分析过程，此处不予赘述。

4.4 协方差分析

在前面所述的方差分析过程中，可以看到各种定性变量对因变量的影响。其中，作为因

素的定性变量的水平个数是有限个，也是可以控制的，即可以控制某个因素的某个水平，观察因变量的变动状况。如在例 4-2 中，可以控制学历因素的水平，只观察硕士以上学历对购房面积的需求。但事实上，有些因素的不同水平很难人为控制，但这些因素对因变量产生了显著影响。在方差分析中，如果忽略这些因素的存在，而只分析其他可控因素对因变量的影响，往往会夸大或缩小这些因素的影响效应，使得分析结论与真实结果有偏差。

以例 4-2 为例，在研究住房使用面积需求时，仅考虑诸如收入、学历、工作单位、户型等可控因素，而不考虑消费者家庭人口数的影响(无法控制消费者家庭人口数量以进行观测)，显然是不全面的。因此，为了更加准确地研究控制变量即因素的不同水平对因变量的影响，应尽量排除其他因素对分析的影响作用。在实际分析操作中，为了达到这个目的，除了前面介绍的、根据效应显著性进行的因素筛选之外，还可以利用协方差分析进行筛选。

协方差分析是将那些难以控制的因素当作协变量，在排除协变量影响的条件下，分析可控因素对因变量的影响，从而更加准确地对可控因素进行评价。

在上述方差分析中，因素都是定性变量，而协方差分析中的协变量是定量变量，即连续数值型变量。在进行协方差分析过程中，协变量之间通常没有交互效应，且与因素变量之间也没有交互效应。

考虑协变量的方差分析模型的一般形式如下。

因变量 = 因素主效应 + 因素间的交互效应 + 协变量 + 随机误差

对于该模型，同样可以用广义线性模型的形式进行方差分析。



例 4-3

某笔记本电脑销售商为考察其不同卖场的店面一周之内的笔记本销售情况，收集了表 4-5 所示的数据（详见 Sale\_Points.sas7bdat），试分析各种因素对销售额的影响。设显著性水平  $\alpha = 0.05$ 。

表 4-5 笔记本电脑的销售状况

卖场名称 (market)	售后服务 (service)	销售额 (万元) (sales)	返点 (%) (points)
百脑汇	一年	26.00	1.80
百脑汇	一年	22.00	1.10
百脑汇	一年	21.80	0.90
百脑汇	一年	33.10	2.20
中关村 e 世界	一年	22.00	2.00
中关村 e 世界	一年	19.00	1.50
中关村 e 世界	一年	17.50	2.00
中关村 e 世界	一年	26.00	2.10
海龙	一年	23.00	1.20
海龙	一年	25.00	1.30
海龙	一年	32.00	1.90
海龙	一年	30.00	1.80
百脑汇	三年	36.00	2.00
百脑汇	三年	32.00	2.15
百脑汇	三年	28.00	1.21

续表

卖场名称 (market)	售后服务 (service)	销售额 (万元) (sales)	返点 (%) (points)
百脑汇	三年	30.00	1.91
中关村 e 世界	三年	28.00	1.50
中关村 e 世界	三年	23.00	1.20
中关村 e 世界	三年	24.50	1.60
中关村 e 世界	三年	30.00	1.80
海龙	三年	41.00	1.20
海龙	三年	46.00	1.81
海龙	三年	48.50	1.70
海龙	三年	41.30	1.30

本例要对销售额进行分析，因此“sales”变量可作为因变量，其受到两个因素的影响，即“market”和“service”变量。这两个变量是定性变量，分别有 3 个和 2 个水平，进行观测时可以选择其中任何一种情况进行考察，在方差分析模型中可以将它们设定为因素；而返点变量“points”表示对销售人员的激励，根据销售人员的实际情况给出一定比例的提成。该变量的数值因销售人员表现（如服务态度、销售业绩等）和企业规章制度而定，属于连续型的定量变量，所以在方差分析模型中可将其作为协变量考虑。因此，本例分析的目的便是要排除协变量即返点的作用之后，分析卖场因素和售后服务因素对销售量的影响，为企业店面选址和制定售后服务条款提供准确的参考意见。

**STEP 1** 进入 SAS/Analyst，打开 Sale\_Points.sas7bdat 数据集。协方差分析主要使用广义线性模型对因素效应进行分析，故选择系统菜单“Statistics→ANOVA→Linear Models”，弹出广义线性模型对话框，如图 4-23 所示。

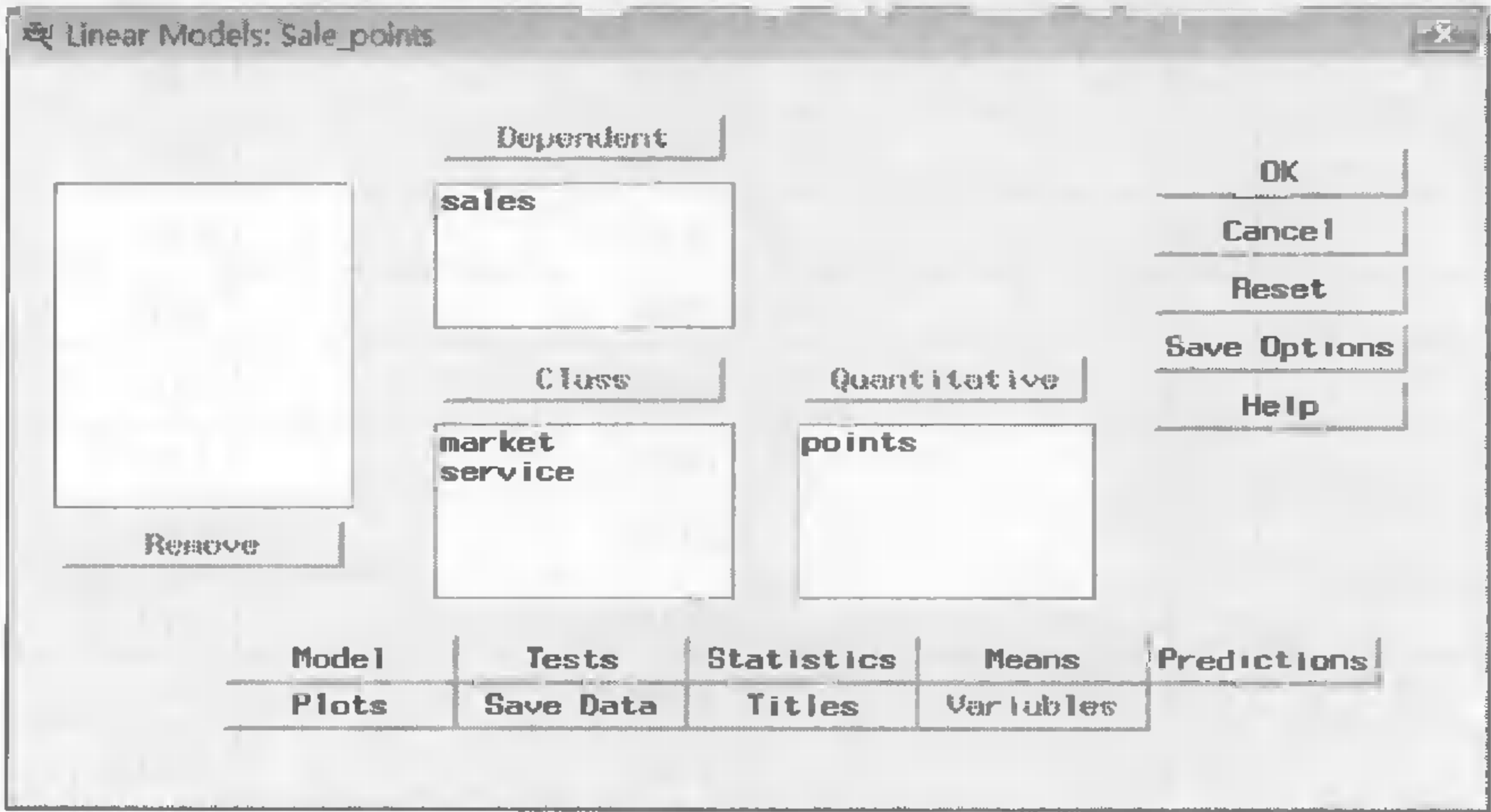


图 4-23 协方差分析的“Linear Models”对话框

**STEP 2** 因素和因变量的设定方法同多因素方差分析，本例中增加了协变量，因此在变量选择区域选中“points”变量，单击“Quantitative”按钮，将其设置为协变量。接下来的设置过程与多因素方差分析的设置过程完全一样。本例除考虑因素主效应外，还考虑了“market”和“service”的交互效应，并且对协方差分析的模型进行含有截距项的参数估计。设置好模

型和分析过程后，单击“OK”按钮，可以得到图 4-24 所示的结果。

The GLM Procedure					
Dependent Variable: sales    销售额					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	6	1469.291017	244.881836	47.05	<.0001
Error	17	88.478566	5.204622		
Corrected Total	23	1557.769583			
	R-Square	Coeff Var	Root MSE	sales Mean	
	0.943202	7.758642	2.281364	29.40417	
Source	DF	Type III SS	Mean Square	F Value	Pr > F
market	2	723.7019370	361.8509685	69.52	<.0001
points	1	196.5239340	196.5239340	37.76	<.0001
service	1	544.8213224	544.8213224	104.68	<.0001
market*service	2	195.9279051	97.9639526	18.82	<.0001
Parameter	Estimate		Standard Error	t Value	Pr >  t
Intercept	4.809189790 B		2.88984184	1.66	0.1144
market	百脑汇	8.034907413 B	1.70727194	4.71	0.0002
market	梅龙	9.380543986 B	1.68568772	5.56	<.0001
market	中关村e世界	0.000000000 B	.	.	.
points		8.587268532	1.39746808	6.14	<.0001
service	三年	8.470225699 B	1.69615439	4.99	0.0001
service	一年	0.000000000 B	.	.	.
market*service	百脑汇 三年	-5.421683458 B	2.47813541	-2.19	0.0429
market*service	百脑汇 一年	0.000000000 B	.	.	.
market*service	梅龙 三年	8.637669556 B	2.32681845	3.71	0.0017
market*service	梅龙 一年	0.000000000 B	.	.	.
market*service	中关村e世界 三年	0.000000000 B	.	.	.
market*service	中关村e世界 一年	0.000000000 B	.	.	.

图 4-24 协方差分析的结果

从图 4-24 所示的结果中可以看出，协变量对因变量的影响非常显著，其  $P$  值（“Pr>F”）小于 0.0001 故远远小于  $\alpha(0.05)$ ，因此对于分析过程而言，考虑剔除其影响的协方差分析是非常合理的。

对于考虑剔除协变量影响的效果，可以通过对比不考虑协变量的方差分析模型来得出结论，如图 4-25 所示。

	R-Square	Coeff Var	Root MSE	sales Mean		
	0.817045	13.53254	3.979130	29.40417		
Source	DF	Type III SS	Mean Square	F Value	Pr > F	
market	2	593.1608333	296.5804167	18.73	<.0001	
service	1	512.4504167	512.4504167	32.37	<.0001	
market*service	2	167.1558333	83.5779167	5.28	0.0157	

图 4-25 不考虑剔除协变量影响的方差分析结果

对图 4-24 和图 4-25 进行对比分析，通过各因素的  $P$  值可以看到，考虑剔除协变量影响的协方差分析模型的交互效应要比不考虑剔除协变量影响的方差分析模型的交互效应更显著。因此，由图 4-25 所示结果得到的分析更加准确。

观察图 4-25 中各因素主效应及其交互效应的参数估计值，可知海龙卖场对销售额影响最大，而百脑汇的销售额也比较大，这两个卖场的销售额在排除返点因素影响后，要比中关村 e 世界的销售额多出 8~9 万元，而这种差异是非常显著的。此外，笔记本电脑的售后服务时间也是值得关注的重点问题，在剔除销售人员返点因素之后，不考虑卖场因素，提供 3 年售后服务的笔记本电脑的销售额要比提供一年售后服务的笔记本电脑的销售额高出 8.47 万元。

对于上述分析过程，可以用 SAS 编程语言的 GLM 过程实现，具体程序如下。

```
proc glm data=Sasuser.Sale_points;  
  class market service; /*指定因素变量*/  
  model sales = market point service market*service / ss3 solution; /*设定协方差模型并进行参数估计*/  
run;
```

4.5 本章小结

本章主要介绍了方差分析的基本内容，简要回顾如下：方差分析通过方差比较的方式考察多个总体特征的差异是否显著；方差分析可分为一元单因素方差分析、一元多因素方差分析、多元方差分析及含有定量变量的协方差分析；方差分析可通过多重比较及方差分析模型的参数估计等方法找出影响显著的因素，并对这种影响进行定量衡量；影响因变量的各种因素之间可能存在交互效应；在 SAS 系统中可以使用 SAS/Analyst 菜单操作和 ANOVA、GLM 等过程进行编程实现方差分析。

第 5 章

非 参 数 检 验

前面所介绍的简单统计推断都是在给定或假设总体服从一定分布的前提条件下进行的。在给定分布的条件下，可以根据样本量大小、总体方差是否已知等情况，使用  $Z$  检验、 $T$  检验或  $F$  检验或  $X^2$  检验来推断总体的某些特征，或是对不同总体在某些方面的特征进行比较，这个过程叫做参数检验过程。

而在实际生活中，有许多情况是总体分布未知的，这时在给定分布条件下所进行的  $Z$  检验、 $T$  检验或  $F$  检验或  $X^2$  检验就不再适用。但是这并不意味着就此不能对总体特征进行推断了，这种在总体分布未知或与总体分布无关的情况下进行统计推断的过程称为非参数检验 (Non-Parametric Test)。

5.1 非参数检验的基本问题

在实际分析活动中，针对来自于不同总体的、不同类型的数据，可供选择的非参数检验分析方法非常多。为了避免分析上的混淆，应当首先搞清楚非参数检验的一些基本问题。

依据检验目的不同，非参数检验大体上可以分为对总体分布形式的检验（即拟合优度检验）和对总体分布位置或形状的检验（即位置检验）。在这两类检验方法中，前者检验样本所在的总体是否服从某个已知的理论分布，后者检验样本所在总体的分布位置或形状是否相同。由于总体分布未知，后者通常是对中位数进行检验。

依据所检验样本反映的总体数目不同，非参数检验又可以分为对单个样本的检验、对两个样本的检验和对多个样本的检验等。

在这些众多的检验方法中，秩 (Rank) 是非参数检验中最为常用的概念。它在非参数检验中十分重要，很多检验方法都会用到它。

秩的概念非常简单，从其英文单词的含义就可以明显看出来。所谓秩，就是将一个数列按照由小到大的顺序排列后，每个数值所获得的位置序号。例如有一个数列：

8    12   5    17   26   3    31   19   18   20


把这些数值按照从小到大的顺序进行排列，并把各个数值所处的位置次序标注于其下，可得到：

数值：	3	6	8	12	17	18	19	20	26	31
次序（秩）：	1	2	3	4	5	6	7	8	9	10

这样数值 3 处于第 1 位，其秩为 1；而 6 处于第 2 位，秩为 2；8 的秩为 3……以此类推，每个数值都可以有其对应的秩。在某些情况下，数据的秩也可以被称为等级。

非参数检验方法的适用范围比较广，无论样本所在的总体分布形式如何，对于一些非精

确测量的资料或等级资料数据均适用。



对于符合用参数检验的数据，如用非参数检验，则可能会丢失信息，导致检验效率下降。

对于不同研究目的和样本数据有多种可供选择的非参数检验方法，本节不再统一讲解非参数检验的基本原理，而是把该部分的理论内容放置在后续针对各种数据的分析过程中进行讲解。

5.2 单样本非参数检验


单样本检验可对样本数据来自于何种位置和形状的总体或是否具有随机性进行检验，主要方法有 Wilcoxon 符号秩检验、K-S 检验、游程检验等。

5.2.1 单样本均值的 Wilcoxon 符号秩检验

符号检验 (Sign Test) 是利用正、负号的数目对某种假设做出判定的一种非参数统计方法。Wilcoxon 符号秩检验也是符号检验法的一种。简单的符号检验法只是利用符号的正负来说明差异的存在，但是并没有考虑到差异的大小。而 Wilcoxon 符号秩检验对此进行了改进，在该种检验方法中，要求样本来自于连续且对称的总体。

Wilcoxon 符号秩检验的基本思想是：假设总体的中位数为  $M_0$ ，即原假设  $H_0:M=M_0$ 。从总体中得到一个样本，样本的观测值为  $x_1,x_2,\dots,x_n$ ，计算  $D_i=x_i-M_0(i=1,2,\dots,n)$ ，并按照  $|D_i|$  进行排序，每个  $|D_i|$  得到一个相应的秩。然后把  $D_i$  的符号加到相应的秩上，对带有负号的秩的绝对值求和，记为  $W^-$ ；对带正号的秩的绝对值求和，记为  $W^+$ 。如果  $M=M_0$  确实是中位数， $W^-$  和  $W^+$  应当基本上相等。而当  $W^-$  和  $W^+$  相差很大时，则可以拒绝对总体的中位数的原假设。通常取  $W=\min(W^-,W^+)$  作为 Wilcoxon 统计量进行检验。

由于 Wilcoxon 符号秩检验要求假定总体是连续且对称的，对总体中位数的检验等价于对总体均值的检验。此外，Wilcoxon 符号秩检验的基本思想还可用于两个样本的检验，以检验两样本在实验前后是否存在明显的变化（详见 5.3 节）。



**例 5-1** 某瓶装纯净水厂商生产的产品标称净含量为 600ml，现质量监督管理部门对该产品是否合格进行抽检，得到表 5-1 所示的抽检数据（详见 Water.sas7bdat）。试根据抽检结果对该产品质量进行评价。

表 5-1 瓶装纯净水的抽检结果（单位：ml）

598.78	602.14	599.15	598.74	596.74	598.19	598.47	598.86	598.05
599.98	596.60	601.07	600.20	603.46	597.87	600.60	601.65	596.44
600.48	598.45	599.66	598.72	595.37	599.13	602.84	600.15	595.94

在该数据集中，纯净水净含量的变量为 Net。如果该厂商生产的产品合格，那么经过抽检的瓶装纯净水的平均净含量应当不小于 600ml。

由于事先没有给定该品牌纯净水净含量的总体分布，因此可以考虑利用非参数的 Wilcoxon 符号秩检验。表 5-1 中的中位数  $M$  为 598.86ml，小于标称的 600ml，据此可以提出

该问题的原假设和备择假设如下。

$$H_0 : M \geq 600 \qquad H_1 : M < 600$$

对于单总体均值的 Wilcoxon 符号秩检验，SAS 系统提供了前面介绍过的 UNIVARIATE 过程，只需在 UNIVARIATE 语句的选项中指定原假设 “mu0=” 即可。本例具体程序如下。

```
proc univariate data=Sasuser.Water mu0=600;    /*调用 UNIVARIATE 过程，并用关键字 “mu0=” 指定
原假设*/
    var Net;                                     /*指定分析变量*/
run;
```

运行程序之后，除了可得到前面章节介绍过的 UNIVARIATE 过程的结果之外，还可以输出 Wilcoxon 符号秩检验的结果，如图 5-1 所示。

Tests for Location: Mu0=600				
Test	-Statistic-		-----p Value-----	
Student's t	t	-2.10758	Pr >  t	0.0449
Sign	M	-4.5	Pr >=  M	0.1221
Signed Rank	S	-83	Pr >=  S	0.0438

图 5-1 Wilcoxon 符号秩检验的结果

图 5-1 从上至下分别给出了学生  $T$  检验、符号检验和符号秩检验等三种检验方法的双侧检验结果。由于本例中的备择假设中的符号为 “<”，因此应当计算单侧检验的  $P$  值。第 3 章中介绍过的 SAS 系统单侧  $P$  值计算方法如下所示。

如果备择假设取 “<” 或 “>” 符号，则单侧  $P$  值应当按照以下原则计算。

● 如果备择假设取 “<” 符号：

当  $t \geq 0$  时，进行判定的单侧  $P$  值为  $1-(Pr>|S|)/2$ 。

当  $t < 0$  时，进行判定的单侧  $P$  值为  $(Pr>|S|)/2$ 。

● 如果备择假设取 “>” 符号：

当  $t \geq 0$  时，进行判定的单侧  $P$  值为  $(Pr>|S|)/2$ 。

当  $t < 0$  时，进行判定的单侧  $P$  值为  $1-(Pr>|S|)/2$ 。

根据最后一行的符号秩检验对应的统计量值为 -83，及本例备择假设符号为 “<”，故其单侧  $P$  值为： $0.0438/2=0.0219$ ，因此可以在给定的显著性水平  $\alpha = 0.05$  条件下拒绝原假设，认为该厂商生产的该种瓶装纯净水的产品质量不合格。

5.2.2 单样本的 Kolmogorov-Smirnov 检验

Kolmogorov-Smirnov 检验简称 K-S 检验，主要用来检验样本数据所反映的总体是否服从某种理论分布族，即用样本数据的累计分布与某个特定的理论分布相比较，若两者间的差距很小，则推断该样本来自于某特定分布族。

设总体累积分布为  $F(x)$ ，理论分布族为  $F_0(x)$ ，则 K-S 检验的问题转化为以下的原假设和备择假设。

$$H_0 : F(x) = F_0(x) \qquad H_1 : F(x) \neq F_0(x)$$

其中  $F_0(x)$  可以是需要进行检验的分布。SAS 系统中可提供 K-S 检验方法对指数分布、 $\beta$  分布、 $\gamma$  分布、正态分布、对数正态分布、Weibull 分布进行检验。



例 5-2

某调查公司在某项调查中收集到 76 个观测值的样本数据（详见 KS.sas7bdat），如表 5-2 所示。试分析该数据的总体分布是何种分布。

表 5-2 76 个观测值的样本数据

77	75	77	86	77	93	90	86	89	84	78
92	79	88	83	73	94	61	77	94	85	90
90	82	85	92	70	87	78	69	83	73	90
71	97	72	79	65	89	87	72	93	85	82
74	91	85	79	75	76	82	91	71	92	61
80	88	80	89	77	90	89	77	92	93	80
83	88	85	73	67	76	92	61	87	86	

在该数据集中，观测值样本数据的变量为 Observed。对于 K-S 检验，在 SAS 系统中仍然可以由 UNIVARIATE 过程中的 HISTOGRAM 语句来实现。

HISTOGRAM 语句可以进行 K-S 检验的分布对应的关键字如下。

- beta:  $\beta$  分布，具有  $\theta$ 、 $\sigma$  参数和  $\alpha$ 、 $\beta$  形状参数。
- exponential: 指数分布，具有  $\theta$ 、 $\sigma$  两个参数。
- gamma:  $\gamma$  分布，具有  $\theta$ 、 $\sigma$ 、 $\alpha$  参数。
- lognormal: 对数正态分布，具有  $\theta$ 、 $\sigma$ 、 $\zeta$  参数。
- normal: 正态分布，具有  $\mu$ 、 $\sigma$  参数。
- weibull: 韦伯分布，具有  $\theta$ 、 $\sigma$ 、 $c$  参数。

对于上述各分布，各参数在 SAS 系统中可用以下关键字表示。

- $\theta$ : THETA
- $\sigma$ : SIGMA
- $\alpha$ : ALPHA
- $\beta$ : BETA
- $\zeta$ : ZETA
- $\mu$ : MU
- $c$ : C

用户可以自行指定总体分布的参数值，也可以使用关键字 EST 指定为样本估计值。本例对样本数据进行常用的分布检验程序如下：

```
proc univariate data=Sasuser.KS noprint;
  var Observed;
  histogram / noplot
    normal(mu=est sigma=est)
    lognormal(zeta=est sigma=est theta=est)
    exponential(sigma=est theta=est)
    weibull(sigma=est c=est theta=est);
run;
```

/\*调用 UNIVARIATE 过程\*/

/\*指定表示分析数据的变量\*/

/\*使用 HISTOGRAM 语句，设置不显示直方图\*/

/\*指定  $H_0$  为正态分布，参数使用估计值\*/

/\*指定  $H_0$  为对数正态分布，参数使用估计值\*/

/\*指定  $H_0$  为指数分布，参数使用估计值\*/

/\*指定  $H_0$  为韦伯分布，参数使用估计值\*/

运行程序之后，可分别得到各分布的 K-S 检验结果，如图 5~2~图 5-5 所示。

The UNIVARIATE Procedure				
Fitted Distributions for Observed				
Parameters for Normal Distribution				
Parameter		Symbol	Estimate	
Mean		Mu	81.96053	
Std Dev		Sigma	8.661702	
Goodness-of-Fit Tests for Normal Distribution				
Test	---Statistic---		-----p Value-----	
Kolmogorov-Smirnov	D	0.11085609	Pr > D	0.021
Cramer-von Mises	W-Sq	0.13947527	Pr > W-Sq	0.034
Anderson-Darling	A-Sq	0.92219442	Pr > A-Sq	0.020

图 5-2 正态分布的参数估计及拟合优度检验结果

Parameters for Lognormal Distribution				
Parameter	Symbol	Estimate		
Threshold	Theta	-122.002		
Scale	Zeta	5.31703		
Shape	Sigma	0.043034		
Mean		81.96468		
Std Dev		8.7816		
Goodness-of-Fit Tests for Lognormal Distribution				
Test	---Statistic---		-----p Value-----	
Kolmogorov-Smirnov	D	0.11604429	Pr > D	0.002
Cramer-von Mises	W-Sq	0.15467999	Pr > W-Sq	0.005
Anderson-Darling	A-Sq	1.03722693	Pr > A-Sq	0.002

图 5-3 对数正态分布的参数估计及拟合优度检验结果

Parameters for Exponential Distribution				
Parameter	Symbol	Estimate		
Threshold	Theta	60.72053		
Scale	Sigma	21.24		
Mean		81.96053		
Std Dev		21.24		
Goodness-of-Fit Tests for Exponential Distribution				
Test	---Statistic---		-----p Value-----	
Kolmogorov-Smirnov	D	0.2943166	Pr > D	<0.001
Cramer-von Mises	W-Sq	2.3719994	Pr > W-Sq	<0.001
Anderson-Darling	A-Sq	11.6756170	Pr > A-Sq	<0.001

图 5-4 指数分布的参数估计及拟合优度检验结果

Parameters for Weibull Distribution				
Parameter	Symbol	Estimate		
Threshold	Theta	-61.2324		
Scale	Sigma	147.0691		
Shape	C	20.52001		
Mean		82.03051		
Std Dev		8.6609		
The UNIVARIATE Procedure				
Fitted Distributions for Observed				
Goodness-of-Fit Tests for Weibull Distribution				
Test	---Statistic---		-----p Value-----	
Kolmogorov-Smirnov	D	0.08443054	Pr > D	0.113
Cramer-von Mises	W-Sq	0.10159024	Pr > W-Sq	0.053
Anderson-Darling	A-Sq	0.59381369	Pr > A-Sq	0.061

图 5-5 Weibull 分布的参数估计及拟合优度检验结果

在图 5-2~图 5-4 中可看到, 在给定的显著性水平 $\alpha=0.05$  条件下, 正态分布、对数正态分布和指数分布的 Kolmogorov-Smirnov 的  $D$  统计量对应的  $P$  值 ( $P>D$ ) 均小于 $\alpha$ , 因此可拒绝该公司所搜集的样本数据来自于上述总体的原假设。在图 5-5 中, Weibull 分布检验的  $D$  统计量对应的  $P$  值为 0.113, 大于 $\alpha$ , 因此不能拒绝样本数据来自于 Weibull 分布的总体的假设。

SAS 系统的 UNIVARIATE 过程除了给出 K-S 检验的  $D$  统计量之外, 还给了另外两种检验分布的统计量, 即 Cramér-von Mises 的  $W^2$  统计量和 Anderson-Darling 的  $A^2$  统计量。

那么, 在何种情况下使用何种统计量进行分布的检验呢? Kolmogorov-Smirnov  $D$  统计量、Cramér-von Mises  $W^2$  统计量和 Anderson-Darling  $A^2$  统计量在给定所检验分布的所有参数已知时均可使用。对于正态分布检验而言, 当 $\theta$ 已知、 $\sigma$ 未知或 $\theta$ 未知、 $\sigma$ 已知时, 使用  $A^2$  和  $W^2$  统计量进行检验; 对于对数正态分布检验而言, 当 $\theta$ 已知、 $\zeta$ 已知、 $\sigma$ 未知或 $\theta$ 已知、 $\zeta$ 未知、 $\sigma$ 已知时, 使用  $A^2$  和  $W^2$  统计量进行检验; 对于 Weibull 分布检验而言, 当 $\theta$ 已知、 $\sigma$ 未知、 $c$ 已知或 $\theta$ 已知、 $\sigma$ 已知、 $c$ 未知或 $\theta$ 已知、 $\sigma$ 未知、 $c$ 未知时, 使用  $A^2$  和  $W^2$  统计量进行检验。

对于单样本的分布检验问题, 在 SAS 系统中, 还可以通过 SAS/Analyst 进行分析。

**STEP 1** 进入 SAS/Analyst, 打开本例所用的 KS.sas7bdat 数据集, 选择系统菜单“Statistics→Descriptive→Distributions”, 弹出分布对话框, 如图 5-6 所示。

**STEP 2** 在变量选择区域中选中变量“Observed”, 单击“Analysis”按钮把其指定为分析变量。然后单击右下角的“Fit”按钮, 弹出拟合对话框, 如图 5-7 所示。

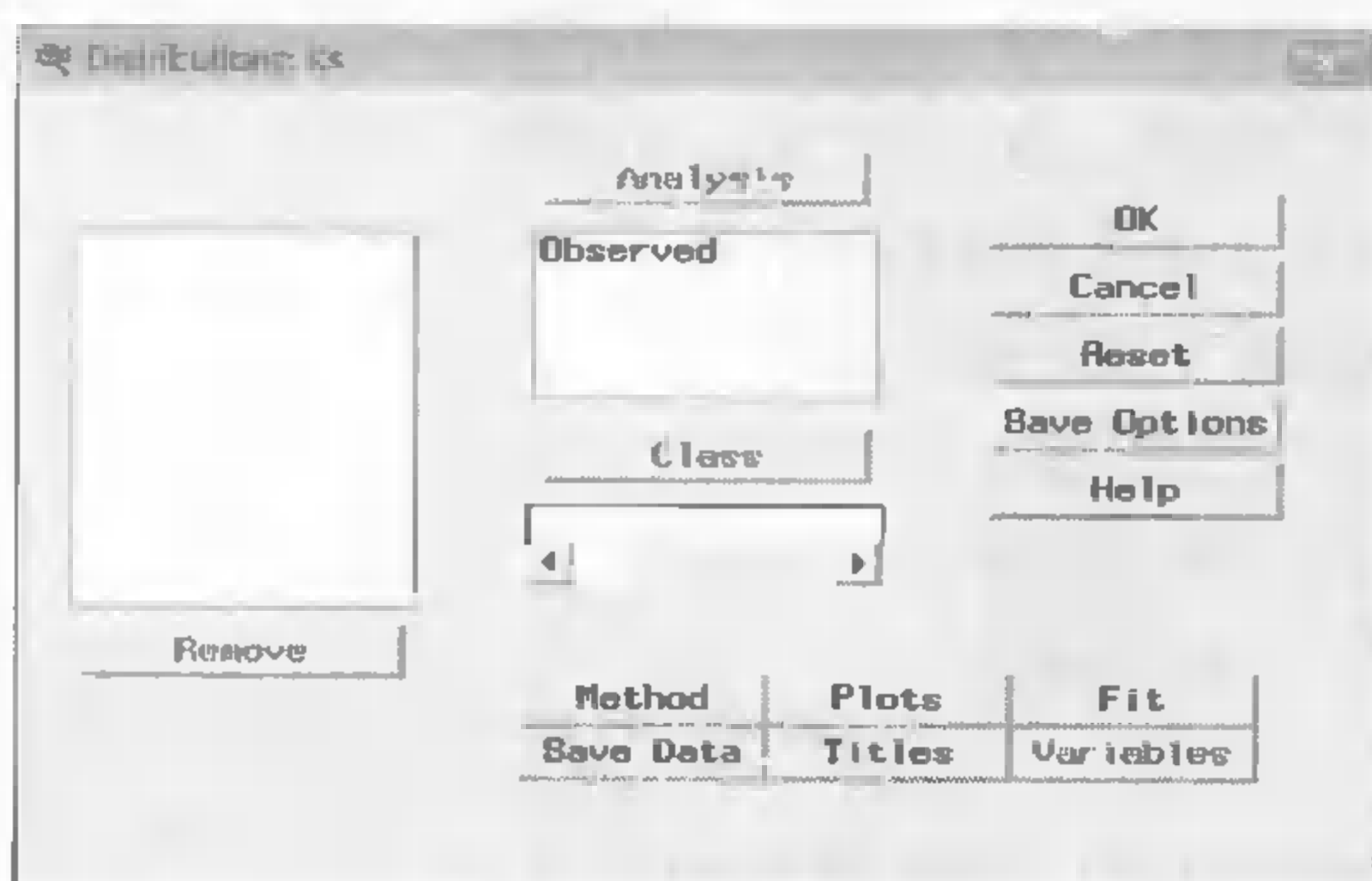


图 5-6 “Distributions”对话框

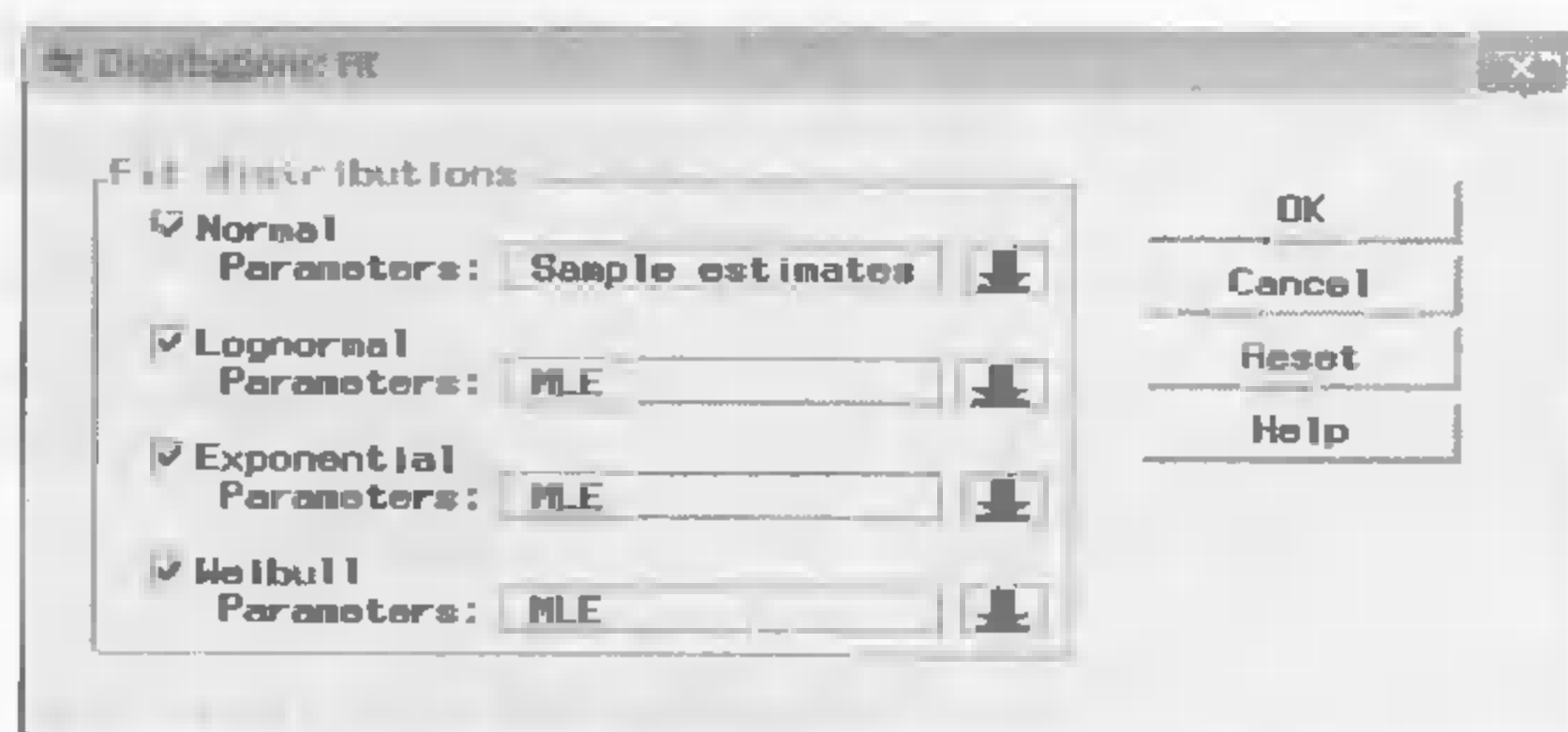



图 5-7 “Fit”对话框

**STEP 3** 在 Fit 对话框中, 系统提供了可用于检验的 4 种总体分布, 即正态分布、对数正态分布、指数分布和 Weibull 分布。对于各种分布的参数, 可以单击对应的  按钮, 选择使用样本估计值或自行指定参数的具体值。本例使用系统自行估计值进行检验。单击“OK”按钮返回图 5-6 所示的对话框。在该对话框中, 单击“OK”按钮, 便可在 SAS/Analyst 的分析结果中找到与编程过程一致的输出结果。

## 5.3 两个样本的非参数检验

对于来自于两个独立总体的两个样本数据, 同样可以利用非参数检验的方法来检验它们之间的差异。

### 5.3.1 两个独立样本中位数比较的 Wilcoxon 秩和检验

当比较两个独立样本的均值差异时, 可以使用 Wilcoxon 秩和检验。Wilcoxon 秩和检验也可被称为 Mann-Whitney-Wilcoxon 检验。

两个独立样本的 Wilcoxon 秩和检验的前提条件是两个总体的分布具有类似的形状。其基本原理与 5.2.1 小节介绍过的单样本 Wilcoxon 秩和检验类似,假定第 1 个样本的容量为  $n_1$ ,第 2 个样本的容量为  $n_2$ ,把两个样本进行合并,则合并之后的样本容量为  $n_1 + n_2$ ,把合并之后的样本数据从大到小进行排序,可得到每个观测值所对应的秩。然后分别把第 1 个样本和第 2 个样本的秩相加,得到第 1 个样本的秩和为  $W_1$ ,第 2 个样本的秩和为  $W_2$ 。如果  $W_1$  和  $W_2$  差异比较大,则可以拒绝两个独立样本中位数相等的原假设。

在 SAS 系统中,可以利用 NPAR1WAY 过程对该检验进行非参数检验。NPAR1WAY 过程的主要语法如下:

```
proc npar1way < 选项 >;
  by 变量;
  class 变量;
  exact 统计选项 </ 计算选项 >;
  freq 变量;
  output <out=数据集名>< 选项 >;
  var 变量;
run;
```

在 NPAR1WAY 过程中,by、class、freq、var 语句的功能与前面章节介绍过的功能相同。exact 语句主要用于进行精确非参数检验,本书对该部分内容不予阐述,有兴趣的读者可参考相关非参数检验的理论书籍。

NPAR1WAY 的主要选项如下。

- DATA=: 指定分析的数据集。
- ANOVA: 对原始数据进行标准的方差分析。
- D: 对于两样本数据,除了生成由经验分布函数(EDF)产生的双侧 K-S D 统计量之外,还计算单侧 K-S  $D^+$ 和  $D^-$ 统计量及其近似  $P$  值。
- MISSING: 把缺失的分类水平按非缺失值计算。
- EDF: 要求计算基于 EDF 的某些统计量,其中包括 Kolmogorov-Smirnov 和 Crammer-Vonmises 统计量。
- SAVAGE: 用 Savage 得分进行分析,它是指数分布的期望次序统计量。
- MEDIAN: 用中位数得分进行分析。中位数得分为 0 或 1,当大于中位数时为 1,反之为 0。
- MOOD: 用 Mood 得分进行分析。
- VW: 用 Van der Waerden 得分进行分析。
- WILCOXON: 用 Wilcoxon 得分进行分析。对于两样本的情况进行的是 Wilcoxon 秩和检验,对于多样本的情况进行的是 Kruskal-Wallis 检验。
- ST: 用 Siegel-Tukey 得分进行分析。
- KLOTZ: 运用 Klotz 得分进行分析。



### 例 5-3

某连锁经营公司在全国范围内有若干连锁店,为考察市场的地域差异性,现对南方市场的 32 家和北方市场的 40 家连锁店进行调查,考察其某月销售额是否有差异,数据(详见 Sales\_District.sas7bdat)如表 5-3 所示。

表 5-3 某连锁经营公司南北方市场某月的销售情况（单位：万元）

销售额 Sales	地区 District	销售额 Sales	地区 District	销售额 Sales	地区 District	销售额 Sales	地区 District	销售额 Sales	地区 District	销售额 Sales	地区 District
87.17	North	87.02	North	114.21	North	101.24	South	117.38	South	120.99	South
88.45	North	112.34	North	88.79	North	112.56	South	99.92	South	105.92	South
93.52	North	86.32	North	121.35	North	112	South	109.47	South	118.98	South
96.17	North	132.19	North	102.79	North	140.65	South	103.98	South	116.64	South
92.68	North	128.78	North	123.52	North	137.85	South	138.81	South	136.82	South
85.65	North	106.91	North	117.26	North	118.75	South	142.58	South	125.77	South
104.03	North	97.62	North	100.3	North	106.73	South	117.89	South	117.47	South
123.34	North	87.03	North	112.77	North	111.25	South	138.51	South	141.99	South
88.25	North	128.52	North	100.09	South	121.14	South	142.62	South	112.06	South
98.25	North	134.63	North	143.09	South	142.16	South	102.34	South	138.62	South
96.64	North	104.42	North	99.78	South	128.59	South	96.21	South	106.13	South
95.03	North	132.17	North	109.5	South	121.25	South	133.7	South	118	South

对于南、北地区某月的销售情况差异，可以比较其中位数是否相等。提出原假设和备择假设如下。

$H_0:M_{north}=M_{south}$        $H_1:M_{north}\neq M_{south}$

利用 NPARIWAY 过程进行分析的具体程序如下。

```
proc npariway data=Sasuser.Sales_District wilcoxon; /*调用 NPARIWAY 过程进行 Wilcoxon 检验*/
  var Sales; /*指定分析变量*/
  class District; /*指定用于区分两个样本的分类变量*/
run;
```

运行程序后，可以得到 Wilcoxon 检验的结果，如图 5-8 所示。

The NPARIWAY Procedure					
Wilcoxon Scores (Rank Sums) for Variable Sales Classified by Variable District					
District	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
North	32	847.0	1168.0	88.242091	26.468750
South	40	1781.0	1460.0	88.242091	44.525000

Wilcoxon Two-Sample Test	
Statistic	847.0000
Normal Approximation Z	-3.6321
One-Sided Pr < Z	0.0001
Two-Sided Pr >  Z	0.0003
t Approximation	
One-Sided Pr < Z	0.0003
Two-Sided Pr >  Z	0.0005

Z includes a continuity correction of 0.5.

图 5-8 两个独立样本中位数的 Wilcoxon 检验结果

图 5-8 中的第一个表格主要列示了两个样本的秩和（即 Wilcoxon 得分）统计量，其中“North”地区的统计量为 847.0，“South”地区的统计量为 1 781.0。“Expected Under H0”和“Std Dev Under H0”列分别给出了两样本的秩和在原假设成立情况下的期望值和标准差。

“Wilcoxon Two-Sample Test”表格主要列示了 Wilcoxon 检验的结果。在该结果中，SAS 系统给出了分别用正态分布近似和 *T* 分布近似的检验统计量及其对应的双侧、单侧检验 *P* 值。本例检验过程为双侧检验，因此根据正态近似的 *P* 值（“Two-Sided Pr>|Z|”）0.0003 和 *T* 分布近似的 *P* 值（“Two-Sided Pr>|Z|”）0.0005，可以拒绝原假设，即可认为该公司南北方市场的销售情况是有差异的。

本例数据也可以使用 NPARIWAY 过程中的其他方法进行非参数检验。

```
proc npar1way data=Sasuser.Sales_District
  wilcoxon      /*进行 Wilcoxon 检验*/
  median        /*进行 Median 中位数检验*/
  vw;           /*进行 Van der Waerden 检验*/
  var Sales;
  class District;
run;
```

运行程序后，可得到类似 Wilcoxon 检验的结论。

在 NPARIWAY 过程中，对于用户指定的每一种检验方法，SAS 系统均会给出单因素方差分析的检验结果。在样本之间没有差异的原假设情况下，该检验被的统计量服从自由度为样本类别数 *r*-1 的近似卡方分布。特别对于 Wilcoxon 检验，该方差分析检验被叫做 Kruskal-Wallis 检验。以本例的 Wilcoxon 检验输出结果为例，Kruskal-Wallis 检验结果如图 5-9 所示。

Kruskal-Wallis Test	
Chi-Square	13.2330
DF	1
Pr > Chi-Square	0.0003

图 5-9 Wilcoxon 检验方法的  
Kruskal-Wallis 检验结果

图 5-9 中的  $P$  值(“Pr>Chi-Square”)显示, Kruskal-Wallis 检验结果显著, 说明样本之间存在差异, 而且这种差异是显著的。

**STEP 1** 利用 SAS/Analyst 也可以进行上述的两样本非参数检验过程。进入 SAS/Analyst, 打开 Sales\_District.sas7bdat 数据集, 选择系统菜单 “Statistics → ANOVA → Nonparametric One-Way ANOVA”, 弹出非参数方差分析对话框, 如图 5-10 所示。

**STEP 2** 在图 5-10 所示对话框的变量选择区域中选中 “Sales”, 单击 “Dependent” 按钮把其指定为要进行分析的因变量; 选中 “District”, 单击 “Independent” 按钮把其指定为自变量或影响因素。单击 “Tests” 按钮, 可以弹出非参数检验方法设置对话框, 如图 5-11 所示。

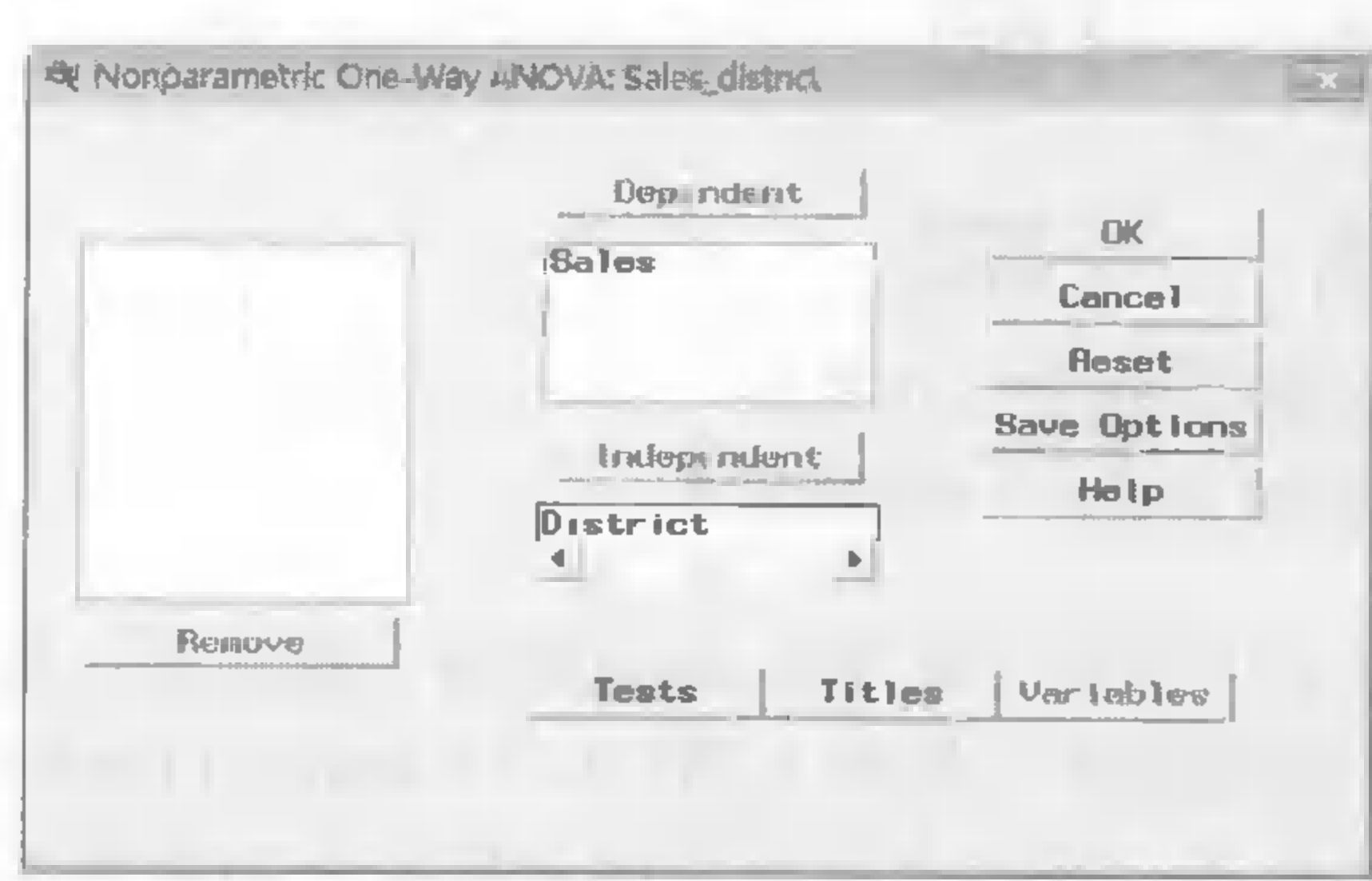


图 5-10 非参数方差分析对话框

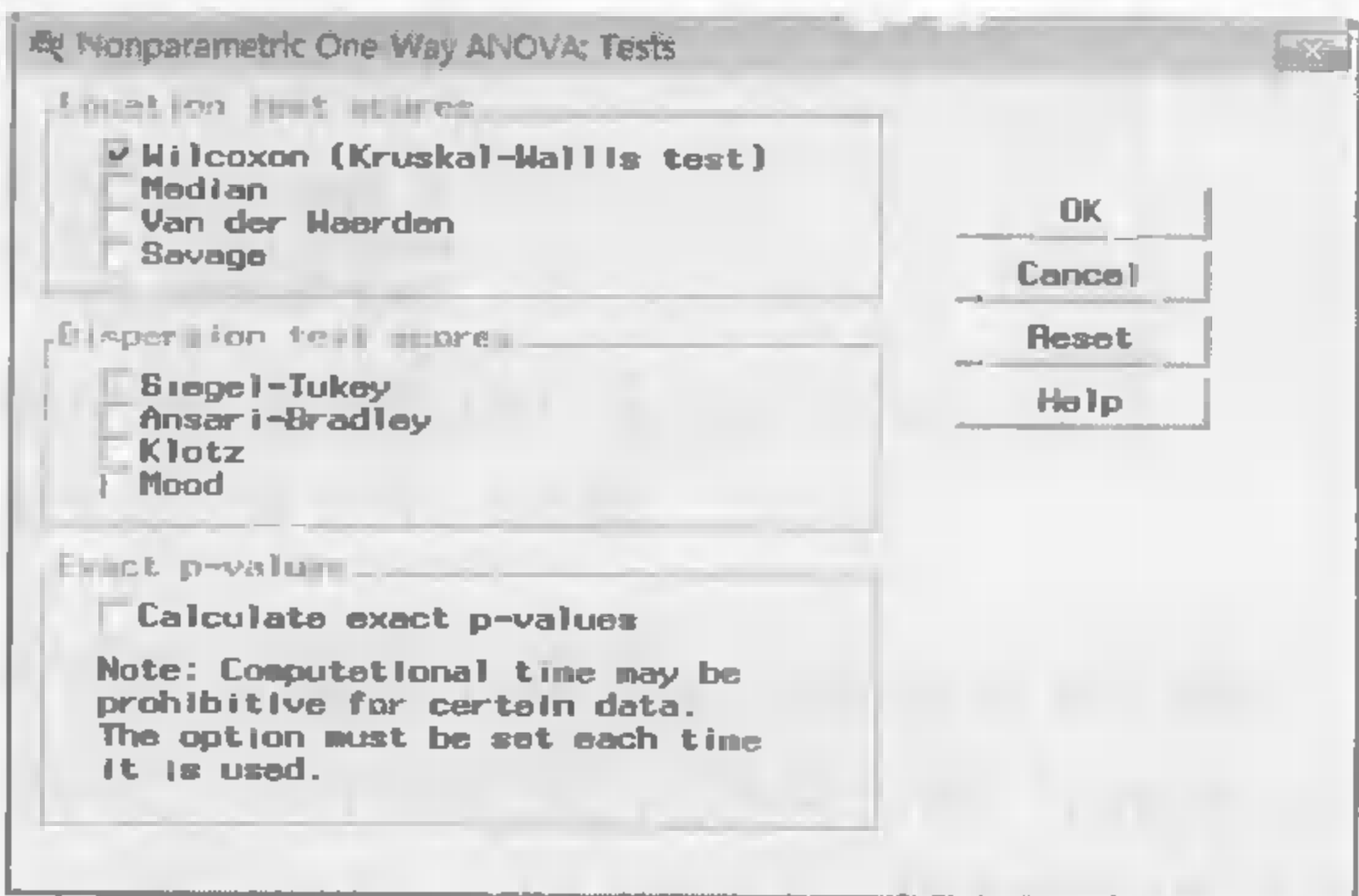


图 5-11 非参数方差分析检验对话框

**STEP 3** 在图 5-11 所示的对话框中, 系统给出了位置检验 (Location Test)、离中趋势 (Dispersion Test) 和精确检验 (Exact Test) 3 类检验方法, 每类方法下提供了一些典型的具体检验方法, 用以计算不同的统计量以进行分析。如本例需要比较两个地区销售状况的中位数, 即对位置平均数进行检验, 在 “Location test scores” 分栏下选择 “Wilcoxon” 复选框, 表示进行 Wilcoxon 检验。此外, 也可以选择如 Median、Van der Waerden 等检验方法。指定好检验方法之后, 单击 “OK” 按钮返回图 5-10 所示对话框。在该对话框中, 单击 “OK” 按钮, 便可以得到与使用 NPAR1WAY 过程编程一致的结果。

5.3.2 两个独立样本分布的 Kolmogorov-Smirnov 检验

两个独立样本分布的 K-S 检验主要检验样本所来自的总体分布是否相同。类似于单样本分布的 K-S 检验, 其原假设和备择假设如下。

$$H_0 : F_1(x) = F_2(x) \quad H_1 : F_1(x) \neq F_2(x)$$

其中  $F_1(x)$  和  $F_2(x)$  分别表示两个总体的分布。



例 5-4

为考察两个城市网吧的经营规模, 对这两个城市的网吧的电脑数量进行了调查, 数据 (详见 Cafe\_Scale.sas7bdat) 如表 5-4 所示。检验这两个城市网吧电脑规模的分布是否相同。

表 5-4 两城市各网吧电脑数量（单位：台）

城市 City	网吧编号 Café_No	电脑数量 Computers	城市 City	网吧编号 Café_No	电脑数量 Computers	城市 City	网吧编号 Café_No	电脑数量 Computers	城市 City	网吧编号 Café_No	电脑数量 Computers
城市 1	1	200	城市 1	25	80	城市 1	48	80	城市 2	71	84
城市 1	2	50	城市 1	26	140	城市 1	49	100	城市 2	72	74
城市 1	3	160	城市 1	27	50	城市 1	50	60	城市 2	73	100
城市 1	4	50	城市 1	28	200	城市 1	51	120	城市 2	74	110
城市 1	5	80	城市 1	29	50	城市 1	52	198	城市 2	75	50
城市 1	6	50	城市 1	30	110	城市 1	53	100	城市 2	76	70
城市 1	7	140	城市 1	31	60	城市 1	54	130	城市 2	77	110
城市 1	8	110	城市 1	32	100	城市 1	55	130	城市 2	78	150
城市 1	9	117	城市 1	33	80	城市 1	56	80	城市 2	79	102
城市 1	10	130	城市 1	34	68	城市 1	57	150	城市 2	80	90
城市 1	11	300	城市 1	35	100	城市 1	58	70	城市 2	81	85
城市 1	12	140	城市 1	36	38	城市 1	59	90	城市 2	82	150
城市 1	13	60	城市 1	37	178	城市 1	60	50	城市 2	83	80
城市 1	14	50	城市 1	38	180	城市 1	61	66	城市 2	84	160
城市 1	15	256	城市 1	39	124	城市 1	62	150	城市 2	85	80
城市 1	16	150	城市 1	40	90	城市 1	63	75	城市 2	86	100
城市 1	17	100	城市 1	41	130	城市 1	64	75	城市 2	87	100
城市 1	18	100	城市 1	42	52	城市 1	65	90	城市 2	88	80
城市 1	19	70	城市 1	43	145	城市 2	66	115	城市 2	89	400
城市 1	20	53	城市 1	44	150	城市 2	67	137	城市 2	90	110
城市 1	21	86	城市 1	45	60	城市 2	68	165	城市 2	91	100
城市 1	22	25	城市 1	46	80	城市 2	69	85	城市 2	92	70
城市 1	23	80	城市 1	47	79	城市 2	70	73	城市 2	93	50
城市 1	24	109									

对于两个样本分布是否一致的非参数检验问题，仍然可以利用 NPAR1WAY 过程进行编程实现，具体程序如下。

```
proc npar1way data=Sasuser.Cafe_Scale; /*调用 NPAR1WAY 过程*/
  var Computers; /*指定用于检验的变量*/
  class City; /*指定用于区分样本的分类变量*/
run;
```

运行程序后，可以得到 K-S 统计量及其对应的检验 P 值，如图 5-12 所示。

The NPAR1WAY Procedure			
Kolmogorov-Smirnov Test for Variable Computers Classified by Variable City			
City	N	EDF at Maximum	Deviation from Mean at Maximum
城市1	65	0.261538	0.461463
城市2	28	0.071429	-0.703095
Total	93	0.204301	
Maximum Deviation Occurred at Observation 34 Value of Computers at Maximum = 68.0			
Kolmogorov-Smirnov Two-Sample Test (Asymptotic)			
KS	0.087208	D	0.190110
KSa	0.841006	Pr > KSa	0.4791

图 5-12 两独立样本分布 K-S 检验结果

从图 5-12 所示的  $D$  统计量对应的  $P$  值 (“Pr>KSa” =0.4791) 可以看出，在显著性水平  $\alpha=0.05$  的条件下， $P$  值远远大于  $\alpha$ ，所以没有理由拒绝原假设，即没有充分证据表明可以否定城市 1 和城市 2 网吧电脑数量的分布相同的假设。

对于两个独立样本分布的检验，NPAR1WAY 过程除了输出 K-S 检验的基本信息之外，还可以输出诸如 Cramer-von Mises 检验等其他检验方法的结果，读者同样可以依据各种检验统计量及  $P$  值进行检验判定。

5.3.3 成对样本中位数的 Wilcoxon 符号秩检验

类似于第 3 章中的成对样本均值是否相等的参数检验，成对样本中位数的非参数检验同样也要对成对样本事先进行数据转换，即先求出构成成对样本配对的观测值的差，然后再利用单样本中位数的非参数检验方法进行检验，它们的基本原理一致。



例 5-5

在例 3-12 中，考察了 2006 年和 2007 年固定样本的北京市居民生活的幸福程度，如表 5-5 所示（数据详见 Happiness.sas7bdat）。现假定对于两个年份幸福程度的分布状况未知，用非参数检验的方法检验幸福指数是否会得到提升。

表 5-5 2006 年和 2007 年北京市居民幸福指数

06 年幸福指数	07 年幸福指数	06 年幸福指数	07 年幸福指数	06 年幸福指数	07 年幸福指数	06 年幸福指数	07 年幸福指数	06 年幸福指数	07 年幸福指数
Happy_06	Happy_07	Happy_06	Happy_07	Happy_06	Happy_07	Happy_06	Happy_07	Happy_06	Happy_07
68.89	75.53	71.37	63.00	65.05	83.81	66.62	69.49	71.53	76.32
80.83	65.55	68.15	54.67	68.77	77.64	67.03	78.52	83.57	71.68
78.00	61.80	71.07	75.77	84.50	71.94	66.87	77.44	82.97	84.44
69.27	68.24	65.11	76.25	77.34	73.86	81.44	73.10	74.85	74.07
74.23	67.02	79.03	63.21	77.95	73.10	72.73	83.72	65.34	81.28
77.41	68.53	75.89	81.83	83.42	77.64	82.67	73.64	69.04	78.24
69.26	60.00	72.71	56.89	75.23	80.92	69.85	84.43	69.55	83.83
77.86	62.13	75.74	72.64	76.25	81.33	75.96	83.50	75.92	68.66
80.66	76.74	75.98	80.42	76.54	73.65	66.48	84.43	65.08	68.66

续表

06 年幸福指数 Happy_06	07 年幸福指数 Happy_07	06 年幸福指数 Happy_06	07 年幸福指数 Happy_07	06 年幸福指数 Happy_06	07 年幸福指数 Happy_07	06 年幸福指数 Happy_06	07 年幸福指数 Happy_07	06 年幸福指数 Happy_06	07 年幸福指数 Happy_07
76.66	53.24	83.26	65.43	68.61	72.25	72.08	67.10	77.19	68.16
82.20	55.72	68.96	65.79	74.90	69.23	74.88	68.94	57.87	76.24
74.15	65.00	81.38	80.04	69.77	57.71	75.01	75.70	74.16	66.36
65.87	54.76	72.02	75.12	72.85	65.67	54.35	77.15	59.98	75.34
71.24	77.13	69.78	76.22	72.12	61.56	60.93	74.99	66.94	80.87
81.27	67.13	77.10	68.62	77.34	74.75	65.51	65.76	54.65	75.12
78.21	69.52	82.30	66.32	65.56	72.12	77.14	77.93	70.99	75.84
71.88	73.48	79.77	69.36	72.85	68.56	75.05	71.58	78.92	77.92
67.44	58.90	76.66	61.78	68.20	63.23	83.69	84.63	65.98	79.91
69.18	56.76	73.48	71.89	81.52	58.79	81.83	67.17	61.40	73.96
83.90	67.85	77.87	80.17	79.56	65.12	73.16	76.37	67.72	75.48
68.08	78.81	80.77	72.68	72.08	83.17	65.56	69.11	82.58	78.57
74.50	67.32	75.59	69.60	66.08	82.24	71.13	72.47	77.35	69.94
83.32	66.01	79.13	74.19	74.41	69.78	66.58	74.80	65.53	72.45
69.73	72.20	68.55	72.57	82.97	66.77	75.87	72.01	73.80	67.69
81.34	70.54	67.28	83.53	71.56	84.16	77.10	69.04	68.31	73.97
68.68	68.37	79.92	78.30	84.55	65.48	70.62	76.23	63.19	70.02
71.66	67.32	81.06	84.22	71.13	69.80	82.48	77.32	61.06	68.23
78.02	70.86	72.66	73.09	84.74	78.29	73.19	84.52	57.77	68.15
71.49	66.03	71.60	82.10	76.23	84.24	75.32	75.63	77.51	69.99
77.01	69.52	72.74	78.71	76.92	82.90	84.40	73.84	69.02	78.86
73.48	68.29	82.05	61.38	72.25	73.44	68.78	73.56	63.64	65.00
71.64	79.16	71.83	71.68	67.40	76.31	75.30	72.39	81.27	68.95
65.09	61.71	75.30	79.20	83.29	65.41	83.13	69.61	61.16	82.81
84.93	76.46	79.76	74.10	65.43	82.16	71.04	65.33	65.56	72.89
75.25	73.90	81.74	65.98	73.22	83.61	82.08	72.07	59.33	73.38
78.06	60.29	68.07	79.86	77.81	74.77	82.54	82.31	66.51	69.39
70.08	65.96	83.43	66.68	69.70	81.51	65.90	84.62	52.23	67.81
71.28	79.35	73.53	66.53	72.32	65.48	84.32	80.15	68.96	82.03
81.57	68.28	71.28	68.28	71.02	65.23	78.51	70.13	82.29	81.50
65.43	65.00	77.52	67.97	77.69	78.18	76.53	73.49	77.31	76.72

成对样本非参数检验要求两个样本的样本量相同。本例的原假设和备择假设如下。

$$H_0 : M_{06} - M_{07} \geq 0 \qquad H_1 : M_{06} - M_{07} < 0$$

其中  $M_{06}$ 、 $M_{07}$  分别表示 2006 年和 2007 年的幸福满意度中位数。

该问题的检验过程：先求出成对样本观测值之差，然后用 UNIVARIATE 过程对差值进行非参数检验。本例具体程序如下。

```
data null;                                /*新建一个临时数据集 null，用于存储样本观测值之差的数据*/
  set Sasuser.Happiness;
  Happy=Happy_06-Happy_07; /*求成对样本观测值的差*/
run;
proc univariate;                          /*调用 UNIVARIATE 过程*/
  var Happy;                             /*指定用于分析的变量*/
run;
```

运行程序后，除了得到常用的一些统计量之外，还可以得到检验结果，如图 5-13 所示。

Tests for Location: Mu0=0				
Test	-Statistic-		-----p Value-----	
Student's t	t	1.423869	Pr >  t	0.1561
Sign	M	12	Pr >=  M	0.1036
Signed Rank	S	1199	Pr >=  S	0.1439

图 5-13 成对样本中位数检验结果

图 5-13 给出的是双侧检验的  $P$  值，本例为单侧检验问题，即备择假设取 “<” 符号。依据图 5-13 和 SAS 系统中单侧检验  $P$  值的计算原则，可以得知本例的 Wilcoxon 符号秩检验  $P$  值为  $1-0.1439/2=0.92805$ ，对于给定显著性水平  $\alpha=0.05$ ， $P$  值远远大于  $\alpha$ ，所以没有理由拒绝原假设，即可以认为 07 年的居民幸福指数相对 06 的水平有所下降。该结论与例 3-12 所得到的参数检验结果一致。

## 5.4 多个样本的非参数检验

对于来自多个总体的样本数据，同样可利用非参数检验方法检验它们之间的差异。

### 5.4.1 多个独立样本位置的 Kruskal-Wallis 检验

该种非参数检验方法可以检验多个独立总体的位置参数（如中位数）是否一致，其原假设是假定各个总体位置参数相同，备择假设为各总体位置参数不全相同。这种非参数检验的基本原理与两个独立样本的 Wilcoxon 检验思想类似，并从样本观测值的秩和入手进行考查，即把来自于  $n$  个总体的样本放在一起进行排序，然后分别把来自于不同总体的样本观测值的秩和  $R_n$  计算出来。如果各个  $R_n$  有显著差异，则可以认为各总体位置参数不同，反之则相同。

在通常情况下，多个独立样本位置秩和检验可以使用近似服从卡方分布的 Kruskal-Wallis 统计量进行检验，并且假定各个总体具有相似形状连续分布。



例 5-6

某饮料企业为考察其生产的某种品牌产品在各个季节的销售情况是否有差异，分别对其各个营销网络销售终端的销量情况进行了调查，得到表 5-6 所示的数据（详见 Drinks.sas7bdat）。试分析季节因素是否会对该品牌的饮料销售状况产生影响。

表 5-6 饮料的季节销售数据（单位：万元）

销售终端编号 (Teminal_No)	季节 (Season)			
	春 季	夏 季	秋 季	冬 季
1	98.96	88.97	88.55	84.34
2	94.43	83.69	124.64	87.79
3	91.23	111.11	108.75	87.74
4	93.2	125.47	121.96	81.41
5	98.76	119.22	83.23	85.58
6	82.28	117.97	83.16	89.81
7	93.72	118.77	122.76	85.87
8	81.41	90.89	97.6	84.74
9	87.56	122.56	81.64	86.26
10	92.9	99.98	124.42	84.74
11	91.42	104.4	85.97	83.32
12	87.12	95.31	102.6	88.6
13	89.81	129.97	86.32	80.63
14	85.19	128.8	80.98	84.43
15	99.08	120.72	108.92	88.18
16	99.52	123.2	90.75	82.78

多个独立样本位置参数的秩和检验仍然可以利用 NPAR1WAY 过程中的 WILCOXON 方法实现，具体程序如下：

```
proc npar1way data=Sasuser.Drinks wilcoxon;
  var Sales;
  class Season;
run;
```

运行程序之后，可以得到非参数检验的结果，如图 5-14 所示。

The NPAR1WAY Procedure					
Wilcoxon Scores (Rank Sums) for Variable Sales Classified by Variable Season					
Season	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
春季	16	494.0	520.0	64.495847	30.87500
夏季	16	777.0	520.0	64.495847	48.56250
秋季	16	545.0	520.0	64.495847	34.06250
冬季	16	264.0	520.0	64.495847	16.50000
Average scores were used for ties.					
Kruskal-Wallis Test					
		Chi-Square	23.9595		
		DF	3		
		Pr > Chi-Square	<.0001		

图 5-14 多个独立样本位置参数检验结果

依据图 5-14 中的 Kruskal-Wallis 检验统计量对应的 P 值（“Pr>Chi-Square” <0.0001），在

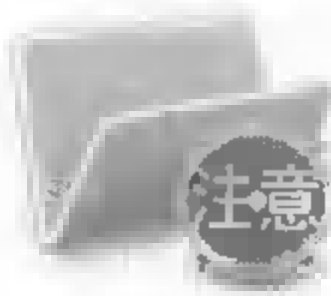
给定的显著性水平 $\alpha=0.05$ 的条件下，可知不同季节所反映的总体位置并不相同，即可以认为不同季节中该品牌饮料的销售量有显著差异。

Kruskal-Wallis 统计量检验过程也可以用 SAS/Analyst 来实现。利用 SAS/Analyst 进行多个独立样本位置检验的设置过程与两个独立总体中位数 Wilcoxon 检验相同如图 5-10 和图 5-11 所示，具体操作不再赘述。

5.4.2 多个独立样本位置的 Jonckheere-Terpstra 检验

该非参数检验方法的原假设与 5.4.1 小节的 Kruskal-Wallis 检验的原假设一样，都是假定各个总体分布的位置参数相同。但 Jonckheere-Terpstra 检验的备择假设为各个总体分布的位置参数存在顺序差异，如 $\tau_1 < \tau_2 < \dots < \tau_n$  或  $\tau_1 > \tau_2 > \dots > \tau_n$ ，其中  $\tau$  表示位置参数。

Jonckheere-Terpstra 检验简称 J-T 检验，在 SAS 系统中可利用 FREQ 过程进行编程实现。



在调用 FREQ 过程时，应当使用 table 语句指定行列变量，且列变量表示将要进行检验的变量，行变量为名义变量或顺序变量，即表示区分各个来自于不同总体样本的分类变量。

对于例 5-6 的季节销售数据，可以用 Jonckheere-Terpstra 方法考察春、夏、秋、冬 4 个季节销售额的位置参数是否有顺序排列的特征。事实上，通过观测 4 个季度的样本数据，可知冬、春、秋、夏季的中位数分别为 85.16、92.16、94.175、118.37。因此可以在 SAS 系统中先指定观测值的分类顺序，即按照冬、春、秋、夏的季度顺序，销售量按照升序进行排列，然后再进行 J-T 非参数检验。具体程序如下。

```
proc format;                                /*调用 FORMATE 过程，为分类观测值挂上顺序标签*/
    value $Season_FMT      '冬'='No1'
                           '春'='No2'
                           '秋'='No3'
                           '夏'='No4';
run;
proc freq order=formatted data=Sasuser.Drinks; /*调用 FREQ 过程，并且用“ORDER=FORMATTED”
关键字指定按照上述 FORMATE 过程指定的顺序进行分类变量排序*/
    table Season*Sales /jt;                  /*用 TABLE 语句指定分类变量和分析变量，选项中的
JT 关键字表示进行 Jonckheere-Terpstra 检验*/
    format Season $Season_FMT.;
run;
```

运行程序之后，除了可以得到 Season 变量和 Sales 变量的交叉分析表格之外，还可以得到 Jonckheere-Terpstra 检验的结果，如图 5-15 所示。

在 SAS 系统的 J-T 检验过程输出的结果中，可输出标准 J-T 统计量及其对应的单侧检验 P 值（“One-sided Pr>Z”）和双侧检验 P 值（“Two-sided Pr>|Z|”）。当 J-T 统计量小于或等于 0 时，FREQ 过程计算的检验 P 值为左单侧检验 P 值，反之计算右单侧检验 P 值。在给定的显著性水平 $\alpha$ 条件下，如左单侧 P

Statistics for Table of Season by	
Jonckheere-Terpstra Test	
Statistic	911.0000
Z	1.7148
One-sided Pr > Z	0.0432
Two-sided Pr >  Z	0.0864
Sample Size = 64	

图 5-15 Jonckheere-Terpstra 非参数检验结果

值小于 $\alpha$ ，表示支持 $\tau_1 > \tau_2 > \dots > \tau_n$ 的备择假设，即各总体的位置参数按降序排序；如右单侧 $P$ 值小于 $\alpha$ ，则表示支持 $\tau_1 < \tau_2 < \dots < \tau_n$ 的备择假设，即各总体的位置参数按升序排序。

从图 5-15 的右单侧检验  $P$  值 ( $1.7148>0$ ) 可以看出，在给定显著性水平 $\alpha = 0.05$ 的条件下， $P$  值为 0.0432，大于 $\alpha$  (0.05)，即可以认为冬、春、秋、夏 4 个季度的销售额的位置参数是按照升序进行排列的。

5.4.3 多个独立样本中位数的 Brown-Mood 检验

Brown-Mood 检验主要用于检验独立样本之间的中位数是否相等，其原假设是各个样本所反映的总体的中位数相等，备择假设为各个样本所反映的总体的中位数不全相等。

在 SAS 系统中，可以调用 NPARIWAY 过程，通过指定 MEDIAN 关键字实现多个独立样本中位数的 Brown-Mood 检验过程。该种非参数检验方法对于总体是拖尾的对称分布情况下的检验比较有效。

仍然以例 5-6 的季度销售额数据为例，进行中位数 Brown-Mood 检验的具体程序如下。

```
proc npariway data=Sasuser.Drinks median; /*调用 NPARIWAY 过程，指定使用 median 方法进行检验*/
var Sales;
class Season;
run;
```

运行程序之后，可以得到各独立样本的基本信息和检验结果，如图 5-16 所示。

The NPARIWAY Procedure					
Median Scores (Number of Points Above Median) for Variable Sales Classified by Variable Season					
Season	N	Sum of Scores	Expected Under H0	Std Dev Under H0	Mean Score
春季	16	10.0	8.0	1.745743	0.6250
夏季	16	14.0	8.0	1.745743	0.8750
秋季	16	8.0	8.0	1.745743	0.5000
冬季	16	0.0	8.0	1.745743	0.0000
Average scores were used for ties.					
Median One-Way Analysis					
Chi-Square		25.5938			
DF		3			
Pr > Chi-Square		<.0001			

图 5-16 中位数 Brown-Mood 非参数检验结果

对于该检验，SAS 系统可以计算出近似 Pearson 卡方统计量以进行检验。从图 5-16 中可以看到，如给定显著性水平 $\alpha = 0.05$ ，检验  $P$  值 (“Pr>Chi-Square”) 小于 0.0001，即  $P$  值远远小于 $\alpha$ ，故可以拒绝原假设，认为各个季节的销售额的中位数是有显著差异的。

5.5 本章小结

本章主要介绍了非参数检验及其在 SAS 系统中的实现，主要内容简要回顾如下：非参数检验是在总体分布未知或与总体分布无关的情况下进行统计推断的过程；秩 (Rank) 是非参

数检验中最为常用的概念，在非参数检验中十分重要，很多检验方法都会用到它；对于符合用参数检验的数据，如用非参数检验，则可能会丢失信息，导致检验效率下降；非参数检验可对总体位置参数（最为常见）中位数、分布情况等内容进行检验；对于单样本、两个样本及多个样本的非参数检验，在 SAS 系统中可以利用 UNIVARIATE、FREQ、NPAR1WAY 等过程实现。

（此处为模糊的正文内容，主要描述非参数检验的应用场景和SAS实现方法）



（此处为模糊的正文内容，继续讨论非参数检验的具体应用和结果解读）

（此处为模糊的正文内容，涉及SAS代码示例或结果输出分析）

## 第 6 章

# 相关与回归分析

现实世界中的任何事物之间都存在或多或少的必然联系，数据之间也不例外。在现实生活中，最常见的是数据之间的函数关系。在数据间的函数关系中，一个（些）数据发生变动，与之对应的另一个（些）数据会严格按照函数关系发生相应的变动，这种变动情况可以根据函数的具体形式进行精确度量。但实际上，数据之间的变动情况还会受到其他没有考虑到或者根本无法考虑的因素的影响，使得数据变动状况很少真正能够用函数的形式来具体描述，因而数据之间的关系往往体现为相互依存的非函数关系。而对于这种关系，可以根据数据本身的特征和自身经验进行大概的判定。

相关分析是分析两个变量或两组变量之间的相互依存关系的一种典型方法，参与相关分析的变量通常是随机变量。而回归分析不仅可以分析两个变量，还可以分析多个变量之间的相互影响。本章将对这两种常用方法进行详细讲解，并结合 SAS 系统进行分析。

### 6.1 相关分析

相关分析（Correlation Analysis）主要分析两个变量之间的相互依存关系。在学习相关分析之前，应当先区分变量或数据之间的两种主要关系。

- 函数关系：当一个或几个变量取一定的值时，另一个变量有确定值与之严格相对应，则称这种关系为函数关系。
- 相关关系：变量之间的影响不能够用具体的函数来度量，但变量之间的关系确实存在，但数量上不是严格对应的相互依存关系，则称之为相关关系。

函数关系是确定性的，往往把发生变动的变量称为“自变量”，而把受自变量变动影响而发生变动的变量称为“因变量”。如牛顿第二定律公式： $F=ma$ ， $m$  代表质量， $a$  代表加速度，当  $m$  不变时， $a$  增加一倍成为  $2a$ ，则代表力的  $F$  变量随之发生变动，也会增加一倍，即变为  $2F$ ；再如北京市出租车 15km 内的单价在 5 点到 23 点之间是 2 元/km，起步价是 3 公里 10 元，如一个人在早上 10 点钟打车走了 8 公里的路程，则可以根据函数关系精确计算出其应付的出租车价格为： $10 + (8-3) \times 2 = 20$  元。

相关关系是不确定的，主要考察变量之间的相互影响。这种影响不存在方向性，即变量 A 与变量 B 相关和变量 B 与变量 A 相关是一致的。相关关系主要体现为变量之间的相互依存关系，如身高和体重之间的关系便是相关关系的一种体现。在通常情况下，一个人的身高比较高，其体重也会相应比较重。但不能说身高增加 1cm，体重就会增加 2kg，因为还有例外，即有些身高比较高但比较瘦的人，其体重反而不如身高比较低的人的体重重。身高和体重这两个变量之间虽然不能用函数关系来描述，但是从总体上来说，这两个变量之间是存在一定的关系的。这种关系便是相互依存的关系。此外，相关分析不具有传递性，即 A 和 C 相

关，B 和 C 相关，但 A 和 B 不一定相关。

根据其分析方法和处理对象不同，相关分析可以分为简单相关分析、偏相关分析和非参数相关分析等，本节将对这些分析过程进行详细介绍。此外，根据相关关系表现形式的不同，相关分析又可以分为线性相关分析和非线性相关分析，本节主要介绍线性相关的内容和分析过程。

6.1.1 简单相关分析

简单相关分析主要分析两个变量之间相互依存的关系，可以通过主观观测和客观测度指标来衡量。

主观观测变量之间的相关关系，主要是通过两个变量之间的散点图进行分析的。而客观测度主要是通过统计分析的方法计算相关系数，利用相关系数数值的符号和大小来判定相关关系的方向和强弱。

1. 用图形描述相关关系

利用散点图可以描绘两个变量的相互影响状况。选定两个要分析的变量，把其中任意一个变量指定为二维坐标轴的横轴，把另一个变量指定为纵轴。因此，可以根据两个变量的每一对数值在二维坐标轴上描点，所有描出来的点共同形成散点图。根据散点图的不同表现情形，主要分为以下几种类型，如图 6-1 所示。

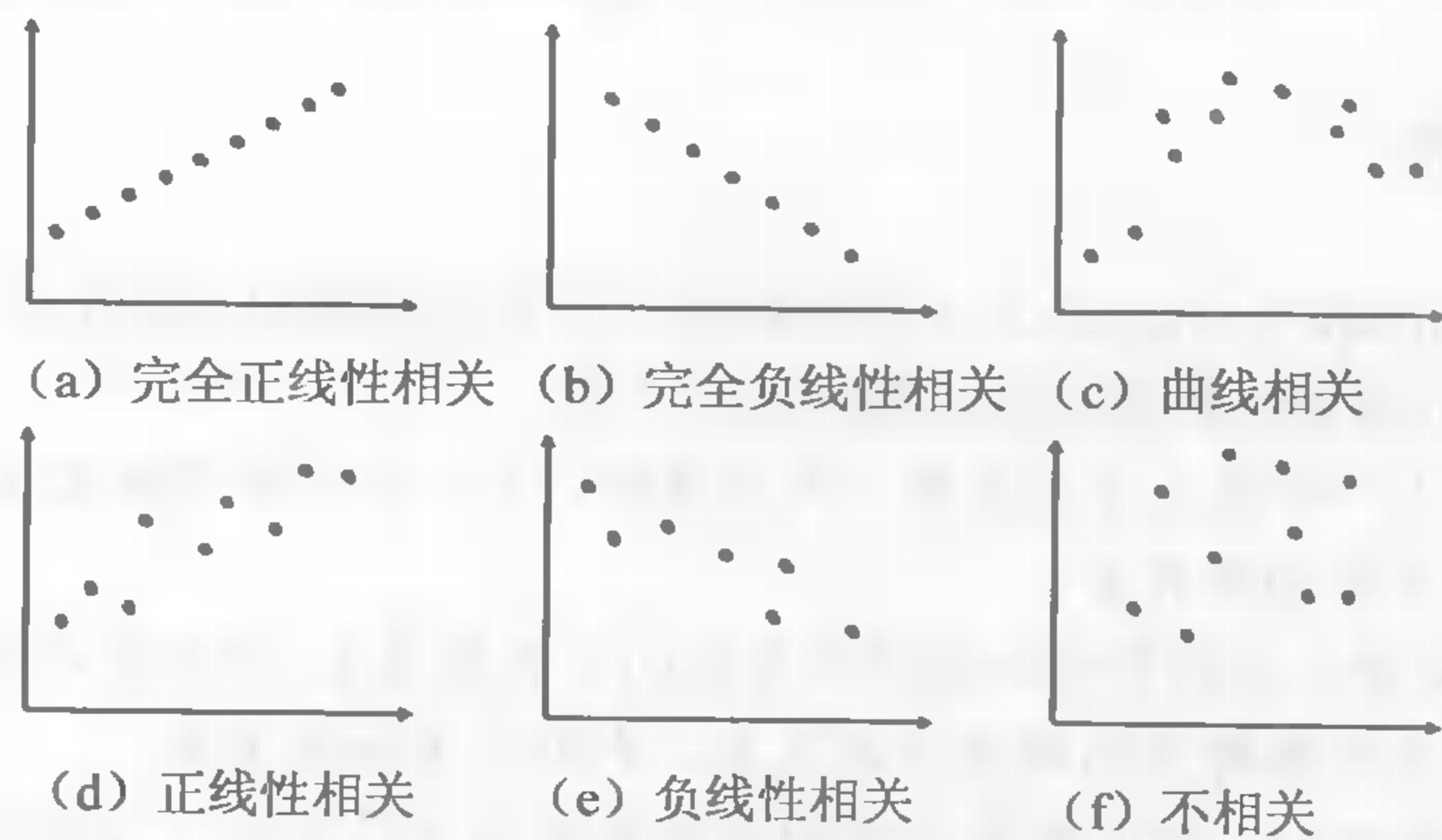


图 6-1 两个变量之间的散点图

图 6-1 中的 (a) 和 (b) 表示了两个变量之间的函数关系，而且这种关系是线性的，可以用一个直线方程来描述两个变量之间一一对应的严格关系。其中 (a) 表示随着一个变量的增大（减小），另一个变量对应地也增大（减小），这种同增同减的情况被称为“正相关”；而 (b) 所描绘的是随着一个变量的增大（减小），另一个变量减小（增大），这种反向变动的情况被称为“负相关”。

而 (c) 中描绘了变量之间的曲线相关关系，变量之间的变动关系随着曲线的形式发生变化，但是这种变动关系同样不能用严格的数学函数表示。

(d) 和 (e) 分别描述的是正线性相关和负线性相关关系。在这两个图中，只能够看到两个变量变动状况的趋势是直线的，但与 (a) 和 (b) 相比，二者之间的变动不能够用直线方程严格对应。在 (f) 中，看不出两个变量之间有相互依存的关系。

根据散点图来描述相关关系比较简单和直观，但是如果要对相关关系进行进一步分析和

下结论,则显得主观性比较强。因此,除了用图形描述相关关系之外,还可以使用相关关系的测度指标——相关系数来衡量。

## 2. 用相关系数测度相关关系

相关系数是描述线性相关程度和方向的统计量,根据样本收集的数据计算的相关系数通常用字母  $r$  表示(通常把  $r$  称为样本相关系数)。 $r$  的正负符号表示相关关系的方向,其绝对值大小表示相关关系的强弱程度。

设有两个变量分别是  $x$  和  $y$ ,根据样本数据计算相关系数的方法,主要采用 Pearson 提出的方法,即 Pearson 相关系数:

$$r = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sqrt{\sum (x - \bar{x})^2 \cdot \sum (y - \bar{y})^2}} = \frac{x \text{ 与 } y \text{ 的协方差}}{x \text{ 标准差与 } y \text{ 标准差的乘积}}$$

如两个变量之间的正向关系可用线性函数表示,则两个变量间的相关系数  $r$  是+1,表示完全正线性相关;如果两个变量之间的负向关系可以用线性函数表示,则两个变量间的相关系数  $r$  是-1,表示完全负线性相关。相关系数  $r$  的取值范围为 $[-1, +1]$ ,具体有以下几种情况。

- ①  $r = +1$ , 表示完全正线性相关。
- ②  $r = -1$ , 表示完全负线性相关。
- ③  $r < 0$ , 表示负线性相关。
- ④  $r > 0$ , 表示正线性相关。
- ⑤  $r = 0$ , 表示不存在线性关系。



当计算出来的  $r = 0$  时,只是表示线性关系不存在,但是变量之间有可能存在其他形式的相关关系(如曲线相关)。

此外,  $|r|$  的大小可以根据经验,表示不同程度的线性相关关系。

- ①  $|r| < 0.3$ , 表示低度线性相关。
- ②  $0.3 \leq |r| < 0.5$ , 表示中低度线性相关。
- ③  $0.5 \leq |r| < 0.8$ , 表示中度线性相关。
- ④  $0.8 \leq |r| < 1.0$ , 表示高度线性相关。

上述这种对相关系数的检验只是从状态上描述了变量之间的相关关系,但是相关系数  $r$  是根据样本数据计算出来的一个统计量,通过从样本数据分析出来的相关关系是否能够对总体数据下结论呢?这需要对相关系数的显著性进行检验。

## 3. 相关系数的显著性检验

对于相关系数的显著性检验问题,主要根据样本数据计算的样本相关系数  $r$  和  $t$  统计量,根据  $r$  服从自由度为  $(n-2)$  的  $t$  分布的假定,对总体相关系数(通常用  $\rho$  表示)是否等于 0 进行假设检验。如果在一定的显著性水平下,拒绝  $\rho \neq 0$  的原假设,则表示样本相关系数  $r$  是显著的。因此,该问题又可以归结为一个假设检验的问题,其原假设和备择假设是:

$$H_0: \rho = 0, H_1: \rho \neq 0$$

该假设检验问题的分析过程和得到结论的方法与第 3 章假设检验的分析过程一致。相关系数显著性的检验也可被用于本章后面讲解的其他相关分析方法。

利用 SAS 系统提供的相关分析菜单和编程语言，可以直接实现对样本相关系数  $r$  的计算和显著性检验。



例 6-1

某杂志为了评价市场上所销售汽车的最高时速与汽车自身的相应指标的影响，收集了各大厂商生产的各种系列和型号的中级汽车的最高时速、车身自重、轮胎尺寸、发动机马力等指标数据（详见 Car\_Corr.sas7bdat），如表 6-1 所示。试对这些指标进行相关分析。假定显著性水平  $\alpha = 0.05$ 。

表 6-1 各车型的指标数据

车型 (Brand_Model)	车身自重 (磅) (Weight)	轮胎尺寸 (英寸) (Circle)	最高时速 (英里) (Max_Speed)	发动机马力 (千瓦) (Horsepower)
Acura Legend V6	3 265.00	42.00	163.00	160.00
Audi 100	2 935.00	39.00	141.00	130.00
BMW 535i	3 640.00	39.00	209.00	208.00
Buick Century	2 880.00	41.00	151.00	110.00
Buick Riviera V6	3 465.00	41.00	231.00	165.00
Cadillac Eldorado V8	3 480.00	42.00	273.00	180.00
Chevrolet Lumina	3 195.00	42.00	151.00	110.00
Chrysler Imperial V6	3 570.00	43.00	202.00	150.00
Chrysler Le Baron Coupe	2 975.00	39.00	153.00	150.00
Chrysler New Yorker V6	3 450.00	42.00	202.00	147.00
Dodge Dynasty	3 080.00	42.00	153.00	100.00
Eagle Premier V6	3 145.00	39.00	180.00	150.00
Ford Taurus	3 015.00	42.00	153.00	90.00
Ford Taurus V6	3 190.00	41.00	182.00	140.00
Ford Thunderbird V6	3 610.00	38.00	232.00	140.00
Hyundai Sonata	2 885.00	41.00	143.00	110.00
Infiniti Q45 V8	4 000.00	42.00	274.00	278.00
Lexus LS 400 V8	3 930.00	40.00	242.00	250.00
Lincoln Continental V6	3 695.00	42.00	232.00	140.00
Lincoln Mark VII V8	3 780.00	43.00	302.00	225.00
Mazda 929 V6	3 480.00	39.00	180.00	158.00
Mercedes-Benz 300E	3 315.00	37.00	181.00	177.00
Nissan Maxima V6	3 200.00	42.00	180.00	160.00
Olds Cutlass Supreme V6	3 220.00	41.00	189.00	135.00
Oldsmobile Cutlass Ciera	2 765.00	42.00	151.00	110.00
Peugeot 505	3 000.00	39.00	132.00	120.00
Saab 9000S	3 065.00	40.00	121.00	130.00
Sterling 827 V6	3 295.00	42.00	163.00	160.00
Toyota Cressida	3 480.00	36.00	180.00	190.00
Volvo 740 GL	3 140.00	37.00	141.00	114.00

在本例中，采用相关分析的方法可以分析各个变量之间的相互依存关系，这里主要考察最高时速与其他因素之间的关系，也可以利用相关分析来判定哪个或哪些因素与最高时速有密切的相关关系。

**STEP 1** 进入 SAS/Analyst，打开 Car\_Corr.sas7bdat 数据集，选择系统菜单 “Statistics→Descriptive→Correlations”，弹出图 6-2 所示的相关分析对话框。

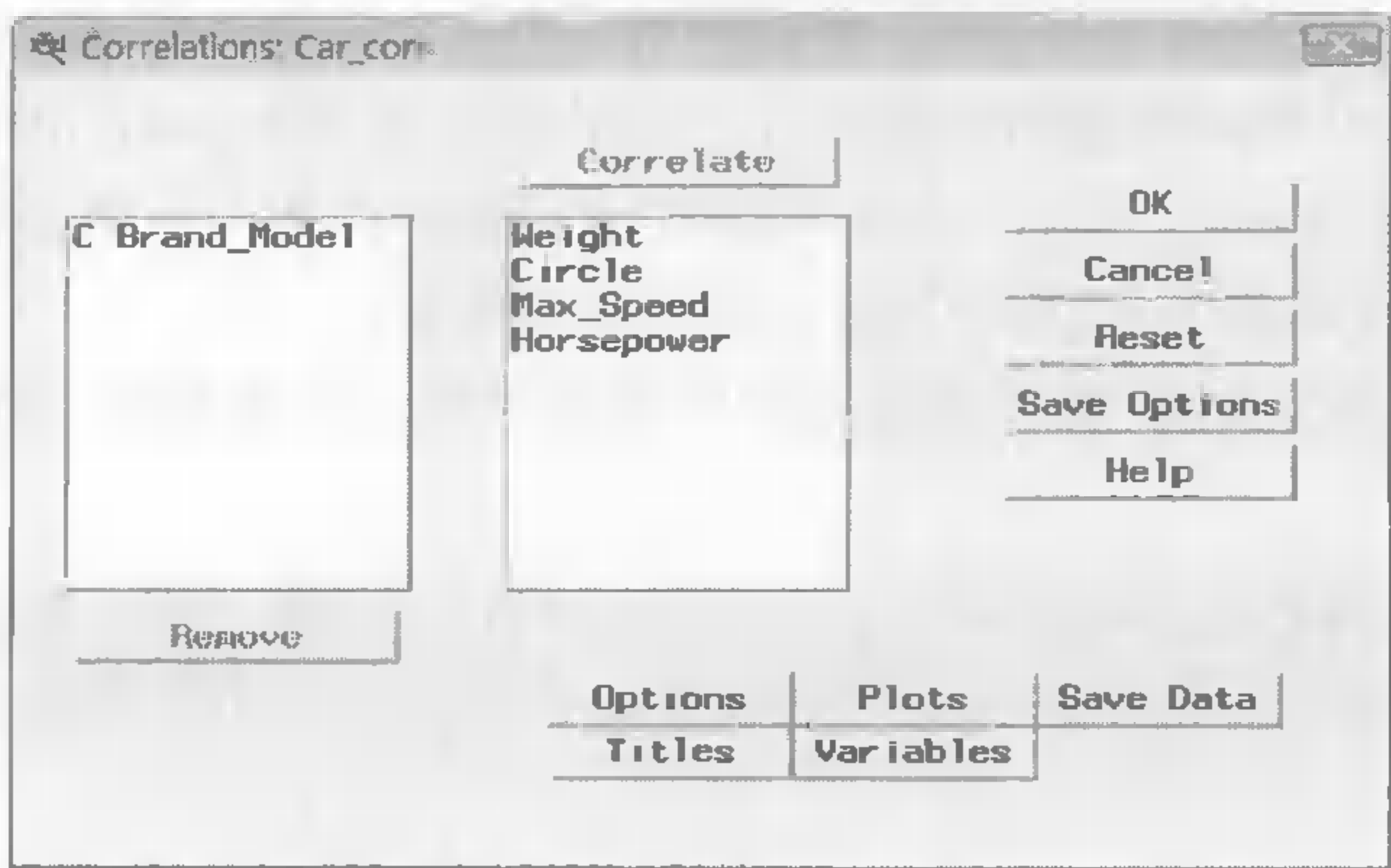


图 6-2 相关分析对话框

**STEP 2** 在相关分析对话框左边的变量选择区域中同时选中 “Weight”、“Circle”、“Max\_Speed” 和 “Horsepower” 变量，单击 “Correlate” 按钮，表示要分析 4 个变量两两之间的相关关系。



相关分析只能够分析两个变量之间的关系，当在图 6-2 所示的对话框中选中两个以上的变量进行分析时，表示要分析这些变量两两之间的相关关系。

**STEP 3** 单击 “OK” 按钮，系统在默认情况下会自动计算并输出分析变量两两之间的 Pearson 样本相关系数值及显著性检验的 P 值，如图 6-3 所示。

The CORR Procedure							
4 Variables: Weight Circle Max_Speed Horsepower							
Simple Statistics							
Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Label
Weight	30	3305	318.58113	99145	2765	4000	车身自重
Circle	30	40.50000	1.88917	1215	36.00000	43.00000	轮胎尺寸
Max_Speed	30	186.23333	45.65061	5587	121.00000	302.00000	最高时速 (英里)
Horsepower	30	152.90000	43.53702	4587	90.00000	278.00000	马力
Pearson Correlation Coefficients, N = 30 Prob >  r  under H0: Rho=0							
	Weight	Circle	Max_Speed	Horsepower			
Weight 车身自重	1.00000	0.07549 0.6918	0.85459 <.0001	0.82559 <.0001			
Circle 轮胎尺寸	0.07549 0.6918	1.00000	0.26369 0.1591	-0.02830 0.8820			
Max_Speed 最高时速 (英里)	0.85459 <.0001	0.26369 0.1591	1.00000	0.75015 <.0001			
Horsepower 马力	0.82559 <.0001	-0.02830 0.8820	0.75015 <.0001	1.00000			

图 6-3 相关分析的输出结果

相关分析的结果首先给出的是各个变量的样本量、样本均值、样本标准差、和、最小值、最大值等样本统计量，变量的标签也会对应地与变量名字一同出现。“Pearson Correlation

Coefficients”表示 Pearson 相关系数矩阵，同时在该表头下也列示了样本相关系数  $r$  显著性检验的原假设，即  $\rho = 0$ 。

在相关系数矩阵中，行列交叉对应的数值即为行变量及其对应列变量之间的 Pearson 相关系数。在每个相关系数的下面，都会给出用于检验其显著性的  $P$  值（“Prob>| $r$ ”）。

在图 6-3 中可以看到，矩阵的主对角线均为 1，表示变量自己与自己之间是函数关系。最高时速(Max\_Speed)变量与车身自重、轮胎尺寸、马力 3 个变量的相关系数分别为 0.85459、0.26369、0.75015。其中最高时速和轮胎尺寸之间的相关系数在  $\alpha = 0.05$  的条件下不显著（即  $P$  值为 0.1591，远远小于  $\alpha$ ）。因此，在不考虑其他因素的作用下，最高时速与车身自重存在高度正线性相关，与发动机马力存在中度正线性相关关系。

简单相关分析过程在 SAS 编程语言过程中非常简单，主要利用 CORR 过程来实现，具体程序如下。

```
proc corr data=Sasuser.Car_Corr pearson;           /*调用 CORR 过程，并计算 Pearson 相关系数*/
  var Weight Circle Max_Speed Horsepower; /*指定相关分析的变量，变量间用空格相隔*/
run;
```

6.1.2 偏相关分析

简单相关分析有时不能够真实反映现象之间的关系。如上述的例 6-1，发动机作为汽车的心脏，可以说对汽车的各项指标均会产生一定的影响。因此，在研究其他指标与最高时速指标之间的相关关系时，会不知不觉地在变量之间加入发动机相关指标，从而会影响对所研究的变量，而由于相关关系的不可传递性，这种影响往往会导致错误的结论。

所以，在进行相关分析时，往往要控制这种变量，剔除其对其他变量的影响之后，再研究变量之间的相关关系。这种剔除其他变量影响之后再进行相关分析的方法被称为偏相关分析（Partial Correlation Analysis）。

仍然以例 6-1 为例，在收集的 4 个指标中，依据常识，发动机马力这个变量对最高时速影响非常大。在考虑最高时速与其他变量的相关关系时，有可能包含了发动机马力因素的影响，因此考虑剔除发动机马力变量影响的偏相关分析。

**STEP 1** 进入 SAS/Analyst，打开 Car\_Corr.sas7bdat 数据集，选择系统菜单“Statistics→Descriptive→Correlations”，在图 6-2 所示的相关分析对话框中，单击“Variables”按钮，弹出偏相关分析的变量设置对话框，如图 6-4 所示。

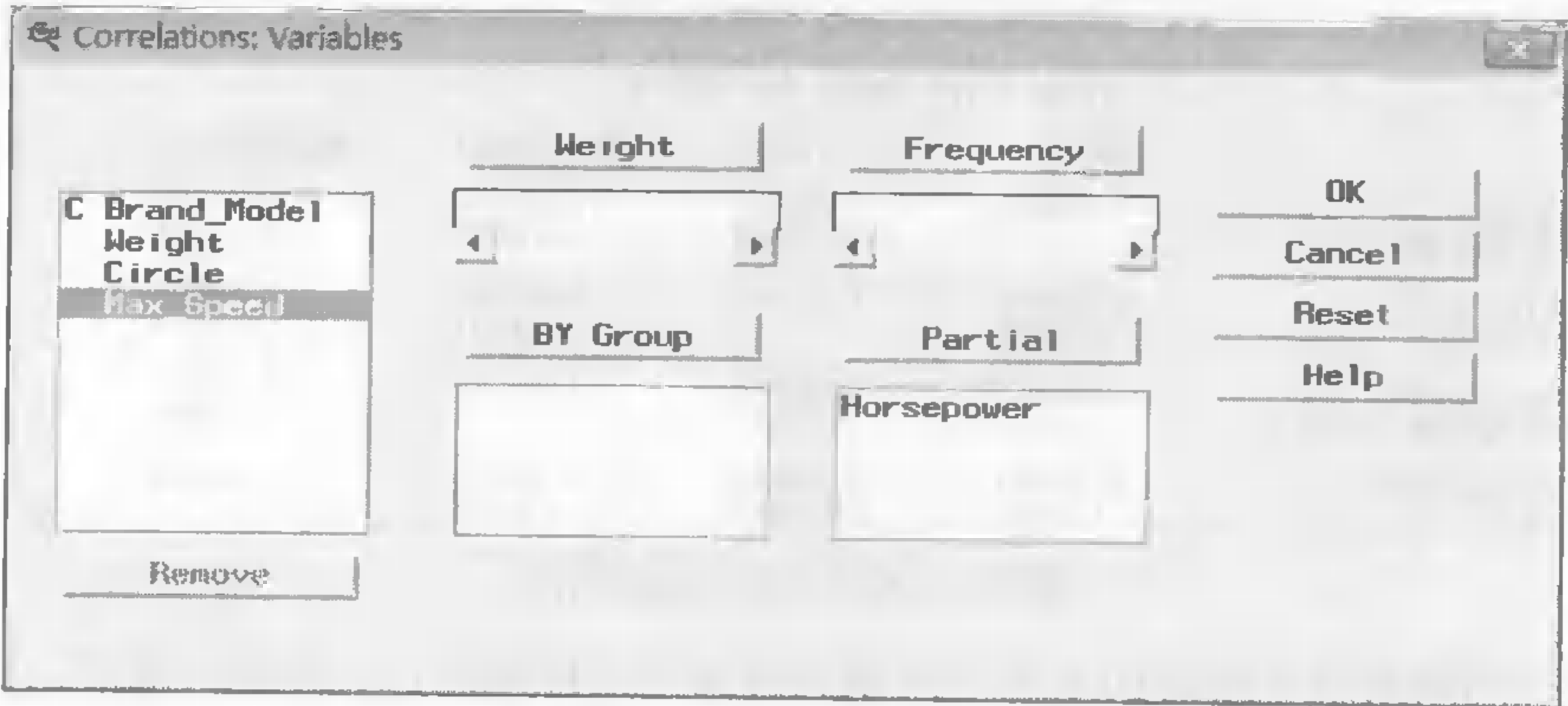


图 6-4 偏相关分析对话框

**STEP 2** 在变量选择区域中选中“Horsepower”变量，单击“Partial”按钮，将其设置为需要排除影响因素的控制变量。单击“OK”按钮返回图 6-2 所示的相关分析对话框，把剩余的“Weight”、“Circle”、“Max\_Speed”变量选中，单击“Correlate”按钮将它们指定为分析变量，再单击“OK”按钮，可得到图 6-5 所示的偏相关分析结果。

Pearson Partial Correlation Coefficients, N = 30 Prob >  r  under H0: Partial Rho=0			
	Weight	Circle	Max_Speed
Weight 车身自重	1.00000	0.17525 0.3632	0.63053 0.0002
Circle 轮胎尺寸	0.17525 0.3632	1.00000	0.43105 0.0196
Max_Speed 最高时速(英里)	0.63053 0.0002	0.43105 0.0196	1.00000

图 6-5 偏相关分析的输出结果

从偏相关分析结果中可以看到，控制住发动机动力变量的影响之后，最高时速与车身自重的相关系数有所降低，具体数值为 0.630 53，处于中度线性相关的范围。而轮胎尺寸与最高时速的相关系数大幅提升，且该相关系数在显著性水平 $\alpha=0.05$  条件下显著。这个分析结果，尤其是轮胎尺寸与最高时速的分析结果，比简单相关分析的结果与实际状况更加接近。汽车轮胎的尺寸增加，在一定程度上可以增加车身的抓地性能，从而提升速度；反过来，汽车速度不断增加，在其他条件不变的情况下，没有一定尺寸的轮胎，其最高速度也很难提升。

用 SAS 编程语言实现偏相关分析时，只要在调用 CORR 过程中加入 partial 语句，并指定控制变量即可，本例具体程序如下：

```
proc corr data=Sasuser.Car_Corr pearson;          /*调用 CORR 过程，并计算 Pearson 相关系数*/
  var Weight Circle Max_Speed;                    /*指定进行相关分析的变量*/
  partial Horsepower;                             /*partial 语句指定控制变量进行偏相关分析*/
run;
```

### 6.1.3 等级相关分析

简单相关分析和偏相关分析通常被广泛应用于定量数据或连续型数据的研究中。对于某些数据，尤其是定性数据的相关分析而言，如果用 Pearson 法计算相关系数，很难得到定性数据的协方差和标准差。因此，可以考虑采用其他方法对定性数据尤其是顺序数据进行相关分析。

对于上述情况的相关分析往往是从数据值的次序入手的，借助了非参数统计分析的思想。次序在数列中代表了某个具体变量值的位置、等级或秩，因此，这类相关分析通常被称为非参数相关分析、等级相关分析或秩相关分析，其计算的相关系数被称为非参数相关系数、等级相关系数或秩相关系数。

在 SAS 系统中，根据计算方法不同，非相关系数主要有 Spearman、Kendall tau-b 和 Hoeffding's D 等级相关系数。

#### 1. Spearman 相关系数

该相关系数主要测度顺序变量间的线性相关关系，在计算过程中只考虑变量值的顺序，而不考虑变量值的大小。

其计算过程为：首先把变量值转换成在样本所有变量值中的排列次序或名词，再利用

Pearson 方法求解转换后的两个变量对应的排列次序（即“秩”或等级）的相关系数。其具体计算公式为：

$$r = \frac{\sum (R_{x_i} - \bar{R}_x)(R_{y_i} - \bar{R}_y)}{\sqrt{\sum (R_{x_i} - \bar{R}_x)^2 \cdot \sum (R_{y_i} - \bar{R}_y)^2}}$$

其中， $R_{x_i}$  和  $R_{y_i}$  分别表示第  $i$  个  $x$  变量和  $y$  变量经过排序后的次序， $\bar{R}_x$  和  $\bar{R}_y$  分别表示  $R_{x_i}$  和  $R_{y_i}$  的均值。

## 2. Kendall tau-b 相关系数

该相关系数与 Spearman 相关系数的作用类似，主要测度顺序变量间的线性相关关系，其计算过程也只考虑变量值的顺序，而不考虑变量值的大小。

在 Kendall 相关系数计算过程中，除对数据进行排序之外，还应当综合考虑该排序与变量值的具体情况。

● 同序对：在两个变量上排列顺序相同的对变量。

● 异序对：在两个变量上排列顺序相反的对变量。

上述对子的数目简称为对子数，设  $P$  为同序对子数， $Q$  为异序对子数， $T_x$  为在  $x$  变量上是同序但在  $y$  变量上不是同序的对子数， $T_y$  为在  $y$  变量上是同序但在  $x$  变量上不是同序的对子数，则 Kendall tau-b 等级相关系数为：

$$\tau_b = \frac{P - Q}{\sqrt{(P + Q + T_x)(P + Q + T_y)}}$$

$\tau_b$  的取值范围与简单相关系数相同，即  $\tau_b \in [-1, +1]$ 。

## 3. Hoeffding's D 相关系数

该相关系数主要用于测度顺序变量或具有等级水平变量间的线性相关关系，其具体计算公式如下。

$$D = 30 \times \frac{(n-2)(n-3)D_1 + D_2 - 2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)}$$

其中：

$$D_1 = \sum (Q_i - 1)(Q_i - 1)。$$

$$D_2 = \sum (R_i - 1)(R_i - 2)(S_i - 1)(S_i - 2)。$$

$$D_3 = \sum (R_i - 2)(S_i - 2)(Q_i - 1)。$$

$R_i$ 、 $S_i$  分别表示变量  $x$ 、 $y$  的排列顺序； $Q_i$  表示 1 加上变量  $x$  和  $y$  的值均小于这两个变量中的第  $i$  个值时的个数，也被称为双变量等级。

上述等级相关系数的计算方法也可被应用于定量数据中，在相关分析中只要除去定量数据的数值意义即可。如例 6-1 所示的数据，同样可用于非参数相关系数的计算，只是计算结果所代表的相关含义会发生相应的变化。

在 SAS 系统中，可以利用菜单和编程的方式实现对上述各种线性相关系数的计算。



例 6-2

为了评价目前我国高等院校研究生的教学和培养效果，首都经贸大学统计学院和研究生部联合对全国各省市 40 多个高等院校的研究生导师及研究生本人进行了研究生培养状况调查。本书从调查原始数据中节选出 1 012 个样本（数据详见 Graduate.sas7bdat 数据集），考察研究生对自身所选专业的兴趣与其他因素之间的相关关系。具体变量情况如表 6-2 所示。

表 6-2 研究生对所选专业的兴趣及其影响因素

变 量	问 题	选 项
Interest	目前你对所学专业方向的兴趣与入学时相比的情况	1—更高；2—没变化；3—下降；4—失去兴趣；5—根本不感兴趣
Major	你认为你所学知识与你专业的方向的相符性	1—完全一致；2—基本一致；3—有点联系；4—不一致；5—完全不一致
Teaching	总体上看，你对所在学校老师教学水平的评价	1—非常满意；2—比较满意；3—一般满意；4—不太满意；5—非常不满意
Tutor	你觉得导师对你学业上的帮助程度	1—非常大；2—比较大；3—一般；4—不大；5—没作用

表 6-2 中的 4 个变量都是顺序变量，即可认为各个变量都可以分为 5 个等级，所代表的顺序值（等级）越大，表示对该问题的否定程度也越大。在这种情形下，可以使用非参数相关系数考察各个变量之间的相关关系。

对于本例中变量的相关分析，SAS 系统提供了上述几种相关系数以进行分析。

**STEP 1** 进入 SAS/Analyst，打开 Graduate.sas7bdat 数据集，选择菜单“Statistics → Descriptive → Correlations”，弹出图 6-2 所示的相关分析对话框，把上述 4 个变量同时选中，单击“Correlate”按钮，把它们设定为分析变量。然后单击“Options”按钮，弹出相关分析选项对话框，如图 6-6 所示。

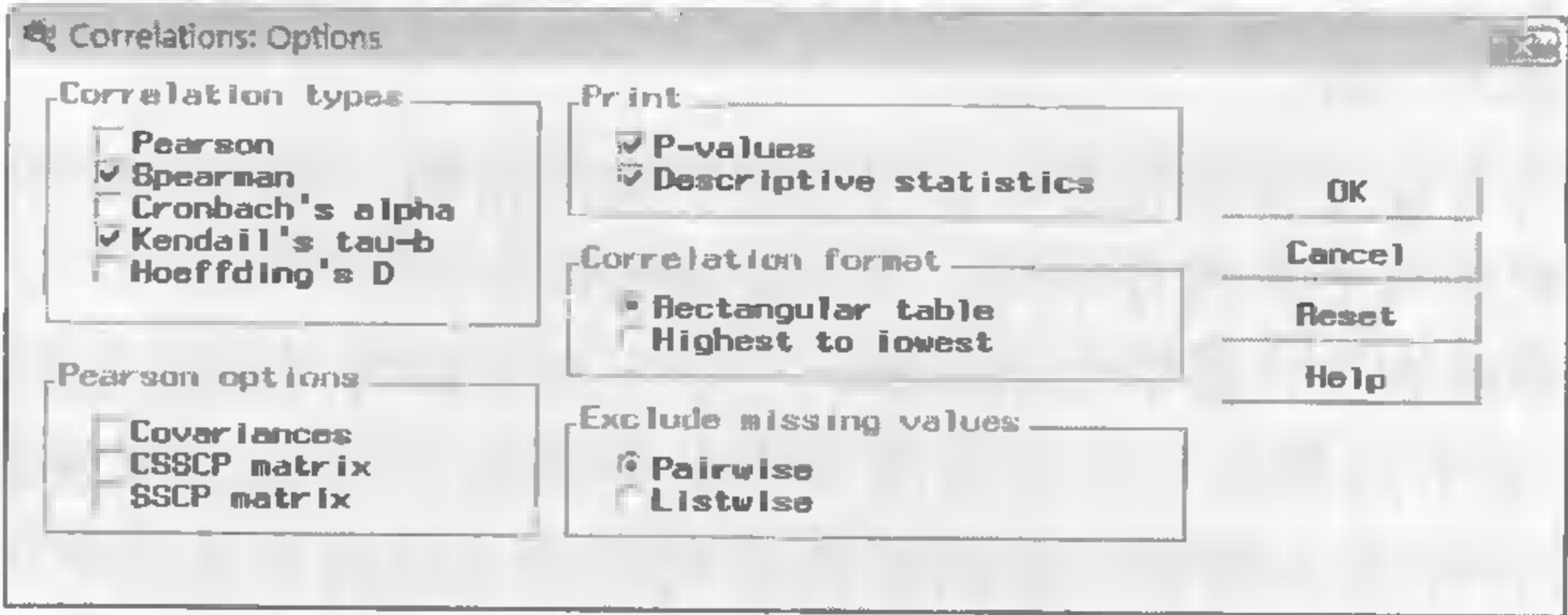


图 6-6 相关分析的选项对话框

在“Correlation types”分栏下，提供了 SAS 系统能够进行计算的相关系数类别，它们的详细功能如下。

- Pearson: 计算 Pearson（皮尔逊）简单相关系数。
- Spearman: 计算 Spearman（斯皮尔曼）相关系数。
- Cronbach’s alpha: 计算 Cronbach（克伦巴赫）一致性系数，常用于问卷信度分析。
- Kendall’s tau-b: 计算 Kendall’s tau-b（肯德尔）相关系数。
- Hoeffding’s D: 计算 Hoeffding’s D（霍夫丁）相关系数。

**STEP 2** 选中对应的复选框，即表示使用该方法计算相关系数。为方便对比，本例选择 Spearman、Kendall's tau-b 方法。在“Print”分栏下，选择对应复选框可以输出用于相关系数显著性检验的 *P* 值（*P*-values）和常用统计量（Descriptive statistics）。“Correlation format”用于指定输出相关系数结果的格式，“Rectangular table”表示用默认的方法输出相关系数表格，而“Highest to lowest”表示输出按照相关系数从大到小排列的表格。单击“OK”按钮返回相关分析对话框。在该对话框中，单击“OK”，可以得到图 6-7 所示的输出结果。

Spearman Correlation Coefficients, N = 1012				
Prob >  r  under H0: Rho=0				
	Interest	Major	Teaching	Tutor
Interest 专业兴趣	1.00000	0.72135 <.0001	0.27430 <.0001	0.80630 <.0001
Major 专业相符性	0.72135 <.0001	1.00000	0.28790 <.0001	0.73184 <.0001
Teaching 教师水平	0.27430 <.0001	0.28790 <.0001	1.00000	0.28440 <.0001
Tutor 导师影响	0.80630 <.0001	0.73184 <.0001	0.28440 <.0001	1.00000
Kendall Tau b Correlation Coefficients, N = 1012				
Prob >  r  under H0: Rho=0				
	Interest	Major	Teaching	Tutor
Interest 专业兴趣	1.00000	0.68259 <.0001	0.24518 <.0001	0.77644 <.0001
Major 专业相符性	0.68259 <.0001	1.00000	0.25715 <.0001	0.67980 <.0001
Teaching 教师水平	0.24518 <.0001	0.25715 <.0001	1.00000	0.25056 <.0001
Tutor 导师影响	0.77644 <.0001	0.67980 <.0001	0.25056 <.0001	1.00000

图 6-7 非参数相关系数输出结果

从斯皮尔曼和肯德尔两个相关系数的大小和方向来看，本例所分析的专业兴趣与其他变量的相关关系状况基本一致。

学生所学知识与专业方向的相符性和导师对学业的帮助，对入学后研究生同学的专业兴趣影响比较大。研究生与本科教育不同，其在校期间的学习和科研工作与导师的指导密不可分，导师对学生指导越到位，对学生帮助越大，学生对专业的兴趣就会越浓厚。

因此，导师对学生的帮助与学生对所学专业的兴趣相关关系程度最强，属于高度相关的层次，而且这种相关关系对于全国所有研究生的总体而言是显著的；而所学知识专业方向的相符性与专业兴趣的相关关系也比较强且是显著的，这是因为研究生教育强调的是研究性的学习和进行创造性的工作，其研究方向往往代表了对应专业领域内的前沿水平。如果专业方向没有把握好或者没有选择好，则会对研究生的科研工作造成较大的压力。所以，这两个变量之间的相关关系也比较强。而教师水平与学生的专业兴趣相关程度不高，这是因为研究生强调科研能力，往往也强调培养其独立科研的能力，通过导师在思想上的适当指导，独立进行论文写作和科研开发，因此教师授课的水平高低基本上与专业兴趣无关。

非参数相关分析类似简单相关分析，也可指定控制变量进行偏相关分析，排除指定变量对相关分析变量的影响。其分析过程、变量指定过程与简单相关分析一致，参见 4.1.2 小节。

非参数相关系数同样可以用 SAS 程序的 CORR 过程来进行计算和分析，具体程序如下。

```
proc corr data=Sasuser.Graduate Spearman Kendall Hoeffding; /*调用相关分析过程，指定分析方法，其中关键字 Spearman、Kendall、Hoeffding 分别表示计算斯皮尔曼、肯德尔、霍夫丁相关系数*/
var Interest Major Teaching Tutor;
run;
```

6.2 典型相关分析

前面讨论了两个变量之间的相关分析，也讨论了为控制第 3 个变量的影响而进行的偏相关分析，均可用简单的计算公式实现。现实生活总是很复杂的，在考虑相关关系时，往往不会仅仅考虑单独两个变量之间的相互依存关系，更多的可能是考虑两组变量之间的关系。而这两组变量中各自又可以由若干变量组成，这便衍生出对多个变量所形成的两个分组之间的关系进行研究，即两组变量之间的典型相关分析（Canonical Correlation Analysis）。

6.2.1 典型相关分析基本原理

典型相关分析是针对两组变量之间的相关关系进行的，这要求在进行分析之前，应当先按照一定的标准对参与分析的变量进行分组。对变量进行分组的一般原则是，处于同一组之内的变量能够在一定程度上共同反映该组所反映的内容。



例 6-3

某省电信公司为了研究该公司管辖范围内的通信技术质量与用户满意度之间的关系状况，委托咨询公司进行相关情况调研。根据原信息产业部的电信行业满意度模型（TCSI），该咨询公司随机收集 500 个用户的满意度各项评价得分及衡量技术质量的典型指标得分以进行分析（详见 Tech\_CSI.sas7bdat），具体变量情况如表 6-3 所示。试分析该公司通信技术质量与用户满意度之间的相关关系。

表 6-3 通信技术质量与满意度评价变量表

考核内容	具体方面	SAS 数据集中的变量
通信技术质量	接通率	Connection
	网络速度	Speed
	数据传输速率	Transmission
	掉线率	Offline
	总体接入质量	Switch
用户满意度	总体满意情况	Overall
	与理想相比的满意情况	Ideal
	与期望相比的满意情况	Expect

在表 6-3 所示的具体变量中，接通率、网络速度等 6 个变量共同反映了该公司的技术水平，构成了一组变量，把这组变量所反映的内容称为“通信技术质量”；同理，总体满意情况变量与其他两个方面的满意情况变量共同反映了用户对该公司的满意度评价，构成了另一组变量，把该组变量反映的内容称为“用户满意度”。



注意

根据各个具体变量定义的“通信技术质量”和“用户满意度”只是代表了两组变量所反映的内容，只是这两组变量的称谓，而不是体现在数据集当中的具体变量。本例只是对“通信技术质量”和“用户满意度”两组变量之间的关系感兴趣。

如果直接对参加分组的这 8 个变量进行两两之间的相关分析，则这两组变量之间的关系就很难得到。

为了达到此分析目的，可以把多个变量与多个变量之间（即两组变量之间）的相关转为两个变量之间的相关。这便要求为每一组变量选取一个综合变量作为该组变量的代表，而这个“综合变量”（也被称为“典型变量”）也能够一定程度上代表该组的大部分信息，然后再对这两个代表之间的关系进行分析。

为各组变量选取代表性的综合变量的最简单形式是，取构成该组的所有变量的线性组合。因为变量之间的线性组合可以有无数种形式，所以选出的这个线性组合必须能够代表各组的含义，同时这个线性组合所得到的结果与另一个线性组合所得到的结果之间的相关系数最大。这个分析过程便是典型相关分析的基本过程。

典型相关分析应当具备以下两个前提条件。

■ 各组变量之间的关系为线性相关关系。

● 各组变量所组成的线性组合与构成该组合的具体变量之间的关系为线性相关。

设第 1 组变量可以由具体的变量  $x_1, x_2, \dots, x_n$  构成；而第 2 组变量可以由具体的变量  $y_1, y_2, \dots, y_m$  构成。则在 SAS 系统的典型相关分析过程中，可以采用以下方法标记线性组合。

●  $V_i = a_0 + a_1x_1 + a_2x_2 + \dots + a_nx_n$ ，表示第 1 组变量的第  $i$  种线性组合。

●  $W_i = b_0 + b_1y_1 + b_2y_2 + \dots + b_my_m$ ，表示第 2 组变量的第  $i$  种线性组合。

典型相关分析可以根据变量线性组合形成的综合变量  $V$  和  $W$  计算相关系数，并分别找出使得相关系数最大的  $V1$  及其对应的  $W1$ 、次大的  $V2$  及其对应的  $W2$ ，以此类推，直至两个综合变量间没有任何关系。对于该部分的分析，SAS 系统默认给出相关关系由大到小排列的 3 组线性组合。

上述分析过程生成多组综合变量( $V1, W1$ )、( $V2, W2$ )……其中  $V1$  和  $W1$  相关性最大，而  $V2$  和  $W2$  次之，以此类推。在  $V1, V2$ ……之间互不相关，且  $W1, W2$ ……之间也互不相关。那么究竟应该选择多少组或者具体哪一组综合变量来对实际问题进行分析呢？这便是典型相关分析中的综合变量选择问题。

对于变量选择问题，在 SAS 系统中可以根据变量之间的关系计算出所谓的特征根，并根据某一对组合的特征根占有所有特征根总和的比重，计算出特征根贡献率。特征根代表了某一对具体组合的信息，而特征根贡献率则代表了这个组合在所有组合中的作用程度。

如果只选择某对最能说明问题的综合变量组合，则应选择特征根贡献最大的变量组合；如需选择一些能够对实际问题进行解释的变量组和，则可以选择累计特征根贡献比较大的组合。

在通常情况下，除了考虑典型相关系数之外，还倾向于采用典型相关系数平方，即综合变量之间的共享方差在其各自方差中的比例。

除综合变量组合之间的典型相关系数之外，还可以计算综合变量  $V$  和  $W$  与其对应的变量  $x, y$  之间的系数，即典型系数。根据该典型系数的大小可以判定计算出来的综合变量所代表的经济含义。变量  $x, y$  与其对应的  $V$  和  $W$  之间的相关系数在 SAS 系统中也可以进行计算。

在 SAS 系统中, 可以利用 SAS/Analyst 分析员模块下的 “Statistics→Multivariate→ Canonical Correlation” 菜单进行典型相关分析。

**STEP 1)** 进入 SAS/Analyst, 打开 Tech\_CSI.sas7bdat 数据集, 选择 Analyst 分析员的系统菜单 “Statistics→Multivariate→ Canonical Correlation”, 打开典型相关分析对话框, 如图 6-8 所示。

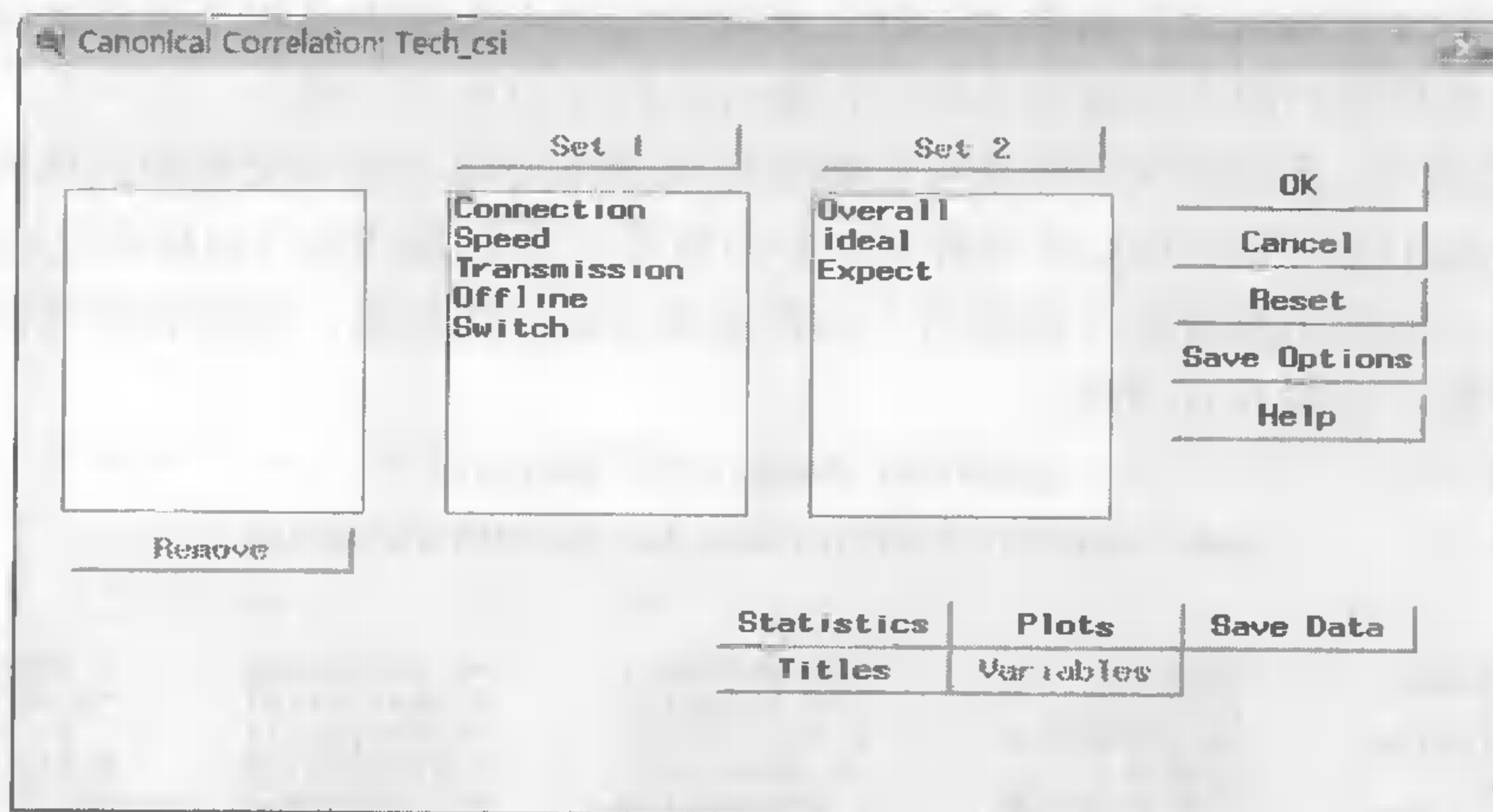


图 6-8 典型相关分析的对话框

**STEP 2)** 典型相关分析对话框中的 “Set 1” 按钮和 “Set 2” 按钮可以分别用于指定两组变量所包含的具体变量。如在变量选择区域中选中 “Connection”、“Speed”、“Transmission”、“Offline” 和 “Switch”, 单击 “Set 1” 按钮, 把上述选中的变量设置为第 1 组变量, 表示通信的技术水平; 同理可以指定 “Overall”、“Ideal” 和 “Expect” 为第 2 组变量, 表示顾客满意度。

在 SAS 系统中, 由 “Set 1” 所代表的综合变量被叫做 “Var 变量”, 由 “Set 2” 所代表的综合变量被称为 “With 变量”。

**STEP 3)** 在图 6-8 所示的对框中, 单击 “OK” 按钮, 可得到典型相关分析的输出结果, 如图 6-9 所示。

The CANCORR Procedure				
Canonical Correlation Analysis				
	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation
1	0.884449	0.883134	0.009748	0.782249
2	0.386433	0.377775	0.038081	0.149331
3	0.118832	0.103867	0.044134	0.014121
Eigenvalues of Inv(E)*H = CanRsq/(1-CanRsq)				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.5924	3.4169	0.9498	0.9498
2	0.1755	0.1612	0.0464	0.9962
3	0.0143		0.0038	1.0000

图 6-9 典型相关系数及特征根

SAS 系统自动给出了从大到小的 3 组典型相关 (Canonical Correlation) 系数及其对应的

特征根 (Eigenvalue) 和特征根贡献率 (Proportion)。此外，系统还给出了各组典型相关系数的调整值及其平方。

从图 6-9 中可以看到，第 1 个典型相关系数为 0.884，典型相关系数的平方为 0.782，其对应特征根贡献率为 0.949 8，即 94.98%，代表了绝大部分的信息。因此，在分析通信技术和满意度的相关性时，只要分析第 1 个典型相关系数即可。该结果表明，从原始变量中综合出来的第 1 对综合变量的相关系数为 0.884，其能够在很大程度上说明了这些变量的相互依存关系，因此技术质量和顾客满意度之间存在着较程度的相关关系。

在 SAS 系统中，典型相关分析还可以输出综合变量与构成综合变量的具体变量之间的典型系数 (Canonical Coefficients)。该系数可以分为原始典型系数 (Raw) 和标准化 (Standardized) 典型系数，其中标准化系数往往能够给人们带来更加直观的印象，更能说明原始变量对综合变量的相对影响，如图 6-10 所示。

Canonical Correlation Analysis					
Raw Canonical Coefficients for the VAR Variables					
		V1	V2	V3	
Connection	接通率	0.0095969211	-0.004325389	0.0999620896	
Speed	网络速度	-0.00080766	0.0494332781	-0.009345822	
Transmission	数据传输速率	0.0057495028	-0.026750111	-0.0404891	
Offline	掉线率	-0.002529989	0.0238652703	0.0177620003	
Switch	总体接入质量	0.0428537143	-0.020842852	-0.055867745	
Raw Canonical Coefficients for the WITH Variables					
		W1	W2	W3	
Overall	总体满意评价	0.0342746695	0.0080255208	-0.099034585	
ideal	与理想相比的满意评价	-0.004209897	0.0648616468	0.0188412023	
Expect	与期望相比的满意情况	0.0240659336	-0.045575939	0.0914328882	
The CANCORR Procedure					
Canonical Correlation Analysis					
Standardized Canonical Coefficients for the VAR Variables					
		V1	V2	V3	
Connection	接通率	0.1760	-0.0793	1.8330	
Speed	网络速度	-0.0164	1.0053	-0.1901	
Transmission	数据传输速率	0.1109	-0.5162	-0.7813	
Offline	掉线率	-0.0585	0.5521	0.4109	
Switch	总体接入质量	0.7882	-0.3833	-1.0275	
Standardized Canonical Coefficients for the WITH Variables					
		W1	W2	W3	
Overall	总体满意评价	0.6376	0.1493	-1.8424	
ideal	与理想相比的满意评价	-0.0776	1.1953	0.3474	
Expect	与期望相比的满意情况	0.4463	-0.8453	1.6957	

图 6-10 原始典型系数和标准化典型系数

因为提取的第 1 组变量对应的特征根贡献率达到了 94.98%，“Var”变量代表“Set 1”指定的综合变量，“With”变量代表“Set 2”指定的综合变量，所以只需对第 1 对综合变量（即 V1 和 W1）进行主要分析即可。根据标准化的典型系数，可以写出 V1 和 W1 综合变量的表达式。

- ①  $V1 = 0.1760 \times \text{接通率} - 0.0164 \times \text{网络速度} + 0.1109 \times \text{数据传输速率} - 0.0585 \times \text{掉线率} + 0.7882 \times \text{总体接入质量}$
- ②  $W1 = 0.6376 \times \text{总体满意评价} - 0.0776 \times \text{与理想相比的满意评价} + 0.4463 \times \text{与期望相比的满意评价}$

在 V1 中，“总体接入质量”的标准化系数最大，所以可以认为 V1 综合变量主要代表总体接入质量，网络速度和掉线率起着副作用（反向作用），但是副作用非常小；在 W1 中，“总

体满意评价”和“与期望相比的满意评价”的标准化系数较大，所以可以认为 *W1* 综合变量代表了满意度的总体评价和与期望相比的满意状况。

此外，在 SAS 系统中，典型相关分析还可以输出综合变量与构成综合变量的具体变量之间的相关系数，如图 6-11 所示。

The CANCORR Procedure					
Canonical Structure					
Correlations Between the VAR Variables and Their Canonical Variables					
		V1	V2	V3	
Connection	接通率	0.8977	-0.0288	0.4041	
Speed	网络速度	0.6458	0.6758	-0.1454	
Transmission	数据传输速率	0.8988	0.0176	-0.1314	
Offline	掉线率	0.4814	0.6384	0.1561	
Switch	总体接入质量	0.9910	0.0655	-0.0632	
Correlations Between the WITH Variables and Their Canonical Variables					
		W1	W2	W3	
Overall	总体满意评价	0.9732	0.1202	-0.1962	
Ideal	与理想相比的满意评价	0.5467	0.7979	0.2539	
Expect	与期望相比的满意情况	0.9453	-0.0329	0.3245	
Correlations Between the VAR Variables and the Canonical Variables of the WITH Variables					
		W1	W2	W3	
Connection	接通率	0.7940	-0.0111	0.0480	
Speed	网络速度	0.5712	0.2612	-0.0173	
Transmission	数据传输速率	0.7949	0.0068	-0.0156	
Offline	掉线率	0.4258	0.2467	0.0185	
Switch	总体接入质量	0.8765	0.0253	-0.0075	
Correlations Between the WITH Variables and the Canonical Variables of the VAR Variables					
		V1	V2	V3	
Overall	总体满意评价	0.8607	0.0464	-0.0233	
Ideal	与理想相比的满意评价	0.4835	0.3084	0.0302	
Expect	与期望相比的满意情况	0.8361	-0.0127	0.0386	

图 6-11 综合变量与原始变量之间的相关系数

在图 6-11 中，系统依次给出了 *Var* 综合变量及构成 *Var* 变量的原始变量之间的相关系数、*With* 综合变量及构成 *With* 变量的原始变量之间的相关系数、构成 *Var* 综合变量的原始变量与 *With* 综合变量之间的相关系数、构成 *With* 综合变量的原始变量与 *Var* 综合变量之间的相关系数。

从上述相关系数中前两种相关系数的分析中可得知，*V1* 与总体接入质量相关系数最大，为 0.9910；*W1* 与总体满意评价、与期望相比的满意评价两者之间的相关系数最大，分别为 0.9732 和 0.9453。该结论与标准化典型系数分析的结论一致。

图 6-11 中的后两种相关系数表明，总体接入质量与总体满意评价、与期望相比的满意评价共同代表的 *W1* 综合变量最相关（相关系数为 0.876 5），而总体满意评价与总体接入质量所代表的 *V1* 综合变量最相关（相关系数为 0.860 7），期望相比的满意评价与总体接入质量所代表的 *V1* 综合变量相关性也很强（相关系数为 0.836 1）。该分析结果与上述分析过程一致。

6.2.2 典型相关系数的显著性检验

典型相关系数与简单相关系数、非参数相关系数一样，都应该进行系数的显著性假设检验。主要根据样本相关系数计算相关的检验统计量，并根据统计量的抽样分布对总体相关系数是否为 0 的假设进行判定。典型相关系数的检验可分为总体显著性检验和降维检验。

1. 相关系数总体显著性检验

总体显著性主要是指所有综合变量之间的总体相关系数是否同时为 0。如果至少有一个

不为 0，则相关系数总体上是显著的。其原假设如下。

$$H_0: CR_1 = CR_2 = \dots = CR_i = 0, (i = 1, 2, \dots, n)$$

在 SAS 系统中，主要有 Wilks' Lambda、Pillai's Trace、Hotelling-Lawley Trace 和 Roy's Greatest Root 等 4 种多元统计量和 F 估计值检验方法。

在 SAS 系统的典型相关分析过程中，系统自动输出相关系数的总体显著性检验结果，如图 6-12 所示。

Multivariate Statistics and F Approximations					
	S=3	M=0.5	N=245		
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.18261816	77.11	15	1358.6	<.0001
Pillai's Trace	0.94570105	45.48	15	1482	<.0001
Hotelling-Lawley Trace	3.78227525	123.82	15	924.11	<.0001
Roy's Greatest Root	3.59240691	354.93	5	494	<.0001

图 6-12 典型相关系数的总体显著性检验

在本例中，4 种检验方法的  $P$  值（“ $Pr>F$ ”）均小于 0.0001，在给定的显著性水平  $\alpha = 0.05$  或 0.01 条件下均非常显著，故拒绝原假设。所以，认为本例的典型相关系数已通过总体显著性检验。

2. 相关系数降维检验

降维检验同属于多元检验方法，可以单独对计算出来的每对综合变量的典型相关系数是否显著进行检验。在实际应用中，仍然可以通过  $P$  值与显著性水平  $\alpha$  的大小进行判定。其原假设如下。

$$H_0: CR_i = 0, (i = 1, 2, \dots, n)$$

在 SAS 系统的典型相关分析过程中，系统自动输出相关系数的降维检验结果，如图 6-13 所示。

Test of H0: The canonical correlations in the current row and all that follow are zero				
Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
0.18261816	77.11	15	1358.6	<.0001
0.83865690	11.33	8	986	<.0001
0.98587902	2.36	3	494	0.0709

图 6-13 典型相关系数的降维显著性检验

在图 6-13 的输出结果中，可以看到本例的前两个典型相关系数的  $P$  值（“ $Pr>F$ ”）小于 0.000 1，是显著的；但是第 3 个典型相关系数  $P$  值为 0.070 9，在  $\alpha = 0.05$  的条件下不是显著的。结合图 6-9 的分析结果来看，其特征根贡献率仅为 0.38%，所以在本例中，该典型相关系数所代表的信息微乎其微，可以忽略不计。

6.2.3 典型相关的冗余分析

在典型相关分析中，所提取的每对综合变量或典型变量都要保证其相关程度达到最大，每个综合变量不仅解释了本组变量的信息，还解释了另一组变量的信息。典型相关系数越大，综合变量解释另一组变量的信息也越多。冗余分析就是要说明综合变量对各组具体观测变量总方差的相互解释程度。

**STEP 1)** 仍然以例 6-3 为例，进入 SAS/Analyst，打开 Tech\_CSI.sas7bdat 数据集，选择 Analyst 分析员的系统菜单 “Statistics→Multivariate→Canonical Correlation”，打开图 6-8 所示的典型相关分析对话框。在该对话框中单击 “Statistics” 按钮，弹出典型相关分析的统计量对话框，如图 6-14 所示。

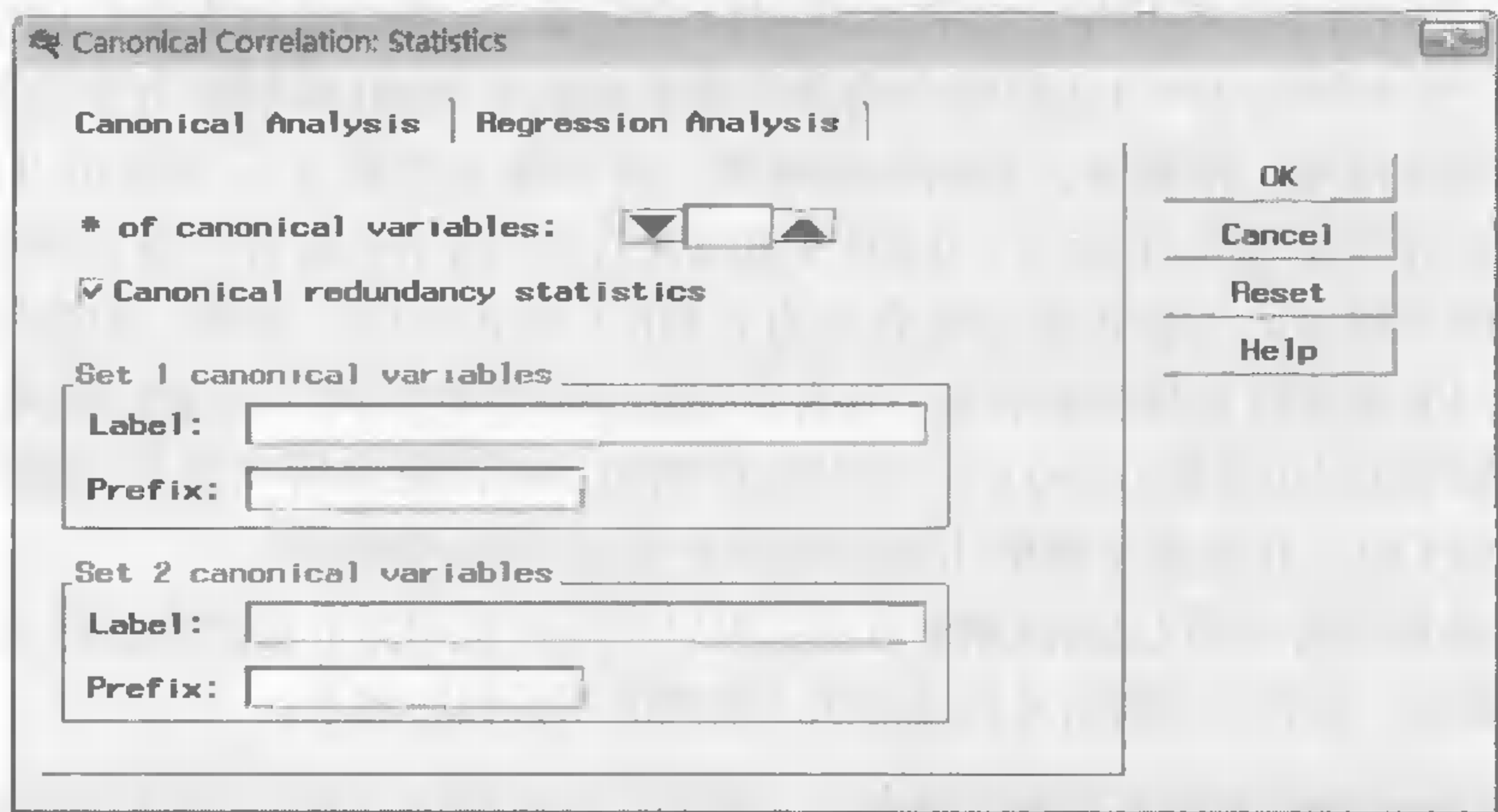


图 6-14 典型相关分析的统计量对话框

**STEP 2)** 在该对话框中，选中 “Canonical redundancy statistics” 复选框，表示指定系统进行典型相关的冗余分析。单击 “OK” 按钮，可以得到以原始数据和标准化数据两种形式表示的冗余分析结果，标准化结果如图 6-15 所示。

The CANCORR Procedure					
Canonical Redundancy Analysis					
Standardized Variance of the VAR Variables Explained by					
Their Own Canonical Variables			The Opposite Canonical Variables		
Canonical Variable Number	Proportion	Cumulative Proportion	Canonical R-Square	Proportion	Cumulative Proportion
1	0.6489	0.6489	0.7822	0.5076	0.5076
2	0.1799	0.8229	0.1493	0.0260	0.5336
3	0.0460	0.8689	0.0141	0.0006	0.5342
Standardized Variance of the WITH Variables Explained by					
Their Own Canonical Variables			The Opposite Canonical Variables		
Canonical Variable Number	Proportion	Cumulative Proportion	Canonical R-Square	Proportion	Cumulative Proportion
1	0.7132	0.7132	0.7822	0.5579	0.5579
2	0.2174	0.9306	0.1493	0.0325	0.5903
3	0.0694	1.0000	0.0141	0.0010	0.5913
Squared Multiple Correlations Between the VAR Variables and the First M Canonical Variables of the WITH Variables					
M			1	2	3
Connection	接通率		0.6304	0.6305	0.6328
Speed	网络速度		0.3262	0.3944	0.3947
Transmission	数据传输速率		0.6319	0.6320	0.6322
Offline	掉线率		0.1813	0.2421	0.2425
Switch	总体接入质量		0.7683	0.7689	0.7690
Squared Multiple Correlations Between the WITH Variables and the First M Canonical Variables of the VAR Variables					
M			1	2	3
Overall	总体满意评价		0.7408	0.7430	0.7435
ideal	与理想相比的满意评价		0.2338	0.3289	0.3298
Expect	与期望相比的满意情况		0.6990	0.6992	0.7007

图 6-15 典型相关分析的冗余分析结果

在图 6-15 所示的前两个表格中，分别分析提取出来的综合变量（*Var* 变量和 *With* 变量）对各组变量的解释能力。如 *V1* 综合变量解释了 64.89% 的组内方差，同时也解释了另外一组变量 55.79% 的方差；而 *W1* 综合变量解释了 71.32% 的组内方差，同时也解释了另外一组变量 50.76% 的方差。由此可见第 1 对综合变量 *V1* 和 *W1* 能够较好地预测另一组变量。

在图 6-15 中的第 3 个表格给出的技术质量综合变量中，各个具体变量与由满意度具体变量提取的前 *M* 个（本例为 *M* = 1,2,3）综合变量的多重相关平方和（即解释方差的累积百分比）。由表中的数据可以得知，接通率、数据传输速率、总体接入质量 3 个变量可以很好地被 *W1* 解释，其累积百分比分别为 0.630 4、0.631 9 和 0.768 3。而 *W1* 综合变量对网络速度和掉线率变量的预测能力较小，其累积百分比分别为 0.326 2 和 0.181 3。同理，根据第 4 个表格的数据也可得知，*V1* 变量对总体满意评价、与期望相比的满意评价两个变量的预测能力较强（其累积百分比分别为 0.740 8 和 0.699 0），而对与理想相比的满意评价变量的预测能力较弱（累积百分比为 0.233 8）。其他两个提取出来的综合变量的分析过程同上。

利用 SAS 编程语言中的 CANCERR 过程也可以实现典型相关分析的全过程。CANCERR 过程的语法比较简单，本书不再给出其具体语法。本例的相关程序如下。

```
proc cancel data=Sasuser.Tech_csi redundancy; /*调用 CANCERR 过程，关键字 redundancy 表示进行
典型相关的冗余分析*/
    var connection speed transmission offline swith; /*var 语句用于指定 Var 综合变量所包含的具体变量*/
    with overall ideal expect; /*with 语句用于指定 With 综合变量所包含的具体变量*/
run;
```

运行程序后，可以得到与上述分析过程一样的结果。

### 6.3 线性回归分析

当变量之间存在相互依存的关系时，还可以进行回归（Regression）分析。“回归”一词来源于高尔顿研究人类身高遗传问题的过程。1870 年，高尔顿在研究人类身高的遗传问题时，发现高个子父母的子女，其身高有低于其父母身高的趋势，而矮个子父母的子女，其身高有高于其父母的趋势，即有“退回”（即 Regression 的原意）到身高均值的趋势。因此，高尔顿首次引入了回归直线、相关系数的概念，始创了回归分析。回归分析的研究领域非常多，有线性回归、非线性回归、定性自变量回归、离散因变量回归等。在社会经济领域的实际应用中，线性回归分析应用非常广泛。

#### 6.3.1 回归分析的基本原理

回归分析与相关分析在理论和方法上具有一致性，变量之间没有关系，就谈不上回归分析或建立回归方程。相关程度越高，回归效果就越好，而且相关系数和回归系数方向一致，可以互相推算。

相关分析中的两个变量之间的地位是对等的，即变量 A 与变量 B 相关等价于变量 B 与变量 A 相关，相关分析的两个变量均为随机变量。而回归分析中要确定自变量和因变量，通常只有因变量是随机变量，人们可以利用回归分析来对研究对象进行预测或控制。

在通常情况下，回归分析往往是通过一条拟合的直线来表示模型的建立。设 *y* 表示因变量，*x* 表示自变量，则有以下回归模型。

$$y = \alpha + \beta x + \varepsilon$$

其中：

- $\alpha$ 和 $\beta$ 是回归模型的参数，被称为回归系数（ $\alpha$ 也可被称为截距项）。
- $\varepsilon$ 是随机误差项或随机扰动项，反映了除  $x$  和  $y$  之间的线性关系之外的随机因素或不可观测的因素。

通常在回归分析中，对 $\varepsilon$ 有以下最为常用的经典假定。

- $\varepsilon$ 的期望值为0。
- $\varepsilon$ 对于所有  $x$  而言具有同方差性。
- $\varepsilon$ 是服从正态分布且相互独立的随机变量。

如果存在多个自变量，则回归模型可以写作为： $y = \alpha + \beta_1x_1 + \beta_2x_2 + K + \beta_ix_i + \varepsilon$

因为上述的直线模型中含有随机误差项，所以回归模型反映的直线是不确定的。回归分析的主要目的是从这些不确定的直线中找出一条最能够代表数据原始信息的直线，并将其作为回归模型来描述因变量和自变量之间的关系。这条直线被称为回归方程，如在只有一个自变量的情况下，可得到直线的形式： $y = \hat{\alpha} + \hat{\beta}x$

因为总体回归模型的参数往往是未知的，人们只能靠样本数据去进行参数估计，所以可用 $\hat{\alpha}$ 和 $\hat{\beta}$ 分别表示回归模型中 $\alpha$ 和 $\beta$ 的参数估计值。方程 $y = \hat{\alpha} + \hat{\beta}x$ 也可被称为估计的回归方程（如图 6-16 所示）。 $\hat{\alpha}$ 是估计的回归直线在  $y$  轴上的截距， $\hat{\beta}$ 是直线的斜率，它表示自变量  $x$  每变动一个单位时，因变量  $y$  的平均变动值， $\hat{y}$  是  $y$  的估计值。

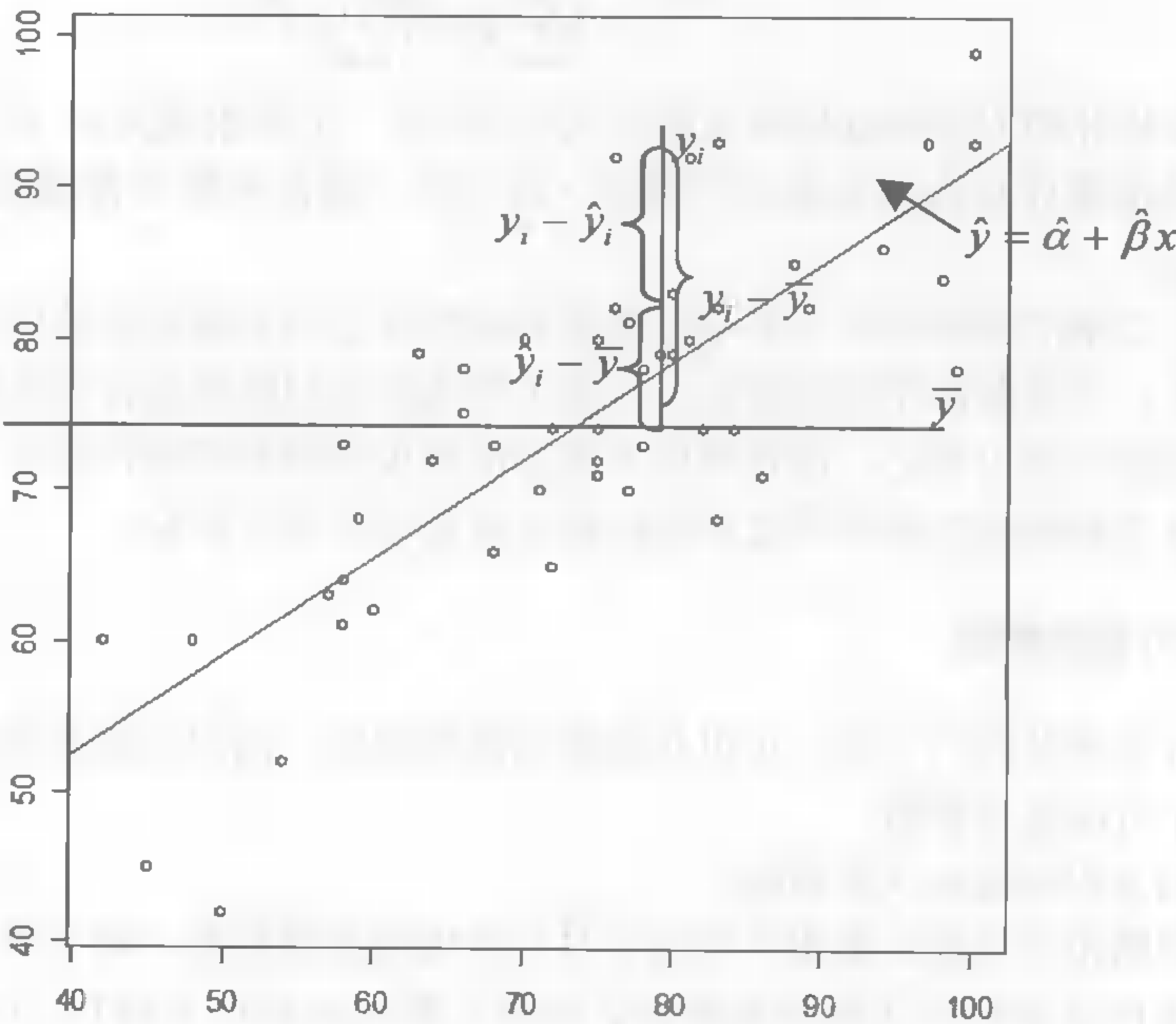


图 6-16 回归方程与散点图

1. 回归方程的参数估计

在通常情况下，把具体某个因变量的观测值记为  $y_i$ ，其均值为  $\bar{y}$ ，把其估计值（即在回归直线上的值）记为  $\hat{y}_i$ 。那么如何在存在随机因素的回归模型中找出最能代表原始数据信息

的回归直线呢？可以通过从因变量的离差入手进行分析。

如图 6-16 所示，因变量的离差为  $y_i - \bar{y}$ ，可以把离差分解为两个部分，即： $y_i - \hat{y}_i$ （残差）和  $\hat{y}_i - \bar{y}$ （回归离差）。每个因变量的离差等于残差与回归离差之和，即  $y_i - \bar{y} = y_i - \hat{y}_i + \hat{y}_i - \bar{y}$ 。对于所有因变量的观测值而言，为了避免正负符号的影响，可以对上述分解出来的 3 个差值求平方得：

$$\sum (y_i - \bar{y})^2 = \sum (y_i - \hat{y}_i)^2 + \sum (\hat{y}_i - \bar{y})^2$$

即总离差平方和（记为 SST）等于残差平方和（SSE）与回归离差平方和（SSR）之和，上述等式可以进行严格教学证明。其证明过程本书不予赘述，请查阅相关统计学原理书籍。

从图 6-16 可以看出，残差越小， $y_i$  就越往回归直线靠近。对于所有的因变量而言，残差平方和越小，观测值就越往回归直线靠近。因此，当残差平方和（SSE）达到极小值时，即  $\sum (y_i - \hat{y}_i)^2 \Rightarrow$  最小值时，估计出来的回归直线能在最大程度上代表原始数值的信息。

当  $\sum (y_i - \hat{y}_i)^2 = \sum [y_i - (\hat{\alpha} + \hat{\beta}x)]^2 \Rightarrow \min$  时，可以利用微分求极值的方法，分别对  $\sum [y_i - (\hat{\alpha} + \hat{\beta}x)]^2$  求  $\hat{\alpha}$  和  $\hat{\beta}$  的偏微分，并使之同时为 0。然后求解联立方程组，便可计算出  $\hat{\alpha}$  和  $\hat{\beta}$  的具体数值，作为回归模型中  $\alpha$  和  $\beta$  的参数估计值，如下所示。

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

$$\hat{\beta} = \frac{n \sum y_i x_i - \sum y_i \sum x_i}{n \sum x_i^2 - (\sum x_i)^2}.$$

这种参数估计的方法和过程是在随机误差项  $\varepsilon$  是一个期望值为 0、对于所有  $x$  而言具有同方差性、服从正态分布且相互独立的假定下进行的，通常被称为普通最小二乘法（Ordinary Least Squares）。

普通最小二乘法同样适用于多个自变量和因变量之间的模型参数估计。在 SAS 系统的回归分析功能中，系统默认使用普通最小二乘法对线性回归模型进行参数估计。

回归分析的主要内容之一便是利用上述方法对模型的参数进行估计。对模型进行参数估计之后，还应当对模型的拟合程度和回归系数的显著性进行检验。

## 2. 回归方程的检验

对于估计出来的回归方程，可以从模型的解释程度、回归方程总体显著性及回归系数的显著性等几个方面进行检验。

### （1）回归方程的拟合优度检验。

回归方程的拟合优度主要用于判定  $\hat{y}_i$  估计  $y$  的可靠性问题，通常用来衡量模型的解释程度。拟合优度检验是建立在模型参数估计时对总离差平方和（SST）分解的基础上的，SST 可以分解为残差平方和（SSE）和回归离差平方和（SSR），通常使用回归离差平方和 SSR 占总离差平方和 SST 的比重来判断模型的解释能力，即：

$$R^2 = \frac{\sum (\hat{y}_i - \bar{y})^2}{\sum (y_i - \bar{y})^2} = 1 - \frac{\sum (y_i - \hat{y}_i)^2}{\sum (y_i - \bar{y})^2}$$

其中  $R^2$  表示拟合优度的判定系数或决定系数, 其取值范围为  $[0, 1]$ 。  $R^2$  越接近于 1, 说明变量之间的相互依存关系越密切, 其相互依存关系就越接近于函数关系, 两变量之间的相关程度越高, 回归方程的拟合程度越好。所以,  $R^2$  越接近 1, 模型的解释程度越好, 模型越精确。

但是,  $R^2$  的数值与自变量的数目有关, 即自变量的个数越多,  $R^2$  越大。这在一定程度上削弱了  $R^2$  的评价能力, 因此可考虑剔除自变量数目影响后的修正  $R^2$ 。

### (2) 回归方程整体显著性检验。

利用普通最小二乘法拟合出来的回归方程都是由样本数据进行的, 那么用它来对总体进行推断是否显著呢? 可以对回归方程整体的显著性进行检验。

回归方程整体显著性检验主要是检验因变量和自变量之间的线性关系是否显著。其原假设和备择假设如下。

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_i = 0$$

$$H_1: \beta_i \text{ 不全为零}$$

对于回归方程整体显著性检验可用  $F$  检验进行, 在 SAS 系统中, 可以计算出用于检验判定的  $P$  值。如果  $P$  值小于理论显著性水平  $\alpha$  值, 可认为在显著性水平  $\alpha$  条件下, 回归方程整体显著。

### (3) 回归方程系数显著性检验。

如果模型的线性关系显著, 还应对模型参数估计的结果, 即回归方程的系数进行显著性检验, 用于考察单个自变量与因变量的线性关系是否成立。其原假设和备择假设如下。

$$H_0: \beta_i = 0 (i = 1, 2, \cdots, k)$$

$$H_1: \beta_i \neq 0 (i = 1, 2, \cdots, k)$$

回归方程系数显著性检验要求对所有估计出来的回归系数分别进行检验 (截距项通常不进行显著性检验), 可以利用  $t$  检验进行。在 SAS 系统中, 可以计算出每个回归系数所对应的  $P$  值。如果某个系数对应的  $P$  值小于理论显著性水平  $\alpha$  值, 可认为在显著性水平  $\alpha$  条件下, 该回归系数是显著的。

在有些情况下, 在没有任何关联的变量之间进行回归分析, 也可能得到显著的检验结果, 从而会对分析过程造成不良的影响。因此, 在进行回归分析之前, 必须考虑好变量之间的关系及其所代表的经济含义。

## 3. 回归方程的预测

回归预测是一种有条件的预测, 依据估计出来的回归方程, 在给定自变量数值的条件下, 对因变量进行预测。其预测的基本公式为:

$$\hat{y}_f = \hat{\alpha} + \hat{\beta}x_f$$

其中  $x_f$  是另外给定的自变量的值,  $\hat{y}_f$  为根据回归方程计算出来的预测值。

## 6.3.2 一元线性回归分析

一元线性回归是回归分析中最简单的一种形式, 主要考察单独一个自变量对因变量的影响。其模型如:  $y = \alpha + \beta x + \varepsilon$

一元线性回归分析的基本步骤如下。

依据变量之间的关系，判断其是否是线性关系。如果是线性关系，可以利用 4.3.1 小节介绍的方法进行回归模型的参数估计，然后根据参数估计的结果进行检验。

在检验过程中，可以先对模型的解释能力进行拟合优度检验。如果拟合优度的判定系数非常小，说明建立的回归方程解释能力较差。在进行回归分析的过程中，可能还有其他重要因素没有加入到模型当中，可以考虑增加有重要影响的自变量；回归方程整体显著性如果不显著，说明变量之间的线性关系不明显，不适合做线性回归；在拟合优度比较高、方程整体显著的情况下，对回归系数进行检验，通过显著性检验的回归系数才对因变量有解释能力。

只有通过检验的模型才能够充分描述变量之间的关系，建立的模型才有现实意义。



例 6-4

在通常情况下，一个国家或地区的犯罪率在很大程度上受到国民素质的影响，而反映国民素质的一个重要指标便是文盲率（或识字率）。在正常逻辑思维当中，文盲率越低，其普法程度就越低，可能会对社会造成的危害越大。为了研究文盲率与犯罪率之间的关系，现在全世界范围内收集到来自于不同区域的 50 个国家或地区的文盲率与谋杀犯罪率的数据（详见 Murder.sas7bdat），如表 6-4 所示。试对文盲率与谋杀犯罪率进行回归分析。

表 6-4 50 个国家或地区的文盲率与犯罪率数据

区域 (Division)	文盲率 (%) (Illiteracy)	谋杀犯罪率 (%) (Murder)	区域 (Division)	文盲率 (%) (Illiteracy)	谋杀犯罪率 (%) (Murder)
East South Central	2.1	15.1	Mountain	0.6	5
Pacific	1.5	11.3	West North Central	0.6	2.9
Mountain	1.8	7.8	Mountain	0.5	11.5
West South Central	1.9	10.1	New England	0.7	3.3
Pacific	1.1	10.3	Middle Atlantic	1.1	5.2
Mountain	0.7	6.8	Mountain	2.2	9.7
New England	1.1	3.1	Middle Atlantic	1.4	10.9
South Atlantic	0.9	6.2	South Atlantic	1.8	11.1
South Atlantic	1.3	10.7	West North Central	0.8	1.4
South Atlantic	2	13.9	East North Central	0.8	7.4
Pacific	1.9	6.2	West South Central	1.1	6.4
Mountain	0.6	5.3	Pacific	0.6	4.2
East North Central	0.9	10.3	Middle Atlantic	1	6.1
East North Central	0.7	7.1	New England	1.3	2.4
West North Central	0.5	2.3	South Atlantic	2.3	11.6
West North Central	0.6	4.5	West North Central	0.5	1.7
East South Central	1.6	10.6	East South Central	1.7	11
West South Central	2.8	13.2	West South Central	2.2	12.2
New England	0.7	2.7	Mountain	0.6	4.5

续表

区域 (Division)	文盲率 (%) (Illiteracy)	谋杀犯罪率 (%) (Murder)	区域 (Division)	文盲率 (%) (Illiteracy)	谋杀犯罪率 (%) (Murder)
South Atlantic	0.9	8.5	New England	0.6	5.5
New England	1.1	3.3	South Atlantic	1.4	9.5
East North Central	0.9	11.1	Pacific	0.6	4.3
West North Central	0.6	2.3	South Atlantic	1.4	6.7
East South Central	2.4	12.5	East North Central	0.7	3
West North Central	0.8	9.3	Mountain	0.6	6.9

本例要研究文盲率对谋杀犯罪率的影响，因此因变量为谋杀犯罪率（Murder），自变量为文盲率（Illiteracy），二者之间的经济含义明显。

在进行回归分析之前，应当对两个变量之间是否是线性关系进行研究。根据本例变量绘制散点图（图略），可发现二者之间在一定程度上呈线性关系。

**STEP 1** 在 SAS/Analyst 中同样可进行回归分析。进入 SAS/Analyst，打开 Murder.sas7bdat 数据集，选择系统菜单 “Statistics → Regression → Simple”，弹出一元回归对话框，如图 6-17 所示。

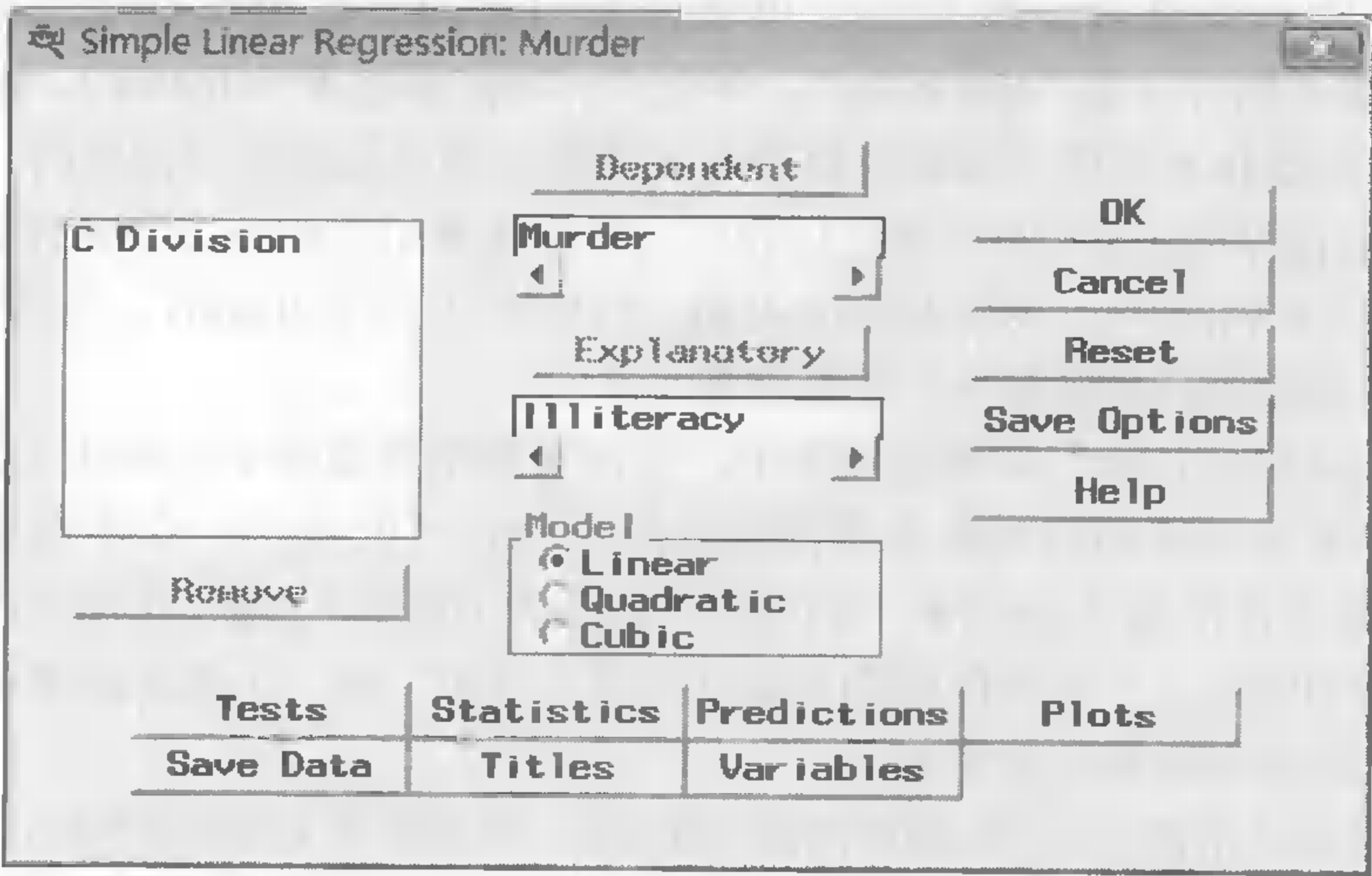


图 6-17 一元回归对话框

在一元回归对话框中的“Model”分栏下，可以指定系统进行 3 种类型的回归。

- Linear: 线性回归，拟合形如  $\hat{y} = \hat{\alpha} + \hat{\beta}x$  的直线。
- Quadratic: 二次项回归，拟合形如  $\hat{y} = \hat{\alpha} + \hat{\beta}_1x + \hat{\beta}_2x^2$  的曲线（即抛物线）。
- Cubic: 三次多项式回归，拟合形如  $\hat{y} = \hat{\alpha} + \hat{\beta}_1x + \hat{\beta}_2x^2 + \hat{\beta}_3x^3$  的曲线。

**STEP 2** 在默认情况下，系统自动指定进行一元线性回归。选中“Murder”变量，单击“Dependent”按钮把其指定为因变量；选中“Illiteracy”变量，单击“Explanatory”按钮，把其指定为自变量。在“Model”分栏下选中“Linear”单选框，表示进行线性回归分析。此外“Predictions”按钮可以设定依据回归方程进行预测（本例不进行预测）。单击“OK”按钮，可以得到图 6-18 所示的回归分析结果。

The REG Procedure						
Model: MODEL1						
Dependent Variable: Murder Murder						
Number of Observations Read				50		
Number of Observations Used				50		
Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	1	329.98270	329.98270	46.89	<.0001	
Error	48	337.76310	7.03673			
Corrected Total	49	667.74580				
Root MSE		2.65268	R-Square	0.4942		
Dependent Mean		7.37800	Adj R-Sq	0.4836		
Coeff Var		35.95397				
Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	2.39678	0.81844	2.93	0.0052
Illiteracy	Illiteracy	1	4.25746	0.62171	6.85	<.0001

图 6-18 一元线性回归的结果

在图 6-18 所示结果的中间部分，可以找到用于拟合优度检验的判定系数  $R^2$ （即“R-Square”）和修正的  $R^2$ （即“Adj R-Sq”）。本例中，判定系数  $R^2 = 0.4942$ ，修正  $R^2 = 0.4836$ ，拟合程度不高。但是在本例所示的截面数据中，该拟合程度还是可以接受的。

在“Analysis of Variance”的方差分析表中，可以计算用于回归方程整体显著性检验的  $F$  统计量的值（即“F Value”）。本例对应的  $P$  值（“Pr>F”）小于 0.0001，所以在  $\alpha = 0.05$  的条件下非常显著，因此回归方程整体上是显著的。

在“Parameter Estimates”参数估计表中，可以看到用普通最小二乘法（OLS）对回归模型的参数估计结果及对应回归系数显著性检验的  $P$  值（“Pr>|t|”）。在本例中，截距项（即 Intercept）的参数估计值为 2.39678，其对应的  $P$  值为 0.0052（通常不对截距项进行显著性检验）；自变量“Illiteracy”对应的回归系数估计值为 4.25746，其对应的  $P$  值小于 0.0001，在给定的显著性水平条件下非常显著。

综上所述，例 6-4 所建立的回归模型拟合程度尚好，模型整体上是显著性的，回归系数也是显著的。因此，可以依据图 6-18 所示的参数估计表的数值写出回归方程式： $Murder = 2.39678 + 4.25746 \times Illiteracy$ ，并根据此方程式对自变量与因变量之间的关系进行分析。

从上述通过拟合优度和显著性检验的方程中可知，当文盲率每增加/降低 1 个单位时，谋杀犯罪率会平均增加/降低 4.25746 个单位。具体而言，文盲率每增加/降低 1 个百分点时，谋杀犯罪率平均增加/降低 4.25746 个百分点。

模型的残差项  $\varepsilon$  应当符合经典假定（详见 4.3.1 小节）。因此，在进行回归分析的过程中，还应当对残差项是否符合假定进行判定。只有在符合假定的前提条件下，上述用 OLS 方法估计出来的回归方程才有解释能力。在 SAS 的一元线性回归模型中，通常可用残差图来判定残差是否与变量相关，用 P-P 图或 Q-Q 图来判定残差项是否符合正态分布。

图 6-17 所示的一元回归对话框中的“Plots”按钮可用于设置系统绘制用于模型残差项假设判定的图形，如图 6-19 所示。

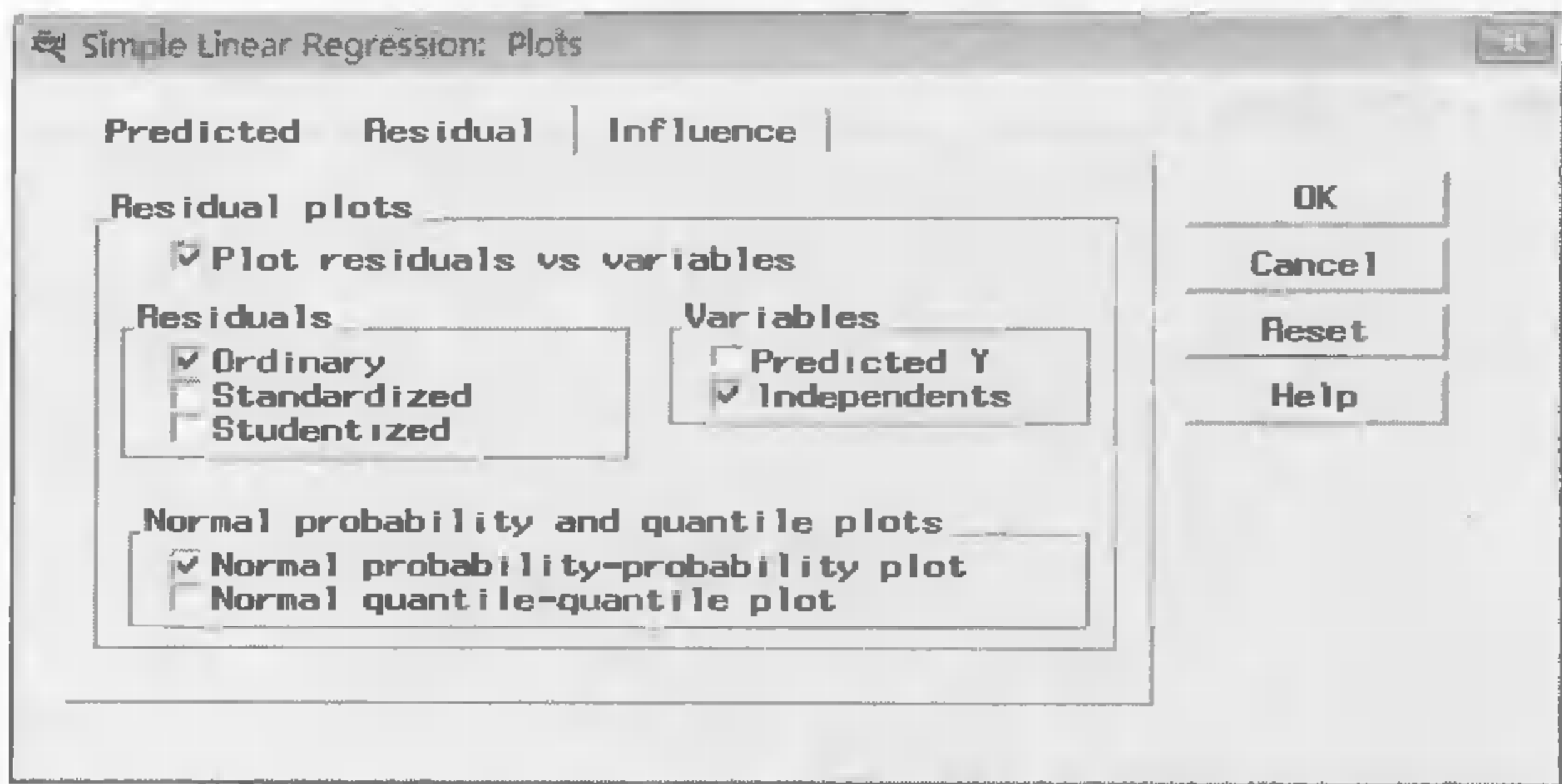


图 6-19 一元回归绘图对话框

在一元回归绘图对话框中，可以按照不同变量绘制 3 类不同的图形。其中，“Predicted”选项卡可以绘制因变量观测值与预测值的散点图，“Residual”选项卡可以依据模型残差项绘制各种不同图形；“Influence”选项卡可以绘制影响统计量和变量之间的散点图。在通常情况下，为了考察残差项是否符合假定，可使用“Residual”选项卡绘制残差和变量之间的散点图。

在“Residual”选项卡中，选中“Plot residuals vs variables”复选框，然后在“Residuals”分栏下选择残差的具体形式（可复选“Ordinary”原始形式、“Standardized”标准化形式、*t* 分布化形式），本例选择残差的原始形式；在“Variable”分栏下选择变量（可复选因变量预测值和自变量），本例选择自变量。在“Normal probalility and quantile plots”分栏下可复选用于判定残差是否服从正态分布的 P-P 图(Normal probability-probability plot)和 Q-Q 图(Normal quantile-quantile plot)，本例选择 P-P 图。单击“OK”按钮返回一元回归对话框。在该对话框中，单击“OK”按钮，除了可以得到图 6-17 所示的回归分析结果之外，还可得到图 6-20 所示的残差图及图 6-21 所示的 P-P 图。

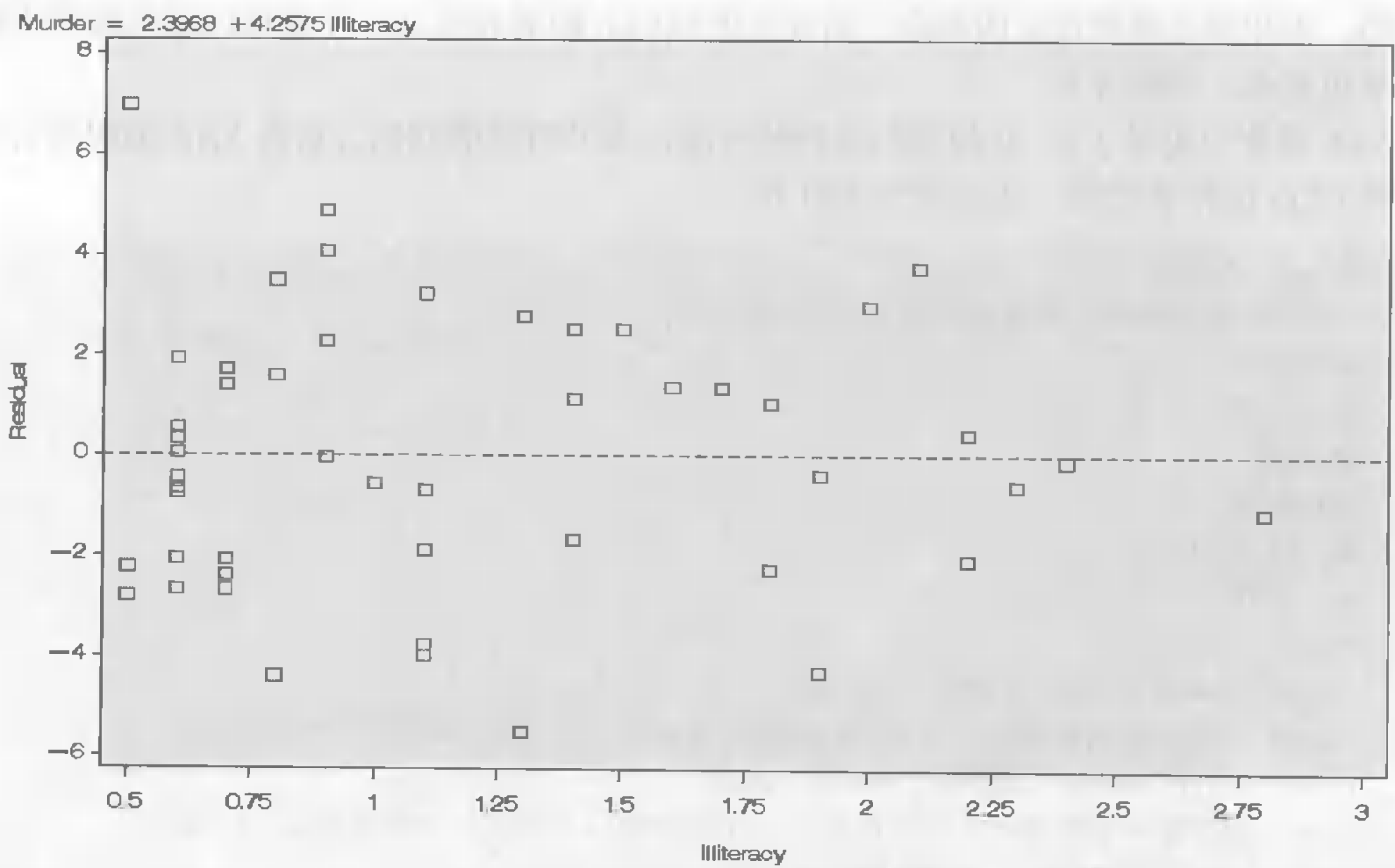


图 6-20 自变量与残差的散点图

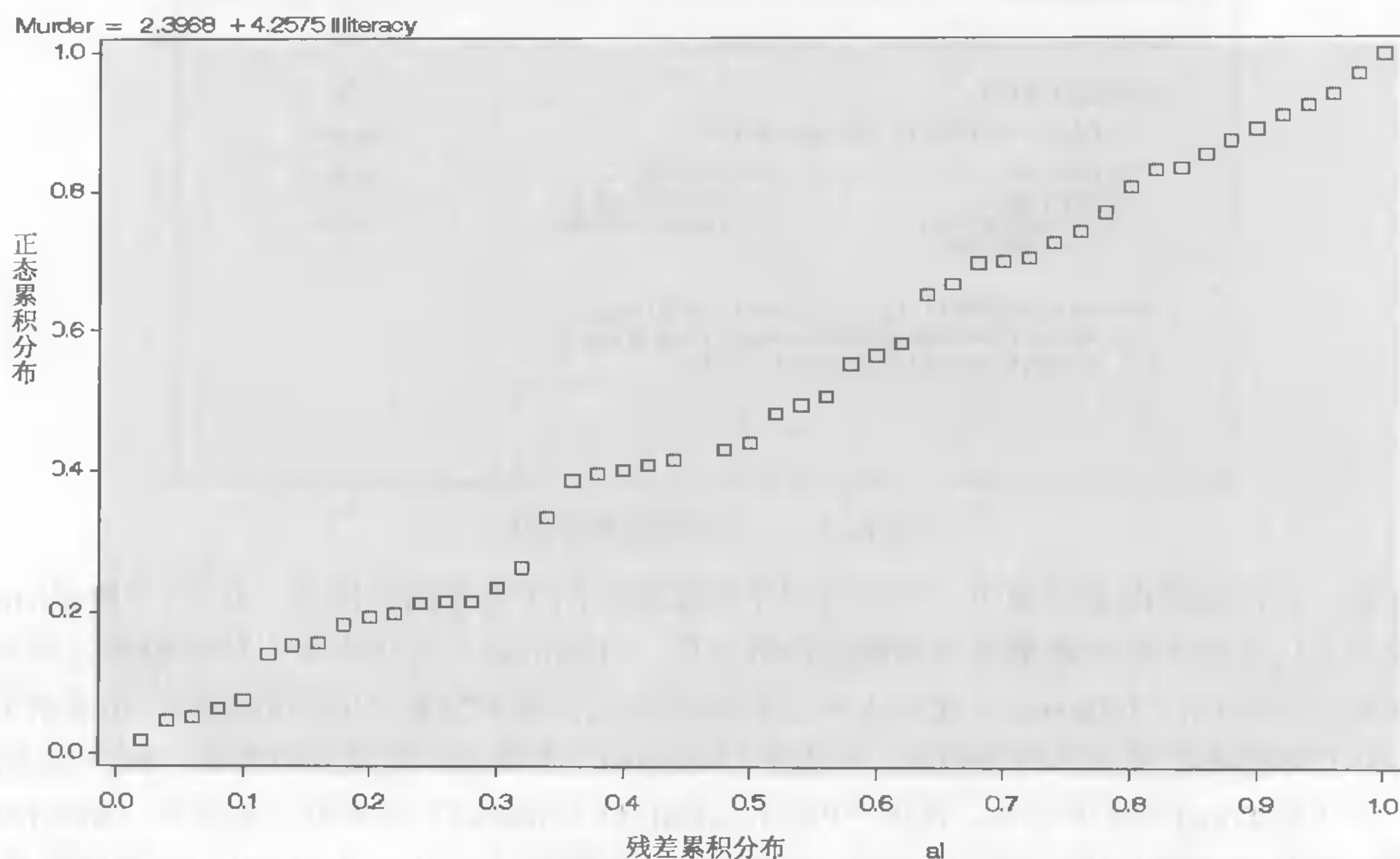


图 6-21 P-P 图

从图 6-20 中可以明显看出，自变量与残差之间的关系不明显，基本上无关，符合 $\varepsilon$ 对于所有 $x$ 而言具有同方差性的假定；而残差大体分布在 $[-6,6]$ 之间，其均值与 0 非常接近，故符合 $\varepsilon$ 零均值的假定。

P-P 图主要用于检验随机变量是否符合正态分布，如果变量符合正态分布，则 P-P 图上的散点表现为集中于一条矩形的对角线上。从图 6-21 所示的结果来看，残差项具有正态分布的趋势，因此符合模型设定的假定，所建立的回归方程更有意义，并且可以用来解释文盲率与谋杀犯罪率之间的关系。

SAS 系统中提供了近 10 种回归分析的方法，本节所述的回归分析在 SAS 编程语言中可以利用 REG 过程来实现，其主要语法如下。

```
proc reg < 选项>;
  <模型标签:> model 因变量=<自变量> </选项>;
  by 变量;
  freq 变量;
  id 变量;
  var 变量;
  weight 变量;
  add 变量;
  delete 变量;
  <标签:> mtest <方程, ..., 方程> </选项>;
  output < out=输出数据集 > 统计量关键字=变量名 < ... 统计量关键字=变量名>;
  paint <条件| allobs> </选项> | < status | undo>;
  plot <y 轴变量*x 轴变量> <=绘图标记> < ...y 轴变量*x 轴变量> <=绘图标记> </选项>;
  print <选项> < anova > < modeldata >;
  refit;
```

```
restrict 方程, ..., 方程;
reweight <条件| allobs> </选项> | <status | undo>;
<标签:> test 方程, <, ..., 方程> </选项>;
```

在 REG 过程中, by、freq、var、weight、id 语句的功能与在前面介绍过的语句的功能一致, 其他常用语句的功能如下。

- MODEL: 指定回归模型, 等号左边为因变量, 等号右边为自变量。其中“Label”可选项可指定模型的标签。
- ADD: 在回归模型中增加自变量。
- DELETE: 从回归模型中剔除自变量。
- MTEST: 在多元因变量模型中进行多元检验。
- OUTPUT: 输出一个新的数据集, 该数据集中包括预测值、残差及其他用于诊断的统计量。
- PAINT: 在散点图中描点。
- PLOT: 绘制散点图。
- PRINT: 显示能够进行模型选项调整的信息。
- REFIT: 重新拟合模型。
- RESTRICT: 在进行模型参数估计时, 设定线性约束。
- REWEIGHT: 从分析中剔除特定观测值或改变所使用观测值的权重。
- TEST: 对参数估计的线性方程进行  $F$  检验。

在上述常用语句中, MODEL 语句是 REG 过程必不可少的语句, 同时也是指定回归模型的最重要的语句。不仅可指定模型中的因变量、自变量, 还可指定模型选项及输出结果选项。



注意

MODEL 语句中的变量必须是所用数据集中的现有变量, 而不能是现有变量的变换变量, 即如果要在模型中加入  $x$  变量的对数形式, 必须先和数据集中生成一个新的变量来代表  $x$  的对数, 如生成 Log\_x 变量来代表  $x$  变量的对数, 则模型中应当引入变量 Log\_x 而不是  $\text{Log}(x)$  函数。

MODEL 语句的常用选项如下。

- NOINT: 不估计模型的截距项。
- STB: 估计模型的标准参数估计结果。
- CLI: 给出因变量预测值的  $1-\alpha$  置信区间上下限 ( $\alpha$  的具体数值可在 REG 过程选项中加入关键字 “Alpha=” 来指定)。
- CLM: 给出因变量期望值的置信区间上下限。
- R: 输出每个样本的因变量预测值、残差及标准误差。
- P: 输出因变量观测值、预测值及残差。

此外, 在进行多元回归分析时, 利用 MODEL 语句的选项还可对变量筛选的方法进行设置 (详见 4.3.3 小节)。

本例利用 REG 过程进行上述回归分析, 具体程序如下。

```
proc reg data=Sasuser.Murder;          /*调用 REG 过程进行线性回归分析*/
  model Murder=Illiteracy;             /*指定回归分析模型, 因变量为 Murder, 自变量为 Illiteracy*/
  plot r.*Illiteracy;                  /*绘制残差与自变量的散点图, 其中关键字 “r.” 表示残差*/
```

```
plot npp.*r.;           /*绘制 P-P 图*/
plot nqq.*r.;           /*绘制 Q-Q 图*/
run;
```

运行程序后，可以得到与上述过程一致的分析结果。

6.3.3 多元线性回归分析

在通常情况下，对因变量产生影响的自变量可能不止一个，有可能有多个。如一个人的体重可能会受到其身高、血型、生活习惯、收入水平等变量的影响。对于多个变量对因变量的影响，可以考虑利用多元线性回归分析的方法进行分析。多元线性回归模型如下。

$$y = \beta_0 + \beta_1x_1 + \beta_2x_2 + \cdots + \beta_nx_n + \varepsilon$$

$\varepsilon$  仍然服从零均值、相互独立且同方差服从正态分布等经典假定。与一元线性回归一样，利用普通最小二乘法可以求出对参数  $\beta_0, \beta_1, \beta_2, \cdots, \beta_n$  的估计值  $\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \cdots, \hat{\beta}_n$ ，即可得到多元回归方程： $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1x_1 + \hat{\beta}_2x_2 + \cdots + \hat{\beta}_nx_n$ 。

多元回归方程的拟合优度检验、方程整体显著性检验与回归系数显著性检验过程与一元回归方程的检验过程一样。



例 6-5

求职人员应聘或跳槽至某个公司，公司 HR 往往会根据其年龄大小、工作经验及学历等诸多因素来决定其起始薪酬。为了考察求职人员起始薪酬的影响因素，现收集了 471 名公司雇员的背景信息（详见 Salary.sas7bdat 数据集，数据来源于 SPSS 示例数据），具体信息如表 6-5 所示。试对求职人员的起始薪酬及其影响因素进行回归分析。

表 6-5 求职人员基本信息变量表

变 量 名	变 量 标 签	属 性
ID	编号	定量变量
Gender	性别	男、女
Education	教育年限	定量数据
Position	职位	普通员工、主管、经理
Current_Salary	目前薪水	定量数据
Begin_Salary	起始薪水	定量数据
Experience	工作经历（周）	定量数据
Age	年龄	定量数据

在表 6-5 所示的背景资料中，“Gender”和“Position”变量都是定性变量。定性自变量的回归有其特殊的处理方法（详见 4.4 节），本例的分析暂不考虑定性变量对因变量的影响。现考虑其余定量变量对起始薪酬的影响。

**STEP 1** 进入 SAS/Analyst，打开 Salary.sas7bdat 数据集，选择系统菜单“Statistics → Regression → Linear”，弹出图 6-22 所示的多元线性回归对话框。

**STEP 2** 选中“Begin\_Salary”变量，单击“Dependent”按钮，把其指定为因变量；选中其余的“Education”、“Current\_Salary”、“Experience”和“Age”定量变量，单击“Explanatory”按钮，把其指定为解释变量。然后单击“OK”按钮，便可得到图 6-23 所示的默认多元线性

回归分析结果。

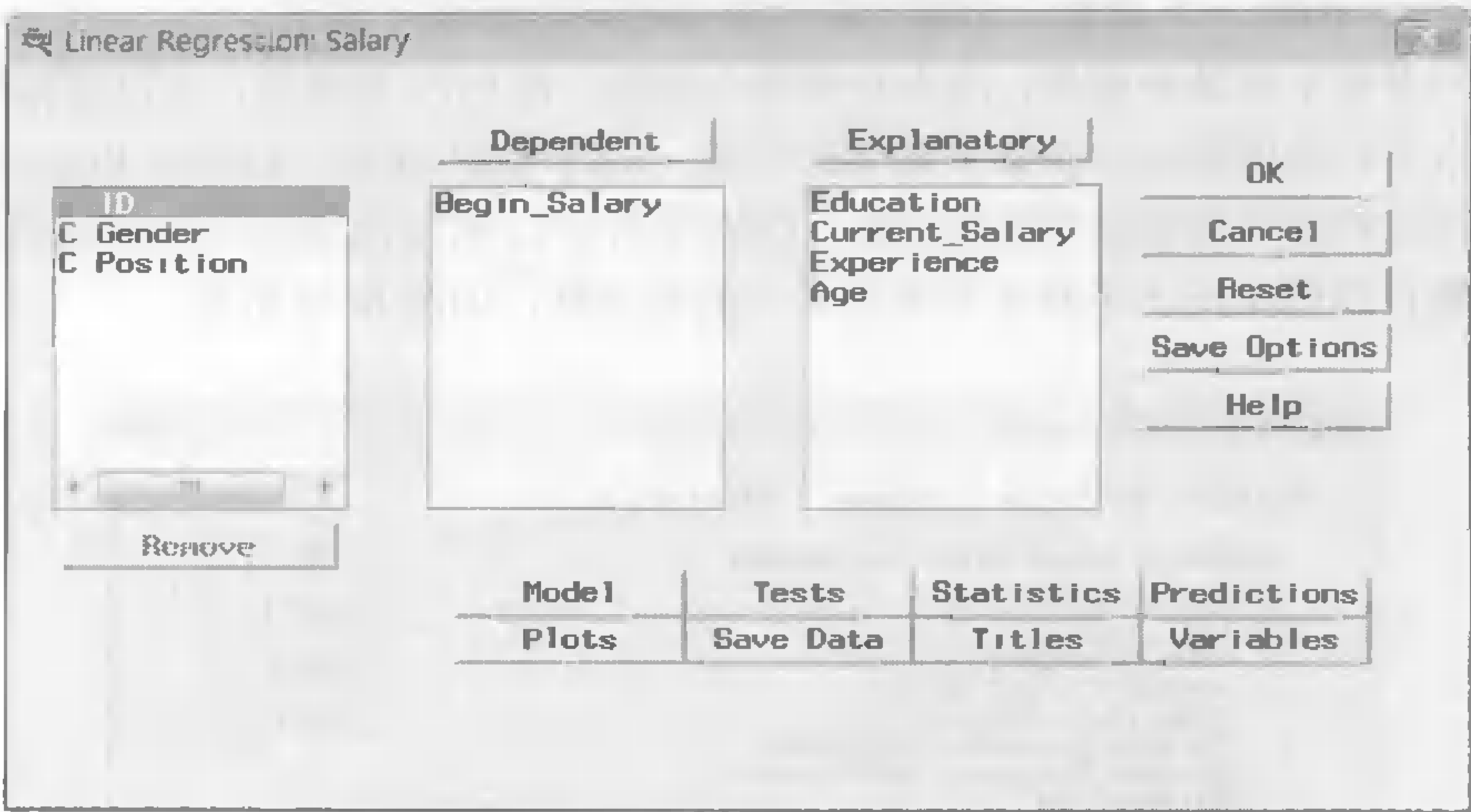


图 6-22 多元线性回归对话框

The REG Procedure						
Model: MODEL1						
Dependent Variable: Begin_Salary 起始薪水						
Number of Observations Read				471		
Number of Observations Used				446		
Number of Observations with Missing Values				25		
Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	4	17053454218	4263363554	432.15	<.0001	
Error	441	4350670512	9865466			
Corrected Total	445	21404124729				
Root MSE		3140.93394	R-Square	0.7967		
Dependent Mean		16998	Adj R-Sq	0.7949		
Coeff Var		18.47865				
Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	-3160.68053	1288.78417	-2.45	0.0146
Education	教育年限	1	516.49720	70.96902	7.28	<.0001
Experience	工作经历(周)	1	8.56698	2.44007	3.51	0.0005
Age	年龄	1	23.38044	21.96196	1.06	0.2876
Current_Salary	目前薪水	1	0.32172	0.01227	26.22	<.0001

图 6-23 多元线性回归分析结果

在多元回归分析结果中，首先看到的结果是样本情况和模型的简单描述，主要内容有模型因变量的变量名、观测样本容量及样本中的缺失值。

在“Analyst of Variance”表格中，输出了回归方程整体显著性检验的过程和结果。在本例中，回归方程整体显著性检验的  $F$  统计量值为 398.81，其对应的  $P$  值（“ $Pr>F$ ”）小于 0.000 1，非常显著。因此，本例所构建的多元回归模型整体上是显著的。此外，回归模型拟合优度的判定系数  $R^2$ （即“R-Square”）和修正后的判定系数（即“Adj R-Sq”）分别为 0.783 4 和 0.781 5，拟合程度较高。

在“Parameter Estimates”表格中，可以得到方程回归系数的估计和显著性检验结果。通过各自变量对应回归系数的  $t$  检验  $P$  值（“ $P>|t|$ ”）的大小，可以判定各个回归系数是否显著（截距项通常不用检验）。在本例中，在显著性水平  $\alpha = 0.05$  下，“Education”、“Experience”

和“Current\_Salary”变量的  $P$  值均远远小于 0.05，因而在模型中影响显著；而“Age”变量在显著性水平  $\alpha = 0.05$  下不显著，说明年龄对起始薪酬的影响不显著。

对于回归系数不显著的变量，应当在模型中剔除。在 SAS 系统中，可以根据用户自行设定的显著性水平自动剔除回归系数不显著的变量。SAS/Analyst 的“Linear Regression”提供了 8 种在建模过程中自动剔除变量的方法，在进行图 6-22 所示的模型变量设置过程中，单击“Model”按钮可以弹出用于选择 8 种方法之一的对话框，如图 6-24 所示。

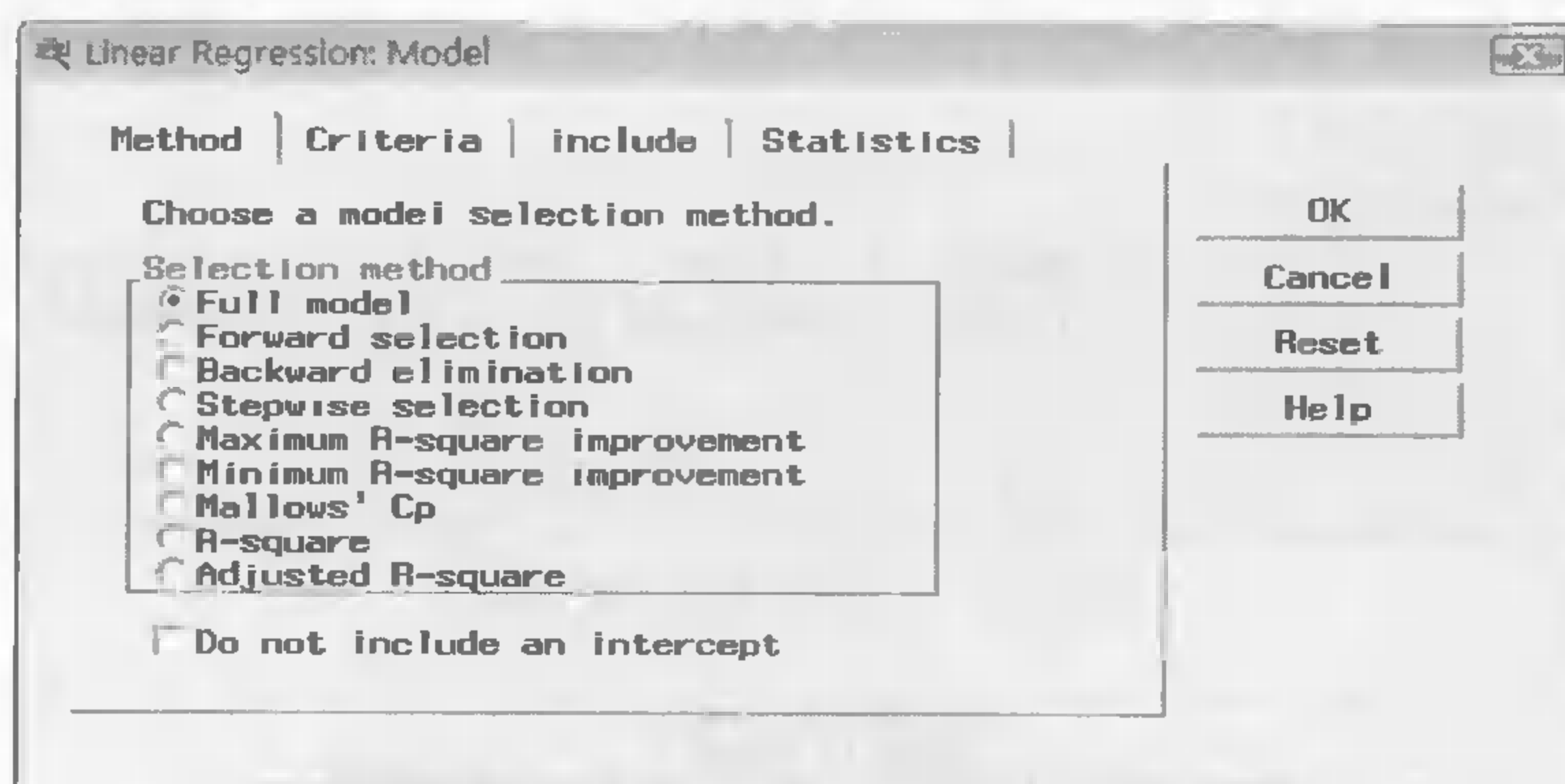


图 6-24 多元线性回归模型设置对话框的 Method 选项卡

在模型设置对话框中的“Method”选项卡下，列示了可供选择的 9 种模型变量选择方法，其中默认的“Full model”表示不进行任何变量剔除和筛选，直接根据用户指定的自变量进行回归分析。其余 8 种方法的具体功能与筛选过程如下。

- Forward selection: 向前引入法。即向模型中逐个引入变量。建模伊始，模型中没有自变量，每当引入一个自变量之后，便计算回归方程  $F$  统计量的值及其对应的  $P$  值，系统根据用户设定的  $P$  值阈值决定该变量是否应当引入。
- Backward elimination: 向后剔除法。即从模型中逐个剔除变量。建模伊始，模型中包含所有的自变量，每当剔除一个自变量之后，便计算回归方程  $F$  统计量的值及其对应的  $P$  值，系统根据用户设定的  $P$  值阈值决定该变量是否应当剔除。
- Stepwise selection: 逐步回归法。即根据用户设定的回归系数显著性检验  $P$  值的引入阈值，逐个向模型引入自变量；然后重新计算模型中所有系数的  $P$  值，根据用户设定的剔除阈值进行变量筛选。
- Maximum R-square improvement: 最大  $R^2$  增量法。即穷尽所有变量组合所构成的模型，找出使得模型拟合优度  $R^2$  增加最大的模型。
- Minimum R-square improvement: 最小  $R^2$  增量法。类似于最大  $R^2$  增量法，不同的是最后选择的模型是使得  $R^2$  增加最小的模型。
- MallowsCp: Mallows Cp 统计量法，即根据 Mallows Cp 统计量进行模型变量选择。
- R-square:  $R^2$  选择法，即根据  $R^2$  进行模型变量选择。
- Adjusted R-square: 修正  $R^2$  选择法，即根据修正  $R^2$  进行模型变量选择。

在本例中，以逐步回归法为例进行分析。

**STEP 1** 在图 6-24 所示的对话框中，选中“Stepwise selection”单选框，然后单击对话框上方的“Criteria”选项卡，可看到图 6-25 所示的内容。

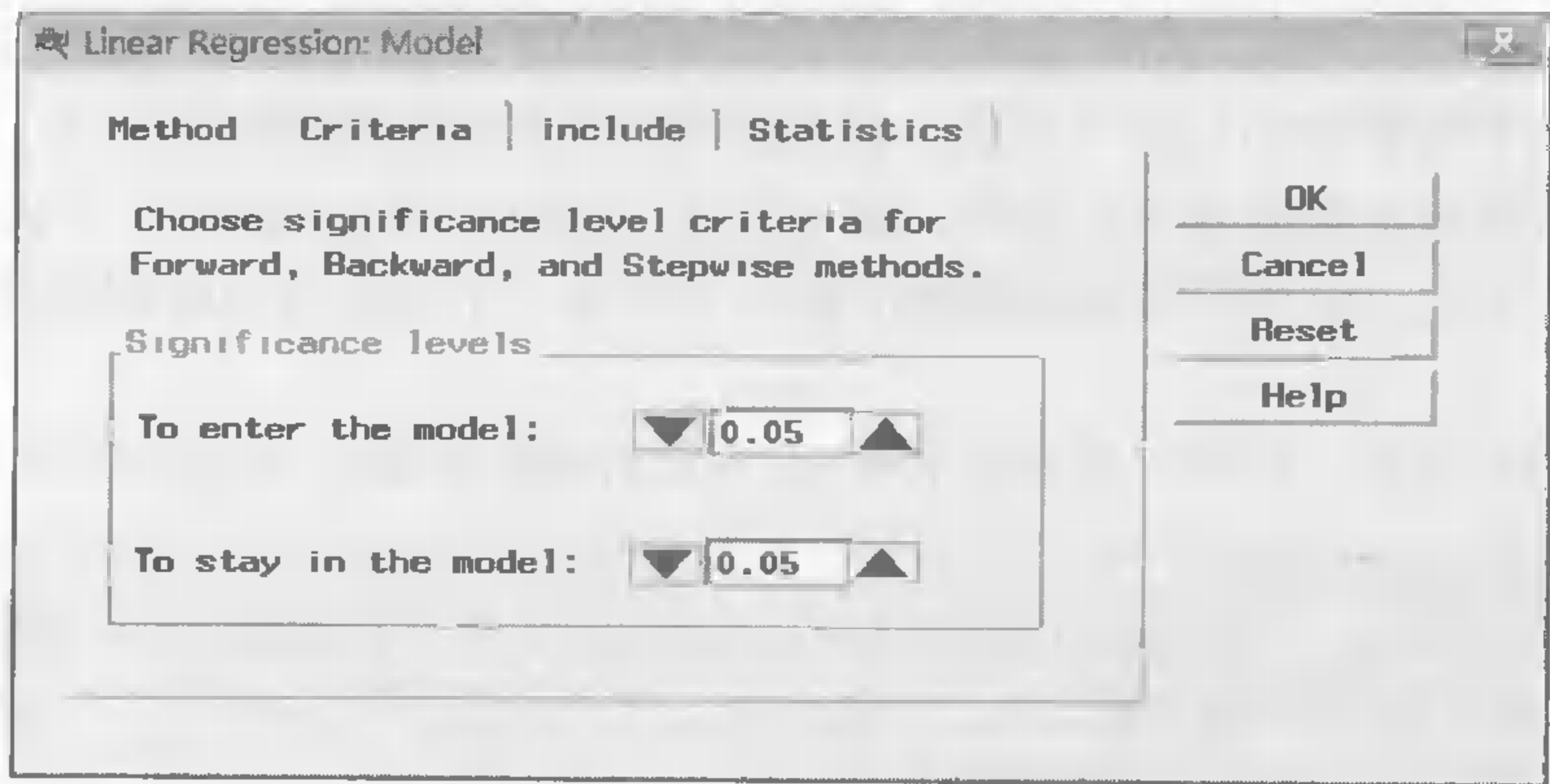


图 6-25 多元线性回归模型设置对话框的“Criteria”选项卡

**STEP 2** 在“Criteria”选项卡中，“Significance levels”栏下的“To enter the model”文本输入框可以指定逐步回归法向模型引入变量的显著性水平，“To stay in the model”文本输入框则可以指定从模型中剔除变量的显著性水平。本例中的两个显著性水平均选择 0.05，单击“OK”按钮返回图 6-24 所示选项卡，再单击“OK”按钮返回图 6-22 所示对话框。在该对话框中，单击“OK”按钮，可以得到图 6-26 所示的输出结果（节选）。

Stepwise Selection: Step 3					
Variable Experience Entered: R-Square = 0.7962 and C(p) = 4.1333					
Analysis of Variance					
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	3	17042273213	5680757738	575.65	<.0001
Error	442	4361851516	9868442		
Corrected Total	445	21404124729			
Variable	Parameter Estimate	Standard Error	Type III SS	F Value	Pr > F
Intercept	-2105.16091	823.48841	64491758	6.54	0.0109
Education	511.66088	70.83415	514905601	52.18	<.0001
Current_Salary	0.32115	0.01226	6771438756	686.17	<.0001
Experience	10.62684	1.48688	504088295	51.08	<.0001
Bounds on condition number: 1.9183, 14.452					
All variables left in the model are significant at the 0.0500 level.					
No other variable met the 0.0500 significance level for entry into the model.					

图 6-26 逐步回归法输出结果（节选）

本例一共经历了 3 次回归建模，最终得到了通过检验的模型。3 次回归的过程均详细地在结果窗口中显示。在 SAS 输出结果的第 3 步（即“Stepwise Selection: Step 3”）的输出结果中，可以看到系统自动根据用户设置的显著性水平进行变量筛选的结果，即最终模型当中含有“Education”、“Current\_Salary”和“Experience”3 个自变量，且均通过了  $F$  检验和回归系数显著性检验。而且系统最终还会提示，在模型中的所有自变量回归系数的显著性（即  $P$  值）都小于 0.05，且在  $\alpha = 0.05$  的条件下没有其他自变量可以被引入到回归模型当中。

至此，根据逐步回归法，可以得到以下的回归方程。

$$\text{Begin\_Salary} = -2105.16091 + 511.66 \times \text{Education} + 0.32 \times \text{Current\_Salary} + 10.63 \times \text{Experience}$$

从回归方程中可以得知，*Education* 变量对因变量的平均影响最大。当学历每增加一个单位时，起始薪酬平均增加 511.66 个单位。目前薪酬状况对起始薪酬影响不大，说明 HR 在给个人制定起始薪酬时主要考虑个人素质，即学历及工作经验等主观因素，个人目前所处的状况则对起薪影响不大。而年龄对起始薪酬的影响不显著，不能作为 HR 制定薪金的主要考虑因素。

对于多元回归分析，同样可以利用 REG 过程进行编程分析，具体程序如下。

```
proc reg data=Sasuser.Salary;
    model Begin_Salary =Education Current_Salary Experience Age /selection=stepwise sle=0.05 sls=0.05;
/*关键字“selection=”指定筛选变量的方法，“sle=”指定进入模型的显著性水平；“sls=”表示剔除或保留变量的显著性水平*/
run;
```

6.4 定性自变量回归分析

在影响因变量的诸多因素中，除了定量变量之外，有时还有一些定性因素，如例 6-5 中的性别、职位等对起始薪酬的影响。在进行回归分析的过程中，需要对定性因素对因变量的影响进行特殊处理，即应当把定性变量转化为虚拟变量（或哑变量）之后再引入回归模型中进行分析。

6.4.1 虚拟变量的设定

虚拟变量的设定是把对变量的定性描述转化成定量数据来进行描述，如性别定性变量有“男”和“女”两种表现，在设定虚拟变量时，可考虑用数字“0”、“1”分别代表“男”、“女”，则性别在 SAS 系统中便可转化为数值型变量以进行分析了。

设定虚拟变量时应当遵循以下原则。

- 对于有 *k* 个表现值的定性变量，只设定 *k*-1 个虚拟变量。
- 虚拟变量的值通常用“0”或“1”表示。
- 对于每个样本而言，同一个定性变量对应虚拟变量的值之和不超过 1。

如性别变量有两个表现值即“男”和“女”（即 *k*=2），因此只需设定 1 个虚拟变量即可。而例 6-5 中的职位变量有 3 个表现值，因此需要设定 2 个虚拟变量来进行分析，如表 6-6 所示。

表 6-6 虚拟变量的设定

	虚 拟 变 量		含 义
	Position1	Position2	
“职位”定性变量	0	0	普通员工
	0	1	主管
	1	0	经理

在表 6-6 所示的虚拟变量中，其具体数值表示何种含义，用户可以根据自身需求进行指定。本例按照表 6-6 所示的关系指定“职位”的虚拟变量（数据详见 Salary.sas7bdat），如果考虑例 6-5 中的所有自变量，则建立的回归方程如下所示。

$$\begin{aligned} Begin\_Salary = & \alpha + \beta_1 \times Gender + \beta_2 \times Education + \beta_3 \times Position1 + \beta_4 \times Position2 \\ & + \beta_5 \times Current\_Salary + \beta_6 \times Experience + \beta_7 \times Age \end{aligned}$$

当虚拟变量 *Gender* 为 0 时，回归方程中不含 *Gender* 变量，表示女性职员起始薪酬的影响状况；而当 *Gender* 为 1 时，回归方程中含有 *Gender* 变量，表示男性职员起始薪酬的影响状况。同理，当 *Position1* 和 *Position2* 同时为 0 时，表示普通员工起始薪酬的影响状况。

6.4.2 含有虚拟变量的回归分析

在图 6-22 所示的多元线性回归对话框中，把定量变量和虚拟变量都作为模型的解释变量放入回归模型中。在“Model”按钮弹出的对话框中先选择“Full Model”，考虑全变量模型，单击“OK”按钮后可得到图 6-27 所示的输出结果。

The REG Procedure						
Model: MODEL1						
Dependent Variable: Begin_Salary 起始薪水						
Number of Observations Read			471			
Number of Observations Used			446			
Number of Observations with Missing Values			25			
Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	7	17801017015	2543002431	309.13	<.0001	
Error	438	3603107714	8226273			
Corrected Total	445	21404124729				
Root MSE		2868.14806	R-Square	0.8317		
Dependent Mean		16998	Adj R-Sq	0.8290		
Coeff Var		16.87381				
Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	988.87206	1307.81654	0.76	0.4500
Gender	性别	1	1498.15081	340.23921	4.40	<.0001
Education	教育年限	1	345.26658	69.14280	4.99	<.0001
Current_Salary	目前薪水	1	0.21925	0.01637	13.39	<.0001
Experience	工作经历(周)	1	7.64285	2.53666	3.01	0.0027
Age	年龄	1	25.27001	21.26728	1.19	0.2354
Position1		1	5263.70101	618.28746	8.51	<.0001
Position2		1	-1626.50144	699.01135	-2.33	0.0204

图 6-27 含有虚拟变量的回归分析结果

在图 6-27 所示结果的“Parameter Estimates”表格中，“Age”变量回归系数不显著，即 P 值（“Pr>|t|”）为 0.235 4，远远大于 0.05，因此考虑从模型中剔除该变量之后，重新进行回归分析，得到图 6-28 所示的输出结果。

Analysis of Variance						
Source	DF	Sum of Squares	Mean Square	F Value	Pr > F	
Model	6	17790689354	2965114892	360.00	<.0001	
Error	440	3623989592	8236340			
Corrected Total	446	21414678946				
Root MSE		2869.90243	R-Square	0.8308		
Dependent Mean		17005	Adj R-Sq	0.8285		
Coeff Var		16.87690				
Parameter Estimates						
Variable	Label	DF	Parameter Estimate	Standard Error	t Value	Pr >  t
Intercept	Intercept	1	2107.17222	874.44976	2.41	0.0164
Gender	性别	1	1398.80717	326.77444	4.28	<.0001
Education	教育年限	1	343.77291	68.98947	4.98	<.0001
Current Salary	目前薪水	1	0.21920	0.01638	13.38	<.0001
Experience	工作经历(周)	1	10.03251	1.52650	6.57	<.0001
Position1		1	5286.31628	616.73498	8.57	<.0001
Position2		1	-1702.05201	696.74617	-2.44	0.0150

图 6-28 剔除不显著变量之后含有虚拟变量的回归分析结果

在剔除不显著的“Age”变量之后，可以看到所有变量的回归系数均在 $\alpha = 0.05$ 条件下显著，且回归方程的判定系数 $R^2 = 0.8308$ ，其总体显著性检验的 $F$ 统计量为360，非常显著。因此，可以对回归方程进行分析。根据图6-28的参数估计结果，回归方程如下。

$$\begin{aligned} \text{Begin\_Salary} = & 2107.17 + 1398.81 \times \text{Gender} + 343.77 \times \text{Education} + 5286.32 \times \text{Position1} \\ & - 1702.05 \times \text{Position2} + 0.22 \times \text{Current\_Salary} + 10.03 \times \text{Experience} \end{aligned}$$

对含有虚拟变量的回归方程进行分析，应当先确定分析的参照方程。参照方程是指当所有虚拟变量为0时的方程，本例的参照方程如下。

$$\text{Begin\_Salary} = 2107.17 + 343.77 \times \text{Education} + 0.22 \times \text{Current\_Salary} + 10.03 \times \text{Experience}$$

因本例中有两个虚拟变量，故所有虚拟变量均为0时的参照方程表示女性（ $\text{Gender} = 0$ ）、职位为普通员工（ $\text{Position1} = \text{Position2} = 0$ ）的起始薪酬影响关系。即女性普通员工的学历每增加1年，起始薪酬平均增加343.77元；工作经验增加1周，则起薪平均增加10.03元；而目前薪酬增加1元，故起薪平均增加0.22元。

对于不同职位的男女员工的起薪影响，可以根据对应虚拟变量的取值来进行分析。如要分析职位为经理的男性起薪状况，即把虚拟变量 $\text{Gender} = 1$ 、 $\text{Position1} = 1$ 、 $\text{Position} = 0$ 代入方程中，如下所示。

$$\begin{aligned} \text{Begin\_Salary} = & 2107.17 + 1398.81 + 343.77 \times \text{Education} + 5286.32 \\ & + 0.22 \times \text{Current\_Salary} + 10.03 \times \text{Experience} \\ = & 2107.17 + 343.77 \times \text{Education} + 0.22 \times \text{Current\_Salary} \\ & + 10.03 \times \text{Experience} + 6685.13 \end{aligned}$$

从该方程可知，职位为经理的男性求职人员的起薪比女性普通员工的起薪平均高出6685.13元。

在利用SAS程序对含有虚拟变量的实际问题进行回归分析时，可调用REG过程，其语法与定量变量回归分析一样，只需把虚拟变量当作解释变量加入模型即可。本例的程序如下。

```
proc reg data=Sasuser.Salary;
    model Begin_Salary = Gender Education Current_Salary Experience Position1 Position2;
run;
```

运行程序之后可以得到图6-28所示的结果。

## 6.5 本章小结

本章主要介绍了相关分析和回归分析的基本原理及其在SAS系统中的实现，主要内容简要回顾如下：简单相关分析、非参数相关分析主要分析两个变量之间的线性相关关系，相关系数为0，表示不存在线性关系，并不表示变量之间没有关系；典型相关分析主要考察两组之间的相互依存关系；线性回归分析主要考察因变量与自变量之间的线性关系，其主要内容包括回归方程的拟合、利用OLS方法对回归模型进行参数估计、对回归方程进行拟合优度检验、方程总体显著性 $F$ 检验及回归系数显著性的 $t$ 检验；自变量中含有定性变量的回归分析，要在模型中引入虚拟变量。

## 第7章

# 因子分析

客观世界是复杂多变的，在社会发展的过程中体现多样性，人们的生活因此而丰富多彩。那么，人们如何简练地从若干个方面去归纳、概括事物发展的历程和特征呢？如何抓住主要矛盾以及抓住矛盾的主要方面？即如何对事物发展过程中呈现出的纷繁芜杂的数据进行简单明了的描述呢？这需要对数据进行精简和概括。

人们往往希望能够找出少数具有代表性的变量来对复杂事物进行描述，这需把反映该事物的很多变量或数据进行高度概括。本章将阐述的因子分析便是如何利用复杂多样的数据来综合描述客观事物特征的分析方法和过程。

### 7.1 数据降维

每个人都会遇到有很多变量的数据，如反映全国或各省市经济、社会发展状况的变量数据、反映一个国家总体发展状况的数据等。这些数据的共同特点是变量很多，在如此多的变量之中，有些变量之间是相关的。同时分析很多个变量是比较困难的，因此人们希望能够找出这些变量的“代表”，以对更多变量进行描述。

如在学校中进行奖学金的评定，需要考虑学生各门课程的学习成绩、与人相处的能力、尊重师长的程度、乐于助人的程度、担任学生干部的努力程度、参加社会实践活动的积极性等因素。假设有一个学生本学期考试的科目有 10 门课程，那么按照上述的参考变量，其参加奖学金评定便会有 15 个变量之多。在实际工作中，不会同时考虑这 15 个变量的数据来进行奖学金评定，通常的做法是把相互关联的变量进行综合，如把上述 15 个变量综合为学习成绩（含 10 门课程）变量、思想品德（含与人相处的能力、尊重师长的程度、乐于助人的程度）变量、工作态度（含担任学生干部的努力程度、参加实践的的积极性）变量等 3 个具有代表性的综合变量，然后依照这 3 个综合变量进行奖学金的评定，从而达到了化繁为简的目的。

#### 7.1.1 数据降维的基本问题

把反映一个事物特征的多个变量用较少的、具有代表性的变量来描述，这个过程被称为数据降维过程。在一般情况下，不同变量是从不同侧面或方面去描述事物特征的，这些不同的方面被称为事物的维度，如从身高、体重、血型 3 个方面反映一个人的特征，则具有 3 个维度。当反映事物的方面太多时，过多的数据会对所描述的对象造成混乱，往往得不到正确的结论。因此，应当把相关的维度进行总结概括，尽量降低数据或变量的维度，简要地对事物特征进行描述。

为了能够简要而不遗漏地反映事物特征，数据降维过程中应当解决以下几个基本问题。

- 能否把多个数据的变量用较少的综合变量来表示。

- 较少的综合变量包含多少原来的信息。
- 能否利用找到的综合变量对事物进行分析。

上述第 1 个问题具体是指，在进行数据降维之前，应当考虑原始变量数据之间的关联性，即变量之间是否具有可提取综合变量的必然联系；而第 2 个问题主要考虑所提取出来的综合变量在多大程度上代表了原始数据的信息，这是利用综合变量进行统计分析，进而得到正确结论的理论基础；第 3 个问题主要阐述了数据降维在统计分析过程中发挥的重要作用，并且在降维得到综合变量的基础之上进行进一步的统计分析活动。

解决好这些基本问题之后，用简化的数据对事物进行描述或判定时，可以在一定的概率保证程度下得到正确的推断结论。

### 7.1.2 数据降维的基本原理

数据降维过程可以从最简单的二维数据降为一维数据开始。先假设只有二维，即只有两个变量，可用二维坐标轴上的横坐标和纵坐标来表示。因此，每个观测值都有相应于这两个坐标轴的两个坐标值，在正态分布的假定下，这些数据在二维坐标轴上形成一个椭圆的分布形状，如图 7-1 所示。

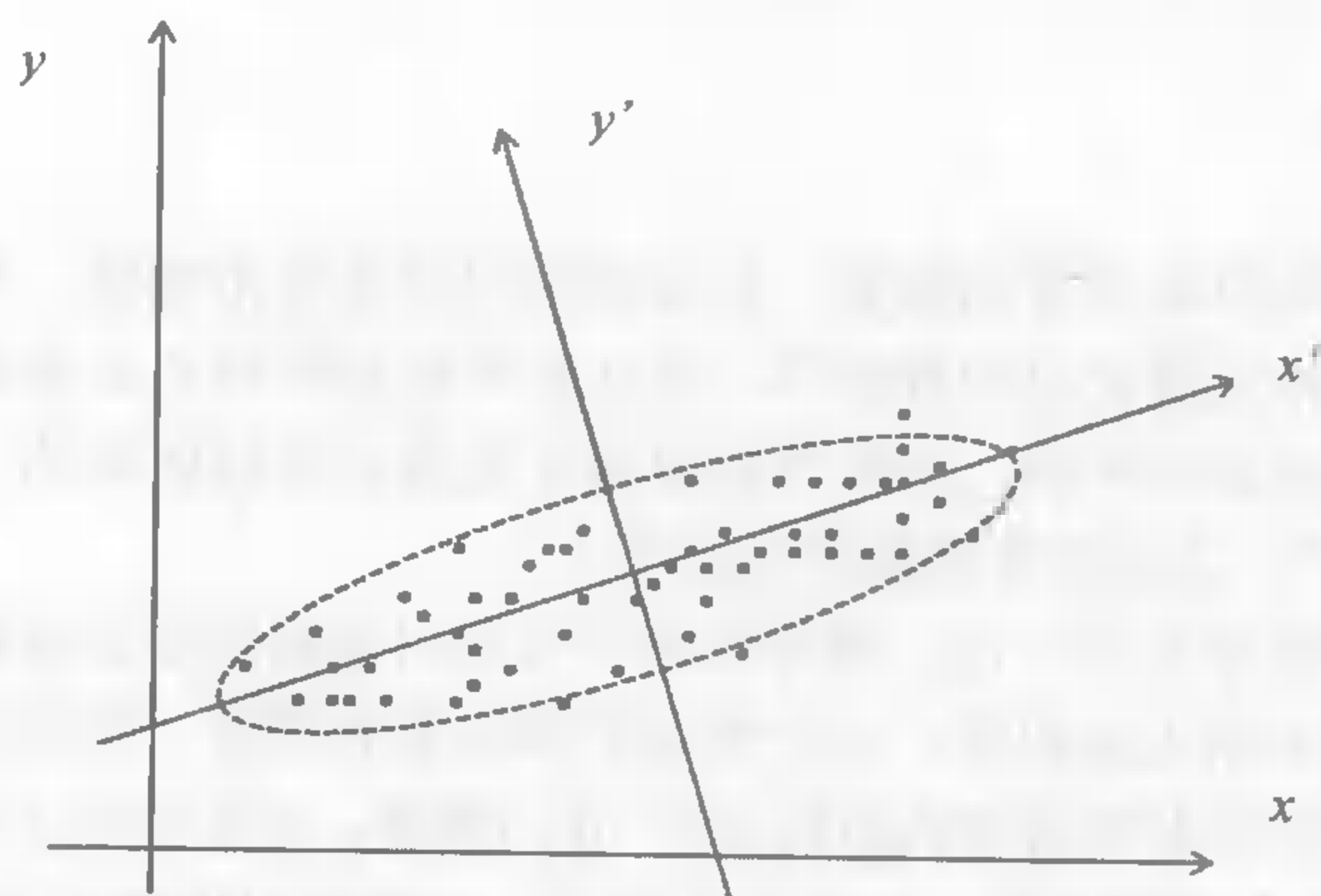


图 7-1 二维正态假定下的分布及坐标变换

众所周知，椭圆有一个长轴和一个短轴，且互相垂直。在短轴方向上，数据变化很少；而在长轴方向上，数据变化的范围较大。在极端的情况下，如果短轴退化成一点，则只有在长轴方向上才能够解释这些点的变化了。因此，长轴就是要找的综合变量。至此，由二维到一维的降维就完成了。

当坐标轴和椭圆的长短轴平行，那么代表长轴的变量就描述了数据的主要变化，而代表短轴的变量则描述了数据的次要变化。但是，坐标轴通常并不和椭圆的长短轴平行，因此需要寻找椭圆的长短轴，即进行坐标平移或旋转变换，使得新变量和椭圆的长短轴平行。如果长轴变量代表了数据包含的大部分信息，则用变量在该轴上的变化代替原先的两个变量（舍去次要的一维），降维就完成了。椭圆的长短轴相差越大，降维效果就越好。

多维变量的情况和二维类似，也有高维的椭球。首先把高维椭球的主轴找出来，再用代表大多数数据信息的、最长的几个轴作为新变量。与二维情况类似，高维椭球的主轴也是互相垂直的。这些互相正交的新变量是原先变量的线性组合，被称为主成分。正如二维椭圆有两个主轴、三维椭球有 3 个主轴一样，有几个变量，就有几个主成分。

究竟要选择多少个主成分，是不是越少越好呢？实际上它有一定的选择标准，即这些被选中的主成分所代表的主轴的长度之和与主轴长度总和的比值，这个比值也被称为“阈值”。有些文献建议，所选主轴的长度之和大约占有所有主轴长度之和的 85%（即阈值为 85%）即可。但在实际应用过程中，要依据研究目的、研究对象和所搜集变量的具体情况而定。

主轴越长，表示变量在该主轴方向上的变动程度越大，亦即方差越大。所以在一般情况下，无需计算主轴的长度，而是计算其主轴方向的方差，根据所选取主轴的方差与所有主轴方向上方差之和的比值，即方差贡献率的大小来判断应该取多少个主成分。

## 7.2 主成分分析

### 7.2.1 主成分分析的基本概念与原理

从 7.1 节介绍的数据降维过程中，已经得知主成分提取的几何意义。主成分是由原始变量提取的综合变量，可以用以下式子表示。

$$\begin{aligned} Y_1 &= \mu_{11}x_1 + \mu_{12}x_2 + L \mu_{1p}x_p \\ Y_2 &= \mu_{21}x_1 + \mu_{22}x_2 + L \mu_{2p}x_p \\ &L \\ Y_p &= \mu_{p1}x_1 + \mu_{p2}x_2 + L \mu_{pp}x_p \end{aligned}$$

其中， $Y$  表示主成分， $x$  为原始变量； $\mu_{ij}$  为系数，有约束条件： $\mu_{k1}^2 + \mu_{k2}^2 + L \mu_{kp}^2 = 1$ ， $\mu_{ij}$  可由原始数据协方差矩阵或相关系数矩阵确定。

在提取出来的各个主成分中， $Y_i$  与  $Y_j$  相互无关。

●  $Y_1$ （第一个主成分）约束条件：是  $x_1, x_2, K, x_p$  的一切线性组合中最大的。

●  $Y_2$ （第二个主成分）是  $x_1, x_2, K, x_p$  的一切线性组合中第二大的。

●  $Y_n$ （第  $n$  个主成分）是  $x_1, x_2, K, x_p$  的一切线性组合中第  $n$  大的。

由原始数据的协方差阵或相关系数矩阵，可计算出矩阵的特征值或特征根。

$$\lambda_1 \geq \lambda_2 \geq L \geq \lambda_p$$

其中， $\lambda_1$  对应  $Y_1$  的方差， $\lambda_2$  对应  $Y_2$  的方差…… $\lambda_p$  对应  $Y_p$  的方差，因此有：

$$\text{阈值} = \frac{\text{被选择的主成分长度}}{\text{主轴成分总和}} = \frac{\text{选择的特征根的和}}{\text{特征根总和}} = (\text{累积}) \text{ 方差贡献率}$$

$\lambda$  对应的特征向量  $\mu$  就是主成分分析线性模型中对应的系数，如： $\lambda_1$  对应的特征向量为  $\mu_{11}, \mu_{11}, K, \mu_{1p}$  为第 1 主成分的线性组合系数，即： $Y_1 = \mu_{11}x_1 + \mu_{12}x_2 + L + \mu_{1p}x_p$ 。

这些系数被称为“主成分载荷”，表示主成分和相应的原先变量的相关系数。相关系数绝对值越大，主成分对该变量的代表性也越大。根据上式计算出来的  $Y$  值被称为主成分得分。

在实际问题中，不同的变量往往有不同的量纲。为了实现不同量纲数据之间的可比性，以保证所提取的主成分与原始变量意义上的一致性，在进行主成分分析之前可按照以下公式将变量标准化。

$$x_i^* = \frac{x_i - E(x_i)}{\sqrt{Var(x_i)}} \quad (i = 1, 2, L, p)$$

其中， $E(x_i)$  表示变量的期望， $Var(x_i)$  表示变量的方差。

7.2.2 主成分分析的基本步骤和过程

在主成分分析的过程中，通常要先把各变量进行无量纲化（即标准化）。把变量进行标准化之后，可按照以下顺序进行主成分分析。

- 选择协方差阵或相关阵计算特征根及对应的特征向量。
- 计算方差贡献率，并根据方差贡献率的阈值选取合适的主成分个数。
- 根据主成分载荷的大小对选取的主成分进行命名。
- 根据主成分载荷计算各个主成分得分。

在 SAS 系统中，在调用主成分分析的过程中，系统会自动对所分析的变量进行标准化，因此用户无需自行对原始变量进行标准化处理。



例 7-1

为评价全国各省/直辖市/自治区的综合发展水平，现收集了全国 24 个地区的人均 GDP、人均可支配收入、人均消费支出等数据（详见 Live.sas7bdat）进行综合考察，如表 7-1 所示。试利用主成分分析方法对各地区综合发展状况进行评价。

表 7-1 各地区社会经济发展状况

地区	人均国内生产总值（亿元） (GDP)	人均可支配收入（元） (Income)	人均消费性支出（元） (Consumption)	城镇就业率（%） (Employment)	人均教育经费（元） (Education)	医疗保健费用（元） (Health)	预期寿命（岁） (Life)
北 京	45 444	17 652.95	13 244.2	0.3937	584.43	1 295.76	76.1
山 西	12 495	8 913.91	6 342.63	0.2554	548.83	538.7	71.65
内 蒙 古	16 331	9 136.79	6 928.6	0.2158	504.77	533.36	69.87
吉 林	13 348	8 690.62	6 794.71	0.1836	502.08	675.77	73.1
黑 龙 江	14 434	8 272.51	6 178.01	0.2418	479.85	613.15	72.37
上 海	51 474	18 645.03	13 773.41	0.2103	1 136.15	796.82	78.14
江 苏	24 560	12 318.57	8 621.82	0.168	656.29	579.32	73.91
浙 江	27 703	16 293.77	12 253.74	0.1936	972.69	831.79	74.7
福 建	18 646	12 321.31	8 794.41	0.2394	531.4	478.41	72.55
山 东	20 096	10 744.79	7 457.31	0.2142	546.64	579.01	73.92
河 南	11 346	8 667.97	6 038.02	0.2439	421.72	472.31	71.54
湖 北	11 431	8 785.94	6 736.56	0.2076	517.28	499.34	71.08
湖 南	10 426	9 523.97	7 504.99	0.1737	582.16	601.34	70.66
广 西	8 788	9 286.7	7 032.8	0.1811	528.13	466.04	71.29
海 南	10 871	8 123.94	5 928.79	0.2	347.11	351.06	72.92
重 庆	10 982	10 243.46	8 623.29	0.1705	772.52	629.32	71.73
四 川	9 060	8 385.96	6 891.27	0.1892	449.68	442.83	71.2
云 南	7 835	9 265.9	6 996.9	0.1885	337.42	663.01	65.49

续表

地区	人均国内生产总值 (亿元) (GDP)	人均可支配收入 (元) (Income)	人均消费性支出 (元) (Consumptiou)	城镇就业率 (%) (Employment)	人均教育经费 (元) (Education)	医疗保健费用 (元) (Health)	预期寿命 (岁) (Life)
西 藏	9 114	9 431.18	8 617.11	0.246	428.09	338.57	64.37
陕 西	9 899	8 272.02	6 656.46	0.2414	701.82	605.31	70.07
甘 肃	7 477	8 086.82	6 529.2	0.2496	505.9	492.23	67.47
青 海	10 045	8 057.85	6 245.26	0.2002	360.52	554.11	66.03
宁 夏	10 239	8 093.64	6 404.31	0.2367	388.3	535.92	70.17
新 疆	13 108	7 990.15	6 207.52	0.328	456.25	499.16	67.41

在本例中，共有 7 个变量可用来综合评价各地区的发展状况。但是，如果从 7 个方面来考察综合发展状况，不免显得过于复杂。因此，可以把 7 个变量进行降维，从中提取出若干个综合变量，利用综合变量所反应的主成分来对地区发展状况进行评价。

**STEP 1** 在 SAS 系统中，可以利用 SAS/Analyst 进行主成分分析。进入 SAS/Analyst，选择“Statistics→Multivariate→Principle Components”，弹出图 7-2 所示的主成分分析对话框。

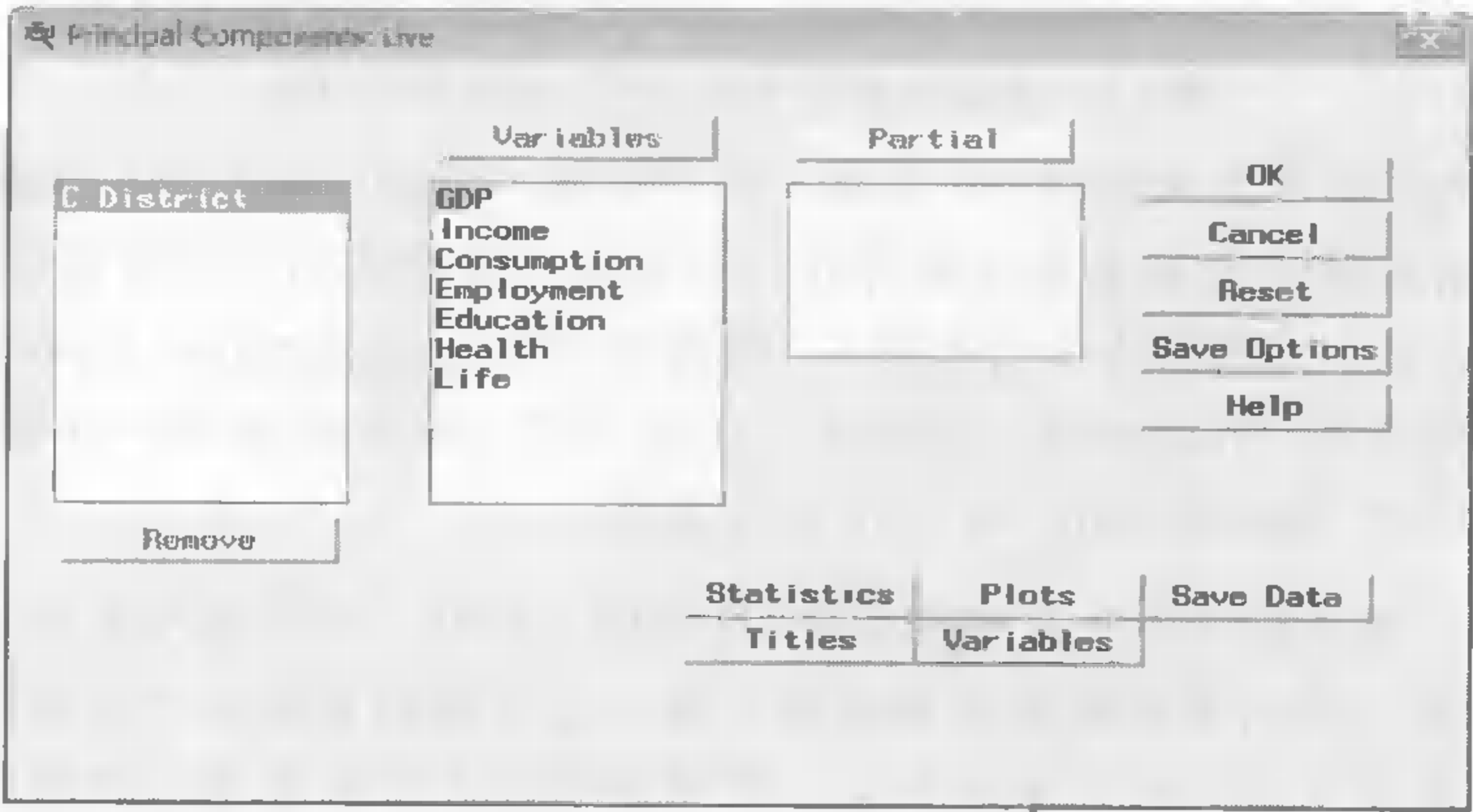


图 7-2 主成分分析对话框

**STEP 2** 在主成分分析对话框中，选中所有要进行分析的变量，单击“Variables”按钮把这些变量指定为将要进行分析的原始变量。

**STEP 3** 单击“Statistics”按钮，弹出图 7-3 所示的统计量设置对话框。在该对话框中，可以对主成分分析所用对象进行设定。

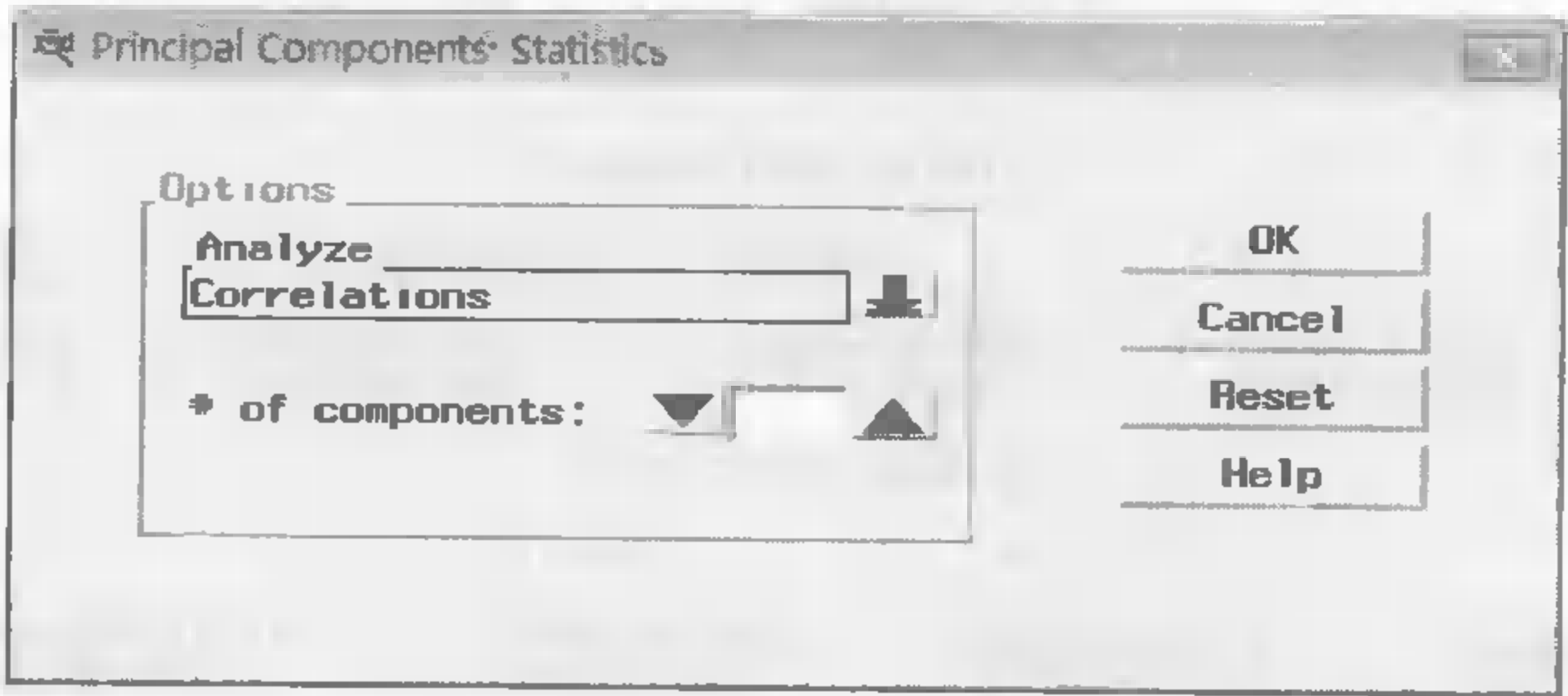





图 7-3 主成分的“Statistics”对话框

**STEP 4** 在图 7-3 所示对话框中的“Options”分栏下，单击“Analyze”下拉选单的  按钮，可选择相关系数矩阵（“Correlations”）、协方差矩阵（“Covariances”）或非修正的相关系数矩阵（“Uncorrected Correlations”）或非修正的协方差矩阵（“Uncorrected Covariances”）进行特征根的计算。

**STEP 5** 如果事先已知要提取主成分的数目，则可在图 7-3 所示对话框中的“# of components”文本输入框中输入所要提取的主成分数目，或者单击 、 按钮进行设定。本例按系统默认的相关系数矩阵（用户也可自行选择使用协方差矩阵）进行特征根计算，不指定所提取主成分的个数。

**STEP 6** 在图 7-2 所示对话框中单击“Save Data”按钮，弹出主成分得分计算对话框，如图 7-4 所示。

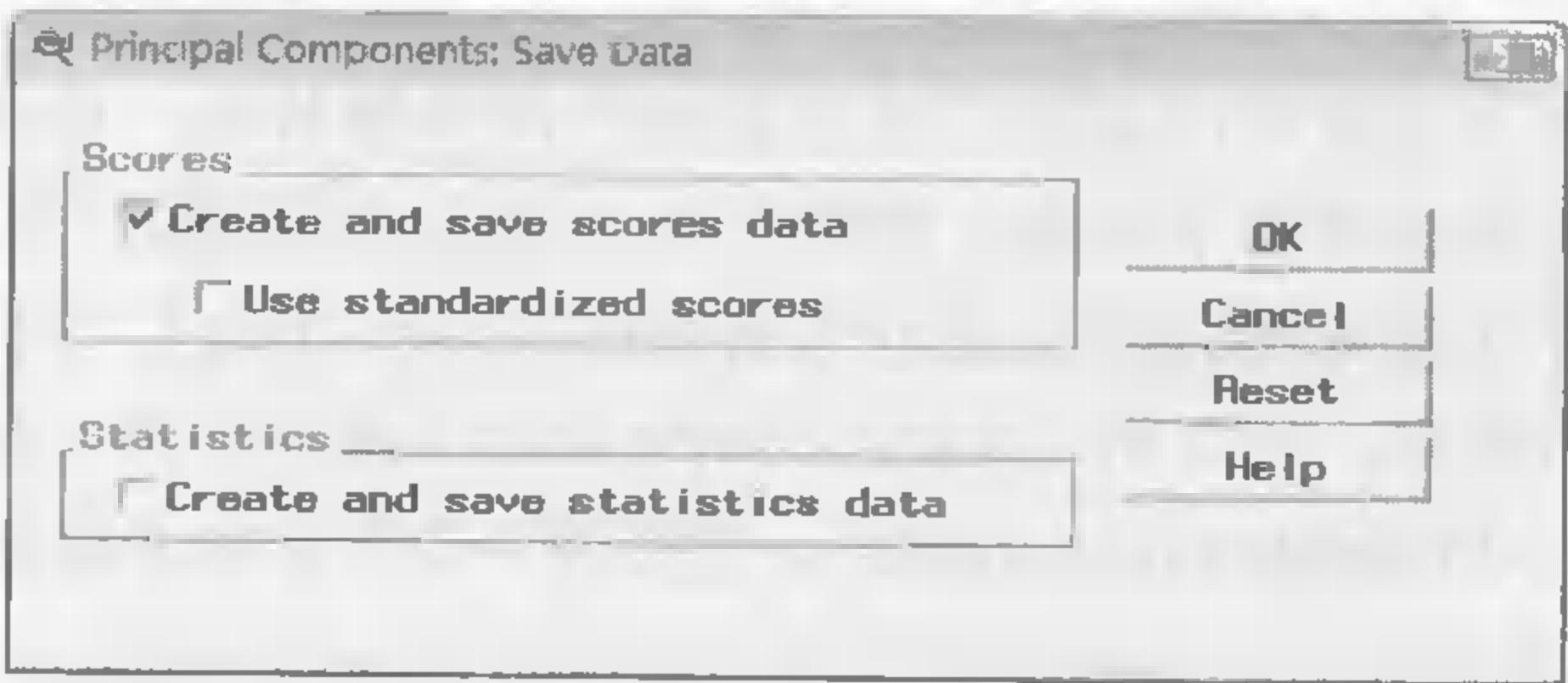


图 7-4 主成分得分及相关统计量计算对话框

**STEP 7** 本例需计算主成分得分。在图 7-4 中选中“Create and save scores data”复选框，即可指定系统自动计算所提取主成分的得分。如需计算标准化的主成分得分（即把主成分得分标准化为均值为 0、标准差为 1 的数列），则选中“Use standardized scores”复选框即可。此外，在该对话框中的“Statistics”分栏下，还可指定系统在数据集中存储一些相应的统计量数据。单击“OK”按钮返回图 7-2 所示的对话框。



单击图 7-2 所示对话框中的“Plot”按钮，可以指定系统绘制特征根的碎石图，用以考察提取主成分的个数，也可绘制主成分之间的散点图，用于考察每个主成分所代表的含义。因用图形的方法来考察上述两个方面的内容略显主观，故本书应用 SAS 系统输出的结果来进行定量分析，所以对上述两种图形不再赘述。

**STEP 8** 返回图 7-2 所示的对话框，单击“OK”按钮，可得到主成分分析的结果。在 SAS 系统中，主成分分析的结果首先展示了原始变量的均值和标准差统计量，如图 7-5 所示。

The PRINCOMP Procedure				
Observations		24		
Variables		7		
Simple Statistics				
	GDP	Income	Consumption	Employment
Mean	16048.00000	10216.90625	7783.388333	0.2238416667
Std	11243.20495	3084.68079	2231.289949	0.0512078452
Simple Statistics				
	Education	Health	Life	
Mean	552.5012500	586.3600000	71.15583333	
Std	188.8018045	190.4346551	3.28473269	

图 7-5 主成分分析结果中原始变量的统计量

利用图 7-5 的输出结果，可以对原始变量进行无量纲化或标准化（SAS 系统会自动进行标准化，故该步骤可以省略）。

**STEP 9** 在主成分的分析结果中，还会输出用于计算特征根的矩阵（协方差矩阵或相关系数矩阵）。本例输出原始变量之间的相关系数矩阵，如图 7-6 所示。

Correlation Matrix				
		GDP	Income	Consumption
GDP	人均国内生产总值	1.0000	0.9382	0.8930
Income	人均可支配收入	0.9382	1.0000	0.9775
Consumption	人均消费性支出	0.8930	0.9775	1.0000
Employment	城镇就业率	0.3345	0.2111	0.2141
Education	人均教育经费	0.6867	0.7557	0.7628
Health	医疗保健费用	0.7498	0.7432	0.7331
Life	预期寿命	0.7421	0.6905	0.5957

Correlation Matrix				
	Employment	Education	Health	Life
GDP	0.3345	0.6867	0.7498	0.7421
Income	0.2111	0.7557	0.7432	0.6905
Consumption	0.2141	0.7628	0.7331	0.5957
Employment	1.0000	-0.1231	0.4341	0.0147
Education	-0.1231	1.0000	0.4653	0.6387
Health	0.4341	0.4653	1.0000	0.5259
Life	0.0147	0.6387	0.5259	1.0000

图 7-6 主成分分析中的相关系数矩阵

**STEP 10** 相关系数矩阵可以列示出变量两两之间的简单相关系数，用于观察变量之间的联系，可在一定程度上考察数据降维是否恰当。此外，在手工计算过程中，可用图 7-6 所示的相关系数矩阵计算特征根及其对应的特征向量。SAS 系统会自动按照给定的矩阵数据进行特征根计算，结算结果如图 7-7 所示。

Eigenvalues of the Correlation Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	4.72549923	3.49115815	0.6751	0.6751
2	1.23434108	0.78567923	0.1763	0.8514
3	0.44866185	0.14254777	0.0641	0.9155
4	0.30611408	0.09235871	0.0437	0.9592
5	0.21375537	0.15318106	0.0305	0.9898
6	0.06057431	0.04952023	0.0087	0.9984
7	0.01105408		0.0016	1.0000

图 7-7 主成分分析中的特征根（依据相关系数矩阵计算）

**STEP 11** 在图 7-7 所示的结果中，一共可输出 4 列结果，即特征根（“Eigenvalue”）、特征根之间的差分（“Difference”）、某个特征根占有所有特征根的比例（“Proportion”）和特征根累计比例（“Cumulative”）。依据 5.2.1 小节的理论分析可知，有多少个原始变量，就可以提取出多少个主成分。因此，图 7-7 所示的结果中可以输出 7 个特征根。

可以依据特征根的贡献率来决定应当取多少个主成分比较合适。在通常情况下，可参照特征根累计贡献率阈值 85% 的标准来进行判定。

**STEP 12** 在图 7-7 所示的结果中，第 1 个特征根的值占有所有特征根值之和的比例（也称为贡献率）为： $4.725 / (4.725 + 1.234 + 0.449 + 0.306 + 0.214 + 0.061 + 0.011) = 0.6751$ ，第 2 个特征根的贡献率为 0.1763。由图 7-7 所示结果的最后一列可知，上述两个特征根的累积贡献率已经达到 0.8514，说明提取两个主成分即可代表原始数据的大部分信息，而且前两个特征根的贡献远远大于其余特征根的贡献。因此，根据累积的特征根贡献率，本例可考虑提取两个主成分进行分析。

在主成分分析中，除考查特征根外，还应当考察如何提取主成分并计算主成分得分的问题，即计算特征根对应的特征向量，以作为原始变量线性组合的系数。在 SAS 系统中，同样

可输出特征根对应的特征向量结果，如图 7-8 所示。

		Eigenvectors						
		Prin1	Prin2	Prin3	Prin4	Prin5	Prin6	Prin7
GDP	人均国内生产总值	0.441618	0.073883	0.083499	0.153700	-.325047	-.796130	0.171583
Income	人均可支配收入	0.447192	-.029164	-.193227	0.036980	-.358861	0.213764	-.765499
Consumption	人均消费性支出	0.435590	-.016302	-.394963	0.035468	-.276029	0.450039	0.611568
Employment	城镇就业率	0.122961	0.827743	0.098366	0.463616	0.231263	0.143879	-.030374
Education	人均教育经费	0.365034	-.397744	-.255464	0.336966	0.718524	-.101607	-.056051
Health	医疗保健费用	0.374018	0.307351	-.001732	-.801352	0.345642	-.062374	-.010942
Life	预期寿命	0.356365	-.235799	0.851325	0.055691	0.011964	0.288153	0.079822

图 7-8 主成分分析中的特征向量

**STEP 13** 根据图 7-8 所示的各主成分对应的特征向量（即系数），可以计算出各主成分的得分。如第一个主成分得分为：

$$Y_1 = 0.442GDP + 0.447Income + 0.436Consumption + 0.123Employment + 0.365Education + 0.374Health + 0.356Life$$

第二个主成分得分为：

$$Y_2 = 0.074GDP - 0.029Income - 0.016Consumption + 0.828Employment - 0.398Education + 0.307Health - 0.236Life$$

**STEP 14** 可以根据主成分计算公式中的系数，即主成分载荷绝对值的大小来判定该主成分主要代表的原始变量的含义。如在第 1 主成分中，除“Employment”变量的系数之外，其余变量的系数均比较大，说明这些变量在第 1 主成分中发挥的影响作用比较大。因此，对于第 1 主成分  $Y_1$ ，归纳除“Employment”变量之外其余变量共同表示的含义，可把第 1 主成分命名为“经济生活成分”。

**STEP 15** 同理，对于第 2 主成分  $Y_2$ ，“Employment”变量的系数显著大于其余变量，说明该变量在第 2 主成分中发挥的影响作用比较大。因此，根据第 2 主成分中发挥作用最大的变量所代表的含义，可以把第 2 主成分命名为“就业成分”。

**STEP 16** 把原始变量进行标准化之后，代入上述公式，便可计算出每个样本在对应主成分中的得分。如果在图 7-4 所示的对话框中选中“Create and save scores data”复选框，则系统会自动依据上述公式计算出所有样本在所有主成分上的得分。

**STEP 17** 在 SAS 系统中，主成分分析过程不会自动把主成分得分存储在所用的数据集中，而是在 SAS/Analyst 的结果索引中用“Scores Table”临时数据集的方式存储原始数据和所计算出来的主成分得分，如图 7-9 所示。

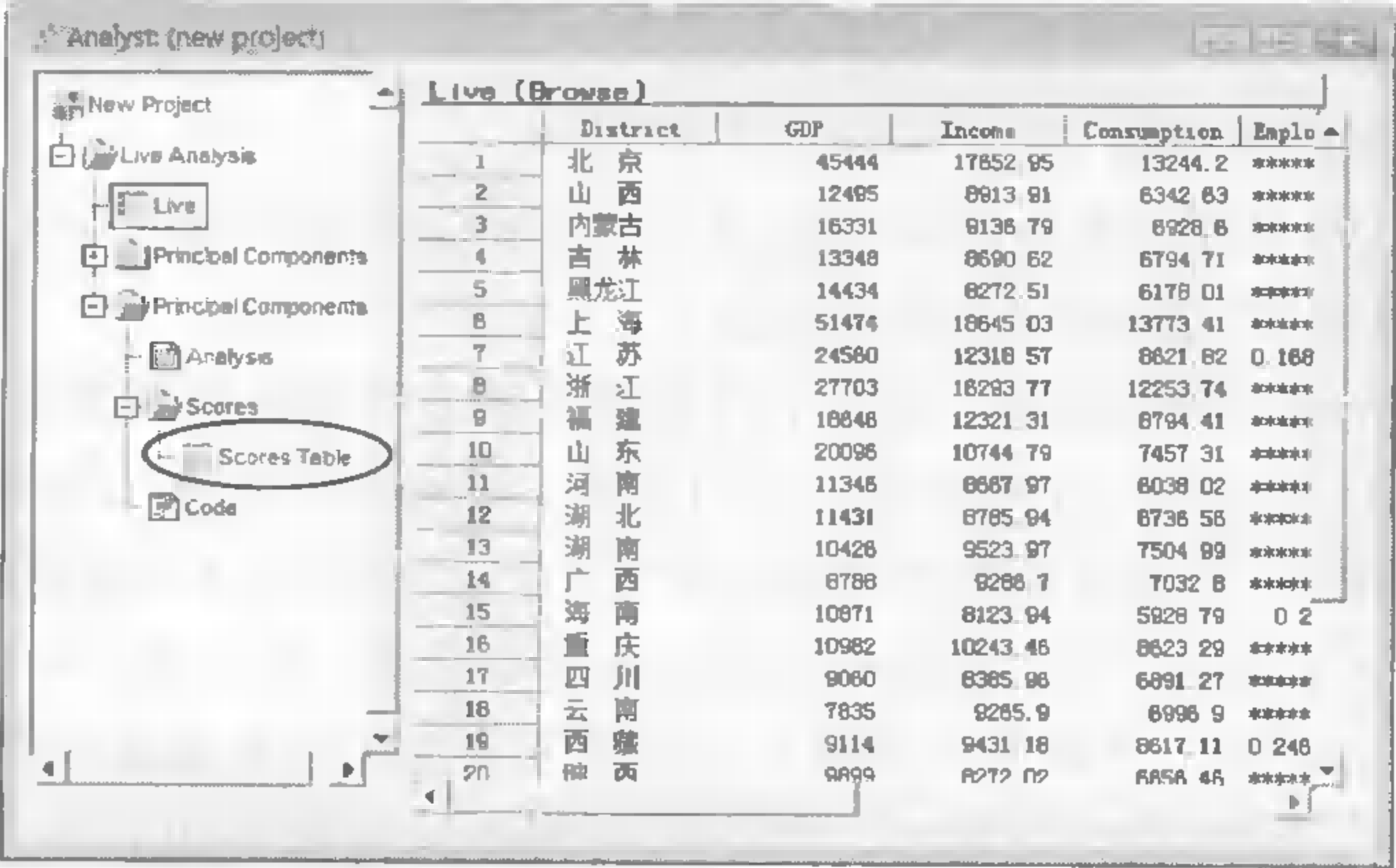


图 7-9 主成分分析的主成分得分结果索引

**STEP 18** 在图 7-9 所示窗口中，双击“Scores Table”结果索引，便会弹出包含原始数据和主成分得分的新窗口。利用该窗口的“File”系统菜单，可以保存该结果（如不保存，所计算出的结果在退出 SAS/Analyst 时会被系统自动清除）。

**STEP 19** 最后，可以把所有样本按照所提取出来的主成分得分进行排序，以考察各个样本的综合情况。此外，也可以利用特征根的贡献率及其对应的主成分得分计算出综合得分（在 SAS 系统中，要手工计算），以对所有样本进行综合评价。一般可采用的综合得分计算公式如下。

$$\text{综合得分} = \sum_{i=1}^n (\text{第}i\text{主成分得分} \times \text{对应特征根贡献率})$$

其中  $n$  是根据阈值确定的主成分个数。

在本例中，根据 85%贡献率阈值选择 2 个主成分进行分析。因此，每个样本的综合得分可按照以下公式进行计算。

$$\text{综合得分} = \text{第1主成分得分} \times 0.6751 + \text{第2主成分得分} \times 0.1763$$

本例各样本的第 1、第 2 主成分及综合得分如表 7-2 所示。

根据表 7-2 所列示的结果，可以分别对第 1、第 2 主成分进行分析，也可以依据综合得分对各地区的发展进行综合分析。

表 7-2 主成分得分及综合得分

地 区	第 1 主成分		第 2 主成分		综 合	
	得 分	排 名	得 分	排 名	得 分	排 名
北 京	5.697979	2	3.551375	1	4.472813	1
山 西	-0.58103	12	0.404957	7	-0.32086	10
内蒙古	-0.66753	13	-0.00435	13	-0.45142	12
吉 林	-0.32792	8	-0.53561	17	-0.3158	9
黑龙江	-0.57168	11	0.418921	6	-0.31209	8
上 海	6.049685	1	-1.50079	24	3.819553	2
江 苏	1.154256	4	-1.30043	22	0.549973	4
浙 江	3.817791	3	-1.24587	21	2.357743	3
福 建	0.540295	5	0.011427	12	0.366768	5
山 东	0.42284	6	-0.3298	14	0.227315	6
河 南	-1.13698	16	0.384593	8	-0.69977	16
湖 北	-0.87939	14	-0.3325	15	-0.6523	15
湖 南	-0.46306	9	-0.84158	19	-0.46098	13
广 西	-0.93805	15	-0.8768	20	-0.78786	18
海 南	-1.59391	22	-0.45978	16	-1.15711	23
重 庆	0.4128	7	-1.3373	23	0.042915	7
四 川	-1.27316	18	-0.60026	18	-0.96533	20
云 南	-1.57886	21	0.373034	9	-1.00012	22
西 藏	-1.63371	23	0.663251	3	-0.98599	21

续表

地 区	第 1 主成分		第 2 主成分		综 合	
	得 分	排 名	得 分	排 名	得 分	排 名
陕 西	-0.4932	10	0.063999	11	-0.32168	11
甘 肃	-1.5033	20	0.600191	4	-0.90907	19
青 海	-1.89646	24	0.330405	10	-1.22205	24
宁 夏	-1.29782	19	0.535104	5	-0.78182	17
新 疆	-1.25958	17	2.027835	2	-0.49283	14

北京、上海等大城市及沿海地区由于地理位置优越及改革开放政策的优惠措施，人才普遍比较集中，其经济、教育等方面的发展比较好，人民生活水平也随之较高，受教育的程度也较高，因此体现为这些地区的第 1 主成分（即经济生活成分）排名靠前；而对于第 2 主成分（即就业成分）而言，反而是经济发展相对落后的地区排名靠前，究其原因，可能是由于经济发展滞后、人才竞争不激烈、流动性不大等原因造成的；从社会经济生活发展的总体情况看，北京、上海、江浙一带总体发展状况相对较好，其综合得分也名列前茅。

主成分分析也可以利用 SAS 编程语言中的 PRINCOMP 过程来实现，其具体语法如下。

```
PROC PRINCOMP <选项>;
  BY 变量;
  FREQ 变量;
  PARTIAL 变量;
  VAR 变量;
  WEIGHT 变量;
```

PRINCOMP 过程中最常用的语句是 VAR 语句，主要用于指定进行主成分分析的原始变量，而 FREQ、WEIGHT 语句与前面讲述的用法一致，PARTIAL 语句主要用于根据偏相关（或协方差）系数矩阵进行特征根及其对应特征向量的计算。

PRINCOMP 过程的主要常用选项如下。

- COV：指定系统利用协方差矩阵计算特征根及特征向量，默认则表示用相关系数矩阵计算。
- N=：指定系统提取主成分的数目，默认则表示有多少个原始变量，就提取多少个主成分。
- PREFIX=：可以根据用户指定的名字对主成分得分变量进行命名。
- STD：指定系统对主成分得分进行标准化。

本例主要使用原始变量的相关系数矩阵进行测算，其程序如下。

```
proc princomp data=Sasuser.Live out=WORK.SCORE; /*指定分析数据集，并把计算结果存储于 Score
临时数据集当中*/
  var GDP Income Consumption Employment Education Health Life; /*指定用于分析的原始变量*/
run;
```

运行程序后，可得到与上述分析过程一致的结果。

7.3 因子分析

因子分析是主成分分析的推广和发展，也是多元统计分析中降维的一种方法。主成分分

析通过线性组合将多个原始变量综合成若干个主成分。在多变量分析中,某些变量之间往往存在相关性。是什么原因使得变量间有关联呢?是否存在不能直接观测到但影响可观测变量变化的公共因子呢?

因子分析(Factor Analysis)就是寻找这些公共因子的模型分析方法,它是在综合原始变量信息的基础上构筑若干意义较为明确的公因子,以它们为框架分解原始变量,以此考察原始变量间的联系与区别。

### 7.3.1 因子分析的基本原理

因子分析的基本目的是用少数几个因子描述许多指标或因素之间的联系,即将比较密切的几个变量归在同一类中,每一类变量即成为一个因子(之所以称其为因子,是因为它往往是不可观测的,类似于隐变量),以较少的几个因子反映原始资料的大部分信息。如在评价学生多门课程的成绩时,可分别从各门课程的共性出发,提取出文科因子和理科因子以对学生进行综合考评。

主成分分析是因子分析的一个特例,通常可采用主成分法估算出因子个数。二者之间的区别和联系主要体现在以下几个方面。

- ❶ 因子分析是把原始变量表示成各因子的线性组合,而主成分分析则是把主成分表示成各原始变量的线性组合。
- ❷ 主成分分析的重点在于解释原始变量的总方差,而因子分析则把重点放在解释原始变量之间的协方差上。
- ❸ 因子分析中的因子个数可根据研究者的需要事先指定,指定因子数量不同,则分析结果可能不同。在主成分分析中,有几个变量,就有几个主成分。
- ❹ 在主成分分析中,当给定的协方差矩阵或相关矩阵的特征值是唯一时,主成分一般是唯一的;而因子分析中的因子不是唯一的,可以旋转得到不同的因子。

#### 1. 因子分析模型

因子分析是从研究变量内部相互依存关系出发,把一些具有错综复杂关系的变量归结为少数几个综合因子的一种多元统计分析方法。它的基本思想是将原始变量进行分类,将相关性较高,即联系比较紧密的变量分在同一类中,而不同类变量之间的相关性则较低,那么每一类变量实际上代表一个基本结构,即公共因子。对于所研究的问题,就是试图用最少数目的、不可观测的所谓公共因子的线性函数来描述所研究的对象。

因子分析的一般模型如下。

$$\begin{aligned} x_1 &= a_{11}F_1 + a_{12}F_2 + L + a_{1m}F_m + \varepsilon_1 \\ x_2 &= a_{21}F_1 + a_{22}F_2 + L + a_{2m}F_m + \varepsilon_2 \\ &\vdots \\ x_p &= a_{p1}F_1 + a_{p2}F_2 + L + a_{pm}F_m + \varepsilon_p \end{aligned}$$

其中,  $F_j (j=1,2,K,m)$  表示不可观测的因子或公因子组成的向量。在 SAS 系统中,可选择  $\alpha$  因子提取法、Harris 成分分析法、主成分法、最大似然法等方法进行因子提取。因子的含义必须结合具体问题的实际意义而定。

$a_{ij} (i=1,2,K,p; j=1,2,K,m)$  被称为因子载荷。因子载荷是第  $i$  个变量与第  $j$  个因子之间的相关系数,反映了第  $i$  个变量在第  $j$  个因子上的重要性,即表示变量  $x_i$  依赖于  $F_j$  的分

量（比重）。

在实际问题中，究竟取多少个因子进行分析呢？这可依据提取出来的主成分方差贡献率来决定，方差贡献率越大，因子分析越有意义。在通常情况下，可参考累积方差贡献率 85% 阈值进行判定。

## 2. 因子旋转

因子分析的目的不仅是找出因子，更重要的是知道每个因子的意义，以便对实际问题进行分析。如果因子的典型代表变量不很突出，为了能够更好地解释公因子  $F$ ，还需要进行因子旋转。通过适当的旋转可得到比较满意的主因子，即使得每个原始变量仅在一个公因子上有较大的载荷，而在其余的公因子上的载荷比较小。

进行因子旋转，就是要使因子载荷矩阵中因子载荷的平方值向 0 和 1 两个方向分化，即使大的载荷更大，小的载荷更小。在因子旋转过程中，按照旋转坐标轴的位置不同，如果主轴相互正交，则称之为正交旋转；如果主轴相互间不是正交的，则称之为斜交旋转。在 SAS 系统中，可供选择的因子旋转方法主要有方差最大化法、四分位最大法、平衡法、正交旋转法等，常用方法是方差最大化正交旋转法。

## 3. 因子得分

在因子分析中，人们往往更愿意用公因子反映原始变量，这样有利于描述研究对象的特征，因而往往将因子表示为原始变量的线性组合，即因子得分函数。

$$\begin{aligned} f_1 &= \beta_{11}x_1 + \beta_{12}x_2 + L + \beta_{1p}x_p \\ f_2 &= \beta_{21}x_1 + \beta_{22}x_2 + L + \beta_{2p}x_p \\ &L \\ f_m &= \beta_{m1}x_1 + \beta_{m2}x_2 + L + \beta_{mp}x_p \end{aligned}$$

用因子得分函数可计算每个样本的因子得分。

但因子得分函数中方程的个数  $m$  小于变量的个数  $p$ ，所以并不能精确计算出因子得分，只能对因子得分进行估计。估计因子得分的方法较多，常用的有回归估计法、Bartlett 估计法、Thomson 估计法等。

### 7.3.2 因子分析的基本步骤和过程

因子分析的核心问题有两个：一是如何构造因子变量，二是如何对因子变量进行命名解释。因此，因子分析的基本步骤和解决思路就是围绕这两个核心问题展开的。通常在进行因子分析之前，也要进行标准化或无量纲化（该过程在 SAS 系统中可自动实现），然后可按照以下顺序进行因子分析。

- 考察原始变量之间的相关性，如果各变量之间是独立的，那么可能不适合使用因子分析。
- 计算变量之间的相关系数矩阵，以作为分析基础。
- 确定提取公因子的方法并根据累积方差贡献率阈值进一步确认提取公因子的数目。
- 进行因子旋转，使公因子更具有可解释性。
- 计算各公因子得分。

- 用各公因子的方差贡献率作为权数,通过对各因子得分进行加权,得到综合评价指标得分。
- 可根据各公因子或综合得分进行排序考察。

本小节将结合例 7-2 详细讲解因子分析方法的过程和步骤。



例 7-2

为评价某省网吧业主对电信公司提供上网服务的满意情况,调查人员在该省范围内随机抽取了 70 家网吧进行上网服务满意度调查。其调查主要内容如表 7-3 所示(数据详见 Internet\_Cafe.sas7bdat)。试利用因子分析方法对该省网吧满意度进行综合评价。

表 7-3 网吧满意度调查情况表

变量名	Connection	Speed	Transformation	Offline	Switch	Timeliness	Initiative
变量 标签	接通率	网络速度	数据传输速率	掉线率	接通率	提供服务及时性	提供服务主动性
变量名	Attitude	Skill	Consideration	Standard	Settlement	Success	Efficiency
变量 标签	服务态度	服务熟练程度	对客户关注程度	服务规范性	解决问题能力	解决问题成功率	服务效率

在本例中,共使用了 14 个原始变量对网吧满意度问题进行分析。如直接利用该 14 个变量对满意度进行分析评价,则显得比较复杂。因此,要分析并提取出该 14 个变量共同反映的东西,即因子,以对研究对象进行综合评价,并根据综合评价得分值对各个网吧样本进行排序,找出评价较低的网吧,然后再进行深入访问,找出满意度较低的原因。

在 SAS 系统中,因子分析方法不能通过菜单操作的方式实现,但可以利用 SAS 编程语言中的 FACTOR 过程实现。FACTOR 过程的主要语法如下。

```
PROC FACTOR <选项>;
  VAR 变量;
  PRIORS 因子的共同度 ;
  PARTIAL 变量;
  FREQ 变量;
  WEIGHT 变量;
  BY 变量;
```

FACTOR 过程的语句类似于主成分分析使用的语句。最常用的是 VAR 语句,主要用于指定进行因子分析的原始变量(如果该语句缺省,则表示分析数据集中的所有变量),FREQ、WEIGHT、BY 语句与前面讲述的用法一致,而 PRIORS 语句通常与 VAR 语句搭配使用,用于指定 VAR 语句所指定变量对应的共同度(共同度的值须在 0~1 之间)。

```
proc factor;
  var    x    y    z;
  priors 0.7  0.8  0.9;
run;
```

在 FACTOR 过程中,用户需考虑的、最主要的问题是其选项的设置。FACTOR 选项可供选择的關鍵字多达 60 余个。在通常的因子分析中,常用的选项如下。

- DATA=: 指定用于因子分析的原始数据。
- OUTSTAT=: 指定用于存储因子分析结果的数据集。
- METHOD=: 指定用于提取公因子的方法。
- ROTATE=: 指定进行因子旋转的方法。
- PERCENT=: 指定选择公因子数目的阈值 (也可用 PROPORTION 语句替代)。
- SCORE: 计算得分系数。

其中, 指定提取公因子方法的 METHOD 语句可供选择的主要方法如下。

- ALPHA:  $\alpha$  因子分析法。
- HARRIS: Harris S-1RS-1 成分分析法。
- IMAGE: 图像协方差矩阵的成分分析法。
- ML: 极大似然法。
- PATTERN: 从数据集中按照不同指定类型读入因子载荷, 这些数据集的类型可以通过在指定数据类型的语句中用 TYPE = FACTOR、TYPE = CORR、TYPE = UCORR、TYPE = COV 或 TYPE = UCOV 指定。
- PRINCIPAL: 主成分分析法。
- PRINT: 迭代主成分因子分析法。
- SCORE: 从含有相关系数或斜方差矩阵的 TYPE = FACTOR、TYPE = CORR、TYPE = UCORR、TYPE = COV 或 TYPE = UCOV 的数据集中读入得分系数。
- ULS: 未加权最小平方法。

在通常情况下, 一般选择主成分分析法进行因子提取。在确定好因子提取的方法之后, 可以利用 PERCENT 语句或 PROPORTION 语句 (要注意 PERCENT 语句的取值在 0~100 之间, PROPORTION 取值为 0~1 之间) 设定累积方差贡献率的阈值, 系统会自动根据阈值确定应该得到的公因子个数。

ROTATE 语句主要用于指定进行因子旋转的方法, 可供选择的主要方法如下。

- BIQUARTIMAX: 复均等方差正交旋转法, 等价于 ROTATE=ORTHOMAX(.5)。
- EQUAMAX: 均等方差正交旋转法。
- FACTORPARSIMAX: 简约因子正交旋转法。
- NONE: 不指定因子旋转方法。
- ORTHCF(p1,p2): 具有两个权数的 Crawford-Ferguson 正交旋转法。
- ORTHGENCF(p1,p2,p3,p4): 具有 4 个权数的广义 Crawford-Ferguson 正交旋转法。
- ORTHOMAX<(p)>: 具有正交权数 p 的一般正交旋转法。
- PARSIMAX: 简约正交旋转法。
- QUARTIMAX: 最大四等分正交旋转法。
- VARIMAX: 方差最大化正交旋转法。

在日常分析活动中, 常常选择方差最大化正交旋转法 (简称方差最大化法) 即 VARIMAX 法进行因子旋转。



注意

在调用 FACTOR 过程时, FACTOR 选项中的 SCORE 语句只能用于在结果窗口中显示因子得分系数, 而不是每个样本的因子得分。

本例初步分析的程序如下。

```
proc factor data=Sasuser.Internet_Cafe /*调用 factor 因子分析过程，并指定用于分析的数据集*/
    method=Principal rotate=Varimax score outstat=IC_out; /*使用主成分法进行因子提取，并利用方差最
大化法进行因子旋转，并把分析和计算得到结果存储于 IC_out 临时数据集当中*/
    var Connection Speed Transformation Offline Switch Timeliness Initiative
        Attitude Skill Consideration Standard Settlement Success Efficiency; /*指定用于分析的原始变量*/
run;
```

运行程序后，可以得到以下一系列的分析结果。

**STEP 1** SAS 系统在“Output”窗口中给出了主成分法由相关系数矩阵计算出的特征根及其贡献率，如图 7-10 所示。

The FACTOR Procedure				
Initial Factor Method: Principal Components				
Prior Communality Estimates: ONE				
Eigenvalues of the Correlation Matrix: Total = 14    Average = 1				
	Eigenvalue	Difference	Proportion	Cumulative
1	7.95334795	5.51006487	0.5681	0.5681
2	2.44328308	1.65802976	0.1745	0.7426
3	0.78525333	0.13317206	0.0561	0.7987
4	0.65208127	0.15658076	0.0466	0.8453
5	0.49550051	0.06833330	0.0354	0.8807
6	0.42716721	0.09507560	0.0305	0.9112
7	0.33209162	0.08873385	0.0237	0.9349
8	0.24335776	0.06082582	0.0174	0.9523
9	0.18253195	0.03947611	0.0130	0.9653
10	0.14305583	0.01688077	0.0102	0.9755
11	0.12617506	0.02207020	0.0090	0.9846
12	0.10410487	0.04463610	0.0074	0.9920
13	0.05946877	0.00688799	0.0042	0.9962
14	0.05258078		0.0038	1.0000
2 factors will be retained by the MINEIGEN criterion.				

图 7-10 因子分析的特征根及其贡献率

**STEP 2** 图 7-10 中给出了所有计算出来的特征根，可以根据特征根的贡献来决定应当选择的公因子数目。从各个特征根的贡献（“Proportion”列）来看，前两个特征根的贡献率分别是 0.5681 和 0.1745，远远大于其他特征根贡献率。二者特征根累积贡献率达到 74.26%，基本上可以在较大程度上反映原始数据的信息。在进行后续的因子分析之前，可以先选择两个因子进行分析（本例系统默认为选择了两个公因子）。

**STEP 3** 因此，可以根据上述前两个特征根的累积贡献率为系统设定阈值，程序修改如下。

```
proc factor data=Sasuser.Internet_Cafe
    method=Principal rotate=Varimax percent=70 score outstat=IC_out;
    var Connection Speed Transformation Offline Switch Timeliness Initiative
        Attitude Skill Consideration Standard Settlement Success Efficiency;
run;
```

在修改的程序中，在 FACTOR 语句的选项中加入了 “percent=” 关键字。

PENCENT 关键字可以为系统指定阈值，并根据该阈值自动为用户选择应该选取的公因子个数。PENCENT 关键字的取值范围在 0~100 之间，等价于取值范围在 0~1 之间的 “proportion=” 关键字，即本例中的 “percent = 70” 等价于 “proportion = 0.7”。

**STEP 4** 运行修改的程序之后，除得到图 7-10 所示的结果之外，还可以得到因子载荷，如图 7-11 所示。

Factor Pattern		Factor1	Factor2
Connection	接通率	0.47346	0.54915
Speed	网络速度	0.49493	0.67766
Transferration	数据传输速率	0.59511	0.69125
Offline	掉线率	0.50854	0.50149
Switch	接通率	0.61266	0.62953
Timeliness	提供服务及时性	0.81700	-0.23957
Initiative	提供服务主动性	0.75175	-0.38517
Attitude	服务态度	0.83362	-0.18528
Skill	服务熟练程度	0.88325	-0.15184
Consideration	对客户关注程度	0.75192	-0.34135
Standard	服务规范性	0.82481	-0.36561
Settlement	解决问题能力	0.93207	-0.16371
Success	解决问题成功率	0.90998	-0.10991
Efficiency	服务效率	0.91975	-0.06828

图 7-11 公因子的因子载荷

**STEP 5** 根据因子载荷，可以利用因子分析模型写出原始变量与公因子之间的关系。

$Connection = 0.47346 \times Factor1 + 0.54915 \times Factor2$

$Speed = 0.49493 \times Factor1 + 0.67766 \times Factor2$

M

$Efficiency = 0.91975 \times Factor1 - 0.06828 \times Factor2$

那么，提取出来的公因子对每个变量的解释程度到底有多大呢？这可从公因子方差表得知，如图 7-12 所示。

Variance Explained by Each Factor						
		Factor1	Factor2			
		7.9533480	2.4432831			
Final Communality Estimates: Total = 10.396631						
Connection	Speed	Transferration	Offline	Switch	Timeliness	Initiative
0.52572660	0.70417572	0.83198563	0.51010938	0.77165884	0.72487734	0.71349100
Attitude	Skill	Consideration	Standard	Settlement	Success	Efficiency
0.72925773	0.80318148	0.68190654	0.81397822	0.89554596	0.84014343	0.85059317

图 7-12 因子分析的公因子方差表

**STEP 6** 在图 7-12 中，系统给出了公因子对所有原始变量的解释能力。这些解释能力的总和是 “Total=” 所代表的数值 (10.396631)，本例中的公因子对每个原始变量的解释能力均较大，其平均值就是所选取的公因子个数对应特征根的累积贡献率，即 0.7426。而 “Variance Explained by Each Factor” 则表明了每个因子的贡献。

**STEP 7)** 为了更好地解释公因子 *Factor1* 和 *Factor2*，可通过因子旋转的方法，使得图 7-11 所示的每个变量仅在一个公因子上有较大的载荷，而在其余的公因子上的载荷比较小。在本例中，在 FACTOR 过程的选项中加入关键字 “Rotate=”，并指定了 “Varimax”，即方差最大化正交旋转法进行因子旋转，可得到图 7-13 所示的因子载荷结果和图 7-14 所示的旋转后的公因子方差表。

The FACTOR Procedure			
Rotation Method: Varimax			
Orthogonal Transformation Matrix			
	1	2	
1	0.87359	0.48667	
2	-0.48667	0.87359	
Rotated Factor Pattern			
		Factor1	Factor2
Connection	接通率	0.14635	0.71015
Speed	网络速度	0.10257	0.83286
Transferration	数据传输速率	0.18347	0.89349
Offline	掉线率	0.20020	0.68559
Switch	接通率	0.22883	0.84811
Timeliness	提供服务及时性	0.83031	0.18832
Initiative	提供服务主动性	0.84417	0.02938
Attitude	服务态度	0.81841	0.24384
Skill	服务熟练程度	0.84549	0.29720
Consideration	对客户关注程度	0.82299	0.06774
Standard	服务规范性	0.89847	0.08202
Settlement	解决问题能力	0.89391	0.31060
Success	解决问题成功率	0.84844	0.34684
Efficiency	服务效率	0.83670	0.38797

图 7-13 方差最大化正交因子旋转的结果

Variance Explained by Each Factor						
		Factor1	Factor2			
		6.6482965	3.7483345			
Final Communality Estimates: Total = 10.396631						
Connection	Speed	Transformation	Offline	Switch	Timeliness	Initiative
0.52572660	0.70417572	0.83198563	0.51010938	0.77165884	0.72487734	0.71349100
Attitude	Skill	Consideration	Standard	Settlement	Success	Efficiency
0.72925773	0.80318148	0.68190654	0.81397822	0.89554596	0.84014343	0.85059317

图 7-14 因子旋转后的公因子方差表

**STEP 8)** 从因子旋转的结果可明显看到，各个原始变量在 *Factor1* 和 *Factor2* 两个因子上的载荷数值差距较图 7-11 所示的数值差距大，因而使得这两个因子的意义显得更加明显。

由图 7-13 中的因子载荷得知，公因子 *Factor1* 与提供服务及时性、提供服务主动性、服务态度、服务熟练程度、对客户关注程度、服务规范性、解决问题能力、解决问题成功率和效率等变量的正相关性很强，载荷均达到了 0.81 以上。而公因子 *Factor2* 与接通率、网络速度、数据传输速率、掉线率和接通率等变量的正相关性很强。

与 *Factor1* 相关性较强的原始变量均是从电信公司服务人员为网吧提供服务的角度来进行满意度衡量的，因此可以把公因子 *Factor1* 命名为“人员服务质量”因子；而与 *Factor2* 相关性较强的原始变量都是从电信公司的技术角度来考察网吧用户满意度的，因此可以把公

因子 *Factor2* 命名为“技术质量”因子。

由图 7-14 得知，经过旋转之后，*Factor1* 和 *Factor2* 的解释能力有所变化，但其总的解释能力（10.396631）与对各原始变量的解释能力不变。

**STEP 9** 对提取出来的公因子进行实际意义上的命名之后，可以根据 SAS 系统估计出的因子得分系数（如图 7-15 所示），并利用因子得分函数计算出每个样本在公因子上的因子得分。

The FACTOR Procedure			
Rotation Method: Varimax			
Scoring Coefficients Estimated by Regression			
Squared Multiple Correlations of the Variables with Each Factor			
		Factor1	Factor2
		1.0000000	1.0000000
Standardized Scoring Coefficients			
		Factor1	Factor2
Connection	接通率	-0.05738	0.22532
Speed	网络速度	-0.08062	0.27258
Transformation	数据传输速率	-0.07232	0.28357
Offline	掉线率	-0.04403	0.21042
Switch	接通率	-0.05810	0.26258
Timeliness	提供服务及时性	0.13746	-0.03567
Initiative	提供服务主动性	0.15929	-0.09172
Attitude	服务态度	0.12847	-0.01524
Skill	服务熟练程度	0.12726	-0.00024
Consideration	对客户关注程度	0.15058	-0.07604
Standard	服务规范性	0.16342	-0.08025
Settlement	解决问题能力	0.13498	-0.00150
Success	解决问题成功率	0.12184	0.01638
Efficiency	服务效率	0.11462	0.03187

图 7-15 标准化的因子得分系数

根据图 7-15 所示的因子得分系数，可以得到以下因子得分函数。

$$\begin{aligned} f_1 = & -0.05738 \times \text{Connection} - 0.08062 \times \text{Speed} - 0.07232 \times \text{Transformation} \\ & - 0.04403 \times \text{Offline} - 0.05810 \times \text{Switch} + 0.13746 \times \text{Timeliness} \\ & + 0.15929 \times \text{Initiative} + 0.12847 \times \text{Attitude} + 0.12726 \times \text{Skill} \\ & + 0.15058 \times \text{Consideration} + 0.16342 \times \text{Standard} + 0.13498 \times \text{Settlement} \\ & + 0.12184 \times \text{Success} + 0.11462 \times \text{Efficiency} \\ f_2 = & 0.22532 \times \text{Connection} + 0.27258 \times \text{Speed} + 0.28357 \times \text{Transformation} \\ & + 0.21042 \times \text{Offline} + 0.26258 \times \text{Switch} - 0.03567 \times \text{Timeliness} \\ & - 0.09172 \times \text{Initiative} - 0.01524 \times \text{Attitude} - 0.00024 \times \text{Skill} \\ & - 0.07604 \times \text{Consideration} - 0.08025 \times \text{Standard} - 0.00150 \times \text{Settlement} \\ & + 0.01638 \times \text{Success} + 0.03187 \times \text{Efficiency} \end{aligned}$$

**STEP 10** 只要把每个样本的原始变量数据代入上述两个式子当中，便可以计算出在两个公因子上的因子得分。

在 SAS 系统中，用户如需得到因子得分，无须按照上述过程进行如此麻烦的手工计算。



在 SAS 系统中进行因子得分的计算过程并不包含在 FACTOR 过程当中，而是使用 SCORE 过程进行因子得分的计算。

SCORE 过程的主要语法如下。

```
PROC SCORE DATA=数据集 < 选项>:
```

```
  BY 变量;
```

```
  ID 变量;
```

```
  VAR 变量;
```

SCORE 过程用于因子得分计算时，应当与 FACTOR 过程配合使用，且应当放在 FACTOR 过程之后。SCORE 语句的选项具有以下主要关键字。

- DATA=: 指定用于因子分析的原始数据集（该数据集需与 FACTOR 过程所用原始数据集一致）。
- SCORE=: 指定存储因子分析结果的数据集（该数据集由 FACTOR 过程中的结果输出语句 OUTSTAT 关键字指定）。
- OUT=: 指定用于存储因子得分的数据集。

在使用 SCORE 过程进行因子得分计算之前，应当在 FACTOR 过程中使用“OUTSTAT=”关键字指定存储因子分析结果的数据集，并把该数据集应用于 SCORE 过程中的“SCORE=”关键字当中，作为计算因子得分的依据。

本例从因子提取到因子旋转再到因子得分计算全过程的程序如下。

```
proc factor data=Sasuser.Internet_Cafe          /*该过程参数设置与前面进行的分析一致*/
  method=Principal rotate=Varimax percent=70 score outstat=IC_out;
  var Connection Speed Transformation Offline Switch Timeliness Initiative
      Attitude Skill Consideration Standard Settlement Success Efficiency;
run;

proc score data=Sasuser.Internet_Cafe score=IC_out out=Sasuser.IC_Score; /*指定与 FACTOR 过程分析
原始数据集一致的分析对象，把 FACTOR 过程中由 OUTSTAT 输出的数据集指定为关键字“score=”分析的数据集，并把因子得分输出结果存储于 Sasuser.IC_Score 数据集当中*/
  var Connection Speed Transformation Offline Switch Timeliness Initiative
      Attitude Skill Consideration Standard Settlement Success Efficiency; /*指定参与得分计算的变量*/
run;

proc print data=Sasuser.IC_Score label; /*调用 PRINT 过程把因子得分结果打印在“Output”窗口中，
并将其设置为显示数据集变量的标签*/
  title 'Scores from Factor Analysis';      /*指定输出结果的名字*/
  var factor1 factor2;                      /*只输出 factor1 和 factor2 变量的值*/
  label factor1='人员服务质量' factor2='技术质量'; /*把因子 factor1 的变量标签命名为“人员服务质量”，把因子 factor2 的变量标签命名为“技术质量”*/
run;
```

运行程序后，本例输出存储因子得分结果的数据集中包含了每个样本原始变量的数值和 *Factor1*、*Factor2* 两个公因子的得分。因此，可以分别按照两个因子所反映的“人员服务质量”和“技术质量”进行因子得分排序，分别找出得分最低的网吧以进行进一步深入调查。如在本例中，变量“No”（编号）为 24 的网吧在“人员服务质量”上得分最低，为-4.7640 分；编号为 38 的网吧在“技术质量”上得分最低，为-3.3982 分。

在实际问题中，还可以根据各公因子得分及其贡献计算每个样本的综合得分。通常可按照以下公式进行综合得分计算。

$$\text{综合得分} = \sum_{i=1}^n (\text{第}i\text{公因子得分} \times \text{对应特征根贡献率})$$

其中  $n$  是根据阈值确定的公因子个数。  
也可直接根据公因子贡献数值大小进行加权，如：

$$\text{综合得分} = \sum_{i=1}^n (\text{第}i\text{公因子得分} \times \text{对应特征根贡献})$$

在使用特征根贡献时，一般使用因子旋转之后的贡献。通过这些方法计算出来的综合得分基本上不会改变各样本排序的相对顺序。本例采用特征根贡献对公因子进行加权计算综合得分，具体程序如下。

```
data Sasuser.IC_Comprehensive;
  set Sasuser.IC_Score;
  keep no factor1 factor2 Comprehensive_Score;
  Comprehensive_Score=factor1*6.6482965+factor2*3.7483345;
run;
proc print data=Sasuser.IC_Comprehensive;
  var no Comprehensive_Score;
run;
```

运行程序后便可生成名为“Sasuser.IC\_Comprehensive”的数据集，“Comprehensive\_Score”变量表示计算出来的综合得分。根据综合得分从大到小的排名情况可以看到，编号为 24 的网吧排名最后。电信公司应当针对排名靠后的网吧进行深入调查，找出改进服务的路径和手段，促进整体满意度的提升。

## 7.4 本章小结

本章主要介绍了主成分分析和因子分析的基本原理及其分析步骤，并用 SAS 系统展示了详细的分析过程，主要内容简要回顾如下：数据降维是分析复杂对象的有利工具之一，主成分分析和因子分析是常用的两种数据降维方法；主成分分析是因子分析的特例，因子分析是主成分分析的推广和发展；可根据方差贡献率阈值来确定应当提取主成分或公因子的个数；主成分分析与因子分析均可以计算综合得分，从而对分析对象进行评价；因子分析可采用因子旋转的方法使得因子的现实解释能力加强；在 SAS 系统的因子分析过程中，FACTOR 过程与 SCORE 过程应当配合使用。

## 第 8 章

# 聚类分析与判别分析

“物以类聚、人以群分”，这往往被人们视为自然的法则。正是由于不同现象之间客观存在的共性，使得大千世界有了界限的划分和质的区别，从而呈现出五花八门的景象。

事物分类思想上最为瞩目是生物分类学的发展，这也是统计分类发展的主要动力。在希腊时期，亚里士多德仅描述了 500 个物种；17 世纪后，人们知道了约 6 000 种植物；而仅仅 100 年后，植物学家又发现了 12 000 个新种。因此，对生物物种进行科学的分类变得极为迫切。林奈把自然界分为 3 界，即动物界、植物界和矿物界，并提出了纲、目、属、种等分类概念。人们可以依照各门类物种的典型特征，把新发现的物种归类至现有的门类当中。

近代统计分析中的聚类和判别分析受到了生物分类学的影响。在现实生活中，需要对复杂的对象依据一定的标准进行分类。有了既定的类别之后，还可涉及到对事物进行归类。因此，本章将重点阐述聚类和判别分析。

### 8.1 聚类分析的基本原理

在通常情况下，人们根据事物现象的一个指标或某一个方面，可以很容易进行分类活动。如按照收入指标把全社会人群划分为高、中、低 3 类。在进行归类时，只需考查新加入的对象在某一个指标上的表现是否符合特定类别即可。

而实际上，需考察的事物或对象往往不只包含单一指标，因此，很可能还需通过许多侧面或指标来进行综合考察。如按照经济发展、教育水平、面积大小、人口等诸多方面对我国地市级以上城市进行分类，按照考试成绩、社会实践、思想品德等方面划分学生奖学金的等级等。这些指标在反映事物特征的作用、量纲、紧密关系等方面可能有所不同，因此很难按照单一指标分类的原则进行分类和归类，而需要考虑采用多元统计分析的方法进行分类。

采用多元统计分析中的分类方法既可以对样本进行分类（记为 Q 型分类），也可以对反映事物特征的指标或变量（记为 R 型分类）进行分类。这两种分类是对等的，在算法上没有任何区别，本书以 Q 型分类为例进行详细讲解。

“近朱者赤，近墨者黑”。在一般情况下，人们往往可根据事物之间的远近距离来判定类别。个体与个体之间的距离越近，其相似程度可能也越高，属于同类的可能性越大。因此，下面介绍分类的原则。

#### 8.1.1 分类的基本原则

首先考虑在没有进行分类之前，所有参加分类过程的个体没有归入任何类别，即每个个

体自成一类。

有了一定的分类原则之后，人们可以根据个体与个体之间的距离长短进行分类。如首先把最近的个体分为同类，然后再根据最短距离继续扩大类别所涵盖的范围，直到把所有个体都分为一个大类为止。整个分类过程就如同对生活在地球上的人进行分类，首先每个人都是自成一类，然后有了人种的区分，最后所有人都可以被分到“人类”这个类别当中，即所有人都是一类。在数据分析过程中，人们通常把类似分类过程称为“系统聚类”方法。

而聚类过程所依据的距离主要有明氏距离、马氏距离等几大类。那么究竟什么是距离呢？设样本数据可以用以下矩阵形式表示。

$$X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}, \text{ 记为 } X = \{x_{ij}\}_{n \times p}$$

设  $d_{ij}$  表示第  $i$  个样本与第  $j$  个样本之间的距离。如果  $d_{ij}$  满足以下 4 个条件，则称其为“距离”。

- $d_{ij} \geq 0$ ，对一切  $i, j$ 。
- $d_{ij} = 0$ ，等价于  $i, j$ 。
- $d_{ij} = d_{ji}$ ，对一切  $i, j$ 。
- $d_{ij} \leq d_{ik} + d_{kj}$ ，对一切  $i, j, k$ 。

第 1 个条件表明聚类分析中的距离是非负的数；第 2 个条件表明个体自身与自身的距离为 0；第 3 个条件表明距离的对等性，即 A 和 B 之间的距离与 B 和 A 之间的距离是一致的；最后一个条件表明两点之间直线距离是最短的。

明氏距离是最常用的距离之一，其计算公式为：
$$d_{ij}(q) = \left( \sum_{k=1}^p |x_{ik} - x_{jk}|^q \right)^{1/q}。$$

明氏距离有以下几种典型情况。

- 当  $q = 1$  时： $d_{ij}(1) = \sum_{k=1}^p |x_{ik} - x_{jk}|$ ，被称为“绝对距离”。
- 当  $q = 2$  时： $d_{ij}(2) = \left( \sum_{k=1}^p |x_{ik} - x_{jk}|^2 \right)^{1/2}$ ，被称为“欧氏距离”。
- 当  $q = \infty$  时： $d_{ij}(\infty) = \max_{1 \leq k \leq p} |x_{ik} - x_{jk}|$ ，被称为“车比雪夫距离”。

但是明氏距离的大小与个体指标的观测单位有关，没有考虑指标之间的相关性。为克服明氏距离的缺点，可以考虑采用马氏距离进行改进。马氏距离是由协方差矩阵计算出来的相对距离，具体计算公式如下。

$$d_{ij} = (X_i - X_j)' \Sigma^{-1} (X_i - X_j)$$

在对事物现象进行分类的过程中，可以依据前述的最短距离原则进行聚类分析。

除了依据最短距离原则进行分类之外，还可以采用相关系数、相似系数、匹配系数等指

标来衡量个体之间的相似性，并以此为依据进行分类。

在分类过程中，为了便于分析，还应当注意以下 3 个重要原则。


- 同质性原则，即同一类中的个体之间有较强的相似性。
- 互斥性原则，即不同类中的个体差异很大。
- 完备性原则，即每个个体在同一次分类过程中，能且只能分在一个类别当中。

同质性原则保证了类别之内个体特征的共性；互斥性原则保证了类别之间的差异性；而完备性原则则说明了每一个个体应当包含在所进行的分类当中，而且每一个个体不能同时被分在不同的类别当中。

在实际应用中，以最短距离原则进行的“系统聚类”比较常用。本节将以此为依据进行详细的分类过程描述。

8.1.2 单一指标的系统聚类过程

为了更好地理解最短距离分类的基本原理，首先考察最简单的单一指标情况。

 **例 8-1**

为考察公司的经营业绩并对其进行分类，可依据它们的年盈利额来进行归类。具体数据如表 8-1 所示。

表 8-1 公司年盈利额

公 司	年盈利（十万元）
甲	1
乙	3
丙	9
丁	14

为直观起见，把表 8-1 的数据排列在数轴上进行分析，并用数轴上的点代表各个公司相应的财务指标，如图 8-1 所示。

直观地看，哪两个点距离最近呢？当然是甲和乙，它们之间的距离是  $2(3 - 1)$ 。如果按照最短距离原则来归类，首先要把甲、乙两点聚合成一类。以后为了方便，称之为“类(甲乙)”。于是就把它们归为一类，如图 8-2 所示。

在图 8-2 所示的分类过程中，可增加一个维度（即增加纵轴）表示两点之间的距离。如甲和乙之间的距离为 2，用横线把代表甲和乙的点连接起来，连线的高度便是纵轴所代表的“2”。

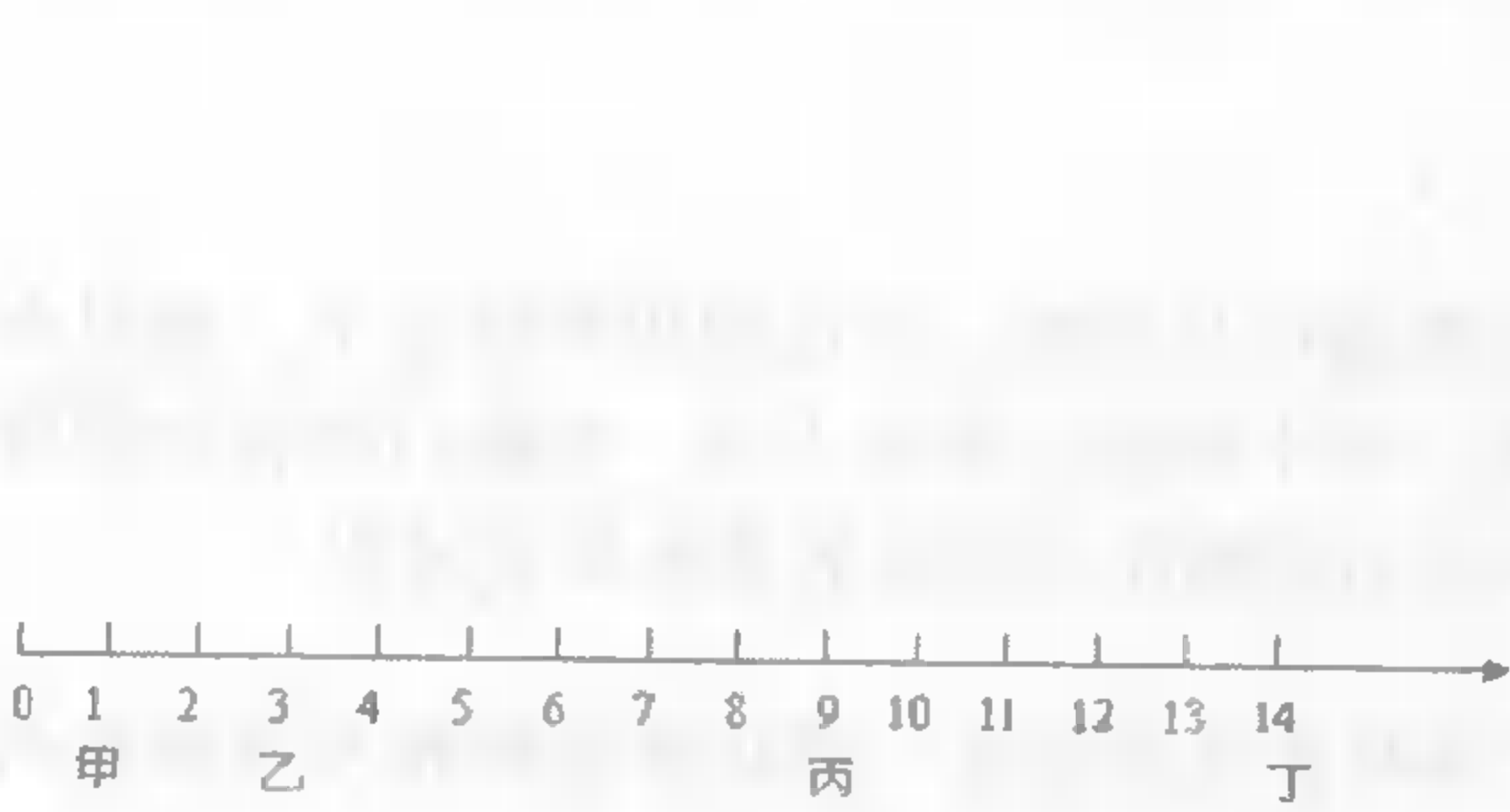


图 8-1 单一指标的分类过程（一）

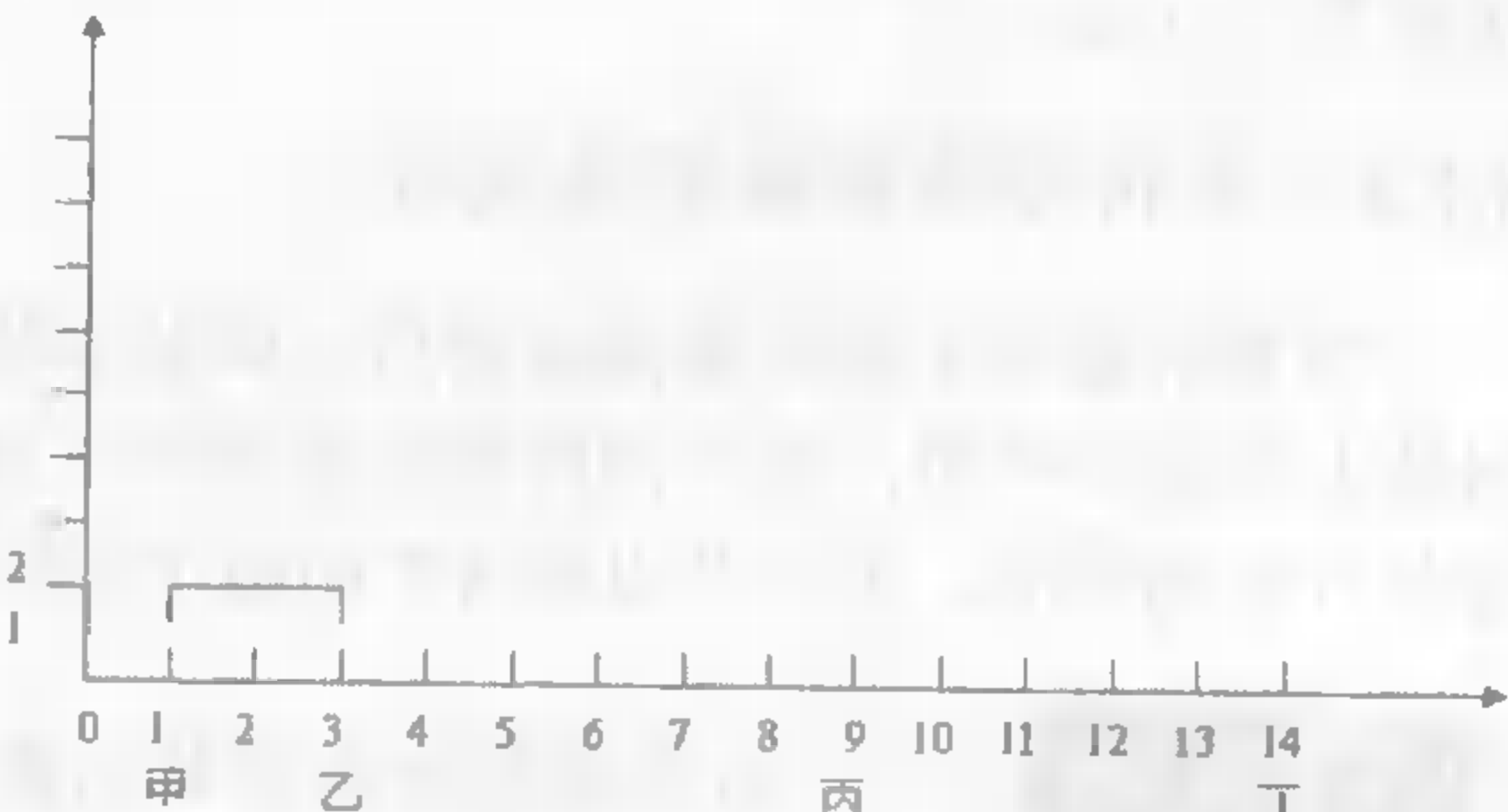


图 8-2 单一指标的分类过程（二）

继续观察，发现在剩下的点中，丙、丁的距离最近，为 5(14-9)，因此可把二者聚为“类(丙丁)”。于是就把它们归为一类，如图 8-3 所示。

这样，该分类过程就包含 2 类，即“类(甲乙)”和“类(丙丁)”。在这两类相互聚合的过程中，可能有 4 个距离，即甲丙、甲丁、乙丙、乙丁，其距离分别是 8(9-1)、13(14-1)、6(9-3)和 11(14-3)。

如果按照最短距离原则来归类，那么上述的乙丙之间的距离是最短距离，因此该距离就代表了“类(甲乙)”和“类(丙丁)”的距离。至此，整个聚类过程就完成了，如图 8-4 所示。

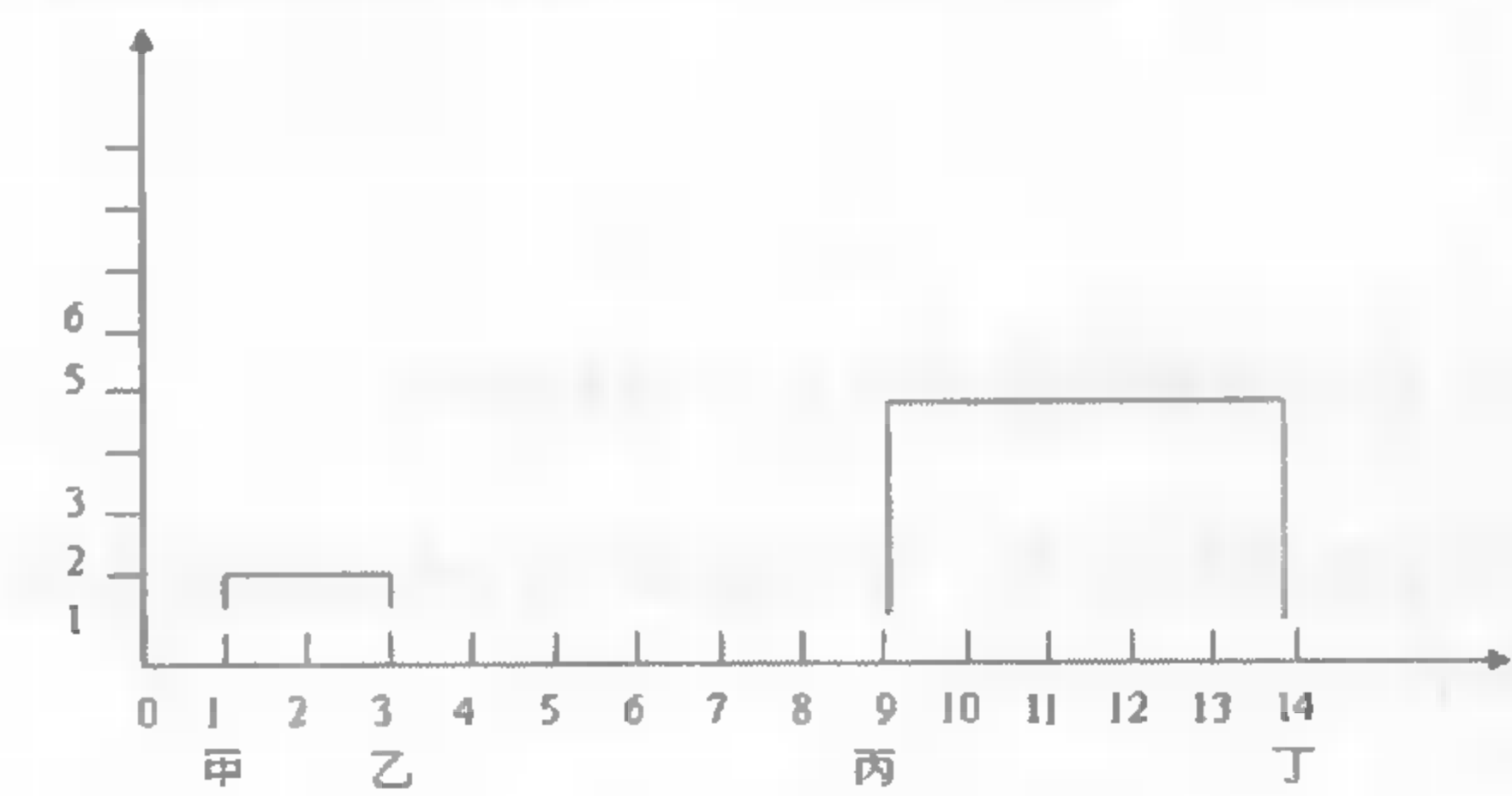


图 8-3 单一指标的分类过程（三）

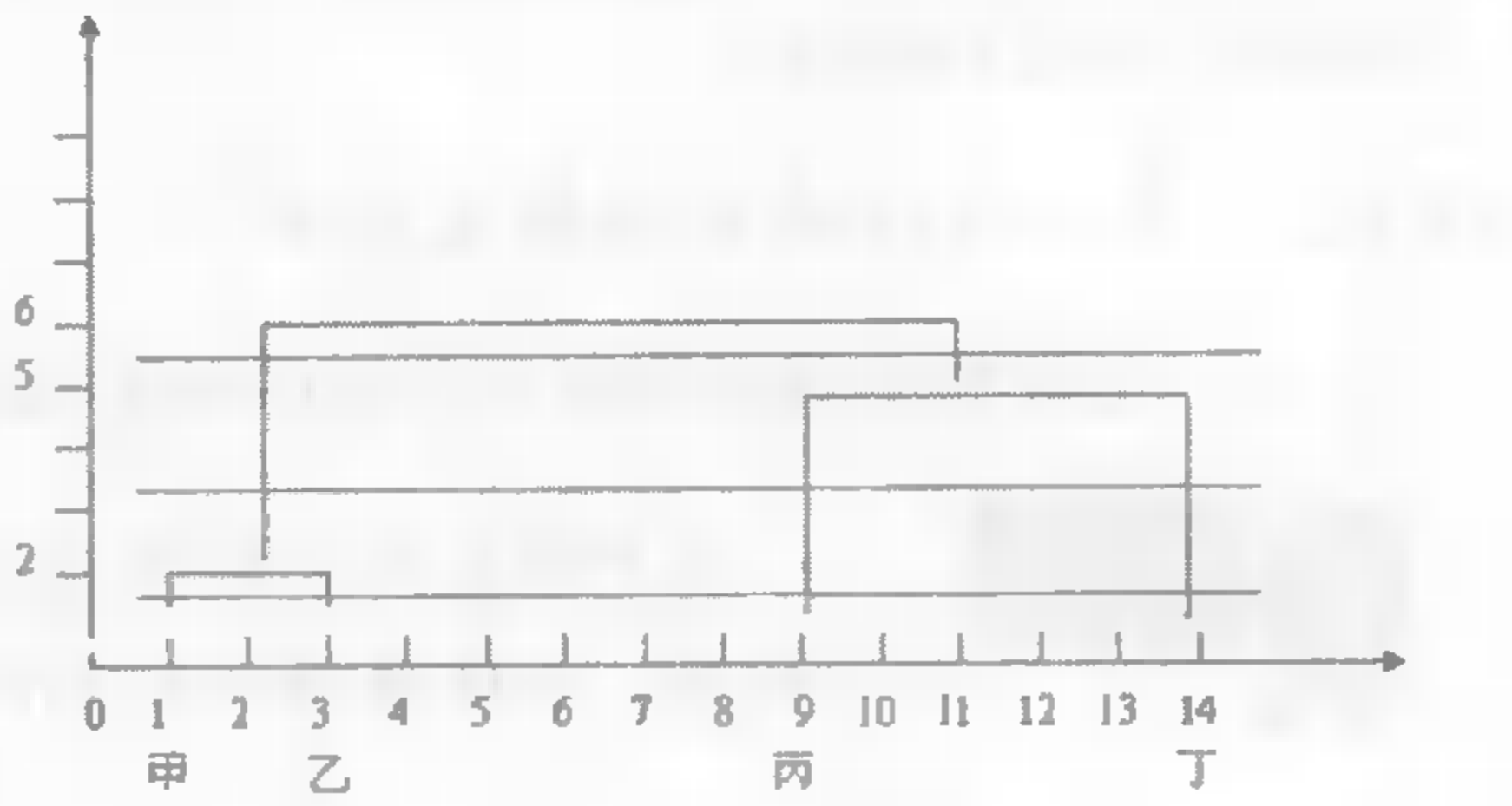


图 8-4 单一指标的分类过程（四）

上述过程可以形象地解答什么叫“聚类”。


聚类的结果（如图 8-4 所示）如同一棵大树的根系，最上面是主根，再往下就是主根的分叉，这种分叉一直分下去，最后就是这一组事物的各个个体。通俗地说，根系的顶端是主根，表明任何一组事物最终聚为一类；反之从根系的末端来看，如果归类达到最详细和最具体，那么各个个体自成一类，即每个个体自身都可看成是一个类。

而描述上述分类过程的图形可以被称为系统聚类过程中的“谱系聚类图”，简称“谱系图”，因而系统聚类又可以被称为“谱系聚类”。

在谱系图中，存在着若干层次的类。如图 8-4 所示，从上往下看，有 3 个层次。即在第 1 层次的水平上可以分为两类（在距离在 5~6 之间的任意位置画一条直线，与整个根系有两个交点），第 2 层次上可分为 3 类（在距离在 2~5 之间的任意位置画一条直线，与根系有 3 个交点），第 3 层次上可以分 4 个类。这样，整个系统分类的过程可以按照不同层次的类别对个体进行不同的分类，每一层次上的分类相对其他层次分类而言不是独立的，即高层次的分类是在低层次的分类基础上完成的。因此，从这个意义上来说，系统聚类也可以被称为“层次聚类”方法。

8.1.3 多指标的系统聚类过程

当要根据多个特征或指标对所反映的事物现象进行分类时，其过程相对较复杂。如对全国的大学进行分类，应当同时综合考虑学生生源、学术声誉、师资力量、学科门类齐全程度等各方面的情况。本小节以例 8-2 的两个指标分类为例描述多指标的系统聚类过程。



**例 8-2** 为考察投资者的盈利能力并对其进行分类，可从资金的投入与回报两个方面进行考察。具体数据如表 8-2 所示。

表 8-2 投资者资金投入与回报情况

投 资 者	资金投入（万元）	回报（万元）
A	35	60
B	15	40
C	30	5
D	80	8
E	90	35

现在要根据“资金投入”和“回报”两个指标对 A、B、C、D、E 5 个投资者进行分类。根据最短距离原则，本例的问题实际上就是以二维空间中的最短距离来进行聚类。

首先用一个二维坐标轴分别表示“资金投入”和“汇报”两个指标，根据对应的指标值，把 5 个投资者所代表的点描绘在图 8-5 所示的二维坐标轴上。

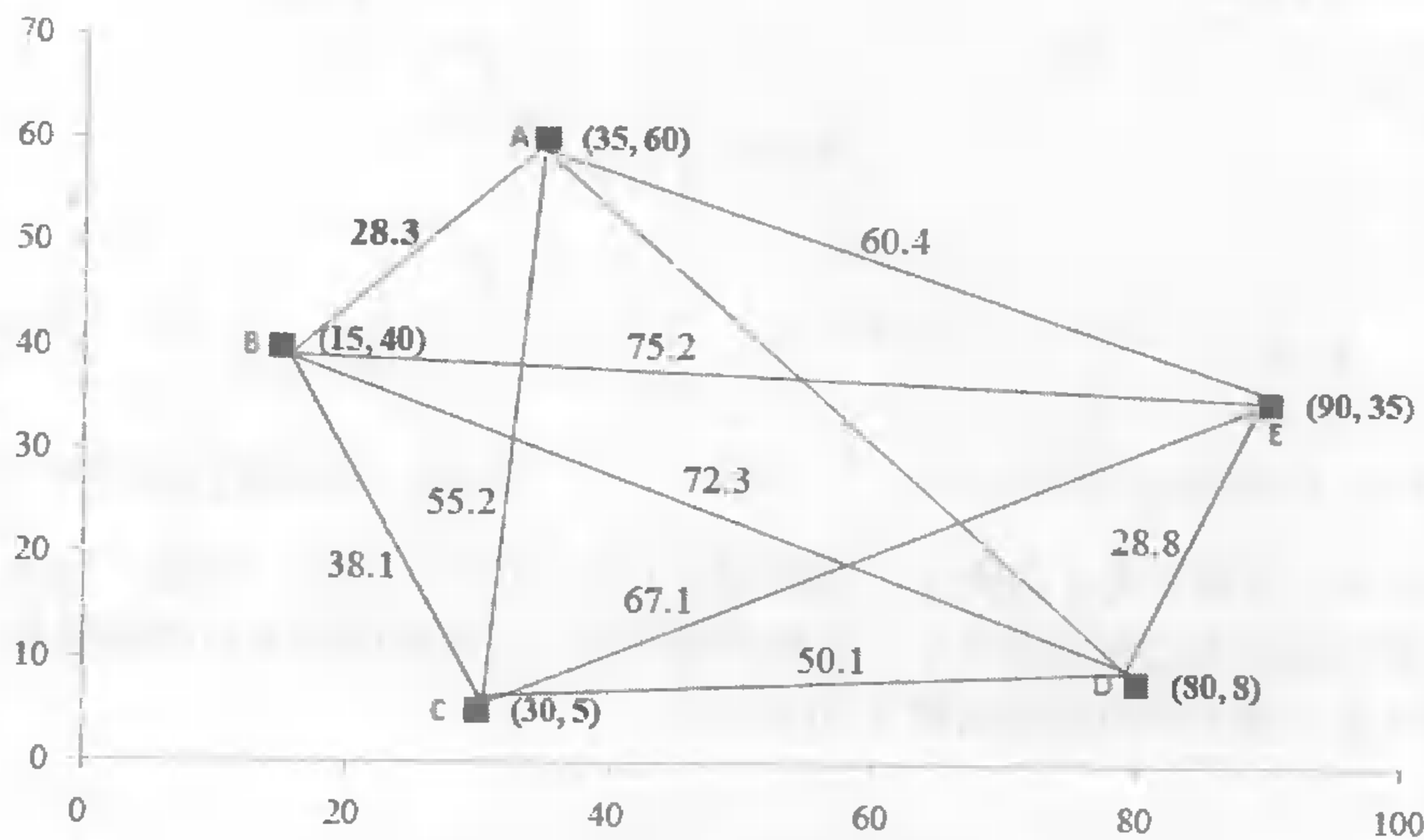


图 8-5 多指标的分类过程（一）

本例采用欧氏距离进行距离计算，如 A、B 两点之间的欧式距离为：

$$\sqrt{(35-15)^2+(60-40)^2}=28.3$$

其余各两点之间的欧氏距离均可被计算出来，并用方框将它们标注在图 8-5 所示的坐标轴上。从图 8-5 中可以直观地看到，A、B 之间的距离最近（28.3）。因此，按照最短距离原则，可以把 A 和 B 聚为一类，记为类 4，如图 8-6 所示。

类别与类别之间考虑到是以最短距离原则聚类的，因此在类 4 与外部的距离中（类 4 含有 A、B 两个点，故其与外部的距离有两个），最短者才是有意义的。所以，可以把图 8-6 中不是最短距离的连线去掉，从而得到图 8-7 所示的分类过程。

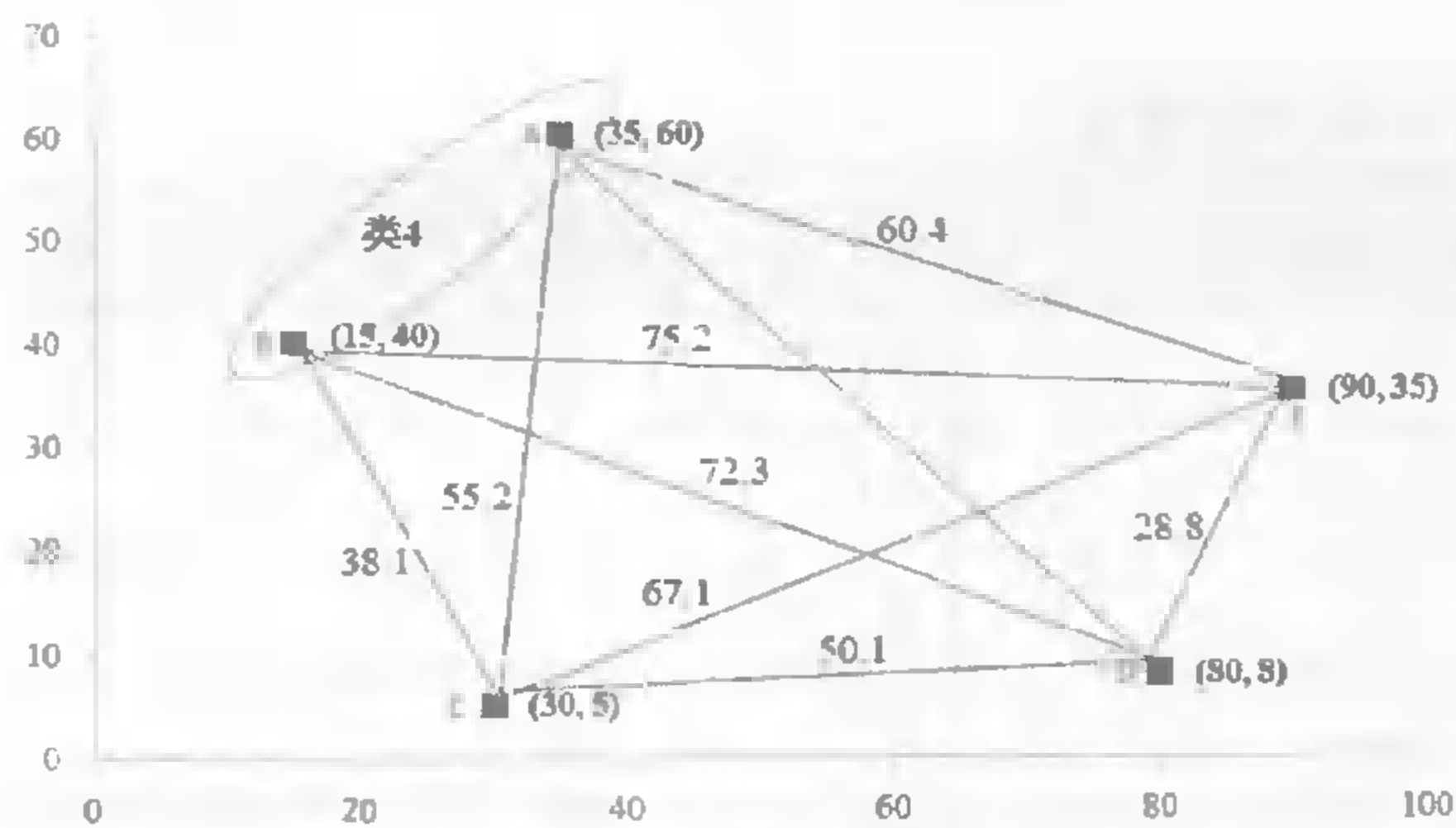


图 8-6 多指标的分类过程（二）

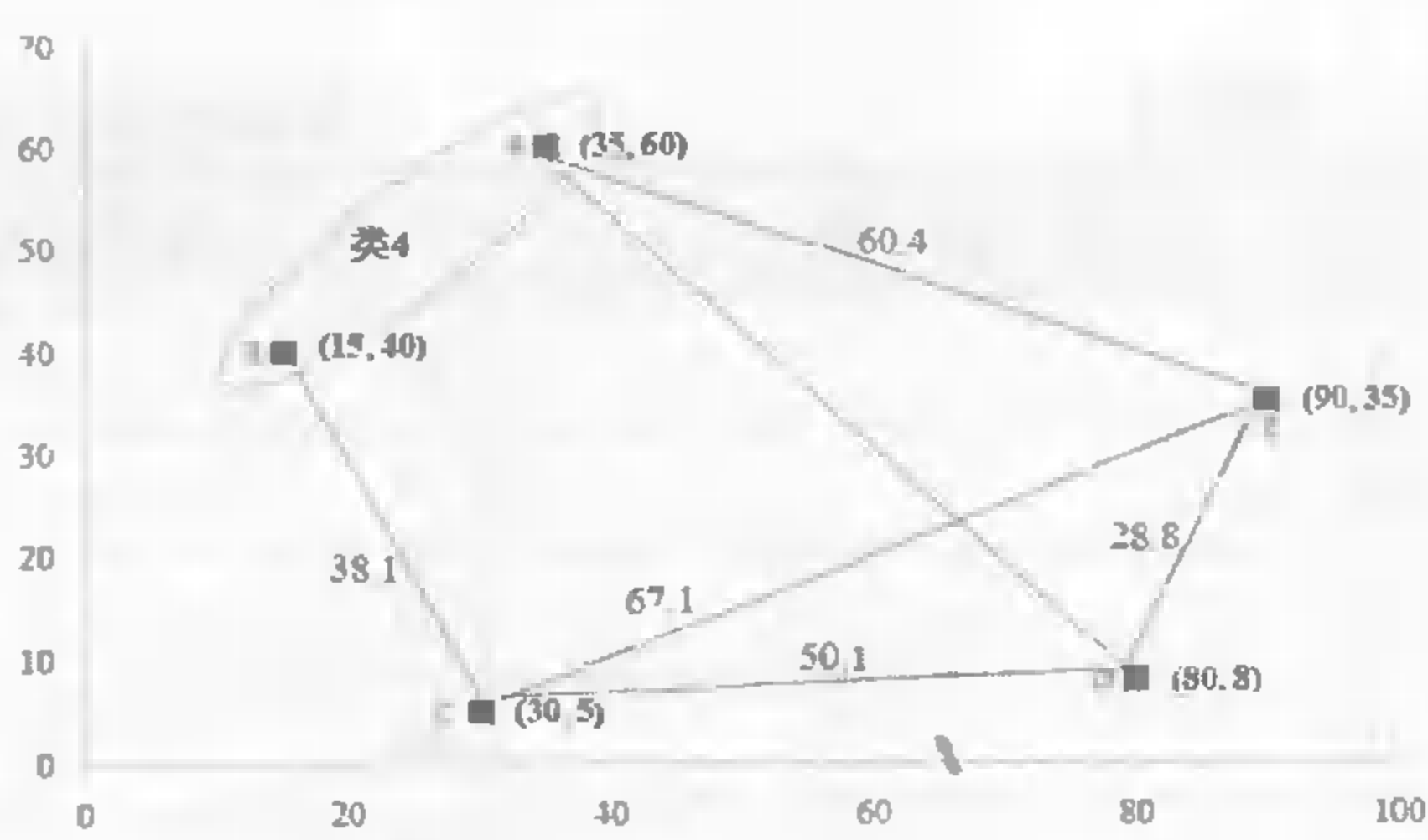


图 8-7 多指标的分类过程（三）

显然，在图 8-7 中剩下的所有距离中，最短距离就是 DE(28.8)。因此，把 D 和 E 聚为一类，记为类 3。把 D 和 E 合并之后，可得到图 8-8 所示的过程。

继续观测图 8-8 中的所有距离，可知 C 到类 4 的距离是最短的 (38.1)，于是又可以把 C 和类 4 聚为一类，记为类 2。按照上述的原则，同样可以得到图 8-9 所示的过程。

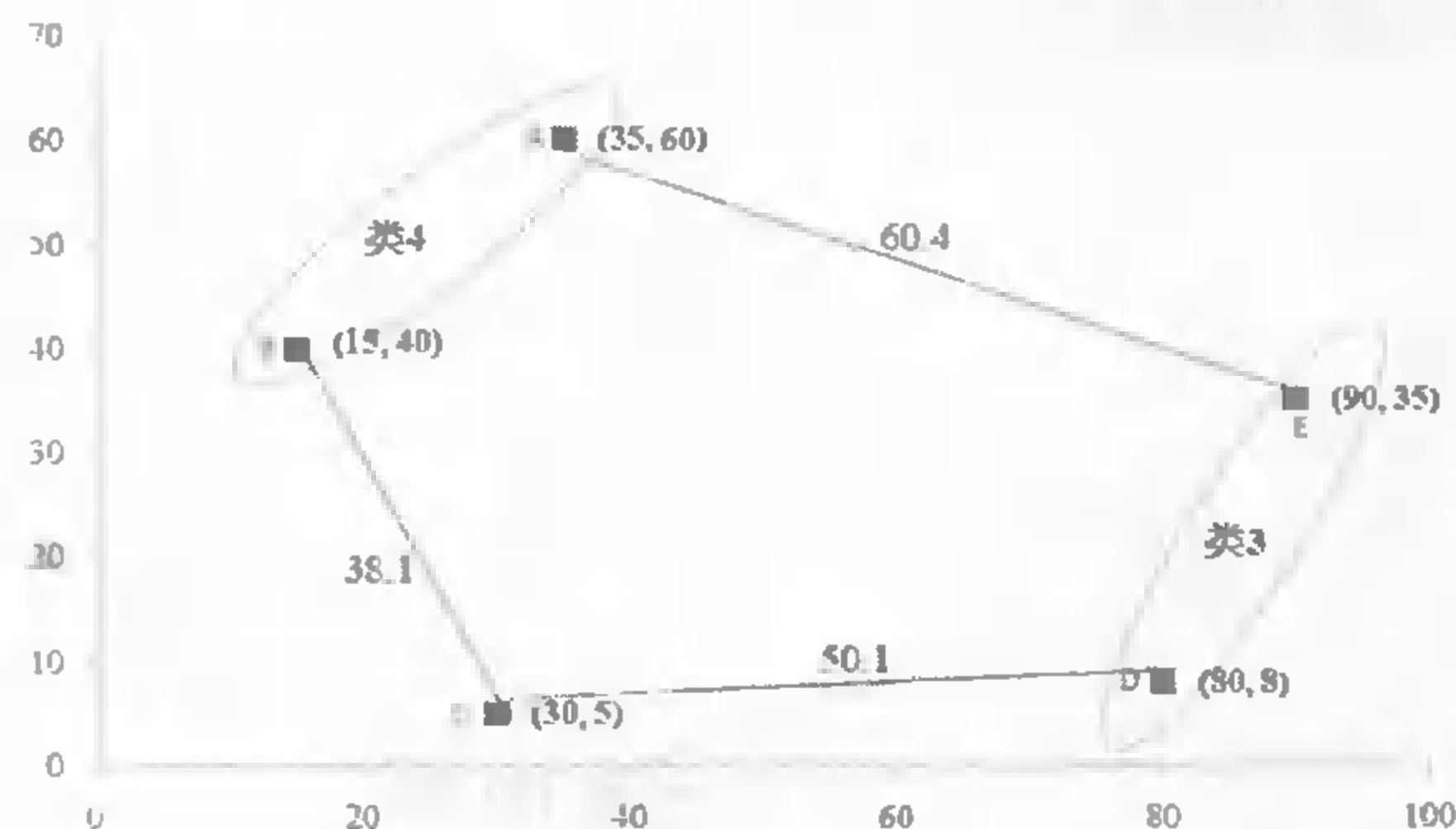


图 8-8 多指标的分类过程（四）

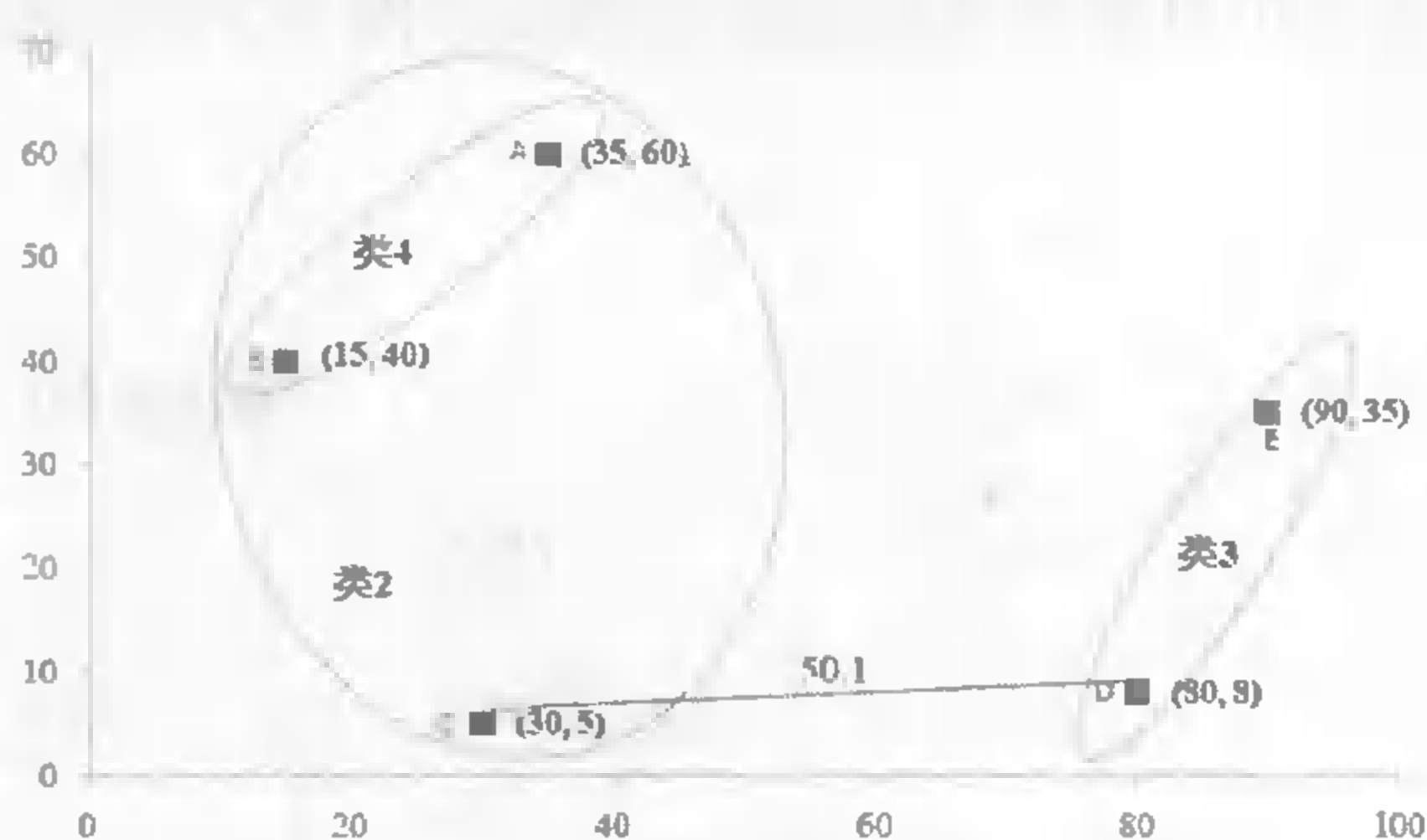


图 8-9 多指标的分类过程（五）

在图 8-9 中，只剩下类 2 与类 3 之间的距离 CD (50.1)。因此，把类 2 与类 3 合并成一大类。至此，所有的样本点都已经处于一定的类别当中，且所有的样本已经被归为一大类，系统聚类过程结束。整个分类过程如图 8-10 所示。

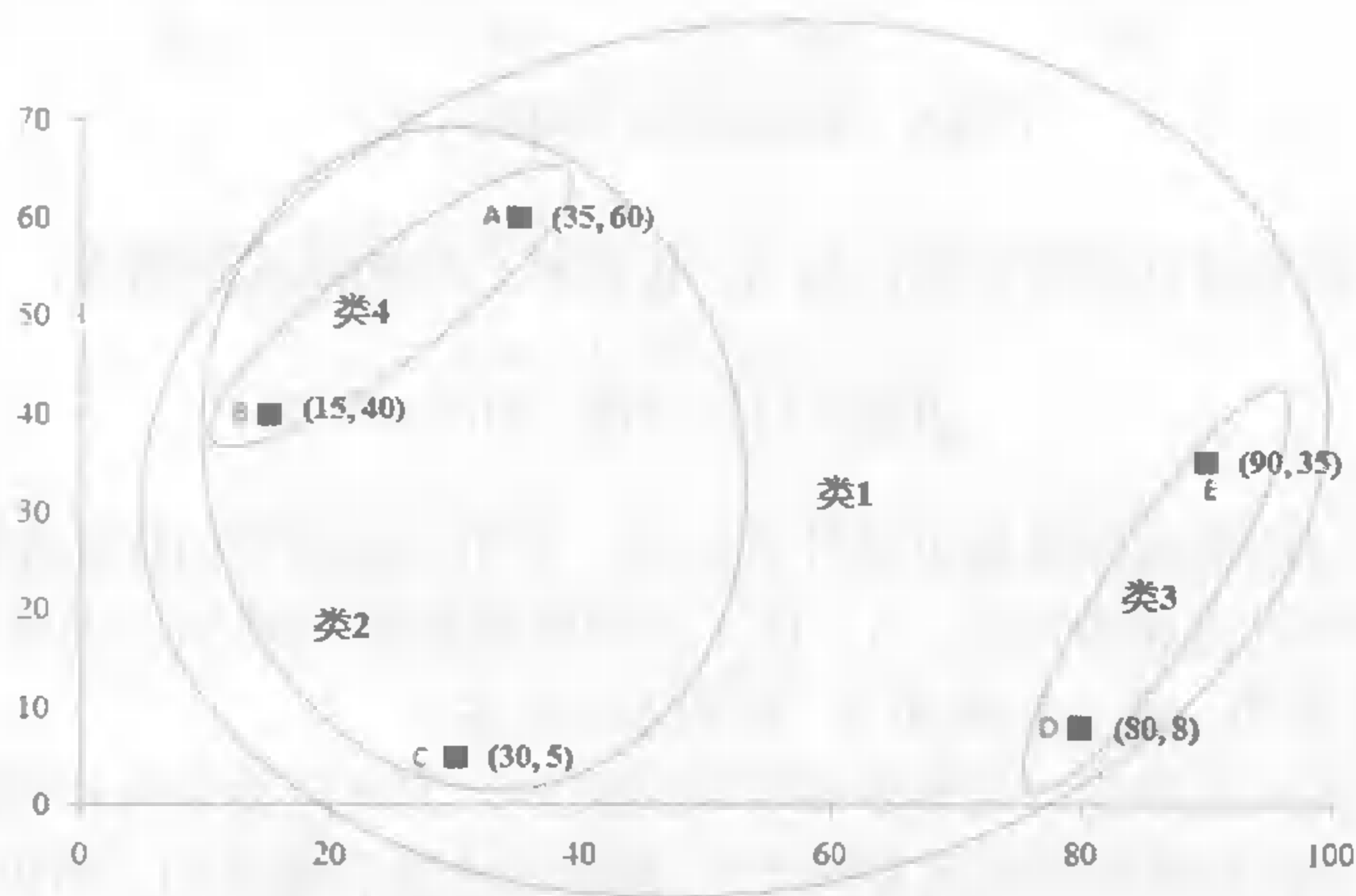


图 8-10 多指标的分类过程（六）

上述分类过程可以用表格的形式来表现。首先，把各样本间的距离列示在表 8-3 中。该表以对角线对称，为简单起见，省略掉重复数值。

表 8-3 样本点之间的距离（一）

	A	B	C	D	E
A					
B	28.3				
C	38.1	55.2			
D	72.3	68.7	50.1		
E	75.2	60.4	67.1	28.8	

观察表 8-3 中各样本点之间的距离，依据最短距离原则找出最短的距离并用灰色标注出来。显然，A、B 要聚成一类，于是可将表 8-3 处理成表 8-4 所示的形式。

表 8-4 样本点之间的距离（二）

	AB (4)		C	D	E
AB (4)					
C	38.1	55.2			
D	72.3	68.7	50.1		
E	75.2	60.4	67.1	28.8	

在 A、B 聚为类 4 后，其内部的距离失去意义，但其与外部的距离还有意义。按照最短距离原则，在这些外部的距离中，只有最短者被保留。因此，表 8-4 中不是最短的距离被删除，整理后如表 8-5 所示。

表 8-5 样本点之间的距离（三）

	AB (4)	C	D	E
AB (4)				
C	38.1			
D	68.7	50.1		
E	60.4	67.1	28.8	

在表 8-5 中列示的所有距离中，D 和 E 的距离最短（28.8），故把 D、E 聚为类 3。同理，删除不是最短的距离，整理后如表 8-6 所示。

表 8-6 样本点之间的距离（四）

	AB (4)	C	DE (3)
AB (4)			
C	38.1		
DE (3)	60.4	50.1	

观察表 8-6，可知 AB（4）与 C 的距离最短（38.1），因此把二者归为类 2，并删除非最短距离，得到表 8-7。

表 8-7 样本点之间的距离（五）

	ABC (2)	DE (3)
ABC (2)		
DE (3)		50.1

最后把 ABC（2）和 DE（3）聚成一类，如表 8-8 所示。

表 8-8 样本点之间的距离（六）

	ABCED (1)
ABCED (1)	

至此，系统聚类过程结束。

对于根据两个以上指标进行分类的过程，也类似于该过程。综合上述分类过程，下面把系统聚类的一般流程总结绘制成流程图，如图 8-11 所示。

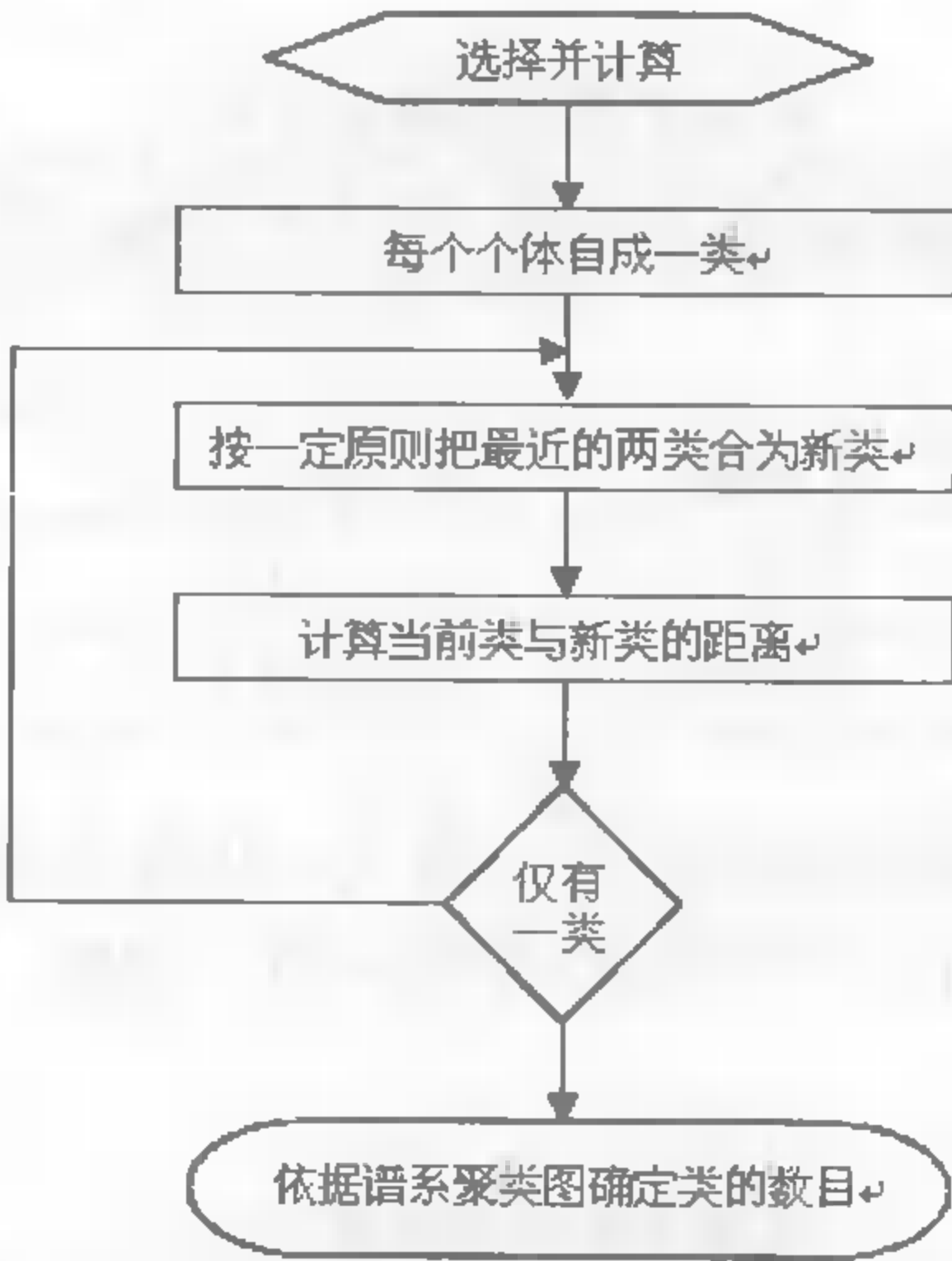


图 8-11 系统聚类流程图

8.2 聚类分析的步骤和过程

在 SAS 系统中，目前只能通过调用过程的方式进行聚类分析，主要通过 CLUSTER、FASTCLUS、VARCLUS 和 TREE4 个过程实现。

8.2.1 系统聚类

依据多个指标进行聚类分析时，可根据图 8-11 所示的流程图进行系统聚类分析。在 SAS 系统中，提供了 11 种系统聚类过程中确定类别与类别之间距离的方法：

- 最短距离法（SINgle linkage）
- 最长距离法（COMplete method）
- 中间距离法（MEDian method）
- 重心法（CENtroid method）

- 类平均法 (AVERage linkage)
- 离差平方和法 (WARD)
- 可变类平均法 (FLEXible-beta method)
- 可变法及 McQuitty 相似分析法 (MCQ)
- 最大似然谱系聚类 (EML)
- 密度估计法 (DEN)
- 两阶段密度估计法 (TWO)

在上述系统聚类距离确定方法中, 本节主要介绍前 6 种常用方法。在 SAS 编程语言中, 如需调用上述方法, 只需在程序中标注出上述对应方法英文单词的前 3 个字母即可。

### 1. 最短距离法 (SINGle linkage)

最短距离法又被称为“单连接聚类法”。如果  $G_p$  和  $G_q$  两类合并为新类  $G_n$ , 在最短距离法中, 新类  $G_n$  与其他任意类  $G_k$  之间的距离系数由下列公式决定。

$$D_{kn} = \text{Min}(D_{kp}, D_{kq})$$

即如果新类与其他类别之间存在多个距离, 则取这些距离当中最小者作为两类之间的距离。在进行分类的过程中, 以最小距离原则进行分类, 其具体并类过程与 6.1 节的分析过程相同。

### 2. 最长距离法 (COMplete method)

该方法也被称为“完全连接法”。如果  $G_p$  和  $G_q$  两类合并为新类  $G_n$ , 在最长距离法中, 新类  $G_n$  与其他任意类  $G_k$  之间的距离系数由下列公式决定:

$$D_{kn} = \text{Max}(D_{kp}, D_{kq})$$

即如果新类与其他类别之间存在多个距离, 则取这些距离当中最大者作为两类之间的距离。

在进行分类的过程中, 首先以最小距离原则把最近的两个样本合并为一个新类, 其余各样本自成一类。刚合并的新类与其他类别之间按最大距离原则确定之后, 再按照最小距离原则把类别之间距离最小的合并为一个新类, 以此类推, 直至把所有样本归为一类。

### 3. 中间距离法 (MEDian method)

如果  $G_p$  和  $G_q$  两类合并为新类  $G_n$ , 在中间距离法中, 新类  $G_n$  与其他任意类  $G_k$  之间的距离系数由下列公式决定。

$$D_{kn} = \frac{D_{kp} + D_{kq}}{2} - \frac{D_{pq}}{4}$$

如新类与其他类别之间存在多个距离, 则把按照上述公式计算的结果作为两类之间的距离。然后再按照最小距离原则把类别之间距离最小的两类合并为一类, 直至把所有样本归为一类。

### 4. 重心法 (CENTroid method)

在以上定义类与类之间的距离时, 没有考虑每一类中所包含的样品个数。重心法可以克

服这个缺点。

由重心法计算的  $G_p$  和  $G_q$  两类之间距离由下列公式决定。

$$D_{pq} = \|\bar{X}_p - \bar{X}_q\|^2$$

该距离即为两类重心（通常可用类内样本均值表示）之间的欧氏距离。当观测距离为欧氏距离  $d(x, y) = |x - y|^2$  时，新类  $G_n$  与其他任意类  $G_k$  之间的距离系数由下列公式决定。

$$D_{kn} = \frac{N_p D_{kp} + N_q D_{kq}}{N_n} - \frac{N_p N_q D_{pq}}{N_n^2}$$

其中， $N_i (i = p, q, n)$  表示各类的样本量。

即在并类过程中，以最小重心距离作为依据并类，直至把所有样本归为一类。

### 5. 类平均法 (AVERage linkage)

重心法虽有较好的代表性，但并未充分利用各个样本的信息，故有人提出用两类样品两两之间平方距离的平均值作为类之间的距离。如  $G_p$  和  $G_q$  两类，可以计算每类中每对样本点之间的平均距离。

$$D_{pq} = \frac{1}{N_p N_q} \sum_{i \in G_p} \sum_{j \in G_q} d(x_i, x_j)$$

若  $d(x, y) = |x - y|^2$ ，则新类  $G_n$  与其他任意类  $G_k$  之间的距离系数由递推公式决定。

$$D_{kn} = \frac{N_p D_{kp} + N_q D_{kq}}{N_n}$$

即在并类过程中，以类别样本点之间的平均距离作为依据并类，直至把所有样本归为一类。

### 6. 离差平方和法 (WARD)

离差平方和法又被称为“Ward 最小方差法”。它的思想来源于方差分析，即如果类分得恰当，同类内样品之间的离差平方和应较小，而类间的离差平方和应当较大。该法要求样品间距离必须采用欧氏距离。

离差平方和法定义类间的平方距离为： $D_{pq}^2 = S_n^2 - S_p^2 - S_q^2$ 。其中， $S_n^2$  是类  $G_p$  和  $G_q$  合并成的  $G_n$  类的类内离差平方和。

当观测距离  $d(x, y) = \|x - y\|^2 / 2$  时，则新类  $G_n$  与其他任意类  $G_k$  之间的距离由下列递推公式决定。

$$D_{kn} = \frac{(N_k + N_p) D_{kp} + (N_k + N_q) D_{kq} - N_k D_{pq}}{N_k + N_n}$$

当采用离差平方和法进行分类时，先让每个样品自成一类，然后并类。每并一类，离差平方和都要增大，选择使其增加最小的两类合并，直到所有的样品归为一类为止。

那么，这么多种方法都可以对样本数据进行聚类分析，究竟采用哪一种方法最好呢？这

个问题至今没有一个明确的答案。目前可以参考经验分析以及研究对象的特征和研究要求，进行经验判断。如给定距离的阈值或是通过计算相应的统计量进行判定，再如 Demirmen (1972) 提出了一些在决定聚类方法取舍时应遵循的原则。

- 任何类必须在邻近的各类中是突出的，即各类重心（常用平均数衡量）之间应该有最大的距离。
- 在确定的类中，各类所包含的元素都不宜过多。
- 分类数目应符合实际。
- 当用许多方法进行分类时，应选出现次数最多的那种分类结果。

实际应用中的通常做法是多用几种不同的方法进行聚类，然后根据现有理论和研究要求挑选出合适的分类类别。



例 8-3

某网站键鼠频道为广大职业玩家及游戏爱好者策划了一次全面的游戏鼠标横向测试，通过专家和消费者打分的形式，收集到了 13 款游戏鼠标的重要参数，即外观及手感、芯片及微动、功能及驱动、兼容性、游戏性等数据，如表 8-9 所示（详细数据见 Mouse\_Cluster.sas7bdat）。要求以这些指标为依据对所收集到的样本进行聚类分析。

表 8-9 13 款游戏鼠标横向评测数据

品牌型号 (Brand)	外观及手感 (Touch)	芯片及微动 (Chips)	功能及驱动 (Driver)	兼容性 (Compatibility)	游戏性 (Game)
Razer 3G	7.5	17.5	7	8	8
Razer 巨腹蛇	7.5	19.5	7	7	9
微软 SideWinder	8.5	18	8.5	8	9.5
罗技 G9	9	18.5	8.5	8	9.5
美心 点击王	7	14	6.5	7	7.5
苹果新概念 MG09V5U	7	16	6.5	7.5	8
双飞燕 XL-750FS	7.5	17	8	7.5	8
微软 Habu	8	17.5	8.5	7.5	8.5
明基 幻影熊	7	16.5	6	8	7
罗技 新版 MX518	7.5	17	7.5	8.5	8
多彩 T2	8	16	6.5	7	7
优派 黑甲鼠	7	15.5	6	8	7
多彩 DLM-615LU	7	17.0	8	7	7

上表数据源有中关村在线 (<http://mouse.zol.com.cn/71/717641.html>)，仅用于分析

本例要求根据鼠标外观及手感、芯片及微动性能等 5 个指标综合对样本进行聚类分析，因此可考虑采用系统聚类法进行聚类。

在 SAS 系统中，主要通过调用 CLUSTER 过程进行系统聚类。CLUSTER 过程的主要语法如下。

```
PROC CLUSTER METHOD = 聚类方法名称 < 选项 >;  
  BY 变量;
```

COPY 变量;  
FREQ 变量;  
ID 变量;  
RMSSTD 变量;  
VAR 变量;

其中 BY、FREQ、VAR 语句的用法与前面所述一致。

COPY 语句主要用于从所有的原始数据集中复制指定的变量并将其存入 CLUSTER 语句所指定的输出数据集中。

ID 语句主要用于在聚类分析的谱系聚类图中对观测样本进行标注。该语句在样本量比较多的情况下比较有用，用户根据标注后的样本便可快速查看样本所属的类别。如果 ID 语句省略，则系统在谱系聚类图中会自动以样本的编号进行标注。

RMSSTD 语句主要用于计算每个类别的标准误差均方根，该计算结果被存储在该语句所指定的变量中。

而聚类方法及聚类结果的控制主要由 CLUSTER 语句的选项来进行调整，CLUSTER 语句的关键字非常多，本小节主要介绍比较常用的关键字。

- DATA =：指定用于聚类分析的数据集。
- OUTTREE =：指定分析结果输出数据集。
- SIMPLE：显示均值、标准差等简单统计量。
- METHOD =：指定聚类分析中类别之间距离的确定方法。

在 CLUSTER 语句的“METHOD =”关键字下，可以指定 11 种系统聚类方法。

- AVERAGE：类平均法
- CENTROID：重心法
- COMPLETE：最长距离法
- DENSITY：密度估计法
- EML：最大似然谱系聚类
- FLEXIBLE：可变类平均法
- MCQUITTY：可变法及 McQuitty 相似分析法
- MEDIAN：中间距离法
- SINGLE：最短距离法
- TWOSTAGE：两阶段密度估计法
- WARD：离差平方和法

上述方法在 SAS 系统中可以用前 3 个字母进行缩写。如“Method = Average”等价于“Method=ave”。

在实际分析过程中，可任意指定其中的方法进行聚类。本例利用 CLUSTER 过程进行聚类分析的程序如下。

```
proc cluster data=Sasuser.Mouse_Cluster
    method=ave outtree=Mouse_Cluster_Out;    /*调用 CLUSTER 过程，指定用 Sasuser.Mouse_Cluster
数据集进行分析，类别之间距离确定方法采用类平均法，并且把结果输出至临时数据集 Mouse_Cluster_Out 当中*/
    var Touch Chips Driver Compatibility Game; /*指定用于聚类的变量*/
```

```
id Brand;                                /*指定标注变量*/
run;
```

运行之后得到图 8-12 所示的基本信息和图 8-13 所示的聚类过程。

The CLUSTER Procedure				
Average Linkage Cluster Analysis				
Eigenvalues of the Covariance Matrix				
	Eigenvalue	Difference	Proportion	Cumulative
1	3.14769719	2.58016059	0.7280	0.7280
2	0.56753660	0.28295506	0.1313	0.8593
3	0.28458154	0.07462236	0.0658	0.9251
4	0.20995918	0.09601575	0.0486	0.9736
5	0.11394343		0.0264	1.0000
Root-Mean-Square Total-Sample Standard Deviation = 0.929916				
Root-Mean-Square Distance Between Observations = 2.940652				

图 8-12 系统聚类的基本信息

本例首先选择类平均法（Average Linkage）进行聚类分析。图 8-12 显示了利用该方法进行分析的协方差矩阵基本信息。由图 8-12 显示的信息可以看到聚类的特征根及其对应的贡献率，也可以得知总体样本标准差均方根（Root-Mean-Square Total-Sample Standard Deviation）为 0.929916，说明所进行分析的数据内部的变异性较小，这是因为该次评测数据的计量单位一致、技术逐渐成熟导致产品同质性日趋明显造成的；样本观测值之间距离的均方根（Root-Mean-Square Distance Between Observations）为 2.940652，说明样本之间距离较近。

Cluster History				
NCL	-----Clusters Joined-----	FREQ	Norm RMS Dist	Tie
12	微软 SideWinder 罗技 G9	2	0.2405	
11	Razer 3G 罗技 新版MX518	2	0.2945	
10	双飞燕 XL-750FS 微软 Habu	2	0.3401	T
9	明基 幻影熊 优派 黑甲鼠	2	0.3401	
8	苹果新概念 MG09V5U CL9	3	0.4499	
7	CL11 CL10	4	0.4957	
6	CL8 多彩 T2	4	0.5286	
5	CL7 多彩 DLM-615LU	5	0.5703	
4	美心 点击王 CL6	5	0.7838	
3	CL5 CL12	7	0.8703	
2	CL3 Razer 巨腹蛇	8	0.9618	
1	CL2 CL4	13	1.1999	

图 8-13 系统聚类的并类过程

图 8-13 完整地显示了样本之间的并类过程。“NCL”表示聚类编号，即并类次序，并类从上至下，最后所有样本归为一类；“Clusters Joined”表示分类的过程，从中可以非常清晰地看出样本和类别之间的并类过程；“FREQ”表示每一次并类时该类别中所包含的样本量；“Norm RMS Dist”表示距离的均方根；最后一列“Tie”则标注了距离均方根相等的数值，当并类过程中距离均方根相等时，系统会自动在该距离的第一个数值上标注“T”字样以提醒用户注意。

在本例中，首先按照距离的大小把“微软 SideWinder”和“罗技 G9”两个鼠标归为第一类；然后重新计算新类与其余样本点之间的平均距离，根据最小的平均距离，把“Razer 3G”和“罗技新版 MX518”两个鼠标归为第二类。以此类推，直至类“CL2”和“CL4”归为一大类为止。

其中的类“CL2”包含的样本为聚类编号“NCL”为“2”的类别对应的样本，具体包括“CL3”和“Razer 巨腹蛇”。把“CL3”根据聚类标号对应至具体的样本，最终可得到“CL2”所包括的样本为“Razer 巨腹蛇”、“多彩 DLM-615LU”、“Razer 3G”、“罗技新版 MX518”、“双飞燕 XL-750FS”、“微软 Habu”、“微软 SideWinder”和“罗技 G9”8 个鼠标。同理，可以找出“CL4”所包含的具体样本。

根据以上各个聚类编号，便可依据一定的层次找出想要的分类结果。但在通常情况下，系统聚类往往是通过谱系聚类图来直观描述分类过程的。

在 SAS 系统中，CLUSTER 过程负责具体分析的具体测算，系统聚类谱系图则可通过 TREE 过程完成。TREE 过程的具体语法如下。

```
PROC TREE < 选项>;
  NAME 变量;
  HEIGHT 变量;
  PARENT 变量;
  BY 变量;
  COPY 变量;
  FREQ 变量;
  ID 变量;
```

TREE 过程中的 NAME 语句用于指定谱系聚类图中类别的名称，HEIGHT 语句用于指定每一类别在谱系图中的高度，PARENT 用于指定每一类别的归属类别名称，其余语句与之前介绍的功能一致。

TREE 语句的选项非常多，这里主要介绍几个常用选项。

- DATA = : 指定用于绘制谱系图的数据集，该数据集应当是 CLUSTER 过程由 OUTTREE 关键字指定的数据集。
- OUT = : 指定 TREE 过程分析结果的输出数据集。
- STANDARD: 把指定变量进行均值为 0、方差为 1 的标准化。
- NCLUSTERS = : 指定由 OUT 设定的输出数据集中的类别个数。
- HORIZONTAL: 指定系统绘制水平的谱系聚类图，缺省则表示绘制垂直的谱系图。

TREE 过程通常与 CLUSTER 过程搭配使用，也可与 VARCLUS 过程（详见 8.3.2 节）搭配使用。在调用过程中，应当将其放在 CLUSTER 或 VARCLUS 过程之后。

如果在 CLUSTER 过程的选项中加入了“OUTTREE = ”关键字以指定聚类结果输出数据集，则在引用 TREE 过程时非常简单，可以省略 TREE 过程的一切语句。如本例需绘制谱系图，具体程序如下。

```
proc cluster data=Sasuser.Mouse_Cluster
  method=ave outtree=Mouse_Cluster_Out;
  var Touch Chips Driver Compatibility Game;
  id Brand;
run;
proc tree horizontal;
run;
```

此处的 proc tree 过程语句等价于以下程序。

```
proc tree data=Mouse_Cluster_Out horizontal;
  name _name_;
  parent _parent_;
  height _height_;
  id brand;
run;
```

因为在 CLUSTER 过程中用 OUTTREE 指定了聚类分析的数据集，所以系统会自动根据聚类分析的过程和结果在数据集中存储默认的、用于绘制谱系图的变量。有兴趣的读者可以打开 OUTTREE 生成的结果输出数据集进行查看。

运行程序后，除可以得到图 8-12 和图 8-13 所示的结果之外，还可以得到水平放置的系统聚类谱系图，如图 8-14 所示。

图 8-14 中的横轴表示用于判定类别的距离或相似系数。本例使用的是类平均法，故横轴表示类别之间的平均距离，而纵轴表示用“Brand”品牌型号变量标注的每个具体样本。

那么究竟如何解读谱系聚类图呢？只需在谱系图中任意地点画一条竖直的直线，则该直线与图中的横线有多少个交点，就可以把样本分为多少类。分类的依据便是这条竖直的直线所对应的横轴距离，与每个交点相连的样本同属于一类。

如在图 8-14 上画了一条竖直的直线，该直线与谱系聚类图有 3 个交点，即可把所有样本分为 3 类。与从上往下的第 1 个交点相连的样本是“Razer 3G”、“罗技新版 MX518”、“双飞燕 XL-750FS”、“微软 Habu”、“多彩 DLM-615LU”、“微软 SideWinder”和“罗技 G9”，这些鼠标构成了一类；而与第 2 个交点相连的只有“Razer 巨腹蛇”单独 1 个样本，则其自成一类；剩下的“美心 点击王”、“苹果新概念 MGO9V5U”、“明基 幻影熊”、“优派 黑甲鼠”和“多彩 T2”与第三个交点相连，形成一类。

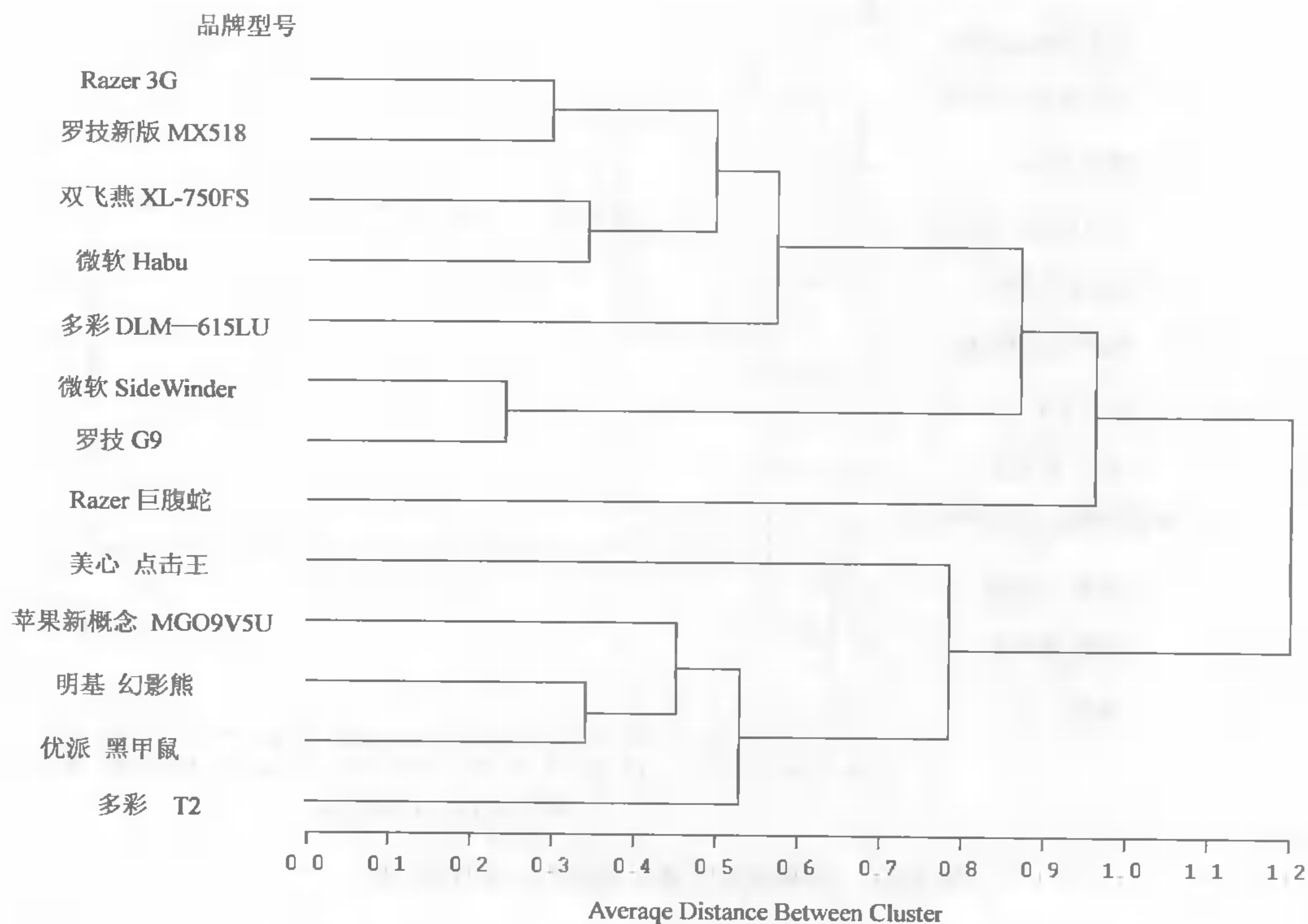


图 8-14 系统聚类的谱系聚类图 (AVE 法)

如果把所画的这条竖直的直线向右平移，可能会与横线产生两个交点，则可根据上述方法分别找出两个交点相连的两类样本。

在本例中，使用类平均法进行分类，无论是分为 3 类还是 2 类，产品类别的特征都不是很明显，而且分为 3 类时还会出现单个样本自成一类的结果。因此，为了更好地对产品进行区分，继续采用其他聚类方法进行分析，作为最终分类结果的参考。如使用 Ward 法进行聚类并绘制谱系图的程序如下。

```
proc cluster data=Sasuser.Mouse_Cluster
  method=ward outtree=Mouse_Cluster_Out;
  var Touch Chips Driver Compatibility Game;
  id Brand;
run;
proc tree horizontal;
run;
```

运行程序之后，可以得到图 8-15 所示的谱系图。

依据图 8-15 所示谱系图，从上至下，可以把所有的鼠标分为 3 类，即把“Razer 3G”、“罗技 新版 MX518”、“双飞燕 XL-750FS”、“微软 Habu”、“多彩 DLM-615LU” 归为第 1 类，把“Razer 巨蝮蛇”、“微软 SideWinder”和“罗技 G9” 归为第 2 类，把“美心 点击王”、“苹果 新概念 MG09V5U”、“明基 幻影熊”、“优派 黑甲鼠”和“多彩 T2” 归为第 3 类。

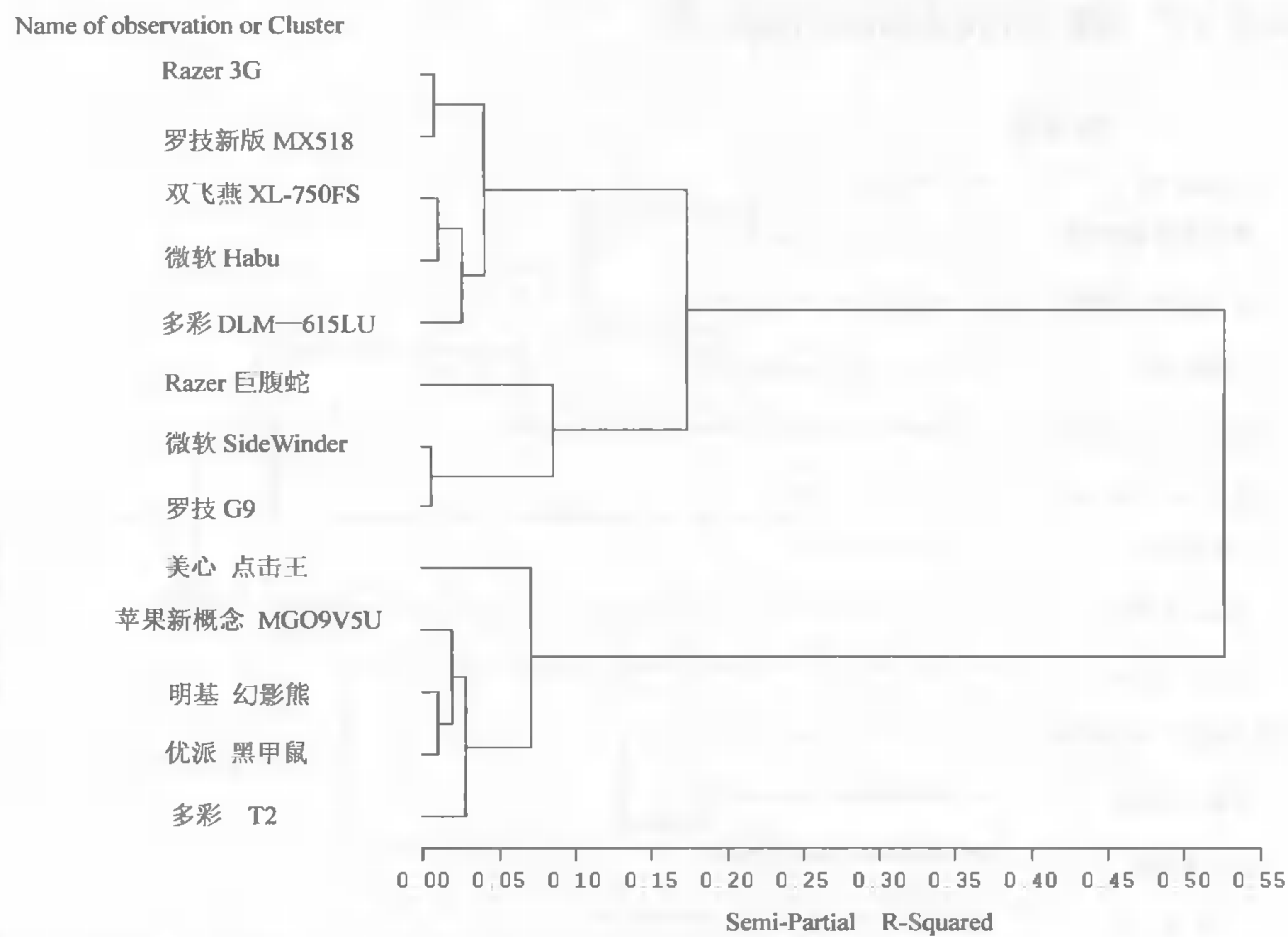


图 8-15 系统聚类的谱系聚类图（WARD 法）

通过对原始数据进行进一步分析可知，罗技 07 年新款游戏鼠标 G9 以总分 53.5 分获得第 1，微软硬件首款专业游戏鼠标 SideWinder 以总分 52.5 排名第 2，Razer 巨蝮蛇以总分 50 分位列第 3。因此，此 3 款鼠标都应该属于性能相对卓著的产品，性能和价格相类似，可以归为一类，即图 8-15 所示谱系图中的第 2 类；通过对第 1 类鼠标组成样本的特征进行分析，发现其属于性能较好、价格适中的中端产品；而第 3 类产品性能一般且价格较低，可以将其定义为中低端产品。

8.2.2 快速聚类

在系统聚类中，研究者事先并不知道对样本要聚成多少个类别，一般通过图 8-14 或图

8-15 所示的谱系聚类图进行类别的划分。在有些情况下,研究者事先知道将研究对象分为几类,即已知类别的个数  $k$ ,只是不知道这些类别当中的具体样本。这时可以考虑采用本小节将介绍的快速聚类方法进行聚类。

快速聚类一般用于大样本情况下的样品聚类。SAS 系统中使用的快速聚类方法以 Anderberg (1973) 提出最近中心归类法为基本算法,主要按照以下原则进行聚类。

**STEP 1** 选择  $k$  个观测值组成初始类别并作为聚类种子。

**STEP 2** 找出聚类种子的中心。

**STEP 3** 把每一个观测值根据最小欧氏距离(观测值与聚类种子中心之间的距离)原则归入各类,构成暂时的类别。

**STEP 4** 计算每个暂时的类中各个变量的均值,以此作为新类别的中心。

**STEP 5** 再次把每一个观测值根据最小欧氏距离原则归入各类,构成新的、暂时的类别。

**STEP 6** 重复第 3 步至第 5 步,当中心的迭代标准达到要求时,聚类过程结束。

上述过程也可被称为 K-Means (K-均值) 聚类,在 SAS 系统中,主要通过 FASTCLUS 过程来实现。聚类种子可凭经验选择(通常以数据集的形式储存,在 FASTCLUS 过程中应当通过 SEED 关键字指定),也可由 FASTCLUS 过程自动选择。FASTCLUS 过程主要输出聚类的结果,不会输出谱系聚类图。

快速聚类 FASTCLUS 过程的主要语法如下。

```
PROC FASTCLUS <选项>;
  VAR 变量;
  ID 变量;
  FREQ 变量;
  WEIGHT 变量;
  BY 变量;
```

其中,VAR、ID、FREQ、WEIGHT、BY 语句与前面章节介绍的功能一致。

FASTCLUS 的主要常用选项如下。

- DATA = : 指定用于快速聚类分析的数据集。
- SEED = : 指定用于快速聚类分析的聚类种子数据集。
- OUT = : 指定用于储存分析过程和结果的输出数据集。
- MAXCLUSTERS =  $k$ : 指定系统自动选择  $k$  个观测值组成初始类别并作为聚类种子,(亦即指定要分的类别数)该语句可以简写为“MAXC =  $k$ ”。
- RADIUS =  $t$ : 指定中心的最短距离,其默认值为 0。
- CONVERGE =  $n$ : 指定迭代收敛标准值,其默认值为 0.02。当中心最大距离小于或等于  $n$  倍初始中心时,迭代终止。
- MAXITER =  $n$ : 指定最大迭代次数,其默认值为 1。
- LIST: 显示所有的观测值、ID 变量标注值、每个类别的聚类编号、观测值与最终聚类种子之间的距离。

在使用 FASTCLUS 过程时,系统不会自动对变量进行标准化(详见 1.4.4 小节)。



如果各变量的量纲不一致,则应当先对各变量进行标准化,用标准化之后的结果进行快速聚类。

此外，为了清晰明显地区别各个样本所属的类别，通常在 FASTCLUS 过程的选项中使用 LIST 选项关键字，用以列示出 ID 变量标注的样本及其对应的类别。

这里为了与 8.2.1 小节中的系统聚类进行对比分析，仍然以例 8-3 为例，假定已知样本可以分为 3 类，在小样本的情况下进行快速聚类。具体程序如下。

```
proc fastclus data=Sasuser.Mouse_Cluster maxclusters=3 list out=Mouse_Fast_Out; /*调用 fastclus 快速聚类过程，指定最大分类数为 3，并把聚类结果存储在临时数据集 Mouse_Fast_Out 中*/
var Touch Chips Driver Compatibility Game; /*指定用于分析的变量*/
id Brand; /*指定用于标注样本的变量*/
run;
```

运行程序后，SAS 系统会自动给出一系列的输出结果。这里为了突出实用性，主要介绍最终聚类结果，其余结果的详细分析请参考相关资料。在 SAS 输出结果索引窗口“Results”中，展开“Fastclus: The SAS System”索引项，然后双击“Cluster Listing”，便可在“Output”窗口中看到快速聚类的最终聚类的结果，如图 8-16 所示。

Cluster Listing			
Obs	Brand	Cluster	Distance from Seed
1	Razer 3G	1	0.8452
2	Razer 巨腹蛇	3	1.7159
3	微软 SideWinder	3	0.9280
4	罗技 G9	3	0.9280
5	美心 点击王	2	1.2802
6	苹果新概念 MG09V5U	2	0.9860
7	双飞燕 XL-750FS	1	0.7559
8	微软 Habu	3	1.5635
9	明基 幻影熊	2	1.5456
10	罗技 新版MX518	1	0.9449
11	多彩 T2	2	1.4907
12	优派 黑甲鼠	2	0.8498
13	多彩 DLM-615LU	1	1.1180
Criterion Based on Final Seeds = 0.5333			

图 8-16 快速聚类的最终聚类结果

图 8-16 中显示了依据样本与聚类种子中心之间的距离进行迭代之后的最终聚类结果，“Cluster”列标注了各个样本应当归属的类别，如“Razer 巨腹蛇”、“微软 SideWinder”、“罗技 G9”和“微软 Habu”4 个鼠标聚为一类（“Cluster”标注为 3）。图 8-16 所示的分类结果与图 8-15 所示的分类结果大体一致。

本例在调用 FASTCLUS 过程中使用了选项中的“OUT=”关键字，生成一个名为 Mouse\_Fast\_Out 的临时数据集。图 8-16 所示结果中的最后两列（即列“Cluster”和列“Distance from Seed”）会作为变量被储存在该输出数据集中，便于用户随时进行查看。

8.2.3 变量聚类

前面章节主要介绍的是对样本进行聚类，而变量的聚类在 SAS 系统中主要通过 VARCLUS 过程来实现。VARCLUS 过程可将数值型的变量进行分离或分层聚类。在此过程中，每一类是该类别之内所有变量的一个线性组合，该线性组合既可以是每个类别的第一主成分，也可以是每个类别的重心成分。其中提取每个类别的第一主成分是 SAS 系统的默认方法，第一主成分具有最大的方差贡献率。VARCLUS 过程的主要语法如下。

```
PROC VARCLUS < 选项 >;
  VAR 变量;
  SEED 变量;
  PARTIAL 变量;
  WEIGHT 变量;
  FREQ 变量;
  BY 变量;
```

其中 SEED 语句用于指定作为初始化类别的聚类种子, PARTIAL 语句用于指定使用偏相关矩阵进行聚类分析。其余语句与前面章节介绍的功能相同。VARCLUS 常用的主要选项如下。

- CENTROID: 使用重心成分进行聚类。当使用未加权标准化数据或非标准化数据的协方差矩阵进行分析时, 应该使用该方法。该关键字缺省表示使用主成分法聚类。
- HIERARCHY: 分析聚类谱系结构。
- DATA =: 指定进行分析的原始数据。
- OUTSTAT =: 指定存储均值、相关系数、聚类得分等统计量的输出数据集。
- OUTTREE =: 指定存储聚类分析结构的输出数据集; 该选项生成的数据集往往作为 TREE 过程的分析数据。
- MAXCLUSTERS =: 指定最大的聚类个数。
- MAXEIGEN =: 指定聚类最大的第二个特征值; 注意不能用于重心法(CENTROID)。
- MINCLUSTERS =: 指定最小的聚类个数。
- PROPORTION =: 指定成分的方差贡献率(用比率表示), 默认值为 0。当选项中有 CENTROID 关键字时, 其默认值为 0.75。
- PERCENT =: 指定成分的方差贡献率(用百分比表示)。

在本小节中, 仍然使用例 8-3 的数据进行变量聚类, 用以考察在评测过程中, 设计的指标可以从哪几个主要方面来反映鼠标的综合性能。例 8-3 变量聚类的程序如下。

```
proc varclus data=Sasuser.Mouse_Cluster;
  var Touch Chips Driver Compatibility Game;
run;
proc varclus hierarchy data=Sasuser.Mouse_Cluster;
  var Touch Chips Driver Compatibility Game;
run;
proc varclus centroid data=Sasuser.Mouse_Cluster outtree=Mouse_Var_Out;
  var Touch Chips Driver Compatibility Game;
run;
proc tree data=Mouse_Var_Out horizontal;
run;
```

上述程序的第 1 个过程没有任何选项, 系统默认选择主成分法聚类; 第 2 个过程加入了“hierarchy”关键字, 要求保证分析过程中不同水平的谱系结构; 第 3 个过程的选项中加入了“centroid”和“outtree”关键字, 表示使用重心法聚类, 并且生成用于绘制谱系聚类图的输出文件。最后的 TREE 过程用于绘制由重心法聚类的谱系图。

运行程序之后, 可以得到非常多的结果, 下面依据过程的顺序一一进行解释。

首先是由系统默认的主成分聚类得到的结果，如图 8-17 所示。

图 8-17 表明解释方差为 3.066239，占总方差的 0.6132(3.066239/5)，系统自动依据其隐含的临界值做出是否应当进行下一步分类的决定。在本例中，图 8-17 的最后一行自动提示已经达到临界值，停止继续分类。因此，本例数据使用主成分聚类法可以把所有的指标或变量分为一大类。但是这种分类结果并不是想要的结果，对变量分类的目的是想了解究竟能从哪几个大的方面综合考察所有反映鼠标各种详细性能的指标。

第 2 个过程加入了“hierarchy”关键字，要求保证分析过程中不同水平的谱系结构的分类过程与图 8-17 所示结果一致。实际上，在 SAS 系统中，该关键字对于分类结果无实质性影响。

Oblique Principal Component Cluster Analysis					
Observations	13	Proportion	0		
Variables	5	Maxeigen	1		
Clustering algorithm converged.					
Cluster Summary for 1 Cluster					
Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained	Second Eigenvalue
1	5	5	3.066239	0.6132	0.9615
Total variation explained = 3.066239 Proportion = 0.6132					
No cluster meets the criterion for splitting.					

图 8-17 主成分变量聚类结果

本例的第 3 个过程使用重心聚类，并用 OUTTREE 语句生成分析结果，且用 TREE 过程绘制谱系聚类图。图 8-18 展示了聚类的过程。

Oblique Centroid Component Cluster Analysis				
Observations	13	Proportion	0.75	
Variables	5	Maxeigen	0	
Clustering algorithm converged.				
Cluster Summary for 1 Cluster				
Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained
1	5	5	2.868328	0.5737
Total variation explained = 2.868328 Proportion = 0.5737				
Cluster 1 will be split because it has the smallest proportion of variation explained, 0.573666, which is less than the PROPORTION=0.75 value.				
Clustering algorithm converged.				
Cluster Summary for 2 Clusters				
Cluster	Members	Cluster Variation	Variation Explained	Proportion Explained
1	4	4	3.013763	0.7534
2	1	1	1	1.0000
Total variation explained = 4.013763 Proportion = 0.8028				

图 8-18 重心法变量聚类的过程

在本例的聚类过程中，由于采用重心法进行变量聚类，系统默认的解释方差比例（Proportion）为 0.75。在把所有变量归集在一起的第一个分类中，其方差解释比例为 0.5737，小于 0.75，故系统自动提示该类的方差解释比例未达到 0.75 的阈值，第一个分类将被拆分，进行进一步分类。进一步的分类结果如图 8-18 所示的下半部分，经过拆分，把所有变量分成了两类，第一类（Cluster 1）含有 4 个变量，第二类（Cluster 2）含有 1 个变量，两类总的方差贡献率为 0.8028，大于 0.75，故分类过程终止，最终的变量聚类结果便是将所有变量划分成了两类。

此外，变量聚类过程还可以输出每个变量与类别之间的相关系数平方，以及用于类别预测的标准化回归系数等分析结果，这里不予赘述。

为了更好地展示上述利用重心法变量聚类的过程，这里使用 TREE 过程绘制谱系聚类图，即本例程序的第 4 个过程。运行程序后得到图 8-19 所示的结果：

在图 8-19 所示的谱系图中，横轴表示聚类类别的个数。在本例中，依据图 8-18 所示的分类结果，可以分为两类，即变量“Touch”、“Chips”、“Drivers”和“Game”分为一类，变量“Compatibility”自成一类。

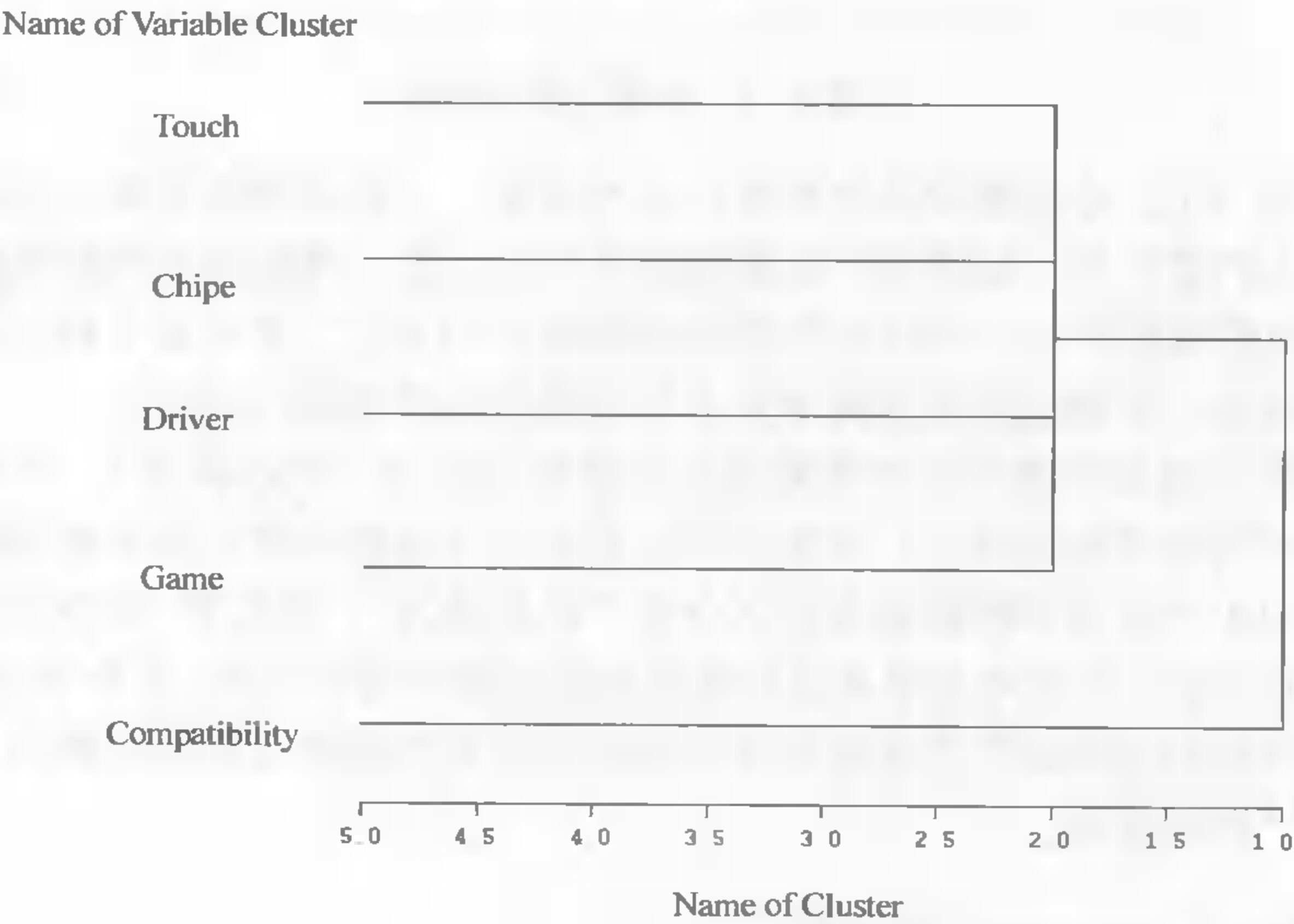


图 8-19 重心法变量聚类谱系图

在聚类分析中，除了用纯粹的数据结构关系考察分类的正确性之外，还可以根据实际问题的需要及变量本身的含义来考察分类的合理性。本例的第一个类别含有“Touch”等 4 个变量，这 4 个变量均从外观手感、芯片及微动、功能及驱动、游戏性能等方面反映了鼠标的综合性能，而且这些性能均可以通过用户的客观感受和专业手段来进行评价，它们是一款鼠标的硬性指标；而第二类的“Compatibility”变量反映的是鼠标的兼容性，主要通过主观感受来评价，而且容易受到系统和软件环境的影响，可以认为它是鼠标的软性指标。因此，通过变量聚类分析的过程和结果，本例实际上是从硬性、软性两大类指标即两个侧面来评判各款鼠标的综合性能。

### 8.3 判别分析的基本思想

在现实生活中，人们不光要对现有事物分门别类，有时还需依据特定特征对现有样本进行分类。在给定现有分类的条件下，要求把新收集的样本依据既定的特征归入现有的某一个类别当中，该过程被称为“归类”。如有  $G_1$  和  $G_2$  两个类别，对于新加入的样本 A，考虑把 A 归入对应类别中，如图 8-20 所示。由于归类过程涉及到对新样本与现有样本特征的判定与识别问题，因此该过程也可被称为“判别分析”。

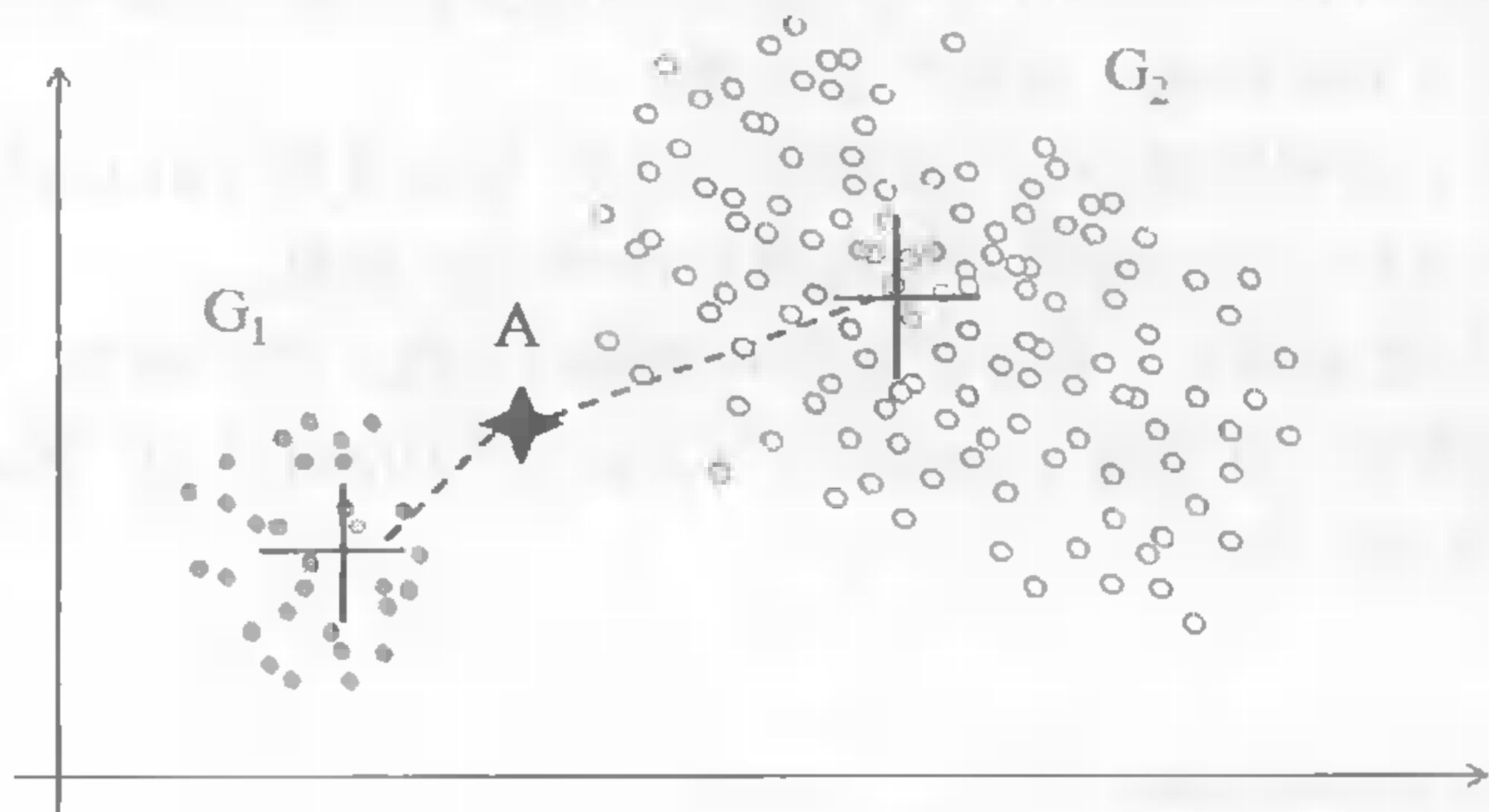


图 8-20 判别分析示意图

判别分析和 8.1.2 小节的聚类分析有什么不同呢？二者之间的主要不同点在于：在聚类分析中，通常人们事先并不知道或一定明确应该分成几类，而是完全根据数据来确定；而在判别分析中，则要求至少有一个已经明确知道类别的“样本”，利用这个样本数据的特征，就可以建立判别准则，并通过预测变量来为未知类别的观测值进行判别。

人们通常把判别分析中已经明确知道的类别样本称为“训练样本”，判别分析过程往往是依据训练样本的特征来进行的。如某企业对其生产产品的消费者购买意愿进行调查。经过调查研究，有 101 个被调查的消费者被划分为“潜在顾客”，另外有 32 个被调查的消费者被划分为“非潜在顾客”。研究者希望从这些被调查的消费者特征出发，从中找出一个分类标准，对那些还没有进行归类的消费者进行定位。而研究者所依据的这些被调查的 133 个消费者的数据就是一个“训练样本”。

## 8.4 判别分析的步骤和过程

判别分析的基本思路是根据从不同总体（设有  $G_1, G_2, \dots, G_i$  个总体）中随机抽取出来的不同样本，在分析样本特征的基础之上建立一定的判别法则，根据新的样本特征和判别法则判别新样本应该来自于哪一个总体。

在判别分析过程中，建立判别法是尤为重要的步骤，也是判别分析的核心所在。根据不同的方法，可以建立不同的判别法则。如果已知或假定总体服从一定的分布（如多元正态分布），则可以使用参数判别规则，反之则可以采用非参数判别规则。

在 SAS 系统中，可以用上述两种判别规则进行判别分析。

参数判别的基本思路具体如下：先根据协方差矩阵计算新样本点到各类中心的距离，并且依据广义距离的大小，把新样本点归入距离最近的一类；或先计算新样本点属于各类的后验概率，然后把新样本归入后验概率最大的一类。

而非参数方法以后验概率为依据进行判别，与参数判别规则不同的是其使用核估计或最近邻估计对概率密度进行估计，这两种估计也需要定义距离。而后验概率通常也可以用距离来表示。

与聚类分析一样，判别规则中的距离同样可以选取不同定义的距离，如欧氏距离、马氏距离等。判别规则所依据的最简单原则是：新样本点离哪一个类别中心的距离最近，那么它就属于哪一类。

除了上述两种主要判别规则和方法之外，在 SAS 系统中还可以使用典型判别法、逐步判别法等多种方法进行判别分析，本节主要介绍常用的几种方法。

### 8.4.1 距离判别

顾名思义，距离判别的基本思想是：样品和哪个总体距离最近，就判它属于哪个总体。由于所有类别已知，所以可求得每个类的中心。这样只要定义了如何计算距离，就可得到任何给定的点到类型中心的距离。这种根据远近判别的方法原理简单、直观易懂，因此距离判别也被称为直观判别法。

在通常情况下，距离判别过程一般采用马氏（Mahalanobis）距离。马氏距离是样本点  $x$  到类中心  $\mu_i$  的一种相对距离。该距离由印度数学家马哈拉比斯于 1936 年依据协方差矩阵  $V$  提出，其计算公式如下。

$$d_i^2 = [x - \mu_i]^T V_i^{-1} [x - \mu_i]$$

马氏距离不受总体空间大小的影响，也不受计量单位的影响，它反映了按平均水平计算被判定样本到中心的相对距离（该距离以方差为单位），实质上就是经过标准化的变量的欧氏距离。

在距离判别过程中，把用来比较样本点到各类中心距离的数学函数称为判别函数。在通常情况下，用线性判别函数进行判别分析非常直观，使用起来最方便，在实际生活中的应用广泛。

### 8.4.2 Bayes 判别

距离判别虽然简单直观、实用，但是在该方法中，没有考虑到每个分类的观察值不同时，每类出现的机会是不同的，也没有考虑误判所造成的损失差异。Bayes 判别可以克服上述缺点，其判别效果更加理想，应用也更广泛。

把对每个样本可能属于某个总体（类别）的可能估计值称为“先验概率”（Prior Probability），并把其记为  $P(G_i)$ 。先验概率可以从经验中得出，也可按每组样本占全部样本的百分比来估计。

每个样本可以根据判别函数计算出得分，在属于  $G_i$  类别条件下判别得分  $S$  的条件概率为  $P(S/G_i)$ 。把样本根据判别函数得分而判为某个类别  $G_i$  的概率称为“后验概率”（Post Probability），根据贝叶斯公式可以计算出后验概率为：
$$P(G_i/S) = \frac{P(S/G_i)P(G_i)}{\sum P(S/G_i)P(G_i)}$$
。可以依

据每个样本被判入某个类别的后验概率进行归类。因而 Bayes 判别的基本思路是：对每个样本，首先计算出判别函数得分，然后根据先验概率  $P(G_i)$  和判别得分  $S$  的条件概率  $P(S/G_i)$ ，计算出该样本被判为每一类的后验概率  $P(G_i/S)$ ，被判入哪一类的后验概率最大，则把该样本判为哪一类。



#### 例 8-4

在例 8-3 中，已经利用相应的聚类方法对各种游戏鼠标依据多个指标进行了分类。现取 Ward 法聚类结果（如图 8-15 所示）把 13 个鼠标分为 3 类（为避免数字作为类别名称混淆输出结果，本例用 A、B、C 表示类别）。现假定这 13 个鼠标的样本来自于已有类别的总体（即已知具体鼠标类别的训练样本）。现又有两款鼠标的评测数据（详见 Mouse\_Discrim.sas7bdat）如表 8-10 所示，试利用判别分析的方法把两款鼠标归入对应的类别。

表 8-10 15 款游戏鼠标横向评测数据

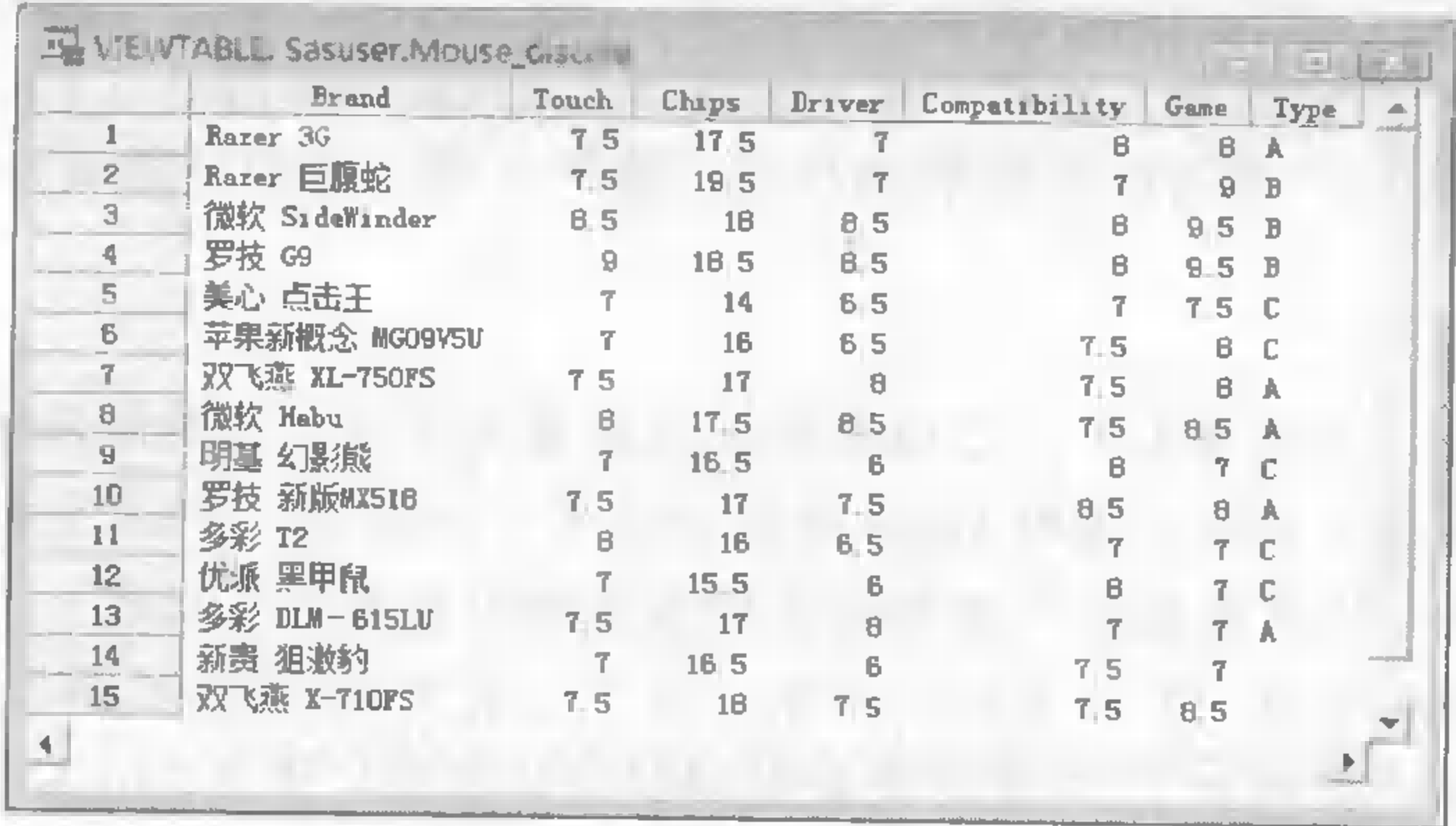
品牌型号 (Brand)	外观及手感 (Touch)	芯片及微动 (Chips)	功能及驱动 (Driver)	兼容性 (Compatibility)	游戏性 (Game)	类别 (Type)
Razer 3G	7.5	17.5	7	8	8	A
Razer 巨腹蛇	7.5	19.5	7	7	9	B
微软 SideWinder	8.5	18	8.5	8	9.5	B
罗技 G9	9	18.5	8.5	8	9.5	B
美心 点击王	7	14	6.5	7	7.5	C
苹果新概念 MG09V5U	7	16	6.5	7.5	8	C
双飞燕 XL-750FS	7.5	17	8	7.5	8	A
微软 Habu	8	17.5	8.5	7.5	8.5	A
明基 幻影熊	7	16.5	6	8	7	C
罗技 新版 MX518	7.5	17	7.5	8.5	8	A
多彩 T2	8	16	6.5	7	7	C
优派 黑甲鼠	7	15.5	6	8	7	C
多彩 DLM-615LU	7	17.0	8	7	7	A
新贵 狙击豹	7	16.5	6	7.5	7	
双飞燕 X-710FS	7.5	18.0	7.5	7.5	8.5	

在表 8-10 中，利用“Type”变量对各个鼠标所归属的类别进行标记。而最后两款鼠标（新贵及双飞燕）是待进行判别对象。

在 SAS 系统中，对于判别分析的原始数据，有两种数据预处理方式。

第一种方式是把已经分好类别的训练样本和未分类的样本放在同一个数据集之中，并且用一个分类变量来标注各个样本所属的类别，如本例中的“Type”变量；而未进行归类或待判别的样本数据对应该分类变量的数值不用任何标记表示。如本例，把所有已经归类和待判数据均放在 Mouse\_Discrim.sas7bdat 数据集中，如图 8-21 所示。

第二种方式是把已有类别的数据和待判数据分别存储为两个数据集，这两个数据集当中作为判别依据的变量的变量名必须相同。如本例把存储已经分好类别样本的数据集命名为“Mouse\_D1”，该数据集即为“训练样本”；而将要进行判别的样本存储为“Mouse\_D2”，如图 8-22 和图 8-23 所示。



	Brand	Touch	Chips	Driver	Compatibility	Game	Type
1	Razer 3G	7.5	17.5	7	8	8	A
2	Razer 巨腹蛇	7.5	19.5	7	7	9	B
3	微软 SideWinder	8.5	18	8.5	8	9.5	B
4	罗技 G9	9	18.5	8.5	8	9.5	B
5	美心 点击王	7	14	6.5	7	7.5	C
6	苹果新概念 MG09V5U	7	16	6.5	7.5	8	C
7	双飞燕 XL-750FS	7.5	17	8	7.5	8	A
8	微软 Habu	8	17.5	8.5	7.5	8.5	A
9	明基 幻影熊	7	16.5	6	8	7	C
10	罗技 新版MX518	7.5	17	7.5	8.5	8	A
11	多彩 T2	8	16	6.5	7	7	C
12	优派 黑甲鼠	7	15.5	6	8	7	C
13	多彩 DLM-615LU	7.5	17	8	7	7	A
14	新贵 狙击豹	7	16.5	6	7.5	7	
15	双飞燕 X-710FS	7.5	18	7.5	7.5	8.5	

图 8-21 判别分析的原始数据

VIEWTABLE: Sasuser.Mouse_d1							
	Brand	Touch	Chips	Driver	Compatibilty	Game	Type
1	Razer 3G	7.5	17.5	7		8	A
2	Razer 巨眼蛇	7.5	19.5	7		7	B
3	微软 SideWinder	8.5	18	8.5		8	B
4	罗技 G9	9	18.5	8.5		8	B
5	美心 点击王	7	14	8.5		7	C
6	苹果新概念 W09V5U	7	16	8.5	7.5	8	C
7	双飞燕 XL-750FS	7.5	17	8	7.5	8	A
8	微软 Haba	8	17.5	8.5	7.5	8.5	A
9	明基 幻影鼠	7	18.5	8	8	7	C
10	罗技 新版MX518	7.5	17	7.5	8.5	8	A
11	多彩 T2	8	16	8.5	7	7	C
12	优派 黑甲鼠	7	15.5	8	8	7	C
13	多彩 DLN-815LU	7.5	17	8	7	7	A

图 8-22 判别分析的训练样本数据集

VIEWTABLE: Sasuser.Mouse_d2							
	Brand	Touch	Chips	Driver	Compatibilty	Game	Type
1	新贵 狙击豹	7	18.5	8	7.5	7	
2	双飞燕 X-710FS	7.5	18	7.5	7.5	8.5	

图 8-23 判别对象数据集

SAS 系统主要通过 DISCRIM 过程进行上述方法的判别。DISCRIM 过程的主要语法如下。

PROC DISCRIM < 选项 >;

CLASS 变量;

BY 变量;

FREQ 变量;

ID 变量;

PRIORS 概率;

TESTCLASS 变量;

TESTFREQ 变量;

TESTID 变量;

VAR 变量;

WEIGHT 变量;

CLASS 语句的主要作用是指定分类的标注标量，即存储类别信息的变量；PRIORS 语句主要用于指定类别之间的先验概率的关系，也可指定每个类别的先验概率，默认为所有类别先验概率相等；TESTCLASS、TESTFREQ、TESTID 语句主要是用 DISCRIM 选项中由 TESTDATA 关键字所指定的数据集中的变量来对观测值进行判别；VAR 语句用于指定进行判别分析的变量；BY、FREQ、ID、WEIGHT 语句与前面章节介绍的功能相同。

DISCRIM 的常用选项如下。

- DATA =：指定用于判别分析的数据集。
- TESTDATA =：指定存储将要被判别的样本的数据集。
- CANONICAL：进行典型判别分析。
- DISTANCE：显示马氏距离平方、F 统计量及其他相关概率。
- METHOD =：指定分类方法。“NORMAL”关键字表示依据参数判别方法使用线性判别函数判别（为默认选项），“NPAR”关键字表示使用非参数方法判别。
- POOL =：该选项有“YES”、“NO”和“TEST”3 个关键字。“POOL=YES”表示使用混合协方差矩阵（即所有类别合并计算的协方差）计算平方距离，并计算线性判别函数；“POOL=NO”表示使用各类间协方差矩阵计算距离，并计算二次判别函数；“POOL=TEST”表示使用 Bartlett’s 类内协方差矩阵同质性的极大似然比检验修正。系统默认选择“YES”关键字。
- LIST：显示对每个样本进行判别分析的后验概率、类别等分类结果。
- TESTLIST：显示对每个待判样本进行判别分析的后验概率、类别等分类结果。
- OUT =：指定存储分类结果的数据集。

- OUTSTAT =： 指定存储协方差矩阵等统计量的数据集。
- CROSSVALIDATE: 指定系统进行交叉核实验证。
- TESTOUT =： 指定存储待判样本分类结果的数据集。

在 DISCRIM 过程中，当 CLASS 语句指定的分类变量有缺失值时，该缺失值对应的样本会被系统自动排除，不参加判别规则的制定。如果样本中只有 CLASS 语句指定的分类变量存在缺失值，其他变量没有缺失值时，则该样本会自动依据判别规则进行归类并在归类结果中显示出来。这也是第一种数据预处理方法的理论依据。

本例使用第一种数据处理方式的程序如下。

```
proc discrim data=Sasuser.Mouse_Discrim list out=Mouse_Discrim_Out distance pool=yes; /* 调用
DISCRIM 过程，使用含有训练样本和待判数据的 Sasuser.Mouse_Discrim 数据集进行分析，显示用于判别的
距离；使用混合协方差矩阵计算线性判别函数，并且把判别结果显示在“Output”窗口和存储在
Mouse_Discrim_Out 临时数据集当中*/
  class Type;                               /*指定分类变量*/
  var Touch Chips Driver Compatibility Game /*指定用于判别依据的变量*/
  id Brand;                                  /*指定标注样本名称的变量*/
run;
```

运行程序后可以得到非常多的输出结果，首先看到的是进行判别分析的一些基本情况。因为本例没有指定样本归类的先验概率，因此系统自动假定归入每一个类别的先验概率相等（本例每类的先验概率为 0.3333）。

接下来，系统给出了每个类别两两配对的马氏距离平方，以及依据马氏距离测度的类别之间差异性是否显著的 F 统计量值及其对应的 P 值。如图 8-24 所示。

图 8-24 中的第一部分为距离平方的测算公式，“Squared Distance to Type”列示了本例中鼠标的 A、B、C 3 个类别之间的距离平方，同时也给出了各类距离差异检验的 F 统计量值，图 8-24 最下部分的“Prob>Mahalanobis Distance for Squared Distance to Type”表格中给出了对应 F 统计量的 P 值，各组差异在显著性水平  $\alpha = 0.05$  的条件下可以通过显著性检验，即可以认为这 3 个类别之间是有差异的，且差异明显，在此基础上进行归类才有意义。

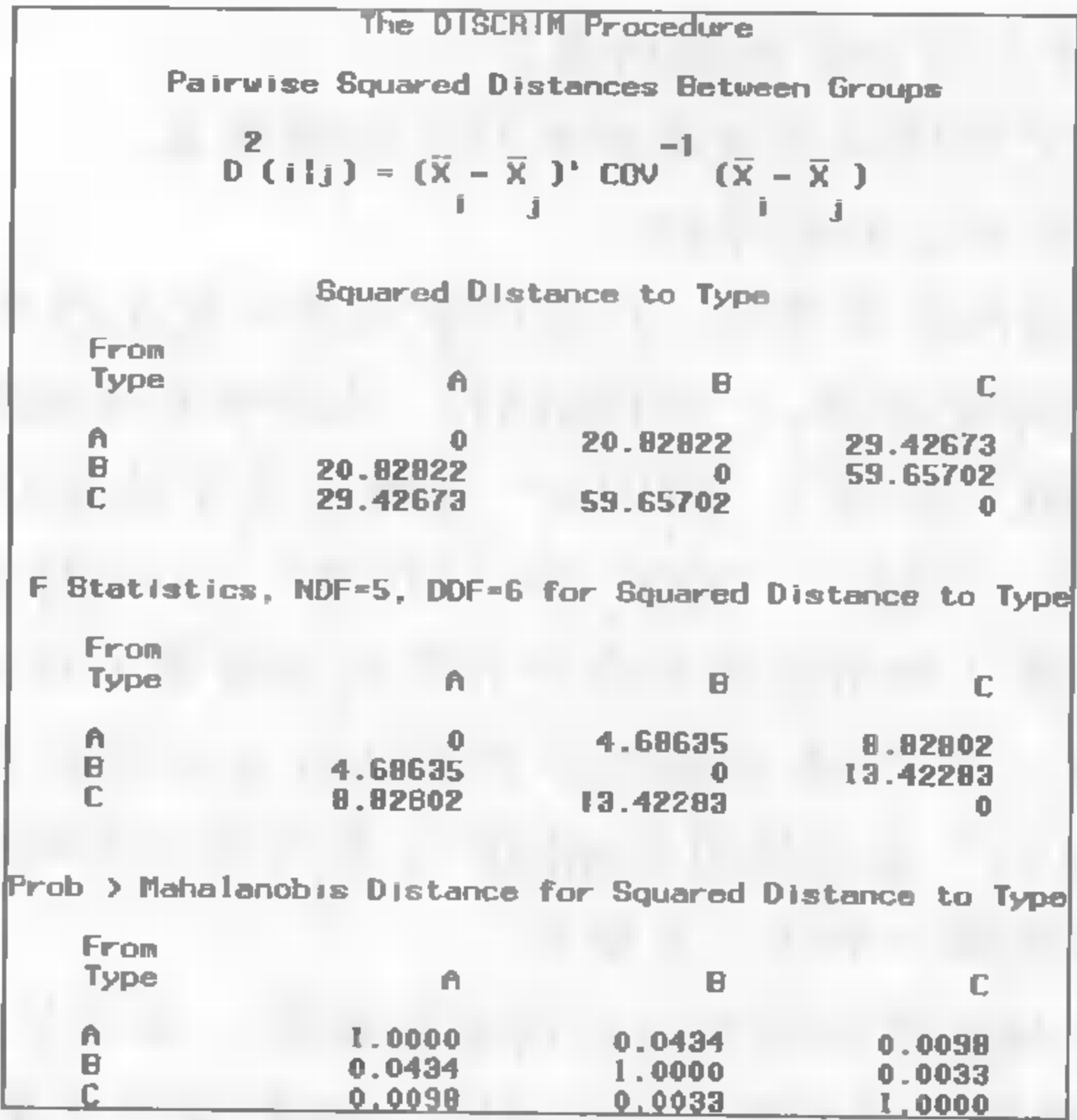


图 8-24 判别分析的基本信息

DISCRIM 过程的输出结果还会给出用于距离判别的判别函数，如图 8-25 所示。

The DISCRIM Procedure

Linear Discriminant Function

Constant =  $-.5 \sum_j \bar{X}_j' \text{COV}_j^{-1} \bar{X}_j$       Coefficient Vector =  $\text{COV}_j^{-1} \bar{X}_j$

Linear Discriminant Function for Type

Variable	Label	A	B	C
Constant		-788.16350	-903.03606	-626.22517
Touch	外观及手感	-7.81528	-1.50078	1.39421
Chips	芯片及微动	51.34464	54.06714	44.29113
Driver	功能及驱动	54.93436	50.29628	40.25220
Compatibility	兼容性	24.40398	21.03482	21.75651
Game	游戏性	17.24008	26.32372	18.45311

图 8-25 判别分析的线性判别函数

根据图 8-25 所示的结果，可以建立以下线性判别函数。

$$\begin{aligned} S_A &= -788.16350 - 7.81528 \times Touch + 51.34464 \times Chips + 54.93436 \times Driver \\ &\quad + 24.40398 \times Compatibility + 17.24008 \times Game \\ S_B &= -903.03606 - 1.50078 \times Touch + 54.06714 \times Chips + 50.29628 \times Driver \\ &\quad + 21.03482 \times Compatibility + 26.32372 \times Game \\ S_C &= -626.22517 + 1.39421 \times Touch + 44.29113 \times Chips + 40.25220 \times Driver \\ &\quad + 21.75651 \times Compatibility + 18.45311 \times Game \end{aligned}$$

根据判别函数及各样本指标值，可以计算各样本在各类的判别函数得分。而且，可以依据判别函数得分可以把需要判别的样本归入其对应的类别中，即哪个判别函数得分最高，就把该样本归入哪一类中。

如对于品牌为“新贵 狙击豹”的待判样本，依据判别函数计算得分如下。

$$\begin{aligned} S_A &= -788.16350 - 7.81528 \times 7 + 51.34464 \times 16.5 + 54.93436 \times 6 \\ &\quad + 24.40398 \times 7.5 + 17.24008 \times 7 \\ &= 637.6327 \\ S_B &= -903.03606 - 1.50078 \times 7 + 54.06714 \times 16.5 + 50.29628 \times 6 \\ &\quad + 21.03482 \times 7.5 + 26.32372 \times 7 \\ &= 622.3721 \\ S_C &= -626.22517 + 1.39421 \times 7 + 44.29113 \times 16.5 + 40.25220 \times 6 \\ &\quad + 21.75651 \times 7.5 + 18.45311 \times 7 \\ &= 648.1967 \end{aligned}$$

因为  $S_C > S_A > S_B$ ，故把该样本归入 C 类。同理，对于品牌为“双飞燕 X-710FS”的鼠标，其判别函数得分分别为 819.0037、817.6515、803.3885，即  $S_A > S_B > S_C$ ，故可把该样本归入 A 类。

此外，根据得分和先验概率，可以计算出每个样本归入每一类的后验概率。SAS 系统输出结果如图 8-26 所示。

图 8-26 中所示的公式分别为马氏距离平方和样本归入每个类别后验概率的计算公式。在图 8-26 中的“Posterior Probability of Membership in Type”表格中，详细地列示了每个样本原始类别（“From Type”）、进行判别分析之后应该归入的类别（Classified into Type）以及每个

样本归入每个类别的后验概率（本例训练样本有 A、B、C3 大类，每个类别的标记对应的列即为归入该类的后验概率）。

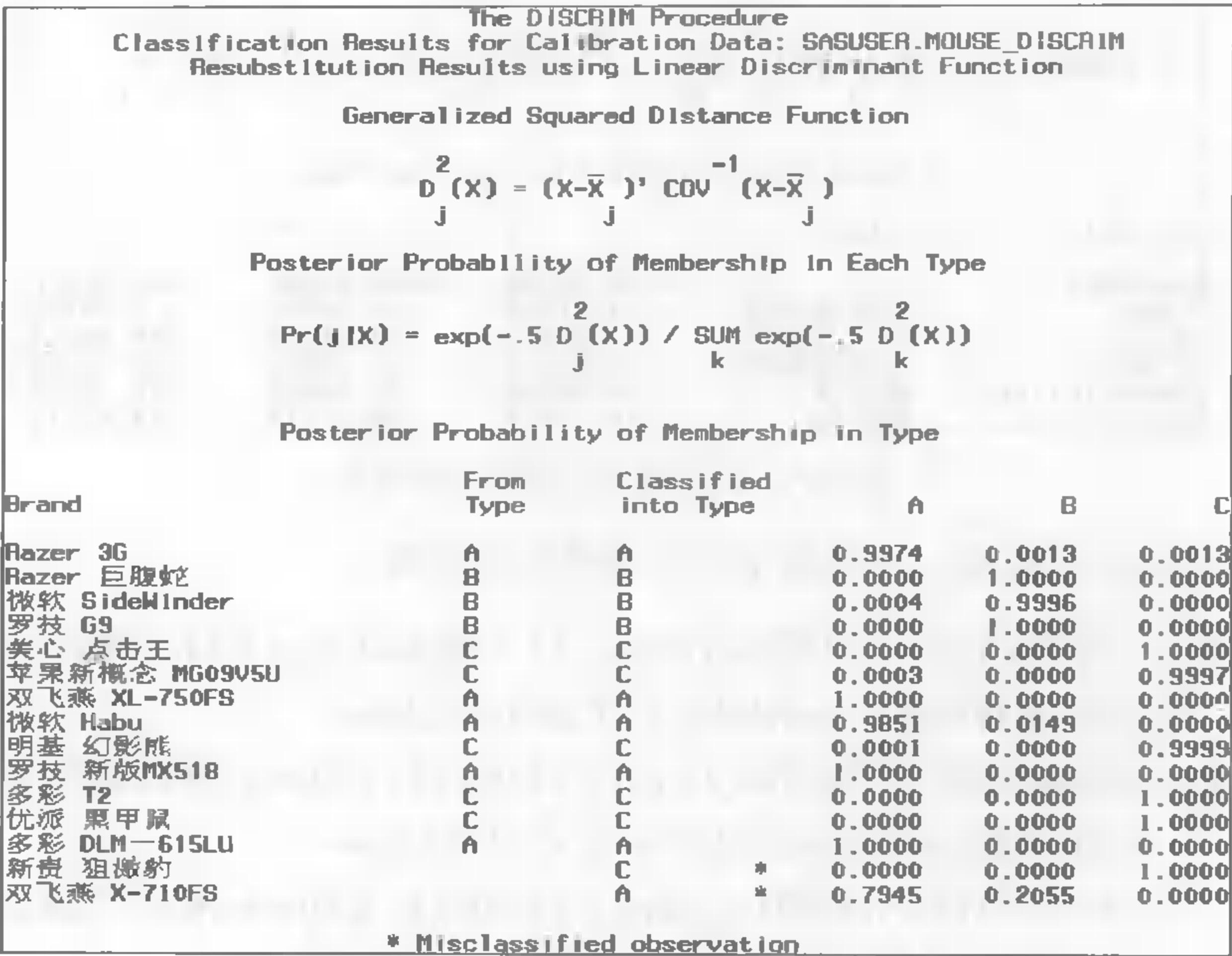


图 8-26 判别分析的判别结果及后验概率

SAS 系统用 “\*” 标注（同时 “\*” 也用于标注经过判别后归类与原有类别不一致的情况）事先没有进行分类即进行判别分析的对象，即图 8-26 中的最后两个样本——“新贵 狙击豹”和“双飞燕 X-710FS”。从这张表格中可以看出，依据后验概率的大小，“新贵 狙击豹”归入 C 类的后验概率为 1，故把其归入 “C” 类；而“双飞燕 X-710FS”归入 A 类的后验概率最大为 0.7945，故把其归入 “A” 类。

每个样本的判别结果及计算出来的后验概率都会保存在 DISCRIM 语句选项 OUT 关键字指定的数据集中，本例程序已经指定 “OUT = Mouse\_Discrim\_Out”，读者可在临时数据库中查看 Mouse\_Discrim\_Out 数据集的结果，其内容与图 8-26 所列示结果类似。

在对事先没有进行分类的样本进行判别的同时，DISCRIM 过程还对原有训练样本中的每个样本进行判别。除了图 8-26 中列示出的进行判别之后的归类结果之外，系统还给出了判别分析的交叉核实过程及错判概率，如图 8-27 所示。

图 8-27 中的 “Number of Observations and Percent Classified into Type” 表示记为交叉核实表（该表也可通过 DISCRIM 选项的 CROSSVALIDATE 关键字得到）。“From Type” 列表示数据集中待判的原有类别（含训练样本原有类别及待判样本类别，其中待判样本类别为空白），而列 “A”、“B”、

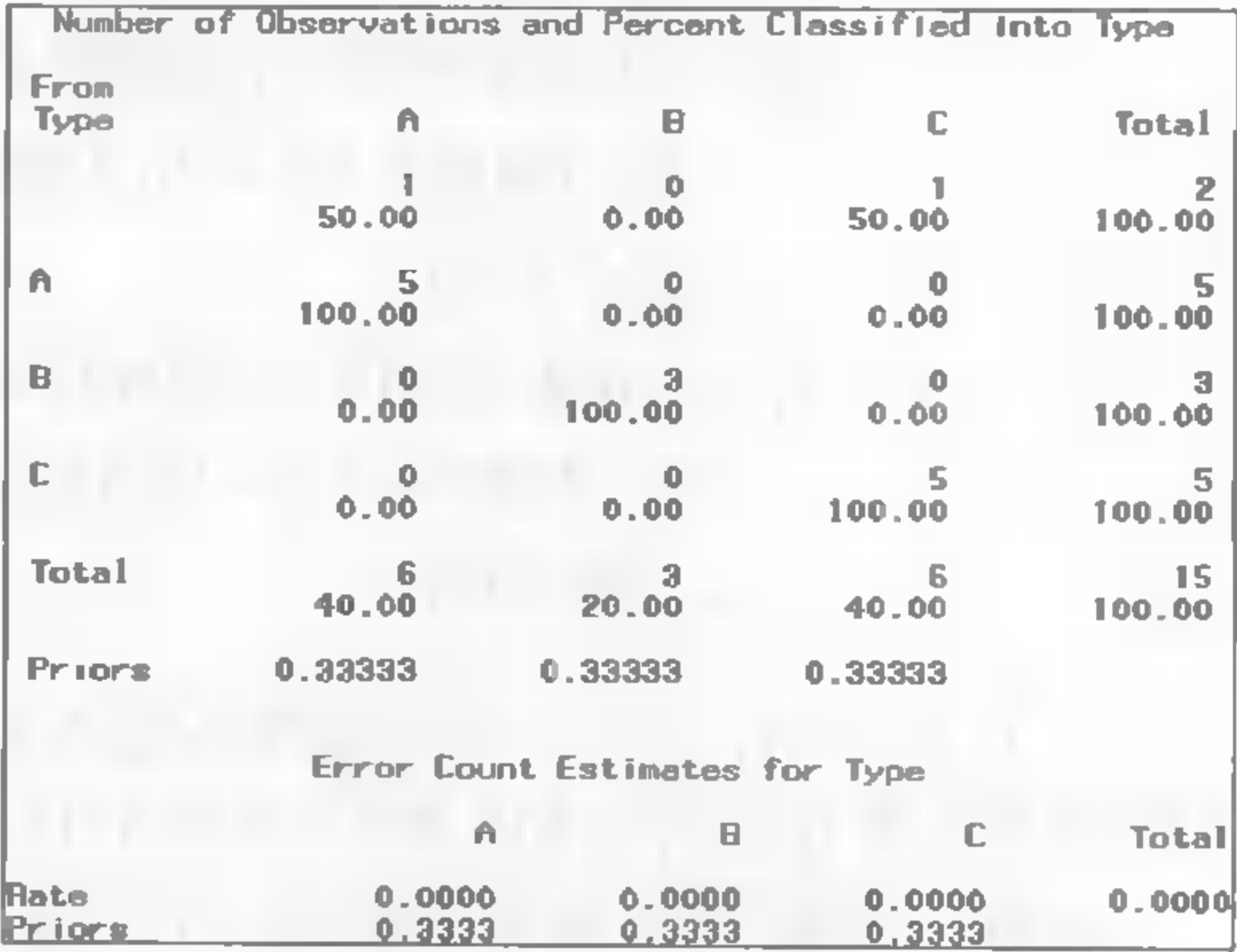


图 8-27 判别分析的交叉核实过程及错判概率

“C”表示对待判样本进行判定的类别，即进行判别分析之后的现有类别。

在图 8-27 中，原有类别为空白的样本一共有两个，分别归入 A 类和 C 类；原有类别为“A”的样本一共有 5 个，经过判别分析之后，全部归入 A 类。依此类推，原有类别为“B”的 3 个样本经过判别后全部归入 B 类，原有类别为“C”的 5 个样本也全部归入到 C 类中。总结归类的全过程：一共有 6 个样本归入了 A 类，占总样本的 40%；3 个样本归入 B 类，占总样本的 20%；6 个样本归入了 C 类，占总样本的 40%。

对于上述判别分析的交叉核实过程，系统自动给出了每个类别的错判概率。简单错判概率由以下公式计算。

$$p = \frac{1}{n} \sum_{i=1}^k \sum_{\substack{j=1 \\ j \neq i}}^k m_{ij}$$

如果考虑先验概率  $p$ ，则错判概率可由以下加权形式计算。

$$P = \sum_{i=1}^k q_i p_i$$

从图 8-27 中的“Error Count Estimates for Type”表格可以看出，本例每一个类别的错判概率均为 0，总错判概率为 0，表明本例所考察的鼠标分类是正确的，且分类精度和可靠性非常高。

如果采用第二种方式进行判别本例数据，即把训练样本和待判样本分为两个数据集进行分析，则利用 DISCRIM 过程的程序如下。

```
proc discrim data=Sasuser.D1 testdata=Sasuser.D2 testlist testout=Mouse_D2_Out;
  class Type;
  testclass Type;
  testid Brand;
  var Touch Chips Driver Compatibility Game;
run;
```

如使用上述的第二种数据预处理方式，要注意在 DISCRIM 选项中，用 DATA 语句指定训练样本数据集，用 TESTDTAT 语句指定待判样本数据集，TESTLIST 语句用于列示待判样本的判别结果，TESTOUT 语句用于指定存储待判样本输出结果的数据集。此外还要在该过程中增加指定待判样本类别分类变量的 TESTCLASS 语句，同时可用 TESTID 语句指定标注待判样本名称的变量。

运行上述程序之后，可以得到与第一种数据预处理方式相同的结果。

前面讨论的距离判别需要估计总体的参数，估计总体参数的前提是已知总体服从什么样的分布。而进行 Bayes 判别时，也应假定总体服从正态分布。在一般情况下，各类总体分布是否就是正态分布，往往是未知的。当总体分布未知时，可以使用非参数判别法。通常可用核方法和近邻方法进行非参数判别分析。

非参数判别仍然使用 Bayes 后验概率作为判别的依据。设有  $n$  个类别，由于各类总体分布未知，故每个类别具有的概率密度函数  $f_n(x)$  未知，于是可对  $f_n(x)$  利用核方法或近邻方法进行估计，将估计的先验概率和密度函数结果代入判别规则中以计算后验概率，然后再按照前面介绍过的方法进行归类。

SAS 系统中的 DISCRIM 过程也可进行非参数判别,通过 DISCRIM 选项中的“METHOD = NPAR”关键字来指定该过程进行非参数判别。选用非参数判别方法时,还应通过“R = 核估计半径”指定系统用核估计方法进行估计,或者通过“K = 近邻个数”指定系统使用近邻估计方法进行估计。

### 8.4.3 Fisher 判别

从距离判别中已经看到了线性判别函数的方便性,由于线性判别函数使用简便,因此人们希望能在更一般的情况下建立一种线性判别函数。而于 1936 年提出的 Fisher 判别法就是根据投影转换坐标轴的思想建立起来的、一种能较好区分各个总体类别的线性判别法。该判别方法对总体的分布不做任何要求,因此也被叫做典型判别。

为了更好地理解 Fisher 判别方法,先考虑两类总体的判别。如图 8-28 所示,有一个已知两个类别的训练样本,把其对应到二维坐标系中,分别用“○”和“●”表示,则按照原有的横纵坐标很难区分这两个类别。

为了更好地用坐标轴区分这两个类别,通过观测,在图 8-28 中寻找能够使得“○”和“●”被明显区分的坐标轴方向,并且用虚线画出来。则每个样本点都沿着重新定位的坐标轴进行投影,将会使得这两个类别分得十分清楚。如果向其他方向投影,则类别区分效果不好。

把上述过程推广到多维的情形,即把 N 类的 m 维数据投影到某一个方向,使得变换后同类别的样本点尽可能聚在一起,不同类别的样本点尽可能分离,以此达到分类的目的。这种将样本首先进行投影,再用判别规则得到判别准则以对待判样本进行归类的判别方法就是 Fisher 判别法。其实质是寻找一个最能反映类与类之间差异的投影方向,即寻找线性判别函数。

在 SAS 系统中, Fisher 判别函数通常以典型函数的形式给出,如果训练样本分类的数据及其指标太多,则一个判别函数是不够的,这时需要寻找第二个甚至第三个线性判别函数。通过投影方法找出的每个典型函数具有对应的判别效率。那么究竟需要多少个判别函数来对样本进行判定呢?通常可以取判别效率 85%的阈值进行判别函数数目的选择。

Fisher 判别或典型判别在 SAS 系统中可用 CANDISC 过程实现,其主要语法如下。

```
PROC CANDISC < 选项 >;
```

```
CLASS 变量;
```

```
BY 变量;
```

```
FREQ 变量;
```

```
VAR 变量;
```

```
WEIGHT 变量;
```

CANDISC 过程的语句比较简单,各语句对应的功能与本书前面章节介绍的一致。其中 CANDISC 的主要选项如下。

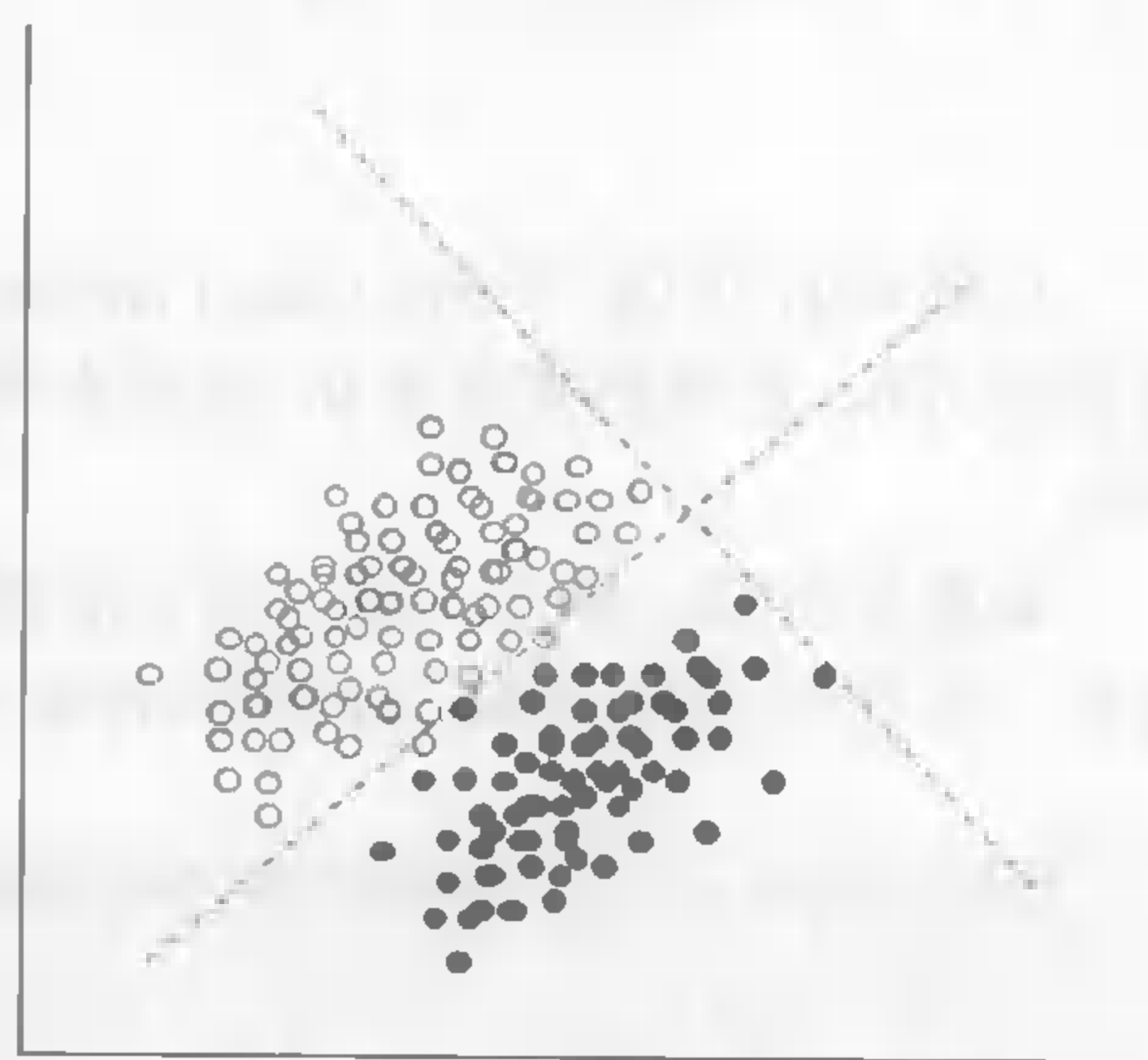


图 8-28 2 类别 Fisher 判别示意图

- DATA=: 指定用于分析的原始数据集。
- OUT=: 指定存储分析结果的数据集。
- OUTSTAT=: 指定存储分析过程中的统计量的数据集。
- SIMPLE: 显示全部样本和类别之间的简单描述统计量。
- DISTANCE: 显示马氏距离平方、*F* 统计量及其他相关概率。

本节仍然以例 8-4 为例，利用 Fisher 法进行判别分析，具体程序如下。

```
proc candisc data=Sasuser.Mouse_Discrim out=Mouse_Can_Out distance simple;
  class Type;
  var Touch Chips Driver Compatibility Game;
run;
proc print data=Mouse_Can_Out;
run;
```

运行程序后可得到非常多的输出结果，这里结合实际问题主要介绍其中的一些重点内容。

运行程序后在“LOG”窗口中，系统提示用于分析数据的分类变量含有缺失值（本例最后两个样本没有进行归类，为待判样本）。在 CANDISC 过程中，不会将分类变量含有缺失值的样本纳入计算判别规则的过程。但是在计算判别函数得分的过程中，系统自动会为分类变量含有缺失值的样本计算判别函数得分。

系统首先给出了全部样本及分类样本的基本统计量，如图 8-29 所示。

The CANDISC Procedure						
Simple Statistics						
Total-Sample						
Variable	Label	N	Sum	Mean	Variance	Standard Deviation
Touch	外观及手感	13	99.00000	7.61538	0.38141	0.6176
Chips	芯片及微动	13	220.00000	16.92308	1.95192	1.3971
Driver	功能及驱动	13	94.50000	7.26923	0.90064	0.9490
Compatibility	兼容性	13	99.00000	7.61538	0.25641	0.5064
Game	游戏性	13	104.00000	8.00000	0.83333	0.9129

图 8-29 Fisher 判别过程的训练样本信息

图 8-29 中的信息有助于计算典型判别函数的得分。该过程也对训练样本各组之间的距离差异进行了 *F* 检验，检验结果同图 8-24，这里不予赘述。此外，系统还可以得到对全部样本进行多元单因素方差分析和对各分类均值向量是否相等进行的 4 种多元检验的结果，如图 8-30 所示。

The CANDISC Procedure					
Multivariate Statistics and F Approximations					
Statistic	S=2		M=1		N=2
	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.01998771	7.29	10	12	0.0010
Pillai's Trace	1.66129339	6.87	10	14	0.0007
Hotelling-Lawley Trace	14.94574519	8.51	10	6.7273	0.0055
Roy's Greatest Root	12.13750781	16.99	5	7	0.0009
NOTE: F Statistic for Roy's Greatest Root is an upper bound.					
NOTE: F Statistic for Wilks' Lambda is exact.					

图 8-30 各分类均值相等的多元检验结果

图 8-30 中的 4 种检验方法的 *P* 值（“Pr > F”）显示，各分类均值向量明显不完全相等。系统还给出了典型相关系数的一些信息，如图 8-31 所示。

The CANDISC Procedure					
	Canonical Correlation	Adjusted Canonical Correlation	Approximate Standard Error	Squared Canonical Correlation	
1	0.961188	0.945752	0.021973	0.923882	
2	0.858727	0.837569	0.075803	0.737411	
Eigenvalues of inv(E)*H = CanRsq/(1-CanRsq)					
	Eigenvalue	Difference	Proportion	Cumulative	
1	12.1375	9.3293	0.8121	0.8121	
2	2.8082		0.1879	1.0000	
Test of H0: The canonical correlations in the current row and all that follow are zero					
	Likelihood Ratio	Approximate F Value	Num DF	Den DF	Pr > F
1	0.01998771	7.29	10	12	0.0010
2	0.26258867	4.91	4	7	0.0332

图 8-31 典型相关系数及特征根

对于典型相关系数的分析和检验问题，请参见 4.2 节，这里也不予赘述。但是，根据图 8-31 中的特征根贡献率可以判定判别函数的判别效率。本例中，第一个判别函数对应的特征根为 12.1375，贡献率为 0.8121，接近于 85%的阈值，因此，可以认为第一个判别函数在一定程度上具有较强的判别效率（第二个判别函数的效率仅为 0.1879），故本例选用第一个判别函数。提取的线性判别函数系数由图 8-32 所示的输出结果给出。

The CANDISC Procedure				
Total-Sample Standardized Canonical Coefficients				
Variable	Label	Can1	Can2	
Touch	外观及手感	-0.465733457	-1.426175661	
Chips	芯片及微动	1.849338859	0.357242080	
Driver	功能及驱动	1.649504787	2.462638955	
Compatibility	兼容性	0.032106365	0.493653582	
Game	游戏性	0.652886747	-1.785468261	
Pooled Within-Class Standardized Canonical Coefficients				
Variable	Label	Can1	Can2	
Touch	外观及手感	-0.351024400	-1.074911942	
Chips	芯片及微动	0.951459634	0.183796180	
Driver	功能及驱动	0.967739715	1.444793333	
Compatibility	兼容性	0.034534870	0.530993215	
Game	游戏性	0.332908226	-0.910413751	
Raw Canonical Coefficients				
Variable	Label	Can1	Can2	
Touch	外观及手感	-0.754121381	-2.309281289	
Chips	芯片及微动	1.323686520	0.255700313	
Driver	功能及驱动	1.738111838	2.594925430	
Compatibility	兼容性	0.063404991	0.974887711	
Game	游戏性	0.715201598	-1.955882485	

图 8-32 典型判别函数系数

SAS 系统的 CANDISC 过程一共给出 3 种判别函数系数：第 1 种是全样本（Total-Sample）标准化典型系数，第 2 种是混合类间（Pooled Within-Class）标准化典型系数，第 3 种是原始形式（Raw）的典型系数。为了实际分析的需要，本书主要介绍根据第 3 种典型系数计算判别函数得分。

依据原始形式的典型系数计算判别函数得分时，需要把原始变量中心化，即求各原始变量与该变量均值之差，得到中心化数据，再根据典型系数求中心化后数据的线性组合得分。对照图 8-29 中的各变量均值与图 8-32 中的原始典型系数，依据特征根贡献率，计算品牌为“新贵 狙击豹”的待判样本的第一个典型判别函数得分如下。

Can1=-0.754121381×(7-7.61538)+1.32368652×(16.5-16.92308)+1.738111838×(6-7.26923)

+0.063404991×(7.5-7.61538)+0.715201598×(7-8)

=-3.02453

同理，可计算品牌为“双飞燕 X-710FS”的待判样本的第一个典型判别函数得分为 2.26391。如果按照本例的第一种数据预处理方法，则系统会自动依据上述判别函数计算训练样本和待判样本的典型判别函数得分，具体结果保存在 CANDISC 过程选项中由 OUT 关键字指定的数据集当中。读者如有兴趣，可打开本例生成的临时数据集 Mouse\_Can\_Out.sas7bdat，即可看到已经计算出来的判别函数得分。

用典型判别函数得分判定样本的归类与用距离判别函数得分判定样本归类的标准不一样。在典型判别中，根据典型变量得分与典型变量类均值（如图 8-33 所示）之间的距离来进行归类。

依据最短距离原则，样本的判别函数得分与哪一类典型变量类均值之间的距离最小，就把该样本判为该类。根据图 8-33 所示的结果和第一个判别函数得分（Can1 变量），计算每个样本到类中心的距离，如表 8-11 所示。

Class Means on Canonical Variables		
Type	Can1	Can2
A	1.234542372	1.761722043
B	3.993558080	-1.873666509
C	-3.630677220	-0.637522138

图 8-33 典型变量类均值

表 8-11 每个样本的第一个典型判别函数得分及 Fisher 判别结果

品 牌	Type	第一个典型判 别函数得分	与类 A 的距离	与类 B 的距离	与类 C 的距离	归类
			1.234542	3.993558	-3.63068	
Razer 3G	A	0.407112653	-0.82743	-3.58645	4.03779	A
Razer 巨腹蛇	B	3.7062823	2.47174	-0.28728	7.33696	B
微软 SideWinder	B	3.994804685	2.760262	0.001247	7.625482	B
罗技 G9	B	4.279587255	3.045045	0.286029	7.910264	B
美心 点击王	C	-5.138791184	-6.37333	-9.13235	-1.50811	C
苹果新概念 MG09V5U	C	-2.10211485	-3.33666	-6.09567	1.528562	C
双飞燕 XL-750FS	A	1.451678736	0.217136	-2.54188	5.082356	A
微软 Habu	A	2.963118023	1.728576	-1.03044	6.593795	B
明基 幻影熊	C	-2.992826611	-4.22737	-6.98638	0.637851	C
罗技 新版 MX518	A	0.646027808	-0.58851	-3.34753	4.276705	A
多彩 T2	C	-3.603140324	-4.83768	-7.5967	0.027537	C
优派 黑甲鼠	C	-4.316513131	-5.55106	-8.31007	-0.68584	C
多彩 DLM-615LU	A	0.704774642	-0.52977	-3.28878	4.335452	A
新贵 狙激豹		-3.024529107	-4.25907	-7.01809	0.606148	C
双飞燕 X-710FS		2.263910135	1.029368	-1.72965	5.894587	A

在表 8-11 所示的归类结果中，待判样本的归类结果与前两节用距离判别和 Bayes 判别进行归类的结果一致。但是在原有训练样本的验证归类过程中，只有原有类别为“A”、品牌为“微软 Habu”的样本经过典型判别之后，将其归入“B”类，总的说来误判率比较低。

8.4.4 逐步判别

在前面几节中的判别分析方法中，都是把所考虑的变量全部作为建立判别规则的依据进

行分析。有时某些变量可能对判别规则的制定没有多大作用，有时则可能在进行分析时忽略了重要的变量，从而影响判别效果。另一方面，如果选择的变量过于复杂，数目过于繁多，则会增加计算量，从而影响参数估计的精度。因此，变量选择问题是判别分析中的一个重要问题。变量选择是否恰当，是判别分析结果有效性的关键。而本小节将介绍的逐步判别法就是解决判别过程中变量选择问题的一种常用方法。在 SAS 系统中，逐步判别主要用于判别分析的变量筛选过程，如果要建立判别函数对样本进行归类，则须配合上几小节介绍的判别方法以进行判别规则的制定。

对于重要变量的选择，SAS 系统提供了向前引入法 (Forward)、向后剔除法 (Backward) 和逐步选择法 (Stepwise) 3 种方法，分别介绍如下。

- 向前引入法：首先判别函数模型中没有变量，在每一个变量的选择过程中，选择 Wilks' Lambda 统计量最小的变量进入模型。再重复该变量选择过程，当不再有未被选入变量的统计量小于入选临界值时，向前引入过程结束。
- 向后剔除法：首先，用户指定的所有变量都在判别模型中。然后根据 Wilks' Lambda 统计量的剔除临界值，剔除模型中判别能力贡献最小的变量。再重复该过程，直到所有留在模型中的变量均符合标准时，向后剔除过程结束。
- 逐步选择法：该法是向前引入和向后剔除的综合，采用有进有出的算法，即每一步都进行检验。首先对各变量进行计算、检验，从中将判别能力最强的变量和最重要的变量引进判别函数中。而对较早进入判别函数的变量，随着其他变量的进入，其显著性可能发生变化，如果其判别能力不强了，则把其从判别函数当中剔除。以此类推，直至所有重要变量都引入判别函数以及所有非重要变量都剔除出判别函数为止。

在逐步判别法对变量的筛选过程中，变量在各分类之间的差异越大，表示该变量越重要。这种差异可用多元方差分析的方法进行判定，故也通常可用  $F$  检验来判断各变量的重要性。

在 SAS 系统中，可用 STEPDISC 过程进行逐步判别。该过程的主要语法如下。

```
PROC STEPDISC < 选项 >;
  CLASS 变量;
  BY 变量;
  FREQ 变量;
  VAR 变量;
  WEIGHT 变量;
```

该过程各语句的功能与上节介绍的判别分析过程相同。

常用的 STEPDISC 选项如下。

- DATA =：指定用于分析的原始数据集。
- METHOD =：指定用于分析的判别方法。其中，“FORWARD”表示向前引入法，“BACKWARD”表示向后剔除法，“STEPWISE”表示逐步选择法。系统默认选择逐步选择法。
- SLENTRY =：指定向前引入法中引入变量的显著性水平阈值，系统默认为 0.15。
- SLSTAY =：指定向后剔除法中保留变量的显著性水平阈值，系统默认为 0.15。

- **START=**: 指定初始模型中的变量数目。当选择向前引入法和逐步选择法时，系统默认值为 0；当选择向后剔除法时，系统默认值为 VAR 语句所指定变量的数目。
- **STOP=**: 指定最终模型中的变量数目，只能用于向前引入法和向后剔除法。当选择向前引入法时，系统默认值为 VAR 语句所指定变量的数目；当选择向后剔除法时，系统默认值为 0。

本小节仍然以例 8-4 为例说明判别分析的变量选择过程，具体程序如下：

```
proc stepdisc method=stepwise data=Sasuser.Mouse_Discrim slentry=0.10 slstay=0.10; /*调用 STEPDISC
过程，使用逐步选择法筛选变量，设定引入显著性水平阈值为 0.1，剔除显著性水平阈值为 0.1*/
class Type;
var Touch Chips Driver Compatibility Game;
run;
```

运行程序后，首先得到逐步判别的一些基本信息，如图 8-34 所示。

The STEPDISC Procedure				
The Method for Selecting Variables is STEPWISE				
Observations	13	Variable(s) in the Analysis		5
Class Levels	3	Variable(s) will be Included		0
		Significance Level to Enter		0.1
		Significance Level to Stay		0.1
Class Level Information				
Type	Variable Name	Frequency	Weight	Proportion
A	A	5	5.0000	0.384615
B	B	3	3.0000	0.230769
C	C	5	5.0000	0.384615

图 8-34 逐步判别的基本信息

因本例指定分析的原始数据集含有两个待判样本，所以其分类变量中含有缺失值。STEPDISC 过程对于含有缺失值的样本不予处理。因此，在图 8-34 中进行分析的样本一共只有 13 个，同时图 8-34 还显示了逐步选择法引入和剔除变量的显著性水平阈值均为程序指定的 0.1。此外，各类别样本构成也一并显示出来。

SAS 输出结果详细地给出了变量筛选过程的每一步。首先是变量筛选的第一个步骤，如图 8-35 所示。

The STEPDISC Procedure					
Stepwise Selection: Step 1					
Statistics for Entry, DF = 2, 10					
Variable	Label	R-Square	F Value	Pr > F	Tolerance
Touch	外观及手感	0.5266	5.56	0.0238	1.0000
Chips	芯片及微动	0.7794	17.67	0.0005	1.0000
Driver	功能及驱动	0.7132	12.43	0.0019	1.0000
Compatibility	兼容性	0.0358	0.19	0.8332	1.0000
Game	游戏性	0.7833	18.08	0.0005	1.0000
Variable Game will be entered.					
Variable(s) that have been Entered					
Game					
Multivariate Statistics					
Statistic	Value	F Value	Num DF	Den DF	Pr > F
Wilks' Lambda	0.216667	18.08	2	10	0.0005
Pillai's Trace	0.783333	18.08	2	10	0.0005
Average Squared Canonical Correlation	0.391667				

图 8-35 逐步选择法的第 1 个步骤

在变量选择法的第一个步骤中，模型中没有任何变量。依据程序中的 VAR 语句，一共有 5 个变量可供选择进入模型，每个模型的  $F$  检验的  $P$  值（“Pr > F”）均列示在图 8-35 的上半部分。依据 0.10 的引入显著性水平阈值，除了变量 “Compatibility” 之外，其余变量均符合引入标准，但逐步选择法的每一步只能选择最有影响的变量进入模型，因此在上述符合引入标准的变量中选择  $F$  值（“F Value”）最大即  $P$  值（“Pr > F”）最小的变量（0.0005），即变量 “Game” 在第一步将被引入模型当中。在该变量选择过程中系统自动提示变量 “Game” 将被引入模型，即提示 “Variable Game will be entered”。接下来是图 3-36 所示的第 2 个步骤。

The STEPDISC Procedure					
Stepwise Selection: Step 2					
Statistics for Removal, DF = 2, 10					
Variable	Label	R-Square	F Value	Pr > F	
Game	游戏性	0.7833	18.08	0.0005	
No variables can be removed.					
Statistics for Entry, DF = 2, 9					
Variable	Label	Partial R-Square	F Value	Pr > F	Toierance
Touch	外观及手感	0.1009	0.50	0.6197	0.5070
Chips	芯片及微动	0.5526	5.56	0.0268	0.4834
Driver	功能及驱动	0.5458	5.41	0.0287	0.5466
Compatibility	兼容性	0.0580	0.28	0.7641	0.9492
Variable Chips will be entered.					
Variable(s) that have been Entered					
Chips Game					
Multivariate Statistics					
Statistic		Value	F Value	Num DF	Den DF Pr > F
Wilks' Lambda		0.096941	9.95	4	18 0.0002
Pillai's Trace		1.004835	5.05	4	20 0.0056
Average Squared Canonical Correlation		0.502417			

图 8-36 逐步选择法的第 2 个步骤

因为在第 1 个步骤中，模型引入了变量 “Game”，所以在第 2 步中，首先依照有进有出的算法对模型中的现存变量进行检验，检验结果如图 8-36 中的 “Statistics for Removal” 项。经过检验，新引入变量 “Game” 的  $P$  值（“Pr > F”）为 0.0005，远远小于 0.1 的剔除显著性水平阈值，故系统提示没有变量将被剔除出模型，即 “No variables can be removed”。

接下来，继续对剩余的变量进行引入筛选。在剩余变量中，“Chips” 的  $P$  值最小（0.0268），小于变量引入显著性水平阈值，故系统提示该变量将被引入模型当中。

依此类推，本例一共进行了 4 次变量选择的过程。限于篇幅，对于中间筛选过程，这里不予一一详细列示。最终筛选结果即第 4 个步骤的输出结果如图 8-37 所示。

The STEPDISC Procedure					
Stepwise Selection: Step 4					
Statistics for Removal, DF = 2, 8					
Variable	Label	Partial R-Square	F Value	Pr > F	
Chips	芯片及微动	0.6119	6.31	0.0227	
Driver	功能及驱动	0.6060	6.15	0.0241	
Game	游戏性	0.5134	4.22	0.0561	
No variables can be removed.					
Statistics for Entry, DF = 2, 7					
Variable	Label	Partial R-Square	F Value	Pr > F	Toierance
Touch	外观及手感	0.3697	2.05	0.1988	0.3500
Compatibility	兼容性	0.0998	0.39	0.6921	0.3737
No variables can be entered.					
No further steps are possible.					

图 8-37 逐步选择法的第 4 个步骤

在第 4 个步骤中，首先对前面步骤引入模型的变量进行了剔除检验，得到的结论是没有变量会被剔除出模型；此外，对剩余的变量继续进行引入检验，但剩余变量的  $P$  值均大于引入显著性水平阈值，故系统提示没有变量将会被引入，逐步选择过程结束。

逐步选择过程结果后，系统会给出变量选择的相关概括信息，如图 8-38 所示。

The STEPDISC Procedure									
Stepwise Selection Summary									
Step	Number In Entered	Removed	Label	Partial R-Square	F Value	Pr > F	Wilks' Lambda	Pr < Lambda	
1	1 Game		游戏性	0.7833	18.08	0.0005	0.21666667	0.0005	
2	2 Chips		芯片及微动	0.5526	5.56	0.0268	0.09694067	0.0002	
3	3 Driver		功能及驱动	0.6060	6.15	0.0241	0.03819668	<.0001	
Average Squared Canonical Correlation									
Step	Number In Entered	Removed					Pr > ASCC		
1	1 Game						0.39166667	0.0005	
2	2 Chips						0.50241735	0.0056	
3	3 Driver						0.72045527	0.0003	

图 8-38 逐步选择法的概括信息

在图 8-38 所示的输出结果中，系统自动对用逐步选择法进行变量筛选的步骤进行了总结。结果发现，根据引入和剔除显著性水平阈值，进入模型的变量有“Game”、“Chips”、“Driver”3 个变量，依照其在判别分析中的重要性这 3 个变量进行了排序。序号越小，表明该变量最先被引入，其作用越大。对引入模型的变量进行检验，系统提示没有任何变量被剔除，即“Removed”列中没有任何变量。

逐步判别法主要用于判别分析的变量筛选。对于依据逐步判别法选择出来的重要变量，可根据前 3 节介绍的判别分析方法构建判别函数或判别规则以对样本进行归类。如本例分别考虑选择 DISCRIM 过程和 CANDISC 过程建立判别规则以进行归类，具体程序如下。

```
proc discrim data=Sasuser.Mouse_Discrim list;
  class Type;
  var Chips Driver Game;
run;
proc candisc data=Sasuser.Mouse_Discrim out=Mouse_Can_Out distance simple;
  class Type;
  var Chips Driver Game;
run;
```

运行程序后，得到的输出结果与前 3 节的分析结果类似。读者可自行观察输出结果，这里不予赘述。

### 8.5 本章小结

本章主要介绍了聚类分析和判别分析的基本原理及其在 SAS 系统的实现，主要内容简要回顾如下：聚类分析可根据一个指标或多个指标对样本数据进行分类，可采用的方法主要有系统聚类法、快速聚类法等；聚类分析也可对指标进行聚类，在 SAS 中可用 VARCLUS 过程实现；判别分析是依据已经分好类的样本进行归类，也可对没有分类的样

本进行归类；判别分析方法主要有距离判别、Bayes 判别、Fisher 判别等，可以依据判别函数或后验概率进行归类；对于判别分析中的变量选择，在 SAS 系统中可用 STEPDISC 过程进行逐步判别。



图 10-1 判别分析结果图



图 10-2 判别分析结果图

图 10-3 判别分析结果图

人们在研究某一个事物或现象的过程中，有时不仅单独考察某一方面的信息，而且还会把几方面的信息联合起来一并考察。如考察某项政策实施之后广大市民对该政策的民意反映，可以单独用一个民意指标“满意状况”来考察。如果把性别指标一并联合起来，考察不同性别人群对该项政策的满意状况，则是用两个指标来衡量一个事物。这两个指标的不同表现可以通过交叉的方式形成若干种状况，如男性对该项政策的满意状况、女性对该项政策的不满意状况等。把性别和满意状况这两个变量交叉联合起来，共同对所研究的问题展开研究，这个过程就叫做“交叉分析”。本章将要讲述的列联分析和对应分析是交叉分析的两种典型形式。

## 9.1 列联分析

对于定类或定序等定性数据的描述和分析，通常可使用列联表进行分析。本节主要介绍基于列联表 $\chi^2$ 检验的列联分析。

### 9.1.1 列联表

两个或两个以上的变量交叉形成的二维频数分布表格被称为“列联表”。如本章引言部分的不同性别对政策实施满意状况的交叉频数分布，设“性别”变量有“男”、“女”两种属性、“满意状况”变量有“满意”、“不满意”两种属性，得到的列联表如表 9-1 所示。

表 9-1

二维列联表的一般形式

人 数		满 意 状 况		合 计
		满 意	不 满 意	
性 别	男	128	117	245
	女	109	96	205
合 计		237	213	450

表 9-1 所示的列联表形式非常简单，只是把两个变量的不同属性进行交叉，计算出各种属性组合的频数，作为表格中的主要数据。

从列联表中可以清楚看到所有人和不同性别的人对该项政策的不同观点分布状况，同时也可以看到所有满意状况及其两种属性表现的性别分布状况。

列联表中变量的属性或取值通常也被叫做“水平”，如性别变量有“男”、“女”两个水平，“满意状况”变量有“满意”和“不满意”两个水平。

列联表行变量的水平个数一般用 R 表示，列变量的水平个数一般用 C 表示，那么一个 R 行 C 列的频数分布表被叫做 R × C 列联表，如表 9-2 所示。

表 9-2 R × C 列联表

频 数		列 变 量				行 合 计
		水平 1	水平 2	L	水平 c	
行变量	水平 1	$f_{11}$	$f_{12}$	L	$f_{1c}$	$\sum_{i=1}^c f_{1i}$
	水平 2	$f_{21}$	$f_{22}$	L	$f_{2c}$	$\sum_{i=1}^c f_{2i}$
	M	M	M	M	M	M
	水平 r	$f_{r1}$	$f_{r2}$	L	$f_{rc}$	$\sum_{i=1}^c f_{ri}$
列 合 计		$\sum_{i=1}^r f_{i1}$	$\sum_{i=1}^r f_{i2}$	L	$\sum_{i=1}^r f_{ic}$	$\sum_{i=1}^r \sum_{j=1}^c f_{ij}$

R × C 列联表中各元素  $f$  就是行列变量进行交叉分类得到的观测值个数所形成的频数分布。行合计表示行变量每个水平在列变量不同水平交叉分类的观测值总数，列合计表示列变量每个水平在行变量不同水平交叉分类的观测值总数。行合计加总应当等于列合计加总，记为总计频数。



例 9-1

某公司欲推行一套新的工资改革方案，为了考查该方案的合理性，提高改革方案在公司各部门推行之后的实际效果，特抽查了市场部、客户服务部、发展战略部、综合部、研发中心等 5 个部门共 220 名员工，以了解员工对该套工资改革方案的态度。以该例数据编制的列联表如表 9-3 所示。

表 9-3 各部门员工对工资改革的态度

人 数		部 门					合 计
		发展战略部	客户服务部	市场部	研发中心	综合部	
态 度	反 对	25	15	20	27	29	116
	支 持	16	21	23	22	22	104
合 计		41	36	43	49	51	220

SAS 系统中有两种数据预处理方式可以输出列联表。

第一种数据预处理方式是以原始调查数据作为数据集，然后利用前面章节介绍过的 FREQ 过程（详见 2.2.4 小节）制表得到列联表；第二种数据预处理方式是输入表 9-3 所示的交叉分组数据，仍然利用 FREQ 过程并在 FREQ 过程中通过 WEIGHT 语句指定交叉分组频数作为权数，也可得到列联表。

第一种数据预处理方式的具体数据格式如图 9-1 所示。

图 9-1 所示的数据内容详见 Salary\_Reform.sas7bdatt 数据集。

**STEP 1** 进入 SAS/Analyst，打开 Salary\_Reform.sas7bdat 数据集，选择系统菜单 “Statistics → Table Analysis”，弹出图 9-2 所示的列联表分析对话框。可以通过该对话框绘制列联表。



ID	Department	Attitude
1	市场部	反对
2	综合部	反对
3	综合部	反对
4	研发中心	反对
5	发展战略部	支持
6	发展战略部	反对
7	研发中心	反对
8	发展战略部	支持
9	研发中心	支持
10	市场部	支持
11	综合部	反对
12	综合部	支持
13	客户服务部	反对
14	市场部	反对
15	研发中心	支持
16	综合部	反对
17	综合部	支持
18	研发中心	反对
19	客户服务部	反对
20	发展战略部	反对
21	市场部	支持
22	客户服务部	支持

图 9-1 列联表的原始数据预处理方式

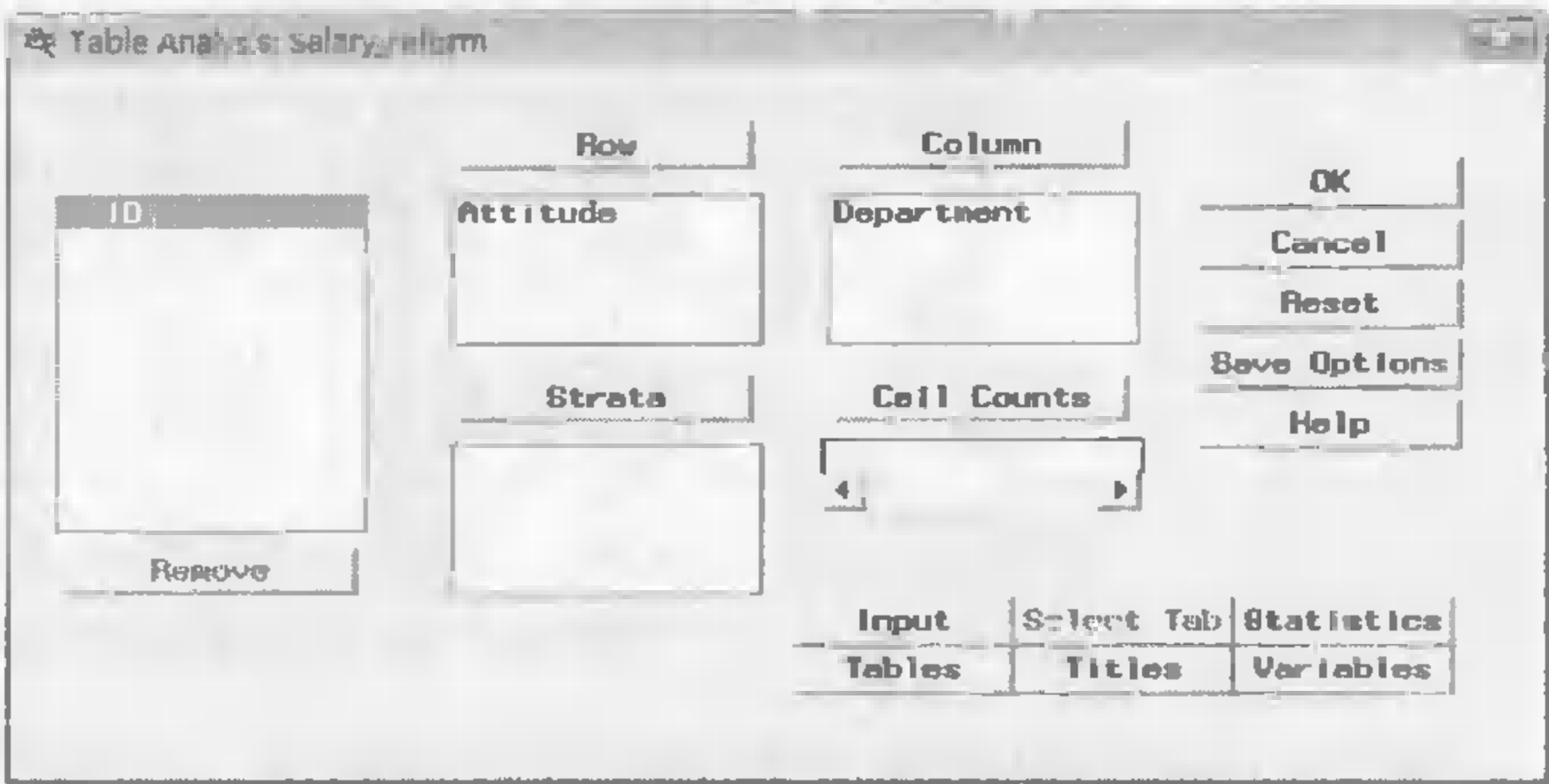


图 9-2 SAS/Analyst 的列联表对话框

**STEP 2** 在图 9-2 所示的变量选择区域中选中 “Attitude” 变量，单击 “Row” 按钮把其指定为行变量；选中 “Department” 变量，单击 “Column” 按钮把其指定为列变量。此外，“Tables” 按钮还可以指定列联表中输出的信息。单击 “Tables” 按钮，弹出表格输出选项对话框，如图 9-3 所示。

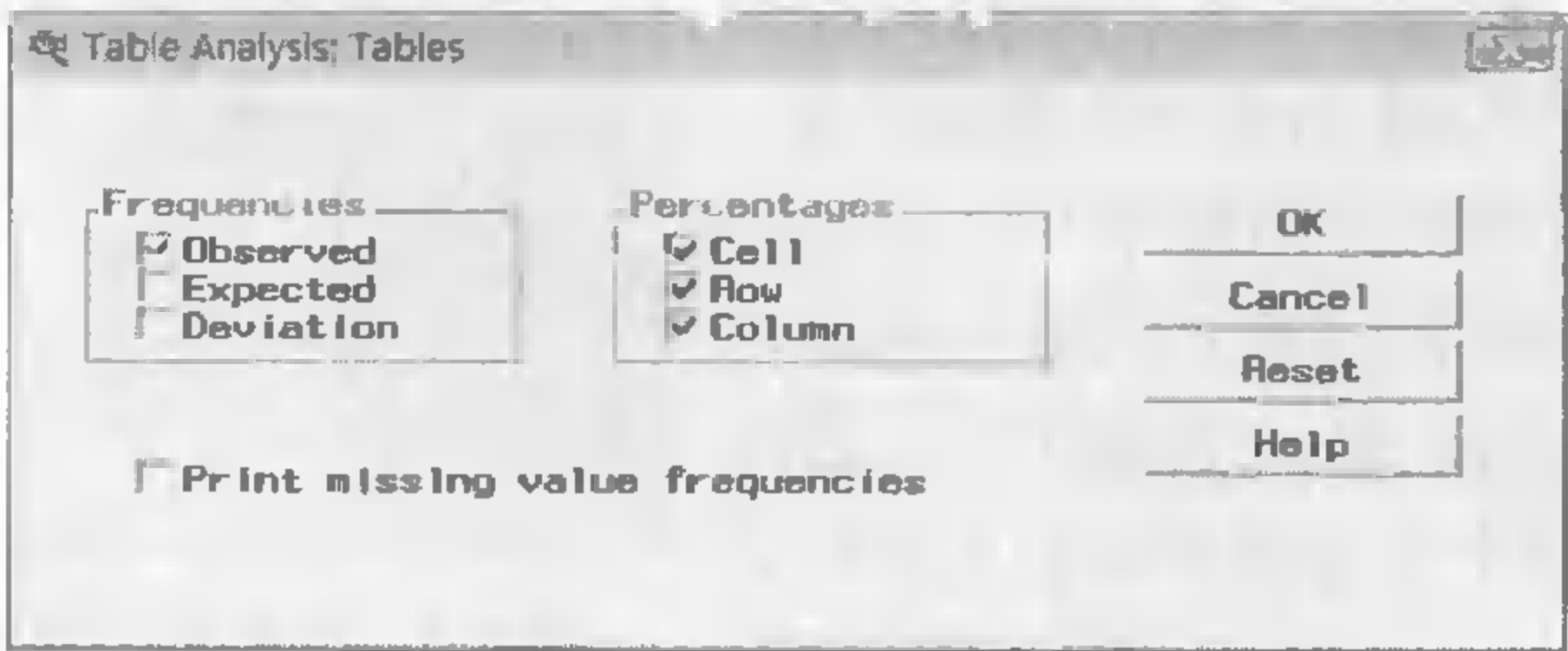


图 9-3 表格输出选项对话框

**STEP 3** 图 9-3 左边的 “Frequencies” 分栏下的复选框可被用于选择交叉分组频数的观测值 (Observed)、理论期望值 (Expected) 和观测值与期望值之差 (Deviation)。该分栏的具体内容详见 7.1.2 小节。

右边的 “Percentages” 分栏下的复选框被用于指定列联表依据交叉频数输出相关的频率，其中 “Cell” 复选框表示百分比，即每个交叉分组频数与总计数目（即样本量）的比值；复选框 “Row” 表示行百分比，即每个交叉分组频数与行合计数目的比值；复选框 “Column” 表示列百分比，即每个交叉分组与列合计数目的比值。最下面的 “Print missing value frequencies” 表示输出缺失值频数。

**STEP 4** 本例选择输出观测值、百分比及行、列百分比，单击 “OK” 按钮返回图 9-2 所示的对话框。在该对话框中，单击 “OK” 按钮，可得到本例的列联表，如图 9-4 所示。

可以看到，在图 9-4 中的交叉分类结果中一共有 4 行数字，表格的左上角标注了每行数字所代表的意思。第 1 行表示交叉分类的频数 (Frequency)，依次往下的 3 行是百分数（省略了 % 号），分别是百分比 (Percent)、行百分比 (Row Pct)、列百分比 (Col Pct)，每行百分比的总和与每列百分比的总和均为 100%。

The FREQ Procedure						
Table of Attitude by Department						
Attitude(Attitude)	Department(Department)					
Frequency Percent Row Pct Col Pct	发展战略部	客户服务部	市场部	研发中心	综合部	Total
反对	25 11.36 21.55 60.98	15 6.82 12.93 41.67	20 9.09 17.24 46.51	27 12.27 23.28 55.10	29 13.18 25.00 56.86	116 52.73
支持	16 7.27 15.38 39.02	21 9.55 20.19 58.33	23 10.45 22.12 53.49	22 10.00 21.15 44.90	22 10.00 21.15 43.14	104 47.27
Total	41 18.64	36 16.36	43 19.55	49 22.27	51 23.18	220 100.00

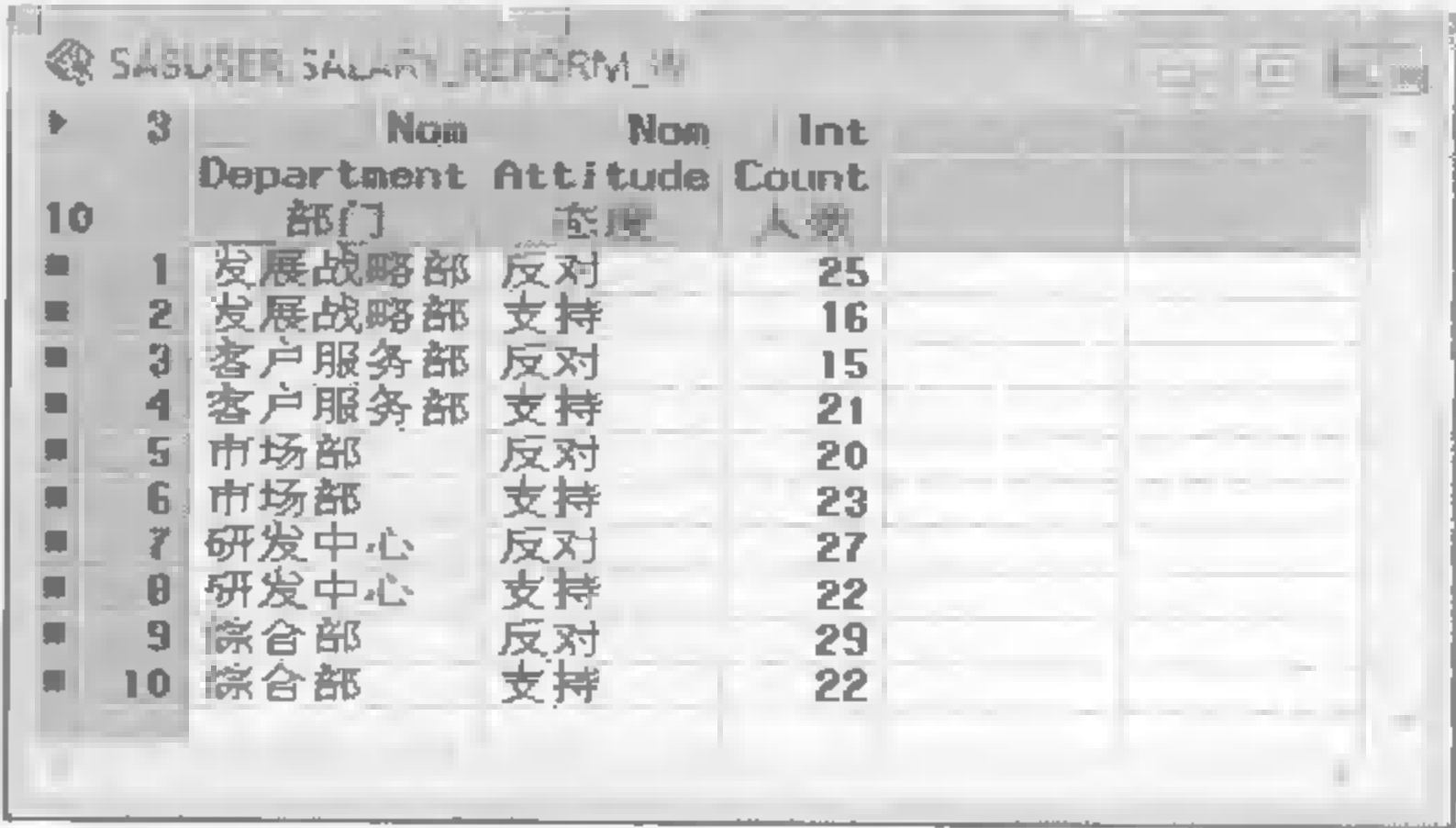
图 9-4 SAS 系统输出的列联表

图 9-4 的输出结果也可用 FREQ 过程实现，具体程序如下。

```
proc freq data=Sasuser.Salary_Reform;  
  table Atti de * Department; /*指定 Atti de 为行变量，Department 为列变量*/  
run;
```

运行程序后可得到与图 9-4 相同的输出结果。  
第二种数据预处理格式如图 9-5 所示。

该种数据预处理方式应用交叉分类汇总的形式进行数据输入，即在向 SAS 系统录入数据之前，已经依据行列变量的水平统计出各种水平交叉分类出现的人数或次数，以数据汇总的方式录入到 SAS 数据集（详见 Salary\_Reform\_W.sas7bdat）中。这样作为出现人数或次数的那个变量便衡量了其对应交叉情形出现的情况，因此成为各种交叉情形的权重。所以，在第二种数据预处理方式下绘制列联表时，要注意权数的应用。



3	Department	Attitude	Count
10	部门	态度	人数
1	发展战略部	反对	25
2	发展战略部	支持	16
3	客户服务部	反对	15
4	客户服务部	支持	21
5	市场部	反对	20
6	市场部	支持	23
7	研发中心	反对	27
8	研发中心	支持	22
9	综合部	反对	29
10	综合部	支持	22

图 9-5 列联表加权形式的数据预处理方式

STEP 1) 进入 SAS/Analyst，打开 Salary\_Reform\_W.sas7bdat 数据集，选择系统菜单“Statistics → Table Analysis”，弹出图 9-2 所示的列联表分析对话框。

STEP 2) 在该对话框中，同样把变量“Attitude”指定为行变量，变量“Department”指定为列变量。然后选中“Count”变量，单击“Cell Coun”按钮，把其指定为权数对行列变量进行加权，其余设置同第二种数据预处理方式。单击“OK”按钮之后，可以得到图 9-4 所示的同样结果。

同样利用 FREQ 过程也可对第二种数据预处理方式绘制列联表，但是要注意用 WEIGHT 语句把计数的变量作为权数进行加权，具体程序如下。

```
proc freq data=Sasuser.Salary_Reform_W;  
  table Atti de * Department; /*指定 Atti de 为行变量，Department 为列变量*/  
  weight Count; /*指定 Count 为权数进行加权*/  
run;
```

运行程序后的结果与图 9-4 相同。

此外，FREQ 过程的 TABLE 语句也提供了类似于图 9-3 所示的输出调整关键字。

- EXPECTED: 输出理论期望频数。
- DEVIATION: 输出观测值与期望值之差。
- NOFREQ: 不输出交叉分类频数。
- NOPERCENT: 不输出百分比。
- NOROW: 不输出行百分比。
- NOCOL: 不输出列百分比。

在列联表中不想输出百分比和行列百分比的程序如下。

```
proc freq data=Sasuser.Salary_Reform_W;  
  table Attitude * Department /nopercnt norow nocol;  
  weight Count;  
run;
```

运行程序后，可以得到类似于表 9-3 所示的结果。



选项关键字放置在“/”之后。

### 9.1.2 列联表的分布

列联表中的分布有两种：一种如表 9-3 或图 9-4 所示，能够直接从样本数据中获得的交叉分类分布，即可以直接观测得到，其行、列合计分别称为行边缘分布和列边缘分布；另一种是期望值的分布，不能直接观测出来，可以通过样本数据和相关理论依据进行计算。

以例 9-1 为例，如果要想了解不同部门的员工对工资改革方案的态度是否存在显著差异，在没有显著差异的假定条件下，各部门员工不同态度的分布即为列联表的理论分布。据此可以计算出各部门态度人数的理论期望频数值。

在本例中，持“反对”态度的员工总人数为 116 人，持“赞成”态度的员工总人数为 104 人。因此对整个公司而言，工资改革方案的反对率应当为： $116/220 = 0.5273$ ，即 52.73%；工资改革方案的支持率应当为： $104/220 = 0.4727$ ，即 47.27%。

现假定各部门对工资改革的态度没有差异，故各部门反对该项政策的人数应当为该部门被调查人数乘以反对率，支持该项政策的人数应当为该部门人数乘以支持率。如发展战略部一共有员工 41 人，持支持态度的理论人数应当为  $41 \times 47.27\% = 19.38$  人。由此计算出来的人数便是列联表的期望值，计算过程及期望值分布如表 9-4 所示。

表 9-4 列联表的期望分布

期 望 人 数		部 门					合 计
		发展战略部	客户服务部	市 场 部	研 发 中 心	综 合 部	
态度	反对	$41 \times 0.5273 = 21.62$	$36 \times 0.5273 = 18.99$	$43 \times 0.5273 = 22.67$	$49 \times 0.5273 = 25.84$	$51 \times 0.5273 = 26.89$	116
	支持	$41 \times 0.4727 = 19.38$	$36 \times 0.4727 = 17.02$	$43 \times 0.4727 = 20.33$	$49 \times 0.4727 = 23.16$	$51 \times 0.4727 = 24.11$	104
合 计		41	36	43	49	51	220

表 9-4 所示的期望分布也可由 SAS 系统自动进行计算。

**STEP 1** 进入 SAS/Analyst, 打开 Salary\_Reform.sas7bdat 或 Salary\_Reform\_W.sas7bdat( 注意需加权 ) 数据集, 选择系统菜单 “Statistics→Table Analysis”, 弹出图 9-2 所示对话框。在该对话框中, 单击 “Tables” 按钮, 弹出图 9-3 所示的 Tables 对话框。在该对话框中的 “Frequencies” 分栏下, 选中 “Expected” 复选框, 表示在列联表中输出期望值。

**STEP 2** 本例同时把 “Observed” (观测值)、“Expected” 和 “Deviation” (观测值与期望值之差) 选中, 并且把右边 “Percentages” 分栏下的所有复选框选中, 单击 “OK” 按钮返回图 9-2 所示对话框。在该对话框中, 单击 “OK” 按钮可得到包含观测值分布、期望值分布等所有信息的列联表, 如图 9-6 所示。

The FREQ Procedure						
Table of Attitude by Department						
Attitude(态度)	Department(部门)					
Frequency						
Expected						
Deviation						
Percent						
Row Pct						
Col Pct	发展战略部	客户服务部	市场部	研发中心	综合部	Total
反对	25 21.618 3.3818 11.36 21.55 60.98	15 18.982 -3.982 6.82 12.93 41.67	20 22.673 -2.673 9.09 17.24 46.51	27 25.836 1.1636 12.27 23.28 55.10	29 26.891 2.1091 13.18 25.00 56.86	116   52.73
支持	16 19.382 -3.382 7.27 15.38 39.02	21 17.018 3.9818 9.55 20.19 58.33	23 20.327 2.6727 10.45 22.12 53.49	22 23.164 -1.164 10.00 21.15 44.90	22 24.109 -2.109 10.00 21.15 43.14	104   47.27
Total	41 18.64	36 16.36	43 19.55	49 22.27	51 23.18	220 100.00

图 9-6 包含观测值分布、期望值分布等全部信息的列联表

在图 9-6 所示的列联表中, 一共有 6 行数字信息, 从上至下分别是观测值频数分布、期望值频数分布、观测值与期望值之差、百分比、行百分比、列百分比。前 3 项为计数数据, 后 3 项为百分数 (单位为%)。

利用 FREQ 过程也可得到图 9-6 所示的结果, 具体程序如下。

```
proc freq data=Sasuser.Salary_Reform_W;  
  table Atti de * Department /expected deviation;  
  weight Count;  
run;
```

9.1.3  $\chi^2$  分布与  $\chi^2$  检验

从 7.1.2 小节中可以得知, 列联表的分布主要有观测值分布和期望值分布, 同时也计算了观测值与期望值之间的偏差。现设  $f_{ij}^o$  表示各交叉分类频数的观测值,  $f_{ij}^e$  表示各交叉分类频数的期望值, 则各交叉分类频数观测值与期望值的偏差为  $f_{ij}^o - f_{ij}^e$ , 故  $\chi^2$  统计量为:

$$\chi^2 = \sum_{i=1}^r \sum_{j=1}^c \frac{(f_{ij}^o - f_{ij}^e)^2}{f_{ij}^e}$$

当样本量较大时,  $\chi^2$  统计量近似服从自由度为  $(R-1)(C-1)$  的  $\chi^2$  (卡方) 分布。 $\chi^2$  值与期望值、观测值和期望值之差有关,  $\chi^2$  值越大表明观测值与期望值的差异越大。因此, 可以由此对 9.1.2 小节中计算期望值的假设进行  $\chi^2$  检验。

上一节在计算期望值分布时,假定各部门对工资改革的态度没有差异,各部门对该项改革方案的支持率或反对率的 $P$ 值均相等,即员工对该项改革方案的态度与其所在部门无关,行、列变量之间是独立的。据此可以提出原假设和备择假设。

$H_0$ : 部门与对改革方案的态度独立;

$H_1$ : 部门与对改革方案态度不独立

然后可利用 SAS 系统对例 9-1 进行 $\chi^2$ 检验。

**STEP 1)** 进入 SAS/Analyst, 打开 Salary\_Reform.sas7bdat 或 Salary\_Reform\_W.sas7bdat(注意需加权)数据集, 选择系统菜单“Statistics→Table Analysis”, 在弹出的图 9-2 所示对话框中单击“Statistics”按钮, 弹出统计量对话框, 如图 9-7 所示。

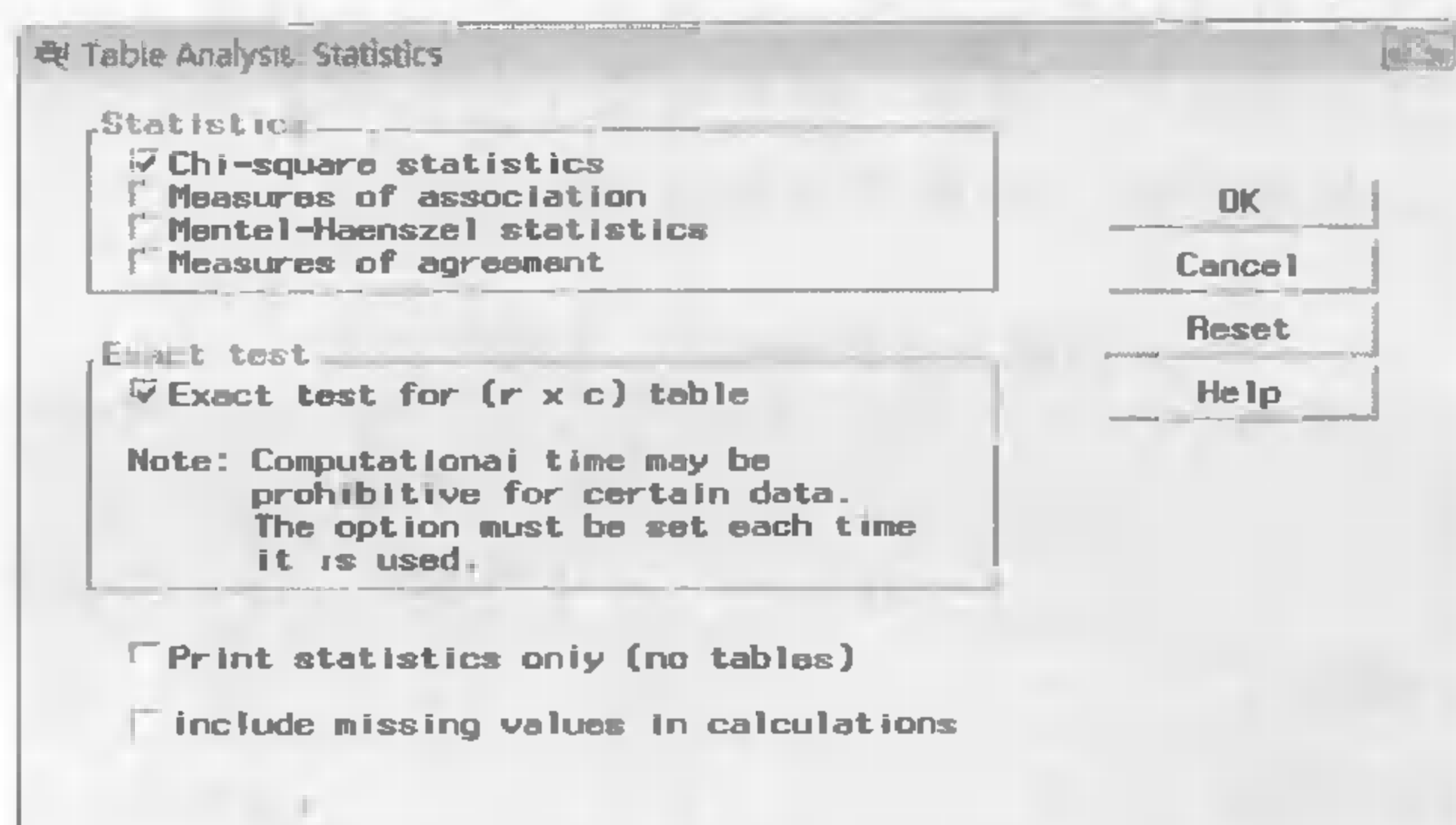


图 9-7 列联分析的统计量对话框

**STEP 2)** 在统计量对话框中的“Statistics”分栏下,选中“Chi-square statistics”复选框,表示输出 $\chi^2$ 统计量并进行 $\chi^2$ 检验;在“Exact test”分栏下,选中“Exact test for (r × c) table”复选框,表示输出 Fisher 精确检验。单击“OK”按钮返回上一级对话框,在该对话框中,单击“OK”按钮,可以得到列联表及检验结果,如图 9-8 所示。

从图 9-8 可以得出, $\chi^2$ 统计量的值为 4.013 3,其对应的 $P$ 值(“Prob”)为 0.404 2。假定理论显著性水平 $\alpha = 0.05$ ,则 $P$ 值远远大于 $\alpha$ 。因此,没有充分理由拒绝行、列变量即部门与态度之间独立的原假设。

SAS 系统还提供了其他几种形式的 $\chi^2$ 统计量进行检验,如似然比 $\chi^2$ 统计量、Mantel-Haenszel $\chi^2$ 统计量等,其检验结果均表明在一定的显著性水平条件下没有充分理由拒绝原假设。

Statistics for Table of Attitude by Department			
Statistic	DF	Value	Prob
Chi-Square	4	4.0133	0.4042
Likelihood Ratio Chi-Square	4	4.0260	0.4025
Mantel-Haenszel Chi-Square	1	0.0603	0.8061
Phi Coefficient		0.1351	
Contingency Coefficient		0.1338	
Cramer's V		0.1351	
Fisher's Exact Test			
Table Probability (P)		6.543E-05	
Pr <= P		0.4105	

图 9-8 列联表检验结果

由于 $\chi^2$ 统计量是一个近似的统计量,因此 SAS 系统还提供了 Fisher 精确检验的过程。该统计量服从超几何分布,在样本量较大时运算量非常大,所以在图 9-7 所示的对话框中指定该统计量进行检验时,系统会提示用户该检验会花大量时间。每次分析均需用户自行指定是否使用精确检验方法。

Fisher 精确检验结果也表明没有充分理由拒绝行、列变量即部门与态度之间独立的原假设。

对于上述分析过程,利用 FREQ 过程中 TABLE 语句选项的“CHISQ”关键字可进行 $\chi^2$ 检验,利用关键字“EXACT”可进行 Fisher 精确检验,具体程序如下:

```
proc freq data=Sasuser.Salary_Reform_W;
  table Atti de * Department / chisq exact expected;
  weight Count;
run;
```

运行程序后可得到图 9-8 所示的结果。

9.1.4 列联表中的关联度分析

通过  $\chi^2$  检验，如果得到行、列变量之间不是独立的结论，则列联分析中还可以对行、列变量之间的相关性进行测量。



例 9-2

某公司推销一种业务。为考察不同收入消费人群对该项业务的购买意向，该公司进行了调查，调查结果（数据详见 Purchase.sas7bdat）如表 9-5 所示，试分析收入和购买意向之间的关系。

表 9-5 不同收入消费群体的业务购买意向

人数 (Count)		收入 (Income)			合 计
		0-低收入	1-中收入	2-高收入	
购买意愿 (Propensity)	2-愿意购买	68	65	37	121
	0-不愿购买	23	29	69	170
	1-暂无打算	23	11	15	49
合 计		114	121	105	340

STEP 1) 进入 SAS/Analyst 后，打开 Purchase.sas7bdat（注意把 Count 变量作为权数），按照例 9-1 分析和检验的步骤可得到图 9-9 所示的检验结果。

多种  $\chi^2$  统计量检验和 Fisher 精确检验的结果表明，在  $\alpha = 0.05$  的显著性水平下，拒绝原假设，即不认为行变量“购买意愿”与列变量“收入”之间是独立的。行、列变量存在相关关系，二者之间的相关程度同样可以在列联表分析中被计算出来。

因行、列变量均为定性变量，为了更好地衡量行、列变量之间相互影响的关系，把收入按照低、中、高以及把购买意向按照购买意愿强烈程度标注顺序，如表 9-5 所示。

STEP 2) 在图 9-7 所示对话框中，选中“Statistics”分栏下的“Measures of association”复选框，便可得到行、列变量之间的关联性分析结果，如图 9-10 所示。

Statistics for Table of Propensity by District			
Statistic	DF	Value	Prob
Chi-Square	4	43.4315	<.0001
Likelihood Ratio Chi-Square	4	43.5675	<.0001
Mantel-Haenszel Chi-Square	1	25.7259	<.0001
Phi Coefficient		0.3574	
Contingency Coefficient		0.3366	
Cramer's V		0.2527	
Fisher's Exact Test			
Table Probability (P)		1.050E-13	
Pr <= P		8.271E-09	
Sample Size = 340			

图 9-9 例 9-2 的列联表检验结果

Statistics for Table of Propensity by Income		
Statistic	Value	ASE
Gamma	-0.4023	0.0659
Kendall's Tau-b	-0.2650	0.0459
Stuart's Tau-c	-0.2517	0.0435
Somers' D C/R	-0.2785	0.0484
Somers' D R/C	-0.2522	0.0438
Pearson Correlation	-0.3021	0.0503
Spearman Correlation	-0.2977	0.0509
Lambda Asymmetric C/R	0.1781	0.0495
Lambda Asymmetric R/C	0.1882	0.0546
Lambda Symmetric	0.1825	0.0430
Uncertainty Coefficient C/R	0.0584	0.0172
Uncertainty Coefficient R/C	0.0645	0.0191
Uncertainty Coefficient Symmetric	0.0613	0.0181
Sample Size = 340		

图 9-10 行列变量关联性分析结果

图 9-10 列示了很多种常用的相关系数，如第 4 章讲述过的 Pearson 相关系数、Spearman 相关系数、Hendall’s Tau-b 等级相关系数等。除此之外，还有 Gamma 系数、Stuart’s Tau-c 系数等均在一定程度上衡量了行、列变量之间的相关性。图 9-10 中的“Value”列表示对应系数的统计量值，“ASE”列表示渐进标准误差，依据给定的理论显著性水平和“ASE”可以计算出对应系数的置信区间。在本例中，如相关系数 Kendall’s Tau-b 的值为-0.2650，表明“收入”和“购买意愿”之间存在低度负相关性，即收入越高，可能购买意愿越不强烈。

**STEP 3** 利用 SAS 编程语言的 FREQ 过程，在该过程的 TABLE 语句选项中加入“MEASURES”关键字，即可得到图 9-10 所示的相关性分析结果，具体程序如下。

```
proc freq data=Sasuser.Salary_Reform_W;  
  table Attitude * Department / measures chisq exact expected;  
  weight Count;  
run;
```

9.1.5  $\chi^2$  分布的期望值准则

在 7.1.3 小节中，已经看到  $\chi^2$  检验是一种近似检验，依据观测值和期望值计算出来的统计量在大样本的情况下近似服从  $\chi^2$  分布。因此，要求在进行列联表检验过程中，样本量应当足够大，而且每个交叉分类的期望频数不能偏小，否则  $\chi^2$  检验可能会出错。

进行  $\chi^2$  检验时， $\chi^2$  分布的期望值准则主要有两条。

- ❶ 当数据交叉分类为两类时，要求每一类别的期望值不小于 5。
- ❷ 当数据交叉分类为两个以上类别时，期望值小于 5 的比例不应超过 20%，否则应把期望值小于 5 的类别与相邻的类别合并。

在表 9-6 所示的列联表中，有一个期望值数值为 4，小于 5，依据期望值准则，则不能够进行  $\chi^2$  检验。

表 9-6 两个类别的列联表分布

产品合格情况	观测值 ( $f^o$ )	期望值 ( $f^e$ )
合格	123	115
不合格	6	4

当数据交叉分类为两个以上类别且期望值小于 5 的比例超过 20%时，其列联表分布如表 9-7 所示。

表 9-7 两个以上类别的列联表分布

产品质量分类	观测值 ( $f^o$ )	期望值 ( $f^e$ )
A	123	115
B	120	132
C	78	87
D	23	45
E	8	4
F	7	3

表 9-7 中一共有 6 个分类，其 20%为： $6 \times 20\% = 1.2$ ，但其期望值小于 5 的分类个数为 2，超过了 20%的数目。对于此种情况，需把期望值小于 5 的类别与相邻的类别合并，即把 E 类和 F 类合并，如表 9-8 所示。

表 9-8 两个以上类别的合并列联表分布

产品质量情况	观测值 ( $f^o$ )	期望值 ( $f^e$ )
A	123	115
B	120	132
C	78	87
D	23	45
E 和 F 合并	15	7

在进行分析时，用经过合并处理的、形如表 9-8 所示的表格进行  $\chi^2$  检验即可。

## 9.2 对应分析

$\chi^2$  检验可以对行、列变量之间是否有关联进行检验，但是行、列变量之间的关联性具体是如何相互作用的，通过列联表的  $\chi^2$  检验则很难进行判断。

因此，可进一步对行、列变量相互关系进行深入分析，对应分析便是进一步研究交叉分析的一种方式，它利用数据降维方法直观明了地分析行、列变量之间的相互关系。对应分析是在 R 型（样本）和 Q 型（变量）因子分析的基础上发展起来的一种多元统计分析方法，它不仅仅关注行变量或列变量本身的关系，更加关注的是行、列变量之间的相互关系。

对应分析可根据所分析变量的数目分为简单对应分析和多重对应分析。简单对应分析主要用于两个分类变量之间关系的研究，而多重对应分析用于分析 3 个或更多变量之间的关系。由于对应分析可以明确划定行、列变量之间的影响关系，因此在市场研究、市场细分、产品定位等领域使用尤为广泛。

### 9.2.1 对应分析的基本思想

对应分析的基本思想是将一个列联表的行、列变量的比例结构以散点形式在较低维的空间中表示出来。对应分析省去了因子选择和因子轴旋转等中间运算过程，可以从因子载荷图上对样品进行直观的分类，而且能够指示分类的主因子及分类的依据。此外，对应分析最主要是要得到能够同时反映众多样本和众多变量的对应分析图。

为了实现上述基本思想，对应分析通常先找到能够代表行、列变量的行得分与列得分。行得分与列得分互为对方的加权均值，它们之间具有相关性。行、列得分在一个数据中可以得到多组数值，可以根据各组行列得分绘制多个二维散点图，然后把各个散点图堆叠起来，最终形成对应分析图。

为了直观明了地描述行、列变量之间的对应关系，通常选择两对行、列得分，通过两张散点图叠加得到对应分析图。在对应分析中，把衡量行列关系强度的指标称为“惯量”，其累积所占的百分比成为选取行、列得分对的数目的主要依据。同时，惯量所占比例也成为衡量某对行、列得分在对应分析图中的重要性的依据。

### 9.2.2 对应分析的步骤和过程

在进行对应分析之前，应当依据列联分析的知识先判定行、列变量之间是否存在相关性（详见第9.1节）。通过检验，如果存在相关性，为进行更进一步的分析研究，可进行对应分析，以找出行、列变量之间的具体影响关系。对应分析最主要的步骤是依据惯量的累积百分比确定选取行、列得分的数目，然后绘制对应分析图，从图中找出行、列变量之间的对应关系。

在 SAS 系统中，可用 CORRESP 过程进行对应分析，其主要语法如下。

```
PROC CORRESP <选项>;
  TABLES <行变量,> 列变量;
  VAR 变量;
  BY 变量;
  ID 变量;
  SUPPLEMENTARY 变量;
  WEIGHT 变量;
```

其中 TABLES 语句主要用于指定行、列变量。可以只指定列变量，也可指定多个行、列变量，但是要注意行、列变量之间务必要以“,”隔开。

SUPPLEMENTARY 语句用于指定代表图形中行、列空间点的变量，但这些代表变量在列联表中不能够用于行、列变量的定位。该语句指定变量的观测值在进行简单对应分析时被忽略，但在进行多重对应分析时必须用于计算余弦平方。此外，该语句指定的变量应当来自于 TABLES 或 VAR 语句中指定的变量。

VAR、BY、ID、WEIGHT 语句的用法和功能与前面章节介绍的一样。

CORRESP 的选项也非常多，常用的主要选项如下。

- DATA=: 指定用于分析的数据集。
- OUT=: 指定存储惯量、行列得分等统计量的输出数据集。
- OUTC=: 指定存储含有输出对应分析图坐标（即行列得分）的数据集，该输出数据集可被用于绘制对应分析图。
- OUTF=: 指定存储含有输出频数的数据集。
- DIMENS=: 指定维度或坐标轴的数目（即指定行、列得分对的数目）。
- MCA: 进行多重对应分析。
- PROFILE=: 对行、列得分进行标准化。
- ALL: 输出所有的分析过程与信息。
- OBSERVED: 输出行、列变量列联表的观测值。
- EXPECTED: 输出列联表的期望值。
- DEVIATION: 输出列联表观测值与期望值之差。

在 SAS 系统中，使用 CORRESP 过程可以得到对应分析的过程与相应的输出信息。要想得到依据行、列得分绘制的对应分析图，还需要用 %PLOTIT 宏进行图形绘制。

- %PLOTIT 宏在对应分析过程中的主要语法如下。
- %PLOTIT (选项)

该宏的主要选项关键字如下。

- DATA=: 指定用于绘图的数据集。
- DATATYPE=: 指定数据集的类型。在对应分析中，应当指定“corresp”类型。
- PLOTVARS=: 指定代表纵、横坐标值的变量，变量之间用空格隔开。
- HREF=: 指定某个得分绘制水平参考线。
- VREF=: 指定某个得分绘制垂直参考线。



在使用%PLOTIT 绘制对应分析图时，如果 DATA 关键字指定的数据集是由 CORRESP 过程的 OUT 关键字生成的数据集，则需在 PLOTVARS 关键字中指定 DIM1、DIM2 作为图形的纵轴和横轴；如果 DATA 关键字指定的数据集是由 CORRESP 过程的 OUTC 关键字生成的数据集，则 PLOTVARS 关键字可被省略。



例 9-3

某品牌手机研发部门现新开发一款手机，为了使该款新手机定位更加准确，快速找准消费群体，该部门委托市场研究公司进行了调查，主要了解不同收入群体在购买手机时最主要考虑的影响因素。通过发放问卷，在全国范围内收集到 7 934 份有效样本，调查汇总数据（详见 CellPhone.sas7bdat）如表 9-9 所示。试对调查结果进行对应分析。

表 9-9 某品牌手机功能与不同收入消费者群体的相关性调查结果

人 数		影响因素（Element）											合计
		价格	待机时间	外观款式	功能	IO 接口	网络兼容性	存储容量	品牌	摄像头	质量	操作系统	
月收入 (Income)	1 000 元以下	658	374	528	332	50	104	143	363	90	626	52	3 320
	1 001~3 000 元	665	406	522	323	36	86	165	387	110	637	46	3 383
	3 001~5 000 元	139	108	119	76	9	24	46	93	26	147	19	806
	5 001~8 000 元	26	19	26	13	8	10	5	20	16	27	10	180
	8 001~10 000 元	13	10	14	11	9	9	8	12	13	15	12	126
	10 000 元以上	13	14	11	6	11	15	7	11	6	18	7	119
合计		1 514	931	1 220	761	123	248	374	886	261	1 470	146	7 934

由于 CellPhone.sas7bdat 数据集中的数据都是根据问卷选项用数字表示的，因此为了方便地区分各个数字对应的选项，可以依据原始问卷的问题和选项为数据集中的变量值挂上标签（详见 1.4.1 小节）。

此外，由于英文 SAS 绘图系统对全角中文字符的兼容性不强，因此为避免绘制对应分析图时出现中文乱码，本例把上述影响因素和月收入变量的具体内容转化为英文。

本例具体程序如下。

```
proc format;                                /*调用 FORMAT 过程，为变量值指定标签*/
  value Income_FMT                          1='Under 1000RMB' /*为收入变量指定值标签*/
    2='1001-3000RMB'
    3='3001-5000RMB'
```

```
4='5001-8000RMB'
5='8001-10000RMB'
6='Upper 10000RMB';
    value Element_FMT          1='Price'          /*为影响因素变量指定值标签*/
2='Duration'
3='Style'
4='Function'
5='IO'
6='Net'
7='Memory'
8='Brand'
9='Camera'
10='Quality'
11='OS';
run;
proc corresp data=Sasuser.CellPhone outc=CellPhone_Out; /*调用 CORRESP 过程，指定输出含有行、
列得分的 CellPhone_Out 临时数据集*/
    tables Income,Element; /*指定行、列变量*/
    format Income Income_FMT. Element Element_FMT.; /*为行、列变量挂上变量值的标签*/
    weight Count; /*本例数据为汇总数据，以人数进行加权*/
run;
%plotit(data=CellPhone_Out, datatype=corresp) /*使用 SAS 宏依据含有得分的数据集绘制对应分析图*/
```

运行程序后，可在“Output”窗口中输出大量结果，并在“Graph”窗口中得到对应分析图。首先查看“Output”窗口中的惯量和卡方统计量构成表，如图 9-11 所示。

The CORRESP Procedure									
Inertia and Chi-Square Decomposition									
Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	17	34	51	68	85
0.17500	0.03062	242.968	84.70	84.70	*****				
0.05837	0.00341	27.032	9.42	94.13	***				
0.03681	0.00136	10.753	3.75	97.88	*				
0.02364	0.00056	4.434	1.55	99.42					
0.01447	0.00021	1.660	0.58	100.00					
Total	0.03615	286.847	100.00						
Degrees of Freedom = 50									

图 9-11 惯量和  $\chi^2$  统计量构成表

在图 9-11 中，第 1 列“Singular Value”为奇异值，即行、列变量进行因子分析所得综合变量的典型相关系数，数值上等于惯量的平方根。第 2 列“Principal Inertia”即为惯量。第 3 列“Chi-Square”为  $\chi^2$  统计量，其值与列联分析中计算的  $\chi^2$  值相等。本例计算出总的  $\chi^2$  统计量为 286.847，远远大于  $\alpha = 0.05$  条件下对应的临界值，表明行、列变量之间有较强的相关性。第 4、5 列“Percent”、“Cumulative Percent”分别为惯量比例与累积惯量比例。

从图 9-11 的惯量比例来看，第一维度的惯量所占比例为 84.70%，其重要性非常大。其余惯量所占比例均比较小，这可以从图 9-11 右边的星线图中可以看出。因此，在多维对应分析图中，主要考察第一维度上的变动情况。

CORRESP 过程的输出结果中还给出了行、列得分对应的坐标，如图 9-12 和图 9-13 所示。

Row Coordinates		
	Dim1	Dim2
1001-3000RMB	-0.0600	0.0161
3001-5000RMB	0.0057	0.0348
5001-8000RMB	0.4865	0.1477
8001-10000RMB	0.8777	0.1983
Under 1000RMB	-0.0308	-0.0282
Upper 10000RMB	0.8597	-0.3387

图 9-12 行变量的坐标

Column Coordinates		
	Dim1	Dim2
Brand	-0.0266	0.0159
Camera	0.3314	0.1997
Duration	-0.0318	-0.0041
Function	-0.0602	0.0121
IO	0.8178	-0.1781
Memory	0.0220	0.0079
Net	0.4018	-0.1752
OS	0.6718	0.1665
Price	-0.0910	-0.0116
Quality	-0.0577	-0.0169
Style	-0.0585	0.0073

图 9-13 列变量的坐标

图 9-12 和图 9-13 所示的坐标计算结果在关键字 OUTC 所指定的输出数据集中用变量 DIM1 和 DIM2 表示。读者可自行打开本例生成的临时数据集 CellPhone\_Out.sas7bdat 进行查看。

依据这两套坐标可以分别绘制行变量的散点图 and 列变量的散点图，如图 9-14 和图 9-15 所示（这两个图形为本书阐述对应分析图绘制过程所用，并非 SAS 系统自动输出的图形）。

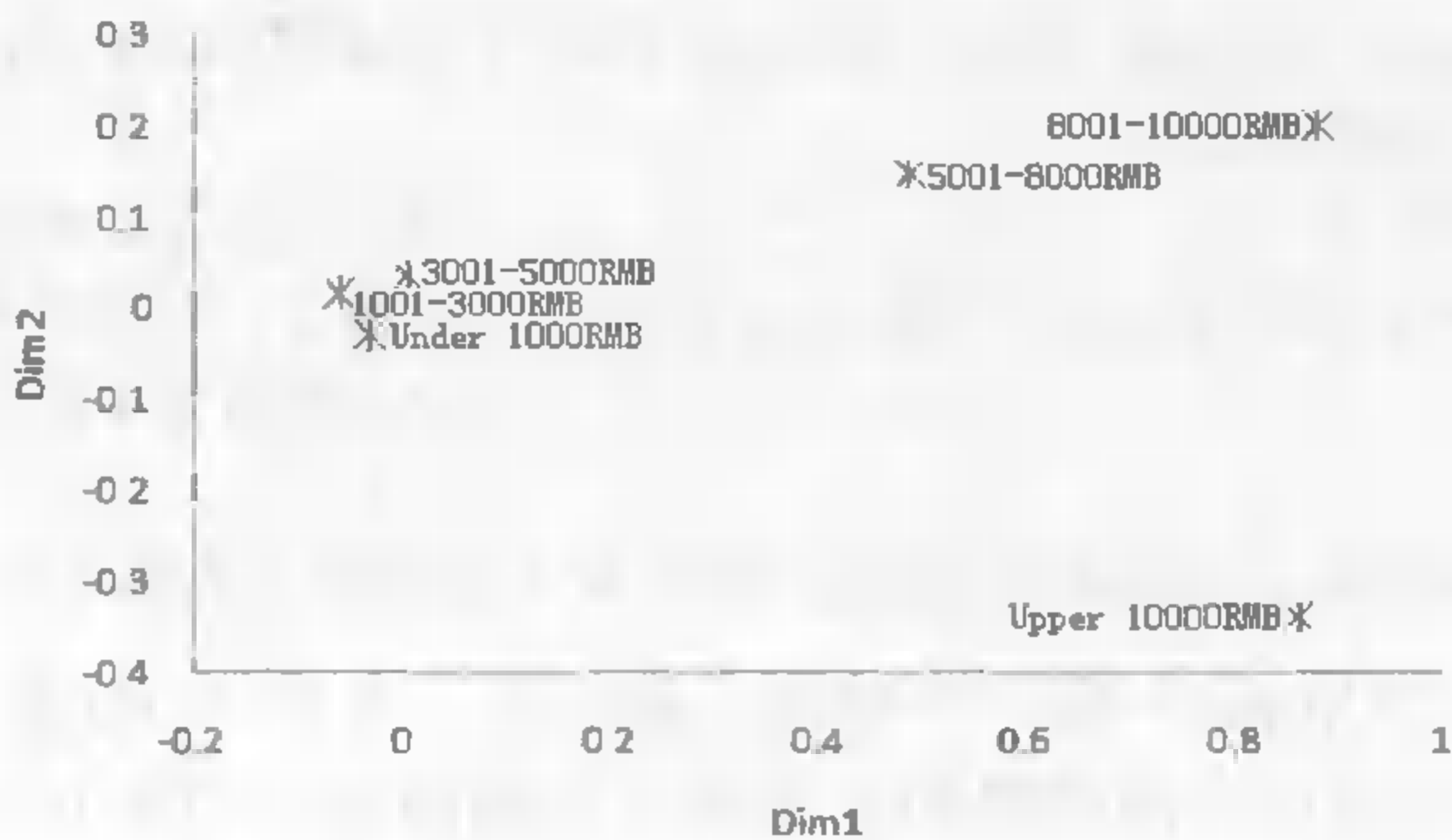


图 9-14 行变量的散点图

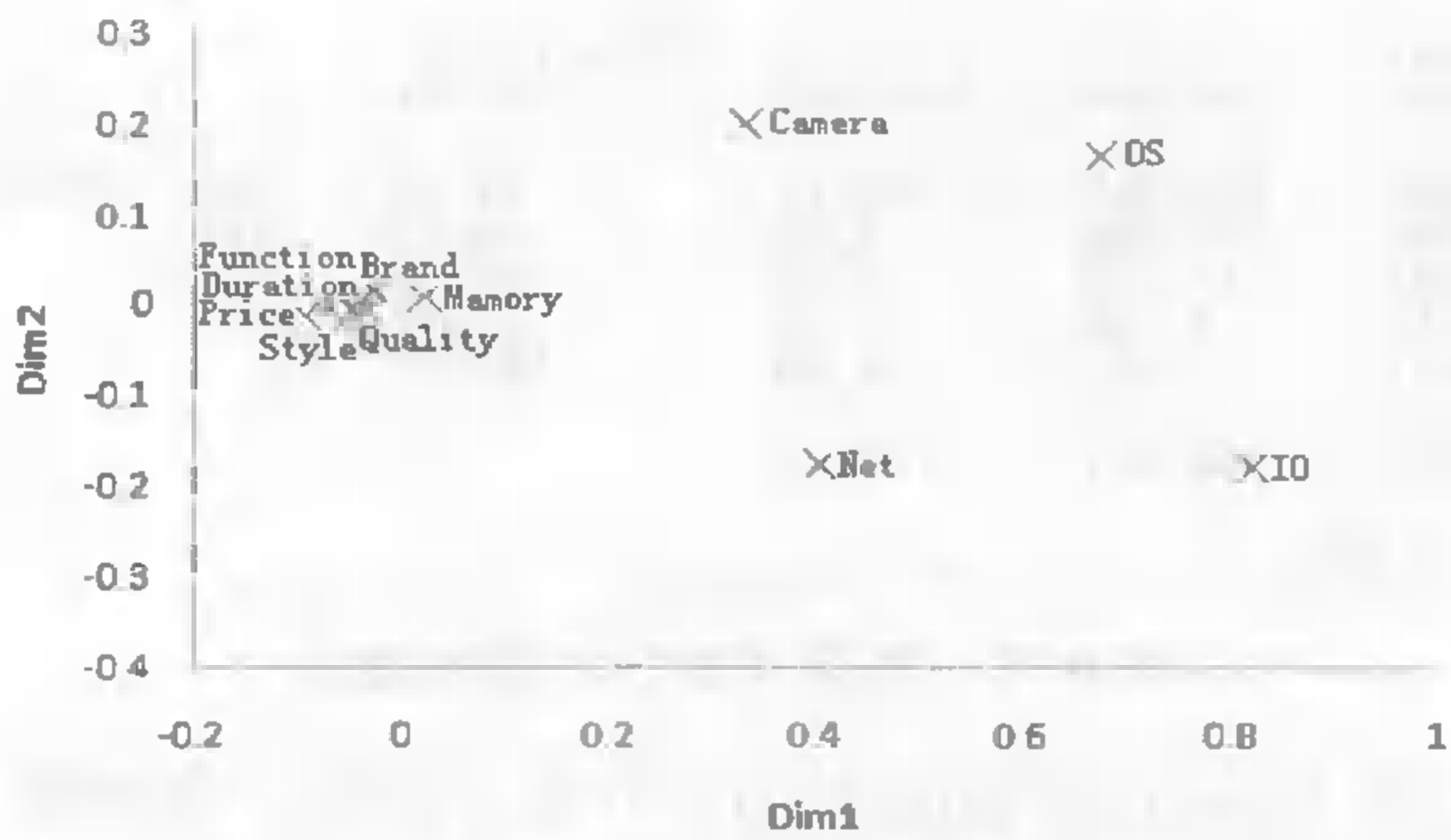


图 9-15 列变量的散点图

把图 9-14 所示的行变量散点图与图 9-15 所示的列变量散点图叠加在一起，便形成了行、列变量的对应分析图（在该过程中，SAS 系统自动进行散点图叠加），如图 9-16 所示。

在 SAS 系统绘制的对应分析图中，标注了各维度的惯量比例，可清楚地从图中进行分析。由于第 1 个维度惯量比例达到 84.70%，故主要观察散点在第一维度上的距离变动情况即可。

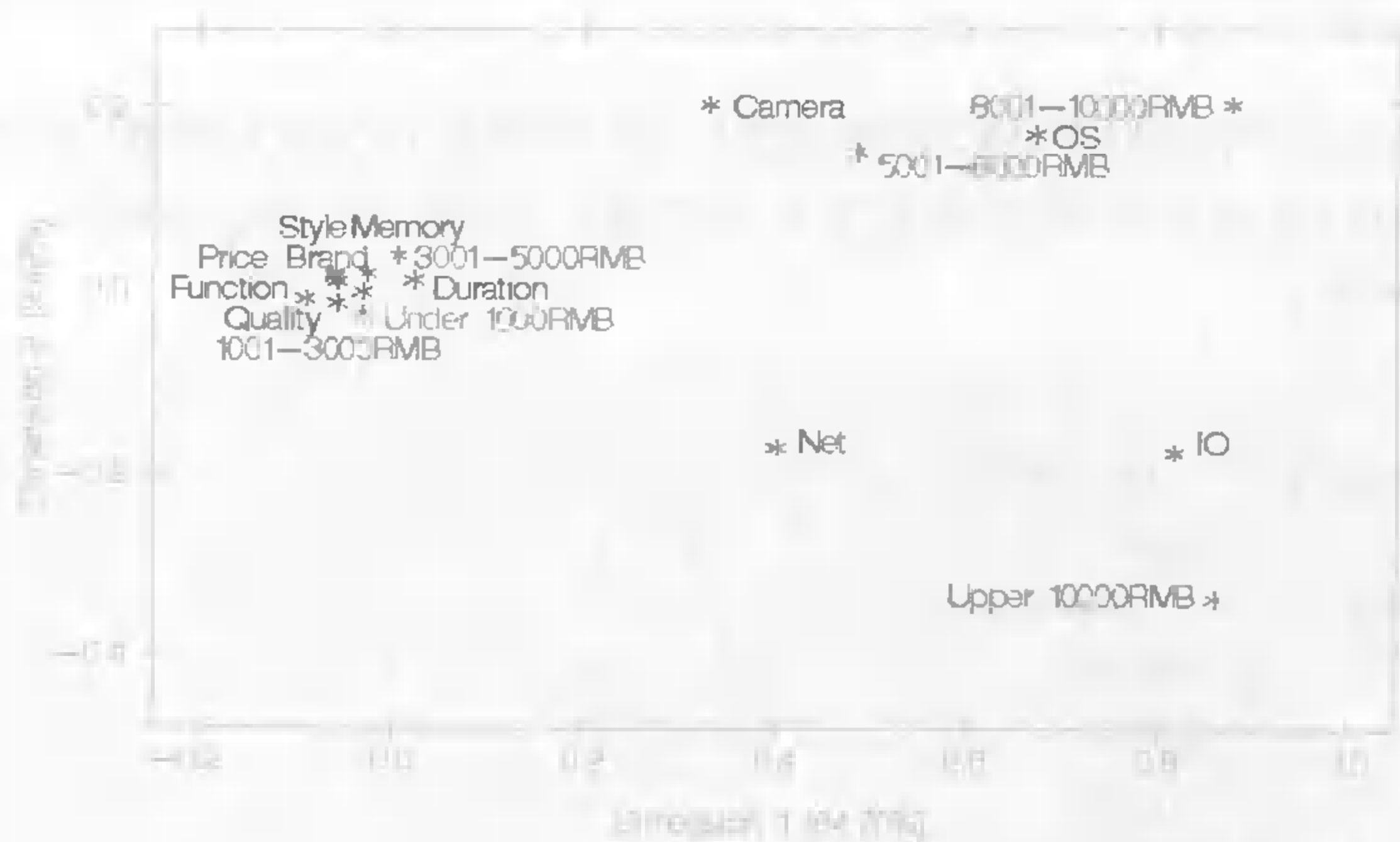


图 9-16 对应分析图

从图 9-16 可以清楚看到，月收入在 5 000 元以下的人群在购买手机时主要考虑待机时间、价格、功能、质量、品牌等因素，因此针对中低收入人群的手机应当考虑从上述方面来吸引客户；而中高收入（即收入在 5 001~10 000 元之间）群体在购买手机时更加关注摄像头、操作系统等方面，因此厂商应该在这两个方面下功夫以满足中高收入群体的需求；对于月收入 10 000 元以上的高端客户，则没有什么影响因素与其对应。

如要分析 3 个或 3 个以上变量之间的相互影响关系，则可用多重对应分析方法进行分析。



例 9-4

随着网络的普及，人们在网上的时间越来越多，网上交流活动越来越频繁，通过网络进行交友或交往的人也越来越多。由于互联网的特殊性，参与网上交流的人们之间可能互不认识且从未谋面，个人信息也可能通过网络发生泄漏，因此滋生了诸多因网络交往而产生的违法犯罪活动。为此，我们特别针对网友之间的交往活动中“与网友见面前是否会将个人真实资料告诉对方”这一具体问题进行了一次问卷调查，搜集到 465 份有效样本，调查汇总数据（详见 CyberFriend.sas7bdat）如表 9-10 所示。试对调查结果进行多重对应分析。

表 9-10 有关“与网友见面前是否会将个人真实资料告诉对方”的调查结果

人数 (Count)	性别 (Gender)	态度 (Attitude)	年龄 (Age)						合 计
			18 岁以下	18-23 岁	23-30 岁	30-40 岁	40-50 岁	50 岁以上	
与网友见面前是否会将真实资料告诉对方	男	会	3	19	23	18	9	7	79
		不会	7	37	44	47	28	13	176
		合计	10	56	67	65	37	20	255
	女	会	3	13	15	4	9	2	46
		不会	8	31	49	37	33	6	164
		合计	11	44	64	41	42	8	210

本例中需要考察性别、态度、年龄 3 个变量之间的相互关系。因此可考虑用多重对应分

析的方法来考察。

多重对应分析同样可使用 CORRESP 过程，但是要在 CORRESP 的选项中加上 MCA 关键字。此外在 TABLES 语句中列示相互关系的变量，变量之间用空格隔开。

本例具体程序如下：

```
proc format;
  value Gender_FMT      0='Female'
                        1='Male';
  value Age_FMT         1='Under 18'
                        2='18-23'
                        3='23-30'
                        4='30-40'
                        5='40-50'
                        6='Upper 50';
  value Atti de_FMT     0='No'
                        1='Yes';
run;
proc corresp mca data=Sasuser.CyberFriend out=CyberFriend_Out; /*mca 关键字表示多重对应分析*/
  tables Gender Age Attitude; /*变量之间用空格隔开*/
  format Gender Gender_FMT. Age Age_FMT. Attitude Attitude_FMT.;
  weight Count;
run;
%plotit(data=CyberFriend_Out,datatype=corresp,href=0,vref=0) /*在坐标(0,0)处绘制纵横参考线*/
```

运行程序后，可得到惯量和卡方统计量构成表，如图 9-17 所示。

The CORRESP Procedure									
Inertia and Chi-Square Decomposition									
Singular Value	Principal Inertia	Chi-Square	Percent	Cumulative Percent	4	8	12	16	20
0.62750	0.39376	555.28	16.88	16.88	*****	*****	*****	*****	*****
0.59824	0.35789	504.71	15.34	32.21	*****	*****	*****	*****	*****
0.57735	0.33333	470.07	14.29	46.50	*****	*****	*****	*****	*****
0.57735	0.33333	470.07	14.29	60.79	*****	*****	*****	*****	*****
0.57735	0.33333	470.07	14.29	75.07	*****	*****	*****	*****	*****
0.55538	0.30845	434.98	13.22	88.29	*****	*****	*****	*****	*****
0.52272	0.27324	385.32	11.71	100.00	*****	*****	*****	*****	*****
Total	2.33333	3290.51	100.00						
Degrees of Freedom = 81									

图 9-17 多重对应分析惯量与卡方统计量构成表

从图 9-17 中可以看出， $\chi^2$  统计量值为 3 290.51，说明变量之间的相互关系显著。前两个维度对应的惯量累积比例为 32.21%，其重要性一般。由于其余惯量比例均比较小，从绘制直观的对对应分析图形考虑，可取前两个维度绘制对应分析图。本例绘制的对应分析图如图 9-18 所示。

图 9-18 中的两个维度对应的惯量比例比较接近，且均要考察散点之间的距离远近。从图 9-18 来看，30 岁以下的人在会见网友之前会告诉对方自己的真实资料，而且 18~23 岁年轻人的点与会公布真实资料的点更加接近，这与实际情况比较相符。由于这部分人群一方面社会阅历尚浅，另一方面出于对互联网络的好奇和追求刺激，往往会向对方公开自己的真实资料。而 40~50 岁的中年女性则比较成熟，社会生活经验丰富，在交往过程中往往趋向于保护自身利益，在见面之前一般不会告诉对方自身的真实资料。

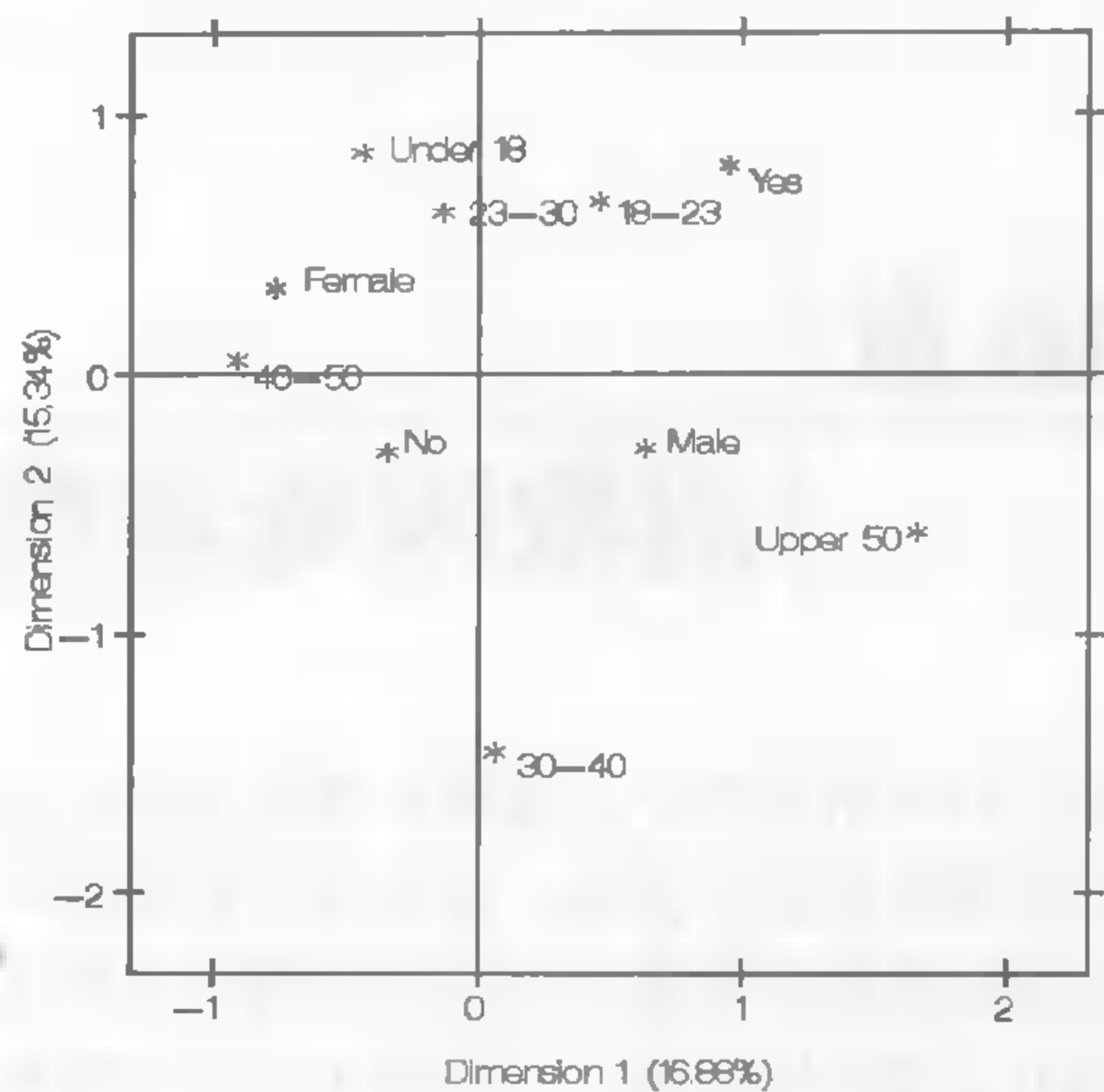


图 9-18 多重对应分析图

### 9.3 本章小结

本章主要介绍了列联分析和对应分析的基本原理及其在 SAS 系统中的实现，主要内容简要回顾如下：列联分析和对应分析可以考察变量之间的相互关系；列联分析主要通过列联表进行，可对行列变量之间的关系进行  $\chi^2$  检验；对应分析是列联分析的深入，主要用于分析变量之间是如何进行相互影响的；对应分析可分为两个行、列变量的简单对应分析和多个变量的多重对应分析，对应分析的结果主要通过对对应分析图来描述。

## 第 10 章

# 离散因变量模型

在第 6 章中介绍的经典回归分析模型中，被解释变量或因变量通常被假定为连续的定量变量，而解释变量中可以含有离散的定性变量，通常用虚拟变量处理之。但在实际的社会经济问题分析过程中，常常会遇到被解释变量可能是离散的定性变量，或受限制变量，如人们对某项政策的态度有支持和不支持两种情况；人们购买手机的意愿也有两种情况，即购买和不购买；人们出行时，可以选择步行、自行车、公交、地铁、私家车等方式。

要考察人们做出某种具体选择的情况及其影响因素时，可把这些离散的定性变量作为因变量进行分析。如出行选择交通工具的种类可能与家庭收入、生活习惯、城市交通状况等因素有关，把交通工具的选择作为因变量，把上述影响因素作为自变量，这样所建立的模型称为离散选择模型。

此外，还有一种情况是因变量是以离散计数的方式描述的，如发生交通事故的次数、运动会上获得的奖牌数等，这些都是以整数计数的形式出现的。可以将分析自变量对计数的因变量的影响时所建立的模型称为计数模型，这也是离散因变量模型的一种典型形式。

上述这些可供选择的选项都是离散定性变量的表现形式，那么是否能够按照第 6 章中的回归思想把定性变量作为因变量与其影响因素进行回归分析呢？这便是本章将要阐述的离散因变量模型的主要内容，该部分内容也属于微观计量经济学的重要研究内容。

离散因变量模型起源于 Fechner 于 1860 年进行的动物条件二元反射研究。1962 年，Warner 首次将它应用于经济研究领域，用以研究公共交通工具和私人交通工具的选择问题。1980 年代之后，该模型被普遍应用于社会、经济决策领域的研究。

### 10.1 线性概率模型

离散选择模型在一般的概率模型框架下展开，并依赖结果是两个或多个选择将模型分为二项选择、多项选择模型和受限因变量模型。因此本节首先介绍线性概率模型。

离散选择模型主要研究选择结果的概率与影响因素之间的关系。

$$\text{Prob}(\text{事件}i\text{发生}) = \text{Prob}(Y = i) = F(\text{影响因素})$$

其中的影响因素可能包含做出选择的主体属性和选择方案属性。如选择何种交通工具出行，既受到选择主体收入程度、生活习惯等属性的影响，也受到交通工具的价格、便捷性等属性的影响。



#### 例 10-1

3G 手机业务目前在国内刚处于起步阶段，为考察其在消费受众中的使用意向，某咨询公司针对不同特征的潜在客户进行了一次小规模摸底调查，对影响 3G 手机购买意向的因素进行分析。

其中, 购买意向为定性变量, 有两种选择, “0”表示不购买, “1”表示购买其影响因素可能有性别、年龄、收入、职位、行业等。

针对上述问题, 设定因变量为  $y$ , 表示是否购买 3G 手机, 则  $y$  有两个值。

$$y = \begin{cases} 0 & \text{不购买} \\ 1 & \text{购买} \end{cases}$$

影响  $y$  的因素记为  $x = (x_1, x_2, L, x_n)$ , 根据多元回归的思想, 可得到以下回归模型。

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + L + \beta_n x_n + \mu$$

其中  $(\beta_1, \beta_2, L, \beta_n)^T = \beta$  表示回归模型中的参数, 即回归系数, 则模型可以简记为  $y = \beta_0 + \beta x + \mu$ 。

在因变量是离散变量的情况下, 不能把  $\beta_i (i=1, 2, M, n)$  理解为保持其他因素不变的情况下对  $y$  的边际影响, 因为  $y$  的取值为 1 或 0。

回忆经典回归分析中  $E(\mu|x)=0$  的假定, 对回归模型左右两边求条件期望, 则有:

$$E(y|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + L + \beta_n x_n$$

因为  $y$  的取值为 1 或 0, 记  $y=1$  的概率为  $p$ , 依据数学期望的定义有:

$$E(y) = 0 \times (1-p) + 1 \times p = p = \text{Prob}(y=1|x)$$

因此, 做出“购买”决策的概率等于  $y$  的期望值, 把上述 2 个式子联立起来:

$$\text{Prob}(y=1|x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + L + \beta_n x_n$$

即做出“购买”决策的概率  $p(y=1|x)$  是  $x_i$  的一个线性函数, 该回归方程亦即线性概率模型 (Linear Probability Model, LPM),  $p(y=1|x)$  可被称为“响应概率”。用普通最小二乘法对模型进行参数估计, 可得到以下回归方程。

$$\text{Prob}(y=1|x) = \hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + L + \hat{\beta}_n x_n = \hat{\beta}_0 + \hat{\beta}x$$

$\hat{y}$  则是  $y=1$  时的概率, 在保持其他因素不变的情况下,  $\beta_i$  表示因  $x_i$  的变化导致“购买”(即  $y=1$ ) 决策概率的变化, 即  $\Delta P(y=1|x) = \beta_i \Delta x_i$ 。

## 10.2 二元选择模型

二元选择模型 (Binary Choice Model) 具体是指离散因变量具有两个选项或两种属性。在通常情况下, 离散因变量的属性是对立的, 如成功和失败、购买和不购买、使用和不使用等。一般可以用“1”表示具备有某种属性, 用“0”表示不具备某种属性, 通常用回归方程式的形式描述变量之间的关系, 因此二元选择模型又可以被称为“0-1 因变量回归模型”。



在应用二元选择模型时, 因变量中的“0”和“1”只是对应属性的标注或符号, 不具备任何数值上的意义, 不能进行直接用于数学运算。

二元选择模型是离散选择模型中最简单的情形, 其研究目的是获得具有某给定特征的个体做出某种选择决策的概率, 其基本原理可从克服线性概率模型的缺陷入手进行

研究。

### 10.2.1 线性概率模型的缺陷与改进

线性概率模型 (LPM) 虽然能够测度因变量事件发生的响应概率, 可以考察因自变量变动导致因变量决策的概率变化。但是该模型存在着几个比较严重的缺陷。

● 解释变量的合理变化会导致概率预测值溢出在  $[0,1]$  区间之外。

- 随机误差项的分布未知。
- 模型误差项具有异方差性, 异方差性使参数估计不具有有效性。
- 即使用加权最小二乘法修正异方差性, 同样无法保证概率预测值在  $[0,1]$  区间之内。

随机误差项分布的缺陷较容易克服, 当样本量充分大时, 其普通最小二乘参数估计量的结果近似服从正态分布。但其他问题却难以解决, 如有回归方程  $\hat{y} = -0.24 + 0.026x_1 + 0.039x_2$ , 当  $x_1 = 15$ 、 $x_2 = 25$  时,  $\hat{y} = 1.125$ 。因  $\hat{y}$  为一个概率值, 不可能大于 1, 故模型设定有误。造成这种缺陷的重要原因是在模型设定时就假设响应概率与自变量之间是线性关系。因此需要对变量之间的线性关系进行变换, 使得自变量  $x$  对应的所有概率值都在  $[0,1]$  区间范围之内, 且自变量  $x$  变化时,  $y$  单调变化。

显然, 选择对  $\beta_0 + x\beta$  的适当变换成为最为关键的问题。依据上述要求, 最为常见且符合要求的形式便是利用分布函数  $F()$  进行变换。因为分布函数对于所有的自变量而言, 其值均在区间  $[0,1]$  范围之内, 且当自变量发生变化时, 分布函数值单调变化。

### 10.2.2 二元选择模型的基本原理

依据分布函数的基本特征, 二元选择模型可从一个满足经典线性模型假定的隐变量模型中推导出来。所谓隐变量便是不能够直接进行观测但可以通过其他直接观测得到的变量 (显变量) 进行描述和反映的变量。

#### 1. 模型构建和参数估计过程

设  $y^*$  是一个由  $y^* = \beta_0 + x\beta + \mu^*$ 、 $y = \begin{cases} 1, & y^* > 0 \\ 0, & y^* \leq 0 \end{cases}$  所决定的、不能够直接观测得到的隐

变量, 隐变量通常是解释变量或影响因素的线性函数; 且假定  $\mu^*$  与  $x$  相互独立,  $\mu^*$  服从某种对称于 0 的分布。则  $y^* = \beta_0 + x\beta + \mu^*$  也可被称为隐变量模型。

记  $F$  为分布函数, 因为  $\mu^*$  对称于 0, 则有:  $F(z) = 1 - F(-z)$ , 其中  $z$  是随机变量。

依据  $y^*$  与  $y$  的对应关系和对称分布函数性质, 则  $y$  的响应概率为:

$$\begin{aligned} P(y=1|x) &= P(y^* > 0|x) = P((\beta_0 + x\beta) + e > 0|x) = P(e > -(\beta_0 + x\beta)|x) = P(e < (\beta_0 + x\beta)|x) \\ &= F(\beta_0 + x\beta) \\ &= 1 - F(-(\beta_0 + x\beta)) \end{aligned}$$

如果已知  $\mu^*$  分布函数  $F(z)$  的具体形式, 那么决定  $y$  的响应概率的模型便确定了。

在二元选择模型中, 通常给定  $\mu^*$  服从的分布具有表 10-1 所示的、两种最为常用的形式。

表 10-1 二元选择模型的分布及对应的模型

$\mu^*$ 的分布	分布函数 $F(z)$	二元选择模型	模型具体形式
标准正态分布	$\Phi(z)$	Probit 模型	$P(y=1 x)=F(\beta_0+x\beta)=\Phi(\beta_0+x\beta)$
逻辑 (Logistic) 分布	$\frac{\exp(z)}{1+\exp(z)}$	Logit 模型	$P(y=1 x)=F(\beta_0+x\beta)=\frac{\exp(\beta_0+x\beta)}{1+\exp(\beta_0+x\beta)}$

$$\begin{aligned} \because E(y) &= P(y=1|x) = F(\beta_0+x\beta) \\ \therefore y &= E(y) + (y-E(y)) = E(y) + \mu = F(\beta_0+x\beta) + \mu \end{aligned}$$

在二元选择模型中，一般选用极大似然法进行参数估计，其似然函数如下。

$$L = \prod_{i=1}^n [F(\beta_0+x\beta)]^{y_i} [1-F(\beta_0+x\beta)]^{1-y_i}$$

其对数似然函数如下。

$$\ln L = \sum_{i=1}^n \left\{ y_i \ln F(\beta_0+x\beta) + (1-y_i) \ln [1-F(\beta_0+x\beta)] \right\}$$

求对数似然函数对每个参数的偏导数，使之均为 0，再求解方程组即可解得模型中的参数估计值。在 SAS 系统中可自动实现该过程。



二元选择模型的参数估计结果不能被理解为自变量变动对因变量的边际影响，而应当被理解为自变量的变动对因变量取“1”的概率的影响，即：

$$\frac{\partial P(y=1|x)}{\partial x_i} = \frac{\partial F(\beta_0+x\beta)}{\partial (\beta_0+x\beta)} \beta_i = f(\beta_0+x\beta) \beta_i, \text{ 其中 } f(\cdot) \text{ 是分布函数 } F(\cdot) \text{ 的密度函数。}$$

## 2. 模型检验

二元选择模型与经典回归模型一样，同样可以用 Z 统计量对回归系数进行显著性检验，该 Z 统计量由极大似然法给出。对于多个系数的约束显著性检验，还可以计算 Wald 统计量，利用 Wald 统计量近似服从  $\chi^2$  分布进行  $\chi^2$  检验。而对于模型的拟合优度检验，则可以利用 LR 似然比进行  $\chi^2$  检验。此外，还可以计算模型的 AIC、BIC 等信息指数以对模型进行评价。

因各统计量计算过程较复杂，且在 SAS 系统中，各统计量可由不同的过程得到，故本书不再讲述检验过程的原理，在具体分析实际事例时再一并讲解。

### 10.2.3 BINARY PROBIT 模型

二元选择 (BINARY PROBIT) 模型对隐变量随机误差项假定服从标准正态分布，其模型具有以下形式 (详见 10.2.2 小节)。

$$P(y=1|x) = F(\beta_0+x\beta) = \Phi(\beta_0+x\beta)$$

在 SAS 系统中，QLIM、GENMOD、LOGISTIC、PROBIT 等 4 个过程均可以建立二元

PROBIT 模型。这些过程除了建立 PROBIT 模型以进行分析之外，同样可用于其他离散选择模型的分析过程当中，因此本书将不再笼统介绍各个具体过程的具体语法，只在对应章节中介绍所调用的过程能够实现本章节内模型建模的功能和主要语法。

### 1. QLIM 过程的二元 PROBIT 建模

QLIM(定性数据和受限因变量模型)可用于本节中的 PROBIT 建模分析。其建立 PROBIT 模型的主要语法如下。

```
PROC QLIM <选项>;
  CLASS 变量;
  MODEL 因变量 = 自变量 /选项;
  ENDOGENOUS 内生变量~DISCRETE (DISTRIBUTION = NORMAL);
  OUTPUT OUT = 输出数据集名 PROBALL;
RUN;
```

其中，CLASS 语句用于指定分析过程的分类变量或定性变量(含因变量和自变量)，MODEL 用于指定模型的因变量和自变量之间的关系，如有多个自变量，则自变量之间用空格隔开。

ENDOGENOUS 语句用于指定模型的内生变量及内生变量的形式及其分布。在单方程的 PROBIT 建模过程中，因变量是离散(Discrete)的内生变量，响应概率所使用的分布函数为正态分布，即指定 Distribution = Normal 或 Distribution = Probit。

OUTPUT 语句用于输出分析结果数据集，PROBALL 关键字表示利用现有样本数据计算各个样本因变量各取值的响应概率。

### 2. PROBIT 过程的二元 PROBIT 建模

PROBIT 过程进行二元 PROBIT 建模的主要语法如下。

```
PROC PROBIT <选项>;
  CLASS 变量;
  MODEL 因变量 = 自变量 /选项;
  OUTPUT OUT = 输出数据集名 PROB = 变量名;
RUN;
```

其中 CLASS 语句和 MODEL 语句的功能同 QLIM 过程。



PROBIT 过程在分析过程中，把离散因变量数值最小的值作为考察响应变量的依据，在通常的 0-1 因变量分析中，其分析结果是表示因变量为“0”值时的概率变动状况。因此，在实际问题分析过程中，考察 PROBIT 过程的参数估计结果时，应当在各个参数估计值前取负值。

OUTPUT 语句用于输出分析结果数据集，PROB 关键字表示利用现有样本数据计算各个样本的响应概率，并指定表示该响应概率的变量名。

### 3. OGISTIC 过程二元 PROBIT 建模

LOGISTIC 主要用于 LOGIT 模型的构建与分析，但是可以通过链接函数指定建立

PROBIT 模型，其主要语法如下。

```
PROC LOGISTIC <选项>;
  CLASS 变量;
  MODEL <(EVENT = '响应值')> 因变量 = 自变量 / LINK = PROBIT;
  OUTPUT OUT = 输出数据集名 PREDICTED = 变量名;
RUN;
```

LOGISTIC 过程 MODEL 语句中的“LINK = PROBIT”表示建立 PROBIT 模型。

与 PROBIT 过程一样，LOGISTIC 过程在分析过程中同样把离散因变量的数值最小的值作为考察响应变量的依据，但是 LOGISTIC 过程可以通过指定 DESCENDING 关键字把离散因变量数值最大的值作为考察响应变量概率的依据。此外，LOGISTIC 过程还可以通过 MODEL 语句中的 EVENT 选项考察因变量不同响应值的建模情况。

OUTPUT 语句用于输出分析结果数据集，PREDICTED 关键字表示利用现有样本数据计算各个样本的响应概率，并指定表示该响应概率的变量名。

4. GENMOD 过程的二元 PROBIT 建模

GENMOD（广义线性模型）过程同样也可用于 PROBIT 建模，其有关的主要语法如下。

```
PROC GENMOD <选项>;
  CLASS 变量;
  MODEL 因变量 = 自变量 / DIST = BINOMIAL LINK = PROBIT;
  OUTPUT OUT = 输出数据集名 PREDICTED = 变量名;
RUN;
```

与 PROBIT、LOGISTIC 过程一样，GENMOD 过程在分析过程中同样把离散因变量的数值最小的值作为考察响应变量的依据，但是可以通过指定 DESCENDING 关键字把离散因变量数值最大的值作为考察响应变量概率的依据。

OUTPUT 语句用于输出分析结果数据集，PREDICTED 关键字表示利用现有样本数据计算各个样本的响应概率，并指定表示该响应概率的变量名。



例 10-2

某快速消费品生产厂商针对其某个品牌的产品进行了一次满意度调查，用以考察消费者的满意状况、对产品的抱怨状况及对产品购买使用的忠诚度状况，依据这些用户特征和态度对消费者是否会持续购买使用该产品进行分析。本次调查搜集了 337 个有效样本（数据详见 Product Usage.sas7bdat），如表 10-2 所示。

表 10-2 快速消费品满意度及持续购买使用状况（节选数据）

满意度 CSI	抱怨情况 Complaint	忠诚度 Loyalty	继续购买 使用意向 Attitude	满意度 CSI	抱怨情况 Complaint	忠诚度 Loyalty	继续购买 使用意向 Attitude	满意度 CSI	抱怨情况 Complaint	忠诚度 Loyalty	继续购买 使用意向 Attitude
9	6	9	1	10	3	10	1	8	3	7	0
9	5	9	1	10	7	10	1	6	6	4	0
7	4	8	1	9	7	7	1	8	7	8	1
6	7	7	0	9	5	10	1	7	5	6	1

续表

满意度 CSI	抱怨情况 Complaint	忠诚度 Loyalty	继续购买 使用意向 Attitude	满意度 CSI	抱怨情况 Complaint	忠诚度 Loyalty	继续购买 使用意向 Attitude	满意度 CSI	抱怨情况 Complaint	忠诚度 Loyalty	继续购买 使用意向 Attitude
8	10	7	1	10	5	6	1	8	10	5	1
7	3	6	1	9	7	10	1	7	6	8	1
9	2	9	1	9	7	7	1	4	7	5	0
6	5	9	0	8	4	8	1	4	8	3	0
9	5	10	1	7	5	10	1	9	6	5	1
7	8	5	0	7	1	10	1	8	5	7	1
8	4	4	0	8	7	6	1	8	4	7	1
9	6	8	1	9	2	8	1	7	5	8	1
9	3	10	1	9	5	7	1	9	5	10	1
8	9	8	0	6	6	3	0	8	4	5	1
10	7	7	1	8	1	9	1	8	8	8	1
9	7	7	1	8	7	6	1	10	7	10	1
6	7	3	0	9	6	4	1	9	8	9	1
...											
9	7	8	1	9	4	10	1	7	6	8	0
10	4	10	1	9	6	10	1	9	6	9	1

在表 10-2 所示的数据中，用户持续购买使用意向是本次调查研究的对象，该变量有两个互斥离散的变量值，即“继续购买”和“不继续购买”，在数据中分别用“1”和“0”来表示。因此用户对于该问题有两个选择，用户选择情况受满意度、抱怨情况及忠诚度因素的影响（以被调查者打分的形式表示），故本例可以建立二元 PROBIT 模型进行分析。

用 QLIM 过程进行 PROBIT 建模，并以此为例对模型检验及评价问题进行分析。使用 QLIM 分析的程序如下。

```
proc format;
  value Attitude_FMT1 = '持续购买'          /*为离散因变量指定变量值标签*/
  0 = '不持续购买';
run;
proc qlim data = Sasuser.Product_Usage;      /*调用 QLIM 过程对 Product_Usage 数据集进行分析*/
  model Attitude = CSI Complaint Loyalty;    /*指定离散因变量及其影响因素（自变量）的模型*/
  endogenous Attitude~discrete (dist = probit); /*指定内生变量，并采用正态分布*/
  format Attitude Attitude_FMT.;           /*为因变量值挂上标签*/
  output out = Qlim_Out proball;            /*指定输出数据集名为 Qlim_Out 的临时数据集*/
run;
```

运行程序后，首先可得到模型中离散因变量的一些基本信息，如图 10-1 所示。

The QLIM Procedure			
Discrete Response Profile of Attitude			
Index	Value	Frequency	Percent
1	不持续购买	80	23.74
2	持续购买	257	76.26

图 10-1 离散因变量的基本信息

在离散因变量的基本信息中，可以看到各变量值的频数和百分比。



该输出结果中的“Index”列所示的数字只代表输出结果的序号，不代表离散因变量的具体取值。

对于模型拟合的基本信息，系统一并给出多种方法测度的结果，如图 10-2 所示。

Model Fit Summary	
Number of Endogenous Variables	1
Endogenous Variable	Attitude
Number of Observations	337
Log Likelihood	-134.37018
Maximum Absolute Gradient	4.61427E-7
Number of Iterations	10
AIC	276.74035
Schwarz Criterion	292.02068

图 10-2 模型拟合基本信息

在图 10-2 中，从上至下依次列示了模型中的内生变量个数、内生变量在数据集中的变量名、观测值数目、对数似然值、最大绝对梯度、迭代次数、AIC 信息指数和 Schwarz 标准。可以依据图 10-2 中的统计量进行模型评价和模型选择，如选择 AIC 较小的模型比选择 AIC 较大的模型更加合适。

模型的拟合优度检验统计量在输出结果中也一并给出，如图 10-3 所示。

Goodness-of-Fit Measures		
Measure	Value	Formula
Likelihood Ratio (R)	100.65	$2 * (\text{LogL} - \text{LogL0})$
Upper Bound of R (U)	369.39	$- 2 * \text{LogL0}$
Aldrich-Nelson	0.23	$R / (R+N)$
Cragg-Uhler 1	0.2582	$1 - \exp(-R/N)$
Cragg-Uhler 2	0.3878	$(1-\exp(-R/N)) / (1-\exp(-U/N))$
Estrella	0.2944	$1 - ((1-R/U)^(U/N))$
Adjusted Estrella	0.2713	$1 - (((\text{LogL}-K)/\text{LogL0})^(-2/N*\text{LogL0}))$
McFadden's LRI	0.2725	$R / U$
Veall-Zimmermann	0.4398	$(R * (U+N)) / (U * (R+N))$
McKelvey-Zavoina	0.4232	
N = # of observations, K = # of regressors		
Algorithm converged.		

图 10-3 模型拟合优度检验

图 10-3 所示的模型拟合优度检验不仅列示了所计算出来的统计量，同时在 Formula 列还给出了得到这些统计量的计算公式。

如本例选择似然比进行拟合优度检验。在图 10-3 中，似然比（Likelihood Ratio）的值为 100.65，因为似然比统计量服从自由度为全模型（含所有考察的自变量）与空模型（不含任

何自变量) 自由参数之差 (本例中该自由度即为  $4-1=3$ ) 的  $\chi^2$  分布。本例中的  $\chi^2$  统计量非常大, 其对应的检验 P 值几乎为 0, 故模型拟合程度非常好。

本例模型的参数估计及其显著性检验结果如图 10-4 所示。

Parameter Estimates				
Parameter	Estimate	Standard Error	t Value	Approx Pr >  t
Intercept	-1.792294	0.531853	-3.37	0.0008
CSI	0.247938	0.063055	3.93	<.0001
Complaint	-0.132856	0.046127	-2.88	0.0040
Loyalty	0.198390	0.044302	4.48	<.0001

图 10-4 QLIM 过程的参数估计及显著性检验

类似于经典回归模型, 对于二元 PROBIT 模型的参数, 同样可用图 10-4 中的  $t$  统计量进行检验。检查各个系数 (通常情况下截距项不进行检验) 所对应的  $P$  值 (“Pr>|t|”), 在  $\alpha = 0.05$  的条件下,  $P$  小于  $\alpha$ , 故该模型中的系数估计结果 (“Estimate”) 是显著的, 即各自变量均对因变量有显著影响。而且, 依据系数估计的结果可以明显看出, 满意度 (“CSI”) 和忠诚度 (“Loyalty”) 的系数为正, 表示随着满意度或忠诚度的增加, 消费者 “持续购买” 的概率也会增加; 而用户抱怨情况 (“Complaint”) 的系数为负, 表示随着用户抱怨的增加, 则消费者 “持续购买” 的概率会降低。



估计出来的系数并不代表随着自变量变动因变量概率变动的绝对数值。但是可以通过参数估计之后所建立的模型对现有观测样本或新的观测样本进行考察或预测。

根据图 10-4 中的参数估计结果, 可以写出本例的 PROBIT 模型。

$$\begin{aligned} P(y=1|x) &= F(\beta_0 + x\beta) = \Phi(\beta_0 + x\beta) \\ &= \Phi(-1.792294 + 0.247938 \times CSI - 0.132856 \times Complaint + 0.198390 \times Loyalty) \end{aligned}$$

如现有对某个消费者的调查访问数据, 调查显示其满意度得分为 8, 抱怨程度得分为 4, 忠诚度得分为 7, 现可以依据所建立模型考察其 “持续购买” 产品的概率。把上述自变量的数值代入以上模型中, 可得到:

$$\begin{aligned} P(y=1|x) &= \Phi(-1.792294 + 0.247938 \times CSI - 0.132856 \times Complaint + 0.198390 \times Loyalty) \\ &= \Phi(-1.792294 + 0.247938 \times 8 - 0.132856 \times 4 + 0.198390 \times 7) \\ &= \Phi(1.048516) \\ &= 0.8528 \end{aligned}$$

即该消费者持续购买该企业产品的概率为 0.8528。

本例的分析过程也可选用 PROBIT 过程进行, 具体程序如下:

```
proc format;
  value Attitude_FMT1 = '持续购买'
    0 = '不持续购买';
run;
proc probit data = Sasuser.Product_Usage; /*调用 PROBIT 过程对 Product_Usage 数据集进行分析*/
  class Attitude; /*指定 Attitude 变量为离散定性或分类变量*/
```

```
model Attitude = CSI Complaint Loyalty; /*指定离散因变量与其对应的自变量模型*/
format Attitude Attitude_FMT.; /*为因变量挂上值标签*/
output out = Probit_Out prob = Estimated_P; /*指定输出含有响应概率的结果数据集,并把为响应概
率对应的变量命名为 Estimated_P*/
run;
```

运行程序后，其分析过程输出的结果大体上与采用 QLIM 过程的结果相同。首先同样输出模型的基本信息，如图 10-5 所示。

Probit Procedure		
Model information		
Data Set	SASUSER.PRODUCT_USAGE	
Dependent Variable	Attitude	持续购买意向
Number of Observations	337	
Name of Distribution	Normal	
Log Likelihood	-134.3701759	
Number of Observations Read	337	
Number of Observations Used	337	

图 10-5 模型基本信息

在图 10-5 中，从上至下列示了分析所用的数据集、因变量、观测值数目、所用分布名称和对数似然值。系统给出关于因变量的信息，如图 10-6 所示。

Class Level information		
Name	Levels	Values
Attitude	2	不持续购买 持续购买
Response Profile		
Ordered Value	Attitude	Total Frequency
1	不持续购买	80
2	持续购买	257
PROC PROBIT is modeling the probabilities of levels of Attitude having LOWER Ordered Values in the response profile table.		
Algorithm converged.		

图 10-6 因变量基本信息

图 10-6 标示了本例的离散因变量名为“Attitude”，有“不持续购买”和“持续购买”两个水平，同时在“Response Profile”表格中标示了这两个水平的样本量。务必注意图 10-6 中的系统提示信息。在提示信息中，系统提示用户使用 PROBIT 过程时，该过程所分析的是离散因变量最小值的响应概率（该提示信息也会在系统“LOG”窗口中显示），即本例因变量的具体值为“1”和“0”，其最小值为“0”，故使用 PROBIT 过程分析的是因变量为“0”即“不持续购买”的响应概率。解决这个问题的方法是，在 PROBIT 过程估计出来的参数估计值之前加负号即可。

本例 PROBIT 过程的参数估计及检验结果如图 10-7 所示。

在 PROBIT 过程中，可使用服从卡方分布的 Wald 统计量对影响因素的显著性进行检验，如在图 10-7 所示的检验过程中，所有因素在  $\alpha = 0.05$  条件下均非常显著。

依据图 10-6 所示的提示信息，把图 10-7 输出的参数估计值取负号，即可得到图 10-4 所示的因变量取值为“1”时的模型参数。

Type III Analysis of Effects							
Effect		DF	Wald Chi-Square		Pr > ChiSq		
CSI		1	15.4613		<.0001		
Complaint		1	8.2957		0.0040		
Loyalty		1	20.0537		<.0001		

Analysis of Parameter Estimates							
Parameter	DF	Estimate	Standard Error	95% Confidence Limits		Chi-Square	Pr > ChiSq
Intercept	1	1.7923	0.5319	0.7499	2.8347	11.36	0.0008
CSI	1	-0.2479	0.0631	-0.3715	-0.1244	15.46	<.0001
Complaint	1	0.1329	0.0461	0.0424	0.2233	8.30	0.0040
Loyalty	1	-0.1984	0.0443	-0.2852	-0.1116	20.05	<.0001

图 10-7 PROBIT 过程的参数估计及检验结果

此外，本例还可采用 LOGISTIC 过程和 GENMOD 过程进行分析，具体程序如下。

```
proc format;
  value Attitude_FMT1 = '持续购买'
0 = '不持续购买';
run;
proc logistic data = Sasuser.Product_Usage;
  model Attitude(event = '持续购买') = CSI Complaint Loyalty /link = probit; /*指定研究因变量为“持续
购买”的响应概率，并指定链接函数为 PROBIT，即进行 PROBIT 建模*/
  format Attitude Attitude_FMT.;
  output out = Logit_Out predicted = Estimated_P; /*指定输出含有响应概率的结果数据集，并把响应概率
对应的变量命名为 Estimated_P*/
run;
proc genmod descending data = Sasuser.Product_Usage;
  model Attitude = CSI Complaint Loyalty /d = binomial link = probit; /*指定分布为 binomial（即
二元离散选择），链接函数为 PROBIT，即进行 PROBIT 建模*/
  output out = Genmod_Out predicted = Estimated_P; /*指定输出含有响应概率
的结果数据集，并把响应概率对应的变量命名为 Estimated_P*/
run;
```

在 LOGISTIC 过程中，在 MODEL 语句指定分析模型时，用 EVENT 关键字为因变量“Attitude”指定 SAS 系统分析“持续购买”的响应概率。如果因变量值挂上了标签，则需在 Event 关键字中使用变量值的标签（如本例）；如果因变量值没有挂上标签，则需在 Event 关键字中使用变量的真实值。如果本例没有为“Attitude”变量值挂标签，则可使用“Event = 1”或“Event = 0”表示研究因变量值为 1 或者因变量值为 0 的响应概率。

在 GENMOD 过程中，系统默认把离散因变量数值最小的值（本例最小值为“0”）作为考察响应变量的依据，但是可以通过 DESCENDING 关键字指定系统把因变量值从大到小进行倒序排列，即分析因变量最大数值的响应概率（本例最大值为“1”）。

运行上述程序后，得到的输出结果与 QLIM 过程和 PROBIT 过程结果类似，且 4 个过程所输出的结果数据集分别为 Qlim\_Out、Probit\_Out、Logit\_Out、Genmod\_Out，其各自响应概率结果均相同，且其模型和结果分析过程相同，此处不予赘述。

在 PROBIT 模型的影响因素中，还可以加入定性变量作为自变量进行研究。



例 10-3

把例 10-1 具体化，即某咨询公司针对不同性别、年龄的潜在客户进行了一次小规模摸底试调查，以对影响 3G 手机购买意向的因素进行分析。调查结果（数据详见 ThreeG.sas7bdat）如表 10-3 所示。

表 10-3 3G 手机购买意向试调查结果

性别 Gender	年龄 Age	购买意向 Purchase	性别 Gender	年龄 Age	购买意向 Purchase	性别 Gender	年龄 Age	购买意向 Purchase	性别 Gender	年龄 Age	购买意向 Purchase
1	35	0	1	39	0	2	58	1	2	59	1
2	44	0	2	49	1	1	50	1	2	38	1
2	45	1	2	42	1	1	32	0	1	39	0
1	47	1	2	50	1	1	52	1	1	43	1
1	51	0	1	45	0	1	35	0	1	52	1
1	47	0	1	47	0	1	51	0	2	39	1
2	54	1	1	30	1	1	51	0	2	31	0
2	47	1	1	39	0	1	47	0	1	39	0
1	35	0	1	51	0	2	54	1	2	34	0
1	34	0	1	45	0	2	47	1	1	46	0
1	48	0	1	43	1	1	35	0	2	58	1
1	56	1	2	39	1	1	34	0	1	50	1
2	46	1	2	31	0	1	48	0	1	32	0
1	59	1	1	39	0	1	56	1	1	52	1
1	46	1	2	34	0	2	46	1	2	38	1
2	59	1	1	52	1	1	59	1	1	46	1
1	46	0									

模型中是否加入定性变量作为自变量，对模型的构建和参数估计方法没有影响，仍然可以利用上述 4 个过程进行建模。



在 PRBOIT 过程中，一定要用 CLASS 语句指定模型中的所有定性因变量和定性自变量。

具体程序如下。

```
proc format;
  value Purchase_FMT      1='购买'
    0='不购买';
  value Gender_FMT        1='女'
    2='男';
run;
proc qlim data = Sasuser.ThreeG;
  class Purchase Gender;
  model Purchase = Age Gender;
  endogenous Purchase ~discrete (dist = probit);
  format Purchase Purchase_FMT.  Gender Gender_FMT.;
/*调用 QLIM 过程*/
/*指定模型所有变量中的定性变量*/
/*指定模型*/
/*指定分布形式，建立 PROBIT 模型*/
```

```
output out = Qlim_3G_Out proball;
run;
proc probit data = Sasuser.ThreeG;                                /*调用 PROBIT 过程*/
  class Purchase Gender;                                          /*指定模型所有变量中的定性变量*/
  model Purchase = Age Gender;                                    /*指定模型的因变量和自变量*/
  format Purchase Purchase_FMT.  Gender Gender_FMT.;
  output out = Probit_3G_Out prob = Estimated_P;
run;
proc logistic descending data = Sasuser.ThreeG;                  /*调用 LOGISTIC 过程*/
  class Purchase Gender;                                          /*指定模型所有变量中的定性变量*/
  model Purchase = Age Gender /link = probit;                    /*指定模型，并设定链接函数为 PROBIT*/
  format Purchase Purchase_FMT.  Gender Gender_FMT.;
  output out = Logit_3G_Out predicted = Estimated_P;
run;
proc genmod descending data = Sasuser.ThreeG;                    /*调用 GENMOD 过程*/
  class Purchase Gender;                                          /*指定模型所有变量中的定性变量*/
  model Purchase = Age Gender /d = binomial link = probit;       /*指定模型并设分布形式和链接函数*/
  format Purchase Purchase_FMT.  Gender Gender_FMT.;
  output out = Genmod_3G_Out predicted = Estimated_P;
run;
```

运行程序后，4 个过程得到的输出结果大体一致，分析方法和依据同例 10-2。

但是要注意，最主要的是 QLIM 过程、PROBIT 过程和 GENMOD 过程得到的参数估计结果相同（PROBIT 过程参数估计值正负符号相反），如图 10-8 所示。

Parameter Estimates				
Parameter	Estimate	Standard Error	t Value	Approx Pr >  t
Intercept	-4.636336	1.168080	-3.97	<.0001
Age	0.096216	0.025254	3.81	0.0001
Gender 男	0.798587	0.356136	2.24	0.0249
Gender 女	0	.	.	.

图 10-8 含有定性自变量的 PROBIT 模型参数估计及检验结果

在图 10-8 中可以看到，“Gender”和“Age”变量作为自变量对消费者购买 3G 手机的影响比较显著，因为其 P 值（“Pr>ChiSq”）均比较小。在“Parameter Estimates”表格中，截距项“Intercept”的参数估计值为 4.6363，其代表定性变量“Gender”取值为“女”时对响应概率的影响（故“Gender”变量取值为“女”时，其对应的参数估计值为“0”）。从图 10-8 还可以看出，男性消费者购买 3G 手机的概率要比女性消费者大。

如有一性别为“男”、年龄为 45 岁的样本，其购买 3G 手机的概率如下。

$$\begin{aligned} P(y=1|x) &= \Phi(-4.636336 + 0.096216 \times Age + 0.798587 \times (Gender = 男)) \\ &= \Phi(-4.636336 + 0.096216 \times 45 + 0.798587 \times 2) \\ &= \Phi(1.290558) \\ &= 0.901572 \end{aligned}$$

而性别为“女”、年龄同样为 45 岁的样本，其购买 3G 手机的概率如下。

$$\begin{aligned} P(y=1|x) &= \Phi(-4.636336 + 0.096216 \times 45) \\ &= \Phi(-0.30662) \\ &= 0.379568 \end{aligned}$$

LOGISTIC 过程得到的参数结果不同于以上 3 个过程，如图 10-9 所示。

The LOGISTIC Procedure					
Type 3 Analysis of Effects					
Effect	DF		Wald Chi-Square	Pr > ChiSq	
Age	1		14.3539	0.0002	
Gender	1		4.9890	0.0255	
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-4.2368	1.1458	13.6730	0.0002
Age	1	0.0962	0.0254	14.3539	0.0002
Gender 男	1	0.3993	0.1788	4.9890	0.0255

图 10-9 含有定性自变量的 PROBIT 模型的 LOGISTIC 过程参数估计及检验结果

在图 10-9 中，变量“Age”的参数估计值与图 10-8 中的参数估计值相同。其余变量和截距项的参数估计值不同。这是由于 LOGISTIC 过程在定性自变量处理过程中赋予的定性自变量值不同于其他 3 个过程的变量值造成的。在输出结果中，找到“Class Level Information”变量水平信息表格，如图 10-10 所示。

Class Level Information		
Class	Value	Design Variables
Gender	男	1
	女	-1

图 10-10 分类变量水平信息

虽然分类变量处理方法不一样，但是 LOGISTIC 过程对响应概率的计算结果与其他 3 个过程是一样的，有兴趣的读者可以打开 4 个过程中由 OUTPUT 语句生成的 Qlim\_3G\_Out、Probit\_3G\_Out、Logit\_3G\_Out 和 Genmod\_3G\_Out 等 4 个临时数据集进行查看。



Probit\_3G\_Out 数据集中的计算结果是因变量为“不购买”的响应概率，“购买”响应概率=1-“不购买响应概率”。

10.2.4 BINARY LOGIT 模型

二元选择 LOGIT 模型对隐变量随机误差项假定服从标准正态分布，其模型具有以下形式（详见 10.2.2 小节）。

$$P(y=1|x)=F(\beta_0+x\beta)=\frac{\exp(\beta_0+x\beta)}{1+\exp(\beta_0+x\beta)}$$

在 SAS 系统中，QLIM、GENMOD、LOGISTIC、PROBIT、CATMOD 等 5 个过程均可以建立二元 LOGIT 模型。这些过程除了建立 LOGIT 模型以进行分析之外，同样可被用于其他离散选择模型分析过程当中，因此本小节将与 10.2.3 节一样，将不再笼统介绍各个具体过程的具体语法，只在对应章节中介绍所调用的过程能够实现本章节内模型建模的功能和主要语法，且这些语法与建立二元 PROBIT 模型类似。

1. QLIM 过程的二元 LOGIT 建模

QLIM（定性数据和受限因变量模型）建立 LOGIT 模型的主要语法如下。

```
PROC QLIM <选项>;
  CLASS 变量;
  MODEL 因变量 = 自变量 /选项;
```

```

ENDOGENOUS 内生变量~DISCRETE (DISTRIBUTION = LOGIT);
OUTPUT OUT = 输出数据集名 PROBALL;
RUN;

```

其中, CLASS 语句用于指定分析过程的分类变量或定性变量(含因变量和自变量), MODEL 用于指定模型的因变量和自变量之间的关系, 如有多个自变量, 则自变量之间用空格隔开。

ENDOGENOUS 语句用于指定模型的内生变量及内生变量的形式及其分布。在单方程的 Logit 建模过程中, 因变量是离散的内生变量, 响应概率所使用的分布函数为逻辑分布, 即指定: Distribution = Logit。

OUTPUT 语句用于输出分析结果数据集, PROBALL 关键字表示利用现有样本数据计算各个样本因变量各取值的响应概率。

## 2. PROBIT 过程的二元 LOGIT 建模

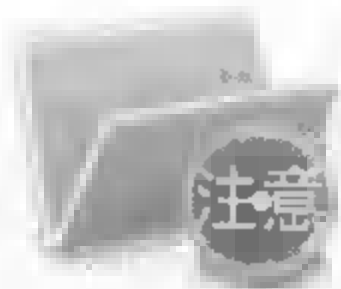
PROBIT 过程进行二元 LOGIT 建模的主要语法如下。

```

PROC PROBIT <选项>;
  CLASS 变量;
  MODEL 因变量 = 自变量 / DIST = LOGISTIC;
  OUTPUT OUT = 输出数据集名 PROB = 变量名;
RUN;

```

其中 CLASS 语句和 MODEL 语句的功能同 QLIM 过程。在 MODEL 语句的选项中加入 DIST = LOGISTIC 关键字, 表示建立的是 LOGIT 模型。



PROBIT 过程在分析过程中, 把离散因变量数值最小的值作为考察响应变量的依据, 在通常的 0-1 因变量分析中, 其分析结果是表示因变量为“0”值时的概率变动状况。因此, 在实际问题分析过程中, 考察 LOGIT 过程的参数估计结果时, 应当将各个参数估计值取负值。

OUTPUT 语句用于输出分析结果数据集, PROB 关键字表示利用现有样本数据计算各个样本的响应概率, 并指定表示该响应概率的变量名。

## 3. LOGISTIC 过程二元 LOGIT 建模

LOGISTIC 主要用于 LOGIT 模型的构建与分析, 其主要语法如下。

```

PROC LOGISTIC <选项>;
  CLASS 变量;
  MODEL <(EVENT = '响应值')> 因变量 = 自变量;
  OUTPUT OUT = 输出数据集名 PREDICTED = 变量名;
RUN;

```

LOGISTIC 过程的 MODEL 语句中省略 DIST 关键字, 表示默认建立 LOGIT 模型。

与 PROBIT 过程一样, LOGISTIC 过程在分析过程中同样把离散因变量数值最小的值作为考察响应变量的依据, 但是 LOGISTIC 过程可以通过指定 DESCENDING 关键字把离散因变量数值最大的值作为考察响应变量概率的依据。此外, LOGISTIC 过程还可以通过 MODEL 语句中的 EVENT 选项考察因变量不同响应值的建模情况。

OUTPUT 语句用于输出分析结果数据集, PREDICTED 关键字表示利用现有样本数据计

算各个样本的响应概率，并指定表示该响应概率的变量名。

4. GENMOD 过程的二元 LOGIT 建模

GENMOD（广义线性模型）过程同样也可用于 LOGIT 建模，其有关的主要语法如下。

```
PROC GENMOD <选项>;
  CLASS 变量;
  MODEL 因变量 = 自变量 / DIST = BINOMIAL LINK = LOGIT;
  OUTPUT OUT = 输出数据集名 PREDICTED = 变量名;
RUN;
```

与 PROBIT、LOGISTIC 过程一样，GENMOD 过程在分析过程中同样把离散因变量数值最小的值作为考察响应变量的依据，但是可以通过指定 DESCENDING 关键字把离散因变量数值最大的值作为考察响应变量概率的依据。

OUTPUT 语句用于输出分析结果数据集，PREDICTED 关键字表示利用现有样本数据计算各个样本的响应概率，并指定表示该响应概率的变量名。

此外 GENMOD 除了通过 MODEL 语句中的 LINK 关键字指定链接函数之外，用户还可以通过 FWDLINK 和 INVLINK 关键字自行构建链接函数形式以进行建模。

5. CATMOD 过程的二元 LOGIT 建模

CATMOD（分类数据模型）过程同样也可用于 LOGIT 建模，其有关的主要语法如下。

```
PROC CATMOD <选项>;
  DIRECT 变量;
  MODEL 因变量 = 自变量 / 选项;
  RESPONSE LOGIT OUT = 输出数据集名;
RUN;
```

CATMOD 过程可以通过选项指定关键字“ORDER = FREQ”，把离散因变量数值出现最多的值作为考察响应变量概率的依据。因此在二元选择模型中，要首先分清楚离散因变量的两个值出现的次数。如果没有 ORDER 关键字，其分析的是离散变量数值最小值（即如果是 0-1 变量，则分析“0”值）的响应概率。

其中的 DIRECT 语句主要用于指定模型中的定量变量。而 RESPONSE 语句用于指定计算响应概率的函数，并可以把计算结果存储在指定的输出数据集当中，其关键字 LOGIT 表示使用条件 LOGIT 链接函数计算边际概率。如果 RESPONSE 语句省略关键字，则系统默认使用条件 LOGIT 函数进行边际概率计算。

由于在 CATMOD 过程的建模过程中，对自变量中含有分类变量或定性变量的模型的约束与前面几个过程对模型的约束不一样，导致其参数估计结果有所不同（如没有定性自变量，则参数估计结果与前面 4 个过程结果相同）。因此，通常不建议使用该过程进行二元 LOGIT 建模。

本节中，仍然以 10.2.3 小节中的例 10-2 和例 10-3 为例进行 LOGIT 建模。

对于例 10-2 的数据，建立二元 LOGIT 模型的具体程序如下。

```
proc format;
  value Attitude_FMT      1='持续购买'      /*为离散因变量指定变量值标签*/
                          0='不持续购买';
```

```
run;
proc qlim data = Sasuser.Product_Usage;          /*调用 QLIM 过程对 Product_Usage 数据集进行分析*/
  class Attitude;                                /*指定 Attitude 变量为定性或分类变量*/
  model Attitude = CSI Complaint Loyalty;          /*指定离散因变量及其影响因素（自变量）的模型*/
  endogenous Attitude~discrete (dist = logit);    /*指定内生变量，并采用逻辑分布*/
  format Attitude Attitude_FMT.;                /*为因变量值挂上标签*/
  output out = Qlim_L_Out proball;                /*指定输出数据集名为 Qlim_L_Out 临时数据集*/
run;
proc probit data = Sasuser.Product_Usage;          /*调用 PROBIT 过程*/
  class Attitude;
  model Attitude = CSI Complaint Loyalty /dist = logistic; /*指定模型的因变量和自变量*/
  format Attitude Attitude_FMT.;
  output out = Probit_L_Out prob = Estimated_P; /*指定输出数据集名为 Probit_L_Out 临时数据集*/
run;
proc logistic descending data=Sasuser.Product_Usage; /*调用 LOGISTIC 过程*/
  class Attitude;
  model Attitude = CSI Complaint Loyalty;          /*指定模型*/
  format Attitude Attitude_FMT.;
  output out=Logit_L_Out predicted=Estimated_P;    /*指定输出数据集名为 Logit_L_Out 临时数据集*/
run;
proc genmod descending data=Sasuser.Product_Usage; /*调用 GENMOD 过程*/
  class Attitude;
  model Attitude = CSI Complaint Loyalty /d=binomial link=logit; /*指定模型并设分布形式和链接函数*/
  format Attitude Attitude_FMT.;
  output out = Genmod_L_Out predicted = Estimated_P; /*指定输出数据集名为 Genmod_L_Out*/
run;
proc catmod order = freq data = Sasuser.Product_Usage; /*调用 CATMOD 过程*/
  direct CSI Complaint Loyalty;                    /*指定模型中的定量变量*/
  model Attitude = CSI Complaint Loyalty;          /*指定模型*/
  format Attitude Attitude_FMT.;
  response logit out = Catmod_L_Out; /*指定用 LOGIT 分布计算响应概率，输出数据集名为 Catmod_L_Out*/
quit;
run;
```

运行上述程序后，5 个过程输出的参数估计及其检验结果基本一致（注意 PROBIT 过程的输出参数估计结果的符号与其他过程相反），如图 10-11 所示。

Testing Global Null Hypothesis: BETA=0					
Test		Chi-Square	DF	Pr > ChiSq	
Likelihood Ratio		102.2179	3	<.0001	
Score		98.7327	3	<.0001	
Wald		62.0028	3	<.0001	
Analysis of Maximum Likelihood Estimates					
Parameter	DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept	1	-3.4956	1.0053	12.0914	0.0005
CSI	1	0.4893	0.1190	16.9143	<.0001
Complaint	1	-0.2347	0.0829	8.0239	0.0046
Loyalty	1	0.3367	0.0783	18.4755	<.0001

图 10-11 LOGIT 模型的参数估计及检验结果（以 LOGISTIC 过程结果为例）

在图 10-11 所示的结果中，对模型总体显著性（“Testing Global Null Hypothesis: BETA = 0”）进行了几种方式的检验，如似然比（“Likelihood Ratio”）检验等。几种检验结果表明，模型总体非常显著（“Pr>ChiSq”值均小于 0.0001）。对于 LOGIT 模型中各参数估计的结果（“Analysis of Maximum Likelihood Estimates”）也较显著，各参数的显著性检验  $P$  值（“Pr>ChiSq”）均较小。

因此，根据图 10-11 的参数估计结果，可得到例 10-2 的 LOGIT 模型如下。

$$P(y=1|x) = F(\beta_0 + x\beta) = \frac{\exp(\beta_0 + x\beta)}{1 + \exp(\beta_0 + x\beta)}$$

$$= \frac{\exp(-3.4956 + 0.4893 \times CSI - 0.2347 \times Complaint + 0.3367 \times Loyalty)}{1 + \exp(-3.4956 + 0.4893 \times CSI - 0.2347 \times Complaint + 0.3367 \times Loyalty)}$$

如现有对某个消费者的调查访问数据，调查显示其满意度得分为 8，抱怨程度得分为 4，忠诚度得分为 7，现可以依据所建立的模型考察其“持续购买”产品的概率。把上述自变量的数值代入以上模型中，可得到：

$$P(y=1|x) = \frac{\exp(-3.4956 + 0.4893 \times 8 - 0.2347 \times 4 + 0.3367 \times 7)}{1 + \exp(-3.4956 + 0.4893 \times 8 - 0.2347 \times 4 + 0.3367 \times 7)}$$

$$= \frac{\exp(1.8369)}{1 + \exp(1.8369)}$$

$$= 0.8626$$

即该消费者持续购买该企业产品的概率为 0.8626，该样本 LOGIT 模型的响应概率与 PROBIT 模型的响应概率（0.8528）略有差异。

上述程序所输出的结果数据集分别为 Qlim\_L\_Out、Probit\_L\_Out、Logit\_L\_Out、Genmod\_L\_Out 和 Catmod\_L\_Out，其各自响应概率结果均相同，且其模型和结果分析过程相同，读者可自行查看对应的结果及输出的数据集以进行对比分析，此处不予赘述。

在 LOGIT 模型的影响因素中，同样可加入定性变量作为自变量以进行研究。如对例 10-3 所示数据进行二元 LOGIT 建模，具体程序如下。

```
proc format;
  value Purchase_FMT      1='购买'
  0='不购买';
  value Gender_FMT        1='女'
  2='男';
run;
proc qlim data = Sasuser.ThreeG;          /*调用 QLIM 过程*/
  class Purchase Gender;                  /*指定模型所有变量中的定性变量*/
  model Purchase = Age Gender;             /*指定模型*/
  endogenous Purchase ~discrete (dist = logit); /*指定分布形式，建立 LOGIT 模型*/
  format Purchase Purchase_FMT.  Gender Gender_FMT.;
  output out = Qlim_3G_L_Out proball;
run;
proc probit data = Sasuser.ThreeG;        /*调用 PROBIT 过程*/
  class Purchase Gender;                  /*指定模型所有变量中的定性变量*/
  model Purchase = Age Gender /dist = logistic; /*指定模型的因变量和自变量，并指定分布形式，建立 LOGIT 模型*/
```

```
format Purchase Purchase_FMT. Gender Gender_FMT.;
output out = Probit_3G_L_Out prob = Estimated_P;
run;
proc logistic descending data = Sasuser.ThreeG; /*调用 LOGISTIC 过程*/
  class Purchase Gender; /*指定模型所有变量中的定性变量*/
  model Purchase = Age Gender; /*指定模型*/
  format Purchase Purchase_FMT. Gender Gender_FMT.;
  output out = Logit_3G_L_Out predicted = Estimated_P;
run;
proc genmod descending data = Sasuser.ThreeG; /*调用 GENMOD 过程*/
  class Purchase Gender; /*指定模型所有变量中的定性变量*/
  model Purchase = Age Gender /d = binomial link = logit; /*指定模型并设分布形式和链接函数*/
  format Purchase Purchase_FMT. Gender Gender_FMT.;
  output out = Genmod_3G_L_Out predicted = Estimated_P;
run;
proc catmod order = freq data = Sasuser.ThreeG; /*调用 CATMOD 过程*/
  direct Age; /*指定模型所有变量中的定量变量*/
  model Purchase = Age Gender; /*指定模型*/
  format Purchase Purchase_FMT. Gender Gender_FMT.;
  response logit out = Catmod_3G_L_Out;
quit;
run;
```

与 PROBIT 模型类似，由于 LOGISTIC 过程的分类变量在模型中的值的设定不同于其他过程，故其参数估计结果不同；而由于 CATMOD 过程的模型约束不同，其参数估计结果也有所不同。其余过程的结果基本一致。

运行程序之后，可得到 QLIM\_3G\_L\_Out、PROBIT\_3G\_L\_Out、LOGIT\_3G\_L\_Out、Genmod\_3G\_L\_Out 和 Catmod\_3G\_L\_Out 等 5 个临时数据集，其中存储依据参数估计结果计算的响应概率。5 个数据集的响应概率值完全一致，读者可自行打开这些数据集以进行查看。

对于社会经济分析过程中的实际二元选择问题，究竟是建立 PROBIT 模型还是建立 LOGIT 模型，这是一个不确定的问题。对于绝大多数的数据而言，两种模型所得到的结果非常接近。

此外，除了利用 SAS 程序进行 LOGIT 建模之外，还可使用 SAS/Analyst 模块进行分析。

**STEP 1)** 以例 10-3 为例，进入 SAS/Analyst，打开 ThreeG.sas7bdat 数据集，选择系统菜单 “Statistics → Regression → Logistic”，弹出 “Logistic” 对话框，如图 10-12 所示。

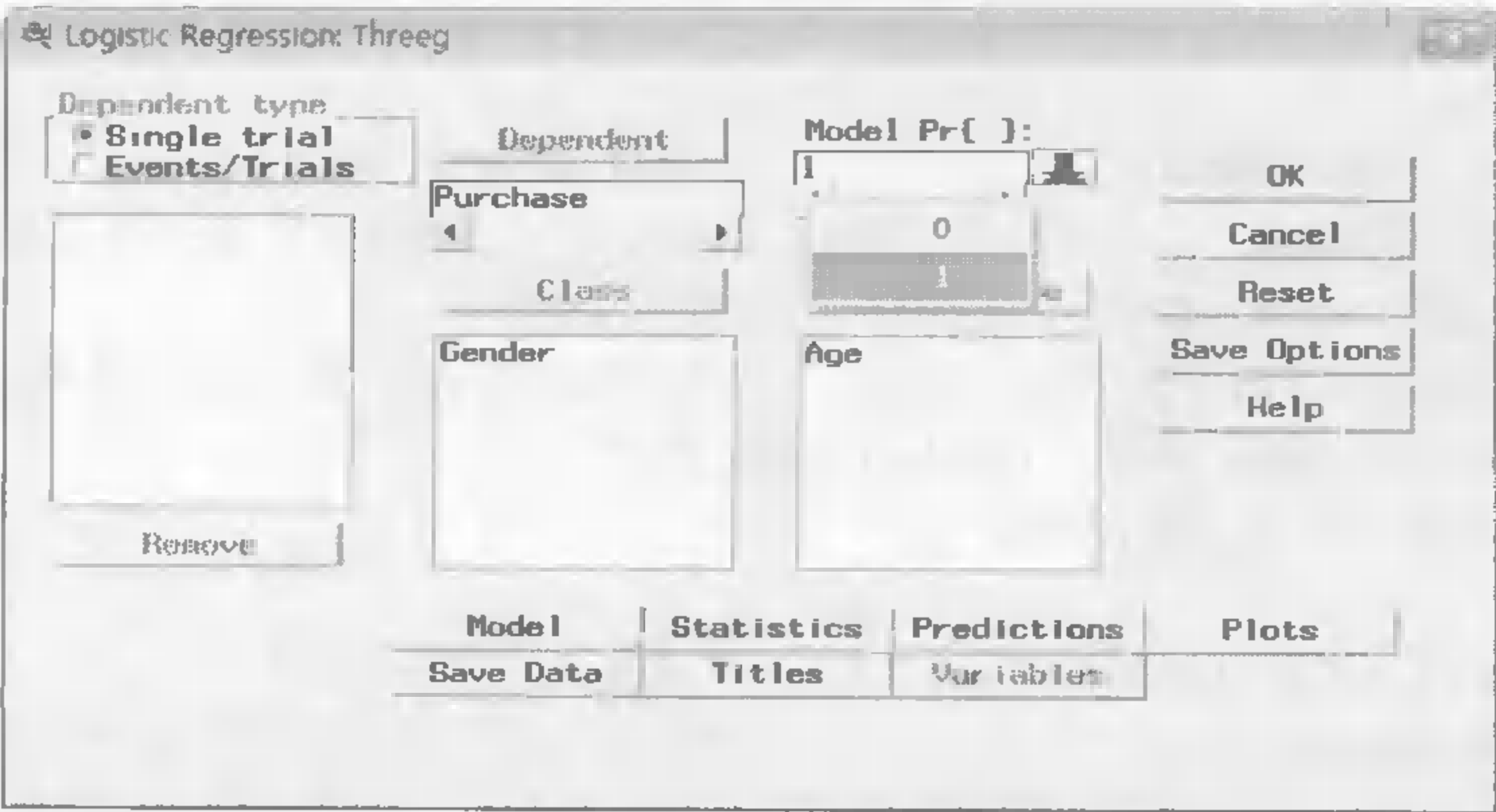



图 10-12 “Logistic” 对话框

**STEP 2** 在“Logistic”对话框左上角的“Dependent type”分栏下选择“Single trial”单选框，表示考察的是离散变量。从变量选择区域中选中“Purchase”变量，单击“Dependent”按钮，将其指定为因变量。然后在“Model Pr{ }:”下拉选单的区域中单击  按钮，在弹出的因变量响应值的选单中选择“1”，表示分析“购买”态度的响应概率（因为因变量 Purchase 值为“1”时表示“购买”，值为“0”时表示不够买）。从变量选择区域中选中“Gender”变量，单击“Class”按钮，将其指定为分类变量；选中“Age”变量，单击“Quantitative”按钮，将其指定为数值型变量。

**STEP 3** 为了得到响应概率的结果，可以选择依据已建立的模型对响应概率进行预测。单击“Predictions”按钮，弹出预测对话框，如图 10-13 所示。

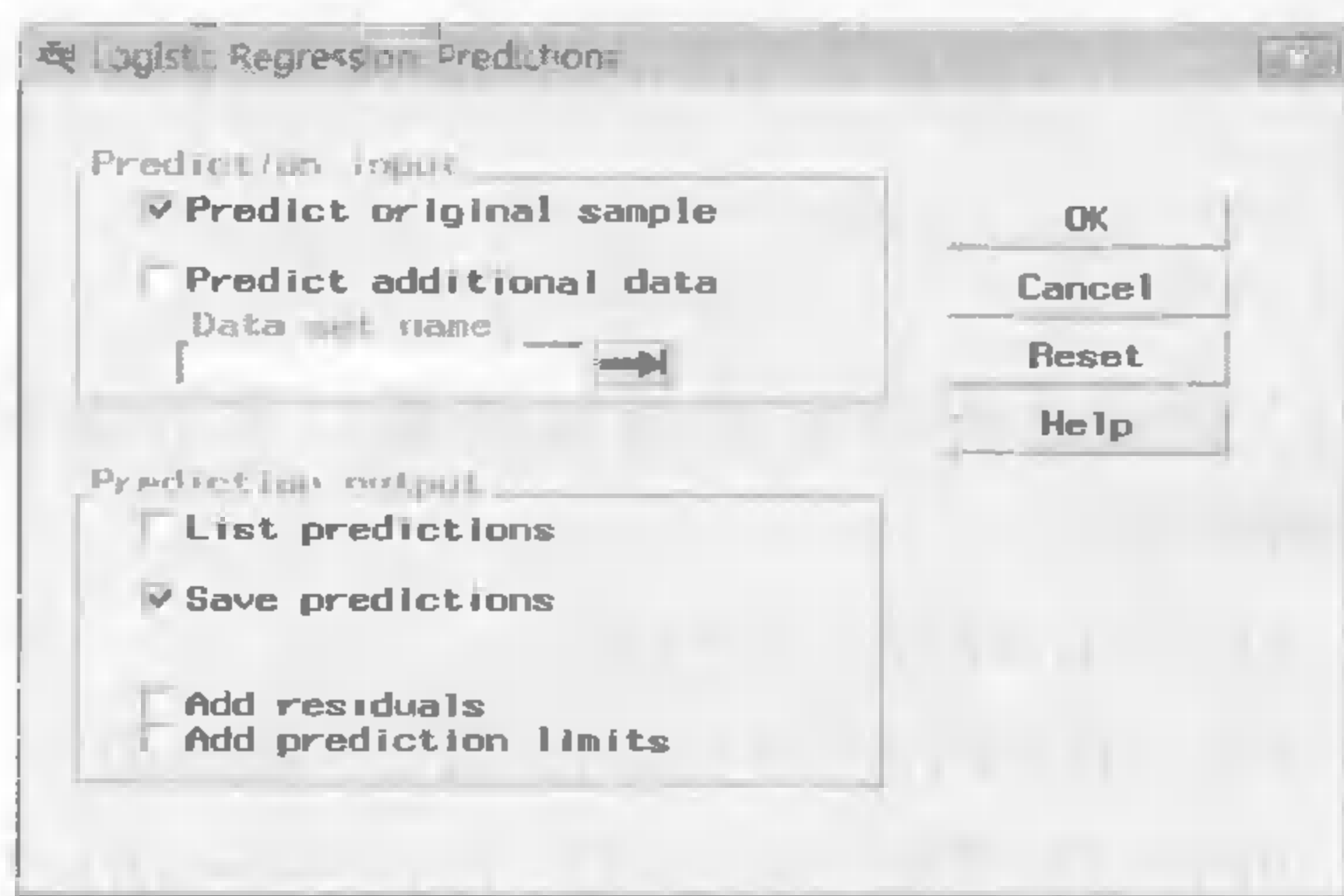


图 10-13 “Predictions”对话框

**STEP 4** 在“Predictions”对话框中的“Prediction input”分栏下，可以选择进行预测的数据来源。“Predict original sample”复选框表示利用现有分析数据集中的数据进行预测，“Predict additional data”复选框表示利用其他给定的数据集进行预测，给定的数据集的路径及名称可以在“Data set name”文本输入框中自行指定。

**STEP 5** “Prediction output”分栏则可用于设定预测结果的显示形式，“List predictions”复选框表示在输出结果中把预测结果显示出来，而“Save predictions”复选框表示可以把输出结果存储到一个数据集当中。此外还可以为预测结果增加残差项和响应概率的预测区间。本例选择利用现有数据进行预测，并把预测结果存储到数据集当中。单击“OK”按钮返回“Logistic”对话框，在该对话框中，单击“OK”按钮，便可得到与利用 SAS 程序的 LOGISTIC 过程编程一致的结果，此处不再赘述。

## 10.3 多重选择模型

前面章节分析的问题都只有两个选项，而更一般的现实问题往往可能有多个选项。如对某项措施的看法，可能有非常满意、满意、不满意、非常不满意等 4 种选择，同样可以用多重选择离散变量进行描述。对于分析有多个选择的离散因变量，可以利用本节介绍的多重选择模型（Multiple Choice）进行具体分析。

### 10.3.1 多重选择模型的基本原理

在多重选择模型中，不同选项或选择结果之间的关系有两种情况。一种情况是各项选择

是有顺序之分的，如上述例子中的非常满意、满意、不满意、非常不满意 4 个选项，这 4 个选项是按照满意的程度依次进行选择的，它反映了被调查者的不同偏好程度；而另一种情况是选项之间没有次序或顺序之分，如选择公交、私家车、自行车、地铁等交通工具出行，各种选择之间没有次序上的联系。

因此，依据离散因变量选项的含义和次序不同，多重选择模型又可以分为顺序选择 (Ordinal) 模型和无序选择 (Multinomial) 模型。

对于顺序选择模型，设有 0、1、2、M 种选择。仍然设  $y^*$  是一个由  $y^* = \alpha + x\beta + \mu^*$ 、
$$y = \begin{cases} 0 & y^* < c_1 \\ 1 & c_1 \leq y^* < c_2 \\ 2 & c_2 \leq y^* < c_3 \\ \vdots & \vdots \\ M & c_M \leq y^* \end{cases}$$
 所决定的、不能够直接观测得到的隐变量，其中  $c_1, c_2, K, c_M$  可被称为临界

值或统称为截距项， $\mu_i^*$  是独立同分布的随机变量，其分布函数为  $F(x)$ ，则选择  $j(j=0,1,2,K,M)$  方案的概率如下。

$$\begin{aligned} P(y=0) &= F(c_1 - \alpha - x\beta) \\ P(y=1) &= F(c_2 - \alpha - x\beta) - F(c_1 - \alpha - x\beta) \\ P(y=2) &= F(c_3 - \alpha - x\beta) - F(c_2 - \alpha - x\beta) \\ &\vdots \\ P(y=M) &= 1 - F(c_M - \alpha - x\beta) \end{aligned}$$

当分布函数  $F(x)$  为标准正态分布时，上述模型被称为 Ordinal Probit 模型；当分布函数  $F(x)$  为逻辑分布时，则被称为 Ordinal Logit 模型。

同理，无序选择模型也可以被细分为 Multinomial Probit 模型和 Multinomial Logit 模型。由于 Multinomial Probit 模型在实际的参数估计过程中实现起来非常困难，在分析时较少采用该类模型，因此本节将主要介绍无序选择模型中的 Multinomial Logit 模型。

对于无序选择模型，选择选项  $j$  的模型基本形式如： $y_j^* = x_j\beta + \varepsilon$

在影响因素  $x_j$  的作用下，选择选项  $j$  表示  $y_j^*$  在所有选项中的效用是最大的。因此，选择选项  $j$  的概率模型为： $P(y=j) = P(y_j^* > y_k^*) \quad \forall k \neq j$

$\varepsilon^*$  是独立同分布的随机变量，一般假定其服从 Weibull 分布，具体的分布函数为  $F(z) = e^{-e^{-z}}$ ，则有：

$$P(y=j) = \frac{e^{x_j\beta}}{\sum_{k \neq j} e^{x_k\beta}}$$

此即为条件 LOGIT 模型。

此外，对于多重选择模型，其估计和检验过程与二元选择模型类似。

顺序选择模型中离散因变量的各种选择通常也是用阿拉伯数字代表，如对于某个产品的偏好程度作出选择“1”-喜欢，“2”-无所谓，“3”-不喜欢。这些数字之间只有顺序意义，没有数值上的意义。即这些所选择的数字仅能代表偏好的程度，如选择“1”所代表的偏好程度

就比选择“3”所代表的偏好程度要高；所选择的数字之间不能够进行数值运算。



在顺序选择模型中，选项含义顺序及选项代码的顺序均必须得到保证，否则会出现模型分析过程中的混乱。如将上述的偏好程度修改成：1-喜欢，2-不喜欢，3-无所谓。虽然代码的顺序得到了保证，但是代码所代表的选项含义的顺序没有得到保证，因为通常就偏好程度而言，一般认为“无所谓”的程度要比“不喜欢”的喜好程度要高。

### 10.3.2 ORDINAL PROBIT 模型

Ordinal Probit 模型是随机误差项分布函数  $F(x)$  为标准正态分布时的顺序选择模型，即选择  $j(j=0,1,2,K,M)$  方案的概率如下。

$$P(y=0) = F(c_1 - \alpha - x\beta)$$

$$P(y=1) = F(c_2 - \alpha - x\beta) - F(c_1 - \alpha - x\beta)$$

$$P(y=2) = F(c_3 - \alpha - x\beta) - F(c_2 - \alpha - x\beta)$$

L

$$P(y=M) = 1 - F(c_M - \alpha - x\beta)$$

因此，顺序选择模型中比较关键的步骤便是确定上述模型中的参数估计值。这些参数估计值不仅仅是隐变量  $y^*$  的参数估计值，更重要、也最容易混淆的是  $y^*$  临界点即截距项  $c_M$  的估计值。在应用截距项求解因变量的响应概率时，切记不要混淆截距项与选项对应的顺序。

为了保证选项的顺序在分析过程中不被混淆，通常在进行顺序选择模型时，可先对离散因变量进行排序，利用排序后的数据进行模型建模分析。

在 SAS 系统中，QLIM 过程、PROBIT 过程和 LOGISTIC 过程均可实现 Ordinal Probit 建模。

#### 1. QLIM 过程的 ORDINAL PROBIT 建模

QLIM 建立 Ordinal Probit 模型的主要语法如下：

```
PROC QLIM <选项>;
  CLASS 变量;
  MODEL 因变量 = 自变量 /选项;
  ENDOGENOUS 内生变量~DISCRETE (<ORDER = DATA> DISTRIBUTION = NORMAL);
  OUTPUT OUT = 输出数据集名 PROBALL;
RUN;
```

其中，CLASS 语句用于指定分析过程的分类变量或定性变量（含因变量和自变量），MODEL 用于指定模型的因变量和自变量之间的关系，如有多个自变量，则自变量之间用空格隔开。

ENDOGENOUS 语句用于指定模型的内生变量及内生变量的形式及其分布。在单方程的建模过程中，因变量是离散的内生变量，响应概率所使用的分布函数为标准正态分布，即指定：Distribution = Normal。QLIM 默认进行 Probit 分析，因此该关键字可以省略。此外，“ORDER = DATA”关键字表示在进行响应概率分析时，指定顺序选择模型分析的顺序（即截距项与选项顺序的对应关系）按照数据集中所指定的顺序进行分析。



使用“ORDER = DATA”关键字时，如果使用 FORMAT 语句定义了因变量的标签，则系统自动按照标签进行响应变量选项排序。这点与其余 PROBIT 和 LOGISTIC 过程不一样，非常容易让人混淆。

OUTPUT 语句用于输出分析结果数据集，PROBALL 关键字表示利用现有样本数据计算各个样本因变量各种选择的响应概率。因此，建议在分析顺序选择模型时，首选该过程。

## 2. PROBIT 过程的 ORDINAL PROBIT 建模

PROBIT 过程进行 Ordinal Probit 建模的主要语法如下：

```
PROC PROBIT <ORDER=DATA> <选项>;
  CLASS 变量;
  MODEL 因变量 = 自变量 / DIST = NORMAL;
  OUTPUT OUT = 输出数据集名 PROB = 变量名;
RUN;
```

其中 CLASS 语句和 MODEL 语句的功能同 QLIM 过程。在 MODEL 语句的选项中加入“DIST = NORMAL”关键字（可省略），表示建立的是 PROBIT 模型。

与二元选择的 PROBIT 模型一样，在实际问题分析过程中，考察 PROBIT 过程的参数估计结果时，应当将各个参数估计值取负值。

OUTPUT 语句用于输出分析结果数据集，PROB 关键字表示利用现有样本数据计算各个样本的响应概率，并指定表示该响应概率的变量名。



如果因变量有 K 种选择，则在 PROBIT 过程中由 OUTPUT 语句输出的含有响应概率的数据集中，只能得到 K-1 个响应概率，不会给出次序最大或最高的选择对应的响应概率。

## 3. LOGISTIC 过程的 ORDINAL PROBIT 建模

LOGISTIC 过程进行 Ordinal Probit 建模的主要语法如下：

```
PROC LOGISTIC <ORDER = DATA> <选项>;
  CLASS 变量;
  MODEL <(EVENT = '响应值')> 因变量 = 自变量 / LINK = PROBIT;
  OUTPUT OUT = 输出数据集名 PREDICTED = 变量名;
RUN;
```

与 PROBIT 过程一样，LOGISTIC 过程在分析过程中同样把离散因变量的最小值作为考察响应变量的依据，但是 LOGISTIC 过程可以通过指定 DESCENDING 关键字把离散因变量数值最大的值作为考察响应变量概率的依据。此外，LOGISTIC 过程还可以通过 MODEL 语句中的 EVENT 选项考察因变量不同响应值的建模情况。

OUTPUT 语句用于输出分析结果数据集，PREDICTED 关键字表示利用现有样本数据计算各个样本的响应概率，并指定表示该响应概率的变量名。与 PROBIT 过程一样，如果因变量有 K 种选择，则在 LOGISTIC 过程中由 OUTPUT 语句输出的含有响应概率的数据集中，只能得到 K-1 个相应概率，不会给出次序最大或最高的选择对应的相应概率。

QLIM 过程、PROBIT 过程和 LOGISTIC 过程中的 ORDER = DATA 关键字可以省略，并且可以依据其运行程序之后的输出结果中的“Response Profile”信息，找到模型参数估计中各截距项所对应的选项信息。此处也非常容易让人混淆，建议进行分析时先对数据按照因变量进行排序，然后加上“ORDER = DATA”关键字，以免在分析的过程中发生错误。



例 10-4

在例 10-3 中，针对不同性别、年龄的潜在客户进行了一次小规模摸底预调查。在预调查中发现了一个问题，即问卷所设计的“购买意向”问题太过于笼统，不能全面反映消费者的真实意愿。因此，为了更好地细分消费者群体，对该问题进行了改进，即对该问题的选项进行细化，主要考察消费者“1-不购买”、“2-无所谓”、“3-购买”等购买意向。调查结果（数据详见 ThreeG\_Multi.sas7bdat）如表 10-4 所示。

表 10-4 3G 手机购买意向改进调查结果

性别 Gender	年龄 Age	购买意向 Purchase	性别 Gender	年龄 Age	购买意向 Purchase	性别 Gender	年龄 Age	购买意向 Purchase
2	54	1	1	50	1	1	39	2
1	35	2	1	45	2	1	51	1
1	59	1	2	31	3	1	46	2
1	35	1	1	34	2	1	46	2
2	38	2	2	47	3	2	50	2
2	34	3	2	39	3	1	35	2
1	56	1	1	48	2	1	52	2
2	54	2	1	59	1	1	39	2
1	47	2	1	51	1	1	39	1
2	46	3	1	45	1	2	47	3
2	58	2	1	32	2	2	44	2
1	52	1	2	39	2	1	56	3
1	30	3	1	39	2	2	34	2
1	47	2	1	50	2	1	34	2
2	59	1	1	52	1	2	49	2
1	46	2	1	43	2	1	47	1
1	46	3	1	47	2	1	43	2
1	39	2	2	58	1	2	38	3
2	45	3	1	51	1	1	51	1
2	31	2	1	35	2	2	46	2
1	32	2	2	59	1	2	42	3
1	52	1	1	48	1			

本例中的“1-不购买”、“2-无所谓”、“3-购买”等购买意向是根据消费的偏好程度由弱至强的顺序进行设置的。因此，在表 10-4 中，购买意向的变量值是有顺序之分的，符合顺序选择模型。

选择 QLIM 过程进行分析，具体程序如下：

```
proc format;
  value Purchase_FMT      1='1-不够买'
                        2='2-无所谓'
                        3='3-购买';
  value Gender_FMT        1='女'
                        2='男';
proc sort data = Sasuser.ThreeG_Multi;
  by Purchase;
run;                                /*为避免混淆选项顺序与截距项顺序，对数据集按照因变量进行排序*/
proc qlim data = Sasuser.ThreeG_Multi; /*调用 QLIM 过程*/
  class Purchase Gender;           /*指定定性变量*/
  model Purchase = Gender Age;      /*指定模型因变量与自变量*/
  endogenous Purchase~ discrete (order = data dist = normal); /*指定模型内生变量及分布*/
  format Purchase Purchase_FMT. Gender Gender_FMT.;
  output out = Qlim_OP_Out proball; /*计算因变量所有选项的响应概率并将它们存储在指定数据集当中*/
run;
```

运行程序后，可得到关于因变量响应的基本信息，如图 10-14 所示。

The QLIM Procedure			
Discrete Response Profile of Purchase			
Index	Value	Frequency	Percent
1	1-不购买	20	30.77
2	2-无所谓	33	50.77
3	3-购买	12	18.46

图 10-14 QLIM 过程中的因变量响应信息

图 10-14 描述了离散因变量的响应情况，“Index”列表示选择的顺序。该顺序与进行参数估计之后模型的截距项顺序是对应的，如图 10-15 所示。

Parameter Estimates					
Parameter		Estimate	Standard Error	t Value	Approx Pr >  t
Intercept		4.033489	0.945150	4.27	<.0001
Gender	男	1.113006	0.327241	3.40	0.0007
Gender	女	0			
Age		-0.083372	0.020181	-4.13	<.0001
Limit2		1.857990	0.278738	6.67	<.0001

图 10-15 QLIM 过程 Ordinal Probit 模型的参数估计及检验结果

图 10-15 所示的参数估计结果经过检验，均非常显著，因此可以对模型进行进一步分析。根据图 10-14 所示的选择顺序和图 10-15 所示的参数估计结果，可得到以下响应概率。

$$P(\text{不购买}) = P(y = 1) = \Phi(c_1 - \alpha - x\beta)$$
$$P(\text{无所谓}) = P(y = 2) = \Phi(c_2 - \alpha - x\beta) - \Phi(c_1 - \alpha - x\beta)$$
$$P(\text{购买}) = P(y = 3) = 1 - \Phi(c_2 - \alpha - x\beta)$$

其中， $\alpha = Intercept = 4.033489$ ， $c_1 = 0$ ， $c_2 = \_Limit2 = 1.857990$ 。

在 QLIM 过程的参数估计结果中，“Intercept”表示模型本身具有的截距项  $\alpha$ ，并且默认第一个选项的临界值或截距项  $c_1$  为 0。此外，把定性自变量转化为 0-1 变量（即 1-男、0-女，因为“女”的参数估计值为 0）的形式。

如有一性别为“男”、年龄为 45 岁的样本，把参数估计结果代入上述模型，得到该消费者购买 3G 手机的响应概率如下。

$$\begin{aligned}
 P(\text{不购买}) &= P(y=1) = \Phi(c_1 - \alpha - x\beta) \\
 &= \Phi(-4.033489 - 1.113006 \times 1 + 0.083372 \times 45) \\
 &= \Phi(-1.39476) \\
 &= 0.0815
 \end{aligned}$$

$$\begin{aligned}
 P(\text{无所谓}) &= P(y=2) = \Phi(c_2 - \alpha - x\beta) - \Phi(c_1 - \alpha - x\beta) \\
 &= \Phi(1.85799 - 4.033489 - 1.113006 + 0.083372 \times 45) - \Phi(-1.39476) \\
 &= \Phi(0.463235) - \Phi(-1.39476) \\
 &= 0.6784 - 0.0815 \\
 &= 0.5969
 \end{aligned}$$

$$\begin{aligned}
 P(\text{购买}) &= P(y=3) = 1 - \Phi(c_2 - \alpha - x\beta) \\
 &= 1 - \Phi(1.85799 - 4.033489 - 1.113006 + 0.083372 \times 45) \\
 &= 1 - \Phi(0.463235) \\
 &= 0.3216
 \end{aligned}$$

对于每个样本的各选项的响应概率，在本例输出的临时数据集 Qlim\_OP\_Out.sas7bdat 中均可看到。

用 PROBIT 过程和 LOGISTIC 过程对本例数据进行 Ordinal Probit 建模的程序如下：

```

proc format;
  value Purchase_FMT      1='1-不购买'
                          2='2-无所谓'
                          3='3-购买';
  value Gender_FMT        1='女'
                          2='男';
proc sort data = Sasuser.ThreeG_Multi;
  by Purchase;
run;
proc probit order = data data = Sasuser.ThreeG_Multi; /*调用 PROBIT 过程*/
  class Purchase Gender; /*指定分类变量*/
  model Purchase = Gender Age; /*指定模型中的因变量和自变量*/
  format Purchase Purchase_FMT. Gender Gender_FMT.;
  output out = Probit_OP_Out prob = Estimated_P; /*指定输出含有响应概率的数据集*/
run;
proc logistic order = data data = Sasuser.ThreeG_Multi; /*调用 LOGISTIC 过程*/
  class Purchase Gender; /*指定分类变量*/
  model Purchase = Gender Age/link = probit; /*指定模型中的因变量和自变量，并确定链接函数为
PROBIT*/
  format Purchase Purchase_FMT. Gender Gender_FMT.;
  output out = Logit_OP_Out prob = Estimated_P; /*指定输出含有响应概率的数据集*/
run;

```

运行程序后可到与 QLIM 过程类似的结果，但是要注意转换 PROBIT 过程的参数估计结果的符号。此外，在这两个过程由 OUTPUT 语句输出的数据集中，如果因变量有 K 种选择，

只能得到  $K-1$  个响应概率，不会给出次序最大或最高的选择对应的响应概率。

### 10.3.3 ORDINAL LOGIT 模型

Ordinal Logit 模型是随机误差项分布函数  $F(x)$  为逻辑分布时的顺序选择模型，即选择  $j(j=0,1,2,K,M)$  方案的概率如下。

$$\begin{aligned} P(y=0) &= \frac{\exp(c_1 - \alpha - x\beta)}{1 + \exp(c_1 - \alpha - x\beta)} \\ P(y=1) &= \frac{\exp(c_2 - \alpha - x\beta)}{1 + \exp(c_2 - \alpha - x\beta)} - \frac{\exp(c_1 - \alpha - x\beta)}{1 + \exp(c_1 - \alpha - x\beta)} \\ P(y=2) &= \frac{\exp(c_3 - \alpha - x\beta)}{1 + \exp(c_3 - \alpha - x\beta)} - \frac{\exp(c_2 - \alpha - x\beta)}{1 + \exp(c_2 - \alpha - x\beta)} \\ &\vdots \\ P(y=M) &= 1 - \frac{\exp(c_M - \alpha - x\beta)}{1 + \exp(c_M - \alpha - x\beta)} \end{aligned}$$

在 SAS 系统中，QLIM 过程、PROBIT 过程和 LOGISTIC 过程均可实现 Ordinal Logit 建模。

#### 1. QLIM 过程的 ORDINAL LOGIT 建模

QLIM 建立 Ordinal Logit 模型的主要语法如下：

```
PROC QLIM <选项>;
  CLASS 变量;
  MODEL 因变量 = 自变量 /选项;
  ENDOGENOUS 内生变量~DISCRETE(<ORDER = DATA> DISTRIBUTION = LOGISTIC);
  OUTPUT OUT = 输出数据集名 PROBALL;
RUN;
```

其中，CLASS 语句用于指定分析过程的分类变量或定性变量（含因变量和自变量），MODEL 用于指定模型的因变量和自变量之间的关系，如有多个自变量，则自变量之间用空格隔开。

ENDOGENOUS 语句用于指定模型的内生变量及内生变量的形式及其分布。在单方程的建模过程中，因变量是离散的内生变量，响应概率所使用的分布函数为逻辑分布，即指定：Distribution = Logistic。此外，“ORDER = DATA”关键字表示在进行响应概率分析时，指定顺序选择模型分析的顺序（即截距项与选项顺序的对应关系）按照数据集中所指定的顺序进行分析。



使用“ORDER = DATA”关键字时，如果使用 FORMAT 语句定义了因变量的标签，则系统自动按照标签进行响应变量选项排序。这点与其余 PROBIT 和 LOGISTIC 过程不一样，非常容易让人混淆。

OUTPUT 语句用于输出分析结果数据集，PROBALL 关键字表示利用现有样本数据计算各个样本因变量各种选择的响应概率。因此，建议在分析顺序选择模型时，首选该过程。

#### 2. PROBIT 过程的 ORDINAL LOGIT 建模

PROBIT 过程进行 Ordinal Logit 建模的主要语法如下：

```
PROC PROBIT <ORDER = DATA> <选项>;
  CLASS 变量;
  MODEL 因变量 = 自变量 / DIST = LOGISTIC;
  OUTPUT OUT = 输出数据集名 PROB = 变量名;
RUN;
```

其中 CLASS 语句和 MODEL 语句的功能同 QLIM 过程。在 MODEL 语句的选项中加入“DIST = LOGISTIC”关键字，表示建立的是 LOGIT 模型。

与二元选择的 LOGIT 模型一样，在实际问题分析过程中，考察 PROBIT 过程的参数估计结果时，应当将各个参数估计值取负值。

OUTPUT 语句用于输出分析结果数据集，PROB 关键字表示利用现有样本数据计算各个样本的响应概率，并指定表示该响应概率的变量名。



如果因变量有 K 种选择，则在 PROBIT 过程中由 OUTPUT 语句输出的含有响应概率的数据集中，只能得到 K-1 个响应概率，不会给出次序最大或最高的选择对应的响应概率。

### 3. LOGISTIC 过程 ORDINAL LOGIT 建模

LOGISTIC 过程进行 Ordinal Logit 建模的主要语法如下：

```
PROC LOGISTIC <ORDER = DATA> <选项>;
  CLASS 变量;
  MODEL <(EVENT = '响应值')> 因变量 = 自变量 / LINK = LOGIT;
  OUTPUT OUT = 输出数据集名 PREDICTED = 变量名;
RUN;
```

LOGISTIC 过程的 MODEL 语句中省略 LINK 关键字，表示默认建立 LOGIT 模型。

与 PROBIT 过程一样，LOGISTIC 过程在分析过程中同样把离散因变量数值最小的值作为考察响应变量的依据，但是 LOGISTIC 过程可以通过指定 DESCENDING 关键字把离散因变量数值最大的值作为考察响应变量概率的依据。此外，LOGISTIC 过程还可以通过 MODEL 语句中的 EVENT 选项考察因变量不同响应值的建模情况。

OUTPUT 语句用于输出分析结果数据集，PREDICTED 关键字表示利用现有样本数据计算各个样本的响应概率，并指定表示该响应概率的变量名。与 PROBIT 过程一样，如果因变量有 K 种选择，则在 LOGISTIC 过程中由 OUTPUT 语句输出的含有响应概率的数据集中，只能得到 K-1 个相应概率，不会给出次序最大或最高的选择对应的相应概率。

在 QLIM 过程、PROBIT 过程和 LOGISTIC 过程中的“ORDER = DATA”关键字可以省略，并且可以依据其运行程序之后的输出结果中的“Response Profile”信息，找到模型参数估计中各截距项所对应的选项信息。此处也非常让人容易混淆，建议在进行分析时，先对数据按照因变量进行排序，然后加上“ORDER = DATA”关键字，以免在分析的过程中发生错误。

同样以例 10-4 的数据为例，利用上述 3 个过程进行 Ordinal Logit 建模，具体程序如下：

```
proc format;
  value Purchase_FMT      1='1-不购买'
                          2='2-无所谓'
                          3='3-购买';
  value Gender_FMT        1='女'
```

```

                2='男';
proc sort data=Sasuser.ThreeG_Multi;
  by Purchase;
run;
proc qlim data = Sasuser.ThreeG_Multi;
  class Purchase Gender;
  model Purchase = Gender Age;
  endogenous Purchase~ discrete (order = data dist = logistic);
  format Purchase Purchase_FMT. Gender Gender_FMT.;
  output out = Qlim_OL_Out proball;
run;
proc probit order = data data = Sasuser.ThreeG_Multi;
  class Purchase Gender;
  model Purchase = Gender Age /dist = logistic;
  format Purchase Purchase_FMT. Gender Gender_FMT.;
  output out = Probit_OL_Out prob = Estimated_P;
run;
proc logistic order = data data = Sasuser.ThreeG_Multi;
  class Purchase Gender;
  model Purchase = Gender Age/link = logit;
  format Purchase Purchase_FMT. Gender Gender_FMT.;
  output out = Logit_OL_Out prob = Estimated_P;
run;
```

运行程序后，可得到 Ordinal Logit 模型的分析过程和结果。以 QLIM 过程为例进行分析，其参数估计和检验结果如图 10-16 所示。

Parameter Estimates					
Parameter		Estimate	Standard Error	t Value	Approx Pr >  t
Intercept		7.544646	1.826958	4.13	<.0001
Gender	男	2.038958	0.602183	3.39	0.0007
Gender	女	0			
Age		-0.157588	0.039041	-4.04	<.0001
Limit2		3.356135	0.568042	5.91	<.0001

图 10-16 QLIM 过程 Ordinal Logit 模型的参数估计及检验结果

图 10-16 所示的参数估计结果经过检验，均非常显著。根据图 10-15 所示的参数估计结果，可得到以下响应概率。

$$\begin{aligned}
P(y=1) &= \frac{\exp(c_1 - \alpha - x\beta)}{1 + \exp(c_1 - \alpha - x\beta)} \\
P(y=2) &= \frac{\exp(c_2 - \alpha - x\beta)}{1 + \exp(c_2 - \alpha - x\beta)} - \frac{\exp(c_1 - \alpha - x\beta)}{1 + \exp(c_1 - \alpha - x\beta)} \\
P(y=3) &= 1 - \frac{\exp(c_2 - \alpha - x\beta)}{1 + \exp(c_2 - \alpha - x\beta)}
\end{aligned}$$

其中， $\alpha = Intercept = 7.544646$ ,  $c_1 = 0$ ,  $c_2 = \_Limit2 = 3.356135$ 。

在 QLIM 过程的参数估计结果中，“Intercept”表示模型本身具有的截距项  $\alpha$ ，并且默认第一个选项的临界值或截距项  $c_1$  为 0。此外，把定性自变量转化为 0-1 变量（即 1-男、0-女，因为“女”的参数估计值为 0）的形式。

同样对性别为“男”、年龄为 45 岁的样本，把参数估计结果代入上述 Ordinal Logit 模型，得到该消费者购买 3G 手机的响应概率如下。

$$\begin{aligned}
 P(\text{不够买}) &= P(y=1) = \frac{\exp(c_1 - \alpha - x\beta)}{1 + \exp(c_1 - \alpha - x\beta)} \\
 &= \frac{\exp(0 - 7.544646 - 2.038958 \times 1 + 0.157588 \times 45)}{1 + \exp(0 - 7.544646 - 2.038958 \times 1 + 0.157588 \times 45)} \\
 &= \frac{\exp(-2.49214)}{1 + \exp(-2.49214)} \\
 &= 0.0764 \\
 P(\text{无所谓}) &= P(y=2) = \frac{\exp(c_2 - \alpha - x\beta)}{1 + \exp(c_2 - \alpha - x\beta)} - \frac{\exp(c_1 - \alpha - x\beta)}{1 + \exp(c_1 - \alpha - x\beta)} \\
 &= \frac{\exp(3.356135 - 7.544646 - 2.038958 \times 1 + 0.157588 \times 45)}{1 + \exp(3.356135 - 7.544646 - 2.038958 \times 1 + 0.157588 \times 45)} - 0.0764 \\
 &= \frac{\exp(0.863991)}{1 + \exp(0.863991)} - 0.0764 \\
 &= 0.7035 - 0.0764 \\
 &= 0.6271 \\
 P(\text{购买}) &= P(y=3) = 1 - \frac{\exp(c_2 - \alpha - x\beta)}{1 + \exp(c_2 - \alpha - x\beta)} \\
 &= 1 - 0.7035 \\
 &= 0.2965
 \end{aligned}$$

对于每个样本的各选项的响应概率，在本例输出的临时数据集 Qlim\_OL\_Out.sas7bdat、Probit\_OL\_Out.sas7bdat 和 Logit\_OL\_Out.sas7bdat 中均可看到。

#### 10.3.4 MULTINOMIAL LOGIT 模型

无序选择模型也可以细分为 Multinomial Probit 模型和 Multinomial Logit 模型。因 Multinomial Probit 模型在实际的参数估计过程（可使用 MDC 过程）中实现起来非常困难，在分析时较少采用该类模型，因此本小节主要介绍 Multinomial Logit 模型。

对于 Multinomial Logit 模型，随机误差项分布函数  $F(x)$  为 Weibull 分布，即选择选项  $j$  的响应概率为：

$$P(y=j) = \frac{e^{x_j\beta}}{\sum_{k \neq j} e^{x_k\beta}}$$

此即为条件 LOGIT 模型。

在 SAS 系统中，可用 CATMOD 过程进行无序选择 Logit 建模分析，具体语法如下。

```

PROC CATMOD <选项>;
  DIRECT 变量;
  RESPONSE LOGITS OUT = 输出数据集名;
  MODEL 因变量 = 自变量 / 选项;
RUN;

```

其中的 `DIRECT` 语句主要用于指定模型中的定量变量。而 `RESPONSE` 语句用于指定计算响应概率的函数，并可以把计算结果存储在指定的输出数据集当中，其关键字 `LOGITS` 表示使用条件 `LOGIT` 链接函数计算边际概率，该选项为系统默认，可以省略。

对于例 10-4，现假定购买意向“Purchase”的选项之间没有顺序关系，“不购买”、“无所谓”、“购买”3 种选择只反映消费者的购买态度，没有任何程度的、偏好上的差异或顺序之分，则该问题便转化为一个无序选择问题。对于无序选择问题，采用 `CATMOD` 过程进行分析，具体程序如下：

```
proc format;
  value Purchase_FMT      1='1-不购买'
                        2='2-无所谓'
                        3='3-购买';
  value Gender_FMT        1='女'
                        2='男';
proc catmod data = Sasuser.ThreeG_Multi;           /*调用 CATMOD 过程*/
  direct Age;                                       /*指定模型中的定量变量*/
  response logits out = Catmod_ML_Out;             /*指定响应概率为条件 LOGIT 函数，并输出含有各样本的各个选择对应的响应概率的数据集*/
  model Purchase = Gender Age;                     /*指定因变量和自变量之间的关系*/
  format Purchase Purchase_FMT. Gender Gender_FMT.;
quit;
run;
```

运行程序后，可得图 10-17 所示的参数估计和检验结果。

The CATMOD Procedure					
Maximum Likelihood Analysis of Variance					
Source	DF	Chi-Square	Pr > ChiSq		
Intercept	2	12.68	0.0018		
Gender	2	10.81	0.0045		
Age	2	13.32	0.0013		
Likelihood Ratio	52	43.62	0.7894		
Analysis of Maximum Likelihood Estimates					
Parameter	Function Number	Estimate	Standard Error	Chi-Square	Pr > ChiSq
Intercept	1	-12.1503	3.8106	10.17	0.0014
	2	-0.1490	2.1047	0.01	0.9436
Gender 女	1	1.7531	0.5533	10.04	0.0015
	2	0.9987	0.3904	6.54	0.0105
Age	1	0.2634	0.0781	11.37	0.0007
	2	0.0290	0.0496	0.34	0.5582

图 10-17 CATMOD 过程的 Multinomial Logit 模型参数估计和检验结果

模型分析过程与前面章节的分析过程类似，此处不再赘述。此外，对于各样本的各种选择对应的响应概率、标准误差及残差，则在输出的临时数据集 `Catmod_ML_Out.sas7bdat` 中可看到。

### 10.4 计数模型

在前面几节中，详细研究了因变量值是离散形式数据的处理方法。在现实生活中，还有一种常见的离散数据类型：当因变量表示事件发生的次数时，它是一个离散形式的整数计数变量。如一届奥运会上某个国家获得的金牌数目、消费者一周光顾商场的次数、轮船发生事

故的次数等。这些变量都是以整数计数为表现形式的，分析这种离散计数因变量的影响因素的模型即为计数模型（Count Model）。

计数模型中的离散因变量的数值是有数值含义的，即次数之间是可以进行运算的，如 2 次与 3 次之间的差距和 4 次与 5 次之间的差距从数值上看是一样的、2 次加上 5 次等于 7 次等。因此，对于只能处理无数值意义的二元选择模型或多重选择模型而言，计数数据不适用。

对于在某个时间、空间等范围之内事情发生的计数数据，一般都认为其近似服从泊松（Poisson）分布。因此，Poisson 分析方法在计数模型中应用非常广泛。

#### 10.4.1 POISSON 回归模型的基本原理

对于因变量和自变量之间的关系可以考虑建立回归模型。Poisson（泊松）回归模型即是考虑变量服从 Poisson 分布而建立一种回归模型，其假定因变量  $y$  服从参数为  $\lambda$  的泊松分布，则该模型的初始方程为：

$$\text{Prob}(y = k) = \frac{e^{-\lambda} \lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

其中， $\lambda$  表示所考察的事件在一定范围之内平均发生的次数， $k$  为整数，表示事件实际观测到的发生次数。

Poisson 回归模型的研究可以从初始方程的惟一参数  $\lambda$  入手。所考察的事件在一定范围内平均发生的次数即参数  $\lambda$  受到各种条件或影响因素  $x$  的影响，因此参数是变化的。经过对数变换，可以写出以下的 Poisson 回归模型。

$$\ln \lambda = \alpha + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \mu = \alpha + x\beta + \mu$$

对于 Poisson 回归模型，仍然可以使用极大似然法进行参数估计，其模型及参数估计值的检验问题类似于前面章节所介绍的内容。

#### 10.4.2 POISSON 回归模型的分析过程和步骤

在 SAS 系统中，主要利用 GENMOD 过程进行 Poisson 回归，具体语法如下。

```
PROC GENMOD <选项>;
  CLASS 变量;
  MODEL 因变量 = 自变量 / DIST = POISSON LINK = LOG;
  OUTPUT OUT = 输出数据集名 PREDICTED = 变量名;
RUN;
```

GENMOD 过程相关语句的作用与前面介绍过的一样，不同的是要在 MODEL 语句中指定服从泊松分布，即指定：DIST = POISSON，而且链接函数 LINK 指定为 LOG（因为在 Poisson 回归中，对  $\lambda$  做了对数变换）。

OUTPUT 语句用于输出分析结果数据集，PREDICTED 关键字表示利用现有样本数据计算各个样本的  $\lambda_i$ 。



#### 例 10-5

为监测某厂家生产的某款激光打印机的质量问题，考察该款打印机发生故障的次数。其发生故障的次数可能会受到打印纸张数量、打印机使用时长或年限、硒鼓是否为原装/组装等因素的影响。现搜集了 30 个调查所得样本数据（详见 Printer.sas7bdat），如表 10-5 所示，试对该款打印机发生故障的情况进行分析。

表 10-5 某款打印机故障次数与其影响因素

故障次数 Counts	已打印页数（千页） Pages	使用时长（千小时） Length	硒鼓类型 Cartridge
5	87.8	44.194	compatible
1	52.2	3.663	genuine
0	0.7	0.331	compatible
1	81.7	18.422	compatible
4	89.9	45.003	compatible
0	24.2	26.281	genuine
6	95.6	34.67	compatible
2	37.4	24.015	compatible
1	11.6	19.063	compatible
0	0.8	1.138	genuine
1	77.4	30.848	genuine
1	49.2	47.457	genuine
1	93.7	33.66	genuine
1	12.5	41.739	compatible
0	16.9	8.867	genuine
2	92	46.228	genuine
5	80.1	41.926	compatible
3	77.5	24.743	genuine
3	12.4	39.455	compatible
0	5.6	28.809	compatible
1	82.1	10.582	genuine
3	65	45.174	compatible
5	87.4	48.326	compatible
6	60.5	66.099	compatible
3	67.5	43.242	compatible
6	65.7	47.178	compatible
0	13.3	17.682	genuine
1	20.9	8.513	compatible
4	72.2	29.129	compatible
1	19.1	31.427	compatible

其中的硒鼓类型有两种：“genuine”表示使用原装硒鼓，“compatible”表示使用兼容硒鼓或自行添加碳粉等其他非原装耗材。

本例中，故障次数为要考察的因变量，它记录了该款打印机发生故障的次数，是一个计数变量。依据此因变量的特点，可以考虑采用 Poisson 回归建立模型，具体程序如下。

```
proc genmod data = Sasuser.Printer;
  class Cartridge;
```

```
/*调用 GENMOD 过程*/
/*指定分类或定性变量*/
```

```
model Counts = Pages Length Cartridge /dist = poisson link = log type3;/* 建立模型，使用泊松分布和
LOG 链接函数，并作影响因素的方差分析检验（type3 型检验）*/
output out = Poisson_Out predicted = Estimated; /*把输出结果存储在 Poisson_Out
临时数据集中，并把预测变量命名为 Esitimated*/
run;
```

运行程序后，可得到自变量对因变量影响的显著性检验结果和模型参数估计结果。自变量对因变量影响是否显著，可从自变量影响的显著性检验结果得到相关结论，如图 10-18 所示。

从图 10-18 中可以看到，3 个自变量的影响显著性检验的  $P$  值（“Pr>ChiSq”）都小于  $\alpha = (0.05)$ ，均非常显著。

LR Statistics For Type 3 Analysis			
Source	DF	Chi-Square	Pr > ChiSq
Pages	1	11.92	0.0006
Length	1	5.72	0.0168
Cartridge	1	8.46	0.0036

图 10-18 自变量对因变量影响的显著性检验

Poisson 回归模型的参数估计及其显著性检验结果如图 10-19 所示。

Analysis Of Parameter Estimates							
Parameter	DF	Estimate	Standard Error	Wald	95% Confidence Limits	Chi-Square	Pr > ChiSq
Intercept	1	-1.8102	0.5243	-2.8379	-0.7826	11.92	0.0006
Pages	1	0.0167	0.0051	0.0067	0.0266	10.67	0.0011
Length	1	0.0240	0.0100	0.0043	0.0437	5.72	0.0167
Cartridge compatible	1	0.9585	0.3593	0.2543	1.6627	7.12	0.0076
Cartridge genuine	0	0.0000	0.0000	0.0000	0.0000	.	.
Scale	0	1.0000	0.0000	1.0000	1.0000	.	.

图 10-19 Poisson 回归模型的参数估计及检验结果

从图 10-19 中可以看到，Poisson 回归模型的参数估计结果均非常显著。而且从图 10-19 中还可以得到以下结论：随着打印机的使用时长、已打印页数的增加，预示着打印机发生故障的次数增加；使用兼容耗材预示着比使用原装耗材的故障发生次数要多。

Poisson 回归模型建立之后，便可利用现有样本数据或新的样本数据进行预测。如现有一个样本数据（表 10-5 中的第 16 个样本），使用的是原装耗材，已打印页数达到 92 000 页，使用时长为 46 228 个小时，根据模型的参数估计结果，把已打印页数和使用时长转化为以“千”为计量单位的数值，即：

$$\begin{aligned} \ln \lambda &= -1.8102 + 0.0167 \times Pages + 0.0240 \times Length + 0.9585 \times \begin{cases} 1, Cartridge = compatible \\ 0, Cartridge = genuine \end{cases} \\ &= -1.8102 + 0.0167 \times 92 + 0.0240 \times 46.228 + 0.9585 \times 0 \\ &= 0.835672 \end{aligned}$$

上式左右两边取指数形式，得： $\lambda = \exp(0.835672) = 2.30$

即在该种情况下估计的打印机出现故障次数为 2.30 次，而样本数据中的真实故障次数为 2 次，说明建立的模型预测精度比较高。在 SAS 系统中，GENMOD 过程会自动为每个样本计算预测的因变量数值，并把预测结果存储在 OUTPUT 语句指定的输出数据集中。

**STEP 1** 对于 Poisson 回归模型，SAS 系统提供了 SAS/Insight 图形操作界面进行分析。

进入 SAS/Insight, 打开 Printer.sas7bdat 数据集, 选择系统菜单“Analyze→Fit”, 弹出 SAS/Insight 的曲线拟合对话框, 如图 10-20 所示。

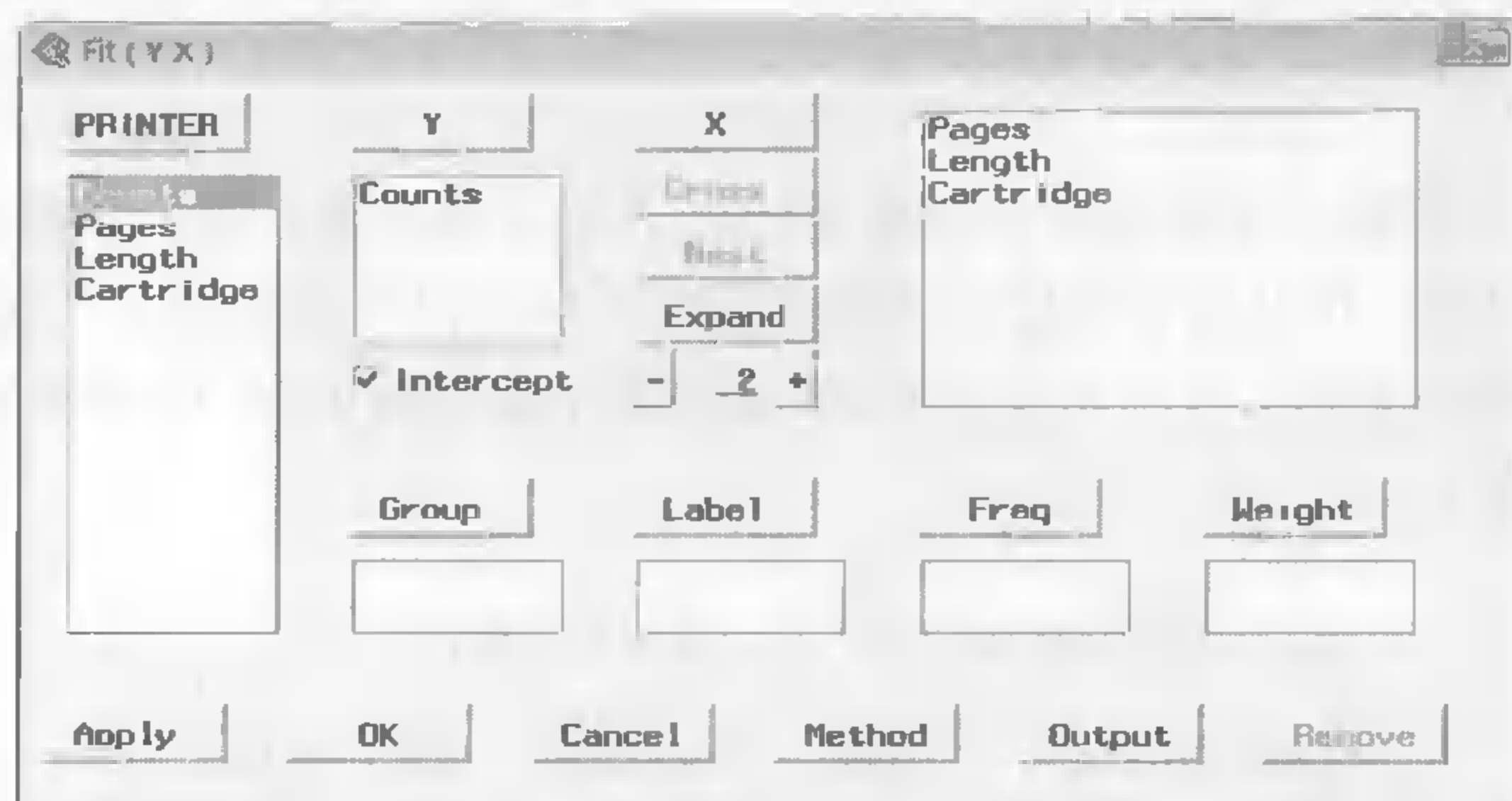


图 10-20 SAS/Insight 的“Fit”对话框

**STEP 2)** 在变量选择区域中选中“Counts”变量, 单击“Y”按钮, 把它指定为因变量; 选中“Pages”、“Length”、“Cartridge”, 单击“X”按钮, 把它们指定为自变量。

**STEP 3)** 单击“Method”按钮, 弹出模型设定、估计、检验方法对话框, 如图 10-21 所示。

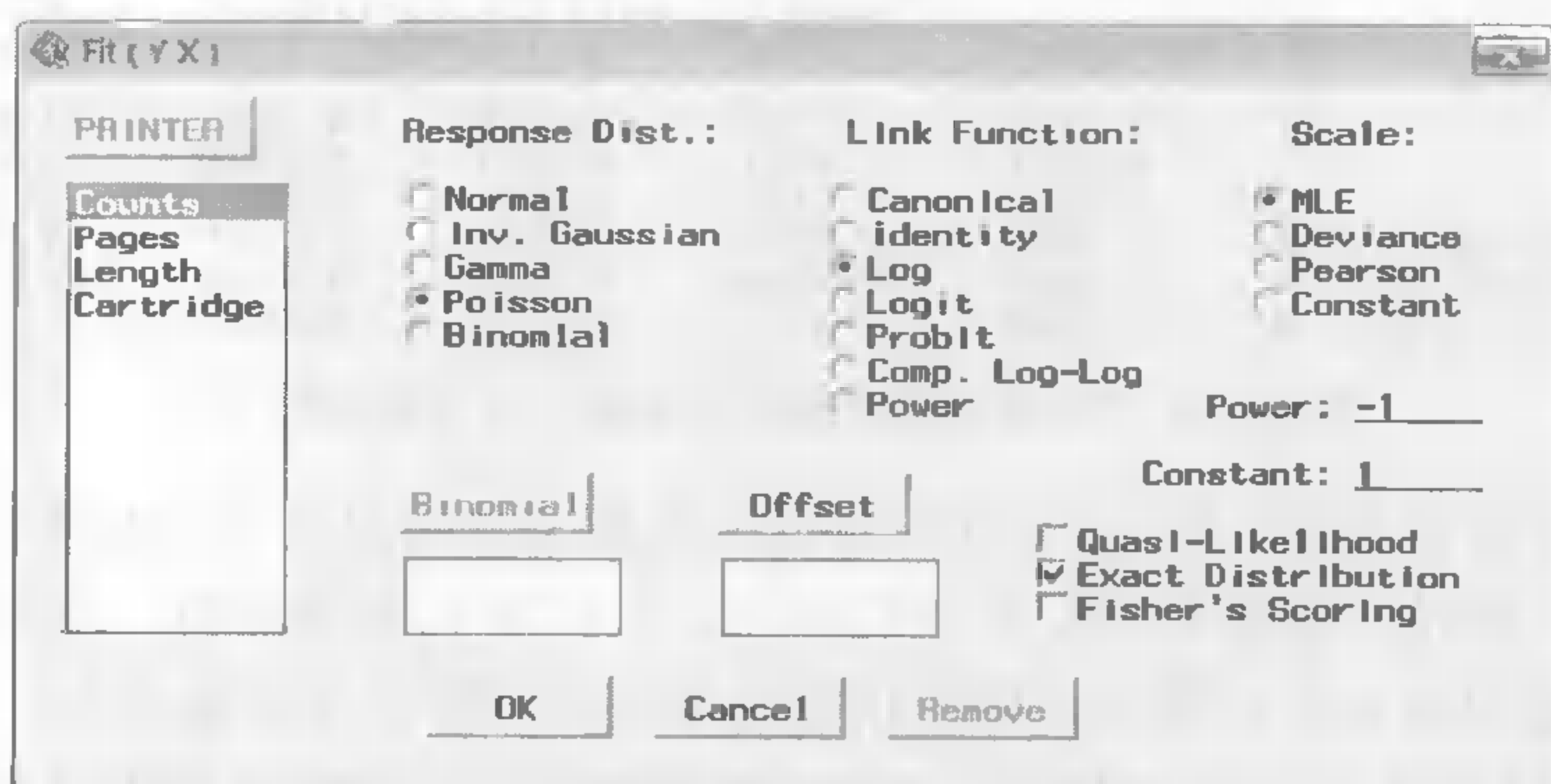


图 10-21 SAS/Insight 的“Method”对话框

**STEP 4)** 在“Method”对话框中, 可以指定要分析的模型。从图 10-21 可以看到, SAS/Insight 的“Fit”功能大体上涵盖了本章所讲述的二元选择模型的大部分内容。如要分析二元选择模型, 只需在响应分布“Response Dist”分栏下选择“Binomial”单选框, 再在链接函数“Link Function”分栏下选择对应的链接函数即可。



**注意** 选择不同的离散选择模型, 务必选择其对应的链接函数。

**STEP 5)** 本例在“Response Dist”分栏下选择“Poisson”单选框, 再在“Link Function”分栏下选择“Log”单选框, 在“Scale”分栏下选择极大似然估计“MLE”单选框, 然后单击“OK”按钮返回“Fit”对话框。

**STEP 6)** 在图 10-20 所示的“Fit”对话框中, 单击“Output”按钮还可以实现统计量、估计值及相关图形的输出, 本例从略。在“Fit”对话框中, 单击“OK”按钮, 便可得到与

GENMOD 过程运行之后类似的结果，如图 10-22 所示。

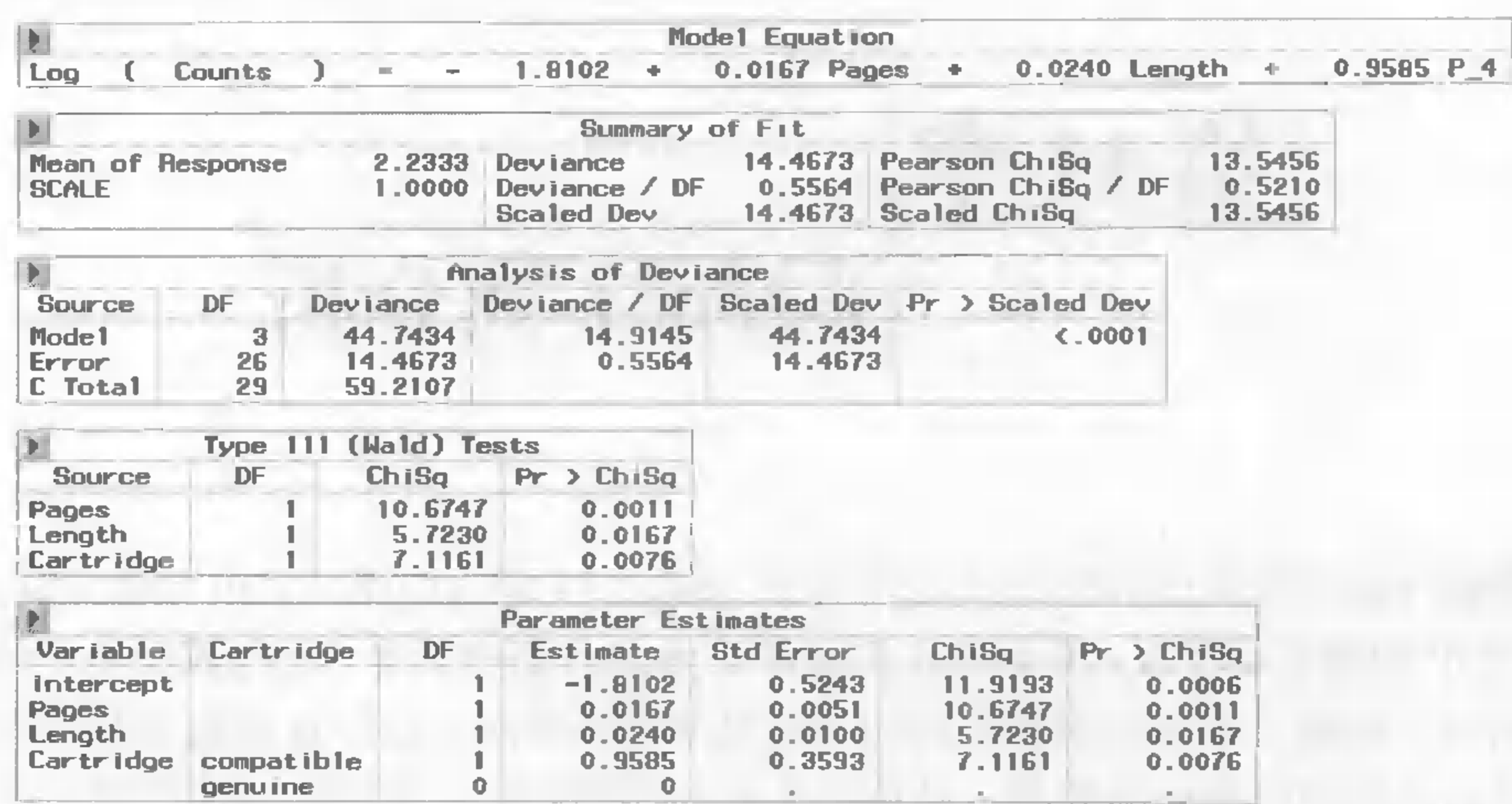


图 10-22 SAS/Insight 的 Poisson 回归分析结果

**STEP 7** SAS/Insight 会自动依据图 10-22 所示的分析模型计算每个现有样本  $\lambda$  的预测值、观测次数与预测值之间的残差，读者可在未关闭 SAS/Insight 输出窗口的情况下，切换至 SAS/Insight 的数据集界面进行查看。

10.5 本章小结

本章主要介绍了离散选择模型的基本原理及其在 SAS 系统中的实现，主要内容简要回顾如下：离散选择模型主要分析的是因变量（定性变量）受自变量的影响情况；离散选择模型依据因变量性质不同，可分为二元选择模型、多重选择模型、计数模型等；二元选择模型、多重选择模型又可以依据隐变量残差的分布不同，可分为 PROBIT 模型和 LOGIT 模型；多重选择模型按照因变量选项的含义不同，又可以分为顺序选择模型和无序选择模型；计数模型主要考察因变量是计数数据的情况下受自变量的影响，通常使用 Poisson 回归进行分析。

本章介绍的SAS过程比较多，能在不同情况下实现相同的功能，如表10-6所示：

表 10-6 离散选择模型与 SAS 实现过程

模    型		具  体  分  类	SAS 9.1.3 实现的过程
二元选择模型（Binary Choice）		Binary Probit	QLIM、PROBIT、LOGISTIC、GENMOD
		Binary Logit	QLIM、PROBIT、LOGISTIC、GENMOD、CATMOD
多重选择模型 （Multiple Choice）	Ordinal Choice	Ordinal Probit	QLIM、PROBIT、LOGISTIC
		Ordinal Logit	QLIM、PROBIT、LOGISTIC
	Multinomial Choice	Multinomial Probit	MDC
		Multinomial Logit	CATMOD
计数模型（Count Model）		Poisson Regression	GENMOD

## 第 11 章

# 时间序列分析

历史数据往往以时间序列的形式呈现出来,如过去 3 年内中国的 CPI 指数走势、最近 100 年来美国 GDP 的增长率情况、某超市在过去 1 年内的月销售额等。这些数据都是随着时间的变化而变化的,反映了事物、现象在时间上的发展变动情况。由于这些数据是相同事物或现象在不同时刻或时期所形成的数据,故称之为时间序列数据,简称时间序列或时序数据。

前面章节所研究的大部分数据都是反映若干事物或现象在同一时刻或时间上所处的状态或特征,或者反映其与时间无关的特征。这些数据反映了事物或现象之间存在的内在数值联系,故称之为横截面数据。

有关横截面数据的研究,大多采用前面章节所介绍过的方法。本章将主要讨论时间序列数据的分析,其主要研究目的是总结过去并预测未来。

### 11.1 时间序列的基本问题

时间序列是将某一个变量或指标在不同时间上的不同数值按照时间的先后顺序排列而成的数列,也被称为时间数列,通常可用  $X_1, X_2, \dots, X_t$  来表示。数据排列的依据可以是年份、季度、月份、天、小时、分钟、秒等表示时间的计量单位。

#### 11.1.1 时间序列的组成部分

由于受到各种偶然因素的影响,时间数列往往表现出某种随机性,彼此之间存在着统计上的依赖关系。



例 11-1

某大型商场为研究其销售总额的情况,现搜集了从 2001 年 1 月至 2008 年 8 月的销售额月度数据(详见 Sales\_Monthly.sas7bdat)进行时间序列分析,如表 11-1 所示。

表 11-1 某商场 2001 年 1 月至 2008 年 8 月的销售额月度数据(单位:万元人民币)

月份 Month	销售额 Sales	月份 Month	销售额 Sales	月份 Month	销售额 Sales	月份 Month	销售额 Sales
2001Jan	814.0	2002Dec	784.6	2004Nov	859.6	2006Oct	890.7
2001Feb	774.8	2003Jan	743.2	2004Dec	824.8	2006Nov	761.2
2001Mar	782.8	2003Feb	689.2	2005Jan	856.0	2006Dec	941.2
2001Apr	772.0	2003Mar	637.5	2005Feb	746.8	2007Jan	967.5
2001May	817.6	2003Apr	692.8	2005Mar	710.8	2007Feb	979.5

续表

月份 Month	销售额 Sales	月份 Month	销售额 Sales	月份 Month	销售额 Sales	月份 Month	销售额 Sales
2001Jun	779.2	2003May	708.4	2005Apr	709.4	2007Mar	876.6
2001Jul	715.6	2003Jun	757.6	2005May	803.2	2007Apr	824.8
2001Aug	637.6	2003Jul	804.4	2005Jun	733.5	2007May	890.9
2001Sep	793.6	2003Aug	786.7	2005Jul	636.4	2007Jun	862.0
2001Oct	878.8	2003Sep	799.5	2005Aug	751.5	2007Jul	864.4
2001Nov	670.0	2003Oct	803.2	2005Sep	833.2	2007Aug	829.5
2001Dec	820.0	2003Nov	778.0	2005Oct	824.8	2007Sep	755.2
2002Jan	889.7	2003Dec	727.6	2005Nov	857.2	2007Oct	875.2
2002Feb	728.8	2004Jan	805.6	2005Dec	862.0	2007Nov	888.3
2002Mar	767.2	2004Feb	809.2	2006Jan	786.4	2007Dec	1,025.2
2002Apr	807.2	2004Mar	770.1	2006Feb	781.5	2008Jan	991.6
2002May	792.4	2004Apr	761.9	2006Mar	739.5	2008Feb	882.3
2002Jun	782.8	2004May	724.0	2006Apr	798.4	2008Mar	902.8
2002Jul	752.8	2004Jun	786.4	2006May	852.3	2008Apr	926.4
2002Aug	739.6	2004Jul	641.2	2006Jun	810.2	2008May	983.5
2002Sep	710.8	2004Aug	610.0	2006Jul	815.2	2008Jun	942.4
2002Oct	854.8	2004Sep	749.2	2006Aug	820.4	2008Jul	989.7
2002Nov	775.6	2004Oct	863.2	2006Sep	810.3	2008Aug	1002.8

该数据集中一共有 92 个按月份先后顺序排列的观测值。为了更好地观测 GDP 的走势状况，把表 11-1 中的数据用趋势图形来表示，如图 11-1 所示。

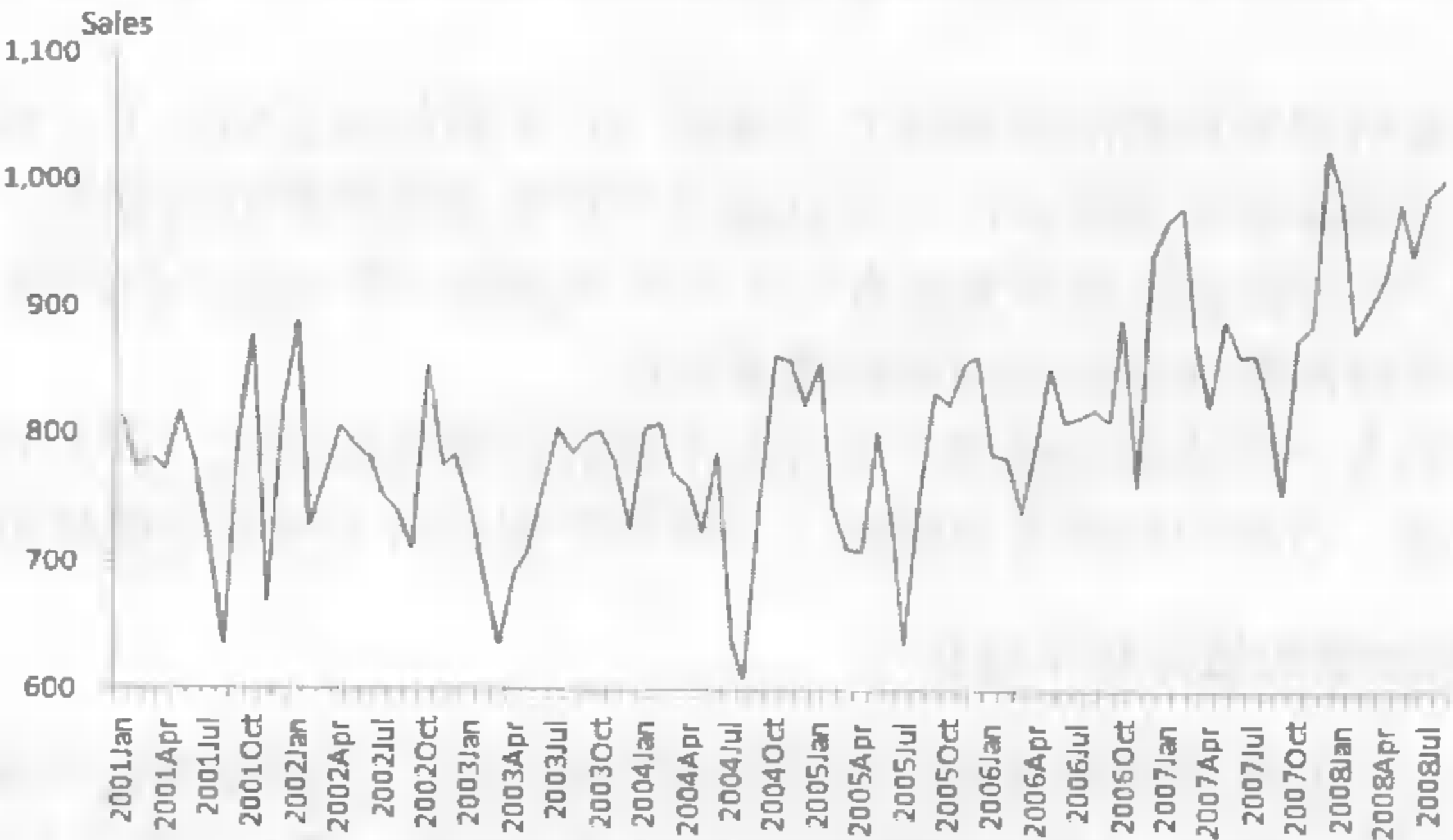


图 11-1 销售额月度数据的时间序列趋势图

从图 11-1 可以看出，销售额月度数据总体上呈直线上升的趋势，但是在上升过程中还有上下波动的情况。

一个时间序列可以由 4 个部分构成，即长期趋势、季节变动、循环波动和不规则变动。长期趋势是指事物或现象在较长时期内持续发展变化的一种趋向或状态，如例 11-1 的数

据具有上升的长期趋势。

季节变动是指事物或现象在一年内随着季节更换形成的有规律变动，如空调的销售量随着季节不同而发生较大的变动，夏季的销量一般都高于冬季的销量。例 11-1 中的数据也表现有季节变动的趋势，如每逢 5、10 月黄金周和年末、节前，销售额均有上升，而在淡季销售额略有下降，但是这种变动并不是很明显。

循环波动是指事物或现象周而复始的变动。循环波动不同于长期趋势和季节变动，它是无固定规律的交替波动，如经济发展过程中有经济周期、金融危机周期等。

不规则变动则是无法用上述组成部分解释或不可控的随机变动。

为了更加深入地研究时间序列的规律，往往可将一个时间序列数据分解为上述的 4 个组成部分。

### 11.1.2 时间序列的平稳性

按照不同的性质和特征，可以对时间序列进行分类。从统计特性上来看，时间序列可以分为平稳时间序列和非平稳时间序列。

#### 1. 平稳性的含义

如果一个时间序列的概率分布与时间  $t$  无关，则称该序列是严格的平稳时间序列。如果时间序列的一、二阶矩存在，而且对任意时刻  $t$  满足均值为常数、协方差为时间间隔的函数，则称该序列为宽平稳时间序列，也叫广义平稳时间序列。反之，不具有平稳性即序列均值或协方差与时间有关的序列被称为非平稳序列，其主要特征表现为在整体上或局部上有明显的上升或下降的趋势。如例 11-1 中的数据，销售额数据与时间有着密切相关的联系，即销售额数值随着时间的推进而不断上升，因此该序列是非平稳的。

严格的平稳时间序列要求比较严格。在通常情况下，如果不明确提出严格平稳，所谓的平稳即指宽平稳，其特性即均值和协方差不随时间变化而变化。本章后续部分将主要研究宽平稳时间序列。

那么为什么要研究平稳时间序列呢？这是因为在平稳的保证情况下，对历史时序数据进行分析的参数估计结果也比较稳定，可以直接用于对未来时序数据的预测。此外，非平稳时间序列在分析时，还可能会出现本来没有什么关系的变量之间出现“伪回归”的情况。因此，平稳性是合理进行时间序列分析和预测的重要保证。

平稳时间序列有一种特殊的情况，即分布不随时间变化而变化，其具有零均值和同方差性，且协方差为零。该序列被称为白噪声。白噪声序列可用于对时序模型拟合进行检验。

#### 2. 时间序列的零均值化和平稳化

在日常生活中，社会经济现象的特征随着时间的推移，大部分都会表现为上升或下降趋势的非平稳时间序列。因此，可以考虑对时间序列进行变换，使非平稳序列转化为平稳序列。为了能够使用 11.2 节中的 Box-Jenkins 法进行 ARIMA 时间序列分析建模，通常将非平稳序列进行零均值化和平稳化，将其转化为零均值平稳时间序列。

零均值化是指对均值不为零的时间序列进行转化，使其均值为零的数据转换过程。通常可用时间序列中的每一个数值  $X_t$  减去该序列的平均值，即  $X_t - \bar{X}$ ，得到的新数列的均值为零。

平稳化是指对非平稳的时间序列进行转化，使之成为平稳时间序列的数据转换过程。通

常可以用每一个数值减去其前面的一个数值，即  $X_t - X_{t-1}$  差分的方法。差分的方法还可以是每一个数值减去其前面的、任何间隔为  $s$  的一个数值，即  $X_t - X_{t-s}$ 。

对原始数据进行一次差分的过程被称为一阶差分。如果数列经过一阶差分之后还是非平稳数列，则可进行二阶差分或高阶差分，即对差分之后的数据再进行差分。在一般情况下，非平稳的时间序列在经过一阶差分或二阶差分之后都可以平稳化。

在有些情况下，还可以通过函数的形式进行零均值化和平稳化，如对时间序列的数值取对数后再进行差分。具体使用什么形式的函数要视具体分析的问题而定。

如对于例 11-1，原始数据呈直线上升趋势，可以进行一阶差分，然后绘制时序图，如图 11-2 所示。

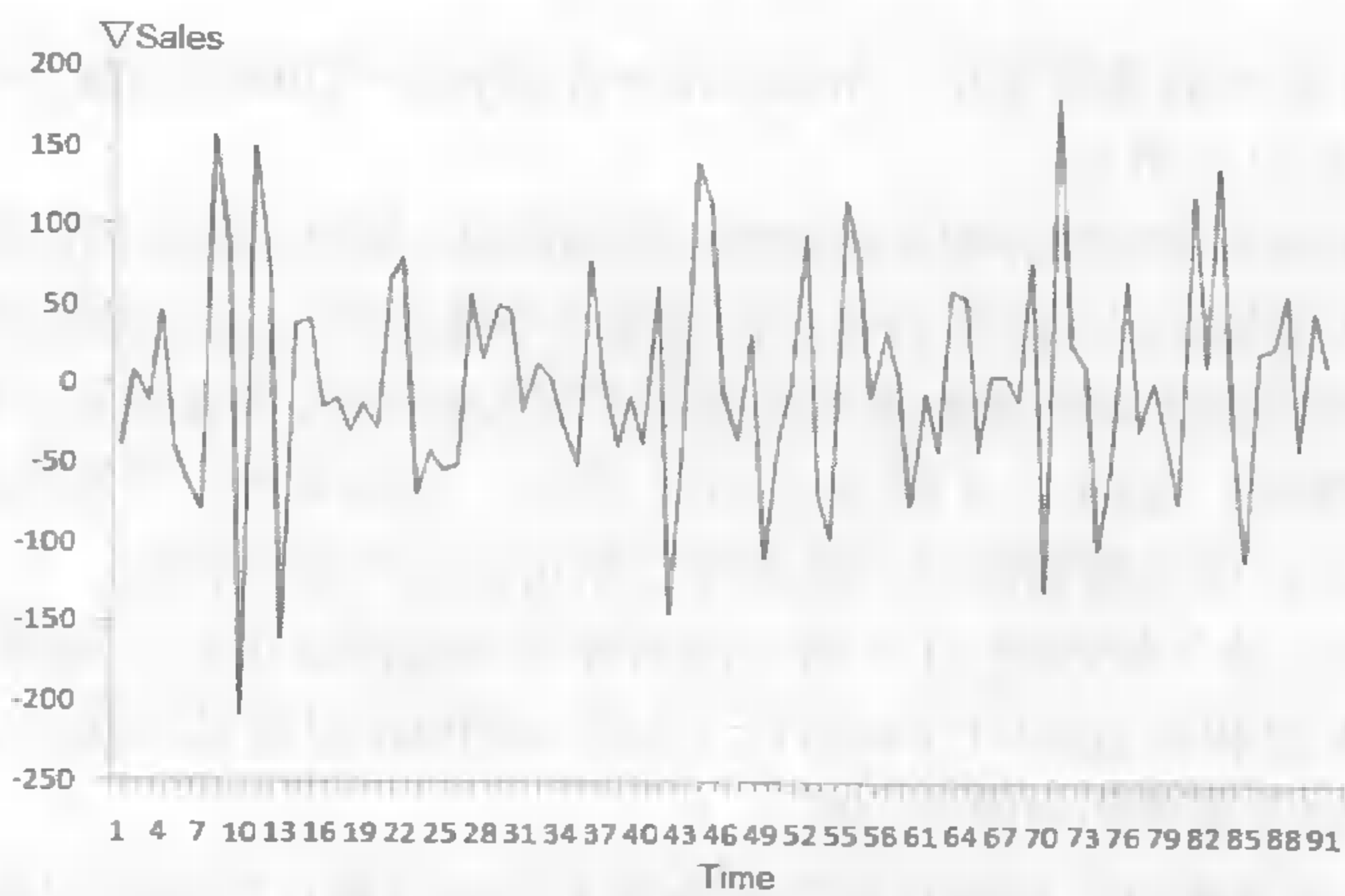


图 11-2 销售额月度数据的一阶差分时间序列图

从图 11-2 中可以看到，一阶差分之后的销售额月度数据在 0 值上下波动，而且已无明显的趋势，因此可以认为它是一个零均值化的平稳序列。

3. 时间序列的平稳性检验

除了利用类似图 11-1 和图 11-2 所示的时序图进行时间序列平稳性的粗略判断之外，还可以利用样本自相关函数及其图形进行进一步判断。

(1) 自相关系数和自相关图检验。

与相关系数类似，自相关系数实际上是构成时序的各个组成元素的相关系数，即通过考察历史数据和未来数据的相关性，可以得知不同时期数据之间的相关程度。其取值范围在 -1 到 1 之间，其绝对值越接近于 1，说明时间序列的自相关程度越高。

对于一个时序数据总体而言，在给定的正整数  $p$  情况下，可以考察  $X_t$  和  $X_{t+p}$  之间的相关系数  $\rho_k$  来度量时间间隔为  $p$  的两部分数据之间的相关性。因此，依据样本数据，可以定义时间序列的  $p$  阶样本自相关系数或自相关函数（ACF，Auto-Correlation Function）如下。

$$r_p = \frac{\sum_{t=1}^{n-p} (X_t - \bar{X})(X_{t+p} - \bar{X})}{\sum_{t=1}^n (X_t - \bar{X})^2}, p = 1, 2, 3, \dots$$

根据各给定的  $p$  计算的自相关系数，可以用自相关系数图来描述。使用例 11-1 中的数据

进行一阶差分之后，可得自相关系数图，如图 11-3 所示。

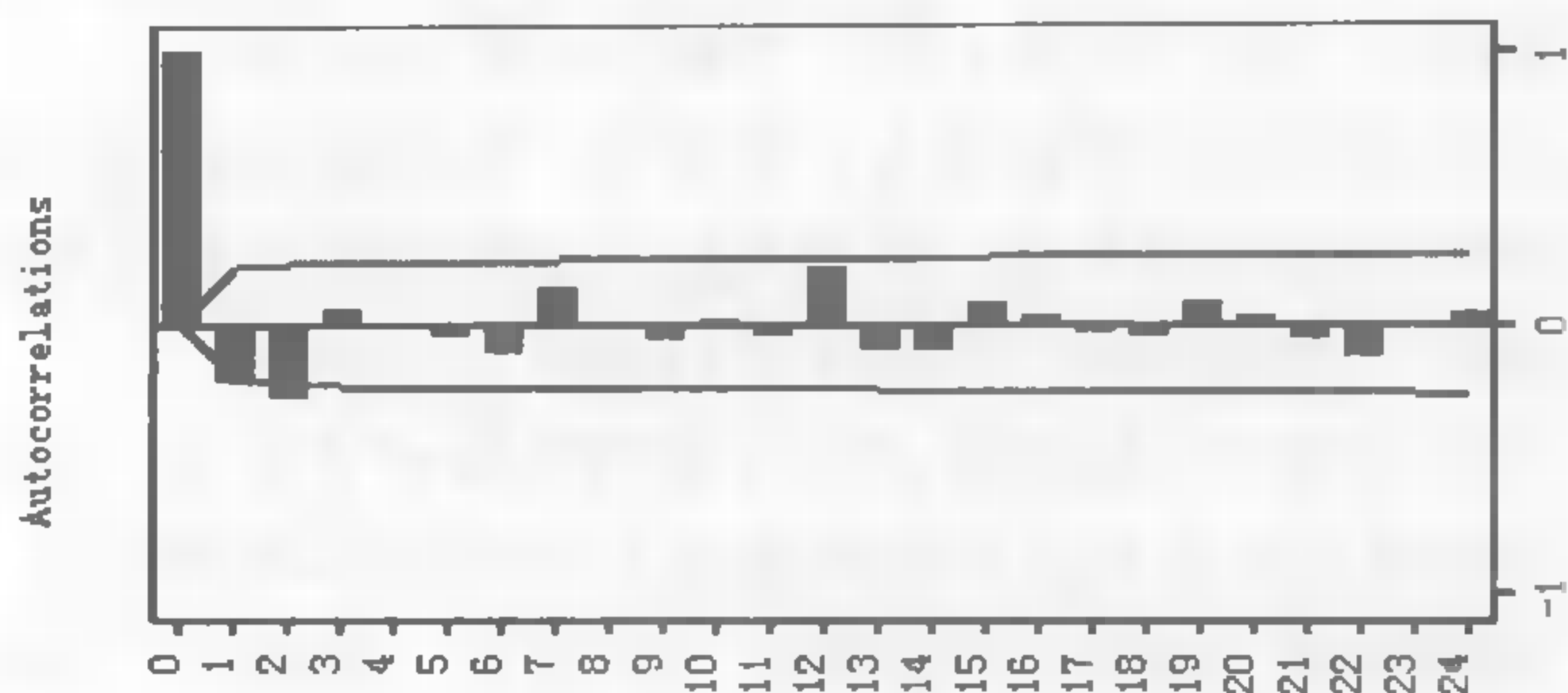


图 11-3 销售额月度数据一阶差分的自相关系数（ACF）图

图 11-3 由 SAS 系统的系统菜单 “Solutions→Analysis→Time Series Forecasting System” 功能模块绘制（详见 11.2 节）。

运用自相关分析图判定时间序列平稳性的一般准则是：若时间序列的自相关系数基本上（通常为  $p>3$  时）都落入置信区间（即图 11-3 中的两条水平线之间），且逐渐趋于零，则该时间序列具有平稳性；若时间序列的自相关系数更多地落在置信区间外面，则该时间序列就不具有平稳性。

依据这个一般准则，在图 11-3 所示的 ACF 图中，在  $p>3$  时，所有的自相关系数均落入了置信区间范围之内，即该数据经过一阶差分之后可以认为是平稳的。

在 SAS 系统中，除了绘制图 11-3 所示的自相关系数图以进行主观的平稳性检验之外，还可以利用 ARIMA 过程中 IDENTIFY 语句（有关 ARIMA 过程的语法，详见 11.2 节）进行自相关系数图及自相关系数的白噪声检验。

如对于例 11-1 中的数据，绘制自相关系数图和进行白噪声检验的程序如下。

```
proc arima data=Sasuser.Sales_Monthly;      /*调用 ARIMA 过程*/
  identify var=Sales;                        /*检验 Sales 序列是否平稳*/
  identify var=Sales(1);                    /*检验 Sales 序列的一阶差分是否平稳*/
run;
```

运行程序后，首先可得到序列的自相关系数及自相关系数图。Sales 序列进行一阶差分之后的检验结果如图 11-4 所示。

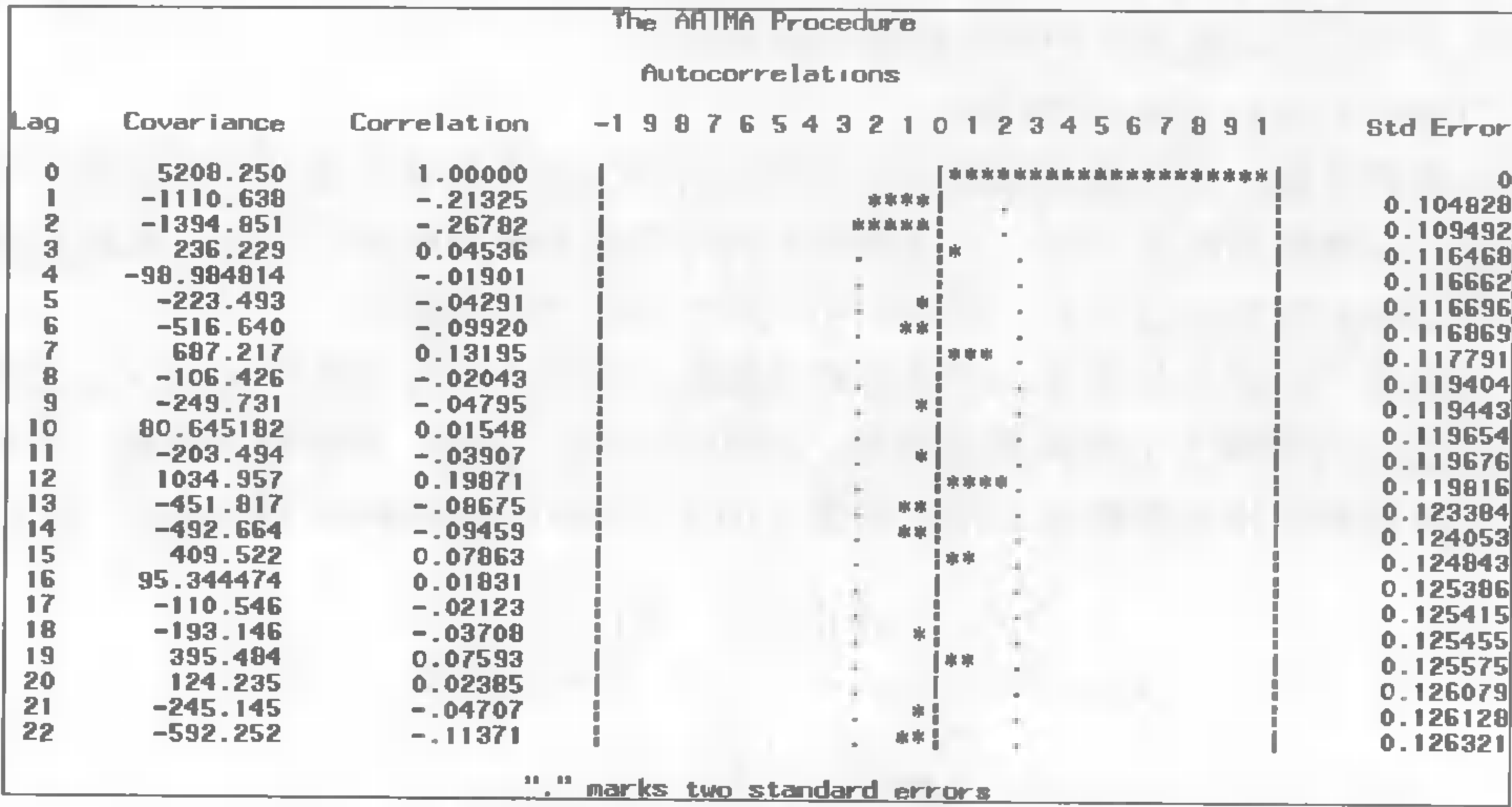


图 11-4 销售额月度数据一阶差分自相关系数及自相关系数图

图 11-4 与图 11-3 类似，只是多了协方差、自相关系数及标准误差的统计量。依据图 11-4 所示，仍然可以判定销售额月度数据进行一阶差分是平稳的。

此外，上述程序还输出了对于自相关系数的白噪声检验，如图 11-5 所示。

Autocorrelation Check for White Noise									
To Lag	Chi- Square	DF	Pr > ChiSq	-----Autocorrelations-----					
6	12.49	6	0.0519	-0.213	-0.268	0.045	-0.019	-0.043	-0.099
12	18.94	12	0.0900	0.132	-0.020	-0.048	0.015	-0.039	0.199
18	21.68	18	0.2466	-0.087	-0.095	0.079	0.018	-0.021	-0.037

图 11-5 销售额月度数据一阶差分自相关系数的白噪声检验结果

如果白噪声检验结果显著，则表明时间序列总体自相关是显著的，即表现为非平稳。当所有的白噪声检验结果均不显著时，则时间序列是平稳的。

本例数据在图 11-5 所示的结果中，在 $\alpha=0.05$ 的条件下，白噪声检验的  $p$  值(“Pr>ChiSq”)均大于  $\alpha$ ，表明白噪声检验不显著，所以销售额月度数据经过一阶差分之后是平稳的。

(2) 单位根检验。

仅从图形描述来对时间序列平稳性进行判断的准确性毕竟有限，一般还可考虑使用单位根检验的方法对时序数据的平稳性进行检验。

一个时间序列如能通过差分的方式平稳化，则可称其具有单位根，即当一个时间序列具有单位根时是非平稳的。在 SAS 系统中，提供了“%DFTEST”宏进行 Dickey-Fuller 单位根检验（即 DF 检验），其原假设和被择假设如下。

$H_0$ : 时间序列具有单位根       $H_1$ : 时间序列是平稳序列（即没有单位根）

对于例 11-1 中的销售额月度数据，可利用“%DFTEST”宏进行单位根检验，程序如下。

```
%dfest(Sasuser.Sales_Monthly,Sales);          /*对原始数据 Sales 的变量进行单位根检验*/
%put P(Sales)=&dfest;                          /*在“LOG”窗口中输出检验 P 值*/
%dfest(Sasuser.Sales_Monthly,Sales,dif=(1));    /*关键字“dif=(1)”表示对 Sales 的一阶差分进行检验*/
%put P(Sales(1))=&dfest;                        /*在“LOG”窗口中输出检验 P 值*/
```

运行程序后，可以在“LOG”窗口中分别得到原始数据和一阶差分数据的单位根检验的  $P$  值，即  $P(\text{Sales})=0.3159464781$ 、 $P(\text{Sales}(1))=0.0000946955$ 。由此可以看出，Sales 变量的单位根检验并不显著，即  $P(\text{Sales})$  值非常大，没有充分的理由可以拒绝原假设，即 Sales 具有单位根，是非平稳的序列；而 Sales 一阶差分的单位根检验  $P$  值非常显著，故可以拒绝原假设，认为 Sales 一阶差分的序列是平稳序列。

11.2 ARIMA 模型的分析过程

时间序列分析的 ARIMA 建模法也叫做 Box-Jenkins 法，它是一种以美国统计学家 George E. P. Box 和英国统计学家 Gwilym M. Jenkins 的名字命名的时间序列预测方法。它主要是在对时间序列进行分析的基础上，选择适当的模型进行预测。

11.2.1 ARIMA 模型

ARIMA 模型也被叫做整合自回归移动平均模型（Auto-Regressive Integrated Moving Average），可分为自回归模型（AR 模型）、移动平均模型（MA 模型）和自回归移动平均模型（ARMA 模型）。

Box-Jenkins 法的基本思想是用时间序列的过去值和现在值的线性组合来预测其未来值。也就是说，将随时间推移而形成的系列数据视为一个随机序列，把时间序列作为一组仅依赖于时间  $t$  的随机变量，这组随机变量所具有的依存关系或自相关性表现了其所观测对象发展的延续性。而这种自相关性一旦被相应的数学模型描述出来，就可以从时间序列的过去值及现在值去预测其未来值。

### 1. AR 模型

AR 模型即自回归模型 (Auto-Regression Model)，其具体表现为某个观测值  $X_t$  与其滞后  $p$  期的观测值的线性组合再加上随机误差项，即：

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + a_t$$

其中  $X_t$  为零均值平稳序列， $a_t$  为随机误差项。为了方便模型的描述，通常把上述模型简记为 AR(p)。

对于 AR(p) 模型而言，有其基本假设：假设  $X_t$  仅与  $X_{t-1}, X_{t-2}, \cdots, X_{t-p}$  有线性关系；在  $X_{t-1}, X_{t-2}, \cdots, X_{t-p}$  已知的条件下， $X_t$  与  $X_{t-p-1}, X_{t-p-2}, \cdots$  无关；且  $a_t$  是一个白噪声。

### 2. MA 模型

MA 模型即移动平均模型 (Average Moving Model)，其具体表现为某个观测值  $X_t$  与先前  $t-1, t-2, \cdots, t-q$  个时刻进入系统的  $q$  个随机误差项即  $a_t, a_{t-1}, \cdots, a_{t-q}$  的线性组合，即：

$$X_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \cdots - \theta_q a_{t-q}$$

通常把上述模型简记为 MA(q)。

对于 MA(q) 而言， $X_t$  仅与  $a_t, a_{t-1}, \cdots, a_{t-q}$  有关，而与  $a_{t-q-1}, a_{t-q-2}, \cdots$  无关，且  $a_t$  是一个白噪声序列。

### 3. ARMA 模型

ARMA 模型即自回归移动平均模型 (Auto-Regressive Moving Average Model)，即观测值  $X_t$  不仅与其以前  $p$  个时刻的自身观测值有关，而且还与其以前时刻进入系统的  $q$  个随机误差存在一定的依存关系，即：

$$X_t = \varphi_1 X_{t-1} + \varphi_2 X_{t-2} + \cdots + \varphi_p X_{t-p} + a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2} \cdots - \theta_q a_{t-q}$$

显然 ARMA(p,q) 模型便是 AR(p) 与 MA(q) 的组合，ARMA(p,0) 便是 AR(p) 模型，ARMA(0,q) 便是 MA(q) 模型。

在进行 ARMA 建模之前，分析的时间序列必须满足平稳性条件。非平稳的时间序列数据则可以按照 11.1.2 小节中介绍的差分方法使之平稳化，并进行平稳性检验。时间序列经过差分平稳化之后，便可建立 ARMA 模型进行分析，待模型进行参数估计之后，再通过数据变换的可逆性，使得模型参数估计结果适应平稳化之前的数据。通过整个过程建立的模型被称为整合的 ARMA 模型，即 ARIMA 模型。如果对原始数据进行了  $d$  次差分，则用差分数据所建立的 ARMA(p,q) 可以记为 ARMA(p,d,q)。

#### 11.2.2 ARMA 模型的识别、估计与预测

建立 ARMA 模型的基本前提是保证时间序列的平稳性，ARIMA 建模过程则是把非平稳

时间序列平稳化,再建立 ARMA 模型。ARMA 的基本形式在 11.2.1 小节中已经详细介绍过,模型中的两个参数  $p$  和  $q$  一旦确定下来,那么 ARMA 模型便可以确定。因此,首先要做的分析工作便是确定  $p$  和  $q$  的具体取值,然后再对 ARMA( $p,q$ )模型进行参数估计及显著性检验,最后利用显著的模型对时间序列进行预测。

对于这一整套模型识别、参数估计与模型预测的建模过程,SAS 系统提供了 ARIMA 过程以进行分析。ARIMA 过程的主要语法如下。

```
PROC ARIMA 选项;
  BY 变量;
  IDENTIFY VAR=变量 选项;
  ESTIMATE 选项;
  OUTLIER 选项;
  FORECAST 选项;
RUN;
```

ARIMA 过程各语句的选项非常多,本小节只介绍最为常用的功能。

BY 语句使用方法与前面章节介绍的功能一致。

IDENTIFY 语句主要用于模型识别,即确定参数  $p$  和  $q$  的步骤,也可以用于考察时间序列的平稳性检验等。编程过程中,该语句可多次使用,其主要选项如下。

- DATA=: 指定用于分析的数据集。如省略,则表示默认使用 ARIMA 过程指定的数据集;
- VAR=: 指定用于分析的时间序列变量。在变量后面加上(n),表示进行该变量的  $n$  阶差分。
- CENTER: 对数据进行零均值化,主要用于分析非零均值的数据。当模型进行参数估计并生成预测结果之后,系统自动会在预测值中加上均值。使用该语句要注意 SAS 的数据处理顺序,即先进行差分,再进行零均值化。
- CLEAR: 清除内存中的已有模型。
- ESACF: 计算扩展的样本自相关函数并使用其估计值进行模型参数  $p$  和  $q$  的识别。
- MINIC: 使用最小信息准则进行模型参数  $p$  和  $q$  的识别。系统默认  $p$  和  $q$  分别从 0 至 5 进行 BIC 指数的识别。
- NLAG: 指定计算自相关系数和互相关系数的滞后期数(即 11.1.2 小节中计算自相关系数的  $P$  值)。
- OUTCOV=: 指定存储自相关系数、偏自相关系数等统计量的数据集。
- P=( $p_{min}:p_{max}$ ): 指定 ARMA 模型中参数  $p$  的最小值和最大值,通常与 MINIC、SCAN 选项搭配使用。
- Q=( $q_{min}:q_{max}$ ): 指定 ARMA 模型中参数  $q$  的最小值和最大值,通常与 MINIC、SCAN 选项搭配使用。
- SCAN: 计算典型相关系数平方的估计值,并用来确定 ARMA 模型参数  $p$  和  $q$  的值。
- NOPRINT: 不输出任何结果。
- STATIONARITY=: 进行时间序列的平稳性检验,用于指定检验方法的关键字有 ADF 或 DICKEY、PP 或 PHILLIPS、RW 或 RANDOMWALK。

ESTIMATE 语句主要用模型的参数估计和指定变量的转换函数,其主要选项如下。

- INPUT=: 指定输入变量及其对应的转换函数。
- METHOD=: 指定模型参数估计的方法, 其中可指定的估计方法关键字为 ML、ULS 和 CLS, 分别表示极大似然法、非条件最小二乘法和条件最小二乘法。CLS 方法是系统默认的方法;
- P=: 指定模型参数  $p$  的值。
- Q=: 指定模型参数  $q$  的值。
- PLOT: 绘制模型残差项的自相关系数图。
- OUTEST=: 指定存储参数估计结果的输出数据集。
- OUTMODEL=: 指定存储模型及模型参数估计结果的输出数据集。
- OUTSTAT=: 指定存储用于模型诊断的统计量的输出数据集。

FORECAST 语句主要用于利用模型参数估计结果对时间序列数据进行预测, 其主要选项如下。

- BACK=: 指定时间序列从最后一个观测值起往前预测的时期。如 “BACK=5” 表示预测最后一个观测值之前 5 期的数值。
- LEAD=: 指定时间序列从最后一个观测值起向后预测的时期。如 “LEAD=5” 表示预测最后一个观测值之后 5 期的数值。
- ID: 指定表示时间序列的日期变量。此语句所指定的变量必须为 SAS 数据类型的日期型。
- INTERVAL=: 指定描述时间序列的时间间隔。可以指定的时间间隔有 YEAR、SEMIYEAR、QTR、MONTH、SEMIMONTH、TENDAY、WEEK、WEEKDAY、DAY、HOUR、MINUTE 和 SECOND, 分别表示年、半年、季度、月份、半月、10 天、周、周天、天、小时、分钟和秒。
- ALPHA=: 指定预测值进行区间估计的显著性水平。
- OUT=: 指定存储预测值等变量的输出数据集。如 FORECAST 语句中没有指定 OUT 输出数据集, 则系统自动把结果存储至 ARIMA 过程的 OUT 选项指定的数据集中。

在 SAS 系统中, ARIMA 过程不同于前面章节中介绍的其他过程。一旦调用 ARIMA 过程, 只要中途分析不用 Quit 语句中断该过程, 就可以一直使用该过程的语句, 而无须重新调用 ARIMA 过程。在通常情况下, 一般不能一步或一次调用 ARIMA 过程对模型进行建模, 而是要通过 ARIMA 过程中提供的各种语句分别对模型进行识别、参数估计、预测等调试。

本书将以例 11-1 为例, 按照以下步骤进行时间序列分析。

### 1. 模型的识别

ARMA 模型的识别主要是针对确定其两个参数  $p$  和  $q$  的具体数值而言的。确定  $p$  和  $q$  具体数值的过程即为模型的识别过程, 也被叫做 ARMA 模型的定阶。如 AR(2)被称为 2 阶 AR 模型, MA(3)被称为 3 阶 MA 模型。

模型的识别是针对平稳数据而言的, 例 11-1 中的数据经过一阶差分之后的数列满足平稳性条件 (详见 11.1.2 小节的分析过程)。

(1) 利用自相关系数、偏自相关系数图进行模型识别。

ARMA 模型的识别可以通过自相关系数和偏自相关系数对应的相关系数图形来进行。自

相关系数在 11.1.2 小节中已经介绍过，它描述的是时间序列观测值与过去值之间的相关性。而偏自相相关系数 (PACF, Partial Autocorrelation Function) 则为在给定中间观测值的条件下，观测值与前面某个间隔的观测值之间的相关系数。偏自相关系数的推导过程较为复杂，其实质是使得残差的方差达到最小的  $k$  阶 AR 模型的第  $k$  项系数。

利用相关系数图进行模型识别，首先应当搞清楚两个基本概念，即截尾和拖尾。

所谓截尾，是指在自相关系数图或偏自相关系数图中，自相关系数或偏自相关系数在滞后的前几期处于置信区间之外，而滞后的系数基本上都落入置信区间内，且逐渐趋于零。如在图 11-3 所示的自相关系数图中，只有滞后前两期的自相关系数处于置信区间之外，其余系数均处于置信区间之内，因此可以称图 11-3 所示情况为截尾。通常把相关系数图在滞后第  $p$  期后截尾的情况叫做  $p$  阶截尾，如图 11-3 又可以被称为 2 阶截尾。

所谓拖尾，是指在自相关系数图或偏自相关系数图中的系数有指数型、正弦型或震荡型衰减的波动，并不会都落入置信区间内。

利用自相关系数图和偏自相关系数图进行模型识别，主要依据以下原则，如表 11-2 所示。

表 11-2 ACF 图和 PACF 图的模型识别

自相关系数图 (ACF 图)	偏自相关系数图 (PACF 图)	模型识别结果
$q$ 阶截尾	拖尾	MA( $q$ )
拖尾	$p$ 阶截尾	AR( $p$ )
拖尾	拖尾	ARMA

在 ACF 图和 PACF 图都拖尾的情况下，ARMA 模型中的  $p$ 、 $q$  参数还需进一步进行确定。

ARMA( $p$ ,  $q$ ) 的偏自相关系数可能在  $p$  阶滞后项前有几项明显高出置信区间，但从  $p$  阶滞后项开始逐渐趋向于零；而其自相关系数则可能在  $q$  阶滞后项前有几项明显高出置信区间，但从  $q$  阶滞后项开始逐渐趋向于零。



利用图形进行定阶只是一种模型识别的辅助手段。实际上，对于一个时间序列的分析，要建立什么样的模型才算正确或合理，需要反复地对模型进行调整。

在 SAS 过程中，可以使用 ARIMA 过程中的 IDENTIFY 语句绘制 ACF 图和 PACF 图以进行模型识别，具体程序如下。

```
proc arima data=Sasuser.Sales_Monthly;      /*调用 ARIMA 过程*/
  identify var=Sales(1);                     /*对 Sales 序列的一阶差分进行模型定阶*/
run;
```

该程序与在 11.1.2 小节中用自相关系数图和自相关白噪声检验平稳性是一致的，其输出的 ACF 图和 PACF 图可用于模型定阶，把两个图放在一起进行考察，如图 11-6 所示。

对于 SAS 系统输出的 ACF 图和 PACF 图，为了对比分析，此处进行了图形截取。SAS 中的 ACF 图是从滞后“0”期开始的，在分析时不计算在参数  $p$  或  $q$  之内。

图 11-6 中的 ACF 图在滞后期  $p = 2$  之后截尾，而 PACF 图随着滞后期扩大，拖尾趋势较为明显。根据表 11-2 所示的模型定阶依据，可以把模型初步定为 MA(2)，由于模型是由一阶差分数据而得，因此也可以记为 ARIMA(0, 1, 2)。

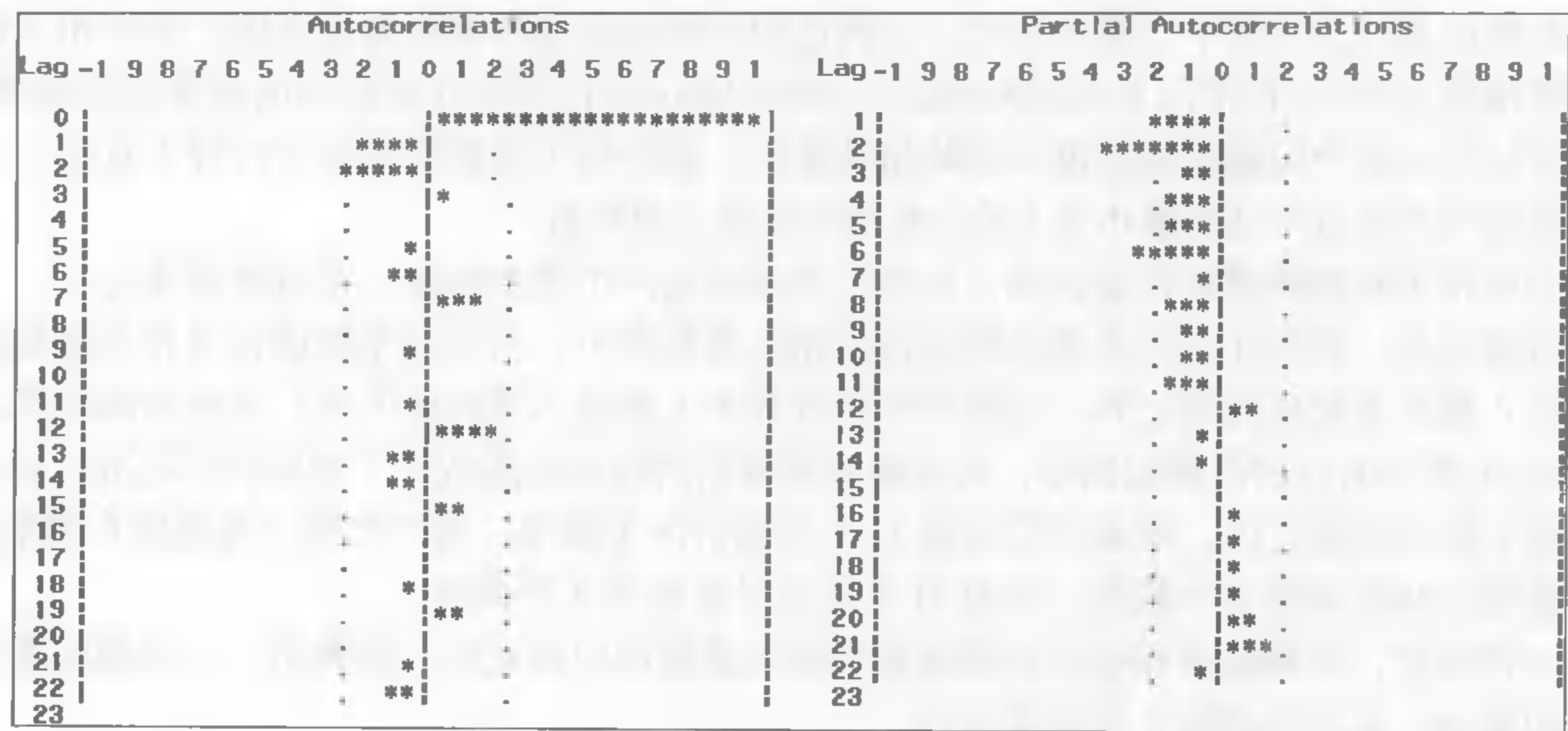


图 11-6 Sales 一阶差分的 ACF 图和 PACF 图

(2) 计算扩展的样本自相关函数并利用其估计值进行模型识别。

利用该方法进行识别的语句是 ARIMA 过程中 IDENTIFY 语句。在该语句的选项中加入 ESACF 关键字，并指定  $p$  和  $q$  的上、下限值。如对例 11-1 的数据进行识别，具体程序如下：

```
identify var=Sales(1) esacf p=(0:6) q=(0:6); /*对 Sales 一阶差分数据进行扩展样本自相关系数估计值模型识别，指定  $p$  和  $q$  的最小值均为 0，最大值均为 6*/
run;
```

因为一旦调用 ARIMA 过程，只要没有 QUIT 语句作为退出调用的标记，则可以一直使用其过程当中的各种语句，故本步骤及后续分析程序中不再加入调用 ARIMA 过程的 PROC 点。

在指定  $p$  和  $q$  的最大值和最小值时，可以依据 ACF 图和 PACF 图的情形自行判断，并没有特殊的规定。

运行程序后，可以得到各类模型之间的扩展自相关系数估计值和系统给出的模型选择依据，如图 11-7 所示。

The ARIMA Procedure							
ESACF Probability Values							
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5	MA 6
AR 0	0.0419	0.0144	0.6970	0.8706	0.7131	0.3960	0.2626
AR 1	<.0001	0.0026	0.5395	0.8820	0.9960	0.5548	0.2480
AR 2	0.0023	0.0006	0.5605	0.9119	0.9598	0.4596	0.4450
AR 3	<.0001	0.1520	0.4303	0.9185	0.4890	0.4762	0.8951
AR 4	<.0001	0.0007	0.8829	<.0001	0.6978	0.4114	0.9043
AR 5	<.0001	0.0302	0.7008	<.0001	0.0692	0.7618	0.9061
AR 6	0.0731	<.0001	0.0030	0.0007	0.1355	0.5199	0.8476

ARMA(p+d,q)  
Tentative  
Order  
Selection  
Tests

---ESACF---  
p+d      q

0	2
1	2
2	2
5	4
6	4

(5% Significance Level)

图 11-7 Sales 一阶差分数据的 ESACF 模型识别

图 11-7 中的第 1 个“ESACF Probability Values”表格表示扩展自相关系数的估计值。主

要考察第 2 个 “ARMA (p + d, q) Tentative Order Selection Tests” 表格，该表依据默认的显著性水平  $\alpha = 0.05$ （也可以通过在 IDENTIFY 语句选项中加入 “ALPHA=” 关键字自行指定），从上至下给出了优先选用的模型识别结果。从该结果中可以看出，“p + d = 0” 和 “q = 2” 的模型是最优选择，其次为 “p + d = 1” 和 “q = 2” 的模型。因为对于 Sales 数据，已经进行了一阶差分，上述结果是针对 Sales 的一阶差分之后的数据而言的，故此时的 d 等于 0。所以 “0” 和 “q=2” 是最优选择，即 MA(2)。

(3) 利用最小信息准则进行模型识别。

在 SAS 系统中，还可以计算 ARMA(p, q) 所有可能模型的 BIC 信息指数，并可根据系统计算出 BIC 指数最小的模型以作为识别依据。

利用最小信息准则进行识别的语句是 ARIMA 过程中 IDENTIFY 语句。在该语句的选项中加入 MINIC 关键字，并指定 p 和 q 的上、下限值。

如对例 11-1 中的数据进行最小信息准则识别，具体程序如下。

```
identify var=Sales(1) minic p=(0:6) q=(0:6); /*对 Sales 一阶差分数据进行最小信息准则模型识别, 指定 p 和 q 的最小值均为 0, 最大值均为 6*/
run;
```

运行程序后，可以得到各种模型的最小信息数值表，如图 11-8 所示。

Minimum Information Criterion							
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5	MA 6
AR 0	8.272563	8.128164	8.104162	8.118319	8.155991	8.203643	8.23339
AR 1	8.216523	8.138399	8.151762	8.152807	8.190966	8.238413	8.256368
AR 2	8.227007	8.175605	8.191151	8.202261	8.222713	8.2701	8.274123
AR 3	8.226148	8.219325	8.230157	8.241269	8.272272	8.317367	8.318899
AR 4	8.246637	8.254711	8.278414	8.290101	8.307148	8.356427	8.36071
AR 5	8.287545	8.301024	8.32602	8.3316	8.351156	8.39249	8.410279
AR 6	8.249197	8.289152	8.322734	8.319321	8.363559	8.401316	8.450027
Error series model: AR(12)							
Minimum Table Value: BIC(0,2) = 8.104162							

图 11-8 Sales 一阶差分的模型信息指数表

在图 11-8 输出的各种模型的 BIC 信息指数中，SAS 系统在该表的最后一行 “Minimum Table Value” 中自动提示 “BIC(0,2) = 8.104162”，即 BIC 指数最小的模型为 ARMA(0,2)，亦即 MA(2)。故例 11-1 的模型可以定为 MA(2)。

(4) 利用典型相关系数平方估计值进行模型识别。

在 SAS 系统中，还可以利用典型相关系数平方估计值进行模型识别，同样使用 ARIMA 过程中的 IDENTIFY 语句进行。在该语句中加入 SCAN 关键字，并指定 p 和 q 的上、下限值。如对例 11-1 中的数据进行识别，具体程序如下。

```
identify var=Sales(1) scan p=(0:6) q=(0:6); /*对 Sales 一阶差分数据进行 SCAN 识别, 指定 p 和 q 的最小值均为 0, 最大值均为 6*/
run;
```

运行程序后，可以得到各类模型之间的典型相关系数平方估计值和用于检验这些估计量显著性的概率值，最后系统同样自动给出的模型选择依据，如图 11-9 所示。

在图 11-9 中的最后一张表格中，系统依据默认的显著性水平  $\alpha = 0.05$ （也可以通过在 IDENTIFY 语句选项中加入 “ALPHA=” 关键字自行指定），从上至下给出了优先选用的模型识别结果。在该结果中可以看到，“p + d = 1” 和 “q = 1” 的模型是最优选择，即 ARMA(1,1)；其次为 “q + d = 0” 和 “q = 2” 的模型，即 MA(2)。

The ARIMA Procedure							
Squared Canonical Correlation Estimates							
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5	MA 6
AR 0	0.0456	0.0723	0.0021	0.0004	0.0019	0.0102	0.0188
AR 1	0.1089	0.0326	0.0001	0.0024	<.0001	0.0139	0.0120
AR 2	0.0142	0.0090	0.0025	<.0001	0.0024	0.0177	0.0035
AR 3	0.0236	0.0011	0.0033	0.0147	0.0174	0.0139	<.0001
AR 4	0.0183	0.0033	<.0001	0.0166	0.0011	0.0028	0.0053
AR 5	0.0646	0.0298	0.0157	0.0143	0.0026	0.0022	<.0001
AR 6	0.0026	0.0160	0.0049	0.0059	0.0100	<.0001	0.0004
SCAN Chi-Square(1) Probability Values							
Lags	MA 0	MA 1	MA 2	MA 3	MA 4	MA 5	MA 6
AR 0	0.0392	0.0130	0.6986	0.8714	0.7153	0.4038	0.2599
AR 1	0.0013	0.1476	0.9223	0.7054	0.9923	0.3611	0.3948
AR 2	0.2598	0.4398	0.6799	0.9542	0.7034	0.3240	0.6767
AR 3	0.1469	0.8019	0.6582	0.3859	0.3310	0.3529	0.9519
AR 4	0.2055	0.6591	0.9569	0.3664	0.8120	0.7127	0.6324
AR 5	0.0165	0.1853	0.3400	0.3760	0.7147	0.7866	0.9864
AR 6	0.6409	0.2757	0.5840	0.5727	0.4635	0.9875	0.9075
ARMA(p+d,q) Tentative Order Selection Tests							
----SCAN----							
p+d                      q							
1                              1							
0                              2							
6                              0							
(5% Significance Level)							

图 11-9 Sales 一阶差分数据的 SCAN 模型识别

依据以上模型识别的方法得到结论：可以认为 MA(2)模型是比较适合的。接下来便可对其进行参数估计和模型检验、评价。

2. 模型参数估计及检验

在对时间序列模型进行识别并确定模型的具体形式之后，便可以利用样本数据进行模型的参数估计并对估计结果进行检验。对于 ARMA 模型，可以对其拟合性和参数估计显著性等方面进行检验。此外，对于一个适当的 ARMA 模型，还应当保证其残差项无自相关性，即对残差项进行白噪声检验。如果模型残差项非白噪声，则需要重新对模型进行识别。

SAS 系统在 ARIMA 过程中提供了 ESTIMATE 语句以进行参数估计并得到相应的检验结果，并可以在其“METHOD=”选项中通过关键字指定参数估计的方法。其中，可指定的估计方法关键字为 ML、ULS 和 CLS，分别表示极大似然法、非条件最小二乘法 and 条件最小二乘法。CLS 方法是系统缺省的默认方法。



利用 SAS 程序进行参数估计时，ESTIMATE 语句必须与 IDENTIFY 语句配合使用。如果 ESTIMATE 语句之前有多个 IDENTIFY 语句，则在模型进行参数估计时，ESTIMATE 语句估计的是离该语句最近的 IDENTIFY 语句指定的时间序列或变量。

对于例 11-1 中的数据，已经知道 Sales 数据经过一阶差分之后是平稳序列，并可以把模型识别为 MA(2)，具体程序如下。

```
identify var=Sales(1) noprint; /*关键字 noprint 表示该语句不输出任何结果与信息*/
estimate p=0 q=2 plot; /*估计 Sales 一阶差分时间序列的 MA(2)模型，并绘制残差自相关系数图*/
run;
```

运行程序之后，除可以得到 IDENTIFY 语句的输出结果之外，ESTIMATE 语句还可以输出的参数估计及其显著性检验结果，如图 11-10 所示。

Conditional Least Squares Estimation					
Parameter	Estimate	Standard Error	t Value	Approx Pr >  t	Lag
MU	1.72539	1.31159	1.32	0.1918	0
MA1,1	0.43648	0.09905	4.41	<.0001	1
MA1,2	0.38249	0.10142	3.77	0.0003	2

图 11-10 ARMA 模型的参数估计结果

从图 11-10 中可以看出，针对 MA(2)模型  $X_t = a_t - \theta_1 a_{t-1} - \theta_2 a_{t-2}$ ，其估计均值项用 MU 表示。当 ARMA 模型参数均为空时（即  $p、q$  均为 0 时），MU 的值为所分析序列的样本均值。模型的两个参数  $\theta_1$  和  $\theta_2$  的参数估计值用“MA1,1”和“MA1,2”表示，具体估计值分别为 0.43648 和 0.38249。由于 ARMA 估计过程中的标准误差是基于大数定律的，因此图 11-10 中的标准误差是近似值，故系统自动用“Approx”进行标注，其  $t$  值及对应的检验  $P$  值在小样本（本例数据经过一阶差分之后样本量为 91，为大样本）条件下不一定可靠，从 MA(2)模型两个参数估计值的显著性  $t$  检验可以看出，它们均非常显著。

与模型拟合程度有关的检验统计量如图 11-11 所示。

图 11-11 所示的拟合统计量主要用于多个模型进行对比分析的场所。其中的“Constant Estimate”是均值项 MU 和自回归参数估计值的函数，仅在 AR 或 ARMA 模型中进行估计，不适用于 MA 模型。“Variance Estimate”和“Std Error Estimate”表示模型残差项的方差和标准差。AIC 和 SBC 用于利用信息准则对模型进行比较，二者值越小，模型拟合得越好。

Constant Estimate	1.725388
Variance Estimate	4020.339
Std Error Estimate	63.40614
AIC	1016.416
SBC	1023.949
Number of Residuals	91
* AIC and SBC do not include log determinant.	

图 11-11 ARMA 模型拟合统计量

对于 ARMA 模型的参数估计和拟合，应当使得估计值后的模型残差项不存在自相关，即模型的残差项是白噪声。因此，还应当对模型的残差项进行白噪声检验，如图 11-12 所示。

Autocorrelation Check of Residuals									
To Lag	Chi-Square	DF	Pr > ChiSq	-----Autocorrelations-----					
6	2.01	4	0.7343	-0.011	-0.004	0.017	-0.064	-0.051	-0.115
12	8.52	10	0.5779	0.095	-0.014	-0.023	0.064	-0.003	0.218
18	10.17	16	0.8576	-0.009	0.001	0.101	0.047	0.045	0.012
24	12.95	22	0.9346	0.078	0.001	-0.057	-0.098	0.016	0.060

图 11-12 模型残差项的白噪声检验

残差项白噪声检验的原假设为残差项是白噪声，备择假设为非白噪声。观察图 11-12 中各滞后期的残差项序列的白噪声检验结果，发现本例中各滞后期的残差项不存在自相关，即“Pr>ChiSq”均远远大于  $\alpha$ “0.05 或 0.1”，可以认为本例建立的 MA(2)模型的残差项为白噪声。因此，MA(2)模型对于 Sales 的一阶差分序列而言是合适的。如果残差项白噪声检验没有通过，则需要对模型重新进行识别。

本例利用 ESTIMATE 语句中 PLOT 关键字绘制残差自相关系数图，如图 11-13 所示。

观察图 11-13 所示各滞后期（Lag = 0 时不考察）的残差自相关系数，可以看出它们均处于置信区间之内在 0 附近波动，是一白噪声序列，这与图 11-12 所示的检验结果一致。因此对于本例的数据，所建立的 MA(2)模型是比较合适的。

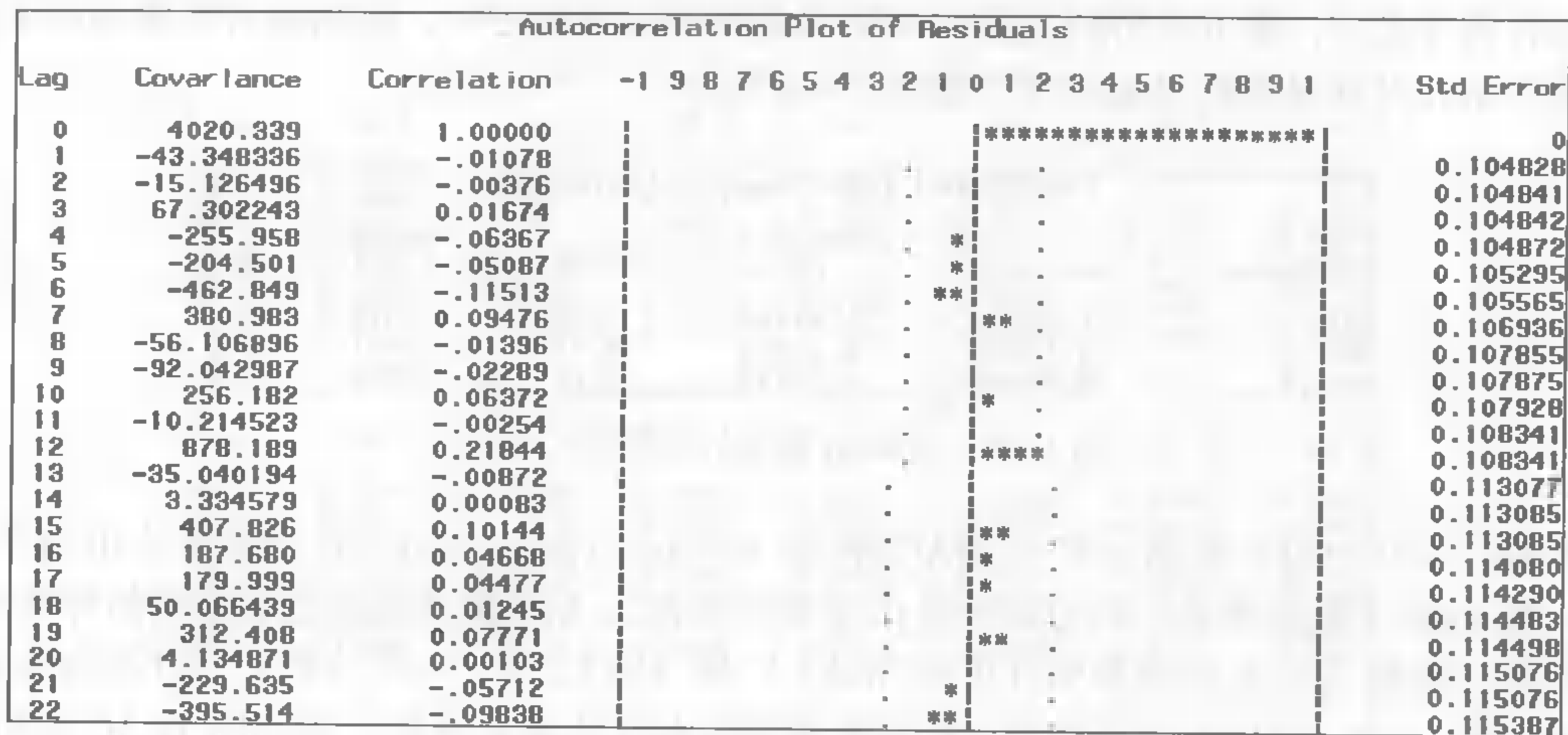
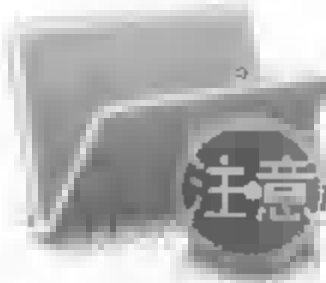


图 11-13 残差项的自相关系数图

3. 模型的预测

经过识别和参数估计并进行相应的检验之后，便可利用所建立的模型进行预测。  
SAS 系统中的 ARIMA 过程提供 FORECAST 语句进行预测。



该语句必须与 ESTIMATE 语句配合使用。如果 FORECAST 语句之前有多个 ESTIMATE 语句，则在进行预测时，FORECAST 预测所用的模型是由离该语句最近的 ESTIMATE 语句估计出来的模型。

如对于例 11-1 中的数据，已经利用 ESTIMATE 语句对 MA(2)模型进行了参数估计，则可利用现有样本数据和模型参数估计的结果，使用以下程序对在 2008 年 9 月至 2008 年 12 月之间的 Sales 数据进行预测。

```
forecast lead=4 out=Sales_Predicted; /*表示向后预测 4 期，并把预测结果存储在 Sales_Predicted 数据集中*/
run;
```

运行程序后，可得到向后 4 期，即 2008 年 9 月至 12 月之间的 Sales 变量的预测值，如图 11-14 所示。

Forecasts for variable Sales				
Obs	Forecast	Std Error	95% Confidence Limits	
93	961.5620	63.4061	837.2882	1085.8357
94	948.5785	72.7808	805.9308	1091.2261
95	950.3039	73.6804	805.8930	1094.7148
96	952.0292	74.5692	805.8764	1098.1821

图 11-14 Sales 序列向后 4 期的预测值

图 11-14 中不仅给出了序列的预测值，还给出了预测标准误差、预测值的置信区间的上下界（置信度可由 FORECAST 语句中的“ALPHA=”关键字自行指定，默认为  $\alpha = 0.05$ ，即 95%置信度）。此外 FORECAST 语句还可以对分析数据集中的所有时期的数值进行预测，在 FORECAST 语句中加入“PRINTALL”关键字，则会在输出结果中输出所有时期的预测值，读者也可打开本例生成的 Sales\_Predicted.sas7bdat 临时数据集进行查看。

为了比较观测值和预测值之间的关系，可以把 Sales 序列的所有观测值和预测值放在一

起, 绘制图 11-15 所示的时间序列图。

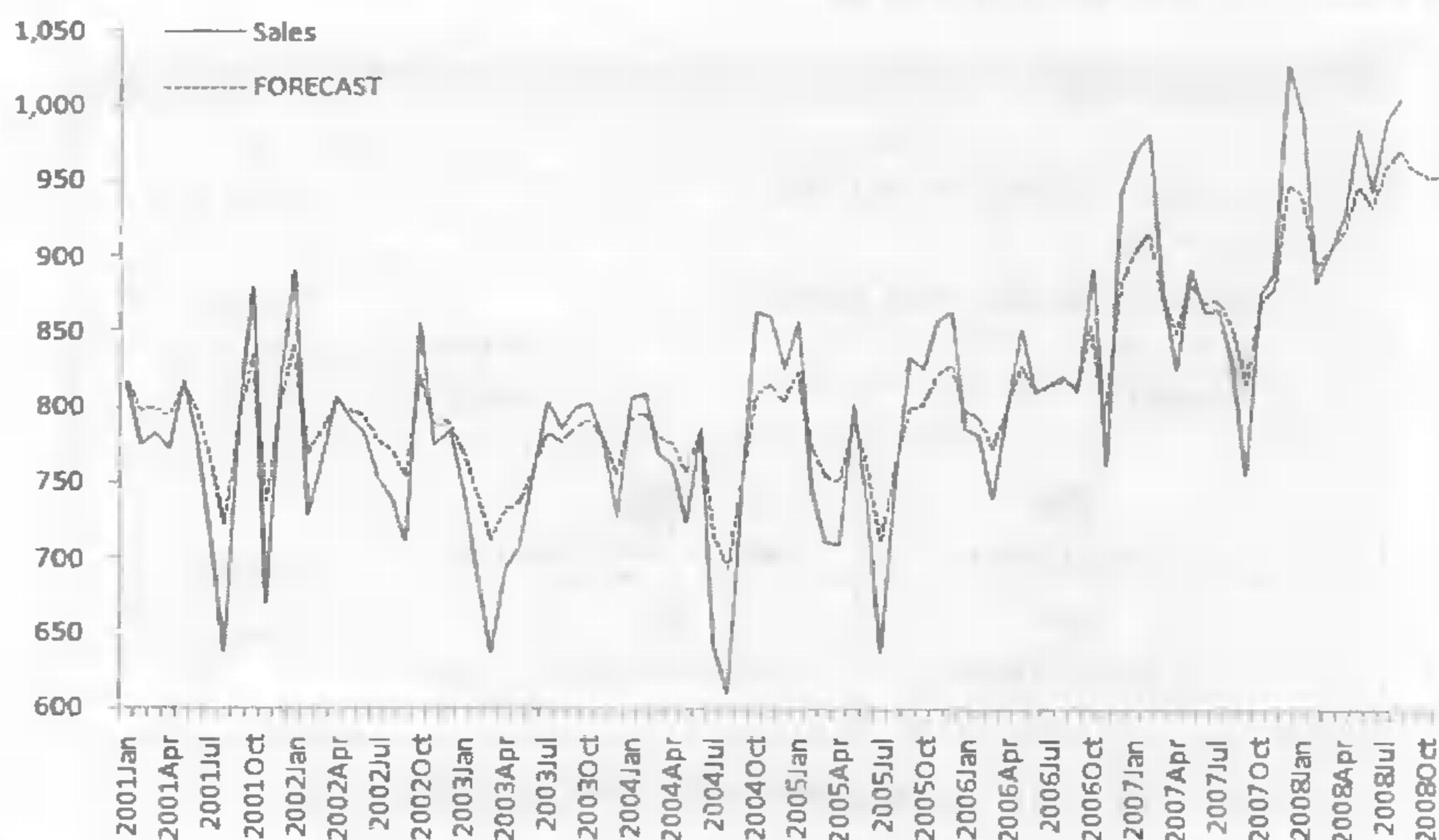


图 11-15 观测值与预测值的时间序列图

从图 11-15 所示的图形结果来看, 观测值与真实值之间还是比较吻合的, 且它们在波动的方向和幅度上基本一致。

综上所述, 对于例 11-1 中的数据进行时间序列分析, 其模型识别、参数估计及检验、预测过程的完整程序如下。

```
ods html;
ods graphics on;
proc arima data=Sasuser.Sales_Monthly;
  identify var=Sales(1);
  estimate p=1 q=1 plot;
  forecast printall lead=4 out=Sales_Predicted;
quit;
run;
ods graphics off;
ods html close;
```

在本章中, 由 ARIMA 过程输出的自相关系数、偏自相关系数图、残差自相关系数图等图形均是由文本拼凑而成的, 从审美角度来看不够美观, 在某些情况下还会影响分析过程。因此, 上述程序调用了 SAS 的 ODS (Output Delivery System) 系统来输出对应的结果和图形。ODS 系统输出的图形比较美观, 便于依据各种图形对时间序列进行分析。这里不再列示上述程序得到的输出内容, 请读者自行查看相关结果。

### 11.2.3 利用 SAS 时间序列预测系统进行菜单操作

在 11.2.2 小节中, 都是用编程的方式进行时间序列分析。除此之外, SAS 系统还提供了图形界面的时间序列预测系统来完成 11.2.2 小节中的模型识别、参数估计及检验和预测过程。

**STEP 1** 选择 SAS 系统菜单 “Solutions → Analysis → Time Series Forecasting System”, 弹

出 SAS 系统的时间序列预测对话框，如图 11-16 所示。可以使用该对话框提供的时间序列分析功能实现 11.2.2 小节中介绍的所有过程。

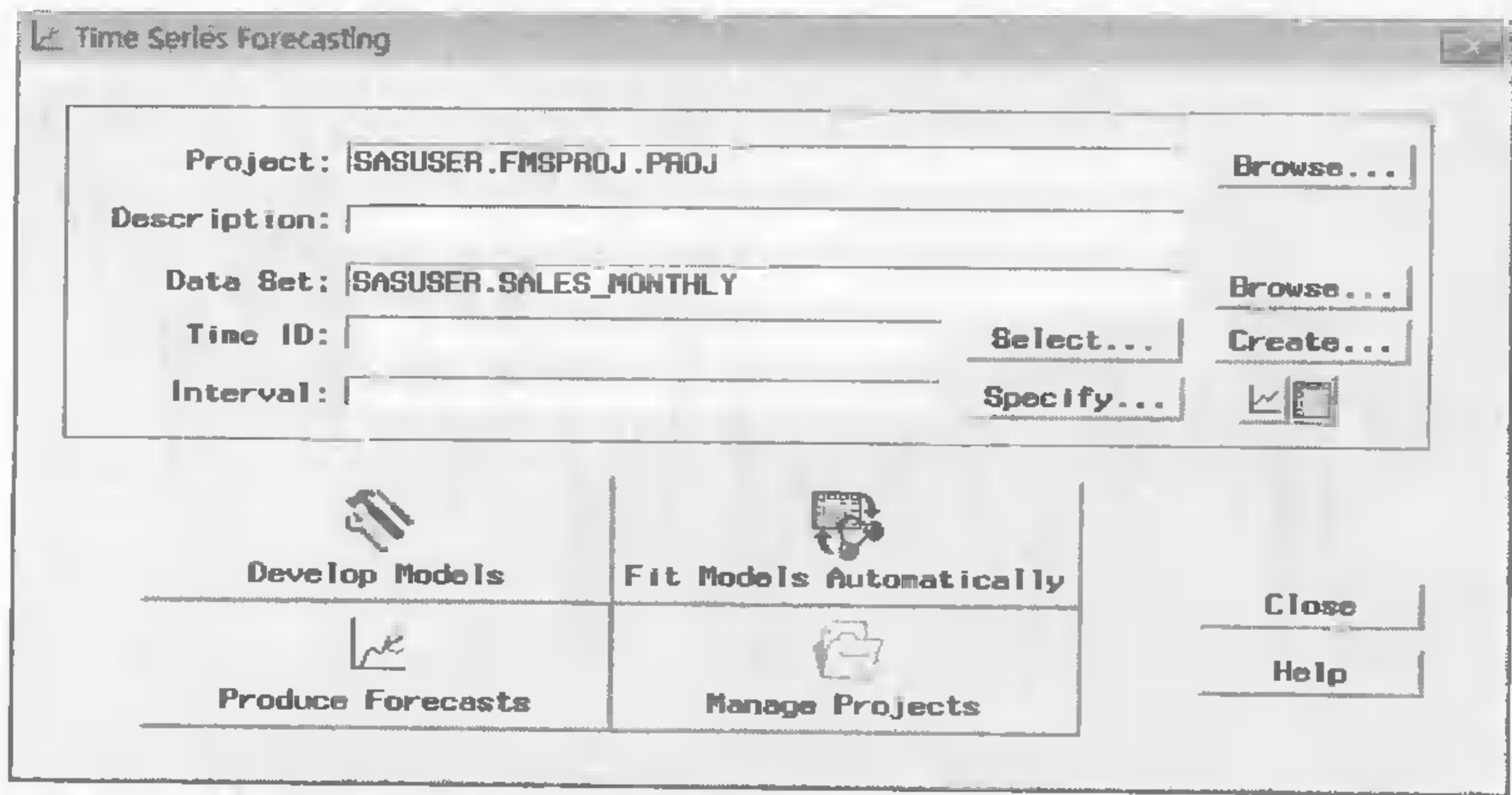


图 11-16 SAS 系统的时间序列预测系统对话框

**STEP 2** 对于时间序列的分析，SAS 系统可以以项目的形式对其分析过程进行管理。因此，在图 11-16 中的“Project”文本输入框中，可以输入分析项目的名称（系统默认为新建项目自动命名）。同时，用户可在“Description”文本输入框中输入一些对于本分析过程的描述。“Data Set”文本输入框用于指定分析所用的数据集，单击该文本输入框右边的“Browse...”按钮，可以弹出 SAS 数据库和数据集浏览器，以方便用户找到想要分析的数据。“Time ID”文本输入框用于指定分析数据集中代表时间的时间序列变量（即日期型变量），单击“Select...”按钮可以在分析数据集中选择已有的日期型变量，单击“Create...”按钮可以新生成一个日期型变量。最后一行的“Interval”文本输入框主要用于指定日期型变量的时间间隔，单击“Specify...”按钮可以选择给定的时间间隔。

**STEP 3** 为了与前面章节中的分析过程对应，本小节仍然使用例 11-1 中的数据进行分析。单击“Data Set”文本输入框右边的“Browse...”按钮，弹出数据集选择对话框，如图 11-17 所示。

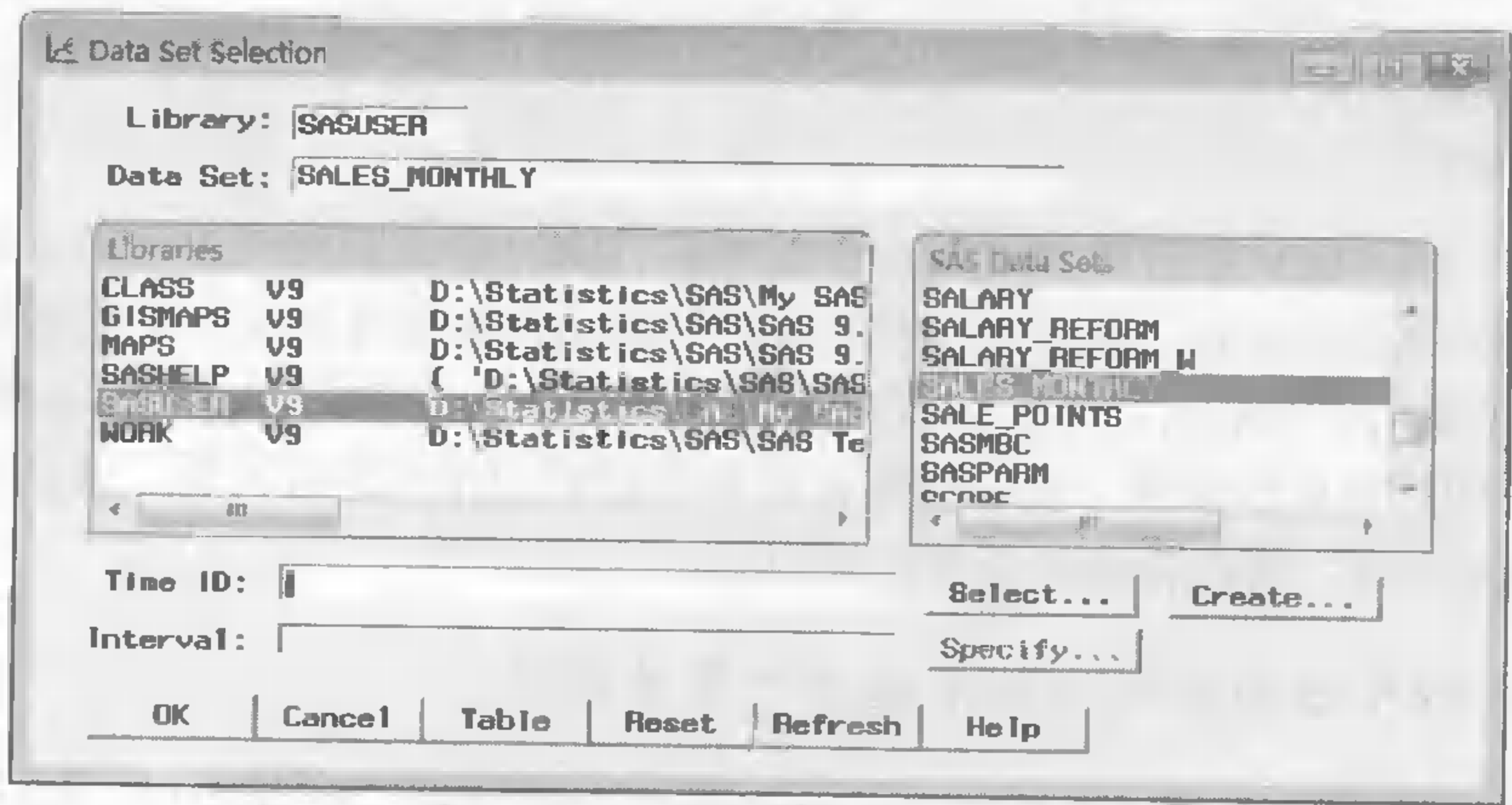


图 11-17 数据集选择对话框

**STEP 4** 本例数据集名为 Sales\_Monthly.sas7bdat，被存储在 Sasuser 数据库中。因此在图 11-17 所示的对话框左边选中“SASUSER”数据库，右边选中“SALES\_MONTHLY”数据集。

在 SAS 时间序列预测系统中，必须指定分析数据集中代表时间的日期型变量。如果数据集中已有现成的日期型变量，则单击图 11-17 所示的对话框右下的“Select...”按钮，即可选择数据集当中的日期型变量。本例数据中虽然有“Month”变量，但是它是一个字符型变量，并不是日期型变量。故可以单击右下的“Create...”按钮，在数据集当中生成一个新的日期型变量。

**STEP 5** 单击“Create...”按钮之后，SAS 系统提供了 4 种生成日期型数据的方法，本例采用第 1 种方法“Create from starting date and frequency...”，即给定起始日期生成日期型变量。选中该种生成变量的方法后，弹出日期变量生成对话框，如图 11-18 所示。

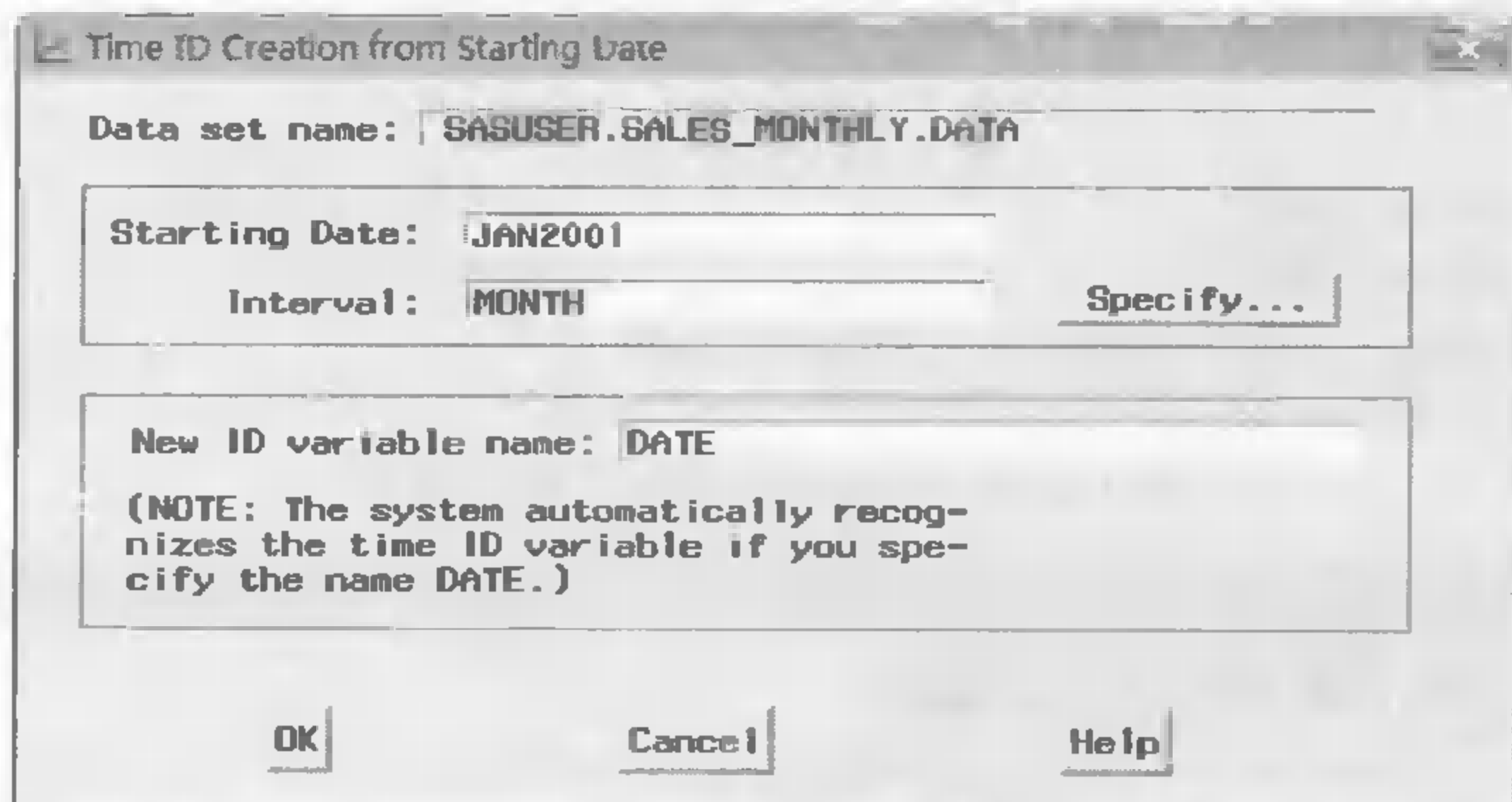


图 11-18 日期变量生成对话框

**STEP 6** 在图 11-18 中的“Starting Date”文本输入框中输入时间的起始点，“Interval”文本输入框用于指定日期变量的时间间隔。本例分析的是自 2001 年 1 月份开始的月度数据，因此在“Interval”中输入“MONTH”，在“Starting Date”中输入“JAN2001”。单击“Specify...”按钮，可以依据 SAS 系统提供的时间间隔定义日期型变量。在“New ID variable name”文本输入框中可以输入新生成的日期型变量的名字，本例系统默认命名为“DATE”。单击“OK”按钮返回图 11-17 所示对话框。在该对话框中，日期型变量及其对应的时间间隔类型会自动与图 11-18 中的设定对应上，再单击“OK”按钮返回图 11-16 所示对话框。至此，已经完成时间序列的数据预处理过程，接下来可进行时间序列平稳性检验、模型识别、参数估计及检验和预测。

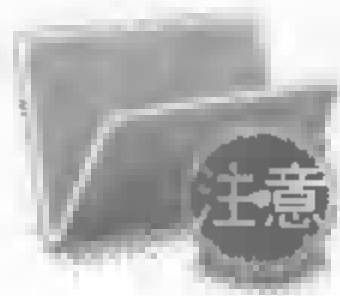


图 11-17 所示的对话框中对应按钮的功能在图 11-16 所示的对话框中同样能够实现。

### 1. 利用图形进行时间序列平稳性检验及模型识别

**STEP 1** 在图 11-16 所示的对话框中，单击 可以查看进行分析的数据集，单击 按钮则可以查看依据数据集绘制的各类图形。单击 按钮之后，弹出指定分析对象（即指定需要分析的时间序列变量）的对话框，如图 11-19 所示。

**STEP 2** 在图 11-19 下方的“Time Series Variables”列表框中，选择“Sales”变量作为需要进行分析的对象，再单击“Graph”按钮，可以得到 SAS 时间序列预测系统的图形操作界面（一旦指定好分析对象之后，以后单击 按钮便可直接弹出图形操作界面），如图 11-20 所示。

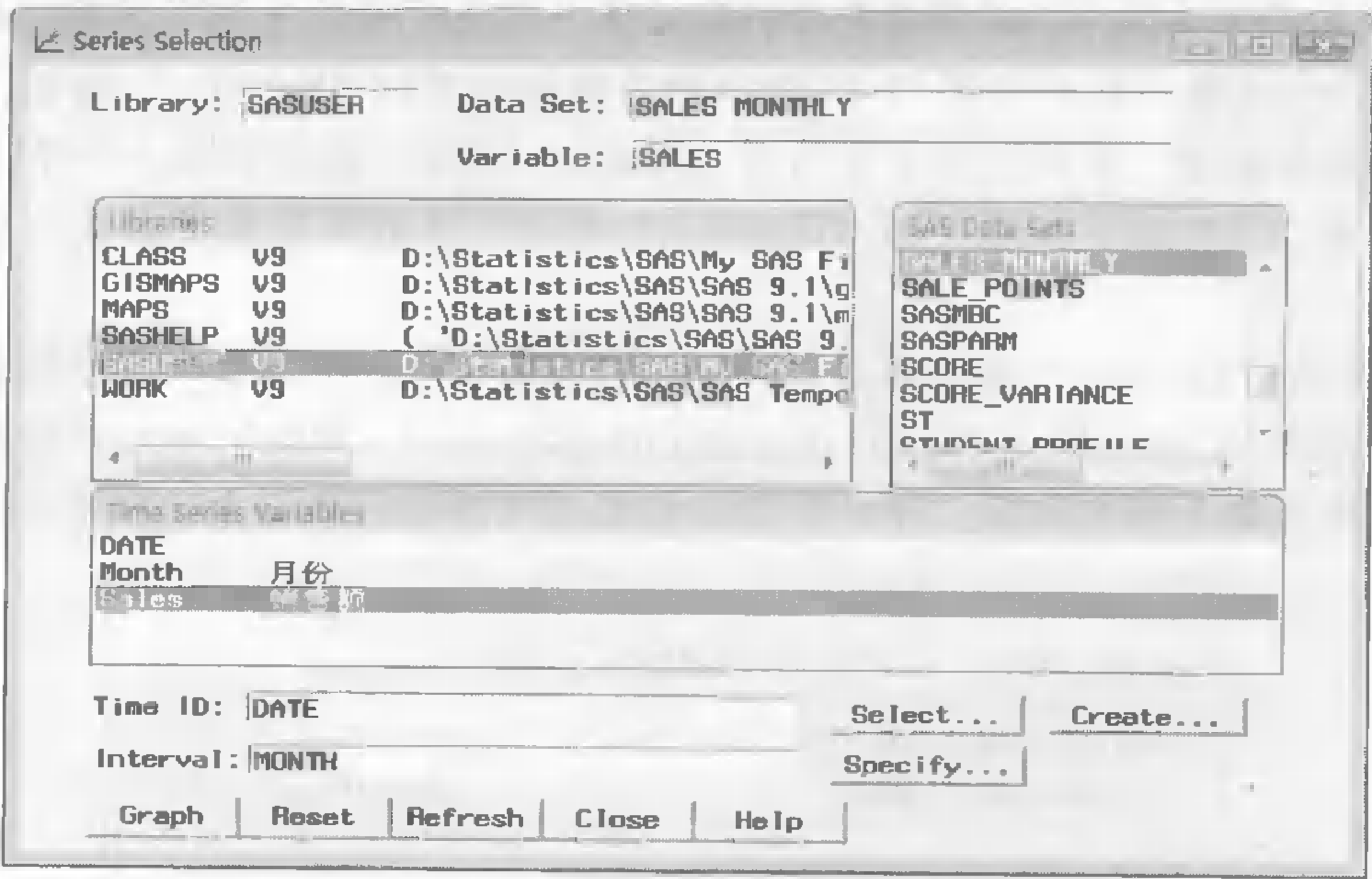


图 11-19 时间序列分析对象对话框

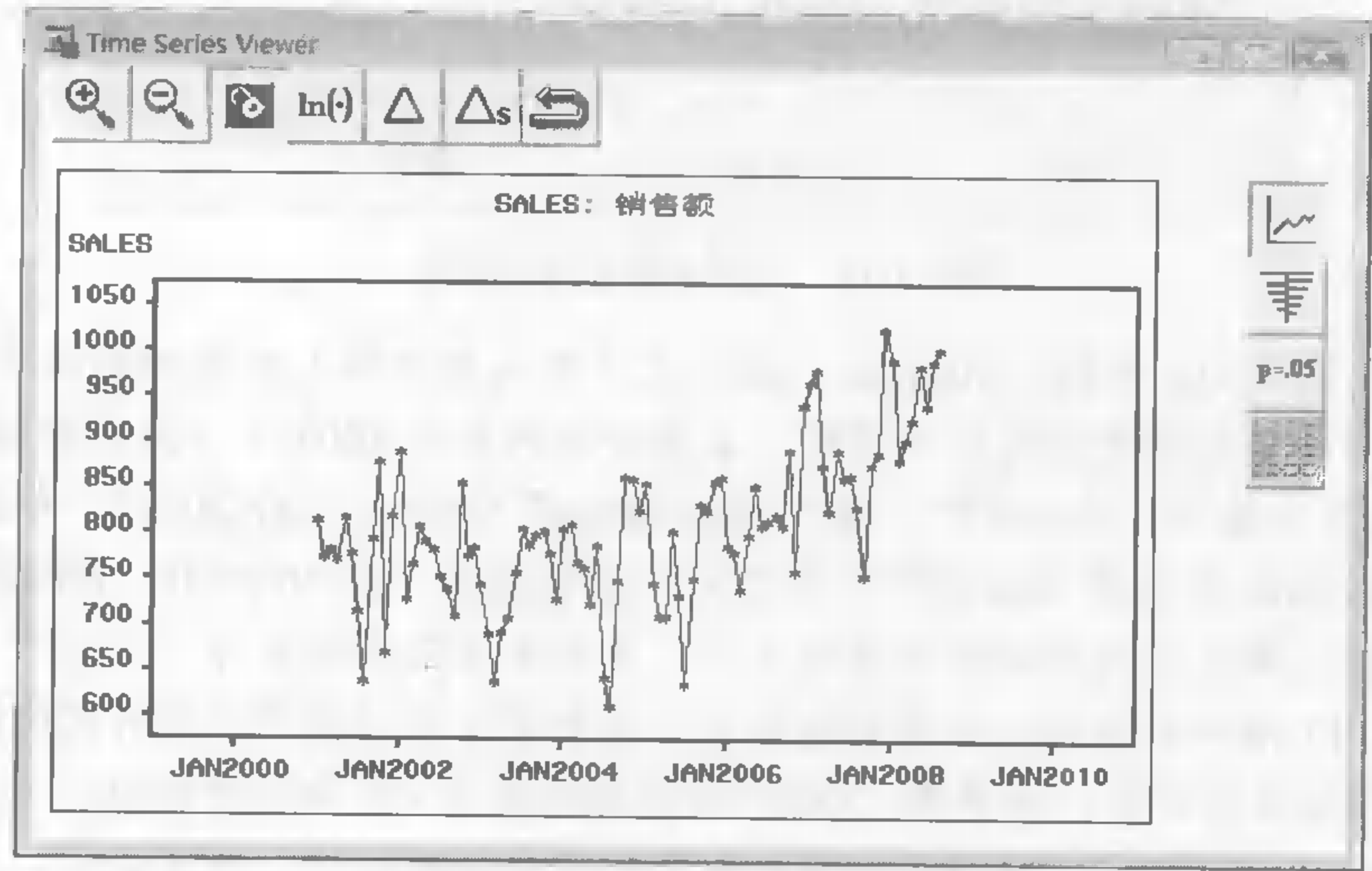


图 11-20 “Time Series Viewer” 图形操作界面

图 11-20 所示的对话框默认显示分析对象原始数据的时间序列图。在该图的右上角可以对图形类型进行切换。

- : 输出原始观测值的趋势图。
  - : 输出自相关系数图、偏自相关系数图和逆自相关系数图。
  - : 输出白噪声和平稳性检验图。
  - : 查看数据集当中的数据。
- 单击图 11-20 上方的按钮，可对分析数据进行数据变换。
- : 对时间序列进行对数变换。
  - : 对时间序列进行一阶差分。
  - : 对时间系列进行季节差分。
  - : 退出图形系统。

如对 Sales 数据进行一阶差分并进行平稳性检验，可单击 按钮，然后再单击 按钮，

即可得到 Sales 数据一阶差分之后的自相关、偏自相关系数图，如图 11-21 所示。

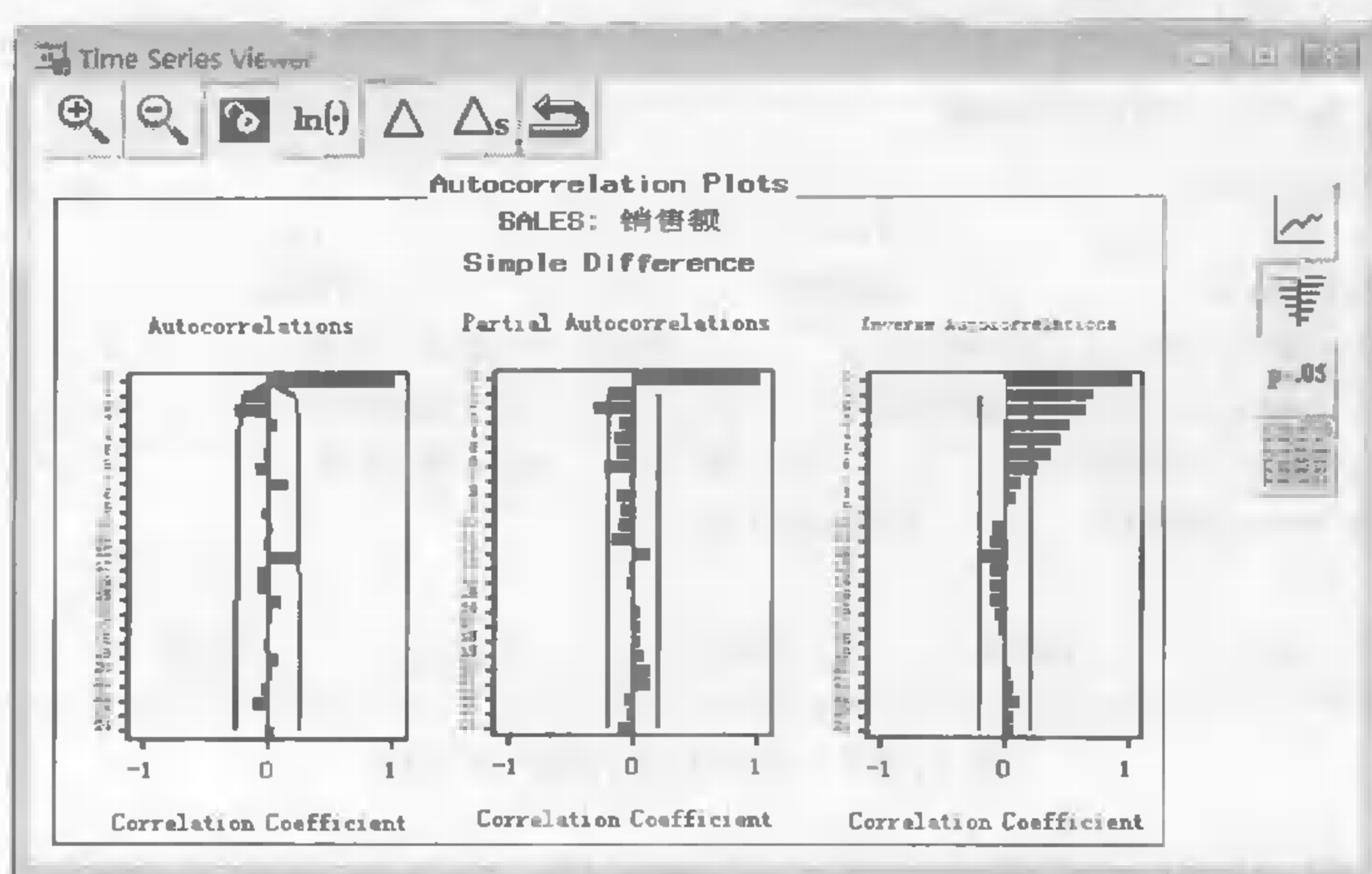


图 11-21 Sales 数据一阶差分之后的自相关、偏自相关系数图

根据图 11-21 所示图形，可以利用 ACF 图判断时间序列的平稳性，也可同时依据 ACF 图和 PACF 图对模型进行识别（关于平稳性判定和模型识别依据详见 11.1.2 小节和 11.2.1 小节）。

## 2. 模型的参数估计、检验和预测

**STEP 1** 对数据进行平稳化和模型识别之后，返回图 11-16 所示的对话框，单击下方的“Develop Models”按钮，弹出模型设置对话框，如图 11-22 所示。

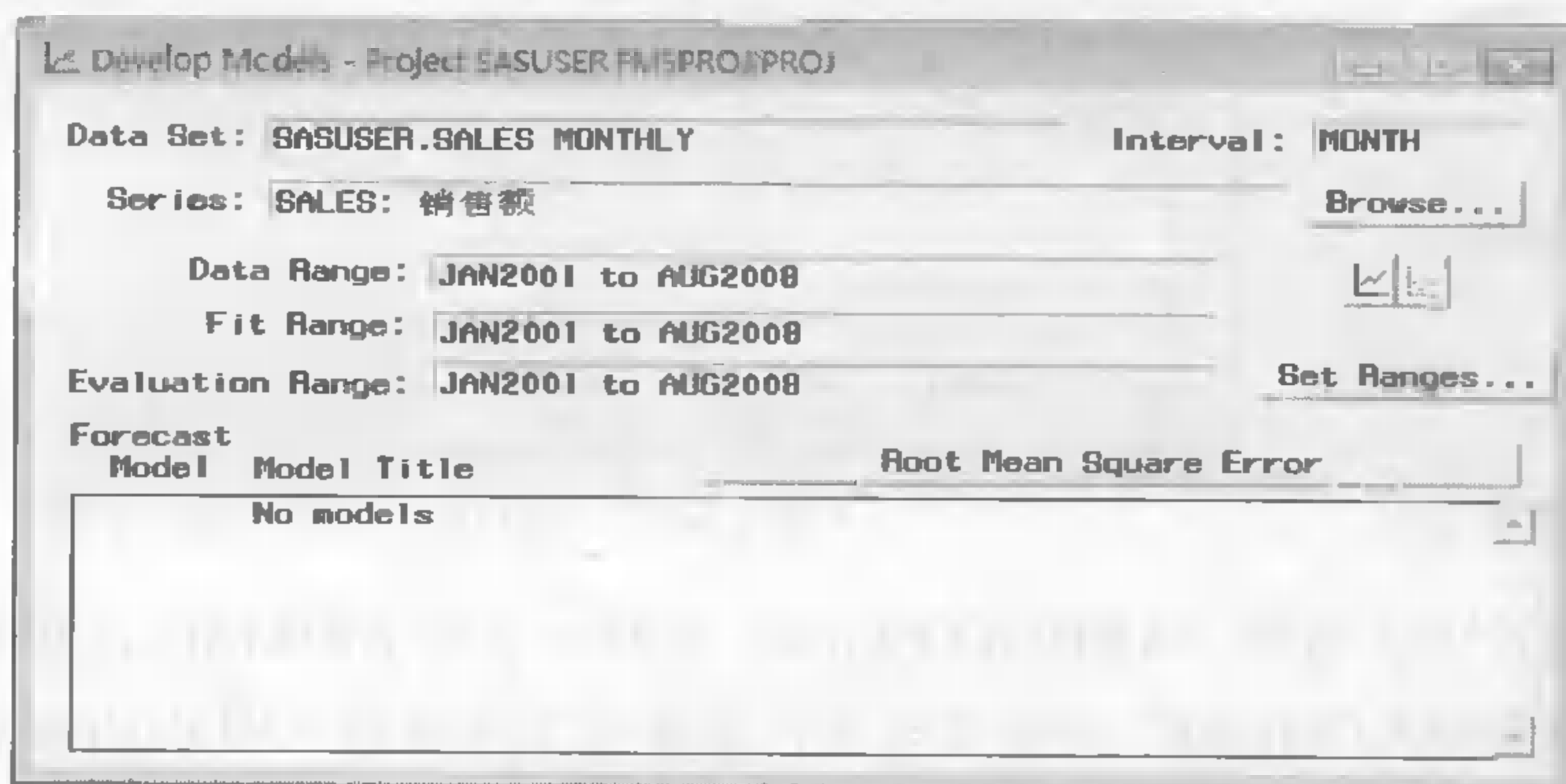


图 11-22 模型设置对话框

**STEP 2** 在图 11-22 中，单击“Set Ranges...”按钮可以设置用于模型参数估计的时间间隔范围，也可以设置利用模型进行预测的时期，如图 11-23 所示。

如想利用例 11-1 中的数据进行建模，并进行为期 1 年的预测，则可以把“Period of Fit”设定为从“JAN2001”到“AUG2008”，把“Forecast Horizon”设定为 12 期（也可以单击对应的▼按钮选择下拉菜单，将其设定为“1 Year”），具体时间为“AUG2009”。在设定时间范围的过程中，可以单击←和→按钮进行年份调整，单击↶和↷按钮进行月份调整。设置好模型估计和预测的时间范围之后，单击“OK”按钮返回图 11-22 所示的对话框。

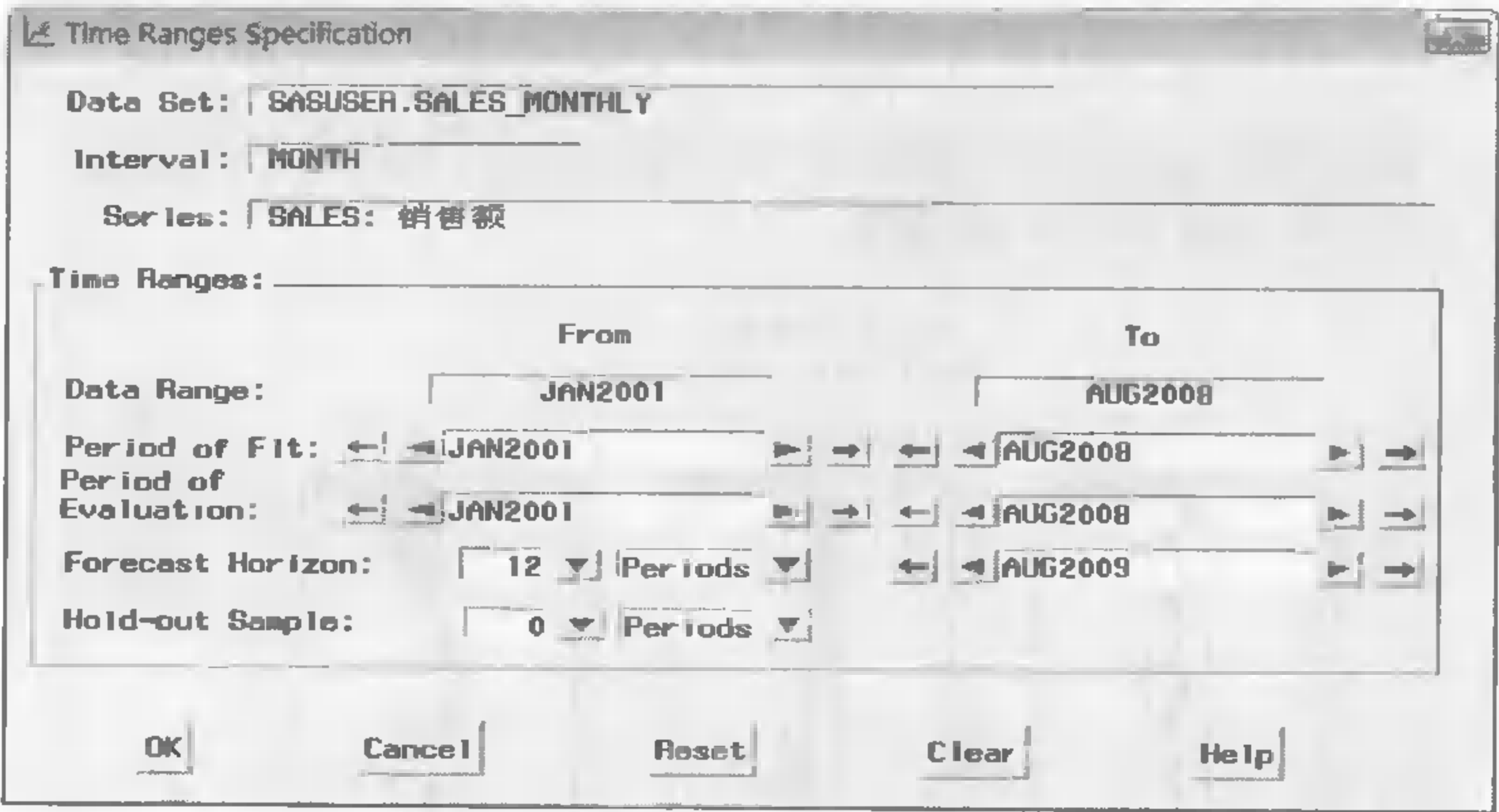


图 11-23 时间区间设定对话框

**STEP 3)** 在图 11-22 中的任意地方单击鼠标右键，或者对话框下方空白处单击鼠标左键，弹出拟合模型菜单，如图 11-24 所示。

**STEP 4)** 如需要建立 ARIMA 模型，则在该菜单中选中“Fit ARIMA Model...”选项，弹出 ARIMA 模型设定对话框，如图 11-25 所示。

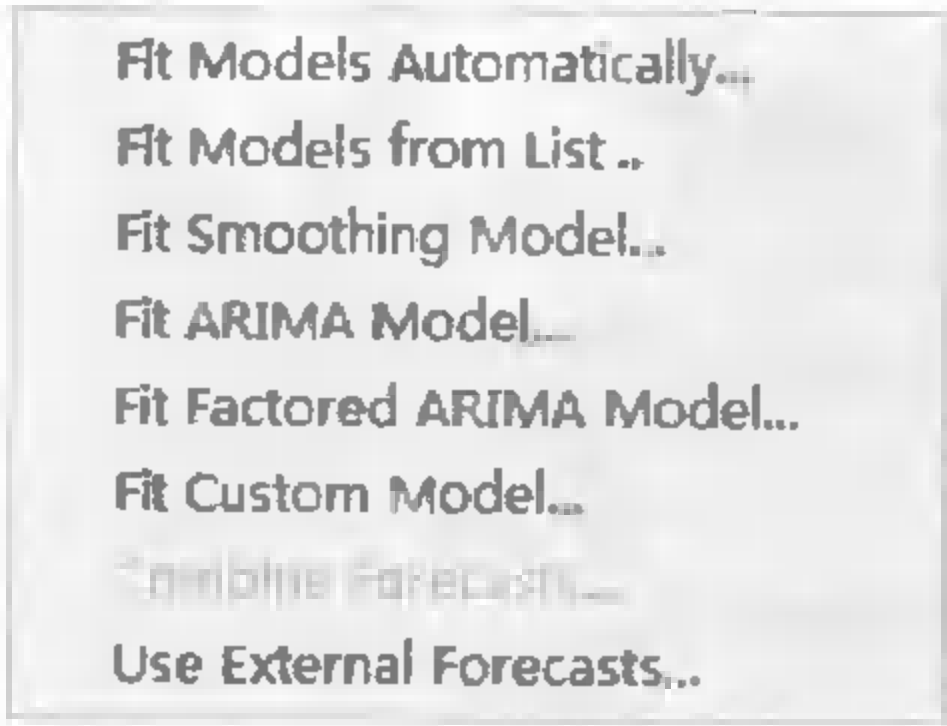


图 11-24 拟合模型菜单

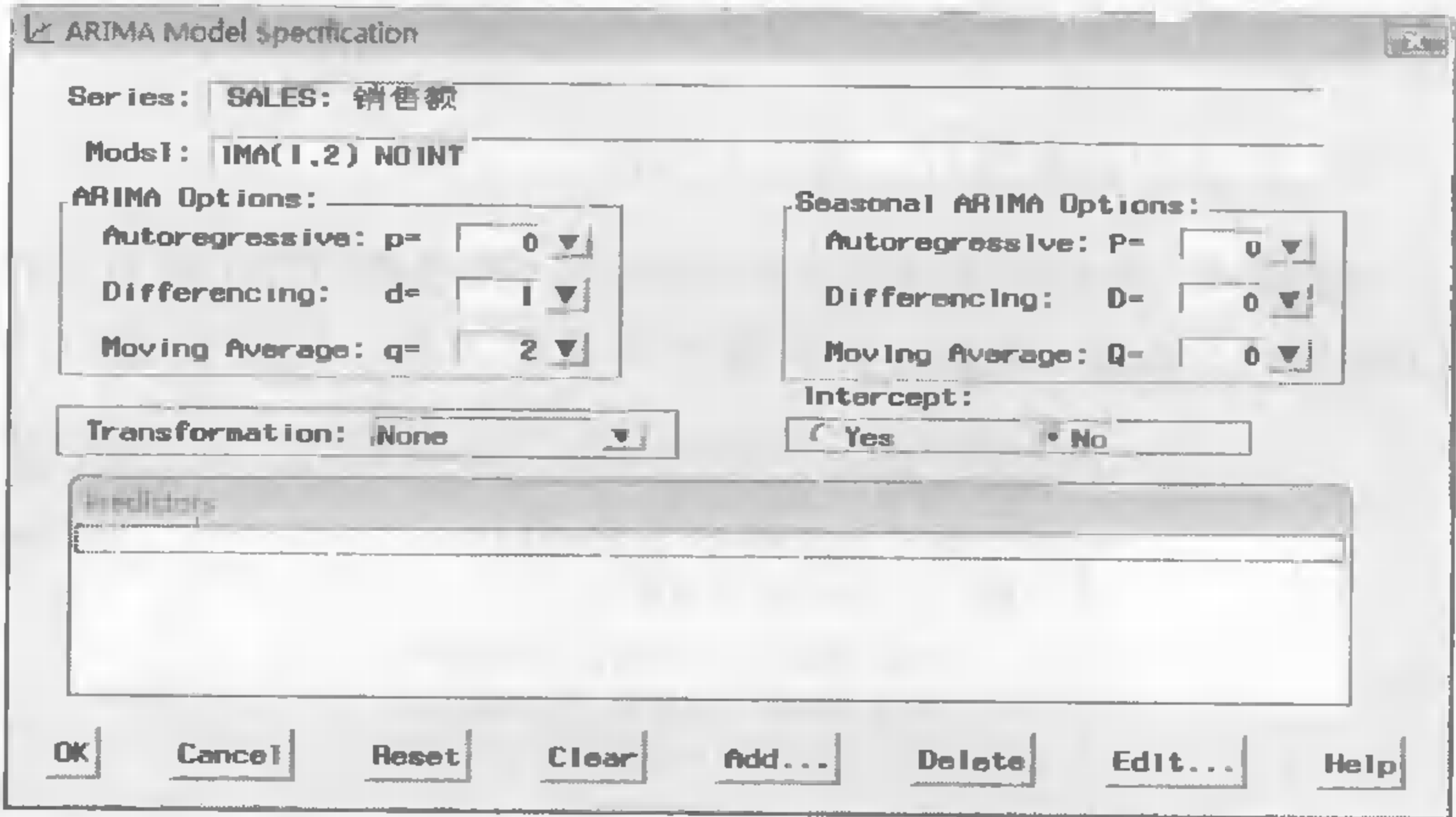


图 11-25 ARIMA 模型设定对话框

**STEP 5)** 图 11-25 左边的“ARIMA Options”分栏可以对  $ARIMA(p,d,q)$  的参数进行设置，右边“Seasonal ARIMA Options”分栏可以进行含有季节成分的  $ARIMA(p,d,q)(P,D,Q)$  模型参数的设置。

在 11.2.2 小节中，已经识别出 Sales 数据经过一阶差分之后的模型为 MA(2)。在图 11-25 中，需进行分析的变量是 Sales 的原始变量，故要对其进行一阶差分，因此在模型中， $d=0$ 。即对 Sales 原始变量进行分析的模型经过识别后，可考虑的模型形式为  $ARIMA(0, 1, 2)$ ，可在图 11-25 中的“ARIMA Options”分栏下把对应的参数设置好。因本例数据的季节性因素不明显，故不对“Seasonal ARIMA Options”选项进行设置。

**STEP 6)** 在“Transformation”下拉列表框中可以选择对原始变量进行对数、开方等数据变换，在“Intercept”下拉列表框中则可以选择模型当中是否包含截距项。在本例数据分析过程中，不用进行数据变换，也不需要估计截距项。单击“OK”按钮，系统自动拟合模型并给出相应的分析结果，如图 11-26 所示。

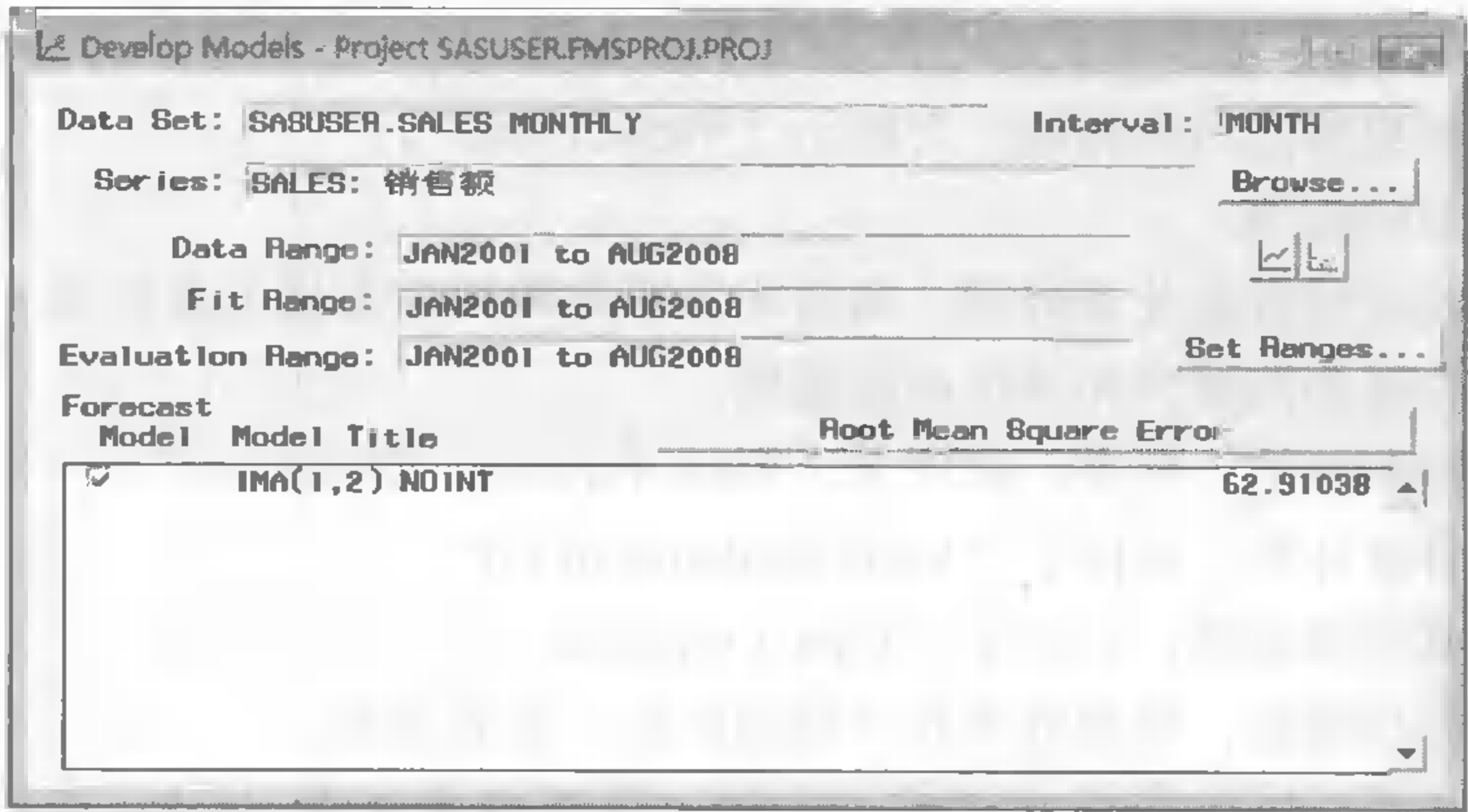


图 11-26 模型设置对话框（已进行模型拟合）

**STEP 7** 图 11-26 与图 11-22 之间的不同之处在于，图 11-26 中下部分的空白处已经具有进行拟合后的模型名称。选中对应的模型，单击鼠标右键，弹出模型查看和调整菜单，如图 11-27 所示。

图 11-27 所示菜单对应的功能如下。

- View Model: 查看模型的基本信息。
- View Parameter Estimates: 查看模型的参数估计结果。
- View Statistics of Fit: 查看模型拟合统计量。
- Edit Model: 编辑模型，即调用图 11-25 所示的对话框重新对模型参数进行设定。
- Refit Model: 重新拟合模型，即重新对模型进行参数估计。
- Reevaluate Model: 重新评价模型。
- Delete Model: 删除已有的模型。
- View Forecasts: 查看模型的预测结果。

**STEP 8** 在该菜单中选择“View Model”、“View Parameter Estimates”、“View Statistics of Fit”和“View Forecast”中的任意一项，均会弹出 SAS 系统的模型查看窗口，如图 11-28 所示。

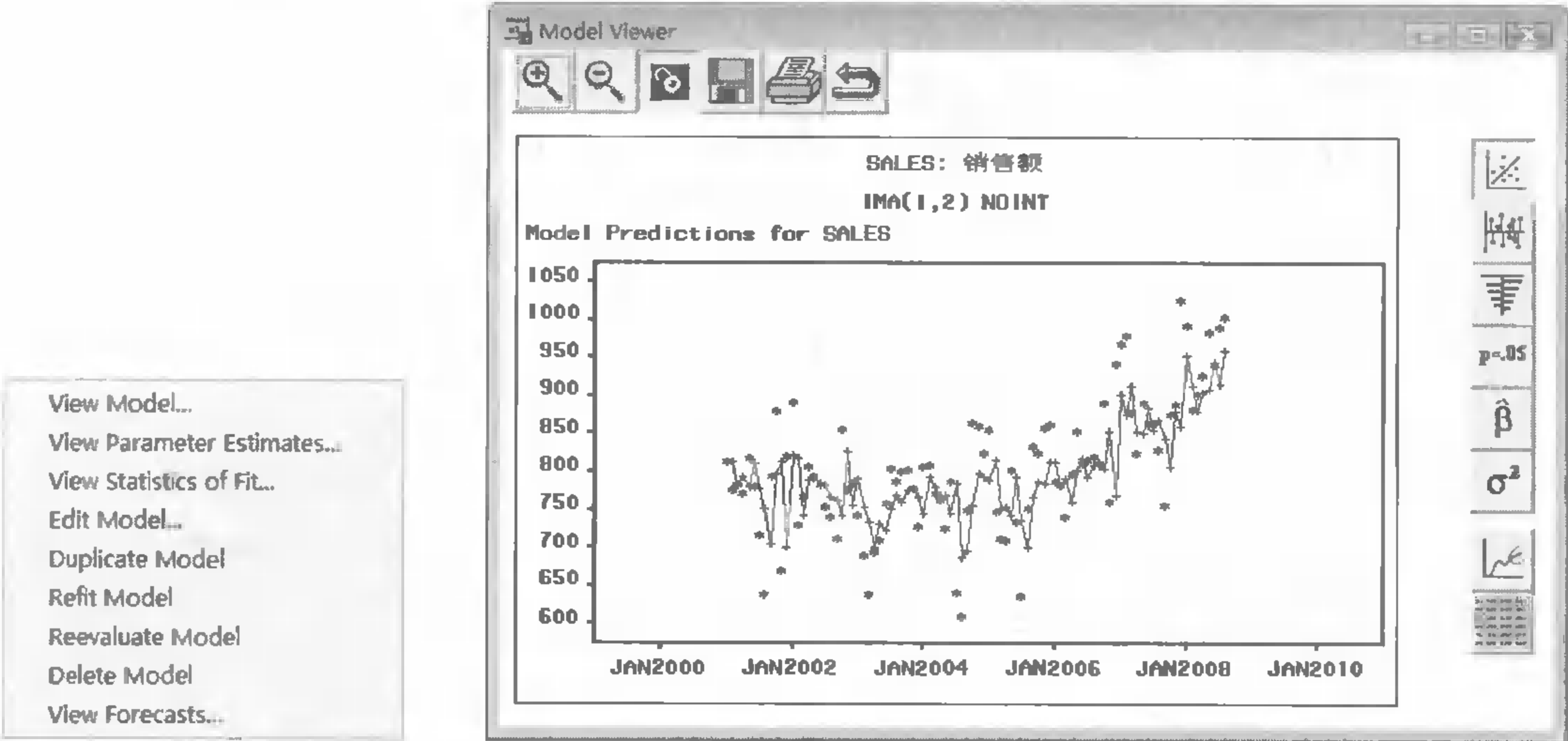












图 11-27 模型查看和调整菜单

图 11-28 “Model Viewer”窗口

在“Model Viewer”窗口中，右边的按钮对应于图 11-27 所示菜单的功能。

- : 查看模型观测值和预测值，对应于“View Model”。
- : 查看模型残差项。
- : 查看残差项的自相关系数图、偏自相关系数图和逆自相关系数图。
- : 查看残差项的白噪声和单位根检验图。
- : 查看模型参数估计结果，对应于“View Parameter Estimates”。
- : 查看拟合统计量，对应于“View Statistics of Fit”。
- : 查看预测预测结果，对应于“View Forecasts”。
- : 查看含有预测值、预测残差和“预测区间”的数据集。

**STEP 9** 在对模型进行参数估计或拟合时，还可以单击图 11-16 中的  “Fit Models Automatically” 按钮进行模型自动拟合。在自动拟合过程中，应当指定系统筛选模型所用的依据。

对于模型的预测，单击图 11-16 中的  “Produce Forecasts” 按钮，可以利用已经经过参数估计的模型对新的数据集进行预测，并可以把指定时间范围的预测结果存储在用户指定的数据集当中。

## 11.3 本章小结

本章主要介绍了时间序列分析中的常用 ARIMA 模型，主要内容简要回顾如下：时间序列是相同事物或现象在不同时刻或时期形成的数据，反映了事物、现象在时间上的发展变化情况；时间序列由长期趋势、季节变动、循环波动和不规则变动 4 个部分组成；对于时间序列的分析和预测，本章主要介绍了 Box-Jenkins 法，即 ARIMA 模型，其基本前提是要把时间序列平稳化，并进行自相关系数图和单位根检验；对平稳时间序列，可以使用图示法、扩展样本自相关函数法、最小信息准则法及典型相关系数平方法等方法进行 ARIMA 模型的识别；在 SAS 系统中，可以使用 ARIMA 过程编程和“Time Series Forecasting System”图形菜单等方式实现 ARIMA 模型识别、参数估计及检验、预测。