



挖掘数据宝藏  
助力决策支持

# SPSS

## 统计分析 从入门到精通

○ 杜强 贾丽艳 编著



人民邮电出版社  
POSTS & TELECOM PRESS

<b>第 1 章 SPSS 15.0 概述</b>	1
1.1 SPSS 简介	1
1.2 SPSS 的安装、启动和退出	3
1.2.1 SPSS 15.0 的安装	3
1.2.2 SPSS 的启动	3
1.2.3 SPSS 15.0 的退出	5
1.3 SPSS 15.0 的界面及设置	5
1.3.1 常用界面	6
1.3.2 通用 General 功能参数	9
1.3.3 Viewer 视图窗口参数	11
1.3.4 Draft Viewer (草稿窗口) 参数	12
1.3.5 Output Labels 输出标签参数	14
1.3.6 Charts 图形参数	14
1.3.7 Interactive 交互图形窗口参数	16
1.3.8 Pivot Table 枢纽表参数	17
1.3.9 Data 数据参数	19
1.3.10 Currency 数值型变量格式参数	20
1.3.11 Scripts 脚本编辑窗口	21
<b>第 2 章 数据文件的建立与操作</b>	22
2.1 数据编辑器与数据文件	22
2.1.1 数据编辑器	22
2.1.2 数据文件	25
2.2 常量、变量、操作符和表达式	25
2.2.1 常量与变量	26
2.2.2 操作符与表达式	30
2.2.3 如何定义一个变量	31
2.2.4 概率事件	35
2.3 输入数据	35
2.3.1 输入数据的方法	36
2.3.2 查看文件信息和变量信息	36
2.4 编辑数据文件	38

2.4.1 在单元格中编辑数据	38
2.4.2 插入变量与删除变量	38
2.4.3 插入观测量与删除观测量	39
2.4.4 数据的剪切、复制和粘贴	39
2.4.5 撤销操作	41
2.5 对数据文件的操作	41
2.5.1 数据文件的打开与保存	41
2.5.2 数据库文件的转换	41
<b>第 3 章 数据文件的操作</b>	50
3.1 数据文件的一般操作	50
3.1.1 数据排序	50
3.1.2 数据文件的分组	51
3.1.3 数据文件的合并	52
3.1.4 数据文件的转置	55
3.1.5 变量取值的求秩	56
3.1.6 变量值的重新编码	58
3.1.7 计算新变量	62
3.2 分类汇总	64
3.2.1 数据描述	65
3.2.2 分类汇总的参数设置	65
3.2.3 分类汇总的结果	67
3.3 观测量的加权	68
3.4 数据文件的结构重组	69
3.4.1 选择数据重组方式	70
3.4.2 变量组到观测量组的重组	71
3.4.3 观测量组到变量组的重组	75
3.4.4 转置重组	78
<b>第 4 章 基本统计分析功能</b>	79
4.1 OLAP 在线分析过程	79
4.1.1 数据描述	79
4.1.2 OLAP 过程的操作和设置	79

4.2 观测的摘要报告分析 .....	83	6.2.2 数据和问题描述 .....	121
4.2.1 观测摘要分析的参数设置 .....	83	6.2.3 卡方检验实例分析 .....	121
4.2.2 输出结果 .....	85	6.3 二项检验 .....	123
4.3 行和列的摘要报告分析 .....	85	6.3.1 原理与方法 .....	123
4.3.1 行形式摘要报告 .....	86	6.3.2 数据和问题描述 .....	123
4.3.2 列形式摘要报告 .....	90	6.3.3 二项检验实例分析 .....	123
4.4 频数分析 .....	93	6.4 游程检验 .....	124
4.4.1 数据描述 .....	93	6.4.1 原理与方法 .....	125
4.4.2 对分类变量的频数分析 .....	93	6.4.2 数据和问题描述 .....	125
4.4.3 对连续变量的频数分析 .....	95	6.4.3 游程检验实例分析 .....	126
4.5 描述性统计分析 .....	97	6.5 Kolmogorov-Smirnov 单样本检验 .....	126
4.5.1 数据描述 .....	97	6.5.1 原理与方法 .....	127
4.5.2 Descriptives 分析 .....	97	6.5.2 数据和问题描述 .....	127
4.6 探索分析过程 .....	98	6.5.3 K-S 单样本检验实例分析 .....	127
4.6.1 数据描述 .....	99	6.6 两独立样本检验 .....	128
4.6.2 Explore 实例分析 .....	99	6.6.1 原理与方法 .....	129
4.7 列联表分析过程 .....	102	6.6.2 数据和问题描述 .....	130
4.7.1 数据描述 .....	103	6.6.3 两独立样本检验实例分析 .....	130
4.7.2 列联表分析的参数设置 .....	103	6.7 $k$ 个独立样本的检验 .....	131
4.7.3 列联表分析的输出结果 .....	106	6.7.1 原理与方法 .....	132
第 5 章 均值比较和 T 检验 .....	108	6.7.2 数据和问题描述 .....	132
5.1 Means 过程 .....	109	6.7.3 $k$ 个独立样本检验实例分析 .....	133
5.1.1 原理与方法 .....	109	6.8 两个相关样本的检验 .....	134
5.1.2 SPSS 实例分析 .....	109	6.8.1 原理与方法 .....	134
5.2 单样本 T 检验 .....	111	6.8.2 数据和问题描述 .....	135
5.2.1 原理与方法 .....	111	6.8.3 两个相关样本检验的实例分析 .....	135
5.2.2 SPSS 实例分析 .....	112	6.9 $k$ 个相关样本的检验 .....	136
5.3 两独立样本 T 检验 .....	113	6.9.1 原理与方法 .....	137
5.3.1 原理与方法 .....	113	6.9.2 数据和问题描述 .....	138
5.3.2 SPSS 实例分析 .....	114	6.9.3 $k$ 个相关样本检验的实例分析 .....	138
5.4 配对样本 T 检验 .....	115	第 7 章 多重响应分析 .....	140
5.4.1 原理与方法 .....	115	7.1 多重响应概述 .....	140
5.4.2 SPSS 实例分析 .....	116	7.2 多重响应变量集的定义 .....	140
第 6 章 非参数检验 .....	118	7.2.1 定义多重响应变量集的实例 .....	141
6.1 非参数检验的简介 .....	118	7.3 多重响应变量集的频数分析 .....	142
6.1.1 非参数检验与参数检验 .....	118	7.3.1 多重响应变量频数分析的实例 .....	142
6.1.2 非参数检验的优点 .....	119	7.4 多重响应变量集的交叉表分析 .....	144
6.1.3 非参数检验的缺点 .....	119	7.4.1 多重响应变量交叉表分析的实例 .....	144
6.2 卡方检验 .....	119	7.5 使用 Tables 过程研究多重响应变量集 .....	146
6.2.1 原理与方法 .....	120		

7.5.1 多重响应变量集的定义 .....	146	8.8.1 加权回归分析简介 .....	202
7.5.2 用 Tables 过程建立包含多重 响应变量集的表格 .....	147	8.8.2 问题描述和数据准备 .....	203
<b>第 8 章 回归分析</b> .....	150	8.8.3 加权回归的参数设置 .....	203
8.1 线性回归 .....	150	8.8.4 案例的结果分析 .....	204
8.1.1 一元线性回归的基本原理 .....	150	8.9 二阶段最小二乘回归 .....	205
8.1.2 多元线性回归的基本原理 .....	152	8.9.1 二阶段最小二乘回归的 基本原理 .....	206
8.1.3 模型假设的其他检验 .....	153	8.9.2 问题描述和数据准备 .....	206
8.1.4 问题描述和数据准备 .....	154	8.9.3 二阶段最小二乘回归的 参数设置 .....	207
8.1.5 线性回归分析的设置和操作 .....	154	8.9.4 案例的结果分析 .....	208
8.1.6 案例的结果分析 .....	159	8.10 最优尺度回归 .....	209
8.2 曲线回归 .....	162	8.10.1 最优尺度回归原理 .....	209
8.2.1 曲线回归的基本原理 .....	162	8.10.2 问题描述和数据准备 .....	209
8.2.2 问题描述和数据准备 .....	163	8.10.3 最优尺度回归的参数设置 .....	210
8.2.3 曲线回归分析的设置和操作 .....	163	8.10.4 案例的结果分析 .....	214
8.2.4 案例的结果分析 .....	165	<b>第 9 章 方差分析</b> .....	217
8.3 非线性回归 .....	166	9.1 方差分析简介 .....	217
8.3.1 非线性回归简介 .....	167	9.1.1 $t$ 检验与方差分析的比较 .....	217
8.3.2 问题描述和数据准备 .....	168	9.1.2 方差分析的基本原理 .....	218
8.3.3 非线性回归的参数设置 .....	169	9.2 单因素方差分析 .....	220
8.3.4 案例的结果分析 .....	173	9.2.1 原理与方法 .....	220
8.4 二元 Logistic 回归 .....	173	9.2.2 单因素方差分析实例 .....	220
8.4.1 二元 Logistic 回归的数学原理 .....	174	9.3 多因素方差分析过程 .....	225
8.4.2 问题描述和数据准备 .....	175	9.3.1 原理与方法 .....	225
8.4.3 二元 Logistic 回归的参数设置 .....	176	9.3.2 二因素方差分析实例 .....	228
8.4.4 案例的结果分析 .....	180	9.3.3 协方差分析实例 .....	236
8.5 多元 Logistic 回归分析 .....	184	9.3.4 交互效应中随机因素的分析 .....	238
8.5.1 多元 Logistic 回归的原理简介 .....	184	9.4 多元方差分析 .....	242
8.5.2 问题描述和数据准备 .....	184	9.4.1 原理与方法 .....	242
8.5.3 多元 Logistic 回归参数设置 .....	185	9.4.2 多元方差分析实例 .....	243
8.5.4 案例的结果分析 .....	189	9.5 重复测量设计的方差分析 .....	244
8.6 Ordinal 回归 .....	191	9.5.1 原理与方法 .....	244
8.6.1 问题描述和数据准备 .....	192	9.5.2 SPSS 实例分析 .....	245
8.6.2 Ordinal 回归的参数设置 .....	192	9.6 方差成分分析 .....	250
8.6.3 案例的结果分析 .....	196	9.6.1 原理简介 .....	250
8.7 概率单位回归分析 .....	197	9.6.2 SPSS 实例分析 .....	250
8.7.1 概率单位回归分析简介 .....	198	9.7 正交实验设计 .....	253
8.7.2 问题描述和数据准备 .....	198	9.7.1 正交实验设计简述 .....	253
8.7.3 概率单位回归的参数设置 .....	199	9.7.2 SPSS 实例分析 .....	254
8.7.4 案例的结果分析 .....	200	9.7.3 正交实验设计的方差分析 .....	256
8.8 加权回归分析 .....	202		



<b>第 10 章 相关分析</b> .....	257	12.3.4 案例的结果分析	295
10.1 相关分析的基本概念	257	12.3.5 对聚类结果的进一步分析	296
10.1.1 相关分析的特点和应用	257	12.4 两阶段聚类分析	298
10.1.2 相关系数的计算	258	12.4.1 两阶段聚类简介	298
10.1.3 SPSS 提供的相关分析功能	259	12.4.2 问题描述和数据准备	299
10.2 两变量相关分析	260	12.4.3 SPSS 两阶段聚类的设置	299
10.2.1 问题描述和数据准备	260	12.4.4 案例的结果分析	304
10.2.2 相关分析的参数设置	260	12.5 一般判别分析	307
10.2.3 案例的结果分析	261	12.5.1 判别分析的基本原理	307
10.3 偏相关分析	262	12.5.2 问题描述和数据准备	308
10.3.1 偏相关分析的基本原理	262	12.5.3 判别分析的参数设置	309
10.3.2 偏相关分析实例	263	12.5.4 案例的结果分析	312
10.4 距离分析	264	12.6 逐步判别分析实例	315
10.4.1 距离分析的基本概念	265	12.6.1 问题描述和数据准备	315
10.4.2 距离分析的参数设置	265	12.6.2 逐步判别的参数设置	316
10.4.3 距离分析实例	269	12.6.3 案例的结果分析	318
<b>第 11 章 因子分析</b> .....	271	12.7 决策树分析	321
11.1 因子分析的原理简介	271	12.7.1 决策树分类的基本原理	321
11.1.1 因子分析的基本思想	271	12.7.2 决策树过程的参数设置	323
11.1.2 因子分析和主成分 分析的联系	272	12.7.3 问题描述和数据准备	338
11.1.3 因子分析的基本步骤	272	12.7.4 案例分析	338
11.2 SPSS 因子分析的应用实例	273	<b>第 13 章 生存分析</b> .....	344
11.2.1 数据描述	273	13.1 生存分析简介	344
11.2.2 SPSS 因子分析过程的设置	274	13.1.1 生存分析的基本概念	344
11.2.3 结果分析	278	13.1.2 生存分析的数据特点	346
<b>第 12 章 分类分析</b> .....	283	13.1.3 生存分析的常用方法	346
12.1 聚类分析的原理简介	283	13.1.4 SPSS 中的生存分析过程	346
12.1.1 聚类分析的基本概念	283	13.2 生命表分析	347
12.1.2 聚类分析的一般原理	284	13.2.1 生命表分析简介	347
12.2 快速样本聚类过程	286	13.2.2 生命表分析的基本步骤	347
12.2.1 快速聚类简介	286	13.2.3 生命表实例分析	348
12.2.2 问题描述和数据准备	286	13.3 Kaplan-Meier 分析	352
12.2.3 SPSS 快速聚类的设置	287	13.3.1 Kaplan-Meier 分析的步骤	352
12.2.4 案例的结果分析	289	13.3.2 生存曲线的比较和检验	352
12.3 分层聚类	291	13.3.3 Kaplan-Meier 分析的实例	353
12.3.1 分层聚类简介	291	13.4 Cox 回归模型	357
12.3.2 问题描述和数据准备	291	13.4.1 Cox 回归模型的原理简介	357
12.3.3 SPSS 分层聚类的设置	292	13.4.2 Cox 回归实例分析	358
		<b>第 14 章 信度分析</b> .....	366
		14.1 信度分析	366

14.1.1 信度分析的基本原理 .....	366	16.3.1 Logit 过程概述 .....	412
14.1.2 问题描述和数据准备 .....	368	16.3.2 问题描述和数据准备 .....	412
14.1.3 信度分析的参数设置 .....	368	16.3.3 Logit 过程的参数设置 .....	412
14.1.4 案例的结果分析 .....	370	16.3.4 案例的结果分析 .....	413
14.2 多维尺度分析 .....	371	16.4 Model Selection 过程 .....	415
14.2.1 多维尺度分析简介 .....	371	16.4.1 Model Selection 过程概述 .....	415
14.2.2 问题描述和数据准备 .....	371	16.4.2 问题描述和数据准备 .....	415
14.2.3 ALSCAL 过程的参数设置 .....	371	16.4.3 层次对数线性模型的操作 过程 .....	416
14.2.4 案例的结果分析 .....	374	16.4.4 案例的结果分析 .....	417
<b>第 15 章 时间序列分析 .....</b>	<b>377</b>	<b>第 17 章 对应分析 .....</b>	<b>419</b>
15.1 SPSS15 的时间序列分析概览 .....	377	17.1 对应分析的基本原理 .....	419
15.1.1 Create Models 的通用 设置选项 .....	378	17.1.1 对应分析与因子分析 .....	419
15.1.2 Apply Models 的通用 设置选项 .....	384	17.1.2 SPSS 中的对应分析 .....	420
15.2 时间序列数据的预分析 .....	384	17.1.3 使用对应分析的注意事项 .....	420
15.2.1 缺失值替换 .....	385	17.2 简单对应分析 .....	421
15.2.2 定义时间变量 .....	385	17.2.1 简单对应分析的数学原理 .....	421
15.2.3 时间序列的平稳化 .....	386	17.2.2 SPSS 简单对应分析实例 .....	422
15.3 指数平滑模型 .....	388	17.3 多元对应分析 .....	427
15.3.1 指数平滑的基本原理 .....	389	17.3.1 多元对应分析基本概念及 其特点 .....	428
15.3.2 指数平滑模型的参数设置 .....	389	17.3.2 多元对应分析的参数设置 .....	428
15.3.3 指数平滑模型实例分析 .....	391	17.3.3 实例的结果分析 .....	435
15.4 ARIMA 模型 .....	395	<b>第 18 章 缺失值分析 .....</b>	<b>438</b>
15.4.1 ARIMA 模型的基本原理 .....	395	18.1 缺失值分析的概念 .....	438
15.4.2 ARIMA 模型的参数设置 .....	397	18.1.1 缺失值的表现方式 .....	438
15.4.3 ARIMA 模型实例分析 .....	398	18.1.2 SPSS 中的缺失值处理方法 .....	439
15.5 季节分解模型 .....	401	18.2 缺失值分析的参数设置 .....	440
15.5.1 季节分解法概述 .....	401	18.3 缺失值分析的实例 .....	444
15.5.2 季节分解模型实例分析 .....	402	<b>第 19 章 统计图形 .....</b>	<b>449</b>
<b>第 16 章 对数线性模型 .....</b>	<b>406</b>	19.1 概述 .....	449
16.1 对数线性模型概述 .....	406	19.1.1 数据和变量的准备 .....	449
16.1.1 简单列联表分析的不足 .....	406	19.1.2 图形构建器的基本操作 .....	451
16.1.2 对数线性模型的基本形式 .....	406	19.1.3 交互式作图和对话框作图 .....	452
16.2 General 过程 .....	407	19.1.4 图形的编辑 .....	453
16.2.1 General 过程概述 .....	407	19.2 条形图 .....	453
16.2.2 问题描述和数据准备 .....	408	19.2.1 数据和问题描述 .....	453
16.2.3 General 过程的参数设置 .....	408	19.2.2 用图形构建器作条形图 .....	453
16.2.4 案例的结果分析 .....	410	19.2.3 交互式条形图 .....	457
16.3 Logit 过程 .....	411		

19.2.4	用对话框创建条形图	459
19.3	线形图	460
19.3.1	数据和问题描述	461
19.3.2	用图形构建器作线形图	461
19.3.3	交互式线形图	462
19.3.4	用对话框创建线形图	464
19.4	面积图	465
19.4.1	数据和问题描述	465
19.4.2	用图形构建器作面积图	465
19.4.3	交互式面积图	467
19.4.4	用对话框创建面积图	467
19.5	饼图	468
19.5.1	数据和问题描述	468
19.5.2	用图形构建器作饼图	468
19.5.3	交互式饼图	469
19.5.4	用对话框创建饼图	470
19.6	高低图	470
19.6.1	数据和问题描述	471
19.6.2	用图形构建器作高低图	471
19.6.3	交互式高低图	472
19.6.4	用对话框创建高低图	473
19.7	帕累托图	477
19.7.1	数据和问题描述	478
19.7.2	用对话框创建帕累托图	478
19.8	控制图	479
19.8.1	数据和问题描述	480
19.8.2	用对话框创建控制图	480
19.9	箱图	486
19.9.1	数据和问题描述	486
19.9.2	用图形构建器作箱图	486
19.9.3	交互式箱图	488
19.9.4	用对话框创建箱图	489
19.10	误差条图	490
19.10.1	数据和问题描述	490
19.10.2	交互式误差条图	490
19.10.3	用对话框创建误差条图	491
19.11	散点图	492
19.11.1	数据和问题描述	492
19.11.2	用图形构建器作高低图	493
19.11.3	交互式散点图	495
19.11.4	用对话框创建散点图	498
19.12	直方图	498

19.12.1	数据和问题描述	499
19.12.2	用图形构建器作直方图	499
19.13	P-P 概率图	500
19.13.1	数据和问题描述	500
19.13.2	用对话框创建帕 P-P 概率图	501
19.14	Q-Q 概率图	502
19.14.1	数据和问题描述	503
19.14.2	用对话框创建 Q-Q 概率图	503
19.15	时间序列图	504
19.15.1	普通序列图	504
19.15.2	自相关序列图	507
19.15.3	互相关序列图	509
19.16	双轴线图	511
19.16.1	数据和问题描述	511
19.16.2	用图形构建器作双轴线图	511

## 第 20 章 上市公司财务危机预警分析

20.1	财务危机预警的应用简介	513
20.1.1	财务危机的定量定义方法	513
20.1.2	财务危机预警的模型选择	514
20.2	数据描述	514
20.2.1	数据说明	514
20.2.2	指标选择	515
20.2.3	补充说明	515
20.3	分析方法概述	516
20.3.1	判别分析	516
20.3.2	logistic 回归方法	516
20.4	SPSS 建模过程和结论分析	517
20.4.1	SPSS 数据筛选操作	517
20.4.2	SPSS 判别分析建模与分析	521
20.4.3	logistic 回归建模与分析	525
20.5	进一步的分析与应用	528
20.5.1	分类结果的应用分析	528
20.5.2	建模方法的改进	529
20.6	建议和推广	529
20.6.1	时间序列研究	529
20.6.2	数据的有效预警期	529
20.6.3	指标的简化方法	529

## 第 21 章 影响汇率的因素分析

21.1	汇率影响因素的简介	531
21.2	数据描述	532

21.3 分析方法概述 .....	533	23.1.2 指标选取 .....	557
21.3.1 探索性分析 .....	533	23.1.3 数据格式 .....	557
21.3.2 多元回归分析 .....	534	23.2 聚类分析法简述 .....	557
21.4 SPSS 建模过程和结论分析 .....	534	23.3 SPSS 建模过程和结论分析 .....	558
21.4.1 数据准备 .....	534	23.3.1 对专科院校进行聚类的 设置操作 .....	558
21.4.2 探索性分析 .....	535	23.3.2 对本科院校的分析 .....	561
21.4.3 多元回归分析 .....	536	23.4 建议和推广 .....	563
21.5 进一步的分析与应用 .....	539	第 24 章 试卷信度的检验与分析 .....	565
21.5.1 剔除存在共线性的外汇 储备变量 .....	540	24.1 试卷信度检验的背景简介 .....	565
21.5.2 回归模型的进一步改进 .....	540	24.1.1 测验内容的自身方面 .....	565
21.5.3 两个回归模型的比较 .....	541	24.1.2 施测过程 .....	565
21.6 建议和推广 .....	542	24.1.3 被测试者的自身因素 .....	566
21.6.1 时间序列研究 .....	542	24.2 数据描述 .....	566
21.6.2 汇率影响因素的定性分析 .....	542	24.3 分析方法概述 .....	566
第 22 章 因子分析在成绩综合评价 中的应用 .....	543	24.3.1 试卷信度的基本计算公式 .....	566
22.1 学生成绩的综合评价简介 .....	543	24.3.2 试卷信度的估计方法 .....	567
22.2 数据描述 .....	544	24.4 SPSS 建模过程和结论分析 .....	568
22.3 分析方法概述 .....	544	24.4.1 SPSS 信度分析的参数设置 .....	568
22.3.1 应用因子分析进行成绩 综合评价的步骤 .....	544	24.4.2 结果分析 .....	569
22.3.2 应用因子分法进行成绩 综合评价的注意事项 .....	545	24.5 建议和推广 .....	570
22.4 SPSS 建模过程和结论分析 .....	546	第 25 章 多因素试验的设计与分析 .....	571
22.4.1 数据准备 .....	546	25.1 试验设计简介 .....	571
22.4.2 SPSS 因子分析建模与分析 .....	548	25.1.1 试验设计的应用 .....	571
22.5 进一步的分析与应用 .....	553	25.1.2 试验设计问题的解决步骤 .....	572
22.6 建议和推广 .....	553	25.2 数据描述 .....	573
22.6.1 高中生的成绩综合评价 .....	553	25.3 分析方法概述 .....	573
22.6.2 对缺失数据的处理 .....	554	25.3.1 正交设计方法 .....	573
22.6.3 多种方法结合的综合 评价模型 .....	554	25.3.2 综合评分方法 .....	575
第 23 章 高等教育办学条件的聚类分析 .....	555	25.4 SPSS 建模过程和结论分析 .....	575
23.1 数据描述 .....	555	25.4.1 数据标准化 .....	576
23.1.1 关于基本办学条件指标 合格与否的判定 .....	555	25.4.2 性能指标权重的确定 .....	577
		25.4.3 利用权重求综合指标 .....	578
		25.4.4 对综合得分的进一步分析 .....	578
		25.5 建议和推广 .....	579



# 第 1 章 SPSS 15.0 概述

SPSS 通过简单的菜单式操作,就可以方便地规范和融合搜集到的原始数据,并能实施从简单的描述性统计分析到复杂的时序分析等多种方法,对数据进行建模,返回有意义的分析结果,比如客户特征的分类、发展趋势的预测等。把这些结果应用于实际,可以帮助读者在发掘潜在客户、制定长远规划等工作上做出更加准确的判断。

本章详细介绍 SPSS 15.0 软件环境的设置内容和设置方式,打造适合自己的 SPSS 15.0 工作环境。

## 1.1 SPSS 简介

业界领先的统计分析软件提供商 SPSS 公司,推出的旗舰统计分析软件 SPSS 15.0 除了继承原有产品的特点之外,还增加了许多显著的新特点。这些增强特性包括:SPSS 15.0 提供了更强大的数据管理功能,能帮助用户更加方便地使用其他应用程序和数据库;新式的图表能够让用户将复杂的信息清晰地表现出来,SPSS 15.0 进一步增强了图形构建器的功能——高度可视化的图表创建界面;SPSS 15.0 还包括了次序回归(Ordinal Regression)分析算法,对两种以上变量的次序输出进行预测,例如预测客户忠诚度及其与客户满意度的相关性。另外,在统计模型与算法、可编程性等方面,SPSS 15.0 都进行了功能更完善、使用更便利等方面的改进。

### 1. SPSS 的特点

(1) 界面友好,操作简单。SPSS 的命令语句、子命令及各种选项绝大部分都包含在各种菜单和对话框中,因此用户无须花大量时间记忆繁杂的命令、过程和选项。在 SPSS 中,大多数操作可以通过菜单和对话框来完成,因此操作便捷,易于学习和使用。

(2) 适用性好,因人而异。虽然大部分统计分析方法可以通过菜单和对话框来完成,但是,对于熟悉 SPSS 语言的用户,也可以在语句窗口中直接编写程序语句,从而灵活地完成各种复杂的统计分析任务。另外,用对话框指定命令、子命令和选项之后,通过单击界面上的 Paste 按钮,可把当前对话框设置对应的语句,自动粘贴到命令语句窗口中,并允许保存为文件。因此 SPSS 能既适用于新用户也适用于老用户。

(3) 算法隐藏。具有第四代语言的特点,只需通过菜单的选择以及对话框的操作告诉系统要做什么,无须告诉系统怎样做。用户只需了解统计分析原理,无须通晓统计分析的各种算法,即可得到统计分析结果。



(4) 接口完善。具有完善的数据转换接口, 其他软件生成的数据文件 (例如 Excel 文件、Access 文件、关系数据库生成的 DBF 文件、文本编辑软件生成的 ASCII 码数据文件等) 均可方便地转换成可供 SPSS 分析的数据文件。

SPSS 支持 OLE 技术和 ActiveX 技术, 生成的表格和交互图对象可以与同类软件进行自动嵌入与链接。它还有内置的 VBA 客户语言 (Sax Basic), 能进行编程。

(5) 功能强大。SPSS 的核心部分是统计功能, 可以完成数理统计分析任务, 提供了从简单的单变量分析到复杂的多变量分析的多种方法。既包括常规的相关分析、回归分析、方差分析、卡方检验、t 检验和非参数检验, 也包括多元回归分析、聚类分析、判别分析、主成分分析和因子分析, 还包括时间序列分析、生存分析和可靠性分析等。

(6) 表格和图形化功能。SPSS 可以直接生成数十种风格的表格 (OLAP cubes), 伴随其他分析过程又可生成一般表、多响应表和频数表等表格。利用专门的编辑窗口或结果查看窗口, 能编辑所生成的表格, 如表 1-1 所示。

表 1-1

SPSS 生成表格

OLAP Cubes						
时段: 6						
	名 称					
	四 川 长 虹		招 商 银 行		总 计	
	收盘价	交易量	收盘价	交易量	收盘价	交易量
合计	26.23	1 586 805	71	3 999 344	97.23	5 586 149
N	4	4	4	4	8	8
均值	6.557 5	396 701.3	17.75	999 836	12.153 8	698 268.6
标准差	0.085 61	13 123.4	0.5	556 136.7	5.991 83	493 885.5
总和的%	0.04	0.038	0.109	0.097	0.149	0.135
合计 N 的%	0.069	0.069	0.069	0.069	0.138	0.138

SPSS 拥有强大的图形功能, 能生成数十种基本图和交互图。基本图包括条形图、线图、面积图、饼图、高低图、帕累托图、控制图、箱图、误差条图、散点图、直方图、P-P 概率图、Q-Q 概率图和时间序列图等。交互图比基本图漂亮, 有二维和三维形式, 包括条形交互图、点形交互图、线形交互图、带形交互图、饼形交互图、箱形交互图、误差条形交互图、直方交互图和散点交互图等。同表格一样, 图形生成以后, 也可以进行编辑。SPSS 的输出图形可以保存为多种格式。

通过直观、漂亮的统计图形, 能形象地显示分析结果, 如图 1-1 所示。

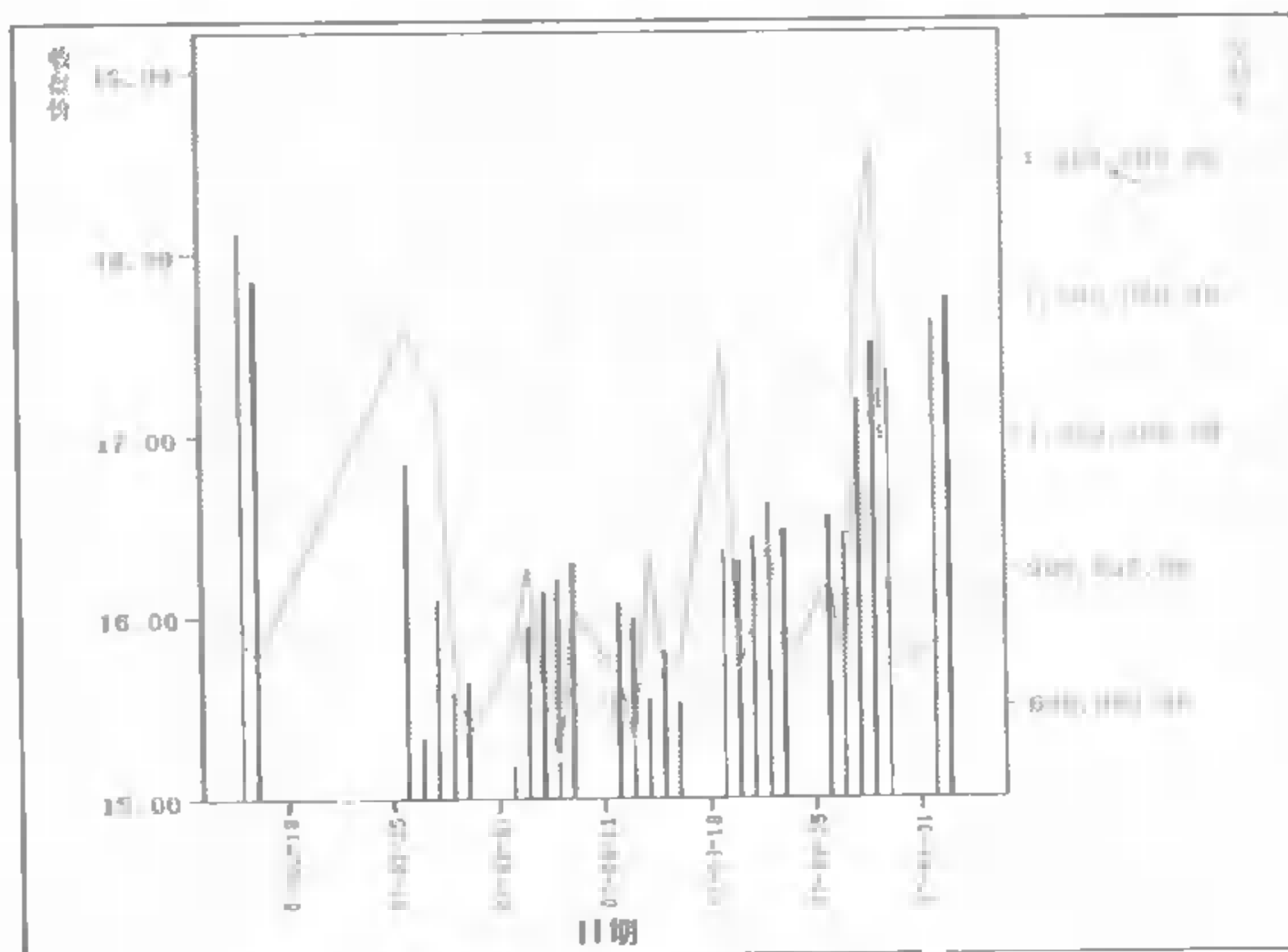


图 1-1 SPSS 生成图形

## 2. SPSS 15.0 的不同版本

对普通用户而言,可能会用到两个版本:个人版 SPSS 15.0 和服务器版 SPSS Server 15.0。

(1) 使用个人版 SPSS,能方便的实现数据过滤、数据筛选、统计分析和特定结果输出等功能,不仅帮助用户挖掘到隐含在大量数据背后的知识,而且能节省用户宝贵的时间。现在的个人电脑一般都能支持 SPSS 15.0 的运行,可适用操作系统为 Microsoft Windows XP 或 Windows 2000。

(2) 把 SPSS Server 15.0 安装在服务器上,把个人版 SPSS 15.0 安装在客户端上,通过分布模式连接服务器,这样用户在客户端就能直接存取服务器上的数据,并能在服务器上运行分析过程。SPSS Server 15.0 配有专属的结果发布系统,能在更广的范围以更方便的方式分享分析结果。

## 1.2 SPSS 的安装、启动和退出

本节介绍 SPSS 安装和启动操作,从这里我们开始踏上 SPSS 应用旅程。

### 1.2.1 SPSS 15.0 的安装

装入 SPSS 15.0 的安装盘,自动运行后弹出如图 1-2 所示的选择界面,单击 Install SPSS (若安装其他组件可单击相应项目),弹出如图 1-3 所示的界面,提示检查当前的系统环境。若光盘内容未自动运行,可右键单击光驱盘符,在弹出的菜单中选择“自动播放”,或者直接双击光驱盘符运行。

待图 1-3 所示的界面自动完成后,弹出如图 1-4 所示的界面,一直单击 Next 按钮,直到完成安装。建议在安装的过程中不要运行其他程序。

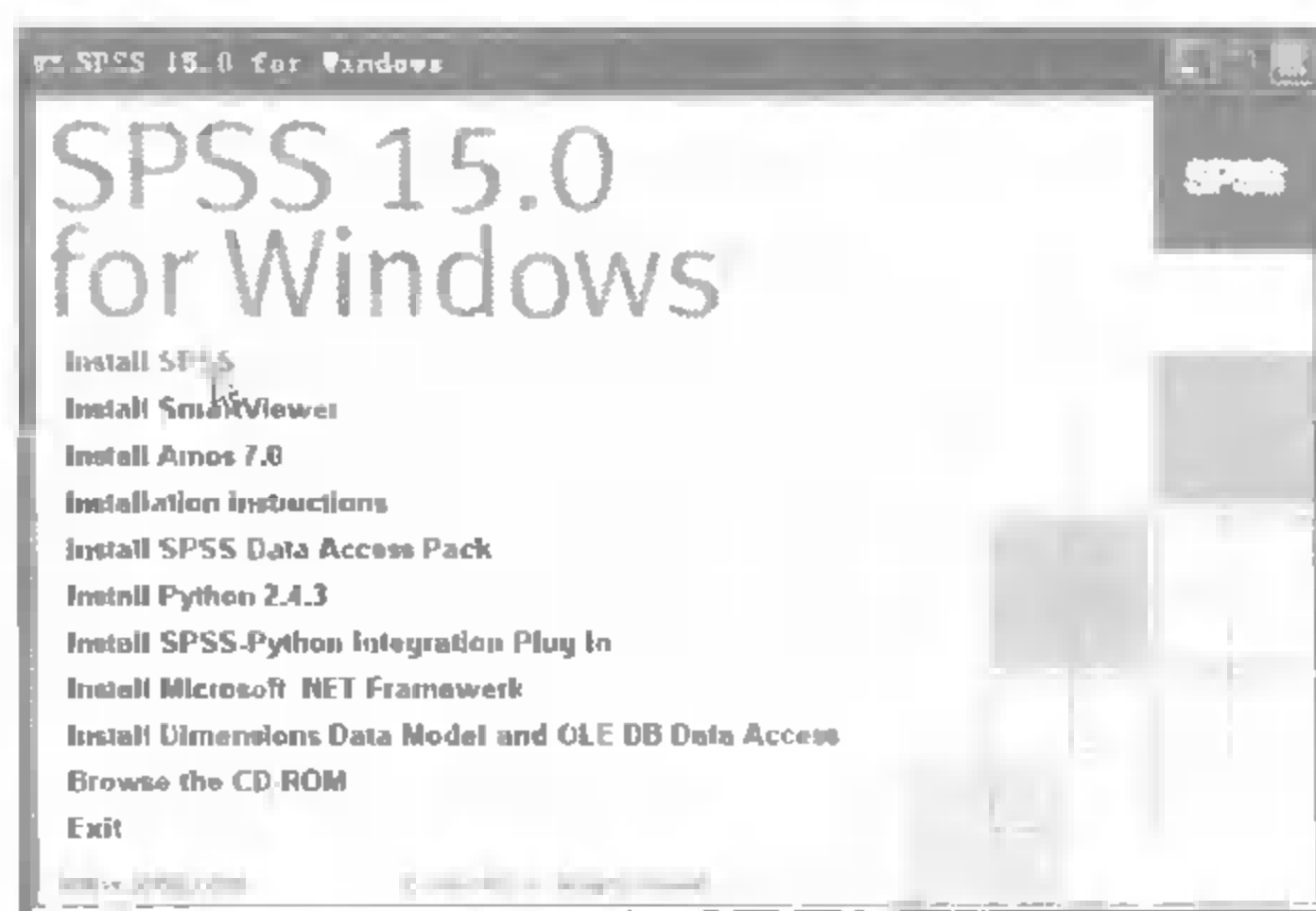


图 1-2 SPSS 安装 1

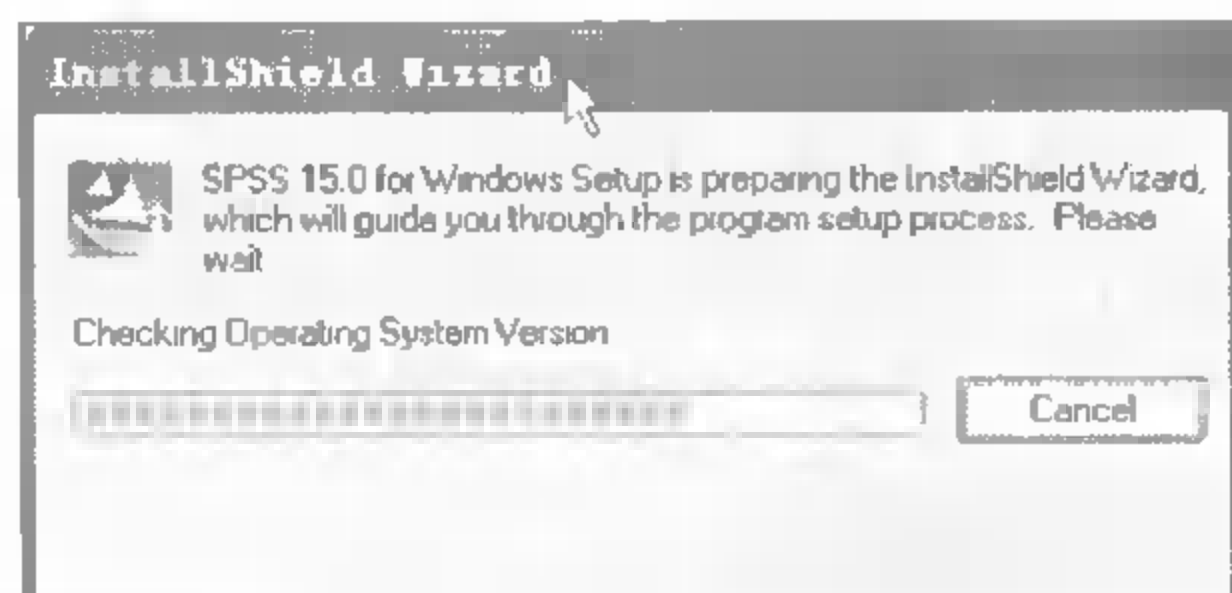


图 1-3 SPSS 安装 2

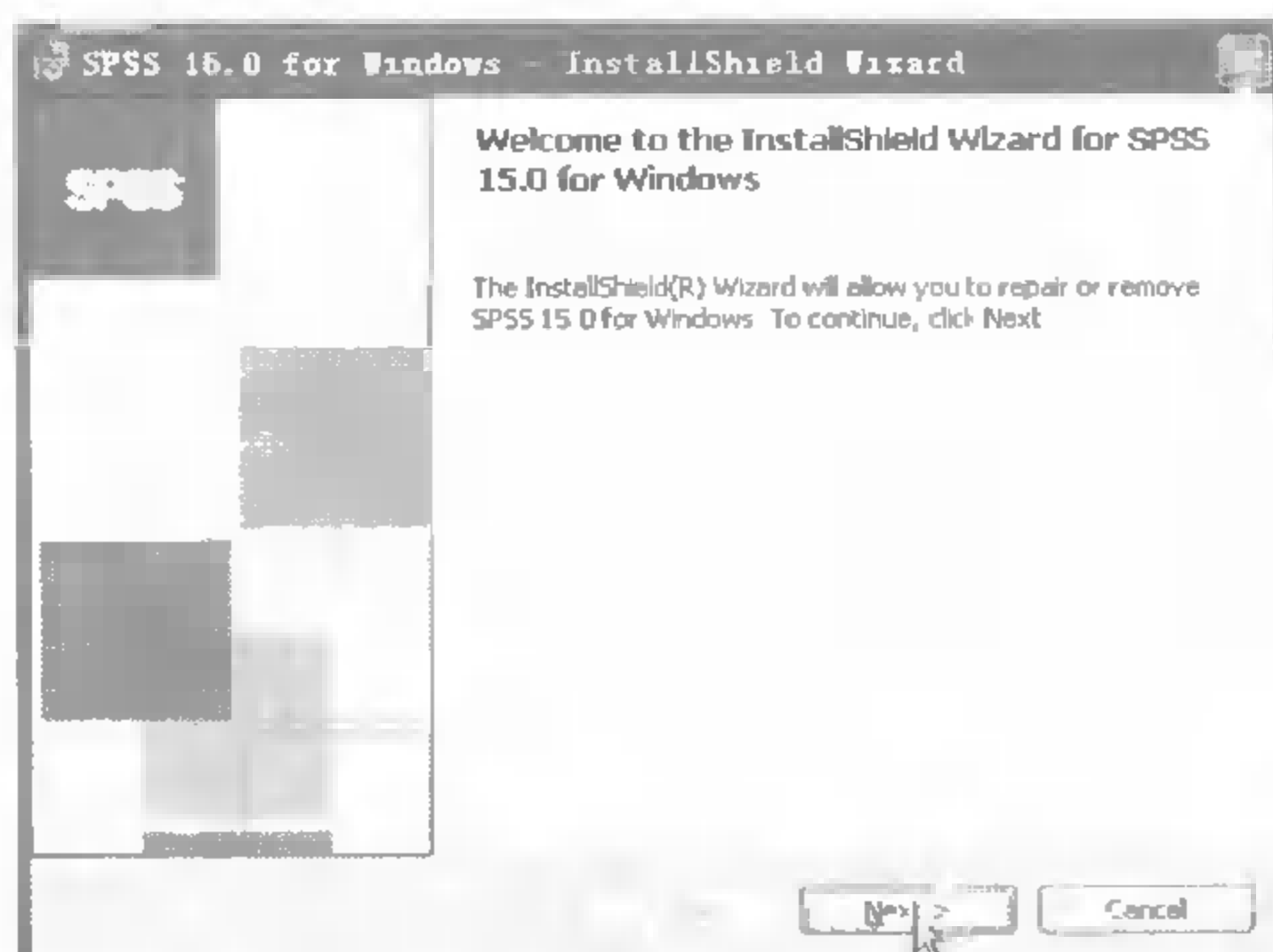


图 1-4 SPSS 安装 3

### 1.2.2 SPSS 的启动

本书主要介绍 SPSS 的窗口菜单运行方式,通过选择窗口、菜单与对话框来完成各种分析过程。

#### 1. 启动

启动 SPSS 程序,可以双击如图 1-5 所示的桌面图标,也可以在“开始”菜单中依次单

击“程序→SPSS for Windows→SPSS15.0 for Windows”。启动后，会出现图 1-6 所示的启动界面，显示本软件的版本和注册信息，如果没有出现界面底部的注册信息，需要按后面介绍的方法进行注册，或者使用试用版本。



图 1-5 SPSS 的桌面图标

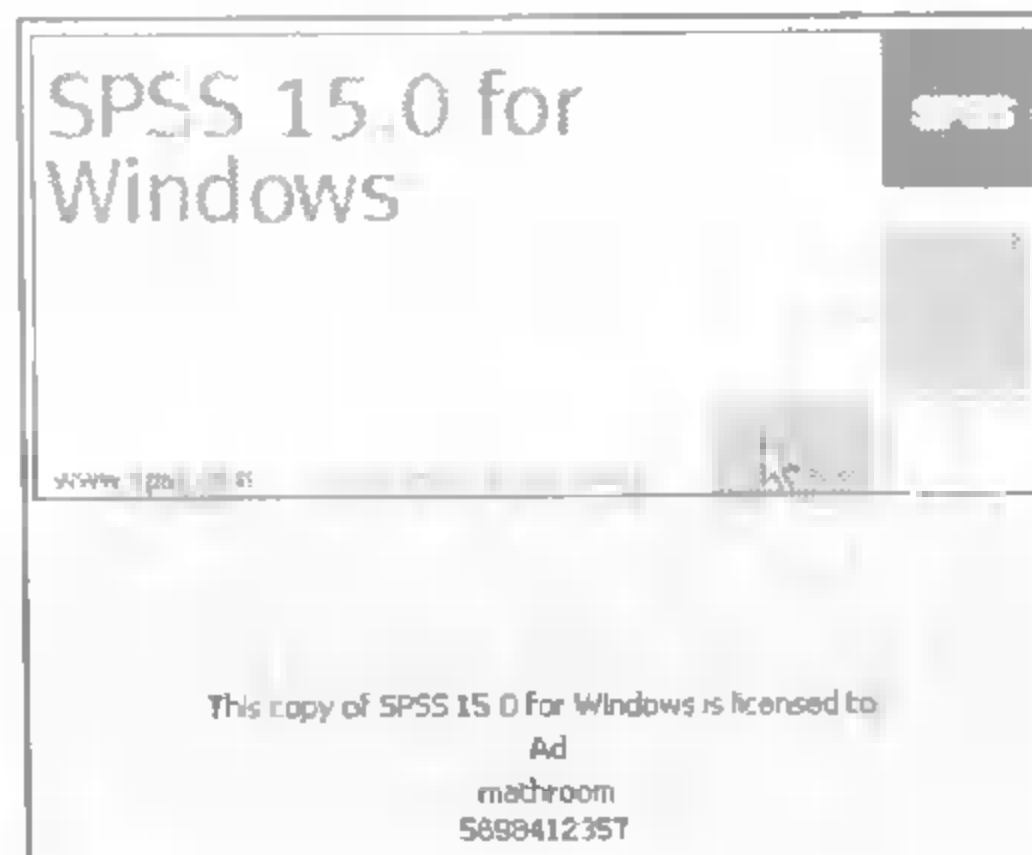


图 1-6 SPSS 的启动界面

之后会出现如图 1-7 所示的选择界面，表明 SPSS 已正常启动。单击选中“Run the tutorial”单选框，再单击“OK”按钮，会打开如图 1-8 所示的图形教程界面，它是一个学习 SPSS 操作的捷径。

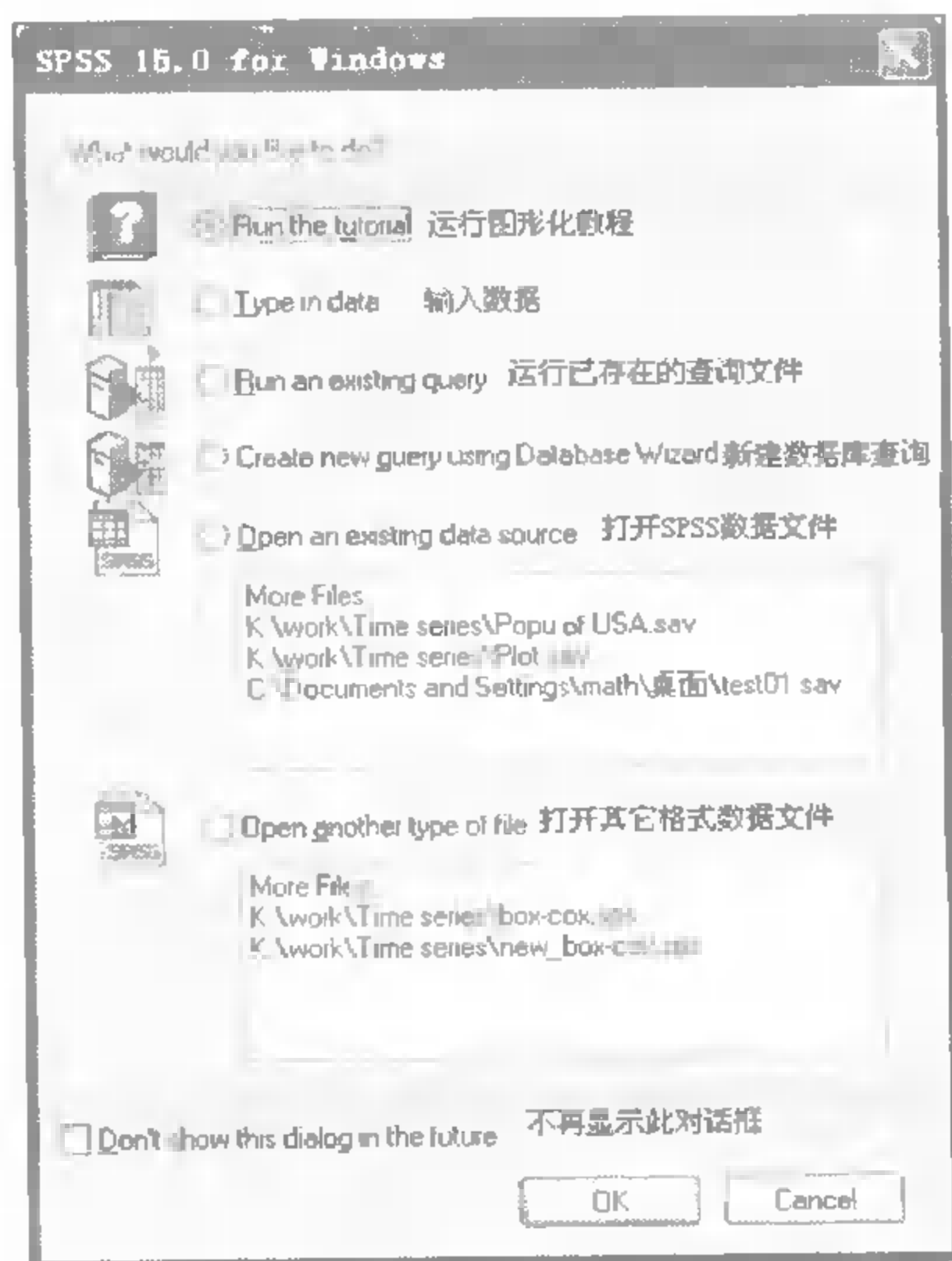


图 1-7 SPSS 启动选项

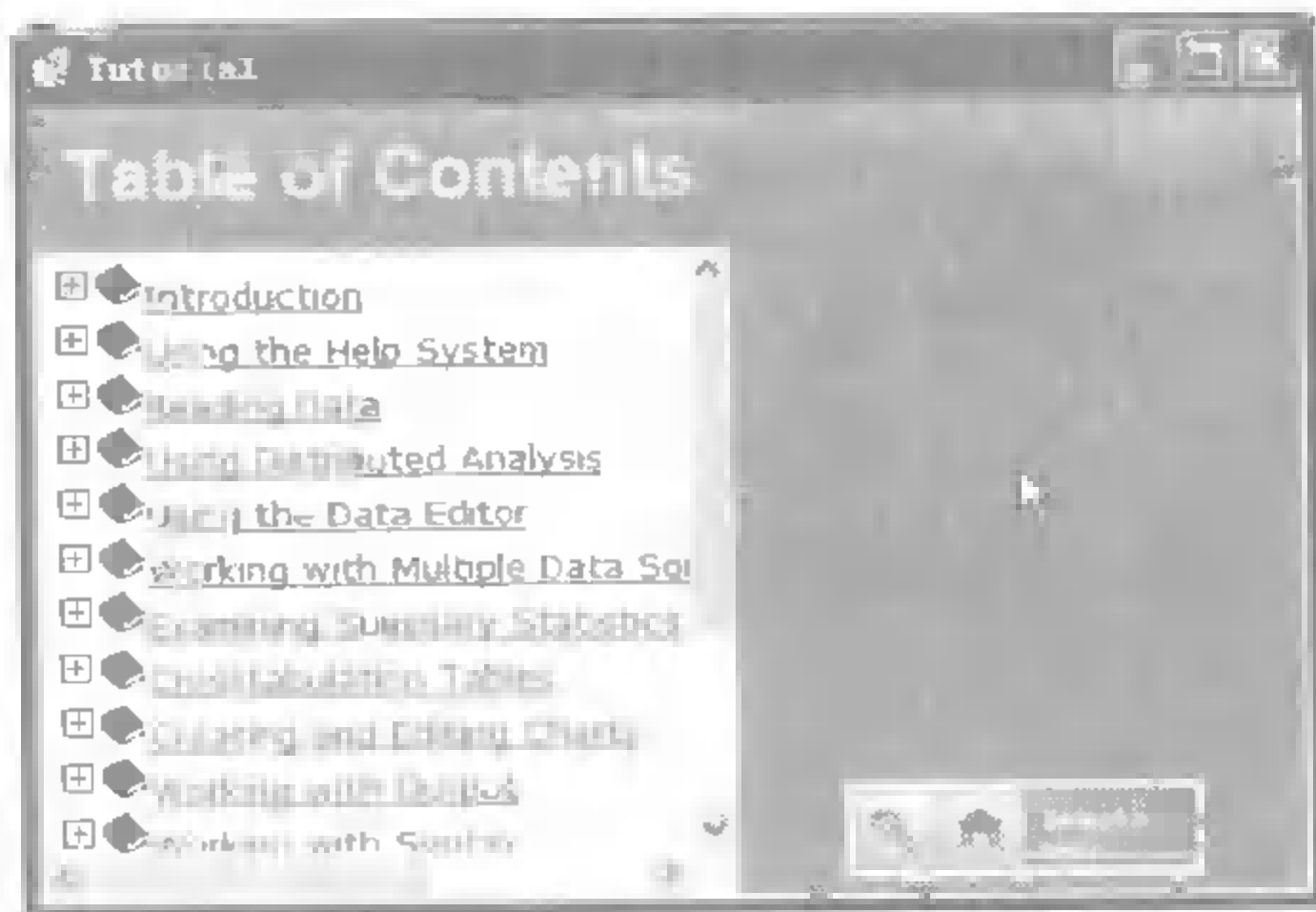


图 1-8 图形教程

## 2. SPSS 的注册方法

在桌面的“开始”菜单中单击如图 1-9 所示的 Wizard 项目，进入如图 1-10 所示的 SPSS 注册界面，单击“Start”后开始注册。



图 1-9 SPSS 注册 1

## 3. 其他 3 种运行方式

(1) 批处理方式。在桌面的“开始”菜单中依次单击“程序→SPSS for Windows→SPSS 15.0

Production Mode Facility”，打开如图 1-11 所示的操作界面，它用来运行指定的 Syntax 程序文件。

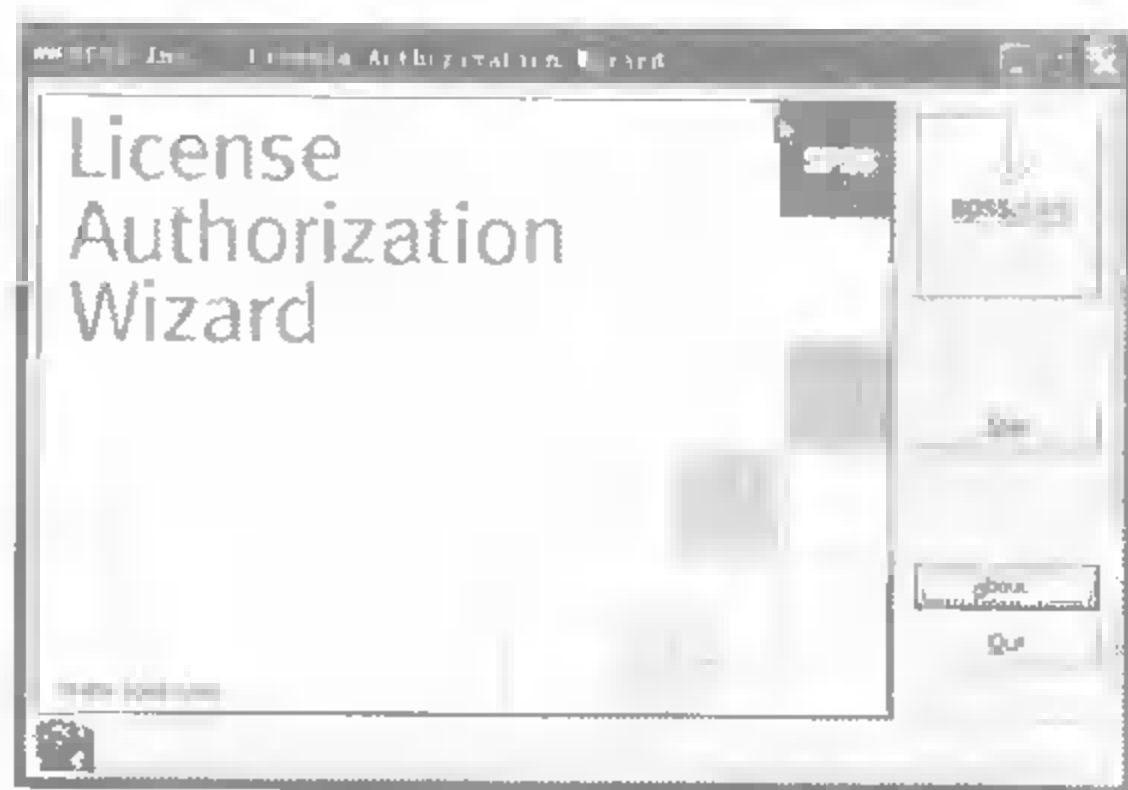


图 1-10 SPSS 注册 2

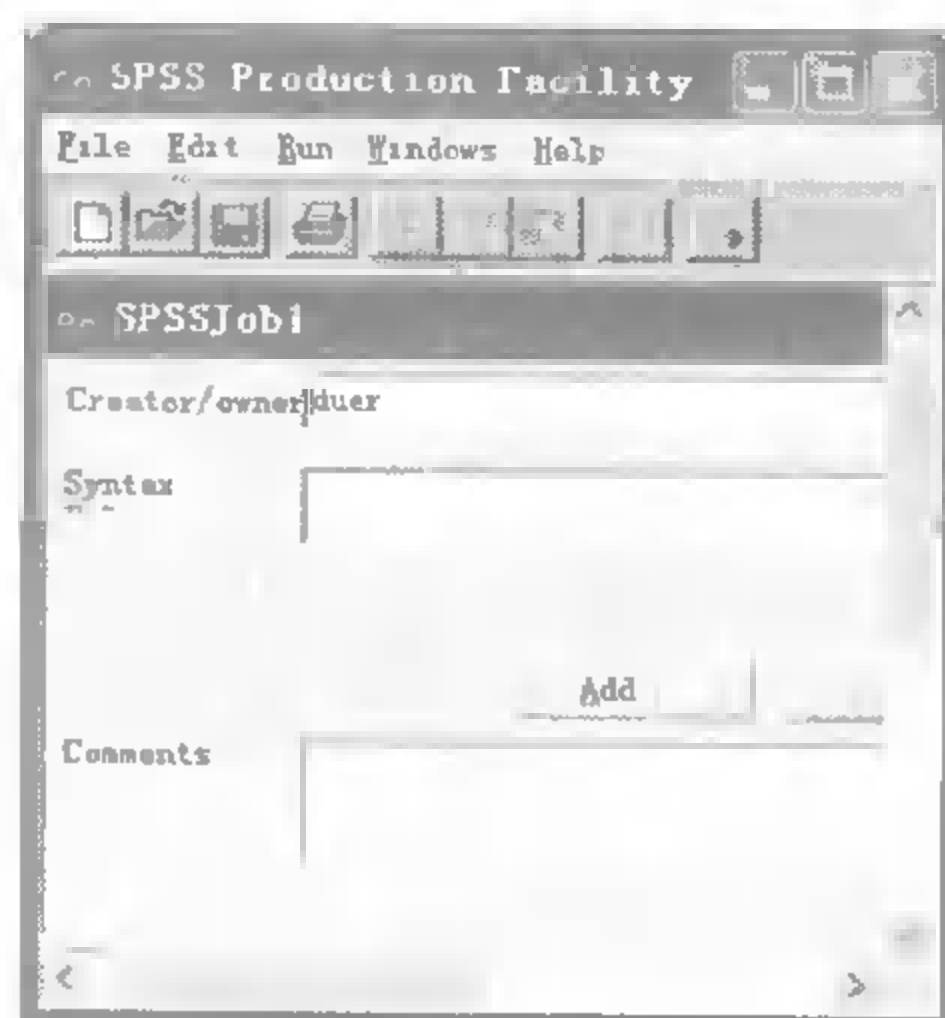


图 1-11 SPSS Production Facility 界面

(2) 程序运行方式。直接在 Syntax 语句窗口或 Script 脚本窗口编辑和运行程序，这种方式要求掌握 SPSS 的 Syntax 语言或 Sax Basic 脚本语言。

(3) 混合运行方式。首先在“完全窗口”方式下的数据编辑窗口中输入数据，或者利用“File”菜单打开已经存在的数据文件；然后利用菜单和对话框操作，设置数据处理的参数；参数设置好后单击设置界面中的 Paste 按钮，将选择的过程及参数转换成相应的命令语句，置于 Syntax 语句窗口；最后，在语句窗口中添加语句、参数，或者修改已有命令中的参数，单击窗口中的 Run 按钮执行分析。

### 1.2.3 SPSS 15.0 的退出

单击菜单“File→Exit”，或者单击 Data Editor 窗口右上角的关闭按钮，都可以退出 SPSS。

第一次退出时会弹出如图 1-12 所示的提示框，提示用户这是退出操作，勾选复选框后，再单击 Yes 按钮，下次退出时就不会再显示这个提示框了。

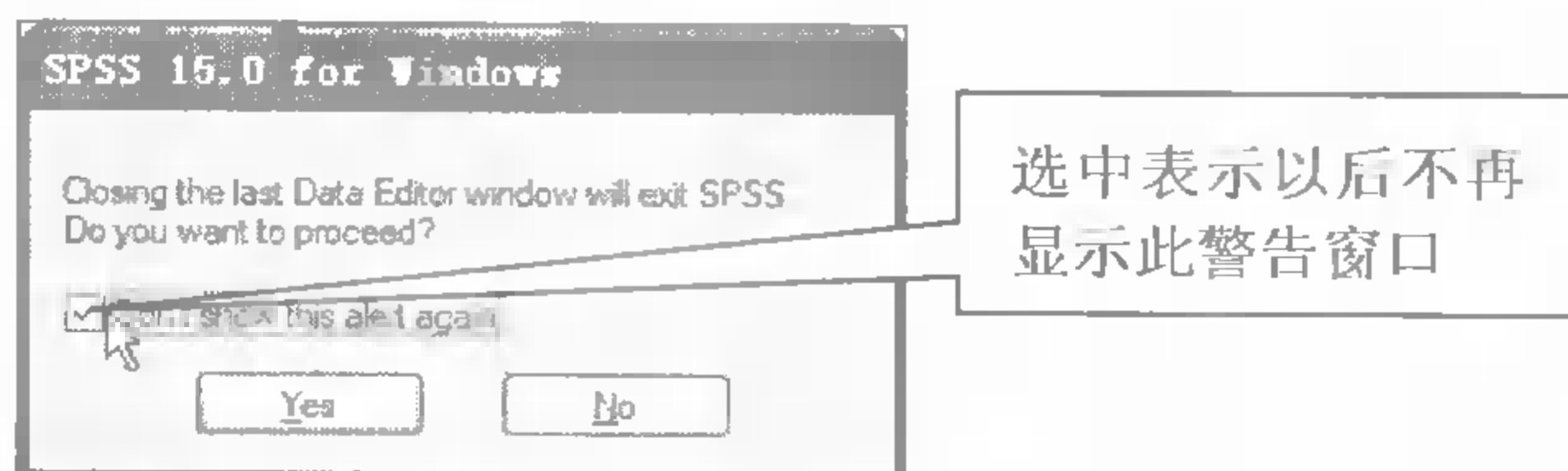


图 1-12 SPSS 退出提示框

如果对文件进行了修改，或者输出了新的结果，退出时还会弹出如图 1-13 所示的文件保存对话框，对相关内容保存后即可正常退出（注意：非正常退出可能引起数据丢失）。



图 1-13 SPSS 提示文件保存对话框

## 1.3 SPSS 15.0 的界面及设置

本节首先介绍 SPSS 常用的几个界面及其功能，对大多数用户来讲，SPSS 是“窗口+对话框”方式的应用工具，熟悉和了解这些界面，对提高使用 SPSS 进行统计分析的工作效率

是非常必要的。随后介绍 SPSS 15.0 的环境参数设置，这些设置将影响常用窗口的运行方式，通过更改它们能够把 SPSS 的工作环境定制为自己喜欢的方式。关于系统的操作、输出和显示等参数，都可以依次单击菜单“Edit→Options...”进行设置。

若无特殊声明，以下几点对每个设置界面均有效：

新的设置仅对应用它们之后产生的输出起作用，而应用新设置之前的输出不会改变；更改参数后，单击“确定”或“应用”按钮即可应用新设置；需要重新设置或暂时不需要进行设置工作时，单击“取消”按钮退出 Options 对话框，返回到 SPSS for Windows 主画面，已设置的参数无效。

### 1.3.1 常用界面

SPSS 的基本界面有数据编辑窗口、结果观察窗口、对象编辑窗口、草稿输出窗口、命令语句窗口和脚本编写窗口，分别介绍如下。

#### 1. 数据编辑（Data Editor）窗口

启动 SPSS 后，在图 1-7 中单击选中 Type in data 或 Open an existing data source，也可单击 Cancel 按钮，进入的第一个窗口便是 Data Editor（数据编辑）窗口，如图 1-14 所示，这是 Data Editor 窗口的数据视图（Data View）。单击底部的“Variable View”标签，可以切换到变量视图，如图 1-15 所示；单击“Data View”标签又可切换回数据视图。在数据编辑窗口中可以进行数据的录入和编辑以及变量属性的定义和编辑，是 SPSS 的基本界面。

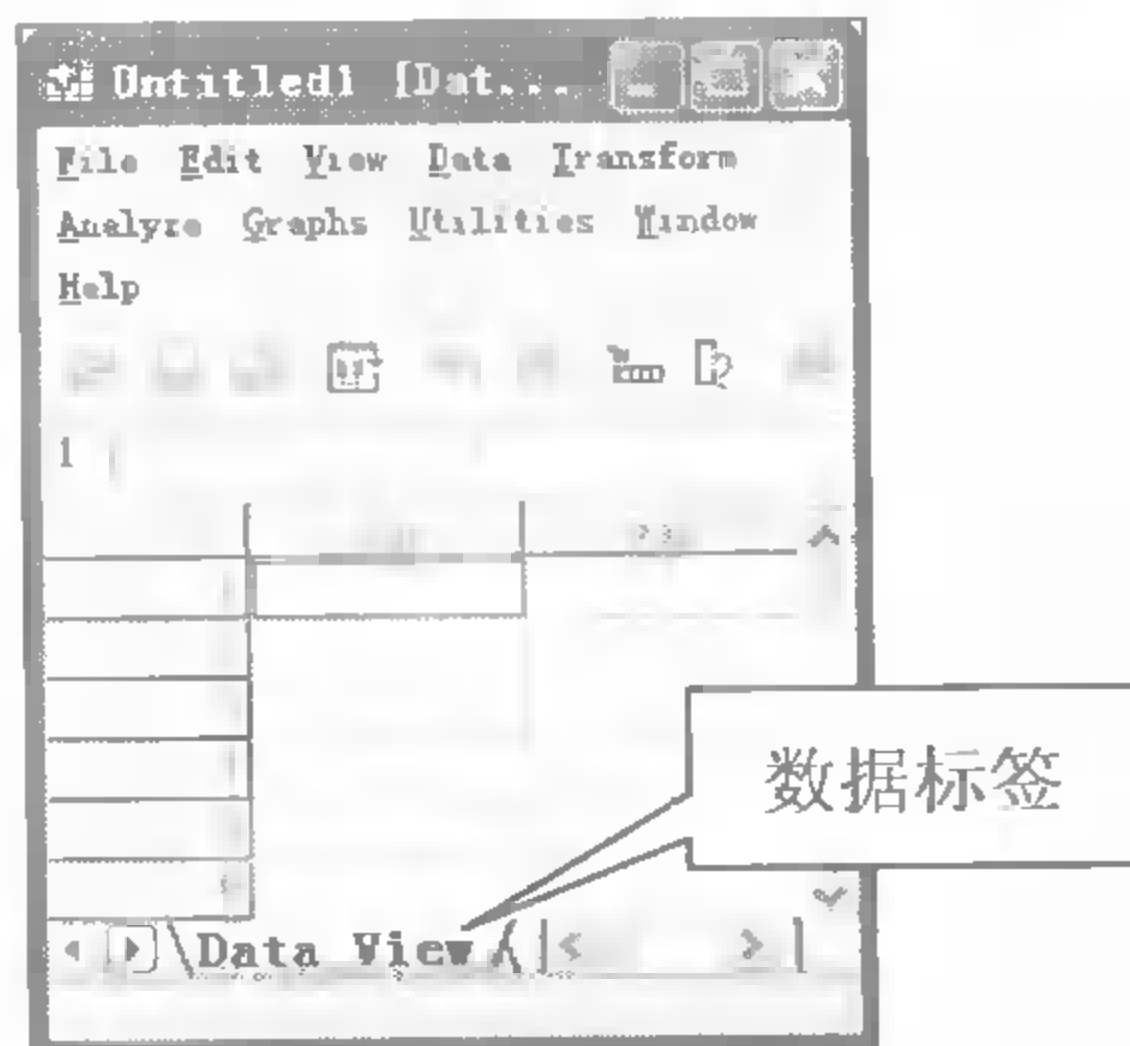


图 1-14 Data Editor 窗口的数据视图

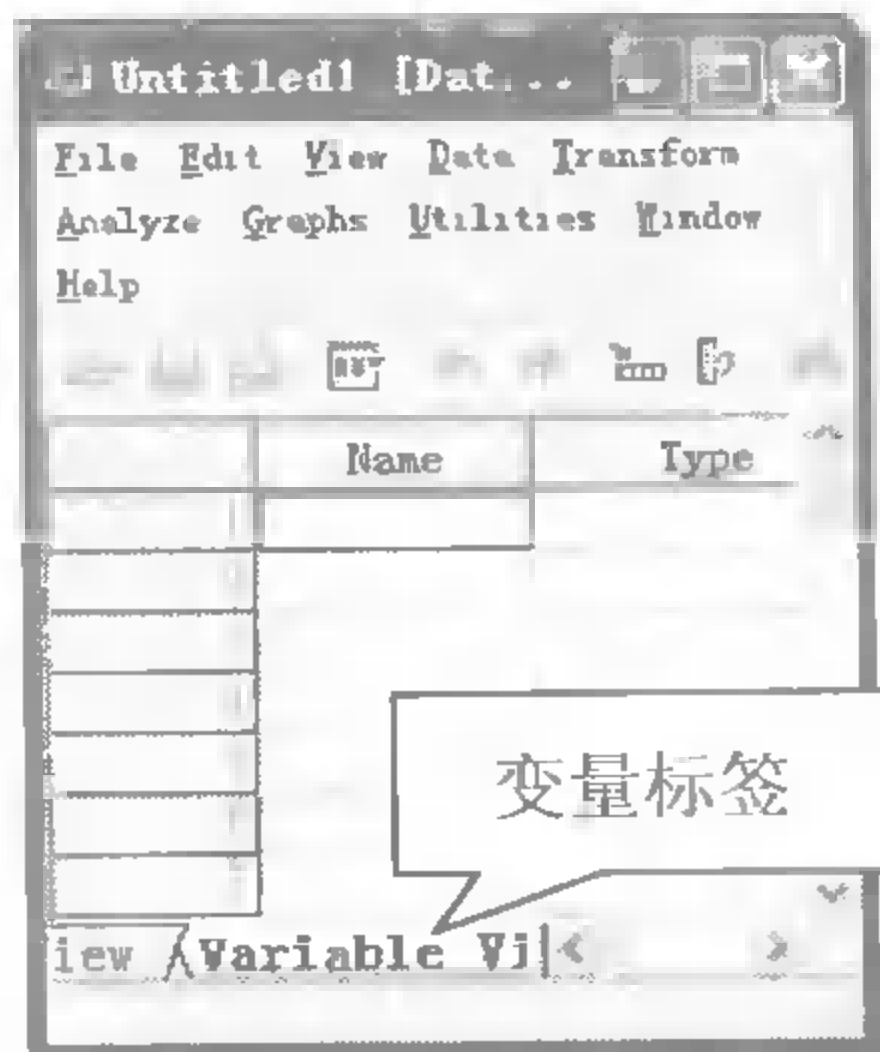


图 1-15 Data Editor 窗口的变量视图

在图 1-7 中，勾选 Don't show this dialog in the future 复选框后，再次启动 SPSS 将直接进入如图 1-14 所示的 Data Editor 窗口。

#### 2. 结果观察（SPSS Viewer）窗口

SPSS 中大多数统计分析结果都将以表或者图的形式在结果观察窗口中显示，如图 1-16 所示。通过设置，当用户进行了某个操作（例如打开文件、OLAP 报告、回归等）后，SPSS Viewer 窗口可以自动弹出；不自动弹出时，相关结果也会在后台显示在此窗口中，只需激活即可看到。双击后缀名为“spo”的 SPSS 输出结果文件，也可以打开本窗口。

在图 1-16 中，右边的显示窗口输出 SPSS 统计分析的结果（包括日志、表格、图形等），左边的导航窗口显示输出结果的目录，单击其中的加、减符号可以显示或隐藏相关内容，在左边窗口选中某项时出现一个红色箭头指向它，所选内容的细节显示在右边的显示窗口。



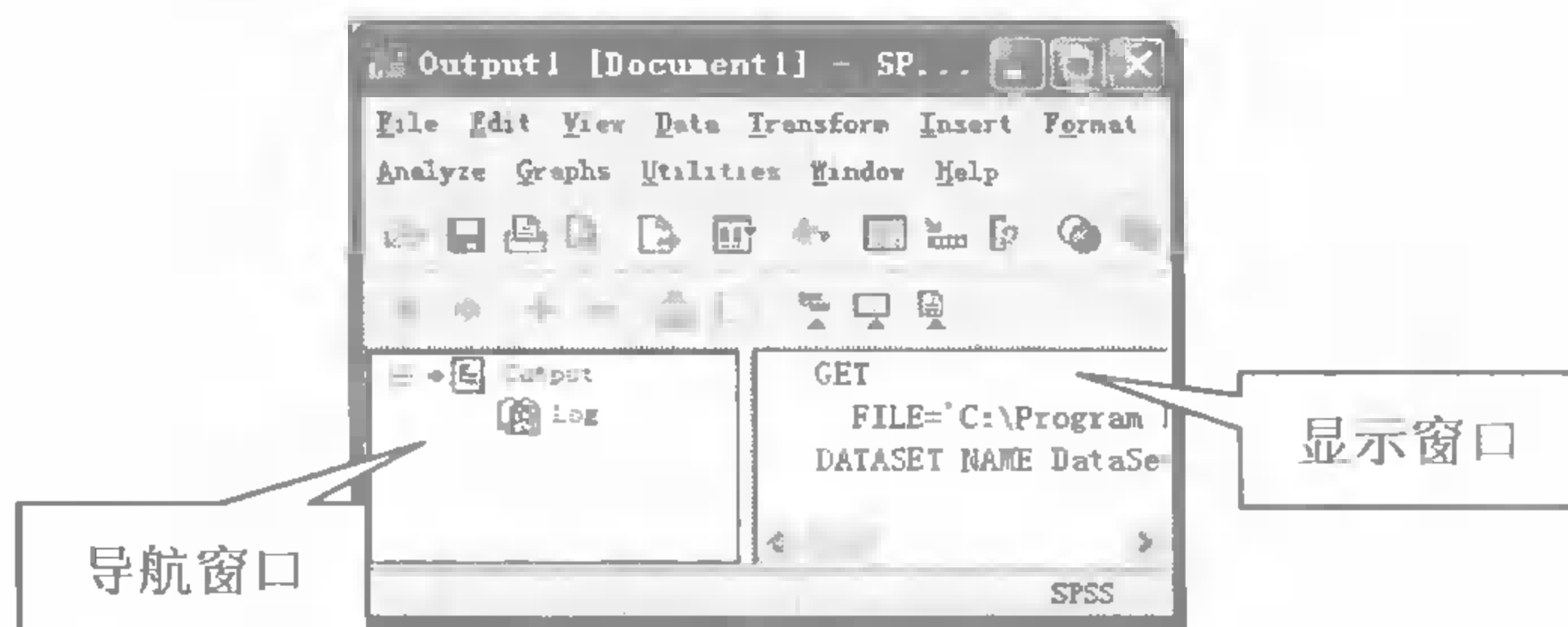


图 1-16 SPSS Viewer 窗口

### 3. 对象编辑 (SPSS Object) 窗口

在图 1-16 的显示窗口里, 用鼠标右键单击某个项目 (表格或图形), 弹出的快捷菜单里有一项“SPSS \*\*\* Object” (此处星号位置可以是 document、pivot table、chart、interactive graph 等), 单击此选项后, 会弹出如图 1-17 所示的 Pivoting Trays (枢纽表) 窗口、图 1-18 所示的 Chart Editor (图形编辑) 窗口等。在图 1-16 的显示窗口中直接双击其中的表格或图形, 也可以打开对应的 Object 窗口。

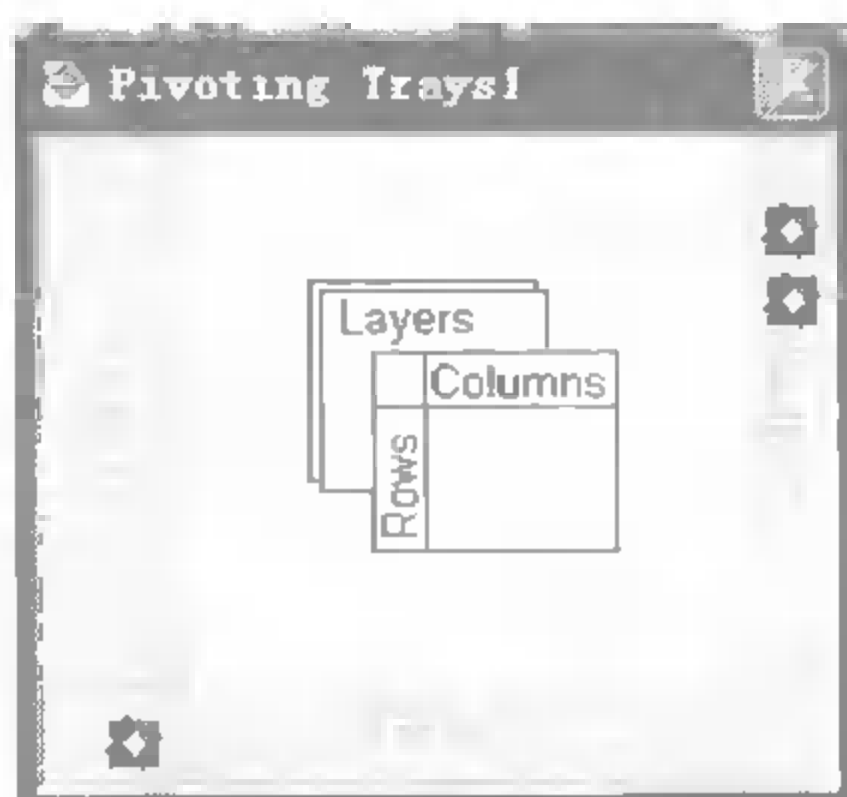


图 1-17 Pivoting Trays 窗口

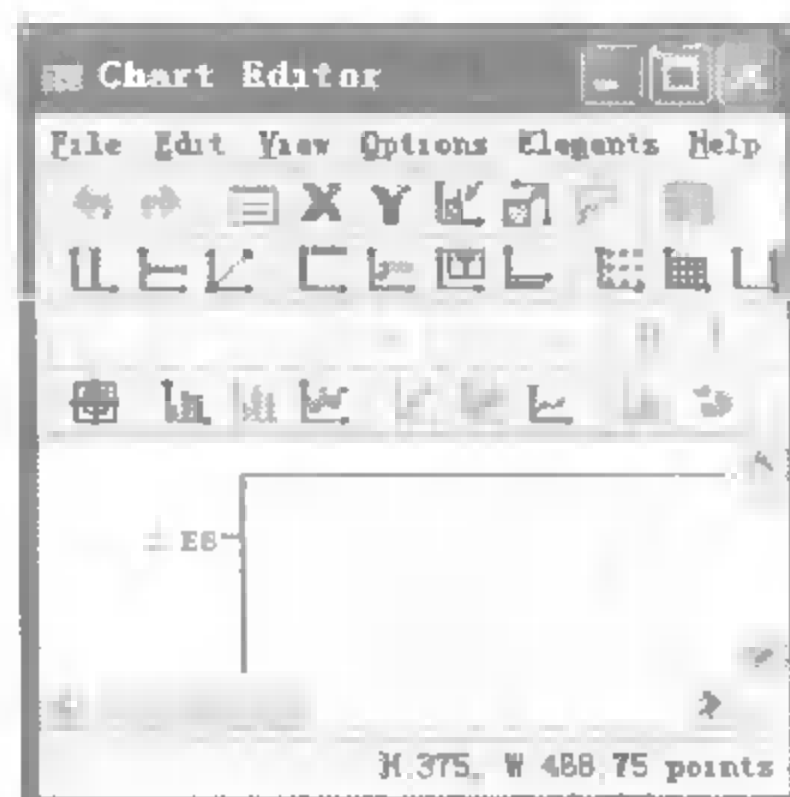


图 1-18 Chart Editor 窗口

在 SPSS Viewer 的显示窗口, 激活 Interactive graph (交互式图形) 对象后, 并不打开新的图形编辑窗口, 而是在原图形位置嵌入它的编辑窗口, 如图 1-19 所示, 图形周边的虚线线框提示这是处于编辑状态时的交互式图形。

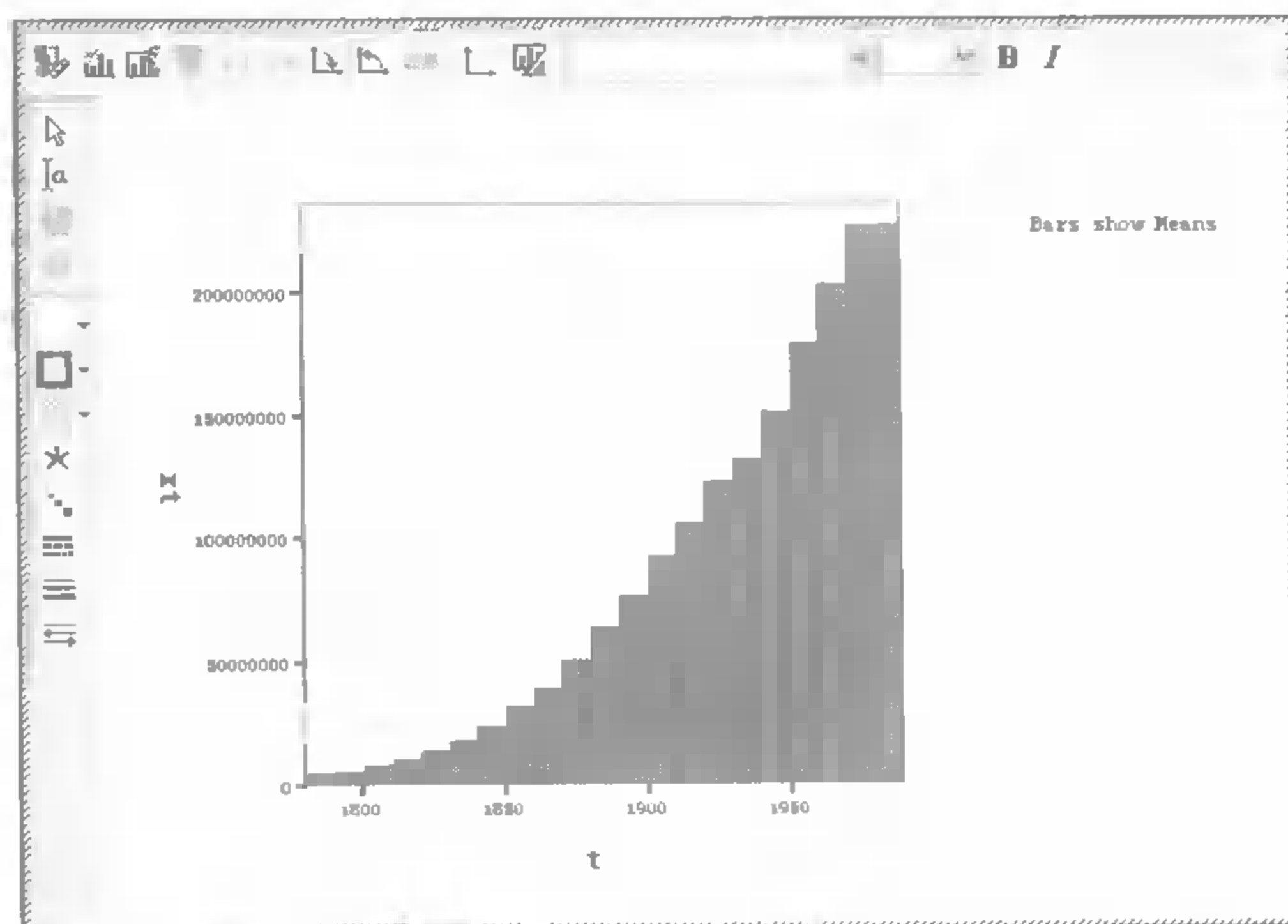


图 1-19 Interactive graph 编辑状态

关于这些 Object 对象的设置和编辑, 后面相关的章节将做具体介绍。

#### 4. 草稿输出 (Draft Viewer) 窗口

SPSS 的 Draft Viewer (草稿输出) 窗口如图 1-20 所示, 可以设置它为默认的结果输出窗口, 这样执行分析过程后会自动显示此窗口。依次单击菜单 “File→New→Draft Output” 或 “File→Open→Output”, 也可以打开此窗口。

在 Draft Viewer 输出窗口中, 枢纽表 (Pivot table) 转换为文本输出, Chart 图形转换为图元文件, 它们都可以编辑, 这一点对于把相关项目应用在类似 Word 的编辑环境中非常有用。

#### 5. 命令语句窗口 (Syntax)

依次单击菜单 “File→New→Syntax” 或 “File→Open→Syntax” 可以打开 Syntax 窗口, 单击任何统计分析对话框上的 “Paste” 按钮, 可自动把对话框设置的各种命令和选项粘贴到 Syntax 窗口中, 如图 1-21 所示。用户可以直接输入 SPSS 语句命令, 或者对复制的内容进行修改, 依次单击菜单 “Run→All” 可执行这些命令。将编写好的 SPSS 程序保存为后缀名为 “sps” 的文件, 可供以后调用。



图 1-20 Draft Viewer 窗口

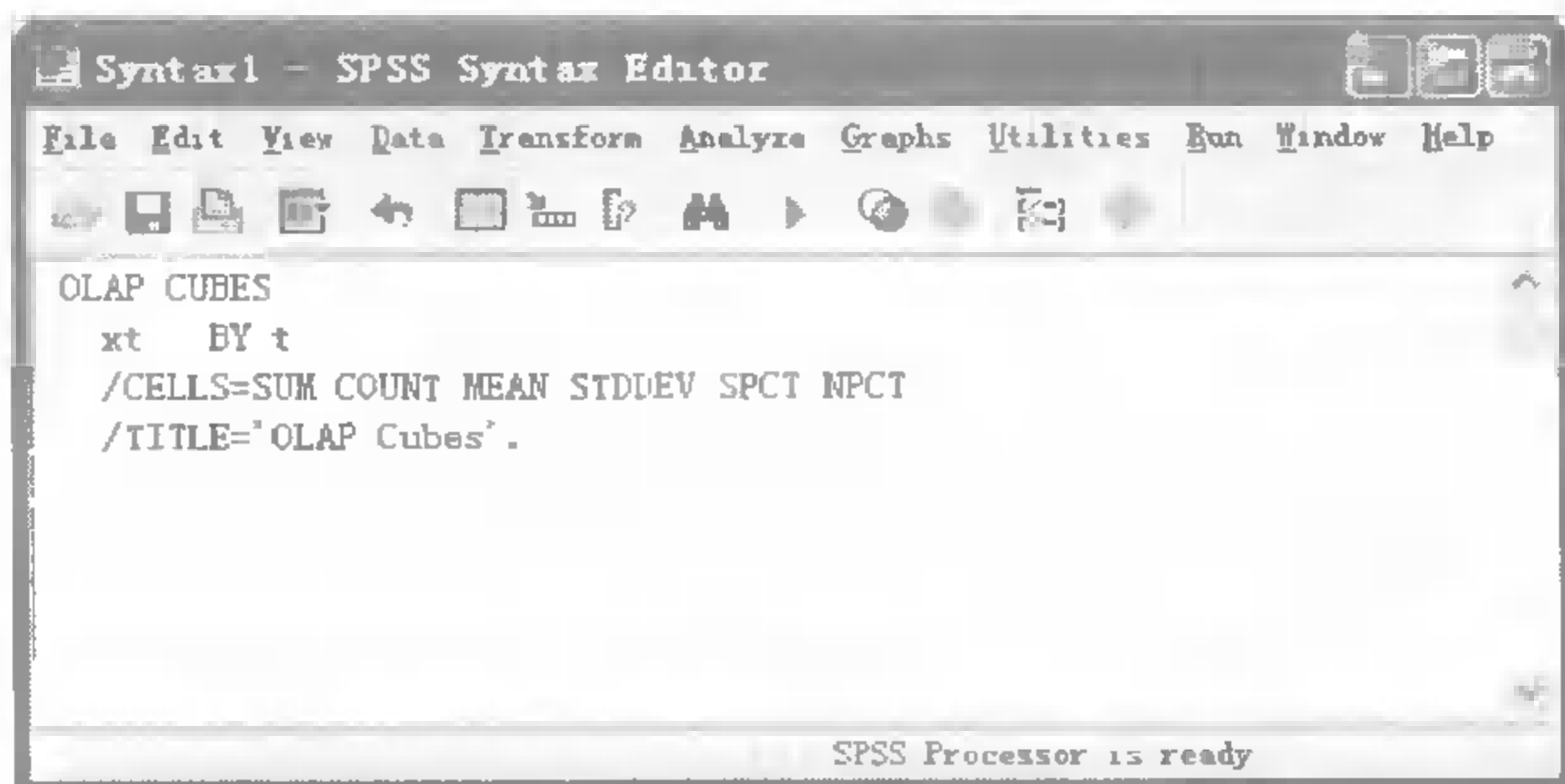


图 1-21 Syntax 窗口

#### 6. 脚本编程 (Script) 窗口

依次单击菜单 “File→New→Script” 或 “File→Open→Script”, 可以打开 Script 窗口, 如图 1-22 所示。在此编写 SPSS 内嵌的 Sax Basic 语言程序 (一种类似 VB 的语言)。一方面可以开发 SPSS 的便捷功能或插件, 另一方面可以编写自动化数据处理的程序。

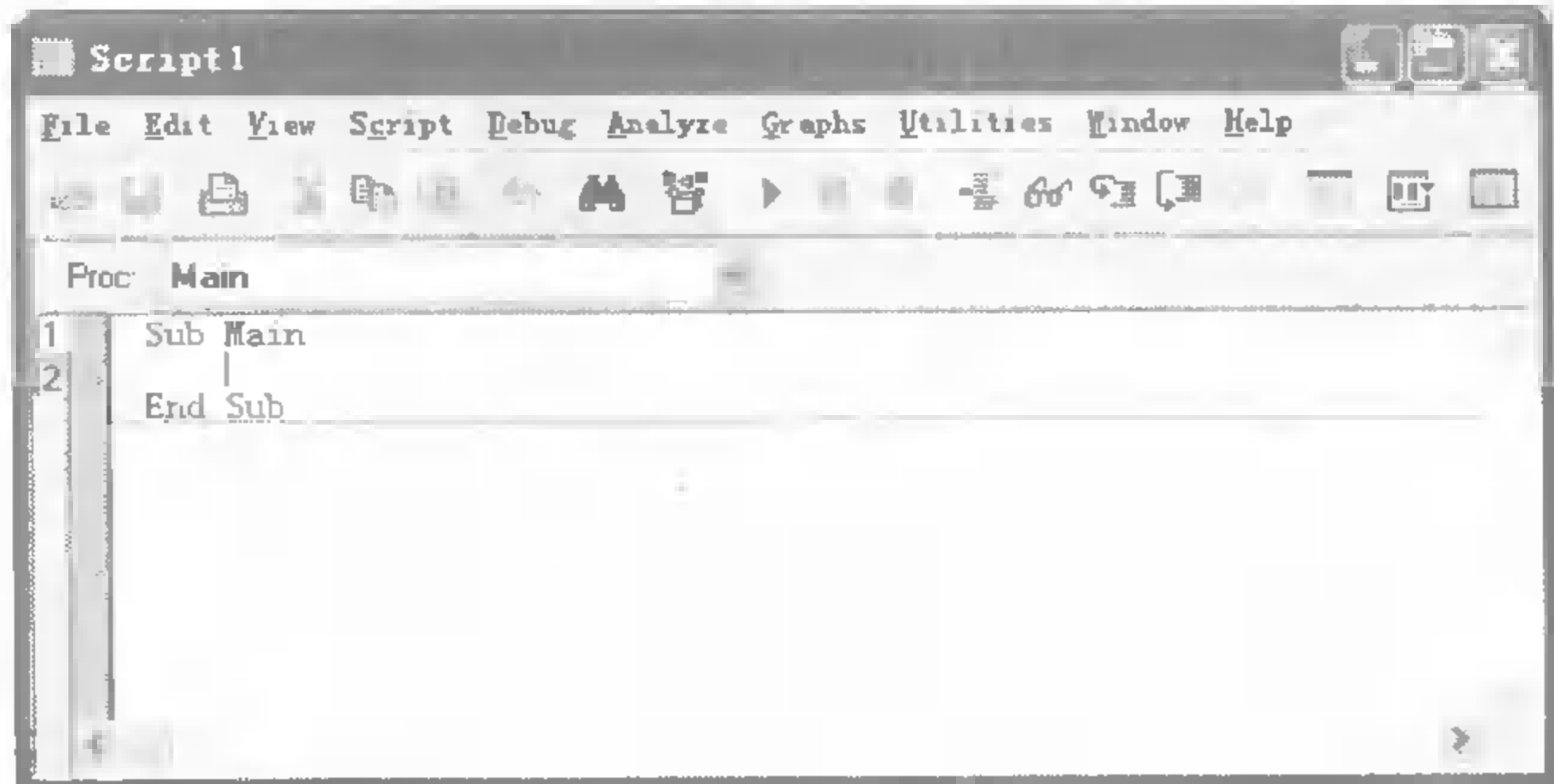


图 1-22 Script 窗口

图 1-11 所示的 SPSS Production Facility 界面和功能即可由 Sax Basic 语言实现。

1.3.2 通用 General 功能参数

General 面板可设置 SPSS 软件系统的各种通用参数（例如启动选项、临时文件夹选项和显示语言等），如图 1-23 所示。所设参数可自动保存，重新启动 SPSS 后不需要重新设置。具体设置内容如下。

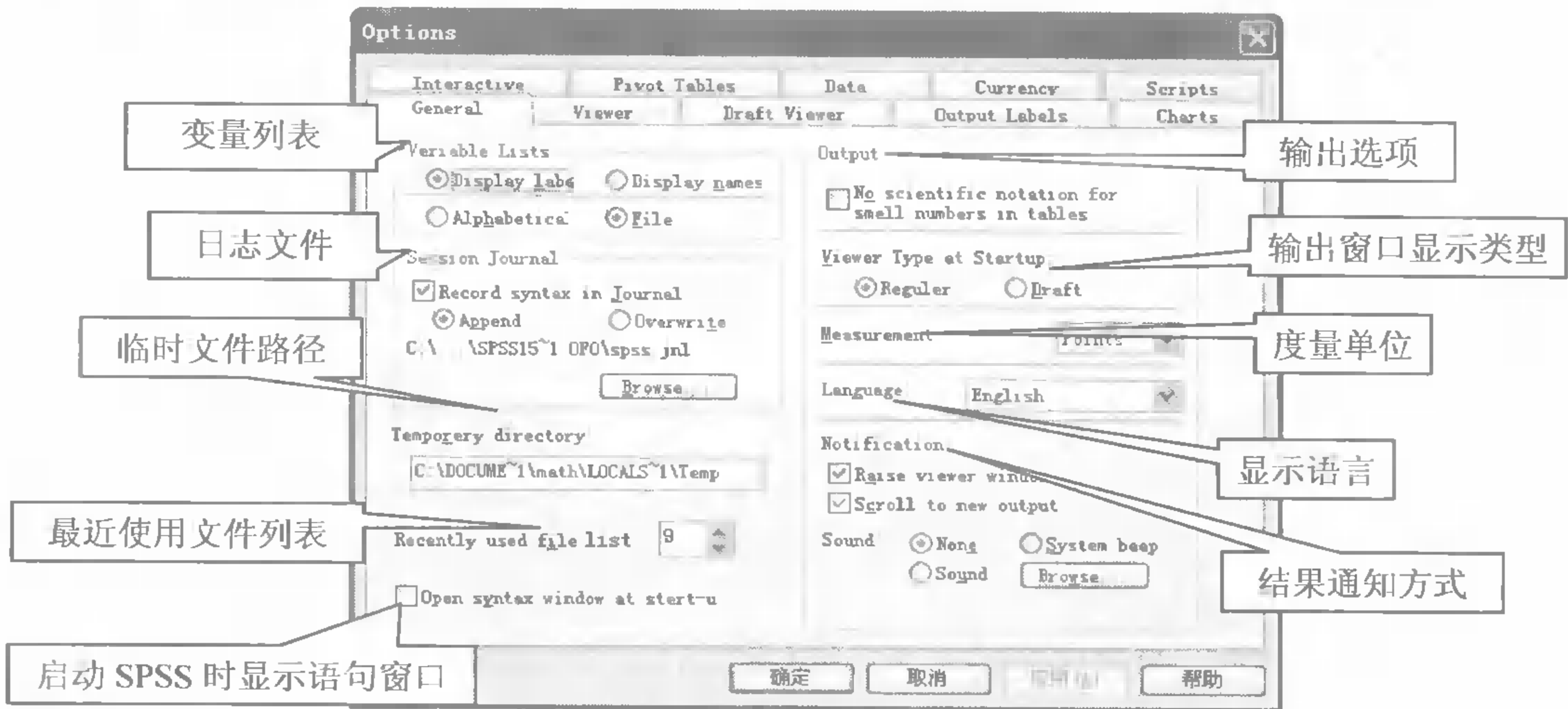


图 1-23 General 标签设置

1. Variable Lists（变量列表）

Variable Lists 子设置栏控制各种统计分析对话框中变量列表的显示方式。例如：在如图 1-24 所示的 OLAP 分析对话框里左侧变量列表的显示方式。

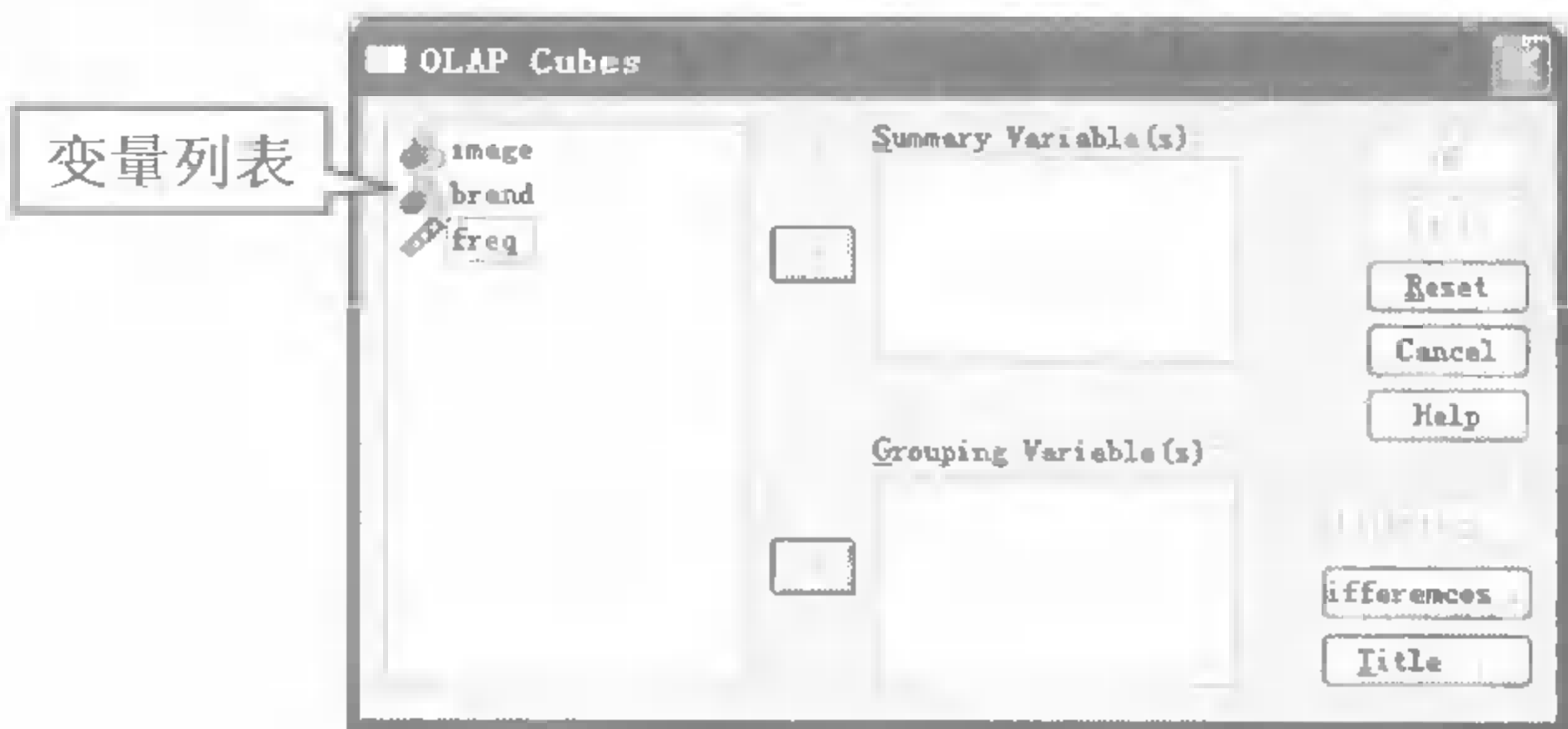


图 1-24 对话框变量列表

Variable Lists 栏的可选参数有显示内容和排列顺序两项，含义如表 1-2 所示。

表 1-2 通用变量列表显示参数

显示内容	排列顺序
Display names: 变量名	Alphabetical: 字母排列
Display labels: 变量标签	File: 按变量在数据文件和数据编辑窗口的显示顺序排列，显示顺序只影响原有变量，不影响由模型产生的目标变量，且一般只反映当前模型所选中的原有变量

2. Session Journal（日志文件）

所有程序段所运行的命令都将保存在一个日志文件里（包括在 syntax 语句窗口中输入和运行的命令和通过对话框产生的命令）。日志文件可以编辑，其中的命令可以直接复制到

Syntax 窗口再次运行，日志文件可以删除、增加、叠加和覆盖，并可选择保存地址和名称。

从日志文件中复制命令语句，存入一个 syntax 语句文档，再结合图 1-11 所示的 SPSS Production Facility 工具或 Script 编程，能够实现自动化处理的功能。

(1) 日志文件的保存方式有如下 2 种：

- Append (附加模式)：每次的运行语句接在前一次运行语句记录后面存入日志文件。
- Overwrite (覆盖模式)：每次运行语句存入日志文件时覆盖前一次存入的内容。

(2) 设定日志文件名及存储路径

单击 Browse 按钮，打开保存文件的对话框，指定日志文件的存储路径和文件名，文件名及其路径将显示在 Browse 按钮的上方。若用户不指定文件名，运行信息将自动保存在默认日志文件中 (C:\Documents and Settings\duer\Local Settings\Application Data\SPSS 15.0 for Windows\spss.jnl)。

### 3. Temporary directory (临时文件路径)

此栏设置数据分析过程所产生临时文件的存放位置，直接在下面的输入框键入临时文件的完整路径即可。临时文件往往需要比较大的空间，所以目标磁盘容量不能太小，建议至少有 512M 或更大的可用空间。如果分析过程中出现无法存取临时目录或临时目录空间已满的提示，需要在此栏修改路径设置。

注意：分布式模式下 (SPSS server 版本)，此栏的设置不会影响临时文件的存放位置，此时的临时文件通过环境变量 SPSSTMPDIR 控制，且只能在运行 server 版本的机器上修改。

### 4. Recently used file list (最近使用文件列表)

最近使用的数据文件(或语句文件)列表会自动显示在菜单“File→Recently used data(或 Recently used file)”中，此子设置栏控制要显示最近使用文件的数目，直接更改输入框中的数字即可。

### 5. Open Syntax window at start-up (启动 SPSS 时显示语句窗口)

如果常常使用命令语句，或是喜欢使用命令语句的有经验的用户，单击将此复选框选中，表示在启动 SPSS 时自动打开一个 syntax 语句窗口。

注意：Student 版本的 SPSS 没有此项设置。

### 6. No scientific notation for small numbers in tables

单击选中此复选框，输出结果中将不显示非常小的小数，而以 0 或 0.000 代替。

### 7. Viewer Type at Start-up (输出窗口显示类型)

此栏设置 Viewer 窗口的显示类型和输出格式，可选项有如下 2 个。

- Regular 单选框：显示 SPSS Viewer 窗口，输出交互式要点表和交互式图形。
- Draft 单选框：显示 Draft Viewer 窗口，把枢纽表转换为文本，图形转换为图元文件。

### 8. Measurement System (度量单位)

此栏设置要点表单元格边距、单元格宽度、打印表格的间隔等内容的度量单位。下拉列表中可选项有：Points (点)、Inches (英寸) 和 Centimeters (厘米)。

### 9. Language (显示语言)

此栏设置输出结果的显示语言，但是对简单文本输出、交互式图形和地图不起作用。下

拉菜单可选的语言种类和 SPSS 的安装版本有关, 如图 1-25 所示。

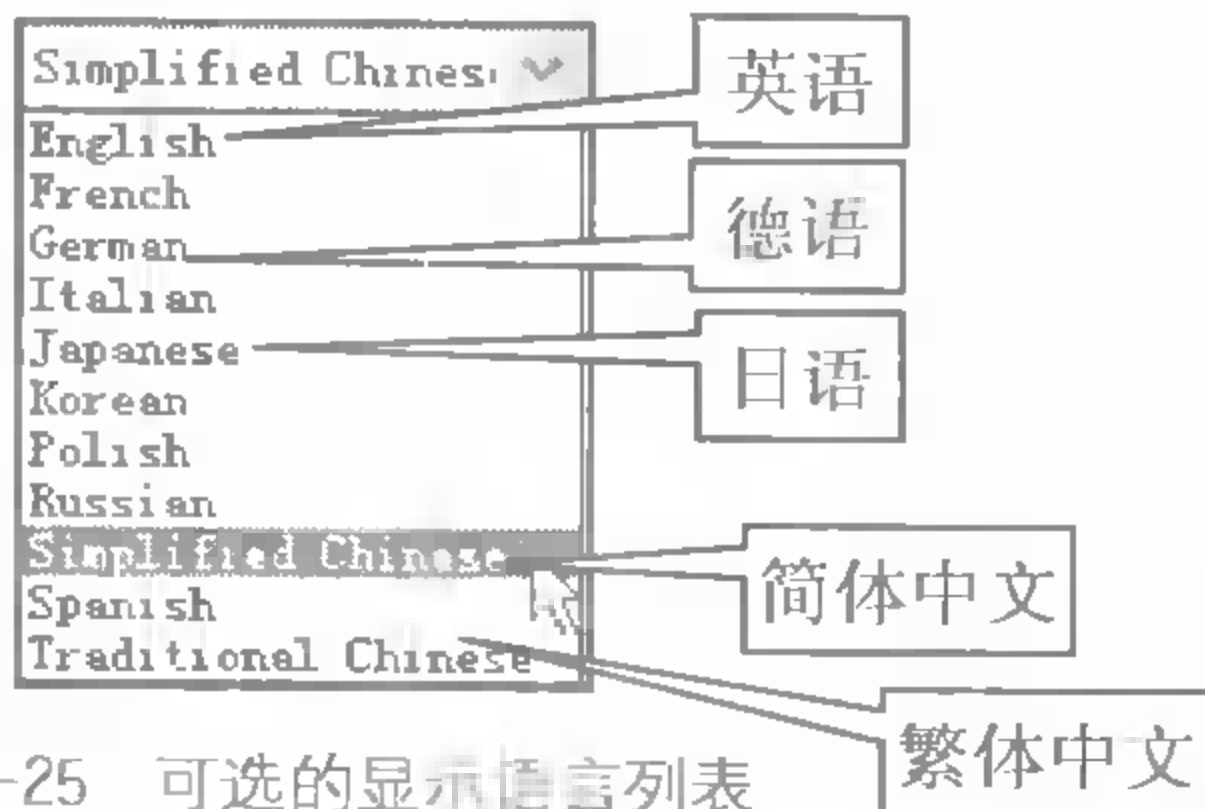


图 1-25 可选的显示语言列表

注意: 修改此项后, 依赖于特定一种语言或字符的自定义脚本可能无法正常显示。

## 10. Notification (结果通知方式)

此栏设置程序运行结果的通知方式, 各设置选项的含义如表 1-3 所示。

表 1-3

程序运行结果通知方式

Raise Viewer Window 复选框	打开 Viewer 窗口
Scroll to new output 复选框	自动在 Viewer 窗口中滚屏至最新的输出结果
Sound 单选框组	指定提示声音, 单击 Browse 按钮可以选择声音文件

### 1.3.3 Viewer 视图窗口参数

Viewer 面板设置如图 1-16 所示的 SPSS Viewer 窗口的显示参数, 包括输出文字的字体、大小和颜色等, 设置界面如图 1-26 所示。可以单独设置不同输出对象 (文本、表格、图形等) 的显示方式和对齐方式。

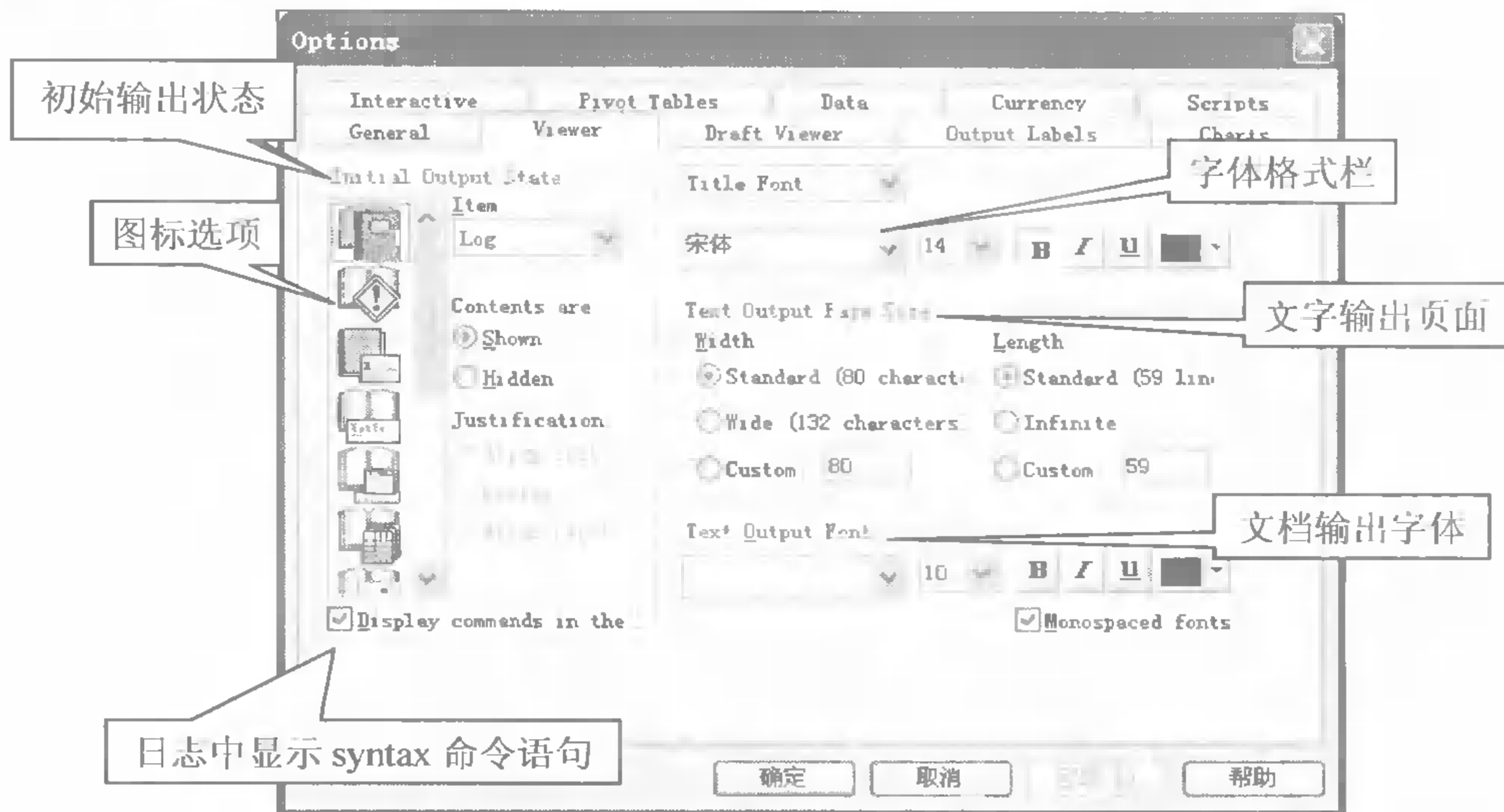


图 1-26 Viewer 标签设置

(1) Initial Output State (初始输出状态)。此子设置栏选择特定输出结果的初始状态参数。首先单击 Item 下拉菜单选中要设置的输出结果, 然后在下面设置所选输出内容的显示参数。

① 选择要设置的输出对象。Item 下拉菜单给出了可选的输出对象, 包括: Log (日志)、Warnings (警告)、Notes (注释)、Titles (标题)、Page Title (页面标题)、Pivot Tables (枢纽



表)、Charts (图表)、Text Output (文本输出)、Graph (图形)、Map (地图)、Tree Model (决策树模型输出) 和 Tree Outline (决策树轮廓), 每个对象在左侧的图标选项列表都有一个与之相应的图标。

② Contents are 栏设置显示方式。可选项有 Shown (显示) 和 Hidden (隐藏)。

③ Justification 栏设置对其方式。可选项有 Align left (左对齐)、Center (居中) 和 Align right (右对齐)。注意: 所有普通 Item 在 Viewer 窗口的对齐方式都是左对齐, 此处的对齐方式只对打印输出有效。

④ Display commands in the log on or off 复选项。勾选此复选框后, 将在日志中显示 SPSS syntax 命令语句, 用户可以从日志中复制命令语句, 并将它们保存在语句文件中。

(2) Title Font (字体) 和 Page Title Font (标题字体)。

单击 Title Font 右端的下拉菜单, 选中 Title Font 选项, 在下面的字体格式栏设置新输出字体的字型、大小和颜色。

单击 Title Font 右端的下拉菜单, 选中 Page Title Font 选项, 在下面的字体格式栏设置新输出页面标题的字型、大小和颜色 (包括由 TITLE 命令、SUBTITLE 命令和 Insert 菜单插入产生的新页面)。

(3) Text Output Page Size (文字输出页面)。Width 栏设置文档输出的页宽, 通过字符数反映。Standard 单选框代表 80 个字符, Wide 单选框代表 132 个字符, Custom 单选框允许用户指定字符个数。

Length 栏设置文档输出的页长, 通过行数反映。Standard 单选框代表 59 行; Infinite 单选框代表无穷多行; Custom 单选框允许用户指定行数。

(4) Text Output Font (文档输出字体)。在下面的字体栏中设置文档输出的字体。单击选中 Monospaced fonts 复选框, 表示使用固定间距的字体 (等宽字体)。

#### 1.3.4 Draft Viewer (草稿窗口) 参数

Draft Viewer 标签设置如图 1-20 所示的草稿输出窗口的显示参数, 包括 Draft 窗口的显示内容、字体、颜色、页宽、页长以及表格的显示方式等, 如图 1-27 所示。



图 1-27 Draft Viewer 标签设置

(1) Display Output Items (显示输出项)。此栏选择在 Draft Viewer 窗口中输出哪些内容,可选项有 Display commands in the log on or off(在日志中显示命令语句)、Warnings (警告)、Tabular Output (列表输出)、Notes (注释)、Titles (标题)、Charts (图表)、Text Output (文本输出, 空格分隔) 和 Log (日志), 其中 Pivot table 将会转换为 Text 文本输出。

(2) Page Breaks Between (分页规则)。此栏设置输出结果的分页规则,可选项有 Procedures 过程(以数据处理过程段分页,如 Frequencies 频数分析、Regression 回归等)和 Items (项目分页,如表格、图形等)。

(3) Font (字体)。此栏设置新输出内容等宽格式的字体(monospaced),保证以空格分割的文本输出正确对齐。

(4) Scale font so wide tables fit printed page。选中此复选框,表示打印页面结果时自动减小字号,适应表格宽度,在 Minimum Size 输入框中填写最小字号。

(5) Tabular Output (列表输出格式)。此栏设置 pivot table (枢纽表)转换为 tabular text (文本列表)时的格式。

枢纽表是交互式表格,用户可以通过表格里的下拉菜单选择显示不同的内容,如图 1-28 所示;文本列表是非交互式表格样式的文本,如图 1-29 所示。枢纽表转换为文本列表时,会把下拉菜单中的每种组合单独列为一个文本列表。

时段						
名称	四川长虹					
	合计	N	均值	标准差	总和的 %	合计 N 的 %
收盘价	28.23	4	6.5575	.08261	4.0%	8.9%
交易量	1586805.00	4	396701.2500	131723.367	3.8%	8.9%

图 1-28 枢纽表

时段	1					
名称	四川长虹					
	合计	N	均值	标准差	总和的 %	合计 N 的 %
收盘价	28.13	5	5.8260	28510	4.3%	8.6%
交易量	3161078.00	5	632215.6000	148006.79816	7.6%	8.6%

图 1-29 文本列表

此栏中可设置的显示格式包括如下几项:

① Separate columns (列分割符)。单击选中 Spaces (空格) 时,可以设置列和单元格的显示方式,可选项如表 1-4 所示。

表 1-4	Draft 窗口的列表显示参数
Display Box Character 复选框	选中表示在单元格周围显示实线;取消选中,激活 Cell 子设置栏,在 Row、Column 输入框中分别输入行和列的线型
Autofit 单选框	自适应列宽
Maximum 单选框	列宽最大值,单击选中它后,在下面的 Character 输入框指定单列最多可以显示的字符数

单击选中 Tabs 时,表示用制表符分割列,输出列表无边框。注意:用 Tabs 作为分隔符

的列表在 Draft Viewer 草稿窗口可能无法正常对齐，但是需要用 Word 处理输出结果时，选中此项会比较方便。

② Repeat column headers 复选框。选中此复选框，表示当输出表格分多页显示时，在每个连续页顶端都重复显示列标题；不选中它时，将只在第一页输出列标题。

(6) Text Output (文本输出格式)。此栏设置非 Pivot Table 转换的文本输出页面格式，Page Width 设置页宽（以字符为单位），Page Length 设置页面长度（以行数为单位）。

### 1.3.5 Output Labels 输出标签参数

Output Labels 标签设置输出结果的标签选项，如图 1-30 所示。变量标签的内容设置可以在 Data Viewer 窗口的 Variable View 视图中进行，通过在输出中显示 Labels（标签）可以方便地观察变量的实际意义。



图 1-30 Output Labels 标签设置

此处可以分别对 Outline Labeling（文本输出标签）和 Pivot Table Labeling（枢纽表标签）进行设置，二者的设置内容相同。

(1) Variables in item labels shown as 下拉列表。此栏设置变量标识，可选项有 3 个：Labels（标签），使用变量标签标识每个变量；Name（名字），使用变量名标识每个变量；Names and Labels（名字和标签），同时使用变量名和变量标签标识每个变量。

(2) Variable values in item labels 下拉列表。此栏设置变量值的显示方式，可选项也有 3 个：Labels（标签），使用变量值标签标识每个变量值；Name（名字），使用变量值本身作为标识；Names and Labels（名字和标签），同时使用变量值本身和变量值标签标识每个变量值。

注意：此处设置只对 Pivot tables 有效，而对 Text 文本输出无效。

### 1.3.6 Charts 图形参数

Charts 选项卡设置输出窗口里的图形默认参数，包括比例、边框、线型、颜色和填充方式等，还可以设置 Java 虚拟机的装载时机，如图 1-31 所示。

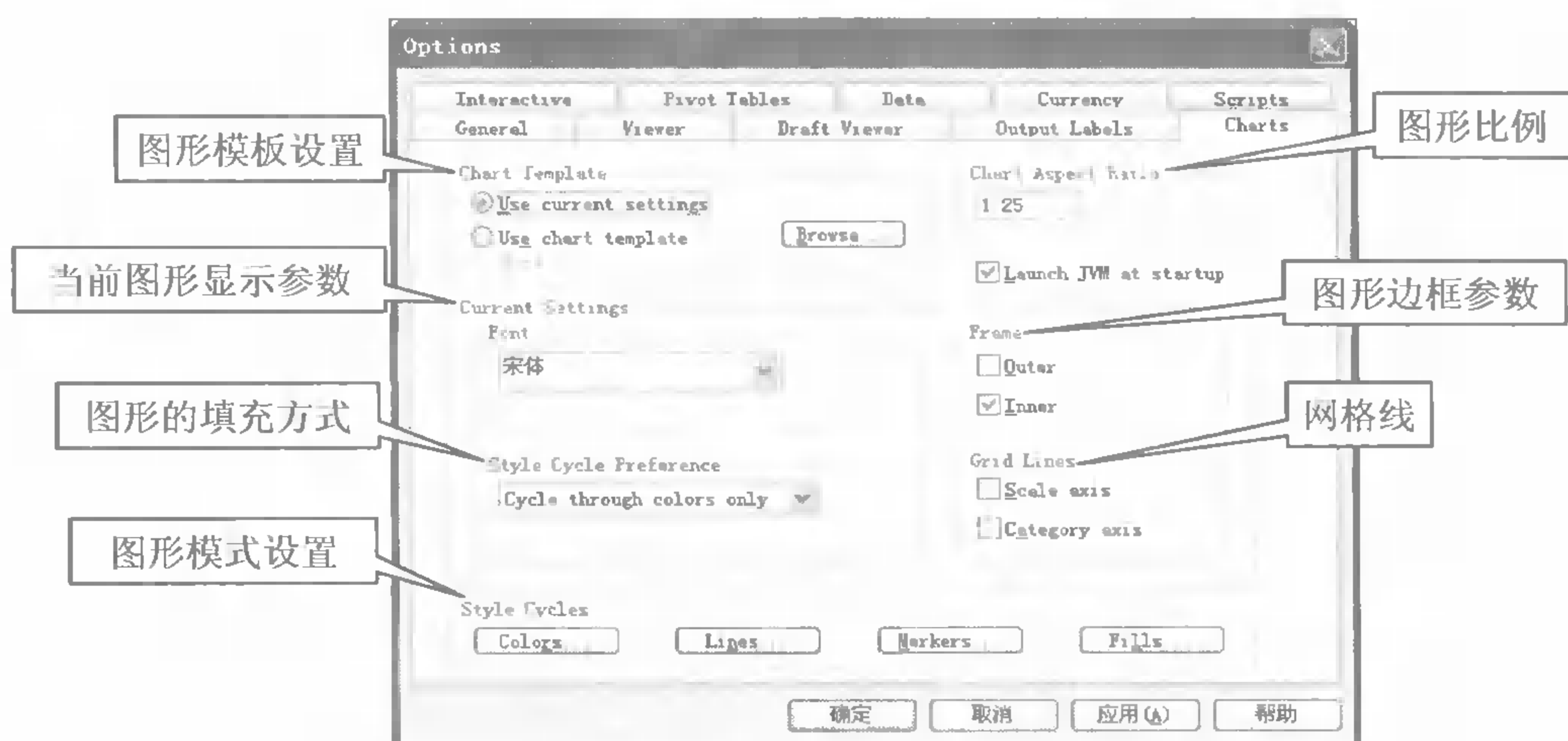


图 1-31 Charts 图形参数设置

### (1) Chart Template (图形模板)。

① 模板选择。单击选中 Use current settings 单选框，表示使用此标签中设置的参数；单击选中 Use chart template 单选框，表示使用指定的图形模板文件中的参数，单击 Browse 按钮选择模板文件。

② 建立模板文件。建立新模板文件的方法是：先在图 1-31 中设置好有关参数，然后在 SPSS View 窗口输出一个图形，双击生成的图形(或在其右键快捷菜单中单击“SPSS Chart Object→Open”)打开 Chart Editor 窗口，如图 1-32 所示。在此编辑图形的显示格式后，依次单击菜单“File→Save chart template...”，就可以保存此图形的显示设置到模板文件。

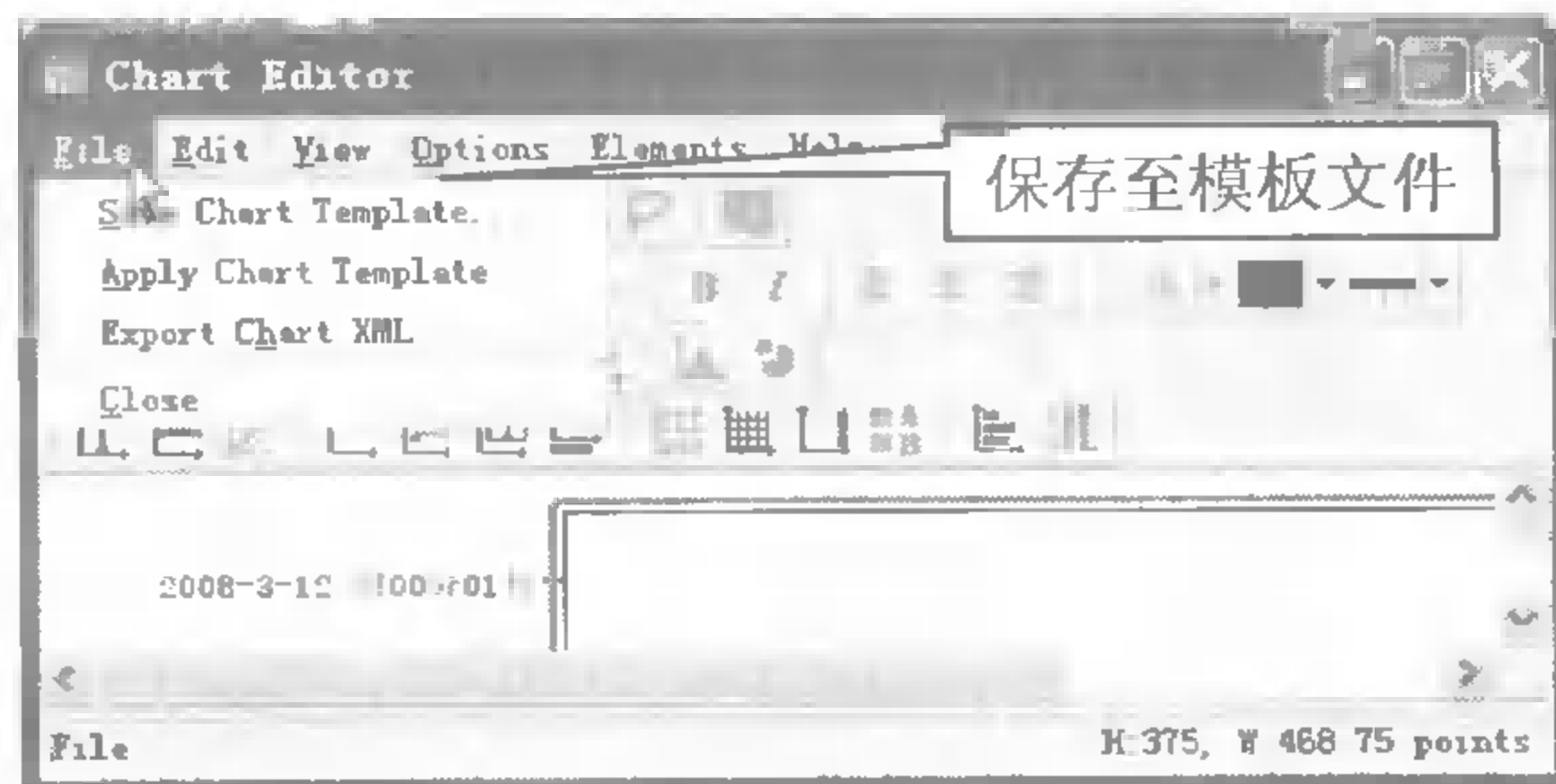


图 1-32 Chart Editor 窗口

(2) Chart Aspect Ratio (图形比例)。此栏设置图形外边框的宽高比例，在输入框中指定数值范围 (0.1~10.0)，小于 1 表示图形显示的高大于宽，反之亦然。

(3) Current Settings (当前图形显示参数)。Font 栏设置新图形中所有文本的格式，在其后的下拉列表选择合适的字体。

(4) Style Cycle Preference 栏设置新生成图形的填充方式，包括区分一个图形中分为多个部分(分组图形 Grouped Chart)时的情况，如复合线型图、复合柱状图等，有如下两个选项：

① Cycle through colors only: 只用不同颜色区分一个图形中的不同元素，不使用模式(线型、值的标识符号、填充方式等)区分。

② Cycle through patterns only: 只用不同模式区分不同元素, 不使用颜色。

(5) Frame (图形边框参数)。可选项有两个: Outer (外边框), 选中后在整个图形 (包含标题、图例等) 的外边加边框; Inner (内边框), 选中后只对作图区域加边框。

(6) Grid Lines (网格线)。可选项有两个: Scale axis 刻度轴 (纵轴), 选中后显示纵轴刻度及水平网格线; Category axis 分类轴 (横轴), 选中后显示横轴刻度及垂直网格线。

(7) Launch JVM at startup (启动时装载 Java 虚拟机)。显示图形特征需要 Java 虚拟机 (Java Virtual Machine, JVM) 的支持, 默认情况下, JVM 随 SPSS 启动自动装载。单击取消此复选框, 能加快 SPSS 的启动时间, 但是在第一次生成图形时会有些慢, 因为这时需要装载 JVM。

(8) Style Cycles (图形模式)。此子设置栏有 4 个按钮: Colors (颜色)、Lines (线型)、Markers (标志) 和 Fills (填充), 分别设置 Style Cycle Preference 栏指定格式内容 (包括颜色和模式) 的具体参数。单击这 4 个按钮打开的对话框布局 and 设置方法完全类似, 下面以 Colors 的设置为例进行介绍。

在图 1-31 中, 单击 Colors 按钮, 弹出如图 1-33 所示的对话框, 设置内容如下:

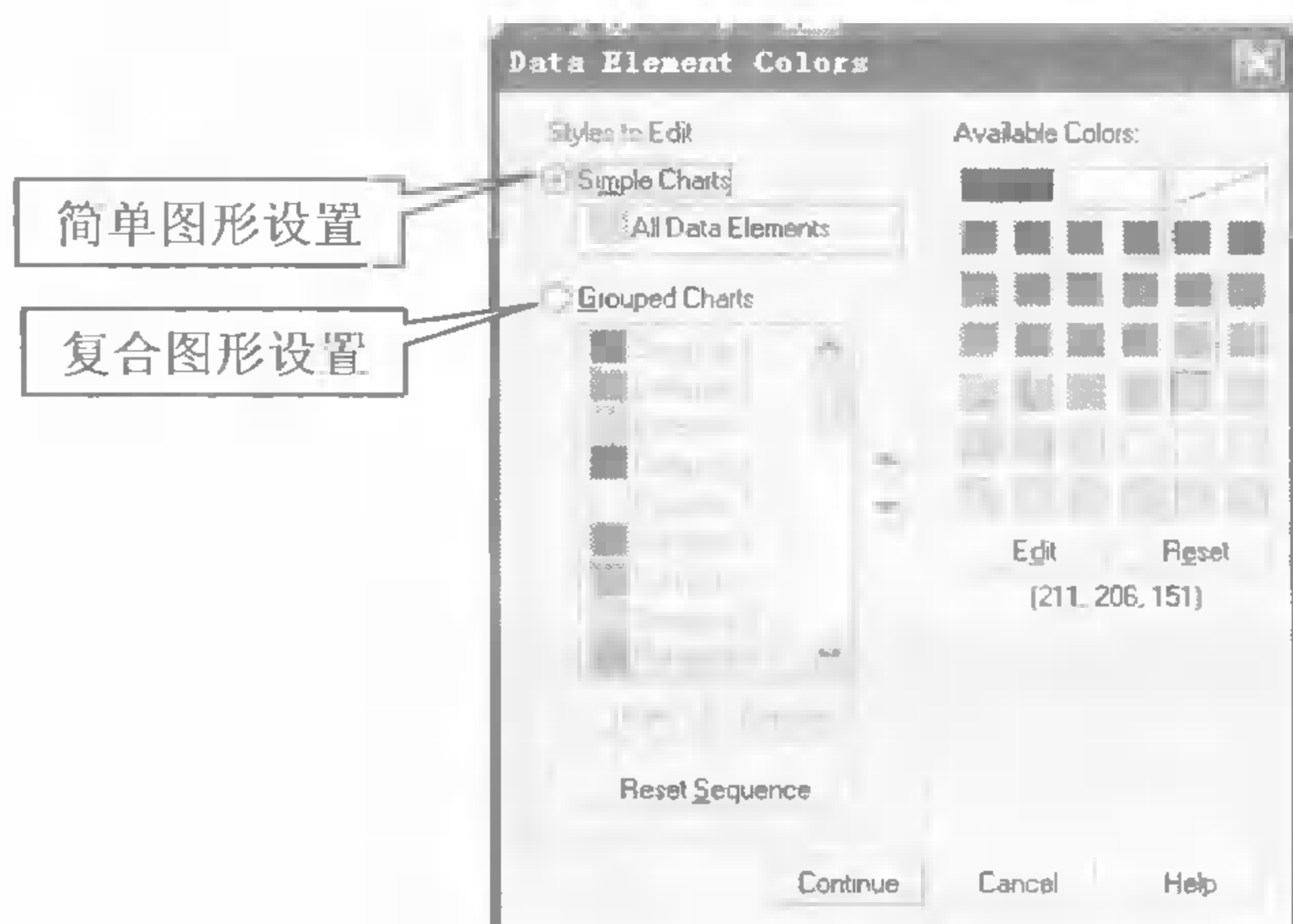


图 1-33 Colors 模式参数

① Simple Charts 单选框设置简单图形的颜色, 选中后在右边的 Available Colors 颜色列表里单击选中一个颜色即可。

② Group Charts 单选框设置复合图形 (分组图形) 的参数, 下面的列表框中给出各类别的名称, 单击选中某个类别后, 在 Available Colors 颜色列表里单击选择一个颜色。下面的 Insert、Remove、Reset Sequence 按钮分别表示插入、删除、重置列表中的类别。

注意: 这里的设置对通过 “Graphs→Interactive” 菜单的子菜单所建立的交互式图形不起作用。

### 1.3.7 Interactive 交互图形窗口参数

Interactive 标签设置 SPSS Viewer 窗口中的 Interactive (交互式) 图形显示参数 (包括外观风格和度量单位等), 如图 1-34 所示。





图 1-34 interactive 标签设置

(1) Chart Look (图表外观风格)。此栏可以从下面的列表单击选中一个外观格式, 此列表默认显示的是 SPSS 安装目录下的外观格式文件, 也可单击 Browse 按钮选择其他的外观格式文件。

如果需要定义新的外观格式文件, 首先进入如图 1-19 所示的 Interactive Graphics Editor, 然后在图形上右击选择 ChartLook 项, 打开如图 1-35 所示的对话框, 再单击 Save As 按钮保存当前图形的外观风格到指定文件。



图 1-35 chart Look 对话框

(2) Data Saved with Chart (图表保存选项)。此栏设置随图形一起保存的数据内容, 可选项有如下两个:

① Save data with the chart 单选框。随图保存所有相关的数据, 这样可以在只打开图形而不打开生成图形的原始数据文件的情况下, 利用随图保存的数据编辑图形, 例如生成新的变量等。虽然这样做很方便, 但它会增加图形文件或 View 视图文件的大小。

② Save only summarized data 单选框。随图形只保存汇总数据, 如总量、总的记录数等。

(3) Print Resolution (显示分辨率选项)。此栏设置交互式图形的显示分辨率。多数情况下, Vector metafile (向量图元文件) 可以显示的更快更好; 而对于 bitmaps (位图), 低分辨率的图表显示更快, 高分辨率的图表显示效果更好。

(4) Measurement Units (度量单位)。此栏设置图形的度量单位, 可选项有 Points (点)、Inches (英尺) 和 Centimeters (厘米)。

(5) Reading Pre-8.0 Data Files (读取旧版本文件选项)。对于旧版 SPSS 的数据文件、其他格式的数据文件和由分析过程生成的新变量, 在此设置数值型变量格式的自动辨认选项。若某个数值变量的不同取值个数少于这里定义的最小值 (由 unique 前的输入框指定), 它将被识别为 nominal 定性变量, 否则为 scale 连续变量。

注意: 此标签的设置除了度量单位外, 其他选项只对通过菜单 “Graphs→Interactive Graphs” 所作的交互式图形起作用。

### 1.3.8 Pivot Table 枢纽表参数

Pivot Table 枢纽表又称为要点表, 能方便地改变表格行、列的交叉显示方式, 其显示参

数设置界面如图 1-36 所示, 可设置的参数包括显示风格、自适应列宽方式和激活方式等。

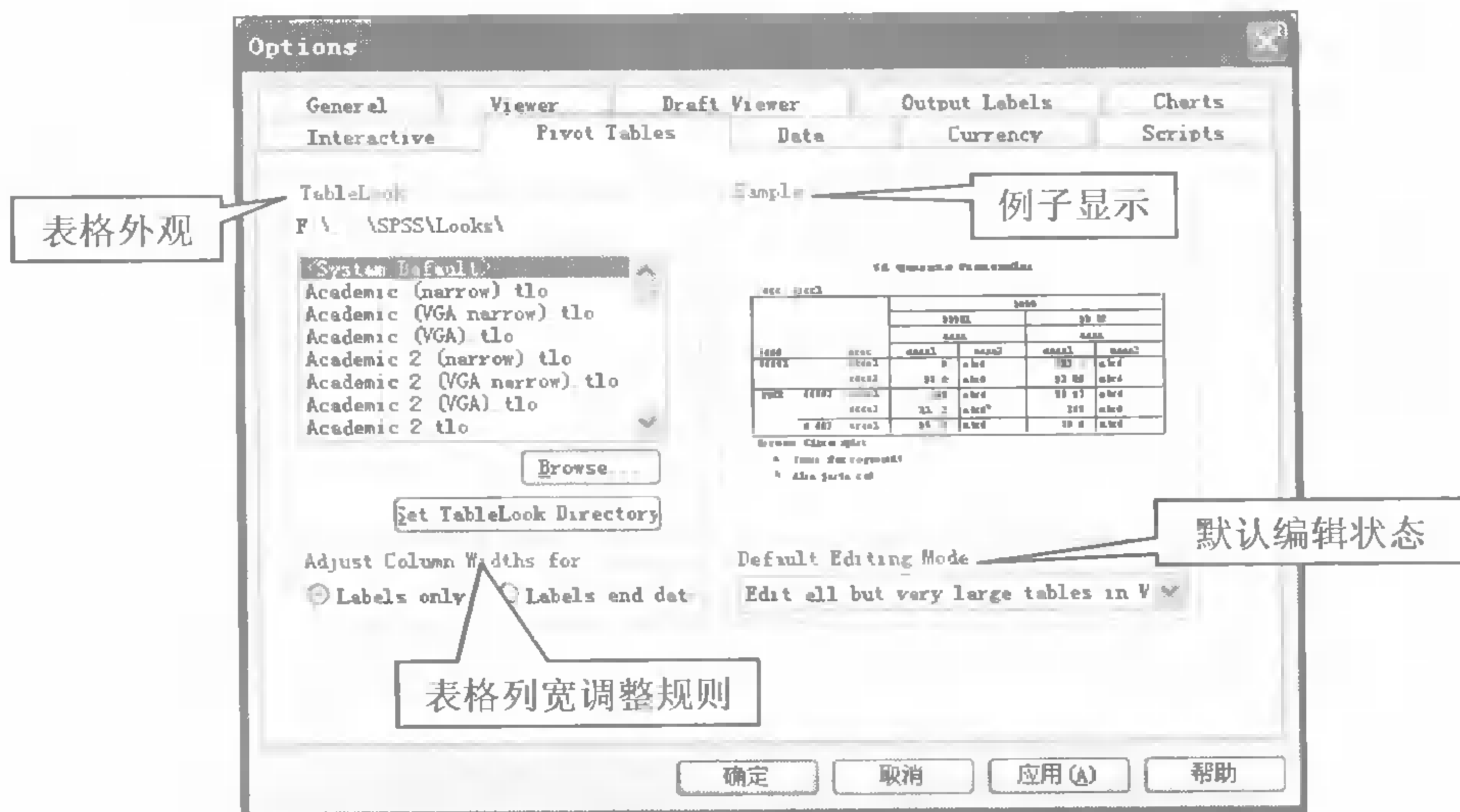


图 1-36 Pivot Table 显示参数设置

(1) Table Look (表格外观)。在下面的列表框单击选中一个外观文件 (\*.tlo), 此列表默认显示 SPSS 安装目录下的外观文件列表。单击下面的 Browse 按钮可以选择用户指定的外观文件。

如果需要定义新的外观文件, 首先在 Viewer 输出窗口双击生成的表格 (或在右键快捷菜单里单击“SPSS Pivot Table Object→Open”项), 打开如图 1-37 所示的 Pivot Table Editor 窗口, 在此编辑表格的显示格式; 然后在图 1-37 中, 依次单击菜单“Format→TableLooks”打开如图 1-38 所示的 Format 保存窗口, 单击 Save As 按钮即可保存当前设置到新的外观文件。

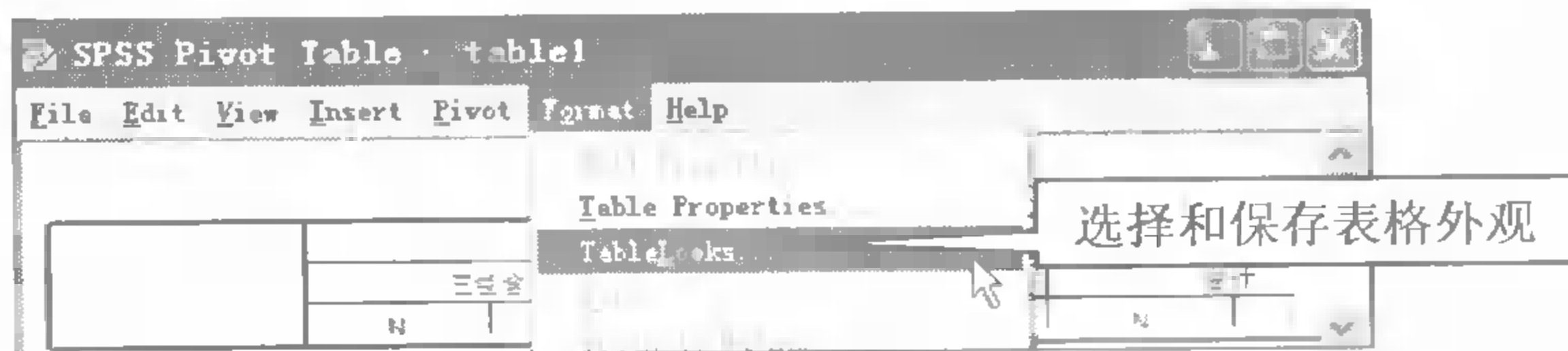


图 1-37 Pivot Table Editor 窗口

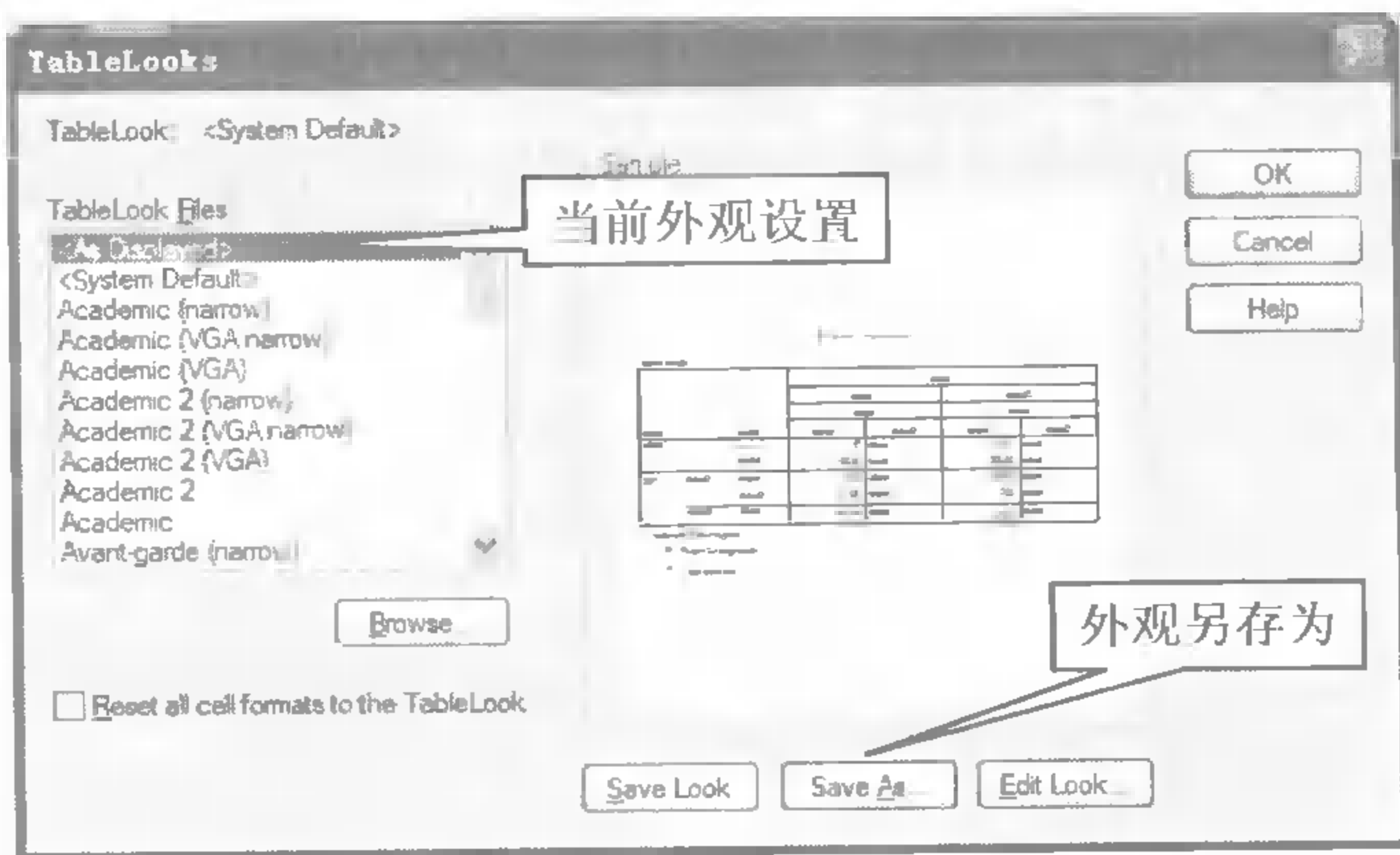


图 1-38 Pivot Table Format 保存窗口

(2) Set TableLook Directory (外观文件列表)。单击此按钮指定一个文件目录, 其中的外

观文件将显示在上面的列表框中。

(3) Adjust Column Widths for (表格列宽调整规则)。此栏设置系统自动调整 Pivot 表格列宽的规则, 可选项有如下两个:

- Labels only (适应标签)。自动调整列宽适合列标签的宽度。此项产生较为紧凑的表格, 比列标签宽的数据值将不被显示出来, 以星号 “\*” 代替。
- Labels and data (适应标签和数据)。自动调整列宽适合列标签宽度和数值宽度中较大的一个。此项产生较为松散的表格, 所有数据值都能正常显示。

(4) Default Editing Mode (默认编辑状态)。此栏设置 pivot tables 在 Viewer 窗口中的激活状态。

默认情况下, 在 Viewer 窗口双击一个 pivot table 会在原始位置嵌入激活状态下的 pivot table。在此栏的下拉列表单击选中 Open all tables in a separate window 项后, 在 Viewer 窗口双击 pivot table 会弹出一个新的编辑窗口。

### 1.3.9 Data 数据参数

Data 标签设置如图 1-39 所示, 在此设置一些数据处理过程 (如 Compute、Count 等) 的更新方式、新变量的显示格式、特定的日期格式以及随机数生成参数等。



图 1-39 Data 数据标签设置

(1) Transformation and Merge Options (转换与合并执行方式)。此栏设置如何执行 Compute、Count、Recode、Read ASCII Data、Add Cases 和 Add Files 命令, 可选择的方式有如下两个:

- Calculate Values Immediately (立即执行方式)  
表示数据转换、文件合并操作在单击“确定”按钮之后立即执行, 此种方式为默认方式。
- Calculate Values before used (延迟执行方式)

表示这些命令直到下次用到相关数据时 (比如用于作图) 才会执行, 在这段保留时间中, 数据编辑器呈失效状态直到这些命令执行完毕。当数据文件很大时选择此项, 可以加快文件中数据的读入速度。通过菜单 Transform 所作的计算不会立即显示在 Data Editor 窗口, 通过依次单击菜单 “Transform→Run Pending Transforms”, 可以立即执行那些未进行的操作。

(2) Display Format for New Numeric Variables (新变量的显示格式)。此栏设置新变量在

数据编辑器中的默认显示格式，在 Width 后指定显示宽度，在 Decimal 后指定小数位数。当数据过长时会自动转为科学计数法，其小数部分仍按照此处的设置显示。此栏对计算时使用的数据精度无影响。

(3) Set Century Range for 2-Digit Years (百年日期格式)。此栏设置日期格式以两个字符显示时，默认的世纪年代（如：10/28/86, 29-OCT-87）。

Automatic 单选框，自动设定系统当前时间向前 69 年，向后 30 年，加上当前时间年代一共 100 年，代表默认的一个世纪。

Custom 单选框，在 Begin year 后输入代表一个世纪的起始年份，End year 后会随输入的起始年份自动延后 100 年。

(4) Random Number Generator (随机数生成器)。

Compatible with SPSS 12 and earLier 单选框。采用 SPSS 12 或之前版本的随机数生成器，如果需要用同样的随机数生成种子 (seed) 生成同一组随机数，以验证以前的某些结论，需要单击选中此项。

Long period Mersenne Twister 单选框。一个新的更加可靠的随机数生成器。

### 1.3.10 Currency 数值型变量格式参数

Currency 标签设置如图 1-40 所示，在此设置常用数值型变量的格式。Custom Output Formats 列表中给出了五种自定义格式，分别命名为 CCA、CCB、CCC、CCD 和 CCE，这些名字不可更改，也不可以添加新的格式，单击选中某个格式名称，设置内容如下：

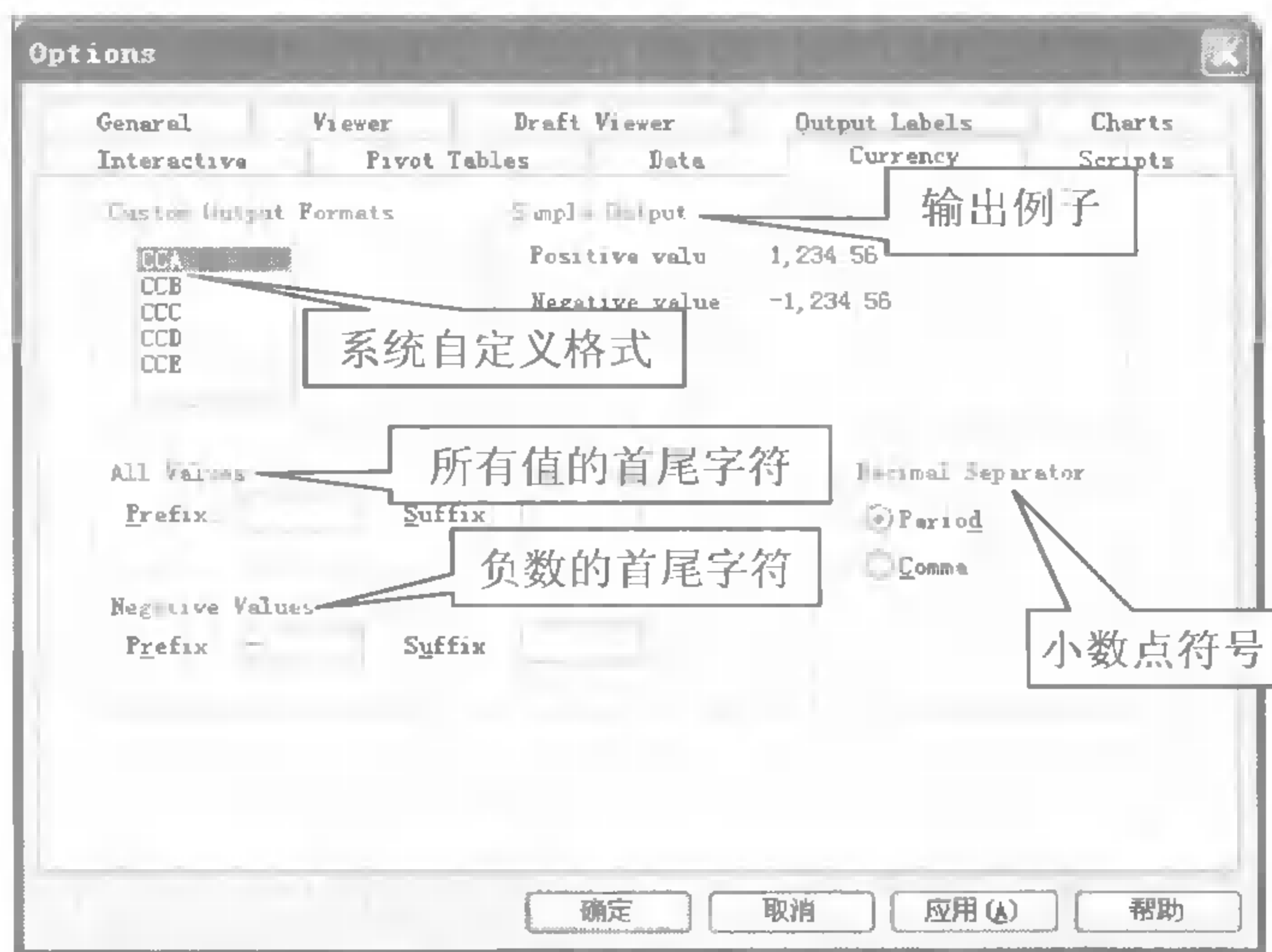


图 1-40 Currency 数值型变量格式标签设置

(1) All Values 子设置栏。此栏设置所有数值的首尾字符，在 Prefix 后面填写前缀字符，默认为空格；在 Suffix 后面填写后缀字符，默认也为空格。

(2) Negative Values 子设置栏。此栏设置负数的首尾字符，在 Prefix 后面填写前缀字符，默认为“-”；在 Suffix 后面填写后缀字符，默认也为空格。

(3) Decimal Separator 子设置栏。此栏设置小数点的符号为 Period (圆点) 或者 Comma (逗号)。

以上参数设置完毕，在 Sample Out 栏会有数字样例的显示，Positive value 是正数样例，



Negative value 是负数样例。选择某个格式后，单击确定按钮，返回 SPSS 主画面，所选格式就可以在定义数值型变量时使用了。

### 1.3.11 Scripts 脚本编辑窗口

Scripts 标签如图 1-41 所示，在此设置一些与 Scripts 脚本运行相关的参数。对于熟悉 Sax Basic 语言以及致力于提高 SPSS 自动化处理能力（比如定值 pivot 表格的显示方式）的用户，可以在此选择相关的参数。

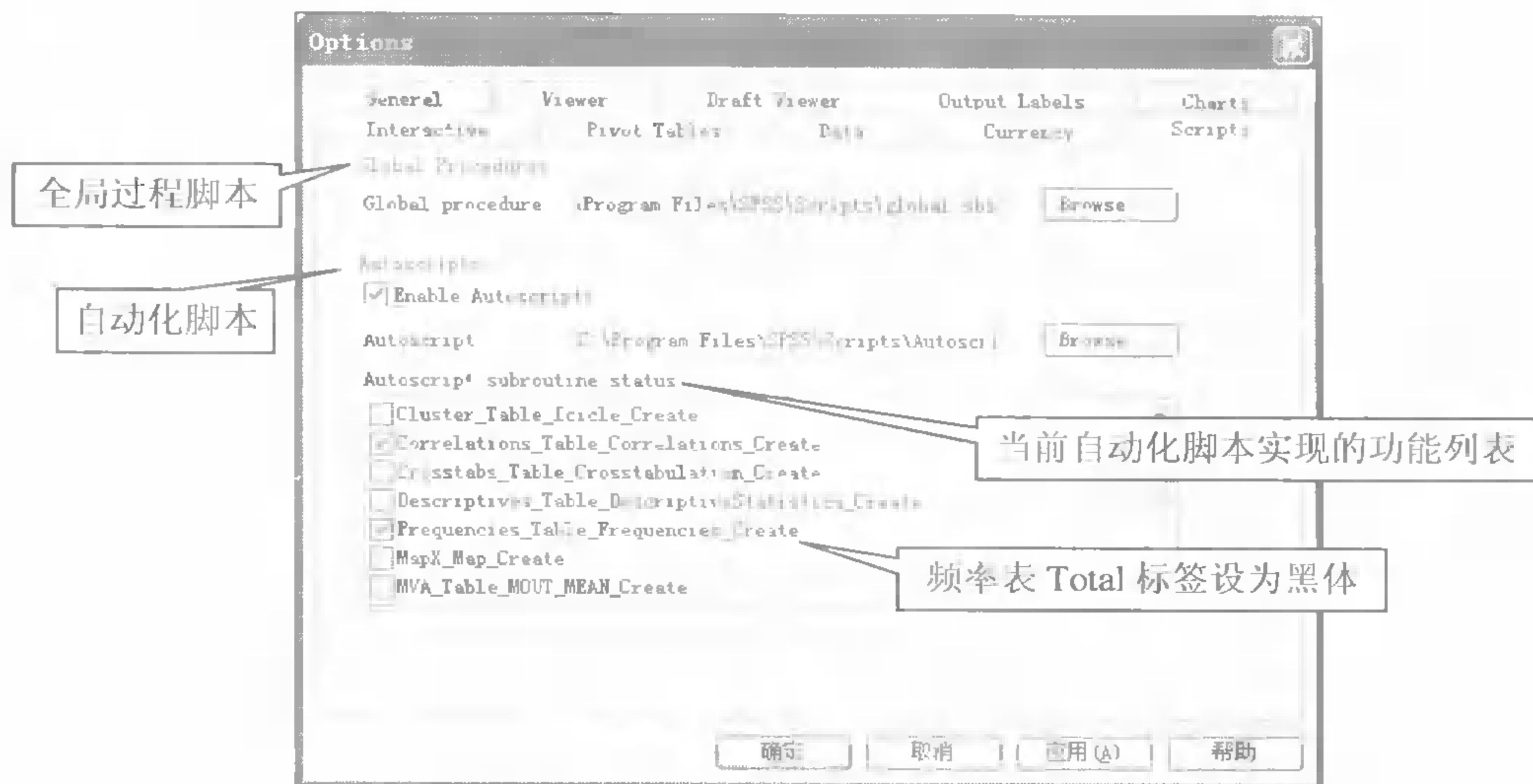


图 1-41 Scripts 脚本编辑窗口参数设置

(1) Global Procedures（全局过程脚本）。此栏指定一个可以被所有 Scripts 脚本（包括自动化脚本）调用的脚本和函数文件库。系统默认的 Global Procedures 文件包含了许多其他过程常用的脚本，更改后可能导致某些功能无法正常工作。

(2) Autocscripts（自动化脚本）。自动化脚本是指当生成某个输出对象时自动运行的脚本文件，它用来实现改变对象的显示格式等功能。

单击选中 Enable 复选框激活自动执行脚本的功能，在 Autocscript 后输入要执行的自动化脚本文件路径（或单击 Browse 按钮选择脚本文件），在下面的列表框就会显示所选脚本可以实现的功能，选中其中的选项，单击应用按钮，即可实现其功能。下面举两个例子：

- ① Frequencies\_Table\_Frequencies\_Create。表示运行 Frequencies 过程并建立频率表后，自动转到频率表格的行标签，把“Total”标签所在的行中的内容设置为黑体。
- ② Crosstabs\_Table\_Crosstabulation\_Create。表示运行 Crosstabs 过程并建立交叉表后，自动将“Total”标签所在的行、列中的相关内容设置为黑体，并且将标签名设置为百分比显示。



# 第 2 章 数据文件的建立与操作

本章介绍数据的准备和预处理工作以及文件导入导出操作，这是整个数据分析过程的开始，熟练掌握了本章内容后，就能充分地发挥和应用 SPSS 强大的分析功能。

## 2.1 数据编辑器与数据文件

数据编辑器（Data Editor）是做统计分析最主要的窗口界面，在此可以观察、录入和编辑数据，或者导入其他格式的数据文件。而且，通过此界面可以执行所有的数据处理和统计分析过程。

### 2.1.1 数据编辑器

启动 SPSS 后，最先见到的就是数据编辑窗口（Data Editor），如图 2-1 所示。在此可对数据和变量进行各种操作，下面详细介绍此界面的各个部分。

#### 1. 标题栏

如图 2-1 所示，标题栏右侧显示“SPSS Data Editor”，表明这是数据编辑窗口。“TEST0.sav”是当前数据文件名，前面的“\*”表示对此文件有修改操作但尚未保存，保存后“\*”消失，做过修改“\*”再度出现，如此反复可以提示用户及时保存文件，以免意外发生时造成数据丢失。紧随文件名的“[DataSet1]”是当前数据集的名字，若再打开或新建一个数据文件就会显示“[DataSet2]”，依次类推，数据集名在各种分析的结果输出中会有所显示，以标识当前分析的数据对象。另外，在编写程序语句时，利用数据集名能够方便地同时处理多个数据集。

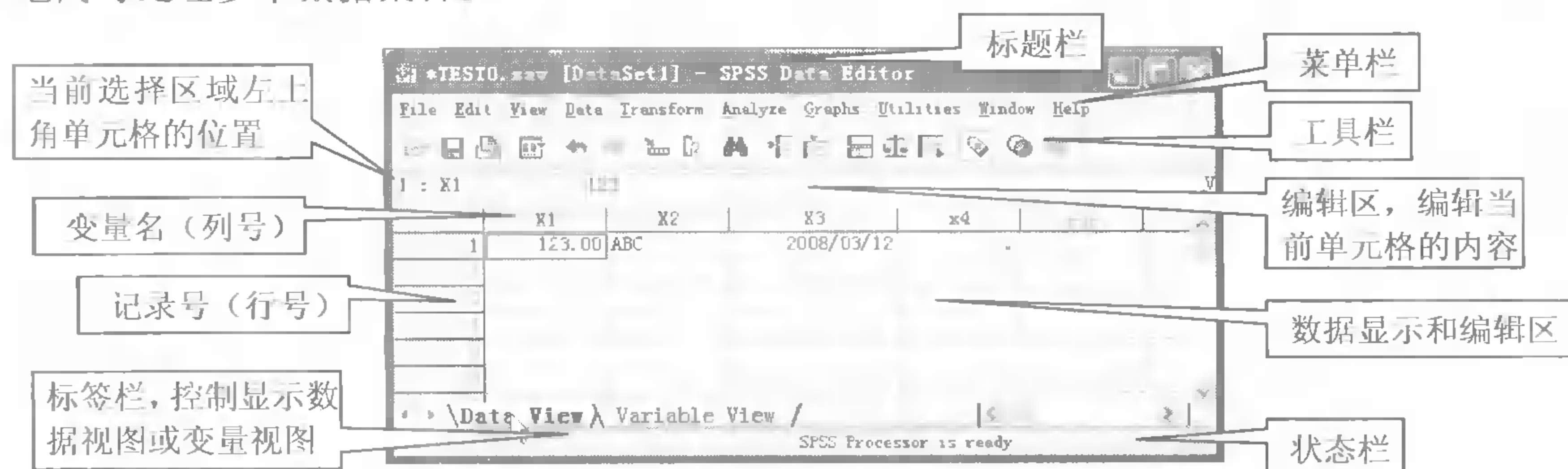


图 2-1 数据录入

## 2. 菜单栏

标题栏下是菜单栏，其各选项的功能如表 2-1 所示。

表 2-1 数据编辑器各菜单功能

File 文件	打开或保存文件、查看文件信息、打印、查看历史打开文件记录等，详细选项参考图 2-22 所示
Edit 编辑	剪切、复制、粘贴等编辑操作，查找、定位特定行的数据，插入数据或变量，在“Edit→Options...”选项卡中可以设置 SPSS 诸多环境变量
View 视图	设定数据编辑窗口的显示方式，可以选择是否显示状态栏、表格线，设置工具栏显示方式。数据表格区的字体，选择当前显示数据窗口或者变量窗口
Data 数据	设置变量性质，定义日期变量，进行转置、排序，分割文件等操作
Transform 转换	用已有变量生成新变量，生成时间序列变量、随机数，缺失值操作等
Analyze 分析	实现各种分析功能，比如 OLAP、Regression、Classify 等
Graphs 图形	做一般图形、交互图形、地图等
Utilities 功能	查看变量信息，运行脚本等
Windows 窗口	切换窗口，改变窗口显示方式
Help 帮助	显示图形化教程帮助、操作帮助、算法帮助等，还可以查看版本、注册信息，链接至 SPSS 官方网站，更新软件等

## 3. 工具栏

菜单栏下面是工具栏，依次单击菜单“View→Toolbars...”打开工具栏的选项设置界面，如图 2-2 所示。单击 Document Type 下边的下拉列表，选中 Data Editor（数据编辑窗口）选项，也可以选择设置其他窗口（如 Viewer，视图窗口）的工具栏显示内容和方式。

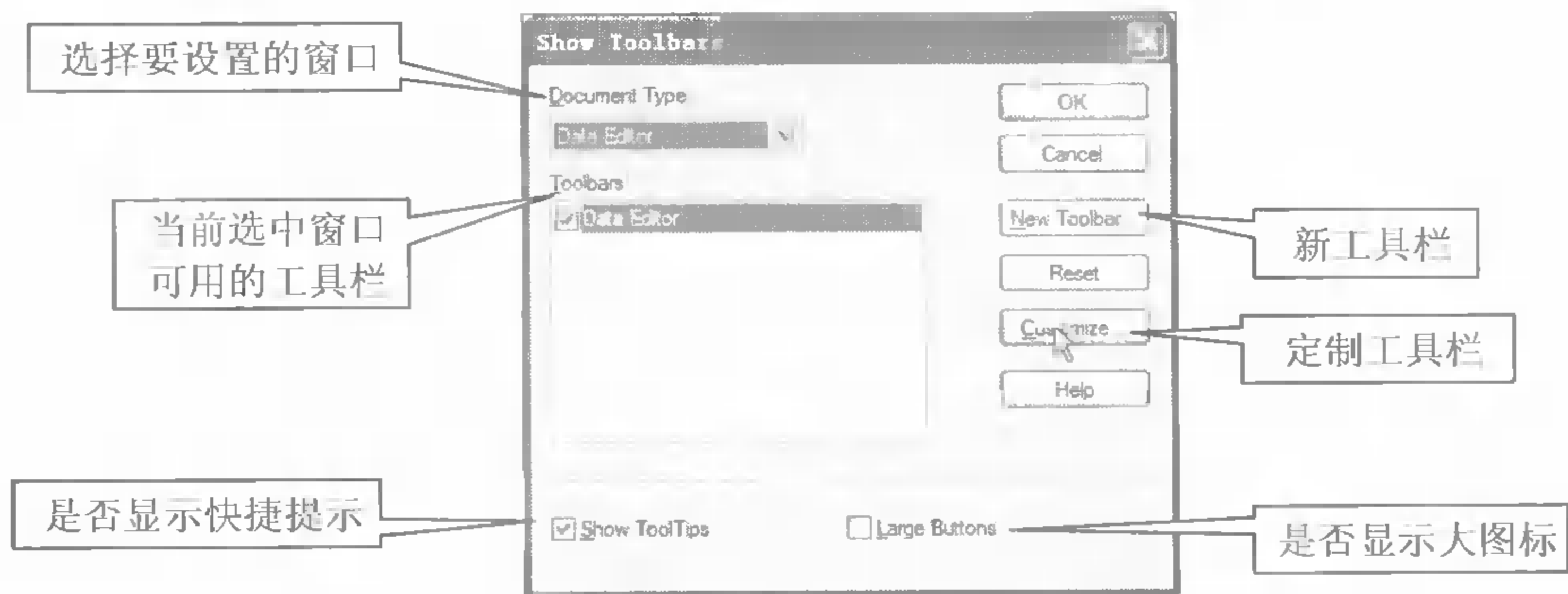


图 2-2 定制工具栏

底部的 Show ToolTips 复选框表示显示快捷提示，即鼠标停留在工具栏某个项目时，自动显示相应的功能提示。Large Buttons 复选框表示在工具栏显示大图标。

单击 New Toolbar 按钮新建工具栏；单击 Reset 按钮恢复默认设置；单击 Customize 按钮编辑当前工具栏。

在图 2-2 中，单击 New Toolbar 按钮，弹出如图 2-3 所示的新建工具栏对话框。在 Toolbar Name 后输入新工具栏的名称“Mytoolbar”，下面的几个复选框表示此工具栏将在哪些窗口显示；单击 Customize 按钮，弹出如图 2-4 所示的工具栏编辑界面，而直接在图 2-2 中单击 Customize 按钮，也会弹出如图 2-4 所示的对话框。

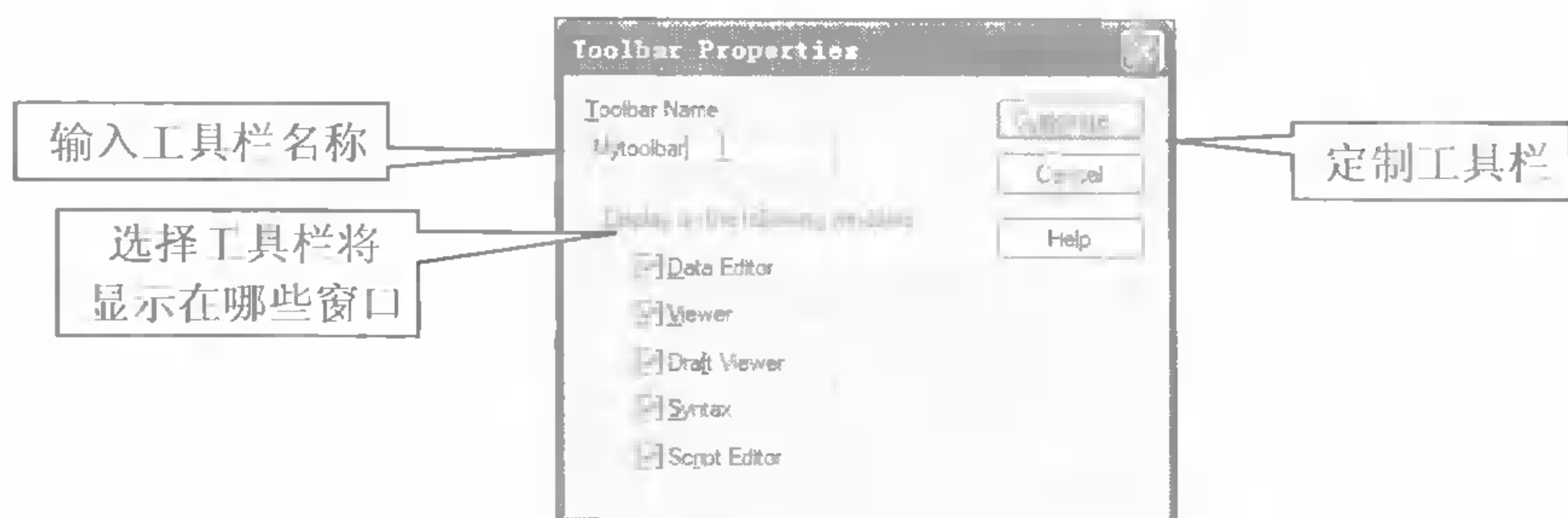


图 2-3 新建工具栏

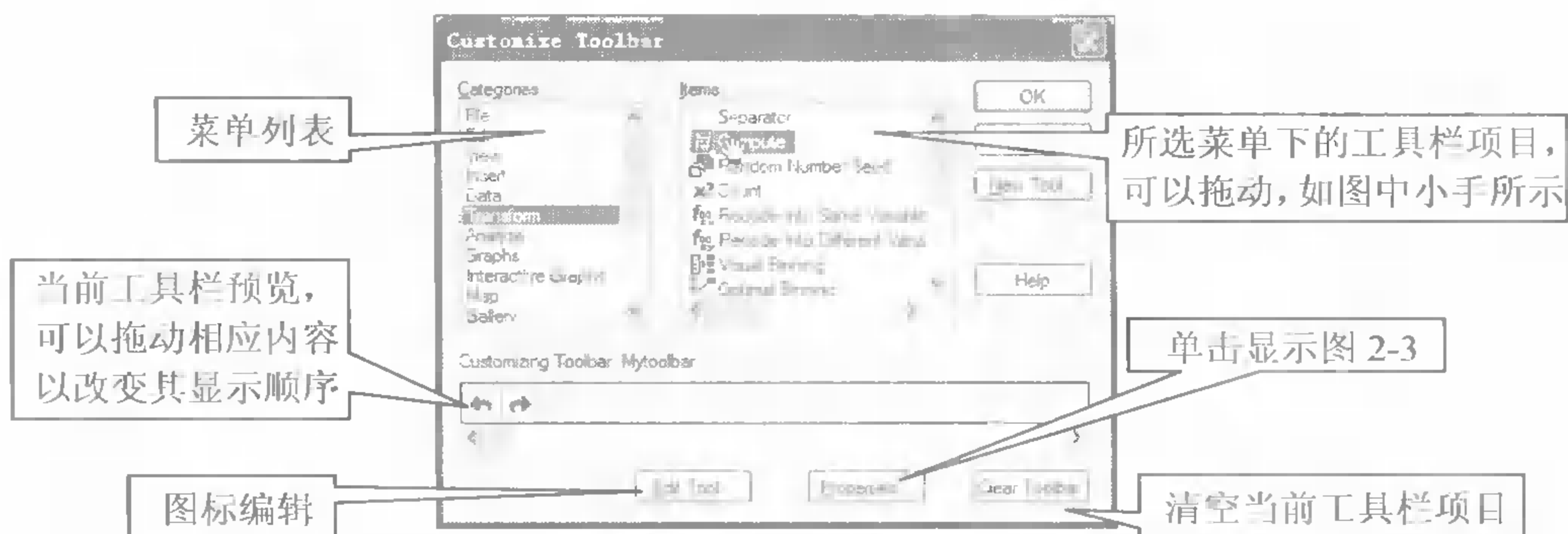


图 2-4 编辑工具栏

在图 2-4 中, 通过鼠标拖放操作可以设置工具栏的显示内容, 左侧的菜单列表显示了各个功能菜单; 右侧的项目列表显示了所选菜单下的工具栏项目, 鼠标指向某个项目会自动变为小手形状, 单击左键并保持左键按下状态, 小手变为握紧形状, 移动鼠标, 将相应项目拖放至对话框底部的工具栏预览区中; 预览区显示当前编辑工具栏的项目, 通过拖放可以改变其项目的显示顺序。单击 Edit Tool 按钮, 弹出图形编辑窗口, 编辑选中按钮的显示图标; 单击 Properties 按钮, 返回图 2-3 所示的界面; 单击 Clear Toolbar 按钮, 清空预览区中的项目。单击 OK 按钮返回 Data Edit 窗口, 原来工具栏下方就会出现新建的工具栏, 如图 2-5 所示。



图 2-5 显示新建工具栏

#### 4. 数据编辑和显示区域

如图 2-1 所示, 在工具栏的下方, 左侧灰色显示的是提示区, 其内容为当前选中单元格的位置, 例如“1:X1”表示第一行(第一条记录)的变量 X1, 如果当前选中的是连续区域, 此处显示所选区域的左上角单元格的位置; 右侧的空白区域是数据编辑区, 显示当前选中单元格的内容, 可以在这里输入或编辑数据内容, 回车后显示在选中的单元格中。

下面的二维表格是数据的显示和编辑区, 顶部显示变量的名字, 左侧显示记录行数, 它的显示风格和操作方式都与 Microsoft Excel 相似。

#### 5. 标签栏

二维表格的下面是 Data Editor 窗口的标签栏, 单击 Data View 标签可以显示 Data 数据窗

口：单击 Variable View 标签可以显示 Variable 变量窗口，如图 2-6 所示，在此显示和编辑变量属性。另外，双击 Data View 窗口中二维表格顶部的某个变量名，可以自动切换到 Variable View 窗口，且双击的变量名被选中。

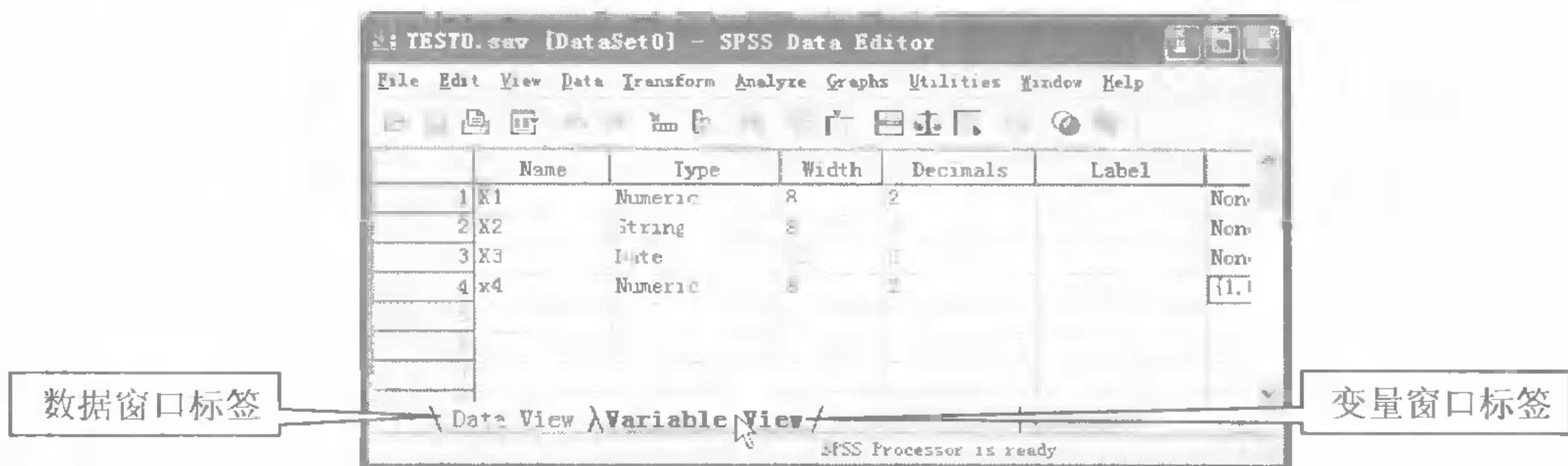


图 2-6 Variable View 标签

### 2.1.2 数据文件

数据文件的操作大多通过 File 菜单实现。在 Data Viewer 视图里，依次单击菜单“File→Save As”打开另存为对话框，如图 2-7 所示，单击保存类型后的下拉菜单，显示出 SPSS 可以保存和读取的文件格式，具体包括 SPSS 数据文件 (\*.sav)、ASCII 码文件 (\*.dat)、Microsoft Excel 文件 (\*.xls) 以及 SAS、Lotus1-2-3、dBase、Stata 等格式的数据文件。另外，通过链接数据库功能，能打开 ODBC 数据源指向的数据库文件，详细操作请参见第 2.5.2 节。



图 2-7 文件另存为对话框

## 2.2 常量、变量、操作符和表达式

本节介绍 SPSS 的各种数据类型和操作符。常量、变量、操作符和表达式是数据操作最基本的概念，在建立数据文件、生成新变量时都要用到，而且他们是 Syntax 命令语句的重要组成部分，只有熟悉了它们，才能更准确有效地进行统计分析。



### 2.2.1 常量与变量

常量就是在一定阶段取值保持不变的量，如圆周率。变量也称为属性，不同的对象（观测行，或记录行）取不同的值。下面分别介绍它们的数据类型和数据格式。

#### 1. SPSS 常量

SPSS 中的常量包括：数值型数字、一个括在单（双）引号中的字符串、按特定格式表示的日期时间，简称为数值型、字符型和日期型。

- 数值型常量数值型常量就是数字，有两种书写方式：普通书写方式，如 123、123.45；科学计数法，多用于表示特别大或特别小的数字，如  $1.23\text{E}10$  表示  $1.23 \times 10^{10}$ ， $1.23\text{E}-11$  表示  $1.23 \times 10^{-11}$ 。
- 字符串常量字符串常量是用单引号或双引号括起来的一串字符，如果其中包含“”（不包括外部的双引号），则该字符串常量必须使用双引号括起来，如：“Jack’s pen”。
- 日期型常量按特定格式给出的确定日期或时间，如：28-OCT-90、10/28/90。

#### 2. 变量及其属性

图 2-8 所示给出了 SPSS 中各种变量所对应的图标。在变量设置、数据分析设置窗口中常常可以在变量名旁边看到这些图标提示，从而方便地判断所选变量的格式。

Measurement Level	Data Type			
	Numeric 数值型	String 字符串型 n/a	Date 日期型	Time 时间型
Scale 数值范围				
Ordinal 序数的				
Nominal 名义的 (无顺序)				

图 2-8 变量图标

反映变量性质的有变量名、变量类型、变量标签与值标签以及变量格式。

##### (1) 变量名。

- 每个变量名都是唯一的，变量名不能复制到其他变量上去。
  - 变量名中不可以有空格。
  - 变量名允许有 64 字节长，首写必须是 24 个英文字母或以下符号之一：@、#或\$。
- 除了首写字符外，随后的字符可以是字母、数字、小数点或其他任意非标点符号的字符。64 个字节意味着可以是 64 个单字节字符，如英文、法文、德文、西班牙文、意大利文、俄文、希腊文等，或者 32 个双字节字符，如中文、日文、朝鲜文等。
- 以“#”为首写的变量名特指草稿型变量。此种类型的变量只能在命令窗口使用，在其他地方不可用。
  - 以“\$”为首写的变量名特指 SPSS 的系统变量。系统变量不可修改，而且在程序中不可用，用户定义的变量不能以“\$”为首写。系统变量存储了 SPSS 运行时的许多参数信息，例如：系统缺失值、系统读取的记录行数、当前系统时间等。常见系统



变量如表 2-2 所示。

表 2-2 系统变量

\$CASENUM	当前记录号
\$SYSMIS	系统缺失值
\$JDATE	系统日期的数字表示，即与 1582 年 12 月 14 日的差
\$DATE	系统日期，格式：dd-mmm-yy
\$DATE11	系统日期，格式：dd-mmm-yyyy
\$TIME	系统时间
\$LENGTH	当前页面长度
\$WIDTH	当前页面宽度

- 避免以 “.” 作为变量名结尾。因为英文句点有时会作为命令的结束标志，定义这样的变量容易引起歧义。只能在 Syntax 语句中定义以英文句点结尾的变量名。
- 避免以 “\_” 作为变量名结尾。因为下划线一般作为由程序或命令自动生成的变量名的结尾，为避免命名冲突，最好不要以下划线作为用户定义变量的结尾。
- 变量名不能与 SPSS 的保留字相同。SPSS 的保留字有 ALL、AND、BY、EQ、GE、GT、LE、LT、NE、NOT、OR、TO、WITH。如果使用了上述保留字作变量名，系统会自动提示。
- 不区分变量名的大写和小写。
- 设置变量的标签名 Variable Label。Variable Label 可以显示在输出窗口，便于查看结果时理解变量的实际意义。如果变量名不能充分反映它所代表的含义，需要额外说明时，或者有必要使用更长的变量名时，都可以用 Variable Label 代替。

(2) 变量类型。

常用变量类型包括数值型、字符型、日期型 3 种，分别介绍如下。

① 数值型变量。数值型变量的长度指变量值所占的字符数，即用字符个数度量的数字宽度，小数点和其他分界符也计算在内。常用的数值型变量有以下几种写法。

- Numeric：标准数值型变量，默认长度为 8，小数位数为 2。
- Comma：带逗号的数值型变量，默认长度为 8，小数位数为 2。显示时整数部分自左向右每隔 3 位用逗号作分隔符，用圆点作小数点。
- Dot：圆点数值型变量，默认长度为 8，小数位数为 2。显示时整数部分自左向右每隔 3 位用圆点作分隔符，用逗号作小数点。
- Scientific Notation：科学计数法，默认长度为 8，小数位数为 2。对于数值很大或很小的变量可以使用科学计数法，输入时表示指数的字母可用 E 也可用 D。下面几种方式都可以被接受：123、1.23E2、1.23D2、1.23E+2、1.23+2。
- Dollar：带美元符号的数值型变量，默认长度为 8，小数位数为 2。其值在显示时，有效数字前面带有 “\$”，输入时可以不输入 “\$”，显示时系统自动加上 “\$” 和分隔符。

对于上述几种数值型变量，输入的小数位超过规定个数时会自动四舍五入。

- Custom Currency：自定义类型。这样的定义只能在命令窗口使用，在有些地方（如生成新变量的对话框）不可用。

数值型变量的格式表示及例子如表 2-3 所示。

表 2-3 数值型变量的格式样例

格式名	描述	样本格式	样本输出	固定输入		自由输入	
				格式	取值	格式	取值
Fw,d	标准数值型	F5.0	1 234	F5.0	1 234	F5.0	1 234
			1.234		1		1
		F5.2	1 234	F6.2	12.34	F6.2	1 234.0
			1.234		1.23		1.23
Nw,d	限制数值型	N5.0	00 123	F5.0	123	F5.0	123
			123				123
		N5.2	12 345	F6.2	123.45	F6.2	12 345
			12.34				
Ew,d	科学计数法	E8.0	1234E3	E10.3	1.234E+06	E10.3	1.234E+06
			1 234		1.234E+03		1.234E+03

② 字符串型变量。字符串变量在使用时，应注意以下几点。

- 字符串中可以包含数字、字母、特殊字符，最长为 32 767 个字符。
- SPSS 区分长字符串和短字符串，一个短字符串最长 8 字节，一个长字符串大于等于 8 字节，长字符串变量不能定义用户缺失值。有些分析过程可以处理短字符串，但不能处理长字符串。
- 系统缺失值不能用于生成字符串变量。
- 当通过转换操作（transformation）或其他过程生成新变量，或者修改了原有变量时，可能产生缺失值或未定义的变量值，这时系统自动赋值为空。变量值以空格表示时，若无特别定义，不能代表缺失值。
- 字符型变量不能参与算术运算。
- 字符串中的大写字母与小写字母，是截然不同的两个字符，这一点在使用时要特别注意，建议用户使用短字符串变量。

③ 日期型变量。SPSS 中的日期型变量（Date）既可以表示日期，也可以表示时间，如表 2-4 所示。

如表 2-4 所示的后两列，可以熟悉常用的日期和时间格式。

表 2-4 日期时间型变量的样例

格式名称	通用格式表述	例子
DATEw	dd-mmm-yy	28-OCT-90
	dd-mmm-yyyy	28-OCT-1990
ADATEw	mm/dd/yy	10/28/90
	mm/dd/yyyy	10/28/1990
EDATEw	dd.mm.yy	28.10.90
	dd.mm.yyyy	28.10.1990
JDATEw	yyddd	90301
	yyyyddd	1990301
SDATEw	yy/mm/dd	90/10/28
	yyyy/mm/dd	1990/10/28

续表

格 式 名 称	通用格式表述	例 子
QYRw	q Q yy	4 Q 90
	q Q yyyy	4 Q 1990
MOYRw	mmm yy	OCT 90
	mmm yyyy	OCT 1990
WKYRw	ww WK yy	43 WK 90
	ww WK yyyy	43 WK 1990
WKDAYw	(name of the day)	SU
MONTHw	(name of the month)	JAN
TIMEw	hh:mm	01:02
TIMEw.d	hh:mm:ss.s	01:02:34.75
DTIMEw	dd hh:mm	20 08:03
DTIMEw.d	dd hh:mm:ss.s	20 08:03:00
DATETIMEw	dd-mmm-yyyy hh:mm	20-JUN-1990 08:03
DATETIMEw.d	dd-mmm-yyyy hh:mm:ss.s	20-JUN-1990 08:03:00

首先给出关于日期型变量格式的几点说明。

- “dd” 是用两位数来表示日期数;
- “ddd” 是用 3 位数来表示从元月一日算起的日数;
- “mm” 是用数字表示的月份数;
- “mmm” 是用英文月份单词的前 3 个字母表示的月份;
- “yy” 是用两位数来表示的年份;
- “yyyy” 是用四位数来表示的年份;
- “hh” 表示小时; “mm” 表示分钟; “ss” 表示秒;
- “m” 用于年与日 (字母 y 与 d) 之间时表示月份; 用于时与秒 (字母 h 与 s) 之间时表示分钟。

指定了日期型变量的格式后, 输入时不一定按指定的格式输入, 可以用 “/” 或 “-” 作为具体日期的分隔符, 回车后系统会自动转换成指定的格式。

(3) 变量标签与值标签。标签应用于变量名和变量取值的辅助说明, 相当方便。

① Variable Labels (变量标签)。变量标签是对变量名的进一步说明, 当变量名较短时, 自身字符数不足以表明其具体含义, 而且当变量比较多时更需要对变量名的含义加以详细解释, 变量标签就起到这样的作用。在统计分析的输出结果中, 可以在与变量名相对应的位置显示该变量的标签, 或者直接以变量标签替代变量名显示, 这有助于理解和分析输出结果。

如果 SPSS 运行在非中文平台上, 不熟悉外文的用户可以给变量名附加中文标签, 这会使统计结果的观察和分析更加方便, 例如表 2-5 所示。

② Value Labels (变量值标签)。变量值标签是对变量取值所做的进一步说明, 分类变量经常需要定义其取值的标签。变量值标签是一个可选择的属性, 可以定义, 也可以不定义。典型的例子就是性别变量, 例如表 2-6 所示。

表 2-5 变量标签示例

变 量 名	对应的变量标签
h	height
w	weight
g	性别
a	年龄

表 2-6 变量值标签示例

变 量 名	值	值 标 签
gender	1	男
	0	女

(4) 变量格式。

变量格式所包含的主要设置内容有如下 3 项。

① 宽度。此处的宽度指在数据编辑窗口中该变量所占的列数。

用户需要明确区分定义变量类型时指定的长度与定义格式时的宽度。在定义变量格式的宽度时，要综合考虑变量类型所定义的长度和变量名所占的宽度，选择较大的一个作为该变量的格式宽度，这样才能保证变量名和变量值都可以正常显示。

② 对齐方式。对齐方式有 3 种：左对齐、右对齐和中间对齐。一般情况下，数值型变量默认的对齐方式为右对齐；字符型变量默认的对齐方式为左对齐；用户可以指定中间对齐方式。

③ Missing Value（缺失值）。在实际工作中常会因为某种原因，出现记录数据失真、没有观测到或没有记录到等数值缺失现象。SPSS 允许用户使用默认的缺失值，或定义自己的缺失值标记。

2.2.2 操作符与表达式

SPSS 的基本运算有 3 种：数学运算、关系运算、逻辑运算。相关的运算符表示方法如表 2-7 所示。

表 2-7 操作符

数学运算操作符		关系运算符		逻辑运算符	
+	加	<(LT)	小于	& (And)	与
-	减	>(GT)	大于	→(Or)	或
*	乘	<=(LE)	小于等于	~ (Not)	非
/	除	>=(GE)	大于等于		
**	幂	1=(EQ)	等于		
()	括号	~=(NT)	不等于		

(1) 数学运算符与算术表达式。数学运算符也就是常用的算术运算符，可以连接数值型的常量、变量、函数，形成算术表达式，运算结果通常为数值。运算符的优先级为：括号>函数>乘方(幂)>乘或除>加或减，同一优先级的符号，位于左侧的优先级较高。例如：  
x+2\*(3-y)-abs(z)/5。

(2) 关系运算符与比较表达式。关系运算符建立两个量之间的比较关系，如果比较关系

成立，关系表达式的值为真（true），否则为假（false）。相互比较的两个量的类型必须一致，无论进行比较的两个量是字符型还是数值型，比较的结果均是逻辑型的。表 2-7 中给出的几个关系运算符均有两种表示方法，括号中的关系运算符与括号前的是等价的，例如：“x>0”和“x GE 0”是等价的，如果 x=2，则表达式“x>0”为真，表达式的值为 1（true）；如果 x=0，表达式“x>0”为假，其值为 0（false）。

（3）逻辑运算符与逻辑表达式。在表 2-7 中，逻辑运算符括号前的符号与括号中的等价，例如：“x & y”与“x and y”等价。逻辑运算符、逻辑型变量或值为逻辑型的表达式（如关系表达式）都称为逻辑表达式，逻辑表达式的值为逻辑型（true 或 false）。逻辑运算规则如表 2-8 所示。

表 2-8 逻辑运算规则

逻辑表达式	结 果	逻辑表达式	结 果
true AND true	= true	true OR true	= true
true AND false	= false	true OR false	= true
false AND false	= false	false OR false	= false
true AND missing	= missing	true OR missing	= true
missing AND missing	= missing	missing OR missing	= missing
false AND missing	= false	false OR missing	= missing

### 2.2.3 如何定义一个变量


定义变量在如图 2-6 所示的 Data Editor 窗口中的 Variable View 选项卡视图中进行。下面介绍此变量视图中关于变量的几条性质和设置方法。

#### 1. Name（变量名）

若用户没有特意地定义变量，比如只是在数据视图中输入过数据，那么变量视图的第一列属性 Name 给出的是默认变量名 VAR00001、VAR00002、VAR00003 等。双击某个变量名可以进入其编辑状态；或者，单击选中某个变量名称，直接键入新变量名，这时单元格中原来的字符自动清空，只显示新键入的内容；编辑好后，键入回车或单击他处地方确认修改，关于变量名的定义规则请参考第 2.2.1 节。依次单击菜单“Edit→Undo”或者使用组合键 Ctrl+Z 可以取消刚刚做过的修改，只是在单元格处于编辑状态时这两个命令均不可用。

注意，只要某个单元格进入过编辑状态（比如对其进行了双击操作），即使没有做任何修改，再单击别处后，仍会在曾处于编辑状态的行上自动生成默认变量，并且在这行和用户自定义的最后一个变量间自动填满默认变量。出现这种情况后，单击选中无用的变量行，在其上单击右键，弹出的快捷菜单中选择 Clear 删除即可。

#### 2. Type（变量类型）

单击选中某个变量的 Type 属性单元格，再单击单元格右侧出现的按钮，弹出如图 2-9 所示的变量类型选择对话框。单击选中相应的单选框，再单击 OK 按钮即可完



成设置。

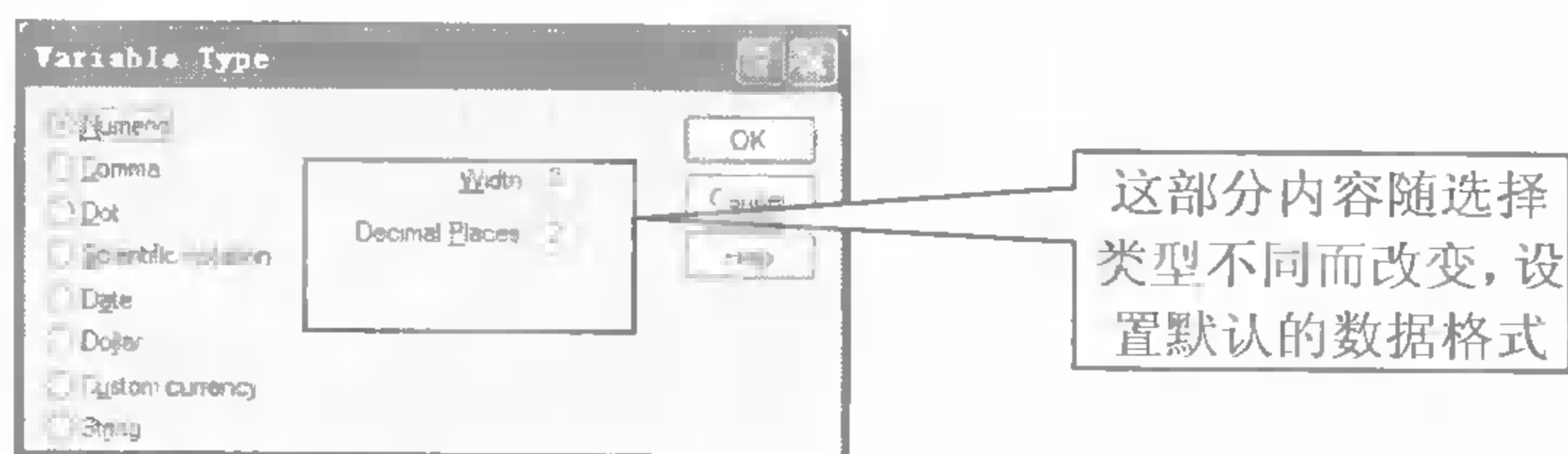


图 2-9 变量类型选择对话框

- **Numeric:** 数值型, 同时定义数值的宽度 (Width), 即整数部分+小数点+小数部分的位数, 默认为 8 位; 定义小数位数 (Decimal Places), 默认为 2 位。
- **Comma:** 加显逗号的数值型, 即整数部分从左侧开始每 3 位数加一个逗号, 其余参数的设置方式同 Numeric。
- **Dot:** 3 位加点数值型, 无论数值大小, 均以整数形式显示, 且每 3 位加一个小点 (不是小数点), 可以定义小数位置, 但都显示为 0, 且小数点用逗号表示, 例如 1.234 56 显示为 12.345 6,00 (实际上是 123456E-4)。
- **Scientific notation:** 科学记数型, 需要同时定义数值宽度 (Width) 和小数位数 (Decimal Places), 在数据编辑窗口中数值以指数形式显示。例如: 定义数值宽度为 9, 小数位数为 2, 则 12 345.678 显示为 1.23E+004。
- **Date:** 日期型, 用户可从系统提供的日期格式中选择合适的类型, 例如选择 “mm/dd/yy”, 则 2008 年 3 月 25 日显示为 “03/25/08”, 其他日期类型请参考表 2-4。
- **Dollar:** 货币型, 用户可从系统提供的形式列表中选择合适的类型, 并定义数值宽度和小数位数, 格式为带有前缀 “\$” 的数值。
- **Custom currency:** 自定义数值型类型, 指定在第 1 章的图 1-35 中预设的类型。
- **String:** 字符型, 同时需要定义字符长度 (Characters)。

### 3. Width (宽度)

此列设置数值变量或字符变量的宽度, 当变量为日期型时无效, 它与在图 2-9 中黑色方框位置进行的设置作用相同, 默认的变量显示宽度为 8。


### 4. Decimals (小数位数)

此列设置数值变量的小数位数, 当变量为非数值型时无效, 它与在图 2-9 中黑色方框位置进行的设置作用相同, 默认的小数位数为 2。

### 5. Label (变量标签)

变量标签是对变量名的进一步描述, 在结果输出中常显示于变量名的旁边, 或代替变量名作为输出, 可以输入中文。

### 6. Values (变量值标签)

值标签是对变量的每个取值的进一步描述, 对于定类变量或定序变量时, 它是非常有用的。单击选中单元格, 再单击单元格右侧出现的  按钮, 弹出如图 2-10 所示的对

话框。

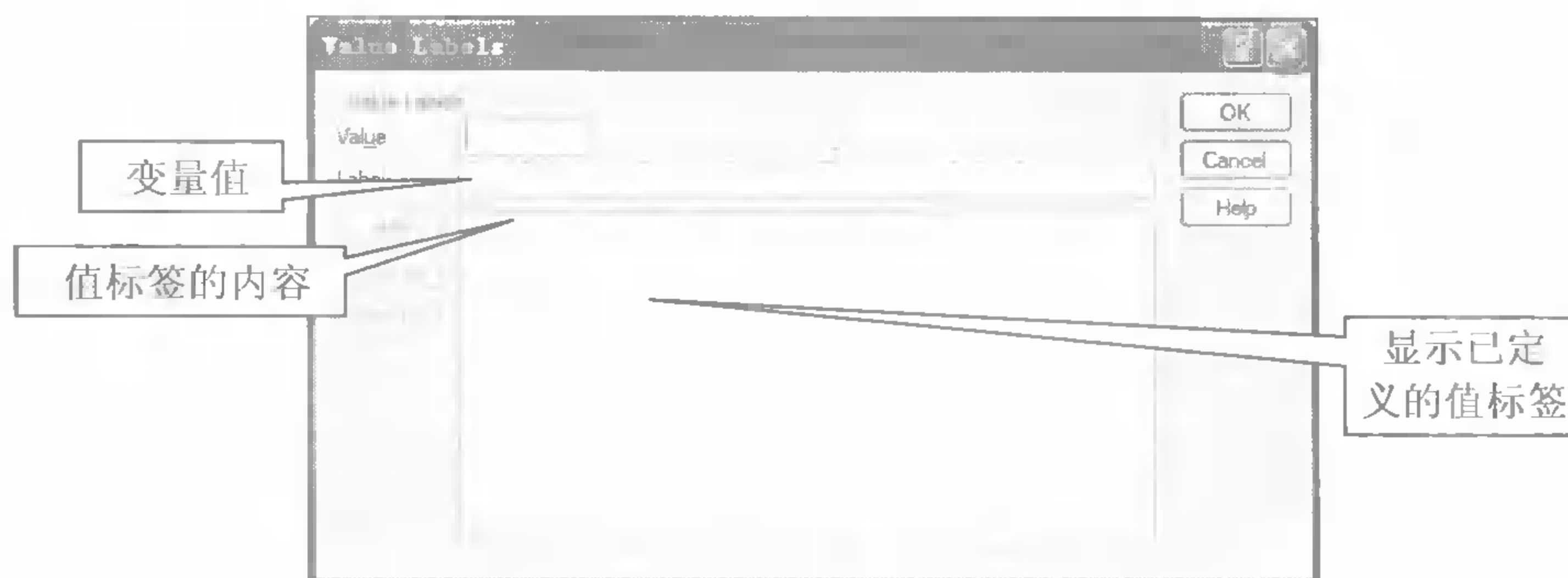



图 2-10 变量值标签定义对话框

在图 2-10 中，Value 处输入预定义标签的变量值，Label 处输入预定义的数据标签内容，单击 Add 按钮添加新标签，单击 Change 按钮改变现有标签，单击 Remove 按钮删除已有标签。设置完毕，单击 OK 按钮确认。

## 7. Missing (缺失值定义)

在 Variable View 视图窗口里，单击选中某变量的 Missing 属性列对应的单元格，再单击单元格右侧出现的  按钮，弹出如图 2-11 所示的定义缺失值对话框。缺失值定义好后会随数据文件一同保存，不需要在每次打开同一文件时重新定义。

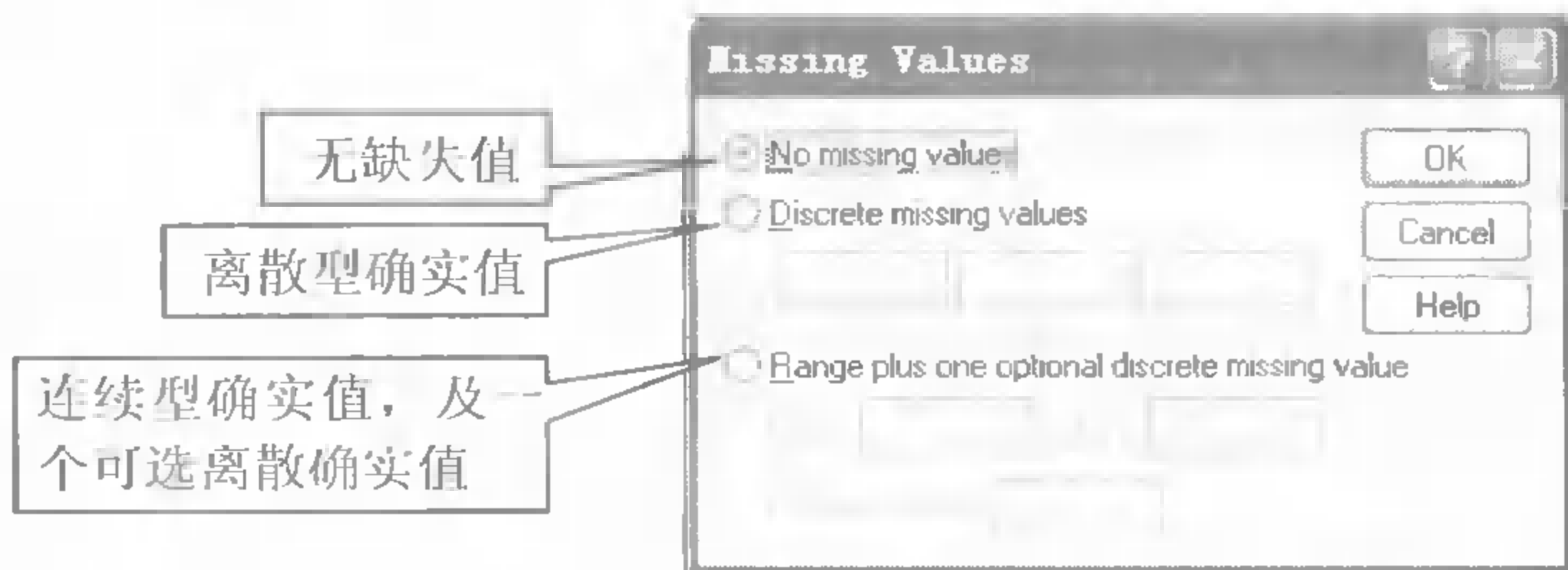


图 2-11 Missing 缺失值定义对话框

如图 2-11 所示，缺失值的定义有如下 3 种方式。

- No missing values 单选框。无缺失值，系统默认方式。如果当前变量的取值很完整，单击选中此项。
- Discrete missing values 单选框。离散缺失值，单击选中该项后激活下面的 3 个输入框，用于指定 3 个可能在变量取值中出现的缺失值，指定缺失值也可以少于 3 个，余下保留空白即可。
- Range plus one optional discrete missing value 单选框。定义缺失值的取值范围，单击选中该项后激活下面的选项。Low、High 输入框分别用于指定缺失值范围的上、下界，而且范围的设置只能针对数值型变量；同时，可以在 Discrete value 输入框指定一个离散形式的缺失值。

注意：不能为长字符串型变量（大于 8 个字符）定义缺失值；对于字符串型变量，所有取值，包括空值（NULL）和空格，默认都是合法的取值，即非缺失值，除非在此处把这些

取值设置为缺失值。设置空值 (NULL) 或空格为缺失值的方法是，在 Discrete missing values 下的某个输入框键入一个单个的空格字符。

## 8. Columns (列的显示宽度)

Columns 属性列设置在 Data View 视图中当前变量所占的列宽度，默认值为 8。用户可以直接键入合适的列宽值，或者单击单元格右侧的上下按钮调节值的大小。

另外，在 Data View 视图中用直接拖动的方式也能设置某列的显示宽度，方法是：鼠标置于二维表格顶部两个列名的相交之处，待光标呈黑体双项箭头时，按下左键不放并向左右拖动，同时，Variable View 视图中相应的 Columns 属性值也会随着拖动设置而自动调整。

如果此处定义的显示宽度小于变量取值宽度，在 Data View 视图中的单元格里，较宽的变量取值会以 “\*” 显示。

## 9. Align (对齐方式)

Align 属性列设置为变量取值在 Data View 视图窗口中显示时的对齐方式。

单击选中某变量的 Align 属性列对应的单元格，再单击单元格右侧出现的的下拉列表，弹出可选的对齐方式有以下 3 种：Left (左对齐)、Center (中间对齐)、Right (右对齐)。

## 10. Measure (变量度量方式)

单击选中某变量的 Measure 属性列对应的单元格，再单击单元格右侧出现的下拉列表，弹出可选的度量方式有以下 3 种。

- Nominal (名义变量)。一种分类变量，并且变量取值之间没有内在的大小可比性。例如：以一个公司不同部门的名称所做的变量；典型的名义变量还有地区、邮编、种族等。
- Ordinal (序数变量)。一种分类变量，但是变量取值之间有内在的大小顺序或等级。例如：把服务满意度作为变量，其取值可以为 1 级、2 级、3 级，数字越大表示越满意。典型的序数变量还有自信度大小、偏好大小等。
- Scale (尺度变量)。也称为刻度变量，一般为有刻度度量的连续型变量，可以在不同的取值之间定义距离。典型的尺度变量有：年份、美元收入、路程等。

注意：对于字符串型的序数变量，分类时将自动按照字母顺序排列。例如：变量取值为 low、medium、high 时，分类比较时的顺序会自动排列为 high、low、medium，这显然是不正确的。所以，建议使用数值型变量代替字符型的序数变量，而将字符设为变量的值标签，例如：1 (low)、2 (medium)、3 (high)。

## 11. 利用菜单定义变量

依次单击菜单 “Data→Defining Variable Properties”，打开设置变量属性的图形向导，如图 2-12 所示。在变量列表中选中要设置属性的变量，单击中间的黑色箭头，将其选入右侧的待设置变量列表；单击 Continue 按钮进入向导的下一步，如图 2-13 所示。

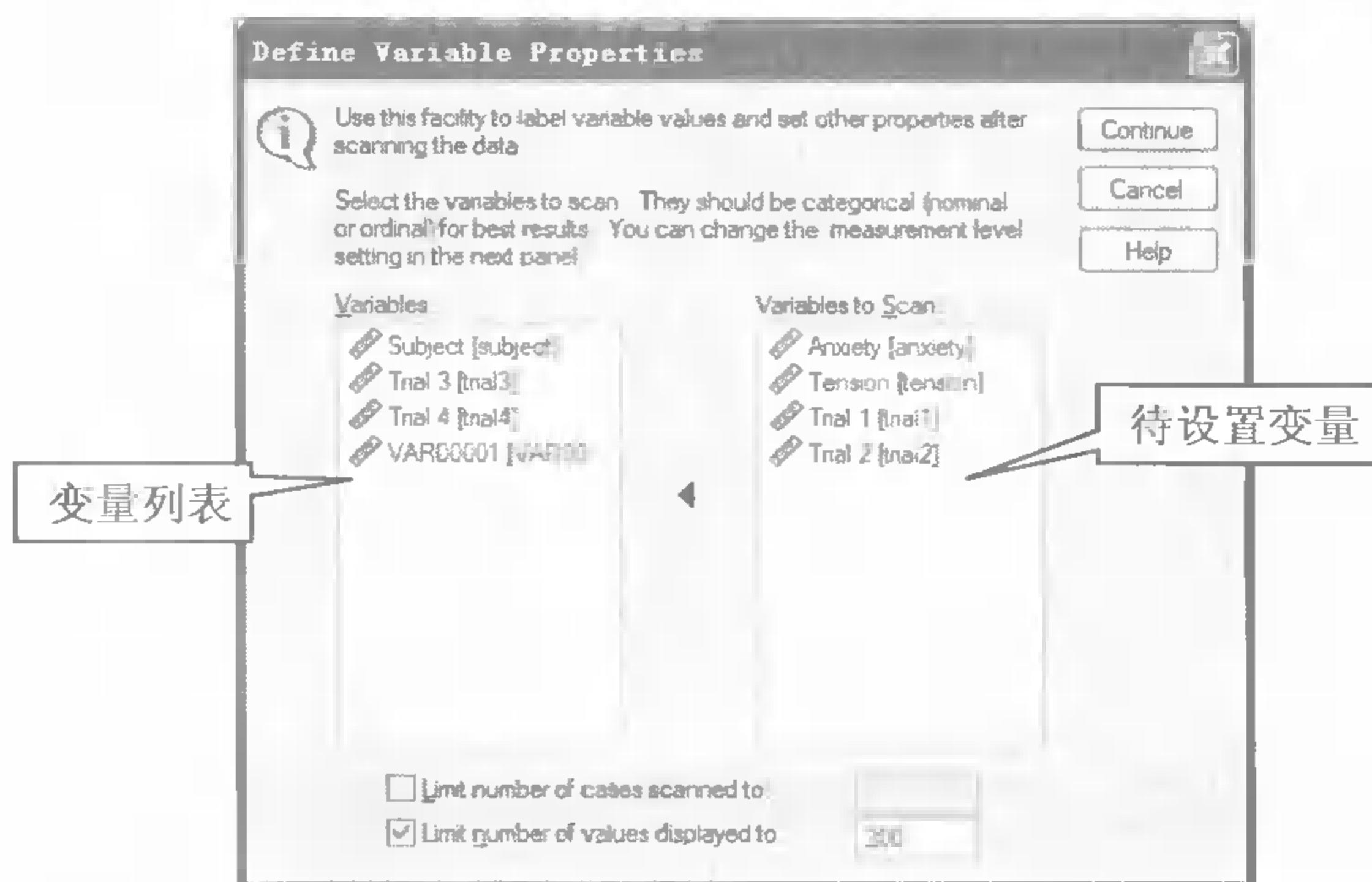


图 2-12 Defining Variable Properties 向导 1

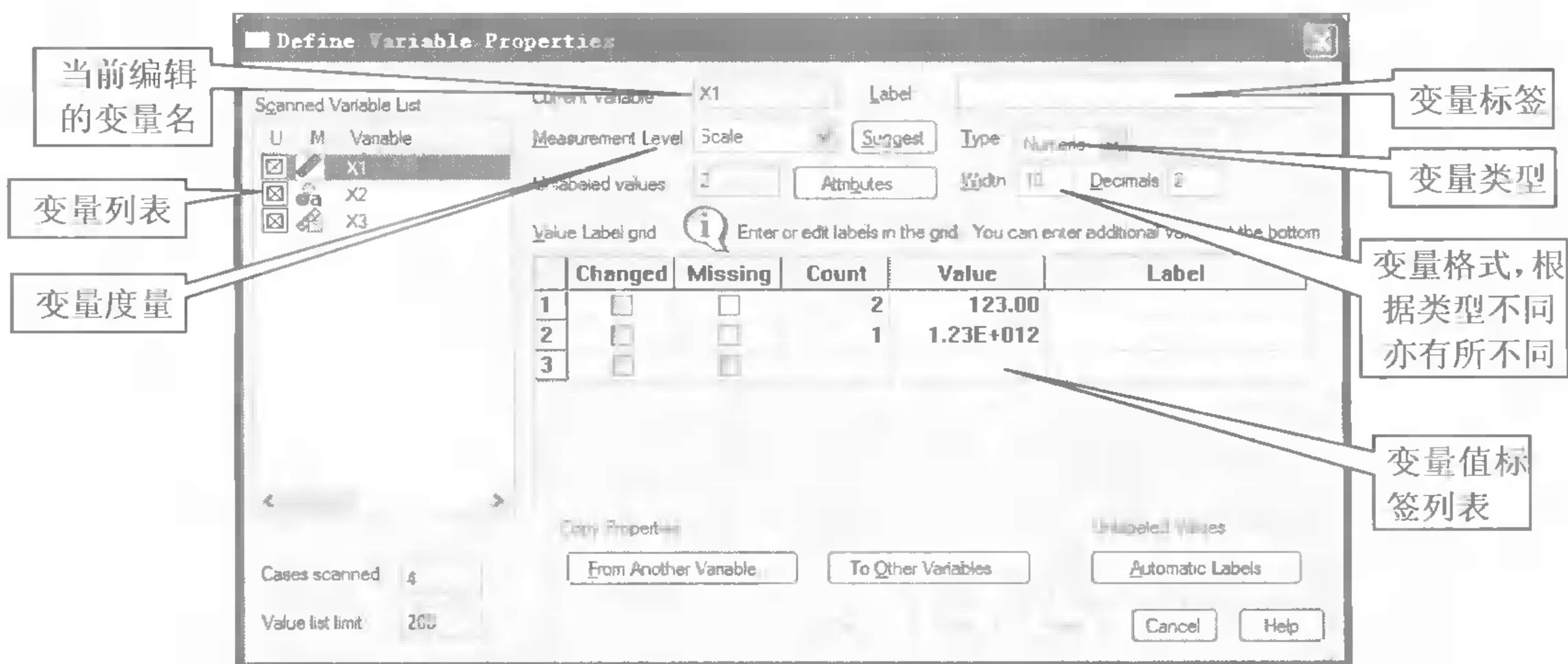


图 2-13 Defining Variable Properties 向导 2

在图 2-13 中参数选项的含义和设置方法, 与 Variable View 变量视图中的方法相似。在左侧的变量列表单击选中某个变量, 然后在右侧的设置区编辑变量属性, 设置完成后, 单击 OK 按钮返回主界面。

### 2.2.4 概率事件

数据编辑窗口的 Data Viewer 视图里, 二维表格的一行是一个观测记录, 在统计学中称作概率事件, 在 SPSS 的功能菜单和帮助信息中用 Cases 表示。

一个 Case 由各变量的一组取值组合而成, 称为一个事件, 它是由一个被观测对象各种特征的实测值组成的观测记录。相对于变量, Cases 可以称为观测量, 单元格中的数据就是某个观测量的某些特征属性的取值, 称之为变量值。

## 2.3 输入数据

本节介绍如何在 Data Editor 窗口输入数据以及 3 种查看文件信息和变量信息的方法。

### 2.3.1 输入数据的方法

在图 2-14 所示的数据编辑窗口 (Data Editor)，可以直接录入和编辑数据。每一列代表一个变量 (Variable) 或被观测量的一个特征，例如问卷上的每一项就是一个变量；每一行代表一个个体、观测或样本，在 SPSS 中称之为事件或记录 (Case)，例如某个人回答的一张问卷就是一个事件；单元格 (Cell) 是观测和变量的交叉位置，记录的是相应观测行的相应列属性的取值。整个数据文件就是一张二维表格，数据范围由观测行和变量列的数目决定，与 Excel 电子表格相似，不同的是此处的单元格只能包括数据值而不能包含公式。在 Data Editor 中，可在任一单元中输入数据，SPSS 将自动将数据表格延长到包含那个单元格的最小范围。

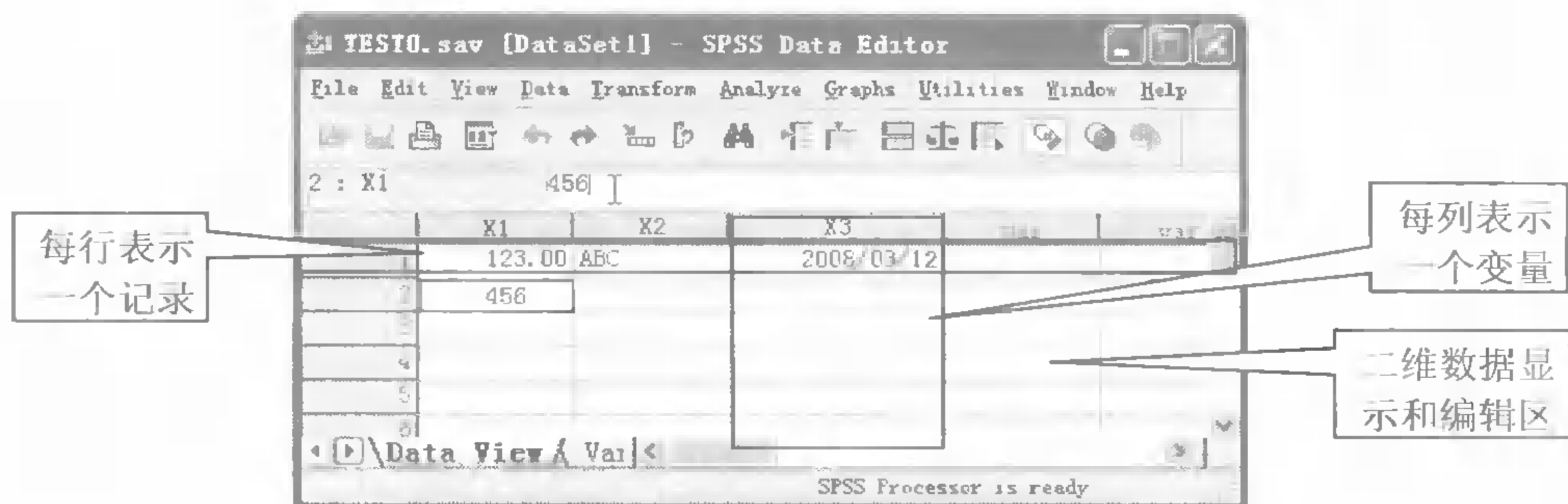


图 2-14 数据录入

例如，在图 2-14 中单击选中第 2 行和第 1 列交叉位置的单元格，此时单元格边框呈加粗显示，表格左上角的提示区显示了当前选中的记录行数和变量名字“2: X1”；然后，在提示区右侧的输入框中键入“456”。回车后数据就会录入选中的单元格里。或者，双击预输入数据的单元格，待光标变为输入提示符（如图 2-14 中数据“456”右侧的指针）时，可以直接在单元格输入数据。

### 2.3.2 查看文件信息和变量信息

#### 1. 在 SPSS Viewer 中查看文件和变量信息

依次单击菜单“File→Display Data File Information→Working File”，可以将当前文件的相关信息输出到 SPSS Viewer 窗口中。例如：在图 2-14 中单击此菜单项，会输出如图 2-15 所示的文件和变量信息。

变量信息							
变量	位置	类型	变量测量	度量	缺失值	初始格式	显示格式
X1	1	<none>	尺度	8	无	F8.2	F8.2
X2	2	<none>	名称	8	无	A8	A8
X3	3	<none>	尺度	11	无	SCATE10	SCATE10

图 2-15 文件信息

依次单击菜单“File→Display Data File Information→External File...”执行查看外部文件信息的功能，弹出如图 2-16 所示的打开文件对话框，在此选择“Anxiety.sav”（SPSS 安装时自带的样本数据）并打开，SPSS Viewer 窗口的输出结果如图 2-17 所示。





图 2-16 Display Data Info 对话框

文件信息		
源	C:\Program Files\SPSS\Anxiety.sav	
类型	SPSS 数据文件	
创建日期	30-NOV-1999 17:34:00	
标签	无	
文件内容	数据类型	
	文档的行数	
	变量集	
	趋势数据信息	
	多重响应定义	
	Data Entry for Windows 信息	
	TextSmart 信息	
	Clementine 信息	
数据信息	个案数	8
	已定义的变量元素数	5
	拖拽变量数	5
	权重变量	无
	已压缩	是

变量信息						
名称	位置	标签	度量尺度	格式	宽度	显示方式
subject	1	Subject	标度	F8	8	右
anxiety	2	Anxiety	标度	F2	2	右
tension	3	Tension	标度	F2	2	右
source	4	Source	标度	F2	2	右
trial	5	Trial	标度	F2	2	右

图 2-17 输出结果

## 2. 在 Data Editor 中查看变量信息

如图 2-18 所示，在数据编辑窗口中单击选中底部的 Variable View 视图标签，即可显示和编辑当前数据文件中的变量信息。

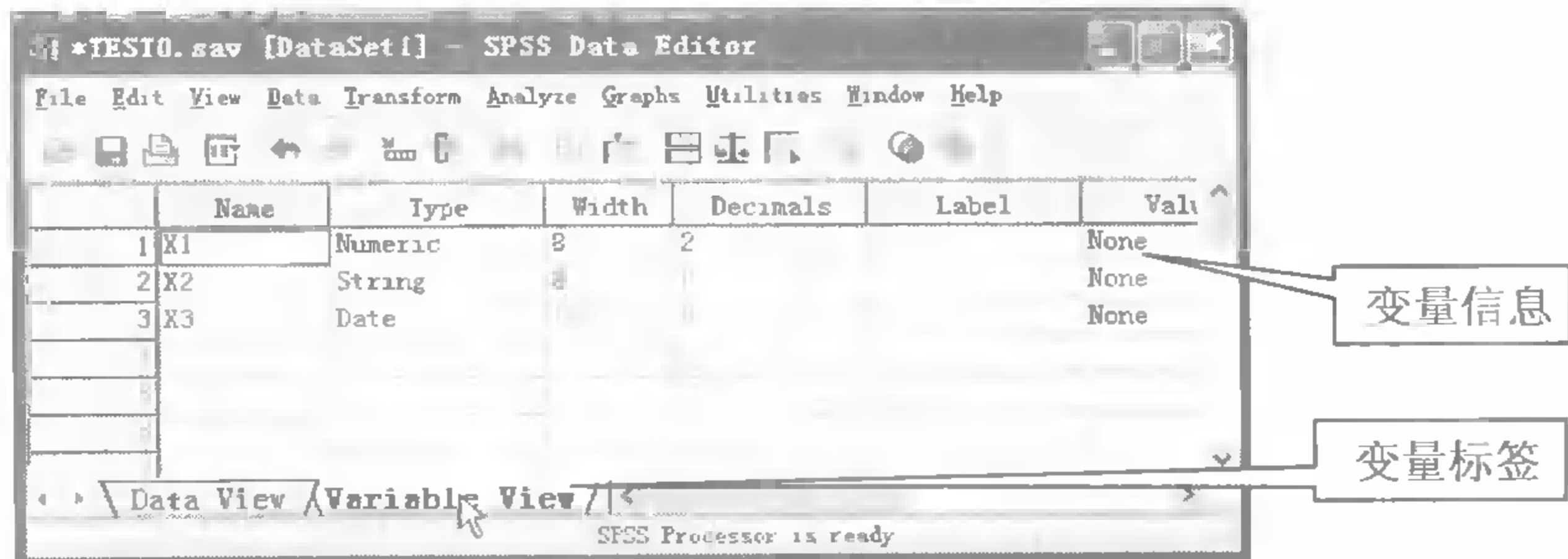



图 2-18 变量信息 1

### 3. 工具栏查看变量信息

在图 2-18 中,依次单击菜单“Utilities→Variables...”,弹出如图 2-19 所示的对话框,显示当前文件中的变量信息;单击工具栏中的按钮也可以打开图 2-19 所示的变量信息对话框。单击选中左侧变量列表里的某个变量,右侧的变量信息区就会显示选中变量的属性信息,且这些信息不可编辑。

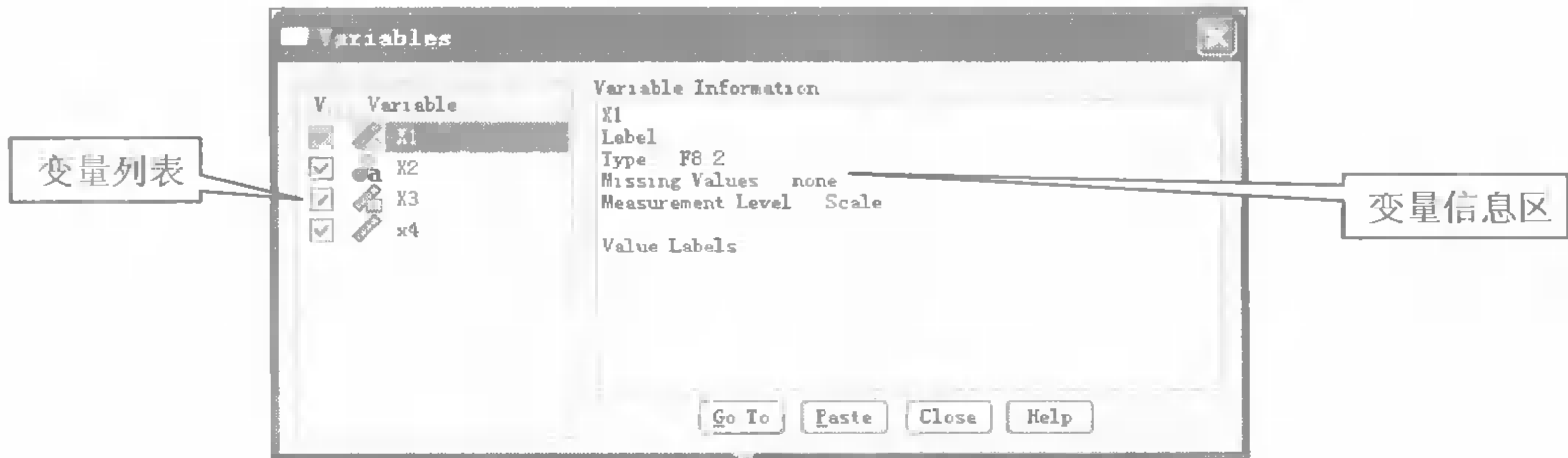


图 2-19 变量信息 2

## 2.4 编辑数据文件

本节介绍如何在 Data Editor 窗口编辑数据和变量,包括数据与变量的插入、复制、粘贴、删除等操作。

### 2.4.1 在单元格中编辑数据

在 Data Editor 窗口的二维表格区中,编辑数据的方法与第 2.3.1 节介绍的输入数据方法相似,具体的操作方式如表 2-9 所示。

表 2-9

数据编辑命令

键 盘 操 作	单元格定位	滚动条操作		窗 口 移 动
↑	向上移动一个单元格	纵向	上箭头	窗口上移一行
↓	向下移动一个单元格		下箭头	窗口下移一行
PgUp	向上移动一屏		单击上箭头与移动块间	向上移动一屏
PgDn	向下移动一屏		单击下箭头与移动块间	向下移动一屏
→ 或 Tab	向右移动一个单元格	横向	上下拖动移动块	窗口不定量上下移动
← 或 Shift+Tab	向左移动一个单元格		左箭头	窗口左移一列
Home	向左移动到行首		右箭头	窗口右移一列
End	向右移动到行末		左右拖动移动块	窗口不定量左右移动

### 2.4.2 插入变量与删除变量

#### 1. 插入变量

在 Data View 视图窗口,把光标移至某个变量名,待鼠标变为向下的黑色箭头,单击选中该列

变量，此时整列变量呈黑白反显；再右击选中变量，在弹出的快捷菜单里单击 Insert Variable 项，即可在选中变量列的左侧插入一个新的默认变量 VAR0000x，此处 x 可能是任何一个整数，为系统定义的变量序号，若之前系统曾自动定义过 n 个变量，那么此处新生成的变量序号就为 n+1。

在 Data View 窗口单击选中任一个单元格，依次单击菜单“Edit→Insert Variable”，也可以完成同样的功能。

在 Data View 窗口，通过输入、复制粘贴新内容到空白列（可以是单个或多个单元格，也可以是整个变量列），就可以在新输入内容的列上自动插入新的变量。

在 Variable View 窗口，同样可以通过上述 3 种方法插入新的变量，只不过这时操作的对象为变量行。

## 2. 删除变量

在 Data View 窗口，单击选中某个整列变量后，依次单击菜单“Edit→Cut（或 Clear）”，即可删除此变量。单击 Cut 项时，选中的变量暂时存入剪贴板，可以通过单击“Edit→Paste”恢复；而单击 Clear 项就是删除操作，可以通过单击“Edit→Undo”来恢复。以上删除操作，均可以通过 Delete 键完成。

在 Variable View 窗口，操作与上类似，只是操作的对象变成变量行。

无论使用哪种方法删除变量，其结果均为被选定的变量消失，其下面的诸变量上移（Variable View 窗口）或右侧的变量左移（Data View 窗口）。

### 2.4.3 插入观测量与删除观测量

#### 1. 插入观测量

插入观测量的操作均在 Data View 窗口进行，观测量的排列次序可以用排序功能整理，故插入位置可以不必计较。

在 Data View 窗口，把光标移动至某个观测记录行号上，待鼠标变为向右的黑色箭头，单击选中整行记录，此时整行记录呈黑白反显，再在选中行上右击，在弹出的快捷菜单中单击 Insert Cases 项，即可在此记录行的上面插入一个新的记录行，所有行号自动重编。

在 Data View 窗口中单击选中任一个单元格，依次单击菜单“Edit→Insert Cases”，也可以完成插入记录行的功能。另外，通过输入、复制粘贴新的内容到空白行（可以是单个或多个单元格，也可以是整行记录的粘贴），就可以自动在新输入内容的行上插入新的记录。

#### 2. 删除观测量

记录行的删除操作仍是在 Data View 窗口进行，单击选中某行记录后，依次单击菜单“Edit→cut（或 clear）”，即可删除此记录。单击 Cut 项时，选中的记录暂时存入剪贴板，可以通过单击“Edit→Paste”恢复；而单击 Clear 项就是删除操作，可以通过单击“Edit→Undo”来恢复。以上删除操作，也可以通过 Delete 键完成。

进行删除操作后，被选中的记录行消失，其下面的各行上移，记录行号自动重编。

### 2.4.4 数据的剪切、复制和粘贴

如图 2-20 所示，首先选中需要操作的数据区域，被选中单元格的顏色呈黑白反显，然后

在选中区域上右击，弹出的快捷菜单有如下几项：Cut（剪切）、Copy（复制）、Paste（粘贴）、（Clear）删除、Grid Font（设置显示字体）。单击它们即可执行相应的操作，也可以通过单击 Edit 菜单的子菜单来执行。

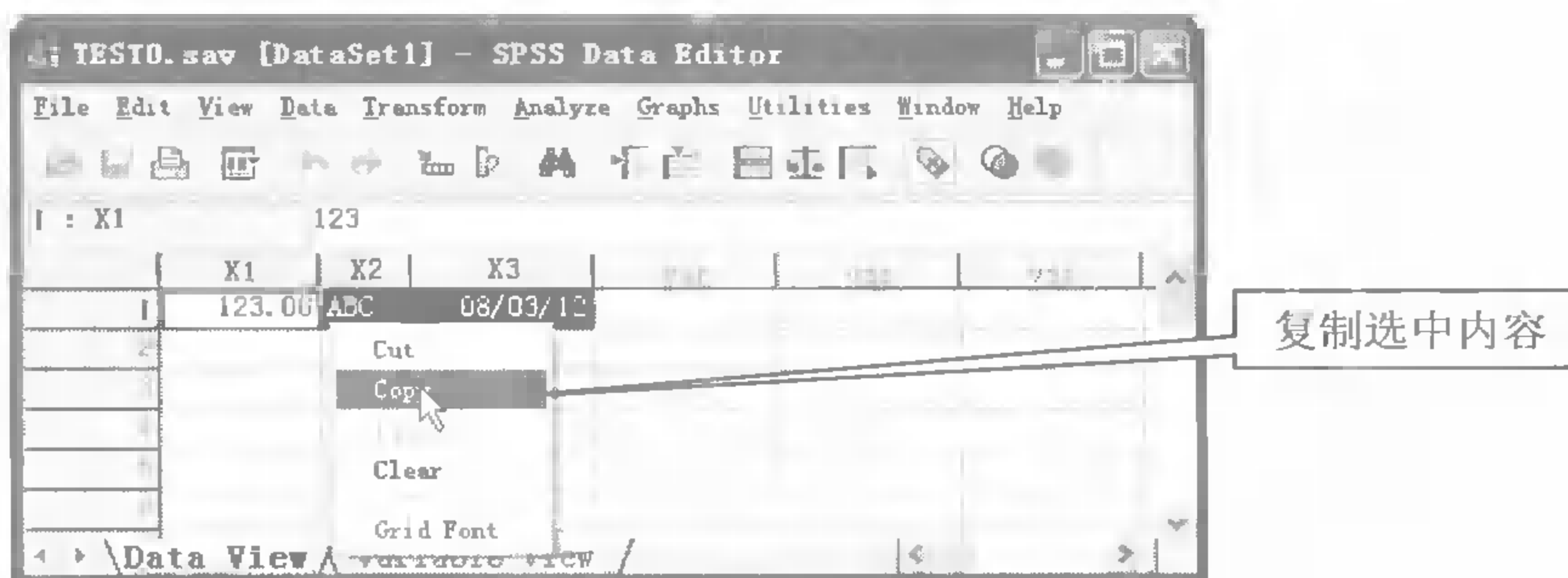


图 2-20 数据编辑

另外，通过快捷键 Ctrl+X、Ctrl+C、Ctrl+V、Ctrl+Z 分别可以执行剪切、复制、粘贴和撤销操作。

下面是进行编辑操作时，需要注意的几个事项。

(1) 在执行粘贴操作时，目的区域的数据格式必须与被粘贴数据的格式一致，否则 SPSS 将按照默认方式处理，如图 2-21 所示。

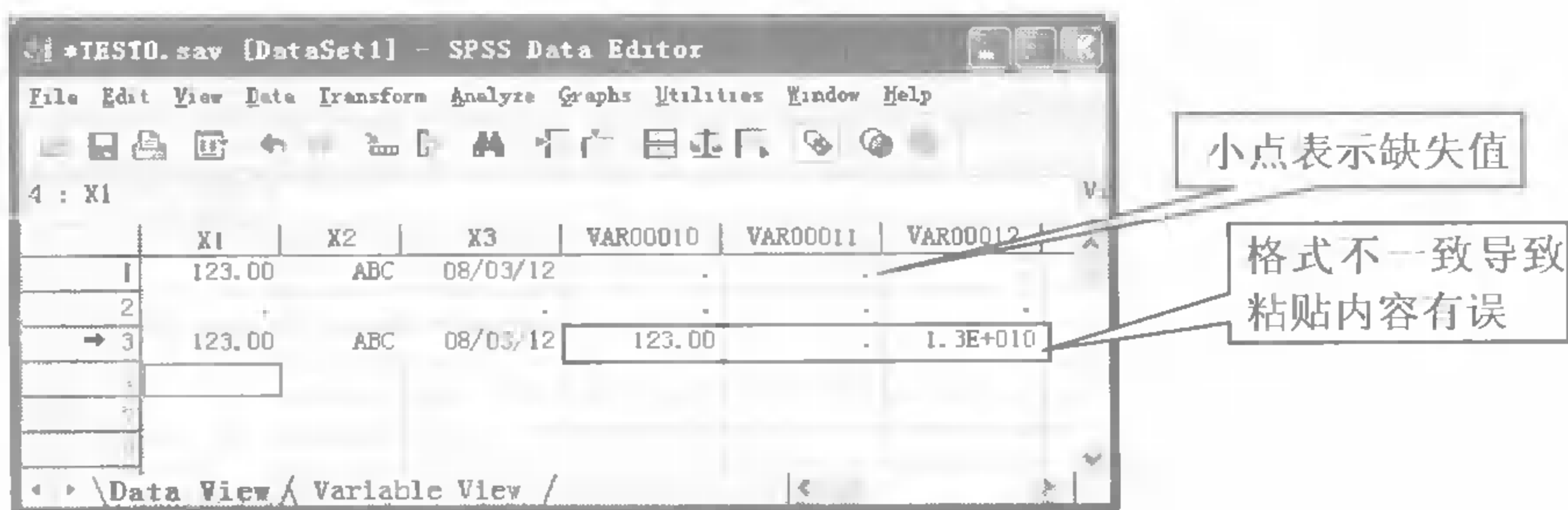


图 2-21 数据粘贴

① 第 3 行第 1-3 列数据，是粘贴在图 2-20 中所复制的内容，粘贴前，单击选中单元格“3: X1”，粘贴后数据与原始数据一致，同时第 2 行出现两个“.”，表示其数据值缺失。在图 2-21 中，默认情况下，数字和日期型变量的缺失值以“.”显示，字符型变量没有缺失值，显示为空 (NULL)。

② 第 3 行第 4-6 列数据，也是粘贴在图 2-20 中所复制的内容，粘贴前，选中了单元格“3: VAR00010”，VAR00010、VAR00011、VAR00012 是系统自动生成的变量名。可以看到，粘贴生成的新变量，默认的数据格式都是数字型，与粘贴内容无关，变量 X2 的内容“ABC”因此就没有被正确的粘贴到目的区域“3: VAR00011”。这提示我们在粘贴前，必须设置好粘贴目的区域的数据格式。

(2) 当需要粘贴整行记录时，可以直接单击相应的行号选中整条记录所有变量，例如在图 2-21 中，当光标移动至第 3 行记录的行号时，会变为一个黑色的向右箭头，此时单击即可选中整个第 3 行记录；选中整行后，再用与前述相仿的方法执行相应的编辑操作。类似地，可以选中整个变量列的内容进行编辑。

(3) 在选择单元格区域时，可以利用 Ctrl、Shift 组合键，快捷的选中连续区域。与 Excel 不同的是，按下 Ctrl 键选择多个单元格时，只能选择和已选中单元格至少有一个边相邻的单元

格，不能任意选择单元格。

(4) 在执行 Clear 删除操作时，若选中的是和图 2-20 所示相似的数据区域，只会删除相应的数据内容，而记录行和变量列仍是保留的，只是数据取值成为默认的缺失值或空值。如果要删除整行记录或整列变量，只能如第(3)点中所述，先选中相应的行或列再执行 Clear 操作。

### 2.4.5 撤销操作

编辑数据或变量属性时，撤销和重复操作通过单击菜单“Edit→Undo”、“Edit→Redo”来实现，而且这两个菜单还会同时提示具体的操作内容，如图 2-22 所示，分别表示：撤销删除变量、重复设置单元格值。另外使用快捷键 Ctrl+Z、Ctrl+R 也可以方便地实现这两个功能。

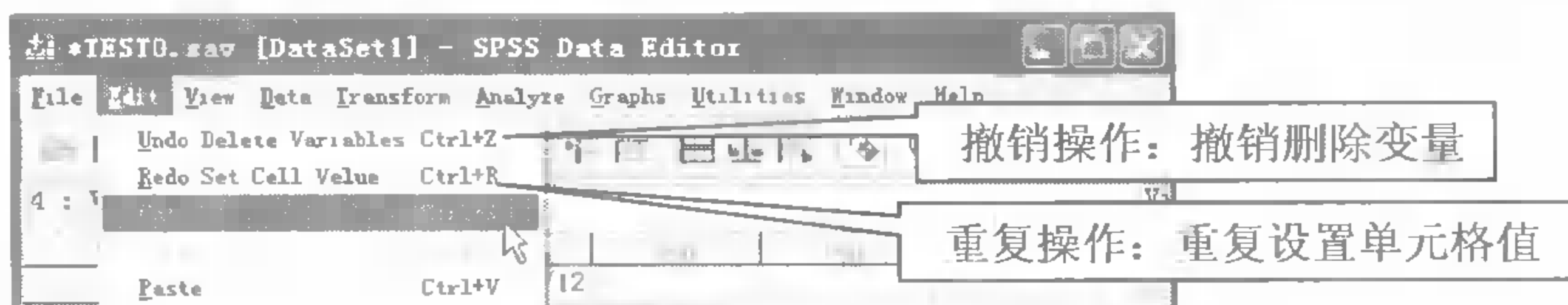




图 2-22 撤销操作

## 2.5 对数据文件的操作

前面几节介绍了如何录入和编辑 SPSS 格式的数据文件，即后缀是“sav”格式的文件，本节来讲解如何快速打开和编辑其他格式的文件，以及如何通过数据库向导进行文件的导入导出操作。

### 2.5.1 数据文件的打开与保存

如图 2-23 所示，通过文件菜单 File 可以实现数据文件的打开与保存操作，图上标出了各个子菜单项的涵义，工具栏最左端的打开按钮和保存按钮也实现相同的功能。

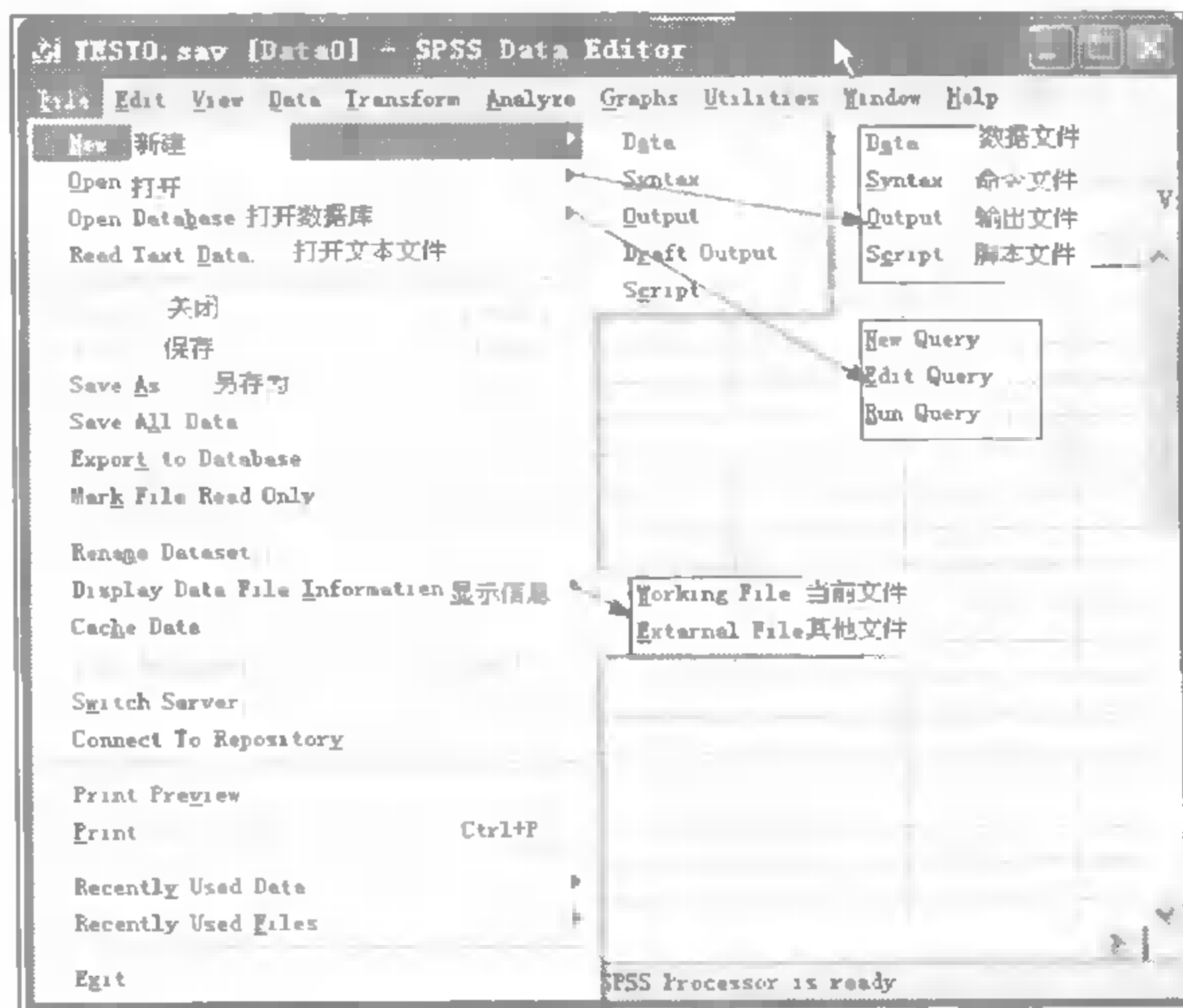


图 2-23 文件菜单

### 2.5.2 数据库文件的转换

下面来介绍如何快速打开其他格式的文件，以及数据库向导的使用方法。



## 1. 快速打开其他格式的数据文件

本节以打开 Excel 文件和文本文件为例，介绍如何打开其他格式的数据文件。

依次单击菜单“File→Open→Data...”，如图 2-23 所示，打开如图 2-24 所示的对话框，在此选中相应的文件即可。

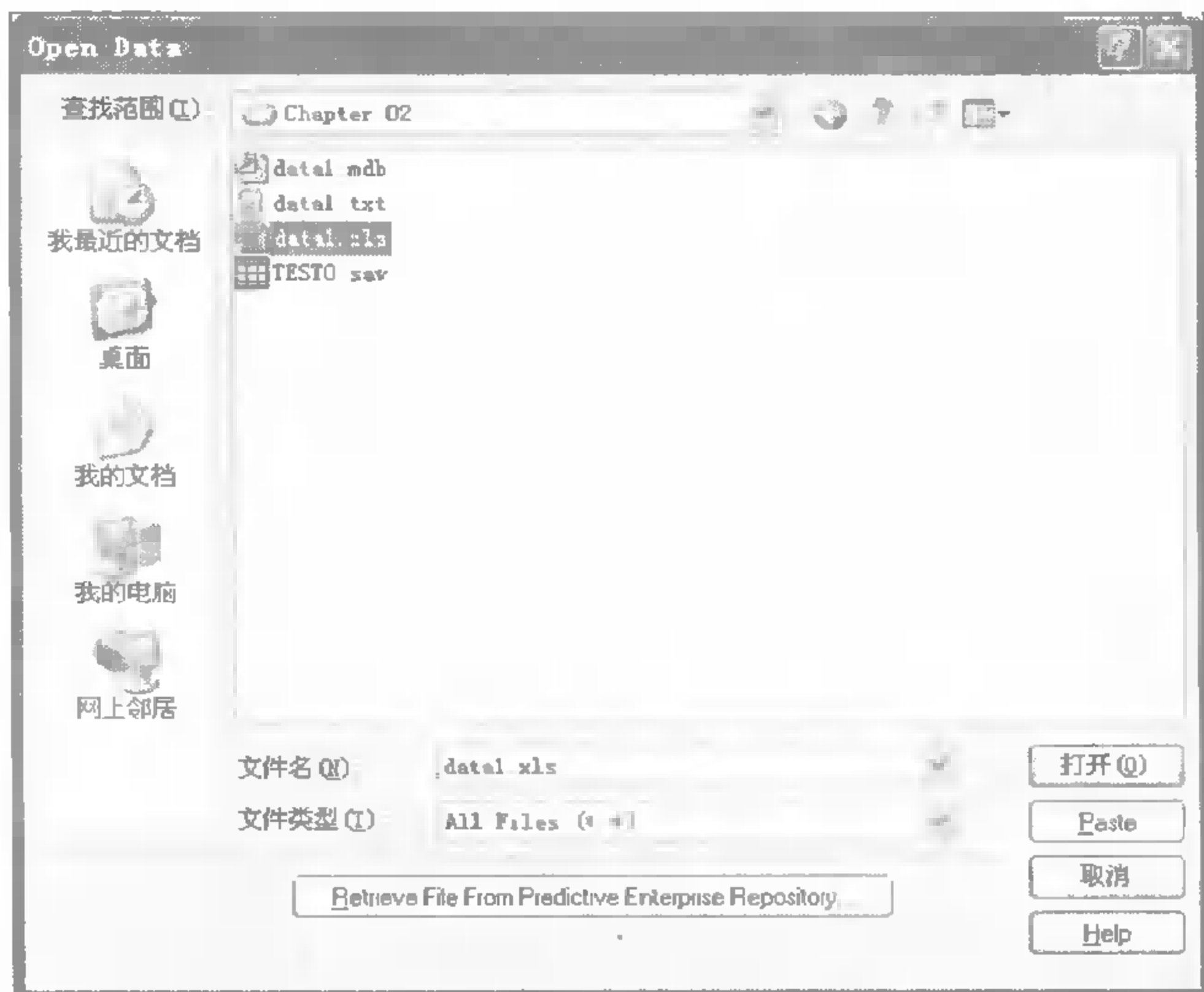


图 2-24 打开数据文件对话框

### (1) 打开 Excel 文件。

双击 Excel 文件，会弹出如图 2-25 所示的对话框，选中其中的 Read 复选框，表示把 Excel 表格中的第一行作为变量名读入，否则将第一行作为数据读入。

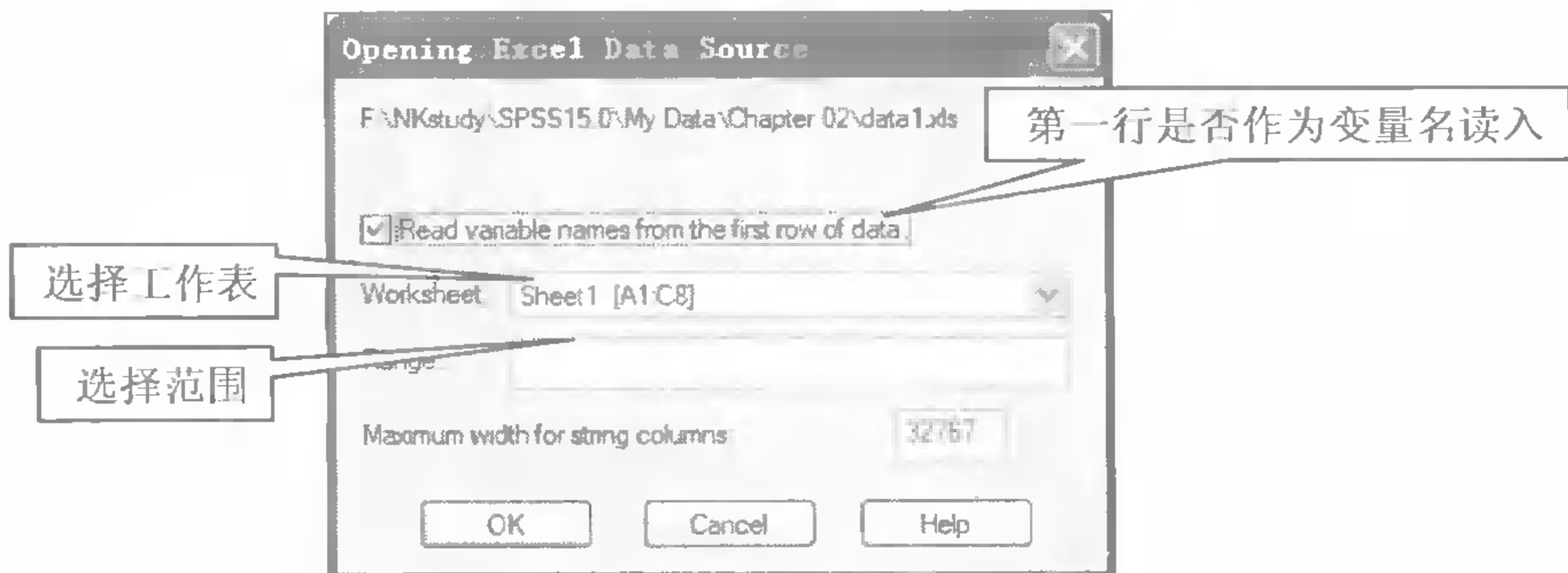


图 2-25 打开 Excel 文件 (1)

单击 OK 按钮，打开如图 2-26 所示的 Data Editor 窗口，显示导入的 Excel 文件的内容。用户未指定变量名时，SPSS 自动为数据按列设定变量名为 V1、V2...，自动设置列的显示宽度，且读入数据的格式自动转换为 SPSS 的数字、字符、日期等格式。依次单击菜单“File→Open→Save (Save As)”，可将当前数据存为其他格式的文件，比如 SPSS 格式；依次单击菜单“File→Export to database...”，可以把当前数据导入指定的数据库中，具体操作方法和随后讲到的 Open Database 过程完全类似。

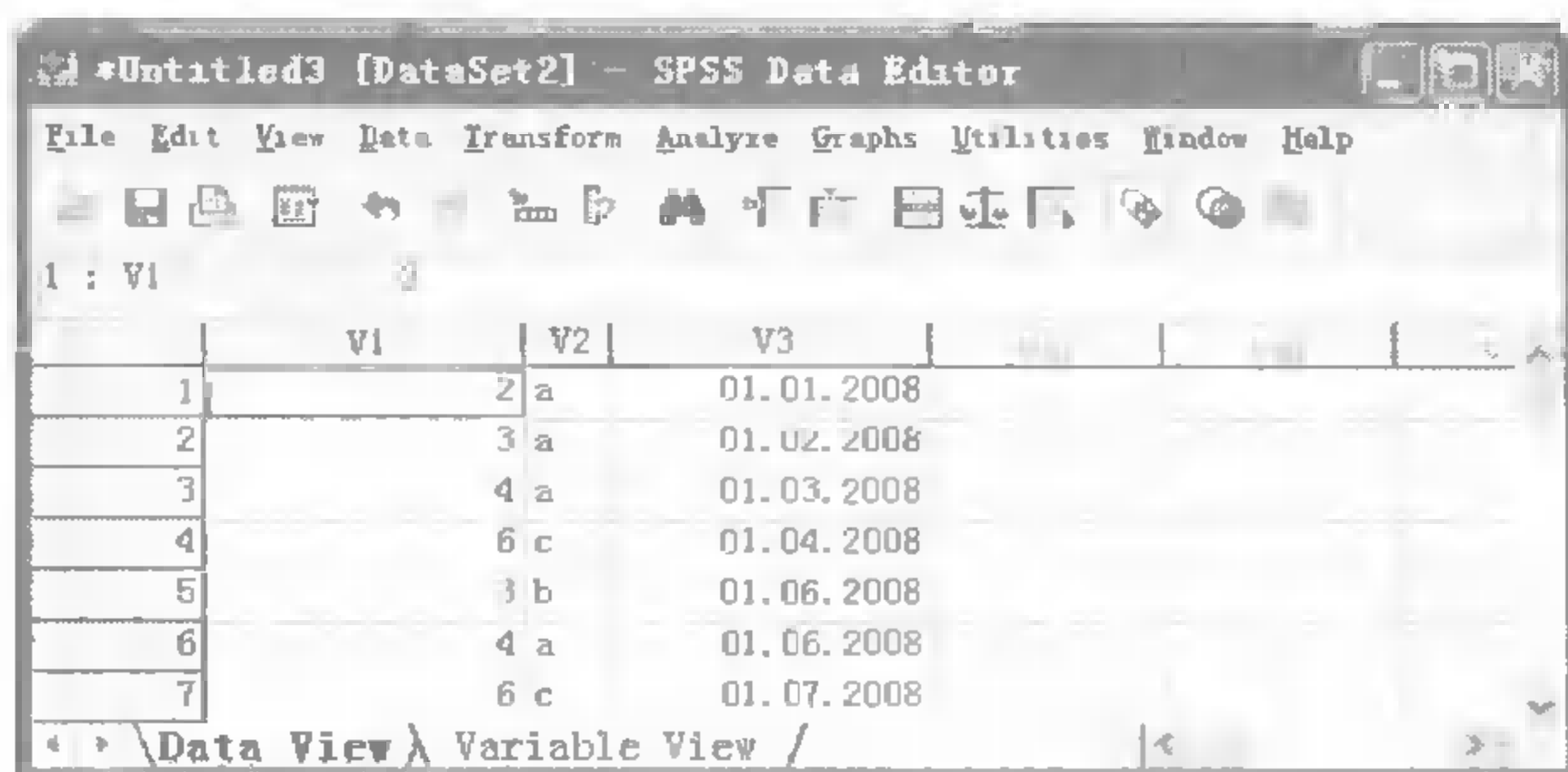


图 2-26 打开 Excel 文件 (2)

## (2) 打开 txt 文本文件。

① 依次单击菜单“File→Open→Data...”或“File→Read Text Data...”，打开如图 2-24 所示的对话框，双击文本文件（例如“Data1.txt”）则弹出如图 2-27 所示的 Text Import Wizard—Step1 of 6 对话框。

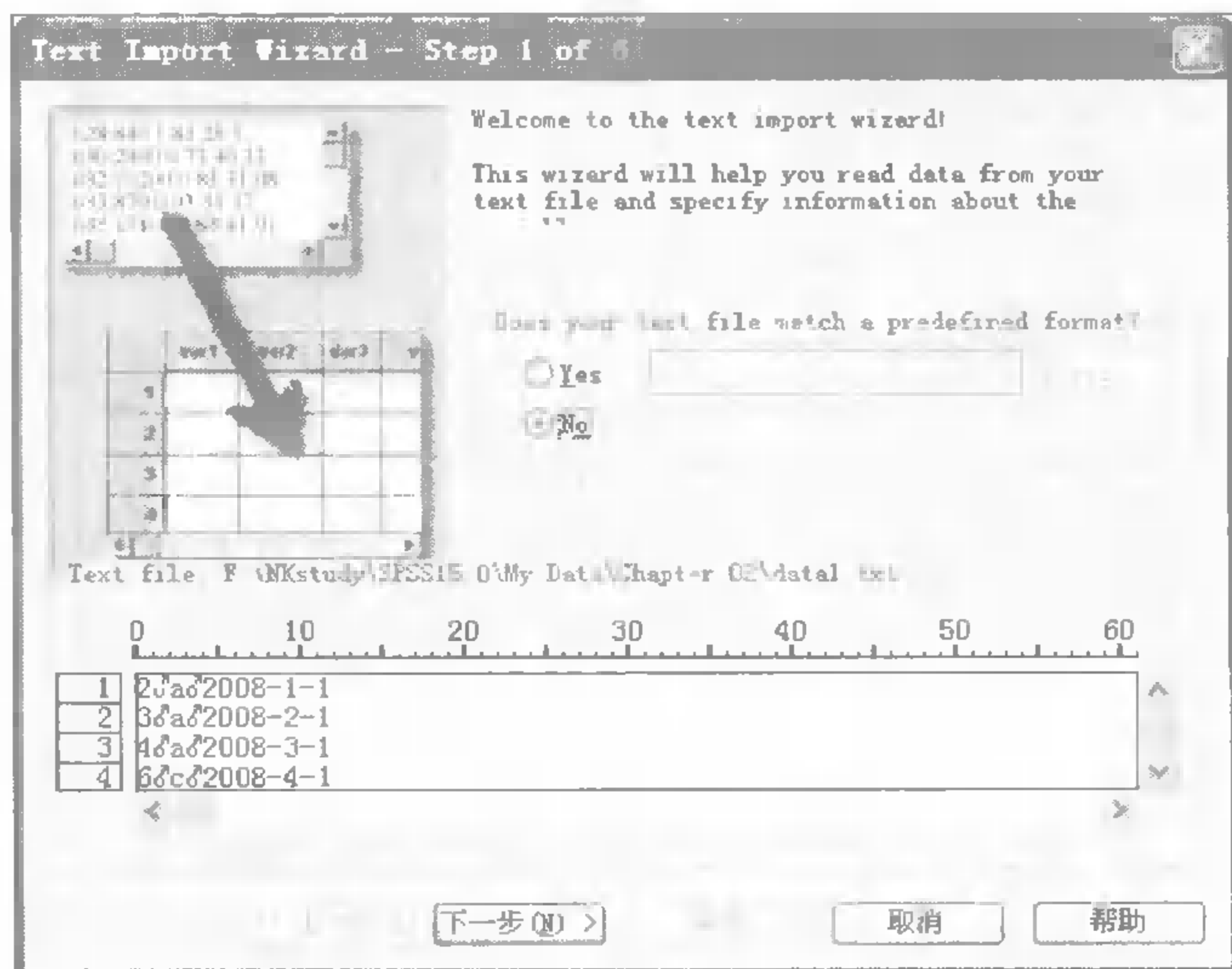


图 2-27 Text Import Wizard—Step1 of 6 对话框

② 单击下一步按钮，进入 Step 2 of 6 对话框，在此选择读入数据的分隔符——Delimited（特定分隔符）或 Fixed width（固定宽度），以及源文件是否包含变量名。

③ 单击下一步按钮进入 Step 3 of 6 对话框，选择读入数据的起始行（The first case）、每行记录包含的变量个数以及要读取的行数。

④ 单击下一步按钮进入 Step 4 of 6 对话框，选择具体的分割符（Delimiters），以及数据是否有引号或其他符号引用。

⑤ 单击下一步按钮进入 Step 5 of 6 对话框，如图 2-28 所示，在此设置读入变量的格式，首先在数据预览区（Data Preview）单击选中 V3 列下的某个单元格，Variable name 栏下会自动显示相应的变量名，Data 栏下给出默认的数据格式 Numeric（数字型），然后单击 Data 栏后的下拉列表，选中代表日期的格式 Data/Time，在右边弹出的变量格式列表框里单击选中“yyyy/mm/dd”，用同样的方法设置变量 V2 的格式为 String（字符串型）。

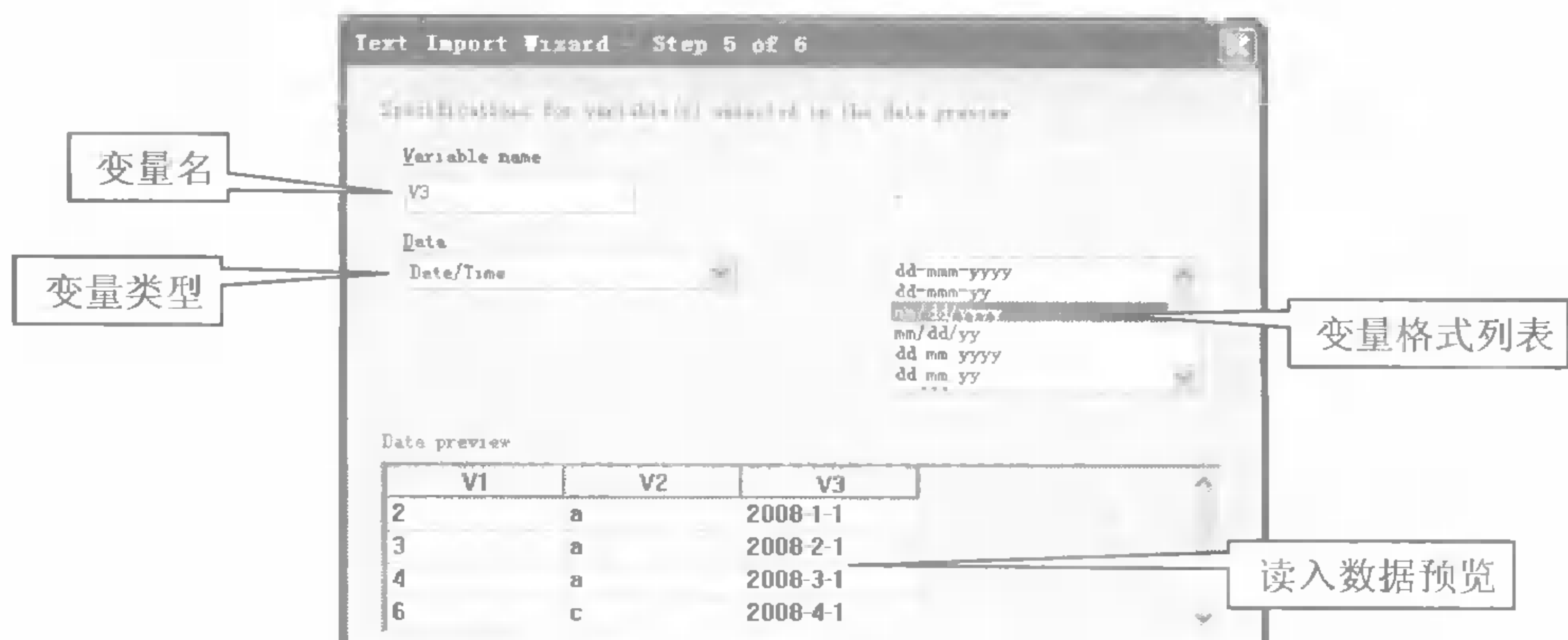


图 2-28 Text Import Wizard - Step5 of 6 对话框

⑥ 单击下一步按钮进入 Step 6 of 6 对话框，如图 2-29 所示，单击完成按钮导入数据，导入后的数据视图与图 2-26 相似。

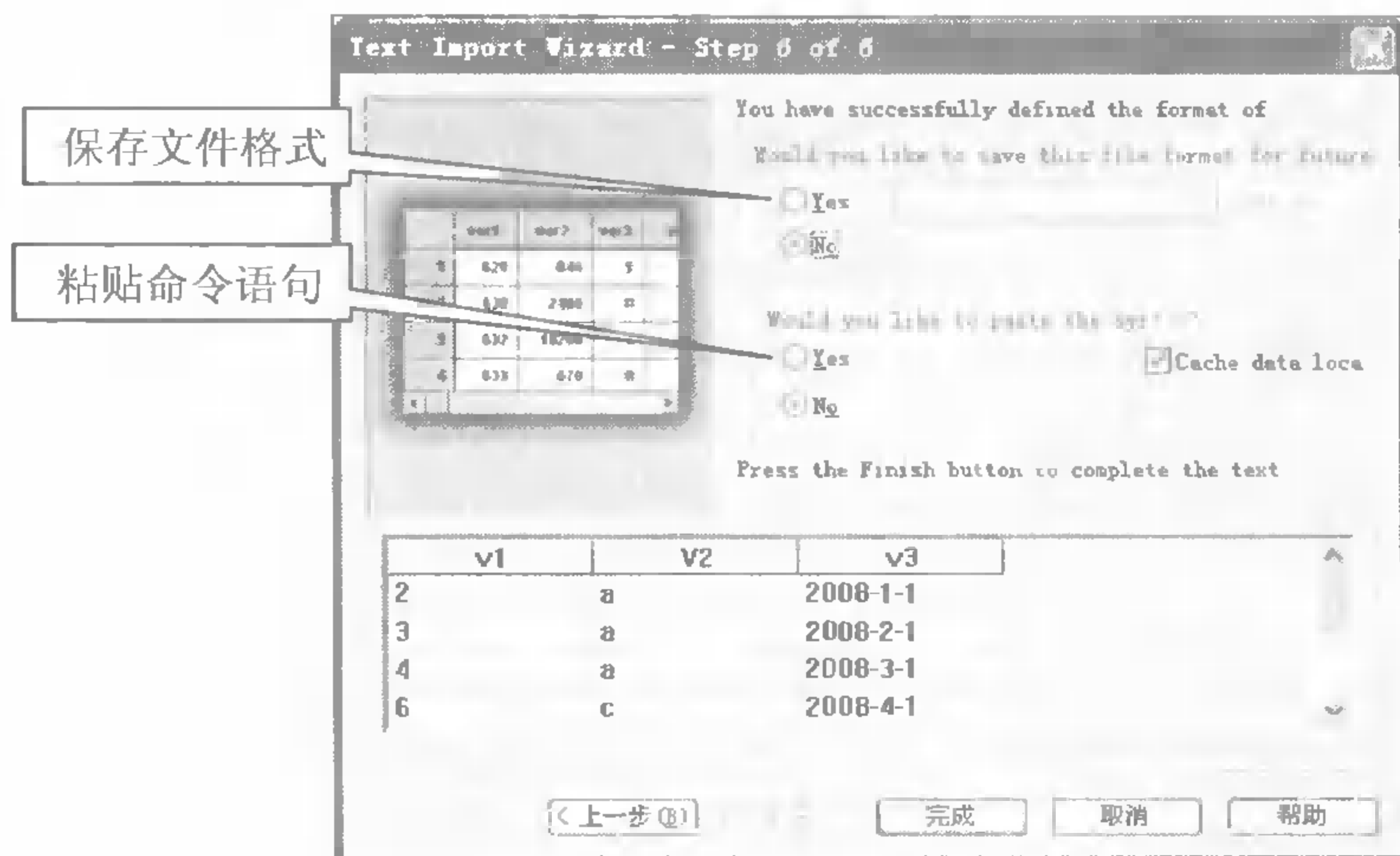


图 2-29 Text Import Wizard - Step6 of 6 对话框

⑦ 最后，同样可以把读入的数据存为其他格式的文件或直接导入指定数据库中。

在图 2-29 中，单击选中保存文件格式栏的 Yes 按钮，保存后的文件可直接用于如图 2-27 所示的 Step 1 of 6 对话框，设置其为预置格式文件（predefined format）作为读入标准；单击选中粘贴命令语句栏的 Yes 按钮，可以将整个数据导入过程的命令语句粘贴到 Syntax Editor 窗口，如图 2-30 所示。

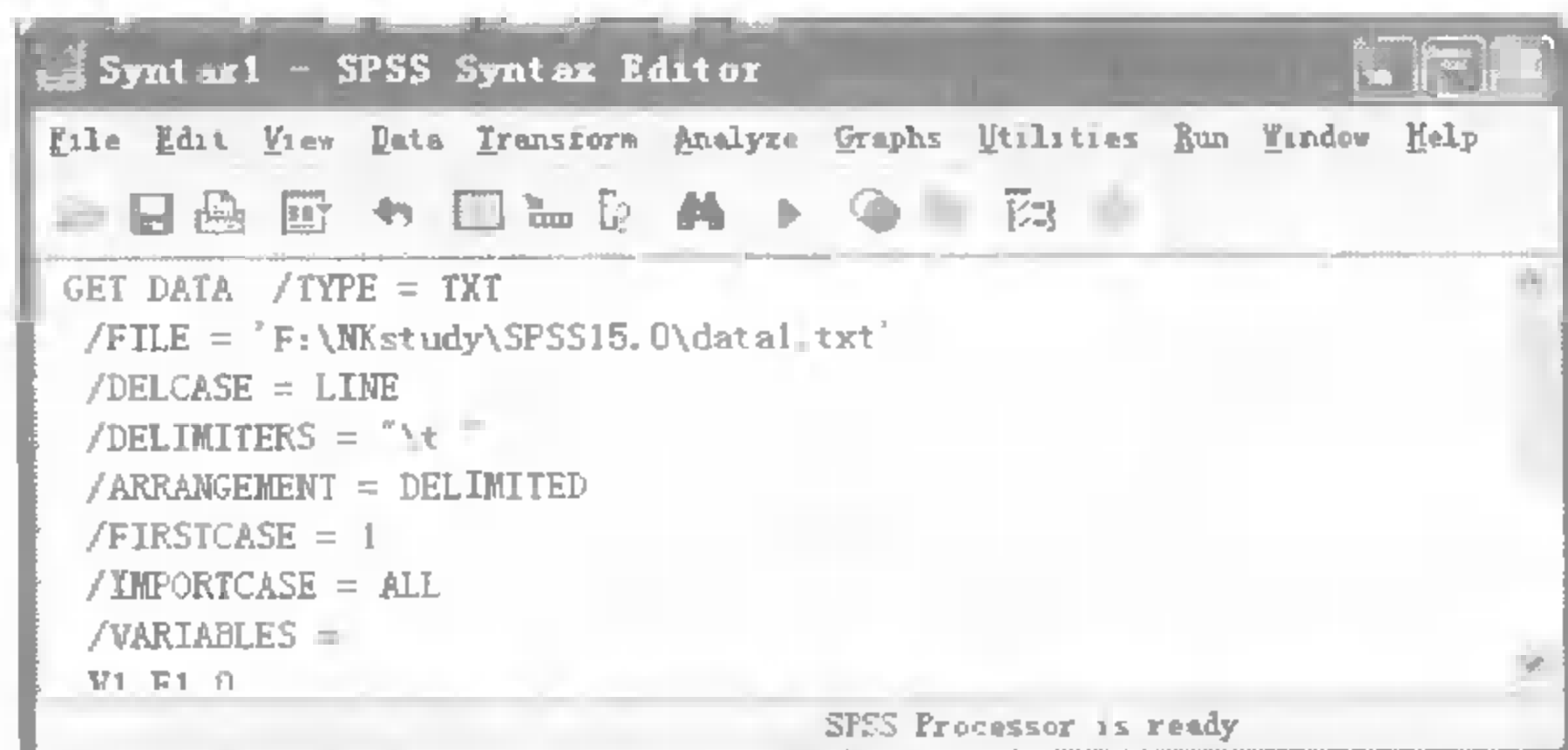


图 2-30 自动粘贴代码

## 2. 利用数据库查询导入数据

依次单击菜单“File→Open Database→New Query...”，打开如图 2-31 所示的数据库导入向导，通过此向导能够建立复杂的 SQL 查询，可以有选择地打开某些数据库中的表，并且允许同时在几个数据库的表之间建立查询。

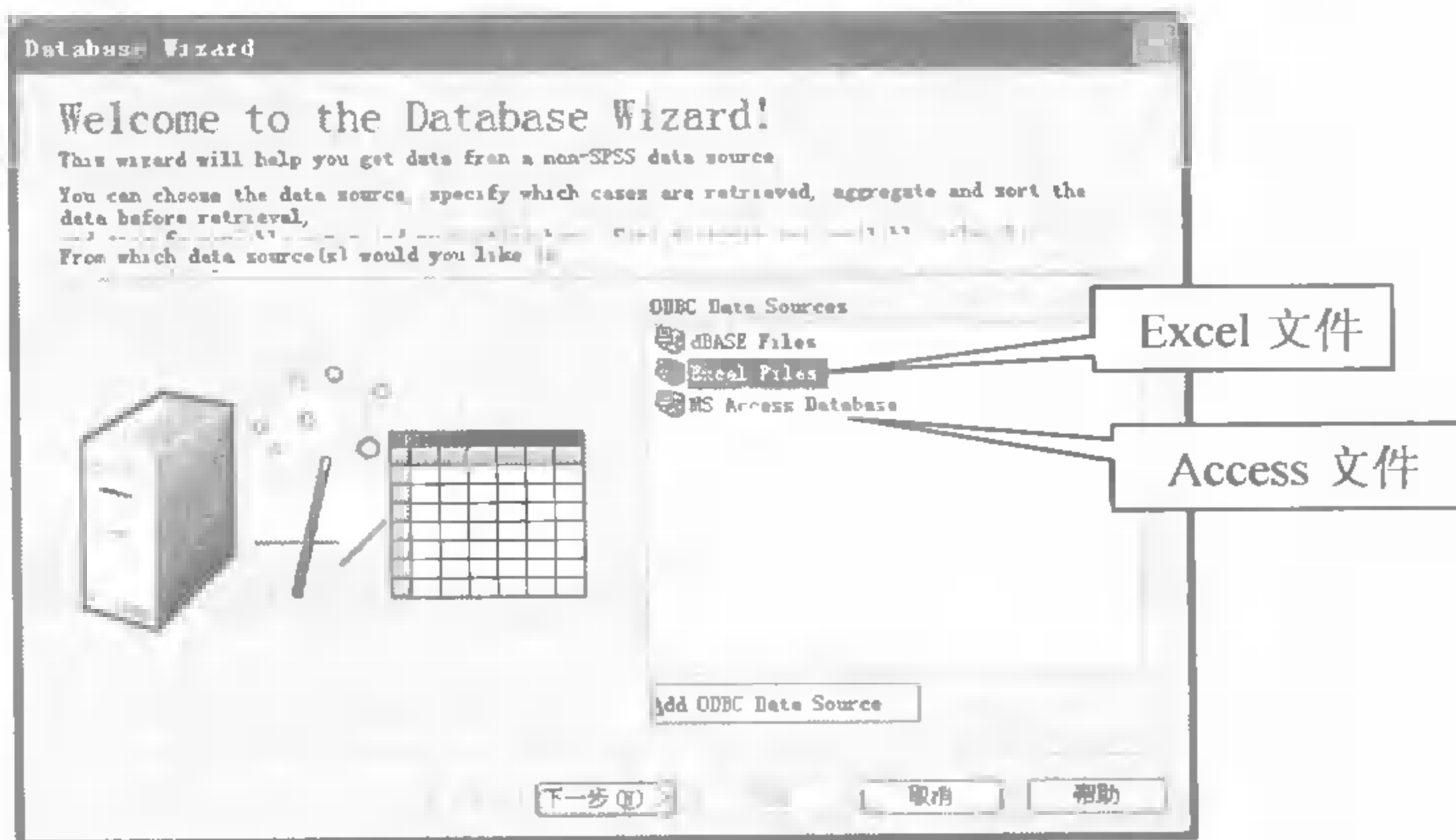


图 2-31 打开数据库向导

SPSS 是通过 ODBC 数据源方式建立和各种数据库的链接的，单击 Add ODBC Data Source...按钮，可以添加用户指定的数据源，通过桌面开始菜单“程序→管理工具→数据源(ODBC)”，也可以建立和编辑与数据源相关的内容。下面仍以打开 Excel 文件为例进行说明。

在图 2-31 中，单击选中 Excel Files 项，单击下一步按钮弹出如图 2-32 所示的数据源文件选择对话框。单击 Browse 按钮弹出的文件选择对话框里，选中文件“Data1.xls”并打开，返回图 2-32，单击 OK 按钮，进入如图 2-33 所示的向导界面。

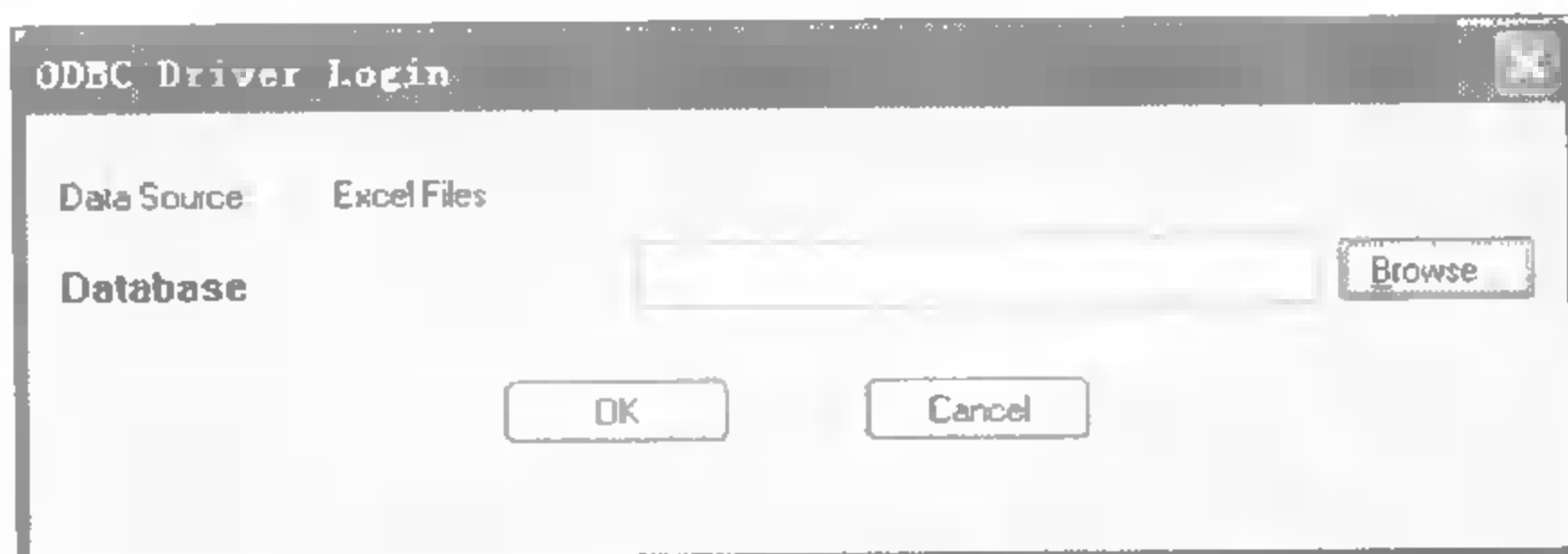


图 2-32 打开数据库向导

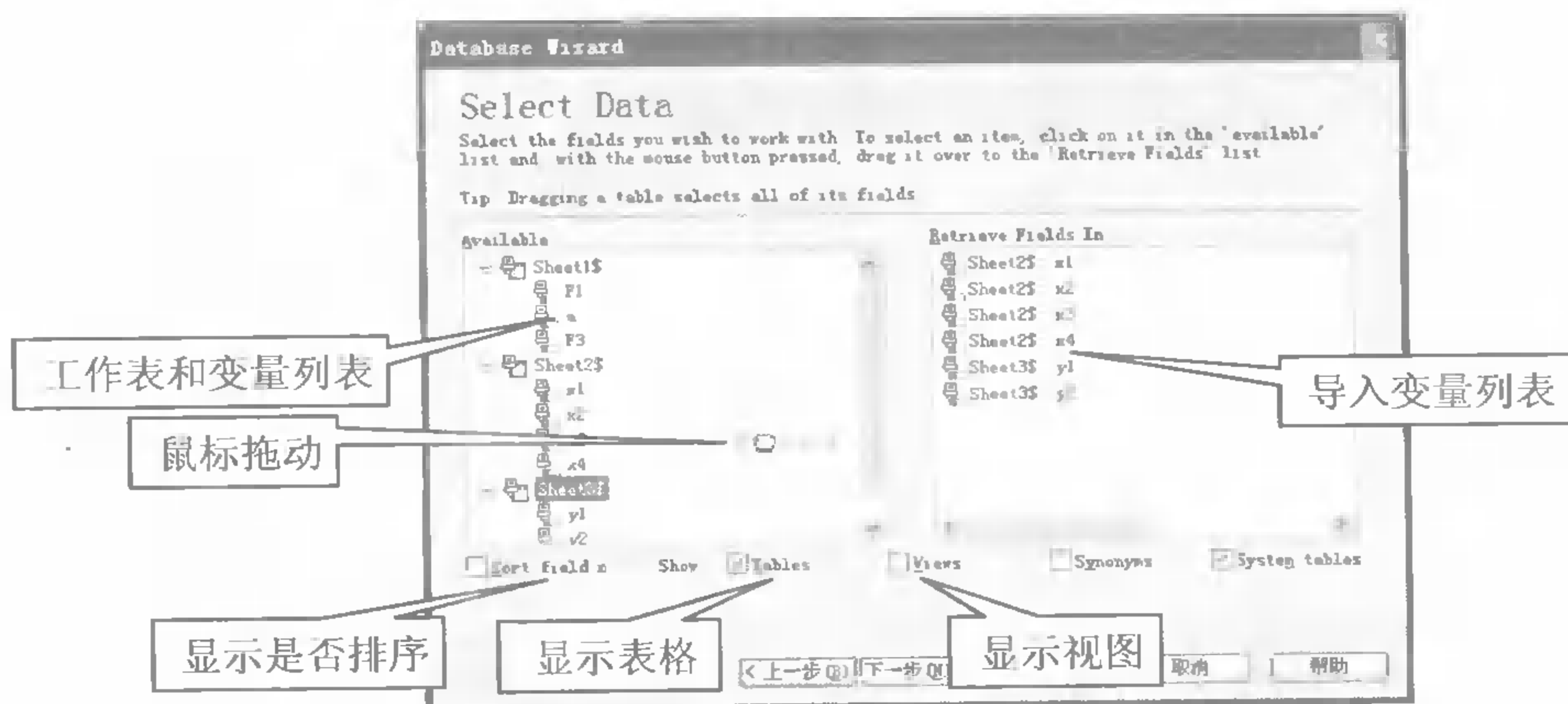


图 2-33 打开数据库向导

在图 2-33 中，左侧的 Available 列表显示读入的 3 个工作表（Sheet1\$~Sheet3\$）以及各表所包含的变量名（注意与快速打开时不同），所有表格的第一行都当作变量名，若第一行数据为非字符型，则自动以“F+列号”作为变量名，例如“F1”。将光标移动至 Available 列表的某个项目上（例如 Sheet2\$），单击并拖动至右侧的 Retrieve 列表框，松开左键，将所选项目作为要导入的变量加入其中。拖动单个变量名可以逐个添加，拖动数据表名可以同时添加此表中的所有变量。

单击下一步按钮，进入如图 2-34 所示的建立表格间链接的对话框。此处也需要用鼠标拖动来建立链接。把光标移动到预建立链接的源表格的变量名上（Sheet2\$ 的第一个 x.Number 的 x 处），待光标变为手型时按住不放，拖动至要建立链接的目标表格的变量名（Sheet3\$ 的 y.Number 的 y 处），松开鼠标后即可在相应的变量名之间出现一条黑色线条，表明链接已经建立。

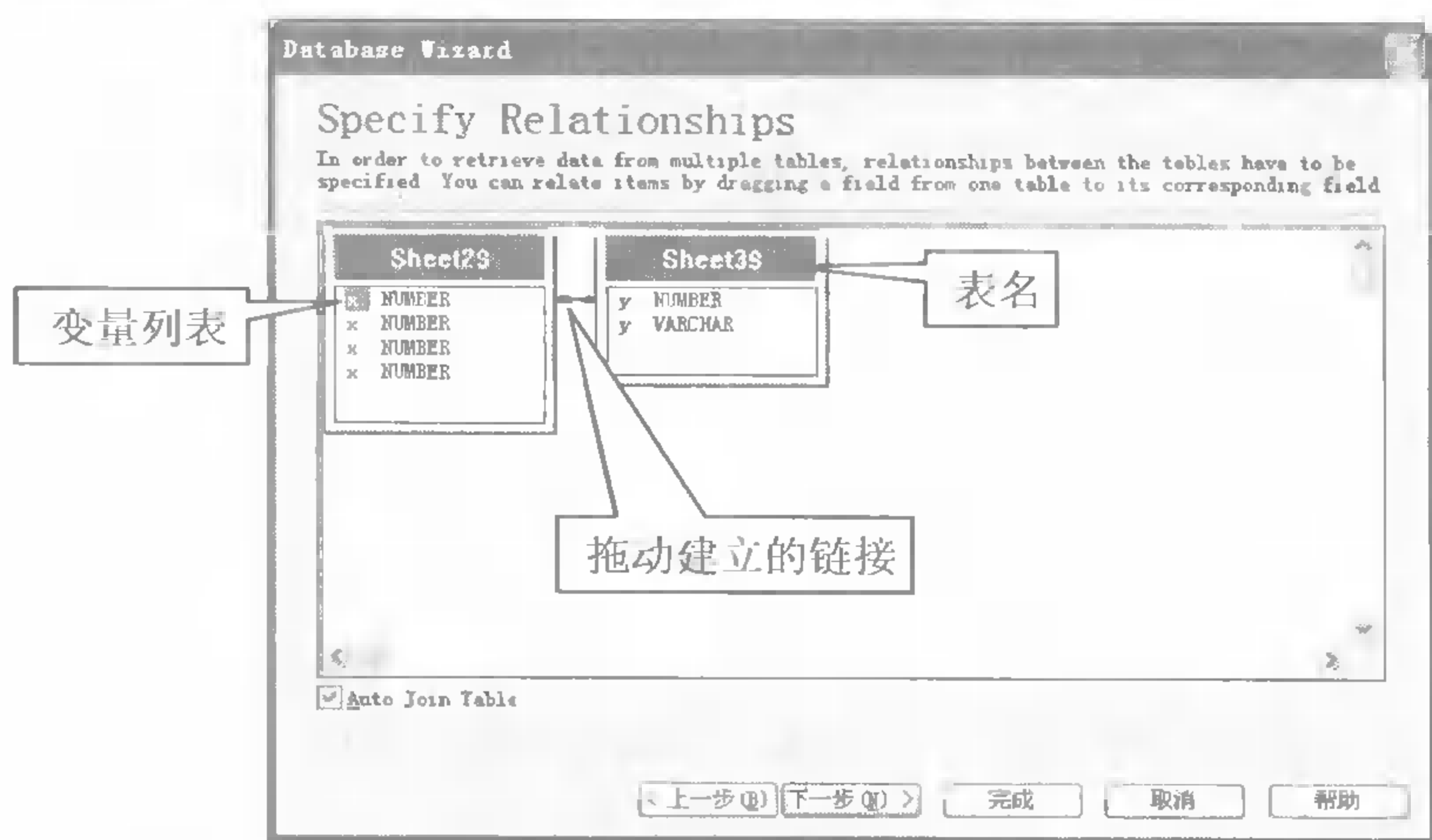


图 2-34 多表连接 (1)

单击代表链接的黑色线条，按 Delete 键可以删除链接。双击此线条，弹出如图 2-35 所示的对话框，在此编辑本条链接的属性，此处有 3 个划有线条的按钮，各自的意义如下：第 1 个表示只包括那些使链接的两个变量取值相等的记录；第 2 个表示包括 Sheet3 全部的记录，以及 Sheet2 中与 Sheet3 建立链接的两个变量相等的记录；第 3 个表示包括 Sheet2 全部的记录，以及 Sheet3 中与 Sheet2 建立链接的两个变量相等的记录。单击相应的按钮即可完成设定，单击 OK 按钮返回图 2-34。

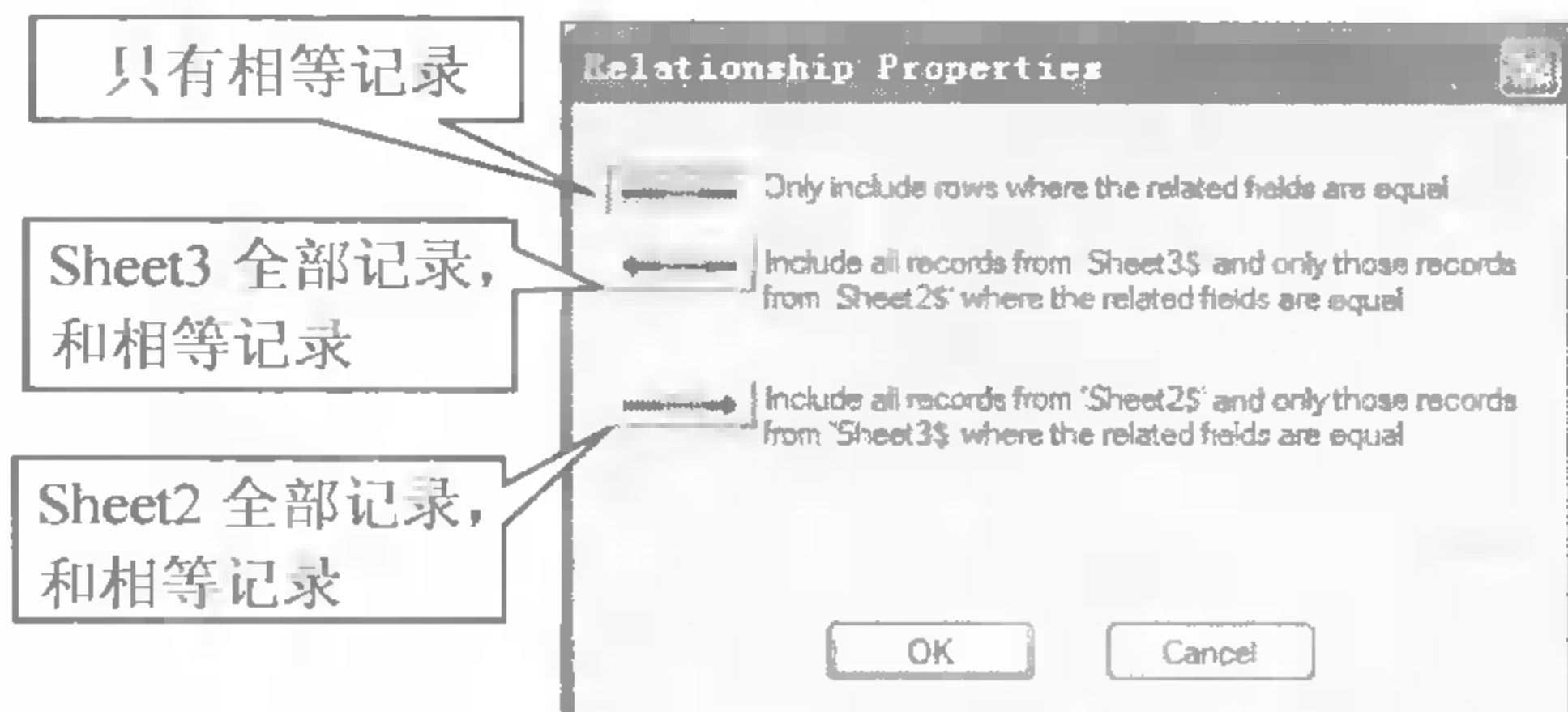


图 2-35 多表连接 (2)

在图 2-34 中单击下一步按钮，弹出图 2-36 所示的建立多表查询复合条件的界面，在下拉列表中选择相应的变量或操作符，就可以建立需要的查询条件，例如 Sheet2\$:x1 = Sheet3\$:y1 或者



直接在表格中输入变量或操作符亦可建立查询条件。左侧的变量列表 (Field) 和函数列表 (Functions) 中的内容, 可以拖放至右侧预输入相同内容的单元格中: 底部的 Use Random Sampling 复选框, 设置关于随机抽样的选项。

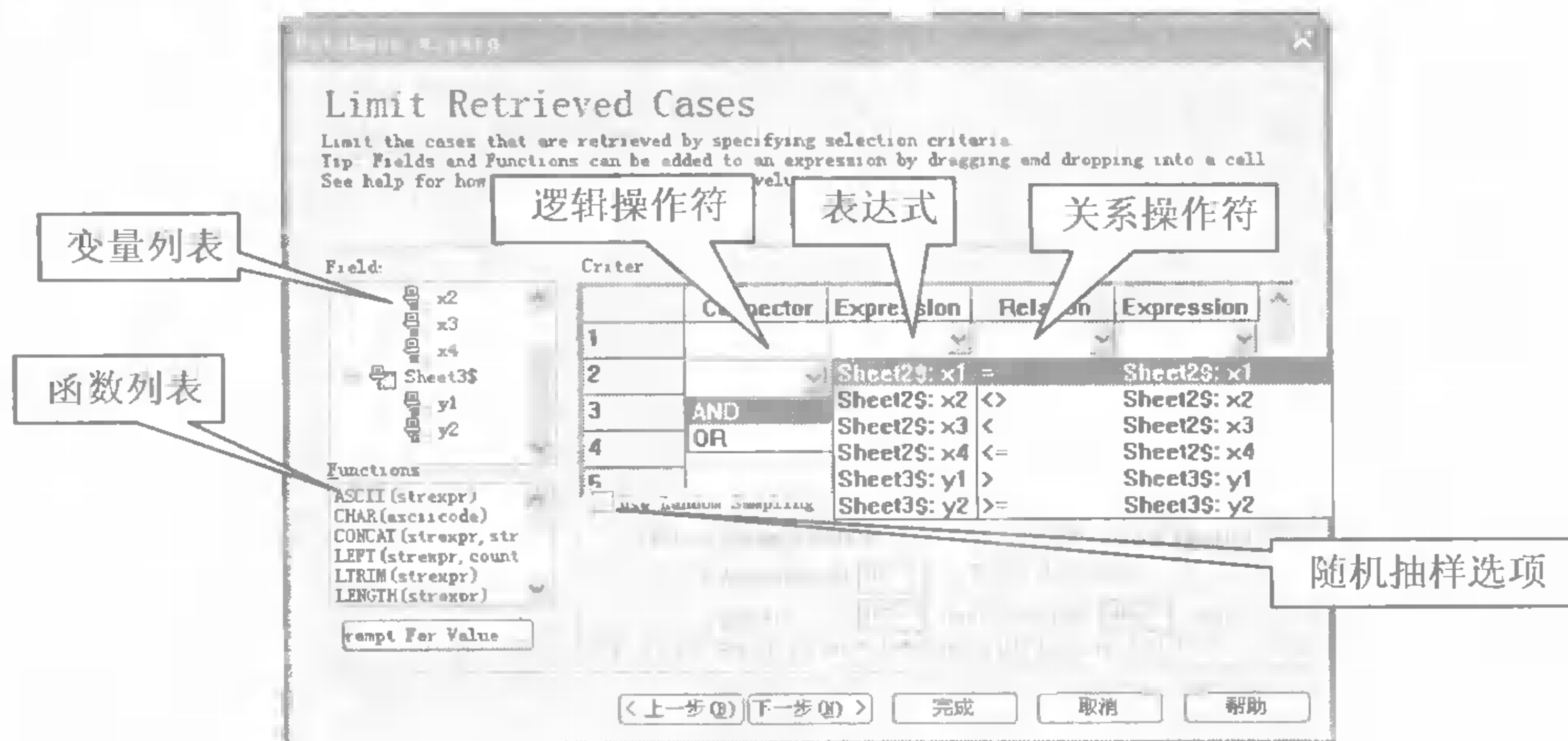


图 2-36 建立查询条件

单击下一步按钮, 进入如图 2-37 所示的变量性质查看和设置对话框, 导入后的变量名 (Result Variable Name) 可以直接在相应的单元格中进行编辑和修改; 数据类型 (Data Type) 不可以修改; Recode to Numeric 复选框表示把字符型变量转换为数值型; 底部的 Width 输入框, 用于指定字符串型变量的宽度。

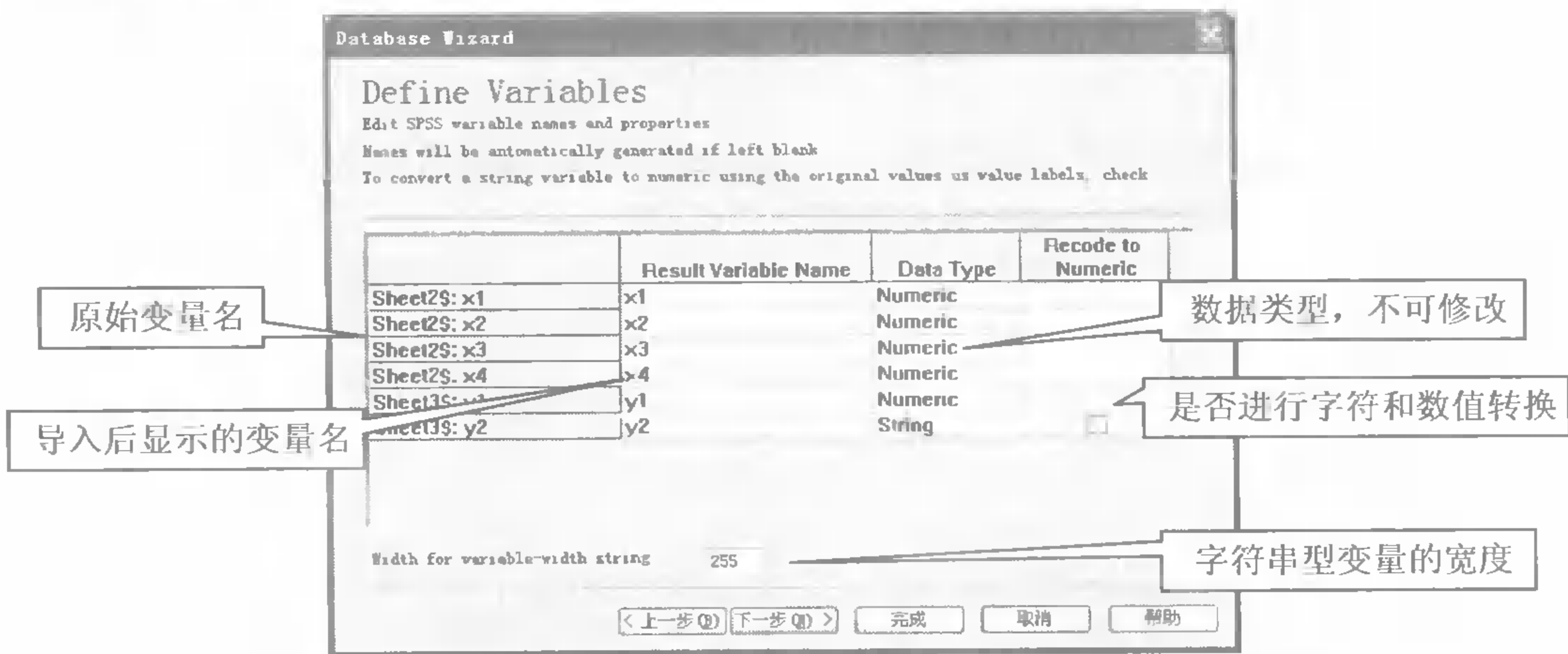


图 2-37 变量性质修改

在图 2-37 中, 单击下一步按钮, 进入如图 2-38 所示的 SQL 语句编辑窗口。SQL query 编辑框记录的就是前几步设置的复合查询条件的 SQL 语句, 若用户熟悉 SQL 语句, 在此可以直接修改或重新编写。Retrieve 单选项, 表示立即执行设置好的查询; Paste 单选项, 表示把整个导入数据库过程的 SPSS 代码粘贴至 Syntax Editor 窗口, 以便修改和保存, 这种自动粘贴代码的功能对编写重复执行的或者自动化数据处理的程序很有用处, 关于 Syntax 命令的语法和实例, 请参考文档 SPSS 15.0 Command Syntax Reference (可由 SPSS.com 获得)。底部的 Save query to file 栏, 指定把查询过程保存至一个 SPSS Query File (\*.spq) 格式的文件, 以备通过菜单 “File→Open Database→Edit Query (Run Query)” 直

接打开、编辑或运行。

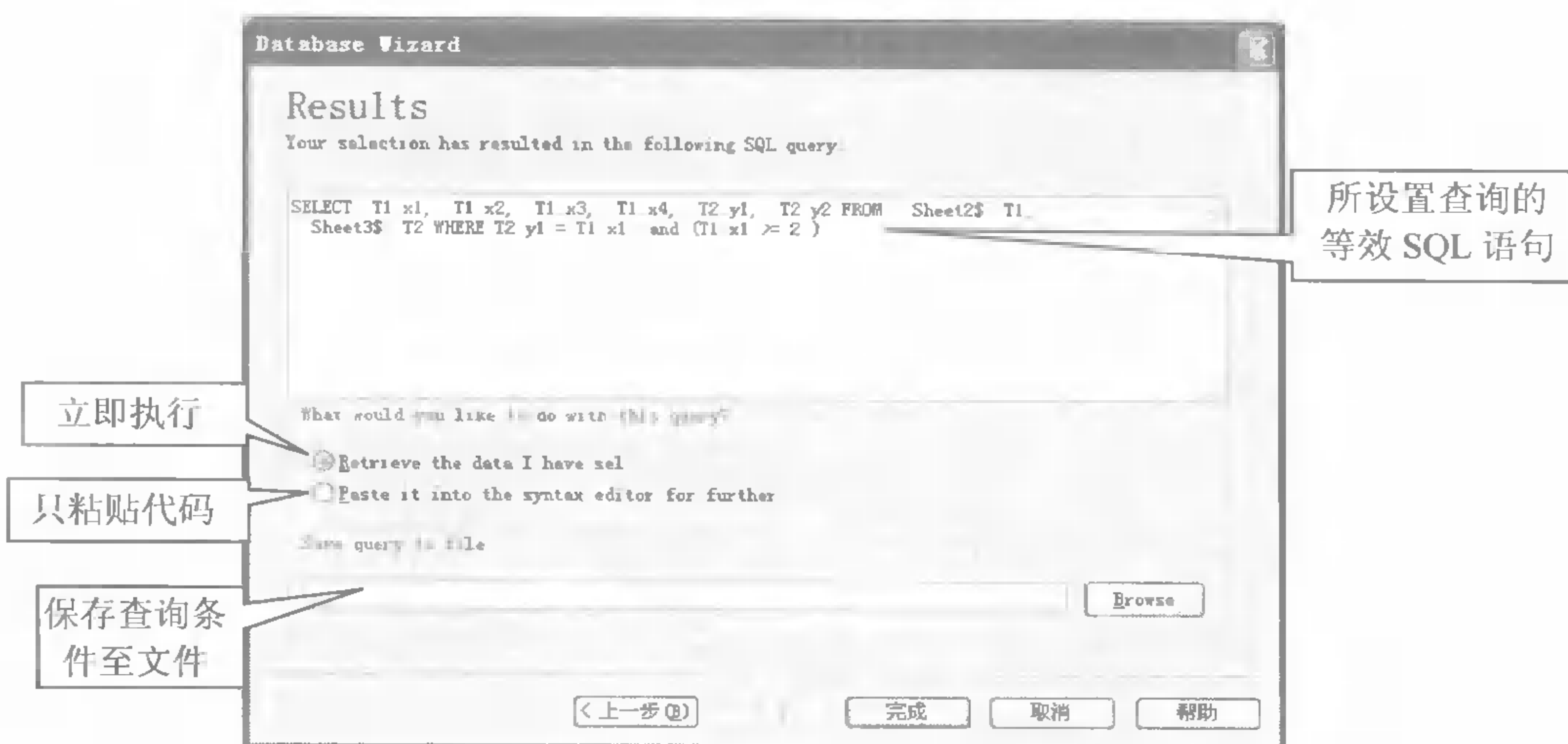


图 2-38 SQL 语句编辑窗口

在图 2-38 中，单击完成按钮，导入 Data Editor 的数据，如图 2-39 所示。

	x1	x2	x3	x4	y1	y2
1	2.00	5.00	44.00	4.00	2.00	a
2	3.00	6.00	24.00	4.00	3.00	b
3	4.00	7.00	54.00	34.00	4.00	d
4	5.00	8.00	24.00	3.00	5.00	c
5	6.00	9.00	64.00	3.00	6.00	a

图 2-39 SQL 语句编辑窗口

对于已经打开的数据文件，可以将其导入其他数据库中的数据表里，下面以图 2-39 所示的数据文件为例，介绍将它导入 Access 数据表的过程。

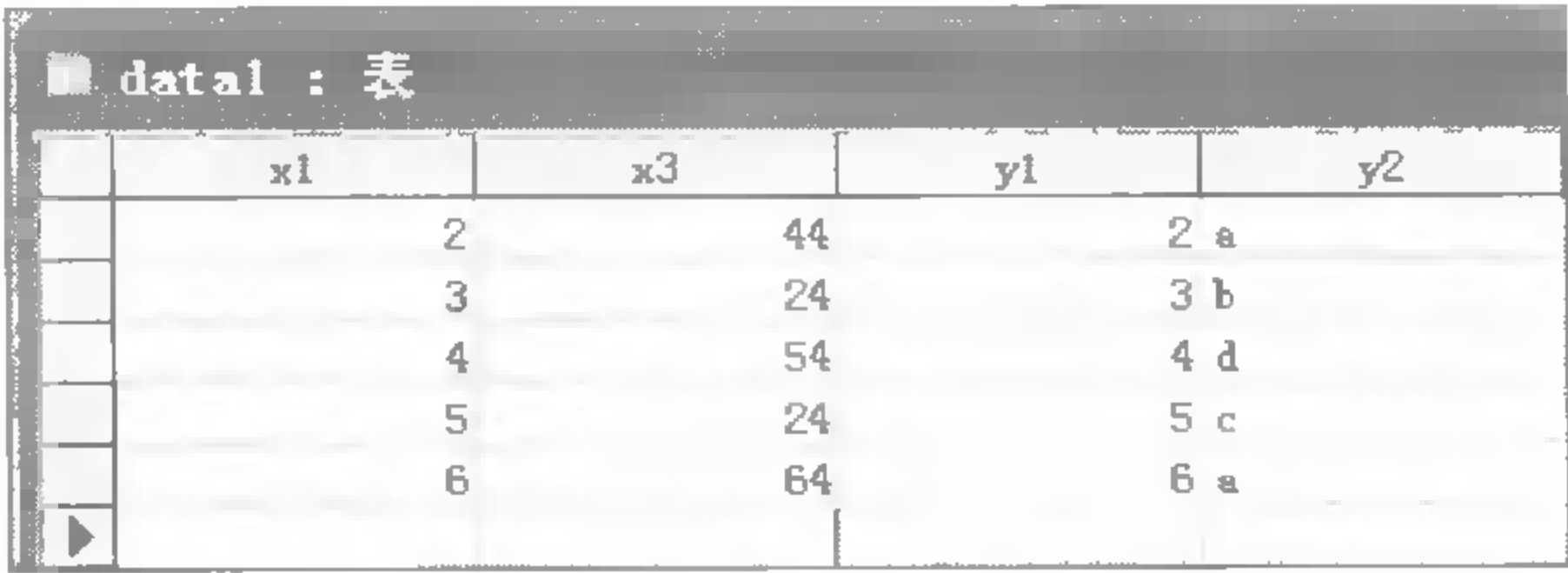
依次单击菜单“File→Export To Database...”打开与图 2-31 相似的对话框，单击选中 MS Access Database 选项，单击下一步按钮，弹出与图 2-32 相似的文件选择对话框。指定一个预导入数据的目标文件，单击 OK 按钮，进入如图 2-40 所示的选择表格对话框，各选项含义如图中标识。



图 2-40 导出至数据库特定表格

单击选中 Create a new table 单选框，并在 Name 后输入目标表格的名称“data1”。单击下

一步按钮，在随后的导出向导中会逐步提示设置导出变量及变量类型、缺失值处理等选项，最终导出后的数据可在 Access 中查看和编辑，如图 2-41 所示。



The image shows a window titled 'data1 : 表' (data1 : Table). Inside is a table with four columns: x1, x3, y1, and y2. The data rows are as follows:

x1	x3	y1	y2
2	44	2	a
3	24	3	b
4	54	4	d
5	24	5	c
6	64	6	a

图 2-41 导出至 Access 的数据

# 第3章 数据文件的操作

进行统计分析之前，需要对数据文件进行适当的整理或转换，使数据格式更加适用于将要用到的分析方法。数据文件的操作主要包括：数据排序、数据文件的分组、数据文件的合并、数据文件的转置、对变量值的求秩、对变量重新编码、计算新变量、数据汇总以及变量加权等。整理数据文件的功能通过 DATA 菜单和 Transform 菜单来完成。

## 3.1 数据文件的一般操作

本节介绍对文件和数据的一些基本操作，包括对数据的排序、文件的合并与分组、文件的转置、计算新变量等内容。


### 3.1.1 数据排序

在做数据分析时，有时需要按照某个变量的取值，重新排列各观测量在数据文件中出现的先后顺序。对观测量排序（升序或降序）的具体操作步骤如下。

（1）数据描述。小王把他 1~6 月的工资收入和支出情况记录了下来，数据文件为“月收入 and 支出数据.sav”，原始数据格式如图 3-1 所示。通过观察发现，原始数据是按照月奖金的升序排列，下面通过 Sort cases 过程对数据按照指定的变量进行排序。

	月工资	月奖金	其它收入	支出	月份
1	1300.00	1200.00	700.00	1000.0	5
2	1300.00	1300.00	800.00	1500.0	6
3	1000.00	1400.00	900.00	700.00	3
4	1000.00	1500.00	600.00	1300.0	2
5	1000.00	1600.00	.00	1400.0	1
6	1300.00	1800.00	.00	900.00	4

图 3-1 小王 1~6 月份的收入支出情况

（2）数据排序的参数设置。依次单击菜单“Data→Sort cases”执行排序过程，其主设置界面如图 3-2 所示。首先在变量列表选中月工资、支出两个变量，然后单击  按钮，将其选入排序变量列表；在 Sort by 列表选中月工资，再单击选中 Sort Order 栏的 Ascending 单选项；在 Sort by 列表选中支出，再单击选中 Sort Order 栏的 Descending 单选项。

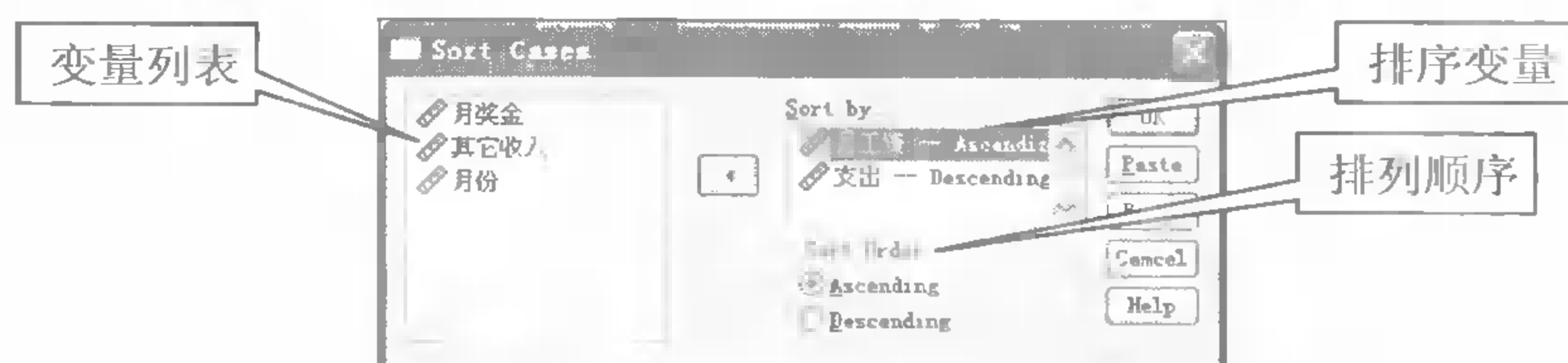


图 3-2 数据排序的设置界面

下面来详细介绍各设置选项的含义。

- Sort by 列表框，用于从变量列表选入排序变量，如果选入了多个排序变量，将自动按照它们在此列表的显示次序，依次对数据进行排序。
- Sort Order 栏，用于选择变量排序的方式：Ascending 升序或 Descending 降序。

(3) 结果显示。

在图 3-2 中，单击 OK 按钮运行，Data Editor 窗口中当前数据集的排序结果如图 3-3 所示，先按照“月工资”升序排列，再对月工资相同时的数据按“支出”降序排列。


	月工资	月奖金	其它收入	支出	月份
1	1000.00	1600.00	.00	1400.0	1
2	1000.00	1500.00	600.00	1300.0	2
3	1000.00	1400.00	900.00	700.00	3
4	1300.00	1300.00	800.00	1500.0	6
5	1300.00	1200.00	700.00	1000.0	5
6	1300.00	1800.00	.00	900.00	4

图 3-3 排序后的结果

### 3.1.2 数据文件的分组

分组就是根据需要对原始数据进行重新排序，使某一变量取值相同的个案集中到一起，便于观察和比较。

(1) 数据描述。本节对某小学 10~13 岁儿童的身高和体重数据进行分析，数据文件为“儿童的身高和体重数据.sav”，原始数据格式如图 3-4 所示。下面介绍对其分组的步骤。

(2) 数据分组的参数设置。依次单击菜单“Data→Split file”执行数据分组的功能，其主设置界面如图 3-5 所示。单击选中 Compare groups 单选项，在变量列表中单击选中性别、年龄两个变量，然后单击  按钮，将其选入分组变量列表。

下面详细介绍各设置选项的含义：

- Analyze all cases, do not create groups(分析全部数据但是不建立分组)，选中此项可恢复未分组前的状态，这是系统默认选项。

	no	gend	age	high	weight
1	06	0	10	1.46	38
2	18	0	11	1.56	48
3	17	0	11	1.50	40
4	07	1	10	1.48	38
5	12	1	10	1.43	43
6	26	1	12	1.64	60
7	15	0	10	1.48	39
8	45	0	10	1.43	35
9	21	1	11	1.55	46
10	27	1	11	1.55	44
11	09	1	11	1.46	40
12	27	1	12	1.59	56
13	04	0	11	1.52	42
14	05	1	10	1.43	41
15	10	0	12	1.60	63
16	14	1	12	1.59	42
17	08	1	11	1.48	40
18	03	0	11	1.55	44
19	20	0	10	1.44	37
20	19	0	12	1.62	56
21	01	1	12	1.60	55
22	02	1	12	1.62	53
23	11	0	11	1.55	55
24	16	0	10	1.44	38
25	13	1	11	1.46	41
26	40	1	12	1.62	49
27	25	1	11	1.55	48

图 3-4 分组前原始数据图

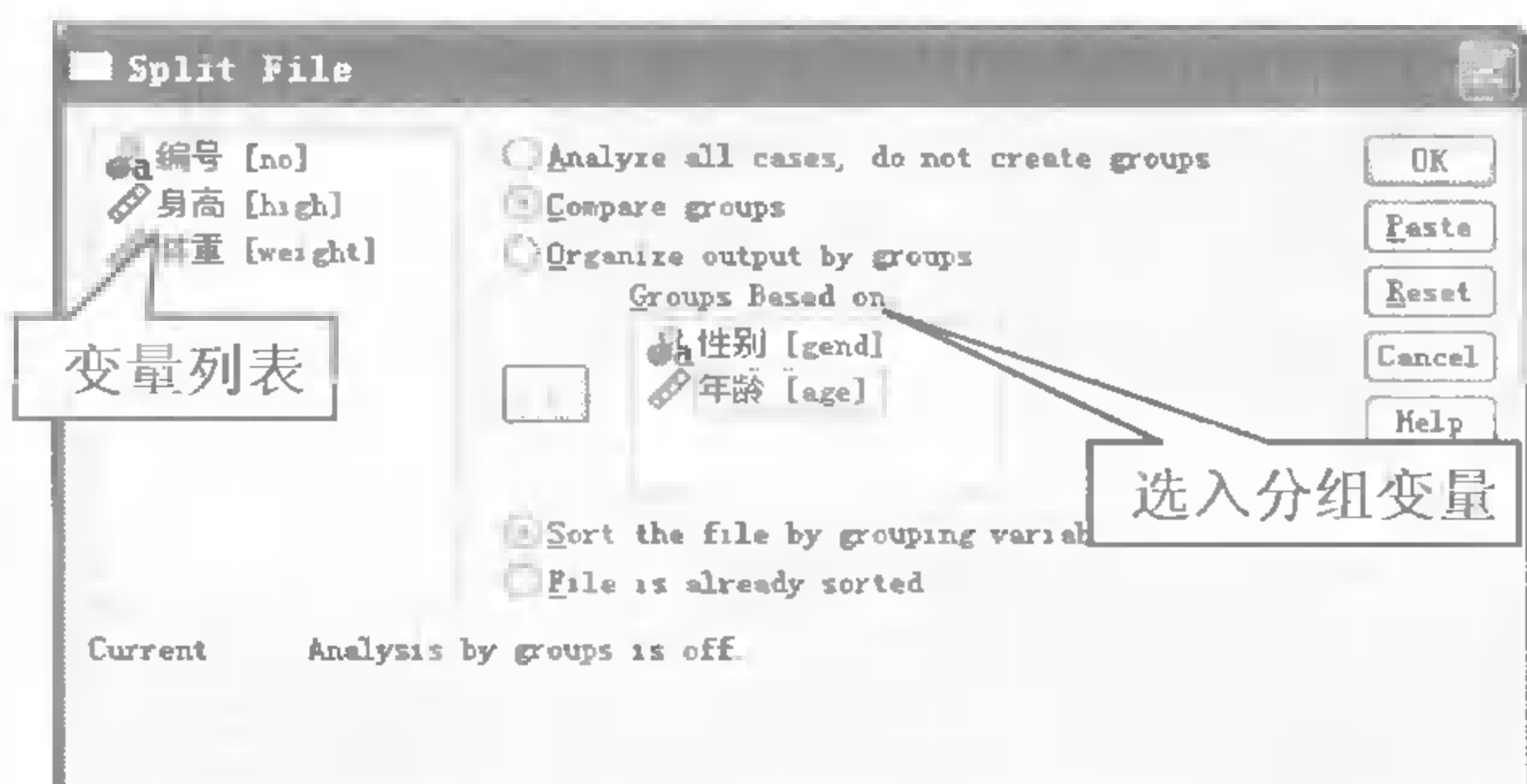


图 3-5 数据分组的参数设置

- Compare groups (对照组选项)，将文件分组以对照组的形式显示。
- Organize output by groups 按分组变量的取值排序输出。



● Groups base on 列表框用于从变量列表选入分组变量。

另外，还需指定排序方式：Sort the file by grouping variables，将数据按照分组变量的取值排序，这是系统默认选项；File is already sorted，标识数据已经按分组变量排序了，不需要重新排序。

### (3) 结果分析。

在图 3-5 中单击 OK 按钮运行，Data Editor 窗口中当前数据集的分组结果如图 3-6 所示，按照性别（gend）和年龄（age）进行了分组。注意：运行完成后，SPSS 会在 Data Editor 窗口的状态栏右侧显示 Split File On 字样，表示当前数据已经分组，且分组信息会随数据文件同时保存。

	no	gend	age	high	weight
1	06	0	10	1.	
2	15	0	10		
3	45	0	10	1.43	36
4	20	0	10	1.44	37
5	16	0	10	1.44	38
6	18	0	11	1.56	48
7	17	0	11	1.50	40
8	04	0	11	1.52	42
9	03	0	11	1.55	44
10	11	0	11	1.55	55
11	10	0	12	1.60	53
12	19	0	12	1.62	56
13	07	1	10	1.48	39
14	12	1	10	1.43	43
15	05	1	10	1.43	43
16	21	1	11	1.55	46
17	27	1	11	1.55	44
18	09	1	11	1.46	40
19	08	1	11	1.48	40
20	13	1	11	1.46	41
21	25	1	11	1.55	48
22	26	1	12	1.64	60
23	14	1	12	1.59	42
24	01	1	12	1.60	55
25	02	1	12	1.62	53
26	40	1	12	1.62	49
27	27	1	13	1.59	55

图 3-6 分组后的数据格式

### 3.1.3 数据文件的合并

SPSS 可以将多个数据文件进行合并，具体合并方式分为对记录的合并和对变量的合并两种，合并记录时要求被合并的文件具有相同的变量。

下面分别以实例操作的方式，简单介绍它们的操作步骤。

(1) 数据描述。对于小王前几个月的工资收入和支出情况，记录在文件“月收入 and 支出数据.sav”中，数据格式如图 3-1 所示。本节增加了小王在 7~10 月份的收入和支出数据，另存在文件“月收入 and 支出记录补充.sav”里，数据格式如图 3-7 所示；还增加了小王在 1~6 月份的生活支出和娱乐支出两个变量，数据存在文件“月收入 and 支出变量补充.sav”里，数据格式如图 3-8 所示。本节就通过这 3 个文件来演示合并操作。

	月工资	月奖金	其它收入	支出	月份
1	1200.00	1500.00	600.00	1400.00	7
2	1200.00	1700.00	600.00	1300.00	8
3	1200.00	1400.00	900.00	700.00	9
4	1300.00	1800.00	900.00	900.00	10

图 3-7 月收入 and 支出的记录补充

	月份	生活支出	娱乐支出
1	1	200.00	100.00
2	2	300.00	300.00
3	3	200.00	200.00
4	4	200.00	100.00
5	5	220.00	200.00
6	6	300.00	200.00

图 3-8 月收入 and 支出的变量补充

#### (2) 对记录的合并。

① 选择文件。打开“月收入 and 支出数据.sav”和“月收入 and 支出记录补充.sav”两个文

件, 切换到“月收入 and 支出数据.sav”文件的 Data Editor 界面, 然后依次单击菜单“Data→Merge file→Add cases”执行文件记录合并功能, 其主设置界面如图 3-9 所示, 可选的合并方式有如下两个: An open dataset 选项, 表示从当前打开的数据集选择合并文件, 下面的列表显示了当前打开的可用数据集名称; An external data file 选项, 表示读取外部的数据文件进行合并, 选中后单击右侧的 Browse 按钮指定文件路径和文件名。

② 变量设置。在图 3-9 中, 单击选中 An open dataset 单选项, 并在下面的列表里选中“月收入 and 支出记录补充.sav”行, 然后单击 Continue 按钮, 进入如图 3-10 所示的设置界面。

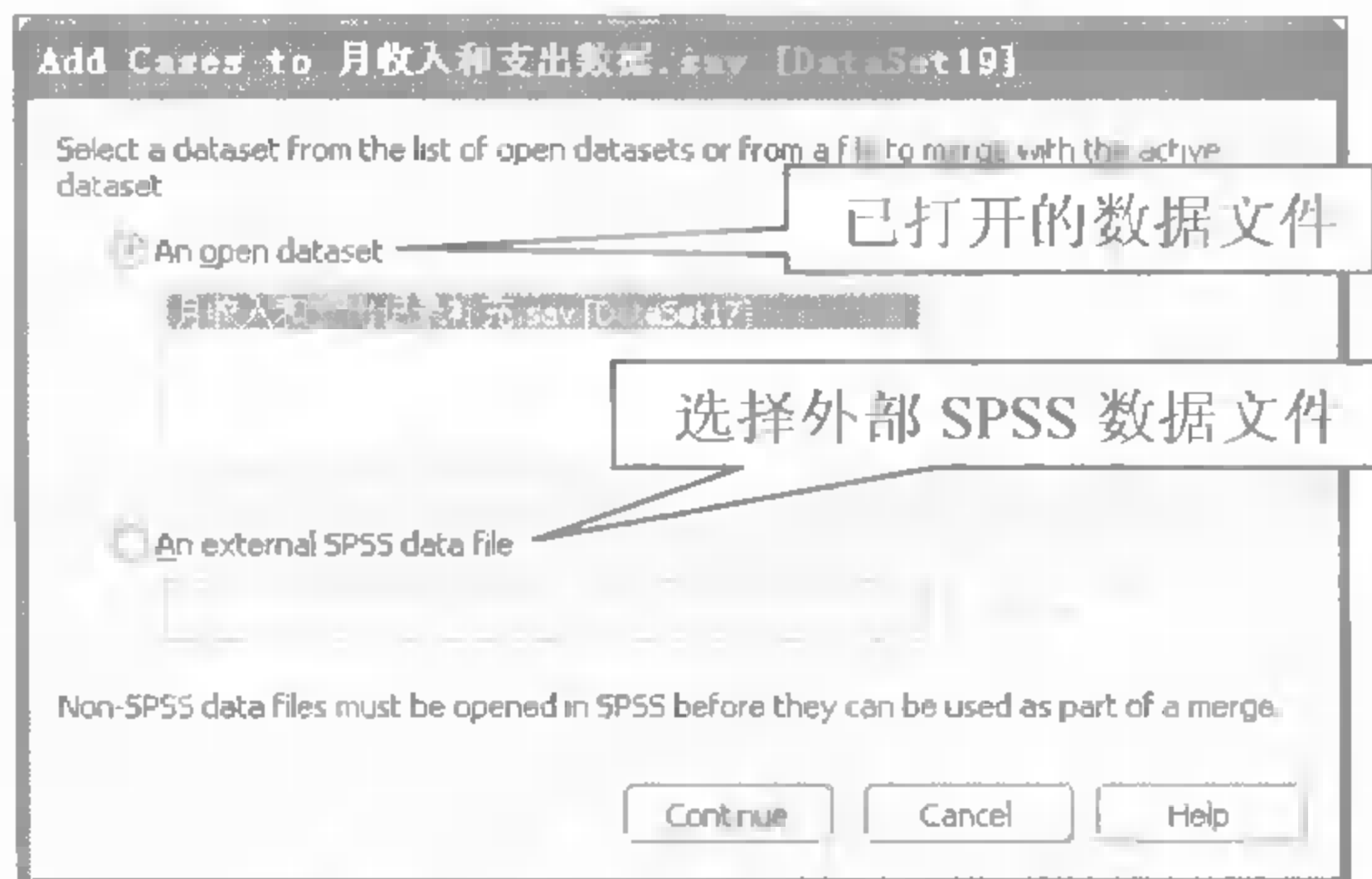


图 3-9 合并记录的文件选择

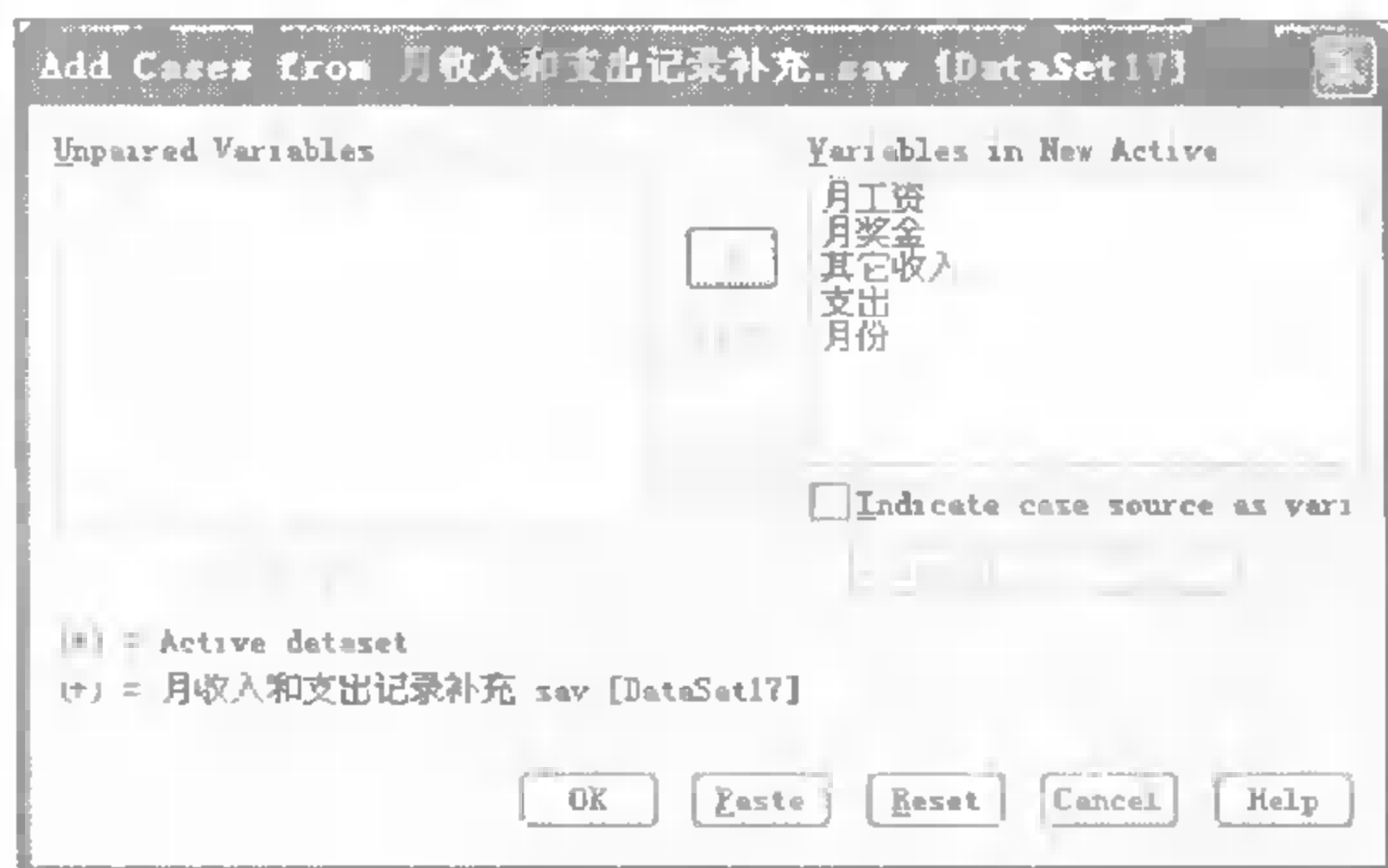


图 3-10 合并记录的参数设置

下面详细介绍各设置选项的含义:

- 变量列表 Unpaired variable 列表显示两个文件中不匹配的变量名 (包括变量名不同的变量和变量名相同而变量定义不同的变量); Variables in new Active Dataset 列表显示合并后的新数据集包含的变量, 默认显示的是两个文件里公有的变量名。
- Indicate case source as variable 复选框选中后, 在合并后的数据集生成一个新变量, 表示每个记录的来源 (取值 0 代表来自源文件, 取值 1 代表来自被合并的文件), 在其下的输入框指定这个新变量的名称。

在图 3-10 中, 列表里显示的变量名后会带下面两个标识: “(\*)” 表示此变量来自源数据文件, 例如“月份(\*)”; “(+)” 表示此变量来自被合并的文件, 例如“月份(+)”。

默认情况下, SPSS 自动把两个文件都有的变量选入右侧的列表。如果要在结果里包含只在一个文件出现的变量, 将其从左侧列表选入右侧的列表即可。如果要在结果里强行合并两个文件中的两个变量, 先在左侧的列表把它们都选中 (两个变量必须来自不同的文件), 然后单击 Pair 按钮, 就可以把强行合并后的变量选入右侧的列表。

③ 记录合并的输出。图 3-10 中两个文件的所有变量都匹配了, 单击 OK 按钮运行, Data Editor 窗口的当前数据集如图 3-11 所示, 两个文件的数据记录行合并到一起了。


注意: 合并后的数据放在当前打开的源数据文件“月收入 and 支出数据.sav”里, 保存后将更新此文件的内容。

### (3) 对变量的合并。

① 选择文件和变量设置。先打开“月收入 and 支出数据.sav”和“月收入 and 支出变量补充.sav”两个文件, 切换到“月收入 and 支出数据.sav”文件的 Data Editor 界面。

	月工资	月奖金	其它收入	支出	月份
1	1300.00	1200.00	700.00	1000.00	5
2	1300.00	1300.00	800.00	1500.00	6
3	1000.00	1400.00	900.00	700.00	3
4	1000.00	1500.00	600.00	1300.00	2
5	1000.00	1600.00	.00	1400.00	1
6	1300.00	1800.00	.00	900.00	4
7	1200.00	1500.00	600.00	1400.00	7
8	1200.00	1700.00	600.00	1300.00	8
9	1200.00	1400.00	900.00	700.00	9
10	1300.00	1800.00	900.00	900.00	10

图 3-11 记录合并后的结果

依次单击菜单“Data→Merge file→Add Variables”执行文件变量合并功能，其设置界面如图 3-12 所示，单击选中 An open dataset 单选项，并在下面的列表里选中“月收入 and 支出变量补充.sav”行，单击 Continue 按钮进入如图 3-13 所示的设置界面。在图 3-13 中勾选 Match 复选框，在 Excluded Variables 列表中单击选中月份(+)变量，单击 Key Variables 列表左侧的  按钮，将月份变量作为主键变量选入，New 列表中的月份(\*)变量消失，如图 3-14 所示。

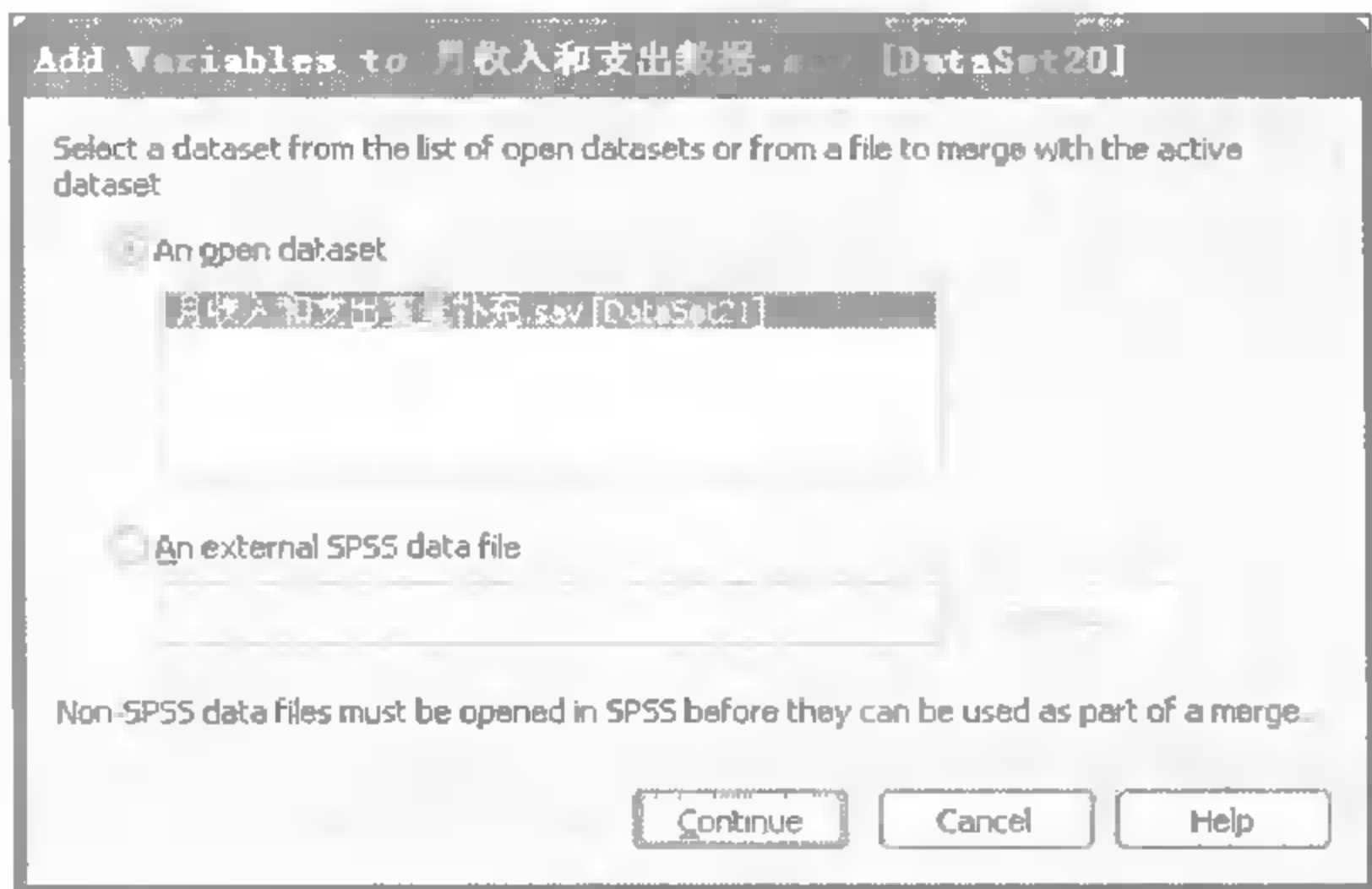


图 3-12 合并变量的文件选择

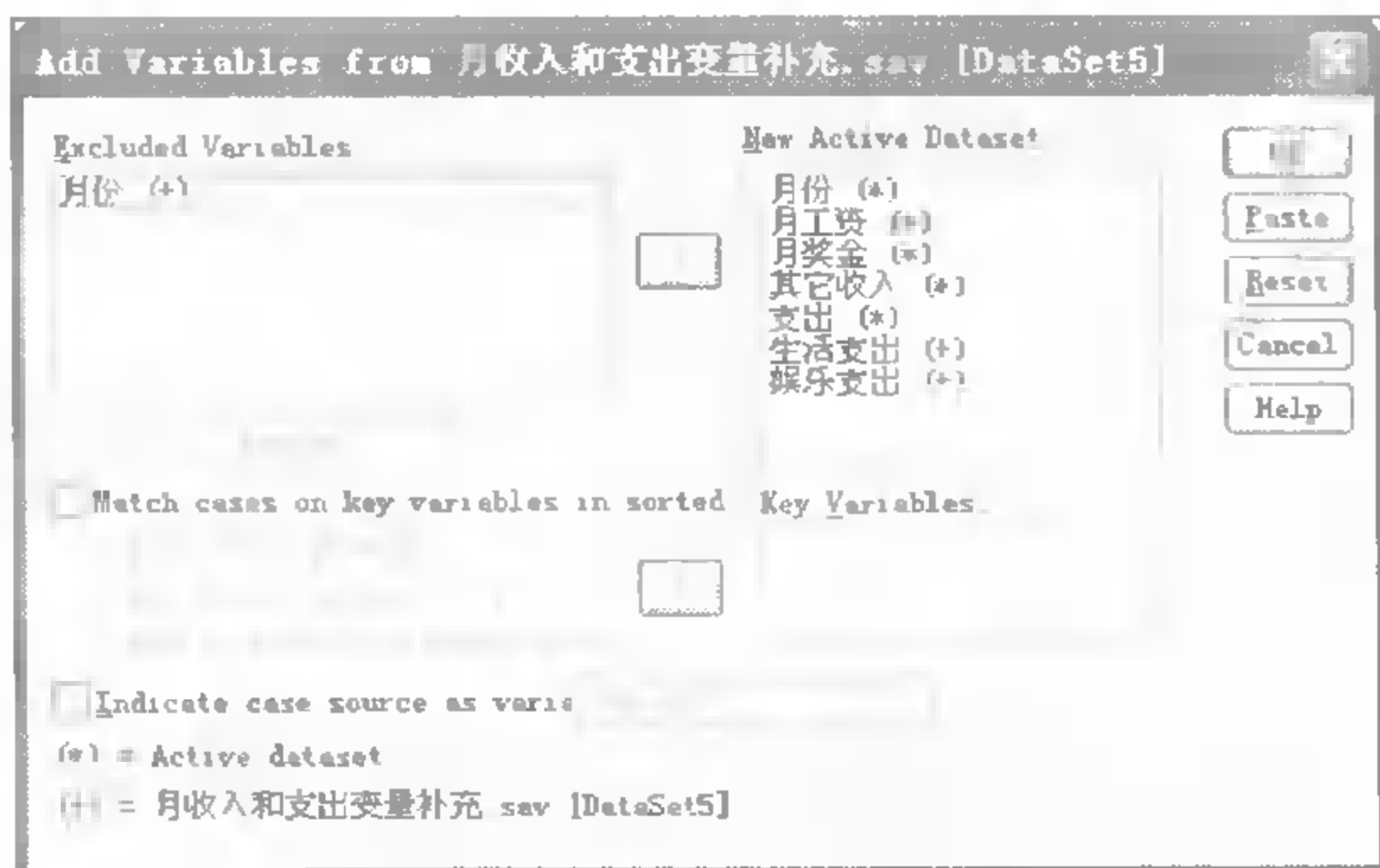


图 3-13 合并变量的参数设置 1

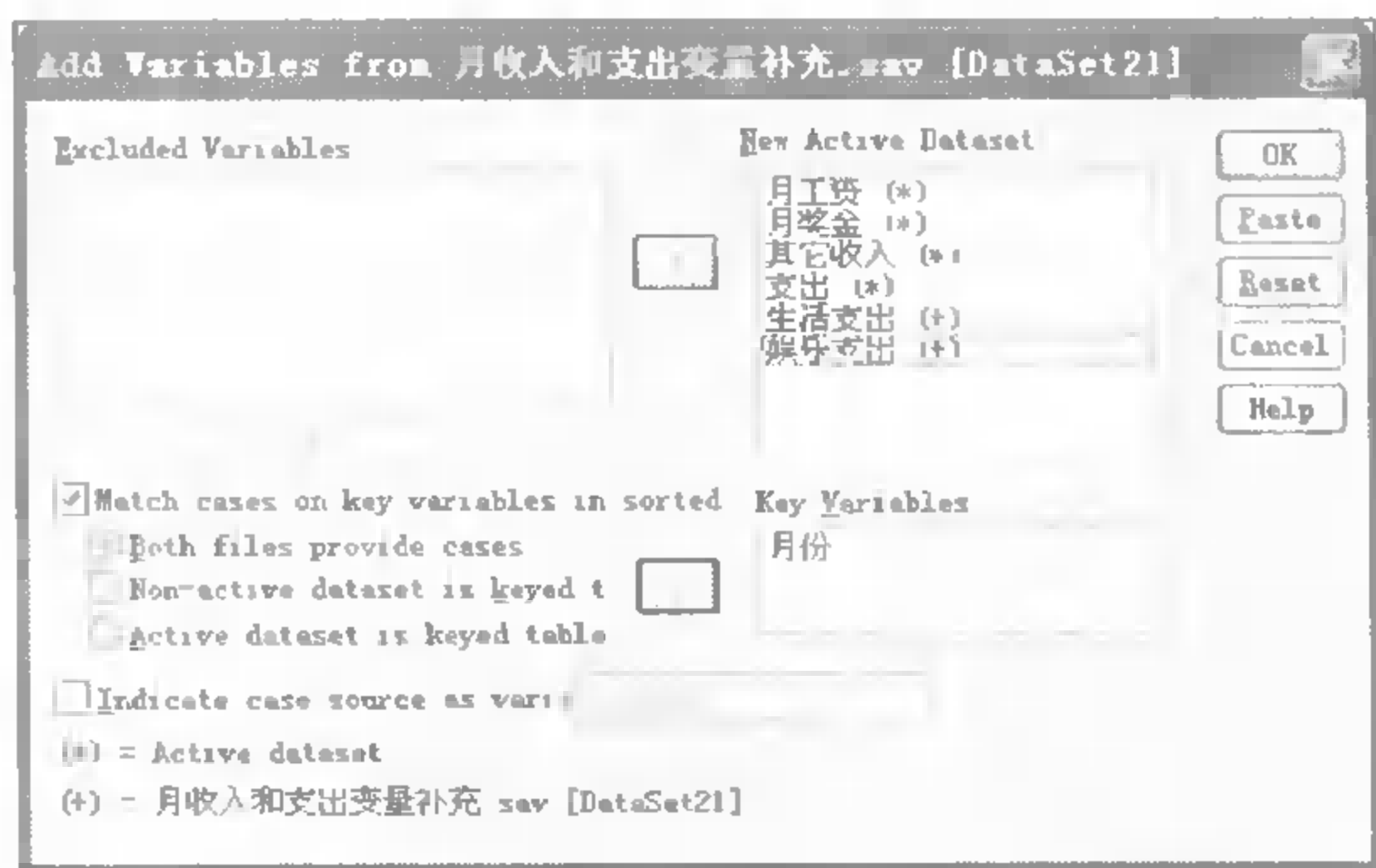


图 3-14 合并变量的参数设置 2

下面详细介绍各设置选项的含义。

在图 3-13 中，Excluded Variables 列表显示出现在 2 个初始文件里但不出现在合并后文件里的变量；New Active Dataset 列表显示合并后的数据集所包含的变量。默认情况下，SPSS 自动把两个文件独有的变量显示在右侧的列表中；如果要在输出集中加入新的变量，先在左侧的列表选中它（单击下面的 Rename 按钮可以对其重新命名），然后单击中间的箭头就可以把选中的变量选入右侧的列表。

对不同文件的变量进行合并时，可以指定合并时的主键（关键变量），它用来标识和匹配不同文件的记录行。选中 Match Case on Key Variable in Sorted 复选框，就可以将主键变量选入 Key Variables 列表，同时还可以指定主键的出现位置，有如下 3 个选项：

- ① Both files provide case，表示主键同时出现在两个文件中。
- ② External file is keyed table，表示主键只出现源文件中，以当前源文件为基准，外部文件匹配源文件的主键值，如果匹配成功，外部文件的新变量值就加入到合并后数据集的新变量中，匹配不成功则不加入。
- ③ Working Data File is keyed table，表示主键只出现在被合并的外部文件中，以外部文件为基准，源文件匹配外部文件的主键值，如果匹配成功，源文件的新变量值就加入到合并后数据集的新变量中，匹配不成功则不加入。

Indicate case source as variable 复选框，选中后在合并后的数据集生成一个新变量，用来表示每个记录的来源（取值 0 代表来自源文件，取值 1 代表来自被合并的外部文件），在右侧的输入框指定这个新变量的名称。

## ② 结果输出。

在图 3-14 中,单击 OK 按钮运行,由于选择了月份作为主键,故弹出如图 3-15 所示的警告框,提示被合并的两个文件必须已经按照主键进行了排序,否则不能正确合并。



图 3-15 关于数据排序的警告框

在图 3-15 中单击“确定”按钮,Data Editor 窗口的当前数据集如图 3-16 所示,两个文件的变量以月份为关键词合并到一起了。

	月工资	月奖金	其它收入	支出	月份	生活支出	娱乐支出
1	1000.00	1600.00	.00	1400.00	1	200.00	100.00
2	1000.00	1500.00	600.00	1300.00	2	300.00	300.00
3	1000.00	1400.00	900.00	700.00	3	200.00	200.00
4	1300.00	1800.00	.00	900.00	4	200.00	100.00
5	1300.00	1200.00	700.00	1000.00	5	220.00	200.00
6	1300.00	1300.00	800.00	1500.00	6	300.00	200.00

图 3-16 变量合并后的结果

注意:合并后的数据放在当前打开的源数据文件“月收入 and 支出数据.sav”里,保存后将更新此文件的内容。



## 3.1.4 数据文件的转置

SPSS 可以将数据编辑器中的数据进行列互换,即原来按行(列)方向排列的数据转换成按列(行)方向排列的数据。

(1) 原始数据描述。文件“地区发展样本数据.sav”里记录了某年份 5 个地区的 6 项经济指标,数据格式如图 3-17 所示。利用 Transpose 过程对其进行转置,以观察转置前后的数据排列方式有何不同。

行标识	area	educat	industry	farm	work	use	gdp	列变量
1	北京	1170593	610.66	1.94	624.3	5178	2011.31	
2	天津	433116.8	587.83	.33	427.0	5209	1336.38	
3	河北	1128001	1822.05	6.16	3382.9	2163	4256.01	
4	山西	600037.5	745.47	5.35	1429.0	1835	1601.11	
5	内蒙古	424137.7	399.42	3.78	1006.8	2141	1192.29	

图 3-17 地区发展样本的数据格式

(2) 数据转置的参数设置。依次单击菜单“Data→Transpose”打开进行数据转置的设置界面,如图 3-18 所示。在变量列表选中除地区外的所有变量,然后单击从上至下第一个  按钮,将其选入转置变量列表;在变量列表选中地区变量,然后单击从上至下第二个  按钮,将其选入标识变量选框;设置结果如图 3-19 所示。

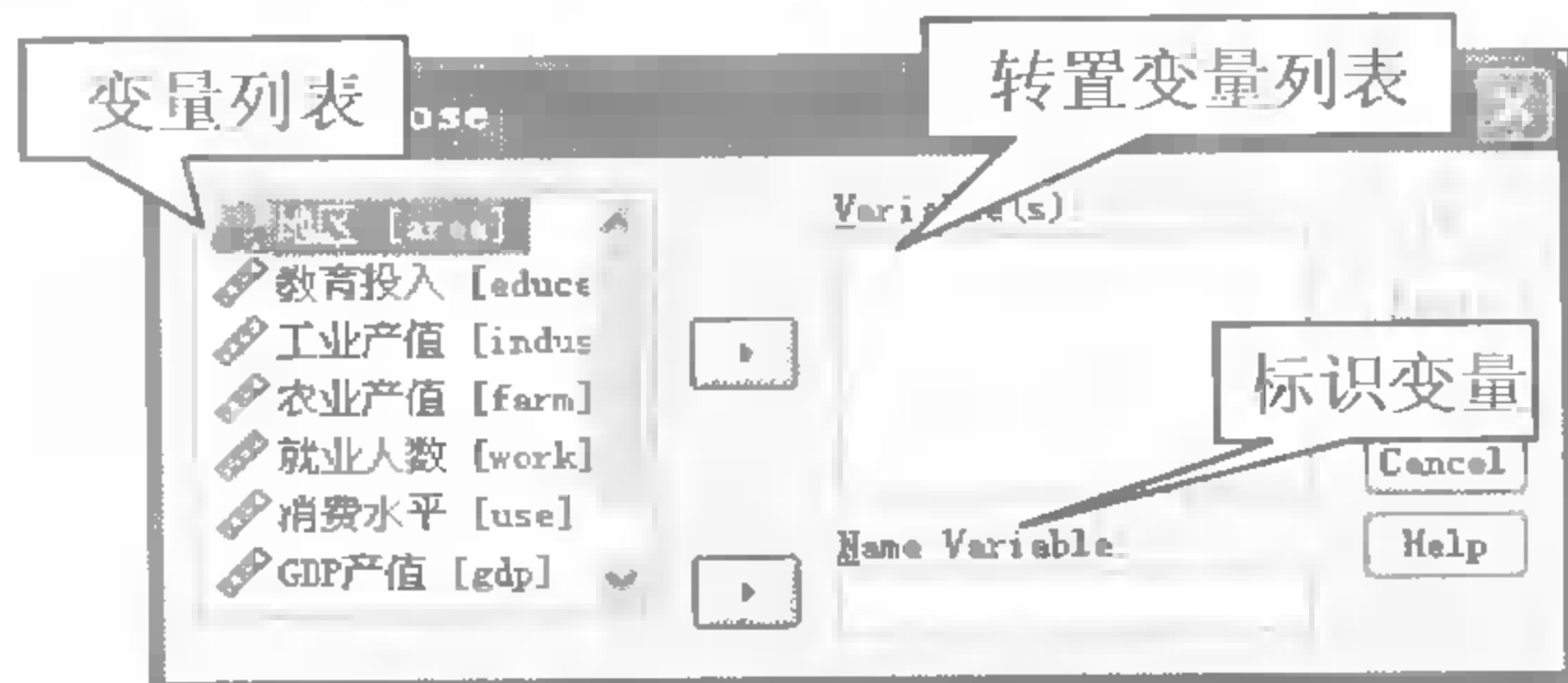


图 3-18 文件转置的设置 1

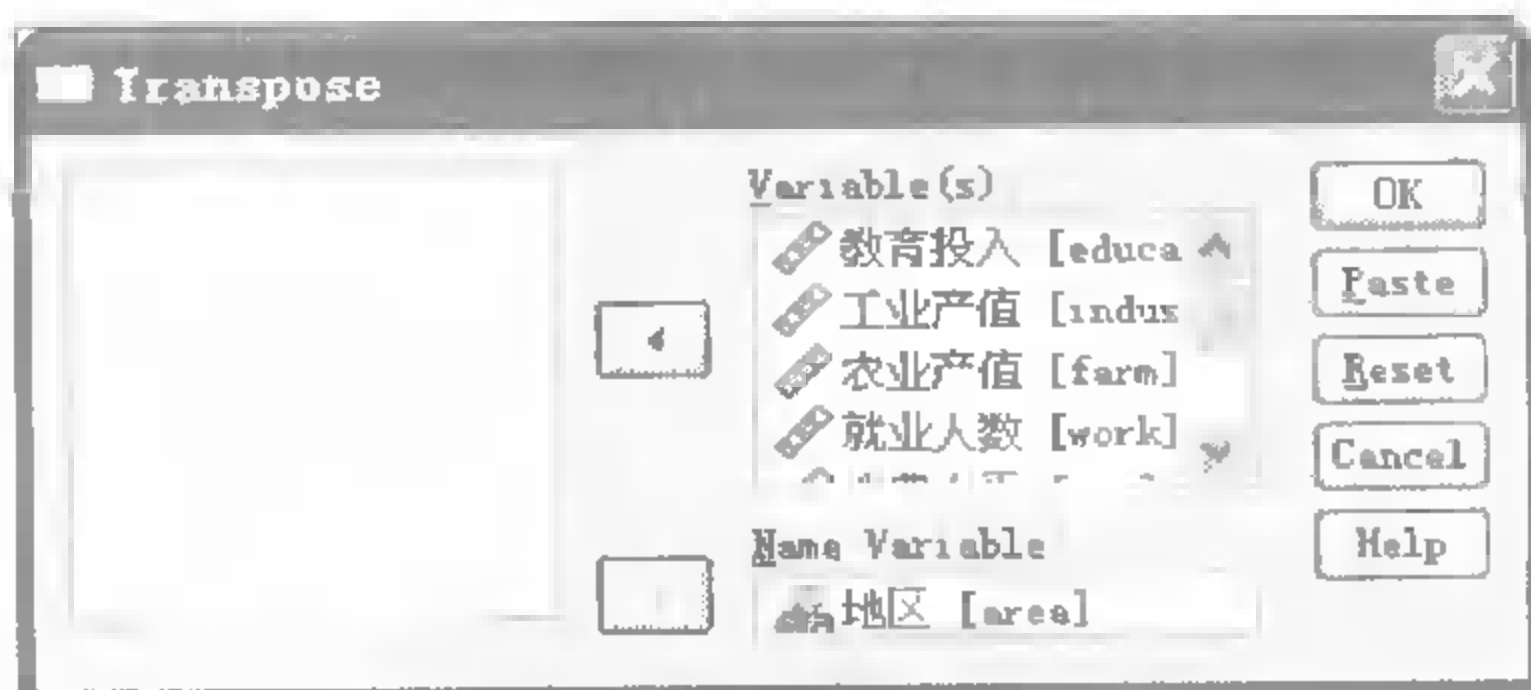


图 3-19 文件转置的设置 2



如图 3-18 所示, 变量列表显示当前数据集中的所有变量; Variable(s)列表用于从左侧的变量列表选入待转置的所有变量; Name Variable 选框用于从左侧的变量列表选入能够标识原始记录的变量, 例如学号、姓名等。

(3) 结果显示。在图 3-19 中, 单击 OK 按钮运行, 转置后生成一个新的数据集, 其数据格式如图 3-20 所示。原来的行标识变量作为新的列变量, 而原来的列变量变为了新的行标识, 且变量名记为 CASE\_LBL; 所有单元格的数据也实现了转置。

	CASE_LBL	北京	天津	河北	山西	内蒙古
1	educat	1170592.60	433116.80	1128000.80	600037.50	424147.78
2	industry	610.66	587.83	1822.05	746.47	189.42
3	farm	1.94	.33	6.16	5.35	3.78
4	work	624.30	427.00	382.90	1429.00	1006.80
5	use	5178.00	5209.00	2163.00	1835.00	2141.00
6	gdp	2011.31	1336.38	4256.01	1601.11	1192.29

图 3-20 转置后的数据格式

(4) 其他设置。图 3-19 中如果选择的被转置变量只是所有变量的一部分, 单击 OK 按钮运行后, 会弹出如图 3-21 所示的警告框, 提示用户未被转置的变量将被丢弃, 单击确定按钮继续转置, 单击取消按钮放弃转置。



图 3-21 警告对话框

### 3.1.5 变量取值的求秩

对变量取值的求秩就是求出变量取值在指定条件下的大小顺序, 使得取值按照一定的顺序进行排列, 秩就反映了取值在这个有序序列里的位置信息。如果在求秩时还指定了分组变量, 则在各个分组内分别计算和输出分析变量的秩。

(1) 数据描述。本节仍使用某小学 10~13 岁儿童的身高和体重数据, 数据文件为“儿童的身高和体重数据.sav”, 原始数据格式如图 3-22 所示。本例先对数据按照体重变量 weight 进行升序排列, 排列后格式如图 3-23 所示。


	no	gend	age	high	weight
1	45	0	10	1.43	35
2	20	0	10	1.44	37
3	06	0	10	1.46	38
4	16	0	10	1.44	38
5	07	1	10	1.48	39
6	15	0	10	1.48	39
7	17	0	11	1.50	40
8	09	1	11	1.46	40
9	08	1	11	1.48	40
10	13	1	11	1.46	41
11	04	0	11	1.52	42
12	14	1	12	1.59	42
13	12	1	10	1.43	43
14	06	1	10	1.43	43
15	27	1	11	1.55	44
16	03	0	11	1.55	44
17	21	1	11	1.55	46
18	18	0	11	1.56	46
19	25	1	11	1.55	48
20	40	1	12	1.62	49
21	10	0	12	1.60	53
22	02	1	12	1.62	53
23	11	0	11	1.55	55
24	01	1	12	1.60	55
25	27	1	13	1.59	55
26	19	0	12	1.62	56
27	26	1	12	1.64	60

图 3-22 原始数据格式

	no	gend	age	high	weight	Rweight	RAN001
1	45	0	10	1.43	35	1.000	27.000
2	20	0	10	1.44	37	2.000	26.000
3	06	0	10	1.46	38	3.000	24.500
4	16	0	10	1.44	38	3.500	24.500
5	07	1	10	1.48	39	5.000	22.000
6	15	0	10	1.48	39	5.500	22.000
7	17	0	11	1.50	40	8.000	20.000
8	09	1	11	1.46	40	8.000	20.000
9	08	1	11	1.48	40	8.000	20.000
10	13	1	11	1.46	41	10.000	18.000
11	04	0	11	1.52	42	11.500	16.500
12	14	1	12	1.59	42	11.500	16.500
13	12	1	10	1.43	43	13.500	14.500
14	06	1	10	1.43	43	13.500	14.500
15	27	1	11	1.55	44	15.500	12.500
16	03	0	11	1.55	44	15.500	12.500
17	21	1	11	1.55	46	17.000	11.000
18	18	0	11	1.56	46	18.500	9.500
19	25	1	11	1.55	48	18.500	9.500
20	40	1	12	1.62	49	20.000	6.000
21	10	0	12	1.60	53	21.500	5.500
22	02	1	12	1.62	53	21.500	5.500
23	11	0	11	1.55	55	24.000	4.000
24	01	1	12	1.60	55	24.000	4.000
25	27	1	13	1.59	55	24.000	4.000
26	19	0	12	1.62	56	26.000	2.000
27	26	1	12	1.64	60	27.000	1.000

图 3-23 排秩后的数据格式



(2) 变量求秩的参数设置。依次单击菜单“Transform→Rank Cases”打开进行变量求秩的主设置界面,如图3-24所示。在变量列表单击选中体重变量,然后单击 Variable 列表左侧的  按钮,将其选入排秩变量列表。

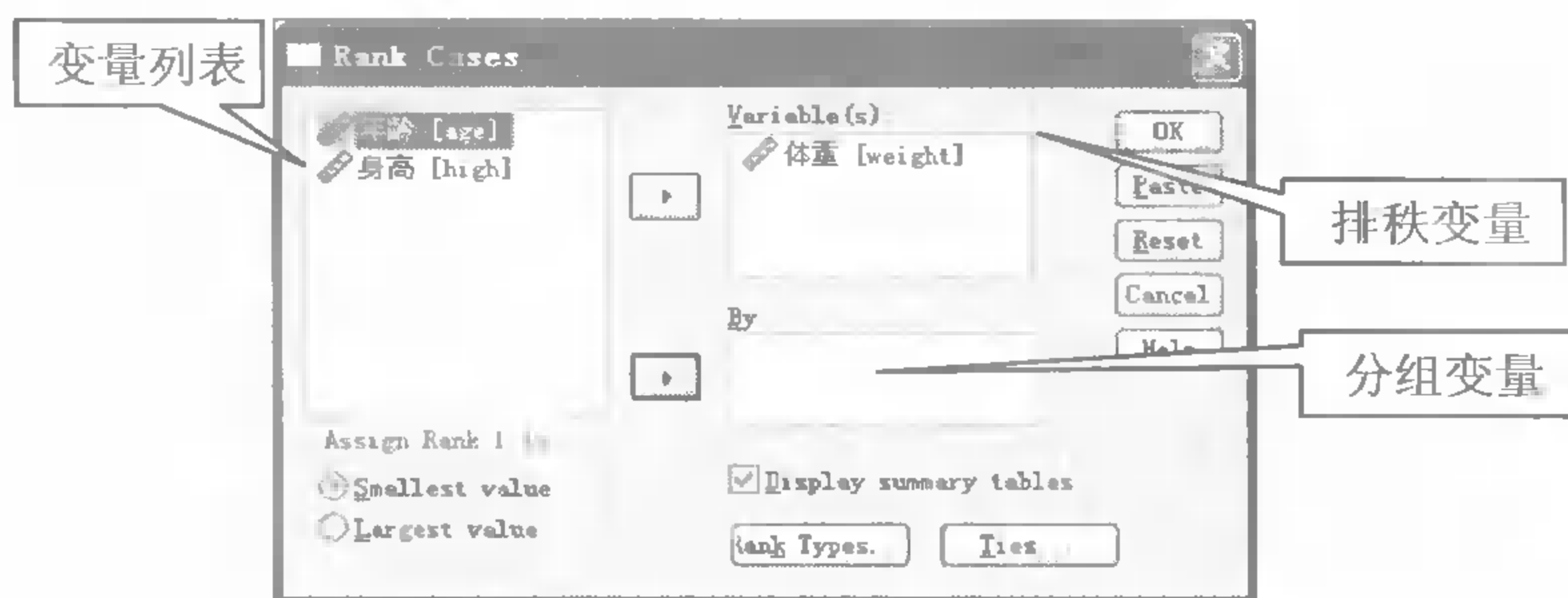


图 3-24 Rank Cases 的主设置界面

- 变量列表显示当前数据集中可用的所有变量; Variable(s)列表用于从左侧的变量列表选入需要求秩的变量; By 列表用于从左侧的变量列表选入分类变量。
- Assign Rank 1 to 栏设定秩的排列顺序,有如下两个选择:
  - ☆ Smallest value 选项,表示最小值用 1 标注,随着取值增大,秩也依次递增。
  - ☆ Largest value 选项,表示最大值用 1 标注,随着取值减小,秩也依次递减。
- 勾选 Display summary tables 复选框,将在 Viewer 窗口输出分析的摘要信息。

### (3) 排秩方法的设置。

在图 3-24 中单击 Rank Types 按钮,弹出如图 3-25 所示的设置对话框,在此设置排秩的方法和参数,单击 Continue 按钮可返回主面板。图 3-25 中可选内容有如下几项。

- Rank 复选框,普通秩次,新变量的取值就是排序后的秩。
- Savage scores 复选框,新变量的值就取以指数分布为基础的原始得分。
- Fractional rank 复选框,用秩除以非缺失观测值的权重和。
- Fractional rank as %复选框,先用秩除以变量有效取值的个数,再乘以 100,得到新变量的取值。
- Sum of case weights 复选框,新变量的取值等于各观测值权重之和,并且新变量在同组中的取值是一个常数。
- Ntile 复选框,以百分位数为基础进行排序,在后面的输入框指定百分位数的个数;默认的百分位个数为 4,此时小于 25%百分位数的变量取值的秩取为 1,25%~50%百分位数之间的变量取值的秩取为 2,50%~75%百分位数之间的变量取值的秩取为 3,大于 75%百分位数的变量取值的秩取为 4。
- Proportion estimate 复选框,计算秩取值分布的累积比例的估计。
- Normal scores 复选框,计算 Proportion estimate 的 Z 得分。

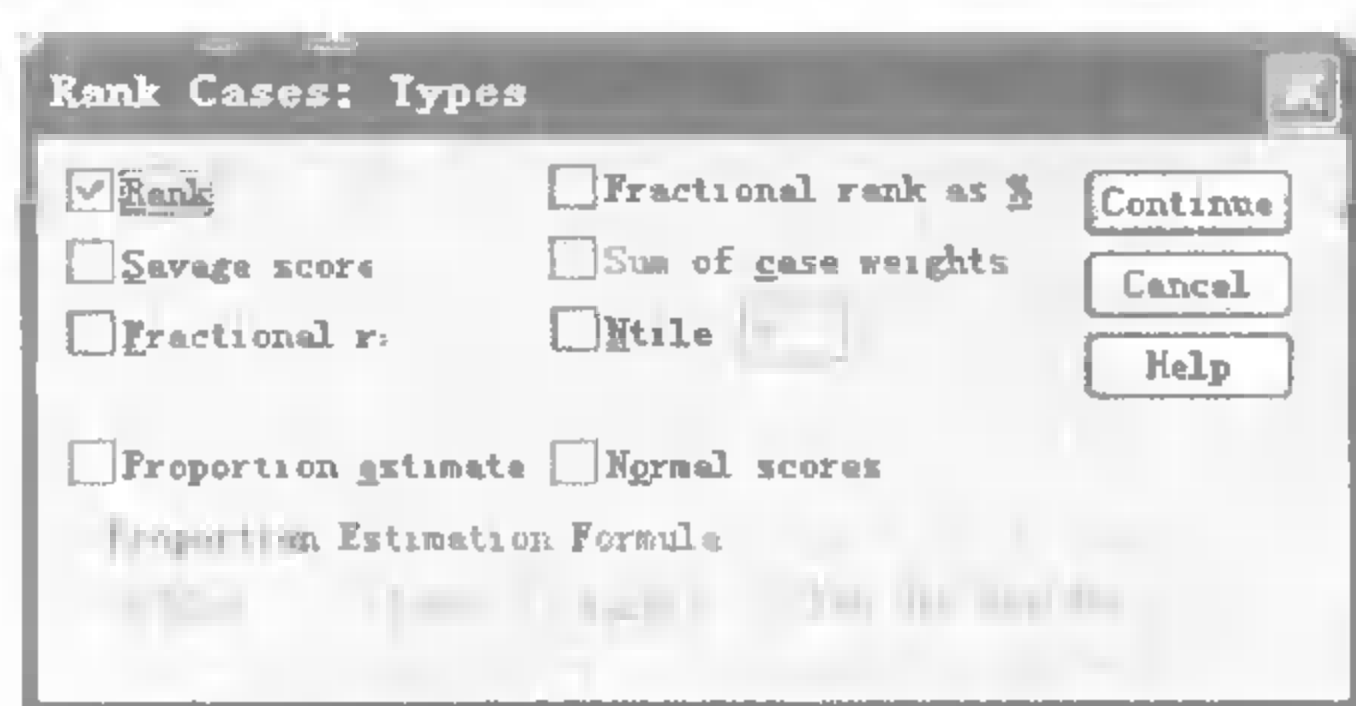


图 3-25 Rank Cases 的排秩方法设置

如果选中 Proportion estimate 或 Normal scores,下面的 Proportion Estimation Formula 栏被激活,用于选择比例估计的公式,可选项有如下 4 个。

- Blom 选项,定义一个新的基于比例估计的变量,公式是  $(r-3/8)/(w+1/4)$ ,其中  $w$  是

观测量权重的总和， $r$  是新求得的秩。

- Tukey 选项，公式是  $(r-1/3)/(w+1/3)$ ，其中  $w$  和  $r$  的含义同上。
- Rankit 选项，公式是  $(r-1/2)/w$ ，其中  $w$  是观测量的个数， $r$  是新求得的秩，且  $r$  的取值范围是  $1 \sim w$ 。
- Van der Waerden 选项，公式是  $r/(w+1)$ ，其中  $w$  是观测量权重的总和， $r$  是新求得的秩，且  $r$  的取值范围是  $1 \sim w$ 。

(4) 对结的处理方法。结就是指由变量取值相同的多个观测形成的组，对结的处理就是要明确如何求这些取值相同的观测的秩。在图 3-24 中，单击 Ties 按钮，弹出如图 3-26 所示的设置对话框，单击 Continue 按钮可返回主面板。

在图 3-26 中，Rank Assigned to Ties 栏给出了如下 4 种处理结的方法：

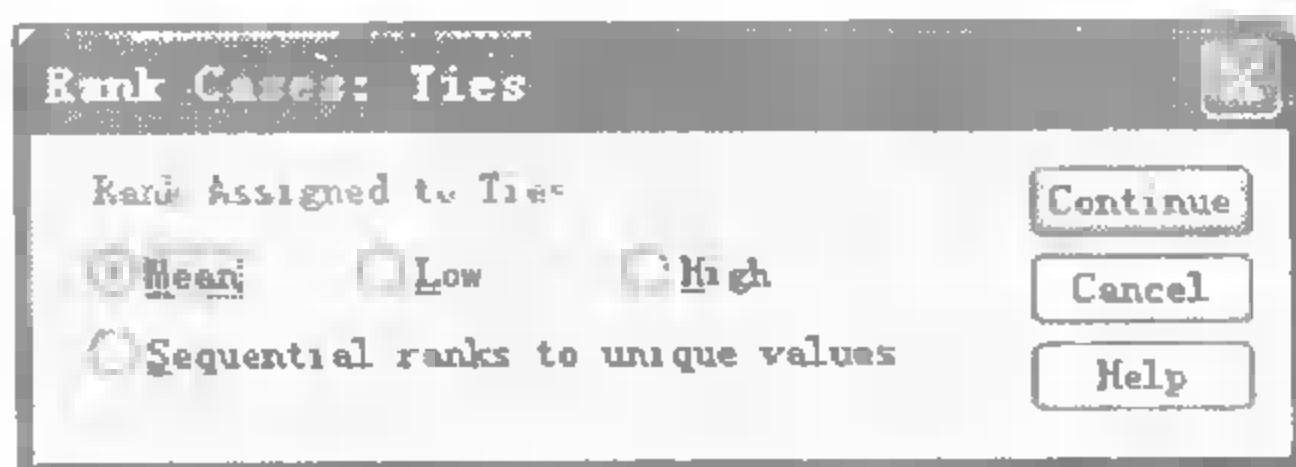


图 3-26 Rank Cases 的结设置

- Mean，取这些相同取值的平均值的秩。
- Low，取这些相同取值的最小值的秩。
- High，取这些相同取值的最大值的秩。
- Sequential ranks to unique values，把这些相同的取值当作一个值进行排序求秩，求得的秩就作为这些取值的秩。

图 3-27 所示是这 4 种处理方法的一个简单示例。

Value	Mean	Low	High	Sequential
10	1	1	1	1
15	2	2	4	2
15	3	2	4	2
15	3	2	4	2
16	5	5	5	3
20	6	6	6	4

图 3-27 对结进行处理的示例

(5) 结果显示。在图 3-24 中单击 OK 按钮运行，Data Editor 窗口的数据格式如图 3-23 所示，其中升序秩列的数据即排序后的秩，可见体重取值越大，秩也越大。

如果在图 3-24 中选中了 Largest value 项，运行后将输出图 3-23 中的降序秩——列的数据，此时体重取值越大秩越小。

### 3.1.6 变量值的重新编码

SPSS 可以对变量的已有取值进行重新编码（或称为赋值）。例如在问卷调查中，有时为了保证问卷的可信度，经常会设定一些实质内容相同但是提问方式不同的问题，通过比较被调查者前后的回答是否一致，来判断问卷的可信程度；但是，当两个问题的答案编码不一致时，是无法进行比较的，这时就需要对某个答案的取值进行重新编码了。

SPSS 里对变量取值重新编码的过程有两个：一个是用新编码直接取代原变量的取值 (Recode into Same Variables)；另一个是将新编码存入新的变量 (Recode into Different Variables)。

#### 1. 数据描述

为了区别正常人和身体不适的人之间的体力差异，记录了几个试验者在一定速率的跑步机上所能坚持的时间，数据文件为“跑步机的测试.sav”，数据格式如图 3-28 (1) 所示。原始数据记录的时间是以秒为单位的，为了便于比较，本节把时间变量重新编码为以分钟为单

位的分段变量。

	group	time
1	1.00	1014
2	1.00	684
3	1.00	810
4	1.00	990
5	1.00	840
6	1.00	978
7	1.00	1002
8	1.00	1110
9	2.00	864
10	2.00	636
11	2.00	638
12	2.00	708
13	2.00	786
14	2.00	600
15	2.00	1320
16	2.00	750
17	2.00	594
18	2.00	750

(1) 原始数据

	group	time
1	1.00	18
2	1.00	12
3	1.00	14
4	1.00	18
5	1.00	14
6	1.00	18
7	1.00	18
8	1.00	20
9	2.00	16
10	2.00	12
11	2.00	12
12	2.00	12
13	2.00	14
14	2.00	10
15	2.00	22
16	2.00	14
17	2.00	10
18	2.00	14

(2) 保持原变量名的编码结果

	group	time	time2
1	1.00	1014	16~18
2	1.00	684	10~12
3	1.00	810	12~14
4	1.00	990	16~18
5	1.00	840	12~14
6	1.00	978	16~18
7	1.00	1002	16~18
8	1.00	1110	18~20
9	2.00	864	14~16
10	2.00	636	10~12
11	2.00	638	10~12
12	2.00	708	10~12
13	2.00	786	12~14
14	2.00	600	10
15	2.00	1320	20~22
16	2.00	750	12~14
17	2.00	594	10
18	2.00	750	12~14


(3) 生成新变量的编码结果

图 3-28 跑步机测试数据编码前后的数据格式

## 2. 保持原变量的编码过程设置

依次单击菜单“Transform→Recode into Same Variables”，打开变量不变式重新编码的主设置面板，如图 3-29 所示，在此选择进行重新编码的变量。

### (1) 变量选择。

在左侧的变量列表单击选中坚持时间变量，然后单击  按钮，将其选入 Numeric 列表作为要编码的变量。

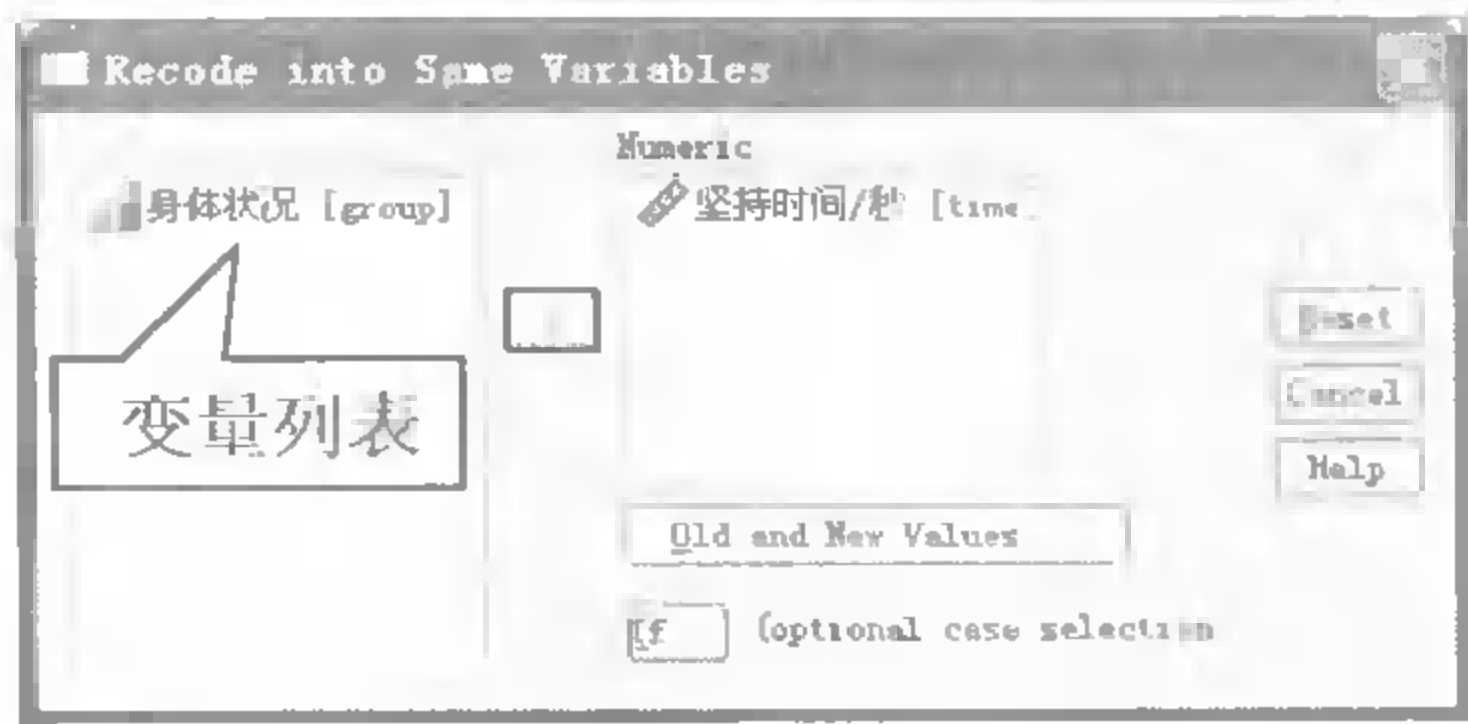


图 3-29 保留变量名的编码设置界面

左侧的列表显示当前数据集中的变量，右侧的 Variables 列表用于选入要进行重新编码的变量，如果同时选入多个变量，则所选变量的类型必须相同（数值型或字符型）。当选入的第一个变量为数值型时，Variables 标签自动变为 Numeric；当选入的第一个变量为字符型时，Variables 标签相应地变为 String Variables。

### (2) 原始观测量的选择。

在图 3-29 中单击 If 按钮，弹出如图 3-30 所示的观测量选择对话框，在此选择需要重新编码的观察量范围，选择方式是编辑一个条件表达式，满足此表达式的观测量将被按照指定的设置进行重新编码。单击 Continue 按钮可返回主界面。

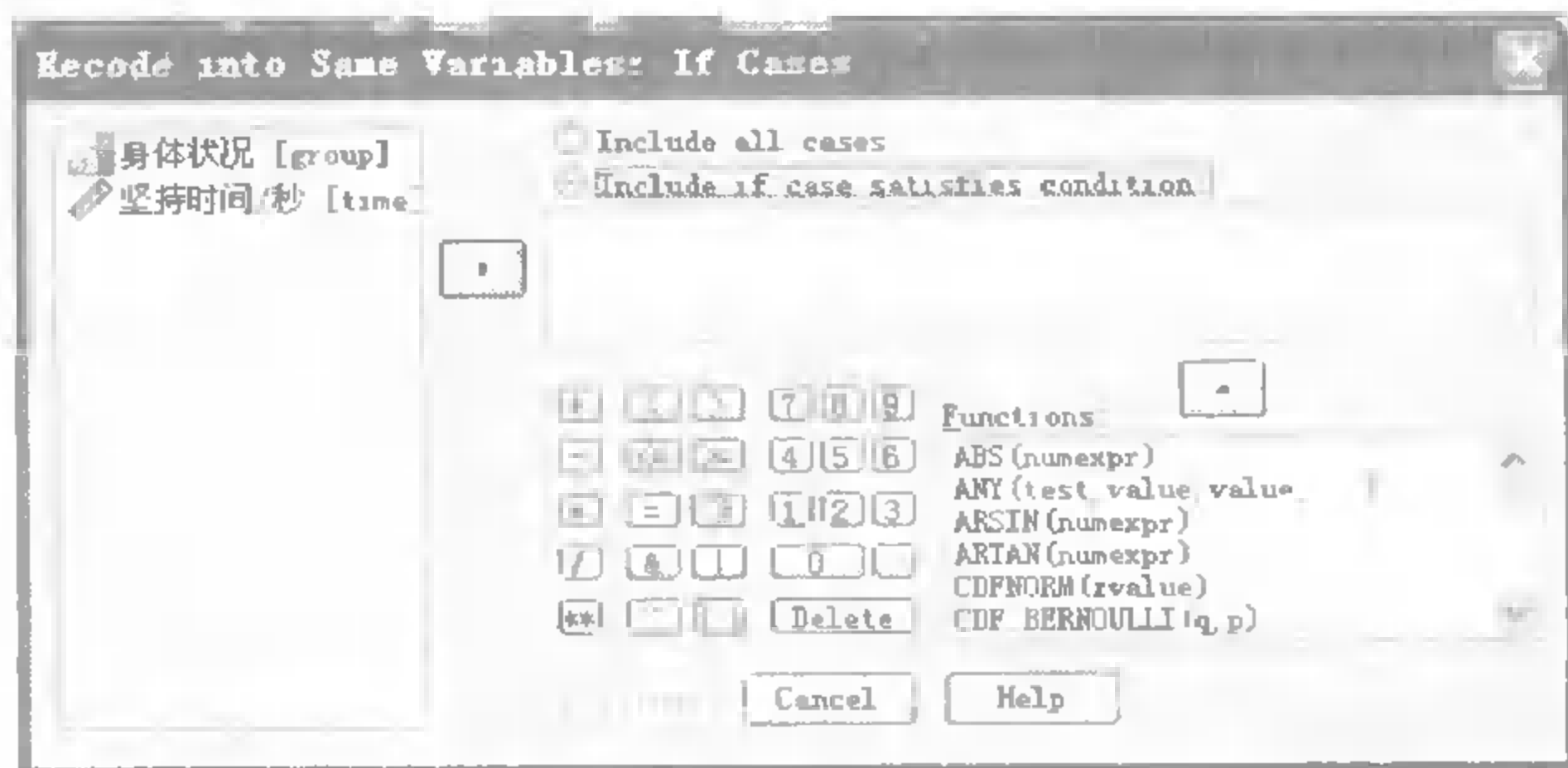


图 3-30 观测量选择界面

- Include all cases 选项，表示对所有原观测量进行重新编码，这是系统默认选项。
- Include if case satisfies condition 选项，只对满足指定条件的观测量进行编码，选中此项激活其他参数选项，在下面的编辑框输入和编辑条件表达式。可以直接在编辑框键入特定的条件表达式，也可以单击面板上的小键盘输入数字和运算符，还可以从左侧的变量列表选入变量名，从右下角的 Functions 列表选入特定的函数表达式。

### (3) 新旧取值的设置。

在图 3-29 中单击 Old and New Values 按钮，打开如图 3-31 所示的子设置界面，在此设置要重新编码的原观测值和新值的对应关系。单击选中 Old Value 栏的 Range, LOWEST through value 选项，并在其下的输入框内键入 600；单击选中 New value 栏的 Value 选项，并在其后的输入框内键入 10；单击 Old--->New 栏下的 Add 按钮，将输入的新旧值对应关系“lowest thro 600--->10”添加到下面的列表中；用相同的方法把其他值的对应关系添加到 Old--->New 列表；单击 Continue 按钮可返回主界面。

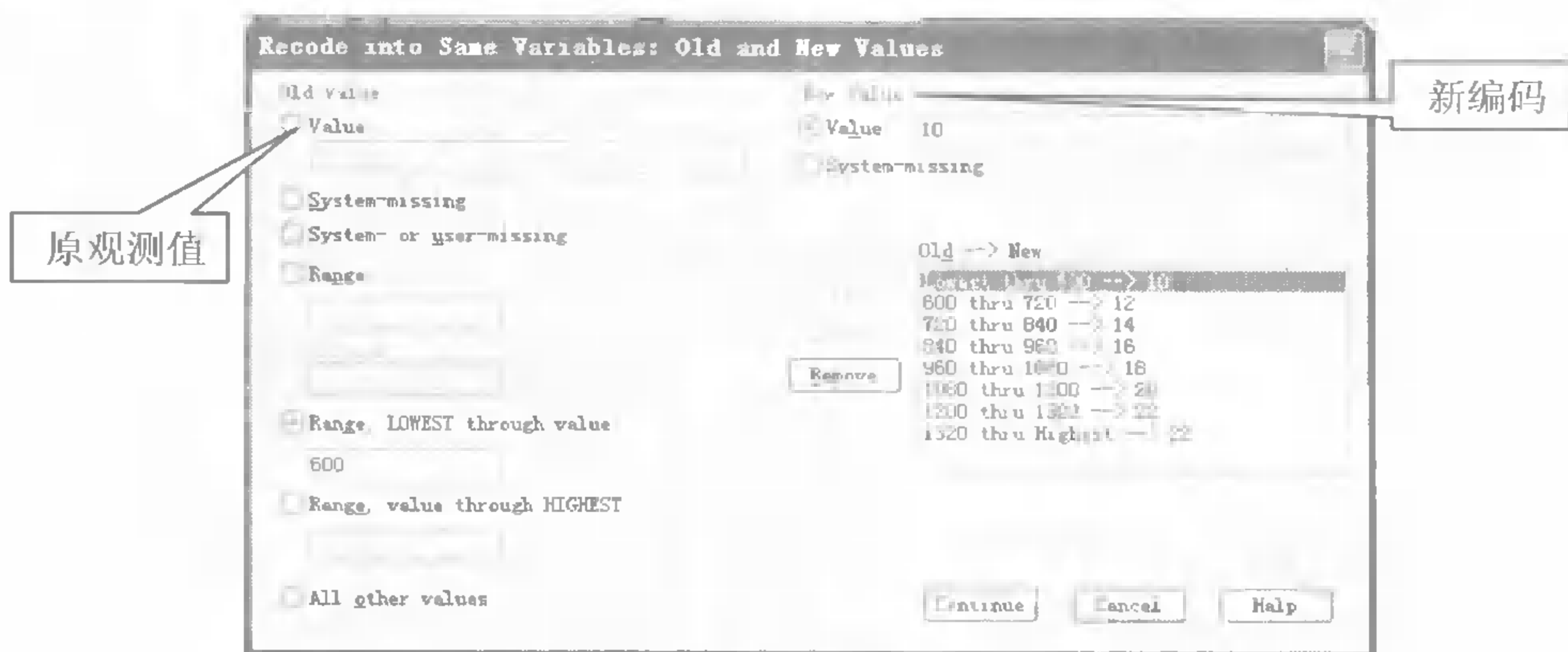


图 3-31 不改变变量名的重新编码设置

#### ① Old Value 栏，用于指定原观测值，可选方式有如下几个。

- Value，输入一个要重新赋值的原观测值。
- System-missing，系统缺失值。
- System-or user-missing，系统或者用户定义的缺失值。
- Range...through...，指定要被重新赋值的原观测值范围，through 前输入区间的下限，through 后输入区间的上限。
- Range, LOWEST through value，指定要被重新赋值的原观测值范围，从最小值到在输入框内指定的值。
- Range, value through HIGHEST，指定要被重新赋值的原观测值范围，从在输入框内指定的值到最大值。
- All other values，所有未定义的观测值。

#### ② New value 栏，用于指定与 Old Value 栏对应的原观测值的新编码，可选方式有如下两个：

- Value，输入一个新的编码值。
- System-missing，系统缺失值。

#### ③ Old--->New 栏的操作

对于选入 Old--->New 列表的新旧值对应关系，单击选中后，可以对其进行编辑和修改，




然后单击 Change 按钮确认修改，或者单击 Remove 按钮将其删除。

注意：如果要重新编码的原变量为数值型变量，则新的编码值也只能输入数值；如果要重新编码的原变量为字符型变量，则新的编码值可以输入任意字符，此时 Old Value 栏只有 Value、System-or user-missing 和 All other values 这 3 个选项可用。

#### (4) 结果显示。

在图 3-29 中单击 OK 按钮运行，Data Editor 窗口显示对体重进行重新编码后的数据格式，如图 3-28 (2) 所示，time 变量已经按照图 3-31 指定的对应关系进行了更新。

### 3. 生成新变量的编码过程设置

依次单击菜单“Transform→Recode into Different Variables”，打开生成新变量的重新编码主设置面板，如图 3-32 所示，选择进行重新编码的变量。在变量列表单击选中“坚持时间/秒”变量，单击  按钮，将其选入右侧的新旧变量对应列表；在 Output Variable 栏的 Name 下框内输入“time2”，Label 下框内输入“坚持时间/分”，单击“Change”按钮确认新变量名的修改。

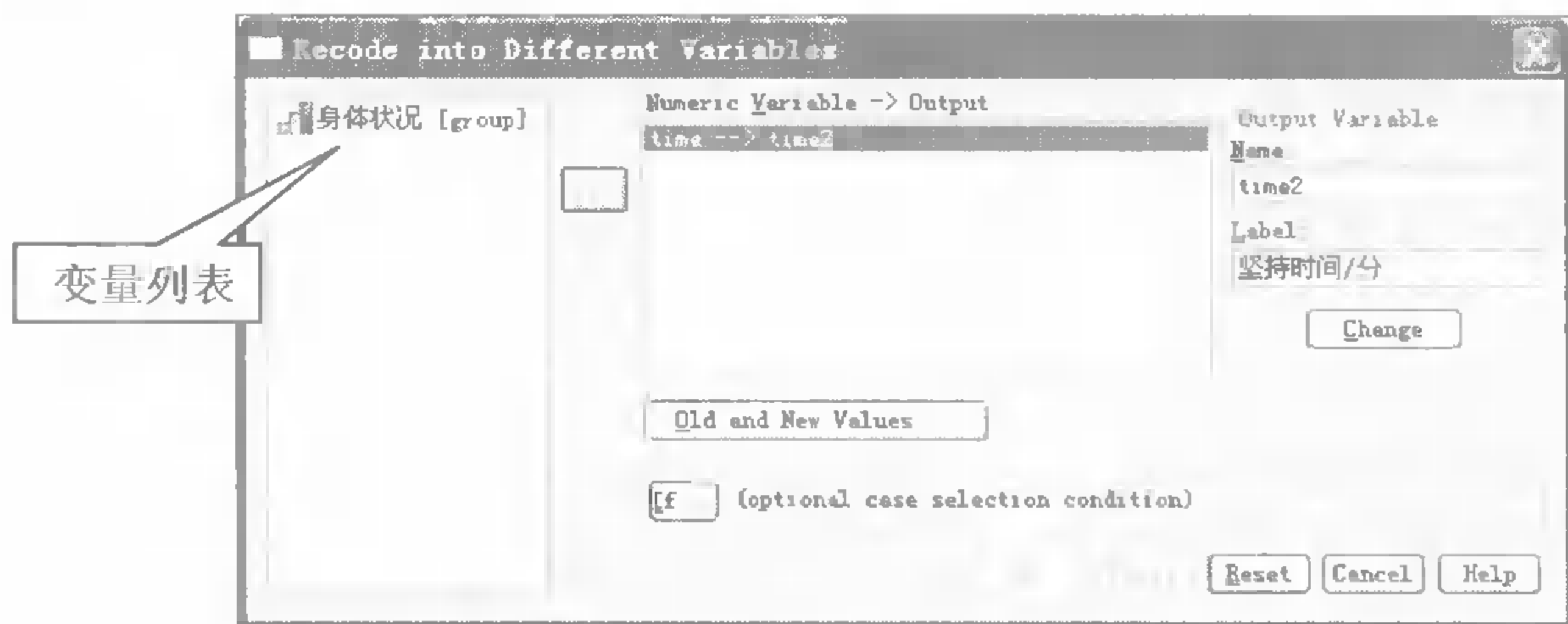


图 3-32 生成新变量的编码设置界面

在图 3-32 中单击 Old and New Values 按钮，打开如图 3-33 所示的子设置界面，设置要重新编码的原观测值和新值的对应关系。单击选中右下角的 Output 复选框，采用与图 3-31 相同的设置方法，把新旧变量的对应取值选入图中的 Old--->New 列表。单击“Continue”按钮可返回主界面。

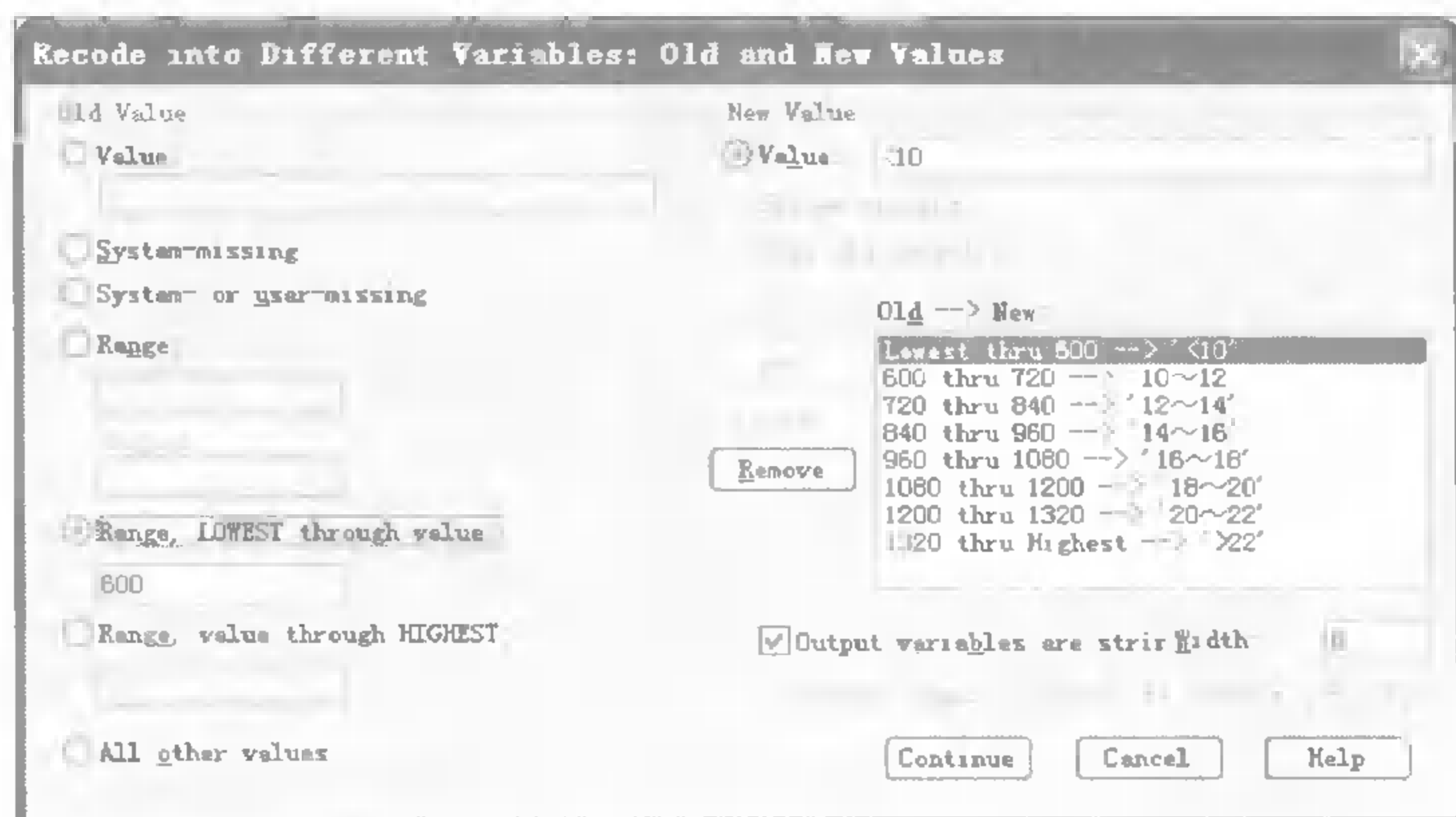


图 3-33 生成新变量的重新编码设置

图 3-32 与图 3-29 相比，右侧多了一个 Output Variable 栏，用于指定新生成变量的变量



名 (Name) 和变量标签 (Label), 在 Name、Label 下框内分别输入变量名、变量标签后, 单击 “Change” 按钮确认修改。

图 3-33 与图 3-31 相比, New value 栏多了如下 3 个可选项。

- Copy old value(s), 直接拷贝对应的原观测值作为新编码。
- Output variables are strings 复选框, 选中后指定输出编码为字符型的, 而不管原变量是数值型的还是字符型的。
- Convert numeric strings to numbers 复选框, 把原观测值中数字形式的字符串重新编码为对应的数值, 对那些非数字形式的字符串用系统缺失值重新编码。

在图 3-32 中单击 OK 按钮运行, Data Editor 窗口显示对体重进行重新编码后的数据格式, 如图 3-28 (3) 所示, time 变量没有改变, 按照图 3-33 指定的对应关系生成了一个新变量 time2。

### 3.1.7 计算新变量

利用已有变量生成新的变量是很常用的功能, 例如分析变量  $x$ 、 $y$  的相关性时, 发现它们并非线性相关, 要研究它们之间的非线性关系, 需要对  $x$  或  $y$  进行各种变换后再进行研究 (比如对数变换、多项式变换等), 这是统计分析中一个解决问题的重要方法。通过生成新变量, 可以完成以下几个方面的任务:

- 创建新变量或更新已存在的变量, 对于新变量, 还可以指定它的类型和标签。
- 在一定的逻辑条件下, 有选择地计算某个数据子集的值。
- 能够使用近 70 种 SPSS 的内在函数进行变量的计算和转换, 包括数值函数、统计函数、分布函数、字符函数等。

前一节介绍的 Recode 过程与 Compute 过程都可以用来产生新变量, 这两种方法的不同之处在于: Recode 方法不能进行运算, 只能对指定的变量取值做数值转换; 而 Compute 过程可以按照任意给定的计算表达式生成新的变量。

SPSS 计算新变量的过程 Compute Variable 功能强大, 是使用较为频繁的一个过程。下面通过一个简单的例子, 来演示 Compute Variable 的操作过程。

#### 1. 数据描述

在第 3.1.1 节中曾使用过小王记录的工资收入和支出情况数据, 数据文件为 “月收入 and 支出数据.sav”, 本节利用它进行分析。数据格式如图 3-34 所示。通过观察发现, 原始数据是按照月份的升序排列的。

	月份	月工资	月奖金	其它收入	支出		
1	1	1000.00	1600.00	.00	1400.00		
2	2	1000.00	1500.00	600.00	1300.00		
3	3	1000.00	1400.00	900.00	700.00		
4	4	1300.00	1800.00	.00	900.00		
5	5	1300.00	1200.00	700.00	1000.00		
6	6	1300.00	1300.00	800.00	1500.00		

图 3-34 小王的工资收入和支出情况

#### 2. 参数设置

依次单击菜单 “Transform→Compute Variable” 执行计算新变量的功能, 其主设置界面如

图 3-35 所示，在这里主要设置的是关于新变量的计算表达式。

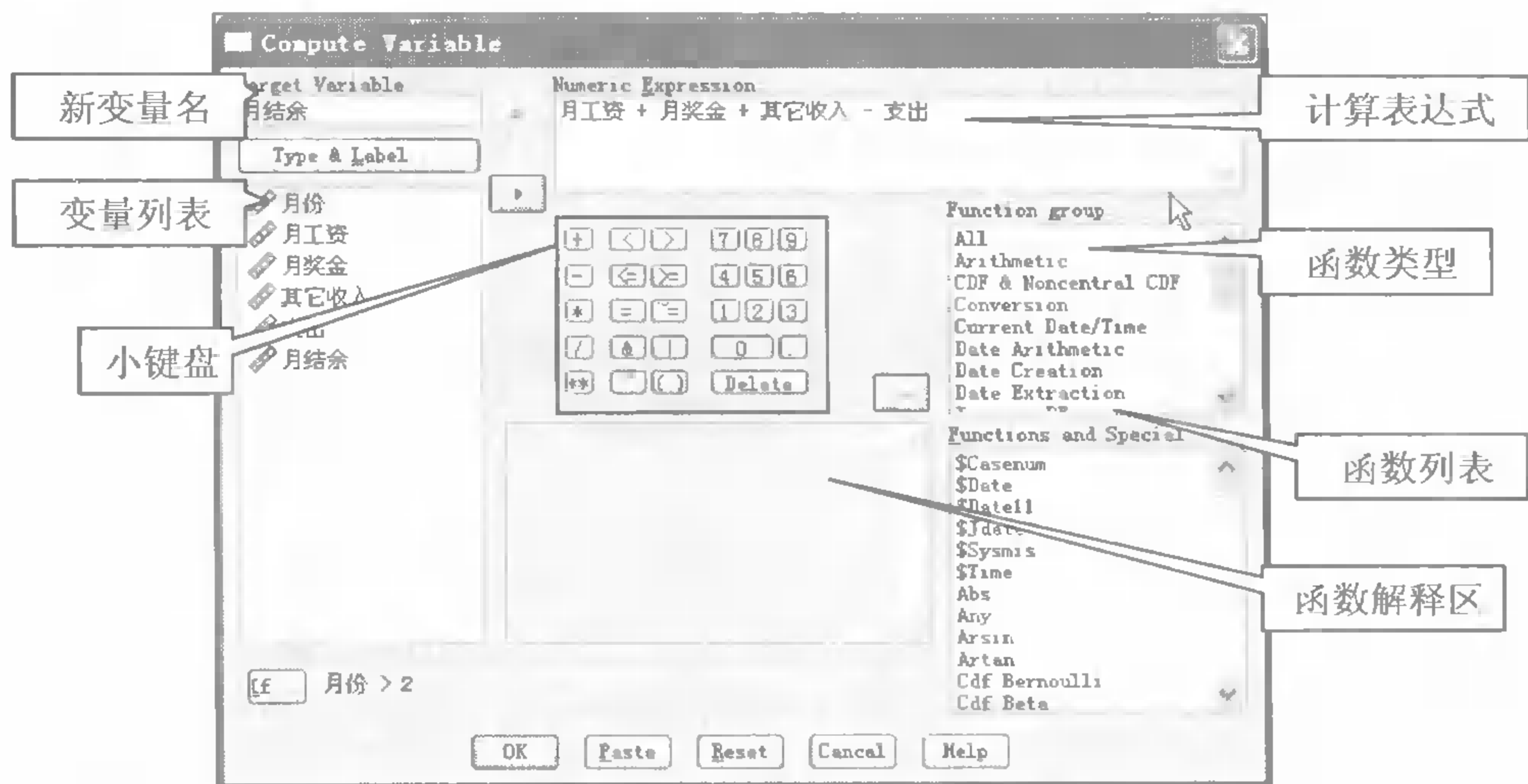


图 3-35 计算新变量的设置界面

#### (1) 新变量的表达式设置。

首先在 Target Variable 栏下的输入框键入新变量的名称“月结余”；然后在 Numeric Expression 编辑框输入“月工资+月奖金+其他收入-支出”。

##### ① Target Variable 栏，用于指定新变量的名称。

可以输入已经存在的变量名，这样运行后将会更新原变量。单击下面的 Type & Label 按钮，弹出如图 3-36 所示的子设置界面，在此设置新变量的标签和变量类型。

- Label 栏设置新变量的标签，有如下两个选择：Label 选项，选中后输入不超过 120 个字符的标签；Use expression as label 选项，用计算表达式的前 110 个字符作为变量标签。

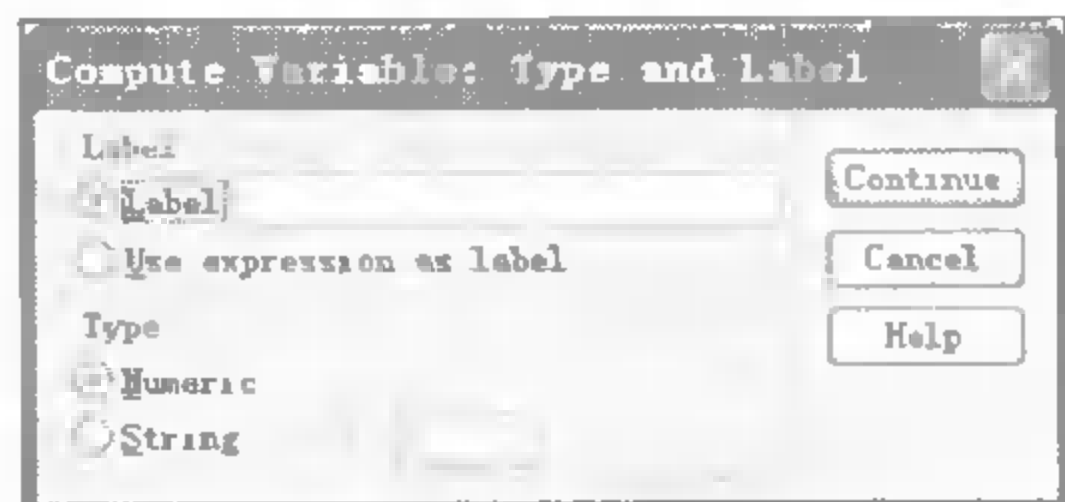


图 3-36 新变量的标签和类型设置

- Type 栏设置新变量的类型，有如下两个选择：Numeric 选项，数值型；String 选项，字符型，选中它还需在 Width 后的输入框中指定字符串的长度，默认值为 8。

##### ② Numeric Expression 栏，用于编辑新变量的计算表达式。

当新变量是字符型时，此处显示的标签是 String Expression。除了可以直接在编辑框键入新变量的计算表达式外，还可以通过鼠标单击此界面上提供的信息，来完成表达式的编辑，可选信息包括如下 3 部分。

- 选中变量列表中的某个变量，然后单击向右的黑色箭头，可把选中的变量名直接填入编辑框中光标所在的位置（双击变量列表中的变量名可达到同样效果）。
- 单击小键盘上的某个按钮，可把相应的数字或运算符填入编辑框中光标所在的位置。
- 函数类型列表给出了不同种类的函数组，选中某一项后，在下面的函数列表会列出所选函数组里的所有可用函数；在函数列表选中某个函数后，会在左侧的函数解释区给出所选函数的解释和用法，单击向上的黑色箭头，可把选中的函

数表达式直接填入编辑框中光标所在的位置（双击函数列表中的函数名可达到同样效果）。

## （2）条件表达式设置。

在图 3-35 的底部单击 If 按钮，弹出如图 3-37 所示的条件设置子界面。单击选中 Include if case satisfies condition 选项，在下面的编辑框输入“月份>2”，单击 Continue 按钮可返回主界面，此时 If 按钮后会显示出当前所设定的条件表达式。

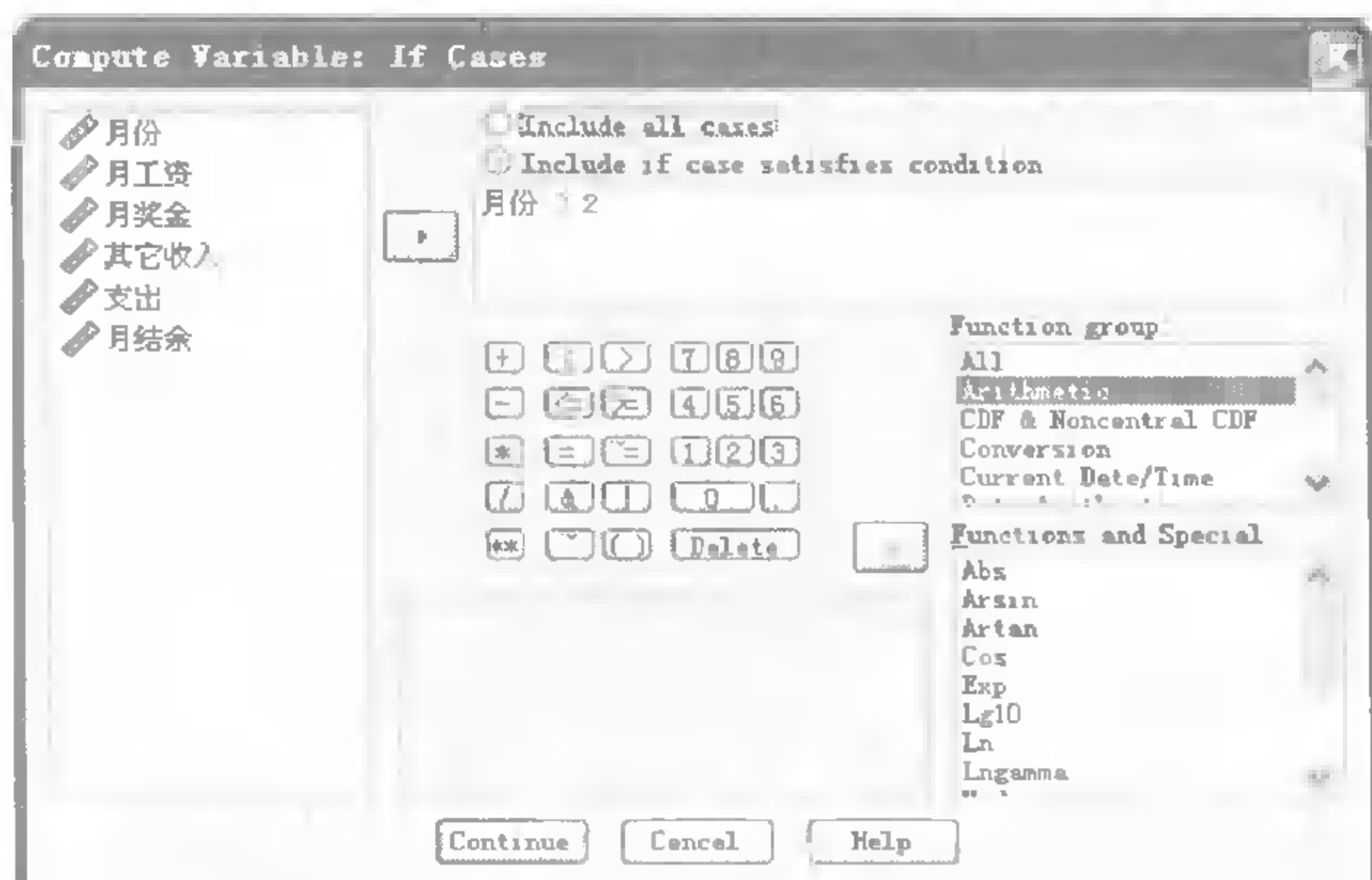


图 3-37 计算新变量的条件表达式设置

下面详细介绍各设置选项的含义。

- Include all cases 选项，表示对所有的观测量都进行新变量的计算，是默认选项。
- Include if case satisfies condition 选项，只对满足指定条件的观测量才计算新变量，选中此项激活其他参数选项。条件表达式在下面的编辑框进行输入和编辑，操作方式与图 3-35 中的计算表达式编辑方法相同。对于那些不满足此处指定条件表达式的观测量，对应的新变量都取系统缺失值。

## 3. 结果显示

在图 3-35 中单击 OK 按钮运行，在当前数据集生成的新变量如图 3-38 所示。

	月份	月工资	月奖金	其它收入	支出	月结余
1	1	1000.00	1600.00	.00	1400.00	.
2	2	1000.00	1500.00	600.00	1300.00	.
3	3	1000.00	1400.00	900.00	700.00	2600.00
4	4	1300.00	1800.00	.00	900.00	2200.00
5	5	1300.00	1200.00	700.00	1000.00	2200.00
6	6	1300.00	1300.00	800.00	1500.00	1900.00

图 3-38 计算新变量的结果显示

其中，新变量表达式为：月结余=月工资+月奖金+其他收入-支出，只对“月份>2”的观测量才按此表达式进行计算，其他情况下新变量都取系统缺失值。

## 3.2 分类汇总

分类汇总就是按指定的分类变量对观测量进行分组，然后计算各分组内的某些变量的描述统计量。汇总结果可以生成新的数据文件，在新文件中对指定分类变量的每个值产生一个观测记录，如果分组变量只有两个值，那么新的汇总文件中将只包含两个观测记录。本节就

用一个简单的实例来介绍如何对数据进行分类汇总。

### 3.2.1 数据描述

本节使用某小学 10~13 岁儿童的身高和体重数据，数据文件为“儿童的身高和体重数据.sav”。本数据曾在第 3.1.2 节使用，数据格式如图 3-4 所示。

下面以性别和年龄为指定的分类变量，对儿童的身高和体重进行汇总。

### 3.2.2 分类汇总的参数设置

依次单击菜单“Data→Aggregate”执行分类汇总功能，打开的主设置界面如图 3-39 所示，在此设置分类汇总的变量、保存等选项。

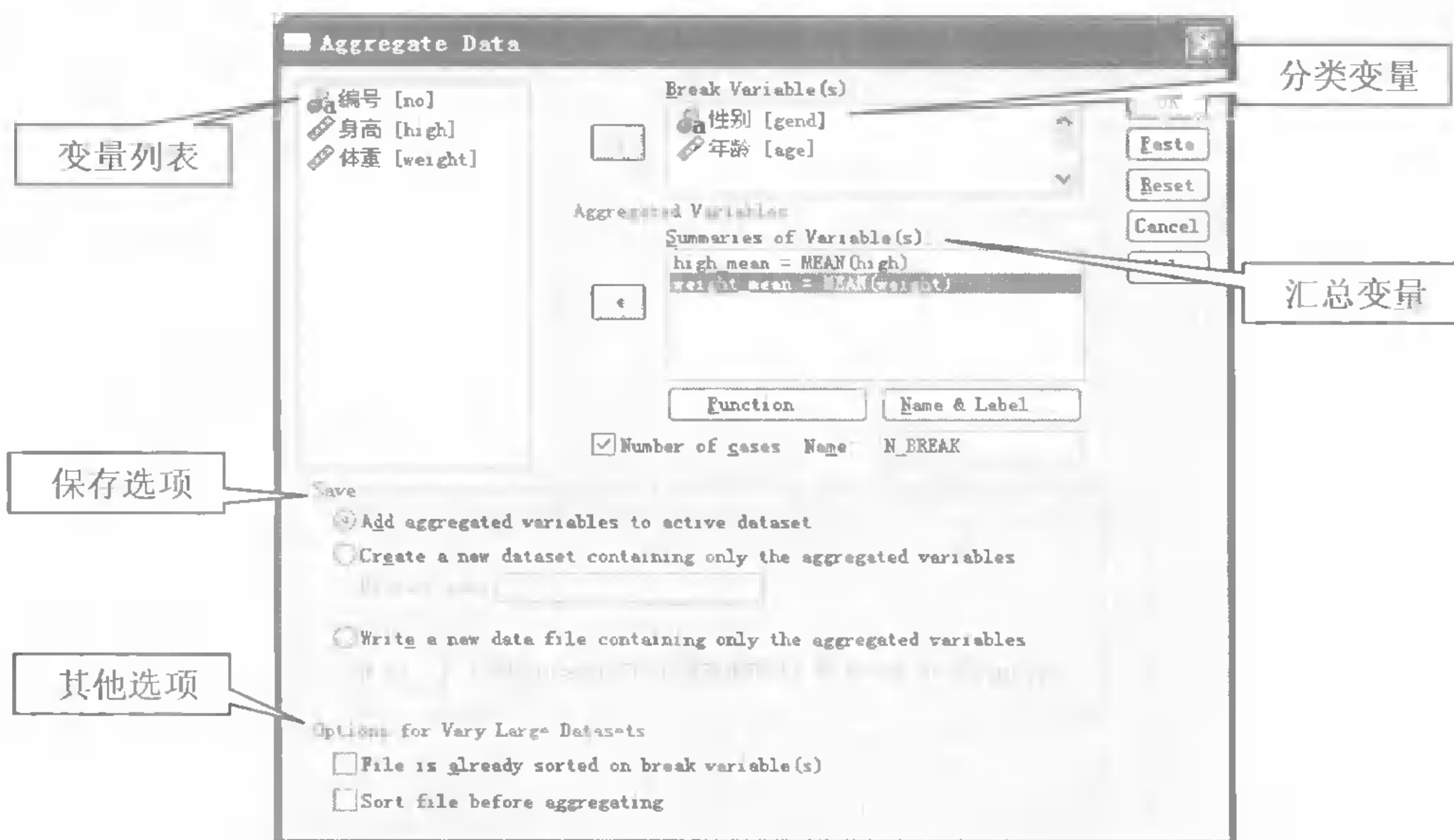




图 3-39 分类汇总的主设置界面

#### 1. 主界面设置

在变量列表选中性别和年龄变量，单击 Break Variable 列表左侧的  按钮，将其选入分类变量列表；在变量列表选中身高和体重变量，单击 Summaries of Variable 列表左侧的  按钮，将其选入汇总变量列表；单击选中 Number of cases 复选框。

下面详细介绍各设置选项的含义。

① Break Variable(s)列表，用于从左侧的变量列表选入汇总的分类变量。

② Summaries of Variable(s)列表。

用于从左侧的变量列表选入汇总变量（要在各分组内进行描述的变量）。选入的汇总变量显示格式为： $y = f(x)$ ，其中  $x$  为选入的变量名， $f$  为进行汇总计算的函数， $y$  为汇总后的变量名。

③ Number of cases 复选框。

选中它后，将在分类结果中用一个变量显示每个类别里的观测量个数。在 Name 后输入此变量的名称，默认为 N\_BREAK。



## ④ Save 子设置栏

设置关于汇总结果的保存选项，有如下 3 个可选方式：

- ① Add aggregated variables to active dataset, 将汇总结果添加到当前数据集中。
- ② Create a new dataset containing only the aggregated variables, 创建一个新的、只包含汇总后变量的数据集，选中后在 Dataset name 后输入新数据集的名称。
- ③ Write a new data file containing only the aggregated variables, 将分类汇总后的变量保存到一个新的数据文件中，选中后单击 File 按钮指定文件路径和名称。

## ⑤ Options for Very Large Datasets 子设置栏

此处设置关于处理较大数据集时的选项，有两个可选内容：

- ① File is already sorted on break variable(s)复选框, 选中表示数据已经按照指定的分类变量排好序了，当数据较大时能节省不少的运行时间。
- ② Sort file before aggregating 复选框, 选中表示要求在分类汇总之前，先按照指定的分类变量对数据进行排序。

## 2. 设置汇总函数

在图 3-39 的 Summaries of Variable(s)列表中选中某个汇总变量后，单击 Function 按钮，弹出如图 3-40 所示的子设置界面，设置选中变量的汇总函数。单击 Continue 按钮可返回主界面。

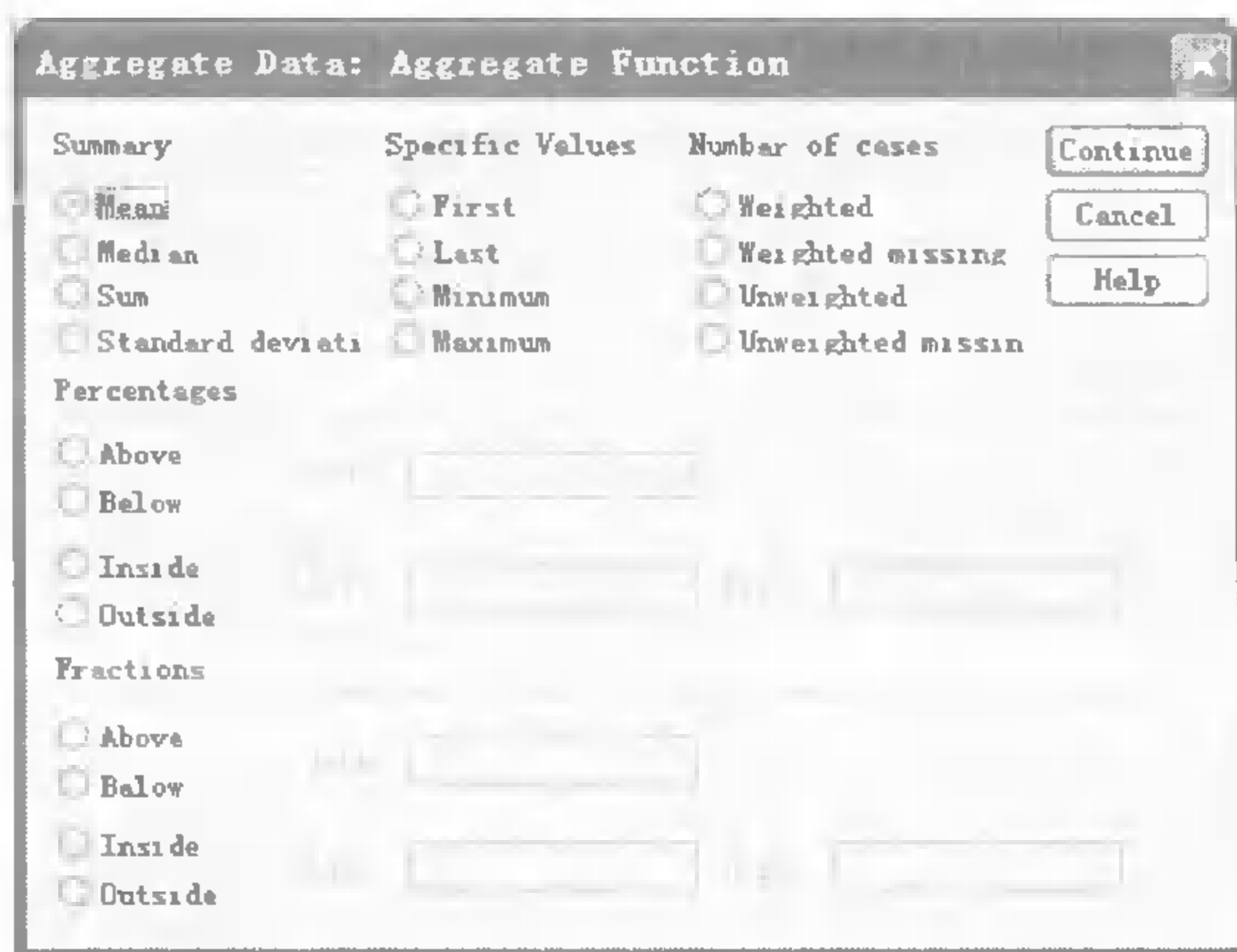


图 3-40 汇总函数的选择对话框

下面详细介绍各设置选项的含义。

① Summary 栏。选择概要型函数，包括如下 4 个：Mean（均值）、Median（中位数）、Sum（求和）和 Standard deviation（标准差）。

② Specific Values 栏。选择特定值函数，包括如下 4 个：First（分组内的第 1 个数值）、Last（分组内的最后 1 个数值）、Minimum（分组内的最小值）和 Maximum（分组内的最大值）。

③ Number of cases 栏。选择与观测量个数有关的汇总项，包括如下 4 个：Weighted（带权重的观测量数目）、Weighted missing（带权重的缺失值数目）、Unweighted（不带权重的观测量数目）和 Unweighted missing（不带权重的缺失值数目）。

④ Percentages 栏。选择百分比形式的函数，包括如下 4 个：Above 选项，变量取值大于指定临界值的观测数占总观测数的百分比，在 Value 后输入此临界值；Below 选项，变量取



值小于指定临界值的观测数占总观测数的百分比，在 Value 后输入此临界值；Inside 选项，变量取值落在指定区间之内的观测数占总观测数的百分比，在 Low 后面输入此区间的下限，在 High 后面输入此区间的上限；Outside 选项，变量取值落在指定区间之外的观测数占总观测数的百分比，在 Low、High 后面分别输入此区间的下限和上限。

⑤ Fractions 栏。选择分数形式的函数，它所包含的 4 个选项的含义与设置方式同 Percentages 栏相似。

### 3. 设置新变量的名称和标签

在图 3-39 的 Summaries of Variables 列表中选中某个汇总变量后，单击 Name & Label 按钮，弹出如图 3-41 所示的子设置界面，设置汇总结果变量的名称和标签。在 Name 后框内输入新变量的名称，在 Label 后框内输入新变量的标签。单击 Continue 按钮可返回主界面。

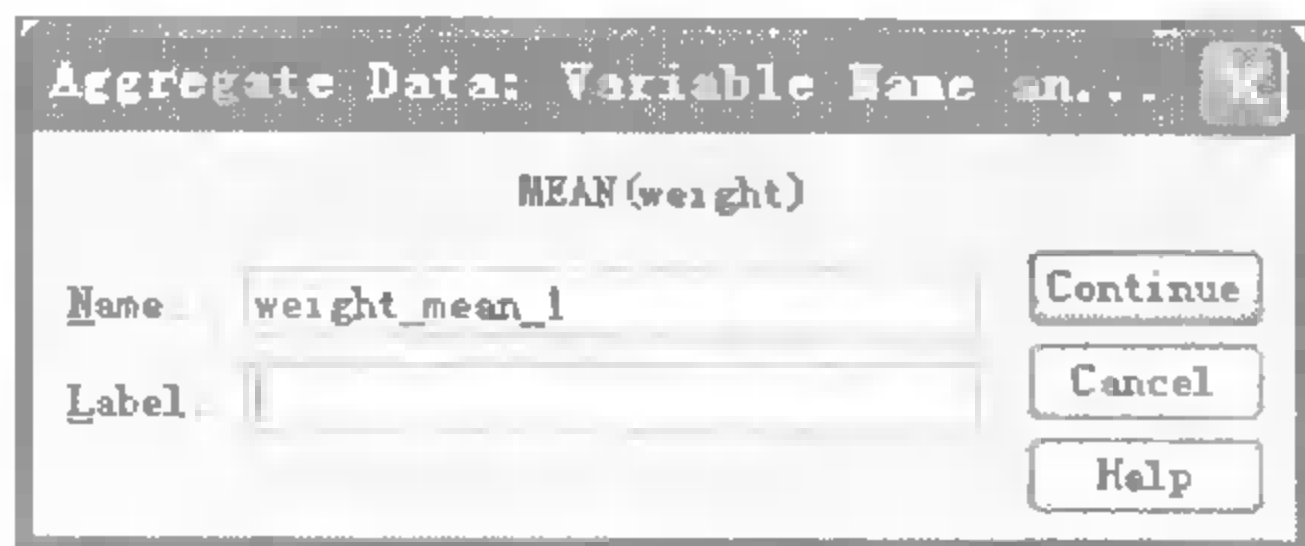


图 3-41 结果变量的名称和标签设置

### 3.2.3 分类汇总的结果

在图 3-39 中单击 OK 按钮运行，输出结果保存在当前数据集中，如图 3-42 所示。

	no	gend	age	high	weight	high_mean	weight_mean_1	N_BREAK
1	06	0	10	1.46	38	1.45	37.40	5
2	18	0	11	1.56	48	1.54	45.80	5
3	17	0	11	1.50	40	1.54	45.80	5
4	07	1	10	1.48	39	1.45	37.40	5
5	12	1	10	1.43	43	1.45	41.67	5
6	26	1	12	1.64	60	1.61	51.80	5
7	15	0	10	1.48	39	1.45	37.40	5
8	45	0	10	1.43	35	1.45	37.40	5
9	21	1	11	1.55	46	1.51	43.17	6
10	27	1	11	1.56	44	1.51	43.17	6
11	09	1	11	1.46	40	1.51	43.17	6
12	27	1	13	1.59	55	1.59	55.00	1
13	04	0	11	1.52	42	1.54	45.80	5
14	05	1	10	1.43	43	1.45	41.67	5
15	10	0	12	1.60	53	1.61	54.50	2
16	14	1	12	1.59	42	1.61	51.80	5
17	08	1	11	1.48	40	1.51	43.17	6
18	03	0	11	1.55	44	1.54	45.80	5
19	20	0	10	1.44	37	1.45	37.40	5
20	19	0	12	1.62	56	1.61	54.50	2
21	01	1	12	1.60	55	1.61	51.80	5
22	02	1	12	1.62	53	1.61	51.80	5
23	11	0	11	1.55	55	1.54	45.80	5
24	16	0	10	1.44	38	1.45	37.40	5
25	13	1	11	1.46	41	1.51	43.17	6
26	40	1	12	1.62	49	1.61	51.80	5
27	25	1	11	1.55	48	1.51	43.17	6

图 3-42 保存在当前数据集的汇总结果

如果在图 3-39 中单击选中 Save 栏的 Create a new dataset 选项，运行后的输出结果保存在一个新的数据集里，如图 3-43 所示，此时对性别和年龄的每个组合单独生成一条汇总后的观测量，记录了身高、体重的均值汇总结果以及它所包含的原始观测量个数。

	gend	age	high_mean	weight_mean_1	N_BREAK
1	0	10	1.45	37.40	5
2	0	11	1.54	45.80	5
3	0	12	1.61	54.50	2
4	1	10	1.45	41.67	5
5	1	11	1.51	43.17	6
6	1	12	1.61	51.80	5
7	1	13	1.59	55.00	1

图 3-43 保存于新数据集的汇总结果

### 3.3 观测量的加权

通过 Data 菜单的 Weight Cases 功能,可对观测量进行加权。对观测量做过加权操作之后,除非用户取消加权或指定其他加权变量,当前的权重将一直保持不变,而且保存数据文件时权重信息也会随之保存。加权操作对于列联表(交叉表)分析显得尤为重要,对于某些计数形式的数据,只有在指定了权重变量后才能进行列联表的分析,对观测量的加权在绘制散点图、直方图以及进行回归分析等过程中都有非常重要的作用。

对于权重变量的使用,需要注意如下 3 点。

- ① 权重变量的取值,应该代表某些类别的观测量的计数信息。
- ② 权重变量取零值、负数、缺失值时,对应的观测量将在其他分析过程中被剔除。
- ③ 权重变量可以取小数,只要它们符合实际意义即可。

#### 1. 数据描述


为了研究抽烟与肺癌的关系,随机采访了 45 个正常人和 55 个肺癌患者,询问并记录了他们是否抽烟,数据文件为“抽烟与肺癌的关系.sav”,数据格式如图 3-44 所示。

	抽烟与否	是否肺癌	人数
1	不抽烟	正常	25
2	不抽烟	肺癌	15
3	抽烟	正常	20
4	抽烟	肺癌	40

此处的人数被作为权重变量使用,本节来介绍如何对其进行加权处理。

图 3-44 抽烟与肺癌的关系数据

#### 2. 对观测量的加权操作

依次单击菜单“Data→Weight Cases”执行加权过程,其设置界面如图 3-45 所示。单击选中 Weight cases by 单选框,在变量列表中选中人数变量,单击  按钮将其选入 Frequency 选框作为加权变量。

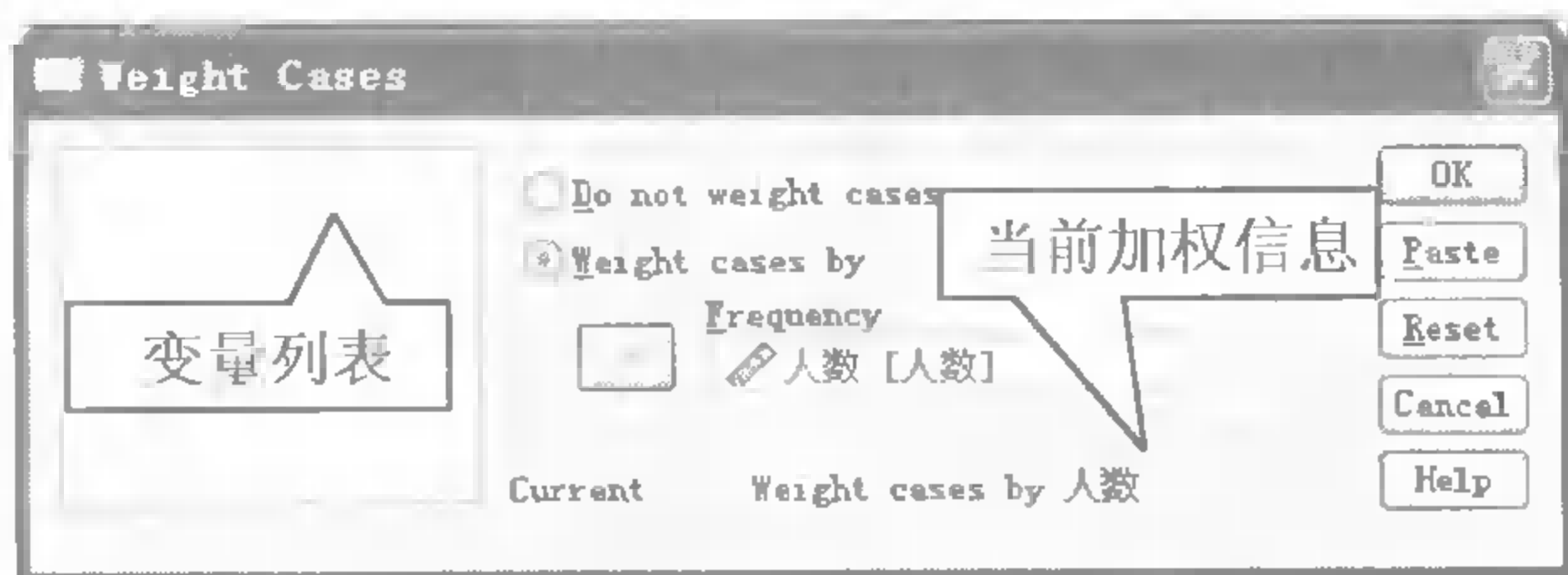



图 3-45 观测量加权的设置对话框



由于权重变量必须为数值型的,所以变量列表只显示当前可用的数值型权重变量。

Do not weight cases 选项,选中表示当前数据集不做加权,可用于对做过加权的数据集取消加权;Weight cases by 选项,按指定变量对数据集进行加权,在 Frequency 栏指定代表频数信息的加权变量;Current 栏显示数据集的当前加权信息,本例中按照人数加权。

在图 3-45 中单击 OK 按钮运行,返回数据窗口 Data Editor,此时底部的状态栏右侧会显示  字样,提示用户当前数据集已经被加权了。

### 3. 进一步分析

加权后的数据集表面上看没有什么变化，但在其他分析过程中却会产生质的不同，下面就以列联表分析为例，解释加权变量的应用。

打开数据文件“抽烟与肺癌的关系.sav”，依次单击菜单“Analyze→Descriptive Statistics→Crosstabs...”执行列联表分析过程，其设置界面如图 3-46 所示。在变量列表中选中抽烟与否变量，单击从上至下第一个  按钮，将其选入 Row 列表作为行变量；在变量列表选中是否肺癌变量，单击从上至下第二个  按钮，将其选入 Column 列表作为列变量。

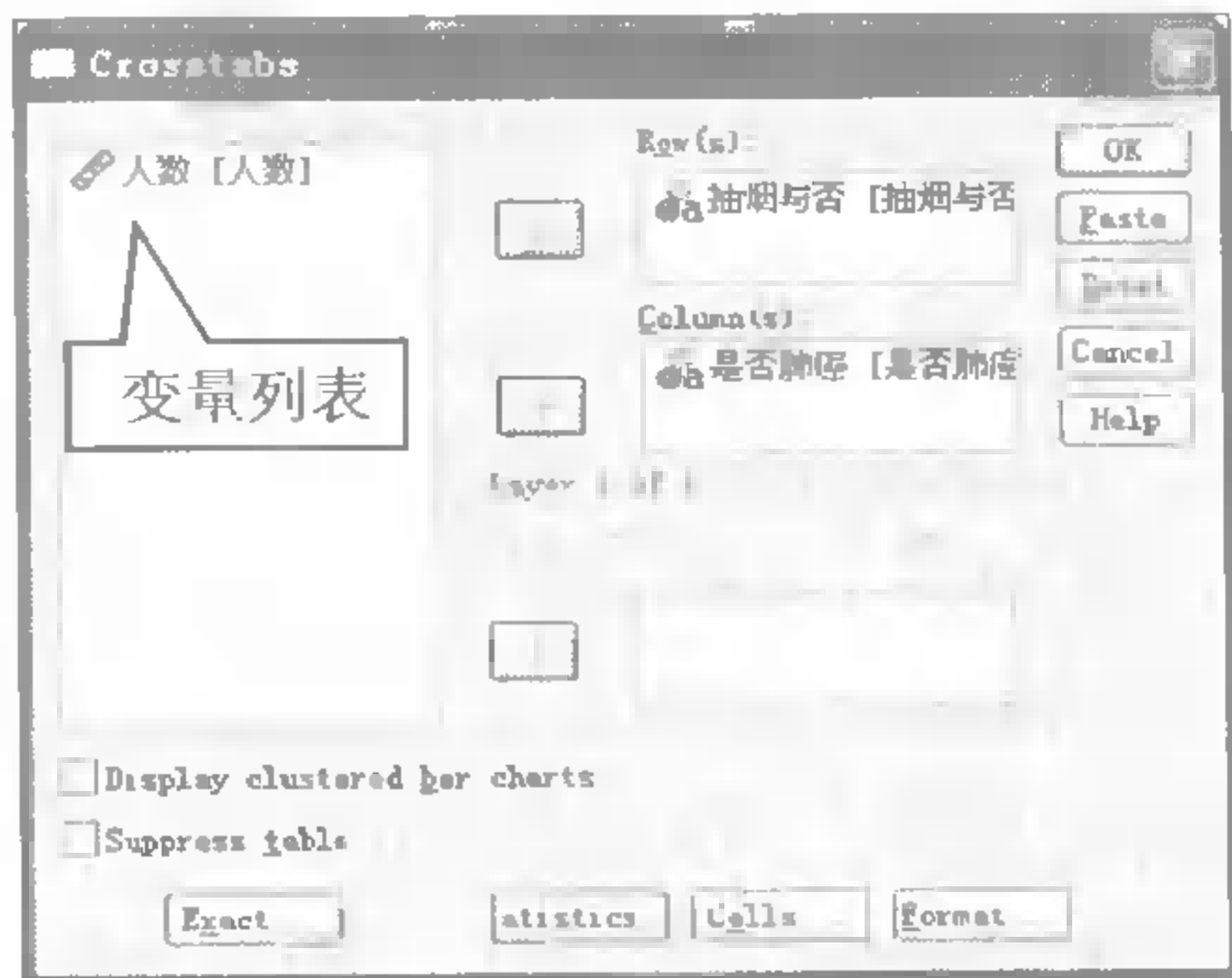


图 3-46 列联表分析的参数设置

在没有对观测量加权之前，按照图 3-46 所示的设置单击 OK 按钮运行，SPSS Viewer 窗口的输出交叉表如图 3-47 所示，没有加权的观测量在统计时，按照实际的记录行数进行计数。

在对观测量按照人数变量进行加权之后，仍按照图 3-46 所示的设置，单击 OK 按钮运行，输出的交叉表如图 3-48 所示，加权之后的观测量在统计时，按照权重变量的取值进行计数。

抽烟与否 * 是否肺癌 交叉制表				
计数		是否肺癌		合计
		肺癌	正常	
抽烟与否	不抽烟	1	1	2
	抽烟	1	1	2
合计		2	2	4

图 3-47 没有加权的交叉表

抽烟与否 * 是否肺癌 交叉制表				
计数		是否肺癌		合计
		肺癌	正常	
抽烟与否	不抽烟	15	25	40
	抽烟	40	20	60
合计		55	45	100

图 3-48 加权之后的交叉表

## 3.4 数据文件的结构重组

在运用 SPSS 进行数据分析时，比较常用的是简单格式的数据文件，即每个观测记录占一行，记录行之间由唯一的标识变量进行区分（例如姓名、ID 号），每个属性或变量占一列。本节就来介绍包括简单格式在内的两种数据文件结构，以及它们之间的转换操作。

下面先以一个例子来介绍两种结构的数据格式。某个销售公司的母公司汇总了 5 个子公司在不同地方的销售业绩，并且业绩是按季度统计的，涉及的数据内容包括子公司的编号（no）、季度（quarter）、销售地点（area）、子公司到销售地点的距离（d）和销售额（sale），对于这部分数据，统计人员有如下两种结构进行组织和保存。

（1）横向结构。如图 3-49 所示，只为每个子公司建立一条观测量记录，把它在 1~4 季度的销售额、销售地点、到销售地点的距离分别作为一个变量加以保存，这种格式称为横向结构，也就是前面提到的简单格式的数据，所用数据文件为“季度销售额的横向格式.sav”。鉴于此格式存储的变量较多，横向结构也被称为变量组结构的数据格式。

	no	quarter1	quarter2	quarter3	quarter4	d	area
1	1	80	90	87	76	1.7	HeBei
2	2	45	67	46	87	1.7	BeiJin
3	3	78	76	88	89	1.9	ShangHai
4	4	78	87	79	76	1.8	TianJin
5	5	98	99	95	94	1.7	HuNan

图 3-49 横向结构的数据格式

(2) 纵向结构。如图 3-50 所示, 为每个子公司的每个季度建立一条观测记录, 也就是将公司编号和季度两个变量作为复合主键, 将销售地点、到销售地点的距离、销售额作为 3 个属性变量, 这种格式称为纵向结构, 所用数据文件为“季度销售额的纵向格式.sav”。鉴于存储的观测记录较多, 纵向结构也被称为观测组结构的数据格式。

通过比较发现, 纵向结构比横向结构冗余信息要多。

不同的分析方法要求使用不同的数据格式, 例如:

General Linear Model 中的 Univariate (单因素方差分析)、Multivariate (多因素方差分析)、Variance Components (方差成分分析), OLAP Cubes 过程、独立样本 T 检验, 以及 Nonparametric Tests (非参数检验) 等分析过程, 一般都要要求数据格式是观测组结构的, 便于进行对分组变量的分析; 而 General Linear Model 的 Repeated Measures (重复测量的方差分析), Cox Regression Analysis 分析中以生存时间为因变量的协方差分析、配对样本 T 检验以及相关样本的非参数检验等分析过程, 则要求数据格式是变量组结构的, 便于对重复测量的分析。当数据文件的结构与分析过程所要求的不一致时, 需要通过对数据文件的重组来改变文件结构。

	no	d	area	quarter	sale
1	1	1.7	HeBei	1	80
2	1	1.7	HeBei	2	90
3	1	1.7	HeBei	3	87
4	1	1.7	HeBei	4	76
5	2	1.7	BeiJin	1	45
6	2	1.7	BeiJin	2	67
7	2	1.7	BeiJin	3	46
8	2	1.7	BeiJin	4	87
9	3	1.9	ShangHai	1	78
10	3	1.9	ShangHai	2	76
11	3	1.9	ShangHai	3	88
12	3	1.9	ShangHai	4	89
13	4	1.8	TianJin	1	78
14	4	1.8	TianJin	2	87
15	4	1.8	TianJin	3	79
16	4	1.8	TianJin	4	76
17	5	1.7	HuNan	1	98
18	5	1.7	HuNan	2	99
19	5	1.7	HuNan	3	95
20	5	1.7	HuNan	4	94

图 3-50 纵向结构的数据格式

### 3.4.1 选择数据重组方式

依次单击菜单“Data→Restructure”执行文件结构重组的功能, 弹出如图 3-51 所示的选择界面。在此, 可以选择如下 3 种类型的重组方式:

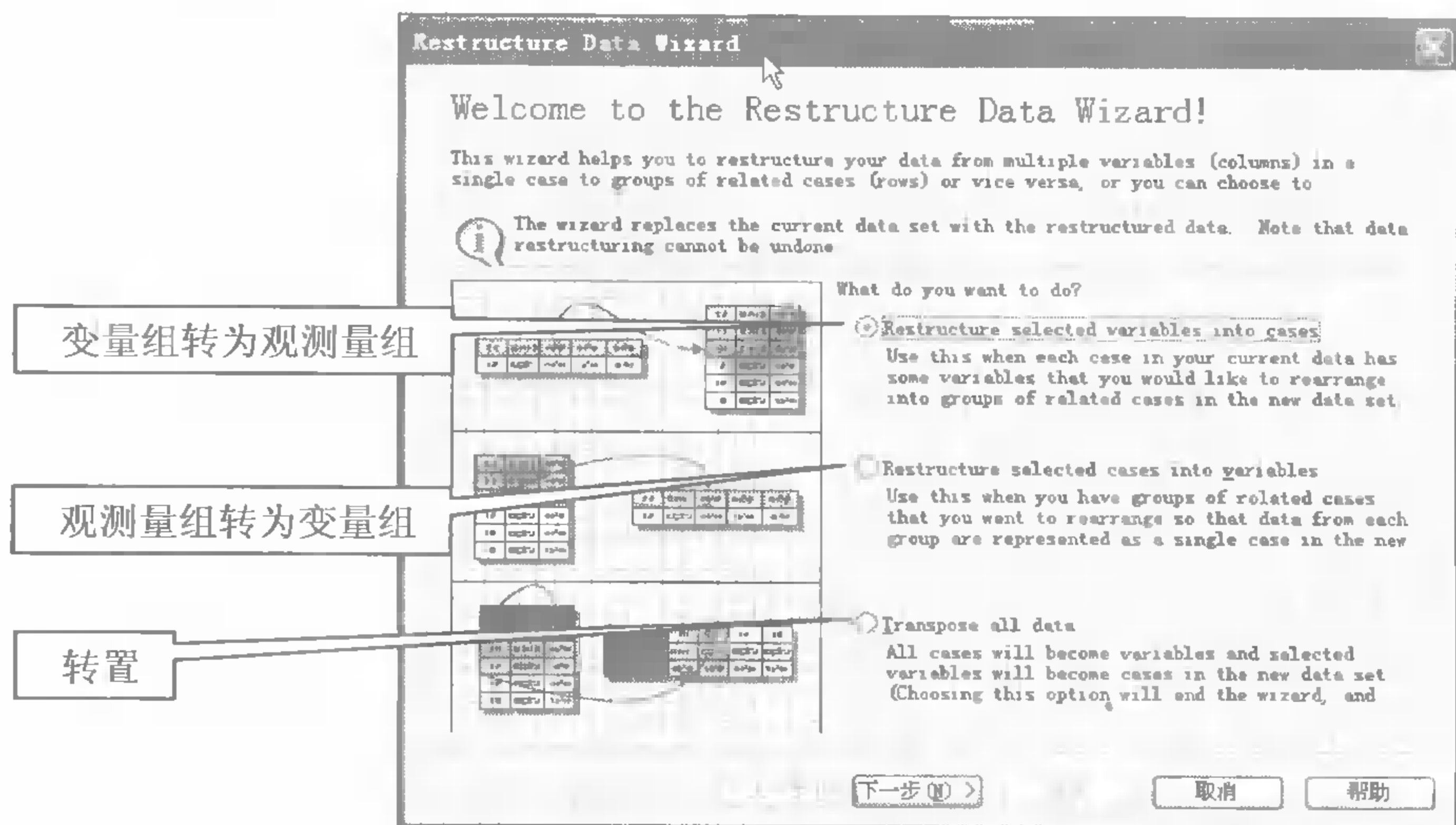


图 3-51 数据文件重组主对话框



- Restructure selected variables into cases 选项，表示将变量组结构转换为观测量组结构，即横向结构到纵向结构的转换。
- Restructure selected cases into variables 选项，表示将观测量组结构转换为变量组结构，即纵向结构到横向结构的转换。
- Transpose all data 选项，表示对所有数据形成的二维矩阵进行转置操作。

下面以图 3-49 和图 3-50 所示的文件为例，对这 3 种重组方式分别介绍。

### 3.4.2 变量组到观测量组的重组

本节演示如何将变量组格式的数据文件转换为观测量组格式的数据文件。首先打开文件“季度销售额的横向格式.sav”，如图 3-49 所示。

#### 1. 变量组个数的选择

在图 3-51 中单击选中 Restructure selected variables into cases 选项，单击下一步按钮，进入如图 3-52 所示的设置界面，单击选中 One 单选框。

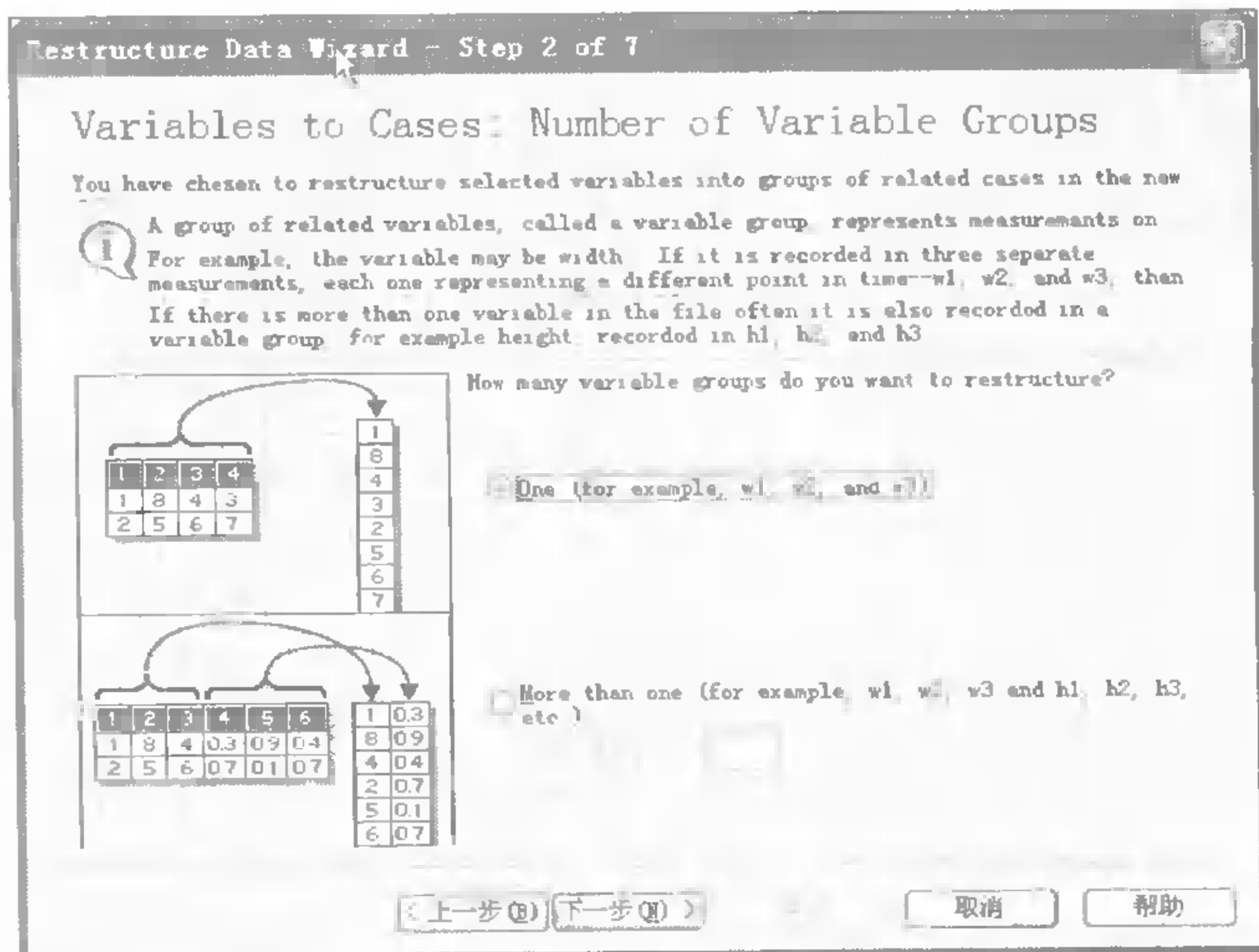


图 3-52 变量组到观测量组的重组第 2 步

此处选择需要进行重组的变量组的个数，有如下两个选项。




- One 单选框，表示只对一个变量组进行重组。
- More than one 单选框，表示要对多个变量组进行重组，选中后在 How many 后的输入框指定变量组的个数，默认值为 2。

本例只对 1 个变量组 quarter 进行重组，故选择 One 单选框；对于选择多个变量组的情况，后面的步骤与选择单个变量组时的情况类似，在随后的介绍中我们将会适时地提醒读者不同之处。

#### 2. 设置要被重组的变量

在图 3-52 中单击下一步按钮，进入图 3-53 所示的设置界面，在这里设置要被重组的变量组里的变量关系。单击 Case Group Identification 栏的下拉列表，选中 Use selected variable



选项，然后在左侧的变量列表选中公司编号变量，单击从上至下第一个  按钮，将其选入 Variable 选框作为标识变量；在 Target 后输入框的键入“sale”；在变量列表选中 1 至 4 季度销售这 4 个变量，单击从上至下第二个  按钮，将其作为变量组成员选入 sale 下的列表框；在变量列表选中距离和销售地点变量，单击从上至下第三个  按钮，将其作为固定变量选入 Fixed 列表。

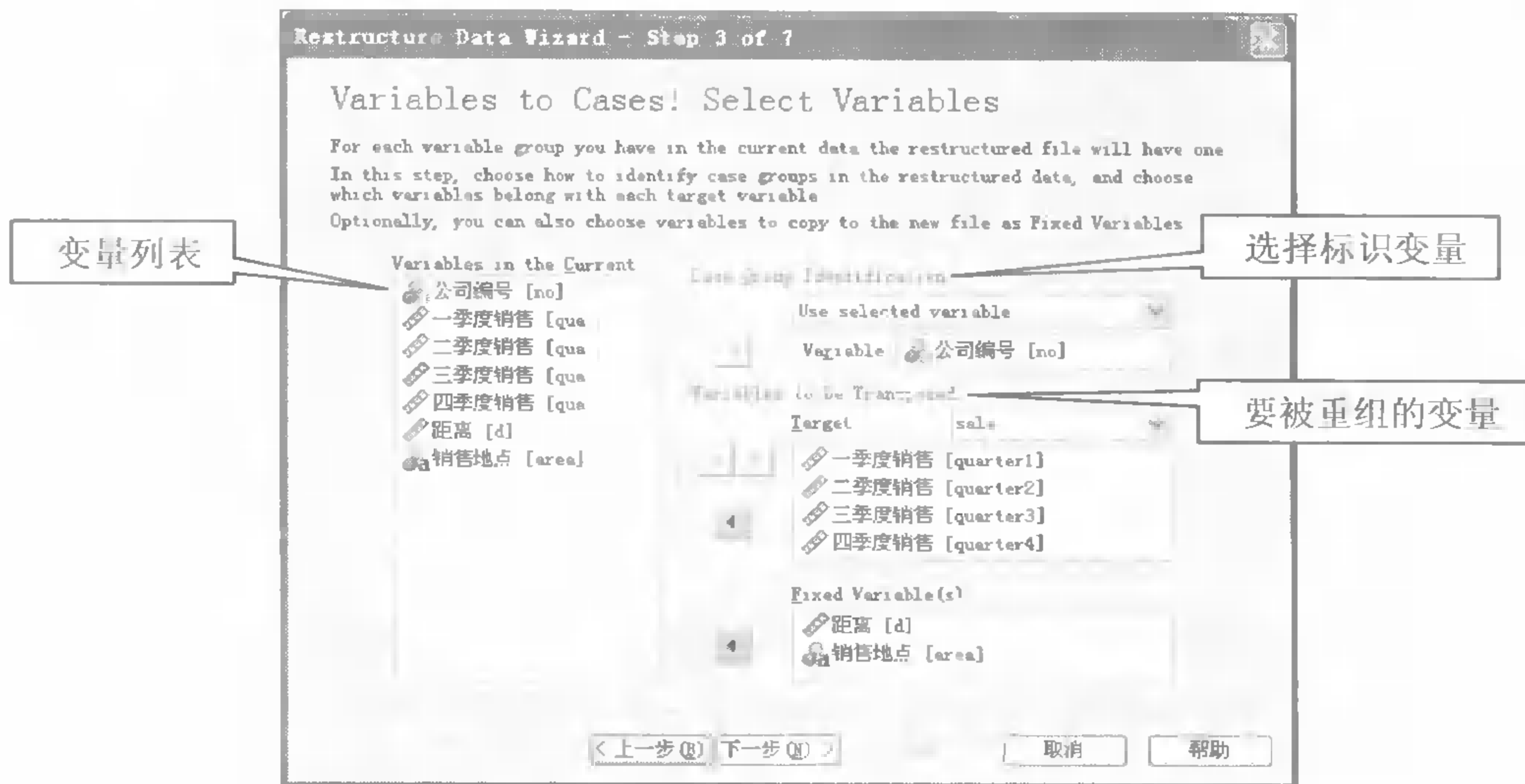




图 3-53 变量组到观测量组的重组第 3 步

下面详细介绍各设置选项的含义。

(1) Case Group Identification 栏。设置在变量组格式的数据中对观测记录的标识变量（例如本例的公司编号）。在下面的下拉列表里有如下 3 种指定标识变量的方法。

- Use case number 选项，使用观测量序号作为标识变量，选中此项后在下拉列表下方出现新的设置选项：，在 Name 后面输入重组后序号变量的变量名（默认为 id），单击 Label 按钮设定重组后序号变量的变量标签。
- Use selected variable 选项，由用户指定一个标识变量，选中此项后在下拉列表下方出现新的设置选项：，用于从左侧的变量列表选入一个指定的标识变量。
- None 选项，不使用标识变量，无新增设置选项。

(2) Variable to be Transposed 栏。用于设置要进行转换的变量组，如果在图 3-52 中指定要重组 N 个变量组，就在 Target 后的下拉列表中给出 trans1、trans2...transN 这 N 个名称来代表 N 个变量组，对默认的变量组名称可以直接进行编辑。先在此下拉列表中指定某个变量组名称，再在 Target 下面的列表框中选入指定变量组所包含的变量。在 Target 下的列表框中选中某个变量后，单击左侧的向上、向下两个黑色箭头可以调整变量的显示顺序。

(3) Fixed Variable(s)列表。固定变量列表，用于从变量列表选入不参与格式重组，但是仍要出现在转换结果中的变量。

### 3. 索引变量的个数设置

在图 3-53 中单击下一步按钮，进入如图 3-54 所示的设置界面，在这里设置重组后所要

生成的索引变量的个数，单击选中 One 单选框。

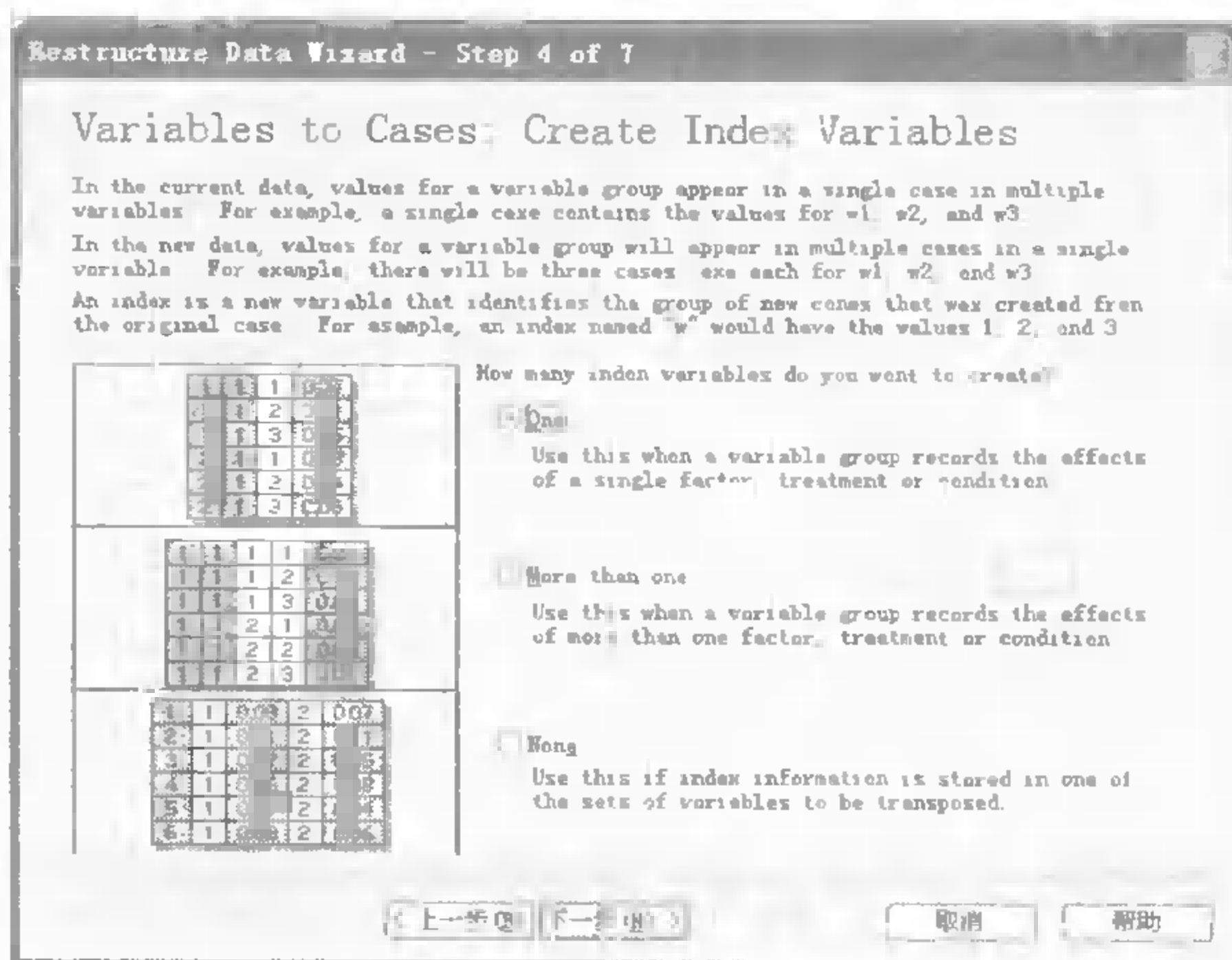


图 3-54 变量组到观测量组的重组第 4 步

索引就是用来区分初始变量组里的各个变量的输出变量，可选项有如下 3 个：

- (1) One，生成一个索引变量，这是默认选项。
- (2) More than one，生成多个索引变量，如果一个变量组记录了多个影响因素作用下的数据，选中此项，并在 How Many 后面输入索引变量的个数。
- (3) None，不生成索引变量，如果索引信息已被存在某个要转换的变量组里，选中此项。

#### 4. 索引变量的参数设置

在图 3-54 中选中 One 之后，单击下一步按钮，进入如图 3-55 所示的设置界面，在这里设置重组后所要生成的索引变量的参数。在 Edit the index 栏中 Name 下输入“quarter”取代原来的内容，Label 下输入“季度”作为变量标签。

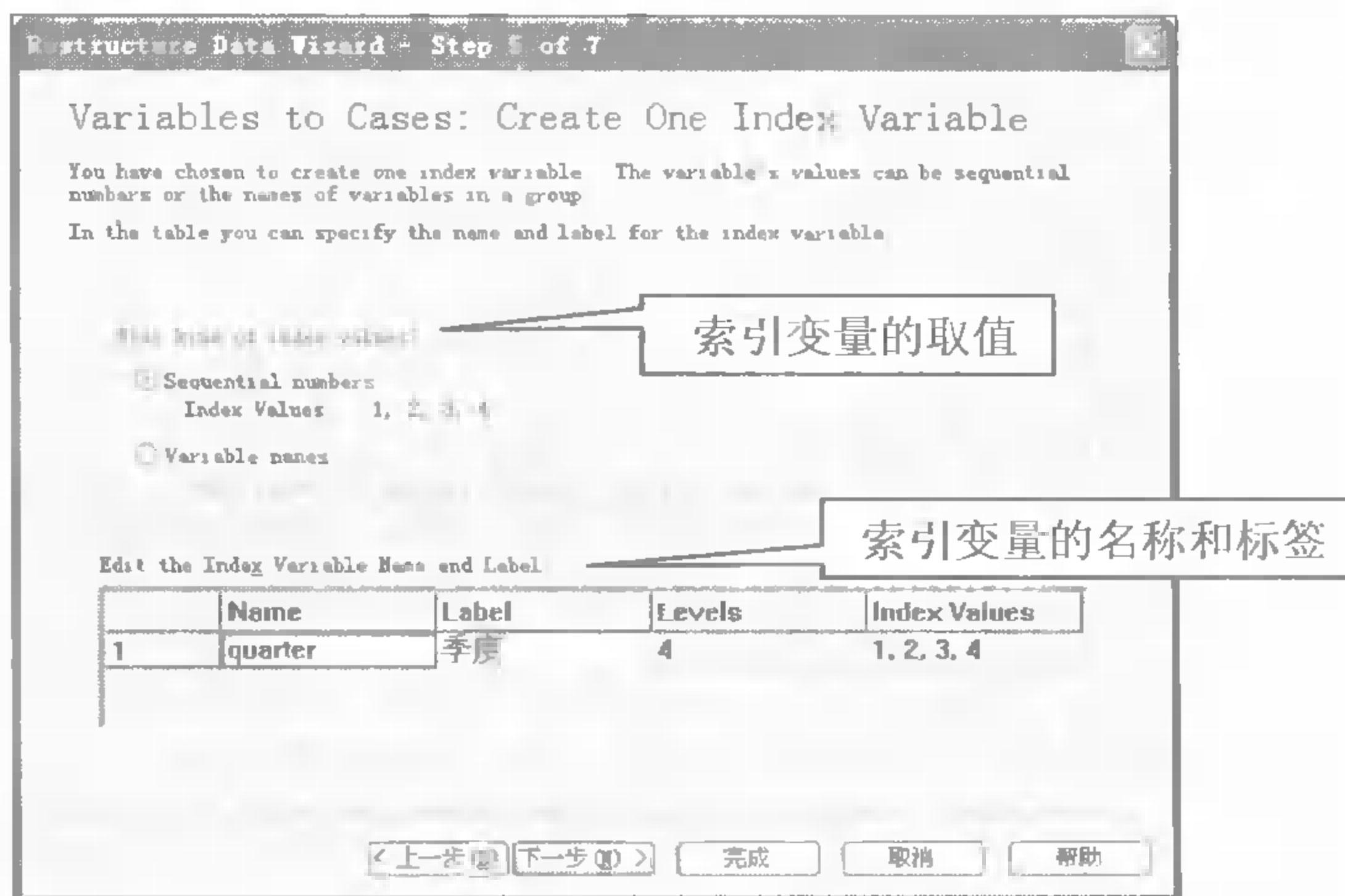


图 3-55 变量组到观测量组的重组第 5 步

索引变量的设置选项包括如下两个部分。

- ① What kind of index values 栏。设置索引变量的取值，方式有如下两种：Sequential

numbers 选项, 索引变量自动赋值为递增的整数序列 (1, 2, 3, ...), 由于本例要转换的 sale 变量组含有 quarter1~quarter4 变量, 故索引自动赋值为 1~4; Variable names 选项, 使用变量组所含各变量的名称作为索引取值, 例如本例中的 quarter1~quarter4。

② dit the index variable name and label 栏。编辑索引变量的变量名和变量标签, 本例只有一个索引变量, 用来区分 4 个季度。在 Name 下指定索引变量的名称; 在 Label 下指定索引变量的标签。

## 5. 其他参数设置

在图 3-55 中单击下一步按钮, 进入如图 3-56 所示的设置界面, 在这里设置文件结构重组的其他参数。

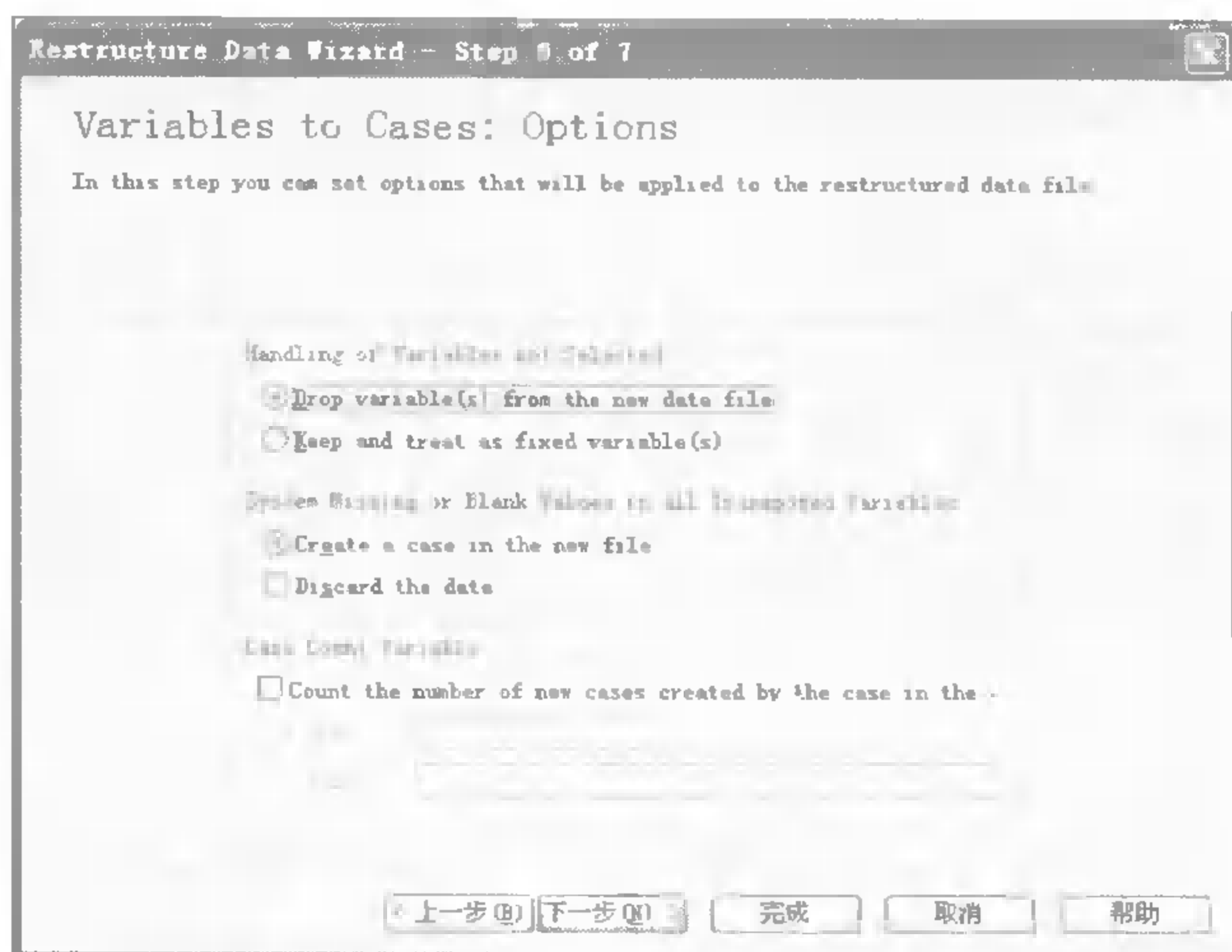


图 3-56 变量组到观测量组的重组第 6 步

设置内容有如下 3 个部分。

① Handling of Variables not selected 栏。设置如何处理原始数据中没有用到的变量, 也就是在图 3-53 中没起任何作用的变量, 对这部分变量的处理方式有如下两种。

- Drop variable(s) from the new data file 选项, 直接丢弃没有用到的变量, 转换后的结果直接删除了这部分变量的信息。
- Keep and treat as fixed variable(s)选项, 将这部分变量作为固定变量对待, 就如同在图 3-53 中选入 Fixed Variable(s)列表里的变量一样。

② System Missing and Blank Values in all Transposed Variables 栏。设置如何对待被重组的变量取值为系统缺失值或空值时的情况, 处理方式有如下两种。

- Create a case in the new file 选项, 在结果中为其单独生成一条观测记录。
- Discard the data 选项, 在结果中删除这部分数据的信息。

③ Case Count Variable 栏。用于选择是否生成计数变量, 当在上面选择 Discard the data 选项丢弃缺失值和空值时, 计数变量起到重要的计数作用。

Count the number of new case created by the case in the current data file 复选框, 选中后表示对原始数据的一个观测记录转换后生成的新观测记录进行计数, 并生成一个计数变量保存计数信息。在 Name 后输入计数变量的名称, 在 Label 后输入计数变量的标签。

6. 完成界面

在图 3-56 中单击下一步按钮，进入如图 3-57 所示的完成界面。

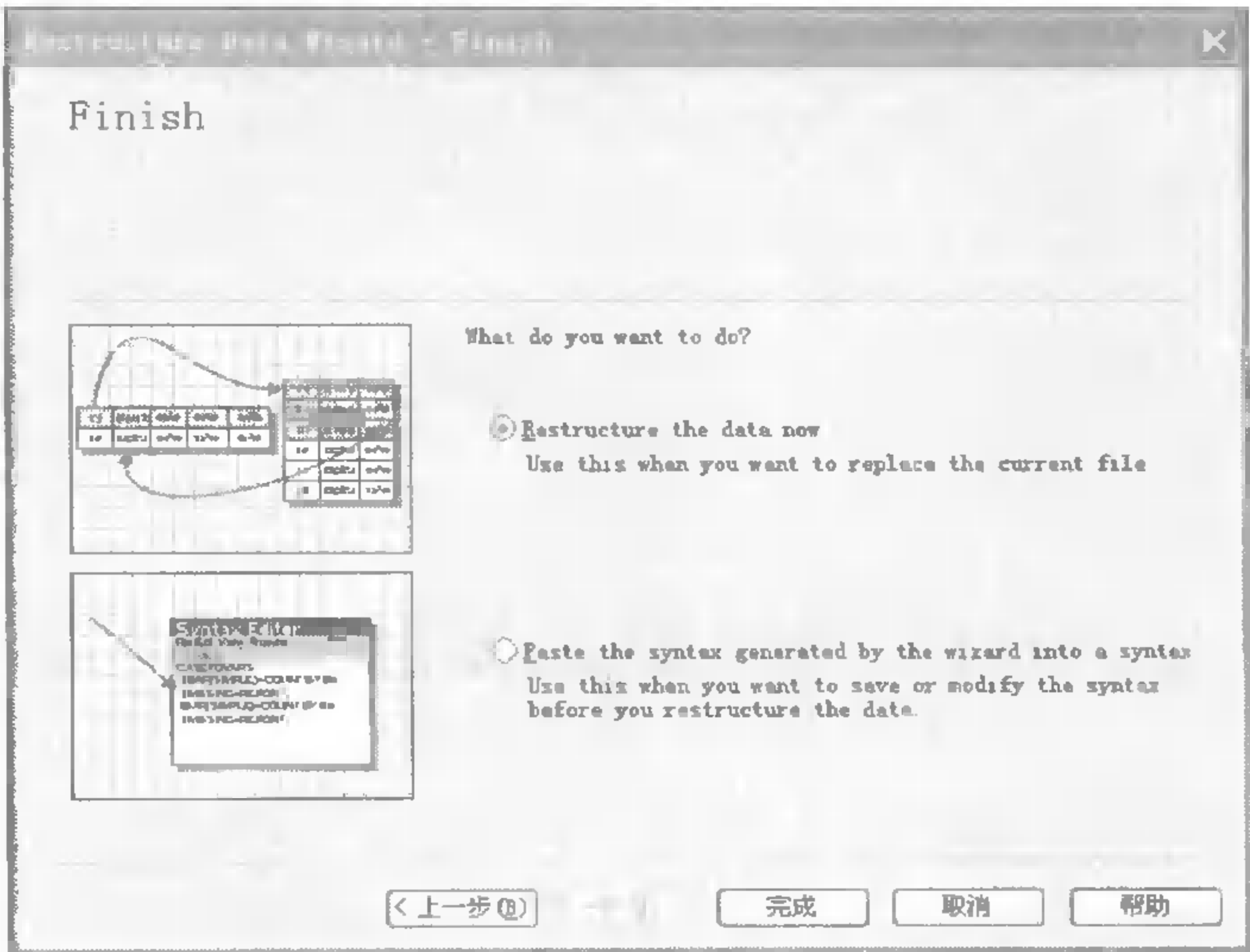


图 3-57 变量组到观测量组的重组第 7 步

在这里设置是否立刻进行文件重组，可选项有如下两个。

- Restructure the data now 选项，表示立即执行数据格式的重组。
- Paste the syntax generated by the wizard into a syntax window 选项，表示将由前面的设置步骤产生的命令语句粘贴到 Syntax 窗口，可做进一步编辑修改，择时运行。

7. 结果显示

在图 3-57 中单击完成按钮，执行数据文件的重组过程，重组后的数据格式如图 3-50 所示。至此，完成了由变量组格式数据到观测量组格式数据的重组任务。



注意：重组后的结果放在了当前数据集里，覆盖了原始数据，建议将重组结果另存为其他文件，以保留原始数据的备份。

3.4.3 观测量组到变量组的重组

在变量组结构转换为观测量组结构之后，如何将其还原回去，这时就要进行观测量组结构到变量组结构的转换。本节演示如何将图 3-50 所示的观测量组格式的数据文件转换为图 3-49 所示的变量组格式的数据文件。

打开文件“季度销售额的纵向格式.sav”，数据格式如图 3-49 所示。

1. 选择被重组的变量

在如图 3-51 所示的选择界面中选中 Restructure selected cases into variables 选项，单击下一步按钮，进入如图 3-58 所示的设置界面。在变量列表中选中公司编号变量，单击从上至下第一个  按钮，将其作为标识变量选入 Identifier 列表框；在变量列表中选中季度变量，单击从上至下第二个  按钮，将其作为索引变量选入 Index 列表框。

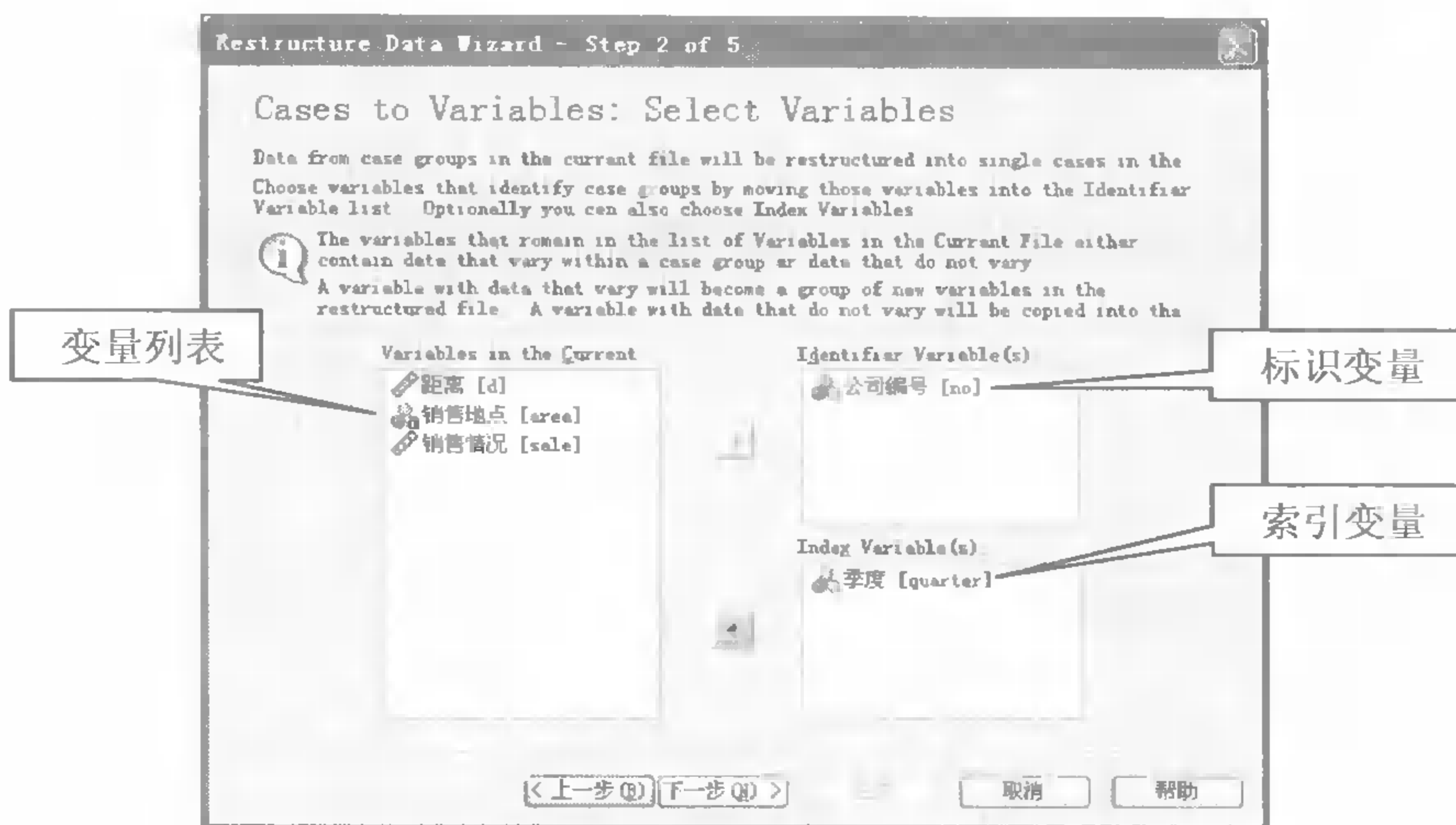


图 3-58 观测量组到变量组的重组第 2 步

此处设置要被重组的变量结构，待设参数有如下两个：Identifier Variable(s)列表，从变量列表选入标识变量，用于在重组后的变量组格式中标识观测记录；Index Variable(s)列表，从变量列表选入索引变量，用于区分重组后某个变量组里的不同变量。

## 2. 设置对原始数据的排序选项

在图 3-58 中单击下一步按钮，进入如图 3-59 所示的设置界面，在这里设置是否要对原始数据进行排序。

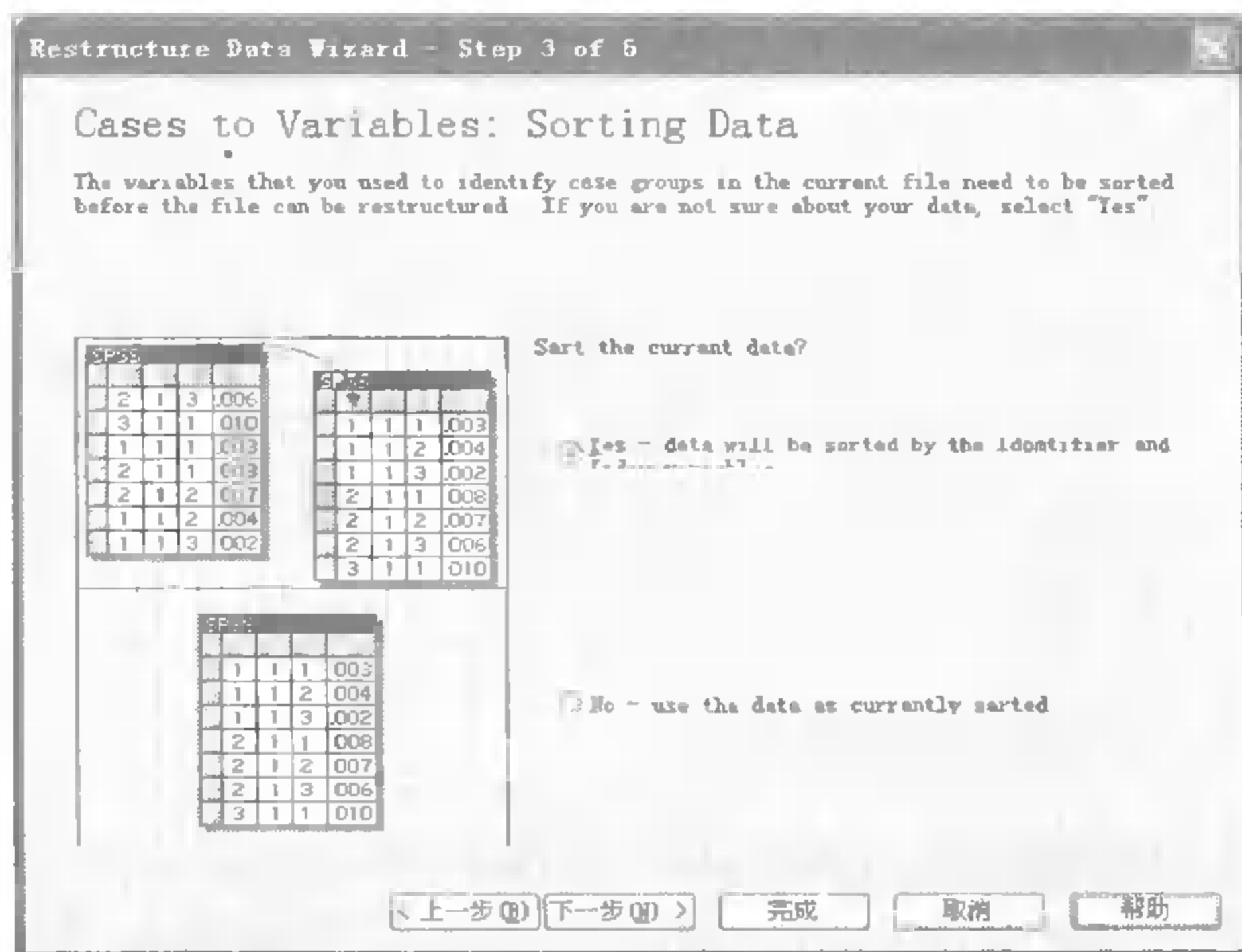


图 3-59 观测量组到变量组的重组第 3 步

对观测量组格式的数据进行重组之前，都要求原始数据按照指定的标识变量（例如本例的公司编号）进行排序，可选项有如下两个：Yes 选项，表示在对数据进行重组之前，按照 Identifier 列表指定的标识变量（保持变量的顺序）对原始数据进行排序；No 选项，表示不进行排序，如果原始数据已经按照指定的标识变量排序了，选中此项。

## 3. 设置关于新变量的参数

在图 3-59 中单击下一步按钮，进入如图 3-60 所示的设置界面，在这里设置一些新变量的参数。



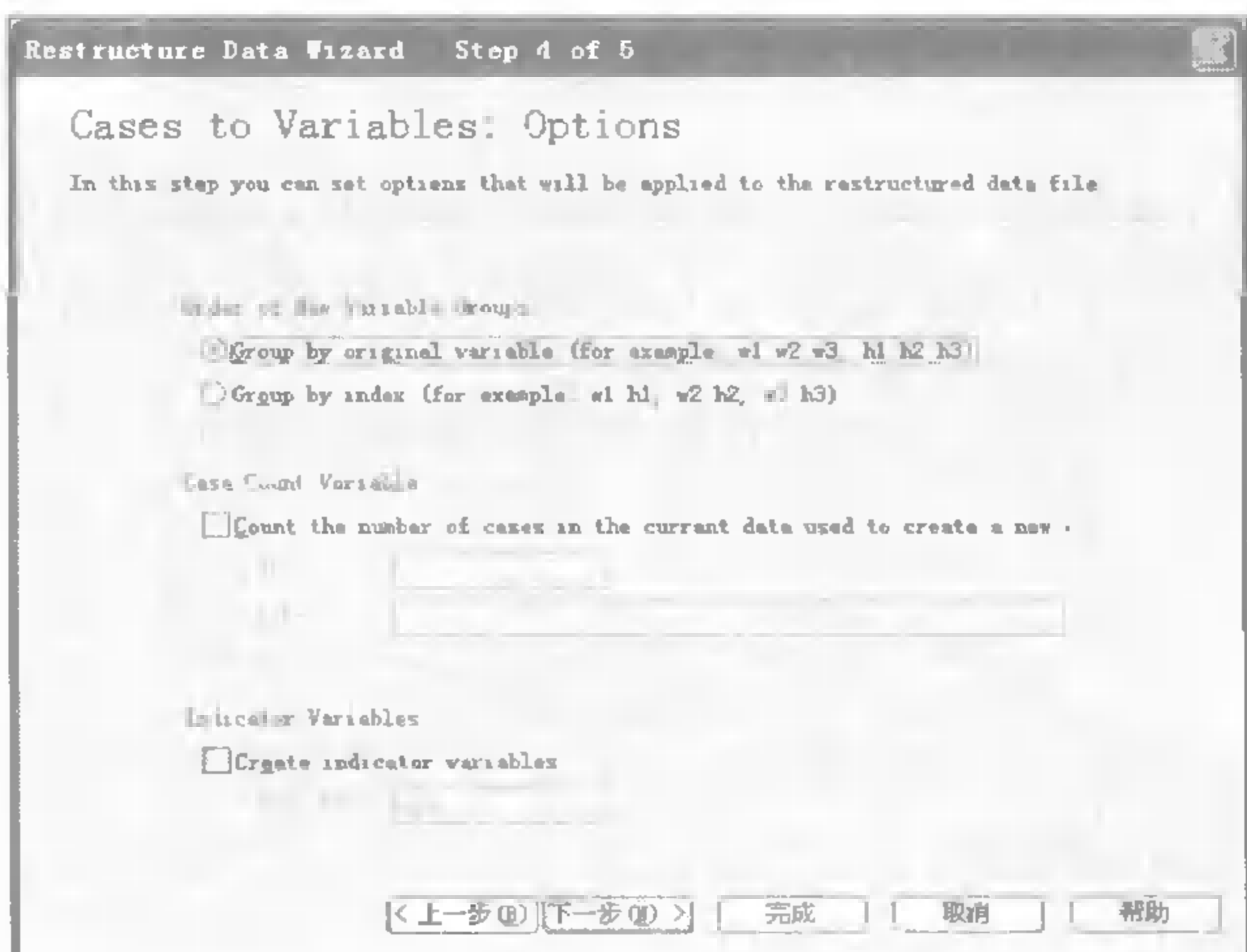


图 3-60 观测量组到变量组的重组第 4 步

设置内容如下 3 个部分。

① Order of New Variable Groups 栏。当重组结果包含多于 1 组的新变量组时，在此设置新变量组的显示顺序，如果只有一个输出变量组，此处两个选项的输出都一样。

- Group by original variable 选项，新变量按原始变量的顺序成组排列，例如有两个被重组的变量 w 和 h，则重组后新变量的显示顺序为：w1, w2, w3, h1, h2, h3。
- Group by index 选项，新变量按索引变量的取值顺序排列，例如有两个被重组的变量 w 和 h，则重组后新变量的显示顺序为：w1, h1, w2, h2, w3, h3。

② Case Count Variable 栏。用于选择是否生成计数变量。选中 Count the number of cases in the current data used to create a new 复选框，表示对转换后生成的每个新观测记录所使用的原始数据的观测记录个数进行计数，并生成一个计数变量保存计数信息。在 Name 后输入计数变量的名称，在 Label 后输入计数变量的标签。

③ Indicator Variables 栏。用于选择是否生成指示变量。选中 Create indicator variable 复选框，表示对索引变量的每个取值生成一个指示变量，用它来记录对应的变量取值是否为空。指示变量取值 1 表示对应的变量取值非空；取值 0 表示对应的变量取值为空。选中此项后，还要求 Root Name 后输入指示变量的前缀。

#### 4. 完成界面

在图 3-60 中单击下一步按钮，进入如图 3-61 所示的完成界面。

在此设置是否立刻进行文件重组，可选项有如下两个。

- Restructure the data now 选项，表示立即执行数据格式的重组。
- Paste the syntax generated by the wizard into a syntax window 选项，表示将由前面的设置步骤所产生的命令语句粘贴到 Syntax 窗口，可做进一步编辑修改，择时运行。

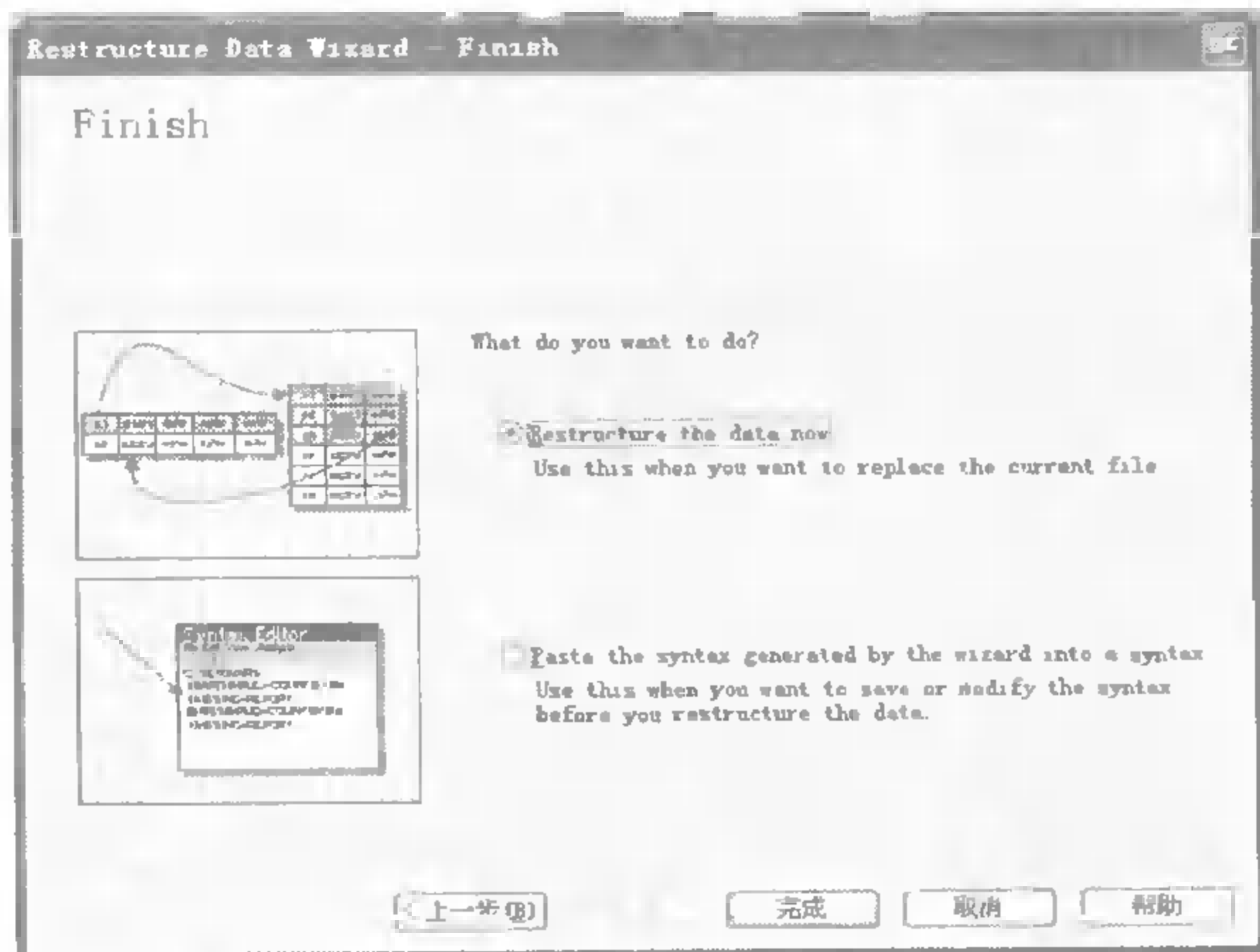


图 3-61 观测量组到变量组的重组第 5 步

## 5. 结果显示

在图 3-61 中单击完成按钮，执行数据文件的重组过程，重组后的数据格式如图 3-62 所示，与图 3-49 所示的数据格式基本相同，只是变量的名称和显示顺序不同。至此，就完成了由观测量组格式数据到变量组格式数据的重组任务。

	no	d	area	sale.1	sale.2	sale.3	sale.4
1	1	1.7	HeBei	80	90	87	76
2	2	1.7	BeiJin	46	67	46	87
3	3	1.9	ShangHai	78	76	88	89
4	4	1.8	TianJin	78	87	79	76
5	5	1.7	HuNan	98	99	95	94

图 3-62 从观测量组到变量组的重组结果

注意：重组后的结果放在了当前数据集里，覆盖了原始数据，建议将重组结果另存为其他文件，以保留原始数据的备份。

### 3.4.4 转置重组

在如图 3-51 所示的选择界面中单击选中 Transpose all data 选项，可以执行对所有数据的转置重组过程，单击完成按钮，弹出关于文件转置（Transpose）的设置对话框，如图 3-18 所示，它也就是单击菜单“Data→Transpose”所弹出的设置对话框。

关于文件的转置（Transpose）操作在第 3.1.4 节有详细的介绍，请读者参考。

# 第4章 基本统计分析功能

在进行统计分析和建模之前，经常需要对数据做一些描述性的工作，为此 SPSS 提供了许多过程，调用它们可以了解数据的基本统计指标，例如：对于定量数据，可以得到均数、标准差、标准误等指标；对于计数数据和一些分类数据，可以得到频率、比率等指标，还可以进行卡方检验等分析。本章就以实例的方式，来逐一介绍这些过程的具体操作方法。

## 4.1 OLAP 在线分析过程

OLAP (Online Analytical Processing, 在线分析过程)，主要用于对统计数据进行分析。它具有快捷、灵活多样的交互方式，能够根据用户的要求自由选择表格的报告方式和报告内容，而且分析结果简洁明了、便于理解。尤其对于多维数据资料，可以从不同的角度给出分析报告。

### 4.1.1 数据描述

对 4 个地区的学生分别实施了不同的教学方法后，收集到部分学生的英语和数学平均成绩，数据格式如图 4-1 所示，所用数据文件为“英语和数学成绩数据.sav”。本节通过 OLAP 过程，分析不同地区、不同教学方法下的成绩分布情况。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	area	String	8		地区类型	None	None	8	Right	Nominal
2	method	String	8		教学方案	None	None	8	Right	Nominal
3	score1	Numeric	8	0	英语成绩	None	None	8	Center	Scale
4	score2	Numeric	8	0	数学成绩	None	None	8	Right	Scale

图 4-1 关于考试成绩的数据格式

### 4.1.2 OLAP 过程的操作和设置

依次单击菜单“Analyze→Reports→OLAP Cubes”，打开 OLAP 分析的主设置对话框，如图 4-2 所示。

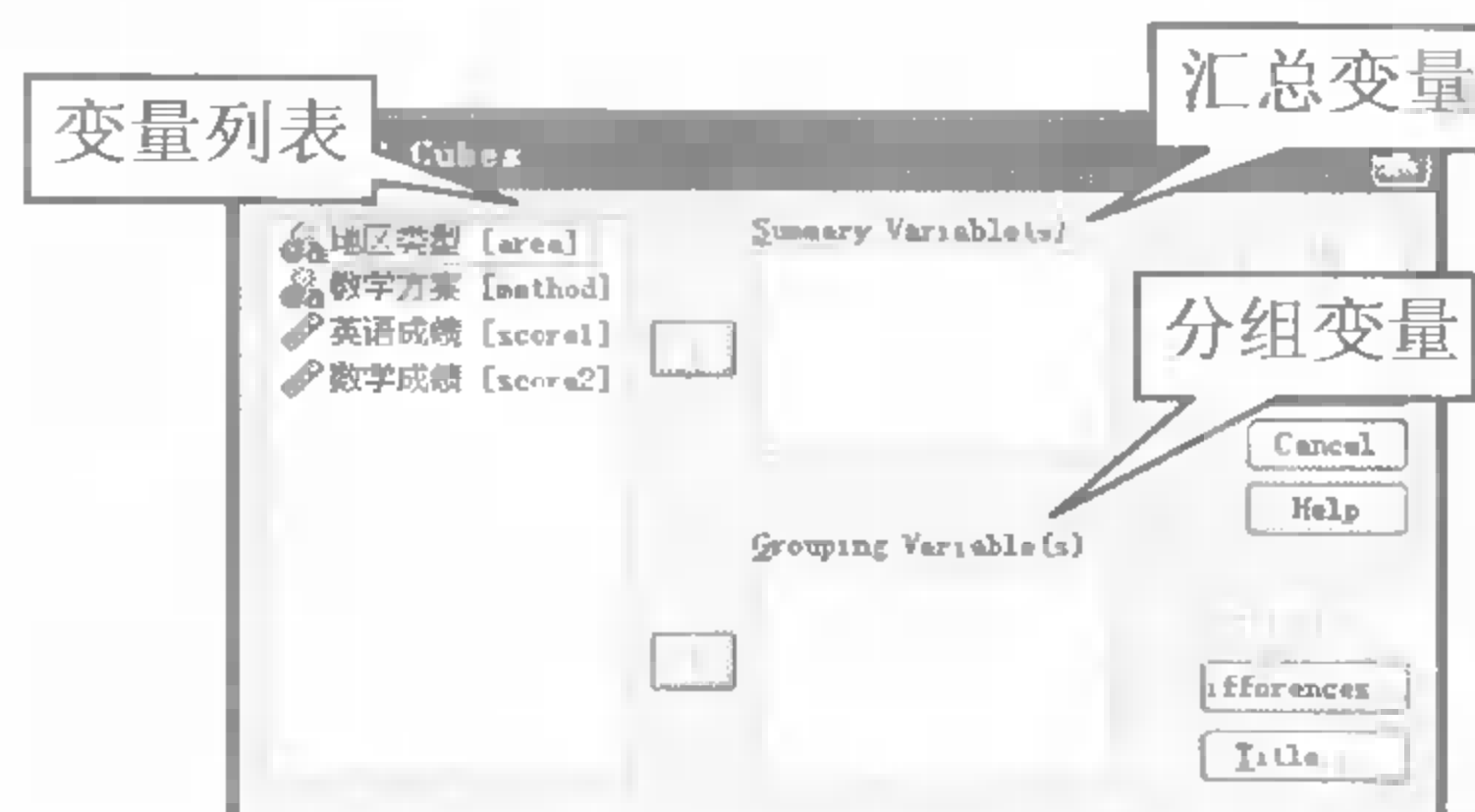




图 4-2 OLAP Cubes 对话框 1

## 1. OLAP 分析的变量设置

首先在变量列表中选中英语成绩、数学成绩两个变量，然后单击汇总变量列表左侧的  按钮，把选中变量选入汇总变量列表；接着在变量列表中选中地区类型、教学方案两个变量，然后单击分组变量列表左侧的  按钮，把选中变量选入分组变量列表。设置好后的界面如图 4-3 所示。

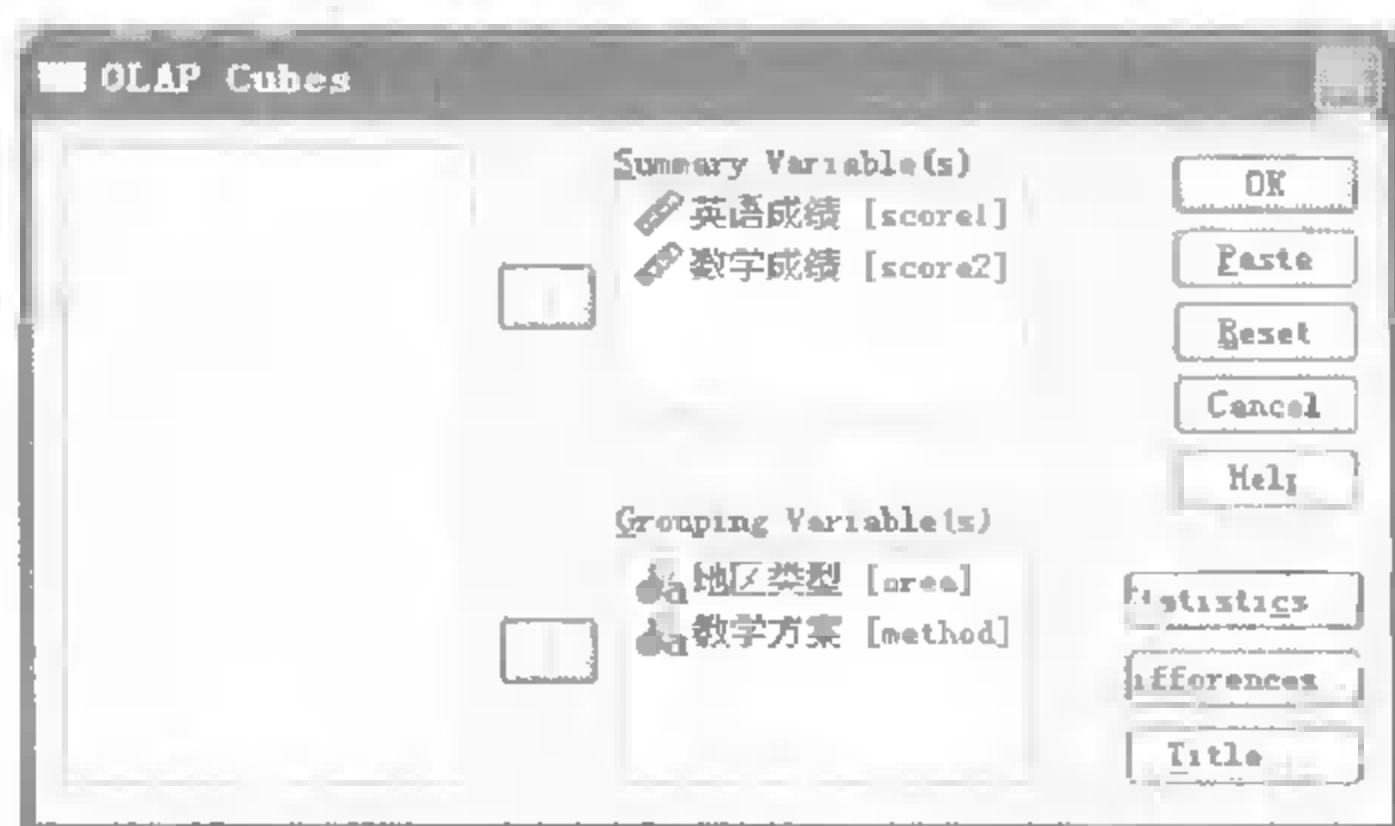



图 4-3 OLAP Cubes 对话框 2

- Summary Variables (s) 列表框：汇总变量。用于从变量列表选入汇总变量，一般要求为连续型变量。
- Grouping Variable (s) 列表框：分组变量。用于从变量列表选入分组变量，以便对汇总变量进行分组统计。

## 2. 分析选项的设置

(1) 统计量设置。在图 4-3 中，单击 Statistics 按钮，弹出统计量设置面板，如图 4-4 所示。首先在 Statistics 列表同时选中 Sum、Number of Cases、Mean、Standard Deviation、Percent of Total Sum 这 5 个选项，然后单击  按钮，将其选入右侧的 Cell 列表；单击 Continue 按钮返回主设置界面。

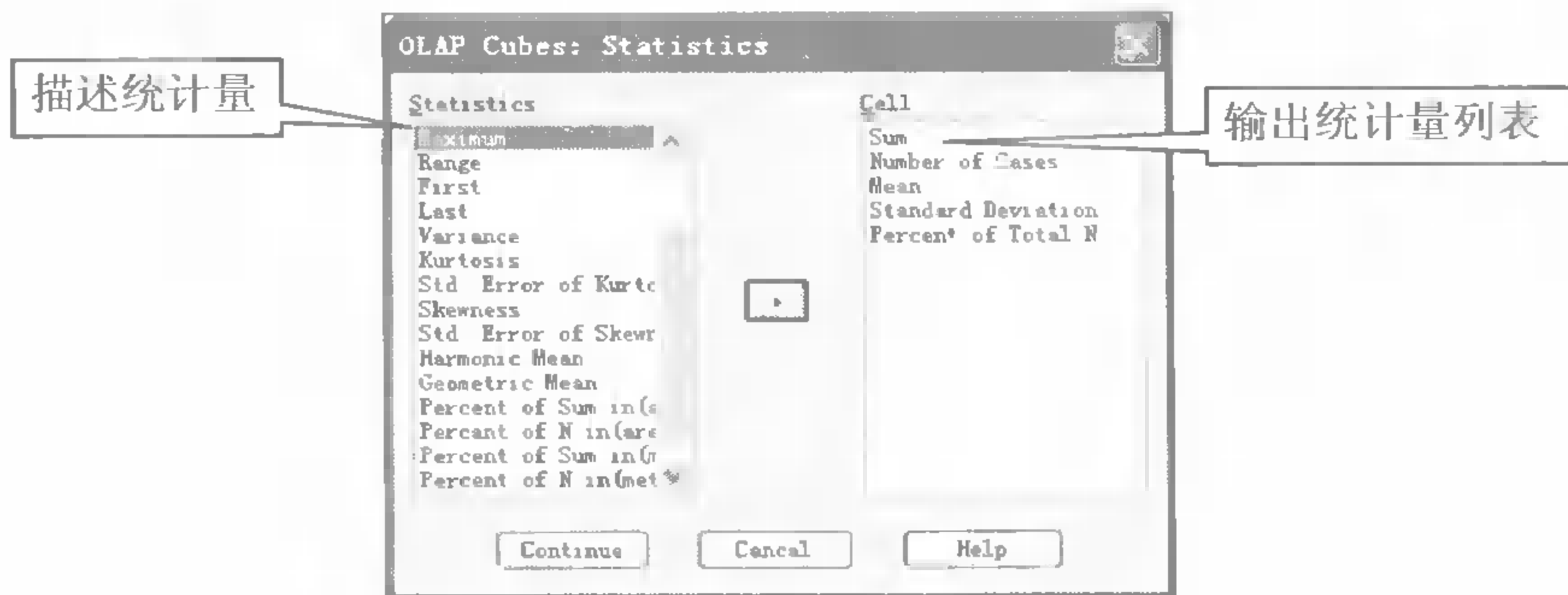



图 4-4 Statistics 对话框

- Statistics 列表显示了可选的统计量，包括：Sum（求和）、Number of Cases（个数）、Mean（均值）、Standard Deviation（标准差）、Percent of Total Sum（求和比例）等。
- Cell 列表的统计量将出现在最终输出的表格里。

(2) 差异统计方式的设置。在图 4-3 中，单击 Differences 按钮，弹出关于变量差异或分组差异统计的子设置面板，如图 4-5 所示。在此设置关于变量之间、或者分组变量的各个类别之间的差异统计选项，包括在图 4-4 中所选的所有输出统计量之间的差异。分别选中

Differences between groups 单选框和 Arithmetic differences 复选框,然后在 Differences between Groups of Cases 栏中的 Grouping 下拉列表选中 method(教学方案),在 Category 后输入 A(教学方案的可取值),在 Minus 后输入 B(也是教学方案的可取值),然后单击右侧的  按钮,把比较对“A-B”选入右侧的 Pairs 列表;设置好后的界面如图 4-6 所示。最后,单击 Continue 按钮返回主设置界面。

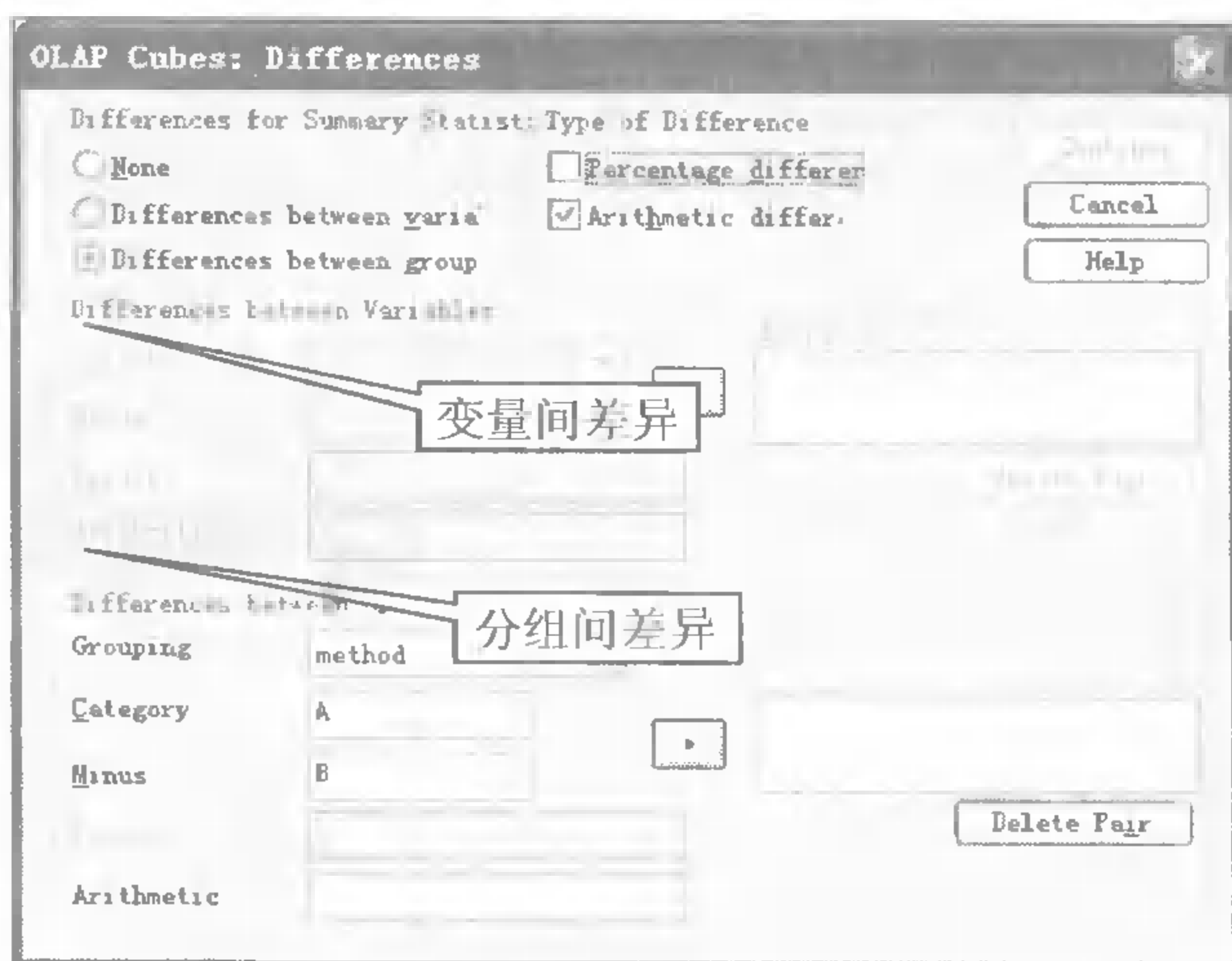


图 4-5 Differences 对话框 1

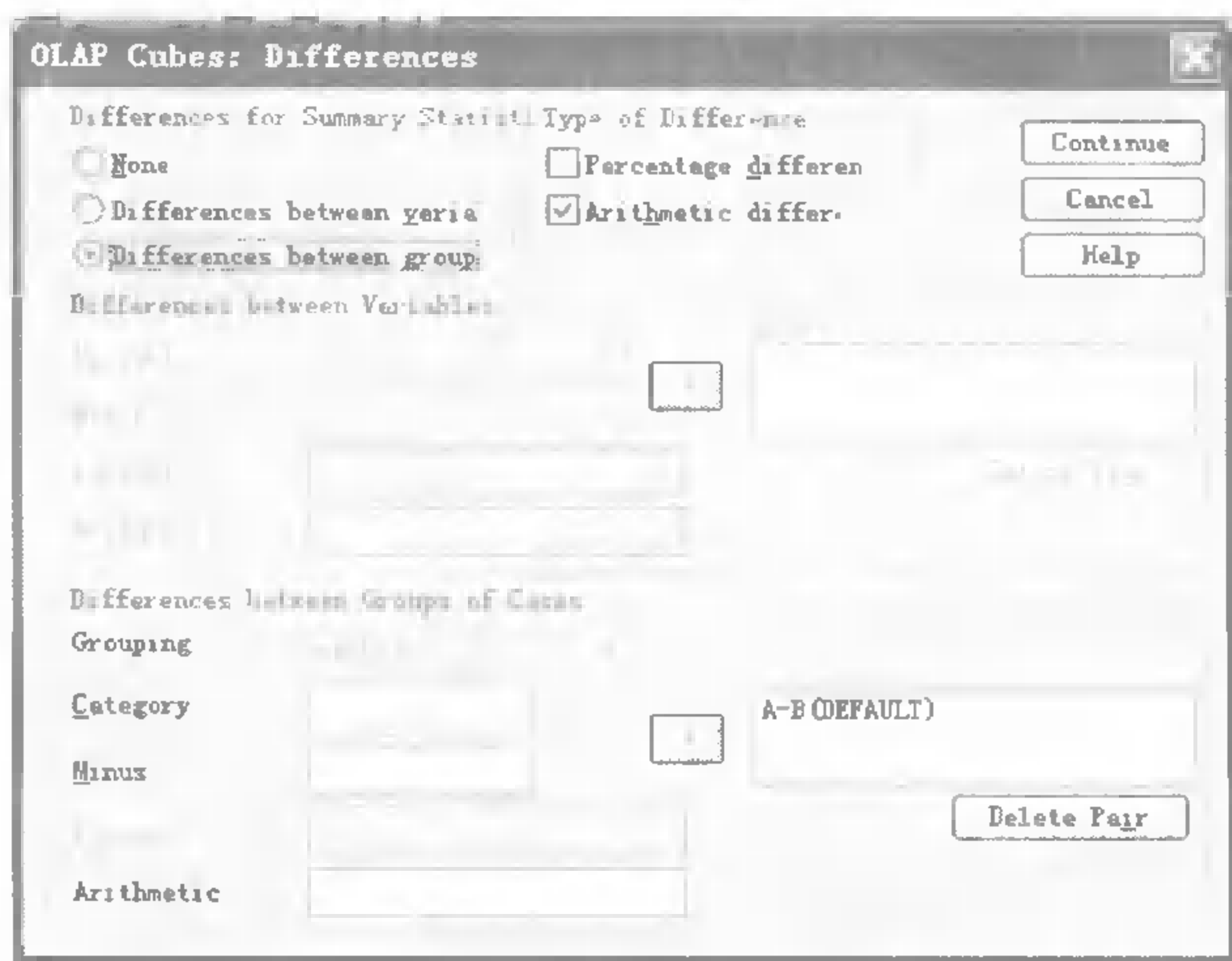


图 4-6 Differences 对话框 2

下面详细介绍各设置选项的含义。

- ① Differences for Summary Statistics 栏, 选择所要计算的差异对象, 有 3 个选项。
  - ① None 不进行差异计算, 默认选项;
  - ② Differences between variables 变量之间的差异, 选中后激活 Differences between Variables 栏的设置;
  - ③ Differences between groups 分组之间的差异, 选中后激活 Differences between Groups of Cases 栏的设置。
- ② Type of Differences 栏, 选择所要计算的差异统计量, 有两个选项: Percentage



differences (百分比差异)、Arithmetic differences (算术差异), 且二者必选其一。

选中 Arithmetic differences 时, 计算时用第一个变量 (或类别) 的统计量减去第二个变量 (或类别) 的统计量; 当选中 Percentage differences 时, 计算比例时用第二个变量 (或类别) 的统计量作为分母。

③ Differences between Variables 栏, 设置关于变量之间差异的选项, 需要至少两个汇总变量。

① Variables 栏, 从下拉列表选择要比较的第一个变量;

② Minus 栏, 从下拉列表选择要比较的第二个变量。

④ Differences between Groups of Cases 栏, 设置关于分组之间差异的选项, 需要至少一个分组变量。

① Grouping 栏, 从下拉列表选择要进行比较的分组变量;

② Category 栏, 输入对选中的分组变量进行比较的第一个类别的取值;

③ Minus 栏, 输入对选中的分组变量进行比较的第二个类别的取值。

(3) 标题设置。在图 4-3 中, 单击 Titles 按钮, 弹出关于输出表格标题的子设置面板, 如图 4-7 所示。在 Title 栏中, 输入表格标题“成绩 \n OLAP cubes”; 在 Caption 栏中, 输入表格注脚“成绩描述表格”。在标题和脚注中, “\n”代表换行符。最后, 单击 Continue 按钮返回主设置界面。

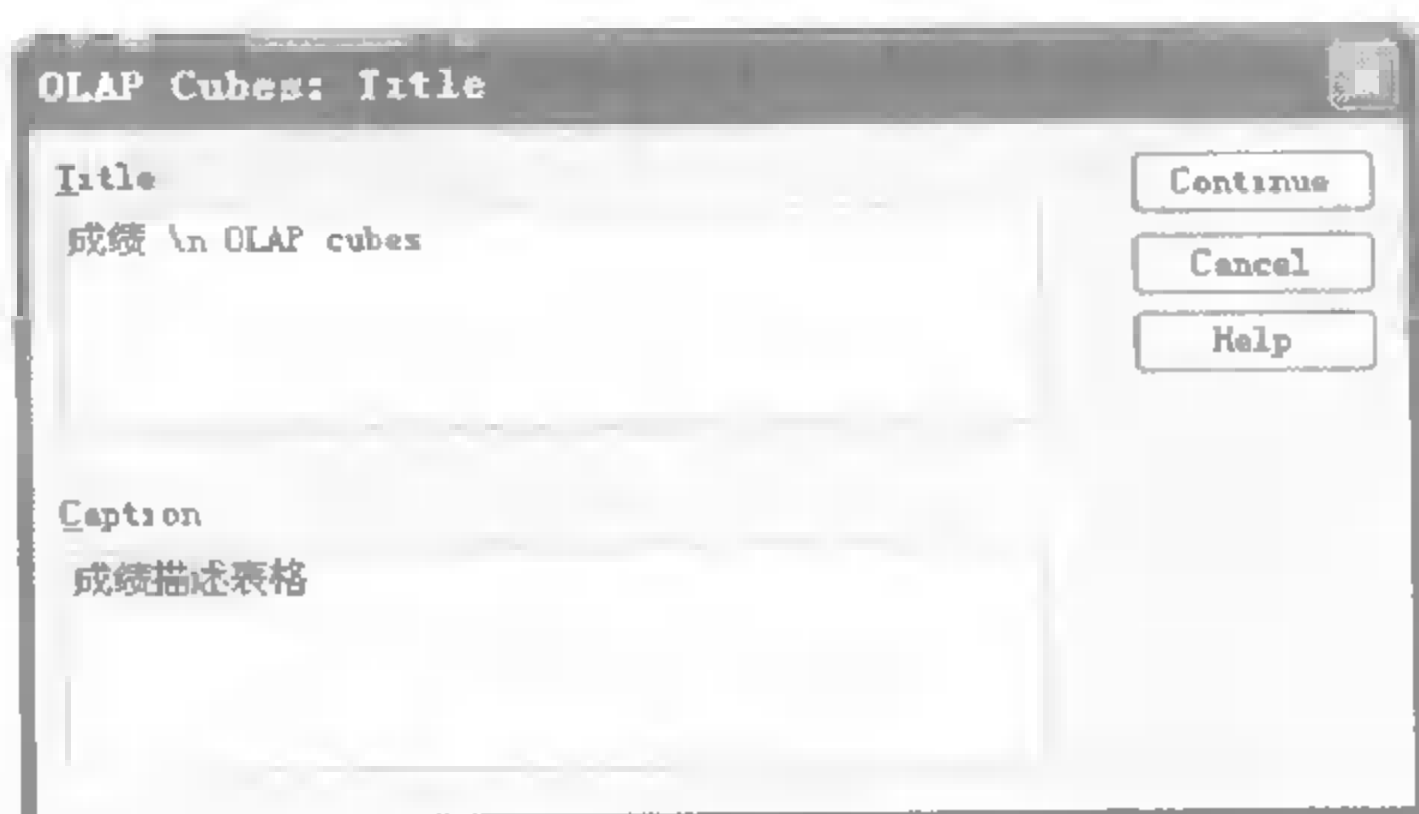


图 4-7 Title 对话框

### 3. 结果分析和交互式操作

在图 4-3 中单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如下。

(1) 案例处理摘要。首先输出的是如图 4-8 所示的案例处理摘要信息, 它给出了分析中用到的案例个数和比例, 以及排除 (例如由于缺失而排除) 的案例个数和比例。可见, 本例的 12 条记录全部用于了统计分析。

案例处理摘要						
	案例					
	已包含		已排除		总计	
	N	百分比	N	百分比	N	百分比
英语成绩 * 地区 类型 * 教学方案	12	100.0%	0	0%	12	100.0%
数学成绩 * 地区 类型 * 教学方案	12	100.0%	0	0%	12	100.0%

图 4-8 处理摘要输出


(2) OLAP 表格。最终输出的是 OLAP 统计表格, 如图 4-9 所示。表格的第一行统计量名称对应于在图 4-4 中所选择的输出统计量; 英语成绩和数学成绩的统计特征显示于相应的单元格里。

成绩 OLAP cubes					
地区类型 教学方案	总计				
	总计				
	合计	N	均值	标准差	合计 N 的 %
英语成绩	1109	12	92.42	21.091	100.0%
数学成绩	1162	12	96.83	16.508	100.0%

成绩描述表格

图 4-9 OLAP 统计表格



钮，将其选入 Variables 列表；接着在变量列表选中教学方案，然后单击 Grouping 列表左侧的  按钮，将其选入 Grouping 列表；设置变量后的界面如图 4-12 所示。

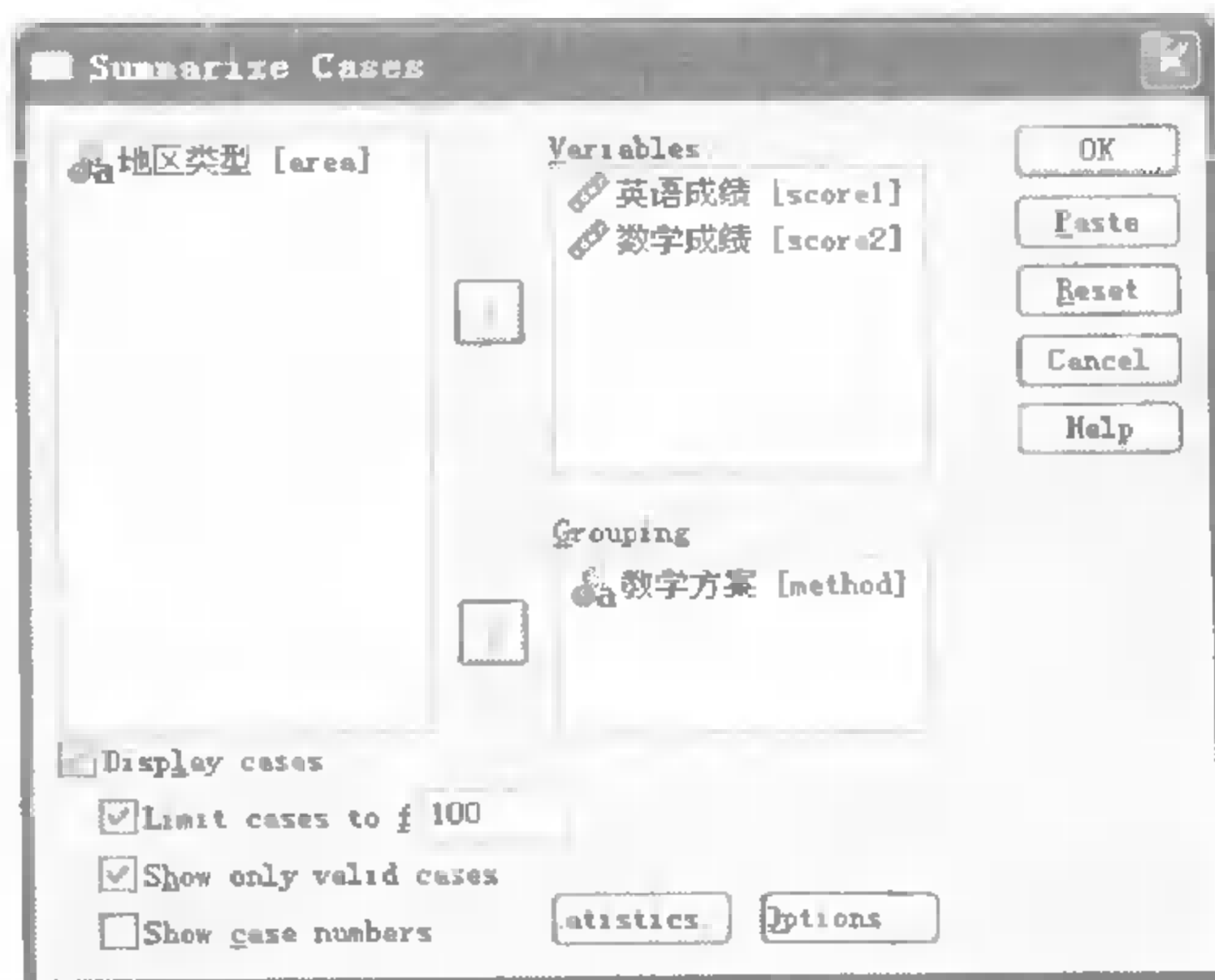



图 4-12 Summarize Cases 对话框 2

① Variables 摘要统计变量列表，选入对数据进行分类汇总的汇总变量；Grouping 分组变量列表，选入对数据进行汇总的分类变量。

② Display cases 复选框，表示在结果中列出数据集里的每个记录行，有 3 个限制显示数量的设置选项。

- Limit cases to first 复选框，显示每个类别的前 n 条记录，n 为后面输入的整数。
- Show only valid cases 复选框，只显示有效的记录。
- Show case numbers 复选框，显示记录的行号。

## 2. 统计量设置

在图 4-12 中，单击 Statistics 按钮，弹出统计量设置面板，如图 4-13 所示。在 Statistics 列表选中 Mean、Standard Deviation 两个选项，单击  按钮，将其选入 Cell 列表，如图 4-14 所示；单击 Continue 按钮返回主设置界面。

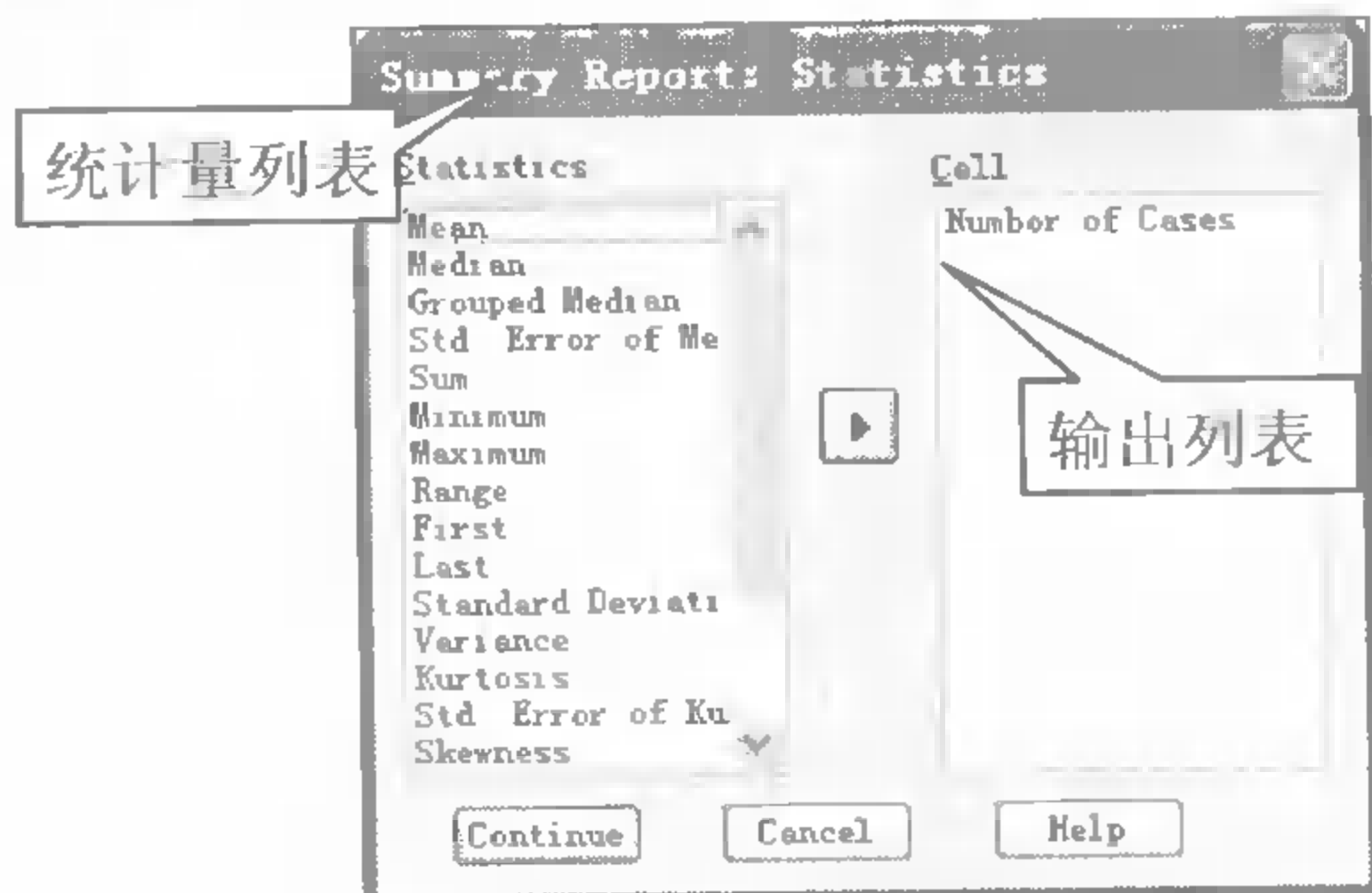


图 4-13 Statistics 设置 1

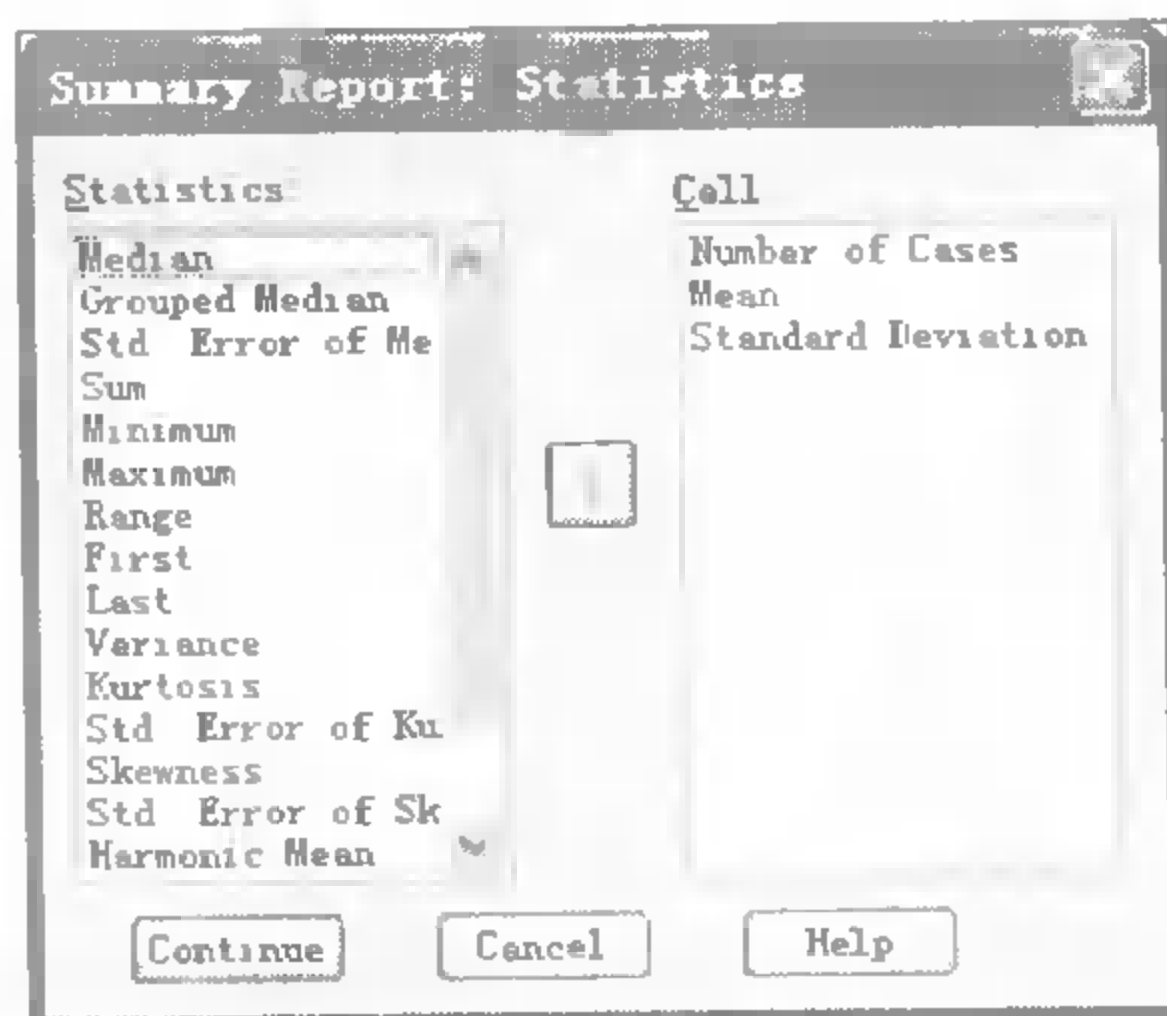


图 4-14 Statistics 设置 2

Statistics 列表显示了可选的统计量；Cell 列表的统计量将出现在结果表格里，有 Number of Cases（个数）、Mean（均值）、Standard Deviation（标准差）3 项。

## 3. Options 选项设置

在图 4-12 中，单击 Options 按钮，弹出选项设置面板，如图 4-15 所示，在此设置表格标

题、缺失值处理方式等内容。在 Title 栏中,输入表格标题“观测摘要输出”;在 Caption 栏中,输入表格注脚“成绩的观测摘要”。单击 Continue 按钮返回主设置界面。

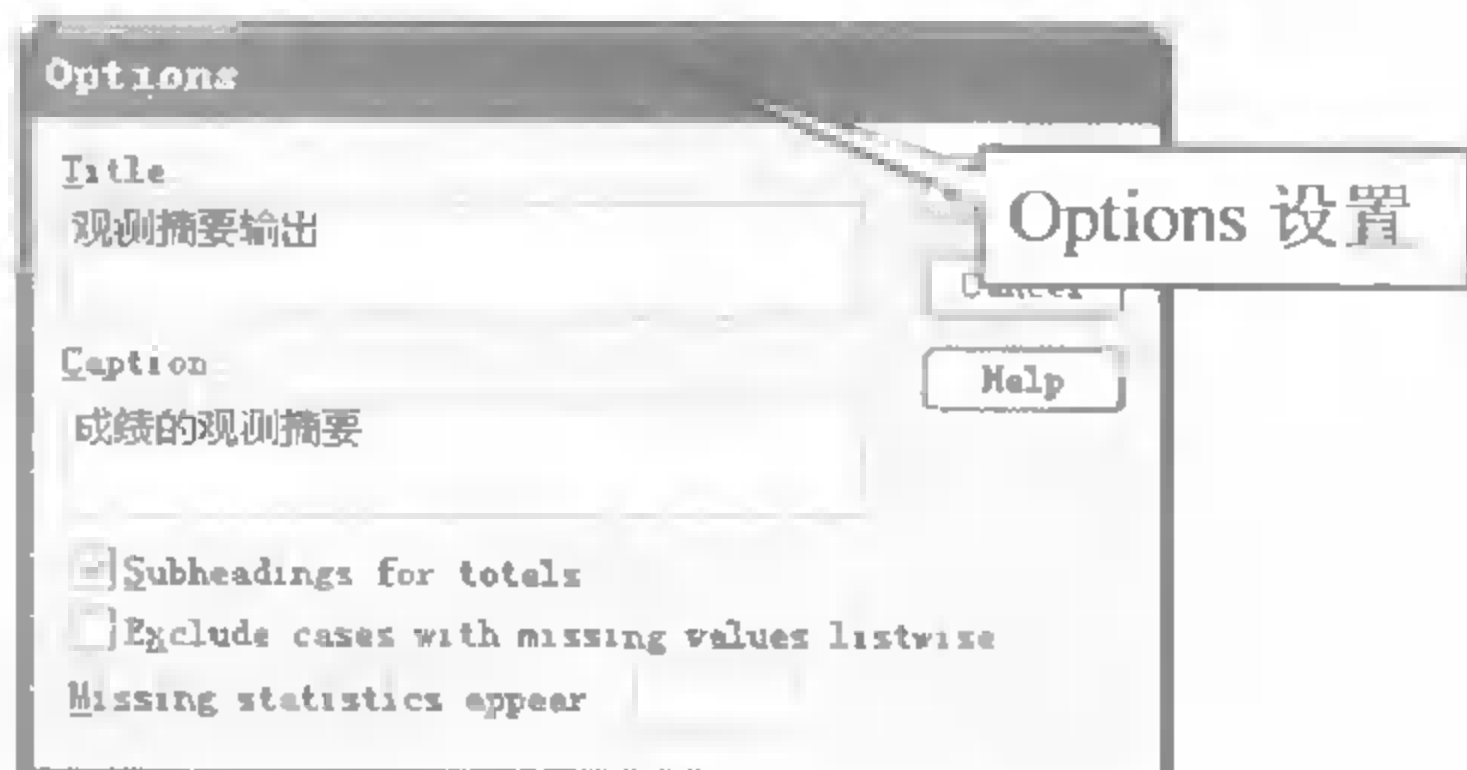


图 4-15 Options 设置对话框

- ④ Subheadings for totals 复选项,表示把统计量的名称作为子标题显示在单元格里。
- ④ Exclude cases with missing values listwise 复选项,表示只要分析中有一个变量取值缺失,就把这条记录从分析中剔除。
- ④ Missing statistics appear 输入框,指定一个在结果中代表缺失值的符号,例如“\*”。

### 4.2.2 输出结果

在图 4-12 中单击“OK”按钮运行,SPSS Viewer 窗口的输出结果如图 4-16 所示,列出了不同教学方案下的所有观测记录,以及它们的汇总信息(包括个数、均值、标准差)。

观测摘要输出 <sup>a</sup>					英语成绩	数学成绩
教学方案	A	1			103	110
		2			82	97
		3			71	116
		4			52	118
		总计	N		4	4
			均值		77.00	110.25
			标准差		21.307	9.465
	B	1			106	102
		2			102	92
		3			100	82
		4			66	71
		总计	N		4	4
			均值		93.50	86.75
			标准差		18.502	13.301
	C	1			118	100
		2			118	106
		3			106	102
		4			66	66
		总计	N		4	4
			均值		106.75	93.50
			标准差		15.564	18.502
	总计	N			12	12
		均值			92.42	96.83
		标准差			21.091	16.508

成绩的观测摘要  
a. 限于前 100 个案例。

图 4-16 观测摘要输出表

## 4.3 行和列的摘要报告分析

行(列)形式的摘要报告分析过程,把数据编辑窗口中的数据重新组织后,按照指定要求罗列在输出窗口,以方便浏览或用于打印,还可以对数据作简单的统计描述。

行形式摘要报告分析过程与上一节的观测摘要过程相比,可给出更为复杂的报告形式,输出格式的设置也更为详细;列形式摘要报告的功能与行形式基本相同,但不能列出原始数据,输出格式也稍有差异。

### 4.3.1 行形式摘要报告

本节继续使用如图 4-1 所示的英语和数学成绩数据，对其做行形式的摘要分析。

依次单击菜单“Analyze→Reports→Report Summaries in Rows”，打开行形式摘要分析的主设置面板，如图 4-17 所示。

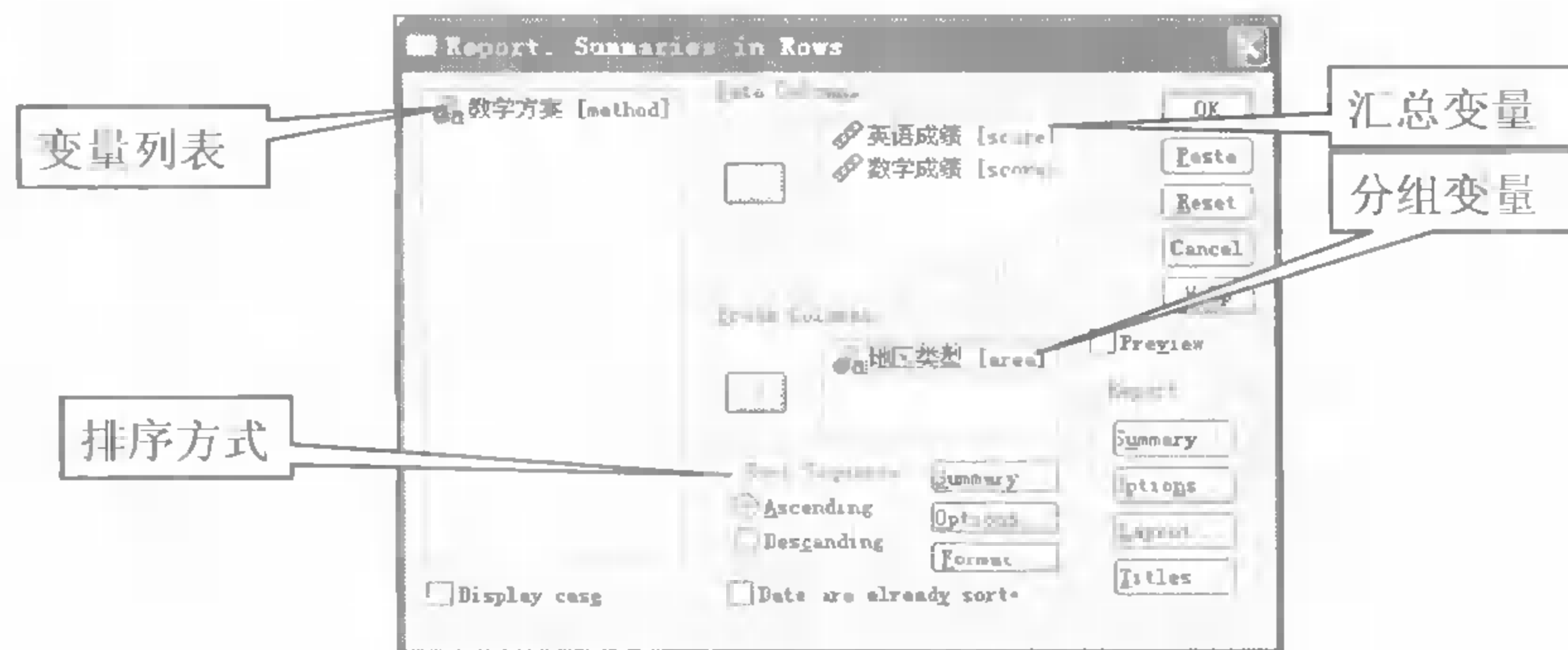


图 4-17 行形式摘要报告主对话框

#### 1. 变量选择设置

首先在变量列表选中英语成绩、数学成绩两个变量，然后单击汇总变量列表左侧的 ☐ 按钮，将其选入汇总变量列表；接着在变量列表选中地区类型，然后单击分组变量列表左侧的 ☐ 按钮，将其选入分组变量列表。

下面详细介绍各设置选项的含义。

- Data Columns 汇总变量列表，一般要求变量为数值型的。
- Break Columns 分组变量列表，用于对汇总变量进行分组统计。
- Display cases 复选框，勾选表示在结果里显示所有的单个记录行。
- Preview 复选框，勾选后将只显示第一页的输出，用来观察实际的输出格式和效果，如果满意，再取消该选项，并对所有数据进行分析。

#### 2. 对指定变量的参数设置

对于选入 Break Columns 栏的分组变量，还可以设置它们的显示顺序和统计参数。

- Sort Sequence 子设置栏，选择分组变量的显示顺序：Ascending（升序）；Descending（降序）。
- Data are already sorted 复选框，使用分组变量进行分析前，如果数据已经按照选入的分组变量值进行了排序，勾选此项可以节省运行时间。
- Break Columns 栏下的 3 个按钮分别用来设置分组变量的不同参数，下面来一一介绍。

(1) Summary 统计量选择。如图 4-17 所示，在分组变量列表选中地区类型变量，然后单击 Break Columns 列表下的 Summary 按钮，弹出如图 4-18 所示的统计量设置面板；勾选 Mean of values、Number of cases、Standard deviation 这 3 个复选框；单击 Continue 按钮返回主设置界面。

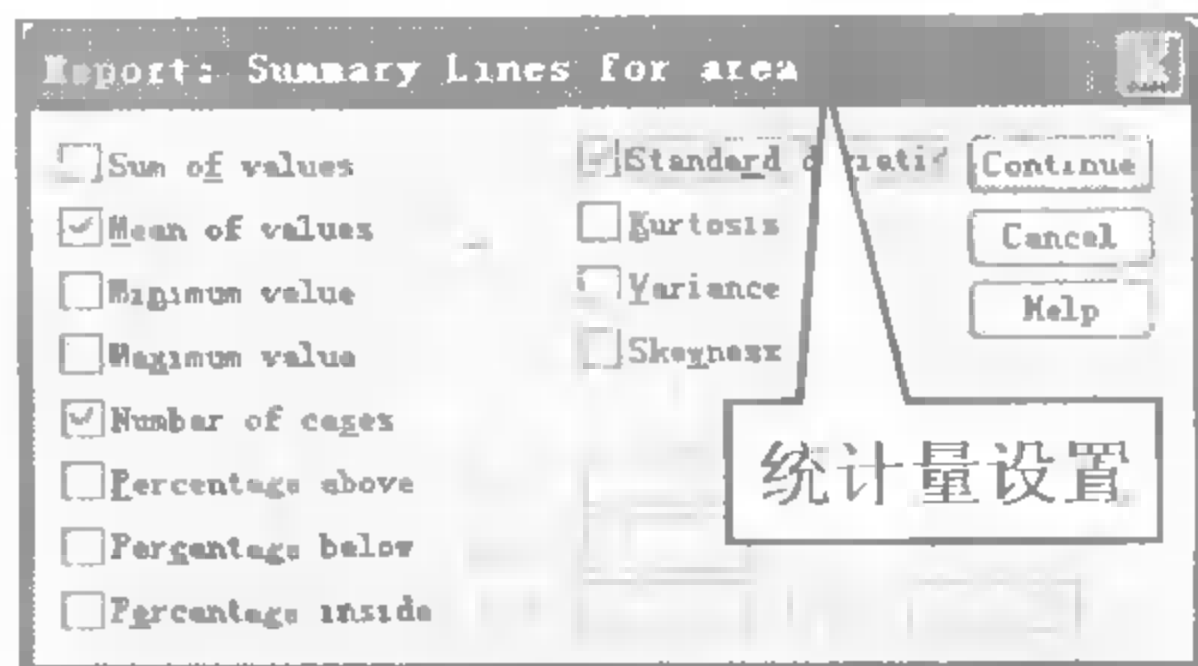


图 4-18 分组变量的统计量设置



可选的统计量包括: sum (求和)、mean (均值)、minimum (最小值)、maximum (最大值)、number of cases (案例个数)、standard deviation (标准差)、kurtosis (峰度)、variance (方差)、skewness (偏度); percentage above (大于指定值的个数比例), 在其后的 value 输入框指定临界值; percentage below (小于指定值的个数比例), 在其后的 value 输入框指定临界值; percentage inside (取值于某个区间的个数比例), 在其后的 low、high 输入框分别指定区间的下限和上限。

(2) Options 选项设置。如图 4-17 所示, 在分组变量列表选中地区类型变量, 然后单击 Break Columns 列表下的 Options 按钮, 弹出如图 4-19 所示的 Options 选项子面板, 在此设置结果显示的分页格式; 单击 Continue 按钮返回主设置界面。

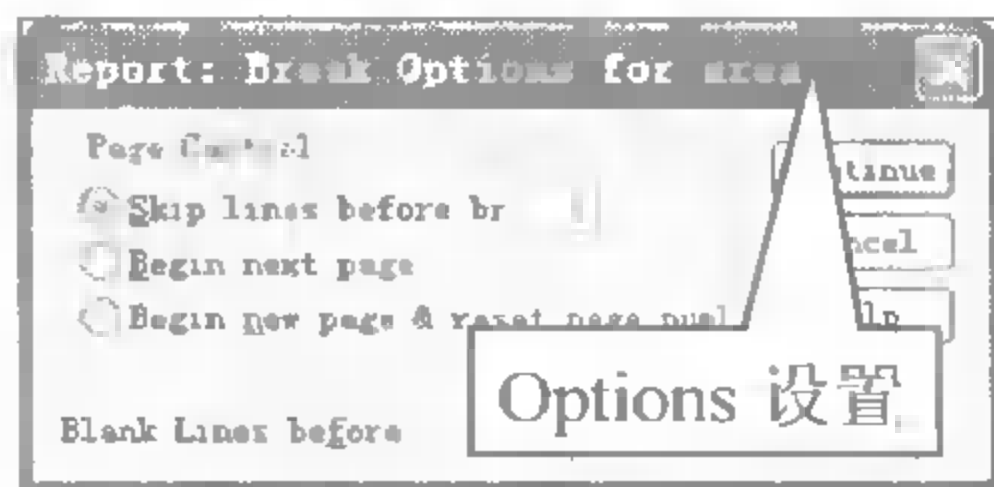


图 4-19 分组变量的 Options 选项设置

① Page Control 栏设置输出页面的格式, 有如下 3 个可选项。

- Skip lines before break 项, 在不同分组类别的输出结果之间插入空行数, 默认值为 1;
- Begin next page 项, 每个分组类别的结果另起一页, 但页码连续;
- Begin new page & reset page number 项, 每个分组类别的结果另起一页, 但页码也会单独标定。

② Blank Lines before 栏, 设置类别标签或数据记录与汇总统计量之间的空行数, 例如: 当同时显示单个记录 and 汇总统计量时, 代表单个记录 and 汇总记录之间的空行数, 默认值为 0 (无空行)。



图 4-20 输出格式设置对话框

(3) Format 格式设置。如图 4-17 所示, 在分组变量列表选中地区类型变量, 然后单击 Break Columns 列表下的 Format 按钮, 弹出如图 4-20 所示的格式设置子面板; 单击 Continue 按钮返回主设置界面。另外, 也可以设置汇总统计量的 Format 格式, 首先在图 4-17 里的汇总变量列表选中一个变量, 然后单击 Data Columns 列表下的 Format 按钮, 也会弹出如图 4-20 所示的格式设置面板。

① Column Title 编辑框, 输入选中变量在结果表格里所对应的标题, 默认为变量标签, 如果没有变量标签就以变量名代替。在下拉菜单选择其对齐方式: Left (左对齐)、Center (居中对齐)、Right (右对齐)。

② Value Position within Column 栏, 设置变量取值或变量标签的对其方式, 有两个选择。

- Offset from left: 左缩进字符数, 在 Offset 输入框指定字符个数; 如果是对汇总变量的格式设置, 此处显示为 Offset from right, 即向右缩进的字符个数;
- Centered within column: 居中显示。

③ Column 输入框, 设定列宽值, 默认以列标题的宽度或最宽的列作为列宽。

④ Column Content 栏, 设置变量取值的输出方式, 可选项有: Values (变量的取值); Value Label (变量的值标签)。

### 3. 对全部数据的统计设置

在图 4-17 的主设置面板里, 右下角的 Report 栏设置对全部数据的统计和输出选项, 它有 4 个按钮, 下面来一一介绍。

(1) Summary 统计量选择。单击 Report 栏中的 Summary 按钮，弹出的界面和 Break Columns 列表下的 Summary 子对话框相同，如图 4-18 所示，参数选项和设置方法也一样。

(2) Options 选项。单击 Report 栏中的 Options 按钮，弹出如图 4-21 所示的对话框，设置对缺失值的处理方式。单击 Continue 按钮返回主设置界面。

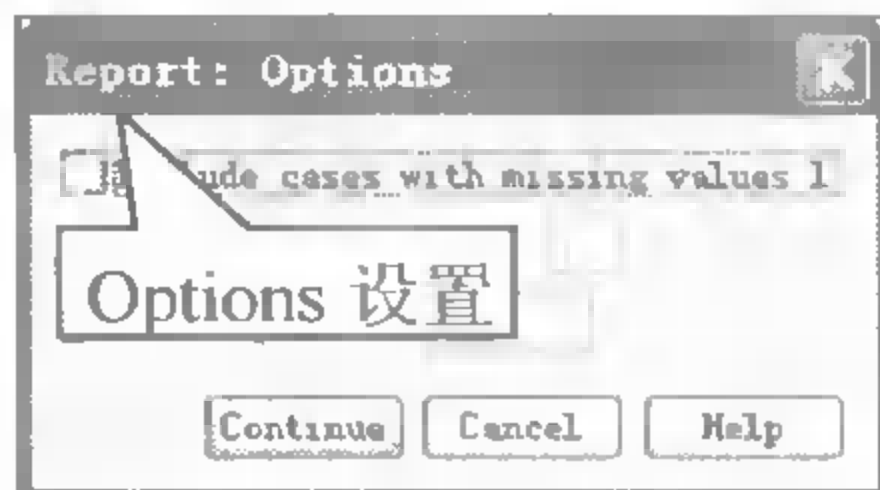


图 4-21 选项设置对话框

- ① Exclude cases with missing values listwise 复选框，勾选表示只要分析中用到的变量有一个取值缺失，就把相应的记录从分析中剔除。
- ② Missing Values Appear as 输入框，指定一个在结果中代表缺失值的符号，例如 “\*”。
- ③ Number Pages From 输入框，指定输出结果的起始页码，默认值为 1。

(3) Layout 布局设置。单击 Report 栏中的 Layout 按钮，弹出如图 4-22 所示的对话框，设置关于输出结果的诸多格式。单击 Continue 按钮返回主设置界面。

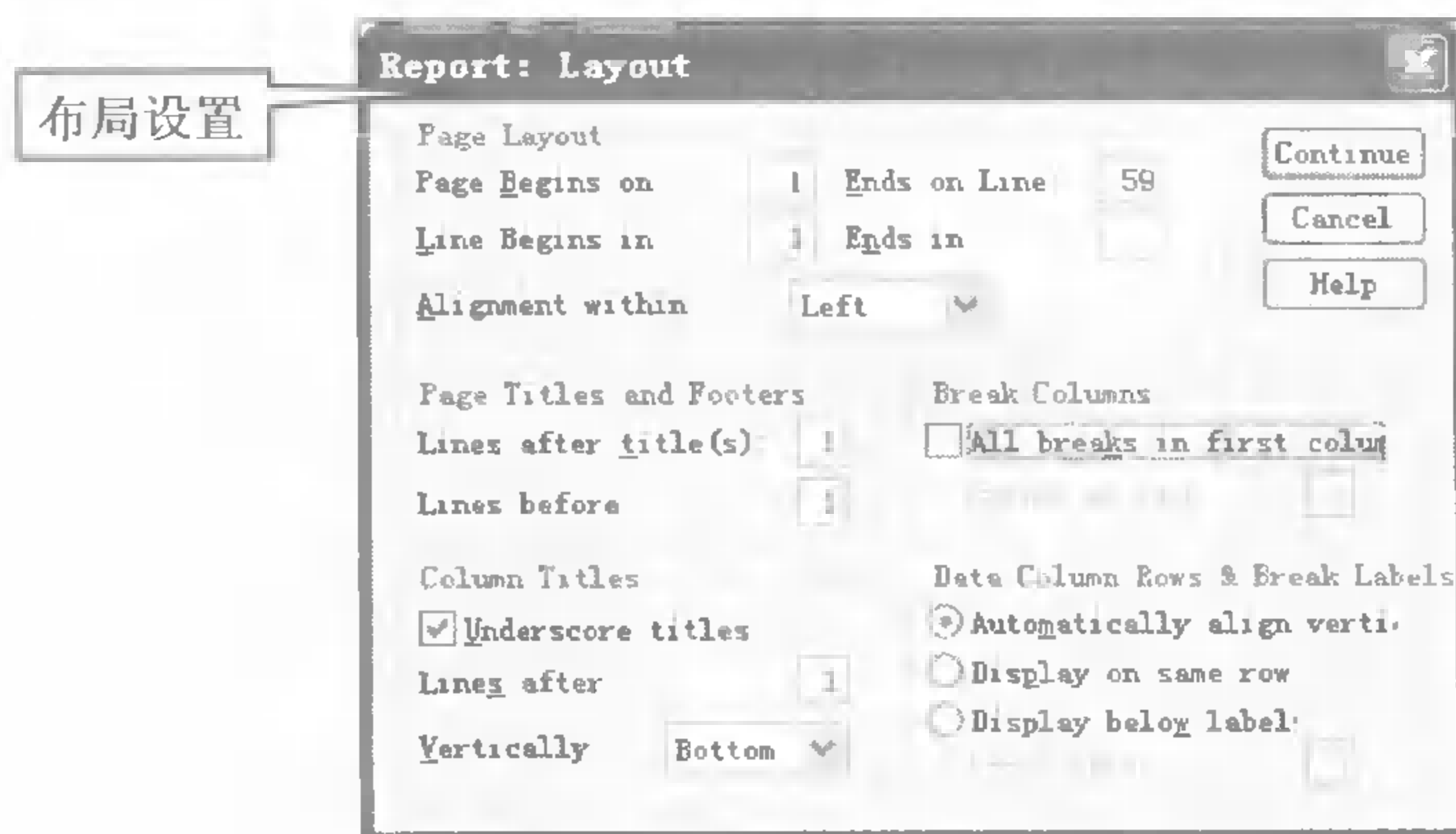


图 4-22 布局设置对话框


- ① Page Layout 栏设置页面布局，可选参数包括如下内容。
  - ① Page Begins on 输入框，指定每页的起始行数（默认为 1）；Ends on Line 输入框，指定每页的结束行数（默认为 59）。
  - ② Line Begins in 输入框，指定每行起始字符的位置（默认为 1）；Ends in 输入框，指定每行终止字符的位置（默认为空）。
  - ③ Alignment within 下拉菜单，设置页面的对齐方式为：Left（左对齐）、Center（居中对齐）或 Right（右对齐）。
- ② Page Titles and Footers 栏设定页面标题、页脚与报告正文之间的距离，有两个选项。
  - ① Lines after titles (s): 指定在标题与报告正文之间插入空行数，默认为 1。
  - ② Lines before: 指定在报告正文与页脚之间插入空行数，默认为 1。
- ③ Column Titles 栏设置列标题的输出格式，有 3 个选项。
  - ① Underscore titles 复选框，表示在每个列标题下加下划线。

- Lines after 栏, 指定在列标题与第 1 行数据之间插入空行数, 默认为 1。
- Vertically 下拉菜单, 设置列标题的垂直对齐方式: Top (顶端对齐) 或 Bottom (底端对齐)。

④ Break Columns 栏设置分组变量的输出位置。选中 All breaks in first column 复选框, 表示所有分组变量都在第 1 列给出, 因而产生较窄的表格, 同时在 Indent at each 输入框指定每一级分组向右缩进字符数, 默认为两个空格; 否则, 每个分组变量各占一列。

⑤ Data Column Rows & Break Labels 栏, 设置各分组内容与分组标签之间的相对位置, 有如下 3 个选择。

- Automatically align vertical 选项, 分组数据从分组变量取值所在的同一行开始列出, 而基本描述统计量在分组变量值的下一行列出。
- Display on same row 选项, 二者都从同一行开始显示。
- Display below label 选项, 二者都显示在分组标签值的下方, 同时在 Lines after labels 输入框指定中间隔空行数。

(4) Titles 标题设置。单击 Report 栏中的 Titles 按钮, 弹出如图 4-23 所示的标题设置面板。在 Special 列表选中 DATE, 单击 Title 栏 Center 左侧的  按钮, 将 “) DATE” 自动填入 Center 输入框; 在 Title 栏的 Right 后输入 “) PAGE”; 在 Footer 栏的 Center 后输入 “行摘要分析脚注”; 单击 Continue 按钮返回主设置界面。

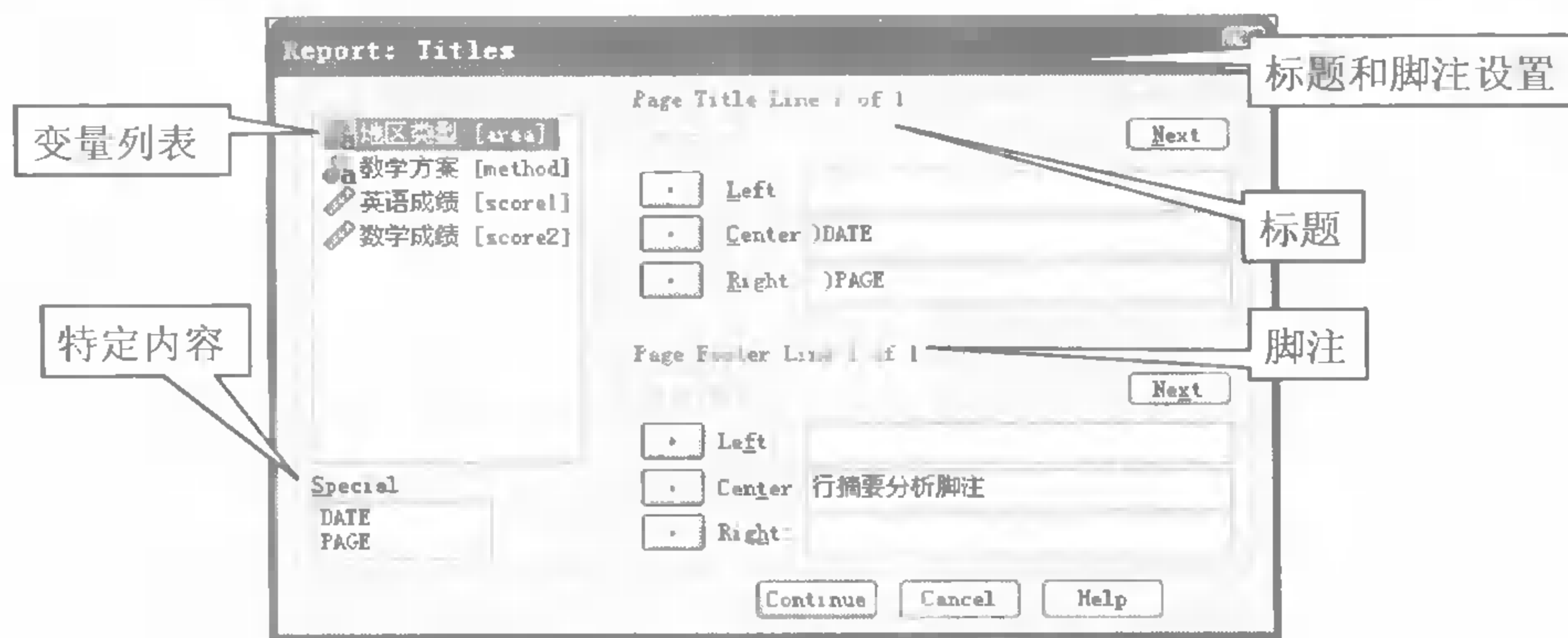


图 4-23 定义标题和页脚对话框

#### ① Page Title Line 栏设置页标题。

可以分别输入标题 Left (左边)、Center (中间) 和 Right (右边) 的显示内容; 最多能够指定 10 行的标题, 方法是输入一组 Left、Center 和 Right 值后, 单击上面的 Next 按钮, 输入框清空, 再输入下一行的内容; 通过单击 Previous (前一行)、Next (后一行) 两个按钮, 可以调节显示和修改每一行的内容。

#### ② Page Footer Line 栏设置页脚注, 设置方法和页标题的设置相仿。

③ 变量列表里列出了当前数据集里的变量, Special 列表给出了两个系统变量 DATE (日期)、PAGE (页码), 选中其中某项后单击黑色箭头就能把它选入对应的标题或脚注里。

### 4. 输出结果

在图 4-17 里单击 OK 按钮运行, SPSS Viewer 窗口的输出报告如图 4-24 所示。

地区类型	英语成绩	数学成绩
北京		
Mean	109	104
N	3	3
StdDev	8	6
河北		
Mean	68	85
N	3	3
StdDev	17	29
上海		
Mean	92	100
N	3	3
StdDev	19	17
天津		
Mean	101	98
N	3	3
StdDev	18	7
Grand Total		
Mean	92	97
N	12	12
StdDev	21	17
...		
行摘要分析脚注		

图 4-24 行形式摘要报告分析的结果

结果对不同地区的两个成绩数据进行了汇总，显示了其均值 (Mean)、个数 (N)、StdDev (标准差)；结果中还标出了在图 4-23 中所设置的标题和脚注内容；另外，图中的省略号并非输出内容，而是由作者编辑添加的，它代表了多个空行。

### 4.3.2 列形式摘要报告

本节继续使用如图 4-1 所示的英语和数学成绩数据，对其做列形式的摘要分析。

列摘要分析过程的操作与行摘要分析类似，本节重点介绍列摘要分析所独有的特点。

依次单击菜单“Analyze→Reports→Report Summaries in Columns”，打开列形式摘要分析的主设置面板，如图 4-25 所示。

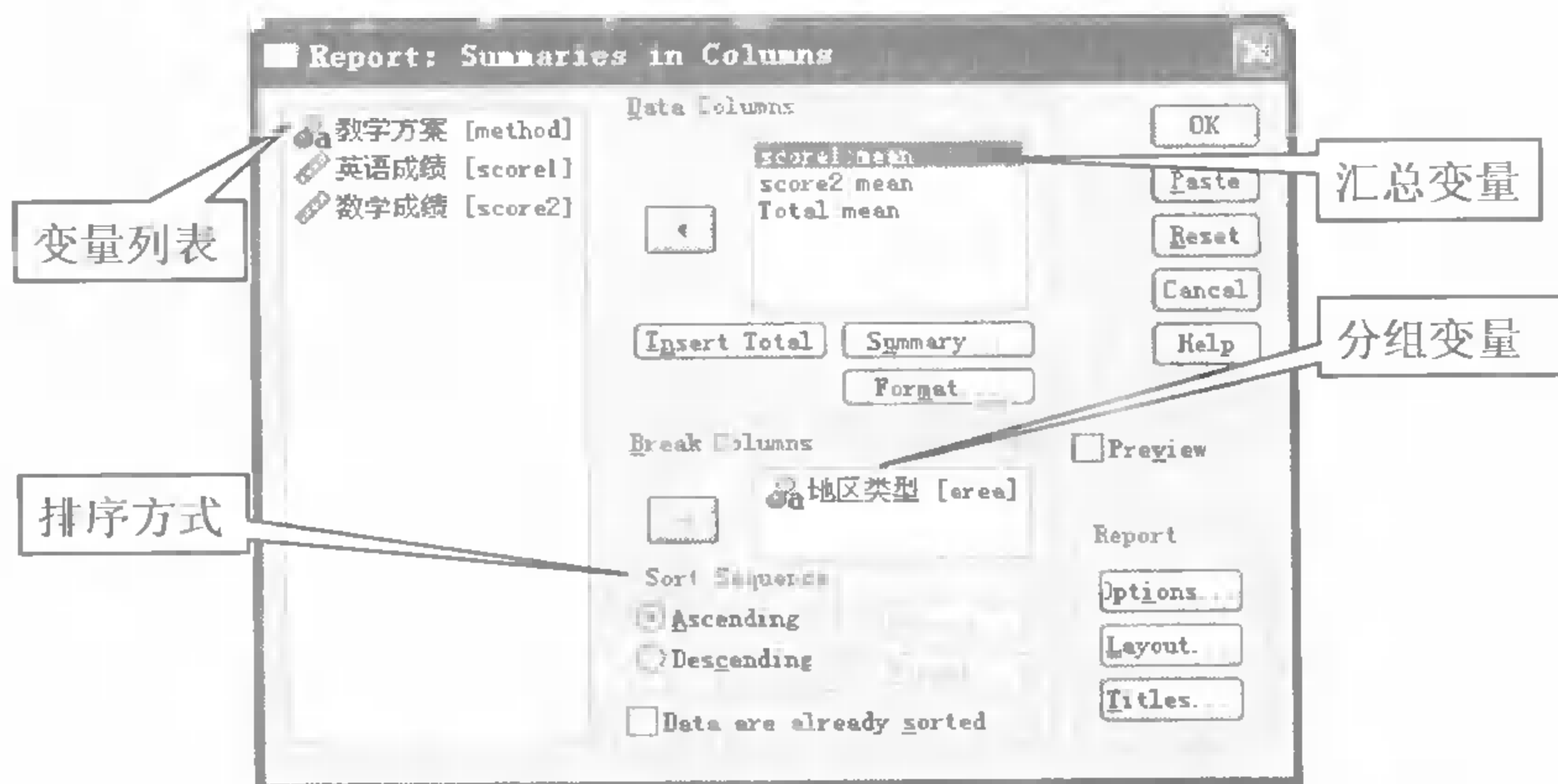


图 4-25 列形式摘要报告主对话框

#### 1. 变量选择设置


首先在变量列表选中英语成绩、数学成绩两个变量，然后单击汇总变量列表左侧的  按钮，



将其选入 Data Columns 列表；单击 Insert Total 按钮，将 Total 变量加入 Data Columns 列表；在变量列表选中地区类型，然后单击分组变量列表左侧的  按钮，将其选入 Break Columns 列表。

- Data Columns 为汇总变量列表；Break Columns 为分组变量列表。
- Insert Total 按钮，单击它会把一个名为 Total 的变量加入 Data Columns 列表，在结果中以列的形式对其它列的数据进行汇总。
- Preview 复选框，勾选后将只显示第一页的输出，用来观察实际的输出格式和效果，如果满意，再取消该选项，并对所有数据进行分析。

## 2. 汇总变量的参数设置

如图 4-25 所示，在汇总变量列表选中 score1（英语成绩），单击其下的 Summary 按钮，弹出如图 4-26 所示的对话框，单击选中 Mean of values 选项，单击 Continue 按钮返回主设置界面；用同样的方法设置 score2（数学成绩）的统计量为 Mean。在汇总变量列表选中 Total 变量，单击其下的 Summary 按钮，弹出如图 4-27 所示的对话框，在 Data Columns 列表选中 score1、score2，单击  按钮将其选入 Summary Column 列表；在 Summary function 下拉列表里选中 Means of columns 项，如图 4-28 所示，单击 Continue 按钮返回主设置界面。

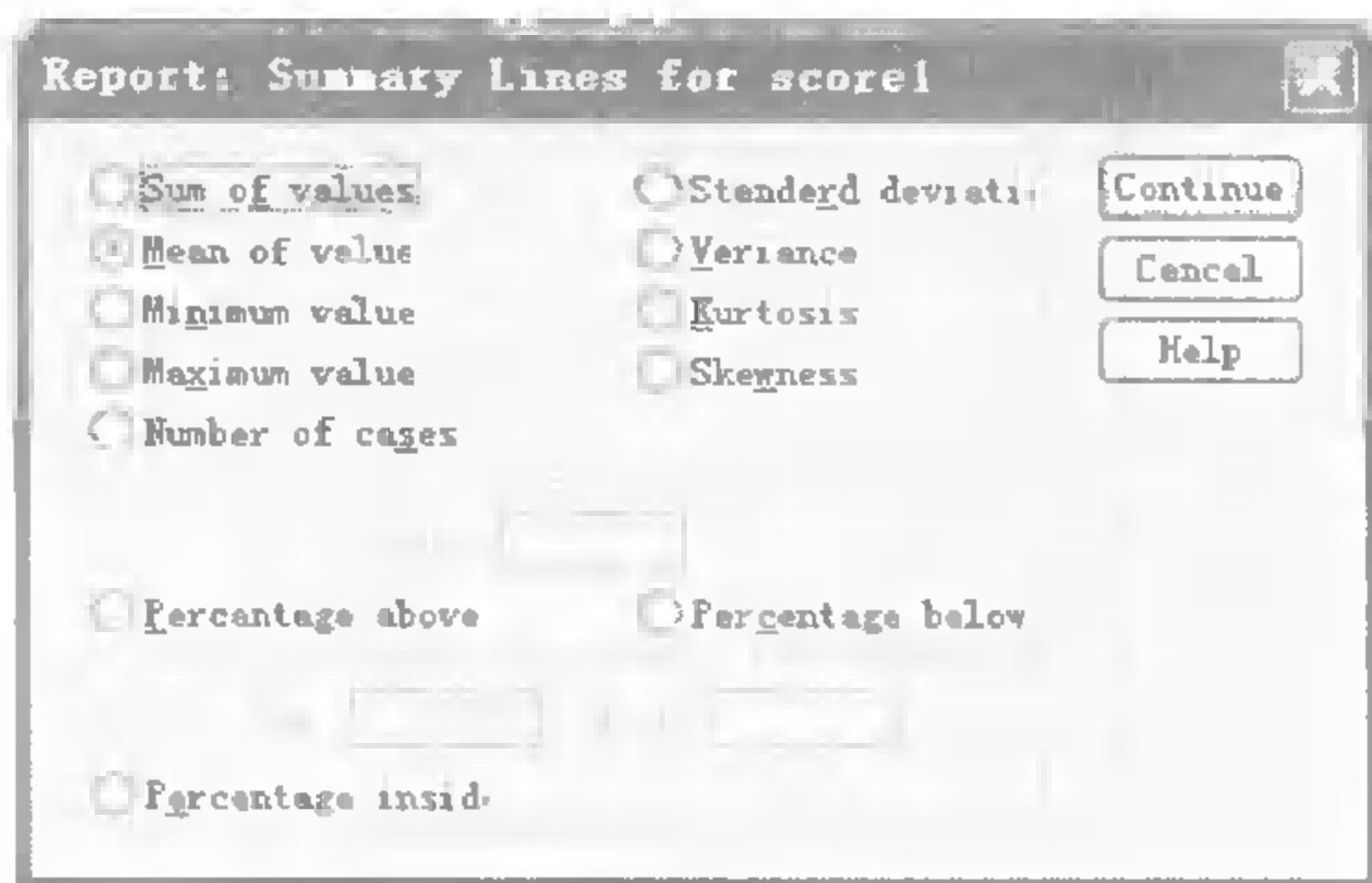


图 4-26 普通变量的 Summary 设置

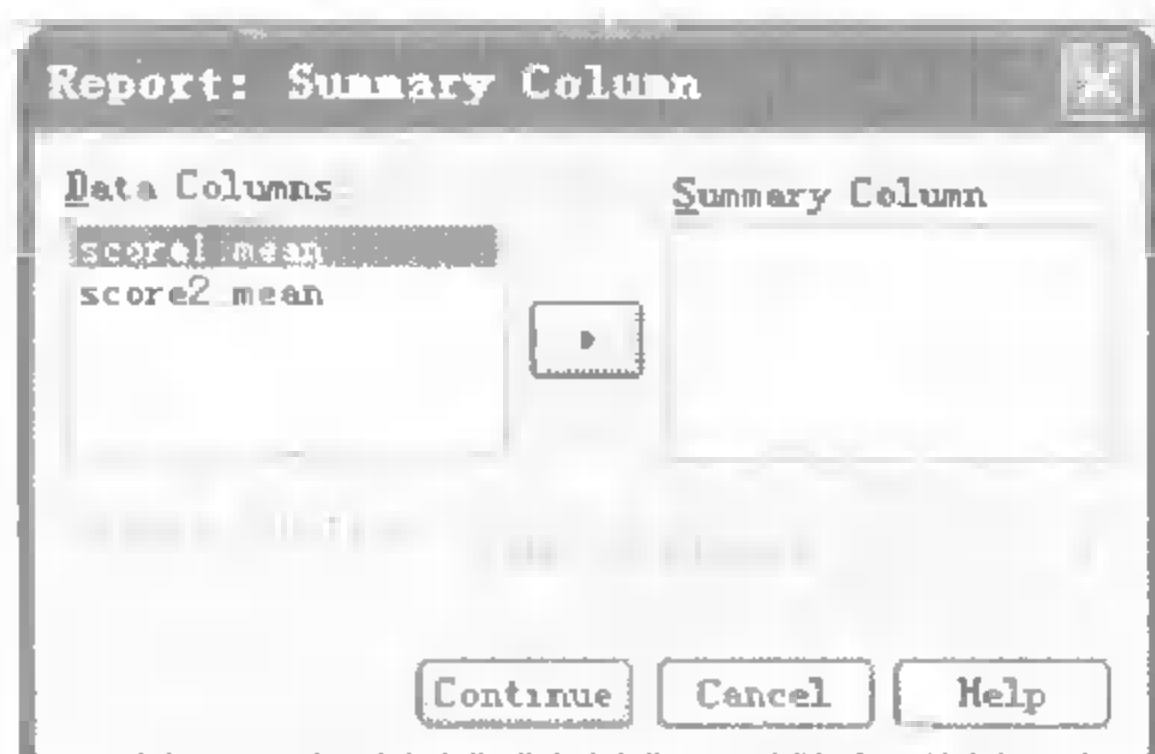


图 4-27 Total 变量的 Summary 设置 1

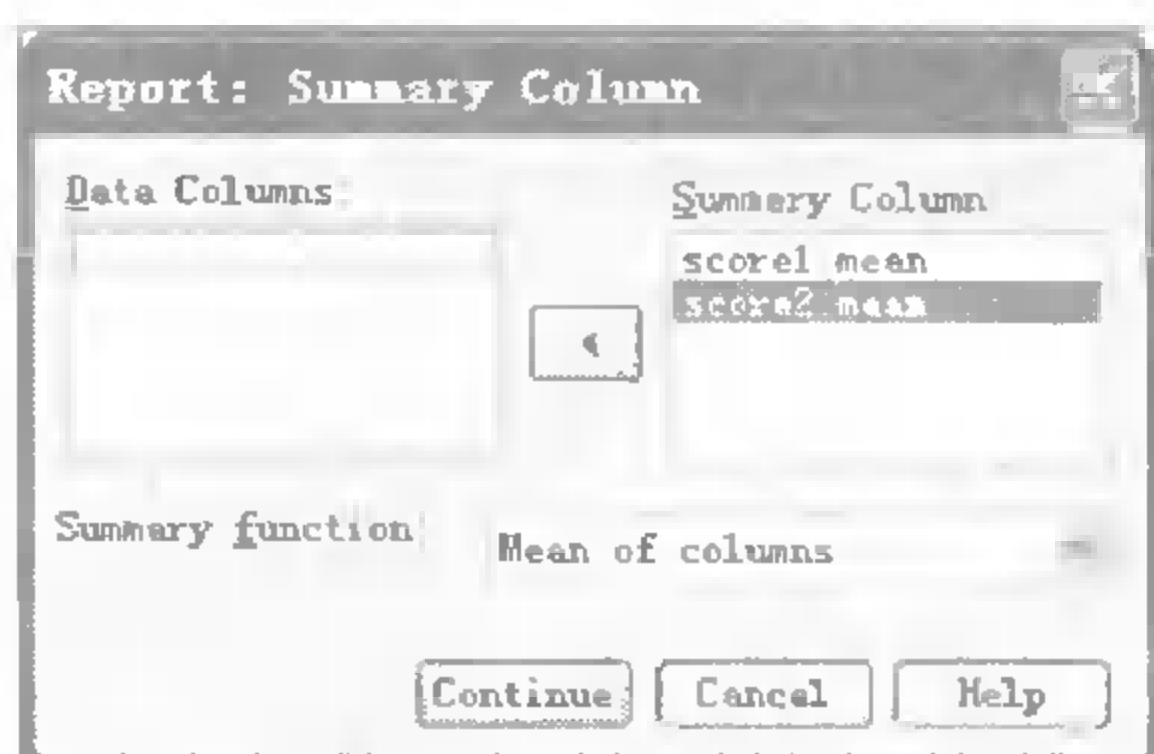


图 4-28 Total 变量的 Summary 设置 2

图 4-26 所示的统计量选择对话框，与图 4-18 所示的统计量设置面板完全相同，只是设置内容都变为了单选项。

在图 4-27 中，Data Columns 列表显示了当前可以输出的列；Summary Column 列表显示的是要进行汇总的列；Summary function 下拉列表，用来指定汇总函数。

返回主界面图 4-25 后，Data Columns 列表的变量名后会显示 mean 字样，表明当前统计的是变量均值，默认统计量为 sum（求和）。

## 3. 对分类变量的参数设置

对于选入 Break Columns 栏的分组变量，还可以设置它们的显示顺序和统计参数。

(1) Sort Sequence 子设置栏，选择分组变量的显示顺序：Ascending（升序）；Descending（降序）。

(2) Data are already sorted 复选框，使用分组变量进行分析前，如果数据已经按照选入的分组变量值进行了排序，选中此项可以节省运行时间。



(3) Break Columns 栏下的 2 个按钮分别用来设置分组变量的不同参数，下面来一一介绍。

- Options 选项设置。在图 4-25 所示的分组变量列表中选中地区类型变量，然后单击 Break Columns 列表下的 Options 按钮，弹出如图 4-29 所示的 Options 选项子面板，单击 Continue 按钮返回主设置界面。

此面板和图 4-19 所示的 Options 选项界面相似，只是多了一个 Subtotal 子设置栏，选中 Display 复选框表示显示分类变量不同类别的子标题，在 Label 下输入子标题的标签。

- Format 格式设置。在图 4-25 所示的分组变量列表中选中地区类型变量，然后单击 Break Columns 列表下的 Format 按钮，弹出如图 4-30 所示的格式设置子面板，单击 Continue 按钮返回主设置界面。

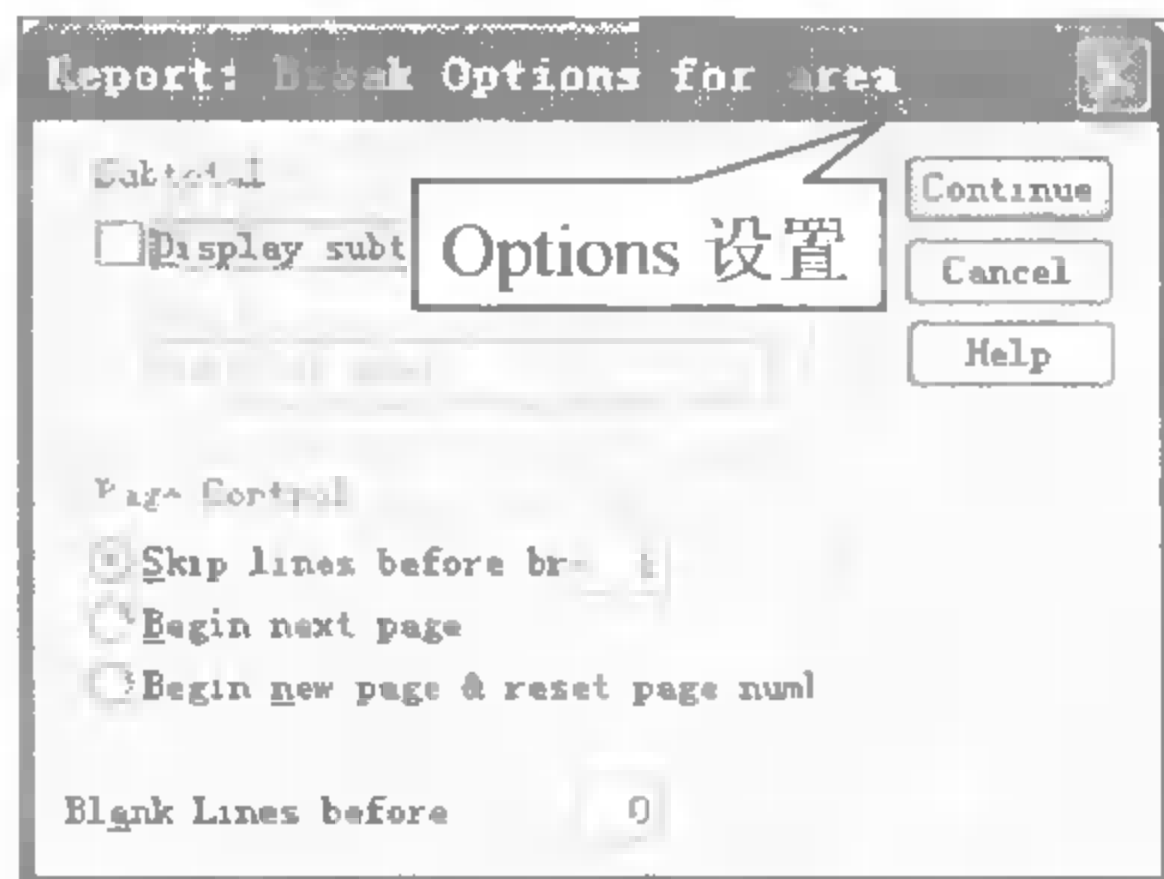


图 4-29 列形式摘要的选项设置

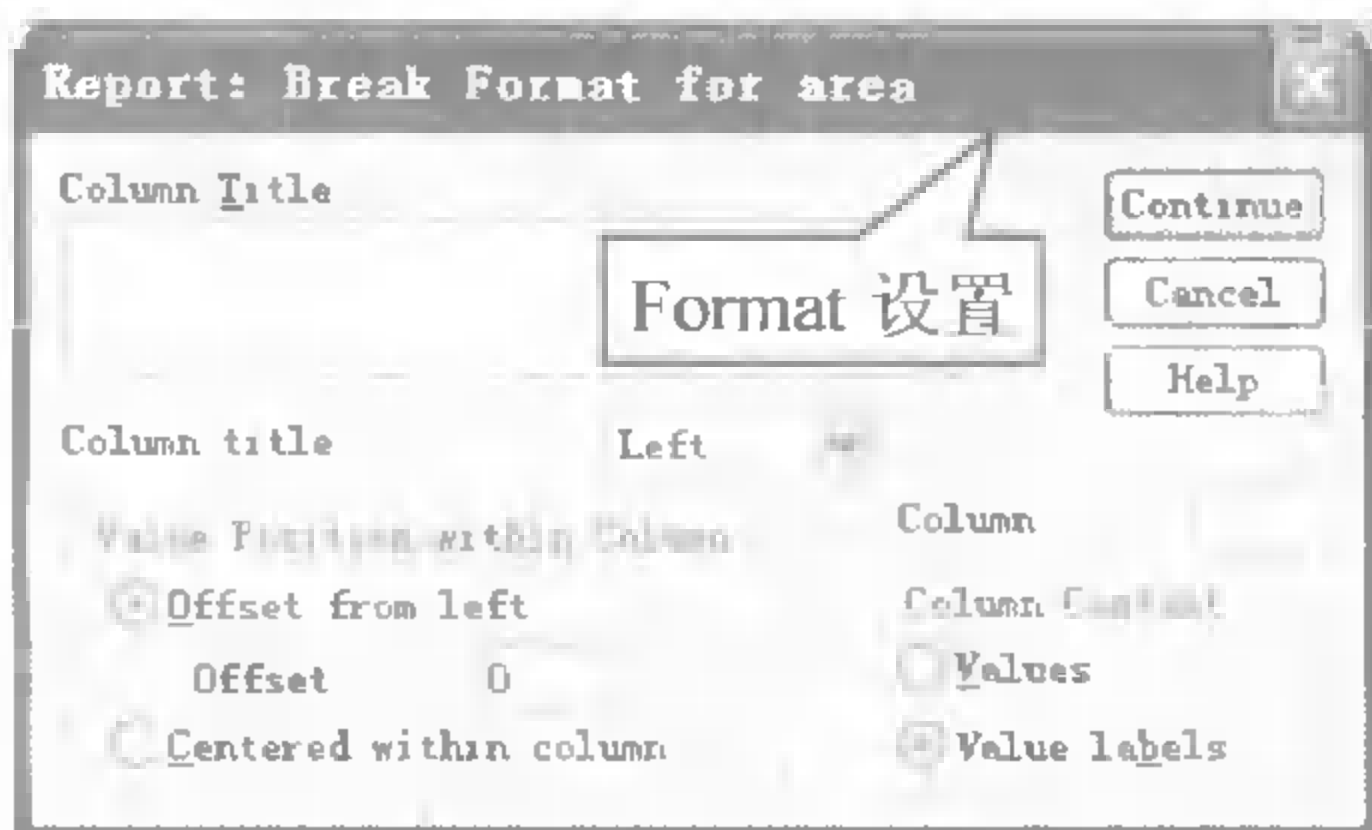


图 4-30 列形式摘要的格式设置

此面板和图 4-20 所示的格式设置界面和设置方法相同。

#### 4. 对全部数据的统计设置

双击图 4-25 所示的 Report 栏中的 Options 按钮，弹出如图 4-31 所示的对话框，勾选 Display 复选框。单击 Continue 按钮返回主设置界面。

Display grand total 复选框表示在输出结果的最后，增加对所有行进行汇总的新行；Label 输入框，指定这个汇总行的行标签。其他选项的含义和设置方法与做行摘要分析时相同。

另外，图 4-25 中右下角的 Report 栏还有其他 2 个按钮：Layout 和 Titles，单击它们弹出的设置界面与做行摘要分析时相同，设置方法也相同。

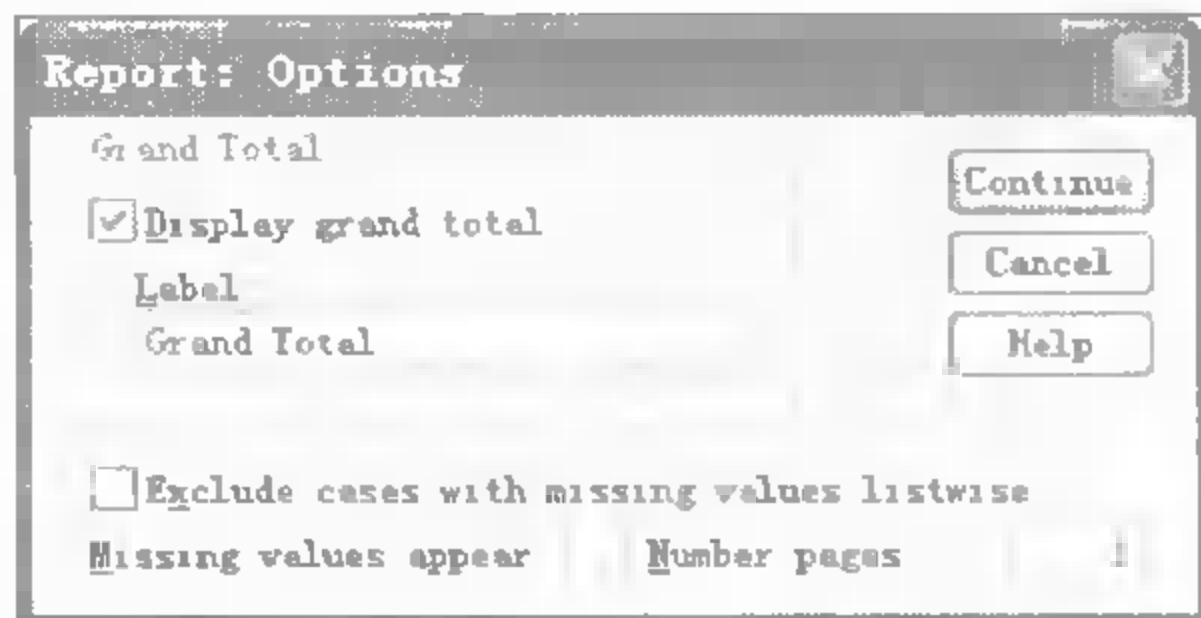


图 4-31 Report 栏中的 Options 设置

#### 5. 结果输出

单击图 4-25 中的 OK 按钮，则 SPSS Viewer 窗口输出如图 4-32 所示的报告。

Page 1			
地区类型	英语成绩 Mean	数学成绩 Mean	Total
北京	109	104	107
河北	88	85	76
上海	92	100	96
天津	101	98	100
Grand Total	95	97	96

图 4-32 摘要报告对话框及结果显示

列 Total 是新加入的汇总列，是同行数据的平均值；行 Total 是新加入的行汇总，是同列数据的平均值。

4.4 频数分析

频数分布法是描述性统计分析的常用方法之一，SPSS 的 Frequencies 过程不仅能够输出详细的频数分布表，而且能够按照用户的要求输出特定的百分位点，和常用的条形图等统计图形。

Frequencies 过程能够处理多种类型的变量，而且对分类变量和连续变量的处理方式也不同。对于分类变量，建议用数值或短字符对取值进行编码。

4.4.1 数据描述

通过调研，我们获得了一些城市和农村学生的心理素质测试得分，数据文件为“心理测试数据.sav”，数据格式如图 4-33 所示。本节通过频数分析，分析所调研的城市、农村学生的个数及其各自得分的分布情况。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	group	Numeric	8	0	组别	{1, 城市学生}	None	8	Right	Scale
2	score	Numeric	8	2	心理测试得分	None	None	8	Right	Scale

图 4-33 心理测试得分的数据格式

4.4.2 对分类变量的频数分析

依次单击菜单“Analyze→Descriptive statistics→Frequencies”，打开频数分析的主设置面板，如图 4-34 所示，在此选择分析变量。

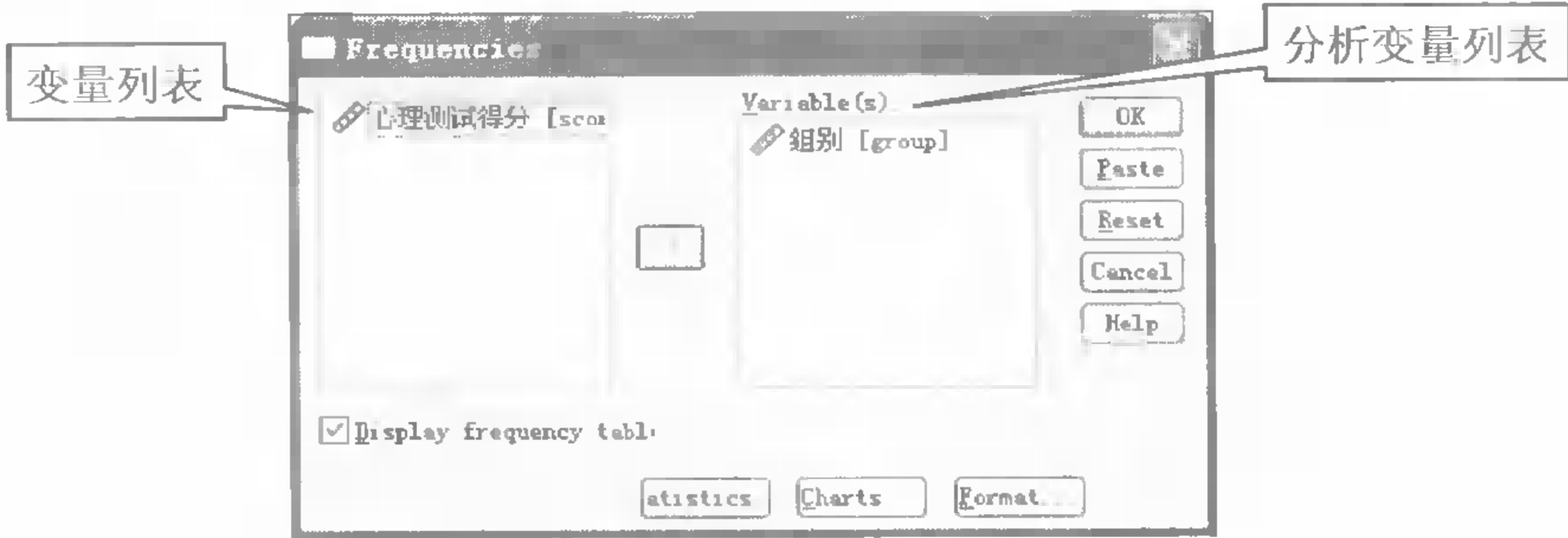


图 4-34 频数分析的主设置界面

1. 变量选择设置

- 在变量列表选中组别（group）变量，然后单击 按钮，将其选入 Variable（s）列表。
- Variable（s）列表，用于从左侧的变量列表选入待分析的变量。
- Display frequency table（s）复选框，表示输出每个分析变量的频数分布表，默认为选中。

2. 统计量选择

在图 4-34 中，单击 Statistics 按钮，弹出如图 4-35 所示的统计量设置界面，在此选择要在结果中显示的统计信息。单击 Continue 按钮返回主设置界面。

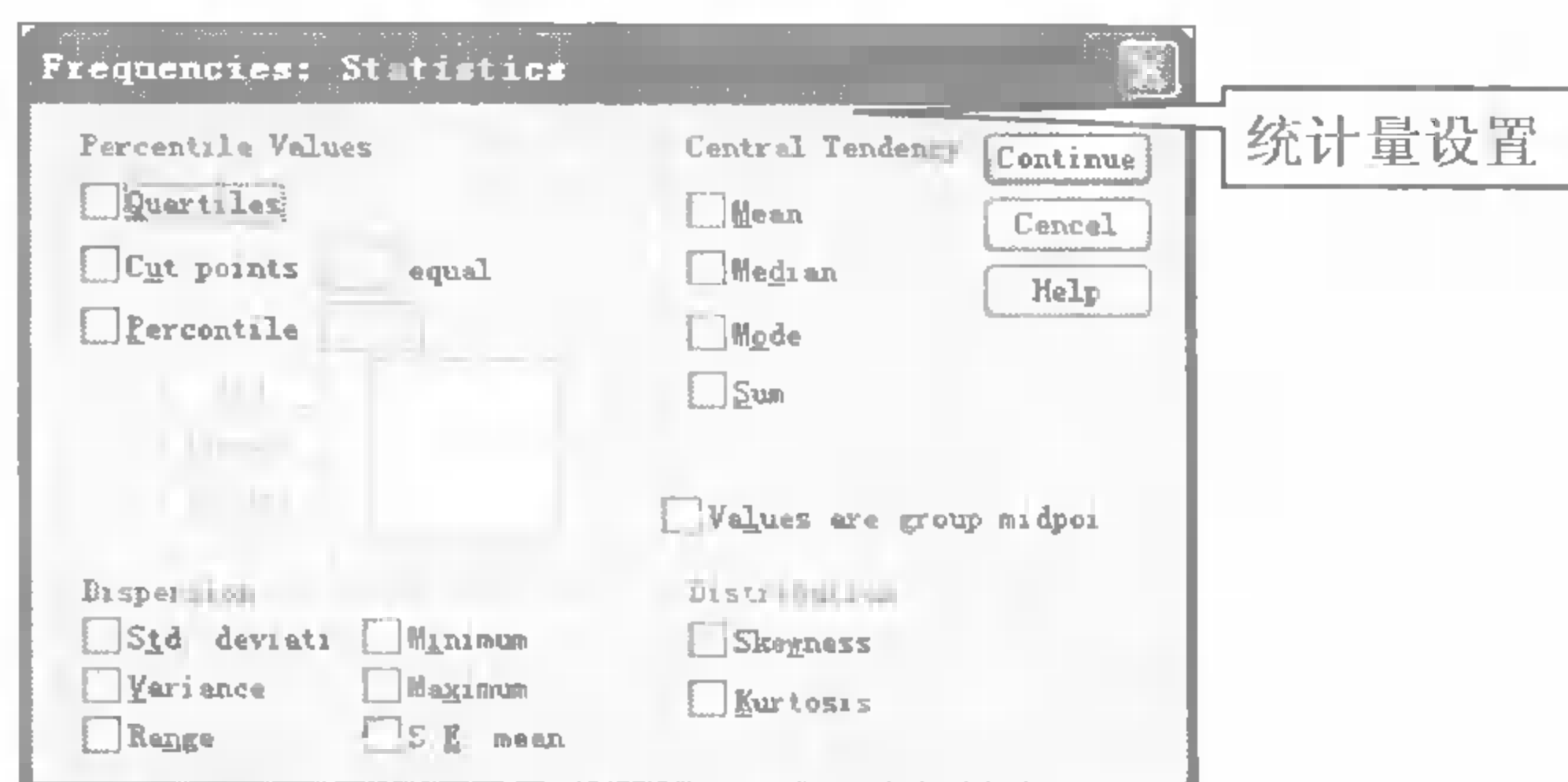


图 4-35 统计量选择

(1) Percentile Values 栏选择要输出的百分位数，有以下 3 个可选项。

- Quartiles 复选框，四分位数。
- Cut points for equal groups 复选框，等间隔的百分位数，在其后的输入框指定 2~100 的间隔值。
- Percentile 复选框，指定百分位点，在其后框输入特定的百分位数后，单击 Add 按钮加入下面的列表，如此重复可以加入多个百分位数；对于列表的数值，选中它后还可以通过单击 Change、Remove 按钮加以修改或删除。

(2) Dispersion 栏设置度量数据离散程度的统计量。可选项有：Std. Deviation (标准差)、Variance (方差)、Range (全距)、Minimum (最小值)、Maximum (最大值)、S.E.mean (均数的标准误)。

(3) Central Tendency 栏设置度量数据集中趋势的统计量。可选项有：Mean (平均数)、Median (中位数)、Mode (众数)、Sum (总和)。

(4) Distribution 栏设置度量数据分布形式的统计量。有 2 个选择：Skewness (偏度)、Kurtosis (峰度)。如果 Skewness 与 Kurtosis 的值都接近 0，测量数据的分布接近正态分布；如果 Skewness 的值为正数，数据倾向于右偏分布；如果 Kurtosis 的数值为正，数据的分布比标准正态分布具有更尖锐的峰型。

(5) Values are group midpoints 复选项。当原始数据记录的是分组频数的数据，并且具体取值是组中值时（例如所有 30~40 岁的人的年龄都记录为了 35），勾选该复选框，SPSS 会自行估计数据的中位数和百分位数。

### 3. 输出图形设置

在图 4-34 中，单击 Charts 按钮，弹出如图 4-36 所示的作图设置界面，在此选择输出哪些图形；在 Chart type 栏单击选中 Pie Chart 单选框，再在 Chart Value 栏单击选中 Percentage 单选框；单击 Continue 按钮返回主设置界面。

(1) Chart type 栏选择图形类型。

有 4 个选择：None (不输出图形)；Bar (条形图)；Pie (饼图)；Histogram (直方图)，仅适用于数值型变量；勾选 With normal curve 复选框表示在输出图形里包含正态曲线。

(2) Chart Value 栏表示在图形中显示何种类型的值。

只对条形图和饼图有效，有 2 个选择：Frequencies (频数)、Percentage (百分比)。

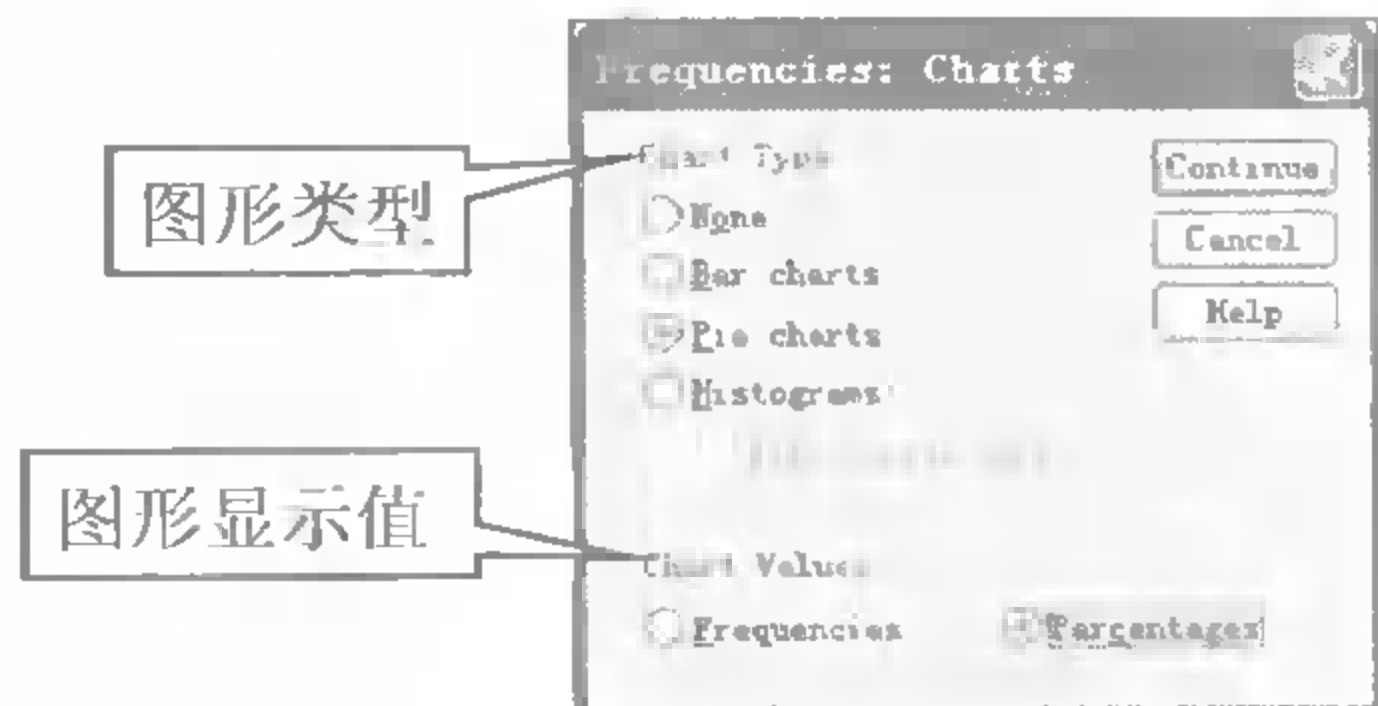


图 4-36 作图设置子面板

#### 4. 输出格式设置

在图 4-34 中，单击 Format 按钮，弹出如图 4-37 所示的格式设置界面；单击 Continue 按钮返回主设置界面。

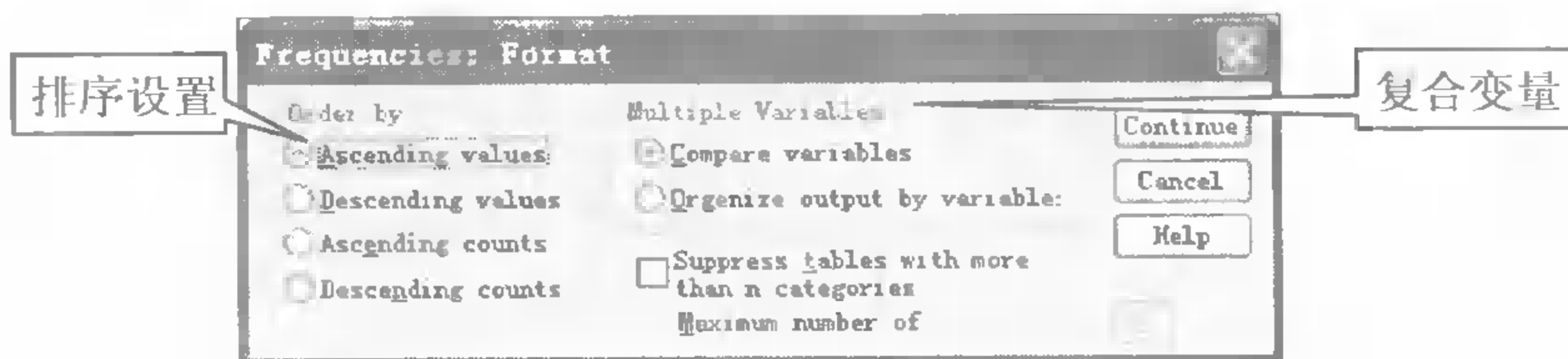


图 4-37 格式设置子面板

下面详细介绍各设置选项的含义。

(1) Order by 栏设置输出表格内容的排序方式，有 4 个选择。

- Ascending Values，按变量值的升序排列，系统默认方式；
- Descending Values，按变量值的降序排列；
- Ascending counts，按频数的升序排列；
- Descending counts，按频数的降序排列。

(2) Multiple variables 栏设置复合变量的输出方式，有 2 个选择。

- Compare variables，将所有变量在一个表格里输出，便于比较；
- Organize output by variables，为每个变量单独输出一个表格。

(3) Suppress tables with more than categories 复选项，控制表格大小。

选中后，在 Maximum number of 后的输入框指定最大能显示的分组个数，当频数表的分组个数大于此临界值时不做输出，避免产生过大的表格。

#### 5. 输出结果

在图 4-34 里单击 OK 按钮运行，SPSS Viewer 窗口的输出如图 4-38、图 4-39 所示。

统计量		
组别		
N	有效	30
	缺失	0

组别					
		频率	百分比	有效百分比	累积百分比
有效	城市学生	16	53.3	53.3	53.3
	农村学生	14	46.7	46.7	100.0
	合计	30	100.0	100.0	

图 4-38 频数分析表

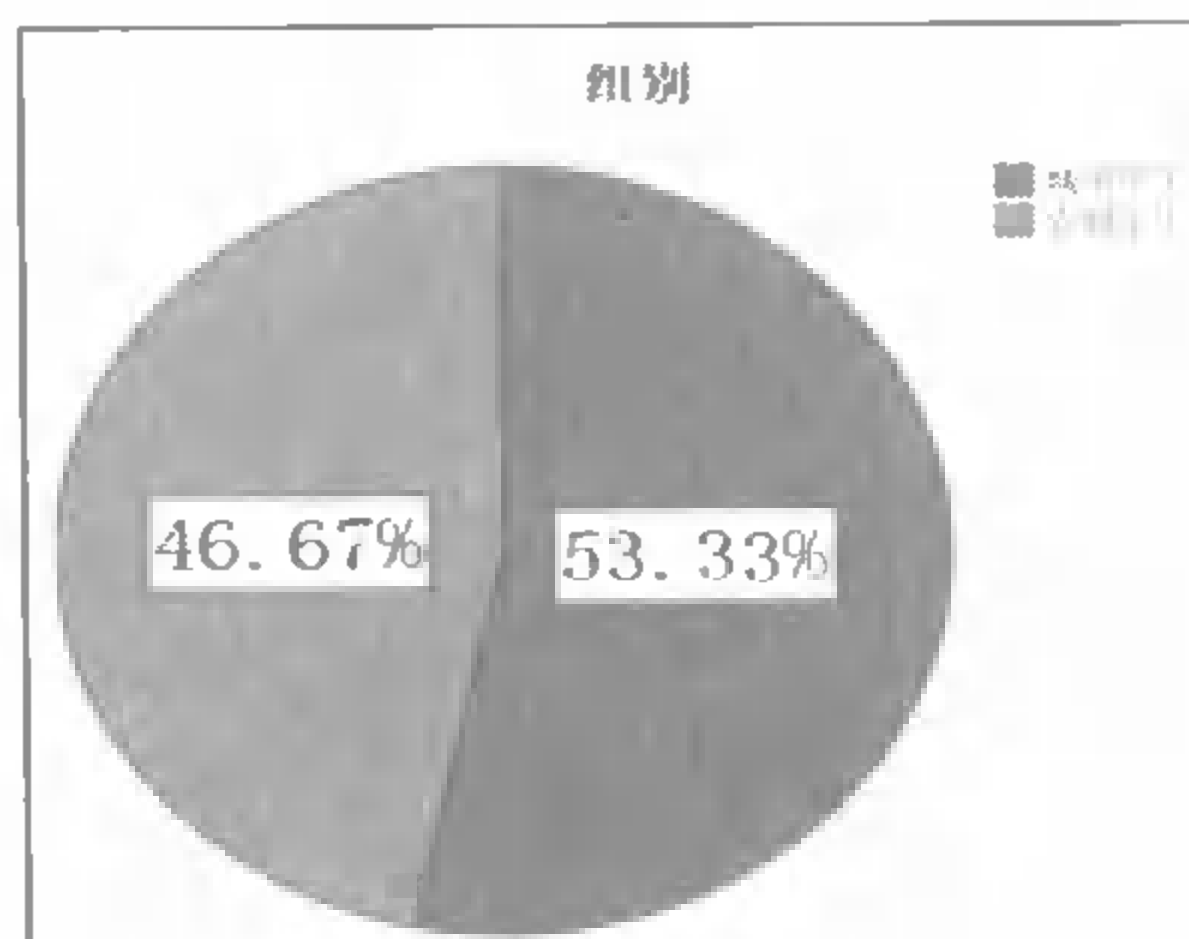


图 4-39 频数分析饼图

“组别”表格给出了城市、农村学生在调研中所占的百分比，分别为 53.3%、46.67%；“组别”饼图则以图形方式直观地显示了这个信息。

#### 4.4.3 对连续变量的频数分析

依次单击菜单“Analyze→Descriptive statistics→Frequencies”，打开频数分析的主设置面板，如图 4-40 所示。

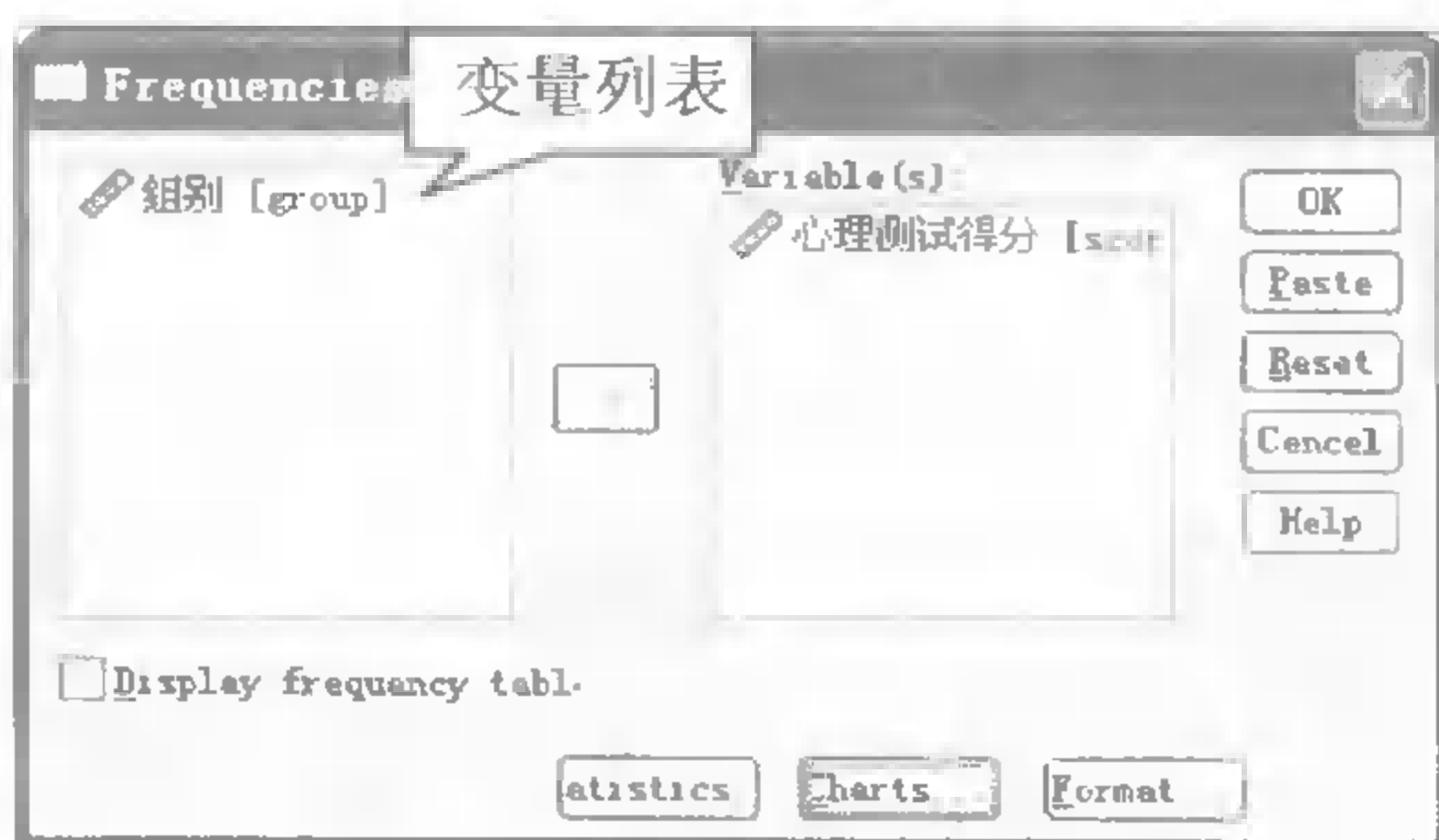


图 4-40 频数分析主界面

## 1. 参数设置

在变量列表选中心理测试得分 (score) 变量, 然后单击 按钮, 将其选入 Variable (s) 列表; 勾选 Display 复选框取消选中。单击 Charts 按钮, 弹出图 4-41 所示的作图设置界面, 勾选 Histogram 单选框和 With normal curve 复选框。

单击 Continue 按钮返回主设置界面。单击 Statistics 按钮, 弹出图 4-42 所示的统计量设置界面, 依次单击选中 Quartiles (四分位数)、Std (标准差)、Mean (均值)、Median (中位数)、Skewness (偏度)、Kurtosis (峰度); 单击 Continue 按钮返回主设置界面。

各界面的参数选项和设置方法, 与上一节中对分类变量做频数分析时相同。

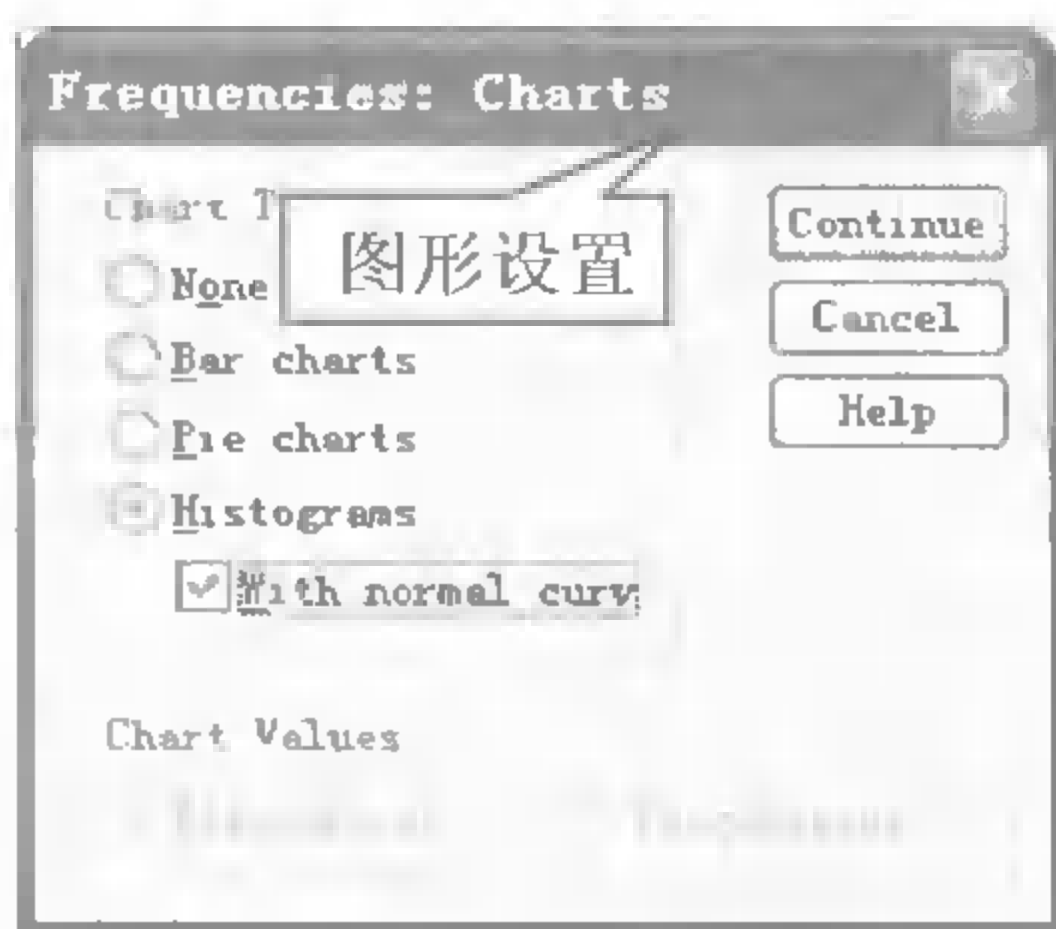


图 4-41 图形设置子界面

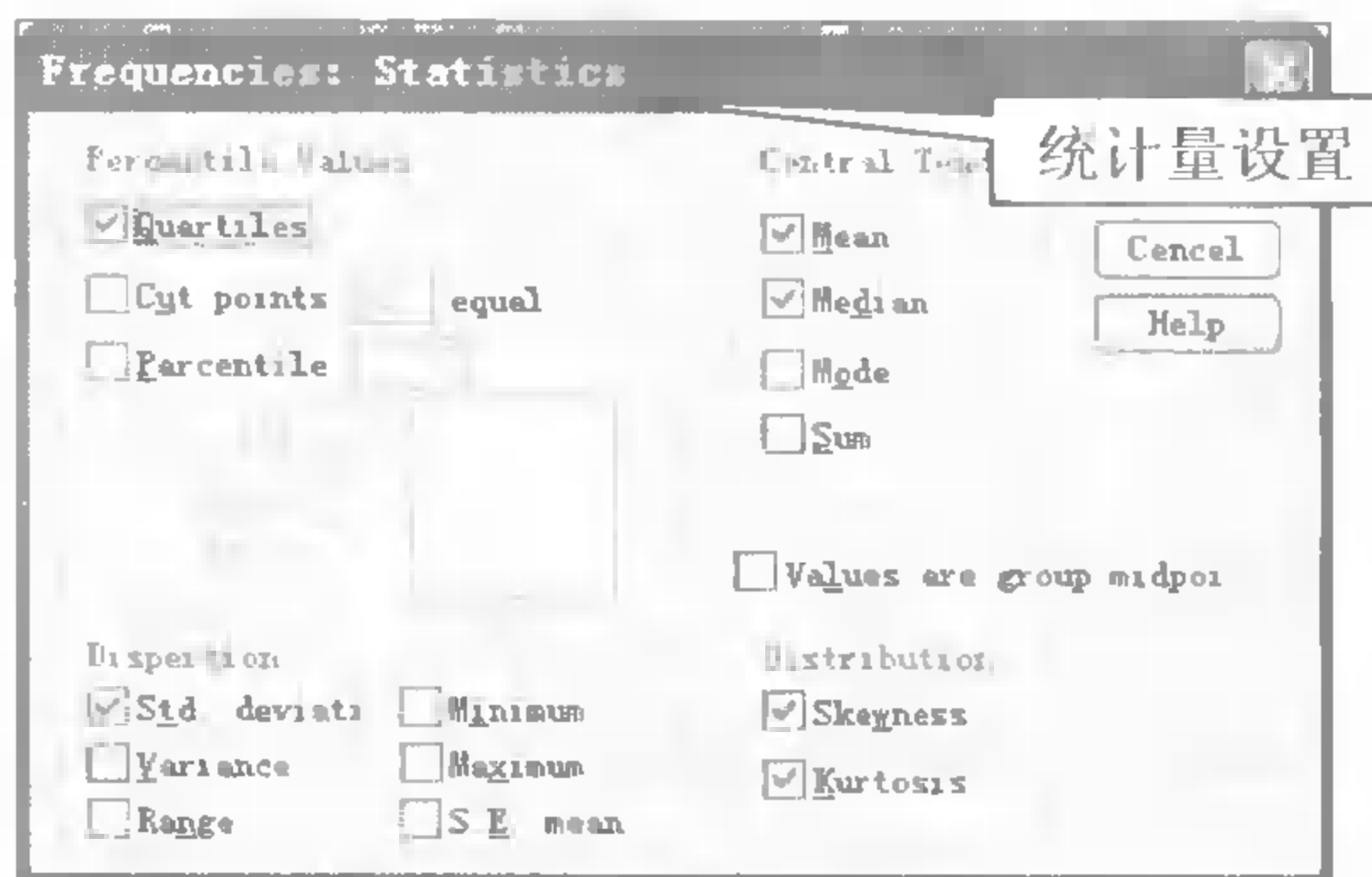


图 4-42 统计量设置子界面

## 2. 结果分析

双击图 4-40 中的 OK 按钮, SPSS Viewer 窗口的输出如图 4-43、图 4-44 所示。

统计量		
心理测试得分		
N	有效	30
	缺失	0
均值		4.0743
中值		4.1750
标准差		1.30649
偏度		-.015
偏度的标准误		.427
峰度		-.638
峰度的标准误		.833
百分位数	25	3.2350
	50	4.1750
	75	4.9375

图 4-43 连续变量的频数分析表

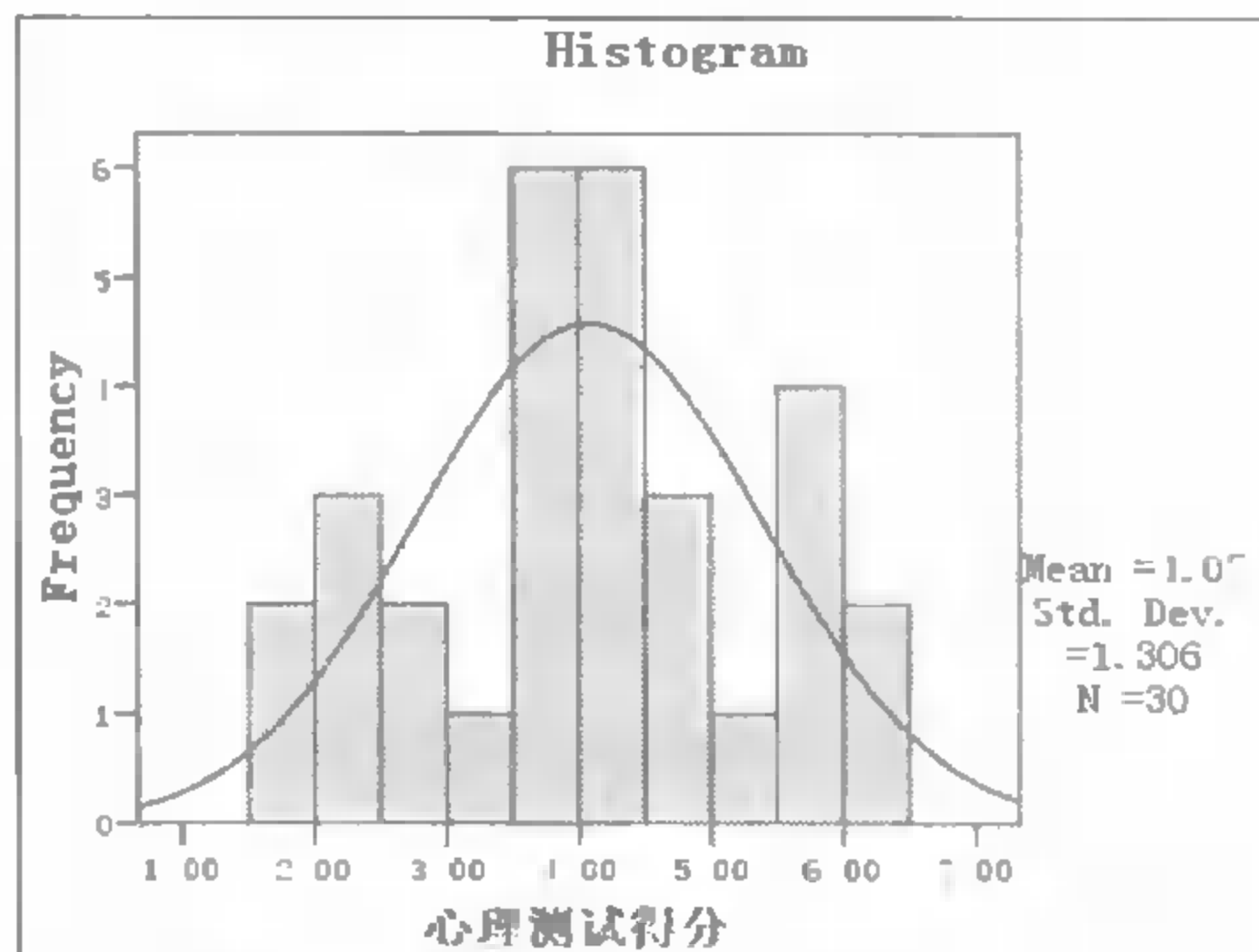


图 4-44 连续变量的频数分析图



“统计量”表格给出了所选统计量的计算结果，包括：均值 4.07、中位数 4.18 等；从偏度和峰度取值都较小看，score 的分布与正态分布较为接近。

从直方图中正态曲线和柱状图的拟合看，score 的分布与正态分布有所差异，建议采集更多数据后再做进一步的分析。

## 4.5 描述性统计分析

Descriptives 过程主要用来对连续变量做描述性分析，可以输出许多类型的统计量；也可以将原始数据转换成标准 Z 分值（标准化数据）并存入当前数据集，标准化后的变量值没有度量衡的差异，更加易于比较，经常应用于其他统计分析过程。

本节通过对几个经济指标的描述性统计分析，来介绍 Descriptives 过程的使用方法。

### 4.5.1 数据描述


文件“各省市发展情况数据.sav”中，记录了某年份全国 31 个省市地区的 6 项经济指标，数据格式如图 4-45 所示。本节通过 Descriptives 描述性分析过程，对这些指标的特征加以研究，分析它们取值的特点和分布情况。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	area	String	8	0	地区	None	None	8	Left	Nominal
2	educat	Numeric	8	1	教育投入	None	None	8	Right	Scale
3	industry	Numeric	8	2	工业产值	None	None	10	Right	Scale
4	farm	Numeric	8	2	农业产值	None	None	10	Right	Scale
5	work	Numeric	8	1	就业人数	None	None	8	Right	Scale
6	use	Numeric	8	0	消费水平	None	None	8	Right	Scale
7	gdp	Numeric	8	2	GDP 产值	None	None	10	Right	Scale

图 4-45 各省市发展情况的数据格式

### 4.5.2 Descriptives 分析

依次单击菜单“Analyze→Descriptive statistics→Descriptives”，打开描述性分析的主设置面板，如图 4-46 所示，在此选择分析变量。

(1) 变量选择。在变量列表选中所有变量，然后单击  按钮，将其选入 Variable(s) 列表，如图 4-47 所示。

- Variable(s) 列表：选入待分析的变量，比较适用于连续变量。
- Save standardized values as variables 复选框，选中表示为每个分析变量计算标准化后的数据 (Z score)，并保存到当前数据集；新变量的命名方式就是在原变量名的前面加“Z”，例如变量 gdp，标准化后的变量名为 Zgdp。新生成的变量可用于其他作图或统计分析过程。

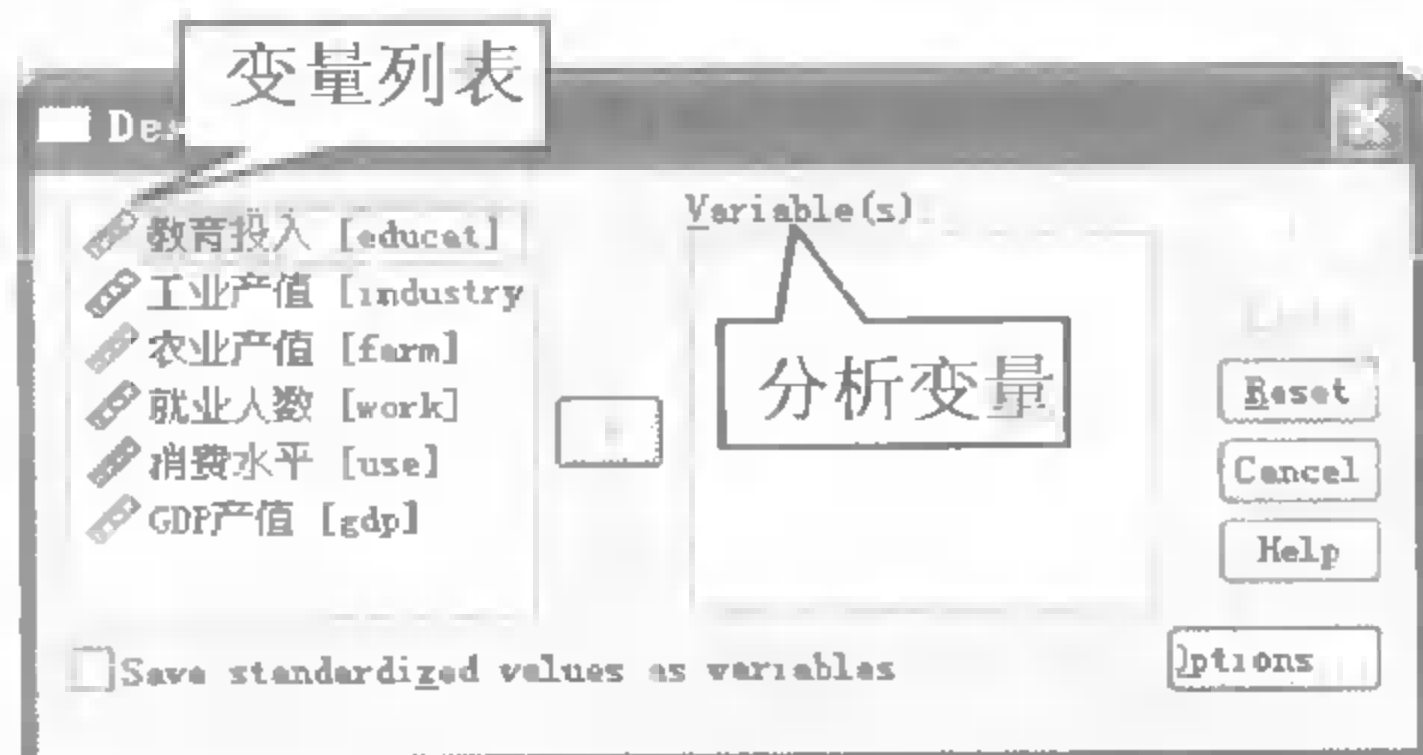


图 4-46 描述统计分析的主界面 1

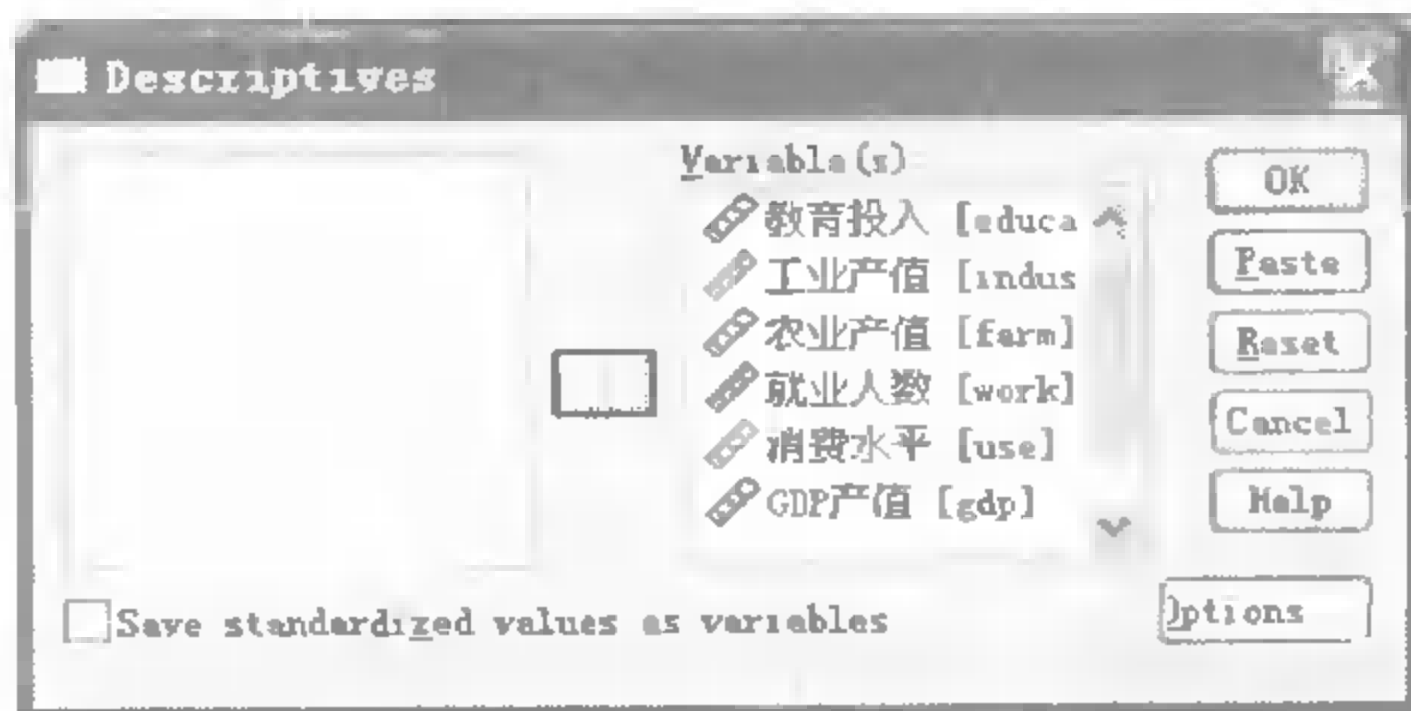


图 4-47 描述统计分析的主界面 2

(2) 选项设置。在图 4-47 中, 单击 Options 按钮弹出图 4-48 所示的子对话框, 用于选择要计算和输出的统计量。勾选 Skewness (偏度)、Kurtosis (峰度) 2 个复选框; 单击 Continue 按钮返回主设置界面。

此界面中关于变量集中趋势、离散程度、分布情况的几个统计量, 与图 4-35 中所示的统计量相同, 请参考在频数分析一节的介绍。

Display Order 栏设置输出结果的显示顺序, 有以下 4 个方法可选。

- Variables list 选项, 按照变量在主设置界面的选择顺序显示, 系统默认方式;
- Alphabetic 选项, 按照变量名的字母排序;
- Ascending means 选项, 按照变量均值的升序排列;
- Descending means 选项, 按照变量均值的降序排列。

显示顺序

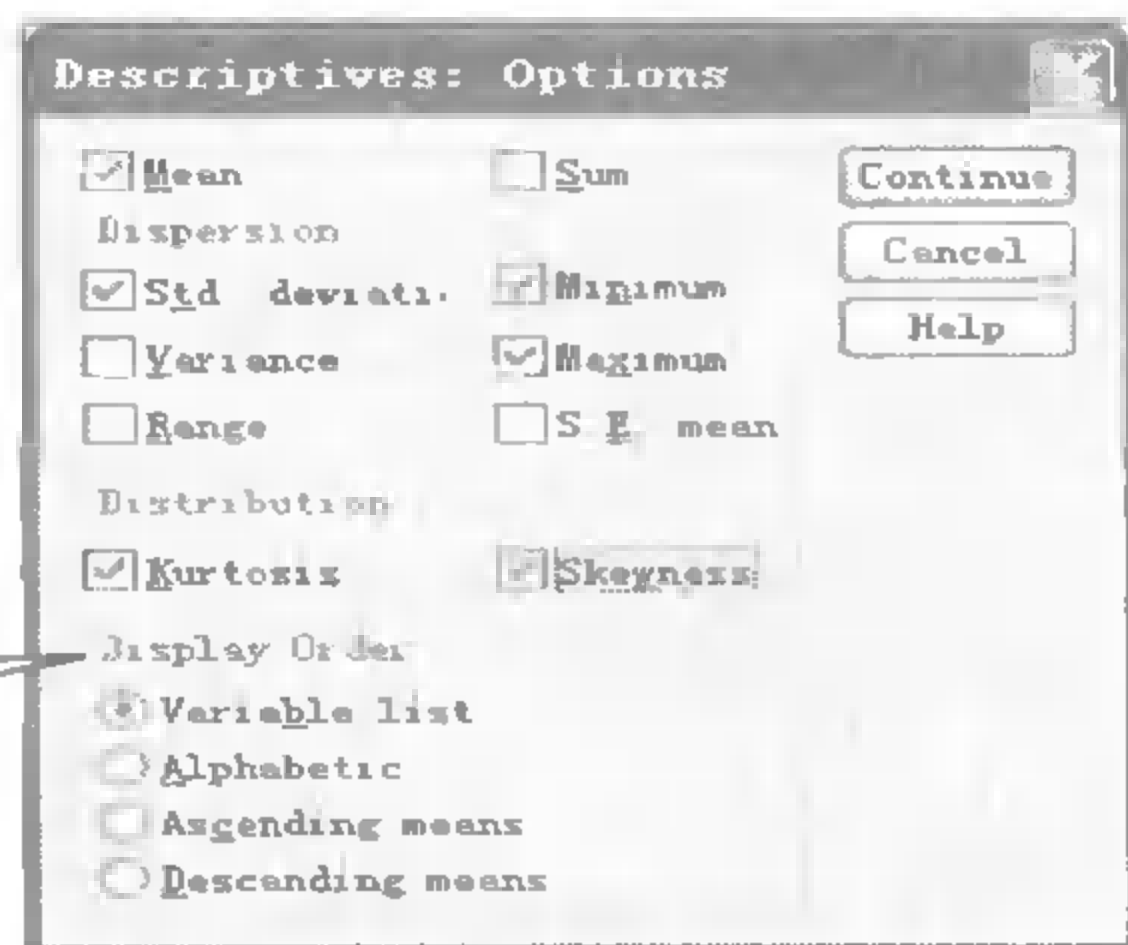


图 4-48 描述统计分析的参数设置

(3) 输出结果。在图 4-47 里单击 OK 按钮运行, SPSS Viewer 窗口的输出如图 4-49 所示。

描述统计量									
	N	极小值	极大值	均值	标准差	偏度		峰度	
	统计量	统计量	统计量	统计量	统计量	统计量	标准误	统计量	标准误
教育投入	31	44793.6	2454230.3	816687.93	552141.45	1.025	.421	1.521	.821
工业产值	31	9.02	3463.12	1076.0710	941.90070	1.148	.421	.670	.821
农业产值	31	33	11018.00	361.1190	1977.8417	5.568	.421	30.999	.821
就业人数	31	118.4	4999.6	2011.619	1399.1021	.554	.421	-.606	.821
消费水平	31	1511	9202	2911.87	1549.938	2.528	.421	8.323	.821
GDP产值	31	91.18	7919.12	2670.3306	2071.6584	1.070	.421	.702	.821
有效的 N (列表状态)	31								

图 4-49 描述性分析结果

通过观察发现, 农业产值的极小、极大值差异很大, 检查原始数据发现只有四川的农业产值大于 10 000, 其他都小于 100, 故确定为输入错误, 建议更正为 11.01; 由此失误也使得农业产值的偏度、峰度都很大, 严重偏离正态分布。

另外, 从消费产值的偏度、峰度看, 它与正态分布的差异也很大; GDP 在数量级 (10e3) 较小的情况下, 标准差却很大, 说明各地区的 GDP 产值之间存在较大差异。

## 4.6 探索分析过程

Explore 过程是对变量进行深入和详尽的描述性统计分析, 称之为探索性分析。它在一般描述性统计指标的基础上, 增加关于数据其他特征的文字与图形描述, 分析结果更加细致与全面, 有助于用户对数据做进一步分析。

探索性分析 (Explore) 过程, 能够生成关于所有个案、或不同分组个案的综合统计量及图形; 可以进行数据筛选工作, 例如检测异常值、极端值、数据缺口等; 还可以进行假设检验。通过探索性分析, 能够帮助用户决定选择何种统计方法进行数据建模, 判断是否需要把数据转换成正态分布, 以及是否需要做非参数统计。

SPSS 的 Explore 过程还适用于对大数据集的分析。

### 4.6.1 数据描述

Explore 过程适用于对数值型的变量（连续型或比率型）进行分析，因素变量应该是取有限个离散值的分类变量（用于对数据进行分组）。此过程对数据的分布没有特定限制。




我们采集了某小学部分 10~13 岁儿童的身高和体重数据，数据文件为“儿童身高体重.sav”，数据格式如图 4-50 所示。本节通过探索性分析对这些数据加以详细描述，并分析这部分儿童的身体特征有何特点。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	no	String	4		编号	None	None	4	Right	Nominal
2	gend	String	4		性别	{0, 男}...	None	4	Right	Nominal
3	age	Numeric	3	0	年龄	None	None	3	Right	Ordinal
4	high	Numeric	4	2	身高	None	None	8	Right	Scale
5	weight	Numeric	2	0	体重	None	None	6	Right	Scale

图 4-50 儿童身高和体重的数据格式

### 4.6.2 Explore 实例分析

依次单击菜单“Analyze→Descriptive statistics→Explore”执行探索性分析的过程，其主设置界面如图 4-51 所示，在这里选择进行分析的各种变量。

（1）变量选择。在变量列表选中体重变量，然后单击因变量列表左侧的  按钮，将其选入 Dependent 列表；在变量列表选中年龄变量，然后单击因素列表左侧的  按钮，将其选入 Factor List 列表；在变量列表选中编号变量，然后单击标签变量选框左侧的  按钮，将其选入 Label Cases 选框。

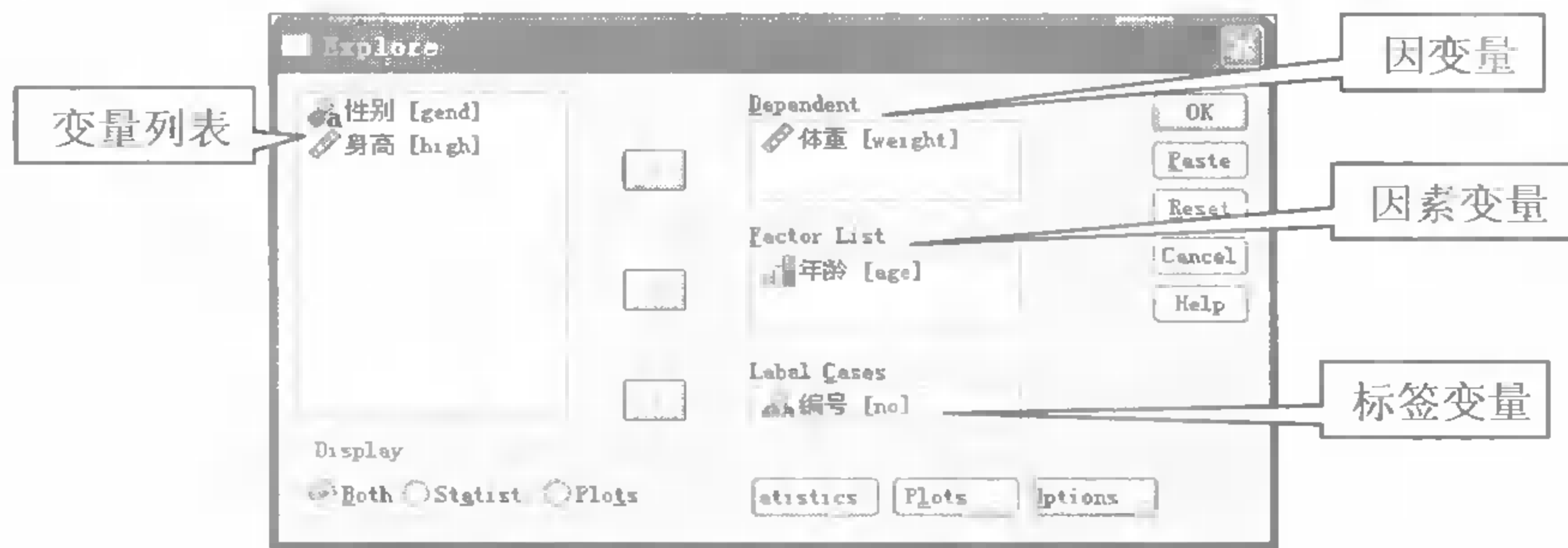


图 4-51 探索性分析的主对话框

① Dependent 列表：用于从变量列表选入因变量，一般为连续变量。

② Factor List 列表：用于从变量列表选入因素变量，一般为分类变量。

如果同时选入了多个因变量和多个因素变量，将对它们之间的两两组合分别进行分析，对于每对因变量和因素变量的搭配，输出结果都是类似的；选入的变量较多时，可能会耗费较长时间并产生很多输出。

③ Label Cases 选框：用于从变量列表选入标签变量，在结果里标识观测量。

④ Display 栏选择输出哪些内容。有 3 个可选项：Statistics（统计量表格）、Plots（图形）、Both（统计量表格和图形）。

（2）统计量设置。单击图 4-51 中的 Statistics 按钮，弹出如图 4-52 所示的统计量设置对话框。单击 Continue 按钮返回主设置界面。

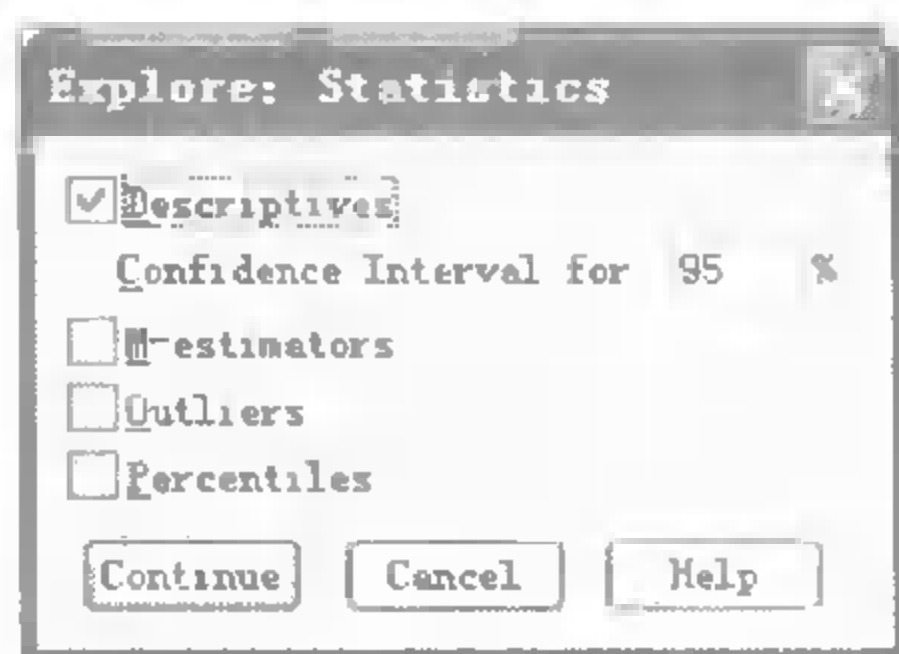


图 4-52 探索性分析的统计量设置

① Descriptives 复选框。选中此项会输出包含如下内容的表格：均值、中位数、众数、5%修正均数、标准误、方差、标准差、最小值、最大值、全距、峰度系数、峰度系数的标准误、偏度系数、偏度系数的标准误等；默认还会显示均值 95% 的置信区间，可以在 Confidence Interval for 后的输入框指定此置信区间的范围。

② M-estimators 复选框。计算并输出比均值和中位数更稳定的数据中心估计值，包括如下 4 个：Huber's、Andrews、Hampel's、Tukey's。

③ Outliers 复选框。输出 5 个最大值与 5 个最小值，包括观测值的标签。

④ Percentiles 复选框。输出第 5%、10%、25%、50%、75%、90%、95% 的百分位数。

(3) 作图设置。单击图 4-51 中的 Plot 按钮，弹出如图 4-53 所示的作图设置对话框。单击取消 Stem-and-leaf 复选项；单击选中 Histogram 复选项；勾选 Normality 复选框；单击选中 Power estimation 单选项。单击 Continue 按钮返回主设置界面。

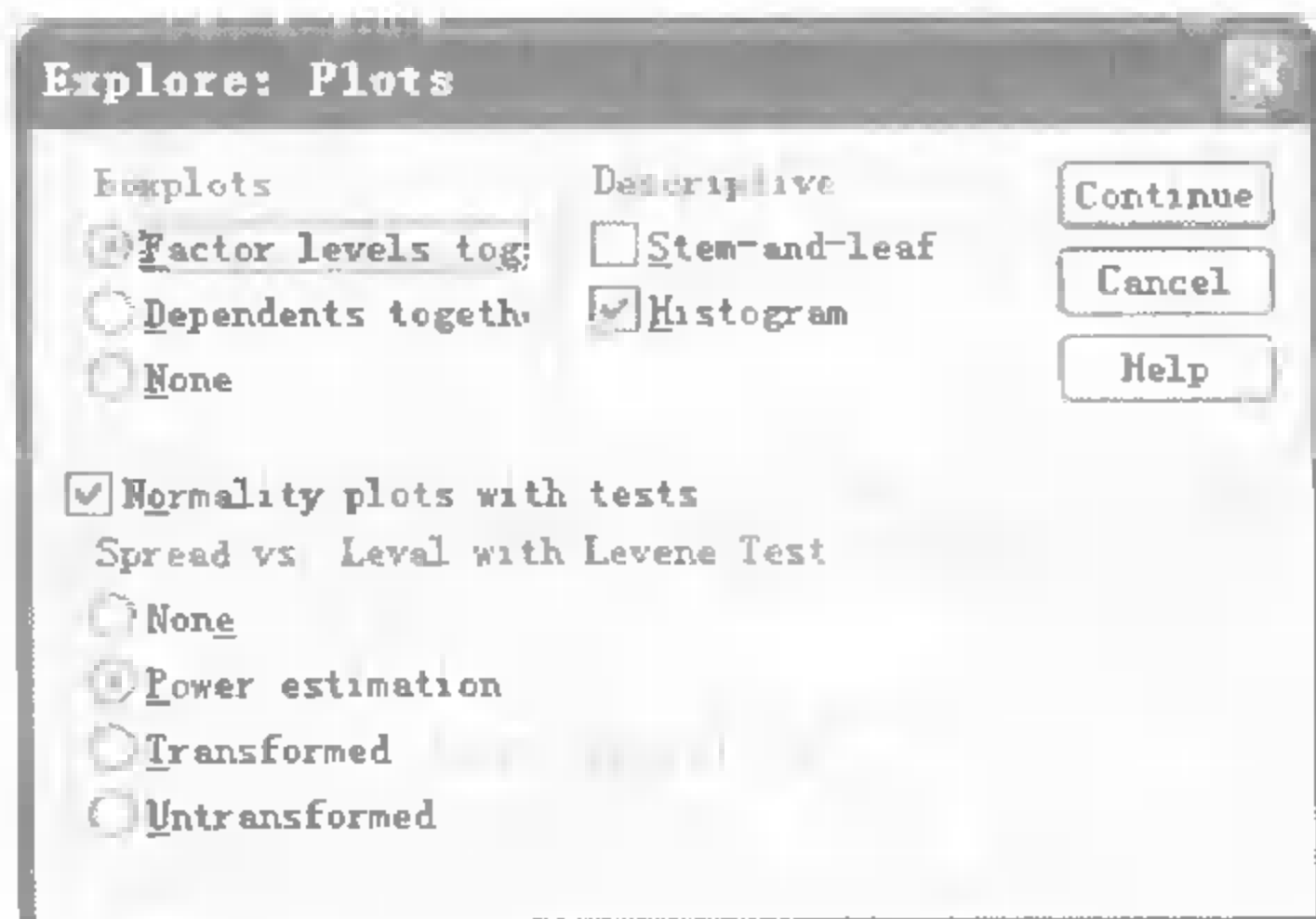


图 4-53 探索性分析的作图设置

① Boxplots 栏设置关于箱图的参数，有如下 3 个选择。

- ① Factor levels together 选项，对每个因素变量，每图只显示一个因变量，该项为默认选项。
- ② Dependents together 选项，对每个因素变量，每图显示所有的因变量。
- ③ None 选项，不绘制箱图。

② Descriptive 栏设置关于数据描述性质的图形输出，有 2 个选择。

- ① Stem-and-leaf 选项，作茎叶图，默认选项。
- ② Histogram 选项，作直方图。

③ Normality plots with tests 复选框。

作正态概率图和去趋势后的正态概率图，并输出检验正态性的 Kolmogorov-Smirnov 统计量及其 Lilliefors 置信水平；如果指定了非整数的权重，并且加权样本数在 3~50 之间时，输出 Shapiro-Wilk 统计量；如果没有指定权重或指定了整数权重，则当加权样本数在 3~5 000 之间时，输出 Shapiro-Wilk 统计量。



④ Spread vs. Level with levene Test 栏设置控制数据转换的散布对水平图。

同时显示回归曲线的斜率和方差齐性检验的 levene 统计量；如果选择了数据转换，将对转换后的数据进行 levene 检验。可选项有如下 4 个。

- ❶ None 不输出散布对水平图。
- ❷ Power estimation 幂次估计，产生四分位数的自然对数，对单元格中位数的自然对数的散布图，以及达到方差齐性要求的幂次估计；根据此散布对水平图，可以估计将各组方差转换成同方差所需的幂次。
- ❸ Transformed 数据变换，在其后的下拉列表选择具体的变换方法，可选方法有：Natural log（自然对数变换），系统默认；1/square root（平方根的倒数变换）；Reciprocal（倒数变换）；Square root（平方根变换）；Square（平方变换）；Cube（立方变换）。
- ❹ Untransformed 选项，不做任何数据变换，将输出产生原始的数据散布图，相当于幂次为 1 的转换。

（4）缺失值设置。单击图 4-51 中的 Options 按钮，弹出如图 4-54 所示的缺失值设置对话框。单击 Continue 按钮返回主设置界面。

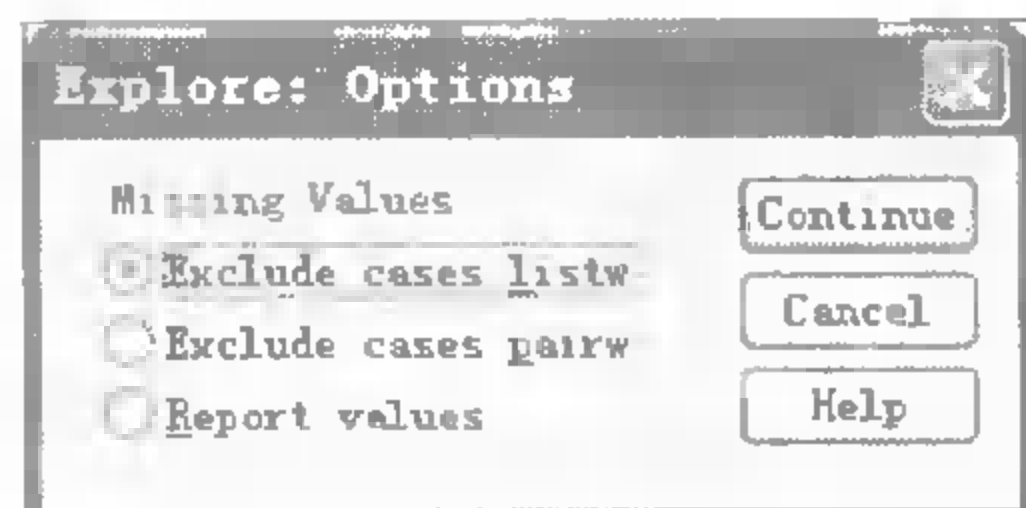


图 4-54 探索性分析的缺失值设置

对缺失值的处理方式有以下 3 种。

- ① Exclude cases listwise。对每个观测记录，只要分析所用到的变量中有 1 个含缺失值，就将该观测记录从所有分析中剔除，系统默认选项。
- ② Exclude cases pairwise。只有分析中用到的变量含缺失值时，才将相应的观测记录从当前分析中剔除，此方法可以最大限度地利用原始数据。
- ③ Report values。将因素变量中含有缺失值的观测作为一个单独的类别进行统计，所有输出结果都将包含这个被标识为缺失的类别。

（5）分析结果。单击图 4-51 中的 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 4-55~图 4-59 所示。

案例处理摘要						
年龄	有效				合计	
	N	百分比	N	百分比	N	百分比
10	8	100.0%	0	0%	8	100.0%
11	11	100.0%	0	0%	11	100.0%
12	7	100.0%	0	0%	7	100.0%
13	4	100.0%	0	0%	4	100.0%

描述				
年龄	统计量	标准差		
均值	39.00	9.82		
均值的 95% 置信区间	下限 31.68			
	上限 46.32			
5% 修整均值	39.00			
中值	38.00			
方差	7.714			
标准差	2.777			
极小值	15			
极大值	43			
范围	8			
四分位距	5			
偏度	480	75.2		
峰度	372	148.1		

图 4-55 摘要和描述性输出

正态性检验						
年龄	Kolmogorov-Smirnov <sup>a</sup>			Shapiro-Wilk		
	统计量	df	Sig.	统计量	df	Sig.
10	.250	8	.150	.900	8	.267
11	.174	11	.200*	.869	11	.072
12	.244	7	.200*	.938	7	.622
13	.237	4		.719	4	.650

a. 这是真实显著水平的下限。  
\*. Lilliefors 显著水平修正。

图 4-56 正态性检验输出

方差齐性检验					
年龄	Levene 统计量	df1	df2	Sig.	
基于均值	982	3	26	.417	
基于中值	735	3	26	.541	
基于中值和带有调整后的 df	735	3	18.575	.544	
基于修整均值	832	3	26	.488	

图 4-57 方差齐性检验输出



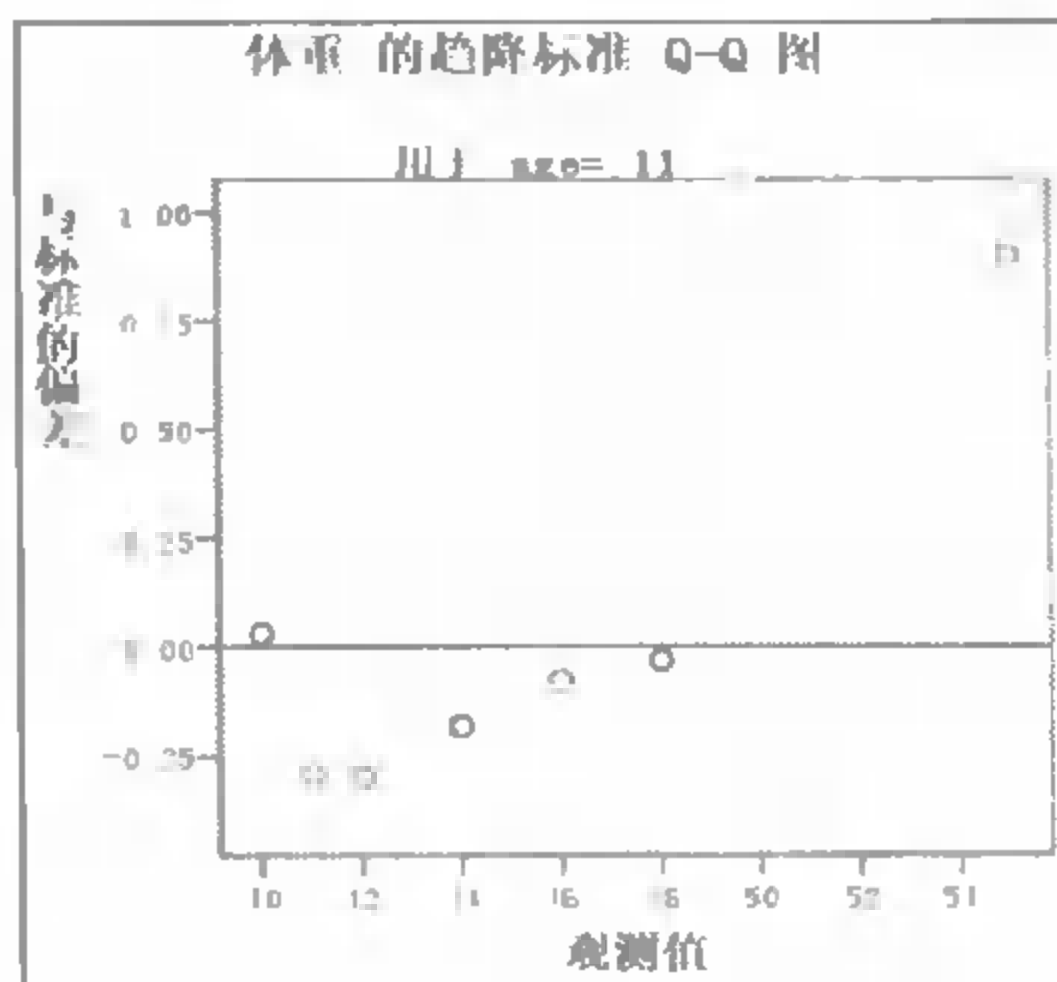
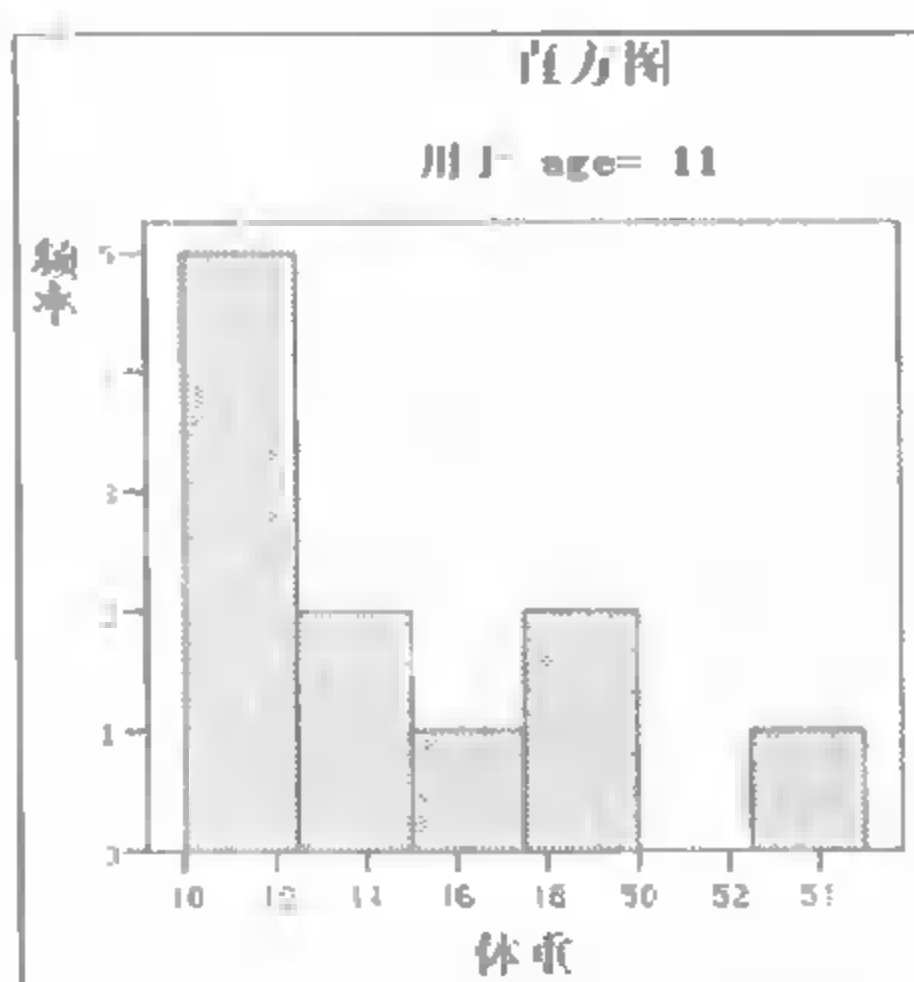


图 4-58 直方图和去除趋势的 QQ 图

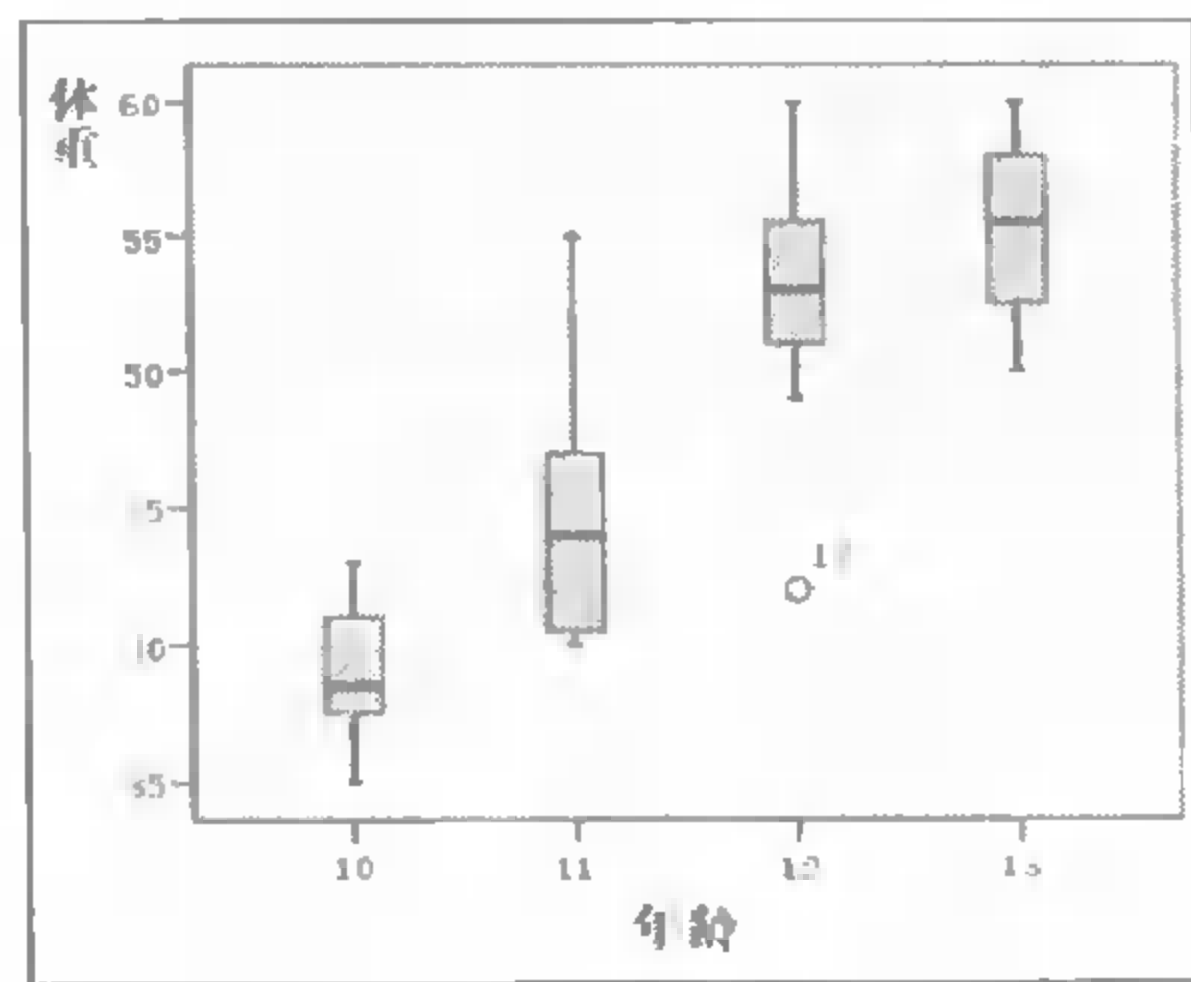


图 4-59 箱图

### ① 案例处理摘要和描述性表格输出。

如图 4-55 所示，摘要表格给出了不同年龄的有效个数和缺失个数，本例没有缺失数据；描述表给出了以年龄 10 为例的描述性统计量输出，包括均值及其 95% 的置信区间、中位数、方差等，其他年龄的输出与此类似。

### ② 正态性检验结果。

如图 4-56 所示，正态性检验的显著性水平 Sig 值都比较大，所以认为每个年龄的体重分布基本都为正态的。

### ③ 方差齐性检验结果。

如图 4-57 所示，4 种 Levene 检验的 Sig 值都大于 0.5，故不能否定方差齐性的假设。

由于本例的样本个数较少，故正态性检验和方差齐性检验的结论仅作参考，建议搜集更多的数据再进行分析，以得到更稳定的结论。

### ④ 直方图和 QQ 图输出。

如图 4-58 所示，是年龄为 11 岁时体重分布的直方图和 QQ 图，其它年龄的输出图形类似。从频率直方图看，体重偏小的人较多；从去除趋势的 QQ 图看，除了有一个点的偏差较大外，其他点都分布在横轴附近，建议搜集更多的数据后再下结论。

### ⑤ 箱图。

如图 4-59 所示，是体重对年龄的箱图，其中 12 岁儿童的体重里，有一个异常数据（偏小），它的编号是 14。

## 4.7 列联表分析过程

列联表是将观测数据按两个或更多属性（分类变量）进行分类时，列出的频数表。

一般情况下，若总体中的个体可按两个属性 A 与 B 分类，A 有  $r$  个等级  $A_1, A_2, \dots, A_r$ ，B 有  $c$  个等级  $B_1, B_2, \dots, B_c$ ，从总体中抽取大小为  $n$  的样本，设其中有  $N_{ij}$  个样本的属性属于等级  $A_i$  和  $B_j$ ， $N_{ij}$  就称为频数，将  $r \times c$  个  $N_{ij}$  排列为一个  $r$  行  $c$  列的二维列联表，简称为  $r \times c$  表。如果所考虑的属性多于两个，可按类似的方式作出它们的列联表，称之为多维列联表。由于属性变量的取值是离散的，因此列联表分析属于离散多元分析的范畴。列联表分析在市场研究中有着广泛的应用。

SPSS 中使用 Crosstabs 过程对计数资料和某些等级资料进行列联表分析，它可以给出 Pearson 卡方检验、似然比卡方检验、Fisher 精确检验 (Fisher's Exact Test)、Yates' corrected chi-square、Pearson's  $r$ 、Spearman's  $\rho$  等等许多统计检验和统计量的输出。

### 4.7.1 数据描述

为了研究客户满意度问题，某零售企业对在 4 个销售点消费的 582 名顾客进行了调查，已知客户服务水平是影响客户满意度的重要因素，通过调查数据，希望分析一下 4 个销售点的客户服务水平是否存在明显的差异。使用的调查数据摘自 SPSS 自带的 Demo 文件“satisf.sav”，所用数据文件为“客户满意度调查数据.sav”，数据格式如图 4-60 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	gender	Numeric	4	0	性别	{0, Male}...	None	6	Right	Nominal
2	agecat	Numeric	4	0	年龄段	{1, 18-24}...	None	6	Right	Ordinal
3	payment	Numeric	4	0	付款方式	{1, Cash}...	None	7	Right	Nominal
4	distance	Numeric	4	0	商场与家的距离	{1, < 1 mile}	None	8	Right	Ordinal
5	store	Numeric	4	0	销售点	{1, 商场1}...	None	6	Right	Nominal
6	price	Numeric	4	0	价格满意度	{1, 很不满意}...	None	6	Right	Ordinal
7	numitems	Numeric	4	0	商品多样性满意度	{1, 很不满意}...	None	8	Right	Ordinal
8	org	Numeric	4	0	组织满意度	{1, 很不满意}...	None	6	Right	Ordinal
9	service	Numeric	4	0	服务满意度	{1, 很不满意}...	None	7	Right	Ordinal
10	quality	Numeric	4	0	质量满意度	{1, 很不满意}...	None	7	Right	Ordinal
11	overall	Numeric	4	0	总体满意度	{1, 很不满意}...	None	7	Right	Ordinal

图 4-60 关于客户满意度调查的数据格式

列联表分析所用到的分类变量，都应为数值型的或者短字符串型（小于等于 8 个字符）的。

下面就利用列联表分析来进行假设检验，检验的零假设是：在 4 个不同的销售点，顾客对商家的服务满意度水平之间没有显著差异。

### 4.7.2 列联表分析的参数设置

依次单击菜单“Analyze→Descriptive statistics→Crosstabs”，打开列联表分析的主设置面板，如图 4-61 所示。

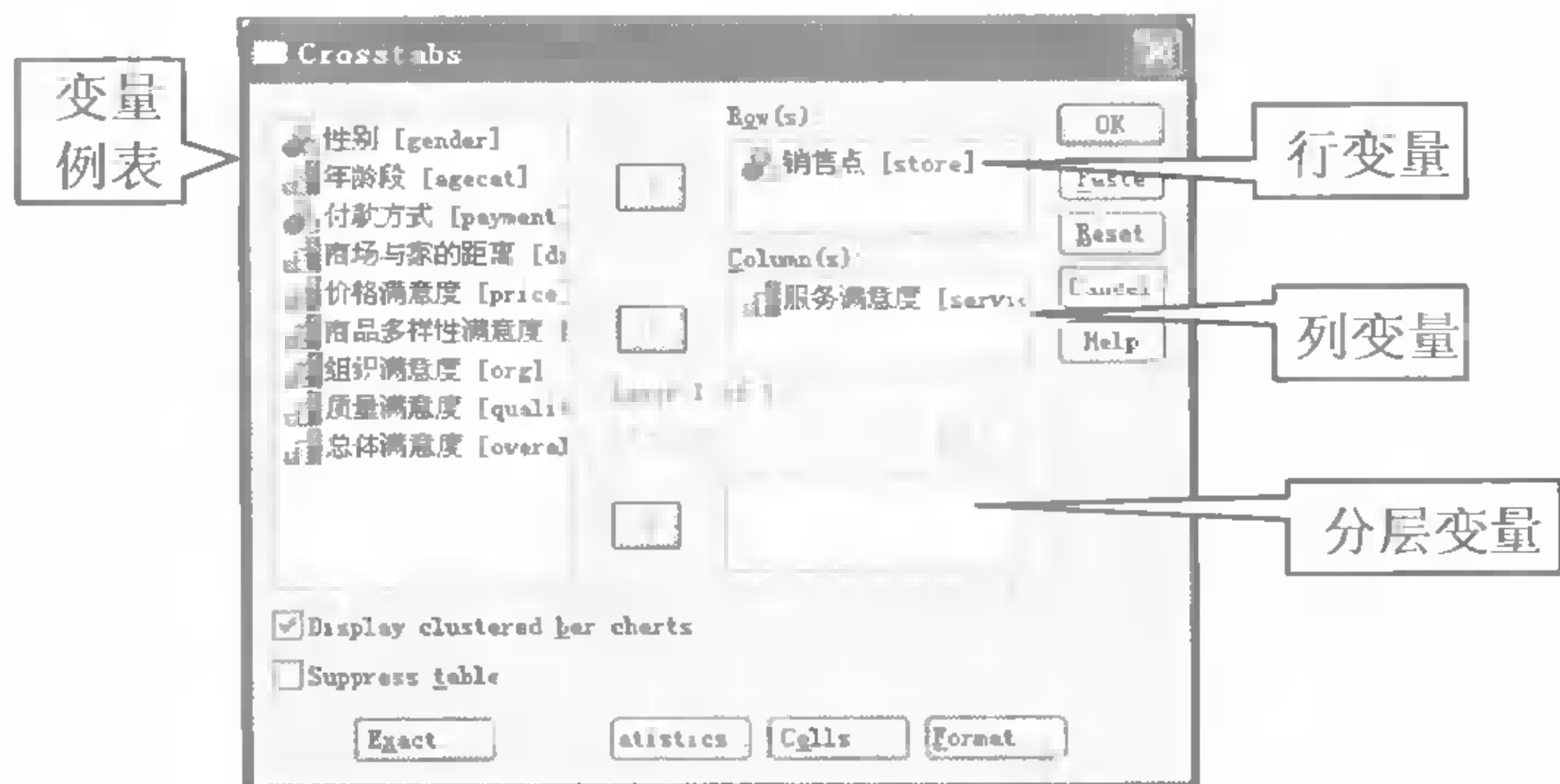




图 4-61 列联表分析的主界面

(1) 变量设置。在变量列表选中销售点变量，然后单击 Row(s) 列表左侧的  按钮，将其选入行变量列表；在变量列表选中服务满意度变量，然后单击 Column(s) 列表左侧的  按钮，将其选入列变量列表；勾选 Display clustered bar chart 复选框。

① Row(s) 列表用于选入行变量；Column(s) 列表用于选入列变量。

② Layer 列表用于选入分层变量。

通过单击 Previous、Next 按钮，可以指定多组分层变量。对分层变量的每个取值（或取值组合），将分别进行关于行、列变量的列联表分析。

③ Display clustered bar chart 复选框，输出关于各类别频数统计的复合条形图。

④ Suppress table 复选框，选中表示不输出频数统计表格。

(2) 精确检验设置。单击图 4-61 中的 Exact 按钮（只有安装了 SPSS 的 EXACT 模块才会有此选项），弹出如图 4-62 所示的精确检验设置对话框，设置计算检验统计量的显著性水平的参数。单击 Continue 按钮返回主设置界面。

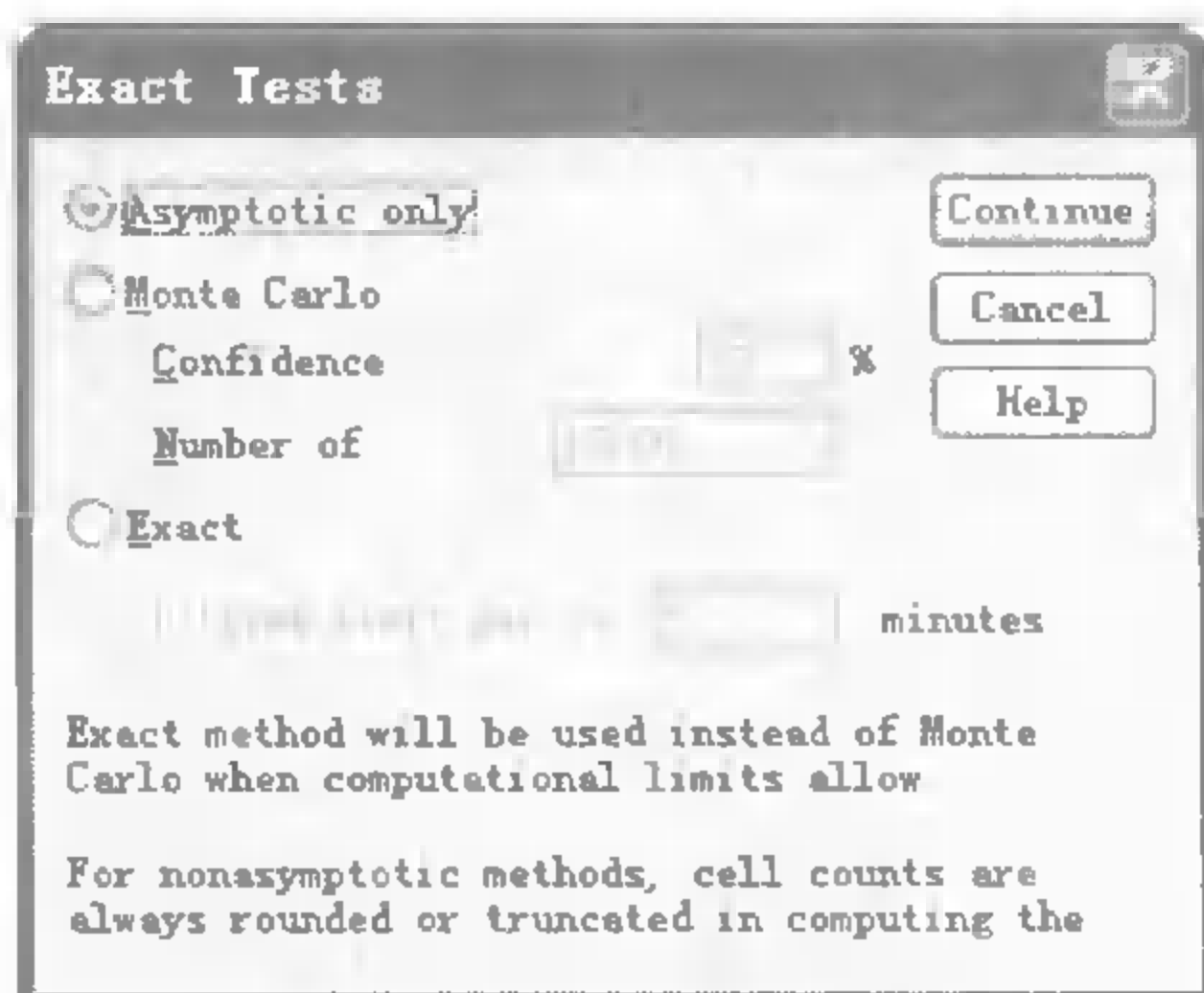


图 4-62 精确检验设置

① Asymptotic only 选项，基于检验统计量的渐进分布计算显著性水平。

由此计算的显著性水平低于 0.05 时，认为是显著的；此方法适用于较大的数据集，当数据量较少或者没有明显的分布特征时，由此方法所得的结论可能会很不稳定。

② Monte Carlo 选项，蒙特卡洛估计方法。

这是对精确显著性水平的无偏估计。它先从一个参考样本重复抽取样本量相同的子样本，再通过子样本的显著性水平推导总体的显著性水平；此方法非常适用于数据量太大，无法使用其他方法进行计算的情况。Confidence 输入框指定置信水平，默认为 99%；Number of 输入框指定抽样的次数，默认为 10 000 次。

③ Exact 选项，精确检验。

一般情况下，由此计算的显著性水平低于 0.05 时被认为是显著的，即认为行、列变量之间存在一定的相关性。选中 Time limit 复选框表示只有当 Exact 方法对单个检验的计算时间低于限制条件时，才用它取代 Monte Carlo 方法；minutes 前的输入框指定时间限制（单位分钟），默认为 5 分钟。

(3) 统计量设置。单击图 4-61 中的 Statistics 按钮，弹出如图 4-63 所示的统计量设置子对话框；勾选 Chi-square 复选框；依次勾选 Nominal 栏下的 4 个复选框；单击 Continue 按钮返回主设置界面。

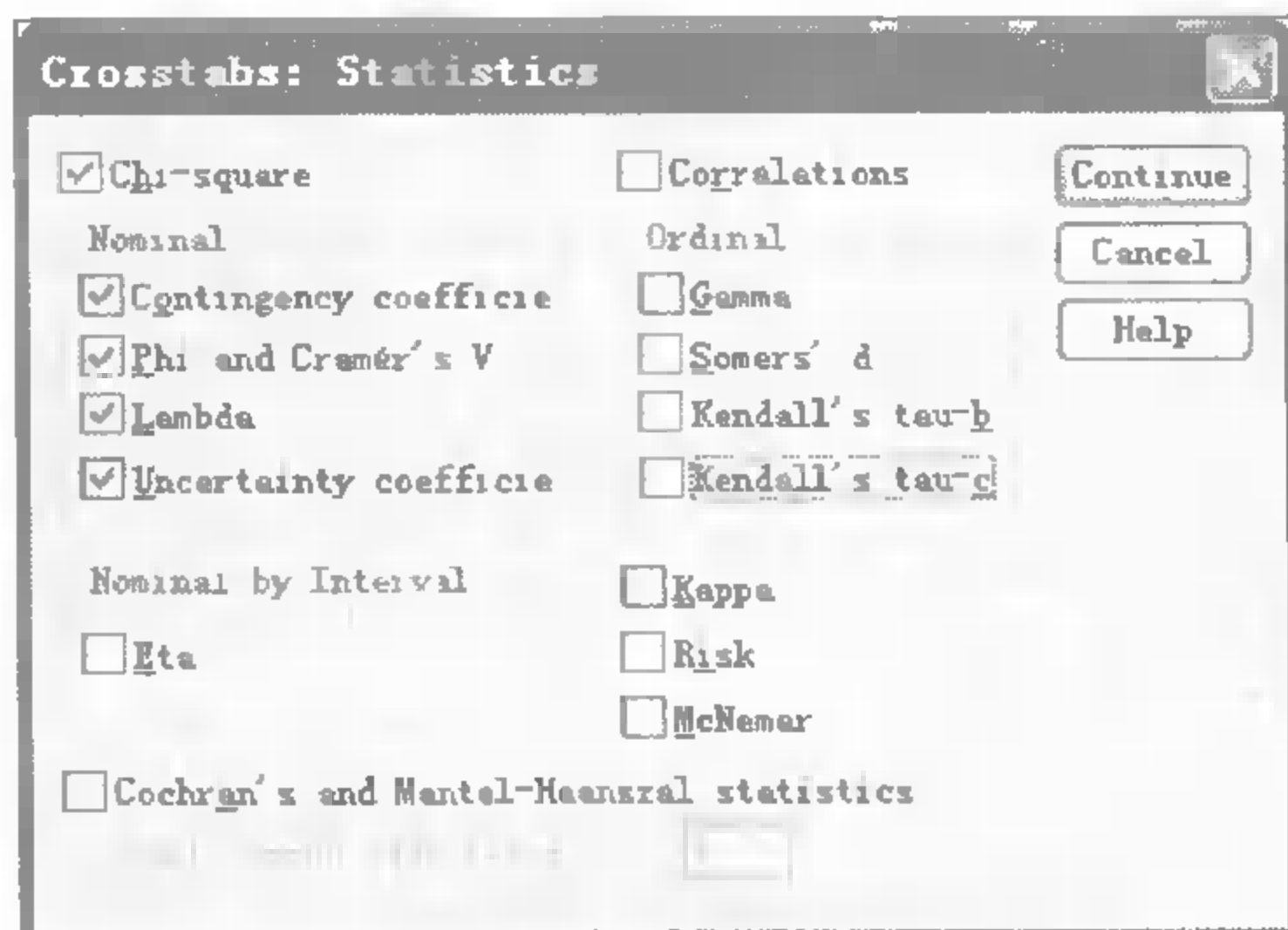


图 4-63 列联表分析的统计量设置

① Chi-square 复选框表示进行卡方检验。卡方检验包括：Pearson 卡方检验、似然比卡方检验、Fisher 精确检验（Fisher's Exact Test）等。

② Correlations 复选框表示进行相关性检验。相关性检验包括行、列变量的 Pearson 相关系数和 Spearman 等级相关系数等。

③ Nominal 栏选择关于名义变量的检验统计量，有 4 个可选项。

● Contingency coefficient 复选框，基于卡方的相关性统计量，其取值界于 0~1 之间，0 表示行、列变量之间没有关系，越接近 1 表示相关性越强。

● Phi and Cramer's V 复选框，两个基于卡方的相关性统计量。

● Lambda 复选框，用于反映由自变量预测因变量时的误差缩减比例，取值为 1 时表明用自变量能完全预测因变量，取值越接近 0 表示自变量对因变量的预测作用越小。

● Uncertainty coefficient 复选框，用于反映由一个自变量预测其他变量时的误差缩减比例，取值越接近 0 表示用这个变量预测其他变量的效果越差。

④ Ordinal 栏选择关于有序变量的检验统计量，有 4 个可选项。

● Gamma 复选框，关于两个有序变量相关性的对称性度量，取值在 -1~1 之间，取值的绝对值越接近 1 表示两个变量的相关性越强，取值接近 0 表示相关性较弱。

● Somers' d 复选框，关于两个有序变量相关性的非对称性度量，取值在 -1~1 之间，取值的绝对值越接近 1 表示两个变量的相关性越强，取值接近 0 表示相关性较弱。

● Kendall's tau-b 复选框，关于有序变量（或秩变量）相关性的非参数统计量，计算时将结（tie）考虑在内，取值在 -1~1 之间，符号表示相关性的方向，绝对值越大表示相关性越强。

● Kendall's tau-c 复选框，关于有序变量相关性的非参数统计量，计算时不考虑结（tie）的问题，取值在 -1~1 之间，符号表示相关性的方向，绝对值越大表示相关性越强。

⑤ Nominal by Interval。当一个变量为分类变量（且必须是数值编码），另一个变量为连续变量时，勾选此栏的 Eta 复选框；Eta 的取值范围为 0~1，取值越接近 1 表示行、列变量的相关性越强。输出 2 个 Eta 值：一个将行变量作为连续变量；一个将列变量作为连续变量。

⑥ Kappa 复选框。输出 Cohen's Kappa 统计量，用于衡量两种方法评价同一个对象时的一致性，取值在 0~1 之间，越接近 1 表示两种方法的评价越一致；只有当表格的行、列变量有相同的取值个数，以及相同的取值范围时才会被输出。

⑦ Risk 复选框。对于 2×2 的表格，此统计量用来衡量某个因素与某事件发生与否的相关性大小，也就是行、列变量的相关性；如果计算所得的置信区间包含 1，则认为此因素与事件发生与否没有显著的相关性。

⑧ McNemar 复选框。关于两个二分变量的非参数检验，它用卡方分布检验响应的改变；经常用来检验对实验进行某项干预之前与之后，所引起响应（某事发生）的变化。

⑨ Cochran's and Mantel-Haenszel statistics 复选框，检验两个二分变量独立性的统计量。

(4) 单元格显示设置。在图 4-61 中，单击 Cells 按钮，弹出如图 4-64 所示的单元格显示设置子对话框，选择要在

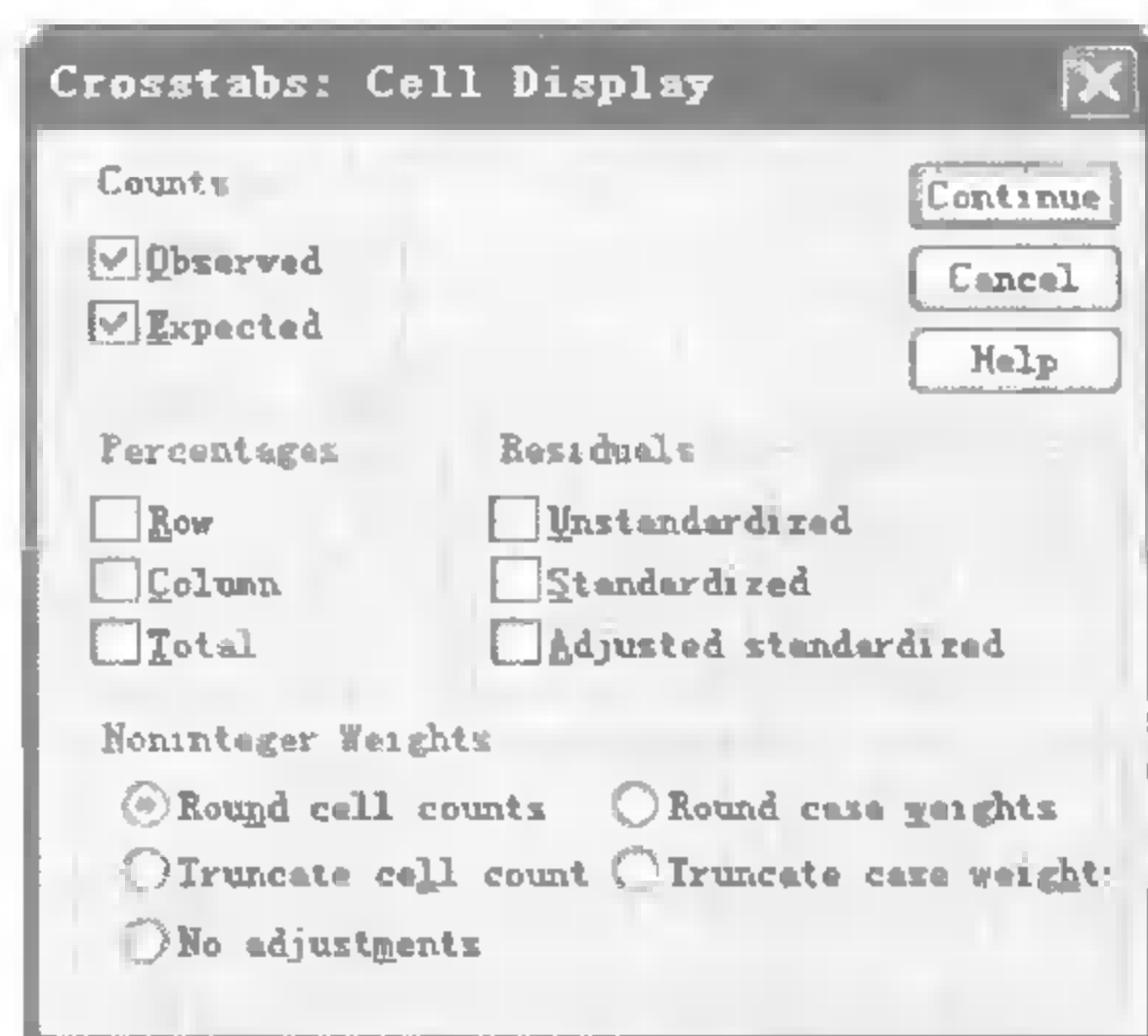


图 4-64 列联表分析的单元格显示设置



输出表格显示的统计量；勾选 Expected 复选框；单击 Continue 按钮返回主设置界面。

① Counts 栏设置关于计数的选项，有：Observed（实际频数）、Expected（期望频数）。

② Percentages 栏设置关于百分比的选项，有 3 个选择：Row（行百分比）、Column（列百分比）、Total（总百分比）。

③ Residuals 栏设置关于残差的选项，有以下 3 个选择。

- ☐ Unstandardized 非标准化残差，即实际频数与期望频数的差。
- ☐ Standardized 标准化残差，由非标准化残差除以残差的标准差估计值，进行标准化后得到。
- ☐ Adjusted standardized 调整的标准化残差，由非标准化残差除以残差的标准误差估计值，进行标准化得到。

④ Noninteger Weights 栏设置关于非整数权重变量的参数。

一般情况下，输出单元格显示整数形式的计数信息，但是如果在数据集中指定了取值可以为小数的权重变量时，在此设置关于非整数权重的处理方式。有如下 5 个可选内容。

- ☐ Round cell counts 选项，单个观测记录的权重按实际值计算，但列联表里的累计权重在用于其他统计量的计算之前，先要进行四舍五入处理。
- ☐ Round case weights 选项，直接对单个观测记录的权重，进行四舍五入。
- ☐ Truncate cell count 选项，单个观测记录的权重按实际值计算，但列联表里的累计权重在用于其他统计量的计算之前，先进行取整运算（即舍去其小数部分）。
- ☐ Truncate case weights 选项，直接对单个观测记录的权重进行取整运算。
- ☐ No adjustments 选项，不做调整，但是进行精确检验之前，列联表里的累计权重都需要被四舍五入或取整。

(5) 行顺序设置。在图 4-61 中，单击 Format 按钮，弹出如图 4-65 所示的行顺序设置子对话框，设置行变量在结果中的显示顺序。单击 Continue 按钮返回主设置界面。

Row Order 栏，指定显示顺序为：Ascending（升序）或 Descending（降序），默认为升序。

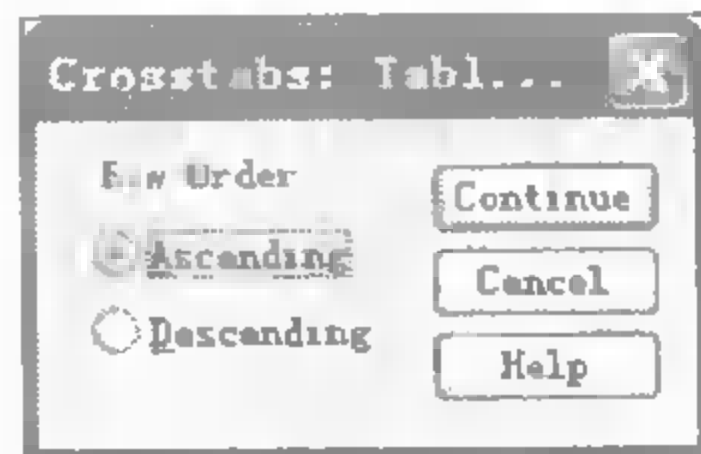


图 4-65 列联表分析的行顺序设置

### 4.7.3 列联表分析的输出结果

在图 4-61 里单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 4-66～图 4-69 所示。

案例处理摘要

	案例					
	有效的		缺失		合计	
	N	百分比	N	百分比	N	百分比
销售点 服务满意度	582	100.0%	0	0%	582	100.0%

销售点·服务满意度交叉制表

		服务满意度					合计
		很不满意	不满意	一般	满意	很满意	
销售点	商场1 计数	25	20	38	30	33	146
	期望的计数	23.3	26.3	39.4	28.1	28.8	146.0
	商场2 计数	26	30	34	27	19	136
	期望的计数	21.7	24.5	36.7	26.2	26.9	136.0
	商场3 计数	15	20	41	33	29	138
	期望的计数	22.1	24.9	37.2	26.6	27.3	138.0
	商场4 计数	27	35	44	22	34	162
	期望的计数	25.9	29.2	43.7	31.2	32.0	162.0
	合计 计数	93	105	157	112	115	582
	期望的计数	93.0	105.0	157.0	112.0	115.0	582.0

图 4-66 摘要和交叉表输出

卡方检验			
	值	df	渐进 Sig. (双侧)
Pearson 卡方	16.293 <sup>a</sup>	12	.178
连续性校正			
似然比	17.012	12	.149
线性和线性组合	.084	1	.772
有效案例中的 N	582		

<sup>a</sup> 1.0 单元格(.0%)的期望计数少于 5。最小期望计数为 21.73。

图 4-67 卡方建议结果





# 第5章 均值比较和 T 检验

统计分析常常采取抽样研究的方法，即从总体中随机抽取一定数量的样本进行研究，并以此推测总体的特性。由于总体中的每个个体间均存在差异，即使严格遵守随机抽样原则，也会由于多抽到一些数值较大或较小的个体致使样本统计量与总体的参数之间有所不同；又由于实验者测量技术的差别或测量仪器精确程度的差别等因素的存在造成偏差，使样本统计量与总体参数之间存在差异。由此可以认为，均值不相等的两个样本不一定来自均值不同的总体。能否用样本均值估计总体均数，两个变量均值接近的样本是否来自均值相同的总体；两个样本某变量的均值不同，其差异是否具有统计意义，它能否说明总体之间存在的差异，这些都是研究工作中经常提出的问题，解决它们就需要进行均值比较和检验。

对于假设检验问题，可以参考如图 5-1 所示的简单分类，本章主要探讨其中的参数检验部分。

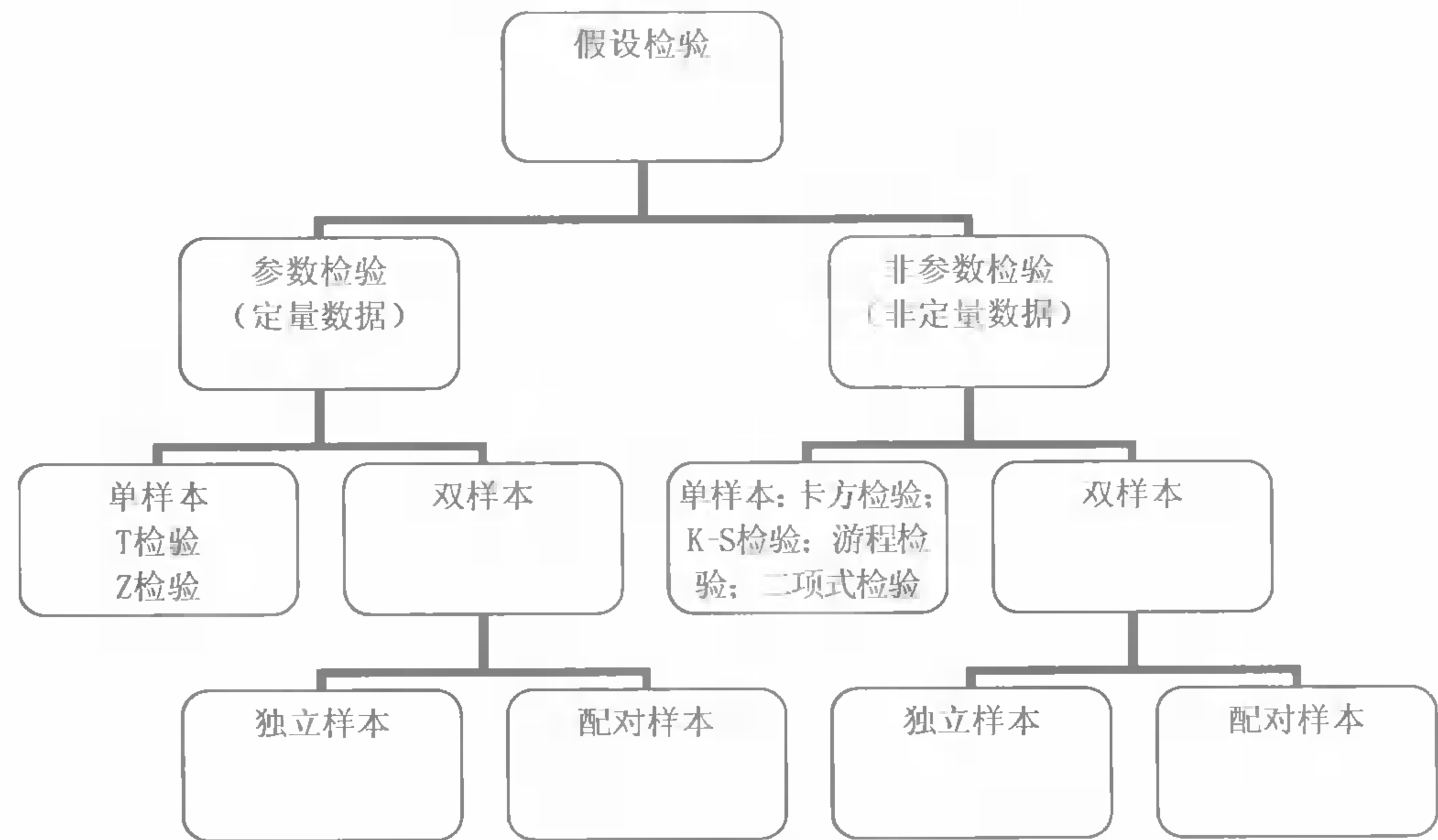


图 5-1 假设检验的简单分类

对来自正态总体的两个样本进行均值比较，常使用 T 检验方法，但两个样本的方差相等时与不等时使用的 T 统计量公式是不同的。对方差的齐次性检验常使用 F 检验方法。

为了正确地进行假设检验，建议先回答以下 3 个问题，根据答案的不同选用不同的统计量和 SPSS 过程进行检验。

- (1) 数据样本有一个还是两个？如果是两个样本，那么各项数据是否都取自同一来源？如果是，那么这两个样本就不是独立样本，而称为配对样本或成对样本。
- (2) 样本方差是否已知？
- (3) 样本是否是大容量的？一般以样本观测数  $n \geq 30$  作为大样本。

## 5.1 Means 过程

### 5.1.1 原理与方法

SPSS 的均值比较过程 (MEANS) 用于分组计算、比较指定变量的描述性统计量，如总和、均值、方差、标准差、观测数等，还可以给出方差分析表和线性检验结果等信息。当观测按一个分类变量分组时，MEANS 过程可以进行分组计算。例如：要计算工作人员上班路程的平均公里数，Sex 变量把工作人员按性别分为男、女两组，MEANS 过程可以分别计算男、女上班路程的平均公里数。

使用 MEANS 过程求若干分组的描述性统计量，目的在于比较，因此必须分组求均值，这是与 Descriptive 过程不同之处。

### 5.1.2 SPSS 实例分析

#### 1. 问题和数据描述

本节通过均值比较过程，分析一组学生身高和体重数据的特征，所用数据文件为“学生身体特征数据.sav”，各变量含义如图 5-2 所示。

	Name	Type	Width	Decim	Label	Values	Missing	Column	Align	Measure
1	no	Numeric	2	0	编号	None	None	3	Right	Scale
2	sex	Numeric	1	0	性别 (男、女)	None	None	2	Right	Nominal
3	age	Numeric	2	0	年龄	None	None	8	Right	Scale
4	h	Numeric	5	2	身高	None	None	5	Right	Scale
5	w	Numeric	3	0	体重	None	None	4	Right	Scale

图 5-2 学生身体特征数据格式

#### 2. SPSS 操作过程

依次单击菜单“Analyze→Compare Means→Means”，执行均值比较过程，弹出的设置界面如图 5-3 所示。

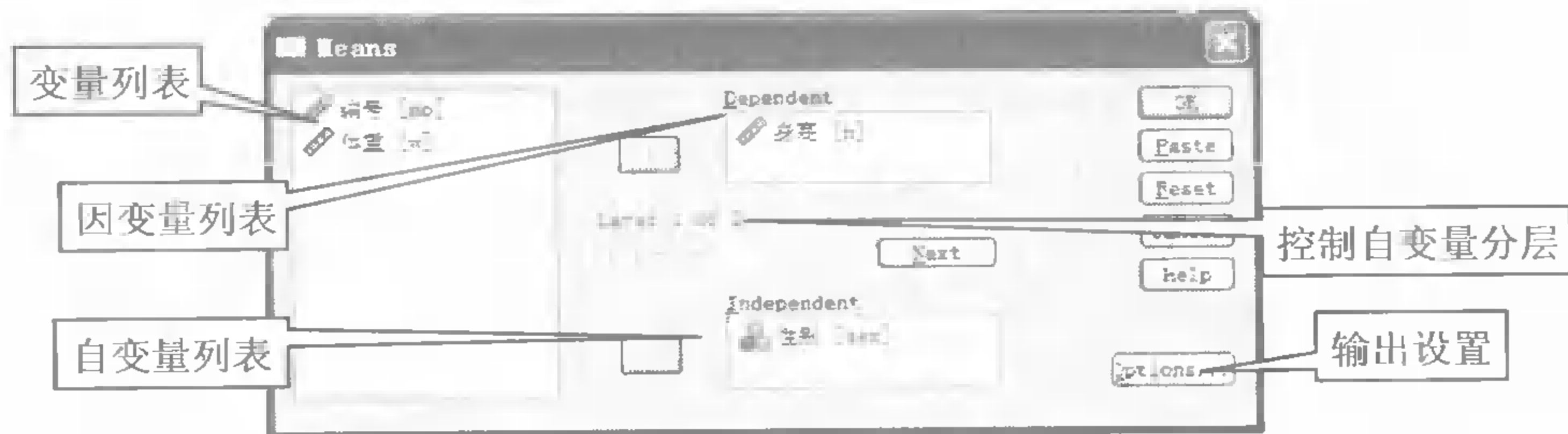




图 5-3 Means 过程主设置面板

- (1) 分析变量的选择。在变量列表单击选中身高变量，单击从上至下第一个 按钮，

将其作为因变量选入 Dependent 列表；在变量列表单击选中年龄变量，单击从上至下第二个  按钮，将其作为第一层(Layer)的自变量选入 Independent 列表；单击 Next 按钮，Independent 列表被清空；在变量列表单击选中性别变量，单击从上至下第二个  按钮，将其作为第二层的自变量选入 Independent 列表。

- 变量列表。Dependent 因变量列表，用于选入待输出统计特征值的数值型变量，如果同时选入多个，将分别单独处理；Independent 自变量列表，用于选入分组变量，可以是数值型或短字符型。
- 自变量分层的操作和意义。自变量列表中同时输入的一组分类变量被当作一层。例如：已知两个分类自变量  $x$ 、 $y$  分别有  $a$ 、 $b$  个取值，因变量只有一个。若两个自变量出现在同一层中，它们将分别和因变量结合进行分组统计，即对因变量做“ $a+b$ ”次统计；若两个自变量分别出现于两个层中，则先将它们自己的取值进行配对组合，再与因变量结合进行分组统计，即对因变量做“ $a \times b$ ”次统计。

单击 Previous、Next 就可以显示与编辑不同的分层，清空某层的自变量列表将删除此层。自变量分层处显示的 Layer M of N 表示共有 N 层自变量，当前显示第 M 层，

(2) 选项参数设置。在图 5-3 中，单击 Options 按钮，弹出如图 5-4 所示的输出设置面板，勾选底部的 Anova、Test 两个复选框。单击 Continue 按钮返回主界面。

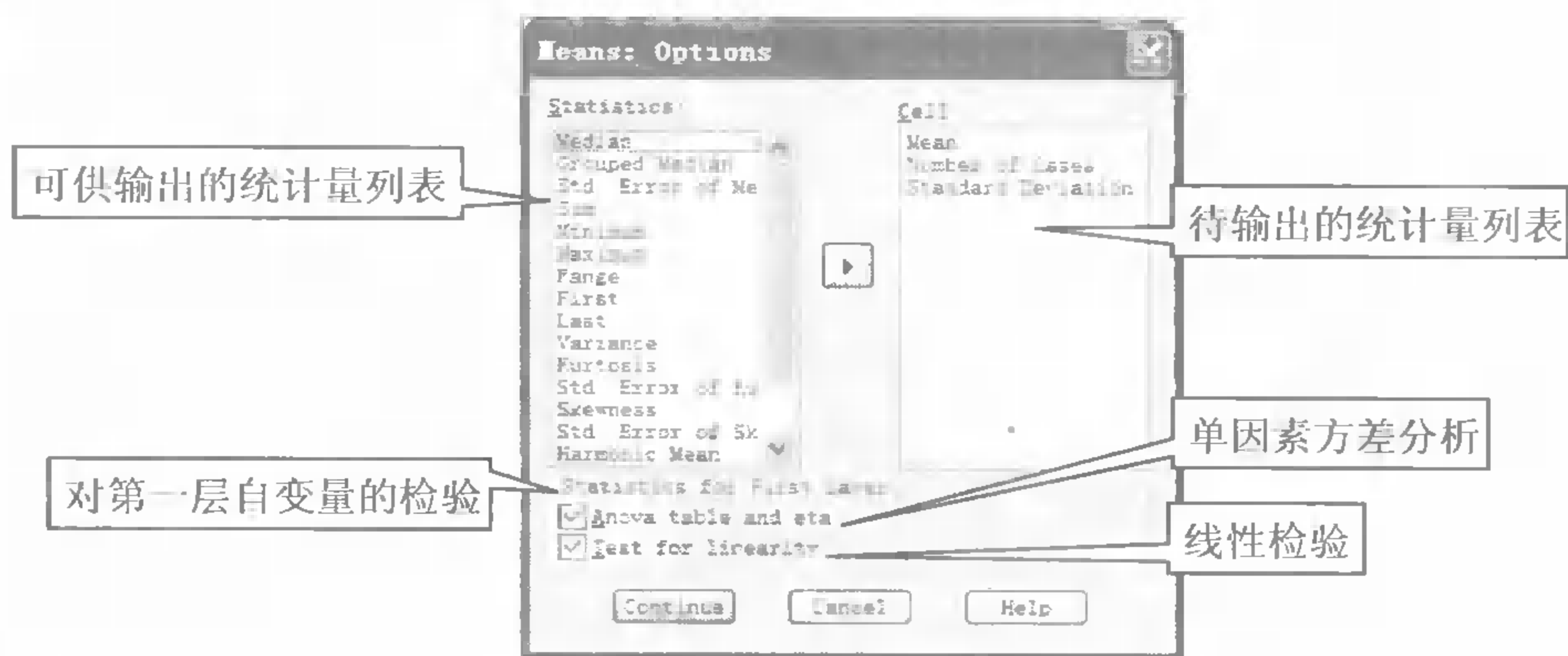


图 5-4 Means 过程 Options 设置面板

- 可供输出的统计量有：sum（总和）、number of cases（记录数）、mean（均值）、geometric mean（几何平均值）、harmonic mean（调和平均值）、standard deviation（标准差）、variance（方差）、standard error of the mean（均数的标准误）、median（中位数）、grouped median（频数表资料中位数）、minimum（最小值）、maximum（最大值）、range（全距）、kurtosis（峰度系数）、standard error of kurtosis（峰度系数的标准误）、skewness（偏度系数）、standard error of skewness（偏度系数的标准误）、percentage of total sum（总和的百分比）、percentage of total N（样本例数的百分比）等。
- Anova table and eta 复选框：对第一层中每一个分组变量进行单因素方差分析，并计算用于度量变量相关程度的 eta 统计量  $\eta$ 、eta Square 统计量  $\eta^2$ 。方差分析的零假设是：对第一层分类自变量的各个取值，因变量的均值都相等。
- Test for linearity 复选框：检验第一层变量的线性相关性，前提是因变量均值是第一层的自变量的线性函数。当自变量是短字符型时不予计算，可以输出平方和、自由

度、均方、F 检验的 F 值、 $R^2$  等统计量，其中  $R^2$  是检验线性拟和优度的统计量，只有在因变量为连续型、且分类的自变量有 3 个以上取值时才会计算。

(3) 输出结果。单击图 5-3 中的 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 5-5 所示。

案例处理摘要

	案例					
	已包含		已排除		总计	
	N	百分比	N	百分比	N	百分比
身高 * 性别 * 年龄	27	100.0%	0	0%	27	100.0%

报告

身高

性别	年龄	均值	N	标准差
女	10	1.4500	5	.02000
	11	1.5383	6	.02317
	12	1.6100	2	.01414
	总计	1.5154	13	.06253
男	10	1.4467	3	.02887
	11	1.5000	5	.04637
	12	1.6140	5	.01949
	13	1.5900	1	
	总计	1.5357	14	.07623
总计	10	1.4488	8	.02167
	11	1.5209	11	.03910
	12	1.6129	7	.01704
	13	1.5900	1	
	总计	1.5259	27	.06941

ANOVA 表<sup>a</sup>

	平方和	df	均方	F	显著性
身高 * 性别 组间 (组合)	.003	1	.003	5.69	.458
组内	122	25	.005		
总计	125	26			

a. 少于三组，无法计算 身高 \* 性别 的线性度量。

相关性度量

	Eta	Eta 方
身高 * 性别	.149	.022

图 5-5 学生身高对性别和年龄的均值比较输出

首先输出的是案例处理摘要表，给出了所用到的有效数据个数和比例。

然后是统计报告表，在此将性别和年龄的取值交叉组合后，分别给出了各组的均值、个数、标准差等统计值量。

最后是单因素方差分析表 (ANOVA) 和相关性度量表，由于年龄的取值少于 3 个，故不能给出其线性度量。

## 5.2 单样本 T 检验

### 5.2.1 原理与方法

单样本 T 检验，用于检验单个变量的均值与给定的常数（指定的检验值）之间是否存在显著差异，样本均值与总体均值之间的差异显著性检验，也属于单样本 T 检验。例如：关于某科目成绩的平均得分与 80 分的差异显著性的检验问题。

单样本 T 检验要求样本来自正态分布总体，它基本的理论计算步骤如下。

(1) 提出零假设与备择假设， $H_0: \mu = \mu_0$ ， $H_A: \mu \neq \mu_0$ ，其中  $\mu$  为样本所在总体的平均数的估计值， $\mu_0$  为已知的总体平均数。



(2) 计算  $t$  统计量, 公式为:  $t = \frac{\bar{x} - \mu_0}{S_x}$ , 自由度为:  $df = n - 1$ 。其中:  $n$  为样本量,  $S_x = \frac{S}{\sqrt{n}}$  为样本标准差。

(3) 由  $df = n - 1$  确定临界值  $t_{0.05}$  和  $t_{0.01}$ , 作出统计推断。若  $|t| < t_{0.05}$ , 不能否定零假设, 表明样本平均数  $\bar{x}$  与总体平均数  $\mu_0$  的差异不显著, 可以认为样本是取自该总体的; 若  $t_{0.05} \leq |t| < t_{0.01}$ , 则否定零假设, 表明样本平均数  $\bar{x}$  与总体平均数  $\mu_0$  的差异显著, 以 95% 的概率认为样本不是取自该总体; 若  $t_{0.01} \leq |t|$ , 否定零假设, 表明  $\bar{x}$  与  $\mu_0$  之间的差异极其显著, 以 99% 的概率认为样本不是取自该总体。

## 5.2.2 SPSS 实例分析


### 1. 数据描述

本节通过单样本 T 检验, 分析周岁儿童的实际身高与人们给出的经验值是否存在显著差异, 所用数据文件为“周岁儿童身高数据.sav”, 数据格式如图 5-6 所示。

	Name	Type	Width	Decima	Label	Values	Missing	Columns	Align	Measure
1	sg	Numeric	8	2	周岁儿童的身高	None	None	8	Right	Scale
2	cs	Numeric	8	2	城市标记	[1.00, 北京]	None	8	Right	Scale

图 5-6 周岁儿童身高数据

### 2. 参数设置

依次单击菜单“Analyze→Compare Means→One-Samples T Test”, 执行单样本 T 检验过程, 其主设置界面如图 5-7 所示。在变量列表单击选中身高 (sg) 变量, 单击  按钮, 将其选入 Test 列表框; 在底部的 Test 输入框键入“70”作为总体均值。单击 Options 按钮, 弹出如图 5-8 所示的设置面板, 单击 Continue 按钮返回主界面。

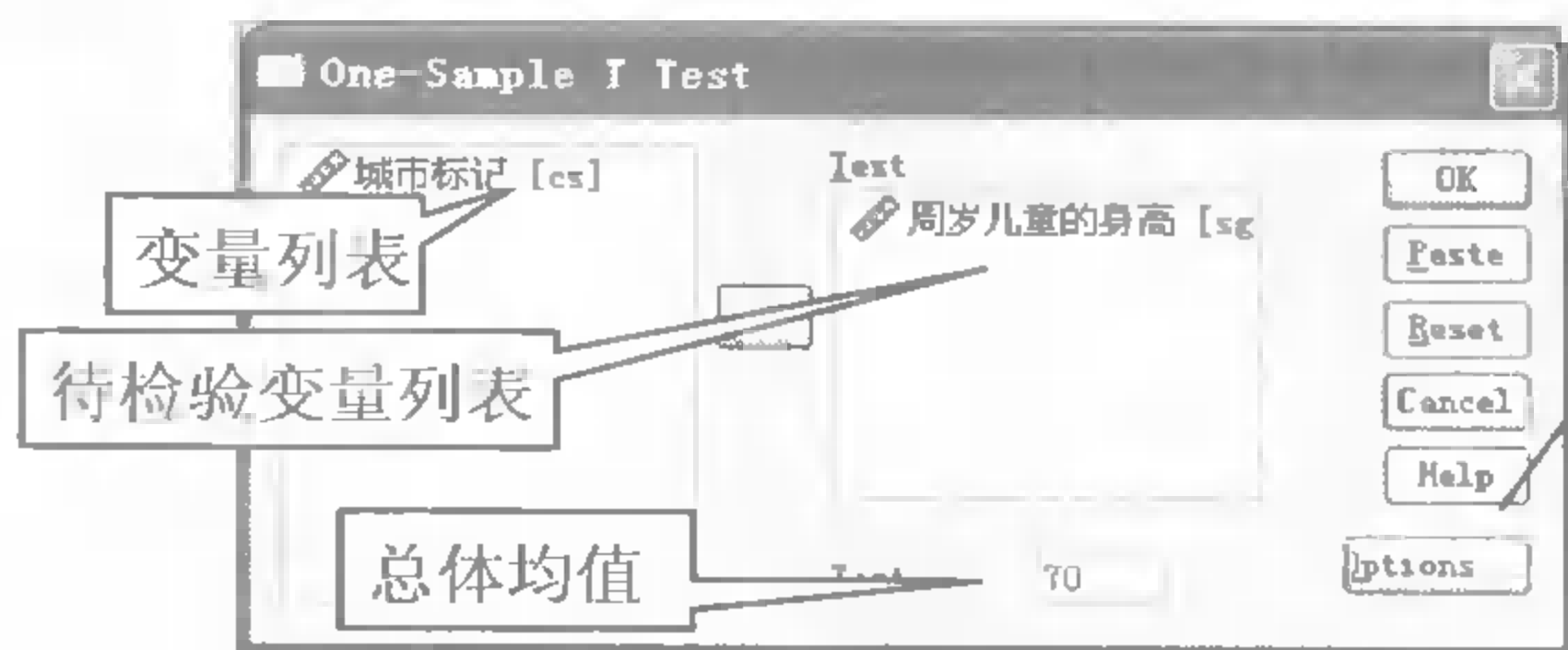


图 5-7 单样本 T 检验的主设置界面

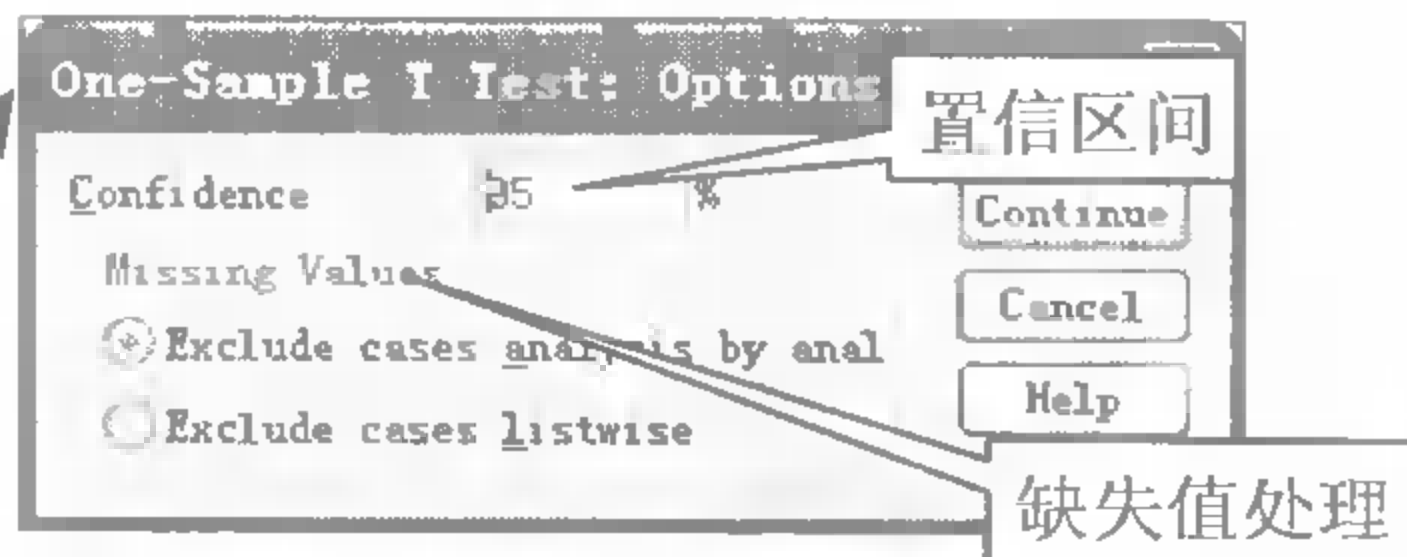


图 5-8 单样本 T 检验的选项设置

在图 5-7 中: Test 列表框, 用于从变量列表选入待检验的变量, 如果选入了多个, 它们都将对同一个总体均值进行检验; Test 输入框, 用于指定待检验的总体均值, 默认为 0。

在图 5-8 中: Confidence 输入框, 指定样本均值与总体均值之差的置信区间, 默认为 95%; Missing Values 栏用于设置缺失值的处理方式, 有如下 2 个选择。

- Excludes cases analysis by analysis: 若选入了多个待检验变量, 在检验某个变量时, 只忽略当前变量中含缺失值的记录, 因此每个变量所用的记录数可能不一样。默认选项, 能够充分地利用原始数据。
- Excludes cases listwise: 只要一个变量含缺失值, 则在所有分析中忽略这个记录。

3. 结果分析

在图 5-7 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出表格如图 5-9 所示。

单个样本统计量				
	N	均值	标准差	均值的标准误
周岁儿童的身高	21	71.8571	3.97851	.86818

单个样本检验						
	检验值 = 70					
	t	df	Sig. (双侧)	均值差值	差分的 95% 置信区间	
					下限	上限
周岁儿童的身高	2.139	20	.045	1.85714	.0461	3.6681

图 5-9 周岁儿童身高单样本 T 检验结果

“单个样本统计量”表格给出了关于样本的几个统计特征：样本量 (N)、均值等。

“单个样本检验”表格给出了 T 检验的结果，包括检验的总体均值 (70)、t 统计量 (2.139) 等信息。本例的双测 Sig 值为  $0.045 < 0.05$ ，故而认为在 0.05 的显著性水平下，测量身高与 70 有显著差异，也就是以 95% 的概率接受周岁儿童平均身高大于 70 的结论。但是，在 0.01 的显著性水平下 ( $0.045 > 0.01$ )，认为测量身高与 70 无显著差异，即不能以 99% 的概率否认周岁儿童平均身高等于 70 的结论。

5.3 两独立样本 T 检验

5.3.1 原理与方法

两独立样本的 T 检验，用于检验两个样本是否来自具有相同均值的总体。

例如：检验同龄农村大学生和城市大学生的平均身高是否具有显著差异；如果需要比较两种试验的结果，把试验单位随机地分成两组样本，然后对它们分别随机地施加一种处理，如此得到的两个试验样本是相互独立的。

两独立样本 T 检验所适用的数据格式，一般如表 5-1 所示。

表 5-1 两独立样本数据的一般形式

处 理	观测值 $x_{ij}$	样本含量 $n_i$	平均数 $\bar{x}$	总体平均数
1	$x_{11} \quad x_{12} \quad \dots x_{1n1}$	$n_1$	$\bar{x}_1 = \sum x_{1j} / n_1$	$\mu_1$
2	$x_{21} \quad x_{22} \quad \dots x_{2n2}$	$n_2$	$\bar{x}_2 = \sum x_{2j} / n_2$	$\mu_2$

两独立样本 T 检验的基本步骤如下。

(1) 提出无效假设  $H_0: \mu_1 = \mu_2$ ，和备择假设  $H_1: \mu_1 \neq \mu_2$ 。

(2) 计算 t 统计量，公式为： $t = \frac{\bar{x}_1 - \bar{x}_2}{S_{\bar{x}_1 - \bar{x}_2}}$ ，自由度为： $df = (n_1 - 1) + (n_2 - 1)$ 。 $S_{\bar{x}_1 - \bar{x}_2}$  为两样本

均值之差的标准误差，理论计算公式为： $S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2}{(n_1 - 1) + (n_2 - 1)} \times (\frac{1}{n_1} + \frac{1}{n_2})}$ ，当

$n_1 = n_2 = n$  时,  $S_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sum (x_1 - \bar{x}_1)^2 + \sum (x_2 - \bar{x}_2)^2}{n(n-1)}} = \sqrt{\frac{S_1^2}{n} + \frac{S_2^2}{n}} = \sqrt{S_{\bar{x}_1}^2 + S_{\bar{x}_2}^2}$ , 其中:  $n_1$ 、 $n_2$ ,

$\bar{x}_1$ 、 $\bar{x}_2$ ,  $S_1^2$ 、 $S_2^2$  分别为两个样本的样本量、样本均值、样本方差。

(3) 根据  $df=(n_1-1)+(n_2-1)$ , 确定临界  $t$  值  $t_{0.05}$  和  $t_{0.01}$ , 作出统计推断。

使用两独立样本的 T 检验时, 不仅要求两个样本相互独立, 而且要求它们的总体分布都必须为正态的。如果分组样本彼此不独立, 例如: 测量的是在长跑锻炼前后某项体能 (比如肺活量) 的成绩, 要求比较锻炼前后其均值是否有显著性差异, 应该使用配对样本 T 检验的功能 (Paired Sample T test)。如果分组样本不止两个, 应该使用一元方差分析过程 (One-Way ANOVA) 进行单变量方差分析。如果试图比较的样本变量的取值不服从正态分布, 应该考虑使用非参数检验过程 (Nonparametric test) 进行分析。如果想要比较的变量是分类变量, 则应该使用 Crosstabs ( $X^2$  检验) 功能。

5.3.2 SPSS 实例分析



1. 数据描述

本节通过两独立样本 T 检验过程, 分析某超市在促销前后的日销售额是否有显著变化, 所用数据文件为 “促销比较\_两独立样本 t 检验.sav”, 各变量含义如图 5-10 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	income	Numeric	4	2	日销售额 (万元)	None	None	8	Right	Scale
2	type	Numeric	8	0	类型	None	None	8	Right	Scale

图 5-10 超市促前后的日销售额数据

2. 参数设置

依次单击菜单 “Analyze→Compare Means→Independent-Samples Test”, 执行两独立样本 T 检验过程, 它的主设置界面与图 5-11 所示。在变量列表单击选中日销售额变量, 单击 Test 列表左侧的  按钮, 将其作为检验变量选入 Test 列表框; 在变量列表单击选中类型 (type) 变量, 单击 Grouping 栏左侧的  按钮, 将其作为分类变量选入 Grouping 选框。单击 Define Groups 按钮, 弹出如图 5-12 所示的对话框, 在两个 Group 输入框分别键入要分析的 type 变量的两个类别取值 “1 和 2”; 单击 Continue 按钮返回主界面。单击 Options 按钮, 弹出如图 5-8 所示的设置面板, 单击 Continue 按钮返回主界面。

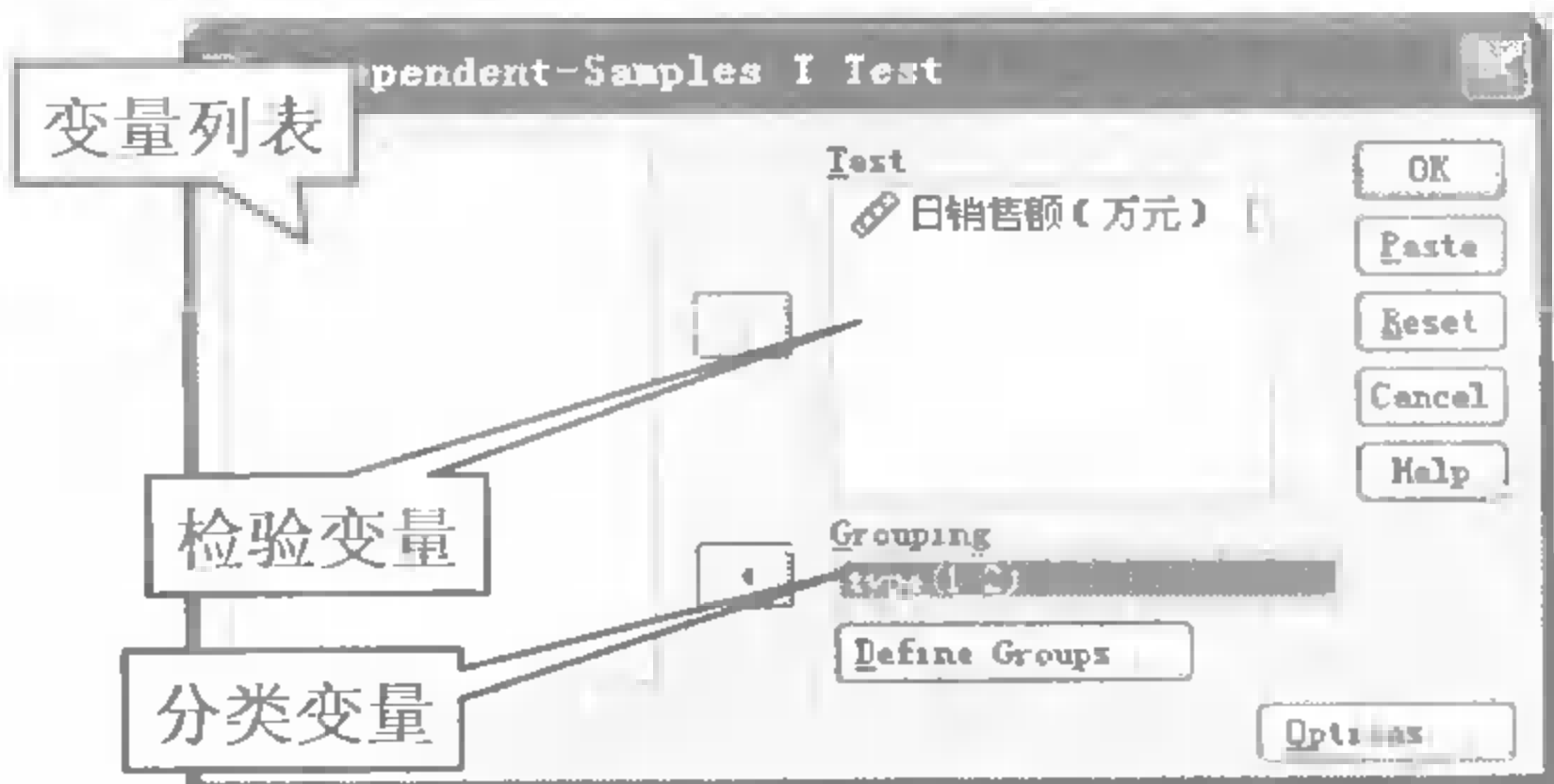


图 5-11 两独立样本 T 检验的主设置界面

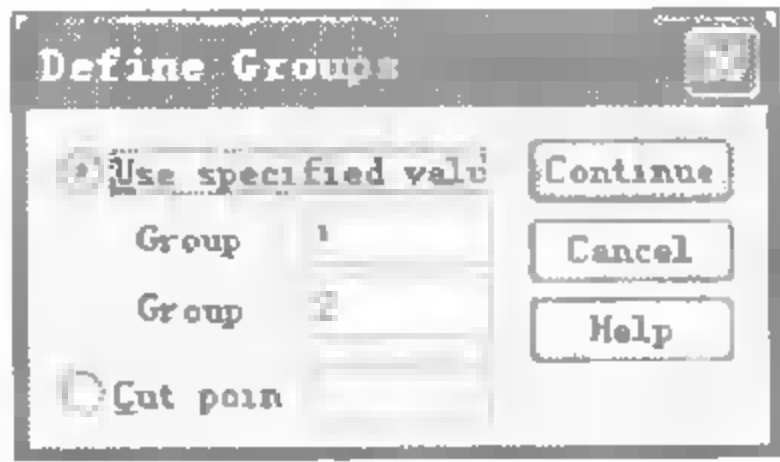


图 5-12 定义取值范围

在图 5-11 中: Test 列表框, 用于从变量列表选入待检验的变量; Grouping 选框, 用于选

入把数据分为两个类别的分类变量。

在图 5-12 中设置类别变量的取值，有如下两种可选方式。

① Use specified values 选项，根据特定取值定义两个类别，在下面的两个 Group 输入框指定类别变量的两个要进行检验的取值，数据中类别变量取其它值的记录将不记入分析，此处可以输入数字（整数或小数）或者字符。

② Cut point 选项，设定分解值，分类变量取值小于此处输入值的记录设为一类，其他记录设为另一类。如果指定的分类变量是字符型的，此项不可用。

### 3. 结果分析

在图 5-11 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出表格如图 5-13 所示。

组统计量					
	类型	N	均值	标准差	均值的标准误
日销售额(万元)	1	18	445.4444	132.44631	31.21790
	2	16	529.8750	138.20076	34.55019

独立样本检验									
		方差方程的 Levene 检验		均值方程的 t 检验					
		F	Sig.	t	df	Sig. (双侧)	均值差值	标准误差值	差分的 95% 置信区间
日销售额(万元)	假设方差相等	225	.638	-1.82	32	.078	-84.4	46.445	-179 10.17
	假设方差不相等			-1.81	31.16	.079	-84.4	46.565	-179 10.52

图 5-13 超市促销前后日销售额 T 检验结果

“组统计量”表格给出了促销前后日销售额的一些统计特征，包括：样本量（N）、均值、标准差等。

“独立样本检验”表格给出了关于方差齐性的 Levene 检验和关于均值相等的  $t$  检验结果。从 F 统计量的 Sig 值  $0.638 > 0.10$  看，不能否认方差相等的假设，所以应该参考第一行的  $t$  检验结果；第一行  $t$  检验的双侧 Sig 值  $0.078 < 0.10$ ，即在 0.10 的显著性水平上，认为促销能够显著地提高日销售额。

## 5.4 配对样本 T 检验

### 5.4.1 原理与方法

配对样本 T 检验，用于检验两个相关的样本（配对资料）是否来自具有相同均值的总体。配对数据来源的方式有两种：自身配对与同源配对。

（1）自身配对：指同一个试验对象，在二个不同时间上分别接受前、后两次处理，用其前后两次的观测值进行对照和比较；或者，对同一试验对象，取其不同部分的观测值或不同方法处理后的观测值进行自身对照和比较。例如：检验某种病畜治疗前后临床检查结果的变化；检验用两种方法测定的食物中药物残留量的区别。

（2）同源配对：指将来源相同、性质相同的两个个体配成一对，如将品种、性别、年龄、体重相同的两个试验动物配成一对，然后对配对的两个个体随机地实施不同处理，再根据所得的试验数据检验两种处理方法的效果。

配对样本 T 检验所适用的数据格式，一般如表 5-2 所示。

表 5-2 配对设计试验资料的一般形式

处 理	观测值 $x_{ij}$				样 本 含 量	样本平均数	总体平均数
1	$x_{11}$	$x_{12}$	...	$x_{1n}$	$n$	$\bar{x}_1 = \sum x_{1j} / n$	$\mu_1$
2	$x_{21}$	$x_{22}$	...	$x_{2n}$	$n$	$\bar{x}_2 = \sum x_{2j} / n$	$\mu_2$
$d_j = x_{1j} - x_{2j}$	$d_1$	$d_2$	...	$d_n$	$n$	$\bar{d} = \bar{x}_1 - \bar{x}_2$	$\mu_d = \mu_1 - \mu_2$

配对样本 T 检验的基本步骤如下：

(1) 提出无效假设  $H_0: \mu_d = 0$ ，和备择假设  $H_A: \mu_d \neq 0$ 。其中  $\mu_d$  为两配对样本的取值之差的总体平均数，它等于两样本所属总体的平均数  $\mu_1$  与  $\mu_2$  之差，即  $H_d = \mu_1 - \mu_2$ 。

(2) 计算  $t$  统计量，公式为： $t = \bar{d} / S_{\bar{d}}$ ，自由度为： $df = n - 1$ 。 $S_{\bar{d}}$  为两样本均值差的标准误，计算公式为： $S_{\bar{d}} = \frac{S_d}{\sqrt{n}} = \sqrt{\frac{\sum (d - \bar{d})^2}{n(n-1)}} = \sqrt{\frac{\sum d^2 - (\sum d)^2 / n}{n(n-1)}}$ ，其中： $d$  为两样本各对数据

之差，即  $d_j = x_{1j} - x_{2j}$ ，( $j=1, 2, \dots, n$ )； $\bar{d} = \sum d_j / n$ ； $S_d$  为  $d$  的标准差； $n$  为样本的样本量。

(3) 根据  $df = n - 1$  确定临界  $t$  值  $t_{0.05}(n-1)$  和  $t_{0.01}(n-1)$ ，作出统计推断。

## 5.4.2 SPSS 实例分析


### 1. 数据描述

本节通过配对样本 T 检验，分析喝减肥茶能否显著地起到减少体重的作用，所用数据文件为“减肥茶检验\_两配对样本 t 检验.sav”，各变量含义如图 5-14 所示。

	Name	Type	Width	Decima	Label	Values	Missing	Columns	Align	Measure
1	hcq	Numeric	8	2	喝茶前体重	None	None	8	Right	Scale
2	hch	Numeric	8	2	喝茶后体重	None	None	8	Right	Scale

图 5-14 减肥茶配对数据

### 2. 参数设置

依次单击菜单“Analyze→Compare Means→Paired-Samples Test”，执行两配对样本 T 检验过程，它的主设置界面如图 5-15 所示。在变量列表同时选中喝茶前体重 (hcq)、喝茶后体重 (hch)，单击  按钮，将其选入右侧的成对列表框；单击 Options 按钮，弹出如图 5-16 所示的设置面板，单击 Continue 按钮返回主界面。

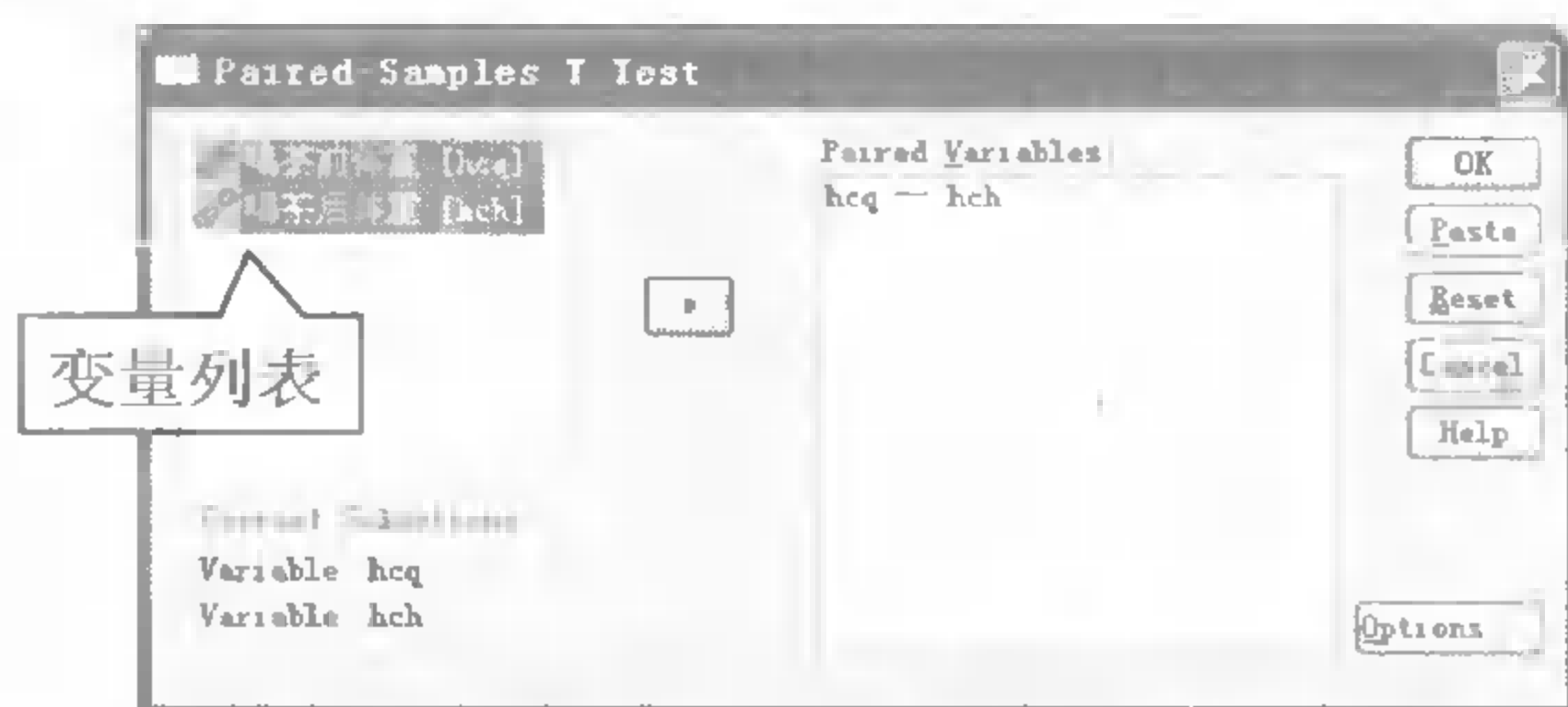


图 5-15 两配对样本 T 检验的主设置界面

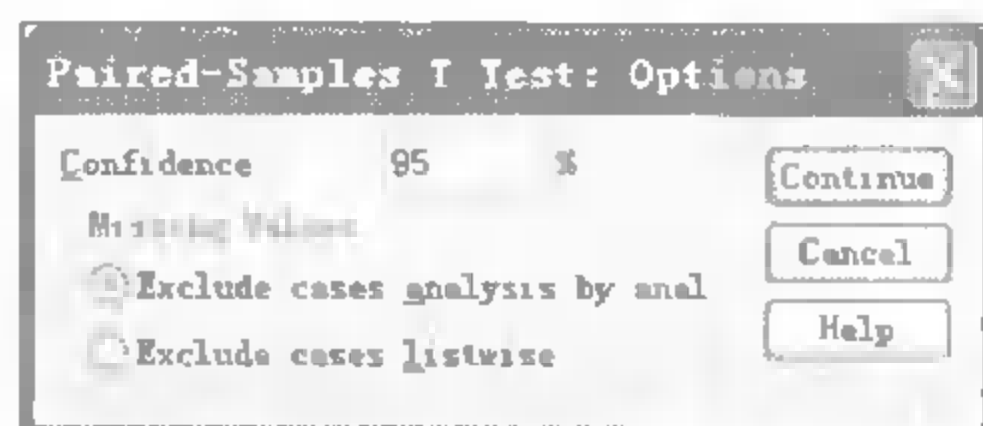


图 5-16 选项参数设置对话框



Paired-variables 列表框用于选入成对的变量组，例如本例的“**hcg -- hch**”，选入多对变量时，将分别对它们做配对检验。在选择检验变量组时，必须在变量列表同时选中配对的两个变量，再单击 **➤** 按钮进行选择。

3. 结果分析

在图 5-11 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出表格如图 5-17 所示。

成对样本统计量

	均值	N	标准差	均值的标准误
对 1 喝茶前体重	81.9167	12	11.61080	3.35175
喝茶后体重	74.8333	12	12.05920	3.48119

成对样本相关系数

对 1	N	相关系数	Sig.
喝茶前体重 & 喝茶后体重	12	.817	.001

成对样本检验

		成对差分				t	df	Sig. (双侧)	
		均值	标准差	均值的标准误	差分的 95% 置信区间				
					下限	上限			
对 1	喝茶前体重 - 喝茶后体重	7.08333	7.17899	2.07240	2.52202	11.64465	3.418	11	.006

图 5-17 减肥茶减肥效果检验结果

“成对样本统计量”表格给出了喝茶前后体重的一些统计特征，包括：样本量（N）、均值、标准差等。

“成对样本相关系数”表格给出了喝茶前后体重的相关性检验结果，可见在 0.01 的显著性水平上，认为喝茶前后的体重具有较强的相关性（相关系数 0.817）。

“成对样本检验”表格给出了关于差分的 t 检验结果，从 t 检验的双侧 Sig 值  $0.006 < 0.01$  看，认为这种减肥茶能够显著地减小人的体重。

## 第 6 章 非参数检验

非参数检验的内容十分丰富，主要有：卡方检验、二项分布检验、游程检验、单样本 K-S 检验、两个独立样本检验、多个独立样本检验、两个相关样本检验、多个相关样本检验。非参数检验方法不依赖于总体的分布，是在总体分布情况不明时，用来检验不同样本是否来自同一个总体的统计推断方法，由于这些方法一般不涉及总体参数而得名。为了便于读者掌握这些检验方法，本章结合大量的实例进行讲解。

在 SPSS 中进行非参数检验，主要通过选择主窗口菜单“Analyze→Nonparametric Tests”的子菜单执行指定的过程，可选方法有如下 8 个：

- Chi-square test (卡方检验)
- Binomial test (二项分布检验)
- Runs test (游程检验)
- 1-Simple K-S test (单样本柯尔莫哥洛夫-斯米诺夫检验)
- 2 Independent Samples test (两独立样本检验)
- K Independent Samples test (多个独立样本检验)
- 2 Related Sample test (两相关样本检验)
- K Related Sample test (多个相关样本检验)

### 6.1 非参数检验的简介

#### 6.1.1 非参数检验与参数检验

非参数检验是相对于参数检验而言的，这两种检验方法在实际中都有广泛的应用，但它们有着不同的数理统计原理和应用场合。

在统计学的发展过程中，最先出现的推断统计方法都对样本所属总体的性质作出若干假设，即对总体的分布形状作某些限定，例如 Z 检验、t 检验，假设样本的总体是正态分布的，或者假设两个样本都取自具有相同方差的总体。这类方法对总体的分布加以某些限定，把所要推断的总体数字特征看作未知的“参数”进行推断，称之为参数统计方法 (parameter statistical methods) 或限定分布统计方法 (distribution-specified statistical methods)，基于此所做的假设检验就称为参数检验 (parametric test)。常用的统计检验如 t 检验、Z 检验、F 检验等都是参数检验。

参数检验只有在关于总体分布的假设成立时，所得出的结论才是正确的，所以它在很多

场合下不便应用，于是统计学家发展了许多对总体不作太多或严格限定的统计推断方法，这些方法一般不涉及总体参数的假设，与之相对应的统计方法通常称为非参数统计（nonparametric statistics）或自由分布统计方法（distribution-free statistical methods），基于此所做的假设检验则称为非参数检验（nonparametric test）或自由分布统计检验（distribution-free statistical test）。非参数检验的前提假设比参数检验方法少很多，也容易满足，适用于已知信息相对较少的数据资料，而且它的计算方法也简便易行。

对于多数参数检验方法，都有一种或几种相对应的非参数检验方法，如表 6-1 所示。

表 6-1 参数检验与非参数检验方法的对应表

参数检验方法	非参数检验方法
$t$ 检验法	两个独立样本的中位数检验
	两个独立样本的秩和检验
$t$ 检验法（配对样本）	成对比较、单样本正负号检验
	成对比较、单样本符号秩检验
单因素方差分析	K 个独立样本的 H 检验法
多因素方差分析	Friedman 检验法
相关系数	Spearman 秩相关系数

6.1.2 非参数检验的优点

与参数检验方法对比，非参数检验方法具有以下优点。

- （1）检验条件宽松，适应性强。参数检验假定总体分布为正态、近似正态或以正态分布为基础而构造的  $t$  分布或  $\chi^2$  分布；非参数检验不受这些条件的限制，弥补了参数检验的不足，对于非正态的、方差不等的以及分布形状未知的数据都适用。
- （2）检验方法灵活，用途广泛。非参数检验不但可以应用于定距、定比等连续变量的检验，而且适用于定类、定序等分类变量的检验。对于那些不能直接进行四则运算的定类数据和定序数据，运用符号检验、符号秩检验都能起到好的效果。
- （3）非参数检验的计算相对简单，易于理解。由于非参数检验更多地采用计数的方法，其过程及结果都可以被直观地理解，为使用者所接受。

6.1.3 非参数检验的缺点

非参数检验也有一些不可避免的缺点。

- （1）非参数检验方法对总体分布的假定不多，适应性强，但方法本身也就缺乏针对性，其功效不如参数检验。
- （2）非参数检验使用的是等级或符号秩，而不是实际数值，方法虽简单，但会失去许多信息，因而检验的有效性也就比较差。例如对于一批适用于  $t$  检验的配对资料，如果采用符号秩检验处理，其功效将低于  $t$  检验，如果用符号检验处理则效率更低，因为它对信息的利用更不充分。当然，如果假定的分布不成立，那么非参数检验就是更值得信赖的。

6.2 卡方检验

在某些统计方法中，往往事先假定总体服从正态分布，然后再对其均值或方差作检验，

但某个随机变量是否服从某种特定的分布是需要进行检验的。卡方检验（Chi-square test）就是一种用来检验给定的样本数据是否来自特定分布的方法。

检验的过程，通常是先根据以往的经验或实际观测数据的分布情况，推测总体服从某种分布，分布函数为  $F(x)$ ，然后再利用样本数据检验该总体的分布函数是否就是  $F(x)$ 。

6.2.1 原理与方法

卡方检验的零假设  $H_0$  为：样本所属总体的分布与指定的理论分布无显著差异。卡方检验直接检验的是实际频数与指定分布的频数是否相符。

1. 卡方统计量  $\chi^2$

$\chi^2$  统计量：
$$\chi^2 = \sum_{i=1}^k \frac{(f_{oi} - f_{ei})^2}{f_{ei}}$$
其中  $k$  是样本分类的个数， $f_{oi}$  表示实际观察到的频数， $f_{ei}$  表示指定理论分布下的频数。观察频数与理论频数越接近，则  $\chi^2$  值越小，根据皮尔逊定理，当  $n$  充分大时， $\chi^2$  统计量渐近服从于  $\chi^2(k-1)$  分布。

根据给定的显著性水平  $\alpha$  和卡方分布的自由度确定检验的临界值  $\chi^2_\alpha$ 。如果  $\chi^2 < \chi^2_\alpha$ ，则不能拒绝  $H_0$ ，即认为样本所属的总体分布与指定的分布无显著差异；反之亦然。

由于奠定  $\chi^2$  检验基础的皮尔逊定理，要求样本量是充分大的，所以使用时建议样本容量应不小于 30，同时每个单元中的期望频数不能太小，如果有类别的频数小于 5，则建议将它与相邻的类别合并，如果有 20% 的单元期望频数都小于 5，就不能再使用  $\chi^2$  检验了。

2. 拟合优度检验

利用随机样本资料，对总体是否服从某种理论分布的检验，检验步骤如图 6-1 所示。

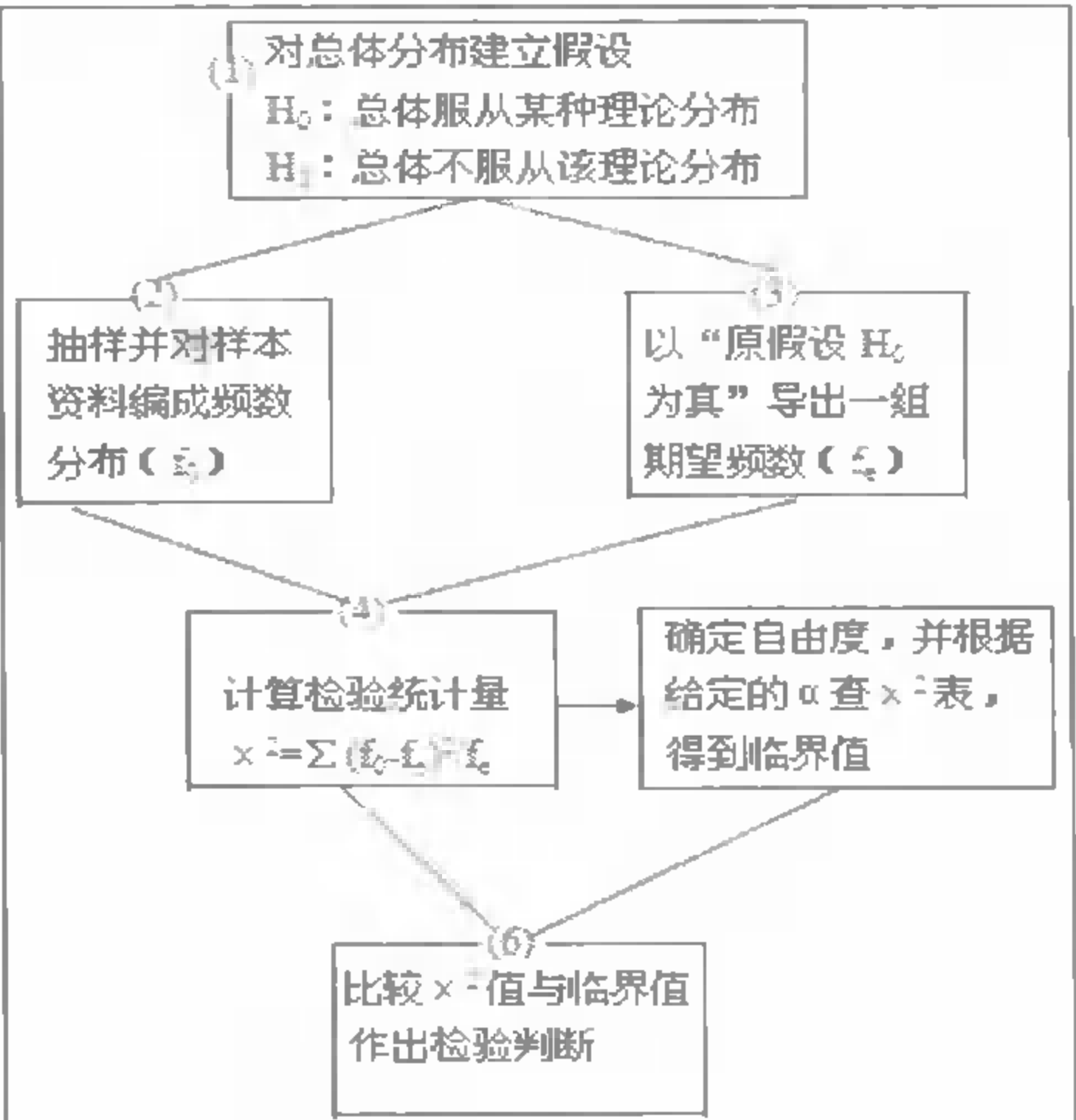


图 6-1 拟合优度检验步骤

3. 独立性检验

利用样本数据，判断总体的两个变量是否彼此独立的检验，检验步骤如图 6-2 所示，其中  $\chi^2$  分布的自由度为： $df = (r-1)(c-1)$ 。

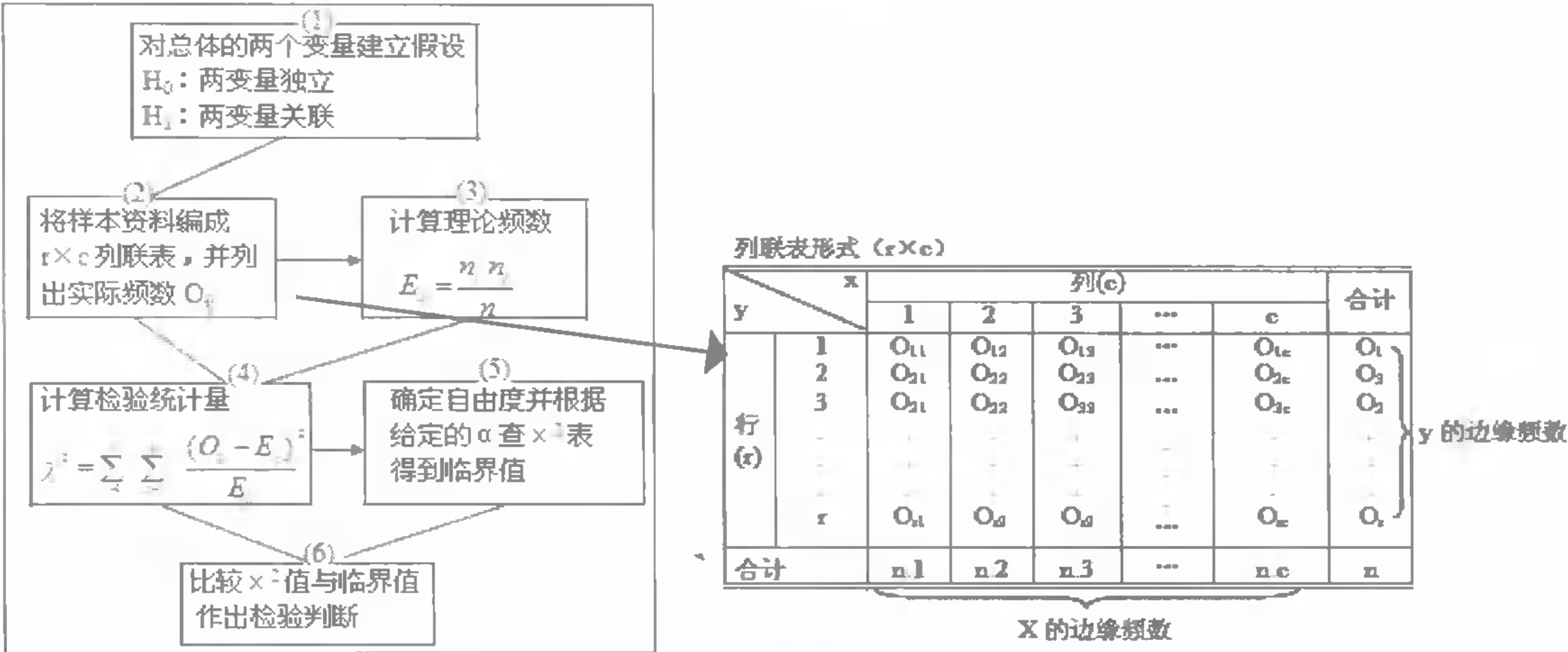


图 6-2 独立性检验步骤

6.2.2 数据和问题描述

本节利用某企业的生产线, 在星期一至星期五产生的不合格产品数量, 检验五个不同工作日的产品不合格率是否相同, 所用数据文件为“不合格产品数量卡方检验.sav”, 数据格式如图 6-3 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	weekday	Numeric	8	2	星期	{1 00 星期1}	None	8	Right	Ordinal
2	count	Numeric	8	2	不合格个数	None	None	8	Right	Scale

图 6-3 关于不合格产品的数据格式

本例检验的假设是,  $H_0$ : 样本所属总体的分布是均匀分布;  $H_1$ : 样本所属总体的分布不是均匀分布。

6.2.3 卡方检验实例分析

依次单击菜单“Analyze→Nonparametric Tests→Chi-square test”, 执行卡方检验过程, 其主设置界面如图 6-4 所示。

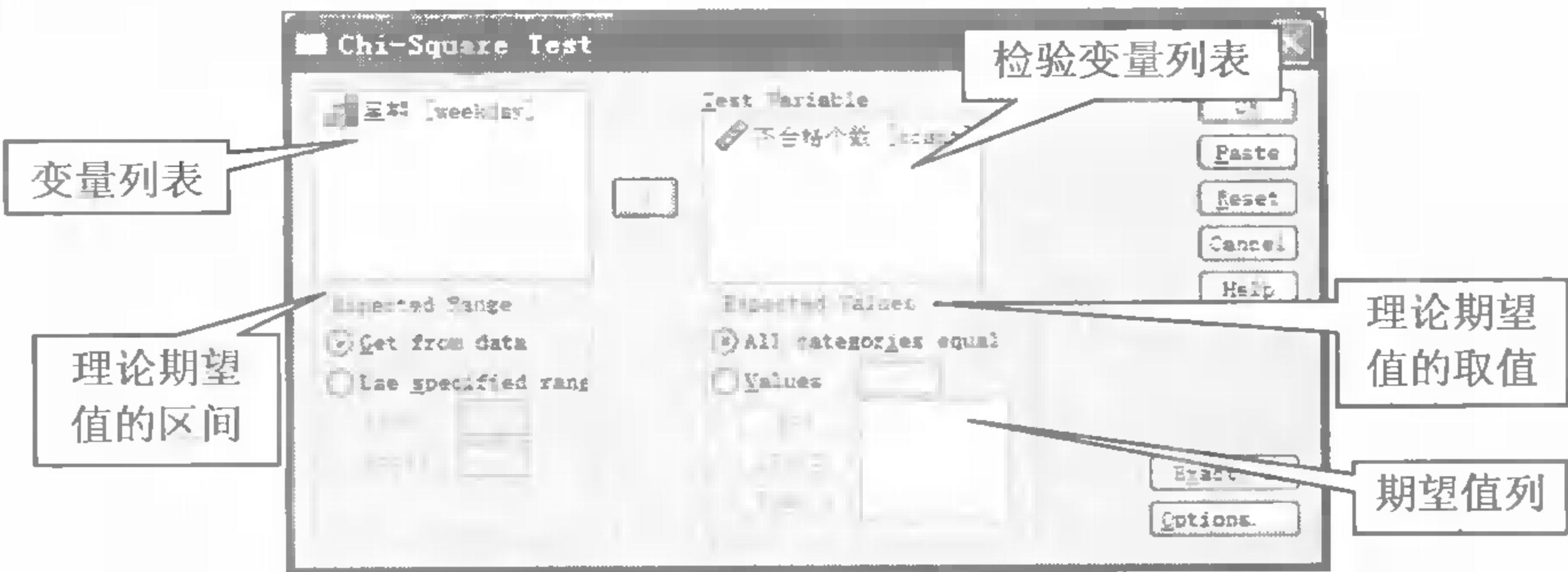


图 6-4 卡方检验的主设置界面

1. 变量设置

在变量列表单击选中不合格个数变量, 单击 按钮, 将其作为检验变量选入 Test 列表。

① Test Variable 列表用于从变量列表选入检验变量, 必须为数值型的分类变量, 若选入



多个，将分别单独处理。

② Expected Range 栏设置检验变量取值的区间范围，有两个可选方式：

- Get from data，表示检验变量每个唯一的取值都作为一个类别，默认选项；
- Use specified range，由用户设置特定的范围，需要在 Lower、Upper 输入框中分别指定检验变量的最小、最大取值，超过这个范围的观测将忽略不计。

③ Expected Values 栏设置待检验理论期望值的具体取值，有两个可选方式：

- All categories equal，表示每个类别的期望取值都相等，即检验样本是否为均匀分布，默认选项；
- Values 选项，由用户设置特定的期望值，先在右侧的输入框指定一个期望值，然后通过单击按钮 Add、Change、Remove 来添加、改变、删除指定期望值的取值；这里输入的顺序是非常重要的，每个新输入的期望值自动显示在期望值列表的底部，而列表中的期望值以从上至下的顺序，对应了样本取值从小到大的顺序，例如期望值列表中的第一个取值对应的是样本中取值最小的观测。

## 2. Options 选项设置

在图 6-4 中单击 Options 按钮，弹出如图 6-5 所示的设置面板，单击 Continue 按钮返回主界面。

- Statistics 栏选择要输出的统计量，可选项有两个：Descriptive 描述性统计量，包括均值、标准差、最大最小值、无缺失数据的观测数等；Quartiles 四分位数。
- Missing Values 栏设置缺失值的处理原则，有两个选项：Exclude cases test-by-test，当有多个待检验变量时，只忽略当前检验变量含缺失值的观测记录；Exclude cases listwise，若某观测的任一变量含有缺失值，则所有检验过程都忽略此观测记录。

## 3. 精确检验的参数设置

在图 6-4 中，单击 Exact 按钮，弹出如图 6-6 所示的精确检验子设置对话框，单击 Continue 按钮返回主界面。

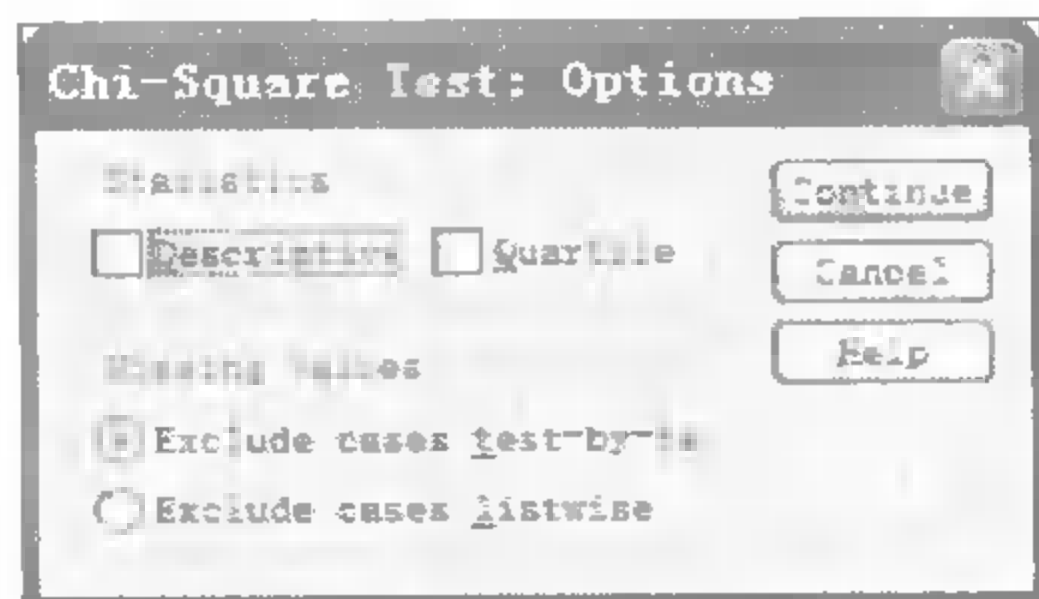


图 6-5 卡方检验的选项设置面板

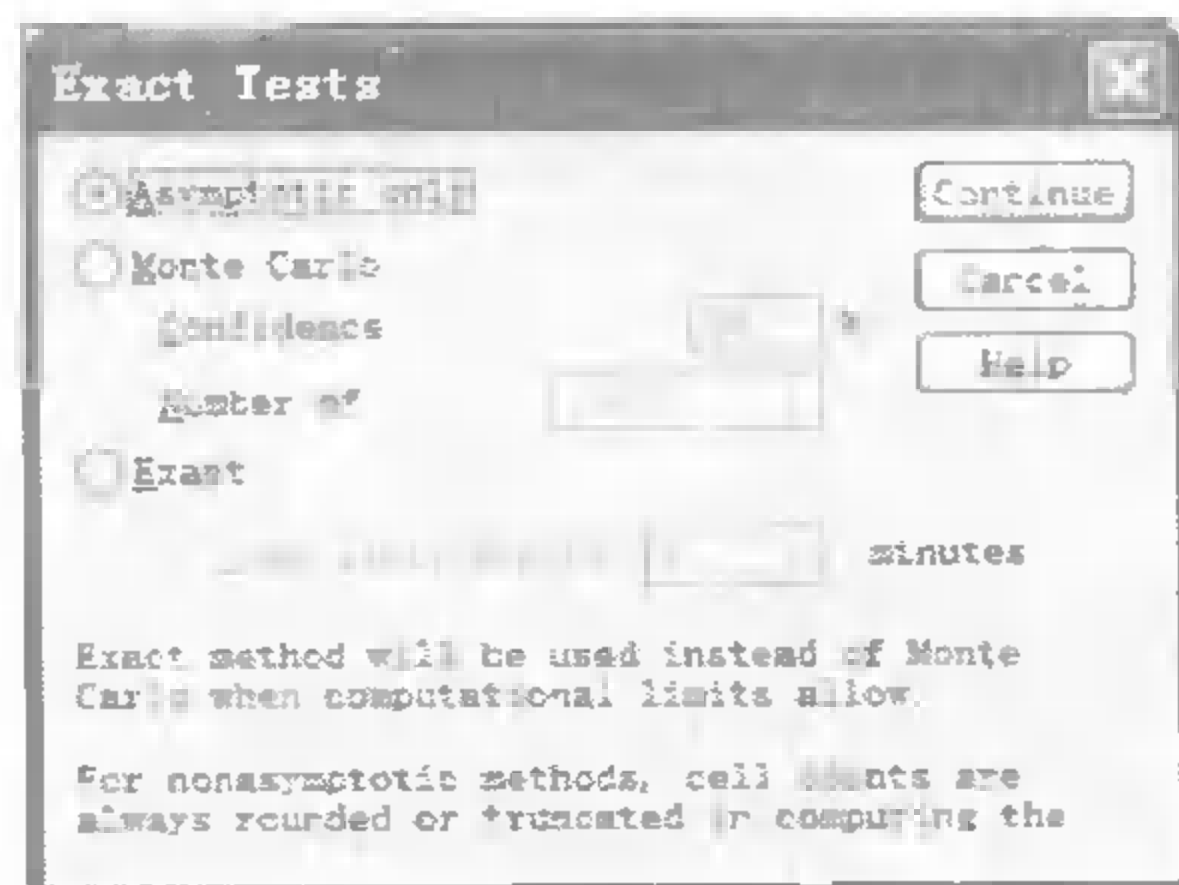


图 6-6 非参数建议精确检验设置面板

当应用卡方检验的前提条件不满足时，例如有多于 20% 的单元期望频数小于 5 时，可以在此界面设置采用其他的检验方法，包括：Exact 精确检验和 Monte Carlo 方法。

## 4. 输出结果

在图 6-4 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出表格如图 6-7 所示。

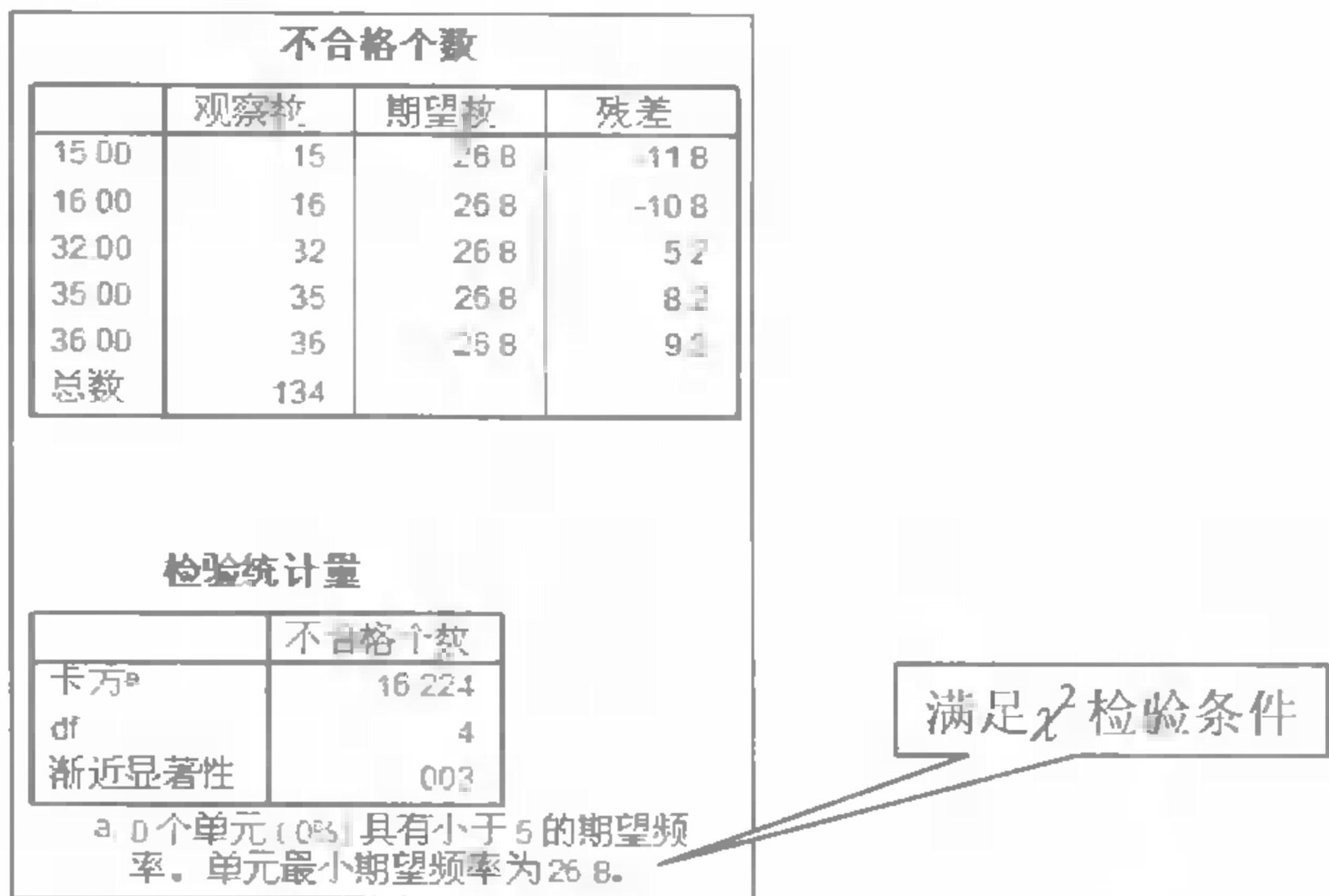


图 6-7  $\chi^2$  检验输出

由于渐进显著性（ $\chi^2$  检验显著性）的取值  $0.003 < 0.01$ ，故而在 0.01 的显著性水平上否定零假设，即认为五个工作日内各天的产品不合格率是不相同的。

6.3 二项检验

在处理实际问题时，有些数据的取值只能划分为两类，比如：医学中的生与死、患病的有与无。从这种二分类总体中抽取的样本，要么是对立分类中的这一类，要么是另一类，其频数分布服从二项分布。二项检验（Binomial test），就是一种用来检验样本是否来自参数为  $(n, p)$  的二项分布总体的方法。

6.3.1 原理与方法

二项检验通过对二值变量的单个取值作检验，能够判断总体中两个类别个体的比例是否分别为  $p$  和  $1-p$ 。参数为  $(n, p)$  的二项分布满足： $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ ，其均值和标准差分别为  $np$  和  $\sigma_X = \sqrt{np(1-p)}$ 。其中  $P(X = k)$  表示所占比例为  $p$  的类别出现  $k$  次的概率。

SPSS 中二项检验的统计量定义为： $p_1 = (n_1 - np) / \sigma_X$ ，其中  $n_1$  为第一个类别的样本个数。随着样本量  $n$  的增大， $p_1$  渐进趋于正态分布，故可用正态分布来检验统计量  $p_1$  的显著性。

6.3.2 数据和问题描述

利用某企业生产某种产品的合格率数据，检验其合格产品数量是否服从参数为  $p$  的二项分布。所用数据文件为“产品合格率二项检验.sav”，数据格式如图 6-8 所示。

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1 是否合格	Numeric	8	2		00 不合格	None	8	Right	Scale

图 6-8 产品合格率二项检验数据

本例检验的原假设为  $H_0$ ：合格产品比率等于 0.8；备则假设为  $H_1$ ：合格产品比率不等于 0.8。

6.3.3 二项检验实例分析

依次单击菜单“Analyze→Nonparametric Tests→Binomial...”，执行二项检验过程，其主设置界面如图 6-9 所示。



的重要程度会远远大于总体参数的重要程度。

游程检验的目的，就是检验取值为二分类并且按某种顺序（例如时间顺序）排列的数据资料，是否确实是随机出现的。

6.4.1 原理与方法

所谓游程，是指二分类变量有相同取值的几个连续记录。以投硬币试验为例，假设以 1 表示正面，0 表示反面，在进行了若干次投掷后，将得到一个以 1、0 组成的数据序列，如：11100110110001，最前面的三个 1 为一个游程（run），游程的长度为 3；随后的两个 0 为第二个游程，游程长度为 2，依次类推，这个序列包含七个游程。把出现 1 的次数记作  $n_1$ ，出现 0 的次数记作  $n_2$ ，则有  $n_1 + n_2 = n$ ；游程的个数记为  $r$ ，它是游程检验的基本统计量。

如果游程的总数极少，就意味样本内部存在着一定的趋势或结构，这可能是由于观察值之间是不独立的，或者样本是来自不同总体的，极端的数据序列可能是：1111111100000000；若样本中存在极大量的游程，则可能有系统的短周期波动影响着观察结果，同样不能认为序列是随机的，极端的数据序列可能是：1010101010101010。因此，出现太少或太多的游程都表明相应变量值的出现不是随机的。

1. 游程统计量

取游程检验统计量  $Z = \frac{r - E(r)}{\sigma_r}$ ，其中： $E(r) = \frac{2n_1n_2}{n_1 + n_2} + 1$ ， $\sigma_r = \sqrt{\frac{2n_1n_2(2n_1n_2 - n_1 - n_2)}{(n_1 + n_2)^2(n_1 + n_2 - 1)}}$

在游程检验中，将样本各个观测归属于两种类别之中，于是各观测出现的分布服从二项分布，随着样本量的增大，游程个数  $r$  的分布近似服从于正态分布。

游程检验的原假设  $H_0$  为：检验变量的取值是随机出现的。根据统计量  $Z$  的取值，与标准正态分布在特定显著性水平下的取值比较，作显著性检验，即可接受或推翻原假设。

2. 游程检验典型问题

- （1）检验两个总体的分布是否相同。将从两个总体中独立抽取的两个样本的观察值混合后，记录游程个数，进行关于随机性的假设检验。
- （2）检验样本的随机性。将取自某一总体的样本观察值按照从小到大的顺序排列，找出中位数（或平均数），把样本分为大于和小于中位数的两个部分，用由这两个部分上下交错形成的游程个数来检验样本是否是随机的。

6.4.2 数据和问题描述

本节利用投掷硬币 20 多次所得的数据，来验证出现正面、背面的几率是否随机，出现正面记为 1，出现背面记为 0。所用数据文件为“投硬币试验.sav”，数据格式如图 6-11 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	records	Numeric	8	0	记录	{0 背面}	None	8	Right	Scale

图 6-11 投掷硬币数据格式

注意：数据输入的顺序不可以改变，否则会改变数据的游程数，从而使得检验结果不可信。本例检验的原假设是  $H_0$ ：出现正面、背面的几率是随机的；备则假设是  $H_1$ ：出现正面、

背面的几率不是随机的。

### 6.4.3 游程检验实例分析

依次单击菜单“Analyze→Nonparametric Tests→run...”，执行游程检验过程，其主设置界面如图 6-12 所示。

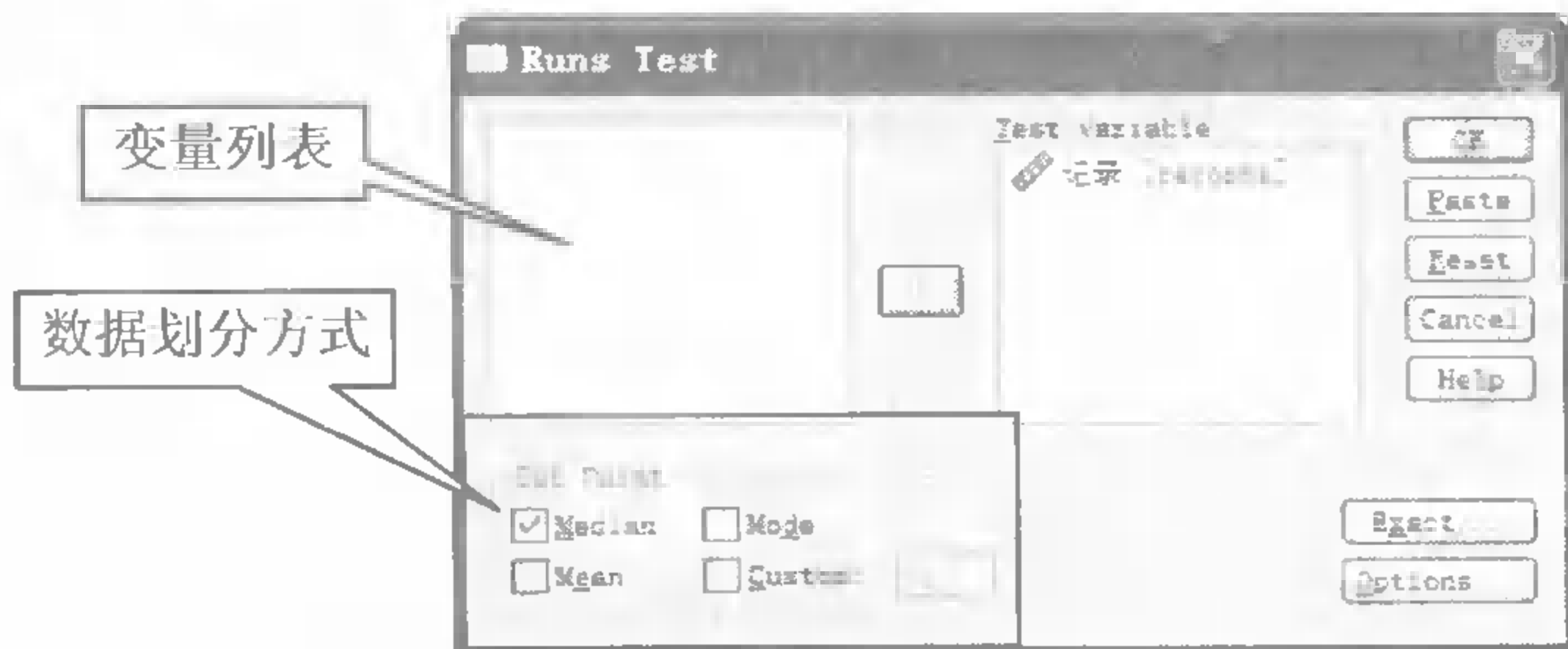


图 6-12 游程检验设置面板

#### 1. 参数设置

在左侧的变量列表单击选中记录变量，单击 按钮，将其作为检验变量选入 Test Variable 列表。

- Test Variable 列表，用于从左侧的变量列表选入检验变量，且必须为数值型分类变量，若同时选入了多个，将分别单独处理。
- Cut Point 栏，设置把数据划分为两个类别的临界点，有 4 种方式：Median 样本中位数（默认选项）；Mode 样本众数；Mean 样本均值；Custom 用户自定义，在后面的输入框指定任意的临界值。样本取值小于此处设定值的归为一类，其他的归为另一类；若选择了多个 Cut Point，则每个待检验变量将分别对每个 Cut Point 的取值单独做一次统计检验。

单击 Options 按钮，将弹出如图 6-5 所示的选项设置对话框；单击 Exact 按钮，将弹出如图 6-6 所示的精确检验设置对话框。其设置选项的含义和设置方法同前。

#### 2. 结果分析

在图 6-12 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出表格如图 6-13 所示。由于渐进显著性的取值  $0.816 > 0.10$ ，故不能否定零假设，即认为出现正面、背面的几率是随机的。

Runs 检验	
	记录
检验值	1
案例 < 检验值	12
案例 ≥ 检验值	14
案例总数	26
Runs 数	15
Z	232
渐近显著性(双侧)	816
a. 中值	

图 6-13 游程检验的结果输出

## 6.5 Kolmogorov-Smirnov 单样本检验

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



名的，它是一种拟合优度检验，用来研究样本观察值的分布和指定的理论分布是否吻合。K-S 检验通过对两个分布之间的差异的分析，判断样本的观察结果是否来自指定分布的总体。

6.5.1 原理与方法

K-S 检验的基本思路是：先将顺序分类资料数据的理论累计频率分布，同观测的经验累计频率分布加以比较，求出它们最大的偏离值，然后在给定的显著性水平上检验这种偏离值是否是偶然出现的。

1. 统计量

设  $S_n(x)$  是随机样本观察值的累积概率分布函数，即经验分布函数，样本量为  $n$ ； $F_0(x)$  是一个特定的累积概率分布函数，即理论分布函数。定义  $D = |S_n(x) - F_0(x)|$ ，如果对于每一个  $x$  值， $S_n(x)$  与  $F_0(x)$  都十分接近，则表明经验分布函数与理论分布函数的拟合程度很高，有理由认为样本数据来自服从该理论分布的总体。

K-S 检验主要考察的是绝对值  $D = |S_n(x) - F_0(x)|$  中那个最大的偏差，即使用如下的统计量做检验： $D_{\max} = \max |S_n(x) - F_0(x)|$ 。

2. K-S 检验的步骤

- (1) 提出假设： $H_0 : S_n(x) = F_0(x)$ ， $H_1 : S_n(x) \neq F_0(x)$ 。
- (2) 计算统计量  $D_{\max}$ 。
- (3) 根据给定的显著性水平  $\alpha$  和样本数据个数  $n$ ，确定单样本 K-S 检验的临界值  $D_\alpha$ 。
- (4) 若  $D_{\max} < D_\alpha$ ，则在  $\alpha$  的显著性水平上，不能拒绝  $H_0$ ；否则，拒绝  $H_0$ 。

3. 卡方检验与 K-S 检验的比较

这两者都是拟合优度检验， $\chi^2$  检验常用于分类数据，而 K-S 检验还可以运用于顺序数据。当预期频数较小时， $\chi^2$  检验需合并邻近的类别才能计算，K-S 检验不需要如此，因而它能比  $\chi^2$  检验保留更多的信息。

6.5.2 数据和问题描述

本节使用文件“儿童身高体重检验.sav”提供的数据，检验 10~13 岁儿童的身高和体重数据是否服从正态分布，数据格式如图 6-14 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	no	String	4		编号	None	None	4	Right	Nominal
2	gend	String	4		性别	{0 男}	None	4	Right	Nominal
3	age	Numeric	3	0	年龄	None	None	3	Right	Scale
4	high	Numeric	4	2	身高	None	None	8	Right	Scale
5	weight	Numeric	2	0	体重	None	None	6	Right	Scale

图 6-14 儿童的身体特征数据

本例检验的原假设为  $H_0$ ：儿童的身高（体重）服从正态分布；备则假设为  $H_1$ ：儿童的身高（体重）不服从正态分布。

6.5.3 K-S 单样本检验实例分析

依次单击菜单“Analyze→Nonparametric Tests→1-Simple K-S”，执行 K-S 单样本检验过

程，其主设置界面如图 6-15 所示。

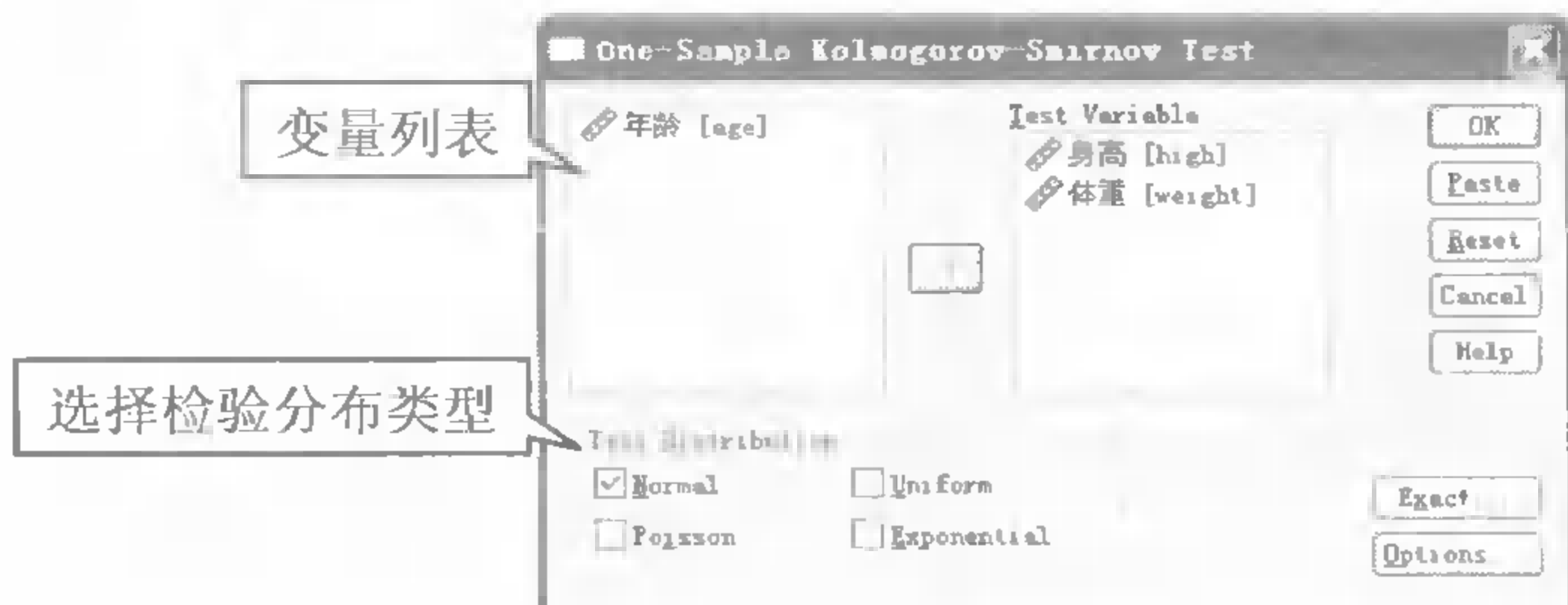


图 6-15 K-S 检验设置面板

## 1. 参数设置

在左侧的变量列表选中身高和体重变量，单击 按钮，将其作为检验变量选入 Test Variable 列表。

- Test Variable 列表，用于从左侧的变量列表选入检验变量，且必须为数值型分类变量，若同时选入了多个，将分别单独处理。
- Test Distribution 栏，设置要检验的理论分布类型，有如下 4 种分布：Normal 正态分布（默认选项）、Uniform 均匀分布、Poisson 分布、Exponential 指数分布。SPSS 假设这些分布的参数都是预先确定的，并且将通过样本估计取得这些参数。

单击 Options 按钮，将弹出如图 6-5 所示的选项设置对话框；单击 Exact 按钮，将弹出如图 6-6 所示的精确检验设置对话框。其设置选项的含义和设置方法同前。

## 2. 结果分析

在图 6-15 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出表格如图 6-16 所示。

单样本 Kolmogorov-Smirnov 检验			
		身高	体重
N		27	27
正态参数 <sup>a</sup>	均值	152.59	45.30
	标准差	0.341	6.960
最极端差别	绝对值	.154	.166
	正	.153	.166
	负	-.154	-.125
Kolmogorov-Smirnov Z		.801	.865
渐近显著性(双侧)		.542	.443

a. 检验分布为正态分布。  
b. 根据数据计算得到。

图 6-16 K-S 检验的结果输出

由于身高和体重的双侧渐进显著性取值都大于 0.10，故不能否定零假设，即认为儿童的身高和体重都是服从正态分布的。

## 6.6 两独立样本检验

有时样本所属的总体分布类型是未知的，但用户还是想判断在这种情况下两个独立的样本是否来自相同分布的总体，两独立样本检验（test for two independent samples）就是用来处理此类问题的一种有效方法。

### 6.6.1 原理与方法

两独立样本检验，通过对两个独立样本的均值、中位数、离散趋势、偏度等进行差异性检验，分析它们是否来自相同分布的总体。SPSS 提供了 4 种检验方法。

(1) Mann-Whitney U 检验。Mann-Whitney U 检验等同于对两组数据的 Wilcoxon 秩和检验和 Kruskal-Wallis 检验，它检验两个样本的总体在某些位置上是否相等。

检验的基本思路是，将来自两个组的样本合并后赋予秩，有结（取值相同的情况）的样品被赋予平均秩，结的数量相对于观测的数量应该是较少的。如果总体在位置上是相同的，则秩应该被随机的混合在两个样本里，计算第一组样本中每个数据的秩大于第二组样本中每个数据的秩的次数，再计算第二组样本中每个数据的秩大于第一组样本中每个数据的秩的次数，Mann-Whitney U 检验取这两个次数中较小的一个进行分析。SPSS 的 Mann-Whitney U 检验过程，同时会输出关于秩和较小的样本的 Wilcoxon 秩和统计量  $W$ 。

(2) Kolmogorov-Smirnov Z 检验和 Wald-Wolfowitz runs 检验。这两个检验是更普通的检验两个样本在位置、分布形状方面的差异的方法。

Kolmogorov-Smirnov Z 检验，建立在两个样本累计分布函数之间的最大绝对差异基础上，当这个差异显著的大时，两个分布认为是有差异的。

Wald-Wolfowitz runs 检验，对两组样本合并后赋秩，如果两个样本是来自相同的总体，则它们应该被随机的分散赋秩。当两个样本各自的秩和相差较大时，被认为是有差异的。

(3) Moses extreme reactions 检验。Moses 检验假设实验变量的变化会影响其他变量在相同或相反方向上的变化。

它分析实验组与控制组相比较时的极值分布，关注于控制组的跨度（span），当实验组和控制组合并时，以跨度的变化来度量实验组里有多少极值。计算方法是：将来自两个组的样本合并和赋秩，控制组的跨度用组里的最大值、最小值所对应的秩的差来定义；为剔除偶然因素引起的跨度波动，取值极高、极低的两端各 5% 的样本被忽略。

在 4 种方法中，Mann-Whitney U 检验是最常用的，下面以它为例，简述两独立样本检验的步骤，如图 6-17 所示。

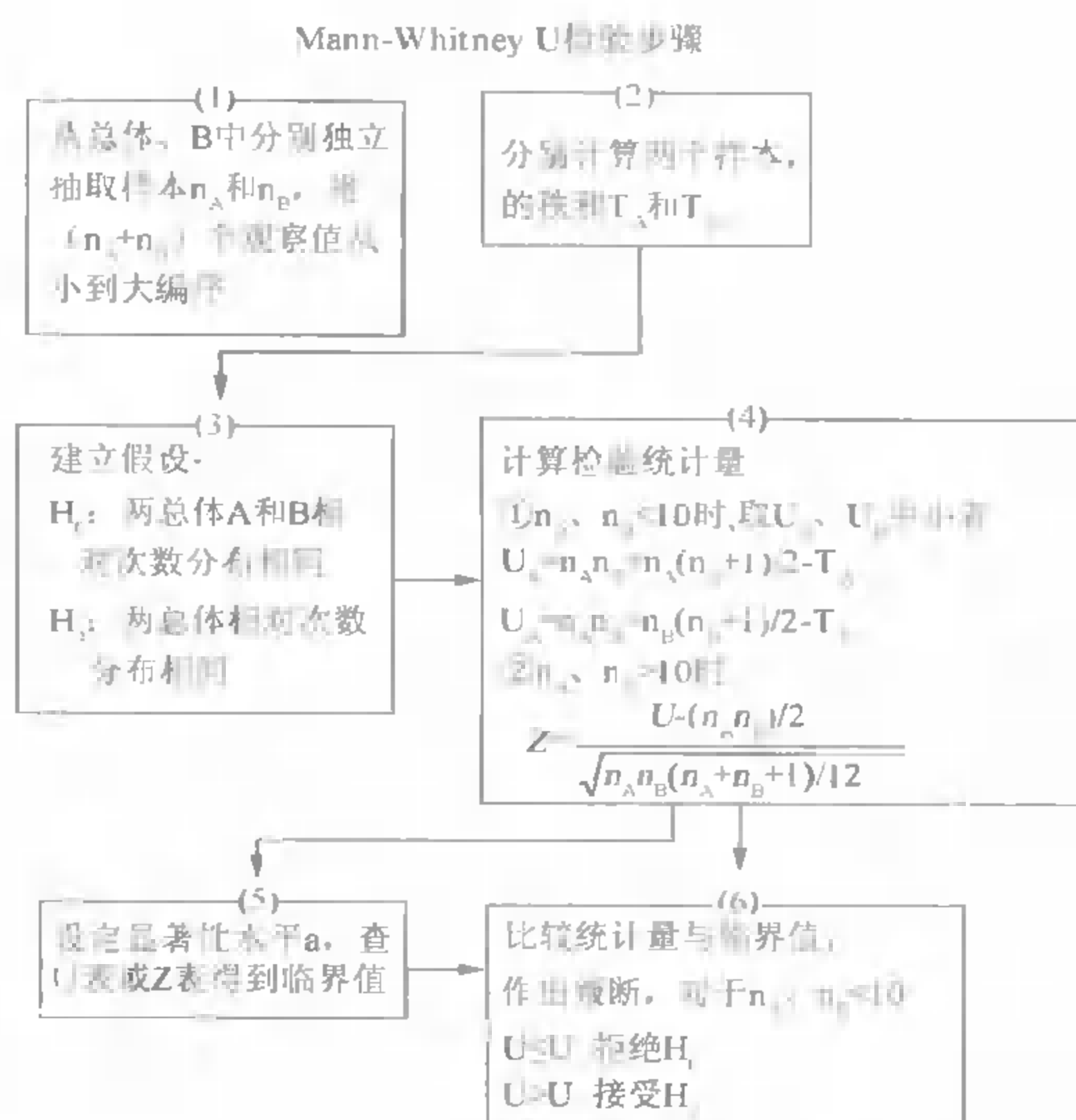


图 6-17 Mann-Whitney U 检验步骤

### 6.6.2 数据和问题描述

对同一种产品采用了两种不同的生产工艺方法，本节来分析其使用寿命是否具有相同的分布。所用数据文件为“使用寿命数据 2 独立检验.sav”，数据格式如图 6-18 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	sm	Numeric	8	2	产品寿命	None	None	8	Right	Scale
2	gy	Numeric	8	2	生产工艺	None	None	8	Right	Scale

图 6-18 不同生产工艺生产的产品使用寿命数据格式

本例检验的原假设为  $H_0$ ：两种工艺生产的产品使用寿命服从同一分布；备则假设为  $H_1$ ：两种工艺生产的产品使用寿命不服从同一分布。

### 6.6.3 两独立样本检验实例分析

依次单击菜单“Analyze→Nonparametric Tests→2 Independent Samples”，执行两独立样本检验过程，其主设置界面如图 6-19 所示。

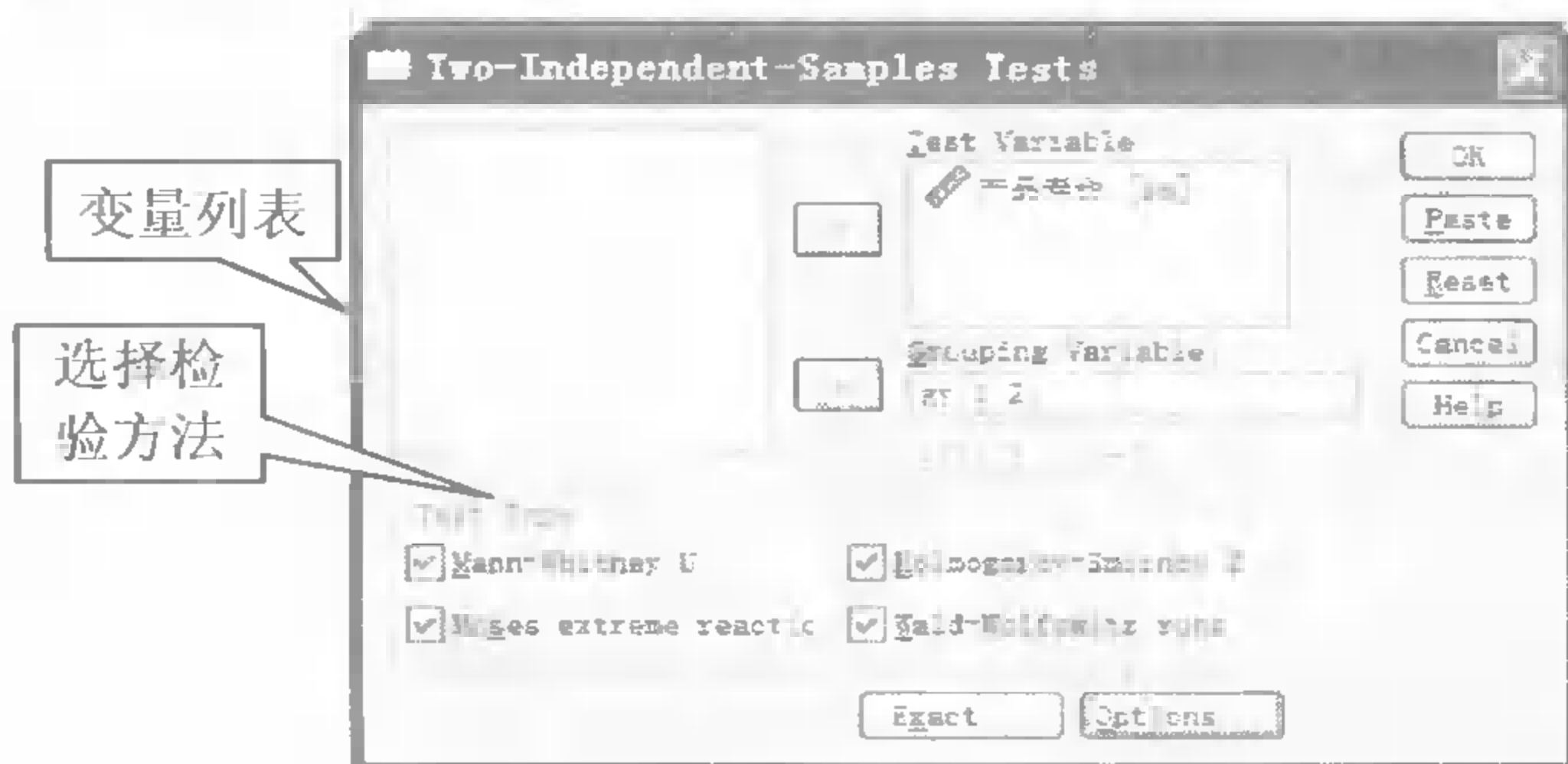


图 6-19 两独立样本检验的主设置面板

#### 1. 参数设置



在左侧的变量列表单击选中产品寿命变量，单击从上至下第一个  按钮，将其作为检验变量选入 Test Variable 列表；在左侧的变量列表单击选中生产工艺（gy）变量，单击从上至下第二个  按钮，将其作为分类变量选入 Grouping Variable 选框；单击 Define Groups 按钮，弹出如图 6-20 所示的取值定义对话框，在 Group1 后输入“1”，在 Group2 后输入“2”，单击 Continue 按钮返回主界面。分别勾选 Test Type 栏下的 4 个复选框。



图 6-20 类别变量的取值定义

(1) Test Variable 列表用于从左侧的变量列表选入检验变量，且必须为数值型分类变量。

(2) Test Type 栏选择检验方法，可选项有 4 个：Mann-Whitney U 方法（默认选中）；Kolmogorov-Smirnov Z 方法；Moses extreme reactions 方法；Wald-Wolfowitz runs 方法。

(3) Grouping Variable 栏用于从变量列表选入把样本分为两类的分类变量。在单击

Define Groups 按钮弹出的图 6-20 里,分别在两个输入框里指定两个类别的取值,其中:Group 1 将作为第一个类别,也就是 Moses extreme reactions 检验法中的控制组。

(4) 其他设置。单击 Options 按钮,将弹出如图 6-5 所示的选项设置对话框;单击 Exact 按钮,将弹出如图 6-6 所示的精确检验设置对话框。其设置选项的含义和设置方法同前。

2. 结果分析

在图 6-19 中,单击 OK 按钮运行,SPSS Viewer 窗口的输出结果如图 6-21 所示。

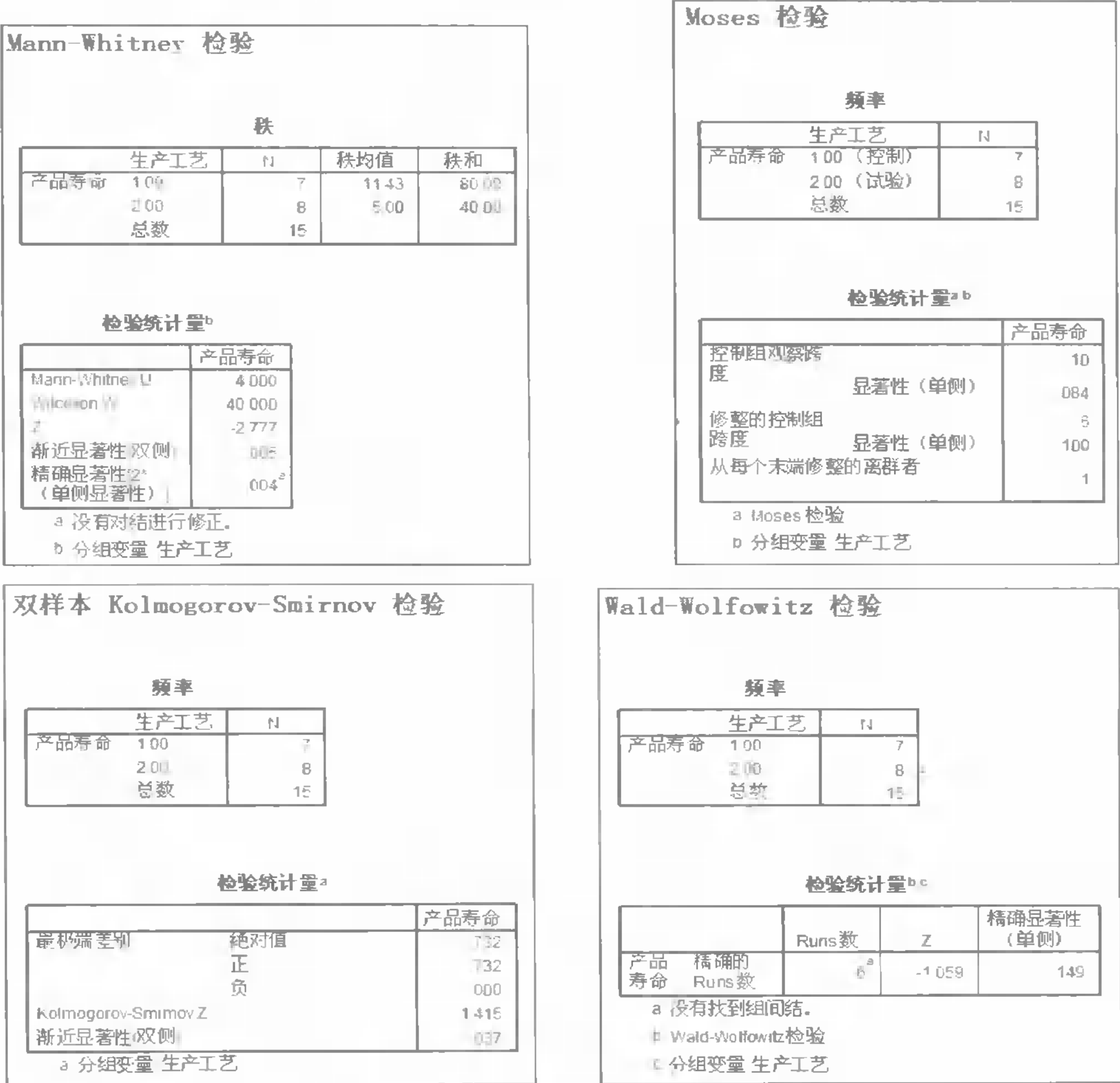


图 6-21 两独立样本检验的输出结果

其中 Mann-Whitney 检验和 K-S 检验统计量的显著性取值都小于 0.05,另外两种方法的显著性取值都大于 0.05。综合考虑 4 种检验的结果,建议否定零假设,即认为两种工艺生产的产品使用寿命不服从同一分布。

6.7 k 个独立样本的检验

要解决多于两个的独立样本之间是否具有相同分布的问题,需借助于多个独立样本检验



(test for several independent samples)方法, 它的基本原理与两独立样本检验相同, 两独立样本检验是多个独立样本检验中最基本的形式。

### 6.7.1 原理与方法

多个独立样本检验方法主要有: Kruskal-Wallis H 检验、中位数 (Median) 检验和 Jonckheere-Terpstra 检验。

Kruskal-Wallis H 检验为 Mann-Whitney U 检验的扩展, 类似非参数一维方差分析, 它研究分布位置上的差异, 利用多个样本的秩和统计量推断它们所代表的总体分布是否相同, 此方法还假设抽样的总体是连续的和相同的。Median 方法用于检验多个样本是否来自具有相同中位数的总体, 它研究总体分布在位置和形状上的差异, 效率相对较低。这两种方法都假设  $k$  个样本是从预先没有排序的总体中抽样所得。

当总体有先验的顺序排列 (升序或降序) 时, Jonckheere-Terpstra 检验法比前面两种方法更为有效。例如:  $k$  个样本分别对应了  $k$  个不同的温度值, 检验的零假设是不同温度下某化学反应的速度分布相同, 备则假设是温度越高反应越快, 这里的两个假设就是有序的, 因此 Jonckheere-Terpstra 检验是最适当的。另外, 只有安装了 SPSS Exact Tests 模块后, Jonckheere-Terpstra 检验选项才是可用的。

下面以 Kruskal-Wallis H 检验为例, 介绍多个独立样本检验的步骤。

(1) 提出零假设与备择假设。

$H_0$ : 各样本代表的总体分布相同;  $H_1$ : 各样本代表的总体分布不完全相同。

(2) 求各样本的秩和统计量。

将各个样本的所有观测值混合后, 按照由小到大的顺序排成从  $1 \sim n$  的秩次。不同样本的相同观测值 (结), 取其平均秩次; 一个样本内的相同观测值, 不求平均秩次。按样本把每个观测值的秩次一一相加, 求出各样本的秩和统计量。

(3) 求 H 统计量。

公式为:  $H = \frac{12}{n(n+1)} \sum \frac{R_i^2}{n_i} - 3(n+1)$ ; 当结较多时, 校正的公式为:  $H_C = \frac{H}{\left[1 - \frac{\sum (t_i^3 - t_j)}{n^3 - n}\right]}$

其中:  $R_i$  为第  $i$  个样本的秩和;  $n_i$  为第  $i$  个样本的样本量,  $N = \sum n_i$ ;  $t_j$  表示某个观测值重复的次数。

(4) 统计推断。

当样本数  $k > 3$ ,  $n_i > 5$  时,  $H$  近似地呈自由度为  $k-1$  的  $\chi^2$  分布, 可对  $H$  进行  $\chi^2$  检验; 当样本数较少时, 有专门的 H 检验统计表供查询临界值。当  $H > H_{0.05}$  或  $P > P_{0.05}$  时否定  $H_0$ , 即认为在 0.05 的显著性水平下, 各样本代表的总体分布不完全相同。

### 6.7.2 数据和问题描述

本节仍利用文件“儿童身高体重检验.sav”提供的数据, 来检验不同年龄儿童的身高、体重是否来自具有相同分布的总体, 数据格式同第 6.5 节的图 6-14 所示。

本例检验的原假设为  $H_0$ : 不同年龄儿童的身高、体重来自具有相同分布的总体; 备则假设为  $H_1$ : 不同年龄儿童的身高、体重不是全部来自具有相同分布的总体。

6.7.3 *k* 个独立样本检验实例分析

依次单击菜单“Analyze→Nonparametric Tests→K Independent Samples”，执行 *k* 个独立样本检验过程，其主设置界面如图 6-22 所示。

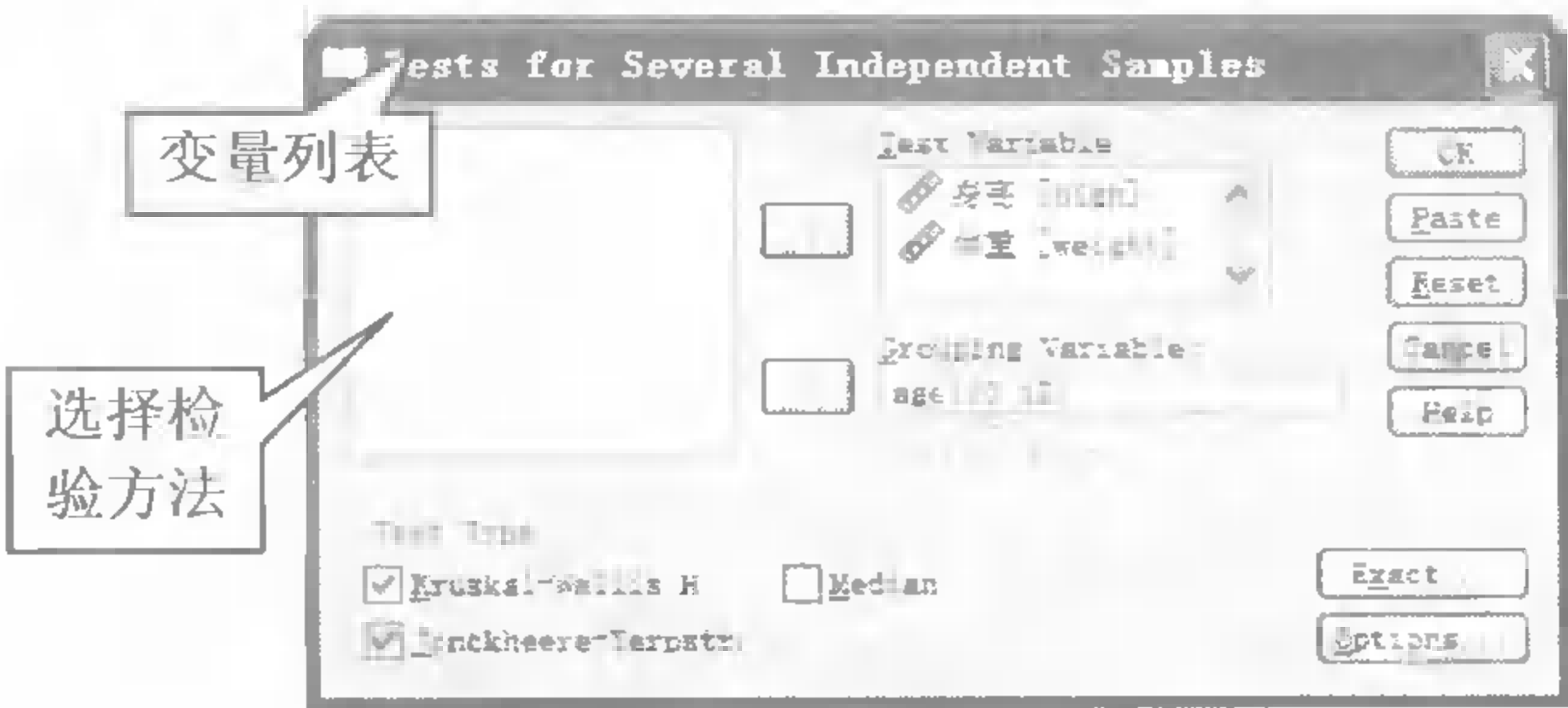




图 6-22 *k* 个独立样本检验的主设置面板

(1) 参数设置。在左侧的变量列表选中身高和体重变量，单击从上至下第一个  按钮，将其作为检验变量选入 Test Variable 列表；在左侧的变量列表单击选中年龄（age）变量，单击从上至下第二个  按钮，将其作为分类变量选入 Grouping Variable 选框；单击 Define Groups 按钮，弹出如图 6-23 所示的取值定义对话框，在 Minimum 后输入“10”，在 Maximum 后输入“13”，单击 Continue 按钮返回主界面。分别勾选 Test Type 栏下的 2 个复选框：Kruskal-Wallis H 和 Jonckheere-Terpstra。

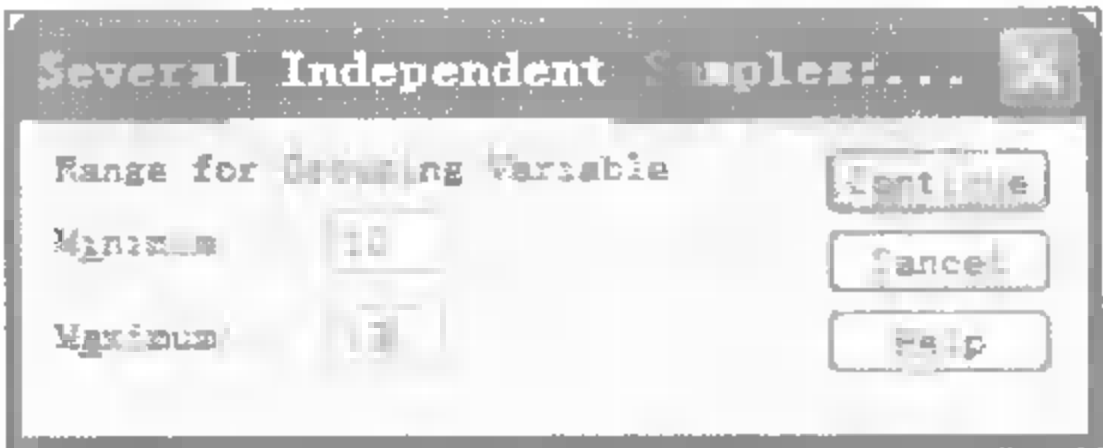


图 6-23 分类变量的取值定义

在图 6-22 中：Test Variable 列表用于从左侧的变量列表选入检验变量，且必须为数值型分类变量；其他选项的设置方法同图 6-19 所示的两独立样本检验面板的设置相似。

在图 6-23 中：Minimum、Maximum 两个输入框，分别用于指定类别变量要检验的最小取值和最大取值，不在此范围内的类别不参与检验。

(2) 结果分析。在图 6-22 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 6-24 所示。

# Kruskal-Wallis 检验

秩

	年龄	N	秩均值
身高	10	8	5.25
	11	11	13.45
	12	7	23.93
	13	1	20.50
	总数	27	
体重	10	8	6.00
	11	11	14.05
	12	7	21.64
	13	1	24.00
	总数	27	

检验统计量<sup>a,b</sup>

	身高	体重
卡方	21.656	16.280
df	3	3
渐近显著性	.000	.001

a. Kruskal-Wallis 检验

b. 分组变量: 年龄

Jonckheere-Terpstra 检验<sup>a</sup>

	身高	体重
年龄中的水平数	4	4
N	27	27
J-T 观察统计量	234.500	222.000
J-T 统计量均值	123.500	123.500
J-T 统计量的标准差	22.362	22.450
标准 J-T 统计量	4.964	4.388
渐近显著性(双侧)	.000	.000

a. 分组变量: 年龄

图 6-24 *k* 个独立样本检验的输出结果

由于两种检验统计量的渐进显著性取值都远小于 0.01，故可以非常显著地否定零假设，接受备则假设，即认为不同年龄儿童的身高、体重不是全部来自具有相同分布的总体。

## 6.8 两个相关样本的检验

两个相关样本检验过程（2 Related Samples test）可对两个相关样本资料（例如配对、配伍资料）进行秩和检验。

### 6.8.1 原理与方法

两个相关样本检验的方法主要有：Wilcoxon 检验、Sign（符号）检验、McNemar 检验和 Marginal Homogeneity 检验等。Wilcoxon 检验用于检验两个相关样本是否来自相同的总体，但对总体分布形式没有限制；Sign 检验通过分析两个样本的正负符号个数判断它们是否来自相同的总体；McNemar 检验用于两个相关二分变量的检验；Marginal Homogeneity 检验用于两个相关定序变量的检验，是 McNemar 检验的扩展。

#### 1. Sign 符号检验

配对资料的符号检验，通过分析两个样本各对数据之差的正负符号的个数，来判断两个总体分布的异同，而不考虑差值的实际大小。本方法适用于相关样本资料和定性变量。

配对数据之差为正值用“+”表示，负值用“-”表示。若两组数据的分布没有显著差异，那么差值为“+”、“-”号的个数应大致相等，即出现“+”或“-”的概率都为 0.5。如果某次随机抽样的配对数据中，“+”号出现过多或过少，就可以在一定的显著性水平 $\alpha$ 上，推断这两组数据的中值水平或总体分布是不相同的。可见，配对符号检验是二项检验的一种应用，当 $n>25$ 时，可按正态分布近似处理。符号检验的步骤如图 6-25 所示。

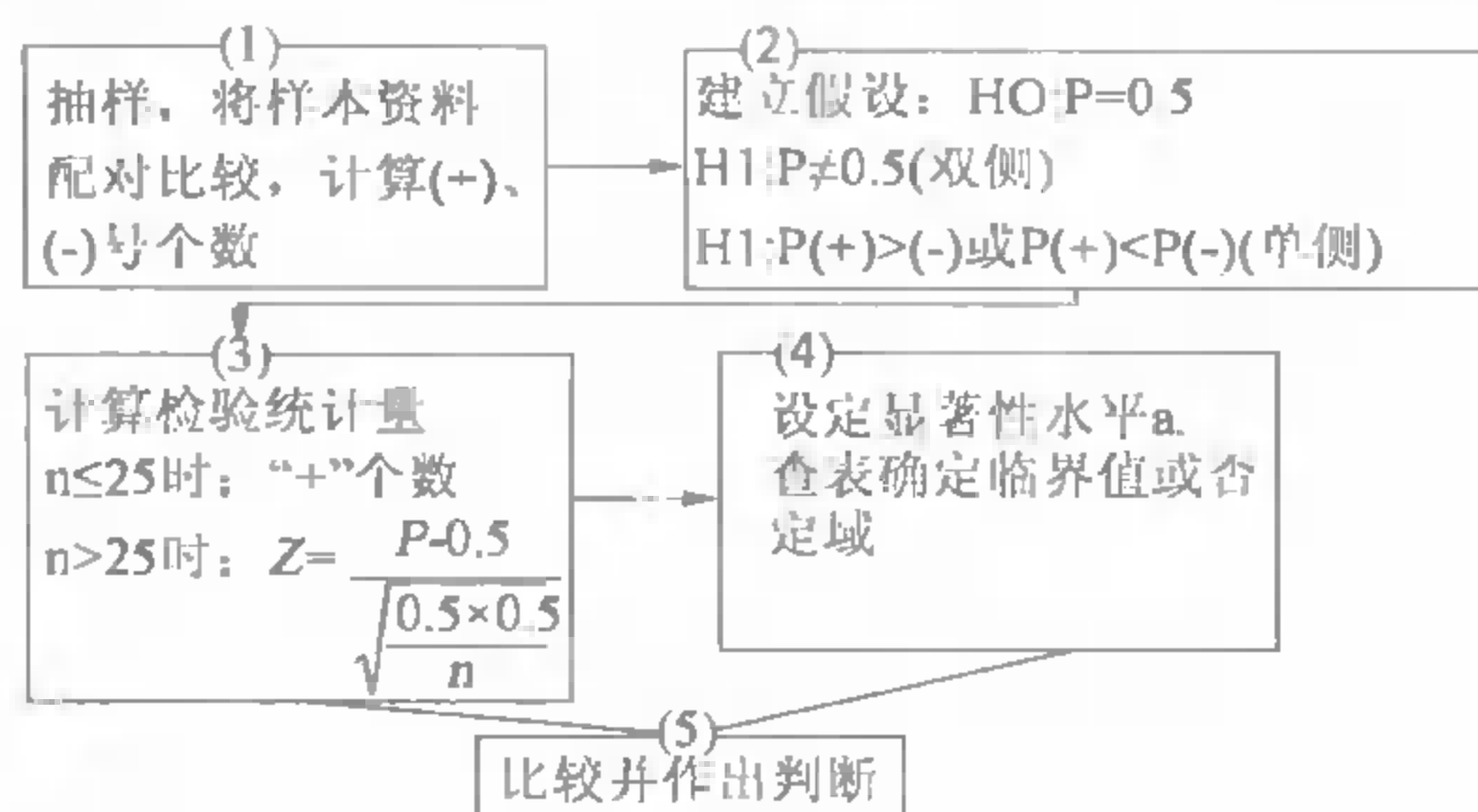


图 6-25 配对资料符号检验的步骤

注意：这种检验比较的是中位数而不是平均数，当分布对称时，中位数与平均数相等。

#### 2. Wilcoxon 检验

Wilcoxon 秩和检验是一种改进后的符号检验，它不仅考虑两组配对数据之差的正负号，而且还利用了其差异大小的信息，因此是一种更为有效的检验方法。

若关联样本的两组数据没有显著差异，则不仅其差值的正负符号应大致相等，而且将差值按大小顺序排列且编自然序号（即秩）后，其正号的秩和（记为 $T^+$ ）与负号的秩和（记为

$T^-$ )也应该大致相等,且这二者中较小的秩和应趋近于总秩和的平均数 $\bar{T} = \frac{n(n+1)}{4}$ 。若正秩和( $T^+$ )与负秩和( $T^-$ )相差太大,使它们之中较小的秩和偏离 $\bar{T}$ 较远,以致超过给定显著性水平 $\alpha$ 所确定的临界点,就推断这两组数据之间存在显著差异,即总体的分布不相同。

Wilcoxon 检验的步骤如图 6-26 所示。

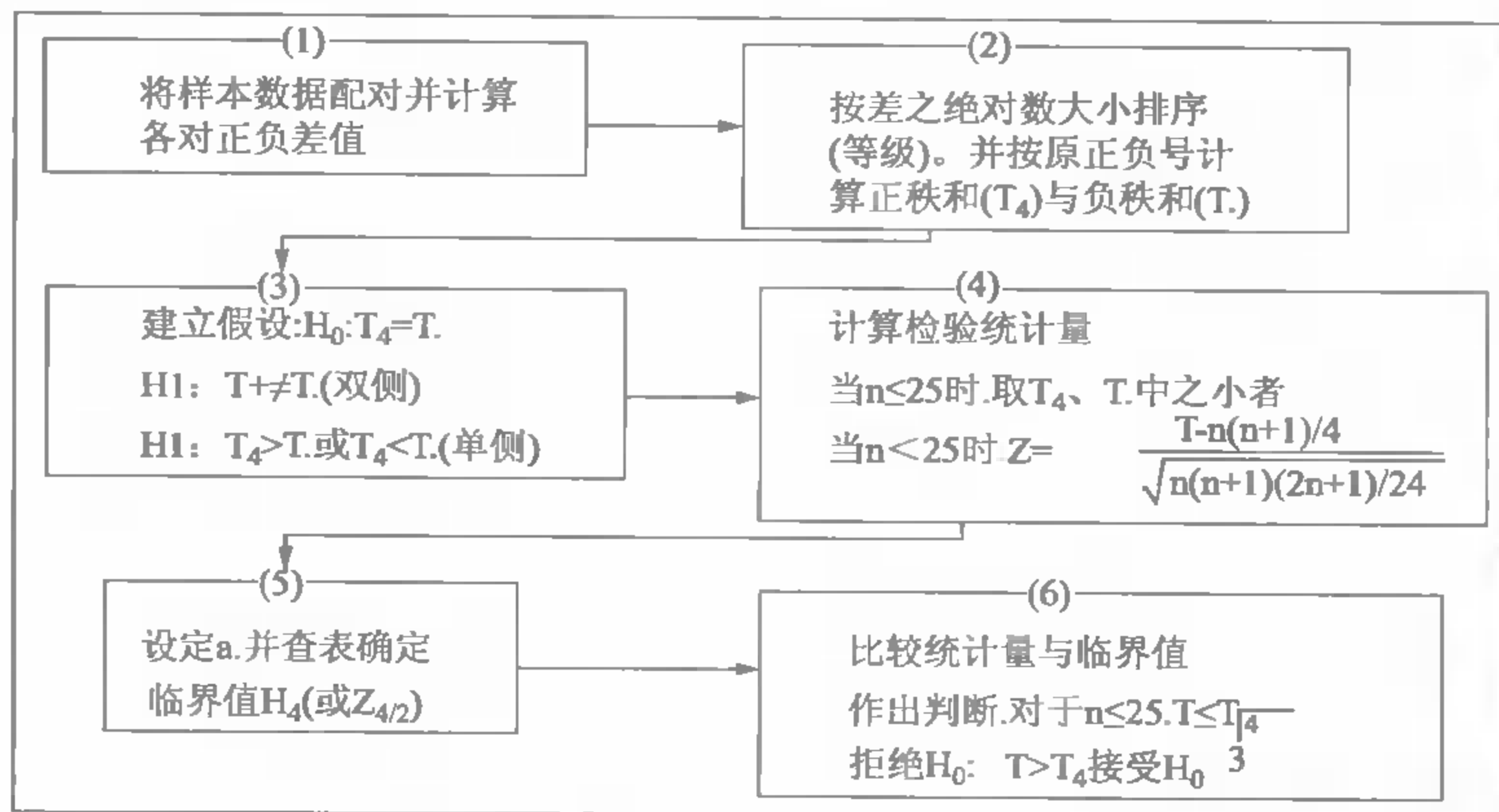


图 6-26 Wilcoxon 检验的步骤

### 3. 其他检验

如果数据是二分类的,应该使用 McNemar 检验。此时,对每个试验对象观测两次,分别在指定事件发生的前、后。此方法能够检验初始的观测比率(事件前)是否等于最终的观测比率(事件后),可用于研究特定事件对试验对象的影响效果。

如果数据是多分类的,应该使用 Marginal Homogeneity 检验,它是 McNemar 检验从二分类事件向多分类事件的推广。此方法使用卡方分布检验事件发生前及发生后观测数据的变化。另外,只有安装了 Exact Tests 模块, Marginal Homogeneity 检验才可用。

### 6.8.2 数据和问题描述

在跳远运动员经过特定训练项目的前、后时段,分别测量记录他们的成绩,以检验训练前后的成绩是否有显著差异,即检验训练前后的成绩是否来自同一个分布的总体,并由此判断训练是否卓有成效。所用数据文件为“跳远成绩配对检验.sav”,数据格式如图 6-27 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	xlq	Numeric	8	2	训练前	None	None	8	Right	Scale
2	xlh	Numeric	8	2	训练后	None	None	8	Right	Scale

图 6-27 跳远运动员成绩数据

本例检验的原假设为  $H_0$ : 训练前后的成绩来自同一个分布的总体;备则假设为  $H_1$ : 训练前后的成绩不是来自同一个分布的总体。

### 6.8.3 两个相关样本检验的实例分析

依次单击菜单“Analyze→Nonparametric Tests→2 Related Samples”,执行两个相关样本检验的过程,其主设置界面如图 6-28 所示。

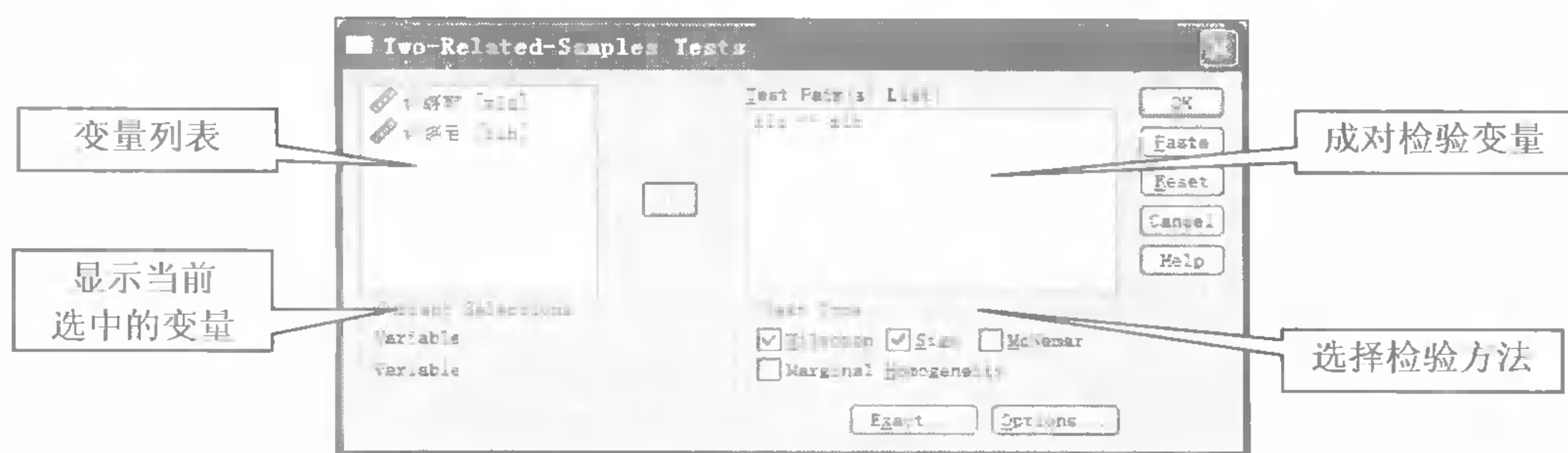


图 6-28 Two Related Sample tests 的主设置面板

## 1. 参数设置

在变量列表同时选中训练前和训练后变量，单击 按钮，将其作为一对检验变量选入 Test Pair(s)列表；分别勾选 Test Type 栏下 Wilcoxon 和 Sign 复选框。

在图 6-28 中，Test Pair(s) List 列表用于从左侧的变量列表选入数值型的检验变量，且必须成对选择，选入的变量也是成对出现，中间以连接符连接，例如“x1q--x1h”；其他选项的设置方法同图 6-19 所示的两独立样本检验面板的设置相似。

## 2. 结果分析

在图 6-28 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 6-29 所示。

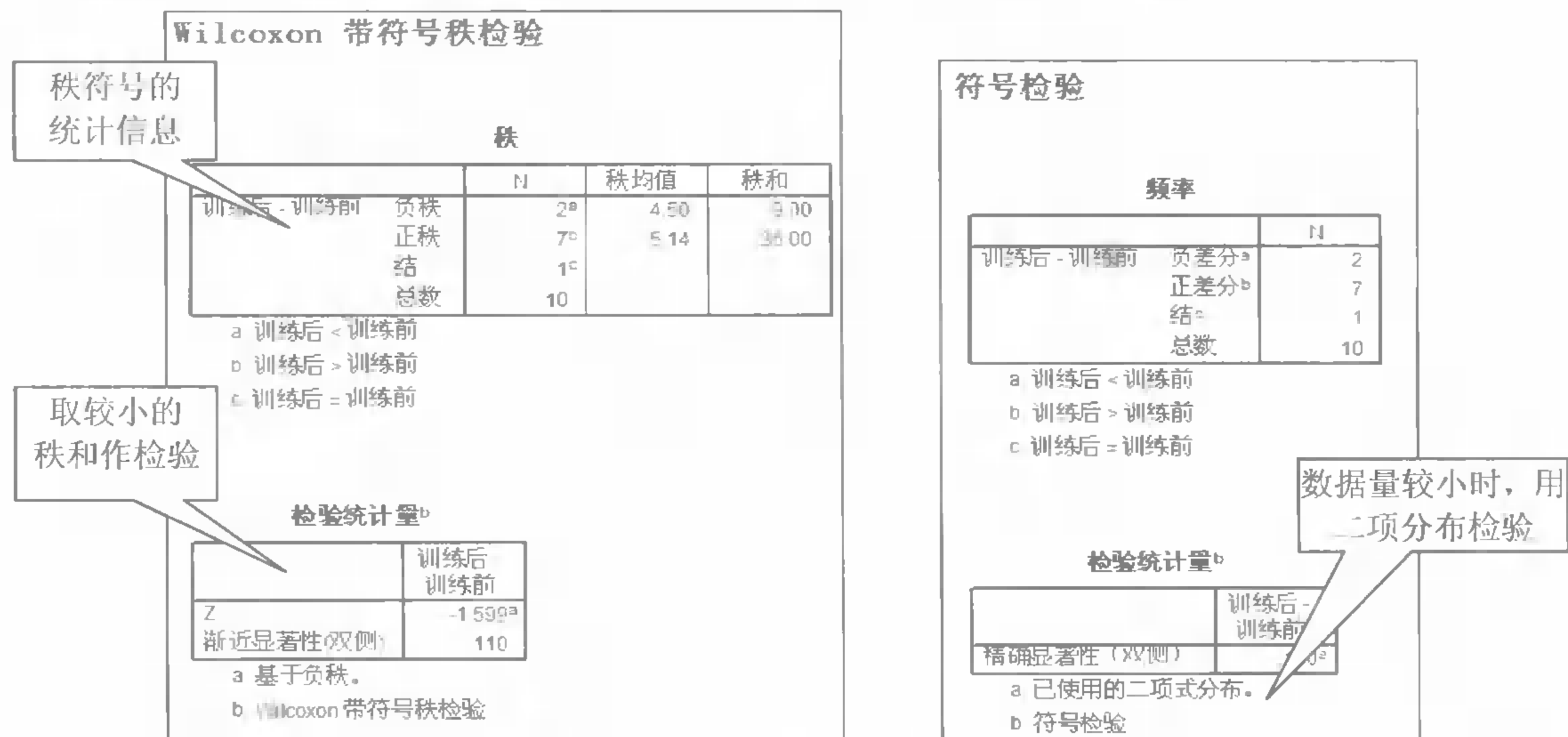


图 6-29 Two Related Sample tests 结果输出

由于 Wilcoxon 检验和 Sign 检验的显著性取值都大于 0.1，故不能否定零假设，所以认为训练前后的成绩是来自同一个分布总体的，训练效果不显著。

## 6.9 k 个相关样本的检验

要解决多个相关样本（例如配对、配伍资料）是否是来自同一个总体的问题，需要借助于多个相关样本检验（test for several related samples）方法。两相关样本检验，是多个相关样



本检验中最基本的形式。

### 6.9.1 原理与方法

多个相关样本的检验方法有：Friedman 检验、Kendall W 检验和 Cochran Q 检验等。

Friedman 检验为双向方差分析，考察多个相关的样本是否来自同一总体；Kendall W 检验，通过计算 Kendall 和谐系数 W，检验多个相关样本是否来自同一分布的总体；Cochran Q 检验作为两相关样本 McNemar 检验的多样本推广，特别适用于定性变量和二分字符变量。

#### 1. Friedman 秩和检验

记第一个因子（称为处理，treatment）有  $k$  个水平，第二个因子（称为区组）有  $b$  个水平，一共就有  $k \times b$  个观测值，这种描述类似于两因子方差分析。形式上，假定这些样本有连续的分布函数  $F_1, F_2, \dots, F_k$ ，零假设为  $H_0: F_1 = F_2 = \dots = F_k$ ，备选假设为  $H_A: F_i(x) = F(x + t_i)$ ，( $i = 1, 2, \dots, k$ )，这里  $F$  为某连续的分布函数，而且各参数  $t_i$  不相等。

Friedman 秩和检验与 Kruskal-Wallis 检验相似，但是由于区组的影响，首先要在每一个区组中计算各个处理的秩，再把每一个处理在各区组中的秩相加。以  $R_{ij}$  表示在  $j$  个区组中第  $i$  个

处理的秩，把秩按照处理求和就得到： $R_i = \sum_{j=1}^b R_{ij}$ ， $i = 1, \dots, k$ ，这样做的目的是要在每个区组内

比较不同的处理。例如：在同个年龄段的人群中比较某药品的疗效，要比不分年龄段的比较来得合理。基于以上陈述，将 Friedman 统计量定义为：

$$Q = \frac{12}{bk(k+1)} \sum_{i=1}^k \left( R_i - \frac{b(k+1)}{2} \right)^2$$

它近似地服从  $k-1$  个自由度的  $\chi^2$  分布，如果各处理水平的取值很不一样，此统计量就会显著得大。

#### 2. Kendall W 检验

在实践中，经常需要按照某些特别的性质，多次对一些个体进行评估或排序，比如：有  $m$  个评估机构，分别要对  $n$  个学校的教育水平进行排序。需要分析的是，这些机构的不同评估结果是否一致，如果不一致，就说明这些评估有较大的误差，评估的意义不大。换句话说，这里要检验的零假设是：对指定学校的多个排序是不相关的或者是随机的；备选假设为：对指定学校的多个排序是正相关的或者是较为一致的。

一个机构对诸个体（学校）排序的秩（次序）的和为  $1 + 2 + \dots + n = n(n+1)/2$ ；所有  $m$  个机构对所有个体评估的总秩和为  $mn(n+1)/2$ ；于是，每个个体的平均秩为  $m(n+1)/2$ 。记每一个个体的  $m$  个秩的和为  $R_i$  ( $i = 1, \dots, k$ )，那么如果评估是随机的，这些  $R_i$  与平均秩的差

别不会很大，反之差别会很大。定义  $S$  为： $S = \sum_{i=1}^n \left( R_i - \frac{m(n+1)}{2} \right)^2$ ，代表个体的总秩与平均

秩的偏差平方和。 $S$  和 Kendall 协同系数 (Kendall's Coefficient of Concordance) 是成比例的，

Kendall 协同系数  $W$  (Kendall's  $W$ ) 的定义为： $W = \frac{12S}{m^2(n^3 - n)}$

#### 3. Cochran Q 检验

当观测量只能取两个值时（例如 0、1），由于有太多相同的取值，就会造成结很多，这

样排序的意义就不明显了。Cochran 检验就是用来解决这个问题的。

举个例子：关于瓶装饮用水的调查，有  $n$  名顾客对  $m$  种瓶装饮用水进行了认可（记为 1）和不认可（记为 0）的表态，检验的零假设是：这  $n$  种瓶装水（作为处理）在顾客（作为区组）眼中没有区别。用  $N_i$  表示第  $i$  个处理得到的“1”的个数，用  $L_j$  表示第  $j$  个区组所给的“1”的个数，出现“1”的总数记为  $N$ 。如果  $N_i$  和所有  $N_i$  均值的差距很大，就可以推断这些处理之间很不一样，Cochran 检验就是基于这个思想的。

假定有  $k$  个处理和  $b$  个区组，Cochran 检验统计量（Cochran's Q）的计算公式为：

$$Q = \frac{k(k-1) \sum_{i=1}^k (N_i - \bar{N})^2}{kN - \sum_{j=1}^b L_j^2} = \frac{k(k-1) \sum_{i=1}^k N_i^2 - (k-1)N^2}{kN - \sum_{j=1}^b L_j^2}, \text{ 其中 } \bar{N} = \frac{1}{k} \sum_{i=1}^k N_i, \text{ 各符号的含义参考}$$

瓶装饮用水的例子。若  $k$  固定， $Q$  在  $b$  很大时近似服从于自由度为  $k-1$  的  $\chi^2$  分布。

6.9.2 数据和问题描述

某商场采用了三种促销方式，并对不同商品的销售额做了记录，本节来分析不同的促销形式是否有显著的优劣之分。所用文件为“促销形式效果检验.sav”，数据格式如图 6-30 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	x1	Numeric	8	2	促销形式1	None	None	8	Right	Scale
2	x2	Numeric	8	2	促销形式2	None	None	8	Right	Scale
3	x3	Numeric	8	2	促销形式3	None	None	8	Right	Scale

图 6-30 不同促销形式下的商品销售额数据

本例检验的原假设为  $H_0$ ：三种促销形式下的商品销售额来自同一个分布的总体；备则假设为  $H_1$ ：三种促销形式下的商品销售额不是来自同一个分布的总体。

6.9.3  $k$  个相关样本检验的实例分析

依次单击菜单“Analyze→Nonparametric Tests→K Related Samples”，执行  $k$  个相关样本检验的过程，其主设置界面如图 6-31 所示。



图 6-31 Tests for Several Related Samples 设置面板

1. 参数设置

在变量列表选中促销形式这 3 个变量，单击 按钮，将其作为检验变量选入 Test 列表；分别勾选 Test Type 栏下 Friedman 和 Kendall's W 复选框。

在图 6-28 中，Test 列表用于从左侧的变量列表选入多于一个的检验变量，且必须为数值型变量；其他选项的设置方法同图 6-19 所示的两独立样本检验面板的设置相似。

2. 结果分析

在图 6-31 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 6-32 所示。

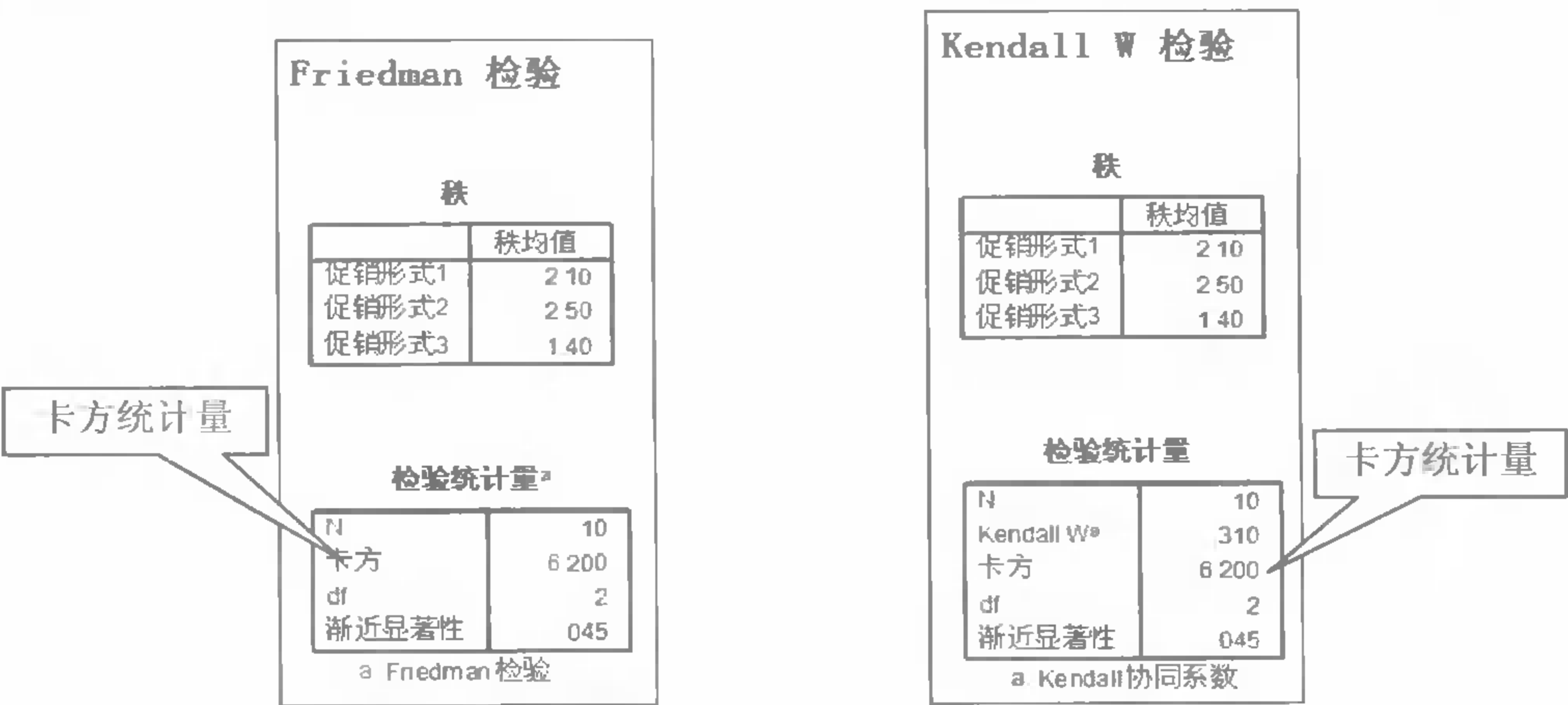


图 6-32 Tests for Several Related Samples 的结果输出

Friedman 检验和 Kendall's W 检验均采用了卡方统计量，且渐进显著性的取值都小于 0.05，故可在 0.05 的显著性水平下否定零假设，即认为三种促销形式下的商品销售额不是来自同一个分布总体的，三种促销方式的效果是显著不同的。

# 第 7 章 多重响应分析

多重响应分析 (Multiple Response)，也称之为多重应答分析或多响应变量分析。

多重应答，又称多选题，即针对同一个问题被访者可能回答出多个有效的答案，这是市场调查研究中十分常见的数据形式。对多选题形式的数，可以使用 SPSS 中的多重响应分析 (Multiple Response) 过程进行频数分析和交叉表分析，还可以使用最优尺度过程 (Optimal Scaling) 进行多重对应分析，研究该数据与其他若干个变量之间的相互关系。

## 7.1 多重响应概述

在网站访问的调查中，经常要求用户选择最喜欢的网站版块，供选择的可能有新闻板块、娱乐板块、用户编辑交流板块等，受访者可以选择单个答案或者多个答案的组合，对这类多项选择题的统计，SPSS 称之为多重响应 (Multiple Response)。

多重响应的数据本质上属于分类数据，但由于各选项均是对同一个问题的回答，故而它们之间会存在一定的相关性，所以对各选项单独进行分析就会显得不恰当。首先，需要对多选题的结果进行数据转换，转换的方式有如下两种。

(1) 多项选择的二分法 (Multiple dichotomy method)。把多项选择题的每一个选项当作一个单独的二元变量来定义，取值 0 代表没有被选中，取值 1 代表被选中。这样，多项选择题的答案有几个选项，就会转换为几个单选变量。

(2) 多项选择的分类法 (Multiple category method)。根据被访者可能提供的答案数量，设置相应个数的单选变量。假设被访者最多只能选择  $n$  个不同答案，就需要采用  $n$  个单选变量来记录本多选题的回答数据。每个单选变量的可能取值都和多项选择题的可选项相同，代表了被调查者的一次选择，记录的是反映被选中的多选题选项的代码。

多重应答资料因其特殊性，不方便应用传统的多元统计分析方法进行研究，利用如上的两种数据转换方式可以极大地丰富对其建模的方法。SPSS 的多重响应分析功能，通过定义变量集的方式，能够对多选题选项进行频数分析和交叉表分析；除此之外，还可以对其进行回归分析、因子分析等操作。

## 7.2 多重响应变量集的定义

SPSS 中的定义多重响应变量集 (Define Multiple Response Sets) 过程，能够将多个基本

变量定义为多重响应的数据类型：多重二分类变量集或多重多分类变量集。这是进行多重响应的频数分析和交叉表分析前，必须要进行的准备工作。

由此定义的多重响应变量集只能在多重响应分析过程（频数分析和交叉表分析）中使用，其它分析过程不能访问。

7.2.1 定义多重响应变量集的实例

1. 问题和数据描述

某电信公司为客户提供了各种各样的服务，包括：多线路使用、语音邮箱、寻呼业务、Internet 服务、来电显示、呼叫等待、呼叫转移、3 方通话和电子帐单等，许多客户经常使用其中的多项服务，所以该电信公司希望对这些服务数据建立一个多响应变量集，并以此来研究这些客户使用模式的特点和规律。所用数据摘自 SPSS 自带的 Demo 文件“telco.sav”，数据文件为“电信客户消费模式数据.sav”，数据格式如图 7-1 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	region	Numeric	4	0	地区	{1, Zone 1}...	None	6	Right	Nominal
2	tenure	Numeric	4	0	在网月数	None	None	6	Right	Scale
3	age	Numeric	4	0	年龄	None	None	6	Right	Scale
4	marital	Numeric	4	0	婚否	{0, 未婚}...	None	7	Right	Nominal
5	income	Numeric	8	2	家庭收入（千）	None	None	10	Right	Scale
6	gender	Numeric	4	0	性别	{0, 男}...	None	6	Right	Nominal
7	multiline	Numeric	4	0	多线路使用	{0, No}...	None	8	Right	Nominal
8	voice	Numeric	4	0	语音邮件	{0, No}...	None	6	Right	Nominal
9	pager	Numeric	4	0	寻呼业务	{0, No}...	None	6	Right	Nominal
10	internet	Numeric	4	0	Internet 服务	{0, No}...	None	8	Right	Nominal
11	callid	Numeric	4	0	来电显示	{0, No}...	None	6	Right	Nominal
12	callwait	Numeric	4	0	呼叫等待	{0, No}...	None	8	Right	Nominal
13	forward	Numeric	4	0	呼叫转移	{0, No}...	None	7	Right	Nominal
14	confer	Numeric	4	0	3 方通话	{0, No}...	None	6	Right	Nominal
15	ebill	Numeric	4	0	电子帐单	{0, No}...	None	6	Right	Nominal
16	custcat	Numeric	8	0	客户种类	{1, Basic ser	None	10	Right	Nominal
17	churn	Numeric	4	0	是否流失	{0, No}...	None	6	Right	Nominal

图 7-1 电信客户消费模式数据格式

2. 定义多响应变量集的参数设置

依次单击菜单“Analyze→Multiple Response→Define Sets...”执行定义多重响应变量集的功能，其主设置界面如图 7-2 所示。

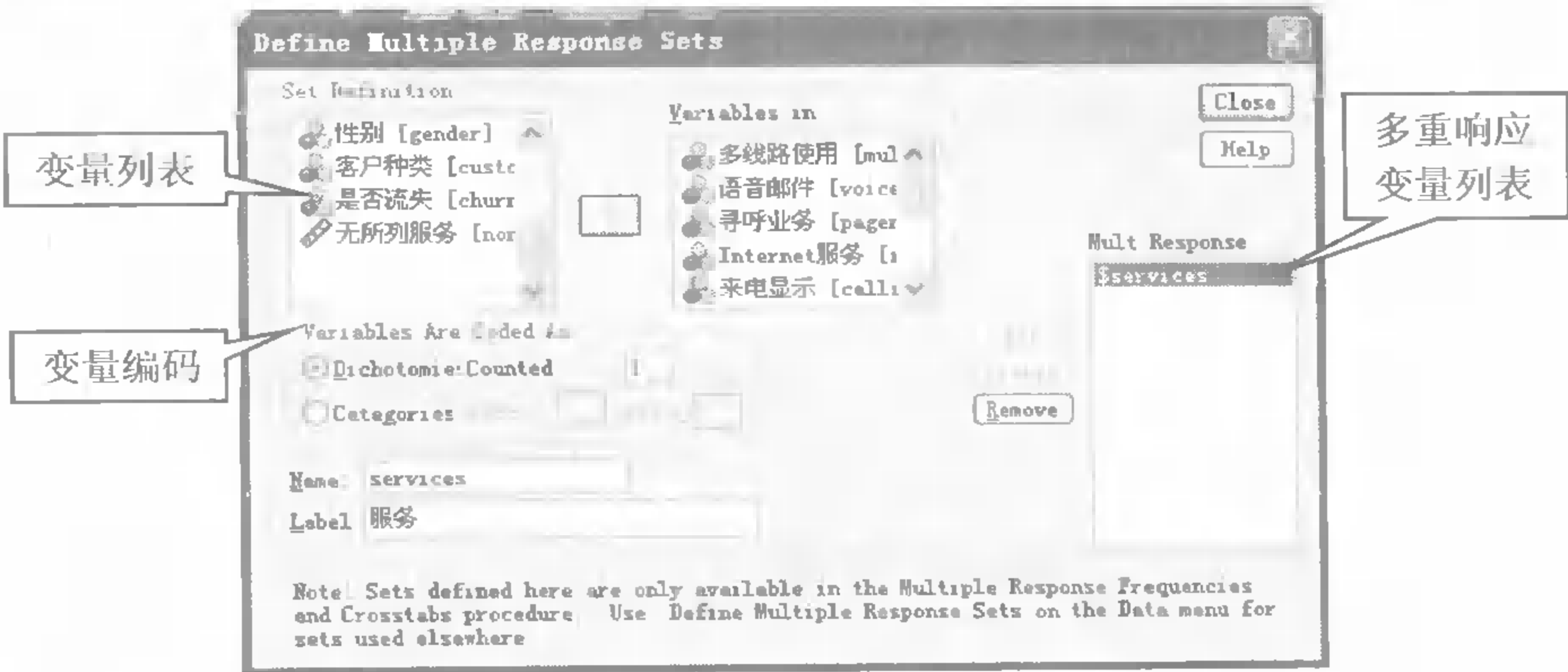


图 7-2 Define Sets 参数设置

在变量列表选中从多线路使用至电子账单的 9 个变量，单击  按钮，将其选入 Variables



in Set 列表；在单选项 Dichotomises Counted value 后输入“1”；在 Name 后输入“services”，在 Label 后输入“服务”；单击 Add 按钮，将这个定义好的多重响应变量集选入右侧的 Mult Response Sets 列表。

① Variables in Set 列表框用于从左侧的变量列表选入同属于一个问题的多个答案变量，并将根据该列表中的变量来定义多重响应变量集。

② Variables Are Coded As 栏选择对当前多重响应变量集的编码方式，有如下两个选项。

☒ Dichotomises Counted value 单选框，表示使用二分变量的计数值进行编码，多选题的每一个选项被当作一个单独的二元变量。Variables in Set 列表选择了几个变量就表示多选题有几个选项；在后面输入框指定选项被选中时的二元变量取值，比如“1”就指当某个单独变量取值为 1 时，他所代表的多选题选项被选中。

☐ Categories 单选框，表示使用分类变量进行编码，为多选题设定与其最多答案个数相等的单选变量；每个单选变量的可能取值都和多选题的可选项相同，它代表被选中的多选题选项的代码。Range、through 输入框分别用于指定可选答案代码的起始值和终止值。

③ 定义变量名和变量标签。Name 输入框指定当前多重响应变量集的名称，系统将自动在指定的名称前加上字符\$；Lable 输入框指定当前多重响应变量集的标签。

④ Mult Response Sets 列表。用于选入定义好的多重响应变量集。单击 Add、Change、Remove 按钮分别添加、修改、删除当前指定的多重响应变量集。

### 3. 结果分析

在图 7-2 中，单击 Close 按钮返回 Data Editor 窗口。

当前数据集没有什么变化，但是再次单击菜单“Analyze→Multiple Response”，会发现 3 个子菜单都可用了，说明当前数据集已经定义了某些多重响应变量集（如：services）。

## 7.3 多重响应变量集的频数分析

对多重响应变量集做频数分析，就是为代表多选题答案的变量集生成频数表。


只有在成功定义了多重响应变量集后，才能进行对变量集的频数分析，所以本节接着上一节的例子，用实例来说明如何做多选题的频数表。

### 7.3.1 多重响应变量频数分析的实例

#### 1. 问题和数据描述

本节接着第 7.2 节的例子进行分析，所用数据的格式如图 7-1 所示。前面已经定义了多重响应变量集 services，随后就来做关于变量集 services 的频数分析。

#### 2. 多重响应 Frequencies 过程的参数设置

依次单击菜单“Analyze→Multiple Response→Frequencies...”执行多重响应变量集的频数分析功能，其主界面如图 7-3 所示。在变量集列表选中服务变量，单击  按钮，将其作为分析变量集选入 Table(s) for 列表框。

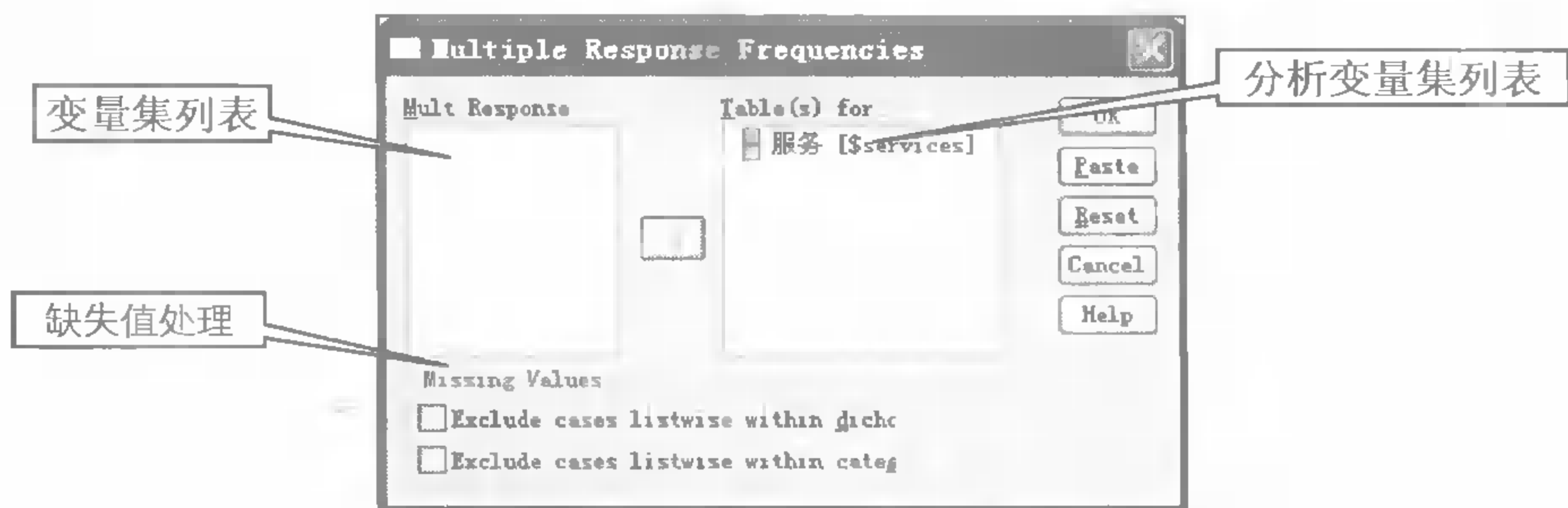


图 7-3 Frequencies 过程参数设置

Mult Response Sets 列表框，显示当前已经定义的多重响应变量集；Table(s) for 列表框，从变量集列表框选入要进行频数分析的多重响应变量集；Missing Values 栏，用于选择处理缺失值的方法，有如下两个可选项。

① Exclude cases listwise within dichotomies 复选框。适用于采用多项选择二分法编码的多重响应变量集，缺失记录不进入分析。默认情况下，如果多重响应变量集中的所有二元变量都没有取代表选中状态的计数值（例如本例“1”），相应的观测记录就被认为是缺失的；当至少有一个二元变量取计数值时，即使其他个别（不是全部）二元变量的取值有缺失，相应的观测记录也不会被当作缺失处理。

② Exclude cases listwise within categories 复选框。适用于采用多项选择分类法编码的多重响应变量集，缺失记录不进入分析。默认情况下，只有当多重响应变量集中每个变量的取值都不在定义范围内时，相应的观测记录才被当作缺失记录处理。

### 3. 结果分析

单击图 7-3 中 OK 按钮运行，SPSS Viewer 窗口的输出表格如图 7-4 所示。

个案摘要						
	个案					
	有效的		缺失		总计	
	N	百分比	N	百分比	N	百分比
\$services <sup>a</sup>	889	88.9%	111	11.1%	1000	100.0%
a. 值为 1 时制表的二分组。						

Services 频率				
	响应			
		N	百分比	个案百分比
服务 <sup>a</sup> 多线路使用		475	12.7%	53.4%
语音邮件		304	8.1%	34.2%
寻呼业务		261	7.0%	29.4%
Internet 服务		368	9.8%	41.4%
来电显示		481	12.9%	54.1%
呼叫等待		485	13.0%	54.6%
呼叫转移		493	13.2%	55.3%
3方通话		502	13.4%	56.5%
电子账单		371	9.9%	41.7%
总计		3740	100.0%	420.1%
a. 值为 1 时制表的二分组。				

图 7-4 个案摘要表和频率表

(1) 个案摘要表。显示了对有效数据和缺失数据的基本统计信息，在本例的 1 000 个案例中，有 111 个案例被认为是缺失的，即这 111 个案例是没有定制任何服务的客户。

(2) 频数表。它与对单个变量进行频数分析的输出表格很相似，但更加紧凑，并提供了额外的信息。

N 列表示使用指定单项服务的客户数目，最后的总计为所有行的求和，它可能大于总的观测数，因为同一个客户可以选择多项服务。响应百分比列，使用指定单项服务的客户数占使用服务总频率的百分比，也就是同行的 N 除以对 N 的总计得到的比例，这在对单个变量的

频数分析表中是没有的。个案百分比列，使用指定单项服务的客户数占总客户数的百分比，也就是同行的 N 除以有效的总客户个数得到的比例。

## 7.4 多重响应变量集的交叉表分析

对多重响应变量集做交叉表分析，就是为代表多选题答案的变量集生成二维交叉表。

只有在成功定义了多重响应变量集后，才能进行对变量集的交叉表分析，所以本节接着上一节的例子来说明如何做多选题的交叉表。

### 7.4.1 多重响应变量交叉表分析的实例

#### 1. 问题和数据描述

本节接着第 7.2 节的例子进行分析，所用数据的格式如图 7-1 所示。前面已经定义了多重响应变量集 services，下面就来做关于变量集 services 的交叉表分析。

#### 2. 多重响应 Crosstabs 过程的参数设置

依次单击菜单“Analyze→Multiple Response→Crosstabs...”执行多重响应变量集的交叉表分析功能，其主设置界面如图 7-5 所示。

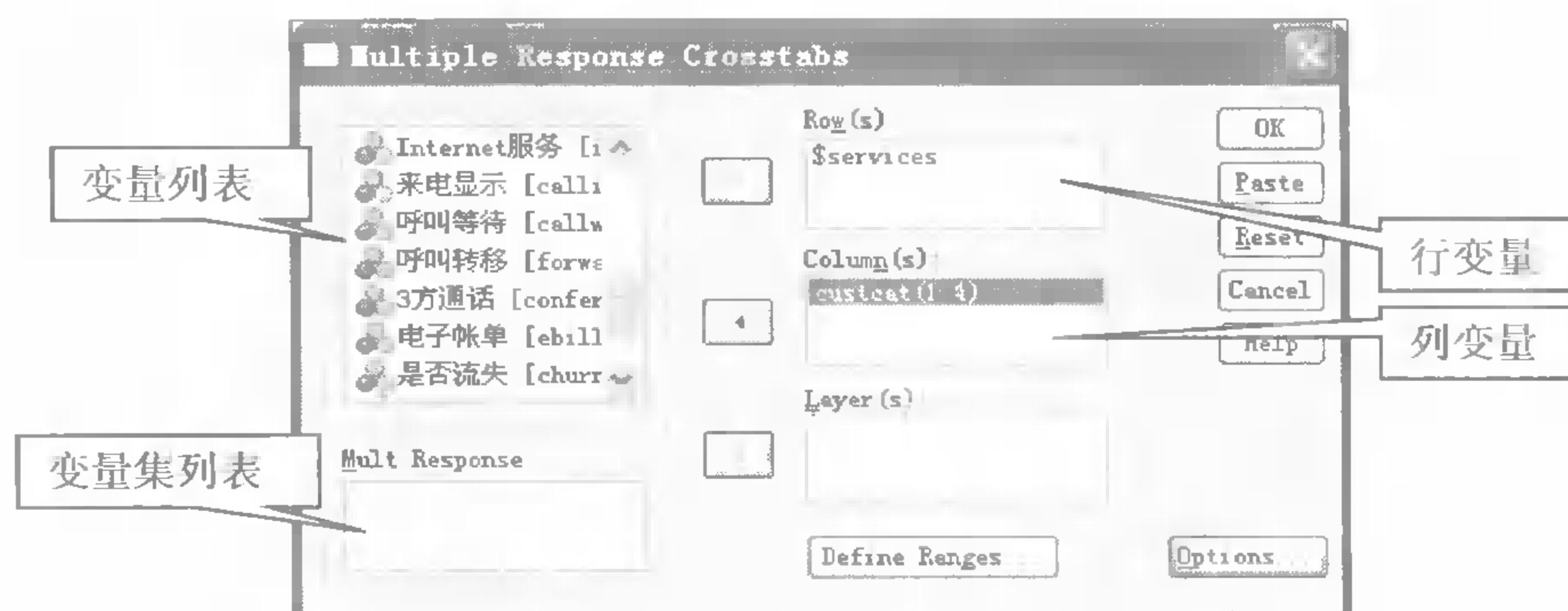




图 7-5 多重响应 Crosstabs 过程的参数设置

(1) 分析变量设置。在变量集列表单击选中服务（\$services）变量集，单击从上至下第一个  按钮，将其作为行变量选入 Row 列表；在变量列表单击选中客户种类（custcat）变量，单击从上至下第二个  按钮，将其作为列变量选入 Column 列表。

选中 Column 列表中的 custcat 变量，单击 Define Ranges 按钮，弹出如图 7-6 所示的取值定义对话框，在 Minimum 和 Maximum 后分别输入“1”、“4”，单击 Continue 按钮返回主界面。

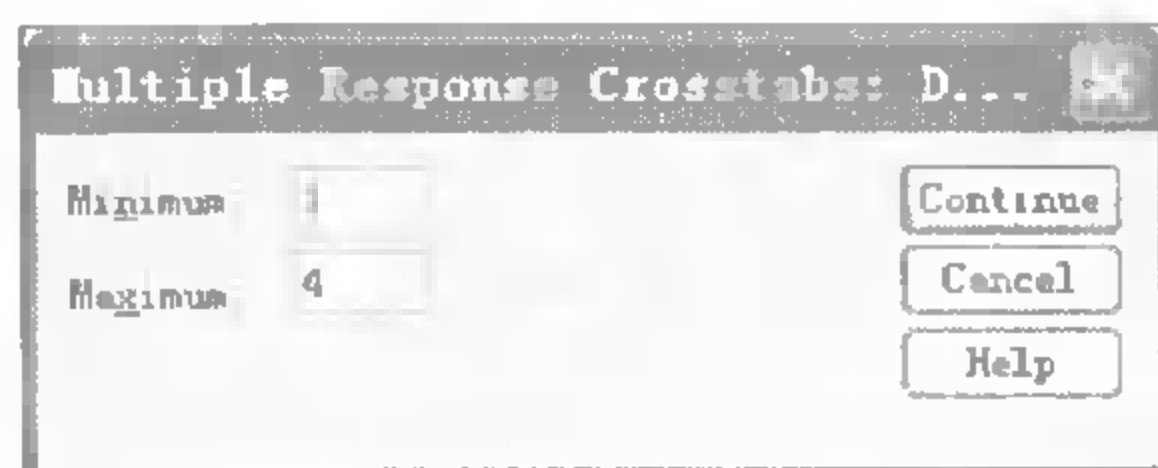


图 7-6 取值范围定义

变量列表显示了当前数据集中的可用变量；

- Mult Response 列表显示了当前定义的所有多重响应变量集;
- Row(s)列表用于选入输出表格的行变量;
- Column(s)列表用于选入输出表格的列变量;
- Layer(s)列表用于选入输出表格的分层变量,对分层变量的每个取值(或取值组合),将输出一个相应行列变量的二维交叉表。普通变量、多重响应变量集都可以作为行变量、列变量、分层变量中的任意一个。

选入 Row(s)、Column(s)和 Layer(s)列表框的普通变量,还必须为其设置取值范围,但不能设置变量集的取值范围。取值范围的定义在图 7-6 所示的对话框中进行。

在图 7-6 中, Minimum、Maximum 输入框分别用于指定要在输出表中显示的变量取值的最小值和最大值。设置后将在相应的变量名后用括号显示它的取值范围,如 custcat(1 4)。

(2) Options 选项设置。在图 7-5 中单击 Options 按钮,弹出如图 7-7 所示的选项设置对话框,勾选 Cell Percentages 栏下的 Column 复选框,单击 Continue 按钮返回主界面。

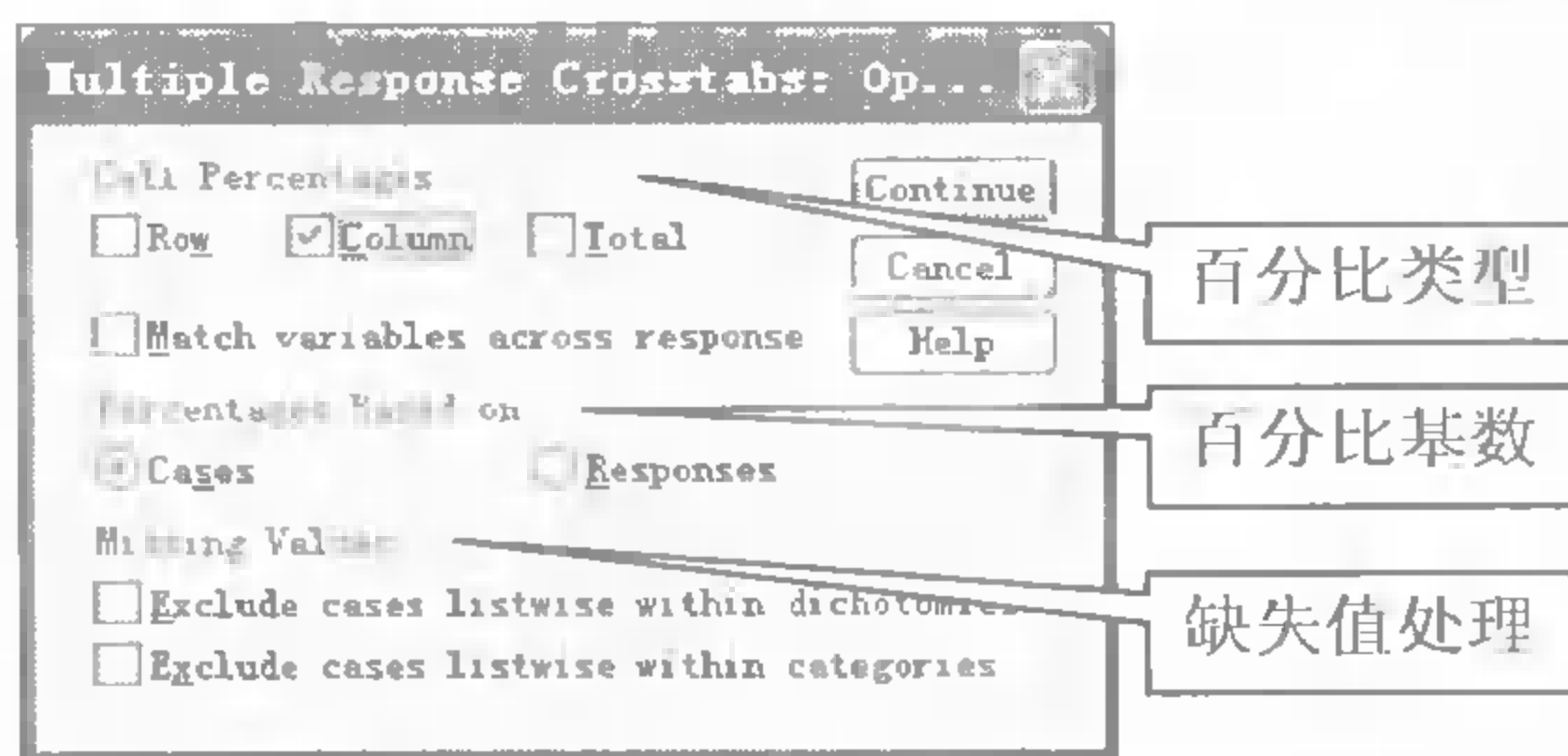


图 7-7 选项设置

- Cell Percentages 栏。选择在单元格显示哪些类型的百分比,有 3 个可选项: Row, 行百分比; Column, 列百分比; Total, 总百分比。另外,单元格总会显示观测的统计个数。
- Match variables across response sets 复选框。这是仅对多响应分类变量集起作用的选项。勾选它,表示把第 1 个变量集中的第 1 个变量和第 2 个变量集中的第 1 个变量作为一对,把第 1 个变量集中的第 2 个变量和第 2 个变量集中的第 2 个变量作为一对,依次类推;且单元格中的百分比以答案总数而不是回答者总数为基数。
- Percentages Based on 栏。设置计算百分比的基数,可选项有如下两个。
  - ☆ Cases 观测数,表示以回答人数为计算百分比的基数。
  - ☆ Responses 响应,表示以总的回答数为计算百分比的基数,由于是多项选择题,所以总的回答数一般要多于回答问题的人数。当勾选 Match 复选框后,只能使用此选项。
- Missing Values 栏。设置处理缺失值的方式,有如下两种备选方案。
  - ☆ Exclude cases listwise within dichotomies 复选框,适用于采用多项选择二分法编码的多重响应变量集,缺失记录不进入分析。默认情况下,如果多重响应变量集中的所有二元变量都没有取代表选中状态的计数值(例如本例“1”),相应的观测记录就被认为是缺失的;当至少有一个二元变量取计数值时,即使其他个别(不是全部)二元变量的取值有缺失,相应的观测记录也不会被当作缺失处理。
  - ☆ Exclude cases listwise within categories 复选框,适用于采用多项选择分类法编码的多重响应变量集,缺失记录不进入分析。默认情况下,只有当多重响应变量集中每个变量的取值都不在定义范围内时,相应的观测记录才被当作缺失记录处理。



### 3. 结果分析

单击图 7-5 中的 OK 按钮运行，SPSS Viewer 窗口的输出表格如图 7-8 所示。

个案摘要

	个案					
	有效的		缺失		总计	
	N	百分比	N	百分比	N	百分比
Sservices*custcat	889	88.9%	111	11.1%	1000	100.0%

Services\*custcat交叉制表

			客户种类				总计
			Basic service	E-service	Plus service	Total service	
服务 <sup>a</sup> 多线路使用	计数		0	217	94	164	475
	custcat 内的 %		0%	100.0%	33.5%	69.5%	
语音邮件	计数		17	28	53	206	304
	custcat 内的 %		11.0%	12.9%	18.9%	87.3%	
寻呼业务	计数		14	16	22	209	261
	custcat 内的 %		9.0%	7.4%	7.8%	88.6%	
Internet 服务	计数		63	110	22	173	368
	custcat 内的 %		40.6%	50.7%	7.8%	73.3%	
来电显示	计数		24	16	234	207	481
	custcat 内的 %		15.5%	7.4%	83.3%	87.7%	
呼叫等待	计数		23	19	239	207	485
	custcat 内的 %		14.8%	8.8%	85.1%	86.4%	
呼叫转移	计数		26	19	238	210	493
	custcat 内的 %		16.8%	8.8%	84.7%	89.0%	
三方通话	计数		32	27	238	205	502
	custcat 内的 %		20.6%	12.4%	84.7%	86.9%	
电子账单	计数		82	109	14	166	371
	custcat 内的 %		52.0%	50.2%	5.0%	70.3%	
总计	计数		155	217	281	236	889

百分比和总计以响应者为基础。

a. 值为 1 时制表的二分组。

图 7-8 多重响应 Crosstabs 过程的输出结果

(1) 个案摘要表。显示了对有效数据和缺失数据的基本统计信息，本例的 1 000 个案例中，有 111 个案例被认为是缺失的，即这 111 个案例是没有定制任何服务的客户。

(2) 交叉分析表。每个单元格显示了使用各种服务的不同种类的人数，以及以客户数为基数的列百分比。

以电子账单和基本服务的交叉单元格为例，表示此类客户共有 82 例，占基本服务类型客户总数（155 例）的 52.9%。其它单元格的解读方式与此类似。

## 7.5 使用 Tables 过程研究多重响应变量集


SPSS 的 Tables 过程也提供了对多重响应变量集进行定义和分析的功能。使用时也需要先建立一个多重响应变量集，然后再使用普通 Tables 过程进行分析，所起的作用和效果和前节的交叉表分析类似。注意：在 Tables 过程建立的多重响应变量集不能在其它分析过程（如 Multiple Response 过程的频数分析和交叉表分析）中被识别和使用。

本节仍使用第 7.2 节的例子数据介绍通过 Tables 过程做多重响应变量集分析的过程，数据格式如图 7-1 所示。

### 7.5.1 多重响应变量集的定义

依次单击菜单“Analyze→Tables→Multiple Response Sets...”执行 Tables 过程的定义多重



响应变量集功能，其主设置界面如图 7-9 所示。在变量列表选中从多线路使用至电子账单的 9 个变量，单击  按钮，将其选入 Variables in Set 列表；在单选项 Dichotomises Counted value 后输入“1”；在 Name 后输入“services”，在 Label 后输入“服务”；单击 Add 按钮，将这个定义好的多重响应变量集选入右侧的 Mult Response Sets 列表。单击 OK 按钮完成关于变量集的定义。

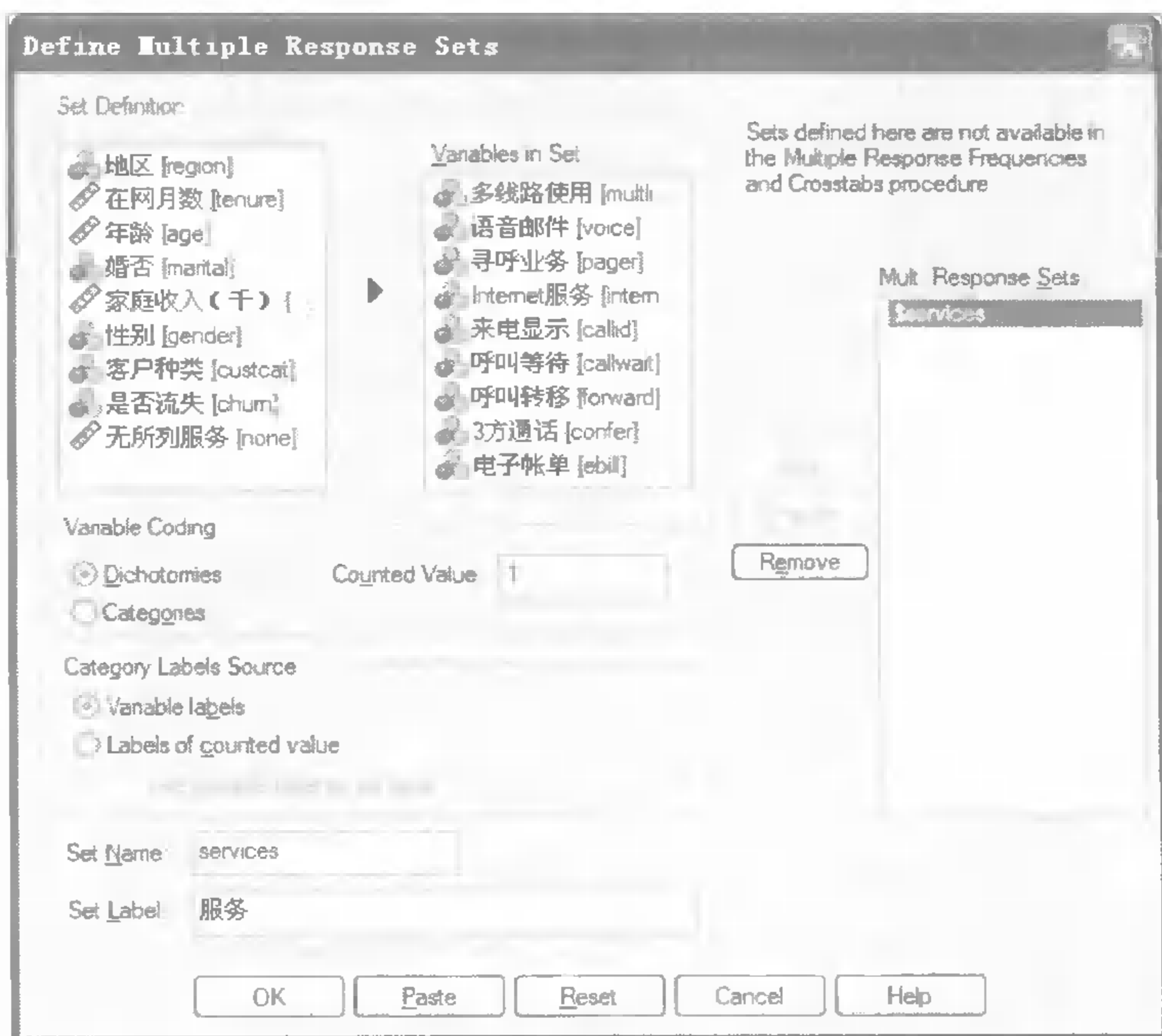


图 7-9 Tables 过程的定义多重响应变量集界面

此设置界面和图 7-2 所示的变量集定义界面基本相同，只是多了一个名为 Category Label Source 的子设置栏，在此设置关于多重响应二分类变量集的输出表格的标签格式，有如下两个选项。

(1) Variable labels 变量标签。使用变量标签（没有定义标签时使用变量名）作为变量集所包含变量的标签。例如：当变量集中有多个变量的取值相同（如都能取“Yes”）时，如果对这个相同的取值（如“Yes”）定义了相同的值标签（或者没有定义值标签），就应该选中此选项。

(2) Labels of counted values 值标签。使用响应变量值（表示选中状态的取值，如本例中的“1”）的标签作为变量集所包含变量的标签，如果对变量集中的所有响应变量值都定义了不同的值标签，选中此选项。激活 Use variable label as set label 复选框，表示用变量集中第一个变量的标签作为当前变量集的标签，如果变量集中的所有变量都没有定义标签，就使用第一个变量名作为当前变量集的标签。

### 7.5.2 用 Tables 过程建立包含多重响应变量集的表格

上节通过 Tables 菜单的 Multiple Response Sets 功能建立了多重响应变量集“services”，随后就利用 Tables 过程来对它进行分析。依次单击菜单“Analyze→Tables→Custom Tables...”打开用户定制表格的对话框，其主设置界面如图 7-10 所示。

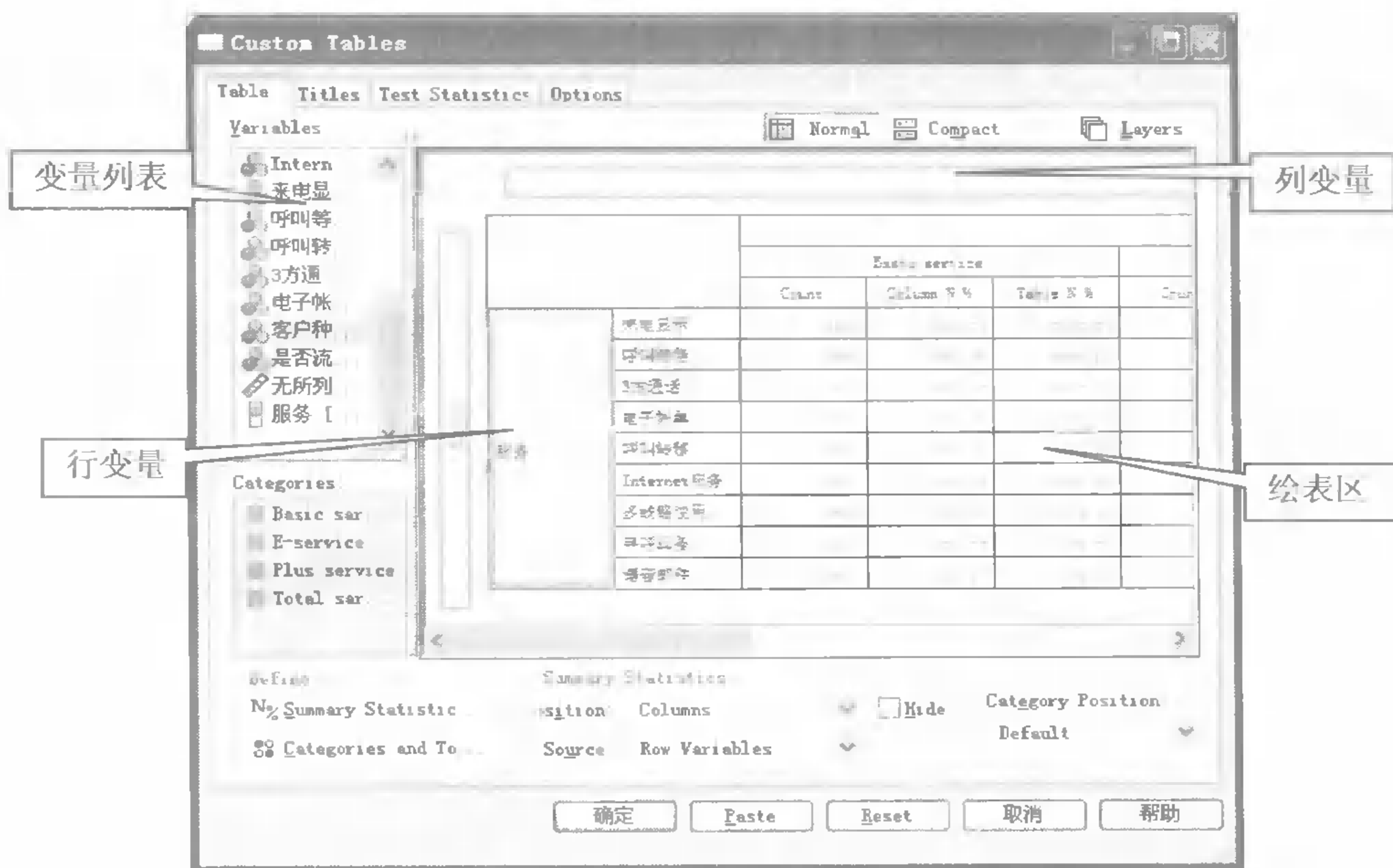


图 7-10 Custom Tables 过程的主设置界面

## 1. 参数设置

在变量列表中单击选中服务（\$services）变量集，将其拖动至绘表区的 Rows 行变量区；在变量列表中单击选中客户种类（custcat）变量，将其拖动至绘表区的 Columns 列变量区。

在绘表区右击服务变量所在的单元格，在弹出的快捷菜单里单击 Summary Statistics 选项，弹出如图 7-11 所示的统计量选择对话框，设置将要在输出表格中显示的统计量。在左侧的统计量列表选中 Column N%（列比例）和 Table N%（总比例）后，单击右侧的黑色箭头，将其选入 Display 列表；单击 Apply to Selection 按钮确认修改，并返回主界面。

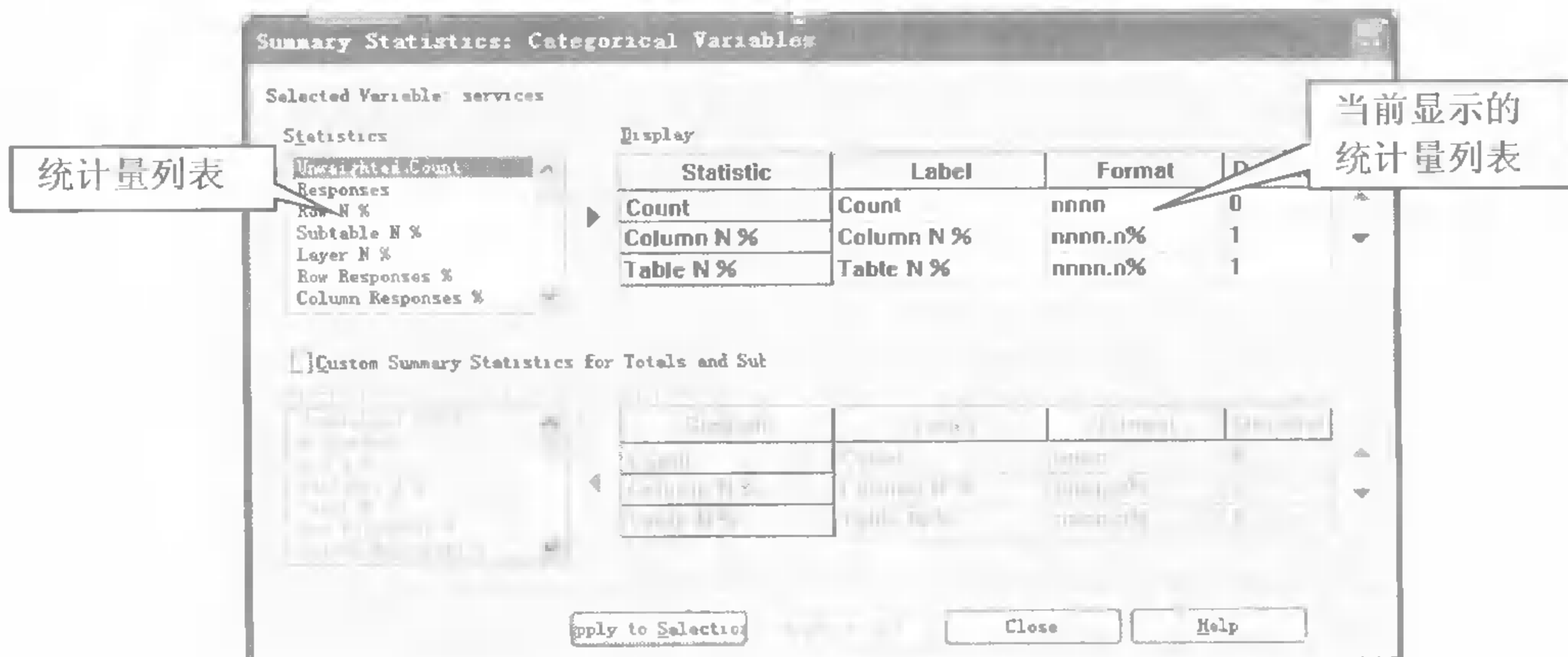


图 7-11 Custom Tables 过程的统计量选择对话框

## 2. 输出结果

在图 7-10 中，单击确定按钮运行，SPSS Viewer 窗口的输出表格如图 7-12 所示。其中“计数”和“列 N%”对应单元格的统计信息与图 7-8 完全一样。

		客户种类											
		Basic service			E-service			Plus service			Total service		
		计数	列 N°	表 N°	计数	列 N°	表 N°	计数	列 N°	表 N°	计数	列 N°	表 N°
服务	多线路使用	9	0%	0%	217	100.0%	34.4%	91	33.3%	10.0%	164	66.7%	18.4%
	语音邮件	17	11.0%	1.8%	28	12.9%	3.1%	16	12.5%	6.0%	308	47.3%	23.2%
	寻呼业务	14	9.0%	1.0%	16	7.4%	1.8%	22	16.8%	2.3%	209	88.6%	23.3%
	Internet服务	93	40.4%	1.1%	110	50.7%	12.4%	22	16.5%	1.3%	173	75.3%	17.3%
	来电显示	24	15.3%	2.7%	46	21.0%	1.8%	134	81.3%	26.3%	207	87.7%	23.3%
	呼叫等待	13	14.8%	2.6%	17	8.8%	2.1%	119	85.1%	26.9%	204	86.4%	22.9%
	呼叫转移	26	16.8%	2.9%	19	8.8%	2.1%	158	84.7%	26.8%	210	89.4%	23.0%
	三方通话	30	20.6%	3.6%	27	12.4%	3.0%	158	84.7%	26.8%	200	86.9%	23.1%
	电子帐单	82	52.8%	9.2%	109	50.7%	11.3%	13	13.0%	1.6%	166	70.5%	18.7%

图 7-12 Tables 过程关于多重响应变量集的表格输出

回归分析是通过试验和观测来寻找变量之间关系的一种统计分析方法，它的理论比较成熟，而且应用十分广泛。回归分析的主要目的在于了解自变量（independent variable）与因变量（dependent variable）之间的数量关系，它的研究内容包括：探索和确定变量之间的相关关系和相关程度；建立回归模型，检验变量之间的相关程度；用回归模型进行估计和预测等。

当研究因变量  $Y$  与自变量  $x$  之间的相关关系时， $Y$  常常是随机变量，它对于给定的  $x$  值有特定的分布，不妨用  $F(y|x)$  表示  $x$  取某一确定的值时， $Y$  所对应的条件分布函数。如果掌握了  $F(y|x)$  随着  $x$  取值的变化而变化的规律，也就完全掌握了  $Y$  与  $x$  之间的关系，然而这样做往往非常复杂甚至是不可能的。作为一种近似，可以转而研究  $Y$  的期望，如果  $Y$  关于  $x$  的条件数学期望  $E(y|x)$  存在，并且是  $x$  的函数，记为： $\mu(x)$ ，称之为  $Y$  关于  $x$  的回归函数，那么讨论  $Y$  与  $x$  相关关系的问题，就转化为讨论  $E(Y|x) = \mu(x)$  这个关于  $x$  的函数问题了。根据  $\mu(x)$  的形式不同，可以定义出线性回归、非线性回归等多种回归分析方法。

## 8.1 线性回归

当采用模型  $E(Y) = \mu(X) + \varepsilon$ ,  $X = (x_1, x_2, \dots, x_n)$  研究  $Y$  与  $X$  之间的关系时，如果  $\mu$  是一个线性函数，则进行的回归分析就是线性回归，其中： $Y$  是因变量， $X$  是自变量， $\varepsilon$  是随机变量（或称为随机误差）。

### 8.1.1 一元线性回归的基本原理

在实际应用中，最简单的情形就是研究两个变量之间的相关关系，即一元线性回归。

#### 1. 一元线性回归方程

设  $(x_1, Y_1), (x_2, Y_2), \dots, (x_n, Y_n)$  是取自总体  $(x, Y)$  的一组样本， $x_1, x_2, \dots, x_n$  是取定的不完全相同的数值（自变量）， $y_1, y_2, \dots, y_n$  是  $Y_1, Y_2, \dots, Y_n$  在试验或观测后的测量值（因变量），由此抽样的结果可以取得  $n$  对样本数据  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ 。于是，一元线性回归方程的形式为： $y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$ ,  $i = 1, 2, \dots, n$ ，其中  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$  需要满足如下 4 个假设条件。

① 正态性假设，即  $\varepsilon_i$  是服从正态分布的随机变量。

② 无偏性假设，即： $E(\varepsilon_i) = 0$ 。

③ 同共方差性假设，即所有  $\varepsilon_i$  的方差都相同；同时也说明了  $\varepsilon_i$  与自变量、因变量之间都是相互独立的。

④ 独立性假设,  $\varepsilon_i$  之间相互独立, 且满足  $COV(\varepsilon_i, \varepsilon_j) = 0, (i \neq j)$ 。

线性回归分析, 就是要根据已有样本的观察值, 寻求  $\beta_0, \beta_1$  的合理估计值  $\hat{\beta}_0, \hat{\beta}_1$ 。对样本中的每个  $x_i$ , 由一元线性回归方程可以确定一个关于  $y_i$  的估计值  $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_i$ , 称之为  $Y$  关于  $x$  的线性回归方程或经验回归公式, 鉴于其线性性质, 也称之为回归直线。

## 2. 一元回归方程系数的普通最小二乘估计

实际观察值  $y_i$  与估计值  $\hat{y}_i$  之差  $y_i - \hat{y}_i = y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i$ , 反映了观察值  $y_i$  与回归直线  $\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$  的偏离程度。令:  $Q(\beta, \beta) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$ ,  $Q$  表示所有的观察值  $y_i$  与回归直线  $\hat{y}_i$  的偏离平方和, 刻画了所有观察值与回归直线的偏离程度。最小二乘法就是寻求使得  $Q(\beta, \beta)$  达到最小的参数估计值  $\hat{\beta}_0, \hat{\beta}_1$ 。

要使  $Q$  取到极小值, 求它关于  $\hat{\beta}_0, \hat{\beta}_1$  的偏导数并令其为零, 即可求得  $\hat{\beta}_0, \hat{\beta}_1$  的最小二乘估计

$$\text{为} \begin{cases} \hat{\beta}_0 = \bar{y} - \bar{x}\hat{\beta}_1 \\ \hat{\beta}_1 = L_{xy}/L_{xx} \end{cases}, \text{其中:} \begin{cases} L_{xx} \stackrel{\text{def}}{=} \sum_{i=1}^n (x_i - \bar{x})^2 = \sum_{i=1}^n x_i^2 - n\bar{x}^2 \\ L_{xy} \stackrel{\text{def}}{=} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - n\bar{x}\bar{y} \end{cases}, \begin{cases} \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \\ \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i \end{cases}$$

由此就得到了关于样本的一元线性回归方程  $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 x$ 。

## 3. 最小二乘估计的性质

若  $\hat{\beta}_0, \hat{\beta}_1$  为  $\beta_0, \beta_1$  的最小二乘估计, 则  $\hat{\beta}_0, \hat{\beta}_1$  分别是  $\beta_0, \beta_1$  的无偏估计, 且分布形式分别为:

$$\hat{\beta}_0 \sim N\left(\beta_0, \sigma^2 \left(\frac{1}{n} + \frac{\bar{x}^2}{L_{xx}}\right)\right), \quad \hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{L_{xx}}\right).$$

## 4. 回归方程的检验

由线性回归模型  $Y = \beta_0 + \beta_1 x + \varepsilon$ ,  $\varepsilon \sim N(0, \sigma^2)$  可知, 当  $\beta_1 = 0$  时, 就认为  $Y$  与  $x$  之间不存在线性回归关系, 故需要检验如下的假设:  $H_0: \beta_1 = 0, H_1: \beta_1 \neq 0$ 。对  $H_0$  的检验有三种本质相同的检验方法: 相关系数检验、 $F$  检验、 $t$  检验。

首先, 记  $SST = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2 + \sum_{i=1}^n (y_i - \hat{y}_i)^2 = SSR + SSE$ , 此式称为总体平方和的分解公式, 其中:  $SST$  称为总体平方和, 代表原始数据所反映的总偏差的大小;  $SSR$  称为回归平方和 (可解释误差), 它是由变量  $x$  引起的偏差, 反映了  $x$  的重要程度;  $SSE$  称为剩余平方和 (不可解释误差), 它是由试验误差以及其它未加控制因素引起的偏差, 反映了试验误差及其它随机因素对试验结果的影响。

(1) 相关系数检验。相关系数定义为可解释误差  $SSR$  和总误差  $SST$  之比, 即  $r^2 = SSR/SST = 1 - SSE/SST$ 。它反映了由于使用  $Y$  与  $X$  之间的线性回归模型来估计  $y_i$  的均值, 而导致总体平方和  $SST$  减少的程度, 从而代表了  $Y$  与  $X$  之间的线性相关程度及回归模型的拟合优良程度。 $r^2$  与  $SSR$  成正比,  $r^2$  越大, 说明  $Y$  与  $X$  之间的线性相关程度越高, 也就说明模型的拟合优度较好;  $r^2$  越小, 说明  $Y$  与  $X$  之间的线性相关程度越低, 即模型的拟合优度较差。



(2)  $F$  检验。在  $H_0$  成立时,  $F$  统计量  $F = \frac{\sum(\hat{y}-\bar{y})^2/1}{\sum(y-\hat{y})^2/(n-2)} = \frac{SSR}{SSE}(n-2) = \frac{r^2}{1-r^2}(n-2)$ , 服从自由度为  $(1, n-2)$  的  $F$  分布  $F(1, n-2)$ 。给定显著性水平  $\alpha$  后, 可以确定临界值  $F_\alpha(1, n-2)$ , 再根据试验数据  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  计算得到  $F$  统计量的取值, 若  $F > F_\alpha(1, n-2)$  时, 拒绝  $H_0$ , 表明回归效果显著; 若  $F \leq F_\alpha(1, n-2)$  时, 接受  $H_0$ , 此时回归效果不显著。

(3)  $t$  检验。由于  $\hat{\beta}_1 \sim N\left(\beta_1, \frac{\sigma^2}{L_{xx}}\right)$ , 以及  $\hat{\sigma}^2 = SSE/(n-2)$  为  $\sigma^2$  的无偏估计, 当  $H_0$  成立时, 可取如下的  $t$  统计量来做检验:  $t = \frac{\hat{\beta}_1}{\hat{\sigma}} \sqrt{L_{xx}} \sim t(n-2)$ 。给定显著性水平  $\alpha$  后, 可以确定临界值  $t_{\alpha/2}(n-2)$ , 再根据试验数据  $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$  计算  $t$  统计量的值, 当  $|t| > t_{\alpha/2}(n-2)$  时, 拒绝  $H_0$ , 这时回归效果显著; 当  $|t| \leq t_{\alpha/2}(n-2)$  时, 接受  $H_0$ , 此时回归效果不显著。

### 8.1.2 多元线性回归的基本原理

在许多实际问题中, 常常需要研究一个因变量与多个自变量之间的相关关系, 例如: 某种产品的销售额不仅受到投入广告费用的影响, 通常还与产品的价格、消费者的收入状况、社会保有量以及其它可替代产品的价格等因素有关系。此时, 最常使用的统计分析方法就是多元线性回归分析, 它是一元线性回归分析的推广形式, 两者在参数估计、显著性检验等方面都有许多相似之处。

#### 1. 多元线性回归的数学模型

设影响因变量  $Y$  的自变量个数为  $p$ , 并分别记为  $x_1, x_2, \dots, x_p$ , 所谓多元线性模型是指这些自变量对  $Y$  的影响是线性的, 即有  $Y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + \varepsilon$  成立, 其中  $\varepsilon \sim N(0, \sigma^2)$ ,  $\beta_0, \beta_1, \beta_2, \dots, \beta_p, \sigma^2$  是与  $x_1, x_2, \dots, x_p$  无关的未知参数, 称上式为因变量  $Y$  对自变量  $x_1, x_2, \dots, x_p$  的多元线性回归方程。

记  $n$  组样本分别是  $(x_{i1}, x_{i2}, \dots, x_{ip}, y_i), (i=1, 2, \dots, n)$ , 令

$$Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix}, \quad X = \begin{pmatrix} 1 & x_{11} & x_{12} & \cdots & x_{1p} \\ 1 & x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{np} \end{pmatrix}, \quad \beta = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \varepsilon = \begin{pmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{pmatrix},$$

则多元线性回归方程的矩阵形式为  $Y = X\beta + \varepsilon$ , 其中  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$  仍然服从第 8.1.1 节提出的正态性、无偏性、同方差性、独立性四个假设。

#### 2. 最小二乘估计

与一元线性回归类似, 可以采用最小二乘法估计参数  $\beta_0, \beta_1, \beta_2, \dots, \beta_p$ , 先引入偏差平方和  $Q(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \dots - \beta_p x_{ip})^2$ 。最小二乘估计就是求使得  $Q$  达到最小的  $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)^T$ , 由于  $Q(\beta_0, \beta_1, \dots, \beta_p)$  是  $\beta_0, \beta_1, \dots, \beta_p$  的非负二次型, 故其最小值一定存在。根据多元微积分的极值原理, 对  $Q$  求偏导数并令其为零, 可解得  $\hat{\beta} = (X^T X)^{-1} X^T Y$ , 这就是  $\beta$  的最小二乘估计。

### 3. 显著性检验

由于事先并不能确定Y和X的相关关系为何种类型,只是假设它们之间存在着线性关系,所以在建立了多元线性回归方程之后,还必须对因变量与自变量之间存在线性关系的假设进行显著性检验,或者说对多元线性回归方程的成立与否进行显著性检验。

此时,仍有关于总体平方和的分解公式成立:  $SST = SSR + SSE$ , 各部分的解释同一元回归时的情形相同;而且,这3个部分的自由度分别为:  $df_T = n-1$ ,  $df_R = p$ ,  $df_E = n-p-1$ , 其中:  $p$  为自变量的个数,  $n$  为实际观测数据的组数。进一步可以得到回归均方和  $MSR$  与残差均方和  $MSE$ :  $MSR = SSR/df_R$ ,  $MSE = SSE/df_E$ 。

(1) 相关系数检验同一元回归时的情形类似。

(2)  $F$  检验。检验多元线性回归方程是否显著成立,即检验各自变量的总体回归系数  $\beta_i (i=1,2,\dots,p)$  是否同时为零,零假设与备择假设分别为:  $H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0$ ,  $H_1: \beta_1, \beta_2, \dots, \beta_p$  不全为零。当  $H_0$  成立时,取  $F$  统计量  $F = MSR/MSE \sim F(df_R, df_E)$ , 由此进行  $F$  检验即可推断多元线性回归关系的显著性。

相关系数检验和  $F$  检验实质上检验的都是因变量与所有自变量的综合线性关系是否显著成立,即使它验证了多元线性回归方程是显著的,也不能确定每一个自变量与因变量的线性关系都是显著的,无法区分哪些自变量对因变量的线性影响是显著的。因此,当多元线性回归关系经显著性检验为显著时,还必须逐一对各回归系数进行显著性检验,以发现和剔除不显著的回归系数所对应的自变量。

(3) 回归系数的  $t$  检验。对单个回归系数的显著性检验,建立相应的零假设与备择假设分别为,  $H_0: \beta_i = 0$ ,  $H_1: \beta_i \neq 0$ , ( $i=1,2,\dots,p$ )。

已知  $COV(\hat{\beta}) = \sigma^2 (X'X)^{-1}$ , 以  $c_{ii}$  记矩阵  $(X'X)^{-1}$  主对角线上的第  $i$  个元素,于是参数估计量的方差为  $Var(\hat{\beta}_i) = \sigma^2 c_{ii}$ ; 另外,取  $MSE = SSE/df_E$  作为  $\sigma^2$  的估计量。在  $H_0$  成立时,取  $t$  统计量

$t_i = \frac{\hat{\beta}_i}{\sqrt{MSE \cdot c_{ii}}} \sim t(n-p-1)$ , 由此进行  $t$  检验即可推断多元线性回归系数  $\hat{\beta}_i (i=1,2,\dots,p)$  的显著性。

#### 8.1.3 模型假设的其他检验

在线性回归模型基本假设下,应用普通最小二乘法可以得到无偏的、有效的参数估计量。但在实际生活中,完全满足这些基本假设的情况并不多见,如果违背了其中的某一项,那么应用普通最小二乘法就不能得到无偏的、有效的参数估计值了,就需要发展新的方法估计模型。这里的假设主要是指  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n$  的正态性、无偏性、同方差性、独立性四个假设,下面对这些假设的验证方法加以简单描述。

(1) 残差的正态性检验。回归模型假设随机误差  $\varepsilon_i$  服从正态分布,这一点可以通过建立标准残差  $E_j = \varepsilon_j / \hat{\sigma}_\varepsilon$  的直方图来检验。理论上,  $E_j$  应服从标准正态分布  $N(0,1)$ , 所以应有近 50% 的  $E_j$  为正, 50% 的  $E_j$  为负; 68% 的  $E_j$  落在 -1 与 +1 之间, 96% 的  $E_j$  落在 -2 与 +2 之间。当样本的容量不太大时,  $E_j$  在理论上应服从于自由度为  $n-p-1$  的  $t$  分布。

另外,  $P-P$  图也是用来检验变量分布与指定分布是否一致的好方法。

(2) 残差的方差齐性检验。回归模型假设随机误差  $\varepsilon_i$  具有相同的方差,这一点可以通过残差散点图来验证。以残差  $\varepsilon_i$  为纵坐标,以估计值为横坐标作图,如果观察点随机地散布在

横轴的周围,就说明残差基本符合同方差性假设。当此假设被否定,残差出现了异方差的情况时,就需要先对原始数据进行适当的变量转换,再利用回归模型进行估计和预测,使方差趋于稳定。

(3) 残差的独立性检验。回归模型还假设随机误差  $\varepsilon_i$  之间相互独立,这一点也可以通过残差散点图来验证。采用和方差齐性检验中相同的图形观察和分析点的散布情况,如果观察点在横轴的周围显示出周期性或趋势性的变化,就说明残差不符合独立性的假设。

(4) 多重共线性检验。建立多元线性回归模型时,如果有两个或两个以上的自变量之间存在线性相关关系,就会产生多重共线性现象。在这种情况下,用最小二乘法估计的模型参数就会很不稳定;而且,当模型中增加或减少一个变量时,已进入模型中的变量的系数也会发生较大变化。在多重共线性现象较为严重的情况下,回归系数的估计值很容易引起误导或导致错误的结论。如果自变量完全线性相关,那么参数就成为不确定的了。

通过容许度 (Tolerance):  $Tol_i = 1 - R_i^2$  或方差膨胀因子 (VIF):  $VIF_i = 1/(1-R_i^2)$ , 可以检验共线性的存在,其中  $R_i^2$  是用其他自变量预测第  $i$  个变量的复相关系数。显而易见 VIF 为 Tol 的倒数, Tol 的值越小, VIF 的值越大,自变量  $x_i$  与其他自变量之间存在共线性的可能性越大。

当确定自变量之间存在明显的共线性时,可用如下几种方法加以处理。

- ① 从有共线性问题的变量里删除不重要的变量。
- ② 增加样本量或重新抽取样本。
- ③ 采用其他方法拟合模型,如岭回归法、逐步回归法、主成分分析法等。

#### 8.1.4 问题描述和数据准备

本节通过对某些汽车的销售量及这些汽车的一些特征数据拟合多元线性回归模型,分析汽车特征与销售量之间的关系,并利用回归结果给出改进汽车设计方案的建议,以促进销售量的提高。数据摘自 SPSS 自带的 Demo 文件“car\_sales.sav”,所用数据文件为“汽车销售数据.sav”,数据格式如图 8-1 所示。其中,“销售量”的数据为对数转换形式,其分布近似为正态分布,如此能更好地拟合线性回归模型。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	lnsales	Numeric	8	2	销售量	None	None	8	Right	Scale
2	type	Numeric	11	0	车型	{0, Automobile}..	None	8	Right	Ordinal
3	price	Numeric	11	3	价格	None	None	8	Right	Scale
4	engine_s	Numeric	11	1	发动机规格	None	None	8	Right	Scale
5	horsepow	Numeric	11	0	马力	None	None	8	Right	Scale
6	wheelbas	Numeric	11	1	轴距	None	None	8	Right	Scale
7	width	Numeric	11	1	宽度	None	None	8	Right	Scale
8	length	Numeric	11	1	长度	None	None	8	Right	Scale
9	curb_wgt	Numeric	11	3	净重	None	None	8	Right	Scale
10	fuel_cap	Numeric	11	1	燃料箱容量	None	None	8	Right	Scale
11	mpg	Numeric	11	0	燃料效率	None	None	8	Right	Scale

图 8-1 汽车销售分析数据格式

#### 8.1.5 线性回归分析的设置和操作

依次单击菜单“Analyze→Regression→Linear...”执行线性回归分析的功能,其主设置界面如图 8-2 所示,在此选择进行分析的变量。



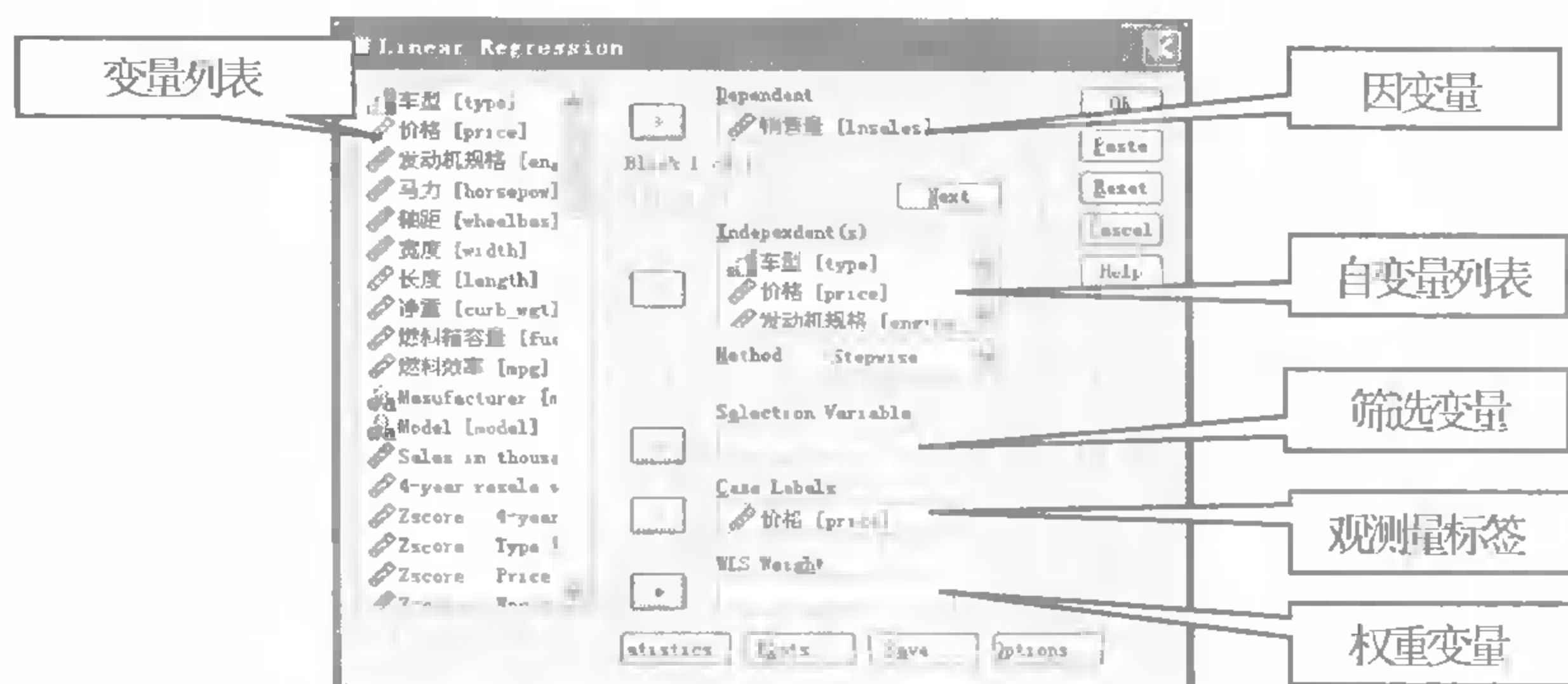





图 8-2 线性回归分析的主设置界面

### 1. 变量设置

在变量列表中单击选中销售量变量，单击从上至下第一个  按钮，将其作为因变量选入 Dependent 选框；在变量列表选中从车型至燃料效率的 10 个变量，单击从上至下第二个  按钮，将其作为自变量选入 Independent(s) 列表；在变量列表单击选中价格变量，单击从上至下第四个  按钮，将其作为标签变量选入 Case Labels 选框；单击 Method 后的下拉列表，选中 Stepwise 选项。

(1) 指定分析变量。Dependent 栏用于选入线性回归分析的一个因变量；Independent(s) 栏用于选入分析的自变量；Case Labels 栏选入标签变量，用于在图形中对观测记录进行标注，最典型的情况就是用观测记录的 ID 号作为标签变量；WLS weight 栏选入权重变量，主要用于加权最小二乘法。

(2) 关于自变量的分组设置。Block 栏由 Previous 和 Next 两个按钮组成，用于对其下的 Independent(s) 栏中指定的自变量进行分组。多元回归分析中自变量的选入方式有强行进入法、向前选择法、逐步法等，如果需要对不同的自变量采用不同的选入方法，就需要使用该设置栏将自变量进行分组。

具体操作步骤如下：先将自变量 x1、x2 选入 Independent(s) 列表，并在 Method 栏选中 Enter 方法；然后单击 Next 按钮，Independent(s) 列表被清空，重新选入自变量 x3、x4，并在 Method 栏选中 Stepwise 方法；这样，就建立了两个自变量组 (Block)，建模时每组自变量将会采用它们各自不同的进入方法 (Method)；单击 Previous 和 Next 按钮，可在不同的自变量组之间切换和编辑。

(3) 指定建模过程的变量选择方法。Method 栏用于指定建模时的变量选择方法，其后的下拉菜单有如下几个选项。

- Enter 强行进入法，Independent(s) 栏中所有的自变量全部进入回归模型，是默认方式。
- Remove 强制剔除法，建立回归方程时，根据设定的条件直接剔除部分自变量。
- Backward 向后消去法，先建立饱和模型，然后根据在 Options 对话框中所设定的参数，每次剔除一个不符合进入模型条件的变量。
- Forward 向前选择法，模型从没有自变量开始，根据 Options 对话框中所设定的参数，每次将一个最符合条件的变量引入模型，直至所有符合条件的变量都进入模型为止，第一个引入回归模型的自变量应该是与因变量最为相关的。
- Stepwise 逐步回归法，是向前选择法和向后消去法的结合。根据 Options 对话框中所设定的参数，先选择对因变量贡献最大且符合判断条件的自变量进入回归方程，再将模型中不符合设定条件的变量剔除。当没有变量被引入或删除时，得到最终回归方程。

(4) 指定筛选变量。Selection Variable 栏用于选入一个对样本的筛选变量，只有满足指

定条件的观测记录才会进入回归分析过程。选入变量后单击 Rules 按钮, 弹出如图 8-3 所示的设置对话框。

Define Selection Rule 下面显示当前选入的筛选变量名(如“mpg”); 在随后的下拉菜单中指定逻辑条件, 可选项包括: equal to (等于)、not equal to (不等于)、less than (小于)、less than or equal to (小于或等于)、greater than (大于) 和 greater than or equal to (大于或等于); Value 下面的输入框指定逻辑条件需要满足的临界值。设置好后就形成一个完整的筛选条件, 比如“mpg greater than 20”就表示只选择那些满足“mpg>20”的观测记录进行分析。

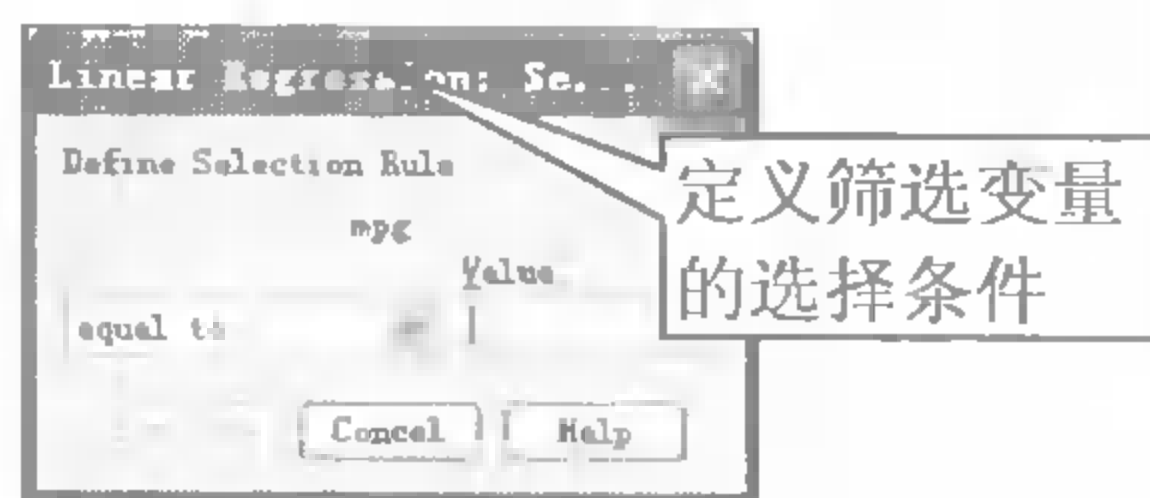


图 8-3 选择条件的定义对话框

## 2. 统计量设置

在图 8-2 中, 单击 Statistics 按钮, 弹出如图 8-4 所示的统计量设置子对话框。依次勾选如下几个复选框: Estimates、Covariance matrix、Model fit、Collinearity diagnostics、Casewise diagnostics; 单击 Continue 按钮返回主界面。

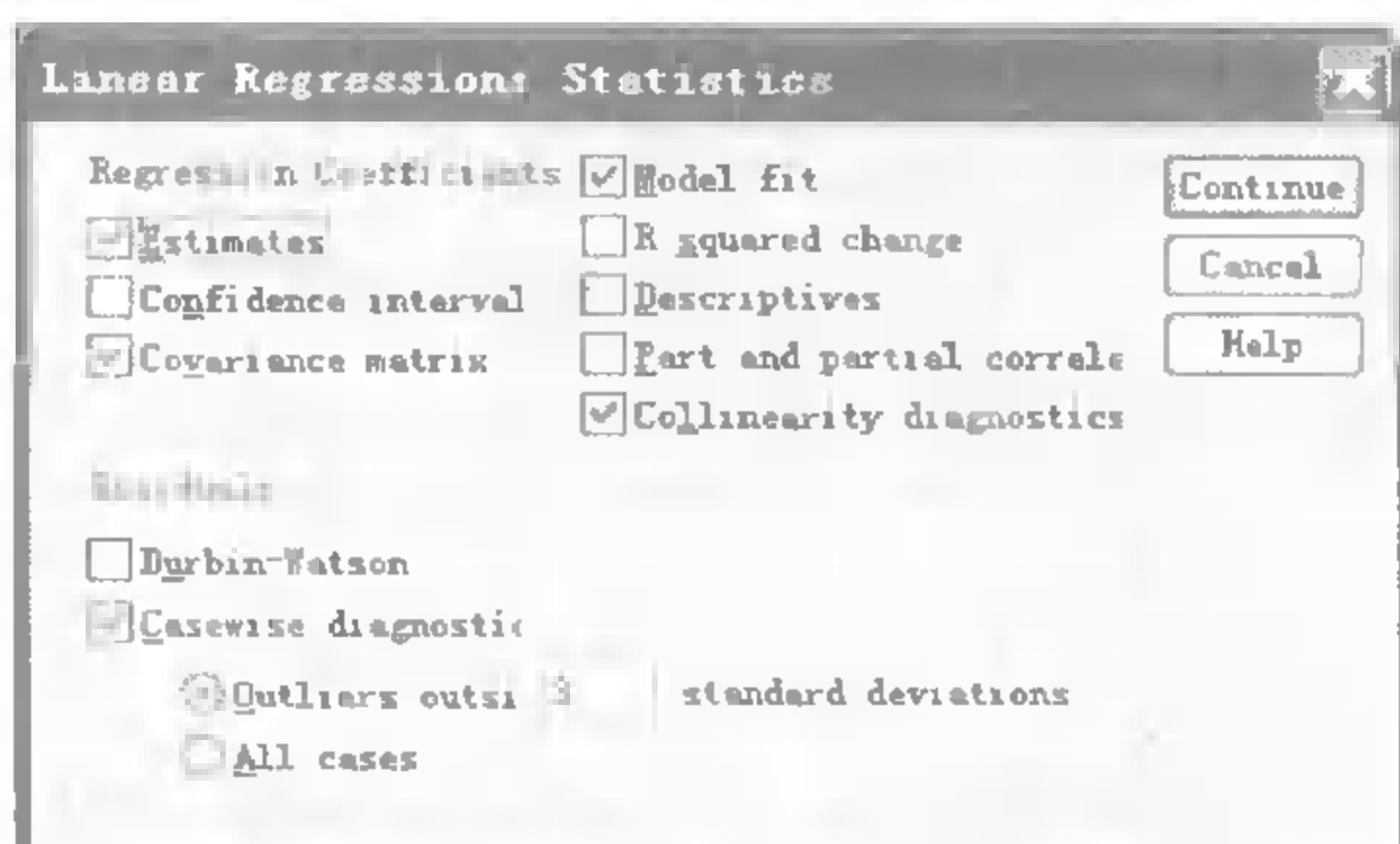


图 8-4 线性回归分析的统计量设置子对话框

(1) Regression Coefficients 子设置栏。用于选择关于回归系数的输出情况, 有如下 3 个可选项。

① Estimates 估计值, 输出回归系数、回归系数的标准误、标准化的回归系数、回归系数的  $t$  检验值及其双侧显著性水平 (Sig 值) 等内容。

② Confidence intervals 置信区间, 输出每个回归系数的 95% 的置信区间。

③ Covariance matrix 协方差阵, 输出回归系数的方差、协方差阵, 同时输出相关系数阵。

(2) 其他可选参数。

① Model fit 复选框, 输出模型中引入或剔除的自变量统计信息、拟合优度统计量、复相关系数  $R$  和  $R^2$  及其修正值、估计值的标准误及 ANOVA 方差分析表。

②  $R$  squared change 复选框, 输出模型中引入或剔除一个自变量所产生的  $R^2$  改变量。 $R^2$  改变量越大, 表明该自变量对模型的贡献越大, 说明其可能是一个较好的回归自变量。

③ Descriptives 复选框, 输出描述性统计量, 包括分析中每个变量的有效个案例数、平均数、相关系数矩阵及其单侧显著性水平等。

④ Part and partial correlations 复选框, 输出部分相关系数和偏相关系数。

⑤ Collinearity diagnostics 复选框, 输出共线性诊断的结果, 包括特征根 (Eigenvalues) 和方差膨胀因子 (VIF) 等。

(3) Residuals 子设置栏设置关于残差诊断的选项, 可选项有如下 2 个。

① Durbin-Watson (D-W 检验统计量)

用来检测回归分析中的残差项是否存在自相关现象, 同时会输出可能是异常值的观测值诊断表。D-W 统计量的取值范围是 0~4, 当残差一阶正相关时 D-W 接近 0, 当残差一阶负相关时 D-W 接近 4, D-W 接近 2 时残差独立。

② Casewise diagnostics (个案诊断)

① Outlier outside  $n$  standard deviations 选项, 设置判定异常值的依据。只有残差超过  $n$



倍标准差的观测才被当作是异常值,  $n$  为在后面的输入框指定的数字, 默认值为 3。

- All cases 选项, 输出所有观测的残差值。

### 3. 图形设置

在图 8-2 中, 单击 Plots 按钮, 弹出如图 8-5 所示的图形设置子对话框, 在此选择需要绘制的回归分析诊断图或预测图。在变量列表单击选中 “\*SDRESID” 变量, 单击 Y 选框左侧的 ☐ 按钮, 将其选入作为绘图的 Y 轴变量, 在变量列表单击选中 “\*ZPRED” 变量, 单击 X 选框左侧的 ☐ 按钮, 将其选入作为绘图的 X 轴变量; 单击 Next 按钮进入下一组绘图变量的选择, 在变量列表中单击选中 “\*ZRESID” 变量, 单击 Y 选框左侧的 ☐ 按钮, 将其选入作为绘图的 Y 轴变量, 在变量列表中单击选中 “\*ZPRED” 变量, 单击 X 选框左侧的 ☐ 按钮, 将其选入作为绘图的 X 轴变量。勾选 Histogram 复选框; 单击 Continue 按钮返回主界面。

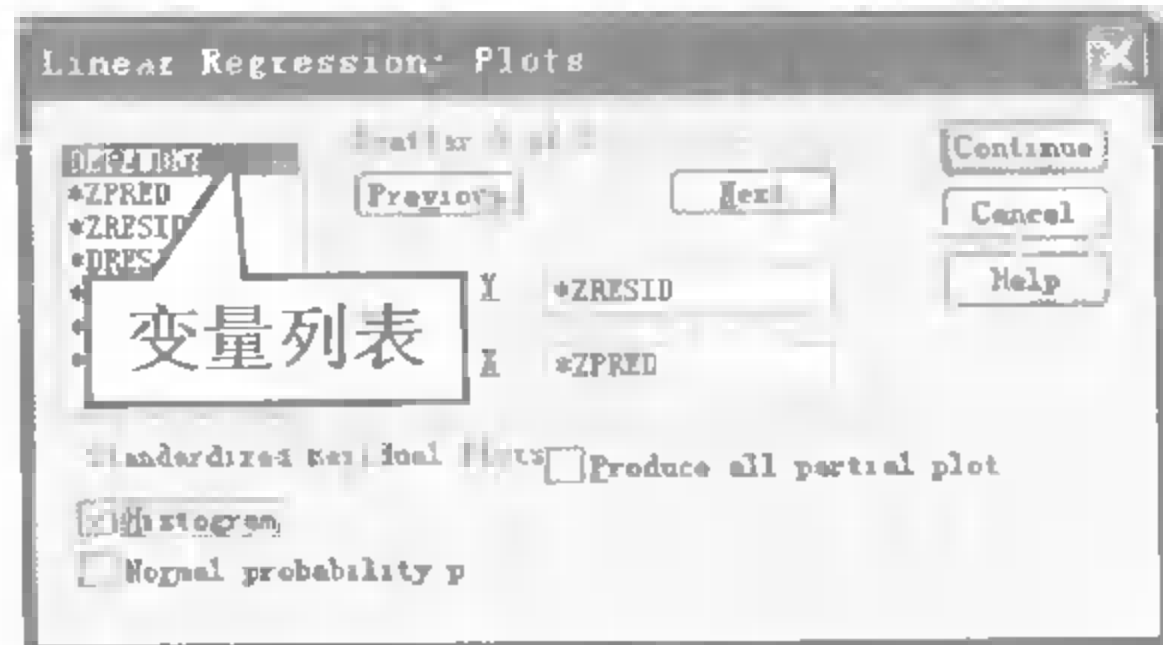


图 8-5 线性回归分析的作图设置子面板

(1) 可选的作图元素。左侧的变量列表给出了可以选择的作图元素, 包括 DEPENDENT (因变量)、\*ZPRED (标准化预测值)、\*ZRESID (标准化残差)、\*DRESID (剔除残差)、\*ADJPRED (修正后预测值)、\*SRESID (学生化残差) 和 \*SDRESID (学生化剔除残差)。

(2) Standardized Residual Plots 子设置栏。在此选择输出的标准化残差图的格式, 有如下 2 个可选项。

- Histogram 直方图, 输出关于标准化残差的直方图, 并带有标准正态曲线。
- Normal probability plot 正态概率图, 输出 P-P 图 (关于残差的正态概率图), 此图可用来检验残差的正态性。

(3) Produce all partial plots 复选框。输出关于每个自变量的偏残差图, 是由包括某个变量和不包括它的其它变量分别进行回归, 得到的 2 个残差所作的散点图。至少有两个自变量引入回归方程时, 才能产生偏残差图。

### 4. 保存设置

单击图 8-2 中的 Save 按钮, 弹出如图 8-6 所示的保存设置对话框, 在此设置关于回归分析过程的保存选项。依次勾选如下几个复选框: Cook's、Leverage values、Mean、Individual。单击 Continue 按钮返回主界面。

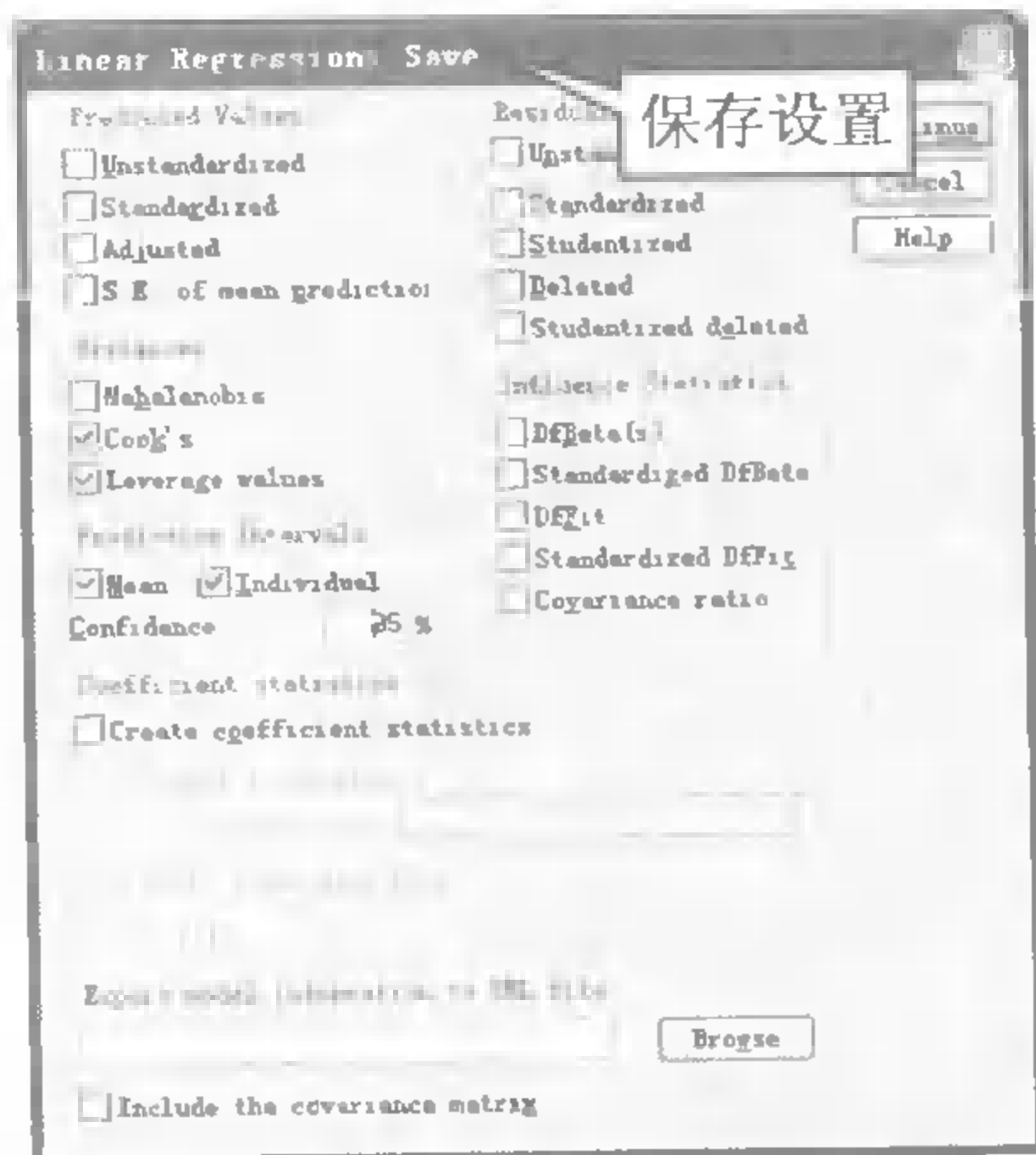


图 8-6 线性回归分析的保存设置

(1) Predicted Values 子设置栏。在此选择关于预测值的保存选项, 有如下 4 个可选项。

- Unstandardized 非标准化的预测值, 回归模型中对因变量的预测值。
- Standardized 标准化预测值, 将每个预测值转换成标准化形式, 用单个预测值减去所有预测值的平均值, 再除以所有预测值的标准差得到的标准化数值。
- Adjusted 调整预测值, 当某个观测记录没有参与对回归方程系数的估计时, 所得到的回归方程对这个观测

记录的预测值。

- S.E.of mean prediction 预测值的均值标准误。

(2) Distances 子设置栏。在此选择关于距离的保存选项，有如下 3 个可选项。

- Mahalanobis 选项，计算 Mahalanobis 距离，即自变量个案值与所有个案平均值的距离，当 Mahalanobis 距离过大时，表明该个案的一个或多个自变量的取值有异常。
- Cook's 选项，计算 Cook 距离，表示把一个个案从计算回归系数的样本中去除时，所引起的残差变化的大小。Cook 距离越大，表明该个案对回归系数的影响也越大。
- Leverage values 选项，计算杠杆值，用以测量单个观测对拟合效果的影响程度。杠杆值的取值范围是  $0 \sim n/(n-1)$ ，取 0 表示此观测对拟合无影响。

(3) Prediction Intervals 子设置栏。用于选择关于预测值置信区间的保存选项，包括 Mean（平均预测值）的上下置信限和 Individual（单个观测）的预测置信区间。勾选了 Mean 或 Individual 复选框后，激活 Confidence 输入框，在此指定 1~99.99 之间的任意数值作为上述两个预测区间的置信度，默认值为 95。

(4) Residuals 子设置栏。在此选择关于残差的保存选项，有如下 5 个可选项。

- Unstandardized 非标准化残差，观察值与模型预测值之差。
- Standardized 标准化残差，其均值为 0，标准差为 1。
- Studentized 学生化残差，用残差除以关于残差标准差的估计值，这个估计值取决于当前个案自变量的取值与自变量均值之间的距离。
- Deleted 剔除残差，表示把某个个案从计算回归系数的样本中去除时，回归后计算所得的关于当前个案的残差，即观测值与调整预测值的差。
- Studentized deleted 学生化剔除残差，用剔除残差除以单个个案的标准误，学生化残差和学生化剔除残差之间的不同，能反应被剔除观测在预测其自身时的作用大小。

(5) Influence Statistics 子设置栏。用于保存单个个案对回归分析影响程度的统计量，也就是把这个个案从回归样本中剔除后计算得到的一些统计量，包括如下 5 个可选项。

- DfBeta(s) (DFBeta 值)，剔除一个个案后回归系数的改变（包括常数项）。
- Standardized DfBeta(s) (标准化 DfBeta 值)，剔除一个个案后回归系数改变量标准后的取值（包括常数项）。当它大于  $2/\text{SQRT}(N)$  时，当前被剔除的个案可能是对回归系数有较大影响的点，这里  $N$  为观测的个案数目。
- DfFit (DfFit 值)，剔除一个个案后预测值的改变量。
- Standardized DfFit (标准化 DfFit 值)，剔除一个个案后预测值改变量标准后的取值。当它大于  $2/\text{SQRT}(p/N)$  时，当前被剔除的个案可能是对回归系数有较大影响的点，这里  $N$  为观测的个案数目， $p$  为当前模型中的参数个数。
- Covariance ratio (协方差矩阵的比率)，剔除一个个案后协方差矩阵的行列式与原协方差矩阵行列式的比值。它的取值接近 1，表明该个案对协方差矩阵的影响不大。

(6) Create coefficient statistics 复选框。设置将回归系数保存至指定数据集或文件的选项，勾选后激活如下两个可选项。

- Create a new dataset，建立一个新的数据集，在 Dataset name 后指定数据集名称。
- Write a new data file，将回归系数保存到新文件中。单击 Files 按钮指定保存路径。

(7) Export model information to XML file 子设置栏。设置将模型信息输出到 XML 格式文件的选项，保存结果可以直接用于 SmartScore 和 SPSS Server，单击 Browse 按钮指定文件名称

及路径。勾选 Include the covariance matrix 复选框, 表示保存协方差阵在如上的 XML 文件中。

## 5. 选项设置

在图 8-2 中, 单击 Options 按钮, 弹出如图 8-7 所示的选项设置对话框, 在此设置关于逐步回归的参数和缺失值的处理方式。单击 Continue 按钮返回主界面。

(1) Stepping Method Criteria 子设置栏。设置逐步回归方法的变量选取准则, 该准则将应用于 Stepwise (逐步回归法)、Backward (向后消去法) 和 Forward (向前选择法), 可选方式有如下 2 个。

- Use probability of F 单选项, 使用 F 的显著性取值 (Sig) 作为标准, 当 Sig 值小于 Entry 后指定的临界值时, 该变量将进入回归方程; 当 Sig 值大于 Removal 后指定的临界值时, 将其剔除。Entry 值必须小于 Removal 值, 且均为正数。若想更多的变量进入模型, 可增大 Entry 值; 反之, 若想在模型中剔除更多的变量, 可以降低 Removal 值。
- Use F value 单选项, 直接使用 F 统计量的取值作为判断依据, Entry、Removal 输入框分别用于指定引入和删除变量的临界 F 值。

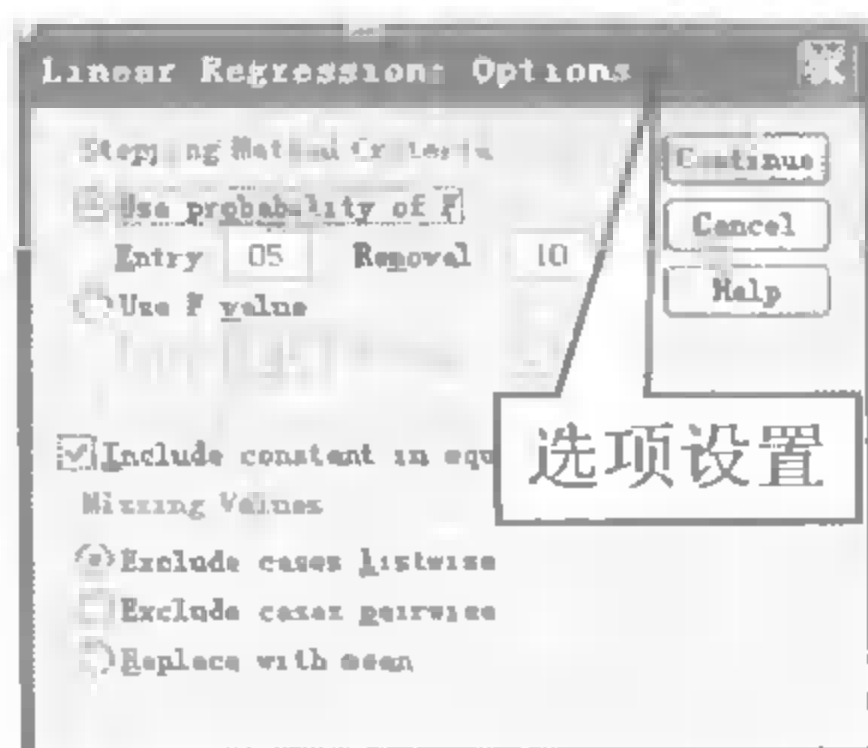


图 8-7 线性回归分析的 Options 设置

(2) Include constant in equation 复选框。勾选它表示在回归方程中包含常数项, 否则回归方程经过原点, 默认为选中状态。

(3) Missing Values 子设置栏。此栏设置关于缺失值的处理方式, 有如下 3 个可选项。

- Excludes cases listwise: 当一次选择多个变量进行分析时, 只要某个变量含有缺失值, 就在所有分析过程中将该记录删除。
- Excludes cases pairwise: 成对剔除带有缺失值的观测量, 只有计算过程中用到的某个变量有缺失值时, 才将相应的记录删除。比如计算两个变量的相关系数时, 只把这两个变量中带有缺失值的记录行剔除, 如果某个记录的这两个变量没有缺失值, 而其他变量中有, 那么此记录仍用于当前相关系数的计算。
- Replace with mean: 用该变量的均值代替其缺失值。

## 8.1.6 案例的结果分析

单击图 8-2 中的 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 8-8~图 8-15 所示。

输入 / 移去的变量 <sup>a</sup>			
模型	输入的变量	移去的变量	方法
1	价格		逐步 (准则 F-to-enter 的概率 ≤ .050, F-to-remove 的概率 ≥ .100)。
2	轴距		逐步 (准则 F-to-enter 的概率 ≤ .050, F-to-remove 的概率 ≥ .100)。

a. 因变量 销售量

模型摘要 <sup>c</sup>				
模型	R	R <sup>2</sup>	调整的 R <sup>2</sup>	估计的标准差
1	.657 <sup>a</sup>	.434 <sup>a</sup>	.422	1.01553
2	.655 <sup>b</sup>	.430 <sup>b</sup>	.422	1.01337

a. 预测变量 (常量) 价格。

b. 预测变量 (常量) 价格 轴距。

c. 因变量 销售量

图 8-8 逐步回归法模型摘要

ANOVA <sup>c</sup>						
模型		平方和	df	均方	F	显著性
1	回归	81.720	1	81.720	65.670	.000 <sup>a</sup>
	残差	186.662	150	1.244		
	合计	268.383	151			
2	回归	115.311	2	57.656	56.122	.000 <sup>b</sup>
	残差	153.072	149	1.027		
	合计	268.383	151			

a. 预测变量 (常量), 价格。

b. 预测变量 (常量), 价格, 轴距。

c. 因变量: 销售量。

图 8-9 ANOVA 方差分析表

系数 <sup>a</sup>							
模型	非标准化系数		标准化系数	t	显著性	共线性统计量	
	B	标准误差	Beta			容差	VIF
1	(常量)	4.684	.194		24.090	.000	
	价格	-.051	.006	-.552	-8.104	.000	1.000
2	(常量)	-1.822	1.151		-1.583	.116	
	价格	.055	.006	.590	9.487	.000	.988
	轴距	.061	.011	.356	5.718	.000	.988

● 因变量 销售量

图 8-10 回归系数的估计值表

已排除的变量 <sup>a</sup>								
模型 <sup>d</sup>		Beta In	t	显著性	偏相关	共线性统计量		
						容差	VIF	最小容差
1	车型	.251 <sup>a</sup>	3.854	.000	.301	.998	1.002	.998
	发动机规格	.342 <sup>a</sup>	4.128	.000	.320	.611	1.638	.611
	马力	.257 <sup>a</sup>	2.082	.041	.167	.293	3.417	.293
	轴距	.356 <sup>a</sup>	5.718	.000	.424	.988	1.012	.988
	宽度	.244 <sup>a</sup>	3.517	.001	.277	.892	1.121	.892
	长度	.308 <sup>a</sup>	4.790	.000	.365	.976	1.025	.976
	净重	.346 <sup>a</sup>	4.600	.000	.353	.722	1.385	.722
	燃料箱容量	.266 <sup>a</sup>	3.687	.000	.289	.820	1.219	.820
	燃料效率	-.198 <sup>a</sup>	-2.584	.011	-.207	.758	1.319	.758
2	车型	.129 <sup>b</sup>	1.928	.056	.157	.835	1.197	.827
	发动机规格	.145 <sup>b</sup>	1.576	.117	.128	.445	2.246	.445
	马力	.028 <sup>b</sup>	.229	.819	.019	.256	3.910	.256
	宽度	-.025 <sup>b</sup>	-.275	.784	-.023	.470	2.126	.470
	长度	.027 <sup>b</sup>	.237	.813	.020	.290	3.448	.290
	净重	.105 <sup>b</sup>	1.028	.306	.084	.365	2.741	.365
	燃料箱容量	.002 <sup>b</sup>	.024	.981	.002	.443	2.259	.443
	燃料效率	.014 <sup>b</sup>	.164	.870	.014	.559	1.790	.559

<sup>a</sup> 模型中的预测变量 (常量), 价格。
   
<sup>b</sup> 模型中的预测变量 (常量), 价格, 轴距。
   
<sup>c</sup> 因变量 销售量

图 8-11 已排除的变量统计信息表

系数相关 <sup>a</sup>				
模型		价格	轴距	
1	相关性	价格	1.000	
	协方差	价格	3.96E-005	
2	相关性	价格	1.000	-.108
		轴距	-.108	1.000
	协方差	价格	3.31E-005	-6.7E-006
		轴距	-6.7E-006	.000

<sup>a</sup> 因变量 销售量

图 8-12 系数相关矩阵

共线性诊断 <sup>a</sup>						
模型	维	特征值	条件索引	方差比例		
				(常量)	价格	轴距
1	1	1.885	1.000	.06	.96	
	2	.115	4.051	.94	.94	
2	1	2.847	1.000	.90	.02	.09
	2	.150	4.351	.01	.97	.91
	3	.003	33.403	.99	.00	.99

<sup>a</sup> 因变量 销售量

案例诊断 <sup>a</sup>					
案例数	价格	标准残差	销售量	预测值	残差
64	25.450	-4.905	-2.21	2.7638	-4.97111
109	18.145	-3.610	.11	3.7651	-3.65892

<sup>a</sup> 因变量 销售量

图 8-13 共线性诊断表和案例诊断表

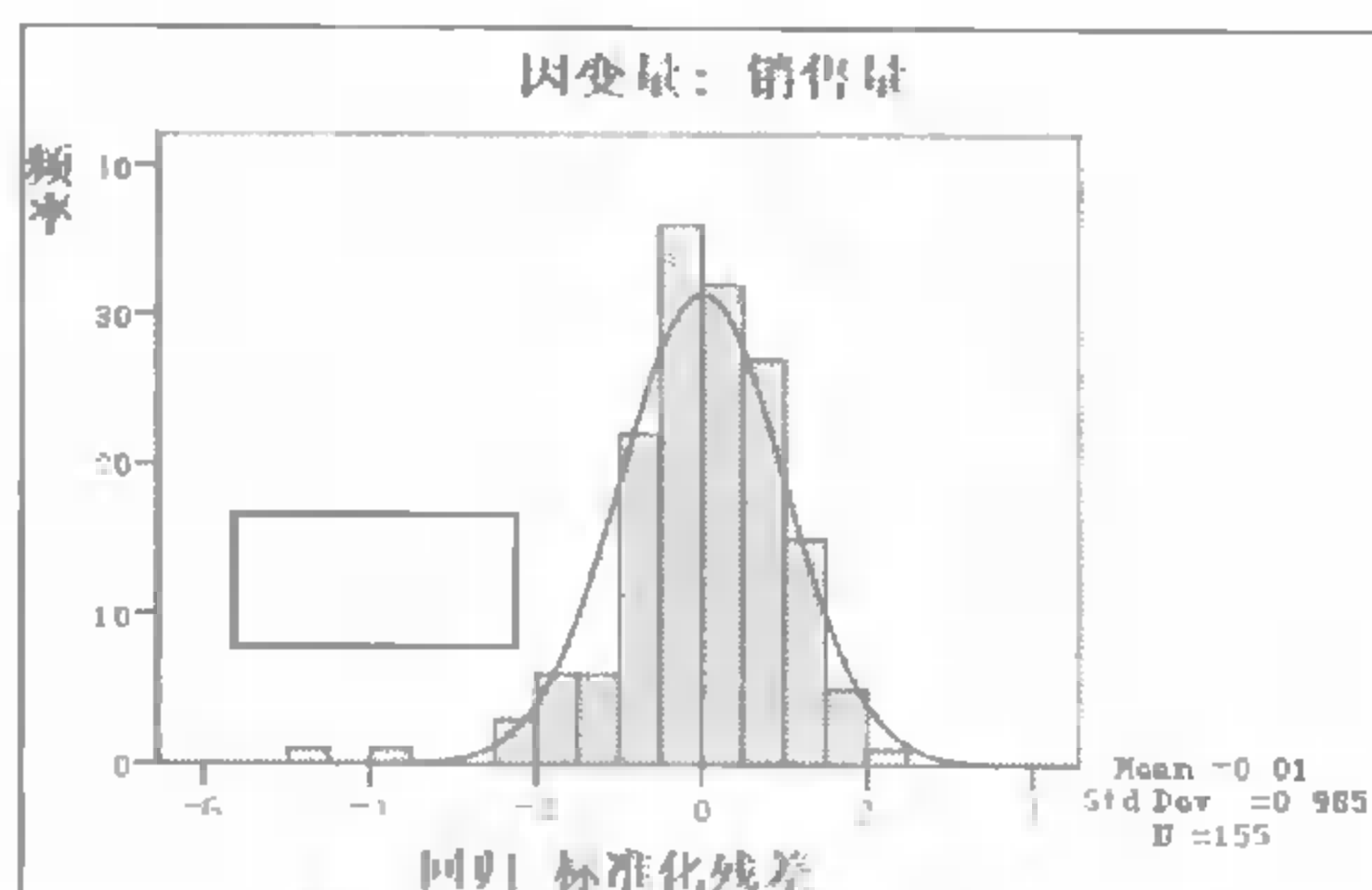


图 8-14 回归残差的直方图



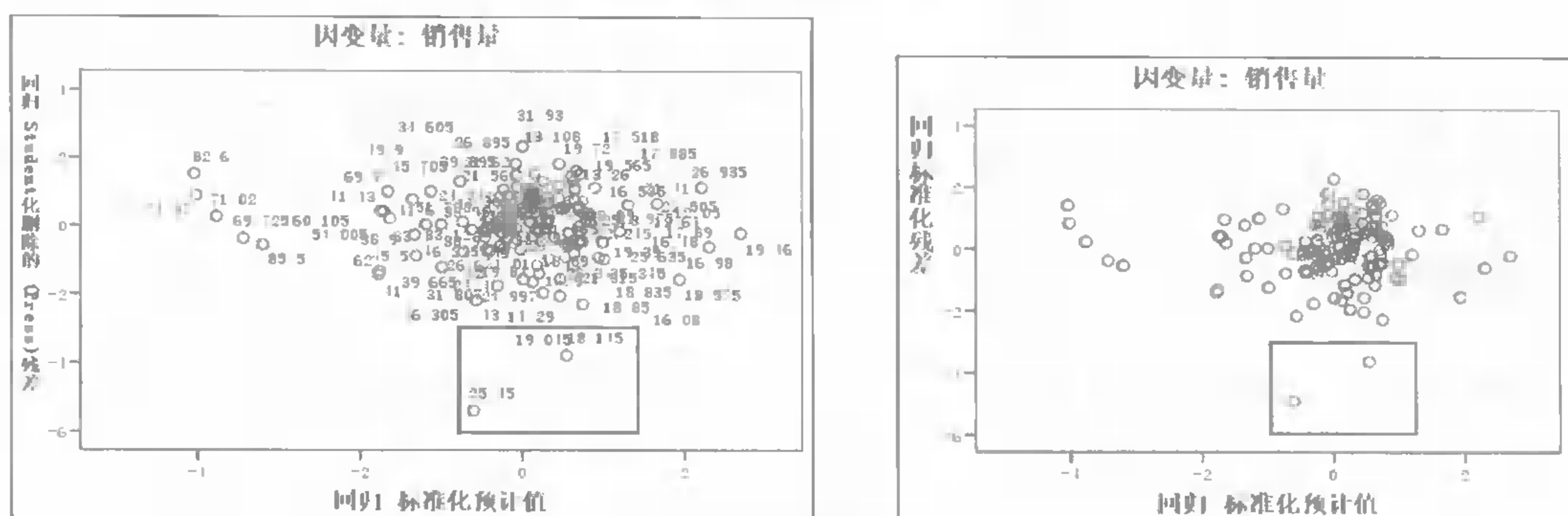


图 8-15 回归残差的散点图

(1) 变量筛选过程。如图 8-8 所示,“输入/移去的变量”表格给出了逐步回归过程里变量的引入和剔除过程及其准则。最先引入了变量价格,建立了模型 1;接着引入了变量轴距,建立了模型 2;没有变量剔除,所以模型 2 中包含两个变量:价格和轴距。

(2) 模型摘要信息。如图 8-8 所示,“模型摘要”表格给出了关于模型的拟合情况。从表中可看出,模型 2 的调整  $R^2$  为 0.430,大于模型 1 调整  $R^2$  值 0.304,说明模型可解释的变异占总变异的比例越来越大,引入方程的变量轴距是显著的。

(3) 方差分析表。如图 8-9 所示,ANOVA 表格给出了回归拟合过程中每一步的方差分析结果。

第一列中用蓝色线框标识的内容,是模型 2 的回归平方和与残差平方和,二者的大小较为接近,说明线性模型解释了总平方和的一半,拟合效果不太理想。棕色线框的标识表明,当回归方程包含不同的自变量时,其显著性概率值均远小于 0.01,所以可以显著地拒绝总体回归系数为 0 的原假设。注意:由此 ANOVA 方差分析表只能说明销售量与价格和轴距之间存在着线性关系,但不能直接说明这线性关系的强弱。

(4) 回归系数的估计。如图 8-10 所示,“系数”表格给出了所有模型的回归系数估计值,根据模型 2 建立的多元线性回归方程为销售量 =  $-1.822 - 0.055 \times \text{价格} + 0.061 \times \text{轴距}$ 。自变量价格的系数小于 0,说明随着价格的升高,销售量下降;轴距的系数大于零,说明轴距越大,销售量越高。经  $t$  检验,价格和轴距的显著性  $P$  值都远小于 0.01,因而均有显著性意义;但常数项的  $P$  值为  $0.116 > 0.10$ ,所以常数项不能通过显著性检验,改进方法是在如图 8-7 所示的 Options 设置界面,单击取消 Include constant in equation 复选框。

“系数”表的最右一列为共线性诊断统计量,两个自变量的膨胀因子(VIF)都为 1.102 (小于 5),所以模型 2 中的两自变量之间没有出现共线性。

(5) 已排除变量的统计信息表。图 8-11 给出了各个模型中已排除变量的统计信息。可见,模型 2 中各变量  $t$  检验的显著性概率  $P$  值全都大于 0.05,所以它们不能被引入模型。

(6) 系数的相关矩阵。图 8-12 给出了各模型中自变量之间的相关系数矩阵。从关于模型 2 的输出看,价格和轴距的相关系数为 -0.108,价格和轴距之间的协方差为  $-6.7E-0.05$ ,取值都非常小,所以价格和轴距之间可以认为是不相关的。

(7) 案例诊断信息。图 8-13 所示,“案例诊断”表格给出了 2 个异常值的编号(84 和 109)、标准化残差等统计量。因为它们的残差绝对值大于 3 倍的残差标准差(在图 8-4 中指定)。

(8) 关于残差的直方图。图 8-14 是标准化残差的直方图,同时绘制了正态分布曲线。可见残差基本符合正态分布,但是也存在着一个或两个非常大的负偏差,如蓝色线框标识。



(9) 关于残差的散点图。如图 8-15 所示，左侧为学生化删除残差对标准化预测值的散点图，学生化删除残差大多分布在-4 到 4 之间，但也存在个别的奇异点，如蓝色线框标识。右侧为标准化残差对标准化预测值的散点图，用蓝色线框标识的是异常点。

8.2 曲线回归

线性回归可以满足很多数据分析的需要，然而它不能对所有的问题都适用。有时，因变量与自变量是通过一个已知或未知的非线性函数关系相联系的，尽管有可能通过一些函数转换方法在一定范围内将它们的关系转变为线性关系，但这种转换有可能导致更为复杂的计算或数据失真。很多时候，用户研究的只有两个相关的变量，如果不能马上根据观测数据确定一种最佳模型，可以利用曲线估计在众多的回归模型中寻找一个简单而又比较适合的模式。

8.2.1 曲线回归的基本原理

如果充分了解数据的特点，可以选择与其相适应的函数模型。但是在大多数情况下，研究者对已有数据的认识往往是不完整的，并不能辨别变量之间的准确关系。这时，可以先将数据绘制成散点图，观察数据在图中的分布情况，再根据图形的特点来确定应采用的模型形式。如果数据在图中呈线性分布，可以选择线性模型；否则，可以同时引入多种非线性模型。由于有些函数的图形十分接近，即使对不同函数的图形特点有所了解，也可能在判断上产生误差，一个比较直接的方法是从  $R_2$  值的大小和图形本身进行比较，直至找到最佳模型。

SPSS 里的曲线回归要求自变量与因变量的类型都为数值型的连续变量。如果选择了时间作为自变量，曲线估计过程将自动生成一个时间变量，其在各观测记录之间的间隔是等长的，同时要求因变量也是时间序列数据。

SPSS 的曲线估计（Curve Estimation）模块能够自动拟合包括线性模型、对数曲线模型、二次曲线模型和指数曲线模型在内的十几种曲线模型。输出的统计量包括模型的回归系数、复相关系数、调整  $R$  方和方差分析表等。

由于曲线估计的内容比较复杂，所以经常通过变量替换的方法把不满足线性关系的数据转换为符合线性回归模型的数据，再利用线性回归进行估计。实际上，SPSS 的 Curve Estimation 过程正是按照指定的要求，先进行这样的变量转换，再通过线性回归的估计方法来估计未知参数的。下面给出几种常用的转换方法，如表 8-1 所示。

表 8-1 回归分析中的变量转换方法

曲 线	变 换	变换后的线性式
幂函数 $y=\alpha x^{\beta}$	$y'=\ln y, x'=\ln x$	$y'=\ln \alpha+\beta x'$
指数函数 $y=\alpha e^{\beta x}$	$y'=\ln y$	$y'=\ln \alpha+\beta x$
双曲函数 $y=\frac{x}{\alpha x+\beta}$	$y'=\frac{1}{y}, x'=\frac{1}{x}$	$y'=\alpha+\beta x'$
对数函数 $y=\alpha+\beta \ln x$	$x'=\ln x$	$y'=\alpha+\beta x$
指数函数 $y=\frac{\beta}{\alpha e^x}$	$y'=\ln y, x'=\frac{1}{x}$	$y'=\ln \alpha+\beta x'$
S 型曲线 $y=\frac{1}{\alpha+\beta e^{-x}}$	$y'=\frac{1}{y}, x'=e^{-x}$	$y'=\alpha+\beta x'$

8.2.2 问题描述和数据准备

某零售商收集了以往的销售数据和相应的广告支出，本节通过曲线回归来分析零售商的广告费用支出与产品销售量之间的关系。所用数据均摘自 SPSS 自带的 Demo 文件“advert.sav”，所用数据文件为“广告费用与销售数据.sav”，数据格式如图 8-16 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	advert	Numeric	8	2	广告费用	None	None	8	Right	Scale
2	sales	Numeric	8	2	销售量	None	None	8	Right	Scale
3	PRED_	Numeric	8	2	Predicted Val	None	None	10	Right	Scale
4	RESID	Numeric	8	2	Residuals	None	None	10	Right	Scale

图 8-16 广告费用与销售量的数据格式

使用 Curve Estimation 过程来比较线性模型和二次曲线模型对数据的拟合效果，并分析能促进销售量增加的广告费用的合理支出范围。

8.2.3 曲线回归分析的设置和操作

依次单击菜单“Analyze→Regression→Curve Estimation...”执行曲线回归分析的功能，其主设置界面如图 8-17 中所示，在此设置分析变量和曲线回归的函数形式。

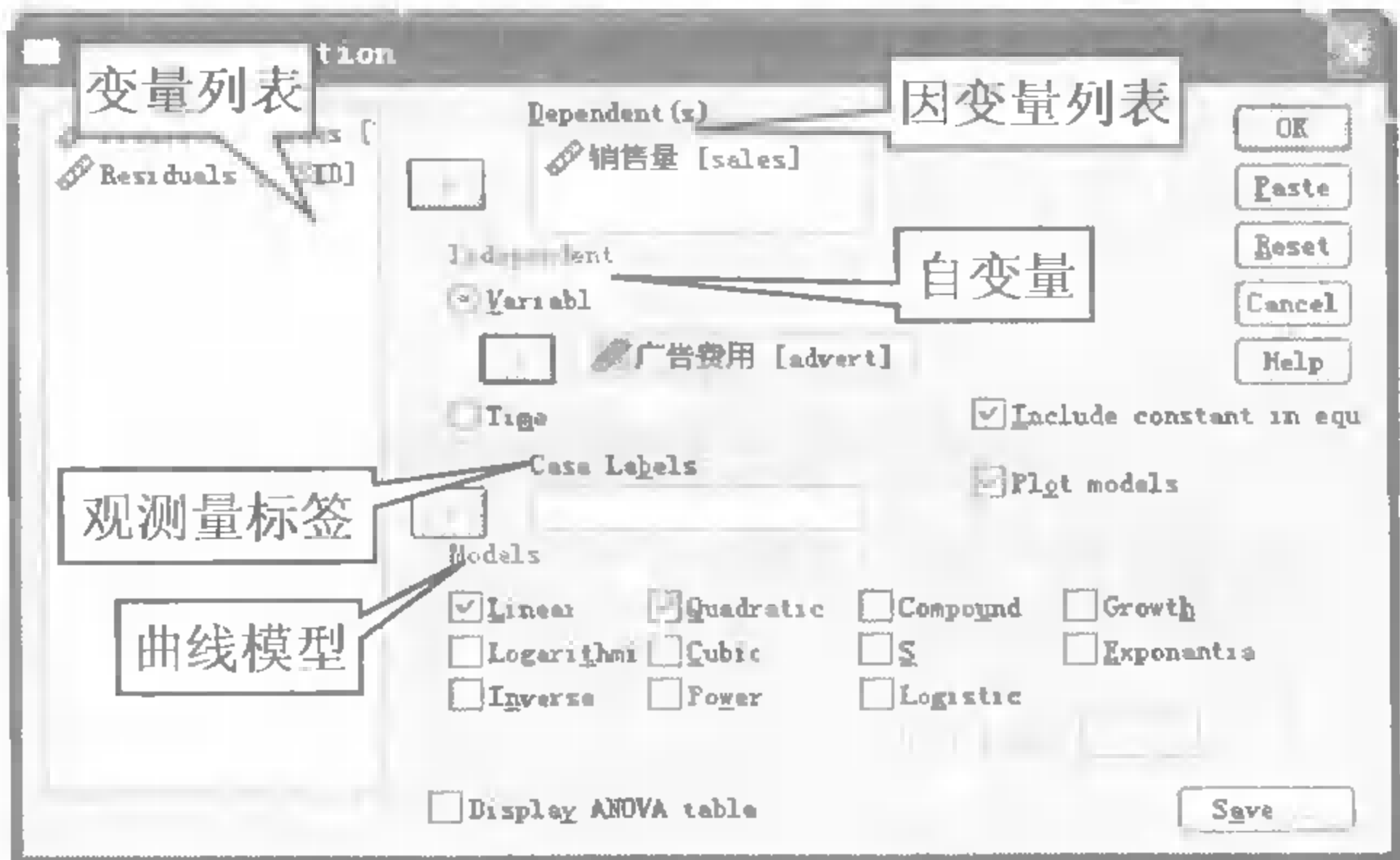


图 8-17 曲线回归分析的主设置界面

1. 变量及模型设置

在变量列表中单击选中销售量变量，单击从上至下第一个 按钮，将其作为因变量选入 Dependent(s)列表框；在变量列表中单击选中广告费用变量，单击从上至下第二个 按钮，将其作为自变量选入 Independent 栏；在 Models 栏中勾选如下 2 个复选框：Linear（线性模型）和 Quadratic（二次曲线模型）。

(1) 变量选择。

- Dependent(s)列表框，用于选入连续变量作为因变量，如果选入了多个，将分别对每个因变量进行模型拟合。
- Independent 栏，用于选入一个自变量，有如下两种选择方法。
  - ☆ Variable 单选项，把普通的自变量从变量列表中选入其下的选框。

☆ Time 单选项，直接使用时间序列自变量。

● Case Labels 栏，用于选入对观测的标签变量，可用于在散点图中标识观测记录。

(2) Include constant in equation 复选框，勾选它表示在回归方程中包含常数项。

(3) Plot models 复选框，做图选项，包括原始数据的散点图和拟合模型的曲线图。

(4) Model 子设置栏。在此选择曲线拟合的函数形式，可选项包括 Linear（线性模型）、Quadratic（二次曲线模型）、Compound（混合曲线模型）、Growth（生长曲线模型）、Logarithmic（对数曲线模型）、Cubic（三次曲线模型）、S（S 型曲线模型）、Exponential（指数曲线模型）、Inverse（逆曲线模型）、Power（幂函数曲线模型）和 Logistic（逻辑曲线模型）。若选择了 Logistic 曲线模型，激活 Upper bound 输入框，用于指定回归方程的上限值，它必须为正数并大于因变量的最大值。具体的函数形式如图 8-18 所示。

(1) Linear	$E(Y_t) = \beta_0 + \beta_1 t$
(2) Logarithmic	$E(Y_t) = \beta_0 + \beta_1 \ln(t)$
(3) Inverse	$E(Y_t) = \beta_0 + \beta_1 / t$
(4) Quadratic	$E(Y_t) = \beta_0 + \beta_1 t + \beta_2 t^2$
(5) Cubic	$E(Y_t) = \beta_0 + \beta_1 t + \beta_2 t^2 + \beta_3 t^3$
(6) Compound	$E(Y_t) = \beta_0 \beta_1^t$
(7) Power	$E(Y_t) = \beta_0 t^{\beta_1}$
(8) S	$E(Y_t) = \exp(\beta_0 + \beta_1 / t)$
(9) Growth	$E(Y_t) = \exp(\beta_0 + \beta_1 t)$
(10) Exponential	$E(Y_t) = \beta_0 e^{\beta_1 t}$
(11) Logistic	$E(Y_t) = (\frac{1}{u} + \beta_0 \beta_1^t)^{-1}$

图 8-18 曲线拟合函数表

(5) Display ANOVA table 复选框，勾选后将输出模型检验的方差分析表。

## 2. 保存设置

单击图 8-17 中的 Save 按钮，弹出如图 8-19 所示的保存设置对话框，用于选择曲线回归过程所要保存的对象。依次勾选 Predicted values 复选框和 Residuals 复选框；单击 Continue 按钮返回主界面。

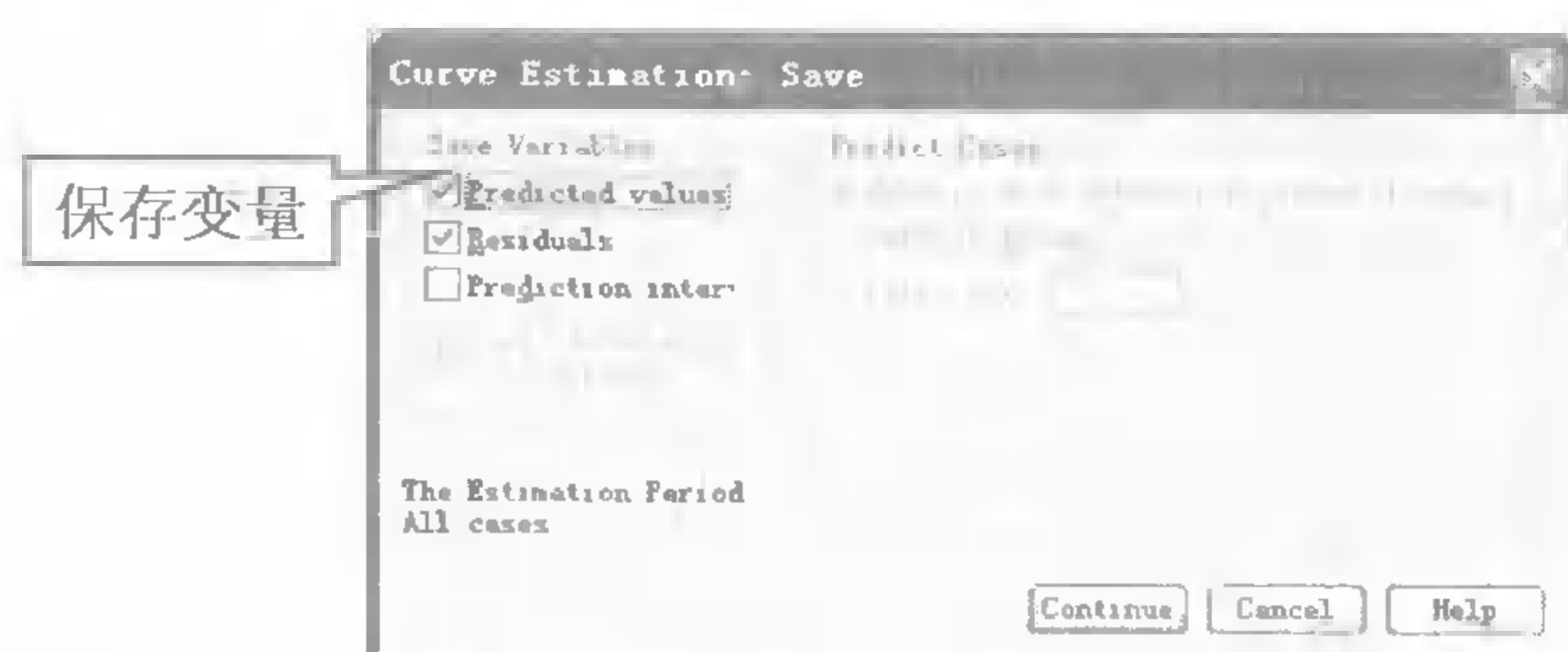


图 8-19 曲线回归的保存设置

① Save Variables 栏。选择需要保存的变量，可选项包括：Predicted values，因变量的预测值；Residuals，残差值；Prediction intervals，因变量的预测区间，需要在 Confidence interval 输入框指定预测区间的置信水平，可以选择 90%、95%或 99%，默认值为 95%。

② Predict cases 栏。用于设置关于预测值的保存选项。如果在图 8-17 中选择了 Time 时间变量作为自变量，在此可以设定超出当前数据时间序列范围的预测周期。

● Predict from estimation period through last case 单选项。保存从估计范围至最后一个观测记录的预测值，所谓估计范围，就是指估计模型参数所用到的样本范围。依次单击菜单“Data→Select Cases”执行样本选择功能，其主设置界面如图 8-20 所示。单击选中 Based on time or case range 后，单击其下的 Range 按钮，弹出如图 8-21 所示的指定范围子对话框；分别在 First、Last 输入框指定估计范围的起始时间和终止时

间（或者是起始观测的序号和终止观测的序号），单击 Continue 按钮返回图 8-20。在图 8-20 中，单击 OK 按钮返回数据编辑窗口，完成估计范围的设定。

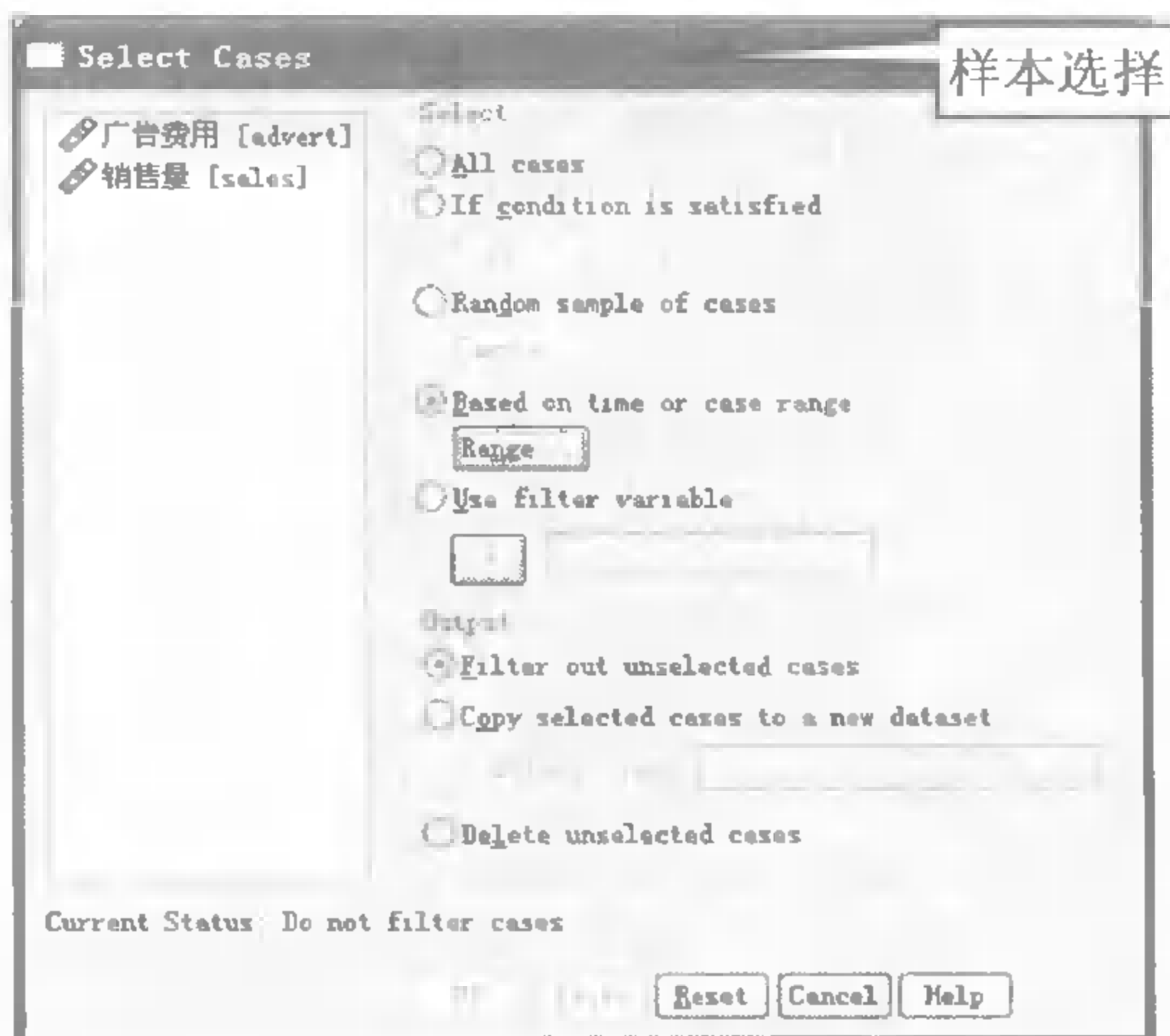


图 8-20 定义估计周期的条件设置



图 8-21 定义估计周期的范围设置

如图 8-19 所示，左下角会显示出当前的估计范围，“The Estimation Period: All cases”表示使用所有观测记录来估计模型参数。

- Predict through observation 单选项。直接指定预测的范围，如果要预测的时间超出了数据集中的时间范围，应该选择该项。Observation 输入框用于指定一个代表预测范围最末端的数值，它可以是日期、时间或观测序号，回归过程将计算和保存从估计范围到这个指定最末端的所有预测值。

#### 8.2.4 案例的结果分析

在图 8-17 中单击 OK 按钮，首先弹出如图 8-22 所示警告对话框，提示将在当前数据文件里增加 4 个变量，单击“确定”会运行并保存这 4 个变量，单击“取消”放弃保存。

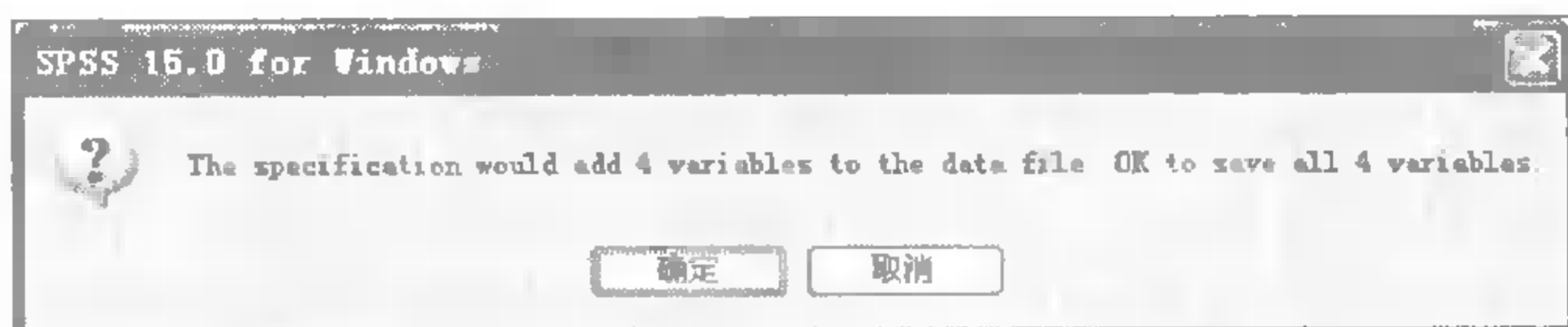


图 8-22 警告对话框

单击“确定”按钮，SPSS Viewer 窗口的输出结果如图 8-23～图 8-25 所示。

模型描述	
模型名称	MOD_1
因变量	销售量
方程	线性
	二次
自变量	广告费用
常数	包含
其值在图中标记为观测值的变量	未指定
用于在方程中输入项的容差	.0001

个案处理摘要	
	N
个案总数	24
已排除的个案 <sup>a</sup>	0
已预测的个案	0
新创建的个案	0

<sup>a</sup> 从分析中排除任何变量中带有缺失值的个案。

图 8-23 模型描述和个案处理摘要



模型汇总和参数估计值								
因变量 销售量								
方程	模型汇总					参数估计值		
	R方	F	df1	df2	Sig.	常数	b1	b2
线性	.839	114.548	1	22	.000	6.584	1.071	
二次	.908	104.113	2	21	.000	3.903	2.854	-.245

自变量为广告费用。

图 8-24 模型汇总和参数估计

(1) 模型描述和个案处理摘要。如图 8-23 所示,“模型描述”表汇总了关于两个模型的介绍信息;“个案处理摘要”表格给出了原始数据中的有效、缺失记录的统计情况,本例的 24 个记录全部用于分析。

(2) 模型检验信息。图 8-24 给出了关于模型成立的 F 检验结果和对模型参数的估计值。

线性模型为销售量=6.584+1.071×广告费用,  $b_1$  的系数 1.071>1, 说明随着广告费用的增加,销售量也会有明显增加,所以销售商应尽可能的增大广告费用,这样就可以收回投资并增加收入。但事实上,市场对广告是有一个饱和度的,当广告投入超过某一点后销售量就会减少,因此这个线性模型的经济意义并不值得借鉴。

二次模型为销售量=3.903+2.854×广告费用-0.245×广告费用<sup>2</sup>,  $b_2$  的系数-0.245<0, 说明当广告费用超过一定数量后,销售量就会转为下降了,并且可以求解得到这个转折点为  $2.854/(2 \times 0.245) = 5.824$ , 它就是商家要寻找的最优广告费用。由此可见,从经济意义出发,二次曲线模型比线性模型来得要好。

从两个模型的 F 检验结果看,它们的 Sig 值都远小于 0.01, 说明模型成立的统计学意义都非常显著。从  $R^2$  统计量看,二次曲线模型(0.908)略优于线性模型(0.839),由此也可以推断二次曲线模型的拟合效果要比线性模型稍好。

(3) 图形输出。图 8-25 所示是两个模型的拟合效果图,小圆圈代表原始观测记录,实线代表线性回归的预测值,虚线代表二次回归的预测值。直观上看,二次曲线模型能更好的拟合数据变化的趋势;线性模型过高地估计了广告费用取值较小和较大时的销售量,而对广告费用居中时销售量的估计又有所偏低。

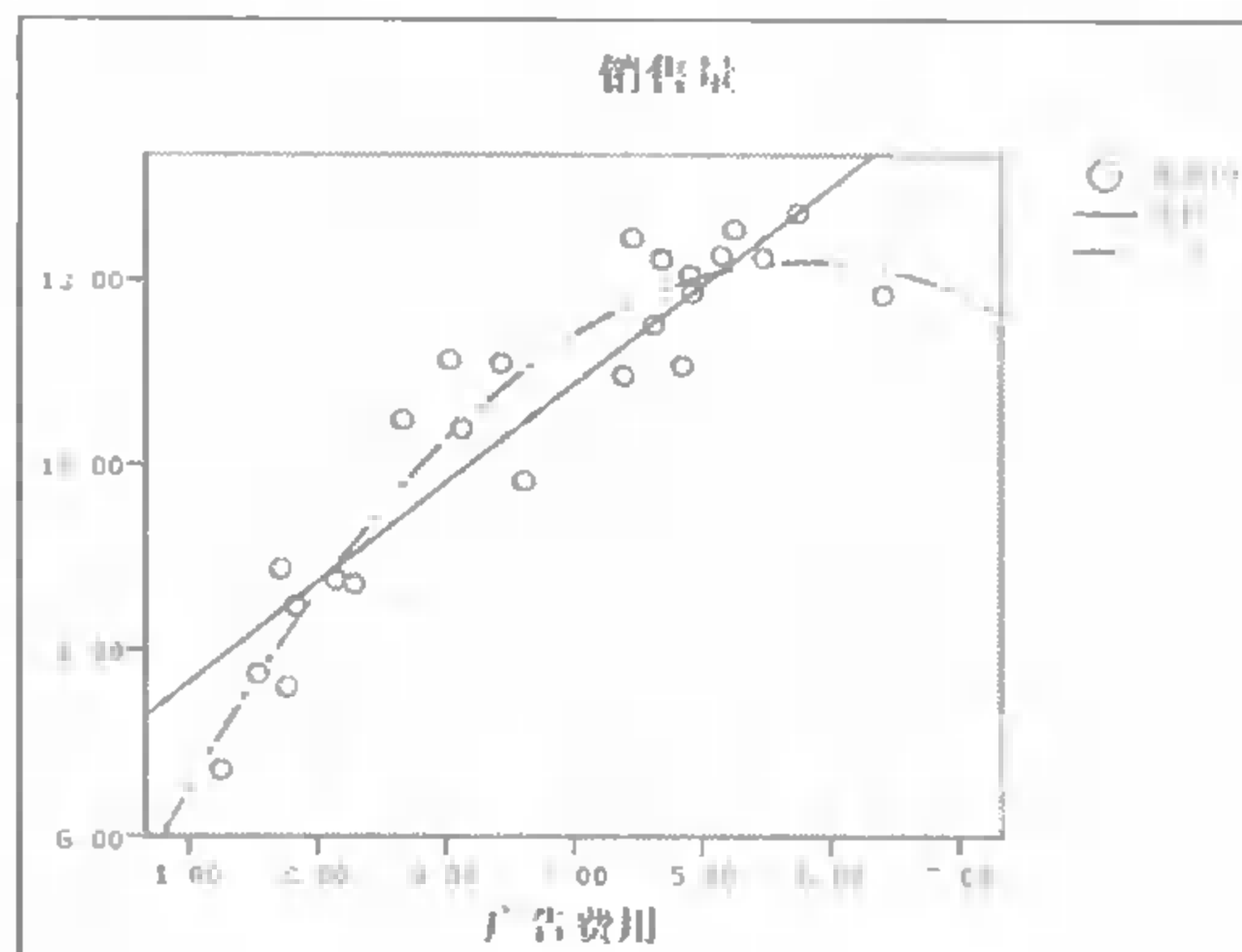


图 8-25 图形输出

### 8.3 非线性回归

非线性回归过程用来建立因变量与一组自变量之间的非线性关系,它不像线性模型那样有众多的假设条件,可以在自变量与因变量之间建立任意形式的模型。

如果已经了解待估方程中的参数取值范围,但是方程式不能写成简单的函数关系式时,建议使用非线性回归分析,因为非线性回归模型里的参数是用迭代算法估计的。例如:健康研究问题中,财政赤字对寿命的影响;社会科学研究中,人口增长与时间的关系;生物学与生理学研究中,有关动物骨骼成长与时间和营养的关系等,都是非线性关系。



8.3.1 非线性回归简介

对于看起来是非线性，但是能够通过变量转换化为线性的模型，称之为本质线性模型（例如曲线回归中所列的几个模型），对于此类模型，通常是用线性回归的方式处理转换后的模型。

有的非线性模型不能通过简单地变量转换化为线性模型，它们称之为本质非线性模型，例如  $y = b_0 + e^{b_1x_1} + \dots + e^{b_nx_n} + \varepsilon$ 。在非线性回归过程中，首先需要估算模型中参数的起始值及其取值范围，再利用迭代算法寻找使得残差平方和达到最小的参数估计值（NLR 算法）；另外还有 CNLR 算法，它的目标是最小化一个光滑的非线性损失函数。

表 8-2 所示是 SPSS 中一些经常使用的非线性模型，可以当作参考，但不能随意套用。在建立非线性模型时，恰当地确定参数初始值是相当重要的，即使准确地选择了函数关系，但设置的初始值不够理想，也有可能得不到一个较好的方程式，即使得到了一个不错的方程式，它所适用的数据范围也有可能是局部的，而不是全体的。

表 8-2 常用非线性模型的函数形式

名 称	模型表达式
Asymptotic Regression	$b_1 + b_2 \cdot \exp(b_3 \cdot x)$
Asymptotic Regression	$b_1 - (b_2 \cdot (b_3 \cdot x))$
Density	$(b_1 + b_2 \cdot x)^{-1/b_3}$
Gauss	$b_1 \cdot (1 - b_3 \cdot \exp(-b_2 \cdot x^2))$
Gompertz	$b_1 \cdot \exp(-b_2 \cdot \exp(-b_3 \cdot x))$
Johnson-Schumacher	$b_1 \cdot \exp(-b_2/(x + b_3))$
Lon-Modified	$(b_1 + b_3 \cdot x)^{b_2}$
Log-Logistic	$b_1 - \ln(1 + b_2 \cdot \exp(-b_3 \cdot x))$
Metcherlich Law of Diminishing	$b_1 + b_2 \cdot \exp(-b_3 \cdot x)$
Returns	
Michaelis Menten	$b_1 \cdot x/(x + b_2)$
Morgan-Mercer-Florin	$(b_1 \cdot b_2 + b_1 \cdot x^2 \cdot b_4)/(b_2 + x^2 \cdot b_4)$
Peal-Reed	$b_1/(1 + b_2 \cdot \exp(-(b_3 \cdot x + b_4 \cdot x^2 + b_5 \cdot x^3)))$
Ratio of Cubics	$(b_1 + b_2 \cdot x + b_3 \cdot x^2 + b_4 \cdot x^3)/(b_5 \cdot x^3)$
Ratio of Quadratics	$(b_1 + b_2 \cdot x + b_3 \cdot x^2)/(b_4 \cdot x^2)$
Richards	$b_1/((1 + b_3 \cdot \exp(-b_2 \cdot x))^{1/b_4})$
Verhulst	$b_1/(1 + b_3 \cdot \exp(-b_2 \cdot x))$
Von Bertalanffy	$(b_1 \cdot (1 - b_4) - b_2 \cdot \exp(-b_3 \cdot x))^{1/(1 - b_4)}$
Weibull	$b_1 - b_2 \cdot \exp(-b_3 \cdot x^{b_4})$
Yield Density	$(b_1 + b_2 \cdot x + b_3 \cdot x^2)^{-1}$

下面对 SPSS 的非线性回归处理过程的步骤和要点作以简单总结。

- （1）数据要求。因变量与自变量都要求为数值型的连续变量，对于分类变量，应该将其重新编码为数值型。
- （2）函数形式。根据经验或已有信息，确定一个本质非线性的模型。
- （3）损失函数。在非线性回归里，损失函数是用来最小化的目标函数，SPSS 默认将残差平方和作为损失函数，这类似于线性回归中最小二乘法的目标，SPSS 允许用户自定义损失函数。
- （4）参数的初始值及其取值范围。由于非线性回归采用迭代算法估计参数，故需要事先

指定参数的初始值和取值范围。合适的初始值能使迭代过程正常、迅速收敛，同时避免只得到局部最优解，常用方法有如下 4 个。

- ① 先通过图形确定参数的取值范围，然后在这个范围里选择初始值。
- ② 根据非线性方程的数学特性进行某些变换后，再通过图形帮助判断初始值的范围。
- ③ 先使用固定的数替代某些参数，以此来确定其他参数的取值范围。
- ④ 通过变量转换，使用线性回归模型来估计参数的初始值。

参数的取值范围指在迭代的过程中，将参数限制在有意义的范围区间，这可以防止得到违背常理或不可解释的结果，有如下两种对参数范围的约束方法。

- ① 线性约束，在约束表达式里只有对参数的线性运算。
- ② 非线性约束，在约束表达式里，至少有一个参数与其它参数进行了乘、除运算，或者自身的幂运算。

### 8.3.2 问题描述和数据准备

在第 8.2 节，使用 Curve Estimation 过程研究了零售商的广告费用支出与产品的销售量之间的关系，结果显示二次函数模型明显优于线性模型。二次模型拟合的结果说明提高广告费用有可能导致销售量的急速降低，所以有些零售商认为二次函数也许不是最合适的模型。

本节再来对此案例进行分析，其数据格式仍然如图 8-16 所示，目的是用 Nonlinear 过程拟合更合适的销售量随广告费用变化的模型。

#### 1. 对数据的初步分析

依次单击菜单“Graphs→Chart Builder...”打开图形构建器界面，如图 8-26 所示。在 Choose from 列表单击 Scatter/dot，然后双击右侧的简单散点图图标，激活上面的图形预览区；从变量列表把销售量变量拖动至图形预览区的 Y 轴位置，从变量列表把广告费用变量拖动至图形预览区的 X 轴位置。单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 8-27 所示。

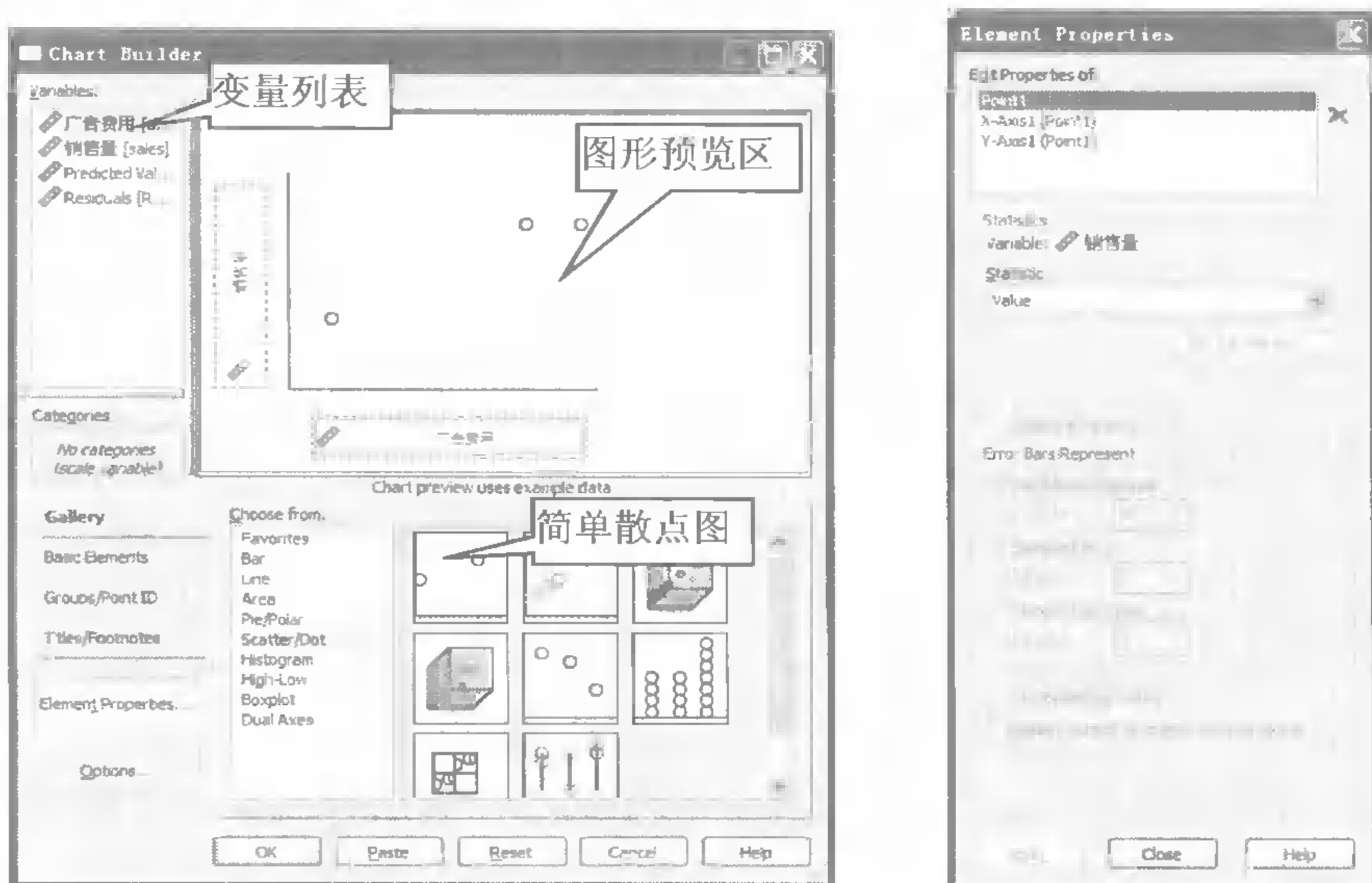


图 8-26 Chart Builder 作图设置

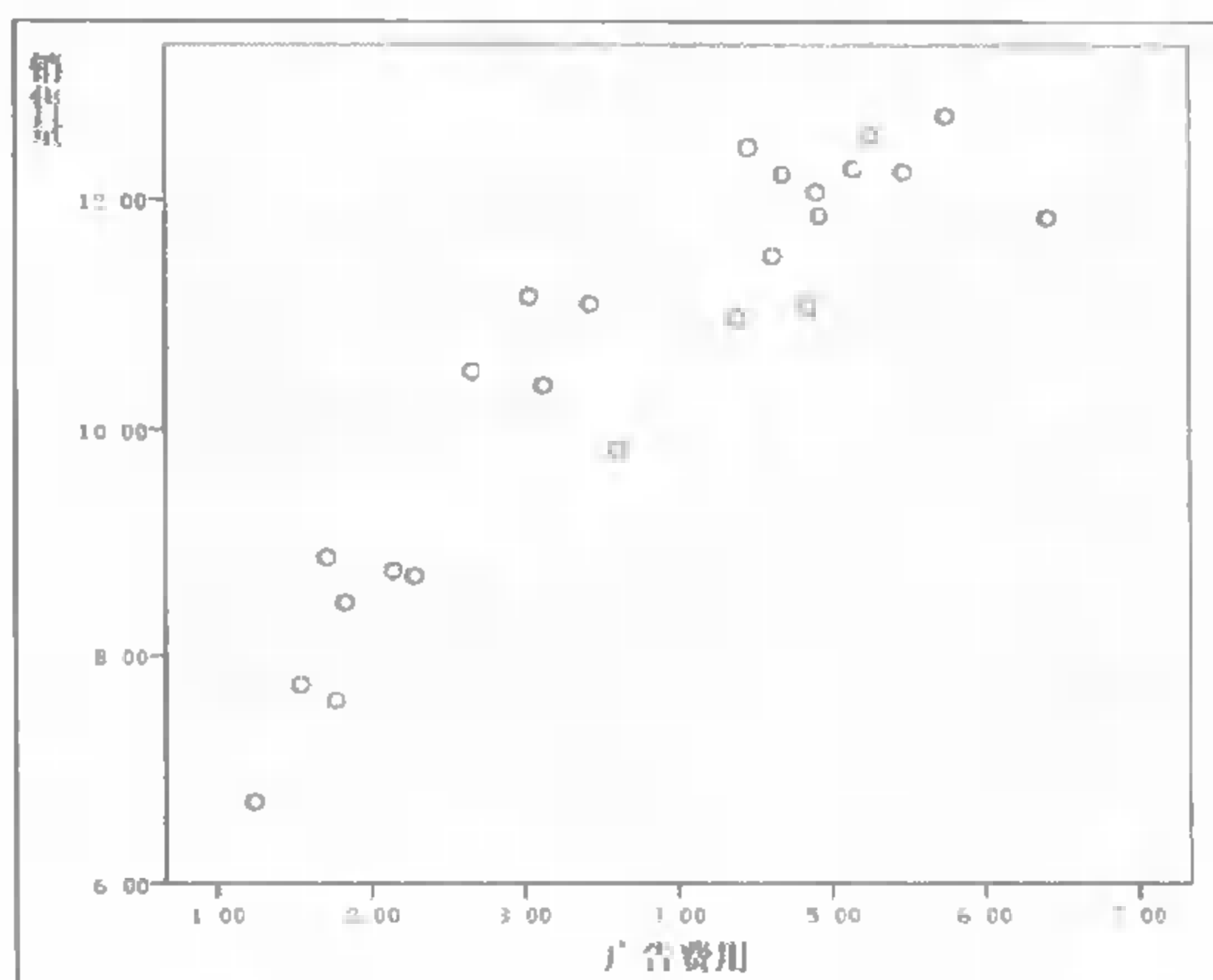


图 8-27 销售量对广告费用的散点图

图 8-27 所示是销售量对广告费用的散点图，通过观察，建议对此数据采用如下的非线性模型（称为 Misticerlich 模型）： $y = b_1 + b_2 e^{b_3 x}$ ,  $b_1 > 0, b_2 < 0, b_3 < 0$ ，此模型符合效益递减规律。最初，当  $x$  的值增加时， $y$  的值迅速增加；但是随着  $x$  的继续增加， $y$  的增速减慢，并最终趋近于  $b_1$  值的水平。

## 2. 参数初始值的选择

Nonlinear 过程需要设置待估参数的初始值，初始值的大小直接影响着模型的收敛性。参考图 8-27 所示的散点图，对初始值进行如下的计算和设置。

(1)  $b_1$  代表了销售量上升趋势的终点，观察散点图，发现销售量的最大值接近于 13，因此建议设定  $b_1$  的初始值为 13。

(2)  $b_2$  为当  $x=0$  时的  $y$  值与  $y$  的最大值（上限）之差，因此，可以用  $y$  的最小值减去  $b_1$  作为  $b_2$  的初始值，即  $b_2 = 7 - b_1 = 7 - 13 = -6$ 。

(3)  $b_3$  的初始值可以用图中两个分离点的斜率来表示。取两个点 ( $x=2, y=8$ ) 和 ( $x=5, y=12$ )，它们之间的斜率为  $(12-8)/(5-2)=1.33$ ，所以  $b_3$  的初始值可以设为 -1.33。

### 8.3.3 非线性回归的参数设置

依次单击菜单“Analyze→Regression→Nonlinear...”执行非线性回归分析的功能，其主设置界面如图 8-28 所示，在此设置分析变量和模型的函数形式。

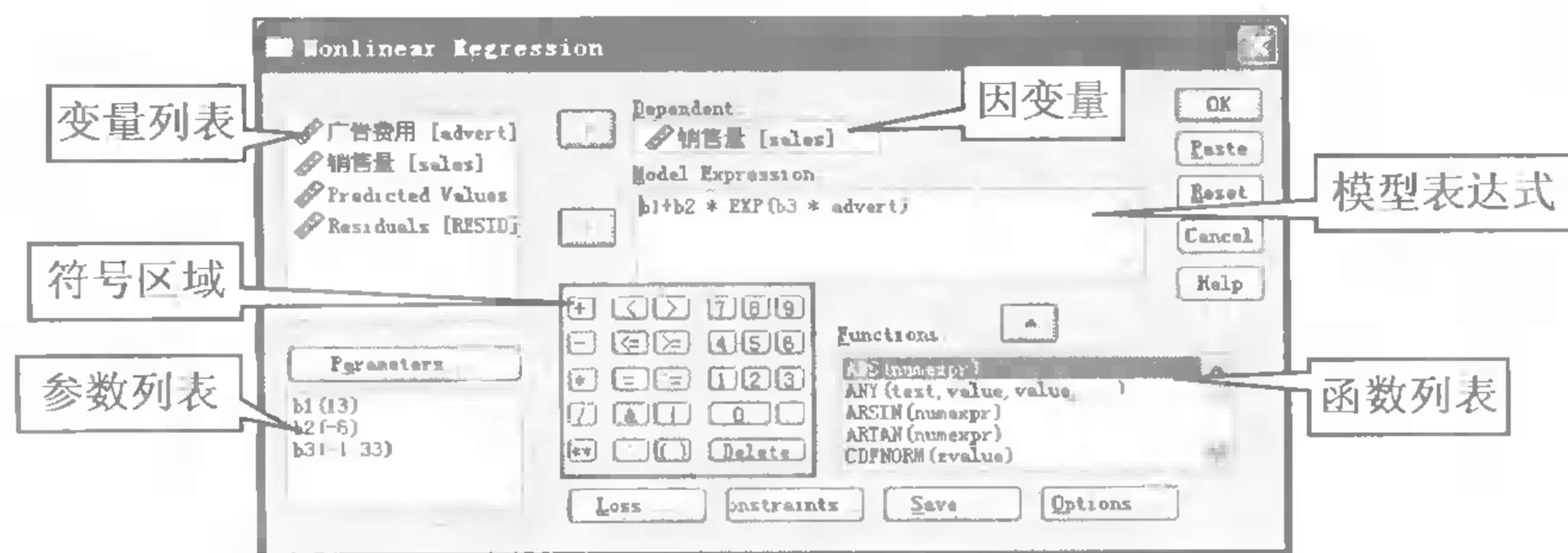





图 8-28 非线性回归分析的主设置界面

### 1. 变量选择及模型设置

在变量列表单击选中销售量变量，单击从上至下第一个  按钮，将其作为因变量选入 Dependent 选框；在模型表达式编辑框输入  $b1+b2*EXP(b3*advert)$ 。

① Dependent 选框，用于从变量列表选入一个数值型变量作为因变量。

② Model Expression 编辑框。

用于设置模型的函数表达式，可以直接输入和编辑函数表达式；还可以从变量列表中选入自变量的名称（单击旁边的  即可）；从符号区域选入数字或运算符（单击相应按钮即可）；从 Functions 列表选入指定的函数（单击  按钮或双击函数名即可）。

### 2. 模型参数的设置

在图 8-28 中，单击参数列表上方的 Parameters 按钮，弹出如图 8-29 所示的参数设置子对话框。在 Name 后输入“b1”，在 Starting 后输入“13”，单击 Add 按钮将其加入参数列表；用同样的方法添加参数“b2(-6)”和“b3(-1.33)”；单击 Continue 按钮返回主界面，此时 Parameters 按钮下的列表将会显示刚刚设定的 3 个参数。

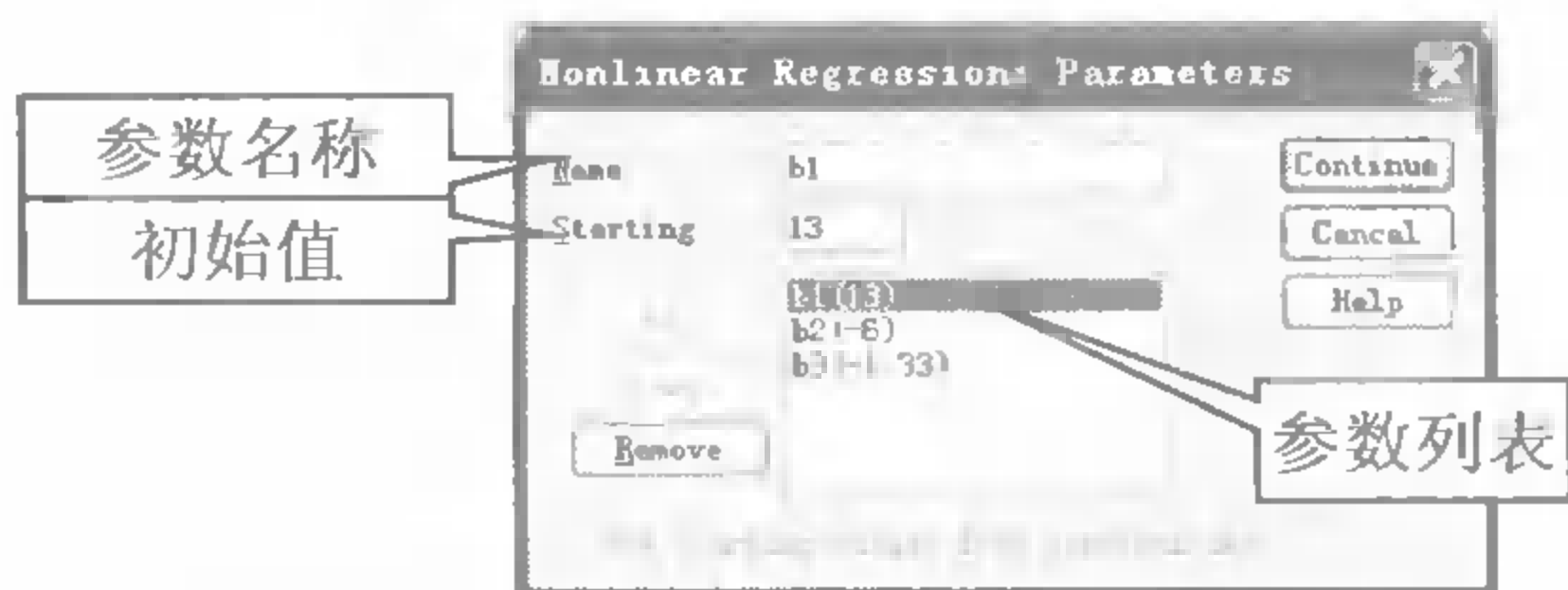


图 8-29 非线性回归分析的模型参数设置

- Name 输入框指定参数名称，而且要与主界面里设置的函数表达式中所使用的名称一致。
- Starting 输入框指定参数的初始值，应尽可能地接近最终结果的值，不适当的初始值可能导致收敛失败、局部收敛或者产生不可预料的结果。
- 输入一个参数名称及其初始值后，单击 Add 按钮将其添加到参数列表框。在列表里选中某个参数，可以单击 Change 或 Remove 按钮对其进行编辑或删除。
- Use starting values from previous analysis 复选框，若已经使用此对话框进行过非线性回归，可勾选此项以直接使用上次分析中所使用的初始值。

### 3. 损失函数的设置

在图 8-28 中，单击 Loss 按钮，弹出如图 8-30 所示的损失函数设置对话框，单击 Continue 按钮返回主界面。

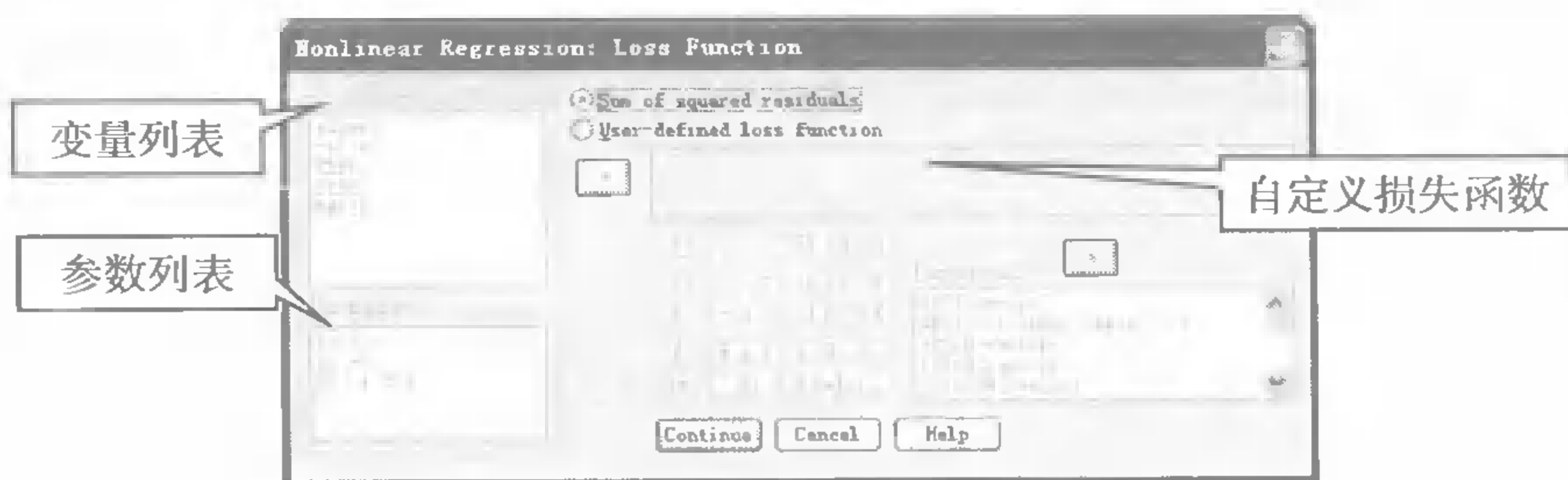


图 8-30 非线性回归分析的损失函数设置

在此，SPSS 给出了两种定义损失函数的方法。



(1) Sum of squared residuals 单选框, 指定使用残差平方和作为损失函数, 是默认选项。

(2) User-defined loss function 单选框, 由用户自行定义损失函数的表达式, 表达式的设置和编辑方式同图 8-28 中函数表达式的编辑相同。

#### 4. 参数约束设置

在图 8-28 中, 单击 Constraints 按钮, 弹出如图 8-31 所示的对话框, 用于设置估计参数的取值范围。

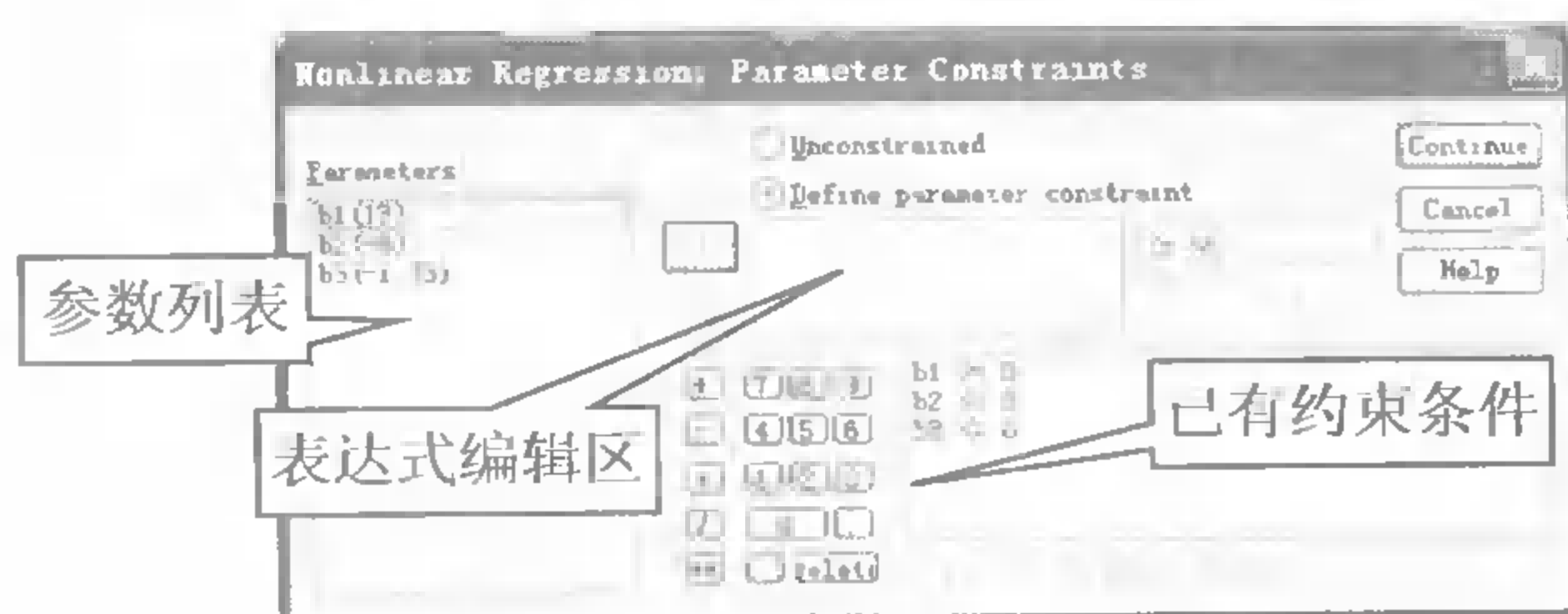



图 8-31 非线性回归分析的参数约束设置

单击选中 Define parameter constraint 单选项, 激活其他设置内容; 在参数列表选中 b1(13), 单击  按钮将其选入表达式编辑区, 再单击右侧的逻辑符号下拉列表选中 “>=”, 然后在右侧的输入框键入 “0”, 单击底部的 Add 按钮将 “b1>=0” 加入约束条件列表; 采用同样的方法加入约束条件 “b2<=0” 和 “b3<=0”。单击 Continue 按钮返回主界面。

在此, SPSS 给出了如下两种对参数取值范围的约束方式。

(1) Unconstrained 单选项, 对所有参数的取值范围都不作任何约束。

(2) Define parameter constraint 单选项。由用户自行定义对参数的约束条件。使用时, 先在表达式编辑区设置约束表达式, 方法同图 8-28 中的函数表达式的编辑相同; 然后在逻辑运算符下拉列表中选择逻辑条件, 可选项有大于等于 ( $\geq$ )、小于等于 ( $\leq$ ) 和等于 ( $=$ ); 接着在右侧的输入框指定逻辑运算的临界值。有了这 3 项, 就构成一个完整的约束条件, 单击 Add 按钮添加到下面的列表。在列表里选中某个约束条件后, 可以单击 Change 或 Remove 按钮对其进行编辑或删除。

#### 5. 估计方法的设置

在图 8-28 中单击 Options 按钮, 弹出如图 8-32 所示的选项设置对话框, 在此设置回归算法的相关参数。默认使用 Sequential quadratic programming 方法, 单击 Continue 按钮返回主界面。

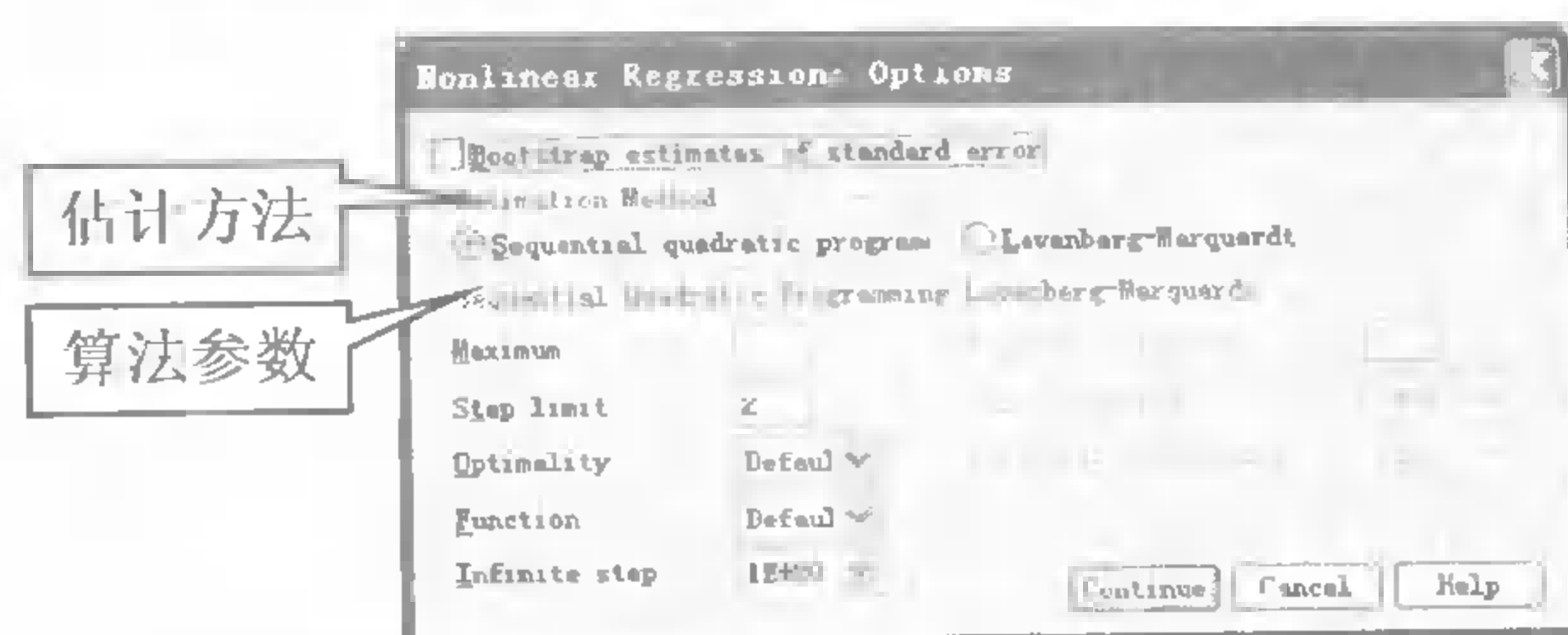


图 8-32 非线性回归分析的 Options 选项设置

(1) Bootstrap estimates of standard error 复选框。勾选它后将只能使用顺序二次规划法,



Bootstrap 通过从原始数据不断的抽样对标准误进行估计。它先使用有重复抽样，从原始数据集获得多个样本量相等的抽样样本然后对每个抽样样本进行非线性回归方程的估计；由此就得到关于每个参数的多个估计值，这些估计值的标准误就是对应参数的标准误的 Bootstrap 估计。用所有原始数据拟合的参数估计值将作为对每个抽样样本进行估计时的初始值。

(2) Estimation Method 子设置栏。在此，SPSS 给出了如下两种可选的参数估计方法。

#### ① Sequential quadratic programming (顺序二次规划法)

此方法适用于限制模型与非限制模型。如果要拟合的是限制模型或者是用户自行定义了损失函数的模型，在默认情况下都使用该选项。它利用双重迭代法进行求解，每一步迭代都建立一个二次规划算法，以此确定寻优的方向，把估计参数不断地代入损失函数求值，直到其满足指定的收敛条件。供设置的参数选项有如下 5 个。

- Maximum iterations 输入框，指定最大迭代次数。
- Step limit 输入框，指定步长限制，输入一个正数作为参数单步变化的最大允许量。
- Optimality tolerance 最优容限，可理解为目标函数的精确度（有效位数），如果最优容限为  $1E-6$ ，表示目标函数应保留 6 位有效数字，最优容限应大于函数精度。
- Function precision 函数精度，取值范围 0~1，函数值较大时作为相对精确度，函数值较小时作为绝对精确度，函数精度要小于最优容限。
- Infinite step size 单步变化限制，如果某步迭代中参数的变化大于此值，迭代终止。

#### ② Levenberg-Marquardt (简记为 L-M 方法)

此方法只适用于非限制模型，供设置的参数选项有如下 3 个。

- Maximum iterations 输入框，指定最大迭代次数。
- Sum-of-squares convergence，如果残差平方和的变化量小于此设置值，迭代终止。
- Parameter convergence，如果任何一个参数的变化量小于此设置值，迭代终止。

### 6. 保存选项的设置

在图 8-28 中，单击 Save 按钮，弹出如图 8-33 所示的保存设置对话框。勾选 Predicted values 复选框和 Residuals 复选框；单击 Continue 按钮返回主界面。

(1) Predicted values 复选框，保存因变量的预测值，变量名记为 Pred\_。

(2) Residuals 复选框，保存残差，变量名记为 Resid。

(3) Derivatives 复选框，保存每个参数的导数值，变量命名规则是用前缀“d.”接参数名的前 6 个字母。

(4) Loss function values 复选框，保存损失函数值，只有当用户指定了自定义的损失函数时，才会保存损失函数的计算值，其变量名记为 Loss\_。

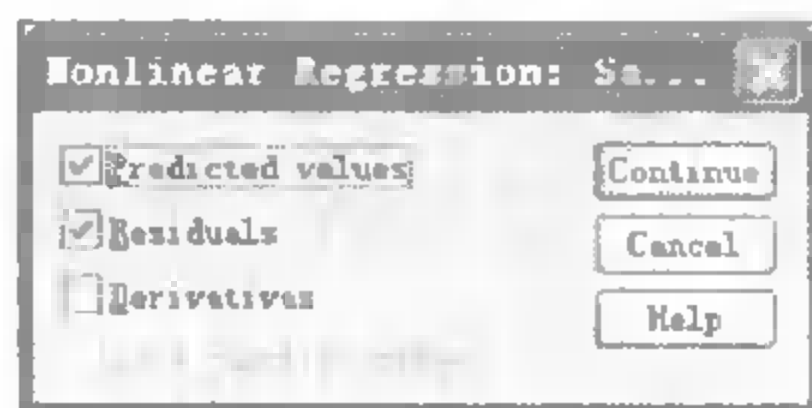


图 8-33 非线性回归分析的保存选项设置

### 7. 提示选择模型方法的对话框

如果在图 8-30 所示的损失函数子设置界面指定了自定义损失函数，或者在图 8-31 所示的参数约束子设置界面指定了某些约束条件，当单击它们的 Continue 按钮返回主界面时，会弹出如图 8-34 所示的警告对话框，提示用户将采用顺序二次规划法进行参数的估计，单击确定按钮接受建议，单击取消按钮返回子设置界面。

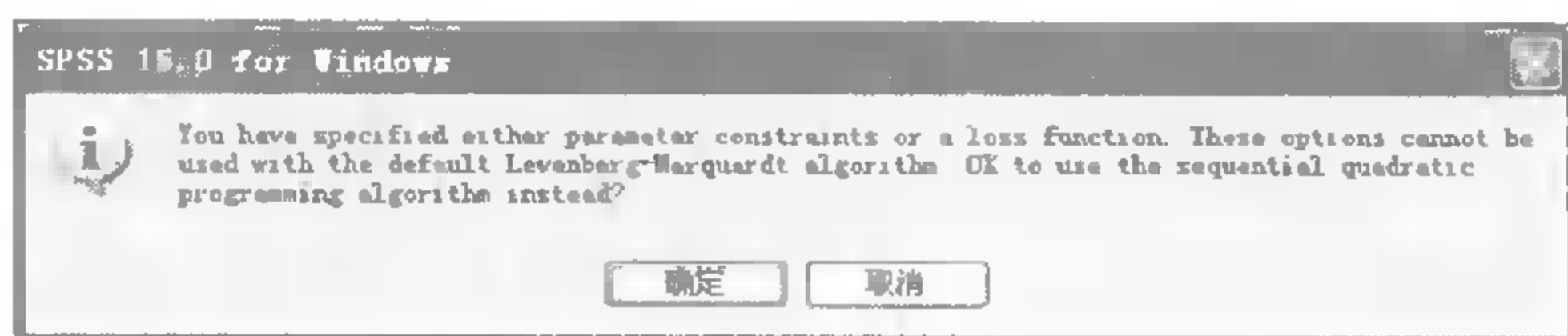


图 8-34 参数约束设置警告框

### 8.3.4 案例的结果分析

单击图 8-28 中的 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 8-35 所示。



图 8-35 Nonlinear 过程的输出结果

(1) 迭代历史记录。“迭代历史记录”表格给出了模型的迭代过程，共迭代了 15 次后得到最优解，最后一行是最终模型的参数估计值。“参数估计值”表格对结果参数有更详细的说明，由此得到的非线性模型为  $y = b_1 + b_2 e^{b_3 x} = 12.904 - 11.268 e^{-0.496x}$ 。

在此模型中， $b_1$  表示广告费用趋于无限大时产品的最大销售额，从“参数估计值”表中得到  $b_1$  的标准误很小 (0.610)，因此判断  $b_1$  的估计值是满可信的； $b_2$  为销售量的最大值与没有广告投入时的销售量的差，反映了广告投入可能带来的最大效益，它的标准误较大 (1.581)，其置信区间也很宽泛，所以  $b_2$  的可建议性不强； $b_3$  控制着  $y$  达到最大值的速度，其标准误也很小 (0.138)，因此  $b_3$  的估计值也较为可信。

(2) 方差分析表。“ANOVA”表格给出了关于方差分析的结果，从  $R$  方大于 0.9 看，拟合模型能够解释因变量大于 90% 的变异，说明模型的拟合效果还是不错的。

## 8.4 二元 Logistic 回归

二元 Logistic 回归是指因变量为二分类变量时的回归分析。在医学研究中经常会遇到二元变量的情况，比如：分析死亡与否的概率与病人生理状况、疾病严重程度之间的关系；研究对某种疾病易感性的概率与个体性别、年龄、免疫水平之间的关系等。对这类问题建立回

归模型时,目标概率的取值在 0~1 之间,但是回归方程的因变量取值却落在实数集当中,这是不能接受的。因此,可以先将目标概率做 Logit 变换,这样它的取值区间就变成了整个实数集,再作回归分析就不会有问题了,采用了这种处理方法的回归分析就是 Logistic 回归。

Logistic 模型有很多形式,除了二元 Logistic 模型外,还有配对 Logistic 模型、多元 Logistic 模型和随机效应的 Logistic 模型等,我们将在下一节介绍多元 Logistic 模型。

注意: Logistic 回归分析与另一个概念 Logistic 曲线模型(即 S 或倒 S 形曲线)是不一样的。如果用户需要拟合的是 Logistic 曲线模型,应该调用第 8.2 节介绍的 Curve Estimation 过程,那里提供了 11 种曲线模型,其中就有 Logistic 曲线模型。

### 8.4.1 二元 Logistic 回归的数学原理

#### 1. Logistic 回归模型

设因变量为  $y$ , 其取值 1 表示事件发生, 取值 0 表示事件未发生; 影响  $y$  的  $m$  个自变量分别记为  $x_1, x_2, \dots, x_m$ 。

记事件发生的条件概率为  $P(y=1|x_i)=p_i$ , 可以得到如下的 Logistic 回归模型

$$p_i = \frac{1}{1 + e^{-(\alpha + \sum_{i=1}^m \beta_i x_i)}} = \frac{e^{\alpha + \sum_{i=1}^m \beta_i x_i}}{1 + e^{\alpha + \sum_{i=1}^m \beta_i x_i}}, \quad 1 - p_i = 1 - \frac{e^{\alpha + \sum_{i=1}^m \beta_i x_i}}{1 + e^{\alpha + \sum_{i=1}^m \beta_i x_i}} = \frac{1}{1 + e^{\alpha + \sum_{i=1}^m \beta_i x_i}},$$

其中  $P_i$  代表在第  $i$  个观测中事件发生的概率,  $1 - p_i$  代表在第  $i$  个观测中事件不发生的概率, 它们都是由自变量  $x_i$  构成的非线性函数。

事件发生与不发生的概率之比  $p_i/(1 - p_i)$  被称为事件的发生比, 简记为 Odds。Odds 一定为正值(因为  $0 < p_i < 1$ ), 并且没有上界, 对 Odds 做对数变换, 就能够得到 Logistic 回归模型的线性模式  $\ln(\frac{p_i}{1 - p_i}) = \alpha + \sum_{i=1}^m \beta_i x_i$ 。

#### 2. Logistic 回归模型的参数估计

对 Logistic 回归模型的参数估计可以采用最大似然法或者迭代法。

最大似然估计的基本思想是先建立似然函数(或对数似然函数), 然后求使得似然函数达到最大的参数估计值。对于已有样本, 可建立样本似然函数为  $L = \prod_{i=1}^n P_i^{Y_i} (1 - P_i)^{1 - Y_i}$ , 于是样

本的对数似然函数为  $\ln L = \sum_{i=1}^n [Y_i \ln P_i + (1 - Y_i) \ln(1 - P_i)]$ 。根据最大似然原理, 应求使(对数)似然函数达到最大值的参数值, 对  $\ln L$  求一阶导数并令其为 0, 再用 Newton-Raphson 迭代方法求解方程组, 即可得出参数的最大似然估计值及其标准误。

#### 3. Logistic 回归模型的假设检验

常用的检验方法有似然比检验(likelihood ratio test)和 Wald 检验。

(1) 似然比检验。似然比检验的基本思想是比较在两种不同假设条件下, 对数似然函数值的差别大小。检验的零假设为两种条件下的对数似然函数值无显著差别, 检验的具体步骤如下。

- 先拟合不包含待检验因素的 Logistic 模型，求对数似然函数值  $\ln L_0$ 。
- 再拟合包含待检验因素的 Logistic 模型，求新的对数似然函数值  $\ln L_1$ 。
- 最后，比较两个对数似然函数值的差异，若两个模型分别包含  $l$  个自变量和  $P$  个自变量，记似然比统计量  $G$  的计算公式为  $G = 2(\ln L_p - \ln L_l)$ 。在零假设成立的条件下，当样本含量  $n$  较大时， $G$  统计量近似服从自由度为  $v = p - l$  的  $\chi^2$  分布，如果只是对一个回归系数（或一个自变量）进行检验，则  $v = 1$ 。

(2) Wald 检验。用  $u$  检验或  $\chi^2$  检验，推断各参数  $\beta_j$  是否为 0，其中  $u = b_j / S_{b_j}$ ， $\chi^2 = (b_j / S_{b_j})^2$ ， $S_{b_j}$  为回归系数的标准误。

#### 4. 逐步回归中的变量筛选

Logistic 逐步回归的变量筛选过程与线性逐步回归过程的变量筛选极为相似，但其中所用的检验统计量不再是  $F$  统计量，而是似然比统计量和 Wald 统计量等。

例如：使用似然比统计量  $G = 2(\ln L_1^{(l)} - \ln L_0^{(l)})$  作为变量筛选标准，在进行到第  $l$  步时，通过比较含有  $X_j$  和不含  $X_j$  的模型，决定  $X_j$  是否引入模型。

### 8.4.2 问题描述和数据准备

#### 1. 数据描述

本节来研究关于银行客户的贷款拖欠问题。通过分析银行掌握的一些客户资料和交易信息，推断指定客户的预期信誉。所用数据来源于 SPSS 自带的 Demo 文件 “bankloan.sav”，所用数据文件为 “银行贷款数据.sav”，数据格式如图 8-36 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	age	Numeric	4	0	年龄	None	None	4	Right	Scale
2	ed	Numeric	4	0	教育水平	{1, Did not	None	4	Right	Ordinal
3	employ	Numeric	4	0	工龄	None	None	6	Right	Scale
4	address	Numeric	4	0	居住年限	None	None	7	Right	Scale
5	income	Numeric	8	2	家庭收入	None	None	8	Right	Scale
6	debtinc	Numeric	8	2	贷款收入比	None	None	8	Right	Scale
7	creddebt	Numeric	8	2	信用卡欠款	None	None	8	Right	Scale
8	othdebt	Numeric	8	2	其他债务	None	None	8	Right	Scale
9	default	Numeric	4	0	是否拖欠	{0, No}...	None	7	Right	Nominal
10	validete	Numeric	8	2	有效数据	None	None	10	Right	Scale

图 8-36 银行贷款数据

本例数据集中的前 700 个案例是先前申请过贷款的用户，我们将利用其中的一个随机样本拟合一个二元逻辑回归模型，然后用拟合的模型对后 150 名预期用户进行信誉分类。因变量为是否拖欠 (default)，取值为 0 (值标签 No) 时，表示没有拖欠贷款，取值为 1 (值标签 Yes) 时，表示有拖欠贷款。

#### 2. 抽取分析样本

(1) 指定随机种子。依次单击菜单 “Transform → Random Number Generators...” 打开生成随机数的设置界面，如图 8-37 所示，在此设置随机抽样的随机种子。勾选

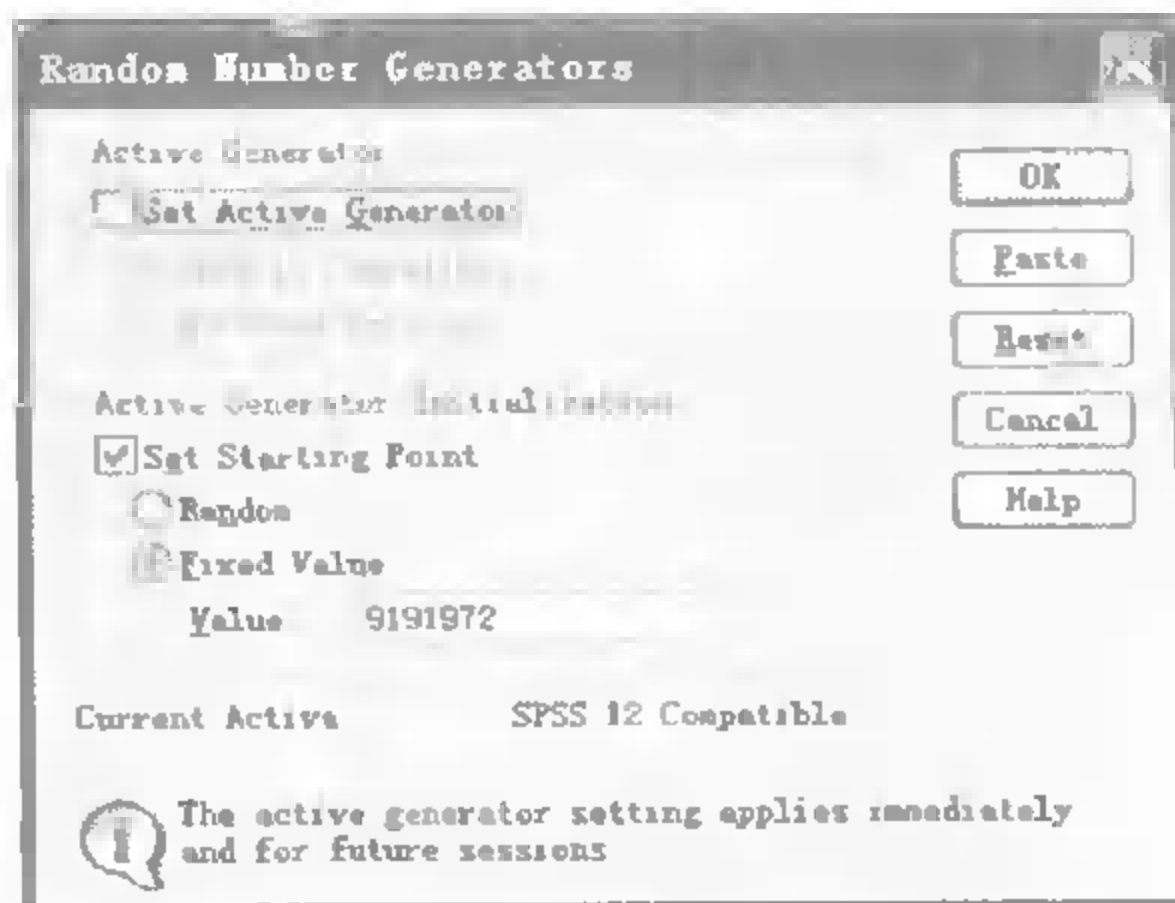


图 8-37 产生随机样本对话框



Set Starting Point 复选框，并单击选中 Fixed Value 单选项，在 value 后输入“9191972”。单击 OK 按钮，返回 Data Editor 窗口。

以后，只要采用与此相同的设置，就可以多次生成相同的随机样本。

(2) 计算筛选变量。依次单击菜单“Transform→Compute Variable...”打开计算新变量的主设置界面，如图 8-38 所示。在 Target Variable 栏输入变量名“validate”，在 Numeric Expression 编辑框中输入表达式 `rv.bernoulli(0.7)`，表示筛选变量 validate 的取值服从参数为 0.7 的 bernoulli 分布。单击左下角的 if 按钮，弹出如图 8-39 所示的条件设置子界面，单击选中 Include if case satisfies condition 单选框，并在下面的公式编辑区输入条件表达式 `MISSING(default)=0`，单击 Continue 按钮返回主界面。

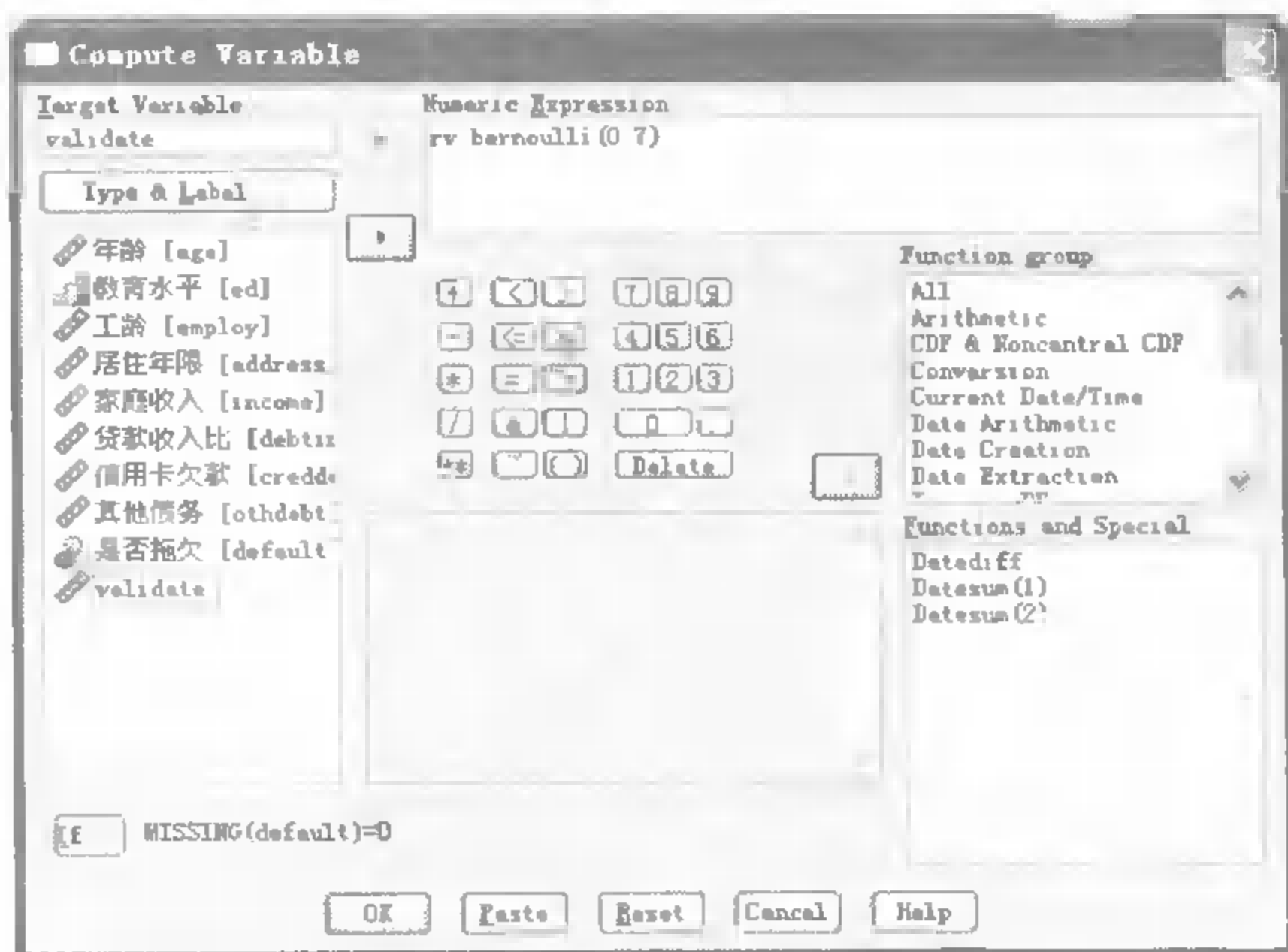


图 8-38 计算新变量的主界面

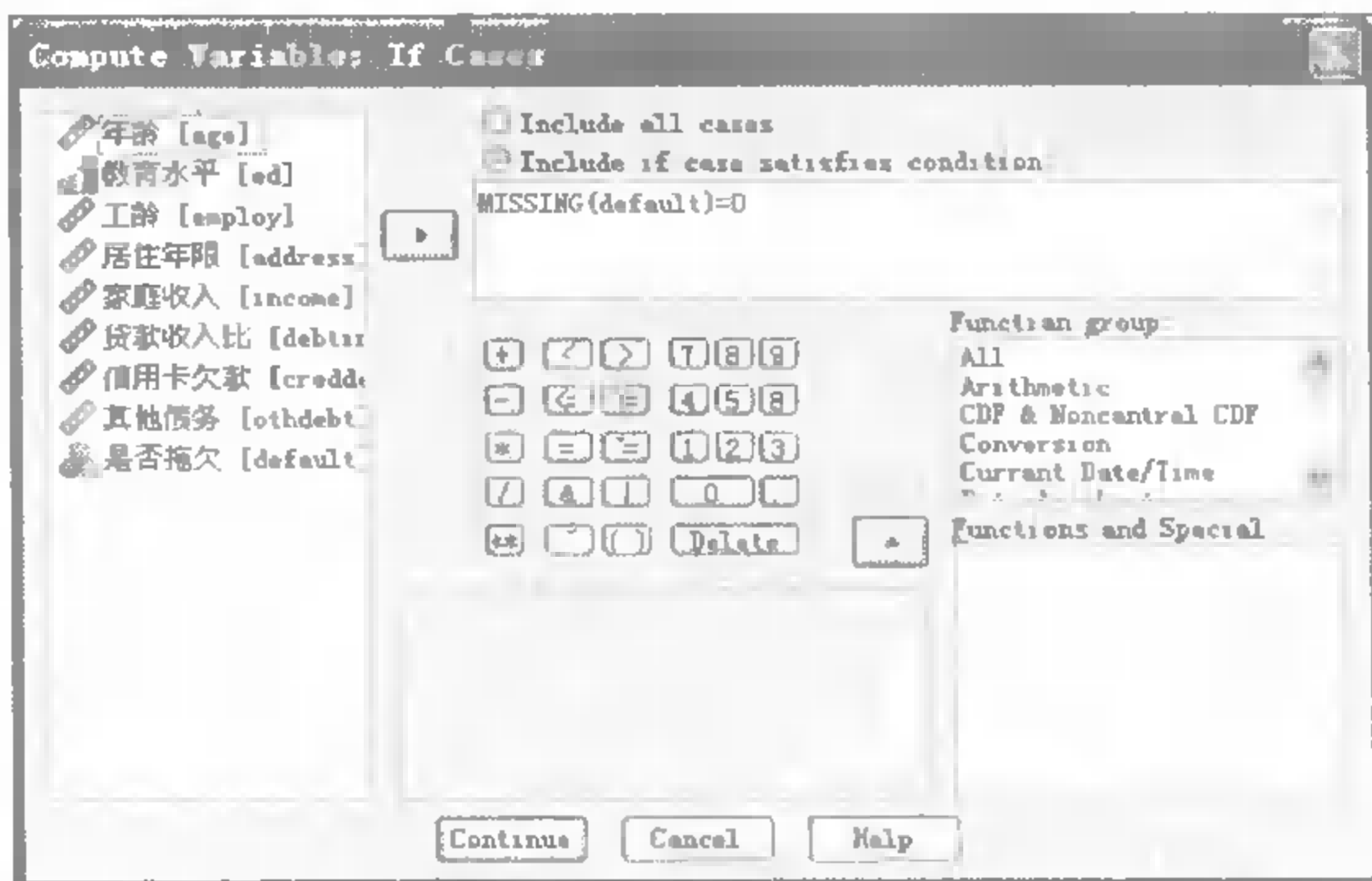


图 8-39 计算新变量的条件设置子界面

在图 8-38 中，单击 OK 按钮运行，当前数据集中增加名为 validate 的变量，它将用于随后的分析之中。对于 700 个历史客户（default 非缺失），validate 以指定的 bernoulli 分布随机取值为 1 或 0；对那些 default（是否拖欠）变量取缺失值的记录，validate 也取缺失值；随后的模型建立过程将只采用那些 validate 取值为 1 的记录，其他记录将应用于验证或预测。

### 8.4.3 二元 Logistic 回归的参数设置

依次单击菜单“Analyze→Regression→Binary Logistic...”执行二元逻辑回归分析过程，主设置界面如图 8-40 所示，在此选择分析变量、筛选变量和回归模型所采用的算法。



图 8-40 Binary Logistic 过程的主设置界面



## 1. 变量设置





在变量列表单击选中是否拖欠变量，单击从上至下第一个  按钮，将其作为因变量选入 Dependent 选框；在变量列表选中从年龄至其他债务的 8 个变量，单击从上至下第二个  按钮，将其作为协变量选入 Covariates 列表框；单击 Method 后的下拉列表，选中 Forward:LR 选项。在变量列表单击选中 validate 变量，单击从上至下第三个  按钮，将其作为筛选变量选入 Selection Variable 选框；单击后面的 Rule 按钮，弹出如图 8-41 所示的条件设置对话框，单击 validate 下的下拉列表中选中“equals”选项，在 Value 下拉框中输入“1”，单击 Continue 按钮返回主界面。



图 8-41 筛选条件的设置对话框

下面详细介绍各设置选项的含义。

(1) Dependent 栏。用于从变量列表选入一个二分变量作为因变量，可以是数值型变量或短字符型变量。

(2) Covariates 列表框。用于从变量列表选入协变量。除了可以选入单个变量，还可以选入变量之间的交互项，方法是在变量列表同时选中多个变量后，单击  按钮，这些选中变量的所有交互作用就被选入 Covariates 列表了。

(3) Blocks 1 of 1 子设置栏。通过单击 Previous 和 Next 两个按钮，可以添加和编辑多个不同的变量组，并为它们指定不同的变量选择方法，设置方式与图 8-2 所示的线性回归设置界面相同。

(4) 指定筛选变量。Selection Variable 栏用于选入一个对样本的筛选变量，只有满足指定条件的观测记录才会进入回归分析过程。选入变量后单击 Rules 按钮，弹出如图 8-41 所示的设置对话框。

Define Selection Rule 下面显示当前选入的筛选变量名（如“validate”）；下拉列表提供了可选的逻辑条件，包括 equals（等于）、not equal to（不等于）、less than（小于）、less than or equal to（小于或等于）、greater than（大于）和 greater than or equal to（大于或等于）；Value 下面的输入框指定逻辑条件需要满足的临界值。设置好后就形成一个完整的筛选条件，比如：“validate equals 1”就表示只选择那些满足“validate=1”的观测记录进行分析。


(5) Method 下拉菜单。用于指定协变量进入回归模型的方法，SPSS 给出了如下 7 种可选方案。

- ① Enter 强迫进入法，协变量全部进入模型。
- ② Forward:Conditional 向前逐步法（条件似然比），变量引入的根据是得分统计量的显著性水平；变量被剔除的依据是条件参数估计所得的似然比统计量的概率值。
- ③ Forward:LR 向前逐步法（似然比），变量引入的根据是得分统计量的显著性水平；变量被剔除的依据是最大偏似然估计所得的似然比统计量的概率值。
- ④ Forward:Wald 向前逐步法（Wald 法），变量引入的根据是得分统计量的显著性水平；变量被剔除的依据是 Wald 统计量的概率值。
- ⑤ Backward:Conditional 向后逐步法（条件似然比），将变量剔除出模型的依据是条件参数估计所得的似然比统计量的概率值。
- ⑥ Backward:LR 向后逐步法（似然比），将变量剔除出模型的依据是最大偏似然估计所

得的似然比统计量的概率值。

- Backward:Wald 向后逐步法 (Wald 法), 将变量剔除出模型的依据是 Wald 统计量的概率值。

## 2. 对分类变量的设置

单击图 8-40 中的 Categorical 按钮, 弹出如图 8-42 所示的对话框, 在此设置对分类变量的处理方式。在变量列表中单击选中教育水平 (ed) 变量, 单击  按钮, 将其作为分类变量选入 Categorical 列表框, 单击 Continue 按钮返回主界面。

下面详细介绍各设置选项的含义。

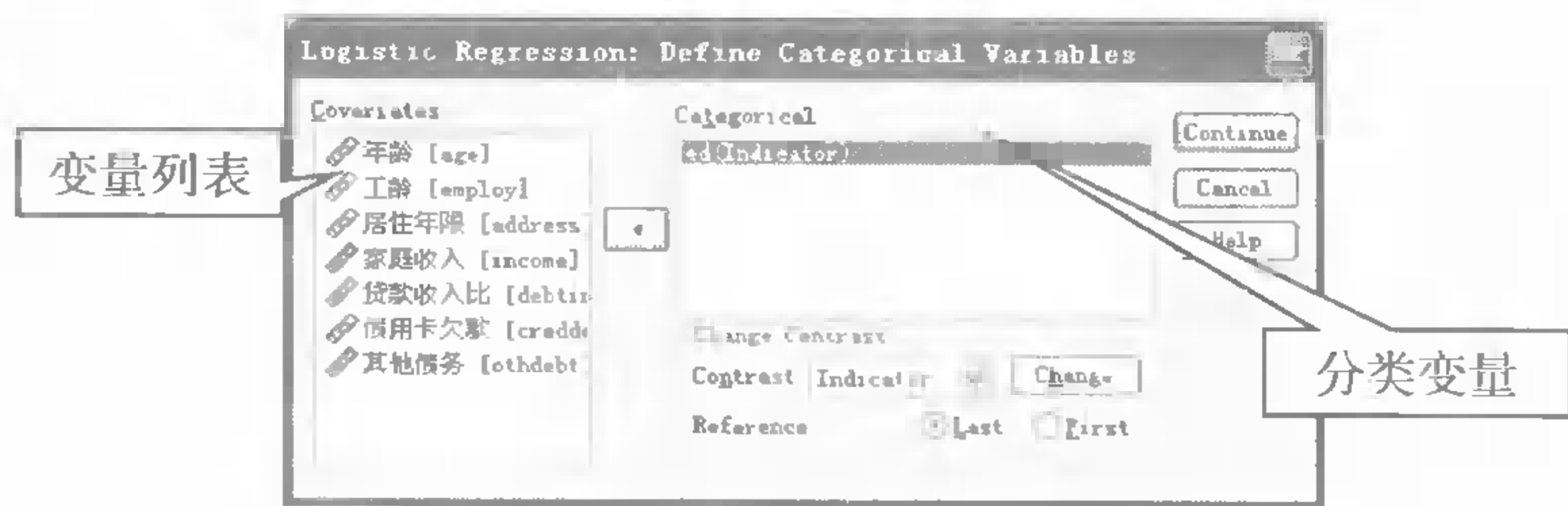


图 8-42 Binary Logistic 过程的分类变量设置对话框

① Covariates 列表框。显示在主面板中选入的全部协变量和交互项, 对于其中的字符串变量或分类变量, 在 Logistic 回归中必须被当作分类协变量来处理。

② Categorical Covariates 列表框。显示当前选择的分类变量, 字符串变量将被自动识别并选入 Categorical 列表。

③ Change Contrast 子设置栏。用于选择分类协变量各水平的对照方式。先在 Categorical 列表中选中需要更改对照方式的分类协变量 (可以同时选中多个), 然后单击 Contrast 下拉列表选择对照方法, 最后单击 Change 按钮确认修改。设置好后, 会在 Categorical 列表里的变量名后以括号方式显示当前变量正在使用的对照方法。SPSS 给出的对照方法有如下 7 种。

- Indicator 指示器, 用于指示是否属于某一个分类, 参考分类在对比矩阵中整行均为 0。
- Simple 简单比较, 预测变量的每个分类 (参考分类除外) 都与参考分类进行比较。
- Difference 差分比较, 除第 1 类外, 预测变量的每个分类都与其前所有分类的平均效应进行比较, 也叫逆 Helmert 比较。
- Helmert (Helmert 比较), 除最后 1 类外, 预测变量的每个分类都与后面所有分类的平均效应进行比较。
- Repeated 重复比较, 除第 1 类外, 预测变量的每个分类都与其前的所有类别进行比较。
- Polynomial 多项式比较, 此方法假设各类别的间距相等, 仅适用于数值型变量。
- Deviation 差别比较, 预测变量的每个分类 (参考分类除外) 都与总体效应进行比较。

④ Reference 栏。用来指定参考分类, 如果选择了 Deviation、Simple 或 Indicator 方法, 就需要指定 1 个参考类别, 可选项有 First (第 1 类) 和 Last (最后 1 类); 默认的参考类都是 Last, 修改后需要单击 Change 按钮确认。对于选择了 First 作为参考类的变量, 其在 Categorical 列表的名称后面会以嵌套括号的方式显示 “First” 字样, 例如: x1 (simple (first)), 表示分类协变量的对照方式是 simple, 参考类为 first; 选择 Last 参考类时不作提示。

### 3. 保存设置

单击图 8-40 中的 Save 按钮,弹出图 8-43 所示的保存设置对话框。依次勾选如下几个复选框: Probabilities、Studentized、Cook's 和 Include the covariance matrix; 单击 Continue 按钮返回主界面。

下面详细介绍各设置选项的含义。

① Predicted Values 栏。设置保存模型的预测值,其可选项有: Probabilities 目标概率,即事件发生的预测概率; Group Membership 预测分类,根据预测概率得到的每个观测的预测分类。

② Influence 栏。设置保存对单个观测记录进行预测时的影响力指标,其可选项有如下 3 个。

- Cook's (Cook 距离),表示把一个个案从计算回归系数的样本中去除时所引起的残差变化的大小, Cook 距离越大,表明该个案对回归系数的影响也越大。
- Leverage values (杠杆值),用来衡量单个观测对回归效果的影响程度,取值范围在 0 到  $n/(n-1)$  之间,取 0 时表示当前记录对模型的拟合无影响。
- DfBeta (s) (DFBeta 值),剔除一个个案后回归系数的改变(包括常数项)。

③ Residuals 栏。设置关于残差的保存选项,可选项有如下 5 个。

- Unstandardized 非标准化残差,观察值与模型预测值之差。
- Logit 逻辑残差,残差除以“预测概率  $\times$  (1-预测概率)”。
- Studentized 学生化残差,用残差除以关于残差标准差的估计值,这个估计值取决于当前个案自变量的取值与自变量均值之间的距离。
- Standardized 标准化残差,其均值为 0,标准差为 1。
- Deviance 变异残差,基于模型变异的残差。

④ Export model information to XML file 栏。设置将模型信息输出到 XML 格式文件的选项,保存结果可以直接用于 SmartScore 和 SPSS Server,单击 Browse 按钮指定文件名称及路径。选中 Include the covariance matrix 复选框,表示保存协方差阵于如上的 XML 文件中。

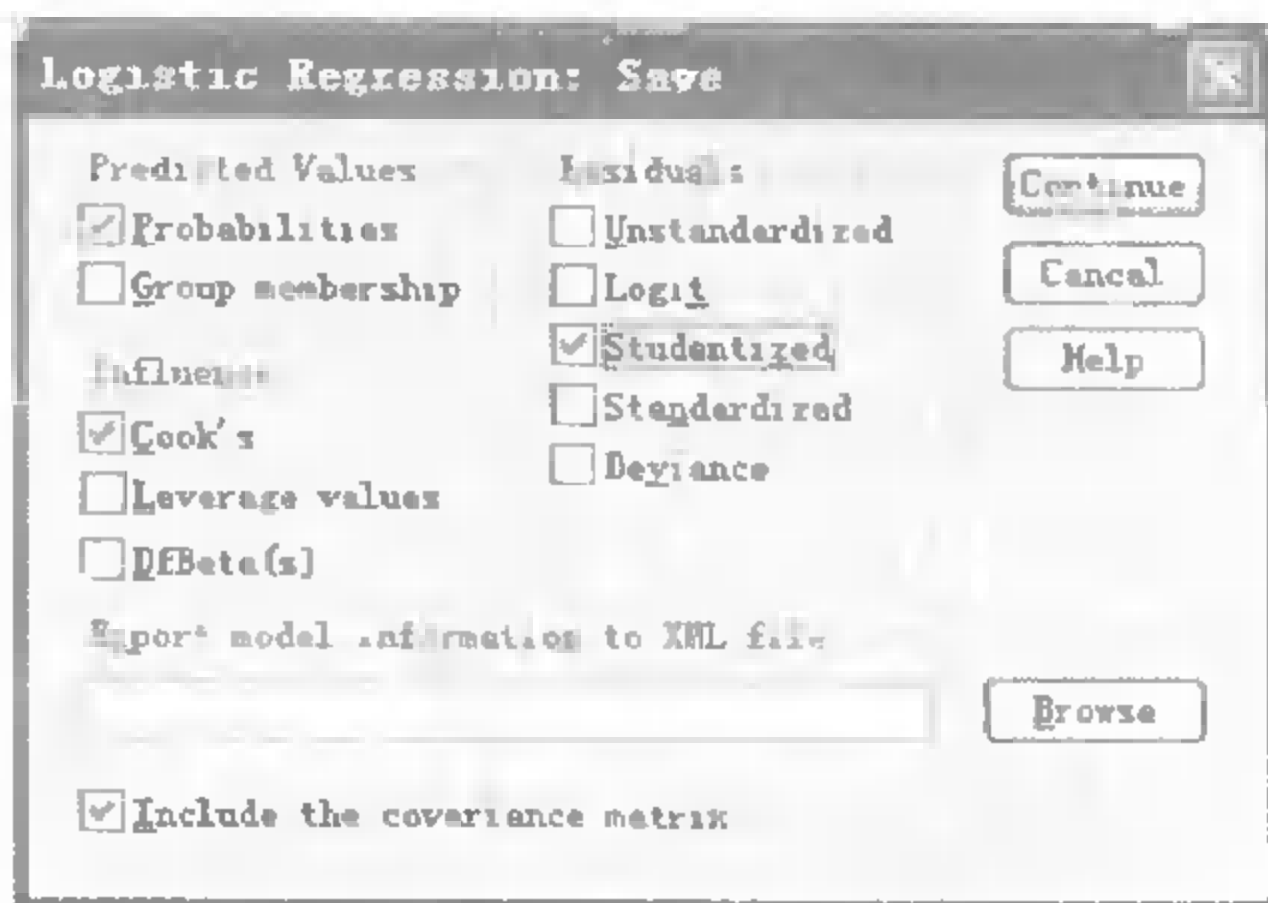


图 8-43 保存选项设置

### 4. 选项设置

单击图 8-40 中的 Options 按钮,弹出图 8-44 所示的对话框,用于设置关于输出和显示的选项。勾选 Classification plots 复选框和 Hosmer-Lemeshow 复选框;单击 Continue 按钮返回主界面。

① Statistics and Plots 栏。用于选择输出哪些统计量和图形,可选的设置内容有如下 6 个。

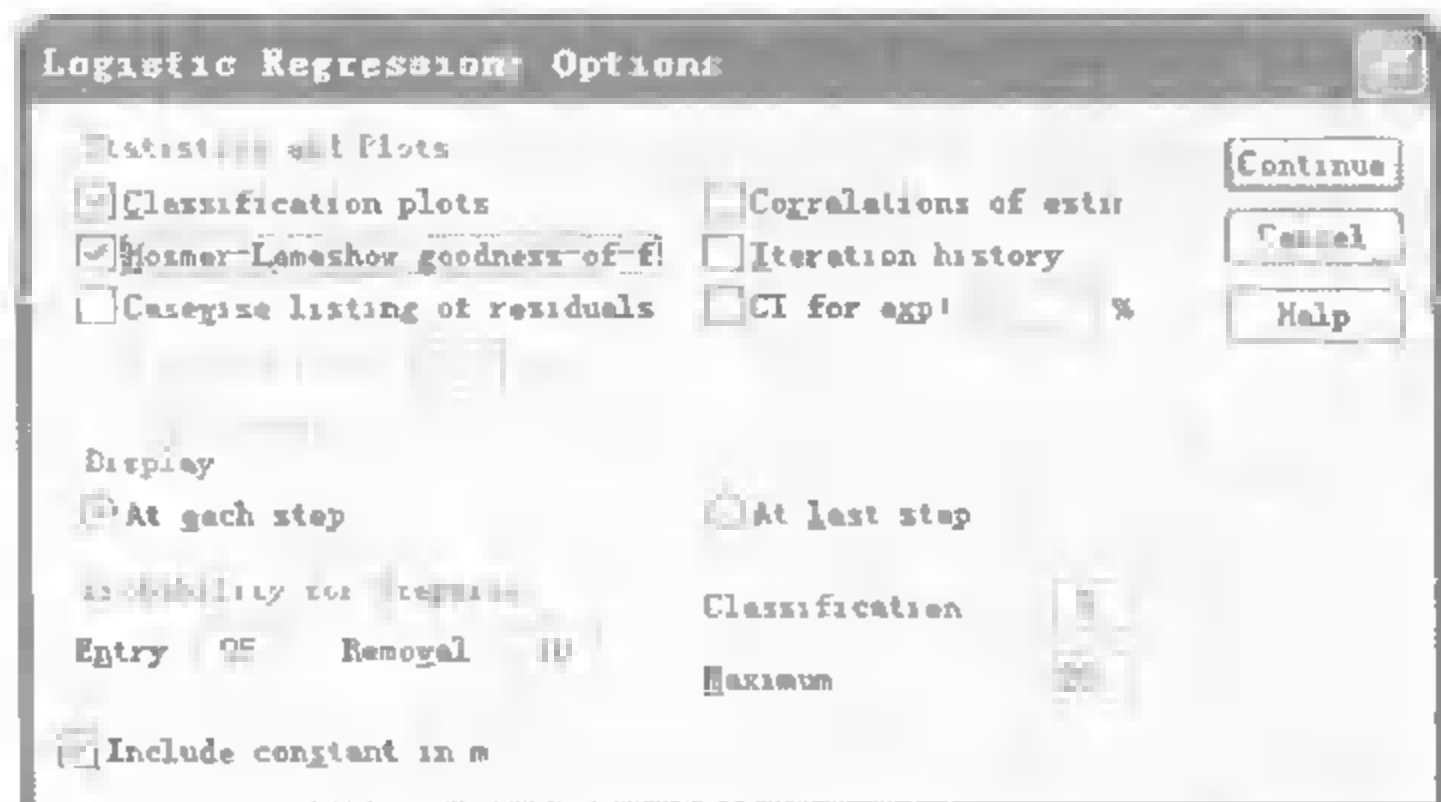


图 8-44 输出选项设置

● Classification plots 分类图,因变量的预测值与观察值的分类直方图。

● Hosmer-Lemeshow goodness-of-fit (Hosmer-Lemeshow 拟合优度),它比传统 Logistic 回归分析的拟合优度更稳定,特别是对含有连续型协变量的模型和对小样本的研究。

● Casewise listing of residuals 残差案例清单,包括非标准化残差、预测概率、观测量的实际与预

测分组水平。Outliers outside  $n$  std.Dev 选项,  $n$  为输入框中指定的正数, 表示只对那些标准化残差大于  $n$  倍标准差的观测量, 输出指定的统计量; All cases 选项, 输出对所有观测的各种统计量。

- ① Correlations of estimates 复选框, 输出参数估计值的相关系数矩阵。
- ② Iteration history 复选框, 输出每一步迭代的相关系数和对数似然比。
- ③ CI for exp(B) 复选框, 设置指数域的置信区间, 在输入框指定一个 1~99 的数值。

② Display 栏。在此选择输出结果的范围, 有如下两个可选项。

- ① At each step, 表示在每一步迭代过程都输出相关的表格、统计量和图形。
- ② At last step, 表示只输出与最终方程有关的表格、统计量和图形。

③ Probability for Stepwise 栏。用于设置变量进入模型和从模型中剔除的依据。如果某变量得分统计量的概率值小于 Entry 处的设置值 (默认 0.05), 那么此变量进入模型; 如果这个概率值大于 Removal 处的设置值 (默认 0.10), 该变量从模型中删除。而且指定的 Entry 值必须小于 Removal 值。

④ Classification cutoff 输入框。用于指定对观测量进行预测分类的临界值, 预测值大于指定值的观测量被归于一类, 其余的观测量被归于另一类, 可设置的范围为 0.01~0.99, 默认值为 0.5。

⑤ Maximum Iterations 输入框, 用于指定模型允许的最大迭代步数。

⑥ Include constant in model 复选框, 勾选它表示在模型中包括非零的常数项。

#### 8.4.4 案例的结果分析

单击图 8-40 中的 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 8-45~图 8-51 所示。

案例处理摘要

未加权的案例 <sup>a</sup>	N	百分比
已选定的案例	499	58.7
包括在分析中	0	0
缺失案例	499	58.7
总计	850	100.0

a. 如果权重有效, 请参见分类表以获得案例总数。

因变量编码

初始值	内部值
No	0
Yes	1

分类变量编码

		频率	参数编码			
			(1)	(2)	(3)	(4)
教育水平	Did not complete high school	266	1.000	.000	.000	.000
	High school degree	134	.000	1.000	.000	.000
	Some college	69	.000	.000	1.000	.000
	College degree	25	.000	.000	.000	1.000
	Post-undergraduate degree	5	.000	.000	.000	.000

图 8-45 案例处理摘要和变量编码信息

模型摘要			
步骤	-2 对数似然值	Cox & Snell R 方	Nagelkerke R 方
1	498.012 <sup>a</sup>	.116	.172
2	447.301 <sup>b</sup>	.201	.299
3	411.553 <sup>b</sup>	.257	.381
4	394.721 <sup>c</sup>	.281	.417

<sup>a</sup> 因为参数估计的更改范围小于 .001, 所以估计在迭代次数 4 处终止。

<sup>b</sup> 因为参数估计的更改范围小于 .001, 所以估计在迭代次数 5 处终止。

<sup>c</sup> 因为参数估计的更改范围小于 .001, 所以估计在迭代次数 6 处终止。

图 8-46 R 方统计量表



Hosmer 和 Lemeshow 检验						
步骤	卡方	df	显著性			
1	3.292	8	.915			
2	11.866	8	.157			
3	9.447	8	.306			
4	4.027	8	.855			

Hosmer 和 Lemeshow 检验的随机性表						
		是否拖欠 = No		是否拖欠 = Yes		总计
		观察值	期望值	观察值	期望值	
步骤 1	1	50	49.778	0	.222	50
4	2	49	48.995	1	1.005	50
	3	47	47.549	3	2.451	50
	4	45	45.495	5	4.505	50
	5	46	42.992	4	7.008	50
	6	39	39.783	11	10.217	50
	7	32	35.801	18	14.199	50
	8	33	30.474	17	19.526	50
	9	24	23.443	26	26.557	50
	10	10	10.689	39	38.311	49

图 8-47 Hosmer-Lemeshow 检验表

分类表<sup>d</sup>

观察值			预测值				
			已选定的案例 <sup>a</sup>		未选定的案例 <sup>b,c</sup>		
			是否拖欠		是否拖欠		
			No	Yes	No	Yes	
步骤 1	是否拖欠	No	361	14	96.3	137	5
		Yes	100	24	19.4	45	14
	总百分比				77.2		
							75.1
步骤 2	是否拖欠	No	351	24	93.6	136	6
		Yes	80	44	35.5	36	23
	总百分比				79.2		
							79.1
步骤 3	是否拖欠	No	348	27	92.8	135	7
		Yes	72	52	41.9	28	31
	总百分比				80.2		
							82.6
步骤 4	是否拖欠	No	352	23	93.9	130	12
		Yes	67	57	46.0	27	32
	总百分比				82.0		
							80.6

a 已选定的案例 有效数据 EQ 1

b 未选定的案例 有效数据 NE 1

c 由于自变量中有缺失值，或分类变量中的值超出选定案例的范围，所以未对某些未选定的案例进行分类。

d 切割值为 .500

图 8-48 观测量分类表

方程中的变量							
步骤		B	S.E.	Wald	df	显著性	Exp(B)
步骤 1	debtinc	.121	.017	52.676	1	.000	1.129
	常量	-2.476	.230	116.315	1	.000	.084
步骤 2	employ	-.140	.023	36.158	1	.000	.869
	debtinc	.134	.018	54.659	1	.000	1.143
	常量	-1.621	.259	39.036	1	.000	.198
步骤 3	employ	-.244	.033	54.676	1	.000	.783
	debtinc	.069	.022	9.809	1	.002	1.072
	creddebt	.506	.101	25.127	1	.000	1.658
	常量	-1.058	.280	14.249	1	.000	.347
步骤 4	employ	-.247	.034	51.826	1	.000	.781
	address	-.089	.023	15.109	1	.000	.915
	debtinc	.072	.023	10.040	1	.002	1.074
	creddebt	.602	.111	29.606	1	.000	1.826
	常量	-.605	.301	4.034	1	.045	.546

a 在步骤 1 中输入的变量 debtinc  
b 在步骤 2 中输入的变量 employ  
c 在步骤 3 中输入的变量 creddebt  
d 在步骤 4 中输入的变量 address

图 8-49 方差中变量系数的参数估计

不在方程中的变量				
步骤	变量	概率	df	显著性
4	age	1 671	1	196
	ed	2 095	4	718
	ed(1)	713	1	398
	ed(2)	939	1	733
	ed(3)	099	1	753
	ed(4)	643	1	423
	income	856	1	355
	othdebt	156	1	665
总统计量		6 178	7	519

图 8-50 模型外变量统计表

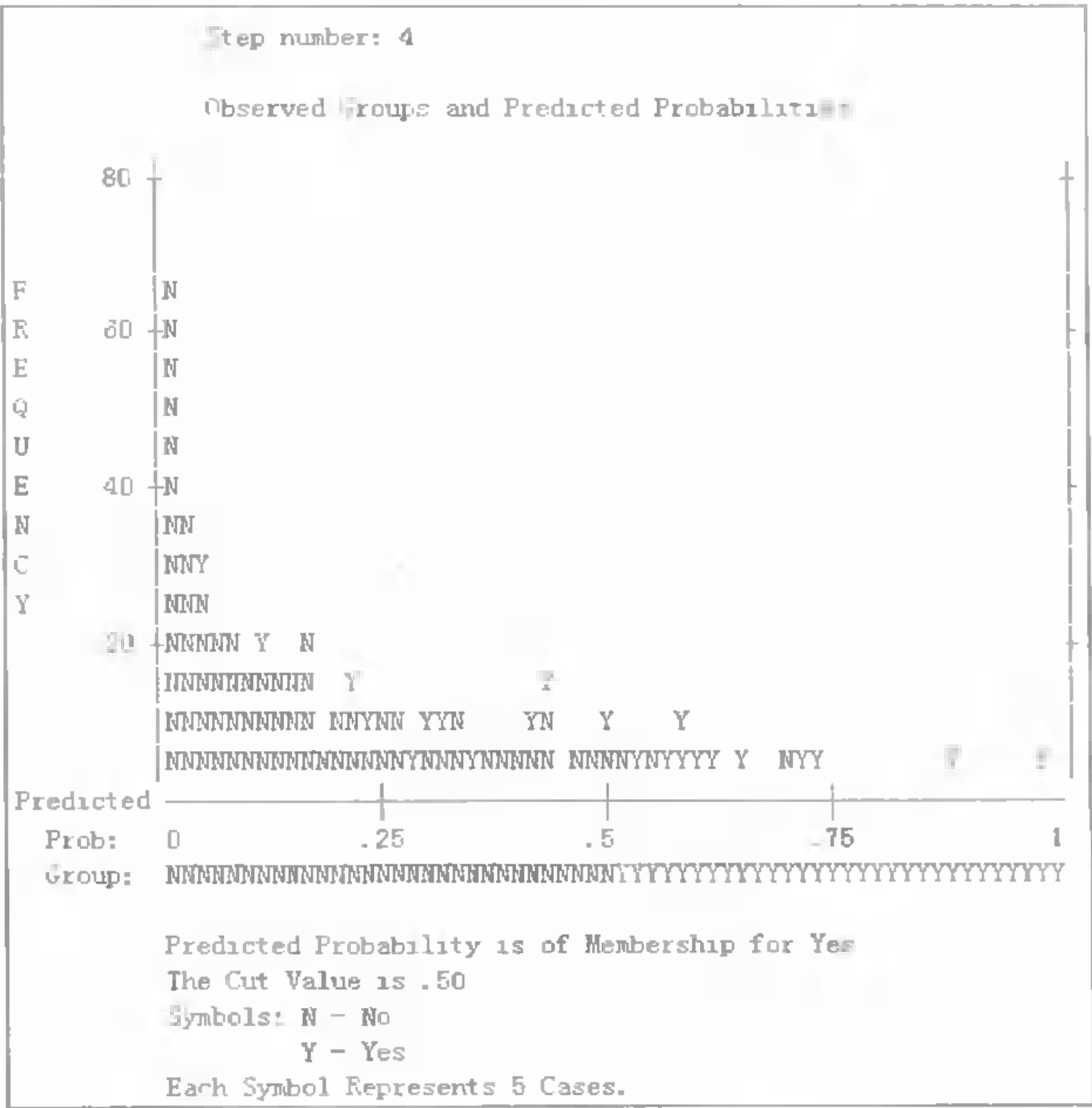


图 8-51 根据预测概率的观测量分组

(1) 案例处理摘要和变量编码信息。如图 8-45 所示，“案例处理摘要”表格显示有 499 个（58.7%）观测记录用于模型的估计。“因变量编码”和“分类变量编码”表格分别列出了是否拖欠和教育水平两个变量的编码情况。

(2) 模型概要。如图 8-46 所示，“模型摘要”表格中以 Cox and Snell's R 方和 Nagelkerke's R 方两个统计量取代了线性回归中的 R 方统计量。本例中它们的取值分别为 0.281 和 0.417，只看这一点，模型拟合的并不太理想；但这两个统计量一般用于不同模型之间的比较，R 方值（小于 1）越大的模型拟合效果越好。

(3) Hosmer-Lemeshow 检验结果。如图 8-47 所示，“Hosmer 和 Lemeshow 检验”表格的零假设使模型能够很好的拟合数据。从其显著性检验的 Sig=0.855>0.5 看，接受零假设，认为模型能够很好的拟合数据。

“Hosmer 和 Lemeshow 检验的随机性”表格根据目标变量的预测概率，把结果分为个数大致相等的 10 个组；“总计”列中是每组的观测数，由于预测值相等的观测被分在一起，所以各组的观测数不一定相同。此表直观地反应了模型预测的效果，可以看出各行的观测值和预测值都大致相同，所以模型的拟合效果不错。

(4) 预测分类结果。如图 8-48 所示，“分类表”给出了关于观测值和预测值的列联表。“已选定的案例”（validate=1）列表示对建模所用数据的回判分类结果；“未选定的案例”（validate=0）列表示对未使用的验证数据的判别分类结果。另外，如果预测概率大于 0.5，预测为 Yes（有拖欠），反之预测为 No（没有拖欠）。

对于最终模型，建模用的 124 个拖欠贷款用户中有 57 个判断正确，正确率为 46.0%；建模用的 375 无拖欠贷款用户中有 352 个判断正确，正确率为 93.9%；对建模数据总的回判正确率为 82.0%，这说明模型的预测效果不错，尤其是对那些无拖欠贷款用户的预测。

由于验证数据没有参与建模，所以用对它的分类结果来验证模型效果更有参考意义，能保证模型的稳定性和通用性。本例的总验证正确率达到 80.6%，说明模型较为稳定。

(5) 逐步回归过程。如图 8-49 所示,“方程中的变量”表格给出了每一步回归的参数估计信息。

以步骤 4 的最终模型为例,由 B 列的系数可得二元 Logistic 模型为  $p=1/(1+e^{-z})$ , 其中  $z=-0.605-0.247\text{employ}-0.089\text{address}+0.072\text{debtinc}+0.602\text{creddebt}$ 。由于 B 列的系数是线性的,所以用来检验其显著性比较方便;但在 Logistic 回归里,Exp(B)列的系数更易于解释,它反应了自变量变动 1 个单位而引起的发生比 Odds 的变化率。以图中棕色线框标识的 Exp(B)为例,表示其他情况不变的条件下,工龄为 2 年的用户拖欠贷款发生比 Odds 是工龄为 1 年用户的 0.781 倍。另外, Wald 统计量的 Sig 值全部小于 0.05,说明参数估计值都显著地不为 0。

利用此处得到的最终模型,就可以对 150 名预期用户进行信誉分类,当其预测概率大于 0.5 时,说明此客户信誉不高,可能会拖欠贷款;反之则推断他不会拖欠贷款。

(6) 不在方程中变量的统计信息。如图 8-50 所示,“不在方程中的变量”表格给出了逐步回归的最后一步中没有进入方程的变量信息,这些变量的显著性值都大于 0.05,故而不是不显著的。

(7) 预测概率的直方图。如图 8-51 所示,横轴是对拖欠贷款概率的预测概率值,纵轴是观测的频数。图中的符号指示观测量实际归属的类别, Y 代表 Yes (拖欠贷款), N 代表 No (不拖欠贷款)。如果模型对原数据成功地进行了模拟,则发生拖欠贷款事件的观测(Y)应分布在图形的右侧,其它观测(N)应更多地地位于左侧;而且两类观测越是分布在两端,预测效果越好。本例中绝大部分观测集中在小于 0.5 的一侧,并且不同性质的观测基本适当地分布于两端,只有少数拖欠贷款的观测错误地分在无拖欠贷款的一端,总体来看模型的拟合效果不错。

(8) 检测异常值的图形。运行结束后,还会在当前数据集新增 3 个变量:预测概率(PRE\_1)和学生化残差(SRE\_1)和 Cook 距离(COO\_1)。

以学生化残差的平方(反应了模型的变异信息)为纵轴,预测概率为横轴,所作图形如图 8-52 所示,纵轴(残差)取值较大的点表示模型对这些点的拟合效果较差。左低右高的蓝色曲线代表的是因变量取 0 (没有拖欠贷款)的观测的残差变化,说明这类观测的预测概率越大,拟合效果越差;左高右低的绿色曲线的含义相仿。

图 8-53 所示是 Cook 距离对预测概率的散点图,图中有较少的几个奇异值,它们的 Cook 值都很大,可能影响了分析,可以进一步对它们进行单独研究。

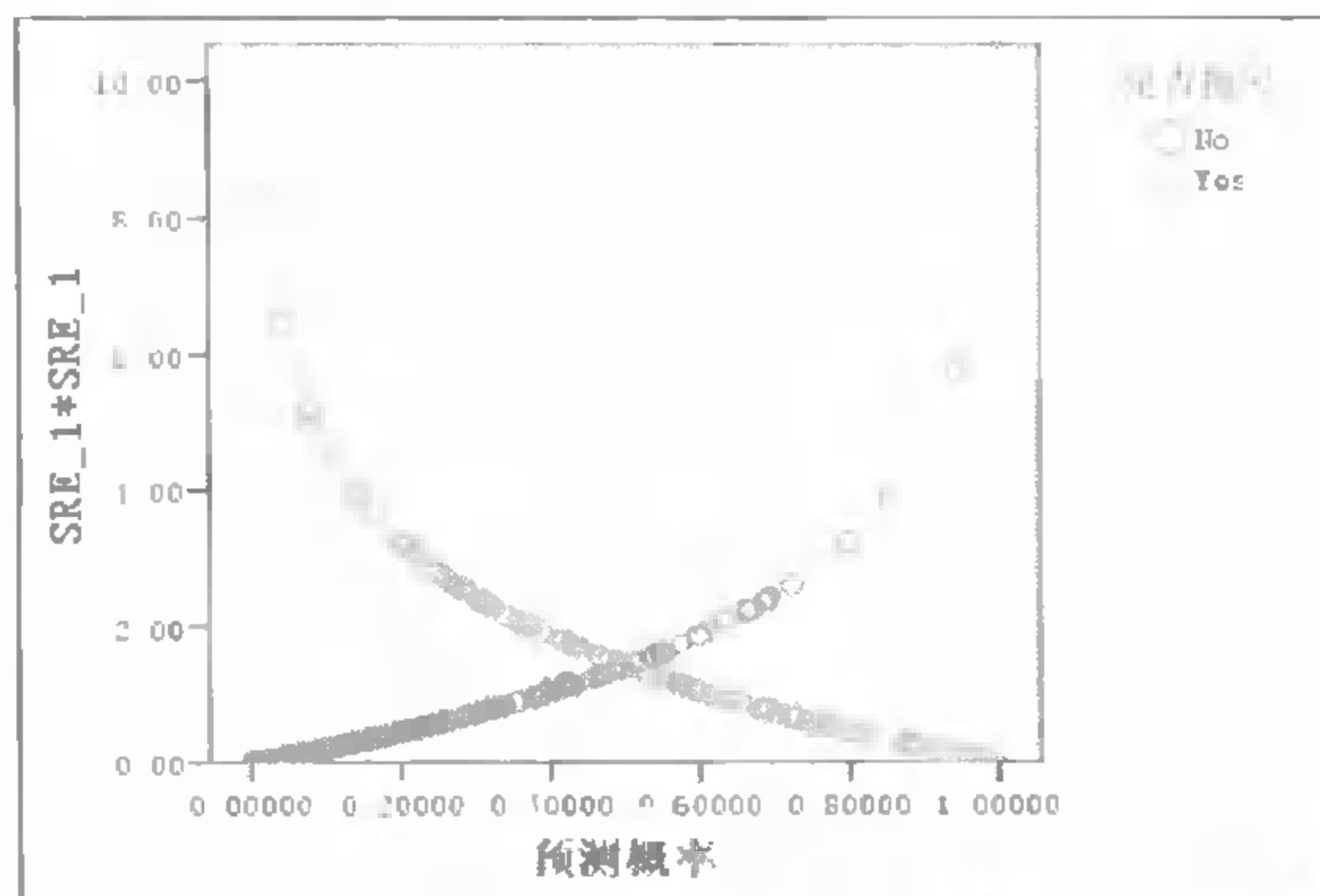


图 8-52 关于残差的散点图

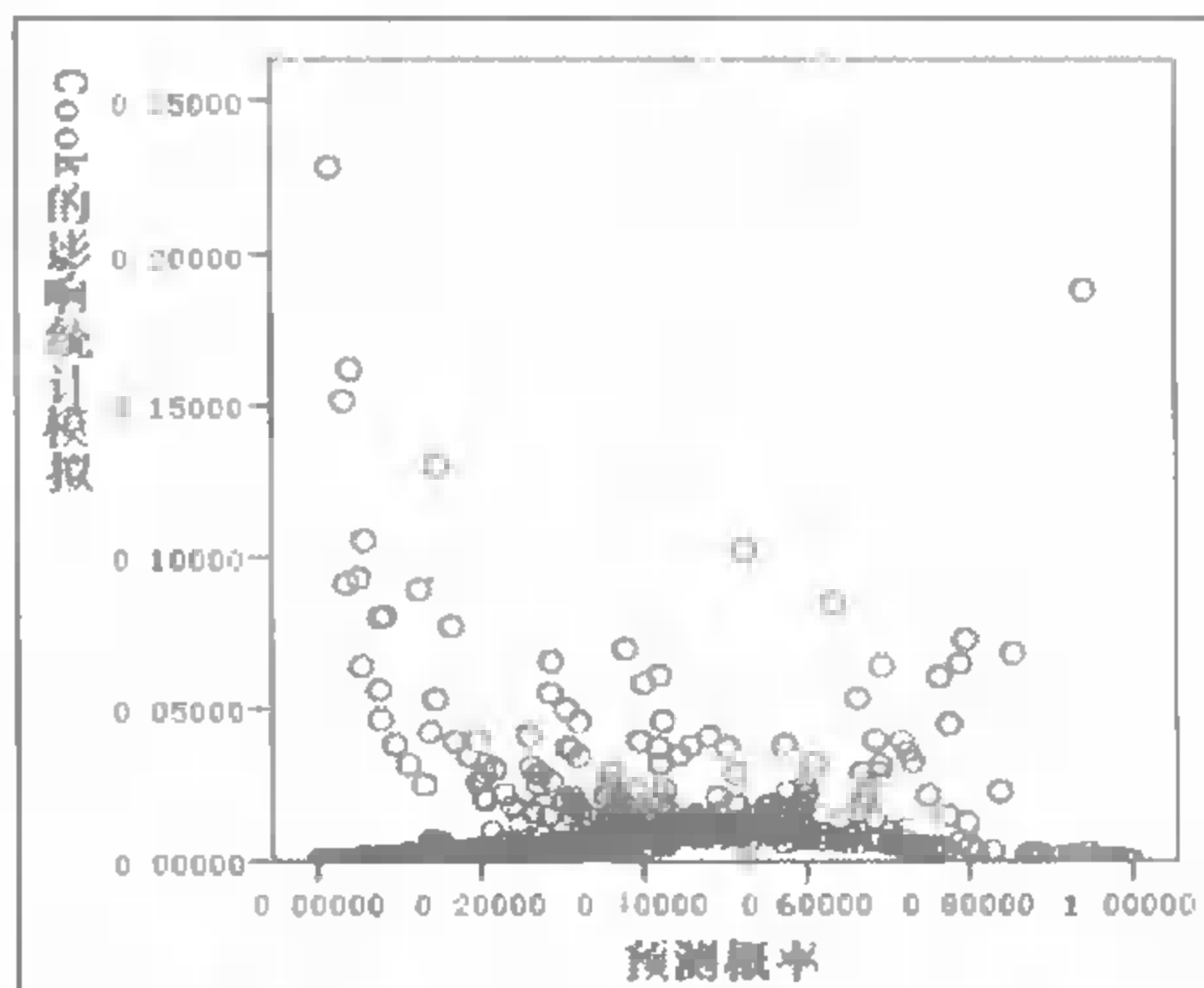


图 8-53 Cook 距离对预测概率的散点图

8.5 多元 Logistic 回归分析

在实际生活中，有时会遇到因变量是多分类（多元）变量的情况。多元变量又可分两种情况：无序多元变量，如胃病可以分为胃炎、不典型增生和胃癌；有序多元变量，如疾病的疗效结果可能是治愈、好转、无效。对于这种分类资料，不能直接使用二元 Logistic 回归分析来处理，而要使用多元 Logistic 回归分析。

多元 Logistic 回归分析实际上就是用多个二元 Logistic 回归分析模型来描述各个类别与参照类别相比较时的作用大小。例如：对于一个三分类的因变量（治愈、好转、无效），可建立两个二元 Logistic 回归分析模型，分别描述好转与治愈相比时、无效与治愈相比时各种疗法的作用，但是在估计这些模型的参数时，所有对象都是一起估计的，一些参数的意义、模型的筛选过程等都和二元 Logistic 回归分析很相似。

8.5.1 多元 Logistic 回归的原理简介

1. 模型的基本形式

对于自变量是连续型变量或计数型变量，且因变量每个取值的概率范围均为 0~1 的情况，都可以用 Logistic 回归方法对因变量的概率取值建立回归模型。设因变量有  $j$  个取值水平，可以对其中的  $j-1$  个水平，各做一个回归方程。

因变量取第  $i$  个水平时的 Logistic 回归模型设为  $\ln(\frac{p_j}{1-p_j}) = \alpha_{i0} + \sum_{p=1}^m \beta_{ip}x_p$ 。这样，对于建立的每一个 Logistic 模型都将获得一组回归系数，如果因变量具有 3 种分类，就将获得两组非零的回归参数。

2. 模型检验

（1）拟合检验。Pearson 卡方统计量，常用在多维表中检验观测频数与预测频数之间的差异。如果卡方值越大，显著性水平越低，模型拟合效果越不好。另一个检验模型拟合优度的指标为卡方偏差统计量，大样本数据的这两个统计量的取值很相近。

（2）伪 R 方统计量。指 McFadden 统计量，计算公式为  $R^2_{McFadden} = \frac{l(0)-l(B)}{l(0)}$ 。其中  $l(B)$  为模型中对数似然比的和， $l(0)$  为只包括截距的模型的对数似然比的和。

8.5.2 问题描述和数据准备

某快餐公司为了提高其早餐的市场份额，对 880 名消费者做了一份调查，本节利用多元 Logistic 回归方法分析 3 种早餐的市场销售情况。数据来自 SPSS 自带的 Demo 文件“cereal.sav”，所用数据文件为“早餐偏好调查数据.sav”，数据格式如图 8-54 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	agecat	Numeric	4	0	年龄段	{1, 低于31}...	None	8	Right	Ordinal
2	gender	Numeric	4	0	性别	{0, 男}...	None	8	Right	Nominal
3	active	Numeric	4	0	生活方式	{0, 消极}...	None	8	Right	Nominal
4	bfast	Numeric	4	0	早餐	{1, 不吃}...	None	8	Right	Nominal
5	marital	Numeric	4	0	婚否	{0, 未婚}...	None	8	Right	Nominal

图 8-54 早餐偏好调查数据格式



调查设计的问卷提出了如图所示的4个问题,其中早餐变量取值1表示不吃,取值2表示吃麦片,取值3表示吃谷类。

### 8.5.3 多元 Logistic 回归参数设置

依次单击菜单“Analyze→Regression→Multinomial Logistic...”执行多元 Logistic 回归分析的功能,其主设置界面如图 8-55 所示。

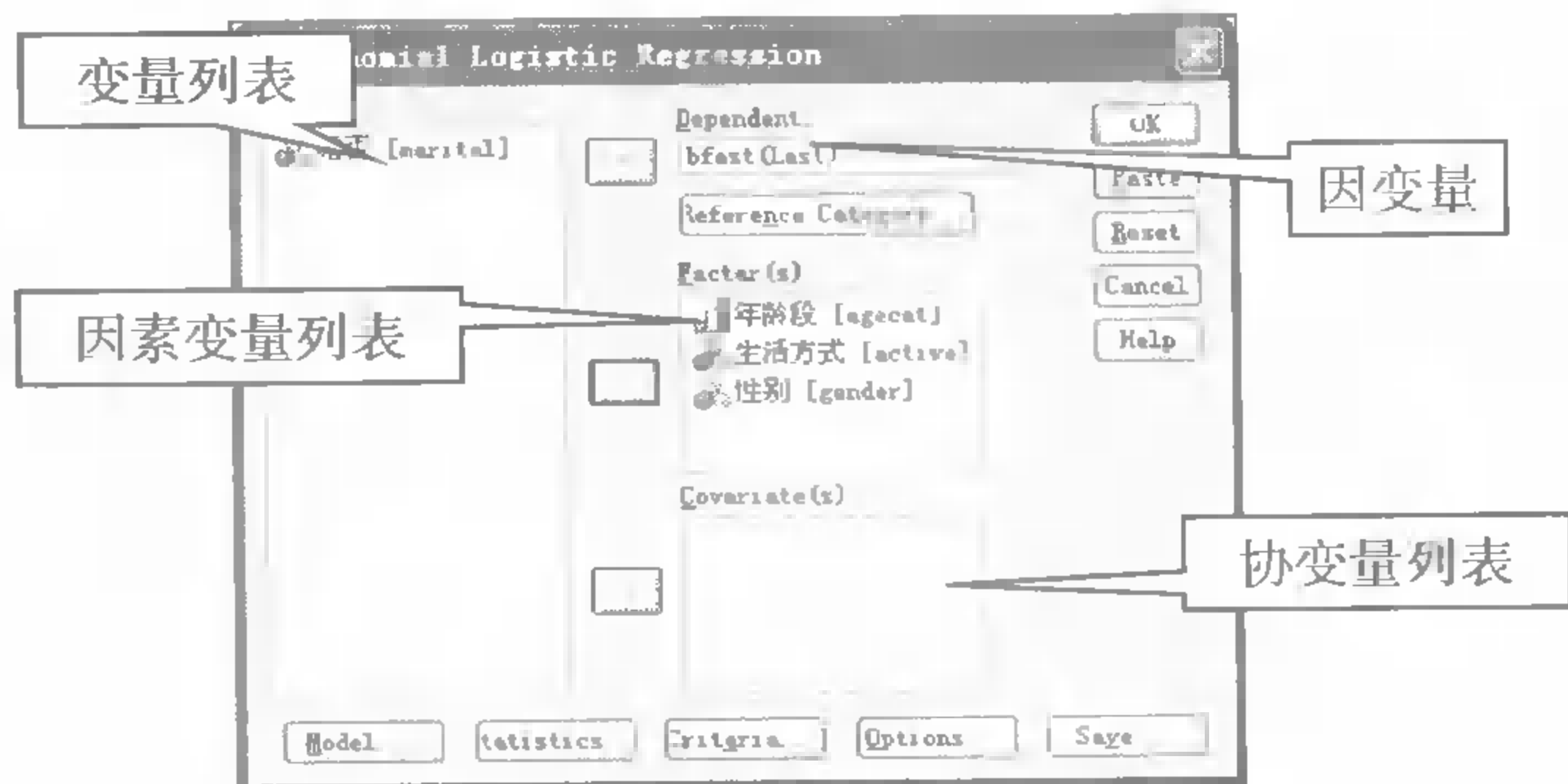




图 8-55 多元 Logistic 分析的主设置界面

#### 1. 变量设置

如图 8-55 所示,在变量列表中单击选中早餐 (bfast) 变量,单击从上至下第一个  按钮,将其作为因变量选入 Dependent 选框;在变量列表选中年龄段、生活方式和性别

变量,单击从上至下第二个  按钮,将其作为因素变量选入 Factor(s) 列表框。单击 Reference Category 按钮,弹出如图 8-56 所示的对话框,在此设置因变量的参考类,单击 Continue 按钮返回主界面。

下面详细介绍各设置选项的含义。

(1) 指定分析变量。Dependent 栏用于从变量列表选入一个分类变量作为因变量;Factor(s) 栏用于选入分类自变量作为因素变量;Covariates 栏用于

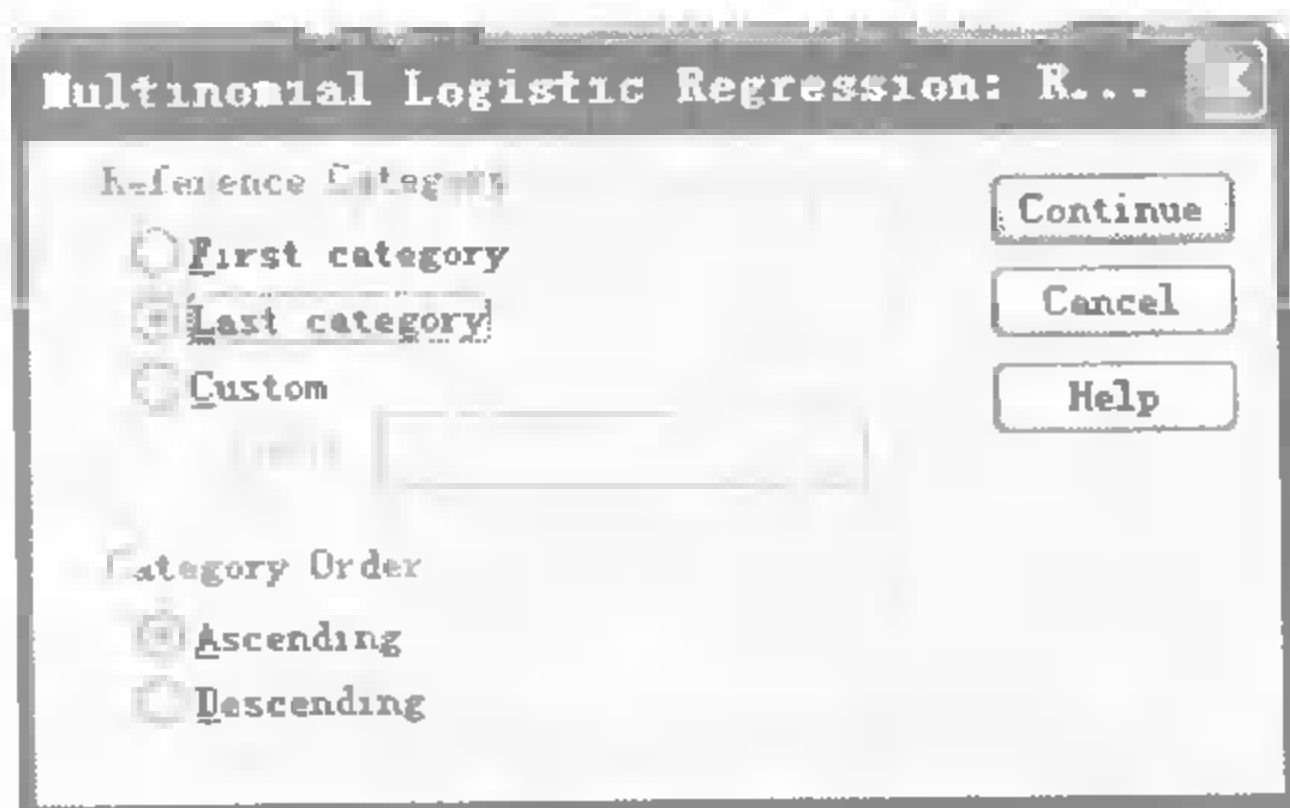





图 8-56 多元 Logistic 分析的参考类设置



于选入连续自变量作为协变量。

(2) 指定参考类。如图 8-56 所示,默认使用因变量的最大取值作为参考类,设置内容如下两个部分。


① Reference Category 栏,用于设置参考类的取值,有如下 3 个选择。

-  First category, 指定第一类为参考类。
-  Last category, 指定最后一类为参考类。
-  Custom category, 由用户指定参考类,需要在 Value 后输入它的取值。

② Category Order 栏,设置区分第一类和最后一类的顺序,有如下两个选择。

-  Ascending 升序,分类变量中取值最小的类为第一类,取值最大的类为最后一类。
-  Descending 降序,分类变量中取值最大的类为第一类,取值最小的类为最后一类。

## 2. 模型设置

在图 8-55 中单击 Model 按钮，弹出如图 8-57 所示的对话框，在此设置模型的具体形式。单击选中 Custom/Stepwise 单选项；在 Factors 列表框选中 agecat 和 active 变量，单击从上至下第一个下拉列表并选中 Main effects 选项，单击其上方的  按钮，将选中的两个主效应选入 Forced 列表框；单击 Continue 按钮返回主界面。

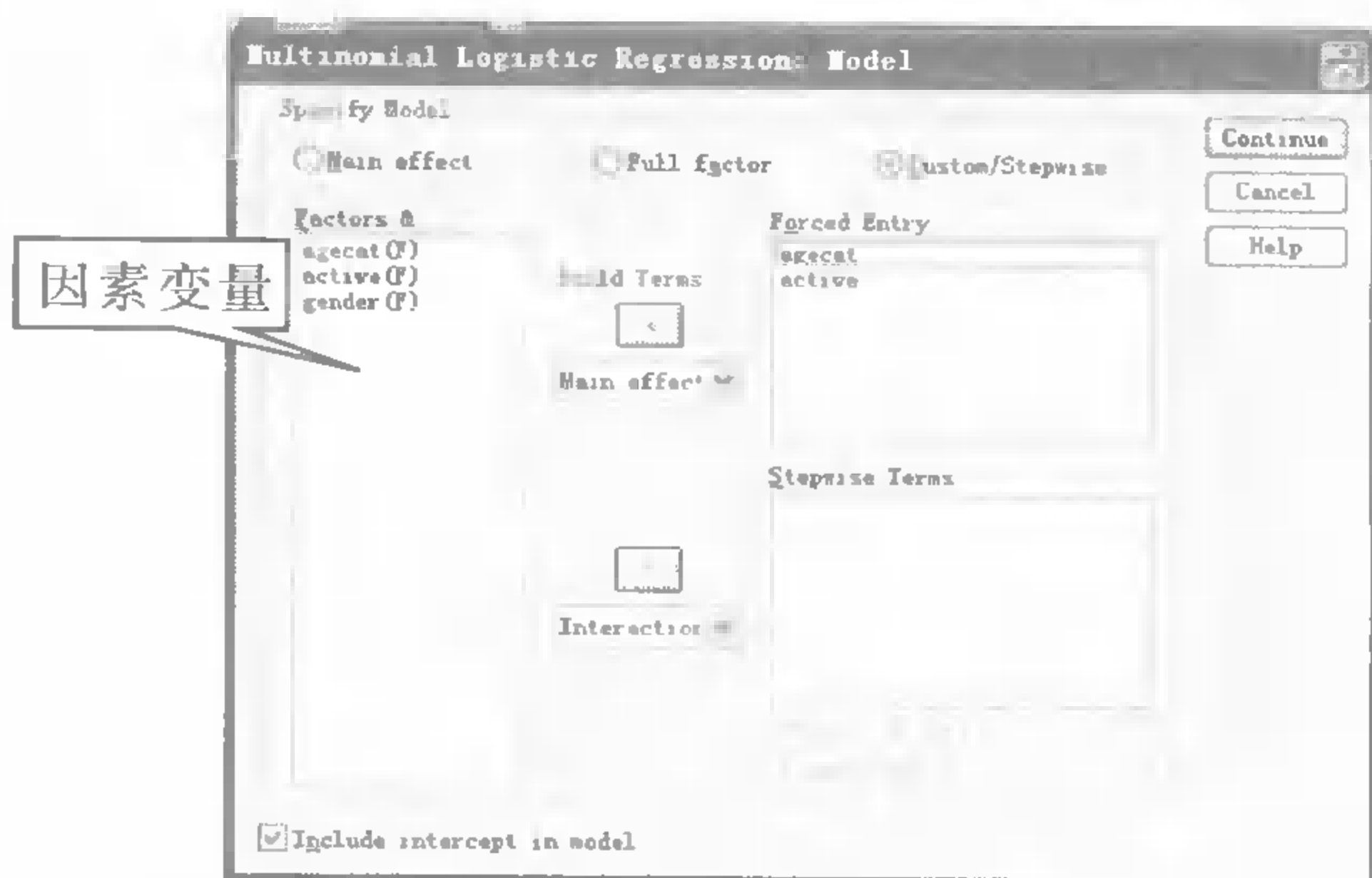


图 8-57 多元 Logistic 分析的模型设置

下面详细介绍各设置选项的含义。

(1) Specify Model 栏。在此选择如何指定回归模型的效应，可选方式有：Main effects 主效应，表示模型中只包括协变量和因素变量的主效应，不包括任何交互效应；Full factorial 全因素，表示模型中包含所有主效应以及它们之间所有可能的交互效应；Custom/Stepwise 自定义/逐步模型，由用户自行选择使用那些效应进行分析，选中它后，激活下面的设置选项。

(2) Factors and Covariates 列表框，显示主面板中选入的协变量 (C) 和因素变量 (F)。

(3) Include intercept in model 复选框，勾选后要求在模型中包含截距项。

(4) Build Team 栏。有两个下拉列表，都用来指定效应的种类，可选项有 6 个：Main effects, 主效应；Interaction, 交互效应；All n-way, 所有  $n$  维交互效应 ( $n=2, 3, 4, 5$ )。

(5) Forced Entry Terms 强制进入列表，选入此列表的效应将强制出现在模型中。

(6) Stepwise Terms 逐步选入列表，选入此列表的效应将以逐步回归的方式加入模型。

(7) Stepwise Method 栏

用于设置 Stepwise Terms 列表里的变量逐步进入模型的方法，下拉列表的可选项有 Forward entry (向前进入法)、Backward elimination (向后消去法)、Forward stepwise (向前逐步法) 和 Backward stepwise (向后逐步法)。

## 3. 统计量设置

在图 8-55 中单击 Statistics 按钮，弹出如图 8-58 所示的对话框，在此选择分析过程的输出统计量。依

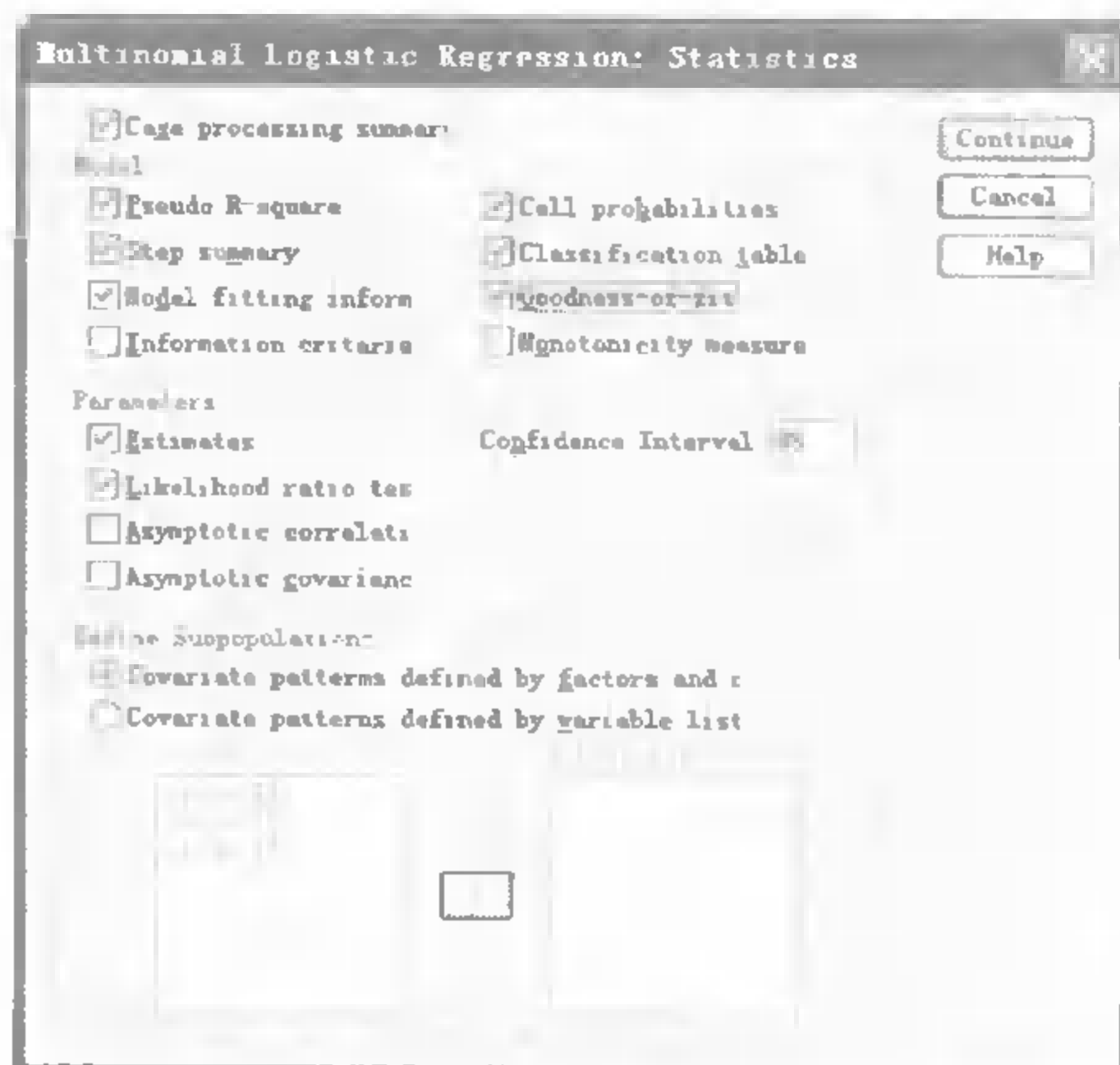


图 8-58 多元 Logistic 分析的统计量设置

次勾选如下3个复选框: Cell probabilities、Classification table 和 Goodness of fit; 单击 Continue 按钮返回主界面。

(1) Case processing summary 复选框, 输出个案处理摘要, 主要指分类变量的综合信息。

(2) Model 子设置栏, 用于选择关于统计模型的统计量, 可选内容有如下8个。

● Pseudo R-square 伪  $R^2$  统计量, 输出 Cox&Snell  $R^2$ 、Nagelkerke  $R^2$  和 McFadden  $R^2$  这3个  $R$  方统计量。

● Step summary 步骤摘要, 如果模型选择了逐步方法, 此选项要求显示每一步的变量进入或剔除出模型的效应表。

● Model fitting information 选项, 输出拟合模型信息以及只包含截距项的模型信息。

● Information criteria 逐步回归的判别准则, 输出 Akaike 信息标准 (AIC) 和施瓦兹-贝叶斯信息标准 (BIC)。

● Cell probabilities 复选框, 输出观测频数和期望频数表 (带残差)、协变量比率和响应分类表。

● Classification table 复选框, 输出关于最终预测分类的统计信息表。

● Goodness of fit 拟合优度统计量, 输出 Pearson 卡方和似然比卡方统计量。

● Monotonicity measures 复选框, 输出协调对、不协调对和约束对的个数统计信息, 输出表中还包括 Somers'D、Goodman&Kruskal's Gamma、Kendall's tau-a 和 Concordance Index C 这些统计量。

(3) Parameters 子设置栏, 用于选择关于模型参数的输出统计量, 有如下4个可选项。

● Estimates 复选框, 输出模型参数的估计值, 包括估计值的置信区间, 在 Confidence Interval 输入框指定置信区间的范围, 默认为 95%。

● Likelihood ratio test 似然比检验, 输出关于模型偏效应的似然比检验。

● Asymptotic correlations 渐近相关系数矩阵, 输出参数估计值的相关系数矩阵。

● Asymptotic covariances 渐近协方差矩阵, 输出参数估计值的协方差矩阵。

(4) Define Subpopulations 栏, 设置分组定义, 有如下两个选择。

● Covariates pattern defined by factor and Covariate 单选项, 对所有的因子变量和协变量计算单元概率, 并进行拟合优度检验, 此为默认选项。

● Covariates pattern defined by variable list below 单选项, 指定要求输出单元概率和拟合优度检验的变量, 从左侧的列表框将输出变量选入右侧的 Subpopulations 列表框即可。

#### 4. 收敛标准设置

在图 8-55 中单击 Criteria 按钮, 弹出如图 8-59 所示的对话框, 在此设置逐步回归的收敛标准, 单击 Continue 按钮返回主界面。

(1) Iterations 子设置栏指定收敛标准, 可选参数有如下6个。

● Maximum iterations 输入框, 指定最大迭代次数。

● Maximum step-halving 输入框, 指定最大等分。

● Log-likelihood convergence 下拉列表,

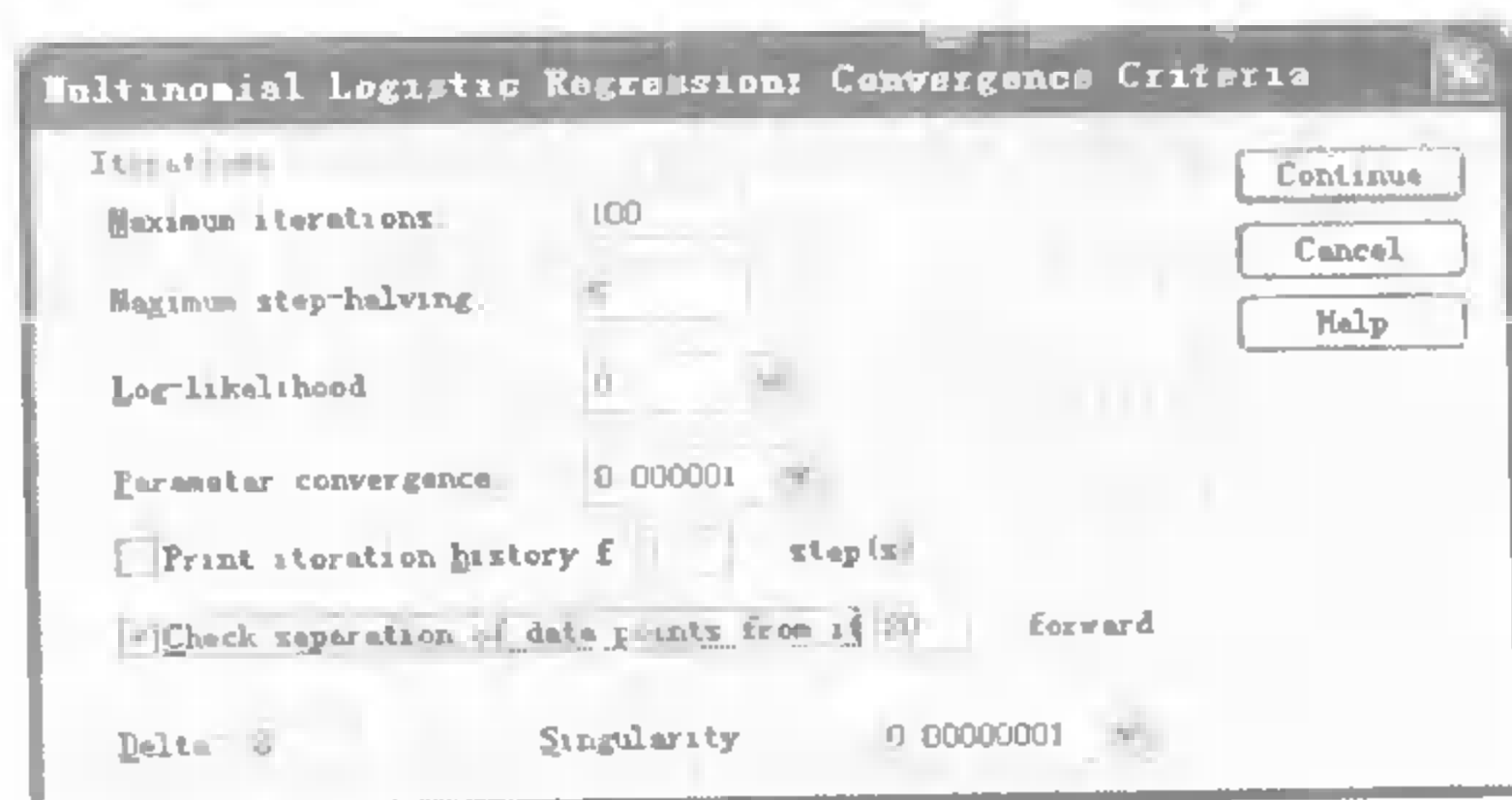


图 8-59 多元 Logistic 分析的收敛标准设置

指定关于对数似然比的收敛依据，如果在逐步回归过程中，对数似然比函数的绝对变化值小于此值，则迭代终止。默认值为 0，表示不使用此准则。

- Parameter convergence 下拉列表，指定关于参数的收敛依据，如果在逐步回归过程中参数估计的绝对变化值小于此值时，迭代终止；设置为 0 时，表示不使用此准则。
- Print iteration history for every n step(s) 复选框，迭代时每隔  $n$  步输出一次， $n$  为输出间隔，在 step(s) 前的输入框指定。
- Check separation of data points from it n forward 复选框，设置开始检查数据被完全分割的起始迭代步骤，在 forward 前的输入框指定起始步骤数。

(2) Delta 输入框。指定一个小于 1 的正数，此值将被添入分类变量交叉表的空单元格中，这有助于稳定算法，防止出现较大的估计偏差。

(3) Singularity tolerance 输入框，指定检验奇异性的容许值。

## 5. Option 选项设置

在图 8-55 中单击 Option 按钮，弹出如图 8-60 所示的对话框，在此设置关于逐步回归的参数。单击 Continue 按钮返回主界面。

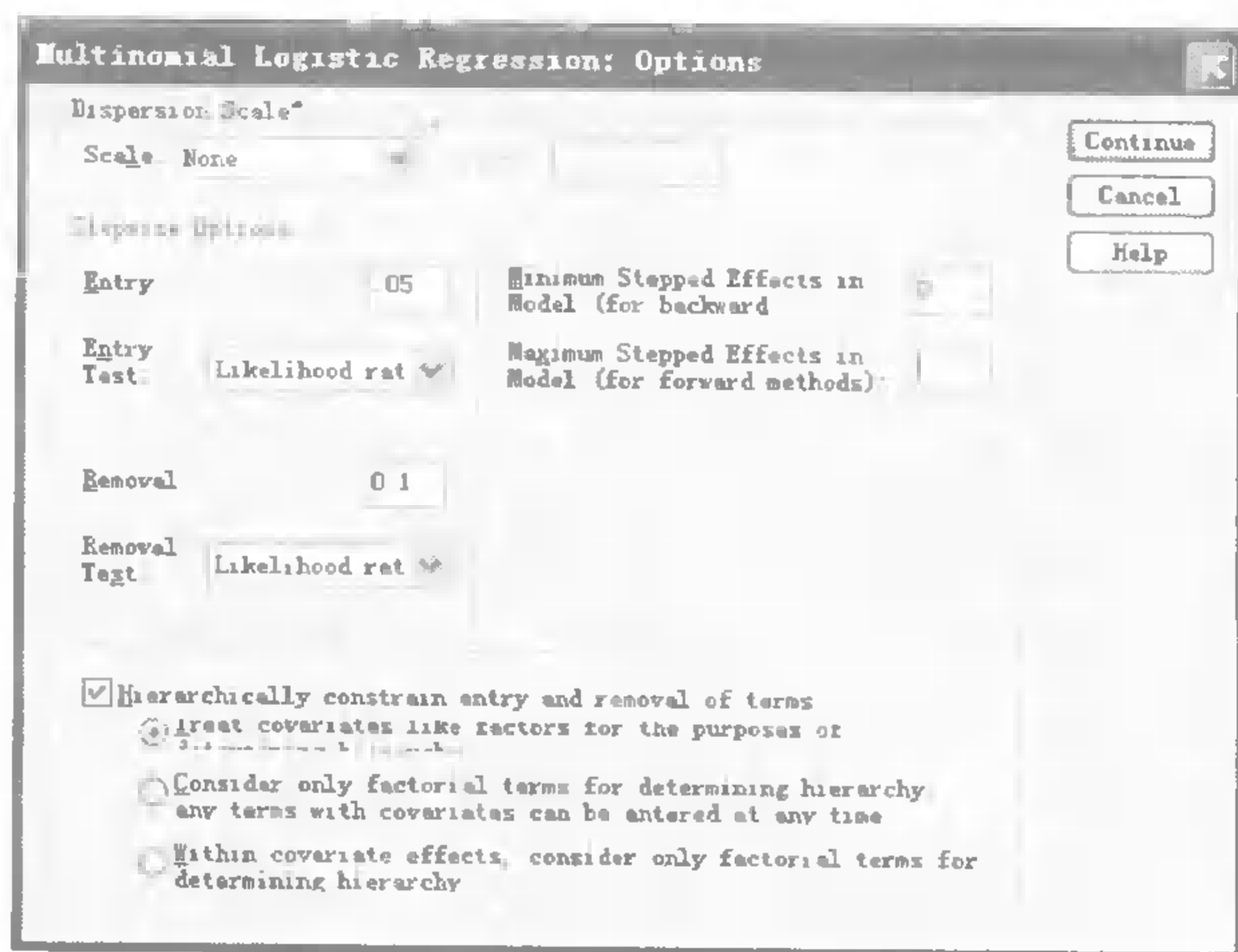


图 8-60 多元 Logistic 分析的 Option 设置

(1) Dispersion Scale 子设置栏，指定离散尺度，用于校正参数协方差阵的估计值。

Scale 下拉列表给出了如下一些尺度参数：None，不指定；Use-defined，用户指定；Pearson，Pearson 卡方统计量；Deviance，偏差函数统计量，即最大似然比卡方统计量。如果选择了 Use-defined、Pearson 或 Deviance 中的一个，还需在 Value 后的输入框指定适合的数值。

(2) Stepwise Options 子设置栏，设置逐步回归模型的有关判别准则，具体设置内容如下。

- Entry Probability 引入概率，指定将变量引入模型的检验统计量的概率临界值，此概率越大，越容易将变量引入模型，用于向前选择法、向前逐步法和向后逐步法。
- Entry test 引入检验，指定在逐步法中引入变量的检验方法，可选项有 Likelihood-ratio（似然比检验）和 Score（得分检验），用于向前选择法、向前逐步法和向后逐步法。
- Removal Probability 剔除概率，指定将变量从模型剔除的检验统计量的概率临界值，此概率越大越容易将变量保留在模型中，用于向后消去法、向前逐步法和向后逐步法。
- Removal test 剔除检验，指定在逐步法中剔除变量的检验方法，可选项有



Likelihood-ratio (似然比检验) 和 Wald (Wald 检验), 用于向后消去法、向前逐步法和向后逐步法。

- Minimum Stepped Effect in Model 模型的最小逐步效应, 当使用向后消去法或向后逐步法时, 在此指定模型所要包含的最小项目数 (截距项不计数)。
- Maximum Stepped Effect in Model 模型的最大逐步效应, 当使先前选择法或向前逐步法时, 在此指定模型所要包含的最大项目数 (截距项不计数)。

(3) Hierarchically constrain entry and removal of terms 复选框, 设置对模型对效应项的限制条件, 勾选它后, 如果模型要加入一个高阶效应, 则模型必须先包含组成这个高阶效应的主效应。例如: 模型要想加入  $x_1 * x_2$  效应, 必须先包含  $x_1$ 、 $x_2$  这两个主效应。下面的 3 个选项, 指定协变量 (因素变量) 在这种限制中的作用方式。

- Treat covariates like factors for the purpose of determining hierarchy, 对协变量和因素变量都采用这样的限制方式。
- Consider only factorial terms for determining hierarchy; any terms with covariates can be entered at any time, 只对因素变量加此限制, 包含协变量的效应可随意进入模型。
- Within covariate effects, consider only factorial terms for determining hierarchy, 加入包含协变量的高阶效应时, 只要求其中的因素变量效应必须先存在模型中。

## 6. 保存设置

在图 8-55 中单击 Save 按钮, 弹出如图 8-61 所示的保存设置对话框。单击 Continue 按钮返回主界面。

(1) Saved variables 子设置栏, 选择需要保存到数据集中的变量, 可选项有如下 4 个。

- Estimated response probabilities 估计响应概率, 把观测记录按响应变量进行分类的估计概率, 响应变量有几个水平就将保存几个变量, 但最多只可保存 25 个。
- Predicted category 预测分类, 保存模型的预测响应分类。
- Predicted category probabilities 预测分类概率, 保存最大的估计响应概率。
- Actual category probability 实际分类概率, 保存预测正确时的估计响应概率。

(2) Export model information to XML file 子设置栏, 设置将模型信息输出到 XML 格式文件的选项, 保存结果可以直接用于 SmartScore 和 SPSS Server, 单击 Browse 按钮指定文件名称及路径。勾选 Include the covariance matrix 复选框, 表示保存协方差阵于如上的 XML 文件中。



图 8-61 多元 Logistic 分析的保存设置

## 8.5.4 案例的结果分析

在图 8-55 中, 单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 8-62~图 8-66 所示。

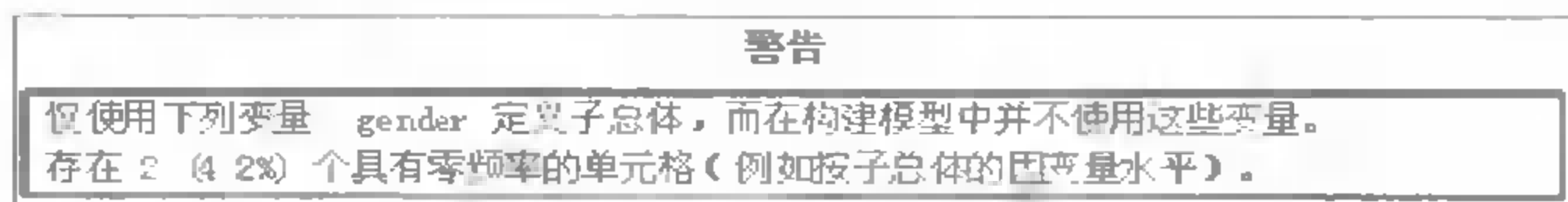


图 8-62 警告信息

案例处理摘要			
		N	边际百分比
早餐	不吃	231	26.3%
	麦片	310	35.2%
	谷类	339	38.5%
年龄段	低于31	181	20.6%
	31-45	208	23.4%
	46-60	231	26.3%
	高于60	262	29.8%
生活方式	消极	474	53.9%
	积极	406	46.1%
性别	男	424	48.2%
	女	456	51.8%
有效		880	100.0%
缺失		0	
总计		880	
子总体		16	

模型拟合信息				
模型	模型拟合标准	似然比检验		
	-2倍对数似然值	卡方	df	显著水平
模型	111.657			
模型	155.915	37.712	3	.000

拟合优度			
	卡方	df	显著水平
Pearson	18.075	22	.641
偏差	21.821	22	.412

伪 R 方	
Cox and Snell	.348
Nagelkerke	.591
McFadden	.197

图 8-63 案例处理摘要和拟合优度检验

似然比检验				
效应	模型拟合标准	似然比检验		
	简化后的模型的 -2 倍对数似然值	卡方	df	显著水平
截距	135.915 <sup>a</sup>	.000	0	
agecat	451.066	315.151	6	.000
active	160.949	25.034	2	.000

卡方统计量是最终模型与简化后模型之间在 -2 倍对数似然值中的差值。通过从最终模型中省略效应而形成简化后的模型。零假设就是该效应的所有参数均为 0。

<sup>a</sup> 因为省略效应不会增加自由度，所以此简化后的模型等同于最终模型。

图 8-64 似然比检验结果

参数估计									
早餐 <sup>a</sup>		B	标准误差	Wald	df	显著水平	Exp(B)	Exp(B) 的置信区间 95%	
								下限	上限
不吃	截距	- 744	287	6 707	1	.010			
	[agecat=1]	938	313	8 989	1	.003	2 555	1 364	4 719
	[agecat=2]	1 047	311	11 333	1	.001	2 848	1 549	5 239
	[agecat=3]	263	332	629	1	.428	1 301	.679	2 494
	[agecat=4]	0 <sup>b</sup>			0				
	[active=0]	- 786	181	18 945	1	.000	.456	.320	.649
	[active=1]	0 <sup>b</sup>			0				
麦片	截距	1 022	195	27 478	1	.000			
	[agecat=1]	-4 256	533	63 770	1	.000	.014	.005	.040
	[agecat=2]	-2 461	275	80 174	1	.000	.065	.050	.146
	[agecat=3]	-1 115	208	28 727	1	.000	.328	.218	.493
	[agecat=4]	0 <sup>b</sup>			0				
	[active=0]	.178	.187	.902	1	.342	1.195	.828	1 724
	[active=1]	0 <sup>b</sup>			0				

a 参考类别是 谷类。

b 因为此参数冗余，所以将其设为零。

图 8-65 参数估计结果

分类				
观察值	预测值			百分比校正
	不吃	麦片	谷类	
不吃	116	34	79	51.1%
麦片	14	251	45	61.0%
谷类	98	116	127	37.5%
总百分比	25.9%	45.6%	28.5%	56.4%

观察频率和预测频率								
性别	生活方式	年龄段	早餐	频率			百分比	
				观察值	预测值	Pearson 残差	观察值	预测值
男	消极	低于31	不吃	12	11.063	.348	37.5%	24.6%
			麦片	0	.941	-.985	0%	3.9%
			谷类	20	19.996	.001	61.5%	71.5%
		31-45	不吃	16	13.955	.668	37.2%	32.5%
			麦片	6	6.416	-.178	14.0%	14.9%
			谷类	21	22.829	-.497	46.6%	52.6%
		46-60	不吃	8	7.724	.106	12.3%	11.9%
			麦片	28	29.857	-.462	43.1%	45.9%
			谷类	29	27.418	.397	44.6%	42.2%
	高于60		不吃	1	4.581	-1.714	1.0%	4.6%
			麦片	71	70.262	.170	74.0%	73.2%
			谷类					

图 8-66 预测分类的结果表格

(1) 警告信息。如图 8-62 所示,“警告”表格给出了如下一些提示信息:首先,变量 gender 只用于定义子总体,而没有包含在模型中;其次,有两个零频率的单元格,由于拟合优度统计量在大样本的假设下才有效,出现零频率单元格可能导致相关统计量失效,此处只有 4.2% 的零频率单元格,因此可以认为拟合优度统计量有效。

(2) 案例处理摘要和拟合优度检验。如图 8-63 所示,“案例处理摘要”表格给出了分类变量各水平下的案例数和边际百分比,以及有效案例和缺失案例的个数统计。

“模型拟合信息”表格给出了最终模型和模型中只包含截距项(其他参数系数全为 0)时的似然比检验结果,此处卡方统计量就是前面的两个 -2 倍对数似然值的差,卡方检验的 Sig 值远小于 0.01,说明最终模型要优于只含截距的模型,即最终模型显著成立。

“拟合优度”表格,检验的零假设是模型能很好的拟合原始数据,从表中的 Pearson 统计量和偏差统计量的 Sig 都大于 0.1 来看,不能否定零假设,即模型的拟合效果还是挺好的。

(3) 似然比检验结果。如图 8-64 所示,“似然比检验”表格给出了最终模型中每个效应(包括截距、年龄和生活方式)的似然比检验结果,零假设为某效应从模型中剔除后系数没有变化。由于卡方检验的 Sig 值都远小于 0.01,故不能否定零假设,即 3 个效应对系数的影响都是显著的,不能被剔除。

(4) 参数估计结果。如图 8-65 所示,参考类的早餐类型为谷类。右数第 4 列为 Wald 检验的显著性水平,若此值小于 0.05,那么对应因素的系数估计显著地不为 0;可见,不吃早餐一栏的 agecat=3 和吃麦片一类的 active=0 两个水平的 Wald 检验 Sig 值都大于 0.10,说明这两个因素对模型的贡献无显著意义;另外,冗余因素参数被设为 0,图中以 0<sub>b</sub> 表示。

此处各估计值的解释同二元 Logistic 回归时的情形类似,如果某个因素的系数估计(B)显著地为正,则在其它因素不变的情况下,取此因素水平的调查者属于当前类别的概率要比属于参考类别的概率大;反之亦然。例如:对不吃早餐一栏的因素水平 agecat=1, B 的值为正,这表示低于 31 岁的人早餐不吃饭的概率要比吃谷类的概率大。

对于 Logistic 回归,Exp(B)一列的信息更加容易解释,例如:对吃麦片一类的因素水平 active=0, Exp(B)的值为 1.195,这说明相对于早餐吃谷类而言,生活方式积极的人早餐吃麦片的发生比 Odds 是生活方式消极的人早餐吃麦片的 Odds 比的 1.195 倍。

(5) 分类结果。如图 8-66 所示,“分类”表格是基于观测频率和预测频率统计而得的。表中对角线上的单元格代表判断正确的个数或概率,非对角线处的单元格为判错的个数或概率。以第 1 行统计数字为例,初始观测中 231 个不吃早餐的人,经过预测有 118 人被分为不吃早餐,即有 51.1% 的判断准确率;同样,在 310 个吃麦片的人中有 251 个判断正确;339 个选择吃谷类的人中有 127 个判断正确。对总体样本判断正确的概率为 56.4%,由此可见模型还有改进的余地。

“观察频率与预测频率”表格是截取对低于 31 岁且生活方式消极的男性的预测统计信息,它比汇总的分类表格更加细致。以第 1 行统计数字为例,表示此类人不吃早餐的原始观测有 12 例,预测他们不吃早餐的观测有 11.063 例,随后是其 Pearson 残差统计量,以及这部分人里不吃早餐的观测值、预测值所占的比例。

## 8.6 Ordinal 回归

如果因变量是有序的分类变量,就可以应用有序回归(Ordinal Regression)分析方法,

又称之为等级回归分析。有序回归可用于对药物疗效的分析，所谓疗效可以分为无效、缓解、好转、治愈 4 个等级，其中缓解与好转是病人的主观体验，难以测量与量化，而用有序回归就可以分析这样的有序变量。

SPSS 的 Ordinal Regression 过程输出的统计量与图形包括协变量中每个分类变量的观测频数、预测频数、累计频数、频数与累计频数的 Pearson 残差、观察概率与预测概率和累积概率；还有参数估计值的渐近相关矩阵与协方差矩阵、Pearson 卡方统计量、似然比卡方统计量、拟合优度统计量、迭代历史、参数估计值、标准误和 Cox&Snell R 方统计量等。

### 8.6.1 问题描述和数据准备

#### 1. 数据描述

本节通过有序回归过程来分析债权人如何确定申请者信用风险的问题，数据均摘自 SPSS 自带的 Demo 文件“german\_credit.sav”，所用数据文件为“信用评价数据.sav”，数据格式如图 8-67 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	age	Numeric	11	2	年龄	None	None	8	Right	Scale
2	chist	Numeric	8	2	帐目情况	{1.00, 没有贷款历史	None	8	Right	Nominal
3	numcred	Numeric	11	2	现有存款	None	None	8	Right	Ordinal
4	othnstal	Numeric	8	2	其他贷款	{1.00, 银行}...	None	8	Right	Nominal
5	housing	Numeric	8	2	住房供给	{1.00, 租房}...	None	8	Right	Nominal
6	duration	Numeric	11	2	持续时间	None	None	8	Right	Scale

图 8-67 信用评价数据结构表

因变量 chist（账目情况）为有序分类变量（ordinal），共有 5 个取值水平：没有贷款历史、现在没有贷款、正在偿还、逾期偿还和拖欠贷款，分别赋值 1~5。另外，还给出了申请者多方面的金融及个人特征，包括现有存款、其他贷款和年龄等。

#### 2. 初步分析

依次单击菜单“Graphs→Chart Builder...”打开图形构建器界面，在 Choose from 列表中选择做 Histograms（简单直方图），并以账目情况作为横轴，默认的 count（计数）作为纵轴作图，SPSS Viewer 窗口的输出图形如图 8-68 所示。

关于账目情况的频数直方图可用来直观地描述账目情况的取值分布。可见，第 3 类（正在偿还）和第 5 类（拖欠贷款）人出现的频率最大，并且类别编码越大，拖欠贷款的可能性也越大，所以建议选择 Complementary log-log 连接函数（参考表 8-3），此函数更关注编码较大的类别，也可以选择 Cauchit 连接函数。

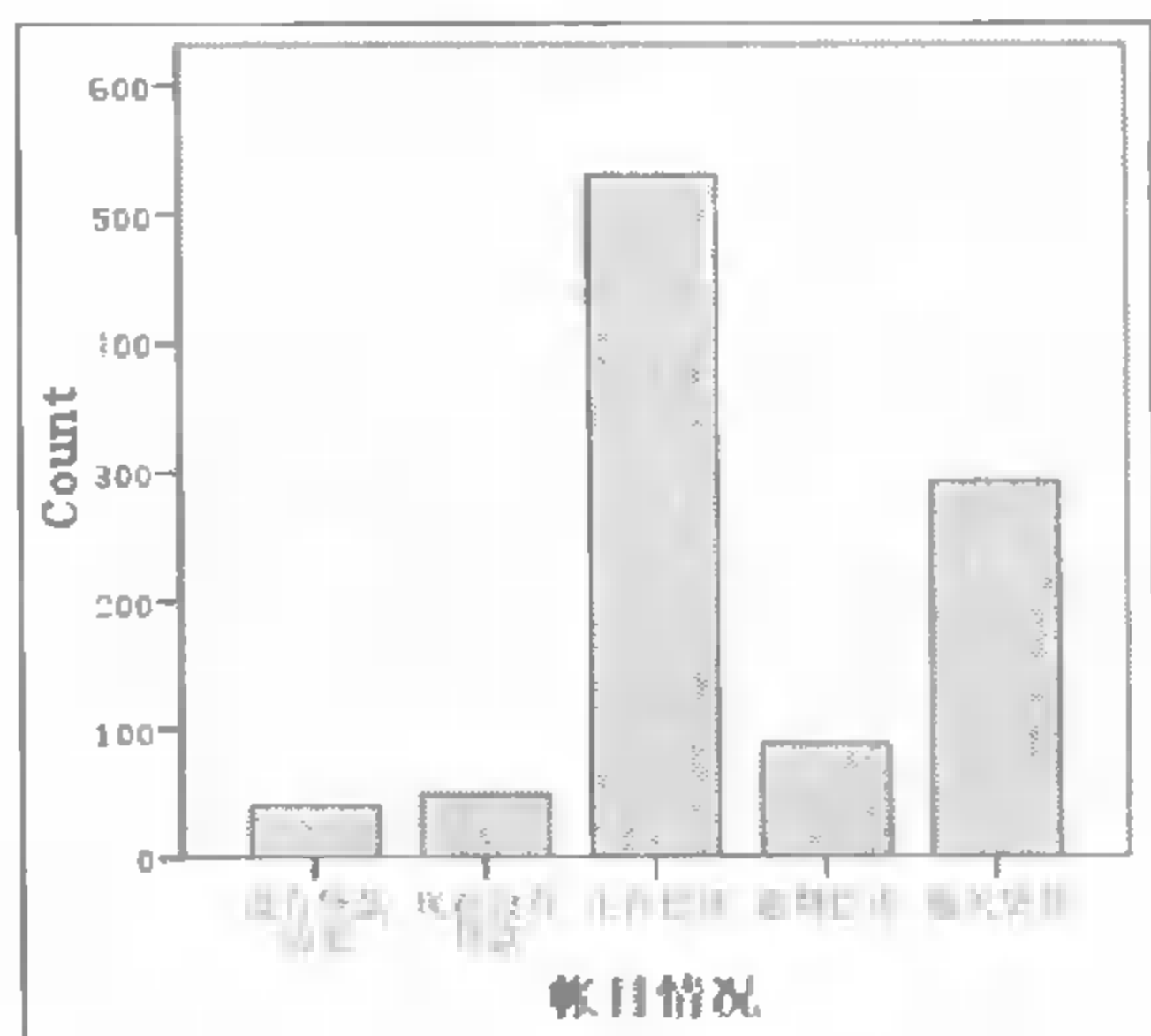


图 8-68 账目情况频率的柱形图

### 8.6.2 Ordinal 回归的参数设置

依次单击菜单“Analyze→Regression→Ordinal...”执行 Ordinal 回归分析的功能，其主设置界面如图 8-69 所示，在此选择进行分析的各种变量。



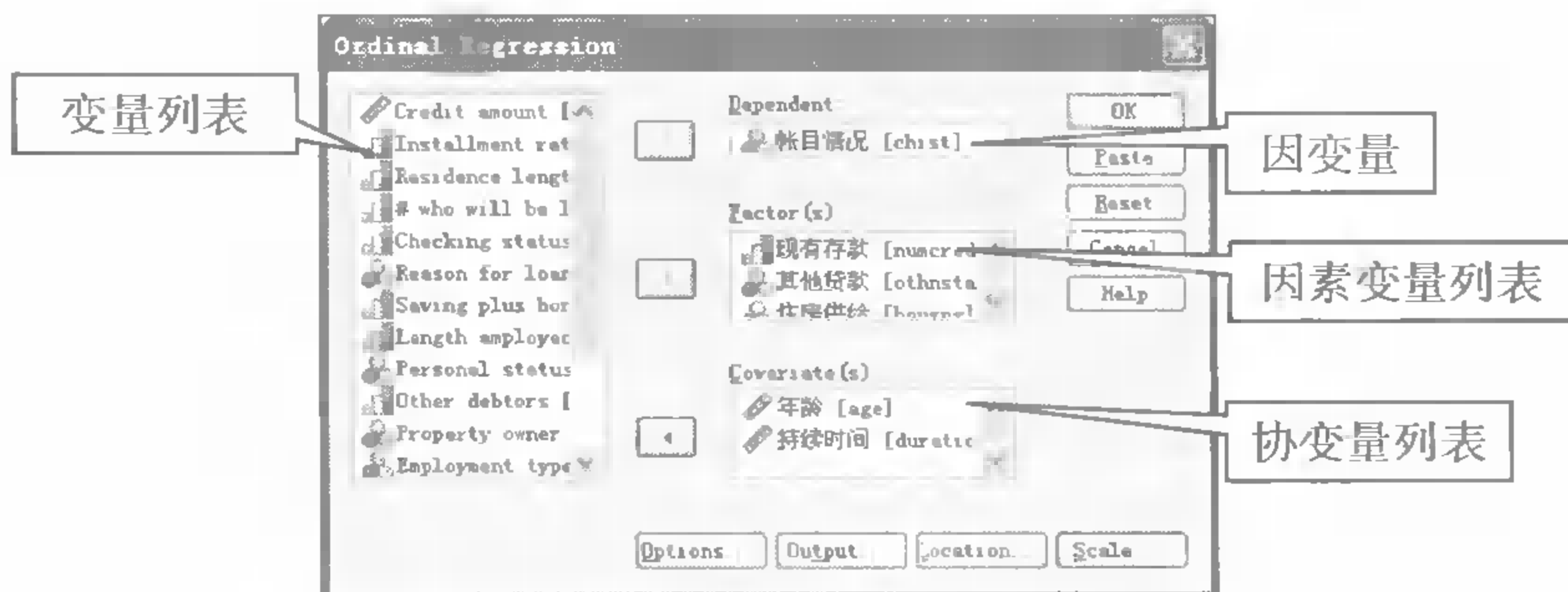





图 8-69 Ordinal 回归的主设置界面

### 1. 变量设置

在变量列表中单击选中账目情况变量，单击从上至下第一个  按钮，将其作为因变量选入 Dependent 选框；在变量列表选中现有存款、其他贷款和住房供给变量，单击从上至下第二个  按钮，将其作为因素变量选入 Factor(s) 列表框；在变量列表选中年龄和持续时间变量，单击从上至下第三个  按钮，将其作为协变量选入 Covariate(s) 列表框。

- Dependent 选框，用于选入一个有序分类变量 (ordinal) 作为因变量，可以是数值型或字符串型的，因变量的取值将自动按照升序排列，最小的值指定为第 1 类。
- Factor(s) 列表框，用于从变量列表选入分类变量作为因素变量。
- Covariate(s) 列表框，用于从变量列表选入数值型变量作为协变量。

### 2. 选项设置

在图 8-69 中单击 Option 按钮，弹出如图 8-70 所示的对话框，用于设置关于迭代参数的选项。单击底部的 link 下拉列表，选中 Complementary log-log 选项；单击 Continue 按钮返回主界面。

(1) Iterations 子设置栏，用于指定迭代终止条件，有如下 4 个选项可以设置。

- Maximum iterations 输入框，指定最大迭代次数；若指定为 0，将只输出初始值。
- Maximum step-halving 输入框，指定最大等分值。
- Log-likelihood convergence 下拉列表，指定关于对数似然比的收敛依据，如果在逐步回归过程中，对数似然比函数的绝对变化值小于此值，则迭代终止。默认为 0，表示不使用此准则。
- Parameter convergence 下拉列表，指定关于参数的收敛依据，如果在逐步回归过程中，每个参数估计的绝对变化值都小于此值时，迭代终止；设为 0 时，表示不使用此准则。

(2) Confidence interval 栏，指定置信区间，范围 0~100，默认值为 95%。

(3) Delta 输入框用于指定一个小于 1 的正数，此值将被添入分类变量交叉表的空单元格中，这有助于稳定算法，防止出现较大的估计偏差。

(4) Singularity tolerance 输入框，指定检验奇异值（因变量的过高预测值）的容许度。

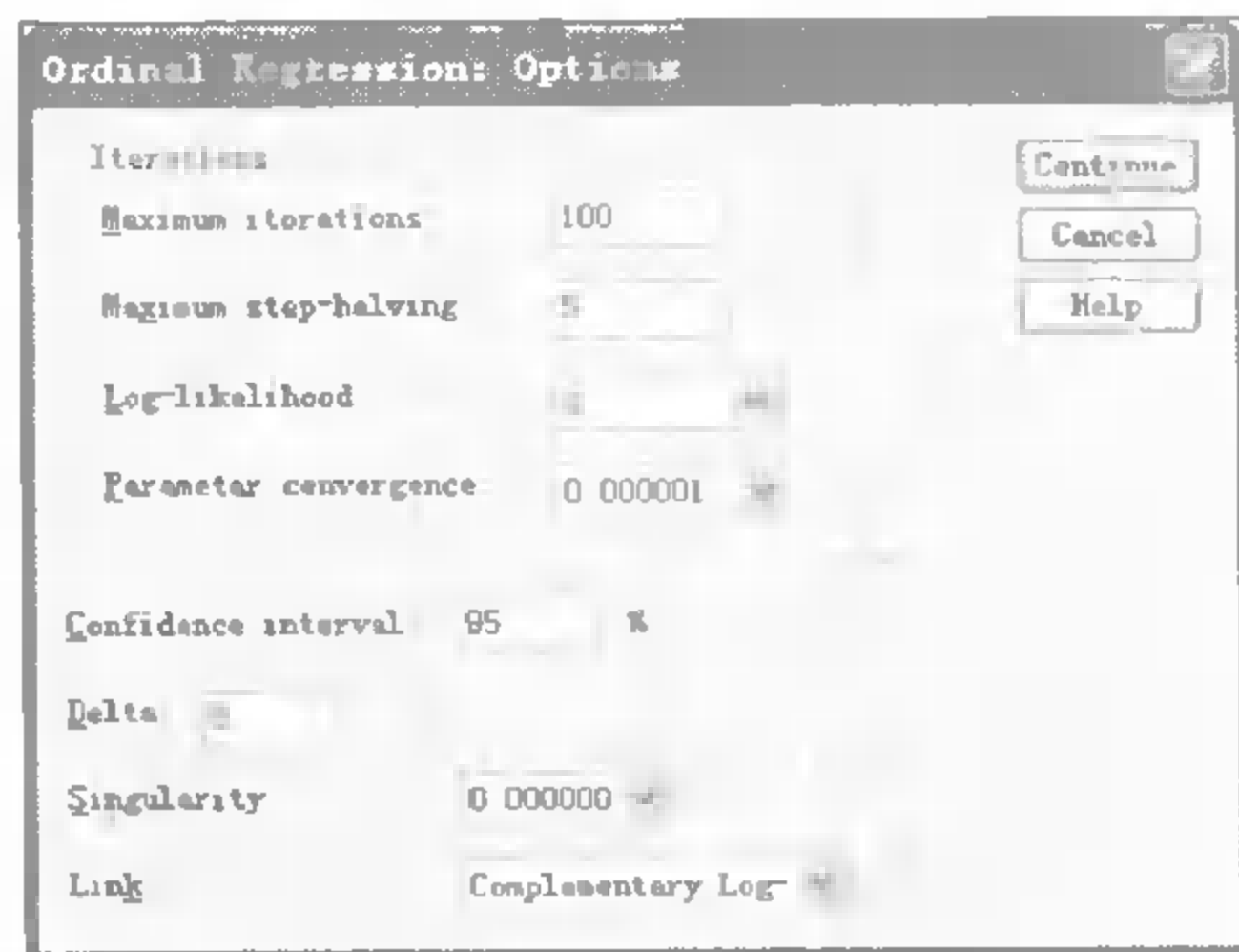


图 8-70 Ordinal 回归的选项设置

(5) Link Function 下拉列表, 指定连接函数, 即对模型估计中的累积概率的转换函数, SPSS 给出了 5 种连接函数, 如表 8-3 所示。

表 8-3

Link Function 连接函数表

函数名称		函数形式	典型应用
Logit	Logit 连接函数	$\log(\xi/(1-\xi))$	因变量接近均匀分布的情况
Complementary log-log	补充对数-对数连接函数	$\log(-\log/(1-\xi))$	因变量取值越大, 概率越大的情况
Negative log-log	负对数-对数连接函数	$-\log(-\log(\xi))$	因变量取值越小, 概率越大的情况
Probit	概率单位连接函数	$\Phi^{-1}(\xi)$	潜在变量为正态分布的情况
Cauchit (inverse Cauchy)	Cauchit 连接函数	$\tan(n(\xi-0.5))$	潜在变量有较多极端值的情况

### 3. 输出设置

在图 8-69 中单击 Output 按钮, 弹出如图 8-71 所示的对话框, 用于选择模型的输出选项。依次勾选 Test of parallel lines 复选框和 Predicted category 复选框; 单击 Continue 按钮返回主界面。

(1) Display 子设置栏, 选择要输出的结果, 可选内容有如下 8 个。

- ☒ Print iteration history for every n step(s), 输出每隔 n 步的迭代记录, n 为在 step(s) 前的输入框指定的数值; 而且总是会输出迭代的第一步和最后一步。
- ☒ Goodness-of-fit statistics 拟合优度统计量, 输出 Pearson 卡方和似然比卡方统计量, 它们的计算都基于分析变量的分类信息。
- ☒ Summary statistics 统计量摘要, 输出 Cox&Snell 卡方、Nagelkerke 卡方和 McFadden 卡方这 3 个统计量。
- ☒ Parameter estimates 参数估计, 输出参数估计值、估计值的标准误和置信区间。
- ☒ Asymptotic correlation of parameter estimates, 输出参数估计值的相关系数矩阵。
- ☒ Asymptotic covariance of parameter estimates, 输出参数估计值的协方差矩阵。
- ☒ Cell information 单元格信息, 输出观察与期望频数、累积频数、频率与累积频率的 Pearson 残差、观测概率和预测概率等内容。
- ☒ Test of parallel lines 平行性检验, 检验参数估计(斜率系数)在各响应类别中是否相同, 此选项仅适用于位置模型(location-only)。

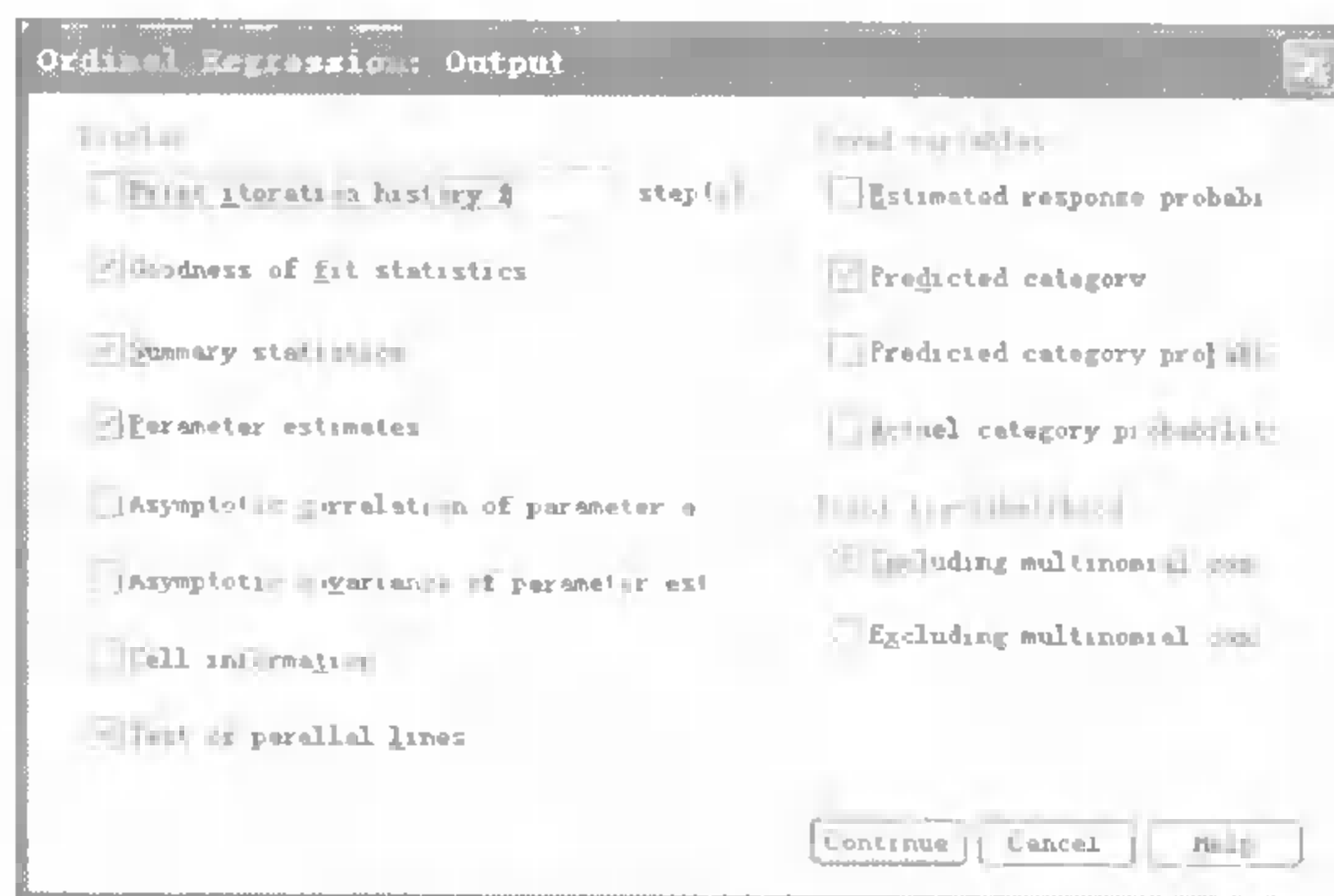


图 8-71 Ordinal 回归的输出设置

(2) Saved variables 子设置栏, 选择要在当前数据集中保存的变量, 可选内容有如下 4 个。

- ☒ Estimated response probabilities 估计响应概率, 把观测记录按响应变量进行分类的估计概率, 响应变量有几个水平就将保存几个变量。
- ☒ Predicted category 预测分类, 保存模型的预测响应分类。
- ☒ Predicted category probabilities 预测分类概率, 保存最大的估计响应概率。
- ☒ Actual category probability 实际分类概率, 保存预测正确时的估计响应概率。

(3) Print log-likelihood 子设置栏, 选择对数似然统计量的输出, 有如下两个可选项。

- Including multinomial constant 选项, 包含常数项, 是默认选项。
- Excluding multinomial constant 选项, 如果要比较不包含常数项的预测结果, 选中此项。

#### 4. 定位模型设置

在图 8-69 中单击 Location 按钮, 弹出如图 8-72 所示的对话框, 用于指定定位模型中的各种效应 (主效应和交叉效应)。单击 Continue 按钮返回主界面。

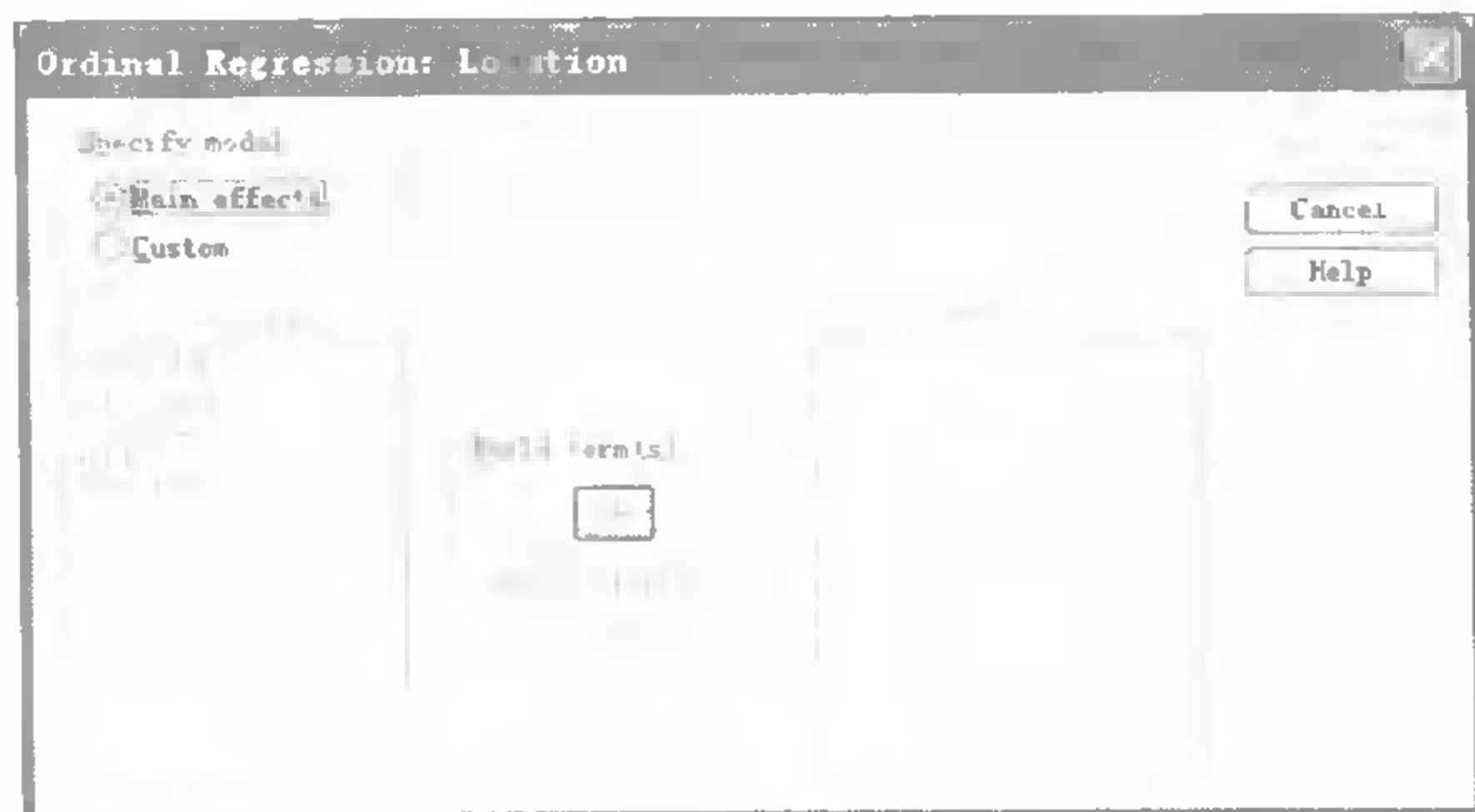


图 8-72 Ordinal 回归的定位模型设置

Specify model 栏用于指定模型类型: Main effects 单选框, 表示采用主效应模型, 只包含协变量和因素变量的主效应, 不包括任何交互效应; Custom 单选框, 表示采用自定义模型, 由用户指定模型分析中的各种效应。选中 Custom 选项, 激活下面的设置内容。

- Factors/covariates 列表框, 显示在主面板选择的因子变量 (F) 和协变量 (C)。
- Location model 列表框, 用于选入在模型分析中要用到的各个效应。
- Build term(s) 下拉列表, 指定效应的种类, 可选项有 Main effects (主效应)、Interaction (交互效应) 和 All n-way (所有 n 维交互项,  $n=2, 3, 4, 5$ )。

选入某个效应的步骤是: 先在 Factors/covariates 列表选中与此效应相关的所有变量, 然后在 Build Term(s) 栏的下拉列表中选择效应类型, 再单击中间的黑色箭头即可把所选变量的指定效应加入 Location model 列表里。

#### 5. 尺度模型设置

在图 8-69 中单击 Scale 按钮, 弹出如图 8-73 所示的对话框, 设置与尺度模型有关的参数, 单击 Continue 按钮返回主界面。

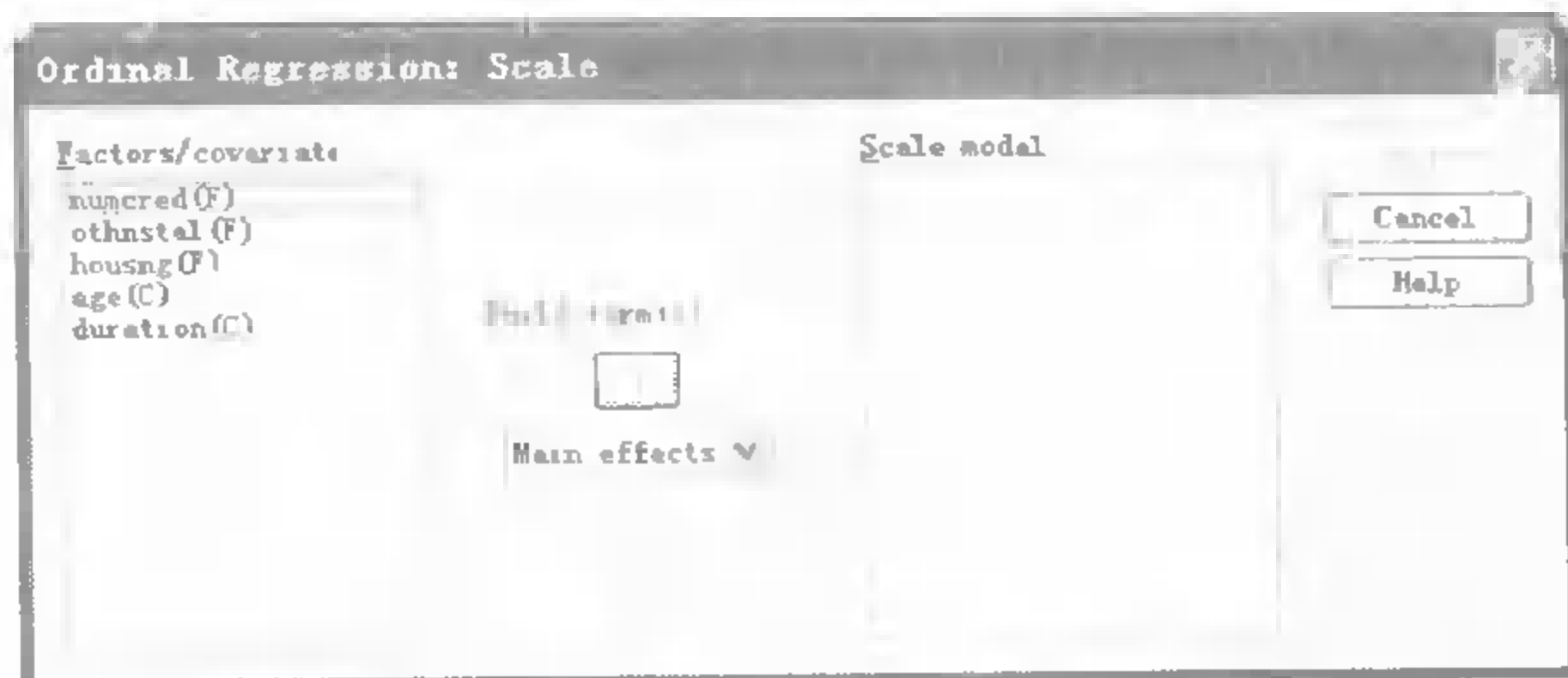


图 8-73 Ordinal 回归的尺度模型设置

此对话框的参数设置与图 8-72 所示的对话框设置方式相似。

### 8.6.3 案例的结果分析

在图 8-69 中,单击 OK 按钮运行,SPSS Viewer 窗口的输出结果如图 8-74~图 8-77 所示。

警告	
有	3078 (78.3%) 个频率为零的单元格 (即通过合并预测变量值构成的因变量水平)。

图 8-74 警告框

案例处理摘要		
帐目情况	N	边际百分比
没有贷款历史	40	1.0%
现在没有贷款	49	1.2%
正在偿还	350	8.7%
逾期偿还	88	2.2%
拖欠贷款	293	7.3%
现有存款	433	10.7%
1.00	183	4.5%
2.00	28	.7%
3.00	6	.0%
其他贷款	159	3.9%
银行股票	17	.4%
无	314	7.7%
住房供给	179	4.4%
租房	113	2.8%
自给	108	2.7%
免费	1000	25.0%
有效缺失合计	0	

模型拟合信息				
模型	-2 对数似然值	卡方	df	显著性
仅截距	2239.858			
最终	1896.552	3513.386	9	.000

联接函数: 辅助对数对数。

拟合度			
	卡方	df	显著性
Pearson	4688.715	3131	.000
偏差	1795.915	3131	.000

联接函数: 辅助对数对数。

伪 R 方	
Cox and Snell	.238
Nagelkerke	.328
McFadden	.149

联接函数: 辅助对数对数。

图 8-75 案例处理摘要和模型拟合信息

参数估计值								
		估计	标准误	Wald	df	显著性	95% 置信区间	
							下限	上限
阈值	[chast = 1.00]	-3.549	.687	28.323	1	.000	-4.856	-2.242
	[chast = 2.00]	-2.720	.656	17.167	1	.000	-4.006	-1.433
	[chast = 3.00]	-.137	.649	.044	1	.833	-1.408	1.135
	[chast = 4.00]	.199	.649	.094	1	.759	-1.072	1.471
位置	age	.015	.004	15.128	1	.000	.007	.023
	duration	-.002	.003	.379	1	.538	-.009	.005
	[numcred=1.00]	-1.134	.594	3.645	1	.056	-2.298	.030
	[numcred=2.00]	.367	.596	.376	1	.540	-.805	1.538
	[numcred=3.00]	.981	.711	1.902	1	.168	-.413	2.374
	[numcred=4.00]	0 <sup>a</sup>			0			
	[othnstal=1.00]	-.397	.116	11.369	1	.001	-.627	-.166
	[othnstal=2.00]	-.469	.193	5.913	1	.015	-.848	-.091
	[othnstal=3.00]	0 <sup>a</sup>			0			
	[housng=1.00]	-.082	.165	.249	1	.617	-.406	.241
	[housng=2.00]	.132	.139	.897	1	.344	-.141	.404
	[housng=3.00]	0 <sup>a</sup>			0			

联接函数: 辅助对数-对数。

a. 因为该参数为冗余的, 所以将其置为零。

图 8-76 参数估计表

平行线检验 <sup>c</sup>				
模型	-2 对数似然值	卡方	df	显著性
零假设	1896.552			
广义	1588.614 <sup>a</sup>	307.938 <sup>b</sup>	27	.000

零假设规定位置参数 (斜率系数) 在各响应类别中都是相同的。

a. 在达到最大步骤对分次数后, 无法进一步增加对数似然值。

b. 卡方统计量的计算基于广义模型最后一次迭代得到的对数似然值。检验的有效性是不确定的。

c. 联接函数: 辅助对数对数。

图 8-77 平行线检验结果



(1) 警告信息。图 8-74 所示是程序运行的警告框，提示用户频率为 0 的单元格有 3 000 多个，出现此框的原因是模型中包含了连续变量。例如把观测中拥有如下特征的申请者组合为一个单元格：现在正在偿还贷款、在银行中有存款、拥有住房、没有其他债务、49 岁、申请 12 月的贷款，由于持续时间和年龄都为连续型变量，所以类似这样的单元格多数为空。空单元格较多时会影响统计量的计算和有效性，所以评价此模型时要慎重使用基于卡方检验的拟合优度统计量。

(2) 案例处理摘要和拟和优度检验。如图 8-75 所示，“案例处理摘要”表格给出了分类变量各水平下的案例数和边际百分比，以及有效案例和缺失案例的个数统计。

“模型拟合信息”表格给出了最终模型和模型中只包含截距项（其他参数系数全为 0）时的似然比检验结果，此处卡方统计量就是前面的两个 -2 倍对数似然值的差，卡方检验的 Sig 值远小于 0.01，说明最终模型要优于只含截距的模型，即最终模型显著成立。

“拟合优度”表格，检验的零假设是模型能很好的拟合原始数据。由于 Pearson 统计量和偏差统计量对空单元格都非常敏感，而本例中的两个连续变量又导致大量空单元格的出现，以至于这两个统计量的检验结果不太可信，不建议采纳。

(3) 参数估计结果。如图 8-76 所示，右数第 3 列为 Wald 检验的显著性水平，若此值小于 0.05，则对应因素的系数估计显著地不为 0。对此，本例的大多因素都不够显著，原因可能是因变量的分类顺序不对或者连接函数选择不理想。

由于转换函数的存在，使得对参数估计值的解释变得困难许多。如果一个协变量的参数估计值为正，那么对此变量取值越大的观测目标类别取值也越大，反之亦然；对于因素变量，参数估计值越大的取值水平，预测目标类别的取值也越大。例如：age 变量的 Wald 检验是显著的（Sig<0.01），且参数估计值为正，说明年龄越大，拖欠贷款的概率也越大。

(4) 平行性检验结果。如图 8-77 所示，平行性检验的零假设是位置参数（斜率系数）在各个响应类别中都是相等的，因为显著性值远小于 0.01，所以否定零假设。

(5) 进一步分析。举个例子说明一下如何利用拟和模型进行应用和预测。设某申请者的个人信息为：申请 48 个月的贷款（duration）、22 岁（age）、有银行存款（numcred）、没有其他贷款（othnstal）、拥有住房（housng），下面就利用本节建立的模型来评估他的信誉水平。

把个人信息数据带入模型预测方程中（除了最后一个分类，每个分类都有一个方程），得到的估计值分别为 -2.78、-1.95、0.63 和 0.97；再把这些估计值代入 Complementary log-log 连接函数的逆函数，得出累积概率值 0.06、0.13、0.85 和 0.93（最后一个分类的累积概率为 1.0）；对这几个累积概率求差分，得出对每个类别的预测概率：第 1 类 0.06，第 2 类  $0.13-0.06=0.07$ ，第 3 类  $0.85-0.13=0.72$ ，第 4 类  $0.93-0.85=0.08$ ，第 5 类  $1.0-0.93=0.07$ 。于是，推断此申请者最有可能为第 3 类人（正在偿还），且归为此类的概率为 72%，还可以推断这个申请者将继续偿还贷款，其账户不会出现危机。

## 8.7 概率单位回归分析

概率单位回归（Probit Analysis）是一种用来分析反应比例与刺激强度之间的关系的方法。例如：对于指定数量的病人，分析他们的给药剂量与治愈比例之间的关系。Probit 回归是 SPSS 中的专业统计分析过程。

### 8.7.1 概率单位回归分析简介

概率单位回归的典型应用是分析杀虫剂浓度和杀死害虫数量之间的关系，并据此判断什么样的剂量浓度是最佳的。在药学研究中，此方法常用于半数效应分析（Median effect dose），寻求达到 50% 输出响应的输入刺激量。

#### 1. 数学原理

同 Logistic 回归分析中的 Logit 变换类似，由于线性模型的某些限制，概率单位回归需要把取值分布在实数范围内的变量，通过累积概率函数  $f$  转换成取值分布在  $(0,1)$  区间的概率值，所得概率分布的表达式为  $P_i = f(\alpha + \beta x_i) = f(\varepsilon_i)$ ，常用的累积概率函数有如下两个。

(1) Logit 概率函数  $P_i = F(\varepsilon_i) = F(\alpha + \beta x_i) = \frac{1}{1 + e^{-\varepsilon_i}} = \frac{1}{1 + e^{-(\alpha + \beta x_i)}}$ ，通过变换可以得到等价的另一种形式  $\ln\left(\frac{P_i}{1 - P_i}\right) = \alpha + \beta x_i$ 。

(2) 标准正态累积概率函数  $P_i = F(\varepsilon_i) = \int_{-\infty}^{\varepsilon_i} e^{-s^2/2} dx$ ，式中  $P_i$  代表事件发生的概率， $s$  是服从标准正态分布的随机变量。此概率值是标准正态分布函数曲线从负无穷到  $\varepsilon_i$  之间的面积，所以  $\varepsilon_i$  的值越大，事件就越可能发生。

#### 2. 数据要求

- (1) 响应变量应该是计数信息，记录在指定的自变量条件下，有响应的观测个数。
- (2) 因素变量必须是分类变量，且须用整数编码。
- (3) 观测量应该是独立的，否则卡方检验和拟合优度检验是不适宜的。

#### 3. Probit 回归与 Logistic 回归的关系

Probit 回归与 Logistic 回归十分接近，事实上，当 Probit 过程选择了 Logit 转换时，进行的统计分析过程就是 Logistic 回归。

一般情况下，Probit 回归更适用于从有计划的试验中获得的数据；而 Logistic 回归更适用于直接的观测数据。输出结果上的差异也能说明它们在侧重点上的不同，Probit 回归输出对各种响应比例有效值的估计；而 Logistic 回归输出对自变量的发生比（odds ratios）的估计。

### 8.7.2 问题描述和数据准备

某零售公司在不同的营业场所（网上、货架和店铺）采用了不同的促销价格。本节来对不同促销价格和对促销有反馈的顾客数量之间的关系进行分析，使用概率单位回归方法拟合响应模型。这些数据均摘自 SPSS 自带的 Demo 文件“offer.sav”，所用数据文件为“促销效果评价数据.sav”，数据格式如图 8-78 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	response	Numeric	4	0	响应数	None	None	8	Right	Scale
2	site	Numeric	1	0	促销地点	{1, 网上}	None	4	Right	Scale
3	value	Numeric	3	0	促销价格	None	None	5	Right	Scale
4	nsubj	Numeric	4	0	促销的商品数	None	None	5	Right	Scale
5	stratum	Numeric	2	0	Stratum	None	None	7	Right	Scale

图 8-78 促销效果评价数据的格式

### 8.7.3 概率单位回归的参数设置

依次单击菜单“Analyze→Regression→Probit...”执行 Probit 回归分析的功能，其主设置界面如图 8-79 所示，在此选择分析中的变量。

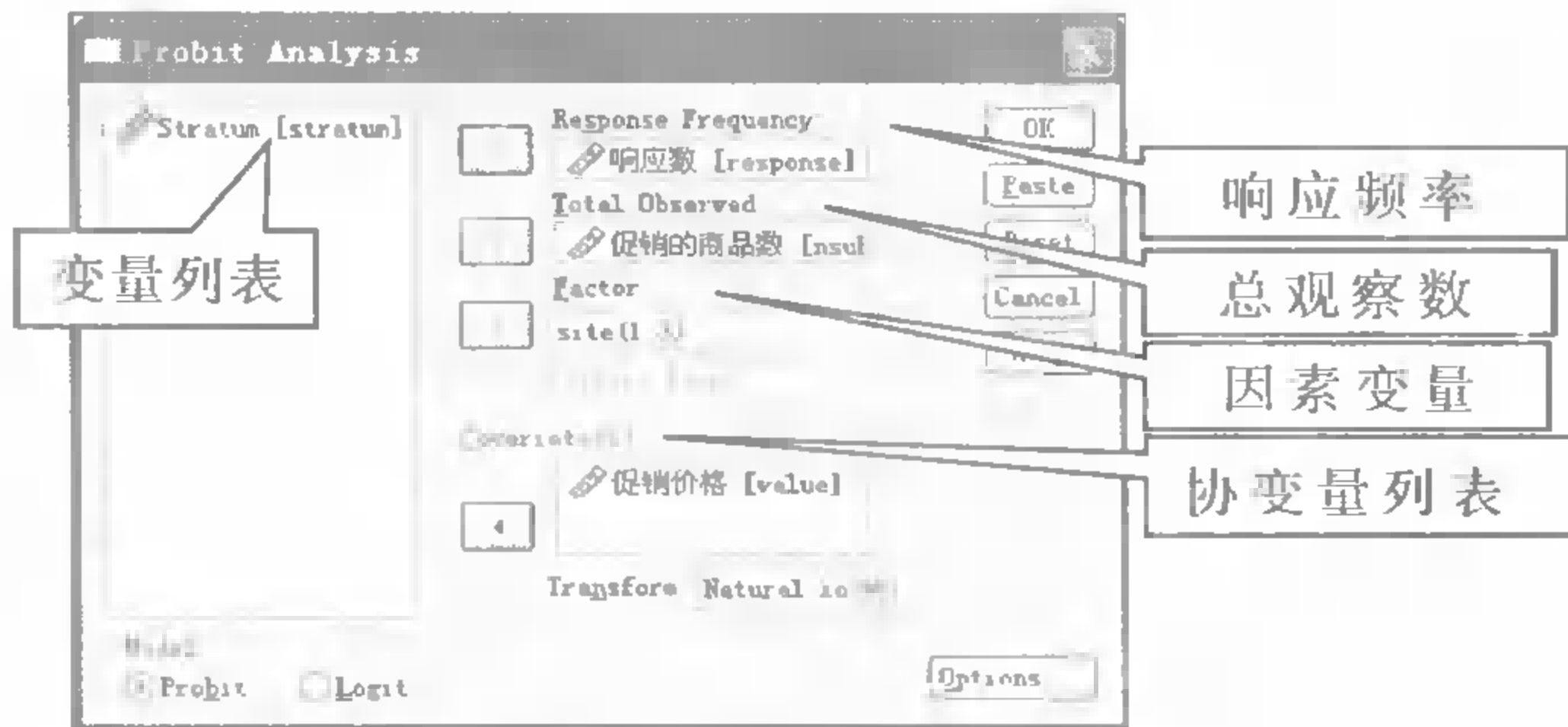






图 8-79 Probit 回归的主设置界面

#### 1. 变量及模型设置

在变量列表中单击选中响应数变量，单击从上至下第一个  按钮，将其作为响应频数变量选入 Response 选框；在变量列表中单击选中促销的商品数变量，单击从上至下第二个  按钮，将其作为总观测变量选入 Total 选框；在变量列表中单击选中促销地点（site）变量，单击从上至下第三个  按钮，将其作为因素变量选入 Factor 选框，单击 Define Range 按钮弹出如图 8-80 所示的对话框，在 Minimum、Maximum 后分别输入“1”和“3”。单击 Continue 按钮返回主界面；在变量列表中单击选中促销价格变量，单击从上至下第四个  按钮，将其作为协变量选入 Covariate 列表框；单击 Transform 下拉列表，选中 Natural log 选项。

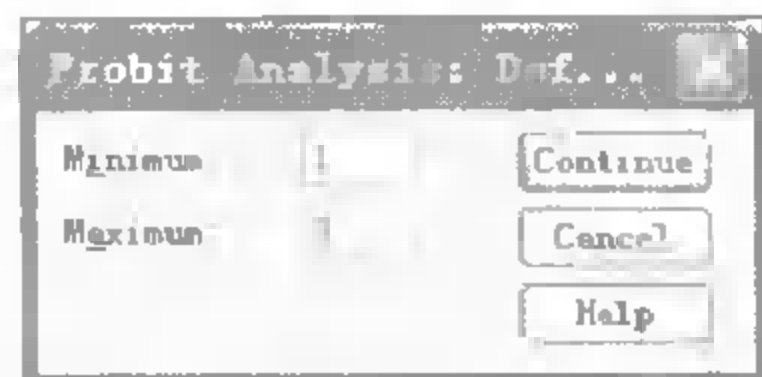


图 8-80 定义取值范围

(1) Response Frequency 栏，用于从变量列表选入一个响应频数变量，它代表在自变量的指定水平下，对有响应的观测的计数信息，取值不能为负。

(2) Total Observed 栏，用于从变量列表选入一个总观测变量，它代表在自变量的指定水平下，总的观测计数，取值不能小于相应的响应频数变量值。

(3) Factor 栏，用于从变量列表选入一个因素变量（必须是整数编码的分类变量）。

选入后单击 Define Range 按钮，弹出如图 8-80 所示的范围定义对话框，在 Minimum、Maximum 输入框分别指定所选因素变量的最小值与最大值。

(4) Covariate(s) 栏，用于指定协变量（自变量），该变量的取值代表不同的试验刺激条件。

(5) Transform 下拉列表，用于设置变量转换函数，当协变量和概率之间不存在线性关系时，需要在此选择对协变量的转换方式：None，不进行转换，系统默认，且自动给出控制组；Log base 10，使用以 10 为底的对数转换；Natural log，使用以 e 为底的自然对数转换。

(6) Model 子设置栏，用于指定一种回归算法，可选项有：Probit 单选框，用标准正态累积概率函数的反函数来转换响应比例；Logit 单选框，对响应比例应用 Logit 转换。

2. 选项设置

在图 8-79 中单击 Option 按钮，弹出如图 8-81 所示的对话框，在此设置回归过程的相关参数。勾选 Parallelism Test 复选框；单击选中 Calculate from Data 单选框；在 Maximum iterations 输入框中键入“100”；单击 Continue 按钮返回主界面。

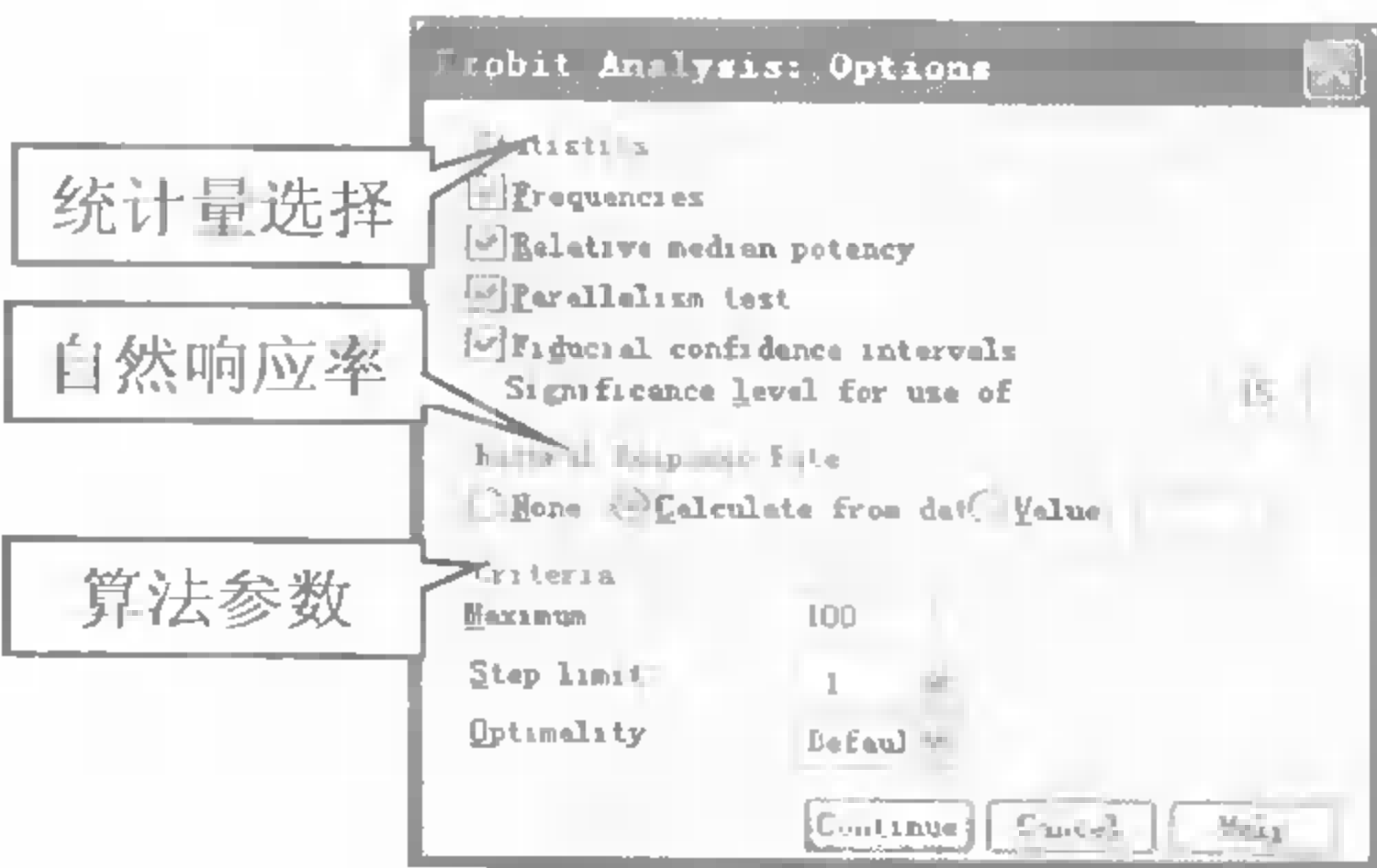


图 8-81 Probit 回归的参数设置

(1) Statistics 子设置栏，用于选择输出哪些统计量，有如下 4 个可选项。

- Frequencies (频数)，输出观测频数、预测频数、观测值的残差等信息。
- Relative Median Potency (相对半数效应)，输出因素变量各水平间的半数效应及其 95% 的置信区间；若模型中没有因素变量，此选项不可用。
- Parallelism Test (平行性检验)，零假设是因素变量各分组的回归方程具有相同的斜率。
- Fiducial Confidence Intervals (基准置信区间)，输出指定响应比例的刺激剂量的置信区间，选中后在 Significance level for use of heterogeneity factor 输入框指定置信水平。

当选入多个协变量时，Fiducial confidence intervals 和 Relative median potency 不可用；只有选入一个因素变量时，Relative median potency 和 Parallelism test 才可用。

(2) Natural Response Rate 子设置栏，指定在没有刺激的情况下（即剂量为 0 时），是否有自然的响应率，有如下 3 个选择。

- None 选项，无自然响应率。
- Calculate from Data 选项，表示从样本数据估计其自然响应率。
- Value 选项，由用户在后边的输入框指定自然响应率，取值必须小于 1。

(3) Criteria 子设置栏，设置与参数估计的迭代算法有关的选项。

- Maximum iterations 输入框，指定最大迭代次数，默认值为 20。
- Step limit 下拉列表，指定单步容许的参数向量的最大变化量。
- Optimality tolerance 下拉列表，指定最优容许度，默认选项为 Default。

8.7.4 案例的结果分析

在图 8-79 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 8-82～图 8-86 所示。

参数估计值							
参数	估计	标准误	z	Sig.	95% 置信区间		
					下限	上限	
PROBIT <sup>a</sup> 促销价格	1.860	.216	8.719	.000	1.457	2.303	
截距 <sup>b</sup> 网上	-7.219	.861	-8.384	.000	-8.081	-6.358	
	货架	-7.641	.888	-8.590	.000	-8.520	-6.743
	店铺	-7.982	.928	-8.601	.000	-8.910	-7.054

<sup>a</sup> PROBIT 模型 PROBIT (p) = 截距 + BX (协变量 X 使用底数为 2.718 的对数来转换。)

<sup>b</sup> 对应于分组变量 site。

图 8-82 参数估计结果



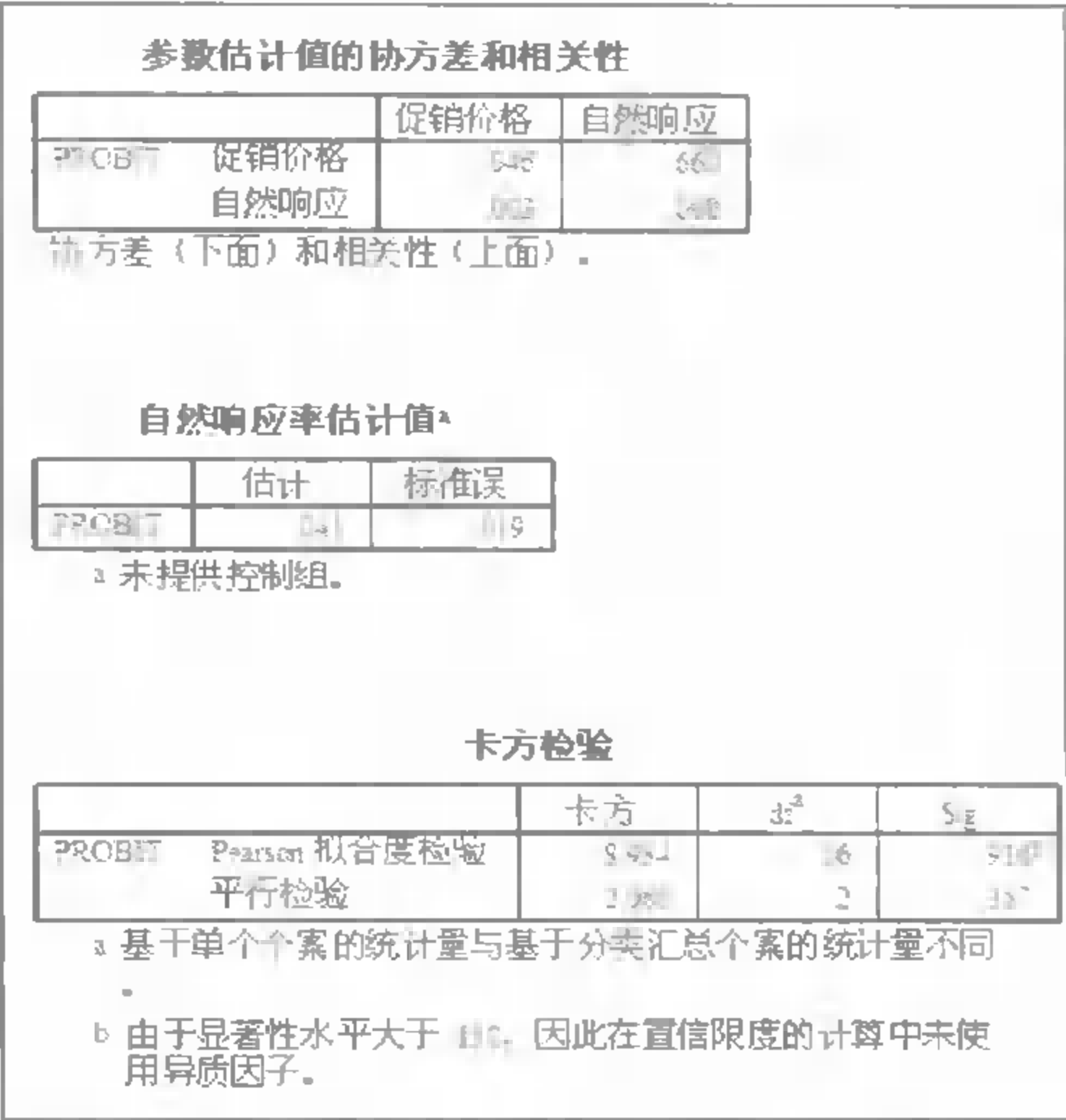


图 8-83 自然响应率估计值和检验信息



图 8-84 置信区间表



图 8-85 与相对众位数强度估计表

(1) 参数估计结果。如图 8-82 所示，所有参数估计的显著性检验 Sig 值都远小于 0.01，说明协变量和三个截距项对方程的作用都有显著意义。由此可得，对 3 种营业场所的 Probit 回归方程如下。

- 网上方程， $\text{Probit}(p) = -7.219 + 1.88 \times (\log_{2.718}(\text{促销价格}))$ 。
- 货架方程， $\text{Probit}(p) = -7.631 + 1.88 \times (\log_{2.718}(\text{促销价格}))$ 。
- 店铺方程， $\text{Probit}(p) = -7.982 + 1.88 \times (\log_{2.718}(\text{促销价格}))$ 。

(2) 自然响应率的估计值和模型检验信息。如图 8-83 所示，“自然响应估计值”表格给出了对自然响应率的估计值，可见在没有促销活动的情况下，总顾客中仍会有 4.1% 的人购买产品。

“卡方检验”表格中，Pearson 拟合优度卡方检验的零假设为模型能够很好的拟合数据。Pearson 检验的显著性  $\text{Sig} > 0.1$ ，故不能否定零假设，即模型对数据的拟合确实较好。平行性检验的显著性值  $\text{Sig} > 0.1$ ，所以认为因素变量各分组的回归方程具有相同的斜率。

(3) 置信区间表。如图 8-84 所示，显示的是指定销售地点的响应概率，在达到各百分位点时的促销价格估计值及其 95% 的置信区间，图中只截取了关于网上促销的部分结果。可见，

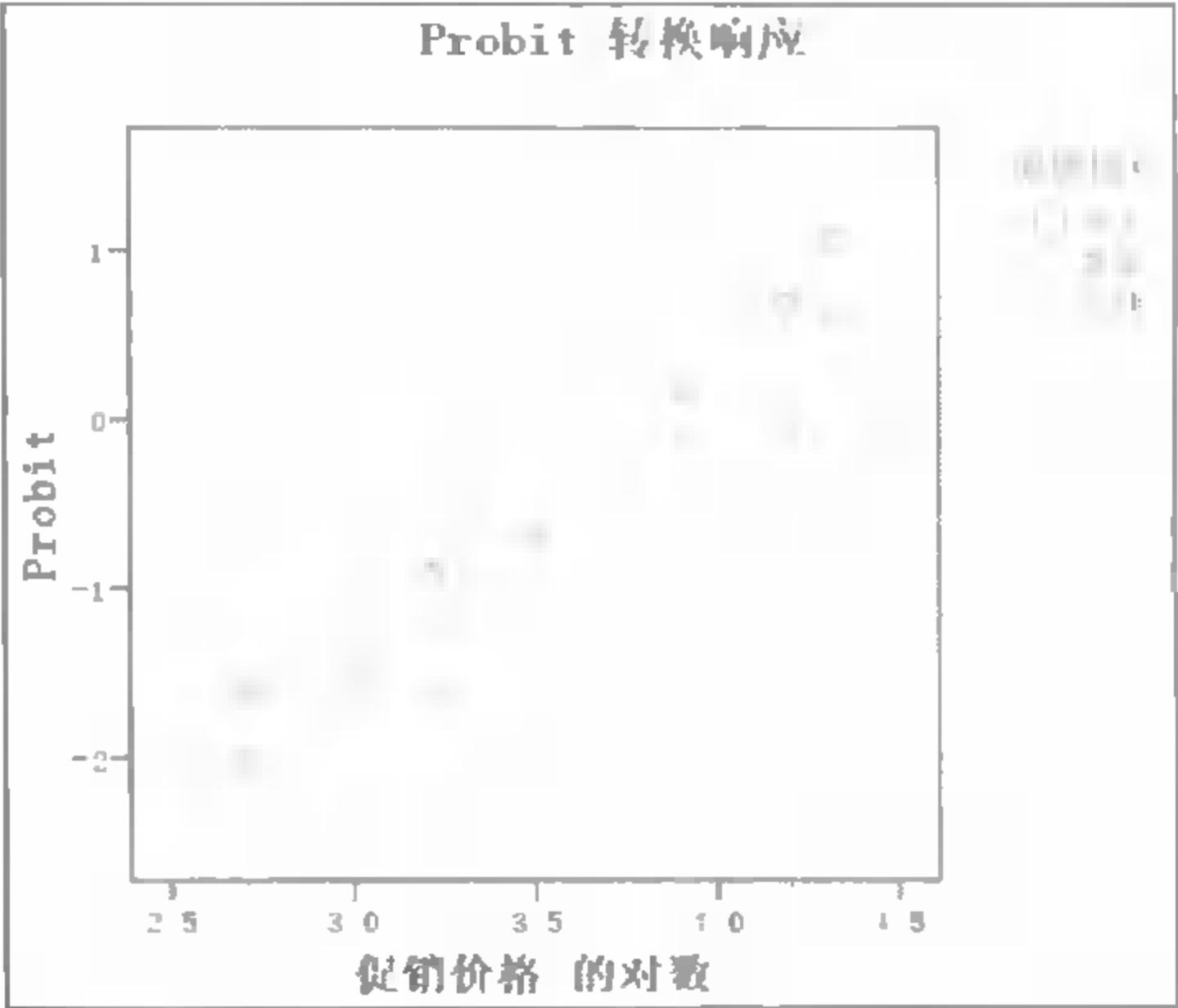


图 8-86 散点图

响应概率  $\text{Probit}=0.500$  时，网上促销价格的半数响应估计值为 46.518；同理可知，货架促销和店铺促销的价格半数响应估计值分别为 57.905 和 69.800。由此也可以得到如下结论：网上促销比货架促销更有效，货架促销比店铺促销更有效。

(4) 相对中位数强度估计值。如图 8-85 所示，这就是常说的相对半数效应表，下面以第一行为例来说明如何读此表。第一行显示的是网上促销（地点取值为 1）对货架促销（地点取值为 2）的相对半数效应，网上与货架半数效应比值的估计值为 0.803，且其 95% 的置信区间（0.660, 0.942）不包括 1，说明二者是有明显区别的，也就是说网上促销更有效，因为它能以较小的促销价格促使 50% 的客户购买产品（即达到 50% 的响应率）。通过此表也可以发现，网上促销比货架促销更有效，货架促销比店铺促销更有效。

(5) 散点图。如图 8-86 所示，是响应概率（Probit）对对数促销价格的散点图。

通过观察可以发现，响应概率与对数促销价格呈明显的线性关系，说明对促销价格所作的自然对数转换是比较合适的，如果散点图没有呈现明显的线性趋势，可以再采用其它转换方法进行分析。另外，网上促销的 Probit 值普遍大于货架促销，货架促销的 Probit 值普遍大于店铺促销，这也反应了网上促销比货架促销更有效，货架促销比店铺促销更有效。

## 8.8 加权回归分析

在线性回归模型中，有一个同共方差性的假设，就是要求所有观测对回归模型的变异具有相同的贡献，以此为基础的方法称之为普通最小二乘法（OLS）。当因某些观测的变异较其他观测大而导致样本的方差不等时，就不能使用 OLS 方法了，如果观测的变异是可以通过其他变量进行预测的，就可以使用加权最小二乘法（WLS）来拟合线性回归模型。WLS 法实际上是在回归过程中按观测量方差的倒数对观测进行加权，这样就会降低具有较大方差的观测记录对计算过程的影响。

### 8.8.1 加权回归分析简介

在研究通货膨胀和失业率对股票价格的影响时，考虑到高市值的股票较低市值的具有更高的变异性（价格波动大），使用 OLS 法便不能很好地反映指定因素对变异性较大的股票的影响，这个时候就需要使用 WLS 方法来解决这个问题。

SPSS 加权回归过程对数据的要求和假设包括：自变量和因变量应该是数值型变量；类似宗教、民族和地区这样的分类变量应该重新编码成二分变量或其他类型的对照（contrast）变量；加权变量必须是与因变量有关的数值型变量；对于自变量的每个取值，对应因变量的取值分布必须是正态的；因变量和每一个自变量的相关关系应该是线性的；所有的观测量之间相互独立；各观测的方差可以不同，但是这些差异可以通过加权变量进行预测。

使用加权最小二乘法时，主要过程分为方差诊断和权重估计两个步骤。

(1) 方差诊断。先利用 OLS 方法对原始数据建立简单线形模型，并绘制其残差对预测值的散点图，如果残差均匀分布在某条与横轴平行的横线附近，说明样本的方差基本相等；反之，如果方差呈现明显的喇叭口形状或其他不规则形状，说明样本方差不相等，有必要进行 WLS 估计。

如果只有一个自变量，可以直接作因变量对自变量的散点图，观察因变量的分布是否均匀，判断方法与残差图相似。

(2) 估计权重。如果认为因变量的方差与其它变量之间存在着相关关系，就可以使用 WLS 方法来估计权重，常用的估计方法有如下两种。

① 利用数据的复制集来估计权重。要使用 WLS 估计回归模型，就需要先计算每一个观测的变异性。一种比较好的方法是具有相同特点或近似特点的数据进行编组（数据的复制集），然后计算因变量在各编组中的方差，并以此方差的倒数作为相应编组中观测的权重。

② 利用变量估计权重。利用方差与其他变量的相关关系估计权重。因变量的方差经常与自变量有关，例如：高市值的股票价格具有较大的方差，具有研究生学历的人员的工资方差要比那些没有获得学位人员的工资方差高出许多。

## 8.8.2 问题描述和数据准备

### 1. 数据描述

某开发商计划利用历史数据预测新建一个商业街的成本，本节就用加权回归来进行分析。数据摘自 SPSS 自带的 Demo 文件“mallcost.sav”，所用数据文件为“建设商业街成本数据.sav”，数据格式如图 8-87 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	sqft	Numeric	8	2	面积	None	None	8	Right	Scale
2	inorout	Numeric	8	2	商业街种类	{.00, 室内}	None	8	Right	Scale
3	yrexp	Numeric	8	2	建筑师从业年数	None	None	8	Right	Scale
4	cost	Numeric	8	2	建筑成本	None	None	8	Right	Scale

图 8-87 建设商业街成本数据格式

本例中因变量为建筑成本，其它变量为自变量，权重变量为面积。其中商业街种类为分变量，取值为 0 时表示室内，取值为 1 时表示室外。

### 2. 初步的残差分析

在进行 WLS 分析之前，先来看看利用 OLS 回归对这个问题所作的残差分析图，由此可以判断是否有作 WLS 的必要。分析的操作过程请参考第 8.1.5 节的相关介绍。

如图 8-88 所示，是用 OLS 回归分析后的标准化残差对标准化预测值的散点图。可见，随着预测值的增大，残差也有增大的趋势，故而可以否定 OLS 中关于同方差的假设，建议采用 WLS 方法对这个问题再进行分析。

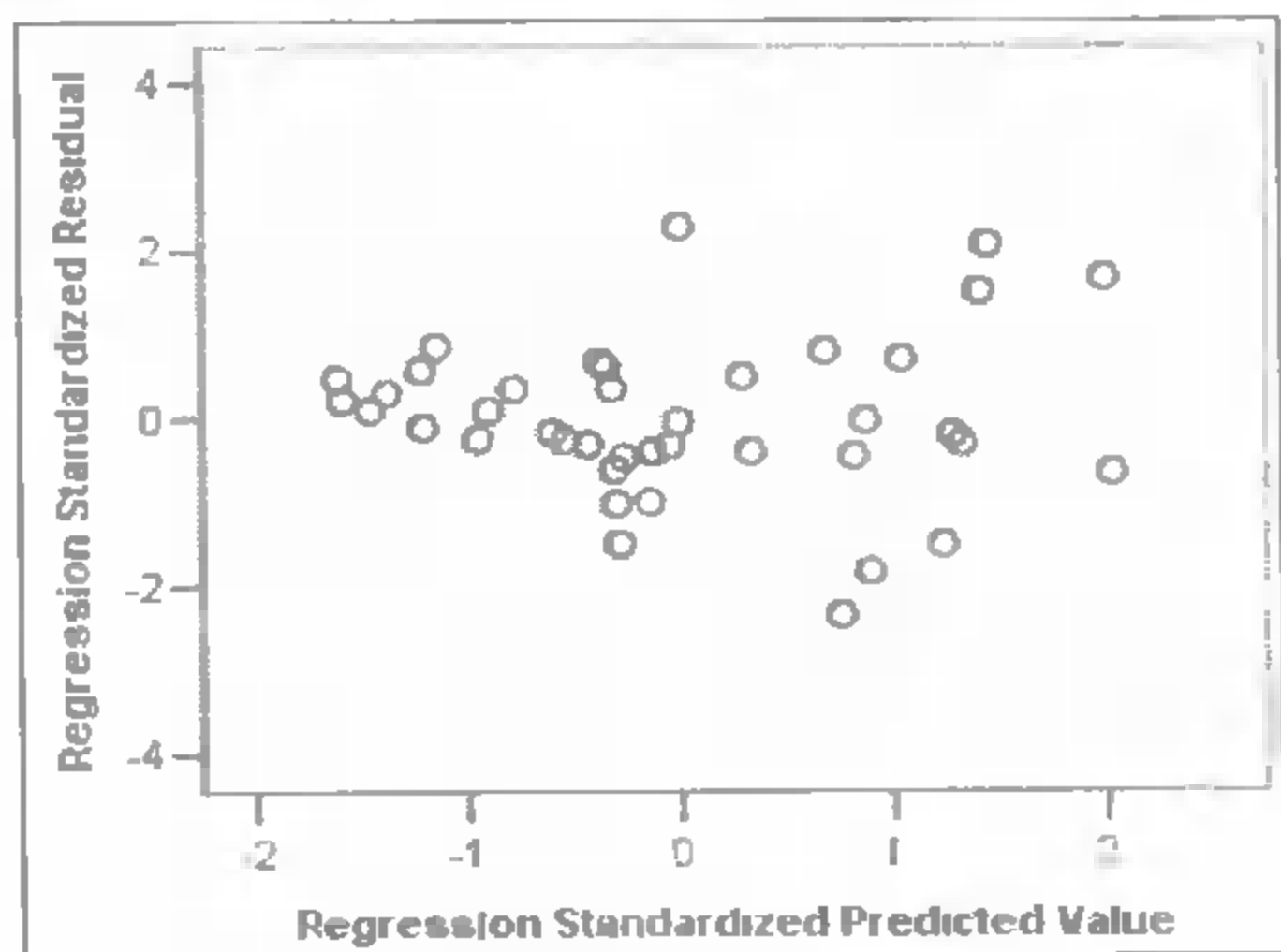


图 8-88 OLS 回归的残差分析图

## 8.8.3 加权回归的参数设置

依次单击菜单“Analyze→Regression→Weight Estimation...”执行加权回归分析的功能，其主设置界面如图 8-89 所示，在此选择进行回归分析的各种变量。

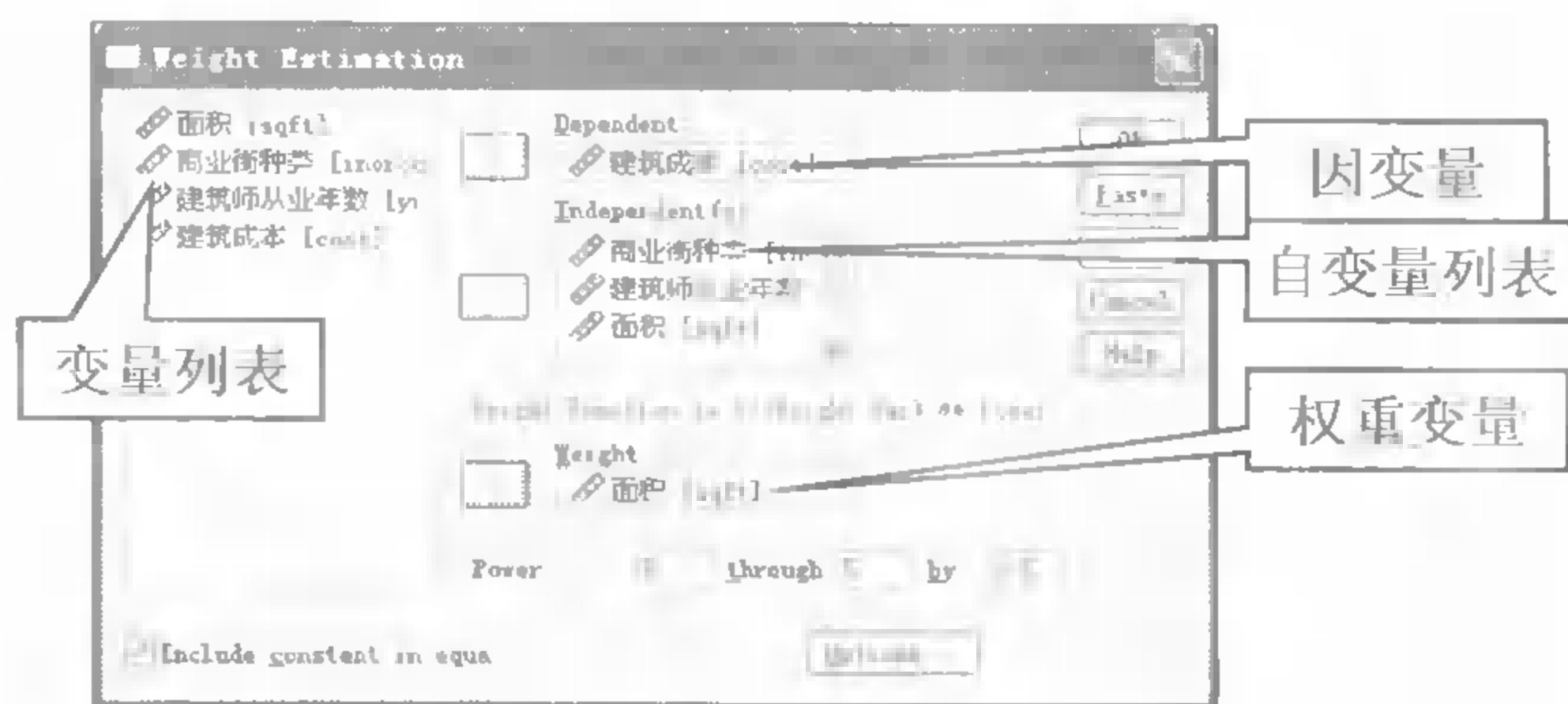


图 8-89 Weight Estimation 过程的主设置界面

## 1. 变量设置

在变量列表中单击选中建筑成本变量，单击从上至下第一个 ☐ 按钮，将其作为因变量选入 Dependent 选框；在变量列表中选中面积、商业街种类和建筑师从业年数变量，单击从上至下第二个 ☐ 按钮，将其作为自变量选入 Independent(s) 列表框；在变量列表中单击选中面积变量，单击从上至下第三个 ☐ 按钮，将其作为权重变量选入 Weight 选框。在 Power 输入框中键入“0”；在 through 输入框中键入“5”；在 by 输入框中键入“0.5”。

(1) Dependent 选框，用于从变量列表选入因变量。

(2) Independent(s) 列表框，用于从变量列表选入自变量。

(3) Weight Function is  $1 / (\text{Weight Var})^{**\text{Power}}$  子设置栏。

Weight 选框，用于从变量列表选入一个加权变量；Power 输入框，指定权重指数的初始值；through 输入框，指定权重指数的结束值；by 输入框，指定权重指数的变化步长。

此处指定的指数范围必须在 -6.5 至 7.5 之间，而且需满足“ $(\text{初始值} - \text{结束值}) / \text{步长} \leq 150$ ”。设置好后，当前的权重变量就是  $1 / (\text{Weight Var})^{\text{Power}}$ ，Power 在指定的范围内取值。

(4) Include constant in equation 复选框，勾选它要求在模型中包括常数项。

## 2. 选项设置

在图 8-89 中单击 Option 按钮，弹出如图 8-90 所示的对话框，在此设置回归过程的其他参数。勾选 Save 复选框；单击 Continue 按钮返回主界面。

(1) Save best weight as new variable 复选框，表示将最佳权重值保存至当前数据集，新变量名为 WGT\_n，n 是用来区分多个类似变量名的数字。

(2) Display ANOVA and Estimates 栏设置方差和估计值的输出形式，可选项有如下两个。

☐ For best power 单选项，只输出最终的方差分析表和指数估计值。

☐ For each power value 单选项，输出在主设置面板指定的指数范围内所有的方差分析表和指数估计值。



图 8-90 Options 选项设置

## 8.8.4 案例的结果分析

在图 8-89 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 8-91～图 8-93 所示。



对数似然值 <sup>b</sup>	
幂	000 -218 675
500	-215 626
1 000	-212 836
1 500	-210 356
2 000	-208 251
2 500	-206 606
3 000	-205 529
3 500	-205 143 <sup>a</sup>
4 000	-205 563
4 500	-206 889
5 000	-209 085

a 选择对应幂以用于进一步分析，因为它可以使对数似然函数最大化。  
b 因变量 cost，源变量 sqft

模型描述	
因变量	cost
自变量	1 2 3
权重	源
幂值	3 500
模型 MOD_1	

模型摘要	
复相关系数	863
R 方	745
调整 R 方	724
估计的标准误	46 730
对数似然函数值	-205 143

图 8-91 权重估计输出

ANOVA					
	平方和	df	均方	F	Sig.
回归	229425.00	3	76476.001	35.022	.000
残差	78612.250	36	2183.674		
总计	308040.25	39			

图 8-92 模型摘要及 ANOVA 方差分析表

系数						
	非标准化系数 <sup>a</sup>		标准化系数		t	Sig.
	B	标准误	Beta	标准误		
(常数)	53.438	16.988			3.146	.003
inorout	-26.533	11.086	-.218	.091	-2.393	.022
sqft	149.273	15.425	.864	.089	9.678	.000
yrexp	-2.209	.941	-.205	.087	-2.343	.024

图 8-93 系数估计值表

(1) 对数似然值和模型摘要。如图 8-91 所示，“对数似然值”表格给出了指定 power 范围内的所有对数似然值，使得这个对数似然值达到最大的指数就为最佳指数，表中以右上角的小 a 表示最优值 (3.5)。

“模型描述”表格给出了加权估计模型的概要信息，包括因变量、自变量、权重变量和最优权重指数。“模型摘要”表格给出采用最佳指数建立的加权回归模型的拟合优度检验结果，可见 R 方和调整 R 方都很大，说明权重指数为 3.5 时的加权回归模型拟合效果不错。

(2) 方差分析表。如图 8-92 所示，是权重指数为 3.5 时所建立的加权回归模型的 ANOVA 方差分析表，从 F 统计量的显著性 Sig 值远小于 0.01 看，由加权回归模型所解释的变异显著地大于由残差所解释的变异，也就是说回归效果很好。

(3) 参数估计结果。如图 8-93 所示，是权重指数为 3.5 时所建立的加权回归模型的参数估计值表。本例各自变量的 t 检验都显著地不为 0 (Sig<0.05)，所以它们对模型的成立都是有显著作用的。最终得到的回归方程为  $\text{cost}=53.438-26.533 \times \text{inorout}-2.209 \times \text{yrexp}+149.273 \times \text{sqft}$ 。

## 8.9 二阶段最小二乘回归

在研究有关时间序列的宏观经济数据时，各分析变量之间存在着较为复杂的内部关系，误差项也就容易与某些预测变量相关。当这种情况发生时，使用普通最小二乘法 (OLS) 所获得的模型就会产生偏差，本节利用二阶段最小二乘法来解决此类问题。

### 8.9.1 二阶段最小二乘回归的基本原理

当某些预测变量与误差项相关联时，二阶段最小二乘法（2SLS）是一种有效的解决方案。在 2SLS 中经常使用如下 3 种类型的变量。

（1）内生变量。在回归分析中随着其他变量的变化而变化的变量和在有反馈作用的情况下具有反馈关系的变量都是内生变量。

（2）工具变量。在回归模型中不受其他变量影响，但是影响其他变量的变量，它们既可以是也可以不是模型中的一部分，与模型中的其它变量不存在因果关系。其特点是对因变量的预测与内生变量拥有相似的效果，但它们与理论误差项是不相关的。在实际操作中要想确定工具变量是否与误差项存在关联是比较困难的，当没有合适的工具变量被引入模型时，具有“滞后”特点的内生变量就会被当成工具变量来使用，虽然其值有“滞后”的特点，但却可能与误差项没有关联。

（3）解释变量指回归方程中的自变量，其范围包括内生变量。当使用二阶段最小二乘法预测因变量时，面临的是内生变量与理论误差项之间存在相关关系的问题。二阶段最小二乘法的第一阶段，就是利用与误差项不存在关联的工具变量，推算可能与误差项存在关联的自变量的值，然后再推算因变量的取值。

SPSS 的 2SLS 回归过程对数据的要求包括：因变量与自变量必须是数值型变量；分类变量必须被重新编码为二分变量或其他类型的对照（contrast）变量；内生自变量必须为连续型变量；对自变量的每个取值，相应因变量取值的分布都必须为正态的；对于自变量的不同取值，因变量分布的方差应该是一个常数；因变量与自变量之间应该呈线性关系。

### 8.9.2 问题描述和数据准备

#### 1. 数据描述

某商品邮寄公司有一个 CD 俱乐部和—个书籍俱乐部，每个月公司都会为俱乐部会员提供一些特殊商品（如家庭用具和普通用具）。此公司想根据会员的书籍购买量、CD 购买量和为会员提供的服务种类来预测他在每个月的特殊商品总购买量。用于购买特殊商品的钱就不能购买书籍和 CD，于是因变量（特殊商品购买量）与解释变量（CD 购买量、书籍购买量）就构成了一种反馈状态，所以本案例适于建立 2SLS 回归模型进行分析。

本节数据摘自 SPSS 自带的 Demo 文件“cross\_sell.sav”，数据文件为“交叉销售数据.sav”，数据格式如图 8-94 所示。此数据记录了 99 个月里，会员在每个月购买如下商品的消费情况：家庭用具商品、普通用具商品、CD 折扣、CD 折扣对数、书折扣和书折扣对数。由于所给的折扣数据与特殊商品的购买是无关的，却影响着 CD 和书籍的购买量，因此建议把 CD 购买量的滞后变量、书籍购买量的滞后变量和两个折扣对数变量都作为工具变量。

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1 buyoff	Numeric	8	2	特殊商品购买量	None	None	10	Right	Scale
2 buycd	Numeric	8	2	CD 购买量	None	None	10	Right	Scale
3 buybk	Numeric	8	2	书籍购买量	None	None	10	Right	Scale
4 offer	Numeric	4	0	特殊商品种类	{1, 家庭用具}	None	4	Right	Nominal
5 discd	Numeric	4	0	CD 的折扣	None	None	5	Right	Scale
6 discbk	Numeric	4	0	书的折扣	None	None	5	Right	Scale
7 offer_type1	Numeric	4	0	家庭用具商品	None	None	11	Right	Nominal
8 offer_type2	Numeric	4	0	普通用具商品	None	None	11	Right	Nominal
9 lndiscd	Numeric	8	2	CD 折扣的对数	None	None	10	Right	Scale
10 lndiscbk	Numeric	8	2	书折扣的对数	None	None	10	Right	Scale

图 8-94 交叉销售数据格式

## 2. 计算 CD 销售量和书籍销售量的滞后变量

依次单击菜单“Transform→Create Time Series...”执行生成时间序列变量的过程，其主界面如图 8-95 所示。在变量列表中选中 CD 购买量（buycd）和书籍购买量（buybk），单击  按钮，将其选入 New 列表框。在 New 列表框中单击选中 buycd\_1，单击 Function 下拉列表，选中 Lag 选项，单击 Change 按钮确认修改；用同样的方法对 buybk\_1 变量进行设置。

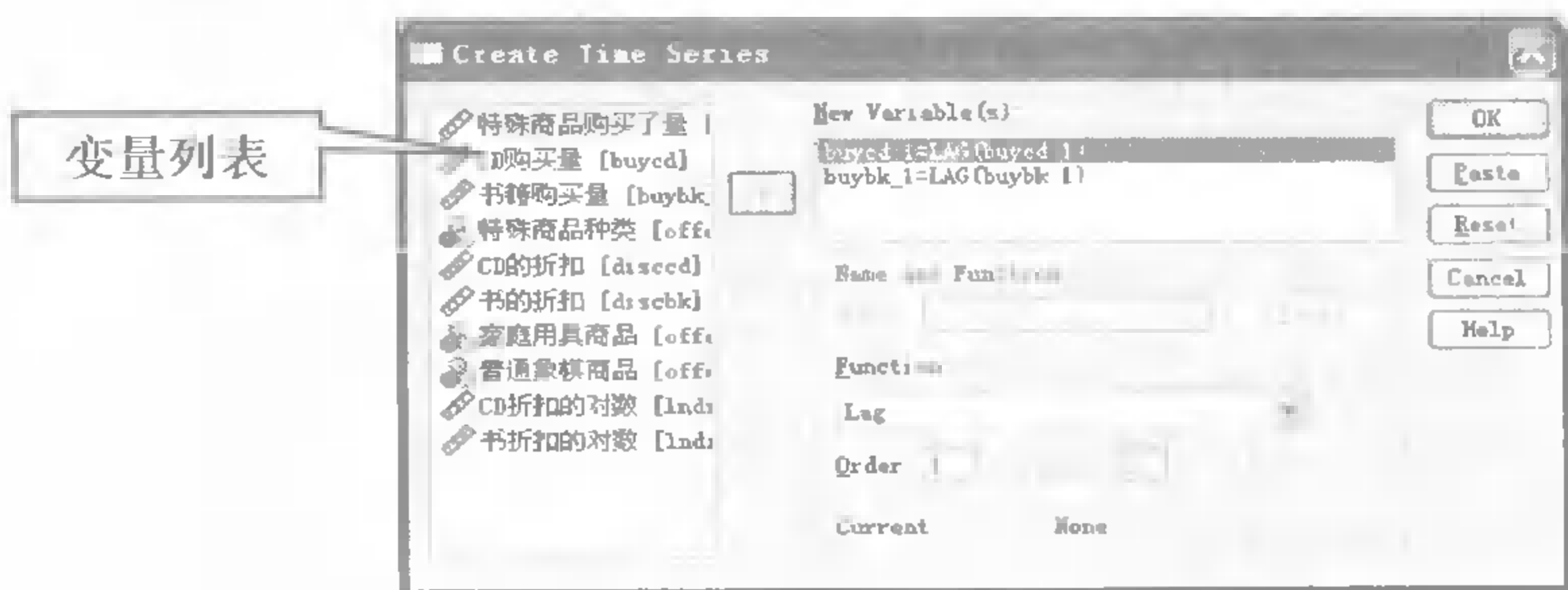


图 8-95 计算 CD 销售量和书籍销售量的滞后变量操作

在图 8-95 中，Function 列表的 Lag 选项就是指定的滞后函数，Order 输入框指定的是滞后阶数。单击 OK 按钮运行，在当前数据集产生两个新的变量：CD 销售量的 1 期滞后 buycd\_1 和书籍销售量的 1 期滞后变量 buybk\_1，它们将在随后的 2SLS 分析中用到。

### 8.9.3 二阶段最小二乘回归的参数设置

依次单击菜单“Analyze→Regression→2-Stage Least Squares...”执行二阶段最小二乘回归分析的功能，其主界面如图 8-96 所示，在此指定进行分析的各种变量。

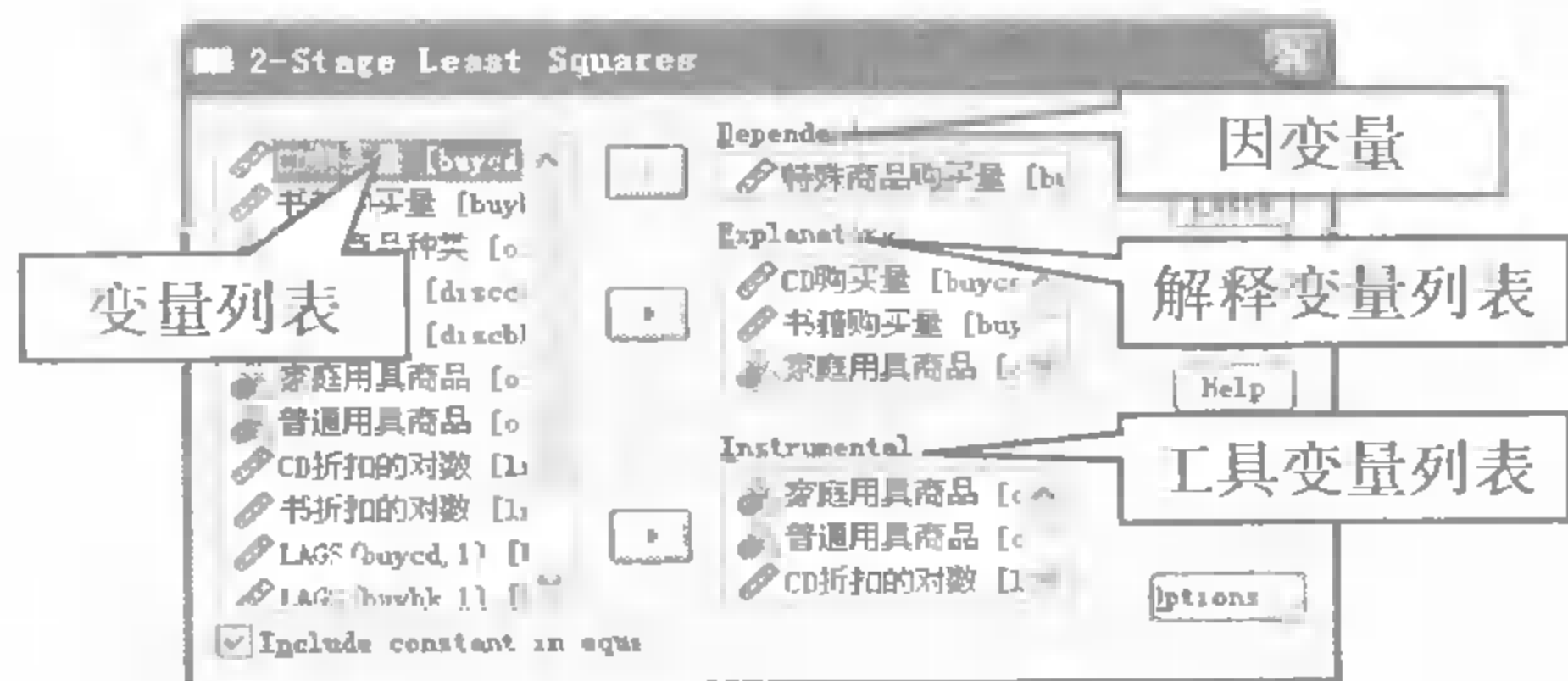





图 8-96 2-Stage Least Squares 过程参数设置

#### 1. 变量设置

在变量列表中单击选中特殊商品购买量，单击从上至下第一个  按钮，将其作为因变量选入 Dependent 选框；在变量列表中选中 CD 购买量、书籍购买量、家庭用具商品和普通用具商品，单击从上至下第二个  按钮，将其作为解释变量选入 Explanatory 列表框；在变量列表中选中从家庭用具商品到 LAGS(buybk,1)的 6 个变量，单击从上至下第三个  按钮，将其作为工具变量选入 Instrumental 列表框。

- ② Dependent 选框，用于从变量列表选入因变量。
- ③ Explanatory 列表框，用于从变量列表选入解释变量。

- Instrumental 列表框，用于从变量列表选入工具变量，它们将被用来预测内生变量的值。要求工具变量的个数至少要与解释变量的个数相等；如果解释变量与工具变量完全相同，则回归结果与线性回归的结果完全相同；没有被指定为工具变量的解释变量可以看做内生变量。
- Include constant in equation 复选框，勾选它要求模型中包括常数项。

## 2. 选项设置

在图 8-96 中单击 Option 按钮，弹出如 8-97 所示的选项设置对话框，勾选 Predicted 复选框，单击 Continue 按钮返回主界面。

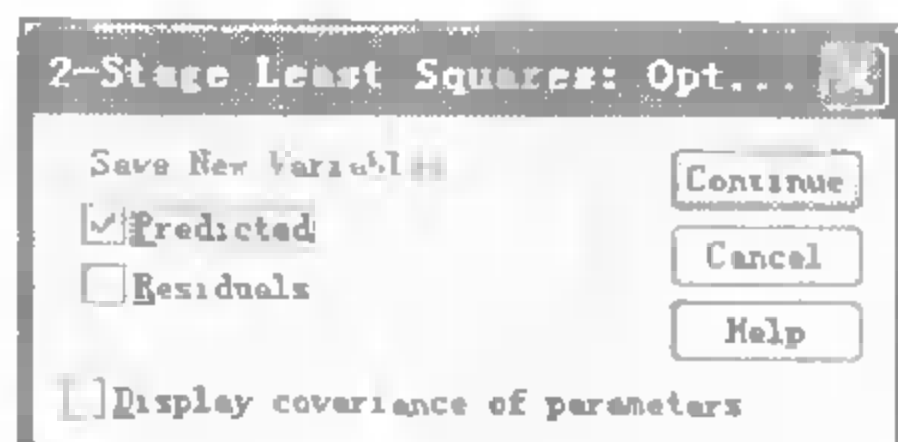


图 8-97 Option 选项设置

- Save New Variable 子设置栏，用于选择要保存到当前数据集的新变量，有如下两个选项：Predicted 复选框保存预测值；Residuals 复选框保存残差。
- Display covariance of parameters 复选框，勾选要求输出参数估计值的协方差阵。

### 8.9.4 案例的结果分析

在图 8-96 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 8-98~图 8-100 所示。

模型描述		
方程	变量	变量类型
1	buyoff	因变量
	buycd	预测值
	buybk	预测值
	offer_type1	预测值与工具
	offer_type2	预测值与工具
	lndiscd	工具
	lndiscbk	工具
	buycd_1	工具
	buybk_1	工具

模型汇总		
方程 1	复相关系数	383
	R 方	147
	调整 R 方	110
	估计的标准误	340

图 8-98 模型概述和模型的拟合优度

ANOVA					
方程 1	平方和	df	均方	F	Sig.
回归	1 851	4	463	3 994	005
残差	10 772	93	116		
总计	12 623	97			

图 8-99 ANOVA 方差分析表

系数					
	未标准化系数		Beta	t	Sig.
	B	标准误			
方程 1 (常数)	-1 511	1 317		-1 147	254
buycd	353	106	1 090	3 336	001
buybk	189	116	542	1 626	107
offer_type1	130	991	117	1 425	158
offer_type2	303	105	300	2 899	005

图 8-100 参数估计值输出

(1) 模型概述和模型汇总。如图 8-98 所示，“模型描述”表格给出模型所用变量的相关信息，变量类型为“预测值”的变量，将用类型为“工具”的变量进行预测，并且用这些预测值取代原来的观测值进行回归模型的估计；变量类型为“预测值与工具”的变量，既用要



它们预测类型为“预测值”的变量值，也要用它们的原始观测值进行回归模型的估计；变量类型为“工具”的变量，只用它们预测类型为“预测值”的变量值，而不用用于最终回归方程的估计。

在“模型汇总”表格里，复相关系数测量的是因变量观测值和预测值之间的相关性大小，较小的值说明二者相关性不强；R 方就是复相关系数的平方，表示当前回归模型解释了因变量差异的 14.7%；调整 R 方用来比较不同的模型，调整 R 方越大，模型拟合效果越好。

(2) 方差分析表。如图 8-99 所示，本例的回归平方和比残差平方和小很多，说明模型只解释了因变量变异的一小部分；F 检验的 Sig 值小于 0.01，说明由模型所解释的那部分变异并不是随机的。

(3) 参数估计值。如图 8-100 所示，由系数的估计值可得回归方程为  $\text{buyoff} = -1.511 + 0.353 * \text{buycd} + 0.189 * \text{buybk} + 0.130 * \text{offer\_type1} + 0.303 * \text{offer\_type2}$ 。但是有几个系数的显著性检验 Sig 值大于 0.1，关于这些变量对模型的贡献，有必要作进一步的探讨和分析。

## 8.10 最优尺度回归

实际工作中经常遇到有序而非数值型的数据（例如描述病情的好转、正常、恶化，描述学历水平的高中、大学本科、硕士研究生等）。大多情况下，这种分类数据的度量起点（零点）是比较难确认的，各取值水平之间的可比关系也是较为模糊的，虽然可以将其取值水平进行重新编码，但是它们相互之间的真实距离仍是不能明确的。

普通回归方法可用来预测分类变量，并且能估计不同类别之间的相关性，但其使用前提是对分类变量进行适当的编码处理，于是不同的编码方案就可能产生不同的回归结果。最优尺度（Optimal Scaling）回归方法能自动将分类变量转换为数值型进行分析，它的常用缩写为 CATREG（Category Regression），即分类回归。

### 8.10.1 最优尺度回归原理

最优尺度回归是标准回归方法的扩展，它按照比例换算名义变量、有序变量以及数值型变量，使用定量化的方法反映各种变量的属性，并利用非线性转换求解最佳的回归方程。

最优尺度回归使用整数对名义变量或者有序变量进行重新编码，默认使用 1 作为每个分类变量取值的起始点（零点）。如果是对数值型分类变量的重新编码，就把改变量的每个原始值都减去其原始的最小值，再加 1 然后取整。

SPSS 的最优尺度回归只允许设置 1 个因变量，最多可设置 200 个自变量。数据中应至少包含 3 个有效的观测记录，并且有效观测的数量必须超过自变量的个数加 1。

### 8.10.2 问题描述和数据准备

某吸尘器生产商调查了影响消费者偏好的 5 个因素，包括包装设计（A\*、B\*、C\*）、商标名称（K2R、Glory 和 Bissell）、价格、是否经过认证（是或否）和退款保证（是或否）。指定这 5 个因素取值的 22 种组合，对应了 22 种类型的产品，请每位消费者对这 22 种产品进行偏好排序，序号越小表示越喜欢。把产品每次被排序的序号作为得分，以一个偏好变量（pref）记每种产品的平均得分。生产商希望通过这些数据分析某种类型的吸尘器的市场前景。

本节通过最优尺度回归来研究偏好与这 5 个因素之间的关系。所用数据摘自 SPSS 自带的 Demo 文件“carpet.sav”，所用文件为“吸尘器偏好数据.sav”，数据格式如图 8-101 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	package	Numeric	4	0	包装设计	{1, A*}...	None	8	Right	Nominal
2	brand	Numeric	4	0	商标名称	{1, K2R}..	None	8	Right	Nominal
3	price	Numeric	4	0	价格	{1, \$1.19}	None	8	Right	Nominal
4	seal	Numeric	4	0	是否经过认证	{1, No}...	None	8	Right	Nominal
5	money	Numeric	4	0	退款保证	{1, No}...	None	8	Right	Nominal
6	pref	Numeric	4	0	偏好	None	None	8	Right	Ordinal

图 8-101 吸尘器偏好数据格式

### 8.10.3 最优尺度回归的参数设置

依次单击菜单“Analyze→Regression→Optimal Scaling...”执行 Optimal Scaling 回归分析过程，其主界面如图 8-102 所示，在此进行分析变量的选择和设置。

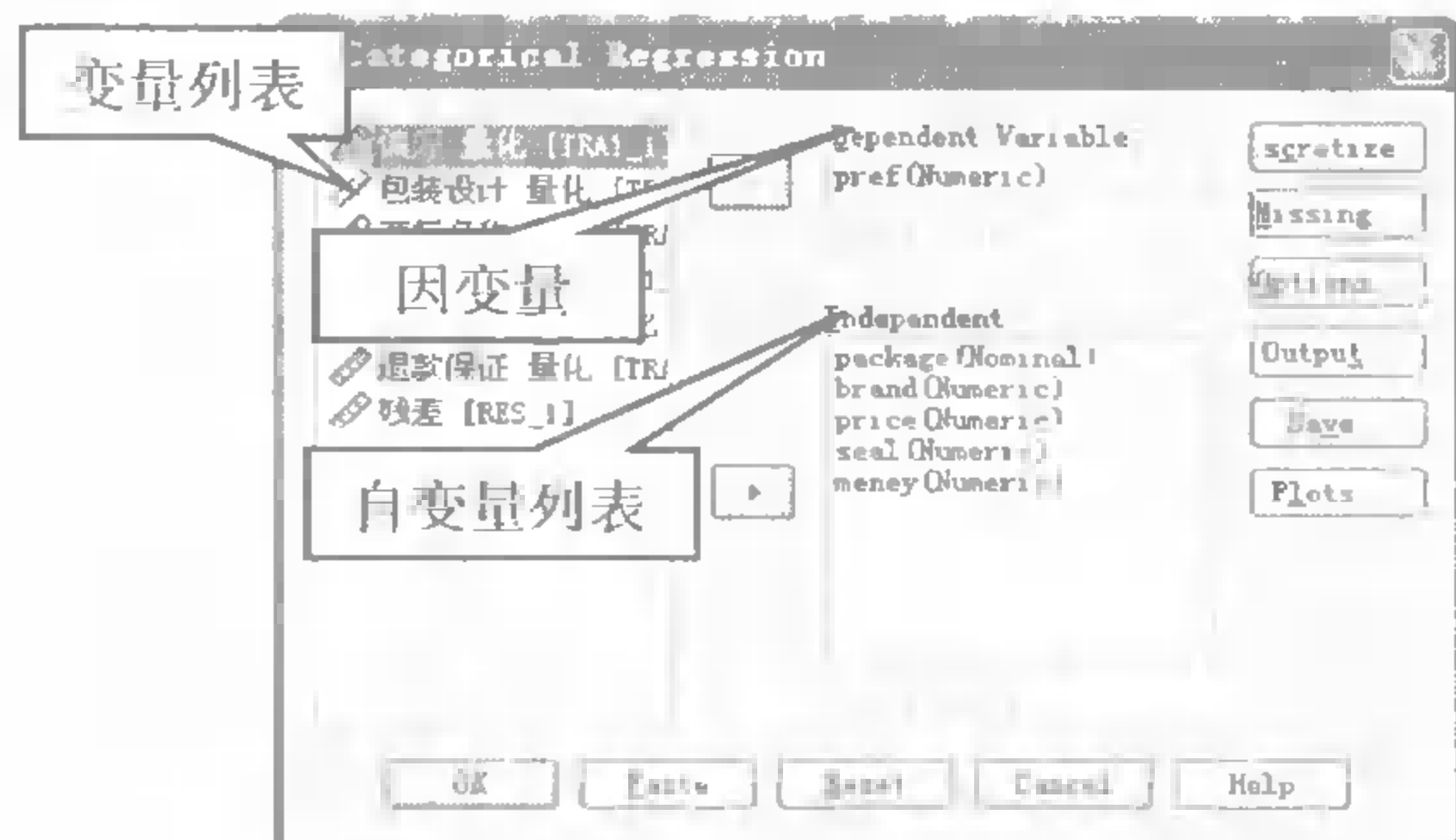




图 8-102 最优尺度回归的主设置界面

#### 1. 变量设置

在变量列表中单击选中偏好(pref)变量，单击从上至下第一个  按钮，将其作为因变量选入 Dependent 选框；在变量列表中选中从包装设计到退款保证的 5 个变量，单击从上至下第二个  按钮，将其作为自变量选入 Independent 列表框。单击选中 Dependent 选框的 pref 变量，再单击其下的 Define Scale 按钮，弹出如图 8-103 所示的对话框，单击选中 Numeric 单选框，单击 Continue 按钮返回主界面；单击选中 Independent 列表的 package 变量，再单击其下的 Define Scale 按钮，弹出如图 8-103 所示的对话框，单击选中 Nominal 单选框，单击 Continue 按钮返回主界面；采用同样的方法将 Independent 列表的其他 4 个变量的最优尺度设置为 Numeric 型的。

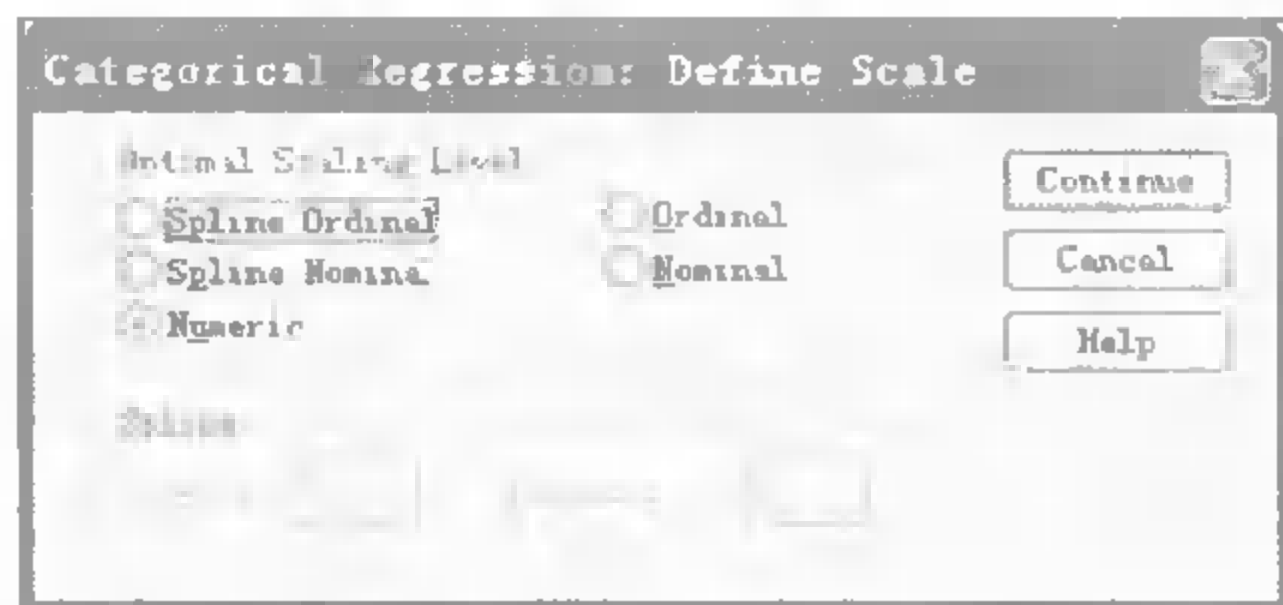


图 8-103 最优尺度回归的取值范围定义对话框

① Dependent Variable 选框，用于从变量列表选入因变量。

② Independent Variable(s)列表框，用于从变量列表选入自变量。

③ 设置指定变量的最优尺度。对于选入的因变量或自变量，都可以定义变量取值的最优尺度。先选中需要设置的变量，再单击相应的 Define Scale 按钮，在弹出的如图 8-103 所示的对话框里进行设置。Optional Scaling Level 子设置栏用于选择变量取值最优尺度的度量方

式, 共有如下 5 个可选项。

- Spline Ordinal 有序样条尺度。最优尺度变量将保持观测变量的取值顺序, 分类点将被置于通过原点的一条直线(或向量)上。结果转换是一个分段光滑且单调的多项式函数, 每个分段多项式的阶数在 Spline 栏的 Degree 输入框指定, 默认值为 2; 分段个数通过在 Spline 栏的 Interior knots 输入框指定的结点个数来确定, 默认值为 2; 分段的位置由程序自动判断。
- Spline Nominal 名义样条尺度。最优尺度变量只保持观测变量对样本的分类结果, 但不保持观测变量的取值顺序, 分类点将被置于通过原点的一条直线(或向量)上。结果转换是一个分段光滑的多项式函数, 每个分段不一定再单调; 分段多项式的阶数、分段个数和分段位置的设置方法同 Spline Ordinal 选项。
- Ordinal 有序尺度。最优尺度变量将保持观测变量的取值顺序, 分类点将被置于通过原点的一条直线(或向量)上。结果转换的拟合效果比 Spline Ordinal 好, 但是光滑性(smooth)要差一些。
- Nominal 名义尺度。最优尺度变量只保持观测变量对样本的分类结果, 但不保存观测变量的取值顺序, 分类点将被置于通过原点的一条直线(或向量)上。结果转换的拟合效果比 Spline Nominal 好, 但是光滑性(smooth)要差一些。
- Numeric 数值尺度。此方法认为分类变量的取值是有序且等间隔的。最优尺度变量将保持观测变量的取值顺序及其相等间隔, 分类点将被置于通过原点的一条直线(或向量)上。如果所有变量都采用数值尺度, 此分析过程就与标准的主成分分析非常类似。

设置好变量的最优尺度后, 在主界面的因变量、自变量名称之后将以括号括住的方式显示当前变量采用的最优尺度, 例如: pref(Numeric)表示 pref 变量的最优尺度为 Numeric。

## 2. 缺失值设置

在图 8-102 中单击 Missing Value 按钮, 弹出如图 8-104 所示的缺失值设置对话框, 单击 Continue 按钮返回主界面。

Analysis Variables 列表框显示当前分析用到的变量, 每个变量后面括号内的说明就是当前变量的缺失值处理方法。如果需要更改对某些变量的处理方式, 先在此列表选中需要改变的变量, 然后在 Strategy 栏选择一种缺失值处理方法, 再单击 Change 按钮确定更改。

Strategy 栏用于选择缺失值处理的方法, 系统给出如下两个选项。

- ① Exclude objects with missing values on this variable 单选框, 如果指定变量取缺失值, 则相应的观测不参与分析。
- ② Impute missing values 单选框, 表示用估计值替代缺失值, 有如下两种可选方法。
  - Mode 众数, 用出现频数最多的值代替缺失值, 如果存在多个众数, 取类别编号最小的那个众数替代缺失值。
  - Extra category 选项, 把缺失值作为单独的一类, 并对其进行编码。

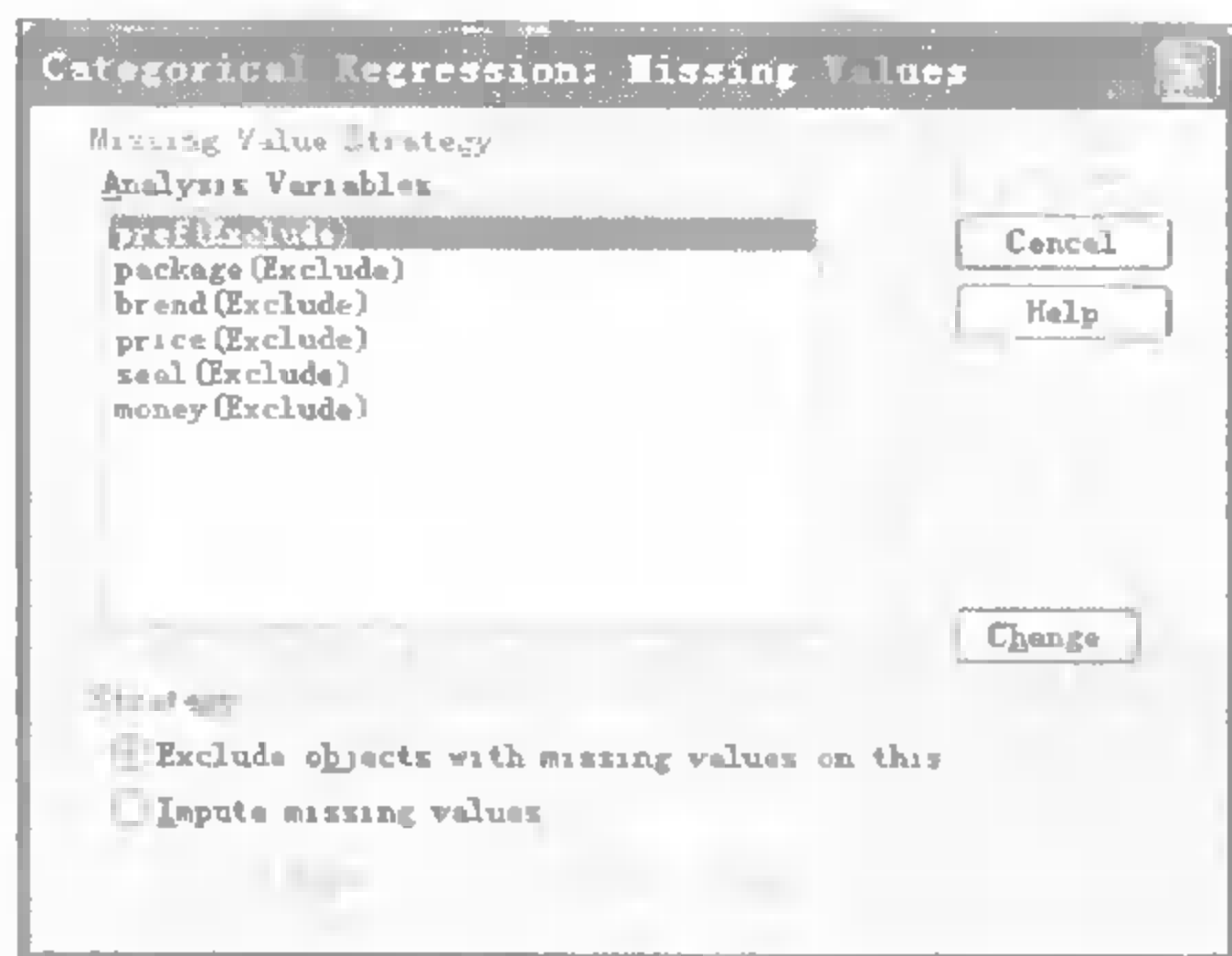


图 8-104 最优尺度回归的缺失值设置

### 3. Options 选项设置

在图 8-102 中单击 Options 按钮,弹出如图 8-105 所示的选项设置对话框,在此对分析过程的多个参数加以设置。单击 Continue 按钮返回主界面。

① Supplementary Objects 栏,用于指定数据中的增补对象,有如下两种可选方式。

- ① Range of cases 指定观测范围,在 First、Last 输入框分别指定增补对象的起始观测和终止观测序号,单击 Add 进行添加。
- ② Single case 指定单个记录,在输入框指定单个观测的序号,单击 Add 进行添加。对于已经选入的增补对象,可通过单击 Change、Remove 按钮加以修改或删除。

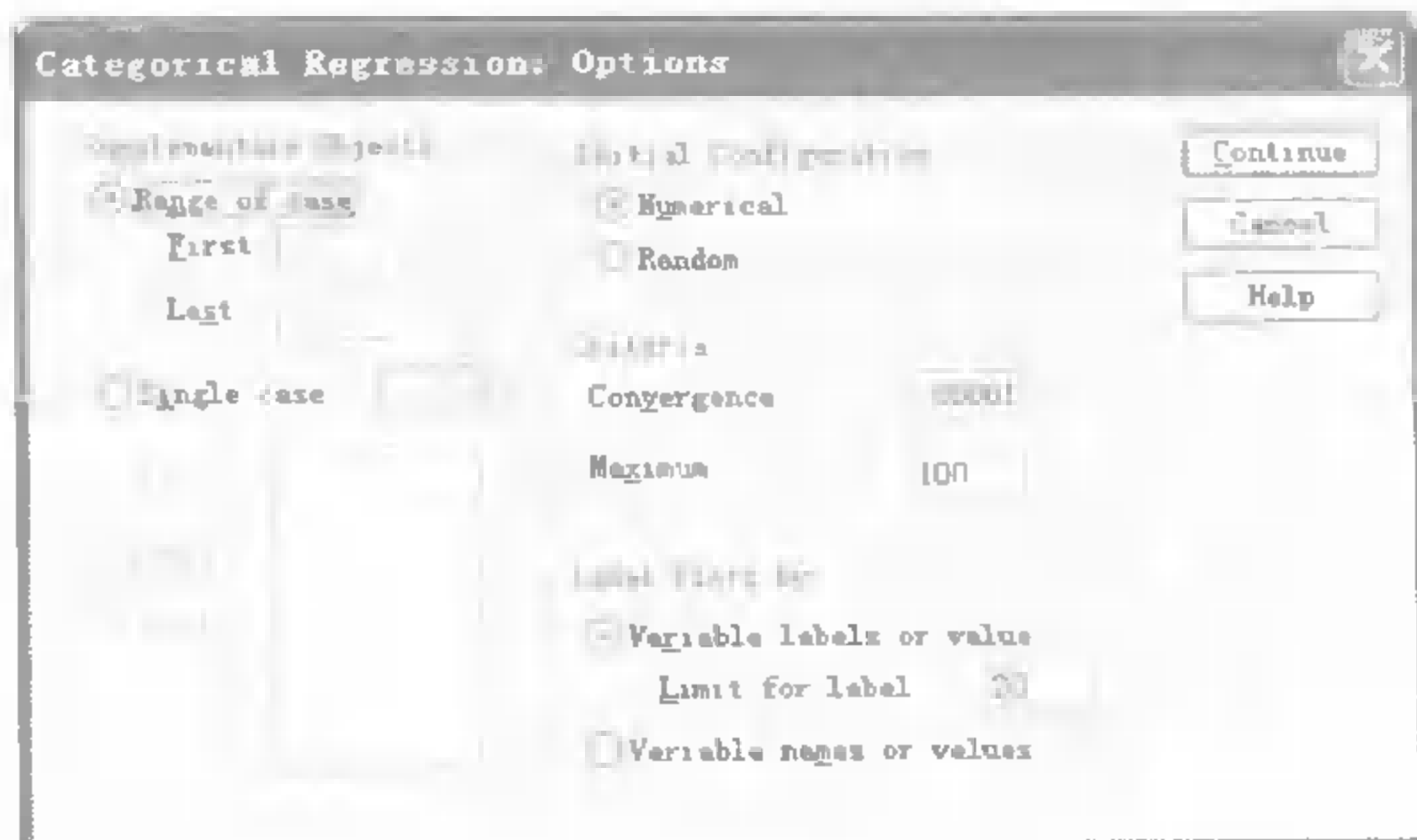


图 8-105 最优尺度回归的选项设置

② Initial Configuration 栏,用于指定对变量的先验认识,有两个可选项。

- ① Numerical 数值,如果分析变量里没有指定名义变量(nominal),选择此项。
- ② Random 随机,如果分析变量里至少有一个名义变量(nominal),选择此项。

③ Criteria 栏,用于设置迭代过程的收敛依据,有如下两个设置内容。

- ① Convergence 输入框,指定回归过程收敛依据的临界值,默认值为 0.000 01,如果两个相邻迭代模型的拟合度差值小于此处所设置的临界值,则回归过程结束。
- ② Maximum iterations 输入框,指定回归过程的最大迭代步数,必须为正数,默认值为 100。

④ Label Plots By 栏,用于设置作图时对变量的标识方式,有两个可选项。

- ① Variable labels or vaule vaules 单选框,在图形中显示变量标签和值标签,同时可在 Limit for lable length 后指定标签的最大长度值。
- ② Variable names or vaules 单选框,在图形中显示变量名和观测值。

### 4. 变量编码设置

在图 8-102 中单击 Discretization 按钮,弹出如图 8-106 所示的设置对话框,在这里设置对变量进行离散化的编码方式。单击 Continue 按钮返回主界面。

(1) Varizbles 列表框显示当前可以进行重新编码的变量,变量名后的括号内说明了当前

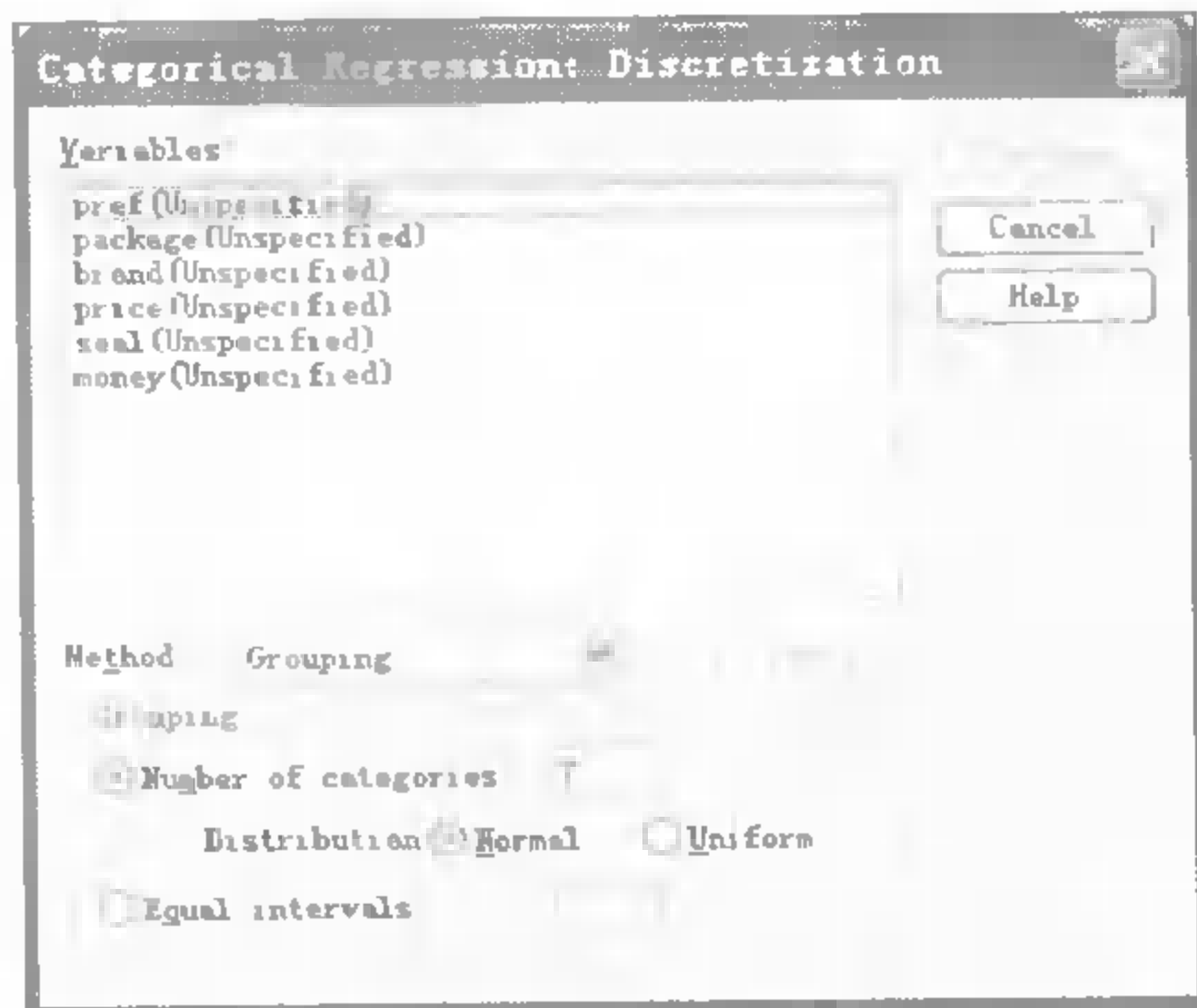


图 8-106 最优尺度回归的变量编码设置

变量使用的离散化方法。如果需要更改某些变量的离散化方法,先在此列表选中需要改变的变量,然后在 Method 栏指定一种离散化处理方法,再单击 Change 按钮确定更改。

(2) Method 下拉列表给出了如下 4 种对变量进行离散化编码的方式。

- ① Unspecified 选项,不进行离散化处理。
- ② Grouping 分组法,将原始变量离散化为指定取值个数或取值间隔的类别变量。

选中后,激活 Grouping 栏的设置选项: Number of



categories 输入框,指定离散化后的取值个数,还可以指定这些取值所服从的分布类型是 Normal (正态分布) 还是 Uniform (均匀分布); Equal intervals 输入框,指定离散化后取值的等间隔长度。

③ Ranking 排序法,通过对观测进行排序,取秩统计量进行离散化。

④ Multiplying 倍增法,对变量的当前取值先进行标准化,再乘以 10 后四舍五入,然后加上一个常数,使其最小取值为 1。

## 5. 输出设置

在图 8-102 中单击 Output 按钮,弹出图 8-107 所示的对话框,在此选择分析过程的输出选项。勾选 Correlations of the original variables 复选框和 Correlations of the transformed variables 复选框;单击取消选中 ANOVA 复选框;单击 Continue 按钮返回主界面。

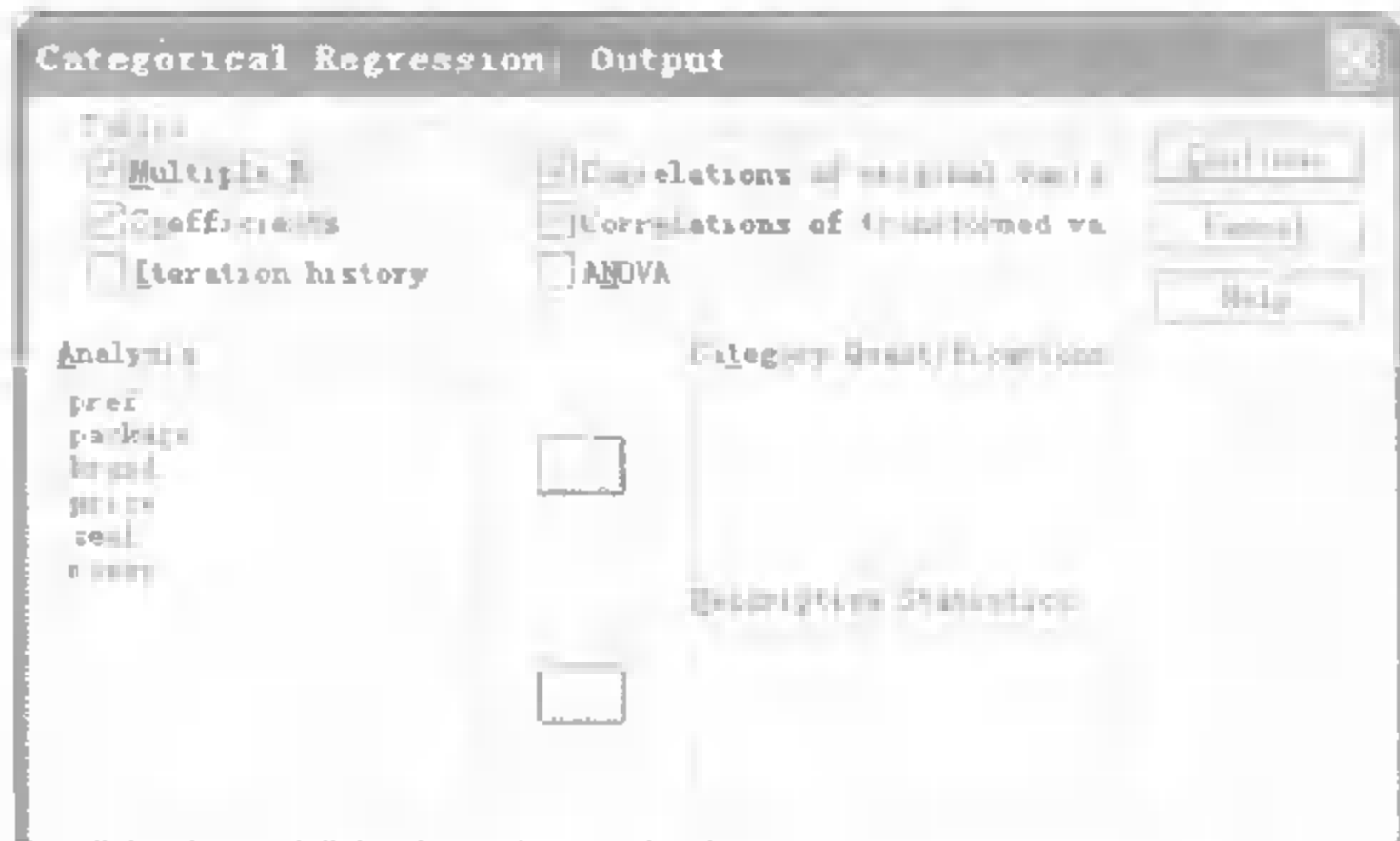


图 8-107 最优尺度回归的输出设置

① Tables 栏用于选择要输出的表格,共有 6 项。

② Multiple R 复相关系数,包括 R 方统计量、调整 R 方等。

③ Coefficients 系数选项,输出以下三个表:回归系数表,包括参数估计值 B、B 的标准误、t 检验统计量及其显著性统计量;最优尺度系数表;相关系数与容许度表。

④ Iteration history 迭代历史,输出每一步迭代初始值、R 方统计量和回归误差等信息。

⑤ Correlations of the original variables 选项,输出转换前变量之间的相关系数矩阵。

⑥ Correlations of the transformed variables 选项,输出转换后变量之间的相关系数矩阵。

⑦ ANOVA 输出方差分析表,包括回归平方和、残差平方和和 F 统计量等信息。

⑧ Analysis Variables 列表显示了当前分析中的变量。

⑨ Category Quantifications (分类量化) 列表,可从 Analysis Variables 列表中选入变量,输出所选变量经过转换后的变量取值。

⑩ Descriptive Statistics (描述性统计量) 列表,可从 Analysis Variables 列表中选入变量,输出所选变量的频数、缺失值和众数等描述性信息。

## 6. 保存设置

在图 8-102 中单击 Save 按钮,弹出图 8-108 所示的对话框,在此选择需要保存的信息。

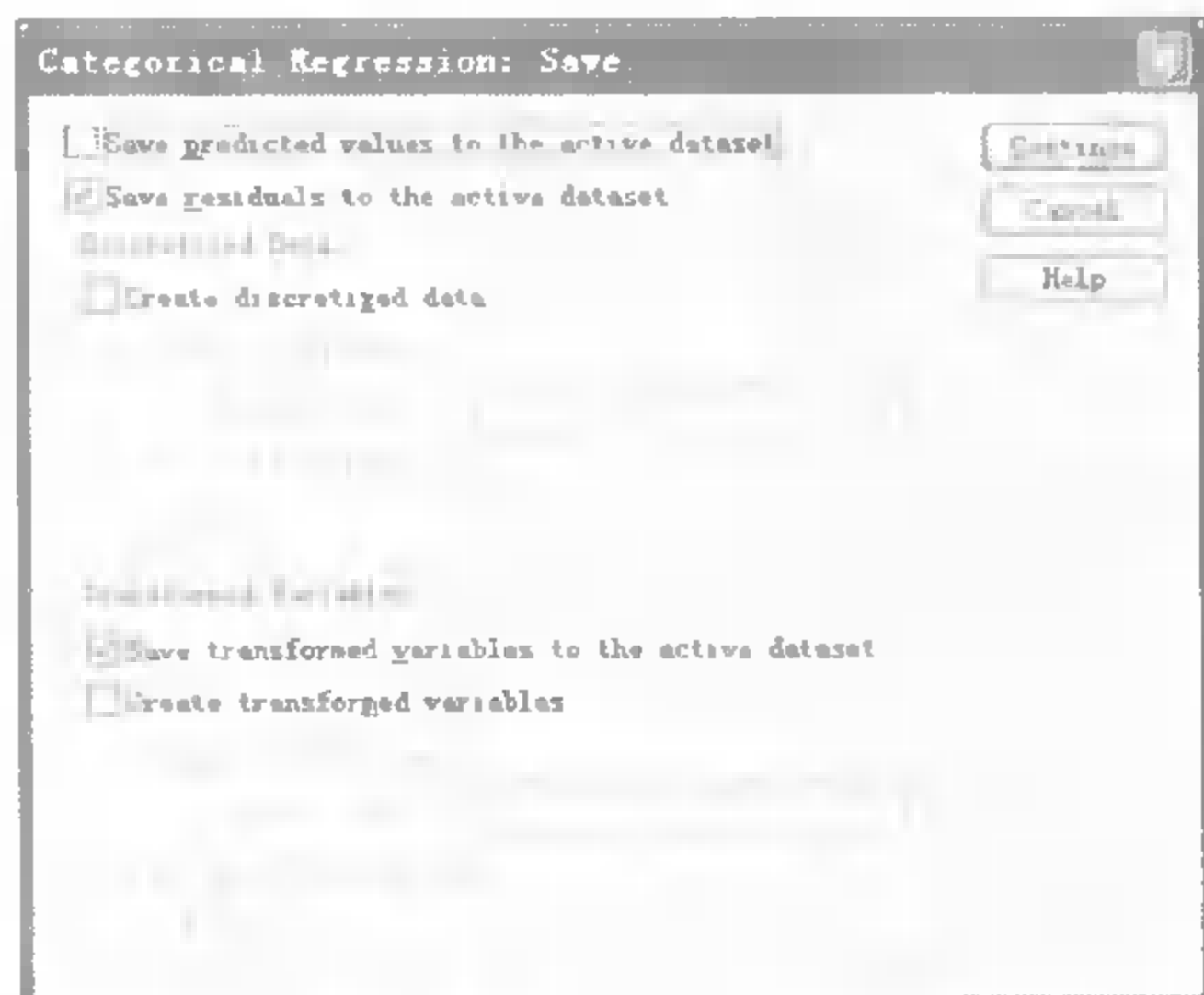


图 8-108 最优尺度回归的保存设置

勾选 Save residuals 复选框和 Save transformed 复选框;单击 Continue 按钮返回主界面。

① Save predicted values to the active dataset 复选框,将模型预测值保存到当前数据集中。

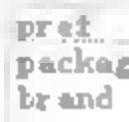
② Save residuals to the active dataset 复选框,将模型预测的残差保存到当前数据集中。

③ Discretized Data 栏,用于保存离散化后的数据,勾选 Create 复选框激活下面的选项。

④ Create a new dataset, 建立一个新的数据集,在 Dataset name 后输入数据集名称。

- Write a new data file, 建立一个新的数据文件, 单击 File 按钮指定文件路径和文件名。
- ④ Transformed Variables 栏, 用于保存转换后的变量。
- Save transformed variables to the active dataset 复选框, 保存到当前数据集中。
- 勾选 Create 复选框激活如下 2 项: Create a new dataset 建立一个新数据集, 并在 Dataset name 后输入数据集名称; Write a new data file 建立一个新的数据文件, 单击 File 按钮指定文件路径和文件名。

## 7. 图形设置

在图 8-102 中单击 Plots 按钮, 弹出图 8-109 所示的对话框, 在此选择需要输出的图形种类。在变量列表中选中 package 和 price 变量, 单击从上至下第一个  按钮, 将其选入 Transformation Plots 列表框; 单击 Continue 按钮返回主界面。

① Transformation Plots 列表框, 对于从变量列表选入其中的变量, 输出它们的转换图形。转换图的横轴代表变量转换前的观测值, 纵轴代表转换后的量化值。

② Residual Plots 列表, 对于从变量列表选入其中的变量, 输出它们的残差图形。残差图的纵轴代表残差值, 横轴代表当前分类变量的取值。

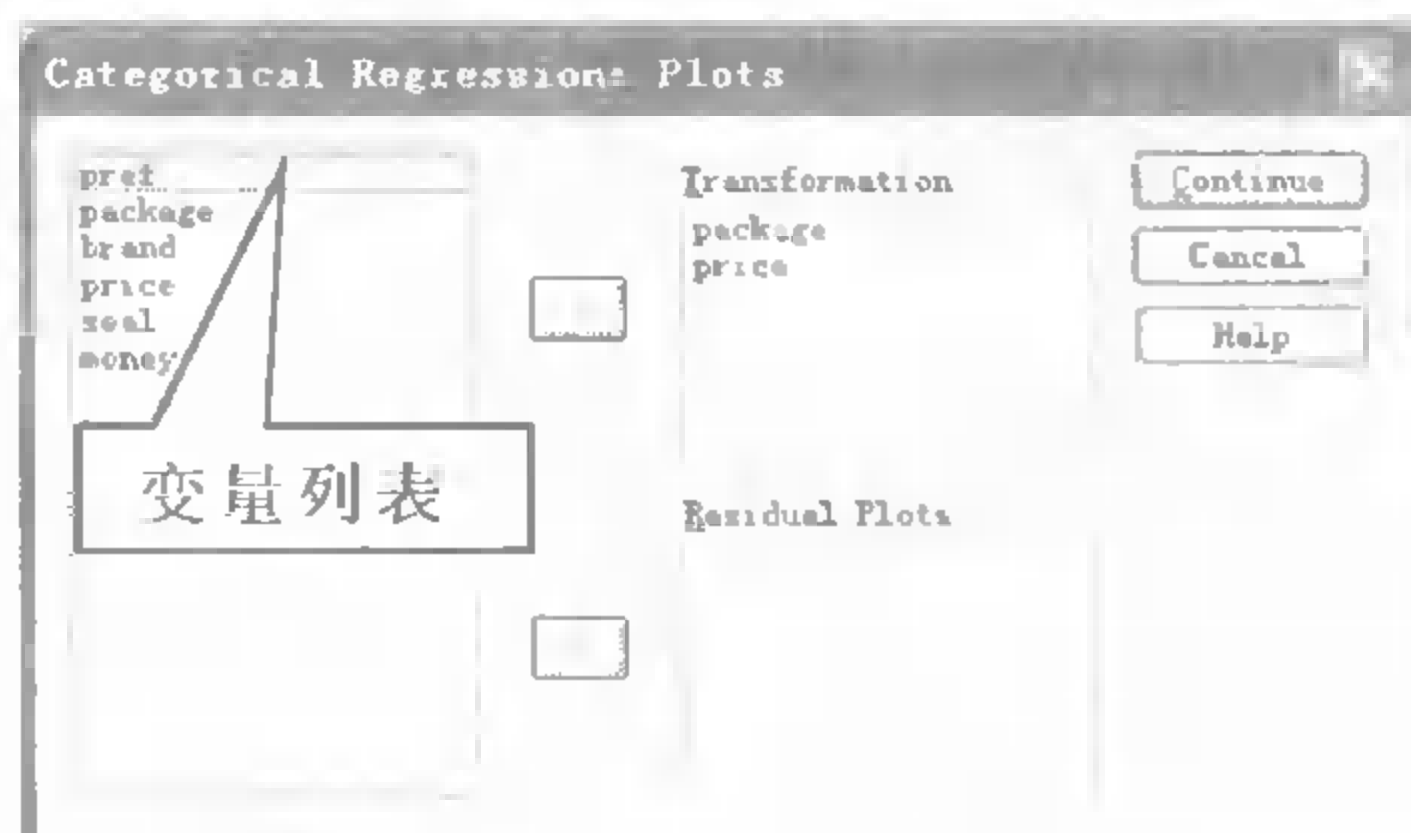


图 8-109 最优尺度回归的图形设置

## 8.10.4 案例的结果分析

单击图 8-102 中的 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 8-110~图 8-114 所示。

信誉		案例处理摘要	
Catreg		有效的活动案例	22
Version 2.1		具有缺失值的活动案例	0
by		补充案例	0
Data Theory Scaling System Group (DTSS)		总计	22
Faculty of Social and Behavioral Sciences		分析中使用的案例	22
Leiden University, The Netherlands			

图 8-110 信誉和案例处理摘要

初始变量的相关系数					
	包装设计	商标名称	价格	是否经过认证	退款保证
包装设计	1.000	-.189	-.126	.031	.066
商标名称	-.189	1.000	.045	-.042	-.034
价格	-.126	.045	1.000	.000	.000
是否经过认证	.031	-.042	.000	1.000	-.039
退款保证	.066	-.034	.000	-.039	1.000
维	1	2	3	4	5
特征值	1.291	1.038	.980	.905	.783

已转换变量的相关系数					
	包装设计	商标名称	价格	是否经过认证	退款保证
包装设计	1.000	-.186	-.089	.032	.102
商标名称	-.186	1.000	.005	-.042	-.034
价格	-.089	.005	1.000	.000	.000
是否经过认证	.032	-.042	.000	1.000	-.039
退款保证	.102	-.034	.000	-.039	1.000
维	1	2	3	4	5
特征值	1.248	1.043	.983	.905	.821

图 8-111 相关系数表

模型摘要		
多 R	R 方	调整的 R 方
.971	.948	.927
因变量 偏好		
预测值 包装设计 商标名称 价格 是否经过认证 退款保证		

	标准系数		df	F	显著性
	Beta1	标准误			
包装设计	-.748	.060	2	155.288	.000
商标名称	.045	.060	1	.578	.459
价格	.371	.059	1	38.312	.000
是否经过认证	-.000	.059	1	.000	.999
退款保证	-.159	.059	1	7.175	.017
因变量 偏好					

图 8-112 模型摘要表和参数估计值表

相关性和容差						
	相关性			重要性	容差	
	零阶	偏	部分		转换后	转换前
包装设计	-.816	-.955	-.733	.624	.959	.942
商标名称	.006	.193	.045	.010	.971	.961
价格	.440	.651	.369	.172	.939	.912
是否经过认证	.370	-.538	.149	.139	.996	.991
退款保证	-.223	-.569	.158	.037	.987	.985
因变量 偏好						

图 8-113 相关性和容差表

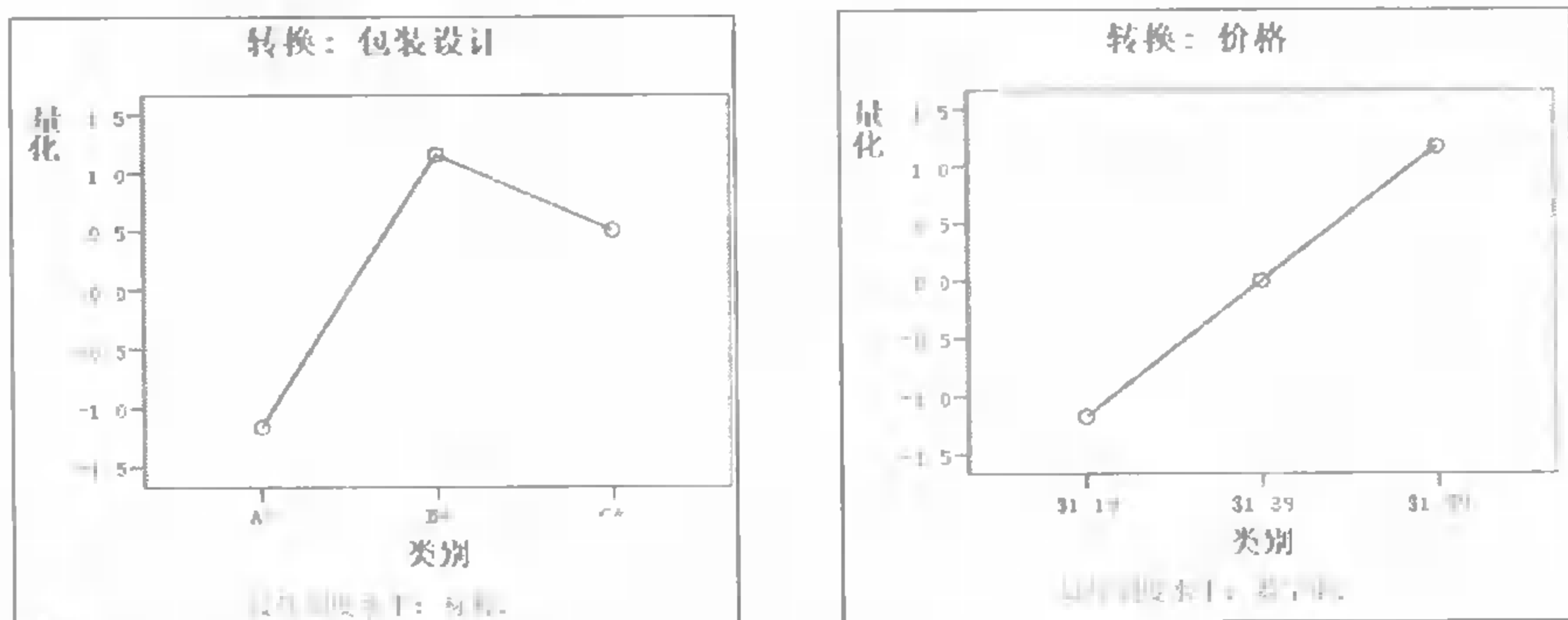


图 8-114 转换图形

(1) 信誉和案例处理摘要输出。如图 8-110 所示, 给出了 Catreg 模块的版权信息及关于案例有效个数等的统计信息。

(2) 相关系数输出。如图 8-111 所示, 给出了初始变量间的相关系数和转换后变量间的相关系数表。可见, 两个表格里的相关系数全都接近于 0, 因此各变量之间可认为是相互独立的, 不存在多重共线性。通过对比两个表格的内容发现, 只有与包装设计有关的相关系数发生改变, 这是因为分析中它的最优尺度设为 Nominal, 而其它变量的最优尺度都为 Numeric。

(3) 模型摘要输出和参数估计输出。如图 8-112 所示, “模型摘要”表中最优尺度回归方程的 R 方和调整 R 方都大于 0.9, 说明回归模型能解释 90% 以上的总变异, 拟合效果不错。而普通回归方程 (操作请参考第 8.1.5 节的介绍) 的 R 方和调整 R 方分别为 0.707、0.615, 可见尺度回归的改进效果不错。

“系数”表格给出了参数估计的结果, 由于尺度回归对变量进行了标准化处理, 所以得到的系数也是标准化的。从 F 检验的显著性 Sig 值看, 除了商标名称不显著外, 其它变量对回归方程的贡献都是显著的。本例包装设计的标准化系数为 -0.748, 表示它的标准化取值增加 1 个单位, 预测值就减小 0.748; 但是因为包装设计的最优尺度为 Nominal, 所以它量化值的增加不一定就反应了原始值的增加。

(4) 相关性和容差。如图 8-113 所示，它包括了偏相关系数、部分相关系数、转换前容差和转换后容差等统计量，这些都可以反映自变量对因变量的影响程度。零阶相关性给出的是转换后的自变量和因变量之间的相关系数；包装设计的偏相关系数最大为-0.955，表示不考虑其它变量的影响时，包装设计解释了因变量 $(-0.955)^2=0.91=91\%$ 的变异；包装设计的部分相关系数为-0.733，表示从包装设计中去除了其它 4 个因素的影响后，剩余部分解释了因变量 $(-0.733)^2=0.54=54\%$ 的变异；重要性取值越大的变量对回归方程的贡献也越大。

自变量之间存在相关关系时会导致模型的不稳定，容差可以反应自变量之间的线性相关程度，它表示的是单个变量不能被其它变量解释的变异比例，接近 1 表示它不能被其它变量预测，本例各变量的容差都很大（大于 0.9），说明变量之间没有明显的线性关系。

(5) 转换图形。如图 8-114 所示，是包装设计和价格的量化转换图形，图中直观地显示了各变量转换前后取值的对应关系。对价格采用的是 numeric 尺度转换，\$1.19 与 \$1.39 转换后的距离和 \$1.39 与 \$1.59 转换后的距离一样，这表示类别 1、2 的距离与类别 3、2 的距离是相等的；包装设计的转换没有保持类别间的等间隔性，因为它选择的是 nominal 尺度转换。



方差分析 (Analysis of Variance, ANOVA) 是由英国统计学家 R.A.Fisher 于 1923 年提出的, 它是一种利用试验获取数据并进行分析的统计方法, 经常用于研究不同效应对指定试验的影响是否显著。常用的方差分析方法包括单因素方差分析、多因素方差分析、多元方差分析、协方差分析、重复设计方差分析。

通过对试验进行精心的设计, 能够在有限的物质条件下 (时间、金钱、人力等), 从尽可能少的试验中获取数据, 并最大限度地包含有用的信息, 而方差分析就是从相应的试验数据中提取这种信息的统计分析方法。在科学试验和现代工业质量控制中, 这套统计方法都得到了广泛的应用, 并产生了很好的效果。

## 9.1 方差分析简介

方差分析把观测总变异的平方和及自由度分解为对应于不同变异来源的平方和及自由度, 以此获得不同来源的变异的估计值, 从而发现各个因素在总变异中的重要程度。通过计算这些变异估计值的适当比值还可以做某些假设检验, 例如检验各样本所属总体的平均数是否相等。

方差分析实质上是关于观测变异原因的数量分析, 它在科学研究中的应用十分广泛。

### 9.1.1 $t$ 检验与方差分析的比较

$t$  检验适用于样本平均数与总体平均数或者两样本平均数之间的差异显著性检验, 但在实际生产和科学研究中经常会遇到多个样本的情况, 这就需要进行多个平均数之间的差异显著性检验, 此时再采用  $t$  检验法就不适宜了, 理由有如下 3 个。

(1) 检验过程繁琐。例如, 某试验包含 5 个处理, 采用  $t$  检验法要进行  $C_5^2=10$  次两两平均数之间的差异显著性检验; 如果有  $k$  个处理, 就要作  $k(k-1)/2$  次类似的检验。

(2) 误差估计的精确性。对同一试验的多个处理进行比较时, 应有统一的试验误差估计值, 如果用  $t$  检验做两两比较, 每次比较都要计算一个  $S_{x_1-x_2}$ , 所以多次比较的误差估计值不能统一。

另外,  $t$  检验不能充分利用数据资料所提供的信息提高误差估计的精确性。例如, 某试验有 5 个处理, 每个处理重复 6 次, 就得到 30 个观测值; 进行  $t$  检验时, 每次比较只能利用 2 个处理共 12 个观测值来估计试验误差, 自由度为  $2 \times (6-1)=10$ ; 如果利用整个试验的 30 个观测值来估计试验误差, 显然估计的精确性要高, 且误差自由度为  $5 \times (6-1)=25$ 。可见使用  $t$  检验时, 由于误差估计的精确性低, 误差自由度小, 因此检验的灵敏性低, 容易掩盖差异的显著性。

(3) 推断的可靠性低。用  $t$  检验进行多个平均数之间的差异显著性检验时, 由于没有考

考虑相互比较的两个平均数的秩次问题,因而会增大犯 I 型错误的概率,降低推断的可靠性。

鉴于以上理由,多个平均数之间的差异显著性检验不宜采用  $t$  检验方法,而须采用方差分析法。方差分析法是科学研究工作中的一个十分重要的工具。

### 9.1.2 方差分析的基本原理

方差分析有很多类型,但其基本原理与步骤是相似的,本节以单因素完全随机试验资料的方差分析为例进行简单介绍。方差分析过程可分为平方和与自由度的分解、F 值检验两个步骤。

#### 1. 平方和与自由度的分解

方差分析使用样本方差(mean squares)来度量数据资料的变异程度,它是变异平方和除以自由度的商。要将试验数据资料的总变异分解为不同来源的变异,首先就要将总变异平方和与总自由度分解为不同变异来源的相应部分。

假设某试验的指定因素可取  $k$  个处理水平,每个处理有  $n$  个观察值,那么试验数据就一共有  $n \times k$  个观察值,其数据组织格式一般如表 9-1 所示。

表 9-1  $k$  个处理每个处理有  $n$  个观测值的数据模式

处 理		观 测 值				合计 $x_{i.}$		平均 $\bar{x}_{i.}$
$A_1$	$x_{11}$	$x_{12}$	...	$x_{1j}$	...	$x_{1m}$	$x_{1.}$	$\bar{x}_1$
$A_2$	$x_{21}$	$x_{22}$	...	$x_{2j}$	...	$x_{2m}$	$x_{2.}$	$\bar{x}_2$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$	$\vdots$
$A_i$	$x_{i1}$	$x_{i2}$	...	$x_{ij}$	...	$x_{im}$	$x_{i.}$	$\bar{x}_i$
$\vdots$	$\vdots$	$\vdots$	...	$\vdots$	...	$\vdots$	$\vdots$	$\vdots$
$A_k$	$x_{k1}$	$x_{k2}$	...	$x_{kj}$	...	$x_{km}$	$x_{k.}$	$\bar{x}_k$
合计							$x_{..}$	$\bar{x}_{..}$

表中  $x_{ij}$  表示第  $i$  个处理的第  $j$  个观测值 ( $i=1,2,\dots,k; j=1,2,\dots,n$ );  $x_{i.} = \sum_{j=1}^n x_{ij}$  表示第  $i$  个处理的  $n$  个观测值的和;  $x_{..} = \sum_{i=1}^k \sum_{j=1}^n x_{ij} = \sum_{i=1}^k x_{i.}$  表示全部观测值的总和;  $\bar{x}_{i.} = \sum_{j=1}^n x_{ij} / n = x_{i.} / n$  表示第  $i$  个处理的观测平均值;  $\bar{x}_{..} = \sum_{i=1}^k \sum_{j=1}^n x_{ij} / kn = x_{..} / kn$  表示全部观测的总平均值。

#### (1) 总平方和的分解。

总平方和反映了全部观测的总变异情况,它是各观测值  $x_{ij}$  与总平均值  $\bar{x}_{..}$  的离差平方和,记为:  $SS_T = \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x}_{..})^2$ , 以下是对此平方和的分解过程。

$$\begin{aligned}
 \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x}_{..})^2 &= \sum_{i=1}^k \sum_{j=1}^n [(\bar{x}_{i.} - \bar{x}_{..}) + (x_{ij} - \bar{x}_{i.})]^2 \\
 &= \sum_{i=1}^k \sum_{j=1}^n [(\bar{x}_{i.} - \bar{x}_{..})^2 + 2(\bar{x}_{i.} - \bar{x}_{..})(x_{ij} - \bar{x}_{i.}) + (x_{ij} - \bar{x}_{i.})^2] \\
 &= n \sum_{i=1}^k (\bar{x}_{i.} - \bar{x}_{..})^2 + 2 \sum_{i=1}^k [(\bar{x}_{i.} - \bar{x}_{..}) \sum_{j=1}^n (x_{ij} - \bar{x}_{i.})] + \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x}_{i.})^2,
 \end{aligned}$$

由于:  $\sum_{j=1}^n (x_{ij} - \bar{x}_i) = 0$ , 可得:  $\sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x}_{..})^2 = n \sum_{i=1}^k (\bar{x}_i - \bar{x}_{..})^2 + \sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x}_i)^2$ , 在

此等式右侧: 前半部分为各处理平均数  $\bar{x}_i$  与总平均数  $\bar{x}_{..}$  的离差平方和, 再乘以试验次数  $n$  的积, 反映了  $k$  次处理之间的变异, 称为处理间平方和, 记为  $SS_t$ ; 后半部分为各处理内部离差平方和之和, 反映了各处理内部的变异情况(误差), 称为处理内平方和或误差平方和, 记为  $SS_e$ 。于是, 单因素方差分析总平方和的分解关系式就为:  $SS_T = SS_t + SS_e$ , 各部分的计算公式如下。

$$\begin{cases} SS_T = \sum_{i=1}^k \sum_{j=1}^n x_{ij}^2 - C \cdots \cdots \cdots \text{总平方和} \\ SS_t = \frac{1}{n} \sum_{i=1}^k x_i^2 - C \cdots \cdots \cdots \text{处理间平方和, 其中: } C = \frac{x_{..}^2}{k \times n} \text{ 称为校正数} \\ SS_e = SS_T - SS_t \cdots \cdots \cdots \text{处理内平方和} \end{cases}$$

(2) 自由度分解。

在计算总平方和时, 资料中的所有观测要受  $\sum_{i=1}^k \sum_{j=1}^n (x_{ij} - \bar{x}_{..}) = 0$  这一条件的约束, 故总自由度等于资料中观测的总个数减 1, 记为:  $df_T = nk - 1$ 。

在计算处理间平方和时, 各处理均数  $\bar{x}_i$  要受  $\sum_{i=1}^k (\bar{x}_i - \bar{x}_{..}) = 0$  这一条件的约束, 故处理间自由度为处理的个数减 1, 记为:  $df_t = k - 1$ 。

在计算处理内平方和时, 要受到  $k$  个条件的约束:  $\sum_{j=1}^n (x_{ij} - \bar{x}_i) = 0$  ( $i=1, 2, \cdots, k$ ), 故处理内自由度为资料中观测的总个数减  $k$ , 记为:  $df_e = kn - k = k(n - 1)$ 。

由:  $nk - 1 = (k - 1) + (nk - k) = (k - 1) + k(n - 1)$ , 得自由度的分解关系式:  $df_T = df_t + df_e$ 。

## 2. F 检验

用不同的平方和除以各自的自由度, 便得到如下的 3 种均方误差, 但总均方误差一般不等于处理间均方误差与处理内均方误差的和。

$$\begin{cases} MS_T = S_T^2 = SS_T / df_T \cdots \cdots \cdots \text{总均方误差} \\ MS_t = S_t^2 = SS_t / df_t \cdots \cdots \cdots \text{处理间均方误差} \\ MS_e = S_e^2 = SS_e / df_e \cdots \cdots \cdots \text{处理内均方误差} \end{cases}$$

由此可得  $F$  统计量:  $F = S_t^2 / S_e^2$ , 它服从自由度为  $df_1 = df_t = k - 1$  和  $df_2 = df_e = k(n - 1)$  的  $F$  分布。 $F$  分布的概率密度曲线是随两个自由度变化的一簇偏正态曲线, 如图 9-1 所示。

方差分析中  $F$  检验的目的在于推断各处理间的差异是否显著, 检验的零假设为: 各处理所得的数据之间无显著差异。若  $F < F_{0.05}$  (或  $P > 0.05$ ), 不能否定零假设; 若

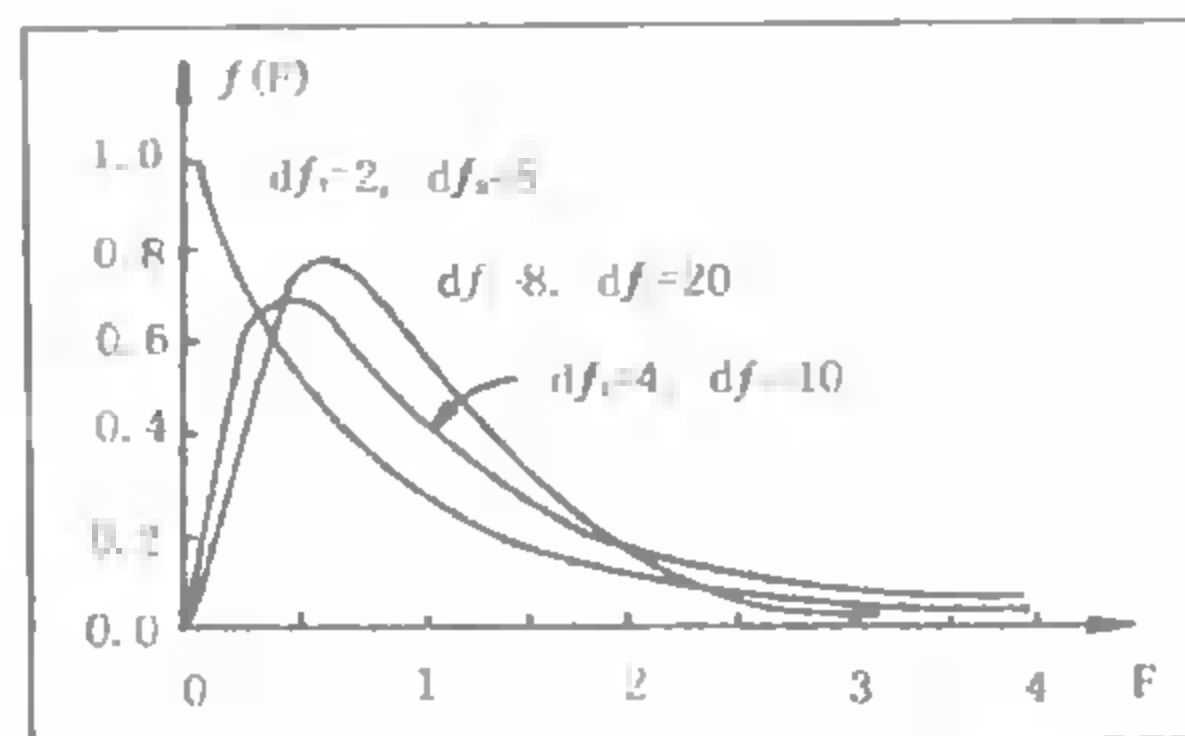


图 9-1  $F$  分布的概率密度曲线

$F_{0.05} \leq F < F_{0.01}$  (或  $0.01 < P \leq 0.05$ ), 否定零假设, 认为各处理的结果之间存在显著差异, 在结果中常以 “\*” 标记这种情况; 若  $F_{0.01} \leq F$  (或  $P \leq 0.01$ ), 否定零假设, 认为各处理结果之间存在极其显著的差异, 在结果中常以 “\*\*” 标记这种情况。

9.2 单因素方差分析

单因素方差分析也称作一维方差分析, 它可用于检验单个因素取不同水平时某因变量的均值是否有显著地变化, 还可进一步用于因变量均值的多重比较, 即在指定因素的若干取值水平中, 检验哪些水平下的试验结果具有区别于其他水平的显著差异。

9.2.1 原理与方法

第 9.1.2 节以单因素方差分析为例解释了方差分析的基本原理, 下面做以简单总结。

1. 假设和要求

- (1) 指定因素的单个处理水平下的样本是来自正态总体的。如果因变量的分布明显是非正态的, 应该选择使用非参数检验方法。
- (2) 各个处理水平间的样本方差相等, 这一点可通过方差齐性检验加以验证。
- (3) SPSS 还要求因素变量的取值必须为整数型的, 分析变量 (因变量) 须为数值型。

2. 计算公式小结

设某试验的指定因素可取  $k$  个处理水平, 每个水平下的观测数分别为  $n_1、n_2、\cdots、n_k$ , 进行方差分析时所用到的有关公式, 如表 9-2 所示。

表 9-2 组内观察值数目不等的单因素方差分析计算公式表

变 因	SS	DF	$S^2$	F
处理间	$SS_T = \sum_1^k n_i (\bar{x}_i - \bar{x})^2 = \sum_1^k \left( \frac{T_i}{n_i} \right)^2 - C$	$DF_T = k - 1$	$S_T^2 = \frac{SS_T^2}{DF_T}$	$F = S_T^2 / S_e^2$
误差	$SS_e = \sum_1^k \sum_1^{n_i} (x - \bar{x}_i)^2 = SS_T - SS_T$	$DF_e = \sum n_i - k$	$S_e^2 = \frac{SS_e^2}{DF_e}$	
总变异	$SS_T = \sum_1^{n_i} (x - \bar{x})^2 = \sum x^2 - C, C = \frac{T^2}{\sum n_i}$	$DF_T = \sum n_i - 1$		

9.2.2 单因素方差分析实例

1. 数据和问题描述

本节利用单因素方差分析, 来检验亩产量的多少是否与施肥量有关。所用数据文件为“施肥量与亩产量关系数据.asv”, 数据格式如图 9-2 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	prod	Numeric	8	2	亩产量	None	None	8	Right	Scale
2	sh	Numeric	8	2	施肥量	1(0), 2(1)	None	8	Right	Scale

图 9-2 施肥量与亩产量数据格式



## 2. SPSS 单因素方差分析的参数设置

依次单击菜单“Analyze→Compare Means→One-Way ANOVA...”执行单因素方差分析过程，其主设置界面如图 9-3 所示，在此选择进行分析的各种变量。

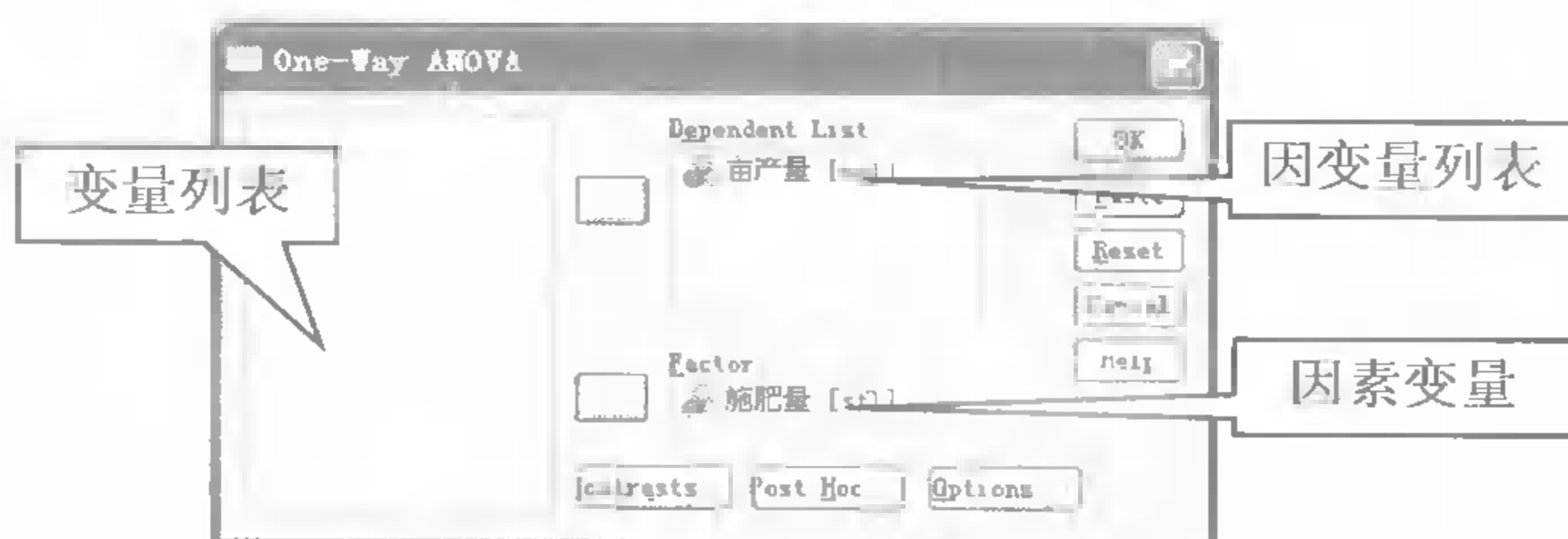




图 9-3 单因素方差分析的主设置界面

### (1) 变量设置。

在变量列表中单击选中亩产量，单击从上至下第一个  按钮，将其作为因变量选入 Dependent List 列表框；在变量列表中单击施肥量，单击从上至下第二个  按钮，将其作为因素变量选入 Factor 选框。

Dependent List 列表框中是因变量，且必须为数值型的连续变量，如果选入了多个变量，它们将分别对指定的因素变量作单因素方差分析。

Factor 选框中是因素变量，变量的取值需为整数。

### (2) 对照选项的设置。

单击图 9-3 中的 Contrasts (对比) 按钮，弹出如图 9-4 所示的对话框，在此设置关于均值对照的比较选项。单击选中 Polynomial (多项式) 复选框；在 Coefficients 后输入“1.2”，单击 Add 按钮将其加入下面的列表，用同样的方法加入“0”、“1”；单击 Continue 按钮返回主界面。

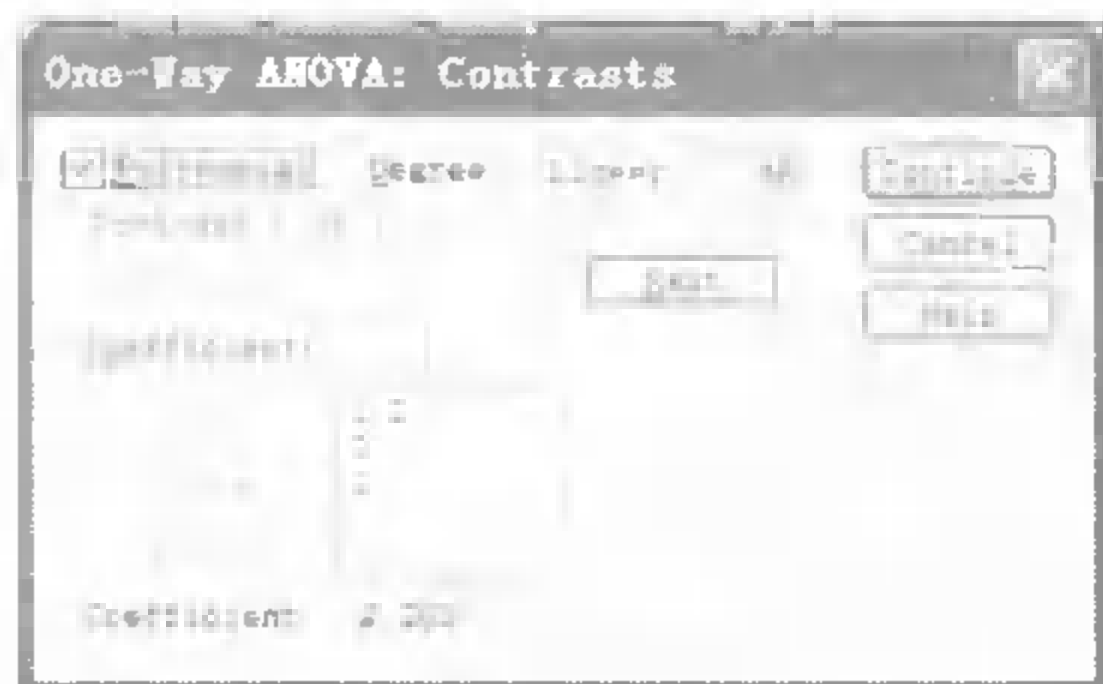


图 9-4 Contrasts 设置

① Polynomial 复选项。设定将组间方差平方和分解为何种形式的趋势成分。

Degree (次数) 下拉列表给出的可选项有：Linear (线性)、Quadratic (二次多项式)、Cubic (三次多项式)、4<sup>th</sup> (四次多项式)、5<sup>th</sup> (五次多项式)。SPSS 会在输出结果里给出指定阶次和低于指定阶次的各阶平方和分解结果，并给出各阶次的自由度、F 统计量和 F 检验结果。

② Contrast 1 of 1 栏。设置一组先验的对照值，由 SPSS 以 *t* 检验进行验证。

先在 Coefficients 输入框中指定一个系数，再单击 Add 按钮加入其下的列表框。因素变量有几个取值水平就输入几个系数；如果只比较第一水平与第四水平的均值，则必须把第二个、第三个系数输入为 0；如果只比较第一水平与第二水平的均值，则只需输入前两个系数。对于已加入列表的系数，可以通过单击 Change 按钮和 Remove 按钮进行修改或删除。

此处还可以建立多组对照关系，输入一组系数后，单击 Next 按钮，系数列表清空，进行新一组对照系数的设置。单击 Previous 或 Next 按钮，可以在不同的对照组之间切换和编辑。

注意：输入系数的顺序要和因素变量取值的升序相对应，列表里的第一个系数需要与最小的因素变量水平相对应。例如：因素变量有 5 个取值，对照系数为 1、0、0、0.5、0.5 时，表示进行第 1 水平和第 4、5 水平的比较。如图 9-4 所示，表示的对照比较为：“1.2\*Mean1-1\*Mean3”。

对应于  $t$  检验的零假设为：第一水平均值的 1.2 倍，与第三水平的均值无显著差异。

### (3) Options 选项设置。

单击图 9-3 中的 Options 按钮，弹出如图 9-5 所示的对话框，在此设置输出选项及缺失值的处理方式。依次勾选复选框 Descriptive、Homogeneity、Brown-Forsythe 和 Means plot。单击 Continue 按钮返回主界面。

下面来详细介绍各设置选项的含义。

#### ① Statistics 栏，选择输出哪些统计量。

- Descriptive，描述统计量（均值、标准差、标准误、最大值、最小值等）。
- Fixed and random effects，固定效应模式和随机效应模式的相关统计量（标准差、标准误和 95% 的置信区间）。
- Homogeneity-of-variance，方差齐次性检验结果（用 Levene test 检验）。
- Brown-Forsythe 和 Welch，这两个统计量用于检验各组均值是否相等，当方差齐性的假设不能确定时，它们要优于 F 统计量。

#### ② Means plot 复选框，输出均值分布图，即因变量均值对因素变量取值水平的折线图。

#### ③ Missing Value 栏，设置对缺失值的处理方法，可选项有 2 个。

- Exclude cases analysis by analysis 选项，剔除那些分析中用到的变量（因变量和因素变量）含缺失值的观测记录，并剔除因素变量取值超出指定范围的记录。
- Exclude cases listwise 选项，只要某条记录有一个因素变量或因变量含缺失值，就在所有分析中剔除这条观测记录。

### (4) Post Hoc 选项设置。

单击图 9-3 中的 Post Hoc 按钮，弹出如图 9-6 所示的对话框，在此指定多重比较方法。

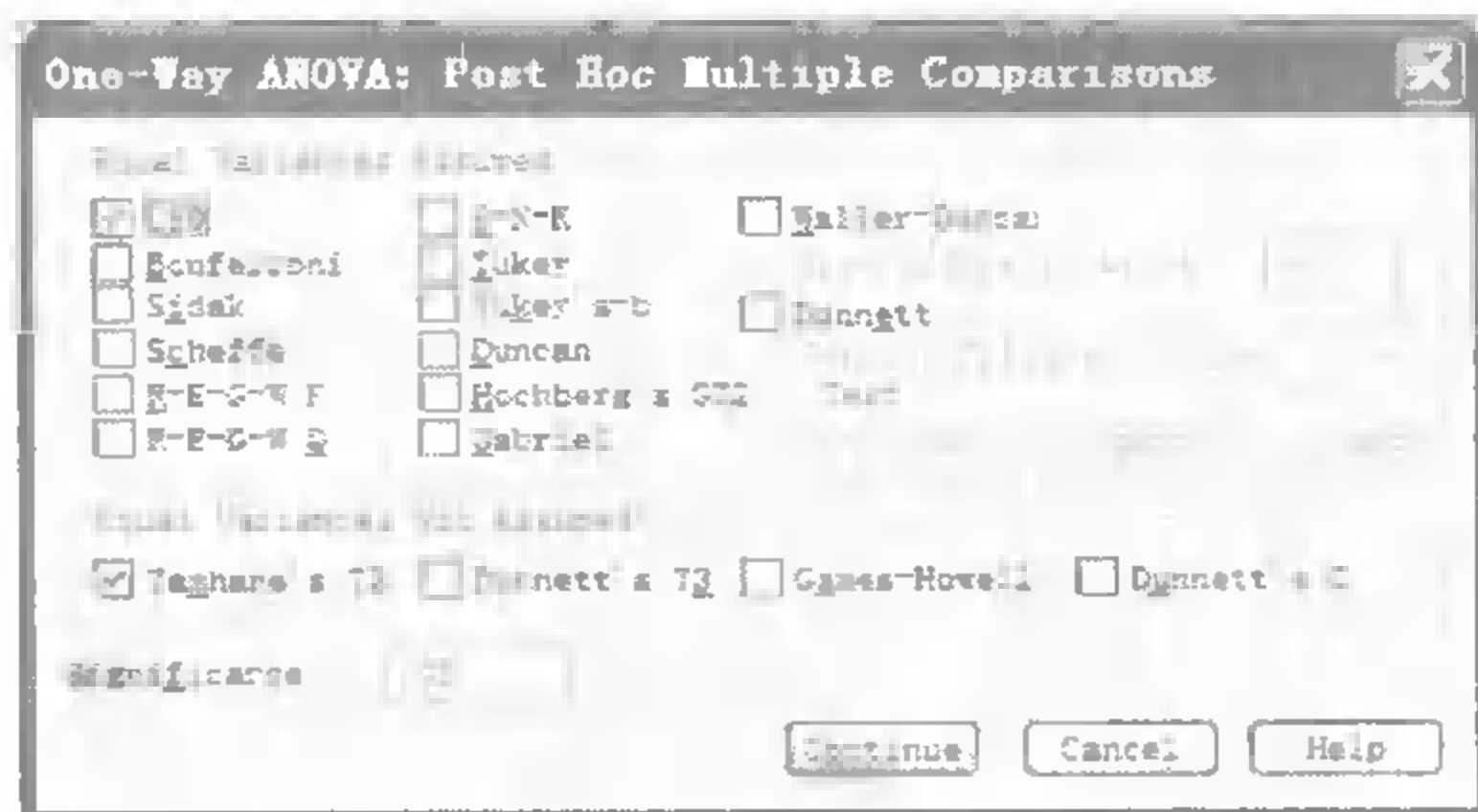


图 9-6 单因素方差分析的 Post Hoc 设置

勾选 LSD 复选框和 Tamhane's T2 复选框；单击 Continue 按钮返回主界面。

所谓多重比较，指的是对因素变量每两个取值水平下的均值做如下比较： $MEAN(j) - MEAN(i) \geq 4.6625 \times RANGE \times SORT(1/N(i) + 1/N(j))$ ，其中： $i, j$  代表不同的水平， $MEAN(i)$ 、 $MEAN(j)$  分别是第  $i, j$  水平下的均值， $N(i)$ 、 $N(j)$  分别为第  $i, j$  水平下的观测个数，对于不同的算法，RANGE 值也不相同。

#### ① 方差具有齐次性时（Equal variance assumed），有以下几个可选方法。

- Least-significant difference (LSD)，用  $t$  检验完成各组均值间的配对比较，对多重比较误差率不进行调整。
- Bonferroni (LSDMOD)，用  $t$  检验完成各组间均值的配对比较，但通过设置每个检验的误差率来控制整个误差率。
- Sidak，基于  $t$  统计量进行多重配对比较，可以调整显著性水平，比 Bonferroni 方法的界限要小。
- Scheffe，使用样本的 F 分布，对所有可能的均值组合进行同步的配对比较，还可检

验分组均值的所有线性组合，而不仅仅是配对比较。

- R-E-G-W-F(Ryan-Einot-Gabriel-Welsch F)，基于 F 检验的多重比较。
  - R-E-G-W Q(Ryan-Einot-Gabriel-Welsch Q)，基于学生化范围的多重比较。
  - S-N-K(Student Newman Keuls)，用学生化范围分布 (Studentized range distribution) 进行组间均值的配对比较。如果各组的样本含量相等，还会逐步比较同类子集的均值。各组的均值从大到小按顺序排列，最先比较差异最大的组对。
  - TUKEY(Tukey's honestly significant difference)，用学生化范围统计量 (Studentized range statistic) 进行组间均值的配对比较。用所有配对比较的误差率估计实验误差率。
  - Tukey's-b，用学生化范围分布进行组间均值的配对比较，其值实际上就是 Tukey's 检验和 Student-Newman-Keuls 检验统计量的均值。
  - Duncan(Duncan's multiple range test)，用与 SNK 检验相似的逐步过程进行多重比较，但会设置对所有检验误差率的保护水平。使用的是学生化范围统计量。
  - Hochberg's GT2，用学生化最大系数 (Studentized maximum modulus) 进行多重比较，类似于 Tukey 检验。
  - Gabriel，用学生化最大系数进行配对比较。当单元格频数不等时，比 Hochberg's GT2 检验法更为有效。
  - Waller-Duncan，用  $t$  统计量进行多重比较检验，使用的是贝叶斯方法。
  - Dunnett，此方法指定一个控制组，其他组都与控制组进行多重配对  $t$  检验。选中后激活下面的选项：Control Category 下拉列表，指定控制组为 First (第一组) 或 Last (最后一组)；Test 栏，指定检验方式为 “2-sided” 双边检验、“<Control” 单边检验或 “>Control” 单边检验。
- ② 方差不具有齐次性时 (Equal variance not assumed)，有以下几个可选方法。
- Tamhane's T2，基于  $t$  检验的保守的配对比较。
  - Dunnett's T3，基于学生化最大系数 (Studentized maximum modulus) 的配对比较。
  - Games Howell，方差不齐时的一种比较灵活的配对比较。
  - Dunnett's C，基于学生化范围 (Studentized range) 的配对比较。
- ③ Significance 输入框，用于指定各种检验的显著性水平。

### 3. 案例的结果分析

单击图 9-3 中的 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 9-7~图 9-11 所示。

描述								
亩产量								
	N	均值	标准差	标准误	均值的 95% 置信区间		极小值	极大值
					下限	上限		
较少	5	874.1667	11.80537	4.81952	861.7777	886.5556	859.00	891.00
一般	5	952.6667	25.57082	10.43924	925.8317	979.5016	921.00	986.00
较多	5	962.8333	19.54908	7.98088	942.3178	983.3488	938.00	985.00
总数	15	929.8889	44.80794	10.56133	907.6064	952.1714	859.00	986.00

方差齐性检验			
亩产量			
Levene 统计量	df1	df2	显著性
3.009	2	15	.080

图 9-7 单因素方差分析的基本统计量输出

ANOVA						
亩产量			方差不齐时的 F 统计量检验结果			
		平方和	df	均方	F	显著性
组间	(组合)	28254.178	2	14127.389	36.058	.000
	线性项	23585.333	1	23585.333	60.197	.000
	对比	4669.444	1	4669.444	11.918	.004
组内		5877.900	15	391.860		
总数		34131.778	17			

均值相等性的健壮性检验				
Brown-Forsythe 统计量检验结果				
亩产量				
	统计量 <sup>a</sup>	df1	df2	显著性
Brown-Forsythe	36.058	2	11.649	.000

<sup>a</sup> 渐近 F 分布。

图 9-8 单因素方差分析的检验结果

对比系数			
对比	施肥量		
	较少	一般	较多
1	1	2	1

对比检验						
亩产量	对比	对比值	标准误	t	df	显著性 (双侧)
假设方差相等	1	2011.8333 <sup>a</sup>	12.62268	159.182	15	.000
	不假设等方差	2011.8333 <sup>a</sup>	9.85509	204.121	9.116	.000

<sup>a</sup> 对比系数总和不为零。

图 9-9 单因素方差分析的对照比较输出

多重比较								
因变量 亩产量								
	Ⅲ:施肥量	Ⅱ:施肥量	均值差 (Ⅲ-Ⅱ)	标准误	显著性	95% 置信区间		
LSD	较少	一般	-78.50000 <sup>a</sup>	11.42804	.000	-102.8583	-54.1417	
		较多	-88.66667 <sup>a</sup>	11.42804	.000	-113.0249	-64.3084	
	一般	较少	78.50000 <sup>a</sup>	11.42804	.000	54.1417	102.8583	
		较多	-10.16667	11.42804	.388	-34.5249	14.1916	
Tamhane	较少	一般	-78.50000 <sup>a</sup>	11.49807	.001	-114.2633	-42.7367	
		较多	-88.66667 <sup>a</sup>	9.32321	.000	-116.4935	-60.8399	
	一般	较少	78.50000 <sup>a</sup>	11.49807	.001	42.7367	114.2633	
		较多	-10.16667	13.14048	.841	-48.2570	27.9237	
	较多	较少	88.66667 <sup>a</sup>	9.32721	.000	60.8399	116.4935	
		一般	10.16667	13.14048	.841	-27.9237	48.2570	

<sup>a</sup> 均值差的显著性水平为 .05。

图 9-10 单因素方差分析的多重比较输出

(1) 描述性统计量和方差齐性检验结果。如图 9-7 所示,“描述”表格给出了施肥量不同时的分组样本的均值、标准差等统计量。

“方差齐性检验”表格是 Levene test 检验的输出,由  $0.05 < \text{显著性} (0.08) < 0.10$  推断,在 0.05 的显著性水平下可认为各组的方差无显著差异,但是在 0.10 的显著性水平下可认为各组的方差是有差异的,换句话说,我们能以 90%但不能以 95%的概率承认组间方差不齐,鉴于这点细微的差别,随后应该综合考虑方差齐时和方差不齐时的检验结果。

(2) 方差分析结果。图 9-8 所示是方差分析的输出结果,因 F 检验和 Brown-Forsythe 检验的显著性都小于 0.01,故可以推断施肥量不同时,亩产量的均值是有显著差异的。

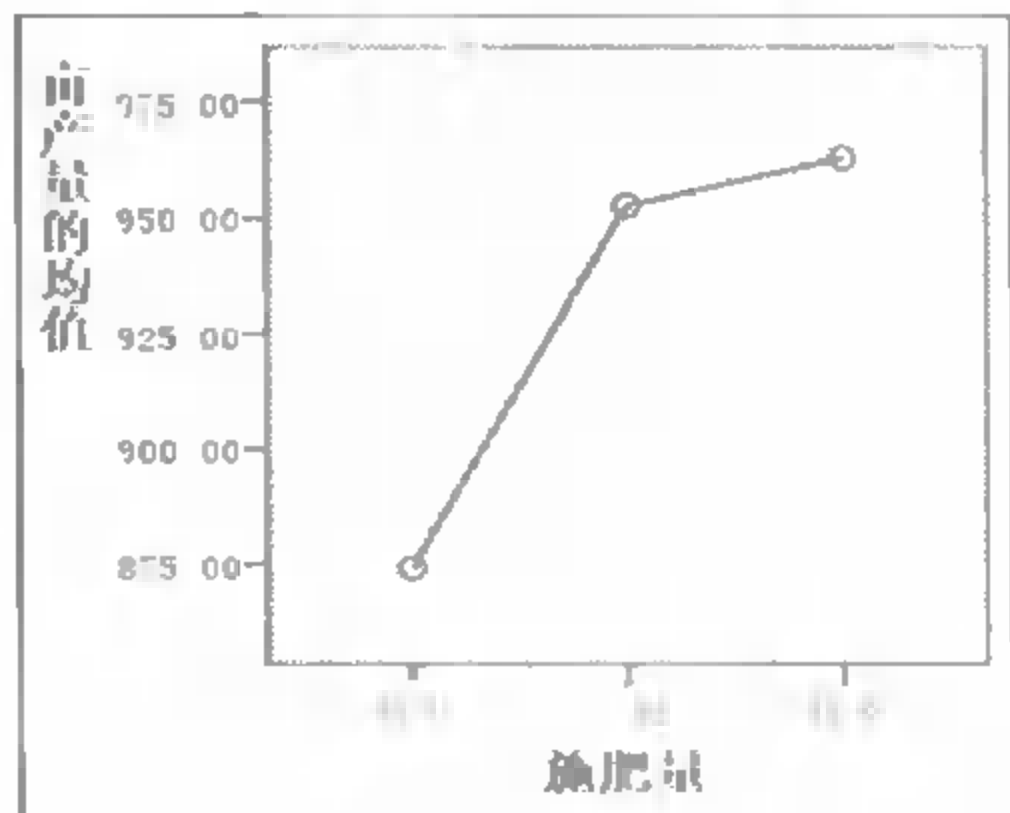


图 9-11 均值折线图



(3) 对照比较结果。图 9-9 所示是对照比较的输出结果。本例比较在施肥较少与施肥较多这两种情况下(施肥量一般时的系数为 0, 不参与比较), 亩产量均值是否服从 1.2:1 的比例。方差齐与不齐时, 检验的双侧显著性都远小于 0.01, 故能够显著的否定 1.2:1 这个先验的假设比例。

(4) 多重比较结果。图 9-10 所示是关于多重比较的输出结果, 分为方差齐与方差不齐两部分。

本例综合考虑方差齐与不齐时的情况, 根据图中标识的显著性(带\*号标识), 推断在 0.05 的显著性水平下, 施肥量较少与一般比较时、施肥量较少与较多比较时, 亩产量的均值是有显著差异的; 而施肥量在一般与较多比较时, 亩产量的均值是没有显著差异的。

(5) 均值折线图。图 9-11 所示是亩产均值对施肥量的折线图, 通过观察可以发现, 各组均值的分布与多重比较的结果是一致的。

### 9.3 多因素方差分析过程

SPSS 的 Univariate 过程, 可以对完全随机设计资料、配伍设计资料、析因设计资料、正交设计资料等进行多因素方差分析或协方差分析。输出的分析结果包括: 描述性统计量、参数估计值、对照系数矩阵、方差齐次性检验结果、水平散点图、残差图等等, 还可以选择执行多项式比较、均值多重比较等功能。

#### 9.3.1 原理与方法

多因素方差分析, 用于研究一个因变量是否受多个自变量(也称为因素)的影响, 它检验多个因素取值水平的不同组合之间, 因变量的均值是否存在显著的差异。多因素方差分析既可以分析单个因素的作用(主效应), 也可以分析因素之间的交互作用(交互效应), 还可以进行协方差分析, 以及各因素变量与协变量之间的交互作用。

##### 1. 假设和要求

多因素方差分析要求因变量是从多元正态总体中随机抽样得来的, 且要求总体中各分组的方差相同, 这一点可以通过方差齐次性检验来验证。

对于 SPSS 的多因素方差分析过程, 因变量和协变量必须是数值型变量, 且它们之间存在着相关关系; 因素变量需是分类变量, 可以是数值型或短字符型(不超过 8 个字符)。固定因素(Fixed Factor)是需要试验加以处理的因素, 随机因素(Random Factor)是随机设置的因素, 它们对实验结果的作用大小可以通过方差成分分析确定。

##### 2. 无交互影响的两因素方差分析

当有两个无交互影响的因素时, 通常采用不重复试验, 即对于两因素取值水平的每种组合只做一次试验。假设试验要考察两个因素 A 和 B, A 因素有 a 个水平, B 因素有 b 个水平, 两者交叉搭配形成 a×b 个水平组合(处理)。这两个因素在试验中处于平等地位, 将试验单位分成 a×b 个组, 每组随机接受一种处理, 每种处理取一个观测值。其数据模式如表 9-3 所示, 其中:

$$x_{i.} = \sum_{j=1}^b x_{ij}, \bar{x}_{i.} = \frac{1}{b} \sum_{j=1}^b x_{ij}, x_{.j} = \sum_{i=1}^a x_{ij}, \bar{x}_{.j} = \frac{1}{a} \sum_{i=1}^a x_{ij}, x_{..} = \sum_{i=1}^a \sum_{j=1}^b x_{ij}, \bar{x}_{..} = \sum_{i=1}^a \sum_{j=1}^b x_{ij} / ab$$

表 9-3 两因素无重复试验数据模式

A 因素	B 因素						合计 $x_{i.}$	平均 $\bar{x}_{i.}$
	$B_1$	$B_2$	.....	$B_j$	.....	$B_b$		
$A_1$	$x_{11}$	$x_{12}$	.....	$x_{1j}$	.....	$x_{1b}$	$x_{1.}$	$\bar{x}_{1.}$
$A_2$	$x_{21}$	$x_{22}$	.....	$x_{2j}$	.....	$x_{2b}$	$x_{2.}$	$\bar{x}_{2.}$
$\vdots$	$\vdots$	$\vdots$	.....	$\vdots$	.....	$\vdots$	$\vdots$	$\vdots$
$A_i$	$x_{i1}$	$x_{i2}$	.....	$x_{ij}$	.....	$x_{ib}$	$x_{i.}$	$\bar{x}_{i.}$
$\vdots$	$\vdots$	$\vdots$	.....	$\vdots$	.....	$\vdots$	$\vdots$	$\vdots$
$A_u$	$x_{u1}$	$x_{u2}$	.....	$x_{uj}$	.....	$x_{ub}$	$x_{u.}$	$\bar{x}_{u.}$
合计 $x_{.j}$	$x_{.1}$	$x_{.2}$	.....	$x_{.j}$	.....	$x_{.b}$	$x_{..}$	$\bar{x}_{..}$
平均 $\bar{x}_{.j}$	$\bar{x}_{.1}$	$\bar{x}_{.2}$	.....	$\bar{x}_{.j}$	.....	$\bar{x}_{.b}$		

与单因素方差分析类似，通过总平方和分解与自由度分解，可得如表 9-4 所示的两因素无交叉效应方差分析表，利用此表进行 F 检验，就可以推断因素 A、B 的效应是否是显著的。

表 9-4 两因素无交叉效应方差分析表

变异来源	SS	DF	$S^2$	F	$S_x$
A 因素	$SS_A = b \sum_{i=1}^a (\bar{x}_{i.} - \bar{x}_{..})^2 = \frac{1}{b} \sum_{i=1}^a x_{i.}^2 - C$	$a - 1$	$SS_A / DF_A$	$S_A^2 / S_e^2$	$\sqrt{S_e^2 / b}$
B 因素	$SS_B = a \sum_{j=1}^b (\bar{x}_{.j} - \bar{x}_{..})^2 = \frac{1}{a} \sum_{j=1}^b x_{.j}^2 - C$	$b - 1$	$SS_B / DF_B$	$S_B^2 / S_e^2$	$\sqrt{S_e^2 / a}$
误差	$SS_e = SS_T - SS_A - SS_B$	$(a - 1)(b - 1)$	$SS_e / DF_e$		
总变异	$SS_T = \sum_{i=1}^a \sum_{j=1}^b (x_{ij} - \bar{x}_{..})^2 = \sum_{i=1}^a \sum_{j=1}^b x_{ij}^2 - C$	$ab - 1$		其中: $C = x_{..}^2 / ab$	

3. 有交互影响的两因素方差分析

在两因素方差分析中，当两个因素之间存在着交互作用时，需要对因素取值的每种水平组合进行多次重复试验，以便将因素之间交互作用的平方和从误差平方和中分离出来。但是有重复试验的数据量就大大增加了，其数据模式如表 9-5 所示，其中：

$$x_{ij.} = \sum_{l=1}^n x_{ijl}$$
$$x_{i..} = \sum_{j=1}^b \sum_{l=1}^n x_{ijl}$$
$$x_{.j.} = \sum_{i=1}^a \sum_{l=1}^n x_{ijl}$$
$$x_{...} = \sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n x_{ijl}$$

$$\bar{x}_{ij.} = \sum_{l=1}^n x_{ijl} / n$$
$$\bar{x}_{i..} = \sum_{j=1}^b \sum_{l=1}^n x_{ijl} / bn$$
$$\bar{x}_{.j.} = \sum_{i=1}^a \sum_{l=1}^n x_{ijl} / an$$
$$\bar{x}_{...} = \sum_{i=1}^a \sum_{j=1}^b \sum_{l=1}^n x_{ijl} / abn$$

表 9-5 两因素有重复试验数据模式

A 因素		B 因素				$A_i$ 合计 $x_{i.}$	$A_i$ 平均 $\bar{x}_i$
		$B_1$	$B_2$	.....	$B_b$		
$A_1$	$x_{1j}$	$x_{111}$	$x_{121}$	.....	$x_{1b1}$	$x_{1.}$	$\bar{x}_{1..}$
		$x_{112}$	$x_{122}$	.....	$x_{1b2}$		
		$\vdots$	$\vdots$	$\vdots$	$\vdots$		
		$x_{11n}$	$x_{12n}$	.....	$x_{1bn}$		
	$\bar{x}_{1.}$	$\bar{x}_{11.}$	$\bar{x}_{12.}$	.....	$\bar{x}_{1b.}$		
$A_2$	$x_{2j}$	$x_{211}$	$x_{221}$	.....	$x_{2b1}$	$x_{2.}$	$\bar{x}_{2..}$
		$x_{212}$	$x_{222}$	.....	$x_{2b2}$		
		$\vdots$	$\vdots$	$\vdots$	$\vdots$		
		$x_{21n}$	$x_{22n}$	.....	$x_{2bn}$		
	$\bar{x}_{2.}$	$\bar{x}_{21.}$	$\bar{x}_{22.}$	.....	$\bar{x}_{2b.}$		
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$A_a$	$x_{aj}$	$x_{a11}$	$x_{a21}$	.....	$x_{ab1}$	$x_{a.}$	$\bar{x}_{a..}$
		$x_{a12}$	$x_{a22}$	.....	$x_{ab2}$		
		$\vdots$	$\vdots$	$\vdots$	$\vdots$		
		$x_{a1n}$	$x_{a2n}$	.....	$x_{abn}$		
	$\bar{x}_{a.}$	$\bar{x}_{a1.}$	$\bar{x}_{a2.}$	.....	$\bar{x}_{ab.}$		
$B_i$ 合计 $x_{.i}$		$x_{.1.}$	$x_{.2.}$	.....	$x_{.b.}$	$x_{..}$	
$B_i$ 平均 $\bar{x}_{.j}$		$\bar{x}_{.1.}$	$\bar{x}_{.2.}$	.....	$\bar{x}_{.b.}$		$\bar{x}_{...}$

与单因素方差分析类似，通过总平方和分解与自由度分解，可得如表 9-6 所示的两因素有交叉效应方差分析表。此时，交叉作用的分析非常重要，通常首先由  $F = s_{AB}^2 / s_e^2$  检验交互作用的显著性；如果交互作用不显著，就对因素 A、B 主效应的显著性进行检验；如果交互作用是显著的，对 A、B 主效应的检验就不太重要了。

表 9-6 两因素有交叉效应方差分析表

变异来源	SS	DF	$S^2$	F	$S_x$
A、B 因素组合	$SS_{AB} = \frac{1}{b} \sum x_{.j}^2 - C$	$df_{AB} = ab - 1$	$S_{AB}^2$	$S_{AB}^2 / S_e^2$	
A 因素	$SS_A = \frac{1}{bn} \sum x_{i.}^2 - C$	$df_A = a - 1$	$S_A^2$	$S_A^2 / S_e^2$	$\sqrt{S_e^2 / bn}$
B 因素	$SS_B = \frac{1}{an} \sum x_{.j}^2 - C$	$df_B = b - 1$	$S_B^2$	$S_B^2 / S_e^2$	$\sqrt{S_e^2 / an}$
A×B 交互	$SS_{A \times B} = SS_{AB} - SS_A - SS_B$	$df_{A \times B} = (a - 1)(b - 1)$	$S_{AB}^2$	$S_{AB}^2 / S_e^2$	$\sqrt{S_e^2 / n}$
试验误差	$SS_e = SS_T - SS_{AB}$	$df_e = ab(n - 1)$	$S_e^2$		
总变异	$SS_T = \sum \sum \sum x_{ijl}^2 - C$	$df_T = abn - 1$		其中： $C = x_{..}^2 / abn$	

9.3.2 二因素方差分析实例

1. 数据和问题描述

本节利用二因素方差分析，来检验不同的包装方式和柜台种类对超市销售额是否有显著影响，所用数据文件为“包装与柜台对销售额影响数据.sav”，数据格式如图 9-12 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	xse	Numeric	8	2	销售额	None	None	8	Right	Scale
2	bzfs	Numeric	8	2	包装方式	None	None	8	Right	Scale
3	gtzl	Numeric	6	2	柜台种类	None	None	6	Right	Scale

图 9-12 包装与柜台对销售额影响的数据格式

2. SPSS 实例操作

依次单击菜单“Analyze→General Linear Model→Univariate...”执行多因素方差分析过程，其主设置面板如图 9-13 所示，在此选择分析的因变量和因素变量。

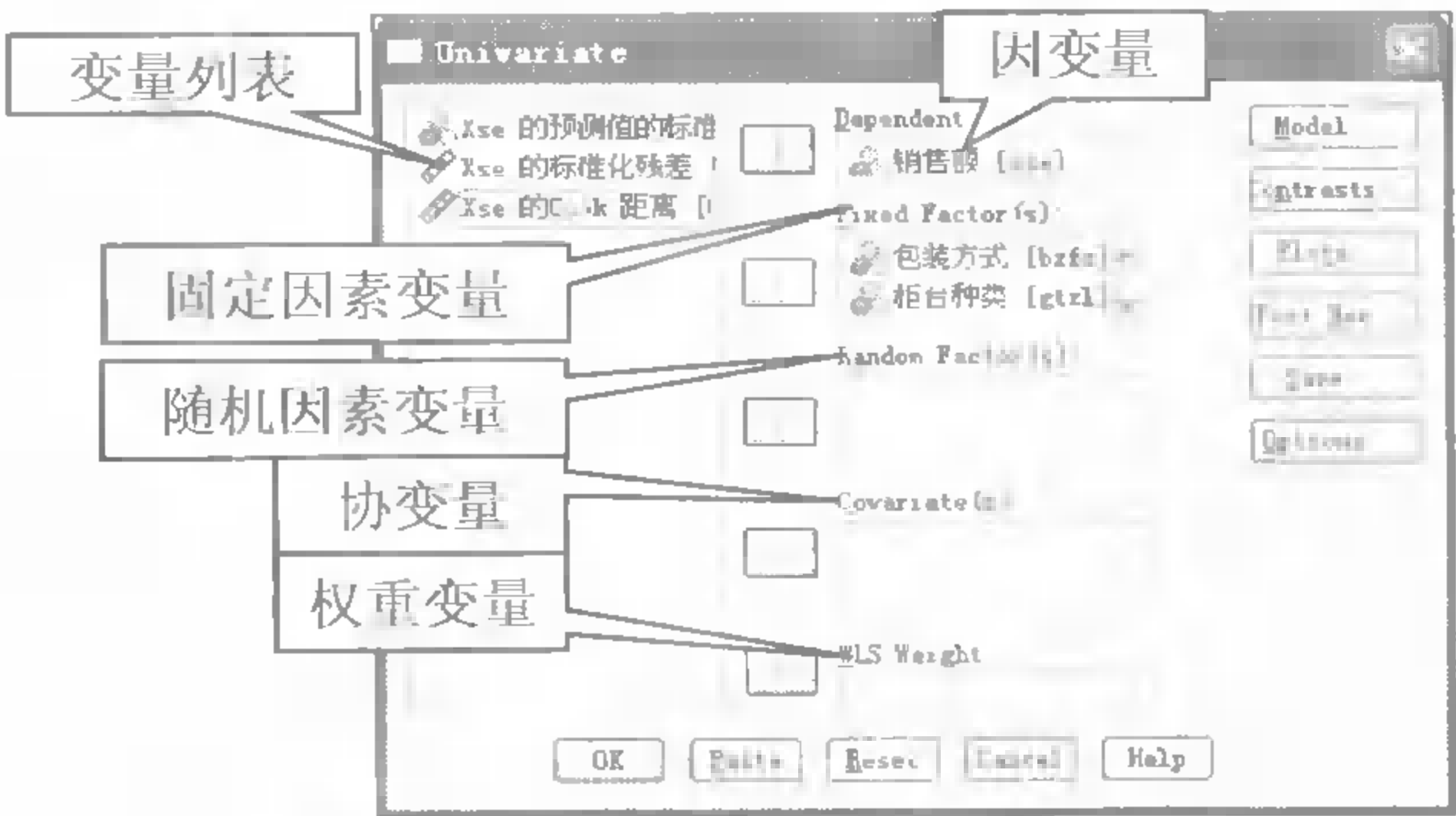






图 9-13 多因素方差分析的主界面

（1）变量设置。在变量列表单击选中销售额变量，单击从上至下第一个  按钮，将其作为因变量选入 Dependent 选框；在变量列表选中包装方式和柜台种类变量，单击从上至下第二个  按钮，将其作为因素变量选入 Fixed 列表框。

-  Dependent 选框，用于从变量列表选入待分析的因变量；Fixed Factor(s)列表框，用于从变量列表选入固定因素变量；Random Factor(s)列表框，用于从变量列表选入随机因素变量；Covariate(s)列表框，用于从变量列表选入协变量。
-  WLS Weight 选框，从变量列表选入权重变量，用于加权的最小平方分析。权重变量可用于为观测记录赋以不同的权重，也可用于给不同的测量精度以适当的补偿。

（2）对照选项设置。在图 9-13 中，单击 Contrasts 按钮，弹出如图 9-14 所示的对话框，在此设置对照比较的选项。在 Factors 列表选中 bzfs（包装方式）和 gtzl（柜台种类）变量，单击 Contrast 下拉列表选中 Simple 选项，再单击 Change 按钮确认更改。单击 Continue 按钮返回主界面。

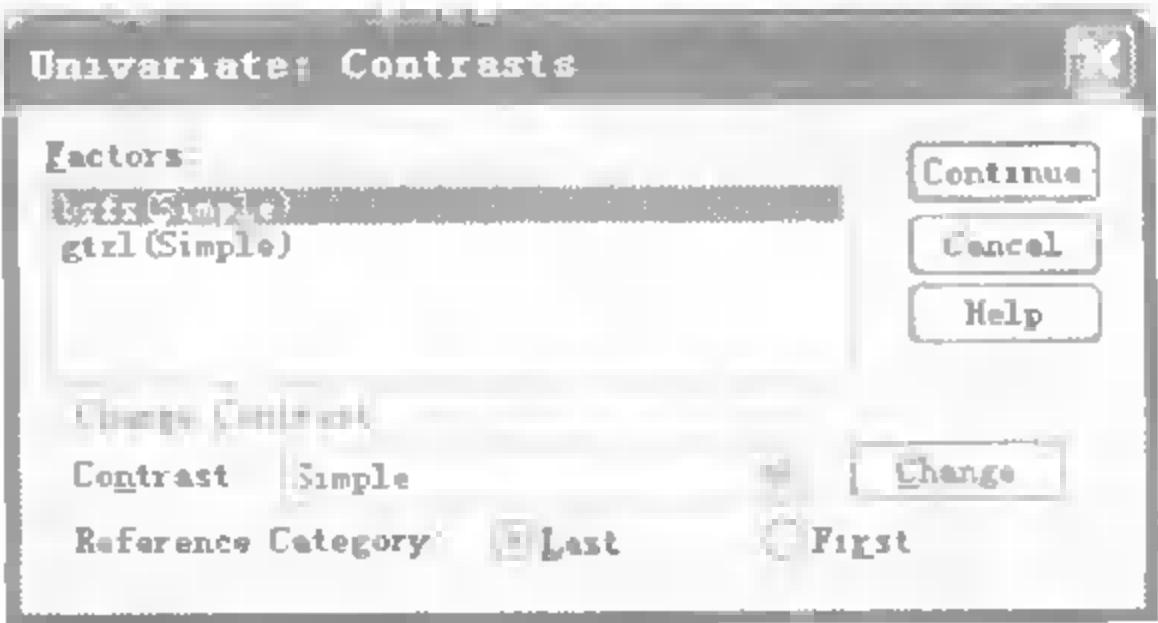


图 9-14 多因素方差分析的对照设置



① **Factors** 列表框，显示在主界面选入的因素变量，变量名后的括号是当前的对比方法。

② **Contrast** 下拉列表，指定对照比较的方法，可选项有如下 7 个。

- ① **None**，不进行均数比较。
- ② **Deviation** 差别比较，因素变量的每个分类（参考分类除外）都与总体效应进行比较。
- ③ **Simple** 简单比较，因素变量的每个分类（参考分类除外）都与参考分类进行比较。
- ④ **Difference** 差分比较，除第 1 类外，因素变量的每个分类都与其前所有分类的平均效应进行比较，也叫逆 **Helmert** 比较。
- ⑤ **Helmert** (**Helmert** 比较)，除最后 1 类外，因素变量的每个分类都与后面所有分类的平均效应进行比较。
- ⑥ **Repeated** 重复比较，除第 1 类外，因素变量的每个分类都与其前的所有类别进行比较。
- ⑦ **Polynomial** 多项式比较，此方法假设各类别的间距相等，仅适用于数值型变量。

③ **Reference** 栏用来指定参考分类，如果选择了 **Deviation** 或 **Simple** 方法，就需要指定一个参考类别，可选项有：**First**，第 1 类；**Last**，最后 1 类（默认的参考类），修改后需要单击 **Change** 按钮确认。

(3) 模型设置。在图 9-13 中，单击 **Model** 按钮，弹出如图 9-15 所示的模型设置对话框。单击 **Continue** 按钮返回主界面。

① **Specify Model** 栏，指定使用模型的类别，有如下两个选项：**Full factorial** 默认选项，指定模型包括所有因素变量的主效应、所有协变量的主效应、所有因素之间的交互效应，不包括协变量与其他因素的交互效应；**Custom** 自定义选项，选中后激活下面的设置内容。

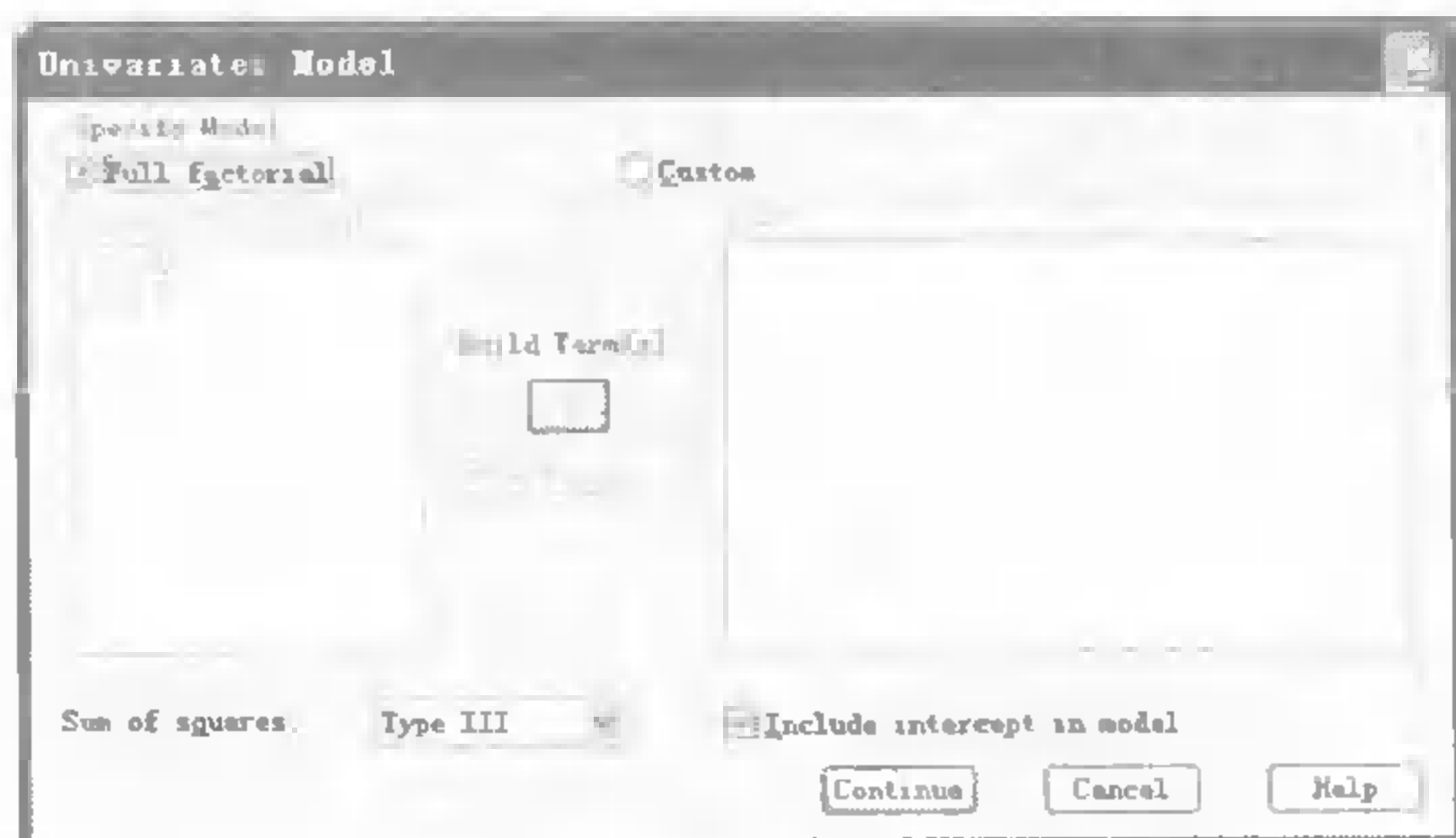




图 9-15 多因素方差分析的模型设置

② **Factors and Covariates** 列表框，显示因素变量的变量名，固定因素变量、随机因素变量、协变量的变量名后分别用括号扩住的字符 **F**、**R**、**C** 加以区分。

③ 选入指定效应的方法。

选择模型主效应：在 **Factors and Covariates** 列表选中某些因素变量，单击  按钮，将选中的变量选入 **Model** 列表框；选择模型交互效应：在 **Factors and Covariates** 列表选中某些因素变量，单击 **Build Term(s)** 下拉列表指定一种交互方式，再单击  按钮，将选中变量的指定效应选入 **Model** 列表框，选入的交互效应项在各因素变量之间以 “\*” 连接，例如 “**device\*light**”。

**Build Term(s)** 下拉列表中可选的交互效应有：**Main effects**（主效应）、**Interaction**（交互效应）、**All n-Way**（所有  $n$  维交互效应）。

④ **Sum of squares** 下拉列表，指定方差平方和分解的方法，可选项有如下 4 个。

- ① **Type I**，平方和的等级分解，适用于如下两类模型：平衡的 **ANOVA** 模型，要求在指定一阶交互效应前指定主效应，在指定二阶交互效应前指定一阶交互效应，依次

类推；嵌套模型，要求一阶效应嵌套在二阶效应里，二阶效应嵌套在三阶效应里，嵌套的形式只能使用命令语句指定。

- Type II，适用于：平衡的 ANOVA 模型、主因子效应模型、回归模型、嵌套设计。
- Type III，默认方法，它的优势是与单元格的频数无关，适用于：Type I、Type II 所列的模型、没有空单元格的平衡和不平衡模型。
- Type IV，此方法是为有缺失单元格的情况设计的。对于某效应 A，如果 A 不包含在其他效应里，则  $Type IV = Type III = Type II$ ；如果 A 包含在其他效应里，Type IV 把对 A 中参数的对照公正地分散于较高水平的效应里。适用于：Type I、Type II 所列的模型、有空单元格的平衡和不平衡模型。

⑤ Include Intercept in model 复选框，表示在模型中包含截距项，默认为选中状态。

(4) 图形选项设置。单击图 9-13 中的 Plots 按钮，弹出如图 9-16 所示的对话框，在此设置关于输出图形的选项。在变量列表单击选中 bzfs (包装方式)，单击从上至下第一个 ☐ 按钮，将其作为横轴变量选入 Horizontal 选框；在变量列表单击选中 gtzl (柜台种类)，单击从上至下第二个 ☐ 按钮，将其作为分线变量选入 Separate Lines 选框；单击 Add 按钮将 “bzfs\*gtzl” 选入下面的列表框；单击 Continue 按钮返回主界面。

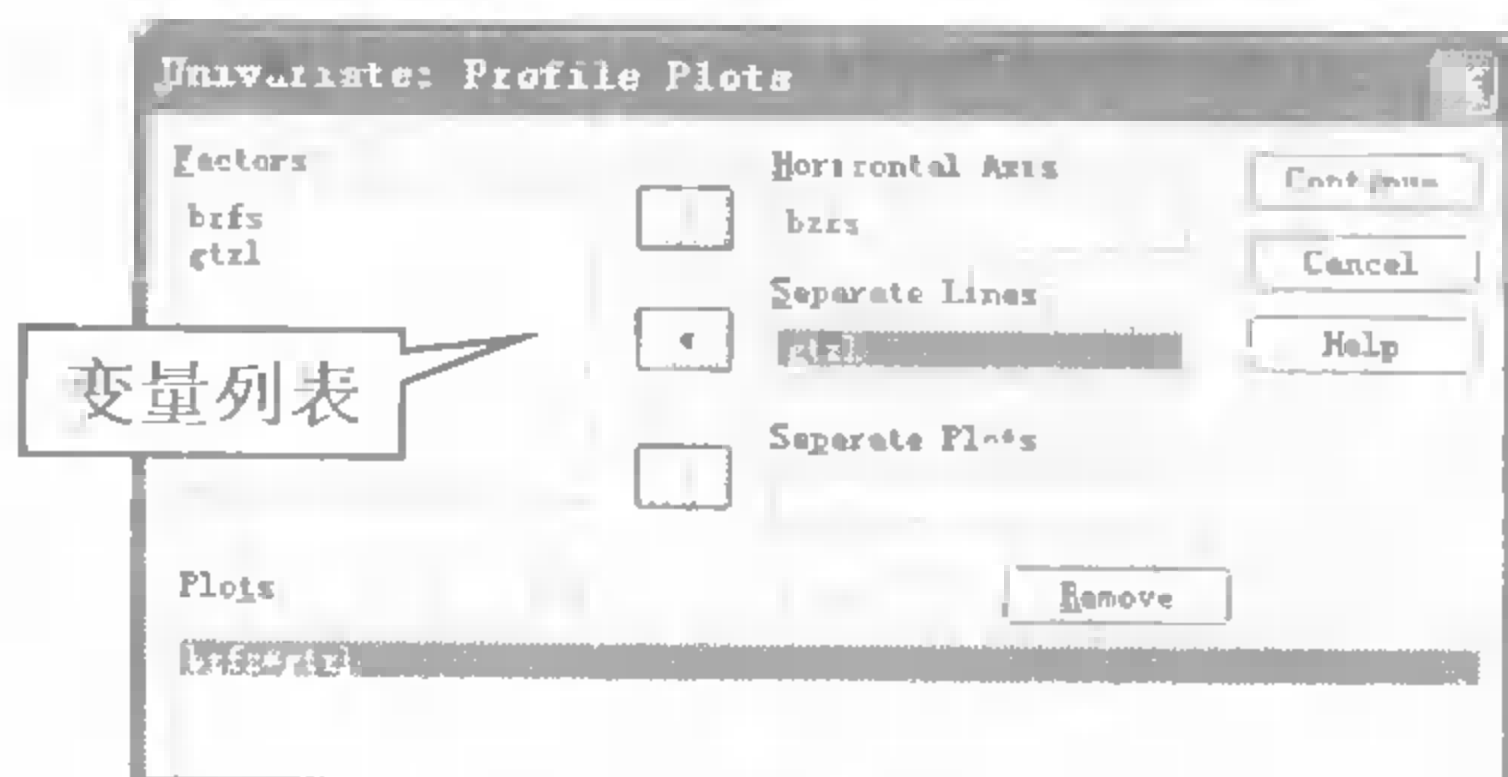


图 9-16 多因素方差分析的作图设置

边际均值散点图 (Profile)，是指以某个因素变量为横轴、因变量边际均值的估计值为纵轴所作的图形；如果指定了协变量，这里的均值就是经过协变量调整后的均值。在单因素方差分析里，边际图用来表现指定因素各个水平的因变量均值；在多因素边际均值图里，相互平行的线表明在相应因素之间无交互效应，反之亦然。

- Factors 列表框，显示在主对话框中所选的因素变量名称。
- Horizontal Axis 水平坐标选框，用于从 Factors 列表选入因素变量。
- Separate Lines 分线变量选框，用于从 Factors 列表选入一个因素变量，对它的每个取值水平，在图中单独作一条直线；Separate Plots 分图变量选框，用于从 Factors 列表选入一个因素变量，对它的每个取值水平，分别输出一个图形。
- 设置方法。水平坐标、分线变量、分图变量不一定同时选择，但指定顺序必须是从上至下依次选择的（否则 Add 按钮不可用）。单击 Add 按钮，将指定作图模型选入下面的 Plots 列表框，此处可能出现 3 种作图模式：单因素边际均值图（显示一个因素变量名）、两因素分线图（用 “\*” 连接的两个因素变量名）、三因素边际图（用 “\*” 连接的三个因素变量名）。在 Plots 列表框选中某个作图模型，单击 Remove 按钮可将其删除；或作一定更改后单击 Change 按钮确认修改。

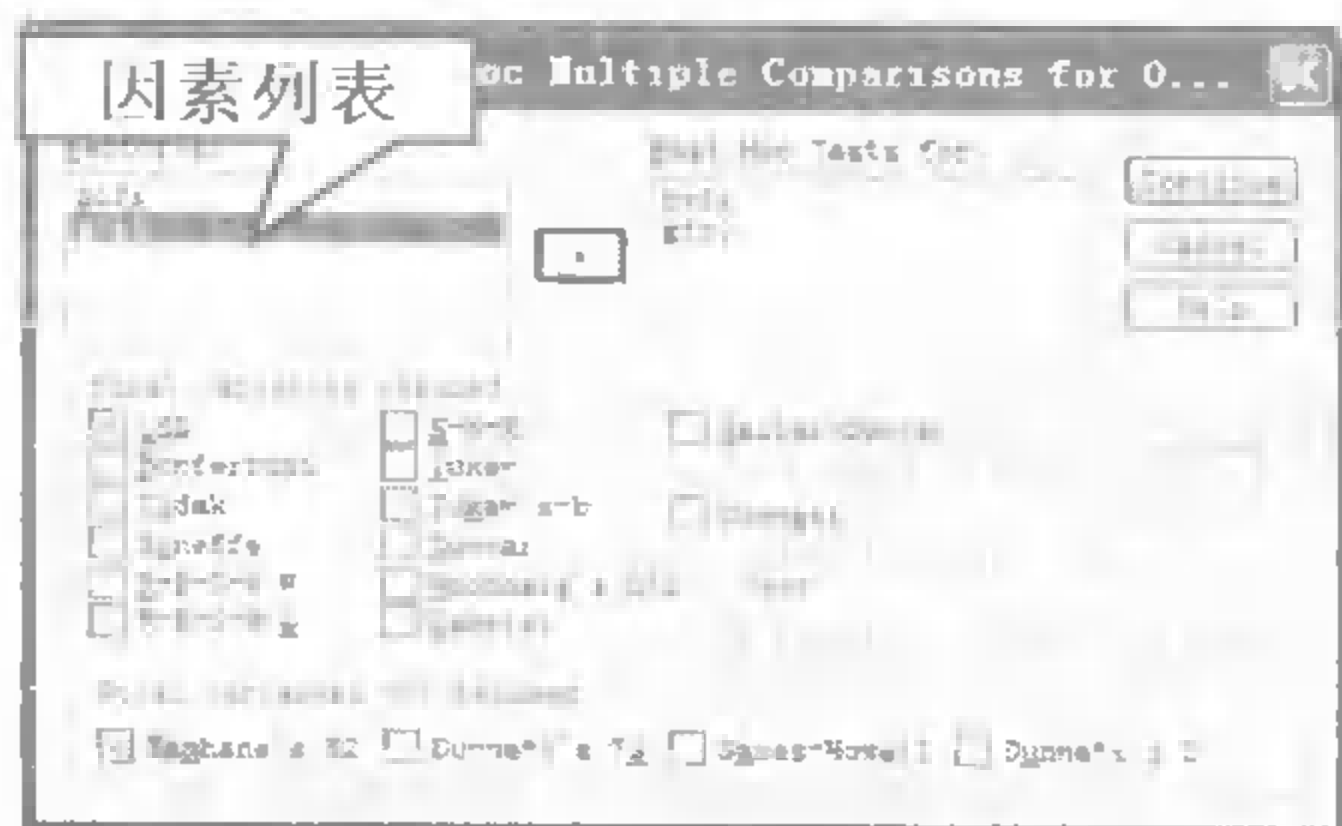


图 9-17 多因素方差分析的多重比较设置

(5) 多重比较选项设置。单击图 9-13 中的 Post Hoc 按钮，弹出如图 9-17 所示的对话框。在此设置多重比较的方法和参数。在因素列表选中 bzfs 和 gtzl 变量，单击 ☐ 按钮，将其选入 Post Hoc Test for 列表框；分别勾选 LSD 复选框和 Tamhane's T2 复选框；单击 Continue 按钮返回主界面。

此设置界面与图 9-6 所示单因素方差分析的多重比较设置界面相同, 请参考第 9.2.2 节。

- **Factor(s)**列表框, 显示在主面板指定的因素变量; **Post Hoc Test for**列表框, 用于选入需要进行多重比较的变量, 下面的设置选项对此处所有选入的变量都起作用。

(6) 保存选项设置。单击图 9-13 中的 **Save** 按钮, 弹出如图 9-18 所示的对话框, 在此设置保存选项。分别勾选复选框: **Standard error**、**Standardized**、**Cook's distance** 和 **Create Coefficient Statistics**; 在 **Dataset name** 后输入“outdataset”; 单击 **Continue** 按钮返回主界面。

在此可以设置将预测值、残差和检测值等数据作为新变量保存到指定的数据文件, 以便在其他分析过程中使用。下面详细介绍各设置选项的含义。

① **Predicted Values** 预测值栏, 可选项有: **Unstandardized**, 非标准化的预测值; **Weight**, 如果在主界面指定了 **WLS(Weighted Least Squares)**权重变量, 勾选该复选框, 表示保存加权的非标准化预测值; **Standard error**, 表示自变量取固定值时, 因变量均值的标准差的估计值。

② **Diagnostics** 诊断值栏, 可选项有: **Cook's distance** (Cook 距离), 表示把一个个案从计算回归系数的样本中去除时, 所引起的残差变化的大小, Cook 距离越大表明该个案对回归系数的影响也越大; **Leverage values** (非中心化杠杆值), 它可用与衡量单个观测对模型拟合效果的影响程度。

③ **Residuals** 残差栏, 可选项有: **Unstandardized**, 非标准化残差值, 即观测值和预测值之差; **Weight**, 加权的非标准化残差 (需在主界面指定了 **WLS** 变量); **Standardized**, 标准化残差 (又称 **Pearson** 残差), 均值为 0, 标准差为 1; **Studentized**, 学生化残差, 用残差除以关于残差标准差的估计值, 这个估计值取决于当前个案自变量的取值与自变量均值之间的距离; **Deleted**, 剔除残差, 表示把某个个案从计算回归系数的样本中去除时, 回归后计算所得的关于当前个案的残差, 即观测值与调整预测值的差。

④ **Coefficient Statistics** 栏, 设置与参数估计值相关的保存选项, 包括: 参数估计值及其置信区间、估计值的方差-协方差矩阵等。

勾选 **Create coefficient statistics** 复选框, 激活如下的两个可选项。

- **Create a new dataset**, 建立一个新的数据集, 在 **Dataset name** 后指定数据集名称。

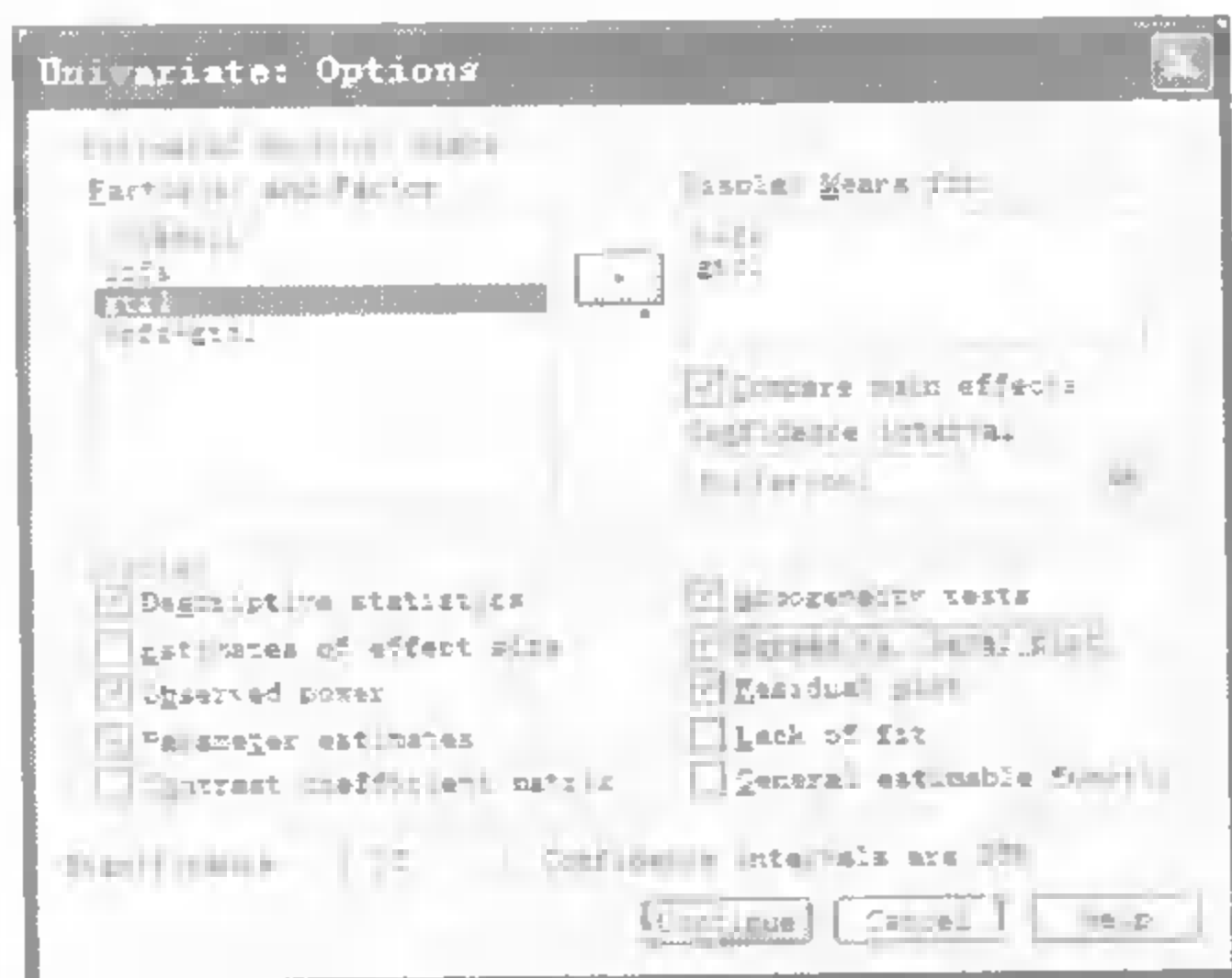


图 9-19 多因素方差分析的 Options 选项设置

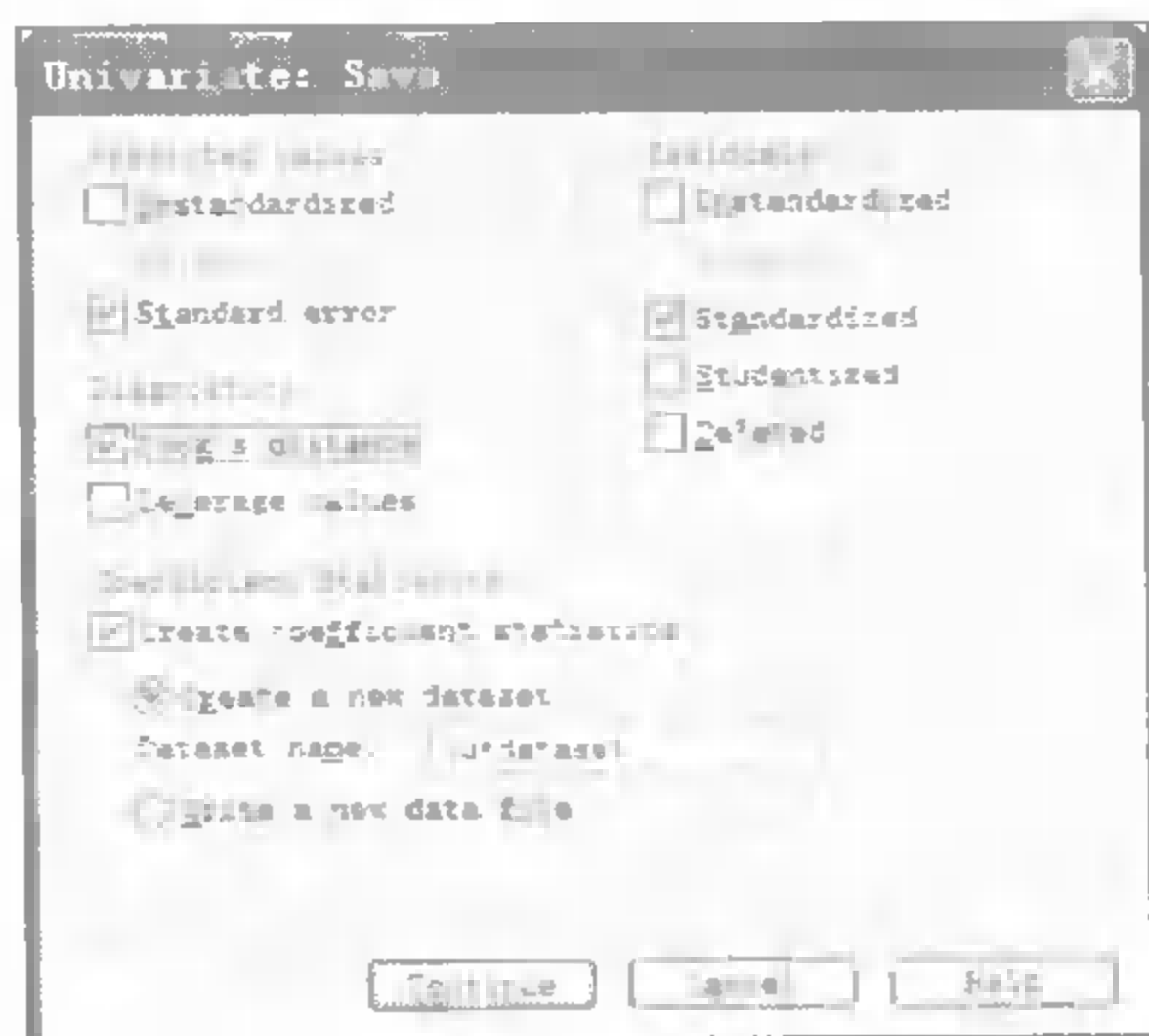



图 9-18 多因素方差分析的保存设置

- **Write a new data file**, 将相关数据保存到新文件中, 单击 **Files** 按钮指定保存路径。

(7) **Options** 选项设置。在图 9-13 中, 单击 **Options** 按钮, 弹出如图 9-19 所示的选项设置对话框。在 **Factor** 列表选中 **bzfz** 和 **gtzl**, 单击  按钮将其选入 **Display** 列表; 勾选 **Compare** 复选框, 单击 **Confidence interval** 下拉列表并选中 **Bonferroni** 选项; 分别勾选复选框: **Descriptive**、**Observed**、**Parameter**、**Homogeneity**、**Spread**、**Residuals**; 单击 **Continue** 按钮返回主界面。



- ① Estimated Marginal Means 子设置栏，在此指定估计估计边际均值的有关参数。
- Factor(s)列表框，显示了在主界面中指定的所有因素变量，及其交互作用项。
  - Display Means for 列表框，用于从 Factor(s)列表选入要输出边际均值的因素变量。
  - Compare main effects 复选项，表示对选入的主效应进行多重均值比较。
  - Confidence interval 下拉列表，指定多重比较的方法，有 3 个可选项：LSD(none)、Bonferroni 和 Sidak，其含义与图 9-6 所示的单因素方差分析时的情况相同。
- ② Display 子设置栏，在此指定要输出的统计量，可选项包括以下内容。
- Descriptive statistics 描述统计量（均值、标准差等）；
  - Estimates of effect size 效应量估计，给出偏 eta-Square 值，反映了每个效应与每个参数估计值可以归于某个因素的变异的大小；
  - Observed power 给出计算功效的显著性 Alpha 值，该值应该在 0.01 到 0.99 之间（系统默认的显著性水平是 0.05）；
  - Parameter estimates 参数估计，给出各因变量与自变量的回归系数、标准误、 $t$  检验的  $t$  值、显著性概率和 95% 的置信区间；
  - Contrast coefficient matrix 显示变换系数矩阵或 L 矩阵；
  - Homogeneity tests 方差齐次性检验；
  - Spread vs. level plot 绘制观测量均值 - 标准差图、观测量均值 - 方差图；
  - Residuals plot 绘制残差图，给出观测值、预测值散点图和观测量数目对标准化残差的散点图，还有正态和标准化残差的正态概率图；
  - Lack of fit 拟合度不足的检验，检查独立变量和非独立变量间的关系是否被充分描述；
  - General estimable function 广义估计函数复选项，可以根据一般估计函数自定义假设检验，对比系数矩阵的行与一般估计函数是线性组合的。
- ③ Significance 输入框，用于指定多重比较的显著性水平，默认为 0.05。

### 3. SPSS 输出结果

单击图 9-13 中的 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 9-20～图 9-25 所示。

描述性统计量				
因变量 销售额				
包装方式	柜台种类	均值	标准差	N
1.00	1.00	5.9000	45589	3
	2.00	6.3333	40415	3
	3.00	6.1333	56862	3
	总计	6.1222	45216	9
2.00	1.00	5.4000	60900	3
	2.00	5.4000	17321	3
	3.00	5.4667	48092	3
	总计	5.4222	38658	9
3.00	1.00	6.4333	66583	3
	2.00	5.9667	66583	3
	3.00	6.3667	20817	3
	总计	6.2556	52941	9
总计	1.00	5.9111	66978	9
	2.00	5.9000	57009	9
	3.00	5.9889	55327	9
	总计	5.9333	57779	27

主体间因子				
		N		
包装方式	1.00	9		
	2.00	9		
	3.00	9		
柜台种类	1.00	9		
	2.00	9		
	3.00	9		

误差方差等同性的 Levene 检验 <sup>a</sup>				
因变量 销售额				
F	df1	df2	Sig.	
1.278	8	18	.314	

检验零假设，即在所有组中因变量的误差方差均相等。  
a. 设计 截距+bxfs+gtzl+bxfs\*gtzl

图 9-20 基本统计量和同方差检验



主体间效应的检验							
因变量 销售额							
源	III 型平方和	df	均方	F	Sig.	非中心 λ 参数	观测到的幂 <sup>a</sup>
校正模型	4 280 <sup>b</sup>	8	535	2 189	.080	17 509	.678
截距	950 520	1	950 520	3888 491	.000	3888 491	1 000
bzfs	3 607	2	1 803	7 377	.005	14 755	.894
gtzl	042	2	021	086	.918	173	.061
bzfs * gtzl	631	4	158	645	.637	2 582	.173
误差	4 400	18	244				
总计	959 200	27					
校正的总计	8 680	26					

a. 使用 alpha 的计算结果 = .05  
b. R 方 = .493 (调整 R 方 = .288)

图 9-21 效应检验输出

参数估计								
因变量 销售额								
参数	B	标准误	t	Sig.	95% 置信区间		非中心 λ 参数	观测到的幂 <sup>a</sup>
截距	6 367	233	22 504	.000	5 906	6 866	22 504	1 000
[bzfs=1.00]	-.233	.404	-.578	.570	-1.081	.615	-.578	.085
[bzfs=2.00]	.909	.404	2.229	.039	-.148	2.229	2.229	.381
[bzfs=3.00]	.000							
[gtzl=1.00]	.067	.404	.169	.871	-.731	.915	.169	.003
[gtzl=2.00]	-.400	.404	-.991	.335	-1.248	.448	-.991	.151
[gtzl=3.00]	.000							
[bzfs=1.00] * [gtzl=1.00]	-.300	.571	-.528	.606	-1.459	.859	-.528	.075
[bzfs=1.00] * [gtzl=2.00]	.000	.571	1.051	.307	-.580	1.580	1.051	.189
[bzfs=1.00] * [gtzl=3.00]	.000							
[bzfs=2.00] * [gtzl=1.00]	-.133	.571	-.234	.818	-1.538	1.066	-.234	.056
[bzfs=2.00] * [gtzl=2.00]	.553	.571	.984	.367	-.566	1.323	.984	.086
[bzfs=2.00] * [gtzl=3.00]	.000							
[bzfs=3.00] * [gtzl=1.00]	.000							
[bzfs=3.00] * [gtzl=2.00]	.000							
[bzfs=3.00] * [gtzl=3.00]	.000							

a. 使用 alpha 的计算结果 = .05  
b. 此参数为冗余参数，将被设为零。

图 9-22 参数估计结果

定制假设检验指数			
1	对比系数 (L 矩阵) 转换系数 (M 矩阵) 对比结果 (K 矩阵)	包装方式的简单对比 (参考类别 = 3) 单位矩阵 零矩阵	
2	对比系数 (L 矩阵) 转换系数 (M 矩阵) 对比结果 (K 矩阵)	柜台种支的简单对比 (参考类别 = 3) 单位矩阵 零矩阵	

对比结果 (K 矩阵)			
包装方式简单对比 <sup>a</sup>			因变量 销售额
级别 1 和级别 3	对比估算值		133
	假设值		0
	差分 (估计 - 假设)		-133
	标准误		233
	Sig.		.574
	差分的 95% 置信区	下限	-623
	间	上限	356
级别 2 和级别 3	对比估算值		-833
	假设值		0
	差分 (估计 - 假设)		-833
	标准误		233
	Sig.		.002
	差分的 95% 置信区	下限	-1323
	间	上限	344

a. 参考类别 = 3

检验结果							
因变量 销售额							
源	平方和	df	均方	F	Sig.	非中心 λ 参数	观测到的幂 <sup>a</sup>
对比	3 607	2	1 803	7 377	.005	14 755	.894
误差	4 400	18	244				

a. 使用 alpha 的计算结果 = .05

图 9-23 对照比较输出

估计						
因变量 销售额						
包装方式	均值	标准误	95% 置信区间		下限	上限
1.00	5.122	.165			5.776	6.468
2.00	5.412	.164			5.076	5.748
3.00	6.256	.165			5.902	6.609

成对比较						
因变量 销售额						
① 包装方式	② 包装方式	均值差值 (I-J)	标准误	Sig. <sup>a</sup>	差分的 95% 置信区间 <sup>a</sup>	
					下限	上限
1.00	2.00	-.290*	.233	.023	-.755	.175
	3.00	1.133	.233	.000	.648	1.618
2.00	1.00	.700*	.233	.023	-.135	.535
	3.00	-.883*	.233	.006	-1.448	-.318
3.00	1.00	1.133	.233	.000	.648	1.618
	2.00	-.833*	.233	.006	-1.318	-.353

基于估算边际均值  
\* 均值差值在 .05 级别上较显著。  
a 对多个比较的调整: Bonferroni.

单变量检验							
因变量 销售额							
	平方和	df	均方	F	Sig.	非中心参数	观测到的幂 <sup>a</sup>
对比	3.647	2	1.823	7.377	.003	14.754	.994
误差	4.404	18	.244				

F 检验 包装方式的效应。该检验基于估算边际均值间的线性独立成对比较。  
a 使用 alpha 的计算结果 = .05

图 9-24 边际均值输出

多个比较							
因变量 销售额							
	① 包装方式	② 包装方式	均值差值 (I-J)	标准误	Sig.	95% 置信区间	
						下限	上限
LSD	1.00	2.00	-.290*	.23307	.003	-.7551	.1751
		3.00	1.133	.23307	.000	.6481	1.6181
	2.00	1.00	.700*	.23307	.003	-.1351	.5351
		3.00	-.883*	.23307	.002	-1.3181	-.3531
	3.00	1.00	1.133	.23307	.000	.6481	1.6181
		2.00	-.833*	.23307	.002	-.3531	-.3531
Tamhane	1.00	2.00	-.290*	.19430	.009	-.7011	.1229
		3.00	1.133	.19430	.000	.6481	1.6181
	2.00	1.00	.700*	.19430	.009	-.1239	.5239
		3.00	-.833*	.19430	.005	-.3531	-.3531
	3.00	1.00	1.133	.19430	.000	.6481	1.6181
		2.00	-.833*	.19430	.005	-.3531	-.3531

基于观测到的均值。  
\* 均值差值在 .05 级别上较显著。

多个比较							
因变量 销售额							
	① 柜台种类	② 柜台种类	均值差值 (I-J)	标准误	Sig.	95% 置信区间	
						下限	上限
LSD	1.00	2.00	-.011	.23307	.965	-.4755	.4535
		3.00	-.078	.23307	.747	-.5874	.4314
	2.00	1.00	-.011	.23307	.965	-.4755	.4535
		3.00	-.089	.23307	.707	-.5785	.4095
	3.00	1.00	-.078	.23307	.747	-.4755	.4535
		2.00	-.089	.23307	.707	-.5785	.4095
Tamhane	1.00	2.00	-.011	.20313	1.000	-.7724	.7504
		3.00	-.078	.20313	.901	-.8526	.6970
	2.00	1.00	-.011	.20313	1.000	-.7724	.7504
		3.00	-.089	.20313	.893	-.8546	.6806
	3.00	1.00	-.078	.20313	.893	-.8526	.6970
		2.00	-.089	.20313	.901	-.8526	.6970

基于观测到的均值。

图 9-25 多重比较的输出

(1) 描述性统计量和方差齐性检验结果。如图 9-20 所示,“主体间因子”表格给出了各主效应不同取值水平下的样本个数统计;“描述性统计量”表格给出了观察样本各个分组的基本统计特征,包括均值、标准差等。

“Levene 检验”表格输出的是方差齐性检验结果,由显著性检验的 Sig 值  $0.314 > 0.10$  推断,在 0.10 的显著性水平上,认为各组方差是无显著差异的。

(2) 方差分析结果。如图 9-21 所示,是各效应检验的输出结果,从各项显著性检验的 Sig 值可以看出,截距项、包装方式 (bzfs) 对销售额的影响,在 0.05 的显著性水平上是比较显著的;但是,柜台种类 (gtzl) 对销售额的影响并不显著。

(3) 参数估计。如图 9-22 所示,是参数估计的输出结果,以 [bzfs=2] 行为例,表示在其他条件相同的情况下,采用包装方式 2 的销售额要比用包装方式 3 的销售额显著地少 0.900。

(4) 对照比较结果。如图 9-23 所示,是对包装方式变量的简单对照比较输出,关于柜台种类的输出与此类似。

“定制假设检验指数”表格给出对照比较的规则，表示两个变量分别与各自的最后一个类别进行比较。“对比结果”表格是包装方式的对照比较结果，级别1和级别3的显著性检验Sig值 $0.574 > 0.10$ ，故推断这两种包装方式下的销售额无显著区别；而包装方式在级别2和级别3时（Sig值为 $0.002 < 0.01$ ），销售额是有显著差异的。“检验结果”表格是对包装方式所有三个类别的总检验，其内容与图9-21中bzfs（包装方式）一行的输出完全相同。

同理可知，销售额在柜台种类的不同取值间没有显著的差异。

从对照比较的结果可以得到比图9-21所示的方差分析结果更多的信息，比如各因素变量具体在哪些取值水平上对因变量的影响有显著作用，或没有作用。

（5）边际均值相关的输出。如图9-24所示，是包装方式的均值估计和成对比较结果，关于柜台种类的输出与此类似。

“估计”表格中，各包装方式下销售额的均值估计都是其样本观测的均值，与图9-20中相应的均值相同。“成对比较”表格显示，在0.05的显著性水平上，包装方式1和2之间、包装方式2和3之间的销售额都是有显著差异的，这比图9-23中定制的对照比较又多了一些信息。“单变量检验”表格的内容与图9-23中的“检验结果”表格完全一样。

同理可知，柜台种类各取值水平对销售额都没有显著影响。

（6）多重比较结果。如图9-25所示，分别给出了关于包装方式和柜台种类的多重比较输出，其中LSD栏是假设方差齐性的检验结果，Tamhane栏是方差不齐时的检验结果：由图9-20已经推断接受方差齐性的假设，故本例只参考LSD栏的检验结果。

表格中的“均值差异”一列带有“\*”号标识的，表示在0.05的显著性水平上，相应的两个因素水平之间的销售额差异是显著的，其中关于包装方式的结果与图9-24中的输出一致。

（7）分布和水平图。如图9-26所示，是销售量关于标准差和方差的分布和水平图。其中标准差对均值的分布图，就是用图9-20中“描述性统计量”表格给出的均值和标准差所作，通过观察各点在此图中的分布有无明显规律，可以直观地检验方差齐性的假设是否成立。

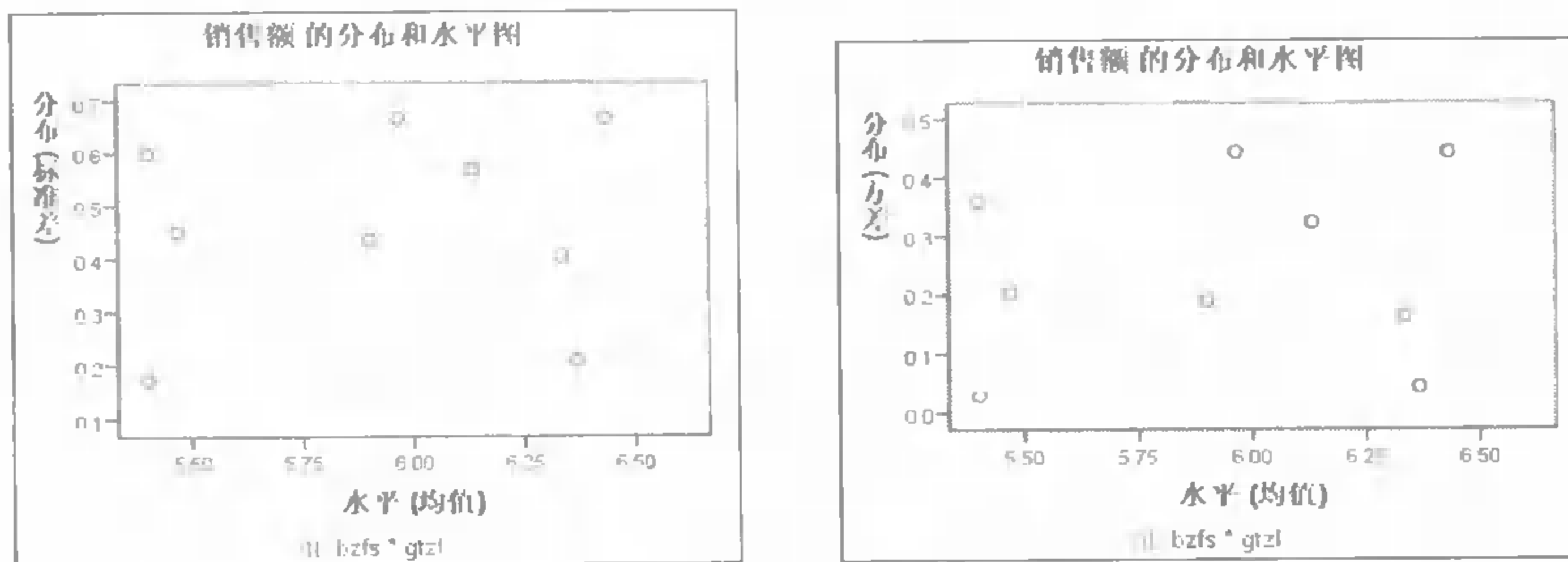


图9-26 销售量的分布和水平图

（8）边际均值图。如图9-27所示，“因变量：销售额”矩阵图是关于残差的两两散点图，包括：观测的、预测的和标准残差。“估算边际均值”图是以包装方式分线的对柜台种类的边际均值图，包装方式的水平1和3有交叉，说明它们之间的销售额差异不太显著；而包装方式在水平1和2之间、2和3之间都没有交叉，说明其销售额差异比较显著，这和前面得出的结论是一致的。

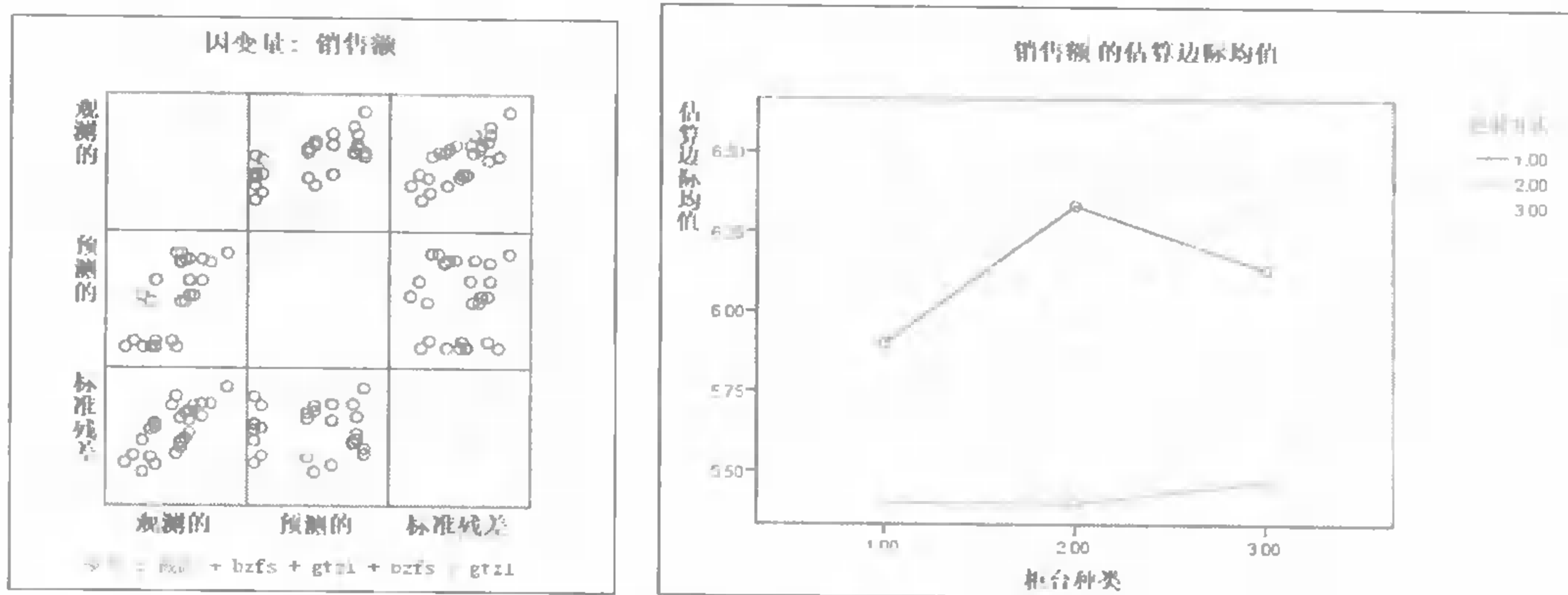


图 9-27 边际均值图

### 9.3.3 协方差分析实例

在进行方差分析时，除了感兴趣的研究因素外，应尽量保证其他条件的一致，这就要用到协方差分析。协方差分析的特点可以简要概括为：消除不可控因素的影响，再进行方差分析。

#### 1. 协方差分析介绍

协方差分析是利用线性回归消除混杂因素的影响后，再进行的方差分析。例如：研究一种药物对患者某个生化指标的影响，需要比较实验组与对照组该指标变化的均值是否有显著差异来确定该药物的有效性，同时还应考虑患者的年龄、病程长短以及原指标水平等对疗效的影响；只有在消除其他因素的影响后再考虑药物的疗效（即指定生化指标的变化），才是科学的分析方法。如果在选择研究对象时，令混杂因素的取值水平都相同，就可以使用一般的方差分析方法。这对于动物实验比较容易控制，比如选择了同品种、同一胎的大白鼠，对其分组后在相同的饲养条件下进行实验，就可以避免许多混杂因素的影响。

协方差分析采用线性回归方法，寻找各分组的因变量  $Y$  与协变量  $X$  之间的数量关系，求出假定  $X$  相等时的修正均值，然后用方差分析比较修正均值之间的差别；与回归分析相比，它侧重于求修正均值，其次才是比较。

从因素变量和协变量的个数，可以把协方差分析分为：单因素协方差分析、多因素协方差分析和多协变量协方差分析等；从试验设计方式，又可以分为：完全随机设计的协方差分析、随机区组设计的协方差分析和析因协方差分析。

协方差分析的假设条件有： $X$  与  $Y$  的线性关系在各个分组都成立，且各组之间的回归系数近似相等； $X$  的取值范围不宜过大，否则修正均值在回归直线的延长线上不能确定是否仍然满足线性关系和平行性的假设条件，协方差分析的结论可能也不正确。

#### 2. 据描述

本节利用单因素协方差分析，研究三种猪饲料的增重效果是否有显著差异。

所用数据文件为“猪饲料增重效果数据.sav”，数据格式如图 9-28 所示，本例需要消除的协变量因素是：猪的喂养前体重。



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	wyq	Numeric	8	2	喂养前体重	None	None	8	Right	Scale
2	wyh	Numeric	8	2	喂养后体重	None	None	8	Right	Scale
3	sl	Numeric	8	2	饲料种类	None	None	8	Right	Ordinal

图 9-28 猪饲料增重效果的数据格式

### 3. 参数设置

依次单击菜单“Analyze→General Linear Model→Univariate...”，执行协方差分析过程，主设置界面如图 9-29 所示，其参数设置与第 9.3.2 节介绍的二因素方差分析过程一样。

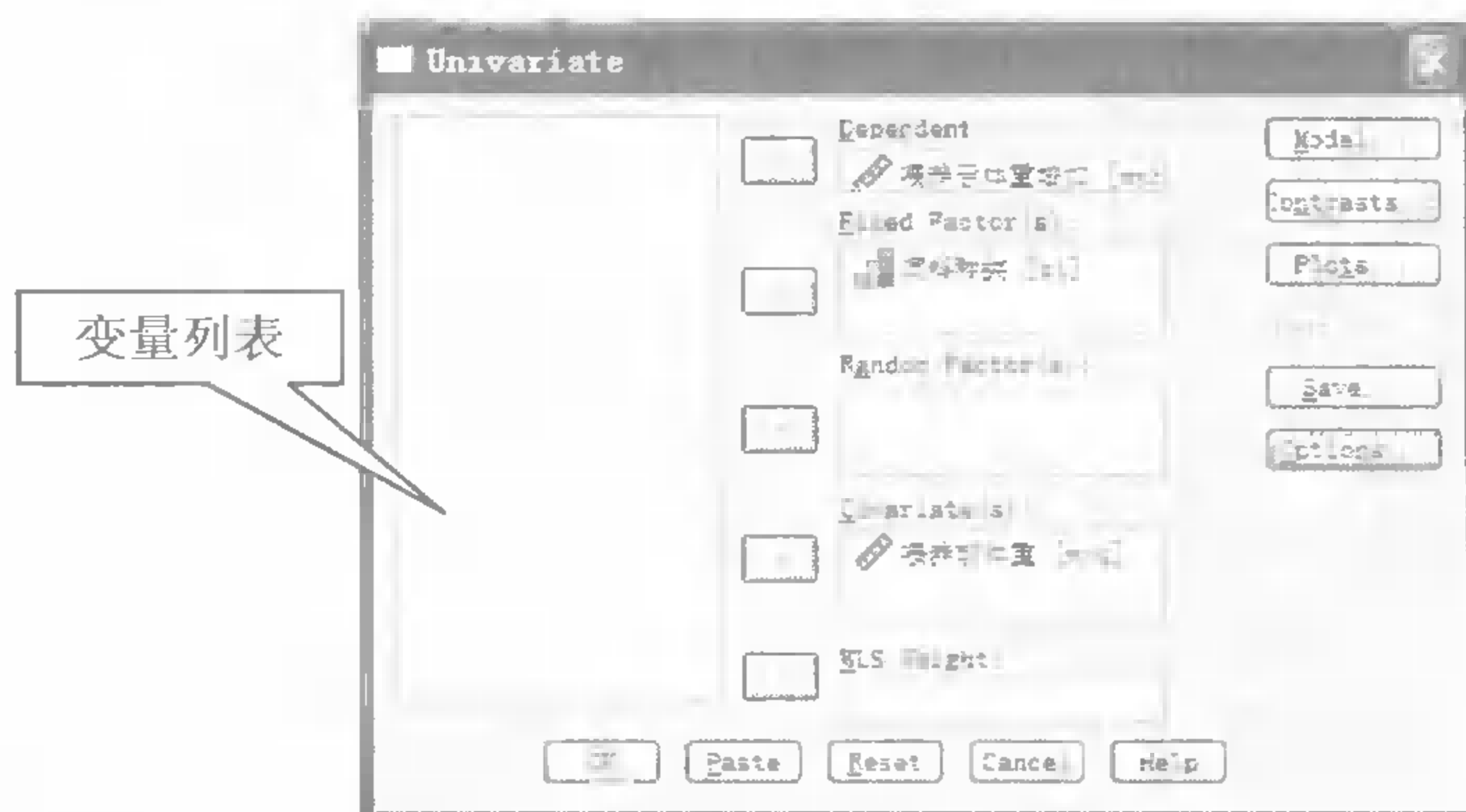






图 9-29 单因素协方差分析的主界面

(1) 在变量列表单击选中喂养后体重增加变量，单击从上至下第一个  按钮，将其作为因变量选入 Dependent 选框；在变量列表单击选中饲料种类变量，单击从上至下第二个  按钮，将其作为因素变量选入 Fixed 列表框；在变量列表单击选中喂养前体重变量，单击从上至下第四个  按钮，将其作为协变量选入 Covariate 列表框。

(2) 单击 Options 按钮，弹出如图 9-30 所示的选项设置对话框，在 Factor 列表选中 sl（饲料种类），单击  按钮将其选入 Display 列表：分别勾选复选框：Descriptive、Parameter 和 Homogeneity；单击 Continue 按钮返回主界面。

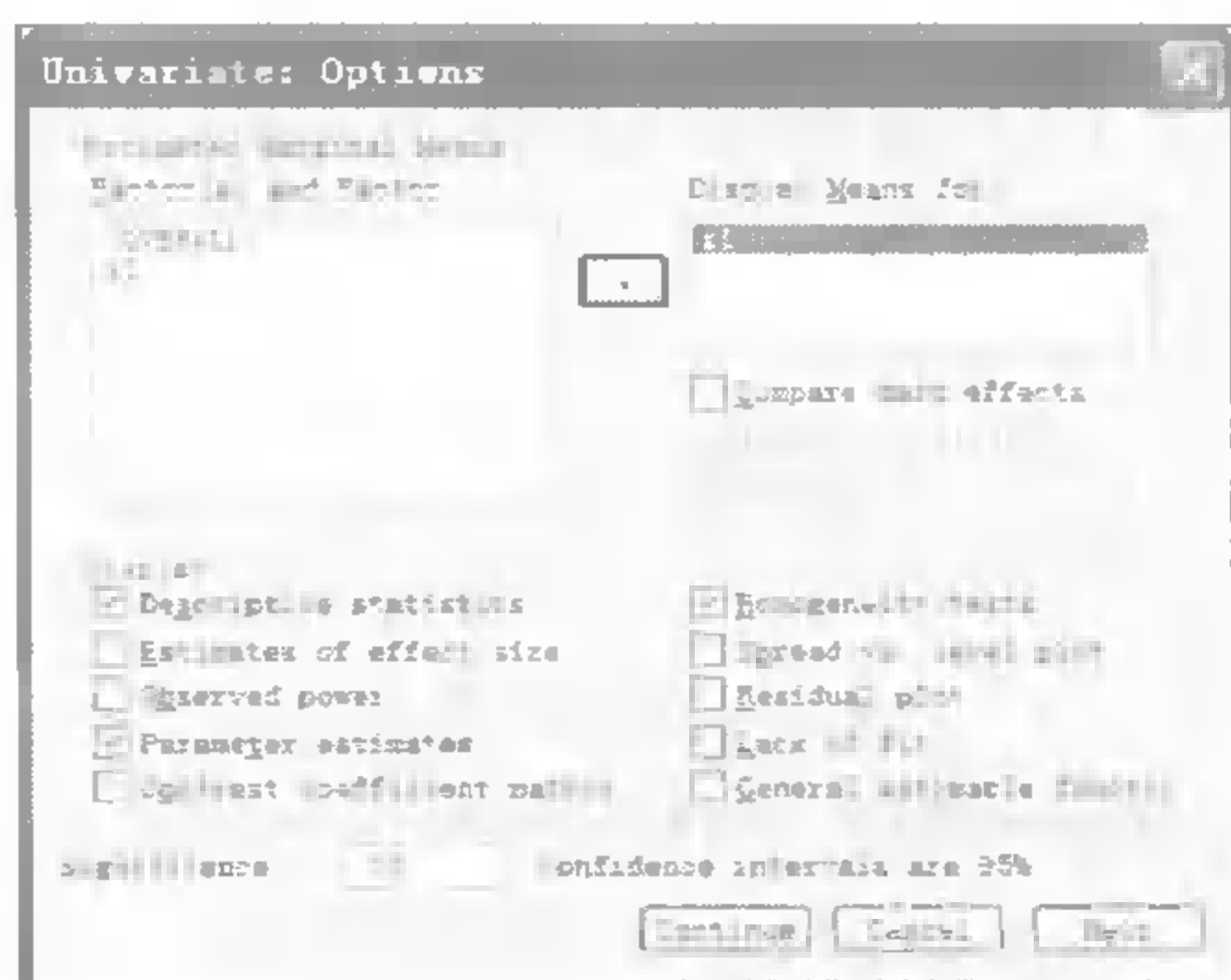


图 9-30 单因素协方差分析的选项设置

### 4. 案例的结果分析

单击图 9-29 中的 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 9-31～图 9-34 所示。

主体间因子			
饲料种类	1.00	2.00	3.00
均值	92.7083	95.8750	98.9000
标准差	10.54176	8.99901	5.12636
N	8	8	8

图 9-31 描述性统计量输出

误差方差等同性的 Levene 检验 <sup>a</sup>				
因变量 喂养后体重增加	F	df1	df2	Sig.
	.74	2	21	.486
检验零假设，即所有组中因变量的误差方差均相等。				
a. 设计 截距-wyq-sl				

图 9-32 方差齐性检验结果

主体间效应的检验					
因变量 喂养后体重增加					
源	III 型平方和	df	均方	F	Sig.
校正模型	2328.344 <sup>a</sup>	2	776.115	68.196	.000
截距	980.448	1	980.448	86.150	.000
wyq	1010.760	1	1010.760	88.813	.000
sl	707.219	2	353.509	31.071	.000
误差	227.615	20	11.381		
总计	206613.000	24			
校正的总计	2555.958	23			

<sup>a</sup> R 方 = .911 (调整 R 方 = .898)

协变量效应检验

因素变量效应检验

图 9-33 效应检验的输出

参数估计						
因变量 喂养后体重增加						
参数	B	标准误	t	Sig.	95% 置信区间	
					下限	上限
截距	35.935	6.575	5.455	.000	22.219	49.651
wyq	2.402	.255	9.424	.000	1.870	2.933
[sl=1.00]	12.793	3.409	3.753	.001	5.682	19.904
[sl=2.00]	17.336	2.409	7.196	.000	12.310	22.361
[sl=3.00]	0 <sup>a</sup>					

<sup>a</sup> 此参数为冗余参数，将被设为零。

估算边际均值				
因变量 喂养后体重增加				
饲料种类	均值	标准误	95% 置信区间	
			下限	上限
1.00	94.959 <sup>a</sup>	1.840	91.120	98.798
2.00	99.501 <sup>a</sup>	1.203	96.991	102.011
3.00	82.165 <sup>a</sup>	1.964	78.255	86.075

<sup>a</sup> 模型中出现的协变量在下列值处进行评估 喂养前体重 = 19.2500

消除协变量影响

图 9-34 参数估计和边际均值

(1) 描述性统计输出。如图 9-31 所示，“主体间因子”表格给出了饲料种类不同取值水平下的样本个数；“描述性统计量”表格给出了因变量在各个分组里的基本统计特征，包括均值、标准差等。

(2) 方差齐性检验结果。如图 9-32 所示，“Levene 检验”表格输出的是方差齐性检验结果，由显著性检验的 Sig 值  $0.486 > 0.10$  推断，在 0.10 的显著性水平上，认为各组方差是无显著差异的。

(3) 各效应检验结果。如图 9-33 所示，协变量的效应检验非常显著 ( $\text{Sig} < 0.01$ )，故认为喂养前体重和喂养后体重增加量之间存在着比较强的线性关系，故而进行协方差分析是有必要的。而且，因素变量的检验结果也很显著 ( $\text{Sig} < 0.01$ )，说明三种饲料对猪的体重增加也是有显著差异的。

(4) 参数估计和边际均值估计。如图 9-34 所示，“参数估计”表格给出了因变量（体重增加）对协变量（喂养前体重）的回归系数 ( $B=2.402$ )，表示喂养前体重越大，则喂养后体重增加量也越大。

“饲料种类”表格给出了消除协变量影响后的边际均值的估计值，可见饲料 2 的增重效果最好，其均值 (99.501) 最大、且标准误 (1.203) 最小。

#### 9.3.4 交互效应中随机因素的分析

随机因素方差分析的基本原理与固定因素类似，只是在计算 F 统计量时所采用的误差平方和稍有不同。以两因素方差分析为例：在无交叉效应且无重复实验时，F 统计量的计算方法与表 9-4 所示固定因素的计算相同；有交互效应时，固定因素的 F 统计量的计算公式如表 9-6 所示，随机效应的 F 统计量的计算公式如表 9-7 所示。

表 9-7

两随机因素有重复实验的方差分析表

变异来源	SS	DF	S <sup>2</sup>	F	
				交互作用 为 0 时	交互作用 不为 0 时
A、B 因素组合	$SS_{AB} = \frac{1}{b} \sum x_{ij.}^2 - C$	$df_{AB} = ab - 1$	$S_T^2$		$S_r^2 / S_e^2$
A×B 互作	$SS_{A \times B} = SS_{AB} - SS_A - SS_B$	$df_{A \times B} = (a - 1)(b - 1)$	$S_{AB}^2$		$S_{AB}^2 / S_e^2$
A 因素	$SS_A = \frac{1}{bn} \sum x_{i..}^2 - C$	$df_A = a - 1$	$S_A^2$	$S_A^2 / S_R^2$	$S_A^2 / S_{AB}^2$
B 因素	$SS_B = \frac{1}{an} \sum x_{.j.}^2 - C$	$df_B = b - 1$	$S_B^2$	$S_B^2 / S_R^2$	$S_B^2 / S_{AB}^2$
试验误差	$SS_e = SS_T - SS_{AB}$	$df_e = ab(n - 1)$	$S_e^2$		
其中: $C = x_{...}^2 / abn$					
总变异	$SS_r = \sum \sum \sum x_{ijr}^2 - C$	$df_e = abn - 1$		$S_R^2 = \frac{(SS_{A \times B} + SS_i)}{(a - 1)(b - 1) + ab(n - 1)}$	

### 1. 数据描述

本节仍利用 Univariate 过程，对多于两个的因素变量进行方差分析，所用数据文件为“心理实验数据.sav”，格式如图 9-35 所示。本例有 3 个因素变量：目标、设备、光线，目的是研究实验者在这三个因素不同取值水平的组合下，测得的分数是否有显著差异。



	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	target	Numeric	8	0	目标	{1 1}	None	8	Right	Scale
2	device	Numeric	8	0	设备	{1 d1}	None	8	Right	Scale
3	light	Numeric	8	0	光线	{1 1}	None	8	Right	Scale
4	score	Numeric	8	0	得分	None	None	8	Right	Scale

图 9-35 心理实验数据格式

在上述三个因素里，目标和设备都是固定的、容易控制的，本例将其作为固定因素变量；而光线具有较大的随机性，不易控制，随后将分别把它作为随机因素和固定因素加以处理，并观察分析两者之间的区别。

### 2. 参数设置

依次单击菜单“Analyze→General Linear Model→Univariate...”，执行单因变量分析过程，主设置界面如图 9-36 所示，其参数设置与第 9.3.2 节介绍的二因素方差分析过程一样。

(1) 指定分析变量。在变量列表单击选中得分变量，单击从上至下第一个  按钮，将其作为因变量选入 Dependent 选框；在变量列表选中目标和设备变量，单击从上至下第二个  按钮，将其作为固定因素选入 Fixed 列表框。



在变量列表单击选中光线变量，单击从上至下第二个  按钮，将其作为固定因素选入 Fixed 列表框（如图 9-36 所示）；或者，在变量列表单击选中



图 9-36 参数设置（光线作为固定因素）

光线变量，单击从上至下第三个  按钮，将其作为随机因素选入 Random 列表框（如图 9-37 所示）。

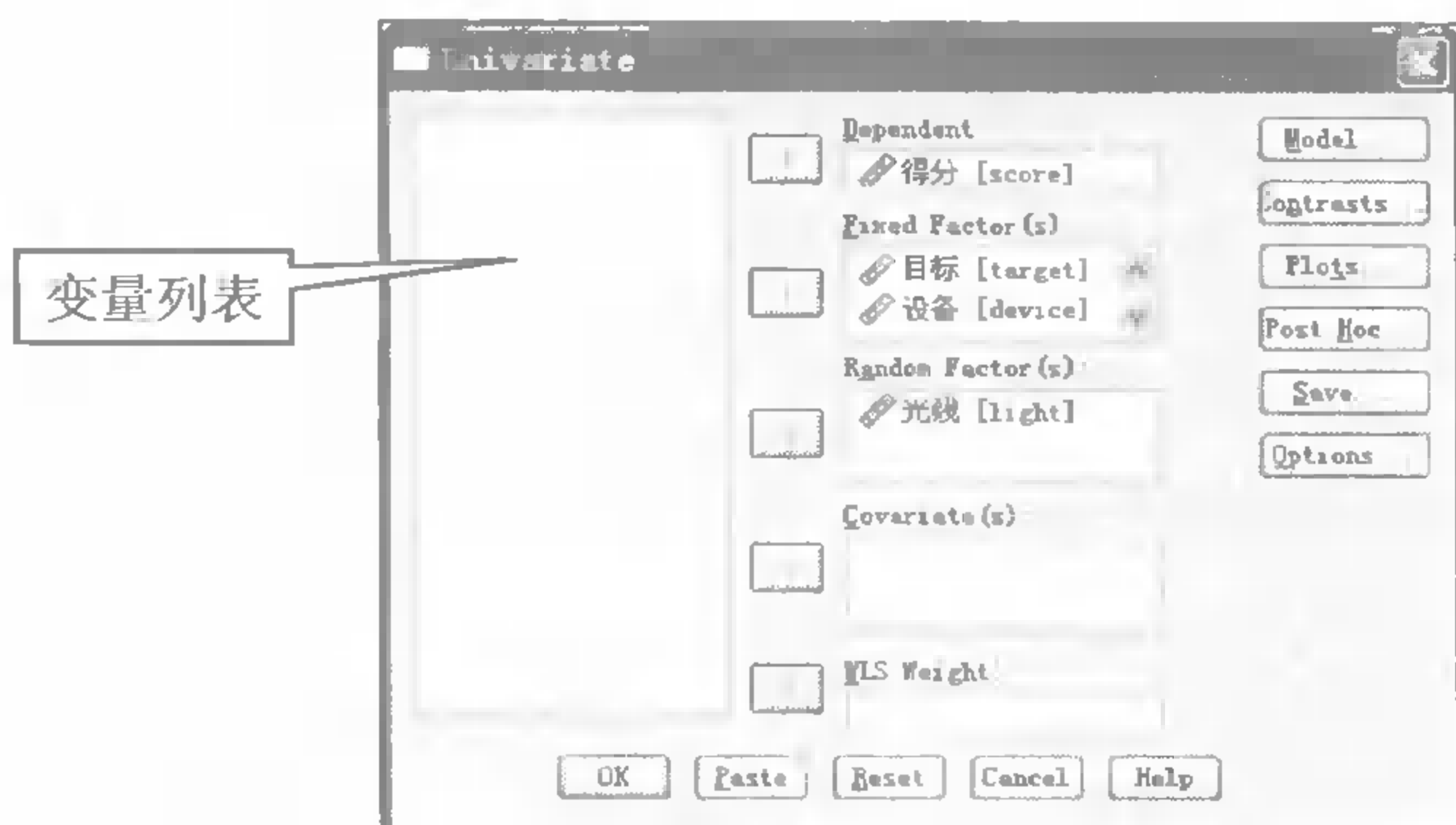





图 9-37 参数设置（光线作为随机因素）

（2）作图选项设置。在图 9-36 中，单击 Plots 按钮，弹出如图 9-38 所示的作图设置对话框。在 Factors 列表单击选中 target（目标）变量，单击从上至下第一个  按钮，将其作为横轴变量选入 Horizontal 选框；在 Factors 列表单击选中 device（设备）变量，单击从上至下第二个  按钮，将其作为分线变量选入 Separate Lines 选框；在 Factors 列表单击选中 light（光线）变量，单击从上至下第三个  按钮，将其作为分图变量选入 Separate Plots 选框；单击 Add 按钮将 “target\*device\*light” 选入下面的列表框；单击 Continue 按钮返回主界面。

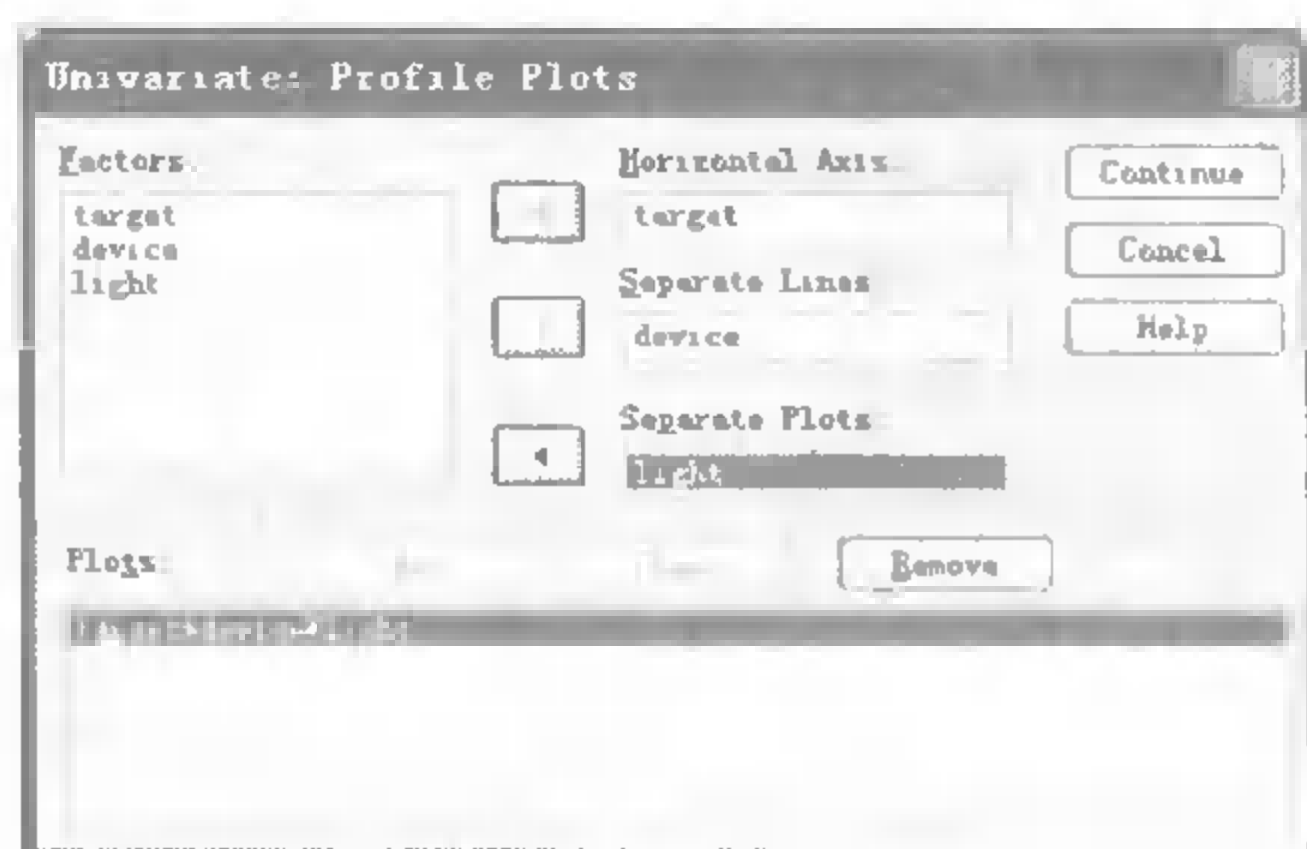


图 9-38 选项作图设置

（3）Options 选项设置。在图 9-36 中单击 Options 按钮，弹出如图 9-39 所示的选项设置对话框。分别勾选复选框：Descriptive statistics、Homogeneity tests；单击 Continue 按钮返回主界面。

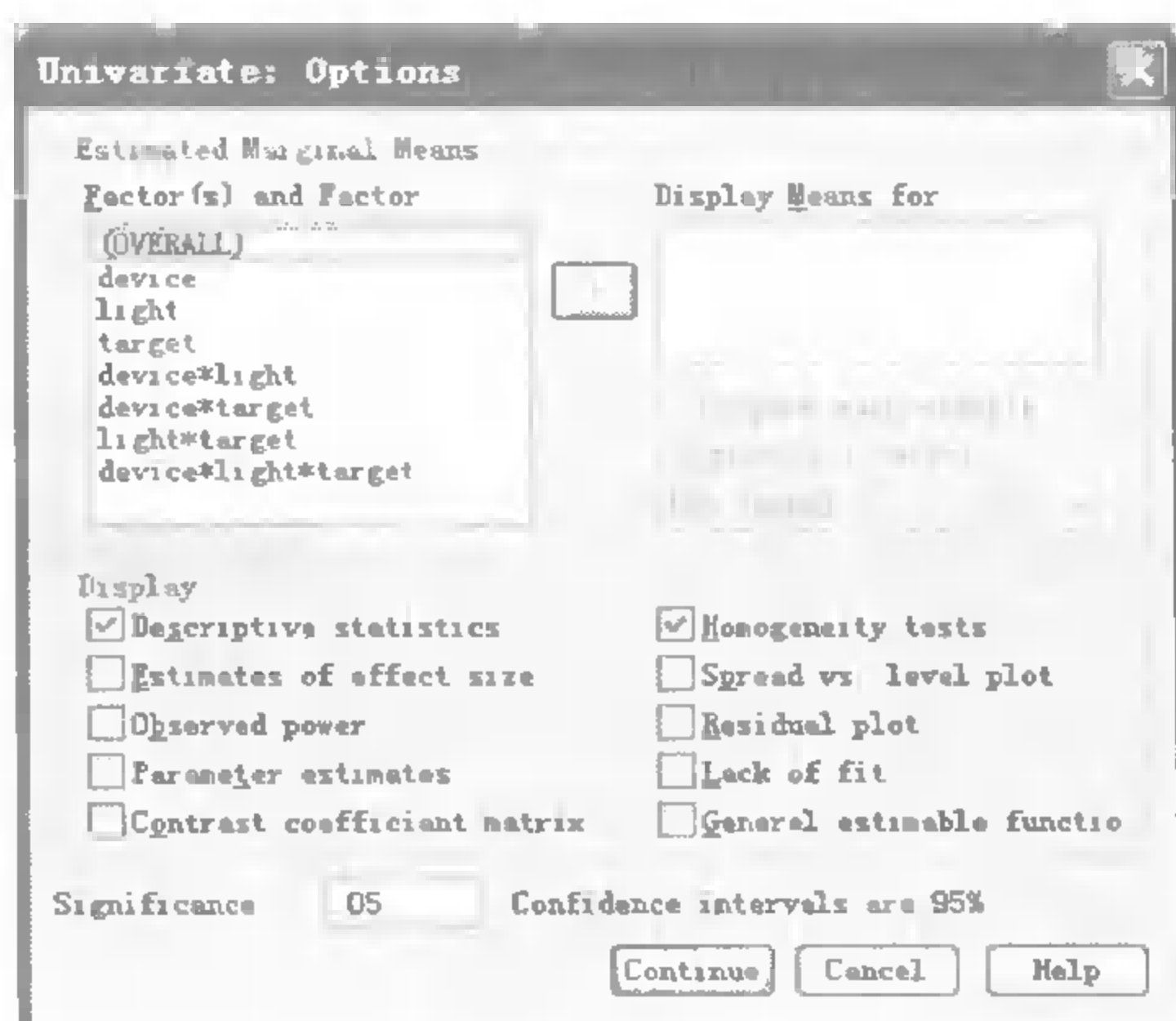


图 9-39 Options 选项设置

### 3. 案例的结果分析

单击图 9-36 中的 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 9-40 和图 9-41 所示。



**误差方差等同性的 Levene 检验<sup>a</sup>**

因变量 得分

F	df1	df2	Sig.
.836	23	96	.680

检验零假设, 即所有组中因变量的误差方差均相等。

a. 设计: 截距+device+light+target+device\*light+device\*target+light\*target+device\*light\*target

**主体间效应的检验**

因变量 得分

源	平方和	df	均方	F	Sig.
校正模型	783.467 <sup>a</sup>	23	34.064	46.481	.000
截距	3162.133	1	3162.133	4312.008	.000
device	86.467	2	43.233	58.055	.000
light	76.800	1	76.800	104.727	.000
target	233.200	3	77.400	106.909	.000
device * light	12.600	2	6.300	8.591	.000
device * target	104.200	6	17.367	23.682	.000
light * target	93.867	3	31.289	42.667	.000
device * light * target	174.333	6	29.056	39.621	.000
误差	70.400	96	.733		
总计	4016.000	120			
校正的总计	853.867	119			

a. R 方 = .918 (调整 R 方 = .905)

图 9-40 光线 (light) 作为固定因素的检验结果

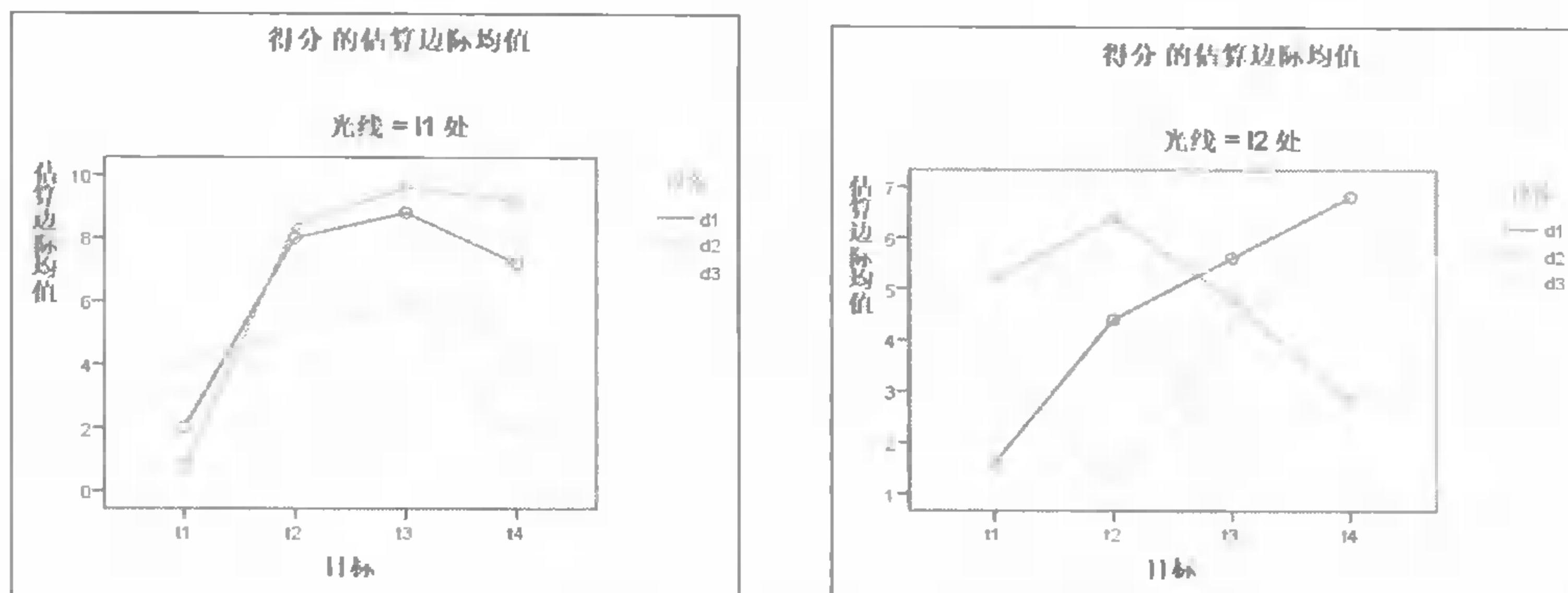


图 9-41 光线 (light) 作为固定因素的边际均值图

在图 9-37 中单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 9-41 和图 9-42 所示。

**误差方差等同性的 Levene 检验<sup>a</sup>**

因变量 得分

F	df1	df2	Sig.
.836	23	96	.680

检验零假设, 即所有组中因变量的误差方差均相等。

a. 设计: 截距+device+target+light+device\*target+device\*light+light\*target+device\*target\*light

**主体间效应的检验**

因变量 得分

源	平方和	df	均方	F	Sig.
截距	3162.133	1	3162.133	4312.008	.000
device	86.467	2	43.233	58.055	.000
target	233.200	3	77.400	106.909	.000
light	76.800	1	76.800	104.727	.000
device * target	104.200	6	17.367	23.682	.000
device * light	12.600	2	6.300	8.591	.000
target * light	93.867	3	31.289	42.667	.000
device * target * light	174.333	6	29.056	39.621	.000
误差	70.400	96	.733		

a. MS(light)  
b. 1/5 MS(device \* light)  
c. 1/3 MS(target \* light)  
d. 1/100 MS(device \* light) + 1/100 MS(target \* light) + 1/400 MS(device \* target \* light)  
e. MS(device \* target \* light)  
f. MS(错误)

图 9-42 光线 (light) 作为随机因素的检验结果

(1) 方差齐性检验和效应检验的输出。如图 9-40 所示, 是把光线作为固定因素时的检验结果; 如图 9-42 所示, 是把光线作为随机定因素时的检验结果; 下面来比较这两种输出结果的异同。

二者关于方差齐性的 Levene 检验输出相同, 由显著性检验的 Sig 值  $0.680 > 0.10$  推断, 各

因素不同取值水平的组合之间，得分变量的方差是无差异的。

光线分别作为固定因素、随机因素时，其效应检验的结果却大相径庭。光线作为固定因素时，各效应的检验都是非常显著的（Sig 值都远小于 0.01）；光线作为随机因素时，只有三阶交叉效应的检验比较显著，其他效应的检验都不显著。可见，本例对光线变量的两种处理方式引起了很大差异。把光线作为随机因素时，由于只考虑了它的两个取值水平 11 和 12（太少），所作的分析就可能会有偏差，所以建议采用把光线作为固定因素时的分析结果。

本例应引起我们对固定因素和随机因素之间区别的重视。在实际应用中，要对所研究的问题作充分的理解和剖析，以清晰地辨别各因素的特征。

（2）边际均值图形输出。光线分别作为固定因素、随机因素时，边际均值图的输出都是一样的，如图 9-41 所示。

这两个图表示光线分别为 11、12 时，不同设备的边际均值随目标变化的折线图，从折线的交叉比较频繁可以推断，在给定的光线水平下，设备和目标之间存在着一定的交互效应。

## 9.4 多元方差分析

多元方差分析的特点是所研究问题的因变量不止一个，这在现实生活和科学研究中经常遇到，例如：研究某些因素对儿童生长过程的影响程度，则身高、体重等都可以作为衡量生长程度的指标，即都可作为因变量；研究片状钛合金在不同温度、不同拉伸速度下的抗拉性能是否一致，可以分别测量钛合金的两边（S1 和 S2）、角（CO）和中心（CE）这四个部位在试验中的抗拉强度，并用这四个测量指标作为因变量。

### 9.4.1 原理与方法

在某些试验中，只测量一个因变量往往不足以得出科学的结论，而需要通过分析多个因变量才能得出合理结论，多元方差分析就是用来研究多个因变量之间是否存在显著差异的方法。它的基本原理与单因变量的方差分析相似，都是通过检验两个或多个样本均值之间的差异是否显著而得出有关结论的统计方法。

多元方差分析还使用了协方差所提供的信息，原因在于因变量之间可能存在着一定的相关性，这在进行统计检验时必须加以考虑。如果只是对同一个因变量测量两次，那得不到更多新的信息；如果测量的是两个相关的因变量，虽然能得到一些新信息，但是同时会有一些重复的信息，这些重复信息就可以在变量之间的协方差中表现出来。

SPSS 的 GLM Multivariate 多因变量分析过程可用于检验平衡的或不平衡的模型，所谓平衡模型就是指每个单元格都包含相同数目的观测。在多因变量模型中，效应平方和与误差平方和都是矩阵形式的，而非单因变量模型中的格式，这些矩阵称作 SSCP（sums-of-squares and cross-products，平方和和叉积）矩阵。如果指定了多个因变量，它将使用 Pillai's trace、Wilks' lambda、Hotelling's trace 等统计量进行分析，正如对单个因变量的方差分析。

GLM Multivariate 分析过程能输出许多结果，包括：多变量检验表、方差分析表、均数比较结果、均数的多变量检验结果、均数的单变量检验结果、参数估计等。

使用 GLM Multivariate 过程，需要验证如下的几个条件。

- 因变量应为数值型的；因素变量应为分类变量；协变量应为与因变量相关的数值变量。
- 多个因变量的取值向量应来自服从多维正态分布的总体。

总体中各单元格的方差-协方差矩阵都应相同。

## 9.4.2 多元方差分析实例

### 1. 问题描述和数据格式

本节对三组贫血患者的血红蛋白浓度和红细胞计数两个变量进行多元方差分析,研究三组患者的身体健康情况。所用数据文件为“贫血患者检测数据.sav”,数据格式如图 9-43 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	group	String	8	0	分组	None	None	8	Left	Nominal
2	h1	Numeric	8	2	血红蛋白浓度(%)	None	None	8	Right	Scale
3	w2	Numeric	8	2	红细胞计数(万/mm <sup>3</sup> )	None	None	8	Right	Scale

图 9-43 贫血患者的检测数据

### 2. SPSS 的多元方差分析设置

依次单击菜单“Analyze→General Linear Model→Multivariate...”,执行多元方差分析过程,其主设置面板如图 9-44 所示,其参数设置与第 9.3.2 节介绍的二因素方差分析过程相仿。

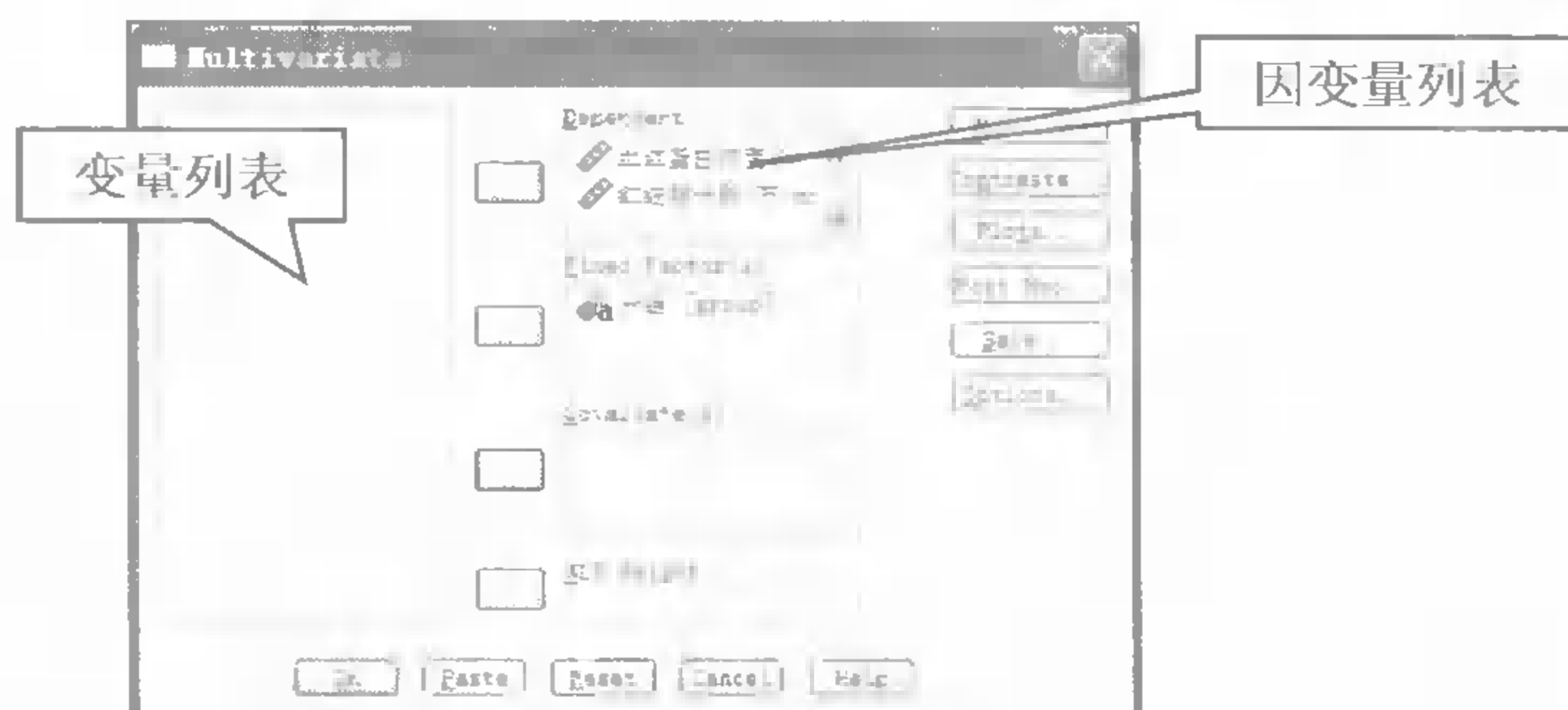




图 9-44 多元方差分析设置面板

在变量列表选中血红蛋白浓度和红细胞计数变量,单击从上至下第一个  按钮,将其作为因变量选入 Dependent 列表框;在变量列表单击选中分组变量,单击从上至下第二个  按钮,将其作为固定因素选入 Fixed 列表框。

### 3. 输出结果分析

单击图 9-44 中的 OK 按钮运行,SPSS Viewer 窗口的输出结果如图 9-45 和图 9-46 所示。

多变量检验 <sup>c</sup>						
效应		值	F	假设 df	误差 df	Sig.
截距	Pillai 的跟踪	.987	1001.859 <sup>a</sup>	2 000	26 000	.000
	Wilks 的 Lambda	.013	1001.859 <sup>a</sup>	2 000	26 000	.000
	Hotelling 的跟踪	77.066	1001.859 <sup>a</sup>	2 000	26 000	.000
	Roy 的最大根	77.066	1001.859 <sup>a</sup>	2 000	26 000	.000
group	Pillai 的跟踪	.566	5.323	4 000	54 000	.001
	Wilks 的 Lambda	.503	5.335 <sup>a</sup>	4 000	52 000	.001
	Hotelling 的跟踪	.853	5.333	4 000	50 000	.001
	Roy 的最大根	.642	8.662 <sup>b</sup>	2 000	27 000	.001

<sup>a</sup> 精确统计量  
<sup>b</sup> 该统计量是 F 的上限,它产生了一个关于显著性级别的下限。  
<sup>c</sup> 设计 截距+group

主体间因子		
分组		N
A		12
B		10
C		8

图 9-45 多变量检验结果

主体间效应的检验						
源	因变量	III 型平方和	df	均方	F	Sig.
校正模型	血红蛋白浓度(%)	7.926 <sup>a</sup>	2	3.963	7.302	.003
	红细胞计数(万/mm <sup>3</sup> )	1375.958 <sup>b</sup>	2	687.979	3.915	.032
截距	血红蛋白浓度(%)	513.898	1	513.898	946.778	.000
	红细胞计数(万/mm <sup>3</sup> )	1757050.816	1	1757050.8	1000.302	.000
group	血红蛋白浓度(%)	7.926	2	3.963	7.302	.003
	红细胞计数(万/mm <sup>3</sup> )	1375.958	2	687.979	3.915	.032
误差	血红蛋白浓度(%)	14.637	27	.542		
	红细胞计数(万/mm <sup>3</sup> )	47426.642	27	1756.120		
总计	血红蛋白浓度(%)	520.100	30			
	红细胞计数(万/mm <sup>3</sup> )	1818407.000	30			
校正的总计	血红蛋白浓度(%)	22.579	29			
	红细胞计数(万/mm <sup>3</sup> )	61180.000	29			

a. R 方 = .85 (调整 R 方 = .803)

b. R 方 = .225 (调整 R 方 = .167)

图 9-46 效应检验结果

(1) 多变量检验输出。如图 9-45 所示,“主体间因子”表格是贫血患者在不同分组里的频数统计情况。

“多变量检验”表格中,首先由显著性检验的 Sig 值都小于 0.01,推断 group (分组)效应对模型的影响是显著的;再观察 group 效应的 Pillai's trace 值(Pillai 的跟踪)和 Hotelling's trace 值(Hotelling 的跟踪),如果它们很接近就说明指定效应的作用并不大,本例二者取值的差异比较大(分别为 0.566 和 0.853),说明 group 效应对模型的影响作用是比较大的。

(2) 效应检验输出。如图 9-46 所示,是多元方差分析的效应检验输出,可见在 0.05 的显著性水平上,group (分组)对血红蛋白浓度和红细胞计数的影响都是显著的。

## 9.5 重复测量设计的方差分析

重复测量设计(Repeated Measure),指对同批研究对象先后施加不同的实验处理后进行测量,或者在不同场合(Occasion,如地点和时间)对其进行至少两次的测量。它可以分为两类:一类是对相同时间的不同因素水平组合的测量;另一类是对不同时间上的重复测量。

### 9.5.1 原理与方法

重复测量设计的方差分析中,可以是在相同条件下进行的重复测量,如此在研究不同处理之间是否存在显著差异的同时,也能研究被试者之间的差异;或者是不同条件下进行的重复测量,如此在研究不同处理之间是否存在显著差异的同时,也能研究重复测量的条件之间的差异,以及这些条件与处理之间的交互效应。

#### 1. 重复测量设计的优缺点

- 优点:把单个个体作为自身的对照,克服了个体之间的变异,分析时能更好地集中于研究效应;同时,把自身当作对照,研究所需的个体就相对较少了。
- 缺点:滞留效应(Carry-over effect),前面处理的效应有可能滞留到下一次的处理;潜隐效应(Latent effect),前面处理的效应有可能激活原本不活跃的效应;学习效应(Learning effect),由于逐步熟悉了实验方式,研究对象的反应能力在后面的处理中也可能逐步提高。



## 2. 重复测量设计方差分析的条件

- 正态性：不同处理水平下的个体取自相互独立的随机样本，其总体均数服从正态分布。
- 方差齐性：不同处理水平下的总体方差是相等的。
- 因变量的方差-协方差矩阵（Variance-Covariance matrix）满足球形（Sphericity）假设，即两个对象的协方差应该等于它们方差的均值减去一个常数。

如果球形假设不能满足，则相关的 F 统计量是有偏的，会造成过多地拒绝本来为真的假设（即增加了 I 型错误）；此时在计算 F 统计量时会对分子、分母作一定的调整。

## 3. 重复测量设计方差分析的假设检验

假设对同一组观测对象在  $k$  个不同的条件下进行了重复测量，获得  $k$  个样本。

零假设  $H_0$  为： $k$  个样本分别来自具有相同均值（记  $\mu$ ）和方差（记  $\sigma^2$ ）的相互独立的总体。SPSS 将  $k$  次重复测量的样本看作  $k$  个因变量，作多元检验，如果 F 统计量的值大于临界值，就否定零假设，反之亦然。

如果试验中还定义了组间因素变量，那么组间偏差平方和就反应了该分组变量各水平间的差异。此时的零假设  $H_0$  为：该分组变量各取值水平下的样本来自均值相同的总体。若组间均方和的取值远大于误差均方和，使 F 统计量的值大于临界值，就否定零假设，反之亦然。

## 4. 数据文件的结构

因变量应为数值型的；因素变量应为分类变量；协变量应为与因变量相关的数值变量。

在记录重复测量设计资料的数据文件中，若干次重复测量结果是作为不同因变量出现的，这是与一般方差分析的数据文件最大的不同。它还要求定义一个组内因素，其取值水平个数与重复测量的次数相同，例如：研究要求分别测量研究对象在某 5 天的体重，那么组内因素就应设为有 5 个取值的天数。

## 9.5.2 SPSS 实例分析

### 1. 问题描述和数据格式

在某个对视觉刺激反应时间的研究中，把视觉刺激作为因素变量，它有 3 个取值水平，将 12 位受试者随机分到 3 个刺激级别的实验组中，给每人一个编号，在相同的外界条件下对每人测试三次。经过整理的数据格式如图 9-47 所示，所用数据文件为“刺激反应测量数据.asv”。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	vision	Numeric	8	0	视觉刺激	{1 刺激1}	None	8	Right	Nominal
2	number	Numeric	8	0	编号	None	None	8	Right	Scale
3	time1	Numeric	8	2	测量1	None	None	8	Right	Scale
4	time2	Numeric	8	2	测量2	None	None	8	Right	Scale
5	time3	Numeric	8	2	测量3	None	None	8	Right	Scale

图 9-47 视觉刺激反应数据格式

本节研究在 3 种刺激水平下，人的平均反应时间是否有显著的差异，相应的零假设为：3 种视觉刺激下的平均反应时间无显著差异。

## 2. 参数设置

依次单击菜单“Analyze→General Linear Model→Repeated Measures...”，执行重复测量设计的方差分析过程，首先打开的是如图 9-48 所示的定义因素设置面板。

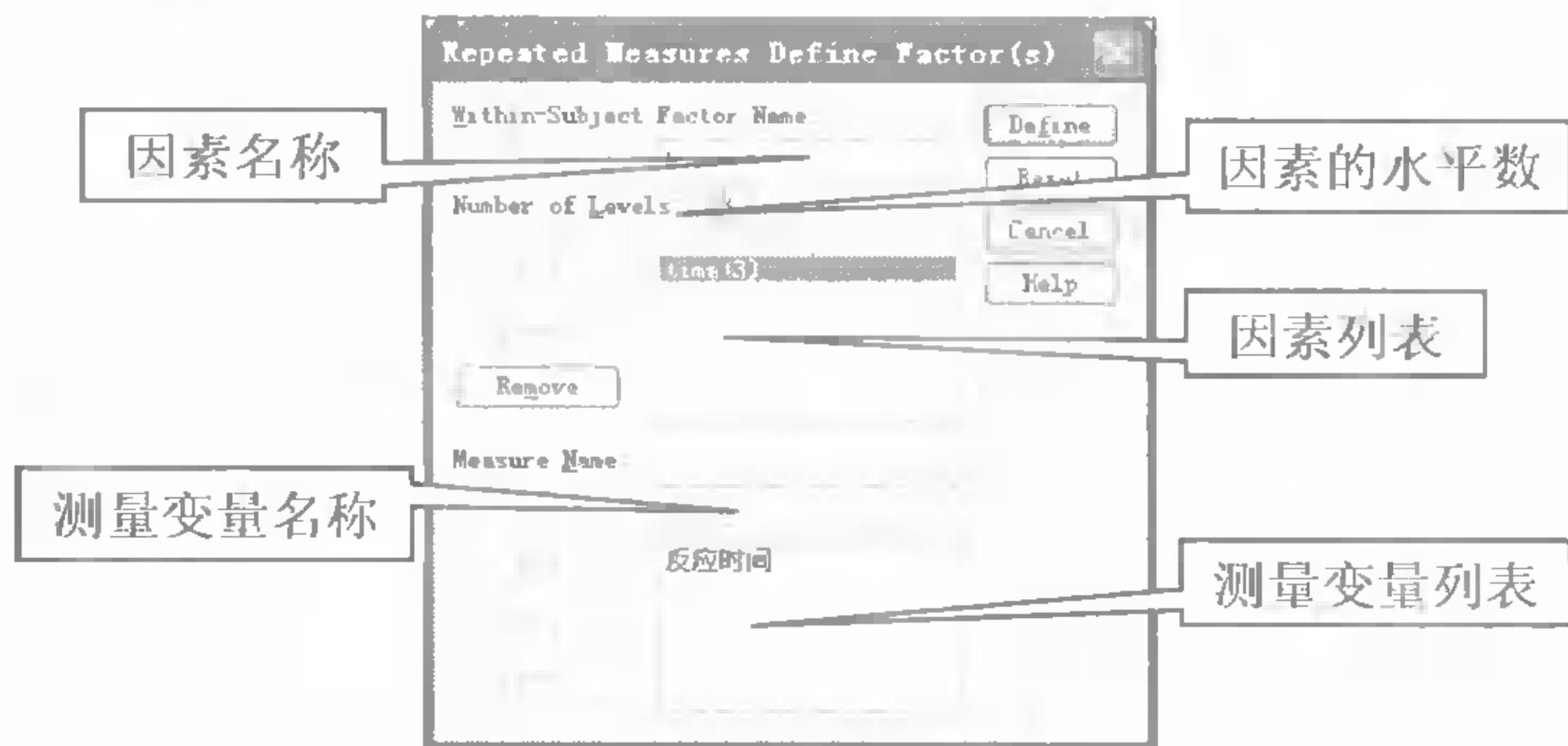




图 9-48 重复测量设计的定义因素设置面板

(1) 重复测量设计的因素定义。在 Factor Name 输入框键入“time”，在 Number of Levels 输入框键入“3”，单击随后的 Add 按钮，将 time(3)加入下面的因素列表框；在 Measure Name 输入框键入“反应时间”，单击随后的 Add 按钮，将其加入下面的测量变量列表框。

- ④ Within-Subject Factor Name 输入框，用于指定组内因素的名称。
- ④ Number of Levels 输入框，用于指定组内因素的取值水平个数。
- ④ Measure Name 输入框，用于指定测试变量的名称。

单击 Number of Levels 输入框下面的 Add 按钮，可以把指定因素添加到因素列表；在因素列表框选中某个因素后，可以对其进行编辑后再单击 Change 确认修改；或者直接单击 Remove 按钮将其删除。对测试变量的设置方法与此类似。

(2) 重复测量设计的主设置界面。单击图 9-48 中的 Define 按钮完成因素定义，并自动进入 Repeated Measures 过程的主设置界面，如图 9-49 所示。在变量列表单击选中测量 1 (time1) 变量，单击从上至下第一个  按钮，将其作为第一个测量变量选入右侧的列表框；用同样的方法依次选入变量测量 2 (time2)、测量 3 (time3)；在变量列表单击选中视觉刺激变量，单击从上至下第二个  按钮，将其作为组间因素选入 Between-Subjects Factor(s)列表框。

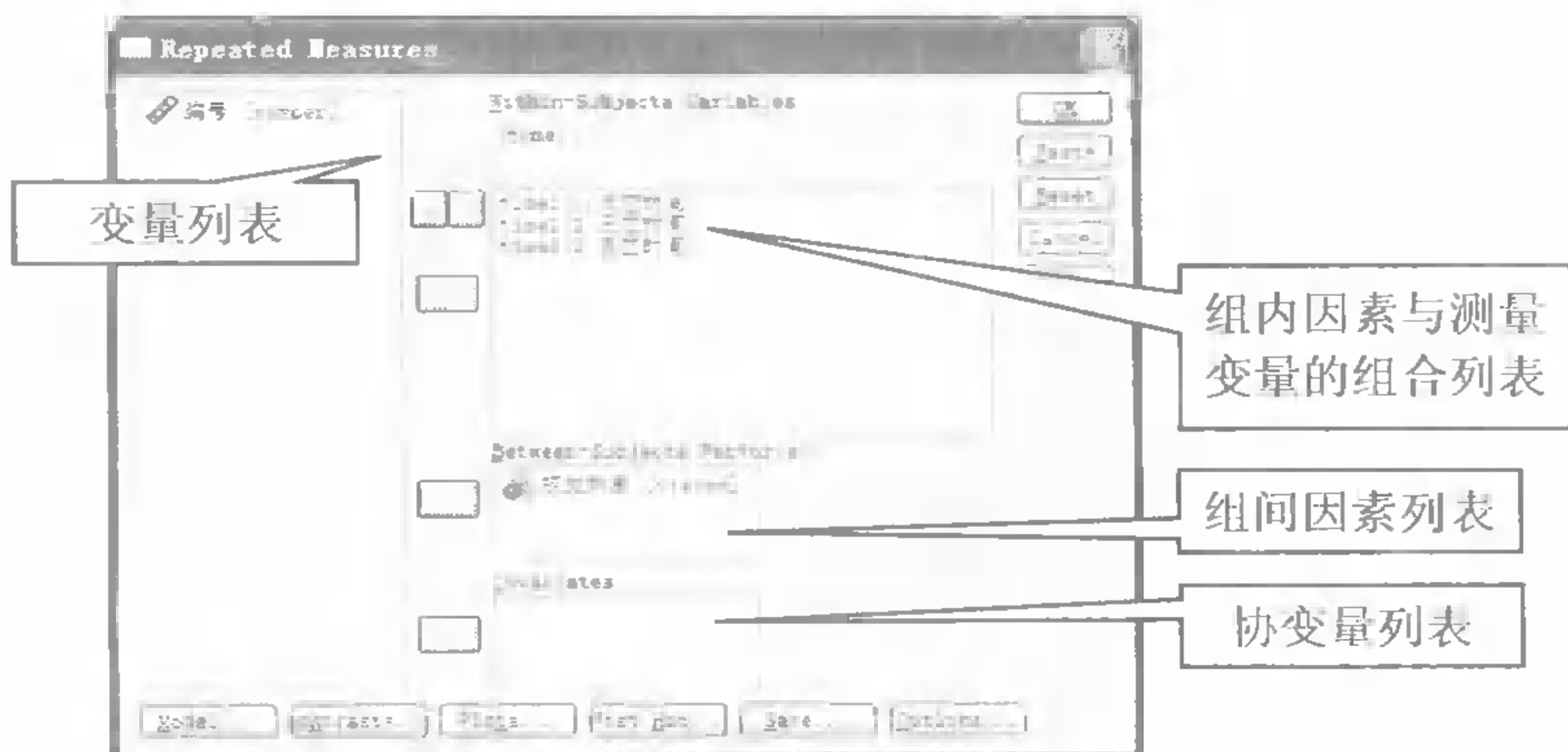



图 9-49 重复测量数据的方差分析设置主面板

- 指定分析变量。Within-Subjects Variables 列表框，用于选入不同的重复测量变量；Between-Subjects Factor(s)列表框，用于选入组间因素变量；Covariates 列表框，用于选入协变量。
- 设置组内因素取值水平与测量变量的对应关系。在 Within-Subjects Variables 列表框中，初始显示的是一列格式为“\_\_?\_\_(n,A)”（无外部引号）的列表，其中：n 表示组内因素的第 n 个水平，A 表示测量变量的名称，这些都是在图 9-48 所示的对话框里所定义的。选入测量变量后，将显示类似于“time1(1,反应时间)”的列表，表示将 time1 变量作为对反应时间的第一次测量记录。

注意：选入的测量变量要与相应的组内因素水平相对应，如果变量名称与括号里的因素水平不对应，可以先选中某个观测变量的名称，再单击  按钮调节它的显示顺序。

(3) 模型设置。单击图 9-49 中的 Model 按钮，弹出如图 9-50 所示的模型设置对话框。单击 Continue 按钮返回主界面。

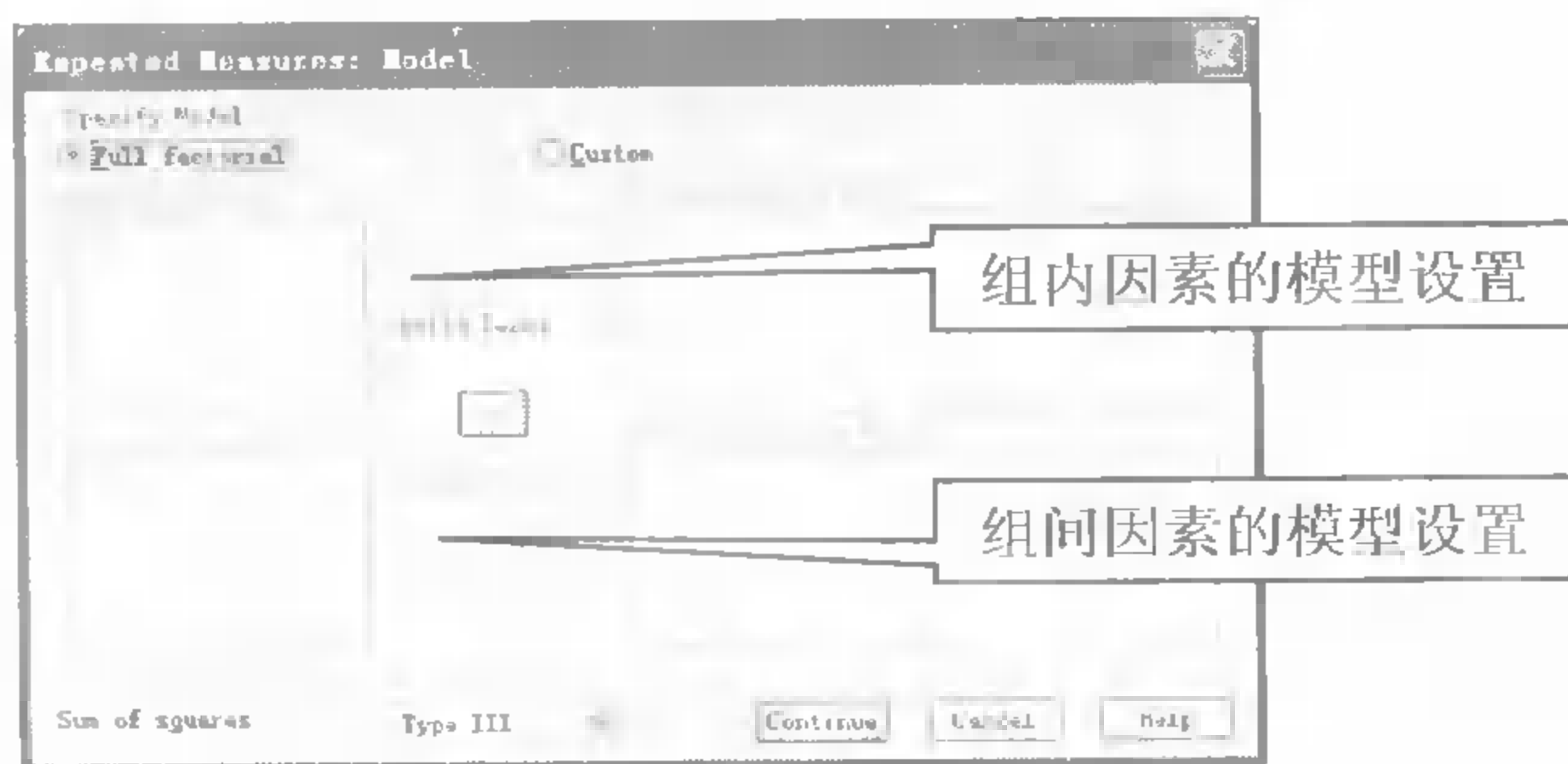


图 9-50 重复测量数据的方差分析的模型设置

此模型子界面的参数设置，与第 9.3.2 节介绍的二因素方差分析过程的 Model 设置相仿（如图 9-15 所示）。只是此处把因素变量分为了组间（Between-Subjects）和组内（Within-Subjects）两部分，需要分别对其进行设置，本例选择默认的 Full Factorial 方式。

如果模型包含多个协变量，是不能定义协变量之间交互作用的。但是，可以先依次单击菜单“Transform→Compute”计算某些协变量的乘积，再把乘积变量作为协变量引入分析。

(4) Options 选项设置。单击图 9-49 中的 Options 按钮，弹出如图 9-51 所示的对话框，分别勾选复选框：Descriptive、Parameter 和 Homogeneity；单击 Continue 按钮返回主界面。

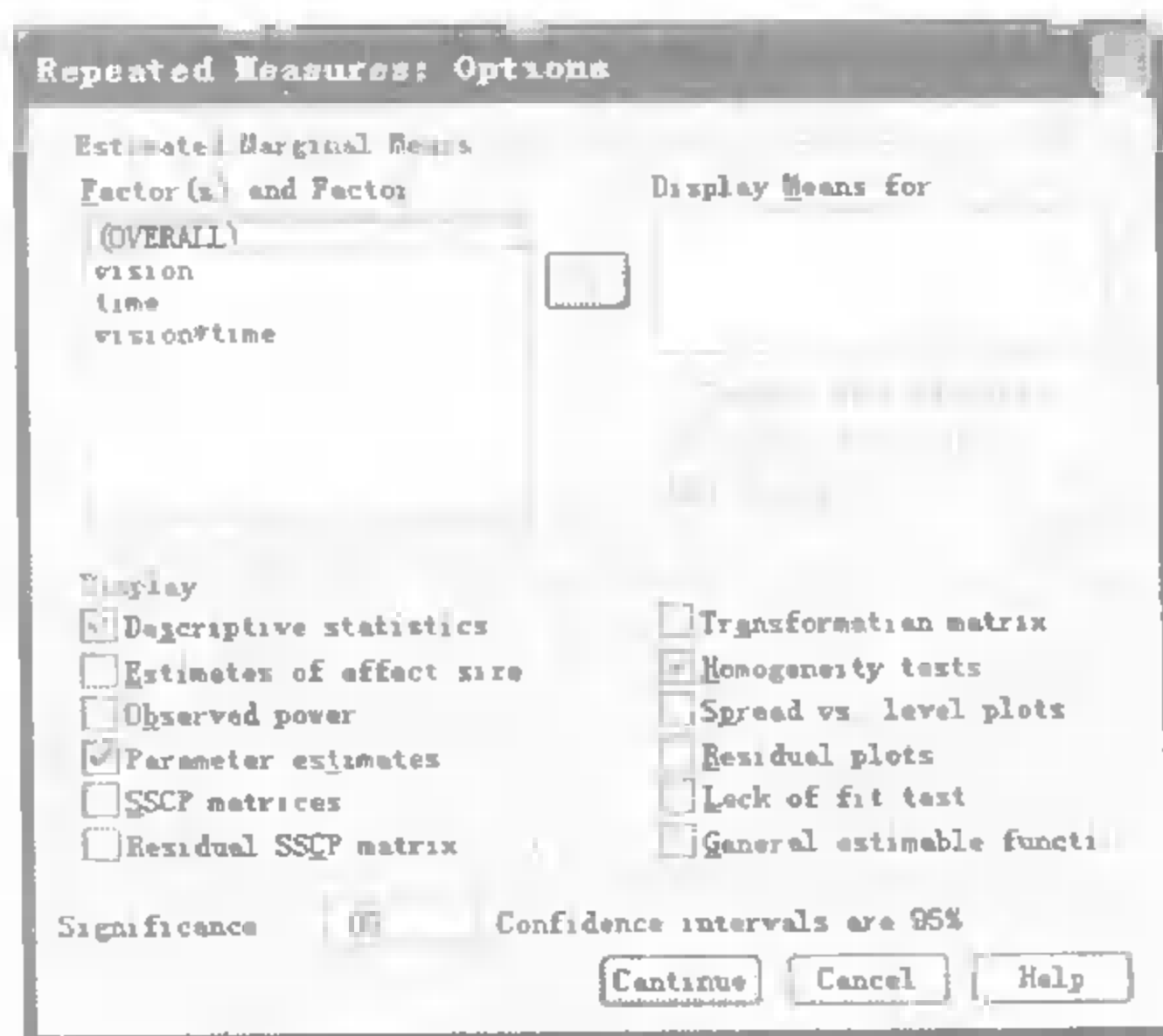


图 9-51 Options 选项设置

(5) 其他设置。在图 9-49 中，其他参数的设置（包括 Contrast、Plots、Post Hoc、Save），都与第 9.3.2 节介绍的二因素方差分析过程的有关设置相仿。

### 3. 输出结果分析

单击图 9-48 中的 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 9-52～图 9-57 所示。

主体内因子			描述性统计量			
度量 反应时间			视觉刺激	均值	标准差	N
time	因变量		刺激1	9250	41932	3
1	time1		刺激2	24000	40825	3
2	time2		刺激3	15750	41130	3
3	time3		总计	16333	73278	12
主体间因子			刺激1	11000	21602	4
视觉刺激	值标签	II	刺激2	28250	36856	4
1	刺激1	1	刺激3	16000	31623	4
2	刺激2	2	总计	18417	80618	12
3	刺激3	3	刺激1	7250	17078	4
			刺激2	24750	38622	4
			刺激3	16000	54772	4
			总计	16000	82902	12

图 9-52 重复测量数据基本统计特征

多变量检验 <sup>c</sup>						
效应		值	F	假设 df	误差 df	Sig
time	Pillai 的跟踪	421	2908 <sup>a</sup>	2 000	8 000	.112
	Wilks 的 Lambda	.579	2908 <sup>a</sup>	2 000	8 000	.112
	Hotelling 的跟踪	127	2908 <sup>a</sup>	2 000	8 000	.112
	Roy 的最大根	.727	2908 <sup>a</sup>	2 000	8 000	.112
time * vision	Pillai 的跟踪	.418	1.189	4 000	18 000	.348
	Wilks 的 Lambda	.519	1.083 <sup>a</sup>	4 000	16 000	.398
	Hotelling 的跟踪	.555	.971	4 000	14 000	.454
	Roy 的最大根	.407	1.032 <sup>a</sup>	2 000	8 000	.215

a. 精确统计量

b. 该统计量是 F 的上限，它产生了一个关于显著性级别的下限。

c. 设计截距+vision  
主体内设计: time

图 9-53 协方差检验和多变量检验输出

Mauchly 的球形度检验 <sup>b</sup>							
度量 反应时间							
主体内效应	Mauchly 的 W	近似卡方	df	Sig	Epsilon <sup>a</sup>		
					Greenhouse-Geisser	Huynh-Feldt	下限
time	.857	1.141	2	.565	.883	1.000	.500

检验零假设，即标准正交变换因变量的误差协方差矩阵与一个单位矩阵成比例。

a. 可用于调整显著性平均检验的自由度。在主体内效应检验表格中显示修正后的检验。

b. 设计截距+vision  
主体内设计: time

图 9-54 球形检验的结果

主体内效应的检验						
度量 反应时间						
源		III 型平方和	df	均方	F	Sig
time	采用的球形度	.412	2	.206	3.255	.062
	Greenhouse-Geisser	.412	1.785	.231	3.255	.070
	Huynh-Feldt	.412	2.000	.206	3.255	.062
	下限	.412	1 000	.412	3.255	.105
time * vision	采用的球形度	.283	4	.071	1.120	.378
	Greenhouse-Geisser	.283	3.541	.080	1.120	.377
	Huynh-Feldt	.283	4 000	.071	1.120	.378
	下限	.283	2 000	.142	1.120	.368
误差 (time)	采用的球形度	1.138	18	.063		
	Greenhouse-Geisser	1.138	15.888	.072		
	Huynh-Feldt	1.138	18 000	.063		
	下限	1.138	9 000	.126		

主体内对比的检验						
度量 反应时间						
源	time	III 型平方和	df	均方	F	Sig
time	线性	.40	1	.401	1.78	.189
	二次	.40	1	.405	6.69	.015
time * vision	线性	.085	2	.043	.705	.496
	二次	.085	2	.043	1.105	.273
误差 (time)	线性	.547	9	.061		
	二次	.591	9	.066		

图 9-55 组内效应检验的输出



误差方差等同性的 Levene 检验 <sup>a</sup>				
	F	df1	df2	Sig.
测量1	.020	2	9	.980
测量2	.379	2	9	.685
测量3	6.911	2	9	.015

检验零假设，即所有组中因变量的误差方差均相等。

a.

设计: 截距+vision  
主体内设计: time

主体间效应的检验					
度量: 反应时间 转换的变量: 平均数					
源	III 型平方和	df	均方	F	Sig.
截距	103.022	1	103.022	346.079	.000
vision	16.515	2	8.258	27.739	.000
误差	2.579	9	.298		

图 9-56 误差的同方差检验和组间效应的检验

参数估计							
因变量	参数	B	标准误	t	Sig.	95% 置信区间	
						下限	上限
测量1	截距	1.373	.206	7.627	.000	1.168	1.592
	[vision=1]	-.630	.292	-2.226	.033	-1.321	.061
	[vision=2]	.825	.292	2.825	.000	.264	1.486
	[vision=3]	.0 <sup>a</sup>					
测量2	截距	1.600	.153	10.428	.000	1.253	1.947
	[vision=1]	-.300	.217	-1.304	.047	-.891	.290
	[vision=2]	1.123	.217	5.146	.000	.733	1.716
	[vision=3]	.0 <sup>a</sup>					
测量3	截距	1.600	.200	8.014	.000	1.148	2.052
	[vision=1]	-.875	.282	-3.099	.013	-1.514	-.236
	[vision=2]	.875	.282	3.099	.013	.236	1.514
	[vision=3]	.0 <sup>a</sup>					

a. 此参数为冗余参数，将被设为零。

图 9-57 参数估计

(1) 基本统计信息输出。如图 9-52 所示，“主体内因子”和“主体间因子”表格给出了各因素不同取值水平下的样本个数统计信息；“描述性统计量”表格给出了各个分组的观察样本的基本统计特征，包括均值、标准差等。

(2) 协方差检验和多变量检验结果。如图 9-53 所示，“Box 检验”表格中显著性检验的 Sig 值  $0.308 > 0.10$ ，所以推断因变量在各分组中的协方差矩阵没有显著差异。

“多变量检验”表格给出了对组内因素 (time)、交互效应 (time\*vision) 的检验。此处分别采用了四种不同的算法，但它们显著性检验的 Sig 值都大于 0.05，由此可得出结论，组内效应 (time) 对造成视觉刺激反应时间的差异没有显著意义，组间与组内的交互效应对造成视觉刺激反应时间的差异也没有显著意义。

(3) 球形检验的输出。如图 9-54 所示，“Mauchly 的球形度检验”表格中，对 Mauchly W 统计量的近似卡方检验的显著性 Sig 值  $0.565 > 0.10$ ，故而不能否定球形假设。

(4) 组内效应的检验和比较。如图 9-55 所示，在“主体内效应的检验”表格中，对于每个效应的检验，第一行是在满足球形假设的条件下，不对 F 统计量的分子、分母做调整时的检验结果；随后三行是在不满足球形假设时，对 F 统计量的分子、分母做了不同调整后的检验结果。

本例满足球形假设，故参考第一行的显著性检验结果，在 0.05 的显著性水平上，不能否定组内因素 (time) 对视觉刺激反应时间无影响的假设。

(5) 误差的同方差检验和组间效应的检验。如图 9-56 所示，“Levene 检验”表格显示，在 0.05 的显著性水平上，测量 1、测量 2 在所有分组中的误差方差都无显著差异，而测量 3 在所有分组中的误差方差显著不同。

“主体间效应的检验”表格是对组间效应的方差分析结果，从 F 检验的显著性 Sig 值远小于 0.01 可以推断，不同刺激水平下的视觉反应时间有着非常显著的差异。

最后给出的是如图 9-57 所示的参数估计结果。

## 9.6 方差成分分析

方差成分分析适用于混合模型的分析，可以研究模型中的随机效应对因变量变异的贡献。

方差成分分析过程提供了 4 种方法：最小标准二次无偏估计（MINQUE）、方差分析（ANOVA）、极大似然估计（Maximum likelihood，简称 ML）、限制的极大似然估计（Restricted maximum likelihood，简称 REML）。

### 9.6.1 原理简介

SPSS 的方差成分分析过程，要求因变量为数值变量；因素变量为分类变量，可以是数值型或短字符型（不超过 8 个字符）；至少有一个因素变量是随机的（Random Factors），随机因素的取值水平是由随机采样得来的；协变量是数值型变量，并与因变量有相关关系。

方差成分分析过程需要验证的假设包括：随机效应的参数的均值为零、方差为有限常数，同一效应的不同参数互不相关，不同随机效应的参数也不相关；残差项也需满足零均值和有限方差的假设，且它与任意随机效应的参数都不相关，不同观测的残差之间也不相关。基于这些假设，随机因素相同取值水平上的观测值是彼此相关的，这与普通线性模型是不同的。

ANOVA 和 MINQUE 这两种方法不严格要求参数和残差服从正态分布，它们能缓解违反正态假设所带来的影响；ML 和 REML 方法都要求模型参数和残差项需服从正态分布。

在做方差成分分析之前，建议先使用 Explore 过程观察数据的基本特征，并使用 GLM 菜单下的 Univariate、Multivariate 或 Repeated Measures 几个过程进行假设检验。




### 9.6.2 SPSS 实例分析

本节对第 9.3.4 节使用的心理实验数据进行方差成分分析，数据格式如图 9-35 所示。

在此把目标和设备变量作为固定因素变量；把光线变量作为随机因素变量，表示样本中光线的取值水平是从某个总体中随机抽取的。

依次单击菜单“Analyze→General Linear Model→Variance Components...”，执行方差成分分析过程，其主设置面板如图 9-58 所示，它与第 9.3.2 节介绍的二因素方差分析过程的有关设置相仿（如图 9-13 所示）。

#### 1. 变量设置

在变量列表单击选中得分变量，单击从上至下第一个  按钮，将其作为因变量选入 Dependent 选框；在变量列表选中目标和设备变量，单击从上至下第二个  按钮，将其作为固定因素选入 Fixed 列表框；在变量列表单击选中光线变量，单击从上至下第三个  按钮，将其作为随机因素选入 Random 列表框。

#### 2. 模型设置

单击 Model 按钮，打开模型设置子面板，与第 9.3.2 节介绍的二因素方差分析过程的模型设置相仿（如图 9-15 所示），本例选择默认的 Full Factorial 全模型方式。

#### 3. Options 选项设置

在图 9-58 中，单击 Options 按钮，弹出如图 9-59 所示的选项设置面板，在此设置方差成

分分析的方法。单击选中 ANOVA 和 TypeIII 单选框；分别勾选 Sum of square、Expected mean square 复选框；单击 Continue 按钮返回主界面。



图 9-58 方差成分分析的主设置面板

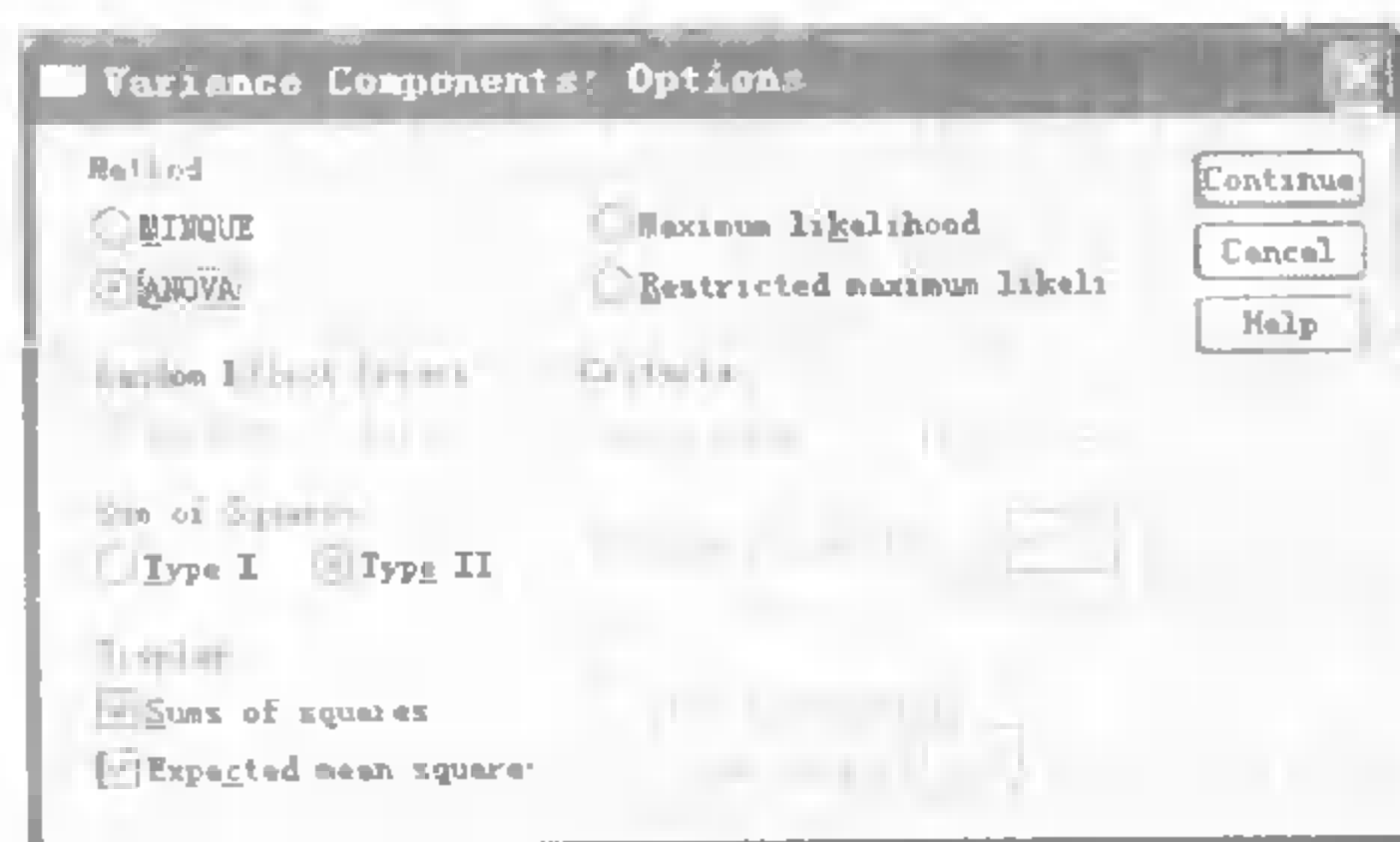


图 9-59 Options 选项设置

(1) Method 子设置栏，用于指定如下 4 种方法中的一个。

① 最小标准二次无偏估计 (MINQUE)。此方法对固定效应的估计是不变的；如果数据服从正态分布并且估计正确的话，使用此方法得到的估计值，是所有无偏估计中方差最小的估计，此为默认方法。

② 方差分析 (ANOVA)。此方法使用 Type I 或 TypeIII 平方和计算每个效应的无偏估计。ANOVA 方法有时会产生负的方差估计，这表明模型不准确或估计方法不适合，或者需要更多的数据。

③ 极大似然 (Maximum likelihood, 简称 ML)。它使用迭代算法产生与实际观测数据最一致的参数估计，但这些估计是有偏的。该方法是接近正态的，它不考虑估计固定效应时的自由度。ML 估计和 REML 估计的结果，对数据转换保持不变。

④ 限制的极大似然 (Restricted maximum likelihood, 简称 REML)。对许多 (并非所有) 平衡数据，该方法比 ANOVA 的估计值要小。因为此方法对固定效应作了调整，计算的标准误可能比 ML 方法要小，它会考虑估计固定效应时的自由度。

(2) Random Effect Priors 子设置栏，只对 MINQUE 方法有效。

● Uniform 选项，指定所有随机效应和残差项对观测量的影响都相等，此项为默认选项。

● Zero 选项，相当于假设随机效应的方差为零。

(3) Sum of Squares 子设置栏，只对 ANOVA 方法有效，指定使用 Type I 或 TypeIII 平方和。

(4) Criteria 子设置栏，只对 ML 或 REML 方法有效。

● Convergence 下拉列表，指定收敛的临界值，可选项为 1E-6 ~ 1E-10。

● Maximum iterations 输入框，指定最大迭代次数。

(5) Display 子设置栏。当选择 ANOVA 方法时，Sum of square 复选框表示在结果中输出平方和；Expected mean square 表示输出期望均方值。

当选择 ML 或 REML 方法时，Iteration history 复选框表示在结果中输出迭代的过程。

#### 4. 保存选项设置

单击图 9-58 中的 Save 按钮，打开如图 9-60 所示的保存设置面板，单击 Continue 按钮返回主界面。

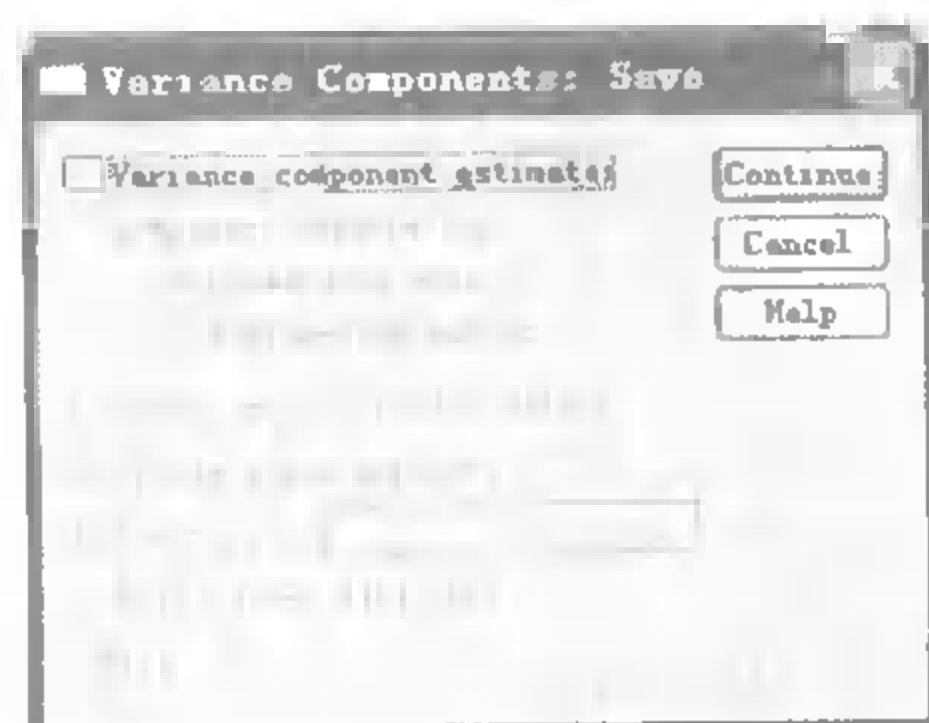


图 9-60 Save 保存选项设置

- Variance component estimates 复选框，指定保存方差成分估计值。
- Component covariation 复选框，只对 ML 或 REML 方法有效。Covariance matrix 单选框，指定保存协方差矩阵，Correlation matrix 单选框，指定保存相关矩阵。
- Destination for created values 下面指定保存位置：Create a new dataset 单选框，表示建立一个新的数据集，在 Dataset name 后指定新数据集的名称；Write a new data file 单选框，表示将相关数据保存到新文件中，单击 Files 按钮指定文件路径和名称。

## 5. 结果输出解释

单击图 9-58 中的 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 9-61 和图 9-62 所示。

因子级别信息			
	值标签	N	
光线	1	11	60
	2	12	60
目标	1	11	30
	2	12	30
	3	13	30
	4	14	30
设备	1	d1	40
	2	d2	40
	3	d3	40
因变量 score			

ANOVA			
源	III 型平方和	df	均方
校正的模型	783.467	23	34.064
截距	2152.133	1	3162.133
light	76.800	1	76.800
target	235.200	3	78.400
device	86.467	2	43.233
light * target	93.867	3	31.289
light * device	12.600	2	6.300
target * device	101.200	6	17.367
light * target * device	174.333	6	29.056
误差	70.400	96	733
总计	4016.000	120	
校正的总计	853.867	119	
因变量 score			

图 9-61 因素水平统计和方差分析表

期望均方差						
源	方差分量					二次项
	Var(light)	Var(light * target)	Var(light * device)	Var(light * target * device)	Var(误差)	
截距	64.000	15.000	30.000	5.000	1.000	截距 target device target * device
light	60.000	15.000	30.000	5.000	1.000	target target * device
target	0.00	11.000	0.00	5.000	1.000	target target * device
device	0.00	0.00	30.000	5.000	1.000	device target * device
light * target	0.00	75.000	0.00	5.000	1.000	target * device
light * device	0.00	0.00	30.000	5.000	1.000	target * device
target * device	0.00	0.00	0.00	5.000	1.000	target * device
light * target * device	0.00	0.00	0.00	5.000	1.000	target * device
误差	0.00	0.00	0.00	0.00	1.000	
因变量 score						
期望均方差根据 III 型平方和计算。						
对于每个源，期望均方差等于单元格中系数之和乘以方差分量，再加上与二次项单元格中的效应相关的二次项。						

方差估计	
分量	估计
Var(light)	11.8
Var(light * target)	149
Var(light * device)	-1.138 <sup>a</sup>
Var(light * target * device)	5.624
Var(误差)	733
因变量 score	
方法 ANOVA (III 型平方和)	
a 对于 ANOVA 和 MINQUE 方法，可能会出现负方差分量估计值，出现负方差分量估计值的可能原因有：(a) 所指定的模型不是正确的模型，或 (b) 方差的真值等于 0。	

图 9-62 期望均值表和方差估计表

(1) 因素水平统计信息和方差分析结果。如图 9-61 所示，“因子级别信息”表格列出了每个因素的取值水平及其值标签，还有相应的观测个数。“ANOVA”表格是方差分析的平方和分解结果，它给出了各效应的误差平方和、自由度及均方值。

(2) 方差估计结果。如图 9-62 所示，“方差估计表”给出各效应的方差估计值。其中 light\*device 的方差估计为负值，可能原因是：所指定的模型不是正确的模型，或方差的真值等于零。鉴于此，光线与设备的交互效应可以不予考虑，因为不仅方差成分估计其为负值，而且平方和分解表中其均方值也很小（6.3），建议更改模型进行第二次方差成分估计。



(3) 模型改进及分析。在图 9-58 中, 单击 Model 按钮, 打开如图 9-63 所示的模型设置子面板, 单击选中 Custom 单选框; 把除了 light\*device 之外的效应全部选入 Model 列表; 单击 Continue 按钮返回主界面。

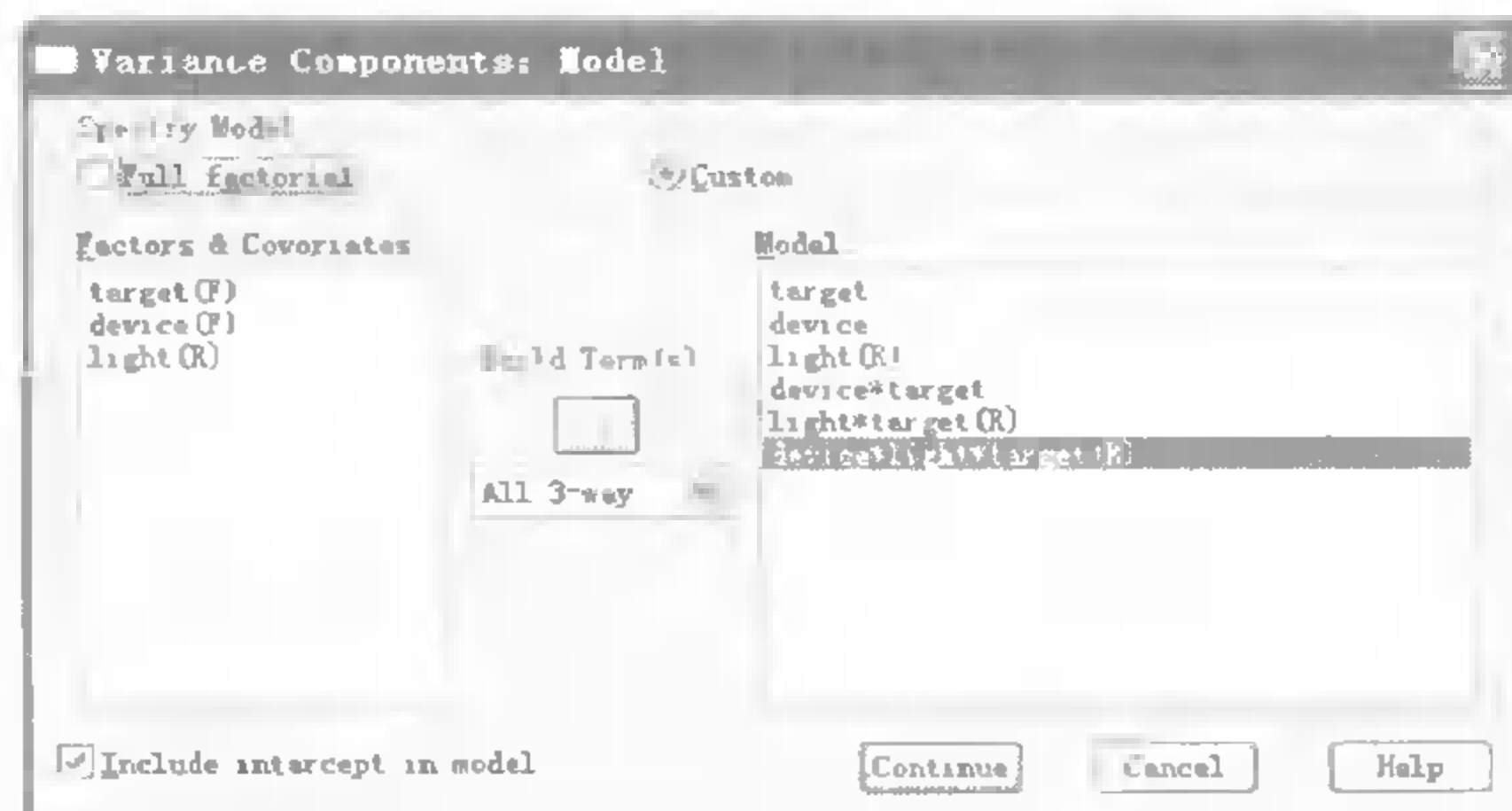


图 9-63 用户定义模型设置

再次在图 9-58 中单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 9-64 所示。

ANOVA			
源	III 型平方和	df	均方
校正的模型	783.467 <sup>a</sup>	28	27.981
截距	3162.133	1	3162.133
target	231.200	3	77.067
device	86.467	2	43.233
light	76.800	1	76.800
target * device	106.200	6	17.700
light * target	93.867	3	31.289
light * target * device	186.933	8	23.367
误差	10.400	36	.290
总计	4016.000	120	
校正的总计	853.867	119	

方差估计	
分量	估计
Var(light * target)	.928
Var(light * target * device)	4.027
Var(light)	.759
Var(误差)	.723

因变量: score  
方法 2: III(均方) (III 型平方和)

图 9-64 自定义模型的方差成分估计结果

对于改进的自定义模型, 方差的最大来源是光线、目标、设备这三个因素的 3 阶交互效应, 而光线因素也是不可忽视的, 应该在试验中作为重要的因素条件加以考虑。

## 9.7 正交实验设计

在实际生产和科学研究中, 经常会遇到多因素试验的问题, 这时往往不需要进行包括所有取值水平组合的全面试验, 只需从这些取值水平的组合搭配中, 选取一小部分做试验就可以了。那么, 怎样选择取值水平的组合以及如何分析试验结果, 才能科学的回答如下问题呢?

- 各研究因素对测量指标的影响, 哪个因素重要? 哪个因素次之?
- 对于单个因素, 它的哪个取值水平较好?
- 不同因素取值水平的哪种组合搭配, 可使试验结果达到最佳?

解决这些问题正是正交试验设计的主要内容。

### 9.7.1 正交实验设计简述

试验设计是数理统计学科的一个重要分支。多数的数理统计方法, 主要用于分析已经得到的试验数据, 而试验设计则是研究如何搜集数据的方法。试验设计主要讨论如下 2 个问题: 如何合理地安排试验; 如何分析试验所得的数据。

进行正交实验设计，一般要使用正交表和相应的交互作用表。SPSS 没有提供这些表格，这里所介绍的，也不是通常的利用给定表格进行表头设计的正交实验设计，而是 SPSS 为满足统计分析的需要而提供的实验设计过程。使用 SPSS 的正交实验设计过程进行模型设计，至少要有有一个分类变量作为因素变量，而且要有明确的实验次数要求。


### 9.7.2 SPSS 实例分析

本节仍对第 9.3.4 节使用的心理实验数据进行研究，数据格式如图 9-35 所示。参考原始数据，随后将通过 SPSS 的实验设计功能生成关于样本中指定因素的正交设计。

#### 1. 观察原始数据的因素水平组合设计

打开数据文件“心理实验数据.sav”，依次单击菜单“Data→Orthogonal Design→Display...”执行 Display 过程，主设置界面如图 9-65 所示。

本过程可以把由 SPSS 生成的正交实验设计或文件中原本存在的实验设计显示出来，输出可以是粗略的统计图表，也可以是详细的单个记录的描述，后者可用于联合分析。

(1) 参数设置。如图 9-65 所示，在变量列表选中 target、device 和 light 变量，单击  按钮，将其作为因素变量选入 Factors 列表；分别勾选 Listing、Profiles 复选框。

其中，Factors 列表框，用于选入文件里的因素变量。Format 栏设置结果的输出格式：Listing for experimenter 复选框，表示输出一个总的列表；Profiles for subjects 复选框，表示为每个记录都输出详细的统计描述。

(2) 输出结果。单击图 9-65 中的 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 9-66 所示。

“卡列表”表格只显示了因素取值水平组合的部分输出；“概要文件编号 n”表格是为每种因素取值组合输出的概要表。

卡列表				
	卡	目标	设备	光线
1		t1	d1	l1
2		t2	d1	l1
3		t3	d1	l1
4		t4	d1	l1
5		t1	d2	l1
6		t2	d2	l1
7		t3	d2	l1
8		t4	d2	l1
9		t1	d3	l1

概要文件编号 1			
卡	目标	设备	光线
	t1	d1	l1

概要文件编号 2			
卡	目标	设备	光线
	t2	d1	l1

图 9-66 Orthogonal Design Display 过程的输出结果

#### 2. 利用 SPSS 进行正交实验设计

对上例的心理实验数据文件中提到的因素进行正交实验设计。

依次单击菜单“Data→Orthogonal Design→Generate...”，执行正交实验设计过程，其主设置界面如图 9-67 所示。

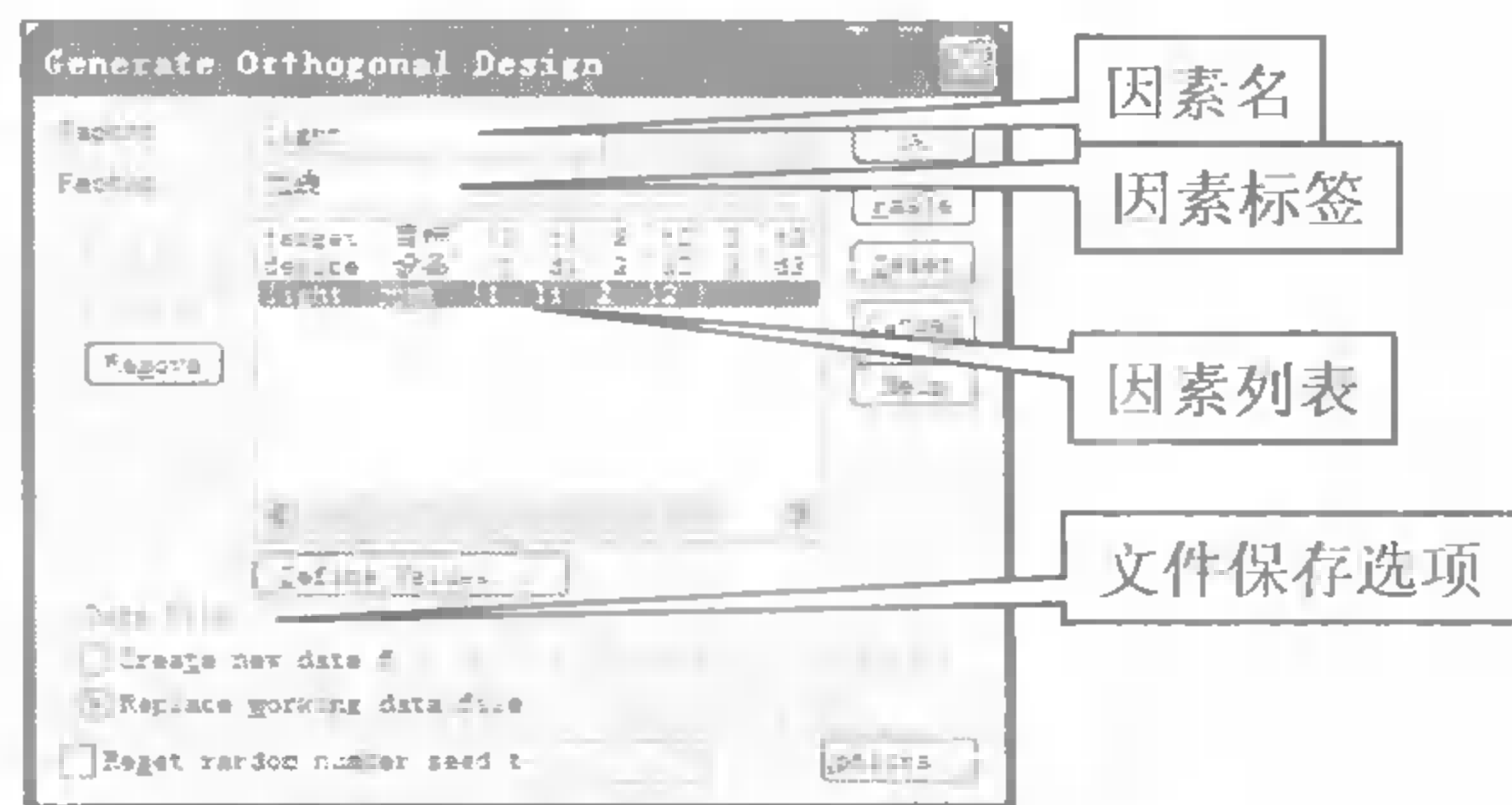


图 9-67 正交实验设计的主设置面板

(1) 因素变量的名称和取值设置。在第一行的 Factor name 输入框键入“light”，在第二行的 Factor label 输入框键入“光线”，单击 Add 按钮将其加入下面的因素列表；用同样的方法将 device（设备）和 target（目标）加入因素列表。单击选中 Replace working data file 单选框。

在因素列表单击选中 light 所在的行，单击 Define Values 按钮，弹出如图 9-68 所示的因素取值定义对话框，在第一列输入 1、2，在第二列输入 t1、t2，单击 Continue 按钮返回主界面；用同样的方法设置 device（1~3）和 target（1~4）的取值。

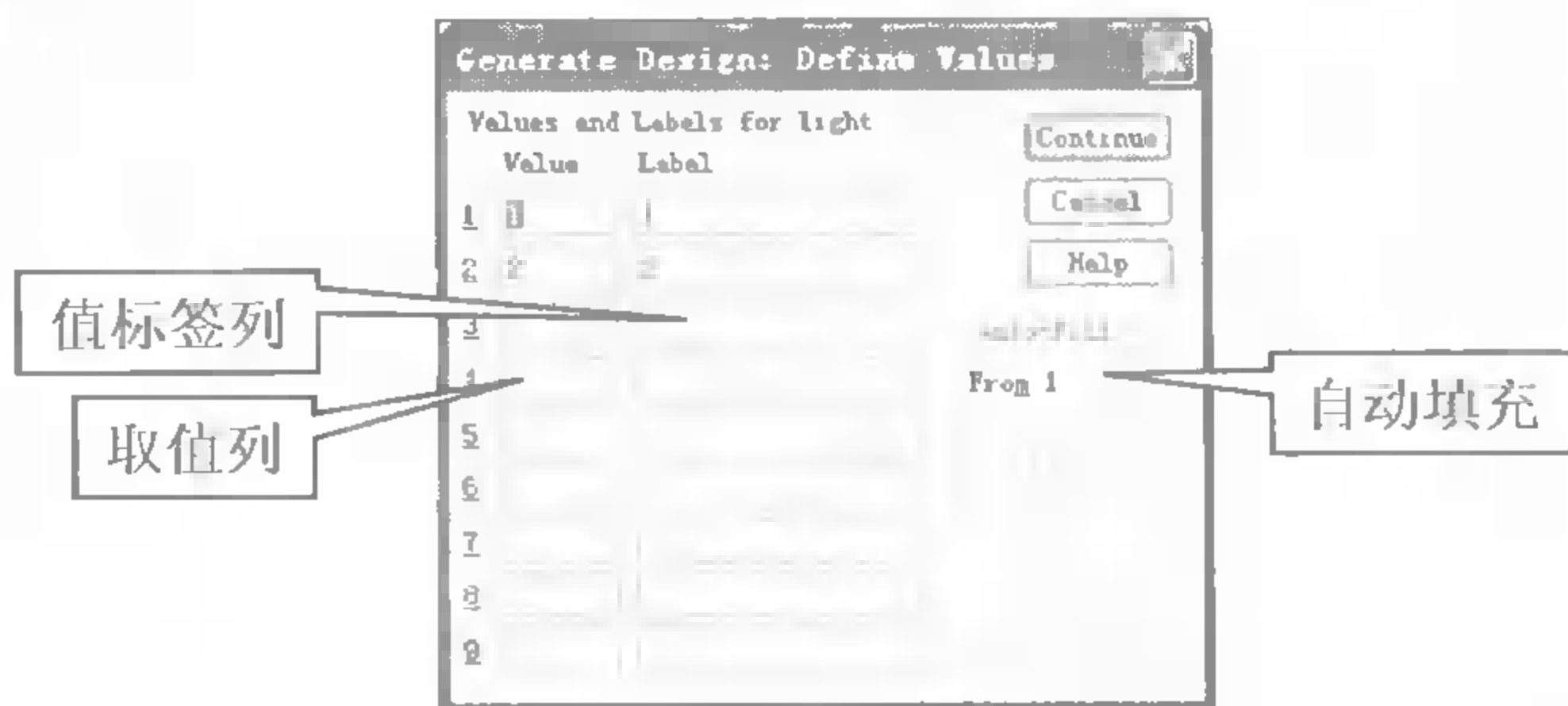


图 9-68 正交实验设计的因素取值设置

① 指定因素变量名和标签。在图 9-67 中，Factor name 输入框用于指定因素变量的名称；Factor label 输入框用于指定因素变量的标签名；但是，此处不能使用 Status\_ 和 Card\_ 作为因素变量名。单击 Add 按钮，可把指定的因素添加至下面的列表框；在列表里选中某个因素名后，可以对其进行重新编辑后单击 Change 按钮确认修改；或者直接单击 Remove 按钮将它删除。

② 指定因素变量的取值。在图 9-68 中，给出了两列输入框：Value 和 Label，分别用于指定变量取值和值标签。变量值只能为数值型的，变量标签可以为数值型的或字符型的。

Auto-Fill 输入框，设置自动填充的方式。例如在此输入 4 后单击 Fill 按钮，就可以在 Value 列自动填入 1~4。如果因素取值不是从 1 开始的连续数字，就不能使用自动填充了。

③ Data File 栏，指定产生的实验设计数据的保存方式，有两个选项。

- ① Create new data file，把数据存至新的文件，单击 File 按钮指定文件名称和路径。
- ② Replace working data file，新生成一个数据集保存正交设计的结果。

④ Reset random number seed to 输入框，指定生成正交实验设计的随机数种子。相同的随机数种子产生相同的实验设计结果，反之依然。如果需要多次重复相同的实验设计方案，就需要在此处设置固定的随机数种子；取值范围为 0~2 000 000 000 的整数。

(2) 观测数限制设置。在图 9-67 中单击 Options 按钮，打开如图 9-69 所示的对话框，在此设置对观测记录数的限制。单击 Continue 按钮返回主界面。

① Minimum number of cases to generate 输入框。用于指定实验设计的最少实验次数，在此输入一个正整数，需不大于全面试验的次数。若此处不作任何输入，SPSS 自动生成完成当前设计所必要的最少试验次数。

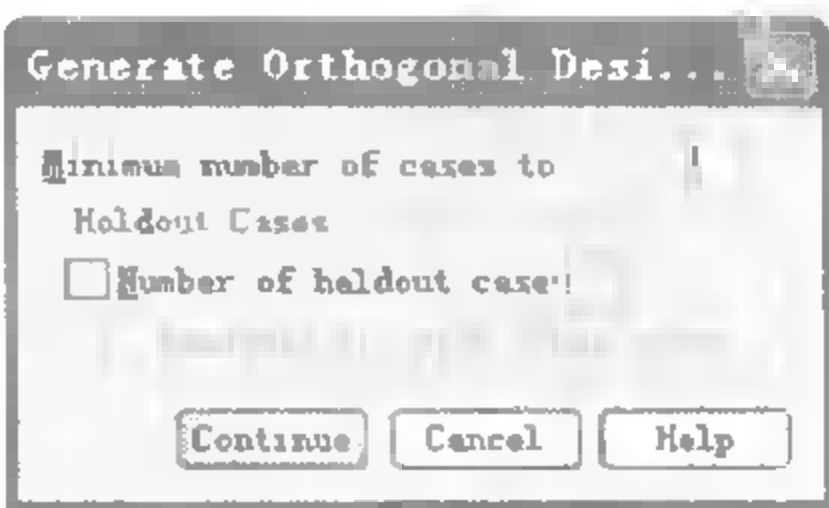


图 9-69 Options 选项设置

② Holdout Cases 子设置栏，设置关于维持观测量的选项。

- Number of holdout cases 复选框，指定除了正常观测之外的维持观测量的个数，Conjoint 过程估计效应时不使用这些额外的数据。选中后在其后面的输入框指定一个正整数，需不大于由所有因素水平组合决定的最大观测数。
- Randomly mix with other cases 复选框，表示将维持观测量与正常的观测量随机混合；如果不选中此项，维持观测量将全部出现在正常观测量的后面。

(3) 结果分析。在图 9-67 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 9-70 所示；新生成数据集的格式如图 9-71 所示。

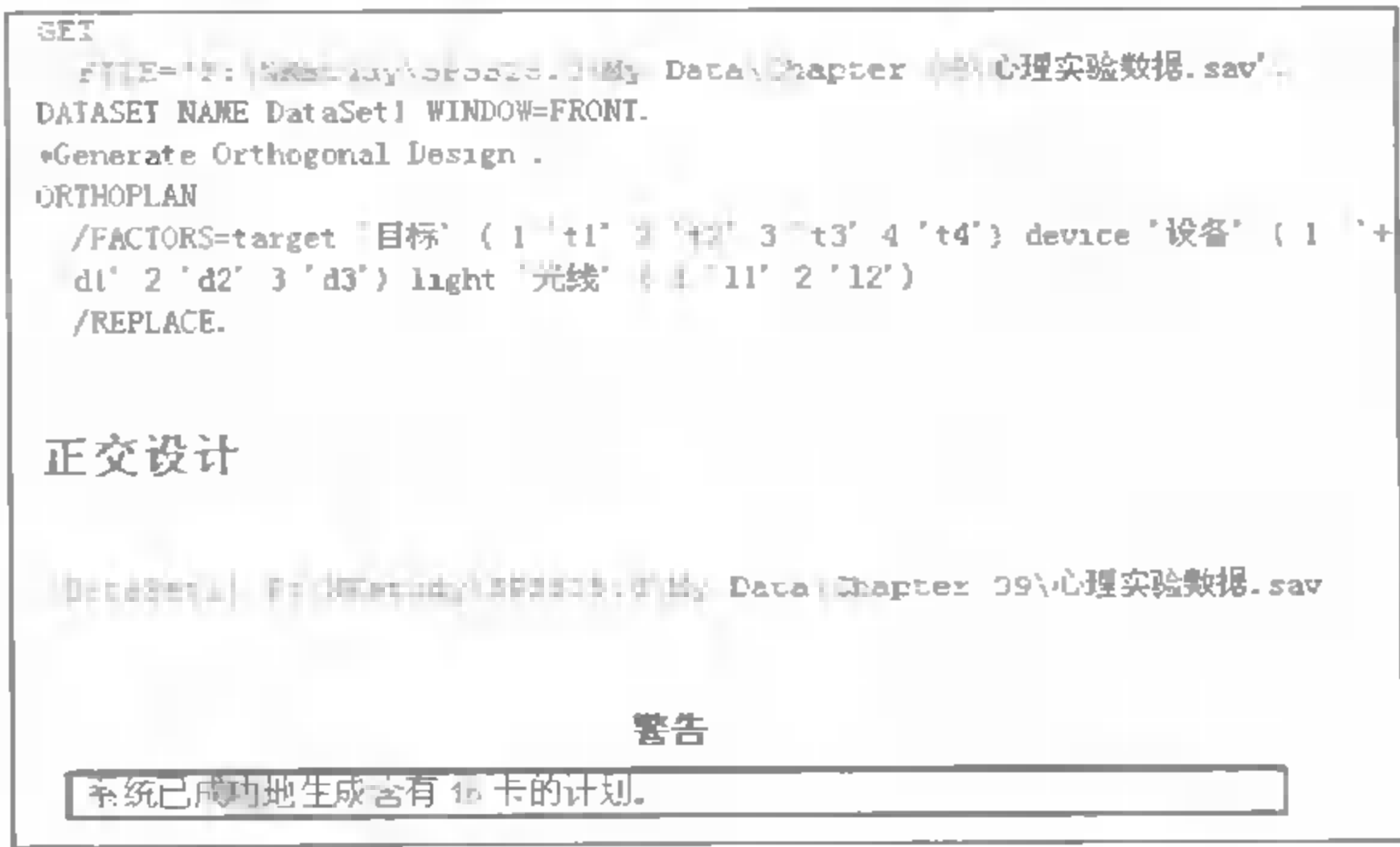


图 9-70 SPSS Viewer 窗口的输出结果

	target	device	light	STATUS	CARD
1	4.00	2.00	2.00	0	1
2	2.00	2.00	1.00	0	2
3	4.00	1.00	2.00	0	3
4	1.00	3.00	1.00	0	4
5	3.00	3.00	2.00	0	5
6	4.00	1.00	1.00	0	6
7	1.00	1.00	1.00	0	7
8	2.00	1.00	2.00	0	8
9	4.00	3.00	1.00	0	9
10	3.00	1.00	2.00	0	10
11	2.00	1.00	1.00	0	11
12	3.00	1.00	1.00	0	12
13	1.00	2.00	2.00	0	13
14	2.00	3.00	2.00	0	14
15	1.00	1.00	2.00	0	15
16	3.00	2.00	1.00	0	16

图 9-71 Data Editor 窗口的正交实验设计结果

① 如图 9-70 所示，“警告”表格提示用户：经过正交实验设计，生成了 16 个因素水平间的组合。本例的 3 个因素取值水平的总组合数为  $4 \times 3 \times 2 = 24$  个。原始数据文件采用的试验方案有 20 种，对 6 位被试者共作了  $20 \times 6 = 120$  次试验，如果采用本例的实验设计方案，一共作  $16 \times 6 = 96$  次试验就可以了，可见正交实验设计是能够较好的提高实验效率的。

② 如图 9-71 所示，前 3 列是实验设计方案的因素取值水平；按照前 3 列指定的条件完成实验后，把测量数据填入“STATUS\_”列的相应单元格中；“CARD\_”列是对 16 次试验的自动编号。

### 9.7.3 正交实验设计的方差分析

根据以上介绍的方法，利用正交实验设计，可以把试验安排地“均衡分散”且“综合可比”，使得只作较少次数的试验就能获得所需的结论，而且操作方法也是简便易行。但是，在任何试验过程中，都存在着随机因素造成的试验误差，通常可以将它们忽略不计，可一旦误差较大就会影响结论的可靠性。

使用方差分析过程，可将由试验误差所引起的指标变动与各研究因素所引起的指标变动区分开来，从而找出影响试验结果的真正重要的因素。对正交实验设计数据的方差分析与多因素方差分析过程相同，请参考第 9.3 节的介绍。



研究一个问题的过程，经常是从对单变量的分析开始，进一步到分析双变量之间的关系，然后再拓展到分析多变量之间的关系。多变量分析与单变量分析最大的不同之处，就是客观事物之间的关联性开始被披露出来。在统计学中，研究客观事物之间相互关联的数量特征具有十分重要的理论意义和实践意义。本章就把相关关系的讨论深入下去，不仅要对相关关系的存在性给出判断，更要对相关关系的强度给出度量和分析。

## 10.1 相关分析的基本概念

提到变量之间的关系，人们很容易想到的是变量间的确定性关系，它的特点是当一个变量值（自变量）确定后，另一个变量值（因变量）也就完全确定了。确定性关系往往可以表示成函数的形式，比如关于圆的半径和面积的关系  $S = \pi r^2$ 。

与确定性关系不同，变量之间还存在着非确定性关系，它的特点是给定了一个变量值后，另一个变量值可以在一定的范围内变化。例如关于家庭的消费支出与家庭收入的关系，同样收入的家庭，其支出可能有很大的差异，因为除了受收入高低的影响外，家庭消费支出还受其他因素的影响；另外还有关于人的身高和体重之间的关系，犯罪与否与年龄之间的关系，吸烟量和寿命之间的关系，校园环境和学生体质之间的关系等。

通常，研究者把非确定性关系称为相关关系，它必须借助于统计手段才能加以研究，故又称为统计相关。

### 10.1.1 相关分析的特点和应用

相关关系是普遍存在的，函数关系仅是相关关系的特例。

#### 1. 相关关系的类型

当事物之间存在相关关系时，不一定是因果关系，也可能仅是伴随关系；但如果事物之间存在因果关系，则它们必然是相关的。相关关系多种多样，归纳起来大致有如下 6 种类型。

- 强正相关关系，其特点是一变量  $X$  增加，导致另一变量  $Y$  明显增加，说明  $X$  是影响  $Y$  的主要因素。
- 弱正相关关系，其特点是一变量  $X$  增加，导致另一变量  $Y$  增加，但增加幅度不明显，说明  $X$  是影响  $Y$  的因素，但不是唯一因素。
- 强负相关关系，其特点是  $X$  增加，导致  $Y$  明显减少，说明  $X$  是影响  $Y$  的主要因素。

- 弱负相关关系，其特点是变量  $X$  增加，导致  $Y$  减少，但减小幅度不明显，说明  $X$  是影响  $Y$  的因素，但不是唯一因素。
- 非线性相关关系，其特点是  $X$ 、 $Y$  之间没有明显的线性关系，却存在着某种非线性关系，说明  $X$  仍是影响  $Y$  的因素。
- 不相关，其特点是  $X$ 、 $Y$  之间不存在相关关系，说明  $X$  不是影响  $Y$  的因素。

## 2. 相关分析的应用

相关分析是研究变量之间相关关系的数理统计方法，它可以从影响某个变量的诸多变量中判断哪些是显著的，哪些是不显著的。而且，在得到相关分析的结果后，还可以用其他统计分析方法对其做更进一步的分析、预测或控制，比如回归分析、因子分析等。

相关分析方法已广泛应用于生物学、心理学、教育学、经济学和医学等各个方面，它对于试验数据的处理、经验公式的建立、管理标准的测定、自然现象和经济现象的统计预报、自动控制中数学模型的确定等，都是一种方便而且有效的统计工具。

举例来说，在一般条件下，适当增加施肥量，农作物产量就会有相应的提高；劳动生产率提高，产品成本将会下降；产品价格的高低变动，会影响其销售量的大小；能源的发展速度提高，会促进整个工业发展的提高等。这些例子中那些起影响作用的因素（自变量）和被影响因素（因变量）之间都存在着一定的相关关系。

### 10.1.2 相关系数的计算

根据数据特点不同，所采用的度量变量间相关程度的统计量也会不同，相应地，相关系数也就有了不同的表现形式。下面介绍最常见的几个相关系数，其中线性相关系数为参数统计方法，而 Spearman 和 Kendall 等级相关系数为非参数统计方法。

#### 1. 线性相关系数

线性相关（linear correlation）又称简单相关（simple correlation），用来度量具有线性关系的两个变量之间相关关系的密切程度及其相关方向，适用于双变量正态分布资料。线性相关系数又称为简单相关系数、Pearson（皮尔森）相关系数或相关系数，有时也称为积差相关系数（coefficient of product-moment correlation）。

常以符号  $r$  表示样本相关系数， $\rho$  表示总体相关系数。

总体相关系数的定义公式是  $\rho_{XY} = \text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X)}\sqrt{\text{Var}(Y)}}$ 。其中  $\text{Cov}(X, Y)$  是随机变量

$X$ 、 $Y$  的协方差， $\text{Var}(X)$  和  $\text{Var}(Y)$  分别代表  $X$  和  $Y$  的方差。总体相关系数是反映两变量之间线性相关程度的一种特征值，表现为一个常数。

样本相关系数的定义公式是  $r_{XY} = \frac{\bar{\sigma}_{XY}}{\sqrt{\bar{\sigma}_{XX}}\sqrt{\bar{\sigma}_{YY}}} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}}$ 。它是根据样本

观测值计算的，抽取的样本不同，其具体的数值也会有所差异。可以证明，样本相关系数是总体相关系数的一致估计量。

判断样本相关系数  $r$  是否来自  $\rho \neq 0$  的总体，需要对它进行显著性检验，此处可以采用  $t$  检验或者  $F$  检验，此时的零假设和备择假设分别为  $H_0: \rho = 0$ ， $H_A: \rho \neq 0$ 。

$t$  检验统计量  $t = r / S_r$ ,  $df = n - 2$ ;  $S_r = \sqrt{(1 - r^2) / (n - 2)}$  称为相关系数的标准误。

$F$  检验统计量  $F = \frac{r^2}{(1 - r^2) / (n - 2)}$ ,  $df_1 = 1$ ,  $df_2 = n - 2$ 。

## 2. Spearman 等级相关系数

Spearman 相关系数相当于 Pearson 相关系数的非参数形式, 它根据数据的秩而不是数据的实际值计算, 适用于有序数据和不满足正态分布假设的等间隔数据。Spearman 相关系数的取值范围也在 -1 到 1 之间, 绝对值越大相关性越强, 取值符号也表示相关的方向。

随机变量  $X$ 、 $Y$  之间的 Spearman 相关系数记为  $r_s$ , 其计算公式为  $r_s = 1 - \frac{6 \sum d^2}{n(n^2 - 1)}$ 。其中  $d$

为分别对  $X$  和  $Y$  取秩之后每对观察值  $(x, y)$  的秩之差,  $n$  为所有观察对的个数。

随后介绍对 Spearman 相关系数  $r_s$  的假设检验, 零假设为  $r_s$  是来自  $\rho_s = 0$  的总体 (即  $X$  与  $Y$  独立)。以显著性水平  $\alpha = 0.05$  为例, 当  $n \leq 30$  或 50 时, 可以查 Spearman's 相关系数表来确定  $P$  值, 此时有: 当  $P \leq 0.05$  时, 拒绝零假设, 说明  $X$  与  $Y$  之间存在着较为显著的相关关系; 当  $P > 0.05$  时, 接受零假设。

## 3. Kendall 等级相关系数

Kendall 相关系数是对两个有序变量或两个秩变量之间相关程度的度量统计量, 因此也属于非参数统计范畴, 它在计算时考虑了结点 (秩相同的点) 的影响。

下面介绍 SPSS 中的 Kendall's Tau (nonparametric correlations algorithms) 算法。两个随机变量  $X$ 、 $Y$  共有  $t$  组观测对  $(x, y)$ , 对任意第  $(i, j)$  个观测数据, 若满足  $i < j$ , 就计算  $d_{ij} =$

$[R(X_j) - R(X_i)][R(Y_j) - R(Y_i)]$ , 令  $S = \sum_{i=1}^{N-1} \sum_{j=i+1}^N \text{sign}(d_{ij})$ , 则 Kendall's tau ( $\tau$ ) 按如下公式计算

$$\tau = \frac{S}{\sqrt{\frac{N^2 - N - \tau_x}{2}} \sqrt{\frac{N^2 - N - \tau_y}{2}}}。 \text{当此式分母为 0 时不能用, 需要按另外的公式计算。}$$

Kendall's tau 相关系数的显著性检验通过统计量  $Z = \frac{S}{\sqrt{d}}$  进行, 在零假设 ( $X$ 、 $Y$  不相关) 成立的条件下, 它近似服从正态分布。

### 10.1.3 SPSS 提供的相关分析功能

SPSS 的相关分析功能集中在 Analyze 菜单的 Correlate 子菜单中, 包括以下 3 个过程。

(1) Bivariate (两两相关分析过程)。此过程用于两个或多个变量之间的参数与非参数相关分析, 如果是对多个变量的分析, 将给出它们之间两两相关分析的结果。这是 Correlate 子菜单中最为常用的一个过程。

(2) Partial (偏相关分析过程)。如果进行相关分析的两个变量取值均受到其他变量的影响, 就可以利用偏相关分析对所谓的其他变量进行控制, 这种方法的思想 and 协方差分析比较类似。

(3) Distances (距离分析过程)。此过程可以在观测记录之间或者不同变量之间进行相似性和不相似性分析。相似性分析可用于检测观测值的接近程度; 不相似性分析常用

于考察各变量的内在联系和结构。该过程一般不单独使用，而是作为因子分析、聚类分析和多维尺度分析等的预分析过程，以帮助了解复杂数据集的内在结构，为进一步的分析做准备。

10.2 两变量相关分析

Bivariate 过程用于进行两个变量之间的相关分析，对双变量正态分布数据，可以选择 Pearson 相关系数；对其他数据，可以选择 Kendall 等级相关系数和 Spearman 等级相关系数。

10.2.1 问题描述和数据准备

本节对某高校的一些学生的身体特征数据进行相关性分析，数据格式如图 10-1 所示，所用数据文件为“肺活量数据.sav”。一般认为，人的肺活量与身高、体重之间有比较明显的相关性，随后就来对体重和肺活量进行两变量相关分析。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	high	Numeric	6	2	身高(cm)	None	None	8	Right	Scale
2	weigh	Numeric	6	2	体重(kg)	None	None	8	Right	Scale
3	vc	Numeric	4	2	肺活量(L)	None	None	8	Right	Ordinal

图 10-1 肺活量数据格式

在做相关分析之前，可以利用散点图初步观察一下两个变量之间有无相关趋势，当从图形能判断它们存在一定的相关趋势时，能使随后的相关分析更有意义。

依次单击菜单“Graphs→Chart Builder”打开图形构建器，选择做散点图（Scatter/Dot），将体重变量作为横轴，肺活量变量作为纵轴作图，输出图形如图 10-2 所示。

观察肺活量对体重的散点图，可以初步判断二者存在一定的正相关关系。故有必要进行下一步的相关分析，以明确这种相关性的存在性及其程度大小。

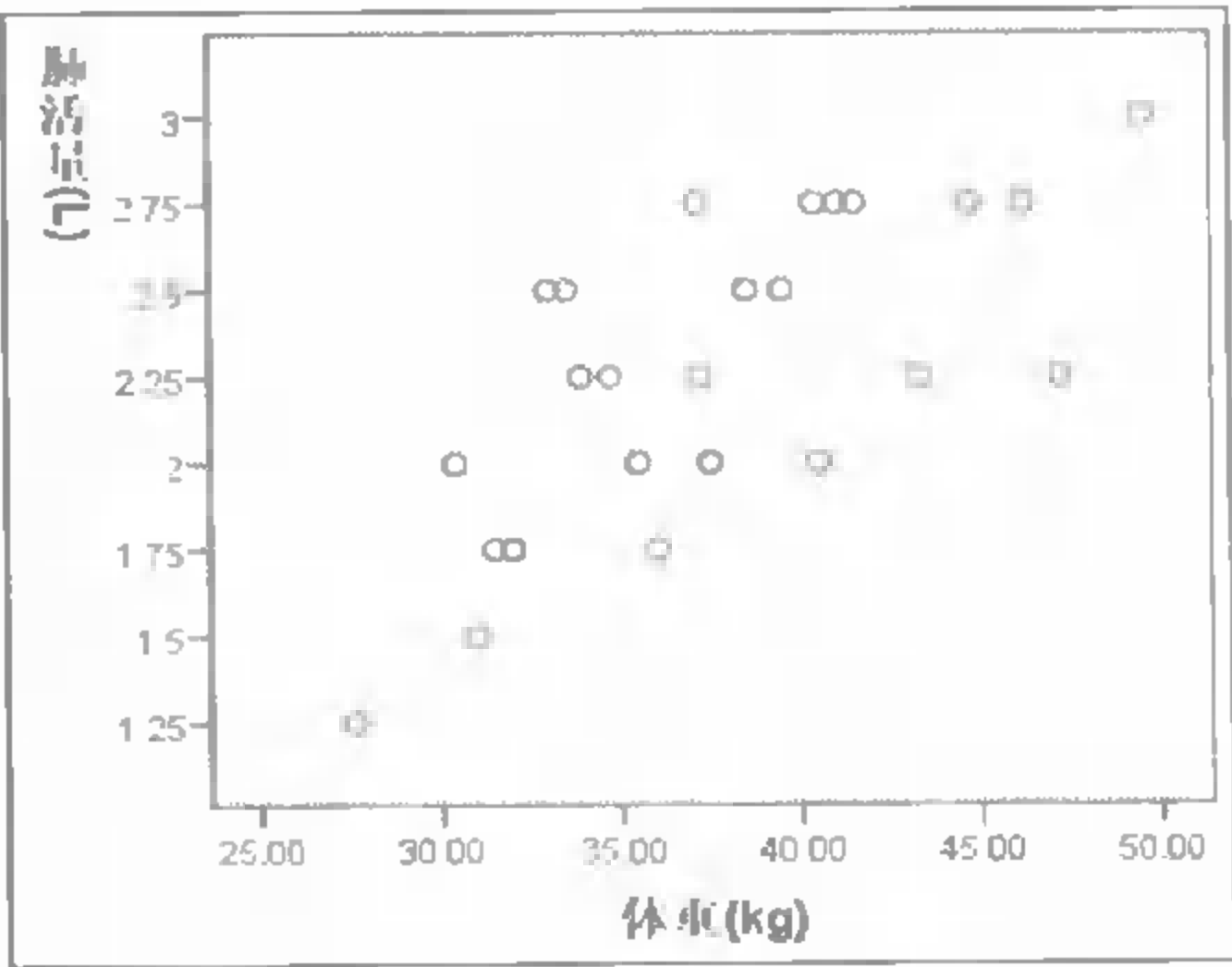


图 10-2 肺活量对体重的散点图

10.2.2 相关分析的参数设置

依次单击菜单“Analyze→Correlate→Bivariate...”执行两变量相关分析过程，其主设置面板如图 10-3 所示。

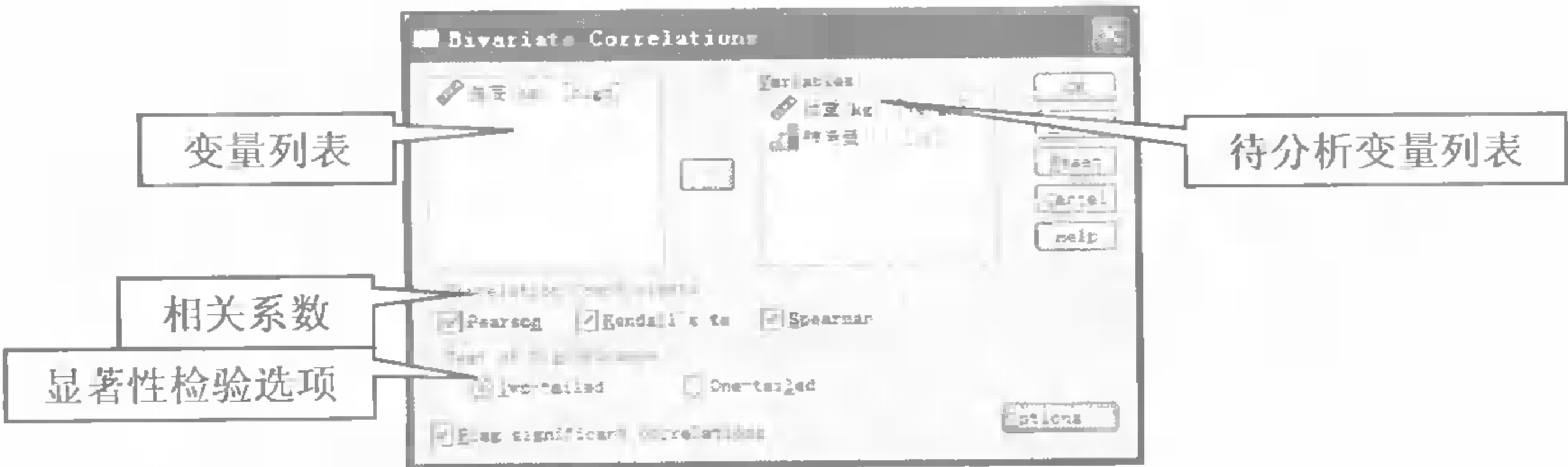



图 10-3 两变量相关分析的主设置面板



## 1. 参数设置

在变量列表中选中肺活量和体重变量，单击  按钮，将其作为分析变量选入 Variables 列表；勾选 Correlation Coefficients 栏的 3 个复选框。

(1) Variables 列表框，用于从变量列表选入要进行相关分析的变量，至少需要选入 2 个；如果选入了多于 2 个的变量，在输出中会以相关矩阵的形式给出两两相关分析的结果。

(2) Correlation Coefficients 子设置栏，在此选择要计算的相关系数类型，有 3 个可选项：Pearson 相关系数（默认选择）、Kendall's tau-b 等级相关系数和 Spearman 等级相关系数。

(3) Test of Significance 子设置栏，设置显著性检验的方式有两种选择。Two-tailed 单选框表示双边检验（默认选项），当事先不知道相关方向（正相关还是负相关）时选中此项；One-tailed 单选框表示单边检验，当事先已经知道相关方向时选中此项。无论选择哪一项，显著性检验的零假设都是总体中两个变量是不相关的。

(4) Flag significant correlations 复选框，勾选它后的输出结果中，相关系数在 0.05 的显著性水平上不为零时，右上角用 “\*” 标识其比较显著；相关系数在 0.01 的显著性水平上不为零时，右上角用 “\*\*” 标识其非常显著。

## 2. Options 选项设置

在图 10-3 中单击 Options 按钮，弹出如图 10-4 所示的设置面板，在此设置输出选项和缺失值处理方式。勾选 Means 复选框；单击 Continue 按钮返回主界面。

(1) Statistics 栏选择输出哪些统计量，有如下两个可选项。

① Means and standard deviations 复选框，输出每个变量的均值和标准差等描述统计量。

② Cross-product deviations and covariances 复选框，输出所有变量的叉积离差矩阵和协方差矩阵。叉积离差等于均值修正变量的积的总和，它就是 Pearson 相关系数的分子；协方差是关于两个变量相关性的非标准化度量，其值等于叉积离差除以  $N-1$ 。

(2) Missing Values 栏设置缺失值的处理方式，有以下两个可选项。

① Exclude cases pairwise 单选项，成对剔除含缺失值的记录，如果观测记录里待分析的两个变量中有一个或两个为缺失值，就剔除这个观测。此方法可以最大程度的利用样本信息，但是当变量多于两个时，计算任意两个变量之间的相关系数时，用到的观测个数有可能不同。

② Exclude cases listwise 单选项，直接剔除含缺失值的观测，如果某个观测的其中一个变量取缺失值，就在所有的分析过程中剔除这个观测。

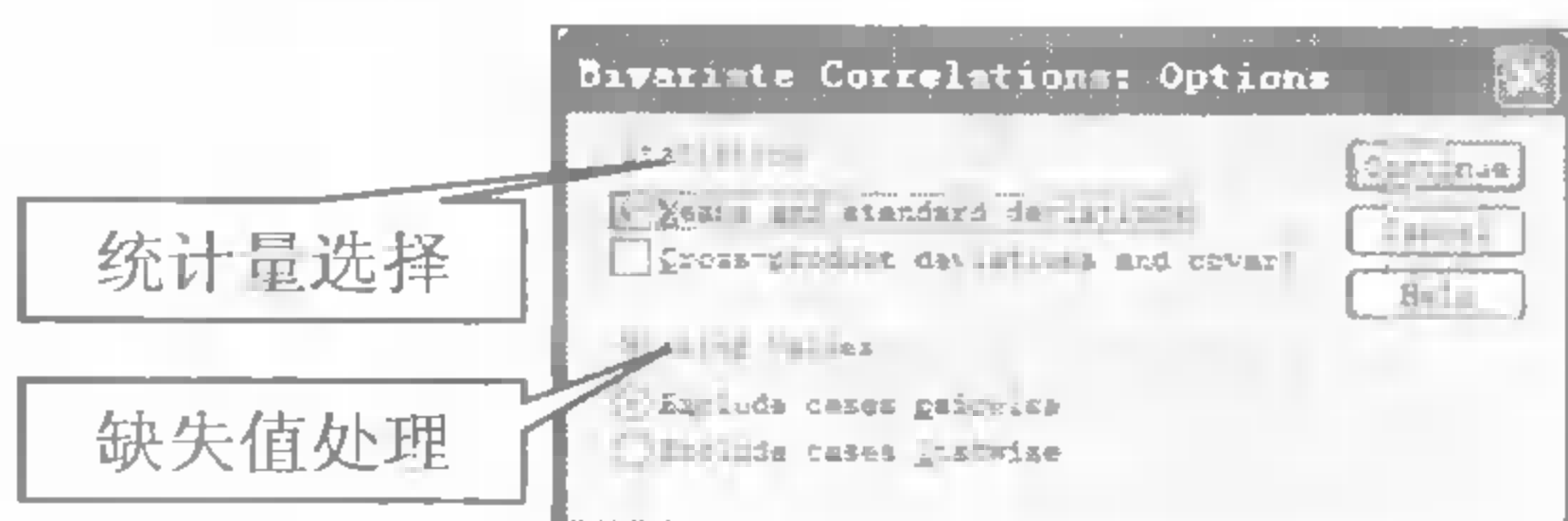


图 10-4 两变量相关分析的选项设置

### 10.2.3 案例的结果分析

单击图 10-3 中的 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 10-5 所示。

描述性统计量			
	均值	标准差	N
体重(kg)	37.1275	5.53275	29
肺活量(L)	2.2969	.44855	29

相关性			
		体重(kg)	肺活量(L)
体重(kg)	Pearson相关性	1	.736**
	显著性(双侧)		.000
	N	29	29
肺活量(L)	Pearson相关性	.736**	1
	显著性(双侧)	.000	
	N	29	29

\*\* 在 .01 水平(双侧)上显著相关。

相关系数				
		体重(kg)	肺活量(L)	
kendall 的 tau_b	相关系数	1.000	.594**	
	Sig.(双侧)		.000	
	N	29	29	
	肺活量(L)	相关系数	.594**	1.000
	Sig.(双侧)	.000		
	N	29	29	
Spearman 的 rho	体重(kg)	相关系数	1.000	.744**
	Sig.(双侧)		.000	
	N	29	29	
	肺活量(L)	相关系数	.744**	1.000
	Sig.(双侧)	.000		
	N	29	29	

\*\* 在置信度(双侧)为 0.01 时,相关性是显著的。

图 10-5 两变量相关分析的结果

(1) 描述性输出。“描述性统计量”表格给出了两个变量的基本统计信息,包括均值、标准差和频率。

(2) 相关性输出。“相关性”表格给出的是 Pearson 相关系数及其检验结果;“相关系数”表格给出的是两个非参数相关系数及其检验结果。可见,3 个相关系数在 0.01 的显著性水平(双边检验)上都非常显著,从而推断体重和肺活量之间存在着明显的正相关关系。

### 10.3 偏相关分析

有时影响一个问题的因素很多,研究者通常先假设其中的某些因素不变,再去考察其他因素对该问题的影响,从而达到简化分析的目的,偏相关分析就是源于这一思想的统计方法。

#### 10.3.1 偏相关分析的基本原理

##### 1. 偏相关系数的含义

线性相关分析计算的是两个变量间的相关系数,它分析两个变量之间线性相关的程度。但是在实际应用中,往往因为第 3 个变量的作用,使相关系数不能真正反映那两个指定变量间的线性相关程度。例如身高、体重与肺活量之间的关系,如果使用 Pearson 相关系数,可以得出肺活量与身高、体重之间分别存在着较强的线性关系;但是对体重相同的人,是否身高越高,肺活量就越大呢?不是的。因为身高与体重有线性关系,体重与肺活量又有线性关系,由此得出身高和肺活量之间存在线性关系的结论是不可信的。偏相关分析能够在研究两个指定变量之间的线性相关关系时,控制可能对其产生影响的其他变量。

在多变量的情况下,变量之间的相关关系是很复杂的,直接研究两个变量间的简单相关系数往往不能正确说明它们之间的真实关系,只有除去其他变量影响后再计算相关系数,才能真正反映它们之间的相关关系;或者说是在其他变量固定不变的情况下,计算两个指定变量之间的相关系数。这样的相关分析就是偏相关分析,经此得出的相关系数叫做偏相关系数。例如要分析身高与肺活量之间的相关性,就要控制体重在相关分析过程中的影响。

##### 2. 偏相关系数的计算

根据固定变量个数的多少,偏相关分析可分为零阶偏相关、一阶偏相关和  $(p-1)$  阶偏

相关, 其中零阶偏相关就是简单相关。

设随机变量  $X$ 、 $Y$ 、 $Z$  之间彼此存在着相关关系, 为了研究  $X$  和  $Y$  之间的关系, 就必须在假定  $Z$  不变的条件下, 计算  $X$  和  $Y$  的偏相关系数, 记为  $r_{xy \cdot z}$ 。由此可见, 偏相关系数是由简单相关系数决定的, 但是在计算偏相关系数时要考虑其他自变量对指定变量的影响, 事实上就是把其他变量当作常数处理了。

以下标 0 代表  $X$ , 下标 1 代表  $Y$ , 下标 2 代表  $Z$ , 则  $X$  与  $Y$  之间的一阶偏相关系数定义为  $r_{01 \cdot 2} = \frac{r_{01} - r_{02}r_{12}}{\sqrt{1-r_{02}^2}\sqrt{1-r_{12}^2}}$ , 其中  $r_{01 \cdot 2}$  是剔除  $Z$  的影响之后  $X$  与  $Y$  的偏相关系数,  $r_{01}, r_{02}, r_{12}$  分

别是  $X$ 、 $Y$ 、 $Z$  之间的两两简单相关系数。如果增加一个变量  $T$  (以下标 3 表示), 则  $X$  与  $Y$  的二阶偏相关系数定义为  $r_{01 \cdot 23} = \frac{r_{02} - r_{03 \cdot 2}r_{13 \cdot 2}}{\sqrt{1-r_{03 \cdot 2}^2}\sqrt{1-r_{3 \cdot 2}^2}}$ 。

一般, 考察多个变量时,  $Y$  与  $X_i (i=1, 2, \dots, p)$  之间的  $p-1$  阶偏相关系数可由如下的递推式定义:  $r_{0i \cdot 12 \dots (i-1)(i+1) \dots p} = \frac{r_{0i \cdot 12 \dots (i-1)(i+1) \dots (p-1)}r_{0ip \cdot 12 \dots (p-1)}r_{ip \cdot 12 \dots (i-1)(i+1) \dots (p-1)}}{\sqrt{1-r_{0p \cdot 12 \dots (p-1)}^2}\sqrt{1-r_{ip \cdot 12 \dots (i-1)(i+1) \dots (p-1)}^2}}$ 。

另外, 对偏相关系数的显著性检验与简单相关系数的情形相似, 用户无须记住繁琐的公式, 只需理解其基本思想。SPSS 能够直接计算出偏相关系数的大小, 以及推断其显著性的  $P$  值。

### 3. 偏相关分析的应用

SPSS 的 Partial 过程用于对变量进行偏相关分析, 它可按用户的要求对指定变量之外的其他变量进行控制, 输出消除其他变量影响之后的偏相关系数。



偏相关分析的一个主要应用是根据观测资料计算样本的偏相关系数, 由此判断哪些自变量对因变量的影响较大, 并选择其作为必须要考虑的因素; 至于那些对因变量影响较小的自变量, 则可舍去。经过这样的选择, 在进行多元回归等分析时, 就可以只考虑那些起主要作用的因素, 如此就能用较少的自变量来描述因变量的变化规律。

偏相关分析的应用也非常广泛, 涉及自然科学和社会科学的各个方面。

#### 10.3.2 偏相关分析实例

本节继续使用如图 10-1 所示的高校学生身体特征数据, 这次利用偏相关分析来研究身高、体重和肺活量这 3 个变量之间的相互关系, 重点分析在排除体重因素后身高与肺活量的相关关系。

##### 1. 参数设置

依次单击菜单 “Analyze→Correlate→Partial...” 执行偏相关分析过程, 其主设置界面如图 10-6 所示, 在变量列表中选中肺活量和身高变量, 单击从上至下第一个  按钮, 将其作为分析变量选入 Variables 列表框; 在变量列表中单击选中体重变量, 单击从上至下第二个  按钮, 将其作为控制变量选入 Controlling 列表框。单击 Options 按钮, 弹出如图 10-7 所示的设置面板, 勾选 Means 复选框和 Zero 复选框; 单击 Continue 按钮返回主界面。

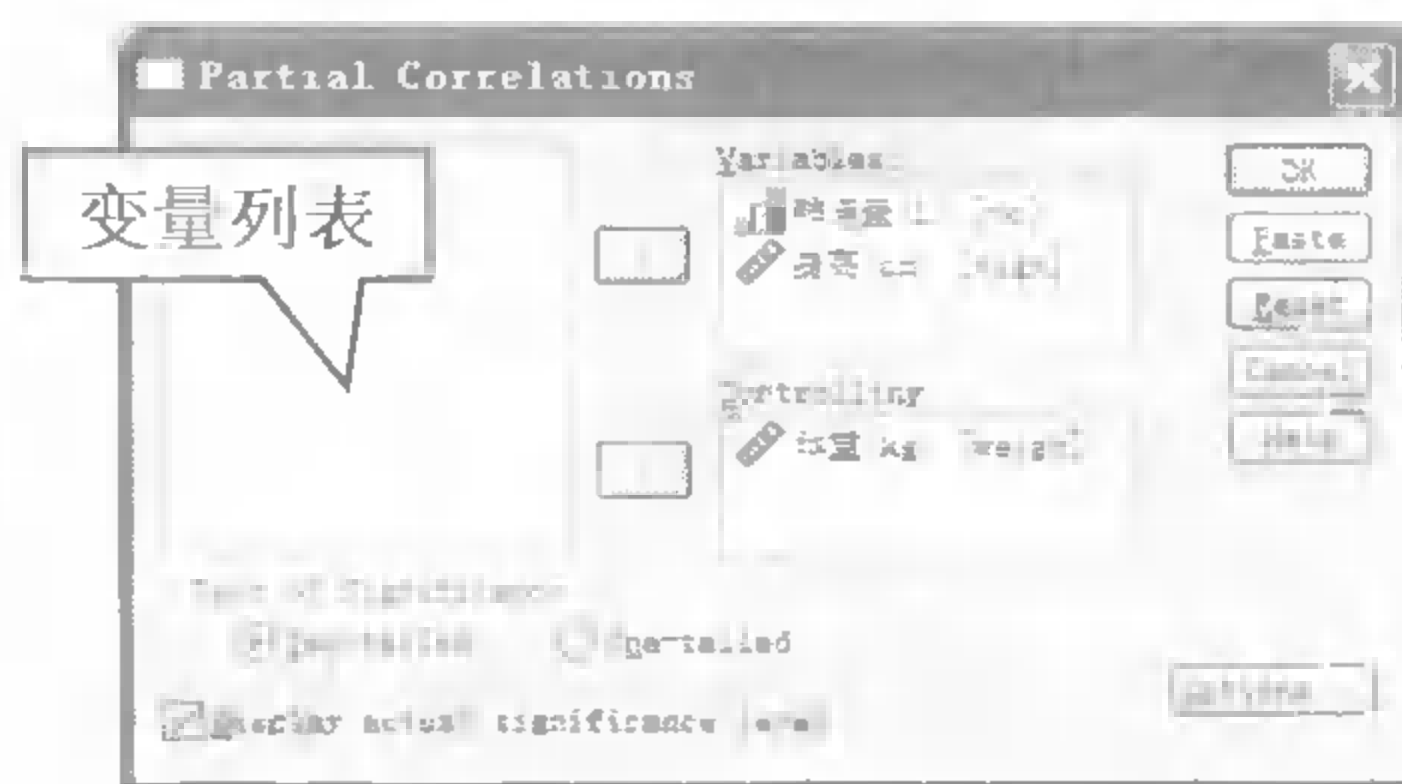


图 10-6 偏相关分析的主设置界面设置

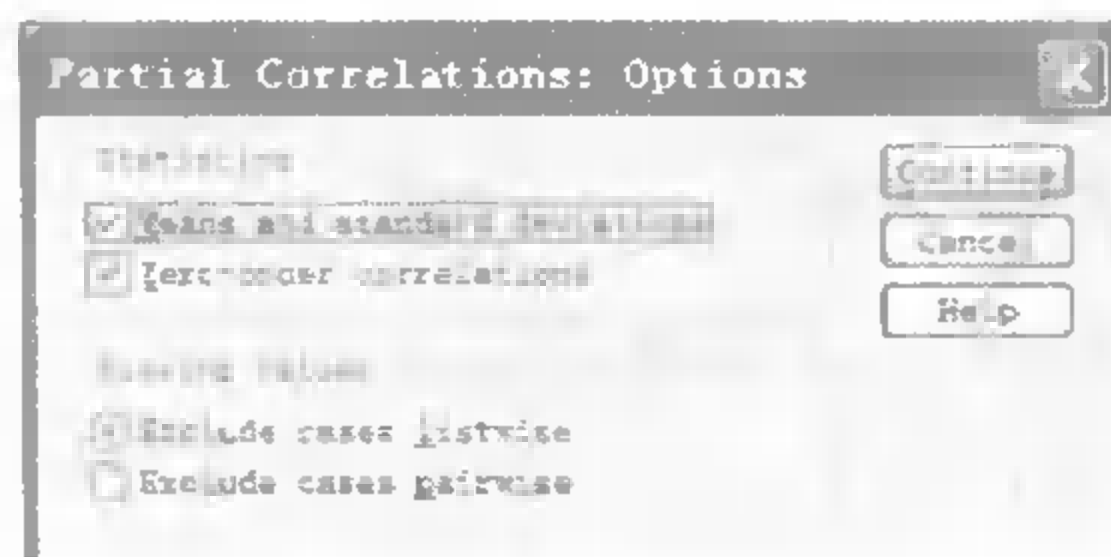


图 10-7 偏相关分析的选项设置

(1) 在图 10-6 中, Variables 列表框用于选入要进行相关分析的变量, 至少需要选入两个; Controlling 列表框用于选入控制变量。

(2) 在图 10-7 中, Zero-order correlations 复选框相当于两变量间 (包括控制变量) 的简单相关系数。

其他设置选项和设置方式与图 10-3 和图 10-4 所示的简单相关分析情形相似。

## 2. 结果分析

在图 10-6 中, 单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 10-8 所示。

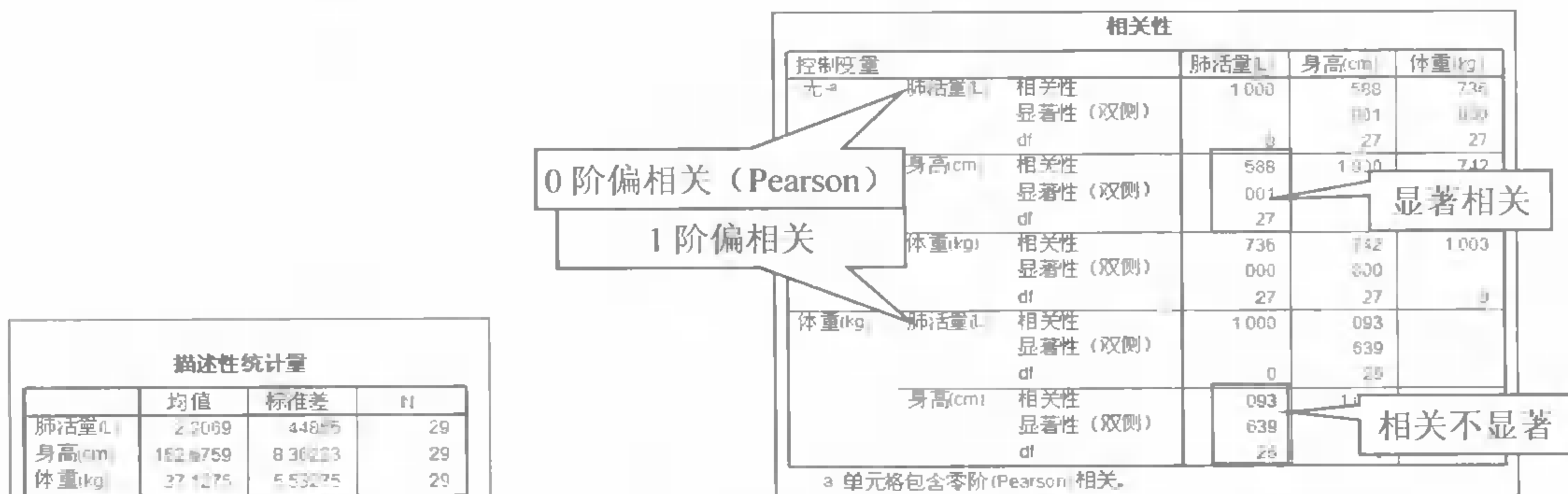


图 10-8 偏相关分析的结果输出

(1) 描述性输出“描述性统计量”表格给出关于三个变量的描述性统计量, 包括均值、标准差和频率。

(2) 相关性输出。“相关性”表格给出了所有变量的 0 阶偏相关 (Pearson 简单相关) 系数和 1 阶偏相关系数的计算结果, 以及它们各自的显著性检验 P 值。如图中标识, 在不控制体重变量时, 身高和肺活量是显著相关的; 但在控制了体重变量后, 身高和肺活量之间的相关性又变得非常不显著了。所以不能简单地判断身高与肺活量之间是否存在着相关关系, 准确一点的结论应该为在体重不变的条件下, 身高与肺活量之间不存在显著的线性相关关系。

## 10.4 距离分析

在偏相关分析中, 我们关心的是某两个变量的相关性, 因此需要控制其他被认为是“次要”变量的影响。实际上, 事情往往比这更复杂, 有时变量多到了无法一一关心的地步, 它



们都携带了一定的信息，但彼此又有所重叠，此时最直接的办法就是将所有变量按照一定的标准进行分类，即聚类分析。但聚类分析是一种比较复杂的多元统计方法，指标太多时计算起来会比较繁琐和费时，如果能事先给点提示就会使其更加简便和易用了。本节介绍的距离分析就是简化数据的一种预分析过程，通过它可以得到初步的分析线索。

### 10.4.1 距离分析的基本概念

SPSS 的 Distances 过程可以用来作距离分析，它能按照指定的统计量计算不同变量（或记录）之间的相似性和不相似性，从而为更深入的分析（比如聚类分析等）提供一定的信息。正是因为这个特点，距离分析不会给出常用的显著性  $P$  值，而只给出各变量（或者记录）之间的距离大小，由用户自行判断其相似的程度。

距离是对观测量之间或变量之间的相似或不相似程度的一种测度，它计算的是 1 对变量之间或 1 对观测量之间的广义距离。这些相似性或距离测度可以应用于其他分析过程，例如因子分析、聚类分析或多维尺度分析等，这样做有助于对复杂数据集的深入分析。

与距离分析有关的统计量分为不相似测度和相似测度两大类。其中不相似性测度包括对等间隔（定距）数据的欧氏距离、欧氏距离平方、Chebychev 距离和 Block 区组距离等；对计数数据使用的  $\chi^2$  距离或  $\phi^2$  距离；对二元（只有两种取值）数据使用欧氏距离、欧氏距离平方、尺寸差异、模式差异和方差等。相似性测度包括对连续变量使用的 Pearson 相关统计量或余弦统计量；对二元数据可使用的统计量有 20 余种。

### 10.4.2 距离分析的参数设置

依次单击菜单“Analyze→Correlate→Distances...”执行距离分析过程，其主设置面板如图 10-9 所示，在这里设置分析变量和方法参数。Compute Distances 栏有两个选项，Measure 栏也有两种测距方式，这两个子设置栏的可选项组合起来共有 4 种处理方式。

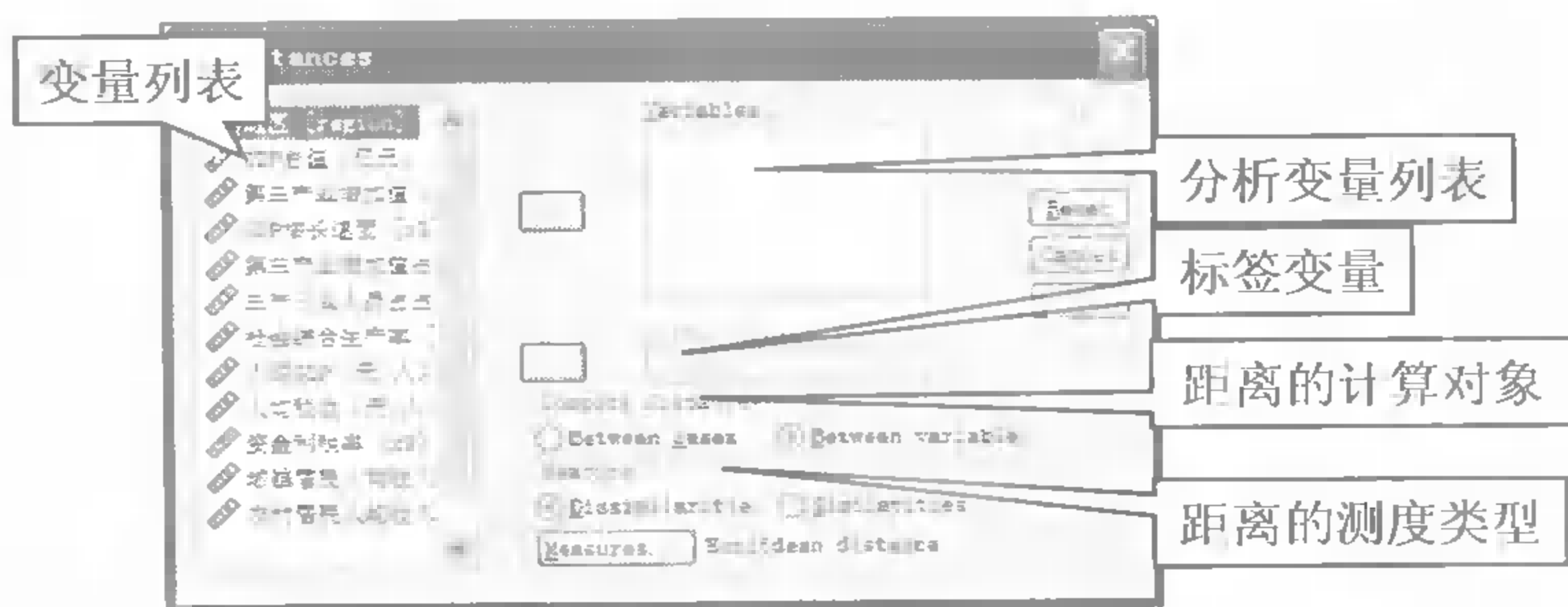


图 10-9 距离分析主设置面板

#### 1. 变量和方法设置

(1) 选择分析变量。Variables 列表框用于从变量列表选入要进行距离分析的变量，至少需要选入两个，可以为连续变量或分类变量。Label Case by 选框用于选入标识观测量的标签变量，只有在 Compute Distances 栏中选择了 Between cases 后，此选框才可用。

(2) Compute Distances 子设置栏。在此选择计算何种对象的距离，有两个可选项。Between cases 单选框表示计算观测量之间的距离；Between Variables 单选框表示计算变量之间的距离。

(3) Measure 子设置栏。在此选择距离测度的类型，有两个可选项。Dissimilarities 单选框表示计算不相似性矩阵，是默认选项，并且默认使用欧氏距离 (Euclidean distance) 测度；Similarities 单选框表示计算相似性矩阵，且默认使用 Pearson 相关系数作为相似性测度。当前设定的测度类型会自动显示在 Measures 按钮的右侧。

## 2. 不相似测度的详细设置

在图 10-9 中，单击选中 Measure 栏的 Dissimilarities 单选框，再单击 Measures 按钮，将弹出如图 10-10 所示的关于不相似测度的子设置界面。

(1) Measure 栏选择测度类型，根据数据类型又分为 Interval、Counts 和 Binary3 类。

① Interval 区间变量 (即连续变量)，其 Measure 下拉列表中可选的测度计算方法如下。

- Euclidean distance 欧氏距离，取两变量 (或观测) 取值之差的平方和的平方根。
- Squared Euclidean distance 欧氏距离的平方，两变量 (或观测) 取值之差的平方和。
- Chebychev 切贝谢夫距离，取两个项目取值之差的最大绝对值。
- Block 布洛克距离，两个项目取值之差的绝对值之和。
- Minkowski 明可夫斯基距离，两项之间的距离是每对取值之差的  $p$  次幂的绝对值之和再开  $p$  次方的根。选择此项后，还需在 Power 下拉列表指定  $p$  值，范围为 1~4。
- Customized 自定义距离，两项之间的距离是每对取值之差的  $p$  次幂的绝对值之和再开  $t$  次方根。选择此项后，还需在 Power 下拉列表指定  $p$  值，在 Root 下拉列表指定  $t$  值，其取值范围都为 1~4。

② Counts 计数变量，其 Measure 下拉列表中可选的不相似测度计算方法如下。

- Chi-square measure  $\chi^2$  距离测度，该不相似性测度基于对两组频数的相等性的卡方检验，它是 Counts 数据类型的默认方法。
- Phi-square measure  $\phi^2$  距离测度，该测度设法把样本的大小考虑进去，以减少观测频数对测度值的影响，它可用前面的卡方测度除以联合频数的平方根得到。

③ Binary 二元变量，首先需指定表征特性存在与否的取值，再指定测度计算方法。

Present 输入框指定表征特性存在的变量值，默认值为 1；Absent 输入框指定表征特性不存在的变量值，默认值为 0。其 Measure 下拉列表中可选的不相似测度的计算方法有以下 7 个。

- Euclidean distance 二元欧氏距离，最小值为 0，最大值无上限。它根据四格表计算  $\sqrt{b+c}$ ，此处的  $b$  和  $c$  是指在一项中出现而在另一项中不出现的对角线上的元素。
- Squared Euclidean distance 二元欧氏距离的平方，最小值为 0，最大值无上限，计算不一致的观测数。
- Size difference，这是一个反映不对称性的指标，取值范围为 0~1。
- Pattern difference，取值范围为 0~1，根据四格表计算  $bc/n^2$ ，其中  $b$  和  $c$  是在一项中出现而在另一项中不出现的对角线上的元素， $n$  是观测量总数。

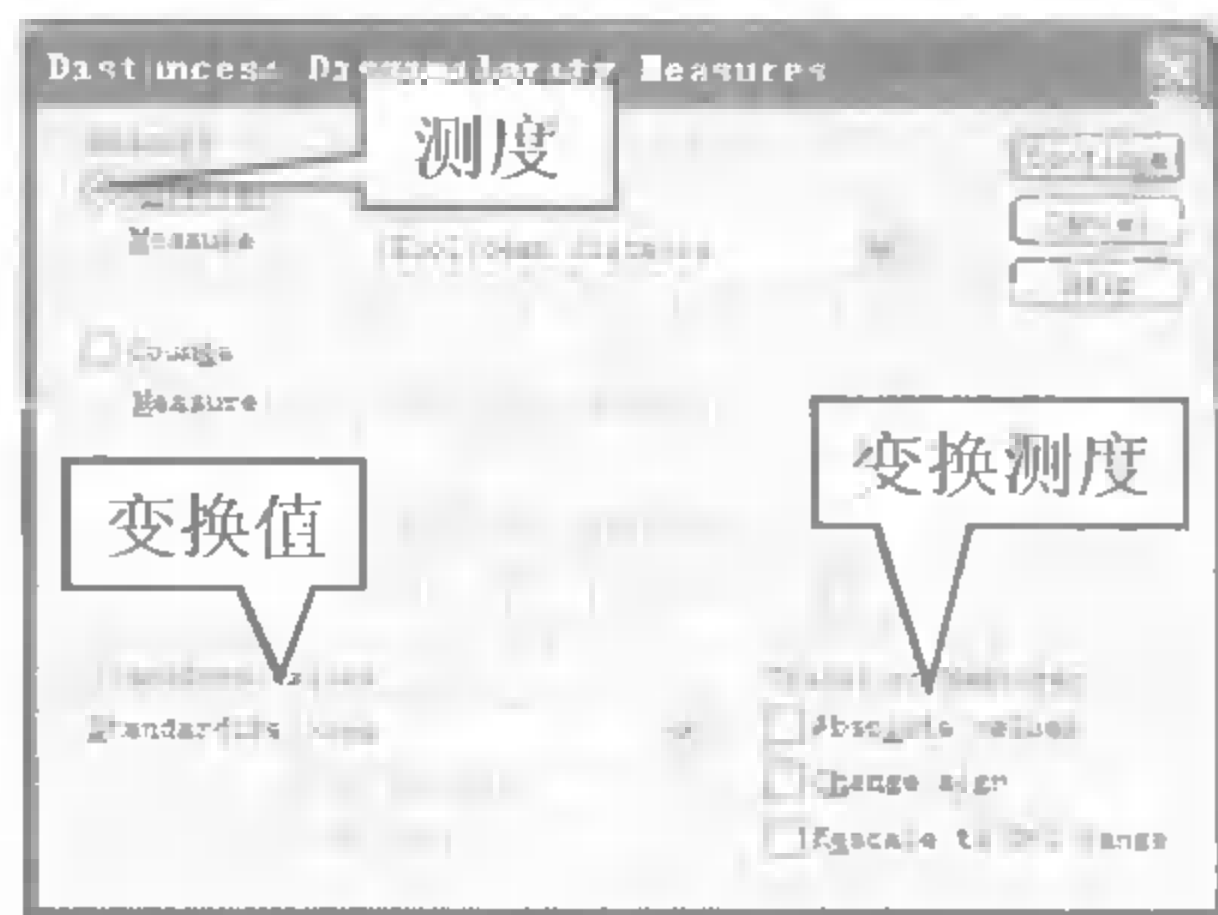


图 10-10 不相似性测度的子设置界面

- Variance 变异指标, 取值范围为  $0 \sim 1$ , 根据四格表计数  $(b+c)/4n$ , 其中  $b$  和  $c$  是在一项中出现而在另一项中不出现的对角线上的元素,  $n$  是观测量总数。
- hape 测度, 范围是  $0 \sim 1$ , 它对配合不当形成的不对称进行惩罚。
- Lance and Williams 不相似性测度 (也称为 Bray-Curtis 系数), 取值范围为  $0 \sim 1$ , 它根据四格表计算  $(b+c)/(2a+b+c)$ , 其中  $a$  表示在两项中都出现的观测对应的单元格,  $b$  和  $c$  是在一项中出现而在另一项中不出现的对角线上的元素。

(2) Transform Values 子设置栏。在此设置计算距离之前对观测量或变量进行标准化的方法, 但是对二元变量不能进行标准化。其 Standardized 下拉列表中可选的标准化方法有如下 7 个。

- None 不进行标准化, 是默认选项。
- Z-Score 标准化 Z 分数, 标准化后的均值为 0, 标准差为 1。
- Range 0 to 1, 标准化后的取值范围为  $0 \sim 1$ 。对被标准化的项目的每一个取值, 减去其最小值, 然后除以范围 (最大值与最小值的差) 得到。
- Range -1 to +1, 标准化后的取值范围为  $-1 \sim +1$ , 它由原始取值除以其范围 (最大值与最小值的差) 得到, 如果范围为 0, 所有的取值保持不变。
- Maximum magnitude of 1, 标准化后的最大值为 1, 它由原始取值除以原始的最大值得到, 如果原始最大值为 0, 则用最小值的绝对值加 1 作除数。
- Mean of 1, 标准化后的均值为 1, 它由原始取值除以均值得到, 如果均值为 0, 直接将所有的原始值加 1, 使得均值为 1。
- Standard deviation of 1, 标准化后的标准差为 1, 它由原始取值除以标准差得到, 如果标准差为 0, 原始值保持不变。

以上除 None 选择项外, 都需要指定标准化的对象, 下拉列表的下方给出了如下两个可选项。By variable 表示对变量进行标准化; By cases 表示对观测量进行标准化。

(3) Transform Measures 子设置栏。在此设置对距离测度的计算结果进行转换的方法, 有如下 3 个可选项。

- Absolute Value 复选框, 表示对距离取绝对值, 有的符号可以表明相关性的方向, 当仅对相关性的方向感兴趣时使用这种转换。
- Change sign 复选框, 表示改变距离的符号, 如此可把相似性测度转换成不相似性测度, 反之亦然。
- Rescale to 0-1 range 复选框, 表示转换后的取值范围为  $0 \sim 1$ 。对已经在 Transform Values 栏按相似方法进行标准化后的测度一般不再使用此方法。

### 3. 相似测度的详细设置

在图 10-9 中, 单击选中 Measure 栏的 Similarities 单选框, 再单击 Measures 按钮, 将弹出如图 10-11 所示的关于相似测度的子设置界面。

(1) Measure 栏选择测度类型, 根据数据类型又分为 Interval 和 Binary 两类。

① Interval 连续变量, 其后的下拉列表

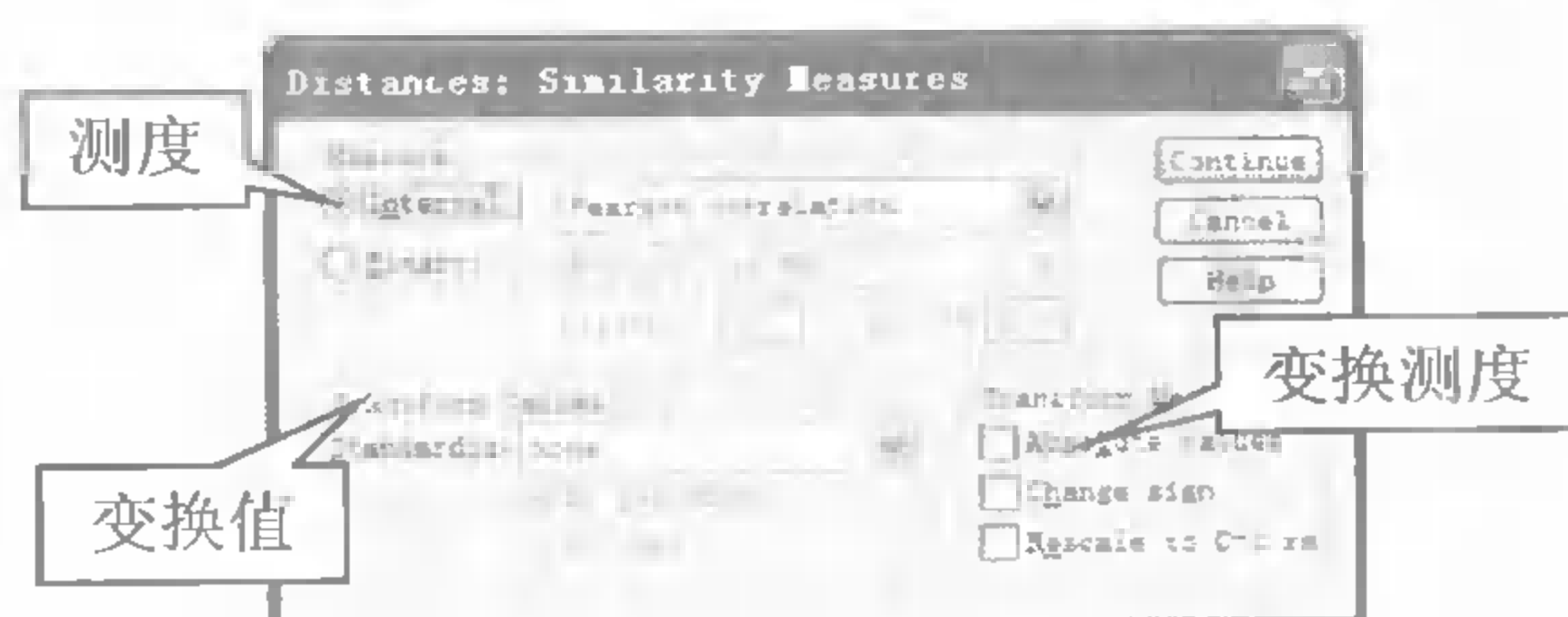


图 10-11 相似性测度的子设置界面

中可选的测度计算方法有以下两个。

- ① Pearson correlation 皮尔逊相关系数，范围为-1~+1，取 0 表示无线性相关，是默认选项。
- ② Cosine 余弦相似度，用两个向量之间的余弦度量其相似性，范围为-1~+1，取 0 表示两个向量正交（相互垂直），即不相关。

② Binary 二元变量，Present、Absent 输入框的设置方法与图 10-10 相同。

SPSS 为每对要计算的项目构造一个  $2 \times 2$  的列联表，其可选的测度计算方法有 20 种之多，可分为如下 4 类：匹配系数、条件概率、可预测性测度和其他测度。

关于匹配系数，有以下 9 种测度方法。

- ① Russell and Rao 二项内积法，它为匹配对与不匹配对赋予相等的权重，是默认选项。
- ② Simple Matching 简单匹配相似性测度，是匹配数与总数的比值，它为匹配与不匹配赋予相等的权重。
- ③ Jaccard 相似性比例指数，不考虑联合缺失项，它为匹配与不匹配赋予相等的权重。
- ④ Dice 相似性系数，不考虑联合缺失项，它为匹配对赋予双倍权重。
- ⑤ Regers and Tanimoto，所有匹配的对均包括在分母中，不匹配的对（包括不成对的）包括在分子中，它为不匹配对赋予双倍权重。
- ⑥ Sokal and Sneath 1，为匹配对赋予双倍权重。
- ⑦ Sokal and Sneath 2，不考虑联合缺失项，它为不匹配对赋予双倍权重。
- ⑧ Sokal and Sneath 3，表示匹配数与不匹配数的比值，下限为 0，无上限。当取值无定义或大于 9 999.999 时，设为 9 999.999。
- ⑨ Kulczynski 1，表示联合出现项与所有不匹配数的比值，下限为 0，无上限。

关于条件概率，有以下 3 种测度方法。

- ① Kulczynski 2，基于某个特性在一项中发生的条件下，该特性在另一项中出现的条件概率。
- ② Sokal and Sneath 4，基于某一项中的特性与其他项中的值相匹配的条件概率。
- ③ Haman，匹配数（即在两项中均出现或均不出现的特性状态）减去不匹配数之差，再除以项目总数的值，范围为-1~+1。

关于条件概率，有以下 4 种测度方法。

- ① Lamda（即 Goodman and Kruskal's Lamda 相似性测度），表示分别在两个方向上进行预测时，用一项预测另一项的误差降低的比例，范围为 0~+1。
- ② Anderberg's D，与 Lamda 方法类似，表示分别在两个方向上进行预测时，用一项预测另一项的误差降低的值，范围为 0~+1。
- ③ Yule's Y（又称为 coefficient of colligation.），它是  $2 \times 2$  表格的交叉率的函数，与表格的边际总数无关，范围为-1~+1。
- ④ Yule's Q，它是 Goodman and Kruskal's gamma 的特殊情况，范围为-1~+1。

其他测度方法有以下 4 种。

- ① Ochiai 方法，它是 Cosin 余弦测度的二元形式，范围为 0~+1。
- ② Sokal and Sneath 5，表示正、负匹配的条件概率的几何平均值的平方，它不依赖于项目编码，取值范围为 0~+1。



- Phi 4 point correlation 方法，它是 Pearson 相关系数的二元形式，取值范围为 $-1 \sim +1$ 。
- Disersion 方法，取值范围为 $-1 \sim +1$ 。

(2) Transform Values 和 Transform Measures 这两个子设置栏的设置方法，与图 10-10 所示的不相似度的转换设置相同。

### 10.4.3 距离分析实例


#### 1. 问题描述和数据准备

用于衡量经济发展水平的指标有很多，于是在作深入分析之前，有必要先了解一下这些指标之间的相似性，本节就用距离分析对这个问题加以研究。包含所用经济指标的数据文件为“地区经济发展水平指标.sav”，数据格式如图 10-12 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	region	String	8	0	地区	None	None	8	Left	Nominal
2	x1	Numeric	5	2	GDP总值(亿元)	None	None	7	Right	Scale
3	x2	Numeric	5	2	第三产业增加值(亿元)	None	None	7	Right	Scale
4	x3	Numeric	5	2	GDP增长速度	None	None	5	Right	Scale
5	x4	Numeric	5	2	第三产业增加值占GDP比重	None	None	5	Right	Scale
6	x5	Numeric	5	2	第三产从业人员占社会劳	None	None	5	Right	Scale
7	x6	Numeric	5	2	社会综合生产率	None	None	5	Right	Scale
8	x7	Numeric	5	2	人均GDP(元/人)	None	None	5	Right	Scale
9	x8	Numeric	5	2	人均税收(元/人)	None	None	7	Right	Scale
10	x9	Numeric	5	2	资金利税率	None	None	5	Right	Scale
11	x10	Numeric	5	2	城镇居民人均收入(元/人)	None	None	8	Right	Scale
12	x11	Numeric	5	2	农村居民人均收入(元/人)	None	None	8	Right	Scale

图 10-12 地区经济发展水平数据

#### 2. 参数设置

依次单击菜单“Analyze→Correlate→Distances...”打开距离分析的主设置界面，如图 10-13 所示，在变量列表中选中除地区以外的所有变量，单击从上至下第一个  按钮，将其作为分析变量选入 Variables 列表；分别单击选中 Between Variables 单选框和 Similarities 单选框。

单击 Measures 按钮，弹出如图 10-14 所示的子设置界面，单击 Transform Values 栏的下拉列表，选中 Z-Scores 选项；单击 Continue 按钮返回主界面。

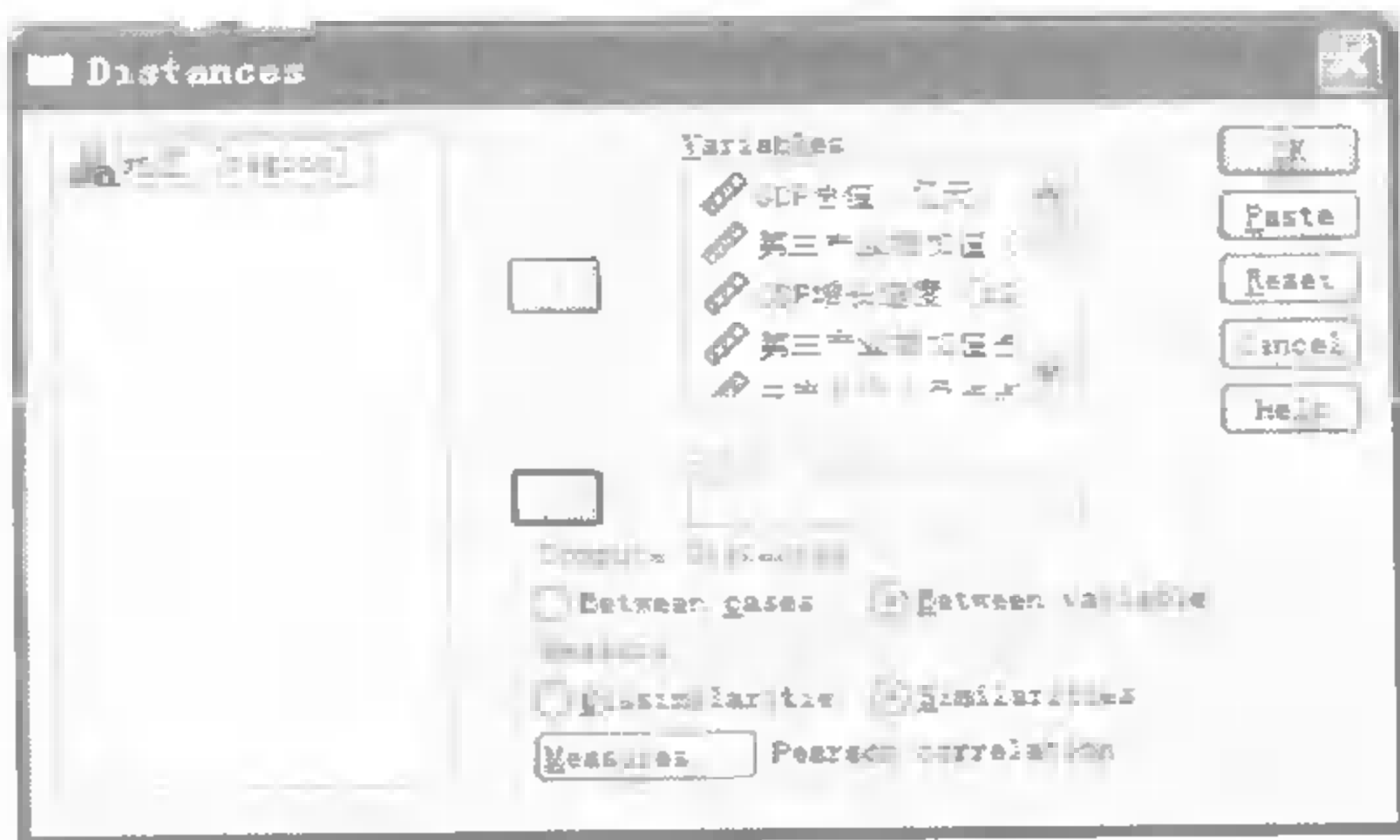


图 10-13 实例的参数设置 1

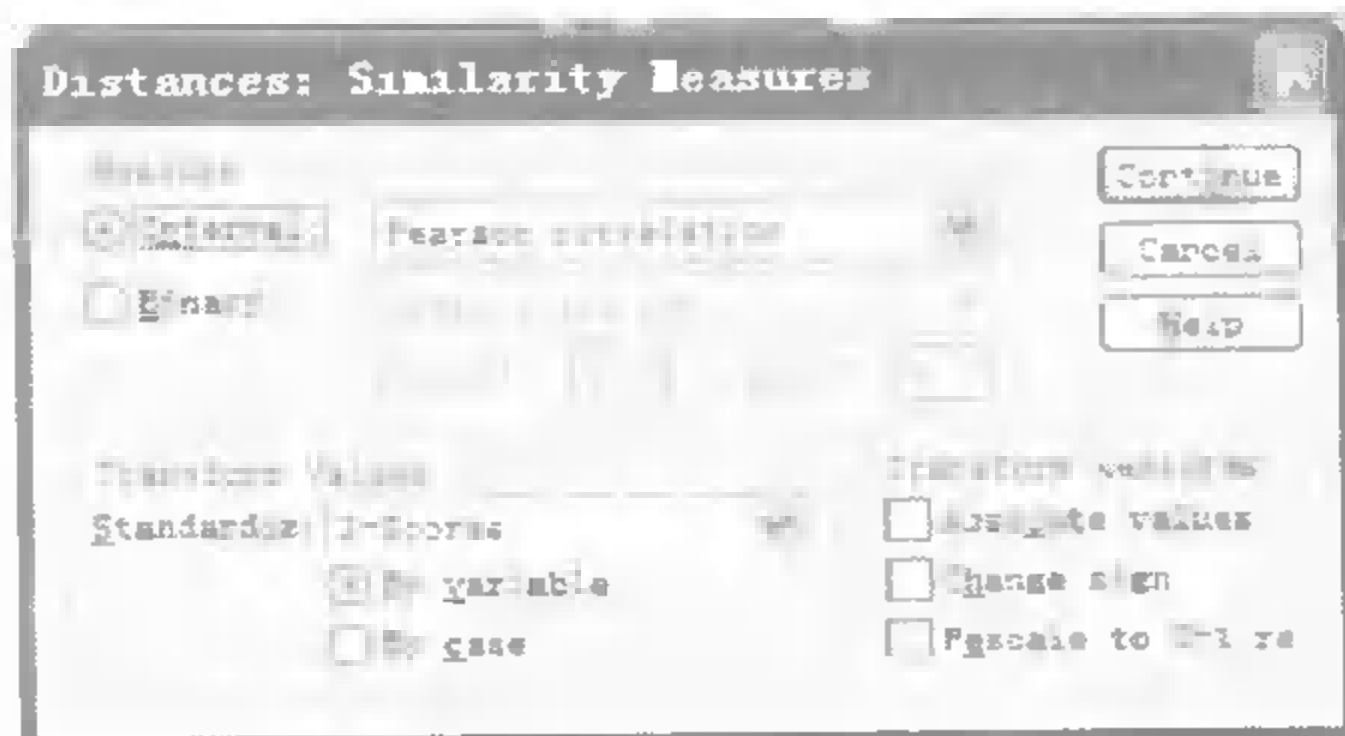


图 10-14 实例的参数设置 2

#### 3. 结果分析

单击图 10-13 中的 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 10-15 所示。



图 10-15 距离分析输出结果

(1) 案例处理摘要。“案例处理摘要”表格给出了数据使用的基本情况，主要是对有无缺失值的统计信息，可见本例的 30 个案例没有缺失，全部用于分析。

(2) 近似矩阵。“近似矩阵”表格给出的是各变量之间的相似矩阵，图中以蓝色线框标注了相关系数较大的几对变量，包括：第三产业增加值和 GDP 总值之间，第三产业从业人员比重分别和社会综合生产率、人均 GDP 之间，社会综合生产率分别和人均 GDP、人均税收之间，人均 GDP 和人均税收之间，农村居民人均收入分别和社会综合生产率、人均 GDP、人均税收这 3 个变量之间。它们在进一步的分析中应重点关注，或者直接对其进行适当的预处理（例如变量约减）。

另外，本例也可以考虑对观测行进行距离的不相似性分析，输出结果为所有地区的不相似矩阵，基于此可以对地区按照经济发展水平作简单的分类。

# 第 11 章 因子分析

许多实际问题不仅涉及的变量众多，而且各变量之间还可能存在着错综复杂的相关关系，这时最好能从中提取少数的综合变量，使其能够包含原变量提供的大部分信息，还要求这些综合变量尽可能地彼此不相关。因子分析就是为解决这一问题而提出的统计分析方法。

因子分析方法能把多个观测变量转换为少数几个不相关的综合指标，这些综合指标往往是不能直接观测到的，但有时却更能反映事物的特点和本质。因此，因子分析在医学、生物学、经济学等诸多领域都得到了广泛地应用。

## 11.1 因子分析的原理简介

### 11.1.1 因子分析的基本思想

“因子分析”于 1931 年由 Thurstone 首次提出，其概念起源于 20 世纪初 Karl Pearson 和 Charles Spearman 等人关于智力测验的统计分析。近年来，随着计算机的高速发展，人们将因子分析方法成功地应用于各个领域，使得因子分析的理论和方法更加丰富。

因子分析的基本目的是用少数几个因子去描述多个变量之间的关系，被描述的变量一般都是能实际观测的随机变量，而那些因子是不可观测的潜在变量。在经济管理科学、心理学、教育学等领域中，例如“态度”、“能力”、“智力”等许多特征实际上是不能直接通过观测来取值的，只能把它们看成是潜在变量或者公共因子；对于那些能够反映这些因子水平的变量（例如“受教育水平”、“考试成绩”、“工作业绩”、“平均收入”等）都是可以观测的。因子分析就是利用这些公共因子（基本特征）来解释可观测变量的一种工具。

因子分析的基本思想是把联系比较紧密的变量归为同一个类别，而不同类别的变量之间的相关性则较低。在同一个类别内的变量，可以想象是受到了某个共同因素的影响才彼此高度相关的，这个共同因素也称之为公共因子，它是潜在的并且不可观测的。因子分析反映了一种降维的思想，通过降维将相关性高的变量聚在一起，不仅便于提取容易解释的特征，而且降低了需要分析的变量数目和问题分析的复杂性。

因子分析的基本原理是以相关性为基础，从协方差矩阵或相关矩阵入手把大部分变异归结为少数几个公共因子所为，把剩余的变异称为特殊因子。至此，每一类的变量实际上就代表了一个公共因子（基本特征），因子分析就是用来寻找和确定这些基本特征的模型。

当对问题的内在体系还不了解时，可利用因子分析把观测变量归并为少数几个公因子，令每个公因子代表空间的一个维度，如果再经过正交或斜交旋转，各个维度之间还可以认为

主成分分析和因子分析都属于多元统计分析中处理降维的方法，因子分析是主成分分析的推广，二者既有联系又有区别。

主成分分析把具有一定相关性的初始变量重新组合成一组不相关的指标，通常所作的处理就是数学上的线性变换。因子分析则是把错综复杂的诸多变量综合为少数几个公因子，并在初始变量和公因子之间建立某种联系，它需要做的工作主要有根据变异的累计贡献率提取一定个数的公因子；对载荷矩阵实施因子旋转；计算因子得分用于进一步的分析。

主成分分析不能作为一个完整的模型加以描述，它只是通常的变量变换，而因子分析需要构造因子模型；在主成分分析中，主成分的个数和变量个数要求相同（因为它只是做了一次线性变换），但因子分析可以构造尽可能少的公因子，从而产生一个简单的模型。

设总体为  $\bar{x} = (X_1, X_2, \dots, X_p)'$ , 其均值向量  $E(\bar{x}) = \bar{\mu}$  和协方差矩阵  $V = (\sigma_{ij})_{p \times p}$  都存在。

一般情况下，公共因子不可能包含总体的所有信息，每个观测变量除了可以由公共因子解释的部分外，还会有一些它们解释不了的部分，称之为相应变量的特殊因子。

[illegible]

其中  $m \leq p$ ,  $F_1, F_2, \dots, F_m$  为初始变量的公共因子,  $\varepsilon_i$  为变量  $X_i$  的特殊因子。如果是正交的因子模型, 还进一步要求公共因子是互不相关的, 特殊因子和公共因子也不相关。

记  $A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \cdots & \cdots & \cdots & \cdots \\ a_{p1} & a_{p2} & \cdots & a_{pm} \end{pmatrix}$ ,  $\hat{F} = (F_1, F_2, \dots, F_m)'$ ,  $\hat{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_p)'$ , 则因子模型的矩阵形式记为

$\bar{x} - \bar{\mu} = A\bar{F} + \bar{\varepsilon}$ 。矩阵  $A$  称为因子载荷矩阵 (component matrix)，系数  $a_{ij}$  称为变量  $X_i$  在因子  $F_j$  上的载荷 (loading)。如果总体是标准化的，有  $a_{ij} = \rho(X_i, F_j)$ ，即变量  $X_i$  在因子  $F_j$  上的载荷  $a_{ij}$  就是  $X_i$  与  $F_j$  的相关系数。

主成分法是估计因子载荷矩阵的一种方法，由于它的估计结果和初始变量的主成分仅相差一个常数倍，故而称为主成分法，它是 SPSS 使用的默认方法。除此之外，常用的方法还有最大似然法、 $\alpha$  因子分析法、加权最小二乘法、映像因子分析法和最小残差法等。



### 3. 因子旋转

至此建立的因子模型还只是一个初始模型, 所得的因子不一定能反映问题的实质特征, 它们所代表的实际意义也不一定容易解释, 因子旋转就是解决这个问题的一种改进方法。

因子旋转的依据是因子模型的不唯一性。设  $T$  是一个正交矩阵, 由于  $TT' = I$ , 所以因子模型  $\bar{x} - \bar{\mu} = A\bar{F} + \bar{\varepsilon}$  与  $\bar{x} - \bar{\mu} = (AT)(T'\bar{F}) + \bar{\varepsilon}$  等价, 而后的载荷矩阵为  $B = AT$ , 公共因子为  $\bar{G} = T'\bar{F}$ 。于是, 如果模型  $\bar{x} - \bar{\mu} = A\bar{F} + \bar{\varepsilon}$  不易于解释, 那么作一个正交变换  $T$ , 把模型变为  $\bar{x} - \bar{\mu} = (AT')(TF) + \bar{\varepsilon} = B\bar{G} + \bar{\varepsilon}$ , 然后再在新模型中寻找因子的合理解释。

因子旋转的常用方法为方差最大化正交旋转 (Varimax)。

### 4. 因子得分函数的估计

在所建立的因子模型中, 已将总体中的原有变量分解为公共因子与特殊因子的线性组合  $X_i = a_{i1}F_1 + a_{i2}F_2 + \cdots + a_{im}F_m + \varepsilon_i, i = 1, 2, \dots, p$ ; 同样地, 可以把每个公共因子表示成原有变量的线性组合  $F_j = b_{j1}X_1 + b_{j2}X_2 + \cdots + b_{jp}X_p, j = 1, 2, \dots, m$ , 称之为因子得分函数, 用它可以计算每个观测记录在各公共因子上的得分, 从而解决公共因子不可测量的问题。获得因子得分函数的关键是求解估计参数  $\hat{b}_j = (\hat{b}_{j1}, \hat{b}_{j2}, \dots, \hat{b}_{jp})'$ , 常用的估计方法有 Thompson 方法等。

### 5. 结果分析和应用

提取出反应原始观测变量特征的公共因子, 并对其实施适当的因子旋转后, 就需要对公因子加以解释, 赋予其实际意义。对因子的解释是否恰当, 不仅与数据本身的性质有关, 还与研究者对专业知识的掌握及因子分析技巧的掌握程度有关。注意: 因子分析是以相关性为基础的, 所以对有些数据是不适用的, 建议用户在做因子分析前对样本数据做一些必要的检验。

因子得分也是重要的输出, 它实际上给出的是各个对象在公共因子上的投影值 (或坐标), 于是以公共因子为坐标轴作图, 就可以按各对象的因子得分标出其在公因子空间的相对位置, 利用此图形就能得到关于原始数据的结构方面的信息。另外, 因子得分还可以看作是对原始数据的降维和约简, 它可以进一步用于其他统计分析过程, 如聚类分析、判别分析等。

## 11.2 SPSS 因子分析的应用实例

本节利用因子分析方法对代表经济发展的多个指标进行研究, 并对不同地区的经济发展特点加以适当描述, 期望以此发现能够体现经济发展水平的潜在因素。

### 11.2.1 数据描述

SPSS 的因子分析过程适用于数值型或比率型的变量, 而不适用于分类变量。本节所用数据文件为 “地区经济发展水平.sav”, 数据格式如图 11-1 所示。

数据来源为某年份我国 30 个省市地区的 11 个经济发展指标, 注意到各个变量的度量尺度并不统一, 有元、亿元和比率等, 所以在分析之前要考虑对变量进行适当的标准化处理, 或者使用相关矩阵进行公因子的计算。我们希望通过对这 11 个变量的因子分析, 发掘出隐藏在它们之后的某些不易被直接观察到但却能够恰当地衡量和解释经济发展水平的公共因子。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	region	String	8		地区	None	None	8	Left	Nominal
2	x1	Numeric	5	2	GDP总值(亿元)	None	None	7	Right	Scale
3	x2	Numeric	5	2	第三产业增加值(亿元)	None	None	7	Right	Scale
4	x3	Numeric	5	2	GDP增长速度	None	None	5	Right	Scale
5	x4	Numeric	5	2	第三产业增加值占GDP比重	None	None	5	Right	Scale
6	x5	Numeric	5	2	三产从业人员占社会劳力	None	None	5	Right	Scale
7	x6	Numeric	5	2	社会综合生产率	None	None	8	Right	Scale
8	x7	Numeric	5	2	人均GDP(元/人)	None	None	5	Right	Scale
9	x8	Numeric	5	2	人均税收(元/人)	None	None	7	Right	Scale
10	x9	Numeric	5	2	资金利税率	None	None	5	Right	Scale
11	x10	Numeric	5	2	城镇居民人均收入(元/人)	None	None	8	Right	Scale
12	x11	Numeric	5	2	农村居民人均收入(元/人)	None	None	8	Right	Scale

图 11-1 某年份的经济发展指标数据

## 11.2.2 SPSS 因子分析过程的设置

依次单击菜单“Analyze→Data Reduction→Factor...”，执行因子分析功能，主设置界面如图 11-2 所示，在此选择参与分析的各种变量。

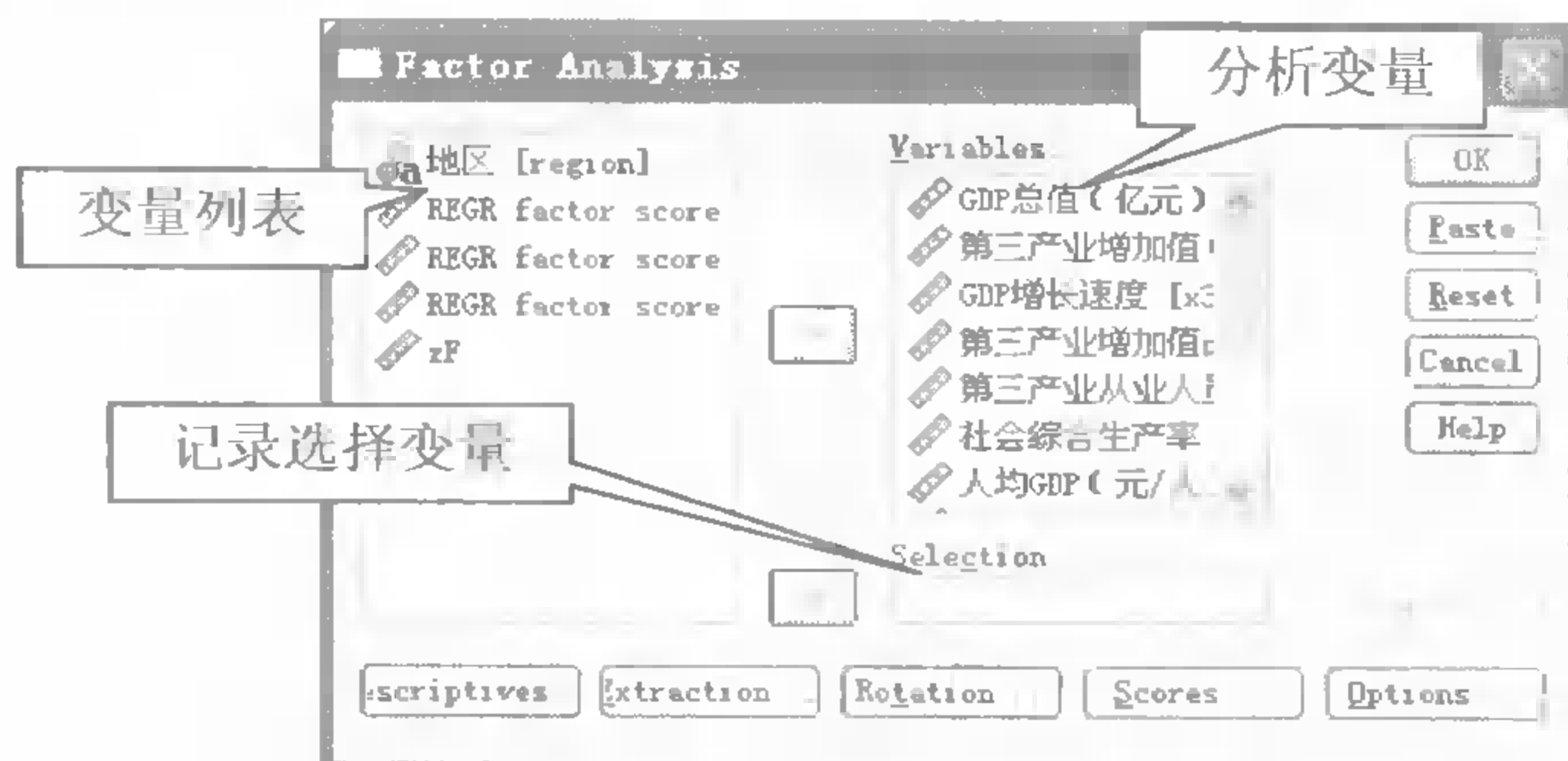


图 11-2 因子分析的主设置面板

### 1. 变量设置

在变量列表中选中从 GDP 总值到农村居民人均收入的 11 个变量，单击从上至下第一个 ☐ 按钮，将其作为分析变量选入 Variables 列表框。

(1) Variables 列表框，用于从变量列表选入待分析的原始变量。

(2) Selection 选框，用于从变量列表选入过滤样本子集的变量。

如果需要选择只满足某个条件的样本进行分析，在此选入指定变量后，单击右侧的 Value 按钮，弹出如图 11-3 所示的对话框。在 Value for selection 输入框指定此变量的某个取值，则只有此变量取这个值的样本才会进入分析过程，单击 Continue 按钮返回主界面。

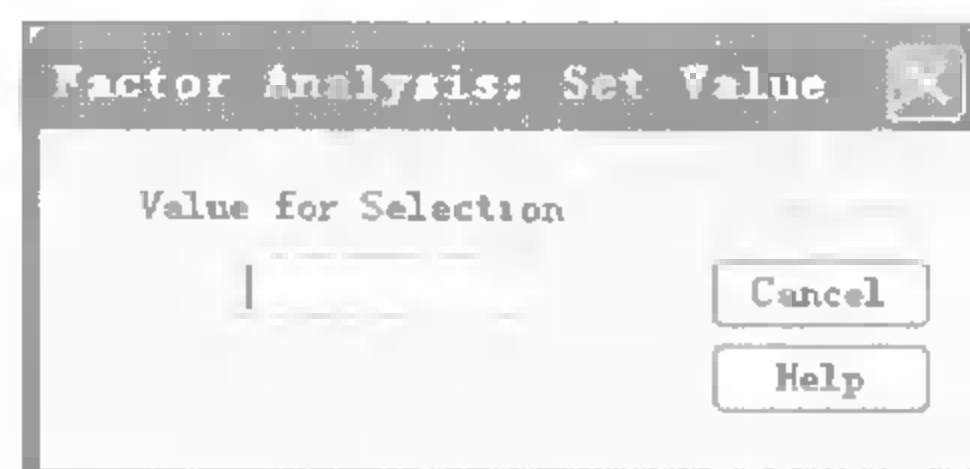


图 11-3 对观测记录的选择设置

### 2. Options 设置

在图 11-2 中单击 Options 按钮，弹出如图 11-4 所示的对话框，勾选 Sorted by size 复选框，单击 Continue 按钮返回主界面。

(1) Missing Values 栏，设置对缺失值的处理方式，有如下 3 个选项。

☐ Excludes cases listwise，当选入了多个变量进行分析时，只要其中的某个变量取缺失

值，就在所有分析过程中将对应的记录删除。

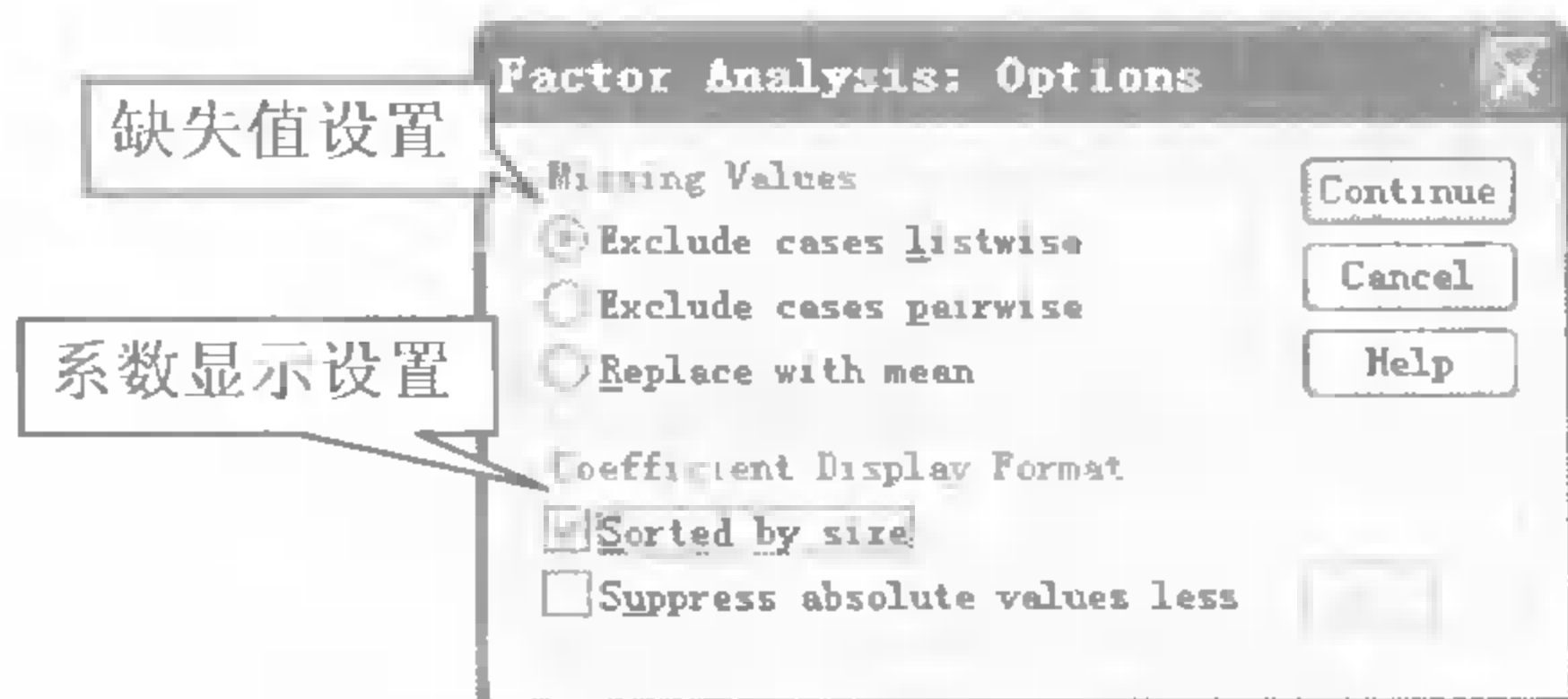


图 11-4 因子分析的 Options 设置

- **Exclude cases listwise**，成对剔除带有缺失值的观测量，在计算某个特定的统计量时，只有当前用到的某个变量有缺失值时，才将相应的记录删除，比如计算两个变量的相关系数时，只把这两个变量中带有缺失值的记录剔除，如果某个记录的这两个变量没有缺失值，而其他变量中有，那么此记录仍用于当前相关系数的计算。
- **Exclude cases pairwise**，成对剔除带有缺失值的观测量，在计算某个特定的统计量时，只有当前用到的某个变量有缺失值时，才将相应的记录删除，比如计算两个变量的相关系数时，只把这两个变量中带有缺失值的记录剔除，如果某个记录的这两个变量没有缺失值，而其他变量中有，那么此记录仍用于当前相关系数的计算。
- **Replace with mean**，用变量的均值取代其缺失值。
- (2) **Coefficient Display Format** 栏，选择载荷系数的显示格式，有两个可选项。
  - **Sorted by size**，载荷系数按取值大小排列，使载荷矩阵中的同一因子上具有较高载荷的变量排在一起，便于观察和分析。
  - **Suppress absolute values less than**，不显示那些绝对值小于指定值的载荷系数，勾选此复选框后，在其后的输入框中键入 0~1 之间的数作为临界值，默认值为 0.10。

### 3. 统计输出设置

在图 11-2 中单击 Descriptives 按钮，弹出如图 11-5 所示的输出设置对话框。分别勾选如下复选框：Univariate descriptive、Coefficients、Significance levels 和 KMO and Bartlett's test of Sphericity；单击 Continue 按钮返回主界面。

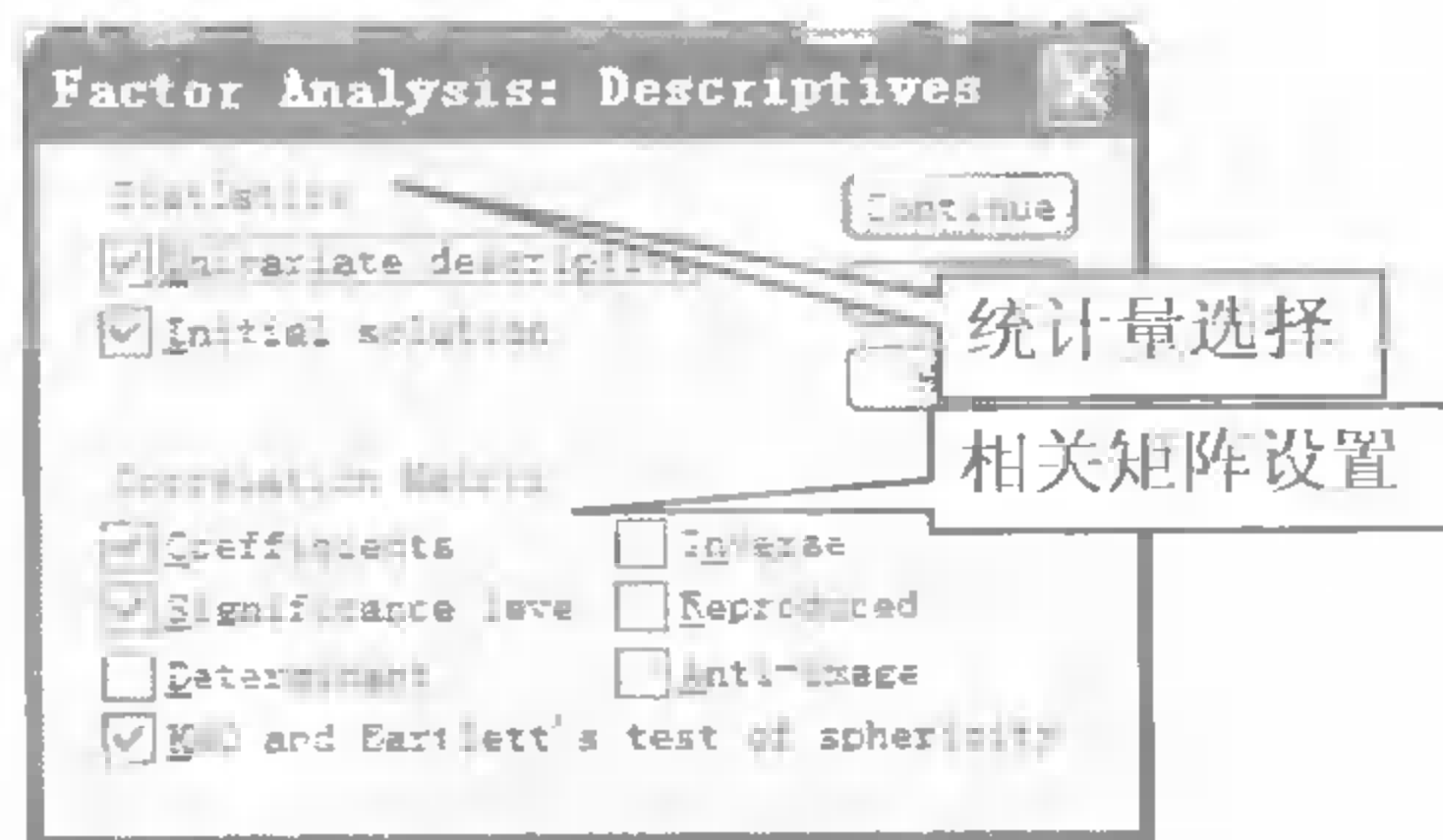


图 11-5 因子分析的 Descriptives 设置

- (1) **Statistics** 栏，选择输出哪些统计量，有如下两个复选框。
  - **Univariate descriptive**，单变量的描述性统计量，勾选它表示输出参与分析的每个初始变量的均值、标准差和有效取值个数等。
  - **Initial solution**，输出包括初始公共因子、初始特征根和初始方差贡献率等，对主成分分析来说，这些值包括了分析变量的相关矩阵或协方差矩阵的对角元素；对因子分析来说，这些值又包括了每个变量在各因子上的载荷的平方和。
- (2) **Correlation Matrix** 栏，选择与相关矩阵有关的输出，有如下 7 个复选框。
  - **Coefficients** 相关系数，输出初始分析变量之间的相关系数矩阵。

- Significance levels 显著性水平，输出每个相关系数关于单侧假设检验的显著性 P 值（零假设是相关系数为 0，即指定变量不相关）。
- Determinant，相关系数矩阵的行列式。
- Inverse，相关系数矩阵的逆矩阵。
- Reproduced 再生相关矩阵，输出因子分析后的相关矩阵，还给出残差（即原始相关与再生相关之间的差值）。
- Anti-image（反象相关矩阵）包括偏相关系数的负数，反象协方差矩阵包括偏协方差的负数。一个好的因子模型对角线上的元素应较大，非对角线元素应较小。
- KMO and Bartlett's test of Sphericity（KMO 检验和球形 Bartlett 检验），此选项输出对采样充足度的 Kaisex-Meyer-Olkin 测度，以检验变量间的偏相关是否很小；Bartlett 球形检验则用来验证相关矩阵是否是单位阵，即各变量间是否独立，若不能否定相关矩阵为单位阵，就说明各变量可能独自提供了一些信息，再采用因子模型就不合适了。

#### 4. 计算公因子的方法设置

在图 11-2 中单击 Extraction 按钮，弹出图 11-6 所示的对话框，在此设置与因子提取方法相关的参数。在 Method 下拉列表保留默认的 Principal components 主成分法；勾选 Scree plot 复选框；单击 Continue 按钮返回主界面。

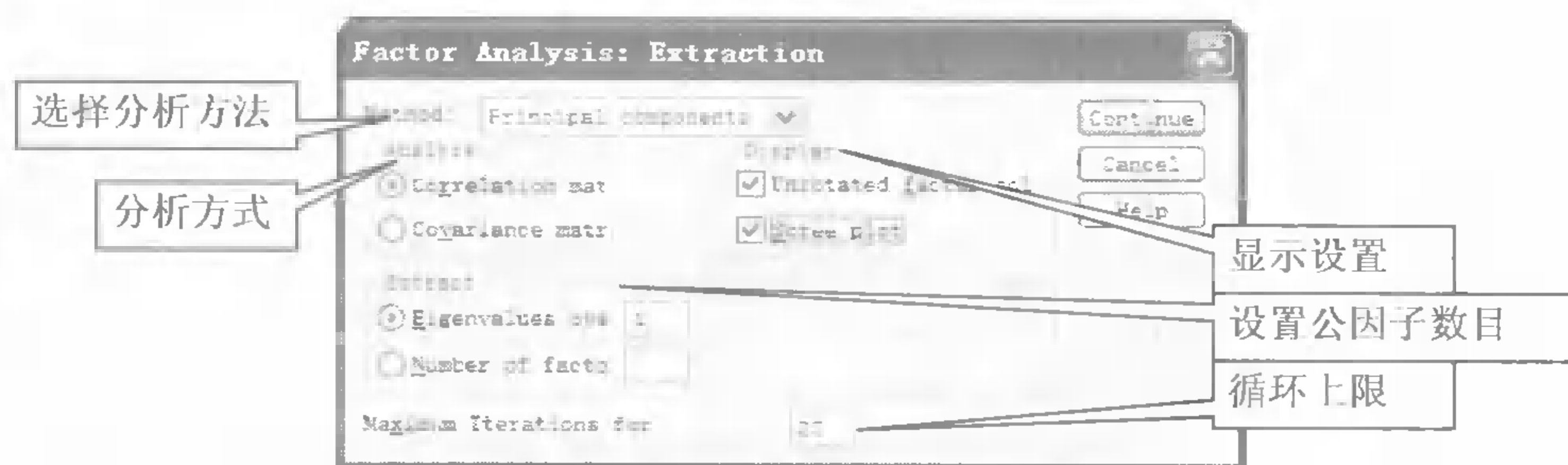


图 11-6 提取公因子的方法设置

(1) Method 下拉列表，SPSS 在此提供了如下 7 种公因子的提取方法。

- Principal components 主成分法，该方法假设变量是因子的线性组合，第一主成分有最大的方差，后续主成分所解释的方差逐个递减，各主成分之间互不相关，主成分法通常用来计算初始公因子，它也适用于相关矩阵为奇异时的情况。
- Unweighted least Square 不加权最小平方法，该方法使得观测的相关矩阵和再生的相关矩阵之差的平方和最小，不计对角元素。
- Generalized least squares 加权最小平方法，使观测和再生相关矩阵之差的平方和最小，并以变量单值的倒数（the inverse of the uniqueness of the variables）对相关系数加权。
- Maximum Likelihood 极大似然法，如果样本来自多元正态总体，参数估计过程与计算初始变量的相关矩阵极为相似；以变量单值的倒数对相关系数加权，并使用迭代算法。
- Principal Axis factoring 公因素分析法，从初始相关矩阵提取公因子，并把多元相关系数的平方置于对角线上（作为变量共同度的初始估计），再用初始因子载荷估计新的变量共同度（取代对角线上的初始估计），如此重复直至变量共同度在两次相邻迭代中的变化达到临界条件。



- Alpha factoring  $\alpha$  因子提取法, 该方法把当前分析变量看作是所有潜在变量的一个样本, 最大化因子的  $\alpha$  可靠性。

- Image factoring 映象因子提取法, 它把每个变量的主要部分定义为其各变量的线性回归, 而不是潜在因子的函数, 称之为偏映象 (partial image)。

(2) Analyze 栏, 用于选择计算公因子的矩阵, 有两个选择。

- Correlation Matrix, 指定以分析变量的相关矩阵作为提取公因子的依据, 适用于各变量的度量单位不同时的情况。
- Covariance matrix, 指定以分析变量的协方差矩阵作为提取公因子的依据, 适用于各变量的方差不相等时的情况。

(3) Display 栏, 选择与因子提取有关的输出选项, 有如下两个复选项。

- Unrotated factor solution: 输出未经旋转的因子载荷矩阵, 默认为选中状态。
- Scree plot: 输出以按特征值大小排列的因子序号为横轴、特征值为纵轴所作的碎石图, 用来帮助确定保留多少个公因子; 典型的碎石图会有一个明显的拐点, 在该点之前是代表大因子的陡峭折线, 之后是代表小因子的缓坡折线。

(4) Extract 栏, 在此设置提取公因子的规则, 有两种方法。

- Eigenvalues over 单选框, 指定想要提取的公因子的最小特征值, 默认值为 1。
- Number of factors 单选框, 指定想要提取的公因子的数目, 理论上有多少个分析变量, 最多就有多少个公因子。

(5) Maximum iterations for Convergence 输入框, 指定因子提取算法收敛的最大迭代次数, 默认值为 25。

## 5. 因子旋转设置

在图 11-2 中单击 Rotation 按钮, 弹出如图 11-7 所示的因子旋转设置对话框。单击选中 Varimax 单选框; 勾选 Loading plot(s) 复选框; 单击 Continue 按钮返回主界面。

(1) Method 栏, 在此选择因子旋转的方法, 可选项有如下 6 个。

- None 不旋转, 此为默认选项。
- Varimax 方差最大旋转, 这是一种正交旋转方法, 它使得每个因子上具有较高载荷的变量数目最小, 由此可以简化对因子的解释。
- Direct Oblimin 直接斜交旋转, 这是非正交旋转方法, 选中此项后需要在下面指定 Delta 参数, Delta 取 0 (默认) 时倾斜性最大; Delta 小于零时取值越小倾斜性越小, Delta 的取值需小于等于 0.8。
- Quartimax 四次最大正交旋转, 该旋转方法使每个变量中需要解释的因子数最少, 由此简化对初始变量的解释。
- Equamax 平均正交旋转, 它是 Varimax 方法与 Quartimax 方法的结合, 使得在每个公因子上较高载荷的变量数目、解释初始变量的公因子数目都达到最少。
- Promax, 一种斜交旋转方法, 它允许公因子间彼此相关, 计算起来比直接斜交旋转更快, 因此适用于对大数据集的分析。

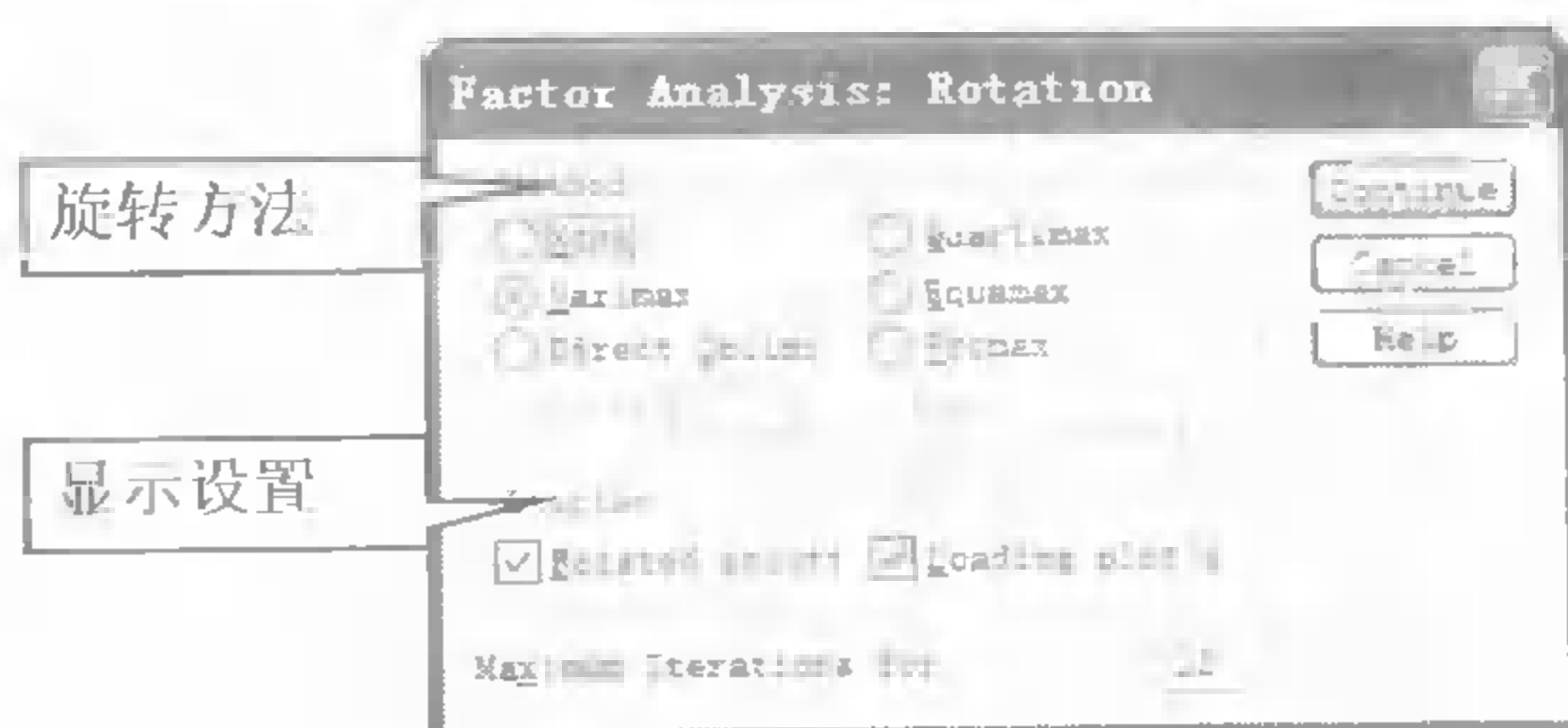


图 11-7 因子旋转设置

(2) Display 栏, 在此选择有关因子旋转的输出, 有如下两个复选项。

Rotated solution 旋转结果, 只有指定了某种旋转方法后此项才可选, 对正交旋转将显示旋转后的模式矩阵、因子转换矩阵; 对斜交旋转将显示旋转后的模式矩阵、因子结构矩阵和因子间的相关矩阵。

Loading plot(s) 因子载荷散点图, 如果有两个公因子, 输出各原始变量在 Factor1、Factor2 坐标系中的散点图; 如果多于两个公因子, 输出前三个公因子的三维因子载荷散点图; 如果只提取了一个公因子, 不作输出; 输出的都是经旋转后的因子载荷图。

(3) Maximum iterations for Convergence 输入框, 指定因子旋转收敛的最大迭代次数, 默认值为 25。

## 6. 因子得分设置

在图 11-2 中单击 Scores 按钮, 弹出图 11-8 所示的因子得分设置对话框。分别勾选 Save 复选框和 Display 复选框; 单击 Continue 按钮返回主界面。

(1) Save as variables 复选框, 表示把每个因子得分作为一个新变量保存到目前数据集。

(2) Method 栏, 选择估计因子得分系数的方法, 可选项有如下 3 种。

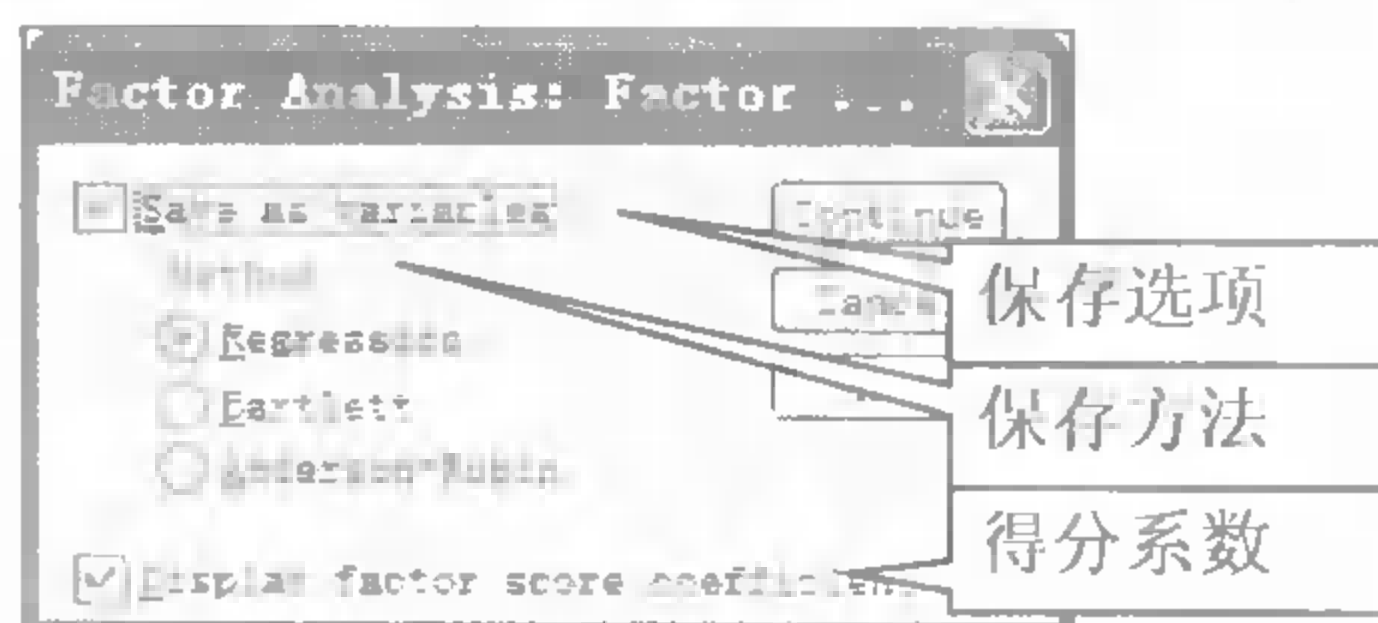


图 11-8 因子得分设置

Regression 回归法, 其因子得分的均值为 0, 方差等于估计因子得分与实际因子得分之间的多元相关的平方, 即使公因子正交时此得分也可能是相关的。

Bartlett 巴特利特法, 因子得分均值为 0。

Anderson-Rubin, 它是 Bartlett 方法的调整, 可以保证估计因子的正交性, 其因子得分的均值为 0、标准差为 1, 且彼此不相关。

(3) Display factor score coefficient matrix 复选框。勾选后, 输出标准化的因子得分系数矩阵, 对原始变量进行标准化后, 可以根据该矩阵计算各观测量的因子得分; 还显示了因子得分变量之间的相关矩阵。

## 11.2.3 结果分析

在图 11-2 中, 单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 11-9~图 11-17 所示。

描述统计量			
	均值	标准差	分析 N
GDP 总值 (亿元)	2078.1203	1852.45232	30
第三产业增加值 (亿元)	875.1800	602.73732	30
GDP 增长速度	110.7000	1.71208	30
第三产业增加值占 GDP 比重	41.967	1.43373	30
第三产业从业人员比重	27.1967	1.16984	30
社会综合生产率	1327.6167	863.47968	30
人均 GDP (元/人)	6941.3000	4635.28118	30
人均税收 (元/人)	929.4295	618.39745	30
资金利税率	6.4243	3.86178	30
城镇居民人均收入 (元/人)	5188.8067	1481.25671	30
农村居民人均收入 (元/人)	2204.3063	557.73676	30

图 11-9 初始变量的描述性输出

(1) 描述性统计输出。图 11-9 所示是关于 11 个初始变量的描述性统计量, 包括均值、标准差和分析用到的取值个数 (N)。

(2) 初始变量的相关性检验。如图 11-10 所示，从初始变量的相关系数矩阵看，多个变量之间的相关系数较大（图中用蓝色线框标识），且其对应的 Sig 值普遍较小，说明这些变量之间存在着较为显著的相关性，进而也说明了有进行因子分析的必要。

相关矩阵										
		GDP增长 速度	第三产业 增加值占 GDP比重	第三产业从 业人员比重	社会综合 生产率	人均GDP (元/人)	人均税收 (元/人)	资金利税率	城镇居民 人均收入 (元/人)	农村居民 人均收入 (元/人)
相关	GDP总值(亿元)	418	191	100	294	398	255	234	540	498
	第三产业增加值(亿元)	425	012	235	432	450	392	205	646	613
	GDP增长速度	1 000	079	153	359	351	282	288	308	422
	第三产业增加值占GDP比重	079	1 000	679	556	521	527	278	421	465
	第三产业从业人员比重	153	679	1 000	834	929	748	200	544	762
	社会综合生产率	359	556	834	1 000	987	940	070	687	919
	人均GDP(元/人)	351	521	929	987	1 000	961	097	734	938
	人均税收(元/人)	282	527	748	940	961	1 000	198	682	873
	资金利税率	288	278	200	070	097	198	1 000	205	110
	城镇居民人均收入(元/人)	308	421	544	687	734	682	205	1 000	774
	农村居民人均收入(元/人)	422	465	762	919	938	873	110	774	1 000
Sig (单侧)	GDP总值(亿元)	011	159	300	056	083	086	106	001	003
	第三产业增加值(亿元)	010	475	106	609	006	016	137	000	000
	GDP增长速度		028	210	026	029	065	061	049	010
	第三产业增加值占GDP比重	238		000	000	000	001	069	010	005
	第三产业从业人员比重	210	000		000	000	000	145	001	000
	社会综合生产率	026	000	000		000	000	357	000	000
	人均GDP(元/人)	029	000	000	000		000	504	000	000
	人均税收(元/人)	065	001	000	000	000		147	000	000
	资金利税率	061	069	145	057	304	147		139	282
	城镇居民人均收入(元/人)	049	010	001	000	000	000	139		000
	农村居民人均收入(元/人)	010	005	000	000	000	000	282	000	

图 11-10 初始分析变量的相关矩阵

(3) KMO 检验和 Bartlett 球形检验。如图 11-11 所示，KMO 检验用于研究变量之间的偏相关性，计算偏相关时由于控制了其他因素的影响，所以会比简单相关系数来得小。一般 KMO 统计量大于 0.9 时效果最佳，0.7 以上可以接受，0.5 以下不宜作因子分析，本例中 KMO 取值 0.692 尚可接受。

KMO和 Bartlett 的检验		
取样足够度的 Kaiser-Meyer-Olkin 度量。		692
Bartlett 的球形 度检验	近似卡方	475.461
	df	55
	Sig	000

公因子方差		
	初始	提取
GDP总值(亿元)	1 000	977
第三产业增加值(亿元)	1 000	972
GDP增长速度	1 000	485
第三产业增加值占GDP比重	1 000	753
第三产业从业人员比重	1 000	347
社会综合生产率	1 000	950
人均GDP(元/人)	1 000	984
人均税收(元/人)	1 000	910
资金利税率	1 000	851
城镇居民人均收入(元/人)	1 000	725
农村居民人均收入(元/人)	1 000	933
提取方法：主成分分析。		

图 11-11 球形检验和公因子方差

Bartlett 球形检验统计量的 Sig<0.01，由此否定相关矩阵为单位阵的零假设，即认为各变量之间存在着显著的相关性，这与从图 11-10 所示的相关矩阵得出的结论相符。

(4) 变量的共同度。如图 11-11 所示，“公因子方差”表格实际给出的就是初始变量的共同度，“提取”列表示变量共同度的取值。共同度取值区间为[0,1]，本例中 GDP 总值的共同度为 0.977，

可以理解为 3 个公共因子能够解释 GDP 总值的方差的 97.7%，其他变量共同度的解释类似。

(5) 方差解释表。如图 11-12 所示，“说明的总方差”表格给出了每个公因子所解释的方差及其累计和。观察“初始特征值”一栏下的“累积%”列，前 3 个公因子解释的累计方差已经达到 85% 以上，故而提取这 3 个公因子就能够比较好地解释原有变量所包含的信息了。

说明的总方差									
成分	初始特征值			提取平方和载入			旋转平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	5.101	55.461	55.461	5.101	55.461	55.461	5.265	47.868	47.868
2	2.304	20.941	76.402	2.304	20.941	76.402	2.672	24.289	72.157
3	1.001	9.099	85.501	1.001	9.099	85.501	1.468	13.345	85.501
4	.095	0.861	91.819						
5	.391	3.558	95.377						
6	.224	2.038	97.416						
7	.176	1.604	99.020						
8	.069	0.628	99.648						
9	.031	0.284	99.932						
10	.005	0.049	99.981						
11	.002	0.019	100.000						

提取方法：主成分分析。

图 11-12 方差解释输出

“提取平方和载入”一栏表示在未经旋转时，被提取的 3 个公共因子各自的方差贡献信息，它们和“初始特征值”栏的前 3 列取值一样，说明前 3 个公因子可以解释总方差的 85.501%，即总体多于 85% 的信息可以由这 3 个公共因子来解释。最后一栏“旋转平方和载入”表示经过因子旋转后得到的新公因子的方差贡献值、方差贡献率和累计方差贡献率，和未经旋转相比，每个因子的方差贡献值有变化，但最终的累计方差贡献率不变。

(6) 特征值碎石图。图 11-13 所示是关于初始特征值（也就是方差贡献）的碎石图。它实际上就是根据图 11-12 中“初始特征值”栏下的“合计”列的数据所作的图形，并将特征值按照降序排列。观察发现，第 3 个公因子后的特征值变化趋缓，故而选取 3 个公因子是比较恰当的。

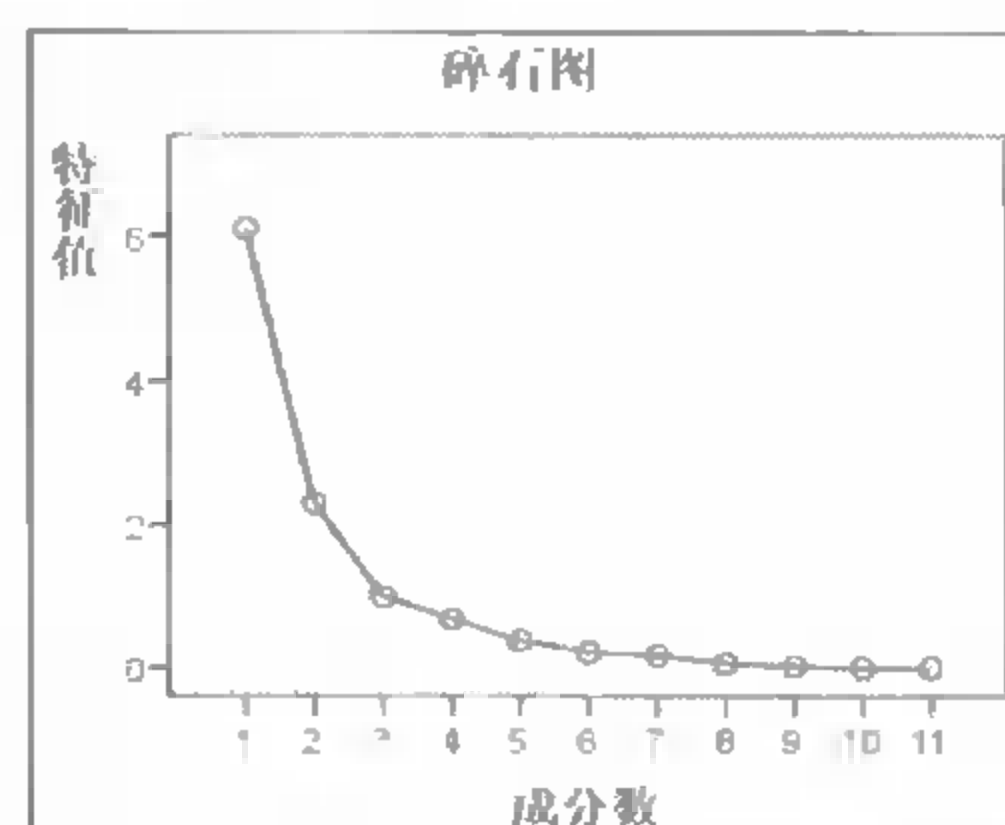


图 11-13 碎石图

(7) 旋转前后的因子载荷阵。如图 11-14 所示，“成分矩阵”是初始的未经旋转的因子载荷矩阵，“旋转成分矩阵”是经过旋转后的因子载荷矩阵。通过观测可以发现，旋转后每个公因子上的载荷分配地更清晰了，因而比未旋转时更容易解释各因子的意义。另外还输出了因子旋转矩阵，如图 11-15 所示。

成分矩阵 <sup>a</sup>			
	成分		
	1	2	3
人均GDP (元/人)	.959	-.180	.111
农村居民人均收入 (元/人)	.965	.025	.011
社会综合生产率	.953	-.194	.105
人均税收 (元/人)	.912	-.162	.229
城镇居民人均收入 (元/人)	.831	.165	-.083
第三产业从业人员比重	.800	-.446	-.088
GDP总值 (亿元)	.900	.754	-.379
第三产业增加值占GDP比重	.958	-.668	.096
第三产业增加值 (亿元)	.630	.659	-.375
GDP增长速度	.431	.477	.258
资金利税率	.132	.537	.738

提取方法 主成分分析法。  
a. 已提取了 3 个成分。

旋转成分矩阵 <sup>a</sup>			
	成分		
	1	2	3
人均GDP (元/人)	.952	.234	.152
社会综合生产率	.943	.220	.137
人均税收 (元/人)	.906	.160	.252
第三产业从业人员比重	.900	.071	-.179
农村居民人均收入 (元/人)	.842	.447	.154
第三产业增加值占GDP比重	.782	.194	-.338
城镇居民人均收入 (元/人)	.653	.528	.144
GDP总值 (亿元)	.056	.976	.150
第三产业增加值 (亿元)	.219	.954	.121
资金利税率	-.043	.062	.920
GDP增长速度	.196	.395	.538

提取方法 主成分分析法。  
旋转法 具有 Kaiser 标准化的正交旋转法。  
a. 旋转在 4 次迭代后收敛。

图 11-14 旋转前后的因子载荷



成分转换矩阵			
成分	1	2	3
1	886	438	156
2	-451	730	514
3	111	-525	844

提取方法 主成分分析法。  
旋转法 具有 Kaiser 标准化的正交旋转法。

图 11-15 因子旋转矩阵

已知因子载荷是变量与公共因子的相关系数，对一个变量来说，载荷绝对值较大的因子与它的关系更为密切，也更能代表这个变量。按照这一观点，第 1 公因子更能代表人均 GDP、社会综合生产率、人均税收、第三产业从业人员比重和农村居民人均收入这几个变量因素；第 2 公因子则更适合代表 GDP 总值和第三产业增加值两个变量；第 3 公因子则较好地代表了资金利税率这个变量。

进一步分析，根据各个变量的特点，可以把第 1 个公因子解释为收入因素，因为它反映了多个代表收入的变量；类似地，把第 2 个公因子解释为生产力因素；把第 3 个公因子解释为利税因素。这样就可以利用新提取出的 3 个潜在因素（收入因素、生产力因素和利税因素）对样本中 30 个地区的经济发展规律加以描述了。

（8）旋转后的因子载荷图。如图 11-16 所示，是旋转后的因子载荷散点图，本来对于 3 个公因子输出的是 3 维图形，但为了显示清晰，此处把它旋转成 2 维图形的形式，即变量关于前 2 个公因子载荷的平面图。

它实际上就是根据图 11-14 里的“旋转成分矩阵”表格中的成分 1 和成分 2 两列数据所作，由此图观察所得的信息与对“旋转成分矩阵”所作的分析一致。例如在图中 x1、x2 可以归于同一个公因子解释，它正是上一步中提到的生产力因素。

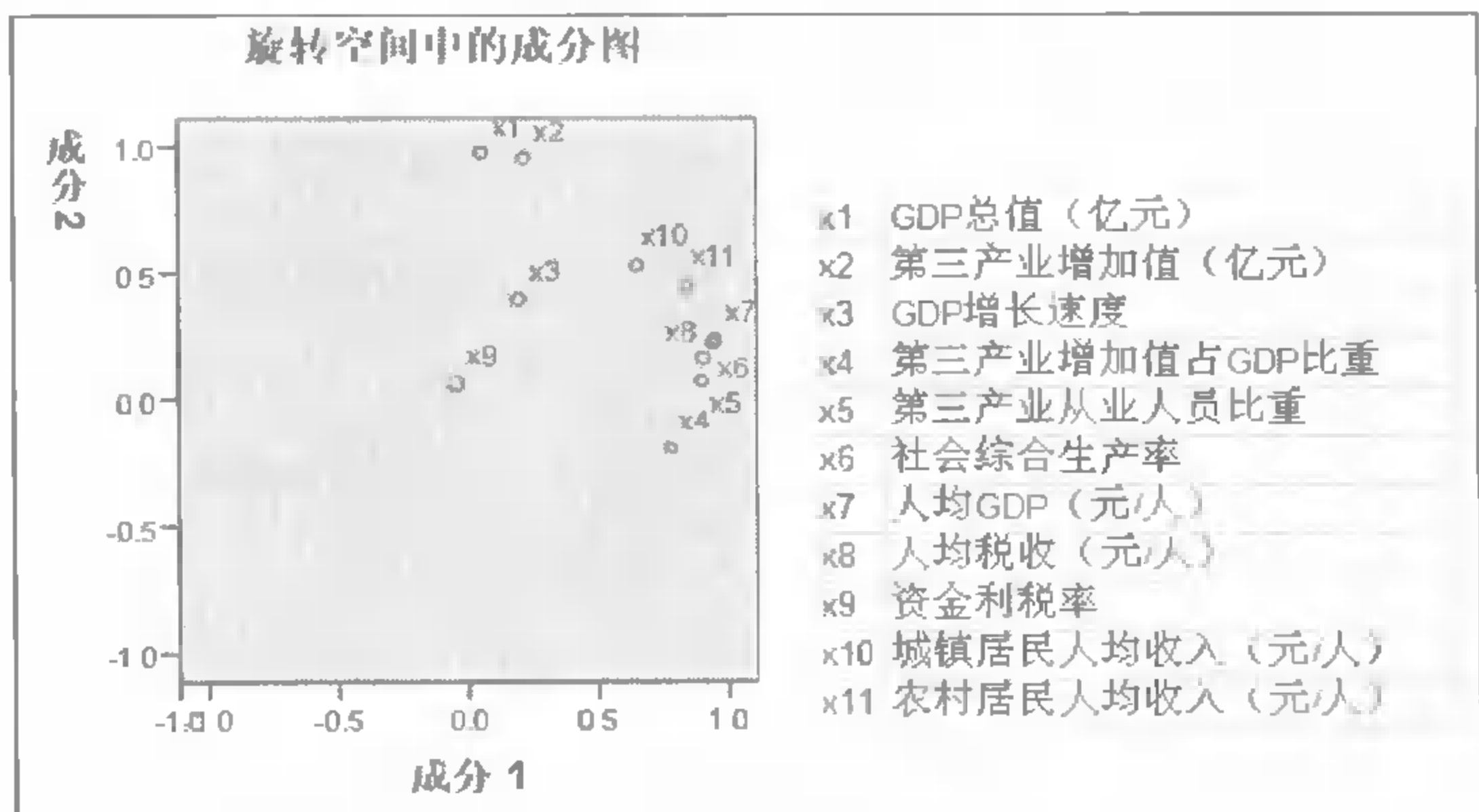


图 11-16 因子载荷散点图

（9）因子得分的系数矩阵。图 11-17 所示是因子得分系数矩阵，由此可得最终的因子得分公式为：

$$F_1 = -0.119 * \text{GDP总值} - 0.079 * \text{第三产业增加值} + \dots + 0.134 * \text{农村居民人均收入};$$

$$F_2 = 0.477 * \text{GDP总值} + 0.451 * \text{第三产业增加值} + \dots + 0.083 * \text{农村居民人均收入};$$

$$F_3 = -0.136 * \text{GDP总值} - 0.153 * \text{第三产业增加值} + \dots + 0.021 * \text{农村居民人均收入}$$

回到 Data Editor 窗口的当前数据集, 会看到文件中增加了 3 列: FAC1\_1 (第 1 因子得分)、FAC2\_1 (第 2 因子得分) 和 FAC3\_1 (第 3 因子得分)。其中北京、上海和天津 3 个直辖市的第 1 因子得分最高, 表明其在收入因素方面拥有绝对的优势; 江苏、山东和广东 3 个省份的第 2 因子得分最高, 正体现了其在生产发展方面较强的实力; 而云南、福建、安徽和黑龙江等省份在资金利税因素上的表现较为突出。

(10) 综合得分的分析。如果研究者还关心各地区的综合实力, 可对 3 个公因子的得分进行加权求和, 权数就取其方差贡献值或方差贡献率, 参考图 11-12 中“旋转平方和载入”栏里的“合计”(方差值)和“方差的%”(方差贡献率)。本例采用方差贡献率作为权重, 3 个旋转后公因子的方差贡献率依次为 47.87%、24.29%和 13.35%, 于是可得地区综合得分的计算公式如下:

$$zF = 47.87\% * FAC1\_1 + 24.29\% * FAC2\_1 + 13.35\% * FAC3\_1$$

依次单击菜单“Transform→Compute Variables”执行计算新变量过程, 其主设置界面如图 11-18 所示, 在 Target Variable 下输入 zF, 在 Expression 下输入如上的 zF 计算公式; 单击 OK 按钮运行, 即可在当前数据集中生成代表综合得分的变量 zF。

成分得分系数矩阵			
	成分		
	1	2	3
GDP总值(亿元)	-.119	.477	-.136
第三产业增加值(亿元)	-.079	.451	.153
GDP增长速度	-.001	.041	.343
第三产业增加值占GDP比重	.201	-.121	-.216
第三产业从业人员比重	.194	-.037	-.153
社会综合生产率	.188	-.048	.070
人均GDP(元/人)	.188	.046	.078
人均税收(元/人)	.190	-.106	.180
资金利税率	-.004	-.208	.746
城镇居民人均收入(元/人)	.079	.155	-.012
农村居民人均收入(元/人)	.133	.083	.021

提取方法: 主成分分析法。  
旋转法: 具有 Kaiser 标准化的正交旋转法。  
构成得分。

图 11-17 因子得分的系数矩阵

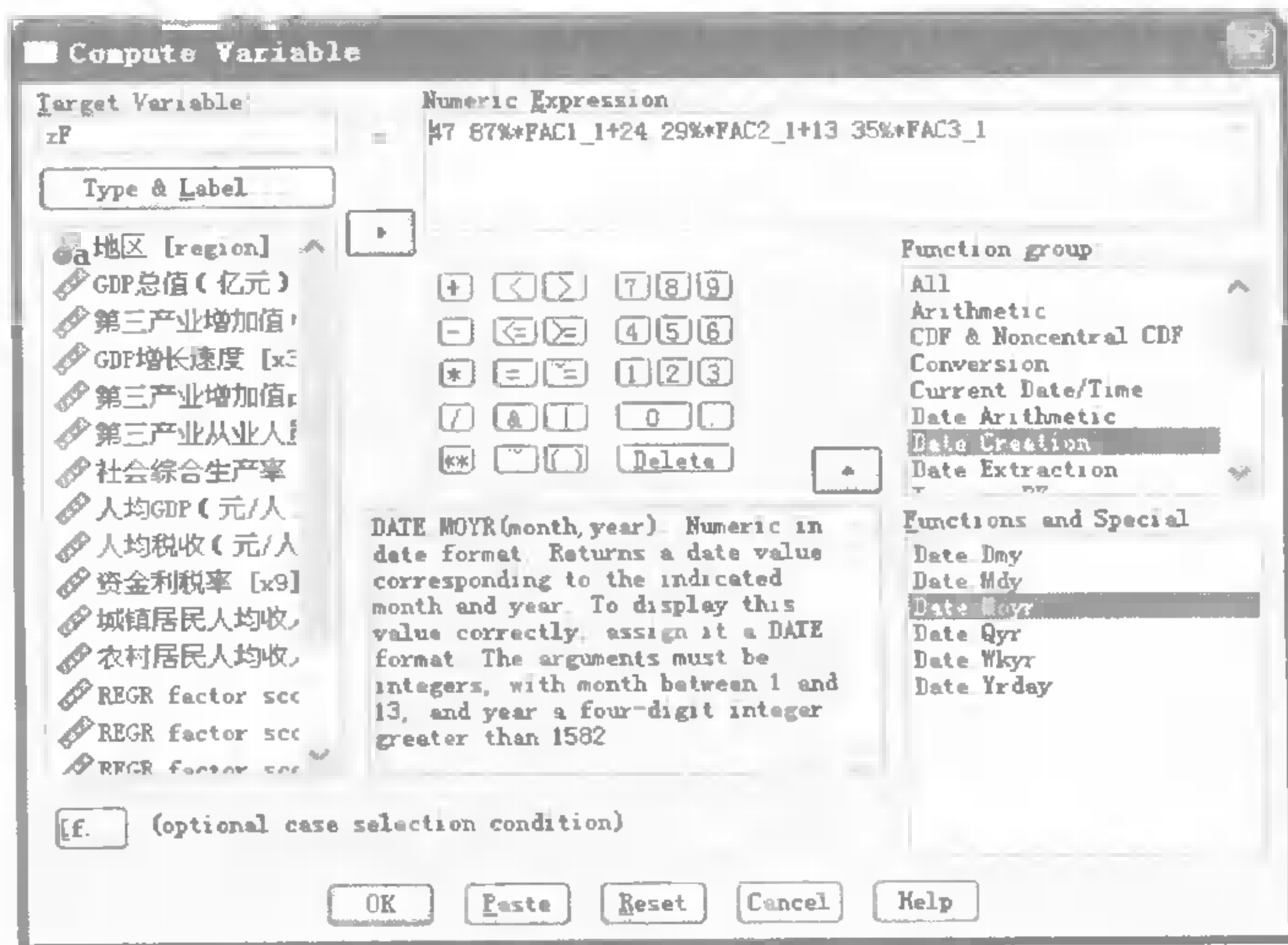


图 11-18 计算新变量的设置界面

在最终得到的综合得分里, 上海、北京、广东、天津、浙江、江苏、浙江几个省市的综合得分占据了前几位, 这与当时的实际情况也是相符的。

# 第 12 章 分类分析

分类是人类认识世界的基本方法，聚类分析和判别分析是多元统计分析里两种基础的分类方法，掌握这两个方法对运用统计手段认识世界具有非常重要的意义。SPSS 中的聚类分析又分为 3 个过程：快速样本聚类、分层聚类和两阶段聚类。判别分析法先根据已知类别归属的观测样本得到一个判别函数，再据此对未知类别的观测进行分类。本章最后一节介绍数据挖掘中常用的分类和预测算法决策树分析法及其 SPSS 的实现过程 Classification Trees。

## 12.1 聚类分析的原理简介

聚类分析是对样品或变量进行分类的一种多元统计方法，目的在于将相似的事物归类。

聚类分析并不是一种纯粹的统计技术，其方法基本上与分布理论和显著性检验无关，一般不用于从样本推断总体的研究。在市场研究中，聚类分析主要用于市场细分、研究消费者行为、寻找新的潜在市场和作为其他统计分析的预处理等。

### 12.1.1 聚类分析的基本概念

聚类 (Clustering) 是将某个对象集划分为若干组 (Class 或 Cluster) 的过程，使得同一个组内的数据对象具有较高的相似度，而不同组中的数据对象是不相似的。相似或不相似的定义基于属性变量的取值确定，一般就采用各对象间的距离来表示。一个聚类 (Cluster) 就是由彼此相似的一组对象所构成的集合，同组的对象常常被当作一个对象加以对待。

聚类分析属于无监督的学习方法，它不依靠事先已知的数据分类，也不依靠标有数据类别的训练样本集合。正因为如此，聚类分析是一种通过观察的学习方法 (Learning by observation)，而不是通过示例去学习规则 (Learning by example)。

#### 1. Q 型聚类和 R 型聚类

在聚类分析中，“性质”是由一组变量 (variables) 来代表的，它用一个  $p$  维的向量表示  $\bar{x} = (X_1, X_2, \dots, X_p)'$ ；某两个观察对象  $\bar{x}_i$  和  $\bar{x}_j$  的“差异”程度由它们之间的距离来度量，根据距离定义的不同，这种差异又分为好多种情况。

当聚类是要把所有的观测记录 (cases) 进行分类时，它把性质相近的观测分在同一个类，性质差异较大的观测分在不同的类，这称之为 Q 型聚类。

当聚类把变量 (variables) 作为分类对象时，称之为 R 型聚类。这种聚类用在变量数目

比较多、且相关性比较强的情形，目的是将性质相近的变量聚为同一个类，并从中找出代表变量，从而减少变量个数以达到降维的效果。

## 2. 聚类分析的应用

在科学研究和社会生产的许多领域（例如模式识别、机器学习、数据挖掘、图像处理和市场分析等）都渗透着聚类分析的研究和应用。

聚类分析的典型应用包括：在商业方面，帮助市场研究人员发现拥有不同特征的顾客组群，并可利用购买模式对其进行描述；在生物方面，可用来获取动物或植物群体内存在的层次结构（taxonomies），还能根据基因功能对其进行分类，由此获得对群体固有结构更深入的了解；它还可以利用地球观测数据库，帮助用户识别具有相似土地使用情况的区域；帮助研究者分类和识别互联网上的文档，以便发现潜在的信息；作为数据挖掘的一项功能，聚类分析还可以作为一个单独使用的工具，用来帮助分析数据的分布、了解数据的特征，找出感兴趣的数据子集作进一步分析；此外，聚类分析也可以作为其他算法的预处理步骤。

作为统计学的一个分支，聚类分析已有多年的研究历史，这些研究主要集中在基于距离的聚类分析方面。现在的大多统计分析软件（例如 S-Plus、SPSS 和 SAS 等）都包含基于 K-均值、K-中心等聚类分析工具。

### 12.1.2 聚类分析的一般原理

本节以最基础的对观测记录的 Q 型系统聚类法为例，简单介绍聚类的一般原理和步骤。

系统聚类是一种逐次合并类的方法，在规定了样品之间的距离和类与类之间的距离后，先让  $n$  个样品各自成为一类；开始时，因每个样品自成一类，类与类之间的距离与样品之间的距离是相等的；然后，将距离最近的两个类合并；如此重复，每次循环减少一个类别，直至所有的样品归为一类为止。然而合并成一个类别就失去了聚类的意义，所以聚类过程应该在达到某个类水平数（即未合并的类数）时停下来，在此得到的聚类就是分析的结果。如何决定聚类个数是一个很复杂的问题，整个聚类过程还可以用二叉树谱系聚类图直观地表示出来。

#### 1. 系统聚类的步骤

把系统聚类的步骤作以简单总结，概括为如下 5 个部分。

- （1）定义样品之间的距离，以及类与类之间的距离。
- （2）令每个观测记录各自成为一个类别。
- （3）计算类与类之间的距离，并将距离最近的两个类合并为一个类，类的数目减 1。
- （4）如果当前的类的数目大于 1，转第③步。
- （5）结束聚类过程。

#### 2. 定义距离的方法

由聚类的一般步骤可见，如何定义样品之间和类之间的距离是关键。根据距离定义的不同，系统聚类又可以分为多种方法，但其执行的基本步骤都是一样的。

（1）样品之间的距离和相似度。

① 以样品  $\bar{x}_i$  和  $\bar{x}_j$  为例，它们之间的距离记为  $d_{ij}$ ，距离越小表示它们越相似，如下是 6



种常用的关于距离的度量方式。

- 欧氏距离 (Euclidian Distance):  $d_{ij} = \sqrt{\sum_{k=1}^p (X_{ik} - X_{jk})^2}$ 。
- 欧氏距离平方 (Squared Euclidian Distance):  $d_{ij} = \sum_{k=1}^p (X_{ik} - X_{jk})^2$ ，这是 SPSS 系统默认的距离定义方式。
- 闵可夫斯基距离 (Minkowski):  $d_{ij}(q) = \left( \sum_{k=1}^p |X_{ik} - X_{jk}|^q \right)^{1/q}$ ,  $q \geq 1$ ,  $q$  可由用户指定。
- 切比雪夫距离 (Chebyshev):  $d_{ij} = \max_{1 \leq k \leq p} \{ |X_{ik} - X_{jk}| \}$ 。
- 布洛克距离 (Block):  $d_{ij} = \sum_{k=1}^p |X_{ik} - X_{jk}|$ 。
- 自定义距离 (Customized):  $d_{ij} = \left( \sum_{k=1}^p |X_{ik} - X_{jk}|^q \right)^{1/r}$ ，其中参数  $q$ 、 $r$  为用户选项。

② 对于样品  $\bar{x}_i$  和  $\bar{x}_j$ ，还可以定义它们之间的相似系数，仍记为  $d_{ij}$ ，相似系数越大表示它们越相似。常用的相似系数有如下两个。

- 皮尔逊相似系数 (Pearson):  $d_{ij} = \frac{\sum_{k=1}^p (X_{ik} - \bar{X}_i)(X_{jk} - \bar{X}_j)}{\sqrt{\sum_{k=1}^p (X_{ik} - \bar{X}_i)^2} \sqrt{\sum_{k=1}^p (X_{jk} - \bar{X}_j)^2}}$ 。
- 夹角余弦 (Cosine):  $d_{ij} = \cos(\theta_{ij}) = \frac{\sum_{k=1}^p X_{ik} X_{jk}}{\sqrt{\sum_{k=1}^p X_{ik}^2} \sqrt{\sum_{k=1}^p X_{jk}^2}}$ 。

## (2) 类别之间的距离。

类 (Class) 指观测记录的集合，两个类之间的距离一般用类中某些特殊点之间的距离定义。设有两个类  $G_a$  和  $G_b$ ，它们之间的距离记为  $D(a, b)$ ，常用的类间距离有如下 5 个。

- 最短法:  $D(a, b) = \min \{ \bar{x}_i \in G_a, \bar{x}_j \in G_b \}$ 。
- 最长法:  $D(a, b) = \max \{ d_{ij} | \bar{x}_i \in G_a, \bar{x}_j \in G_b \}$ 。
- 重心法: 称  $\bar{x}_a = \frac{1}{n_a} \sum_{\bar{x}_i \in G_a} \bar{x}_i$ ,  $\bar{x}_b = \frac{1}{n_b} \sum_{\bar{x}_j \in G_b} \bar{x}_j$  分别为类  $G_a$  和  $G_b$  的重心，其中的  $n_a$  和  $n_b$  分别是  $G_a$  和  $G_b$  所含观测的个数，记  $D(a, b) = d_{ab}$ 。
- 类平均法:  $D(a, b) = \frac{1}{n_a n_b} \sum_{\bar{x}_i \in G_a} \sum_{\bar{x}_j \in G_b} d_{ij}$ 。
- 离差平方和法: 首先定义某个类  $G_s$  的直径为  $D_s = \sum_{\bar{x}_k \in G_s} (\bar{x}_k - \bar{x}_s)'(\bar{x}_k - \bar{x}_s)$ ，设  $G_a$ 、 $G_b$  和  $G_{a+b} = G_a \cup G_b$  的直径分别为  $D_a$ 、 $D_b$  和  $D_{a+b}$ ，记  $D^2(a, b) = D_{a+b} - D_a - D_b$ 。

## 3. 系统聚类的聚类个数

系统聚类最终把所有的观测聚为了一类，而如何确定恰当的聚类个数是一个比较困难的问题，因为分类本身就没有一定的标准。关于这一点，《实用多元统计分析》(王学仁、王松桂，上海科技出版社) 第 10 章给出了一个很好的关于扑克牌分类的例子，可以把扑克牌按花

色分类、按大小点分类、按桥牌的高花色低花色分类等。

有一些决定聚类个数的方法来自方差分析的思想，下面作一些简单介绍。

(1)  $R^2$  统计量。记  $R^2 = 1 - \frac{P_G}{T}$ ，其中  $P_G$  表示聚类数为  $G$  个时的总类内离差平方和， $T$  为所有变量的总离差平方和。 $R^2$  越大说明总的类内离差平方和相对越小，也就是说分为  $G$  个类是合适的。但显然聚类数越多，每个类别越小， $R^2$  也就越大，所以要综合考虑多个条件。取的  $G$  应该使得  $R^2$  足够大，但  $G$  本身比较小，而且  $R^2$  不再大幅度增加。

(2) 半偏相关系数。在把类  $G_K$  和类  $G_L$  合并为下一水平的类  $G_M$  时，定义半偏相关系数  $R^2 = \frac{B_{KL}}{T}$ ，其中  $B_{KL}$  为合并类引起的类内离差平方和的增量，半偏相关系数越大说明这两个类越不应该合并。所以如果在由  $n+1$  类合并为  $n$  类时半偏相关系数很大，就应该保留  $n+1$  个分类。

(3) 其他的选取规则还有双峰性系数、伪  $F$  统计量和伪  $t^2$  统计量等。

## 12.2 快速样本聚类过程

当聚类个数已知时，使用快速聚类过程可以快速地将观测记录分到各类中去，其特点是处理速度快、占用内存少。快速聚类适用于对大样本的聚类分析。

### 12.2.1 快速聚类简介

SPSS 的快速聚类过程使用的是  $k$  均值分类法 (K-Means Cluster)，它允许事先指定聚类个数，也可以指定使聚类过程中止的判据，比如迭代次数等。参与聚类的变量必须是数值型变量，且至少要有 1 个；为了清楚地表明各观测量最后聚到哪一类，还应指定一个对观测量的标识变量，例如编号、姓名等；聚类个数需大于等于 2，但不能大于数据集中的观测量个数。

若有  $n$  个数值型变量参与快速聚类，它们组成一个  $n$  维空间，把每个观测量看作是空间中的一个点，设最后要求的聚类个数为  $k$ 。下面简单介绍一下快速聚类的过程。

首先，选择  $k$  个观测量（由系统自动指定或由用户指定）作为聚类的初始种子，它们就是  $k$  个初始聚类中心点；然后把每个观测量都分派到与这  $k$  个中心距离最小的那个类中，得到第一次迭代形成的  $k$  个类；接着根据组成每一类的观测量计算各变量的均值，每一类的  $n$  个均值在  $n$  维空间中又形成  $k$  个点，这就是第二次迭代的类中心；按照这种方法依次迭代下去，直至达到指定的迭代次数或中止迭代的判据要求时，聚类过程结束。

从上述分析过程可以看出，K-Means Cluster 过程不仅是快速样本聚类，而且是一种逐步聚类分析，即先把被聚对象进行初始分类，然后通过逐步调整得到最终分类。

### 12.2.2 问题描述和数据准备

本节利用快速聚类分析对采用 14 个变量加以描述的 48 名人员进行分类。

某单位对前来应聘的 48 名人员进行了多项测试后，对直接表现其特征的 14 个方面进行了打分，每个单项的打分都采用 10 分制，得分越高说明当事人在此方面表现越好。所用数据文件为“应聘人员打分数据.sav”，数据格式如图 12-1 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Mea
1	no	Numeric	2	0	编号	None	None	3	Right	Scale
2	x2	Numeric	2	0	外貌	None	None	3	Right	Scale
3	x3	Numeric	2	0	学术能力	None	None	4	Right	Scale
4	x4	Numeric	2	0	讨人喜欢	None	None	4	Right	Scale
5	x5	Numeric	2	0	自信程度	None	None	3	Right	Scale
6	x6	Numeric	2	0	精明	None	None	4	Right	Scale
7	x7	Numeric	2	0	诚实	None	None	3	Right	Scale
8	x8	Numeric	2	0	推销能力	None	None	4	Right	Scale
9	x9	Numeric	2	0	经验	None	None	3	Right	Scale
0	x10	Numeric	2	0	积极性	None	None	3	Right	Scale
1	x11	Numeric	2	0	抱负	None	None	3	Right	Scale
2	x12	Numeric	2	0	理解能力	None	None	3	Right	Scale
3	x13	Numeric	2	0	潜力	None	None	3	Right	Scale
4	x14	Numeric	2	0	交际能力	None	None	3	Right	Scale
5	x15	Numeric	2	0	适应性	None	None	3	Right	Scale

图 12-1 应聘人员的测试打分数据

通过聚类,我们希望把这些应聘者分为少数几类,一方面可以了解不同类别的人的特点,另一方面可以根据各类人员所表现的特征,找出其最有可能胜任的岗位。

### 12.2.3 SPSS 快速聚类的设置

依次单击菜单“Analyze→Classify→K-Means Cluster...”执行 K 均值快速聚类过程,其主设置面板如图 12-2 所示,在此指定分析变量、模型方法和初始类中心等参数。

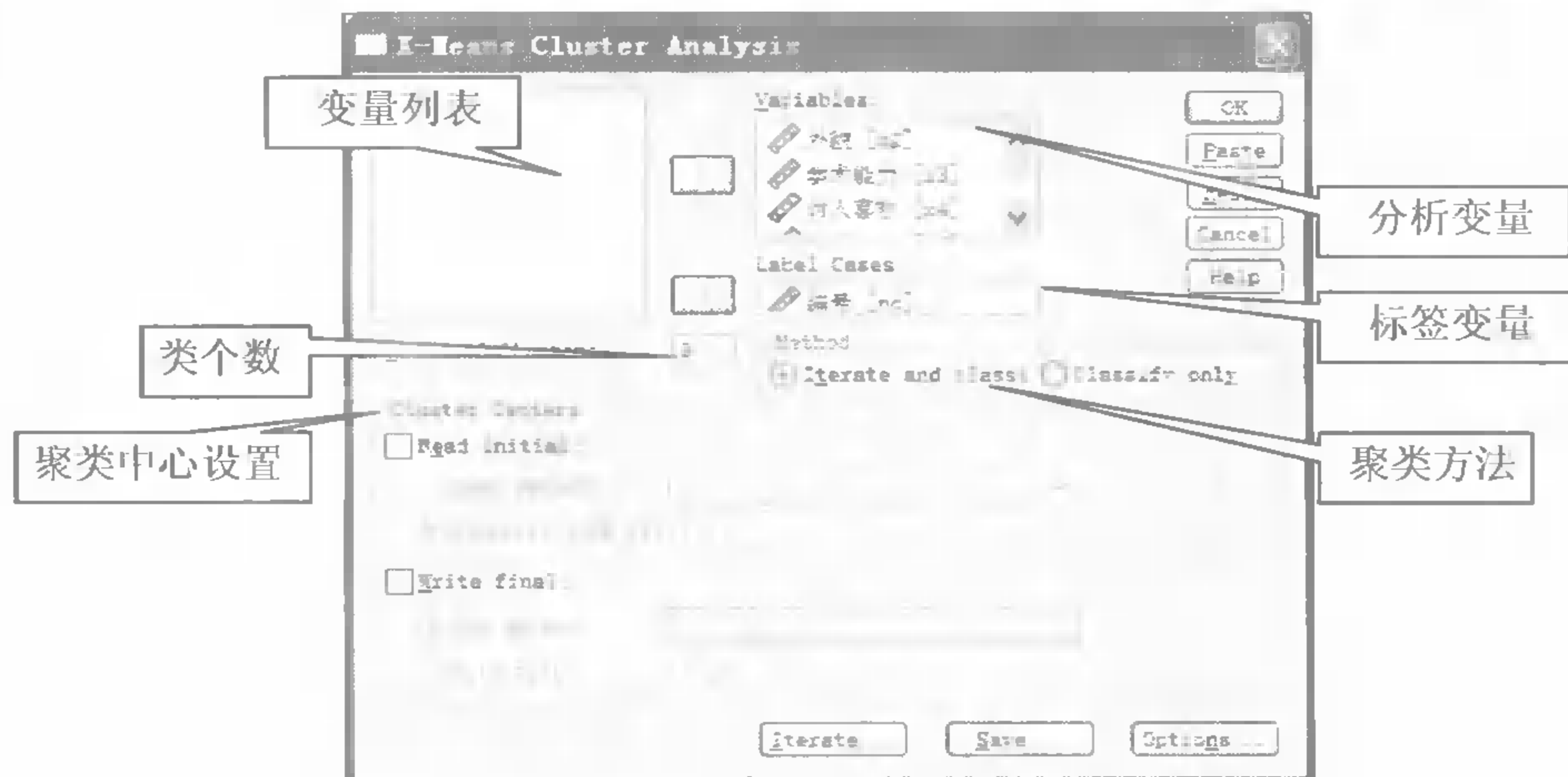




图 12-2 K 均值聚类的主设置面板

#### 1. 主面板的参数设置

在变量列表中选中从外貌(x2)到适应性(x15)的14个变量,单击从上至下第一个  按钮,将其作为分析变量选入 Variables 列表框;在变量列表中单击选中编号变量,单击从上至下第二个  按钮,将其作为标签变量选入 Label Cases 选框;在 Number of Clusters 后输入“3”;在 Method 栏单击选中 Iterate and classify 单选框。

(1) Variables 列表框,用于选入待分析的数值型变量。

(2) Label Cases 选框,用于选入标签变量,在结果中标识观测记录。

(3) Number of Clusters 输入框,指定聚类的个数,默认值为 2。

(4) Method 子设置栏,指定聚类的方法,SPSS 给出了如下两个可选项。

① Iterate and classify 单选框,先指定初始类别中心,然后按 K-means 算法迭代分类。

② **Classify only** 单选框, 选定初始类别中心点后, 只作分类而不再对中心点做任何更改。

综合使用这两个方法, 可以提高分析大型数据的效率, 具体做法是: 先从所有数据中抽取较小的一个样本, 用 **Iterate and classify** 方法进行聚类, 并在下面的 **Write final as** 栏把聚类结果保存起来; 然后对所有原始数据使用 **Classify only** 方法再次聚类, 并在下面的 **Read initial from** 栏指定刚刚存储的包含聚类中心的数据集或文件, 如此可以提高分析效率。

(5) **Cluster Centers** 栏, 设置与聚类中心有关的参数, 又分为如下两个子设置栏。

① **Read initial from** 复选框, 在此选择指定初始类中心的方法, 有两个选择。

☐ **Open dataset** 单选框, 选中后在其后的下拉列表指定一个当前打开的数据集。

☐ **External data file** 单选框, 选中后单击其后的 **File** 按钮指定存有初始类中心的文件。

② **Write final as** 复选框, 在此选择如何保存聚类结果的类中心, 也有两个选择。

☐ **New dataset** 单选框, 建立一个新数据集, 选中后在其后的键入框指定数据集的名称。

☐ **Data file** 单选框, 把结果写入一个外部文件, 单击其后的 **File** 按钮指定文件。

## 2. 迭代设置

在图 12-2 中单击 **Iterate** 按钮, 弹出如图 12-3 所示的迭代参数设置对话框, 单击 **Continue** 按钮返回主界面。

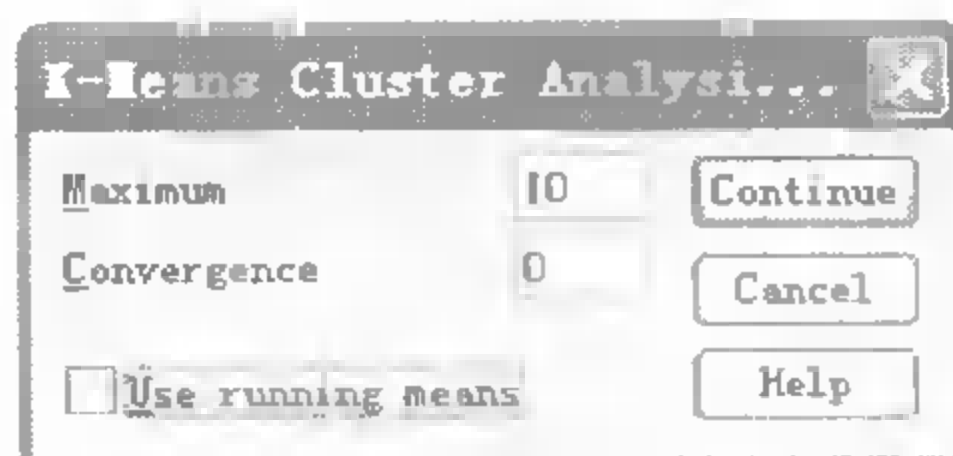


图 12-3 迭代参数设置

(1) **Maximum** 输入框, 指定 K-Means 算法的最大迭代次数, 取值范围 1~999, 默认值为 10。

(2) **Convergence** 输入框, 指定 K-Means 算法的收敛依据, 取值范围 0~1, 默认值为 0。

例如在 **Convergence** 后输入 0.02, 表示当某次迭代后, 类中心之间的距离变化的最小值, 小于初始类中心之间的最小距离的 2% 时, 迭代停止。如果同时设置了以上两个条件, 只要在迭代过程中满足了其中一个条件, 迭代就会停止。

(3) **Use running means** 复选框。勾选此项, 表示在每个观测量被分配到某一类后, 即可计算新的类中心; 不选中此项, 表示在完成了对所有观测量的一次分配后, 再计算新的类中心。不选中它能够节省迭代时间。

## 3. 保存设置

在图 12-2 中单击 **Save** 按钮, 弹出如图 12-4 所示的保存参数设置对话框, 单击 **Continue** 按钮返回主界面。



图 12-4 保存参数设置

在此, 有如下两个选项可以设置。

☐ **Cluster membership** 复选框, 表示用一个新变量 (默认名为 QCL\_1) 保存各观测量最



终被分配到哪一类，取值范围从 1 至聚类个数。

- Distance from cluster center 复选框，表示用一个新变量（默认名为 QCL\_2）保存各观测量到最终所属的类中心的欧氏距离。

#### 4. Options 选项设置

在图 12-2 中单击 Options 按钮，弹出如图 12-5 所示的 Options 选项设置对话框；分别勾选 Initial 复选框和 Cluster 复选框；单击 Continue 按钮返回主界面。

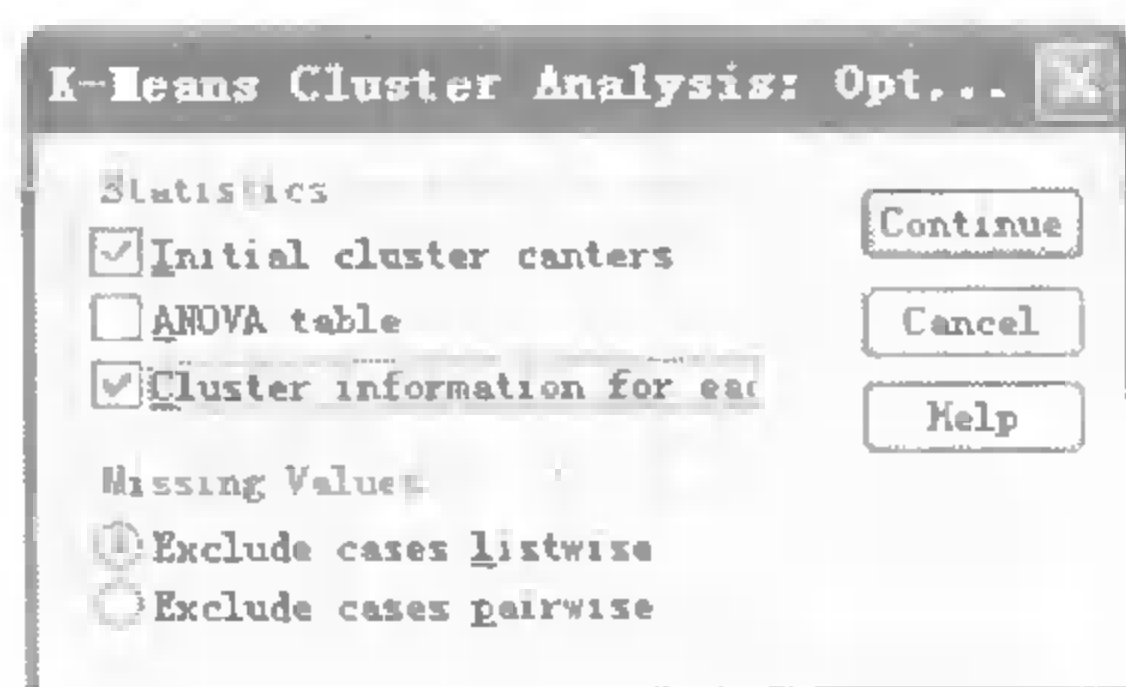


图 12-5 Options 选项设置

(1) Statistics 栏，在此选择输出哪些统计量，可选项有如下 3 个。

- Initial cluster centers 复选框，输出初始的类中心。
- ANOVA table 复选框，输出方差分析表，包括对每个聚类变量的 F 检验。如果所有观测最终被归为一个类别，则不输出任何方差分析表。
- Cluster information for each case 复选框，输出对每个观测的详细分类信息，包括它的所属类别、到所属类中心的距离等。

(2) Missing Values 栏，在此指定处理缺失值的方式，有两种选择。

- Excludes cases listwise: 当同时选入多个变量进行分析时，只要某个变量含有缺失值，就在所有分析过程中将相应的观测记录剔除。
- Excludes cases pairwise: 只有当所有聚类变量都取缺失值时，才将相应的观测从分析中剔除，否则将根据所有非缺失变量的取值把它分配到距离最近的一类中去。

#### 12.2.4 案例的结果分析

单击图 12-2 中 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 12-6~图 12-8 所示。

初始聚类中心			
	聚类		
	1	2	3
外貌	7	6	3
学术能力	10	9	8
讨人喜欢	3	10	0
自信程度	5	9	1
精明	0	10	1
诚实	10	10	0
推销能力	0	10	0
经验	0	10	10
积极性	2	10	0
抱负	2	10	0
理解能力	0	10	0
潜力	0	10	0
交际能力	0	10	0
适应性	0	10	10

迭代历史记录 <sup>a</sup>			
	聚类中心内的更改		
迭代	1	2	3
1	10 721	9 162	5 292
2	734	685	000
3	000	000	000

<sup>a</sup> 由于聚类中心内没有改动或改动较小而达到收敛。任何中心的最大绝对坐标更改为 000。当前迭代为 3。初始中心间的最小距离为 18 815。

图 12-6 初始聚类中心和迭代历史记录

聚类成员			
案例号	编号	聚类	距离
1	1	2	7.595
2	2	2	5.245
3	3	2	5.839

图 12-7 单个观测的归类输出

最终聚类中心			
	聚类		
	1	2	3
外貌	5	8	5
学术能力	7	7	8
讨人喜欢	5	7	3
自信程度	6	9	2
精明	4	8	2
诚实	8	9	2
推销能力	2	7	1
经验	2	5	9
积极性	3	8	1
抱负	4	8	0
理解能力	4	8	2
潜力	4	8	1
交际能力	5	7	1
适应性	3	8	2

最终聚类中心间的距离			
聚类	1	2	3
1		12.713	13.986
2	12.713		21.010
3	13.986	21.010	

每个聚类中的案例数	
聚类	1
	2
	3
有效	48 000
缺失	000

图 12-8 最终聚类中心及各类的简单信息

(1) 初始聚类中心。当没有指定从其他数据集或数据文件读取初始聚类中心时，SPSS 按照如下方法从当前数据集选取初始聚类中心：先拿前  $n$ （聚类个数）个没有缺失值的记录作为开始的聚类中心；逐个扫描余下的记录，若某个记录（记为  $x_1$ ）与开始聚类中心的最小距离小于开始聚类中心之间的最小距离（记为  $x_2$  和  $x_3$  之间的距离）时，就用  $x_1$  取代  $x_2$ 、 $x_3$  之中距离  $x_1$  最小的一个；如此对原始数据扫描完一遍后，就得到了初始的聚类中心。

本例的初始聚类中心如图 12-6 中的“初始聚类中心”表格所示。

(2) 迭代过程。快速聚类的迭代过程会由于聚类中心没有改动或改动较小而达到收敛，本例就是这种情况，由于第 3 次迭代后类中心没有变化导致迭代终止，如图 12-6 中的“迭代历史记录”所示。

(3) 每个观测的归类。如图 12-7 所示，“聚类成员”表格给出了部分单个观测的最终分类信息，包括它的标识变量（编号）、所属类别和距离类中心的距离。

(4) 最终聚类中心。如图 12-8 所示，“最终聚类中心”表格给出了对最终得到的 3 个聚类中心的统计信息。

类别 1 的各项得分都很平均，没有得分很高或很低的项目，且诚实一项的得分最高，鉴于此可以把这类人员归结为脚踏实地型的应聘者，适合于让他们担任财务、物流等方面的岗位。类别 2 的各项得分都较高，只在经验一项上的得分最低，可以推测这类人员多是初出象牙塔的高材生，除了经验，在学术、理解、潜力等方面都有不错的表现，故而把此类归结为激情进取型的应聘者，他们更适合于研发、销售等方面的工作；类别 3 的得分显得有些两级分化，在学术、经验、适应性方面比另外两类应聘者都要优秀，但其他方面的表现相当欠佳，比如外貌打扮、与人交流等，此类应聘者可能是“怪才”、“专才”或者在某些方面经验丰富的“老手”，他们适合于研发、攻关等工作，且在管理上应给予其较多的自由度，这样更能发挥他们的强项。

(5) 聚类结果的其他描述。如图 12-8 所示，从“最终聚类中心间的距离”表格看，这 3 个类别之间的距离都比较远，说明它能够较好地对这些应聘者进行分类和描述。

“每个聚类中的案例数”表格给出了不同种类的人数统计信息，可见“专才”毕竟还是少数，其他两类人才的数量旗鼓相当，这也是比较符合实际的结论。

## 12.3 分层聚类

聚类分析的方法有很多种，除了上节介绍的快速聚类外，比较常用的就是分层聚类。无论哪种聚类方法，其聚类原则都是把距离最近的或最相似的样本聚为一类。

### 12.3.1 分层聚类简介

根据分析过程的不同，分层聚类又分为凝聚法和分解法两种方向相反的聚类方法。

所谓分解法，指聚类开始时先把所有个体（观测量或变量）视为一个大类，然后根据距离或相似性原则逐层分解，直到参与聚类的每个个体自成一类为止。所谓凝聚法，指聚类开始时把参与聚类的每个个体（观测量或变量）视为单独的一类，然后根据两类之间的距离或相似性原则逐步合并，直到成为一个单独的大类为止。

分层聚类过程能实现系统聚类（Q 型聚类或 R 型聚类），而对于系统聚类，无需用户事先确定聚类个数，系统会自动将所有观测纳入计算过程，还可选择执行不同的聚类算法。

通常情况下在进行聚类之前，应先用 Proximities 过程对原始变量做一些诸如标准化的预处理，并计算它们之间的相似性测度或距离测度，然后再用 Cluster 过程对转换后的数据进行聚类分析。在 SPSS 的分层聚类过程里可以同时设置和执行 Proximities 和 Cluster 这两个过程，输出的统计量能帮助用户确定最好的分类结果，而且通过 Cluster 过程的 Plot 选项还能输出两种统计图（Dendrogram 树形图和 Icicle 冰柱图），直观地对聚类结果进行分析。

SPSS 的 Hierarchical Cluster 过程中，分析变量的类型可以为定量的、二分类的或者计数型的，而且变量的取值范围非常重要，它可以极大地影响分类结果。如果多个变量所采用的度量方式不同，应该设置在聚类前做一定的标准化处理。

### 12.3.2 问题描述和数据准备

某汽车制造商想评估当前的市场状况，并判断时下最有竞争力的车型，他们收集了关于多种车型在售价、物理特性等方面的数据。本节就通过分层聚类的方法，对这些车型进行归类和描述，最终给出关于各车型竞争力的建议。案例数据均摘录自 SPSS 自带的 Demo 文件“car\_sales.sav”，所用数据文件为“汽车销售样本数据.sav”，数据格式如图 12-9 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	manufact	String	13	0	制造厂商	None	None	7	Left	Nominal
2	model	String	17	0	车型	None	None	10	Left	Nominal
3	type	Numeric	11	0	车种	{0 轿车;	None	8	Right	Ordinal
4	sales	Numeric	11	3	销量(*1000)	None	None	8	Right	Scale
5	price	Numeric	11	3	价格(*1000)	None	None	8	Right	Scale
6	engine_s	Numeric	11	1	引擎型号	None	None	8	Right	Scale
7	horsepow	Numeric	11	0	马力	None	None	8	Right	Scale
8	wheelbas	Numeric	11	1	轴距	None	None	8	Right	Scale
9	width	Numeric	11	1	宽度	None	None	8	Right	Scale
10	length	Numeric	11	1	长度	None	None	8	Right	Scale
11	curb_wgt	Numeric	11	3	车重	None	None	8	Right	Scale
12	fuel_cap	Numeric	11	1	储油量	None	None	8	Right	Scale
13	mpg	Numeric	11	0	用油效率	None	None	8	Right	Scale



图 12-9 汽车销售数据的格式





本例我们选择了车种为轿车且销量大于 100 000 辆的 11 种车型进行聚类分析。

### 12.3.3 SPSS 分层聚类的设置

依次单击菜单 Analyze→Classify→Hierarchical Cluster... 执行分层聚类分析过程，其主设置面板如图 12-10 所示，在此指定分析变量、聚类方式等参数。

#### 1. 主面板的设置

在变量列表中选中从价格到用油效率的 9 个变量，单击从上至下第一个  按钮，将其作为分析变量选入 Variable(s) 列表框；在变量列表中单击选中车型变量，单击从上至下第二个  按钮，将其作为标签变量选入 Label Cases 选框；在 Cluster 栏单击选中 Cases 单选框；在 Display 栏分别勾选 Statistics 复选框和 Plots 复选框。

- ① Variables 列表框，用于从变量列表选入待分析的聚类变量。
- ② Label Cases 选框，用于选入标签变量，在结果中标识观测记录。
- ③ Cluster 栏，指定聚类分析的类型，有如下两个可选项。
  -  Cases 单选框，表示进行对观测记录的聚类，即 Q 型聚类。
  -  Variables 单选框，表示进行对观测变量的聚类，即 R 型聚类。
- ④ Display 栏，指定聚类分析输出哪些内容，有两个选择。
  -  Statistics 统计量复选框，输出距离矩阵（或相似矩阵）、最终分类信息等。
  -  Plots 图形复选框，输出反映聚类过程的树形图、冰状图等。

#### 2. 聚类方法的设置

在图 12-10 中单击 Method 按钮，弹出如图 12-11 所示的算法参数设置对话框。单击 Cluster 下拉列表，选中 Between-groups linkage 项；单击 Interval 下拉列表，选中 Squared Euclidean distance 项；单击 Standardized 下拉列表，选中 Z scores 项。单击 Continue 按钮返回主界面。

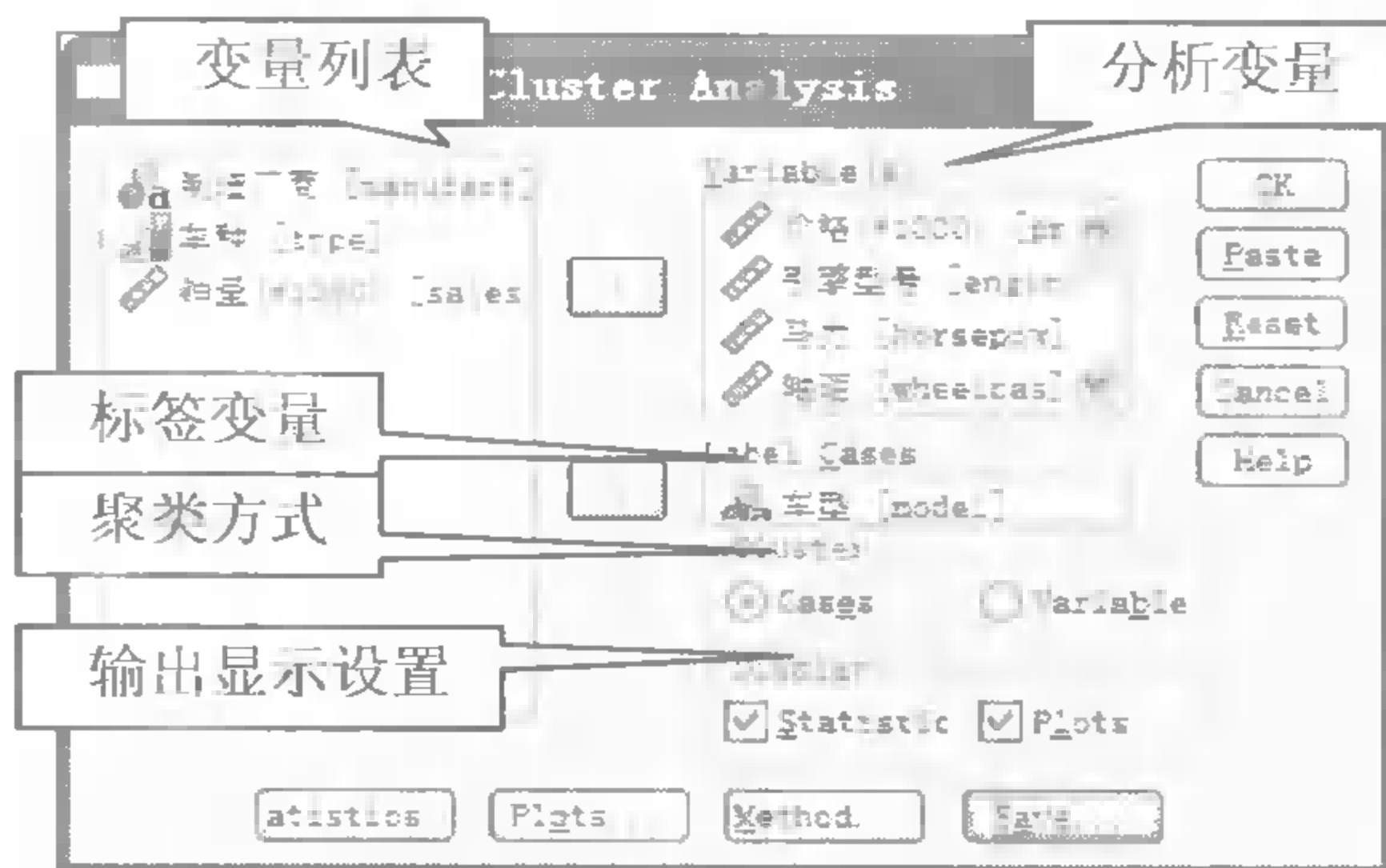


图 12-10 分层聚类的主设置面板

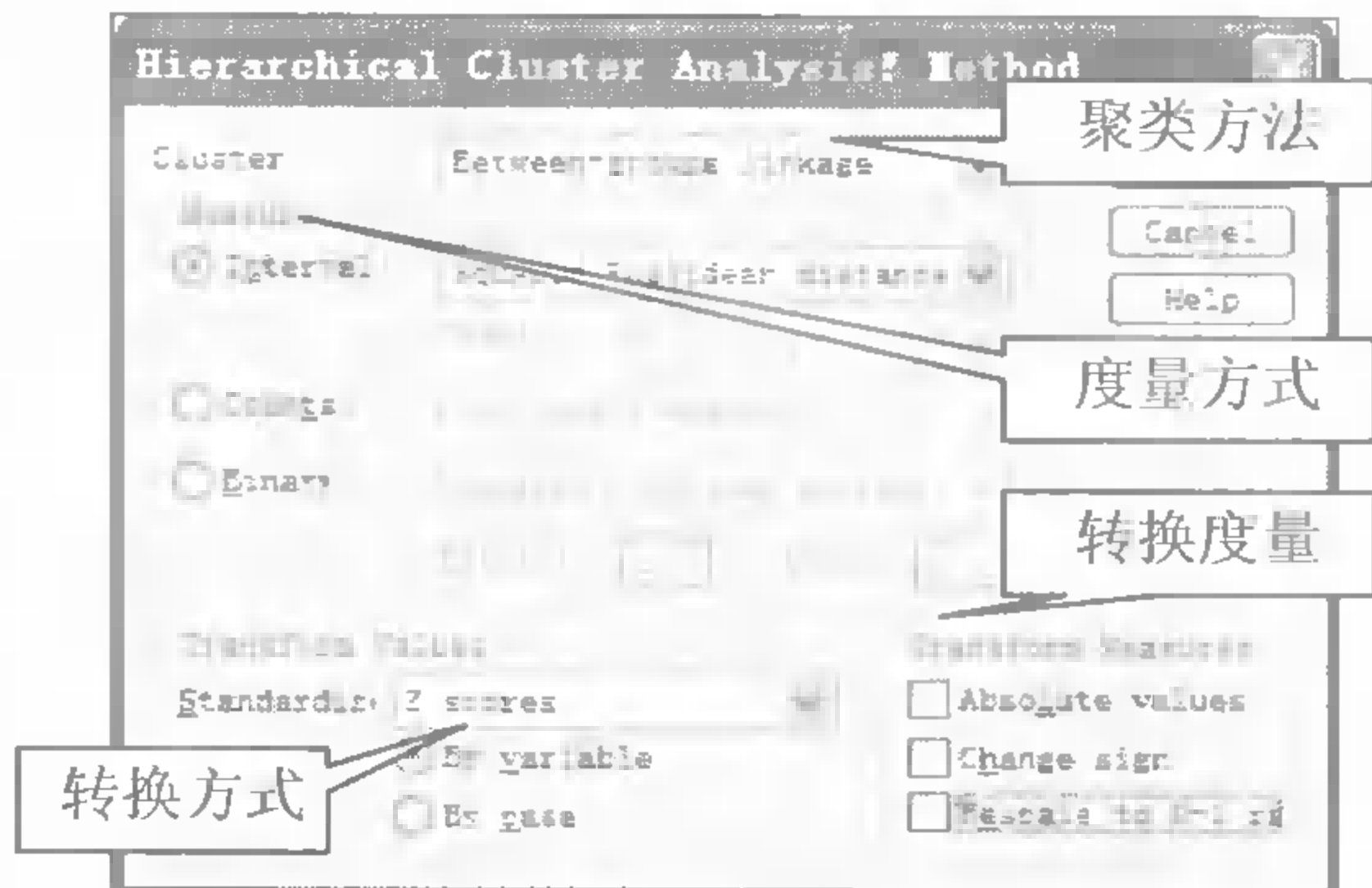




图 12-11 分层聚类的 Method 设置

- ① Cluster 下拉列表，用于指定分层聚类的方法，在此 SPSS 提供了如下 7 种选择。
  -  Between-groups linkage 类间连接法，它合并两类的依据，是使这两个类别里所有两两配对观测的平均距离达到最小，配对的两个观测要属于不同的类。
  -  Within-groups linkage 类内连接法，它合并两类的依据是使两个类别合并后的新类中，所有两两配对观测的平均距离达到最小。



- Nearest neighbor 最近邻法，该方法首先合并最近的或最相似的两个观测，然后用两个类别中的最近点之间的距离代表两个类之间的距离。
- Furthest neighbor 最远邻法，该方法首先合并最近的或最相似的两个观测，然后用两个类别中的最远点之间的距离代表两个类之间的距离，也称之为完全连接法。
- Centroid clustering 重心法，此方法应该与欧氏距离平方一起使用，它先计算各个类别里所有变量的均值，再以这些均值之间的距离代表类别之间的距离。
- Median clustering 中间距离法，它先计算两个类之间所有配对观测的距离，然后取这些距离的中位数代表这两个类之间的距离。
- Ward's method，离差平方和法。

关于各方法中提到的距离计算方式，在第 12.1.2 节中已有介绍。

## ② Measure 子设置栏，指定计算距离的方式。

此栏的设置选项与第 10.4 节距离分析中的距离设置非常相似，它正是把图 10-10 和图 10-11 中的相关选项（分别对应于 Interval、Counts 和 Binary 的选项）综合在了一起，使用户在图 12-11 中可以自由选用分层聚类过程所采用的相似测度或不相似测度。

首先指定数据类型（Interval、Counts 或 Binary）；然后在所选数据类型后的下拉列表中指定相应的距离计算方法。各选项的具体含义请参考第 10.4.2 节对距离分析过程的设置。

③ 数据转换设置。Transform Values 转换数值栏设置对观测量（By variable）或变量（By case）进行标准化的参数，但对二元变量不可用；Standardized 下拉列表用于指定标准化的方法。Transform Measures 转换测度栏设置对距离测度的计算结果进行转换的方法。关于这两个子设置栏的设置内容和设置方法，请参考第 10.4.2 节对距离分析过程的设置。

## 3. 统计量设置

在图 12-10 中单击 Statistics 按钮，弹出如图 12-12 所示的统计量设置对话框。单击选中 Range of solutions 单选框，分别在 Minimum 和 Maximum 后输入“2”和“4”。单击 Continue 按钮返回主界面。

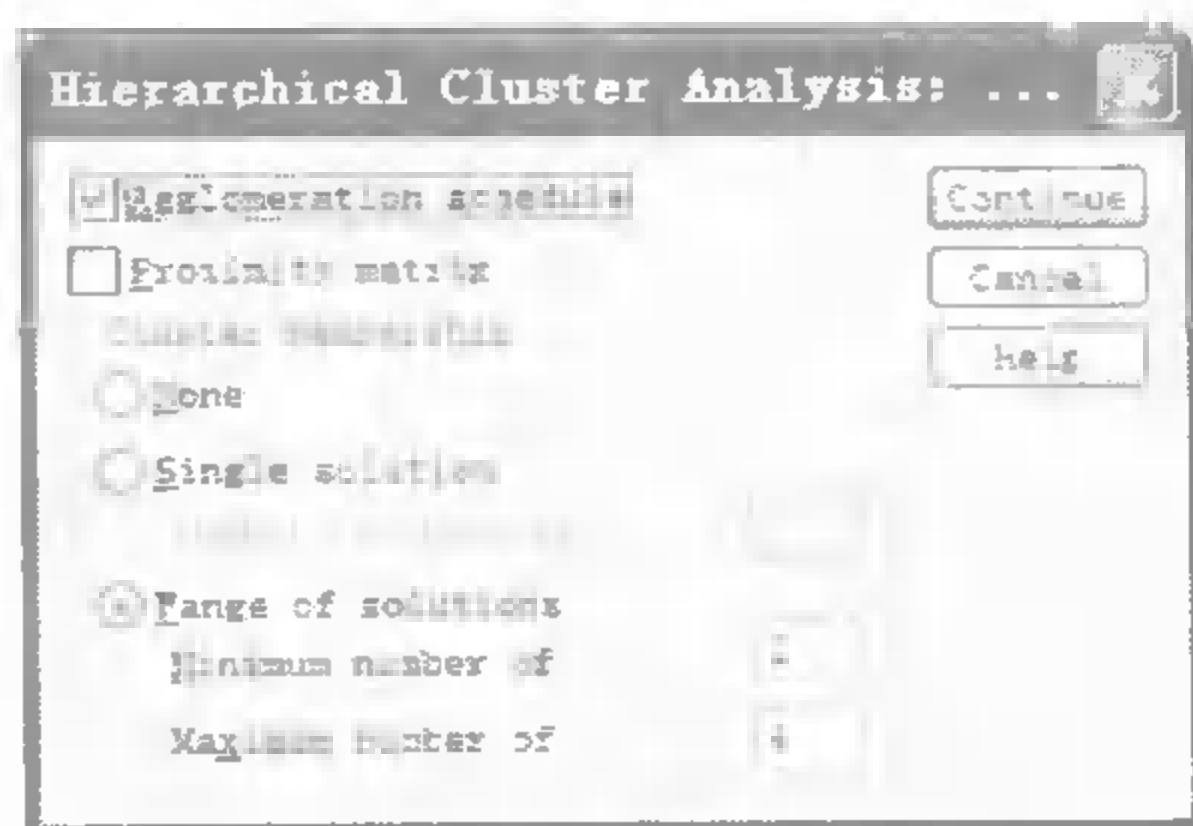


图 12-12 统计量设置

（1）Agglomeration schedule 复选框。输出聚类过程表，包括每一步被合并的类或观测量以及它们之间的距离和新生成的类等信息，根据此表能追踪整个聚类的合并过程，由于每次都是把最相近的两类聚为一类，据此可以查看哪些观测量之间的距离更近。

（2）Proximity Matrix 复选框。输出各项之间的距离矩阵或相似度矩阵，产生什么类型的矩阵（相似性矩阵或不相似性矩阵）取决于在 Method 设置面板中的 Measure 栏的选择。

（3）Cluster Membership 子设置栏。设置类成员表的输出格式，包括每个观测记录的最终分类结果，有如下 3 个可选项。

- None, 不显示类成员表, 系统默认选项。
- Single solution, 输出指定聚类个数时的类成员表, 在右侧的输入框指定聚类个数, 该数值必须大于 1, 且小于等于参与聚类的观测个数和变量个数。
- Range of solutions, 输出聚类个数在某个范围时的类成员表, 在 Minimum 输入框指定一个最小的聚类个数; 在 Maximum 输入框指定一个最大的聚类个数。

#### 4. 作图设置

在图 12-10 中单击 Plots 按钮, 弹出如图 12-13 所示的作图设置对话框。勾选 Dendrogram 复选框; 单击 Continue 按钮返回主界面。

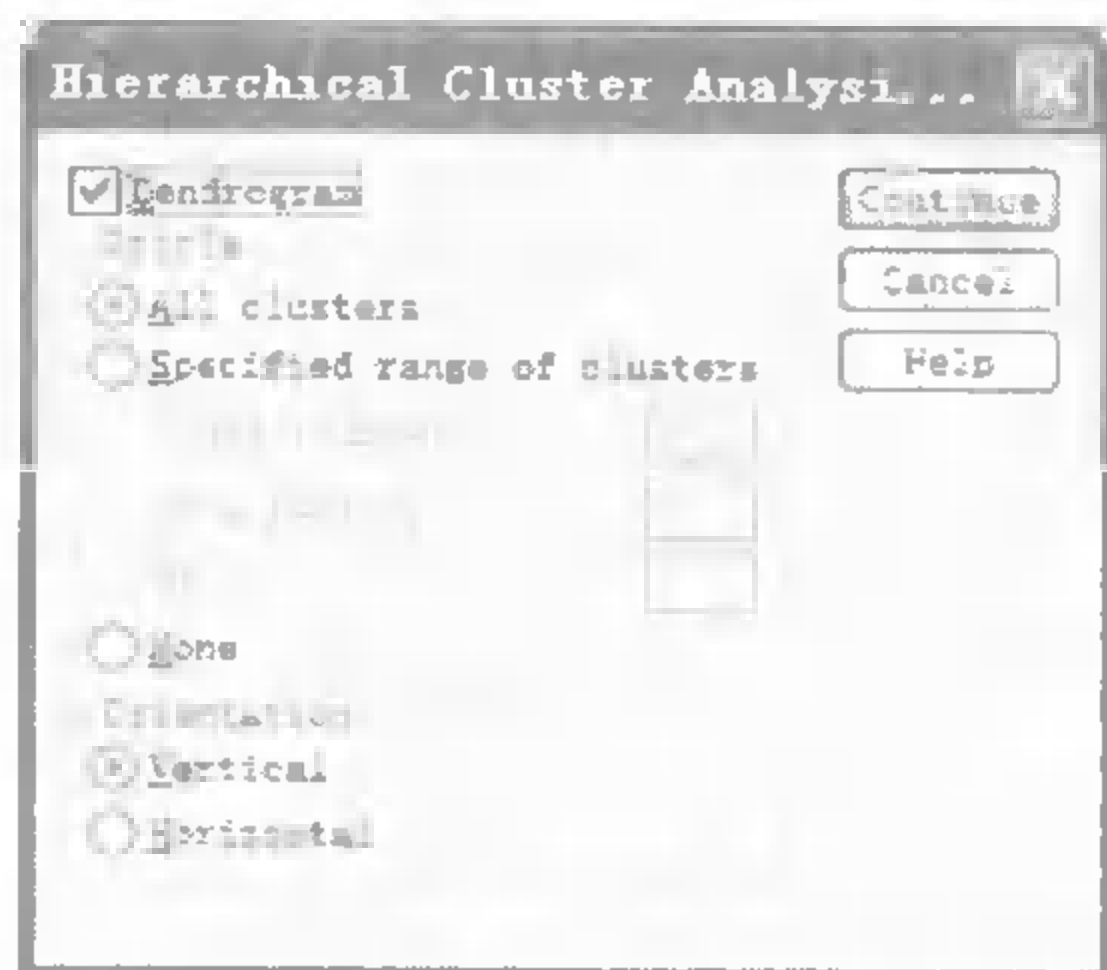


图 12-13 作图设置

Dendrogram 复选框用于输出树形图, 它系统地描绘了聚类的整个过程。

(1) Icicle 栏, 设置关于冰柱图的参数, 可选项有如下 3 个。

- All clusters 单选框, 表示把聚类的每一步都表现在图中, 如此可以查看聚类的全过程; 但如果参与聚类的观测量很多时, 容易使图形变得过大。
- Specified range of clusters 单选框, 指定要显示的聚类个数范围, 选中后需要设置如下的 3 个参数, 它们必须都是正整数。
  - Start 输入框, 指定要显示的起始聚类步数。
  - Stop 输入框, 指定要显示的终止聚类步数。
  - By 输入框, 指定要连续显示的两步聚类步骤之间的步数增量。
- None 单选框, 不生成冰柱图。

(2) Orientation 栏, 设置冰柱图的显示方向, Vertical 纵向显示; Horizontal 水平显示。

#### 5. 结果保存设置

在图 12-10 中单击 Save 按钮, 弹出如图 12-14 所示的保存设置对话框。单击选中 Range of solutions 单选框, 分别在 Minimum 和 Maximum 后输入“2”和“4”。单击 Continue 按钮返回主界面。

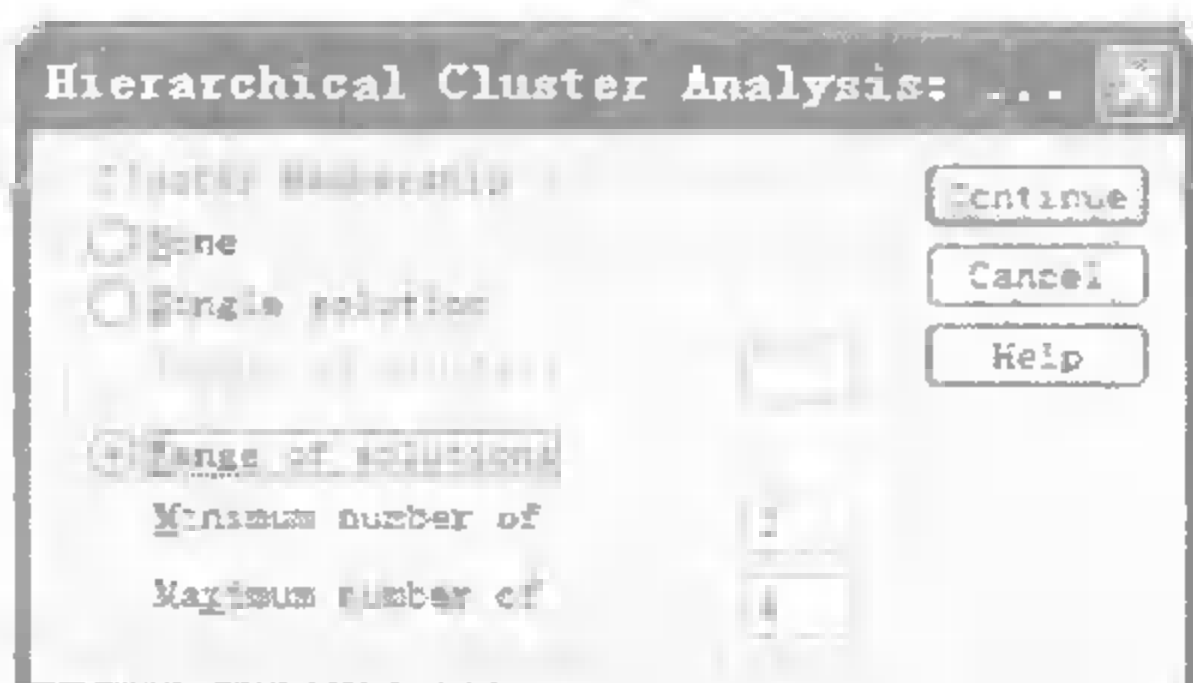


图 12-14 保存设置

在此对话框的 Cluster Membership 设置栏指定保存哪些分类结果，有如下 3 种选择。

① None 单选框，不保存任何结果，是默认选项。

② Single solution 单选框。保存指定聚类个数时的分类结果，在右侧的输入框指定聚类个数，该数值必须大于 1，且小于等于参与聚类的观测个数和变量个数。

③ Range of solutions 单选框。保存聚类个数在某个范围时的分类结果，在 Minimum 输入框指定一个最小的聚类个数；在 Maximum 输入框指定一个最大的聚类个数。

### 12.3.4 案例的结果分析

单击图 12-10 中的 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 12-15~图 12-18 所示。

聚类表						
阶	群集组合		系数	首次出现阶群集		下一阶
	群集 1	群集 2		群集 1	群集 2	
1	8	10	1.260	0	0	3
2	2	3	1.579	0	0	5
3	5	9	1.625	0	0	6
4	6	7	2.619	3	0	6
5	2	4	3.841	2	0	7
6	5	6	5.765	3	4	8
7	1	2	6.010	0	5	10
8	1	8	8.214	6	1	9
9	5	11	11.618	8	0	10
10	1	5	18.711	9	0	0

图 12-15 分层聚类的聚类过程

群集成员				
案例	4群集	3群集	2群集	
1 Cavalier	4	1	1	
2 Focus	1	1	1	
3 Civic	1	1	1	
4 Corolla	1	1	1	
5 Malibu	4	2	2	
6 Impala	2	2	2	
7 Taurus	2	2	2	
8 Accord	3	2	2	
9 Grand Am	2	2	2	
10 Camry	3	2	2	
11 Mustang	4	3	2	

	manufact	model	CLU4_1	CLU3_1	CLU2_1
1	Chevrolet	Cavalier	1	1	
2	Ford	Focus	1	1	
3	Honda	Civic	1	1	
4	Toyota	Corolla	1	1	
5	Chevrolet	Malibu	2	2	2
6	Chevrolet	Impala	2	2	2
7	Ford	Taurus	2	2	2
8	Honda	Accord	3	2	2
9	Pontiac	Grand Am	2	2	2
10	Toyota	Camry	3	2	2
11	Ford	Mustang	4	3	2

图 12-16 每个观测的聚类结果

垂直冰柱图																			
		案例																	
		11 Mustang	10 Camry	8 Accord	7 Taurus	6 Impala	9 Grand Am	5 Malibu	4 Corolla	3 Civic	2 Focus	1 Cavalier							
聚类步数	群集数	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
2		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
3		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
4		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
5		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
6		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
7		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
8		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
9		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X
10		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X	X

图 12-17 分层聚类的冰柱图

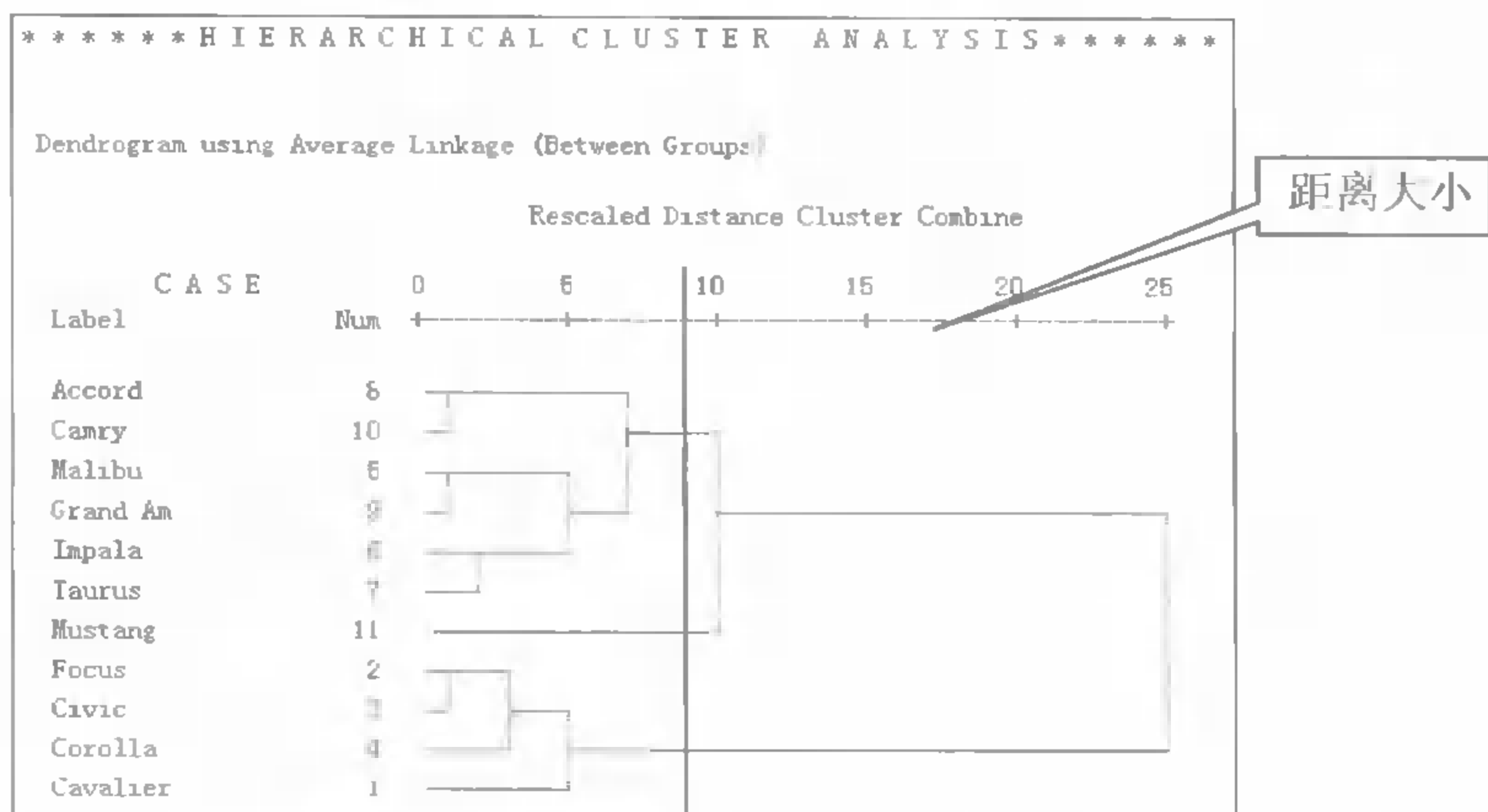


图 12-18 分层聚类的系统树状图

(1) 聚类过程。如图 12-15 所示,“聚类表”表格给出了把样本聚为一类的整个过程,下面以第 5 行为例来说明如何解读此表。“阶”列表示聚类的步骤数(第 5 步);在此步把第 2、4 类(由第 2、3 列给出)合并为一类;其中的第 2 类首次出现在聚类过程的第 2 步(由第 5 列给出),第 4 类是首次(由第 6 列给出)出现;最后一列的数字 7 表示此步的合并结果(仍记为第 2 类)在之后的第 7 步将会再次出现,并与其他类合并。最终,11 个观测经过 10 步聚为了一类。

(2) 各观测的聚类结果。如图 12-16 所示,“群集成员”表格是聚类个数分别为 2~4 个时的类成员表。

数据集窗口显示的是聚类个数分别为 2~4 个时当前数据集的保存信息,可见生成了 3 个新变量,分别保存聚类个数为 2、3、4 时的分类结果。

(3) 冰柱图。如图 12-17 所示的垂直冰柱图以柱状图的方式显示了聚类的整个过程,通过它能很快的发现某个观测所参与的所有聚类步骤。第 1 列显示聚类类别个数(也可以看作聚类的步骤数);行标题中写入观测量标识的列,“x”冰柱填满了整列,表示观测的初始状态;行标题为空的列,冰柱中的空格长度表示当前的聚类步骤数,并在此步骤把这列两边的两个类别聚为一类。

例如第 8 列的“7:Taurus”,它在第 4 步和右侧的观测“6:Impala”合并;然后在第 6 步又和右侧的由“9:Grand Am”和“5:Malibu”在第 3 步合并生成的类合并……如此类推,就得到了整个聚类过程以及每个观测在此过程中的位置。

(4) 聚类过程的系统树状图。如图 12-18 所示,系统树状图更直观地显示出了聚类的整个过程,当要分类的观测(或变量)个数较多时,该图比冰柱图显得清晰了许多;而且树状图还在其靠上的横轴方向给出了各类别之间的相对距离大小,所以建议多使用树状图进行分析。

根据树状图还可以方便地了解指定聚类个数的分类结果,例如当聚类个数为 3 时,在图中有且仅有 3 条横线的地方断开(如蓝色竖线位置所示);断开后,把那些仍然相连的观测分为 1 类,就得到了 3 个类别;第 5~9 个观测归为了一类,第 1~4 个观测归为了一类,第 11 个观测自成一类。



### 12.3.5 对聚类结果的进一步分析

通过分层聚类过程,最终得到了每个观测所属的类别标号,并将其存在了当前数据集中,如图 12-16 所示。下面以聚为 3 类(变量名为 Clu3\_1)为例,说明如何进行进一步的分析。

#### 1. 描述性分析

依次单击菜单“Analyze→Reports→OLAP Cubes...”执行 OLAP 制表过程,其主设置界



面如图 12-19 所示。在变量列表中选中从价格到用油效率的 9 个变量，单击从上至下第一个  按钮，将其作为汇总变量选入 Summary Variable(s) 列表框；在变量列表中单击选中 CLU3\_1（聚类个数为 3 时的分类结果）变量，单击从上至下第二个  按钮，将其作为分类变量选入 Grouping Variable(s) 列表框。单击 OK 按钮运行，会在 SPSS Viewer 窗口输出一个“OLAP Cubes”表格。

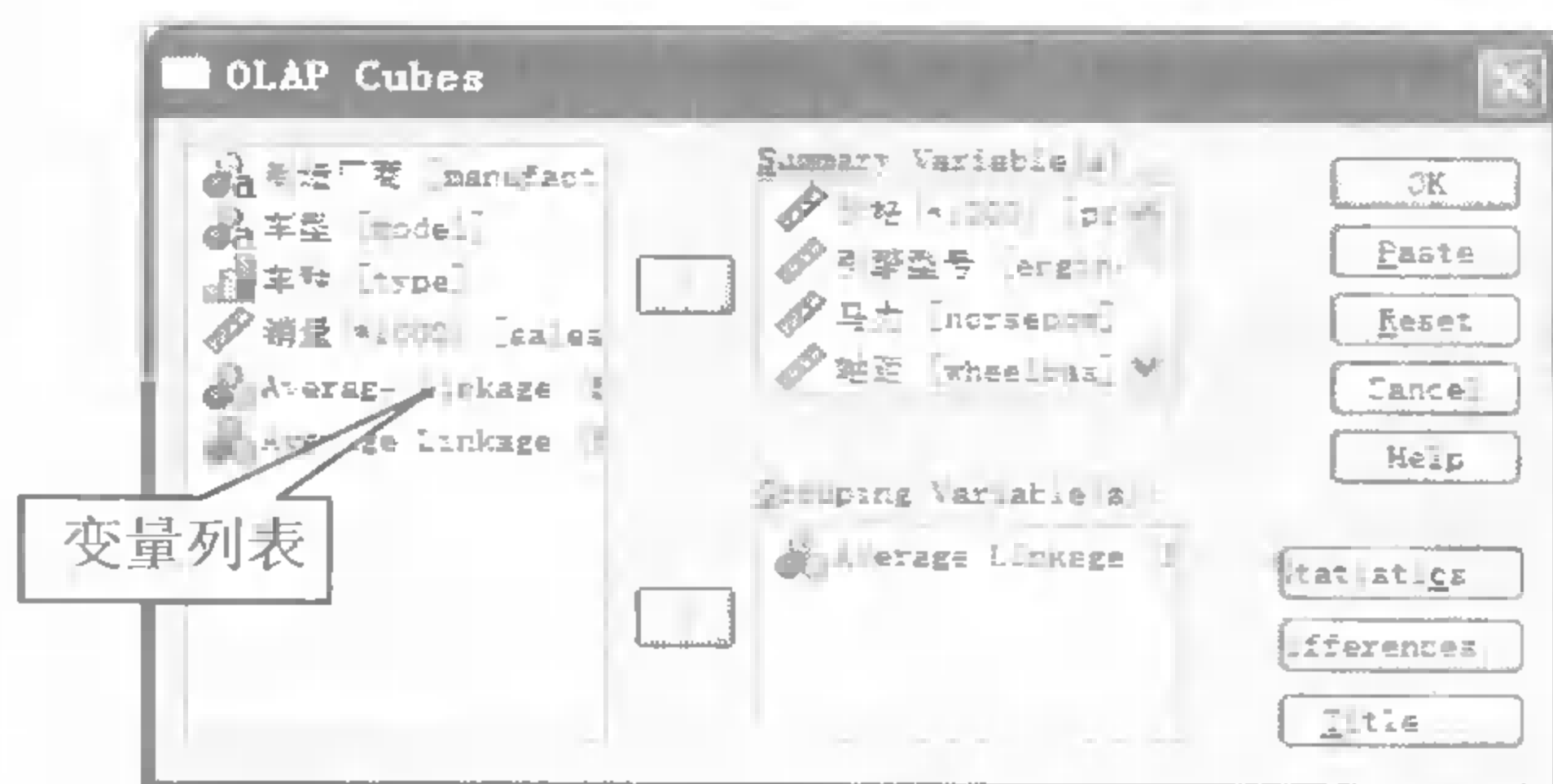



图 12-19 分层聚类结果的 OLAP 分析

在输出的 OLAP Cubes 上双击使其进入编辑状态，接着在表格中右击并选中“Pivoting Trays”菜单项，打开如图 12-20 所示的对话框，拖动图中标识了名称的  到指定位置，返回 SPSS Viewer 窗口，即可得到如图 12-21 所示的 OLAP Cubes。

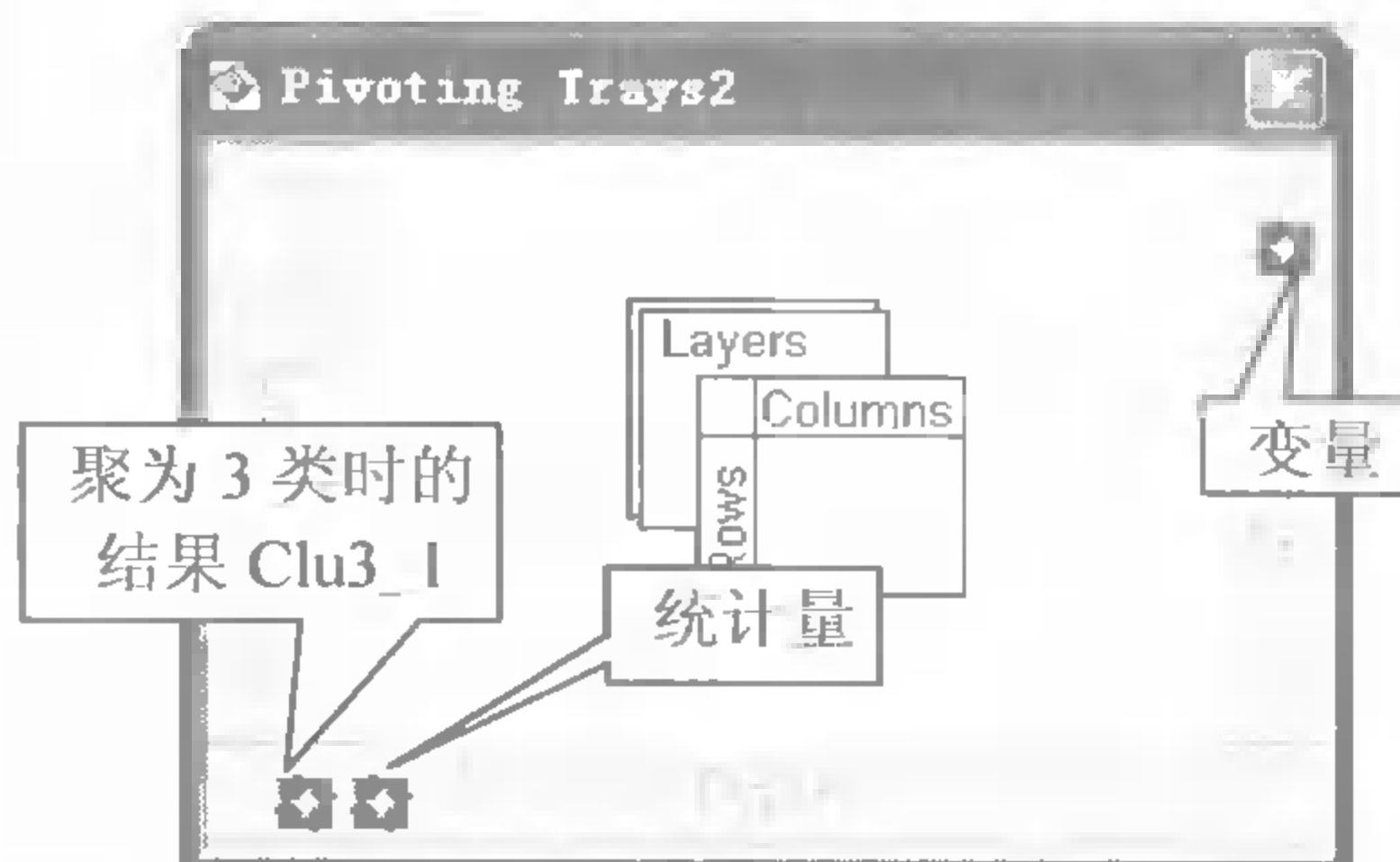


图 12-20 Pivoting Trays 的设置对话框

OLAP Cubes										
Clu3_1 聚类结果		价格 [price]	引擎型号 [engine]	马力 [horsepower]	轴距 [wheelbase]	宽度	长度	车重	储油量	用油效率
1	N	4	4	4	4	4	4	4	4	4
均值		12 892 00	1 900	112 00	101 825	67 150	176 200	2 499 75	13 150	30.50
标准差		414350	2582	6 093	32821	5260	3 1675	149881	9815	2.646
极小值		12 315	1.6	106	97.0	66.7	174.0	2.339	11.9	27
极大值		13 260	2.2	120	104.1	67.9	180.9	2.676	14.3	33
2	N	6	6	6	6	6	6	6	6	6
均值		17 649 67	2 900	158 00	107 517	71 033	191 933	3 138 17	16 467	25.83
标准差		1 576 602	5292	20 396	17 971	1 5629	5 5294	193 760	13 261	1 329
极小值		15 350	2.2	133	105.2	69.4	186.3	2.932	15.0	24
极大值		19 720	3.4	180	110.5	73.0	200.0	3.389	18.5	27
3	N	1	1	1	1	1	1	1	1	1
均值		21 560 00	3 800	190 00	101 300	73 100	183 200	3 203 00	15 700	24.00
标准差		21 560	3.8	190	101.3	73.1	183.2	3 203	15.7	24
极小值		21 560	3.8	190	101.3	73.1	183.2	3 203	15.7	24
极大值		21 560	3.8	190	101.3	73.1	183.2	3 203	15.7	24
总计	N	11	11	11	11	11	11	11	11	11
均值		16 275 09	2 518	144.18	104 882	69 809	185 418	2 911 91	15 191	27.36
标准差		3 130 362	7441	10 987	3 7381	2 4728	8 8458	364 211	1 9588	3 075
极小值		12 315	1.6	106	97.0	66.7	174.0	2.339	11.9	24
极大值		21 560	3.8	190	110.5	73.1	200.0	3.389	18.5	33

图 12-21 分层聚类结果的 OLAP Cubes

其中，第 1 类汽车的多项指标比第 2 类汽车都偏低，只有用油效率偏高，由此推断第 1 类汽车为低端经济型的；第 2 类汽车为中端实用型的；第 3 类只有 Mustang 这 1 种车型，它在体型与第 2 类车没有太大差异的情况下，拥有更足的马力、更高的售价，故把它归为高端车型。其中第 2 类汽车包括的车型最多，有如下 6 个：Malibu、Impala、Taurus、Accord、Grand Am 和 Camry，说明中端实用型的车此时比较受市场青睐，而且在这一领域的竞争也最为激烈。

## 2. 其他分析

描述性分析从均值、标准差等方面简单直观地研究了各类别之间的区别，若需要定量检验各类别间的差异是否在统计意义上显著，可以进一步作方差分析。另外，聚类结果还可以应用于判别分析、回归分析和因子分析等过程。

## 12.4 两阶段聚类分析

两阶段聚类分析 (TwoStep Cluster Analysis) 是一个执行探索性分析功能的过程，用它来揭示原始数据的自然分组或分类，它反映的是数据集内部而不是外观上的分类。

### 12.4.1 两阶段聚类简介

两阶段聚类分析过程假设各分析变量是相互独立的，连续变量服从正态分布，分类变量服从多项式分布；它使用的距离度量方式有欧式距离测度和似然距离测度。经验表明，参与分析的变量时常违反这些假设，但是两阶段聚类能够很好地适应由此造成的干扰。

两阶段聚类分析过程可以输出判别聚类个数所使用的准则 (AIC 或 BIC) 和最终聚类的描述性统计信息等，还能输出关于聚类频数的饼图、条形图及变量重要性图等。

两阶段聚类分析具有如下 4 个比较突出的优势特征。

- 能够同时处理分类变量和连续变量。
- 通过指定的判别准则，自动选择最优的聚类个数。
- 可以有效地分析大样本数据。
- 用户可以自行设置用于计算的内存容量。

### 1. 两阶段聚类的基本步骤

在进行两阶段聚类分析之前，建议使用 SPSS 的如下过程检验需要满足的一些假设条件。Bivariate Correlations 过程检验两个连续变量之间的独立性；Crosstabs 过程检验两个分类变量之间的独立性；Means 过程检验连续变量和分类变量之间的独立性；Explore 过程检验连续变量的正态性；Chi-square (卡方) 过程检验分类变量是否服从多项式分布。

两阶段聚类分析的计算过程分为如下两步。

第 1 步，构建聚类特征树 (Cluster Features Tree, CFT)。开始时，把某个观测量放在树的根节点处，它记录有该观测量的变量信息；然后使用指定的距离测度作为相似性依据，使每个后续观测量根据它与已有节点的相似性，放到最相似的节点中；如果没有找到与某个观测量足够相似的节点，就为它形成一个新节点。

第 2 步，使用凝聚聚类法对聚类特征树的节点进行分组。它通过比较 Schwarz-Bayesian 信息准则 (Schwarz's Bayesian Information Criterion, BIC) 或 Akaike 信息准则 (Akaike Information Criterion, AIC)，确定最优的聚类个数。

## 2. 两阶段聚类中的概念解析

(1) 聚类特征树 (Cluster Features Tree, CFT)。在两阶段聚类的第 1 步由观测量之间的距离所确定的分类结构，每个类别形成一个节点，属于此类的观测量就是该节点的树叶，再由树叶的不断增加构成树枝。

(2) AIC 准则和 BIC 准则。在两阶段聚类的第 2 步中用到的两个确定聚类个数的判断依据。

(3) 调谐算法 (Tuning the Algorithm)。两阶段聚类过程既可以自动聚类，也可以人为控制聚类过程。在人为控制时，需要用户指定参数，在这里称作调谐 (tuning)，参数指定了，聚类特征树的规模就基本确定了。

(4) 噪声处理 (Noise Handling)。构建 CFT 树时，如果指定了聚类个数等参数，而观测量又很多的话，有可能发生 CFT 树长满而不能再长的情况。那些没有长在树上的观测量就称为噪声，可以调整参数重新计算以让 CFT 树容纳更多的观测；也可以直接把它们归入某个类中或者直接丢掉。

(5) 局外者 (Outlier)。对 CFT 树进行噪声处理后，被丢掉的观测量称之为局外者，它们单独构成一类，但不计入聚类结果的类别个数中。

### 12.4.2 问题描述和数据准备

本节仍使用关于汽车制造商的例子 (第 12.3 节分层聚类曾使用)，目的是通过对多种车型在售价、物理特性等数据的聚类分析对这些车型进行归类 and 描述。在分层聚类中，选取了对车种和销量进行特定限制的车型进行分析；在两阶段聚类中，将对所有数据进行分析。

所用数据均摘录自 SPSS 自带的 Demo 文件 “car\_sales.sav”。所用数据文件为 “汽车销售初始数据.sav”，数据格式如图 12-22 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	manufact	String	13		制造厂商			7	Left	Nominal
2	model	String	17		车型			10	Left	Nominal
3	type	Numeric	11	0	车种	轿车	None	8	Right	Ordinal
4	sales	Numeric	11	3	销量 (*1000)	None	None	8	Right	Scale
5	price	Numeric	11	3	价格 (*1000)	None	None	8	Right	Scale
6	engine_s	Numeric	11	1	引擎型号	None	None	8	Right	Scale
7	horsepow	Numeric	11	0	马力	None	None	8	Right	Scale
8	wheelbas	Numeric	11	1	轴距	None	None	8	Right	Scale
9	width	Numeric	11	1	宽度	None	None	8	Right	Scale
10	length	Numeric	11	1	长度	None	None	8	Right	Scale
11	curb_wgt	Numeric	11	3	车重	None	None	8	Right	Scale
12	fuel_cap	Numeric	11	1	储油量	None	None	8	Right	Scale
13	mpg	Numeric	11	0	用油效率	None	None	8	Right	Scale

图 12-22 汽车销售数据的格式

### 12.4.3 SPSS 两阶段聚类的设置

依次单击菜单 “Analyze→Classify→TwoStep Cluster...” 打开两阶段聚类过程的主设置面

板，如图 12-23 所示，在此指定分析变量、聚类个数等内容。

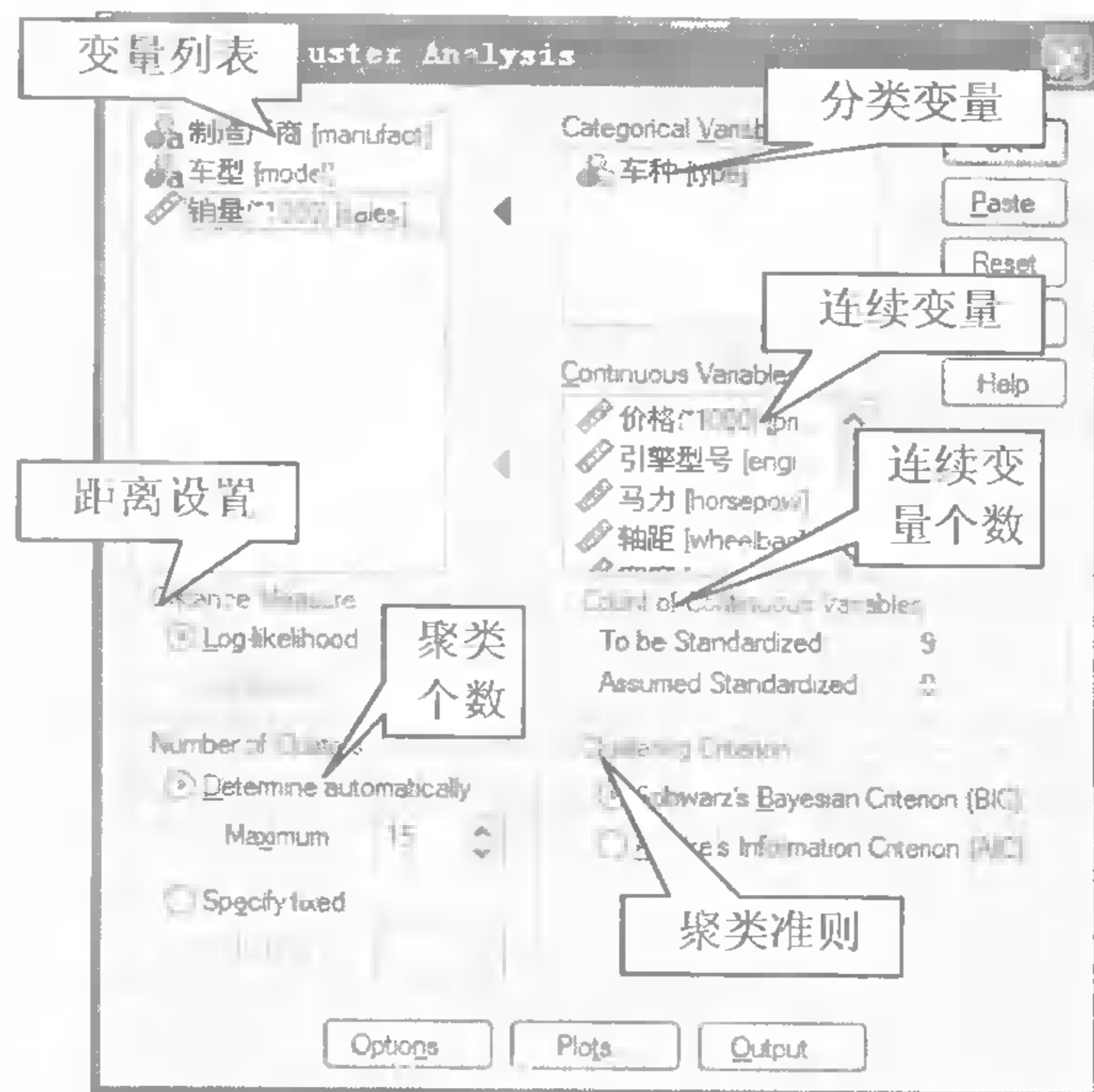










图 12-23 两阶段聚类的主设置面板

### 1. 主面板的设置

在变量列表中选中从价格到用油效率的 9 个变量，单击从上至下第二个  按钮，将其作为连续变量选入 Continuous Variables 列表框；在变量列表中单击选中车种变量，单击从上至下第一个  按钮，将其作为分类变量选入 Categorical Variables 列表框。

下面详细介绍各设置选项的含义。

- ① Categorical Variables 列表框，用于从变量列表选入待分析的分类变量。
- ② Continuous Variables 列表框，用于从变量列表选入待分析的连续变量。
- ③ Count of Continuous Variables 栏，显示对连续变量进行标准化处理的个数统计信息。
  -  To be Standardized 后显示的是要进行标准化处理的连续变量个数。
  -  Assumed Standardized 后显示的是不需要进行标准化处理的连续变量个数，即已经假定它们为标准化后的数据了。对于一个变量是否要被标准化处理，可以在 Options 子面板进行设置。
- ④ Distance Measure 栏，指定度量两个类别之间的距离定义，有如下两个选择。
  -  Log-likelihood 对数似然距离，它假设连续变量服从正态分布，分类变量服从多项式分布，且所有变量都相互独立。
  -  Euclidean 欧氏距离，只有当所有变量都为连续型时才可用。
- ⑤ Number of Clusters 栏，设置如何确定聚类个数，SPSS 给出了两种方法。
  -  Determine automatically 自动确定，按照在 Clustering Criterion 栏设置的准则，由系统自动确定最优的聚类个数；同时，还可以在 Maximum 输入框指定聚类个数的最大值。
  -  Specify fixed 固定值，由用户在 Number 输入框指定聚类个数。
- ⑥ Clustering Criterion 栏



指定自动聚类算法中确定最优聚类个数的准则，从如下两种方法中选择一个：Bayesian Information Criterion (BIC) 贝叶斯准则，Akaike Information Criterion (AIC) 准则。

## 2. 输出设置

在图 12-23 中单击 Output 按钮，弹出如图 12-24 所示的输出设置对话框。分别勾选 Statistics 栏下的 3 个复选框；勾选 Create 复选框；单击 Continue 按钮返回主界面。

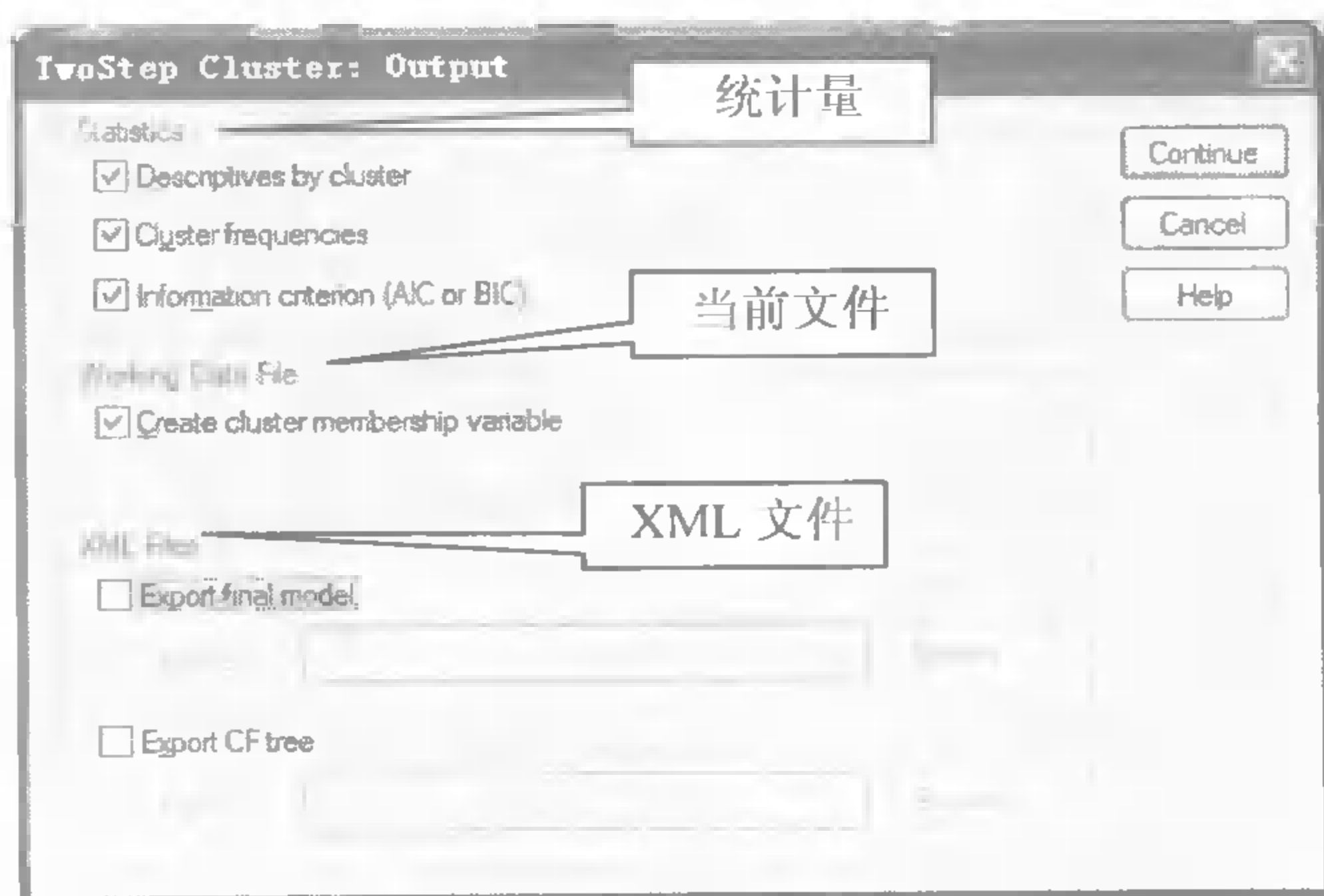


图 12-24 两阶段聚类的输出设置

- ① Statistics 栏，选择输出哪些统计量，它们都将以表格形式输出。
  - Descriptives by cluster 复选框，输出两个表格。一个显示最终类别里连续变量的均值和标准差；另一个显示最终类别里分类变量的频数统计信息。
  - Cluster frequencies 复选框，输出最终分类的观测个数统计表。
  - Information criterion (AIC or BIC) 复选框，输出包含 AIC 或 BIC 的统计表。此表格只有在主面板选择了自动确定聚类个数时才会输出；如果在主面板指定了固定的聚类个数，则此选项不做任何输出。

Descriptive statistics 和 Cluster frequencies 只对最终的聚类结果进行显示；Information criterion 对聚类过程的某个范围进行显示。

- ② Working Data File 栏，指定在当前数据集中保存哪些结果。
  - Create cluster membership variable 复选框，表示保存最终的聚类结果。变量名格式为 TSC\_n，其中 n 为一个正整数，表示在当前过程中执行保存操作的内部运行顺序。
- ③ XML Files 栏，设置以 XML 格式输出最终的聚类模型和 CF 树。
  - Export final model 复选框，将最终聚类模型输出到指定的 XML 文件，它可以直接应用于 SmartScore、SPSS Server 等工具，执行对其他数据打分等操作。单击后面的 Browse 按钮指定文件路径。
  - Export CF tree 复选框，保存当前聚类决策树的状态，以后可以用新数据对它进行修改。单击后面的 Browse 按钮指定文件路径。

## 3. 选项设置

在图 12-23 中单击 Options 按钮，弹出如图 12-25 所示的选项设置对话框。单击 **Advanced** 按

钮展开高级设置选项：单击 Continue 按钮返回主界面。



图 12-25 选项设置面板

下面详细介绍各设置选项的含义。默认情况下，有些高级选项是隐藏的，需要单击按钮 **Advanced >** 展开，随后按钮状态变为 **<= Advanced**，再次单击会将高级选项隐藏起来。

#### ① Outlier Treatment 栏，设置对异常值的处理方式。

如果勾选 **Use noise handling** 复选框，当 CF 树长满后，把稀疏节点合并为一个单独的“噪声”节点，然后重新执行 CF 树生长过程；判定某个节点是稀疏的，只需它的观测个数比最大节点的指定比例还低即可，**Percentage** 输入框用于指定这个比例的临界值，默认值为 25%；CF 树再次长满后，需判断“噪声”节点能否仍留在 CF 树上，如果不能就把它删除掉。

如果不勾选 **Use noise handling** 复选框，当 CF 树长满后，若存在过多异常观测 (Outlier)，就使用更宽松的临界条件重新生长 CF 树；在最终的聚类结果里，那些仍不能归入某个类别的观测就标记为异常观测，它们自成一类并以“-1”作为类号，但不记入聚类的类别个数里。

#### ② Memory Allocation 栏，设置聚类过程所能使用的最内存数。

在后面的输入框指定一个大于等于 4 的数字 (单位 MB)，默认值为 64MB。如果聚类过程使用的内存超过了这个限制，系统将使用硬盘来暂存那些内存无法容纳的信息。由于两阶段聚类经常处理较大的数据集，所以设置此项限制是很有必要的。

#### ③ Standardization of Continuous Variables 栏，设置对连续变量的标准化规则。

默认情况下，所有连续变量都自动选入 **To be Standardized** 列表框，表示对它们都实施标准化处理；对于那些已经是或者假设是标准化数据的变量，将其选入左侧的 **Assumed Standardized** 列表框，表示对其不再进行标准化处理，如此可节省聚类算法的运行时间。

#### ④ CF Tree Tuning Criteria 栏，设置决策树的调整准则，有如下 3 个特定参数。

- **Initial Distance Change Threshold**，指定 CF 树生长的初始临界值，默认值为 0。当向某个节点插入 1 个观测时，如果产生的紧密度小于该临界值，这个叶节点就不被分支；否则，这个节点就被分支成新的节点。
- **Maximum Branches (per leaf node)**，指定单个节点能拥有的最多子节点个数，默认值为 8。

Maximum Tree Depth, 指定 CF 树的最大深度, 默认值为 3。

⑤ Maximum Number of Nodes Possible 栏, 显示当前过程可能产生的最大节点个数。

该数值是基于  $(b^{d+1} - 1)/(b - 1)$  计算的, 其中  $b$  代表最大的分支个数;  $d$  代表 CF 树的最大深度。注意: 一个节点最少要使用 16 个字节的空間, 太大的 CF 树会极大地耗费系统资源, 影响聚类过程的效率。

⑥ Cluster Model Update 栏, 设置关于引入和更新旧模型的选项。

在图 12-24 中的 XML Files 栏可以设置把模型保存到指定的文件。

在此勾选 Import CF Tree XML file 复选框后, 单击 Browse 按钮指定一个 XML 格式的 CF 树文件, 聚类过程将使用当前数据更新导入的旧模型。对于旧模型的更新, 需注意如下 3 点。

- ① 旧模型的参数设置随它一同导入, 当前窗口的设置将被忽略。
- ② 在当前的主设置面板中所选入的分析变量的显示顺序, 需要和导入的旧模型生成时这些变量的显示顺序保持一致。
- ③ 更新的模型仅用于对当前数据的分析, 原始的 XML 文件内容不会被更改, 除非把当前模型输出至与其同名的文件里。

#### 4. 作图设置

在图 12-23 中单击 Plots 按钮, 弹出如图 12-26 所示的作图设置对话框。勾选 Cluster pie chart 复选框; 勾选 Rank 复选框; 单击 Continue 按钮返回主界面。

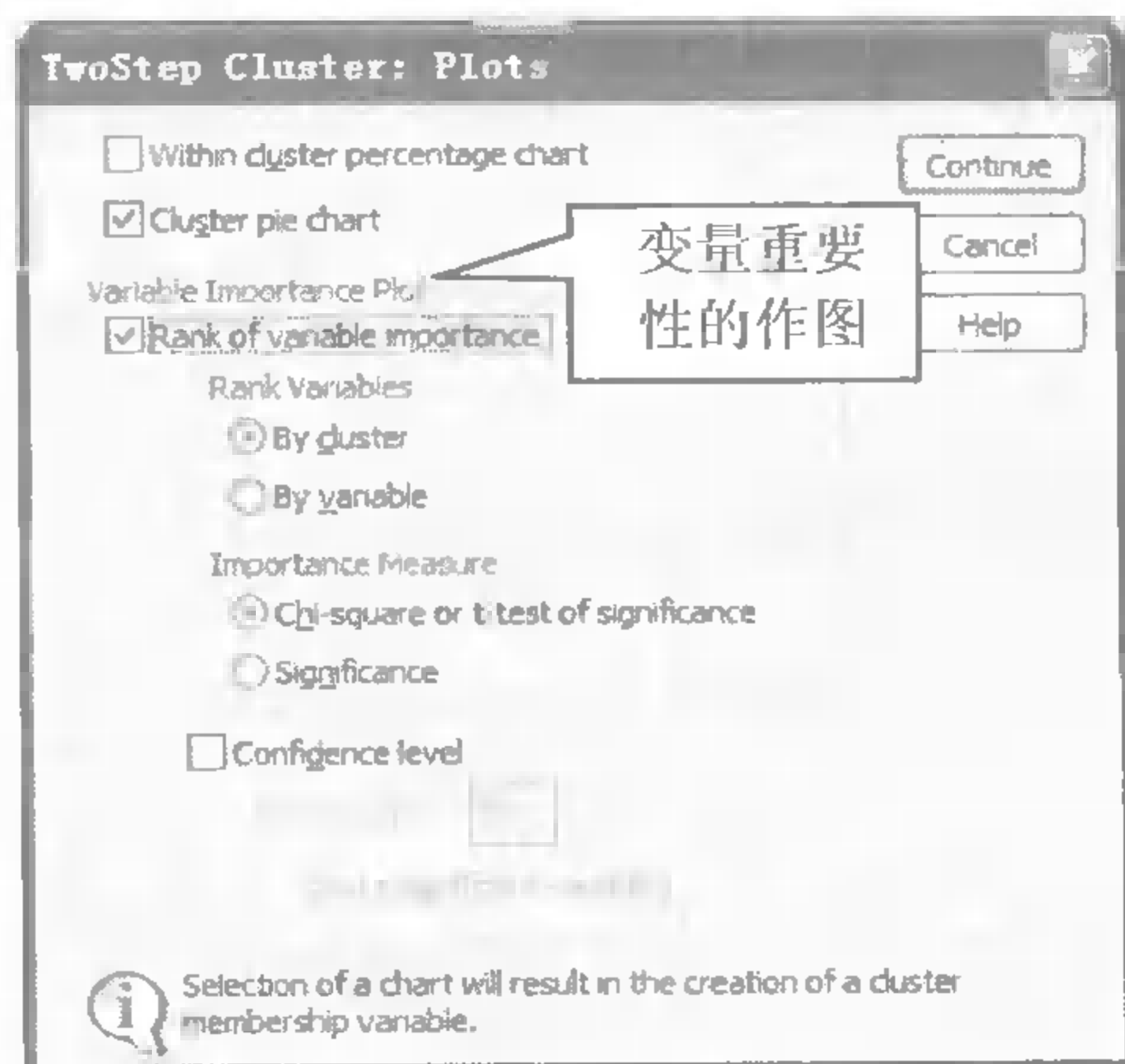


图 12-26 作图设置面板

(1) Within cluster percentage chart 复选框, 输出每个变量在各类内部的变化情况图。

对每个分类变量输出一个复合条形图, 描绘结果类别中关于这个变量的频数统计信息; 对每个连续变量输出一个误差条形图, 描绘各个结果类别的误差差异。

(2) Cluster pie chart 复选框, 输出聚类饼图, 描绘结果类别的频数统计信息。

(3) Variable Importance Plot 栏, 设置关于变量重要性的图形参数。

输出结果将以每个变量的重要性自动排序, 勾选 Rank of Variable Importance 复选框后, 激活下面的设置选项。

① Rank Variables 栏, 设置变量重要图的显示方式, 有两个选择。

- ② By cluster, 表示对每个变量进行作图, 由此比较单个变量在各类中的重要性。

- By variable, 表示对每个类别进行作图, 由此比较单个类别中各变量的重要程度。
- ② Importance Measure 栏, 设置用什么统计量来度量变量的重要性, 有两个选择。
  - Chi-square or t-test of significance 单选框, 对分类变量采用 Pearson 卡方统计量, 对连续变量采用 t 统计量。
  - Significance 单选框, 对连续变量采用由均值检验给出的 1-p 值, 对分类变量采用各水平的期望频数。
- ③ Confidence level 置信水平, 它用于对变量的类内分布与整体分布相等的假设检验。Percentage 输入框指定一个大于等于 50 小于 100 的比例作为置信水平, 默认值为 95%。

如果在上面选择了 By variable 选项或者 Significance 选项, 置信水平的取值将会在重要性图中显示为一条垂直的直线。勾选 Omit insignificant variables 复选框, 表示那些置信度低于指定水平的变量不在变量重要性图中显示。

12.4.4 案例的结果分析

在图 12-23 中, 单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如下。

(1) 两阶段聚类的过程。如图 12-27 所示, “自动聚类”表格给出了两阶段聚类的整个聚类过程。第 1 列表示聚类的步骤数, 其他 4 列为判断最佳聚类个数的统计量。第 2 列的 BIC 统计量取较小值时, 代表了较好的模型; 但有时 BIC 值会随着类数的增加而减少, 从而很难在 BIC 值和聚类个数 (代表了模型复杂性) 之间达到较好的平衡; 这时, 建议最优模型应拥有最大的 BIC 变化比率 (Ratio of BIC Changes) 和最大的距离度量比率 (Ratio of Distance Measures)。

自动聚类				
聚类数	Schwarz 的 Bayesian 准则 (BIC)	BIC 变化 <sup>a</sup>	BIC 变化的比率 <sup>b</sup>	距离度量的比率 <sup>c</sup>
1	1214.377			
2	974.051	-240.326	1.000	1.829
3	885.924	-88.128	367	2.190
4	897.559	11.635	-0.48	1.368
5	931.760	34.201	142	1.036
6	968.073	36.313	-151	1.576
7	1026.000	57.927	-241	1.083
8	1086.815	60.815	-253	1.687
9	1161.740	74.926	-312	1.020
10	1237.063	75.323	-313	1.239
11	1316.271	79.207	-330	1.046
12	1396.192	79.921	-333	1.075
13	1477.199	81.008	-337	1.076
14	1559.230	82.030	-341	1.301
15	1644.366	85.136	-354	1.044

a 变化是相对于表中先前的聚类个数而言。

b 变化的比率与两个聚类解的变化相关。

c 距离度量的比率以当前聚类的个数为基础而不是先前的聚类个数为基础。

聚类分布			
聚类	N	组合 %	总计 %
1	62	40.3%	39.5%
2	39	25.7%	24.8%
3	51	33.6%	32.5%
组合	152	100.0%	96.9%
已排除的案例	5		3.1%
总计	157		100.0%

车种				
聚类	轿车		卡车	
	频率	百分比	频率	百分比
1	61	54.5%	1	2.5%
2	0	0%	39	97.5%
3	51	45.5%	0	0%
组合	112	100.0%	40	100.0%

图 12-27 聚类过程输出和聚类分布输出

本例综合考虑了 BIC 值和距离度量比率来确定最优聚类个数, 建议聚为 3 类时的模型最优, 如图中以蓝色线框标识的一行所示。

(2) 聚类结果的基本统计信息。如图 12-27 所示, “聚类分布”表格给出了最终聚得的 3 类中的观测频数及排除的异常观测的频数, 可见第 1 类的观测数最多, 然后是第 3、2 类。“车



种”表给出了分类变量（车种）的频数统计信息，可见第 1、3 两类中几乎全部为轿车，而第 2 类中全部都是卡车。


（3）连续变量的基本统计信息。随后输出的“质心”表格给出了关于连续变量的均值、标准差等统计信息。在输出的“质心”表格上双击使其进入编辑状态，接着在表格中右击并选中“Pivoting Trays”菜单项，打开图 12-28 中的“Pivoting Trays”对话框，拖动其中标识了名称的到指定位置，关闭此对话框返回 SPSS Viewer 窗口，即可得到图 12-28 中的“质心”表格。



图 12-28 连续变量的基本统计信息

观察发现，第 1 类车的价格便宜，体积、限重和马力都较小，是 3 类车中的低端车型；第 2 类车的价格居中，体积、限重和马力均有明显提高，就是用油效率偏低了点，是 3 类车中较为实用的车型；第 3 类车的价格昂贵，体积、限重与第 2 类相比没有明显变化，而马力提高了许多，用油效率也在平均水平，是 3 类车中比较高端的车型。

（4）频数统计饼图。如图 12-29 所示，是关于聚类结果的频数统计饼图。作此图的数据也在图 12-27 中的“聚类分布”表格给出，饼图中不包含异常类别的信息。

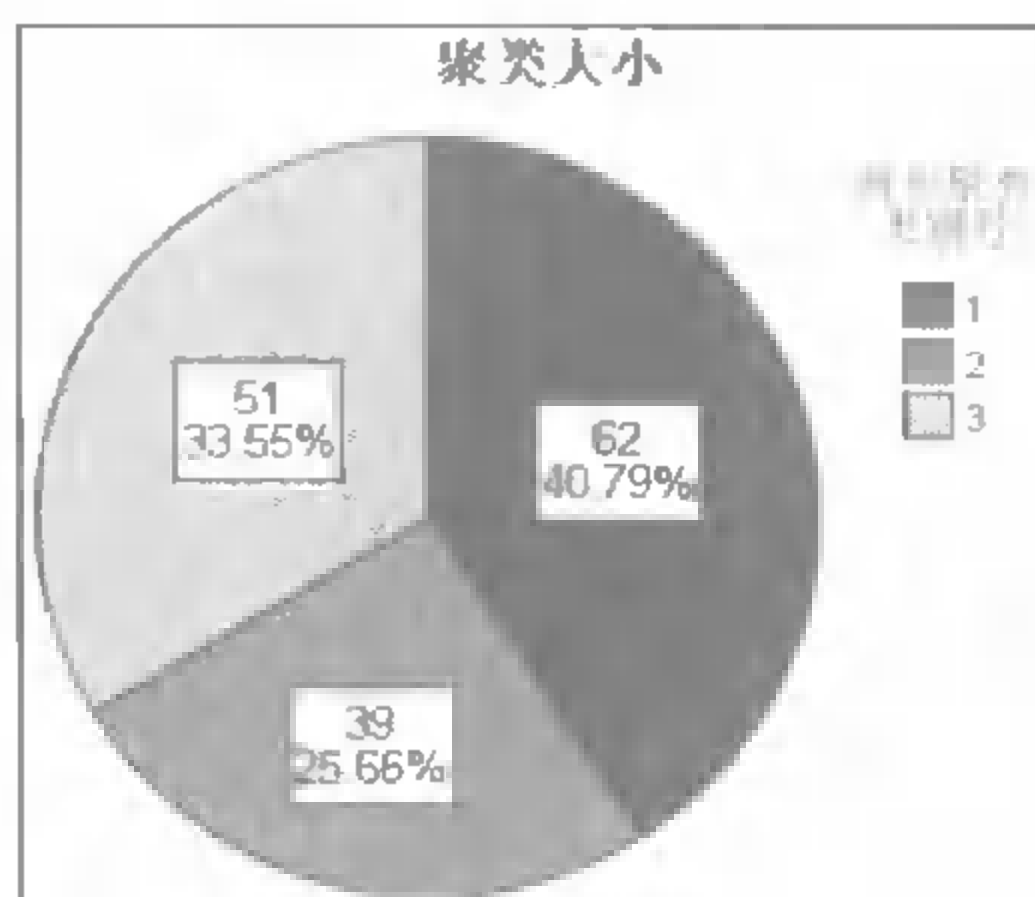


图 12-29 聚类类别的频数统计饼图

（5）均值比较图。如图 12-30 所示，是价格、用油效率两个变量的包含 95% 置信区间的均值比较图，其他变量的输出与此类似。此作图数据也在图 12-28 中的“质心”表格给出，图形直观的反映了和表格数据相同的信息，即第 1、2、3 类车的价格依次增高，第 2 类的用油效率最低，第 3 类的用油效率居中。

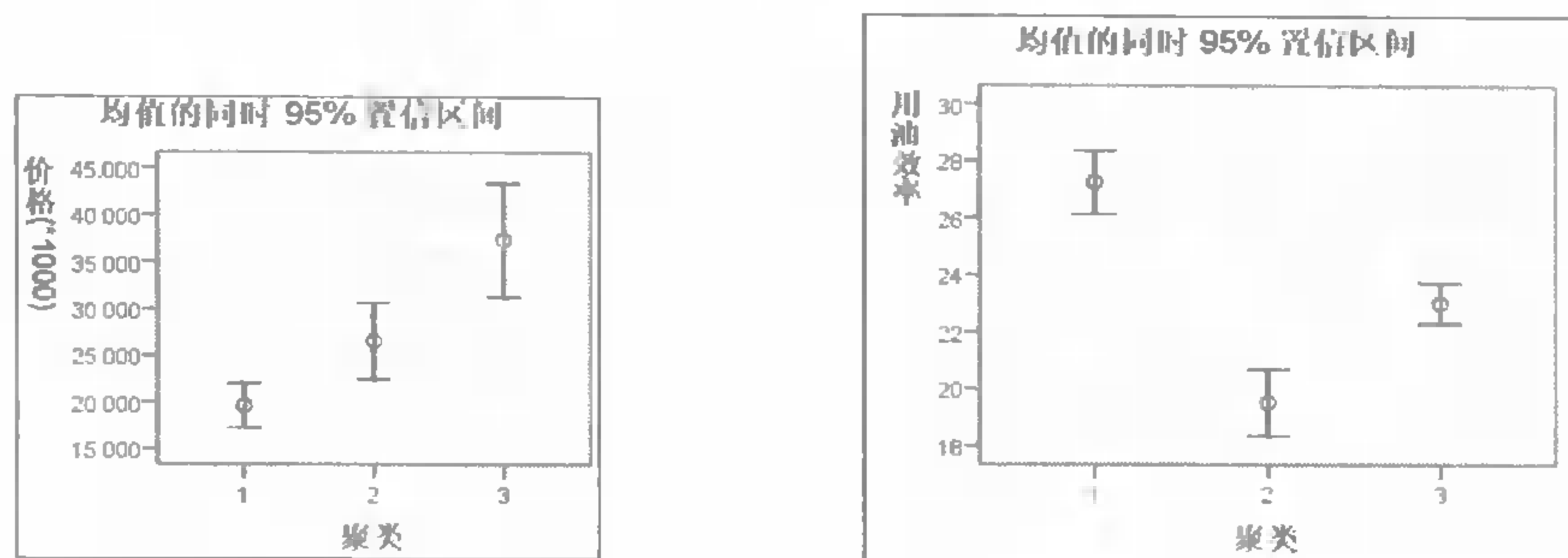


图 12-30 变量在聚类类别中的均值比较图

(6) 以变量区分的重要性图形。如图 12-31 所示，是车种、价格两个变量在各分类中的重要性图形，其他连续变量的输出都与价格的图形类似。可见，分类变量“车种”采用的是卡方检验，它在第 2 个类别里的卡方值最大（重要性最大），因为这类里全是编码为 1 的卡车（轿车编码为 0）；连续变量“价格”采用的是  $t$  检验，它在第 1 类中的重要性最大，小于零的  $t$  值说明价格要低于均值，所以第 1 类的价格是显著地小，第 3 类的价格是显著地大。

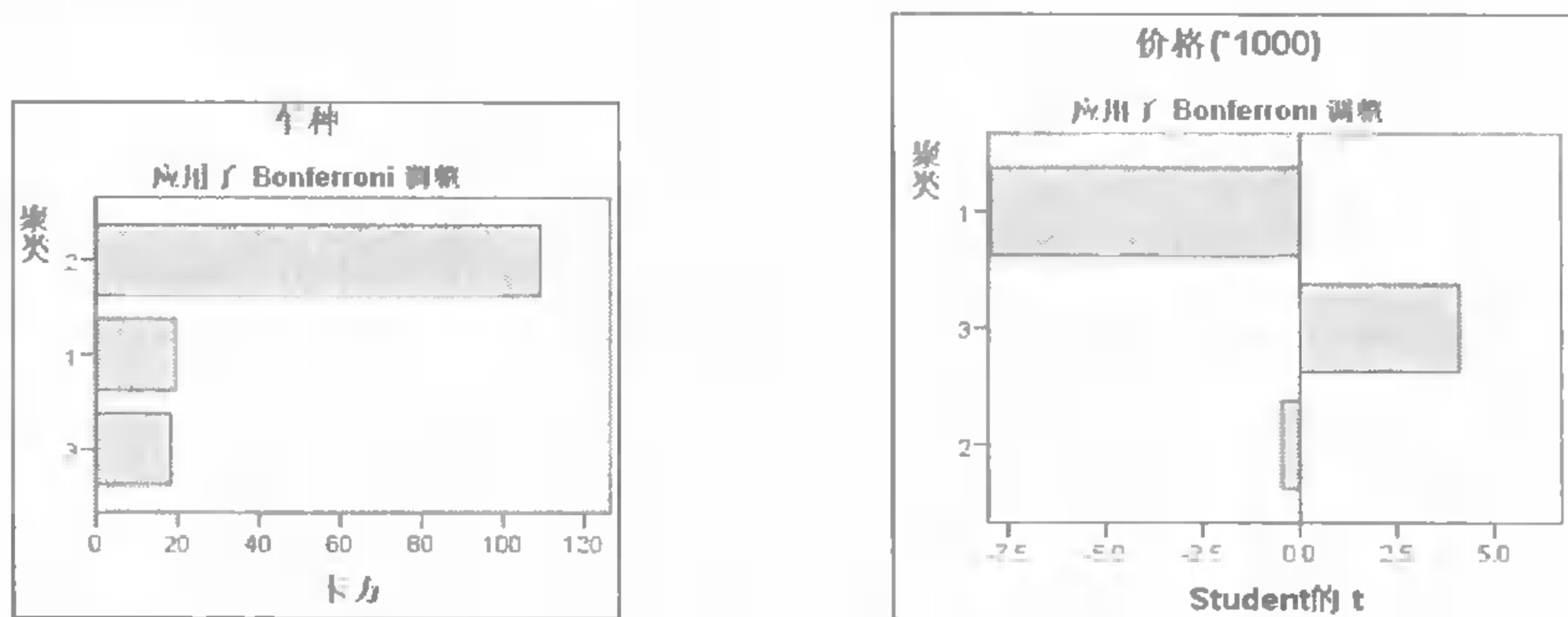


图 12-31 变量在不同类别中的重要性图形输出

(7) 以分类区分的重要性图形。如果在图 12-26 中的 Rank Variables 栏选择的是 By variable 选项，将会输出反映单个类别里不同变量的重要性的图形，如图 12-32 所示。

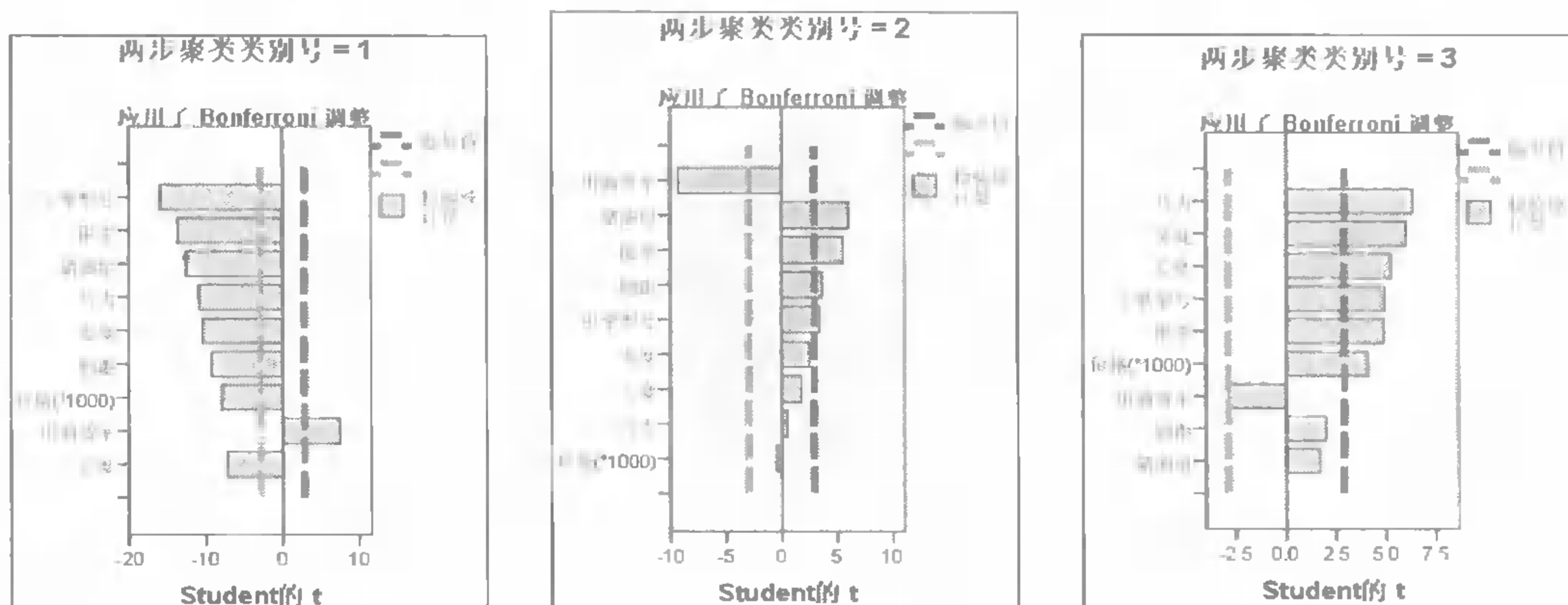


图 12-32 同一类别里不同变量的重要性图形输出

这里给出的只是关于连续变量的  $t$  检验，小于零的  $t$  值说明价格要低于均值，反之亦然；图中的虚线是指定的置信水平，超出此范围的差异认为是显著的。在第 1 类中，所有的连续

变量都较为重要，都对这个类别的形成做出很大贡献，其中用油效率要高于均值，其他指标都要小于均值；在第2类中，宽度、长度、马力和价格4个变量的重要性不显著，它们对这个类别的形成贡献不大；在第3类中，轴距和储油量的重要性不太显著。

## 12.5 一般判别分析

判别分析要处理的问题是事先有一个已知分类的数据集，研究者要把和这个数据集性质相同但未知分类的数据归入已知的分类。例如医生根据各种化验结果、身体特征等判断患者所得的疾病类型；教练根据运动员的体形、运动成绩、生理指标、心理素质等判断他是否应该被选入运动队继续培养。

判别分析有很多方法，例如距离判别、非参数判别和逐步判别等。

### 12.5.1 判别分析的基本原理

判别分析过程基于对预测变量的线性组合，这些预测变量应该能够充分地体现各个类别之间的差异。判别分析从已经确定了观测所属类别的样本中拟合判别函数，再把判别函数应用于由相同观测变量所记录的新数据集，以判断新样本的类别归属。

设在  $P$  维空间中，有  $k$  个关于已知类别的总体  $G_1, G_2, \dots, G_k$ ；单个的观测样本记为  $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ ,  $i = 1, 2, \dots, n$ ，它属于且仅属于  $k$  个总体中的一个，这  $p$  个预测变量也叫做判别指标。判别分析所要解决的问题，就是确定这些观测  $x$  应该属于哪一个总体  $G$ 。

#### 1. 两种判别方法

(1) Bayes 判别。Bayes 判别是一种概率型的判别分析，分析过程开始时，它需要知道观测属于各个类别的先验概率，或者关于各个类别的分布密度；分析过程结束时，计算每个观测归属于某个类别的最大概率或最小错判损失，并以此分类。例如某个观测的判别得分为  $D$ ，则它属于第  $i$  个类别的概率为  $P(G_i | D) = P(D | G_i)P(G_i) / \sum P(D | G_i)P(G_i)$ ，其中  $P(G_i)$  为属于第  $i$  类的先验概率， $P(D | G_i)$  为在第  $i$  类中得  $D$  分的条件概率，而  $P(G_i | D)$  在第  $i$  类中得  $D$  分的后验概率；最后，把观测归入概率  $P(G_i | D)$  最大的类别中。

(2) Fisher 判别。Fisher 判别是一种依据方差分析原理建立的判别方法，它的基本思路就是投影。对  $P$  维空间中的点  $x_i = (x_{i1}, x_{i2}, \dots, x_{ip})$ ,  $i = 1, 2, \dots, n$ ，找一组线性函数  $y_m(x_i) = \sum_j c_j \times x_{ij}$ ,  $m = 1, 2, \dots, m$ ，一般有  $m < p$ ，用它们把  $P$  维空间中的观测点都转换为  $m$  维的，再在  $m$  维空间中对观测集进行分类。降维后的数据应最大限度地缩小同类中观测之间的差异，并最大限度地扩大不同类别观测之间的差异，如此才能获得较高的判别效率。在此采用方差分析的思想，依据使组间均方差与组内均方差之比最大的原则，选择最优的线性函数。

#### 2. 判别分析的一般步骤

执行判别分析过程的步骤，一般分为如下3个部分。

- (1) 依据已知类别的观测集建立一系列分类规则或判别规则。
- (2) 运用所建规则对分析样本、验证样本进行分类检验，得到各样本的判别准确率。
- (3) 选择拥有较高准确率的判别规则，应用于新样本的类别判断。

由以上步骤可见，判别分析过程的输出主要有分类规则和分类结果两个部分。

(1) 分类规则，主要包括典型判别函数 (Canonical Discriminant Function)、衡量预测变量与判别函数之间关系的结构矩阵 (Structure Matrix) 和 Fisher 线性分类函数 (Fisher Classification Function)。典型判别函数是基于 Bayes 判别思想建立的，主要用于考查各类别的观测之间的相关关系，要将其应用于大量的实践操作是不现实的，因为它需要计算关于被分类观测的各种概率，十分繁琐不利于操作。Fisher 线性分类函数则是针对每个类别分别建立的一组函数，它可以方便地应用于对新样本的分类预测。

(2) 分类结果，依据建立的分类规则对原始样本集重新进行分类，通过比较预测分类与原始分类，确定对初始样本的判别准确率。

### 3. 判别分析中的假设检验

如下的几个检验，均要求  $G_i \sim N_p(\bar{\mu}^{(i)}, \Sigma_i), i=1, 2, \dots, k$ ；并且各类的协方差矩阵相等。

(1) 判别函数的有效性检验。检验的原假设  $H_0$  是  $\mu_1^{(i)} = \mu_2^{(i)} = \dots = \mu_k^{(i)}, \mu_j^{(i)} = E_{G_j}(x_i), i=1, 2, \dots, p$ ，即检验所建判别函数能否把  $k$  个类别的样本显著地区分开来。此项检验采用的是威尔克斯  $\lambda$  统计量 (Wilks' lambda)，当原假设为真时，它服从于 Wilks 分布： $\lambda \sim \Lambda(m, n-p, p-1)$ ，这个分布也可以用卡方分布来近似。

(2) 判别指标的显著性检验。原假设  $H_{0i}: \mu_1^{(i)} = \mu_2^{(i)} = \dots = \mu_k^{(i)}, \mu_j^{(i)} = E_{G_j}(x_i), i=1, 2, \dots, p$ ，它逐个检验每个预测变量在各类别中的均值是否有显著差异。检验统计量也采用 Wilks' lambda，当原假设为真时，它服从自由度为  $(m-1, n-m-(p-1))$  的 F 分布，其中  $n$  表示样本容量。当检验结果表明有多个预测变量不显著时，建议考虑使用逐步判别法，它有如回归分析中的逐步回归。

(3) 协方差矩阵相等的 Box 检验。原假设  $H_0: \Sigma_1 = \Sigma_2 = \dots = \Sigma_k$ 。检验统计量采用的是 Box's M，当原假设为真时，该统计量近似服从 F 分布。

### 4. 聚类分析与判别分析的联系及区别

聚类分析和判别分析都是用于分类和预测的方法。判别分析需要从一个已知分类的样本集里总结出判别规则，它是一种有指导（有监督）的学习；聚类分析所面对的样本数据一般都是未知类别的，甚至连分成几类也不知道，它是一种无指导（无监督）的学习。

#### 12.5.2 问题描述和数据准备

研究者采集了 15 名胃病患者的 4 个生理指标，已知其中 14 名患者的病情为如下 3 种中的 1 种：胃癌、萎缩性胃炎和其他胃病，另 1 名患者的病情尚未确定。本节通过判别分析，利用已知病情的 14 名患者的信息建立对病情的判别规则，并对余下 1 名患者的病情加以推断。

所用数据文件为“胃病患者的测量数据.sav”，数据格式如图 12-33 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	gp	Numeric	2	0	类别	{1 胃癌患者}	None	7	Right	Scale
2	x1	Numeric	5	0	铜蓝蛋白	None	None	5	Right	Scale
3	x2	Numeric	5	0	蓝色反应	None	None	5	Right	Scale
4	x3	Numeric	4	0	尿吡啶乙酸	None	None	4	Right	Scale
5	x4	Numeric	4	0	中性硫化物	None	None	4	Right	Scale

图 12-33 胃病患者数据格式



### 12.5.3 判别分析的参数设置

依次单击菜单“Analyze→Classify→Discriminant...”执行判别分析过程，其主设置面板如图 12-34 所示，在此指定分析变量、变量选择方法等内容。

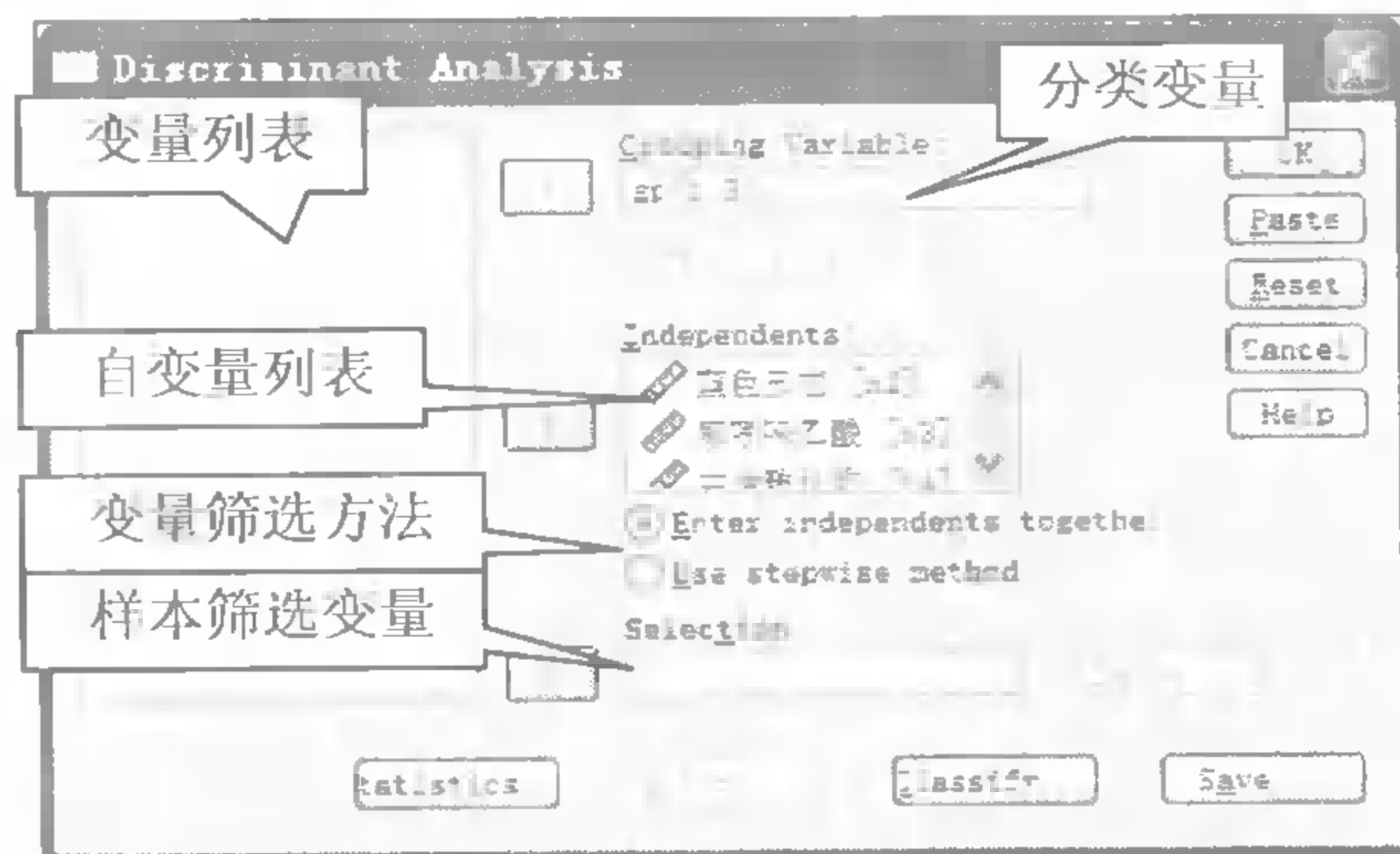




图 12-34 判别分析的主设置界面

#### 1. 变量设置

在变量列表中单击选中类别 (gp) 变量，单击从上至下第一个  按钮，将其作为分类变量选入 Grouping Variable 选框，单击 Define Range 按钮，在弹出对话框的 Minimum 和 Maximum 后面分别输入“1”、“3”，单击 Continue 按钮返回主界面；在变量列表中选中从铜蓝蛋白到中性硫化物的 4 个变量，单击从上至下第二个  按钮，将其作为自变量选入 Independents 列表框。

(1) Grouping Variable 选框，用于选入分类变量，它标识了观测量所属的类别。



选入分类变量后激活 Define Range 按钮，单击它弹出设置分类变量取值范围的对话框，在 Minimum 输入框指定分类变量的最小取值，在 Maximum 输入框指定分类变量的最大取值，再单击 Continue 按钮完成取值范围的设置并返回主界面。

(2) Independents 列表框，用于从变量列表选入进行判别分析的自变量。

(3) Selection 选框，用于选入对样本进行筛选的变量。

选入筛选变量后激活 Value 按钮，单击它弹出定义变量取值的对话框，在 Value for Selection 输入框指定 1 个取值，则只有筛选变量取这个值的观测记录才被用来进行判别函数的推导。单击 Continue 按钮完成设置并返回主界面。

(4) 另外，SPSS 为判别分析的变量选择提供了如下两种方法。

-  Enter independent together 单选框，表示建立包括所有自变量的全模型，当认为所有自变量都能为判别函数的建立提供丰富信息时，选中该选项。
-  Use stepwise method 单选框，指定使用逐步判别法，它需要根据各变量对判别贡献的大小进行选择；选中该项后，底部的 Method 按钮变为可用，单击它会弹出关于逐步判别的参数设置对话框，Method 的具体设置内容请参看第 12.6 节逐步判别分析的介绍。

#### 2. 保存选项设置

在图 12-34 中单击 Save 按钮，弹出如图 12-35 所示的保存设置对话框。单击 Continue 按

钮返回主界面。

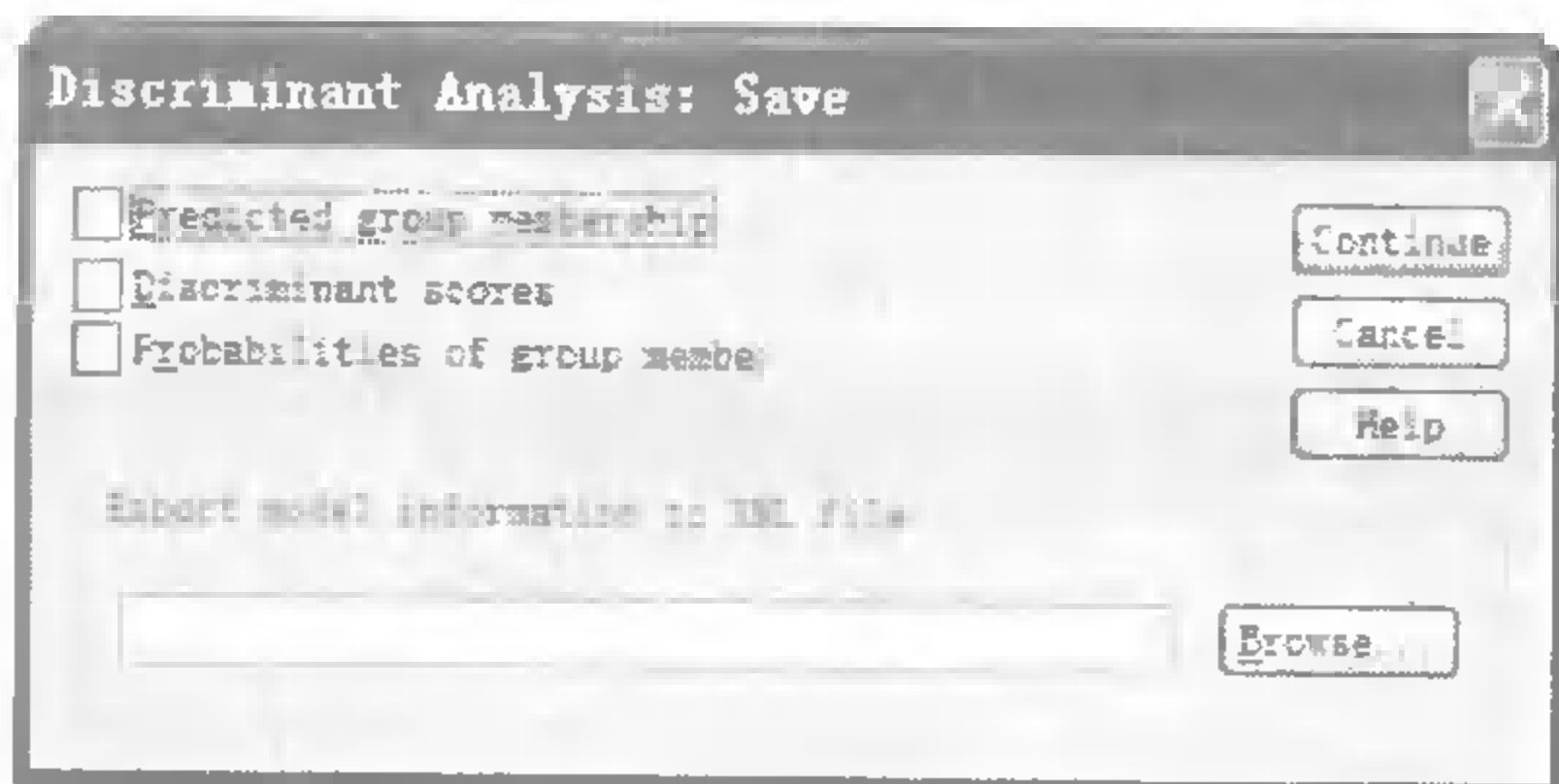


图 12-35 判别分析的保存选项设置

在此，SPSS 提供了如下 4 个设置选项。

(1) Predicted group membership 复选框，保存观测量的预测分类，即根据判别分数把观测量按后验概率最大原则所指派归属的类，新变量的默认变量名为 DIS\_n，其中 n 为一个正整数。

(2) Discriminant score 复选框，保存观测量的判别得分，该分数由未标准化的判别系数乘以自变量的取值再求和后得来；当前模型有几个判别函数，就新建几个得分变量。

(3) Probabilities of group membership 复选框，保存观测记录属于某一类的概率，有几个类别就建立几个新变量。

(4) Export model information to XML file 栏，把模型的设置信息保存至指定的 XML 文件，单击 Browse 按钮选择文件路径和名称。SmartScore 和 SPSS Server 等独立工具可以直接使用这些保存的 XML 文件。

### 3. 输出选项设置

在图 12-34 中单击 Statistics 按钮，弹出如图 12-36 所示的输出设置对话框。分别勾选如下几个复选框：Means、Box's M、Within-groups covariance matrix、Separate-groups covariance matrices、Fisher's 和 Unstandardized；单击 Continue 按钮返回主界面。

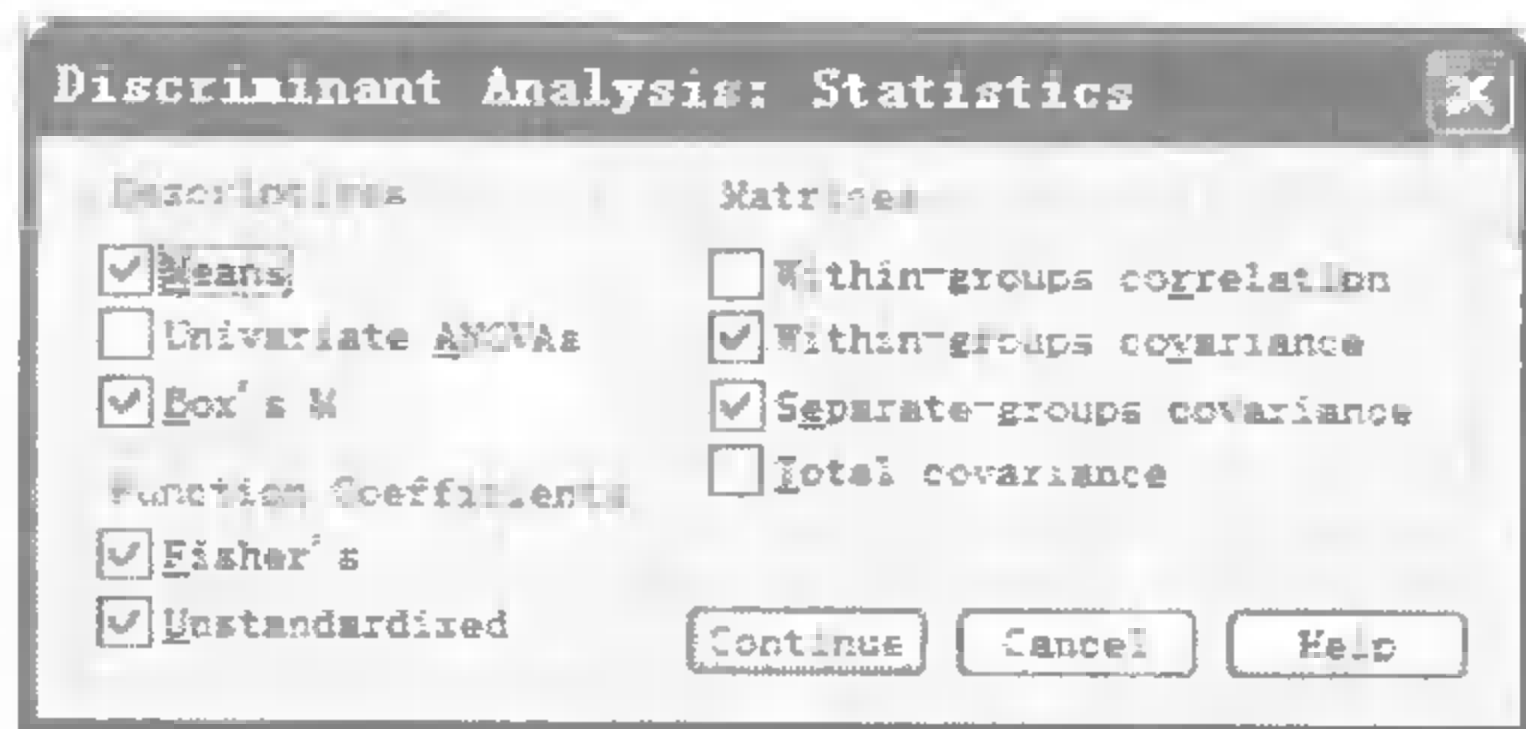


图 12-36 判别分析的输出选项设置

(1) Descriptives 栏，选择输出哪些描述统计量，有如下 3 个选项。

- Means 均值项，输出每个类别中各自变量的均值 (Mean)、标准差 (std Dev)，以及总样本中各自变量的均值和标准差。
- Univariate ANOVAs 方差分析选项，输出单变量的方差分析结果，检验的零假设是单个自变量在各类中的均值都相等。
- Box's M 协方差分析选项，检验各类别的协方差矩阵是否相等。

(2) Function coefficients 栏，选择判别函数系数的输出形式，有如下两个选择。

- Fisher's，可以直接用于对新样本进行判别分类的 Fisher 系数，对每个类别给出一组

系数，把观测量都归入判别得分最大的那一类中。

- Unstandardized, 未经标准化处理的判别系数。

(3) Matrices 栏, 选择关于矩阵的输出, 有 4 个可选项。

- Within-groups correlation matrix 类内相关矩阵, 根据类内协方差矩阵计算的相关矩阵。
- Within-groups covariance matrix 类内协方差矩阵, 它是将每个类别的协方差矩阵求平均后得到的, 不同于总体的协方差阵。
- Separate-groups covariance matrices, 输出每个类别各自的协方差矩阵。
- Total covariance matrix, 输出总样本的协方差矩阵。

#### 4. 分类参数的设置

在图 12-34 中单击 Classify 按钮, 弹出如图 12-37 所示的参数设置对话框。分别单击选中如下两个单选框: Compute from 和 Within-groups; 分别勾选如下几个复选框: Casewise results、Summary table、Combined-groups 和 Territorial map; 单击 Continue 按钮返回主界面。

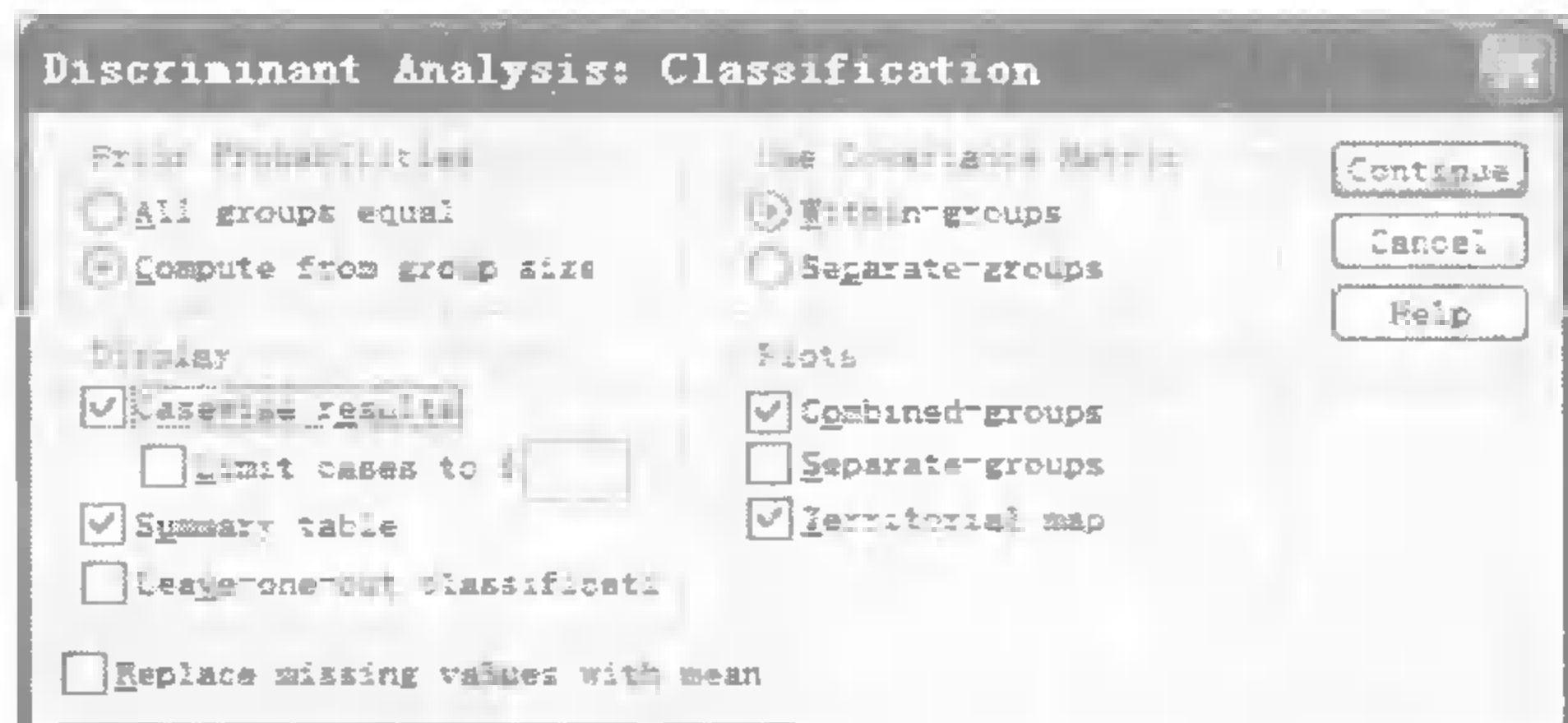


图 12-37 判别分析的分类选项设置

(1) Prior Probabilities 栏, 指定先验概率, 在如下两种方法中选择一个。

- All groups equal, 各类别的先验概率相等, 如果样本有  $n$  类, 它们的先验概率都为  $1/n$ 。
- Compute from groups sizes, 表示各类别的先验概率与其样本量成正比。

(2) Use Covariance Matrix 栏, 设置分类所使用的协方差矩阵, 有两个选择。

- Within-groups, 指定使用合并的类内协方差矩阵进行分类。
- Separate-groups, 指定使用每个类别的协方差矩阵进行分类。由于分类是根据判别函数, 而不是根据原始变量, 因此该选项并不等价于二次判别。

(3) Plots 栏, 选择输出哪些统计图形, 有 3 个选择。

- Combined-groups 联合散点图, 根据前两个判别函数的得分所作的、包括所有类别的散点图; 如果只有一个判别函数, 就输出直方图。
- Separate-groups 多张散点图, 根据前两个判别函数的得分所作的散点图, 总体分为几类就生成几张散点图; 如果只有一个判别函数, 则输出直方图。
- Territorial map 边界图, 根据判别函数的得分所作的、对观测量进行分类的边界图; 此图把平面划分成与分类个数相同的几个区域, 每类占据一个区域, 各类的均值在其区域中用 “\*” 号标出; 如果仅有一个判别函数, 则不作此图。

(4) Display 栏, 设置关于分类结果的输出选项, 有 3 个可选项。

- Casewise results, 输出对单个观测量的详细分类信息。Limits cases to 复选框设置输出的

范围,若输入n,表示只对前n个观测量有输出,当观测数目很大时建议勾选此项。

Summary table,输出分类总结表,包括正确分类的观测数目(原始类和根据判别函数给出的预测类相同)和错分观测数目,以及正确率和错误率。

Leave-one-out classification,输出交互校验信息,也称为“U-method”。由除去单个观测以外的其他观测导出的判别函数预测这个观测的类别,输出如此得到的统计信息。

(5) Replace missing value with mean 复选框,勾选表示用变量的均值代替其缺失值。

#### 12.5.4 案例的结果分析

在图 12-34 中单击 OK 按钮运行,SPSS Viewer 窗口的输出结果如图 12-38~图 12-41 所示。

组统计量					
类别		均值	标准差	有效的 N (列表状态)	
				未加权的	已加权的
胃癌患者	蓝色反应	150.40	16.502	5	5.000
	尿吡啶乙酸	13.80	5.933	5	5.000
	中性硫化物	20.99	12.323	5	5.000
	铜蓝蛋白	188.60	57.138	5	5.000
萎缩性胃炎	蓝色反应	118.75	14.104	4	4.000
	尿吡啶乙酸	7.50	1.732	4	4.000
	中性硫化物	14.25	8.386	4	4.000
	铜蓝蛋白	156.25	47.500	4	4.000
其他胃病	蓝色反应	121.40	13.012	5	5.000
	尿吡啶乙酸	5.00	1.871	5	5.000
	中性硫化物	8.00	7.314	5	5.000
	铜蓝蛋白	151.00	33.801	5	5.000
合计	蓝色反应	131.00	29.203	14	14.000
	尿吡啶乙酸	8.86	5.318	13	13.000
	中性硫化物	14.14	10.726	14	14.000
	铜蓝蛋白	155.93	46.787	14	14.000

图 12-38 关于样本的描述统计输出

协方差矩阵					
类别		蓝色反应	尿吡啶乙酸	中性硫化物	铜蓝蛋白
胃癌患者	蓝色反应	272.300	9.100	39.750	-211.300
	尿吡啶乙酸	9.100	35.200	25.000	-103.350
	中性硫化物	-39.750	-25.000	177.500	402.000
	铜蓝蛋白	-211.300	103.350	-402.000	1284.000
萎缩性胃炎	蓝色反应	198.917	20.500	74.167	338.750
	尿吡啶乙酸	20.500	1.000	12.333	-27.500
	中性硫化物	74.167	12.333	70.111	-119.833
	铜蓝蛋白	338.750	-27.500	-119.833	2255.250
其他胃病	蓝色反应	169.309	8.750	23.750	144.600
	尿吡啶乙酸	8.750	3.500	1.000	-8.750
	中性硫化物	-23.750	-1.000	53.500	117.500
	铜蓝蛋白	144.600	-8.750	117.500	1142.500

图 12-39 协方差矩阵输出

对数行列式		
类别	秩	对数行列式
胃癌患者	4	20.943
萎缩性胃炎	4	15.315
其他胃病	4	28.116
汇总的组内	4	28.116

打印的行列式的秩和自然对数是组内协方差矩阵的秩和自然对数。

a 秩 < 4  
b 案例太少无法形成非奇异矩阵

检验结果 <sup>a</sup>		
箱的 M	近似 F	26.091
	df1	10
	df2	305.976
	Sig.	.345

对相等总体协方差矩阵的零假设进行检验。

a 有些协方差矩阵是奇异矩阵,因此一般程序不会起作用。将相对非奇异组的汇总组内协方差矩阵检验非奇异组。其行列式的对数为 21.390。

图 12-40 Box's M 检验结果

特征值				
函数	特征值	方差的 %	累积 %	正则相关性
1	3.167 <sup>a</sup>	95.2	95.2	.872
2	.159 <sup>a</sup>	4.8	100.0	.370

a 分析中使用了前 2 个规范判别式函数。

Wilks 的 Lambda				
函数检验	Wilks 的 Lambda	卡方	df	Sig.
1 到 2	.207	14.958	8	.060
2	.863	1.398	3	.705

图 12-41 判别函数的方差解释和显著性检验



(1) 样本描述摘要。如图 12-38 所示,“分析案例处理摘要”表格是关于样本的使用信息,包括有效数据和缺失数据(本例中的缺失数据就是未分类的观测量)的统计信息。

“组统计量”表格给出了各个类别的均值、标准差等统计量,通过这些统计数据,可以大概地了解不同病情的患者在这 4 个生理指标上的差异。

(2) 样本的协方差矩阵及其检验。如图 12-39 所示,给出了总样本的协方差矩阵以及各个类别的协方差矩阵。

如图 12-40 所示,“检验结果”表格给出了 Box's M 检验的结果,从  $\text{Sig} > 0.10$  推断,不能否定各类协方差矩阵相等的零假设,建议使用汇聚的组内矩阵(Within-groups)进行计算和分类。如果否定了协方差矩阵相等的假设,就应使用分组的协方差阵(Separate-groups)分析。

(3) 判别函数的检验。如图 12-41 所示,“特征值”表格给出了两个典型判别函数所能解释的方差变异,其中第 1 个函数解释了所有变异的 95.2%,第 2 个函数解释了余下的 4.8%。

“Wilks 的 Lambda”表格用来检验各个判别函数有无统计学上的显著意义,从 Sig 值看,第 1 个函数在 0.1 的显著性水平上是比较显著的,而第 2 个函数并不显著;考虑到第 1 个函数较为显著地解释了 95%以上的方差变异,从而可以接受由此建立的判别规则。若想让所有的判别函数都显著成立,可以考虑使用逐步判别法。

(4) 标准化的典型判别函数系数。如图 12-42 所示,“系数”表格是两个判别函数中各个变量的标准化系数,由此可以判断各函数主要受哪些变量的影响;“结构矩阵”给出的是判别变量和标准化判别函数之间的相关性数据,同样可以用来判断各函数受哪些判别变量的影响最大。两个输出都说明了第 1 个函数受尿吡啶乙酸、蓝色反应的影响较大,第 2 个函数受蓝色反应、中性硫化物的影响较大。

标准化的规范判别式函数系数		
	函数	
	1	2
蓝色反应	505	-753
尿吡啶乙酸	585	532
中性硫化物	347	653
铜蓝蛋白	443	-395

结构矩阵		
	函数	
	1	2
尿吡啶乙酸	523*	309
铜蓝蛋白	229*	-031
蓝色反应	611	-630*
中性硫化物	294	527*

判别变量和标准化规范判别式函数间的汇聚组间相关性  
\* 每个变量和任意判别式函数间最大的绝对相关性

图 12-42 标准化系数和结构矩阵

(5) 未标准化的典型判别函数系数和类别质心函数。图 12-42 中的标准化系数在使用时需要先将原始变量标准化,不太方便;而非标准化判别系数可以直接通过原始变量进行计算,如图 12-43 中的“规范判别式函数系数”表格所示。

规范判别式函数系数		
	函数	
	1	2
蓝色反应	041	-051
尿吡啶乙酸	177	138
中性硫化物	034	055
铜蓝蛋白	009	-005
(常量)	9.023	5.622
非标准化系数		

组质心处的函数		
类别	函数	
	1	2
胃癌患者	2.092	-0.072
萎缩性胃炎	-0.825	0.527
其他胃病	-1.431	-0.349

在组均值处评估的非标准化规范判别式函数

图 12-43 未标准化的判别函数系数和类别质心函数

在图 12-43 中,“组质心处的函数”表格给出的是各类别的重心在平面上的坐标,例如胃癌患者的重心坐标为 (2.092, -0.072)。只要根据前面的典型判别函数(标准化或未标

准化的)计算出每个观测的平面坐标后,再计算它们和各类重心的距离,就可以判断其类别归属了。

(6) Fisher 判别函数。使用典型判别函数(标准化的或未标准化的)时,对每个观测先要计算其平面坐标,然后比较它与各类别重心的距离,再做分类。相比而言,Fisher 判别函数要简单许多,直接用它计算每个观测属于各类的得分,并把此观测归入得分最高的类别中即可。

本例中 Fisher 判别函数的输出,如图 12-44 中的“分类函数系数”表格所示。

组的先验概率			
类别	先验	用于分析的案例	
		未加权的	已加权的
胃癌患者	357	5	5 000
萎缩性胃炎	286	4	4 000
其他胃病	357	5	5 000
合计	1 000	14	14 000

分类函数系数			
	类别		
	胃癌患者	萎缩性胃炎	其他胃病
蓝色反应	780	629	649
尿卟啉乙酸	865	429	201
中性硫化物	131	071	-008
铜蓝蛋白	154	122	122
常量	-81.468	-50.292	-50.128

Fisher 的线性判别式函数

图 12-44 先验概率和 Fisher 判别函数

(7) Territorial map 边界图。边界图根据典型判别函数,按照观测量与各类别重心的距离,在平面上划分类别区域,如图 12-45 所示。图中以红色“\*”号标识各类别的重心,某观测按照典型判别函数计算的坐标落在哪个区域,它就被归为哪个类别。

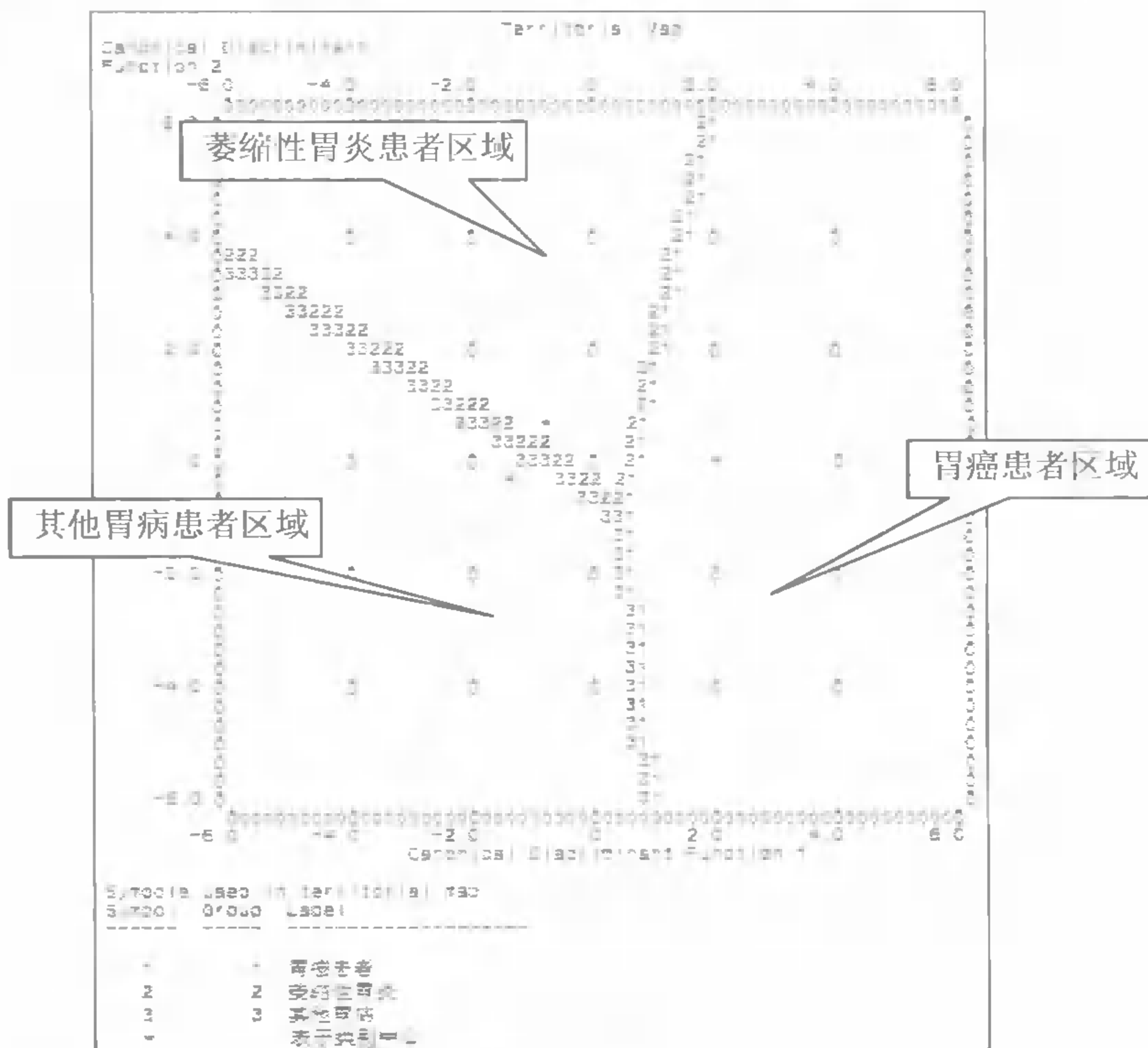


图 12-45 边界图输出

(8) 典型判别的散点图。利用两个典型判别函数,计算所有观测在 2 维平面的坐标,再加上 3 个类别重心的坐标,由此所作的散点图如图 12-46 所示。它可以直观的描绘用典型判别函数进行分类的结果。

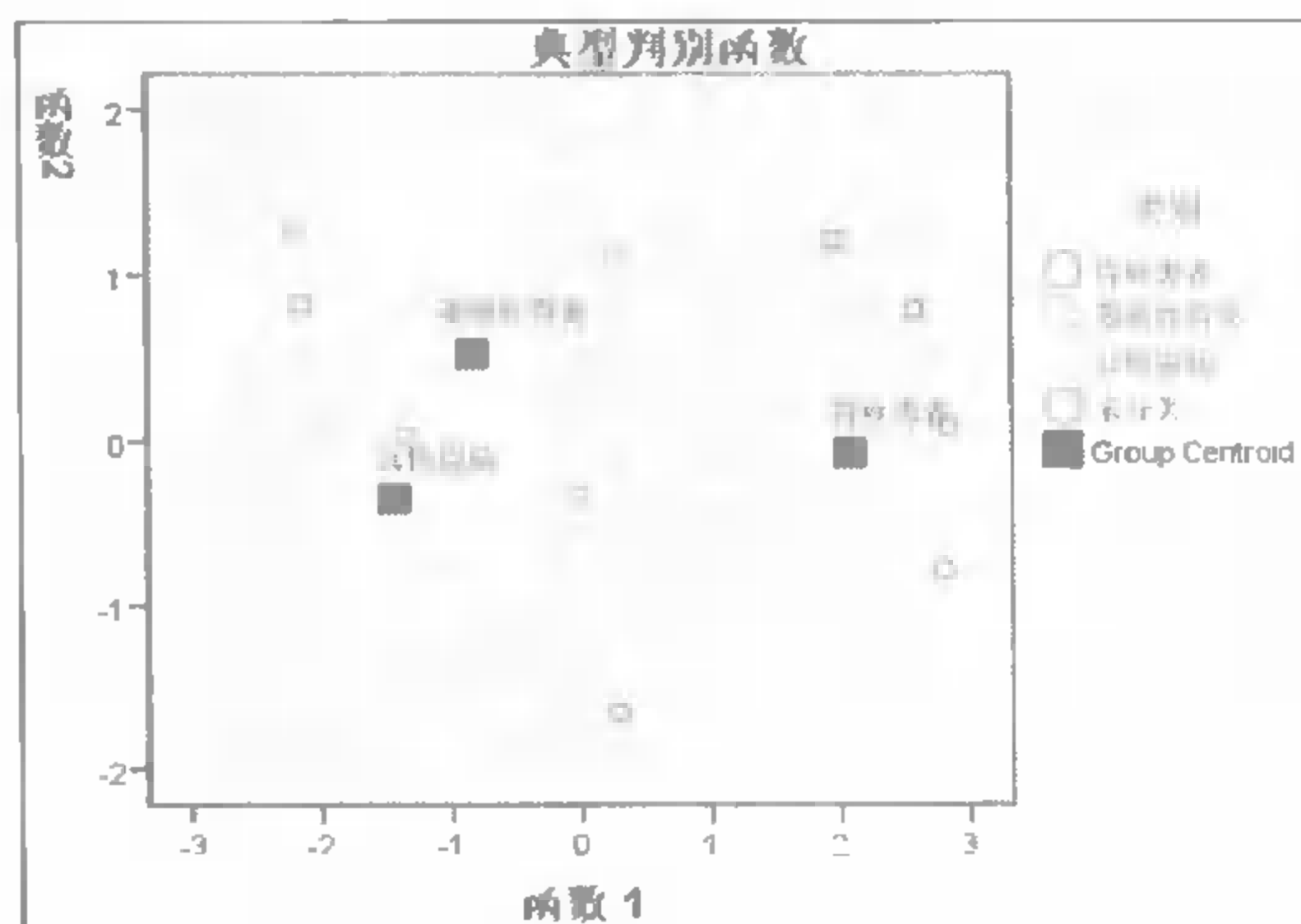


图 12-46 典型判别的散点图

(9) 分类总结表。如图 12-47 所示，“分类结果”表格是用典型判别函数进行预测的统计信息。

分类结果 <sup>a</sup>					
	类别	预测组成员			合计
		胃癌患者	萎缩性胃炎	其他胃病	
初始	计数				
	胃癌患者	4	0	1	5
	萎缩性胃炎	0	3	1	4
	其他胃病	0	1	4	5
	未分组的案例	0	0	1	1
	%				
	胃癌患者	80.0	0	20.0	100.0
	萎缩性胃炎	0	75.0	25.0	100.0
	其他胃病	0	20.0	80.0	100.0
	未分组的案例	0	0	100.0	100.0

<sup>a</sup> 已对初始分组案例中的 78.6% 个进行了正确分类。

图 12-47 判别结果总结表

以第一行为例来说明如何解读此表。原始数据合计有 5 例观测属于胃癌患者；对原始的 5 例胃癌患者，经过典型判别函数的判断，有 4 例仍判为胃癌患者（正确预测），有 1 例被判为其他胃病（错误预测）；其他数据的含义与此类似。最终，对原始观测案例的  $11/14 = 78.6\%$  进行了正确分类，未知类别的一个观测被判断为是其他胃病。

## 12.6 逐步判别分析实例

逐步判别的步骤分为两步。首先，根据自变量和因变量（分类变量）相关性的的大小筛选一部分自变量，这里的相关性是指自变量能否显著地把因变量区分开来；然后，用取定的变量作进一步的判别分析。注意：在模型中保留的自变量不是单独地参考每个自变量和因变量的相关性，而是综合考虑由一部分自变量形成的整体对因变量的区分能力。

### 12.6.1 问题描述和数据准备

某研究人员采集了 17 个企业在固定资产率、资金利率等 7 方面的数据，其中的 15 个企业被归结为如下 3 种类型中的一种：管理型、一般型和资金型，另有两个企业的类型尚未确定。本节通过逐步判别分析，利用已知类别的 15 个企业的信息，建立对企业类型的判别规则，并对余下的 2 个企业加以归类。

所用数据文件为“表征企业类型的数据.sav”，数据格式如图 12-48 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	id	Numeric	2	0	编号	None	None	3	Right	Scale
2	name	String	8		企业名称	None	None	8	Left	Nominal
3	x1	Numeric	3	2	固定资产率 (%)	None	None	6	Right	Scale
4	x2	Numeric	3	2	固定资产利率 (%)	None	None	6	Right	Scale
5	x3	Numeric	3	2	资金利率 (%)	None	None	6	Right	Scale
6	x4	Numeric	3	2	资金利税率 (%)	None	None	6	Right	Scale
7	x5	Numeric	3	0	流动资金周转天	None	None	6	Right	Scale
8	x6	Numeric	3	2	销售收入利税率	None	None	6	Right	Scale
9	x7	Numeric	3	2	全员劳动生产率	None	None	7	Right	Scale
10	group	Numeric	6	0	类别	{1, 管理型}...	None	5	Right	Scale

图 12-48 关于企业类型的数据格式

## 12.6.2 逐步判别的参数设置

依次单击菜单“Analyze→Classify→Discriminant...”执行判别分析过程，其主设置面板如图 12-49 所示，一般判别分析和逐步判别分析都是通过此过程实现的。

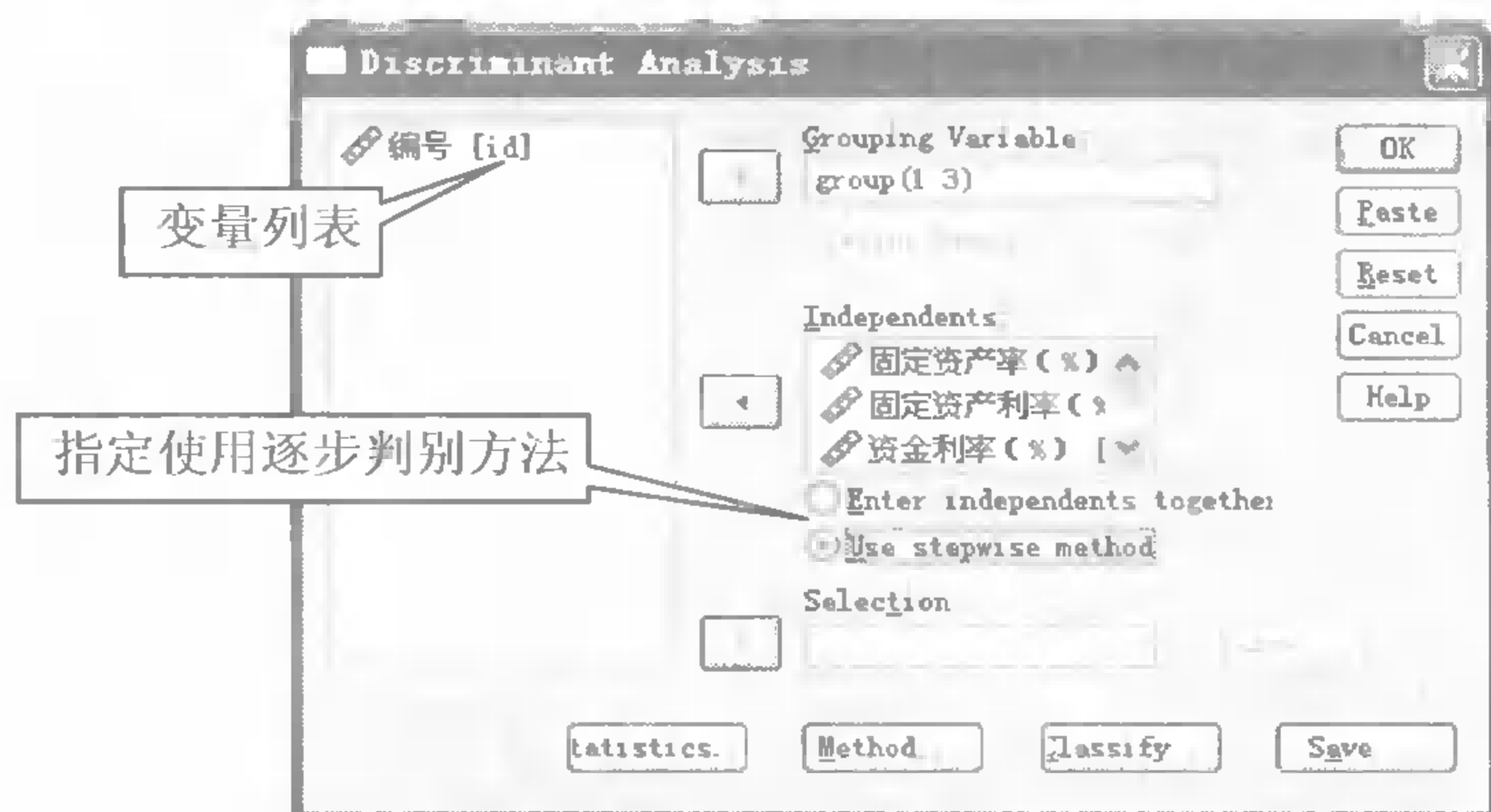


图 12-49 逐步判别的主设置面板

### 1. 变量、输出、和分类方法的设置

在变量列表中单击选中类别 (gp) 变量，单击从上至下第一个 按钮，将其作为分类变量选入 Grouping Variable 选框，单击 Define Range 按钮，在弹出对话框的 Minimum 和 Maximum 后面分别输入“1”、“3”，单击 Continue 按钮返回主界面；在变量列表中选中从固定资产率到全员劳动生产率的 7 个变量，单击从上至下第二个 按钮，将其作为自变量选入 Independents 列表框。单击选中 Use stepwise method 单选框。

此界面曾在一般判别分析中使用（如图 12-34 所示），设置选项请参考第 12.5.3 的介绍。

### 2. 输出选项设置和分类参数设置

在图 12-49 中单击 Statistics 按钮，弹出如图 12-50 所示的输出设置对话框，分别勾选如下 6 个复选框：Means、Box's M、Within-groups covariance、Separate-groups covariance、Fisher's 和 Unstandardized。单击 Continue 按钮返回主界面。



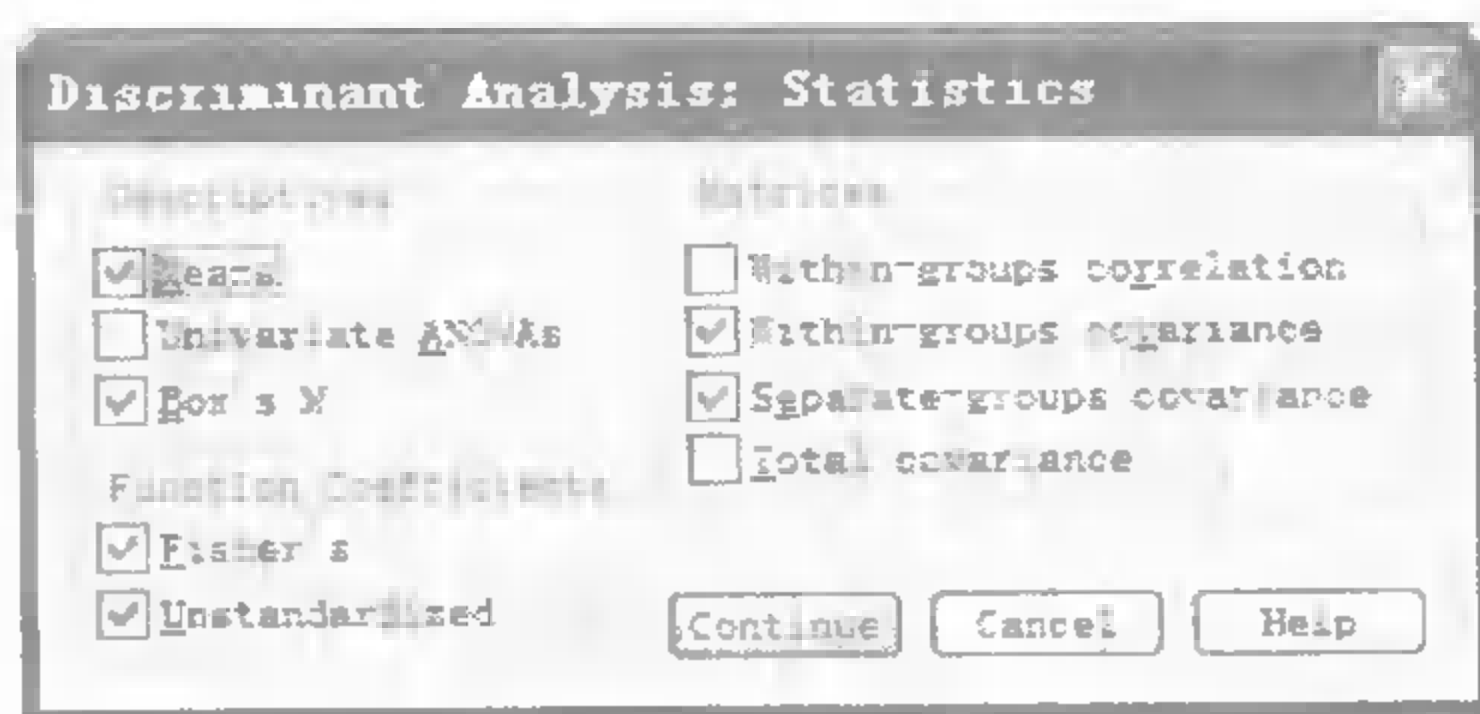


图 12-50 输出选项设置

在图 12-49 中单击 Classify 按钮，弹出如图 12-51 所示的参数设置对话框，分别单击选中如下两个单选框：Compute from 和 Within-groups；分别勾选如下两个复选框：Summary table 和 Combined-groups；单击 Continue 按钮返回主界面。

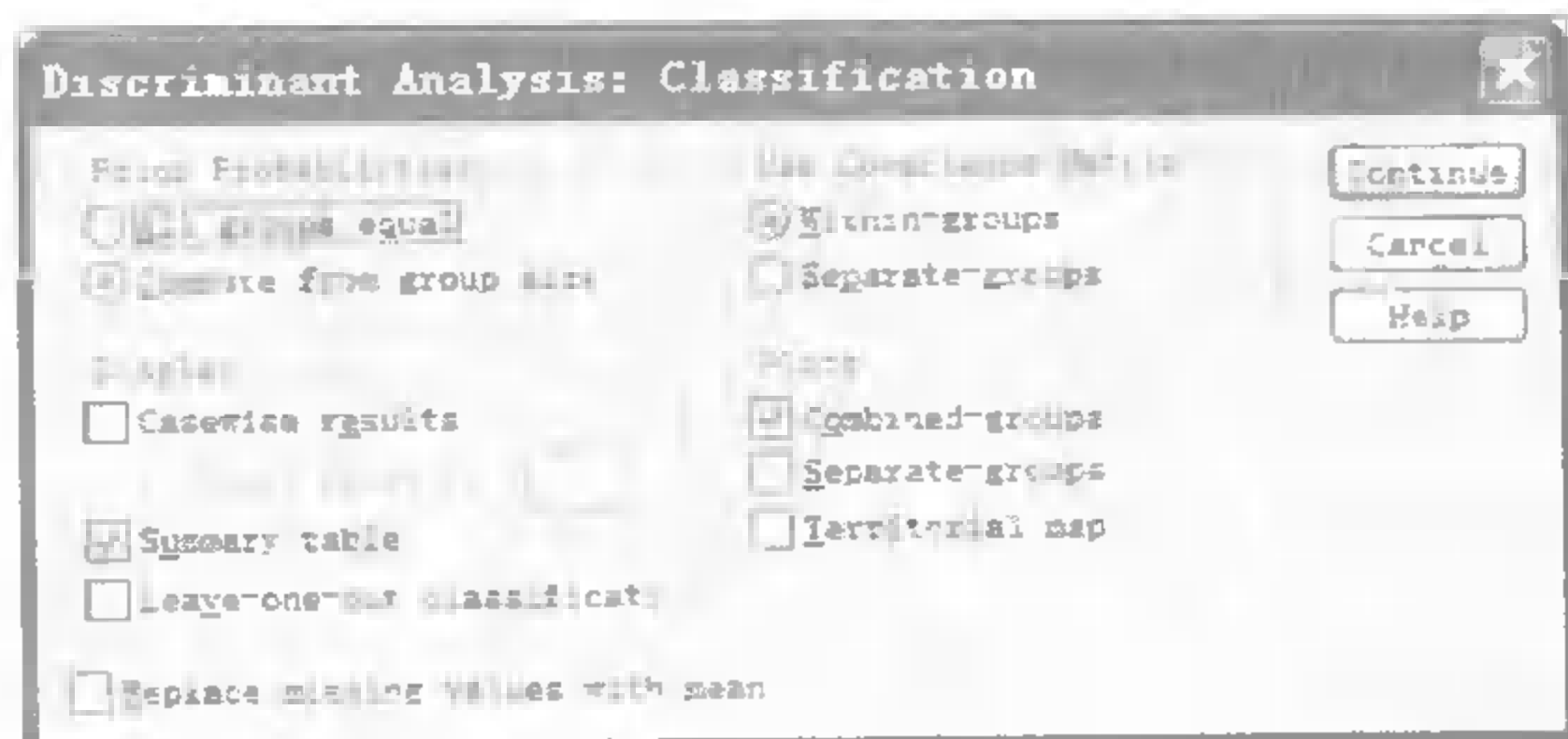


图 12-51 分类参数设置

这两个设置界面以及关于保存选项的设置界面，都与一般判别分析过程的有关设置相同，各选项的具体含义请参考第 12.5.3 的介绍。

### 3. 逐步判别方法的参数设置

在图 12-49 中单击 Method 按钮，弹出如图 12-52 所示的对话框，在此设置有关逐步判别的参数。分别单击选中如下两个单选框：Wilks' lambda 和 Use Probability of F；勾选 Summary of steps 复选框；单击 Continue 按钮返回主界面。

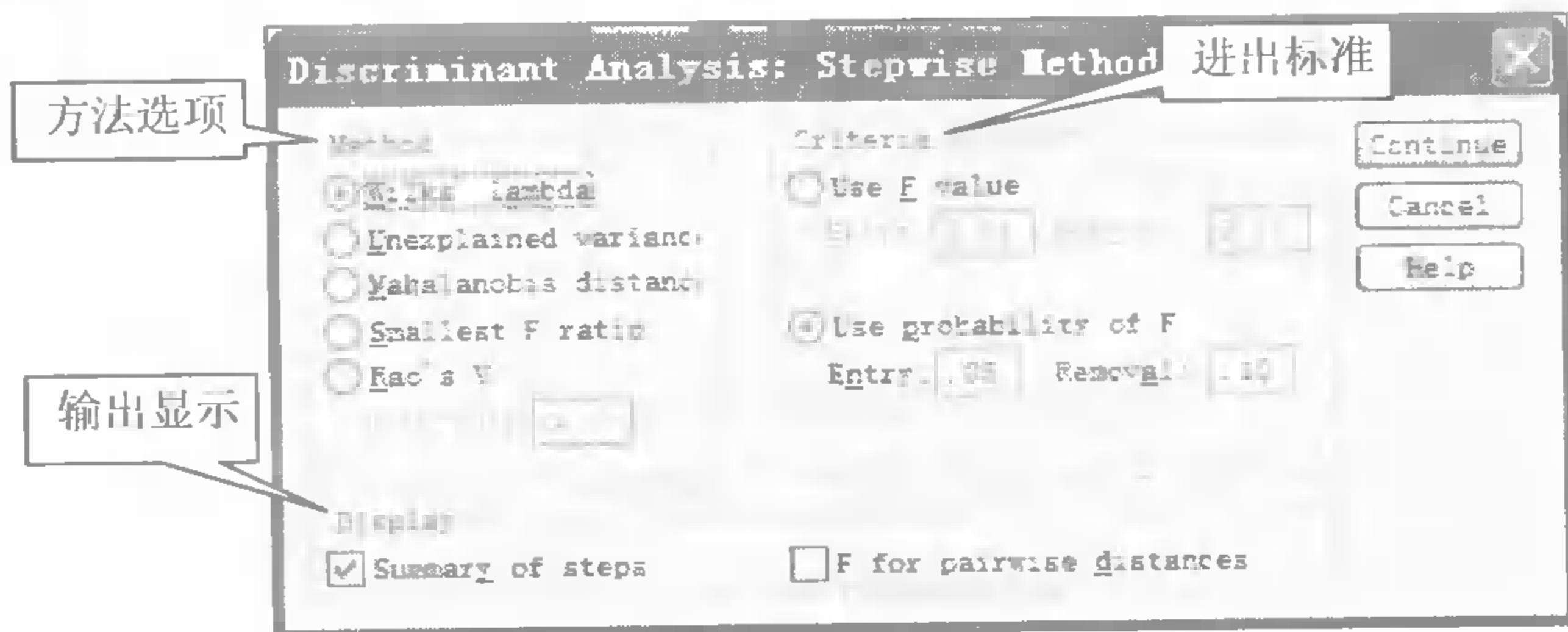


图 12-52 逐步判别方法的参数设置

(1) Method 栏，指定逐步判别分析的方法，在此 SPSS 提供了如下 5 种方法。

- Wilks' lambda，每步都选择使总体的 Wilks'  $\lambda$  统计量达到最小的变量进入判别函数。
- Unexplained variance，每步都选择使各类别间不可解释的方差和达到最小的变量进入判别函数。
- Mahalanobis' distance，每步都选择使靠得最近的两个类别的 Mahalanobis 距离达到最大的变量进入判别函数。

- Smallest F ratio, 每步都选择使基于类间 Mahalanobis 距离计算的一个 F 比率达到最大的变量进入判别函数。
- Rao's V, 每步都选择使 Rao's V 统计量产生最大增量的变量进入判别函数。当某变量导致的 V 值增量大于 V-to-enter 输入框指定的值时, 此变量就进入判别函数。

(2) Criteria 栏, 设置逐步判别过程中保留或删除变量的准则, 可供选择的依据有两个。

- Use F value 选项, 使用 F 值, 是默认方法。

当变量的 F 值大于指定的 Entry 值时, 该变量就会进入模型, 默认的 Entry 值为 3.84; 当变量的 F 值小于指定的 Removal 值时, 该变量就会从模型中剔除, 默认的 Removal 值为 2.71; Entry 值必须大于 Removal 值。要使模型包含更多的变量, 可以减小 Entry 值; 要使模型包含更少的变量, 可以增大 Removal 值。

- Use Probability of F 选项, 使用 F 检验的概率值。

默认的 Entry 值为 0.05, 默认的 Removal 值为 0.10, 其含义与使用 Use F value 时相仿。

(3) Display 栏, 设置输出哪些内容, 有如下两个可选项。

- Summary of steps, 输出逐步判别过程里的每一步的变量统计信息。
- F for Pairwise distances, 输出两两类别之间的 F 比率 (F ratios) 矩阵。

### 12.6.3 案例的结果分析

在图 12-49 中, 单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 12-53~图 12-56 所示。

分析案例处理摘要				组统计量					
未加权案例		N	百分比	类别	均值	标准差	有效的 N (列表状态)		
有效		15	88.2	管理型	60.4360	6.83726	未加权的	5	5.000
排除的	缺失或越界组代码	2	11.8	固定资本率 (%)	21.1660	6.28101	未加权的	5	5.000
	至少一个缺失判别变量	0	0.0	资金利率 (%)	22.1100	6.41639	未加权的	5	5.000
	缺失或越界组代码还有	0	0.0	资金利税率 (%)	14.0280	6.84440	未加权的	5	5.000
	至少一个缺失判别变量	0	0.0	流动资金周转天数	51.0000	5.43139	未加权的	5	5.000
合计		17	100.0	销售收入利税率 (%)	21.1280	6.68469	未加权的	5	5.000
				全员劳动生产率 (万元/人年)	2.2760	5.1389	未加权的	5	5.000

图 12-53 基本统计信息

检验结果		
箱的 M	近似	25.341
F		1.260
	df1	12
	df2	492.007
	Sig	.239

对相等总体协方差矩阵的零假设进行检验。

图 12-54 样本协方差阵相等的检验输出

输入的/删除的变量 a, b, c, d									
步骤	输入的	Wilks 的 Lambda							
		统计量	df1	df2	df3	统计量	df1	df2	Sig
1	流动资金周转天数	.473	1	2	12.000	6.684	2	12.000	.011
2	固定资本率 (%)	.181	2	2	12.000	7.425	4	22.000	.001
3	全员劳动生产率 (万元/人年)	.097	3	2	12.000	7.369	6	20.000	.000

在每个步骤中, 输入了最小化整体 Wilks 的 Lambda 的变量。

a 步骤的最大数目是 14。

b 要输入的 F 的最大显著水平是 .05。

c 要删除的 F 的最小显著水平是 .10。

d F 级、容差或 VIF 不足以进行进一步计算。

图 12-55 逐步判别的变量筛选过程

特征值				
函数	特征值	方差的 %	累积 %	正则相关性
1	4.037 <sup>a</sup>	79.4	79.4	.895
2	1.046 <sup>a</sup>	20.6	100.0	.715

a. 分析中使用了前 2 个规范判别式函数。

Wilks 的 Lambda				
函数检验	Wilks 的 Lambda	卡方	df	Sig.
1 到 2	.097	25.662	6	.000
2	.489	7.877	2	.019

图 12-56 特征值输出和 Wilks' Lambda 检验结果

(1) 基本统计信息。如图 12-53 所示,“分析案例处理摘要”表格是关于样本的使用信息,包括有效数据、缺失数据(本例中的缺失数据就是未分类的观测量)的统计信息。“组统计量”表格是连续变量在各个类别的均值、标准差等统计量,这里只给出了“管理型”的输出,其他类似。

(2) 样本协差阵的检验。如图 12-54 所示,在 Box's M 检验的结果中,从  $\text{Sig} > 0.10$  推断,不能否定各类协方差矩阵相等的零假设,说明使用 Within-groups 选项是合适的。

(3) 筛选变量的过程输出。如图 12-55 所示,是变量筛选的整个过程。可见,第 1 步加入了流动资金周转天数变量;第 2 步加入了固定资产率变量;第 3 步加入了全员劳动生产率变量;这 3 步的 Wilks' Lambda 检验都很显著,说明每一步加入的变量对正确判断分类都是有显著作用的。

(4) 典型判别函数的检验。如图 12-56 所示,“特征值”表格显示第 1 个判别函数解释了所有变异的 79.4%,第 2 个判别函数解释了 20.6%。“Wilks 的 Lambda”表格用来检验各个判别函数有无统计学上的显著意义,由 Sig 值都小于 0.05 推断,这两个判别函数的判别作用都显著地成立。

(5) 典型判别函数的系数输出。如图 12-57 所示,是标准化的典型判别函数系数和结构矩阵;图 12-58 是未标准化的典型判别函数系数和各类别的重心坐标。“结构矩阵”给出的是判别变量和标准化判别函数之间的相关性数据,可以用来判断各判别函数受哪些判别变量的影响较大。

标准化的规范判别式函数系数		
	函数	
	1	2
固定资产率 (%)	-.1316	-.821
流动资金周转天数	.1178	-.073
全员劳动生产率 (万元/人年)	.411	.1139

结构矩阵		
	函数	
	1	2
固定资产率 (%) <sup>a</sup>	-.529 <sup>a</sup>	.171
销售收入利税率 (%) <sup>a</sup>	-.467 <sup>a</sup>	-.217
固定资产率 (%)	-.436 <sup>a</sup>	.229
资金利率 (%) <sup>a</sup>	-.357 <sup>a</sup>	-.348
全员劳动生产率 (万元/人年)	-.284	.682 <sup>a</sup>
流动资金周转天数	.462	-.493 <sup>a</sup>
资金利税率 (%) <sup>a</sup>	-.386	-.300 <sup>a</sup>

判别变量和标准化规范判别式函数之间的汇聚组间相关性  
按函数内相关性的绝对大小排序的变量。  
<sup>a</sup> 每个变量和任意判别式函数间最大的绝对相关性  
<sup>a</sup> 该变量不在分析中使用。

图 12-57 标准化典型判别函数的系数和结构矩阵

规范判别式函数系数		
	函数	
	1	2
固定资产率 (%)	-.163	-.802
流动资金周转天数	.134	-.044
全员劳动生产率 (万元/人年)	.1156	.1208
常量	-.511	.741

非标准化系数

组质心处的函数		
	函数	
类别	1	2
管理型	-.345 <sup>a</sup>	.1067
一般型	.2301	-.016
资金型	-.1503	-.1310

在组均值处评估的非标准化规范判别式函数

图 12-58 非标准化典型判别系数和各类别重心函数

从判别函数系数和结构矩阵可以推断，第 1 个判别函数与固定资产利率、销售收入利税率、流动资金周转天数、固定资产率 4 个变量的相关性较大，第 2 个判别函数与全员劳动生产率、流动资金周转天数、资金利税率 3 个变量的相关性较大。

“组质心处的函数”表格给出各类别的重心在平面上的坐标，例如管理型企业的重心坐标为 (-1.437,1.067)。根据典型判别函数（标准化的或未标准化的）计算出每个观测的平面坐标后，再计算它们和各类别的重心之间的距离，就可以判断其类别归属了。

(6) Fisher 判别函数系数。使用典型判别函数（标准化的或未标准化的）时，对每个观测先要计算其平面坐标，然后比较它与各类别重心的距离，再做分类。相比而言，Fisher 判别函数要简单许多，直接用它计算每个观测属于各类的得分，并把此观测归入得分最高的类别中即可。

本例 Fisher 判别函数的输出，如图 12-59 的“分类函数系数”表格所示。

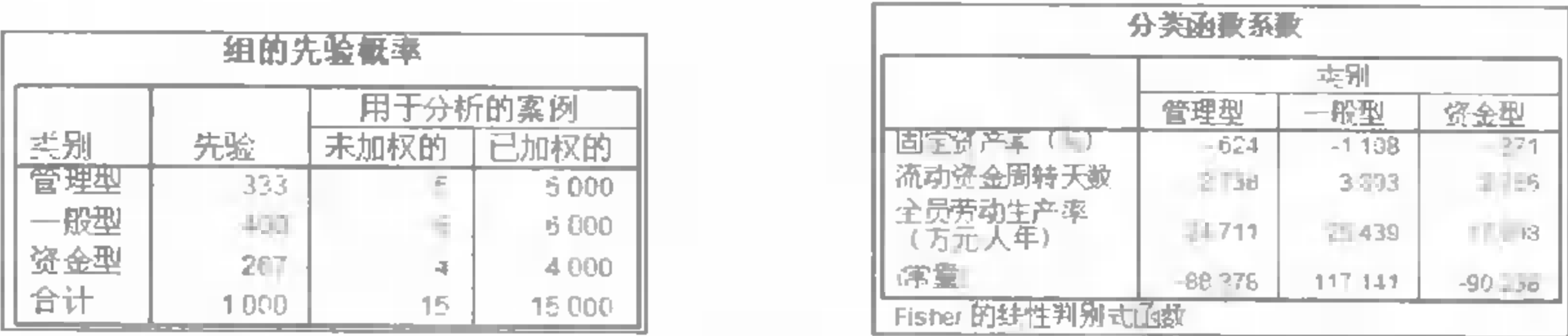


图 12-59 Fisher 判别函数系数的输出

(7) 典型判别的散点图。利用两个典型判别函数，计算所有观测在 2 维平面的坐标，再加上 3 个类别重心的坐标，由此所作的散点图如图 12-60 所示。它可以直观的描绘用典型判别函数进行分类的结果。

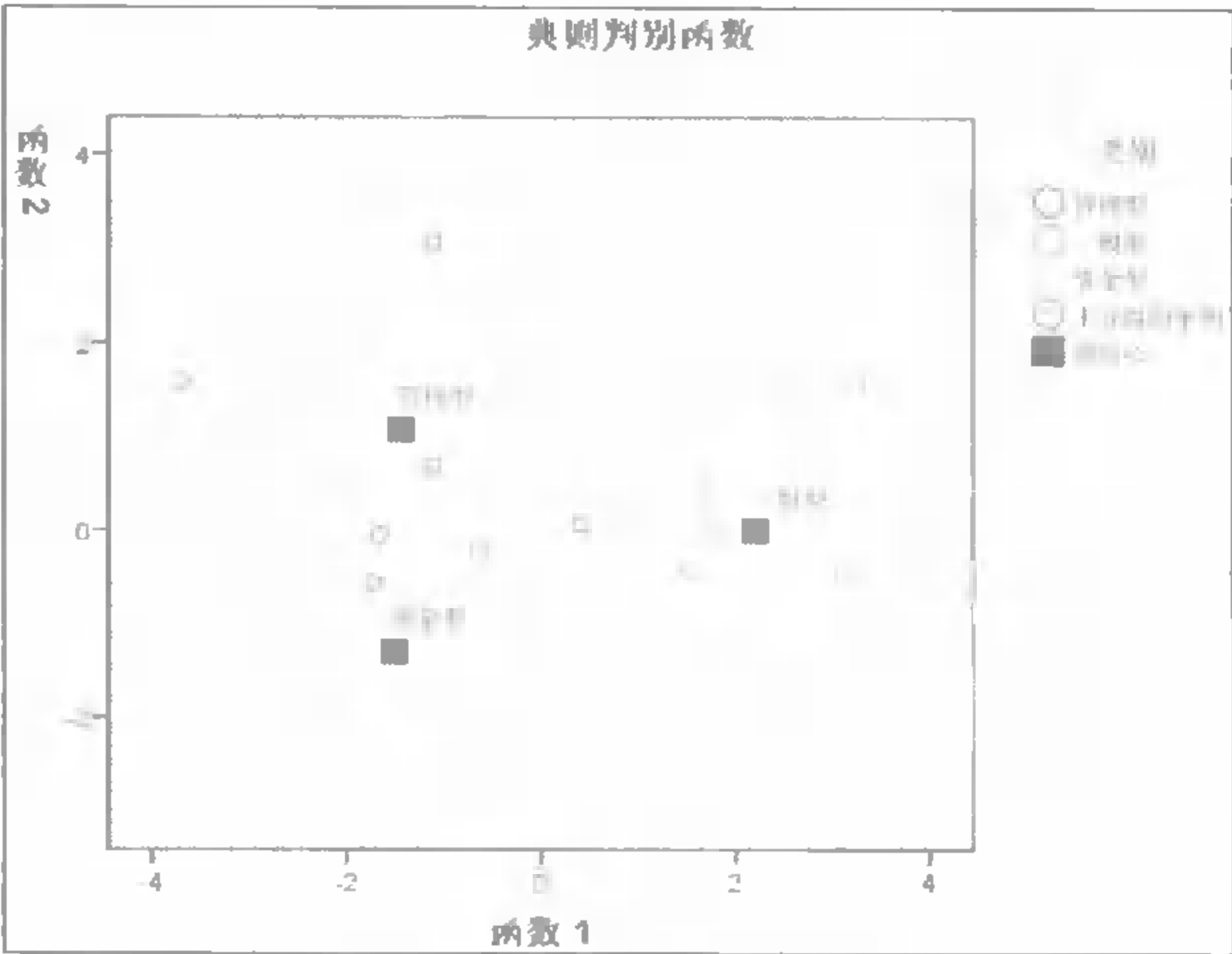


图 12-60 典型判别的散点图

(8) 分类总结表。如图 12-61 所示，“分类结果”表格是用典型判别函数进行预测的统计信息。

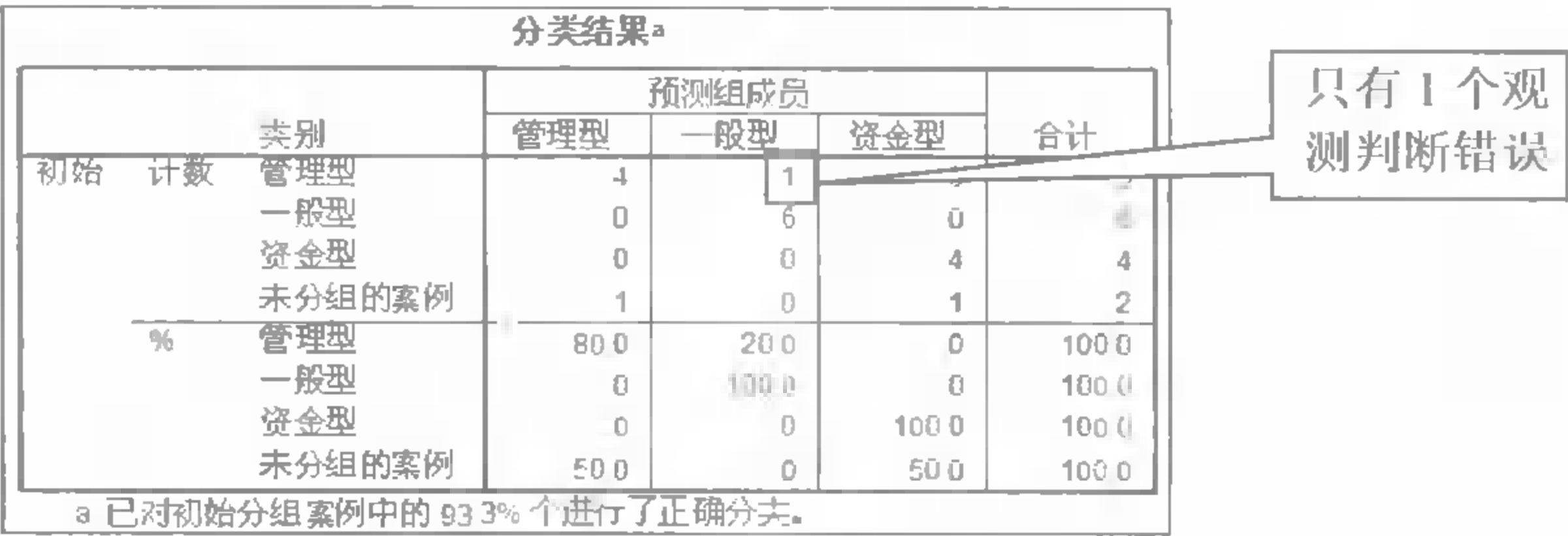


图 12-61 逐步判别的结果总结表



以第一行为例来说明如何解读此表。原始数据合计有 5 例观测属于管理型的；对原始的 5 例管理型企业，经过典型判别函数的判断，有 4 例仍判为管理型（正确预测），有 1 例被判为一般型（错误预测）；其他数据的含义与此类似。最终，对原始观测案例的  $14/15 = 93.3\%$  进行了正确分类，未知类别的两个企业一个归为管理型的、另一个归为资金型的。

## 12.7 决策树分析

SPSS Classification Trees 模块用于创建决策树模型，它可以帮助用户快速准确地识别特定群体、研究群体之间的相关关系以及预测未来事件，决策树模型可应用于数据分类、数据降维、预测、变量筛选、类别合并、连续变量离散化等许多方面。

SPSS 的决策树分析过程，可输出 SPSS Syntax 命令语句、标准 SQL 语句、文本 3 种格式的分类或预测规则，它们既可以输出在 Viewer 窗口中，也可以保存至外部文件以备后用。决策树模型的有关信息也能以 XML 格式导出至指定文件，SPSS Server 等独立工具可以直接使用在此保存的 XML 格式的模型文件。

SPSS 中的决策树生成算法有如下 4 种：CHAID、Exhaustive CHAID、CRT 和 QUEST。

### 12.7.1 决策树分类的基本原理

#### 1. 决策树的理解

从几何意义上可以直观地理解决策树的含义。将训练样本集中的每个观测都看成是  $n$  维（指  $n$  个输入变量）空间的一个点，决策树的建立过程就是它的分枝的形成过程，一个分枝就是在一定规则下对  $n$  维空间的一次区域划分；当决策树建立好以后， $n$  维空间便被划分成了若干个小区域；由于  $n$  维空间不易于观察，一般采用树型结构图的方式表示决策树。

决策树一般分为两大类型。分类决策树主要用于对离散因变量的分类；回归决策树主要用于对连续因变量的预测。可见决策树主要应用于分类和预测分析中，例如判断某些顾客是否为理想的潜在客户；预测具有某种特征的客户在未来的消费金额等。用决策树对一个新的观测作预测时，它自动根据输入变量的取值决定穿越决策树并达到最终叶节点的路径；如果是分类树，就根据最终节点的因变量取值确定对新观测的分类，并给出相应的可信度；如果是回归树，就计算最终节点里的因变量均值作为对新观测的预测值。

决策树模型有各种各样的算法，但各自都有一些优势和不足。一般地，决策树算法主要围绕两大核心问题展开。第一，决策树的生长问题，即利用训练样本集建立决策树的过程；第二，决策树的剪枝问题，即如何对建立的初始决策树进行节点合并及优化处理。下面就对这两个方面加以简要介绍。

#### 2. 决策树的生长

决策树生长的本质是一个对训练样本集不断分组的过程，树上的分枝正是在这个过程中逐渐生长出来的。当所有分枝的数据均无法继续细分时，一棵完整的决策树就形成了。

决策树生长的核心算法就是确定它的分枝准则，这涉及两方面的问题。第一，如何从众多的输入变量中选择一个最佳的分枝变量；第二，如何从指定分枝变量的众多取值中找到一个最佳的分枝阈值。现已有很多算法实现决策树的生长，例如 ID3、C4.5/C5、CHAID、CRT 等，它们大都能够常在常用的数据挖掘软件中找到，用户在使用时，只需要设置或调整几个简

单的参数，就能方便地建立决策树模型，同时完成对决策树的优化处理。

### 3. 决策树的修剪

随着决策树的生长，叶节点含有的样本量不断减少，它们对总体的代表性也不断降低，越深处的节点所体现的特征就越具体，一般性也越差，甚至可能出现如此的结论：只有年收入大于 50 000 元、年龄大于 50 岁、且姓名是张三的人，才是企业的理想客户。

由此可见，虽然一棵完整的决策树能比较准确地反映训练样本的数据特征，但因此也可能失去模型的一般代表性，使它不适用于对新数据的分类或预测。这种现象称之为过度拟合 (Overfitting)，解决这个问题方法之一是对决策树进行必要的修剪，常用的修剪技术有预修剪 (Pre-pruning) 和后修剪 (Post-pruning) 两种。

(1) 预修剪技术。预修剪最直接的方法是事先指定决策树生长的最大深度，使它不能过度生长。但这种方法要求用户对变量的取值分布有较为清晰的了解，并且需要对各种参数的取值反复进行尝试，否则无法给出一个较为合理的深度最大值。如果树的深度太浅，表示过于限制了决策树的生长，容易使它的代表性变得很差，这样也无法实现对新数据的有效分类或预测。

预修剪的其他方法都是采用检验技术来阻止决策树的过度生长，它们通过对树节点的各种检验，决定是否允许相应的分枝继续生长。较为简单的一个检验方法是为防止最终节点的样本量过少，事先给它指定一个最小值；在决策树生长过程中，将不断检验树节点的样本量是否小于所允许的最小值，如果小于就停止分枝的继续生长，否则可以继续分枝。另外，还可以利用统计检验（如卡方检验等）的方法检验树节点内部的差异显著性，以判断是否分枝。

使用预修剪技术的常用算法有 CHAID、ID3、C4.5 等。

(2) 后修剪技术。后修剪技术从另一个角度解决过度拟合的问题，它先让决策树充分生长，再根据一定的规则，剪去那些不具有一般代表性的叶节点或分枝。

后修剪技术是一个边修剪边检验的过程，一般规则是在剪枝的过程中，利用训练样本集或验证样本集，不断检验决策树对目标变量的预测精度，并计算相应的错误率；用户事先指定了一个最大的允许错误率，当剪枝达到某个深度时，如果计算出的错误率高于允许的最大值，就停止剪枝，否则可以继续剪枝。利用训练样本集计算修剪的错误率时，会出现错误率越低决策树复杂程度越高的现象；比较合理的做法是利用验证样本集对剪枝效果进行检验，当错误率明显增大时，再停止剪枝。

使用后修剪技术的常用算法有 CRT 等。

### 4. 决策树的应用和注意事项

除了分类和预测，决策树还可以应用于生成推理规则、寻找最佳变量等方面。

把决策树看作是推理规则的一种图形表示，用它能方便地输出推理规则的其他表现形式；另外，由于决策树的建立过程是一个不断选择最佳分枝变量的过程，一般高层节点比低层节点上的分枝变量对区分因变量的作用要大，所以可以把决策树作为一种衡量变量价值大小的工具。在决策树的应用中，应注意下面的两个问题。

(1) 一般的决策树算法中只能依据单个变量的取值对某个节点进行分枝，无法同时使用多个分枝变量，这在一定程度上限制了决策树的应用范围。一种改进办法是事先利用多个变量计算出新变量，例如比值、多项式求和等，然后再用新变量作为分枝变量。

(2) 决策树所处理的输入变量既可以是连续型的，也可以是分类型的。对于连续自变量，

优势是当数据采用不同的计量单位或存在离群点时，不会给决策树带来显著地影响，因而不会给数据的准备工作造成额外负担；缺点是忽略了数据中所蕴涵的关于分布形态的信息。对于分类自变量，决策树的建树效率会较高；但问题是当分类取值很多且分布又极为分散时，决策树容易长得过于“茂盛”，使最终节点的样本量变得很少；此时，一种改进的方法是将样本量较少的类合并，但由于类间合并有很多可选择的方案，只有穷尽所有方案后才有可能得到较好的合并结果，而穷尽操作的可行性又受到实际应用的限制。

### 5. 带有权重变量的数据集

在 SPSS 中，如果原始数据集里指定了权重变量（Weight variable），Tree 过程将自动对权重变量进行四舍五入处理，于是那些小于 0.5 的权重取值将被重设为 0，而权重为 0 的观测都被剔除掉不参与分析。

## 12.7.2 决策树过程的参数设置

依次单击菜单“Analyze→Classify→Tree...”执行决策树分析过程，在弹出 Tree 过程主设置界面的同时，弹出如图 12-62 所示的警告框。它提示用户在进行 Tree 分析前，必须正确设置分析变量的度量方式，因为度量方式能直接影响某些决策树算法的计算方式；而且如果因变量是分类变量，还必须为其指定值标签。

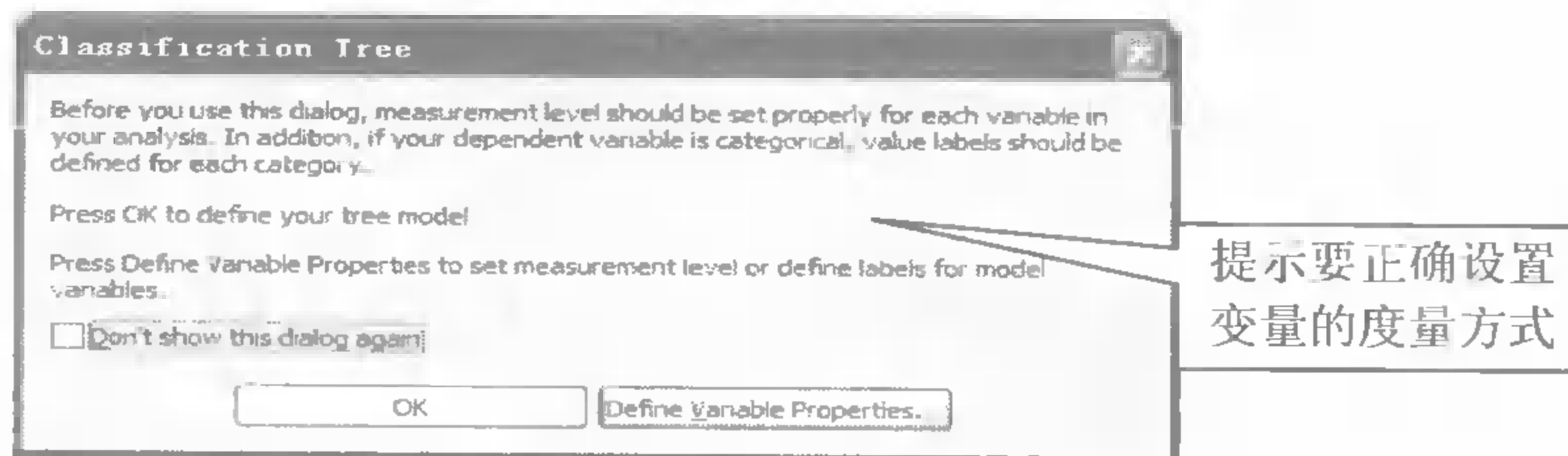


图 12-62 对变量格式设置的警告框

### 1. 主界面的参数设置

在图 12-62 中单击 OK 按钮，进入 Classification Trees 过程的主设置界面，如图 12-63 所示。

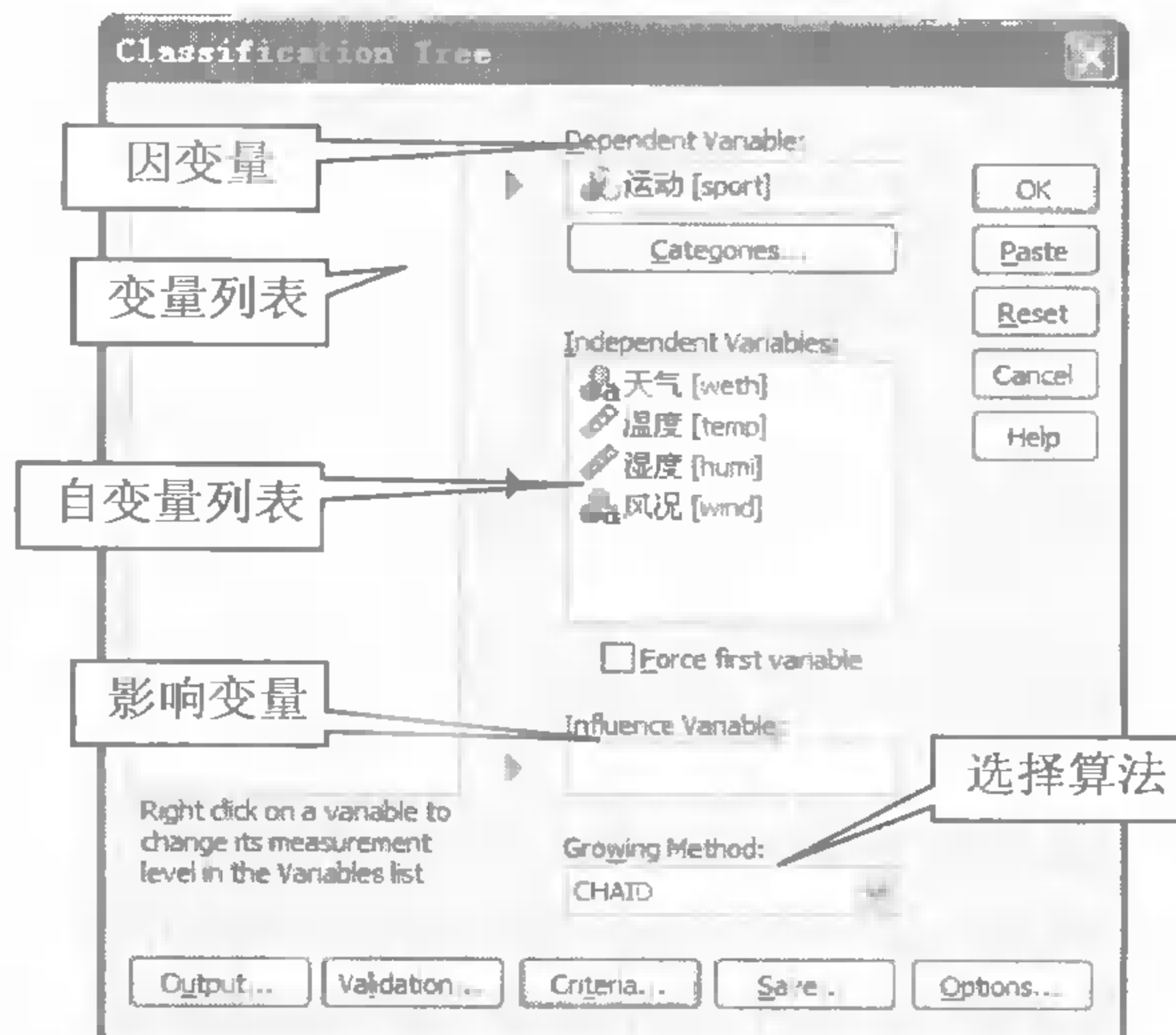


图 12-63 决策树分析的主设置面板



- (1) Dependent Variable 选框，用于从变量列表选入一个因变量。
- (2) Independent Variables 列表框，用于从变量列表选入多个自变量。
- (3) Force first variable 复选框。

勾选它表示直接将 Independent Variables 列表框中的第一个变量，作为决策树生长的开始节点（第 0 个节点）的分枝变量。

- (4) Influence Variable 选框，用于从变量列表选入一个影响自变量。

该变量反应单个观测对决策树生长的影响程度，取值越大影响越大；必须为数值型的变量，且不能设置因变量为影响变量；如果指定了 QUEST 算法，将忽略此变量并输出警告信息。

- (5) Growing Methods 下拉列表，用于指定决策树的生长算法，SPSS 给出如下 4 种方法。

- CHAID (Chi-squared Automatic Interaction Detection) 卡方自动交互检测法，它每一步都选择与因变量相关性最强的自变量作为预测变量，合并那些因变量没有显著差异的预测变量取值。
- Exhaustive CHAID 改进的 CHAID 算法，它检查每个预测变量所有可能的分枝方案。
- CRT (Classification and Regression Trees) 分类回归树，它把数据分到因变量取值尽可能一致的分支，因变量取值都相同的最终节点被称为纯节点。
- QUEST (Quick, Unbiased, Efficient Statistical Tree) 快速、无偏、有效的统计树，它避免其他算法的对多分类变量的青睐；只有因变量为名义变量 (Nominal) 时才可选。

## 2. 目标取值的定义

在图 12-63 中单击选中 Dependent Variable 选框里的变量，然后单击 Categories 按钮，弹出如图 12-64 所示的因变量目标取值定义对话框。

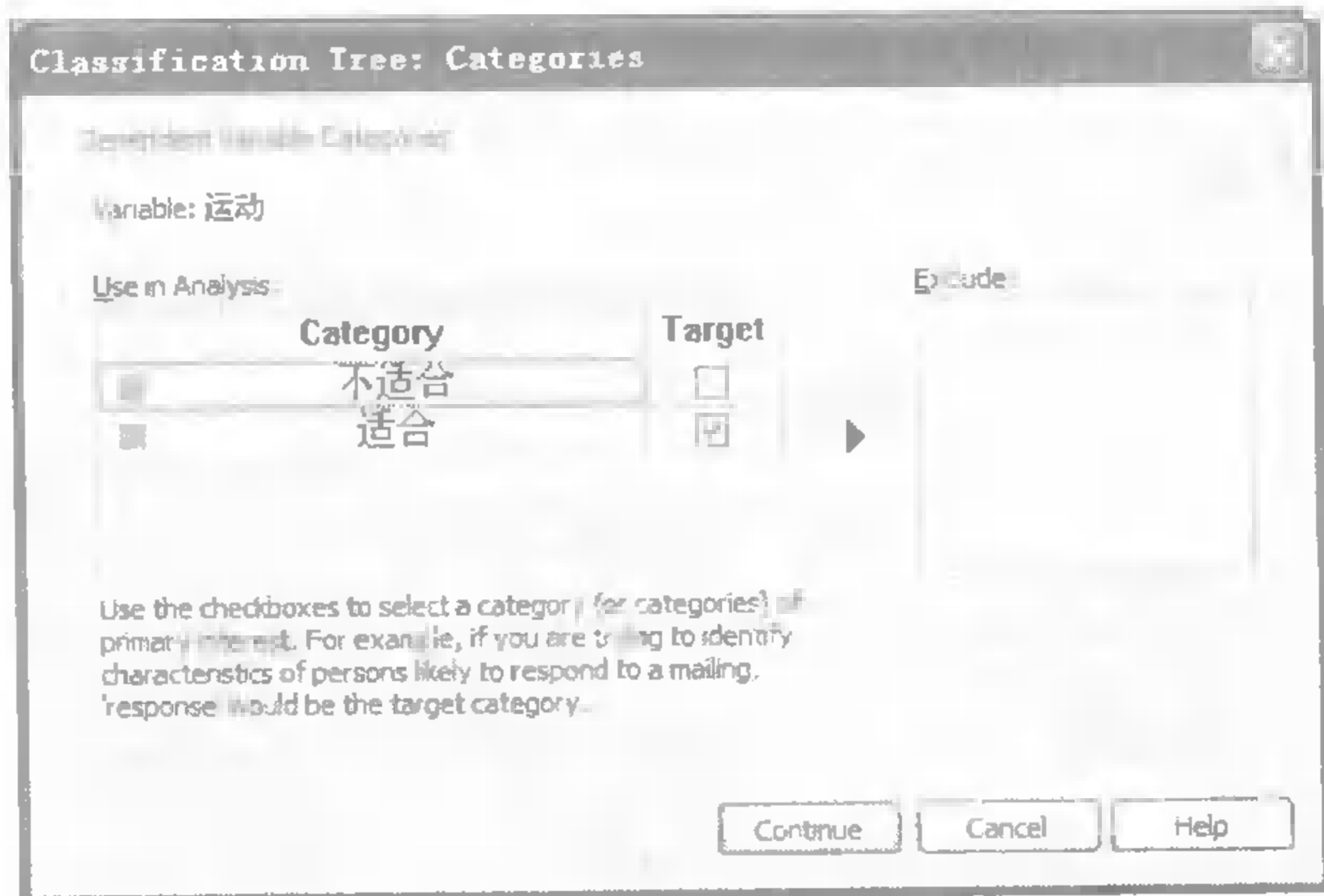


图 12-64 决策树分析的目标定义对话框

- (1) Variable 后显示了当前的因变量名称（运动），不可编辑。
- (2) Use in Analysis 子设置栏。
  - Category 列，给出当前因变量的值标签（必须预先定义）。
  - Target 列，给出一列复选框，用于选择研究者感兴趣的目标取值。

例如在对影响运动的天气进行分析时，用户对适合运动的天气更感兴趣，就在此勾选 Category 列为“适合”的 Target 复选框。允许同时选择多个感兴趣的日标值；默认情况下，



不设置任何目标值，但此时会有一些其他的参数选项变得不可用。

(3) Exclude 列表框，用于选入不参与分析的因变量取值。

默认情况下，用户自定义缺失值将显示在 Exclude 列表框里，可以把它们选入左侧的 Use in Analysis 表格参与分析。

### 3. 输出选项的 Tree 设置

在图 12-63 中单击 Output 按钮，弹出如图 12-65 所示的输出设置界面，它共有 4 个子标签面板。首先面对的是 Tree 设置子面板，在此设置关于决策树的输出格式。



图 12-65 决策树的输出格式设置

(1) Tree 复选框，表示输出图形决策树，勾选后激活 Display 栏的选项。

(2) Display 栏，设置图形决策树的输出格式，有如下 5 个设置选项。

- Orientation 显示方向，有 3 种可选方式：Top down，从上至下（根节点置于顶部）；Left to right，从左至右（根节点置于左边）；Right to left，从右至左（根节点置于右边）。
- Node contents 节点内容，有 3 种可选方式：Table（表格）、Chart（图形）和 Table and Chart（表格和图形）。对于分类变量，节点会显示其频数统计信息或者条图形；对于连续变量，节点会显示均值、标准差、观测数目等统计信息或者柱状图。
- Scale 显示范围，有两种可选方式：Automatic 单选框，对较大的决策树进行自动调整使其适合页面的大小，是默认选项；Custom 单选框，由用户指定决策树的显示比例，Percent 输入框用于指定这个比例的取值，可输入范围为 10~200。
- Independent variable statistics 复选框，对 CHAID 和 Exhaustive CHAID 算法，要求在节点中显示连续变量的 F 统计量值、显著性水平及其自由度，以及分类变量的卡方统计量、显著性水平及其自由度；对 CRT 算法，显示每步的改进值；对 QUEST 算法，显示连续变量和有序变量的 F 统计量值、显著性水平及其自由度，以及名义变量的卡方检验信息。
- Node definitions 节点定义，显示父节点分支时所用的自变量在其每个子节点的取值。

(3) Tree in table format 复选框，勾选后以表格形式输出决策树，包括每个节点的节点统计信息、父节点号码等内容。

#### 4. 输出选项的 Statistics 设置

在图 12-65 中单击 Statistics 标签，弹出如图 12-66 所示的统计量设置子面板。

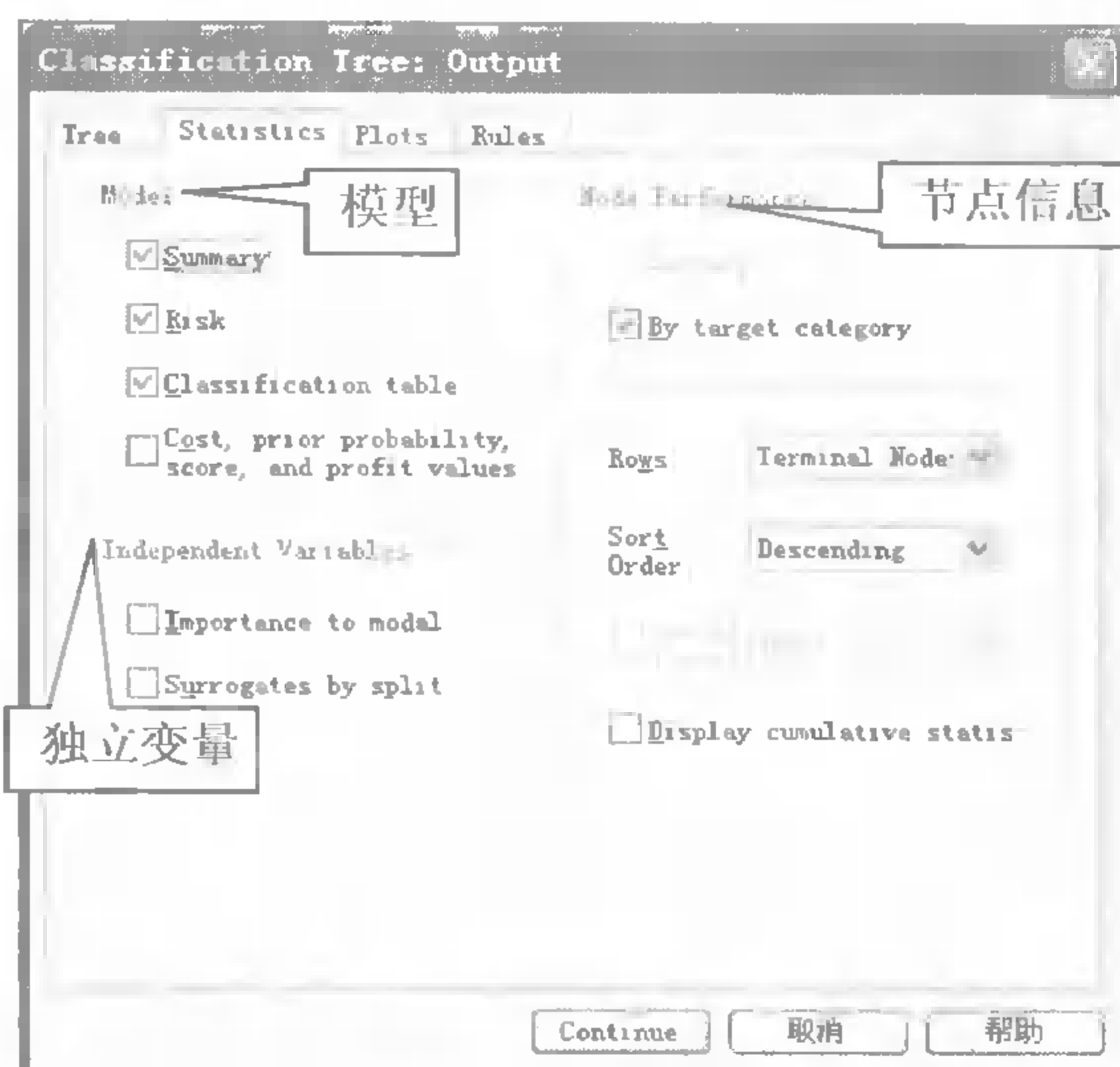


图 12-66 有关统计量的输出设置

(1) Model 栏，设置输出一些关于模型的统计信息，可选项有如下 4 个。

- Summary 摘要，包括模型方法、进入和没进入模型的变量等信息。
- Risk 风险估计及其标准误，用来衡量决策树的预测精度。对于分类因变量，风险估计就是经先验概率和错判损失调整后的错判比率；对连续因变量，风险估计就指节点内的方差。
- Classification table 分类表，对于分类因变量，给出其每个取值水平上的判断正确数和错误数；对于连续因变量，不做任何输出。
- Cost, prior probability, score, and profit values 复选框，对于分类因变量，输出错判损失函数、先验概率、得分和分析所使用的得益函数；对于连续因变量，不做任何输出。

(2) Independent Variables 栏，设置关于自变量的选项，有两个可选项。

- Importance to model 复选框，对于 CRT 方法，把模型中的自变量按其重要性进行排序，此选项对 QUEST 和 CHAID 方法无效。
- Surrogates by split 复选框，对 CRT 和 QUEST 方法，如果模型有可替代的解决方案，就列出所有可能的方案，此选项对 CHAID 方法无效。

(3) Node Performance 栏，设置关于节点的统计信息，有如下 3 个设置选项。

① Summary 摘要表格，对连续因变量，此表包括节点序号、观测数及自变量的均值；对于定义了得益函数的分类因变量，此表包括节点序号、观测数、平均得益和 ROI（投资回报）值；对未定义得益的分类因变量不起作用。

② By target category 复选框，对于定义了目标取值的分类因变量，此表包括得益比例、响应比例、以节点或百分位分组后的 lift 值。它对因变量的每个目标取值输出一个表格；对连续因变量和未定义目标的分类因变量不作输出。

③ Rows 下拉列表, 指定节点信息表的显示方式, 可选项有: Terminal nodes (最终节点)、Percentiles (百分位) 和 Both (两者都有)。如果选择了 Both, 为因变量的每个目标取值输出两个表格; 百分位表按指定顺序依次显示指定百分位处的累计值。

- Sort Order 下拉列表, 如果在 Rows 下拉列表选中了 Percentiles 或 Both, 在此指定百分位表的显示顺序, 有两个选择: Descending (降序) 或 Ascending (升序)。
- Percentile increment 下拉列表, 如果在 Rows 下拉列表选中了 Percentiles 或 Both, 在此指定百分位的递增间隔, 可选项有 1%、2%、5%、10%、20% 和 25%。
- Display cumulative statistics 复选框, 如果在 Rows 下拉列表选中了 Terminal nodes 或 Both, 勾选此项表示在每个最终节点表里增加一列显示累计结果。

## 5. 输出选项的 Plots 设置

在图 12-65 中单击 Plots 标签, 弹出如图 12-67 所示的作图设置子面板, 可选图形依赖于因变量的度量水平、决策树生长方法等参数。

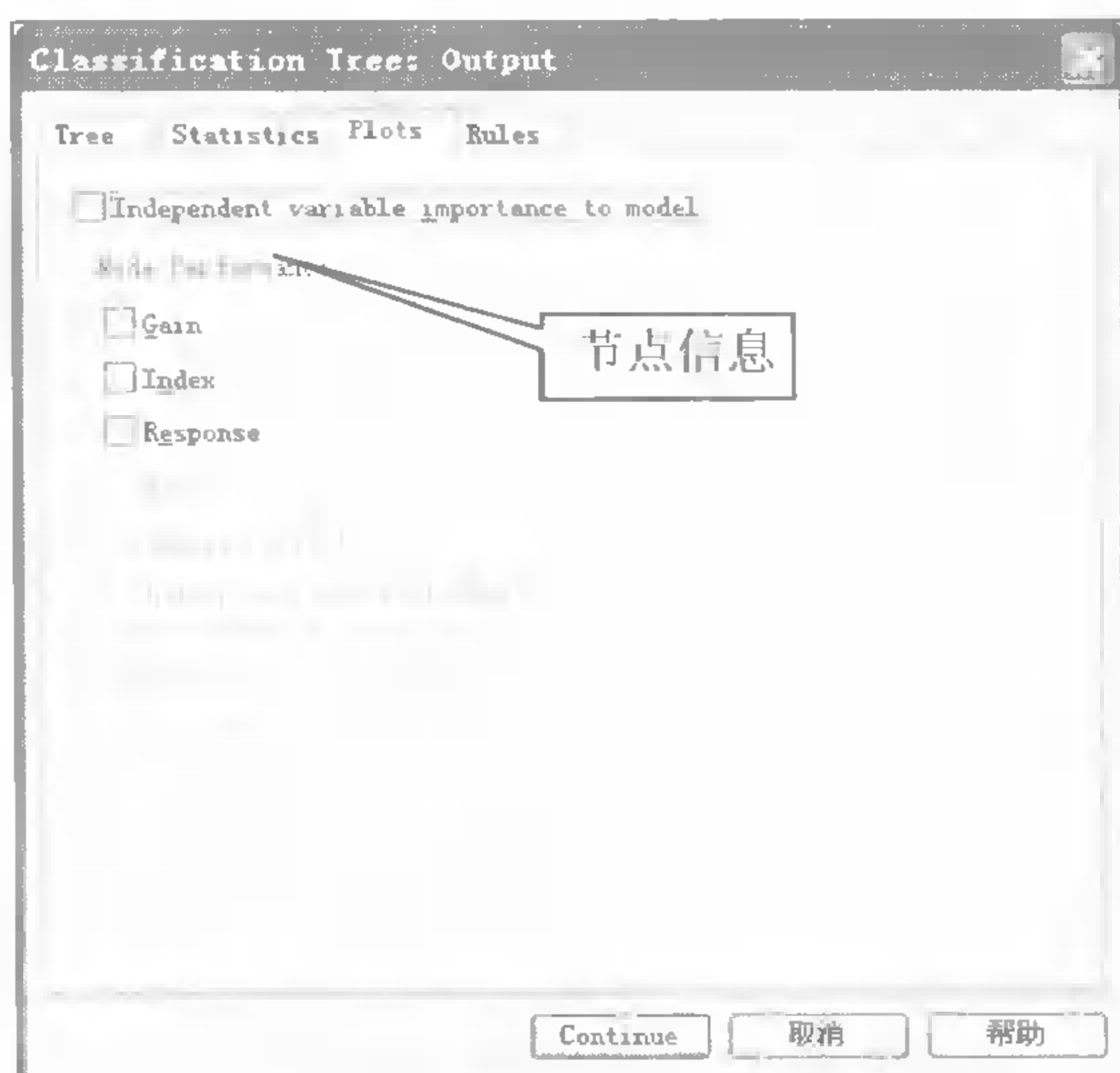


图 12-67 关于作图的输出设置

(1) Independent variable Importance to model 复选框。勾选后输出自变量对模型的重要性条形图, 只对 CRT 方法有效。

(2) Node Performance 栏, 设置关于节点的图形选项, 设置选项有如下 6 个。

- Gain 获利图, Gain 指每个节点中因变量目标取值所占的比例, 计算方式为  $(nd/nt) \times 100$ , 这里 nd 是单个节点的目标取值个数, nt 是总的目标取值个数。获利图就是对指定百分位点的 Gain 累积线型图; 对于因变量的每个目标取值, 单独作一个获利图; 此选项只对定义了目标取值的分类因变量起作用。
- Response 响应图, 作对指定百分位点的累积响应线性图。累积响应的计算方式为  $(tarn/alln) \times 100$ , 这里 tarn 是累积的目标取值个数, alln 是累积的总个数; 此选项只对定义了目标取值的分类因变量起作用。
- Index 指示图, 作对指定百分位点的累积指示线性图。累积指示指标的计算方式为  $(cprp/trp) \times 100$ , 这里 cprp 是累积响应比例, trp 是总样本的响应比例 (即目标取

值的比例); 此选项只对定义了目标取值的分类因变量起作用。

- ④ Mean 均值图, 作对指定百分位点的关于因变量的均值累积线型图, 此选项只对连续型的因变量有效。
- ⑤ Average profit 平均得益图, 作对累积平均得益的线型图, 此选项只对定义了得益函数的分类因变量有效。
- ⑥ Return on investment (ROI) 投资回报图, 作对累积 ROI 的线型图。ROI 指的是得益对开支的比率; 此选项只对定义了得益函数的分类因变量有效。

(3) Percentile increment 下拉列表。在此设置所有与百分位点有关的图形的百分位递增间隔, 可选项有 1%、2%、5%、10%、20%和 25%。

## 6. 输出选项的 Rules 设置

在图 12-65 中单击 Rules 标签, 弹出如图 12-68 所示的规则设置子面板。

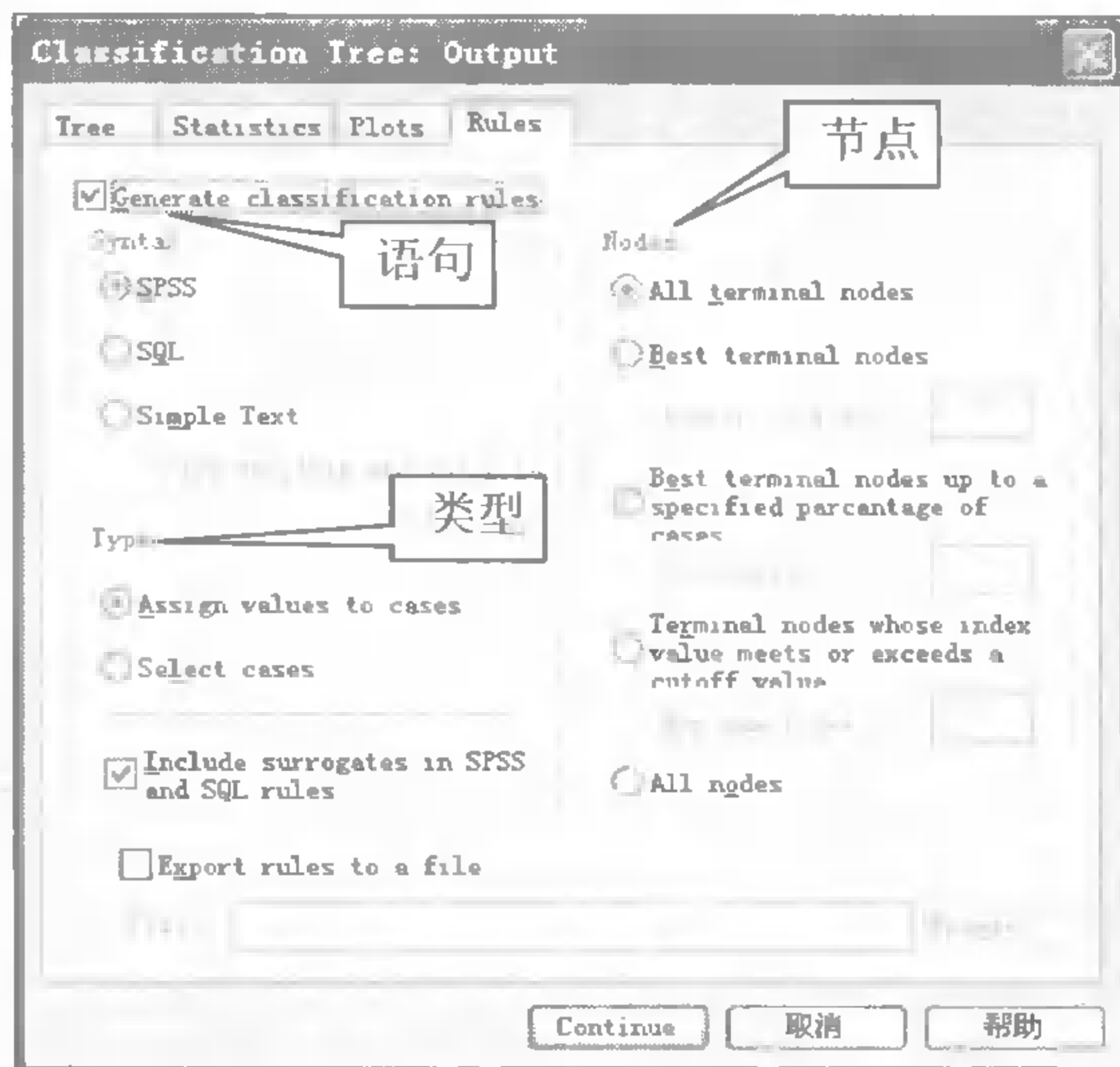


图 12-68 最终决策规则的输出格式设置

(1) Generate classification rules 复选框, 表示输出分类决策规则, 勾选后激活下面的选项。

① Syntax 栏, 设置关于决策规则的语句格式。

Syntax 决策规则可以输出到 SPSS Viewer 窗口或者保存到指定文件, 可选项有 3 个。

- ④ SPSS 选项, 输出 SPSS 命令语句 (Syntax), 决策规则主要以 filter 语句或 Compute 语句给出。通过统计得到的过滤变量, 可以对观测集进行分类和排序。
- ④ SQL 选项, 输出标准的 SQL 语句, 可用于对数据库中的记录进行筛选或赋值。
- ④ Simple text 选项, 简单文本输出, 格式为 “English pseudo-code”, 决策规则是由多组逻辑语句 “if...then” 来描述的。

② Type 栏, 设置关于 Syntax 和 SQL 格式的决策规则的类型, 有 3 个设置选项。

- ④ Select cases 选项, 生成一条符合节点规则的选择语句。
- ④ Assign values to cases 选项, 生成一条符合节点规则的赋值语句。
- ④ Include surrogates in SPSS and SQL rules 复选框。

对于 CRT 方法和 QUEST 方法, 勾选此项表示输出所有可能的替代方案的决策规则, 如



此有可能形成很复杂的规则；如果数据不完整（例如有缺失），可以勾选此项优化决策规则。

③ Nodes 栏，设置要输出决策规则的节点的范围。

SPSS 会对每个指定的节点输出一条决策规则，节点的选择范围有如下 5 个可选项。

- ① All terminal nodes 选项，对每个最终节点输出规则。
- ② Best terminal nodes 选项，对 index 值最好的前 n 个节点输出规则。Number of nodes 输入框用于指定 n 的数值；如果 n 大于最终节点的个数，就对所有最终节点进行输出。
- ③ Best terminal nodes up to a specified percentage of cases 选项，对 index 值最好的前百分之 n 个节点输出规则。Percentage 输入框用于指定 n 的数值。
- ④ Terminal nodes whose index value meets or exceeds a cutoff value 选项，对 index 值大于等于 n 的节点输出规则。Minimum 输入框用于指定 n 的数值。
- ⑤ All nodes 选项，对所有节点都输出决策规则。

(2) Export rules to a file 复选框。设置把决策规则输出至指定的文件，单击 Browse 按钮指定文件路径和文件名。

## 7. 验证选项的设置

在图 12-63 中单击 Validation 按钮，弹出如图 12-69 所示的验证设置子界面。通过检验验证，可以判断模型的稳定性和通用性。



图 12-69 关于模型验证的设置

① None，不进行验证。

② Crossvalidation 交叉验证。

它先把样本分为多个子样本（folds）；然后对每个子样本，用不包含它的其他数据建立一个决策树，再计算此决策树对这个子样本的错判率，以此验证决策树模型对样本的分类效果。此方法最终通过所有数据建立一个决策树模型，此模型的风险估计（错判水平）采用前面那些子模型风险的平均值。

Number of sample folds 输入框指定一个代表子样本个数的整数，不超过 25。

### ③ Split-Sample Validation 样本分离验证。

此方法将样本分为两个子集：训练样本和验证样本。用训练样本拟和决策树模型，用验证样本检验模型。样本分离法应慎重应用于小样本的情况，因为训练集太小时可能产生很差的决策树模型。可用的划分数据集方法有如下两种。

- ① Use random assignment 随机划分，Training sample 输入框用于指定训练集占总样本的比例；验证集占总样本的比例自动显示在 Test sample 后。
- ② Use variable 选项，通过指定变量划分数据集。

Variables 列表框显示可用的变量；Split sample by 选框用于选入划分数据集的变量。指定变量取值为 1 的样本设为训练集，取其他值的样本都设为验证集；不能指定因变量、权重变量、影响变量或强制进入模型的自变量为划分数据集的变量。

### ④ Display results for 栏，设置对哪些样本输出分析结果，有两个选择。

- ① Training and testing samples 选项，对训练集和验证集都输出相关的结果。
- ② Test sample only 选项，只对验证集输出有关结果。

## 8. 保存选项的设置

在图 12-63 中单击 Save 按钮，弹出如图 12-70 所示的保存设置子界面。

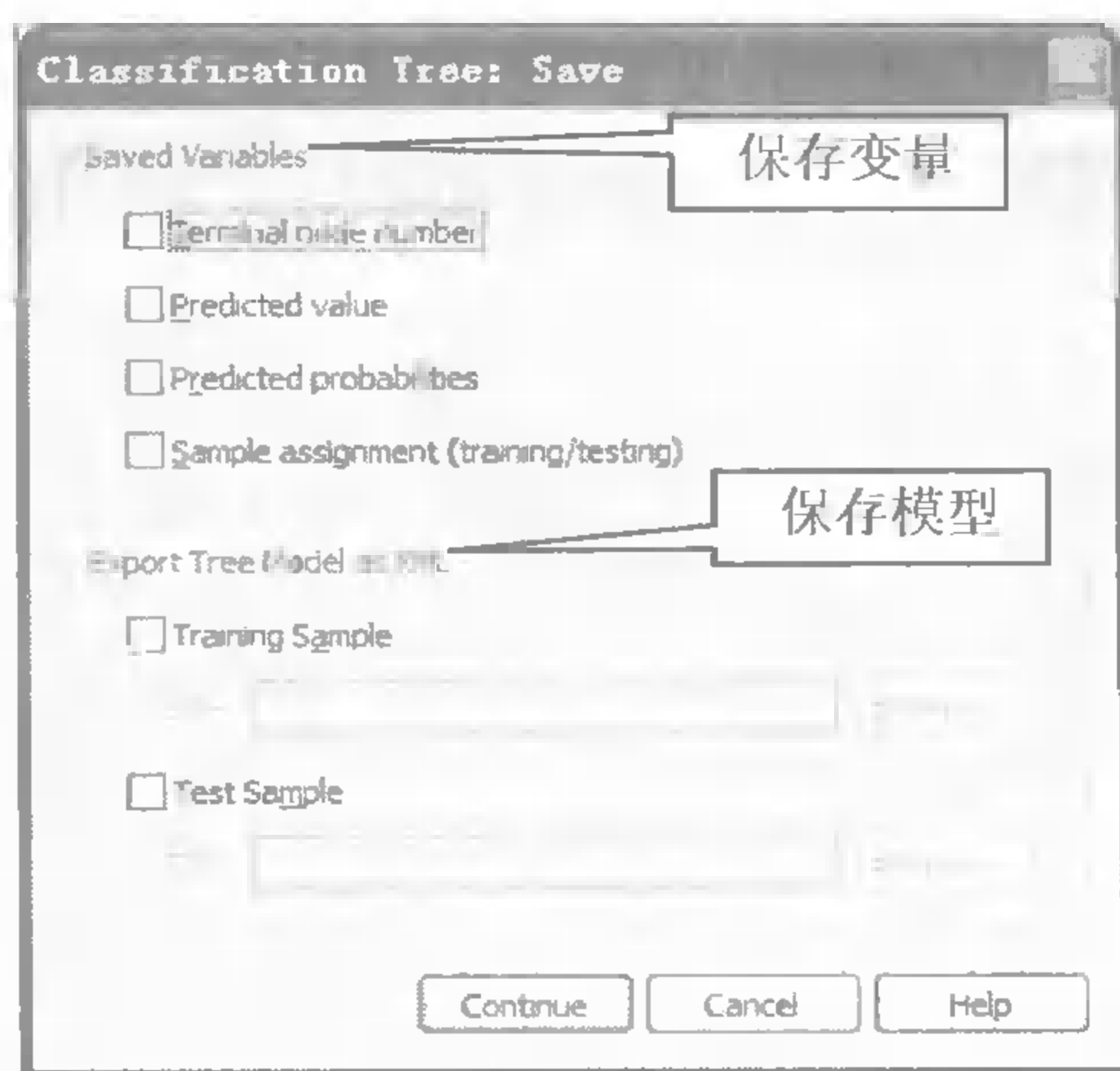


图 12-70 关于保存选项的设置

### ① Saved Variables 栏，设置保存哪些变量，可选内容有如下 4 个。

- ① Terminal node number 节点序号，此变量保存每个观测所属最终节点的序号。
- ② Predicted value 预测值，此变量保存由模型预测的因变量值。
- ③ Predicted probabilities 预测概率，对分类因变量，保存预测分类属于各类别的概率。因变量有几个类别就保存几个变量；对连续因变量不可用。
- ④ Sample assignment (training/testing) 样本类型，此变量记录单个观测是用于训练（取值 1）还是用于验证（取值 0）。只有选择样本分离验证法时此选项才可用。

### ② Export Tree Model as XML 栏，设置把模型格式输出到指定 XML 文件的选项。

在此可以分别设置对训练样本（Training sample）和验证样本（Test sample）的输出，单击其后的 Browse 按钮指定文件路径和文件名。

## 9. Growth Limits 参数的设置

在图 12-63 中单击 Criteria 按钮, 弹出如图 12-71 所示的算法参数设置子界面, 生长算法不同, 此处的设置内容也有所不同。下面先来介绍 CHAID 和 Exhaustive CHAID 算法的 Growth Limits 参数设置。

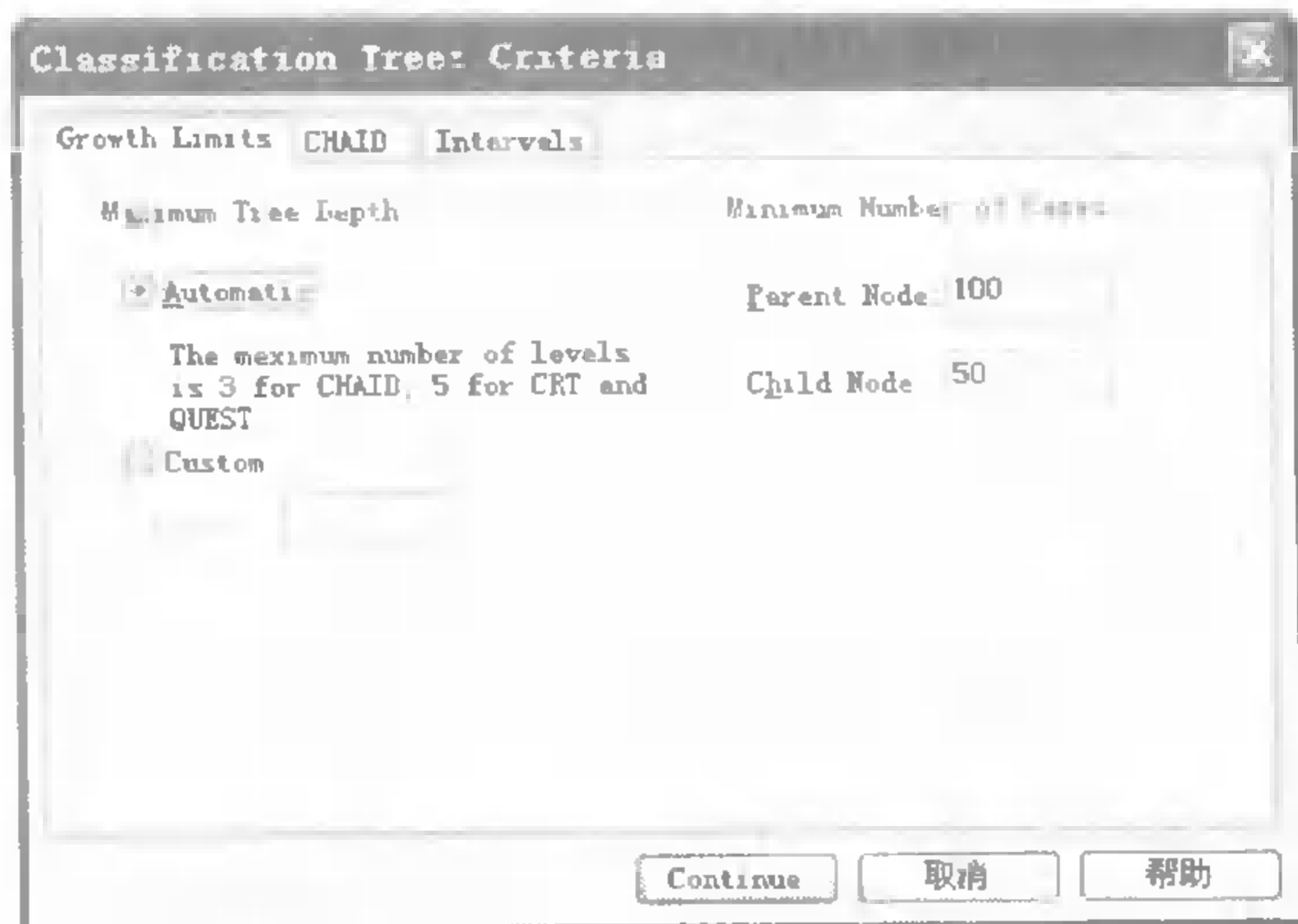


图 12-71 关于生长限制的设置

- ① Maximum Tree Depth 栏, 指定决策树在根节点以下的最大深度 (不含根节点)。
  - Automatic 自动, 对 CHAID 和 Exhaustive CHAID 算法, 最大深度为 3; 对 CRT 和 QUEST 算法, 最大深度为 5。
  - Custom 自定义, 在 Value 输入框指定一个代表最大深度的取值。
- ② Minimum Number of Cases 栏, 设置每个节点需要的最少观测个数。
  - Parent Node 输入框, 指定父节点需要的最少观测数, 默认值为 100。
  - Child Node 输入框, 指定子节点需要的最少观测数, 默认值为 50。

不符合对观测个数的限制的节点将不会被分支, 增大这两个最小值会使决策数的节点减少, 反之亦然。当样本较少时, 如果在此采用默认的最少个数, 可能会输出只有根节点的决策树, 此时输入较小的临界值才能得到更有意义的决策树。

## 10. CHAID 算法的参数设置

在图 12-71 中单击 CHAID 标签, 打开如图 12-72 所示的 CHAID 算法参数设置子界面。

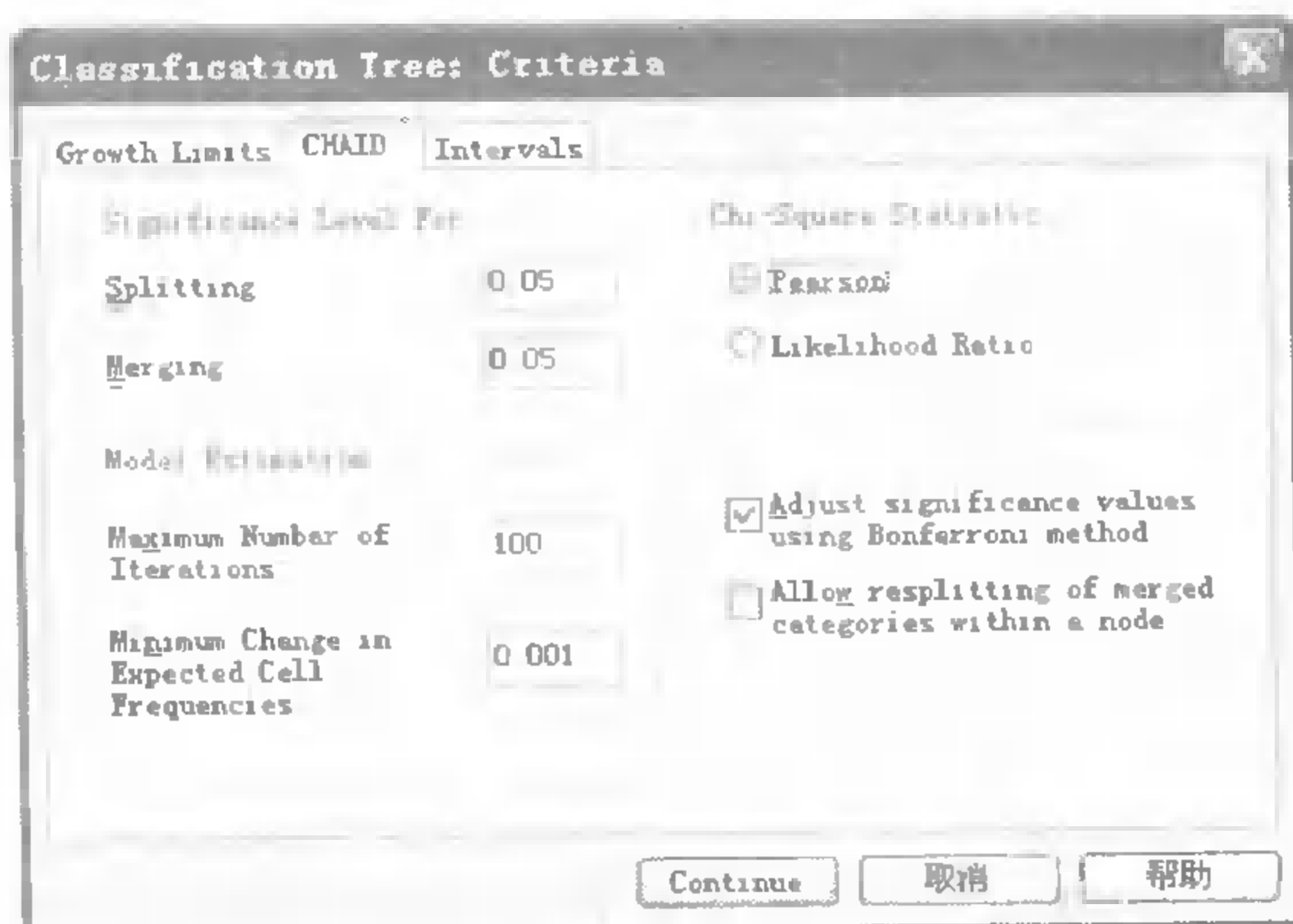


图 12-72 关于 CHAID 算法的设置

① Significance Level For 栏，有两个设置选项。

- Splitting nodes 输入框，指定分割节点的显著性水平临界值。默认值为 0.05，输入值须大于 0 小于 1，较小的临界值输出较少的节点。
- Merging categories 输入框，指定合并节点的显著性水平临界值。默认值为 0.05，输入值须大于 0 小于 1，设为 1 表示禁止节点合并。

② Chi-Square Statistic 栏，设置使用的卡方统计量。

- Pearson 卡方，计算速度快，但用于小样本时要谨慎考虑，是默认方法。
- Likelihood ratio 似然比卡方，比 Pearson 卡方统计量稳定，但计算较费时间，对于小样本此方法比较合适。

对于有序因变量 (ordinal)，合并和分割节点均使用似然比卡方统计量；对于名义因变量 (nominal)，可以在两者中指定 1 个。

③ Model Estimation 栏，对于分类因变量 (nominal 或 ordinal)，可设置如下两项参数。

- Maximum number of iterations 输入框，指定最大迭代次数，默认值为 100。如果决策树生长由于达到最大迭代次数而停止，可以在此输入更大的临界值。
- Minimum change in expected cell frequencies 输入框，指定单元格频数的最小改变量，默认值为 0.05，输入值必须大于 0 小于 1。此值越小，生成的节点越少。

④ Adjust significance values using Bonferroni method 复选框。

对于多重比较，用 Bonferroni 方法调整合并或分割节点时的显著性水平，默认为选中状态。

⑤ Allow resplitting of merged categories within a node 复选框。

除非指定了不进行节点合并，否则将对自变量取值进行可能的合并，以生成最简单的决策数。勾选此项，表示允许对合并的节点进行重新分割以生成更好的决策树。

## 11. CHAID 算法的 Intervals 设置

在图 12-71 中单击 Intervals 标签，打开如图 12-73 所示的 Intervals 参数子界面，在此设置对连续变量的离散化方式。

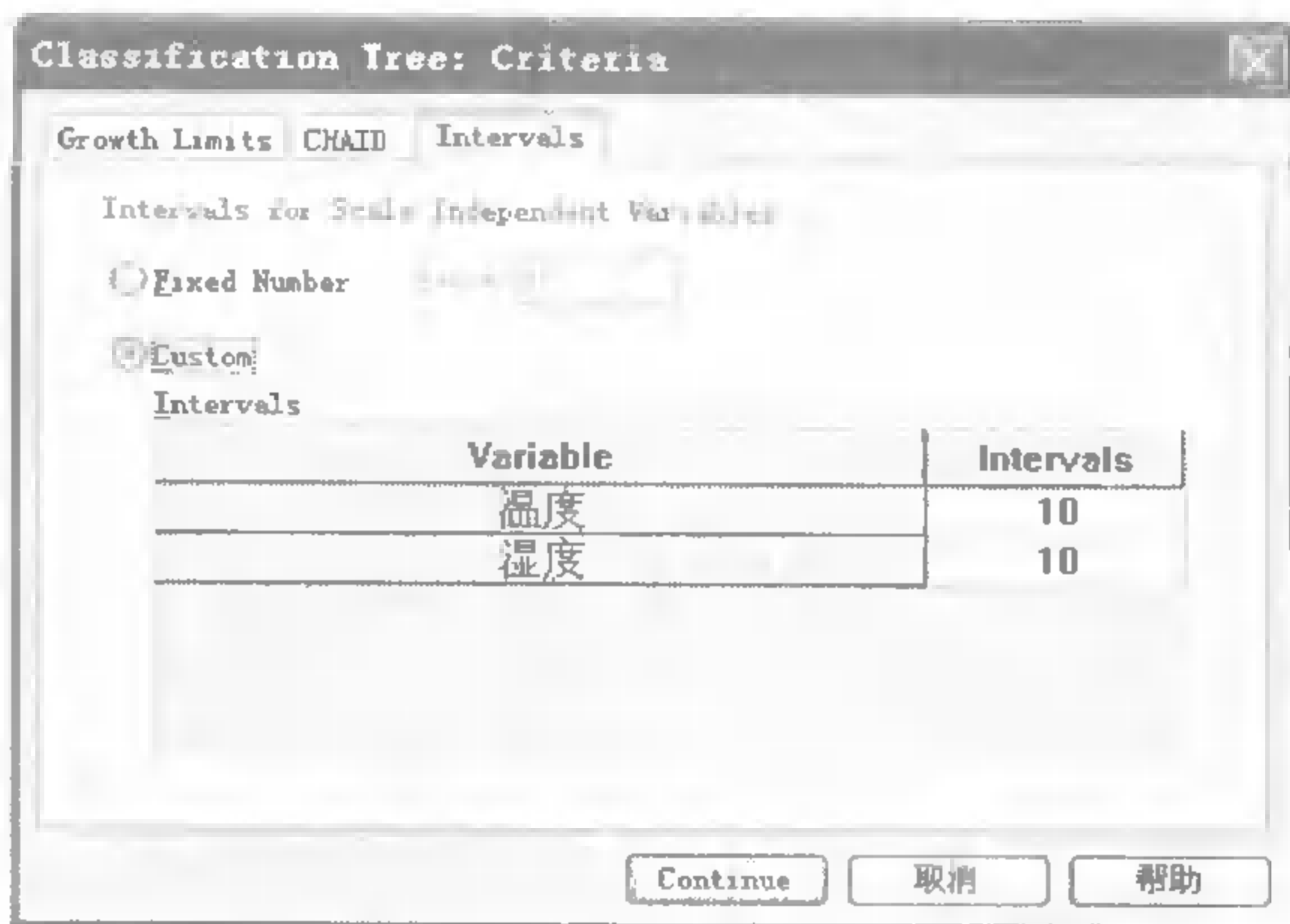


图 12-73 CHAID 算法的 Intervals 设置

在 CHAID 算法中，分析前要把连续自变量重新划分为离散的区间，例如 0-10、11-20、21-30 等；而在计算过程中，这些区间又可能被再度以不同的方式合并；Intervals 标签允许用



户设置连续自变量最初被离散化的分组个数（同时也是最大个数）。

① Fixed number 单选框，指定一个固定值。

Value 输入框用于指定区间个数，默认值为 10；所有连续自变量最初都被分为指定个数的区间。

② Custom 单选框，分别自定义每个变量的参数。

Custom 下的二维表格中，Variable 列显示当前可用的连续自变量；Intervals 列用于输入同行自变量的初始分组个数。

## 12. CRT 算法的参数设置

在图 12-63 中，单击 Growing Methods 下拉列表并选中 CRT 选项；单击 Criteria 按钮，弹出与图 12-71 相似的界面，再单击 CRT 标签打开如图 12-74 所示的 CRT 算法设置界面。

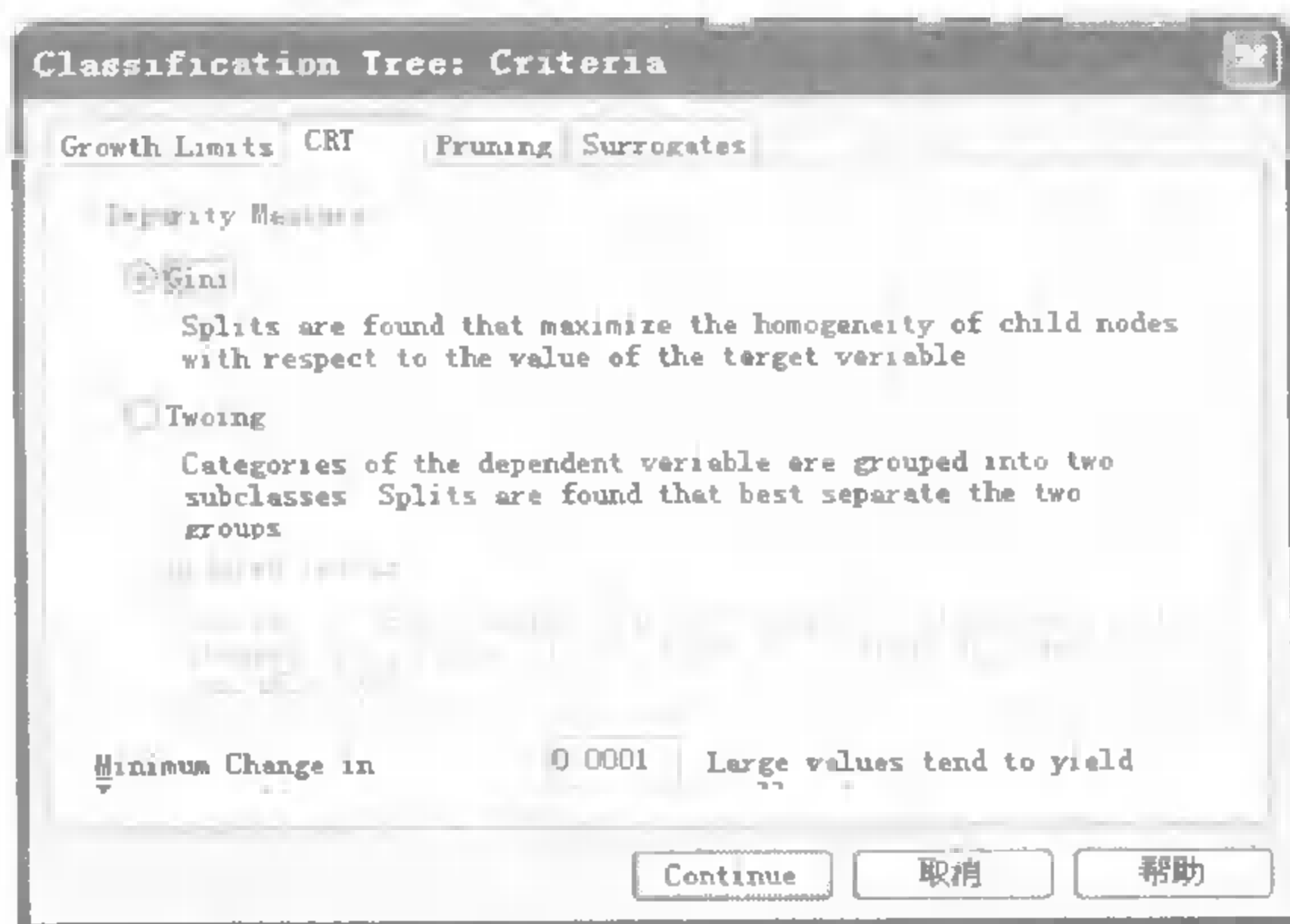


图 12-74 CRT 算法的标签设置

CRT 算法的 Criteria 设置面板有 4 个子标签，Growth Limits 子标签的设置与图 12-71 相同，此外还有 CRT、Pruning、Surrogates 三个子标签。

在图 12-74 中设置 CRT 算法关于分割节点的参数。

① Impurity Measure 栏，设置节点内部的不纯性（impurity）度量。

CRT 算法的基本思想是最大化节点内部的观测之间的相似性，并以此来判断节点是否要进一步分支；因变量取值全部相等的最终节点称为纯节点（pure），不再对其进行分割。对于连续因变量，不纯度以节点内的最小方差平方和（LSD）来度量，并用频率权重或影响变量加以调整；对于分类因变量，可选的不纯性度量方式有以下 3 个。

- Gini，寻求使子节点内部的因变量一致性达到最高的分支方法。它基于因变量各取值水平在节点内的出现比例的平方；对纯节点 Gini 取得最小值 0。是默认选项。
- Twoing，把因变量的取值水平被分为两个子集，寻求使这两个子集分得最开的方案。
- Ordered twoing 选项，与 Twoing 类似，但要求只有因变量相邻的取值才可以合并为一类，此选项只对有序因变量（ordinal）有效。

② Minimum change in improvement 输入框。指定分割一个节点所需要的最小不纯度减少值，默认值为 0.000 1；此临界值越大，输出决策树的节点越少。

### 13. QUEST 算法的参数设置

在图 12-63 中, 单击 Growing Methods 下拉列表并选中 QUEST 选项; 单击 Criteria 按钮, 弹出与图 12-71 相似的界面, 再单击 QUEST 标签打开如图 12-75 所示的 QUEST 设置界面。

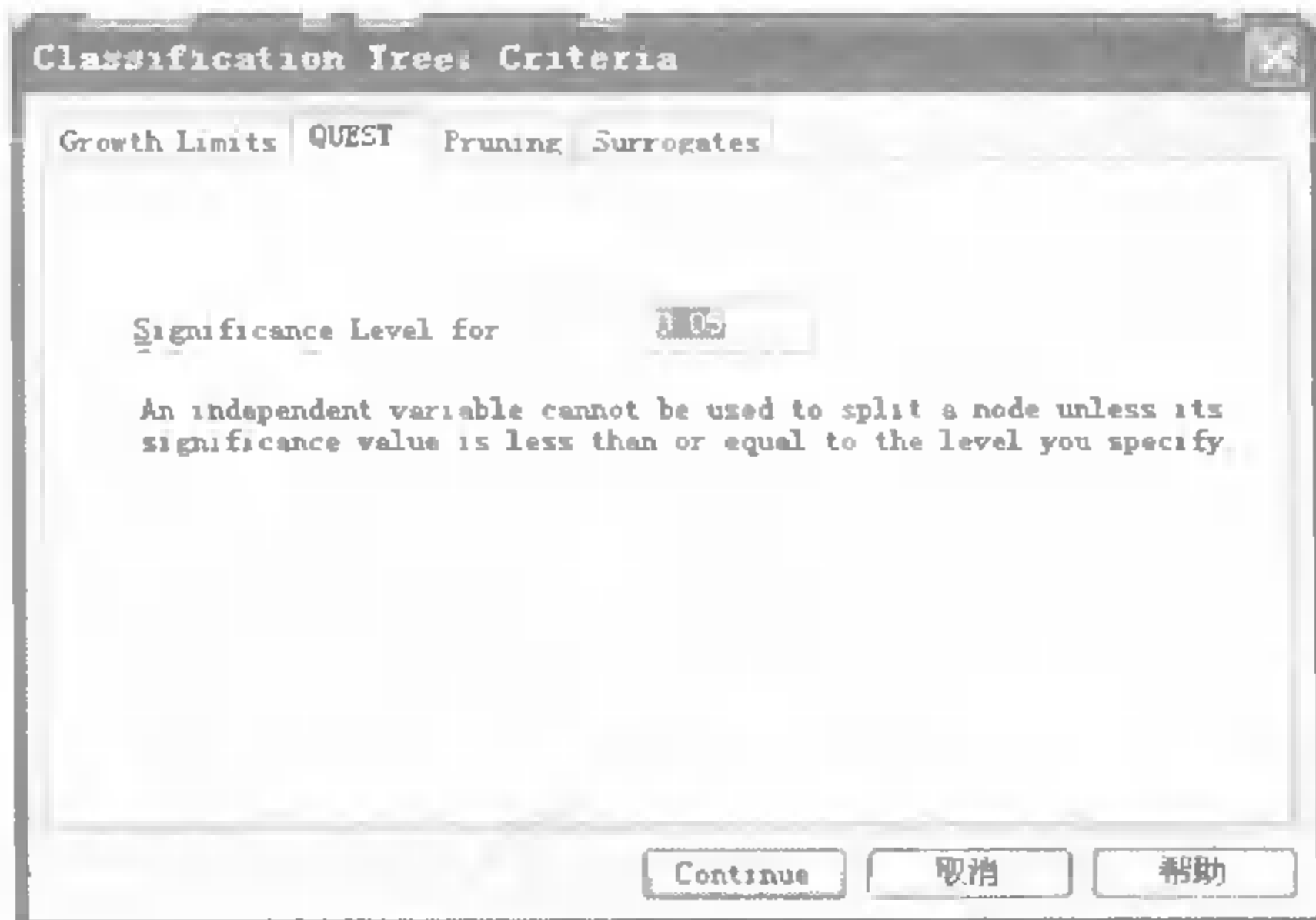


图 12-75 QUEST 算法的标签设置

QUEST 算法的 Criteria 设置面板有 4 个子标签, 且 Growth Limits 子标签的设置与图 12-71 相同, 此外还有 QUEST、Pruning、Surrogates 三个子标签。

在图 12-75 中设置 QUEST 算法关于分割节点的显著性水平临界值。

Significance Level for 输入框, 指定一个大于 0 小于 1 的数值, 默认值为 0.05。只有自变量重要性检验的显著性 P 值小于此处的临界值时, 才有可能用它来分割节点; 越小的显著性水平临界值, 输出决策树所包含的自变量越少。

### 14. CRT 和 QUEST 算法的 Pruning 设置

在图 12-74 (或者图 12-75) 中单击 Pruning 标签, 打开如图 12-76 所示的子标签界面, 在此设置 CRT (或 QUEST) 算法关于剪枝的参数。

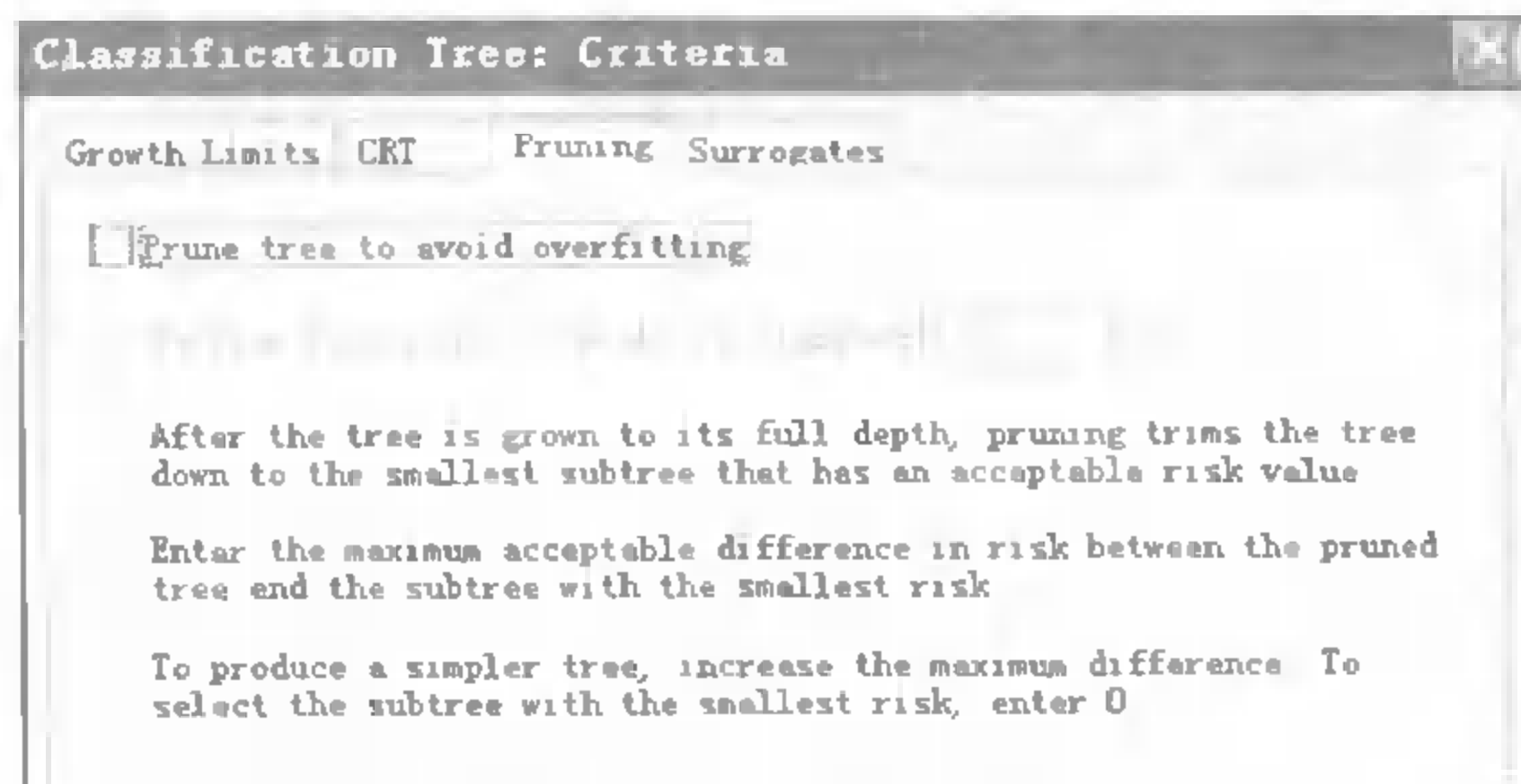


图 12-76 Pruning 标签的设置

① Pruning tree to avoid overfitting 复选框。勾选它表示决策树长满后, 还要对其进行剪枝以避免生长过渡。

② Maximum difference in risk 输入框。用于指定决策树被剪枝前后所允许的风险 (risk) 值的最大差额。它以标准误差的方式表示, 默认值为 1; 增大此值, 将生成更小的决策树; 设为 0, 将输出风险最小的决策树。

## 15. CRT 和 QUEST 算法的 Surrogates 设置

在图 12-74（或者图 12-75）中单击 Surrogates 标签，打开如图 12-77 所示的子标签界面，在此设置 CRT（或 QUEST）算法关于备选方案的参数。所谓备选，指当前用于分支的自变量取缺失值时，可以用其他相关的变量代替（Surrogate）它来进行分支。

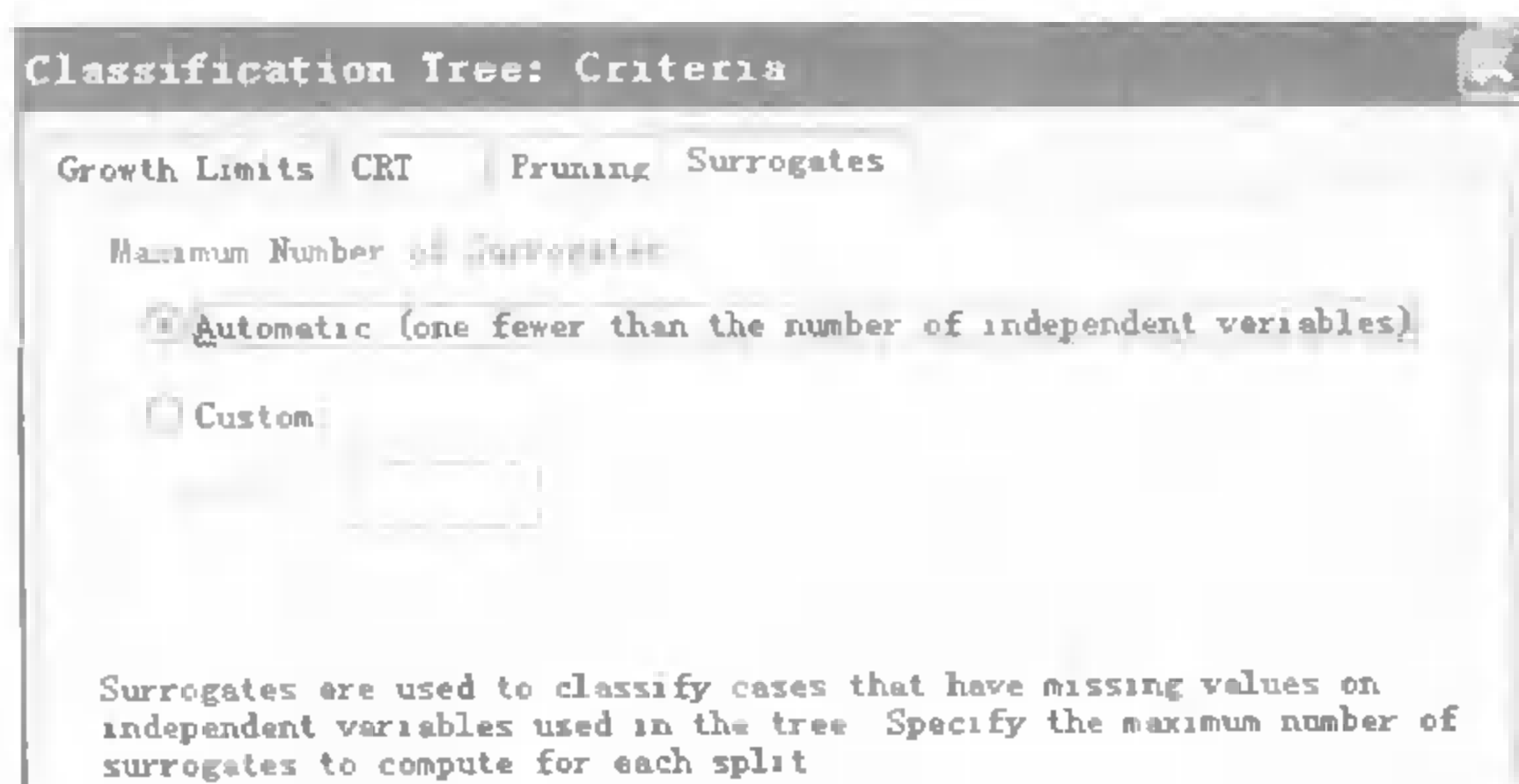


图 12-77 Surrogates 标签的设置

Maximum Number of Surrogates 子设置栏，指定模型中允许使用的备选自变量的最大个数，有如下两种限制方法可选。

① Automatic 自动，是默认方式。备选自变量的最大个数比所有自变量的个数少 1；也就是说，对任一自变量，可以用任意其他变量作为备选方案。

② Custom 自定义，在 Value 后输入备选自变量的最大个数，0 表示不使用备选自变量。

## 16. 关于缺失值的设置

在图 12-63 中单击 Options 按钮，打开如图 12-78 所示的子标签面板，设置关于决策树分析的其他参数。Options 设置面板有 4 个子标签，下面先来看 Missing Values 的设置。

User-Missing Values of Nominal Independent Variables 栏，设置对名义自变量（Nominal）的用户定义缺失值如何处理，有 2 种方法：Treat as missing values 单选项，当作系统缺失值处理；Treat as valid values 单选项，当作正常的有效值处理。

另外，决策树生长算法不同，对缺失值的处理方式也有所不同。

对 CHAID 和 Exhaustive CHAID 算法，自变量的系统缺失值 and 用户缺失值都被当作单独的一类进行处理。对于连续自变量和有序自变量，这两个算法先使用有效数据进行分类；然后判断指定变量含缺失值的观测能否分入离它们最近的类，或者把这些观测归为单独的一类。

对 CRT 和 QUEST 算法，自变量含有缺失值的观测将不用于决策树的生长，如果指定了可以使用备选变量，就使用备选变量对其进行归类。

## 17. 惩罚函数的设置

在图 12-78 中单击 Misclassification Costs 标签，打开如图 12-79 所示的子标签面板，在此设置关于错判惩罚函数的参数。对于分类因变量，可用惩罚函数指定归类错误时的惩罚力度，具体设置内容如下。

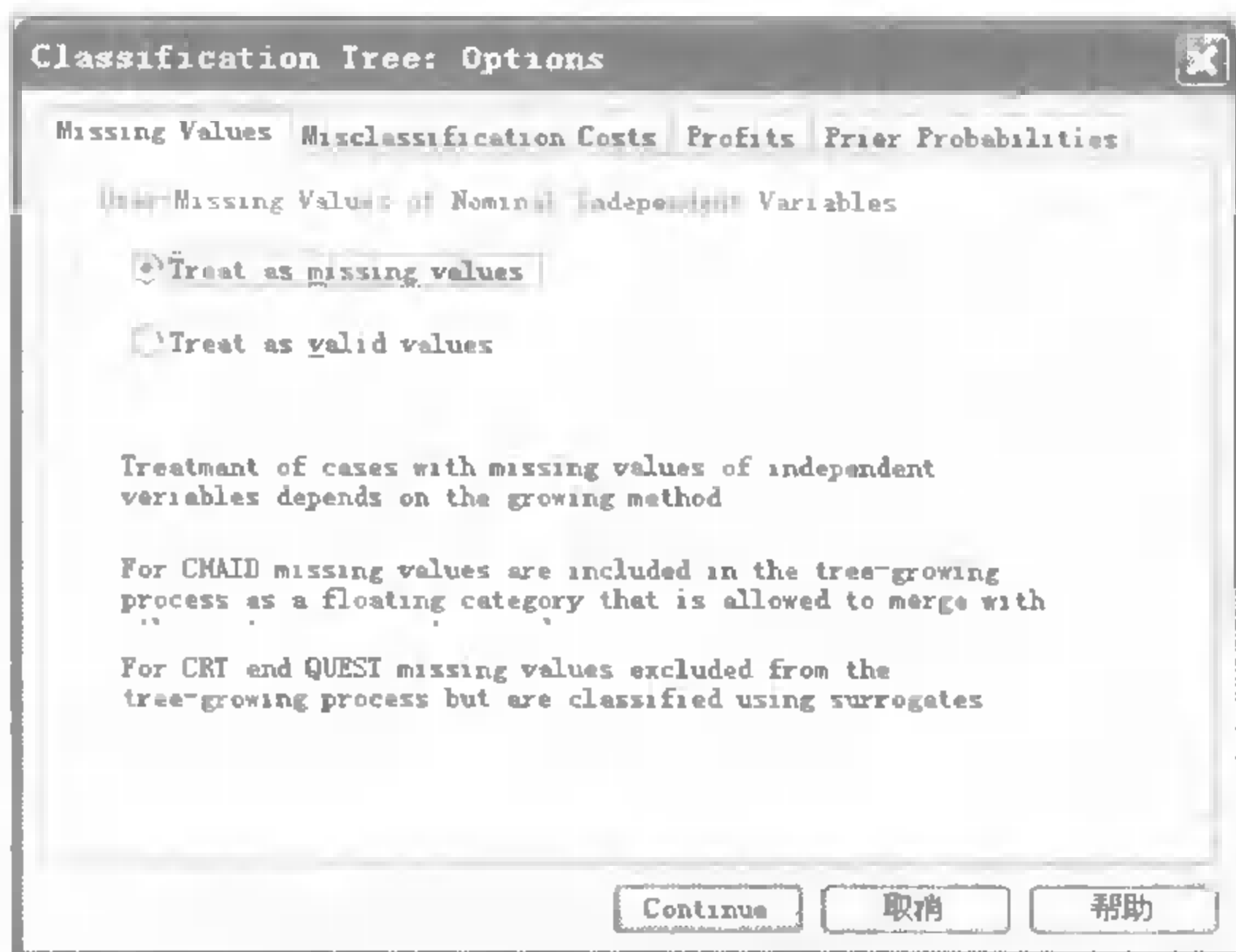


图 12-78 缺失值的处理设置

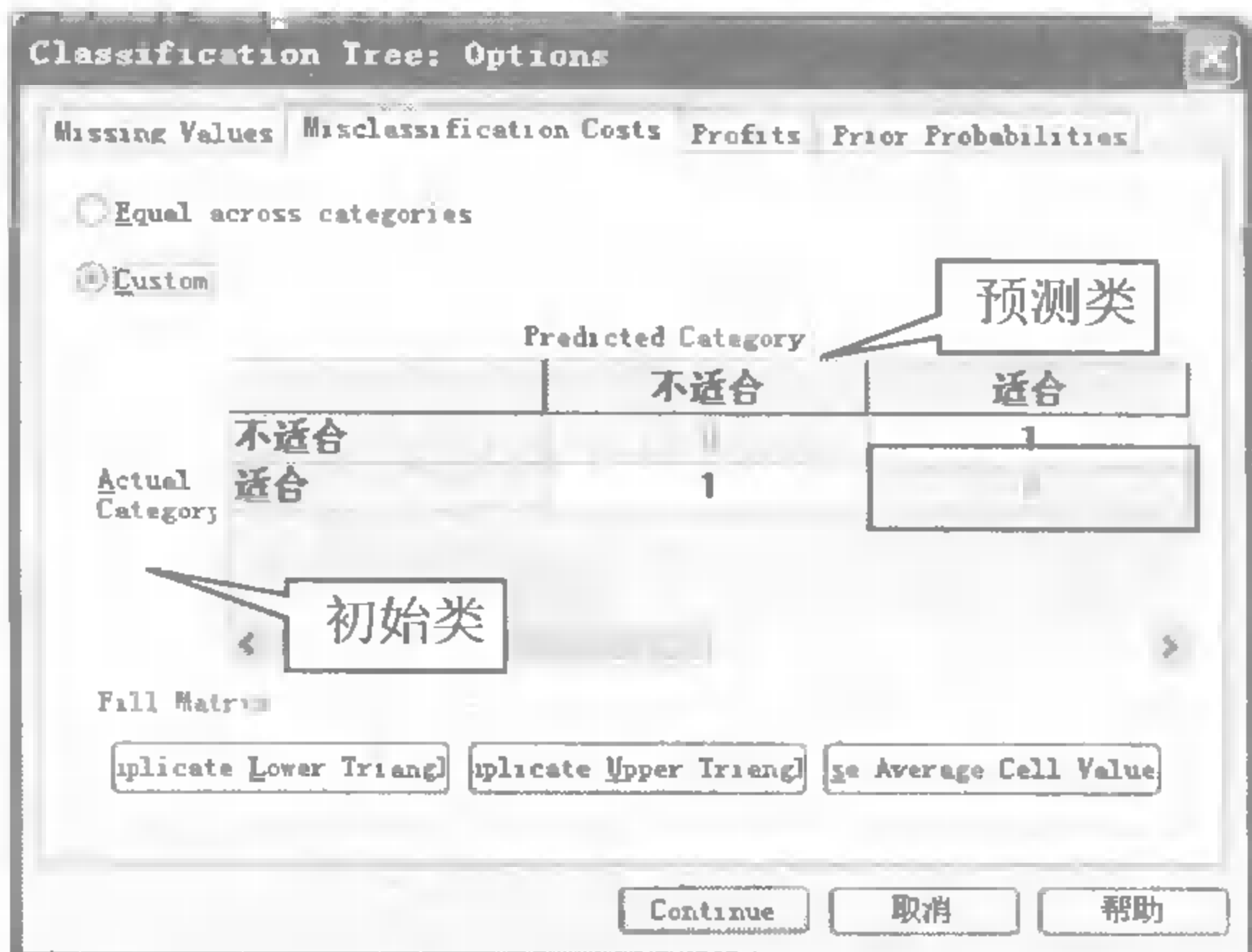


图 12-79 错判惩罚函数的设置

① Equal across categories 选项，表示对各种错判分类的惩罚都一样。

② Custom 选项，由用户自定义错判惩罚函数。

只有当分类因变量至少设置了两个值标签时，此选项才可用。在下面的二维表格里设置具体的惩罚措施。行表示初始分类，列表示预测分类；对角线上是正确预测的情况，惩罚都为 0 且不可编辑；其他单元格都是对预测错误的惩罚，例如以蓝色线框标识的单元格表示把不适合预测为适合的惩罚值为 1，用户指定的惩罚值必须非负。

③ Fill Matrix 栏有 3 个按钮，用以方便地设置如何使惩罚矩阵成为对称的形式。

- Duplicate Lower Triangle 按钮，把矩阵的下三角复制到上三角使之对称。
- Duplicate Upper Triangle 按钮，把矩阵的上三角复制到下三角使之对称。
- Use Average Cell Values 按钮，计算任意两个对称单元格的算术平均值并取代它们，使矩阵对称。

单击如上 3 个按钮中的任一个，都会弹出确认使用的提示框，在提示框里单击“确定”按钮即可应用所选择的方法。

## 18. 收益函数的设置

在图 12-78 中单击 Profits 标签，打开如图 12-80 所示的子标签面板，在此设置预测分类正确时的收益函数的参数。

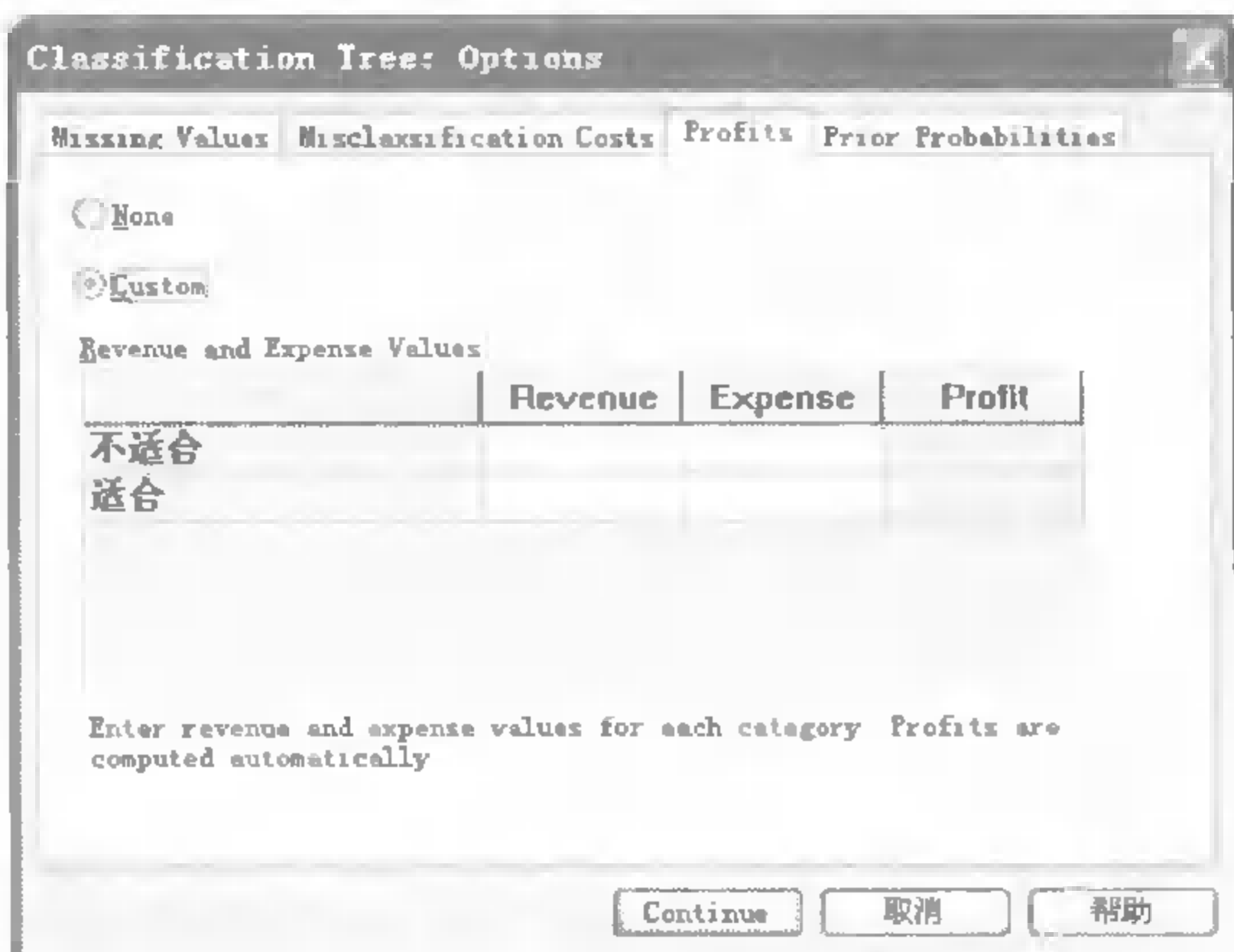


图 12-80 收益和开销函数的设置



① None 选项，表示不使用收益函数。

② Custom 选项，表示由用户自定义收益函数，选中后激活下面的选项。

只有当分类因变量至少设置了两个值标签时，此选项才可用。在下面的二维表格里设置具体的收益取值：第一列显示了当前分类因变量的取值标签；Revenue 列输入对当前行的值标签预测正确时的收入值；Expense 列输入对当前行的值标签预测正确时的开销值；Profit 列表示收益值，自动由公式“Revenue-Expense”计算得出。

选择 Custom 选项后，必须用数值填满所有可编辑的单元格。Profit 值会影响得益表格中的平均收益值和 ROI 值，但不会影响决策树的基本结构。

## 19. 先验概率的设置

在图 12-78 中单击 Priors Probabilities 标签，打开如图 12-81 所示的子标签面板，在此设置关于先验概率的有关参数。对于 CRT 和 QUEST 算法，如果因变量是分类变量，且至少定义了两个值标签，才可以指定其类别取值的先验概率。

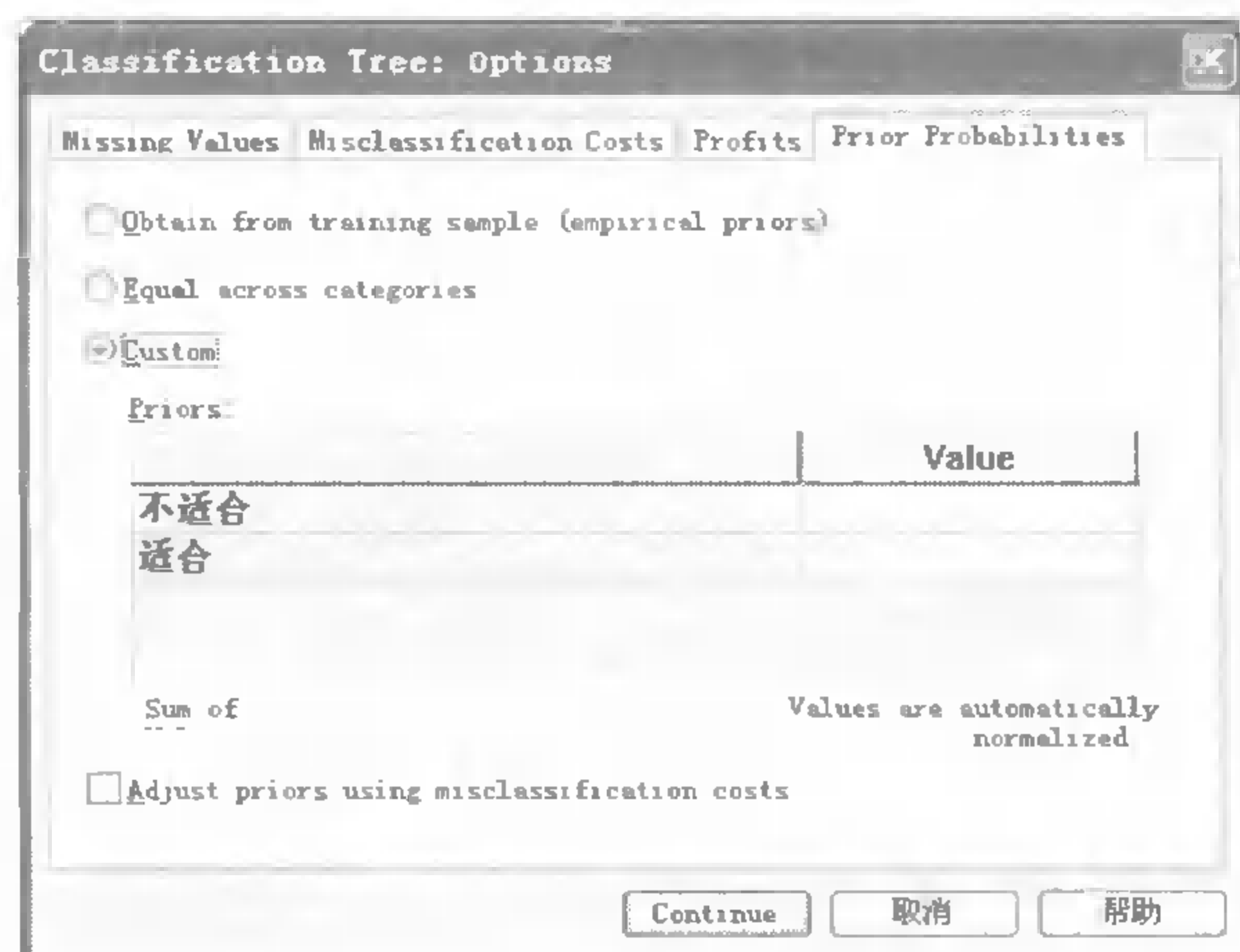


图 12-81 先验概率的设置

单纯由训练样本集出发，所建决策树有时不能很好地反映总体特征，使用先验概率可以对其加以修正。具体设置内容有如下 4 个。

① Obtain from training sample (empirical priors) 单选框，指定经验先验概率。

从训练样本集获得先验概率，当样本能很好的代表总体时选中此项；当采用样本分离法进行验证时，由于训练集是随机抽取的，所以事先也不知道训练集的分布情况；此为默认选项。

② Equal across categories 单选框，指定等先验概率。

当因变量各取值水平所占的比例都很相近时，选中此项。

③ Custom 单选框，由用户自定义先验概率。

在下面的二维表格中，Value 列用于输入当前行标签所对应的先验概率。输入可以是比例、百分比、频数等任何能反映类别取值分布的数值。

④ Adjust priors using misclassification costs 复选框如果定义了错判惩罚函数，可以选中此项，表示用错判矩阵对先验概率进行调整。

### 12.7.3 问题描述和数据准备

对于热爱运动的人来说，天气情况无疑是他们最为关注的信息，下面采集了多个采访对象对在 31 种天气情况下是否适宜运动的看法，这里的天气情况包括天气（晴天、阴天还是有雨）、温度、湿度和是否有风 4 个方面。本节希望通过决策树分析建立一个由这 4 个因素的取值来判断是否适宜运动的规则，作为对其他喜爱运动的人士的建议。

所用数据文件为“是否适宜运动的天气数据.sav”，数据格式如图 12-82 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	weth	String	4		天气	None	None	4	Left	Nominal
2	temp	Numeric	11	0	温度	None	None	10	Right	Scale
3	humi	Numeric	11	0	湿度	None	None	6	Right	Scale
4	wind	String	2		风况	None	None	6	Left	Nominal
5	sport	Numeric	6	0	运动	{0 不适合}	None	6	Right	Nominal

图 12-82 关于运动的天气数据格式



### 12.7.4 案例分析

依次单击菜单“Analyze→Classify→Tree...”打开决策树分析过程的主设置界面，如图 12-83 所示。



图 12-83 决策树分析的主设置界面

#### 1. 参数设置

(1) 指定分析变量。在变量列表中选中运动变量，单击从上至下第一个  按钮，将其作为因变量选入 Dependent Variable 选框；在变量列表中选中从天气到风况的 4 个变量，单击从上至下第二个  按钮，将其作为自变量选入 Independent Variables 列表框；保留 Growing Methods 下拉列表的 CHAID 选项。

(2) 定义目标取值。在图 12-83 中单击选中 Dependent Variable 选框的运动变量，然后单击 Categories 按钮，弹出如图 12-84 所示的因变量目标取值定义对话框。勾选“适合”右侧的复选框；单击 Continue 按钮返回主界面。



图 12-84 因变量目标取值的设置

(3) 验证参数的设置。在图 12-83 中单击 Validation 按钮，弹出如图 12-85 所示的验证参数设置界面。单击选中 Split-sample validation 单选框；单击选中 Use random assignment 单选框，在 Training sample (%) 后输入“80.00”；单击 Continue 按钮返回主界面。

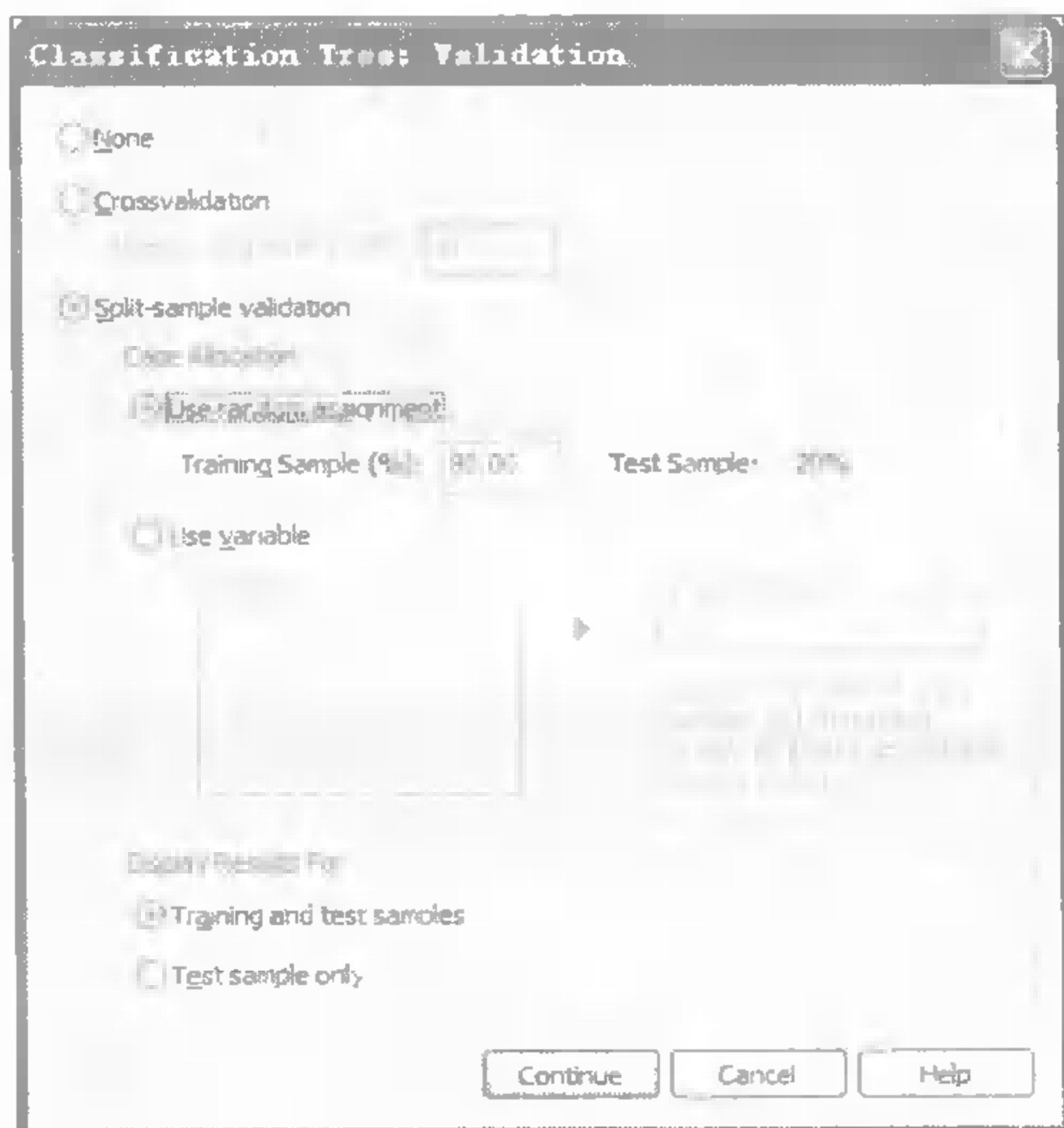


图 12-85 验证参数的设置

(4) 生长参数和 Intervals 参数的设置。在图 12-83 中，单击 Criteria 按钮，弹出如图 12-86 所示的生长参数设置界面。在 Parent Node 后输入“5”，在 Child Node 后输入“2”。

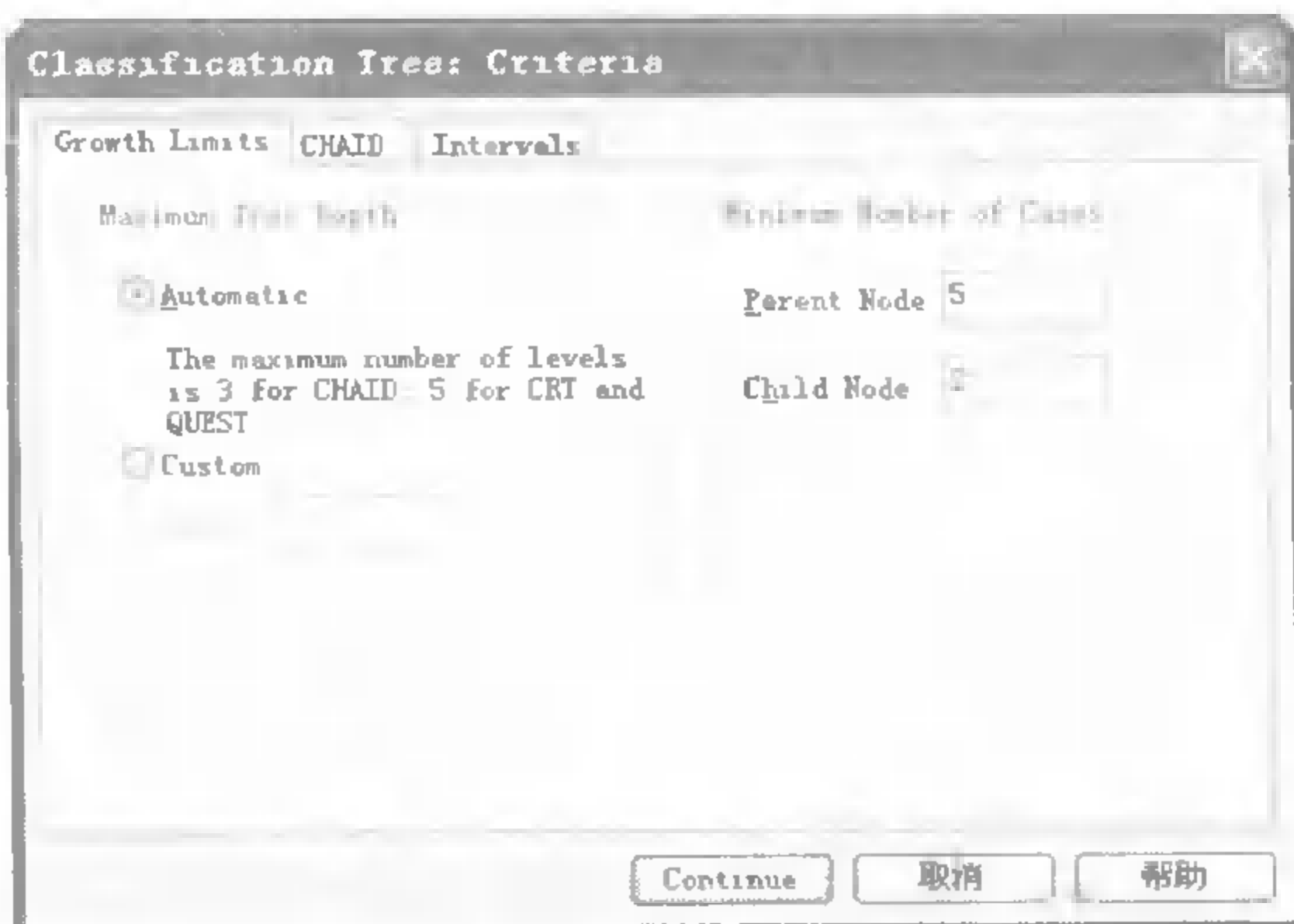


图 12-86 生长参数的设置

在图 12-86 中单击 Intervals 标签，弹出如图 12-87 所示的 Intervals 参数设置界面。单击选中 Fixed number 单选框，在 Value 后输入“20”；单击 Continue 按钮返回主界面。

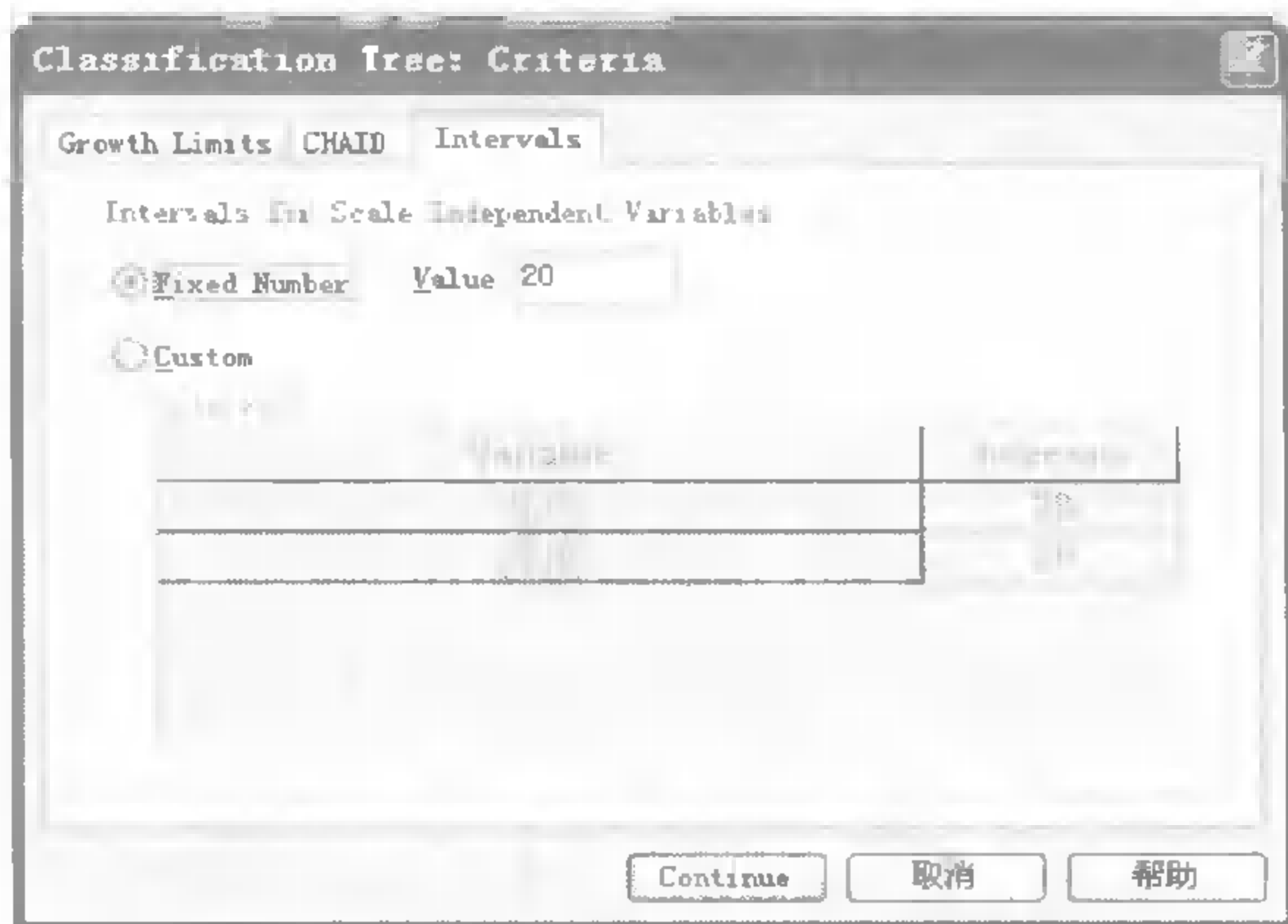


图 12-87 Intervals 设置

(5) 惩罚函数的设置。在图 12-83 中单击 Options 按钮，打开如图 12-78 所示的设置面板，单击 Misclassification Costs 标签，打开如图 12-88 所示的惩罚函数子标签面板。单击选中 Custom 单选框，在“适合”行与“不适合”列的交叉单元格键入“0.8”；单击 Continue 按钮返回主界面。这表示把适合运动的天气判断为不适合运动的错判损失，要比把不适合运动判断为适合运动的错判损失稍小。

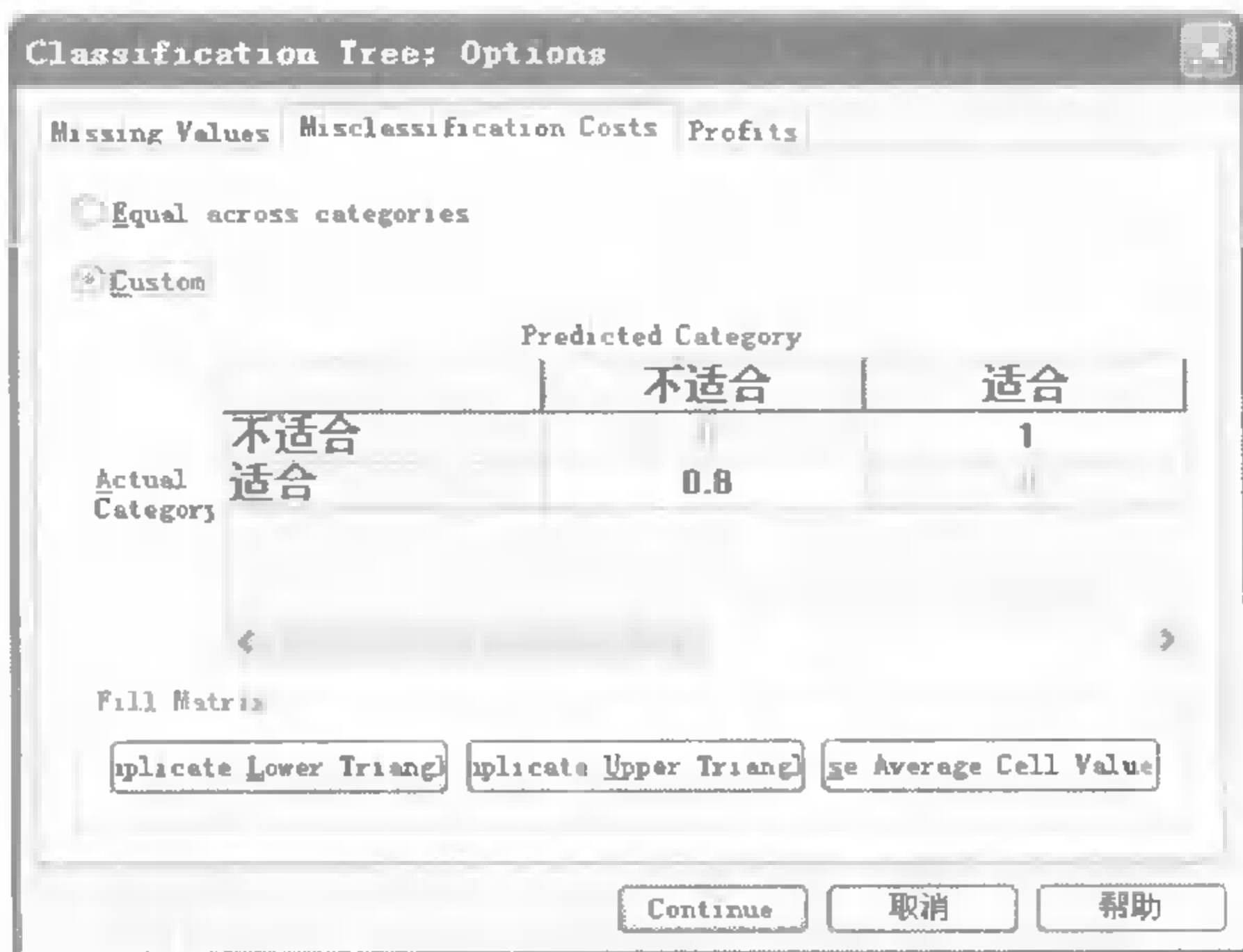


图 12-88 惩罚函数的设置

## 2. 结果分析

在图 12-83 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如下。

(1) 决策树的模型概要信息。如图 12-89 所示，是关于模型汇总的信息，包括因变量、自变量、生长方法和验证方法等；还给出了最终输出的决策树模型的基本信息，包括用到的自变量、节点数、最终节点数和决策树深度。



模型汇总		
指定	增长方法	CHAID
	因变量	运动
	自变量	天气 温度 湿度 风况
	验证	拆分样本
	最大树深度	3
	父节点中的最小个案	5
	子节点中的最小个案	2
结果	自变量已包括	天气 湿度
	节点数	5
	终端节点数	3
	深度	2

图 12-89 决策树的模型汇总表

(2) 图形决策树的输出。如图 12-90 所示，是关于训练样本集（随机抽取的 80%）的图形决策树分类示意图。

如图 12-91 所示，是关于验证样本集（20%）的图形决策树分类示意图。

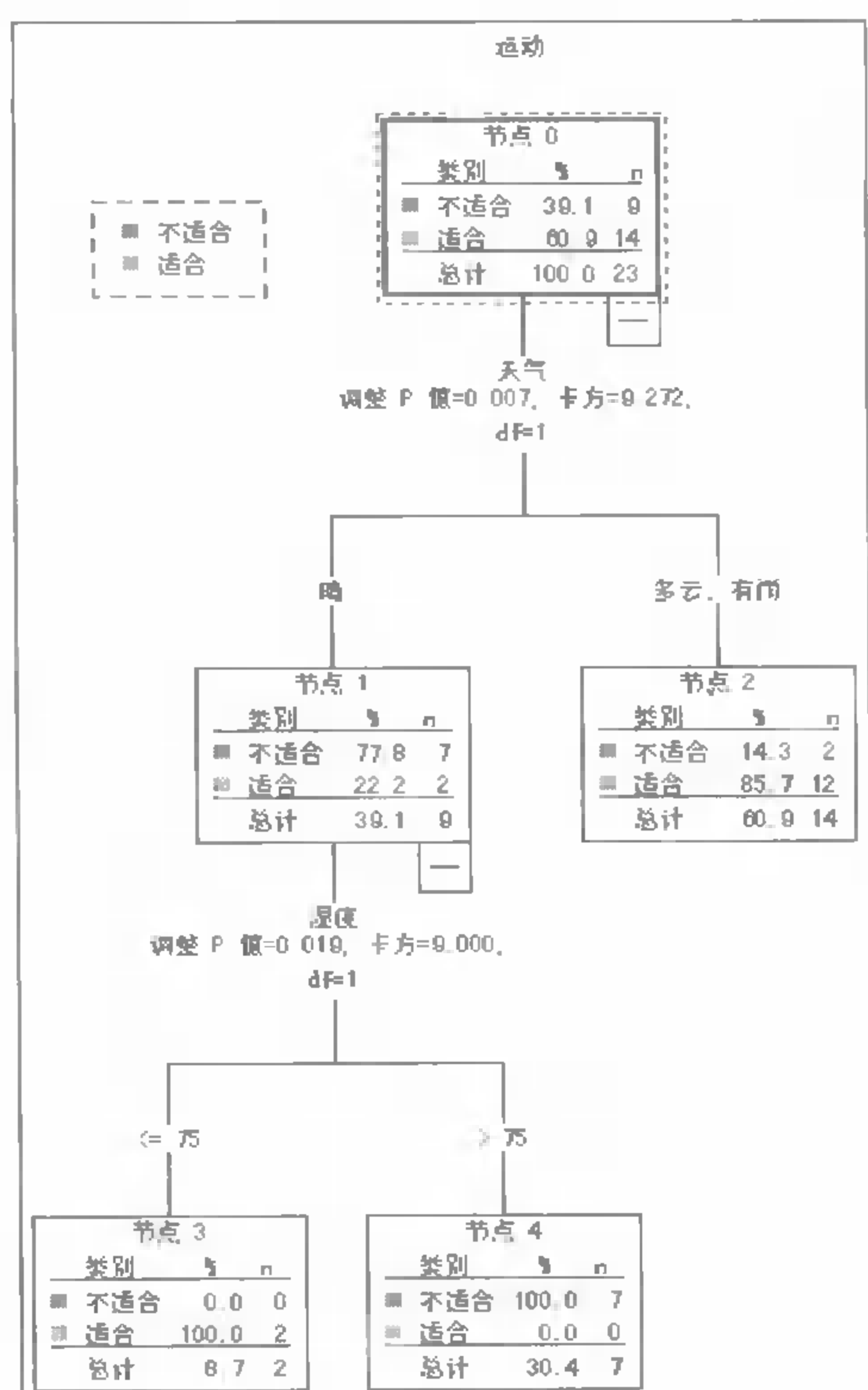


图 12-90 训练样本的决策树输出

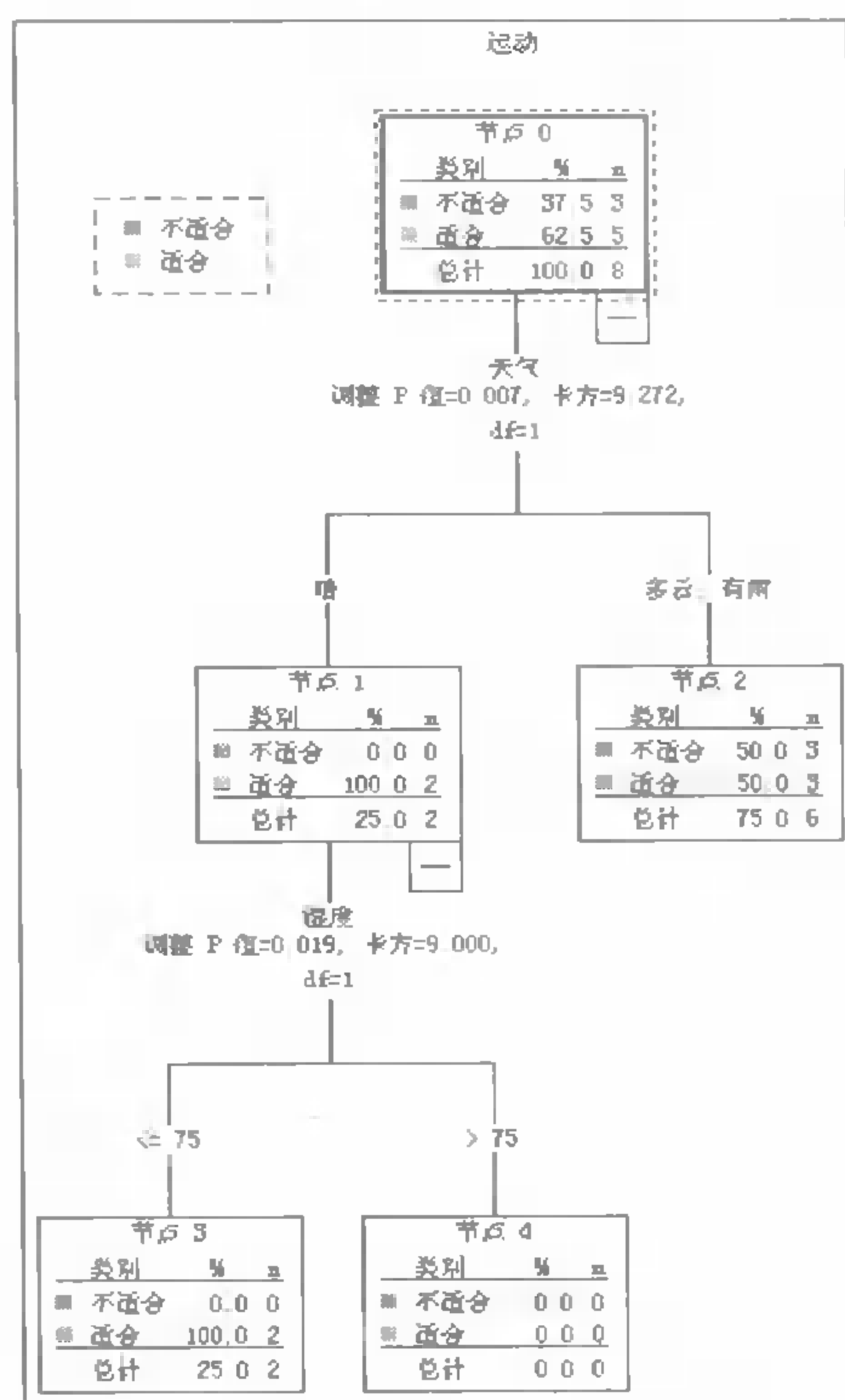


图 12-91 验证样本的决策树输出

### ① 决策树图形的解释

由图 12-90 可知，三个最终节点分别为节点 2、节点 3 和节点 4，其中节点 3 和 4 都是纯节点（因变量只有 1 个取值），节点 2 的预测准确率最高也能达到 85.7%。

再来看图 12-91 所示的验证样本集，纯节点 3 依然保持了较好的预测能力；节点 2 的预测能力有所下降，50%的判断准确率并不能给决策者提供更多有意义的信息。

### ② 决策树图形的编辑

由于设置了 SPSS Viewer 窗口显示中文，这可能使决策树图形不能正常显示，更正方法为在决策树图形上双击，打开如图 12-92 所示的 Tree Editor 窗口；在图 12-92 中双击任意节点，弹出如图 12-93 所示的字体设置对话框，在此字体 (Font) 选择黑体、大小 (Size) 选择

10, 单击 Apply 按钮应用设置, 再单击 Close 按钮关闭此对话框, 就可以正常显示中文了。

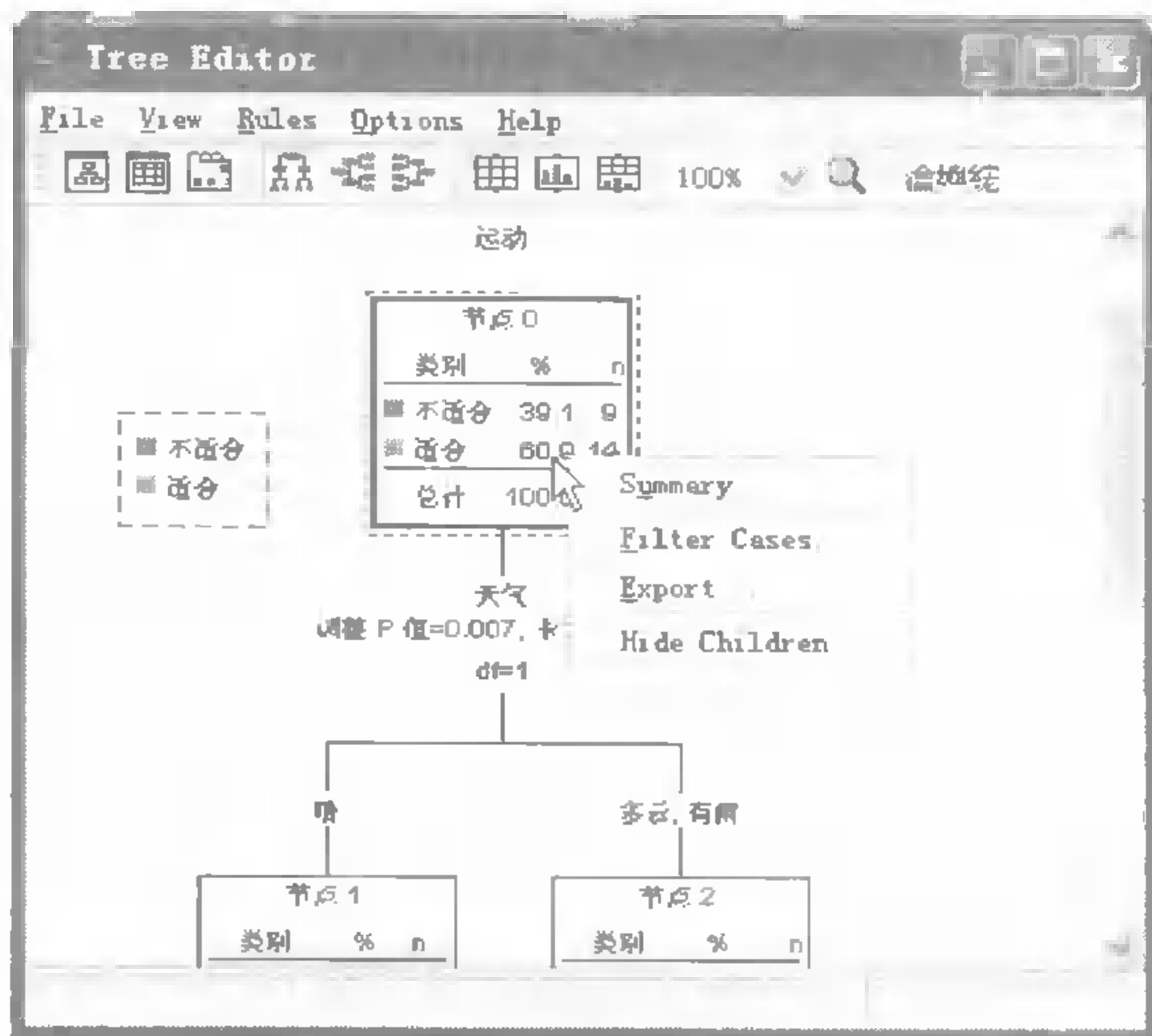


图 12-92 Tree Editor 编辑窗口

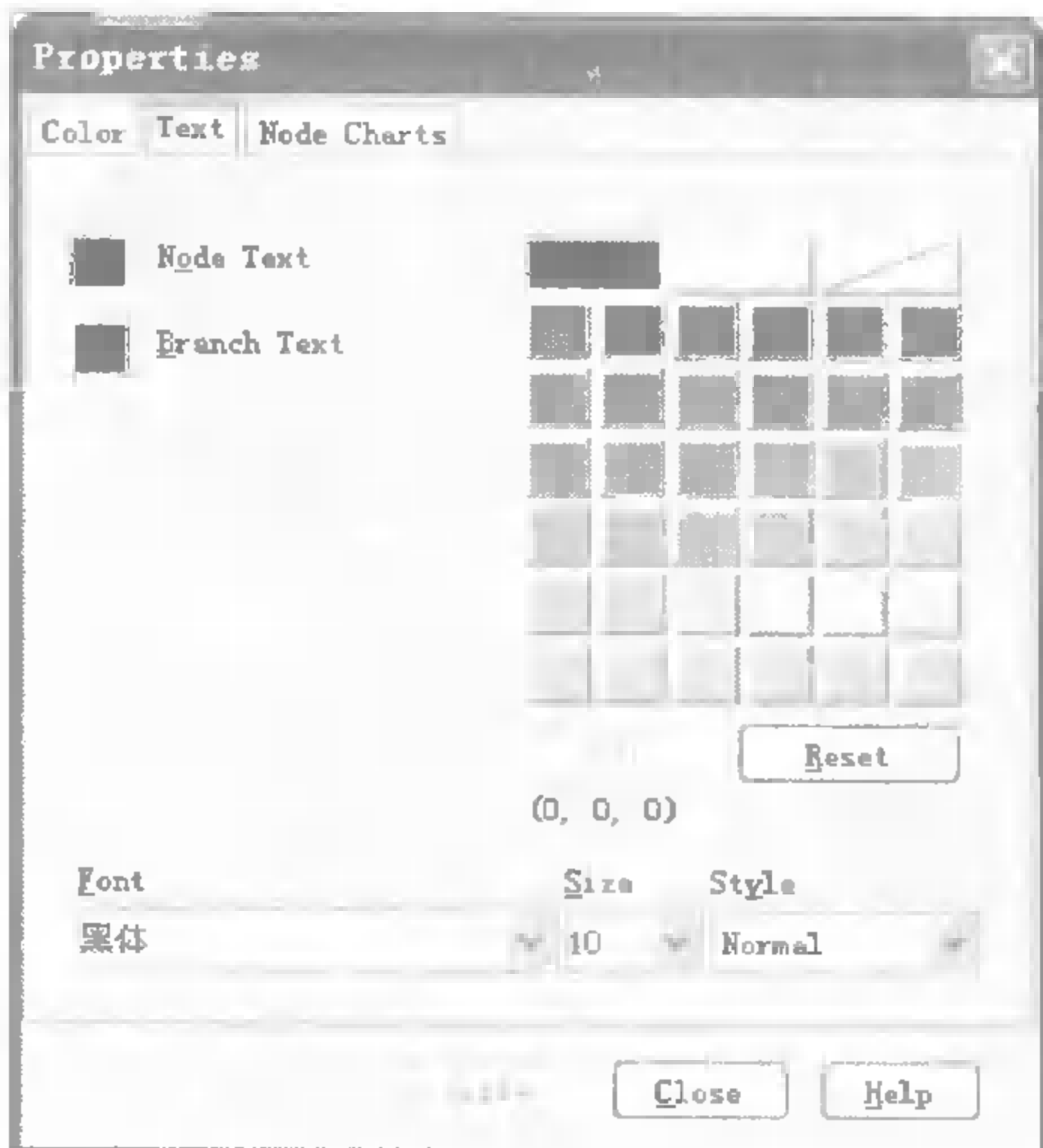


图 12-93 字体编辑窗口

回到图 12-92 所示的 Tree Editor 窗口中, 在任意节点上右击, 都弹出图中所示的快捷菜单, 下面对其加以简单介绍。

- Filter Cases 选项, 选中后弹出利用该节点在当前数据集生成过滤变量的设置对话框, 只需输入新变量的名字即可。
- Export 选项, 选中后弹出输出设置对话框, 把当前节点 SPSS Syntax 或 SQL 格式的决策规则输出到指定文件。
- Summary 选项, 选中后弹出新的节点设置窗口。在那里可以选择用表格、图形或命令语句 (SPSS 或 SQL) 的方式显示当前节点; 还可以输出节点的决策规则到指定文件, 以及直接利用该节点生成过滤变量。
- Hide Children 选项, 选中后隐藏当前节点的所有子节点。单击当前节点右下角的减号可以达到相同效果, 同时减号变为加号, 再次单击加号重新显示当前节点的子节点。

(3) 收益和风险输出。如图 12-94 所示, “收益的收益”表格给出用决策树进行分类的收益信息。其中“收益”列显示了我们感兴趣的因变量目标取值 (本例为适合运动) 的分布情况; “响应”列显示了当前节点中的目标响应; “指数”列显示的是收益百分比和节点百分比的比值。例如训练样本 (Training 行) 的第 3 节点包括了总观测的 8.7%, 所有目标取值 (适合运动) 的 14.3%, 而目标取值占当前节点 (响应) 的比例为 100%。

收益的收益							
样本	节点	节点		收益		响应	指数
		N	百分比	N	百分比		
Training	3	2	8.7%	2	14.3%	100.0%	164.3%
	4	14	60.9%	12	85.7%	85.7%	140.8%
	4	7	30.4%	0	0%	0%	0%
Test	3	2	25.0%	2	40.0%	100.0%	160.0%
	4	3	75.0%	3	60.0%	50.0%	80.0%
	4	0	0%	0	0%		

增长方法 CHAID  
因变量列表 运动

风险		
样本	估计	标准误差
训练	0.87	0.059
检验	375	171

增长方法 CHAID  
因变量列表 运动

图 12-94 收益和风险输出

“风险”表格给出用决策树进行分类的风险信息。根据事先设置的错判矩阵得到使用该

决策树对训练样本和验证样本分别进行分类的风险度为 8.4%、37.5%。

(4) 总体预测分类汇总。如图 12-95 所示,是用最终决策树模型进行分类的汇总表。可见,对训练样本和检验样本的总判断准确率分别为 91.3%和 62.5%;而且如果真实的观测情况是适合运动的话,预测的准确率都能够达到 100%;对于训练样本,预测为适合的后验概率为  $14/(14+2) = 87.5\%$ ;对于验证样本,预测为适合的后验概率为  $5/(5+3) = 62.5\%$ 。

分类				
样本	观测	预测		
		不适合	适合	百分比更正
训练	不适合	7	2	77.8%
	适合	0	14	100.0%
	整体百分比	30.4%	89.6%	91.3%
检验	不适合	0	3	0%
	适合	0	5	100.0%
	整体百分比	0%	100.0%	62.5%
增长方法 CHAID				
因变量列表 运动				

图 12-95 分类汇总表

总体看来,使用该决策树来预测适宜运动的天气是有一定参考价值的。

# 第 13 章 生存分析

生存分析方法广泛应用于医学、社会科学、工业研究等领域，例如：患者经治疗后的生存时间分析、设备使用寿命分析等。这类问题的特点是在研究期间结束时，所要研究的事件还没有发生或者过早终止，使得需要收集的数据发生缺失，这样的数据称为生存数据。生存分析就是处理、分析生存数据的一种方法，又称之为时间-效应分析（Time-Effect Analysis）。

生存分析的主要研究内容包括如下 3 个方面：

- ① 对生存状况进行统计描述，例如生存概率、生存率、中位生存期等。
- ② 寻找影响生存时间的“危险因素”和“保护因素”。
- ③ 估计生存率和生存时间的长短，进行预后分析。

## 13.1 生存分析简介

生存分析（Survival Analysis），即生存数据的统计分析，是近年产生并且发展甚为迅速的一门应用统计分支。“生存”的含意很广，可以指人或动物的存活（相对于死亡），也可以指患者的病情正处于缓解状态（相对于再次复发或恶化）；对于某个系统或产品，“生存”就指它正常工作并完成其规定功能的状态（相对于失效或故障）。尽管生存分析中所讨论的模型，以至所采用的术语大多来自医学和生物学，但它的应用并不局限于这两个领域，实践表明，它在工业技术甚至社会经济学科中都有着广泛的应用。

### 13.1.1 生存分析的基本概念

下面介绍一些在生存分析中经常用到的概念。

#### 1. 生存时间（survival time）

广义的生存时间指从某个起始事件开始，到某个终点事件的发生所经历的时间，也称为失效时间（failure time）。

生存时间的特点如下：

- ① 分布类型不易确定，一般不服从正态分布，有时近似服从指数分布、Weibull 分布、Gompertz 分布等，多数情况下都不服从于特定的分布类型。
- ② 影响生存时间的因素较为复杂，而且不易控制。

根据研究对象的结局，生存时间数据可分为如下两种类型。

- ① 完全数据（complete data）：观察对象在观察期内出现响应（终点事件），这时记录到



的时间信息是完整的。

- 截尾数据 (截尾值、删失数据, censored data): 尚未观察到研究对象出现响应 (终点事件) 时, 即由于某种原因停止了随访, 这时记录到的时间信息是不完整的, 常在数据的右上角以符号 “+” 标识。

## 2. 死亡概率

死亡概率 (mortality probability): 指期初的观察对象在某单位时段内死亡的可能性大小, 记为  $q$ , 计算公式为:  $q = \frac{\text{某单位时段内死亡数}}{\text{该时段期初观察人数}}$ , 若在此时段内有截尾值, 则分母用校正人口数 (期初人数-截尾数) / 2 代替。

死亡率 (mortality rate): 指单位时间内研究对象的死亡频率或强度, 例如平均每百人 (或万人、千人) 中的死亡人数, 记为:  $m = \frac{\text{某单位时段内死亡数}}{\text{该时段平均人口数}} \times 100\%$ , 其中: 平均人口数 = (该时段期初人口数+期末人口数) / 2。

## 3. 生存概率

生存概率 (survival probability): 表示某单位时段开始时, 存活的个体到该时段结束时仍存活的可能性大小, 用  $p$  表示, 计算公式为:  $p = \frac{\text{活满某时段的人数}}{\text{该时段期初观察人数}} = 1 - q$ , 若该时段内有删失数据, 则分母用校正人口数 (期初人数-截尾数) / 2 代替。

生存率 (survival rate): 指研究对象经历  $t$  个时段后仍存活的概率, 即生存时间大于等于  $t$  的概率, 用  $p(T \geq t)$  表示。

## 4. 生存函数

生存率随时间  $t$  的变化而变化, 它是  $t$  的函数, 记为  $S(t)$ , 称之为生存函数 (survival function), 生存函数在某时刻的函数值就是生存率。它的计算公式分为两种情况: 若前  $t$  个时段没有删失数据, 则  $S(t) = P(T \geq t) = \frac{t \text{ 时段结束时仍存活的人数}}{\text{研究期初观察总人数}}$ ; 若观察期内存在删失数据, 假定多个对象在各单位时段内是否生存的事件是相互独立的, 其生存概率分别记为  $p_1, p_2, p_3, \dots, p_t$ , 则有:  $S(t) = p_1 \cdot p_2 \cdot p_3 \cdots p_t = \prod_{ij \leq t} p_j$ 。

生存函数又称为累积生存概率 (cumulative probability of survival), 即将时刻  $t$  尚存活看成是前  $t$  个时段一直存活的累计结果。

## 5. 生存率曲线 (survival curve)

它是指以时间为横轴、生存率为纵轴, 将各时刻的生存率连接在一起的曲线图。曲线形状分为如下两种: 阶梯形, 多为小样本资料用直接法估计得到的生存曲线; 折线形, 多为大样本资料用频数表法估计得到的生存曲线。

## 6. 中位生存期 (median survival time)

指生存时间的中位数, 也称之为半数生存期, 表示生存率等于 50% 时的时刻, 反映了生

存时间的平均水平。

## 7. 危险率函数

危险率函数(hazard function): 指  $t$  时刻存活, 在  $t \sim t + \Delta t$  时段内死亡的概率(条件概率), 用  $h(t)$  表示, 计算公式为:  $h(t) = \lim_{\Delta t \rightarrow 0} \frac{p(t < T < t + \Delta t | T \geq t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{n(t) - n(t + \Delta t)}{n(t) \cdot \Delta t}$ 。危险率函数又称为死亡力(force of mortality)、瞬时死亡率(instantaneous failure rate)等, 它是生存分析的基本函数, 反映了研究对象在某时刻死亡的风险大小。

生存函数与危险率函数的关系可以表示为:  $S(t) = \exp \left[ - \int_0^t h(t) dt \right]$ 。

### 13.1.2 生存分析的数据特点

由于生存数据的特殊性, 关于其数据来源和数据特点, 在此给出几点简单的参考意见。

- (1) 样本需由随机抽样方法获得, 应保证一定的数量, 死亡例数及其比例不能太少。
- (2) 完全数据所占的比例不能太少, 即截尾值不宜太多。
- (3) 截尾值出现的原因应具有无偏性, 这就需要对被截尾研究对象的年龄、职业、地区、病情轻重等情况进行分析。

(4) 生存时间应尽可能精确, 由于许多常用的生存分析方法, 都是在对生存时间排序的基础上做统计处理, 即使小小的舍入误差也可能因改变时间顺序而影响结果。

### 13.1.3 生存分析的常用方法

(1) 非参数法: 其特点是不论资料是什么样的分布形式, 只根据样本提供的顺序统计量对生存率进行估计, 包括乘积极限法和寿命表法等。

(2) 参数法: 其特点是假定生存时间服从于特定的分布, 并根据已知分布的特点对观测样本的生存时间进行分析, 包括指数分布法、Weibull 分布法、对数正态回归分析法和 logit 回归分析法等。

(3) 半参数法: 兼有非参数法和参数法的特点, 主要用来研究影响生存时间和生存率的因素, 属多因素分析方法, 典型方法为 Cox 模型分析法。

### 13.1.4 SPSS 中的生存分析过程

SPSS 提供了比较全面的生存分析处理过程, 依次单击菜单 “Analyze→Survival”, 弹出的子菜单有如下 4 个分析过程。

- Life tables: 适用于分组生存资料, 可求出不同组段的生存率。当样本量较大时, 可把数据按时间段分成几组, 用此分析过程观察不同时间段的生存率。
- Kaplan-Meier: 适用于样本含量较小的情况, 它不能给出特定时刻的生存率, 所以不用担心某些时间段内只有很少的几个观测, 甚至没有观测的情况。
- Cox Regression: 用于拟合 Cox 比例风险模型, 它是多因素生存分析比较常用的一种方法。
- Cox w/Time-Dep Cov: 带时间相依性变量的生存分析, 是 Cox 模型的发展。当所研究的危险因素取值(或者其作用强度)随时间不断变化时, 就要用到这个过程。

13.2 生命表分析

当样本量较大时，通常先将样本数据整理成频数表的形式，再用生命表法计算数据的生存率及其标准误。生命表法采用与编制生命表相似的原理计算生存率，首先求出对患者实施治疗后（或对健康者实施预防措施后）各个时期的生存概率，然后根据概率的乘法法则，将不同时期的生存概率相乘，就得到自观察开始到指定时刻的生存率。

SPSS 的 Life Tables 过程可以用于如下问题的研究。

- 编制寿命表。
- 绘制各种生命曲线，例如生存函数曲线、风险函数曲线等。
- 控制其他因素后，对指定的研究因素在不同水平下的生存时间分布进行比较，包括总体上的比较和不同取值水平之间的两两比较。

13.2.1 生命表分析简介

生命表（Life-Table Method，简称 LT 法）是一种古老的工具，它的出现与人口统计学和人寿保险科学有着密切的联系。现代统计学科的发展使寿命表的理论更趋完善，令其在流行病学、临床医学、遗传学等许多领域都得到广泛的应用。

生命表法通过计算落入单位时间段内的失效观察和删失观察的个数，估计该区间上的死亡概率；并且用该区间及其之前各区间上的生存概率之积估计生存率。就统计方法而言，寿命表分析属于非参数方法的范畴。

当资料是按照固定的时间间隔收集（比如每个月随访一次）时，随访结果只有该年或该月期间的若干观察人数、出现预期观察结果的人数和截尾人数（删失人数），每位患者的确切生存时间是无法知道的，此时就应当使用寿命表法进行研究，这也被称之为分组资料的生存分析。

13.2.2 生命表分析的基本步骤

本节以一个简单的例子来介绍生命表分析的基本步骤。

如表 13-1 所示，第（1）～（4）列是某项手术之后随访收集的资料，下面用生命表法估计其生存率，计算结果列在表的后几列中，各列符号的意义及其计算方法如下。

表 13-1 术后随访资料表

时间 段/年	期内死 亡人数	期内删 失人数	年初观 察人数	校正期初 观察人数	死亡 概率	生存 概率	t+1 年生 存率	生存率 标准误
$t$	$d$	$c$	$n_0$	$n = n_0 - c/2$	$q = d/n$	$p = 1 - q$	$s(t+1)$	$SE[S(t+1)]$
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
0~	68	8	233	229.0	0.296 9	0.703 1	0.703 1	0.030 2
1~	61	7	157	153.5	0.397 4	0.602 6	0.423 7	0.033 2
2~	38	3	89	87.5	0.434 3	0.565 7	0.239 7	0.029 3
3~	16	1	48	47.5	0.336 8	0.663 2	0.158 9	0.025 4
4~	8	0	31	31.0	0.258 1	0.741 9	0.117 9	0.022 6
5~6	23	0	23	23.0	1.000 0	0.000 0	0.000 0	0.000 0

- 第(1)列, 术后年数  $t$ : 以手术后为观察起点, 按术后年数划分组段, 例如“0~”组段表示术后不满1年。
- 第(2)列, 期内死亡人数  $d$ : 表示相应时段内出现结局事件(如死亡)的人数。
- 第(3)列, 期内删失人数  $c$ : 表示相应时段内出现截尾(失访、死于其他疾病等情况)的人数。
- 第(4)列, 年初观察人数  $n_0$ : 表示各组段的下限所对应时刻的观察人数。
- 第(5)列, 校正期初观察人数  $n$ , 计算公式为:  $n = n_0 - c/2$ 。
- 第(6)列, 由定义可知, 相应时段内的死亡概率  $q$  为:  $q = d/n$ 。
- 第(7)列, 生存概率  $p$ :  $p = 1 - q$ 。
- 第(8)列, 生存率  $S(t+1)$ , 表示各组段的上限所对应时刻的生存率, 即研究对象活满  $t+1$  年的概率, 计算公式为:  $S(t+1) = p_0 \cdot p_1 \cdot p_2 \cdots p_{n-1}$ 。
- 第(9)列, 第  $t+1$  年生存率的标准误, 计算公式为:  $SE[S(t+1)] = S(t+1) \sqrt{\sum_{i=1}^{t+1} \frac{q_i}{p_i n_i}}$ 。

如表 13-1 所示, 从死亡概率看, 前 3 年死亡的危险性逐年增加, 而后呈下降趋势, 生存概率从反面说明了这一特性; 再看第(8)列的生存率, 半数以上的病人术后活不到 2 年, 说明此病对生命威胁较大; 由于生存率的标准误都比较小, 表示此处的生存率具有一定代表性。

13.2.3 生命表实例分析

本节以电信数据为例, 用生命表法来研究很多行业都很关心的客户流失问题。SPSS 的 Life Tables 过程要求时间变量为数值型的, 事件变量和因素变量都为分类变量, 且以整数编码。

1. 数据和问题描述

本例的数据摘自 SPSS 自带的 Demo 示例文件“telco.sav”, 所用数据文件为“电信客户流失数据.sav”, 数据格式如图 13-1 所示。其中时间变量为客户的在网月数, 结局事件为客户在最后一个 月是否流失, 客户种类是按照客户消费的服务种类进行划分的。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	tenure	Numeric	4	0	在网月数	None	None	6	Right	Scale
2	custcat	Numeric	8	0	客户种类	{1, 基本服务}	None	6	Right	Nominal
3	churn	Numeric	4	0	是否流失	{0, 没有}...	None	6	Right	Nominal
4	ed	Numeric	4	0	教育水平	{1, 低于高中}	None	6	Right	Ordinal
5	employ	Numeric	4	0	当前工作年限	None	None	6	Right	Scale
6	retire	Numeric	8	2	是否退休	{.00, 没退休}	None	10	Right	Nominal
7	gender	Numeric	4	0	性别	{0, 男}	None	6	Right	Nominal
8	reside	Numeric	4	0	家庭人数	None	None	6	Right	Scale


图 13-1 电信客户流失数据格式

下面通过生命表分析, 研究不同种类的客户, 其流失情况有何差异。

2. Life Tables 分析过程的参数设置

依次单击菜单“Analyze→Survival→Life Tables...”, 执行生命表分析过程, 其主设



置界面如图 13-2 所示,在此设置分析变量及其取值范围。在变量列表单击选中在网月数变量,单击从上至下第一个  按钮,将其作为时间变量选入 Time 选框,在 by 前输入 60,在 by 后输入 3。

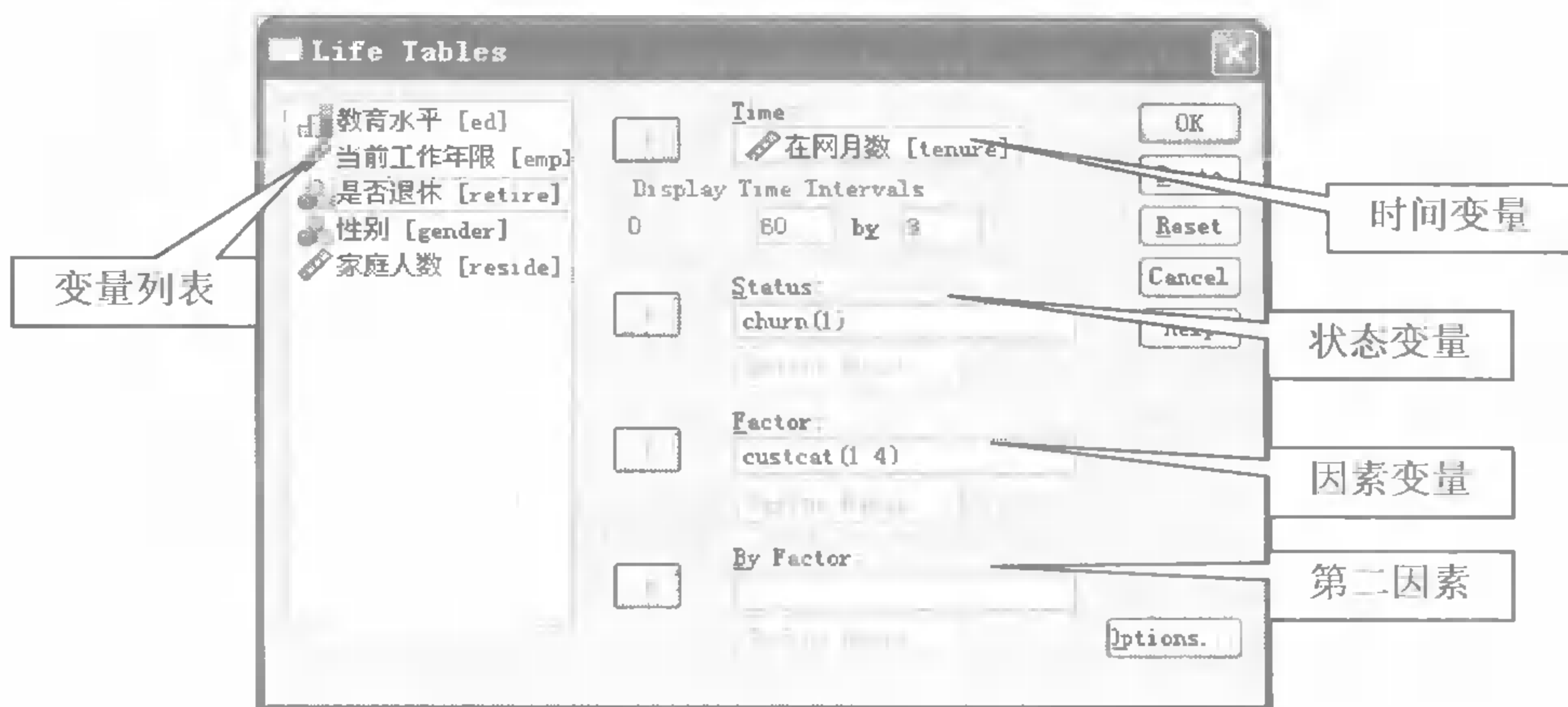



图 13-2 Life Tables 过程的主设置界面

(1) Time 选框,用于选入代表生存时间的变量。

Display Time Intervals 子设置栏,指定在生命表中生存时间的范围及其组距。by 前的输入框指定生存时间的上限,by 后的输入框指定生存时间的组距(单位时间段)。

(2) Status 选框,用于选入定义事件是否发生的生存状态变量。

在变量列表单击选中是否流失变量,单击从上至下第二个  按钮,将其作为状态变量选入 Status 选框;单击 Define Event 按钮,弹出如图 13-3 所示的定义事件对话框,在 Single value 后输入“1”。单击 Continue 按钮返回主界面。

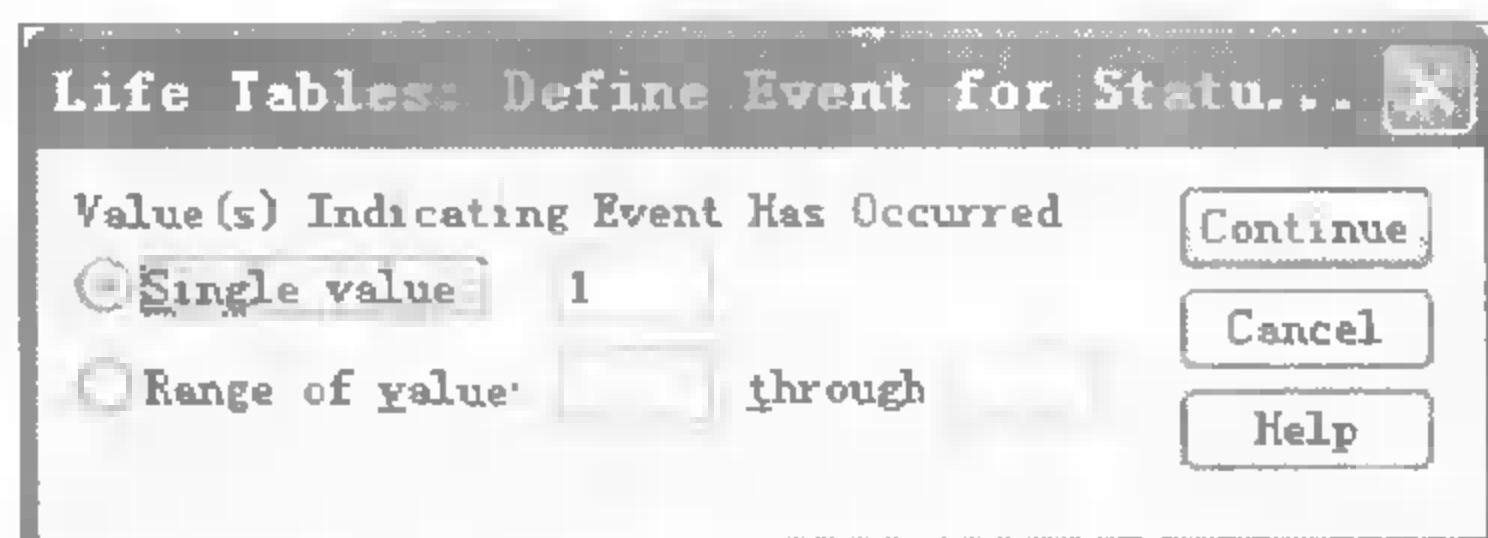



图 13-3 定义最终事件的对话框

在此给出了两种指定事件发生与否的方法。

- Single value 单选框,当生存状态为二元变量时,选中此项,并在后面的输入框指定状态变量的代表事件发生的取值即可。
- Range of values 单选框,当生存状态为多分类变量时,选中此项,并在 through 前的输入框指定取值范围的起始值,在 through 后的输入框指定取值范围的终止值。

(3) Factor 选框,用于选入第一个因素变量(分组因素)。

在变量列表单击选中客户种类变量,单击从上至下第三个  按钮,将其作为因素变量选入 Factor 选框;单击 Define Range 按钮,弹出如图 13-4 所示的定义取值范围对话框,分别在 Minimum、Maximum 后输入“1”和“4”;单击 Continue 按钮返回主界面。

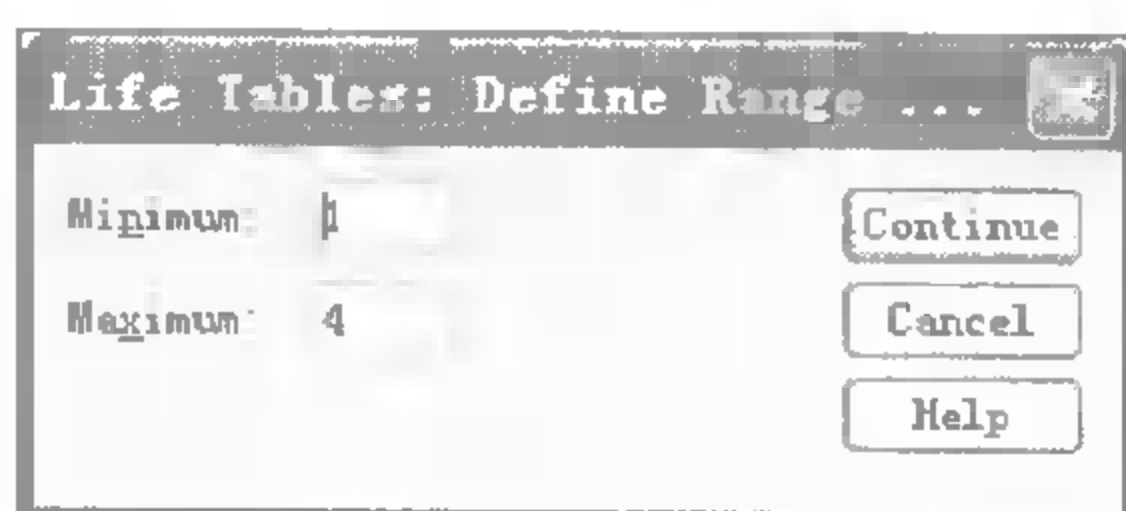


图 13-4 定义因素变量的对话框

如图 13-4 所示, Minimum、Maximum 输入框分别用于指定取值范围的最小值和最大值, SPSS 对于此设定范围外的观测, 将自动忽略。

(4) By Factor 选框, 用于选入第二个因素变量(分层因素), 设置方式同(3)。

(5) 表格和图形输出设置。在图 13-2 中, 单击 Options 按钮, 弹出如图 13-5 所示的选项设置对话框。勾选 Survival 复选框; 单击选中 Pairwise 单选框; 单击 Continue 按钮返回主界面。

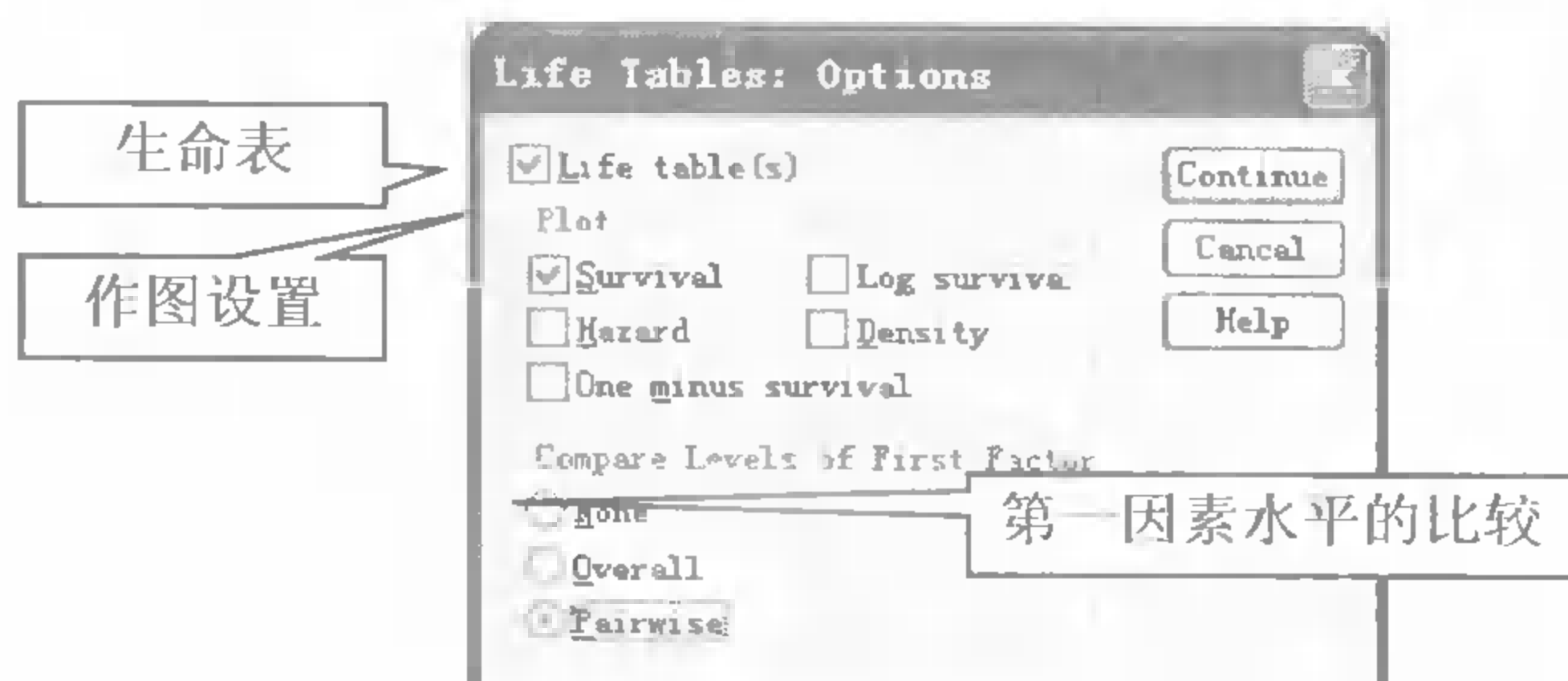


图 13-5 表格和图形输出设置对话框

① Life tables 复选框, 选中它表示在结果里输出生命表, 反之不会输出。

② Plot 栏, 在此选择输出的图形类型, 可选项有如下几个: Survival (累积生存函数曲线)、Hazard (累积风险函数散点图)、One minus survival (生存函数被 1 减后的曲线)、Log survival (对数累积生存函数曲线) 和 Density (密度函数散点图)。

③ Compare Levels of First Factor 栏, 设置对第一个因素不同取值水平的比较方法。此处所作的就是 Wilcoxon (Gehan) 检验, 它用于比较不同分组的生存率, 如果还指定了第二个因素, 就对它的每个取值分别做检验。可选项有如下 3 个。

- ❶ None 不做比较, 系统默认。
- ❷ Overall 整体比较, 其检验的零假设为各分组的生存曲线全部相同, 相当于方差分析中的总体比较。
- ❸ Pairwise 两两比较, 相当于方差分析中的两两比较, 同时也会给出整体比较的结果。

### 3. 结果分析

在图 13-2 中单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 13-6~图 13-8 所示。

年限表														
一阶控制	期初时间	期初记入数	期内退出数	历险数	期间终结数	终结比例	生存比例	期末的累积生存比例	期末的累积生存比例的标准误	概率密度	概率密度的标准误	风险率	风险率的标准误	
客户基本服务种类	000	266	5	264.500	10	04	96	96	01	013	004	01	00	
	7 000	253	10	246.000	17	07	93	90	02	022	005	02	01	
	6 000	226	12	220.000	10	05	95	86	02	014	004	02	00	
	9 000	204	11	198.500	10	05	95	81	02	014	004	02	01	
	12 000	183	13	176.500	6	03	97	76	03	009	004	01	00	
	15 000	164	10	159.000	5	03	97	76	03	008	004	01	00	
	18 000	140	15	141.500	1	01	99	75	03	000	000	00	00	

图 13-6 客户流失生命表

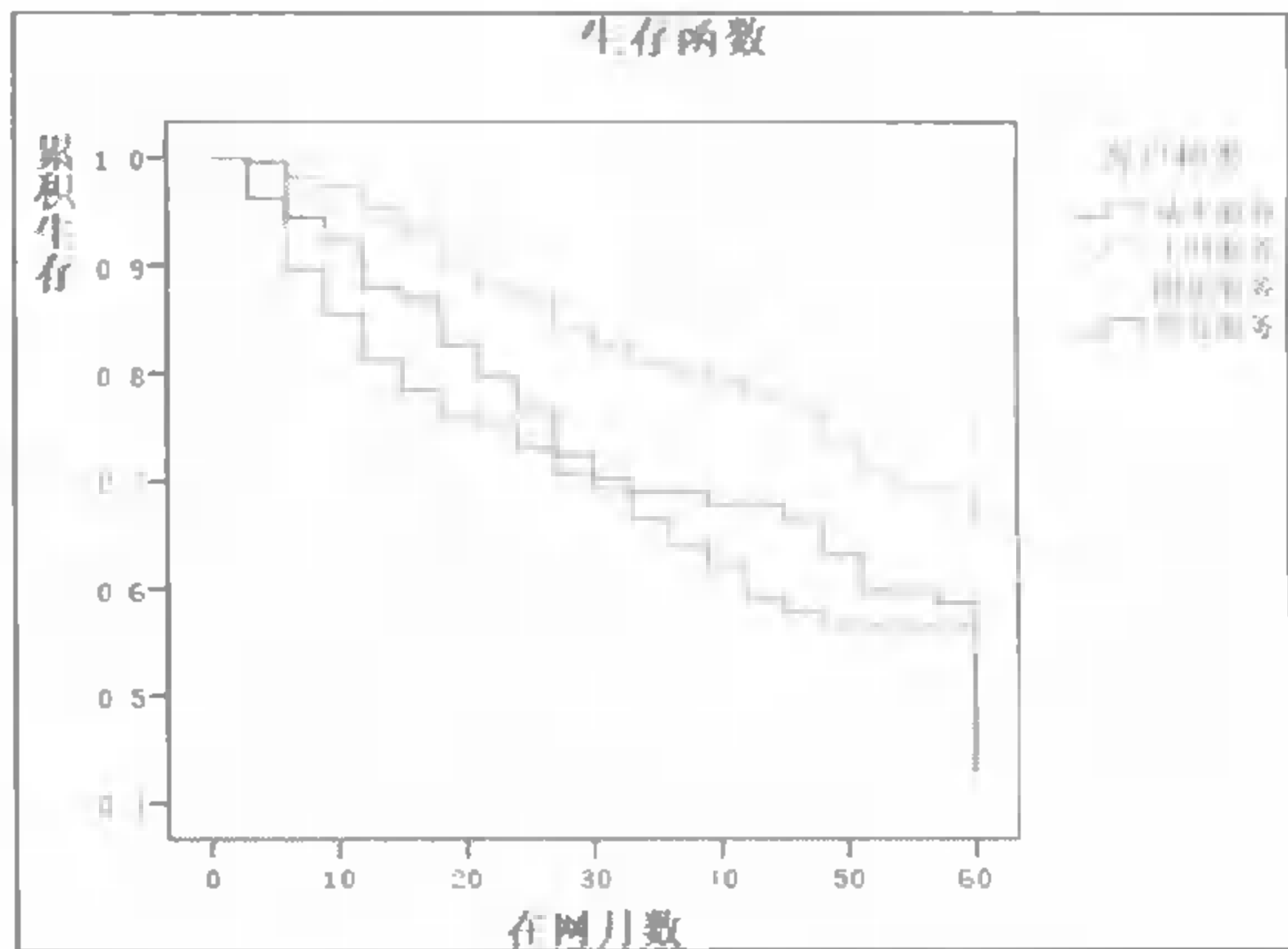


图 13-7 客户流失累积生存函数图

整体比较<sup>a</sup>

Wilcoxon (Gehan) 统计量	df	Sig.
39.179	3	.000

<sup>a</sup> 比较是精确的。

成对比较<sup>a</sup>

(1) 客户类型	(2) 客户类型	Wilcoxon (Gehan) 统计量	df	Sig.
1	2	18.640	1	.000
	3	37.154	1	.000
	4	2.949	1	.086
2	3	18.640	1	.000
	4	5.215	1	.019
	1	9.222	1	.002
3	1	37.154	1	.000
	2	5.515	1	.019
	4	27.229	1	.000
4	1	2.949	1	.086
	2	4.223	1	.002
	3	27.229	1	.000

<sup>a</sup> 比较是精确的。

平均分

比较组	总数	未审查	已审查	已审查的百分比	平均分
1对2	266	83	183	68.8%	-33.902
1对3	266	59	207	77.8%	-47.358
1对4	266	83	183	68.8%	-32.726
2对3	266	44	222	84.3%	-49.911
2对4	266	83	183	68.8%	-14.842
3对4	266	88	178	62.7%	16.729
1对2	266	83	183	68.8%	-20.120
1对3	266	44	222	84.3%	-15.537
1对4	266	83	183	68.8%	-17.871
2对3	266	88	178	62.7%	-23.657
2对4	266	44	222	84.3%	-39.334
3对4	266	88	178	62.7%	-47.072
1	266	83	183	68.8%	-14.842
2	266	59	207	77.8%	27.871
3	266	44	222	84.3%	39.334
4	266	88	178	62.7%	-47.072

整体比较

图 13-8 整体比较的 Wilcoxon 检验和两两比较结果

(1) 生命表输出。如图 13-6 所示，主要消费在于基本服务的客户，有很多在入网后一年的时间内流失，所以对这类客户，建议运营公司在其入网的一年内加强监测和回访，以提高客户满意度。

(2) 累积生存函数图。如图 13-7 所示，是关于客户流失的累积生存函数图，它是对生命表的图形展示，能让用户以更形象的方式分析结果。从图中看出，所有服务和基本服务这两种客户的累计生存函数下降最快，网络服务客户的累计生存函数较附加服务客户的下降要快。要判断这些差异是由随机因素引起的，还是有统计学意义的，就需要分析各水平的相互比较结果了。

(3) 各水平比较结果。如图 13-8 所示，“整体比较”表格的 Wilcoxon 检验结果说明四种客户的生存曲线是显著不同的；“成对比较”表格给出了更详细的结论，如图中蓝色线框标识，基本服务客户与上网、附加这两类客户生存曲线之间的差异是显著的，所有服务客户与上网、附加这两类客户生存曲线之间的差异也是显著的；但基本服务与所有服务、上网服务与附加服务，这两对类别的客户之间的生存曲线差异是不显著的。

### 13.3 Kaplan-Meier 分析

生命表分析适用于大样本的情况，而在处理小样本时，为充分利用每个数据所包含的信息，必须采用更为精确的估计方法，其中应用最多、效率较高的就是 Kaplan-Meier 的乘积极限估计 (Product-Limit Estimates)。

SPSS 的 Kaplan-Meier 过程，适用于如下问题的研究。

- 估计某研究因素不同水平的中位生存时间。
- 比较某研究因素不同水平的生存时间有无差异。
- 控制某分层因素后，对感兴趣的分组因素不同水平的生存时间做比较。

#### 13.3.1 Kaplan-Meier 分析的步骤

乘积极限法由 Kaplan 和 Meier 在 1958 年首先提出，故又称为 Kaplan-Meier 法(K-M 法)。它主要适用于样本含量较小的资料，计算步骤如下。

(1) 将样本量为  $n$  的样本观察值，按照生存时间  $t$  由小到大依次排列，秩记为： $i = 1, 2, 3 \dots n$ 。如果遇到非截尾值与截尾值相同的情况，将非截尾值排在前面。

(2) 列出各时刻（代表某个很短的时间单位）开始时的存活数，记为：期初观察单位数  $n_i$ 。

(3) 计算各时刻的死亡概率  $q$ ，以及生存概率  $p = 1 - q$ 。

(4) 计算存活到各时刻的生存率  $S(t_i)$ ，它等于从时间起点到  $t_i$  之间各生存概率的连乘积。

(5) 计算生存率的标准误： $SE[S(t_i)] = S(t_i) \sqrt{\sum \left[ \frac{1}{(n-i)(n-i+1)} \right]}$ ，其中  $i$  为秩次，

$\sum \left[ \frac{1}{(n-i)(n-i+1)} \right]$  表示把小于等于  $t_i$  的各非截尾值所对应的  $\frac{1}{(n-i)(n-i+1)}$  求和。

(6) 绘制生存率曲线，通常都是成阶梯形的曲线。方法是以非截尾值为横轴、生存率为纵轴作散点图，然后将各点先垂直向下再水平向右连接成阶梯形的。

(7) 按照正态近似法估计总体生存率的置信区间，某时刻  $t_i$  总体生存率的  $(1-\alpha)\%$  置信区间为： $S(t_i) \pm u_{\alpha/2} SE[S(t_i)]$ 。

#### 13.3.2 生存曲线的比较和检验

有时需要对多个样本的生存曲线进行比较，以判断它们之间是否具有显著的差异。

对数秩检验 (log-rank test) 就是用于比较多条生存曲线的一种方法，其零假设为指定的多条生存曲线没有差异。基本思想是：先根据不同分组的期初人数和死亡人数，计算各组的理论死亡数；若零假设成立，则实际死亡数与理论死亡数不会相差太大，否则应认为零假设不成立，即这些生存率曲线之间的差异具有统计学意义。

对数秩检验统计量（近似法）为： $\chi^2 = \sum_{k=1}^m \frac{(A_k - T_k)^2}{T_k}$ ，其中  $A_k$  和  $T_k$  分别是第  $k$  组死亡

的实际数和理论期望数。在  $H_0$  成立的条件下，统计量  $\chi^2$  服从自由度为  $m-1$  的  $\chi^2$  分布， $m$  为分组个数，由此据  $\chi^2$  检验就可作出是否拒绝  $H_0$  的决定。



### 13.3.3 Kaplan–Meier 分析的实例

SPSS 的 Kaplan-Meier 过程要求时间变量为连续型的，事件变量可以为连续变量或分类变量，因素变量和分层变量必须为分类变量。

本节利用 Kaplan-Meier 方法研究不同药品对同一疾病的治疗效果。

#### 1. 数据和问题描述

本例的数据摘自 SPSS 自带的 Demo 示例文件“pain\_medication.sav”，所用数据文件为“止痛药数据调查.sav”，数据格式如图 13-9 所示。其中时间变量为药物起作用的生效时间，事件变量（状态）表示在指定时刻某种药物是否生效，因素变量为治疗方法。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	treatment	Numeric	4	0	治疗方法	{0, 新药}...	None	8	Right	Nominal
2	time	Numeric	8	2	生效时间	None	None	7	Right	Scale
3	status	Numeric	4	0	状态	{0, 非有效}...	None	6	Right	Nominal

图 13-9 新旧止痛药调查的数据格式

下面通过 Kaplan-Meier 分析，研究新药相对于旧药，是否具有更好的治疗效果。

#### 2. Kaplan-Meier 分析过程的参数设置

依次单击菜单“Analyze→Survival→Kaplan-Meier...”。执行 Kaplan-Meier 分析过程，其主设置界面如图 13-10 所示，在此设置分析变量及其取值。

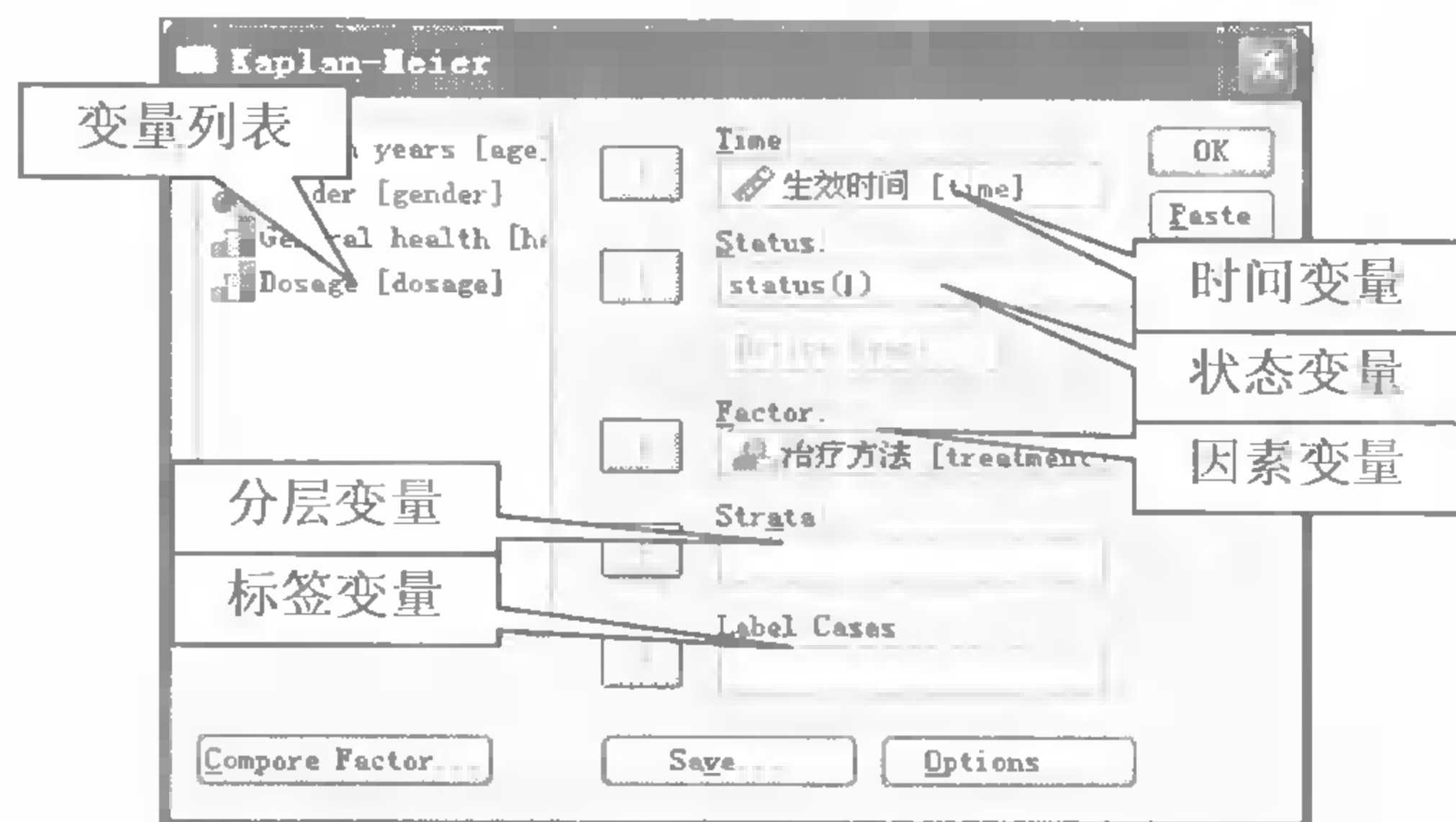





图 13-10 Kaplan–Meier 过程的主设置面板

（1）指定分析变量。在变量列表单击选中生效时间变量，单击从上至下第一个  按钮，将其作为时间变量选入 Time 选框；在变量列表单击选中状态变量，单击从上至下第二个  按钮，将其作为状态变量选入 Status 选框；在变量列表单击选中治疗方法变量，单击从上至下第三个  按钮，将其作为因素变量选入 Factor 选框。

在图 13-10 中，Time 选框用于选入生存时间变量；Status 选框用于选入生存状态变量；Factor 选框用于选入因素变量；Strata 选框用于选入分层因素，可以看作是研究者欲加以控制的混杂因素，SPSS 会对其每个取值水平分别进行分析；Label Cases 选框用于选入观测的标签变量，当有必要关心每名患者在研究队列中的情况时，可以在此选入代表姓名的变量，以便在生命表中输出各个患者的姓名。

(2) 状态变量的取值设置。在图 13-10 中, 单击选中 Status 选框里的变量, 然后单击 Define Event 按钮, 弹出如图 13-11 所示的定义事件对话框, 在 Single value 后输入“1”; 单击 Continue 按钮返回主界面。

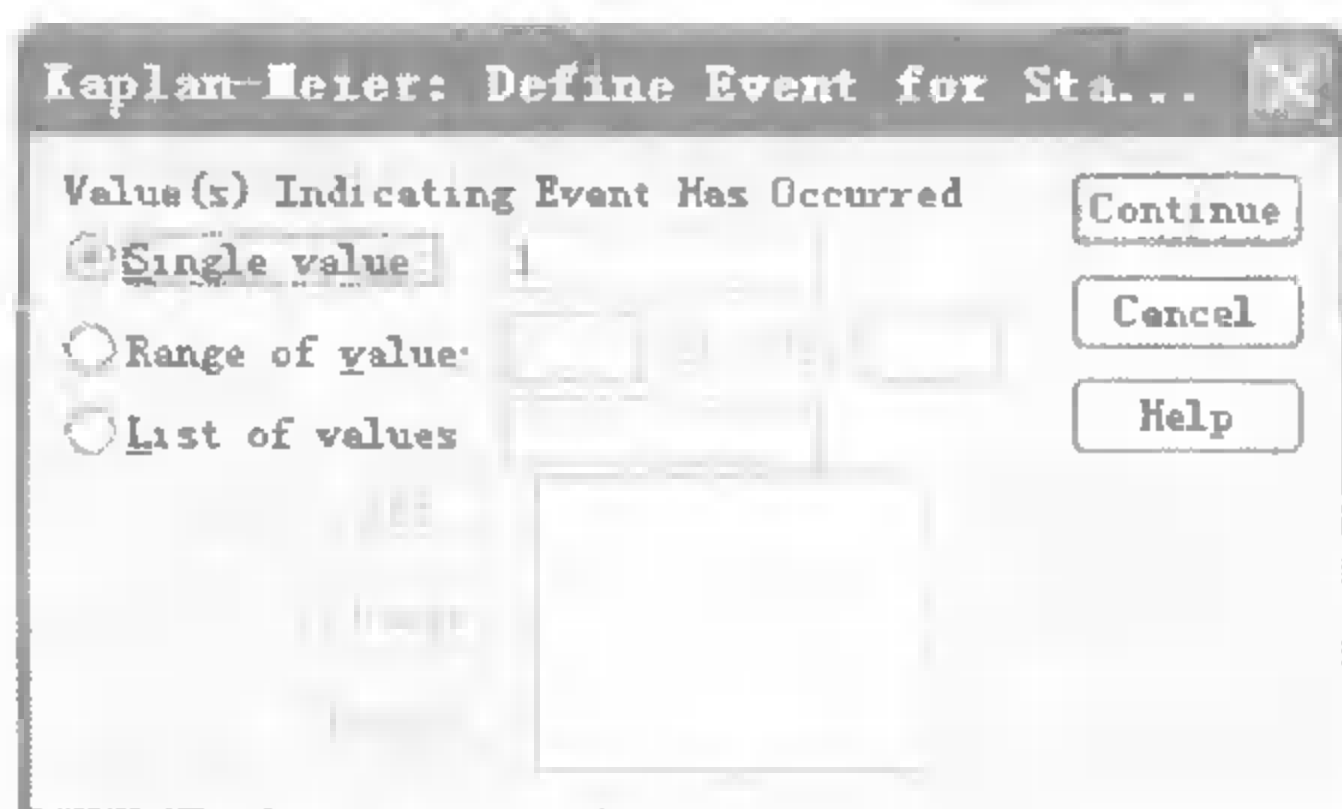


图 13-11 取值定义子对话框

在此给出了 3 种指定事件发生与否的方法。

- ① Single value 单选框, 当生存状态为二元变量时, 选中此项, 并在后面的输入框指定状态变量的代表事件发生的取值即可。
- ② Range of values 单选框, 当生存状态为多分类变量时, 选中此项, 并在 through 前的输入框指定取值范围的起始值, 在 through 后的输入框指定取值范围的终止值。
- ③ List of values 选项, 在其后的输入框填入某个数字后, 单击 Add 按钮将其加入下面的列表里, 如此重复可以指定代表事件发生的多个不同的值; 如果需要更改已填入的值, 先在列表里单击选中它, 然后在 List of values 输入框进行编辑, 最后单击 Change 按钮即可确认修改, 而单击 Remove 按钮将直接删除选中的值。

(3) 因素取值水平的比较设置。在图 13-10 中, 单击 Compare Factor 按钮, 弹出如图 13-12 所示的对话框, 在此设置对因素变量取值水平的比较方法。依次勾选如下 3 个复选框: Log rank、Breslow、Tarone-Ware; 单击选中 Pooled over strata 单选框; 单击 Continue 按钮返回主界面。

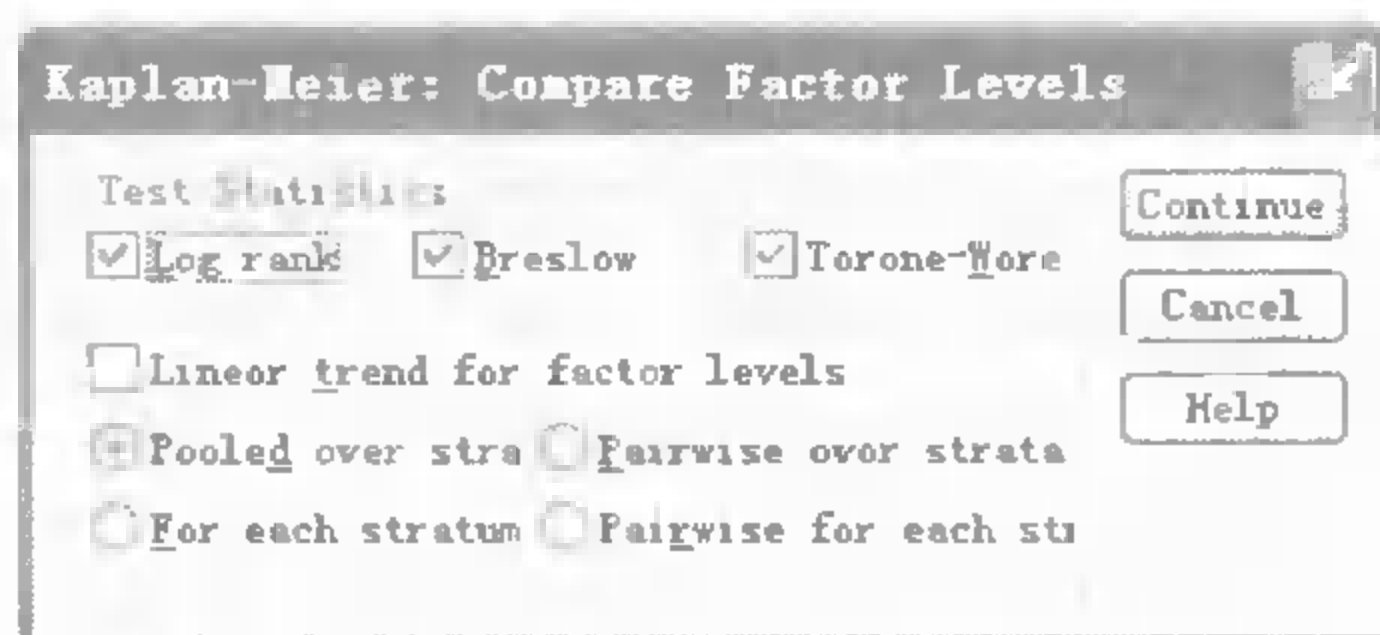


图 13-12 关于因素取值水平的比较设置

① Test Statistics 栏, 在此选择具体的检验统计量, 有如下 3 个选择。

- ① Log rank, 检验各组生存率曲线的分布是否相同, 且各时刻的权重一样。
- ② Breslow, 检验各组生存率曲线的分布是否相同, 并以各时刻的观察例数为权重。
- ③ Tarone-Ware, 检验各组生存率曲线的分布是否相同, 以各时刻观察例数的平方根为权重。

② Linear trend for factor levels 复选框, 用于指定分组因素各水平之间的线性趋势检验。只有当分组因素是有序变量时 (比如疗效的取值: 痊愈、好转、无效), 作线性趋势检验才有实际意义, 在这种情况下, SPSS 假定各水平之间的效应是等距的 (比如: 痊愈与好转之间的差距和好转与无效之间的差距是相同的)。

③ 最后的一组单选框用来指定进行总体比较还是两两比较, 以及对分层变量的处理方式, 可选项有如下 4 个。

- Pooled over strata: 对因素变量各取值水平下的生存曲线作整体比较, 默认选项。
- For each stratum: 按分层变量的不同取值, 对每一层分别进行因素变量各取值水平间的整体比较, 如果没有指定分层变量不作输出。
- Pairwise over strata: 作因素变量各水平之间的两两比较; 此选项对线性趋势检验无效。
- Pairwise for each stratum: 按分层变量的不同取值, 对每一层分别进行因素变量各取值水平间的两两比较; 此选项对线性趋势检验无效。

(4) 保存选项设置。在图 13-10 中单击 Save 按钮, 弹出如图 13-13 所示的对话框, 在此设置保存选项, 单击 Continue 按钮返回主界面。

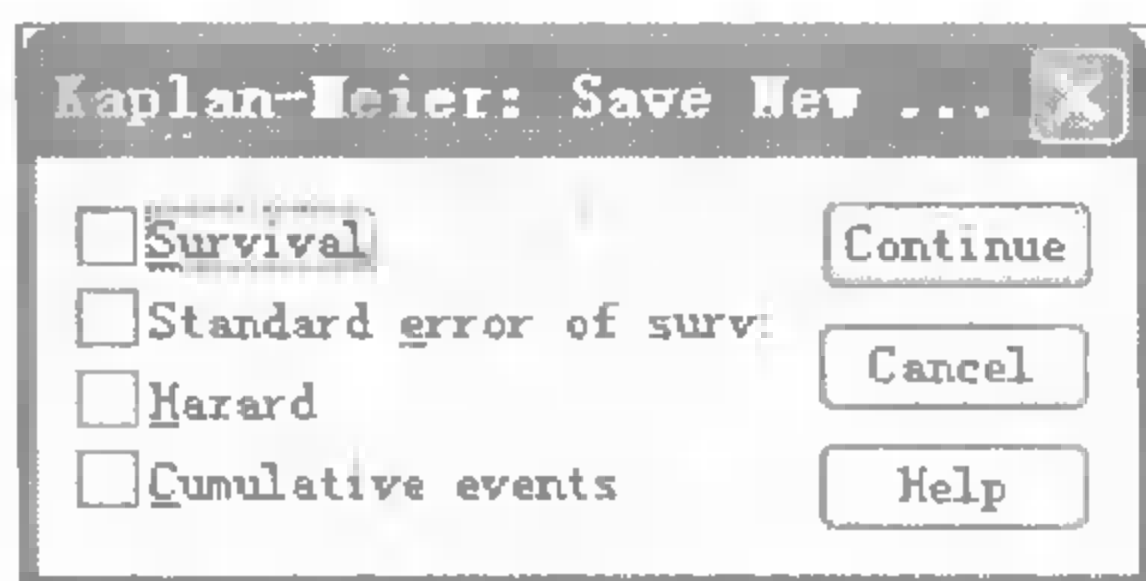


图 13-13 保存选项设置

在图 13-13 中, 可选的保存选项有如下 4 个: Survival (累积生存率的估计值); Standard error of survival (累积生存率估计值的标准误); Hazard (累积风险函数的估计值); Cumulative events (终结事件的累积频数), 按照生存时间和生存状态排序。

(5) 输出选项设置。在图 13-10 中单击 Options 按钮, 弹出如图 13-14 所示的对话框, 在此设置输出选项, 分别勾选如下 4 个复选框: Survival table(s)、Mean and median survival、Quartiles、Survival; 单击 Continue 按钮返回主界面。

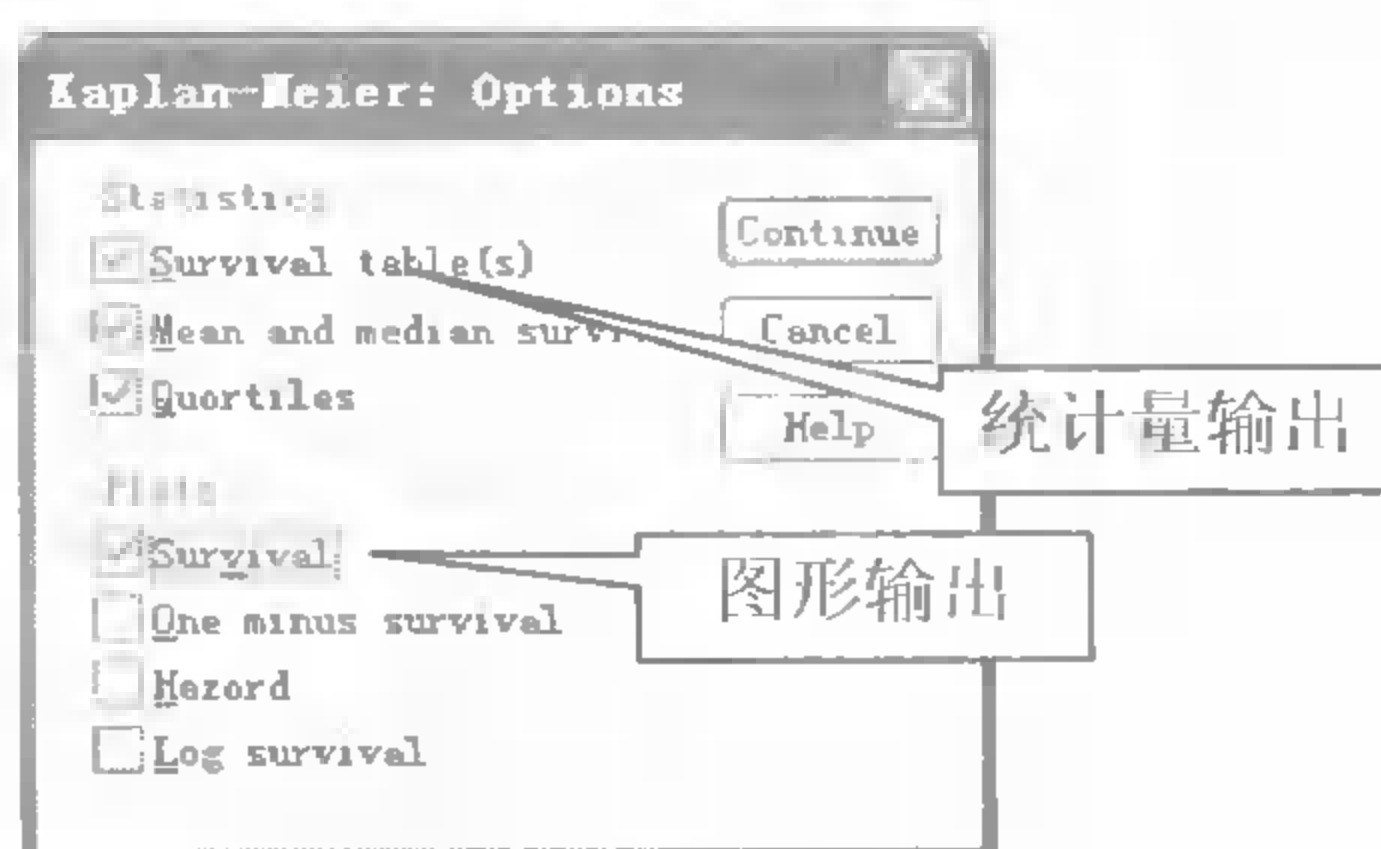


图 13-14 Options 选项设置

- Statistics 栏, 在此选择输出哪些统计量, 可选内容包括: Survival table(s), 生存分析表, 类似于生命表, 只是以个体为单位输出; Mean and median survival, 平均生存时间和中位生存时间, 及其各自的标准误和置信区间; Quartiles, 输出生存时间的三个四分位数。
- Plot 栏, 在此选择输出哪些统计图形, 可选内容包括: Survival, 累积生存函数曲线; One minus survival, 1 减生存函数所得的曲线; Hazard, 累积风险函数的散点图; Log survival, 对数累积生存函数曲线。

### 3. 结果分析

在图 13-10 中单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 13-15~图 13-17 所示。

个案处理摘要					
治疗方法	总数	事件数	删失		
			N	百分比	
新药	104	79	25	24.0%	
旧药	96	74	22	22.9%	
整体	200	153	47	23.5%	

生存表						
治疗方法	时间	状态	此时生存的累积比例		累积事件数	剩余个案数
			估计	标准误		
新药	1	600	有效		1	103
	2	600	有效	.981	2	102
	3	700	有效	.971	3	101
	4	800	有效	.962	4	100
	5	900	有效	.952	5	99
	6	1 100	有效		6	98
	7	1 100	有效		7	97

图 13-15 止疼药分析摘要和生命表输出

生存表的均值和中位数								
治疗方法	均值 <sup>a</sup>				中位数			
	估计	标准误	95% 置信区间		估计	标准误	95% 置信区间	
			下限	上限			下限	上限
新药	4.867	.360	4.162	5.572	3.700	.292	3.126	4.272
旧药	5.185	.350	4.499	5.871	4.100	.1131	1.884	6.316
整体	5.014	.252	4.520	5.507	3.900	.272	3.367	4.433

a. 如果估计值已删失，那么它将限制为最长的生存时间。

治疗方法	25.0%		50.0%		75.0%	
	估计	标准误	估计	标准误	估计	标准误
新药	7.100	.509	3.700	.292	1.900	.226
旧药	7.700	.648	4.100	.1131	2.400	.247
整体	7.300	.371	3.900	.272	2.100	.196

图 13-16 止疼药分析生命表的统计特征

整体比较			
	卡方	df	Sig.
Log Rank (Mantel-Cox)	379	1	.538
Breslow (Generalized Wilcoxon)	748	1	.387
Torone-Ware	705	1	.401

为 治疗方法 的不同水平检验生存分布等同性。

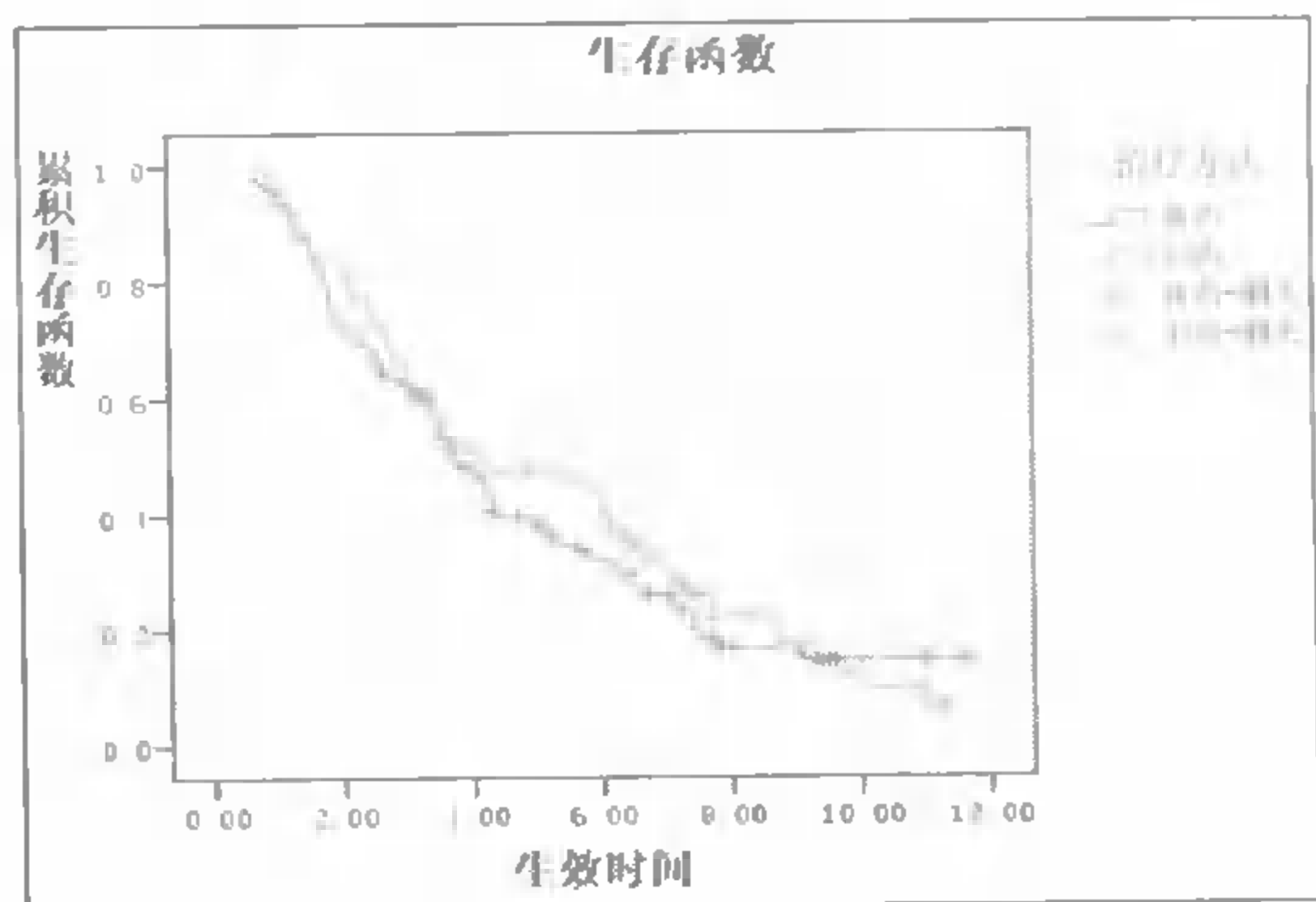


图 13-17 整体比较和生命函数图

(1) 摘要和生命表输出。如图 13-15 所示，“个案处理摘要”表格给出了样本数据的简要统计信息，包括因素变量各取值水平下的事件发生数与未发生数（删失）。

“生存表”给出了类似 Life table 分析中的生命表，只是这里每个观测单独占据一行。

(2) 生命表统计特征输出。如图 13-16 所示，显示的是关于生存表的均值、中位数和百分位数。可见新药、旧药之间，在均值、中位数、四分位数的差异都不是很明显；故可以初步判断，新、旧药品在生效时间上的差异不太明显，更精确的判断需要通过生存函数图和假



设检验完成。

(3) 累积生存函数的图形。如图 13-17 所示, 在“整体比较”中, 三种检验的 Sig 值都很大, 说明新、旧药品之间的生效时间在 0.1 的显著性水平上, 是没有差异的。“生存函数”图是对图 13-15 中的累计生存率的直观描述, 图中显示新药的生存函数多位于旧药生存函数的下面, 说明新药的生效时间要比旧药好一些, 但是从假设检验的结果已知, 这种差异并没有统计学上的显著意义。

## 13.4 Cox 回归模型

Cox 回归模型由英国统计学家 D.R.Cox 于 1972 年提出, 主要用于肿瘤或其他慢性疾病的预后分析, 其优点包括: 是适用于多因素的分析方法、不考虑生存时间的分布形状、能够有效地利用截尾数据。

### 13.4.1 Cox 回归模型的原理简介

生存分析中一个很重要的内容, 就是探索影响生存时间(生存率)的危险因素, 这些因素通过影响各个时刻的死亡风险(危险率)来影响生存率, 例如: 不同特征的人群在某些时刻的危险率函数就是不同的。

#### 1. Cox 模型的基本公式

通常将危险率函数表达为基准危险率函数与相应协变量函数的乘积, 即:  $h(t) = h_0(t) \cdot f(X)$ , 对于协变量函数  $f(X)$ , 最常用的形式是对数线性模型:  $f(X) = \exp\left(\sum_{i=1}^m \beta_i X_i\right)$ 。当基准危险率函数  $h_0(t)$  已知时,  $h(t) = h_0(t) \cdot f(X)$  就为参数模型, 例如:  $h_0(t) = \lambda$  时, 危险率为指数回归模型;  $h_0(t) = \lambda t^{\gamma-1}$  时, 为 Weibull 回归模型;  $h_0(t) = \lambda e^{\alpha t}$  时, 为 Gompertz 模型。

1972 年英国生物统计学家 D.R.Cox 提出在基准危险率函数未知的情况下, 估计模型参数的方法, 称为 Cox 比例风险回归模型 (Cox's proportional hazard regression model), 简称为 Cox 回归模型。该模型的参数估计不依赖于基准危险率的分布类型, 属于一种半参数模型。Cox 模型不直接考察生存函数  $S(t)$  与协变量的关系, 而是用风险率函数  $h(t)$  作为因变量, 并假定:  $h(t, X) = h_0(t) \exp(\beta' X) = h_0(t) \exp(\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_m X_m)$ , 这就是 Cox 模型的基本形式, 其中:

(1)  $h(t, X)$  表示具有协变量  $X$  的个体在时刻  $t$  的危险率(瞬时死亡率);  $X' = (X_1, \cdots, X_m)$  表示可能与生存时间有关的协变量或交互项, 它们可以是定量的或定性的, 且在整个观察期间内不随时间的变化而变化。

(2)  $\beta' = (\beta_1, \beta_2, \cdots, \beta_m)$  为 Cox 模型的偏回归系数, 是需要根据样本估计的参数。 $\beta_i$  表示当其他协变量不变时,  $X_i$  每变化一个单位, 风险率的自然对数变化  $\beta_i$  个单位。若  $\beta_i > 0$ , 该因素为危险因素;  $\beta_i < 0$ , 该因素为保护因素;  $\beta_i = 0$ , 该因素为无关因素。

(3)  $h_0(t)$  是所有危险因素都为 0 时的基础风险率, 它是未知的, 假定它与  $h(t, X)$  是成比例。

(4) 在等式的右侧:  $h_0(t)$  没有明确的定义和假设的分布, 其参数无法估计, 为非参数部分; 而指数部分的参数可以通过样本的实际观察值来估计, 正因为 Cox 模型有非参数和参数

两部分组成，故又称为半参数模型。

## 2. 参数估计和假设检验

利用生存率函数  $S(t, X)$  与风险函数  $h(t, X)$  的关系，可以导出如下的公式成立： $S(t, X) = \exp\left[-\int_0^t h(t, X) dt\right] = \exp\left[-\int_0^t h_0(t) \exp(\beta'X) dt\right] = [S_0(t)]^{\exp(\beta'X)}$ ，它反映了协变量  $X$  与生存函数的关系。

偏回归系数  $\beta$  的估计需要借助于偏似然函数，有了它之后再对基础风险函数和风险函数做出估计。对于模型中变量的取舍原则，有如下几种假设检验方法可供选择。

(1) 似然比检验 (likelihood ratio test)，可用于模型中原有不显著变量的剔除和新变量的引入，以及包含不同变量的各模型的比较。

(2) 得分检验 (score test)，可用于检验一个或多个新变量能否引入模型，也可用于检验变量间的交互作用是否显著。

(3) Wald 检验，用于检验模型中的变量是否应被剔除；它还可按照置信区间的大小来推断模型内的参数是否为 0，方法是当偏回归系数 95% 的置信区间包含 0 时，就认为它与 0 无限制差异。

## 3. Cox 回归模型的分析步骤

首先 Cox 回归基本模型需要满足如下两个前提假设：各危险因素的作用大小不随时间变化而变化；各危险因素之间不存在交互作用。然后按照如下步骤进行分析。

- (1) 明确所研究问题的自变量和因变量。
- (2) 利用样本估计参数，拟合模型。
- (3) 做关于模型中的变量取舍的假设检验，以及模型的拟和优度检验。
- (4) 模型的解释及应用。

### 13.4.2 Cox 回归实例分析

本节仍以电信数据为例来分析客户流失的问题，在第 13.2.3 节，曾用生命表法研究过这个问题，所用数据文件为“电信客户流失数据.sav”，数据格式如图 13-1 所示。

注意：Cox Regression 过程要求时间变量为数值型的；事件变量可以为连续变量或分类变量；自变量 (covariates，协变量) 可以为分类的或连续的，如果是分类的，则必须为虚拟变量 (dummy-coded，哑变量) 或指示变量 (indicator-coded)；该过程还可以设置对分类自变量进行自动编码；分层变量必须为分类变量，取值可以是短字符串型或整数型的。

#### 1. Cox Regression 分析过程的参数设置

依次单击菜单“Analyze→Survival→Cox Regression...”，执行生命表分析过程，其主设置界面如图 13-18 所示，在此设置分析变量及其取值规则。

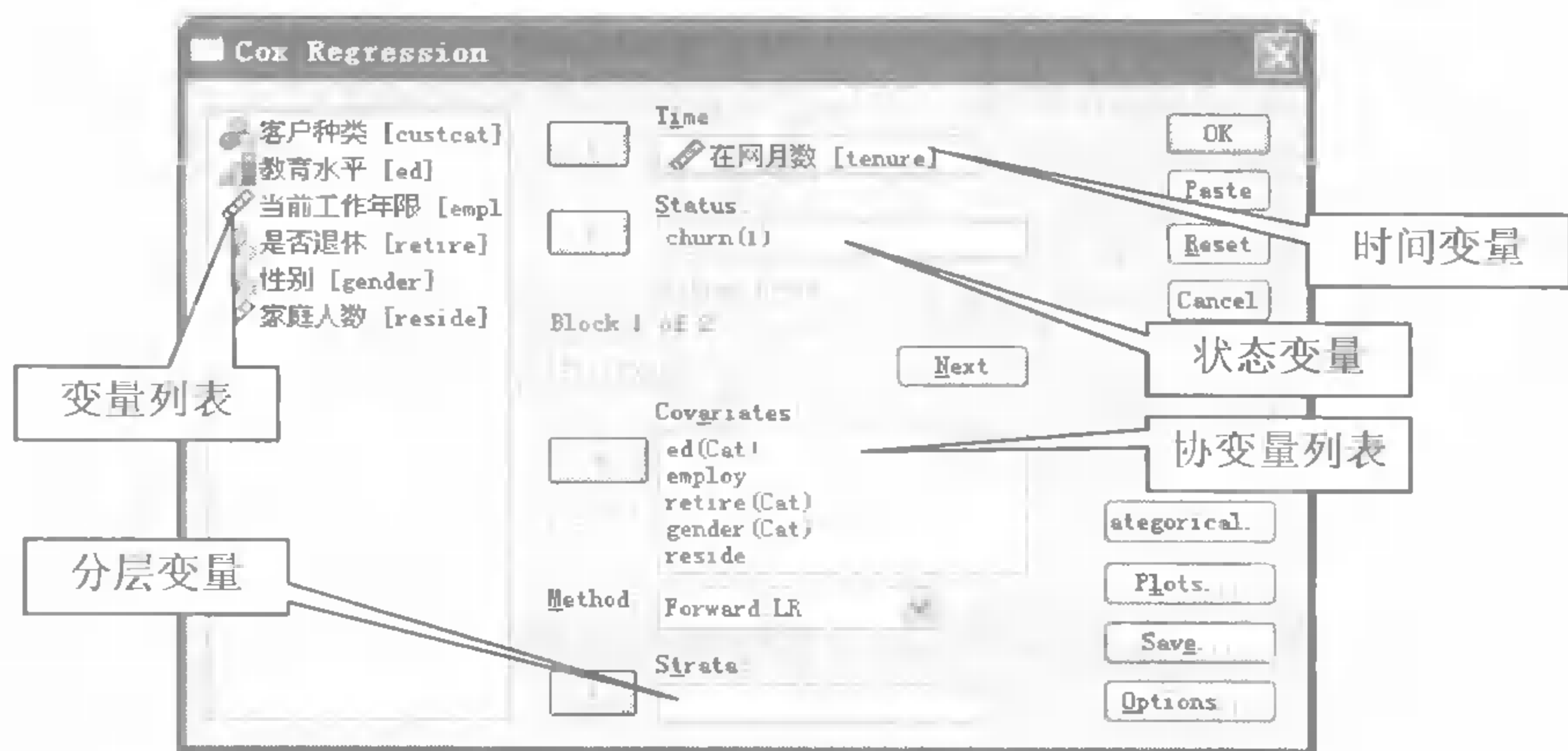




图 13-18 Cox Regression 设置面板

(1) 指定分析变量。在变量列表单击选中在网月数变量，单击从上至下第一个  按钮，将其作为时间变量选入 Time 选框；在变量列表单击选中是否流失变量，单击从上至下第二个  按钮，将其作为状态变量选入 Status 选框；单击 Define Event 按钮，弹出如图 13-19 所示的定义事件对话框，在 Single value 后输入“1”，单击 Continue 按钮返回主界面。

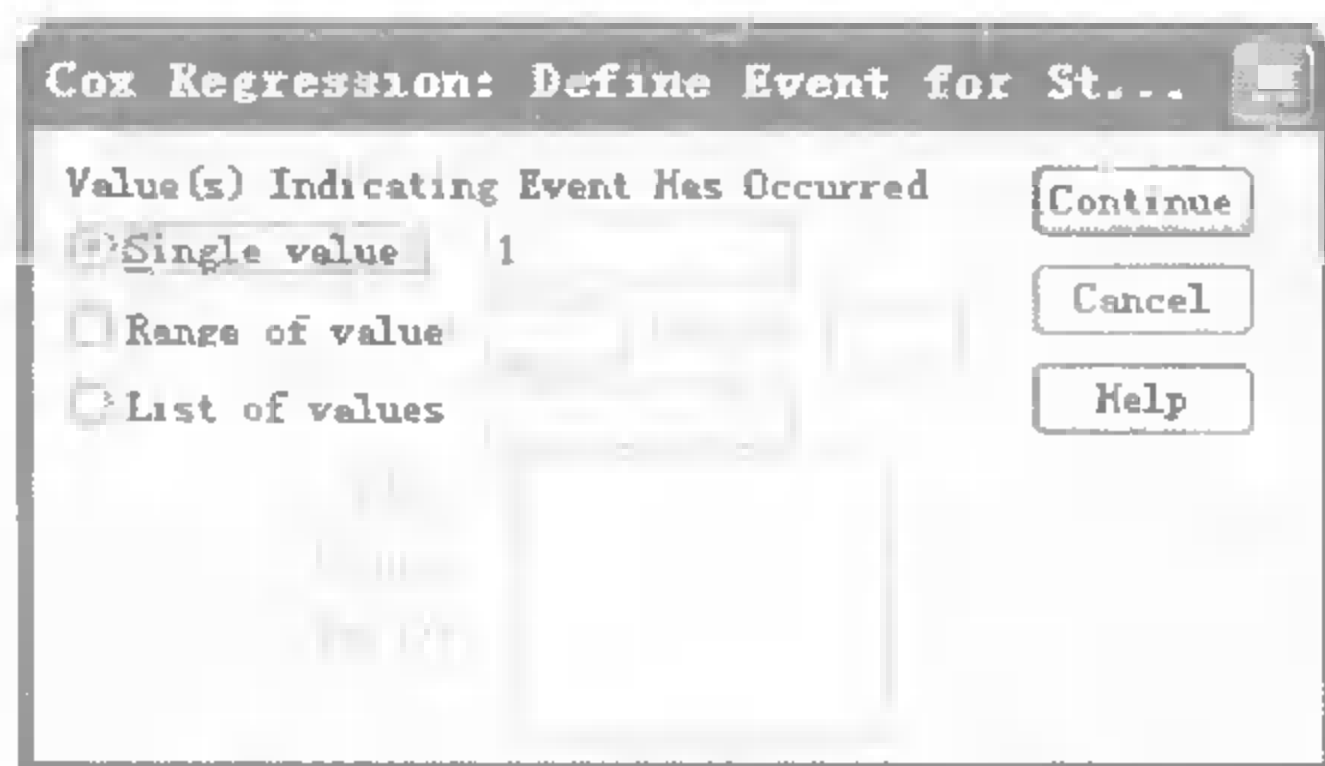




图 13-19 取值定义子对话框

在图 13-18 中：Time 选框，用于选入生存时间变量；Status 选框，用于选入生存状态变量；Covariates 列表框，用于选入协变量；Strata 选框，用于选入分层因素，可以看作是研究者欲加以控制的混杂因素。

图 13-19 所示的定义事件取值的对话框，与图 13-11 完全一样，设置方法也相同。

(2) 协变量设置。在变量列表选中从教育水平到家庭人数的 5 变量，单击从上至下第三个  按钮，将其作为第一组协变量 (Block 1) 选入 Covariates 列表框，单击 Method 下拉列表指定这组协变量的变量选择方法为 Forward LR；单击 Next 按钮打开第二组的 Covariates 列表框，在变量列表单击选中客户种类变量，单击从上至下第三个  按钮，将其作为第二组协变量 (Block 2) 选入 Covariates 列表框，保留 Method 下拉列表的 Enter 选项；单击 Previous 按钮返回第一组协变量的 Covariates 列表。


此处的 Method 下拉列表，用于指定协变量进入回归模型的方式，有如下 7 个可选项。

- ① Enter 强行进入法，同一组中的协变量，一次性地全部进入回归方程。
- ② Forward Condition 向前选择法，通过条件似然检验确定协变量是否能进入回归方程。
- ③ Forward LR 向前选择法，通过似然率检验确定协变量是否能进入回归方程。
- ④ Forward Wald 向前选择法，通过 Wald 检验确定协变量是否能进入回归方程。
- ⑤ Backward Condition 向后消去法，通过条件似然检验确定协变量能否从方程中消去。

➤ Backward LR 向后消去法, 通过似然率检验确定协变量能否从方程中消去。

➤ Backward Wald 向后消去法, 通过 Wald 检验确定协变量能否从方程中消去。

一般来说, 使用向后消去法更可能避免漏掉潜在的有价值的预测因子; 如果要求至少有一个协变量进入模型, 建议使用向前选择法。

(3) 分类协变量设置。在图 13-18 中, 单击 Categorical 按钮, 弹出如图 13-20 所示的对话框, 在此设置如何处理分类协变量。在变量列表选中 custcat、ed、gender 和 retire 这 4 个变量, 单击  按钮, 将其选入 Categorical Covariates 列表框; 单击 Continue 按钮返回主界面。

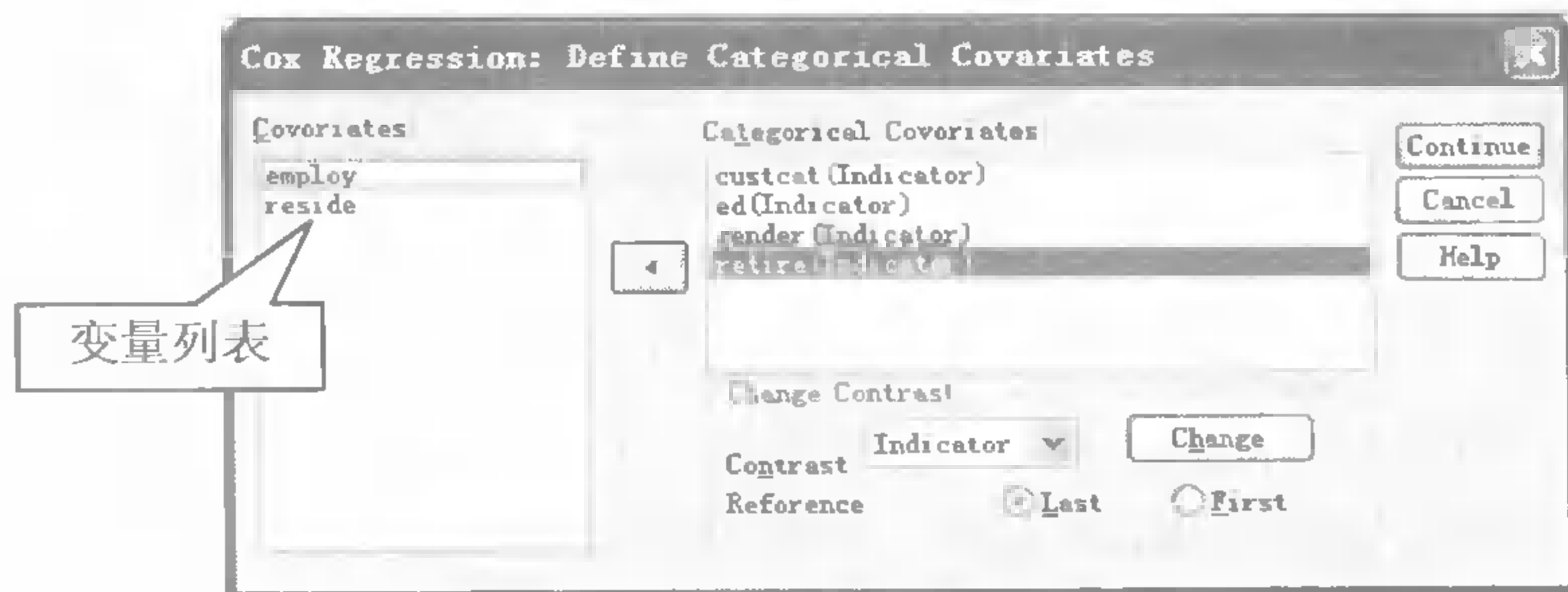


图 13-20 分类协变量的编码设置

① Covariates 列表框, 显示主设置面板中选定的所有变量; Categorical Covariates 列表框, 用于选入指定为分类变量的协变量, 变量名后的括号显示当前变量正在使用的对照方法。

② Change Contrast 栏, 用于设置对指定协变量的对照(编码)方式, 修改后需单击 Change 按钮确认, Contrast 下拉列表给出的对照方式有如下 7 个。

- Indicator 指示器, 用于指示是否属于某一个分类, 参考分类在对比矩阵中整行均为 0。
- Simple 简单比较, 预测变量的每个分类 (参考分类除外) 都与参考分类进行比较。
- Difference 差分比较, 除第 1 类外, 预测变量的每个分类都与其前所有分类的平均效应进行比较, 也叫逆 Helmert 比较。
- Helmert (Helmert 比较), 除最后 1 类外, 预测变量的每个分类都与后面所有分类的平均效应进行比较。
- Repeated 重复比较, 除第 1 类外, 预测变量的每个分类都与其前的所有类别进行比较。
- Polynomial 多项式比较, 此方法假设各类别的间距相等, 仅适用于数值型变量。
- Deviation 差别比较, 预测变量的每个分类 (参考分类除外) 都与总体效应进行比较。

③ Reference 栏, 用于指定参考分类。

如果在上面选择了 Deviation、Simple 或 Indicator 方法, 就需要指定 1 个参考类别, 可选项有: First, 第 1 类; Last, 最后 1 类; 默认的参考类都是 Last, 修改后需要单击 Change 按钮确认。对于选择了 First 作为参考类的变量, 其在 Categorical 列表的名称后面会以嵌套括号的方式显示“First”字样, 例如: x1 (simple (first)), 表示分类协变量 x1 的对照方式是 simple, 参考类为 first; 选择 Last 为参考类时不作提示。

(4) 图形设置。在图 13-18 中单击 Plots 按钮, 弹出如图 13-21 所示的对话框, 在此设置关于输出图形的选项。勾选 Survival、Hazard 复选框; 在 Covariate 列表单击选中 custcat 变量,



单击  按钮，将其选入 Separate 选框；单击 Continue 按钮返回主界面。

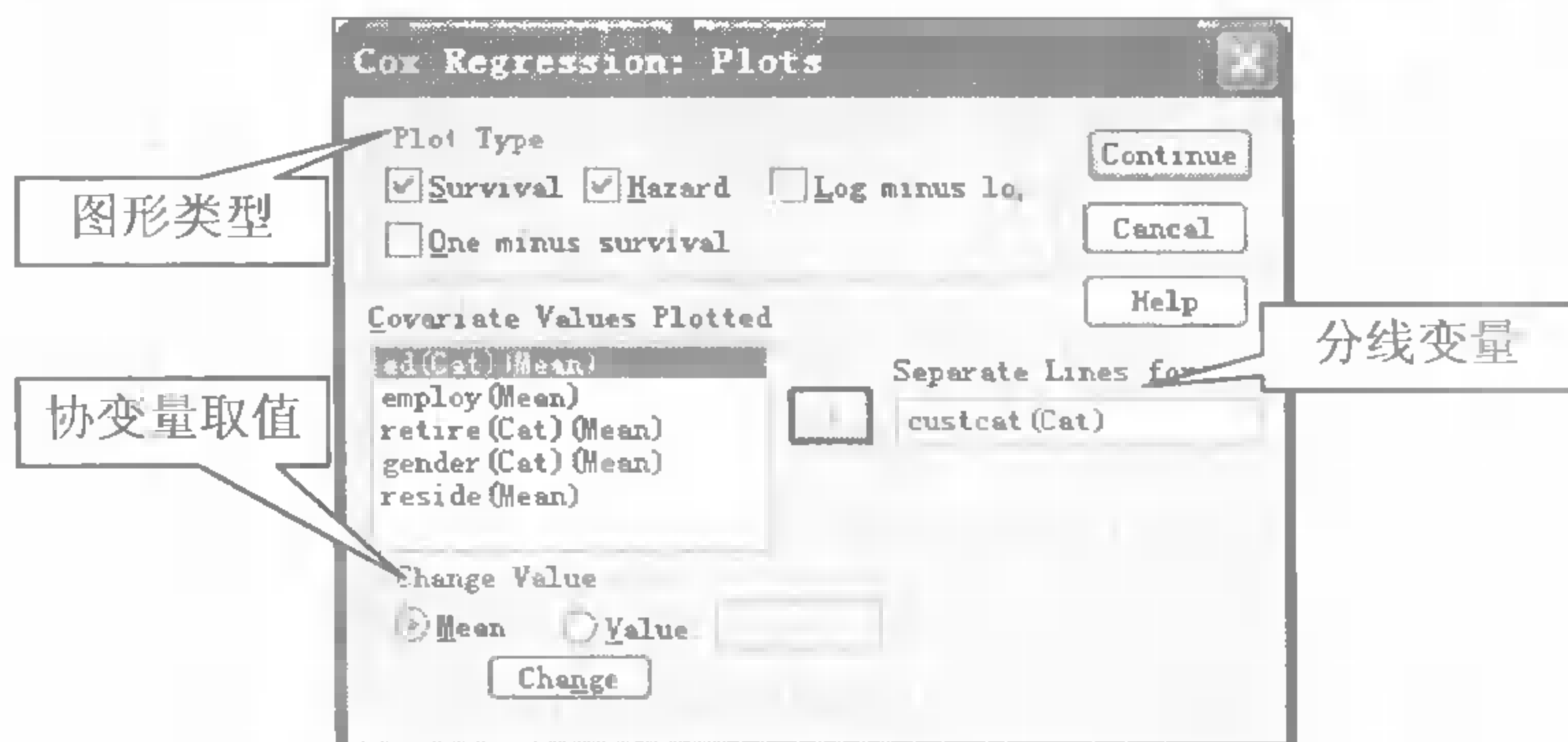


图 13-21 作图选项的设置

① Plot Type 栏，用于选择输出哪些类型的图形，可选项有如下几个：Survival，线性刻度的累计生存函数图；Hazard，线性刻度的累计危险函数图；Log minus log，对数转换（ $\ln(-\ln)$  transformation）后的累计生存函数图；One-minus survival，1 减累计生存函数的图形。

② Covariate Value Plotted 栏，只有指定协变量为固定值时，才能作生存函数（或风险函数）关于时间的图形；默认状态下，这里的固定值取的都是协变量各自的均值（Mean）。需要修改这个固定值时，先在 Covariate 列表选中某个变量，然后在 Change Value 栏单击 Value 选项，并在其后的输入框指定相应的固定值，单击 Change 按钮确认修改即可。

③ Separate Lines for 选框，用于选入一个分类协变量，作图时将它作为分线变量，对其每个取值将分别作一条曲线。

(5) 保存选项设置。在图 13-18 中，单击 Save 按钮，弹出如图 13-22 所示的对话框，在此设置保存选项；单击 Continue 按钮返回主界面。

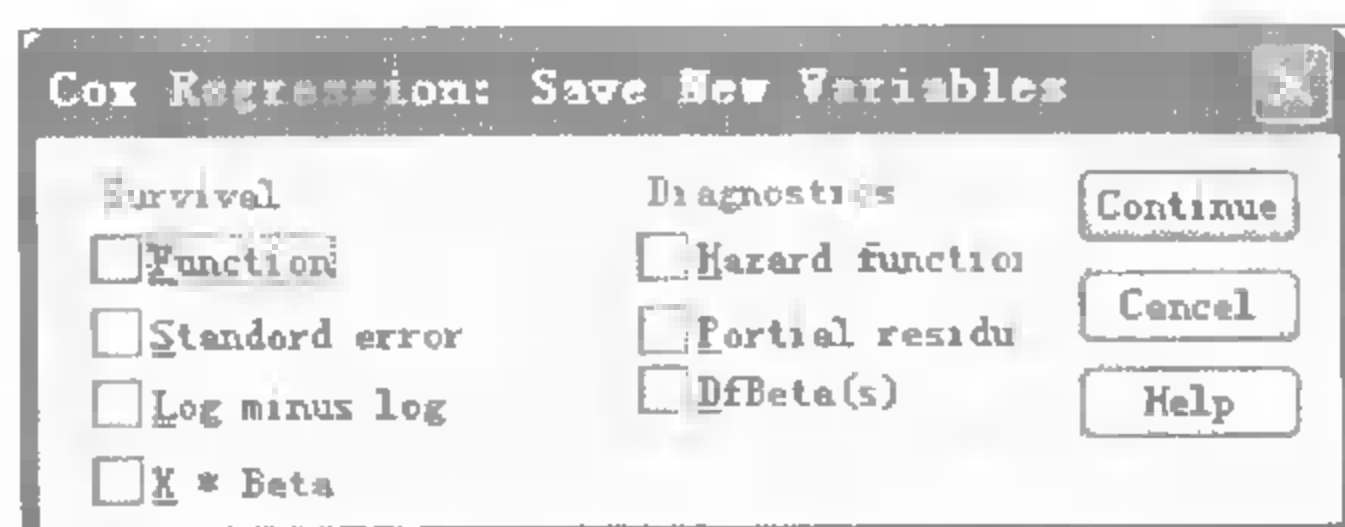


图 13-22 保存选项的设置

在此，SPSS 给出了如下 3 部分可选的保存内容。

- Survival 生存变量栏，可选内容包括：Function，生存函数估计值；Standard error，生存函数估计值的标准误；Log minus log，对数转换（ $\ln(-\ln)$  transformation）后的累计生存函数。
- Diagnostics 诊断变量栏，可选内容包括：Hazard function，累积危险函数估计值；Partial residuals，偏残差，用它对生存时间作图可以检验关于风险函数的比例假设；DfBeta(s)，剔除某个观测后引起的参数估计值的变化，对最终模型的每个协变量都生成一个新变量用于保存。
- X\*Beta 复选框，保存线性预测的得分，由中心化协变量与估计参数相乘后再求和所得。

(6) 输出选项设置。在图 13-18 中，单击 Options 按钮，弹出如图 13-23 所示的对话框，在此设置模型中的一些统计量；单击 Continue 按钮返回主界面。

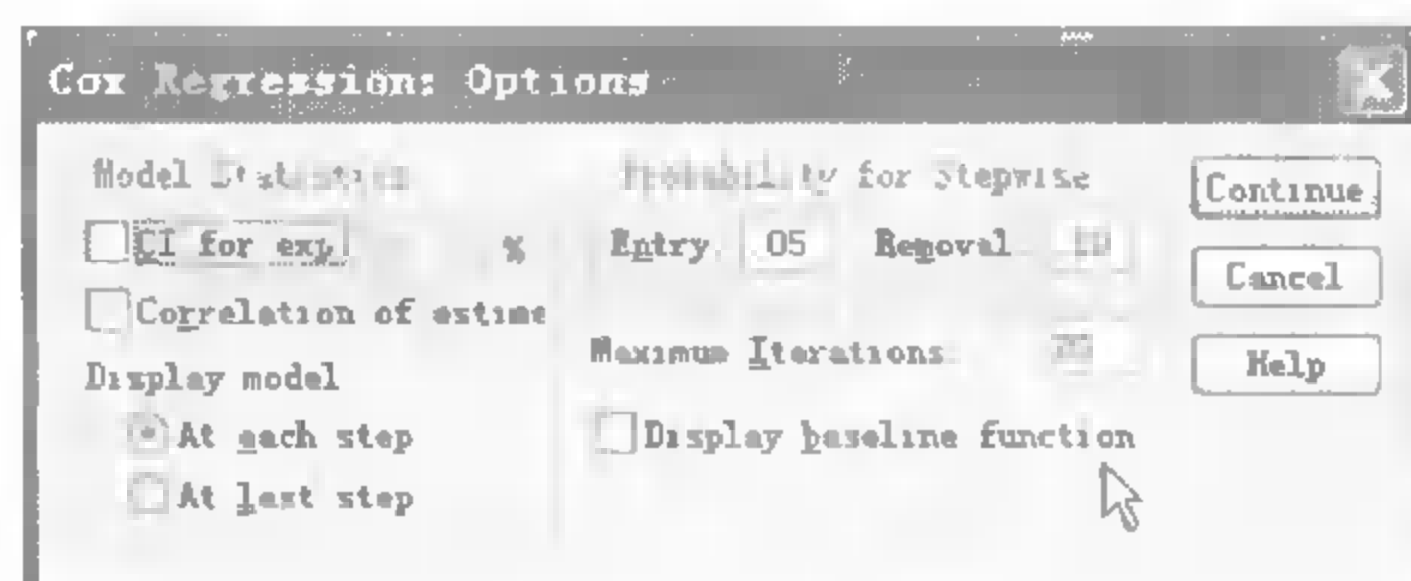


图 13-23 输出选项的设置

① Model Statistics 栏，设置模型统计量，可选项有：CI for exp(B)，exp(B) 的置信区间，默认为 95%；Correlation of estimates，系数估计值的相关矩阵。Display model 栏，指定输出方式为如下两种中的一个：At each step，逐步回归的每一步都输出相关的统计量；At last step，只在逐步回归的最后一步输出相关的统计量。

② Probability for Stepwise 栏，指定协变量进入或剔除出模型的临界概率。Entry 输入框指定变量进入模型的临界值，默认为 0.05；Removal 输入框指定变量移出模型的临界值，默认为 0.10。

③ Maximum Iterations 输入框，指定最大的迭代次数，默认为 20。

④ Display baseline function 生成基准危险函数、协变量均值生存函数和危险函数表。

## 2. 结果分析

在图 13-18 中单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 13-24~图 13-29 所示。

案例处理摘要					
			N	百分比	
分析中可用的案例	事件=		274	27.4%	
	删失		726	72.6%	
	合计		1000	100.0%	
删除的案例	带有缺失值的案例		0	0%	
	带有负时间的案例		0	0%	
	层中的最早事件之前删失的案例		0	0%	
	合计		0	0%	
合计			1000	100.0%	

a. 因变量 = 在网月数

分类变量编码 <sup>c, d, e, f</sup>						
	频率	(1) <sup>a</sup>	(2)	(3)	(4)	
custcat <sup>b</sup>	1=基本服务	266	1	0	0	
	2=上网服务	217	0	1	0	
	3=附加服务	281	0	0	1	
	4=所有服务	236	0	0	0	
ed <sup>b</sup>	1=低于高中	204	1	0	0	0
	2=高中	287	0	1	0	0
	3=大学	209	0	0	1	0
	4=学士	234	0	0	0	1
	5=硕士	66	0	0	0	0
retire <sup>b</sup>	00=没退休	953	1			
	1 00=退休	47	0			
gender <sup>b</sup>	0=男	483	1			
	1=女	517	0			

a. 已经记录了 (0,1) 变量，所以其系数不会与指示符 (0,1) 编码相同。  
b. 示性参数编码  
c. 分类变量 custcat (客户种类)  
d. 分类变量 ed (教育水平)  
e. 分类变量 retire (是否退休)  
f. 分类变量 gender (性别)

图 13-24 数据摘要和分类变量编码输出

模型系数的综合测试 <sup>c, d</sup>										
步骤	-2 倍对数似然值	整体 (得分)			从上一块开始更改			从上一块开始更改		
		卡方	df	Sig.	卡方	df	Sig.	卡方	df	Sig.
1 <sup>a</sup>	3357.318	131.930	1	.000	169.046	1	.000	169.046	1	.000
2 <sup>b</sup>	3344.089	147.561	5	.000	13.229	4	.010	162.275	5	.000

a. 在步骤编号 1 employ 处输入变量  
b. 在步骤编号 2 ed 处输入变量  
c. 起始块编号 0，量初的对数似然函数 -2 倍对数似然值 3526.364  
d. 起始块编号 1 方法 = 向前逐步 (似然比)

图 13-25 block1 中向前逐步法的系数检验

方程中的变量						
		B	SE	Wald	df	Sig.
步骤 1	employ	-100	.009	118.894	1	.000
步骤 2	ed			12.820	4	.012
	ed(1)	-.535	.262	4.165	1	.041
	ed(2)	-.367	.229	2.569	1	.109
	ed(3)	-.106	.231	.209	1	.647
	ed(4)	.080	.216	.135	1	.713
	employ	-.034	.009	102.137	1	.000

不在方程中的变量 <sup>a, b</sup>				
步骤		得分	df	Sig.
步骤 1	ed	13.044	4	.011
	ed(1)	4.956	1	.028
	ed(2)	3.323	1	.068
	ed(3)	.280	1	.597
	ed(4)	7.277	1	.007
	retire	3.852	1	.050
	gender	.001	1	.969
	reside	2.256	1	.133
	custcat	31.142	3	.000
	custcat(1)	12.794	1	.000
步骤 2	custcat(2)	6.417	1	.011
	custcat(3)	13.102	1	.000
	retire	2.897	1	.089
	gender	.019	1	.891
	reside	2.272	1	.132
	custcat	32.368	3	.000
	custcat(1)	22.412	1	.000
	custcat(2)	9.123	1	.003
	custcat(3)	8.876	1	.003

a. 残差卡方 = 带有 10 df Sig. 的 48.283。 = .000  
b. 残差卡方 = 带有 6 df Sig. 的 36.918。 = .000

图 13-26 block1 中的变量

模型系数的综合测试 <sup>a, b</sup>									
-2 倍对数似然值	整体 (得分)			从上一块开始更改			从上一块开始更改		
	卡方	df	Sig.	卡方	df	Sig.	卡方	df	Sig.
3312.741	176.815	8	.000	31.348	3	.000	31.348	3	.000

a. 起始块编号 0，量初的对数似然函数 -2 倍对数似然值 3528.364  
b. 起始块编号 2 方法 = 输入

图 13-27 block2 中输入方法的系数检验

方程中的变量						
	B	SE	Wald	df	Sig.	Exp(B)
ed			12.935	4	.012	
ed(1)	-.613	.278	4.843	1	.028	.542
ed(2)	-.357	.237	2.285	1	.132	.700
ed(3)	-.148	.236	.394	1	.530	.862
ed(4)	.102	.217	.221	1	.638	1.107
employ	-.090	.009	93.480	1	.000	.914
custcat			31.084	3	.000	
custcat(1)	.325	.168	3.742	1	.053	1.384
custcat(2)	-.486	.170	8.204	1	.004	.615
custcat(3)	-.530	.195	7.398	1	.007	.589

不在方程中的变量 <sup>a</sup>			
	得分	df	Sig.
retire	2.227	1	.136
gender	.008	1	.929
reside	1.967	1	.161

a. 残差卡方 = 带有 3 df Sig. 的 4.695。 = .196

图 13-28 block2 中的变量

协变量均值和模式值					
	均值	模式			
		1	2	3	4
ed(1)	204	204	204	204	204
ed(2)	287	287	287	287	287
ed(3)	209	209	209	209	209
ed(4)	234	234	234	234	234
employ	10.987	10.987	10.987	10.987	10.987
retire	953	953	953	953	953
gender	483	483	483	483	483
reside	2.331	2.331	2.331	2.331	2.331
custcat(1)	.266	1.000	.000	.000	.000
custcat(2)	.217	.000	1.000	.000	.000
custcat(3)	.281	.000	.000	1.000	.000

图 13-29 协变量均值和模式值

(1) 数据摘要和分类变量编码表。如图 13-24 所示，“案例处理摘要”表格给出了数据的简要统计信息，其中删失 (censored) 一行表示事件 (客户流失) 没有发生的观测个数 (726)，删失记录不会用于计算回归系数，但要用于计算基准危险率。

“分类变量编码”给出了对分类变量自动编码的结果，它有助于解释分类协变量 (尤其

是二元变量)的回归系数。默认情况下,参考类为分类变量取值的最后一个类别,例如对于性别变量(gender):原始数据中“女”的取值为1,但在回归中“女”的编码为0。

(2) Block1 中变量的系数检验。如图 13-25 所示,是向前逐步回归(Forward:LR 方法)的系数检验结果,如果加入一个变量后卡方更改量的显著性小于 0.05,则加入此变量是合理的,例如此处在第 1、2 步分别加入变量 employ 和 ed 都是合理的;反之,如果删除一个变量后卡方更改量的显著性大于 0.10,则去除此变量是合理的。

(3) Block1 中的变量。如图 13-26 所示,是向前逐步回归估计完成后,模型中取舍的变量情况:经过两步迭代,保留了 employ 和 ed 这两个变量。在“方程中的变量”表格的步骤 2 中:employ 变量的 Exp(B) 列的取值表示,一个用户为当前雇主的工作时间每增加一年,其流失的危险率就降低  $100\% - (100\% \times 0.911) = 8.9\%$ ; 当一个用户为当前雇主的工作时间达到三年后,其流失的危险率就降低了  $100\% - (100\% \times (0.911^3)) = 24.4\%$ ; 对于受教育程度,ed(5)是参考类(硕士,编码为 0),ed(1)行(低于高中)的系数检验 Sig 值小于 0.05,说明它与 ed(5)在流失率上的差异是统计显著地,从 Exp(B) 列看,ed(1)的流失危险率是 ed(5)的 0.586 倍。

(4) Block2 中变量的系数检验。如图 13-27 所示,是强行输入法(Enter 方法)的系数检验结果,从卡方更改量的显著性远小于 0.01 看,加入客户种类(custcat)变量是合理的。

(5) Block2 中的变量。如图 13-28 所示,是强行输入法估计完成后,模型中取舍的变量情况:此时仍保留的变量有 employ、ed 和 custcat,对于客户种类,custcat(4)是参考类(所有服务)。在“方程中的变量”表格里,custcat(1)行(基本服务)的 Exp(B) 系数,表示基本服务类型客户的危险率是所有服务类型客户的 1.384 倍,但是它的显著性检验 Sig 值为 0.053,所以在 0.05 的显著性水平上,认为这个差异是随机的(不显著);对于上网服务和附加服务两类客户,其系数的显著性 Sig 值都小于 0.01,所以它们与参考类之间的差异都是极其显著的,其中上网服务客户的流失率是所有服务客户的 0.615 倍,附加服务客户的流失率是所有服务客户的 0.589 倍。

而在“不在方程中的变量”表格里,所有变量的显著性检验 Sig 值都大于 0.1。

(6) 协变量的均值输出。如图 13-29 所示,给出了每个预测变量的均值;以及在 Plots 子设置面板指定的作图协变量的各个模式,本例为 custcat(客户种类)变量,其他变量的模式统一显示为同行的均值。关于生存函数和风险函数的图形正是根据此均值和模式表所作的。

(7) 生存函数图。如图 13-30 所示,“协变量均值处的生存函数”图是各协变量取均值时(如图 13-29 所示)的累积生存函数图形;“模式 1-4 的生存函数”图是按客户类型分组后的累积生存函数图,可见基本服务和所有服务两类客户的生存函数曲线偏低,这与通过分析图 13-28 所得的关于流失危险率的比较信息是一致的。

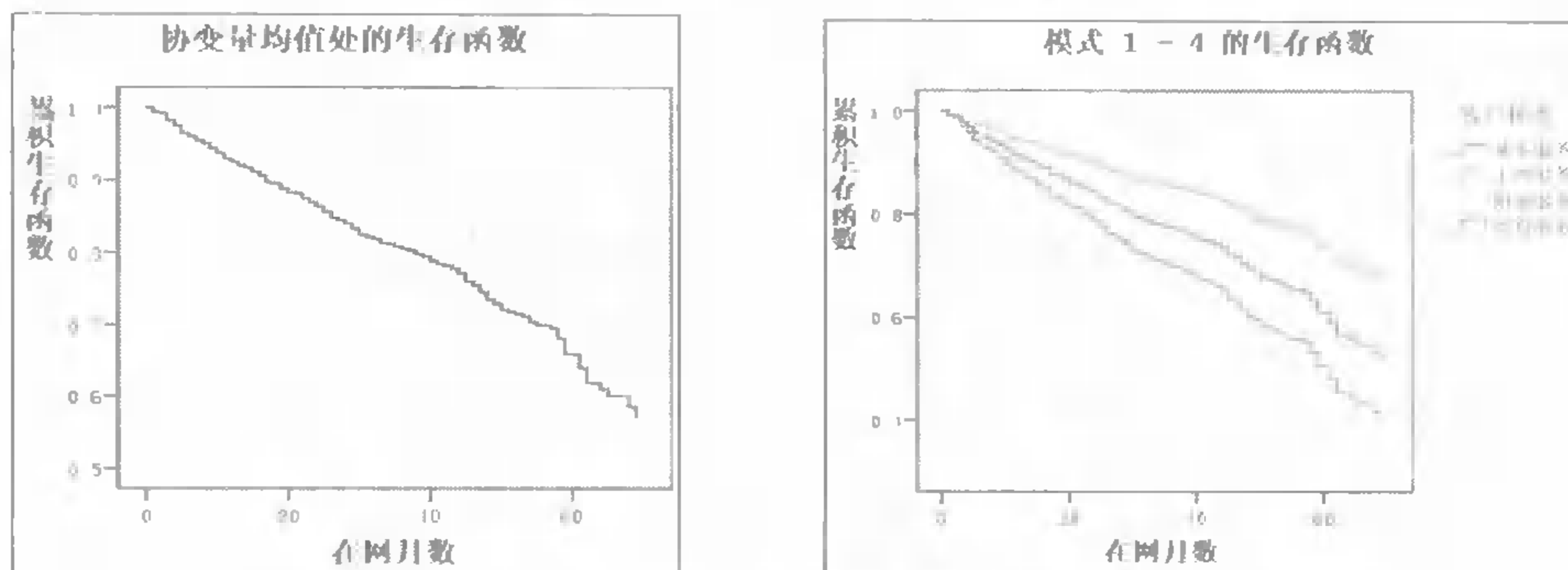


图 13-30 生存函数的图形



(8) 危险函数图。如图 13-31 所示,“协变量均值处的危险函数”图是各协变量取均值时(如图 13-29 所示)的累计危险函数图形;“模式 1-4 的危险函数”图是按客户类型分组后的累计危险函数图。它们所反映的信息与图 13-30 完全类似。

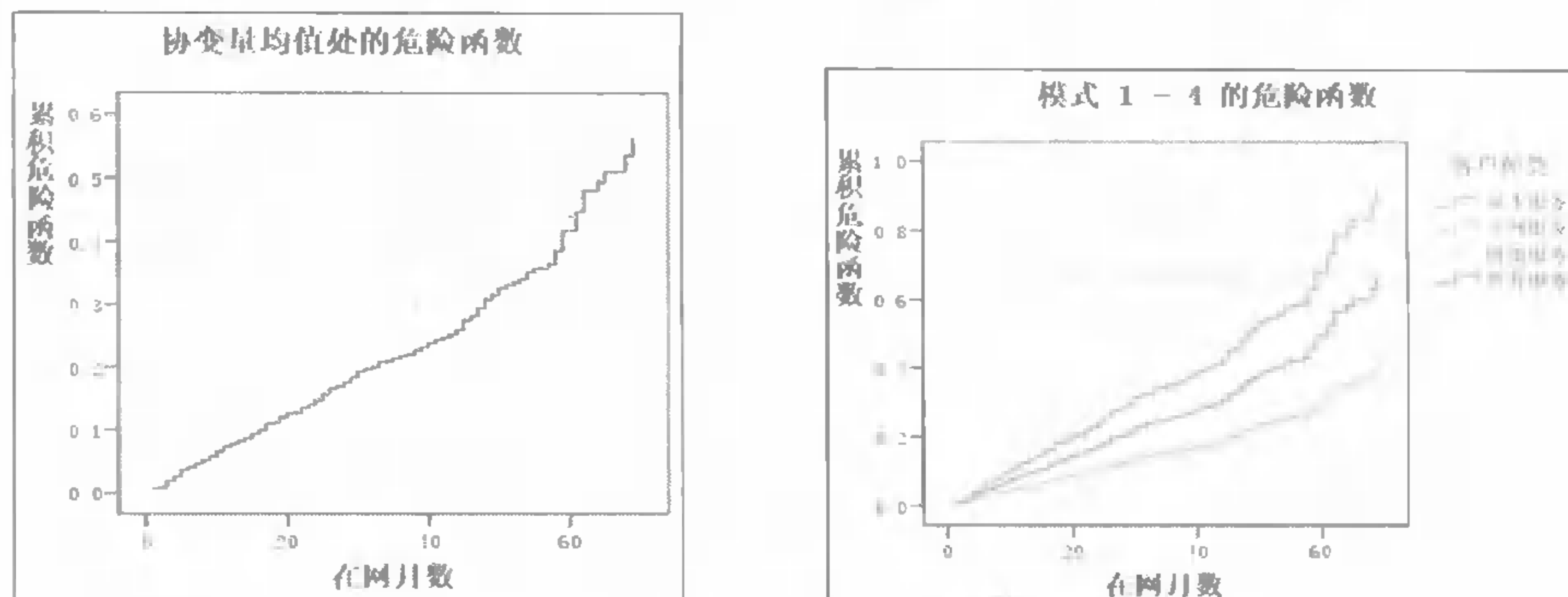


图 13-31 危险函数的图形

调查问卷是收集数据的重要途径，它广泛应用于科学研究和生产实践的方方面面。当我们收回成百上千份的调查问卷后，最关心的是问卷表中的题目能否反映调查意图，以及所得数据的可靠性怎样。如何设计出可靠性好、实用性强的调查问卷表，如何从调查数据中提取真实可靠的信息，是一项十分重要的工作。本章就来介绍调查问卷表的信度分析方法。

## 14.1 信度分析

信度分析用于评价问卷的稳定性或可靠性，它检验用问卷对同一事物进行重复测量后所得结果的一致性程度，还可用于判断问卷中的不同问题是否是针对同一个目标所设的。

### 14.1.1 信度分析的基本原理

信度 (Reliability)，指对同一事物的重复测量结果的一致性程度，它能够反映测量工具的稳定性或可靠性，一般用信度系数表示。信度本身与测量结果的正确与否无关，它的用途在于检验测量本身是否稳定。

#### 1. 信度简介

按照评价对象的不同，信度可以分为如下两类。

(1) 内在信度。衡量调查表中的某一组问题测量的是否是同一个概念，如果内在信度系数达到 0.8 以上，就认为调查表有较高的内在一致性，常用的有 Cronbach  $\alpha$  系数和分半信度。

(2) 外在信度。衡量用同一问卷在不同时间对同一对象进行重复测量，所得结果之间的一致性程度，也称为重测信度。

一般而言，如果量表的信度系数达到 0.9 以上，该测验或量表的信度就较好；信度系数在 0.8 以上，是可以接受的；如果在 0.7 以下，就应该对此量表进行修订；如果低于 0.5，则此量表的调查结果就很不可信了。

信度只是用来衡量一致性或稳定性的指标。一致性高的问卷是指，同一群人接受性质相同、题型相同、目的相同的不同问卷测验后，在各结果之间显示出较强的正相关性。稳定性高的测量工具是指，一群人在不同的时空条件下，接受相同工具的测量后，所得结果的差异很小。测验的信度越高，表示测验结果越可信，但也不能期望两次测验的结果完全一致，信度除受测验质量的影响外，还受很多其他因素的影响。

## 2. 信度分析的方法

下面介绍 SPSS 中的信度估计方法, 应用时需要根据数据类型来选择不同的方法。

### (1) Cronbach $\alpha$ 系数。

$\alpha$  系数的计算公式为:  $\alpha = \frac{k}{k-1} \left( 1 - \frac{\sum_{i=1}^k S_i^2}{S_x^2} \right)$ , 其中: 量表共有  $k$  个题目,  $n$  个观测,  $S_i$  为

第  $i$  题得分的方差,  $S_x$  为测验总得分的方差。它用来衡量调查表中多个问题的得分之间的一致性, 适用于答案为多重记分的问卷; 还可用于测量李克特量表 (Likert-type Scale) 的信度。

$\alpha$  信度系数与量表的题目数量关系密切。一个含有约 10 个题目的量表,  $\alpha$  系数应能达到 0.8 以上; 如果题目增加到多于 20 个时,  $\alpha$  系数会很容易地升至 0.9 以上; 如果量表的题目减少,  $\alpha$  系数也会随之降低。因此判断量表信度时, 首先应当了解该量表所含题目的数量, 再以此为基础判断  $\alpha$  系数是否达到了可以接受的水平。

(2) 分半信度。任何测验都是对所有可能题目的一份取样, 如果抽取的部分不同, 就能编制很多平行的等值测验, 这叫做复本 (即内容和形式都相似的测验), 例如考试时用到的 A、B 卷。

如果一种测验有两个以上的复本, 根据被试者接受两个复本测验的得分, 计算相关系数, 就可以得到复本信度。但建立复本比较困难, 在没有复本且测验只能实施一次的情况下, 通常采用分半法估计信度, 即把测验题目分成对等的两半, 根据各人在这两半测验中的分数, 计算其相关系数作为信度指标。

计算公式为:  $r_{xx} = \frac{2r_{hh}}{1+r_{hh}}$ , 其中:  $r_{hh}$  为两半测验分数之间的相关系数,  $r_{xx}$  为整个测验的

信度估计值。注意: 测验的题目数量较少时, 比如 10 题以下, 就不适合采用这种方法。

计算分半信度时, 要求人为分开的两部分测验题目要尽可能相似, 它们的得分应具有相近的平均值和标准差。当此条件不能满足时, 需要采用下面两个公式来估计信度。

● 弗朗那根公式:  $r = 2 \left( 1 - \frac{S_a^2 + S_b^2}{S_x^2} \right)$ , 其中:  $S_a^2$ 、 $S_b^2$  分别为两部分测验得分的方差,  $S_x^2$  为测验总分数的方差,  $r$  为信度值。

● 卢伦公式:  $r = 1 - S_d^2 / S_x^2$ , 其中:  $S_d^2$  为两部分测验得分之差的方差,  $S_x^2$  为测验总分数的方差,  $r$  为信度值。

(3) Cuttman 系数。如果一个测验全由二值记分 (如 1 和 0) 的题目所组成,  $\alpha$  信度系数公式中单个题目得分的方差, 就等于该题目上的通过率  $p$  (1 的比例) 与未通过率  $q$  (0 的比例) 的乘积。

Cuttman 系数的计算公式为:  $r_{kk} = \frac{k}{k-1} \left( 1 - \frac{\sum p_i q_i}{S_x^2} \right)$ , 其中:  $k$  为题目数,  $p_i$  为通过第  $i$

题的人数比例,  $q_i$  为未通过第  $i$  题的人数比例,  $S_x^2$  为测验总分的方差。

## 3. 平行测验的信度分析

信度, 可以定义为两个平行测验所得分数之间的相关性。此时, 用一个测验上某些题目

的得分，去推断另一个测验上该题目的得分能力，并用这种能力定义测验的信度。平行测验信度估计的条件有：方差具有齐次性；两平行测验的均值相等。

#### 4. SPSS 中的信度分析

SPSS 的 Reliability Analysis 过程用于信度分析，它应满足如下使用条件。

(1) 观测量之间相互独立；各题目的误差之间互不相关；每对题目的得分都应服从二维正态分布。

(2) 用于分析的数据需为数值型的二元变量、有序变量或连续变量。

##### 14.1.2 问题描述和数据准备

某个 TV 工作室想调查用户将来是否继续收看他们的节目，本节对其问卷进行信度分析，以判断由此得出的结论是否可靠。数据摘自 SPSS 自带的 Demo 文件“tv-survey.sav”，数据文件为“对 TV 工作室的调查.sav”，数据格式如图 14-1 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	any	Numeric	1	0	任何原因	{0, NO}...	None	8	Right	Ordinal
2	bored	Numeric	1	0	没有其他节目	{0, NO}...	None	8	Right	Scale
3	critics	Numeric	1	0	评论较好	{0, NO}...	None	8	Right	Scale
4	peers	Numeric	1	0	其他人在看	{0, NO}...	None	8	Right	Scale
5	writers	Numeric	1	0	保留原编剧	{0, NO}...	None	8	Right	Scale
6	director	Numeric	1	0	保留原导演	{0, NO}...	None	8	Right	Scale
7	cast	Numeric	1	0	保留原演员	{0, NO}...	None	8	Right	Scale


图 14-1 对 TV 工作室的调查数据格式

此问卷有如图 14-1 所示的 7 个问题（理由），取值 1 表示用户基于这个理由会在将来继续观看此 TV 工作室的节目，取值 0 表示此理由不能促使他继续观看此节目。

##### 14.1.3 信度分析的参数设置

依次单击菜单“Analyze→Scale→Reliability Analysis...”，执行信度分析过程，其主设置界面如图 14-2 所示。

###### 1. 指定分析变量

在变量列表选中从（任何原因）到（保留原演员）的 7 变量，单击  按钮，将其作为分析变量选入 Items 列表框；保留 Model 下拉列表的 Alpha 选项。

(1) Items 列表框，用于选入代表问题得分的变量，但不能选入总得分。

(2) Model 下拉列表，用于指定要使用的信度系数，可选项有如下 5 个。

☒ Alpha 选项，表示 Cronbach  $\alpha$  系数，默认选项。

☐ Split-half 选项，表示分半信度。

☐ Guttman 选项，表示 Guttman 系数，输出的 Lambda3 实际就是 Cronbach  $\alpha$  系数。

☐ Parallel 选项，表示平行测验的信度估计。

☐ Strict parallel 选项，在平行测验的基础上，要求各变量的均值相等。

(3) Scale Label 输入框，用于指定刻度标签。

###### 2. 统计量设置

在图 14-2 中，单击 Statistics 按钮，弹出如图 14-3 所示的统计量设置面板。分别勾选 Item、



Correlations 复选框。

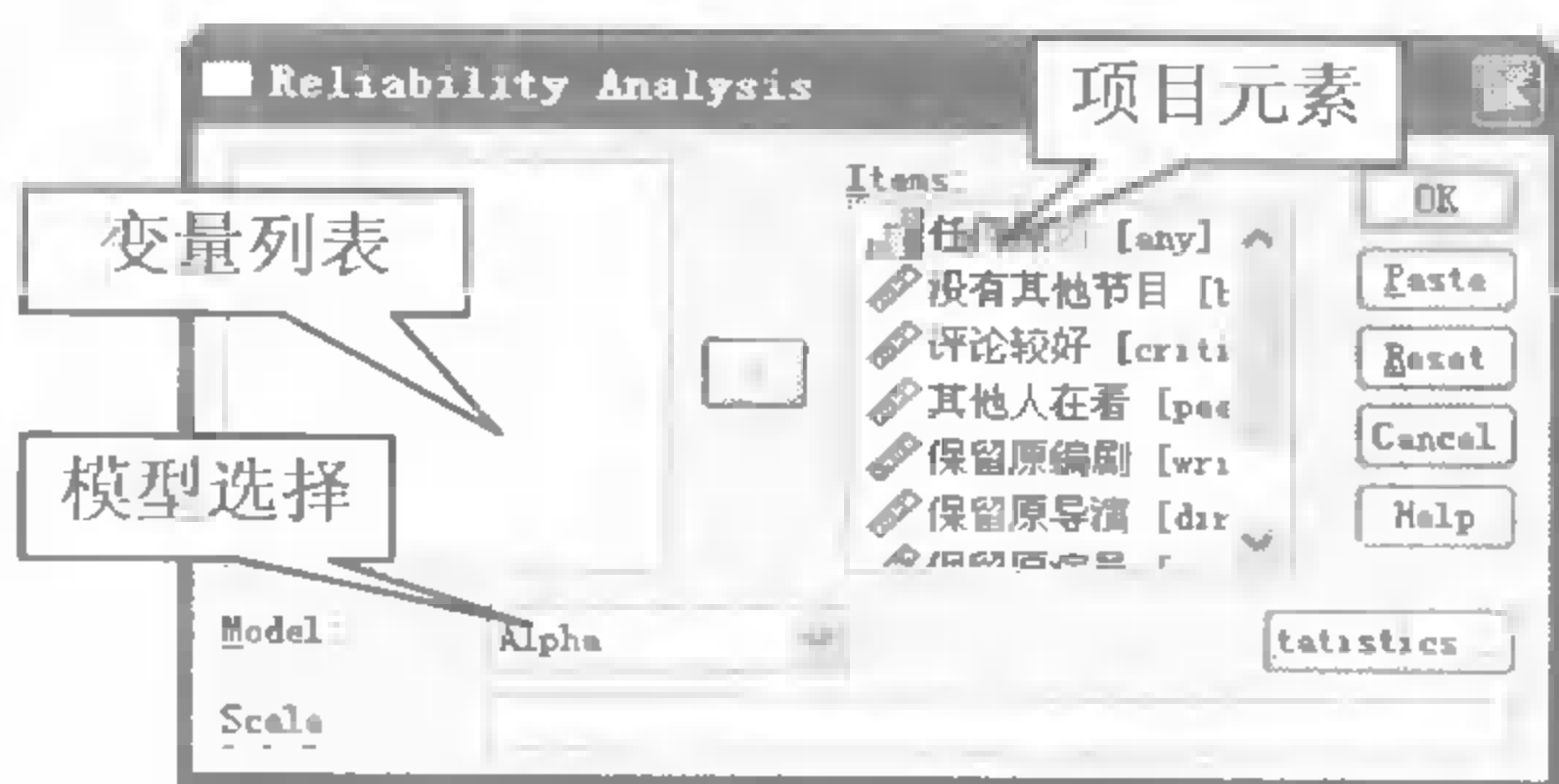


图 14-2 信度分析的主设置面板

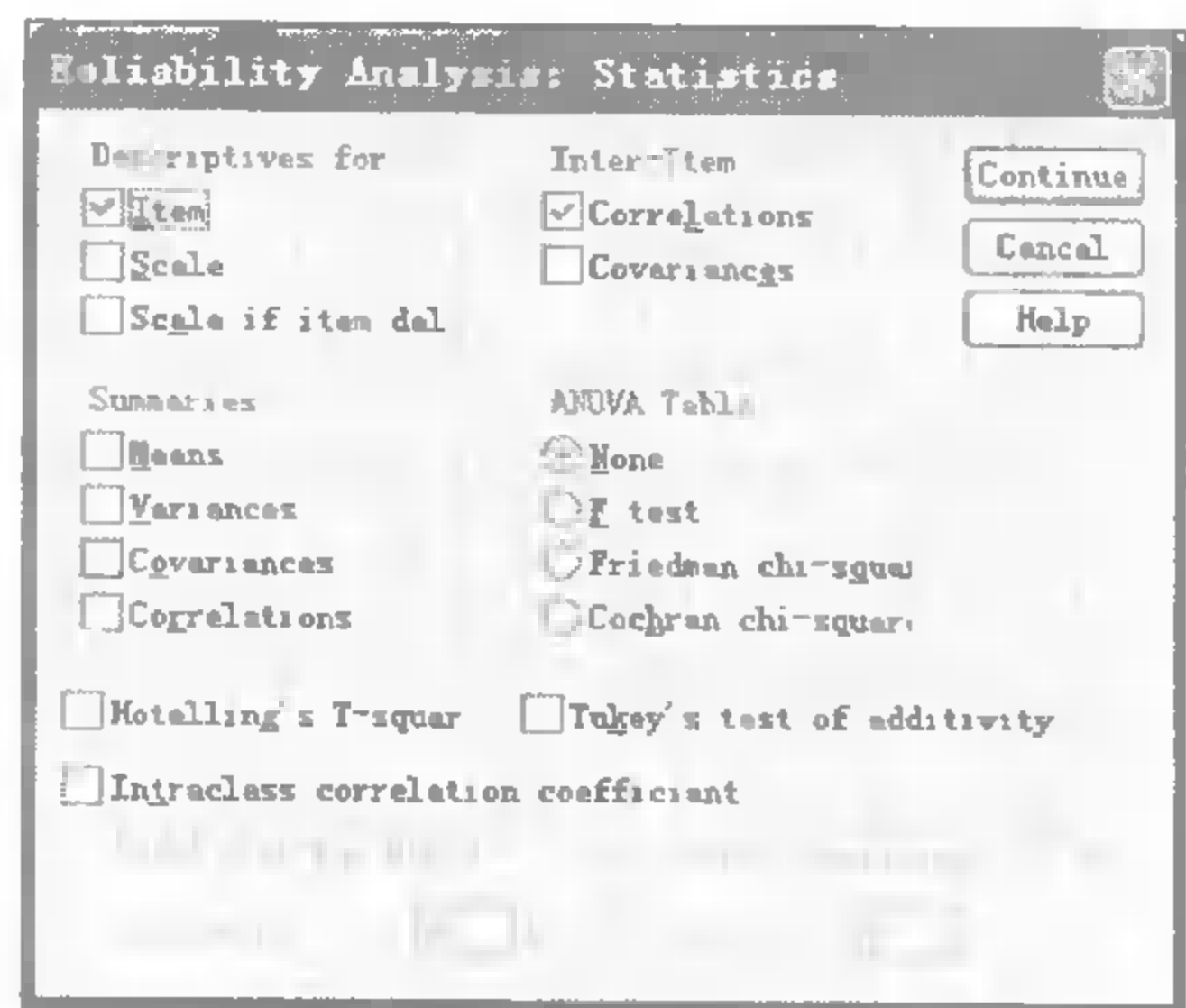


图 14-3 信度分析的统计量设置

(1) Descriptives for 栏，选择输出哪些统计量，可选项有如下 3 个。

- Item 选项，输出各变量的均值、标准差等信息。
- Scale 选项，输出各变量之和（即总分）的均值、方差和标准差等信息。
- Scale if item deleted 选项，输出在问卷中删除指定变量（问题）后，相应统计量的改变值。此项比较重要，可用来对问卷中的各项进行逐一分析，以达到改进问卷的目的。

(2) Inter-Item 栏，设置输出变量间的相关信息。此处有两个可选项：Correlations（相关矩阵）和 Covariances（协方差矩阵）。

(3) Summaries 栏，设置关于各项目的描述统计量的输出，可选内容有如下 4 个。

- Means 选项，输出项目均数的最小、最大、平均值，项目均数的极差和方差，最大项目均数与最小项目均数之比。
- Variances 选项，输出项目方差的最小、最大、平均值，项目方差的极差和方差，最大项目方差与最小项目方差之比。
- Covariances 选项，输出项目协方差的最小、最大、平均值，项目协方差的极差和方差，项目协方差的最大项与最小项之比。
- Correlations 选项，输出项目相关系数的最小、最大、平均值，项目相关系数的极差和方差，项目相关系数的最大项与最小项之比。

(4) ANOVA Table 栏，设置方差分析选项。

它用来分析同一被访者对不同问题的答案是否相关，即不同变量的取值是否相互独立，如果问卷设计得好，这些答案应该是相关的。可选内容有如下 4 个。

- None，不进行分析。
- F test，相当于重复测量的方差分析，该方法适用于数据呈正态分布的情况。
- Friedman chi-square，输出 Friedman 卡方统计量和 Kendall 调谐系数，该方法适用于取秩格式的数据，它可以取代方差分析中的 F 检验。
- Cochran chi-square，对各变量进行 Cochran's 卡方检验，该方法适用于二元变量数据。

(5) Hotelling's T-square 复选框，进行多元检验，零假设为：所有数值变量的均值都相等。

(6) Tukey's test of additivity 复选框，检验各变量之间是否具有显著的交互作用。

(7) Intraclass correlation coefficient 复选框，设置关于组内相关系数的选项。

- Model 下拉菜单，指定计算组内相关系数的模型。

Two-Way Mixed（两方向固定模型），当人为效应及项目效应均为固定时选择此项；Two-Way Random（两方向随机模型），当人为效应及项目效应均为随机时选择此项；One-Way Random（单方向随机模型），当人为效应为随机时选择此项。

- Type 下拉菜单，指定指标（index）的类型，可选项有如下两个：Consistency（一致性）和 Absolute Agreement（绝对一致）性。
- Confidence interval 输入框，指定置信区间，默认为 95%。
- Test value 输入框，指定一个用于和观测相关系数进行比较的待检验数值，输入数值要求在 0~1 之间，默认为 0。

#### 14.1.4 案例的结果分析

在图 14-2 中单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 14-4 和图 14-5 所示。

个案处理摘要		
案例	N	%
有效	906	100.0
排除的 <sup>a</sup>	0	0
总计	906	100.0
<sup>a</sup> 在此程序中基于所有变量的列表方式删除。		
可靠性统计量		
Cronbach's Alpha	基于标准化项的 Cronbach's Alpha	项数
.898	.894	7

项统计量			
	均值	标准差	N
任何原因	49	500	906
没有其他节目	50	500	906
评论较好	50	500	906
其他人在看	53	499	906
保留原导演	83	378	906
保留原演员	89	315	906
保留原编剧	61	389	906

图 14-4 信度分析结果输出 1

项间相关性矩阵							
	任何原因	没有其他节目	评论较好	其他人在看	保留原导演	保留原演员	保留原编剧
任何原因	1.000	.815	.813	.782	.421	.303	.406
没有其他节目	.815	1.000	.826	.807	.423	.307	.422
评论较好	.813	.826	1.000	.804	.453	.336	.458
其他人在看	.782	.807	.804	1.000	.460	.340	.443
保留原导演	.421	.423	.453	.460	1.000	.600	.632
保留原演员	.303	.307	.336	.340	.600	1.000	.625
保留原编剧	.406	.422	.458	.443	.632	.625	1.000

图 14-5 信度分析结果输出 2

（1）信度系数输出。如图 14-4 所示，“摘要”表格给出了初始数据中关于缺失值的统计信息。

“可靠性统计量”表格给出了 Cronbach  $\alpha$  系数的计算结果，其中标准化的  $\alpha$  系数只有在图 14-3 中的 Inter-Item 栏选中某项后才会输出。表中的 0.898 是对真实  $\alpha$  系数的估计（下界），由于它大于 0.8，故可以推断此问卷的可信度还是不错的。

（2）变量统计信息。如图 14-4 所示，“项统计量”表格给出了各变量（题目得分）的基本统计信息。

可见，约 50%的人会在大多数情况下继续观看节目，还有 30%~40%的人在节目变换不大（不改变编剧、导演或演员）的情况下，会继续观看。

（3）相关矩阵。如图 14-5 所示，给出的是各问题得分之间的相关矩阵。可见，前 4 项之间的相关性较高，说明如果用户基于这 4 个原因中的一个继续观看该节目的话，在其他 3 种原因的诱导下，他们同样也会选择继续观看。

## 14.2 多维尺度分析

多维尺度分析 (Multi-dimension Analysis, MDS) 是市场研究的一种有力手段, 它可在低维空间 (通常是二维空间) 展示多个研究对象 (比如品牌) 之间的联系, 利用平面距离来反映研究对象之间的相似程度, 还能够用来识别影响事物相似性的潜在因素。

### 14.2.1 多维尺度分析简介

多维尺度分析, 用于研究多个事物之间的相似 (不相似) 程度, 通过适当的降维方法, 将这种相似 (不相似) 程度在低维度空间中用点与点之间的距离表示出来, 它是市场调查、数据分析的常用统计方法之一。例如, 有 10 个商场, 让消费者排列出它们两两之间的相似程度, 对这些数据应用多维尺度分析, 能够判断哪些商场是消费者真正认为相似的。

由此可见, 多维尺度分析是基于研究对象之间的相似性 (距离) 的, 只要获得了研究对象之间的距离矩阵, 就可以做出它们的相似性感知图。

在实际应用中, 获取距离矩阵的主要有两种方法: 一种是直接评价法, 先把所有评价对象进行两两组合, 然后要求被访者对所有的这些组合直接进行相似性评价; 另一种是间接评价法, 由研究人员根据经验, 事先找出影响研究对象相似性的主要属性, 然后让被访者对这些属性进行逐一评价, 再将这些属性得分当作多维空间的坐标, 计算对象之间的距离。

SPSS 的多维尺度分析过程 (ALSCAL), 对数据的分布没有特定要求, 但是需要正确指定分析变量的度量方式 (Ordinal, Interval, 或者 Ratio)。

### 14.2.2 问题描述和数据准备

一些受访者对几种饮料的相似程度进行了排序, 本节对这些数据进行多维尺度分析, 以判断哪些饮料在受访者看来确实是相似的。数据来源于《SAS 系统与市场调查分析》中的例题, 数据文件为“饮料相似度排序数据.sav”, 数据格式如图 14-6 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	sub	String	8		受试者	{sub1, 受	None	6	Right	Nominal
2	sort	String	8		种类	None	None	6	Right	Nominal
3	milk	Numeric	8	0	牛奶	None	None	4	Right	Scale
4	coffee	Numeric	8	0	咖啡	None	None	6	Right	Scale
5	tea	Numeric	8	0	茶	None	None	5	Right	Scale
6	soda	Numeric	8	0	苏打水	None	None	5	Right	Scale
7	juice	Numeric	8	0	果汁	None	None	5	Right	Scale
8	botwater	Numeric	8	0	矿泉水	None	None	7	Right	Scale
9	beer	Numeric	8	0	啤酒	None	None	5	Right	Scale
0	wine	Numeric	8	0	葡萄酒	None	None	6	Right	Scale

图 14-6 对 TV 工作室的调查数据格式

本例要求每个受试者 (sub) 对 7 种饮料作两两比较, 根据它们之间的相似程度打分, 采用 7 分制, 分值越小相似程度越大, 所以本例数据是不相似数据。

### 14.2.3 ALSCAL 过程的参数设置

依次单击菜单 “Analyze→Scale→Multidimensional Scaling(ALSCAL)...”, 执行 ALSCAL 多维尺度分析过程, 其主设置界面如图 14-7 所示。

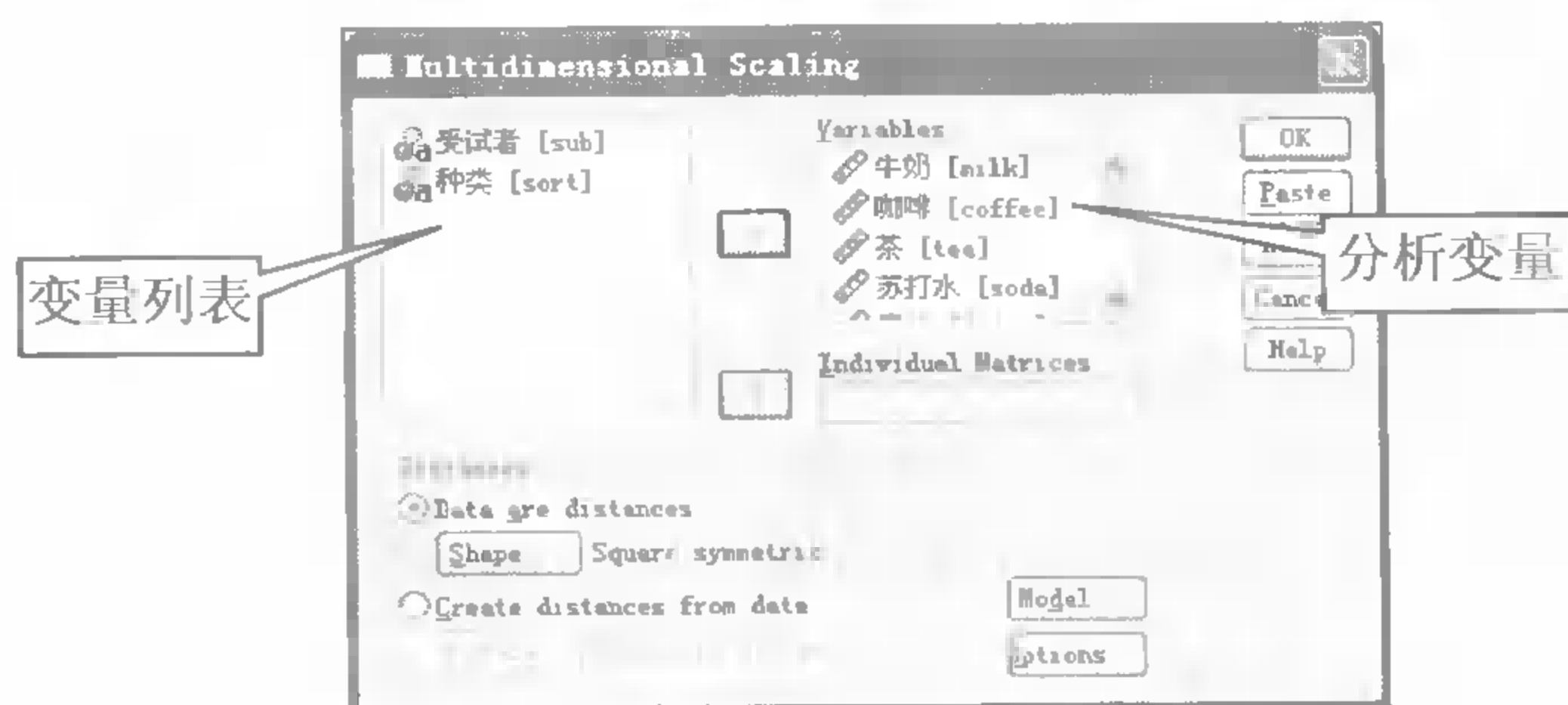


图 14-7 多维尺度分析 (ALSCAL) 设置对话框

## 1. 变量设置

在变量列表选中从牛奶到葡萄酒的 8 变量，单击从上至下第一个 ☐ 按钮，将其作为分析变量选入 Variables 列表框。

(1) Variables 列表框，用于选入表示距离的分析变量。

(2) Individual Matrices for 选框，用于选入分组变量，分析时将会为每一组变量分别计算距离矩阵。选中 Create distances from data 选项时才可用。

(3) Distance 栏，指定距离的计算方法，有两个选择。

① Data are distances 单选框，表示当前数据就是距离（不相似）矩阵，可以直接用于分析。

它的右侧显示当前指定的矩阵形状，单击 Shape 按钮，弹出如图 14-8 所示的子对话框，设置距离矩阵的形状。此处的设置对分析结果有较大的影响，各选项的解释如下。

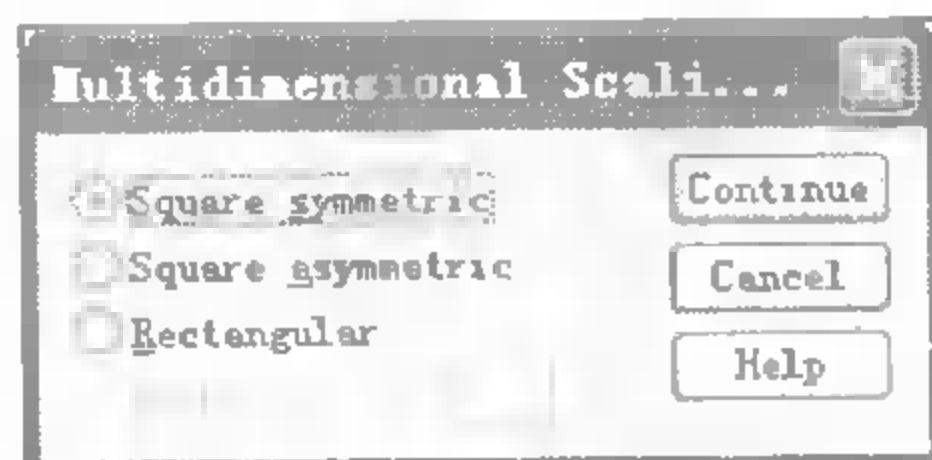


图 14-8 Shape 设置对话框

● Square symmetric 距离矩阵为完全对称方阵，行和列表示相同的项目，如果只录入了半个矩阵，系统会根据对称性自动填充另一半，本例即为这种情况。

● Square asymmetric 距离矩阵为不对称方阵，行和列表示相同的项目。

● Rectangular 距离矩阵为完全不对称的矩形形式，行和列表示不同的项目。SPSS 把有序排列的数据当作矩形矩阵，如果其中含有多个矩形矩阵，需要设置每个矩阵的行数。Number of rows 输入框，用于指定单个矩阵行数，输入值应大于等于 4，并且能够整除数据的所有行数。

② Create distances from data 单选框，表示用户需要自行选择相似矩阵的计算方法。

当数据比较复杂、不能直接用作距离矩阵时选择此项，表示从当前数据出发计算距离矩阵，右侧显示当前指定的距离测量方式。单击 Measure 按钮，弹出如图 14-9 所示的子对话框。

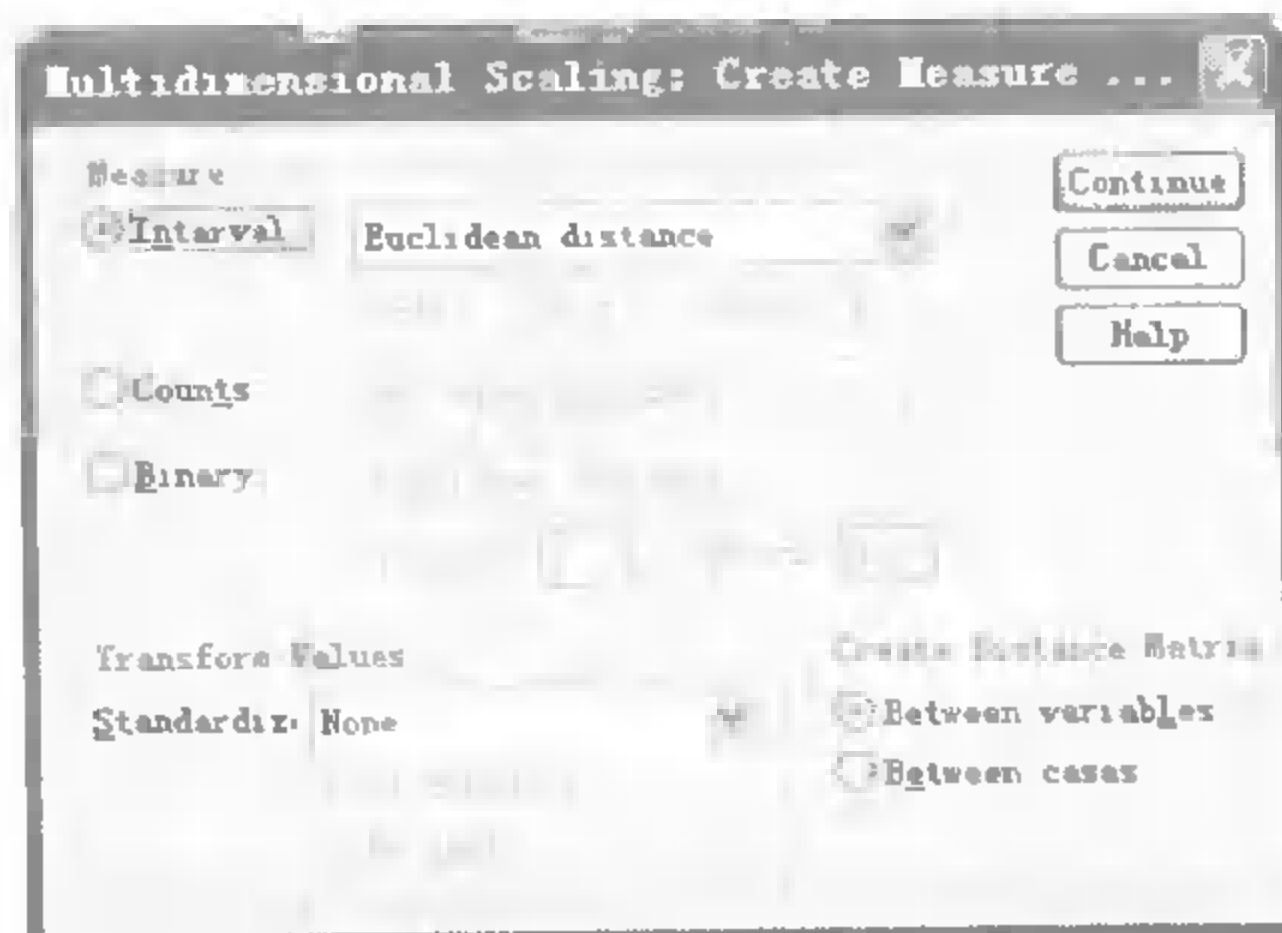


图 14-9 Measure 设置对话框



- Measure 栏, 指定不相似度的测量方法; Transform Value 栏, 指定标准化转换的方法。这两部分的参数设置, 请参考第 10.4.2 节中距离分析的 Measure 设置 (参见图 14-10)。

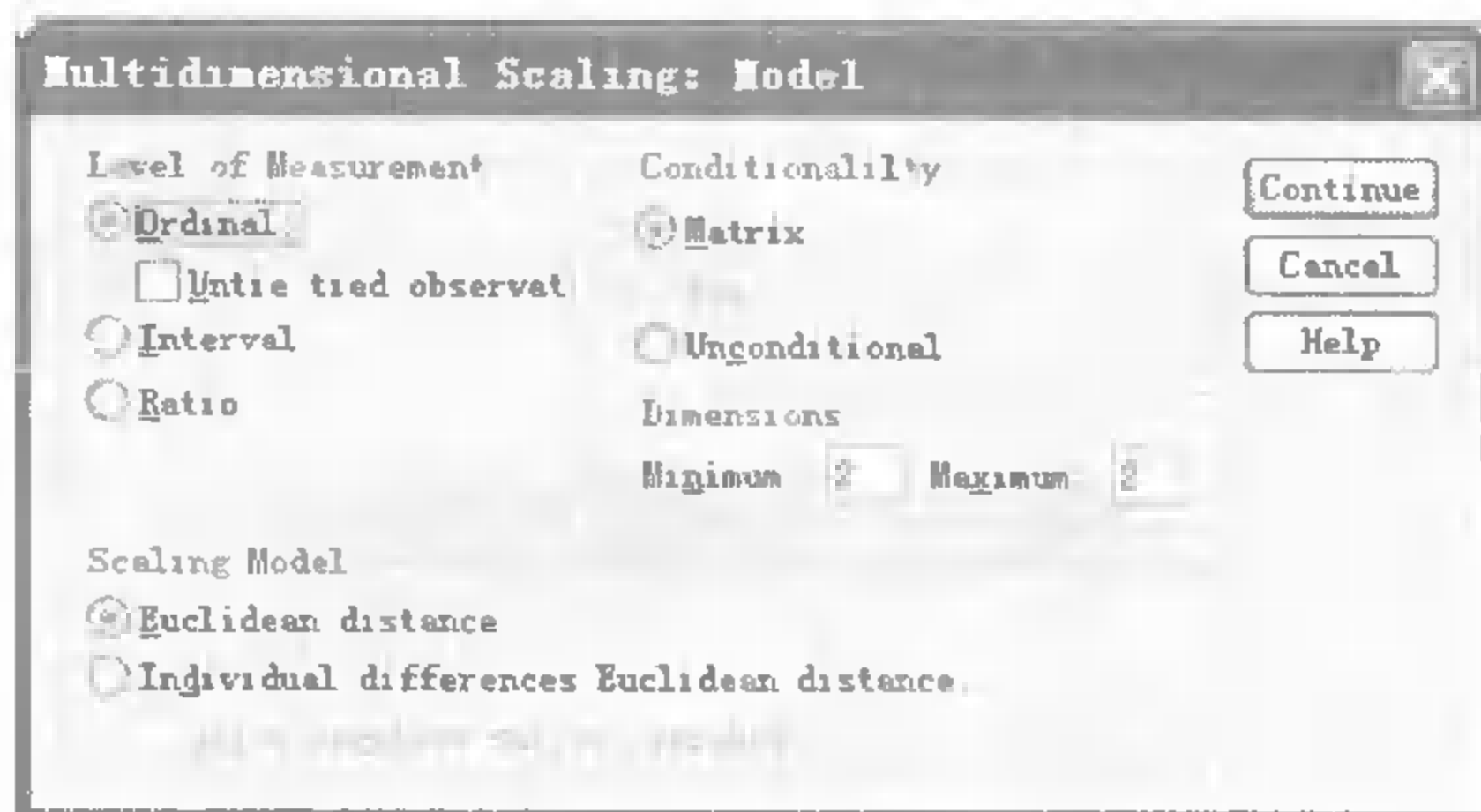


图 14-10 模型参数设置

- Create distance matrix 栏, 指定创建距离矩阵的方式, 有两个选项: Between variables 单选框, 表示计算配对变量之间的不相似性距离矩阵; Between cases 单选框, 表示计算配对观测量之间的不相似性距离矩阵。

## 2. 模型设置

在图 14-7 中, 单击 Model 按钮, 弹出如图 14-10 所示的模型设置子面板。采用默认设置, 单击 Continue 按钮返回主界面。

(1) Level of Measurement 栏, 用于指定数据的测量尺度, 设置内容有如下 3 项。

- Ordinal 选项, 表示有序测量尺度, 即分析数据是有序分类资料。

因为调查数据多是由受访者对相似性所做的主观判断 (打分), 所以大多数数据都为此类型。

Untie tied observations 复选框, 设置对节 (Tie, 评分相同) 的处理方式。默认情况下, 对取值相同的评分赋予相同的秩, 勾选中该复选框, 表示对相同的评分赋予不同的秩。

- Interval 区间尺度, 分析数据是由连续性变量或定量变量组成的资料。

- Ratio 比例尺度, 分析数据是由比例形式的定量变量组成的资料。

(2) Conditionality 栏, 指定哪些比较是有意义的, 可选项有如下 3 个。

- Matrix, 适用于只有一个距离矩阵, 或者每个距离矩阵仅代表单个受访者的情况。它表示单个距离矩阵内部的各个数值意义相同, 是可以相互比较的。
- Row, 适用于距离矩阵为非对称的或矩形的。它表示仅同行数据之间的比较才有实际意义, 同列数据之间无需进行比较。
- Unconditional, 不受任何限制, 任意两个数据之间的比较都是有意义的。

(3) Dimensions 栏, 用于指定尺度分析 (scaling solutions) 的维度, 默认为 2 维。

Minimum、Maximum 两个输入框分别指定维度的最小值和最大值, 它们的输入值都需为 1~6 的整数, 对指定范围内的每个维度分别进行分析。

(4) Scaling Model 栏, 设置尺度模型的距离选项, 可选项有如下 2 个。

- Euclidean Distance, 欧氏距离, 它可用于任何类型的矩阵分析中。如果数据为单一矩阵, 将进行典型多维尺度分析 (CMDs); 否则, 进行重复多维尺度分析 (RMDS)。

- Individual differences Euclidean Distance, 表示使用个体差异的欧氏距离矩阵进行分析。它要求数据包含两个以上的距离矩阵。

Allow negative subject weights 复选框, 如果允许权重变量取负值, 选中该项。

### 3. 输出设置

在图 14-7 中, 单击 Options 按钮, 弹出如图 14-11 所示的输出设置子面板。勾选 Group plots 复选框; 单击 Continue 按钮返回主界面。

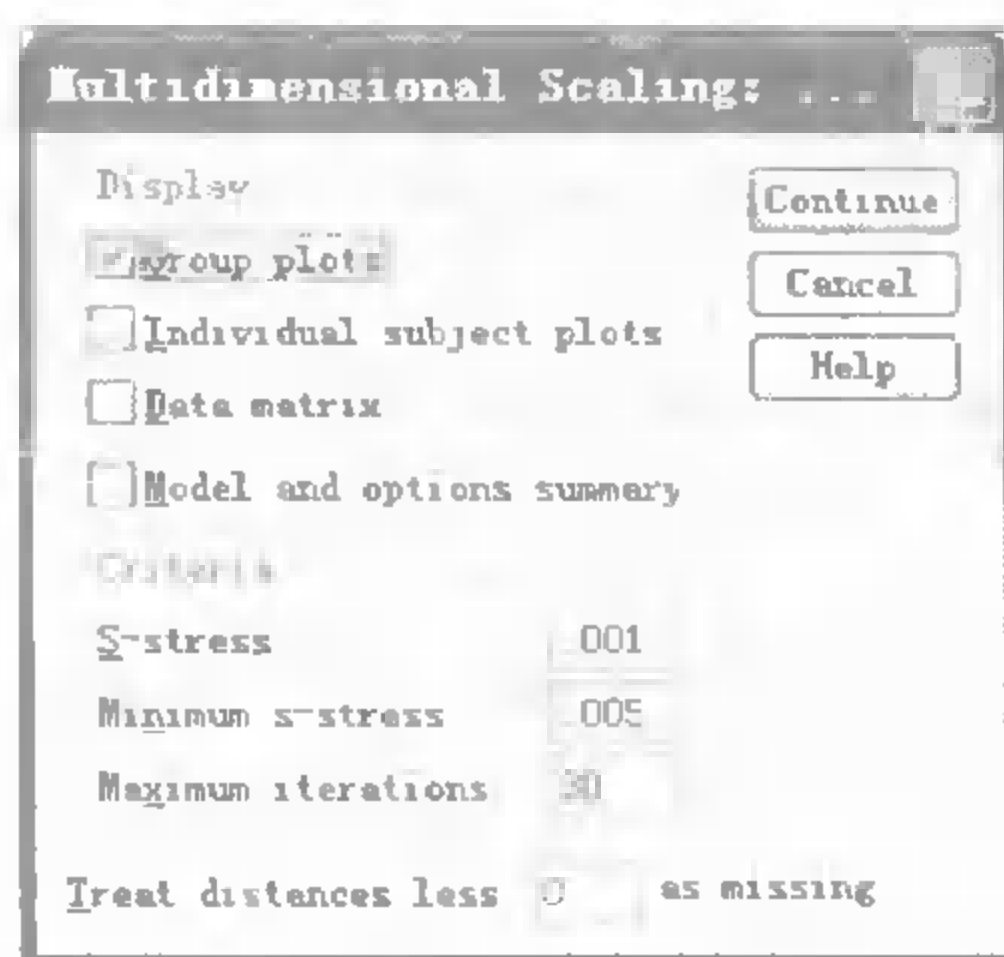


图 14-11 Options 选项参数设置

(1) Display 栏, 选择输出哪些图形和分析结果, 可选项有如下 4 个。

- Group plots 复选框, 多维尺度分析图, 这是最重要的输出结果, 用它可以直接地分析散点间 (项目之间, 或者个体之间) 的相关性的合理解释。
- Individual subject plots 复选框, 为每位受试者分别输出单独的分析图形。
- Data matrix 复选框, 输出每位受试者的数据矩阵。
- Model and options summary 复选框, 输出分析中所用的数据、模型、算法等参数的详情。

(2) Criteria 栏, 设置迭代收敛的依据, 设置内容有如下 3 项。

- S-stress convergence 输入框, 指定 S-stress (S 应力) 的最小改变量, 默认 0.001。当两次相邻迭代的 S-stress 增量小于等于此值时停止迭代。
- Minimum s-stress value 输入框, 指定 S-stress (S 应力) 的最小值, 默认 0.005, 输入值应大于 0 小于 1。当迭代计算出的 S-stress 小于等于此值时停止迭代。
- Maximum iterations 输入框, 指定最大迭代次数, 默认值为 30。

(3) Treat distances less than N as missing 子设置栏, 表示把距离小于 N 值的数据当作缺失值, N 为在 as 前的输入框指定的数值。

### 14.2.4 案例的结果分析

在图 14-7 中单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 14-12~图 14-14 所示。

Iteration history for the 2 dimensional solution (in squared distances)		
Young's S-stress formula 1 is used.		
Iteration	S-stress	Improvement
1	.45654	
2	.41326	.04327
3	.40999	.00328
4	.40936	.00062
Iterations stopped because		
S-stress improvement is less than .001000		

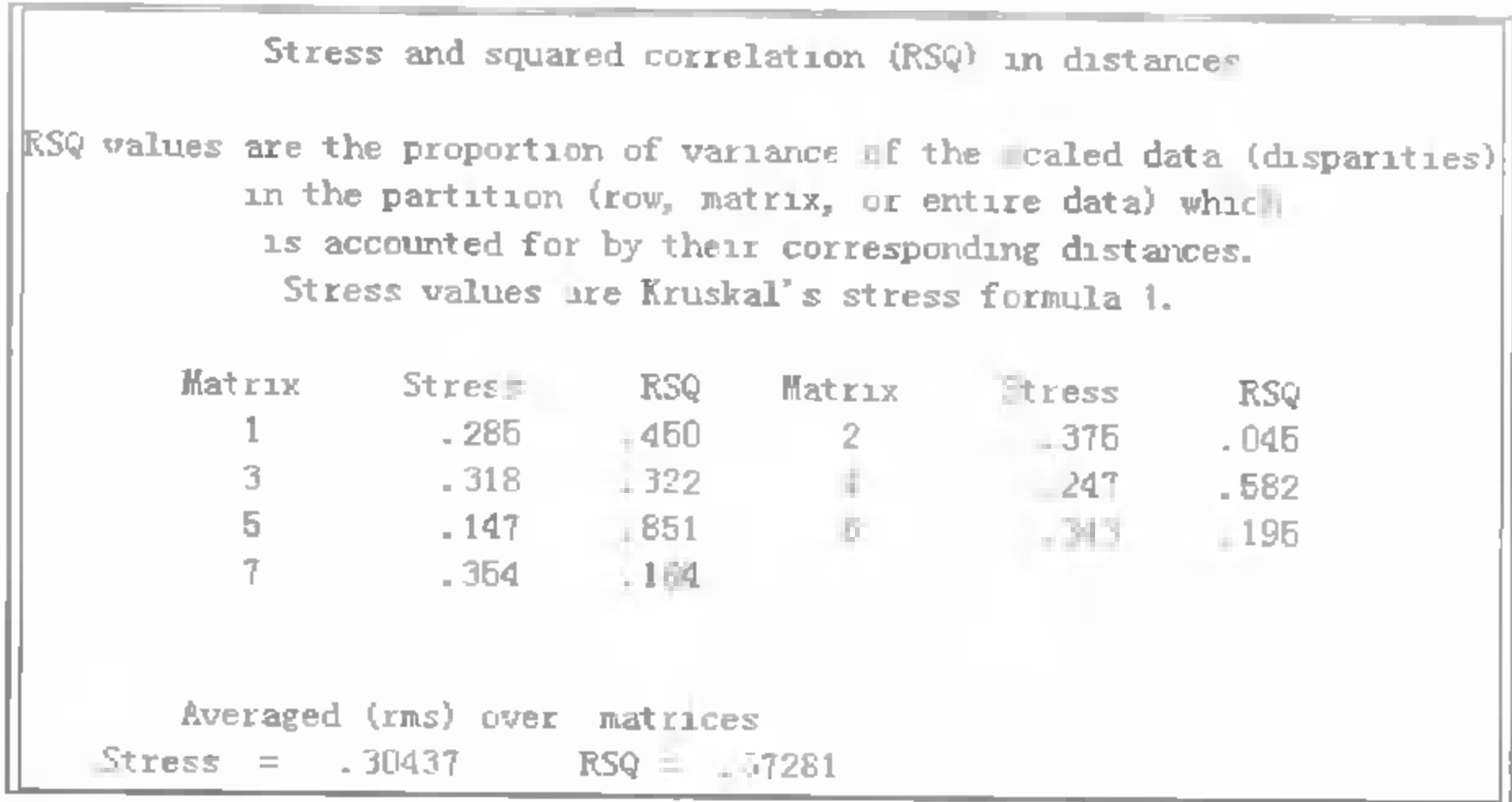


图 14-12 迭代记录和相关性输出

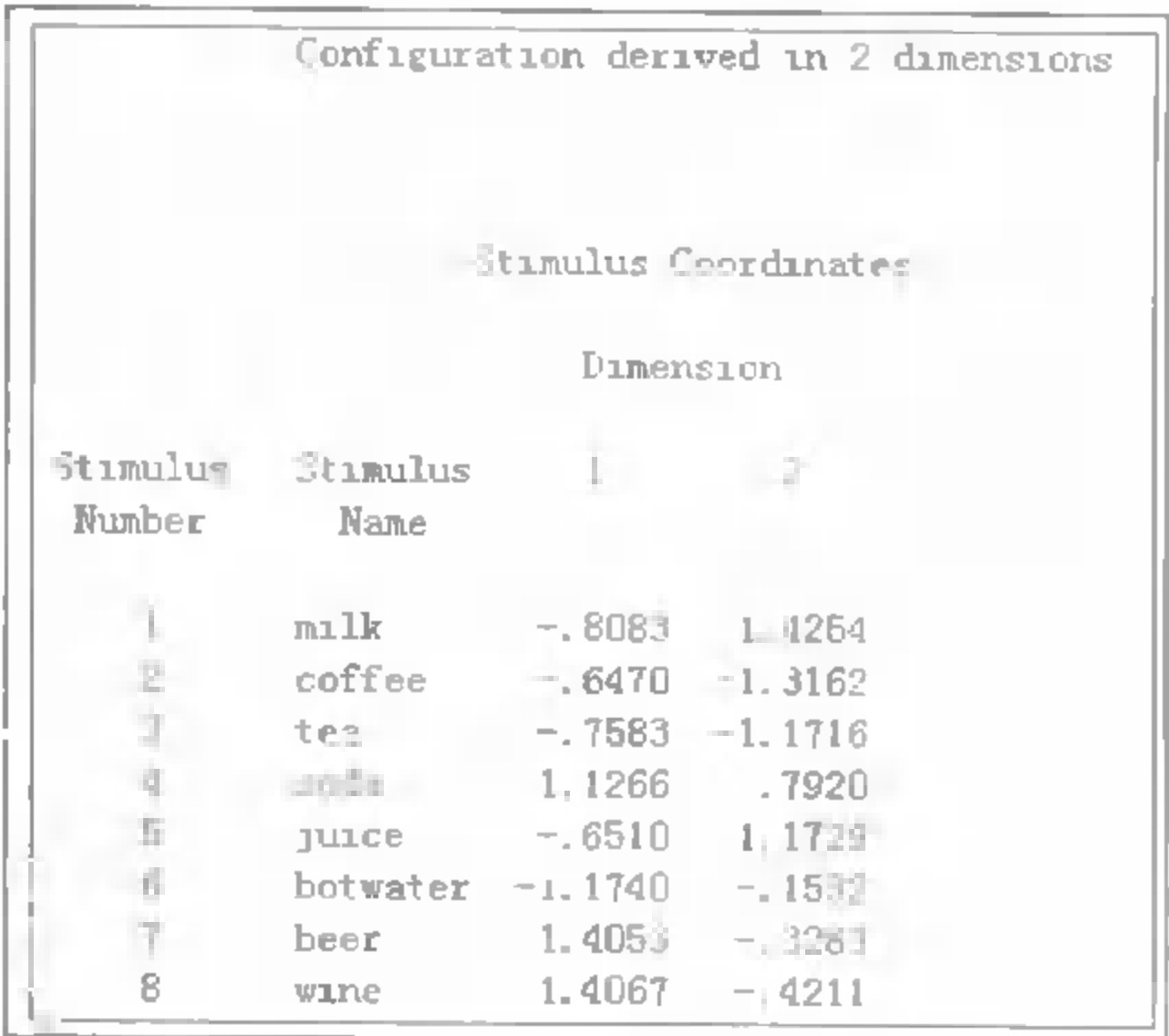


图 14-13 二维导出构形表

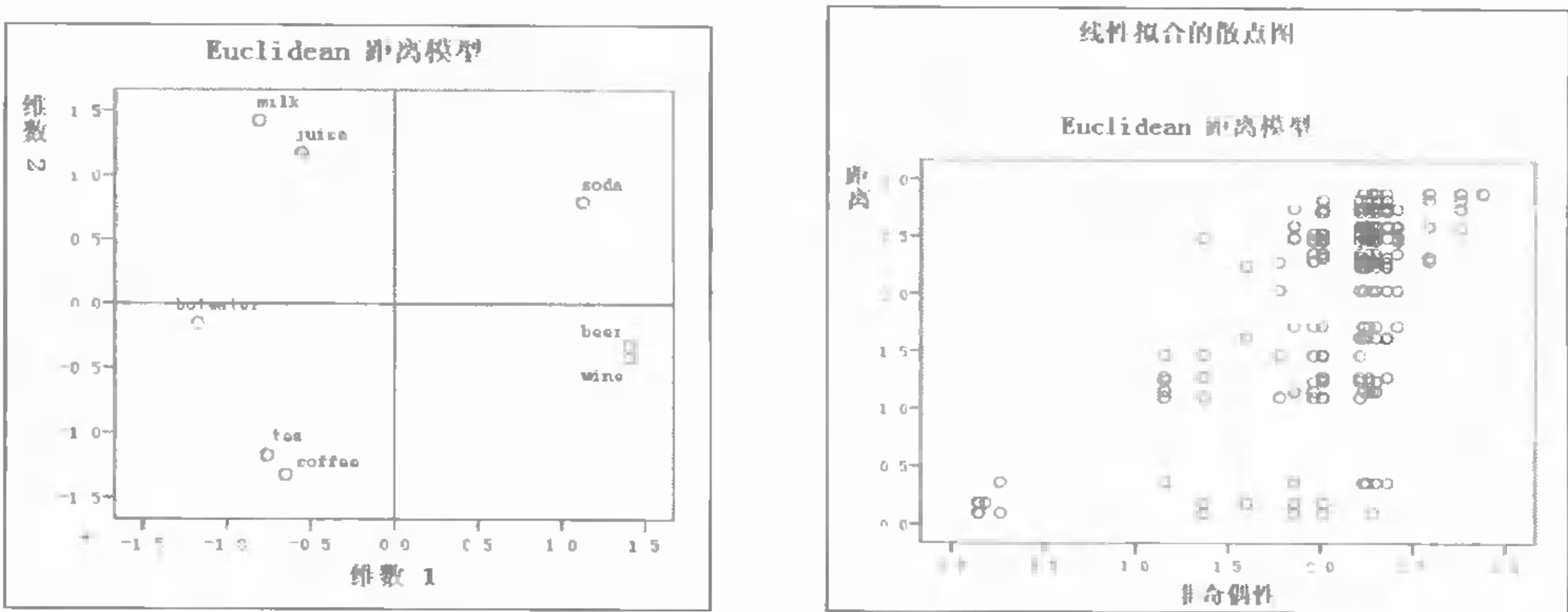


图 14-14 多维尺度分析图

(1) 迭代记录和信度估计。如图 14-12 所示，“Iteration history”表格给出了二维空间中的迭代记录，可见在 4 次迭代之后，S-stress（应力）值的变化小于 0.001，达到收敛标准。

“Stress and squared correlation (RSQ) in distances”表格给出了 Stress 与 RSQ 值的输出，它们是多维尺度分析的信度和效度的估计值。Stress（应力）是拟合量度值，越大说明拟合优度越好；RSQ（平方相关系数）的取值也是越大越理想，一般在 0.60 是可以接受的。本例的 Stress 平均值为 0.304 37，RSQ 平均值为 0.372 81，都说明拟合度不太好。

改进方法有两个：采用 SPSS 的 PROXSCAL 过程进行分析，或者增加受试者。

(2) 二维导出构形表。如图 14-13 所示，是二维导出构形表，“1”、“2”列表示八种饮料在二维空间中的坐标值，可用于作多维尺度分析图。

(3) 多维尺度分析图。如图 14-14 所示，左侧的“Euclidean 距离模型”图形就是常用的多维尺度分析图，它把反应变量之间相似程度的坐标在平面上排列出来，通过观察哪些散点比较接近，将变量进行分类，并寻找散点之间相关性的合理解释。本例有三组聚集点：咖啡（coffee）和茶（tea）是相似的；果汁（juice）和牛奶（milk）是相似的；啤酒（beer）和葡萄酒（wine）是相似的。另外，只从第 2 维度看，还可将这些饮料分为两类：milk、juice、soda 属于营养型饮料；beer、wine、coffee 和 tea 属于提神型饮料。

“线性拟和的散点图”是欧氏距离模型的线形拟合散点图，它是欧氏距离（Distances）对原始数据不一致程度（Disparities）的散点图，如果模型的拟合程度好，所有散点应分布在一条直线的周围。本例各点的分布比较分散，不呈现明显的线性趋势，再次说明模型的拟合效果不好。



# 第15章 时间序列分析

直观地讲，时间序列指随时间变化的、具有随机性的、且前后相互关联的动态数据序列，它是依特定时间间隔而记录的指定变量的一系列取值。大量的经济统计指标都是按照年、季、月或日统计的，随着时间的推移，就形成了这些统计指标的时间序列。

时间序列分析，就是研究时间序列在演变过程中存在的统计规律的方法，研究问题包括：长期变动趋势、季节性变动规律、周期变动规律，以及预测未来时刻的发展和变化等。

时间序列数据的取值方式，有如下两种。

(1) 取某些观测时刻的瞬时值，例如：某城市每日中午的气温值，仓库在月末的存储量，每年7月1日的人口数，每年开学的学生在册人数等。

(2) 取某些观测时刻之间的累积值，例如：每年的工农业总产值，某商场的月销售额，每年的钢总产量，每年的粮食总产量，每年的商品贸易总额等。

上述时间序列取值有一个共同的特点，即都是离散型的时间序列，当然也有连续型时间序列，例如心电图、工业供电仪表的记录等。本章只讨论对离散型时间序列的分析。

时间序列分析又可分为时域分析与频域分析两大部分，它们的研究手段包括：建立时间序列模型、参数估计、最佳预测和控制、谱估计等，特别是对于自回归模型、滑动平均模型、自回归及滑动平均模型有着一套比较完整的统计理论。近年来，时间序列分析已经渗透到交通运输、智能控制、神经网络模拟、生物、医学、水文、气象、经济学、空间科学等众多领域，发挥着无可比拟的作用。

## 15.1 SPSS15 的时间序列分析概览

依次单击菜单“Analyze→Time Series”，弹出如图 15-1 所示的子菜单选项，时间序列分析的各种功能就通过这些子菜单来执行。

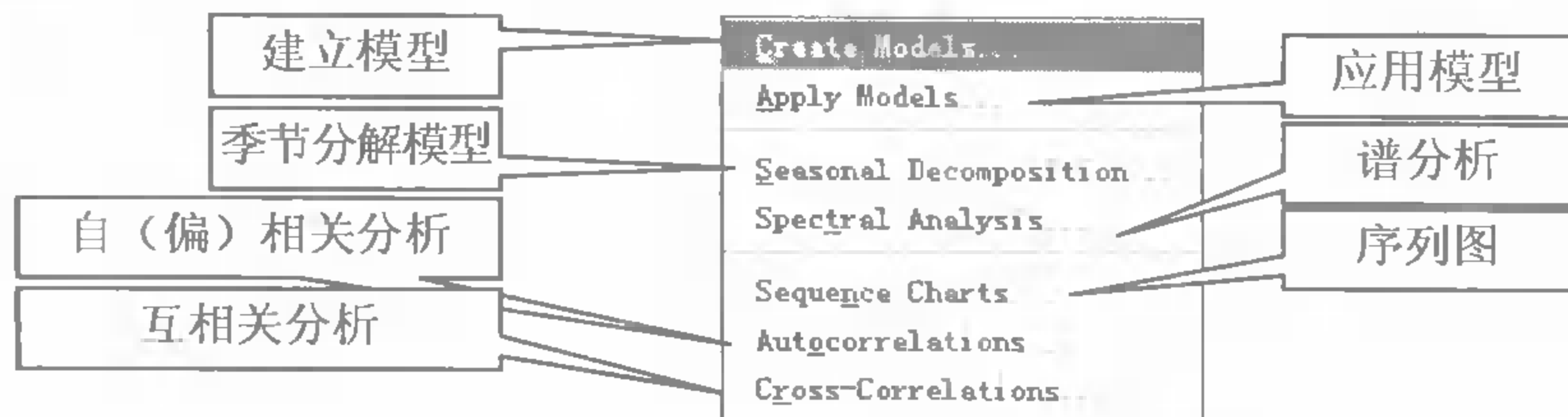


图 15-1 Time Series 的子菜单选项

本节介绍建立模型（Create Models）和应用模型（Apply Models）两个模块，其中 Create

Models 又分为指数平滑模型、自回归滑动平均模型两种方法，对它们的应用详解分别安排在第 15.3 节和第 15.4 节。第 15.5 节介绍季节分解（Seasonal Decomposition）模型的使用。

### 15.1.1 Create Models 的通用设置选项

#### 1. 变量设置和模型选择

依次单击菜单“Analyze→Time Series→Create Models...”，执行建立模型的功能，其主设置面板如图 15-2 所示，在此指定分析变量、选择模型方法。

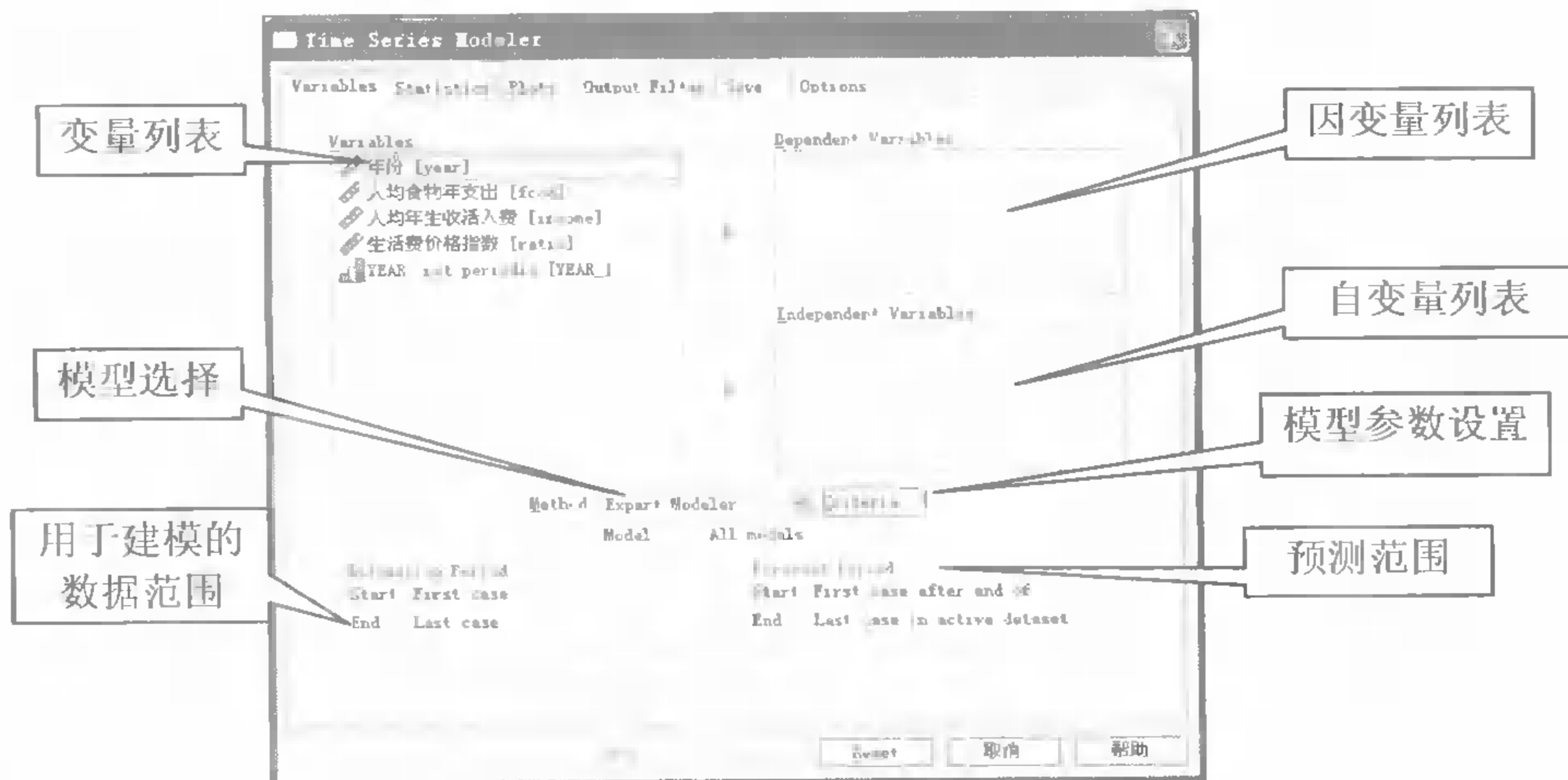


图 15-2 Create Models 的变量设置面板

- (1) Variables 列表框，显示当前可用的变量。
- (2) Dependent Variables 列表框，用于从变量列表选入因变量。
- (3) Independent Variables 列表框，用于从变量列表选入自变量。

注意：自变量的选择只对 ARIMA 模型有效；如果指定了 Expert Modeler 模型，而且同时选入了自变量，则只会按照 ARIMA 模型进行建模。

- (4) Method 下拉列表，用于指定建模方法，有 3 个可选项。
  - Expert Modeler 选项，表示对每个因变量分别自动寻找最优的拟合模型。
  - Exponential Smoothing 选项，用于自定义特定的指数平滑模型。
  - ARIMA 选项，用于自定义特定的 ARIMA 模型。

单击 Criteria 按钮，设置指定模型的有关参数，具体设置内容在随后章节陆续介绍。

- (5) 显示数据范围。

- Estimation Period 估计范围栏，显示模型估计时所用到的数据范围，默认为当前数据集的所有记录。如果需要指定此数据范围，依次单击菜单“Data→Select Cases...”执行数据选择功能，选中后的数据将用于估计模型。
- Forecast Period 预测范围栏，显示要预测的数据范围，默认为当前数据集的所有记录范围。如果需要指定此数据范围，在图 15-2 中的 Options 子标签面板进行设置。

#### 2. 计量选项设置

在图 15-2 中单击 Statistics 标签，显示如图 15-3 所示的子设置界面，在此设置关于统计

量的选项。

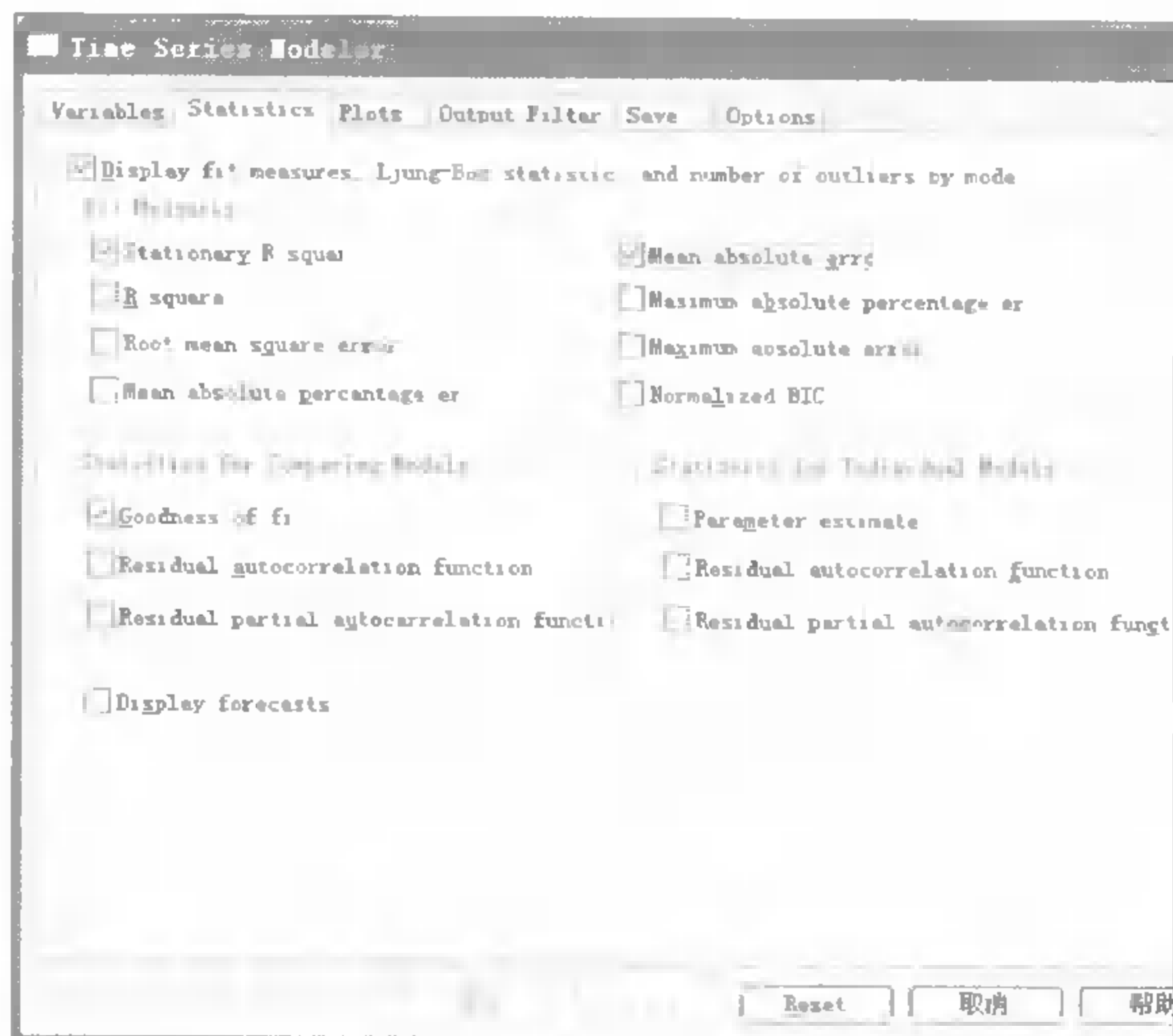


图 15-3 Create Models 的统计量设置

① Display fit...复选框表示输出如下内容：模型拟和方法、Ljung-Box 统计量、由模型定义的异常点个数等。选中后，激活 Fit Measures 子设置栏的选项。

② Display forecasts 复选框，表示为每个估计模型输出预测值及其置信范围。

③ Fit Measures 栏，设置输出哪些反映模型拟和优度的统计量，可选项有如下 8 个。

● Stationary R-squared 平稳 R 方统计量，用来比较模型中的固定成分与一个简单均值模型的差别，当原始序列中有趋势成分或季节成分时，要优于 R 方统计量。它取负值时，表示当前模型没有基本均值模型好；它取正值（小于 1），表示当前模型要优于基本均值模型。

● R-squared 复选框，R 方统计量，用来估计由模型解释的变异在总变异中的比例，当原始序列为平稳序列时，优于平稳 R 方统计量。它取负值时，表示当前模型没有基本均值模型好；它取正值（小于 1），表示当前模型要优于基本均值模型。

● RMSE (Root Mean Square Error) 均方误差，用来度量原始因变量序列与它的模型预测值的差异，度量单位与原序列一致。

● MAPE (Mean Absolute Percentage Error) 绝对比例误差均值，用以度量原始因变量序列与它的模型预测值的差异。此统计量与序列的度量单位无关，因而可以用来比较度量单位不同的序列之间的拟合优度。

● MAE (Mean absolute error) 绝对误差均值，用以度量原始因变量序列与它的模型预测值的差异，度量单位与原序列一致。

● MaxAPE (Maximum Absolute Percentage Error) 最大绝对比例误差，以比例形式表示的最大预测误差。当关注于预测单个记录的最差情况时，使用此统计量。

● MaxAE (Maximum Absolute Error) 最大绝对误差，用来度量最大的预测误差。当关注于预测单个记录的最差情况时，使用此统计量。MaxAPE、MaxAE 可能发生在不同的观测记录上，故而有必要加以区分。

● Normalized BIC (Normalized Bayesian Information Criterion) 正态 BIC 统计量，用来度量模型的拟合优度，同时考虑了模型的复杂程度。它基于均方误差，并包含一个对参数个数和序列长度的惩罚项。

④ Statistics for Comparing Models 栏，设置关于模型比较的统计量输出。

此设置栏的每个选项，都将单独输出一张表格，可选内容有如下 3 个。

- Goodness of fit 拟合优度，把每个模型的拟和优度统计量（见 Fit Measures 栏的内容）汇总在一张表格里输出，便于比较。
- Residual autocorrelation function (ACF) 残差的自相关函数，输出每个模型的残差自相关函数的统计特征和百分位点。
- Residual partial autocorrelation function (PACF) 残差的偏相关函数，输出每个模型的残差偏相关函数的统计特征和百分位点。

⑤ Statistics for Individual Models 栏，设置关于单个模型的输出信息。

此设置栏的每个选项，都将单独输出一张表格，可选内容有如下 3 个。

- Parameter estimates 参数估计值，对指数平滑模型和 ARIMA 模型，分别输出它们各自的参数估计表。对于异常值的估计，将单独输出一张表格。
- Residual autocorrelation function (ACF) 残差自相关函数，输出每个模型残差的自相关序列及其置信区间。
- Residual partial autocorrelation function (PACF) 残差偏相关函数，输出每个模型残差的偏相关序列及其置信区间。

### 3. 作图选项设置

在图 15-2 中单击 Plots 标签，显示如图 15-4 所示的子设置界面，在此设置关于作图选项的参数。

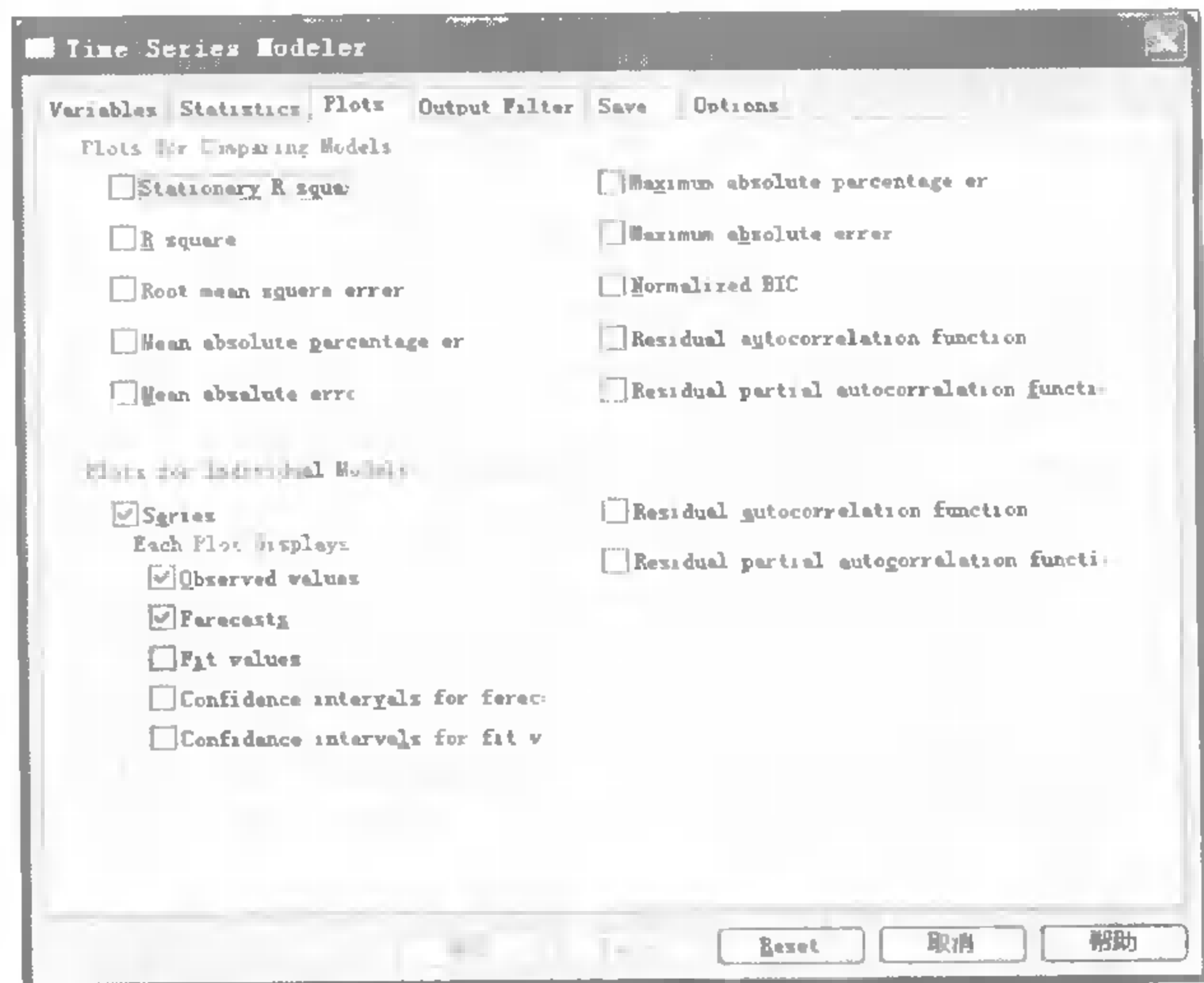


图 15-4 Create Models 的作图设置

(1) Plots for Comparing Models 栏，设置关于模型比较的图形输出。

此设置栏的每个选项，都将单独输出一个图形。设置内容与图 15-3 中的 Fit Measures 栏基本相同，只是增加了两个复选框：Residual autocorrelation function (ACF)，残差自相关函数；Residual partial autocorrelation function (PACF)，残差偏相关函数（PACF）。

(2) Plots for Individual Models 栏，设置关于单个模型的作图选项。

Series 复选框，输出序列图。选中后激活 Each Plot Display 子设置栏的选项，用于指定欲



在序列图中显示的内容，可选项有如下 5 个。

- Observed values 复选框，因变量的原始观测序列。
- Forecasts 复选框，对预测范围的观测的预测值。
- Fit values 复选框，对估计范围的观测的预测值。
- Confidence intervals for forecasts 复选框，预测范围内的置信区间。
- Confidence intervals for fit values 复选框，估计范围内的置信区间。

(3) Residual autocorrelation function (ACF)复选框，输出每个模型的残差自相关序列图。

(4) Residual partial autocorrelation function (PACF)复选框，输出每个模型的残差偏相关序列图。

#### 4. 输出限制选项设置

在图 15-2 中，单击 Output Filter 标签，显示如图 15-5 所示的子设置界面，在此设置关于输出限制选项的参数。

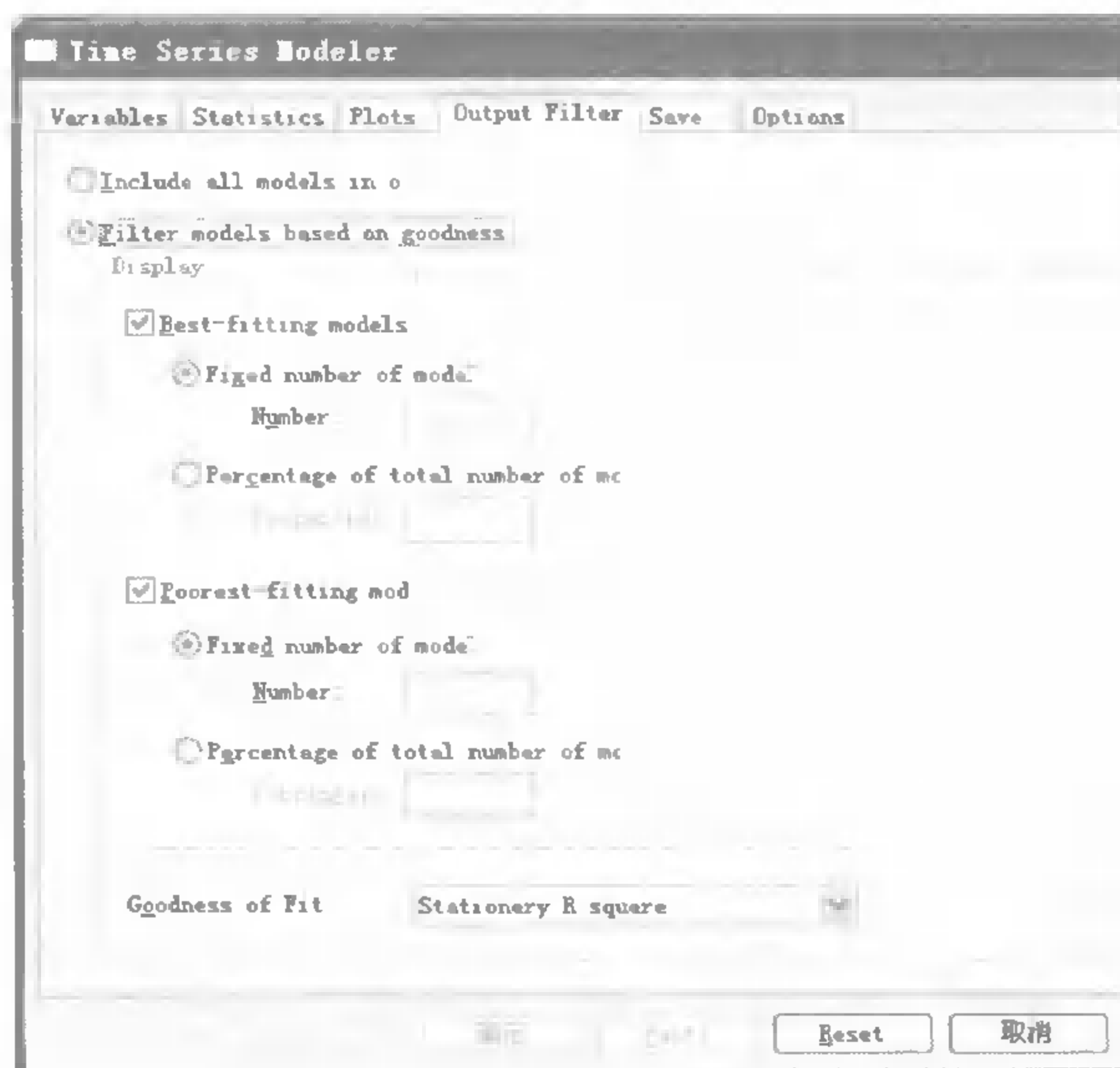


图 15-5 Create Models 的输出设置

(1) Include all models in the output 单选框，表示输出所有模型的分析结果，默认选项。

(2) Filter models based on goodness of fit 单选框，表示只输出某些模型的分析结果。

选中后激活下面的设置选项，只对那些满足一定拟合优度条件的模型，输出分析结果。

① Best-fitting models 复选框，表示输出拟合优度最好的模型。

- Fixed number of models 单选项，在 Number 输入框指定要显示的最好模型个数，若输入值大于总的模型个数，就对所有模型输出。
- Percentage of total number of models 单选项，在 Percentage 输入框指定要显示的最好模型个数占总模型个数的比例。

② Poorest-fitting models 复选框，表示输出拟合优度最差的模型。

- Fixed number of models 单选项，在 Number 输入框指定要显示的最差模型个数，若输入值大于总的模型个数，就对所有模型输出。

- Percentage of total number of models 单选项，在 Percentage 输入框指定要显示的最差模型个数占总模型个数的比例。

③ Goodness of Fit Measure 下拉列表，指定衡量模型优劣的拟和优度统计量。

可选项与图 15-3 中 Fit Measures 栏下的 8 个统计量一样。默认使用的是平稳 R 方统计量

(Stationary R-Squared)，SPSS 给出的计算公式为：
$$R_S^2 = 1 - \frac{\sum_t (Z(t) - \hat{Z}(t))^2}{\sum_t (\Delta Z(t) - \overline{\Delta Z})^2}。$$

## 5. 保存选项设置

在图 15-2 中单击 Save 标签，显示如图 15-6 所示的子设置界面，在此设置关于保存选项的参数。

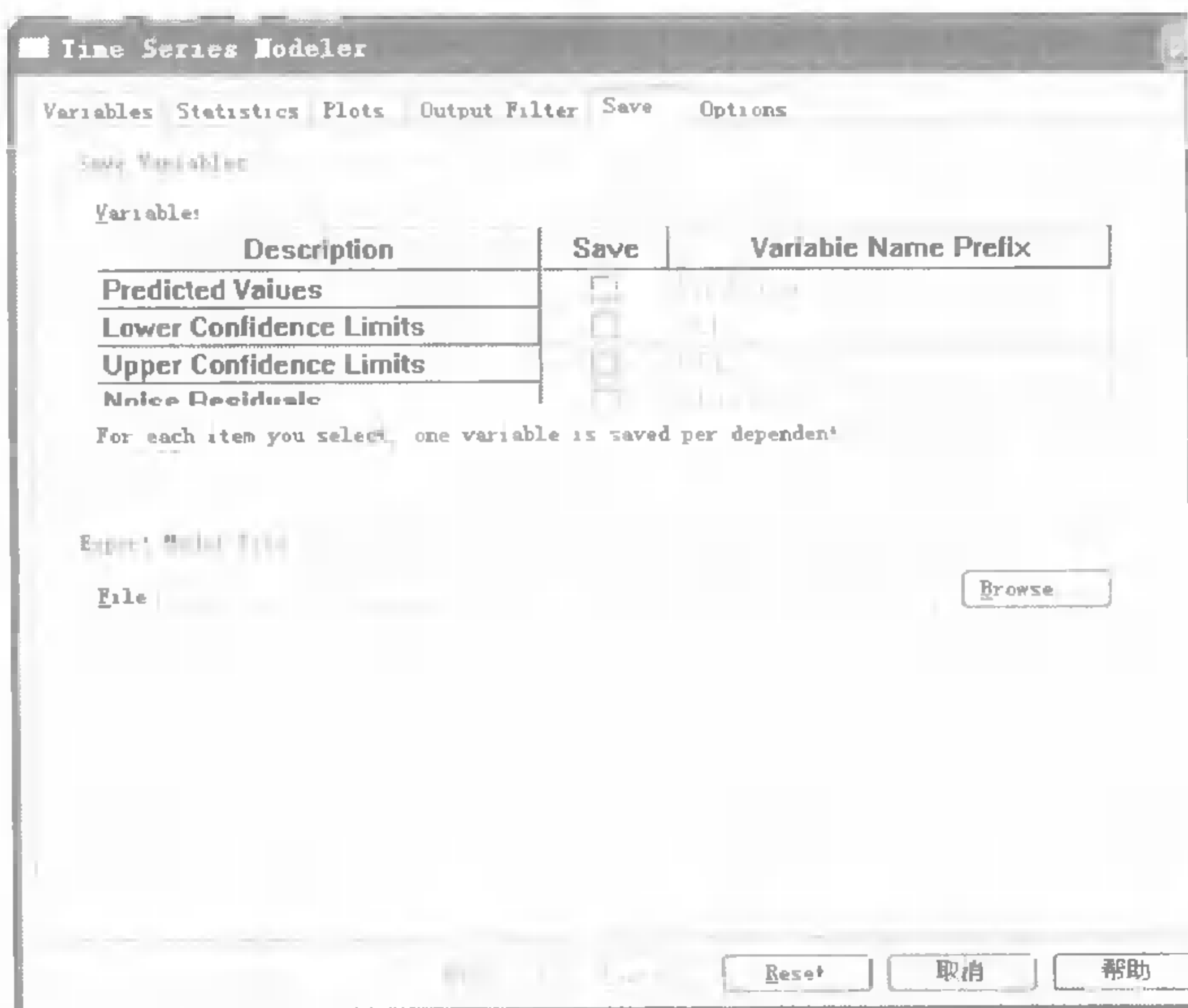


图 15-6 Create Models 的保存设置

(1) Save Variables 栏，设置关于模型预测值的保存选项。

在下面的二维表格中，Description 列给出了待保存的具体内容，可选项有如下 4 个。

- Predicted Values 行，模型预测值。
- Lower Confidence Limits 行，预测值的置信下限。
- Upper Confidence Limits 行，预测值的置信上限。
- Noise Residuals 行，预测值的残差。如果对因变量作了某种变换，此处保存的是变换后序列的残差值。

勾选 Save 列的复选框，就将保存同行的关于预测值的选项；Variable Name Prefix 列，显示默认的变量名前缀，可以对其进行编辑和更改。

(2) Export Model File 栏，设置把所有模型信息输出到指定的 XML 文件。

File 输入框，用于指定 XML 模型文件的路径和名称；或者单击 Browse 按钮打开文件选择对话框指定保存文件。此处保存的模型文件可以应用于 Apply Models 模块，例如当时间序列的记录增加时，可用被保存的模型直接更新预测值，而不必重新进行建模。

## 6. Options 选项的设置

在图 15-2 中，单击 Options 标签，显示如图 15-7 所示的子设置界面，在此设置关于预测范围、缺失值处理方式、置信区间等选项的参数。

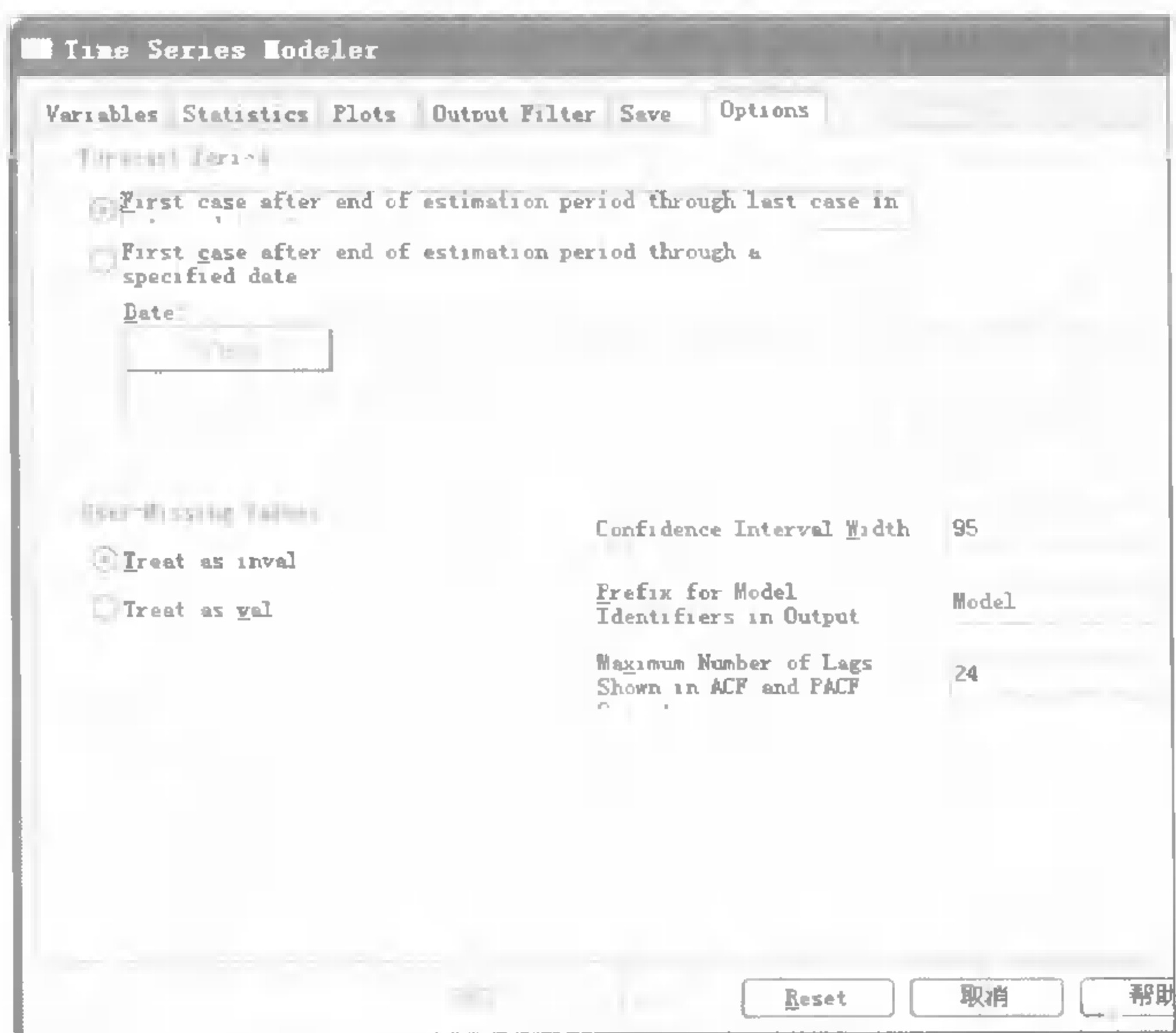


图 15-7 Create Models 的 Options 设置

(1) Forecast Period 栏，设置模型的预测范围，可选方法有如下两种。

☒ First case after end of estimation period through last case in active dataset 单选项

表示预测范围从估计模型所用数据的最后一个记录，到当前数据集的最后一个记录。当估计模型所用数据并非全部数据时选用此项，常用来比较预测值和观测值的误差。

☐ First case after end of estimation period through a specified date 单选项

表示预测范围从估计模型用到数据的最后一个记录，到用户指定的某个日期，常用来预测超过当前数据集的时间范围的记录。Date 栏，用于指定要预测的日期。

如果当前数据集已经定义了日期变量，Date 栏会显示与当前日期格式相对应的 Year、Month（或 Cycle 周期等）输入列；如果当前数据集没有定义日期，则此处只显示一个输入列 Observation，这时输入与当前数据集（Data Editor 窗口）对应的记录行号就可以了。

(2) User-Missing Values 栏，设置缺失值的处理方式，有如下两个选择。

☒ Treat as invalid 单选框，表示把用户定义缺失值当作系统缺失值对待，作为无效数据。

☐ Treat as valid 单选框，表示把用户定义缺失值作为有效据。

(3) Confidence Interval Width (%) 输入框，指定置信区间的宽度，它用于模型预测值和残差的自相关函数，默认为 95 (%)。

(4) Prefix for Model Identifiers in Output 输入框，指定在输出结果中用以区分不同模型的名称前缀，默认为“Model”。对于在图 15-2 中选入的每个因变量，都将生成各自单独的模型，SPSS 会以带前缀的模型名称加以区别，名称的后缀为依次增加的整数。

(5) Maximum Number of Lags Shown in ACF and PACF Output 输入框，指定自相关函数和偏相关函数的最大延迟阶数，默认数值为 24 阶。

### 15.1.2 Apply Models 的通用设置选项

依次单击菜单“Analyze→Time Series→Apply Models...”执行应用模型的功能，其主设置界面如图 15-8 所示。它的功能是应用先前由 Create Models 功能输出的某个模型文件，直接对指定数据集进行预测和分析；模型文件通过设置图 15-6 中的 Export Model File 栏输出。



图 15-8 Apply Models 的主设置面板

(1) Model File 输入框，用于指定 XML 模型文件的路径和名称，可以直接输入，也可以单击 Browse 按钮打开文件选择对话框指定模型文件。

(2) Model Parameters and Goodness of Fit Measures 栏，设置模型估计的参数和模型拟合优度统计量的引入方式，有如下两个可选项。

① Load from model file 单选框，表示从指定的模型文件中直接读取。模型参数不再重新估计，有关的拟合优度统计量也使用指定文件中的设置。

② Reestimate from data 单选框，表示利用当前数据集重新估计模型，但不改变读入的 XML 文件中的模型结构；例如，模型文件记录的是 ARIMA(1,0,1)模型，重新估计只更新其参数值，但会保留此模型的形式。但是，异常值是直接从指定模型文件中读取的，不进行重新检测。

③ Forecast Period 栏，指定模型的预测范围，设置方法与图 15-7 中的 Forecast Period 栏的设置相同。

④ 除了如图 15-8 所示的 Models 子标签外，Apply Models 功能的其他子标签面板(Statistics 标签、Plots 标签等)的设置，与前节介绍的 Create Models 功能相应的子标签设置基本相同。

## 15.2 时间序列数据的预分析

对时间序列数据进行分析之前，需要做许多的准备和检验工作。在 SPSS 中对时间序列数据的整理功能主要是指：定义日期(Define Dates)、创建时间序列(Create Time Series)和



缺失值替换 (Replace Missing Values)。

预分析的一般步骤是：首先，检查数据是否存在缺失值，并进行缺失值替换的操作；其次，SPSS 是不会自动把数据集中的记录识别为时间序列的，必须要加以定义；最后，原始的时间序列数据往往需要经过重新定义和计算，使其成为平稳序列，然后才能用于进一步分析。

### 15.2.1 缺失值替换

时间序列模型一般都要求序列数据是完整无缺的，但实际上数据常常是不完整的。

当序列中存在缺失数据时，如果采用直接剔除的方法，容易使剔除缺失值之后的数据周期发生错位，在这种情况下应当使用 Replace Missing Values 过程，采用适当的方法对缺失值进行替换，并将替换结果保存到新的变量中。

依次单击菜单“Transform→Replace Missing Values...”，执行缺失值替换功能，其主设置界面如所图 15-9 所示。

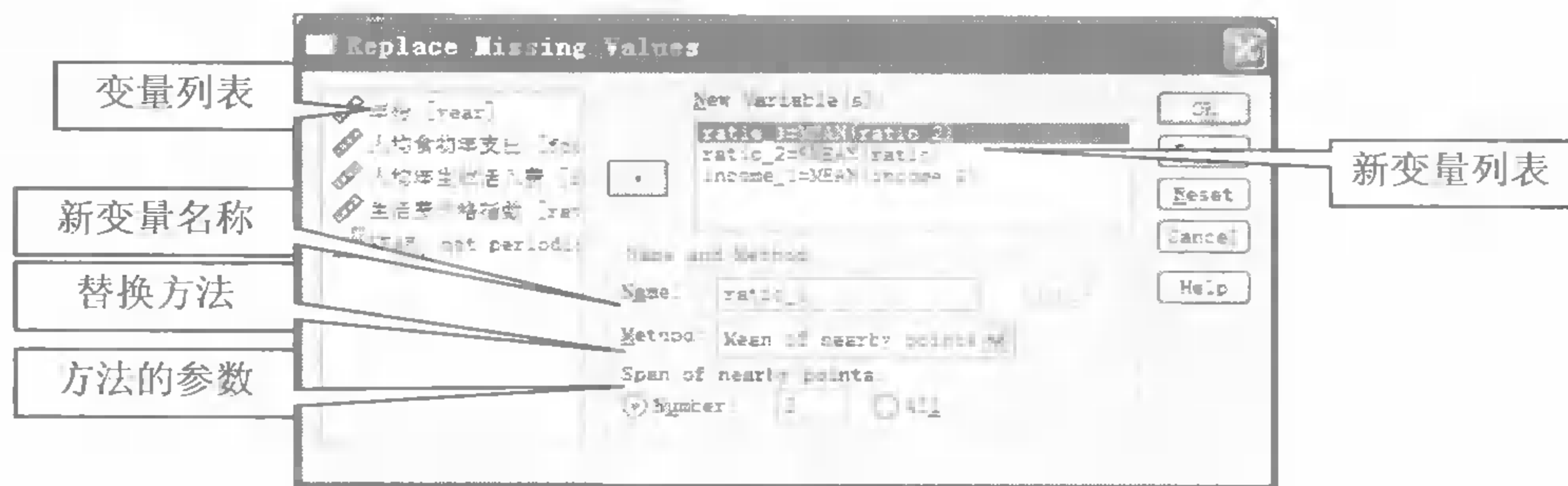


图 15-9 缺失值替换的设置对话框

(1) New Variable(s)列表框，用于从变量列表选入可能含有缺失值的变量。对于同一个变量，可以选入多次，以设置不同的替换方法。

(2) Name and Method 栏，用于设置缺失值替换的参数。

在上面的新变量列表框单击选中某个变量后，在此修改选中变量的缺失值替换方式，修改后需要单击 Change 按钮加以确定。

① Name 输入框，用于指定新变量的名称。

② Method 下拉列表，用于选择替换缺失值的方法，可选项有如下 5 个。

- Series mean 全体序列的均值，为默认选项。
- Mean of nearby points 相邻若干点的均值。在 Number 输入框指定采用当前缺失值前后各多少条记录计算均值；All 单选框，表示使用所有记录。
- Median of nearby points 相邻若干点的中位数。在 Number 输入框指定采用当前缺失值前后各多少条记录计算中位数；All 单选框，表示使用所有记录。
- Linear interpolation 线性内插，使用当前缺失值前后分别最近的两个有效数据计算均值。如果序列的最前或最后记录有缺失值，不对其做任何替换。
- Linear trend at point 该点的线性趋势。将记录号作为自变量，序列值作为因变量进行回归，求得的该点估计值。

### 15.2.2 定义时间变量

时间序列数据的一个明显特点就是记录依某种时间排列。在 SPSS 中，只有经用户定义

了时间变量之后，系统才能识别指定序列的时间特征，比如周期等。

依次单击菜单“Data→Define Dates...”，执行建立时间变量的功能，其主设置界面如图 15-10 所示，单击选中 Cases Are 列表中的“Years Quarters”选项。

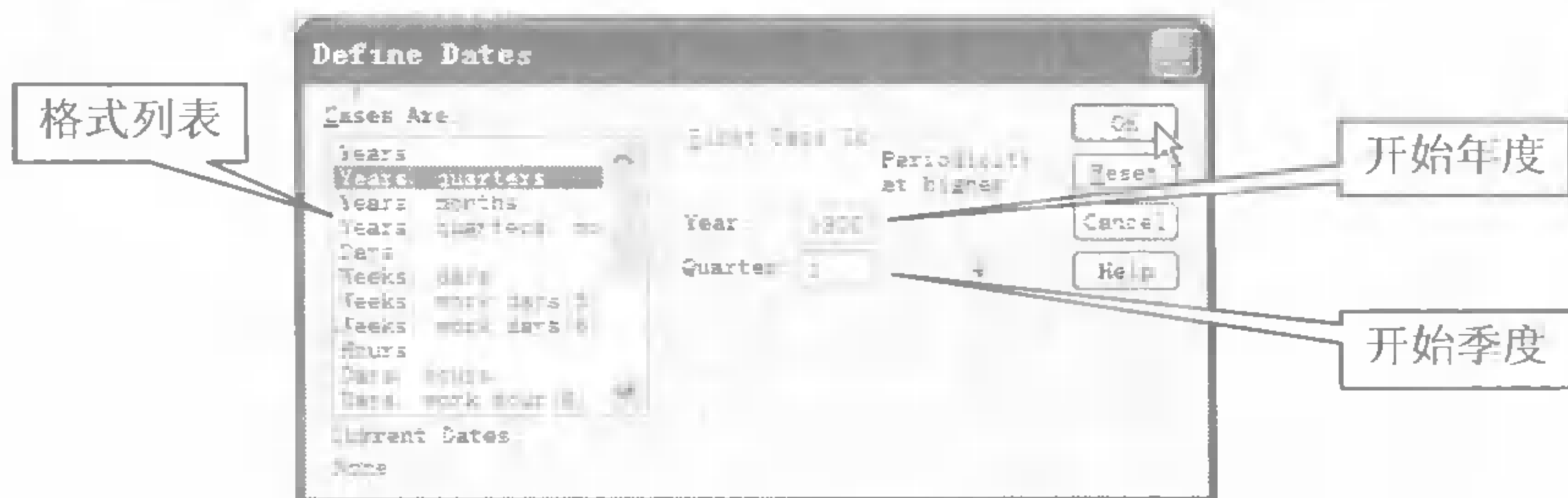


图 15-10 建立时间变量对话框

(1) Cases Are 列表框，给出了多种时间格式供用户选择。

序列的周期由时间格式的最小时间单位决定，例如：“Years Quarters”的周期为 4。

注意：Not dated 格式表示删除文件中已经定义的时间变量，即删除名称为 year\_、quarter\_、month\_ 等的变量；Custom 格式表示直接采用当前数据集中由命令语句所建立的时间变量。

(2) First Case Is 子设置栏，指定序列的起始时间。

与 Cases Are 列表选中的格式不同，此设置栏的显示内容也会有所变动。例如：指定了“Years Quarters”格式，此处显示 Year（年份）、Quarter（季度）两个输入框以待设置。

右侧的 Periodicity at higher level 下面，显示指定时间格式的周期，例如：选择“Years Quarters”格式时，此处显示周期为 4。

Current Dates 栏（在界面左下角），初次定义时间变量时此处显示“None”；定义好时间变量之后，如果再次进入该对话框，此处会显示当前数据集的时间变量信息。

(3) 新建时间变量的示例。在图 15-10 中，单击 OK 按钮运行。返回 Data Editor 窗口，当前数据集增加如图 15-11 所示的新变量，它们的含义如第 5 列的 Label 标签所示。

YEAR_	Numeric	8	0	YEAR not periodic	None	None	10	Right	Ordinal
QUARTER_	Numeric	1	0	QUARTER period 4	None	None	8	Right	Ordinal
DATE_	String	1		Date Format "QQ YYYY"	None	None	9	Left	Ordinal

图 15-11 新增时间变量

### 15.2.3 时间序列的平稳化

定义了时间变量之后，就基本建立了待分析的时间序列数据。但是并非随便建立一个序列就算万事大吉，许多时间序列分析方法都要求序列必须满足平稳性的条件。

#### 1. 平稳性简介

时间序列数据可以看作是随机过程的一个样本，需要根据如下 3 个要求判断它是否平稳。

- 均值不随时间变化。
- 方差不随时间变化。
- 自相关系数只与时间间隔有关，而与所处的具体时刻无关。

实际上，大多数用初始数据建立的时间序列都是不平稳的，所以在着手建模之前应该先

验证序列的平稳性，并把不平稳的序列转化为平稳序列。

SPSS 的 Create Time Series 过程可以对初始序列进行一些预处理，它使用差分、移动平均等变换方法，由原始序列导出一条或多条可能满足平稳性要求的新序列。虽然 SPSS 的某些时间序列分析模块中，也嵌入了诸如差分、对数变换等的运算功能，但 Create Time Series 过程的功能更加强大。

另外，在对序列进行平稳化之前，建议先作一个时间序列观测值对时间的线形图，以观察序列的线性趋势、周期性、方差齐性等特点，再以此为参考选择恰当的方法进行平稳化。

## 2. 参数设置

依次单击菜单“Transform→Create Time Series...”，执行建立新时间序列的功能，其主设置界面如所图 15-12 所示。

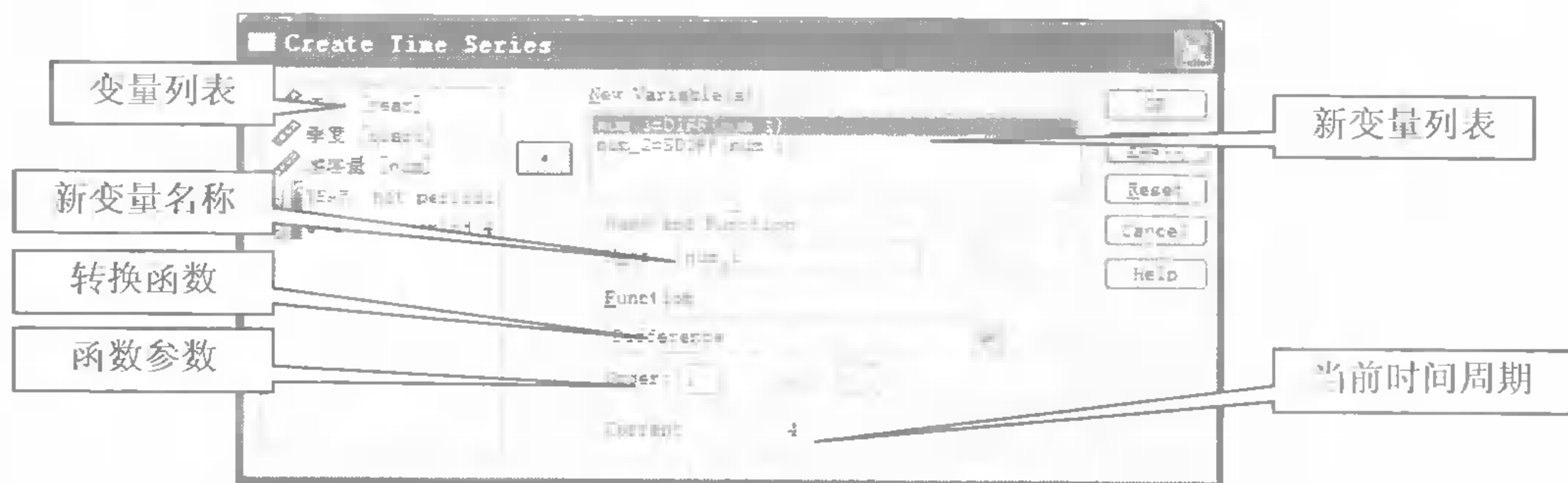


图 15-12 建立新时间序列的对话框

(1) New Variable(s)列表框，用于从变量列表选入原始的时间序列变量。对于同一个变量，可以选入多次，以设置不同的转换方法。

(2) Name and Function 栏，用于设置对变量进行转换的参数。

在上面的新变量列表框单击选中某个变量后，在此修改对选中变量的转换方法和参数，修改后需要单击 Change 按钮加以确定。

① Name 输入框，用于指定新变量的名称。

② Function 下拉列表，用于选择转换原序列的方法，可选项有如下 9 种。

● Difference，计算变量的一般差分（非季节性）。差分是序列平稳化的常用手段，它的作用是消除前后数据的依赖性。下面的 Order 输入框用于指定差分的阶数，默认为 1。差分会损失数据，阶数为  $n$  时数据损失  $n$  个，新变量的前  $n$  个数据用缺失值表示。

● Seasonal difference，季节性差分，表示差分的间距由数据的周期决定。没有定义时间周期的序列不能做季节性差分。下面的 Order 输入框用于指定差分的阶数，默认为 1，阶数为  $n$  时数据损失为季节周期的  $n$  倍，把新变量里排在最前的记录作为缺失值。

● Centered moving average，中心移动平均，以当前值为中心计算指定范围的均值。下面的 Span 输入框用于指定计算均值的范围，默认为 1。如果指定范围（记为 span）是奇数，计算当前值以及前后各  $(span-1)/2$  个数的均值；如果指定范围是偶数，计算时把当前值乘以 2，再加上前后各  $span/2$  个数，然后除以  $(span+2)$  得到均值；如果指定 span 为 1，则均值等于原始值。对于序列最初和最后的几个记录，中心移动平均法无法处理，例如：如果 span 为 5，则开始、结束时的各 2 个记录都不能计算。取移动平均的效果是把序列的噪声部分抵消，保留趋势部分。当数据服从对称的分布特别是正

态分布时，此方法比较合适。

- Prior moving average, 向前移动平均。计算当前值之前的某个范围的均值。下面的 Span 输入框用于指定范围，默认为 1。

- Running median, 移动中位数，以当前值为中心计算指定范围的中位数。下面的 Span 输入框用于指定计算中位数的范围，默认为 1。span 取不同值时的计算方法，与中心移动平均 (Centered moving average) 相似，此方法比移动平均法更为稳健。

☆ Cumulative sum, 累计和，表示以原序列的累计和作为新序列。

☆ Lag 滞后处理，表示让原始序列向后滞留指定的阶数。下面的 Order 输入框用于指定滞后的阶数，默认为 1。

☆ Lead: 提前值，和滞后相反，让原序列提前指定的阶数，阶数由 order 输入框中指定。

☆ Smoothing 光滑处理，表示计算原始序列的 T4253H 平滑序列。

T4253H 是一种复合平滑方法，它先对序列依次作 4 次移动中位数 (running median) 处理，计算范围 (span) 分别为 4、2、5、3；然后以 Hanning 权重再作移动平均处理。这是个对一般序列都很有效的复合平滑器，它先把序列中的异常值剔除，然后再使序列变得更为平滑。T4253H 平滑法要求原序列含有大于三个的记录，而且序列中不能含有缺失值。

假设序列中的任意一点都是由平滑部分加上随机误差组成，那么如果把该点附近的若干点平均一下，误差项就会趋于相互抵消，于是原序列的特征（如波峰、波谷）就更加突出，这就是平滑的作用。均值和中位数是常用的平滑器，但是如何选择合适的平滑器，与数据特征尤其是误差项的特征关系密切。在实际工作中，经常需要对原始序列做多次的移动平均或移动中位数处理，才能得到比较好的平滑曲线，所以如 T4253H 之类的复合平滑器也是很常用的。

(3) Current 栏，显示当前时间变量的周期；如果没有定义时间变量，此处为空。

### 3. 结果示例

假设对变量 num 指定了一般差分 (DIFF) 和中心移动平均 (MA) 两种处理方法，新变量的名称分别为 num\_1、num\_2。如图 15-13 所示，是 Variable View 视图里新建变量的属性。

num_1	Numeric	9	2	DIFF(num 1)	一般差分	None	11	Right	Scale
num_2	Numeric	8	2	MA(num 2 2)	中心移动平均	None	10	Right	Scale

图 15-13 建立新时间序列输出的新变量

## 15.3 指数平滑模型

移动平均的目的是为了去除序列中的误差影响（不规则成分），使原始序列变得较为平滑，于是它就使趋势成分和循环成分变得更为清晰，易于分析。

另一种平滑序列的方法就是指数平滑法 (Exponential Smoothing)，它是加权移动平均法的一种特殊情况，仍然使用特定范围内记录的加权平均值作为预测。与其他加权移动平均法不同的是，指数平滑使用当前时刻之前的全部数据来决定它的平滑值；而且它只需要指定一个参数，即最近时期记录值的权重，其他时期记录值的权重由自动推算得来，而且离预测期越远，权重变得越小。



15.3.1 指数平滑的基本原理

指数平滑法最早是由 C.C.Holt 在 1958 年左右提出的，最初只应用于以无趋势、非季节性作为基本形式的时间序列分析。后经过 Brown、Winter 等统计学家的研究和发展，使它逐步适用于更多类型的数据序列。指数平滑法的估计是非线性的，它的目标是使预测值和观测值之间的均方误差（MSE）达到最小。本节简单介绍指数平滑法的基本公式和预测方程。

1. 简单加权平均

记  $x_1, x_2, \dots, x_t, \dots$  为一时间序列，用前  $t$  期的观测值预测第  $t+1$  期的取值时，设赋予第  $i$  期观测的权重为  $w_{t+1-i} (i=1, 2, \dots, t)$ ,  $w_1 > w_2 > \dots > w_t$ ，则计算公式就为  $\hat{x}_{t+1} = \frac{\omega_1 x_t + \omega_2 x_{t-1} + \dots + \omega_t x_1}{\omega_1 + \omega_2 + \dots + \omega_t}$ ，这就是所谓的加权平均法。此方法需要自行决定权重，主观性较大，且计算较为繁琐。

2. 自动加权平均

自动取权重的思路是：自当前期向前，让各期权重按指数规律下降，把第  $t, t-1, \dots$  期观测的权重依次记为： $\alpha, \alpha\beta, \alpha\beta^2, \dots (\alpha > 0, 0 < \beta < 1)$ ；为使权重之和等于 1，令  $t \rightarrow \infty$  时，有下式成立： $\alpha + \alpha\beta + \alpha\beta^2 + \dots = 1$ 。由此可得，第  $t, t-1, \dots$  期观测的权重依次为： $\alpha, \alpha(1-\alpha), \alpha(1-\alpha)^2, \dots$ 。继续考虑  $t$  充分大时的情形，这时有下式成立： $T_t = \alpha x_t + \alpha(1-\alpha)x_{t-1} + \alpha(1-\alpha)^2 x_{t-2} + \dots$ ，把滞后 1 期的估计值单独提出，可得： $T_t = \alpha x_t + (1-\alpha)T_{t-1}$ ，此式称为指数平滑法的基本公式。这个公式是由递推形式给出的， $\alpha$  叫做平滑常数，且满足： $0 < \alpha < 1$ ； $T_t$  称为时间序列  $x_t$  第  $t$  期的指数平滑值。指数平滑法的预测方程就是： $\hat{x}_{t+1} = T_t$ ，即把第  $t$  期的指数平滑值作为第  $t+1$  期的预测值，它既继承了加权平均法重视近期数据的思想，又能克服权重不易确定的局限性。

15.3.2 指数平滑模型的参数设置

在图 15-2 中，单击 Method 下拉列表并选中 Exponential Smoothing 项，再单击 Criteria 按钮，弹出如图 15-14 所示的指数平滑参数设置对话框。

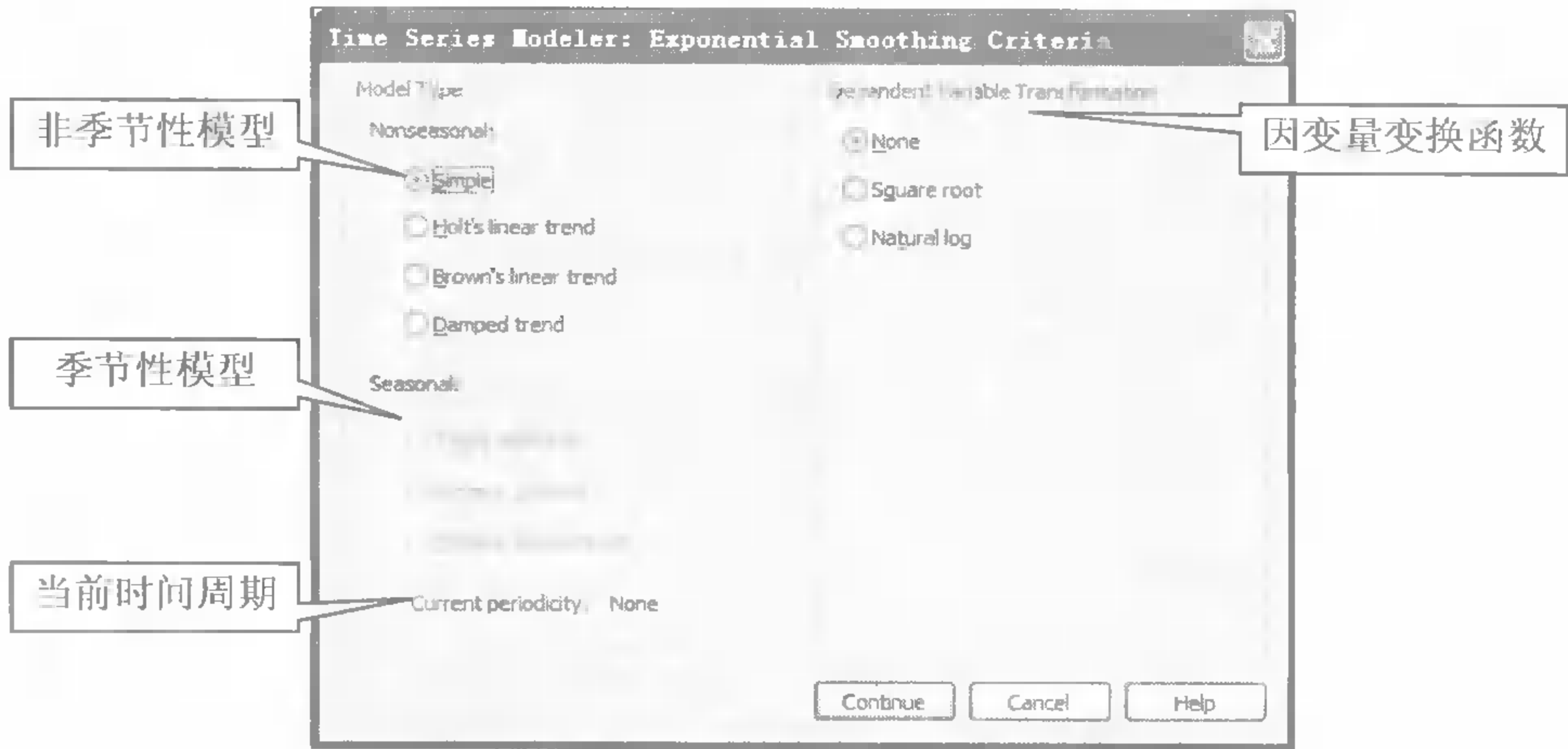


图 15-14 指数平滑模型参数设置

Dependent Variable Transformation 栏, 指定对因变量的变换方法, 可选项有如下 3 个: None, 不作变换; Square root, 开平方变换; Natural log, 自然对数变换。

Current Periodicity 栏, 显示当前数据集的周期。如果没有定义时间变量, 或者指定的时间变量没有周期, 此处显示 “None”。

Model Type 栏, 指定指数平滑模型的类型。Nonseasonal 子设置栏给出的都是无季节成分的模型; Seasonal 子设置栏给出的都是含季节成分的模型。下面依次介绍每个模型的适用情况。

### 1. Simple 模型

简单 (单一) 指数平滑法, 它在分析不含趋势成分和季节成分的序列时, 具有十分良好的效果。Simple 模型与无常数项的 ARIMA (0,1,1) 模型十分相似。

简单指数平滑法使用的就是指数平滑的基本预测公式。

令  $y_t$  为  $t$  时刻的观测数据,  $S_t$  为平滑后的数据,  $A$  为一个介于 0 到 1 之间的实数, 根据基本预测公式, 有:  $S_t = Ay_t + (1-A)S_{t-1}$ 。其中  $A$  称为平滑常数, 越小的  $A$  值 ( $A < 0.1$ ) 适用于波动越明显的时间序列数据, 即不规则效应越明显的序列; 反之依然。

一般地, 有下式成立:  $S_t = Ay_t + A(1-A)y_{t-1} + A(1-A)^2 y_{t-2} + \dots + A(1-A)^{t-2} y_2 + (1-A)^{t-1} y_1$ , 例如:  $A=0.5$ , 则:  $S_t = 0.5y_t + 0.25y_{t-1} + 0.125y_{t-2} + 0.062y_{t-3} + \dots$ 。

由此可知, 每一个平滑后的数据都是由过去的数据加权求和后所得。越接近当期的数据, 其权重越大, 对当期的影响也越大; 反之, 越早期的数据, 其权重越小, 对当期的影响也越小, 这十分符合我们的逻辑直觉。

### 2. Holt 线性趋势模型

霍特线性趋势模型 (Holt's linear trend), 适用于处理具有线性趋势成分、但不含季节成分的时间序列数据。Holt 模型与 ARIMA (0,2,2) 模型十分相似。

Holt 线性趋势模型的基本预测方程为: 
$$\begin{cases} T_t = \alpha x_t + (1-\alpha)(T_{t-1} + b_{t-1}) \\ b_t = \beta(T_t - T_{t-1}) + (1-\beta)b_{t-1} \\ \hat{x}_{t+\tau} = T_t + b_t\tau, \quad \tau = 1, 2, \dots \end{cases}$$
 其中各符号的意义

如下:

$t$ : 当前期;  $\tau$ : 预测超前期数, 也称之为预测步长;

$T_t, T_{t-1}$ : 利用前  $t$  期或前  $t-1$  期数据, 对第  $t$  期或第  $t-1$  期趋势的估计;

$b_t, b_{t-1}$ : 利用前  $t$  期或前  $t-1$  期数据, 对趋势增量  $b$  的估计;

$x_t$ : 第  $t$  期的实际观察值;  $\hat{x}_{t+\tau}$ : 利用前  $t$  期数据, 对第  $t+\tau$  期的预测值;

$\alpha, \beta$ : 平滑常数, 满足  $0 < \alpha, \beta < 1$ ;

利用如上的基本预测方程进行计算时, 除了需要指定两个平滑常数  $\alpha, \beta$  之外, 还需要事先指定两个初值  $T_1$  和  $b_1$ 。

### 3. Brown 线性趋势模型

布朗线性趋势模型 (Brown's linear trend), 适用于处理具有线性趋势成分、但不含季节成分的时间序列数据。Brown 模型是 Holt 模型的一种特殊情况, 它与 ARIMA (0,2,2) 模型十分相似, 并且其二阶滑动平均的参数, 就是一阶滑动参数 1/2 倍的平方。

Brown 模型的基本原理，与线性二次移动平均法相似。它的基本预测方程为：

$$\begin{cases} S'_t = \alpha x_t + (1-\alpha)S'_{t-1} \\ S''_t = \alpha S'_t + (1-\alpha)S''_{t-1} \end{cases}, \begin{cases} a_t = 2S'_t - S''_t \\ b_t = \frac{\alpha}{1-\alpha}(S'_t - S''_t) \end{cases}, F_{t+m} = a_t + b_t m$$

其中：\$S'\_t\$ 为一次指数平滑值，\$S''\_t\$ 为二次指数平滑值；\$m\$ 为预测超前期数，\$F\_{t+m}\$ 为第 \$m\$ 期预测值。

#### 4. 控制趋势模型

控制趋势模型 (Damped trend)，适用于处理具有一个逐渐消失的线性趋势成分、但不含季节成分的时间序列数据。Damped 趋势模型与 ARIMA (1,1,2) 模型十分相似。

#### 5. 简单季节模型

简单季节模型 (Simple seasonal)，适用于处理含有不随时间变化的季节成分、但不含趋势成分的时间序列数据。简单季节模型与 SARIMA (0,1,1) × (0,1,1)<sub>s</sub> 模型十分相似，其中 \$s\$ 为时间序列的周期。

#### 6. Winters 加法模型

温特加法模型 (Winters' additive)，适用于处理包含线性趋势成分、且包含一个不依赖序列水平的季节成分的时间序列数据。Winters 加法模型与 SARIMA (0,1,0) × (0,1,1)<sub>s</sub> 模型十分相似，其中 \$s\$ 为时间序列的周期。

Winters 加法模型的基本预测方程为：

$$\begin{aligned} s_t &= \alpha \frac{x_t}{I_{t-L}} + (1-\alpha)(S_{t-1} + b_{t-1}) & 0 < \alpha < 1 \\ b_t &= \gamma(S_t - S_{t-1}) + (1-\gamma)b_{t-1} & 0 < \gamma < 1, \quad F_{t+m} = (S_t + b_t m)I_{t-L+m} \\ I_t &= \beta \frac{x_t}{S_t} + (1-\beta)I_{t-L} & 0 < \beta < 1 \end{aligned}$$

其中：\$L\$ 为季节长度，\$I\$ 为季节的修正系数；\$m\$ 为预测超前期数，\$F\_{t+m}\$ 为第 \$m\$ 期预测值。

#### 7. Winters 乘法模型

温特乘法模型 (Winters' multiplicative)，适用于处理包含线性趋势成分、且包含一个依赖序列水平的季节成分的时间序列数据。没有与 Winters 乘法模型相似的 SARIMA 模型。

### 15.3.3 指数平滑模型实例分析

#### 1. 数据和问题描述

(1) 数据文件。研究人员收集了 1950~1990 年有关天津食品消费的数据，本节利用指数平滑模型建模，分析这段时间内食品消费的变化情况，数据格式如图 15-15 所示，所用数据文件为“天津食品消费相关数据.sav”。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	year	Numeric	8	0	年份	None	None	8	Right	Scale
2	food	Numeric	8	2	人均食物年支出	None	None	8	Right	Scale
3	income	Numeric	8	2	人均生活费年收入	None	None	8	Right	Scale
4	ratio	Numeric	8	2	生活费价格指数	None	None	8	Right	Scale

图 15-15 天津食品消费相关数据格式

(2) 查看当前日期变量数据文件。依次单击菜单“Data→Define Dates...”，打开定义时间变量的对话框，如图 15-16 所示；单击 Cancel 按钮返回 Data Editor 窗口。

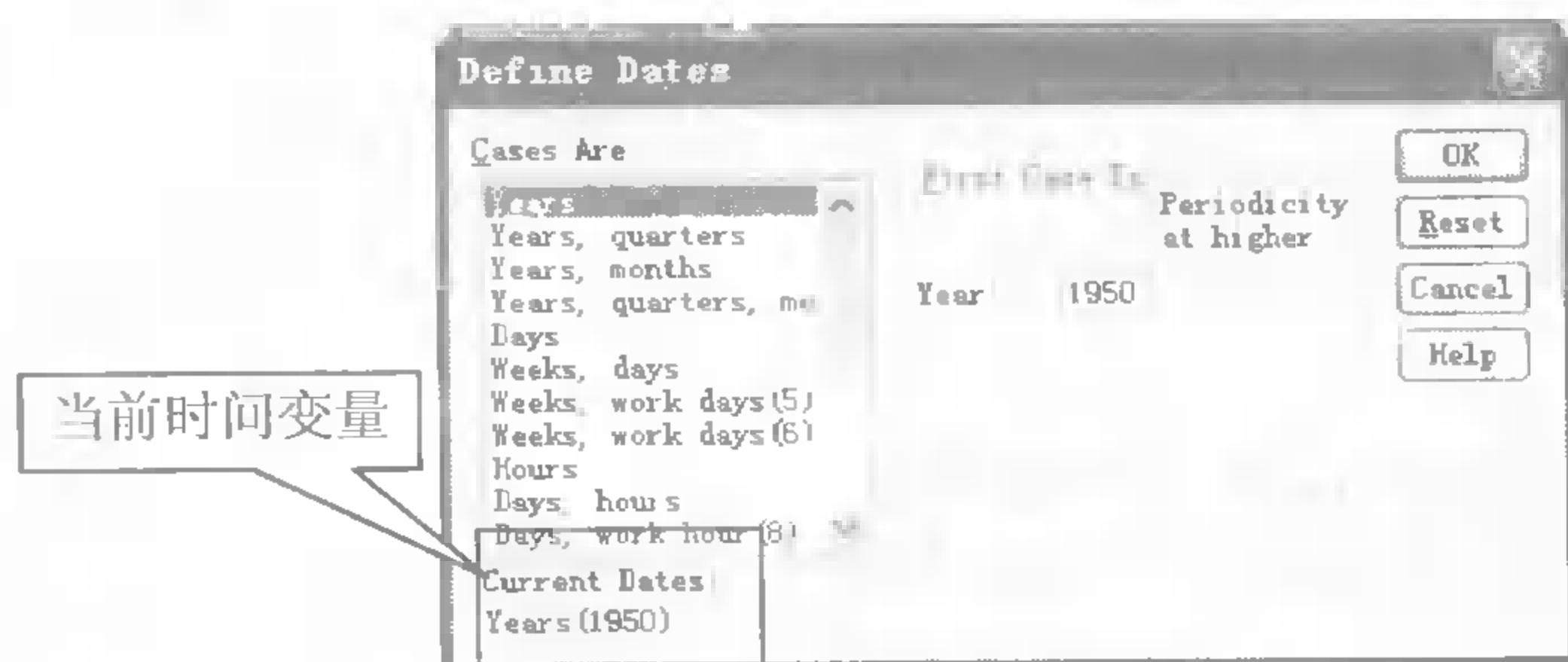


图 15-16 定义时间变量对话框


在 Current Dates 子设置栏，显示了当前时间变量的日期格式为“Years (1950)”，表示观测的起始时间为 1950 年，没有季节周期。

## 2. 参数设置

依次单击菜单“Analyze→Time Series→Create Models...”，打开建立模型的主设置面板，如图 15-17 所示。

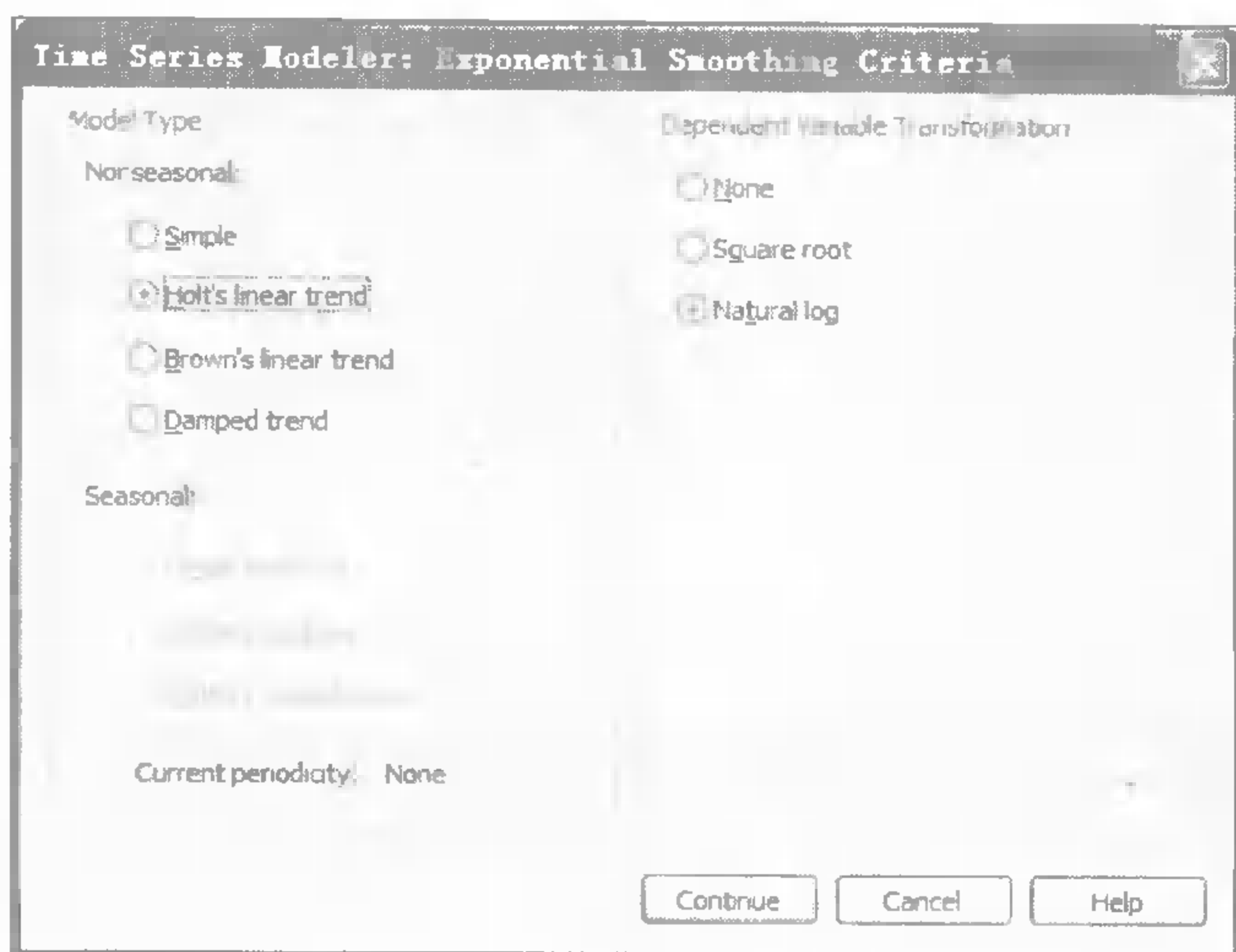


图 15-17 指数平滑案例的参数设置主界面

(1) 变量和模型选择。在 Variables 列表框单击选中人均食物年支出，单击从上至下第一个  按钮，将其作为因变量选入 Dependent Variables 列表框；单击 Method 下拉列表选中 Exponential Smoothing 选项。



(2) 指定模型类型。在图 15-17 中, 单击 Criteria 按钮, 弹出如图 15-18 所示的模型设置对话框。单击选中 Holt's linear trend 单选框; 单击选中 Natural log 单选框; 单击 Continue 按钮返回主面板。



15-18 指数平滑案例的模型设置

(3) 其他设置。在图 15-17 中, 单击 Plots 标签, 打开如图 15-19 所示的作图选项设置界面。勾选 Fit values 复选框; 勾选 Residual autocorrelation function (ACF)复选框和 Residual partial autocorrelation function (PACF)复选框。

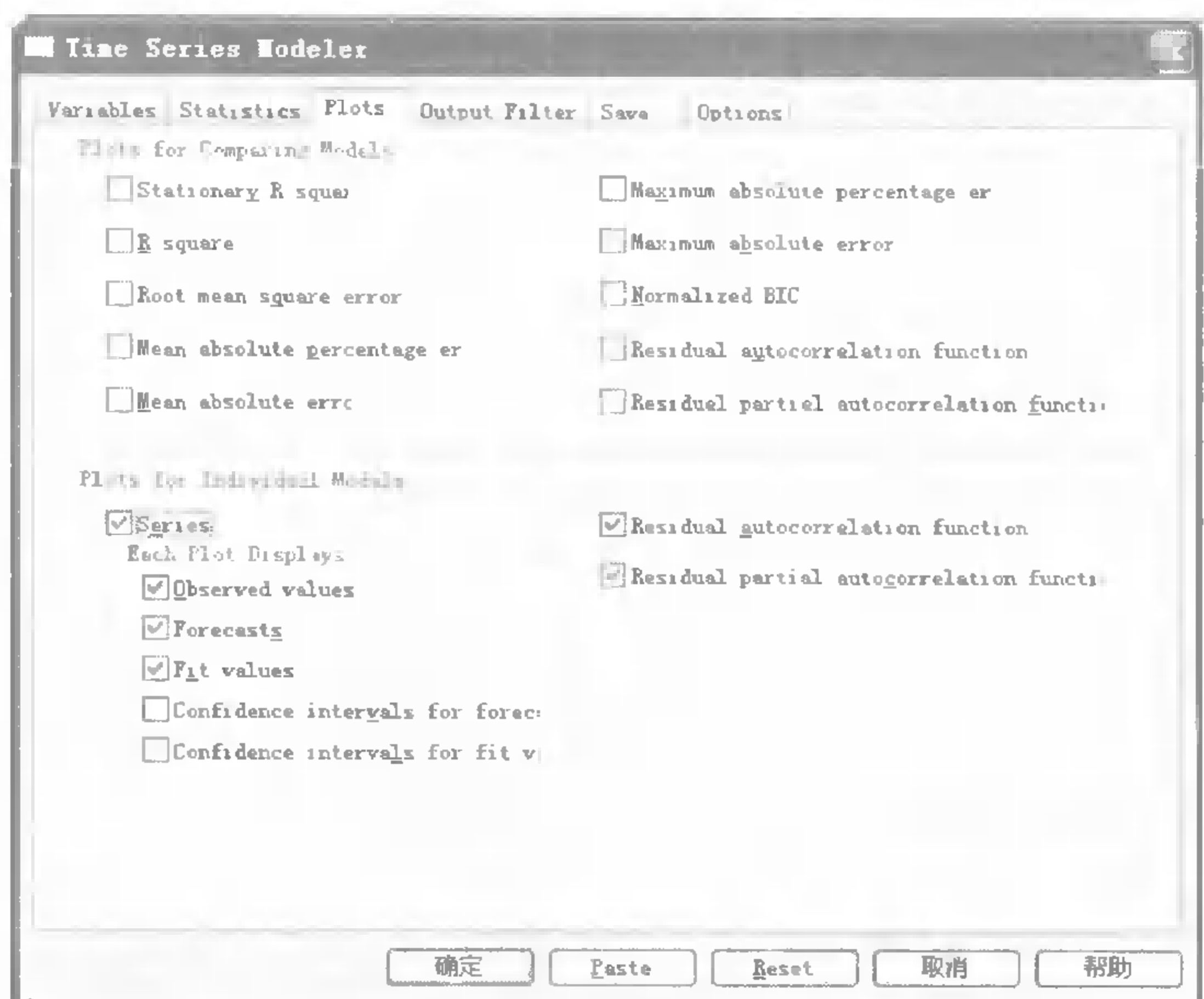


图 15-19 作图选项的设置

在图 15-17 中, 单击 Options 标签, 打开如图 15-20 所示的选项设置界面。单击选中 First case after end of estimation period through a specified date 单选框; 在 Year 下面的单元格输入“1991”; 单击 Variables 标签返回图 15-17 所示的变量设置界面。

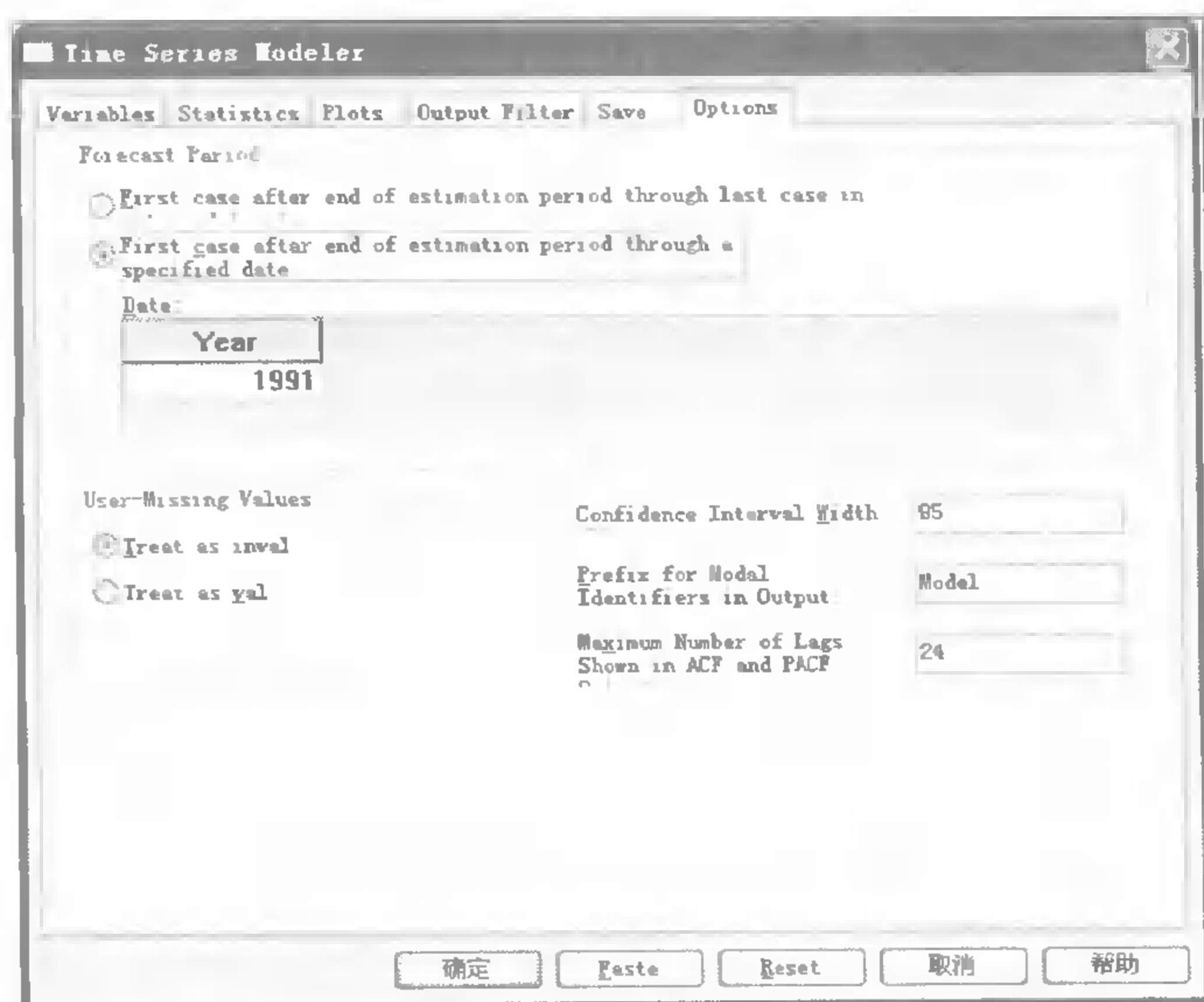


图 15-20 Options 选项的设置

### 3. 结果分析

在图 15-17 中单击“确定”按钮运行，SPSS Viewer 窗口的输出结果如图 15-21~图 15-23 所示。

模型描述											
模型 ID										模型类型	
Model_1 人均食物年支出										Econ	

模型拟合											
拟合统计量	均值	SE	最小值	最大值	百分点						
					5	10	25	50	75	90	95
平稳R方	.470		.470	.470	.470	.470	.470	.470	.470	.470	.470
R方	.994		.994	.994	.994	.994	.994	.994	.994	.994	.994
RMSE	14.351		14.351	14.351	14.351	14.351	14.351	14.351	14.351	14.351	14.351
MAPE	4.388		4.388	4.388	4.388	4.388	4.388	4.388	4.388	4.388	4.388
MaxAPE	15.422		15.422	15.422	15.422	15.422	15.422	15.422	15.422	15.422	15.422
MAE	9.686		9.686	9.686	9.686	9.686	9.686	9.686	9.686	9.686	9.686
MaxAE	31.585		31.585	31.585	31.585	31.585	31.585	31.585	31.585	31.585	31.585
正态化的BIC	5.509		5.509	5.509	5.509	5.509	5.509	5.509	5.509	5.509	5.509

模型统计量						
模型	预测变量数	模型拟合统计量 平稳R方	Ljung-Box Q(18)			离群值数
			统计量	DF	Sig.	
人均食物年支出 Model_1	0	.470	29.324	16	.210	0

图 15-21 关于模型的基本统计信息输出

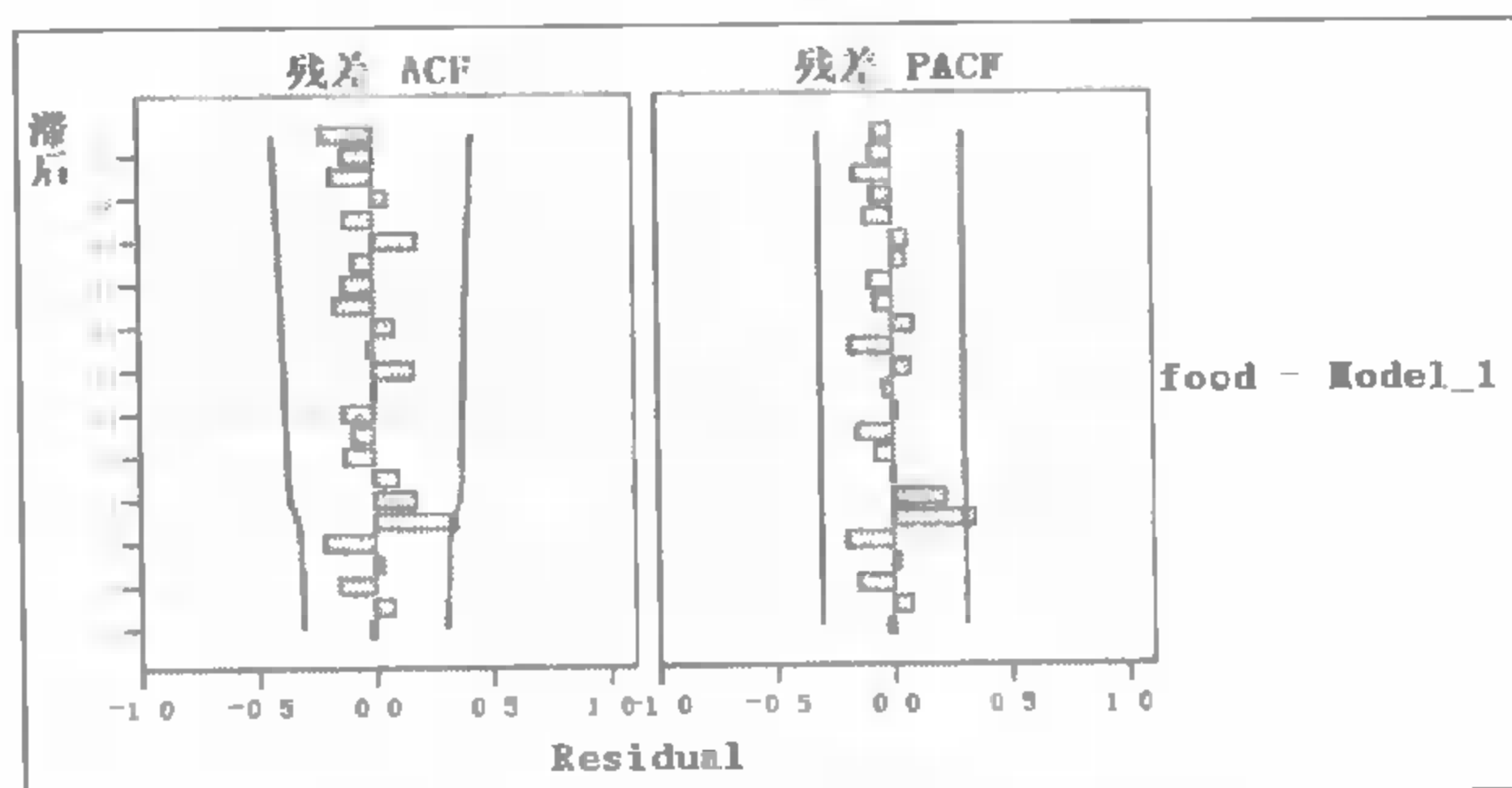


图 15-22 残差的相关函数序列图

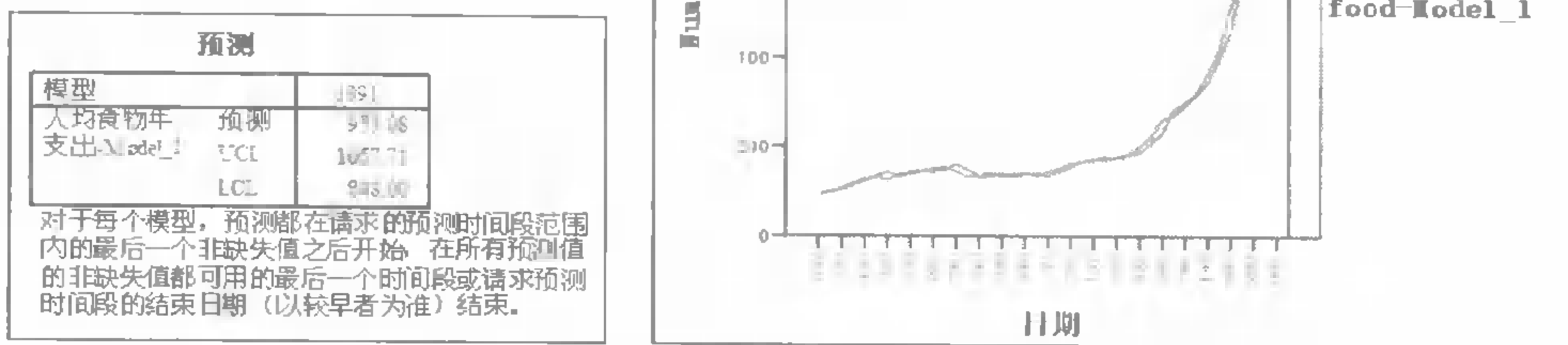


图 15-23 预测结果和图形

(1) 模型基本统计信息。如图 15-21 所示，“模型描述”表格给出了当前模型所使用的分析变量和方法。

“模型拟合”表格，给出了包括平稳 R 方在内的 8 个拟合优度统计量。

从“模型统计量”表格看，平稳 R 方统计量的取值大于 0 (0.470)，说明当前 Holt 线性模型要优于基本的均值模型。

(2) 残差序列图。如图 15-22 所示，是关于残差的自相关 (ACF) 和偏自相关 (PACF) 序列图。可见，两个图形都没有显著的趋势特征（拖尾或截尾），故而可以初步判断本例所用模型是比较恰当的。

(3) 预测结果和拟和图形输出。如图 15-23 所示，“预测”表格给出了因变量在 1991 年的预测值及其置信区间。

右侧的线形图描绘了实际观测序列、模型拟和序列的变化趋势，并加上了 1991 年的预测数据；由观测序列、拟和序列在图中高度相近的特点，可以判断本例使用的模型是较为合理的。

## 15.4 ARIMA 模型

ARIMA（自回归综合移动平均）是时间序列分析中最为常用的模型，也称之为 Box—Jenkins 模型，或称为带差分的自回归移动平均模型。ARIMA 模型可以对含有季节成分的时间序列数据进行分析，它包含 3 个主要的参数——自回归阶数 ( $p$ )、差分阶数 ( $d$ ) 和移动平均阶数 ( $q$ )，一般模型的形式记为 ARIMA ( $p, d, q$ )。

### 15.4.1 ARIMA 模型的基本原理

处理非平稳的时间序列时，可以先建立一个包含趋势成分的模型，对由此初步模型得到的残差项，再使用 ARIMA( $p, d, q$ )模型来拟合。

#### 1. 差分

差分是使序列平稳化的主要手段，常用的有一般性差分和季节性差分两种。

令  $y_t$  为原始时间序列， $B$  为延迟算子，于是有： $By_t = y_{t-1}$ ， $B^d y_t = y_{t-d}$ ，则一阶差分为： $\nabla y_t = (1 - B)y_t = y_t - y_{t-1}$ ； $d$  阶差分为： $\nabla^d y_t = \nabla(\nabla^{d-1} y_t) = (1 - B)^d y_t$ 。

如果  $y_t$  还是一个周期为  $T$  的序列，以  $\nabla_T$  表示季节差分算子，有： $\nabla_T y_t = y_t - y_{t-T}$ 。

这两种差分可以任意组合，直至差分后的序列为平稳的。平稳性可以通过检查差分后序列的自（偏）相关序列图来判断。对于非季节性数据，通常求一阶差分就足够了；对于周期为 12 的季节性数据，当季节效应是相加属性时，通常使用差分算子  $\nabla\nabla_{12}$ ，当季节效应是相乘属性时，通常使用差分算子  $\nabla_{12}^2$ ；对于以季度为周期的数据，通常使用差分算子  $\nabla_4$ 。

## 2. ARIMA 模型的分类

所谓 ARIMA 模型，就是对差分后的序列建立 ARMA 模型。根据参数个数的不同，ARMA 模型可分为如下几个基本类型，它们是相对简单而且被广泛研究的模型，理解了这些基本模型的原理，就能够掌握一般的 ARMA 和 ARIMA 模型。

(1) 自回归模型。自回归模型的一般形式为： $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t$ ，体现了时间序列  $x_t$  的某个时刻  $t$  和它之前  $p$  个时刻间的相互联系，其中： $\varepsilon_t$  假设为白噪声序列，且和  $t$  时刻之前的原始序列  $x_k(k < t)$  互不相关。此式称为  $p$  阶自回归模型，记为 AR( $p$ )。

AR( $p$ )模型的偏自相关函数在  $p$  阶之后应为零，称其具有截尾性；AR( $p$ )模型的自相关函数不能在某一步之后为零（截尾），而是按指数衰减（或呈正弦波形式），称其具有拖尾性。实际应用中，可以根据自（偏）相关函数的这些特征来识别 AR( $p$ )模型。

(2) 移动平均模型。移动平均模型的一般形式为： $x_t = \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$ ，其中  $\varepsilon_t$  假设为白噪声序列，说明时间序列  $x_t$  能表示为若干个白噪声的加权平均和。此式称为  $q$  阶移动平均模型，记为 MA( $q$ )。

MA( $q$ )模型的自相关函数在  $p$  阶之后应为零，称其具有截尾性；MA( $q$ )模型的偏自相关函数不能在某一步之后为零（截尾），而是按指数衰减（或呈正弦波形式），称其具有拖尾性。实际应用中，可以根据自（偏）相关函数的这些特征来识别 MA( $q$ )模型。

(3) 自回归移动平均模型。自回归移动平均模型是自回归模型与移动平均模型的综合，其一般形式为： $x_t = \phi_1 x_{t-1} + \phi_2 x_{t-2} + \cdots + \phi_p x_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$ ，其中  $\varepsilon_t$  假设为白噪声序列，且和  $t$  时刻之前的原始序列  $x_k(k < t)$  互不相关，记为 ARMA( $p, q$ )模型。

ARMA( $p, q$ )模型的自相关函数和偏自相关函数，都具有拖尾性。

(4) 关于序列相关性的总结。AR( $p$ )模型、MA( $q$ )模型都是 ARMA( $p, q$ )模型的特例，有：AR( $p$ )=ARMA( $p, 0$ )，MA( $q$ )=ARMA( $0, q$ )。如表 15-1 所示，是对各种 ARMA 模型相关函数特征的总结。

表 15-1 ARMA 模型相关函数的特征

模 型	AR( $p$ )	MA( $q$ )	ARMA( $p, q$ )
自相关函数	拖尾	截尾	拖尾
偏自相关函数	截尾	拖尾	拖尾

## 3. 建立 ARIMA 模型的一般步骤

建立 ARIMA 模型的一般步骤，可以分为如下 4 个部分。

(1) 通过差分或其他变换，使时间序列满足平稳性（stationary）的要求。

(2) 模型识别（identification），主要是利用 ACF、PACF 和 AIC 等序列估计模型的大致类型，并给出几个初步模型以待进一步验证和完善。

(3) 参数估计和模型诊断（estimation and diagnostic），对识别阶段所给初步模型的参数进行估计及假设检验，并对模型的残差序列作诊断分析，以判断模型的合理性。



(4) 预测 (forecasting)，利用最优模型对序列的未来取值或走势进行预测。

在以上步骤中，模型识别、参数估计及模型诊断的过程通常都是不断反馈、逐渐完善的过程，对模型不断实施修正的过程如图 15-24 所示。

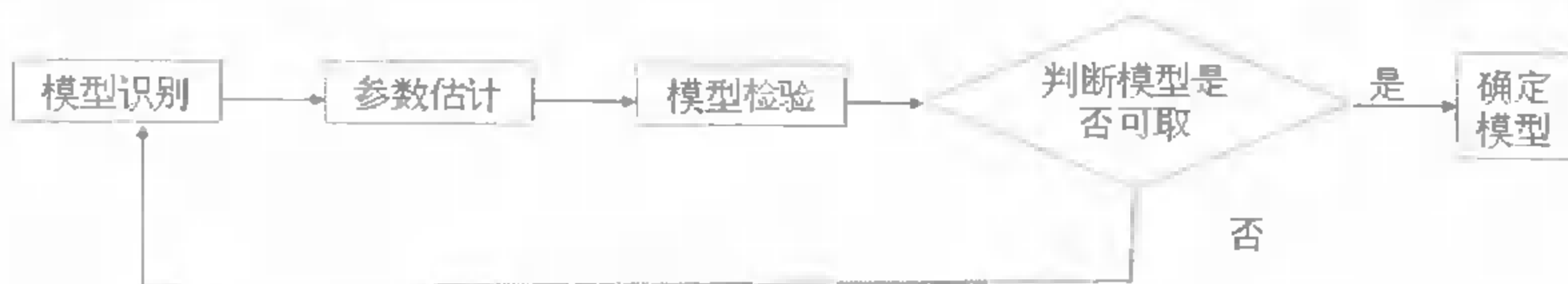


图 15-24 ARIMA 建模的一般步骤

## 15.4.2 ARIMA 模型的参数设置

在图 15-2 中，单击 Method 下拉列表并选中 ARIMA 项，再单击 Criteria 按钮，弹出如图 15-25 所示的模型参数设置面板。

### 1. 模型参数设置

(1) ARIMA Orders 栏，指定模型不同成分的阶数，确定模型的结构 (Structure)。

此栏有 6 个待设参数，分别对应于  $SARIMA(p,d,q) \times (sp,sd,sq)$  模型中的 6 个参数。在 Structure 栏的二维表格中：Nonseasonal 列从上至下的 3 个输入框分别对应于  $p$ 、 $d$ 、 $q$ ；Seasonal 列从上至下的 3 个输入框分别对应于  $sp$ 、 $sd$ 、 $sq$ ；只有定义了序列周期后，Seasonal 列的设置才会生效；二维表格底部的 Current 行，显示了当前序列数据的周期 ( $s$ )。

(2) Dependent Variable Transformation 栏，指定对因变量的变换方法，可选项有如下 3 个。None，不作变换；Square root，开平方变换；Natural log，自然对数变换。

(3) Include constant in model 复选框，选中表示在 ARIMA 模型中包括常数项。

当确信时间序列数据的均值为零，或者已经对其应用了差分算子，就建议在模型中不包括常数项。

### 2. 异常值检测选项的设置

在图 15-25 中，单击 Outliers 标签，显示如图 15-26 所示的子设置界面，在此设置关于异常值检测的选项。

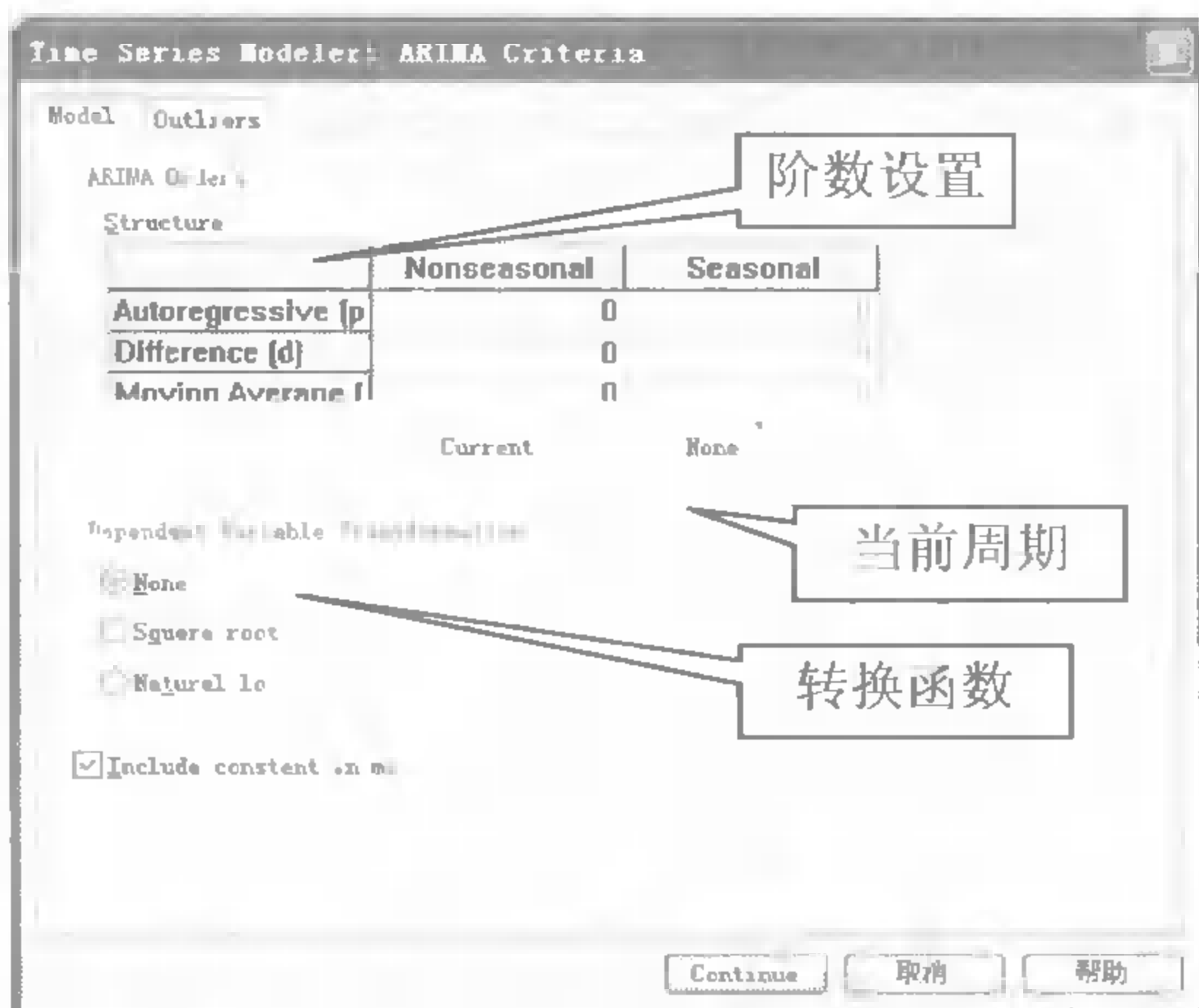


图 15-25 ARIMA 方法的 Model 设置

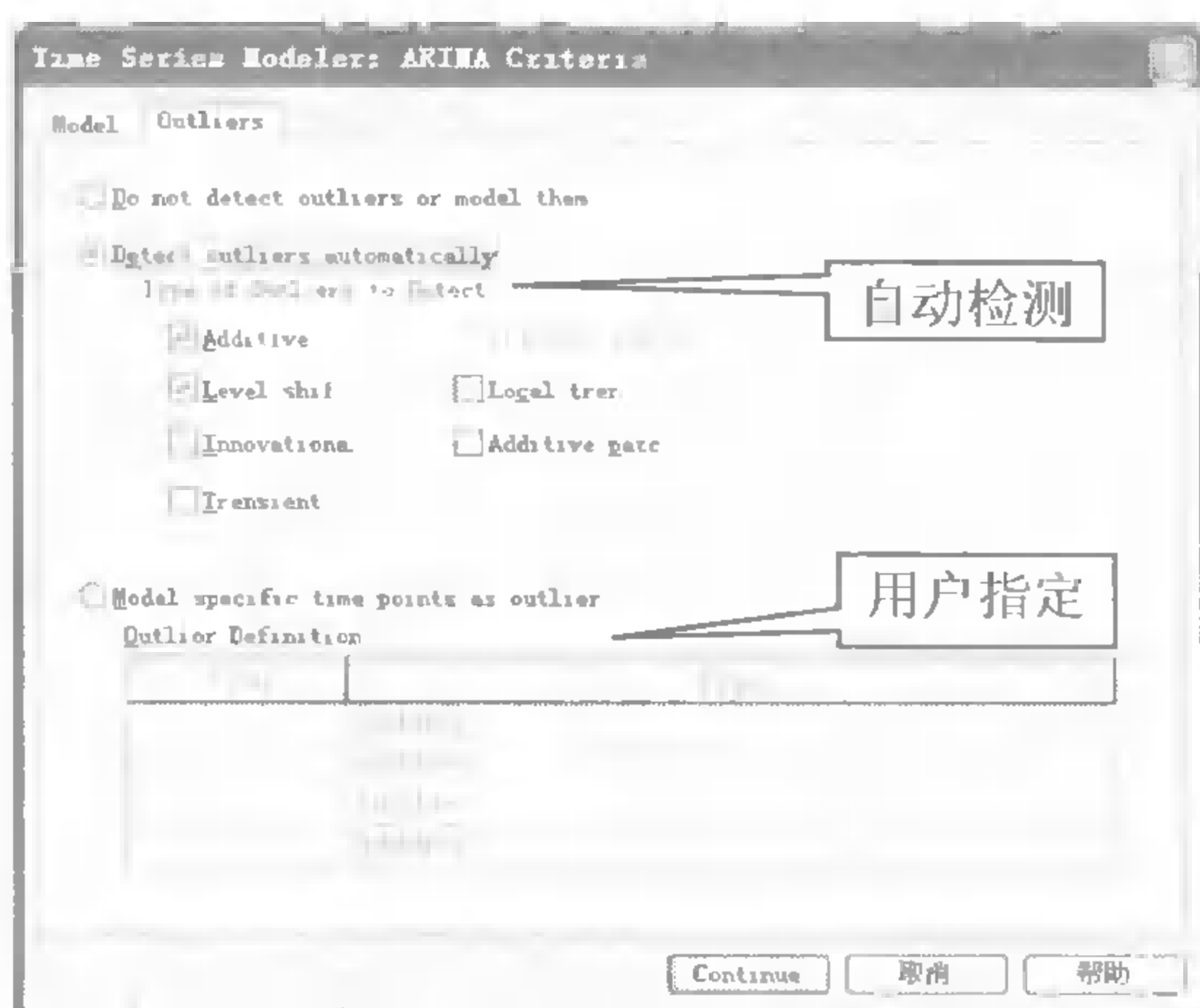


图 15-26 ARIMA 方法的异常值检测设置

(1) Do not detect outliers or model them 单选项，表示不检测异常值。

(2) Detect outliers automatically 单选项，指定自动检测异常值的方法，可选项有如下几个。

- Additive: 表示只影响单个观测记录的异常值，例如：数据编码错误导致的异常值。
- Level shift: 由数据的水平移动引起的异常值。政府政策的变化可能导致这种异常值。
- Innovational: 由于噪声变动形成的异常值。对于平稳过程，它会影响多个观测记录；对于非平稳过程，它可能影响某个时刻之后的所有观测。
- Transient: 对后续观测的影响程度，按指数水平衰减至 0 的异常值。
- Seasonal additive: 周期性的影响某些时刻的异常值，且影响程度对不同时刻的观测是相同的。例如：从某一年开始，每年 1 月份的销售额都异常地高。
- Local trend: 局部的线性异常值，表示某点之后的序列开始出现明显的趋势成分。
- Additive patch: 表示两个或多个连续出现的 Additive 类型的异常值。

(3) Model specific time points as outliers 单选项，设置特定时刻的数据为异常值。

选中此单选框后，在 Outlier Definition 下的二维表格中每行指定一个特定的异常数据，在第一列 (Year) 输入时间点，在第二列 (Type) 从下拉列表选择异常点的类型。如果当前序列中没有定义时间变量，此处只显示一列，用于输入数据异常点在 Data Editor 窗口的行号。

### 15.4.3 ARIMA 模型实例分析

#### 1. 数据和问题描述

本节，仍然使用 1950~1990 年的天津食品消费数据，分析这段时间内的人均生活费年收入的变化情况，数据格式如图 15-15 所示，所用数据文件为“天津食品消费相关数据.sav”。

#### 2. 参数设置


依次单击菜单“Analyze→Time Series→Create Models...”，打开建立模型的主设置面板，如图 15-27 所示。在 Variables 列表单击选中人均生活费年收入，单击从上至下第一个  按钮，将其作为因变量选入 Dependent Variables 列表框；单击 Method 下拉列表，选中 ARIMA 选项。



图 15-27 ARIMA 模型的主设置界面

在图 15-27 中,单击 Criteria 按钮,弹出如图 15-28 所示的子设置面板。在 Nonseasonal 列从上至下的 3 个单元格依次输入 1、1、2,表示模型结构是 ARIMA(1,1,2);单击选中 Natural log 单选框;单击取消选中 Include constant in model 复选框;单击 Continue 按钮返回主界面。

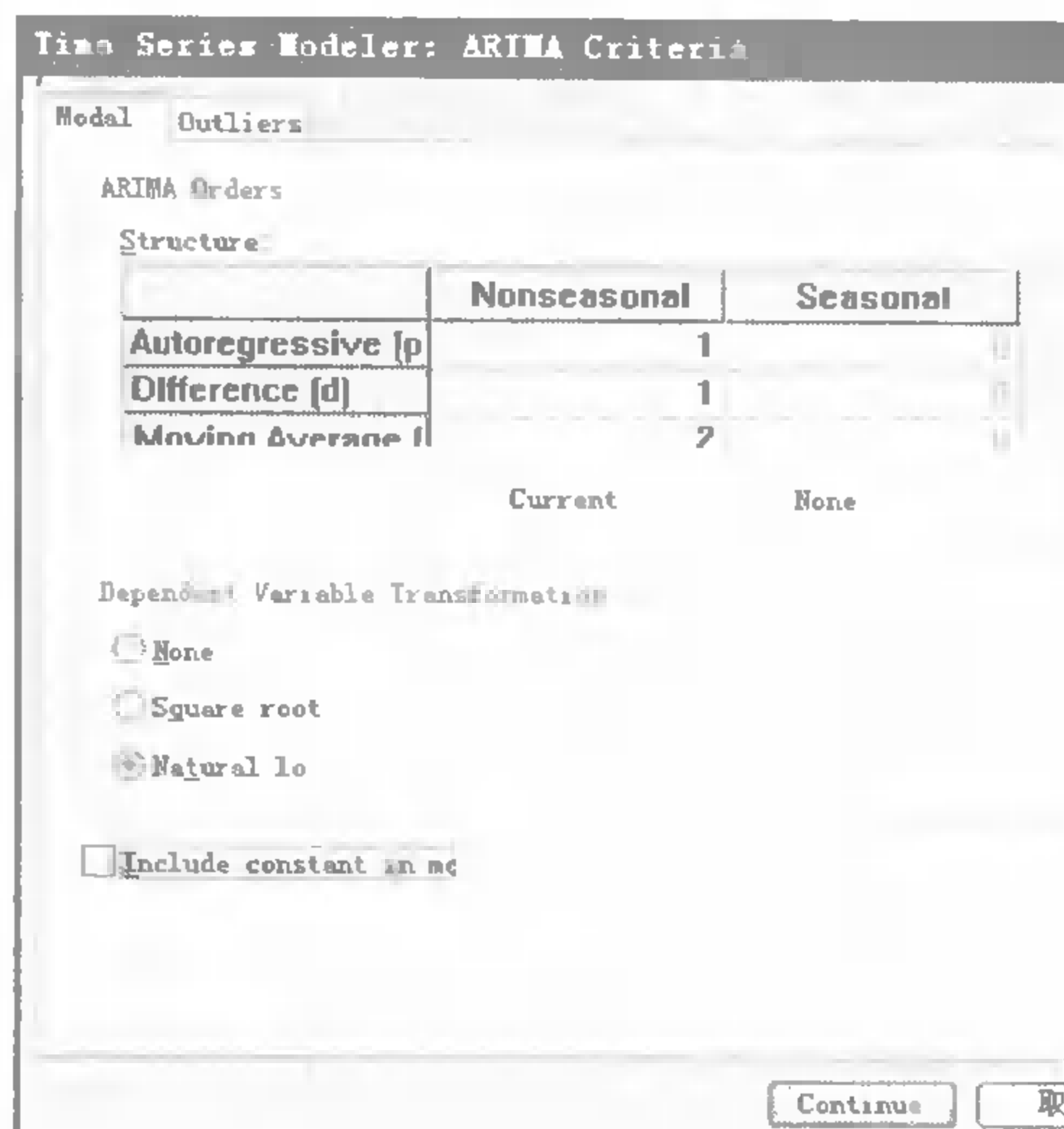


图 15-28 模型参数的设置

在图 15-27 中,单击 Statistics 标签,弹出如图 15-3 所示的子设置界面,勾选 Parameter estimates 复选框;单击 Variables 标签返回图 15-27 所示的主设置界面。

在图 15-27 中,单击 Plots 标签,弹出如图 15-4 所示的子设置界面,分别勾选 ACF、PACF 复选框;单击 Variables 标签返回图 15-27 所示的主设置界面。

在图 15-27 中,单击 Options 标签,打开如图 15-7 所示的选项设置界面。单击选中 First case after end of estimation period through a specified date 单选框;在 Year 下面的单元格输入“1991”;单击 Variables 标签返回图 15-27 所示的变量设置界面。

### 3. 结果分析

在图 15-27 中单击“确定”按钮运行,SPSS Viewer 窗口的输出结果如下。

(1) 模型描述和模型拟和优度统计量。如图 15-29 所示,“模型描述”表格给出了当前模型所使用的分析变量和方法;“模型拟和”表格,给出了包括平稳 R 方在内的 8 个拟合优度统计量。

模型描述											
模型 ID						模型类型					
人均生活费年收入						ARIMA(1,1,2)					

模型拟合											
拟合统计量	均值	SE	最小值	最大值	百分点						
					5	10	25	50	75	90	95
平稳 R 方	.296		.296	.296	.296	.296	.296	.296	.296	.296	.296
R 方	.994		.994	.994	.994	.994	.994	.994	.994	.994	.994
RMSE	27.001		27.001	27.001	27.001	27.001	27.001	27.001	27.001	27.001	27.001
MAPE	3.632		3.632	3.632	3.632	3.632	3.632	3.632	3.632	3.632	3.632
MaxAPE	14.126		14.126	14.126	14.126	14.126	14.126	14.126	14.126	14.126	14.126
MAE	15.899		15.899	15.899	15.899	15.899	15.899	15.899	15.899	15.899	15.899
MaxAE	79.995		79.995	79.995	79.995	79.995	79.995	79.995	79.995	79.995	79.995
正态化的 BIC	6.868		6.868	6.868	6.868	6.868	6.868	6.868	6.868	6.868	6.868

图 15-29 ARIMA 的模型基本信息和模型拟合优度输出

(2) 模型参数输出。如图 15-30 所示, 是 ARIMA(1,1,2)模型的参数估计结果, 从  $t$  统计量的显著性 (Sig 列) 可以看出, 此模型的一阶自回归系数很不显著, 所以有必要对模型结构进行改进, 去掉自回归部分的影响。

ARIMA 模型参数					估计	SE	t	Sig
人均生活费 年收入 -Model_1	人均生活 费年 收入	自然 对数	AR	滞后 1	.413	.202	2.044	.048
			差分		1			
			MA	滞后 1	-.376	.177	-2.123	.041
				滞后 2	-.607	.153	-3.962	.000

图 15-30 模型参数输出

在图 15-28 中, 指定模型结构为 ARIMA(0,1,2), 设置方法同前; 单击 Continue 按钮返回主界面。再次在图 15-27 中单击“确定”按钮运行, 改进模型的参数估计输出如图 15-31 所示, 可见所有参数都显著地不为 0 了。

ARIMA 模型参数					估计	SE	t	Sig
人均生活费年 收入-Model_2	人均生活 费年收入	自然 对数	差分		1			
			MA	滞后 1	-.840	.123	-5.221	.000
				滞后 2	-.895	.133	-5.233	.000

图 15-31 改进模型的参数输出

(3) 改进模型的残差序列图。如图 15-32 所示, 是关于残差序列的自相关 (ACF) 图形和偏自相关图形 (PACF)。可见两个图形都没有显著的趋势特征 (拖尾或截尾), 故可以初步判断这个改进模型是比较恰当的。

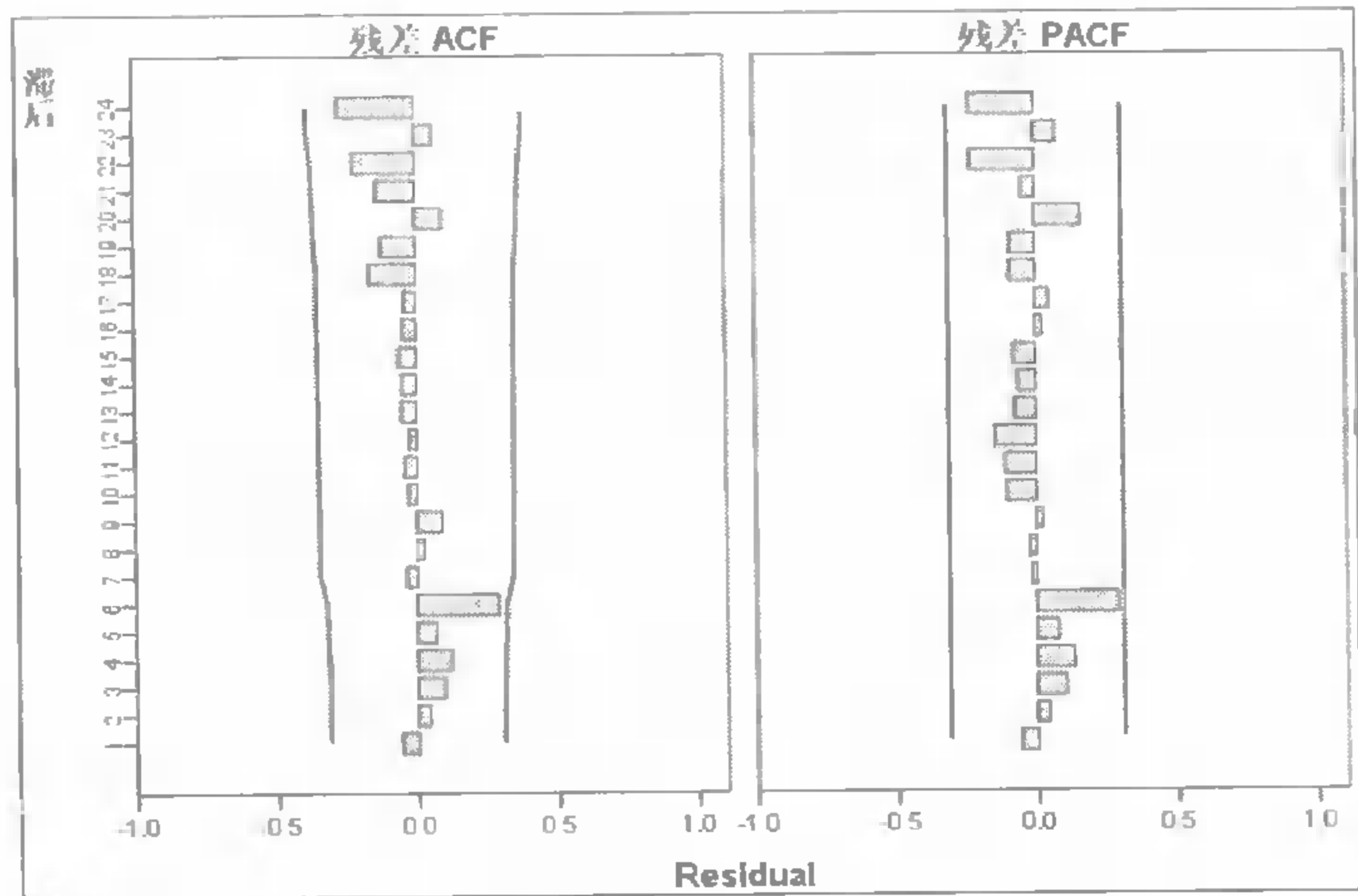


图 15-32 ARIMA 模型残差的相关函数图

(4) 改进模型的预测结果。如图 15-33 所示, 是使用 ARIMA(0,1,2)模型对序列在 1991 年的预测结果。“预测”表格给出了 1991 年的预测值及其置信区间。右侧的线形图描绘了实际观测序列、模型拟合序列的变化趋势, 并加上了 1991 年的预测数据, 从两条曲线高度接近可以推断改进模型较为合理。



预测		
模型		1991
人均生活费年	预测	1622.79
收入_Model_1	UCL	1805.06
	LCL	1454.80

对于每个模型，预测都在请求的预测时间段范围内的最后一个非缺失值之后开始，在所有预测值的非缺失值都可用的最后一个时间段或请求预测时间段的结束日期（以较早者为准）结束。

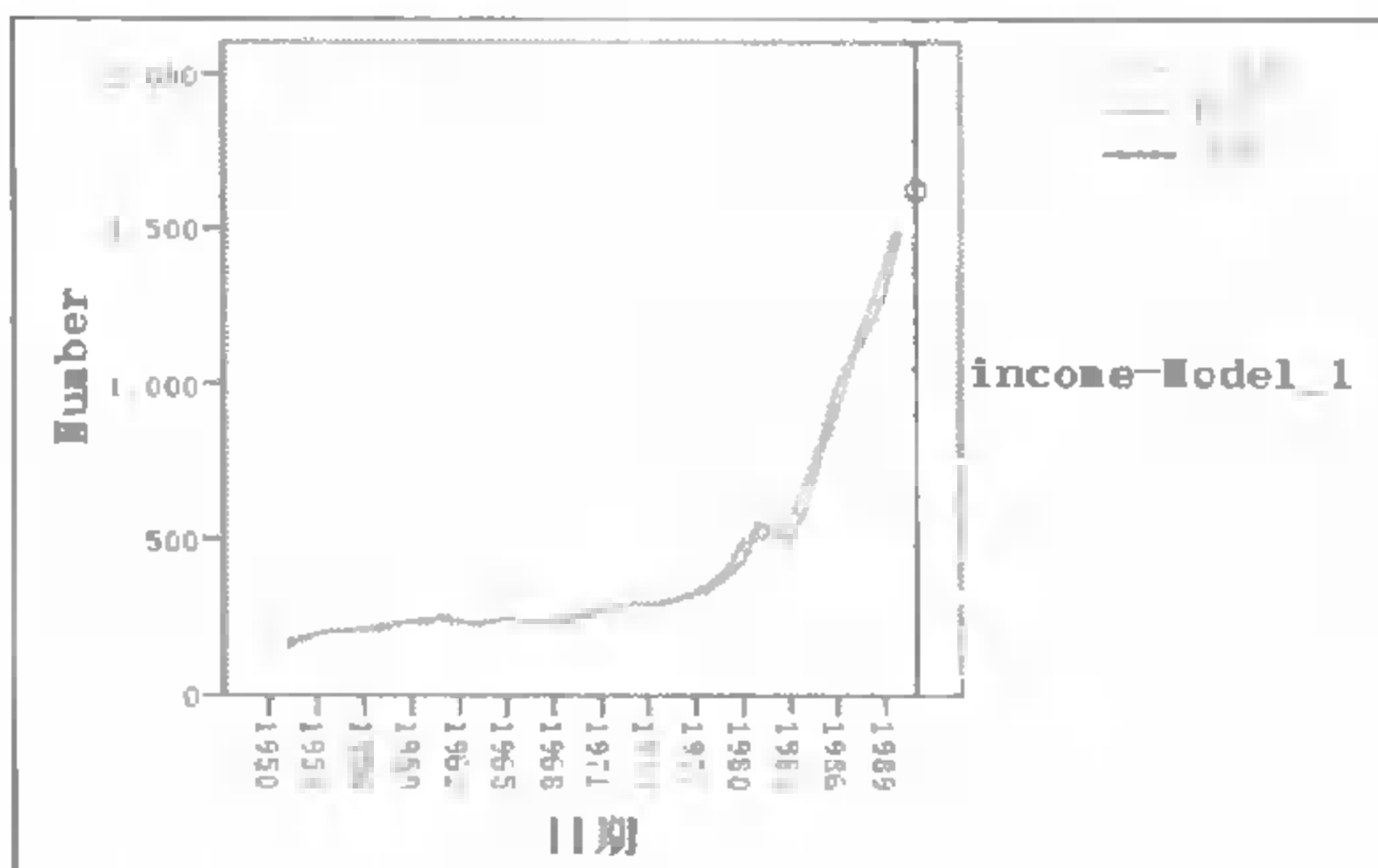


图 15-33 ARIMA 模型预测结果和拟合图形

## 15.5 季节分解模型

时间序列是对某一统计指标，按照指定的时间间隔，搜集整理的一组统计数据。一般认为，一个时间序列可能包含 4 种变动因素：长期趋势变动、季节性变动、循环性变动和不规则变动。但并不是所有的时间序列都会同时含有这 4 种变动因素，例如：年份统计表就不存在季节性变动因素，按照季度统计的数据不一定存在循环变动因素。

### 15.5.1 季节分解法概述

所谓季节分解，就是通过某些手段把时间序列中的 4 种变动趋势分解出来，并分别对其加以分析，再将分析结果综合起来组成一个对原始时间序列的总模型。

#### 1. 时间序列的 4 种成分

(1) 长期趋势 (Long term trend)，记为：T。长期趋势，表示序列取值随时间逐渐增加、减少或不变的长期发展趋势。

例如：全球人口总数随着时间推移，正在逐步增长；人口死亡率，由于医疗技术的进步及生活水平的提高，出现了长期向下的趋势。另外，同一序列在不同时期可能表现出不同的长期趋势，例如：某商品的销量，在产品初期具有向上趋势；在产品成长期有加速向上的趋势；在产品成熟期表现出缓慢增长的趋势；在产品末期呈向下的趋势。

(2) 季节趋势 (Seasonal component)，记为：S。季节趋势，表示由于受到季节因素或某些习俗的影响，而出现的有规则的变化规律。

例如：电风扇和空调的销售量，在夏季多而冬季少；每一天的交通流量，在上下班时间出现高峰，其余时间则较为稳定；圣诞节之前，玩具的销售量总会增加等。

(3) 循环趋势 (Cyclical component)，记为：C。循环趋势，表示序列取值沿着趋势线有如钟摆般循环变动的规律。

循环趋势的周期长短和波动幅度是主要的研究对象。有时一个时间序列的循环是由多个小循环组合而成的，例如：总体经济指标的循环，就是由各个产业的循环组合而成。

(4) 不规则趋势 (Irregular component)，记为 I。不规则趋势，表示把时间序列中的长期趋势、季节趋势和循环趋势都去除后余下的部分。

一般而言，长期趋势、季节趋势和循环趋势都受到规则性因素的影响，只有不规则趋势是随机性的，它发生的原因有：自然灾害、天气突变、人为的意外因素等。

## 2. 季节分解模型的种类

对于时间序列中各变动因素之间的关系，通常有两种不同的假设：加法关系假设和乘法关系假设，相应地就有了时间序列季节分解的加法模型和乘法模型。

(1) 加法模型。加法模型假设：时间序列是由 4 种成分相加而成的；各成分之间彼此独立，没有交互影响。如果以  $Y$  表示某个时间序列，它的加法模型就为： $Y=T+C+S+R$ 。

按照加法模型的假设，季节因素、周期因素和不规则因素都围绕着长期趋势而上下波动，它们可以表现为正值或负值，反映了各自对时间序列的影响方式和程度。

(2) 乘法模型。乘法模型假设：时间序列是由 4 种成份相乘而成的；各成分之间存在着相互依赖的关系。如果以  $Y$  表示某个时间序列，它的乘法模型就为： $Y=T \times C \times S \times R$ 。

按照乘法模型的假设，季节因素、周期因素和不规则因素也围绕着长期趋势而上下波动，但这种波动表现为一个大于或小于 1 的系数，反映它们在长期趋势的基础上对原始序列的相对影响方式和程度。

### 15.5.2 季节分解模型实例分析

#### 1. 数据和问题描述

(1) 数据文件。本节利用季节分解模型，对某城市 5 年内每个季度的游客数量进行分析，以了解其旅游市场的发展变化规律。所用数据文件为“某市游客量时序数据.sav”，数据格式如图 15-34 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	year	Numeric	8	0	年份	None	None	8	Right	Scale
2	quart	Numeric	8	0	季度	None	None	7	Right	Scale
3	num	Numeric	8	2	游客量	None	None	8	Right	Scale
4	YEAR_	Numeric	6	0	YEAR not periodic	None	None	10	Right	Ordinal
5	QUARTER_	Numeric	1	0	QUARTER period 4	None	None	10	Right	Ordinal
6	DATE_	String	7		Date Format "QQ"	None	None	9	Left	Nominal

图 15-34 游客数量时序数据格式

(2) 查看当前日期变量。依次单击菜单“Data→Define Dates...”，打开定义时间变量的对话框，如图 15-35 所示。单击 Cancel 按钮返回 Data Editor 窗口。

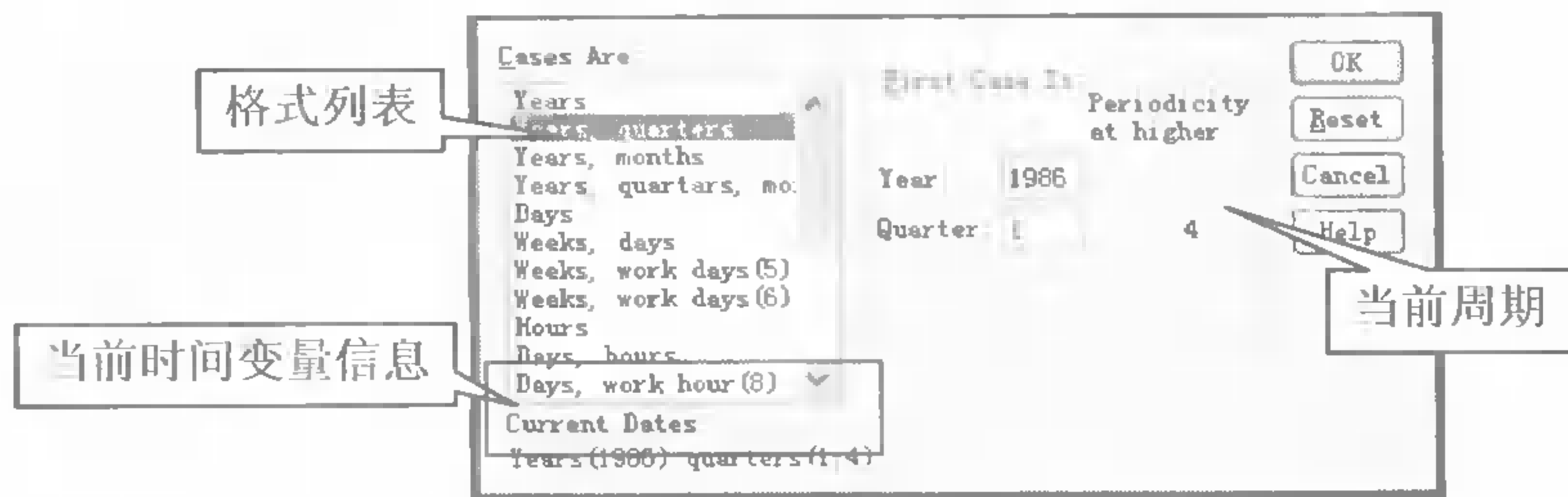


图 15-35 游客数据定义时间序列的设置面板

在 Current Dates 子设置栏，显示了当前时间变量的日期格式为“Years(1986) quarters(1:4)”，表示观测的起始时间为 1986 年第 1 季度，序列周期为 4。

#### 2. 参数设置

依次单击菜单“Analyze→Time Series→Seasonal Decomposition...”，执行时间序列季节分解

的功能。如果当前数据集没有定义周期性的时间成分变量，会弹出如图 15-36 所示的提示对话框，单击确定按钮后，指定有周期的日期格式。正常情况下，进入图 15-37 所示的主设置界面。

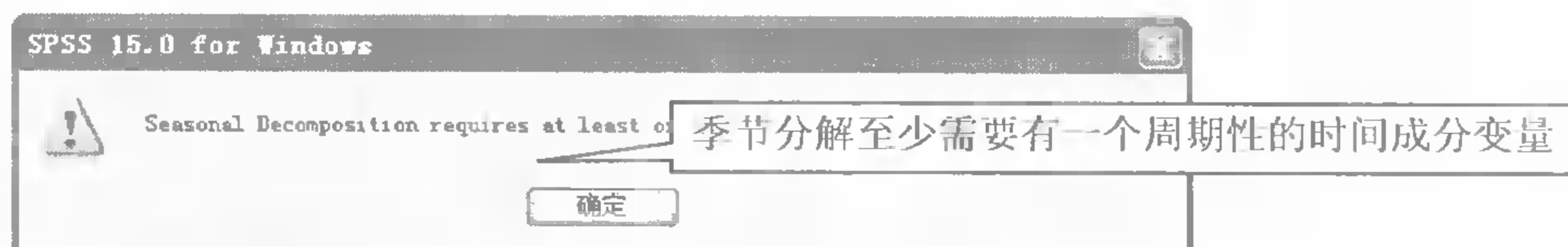


图 15-36 提示定义时间序列对话框

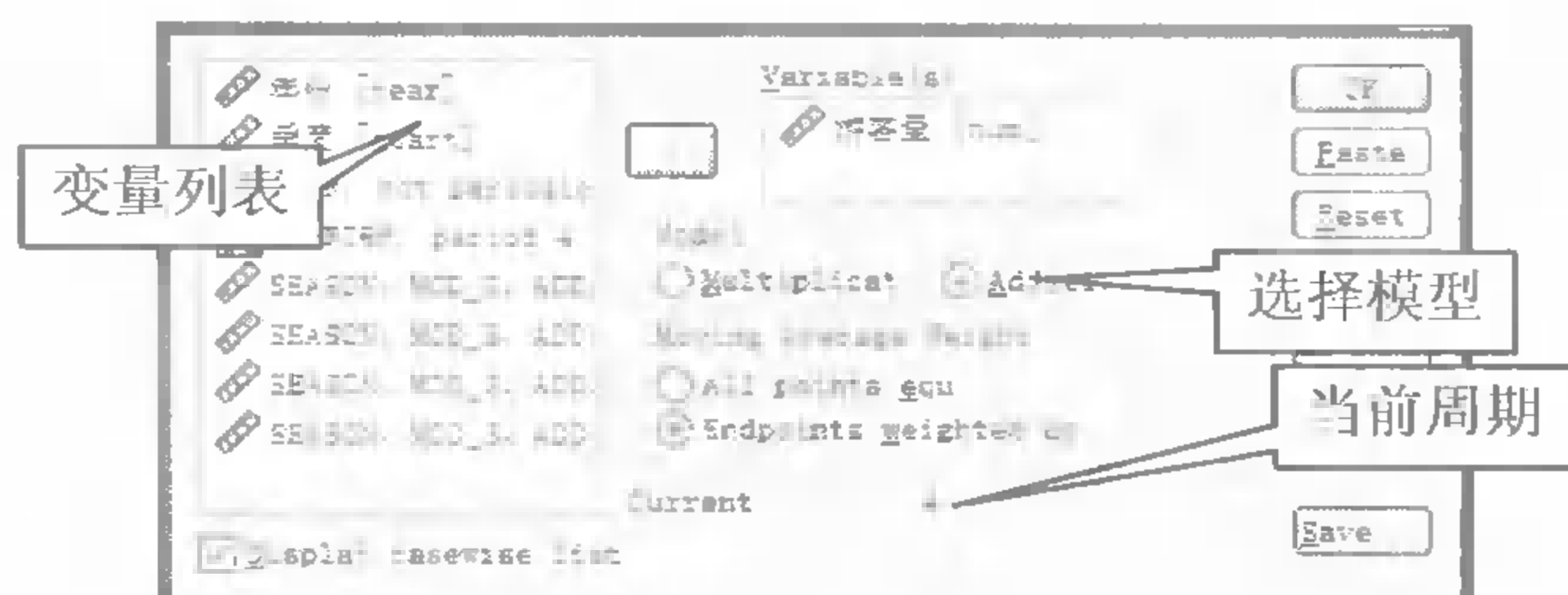


图 15-37 季节分解的主设置面板

在变量列表单击选中游客量，单击 ☐ 按钮，将其作为时序变量选入 Variable(s)列表框；单击选中 Additive 单选框；单击选中 Endpoints weighted by .5 单选框；勾选 Display 复选框。

- (1) Variable(s)列表框，用于选入要进行季节分解的原始序列变量。
- (2) Model 栏，用于指定季节分解的模型类型，可选项有如下两个。  
Multiplicative 单选框，乘法模型；Additive 单选框，加法模型。
- (3) Moving Average Weight 栏，指定计算移动平均时的权重，有两个可选项。  
☒ All points equal，表示等值权重，一般用于周期是奇数的情形。  
☒ Endpoints weighted by .5，表示端点权重为 0.5，一般用于周期是偶数时的情形。
- (4) Current 栏，显示当前序列数据的周期，不可编辑。
- (5) Display casewise listing 复选框，表示输出对每个观测量的季节分解结果。
- (6) 保存选项的设置

单击 Save 按钮，打开如图 15-38 所示的子对话框，设置关于保存选项的参数。

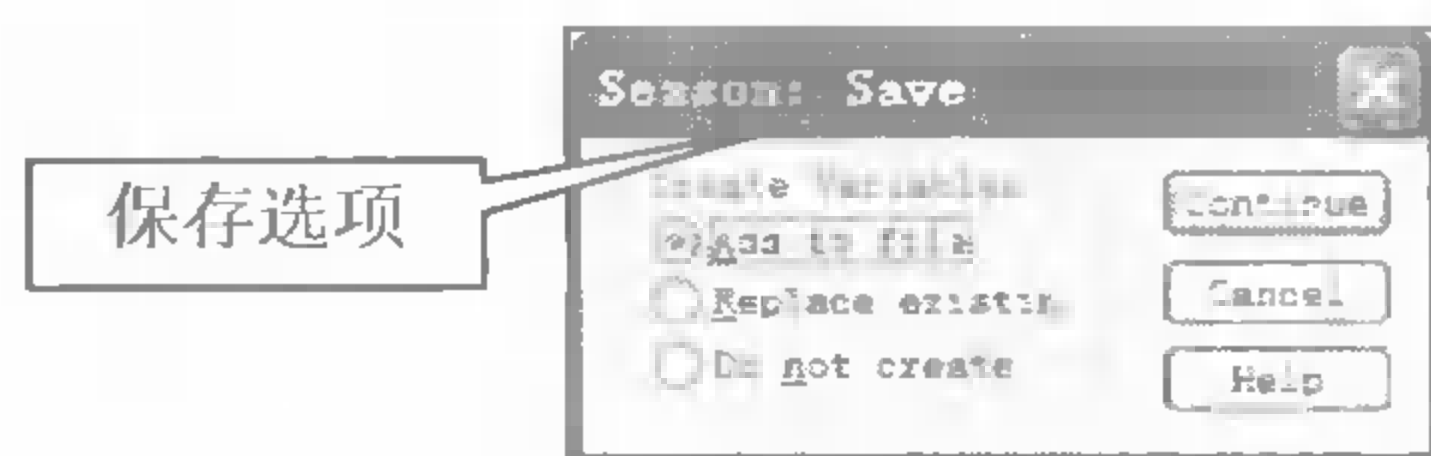


图 15-38 季节分解的保存设置

- ☒ Add to file 单选框，作为永久新增变量添加到当前数据集里。
- ☒ Replace existing 单选框，作为临时新增变量添加到当前数据集里，新模型的输出值将覆盖旧模型保存的变量。
- ☒ Do not create 单选框，不在当前数据集保存模型结果。

季节分解模型将会输出 4 个结果变量，它们的名称（以 3 个字母表示）及含义分别为：SAF，表示序列的季节成分；SAS，表示去除季节成分后的序列；STC，表示序列的趋势和循环成分；ERR，表示序列的不规则成分（随机部分）。

### 3. 结果分析



在图 15-37 中, 单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 15-39 所示。

季节性分解							
序列名称 游客量							
DATE	原始序列	移动平均数序列	原始序列与移动平均数序列的差分	季节性因素	季节性调整序列	平滑的趋势循环序列	不规则(误差)分量
Q1 1986	121.500			6.5547	94.445	96.404	-1.66319
Q2 1986	118.000			19.660	98.340	96.674	1.66581
Q3 1986	91.000	87.8125	-3.18750	-7.258	97.258	97.906	-1.94861
Q4 1986	74.500	94.4125	-19.91250	-15.96	87.977	94.235	-1.24306
Q1 1987	128.000	100.000	28.0000	8.5547	101.445	100.685	0.76515
Q2 1987	213.000	149.3125	63.68750	19.660	193.340	191.712	1.61288
Q3 1987	94.000	102.250	-8.25000	-7.258	101.258	102.958	-1.75583
Q4 1987	83.000	102.3000	-19.30000	-18.96	101.277	102.442	-1.48579
Q1 1988	109.000	113.000	-4.00000	6.5547	102.445	102.458	-0.0136
Q2 1988	129.000	103.6250	25.37500	19.660	105.340	101.313	4.04684
Q3 1988	98.000	104.5000	-6.50000	-7.258	103.258	104.162	-1.10417
Q4 1988	58.000	102.7125	-44.71250	-18.96	101.077	103.664	-1.48734
Q1 1989	117.000	107.2500	9.75000	6.5547	108.445	107.163	1.28338
Q2 1989	131.000	108.4250	22.57500	19.660	111.340	108.934	2.40066
Q3 1989	103.000	119.0000	-16.00000	-7.258	109.258	109.518	-0.51718
Q4 1989	91.000	118.7100	-27.71000	-18.96	109.977	110.684	-0.68479
Q1 1990	113.000	115.5000	-2.50000	6.5547	108.445	111.161	-2.71178
Q2 1990	129.000	120.0000	9.00000	19.660	111.340	119.321	-8.02117
Q3 1990	140.000			-7.258	147.258	134.872	22.38628
Q4 1990	97.000			18.96	115.977	127.315	-31.31524

图 15-39 季节分解模型的输出结果

(1) 模型基本统计信息。如图 15-39 所示, “模型描述” 表格给出了当前模型所使用的分析变量和模型参数。

(2) 季节分解的结果。如图 15-39 所示, “季节性分解” 表格给出了对每个观测的季节分解结果, 各成分的含义如图中的表头所示。这部分分解结果也保存在了当前数据集中。

(3) 作图分析。依次单击菜单 “Analyze→Time Series→Sequence Charts...”, 打开建立普通序列图的设置界面, 如图 15-40 所示。在变量列表选中游客量、SAS\_1 和 STC\_1 三个变量, 单击从上至下第一个  按钮, 将其作为作图变量选入 Variables 列表框; 在变量列表单击选中 Date 变量, 单击从上至下第二个  按钮, 将其作为时间轴变量选入 Time 选框。

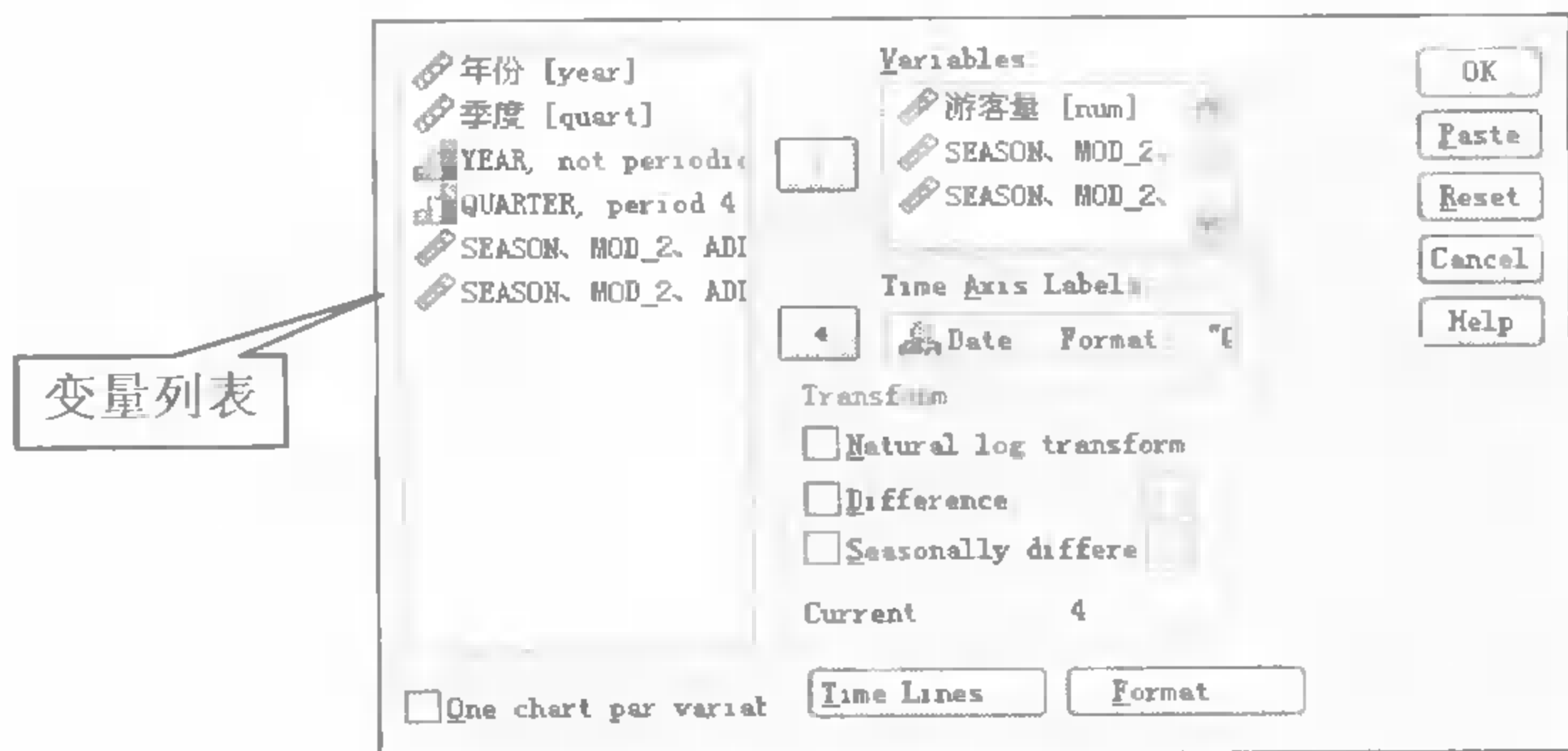


图 15-40 作普通序列图的参数设置

在图 15-40 中, 单击 OK 按钮运行, SPSS Viewer 窗口的输出图形如图 15-41 所示, 它在一个图里描绘了原始序列 (蓝色)、趋势循环序列 (红色) 和季节调整序列 (绿色) 的趋势线, 其中去除季节和误差因素后的趋势循环序列表现出明显的趋势性, 可用回归分析等其他分析方法加以研究。



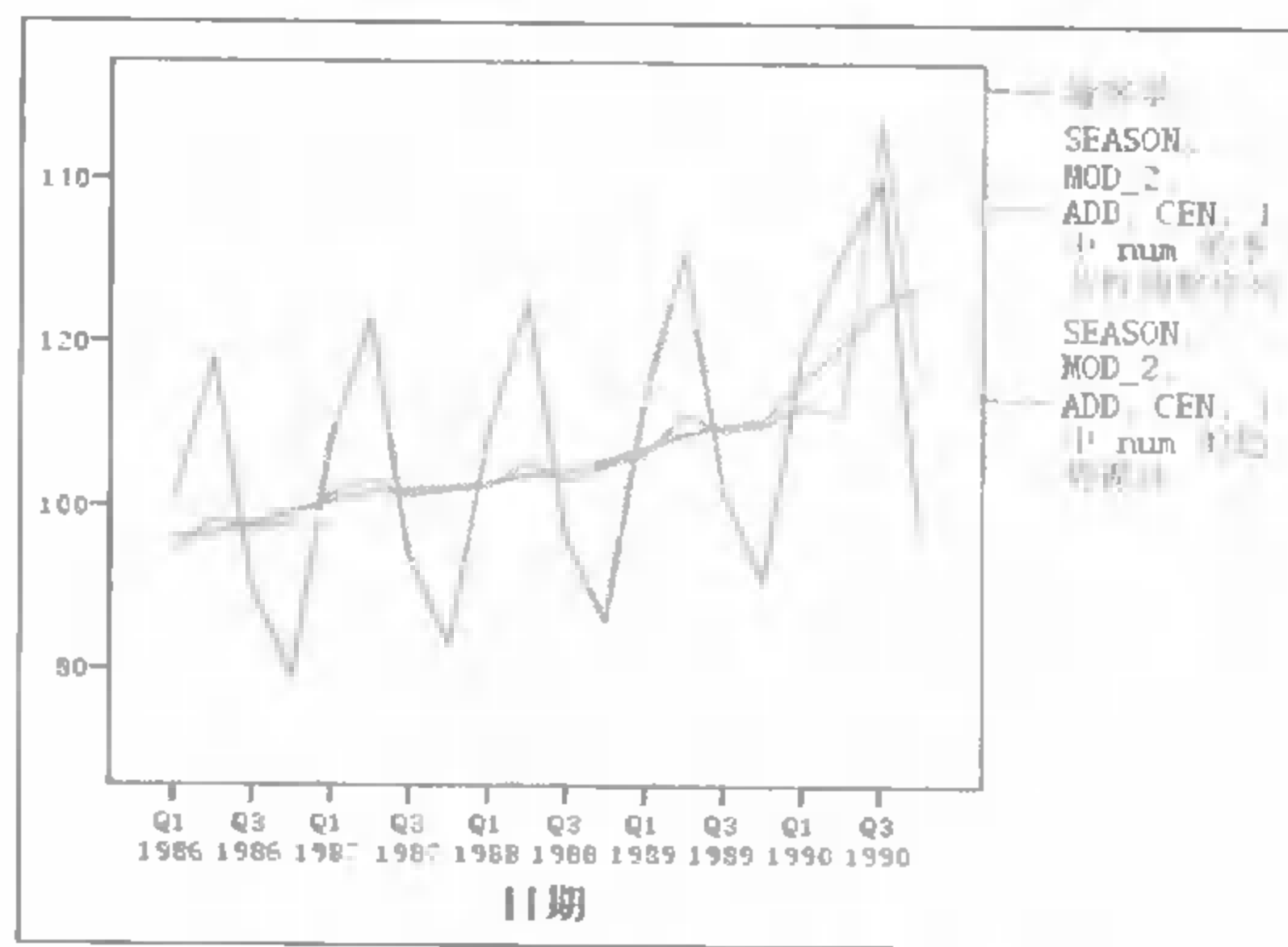


图 15-41 3 条序列的图形显示

# 第 16 章 对数线性模型

对数线性 (Log-linear) 模型, 是一种适用于离散型数据或整理成列联表格式的计数资料的统计分析工具, 它在分析中把用于分类的因素作为自变量, 把列联表单元格中的频数作为因变量。对于列联表资料, 通常作卡方检验, 但卡方检验不能系统地分析变量之间的联系, 也无法估计变量间相互作用的大小, 而对数线性模型正是处理这些问题的最佳方法。

## 16.1 对数线性模型概述

对数线性模型与方差分析模型、线性模型有一定的相似性, 它将概率 (或理论频数) 取对数后分解成主效应和因素之间的交互效应, 并通过交互效应来反映各变量之间的关系。

### 16.1.1 简单列联表分析的不足

对于属性数据, 经常使用列联表来反映变量之间的联合分布, 当只有两个变量 (因素) 时, 称为二维列联表; 三维或更高维的列联表也称为多维列联表。在列联表中, 频数分布受两种效应的影响: 一种反映了某个因素自身的效应, 称之为主效应; 另一种反映了由于因素之间的关联性所产生的效应, 称之为交互效应。对于二维列联表, 其主效应有两个, 交互效应只有一个; 当属性变量的数目增加时, 交互效应的维数也会快速增加, 就相当于有了多张二维列联表; 而当属性变量的取值水平增加时, 每一张二维列联表也会变大。

一般的列联表统计方法, 通常只分析两个变量之间的联系, 例如: 关于受教育程度与生活满意度的列联表, 可以直接从表中读取相应的主效应和交互效应。当做多个变量的属性分析时, 这种方法就无法把握诸多变量之间的关系了, 这时可以一次只分析两个变量之间的交互表, 经过多次两两交互分析, 再将各自的结论拼接成反映多变量之间复杂关系的整体。尽管这种做法能得到一些信息, 然而正如多个简单回归不能代替多元回归一样, 这种缺乏综合性的分析方式是不能以多个个别分析叠加出整体的多元联系的。另外, 如果整个频数分布被分成多张二维交互表, 就只能大致地分析每一张二维表的主效应和交互效应, 而多变量之间的联合交互效应 (或高阶交互作用) 将无法分析, 但有时只有这些高阶交互效应才能真正反映变量之间的关联。Log-linear 模型, 是一种更有效的处理列联表信息的统计方法。

### 16.1.2 对数线性模型的基本形式

假设有 3 个离散变量 ( $A$ 、 $B$ 、 $Y$ ), 由这 3 个因素构成的饱和对数线性模型为:  $\ln m_{ijk} = \lambda + \lambda_i^A + \lambda_j^B + \lambda_k^Y + \lambda_{ij}^{AB} + \lambda_{jk}^{BY} + \lambda_{ik}^{AY} + \lambda_{ijk}^{ABY}$ , 其中  $m$  表示期望频数,  $\lambda_i^A \dots \lambda_{ij}^{AB} \dots \lambda_{ijk}^{ABY}$  分别表示主效应、二维交互效应、三维交互效应。

在饱和对数线性模型中，主效应的大小只代表相应的变量对期望频数的贡献，但它仅仅反映了各变量行或列的边缘合计之间的差别，并不能反映与其它变量的关系。因此，分析主效应的大小不能反映变量之间的关系，变量关系主要通过各交互效应体现。

实际上，饱和对数线性模型使用的情况更多。给定一组分类变量，会有多种不饱和模型可用，那么如何选择模型就是一个问题。选择方法不仅要有说服力，而且还要尽量简单，不应含有无意义的高阶交互作用，常用的模型选择策略有如下 3 种。

(1) 先建立饱和模型，再检查每个系数的标准值 ( $z$  值) 或可信区间，消去无意义的效应。

(2) 日后淘汰法：一开始就把所有效应包含到模型中，逐步从检验概率大于临界值的效应中，淘汰拟合优度变化最小的效应。

(3) 逐一加入法：系统地检查各效应对模型的“贡献”，例如：先建立包含主效应和二阶交互效应的模型，然后建立只有主效应的模型，这两种模型的似然比之差，就是二阶交互效应对模型的贡献；通过检验拟合优度有无差异，就可以推断二阶交互效应能否被去除。

无论采用以上哪种策略，对数线性模型的约束条件都是相同的，即对任何一个脚标求和都得 0，如果一个低阶的交互效应为 0，则包含它的更高阶交互效应全部为 0；当某个高阶交互作用有统计学意义时，即使它包含的低阶效应不显著，也应将其保留在模型里。

SPSS 的对数线性分析功能包括如下 3 个模块：General Loglinear Analysis (广义对数线性模型过程)、Logit Loglinear Analysis (对数线性模型过程)、Model Selection Loglinear Analysis (模型选择对数线性分析过程)。

## 16.2 General 过程

General 过程用于一般对数线性模型的分析，适用于带有证实性的研究，即事先已经对模型有了一定的先验信息，或已知了关于模型的某些假设。这种情况下，一般只对某些特定的效应项感兴趣，利用一般对数线性模型能够检验那些假设是否正确、充分，它可以为模型总体和单个参数给出比较详细的检验结果。

对数线性模型还有一个特点，就是没有具体区分哪些是因变量，哪些是自变量，所有变量在分析中一视同仁，通过研究输出结果由用户对此做出判断。

### 16.2.1 General 过程概述

一般对数线性分析 (General Log-linear Analysis) 过程，用于研究列联表中观察对象的频数统计信息。在列联表里，分类变量称为因子 (或因素)，表格中每个行列的交叉点构成一个单元格，在此记录对应观测的例数 (频数)。

General 过程输出的统计量与图形包括：观测频数、期望频数、初始残差 (raw residual)、调整残差 (adjusted residual) 及偏差残差 (deviance residual)，设计矩阵 (design matrix)，参数估计值 (parameter estimate)，发生比 (odds ratio)，对数发生比 (log-odds ratio)，Wald 统计量 (Wald statistic)，调整残差图，偏差残差图，及正态概率图等。

### 16.2.2 问题描述和数据准备

某杂志出版社每月都会向公司“购买数据库”中的所有用户发邮件进行促销，但是反馈率很低，出版社为了提高杂志定购的反馈率，假设定购报纸的用户更倾向于定购杂志，于是只需向定购报纸的用户发送邮件。本节就利用“定购报纸”和“是否反馈”为分析变量建立对数线性模型，来检验这种假设是否可取。所用数据摘自 SPSS 自带的 Demo 文件“demo.sav”，数据文件为“报纸订阅调查数据.sav”，数据格式如图 16-1 所示。

本例所关心的 2 个问题中，“报纸订阅”取值 1 表示用户定购报纸，取值 0 表示用户不定购报纸；“是否反馈”取值为 1 表示反馈，取值为 0 表示未反馈。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	news	Numeric	4	0	报纸订阅	{0, 1}...	None	8	Right	Scale
2	response	Numeric	4	0	是否反馈	{0, 1}...	None	8	Right	Scale
3	inccat	Numeric	8	2	收入	{1.00, Under	None	8	Right	Ordinal
4	age	Numeric	4	0	年龄	None	None	8	Right	Scale
5	marital	Numeric	4	0	婚否	{0, Unmarried	None	8	Right	Scale

图 16-1 杂志订阅的调查数据格式

### 16.2.3 General 过程的参数设置

依次单击菜单“Analyze→Loglinear→General...”打开 General 过程的主设置面板，如图 16-2 所示，在此选择进行分析的各种变量。

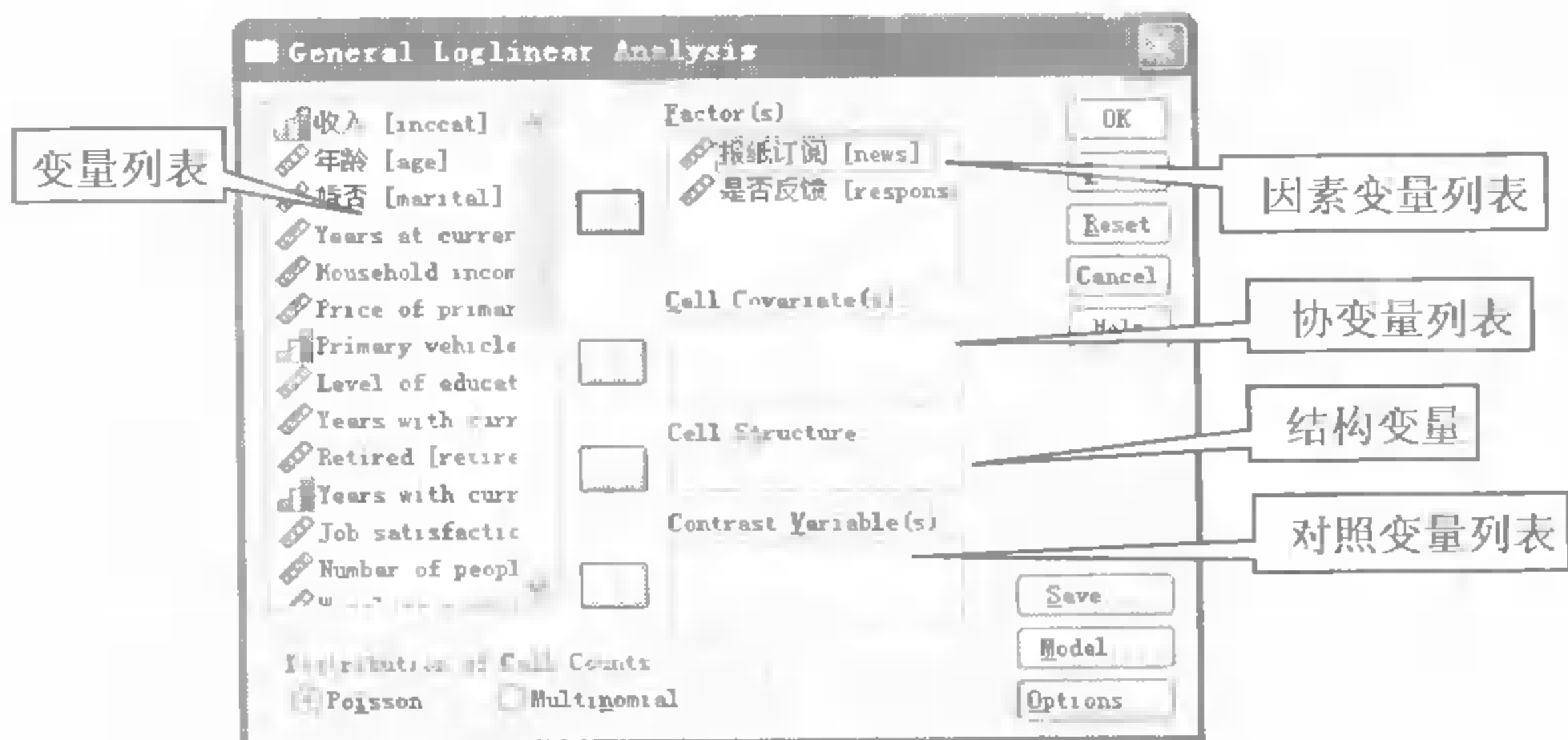



图 16-2 General 过程的主设置界面

#### 1. 变量设置

首先在变量列表选中报纸订阅和是否反馈变量，然后单击从上至下第一个  按钮，将其作为因素变量选入 Factor(s) 列表框。

##### (1) 指定分析变量。


- Factor(s) 列表框，用于选入需要分析的因素变量，最多可选择 10 个。
- Cell Covariate(s) 列表框，用于选入单元格协变量。
- Cell Structure 选框，用于选入单元格结构变量，相当于指定一个权重变量。
- Contrasts Variable(s) 列表框，用于选入对照变量，用于计算广义对数比率 (GLOR)。

(2) Distribution of Cell Counts 子设置栏：指定单元格频数的分布，有两个选择：Poisson 分布，当样本量不固定，并且各单元格相互独立时使用，此项为默认选项；Multinomial 联合



二项分布，当总样本量固定，且各单元格频数不独立时使用。

## 2. 模型设置

在图 16-2 中单击 Model 按钮，弹出如图 16-3 所示的对话框，在此设置模型参数。勾选 Custom 单选框；在因素列表选中 news 和 response 变量，单击 Build Term 下拉列表选中 Main effects 选项，然后单击  按钮，将选中因素的指定效应选入 Terms in Model 列表框。单击 Continue 按钮返回主界面。

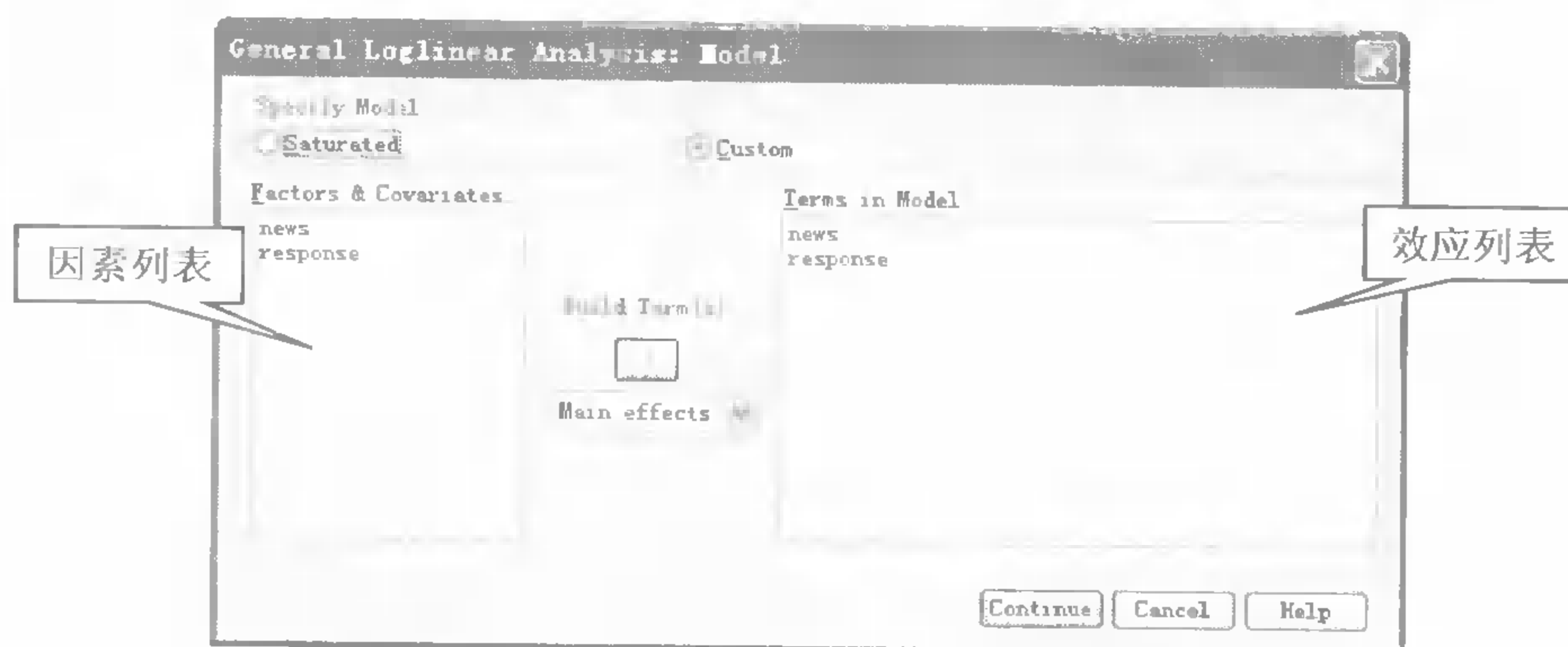


图 16-3 General 过程的 Model 设置

(1) Specify Model 栏，指定使用模型的类别，有如下两个选项。

Saturated 默认选项，指定模型包括所有因素变量的主效应和交互效应，但不包括与协变量有关的效应；Custom 自定义选项，选中后激活下面的设置内容。

(2) Factors and Covariates 列表框，显示因素变量的变量名称。

(3) 选入指定效应的方法。在 Factors and Covariates 列表选中某些因素变量，单击 Build Term(s) 下拉列表指定一种效应类型，再单击  按钮，将选中变量的指定效应选入 Terms 列表框。

Build Term(s) 下拉列表中可选的效应种类有：Main effects（主效应）、Interaction（交互效应）、All n-Way（所有 n 维交互效应）。

## 3. 输出设置

在图 16-2 中单击 Options 按钮，弹出如图 16-4 所示的对话框。勾选 Design matrix 复选框；单击 Continue 按钮返回主界面。

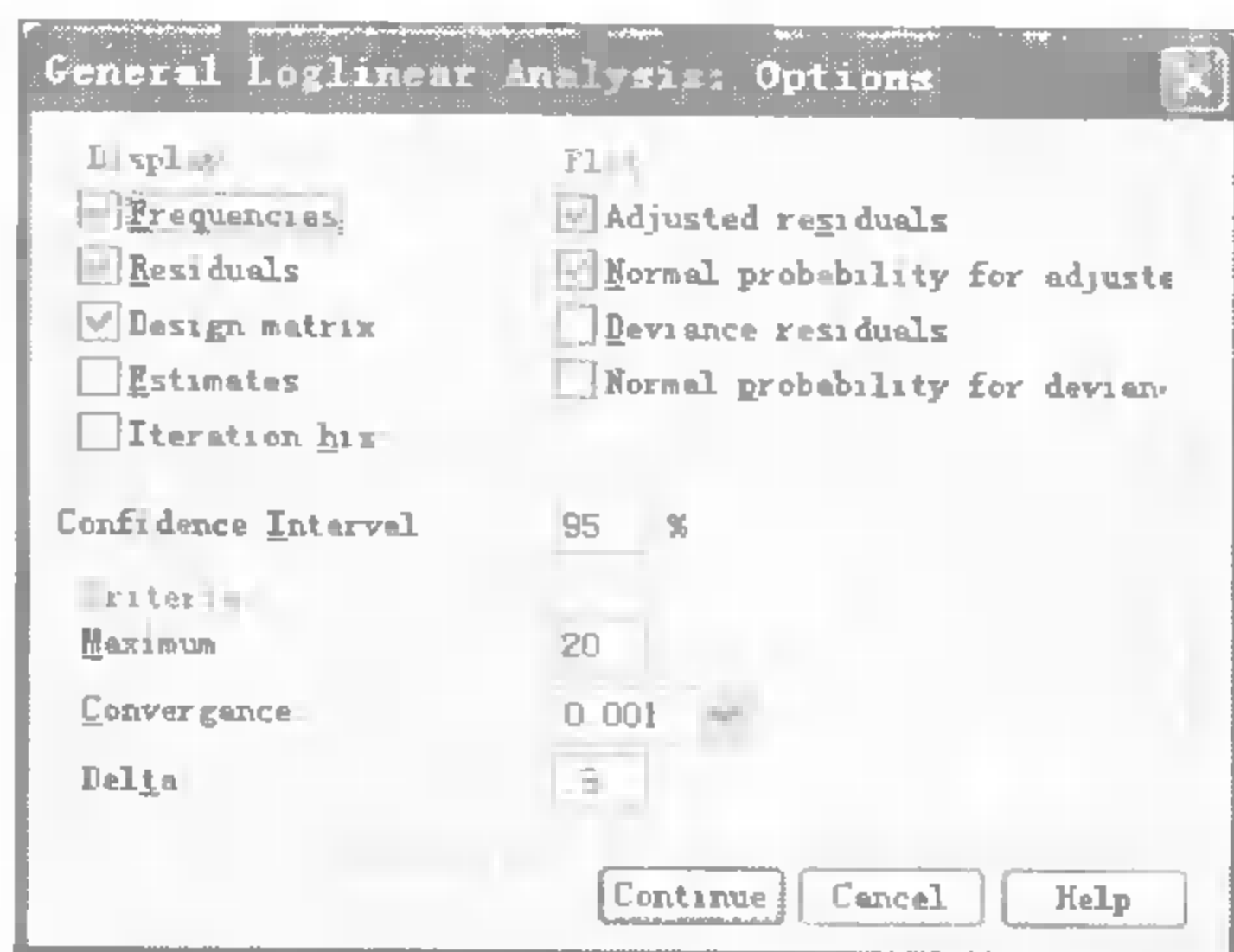


图 16-4 General 过程的 Options 设置

(1) Display 子设置栏。在此选择输出哪些统计信息，可选项有：Frequencies（频数）、Residuals（残差）、Design matrix（设计矩阵）、Estimates（参数估计值）、Iteration history（迭代历史）。

(2) Plot 子设置栏。在此设置与输出图形有关的选项，包括：Adjusted residuals（调整残差图），Normal probability for adjusted（调整残差的正态概率图），Deviance residuals（偏差残差图），Normal probability for deviance（偏差残差的正态概率图）。

(3) Confidence Interval 输入框，指定参数估计值的置信区间，默认为 95%。

(4) Criteria 子设置栏，用于设置与迭代相关的参数，设置内容包括：Maximum iterations 输入框，指定最大迭代次数，默认为 20；Convergence 下拉列表，指定收敛标准，默认为 0.001；Delta 输入框，指定调整系数，默认为 0.5。

#### 4. 保存设置

在图 16-2 中单击 Save 按钮，弹出如图 16-5 所示的保存设置对话框，单击 Continue 按钮返回主界面。

(1) Residual 残差，又称简单残差或原始残差，为单元格观察频数与期望频数之差。

(2) Standardized residuals 标准化残差，用残差除以标准误估计值，又称为 Pearson 残差。

(3) Adjusted residuals 调整残差，如果模型选择正确，它渐进服从标准正态分布，可用于检验标准化残差的正态性。

(4) Deviance residuals 偏差残差，似然比卡方统计量的平方根，且与残差的符号相同，它渐进服从于标准正态分布。

(5) Predicted values 预测值。

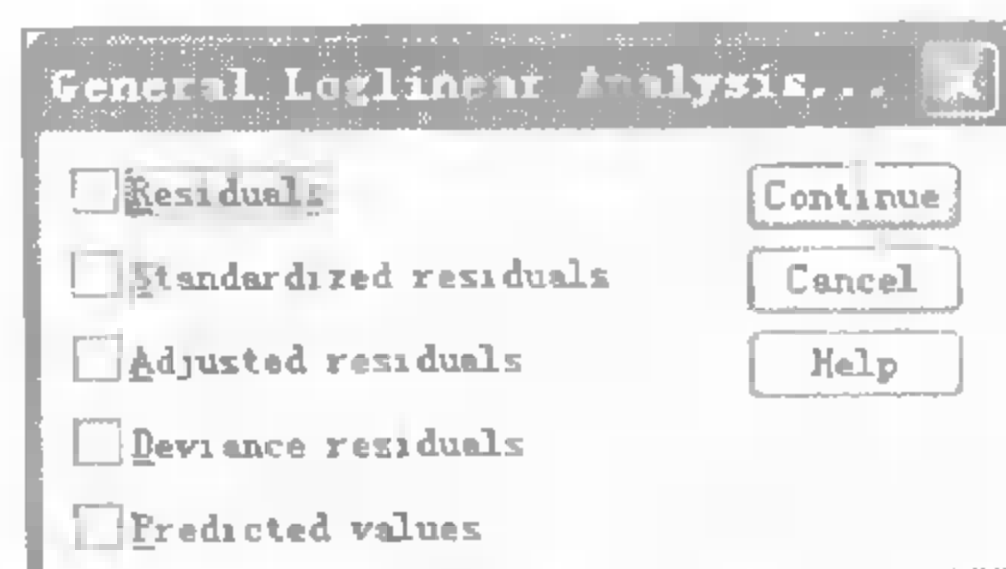


图 16-5 General 过程的保存设置

### 16.2.4 案例的结果分析

在图 16-2 中单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 16-6~图 16-8 所示。

数据信息		
案例	有效	6400
	缺失	0
	加权有效	6400
单元	定义的单元格	4
格	结构中的无效单元	0
	采样无效单元	0
类别	报纸订阅	2
	是否反馈	2

收敛信息 <sup>a,b</sup>	
最大迭代次数	20
收敛公差	.00100
最终最大绝对差值	.00032
最终最大相对差值	.00013
迭代次数	5

a 模型：泊松  
b 设计常量 - news - response  
c 由于参数估计的最大绝对变化小于指定的收敛条件，导致迭代已收敛。

拟合度检验 <sup>a,b</sup>			
	值	df	Sig.
似然比	48.302	1	.000
Pearson 卡方检验	50.781	1	.000

a 模型：泊松  
b 设计常量 - news - response

设计矩阵 <sup>a,b</sup>				
参数	报纸订阅			
	订		不订	
	是否反馈		是否反馈	
	是	否	是	否
单元结构	1	1	1	1
常量	1	1	1	1
[news = 0]	1	1	0	0
[response = 0]	1	0	1	0

设计矩阵的缺省显示已被转置。未显示冗余的参数。  
a 模型：泊松  
b 设计常量 - news - response

图 16-6 基本信息和拟合度检验的结果

单元计数和残差 <sup>a b</sup>									
报纸订阅	是否反馈	观测		期望的		残差	标准化残差	调整残差	偏差
		计数	%	计数	%				
订	是	380	15%	293.664	11.6%	86.332	3.038	7.072	4.817
	否	2388	97.3%	2474.333	98.7%	-86.333	-3.036	-7.072	-4.746
不订	是	299	12%	388.333	15.6%	-88.333	-3.098	-7.072	-4.780
	否	2333	92.1%	2246.667	90.7%	86.333	3.015	7.072	4.509

<sup>a</sup> 模型: 泊松  
<sup>b</sup> 设计常量 = news - response

图 16-7 单元计数和残差

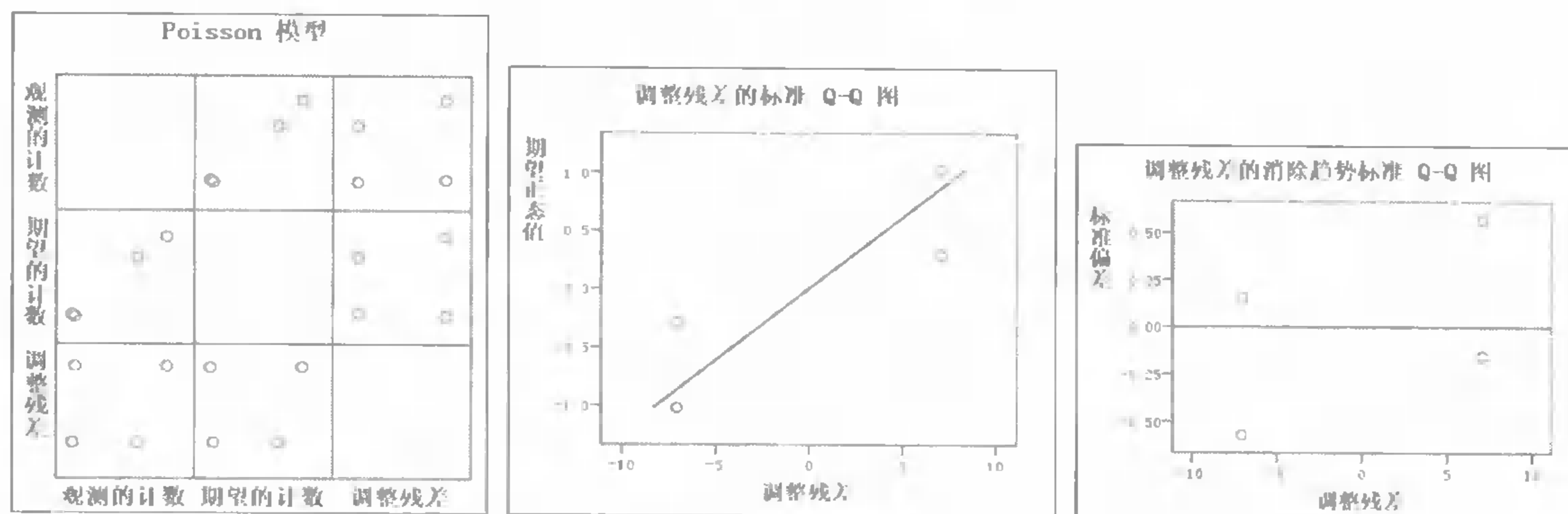


图 16-8 图形输出

(1) 模型基本信息。如图 16-6 所示,“数据信息”表格给出了关于有效和缺失的各种统计信息;“收敛信息”表格给出模型设置的多个收敛临界条件,并说明了当前模型收敛的原因。

(2) 拟合优度检验。如图 16-6 所示,“拟合优度检验”表格给出了似然比检验和 Pearson 卡方检验的结果,检验的零假设是:模型能很好的拟合数据,在本例中相当于说报纸订阅和是否反馈两个变量是独立的。从结果看,这两个检验的显著性 Sig 值都远小于 0,故而否定零假设,认为报纸订阅和是否反馈之间是有一定的相关关系的。

(3) 单元计数和残差表格。如图 16-7 所示,“单元计数和残差”表格反映了对因素变量交叉分类的统计结果,它是计算拟合优度统计量的基础。第一行是订阅报纸的用户反馈的统计结果;第一列是实际观测频数统计结果;“期望的”列是假设模型准确时的期望频数统计结果;后 4 列的残差是观测值和期望值之间的差异,由于残差较大,说明模型不能很好地拟合数据。

(4) 图形输出。如图 16-8 所示,“Poisson 模型”图为四个单元格的观测频数、期望频数和调整残差两两对应的散点图。观察“调整残差的标准 Q-Q 图”和“调整残差的消除趋势标准 Q-Q 图”,发现残差存在着一定的趋势,说明它不服从正态分布,因此所拟合的模型不能完全解释四个单元格频数的分部规律,可能还有有意义的效应未被纳入。

**改进方法:**在如图 16-3 所示的模型设置对话框中,单击选中 Saturated 选项,使用饱和模型进行分析。另外在第 16.4 节还将对这个问题进行研究。

### 16.3 Logit 过程

一般对数线性模型 (General) 能够完成许多分析任务,它的特点是不区分研究因素中的因变量和自变量,对所有变量一视同仁。但有时用户对所研究的问题已经有了一些线索,比如已经确定了何为因何为果,此时再用 General 模型就不能利用这些信息了。在这种情况下,

如果因变量是二元变量，就可以利用 Logit 过程进行分析。

### 16.3.1 Logit 过程概述

Logit 对数线性分析 (Logit Log-linear Analysis)，用于分析因变量与自变量之间的相关关系。与 General 模型相比，Logit 模型更像是方差分析，它明确分出了因变量和自变量，直接服务于分类变量之间的因果关系。SPSS 的 Logit 过程假设研究数据服从多项式分布，因此又称为多项式 Logit 模型，其参数估计使用 Newton—Raphson 方法。

Logit 过程输出的统计量与图形很多，包括：观测频数、期望频数、初始残差 (raw residual)、调整残差 (adjusted residual) 及偏差残差 (deviance residual)，设计矩阵 (design matrix)，参数估计值 (parameter estimate)，发生比 (odds ratio) 和广义对数发生比 (generalized log-odds ratio)，Wald 统计量 (Wald statistic)，调整残差图，偏差残差图，及正态概率图等。

### 16.3.2 问题描述和数据准备



某快餐公司为了提高其早餐的市场份额，对 880 名消费者做了一次调查，本节利用对数线性模型的 Logit 过程，分析所得数据以了解三种早餐的市场销售情况。数据摘自 SPSS 自带的 Demo 文件 “cereal.sav”，数据文件为 “早餐偏好调查数据.sav”，数据格式如图 16-9 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	agecat	Numeric	4	0	年龄段	{1, 低于31}..	None	8	Right	Ordinal
2	gender	Numeric	4	0	性别	{0, 男}...	None	8	Right	Nominal
3	active	Numeric	4	0	生活方式	{0, 消极}...	None	8	Right	Nominal
4	bfast	Numeric	4	0	早餐	{1, 不吃}...	None	8	Right	Nominal

图 16-9 早餐偏好调查数据格式

本例问卷设计了如图所示的 4 个问题，其中“早餐”取值 1 表示不吃，取值 2 表示吃麦片，取值 3 表示吃谷类。

### 16.3.3 Logit 过程的参数设置

依次单击菜单 “Analyze→Loglinear→Logit...”，打开 Logit 过程的主设置面板，如图 16-10 所示。在变量列表单击选中早餐变量，单击从上至下第一个  按钮，将其作为因变量选入 Dependent 列表框；在变量列表选中年龄段、性别和生活方式变量，单击从上至下第二个  按钮，将其作为因素变量选入 Factor (s) 列表框。

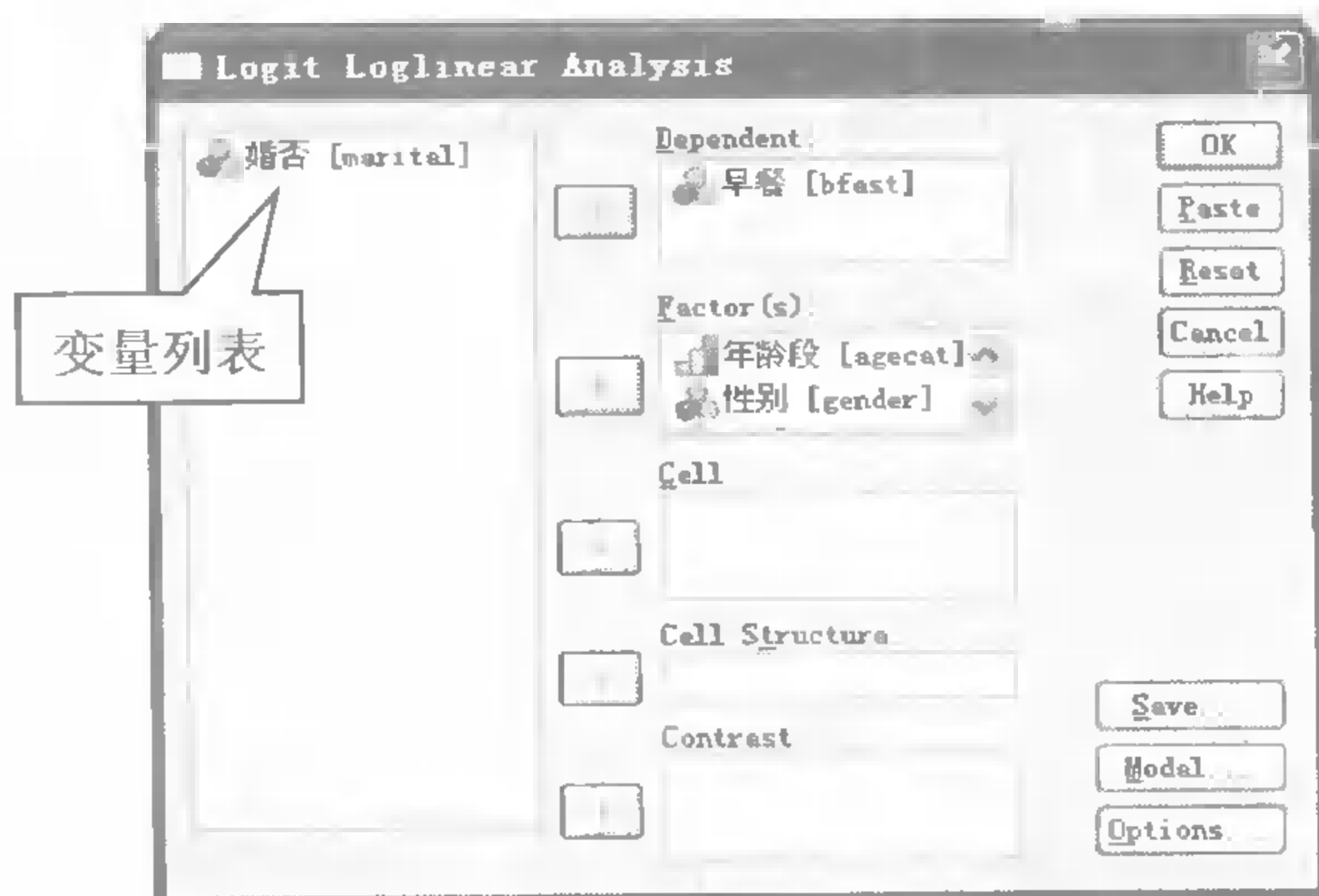



图 16-10 Logit 过程的主设置界面



Dependent 列表框用于选入因变量, 必须为分类变量; Factor(s)列表框用于选入因素变量; 其他选项的设置与图 16-2 相仿。

另外, 单击 Model、Options、Save 按钮所弹出的子设置界面, 分别与第 16.3.3 节介绍的 General 过程的图 16-3、图 16-4 和图 16-5 相仿。

单击 Options 按钮, 弹出如图 16-11 所示的对话框, 勾选 Estimates 复选框; 单击 Continue 按钮返回主界面。单击 Model 按钮, 弹出如图 16-12 所示的对话框, 单击选中 Custom 单选框; 在因素列表选中 active 和 agecat 变量, 单击 Build Term 下拉列表选中 Main effects 选项, 然后单击  按钮, 将选中因素的指定效应选入 Terms in Model 列表框; 单击取消 Include 复选框; 单击 Continue 按钮返回主界面。

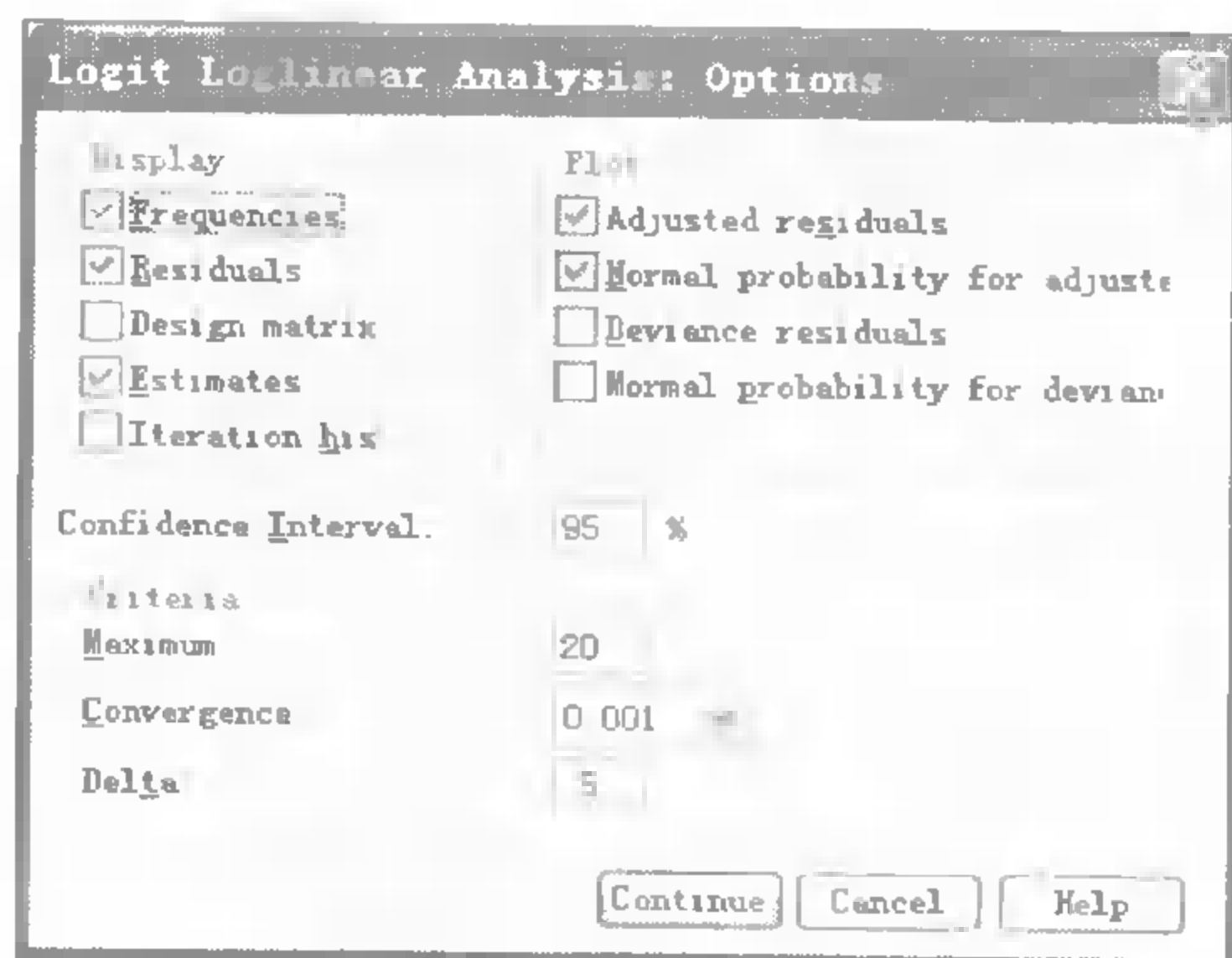


图 16-11 Logit 过程的 Options 设置

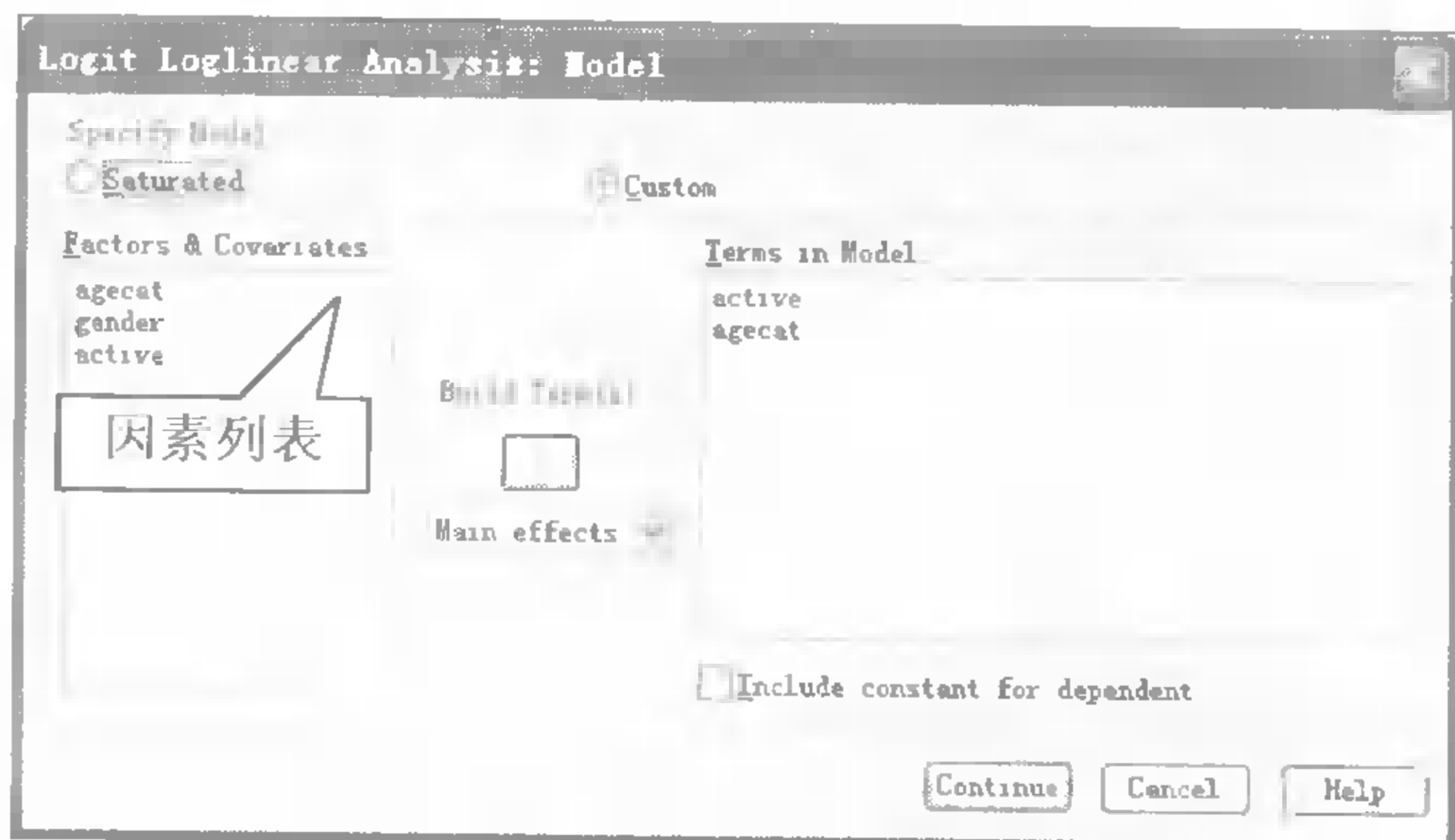


图 16-12 Logit 过程的模型设置

#### 16.3.4 案例的结果分析

在图 16-10 中, 单击 OK 按钮运行, 首先会弹出“变量 gender 在分析中未使用”的警告对话框, 在其中单击确定按钮继续, SPSS Viewer 窗口的输出结果如图 16-13~图 16-16 所示。

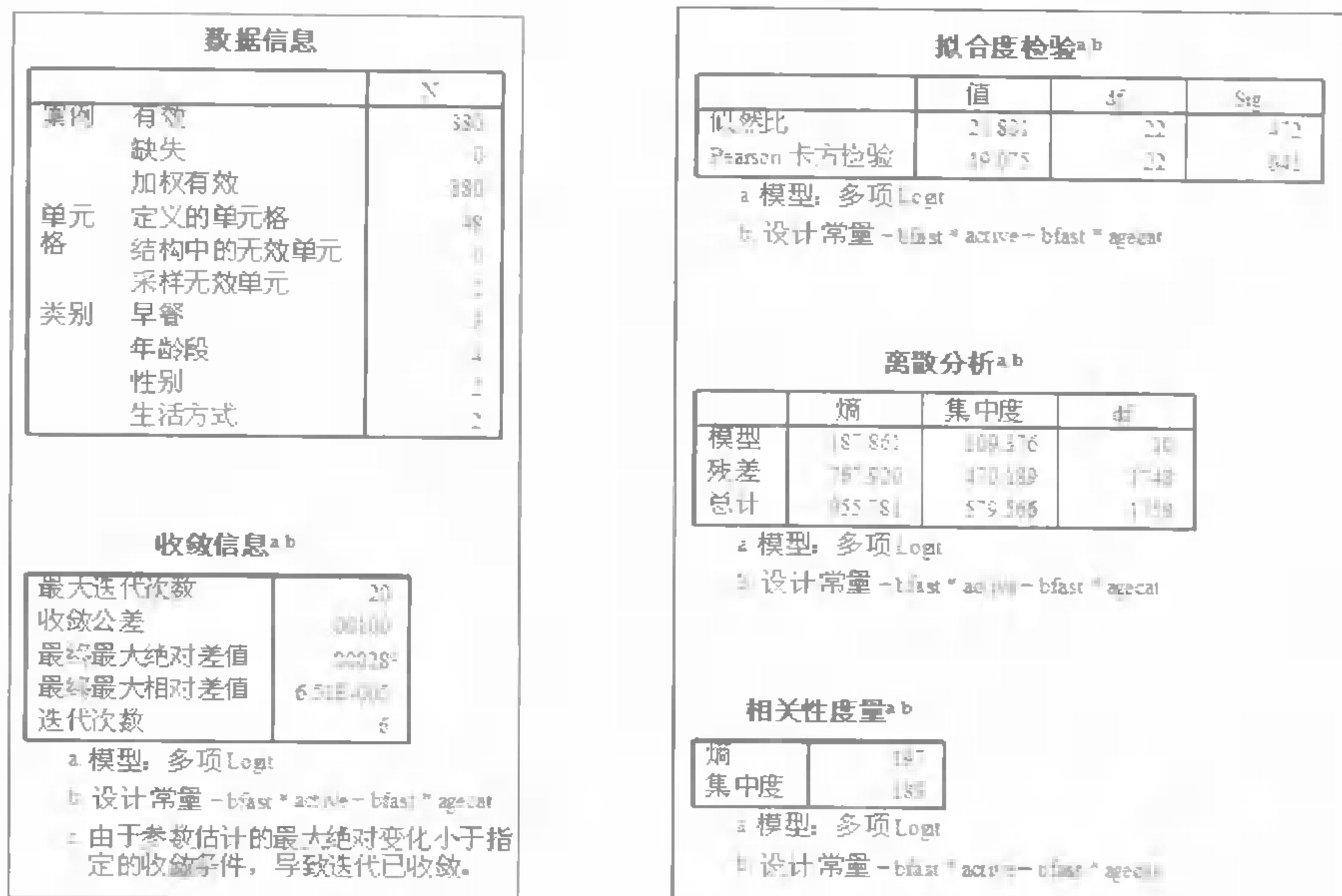


图 16-13 基本信息和拟合优度检验结果

单元计数和残差 <sup>a,b</sup>											
年龄段	性别	生活方式	早餐	观测		期望的		残差	标准化残差	调整残差	偏差
				计数	%	计数	%				
低于31	男	消极	不吃	12	37.5%	11.063	34.6%	.937	.346	.405	1.397
			麦片	0	0%	.941	2.9%	-.941	-.985	1.130	.000
			谷类	23	62.5%	19.996	62.5%	.004	.001	.002	.069
		积极	不吃	33	50.6%	28.553	53.9%	-.553	-.152	-.190	-1.047
			麦片	0	0%	.927	1.7%	-.927	-.971	-1.115	.000
			谷类	25	47.2%	23.520	44.4%	1.480	.409	.509	1.747
	女	消极	不吃	14	36.8%	13.137	34.6%	.863	.294	.354	1.335
			麦片	2	5.3%	1.118	2.9%	.882	.847	1.003	1.526
			谷类	22	57.9%	23.745	62.5%	-1.745	-.585	-.704	-1.633
		积极	不吃	30	51.7%	31.247	53.9%	-1.247	-.328	-.422	-1.563
			麦片	2	3.4%	1.014	1.7%	.986	.987	1.151	1.648
			谷类	26	44.8%	25.739	44.4%	.261	.069	.068	.725
31-45	男	消极	不吃	16	37.2%	13.955	32.5%	2.045	.666	.793	2.092
			麦片	6	14.0%	6.416	14.9%	-.416	-.178	-.213	-.897
			谷类	21	48.8%	22.629	52.6%	-1.629	-.497	-.594	-1.771
		积极	不吃	23	42.6%	28.206	52.2%	-5.208	-1.419	-1.748	-3.064

图 16-14 统计结果摘要

参数估计 <sup>c,d</sup>						
参数	估计	标准误	Z	Sig.	95% 置信区间	
					下限	上限
[bfast = 1] * [active = 0]	-.580	.285	-2.032	.040	-1.158	-.002
[bfast = 1] * [active = 1]	-.744	.287	-2.590	.010	-1.308	-.181
[bfast = 2] * [active = 0]	-.200	.162	-1.241	.219	-.523	.123
[bfast = 2] * [active = 1]	-.022	.195	-.112	.910	-.400	.356
[bfast = 3] * [active = 0]	.0	.	.	.	.	.
[bfast = 3] * [active = 1]	.0	.	.	.	.	.
[bfast = 1] * [agecat = 1]	.938	.313	2.998	.003	.323	1.553
[bfast = 1] * [agecat = 2]	1.047	.311	3.366	.001	.437	1.656
[bfast = 1] * [agecat = 3]	.263	.352	.745	.458	-.437	.964
[bfast = 1] * [agecat = 4]	.0	.	.	.	.	.
[bfast = 2] * [agecat = 1]	-.4356	.335	-1.297	.190	-1.101	.230
[bfast = 2] * [agecat = 2]	-.3461	.275	-1.241	.219	-.899	.207
[bfast = 2] * [agecat = 3]	-.1115	.208	-.536	.592	-.523	.299
[bfast = 2] * [agecat = 4]	.0	.	.	.	.	.
[bfast = 3] * [agecat = 1]	.0	.	.	.	.	.
[bfast = 3] * [agecat = 2]	.0	.	.	.	.	.
[bfast = 3] * [agecat = 3]	.0	.	.	.	.	.
[bfast = 3] * [agecat = 4]	.0	.	.	.	.	.

a 在多项式假设中常量不作为参数使用。因此不计算它们的标准误差。  
b 此参数为冗余参数，因此将被设为零。  
c 模型：多项 Logistic  
d 设计常量 = bfast \* active + bfast \* agecat

图 16-15 参数估计结果

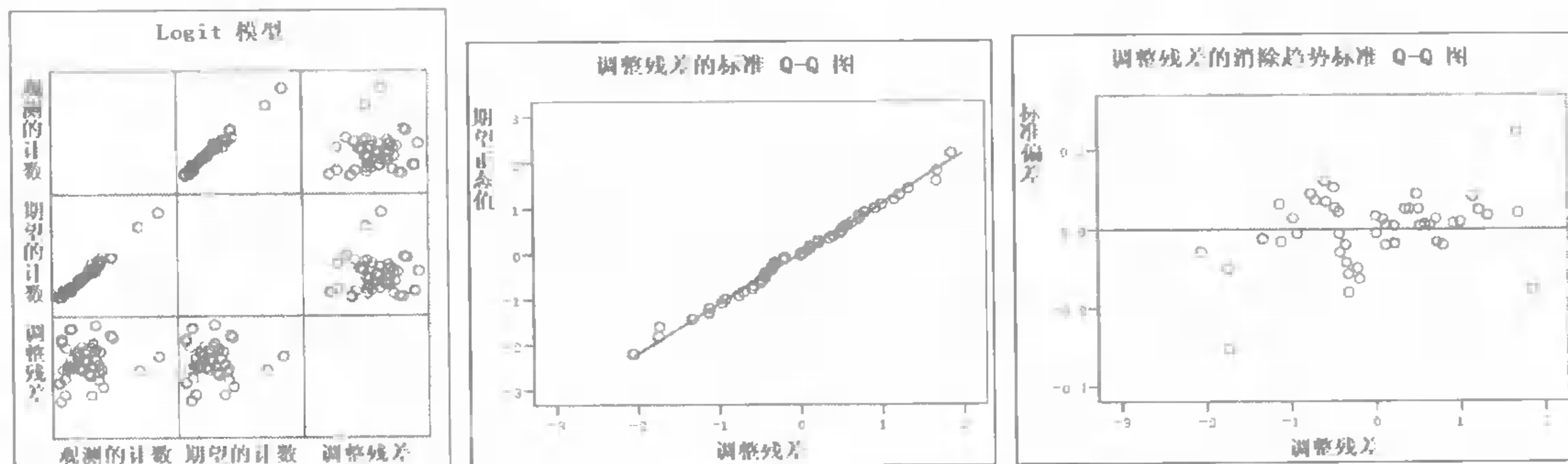


图 16-16 图形输出

(1) 模型基本信息。如图 16-13 所示，“数据信息”表格给出了关于有效和缺失的各种统计信息；“收敛信息”表格给出模型设置的多个收敛临界条件，并说明了当前模型收敛的原因。

(2) 拟合优度检验。如图 16-13 所示，“拟合优度检验”表格给出了似然比检验和 Pearson 卡方检验的结果，检验的零假设是：模型能很好的拟合数据。从结果看，这两个检验的显著性

Sig 值都大于 0.10, 故而接受零假设, 说明三个因素对早餐的偏好有显著的影响。

“相关性度量”中的熵 (entropy) 和集中度 (concentration) 取值越大, 说明模型可解释的离差越多, 最大为 1。

(3) 单元计数和残差表格。如图 16-14 所示, “单元计数和残差”表格给出了因素变量交叉分类的部分统计结果, 它是计算拟合优度统计量的基础。第一行是 31 岁以下消极男性的早餐偏好频数统计信息; “观测”列是实际观测频数统计结果; “期望的”列是假设模型准确时的期望频数统计结果; 后 4 列的残差是观测值和期望值之间的差异, 残差越小, 模型的拟合效果越好。

(4) 参数估计结果。如图 16-15 所示, 是参数估计的部分结果。Z 检验的显著性 Sig 值都很小, 故可推断相应的系数都显著的不为 0, 即生活方式和年龄段这两个因素对早餐的选择都有影响。“估计” (Estimate) 列为参数交互作用的系数值, 如果某个因素水平的系数为正, 则因变量取相应类别的概率要大于取参考类的概率, 反之亦然。例如: [bfast=1]\*[active=0] 的系数为负, 说明不考虑年龄时, 生活方式消极的人不吃早餐的概率较小; [bfast=1]\*[agecat=2] 的系数为正, 说明不考虑生活方式时, 中年人不吃早餐的概率较大。

(5) 图形输出。如图 16-16 所示, “Poisson 模型”图为四个单元格的观测频数、期望频数和调整残差两两对应的散点图。观察“调整残差的消除趋势标准 Q-Q 图”, 并未发现残差存在着某些趋势, 不能否定它服从正态分布, 因此推断模型能较好的解释原始数据。

## 16.4 Model Selection 过程

Model Selection 过程, 可以用来拟合分层对数线性模型 (Hierarchical Model)。如果用户只是设想若干分类变量之间可能会存在一定的关系, 但是并无明确的假设, 比如没有具体区分出因变量和自变量, 此时就适宜采用分层对数线性模型。

### 16.4.1 Model Selection 过程概述

模型选择 (Model Selection) 对数线性模型, 能够对多维列联表进行分析, 它使用迭代比例拟合方法来估计参数, 可以发现哪些分类变量之间存在着联系, 帮助用户进行探索性的分析。SPSS 的 Model Selection 过程可以选择两种方式建立模型: 强迫引入法和向后消去法。

向后消去法指从饱和模型入手, 从高阶交互项开始逐步排除无意义的参数, 直到形成一个最佳的简约模型。但是分层模型只输出饱和模型的参数估计及偏相关分析, 不能输出简约模型的参数估计结果, 故而在用它得到了最佳的简约模型后, 还应当采用 General 一般模型计算具体的参数估计值和检验结果。

对多数用户来说, 该过程的实用性最好, 因为它可以进行属性变量的自动筛选, 类似于多元回归中的逐步回归, 这在对三维以上的列联表进行联合分析时, 可以极大地降低工作量。

Model Selection 过程输出的统计量和图形包括: 频数 (frequencies)、残差 (residual), 参数估计值 (parameter estimate), 标准误差 (standard error), 置信区间 (confidence interval), 偏相关检验 (test of partial association), 残差图及正态概率图等。

### 16.4.2 问题描述和数据准备

本节对杂志订阅的例子进行研究。所用数据文件为“报纸订购调查数据.sav”, 数据格式如图 16-17 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	news	Numeric	4	0	报纸订阅	{0, 1}...	None	8	Right	Scale
2	response	Numeric	4	0	是否反馈	{0, 是}...	None	8	Right	Scale
3	inccat	Numeric	8	2	收入	{1.00, Under	None	8	Right	Ordinal
4	age	Numeric	4	0	年龄	None	None	8	Right	Scale
5	marital	Numeric	4	0	婚否	{0, Unmarried	None	8	Right	Scale

图 16-17 杂志订阅的调查数据格式

在第 16.2 节中，出版社为了提高杂志定购的反馈率，采用的方法是只向定购报纸的用户发送促销邮件；本节进一步假设收入水平会影响反馈率的高低，故而建议只向具有一定收入的用户发送邮件。随后，就用收入、报纸订阅和反馈率这三个因素变量来拟合对数线性模型，并根据模型找到合适的目标客户群，以提高反馈率。

本例所使用的 3 个因素中，“报纸订阅”取值 1 表示用户表定购报纸，取值 0 表示用户不定购报纸；“是否反馈”取值为 1 表示反馈，取值为 0 表示未反馈；inccat（收入）的单位是千美元，取值 1 表示收入在 25 千美元以下，取值 2 表示收入在 25~49 千美元之间，取值 3 表示收入在 50~74 千美元之间，取值 4 为 75 千美元以上。

### 16.4.3 层次对数线性模型的操作过程

依次单击菜单“Analyze→Loglinear→Model Selection...”，打开 Model Selection 过程的主设置面板，如图 16-18 所示。

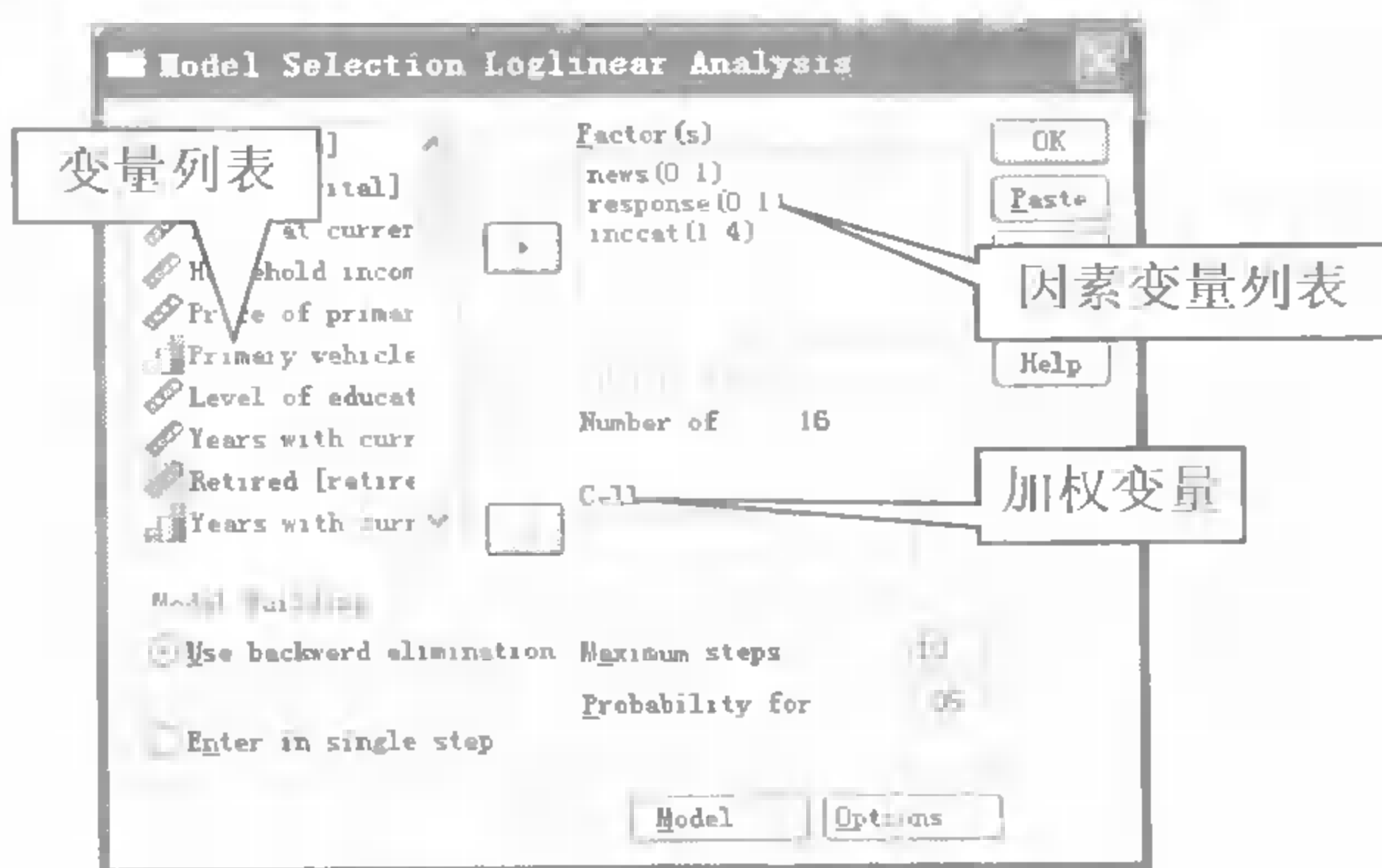



图 16-18 Model Selection 过程的主界面

#### 1. 变量设置

在变量列表选中报纸订阅、是否反馈和收入变量，单击从上至下第一个  按钮，将其作为因素变量选入 Factor(s) 列表框。在 Factor(s) 列表选中 news（报纸订阅）变量，单击 Define Range 按钮，弹出如图 16-19 所示的对话框，分别在 Minimum、Maximum 后输入“0”和“1”，单击 Continue 按钮返回主界面；用同样的方法设置 response 变量的取值范围为 0~1，及 inccat 变量的取值范围为 1~4。

（1）Factor(s) 因素变量列表框，用于选入两个或两个以上的分类变量。

在此列表选中某个变量后，单击 Define Range 按钮，弹出如图 16-19 所示的 Loglinear Analysis: Define Range（对数分析：定义变量）对话框，在此指定选中因素变量的 Minimum（最小值）和 Maximum（最大值）。

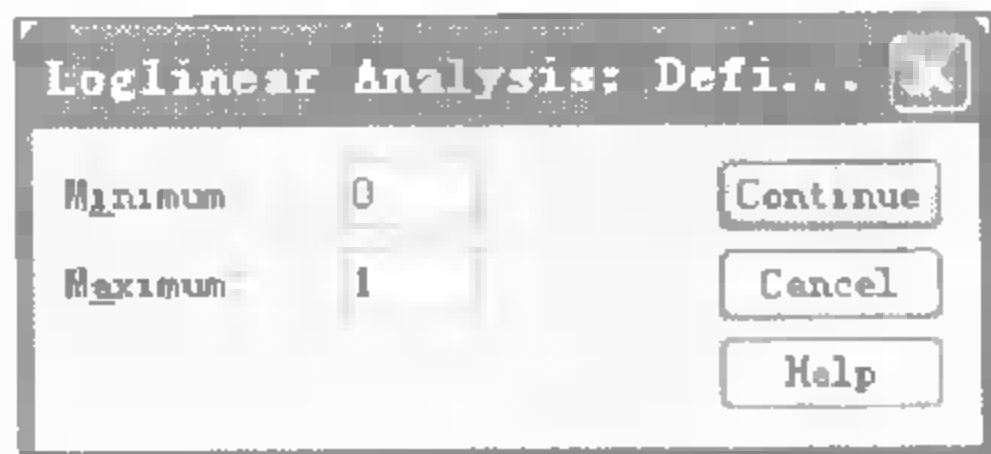


图 16-19 取值范围定义对话框



(2) Cell Weights 选框, 指定对单元格的加权变量。

(3) Model Building 子设置栏, 设置模型拟合的方法, 有如下两个可选项。

- ① Use backward elimination, 向后消去法, 在 Maximum steps 输入框指定最大步骤数, 在 Probability for removal 输入框指定剔除变量的临界概率。默认方法。
- ② Enter in single step, 强行进入法。

## 2. 输出设置

在图 16-18 中, 单击 Option 按钮, 弹出如图 16-20 所示的输出设置对话框, 勾选 Parameter estimates 复选框; 单击 Continue 按钮返回主界面。

(1) Display 栏, 选择输出哪些统计量, 可选项有 Frequencies (频数) 和 Residuals (残差); 在饱和模型中, 观察数和期望数相等, 残差等于 0。

(2) Display for saturated Model 栏, 选择专为饱和模型输出的一些统计量, 包括 Parameter estimates (参数估计值及其置信区间) 和 Association table (偏相关检验表)。由于分层模型靠后的结果不会给出参数的估计和检验结果, 故可以使用最初模型的参数估计和检验结果作为参考。

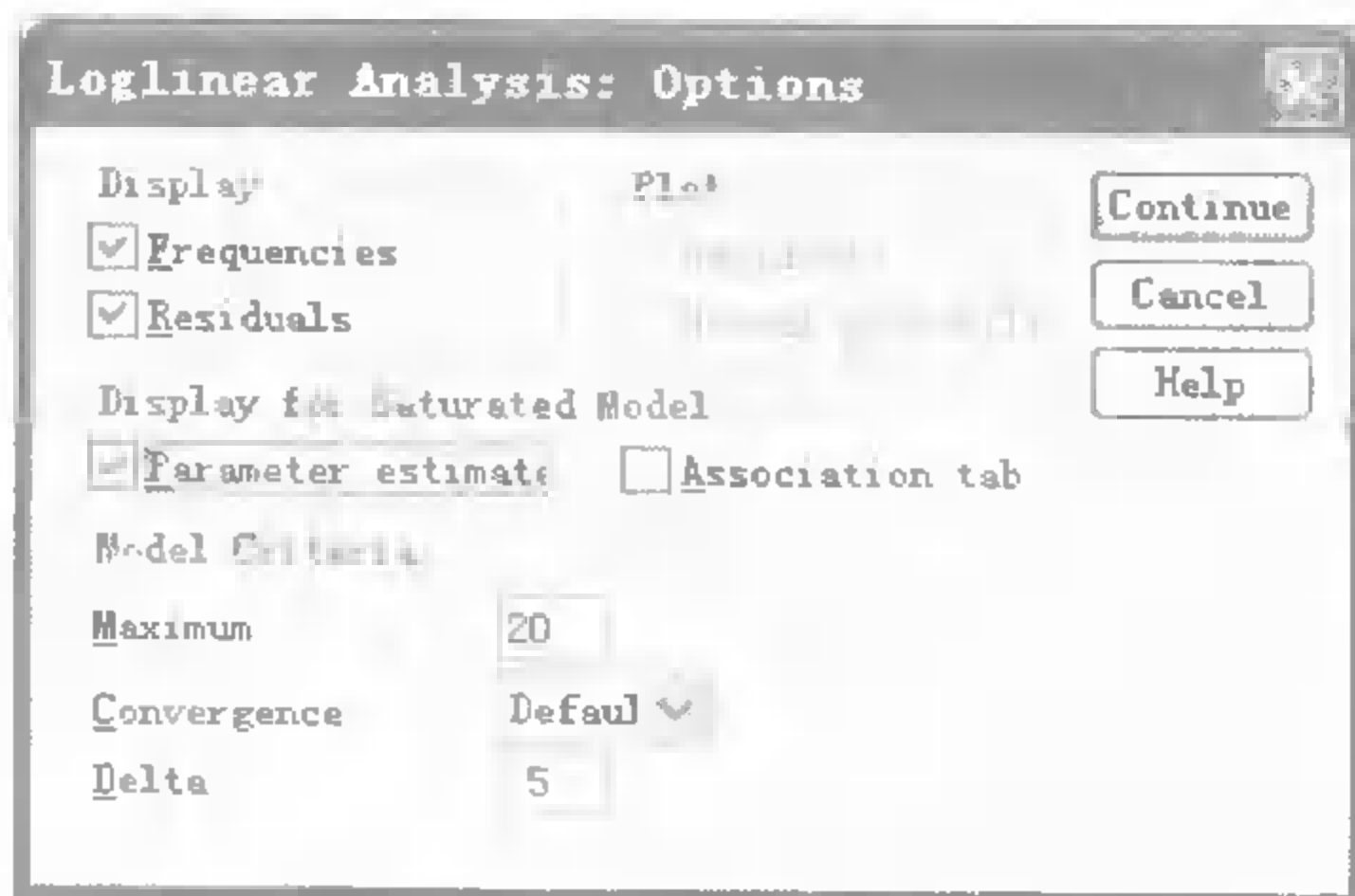


图 16-20 Model Selection 过程的输出选项设置

(3) Plot 栏, 选择输出图形, 包括: Residuals (残差图) 和 Normal probability (正态概率图)。

(4) Model Criteria 子设置栏, 在此设置关于迭代的参数: Maximum iterations 输入框, 指定最大迭代次数, 默认为 20; Convergence 下拉列表, 指定收敛值; Delta 输入框, 指定  $\delta$  值, 默认为 0.5。

## 3. 模型设置

在图 16-18 中, 单击 Model 按钮弹出的对话框与图 16-12 所示的 Logit 对数模型的 Model 子设置对话框相同, 在此均采用默认设置, 单击 Continue 按钮返回主界面。

### 16.4.4 案例的结果分析

在图 16-10 中单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 16-21~图 16-23 所示。

参数估计值							
效果	参数	估计	标准误	Z	Sig.	95% 置信区间	
						下限	上限
news*response*inccat	1	.058	.036	1.578	.115	-.014	.129
	2	-.022	.032	-.669	.500	-.083	.040
	3	.021	.047	.457	.648	-.070	.112
news*response	1	.172	.023	7.601	.000	.128	.217
news*inccat	1	-.139	.036	-3.675	.000	-.205	-.063
	2	-.133	.032	-4.147	.000	-.216	-.050
	3	.036	.047	.779	.436	-.055	.128
response*inccat	1	.278	.036	7.619	.000	.206	.349
	2	.075	.032	2.348	.019	.012	.139
	3	-.174	.047	-3.716	.000	-.265	-.083
news	1	-.003	.023	-.232	.816	-.050	.049
response	1	.1076	.023	47.455	.000	.1121	.1032
inccat	1	.075	.036	2.079	.038	-.117	.004
	2	.504	.032	15.650	.000	.441	.567
	3	-.427	.047	-9.168	.000	-.518	-.335

图 16-21 参数估计值

(1) 参数估计结果。如图 16-21 所示，在参数估计值表中，Z 统计量的值等于“估计”与“标准误”之比，如果 Z 检验的显著性 Sig 值小于 0.05，就说明相应参数显著地不为 0。

(2) 步骤摘要表。如图 16-22 所示，“步骤摘要”表格给出了向后消去法的迭代步骤。第 0 步检验饱和模型（初始模型）的 3 阶交互效应 `inccat*news*response` 是否显著，由于卡方改变量（已删除）的显著性检验 Sig 值  $0.262 > 0.10$ ，因此应在此模型剔除 `inccat*news*response` 效应项。

步骤摘要						
步骤 <sup>a</sup>	效果	卡方 <sup>b</sup>	df	Sig.	迭代数	
0	生成类 已删除 生成类	news*response*inccat news*response*inccat	3.598	3	.262	4
1	已删除 的效果	news*response news*inccat response*inccat news*response news*inccat response*inccat	3.598 3.598 3.598 3.598 3.598 3.598	3 3 3 3 3 3	.262 .262 .262 .262 .262 .262	
2	生成类	news*response news*inccat response*inccat	3.598	3	.262	

a 对于已删除的效果，从模型中删除该效果之后，这是卡方中的更改。

b 在每一步骤中，如果最大显著性水平大于 .050，则删除含有 似然比更改 的最大显著性水平的效果。

c 在步骤 1 之后，将在每一步骤显示最佳模型的统计量。

图 16-22 迭代步骤

单元计数和残差								
模型	反馈	inccat	观测		期望		残差	标准残差
			计数	%	计数	%		
1	是	Under 25	152.000	11.8%	21.475	1.5%	130.525	10.3%
		25+ 549	174.000	13.7%	23.934	1.7%	150.066	11.9%
		550-974	12.000	.9%	18.224	.9%	-6.224	-.5%
		975+	91.000	7.1%	14.084	1.0%	76.916	6.0%
	否	Under 25	314.000	24.5%	228.028	1.7%	85.972	6.7%
		25+ 549	741.000	58.0%	719.066	5.5%	21.934	1.7%
		550-974	434.000	34.0%	429.771	3.3%	4.229	.3%
		975+	154.000	12.0%	335.144	2.6%	-211.144	-16.7%
不拟合	是	Under 25	152.000	11.8%	21.475	1.5%	130.525	10.3%
		25+ 549	174.000	13.7%	23.934	1.7%	150.066	11.9%
		550-974	12.000	.9%	18.224	.9%	-6.224	-.5%
		975+	91.000	7.1%	14.084	1.0%	76.916	6.0%
	否	Under 25	314.000	24.5%	228.028	1.7%	85.972	6.7%
		25+ 549	741.000	58.0%	719.066	5.5%	21.934	1.7%
		550-974	434.000	34.0%	429.771	3.3%	4.229	.3%
		975+	154.000	12.0%	335.144	2.6%	-211.144	-16.7%

拟合优度检验			
	卡方	df	Sig.
似然比	3.598	3	.262
Pearson	4.028	3	.258

图 16-23 单元计数和拟合优度检验

第 1 步，检验在包含所有二阶交互效应和主效应的模型中，每个二阶效应是否显著。卡方改变量（已删除）的显著性检验 Sig 值都小于 0.01，因此可以否定零假设，认为这三个二阶交互效应都应该保留在模型中。

第 2 步，由于所有的二阶交互效应都不能被剔除，无需再做其他检验，此步得到最优模型，它包括主效应和所有二阶效应，这说明报纸订阅和收入都影响着邮件的反馈率。

(3) 单元格统计和残差表。如图 16-23 所示，“单元计数和残差”表格给出了因素变量交叉分类的统计结果，它是计算拟合优度统计量的基础。第一行表示收入在 25 千美元以下，订购报纸并且有反馈的客户的频率统计信息；“观测”列是实际观测频数统计结果；“期望的”列是假设模型准确时的期望频数统计结果；后 4 列的残差是观测值和期望值之间的差异，残差越小，模型的拟合效果越好。

(4) 拟合优度检验。如图 16-23 所示，“拟合优度检验”表格给出了似然比和 Pearson 卡方检验的结果，检验的零假设为：最终模型能很好的拟和原始数据。从显著性检验的 Sig 值都大于 0.10 看，不能否定零假设，即认为模型拟合效果不错。

对应分析 (Correspondence Analysis, CORA), 是由法国人 Jean Paul Benzerc 于 20 世纪 60 年代创立, 直到 80 年代才在英语国家兴起的一种多元相依 (Inter-dependence) 变量统计分析技术。它主要对名义变量和定序变量的多维频度表进行分析, 探索相同变量的不同取值类别之间的差异, 以及不同变量的不同取值类别之间的对应关系。

使用对应分析的条件包括: 变量是名义变量或定序变量; 行变量的类别取值与列变量相互独立; 行、列变量构成的交叉频数表中不能有 0 值或负数。

根据前人的经验, 总结出对应分析的如下 4 个优点。

- 1. 名义变量划分的类别越多, 这种分析的优势越明显。
- 2. 可以将名义变量或定序变量转变为间距变量。
- 3. 揭示行变量类别间与列变量类别间的联系。
- 4. 将变量类别间的联系直观地表现于图形中。

## 17.1 对应分析的基本原理

当研究分类变量之间的关系时, 可以采用卡方检验, 也可以使用对数线性模型。但当分类变量较多, 或各个变量的类别取值较多时, 用以上方法就无法直观而简单地给出各分类之间的联系, 解释起来也略显复杂。此时可以用对应分析方法进行处理, 它能输出简单直观的结果, 当变量个数越多、各个变量的类别取值越多时, 对应分析的优势就越明显。

### 17.1.1 对应分析与因子分析

对应分析也称关联分析、R-Q 型因子分析, 通过分析由定性变量构成的交叉汇总表来揭示变量间的联系, 主要应用于市场细分、产品定位、地质研究以及计算机工程等领域中。它是一种视觉化的数据分析方法, 能够将几组看不出任何联系的数据, 通过直观的定位图展现出来。

普通的 R 型和 Q 型因子分析往往是相互独立的, 分别对样品和属性进行处理, 因此用因子分析研究属性和样品之间的内在联系, 就比较困难, 于是产生了对应分析法。对应分析综合了 R 型和 Q 型因子分析的优点, 将它们统一起来, 使得由 R 型的分析结果很容易得到 Q 型的分析结果, 这同时也克服了 Q 型因子分析计算量大的困难; 更重要

的是它把变量和样品的载荷反映在相同的公因子轴上，将变量和样品联系起来，更加便于解释和推断。

对应分析的基本思想，是将一个列联表的行和列中各元素的比例结构，以点的形式在较低维的空间中表示出来。它最大特点是把众多的样品和众多的变量同时作到同一张图上，将样品的种类及其属性在图上直观地表示出来。另外，它还省去了因子选择和因子轴旋转等复杂的数学运算及中间过程，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要参数（主因子）以及分类的依据，是一种直观、简单、方便的多元统计方法。

对应分析的主要结果是反映变量间相互关系的对应分析图。根据 R 型因子分析和 Q 型因子分析的内在联系，可在同一个图形中将样品和属性同时反映出来，图形中邻近的变量点表示它们关系密切，邻近的样品点也表示它们关系密切；而且属于同一类型的样品点，可以用邻近的变量点来表征。对应分析的目的之一，就是在一个低维空间中描述各个变量间的关系。

对应分析揭示的是环境、结构、行为之间的“对应关系”，能够说明有什么类型的环境和结构，就可能会出现什么类型的行为，而不是反映各个变量间的“因果关系”。它常用于研究多个分类变量（名义变量或定序变量）间的关系，是市场细分、产品定位、品牌形象以及满意度研究等领域常用的一种方法。

### 17.1.2 SPSS 中的对应分析

对应分析根据所用变量的数目分为两种：简单对应分析，用于分析两个分类变量之间的关系，在 SPSS 中使用 Correspondence Analysis 过程执行；多元对应分析（也称多重对应分析）用于分析一组分类变量之间的相关性，在 SPSS 中使用 Optimal Scaling（最优尺度分析）过程来拟合。

使用 SPSS 的简单对应分析功能，输出的统计量与图形包括：对应测度（correspondence measure），行与列的分类信息（profile），奇异值（singular value），行与列的得分（score），行与列得分的置信区间，转换图（transformation plot），行点图（row point plot），列点图（column point plot）及行列点图（biplot）等。

### 17.1.3 使用对应分析的注意事项

虽然对应分析有不少优点，但在某些方面还是有所缺憾，运用时也需注意以下问题。

（1）对应分析不能用于相关关系的假设检验。它虽然可以揭示变量间的联系，但不能说明两个变量之间存在的联系是否显著。因而在做对应分析前，可以用卡方统计量检验两个变量的相关性。

（2）对应分析输出的图形通常是二维的，这是一种降维的方法，将原始的高维数据按一定规则投影到二维图形上。而投影可能引起部分信息的丢失。

（3）对极端值敏感，极端值（异常点）对对应分析的结果影响较大。在进行分析之前，建议先检查列联表中的数据，避免极端值的存在。比如有取值为零的数据存在时，可视情况将相邻的两个状态取值合并。

（4）原始数据的无量纲化处理。运用对应分析法处理问题时，各变量应具有相同的量纲（或者均无量纲）。



## 17.2 简单对应分析

对两个定性变量进行对应分析时,因为变量取值都是离散的,所以将变量取值转换为列联表的形式进行处理。经转换形成的列联表是一个  $n \times p$  的矩阵,其中:第一个变量有  $n$  个取值,第二个变量有  $p$  个取值,或者理解为有  $n$  个观测记录和  $p$  个变量。对应分析就是围绕着这个矩阵进行的。

### 17.2.1 简单对应分析的数学原理

为了要把行变量和列变量关联起来,用两个向量来代表行变量和列变量,分别称为行记分(row score)和列记分(column score)。

令:  $A = [a_{ij}]$  为  $n \times p$  的数据矩阵,行记分为一个  $n$  维向量  $x = [x_i]$ ,列记分为一个  $p$  维向量  $y = [y_j]$ ,称由一个标量和两个向量组成的三元组合  $(r, x, y)$  为对应分析问题  $C_0(A)$  的解,如

$$\text{果它满足条件: } \begin{cases} rx_i = \sum_{j=1}^p \frac{a_{ij} y_j}{a_{i.}} & (i=1, \dots, n) \\ ry_j = \sum_{i=1}^n \frac{a_{ij} x_i}{a_{.j}} & (j=1, \dots, p) \end{cases}$$

此关系式说明三元组合  $(r, x, y)$  需要满足如下 3 条:

- ① 行记分  $x_i$  和列记分  $y_j$  的加权均值成比例;
- ② 列记分  $y_j$  和行记分  $x_i$  的加权均值成比例;
- ③ 数值  $r$  为行列记分的相关(在典型相关的意义上)。

下面记:  $R = \text{diag}(a_{i.})$ ,  $C = \text{diag}(a_{.j})$ ,  $R^{1/2} = \text{diag}(a_{i.}^{1/2})$ , 则上面的条件式可以写为如下形式:  $rx = R^{-1}Ay$ ,  $ry = C^{-1}Ax$ 。这里  $\text{diag}(w_i)$  代表由向量  $w_i$  作为对角线元素的对角矩阵,而  $a_{i.}$ 、 $a_{.j}$  分别代表由  $A$  的行总和、列总和所形成的向量。于是可以验证,上面的两个条件式等价

$$\text{于: } \begin{aligned} rR^{1/2}x &= (R^{-1/2}AC^{-1/2})C^{1/2}y \\ rC^{1/2}y &= (C^{-1/2}A'R^{-1/2})R^{1/2}x = (R^{-1/2}AC^{-1/2})'R^{1/2}x \end{aligned}$$

由此,  $x$  为一个解的条件是如下的特征值问题有解,且最大特征值为 1 是平凡解,两组

$$\text{非零特征值相同: } \begin{aligned} r^2(R^2x) &= (R^{-1/2}AC^{-1/2})(R^{-1/2}AC^{-1/2})'(R^2x) \\ r^2(C^2y) &= (R^{-1/2}AC^{-1/2})'(R^{-1/2}AC^{-1/2})(C^2y) \end{aligned}$$

$$\text{再令: } Z \equiv (R^{-1/2}AC^{-1/2}), v \equiv (R^{-1/2}x, u \equiv C^{-1/2}y), \text{ 则前述的特征值问题可以写为: } \begin{aligned} r^2u &= Z'Zu \\ r^2v &= ZZ'Zv \end{aligned}$$

这是两个特征值问题,它们有同样的非零特征值,如  $U$  是  $Z'Z$  的特征向量,则  $ZU$  是  $ZZ'$  的特征向量,根据线性代数的知识,它们有相同的非零特征根:  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$ 。设  $Z'Z$  相应的特征向量为  $u_1, u_2, \dots, u_n$ ,  $ZZ'$  相应的特征向量为  $v_1, v_2, \dots, v_n$ ,对最大的  $m$  个特征值可得因子载荷阵:

$$F = \begin{bmatrix} u_{11}\sqrt{\lambda_1} & u_{12}\sqrt{\lambda_2} & \dots & u_{1m}\sqrt{\lambda_m} \\ u_{21}\sqrt{\lambda_1} & u_{22}\sqrt{\lambda_2} & \dots & u_{2m}\sqrt{\lambda_m} \\ \vdots & \vdots & \ddots & \vdots \\ u_{p1}\sqrt{\lambda_1} & u_{p2}\sqrt{\lambda_2} & \dots & u_{pm}\sqrt{\lambda_m} \end{bmatrix}, G = \begin{bmatrix} v_{11}\sqrt{\lambda_1} & v_{12}\sqrt{\lambda_2} & \dots & v_{1m}\sqrt{\lambda_m} \\ v_{21}\sqrt{\lambda_1} & v_{22}\sqrt{\lambda_2} & \dots & v_{2m}\sqrt{\lambda_m} \\ \vdots & \vdots & \ddots & \vdots \\ v_{n1}\sqrt{\lambda_1} & v_{n2}\sqrt{\lambda_2} & \dots & v_{nm}\sqrt{\lambda_m} \end{bmatrix}$$

矩阵 F 的头两列（等价于取  $m = 2$ ）所组成的散点图，与矩阵 G 的头两列所组成的散点图叠加，就形成了对应分析图。由于各种模型的选项不同，实际的点图和这两组载荷向量所构成的图形可能会有所不同，但这种不同不会影响对数据进行探索性分析的结果。

## 17.2.2 SPSS 简单对应分析实例

### 1. 问题和数据描述

1992 年美国大选，克林顿击败了老布什和佩罗当选总统，本节来分析一下在这次选举中，不同教育程度选民的选举倾向性有何特点。数据来源于 SPSS 自带的数据集“vote.sav”，整理后的数据格式如图 17-1 所示，所用数据文件为“92 年美总统选举数据.sav”。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	pr92	Numeric	8	0	候选人	{1, 老布什}...	None	4	Right	Scale
2	degree	Numeric	1	0	受教育程度	{0, 高中以下}, 7, 8, 9		6	Right	Nominal
3	age	Numeric	2	0	投票人年龄	None	0, 98, 99	8	Right	Scale
4	agecat	Numeric	8	2	年龄段	{1.00, 不满35岁}	None	8	Right	Nominal
5	educ	Numeric	2	0	受教育年限	None	97, 98, 99	6	Right	Scale
6	sex	String	8		性别	{male, 男}...	None	7	Left	Nominal

图 17-1 1992 年美国大选抽调数据

注意：对应分析适合的变量只能是数值型的无序分类变量，否则需先行进行转换。如果分析变量是有序的，可以使用分类主成分分析法。

### 2. 简单对应分析的操作和设置

依次单击菜单“Analyze→Data Reduction→Correspondence Analysis...”打开对应分析对话框，其主设置面板如图 17-2 所示。

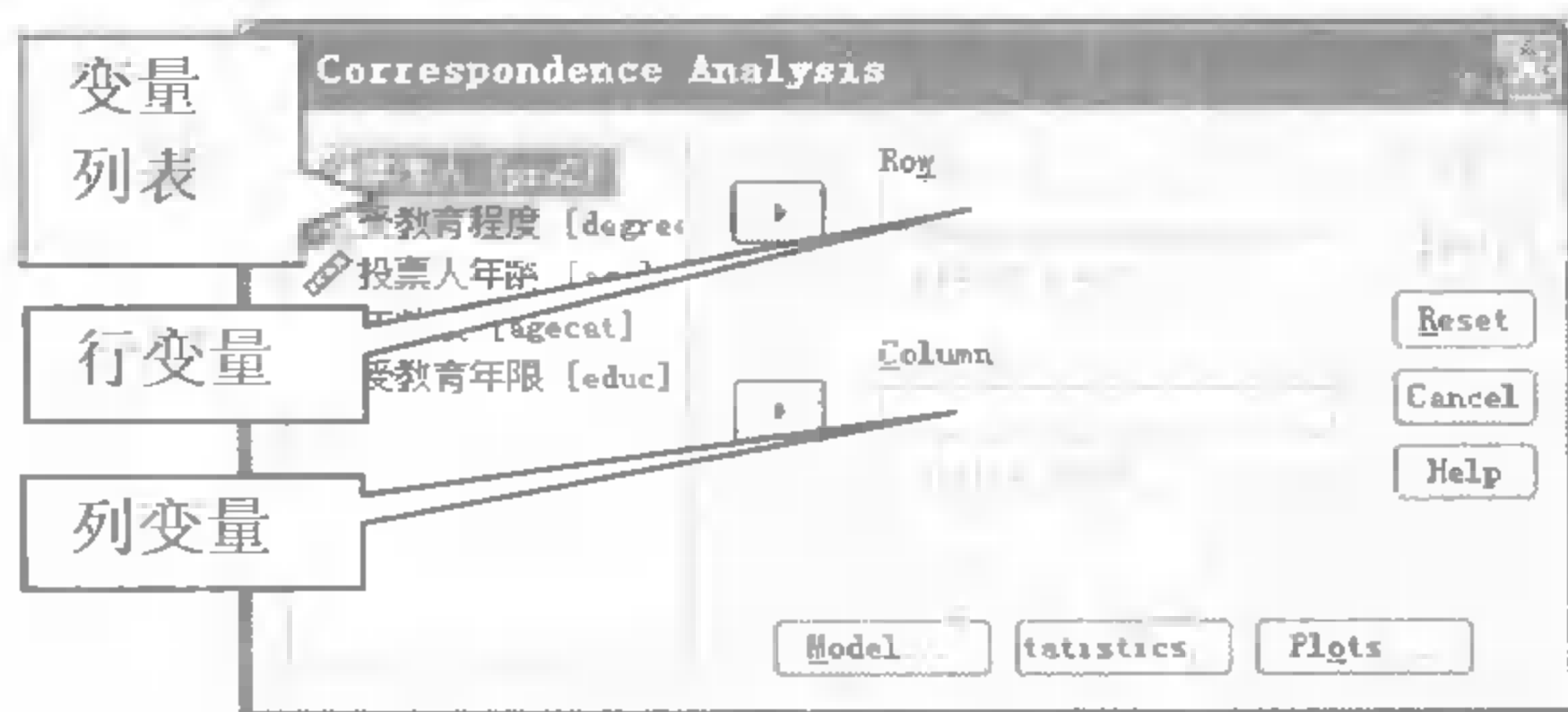




图 17-2 对应分析主设置面板 1

(1) 变量设置。首先在变量列表选中候选人变量，然后单击 Row 左侧的  按钮，将其选入 Row 选框；接着在变量列表选中受教育程度变量，然后单击 Column 左侧的  按钮，将其选入 Column 选框。单击选中 Row 选框，单击它下面的 Define Range 按钮，弹出图 17-3 所示的设置界面，在 Minimum、Maximum 后面分别输入 1、3 后，单击 Update 确认，再单击 Continue 按钮返回主界面；单击选中 Column 选框，单击它下面的 Define Range 按钮，弹出与图 17-3 相同的设置界面，在 Minimum、Maximum 后面分别输入 0、4 后，单击 Update 确认，再单击 Continue 按钮返回主界面。设置好后的界面如图 17-4 所示。

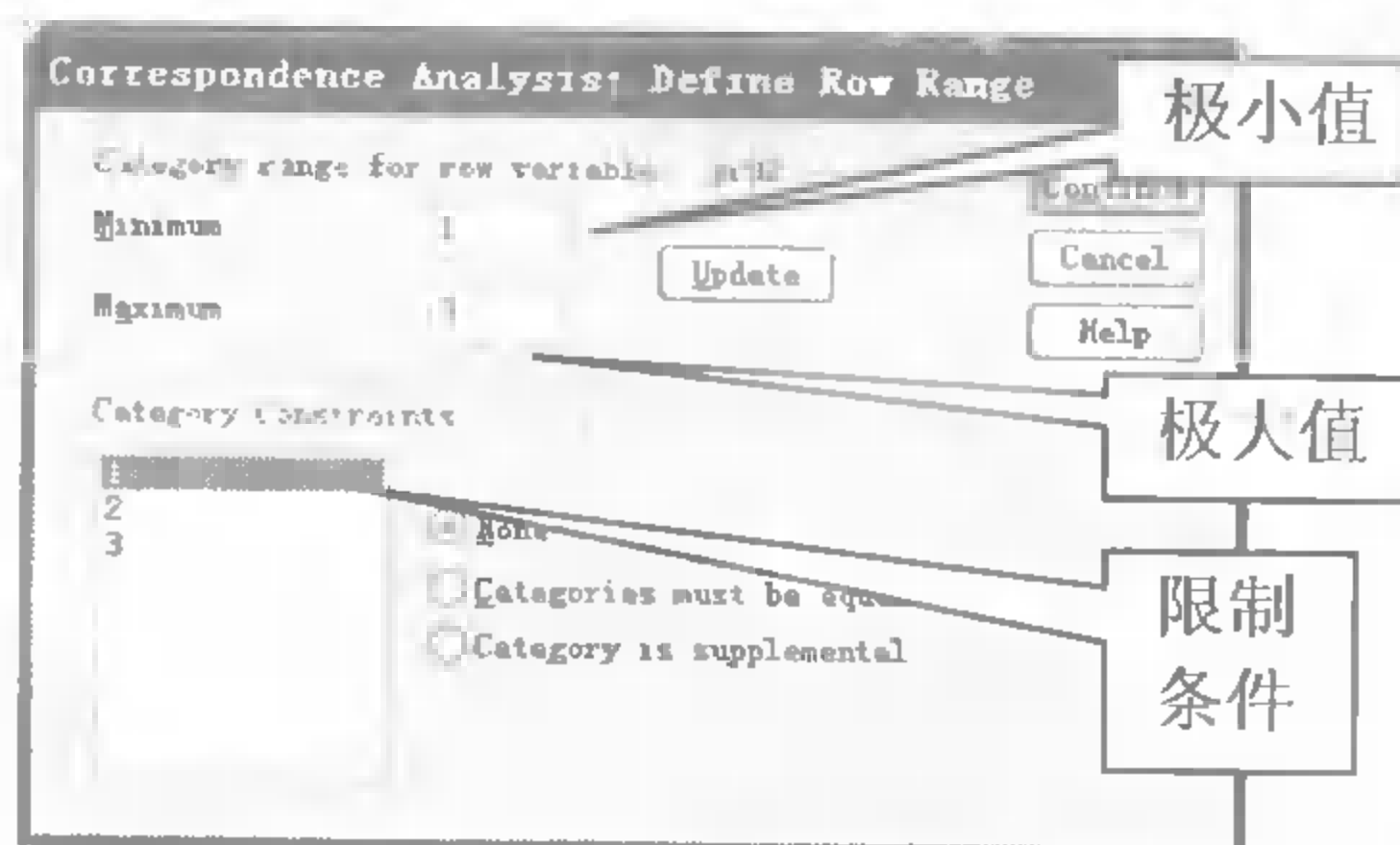


图 17-3 对应分析的 Define Range 设置界面

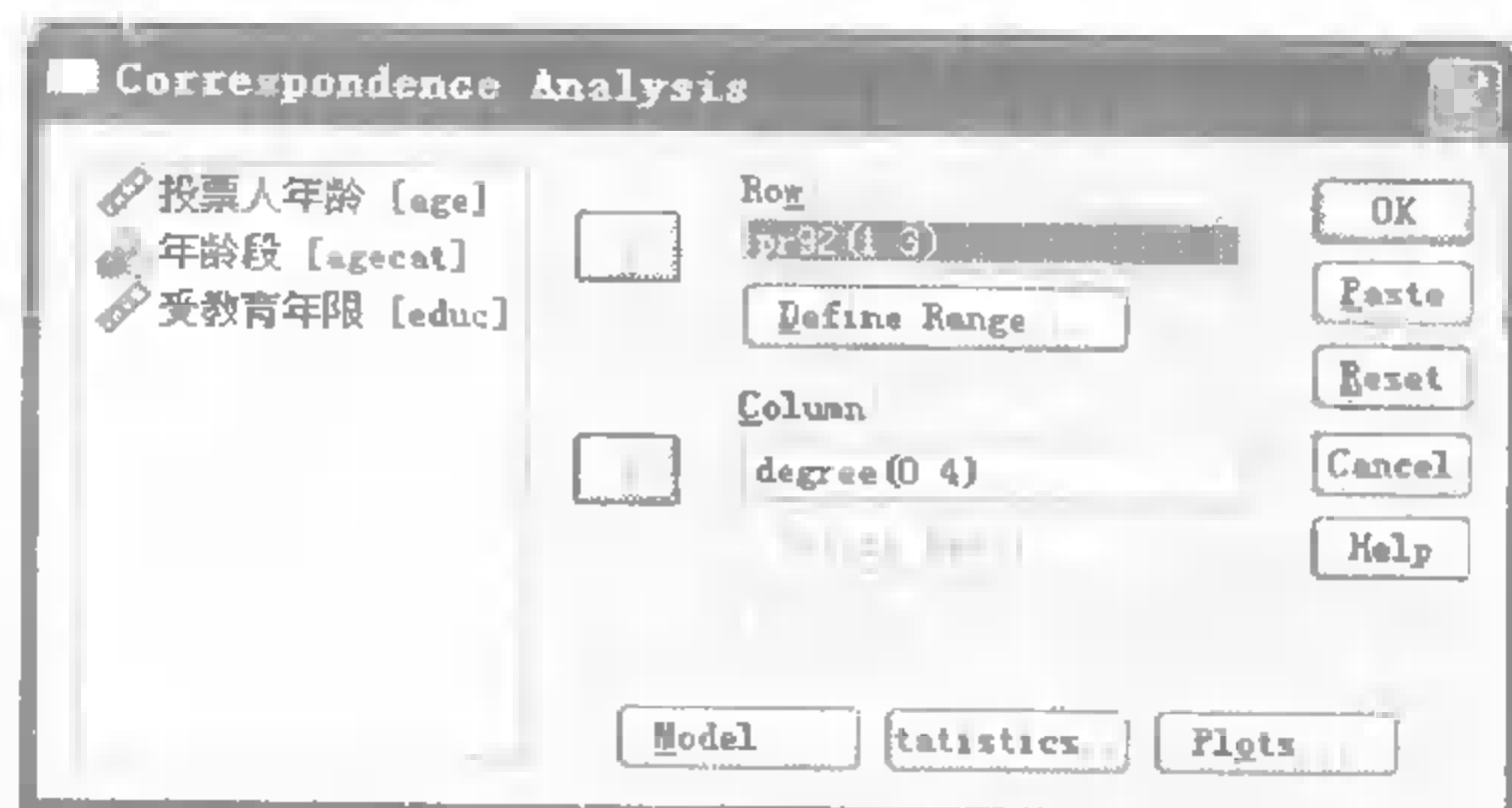


图 17-4 对应分析主设置面板 2

① 主界面的 Row、Column 选框，分别表示对应分析的行变量和列变量。

② 在 Define Range 对话框里，Minimum、Maximum 输入框分别表示相应分类变量的最小值、最大值，此处只能输入整数，其输入后须单击 Update 按钮加以确认，数据集里超出此处设置范围的记录在分析时将被忽略。

③ Define Range 对话框里的 Category Constraints 栏设置分类变量取值的约束条件，其下的列表框显示的是当前分类变量（行变量或列变量）的取值列表。从取值列里选中一个值，通过单击右侧的三个单选框设置其约束条件，如果分类值所代表的分类不符合对应分析的需要，或者其分类的界限是模糊的，就可以使用约束条件对这些取值进行明确的划分。3 个可选约束条件含义如下。

- None 不作任何约束，分类数据保持原状，此项为默认选项；
- Categories must be equal 等同约束，表示各类别必须有相同的得分。如果分类顺序是不合常理或违反直觉的，选择此项约束。行约束的最大个数为行分类总数减 1。要对分类取值进行分组等同约束，需要使用命令语句，例如：变量取值为 1、2、3、4，若设置类别 1、2 满足等同约束，同时类别 3、4 满足等同约束，在命令语句中的参数设置选项为：/EQUAL=VAR((1 2), (3 4))。
- Category is supplemental 增补约束，增补的种类不影响分析过程和种类维数，但会在有效种类的定义空间里被描述。列分类变量的最大增补个数为列分类的总数减 2。

(2) 模型设置。在图 17-4 中，单击 Model 按钮，弹出如图 17-5 所示的模型设置子面板。Dimensions in solution 输入框显示的当前维数为 2，可自行更改；其他选项均采用默认设置；单击 Continue 按钮返回主界面。

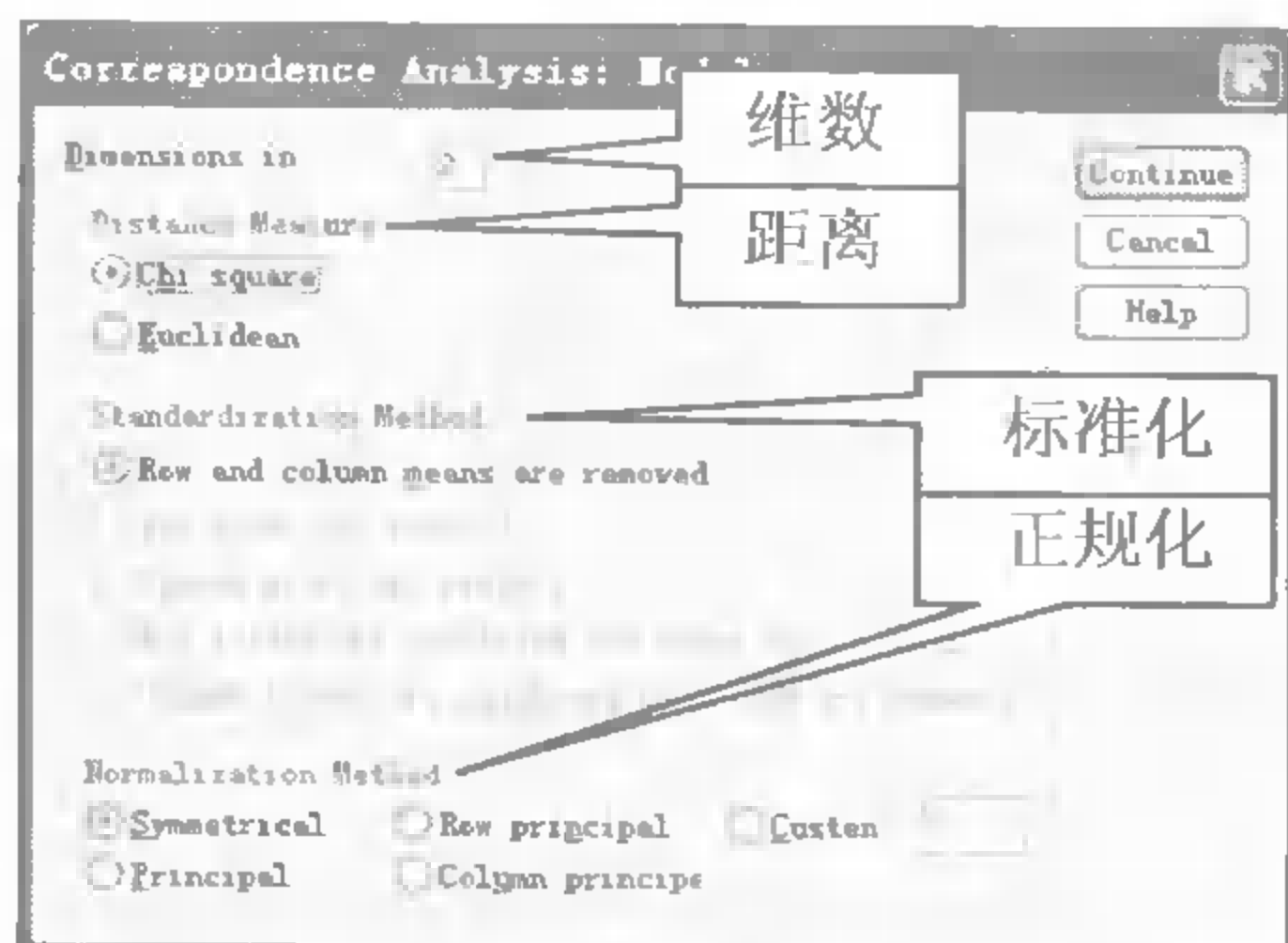


图 17-5 对应分析的模型设置

① Dimensions in solution 输入框, 指定对应分析的维数。建议选择尽可能小的维数来解释所有变差。最大维数取决于有效分类的数目和等同约束的数目, 可能为如下两种情况。

- 有效的行种类减去行等同约束的分类数目, 再加上行约束分类的集合数目;
- 有效的列种类减去列等同约束的分类数目, 再加上列约束分类的集合数目。

② Distance Measure 栏, 选择行、列分类各自的距离测度, 有以下两个选项。

- Chi square 卡方距离, 默认方法。用于度量分类变量间的距离, 此方法适用于标准对应分析。以  $i$ 、 $j$  作为下标来区分变量,  $m$  表示分类的个数, 则  $i$ 、 $j$  之间的

卡方距离按如下公式进行计算:  $d_{ij} = \sum_{k=1}^m \{(x_{ik} - e_{ijk})^2 / e_{ijk} + (x_{jk} - e_{jik})^2 / e_{jik}\}$ , 其中:

$$e_{ijk} = (x_{ik} - x_{jk})T_i / T_{ij}, T_i = \sum_{k=1}^m x_{ik}, T_{ij} = T_i + T_j (k=1, 2, \dots, m; i, j=1, 2, \dots, n)。$$

- Euclidean 欧氏距离, 用于度量分类变量或连续变量间的距离, 取两行 (或两列) 之

间差的平方和再开方作为距离。计算公式为:  $d_{ij} = \sqrt{\sum_{k=1}^m (x_{ik} - x_{jk})^2}$ 。

③ Standardization Method 栏, 设置标准化的方法, 有如下 5 个选项, 其中后 4 个只能在 Distance Measure 栏选中欧氏距离后才可用。

- Row and column means are removed 行、列数据都被中心化 (减去其均值), 当选择卡方距离时, 只能指定该方法;
- Row means are removed 只有行数据被中心化;
- Column means are removed 只有列数据被中心化;
- Row totals are equalized and means are removed 行数据被中心化, 且确定中心之前, 先令行边际都相等;
- Column totals are equalized and means are removed 列数据被中心化, 且确定中心之前, 先令列边际都相等。

④ Normalization Method 栏设置正规化的方法, 有如下 5 个选项。

- Symmetrical 对称法, 对于每个维度, 行得分是列得分除以匹配奇异值的加权平均, 列得分是行得分除以匹配奇异值的加权平均。此方法可以用来检查两个变量之间的差异性或相似性。
- Principal 方法, 如果要检查行或列变量各自内部分类间的距离, 而不是检查行、列间的距离, 选用此方法。
- Row principal 方法, 如果要检查行变量内部分类间的距离, 选用此方法。
- Column principal 方法, 如果要检查列变量内部分类间的距离, 选用此方法。
- Custom 用户自定义, 在输入框指定一个介于 -1 和 1 之间的数字。值 -1 相当于 Column principal 方法, 值 1 相当于 Row principal 方法, 值 0 相当于 Symmetrical 方法。其他值用于指定行、列得分的比重, 本方法适用于输出特定的二维图形。

(3) 输出统计量设置。在图 17-4 中单击 Statistics 按钮, 弹出如图 17-6 所示的统计量设置子面板, 默认 Correspondence table、Overview of row points、Overview of column points 这 3 个复选框处于选中状态; 单击 Continue 按钮返回主界面。



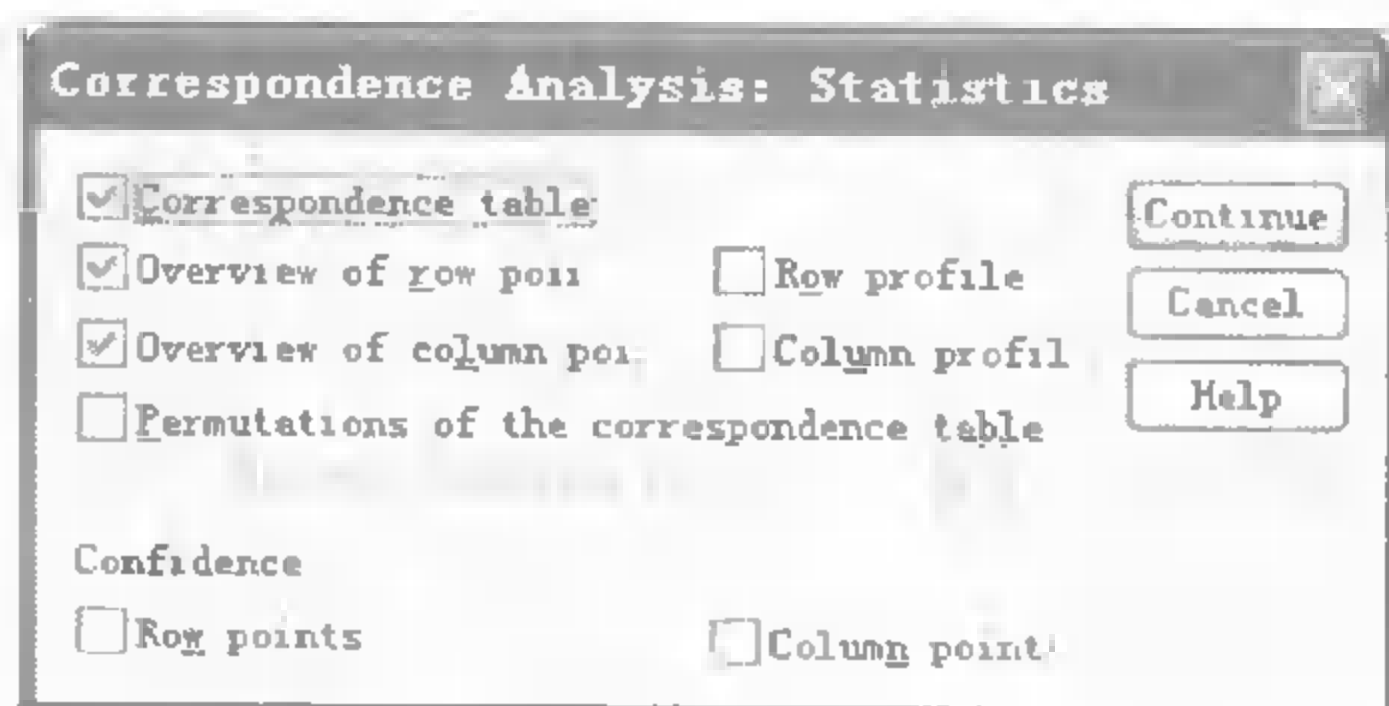


图 17-6 对应分析的统计量设置

此面板用来设置对应分析输出哪些表格，有如下 5 部分内容可选。

- ① Correspondence table 交叉分组列表，含行、列变量的边际总和。
- ② Overview of row points 行详细信息表，包括行变量各分类的得分、质量、惯量、对维度的惯量贡献、维度对分数惯量的贡献。选中 Row profiles 复选框，还会显示每个行变量分类对所有列变量分类的分布情况。
- ③ Overview of column points 列详细信息表，包括列变量各分类的得分、质量、惯量、对维度的惯量贡献、维度对分数惯量的贡献。选中 Column profiles，还会显示每个列变量分类对所有行变量分类的分布情况。
- ④ Permutations of the correspondence table 复选框，输出按照第一个维度上的得分升序排列的行、列对应表。还可以在 Maximum dimension for 输入框指定表格的最大维数，输出结果包括从 1 到指定最大维度的交叉表。
- ⑤ Confidence Statistics for 栏有两个选项：Row points 为所有非增补行输出标准差和相关系数；Column points 为所有非增补列输出标准差和相关系数。

(4) 绘图选项设置。在图 17-4 中，单击 Plots 按钮，弹出如图 17-7 所示的绘图设置子面板，默认情况下 Biplot 复选框、Display all dimensions in the solution 单选框处于选中状态，单击相应的选项可以更改设置；单击 Continue 按钮返回主界面。

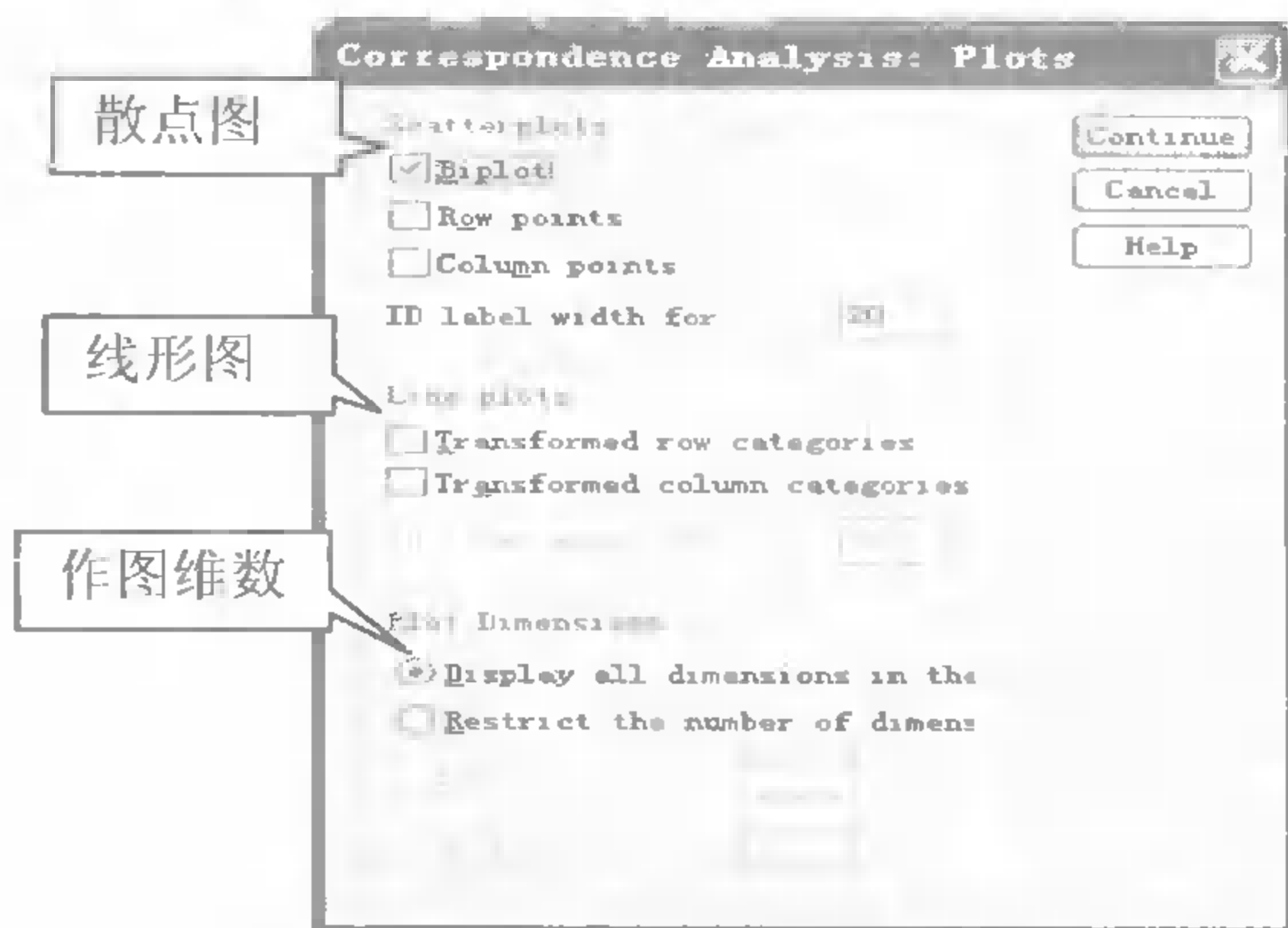


图 17-7 对应分析的绘图设置

此面板选择对应分析输出哪些图形，设置内容包括如下 3 种类型的图形。

- ① Scatterplots 散点图，以矩阵形式输出成对行、列分类取值的散点图。

Biplot 输出行、列的联合分布图，如果选择了 Principal 正规化方法，此选项不可用；Row points 以矩阵形式输出每个行分类的得分图；Column points 以矩阵形式输出每个列分类的得分图。ID label width for scatterplots 输入框，设置散点图 ID 标签的字符个数，默认为 20。

② Line plots 线形图，输出指定变量的每个维度的线形图。

● Transformed row categories 复选框，以行分类的原始取值对行分类的得分作图。

● Transformed column categories 复选框，以列分类的原始取值对列分类的得分作图。

③ Plot Dimensions 栏设置输出维度，对所有输出的多维图形有效。

Display all dimensions in the solution 单选框，表示分析用到的行、列维度都将以交叉矩阵的形式输出；Restrict the number of dimensions 单选框，限制输出指定维度组合的图形，必须在 Lowest、Highest 后指定最小、最大维度，最小维度可以从 1 到总维度数目减 1 的整数，最大维度可以从 2 到总维度数目的整数。

### 3. 结果分析

在图 17-4 中单击“OK”按钮运行，SPSS Viewer 窗口的输出结果如图 17-8～图 17-12 所示。

信号	
CORRESPONDENCE	
Version 1.1	
by	
Data Theory Scaling System Group (DTSS)	
Faculty of Social and Behavioral Sciences	
Leiden University, The Netherlands	

图 17-8 版权信息

对应表						
候选人	受教育程度					有效边际
	高中以下	高中	大专	学士	硕士	
老布什	55	349	48	146	63	661
佩罗	12	159	26	62	19	278
克林顿	122	439	58	178	111	908
有效边际	189	947	132	386	193	1847

图 17-9 版权信息和对应表

摘要								
维	奇异值	惯量	卡方	显著性	惯量比例		置信奇异值	
					考虑情况	累积	标准差	相关性
								2
1	138	019			987	987	021	061
2	016	000			013	1 000	024	
总计		019	35 516	000 <sup>a</sup>	1 000	1 000		

a. 8 自由度

图 17-10 对应分析结果摘要

概述行点 <sup>a</sup>									
候选人	质量	维中的得分		惯量	贡献				
		1	2		点对维惯量		维对点惯量		总计
					1	2	1	2	
老布什	358	193	-157	002	097	545	929	071	1 000
佩罗	151	664	198	009	481	368	990	010	1 000
克林顿	492	-344	053	008	422	087	997	003	1 000
活动总计	1 000			019	1 000	1 000			

a. 对称标准化

概述列点 <sup>a</sup>									
受教育程度	质量	维中的得分		惯量	贡献				
		1	2		点对维惯量		维对点惯量		总计
					1	2	1	2	
高中以下	392	-897	087	011	538	048	999	001	1 000
高中	513	169	018	002	106	010	999	001	1 000
大专	071	362	344	001	068	525	905	095	1 000
学士	209	153	-174	001	036	394	869	131	1 000
硕士	104	-103	-059	004	192	023	998	002	1 000
活动总计	1 000			019	1 000	1 000			

a. 对称标准化

图 17-11 行、列详细信息表

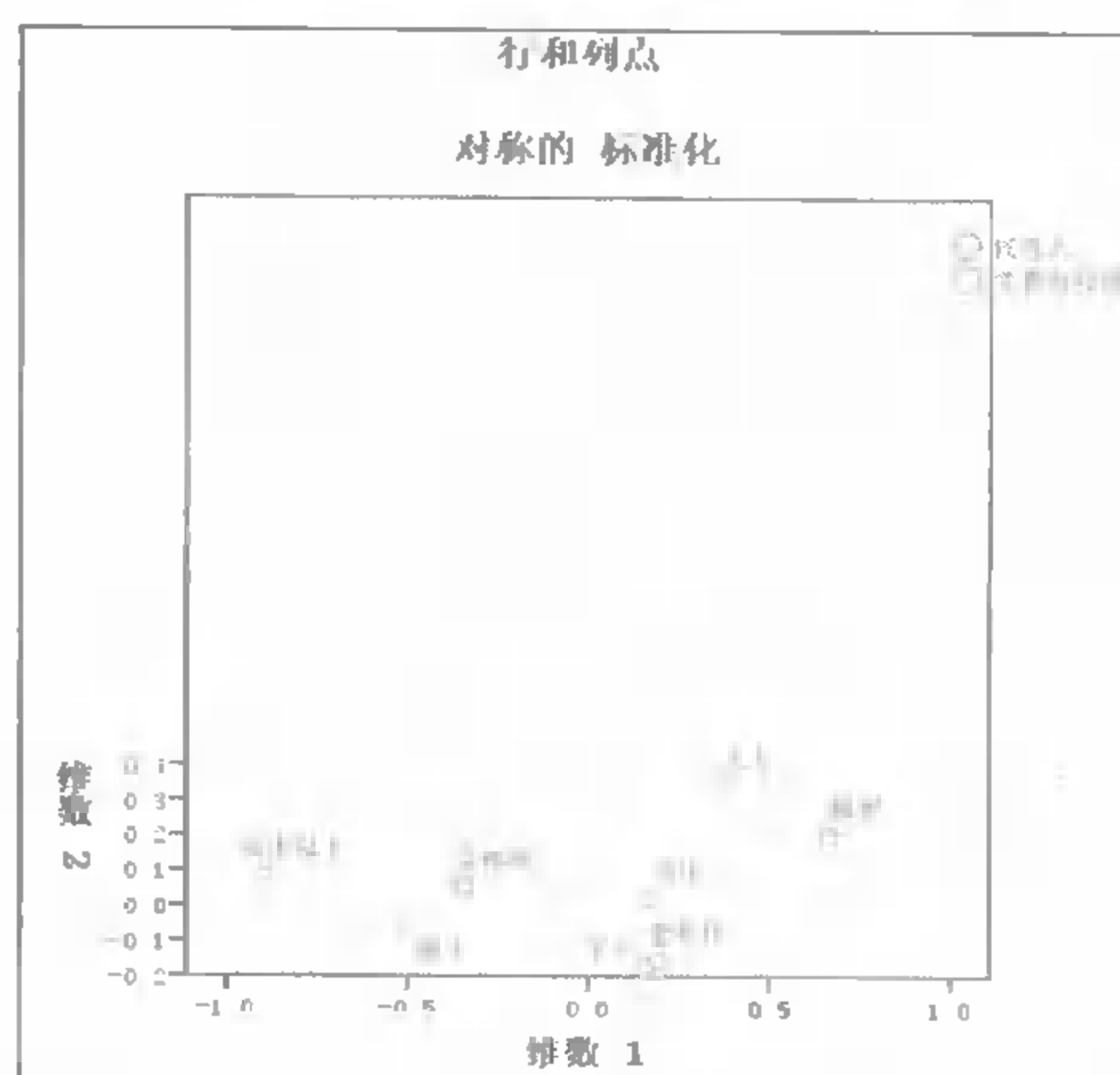


图 17-12 二维对应分析图

(1) 版权信息。图 17-8 所示的“信誉”表格是 SPSS 对应分析模块的版权信息，说明该模块是由荷兰 Leiden 大学 DTSS 课题组编制的，SPSS 通过合同对该程序进行了套装，所以每次都会显示该信息。

(2) 版权信息和对应表。如图 17-9 所示，“对应表”反映了两个变量各类别组合的基本情况，它还可用于检查是否存在数据录入错误。从此表来看，高中学历的人群投克林顿票的人较多。

(3) 对应分析结果摘要表。如图 17-10 所示，是整个对应分析的结果汇总表，它是输出中最为重要的一个，主要用于确定使用多少个维度来对结果进行解释。其中，奇异值就是惯量的平方根，相当于相关分析里的相关系数；而惯量就是常说的特征根，用于说明对应分析的各个维度，能够解释列联表的两个变量之间相互联系的程度。

第一维惯量值为 0.019，第二维为 0.000 26，右侧给出了它们各占的百分比，说明其分别解释了总信息量的 98.7% 和 1.3%，因此二维图形可以完全表示两变量间的信息，并且观察时以第一维度为主。

(4) 行、列详细信息表。如图 17-11 所示，在“概述行点”表格里，质量列为每一类别所占总体的百分比；随后的两列则为坐标值，可见 3 个候选人在第一维上分散的比较好；右侧给出了每个类别对各个维度的贡献量，包括点对维度惯量的贡献和维度对点惯量的贡献两种。“概述列点”表格的结构与“概述行点”表格类似，解读方法也一样。

(5) 对应分析图。如图 17-12 所示，给出的是对应分析图，观察此图遵循如下两步。

① 首先检查各变量在横轴和纵轴方向上的区分情况，如果同一变量不同类别在某个方向上靠得较近，则说明这些类别在该维度上区别不大；

② 然后比较不同变量各个分类间的位置关系，落在邻近区域内的不同变量的分类点，彼此之间的相互联系较为紧密。

本例中，两个变量在第一维度上分的都很开，第二维度区分效果不明显，这和第③节中提到的变异以第一维度为主一致。在投票的倾向性上，高中和学士学历的人更青睐老布什，而研究生学历的人更倾向于克林顿，对佩罗最感兴趣的属大专生了。

### 17.3 多元对应分析

最优尺度分析 (Optimal Scaling) 是独立发展起来的，与对应分析相互独立的两类方法，

只不过它也可以进行多元对应分析,SPSS 的多元对应分析功能正是使用 Optimal Scaling 过程来执行的。SPSS 的最优尺度分析提供了同质性分析、分类变量的主成分分析和非线性典型相关分析三种方法,以满足不同的数据要求,而同质性分析就是本节要介绍的多元对应分析。

多元对应分析的核心目的与简单对应分析相似,即力图在低维度空间描述两个或多个变量之间的关系,这些变量以分类变量为主,也可以为连续性变量。

### 17.3.1 多元对应分析基本概念及其特点

对多个定性变量的研究,其计算方法与两变量时的情况基本相同。多元对应分析的计算结果亦与简单对应分析有相同的特性,比如:有关行分析与列分析的结果是互为对偶的;同秩的主轴对应相同的特征值。因此多元对应分析的计算结果,在低维平面图上亦是可以做叠加观察和分析的。

多元对应分析要比简单对应分析更进一步,主要表现在以下 3 个方面。

(1) 可以同时分析多个分类变量之间的关系,并同样用图形方式表示出来。

(2) 能够处理的变量种类更加丰富,例如可以对无序多分类变量、有序多分类变量和连续型变量同时进行分析。

(3) 最优尺度分析还对多选题的分析提供了支持。

但是,SPSS 的最优尺度分析过程不像多元回归过程一样能够自动筛选变量,因此变量较多时容易使图形显得混乱,可能会掩盖掉真实的变量联系。此时,需要用户根据经验和分析结果进行耐心的筛选,以得到最优结果,这对使用者的分析水平是一个严峻的考验。

### 17.3.2 多元对应分析的参数设置

本节仍然使用如图 17-1 所示的美总统选举数据,这次除了考虑候选人、受教育程度两个变量外,再加入性别和年龄段两个因素。

注意:多元对应分析过程要求把字符串变量按照字母顺序转换为正整数;用户和系统定义的缺失值,以及小于 1 的取值都会被当作缺失值对待;数据集必须含有至少 3 个观测;所有分析变量都需为多分类的名义变量。

与简单对应分析相比,多元对应分析的设置较为复杂,下面逐步加以详细介绍。

#### 1. 分析方法的选择

依次单击菜单“Analyze→Data Reduction→Optimal Scaling...”打开最优尺度分析的选择对话框,如图 17-13 所示,在此选择采用何种最优尺度分析方法。

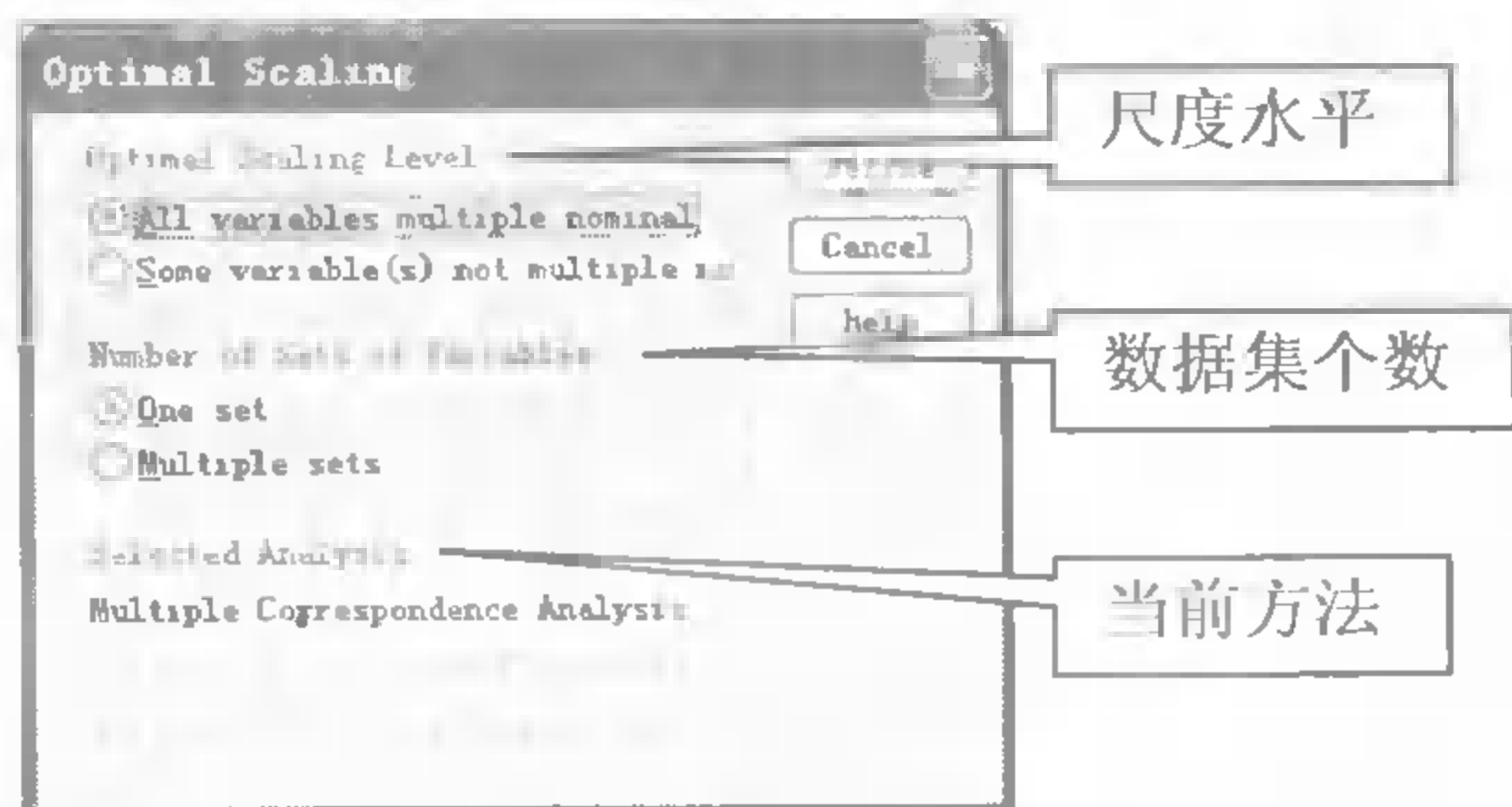


图 17-13 最优尺度分析选择对话框



(1) Optimal Scaling Level 栏, 设置变量的度量类型。

如果所有变量均为无序多分类(名义变量), 应选择 All variables multiple nominal 单选框; 如果有的变量是单分类的名义变量、有序分类变量或者离散的数值型变量, 应选择 Some variable(s) not multiple nominal 单项框。

(2) Number of Sets of Variables 栏, 设置变量集的个数。

One set 选项表示只分析一组变量间的关系; 如果数据集中存在多选题变量集, 即有多个变量是同一道多选题的不同答案, 应当选择 Multiple sets 单选框。

(3) Selected Analysis 栏, 显示当前选项所使用的分析方法, 不可编辑, 有如下 3 种。

- ❶ Multiple Correspondence Analysis 多元对应分析, 当选择 All variables multiple nominal 和 One set 两个选项时使用该方法。它用于分析多个无序分类变量间的关系, 并使用散点图表示出来, 分析过程与简单对内分析非常相似, 但分析的变量可以为多个。默认使用此方法。
- ❷ Categorical Principal Components Analysis(CatPCA)分类变量的主成分分析, 当选择 Some variable(s) not multiple nominal 和 One Set 两个选项时使用该方法。它使用尽量少的主成分来解释尽可能多的原始信息。它就是市场研究中非常重要的多维偏好分析。
- ❸ Nonlinear Canonical Correlation Analysis(OVERALS)非线性典型相关方法, 只要选择了 Multiple Sets 就只会使用此方法。它用于分析两个或多个变量集之间的关系, 允许变量为任何类型: 无序分类、有序分类或连续变量。

## 2. 多元对应分析的操作和设置

在图 17-13 里, 选中 All variables multiple nominal 和 One set 两个选项, 单击 Define 按钮, 弹出如图 17-14 所示的多元对应分析主设置面板。

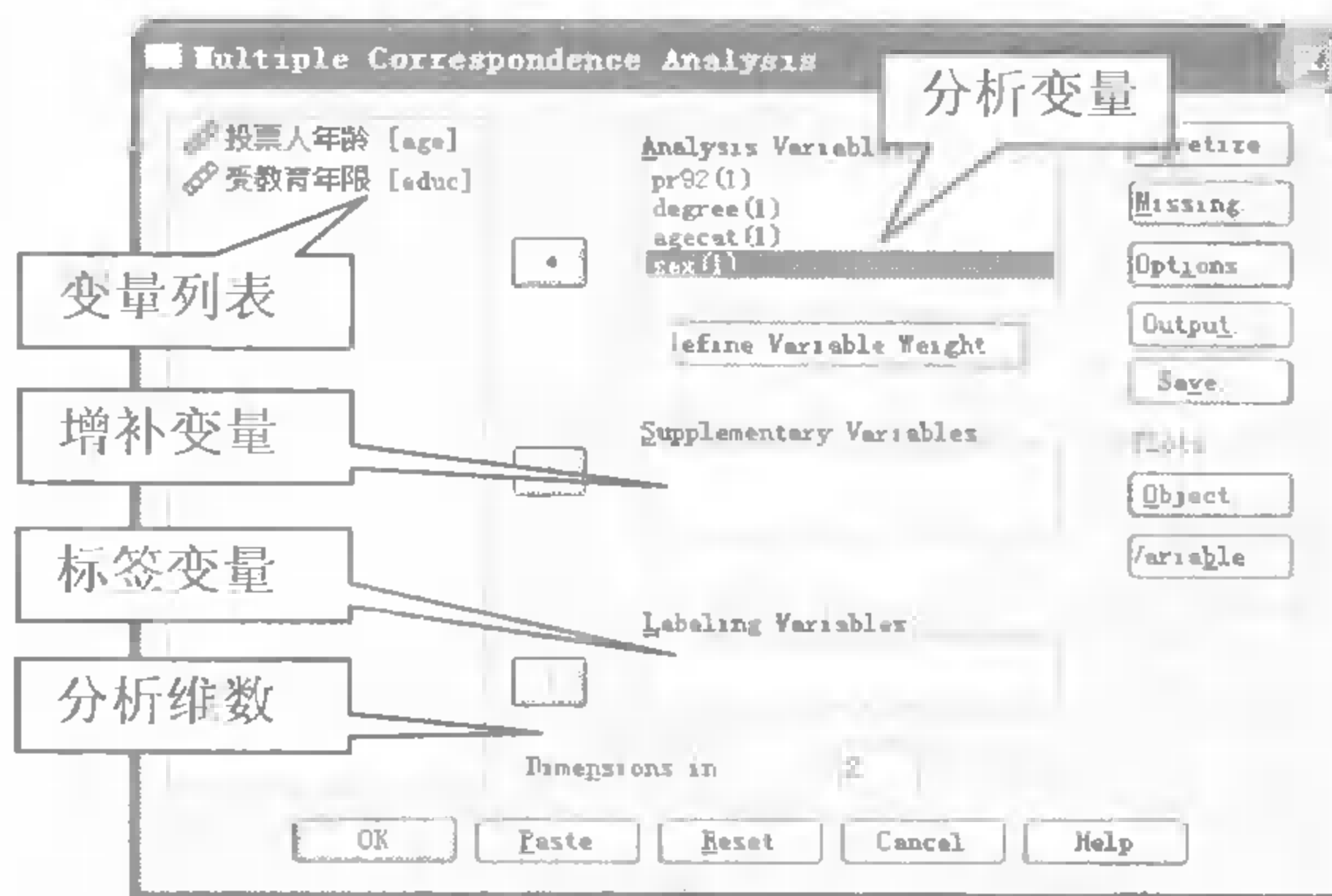



图 17-14 多元对应分析主设置界面

(1) 变量设置。首先在变量列表选中候选人、受教育程度、年龄段、性别这 4 个变量, 然后单击最上面的  按钮, 将其作为分析变量选入 Analysis Variables 列表; 在分析变量列表选中 pr92 变量后, 点击下面的 Define Variable Weight 按钮, 弹出如图 17-15 所示的变量权重设置对话框, 默认权重为 1, 单击 Continue 按钮返回主界面; 其他变量的权重设置方法与此相同, 都采用默认权重 1。

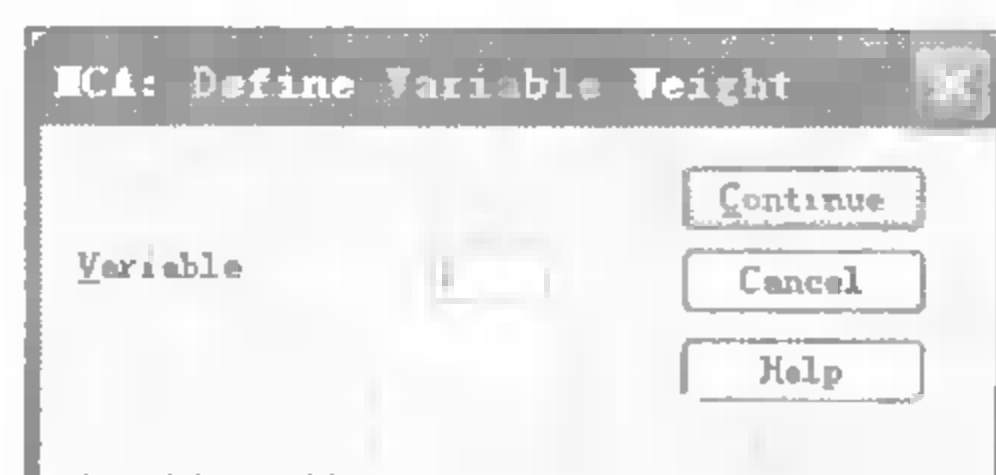


图 17-15 变量权重

- Analysis Variables 分析变量列表，当只选入两个变量时，本方法就相当于 17.2 节的简单对应分析；变量名的显示格式为“name (n)”，其中 name 为变量名，n 为其当前权重；权重通过单击 Define Variable Weight 按钮进行设置。
- Supplementary Variables 增补变量列表，增补变量不用于分析，只用于结果对比和描述。
- Labeling Variables 标签变量，用于在结果里标识各个记录。
- Dimensions in 输入框，设置描述分析结果的低维空间维数，默认为 2。

(2) 分析变量的离散化设置。在图 17-14 中单击 Discretize 按钮，弹出如图 17-16 所示的变量离散化 (Discretization) 设置对话框。在变量列表选中某个变量后，通过 Method 下拉列表选择对其离散化的方法，默认为 Unspecified (不离散化)，修改后需单击 Change 按钮确认；单击 Cancel 按钮（如果做了修改，单击 Continue 按钮）返回主界面。

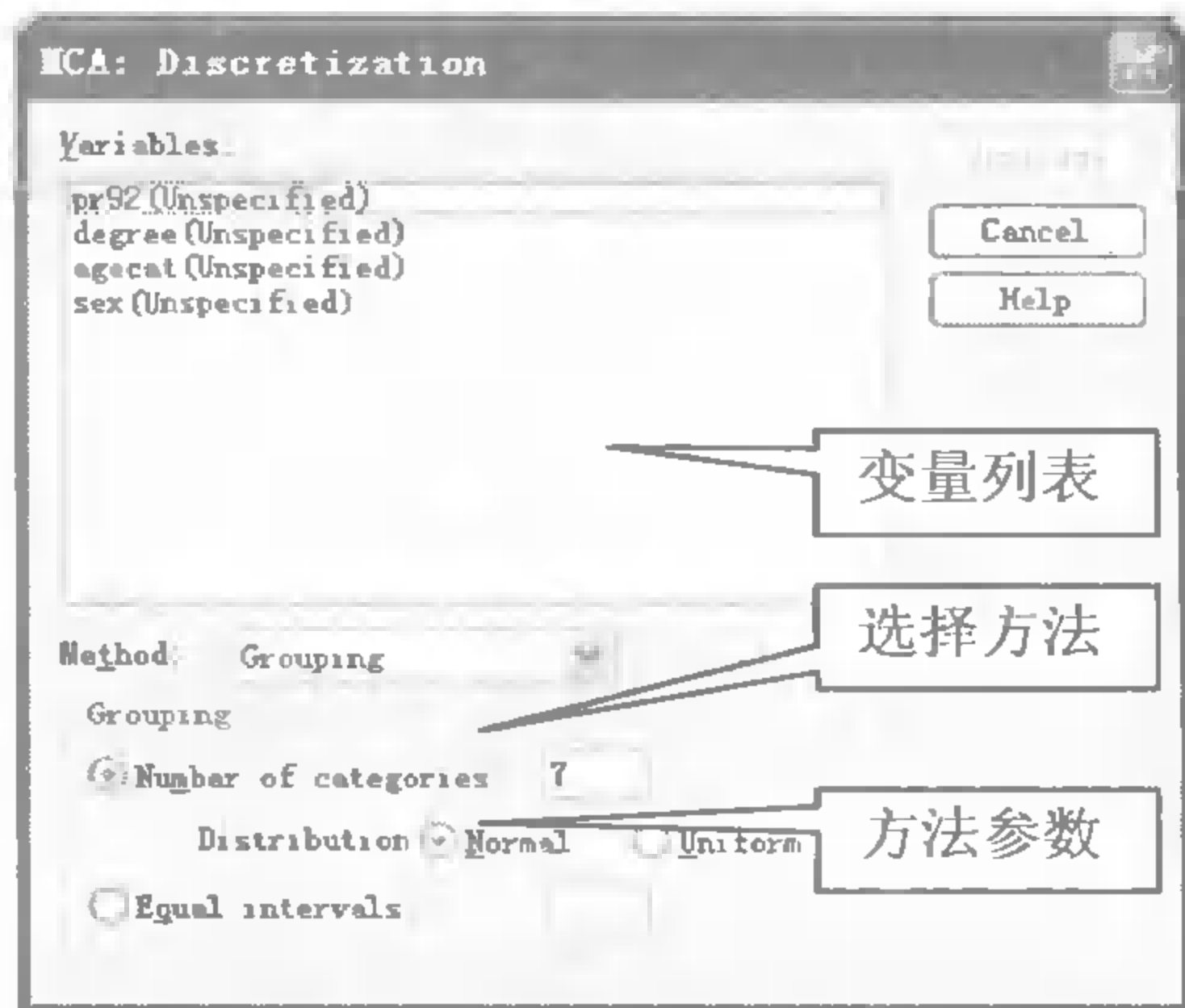


图 17-16 离散化设置对话框

由于多元对应分析要求所有分析变量都是多分类的名义变量，且取值需为正整数，所以有必要对不符合这些条件的变量进行离散化。下面详细介绍各设置选项的含义。

① Variables 变量列表显示的变量格式为“name (f)”，其中 name 为变量名，f 为其当前的离散化方法。

② Method 下拉菜单中，可选的离散化方法有如下 4 个。

- Unspecified 无离散化操作。
- Grouping 分组，将取值重新编码为固定个数或者固定间隔的类别。选中后，Grouping 栏的设置项变为可编辑的，在 Number of categories 后指定分类的个数，同时指定变量的分布情况 (Distribution) 是正态分布 (Normal) 还是均匀分布 (Uniform)；或者，在 Equal intervals 后指定自动分类的间隔大小。
- Ranking 排序，将变量取值排序后，取其秩进行分类。
- Multiplying 加乘，将当前取值标准化后，乘以 10，再取整，最后加上一个常数，使得离散化后的最小值为 1。

(3) 缺失值的处理方式。在图 17-14 中单击 Missing 按钮，弹出如图 17-17 所示的缺失值设置对话框。在 Analysis Variables 或 Supplementary Variables 列表选中某个变量后，再在 Strategy 栏单击相应的单选框更改其缺失值的处理方法，并单击 Change 按钮确认。单击 Cancel 按钮（如果做了修改，单击 Continue 按钮）返回主界面。

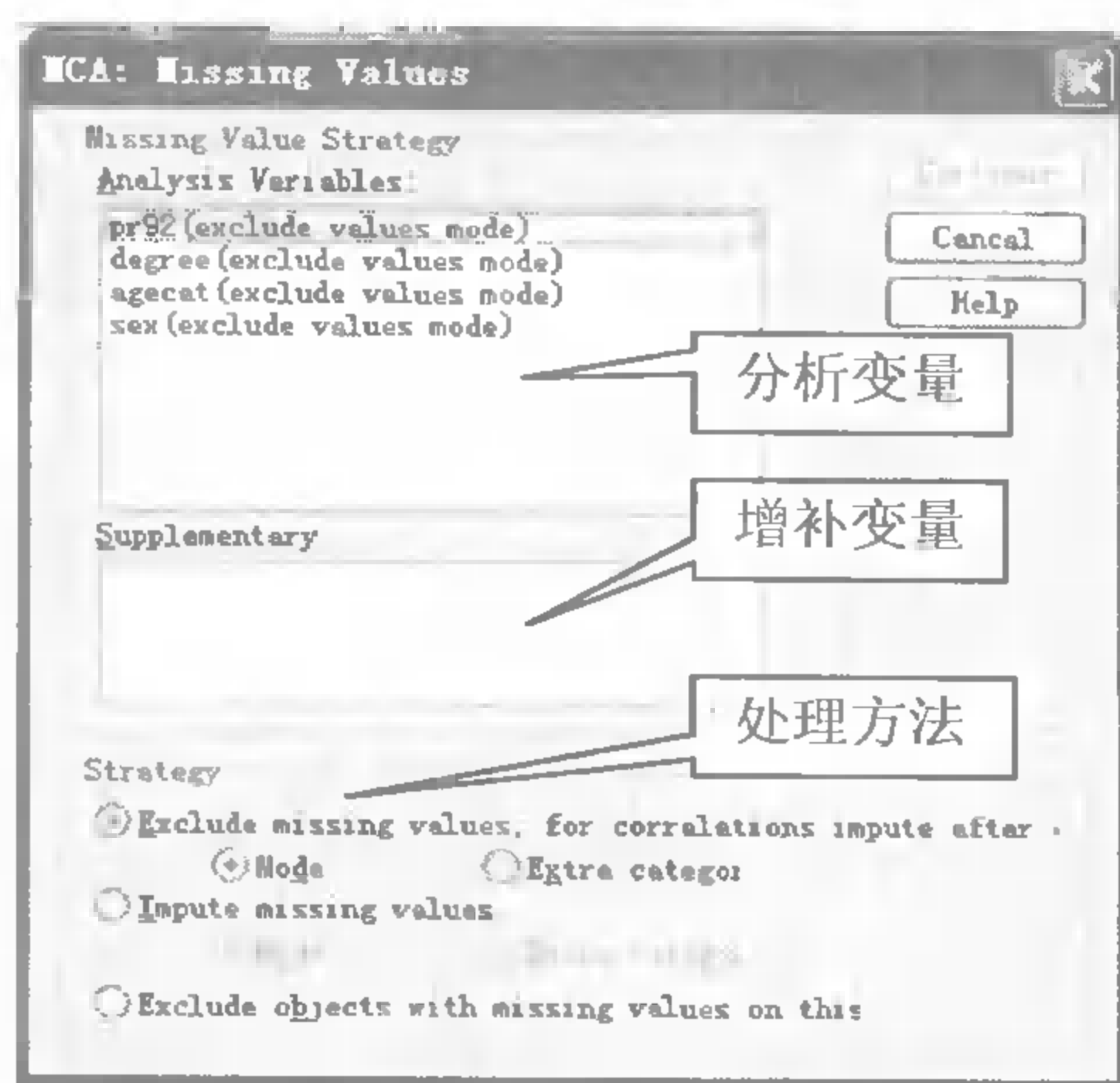


图 17-17 缺失值设置对话框

① Analysis Variables 列表、Supplementary Variables 列表分别显示了当前选入的分析变量和增补变量。变量的显示格式为“name (f)”，其中 name 为变量名，f 为其当前的缺失值处理方法。

② Strategy 栏提供了如下 3 种缺失值处理方法。

- Exclude missing values 排除含缺失值的变量，是被动处理方式。所选变量含缺失值的观测对此变量的分析不作贡献；如果所有变量都采用此方法，则这些变量取值都为缺失的观测，并将被作为增补对象处理。如果输出相关矩阵，则分析后缺失值的替换方式有两种，Mode 表示用类别取值的众数来取代缺失值，如果存在多个众数，取类指示最小的那个；Extra category 表示用 1 个额外的分类值取代所有缺失值。
- Impute missing values 缺失值替换，主动处理方式。
- Exclude objects with missing values on this variable 排除分析变量含缺失值的观测，此方法对增补变量无效。

(4) Options 参数设置。在图 17-14 中，单击 Options 按钮，弹出如图 17-18 所示的参数设置对话框。采用默认设置，单击 Continue 按钮返回主界面。

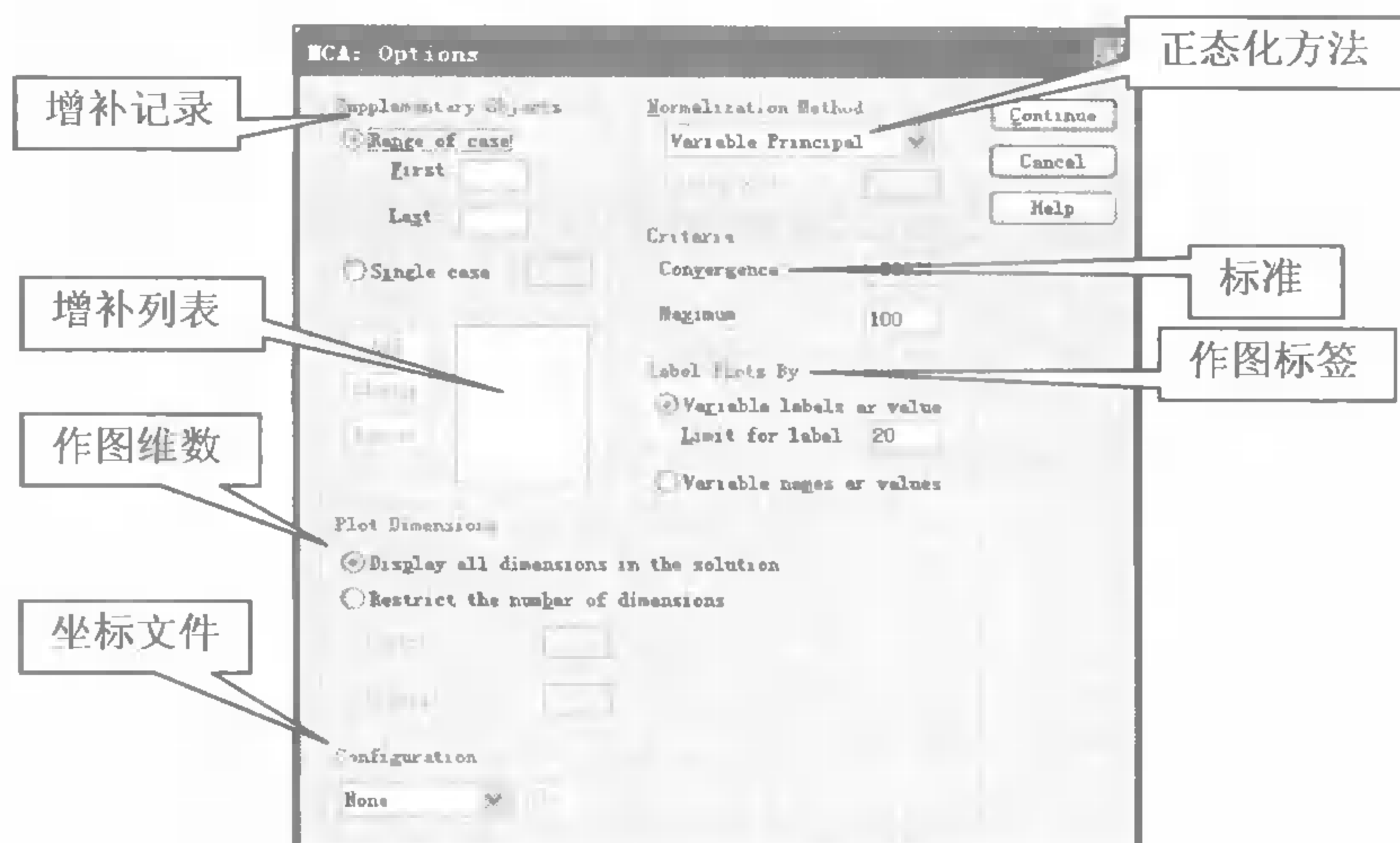


图 17-18 Options 参数设置对话框

① Supplementary Objects 栏设置增补观测在数据集里的记录号（行号）。

可以选择 Range of cases 单项框，指定起始行 First 和结束行 Last，然后单击 Add 按钮加入下面的增补列表；也可以选择 Single cases 单项框，在其后输入特定的行号，单击 Add 按钮加入增补列表；在增补列表选中某一项后，通过单击 Change、Remove 按钮可以对其进行更改或删除。

② Normalization Method 下拉列表用来选择变量或观测得分的正态化方法。

一个分析过程只能指定一个正态化方法，有 5 个可选项，它们分别对应于图 17-5 中简单对应分析的正规化方法（Normalization Method）。

- Variable Principal 变量主成分，相当于简单对应分析的 Column principal 方法；
- Object Principal 观测主成分，相当于简单对应分析的 Row principal 方法；
- Symmetrical 对称，相当于简单对应分析的 Symmetrical 方法；
- Independent 独立，相当于简单对应分析的 Principal 方法；
- Custom 自定义，相当于简单对应分析的 Custom 方法。

③ Criteria 栏设置模型的拟和标准。

Convergence 输入框指定收敛的临界值，若循环求解的最后两个模型的拟和优度之差小于此处的 convergence 值，则停止循环；Maximum number of iterations 输入框指定最大循环次数。



④ Label Plots By 栏设置输出图形的显示方式。

Variables labels or values 单选项，显示变量标签或变量值；Variables names or values 单选项，显示变量名或变量值；Limit for label 输入框，指定变量标签的最大长度。

⑤ Plot Dimensions 栏设置输出图形的维数，其设置方法与图 17-7 所示的简单对应分析 Plot Dimensions 设置相同。

⑥ Configuration 栏设置从一个文件读入坐标的结构信息，单击 File 按钮选择文件。

文件中的第一个变量，对应于当前分析中第一维的坐标；文件中的第二个变量，对应于当前分析中第二维的坐标，依次类推。底部的下拉列表，用来指定文件里的坐标所对应的观测的起始位置：Initial 表示对应于分析中的起始观测；Fixed 固定的，表示文件中的坐标信息将用来装配当前的分析变量，它们被当作增补变量处理。

（5）输出选项设置。在图 17-14 中单击 Output 按钮，弹出如图 17-19 所示的子对话框，设置多元对应分析的输出内容。勾选 Correlations of original variables 复选框；在变量列表选中 pr92（候选人）变量，单击从上至下第一个  按钮，将其选入量化列表；在变量列表选中 sex（性别）变量，单击从上至下第二个  按钮，将其选入描述列表；单击 Continue 按钮返回主界面。

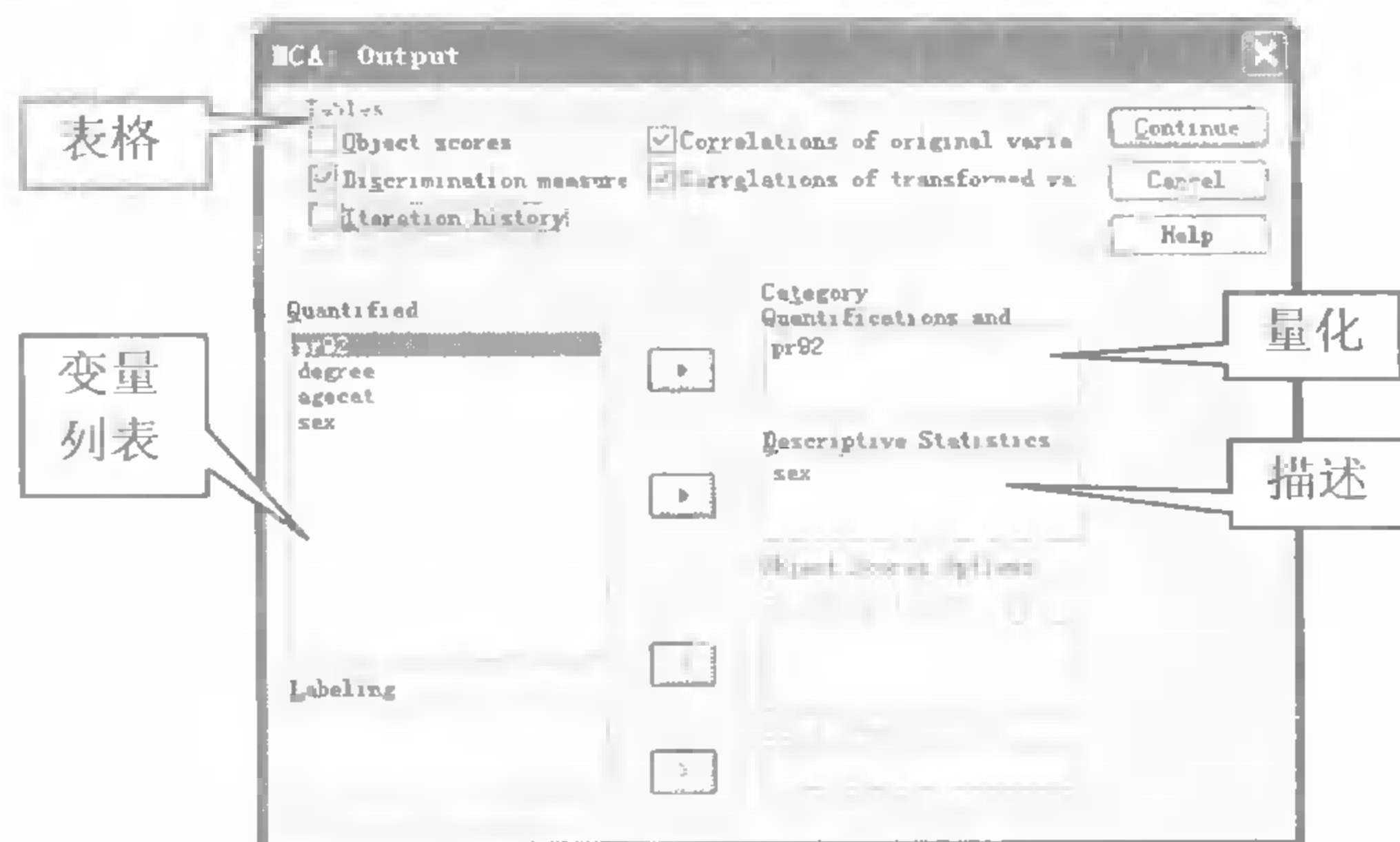


图 17-19 输出选项设置对话框



① Tables 栏选择输出哪些表格，可选项有如下 5 个。

- Object scores 观测得分表，包括质量、收敛标准、坐标等信息。选中后，右下角的 Object Scores Options 栏变为可选状态，其中：Include Categories Of 列表中的分析变量，将在结果表里显示它们的类别信息；Label Object Scores By 选框用于指定标识观测得分的标签变量。
- Discrimination measures 选项，输出每个变量、每个维度的判别度量方式。
- Iteration history 选项，输出迭代过程中方差的变化过程。
- Correlations of original variables 选项，输出初始变量取值的相关系数矩阵及其特征值。
- Correlations of transformed variables 选项，输出变换后变量的相关系数矩阵及其特征值。

② Category Quantifications and Contributions 量化变量列表，对其每一个维度输出类别量化的信息包括：质量、标准、坐标。

③ Descriptive Statistics 描述变量列表，输出其频数、缺失值个数、众数等基本统计信息。

(6) 保存选项设置。在图 17-14 中，单击 Save 按钮，弹出如图 17-20 所示的子对话框，设置多元对应分析的保存选项。保留默认选项，单击 Continue 按钮返回主界面。

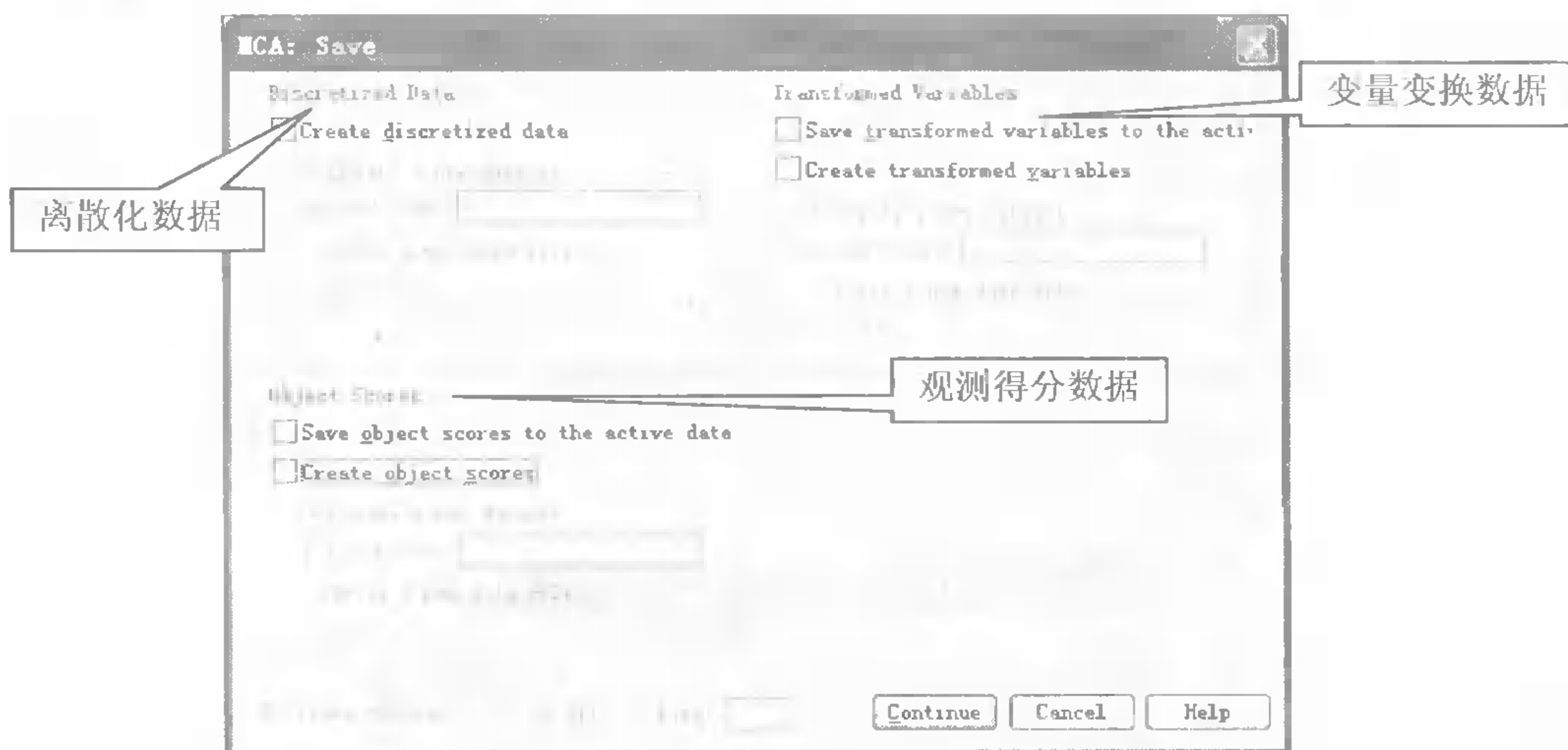


图 17-20 保存选项设置对话框

Discretized Data 栏设置保存离散化数据的选项；Object Scores 栏设置保存观测得分数据的选项；Transformed Variables 栏设置保存变量变换数据的选项。这 3 部分的设置内容相似，需要分别为其指定有关参数，内容有如下 3 个。

- Save object scores (transformed values) to the active dataset 复选框，把指定数据存至当前数据集；选中后激活底部的 multiple nominal dimensions 栏，设置待保存数据的维度，All 单选框表示保存所有维度，First 输入框指定待保存的最大维度。
- Create a new dataset 选项，表示建立一个新的数据集来保存指定数据，在 Dataset name 后的输入框键入数据集的名字。
- Write a new data file 选项，表示建立一个新的文件来保存指定数据，单击 File 按钮

选择文件。

(7) 对象作图的参数设置。在图 17-14 中, 单击 Object 按钮, 弹出如图 17-21 所示的对象作图子设置界面, 单击 Continue 按钮返回主界面。

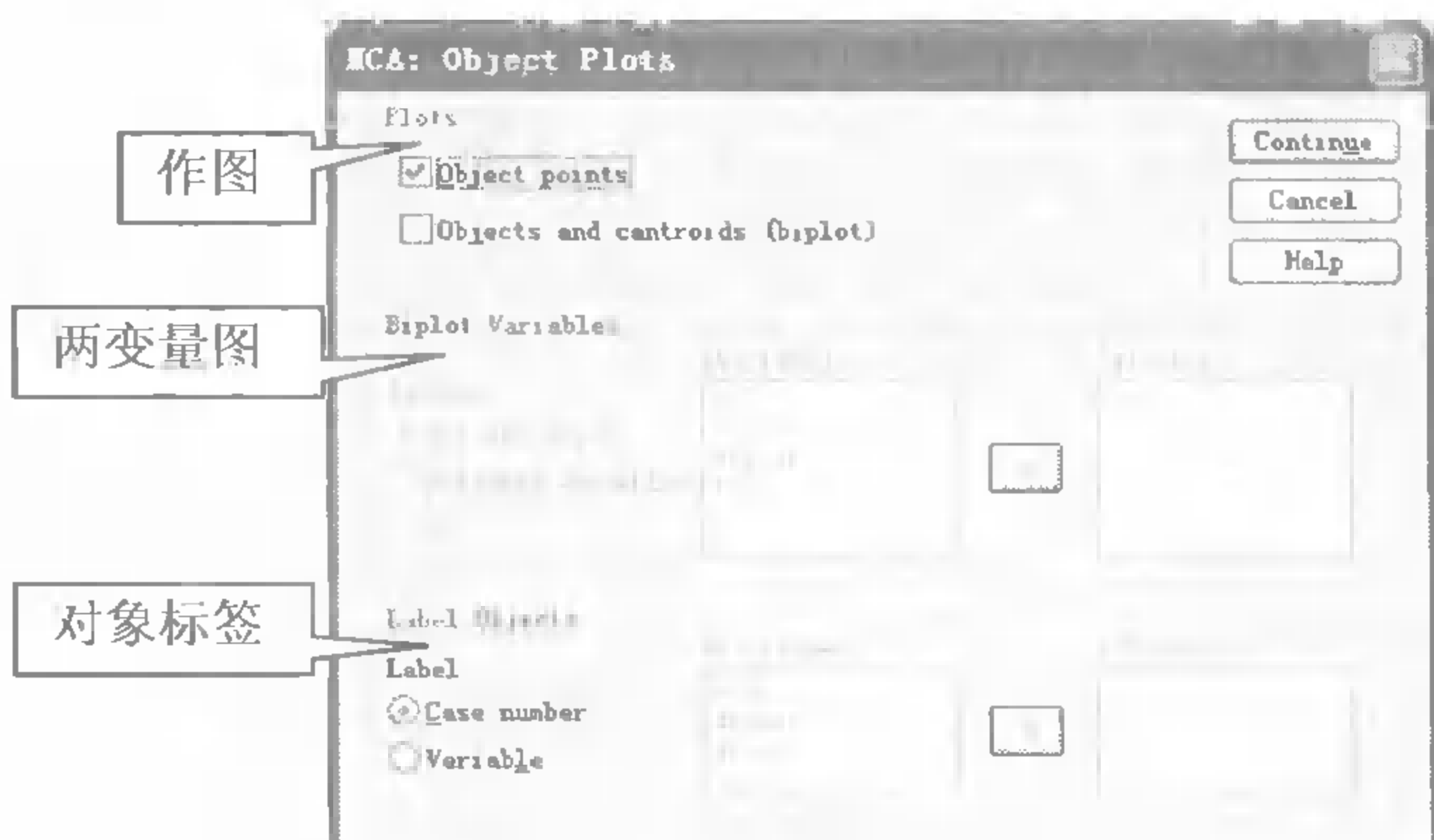


图 17-21 对象作图设置

- Plots 栏设置作图的类别: Object points 选项, 只对对象点作图; Objects and centroids (biplot) 选项, 对对象点及其中心点作图。
- Biplot Variables 栏设置作行、列联合分数图的变量。All variables 单选项, 使用全部变量; Selected variable 单选项, 把需要的变量从 Available 列表选入 Selected 列表。
- Label Objects 栏设置标识对象的标签变量。Case number 单选项, 以行号作为标签; Variable 单选项, 把标签变量从 Available 列表选入 Selected 列表, 可以同时选入多个。

(8) 变量作图的参数设置。在图 17-14 中单击 Variable 按钮, 弹出如图 17-22 所示的变量作图子设置界面。在变量列表选中 degree (受教育程度) 变量, 单击从上至下第一个 按钮, 将其选入 Category Plots 列表; 在变量列表选中所有变量, 单击从上至下第二个 按钮, 将其选入 Joint Category Plots 列表; 单击 Continue 按钮返回主界面。

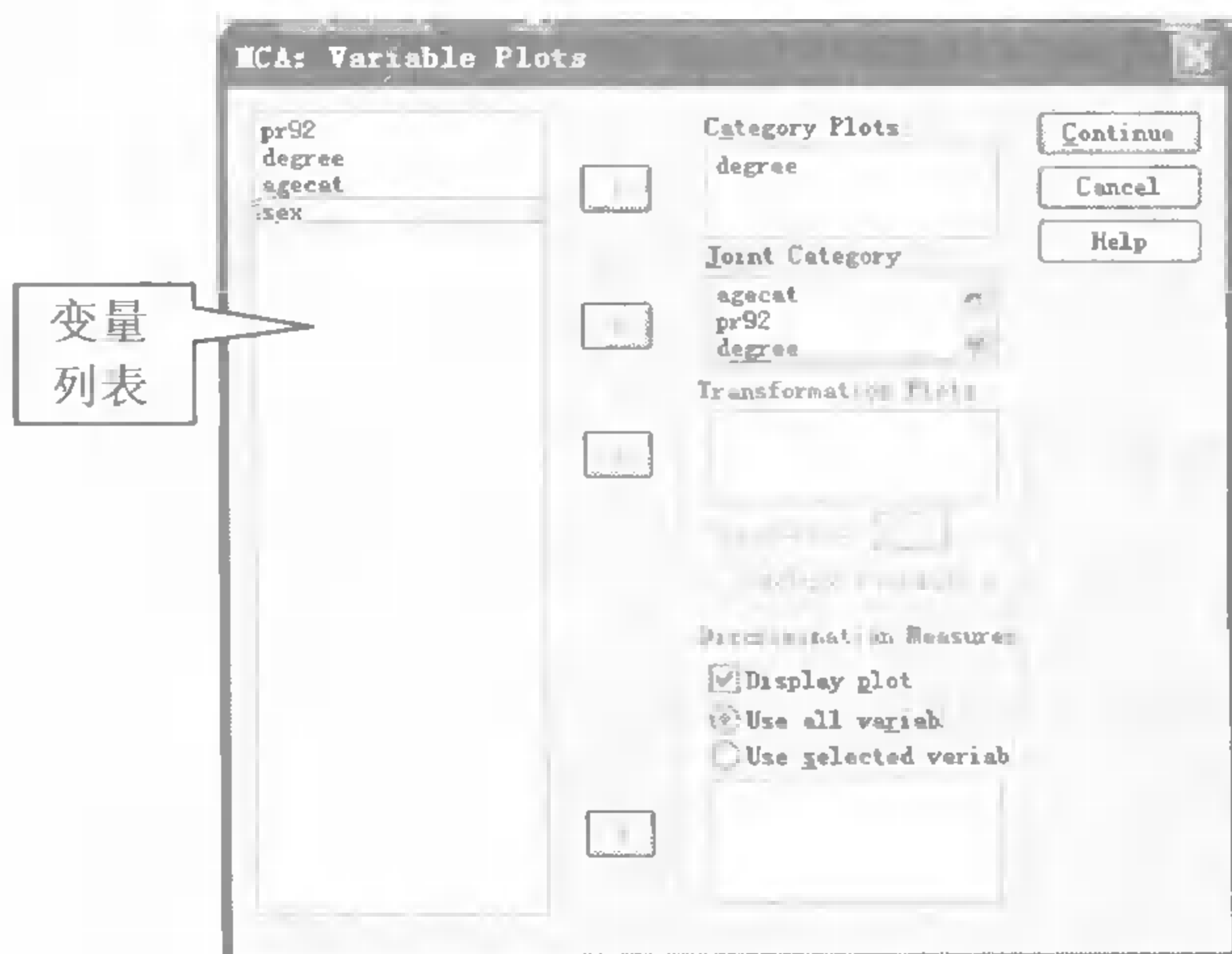


图 17-22 变量作图设置

- Category Plots 列表, 对选入的每个变量作一个图形, 显示其各类别的中心值。

- Joint Category Plots 列表, 在一个图形中显示所有选入变量各类别的中心值。
- Transformation Plots 列表, 对选入变量作最优量化值对类别指示变量的图形; 在 Dimensions 输入框指定作图的维数, 每个维数输出一个图形; 选中 Include residual plots 复选项表示为每个选入的变量输出残差图。
- Discrimination Measures 栏, 选中 Display plot 复选框激活后面的选项, 为指定变量输出区分度量 (就是量化后变量在各维度上的方差) 的图形, 指定变量的方式有两种: Use all variables 单选项, 表示使用全部变量; Use selected variable 单选项, 把需要使用的变量从左侧的变量列表选入下面的列表框。

### 17.3.3 实例的结果分析

在图 17-14 中单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 17-23~图 17-27 所示。

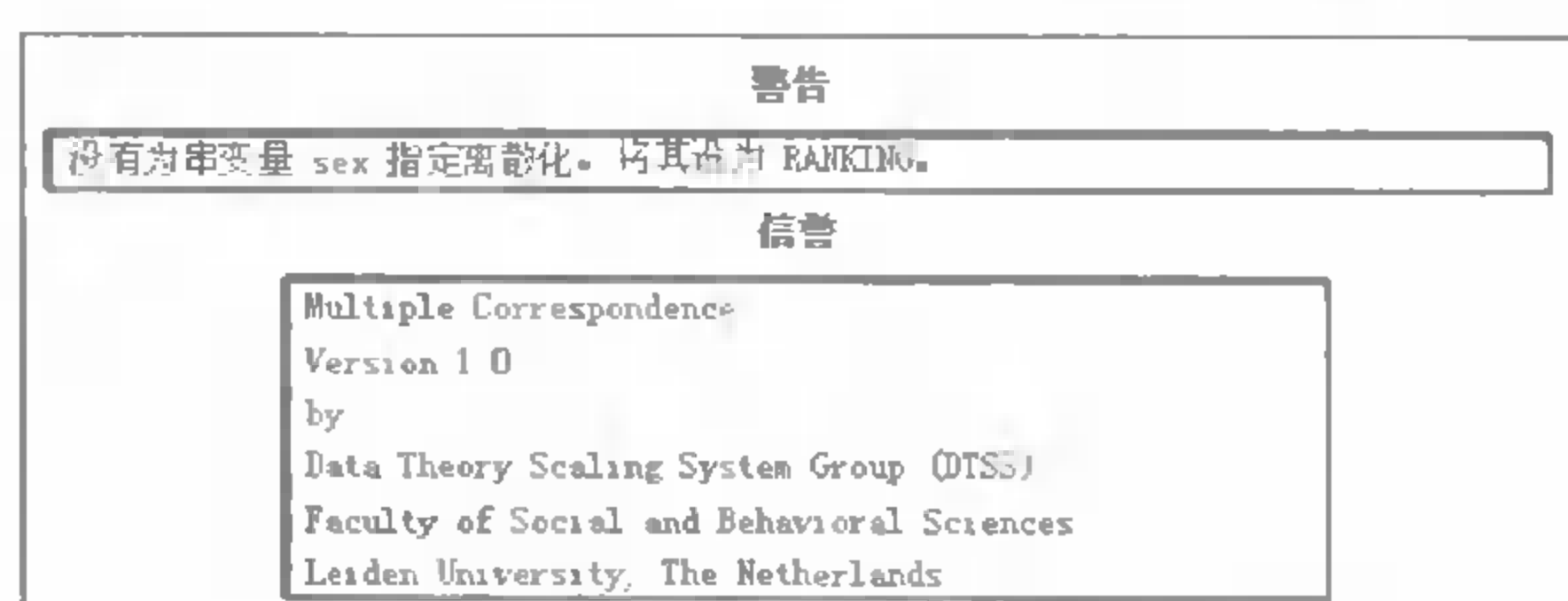


图 17-23 版权信息和处理摘要输出

案例处理摘要	
有效的活动案例	1659
具有缺失值的活动案例	189
补充案例	0
总计	1847
分析中使用的案例	1847

性别			
		离散后的类别 <sup>a</sup>	频率
有效	Female	1	804
	Male <sup>b</sup>	2	1043
	总计		1847

a. 秩  
b. 模式。

图 17-24 案例处理摘要和性别统计信息

迭代历史记录			
迭代次数	方差考虑情况		损失
	总计	增量	
32 <sup>a</sup>	1.259658	.000010	2.740342

a. 因为获得收敛的检验值, 所以迭代过程停止。

模型摘要			
维	Cronbach's Alpha	方差考虑情况	
		总计 (特征值)	增量
1	.329	1.327	.332
2	.215	1.192	.298
总计		2.519	.630
均值	.275 <sup>a</sup>	1.260	.315

a. 总 Cronbach's Alpha 基于平均特征值。

图 17-25 迭代记录和模型摘要输出

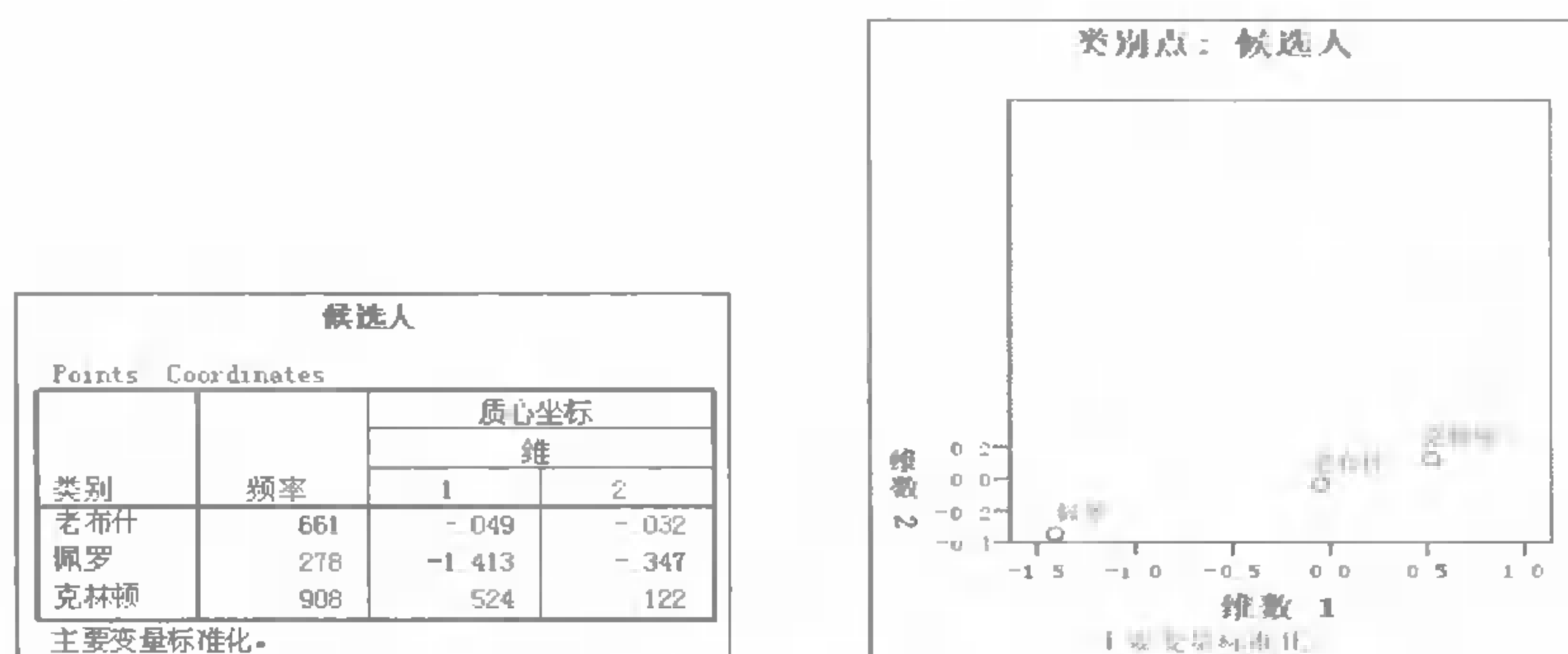


图 17-26 候选人变量的质心坐标及其图形

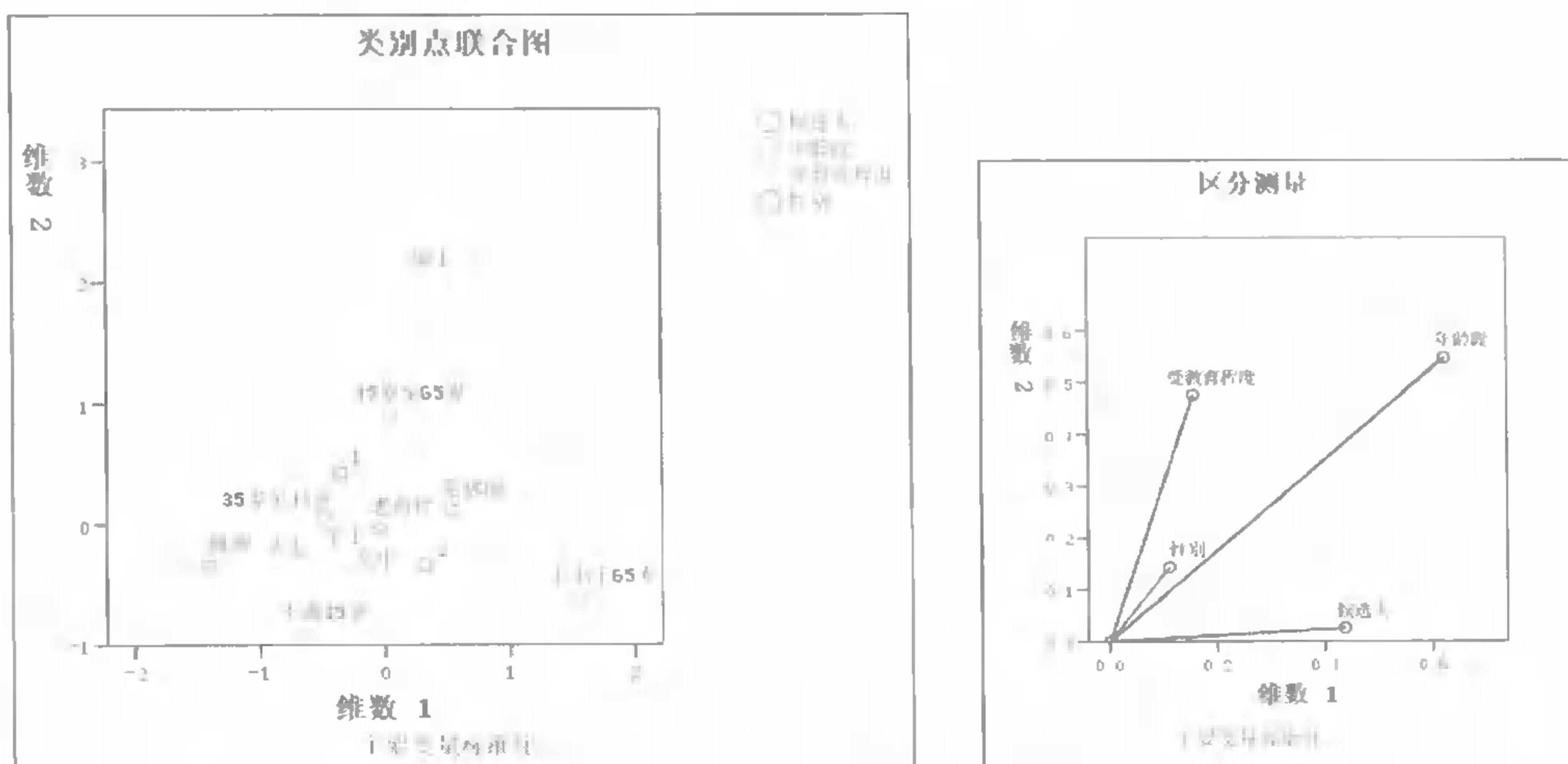


图 17-27 多元对应分析类别点联合图形和区分度量图形

(1) 警告信息和版权信息。如图 17-23 所示，“警告”表格指出没有对字符串变量 Sex 进行离散化设置，系统将使用 Ranking 方法对其实施自动离散化；图 17-24 中的“性别”表格给出了 Sex 变量离散化后的值：1 表示 Female，2 表示 Male。“信誉”表格显示的是 SPSS 对应分析模块的版权信息。

(2) 处理摘要表格。图 17-24 中的“案例处理摘要”表格罗列了原始数据的基本使用情况，包括缺失值观测数、补充案例数等。“性别”表格给出的是性别变量的编码和统计信息。

(3) 迭代记录和模型摘要。如图 17-25 所示，“迭代历史记录”表格给出了最后一次迭代的次数、方差、方差增量等信息，表格下方还说明了迭代终止的原因。“模型概要”表格给出了两个维度的方差总计（特征值）及其惯量信息。

(4) 单个变量的类别中心坐标及其图形。如图 17-26 所示，是候选人变量的质心坐标及其图形。从这样的单个图形可以判断把该变量映射至二维空间后，其各个类别取值的区分程度。其他变量的质心坐标及其图形与此类似。

(5) 所有变量的类别点联合图形和区分度量图形。如图 17-27 所示，“类别点联合图”是把四个分析变量的类别点中心坐标，在一个图形中加以显示的效果，此图形与图 17-12 所示的简单对应分析的二维分析图类似，它根据图形中各点的邻近关系进行分类，只是多了几个变量的信息。

图 17-27 中还给出了关于区分度量的图形。所谓区分度量，相当于变量量化后的值向量与观测得分维度向量的平方相关系数，反应了维度得分与量化后变量值的相关性大小，由此可以判断重点变量在与其相关性较大的维度上的特征，在这个维度上的类别点一般会分得更开。可见受教育程度在维度 2 上值得受较大关注；年龄段在两个维度上都需要关注；而性别变量的区分度量在两个维度上都显小，故可考虑增大性别变量的权重再作分析。

(6) 更改权重后的类别点联合图形和区分度量图形。本节需要更改性别变量的权重后再做分析，首先在图 17-14 中的 Analysis Variables 列表里，选中 sex（性别）变量；然后单击 Define Variable Weight 按钮，弹出如图 17-15 所示的对话框，在输入框中键入 2，单击 Continue 按钮返回主界面。在主界面单击 OK 按钮运行，输出的类别点联合图和区分度量图如图 17-28 所示。



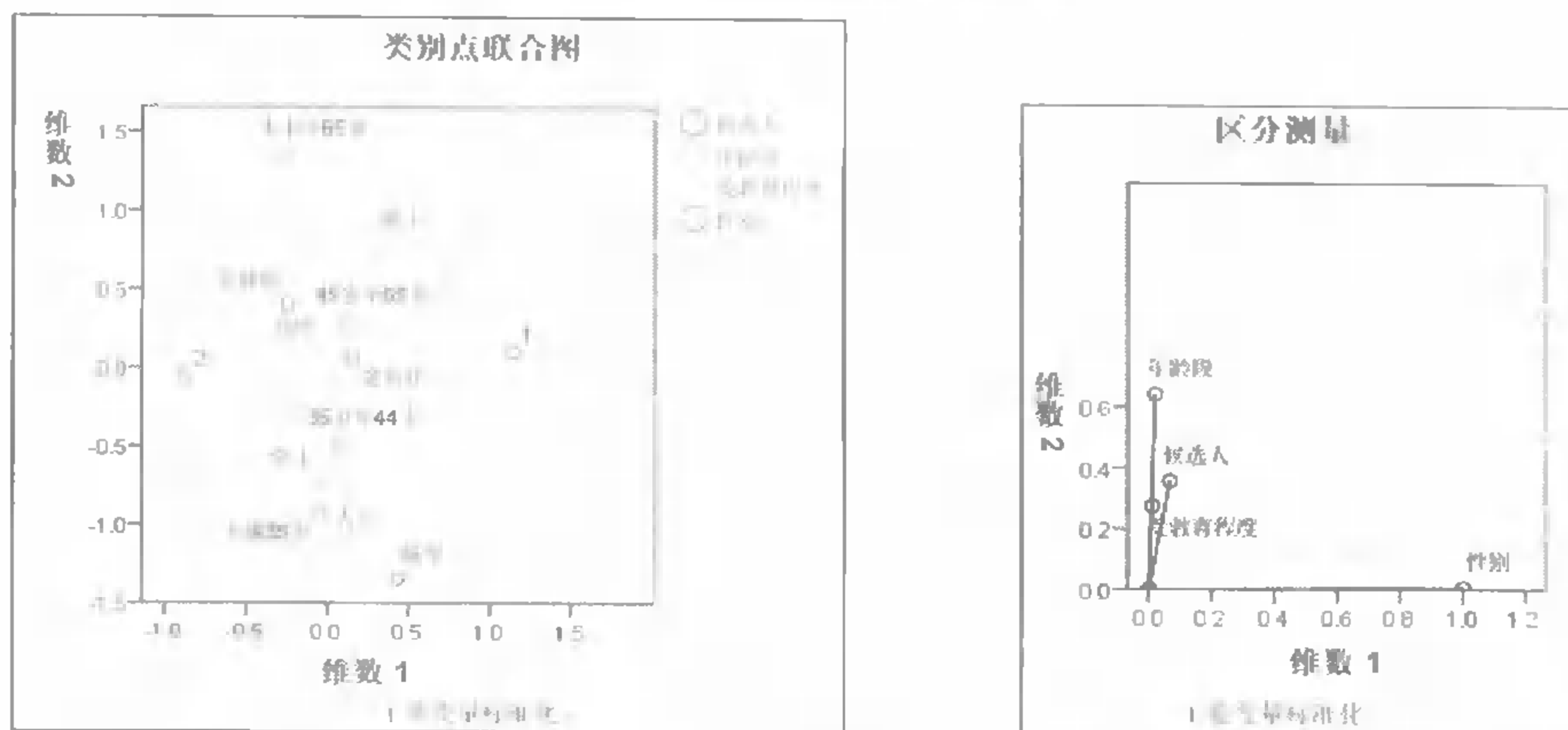


图 17-28 更改权重后的类别点联合图形和区分度量图形

通过更改设置，性别变量的权重为 2，其他变量的权重不变仍为 1，这表示在分析中加大了性别变量的影响。

比较图 17-27 和图 17-28 发现，改进后的类别点联合图区分性更好一些。下面给出一些比较直观的解释：高中学历、45~65 岁的人对克林顿和老布什比较青睐；硕士水平选民更喜欢克林顿；本科学历、35~44 岁的中年人对老布什感兴趣多一些；对于佩罗，支持他的多是大专学历、不满 35 岁的青年一派；从性别上看，女士更喜欢老布什和佩罗，男士支持克林顿多一些。

# 第 18 章 缺失值分析

在日常工作及科学研究中，当处理样本较大的群体调查时，由于多种原因可能会导致所收集的数据不完整，这时的初始数据中就含有缺失值。缺失值带来许多负面影响，比如：含缺失值的观测可以看作是正常观测的系统误差，这会导致计算结果不准确；获得的信息比预期要少，这导致计算统计量的精度降低；许多统计过程的假设是基于完整数据的，数据不完整将导致计算过程无法进行。

在 SPSS 中可以采用多种方式对缺失值进行灵活处理，比如，在各个统计分析过程里加入处理缺失值的选项；或者通过缺失值替换过程在分析前先处理掉缺失值；本章介绍专门用于缺失值分析的过程 Missing Value Analysis。

## 18.1 缺失值分析的概念

缺失值是统计人员和数据采集人员所不愿见到的，但也是无法避免的。在大型的数据采集任务里中，即使有着非常严格的质量控制，含有缺项、漏项的记录也可能很容易地达到 10%；在进行敏感问题的调查时，缺失值问题就更加突出了，比如问卷中涉及到了家庭收入、婚外性伴侣等问题时，许多受访者都会以漏填来避免尴尬。

有些统计分析方法采取将含缺失值的观测记录直接删除的做法，当缺失值较少时，这样做没有太大问题；但当缺失值数量较多时，这样做会直接丢失大量的信息，并有可能导致错误的结论，故而进行更为系统的缺失值分析是非常有必要的。

### 18.1.1 缺失值的表现方式

数据的缺失是有一定规律的，其缺失方式大致可以分为以下三种：完全随机缺失 (missing completely at random, MCAR)、随机缺失 (missing at random, MAR) 和非随机缺失 (missing at non-random, MANR)。

(1) 完全随机缺失。完全随机缺失的含义就是指缺失现象完全是随机发生的，和自身或其他变量的取值无关。这是缺失值问题中处理起来比较简单的一种，可以直接将缺失值删除，无需担心估计偏差，这样做唯一的缺点是会丧失一些信息；也可以采用均值替换等方法处理缺失值，以便充分利用样本信息。要评估 MCAR 假设是否成立，可以通过比较回答者和未回答者的分布情况进行验证，也可以使用单变量 t 检验或 Little's MCAR 检验进行更精确的推断。事实上，完全符合 MCAR 的情况非常少见，而且上述的检验方法都只能证明 MCAR 假设不成立，而不是证明其成立，因此在对缺失情况作评价时一定要相当谨慎，切不可妄下结论。

(2) 随机缺失。这种情况要严重些,但也更加常见,它的含义是指有缺失值的变量缺失情况的发生与数据集中其他无缺失变量的取值有关。此时,缺失值不仅会引起信息损失,还可能导致分析结果的不可信。比如调查人群的血压时发现数据有缺失,但缺失情况是以高龄组为主,这是由于高龄组的受访者因行动不便,不能到场接受深度访谈和检查所致:此时将缺失值直接删除就不一定合适,而应利用已知变量对缺失的数据进行估计,这样才能对总体有一个综合的评价。

(3) 非随机缺失。这是最坏的一种情形,数据的缺失不仅和其他变量的取值有关,也和其自身有关,比如在调查收入时,收入高的人出于各种原因不愿意提供其家庭年收入值。这种情形下,缺失值分析模型基本上是无能为力的,只能做一下粗略的估计。

SPSS 的缺失值分析模块,主要是对 MCAR 和 MAR 的情况进行研究,尤其是后者。研究者应该在进行调查之前,就考虑哪些重要变量可能会有缺失值出现,以及由此引发问题的严重程度;然后在设计问卷时就包括一些与之相关的变量,以使用这些变量来估算缺失值。

### 18.1.2 SPSS 中的缺失值处理方法

对不同情况的数据缺失值,SPSS 提供了多种处理方法。

(1) 删除缺失值。当缺失值较少时可以采用该方法,它不需要单独的分析过程,在多数统计分析过程的 Options 面板或其他设置面板里经常有相应的选项,一般给出的可选内容如下。

- Excludes cases analysis by analysis。如果同时选择了多个变量进行分析,只有当前分析中具体用到的某个变量有缺失值时,才将相应的记录删除。这是多数情况下的默认处理方式,可以最大限度地利用有效信息。
- Excludes cases listwise。如果同时选择了多个变量进行分析,只要其中的某个变量含有缺失值,就在所有分析过程中将相应的记录删除。
- Report values。这是描述性的统计分析过程专有的选项,表示将缺失值作为一个单独的分类加以描述。

(2) Replace Missing Values 过程。主菜单 Transform 下的 Replace Missing Values 子选项,专门用于实现缺失值替换的功能。

Replace Missing Values 过程将所有的记录看成一个序列,然后采用某种统计量对缺失值进行替换或填充,此方法经常用于解决时间序列模型中的缺失值问题;虽然其中的一些替换方法也可以用于普通数据,但可能会出现不准确的结果。

(3) Missing Value Analysis 过程。主菜单 Analyze 下的 Missing Value Analysis 子选项,专门用于进行缺失值分析,本章将要详细介绍的就是 SPSS 的这个功能。Missing Value Analysis 过程可以对缺失值问题进行全面而细致地分析,其主要功能包括如下几项。

- 缺失值的描述和快速诊断。它用诊断报告的形式评估缺失值问题的严重性,用户可以观察到它们在哪些变量中出现,以及出现的比例有多少,还可以推断其出现是否与其他变量的取值有关。通过这些信息,可以帮助用户判断这些缺失值的出现是否会影响到分析结论的准确性。
- 更精确的摘要统计量。它提供了多种方法用于估计含缺失值数据的均值、相关矩阵和协方差矩阵,通过这些方法计算出的统计量将更加可靠。
- 缺失值替换。它可以使用 EM 或回归算法,从未缺失数据的分布情况中,推导出缺失数据的估计值,从而能有效地使用所有数据进行分析,以此提高统计结果的可信度。

## 18.2 缺失值分析的参数设置

缺失值分析过程可以处理任意类型的数据，但要求对非系统定义的缺失值，必须定义为用户缺失值。依次单击菜单“Analyze→Missing Value Analysis...”，打开缺失值分析的主设置面板，如图 18-1 所示。

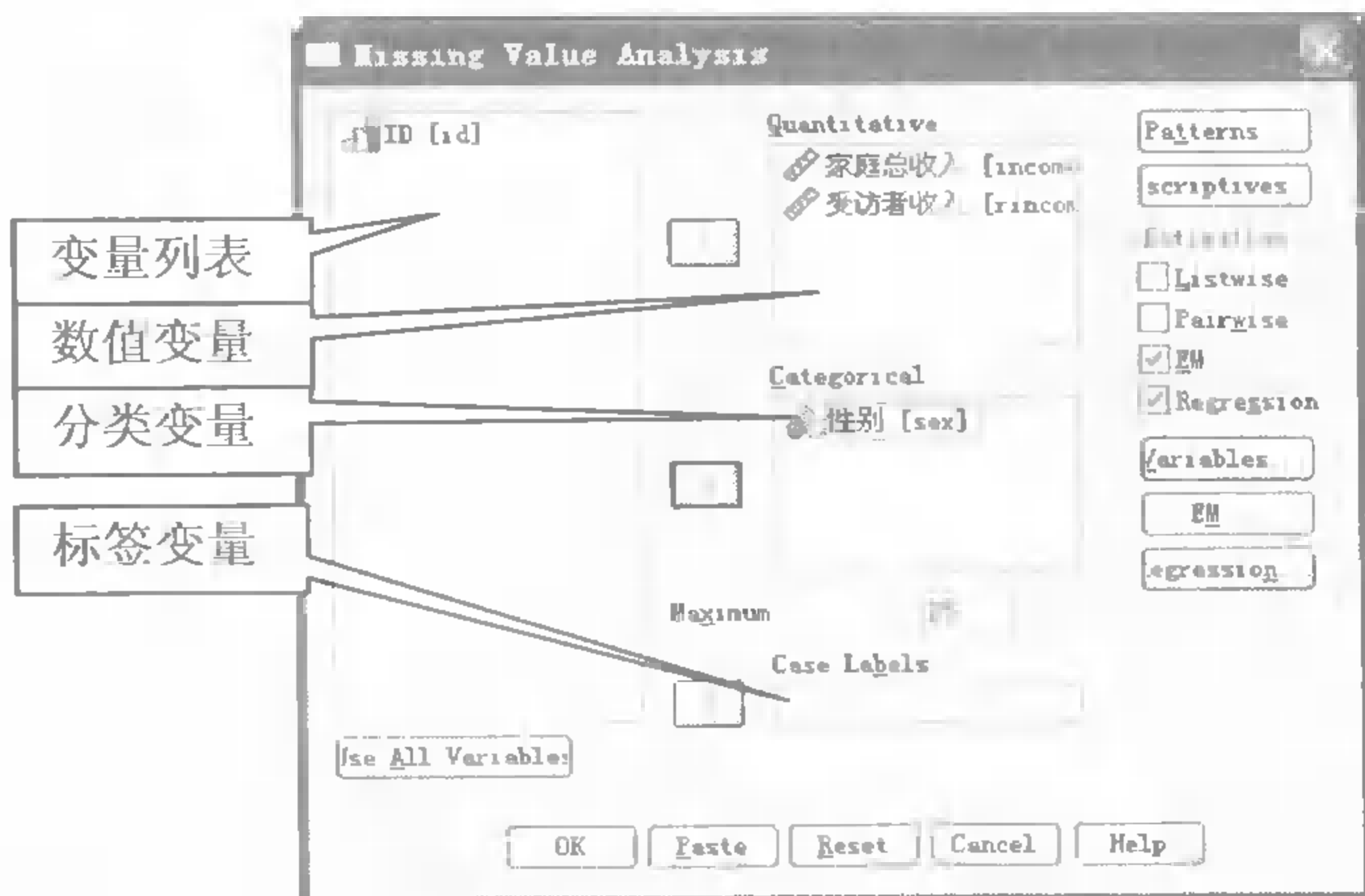


图 18-1 缺失值分析的主设置面板

### 1. 主界面设置

在图 18-1 中，设置与分析变量、缺失值处理方法相关的选项。

#### (1) 指定分析变量。

- Quantitative 列表框，用于选入进行缺失值分析的定量变量（数值型变量）。
- Categorical 列表框，用于选入进行缺失值分析的分类变量。Maximum 输入框，指定分类变量允许的最多分类数，默认为 25，超过此临界值的分类变量将不进入分析，因为太多的分类将大大减慢运算速度，并且对计算机内存有很高的需求。
- Case Labels 选框，用于选入对结果进行标识的标签变量。

没有选入 Quantitative 列表和 Categorical 列表的变量将不会存储到结果数据文件中，如果希望附加一些变量到结果文件，可以将它们指定为分类变量。

#### (2) Use All Variables 按钮。

单击它自动将左侧变量列表的所有变量选入特定的分析列表框，数值型变量全部选入 Quantitative 列表框，字符型变量全部选入 Categorical 列表框。

(3) Estimation 子设置栏，用于选择计算均值、相关矩阵和协方差矩阵等统计量时，对缺失值的处理方法。

- Listwise 复选框，只要分析中的任意一个因变量或分组变量中带有缺失值，则该记录将不被用来作任何分析。
- Pairwise 复选框，只有具体计算时用到的变量含缺失值时，该记录才不进入当前分析。
- EM 复选框，使用 EM（Expectation-maximization，期望最大化）迭代方法估计缺失值，推荐使用该方法。
- Regression 复选框，使用多元线性回归算法估计缺失值。



## 2. 模式设置

在图 18-1 中单击 Patterns 按钮，弹出如图 18-2 所示的模式设置对话框，在此设置关于变量（或记录）的缺失值样式表的格式。

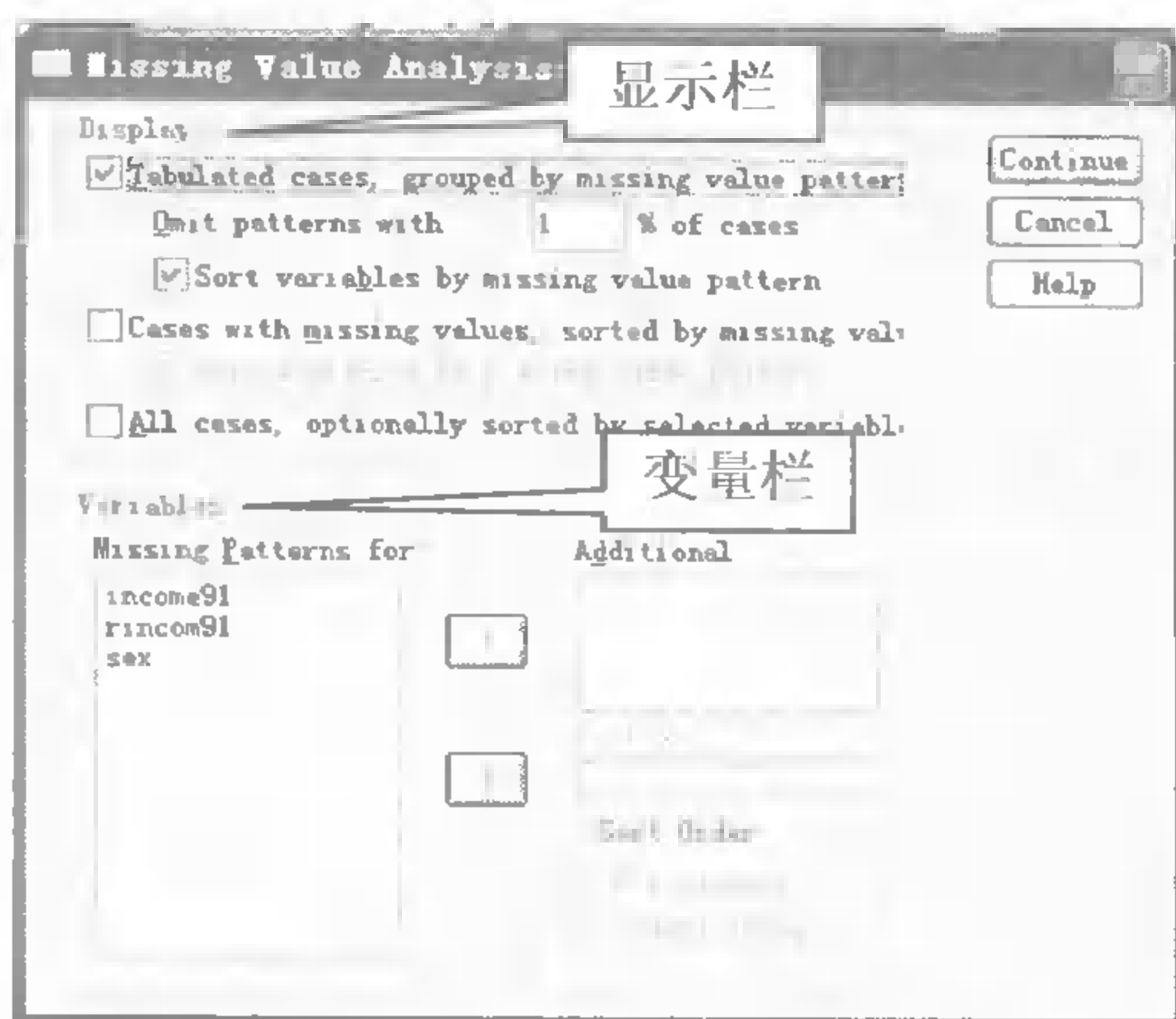


图 18-2 缺失值分析的 Patterns 设置

(1) Display 栏，选择缺失值样式表的类型，有如下 3 个可选项。

① Tabulated cases 方式，为每个分析变量都输出缺失值样式表，缺失值类别用“X”表示。

➤ Omit patterns 输入框，用于指定忽略比例，出现频数小于此比例的缺失模式将不被显示。

➤ Sort Variables 复选框，选中表示按照缺失值模式排序。

② Cases with missing values 方式，为每个含有缺失值的记录给出缺失值样式表。

➤ Sort Variables 复选框，选中表示按照缺失值模式排序。在输出样式表中，系统缺失值、用户定义缺失值 1、用户定义缺失值 2、用户定义缺失值 3 分别用“S”、“A”、“B”、“C”表示；另外，在样式表中按照  $(Q1-1.5 \times IQR, Q3+1.5 \times IQR)$  估计正常值的范围，超出此范围的取值被认为是极大值或极小值，分别用“+”、“-”符号来表示。

③ All cases 方式，列出所有记录的缺失值情况。

输出样式表中的表示符号，与 Cases with missing values 方式给出的含义相同。

(2) Variables 子设置栏。在 Display 栏选中某项后，激活 Variables 栏的设置内容，在此指定样式表中的样式的标签变量和排序方式。

➤ Missing Patterns for 列表框，显示选入的所有分析变量。

➤ Additional information for 列表框，输出所列变量的观测值列表。在样式表中，为定量变量输出其均值，为分类变量输出其各缺失值样式的频数。

➤ Sort by 选框，指定输出观测列表的排序变量（只有在 Display 栏选择 All cases 方式时才有效）。Sort Order 栏，指定其排序方式为：Ascending 升序或 Descending 降序。

## 3. 描述性统计量输出设置

在图 18-1 中，单击 Descriptives 按钮，弹出如图 18-3 所示的描述设置对话框。

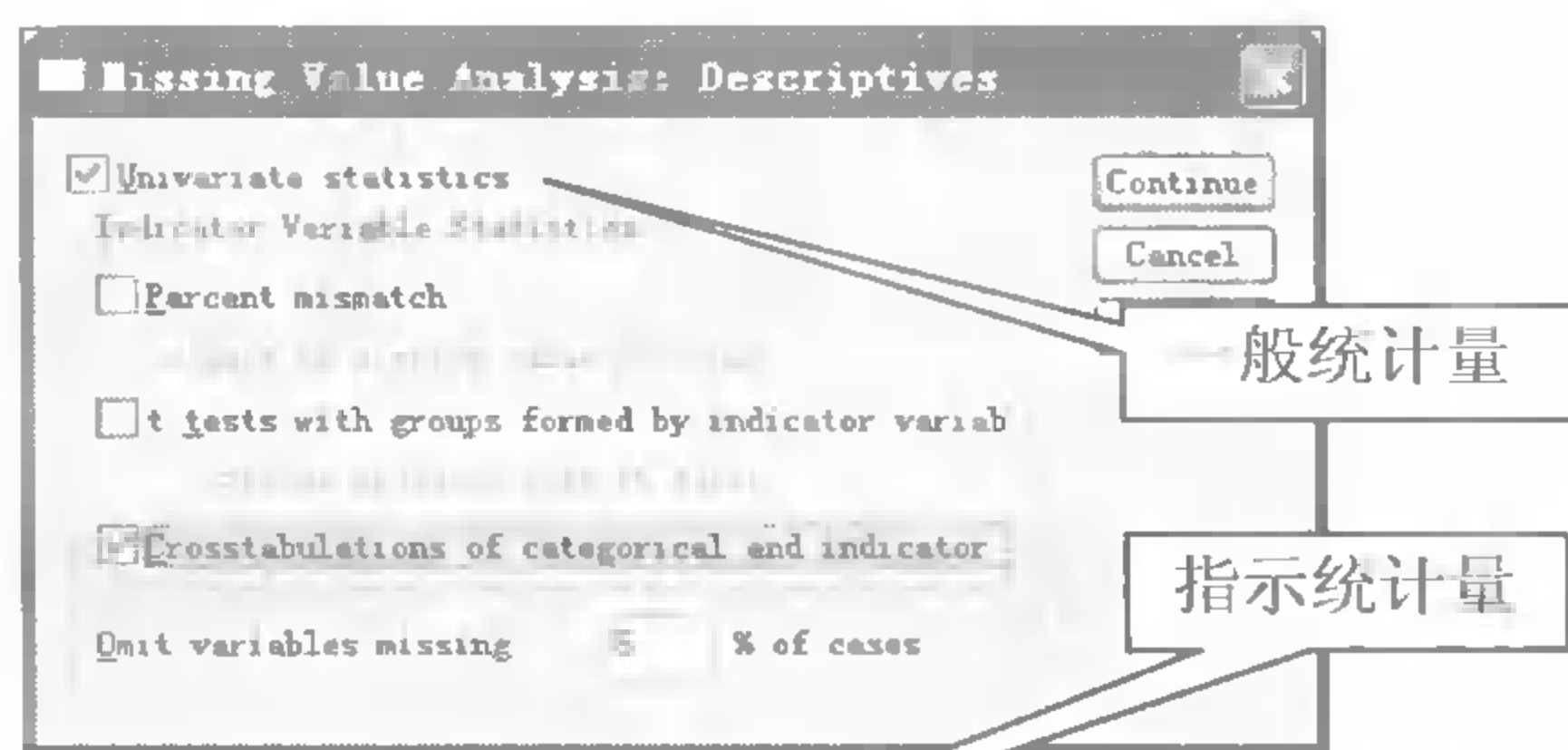


图 18-3 缺失值分析的 Descriptives 设置

(1) Univariate statistics 复选框。表示为每个变量输出非缺失数据的个数、均数、标准差等基本统计量，同时输出缺失值、极大值、极小值的数量和百分比。

(2) Indicator Variable Statistics 子设置栏。SPSS 为每个进入分析的变量生成一个指示变量，用于标识相应数值是否缺失，该变量不会被显示出来，但可以在此处对指示变量进行一定的利用和分析，有如下 3 个设置选项。

- Percent mismatch 复选框，表示对于每对变量，输出其中一个变量缺失、另一个未缺失的记录所占的比例。Sort by 复选框，选中表示按照缺失值样式进行排序。
- t tests 复选框，根据指示变量标识的是否缺失，将记录分为两组，然后对每个数值变量在这两个组的均值进行 t 检验。
- Cross tabulations of categorical and indicator variables 复选框，为分类变量和指示变量生成交叉表，输出非缺失值、各类缺失值的频数信息。

(3) Omit variables 输入框，指定忽略比例，缺失值频数小于此比例的变量将不被显示。

#### 4. EM 和 Regression 方法的变量设置

在图 18-1 里的 Estimation 栏单击选中 EM 或 Regression 后，单击 Variables 按钮，弹出如图 18-4 所示的变量设置对话框。

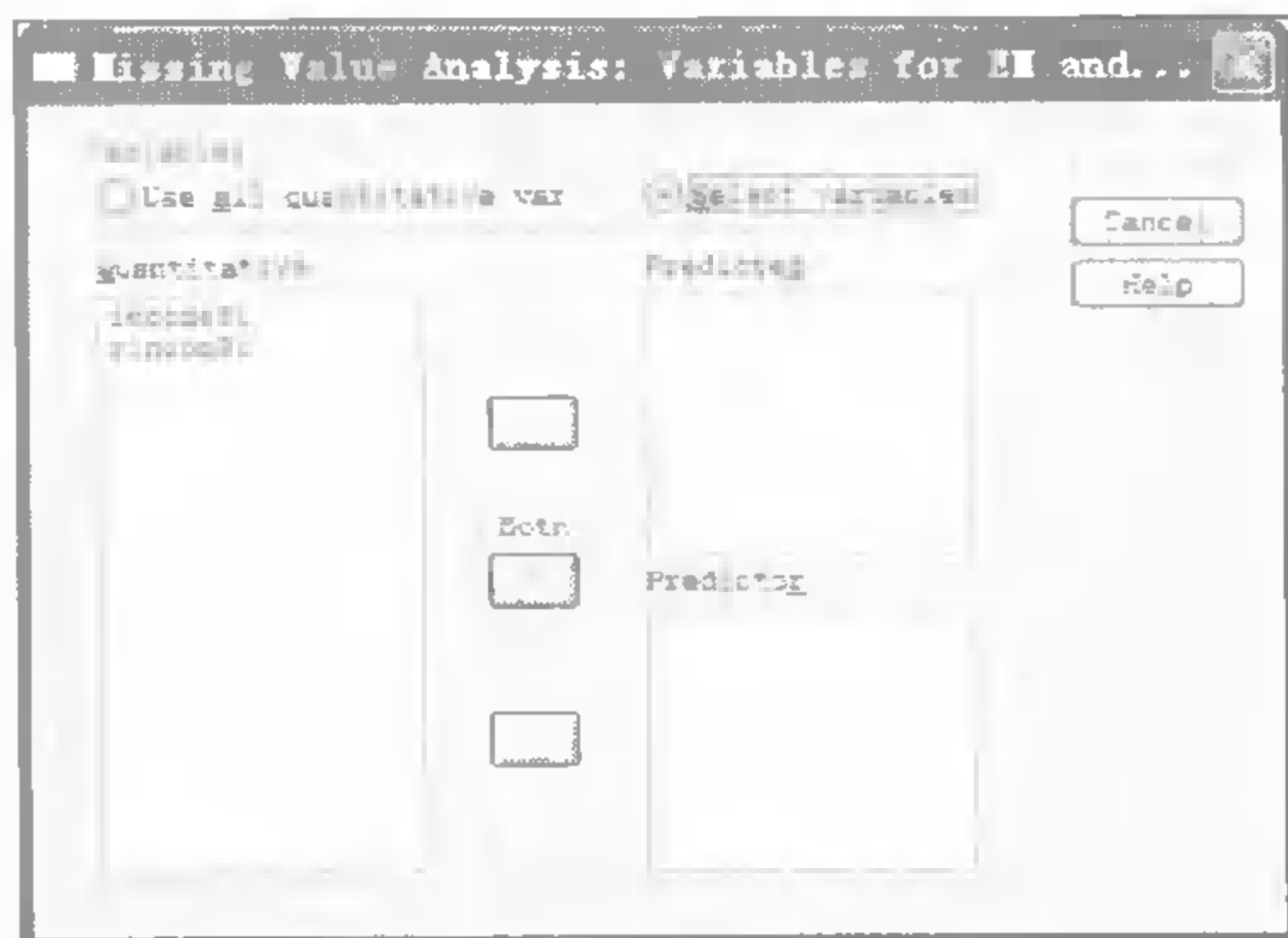


图 18-4 缺失值分析的 Variables 设置

(1) Variables 栏，选择指定变量的方式，有两个可选项。

- Use all quantitative variables 单选框，选中后表示使用所有连续变量。
- Select variables 单选框，由用户指定分析变量，选中后激活下面的设置内容。

(2) Quantitative 列表框，显示所有可用于缺失值估计的连续变量。

(3) Predicted Variables 列表框，用于选入需要估计缺失值的变量（因变量）。

(4) Predictor Variables 列表框，用于选入用来估计其他变量缺失值的变量（自变量）。

(5) Both 按钮，单击它可以把 Quantitative 列表里选中的变量，同时选入 Predicted Variables 列表框和 Predictor Variables 列表框。

## 5. EM 设置

在图 18-1 里的 Estimation 栏单击选中 EM 后，单击 EM 按钮，弹出如图 18-5 所示的对话框，在此设置 EM 算法的相关参数。

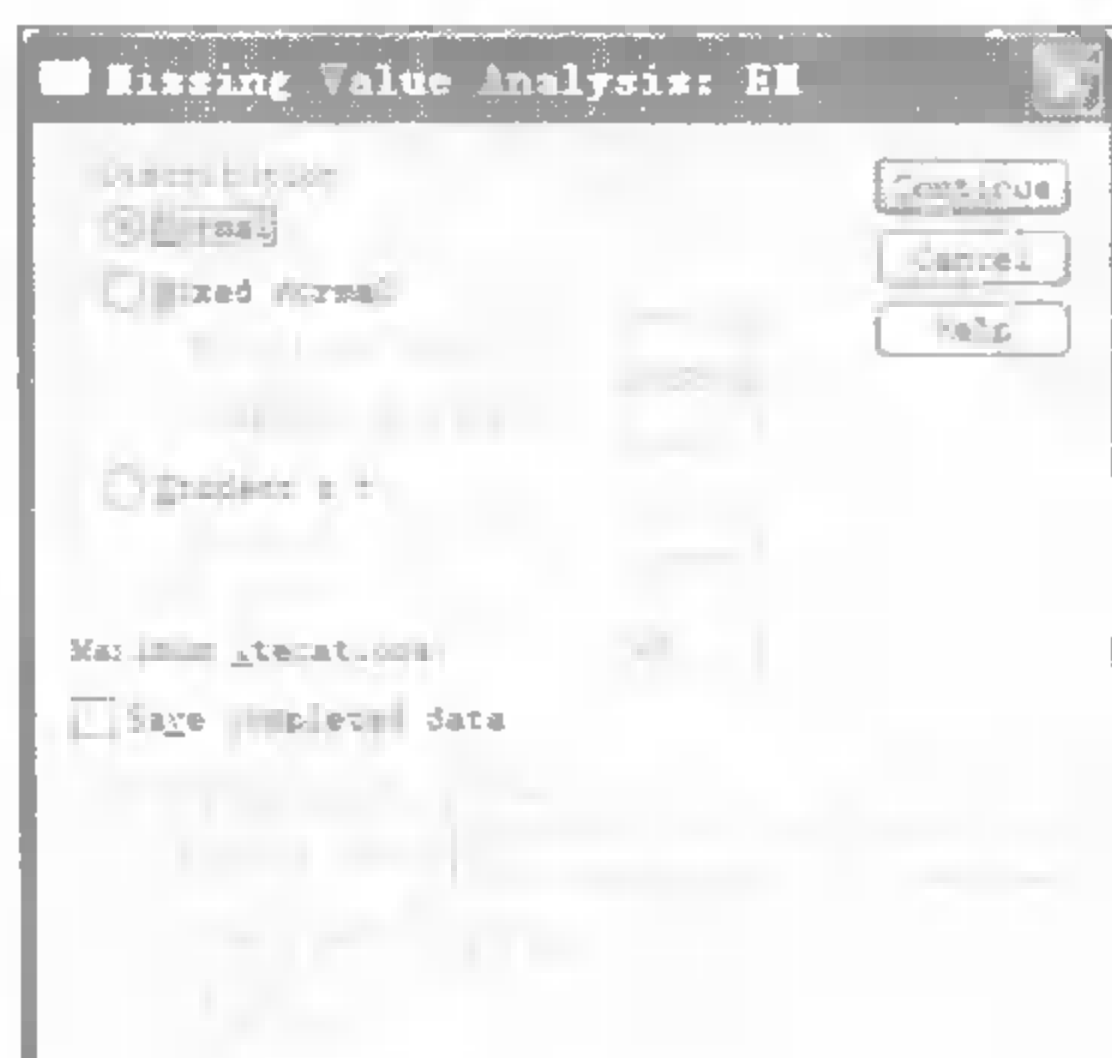


图 18-5 EM 算法的参数设置

(1) Distribution 子设置栏，在此选择总体的分布形式，可选项有如下 3 个。

- Normal 正态分布，默认选项。
- Mixed normal 混合正态分布，在 Mixture proportion 输入框指定混合比例，在 Standard deviation ratio 输入框指定标准差比。
- Student's t 学生 t 分布，需要指定 t 分布的自由度（Degrees of freedom）。

(2) Maximum iterations 输入框，指定最大迭代次数，默认为 25。

(3) Save completed data 复选框，用于保存将缺失值用 EM 算法替换后的数据，可以新建一个数据集（Create a new dataset 选项），在 Dataset name 输入框指定数据集名称；或者新建一个数据文件（Write a new data file 选项），单击 File 按钮指定文件路径和文件名。

## 6. Regression 设置

在图 18-1 里的 Estimation 栏单击选中 Regression 后，单击 Regression 按钮，弹出如图 18-6 所示的对话框，在此设置 Regression 算法的相关参数。

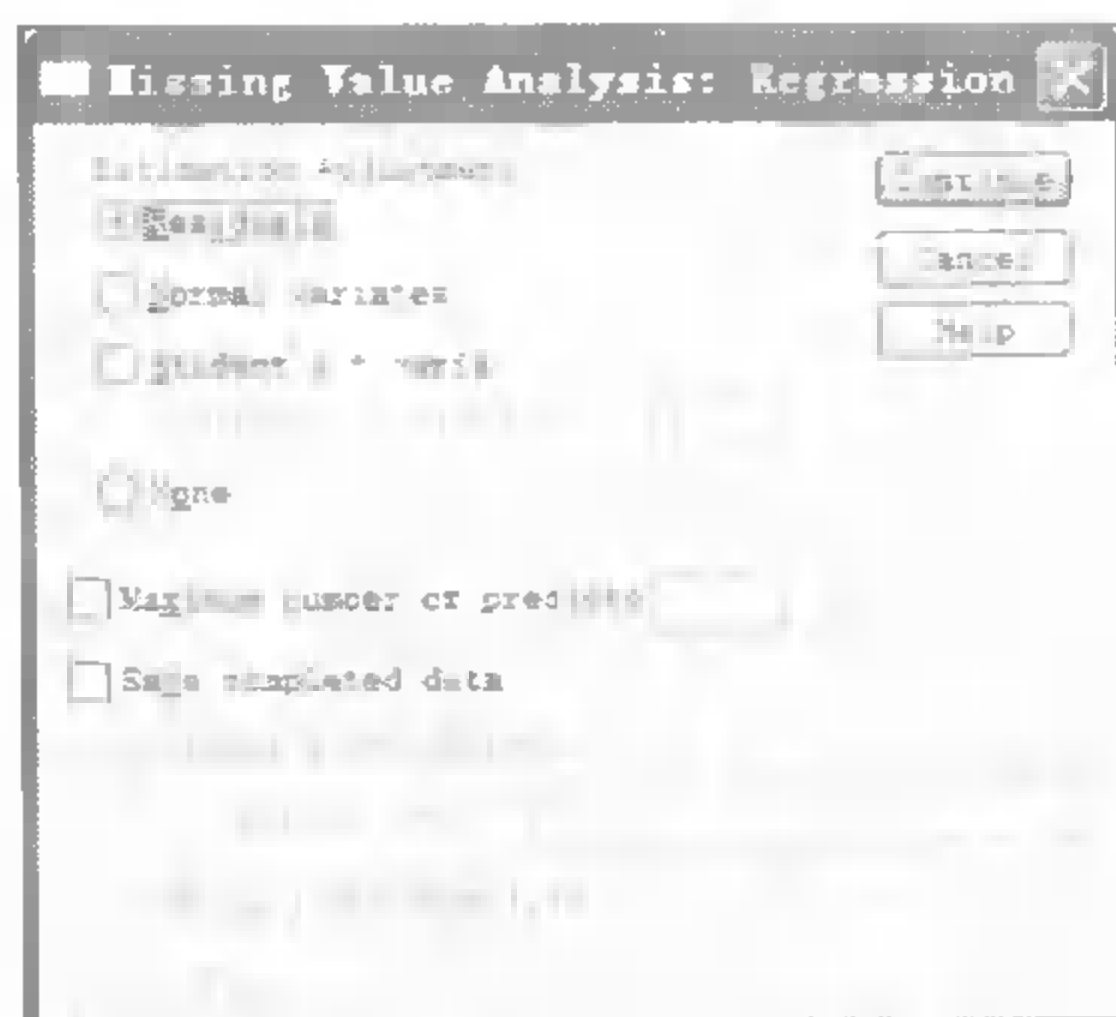


图 18-6 Regression 算法的参数设置

① Estimation Adjustment 子设置栏。在用回归算法计算出估计值后，可以再加上一个随机项，使得它更接近实际情况，此设置栏就是用来指定这个随机项的分布形式的，有如下 4 个可选项。

- Residuals，随机误差项的分布和由方程导出的残差项相同。
  - Normal variates，随机误差项服从正态分布，且均数为 0，标准差为回归方程的均方误差的平方根（RMSE）。
  - Student's t variates，随机误差项服从 t 分布，且均数为 0，标准差为回归方程的均方误差的平方根（RMSE）。在 Degrees of freedom 输入框指定 t 分布的自由度。
  - None，不添加随机误差项，直接用初始估计值替换缺失值。
- ② Maximum number of predictors 输入框，指定能进入回归方程的自变量的最大个数。
- ③ Save completed data 复选框，与图 18-5 所示 EM 算法的保存选项含义相同。

### 18.3 缺失值分析的实例

缺失值分析的数据可以是连续的，也可以是分类的。对每个变量，非系统的缺失值必须定义为用户缺失值，例如：问卷的某个问题以“5”表示答案“不知道”，要把这个取值当作缺失值处理，就必须把“5”定义为用户缺失值。

#### 1. 数据描述




本节对美国 1993 年常规的社会调查数据进行缺失值分析，这些数据摘自 SPSS 自带的 Demo 文件“GSS 93 for Missing Values.sav”。所用数据文件为“美国 93 年常规社会调查.sav”，数据格式如图 18-7 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	id	Numeric	4	0	ID	None	None	6	Right	Ordinal
2	sex	Numeric	1	0	性别	{1 Male}	None	1	Right	Nominal
3	income91	Numeric	2	0	家庭总收入	{0 NAP}	0 98 99	8	Right	Scale
4	rincom91	Numeric	2	0	受访者收入	{0 NAP}	0 98 99	8	Right	Scale

图 18-7 美国 1993 年常规的社会调查数据格式

由于“收入情况”是一个涉及个人隐私的问题，数据中难免包括缺失记录，故而很有必要进行缺失值分析。在原始数据里，家庭总收入、受访者收入两个变量的用户定义缺失值都为离散值 0、98、99，其中：0 代表收入较少不愿透露（取值“NAP”）；99 代表收入较多不愿透露（取值“NA”）；98 代表无法定量化的模糊回答（取值“DK”）。本节就来对这些缺失信息，用 Missing Value Analysis 过程加以分析，以备更深入地了解和完善调查群体的收入情况。

#### 2. 参数设置

依次单击菜单“Analyze→Missing Value Analysis...”，打开缺失值分析的主设置面板，如图 18-8 所示。在变量列表选中家庭总收入和受访者收入变量，单击从上至下第一个  按钮，将其作为连续分析变量选入 Quantitative 列表；在变量列表单击选中性别变量，单击从上至下第二个  按钮，将其作为分类分析变量选入 Categorical 列表；在变量列表单击选中 ID 变量，单击从上至下第三个  按钮，将其作为标签变量选入 Case Labels 选框。分别勾选 Estimation 栏的 3 个复选框：Pairwise、EM 和 Regression。



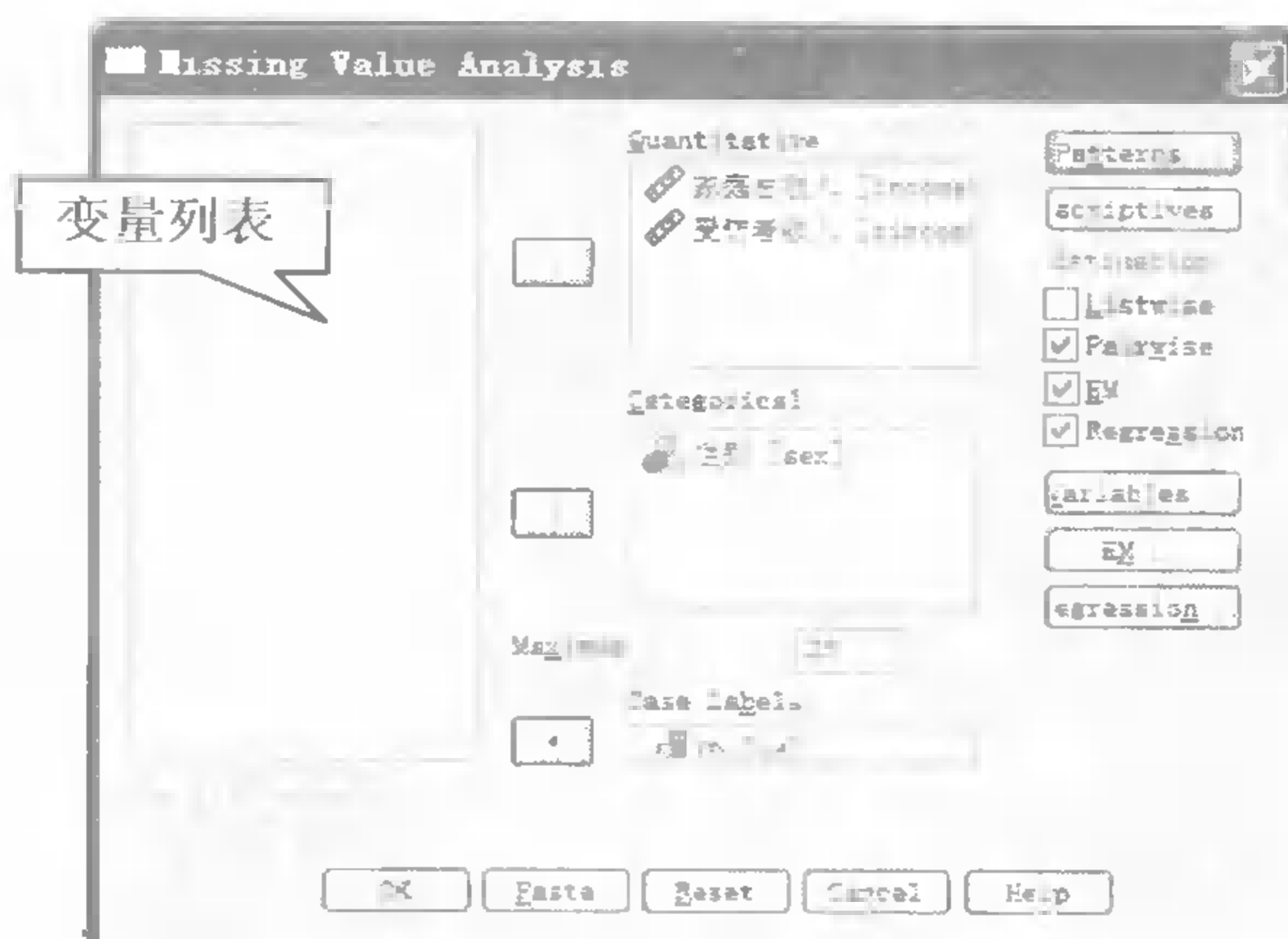


图 18-8 缺失值分析案例的主设置界面

在图 18-8 中，单击 Patterns 按钮，弹出如图 18-9 所示的对话框，勾选 Tabulated cases 复选框；单击 Continue 按钮返回主界面。

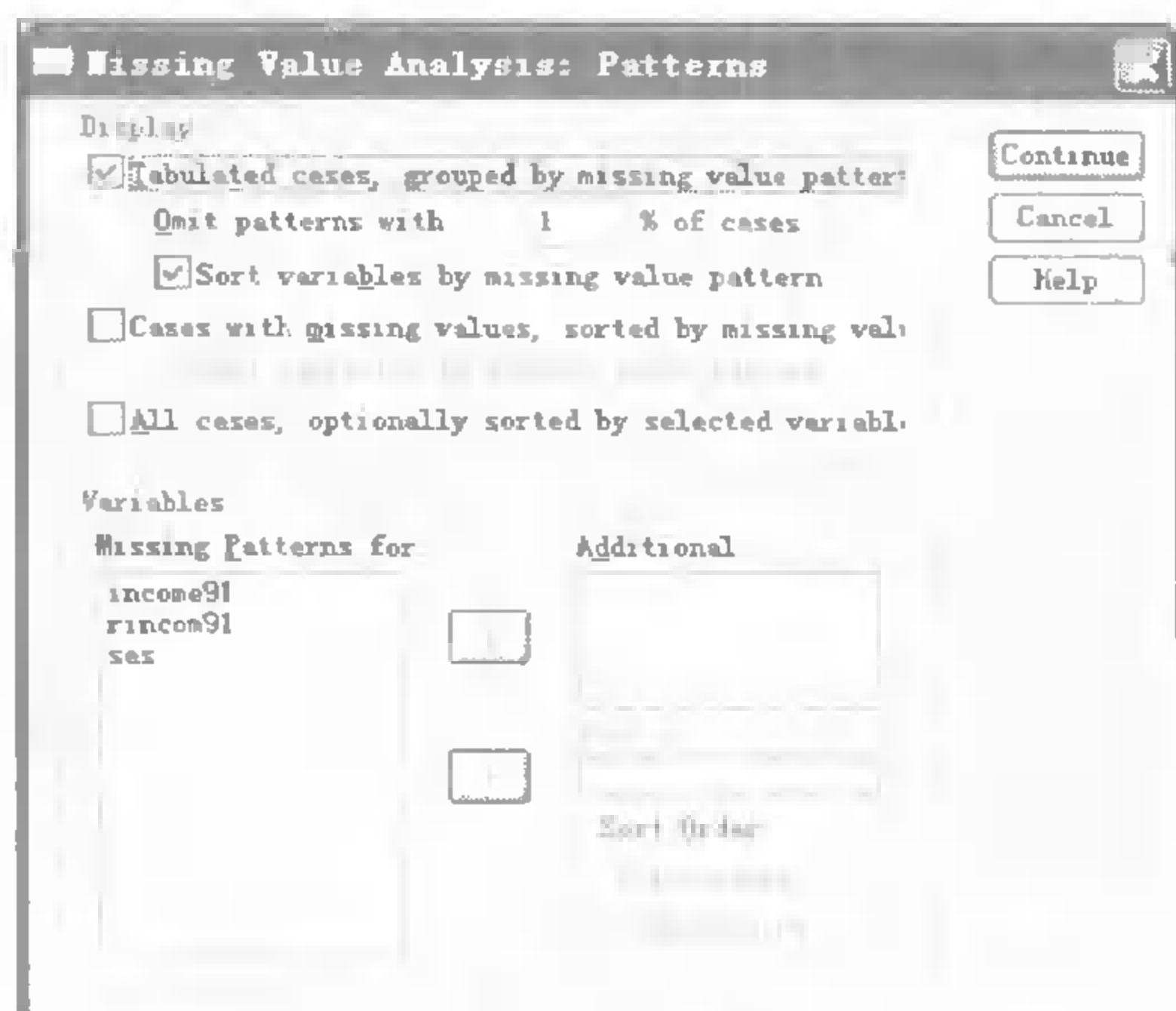


图 18-9 缺失值分析的 Patterns 设置

在图 18-8 中，单击 Descriptives 按钮，弹出如图 18-10 所示的对话框，勾选 Crosstabulations 复选框；单击 Continue 按钮返回主界面。

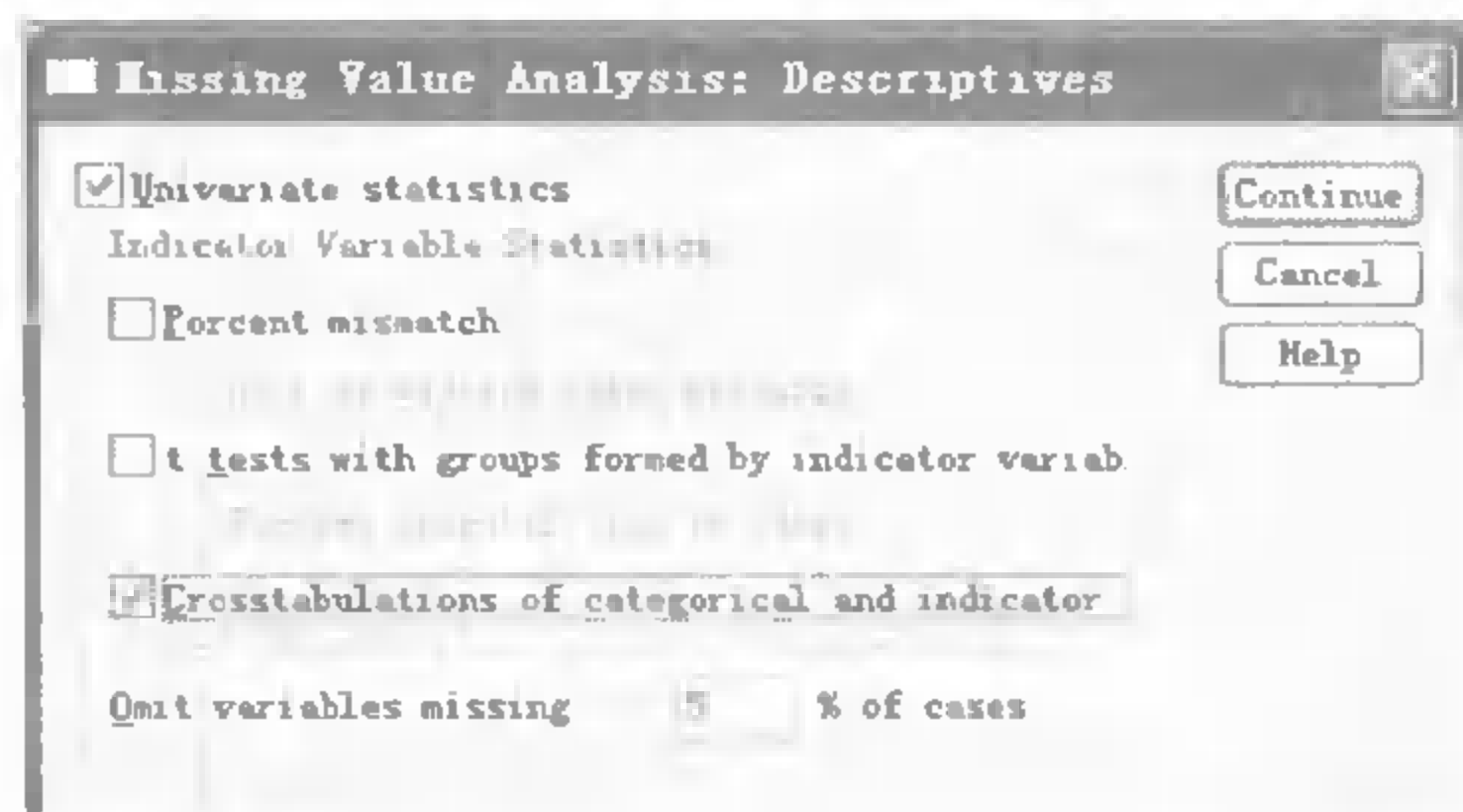


图 18-10 缺失值分析的 Descriptives 设置

### 3. 结果分析

在图 18-8 中单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 18-11~图 18-15 所示。

单变量统计							
	N	均值	标准差	缺失		极值数目 <sup>a</sup>	
				计数	百分比	低	高
income91	1434	14.68	5.462	56	3.4	0	9
rincom91	994	12.80	5.621	506	33.7	0	11
sex	1500			0	0		

a. 超出范围 ( $Q1 - 1.5 * IQR$ ,  $Q3 + 1.5 * IQR$ ) 的案例数。

图 18-11 关于缺失值的基本统计信息

估计均值摘要		
	income91	rincom91
所有值	14.68	12.80
EM	14.66	11.74
回归	14.73	12.10

估计标准差摘要		
	income91	rincom91
所有值	5.462	5.621
EM	5.474	6.118
回归	5.425	5.778

图 18-12 关于缺失值的估计结果

sex					
		Male	Female		
rincom91	存在	482	512		
	计数	994	512		
	百分比	66.3	59.6		
	缺失				
	% NAP	32.9	39.3		
	% DK	6	7		
	% NA	3	3		

用户定义  
缺失值

不显示少于 5% 个缺失值的指示变量。

制表模式			
	缺失模式 <sup>a</sup>		如果 非整数, 则取整
案例数	sex	income91	rincom91
979			
455		X	
51		X	X

不显示少于 1% 个 (15 个或更少) 案例的模式。

a. 以缺失模式排列变量。

b. 完整案例数, 如果未使用该模式 (用 X 标记) 中缺失的变量。

图 18-13 交叉制表和缺失值样式表

成对频率			
	income91	rincom91	sex
income91	1434		
rincom91	979	994	
sex	1434	994	1500

成对均值		
	income91	rincom91
income91	14.68	12.86
rincom91	15.95	12.80
sex	14.68	12.80

存在其他变量时定量变量的均值。

成对标准差		
	income91	rincom91
income91	5.462	5.566
rincom91	4.642	5.621
sex	5.462	5.621

存在其他变量时定量变量的标准差。

成对协方差		
	income91	rincom91
income91	29.828	
rincom91	18.265	31.597

成对相关系数		
	income91	rincom91
income91	1	
rincom91	.707	1

图 18-14 成对统计量的交叉表

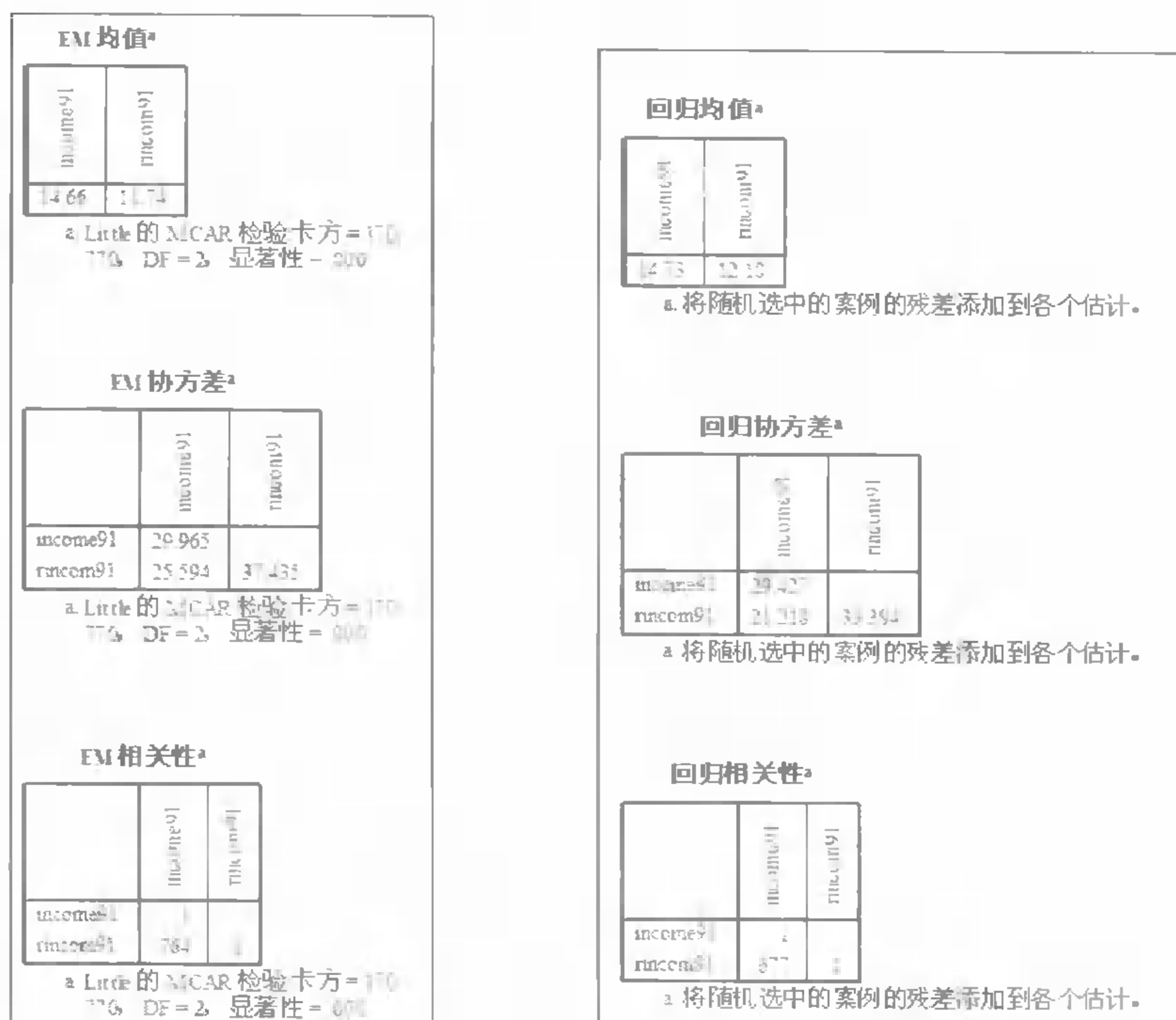


图 18-15 关于缺失值估计算法的输出

① 变量基本统计信息。如图 18-11 所示,“单变量统计”表格给出三个分析变量中未缺失数据的频数、均值、标准差,以及缺失值的个数和百分比,还有极值的统计频数。通过这些指标,可以初步了解数据的概貌特征,例如:income91(家庭总收入)的有效记录有 1 434 条,它们的均值为 14.68,标准差为 5.462,它有 66 条记录缺失,所占比例为 4.4%。

② 缺失值估计结果。如图 18-12 所示,为使用 EM、Regression 两种算法进行缺失值的估计和替换后,总体数据的均值和标准差的变化情况。“所有值”行为原始数据的统计特征,与图 18-11 中对应的数据一样;“EM”、“回归”两行分别是用指定算法估计后,总体数据的统计特征。

可见,EM 算法的估计结果,均值都比初始均值小,且标准差都比初始值大;Regression 算法的估计结果,均值都比 EM 估计的大,且标准差都比 EM 估计的小。由此特点可以根据实际问题的需要,从中选择一个合适的估计算法加以利用。

③ 指示变量与分类变量的交叉制表和缺失值样式表。如图 18-13 所示,“sex”表格为类别变量(sex, 性别)与指示变量 rincom91(受访者收入)的取值交叉表,给出在不同性别下,非缺失的个数(“存在”行)及其比例,和各种缺失值的出现比例,图中标识了三个用户定义缺失值的取值,在不同性别的人中的分布情况。例如:所有男性受访者中,rincom91(受访者收入)变量的有效记录占 75.2%,收入较少而不愿透露(NAP%)的比例为 24.2%。

“制表模式”表格就是缺失值样式表,它给出了缺失值分布的详细信息,例如:3 个变量都没有缺失的观测有 979 例;只有 rincom91(受访者收入)缺失的观测有 455 例,sex(性别)和 income91(家庭总收入)都没有缺失的观测有 1 434 例;只有 rincom91 和 income91 都缺

失的观测有 51 例，sex 没有缺失的观测有 1 500 例。

④ 几个成对统计量的交叉表。如图 18-14 所示，是几个成对变量间的交叉统计表，包括频率、均值、标准差、协方差和相关系数表，由此可以观察和分析变量之间的相互关系与影响。

例如，在成对频率表里，受访者收入（rincom91）和家庭总收入（income91）都没有缺失的记录有 979 例，rincom91 没有缺失的记录有 994 例；在成对均值表里，income91 行和 income91 列的交叉单元格的值，就是图 18-11 中的 income91 的均值 14.68；另外，由于 sex 变量没有缺失值，所以 sex 行的取值和相应的对角线上的取值都相等。

⑤ 几个成对统计量的交叉表。如图 18-15 所示，给出分别用 EM、Regression 算法估计后的均值、协方差和相关性的表格，其中的部分数据在图 18-12 中已有所提及。



# 第 19 章 统计图形

本章介绍如何利用 SPSS 绘制统计图形，以及如何编辑生成的图形。统计图形是用点的位置、线段的升降、直条的长短或面积的大小等描绘资料数据和分析结果的一种方式，其特点是简明、生动、形象易懂。掌握如何利用统计图形来分析问题是很重要的。

统计图形分为许多种，包括：条形图、线型图、面积图等，不同图形可能有着不同的数据要求和适用环境，使用时一定要考虑好每种统计图的功能和特点。

## 19.1 概述

SPSS 中直接绘制统计图形的功能通过 Graphs 菜单实现，它下设的子菜单有：Chart Builder（图形构建器）、Interactive（交互式图形）、Legacy Dialogs（旧版图形）、Map（地图图形），本章主要介绍 Chart Builder、Interactive 和 Legacy Dialogs 的使用。另外，统计图形还可以伴随着其他分析过程而输出，例如回归分析过程、方差分析过程等。

### 19.1.1 数据和变量的准备

#### 1. 打开数据文件

在创建图形之前，Data Editor 窗口必须已有数据，否则单击 Graphs 菜单下的某个作图选项后，会弹出如图 19-1 所示的对话框，提示用户打开数据文件。单击 Open Data File 按钮，打开文件选择对话框，指定包含作图数据的文件路径和名称。



图 19-1 打开文件的提示对话框

#### 2. 定义变量性质

变量类型对建立图形是非常重要的，在开始创建图形之前，建议用户先检查数据类型是否适用于要建立的图形。如果 Data Editor 窗口已有数据，依次单击菜单“Graphs→Chart Builder”，首先弹出如图 19-2 所示的对话框，提示用户要正确定义数据类型。

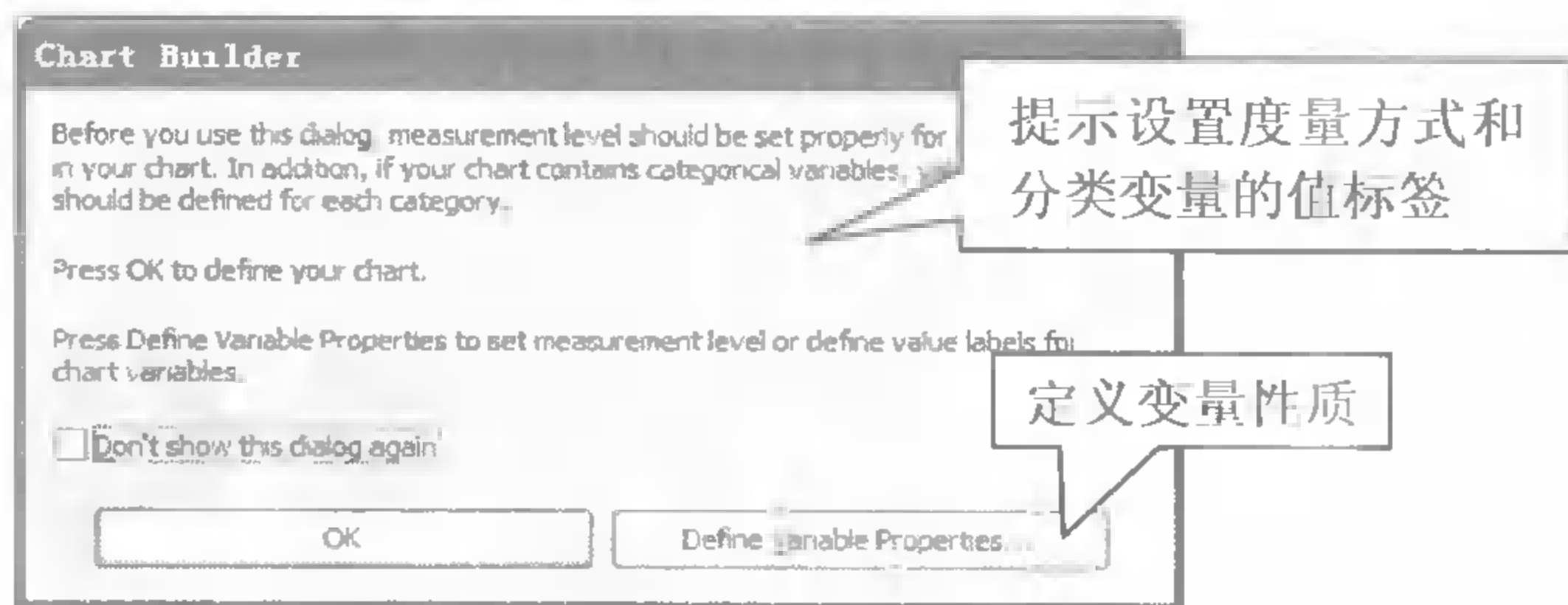


图 19-2 定义变量性质的提示对话框

在图 19-2 中，单击 OK 按钮将进入图形构建器的操作界面；单击 Define 按钮，会弹出如图 19-3 所示的定义变量性质设置面板，此界面也可以通过依次单击菜单“Data→Define variable properties”打开；勾选 Don't 复选框表示不再出现此对话框，否则每次使用 Chart Builder 作图时都会弹出此提示对话框。

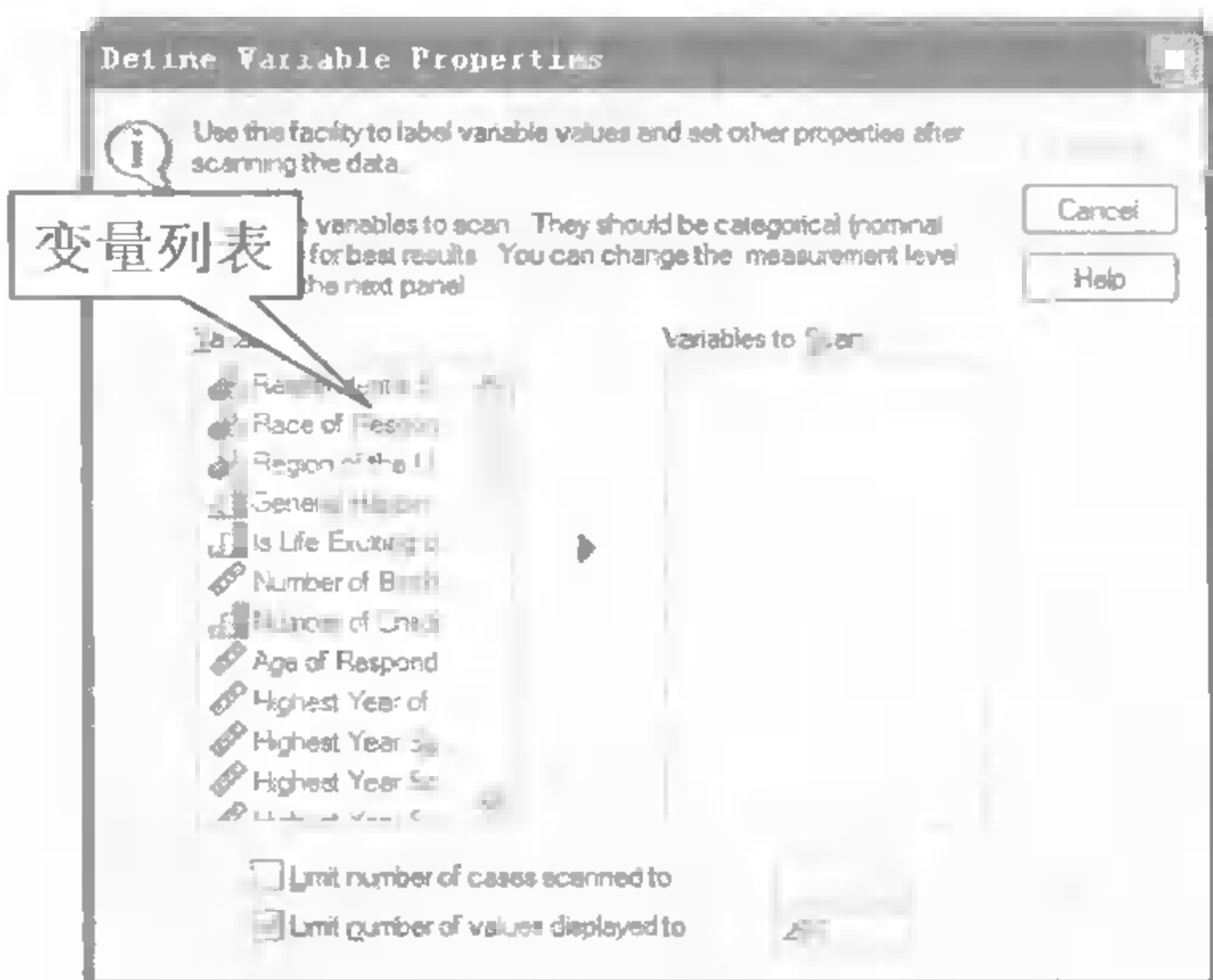


图 19-3 定义变量性质的设置界面

在图 19-3 中，需要设置的变量性质主要包括两点：度量方式和分类变量的值标签。

变量的度量方式用于区分连续变量、序数分类变量、名义分类变量（无顺序）、集合变量等数据类型，其中集合变量又分为两种，它们的图标表示如图 19-4 所示，这些图标可以帮助用户在构建图形时方便地识别变量的度量方式，如图 19-3 中的变量列表所示。

Measurement Level	Data Type			
	Numeric 数值型	String 字符串型	Date 日期型	Time 时间型
Scale 数值范围		n/a		
Ordinal 序数变量				
Nominal 名义变量				
Multiple response set, multiple categories 集合变量，多分类变量				
Multiple response set, multiple dichotomies 集合变量，多二分变量				

图 19-4 变量的度量水平

另外除了变量类型之外，数据文件的组织形式对创建图形也是非常重要的，关于“横向数据”和“纵向数据”对于作图的影响，将在第 19.4 节中以实例的方式加以介绍。

### 19.1.2 图形构建器的基本操作

依次单击菜单“Graphs→Chart Builder”，打开图形构建器的操作界面，如图 19-5 所示，在此主要通过拖曳的方式建立图形，变量列表中的变量、预设图形类型中的图标等都可以用鼠标拖动至图形预览区，在预览区还可以拖动这些元素以改变它们的位置或拖出预览区。

#### 1. 更改变量类型

如图 19-5 所示，在变量列表单击选中某个变量后，下面的 Categories 栏会显示关于选中变量的某些信息：如果选中的是分类变量（Nominal 或 Ordinal），此处显示它的值标签；如果选中的是连续变量，此处显示“No Categories (Scale Variable)”，表示无分类取值预览。

右击变量列表中的某个变量，在弹出的快捷菜单中可以临时更改变量的类型以适合作图，这不会改变数据文件中实际的数据类型。

#### 2. 元素属性设置

在图 19-5 中，单击 Element Properties 按钮，弹出如图 19-6 所示的元素性质设置面板，所作图形不同，此面板的显示内容也会有所不同。

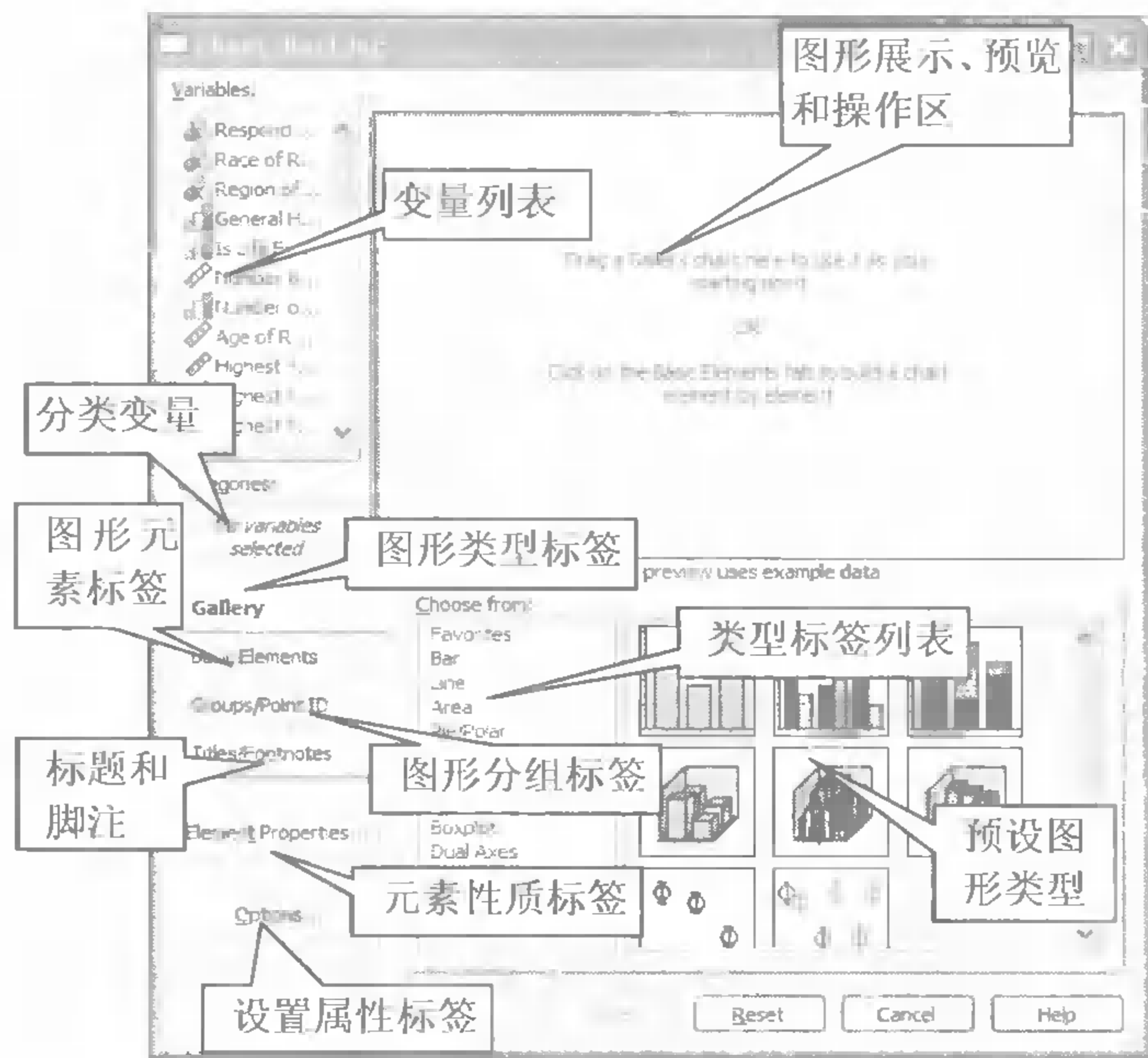


图 19-5 Chart Builder 界面一览



图 19-6 元素属性设置

在此面板中对某项元素的性质做了修改之后，需要单击 Apply 按钮确认更改，否则方才的设置无效；如果没有单击 Apply 按钮，就要转换到其他元素的性质设置界面，会弹出如图 19-7 所示的对话框，提示用户是否应用刚刚进行的设置，单击“是”按钮确认更改。

#### 3. 作图的 Options 选项设置

在图 19-5 中单击 Options 按钮，弹出如图 19-8 所示的选项设置面板，在此设置作图时如何处理缺失值、选用哪些图形面板等内容。

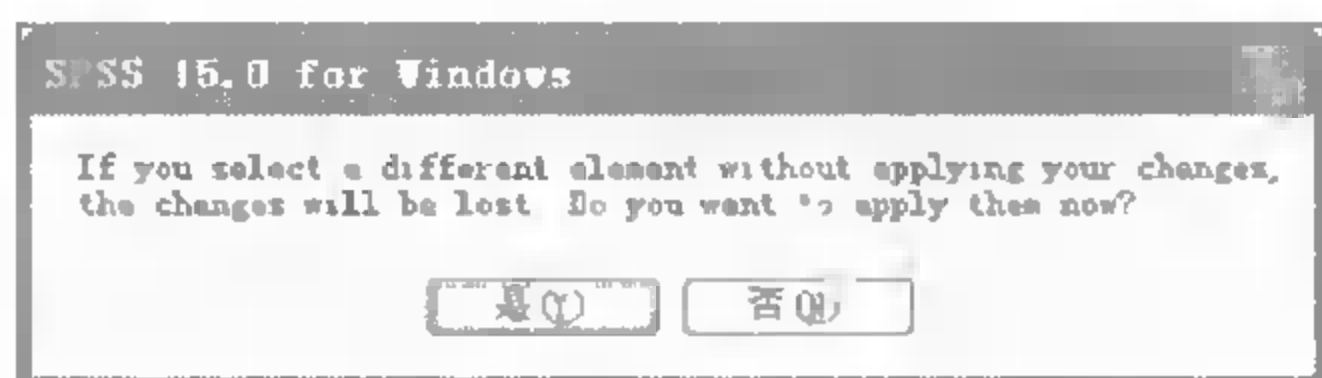


图 19-7 提示应用设置对话框

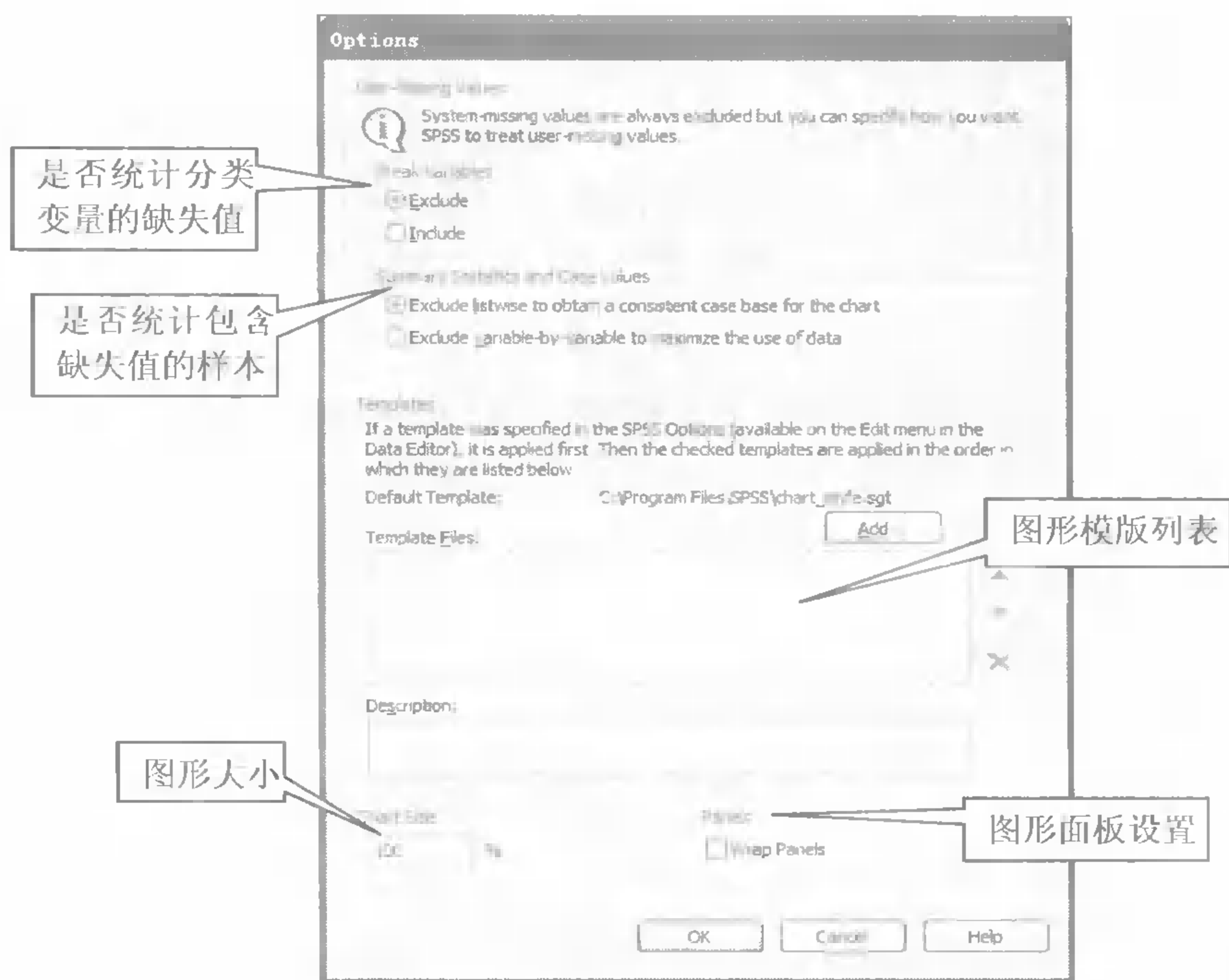


图 19-8 设置属性面板

(1) Break Variables 栏，对于系统缺失值，SPSS 在作图时都不加以统计；对于分类变量的用户定义缺失值，有如下两种处理方式。

- Include 单选框，表示作图时把它们作为一个单独的类别加以统计。
- Exclude 单选框，表示作图时忽略这些用户定义缺失值。

(2) Summary Statistics and Case Values 栏，如果某个观测（Case）变量出现了用户定义缺失值，作图时有以下两种处理方法。

- Exclude listwise 单选框，表示作图时直接忽略这个观测。
- Exclude variable-by-variable 单选框，表示只有包含缺失值的变量用于当前计算和分析时才忽略这个样本。

(3) Template Files 列表框，用于设置关于作图的模板文件。

单击 Add 按钮，打开文件选择对话框，添加指定的预置模板文件。

作图时，最先使用的是系统默认模版，它通过菜单项“Edit→Options”设置，然后会按照此模版列表中显示的顺序依次使用，靠后显示的模板将会覆盖前面的模版效果。

(4) Chart Size 输入框，设置图形显示的大小。

在此指定一个相对于默认图形的比例数字，大于 100 的数值有放大效果，反之亦然。

(5) Wrap Panels 复选框，设置有多列子图形时的显示方式。

选中它表示图形列过多时允许分行显示，这样可以避免把几列图形强制显示在同一行上；否则图形列过多时，每行上的图形会自动缩小以显示在同一行上。

### 19.1.3 交互式作图和对话框作图

关于作交互式图形（Interactive）的设置界面、旧版对话框（Legacy Dialogs）式的作图界面，以及图形构建器的详细使用方法，请参看第 19.2 节的具体操作，在那里以条形图为例



例详细讲解了这些界面的一般操作，其他统计图形的作图过程，也都可以以此为参考。

### 19.1.4 图形的编辑

输出的统计图形，大多以一般图形或交互式图形的方式显示在 SPSS Viewer 输出窗口。

在右击图形弹出的快捷菜单中选择“SPSS \*\*\* Object”选项，如果是一般图形，会自动打开 Chart Editor 窗口，在其中可以对图形作更多的设置和润色；如果是交互式图形，会直接使其进入内嵌于 Viewer 窗口的编辑状态，也可以做更进一步的编辑和修改。

在 Viewer 窗口里双击相应的图形，也可以达到同样的效果。

## 19.2 条形图

条形图用直条的长短体现非连续性资料的特征，适用于描绘分类变量（Nominal 或 Ordinal）的取值大小、取值比例等特点，常用的条形图类型有简单条形图、分类条形图和堆积条形图等。

### 19.2.1 数据和问题描述

某研究者搜集了 500 名司机在过去 5 年内发生车祸次数的数据，数据格式如图 19-9 所示，所用数据文件为“autoaccident.sav”。

	Name	Type	Width	Decimals	Label	Values	Missing	Column	Align	Measure
1	gender	Numeric	2	0	性别	{1 男}	None	8	Right	Nominal
2	age	Numeric	2	0	年龄	None	None	8	Right	Scale
3	accident	Numeric	2	0	5年来车祸次数	None	None	8	Right	Scale
4	age_step	Numeric	1	0	年龄段	{1 <=25}	None	8	Right	Ordinal

图 19-9 车祸数据的变量含义

本节通过创建条形图观察不同年龄段、不同性别的人，发生车祸的次数分布有何特点。

### 19.2.2 用图形构建器作条形图

依次单击菜单“Graphs→Chart Builder”，打开图形构建器的操作界面，如图 19-10 所示，单击 Gallery 标签；在 Choose from 列表框单击选中 Bar，在其右侧列出预设的条形图图标。

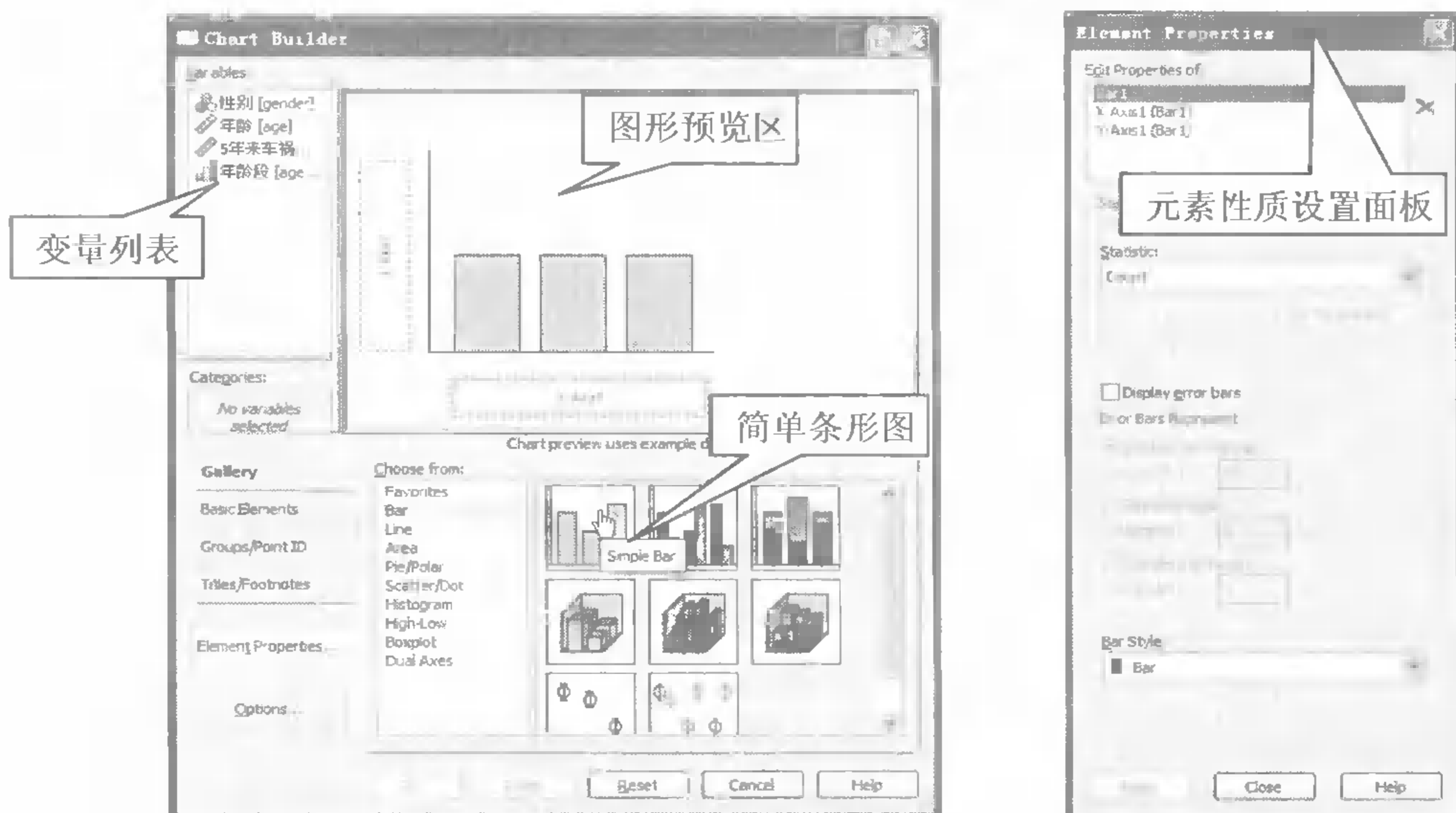



图 19-10 创建简单条形图

## 1. 简单条形图

下面先来作平均年龄随车祸次数变化而变化的简单条形图。

(1) 参数设置。在图 19-10 中, 双击预置图标  (Simple Bar), 就会在图形预览区给出简单条形图的预览, 同时自动弹出元素性质设置面板: 把预置图标拖动至图形预览区, 可以达到相同的效果。

① 指定作图变量。在变量列表中右击 5 年来车祸次数变量, 在弹出的快捷菜单里选中 Ordinal 选项, 指定它为有序分类变量。从变量列表中把变量 (其实是变量标签) 5 年来车祸次数、年龄, 分别拖动至预览区的 X-Axis、Y-Axis 两个虚线框中, 将其分别作为条形图的 X、Y 坐标轴。

图 19-11 所示是设置好了作图变量的操作界面。

② 设置图形元素的属性。图形里各种元素的属性都在 Element Properties 对话框中进行设置, 如图 19-12 所示。

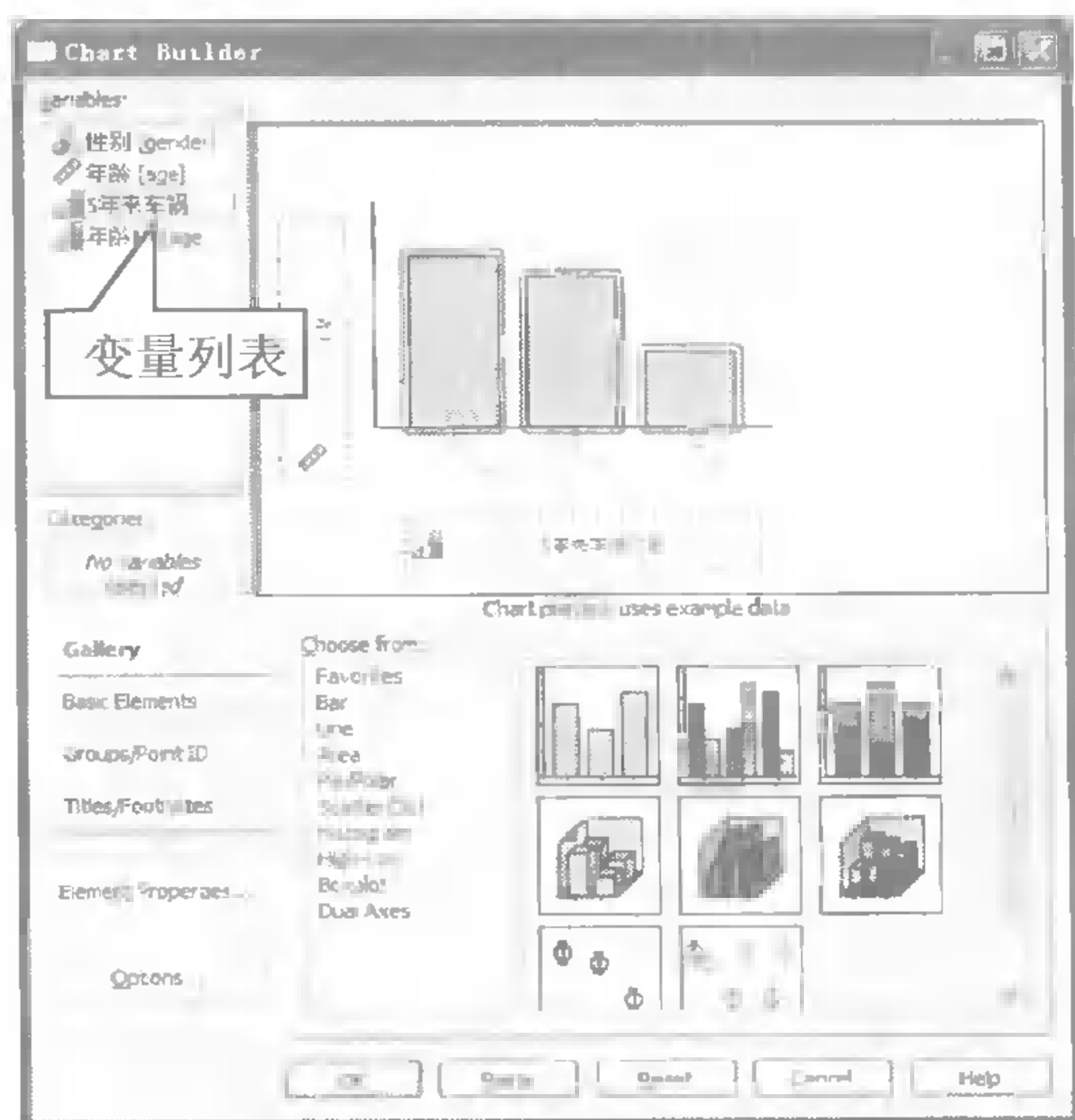


图 19-11 指定作图变量变量

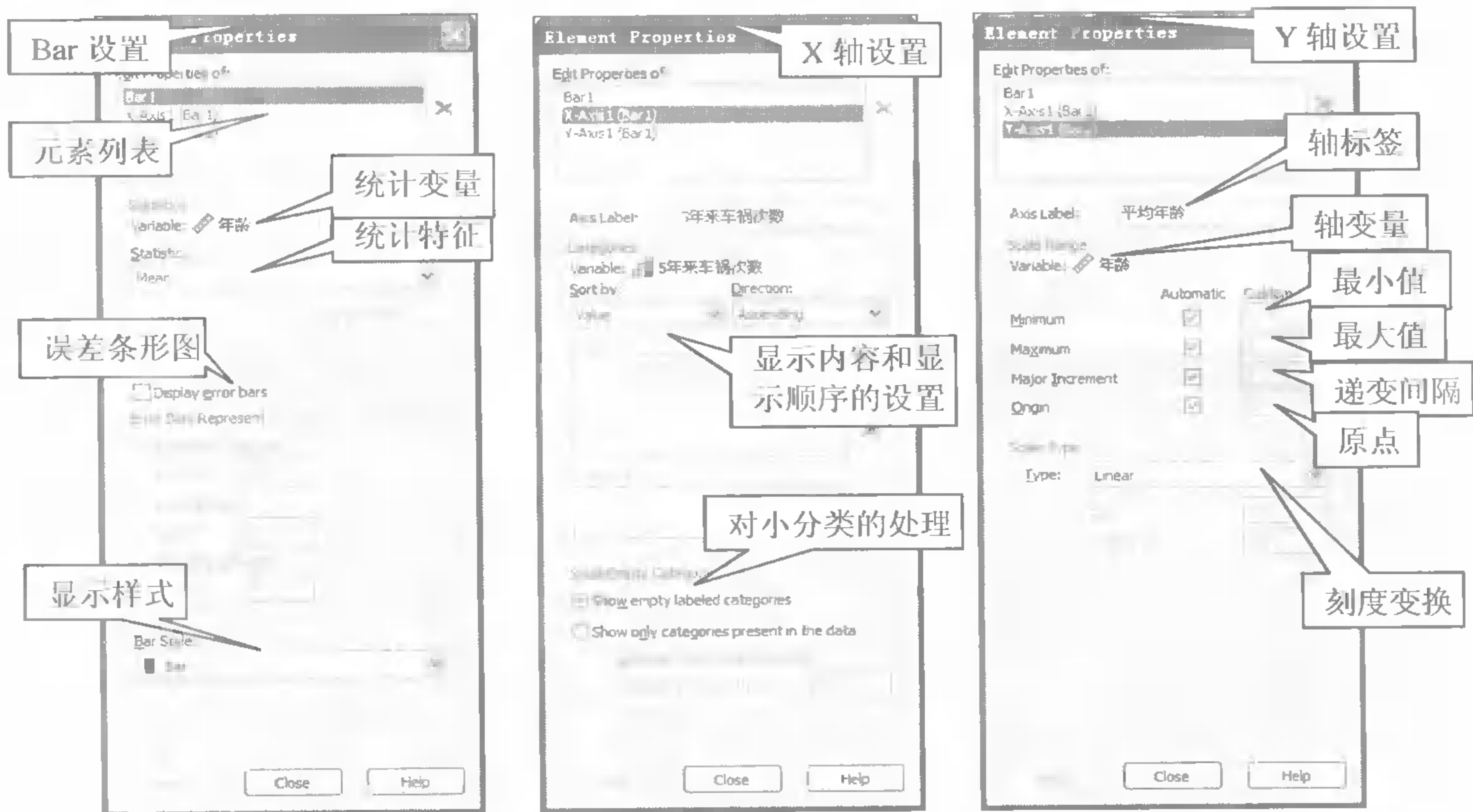


图 19-12 简单条形图的参数设置

在“Bar 设置”界面, 单击选中 Edit 列表框里的 Bar1; 单击 Statistic 下拉列表并选中 Mean 选项, 表示条形图将代表年龄变量的均值大小; 单击 Apply 按钮应用设置。

在“X 轴设置”界面, 单击选中 Edit 列表框里的 X-Axis1; Axis Label 输入框自动显示了当前 X 轴变量的标签; 保留默认设置。

在“Y 轴设置”界面，单击选中 Edit 列表框里的 Y-Axis1；在 Axis Label 输入框键入“平均年龄”作为 Y 轴标签；单击 Apply 按钮应用设置。

③ 添加标题和脚注。在图 19-11 中，单击 Titles/Footnotes 标签，打开如图 19-13 所示的选择界面，分别勾选 Title1、Footnote1 复选框；Element Properties 对话框的 Edit 列表将自动添加 Title1（标题 1）和 Footnote1（脚注 1）两个元素，如图 19-14 所示。

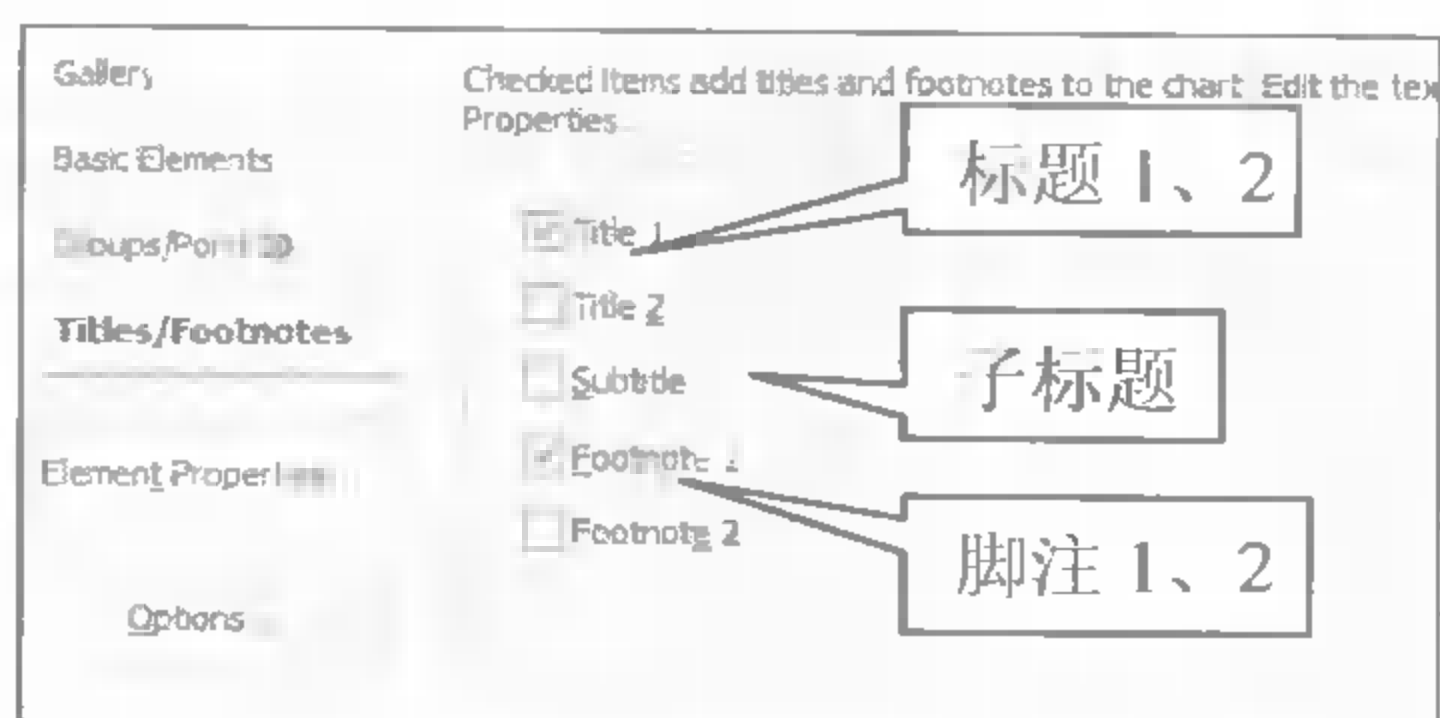


图 19-13 添加标题和脚注

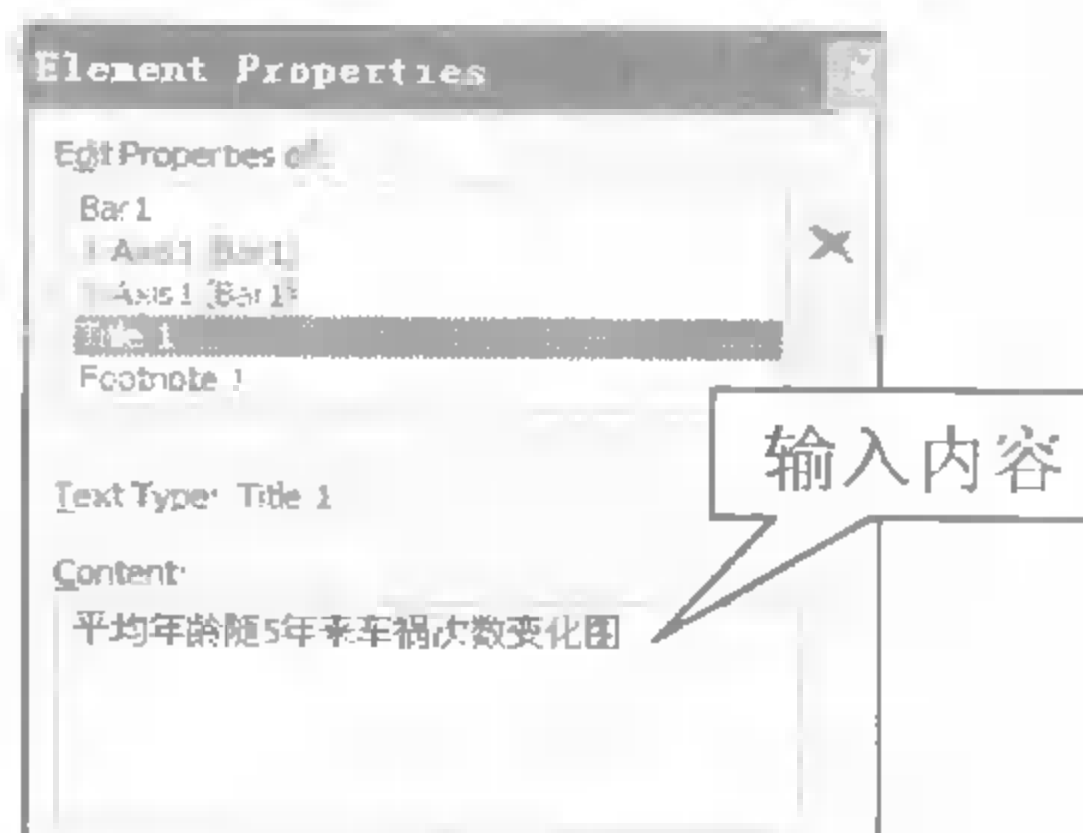


图 19-14 设置标题和脚注

在图 19-14 中，单击选中 Edit 列表框中的 Title1，在 Content 编辑框输入“平均年龄随 5 年来车祸次数变化图”作为标题 1；单击 Apply 按钮（图中未显示）应用设置。用同样的方法设置脚注 1 的内容为“平均年龄 VS 车祸次数”。

(2) 输出图形。在图 19-11 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-15 所示。可见，发生车祸次数最多时的平均年龄最大，而其他情况下的平均年龄并无明显差异。

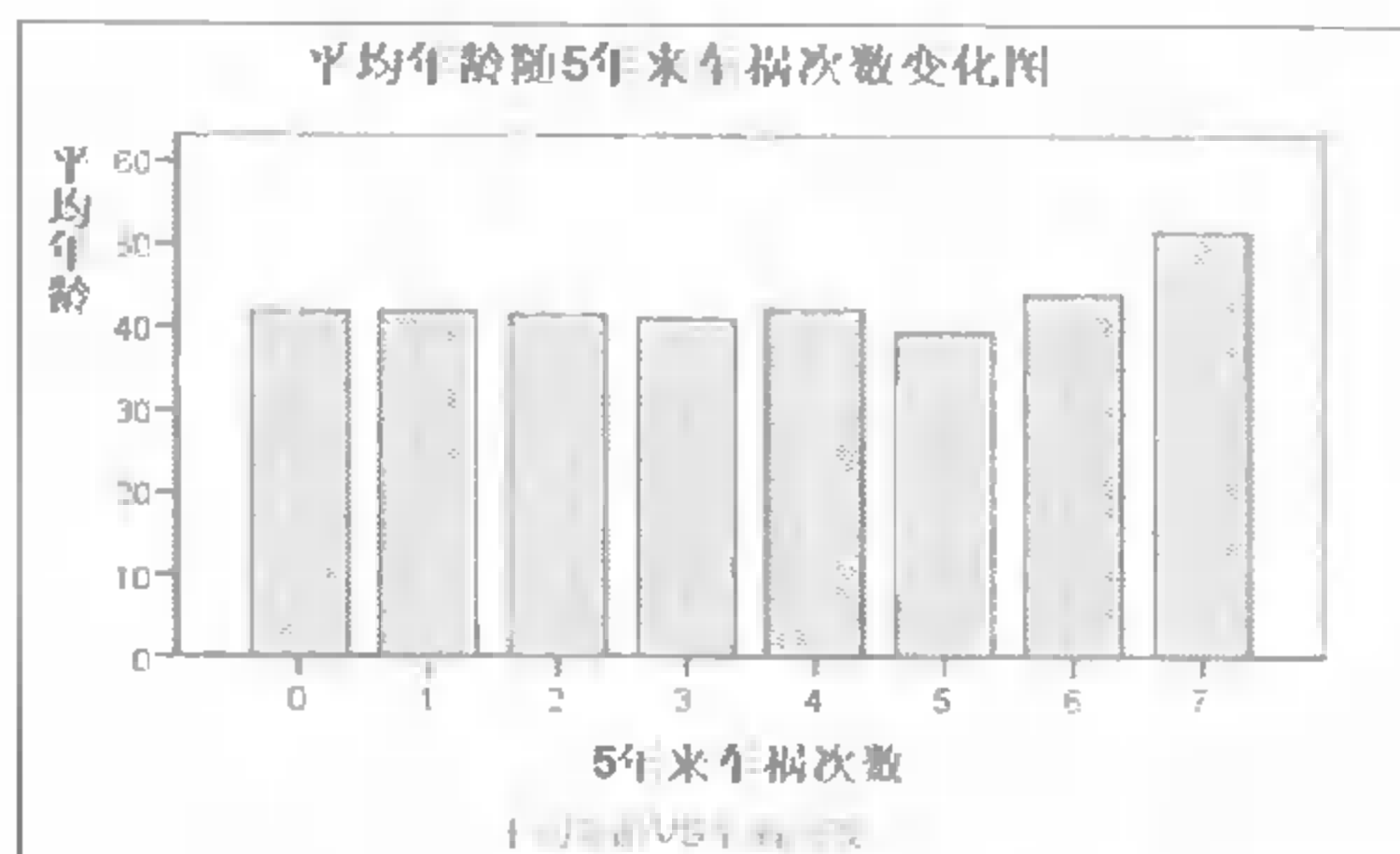


图 19-15 编辑后的简单条形图

## 2. 分类条形图

分类条形图能够反映更多的信息，它对 X 轴的每个取值再按照某个指标进一步细分，并作出关于所得子类别的条形图。随后作出在特定年龄段下不同性别的人发生车祸次数的分布图，观察年龄和性别对发生车祸次数的影响。

(1) 参数设置。在图 19-10 中，双击预置图标  (Clustered Bar)，在图形预览区给出分类条形图的预览，同时自动弹出 Element Properties 对话框；把预置图标拖动至图形预览区可以达到相同的效果。

① 指定作图变量。从变量列表中把变量年龄段、5 年来车祸次数、性别，分别拖动至预览区的 X-Axis、Y-Axis 和 Cluster 三个虚线框中，将其分别作为分类条形图的 X 坐标轴、Y 坐标轴和子类别变量。如图 19-16 和图 9-17 所示，是设置好了作图变量的操作界面。



图 19-16 指定了作图变量的分类条形图预览



图 19-17 子类别变量的性质

② 设置图形元素的属性。图形里各种元素的属性都在 Element Properties 对话框中进行设置。

如图 19-17 所示，单击选中 Edit 列表框里的 GroupColor；Legend Label 输入框自动显示了当前子分类变量的标签；保留默认设置。

单击选中 Edit 列表框里的 Bar1；单击 Statistic 下拉列表并选中 Mean 选项，表示条形图将代表车祸次数的均值大小；单击 Apply 按钮应用设置。

其他元素的属性设置方法，与在图 19-12 中介绍的操作方法相仿。

参考图 19-13 的设置方法，指定此分类条形图的标题 1 为“车祸次数对年龄和性别分布图”。

(2) 输出图形。在图 19-16 中单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-18 所示。可见年龄较小时，男性发生车祸的次数相对较小；随着年龄的增大，男性发生车祸的次数明显增多，而女性则有所下降。

### 3. 堆积条形图

与分类条形图类似，堆积条形图对 X 轴的每个取值都按照某个指标进一步细分，从而反映了更多的信息。不同的是堆积条形图不把子类别分散开来做条形图，而是将其逐次堆积在 Y 轴方向上，如此还能很好地比较总值的大小。

(1) 参数设置。在图 19-16 中，双击预置图标  (Stacked Bar)，在图形预览区给出堆积条形图的预览，同时自动弹出 Element Properties 对话框；把预置图标拖动至图形预览区可以达到相同的效果。

从变量列表中把年龄段、5 年来车祸次数、性别，分别拖动至预览区的 X-Axis、Y-Axis 和 Stack 三个虚线框中，将其分别作为堆积条形图的 X 坐标轴、Y 坐标轴和子类别变量。

其他参数的设置方法，与作分类条形图时完全相同。

(2) 输出图形。在主界面中，单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-19 所示。可见它除了反映出与图 19-18 相同的信息外，还较好地反映了在不同年龄段所有人发生



车祸次数的分布特点：比较平均，无明显波动。

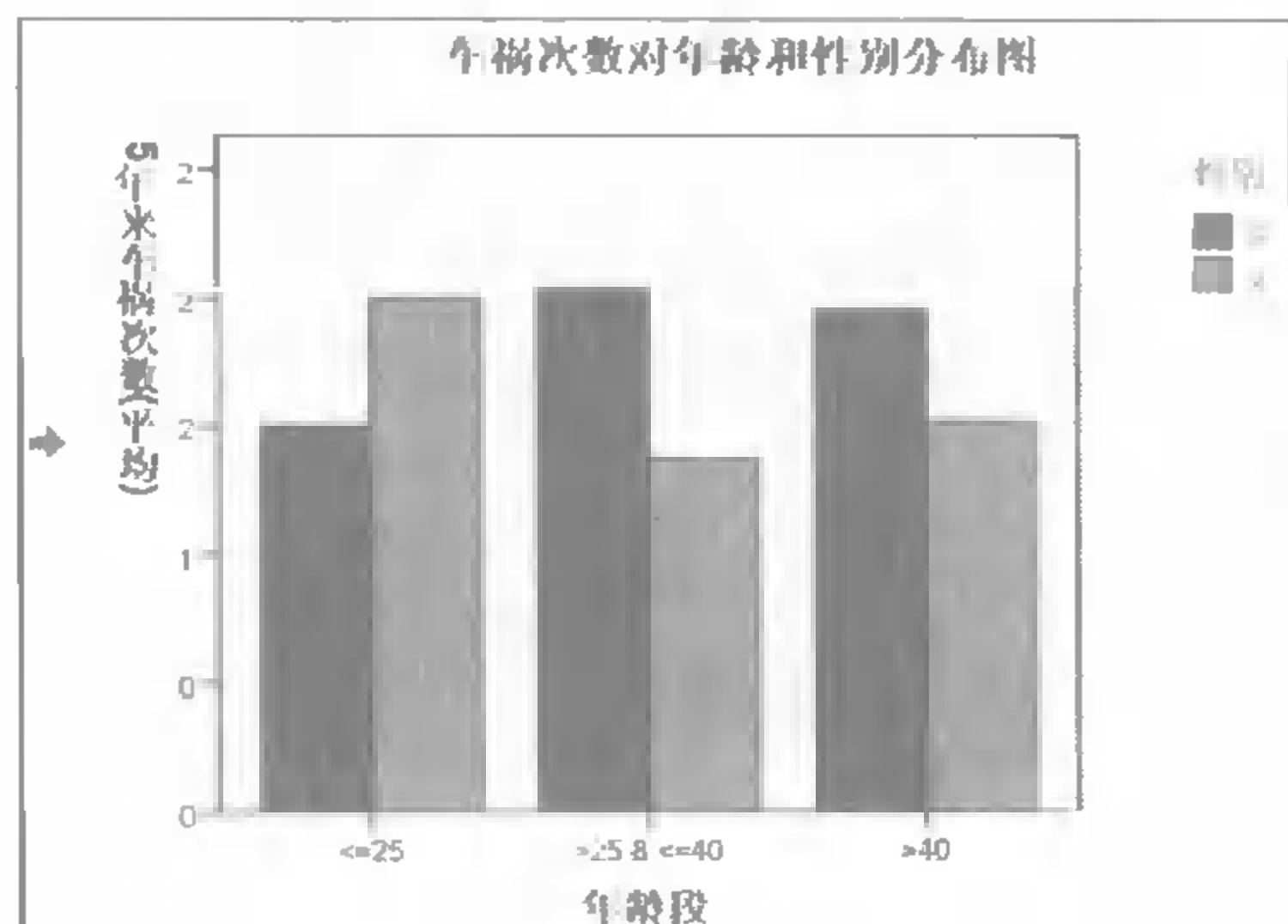


图 19-18 车祸次数对年龄和性别分类条形图

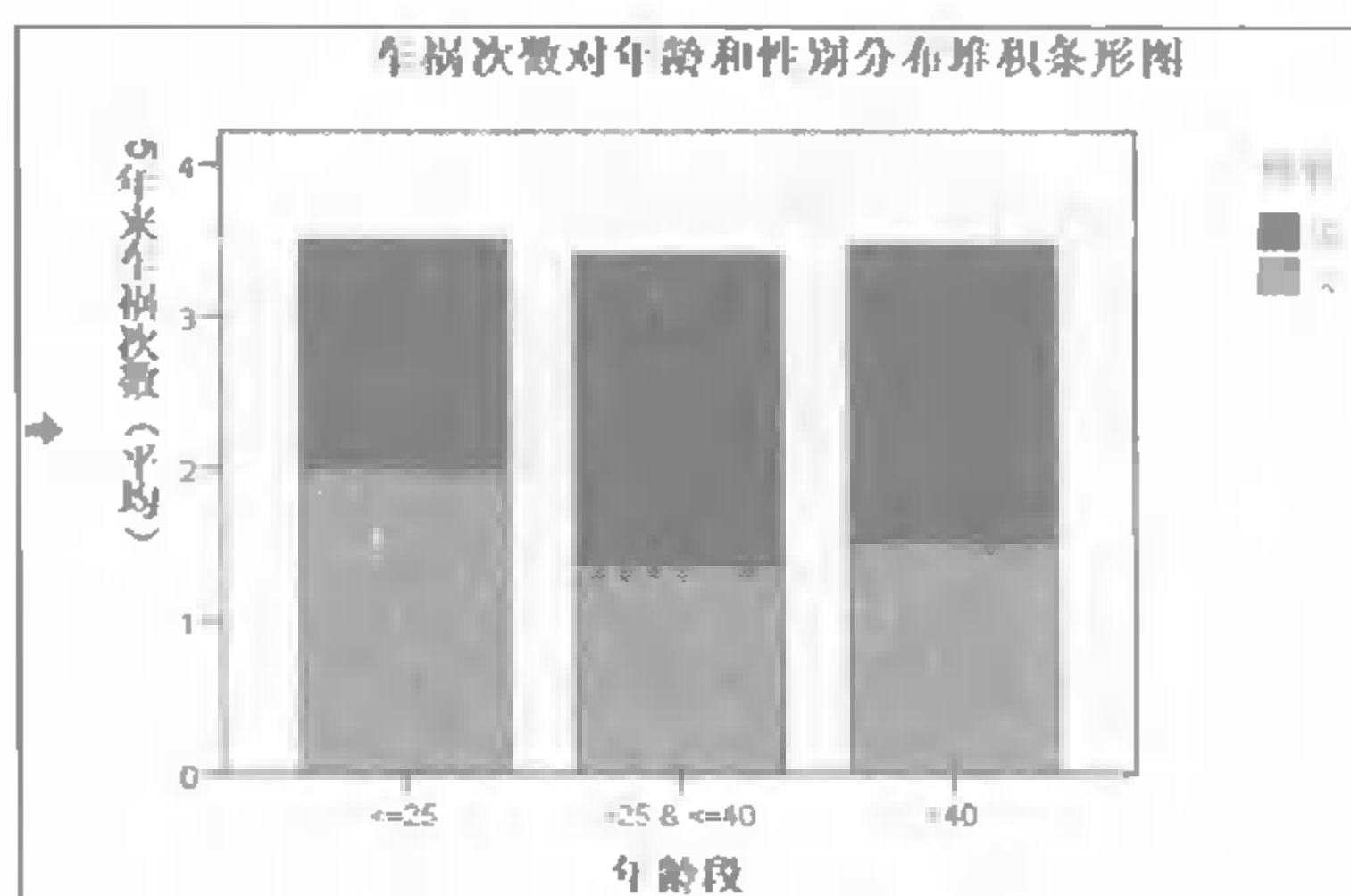





图 19-19 车祸次数对年龄和性别堆积条形图

#### 4. 三维条形图

如图 19-16 所示，在预置图标显示区的第二行，给出了 3 个代表三维条形图的图标：  ，分别表示简单三维条形图、分类三维条形图和堆积三维条形图。三维条形图的做法和二维时的情形完全类似。

#### 19.2.3 交互式条形图

交互式图形能方便地设置图形的个性化属性，使其达到最佳的显示效果。依次单击菜单“Graphs→Interactive→Bar...”，打开交互式条形图的作图界面，如图 19-20 所示。

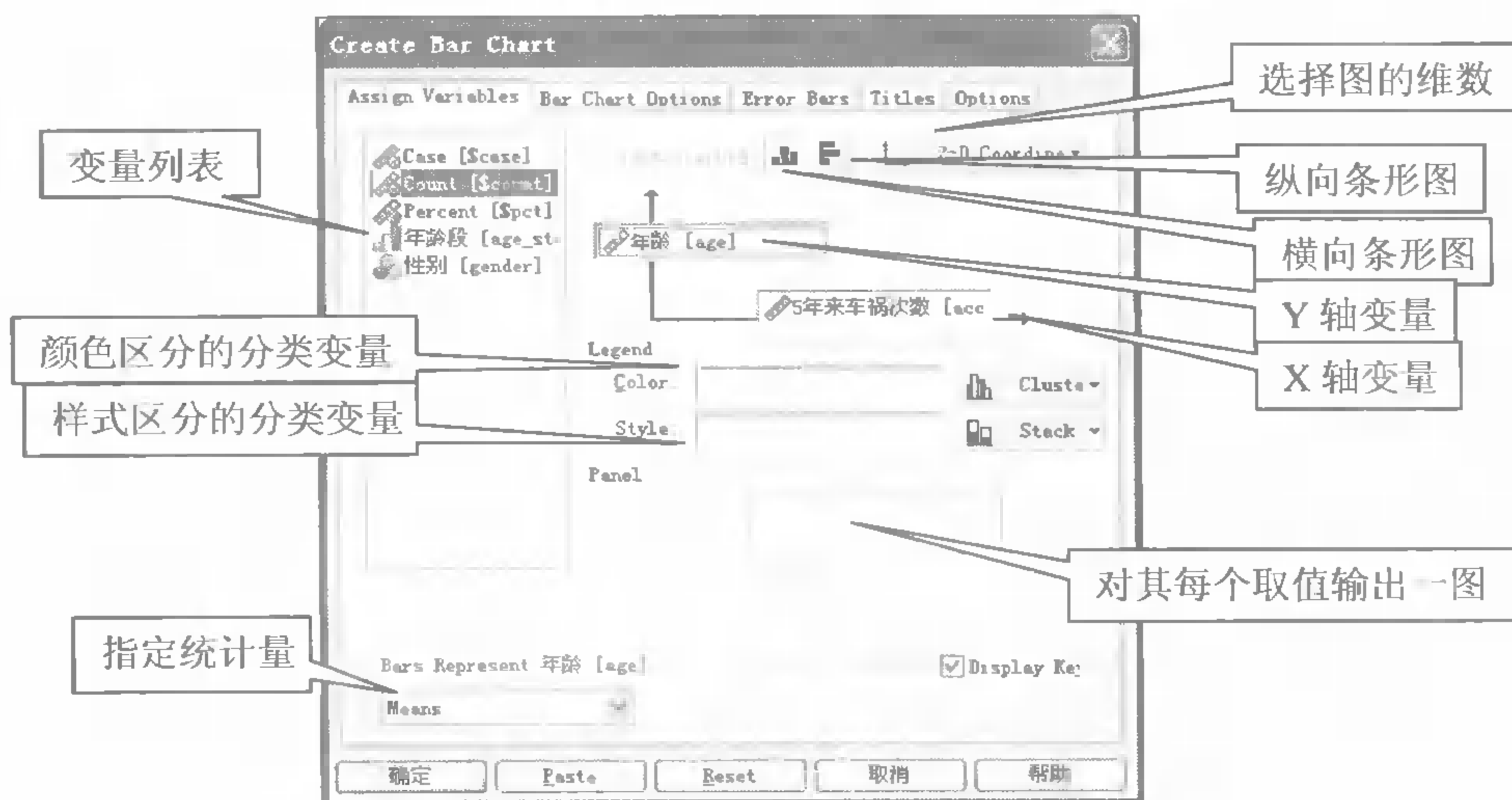


图 19-20 交互式条形图变量设置面板

##### 1. 交互式简单条形图

在此，作一个与第 19.2.2 节的简单条形图相似的交互式图形。

###### (1) 参数设置。

- ① 指定作图变量。从变量列表中把 5 年来车祸次数、年龄，分别拖动至 X 轴变量、Y 轴变量所示的选框，将其分别作为条形图的 X、Y 坐标轴。
- ② 条形图的样式设置。在图 19-20 中，单击 Bar Chart Options 标签，打开如图 19-21

所示的子设置界面，勾选 Value 复选框，表示在图中显示变量取值。

**标题和脚注的设置。**在图 19-20 中，单击 Titles 标签，打开如图 19-22 所示的子设置界面，在 Chart Title 编辑框输入“年龄随车祸次数变化条形图”作为标题；在 Chart Subtitle 编辑框输入“交互式简单条形图”作为子标题；在 Caption 编辑框输入“平均年龄随 5 年来车祸次数变化交互式简单条形图”作为脚注。单击 Assign 标签返回图 19-20 所示的变量设置界面。

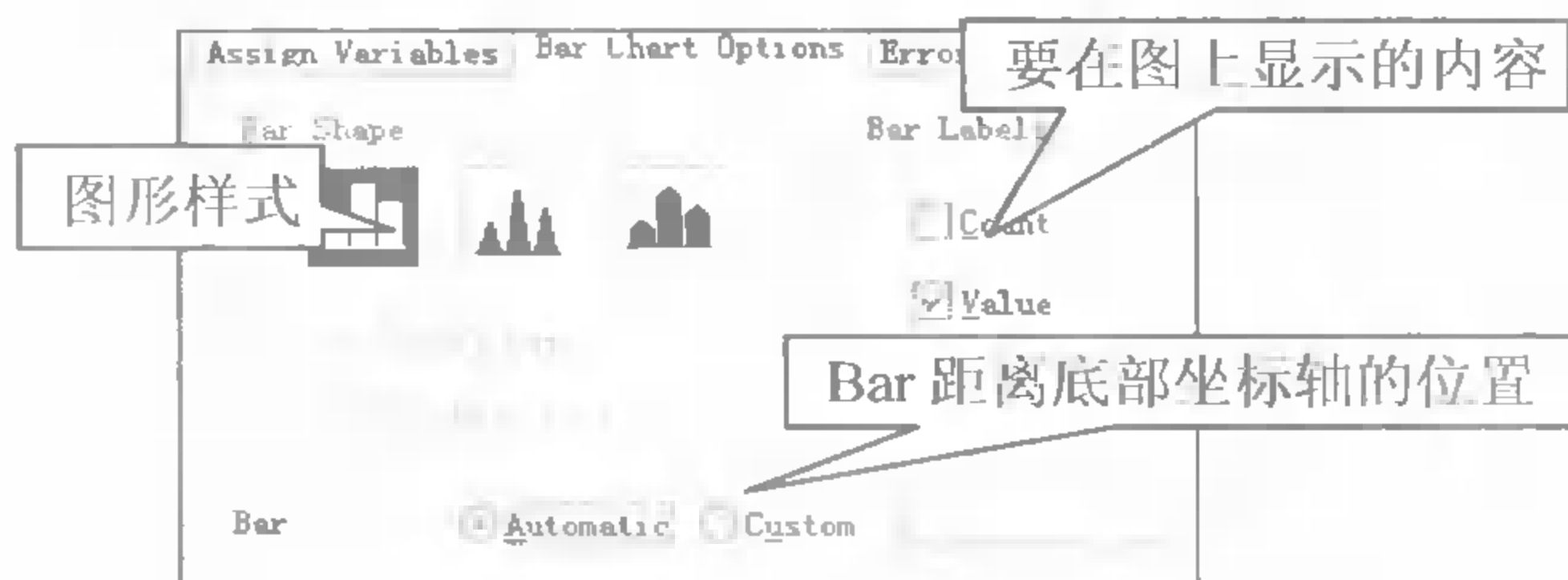


图 19-21 交互式条形图的 Bar 设置

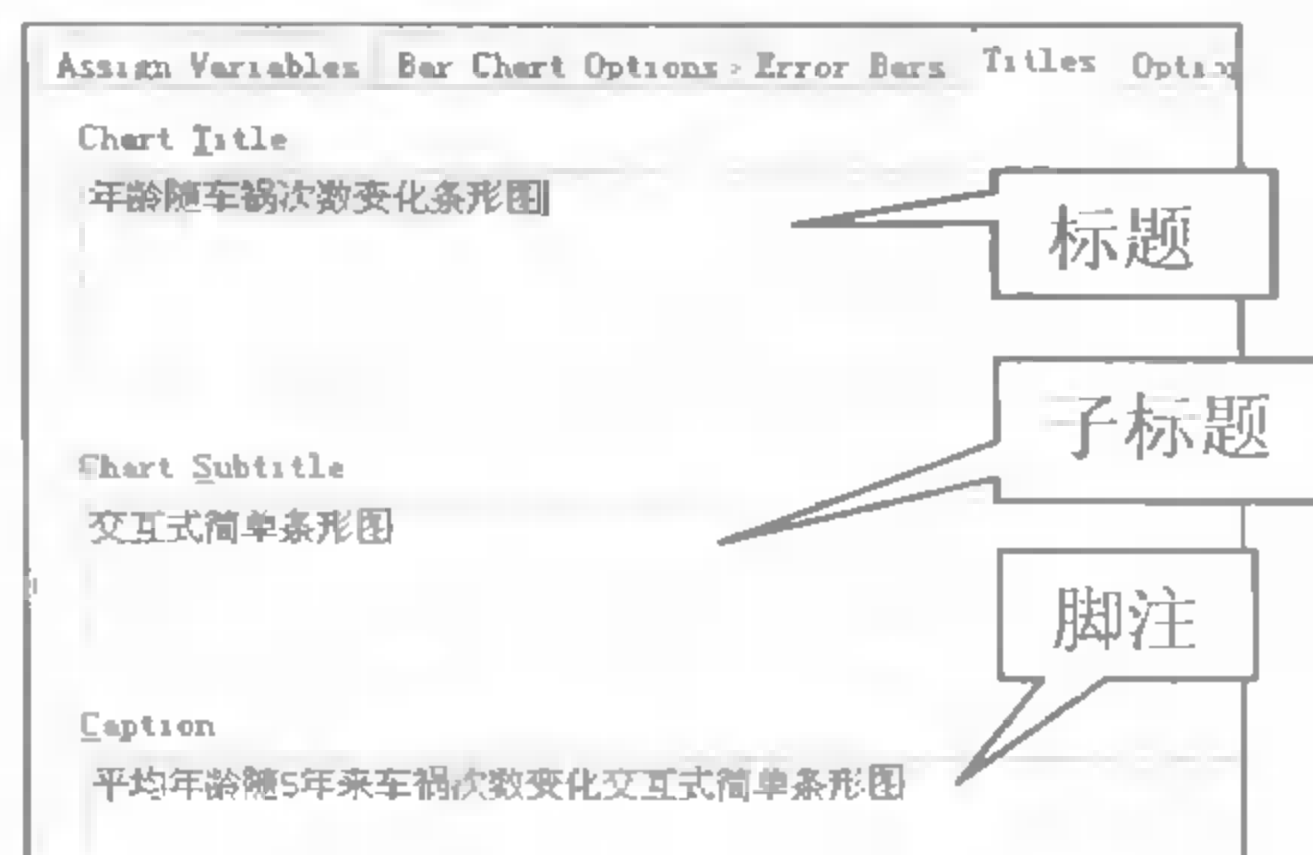


图 19-22 交互式条形图的标题设置

(2) 输出图形。在图 19-20 中，单击确定按钮运行，SPSS Viewer 窗口的输出图形如图 19-23 所示。

## 2. 交互式分类条形图

在图 19-20 中把变量列表中的年龄段、5 年来车祸次数、性别，分别拖动至 X 轴变量、Y 轴变量和 Color 选框里，将其分别作为条形图的 X 坐标轴、Y 坐标轴和子分类变量。指定了作图变量的操作界面，如图 19-24 所示。

单击 Titles 标签，打开如图 19-22 所示的子设置界面，只在 Chart Title 编辑框输入“车祸次数对年龄和性别的分类条形图”作为标题。

在图 19-24 中，单击确定按钮运行，SPSS Viewer 窗口的输出图形如图 19-25 所示。

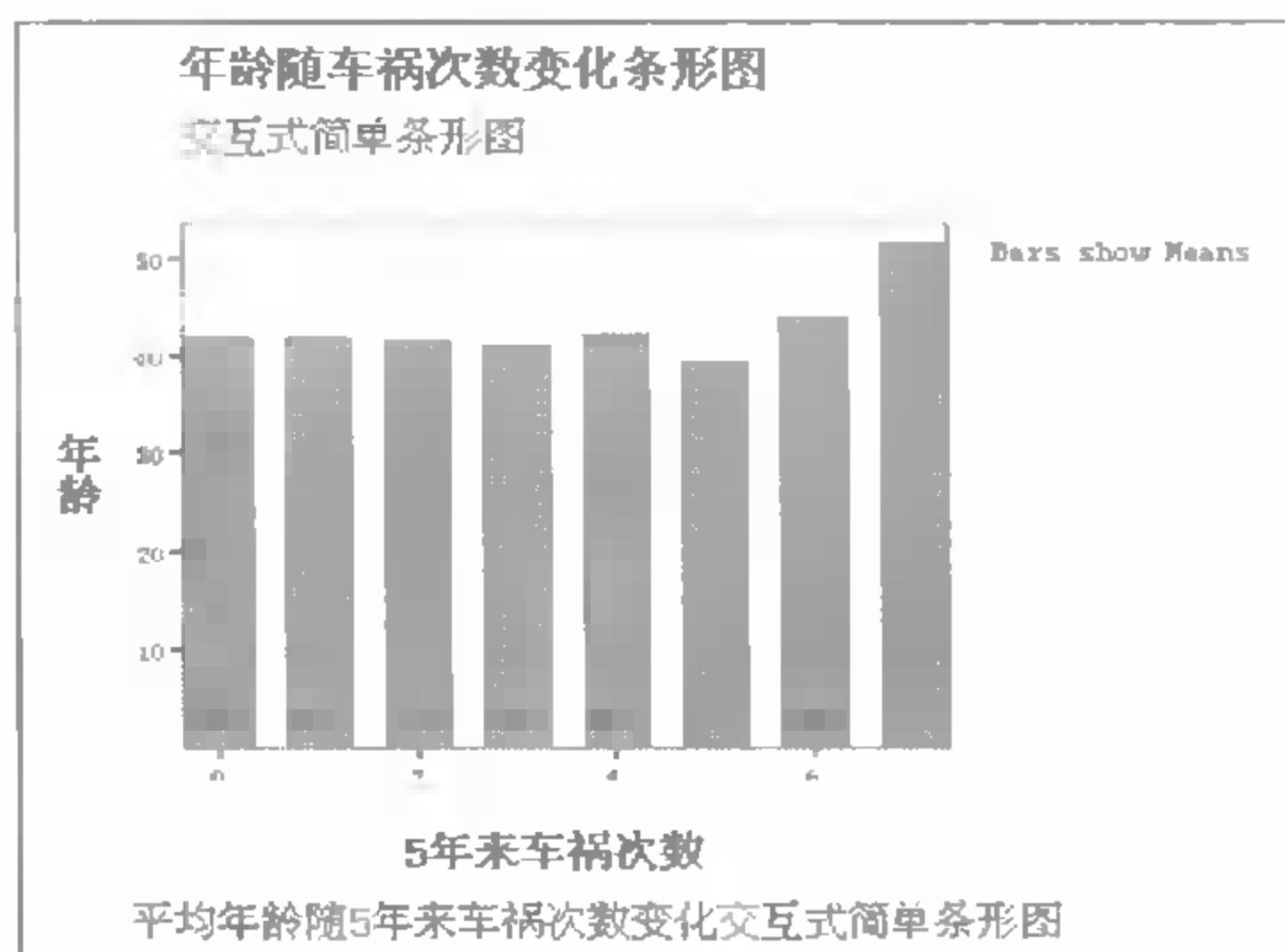


图 19-23 平均年龄对车祸次数的交互条形图



图 19-24 交互式分类条形图的设置

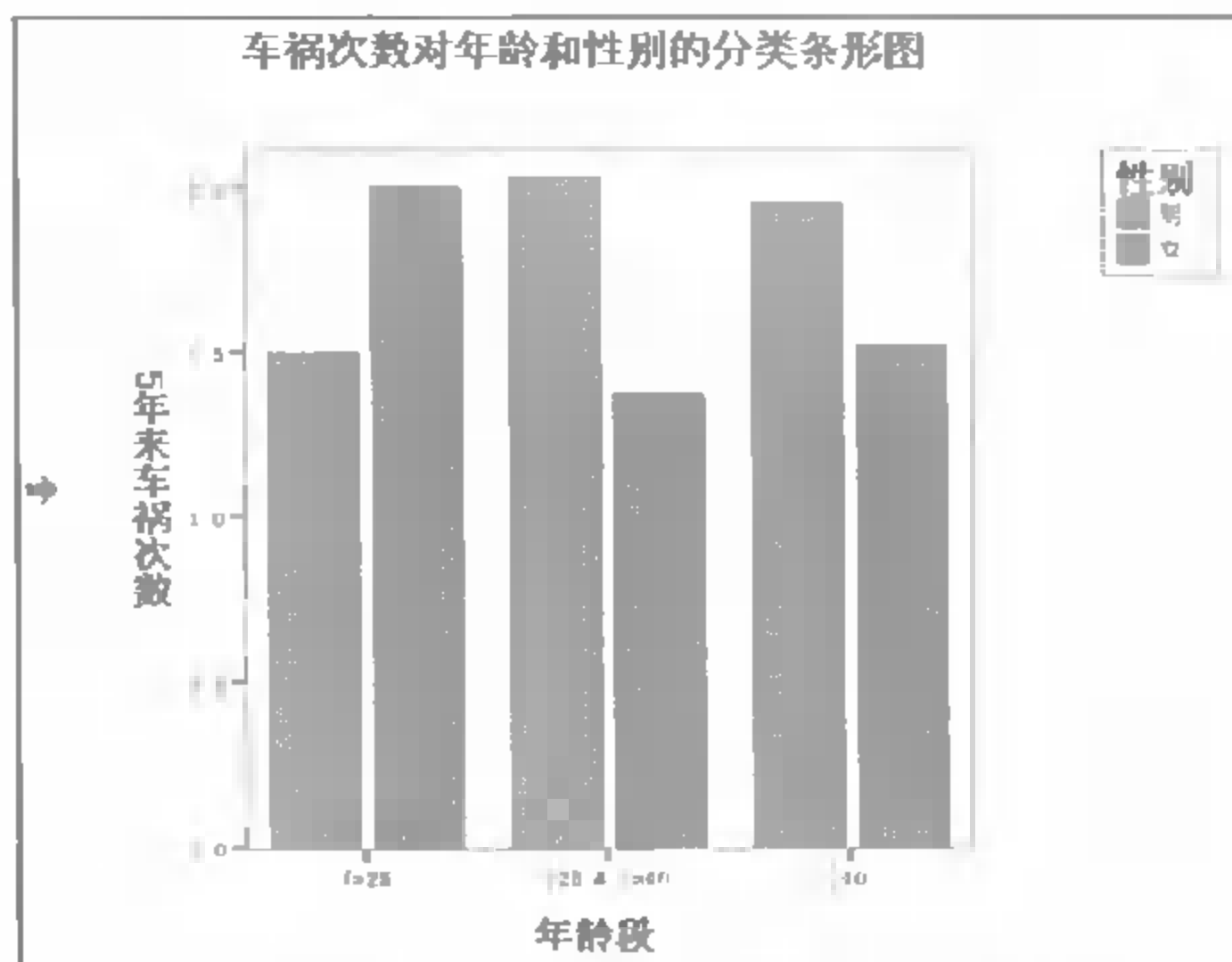


图 19-25 交互式分类条形图的输出

### 3. 交互式堆积条形图

在图 19-20 中，把变量列表中的年龄段、5 年来车祸次数、性别，分别拖动至 X 轴变量、Y 轴变量和 Color 选框里，将其分别作为条形图的 X 坐标轴、Y 坐标轴和子分类变量；单击 Color 选框右侧的 Cluster 下拉列表选中 Stack 选项，如图 19-26 所示。

单击 Titles 标签，打开如图 19-22 所示的子设置界面，只在 Chart Title 编辑框输入“车祸次数对年龄和性别的堆积条形图”作为标题。

在图 19-26 中，单击确定（未显示）运行，SPSS Viewer 窗口的输出图形如图 19-27 所示。

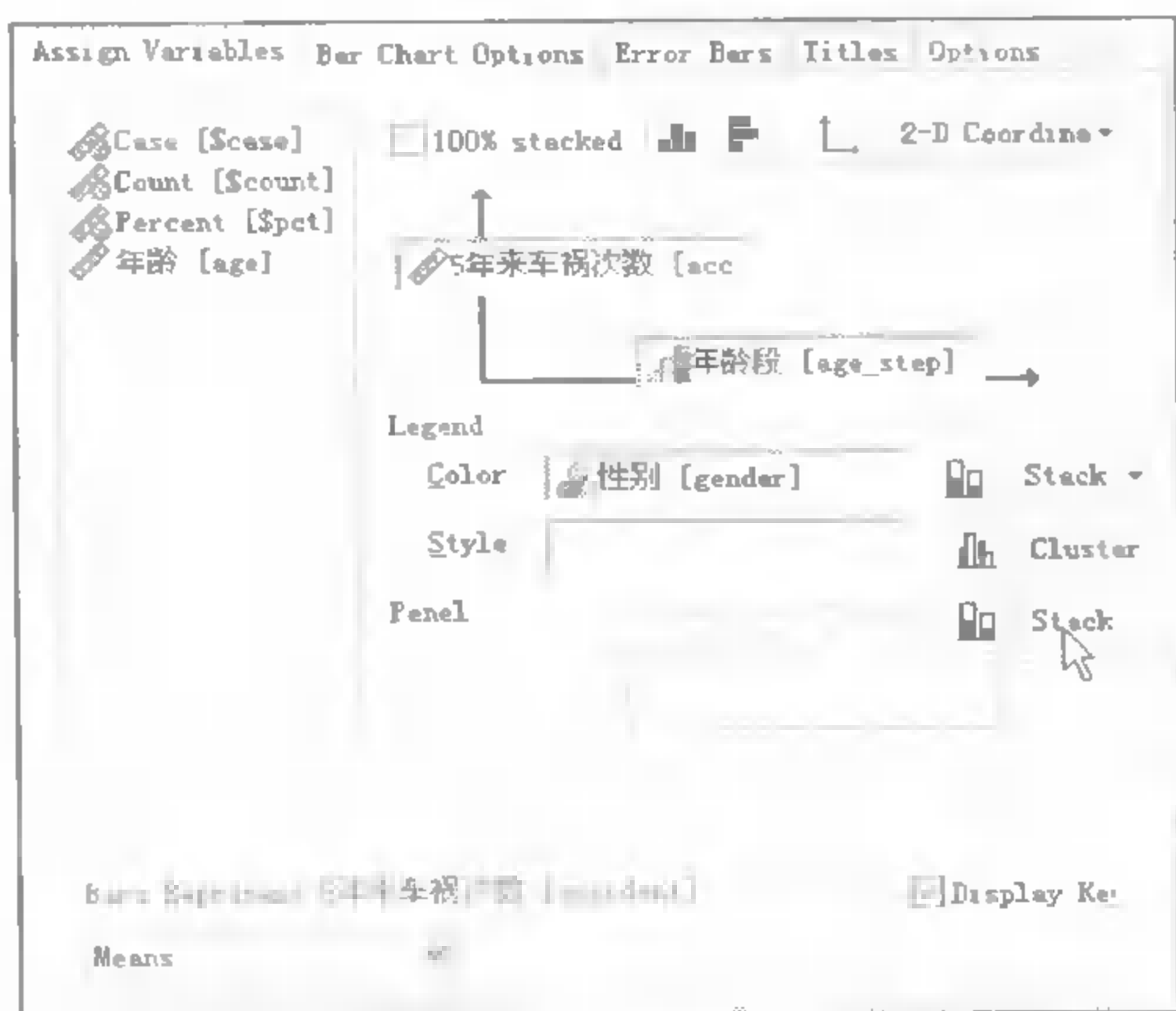


图 19-26 交互式堆积条形图的设置

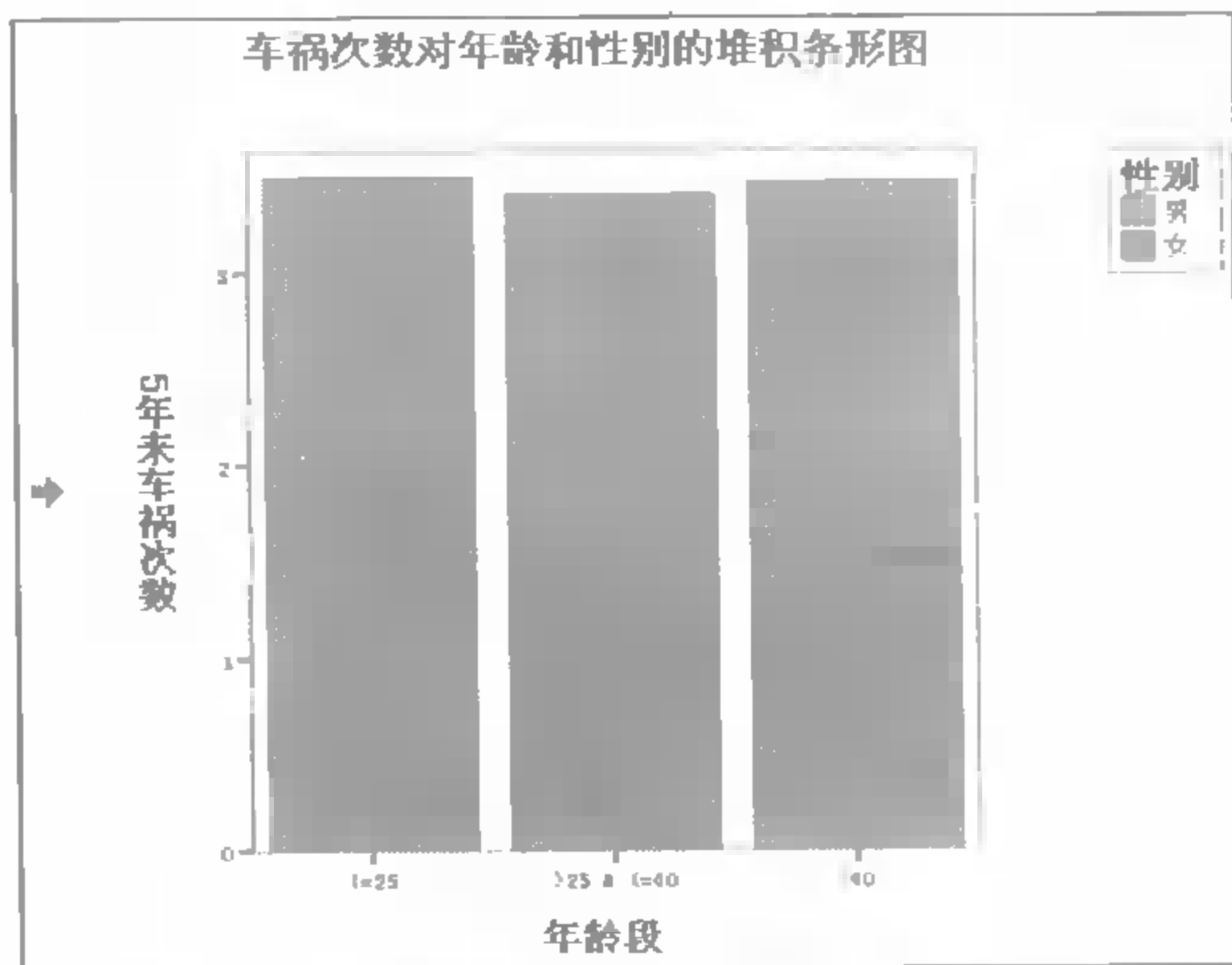


图 19-27 交互式堆积条形图的输出

### 19.2.4 用对话框创建条形图

依次单击菜单“Graphs→Legacy Dialogs→Bar...”，打开利用对话框创建条形图的类型选择界面，如图 19-28 所示。

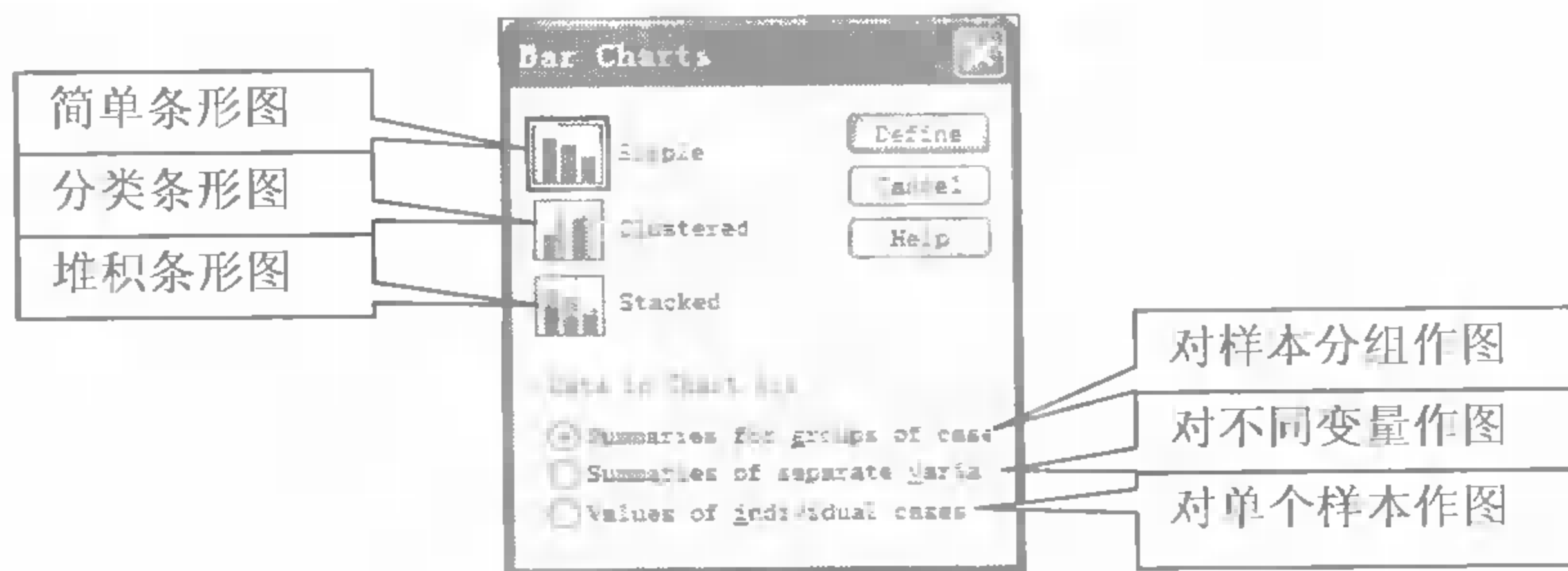


图 19-28 建立条形图选择对话框

Data in Chart Are 子设置栏，用于指定作图的数据对象，3 个可选项的含义如图中标识，关于它们的具体示例如图 19-29 所示。

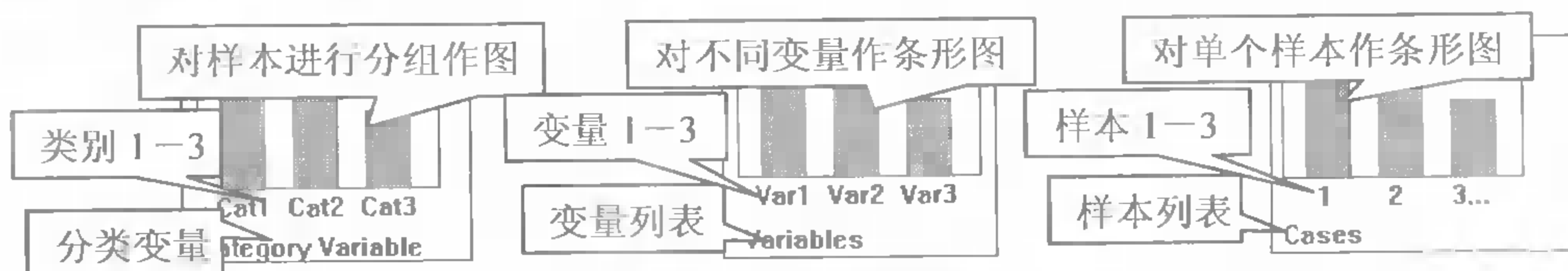


图 19-29 不同作图对象的图形示例

## 1. 简单条形图

在图 19-28 中,单击选中 Simple 图标,单击选中 Summaries for groups 单选框;单击 Define 按钮进入作简单条形图的设置界面,如图 19-30 所示。

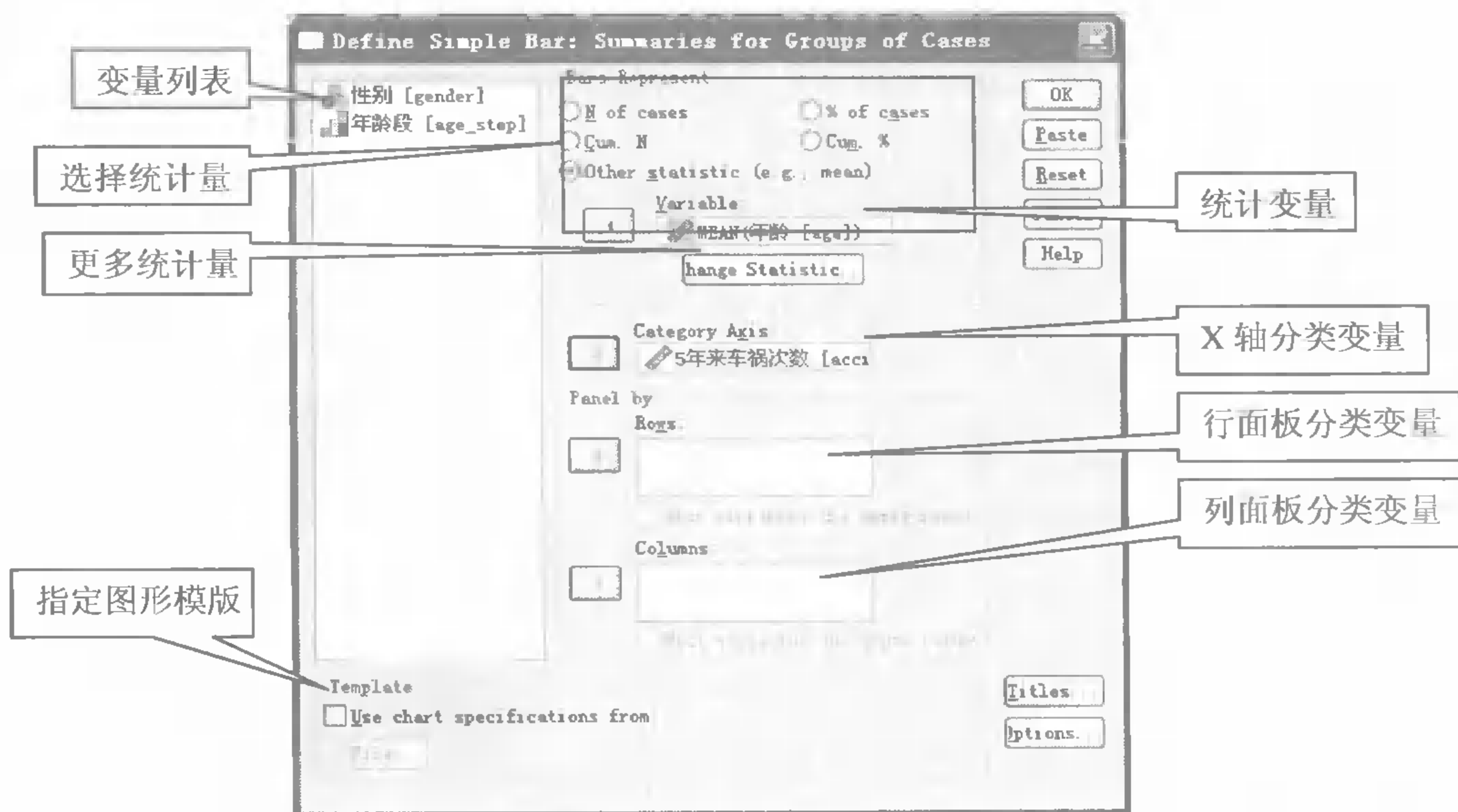




图 19-30 对话框作简单条形图的设置面板

单击选中 Other 单选框;在变量列表单击选中年龄变量,单击从上至下第一个  按钮,将其作为统计变量选入 Variable 选框;在变量列表单击选中 5 年来车祸次数变量,单击从上至下第二个  按钮,将其作为 X 轴分类变量选入 Category Axis 选框。

在图 19-30 中,单击 OK 按钮运行,输出图形与图 19-15 基本相同。

## 2. 分类条形图

在图 19-28 中,单击选中 Clustered 图标;单击 Define 按钮进入作分类条形图的设置界面,它与图 19-30 基本相同,只是多了一个 Define Clusters by 选框,用于指定一个子分类变量。

输出图形与用图形构建器所作的图形一样。

## 3. 堆积条形图

在图 19-28 中,单击选中 Stacked 图标;单击 Define 按钮进入作堆积条形图的设置界面,它与图 19-30 基本相同,只是多了一个 Define Stacks by 选框,用于指定一个子分类变量。

输出图形与用图形构建器所作的图形一样。

## 19.3 线形图

线形图利用线条(直线、折线或曲线)的延伸和波动,反映连续性变量的变化趋势。描述非连续性的资料一般不使用线形图,而使用条形图或直方图。



### 19.3.1 数据和问题描述

用线形图描绘股票数据的变化趋势最合适不过了，本节就对招商银行在 2007 年和 2008 年的部分日股价数据进行作图，所用数据文件为“Stockdata.sav”，数据格式如图 19-31 所示。

	Name	Type	Width	Decr	Label	Values	Missing	Columns	Align	Measure
2	mm	Numeric	8	0	月份	None	None	6	Right	Scale
3	dd	Numeric	8	0	日	None	None	8	Right	Scale
4	date1	Date			日期	None	None	10	Right	Scale
5	startm	Numeric	8	2	开盘价	None	None	8	Right	Scale
6	highm	Numeric	8	2	最高价	None	None	8	Right	Scale
7	lowm	Numeric	8	2	最低价	None	None	8	Right	Scale
8	endm	Numeric	8	2	收盘价	None	None	8	Right	Scale
9	allquantity	Numeric	8	2	交易量	None	None	8	Right	Scale
10	allmoney	Numeric	8	2	交易金额	None	None	8	Right	Scale

图 19-31 股票数据的变量含义

### 19.3.2 用图形构建器作线形图

依次单击菜单“Graphs→Chart Builder”打开图形构建器，如图 19-32 所示。单击 Gallery 标签，在 Choose from 列表单击选中 Line，就在其右侧显示预设的线形图图标。

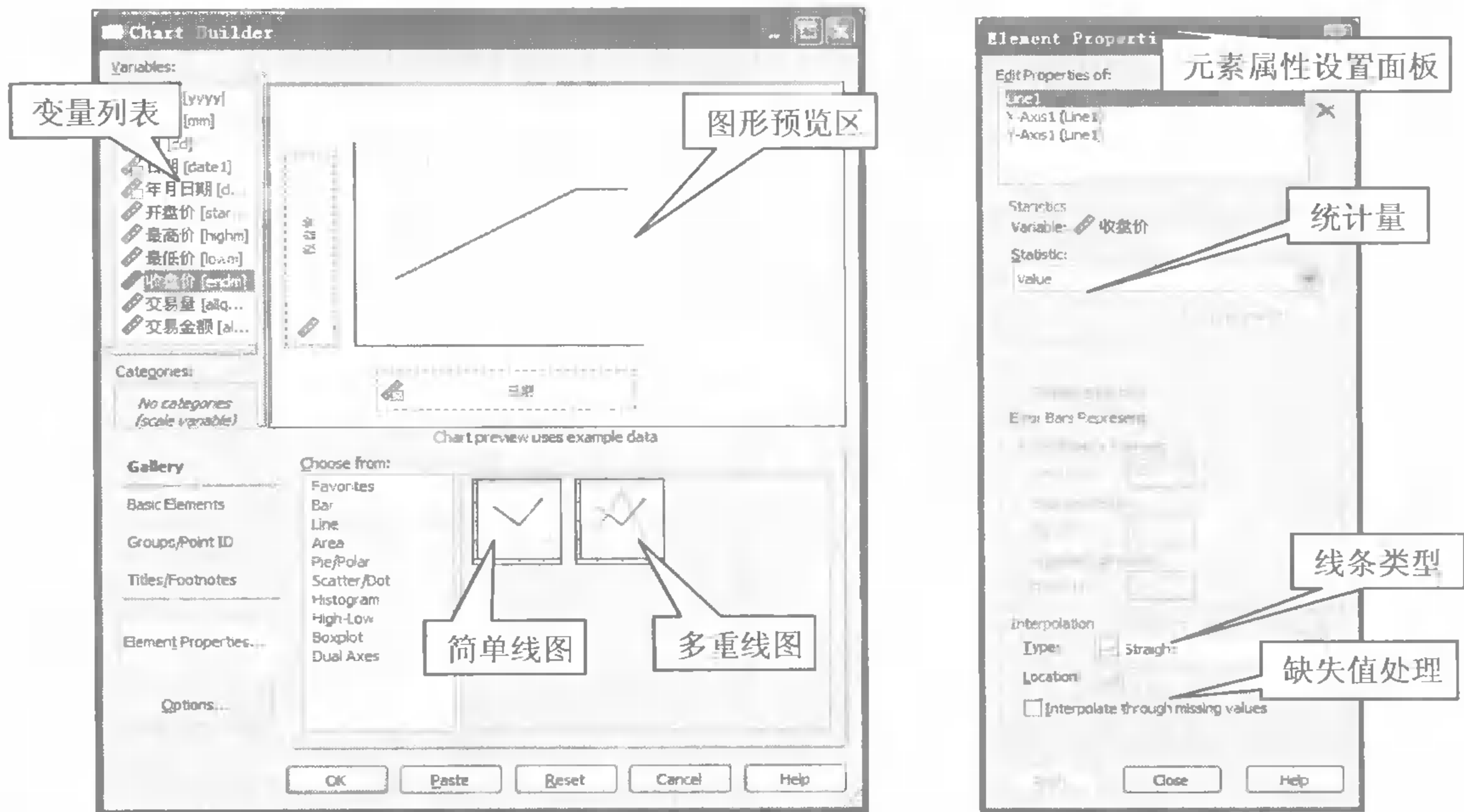



图 19-32 创建简单线形图的设置界面

#### 1. 简单线形图

下面先来作收盘价格随日期而变化的简单线形图。

(1) 参数设置。在图 19-32 中，双击预置图标  (Simple Line)，就会在图形预览区给出简单线形图的预览，同时自动弹出元素属性设置面板；把预置图标拖动至图形预览区，可以达到相同的效果。

从变量列表中把日期、收盘价两个变量，分别拖动至预览区的 X-Axis、Y-Axis 两个虚线

框中，将其分别作为简单线形图的 X 坐标轴和 Y 坐标轴。

(2) 输出图形。在图 19-32 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-33 所示。可见，收盘价表现出了逐步走低的信息。

## 2. 多重线形图

多重线形图在一个图里显示多条趋势线，它需要指定一个分线变量，对其每个取值分别在图里作一条曲线，以便观察和比较不同类别的样本的变化趋势。下面对 2007、2008 两年里的收盘价月均值作多重线形图，观察这两年的收盘价的变化特点。

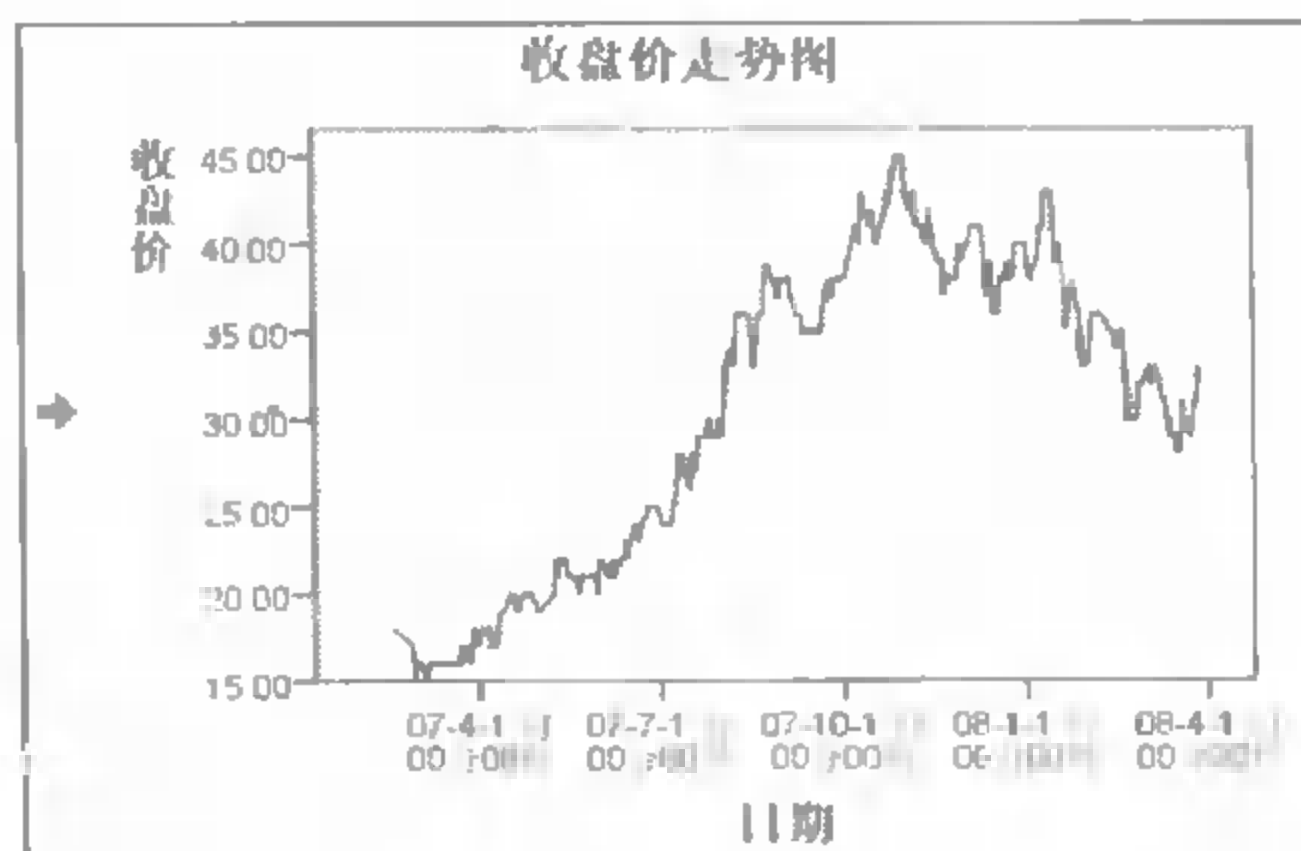



图 19-33 收盘价走势图

(1) 参数设置。在图 19-32 中，双击预置图标  (Multiple Line)，就会在图形预览区给出多重线形图的预览，同时自动弹出元素属性设置面板；把预置图标拖动至图形预览区，可以达到相同的效果。

在变量列表中右击年份变量，在弹出的快捷菜单里选中 Ordinal 选项，指定它为有序分类变量。从变量列表中把月份、收盘价、年份三个变量，分别拖动至预览区的 X-Axis、Y-Axis 和 Set color 三个虚线框中，将其分别作为多重线形图的 X 坐标轴、Y 坐标轴和分线变量。指定了作图变量的图形预览区如图 19-34 所示。

在图 19-32 中的 Element Properties 对话框中，选中 Edit 列表框里的 Line1，单击 Statistic 下拉列表选中 Mean 选项，表示线形图将代表收盘价变量的均值大小；单击 Apply 按钮应用设置。

(2) 输出图形。在图 19-34 中，单击 OK 按钮（未显示）运行，SPSS Viewer 窗口的输出图形如图 19-35 所示。可见，07、08 两年里收盘价的走势呈现出相反的趋势。

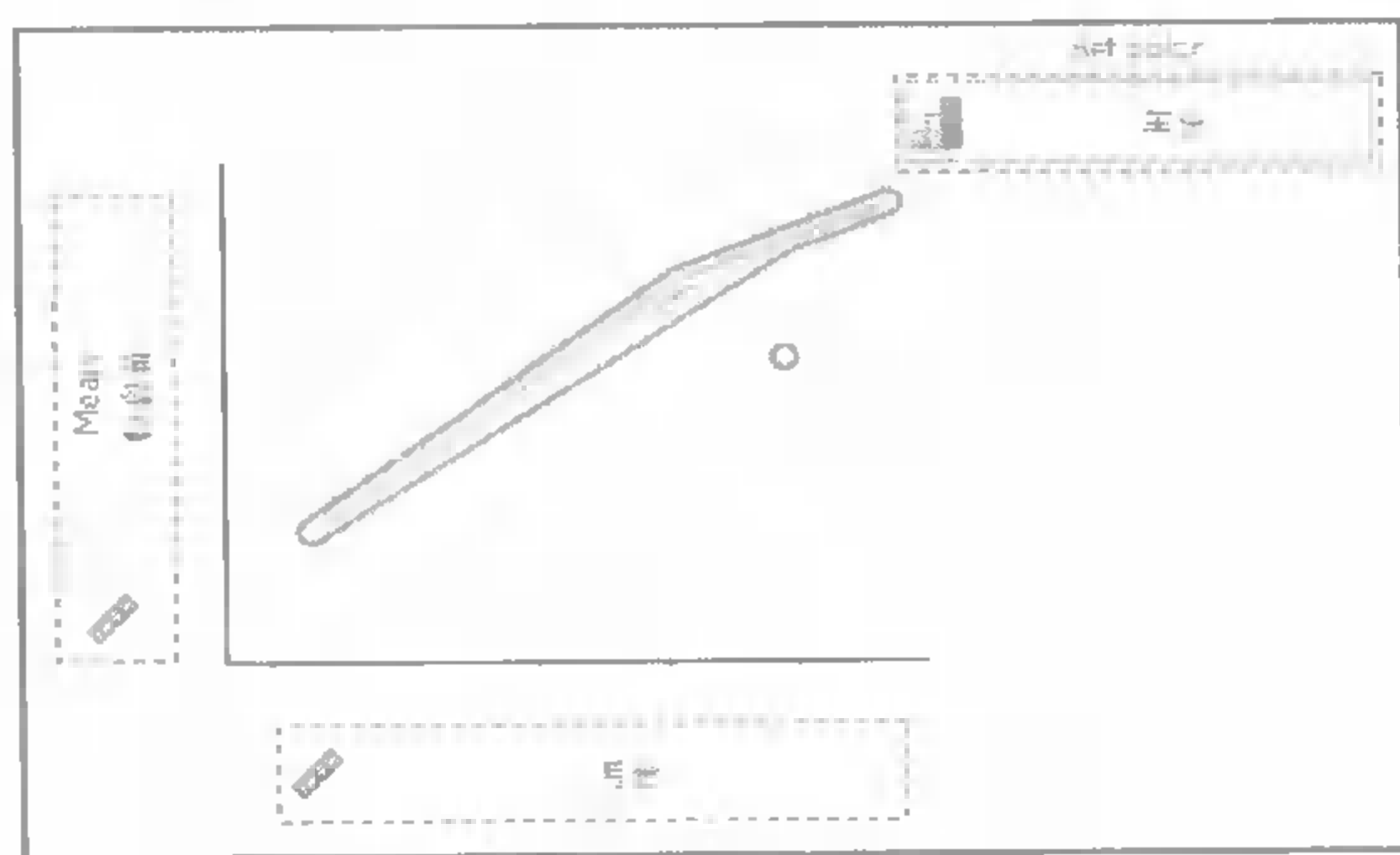


图 19-34 多重线形图的设置预览

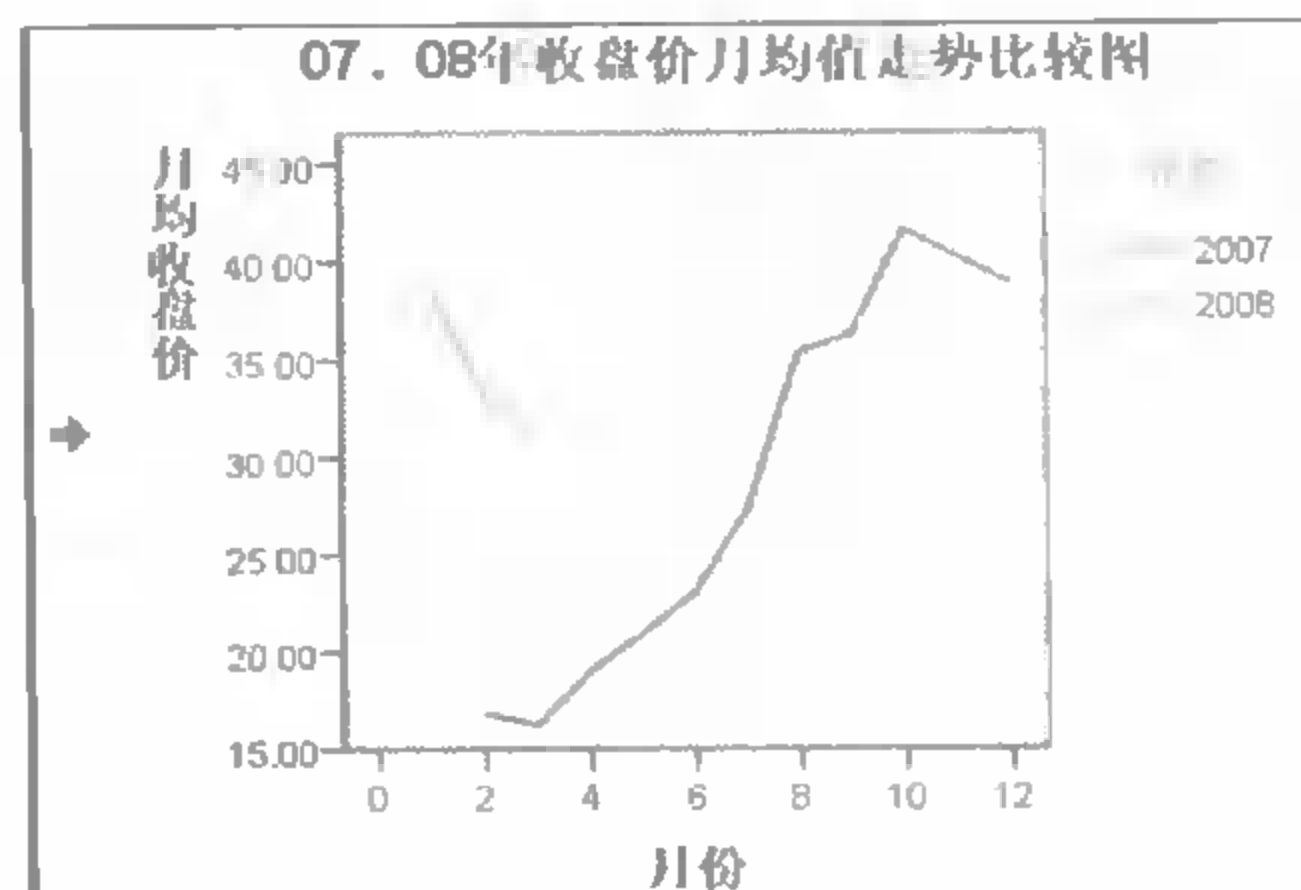


图 19-35 多重线形图的输出

### 19.3.3 交互式线形图

依次单击菜单“Graphs→Interactive→Line...”，打开建立交互式线形图的操作界面，如图 19-36 所示。

#### 1. 交互式简单线形图

在此，要作一个与第 19.3.2 节中的简单线形图相似的交互式图形。

(1) 参数设置。

① 指定作图变量。在图 19-36 中，从变量列表中把日期、收盘价两个变量，分别拖动至

X 轴变量、Y 轴变量所示的选框，将其分别作为线形图的 X、Y 坐标轴。

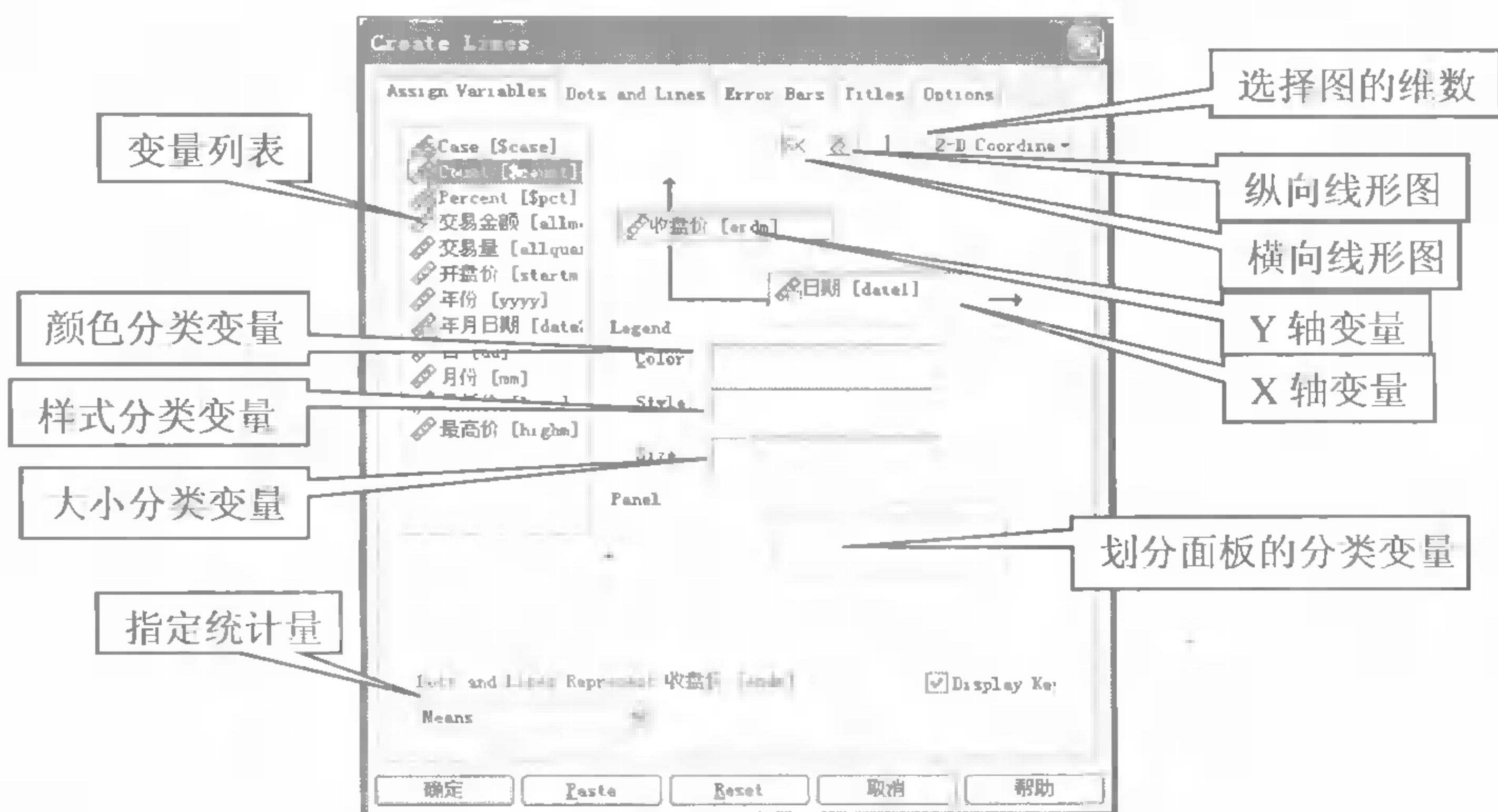


图 19-36 交互式条形图变量设置面板

② 线形图的样式设置。在图 19-36 中，单击 Dots and Lines 标签，打开如图 19-37 所示的子设置界面，在此设置关于线形图显示样式的选项。保留默认设置，单击 Assign 标签返回图 19-36 所示的设置界面。

- Display 栏，选择图形的显示内容：Dots（点）、Drop Lines（连接线）。
- Point Labels 栏，选择点的标签内容：Value（取值）、Percent（比例）、Count（个数）。
- Line Labels 栏，选择线标签内容：Categories（类别）、Percent（比例）、Count（个数）。
- Interpolation 栏，选择在图中所使用的线型，默认为 Straight。
- Break lines at missing values 复选框，选中表示图中的曲线不通过含缺失值的分类。

(2) 输出图形。在图 19-36 中，单击确定按钮运行，SPSS Viewer 窗口的输出图形如图 19-38 所示。可见，输出图形与图 19-33 基本相同。

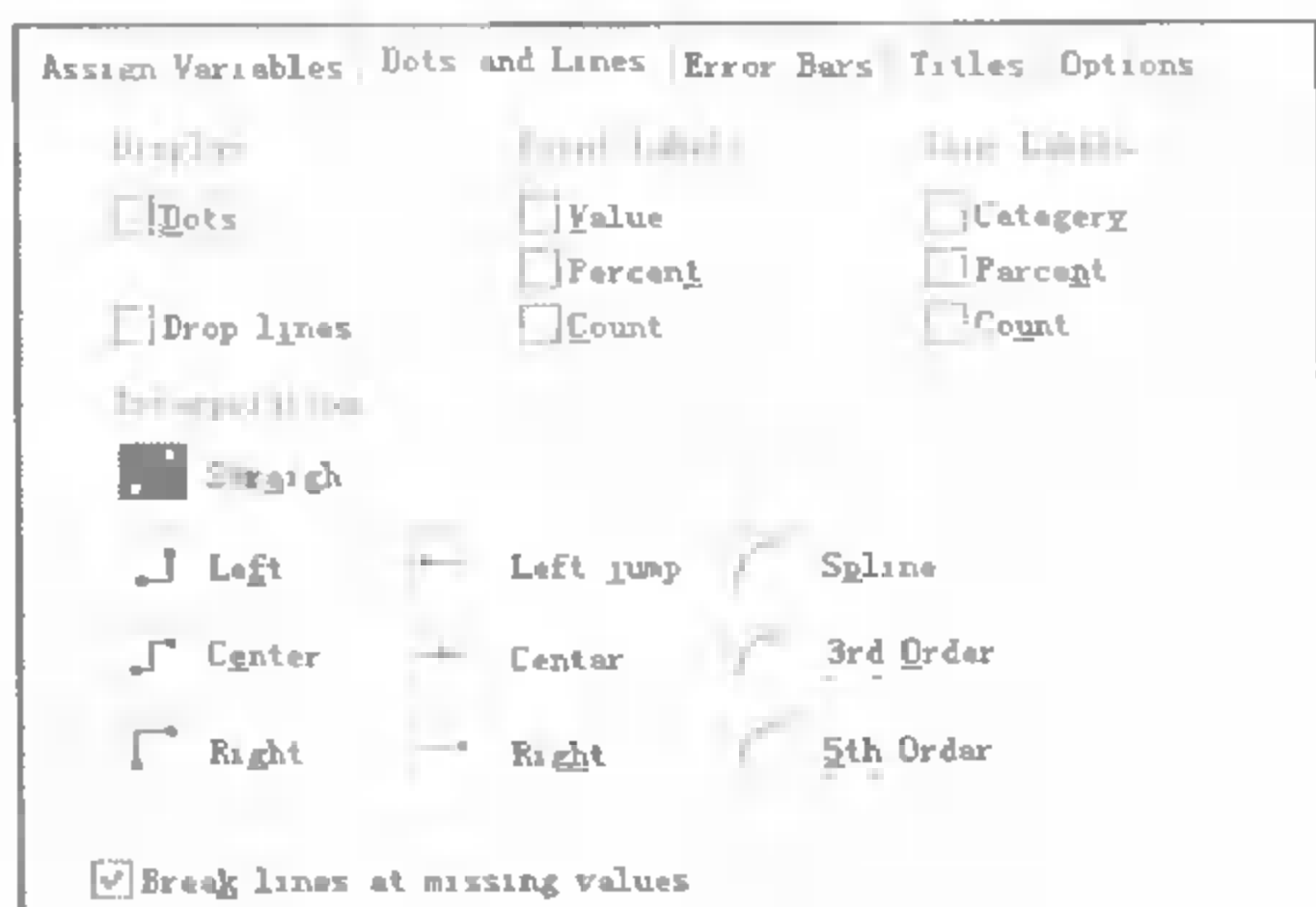


图 19-37 交互式线形图的样式设置

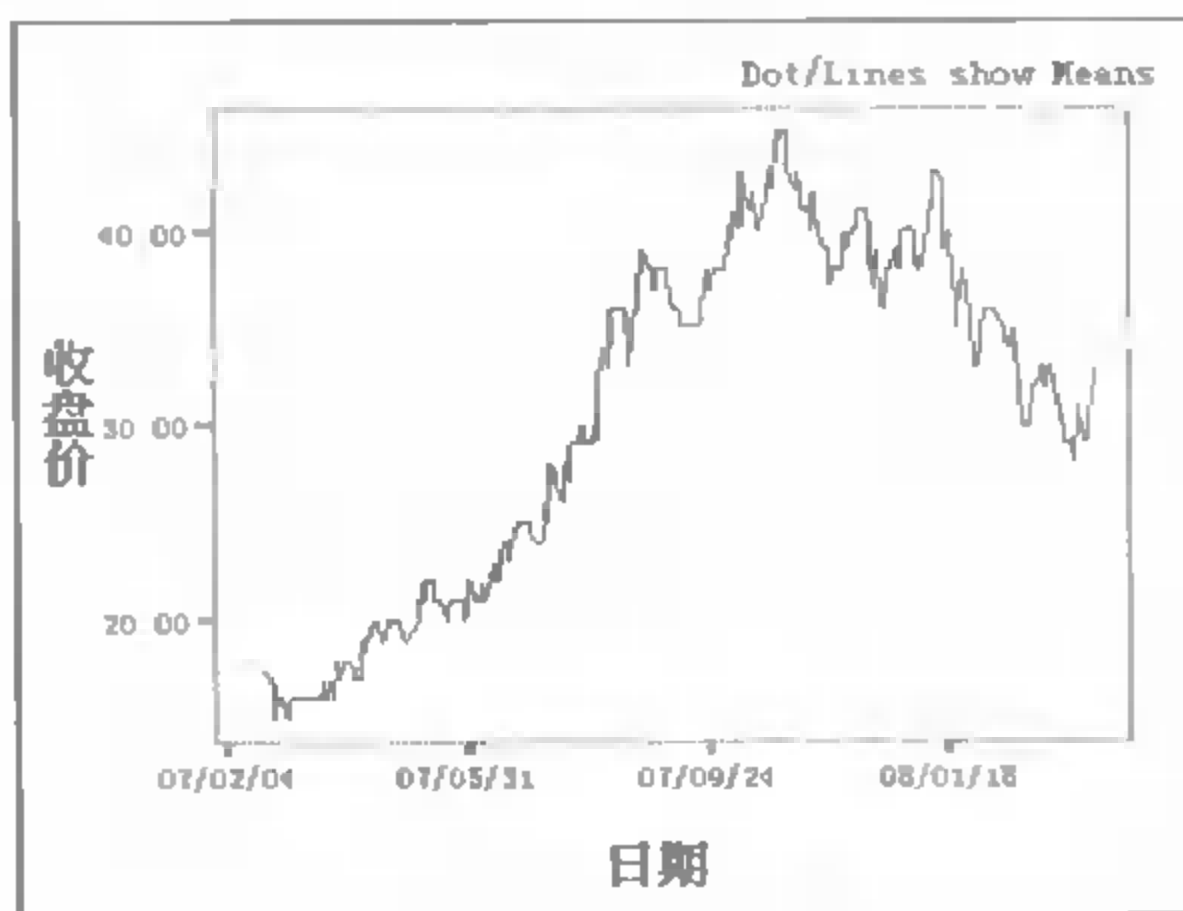


图 19-38 交互式线形图的输出

## 2. 交互式多重线形图

如图 19-36 所示，在变量列表中右击年份变量，在弹出的快捷菜单里选中 Categorical 选项，指定它为有序分类变量。从变量列表中把月份、收盘价、年份，分别拖动至 X 轴变量、Y 轴变量和 Color 选框里，将其分别作为线形图的 X 坐标轴、Y 坐标轴和子分类变量。

单击确定按钮运行，SPSS Viewer 窗口的输出图形与图 19-35 基本相同。

### 19.3.4 用对话框创建线形图

依次单击菜单“Graphs→Legacy Dialogs→Line...”，打开利用对话框创建线形图的选择界面，如图 19-39 所示。Data in Chart Are 子设置栏，用于指定作图的数据对象，关于它们的具体示例如图 19-29 所示。

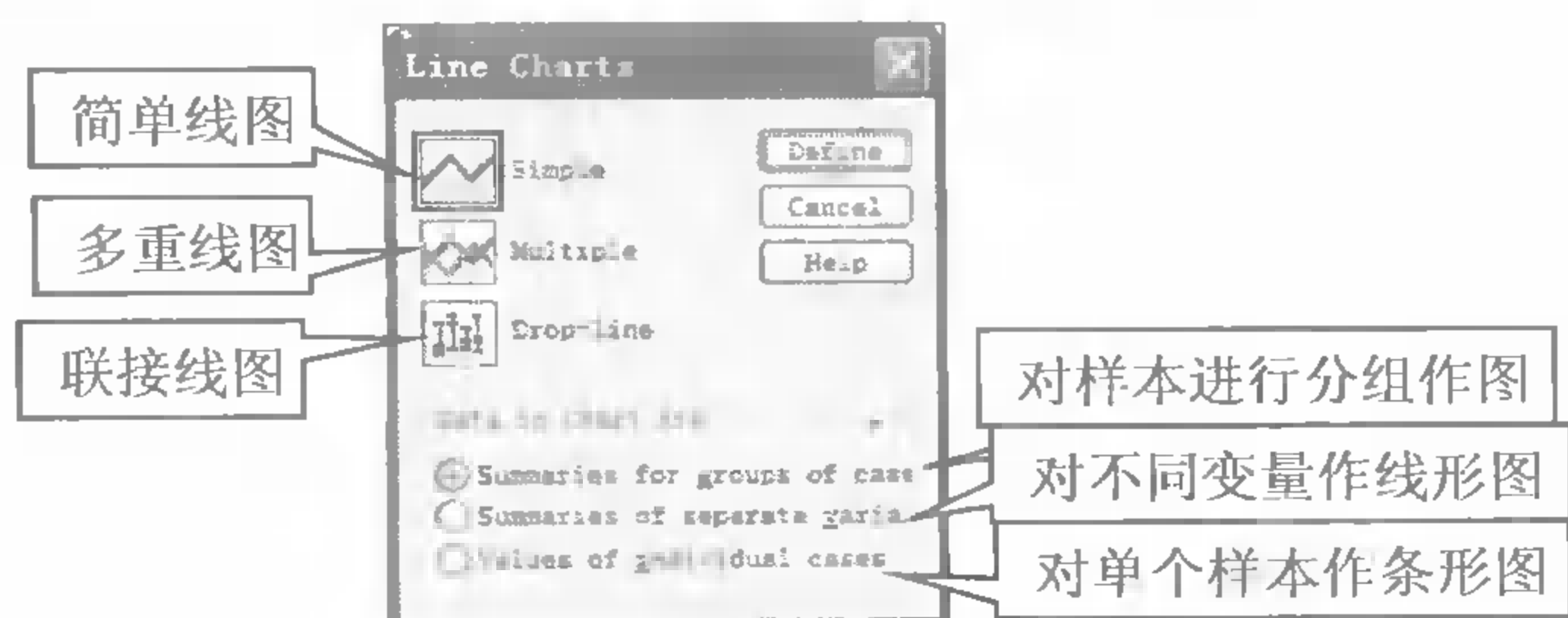


图 19-39 建立线图选择对话框

在此，对三种线形图的设置方法都和第 19.2.4 节条形图的操作类似，在条形图的示例中都是对样本进行分组作图，这次选择对变量进行分组作图，目的是在同一个图里观察开盘价、收盘价、最高价、最低价的月均值走势情况。

#### 1. 参数设置

在图 19-39 中，单击选中 Multiple 图标，单击选中 Summaries of separate variable 单选框；单击 Define 按钮进入作多重线形图的设置界面，如图 19-40 所示。

在变量列表选中从开盘价至收盘价的 4 个变量，单击从上至下第一个 按钮，将其作为作图变量选入 Lines Represent 列表框；在变量列表单击选中年月日期变量，单击从上至下第二个 按钮，将其作为 X 轴分类变量选入 Category Axis 选框。

#### 2. 输出图形

在图 19-40 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-41 所示。可见，在都取月均值的情况下，4 种价格走势完全趋同。

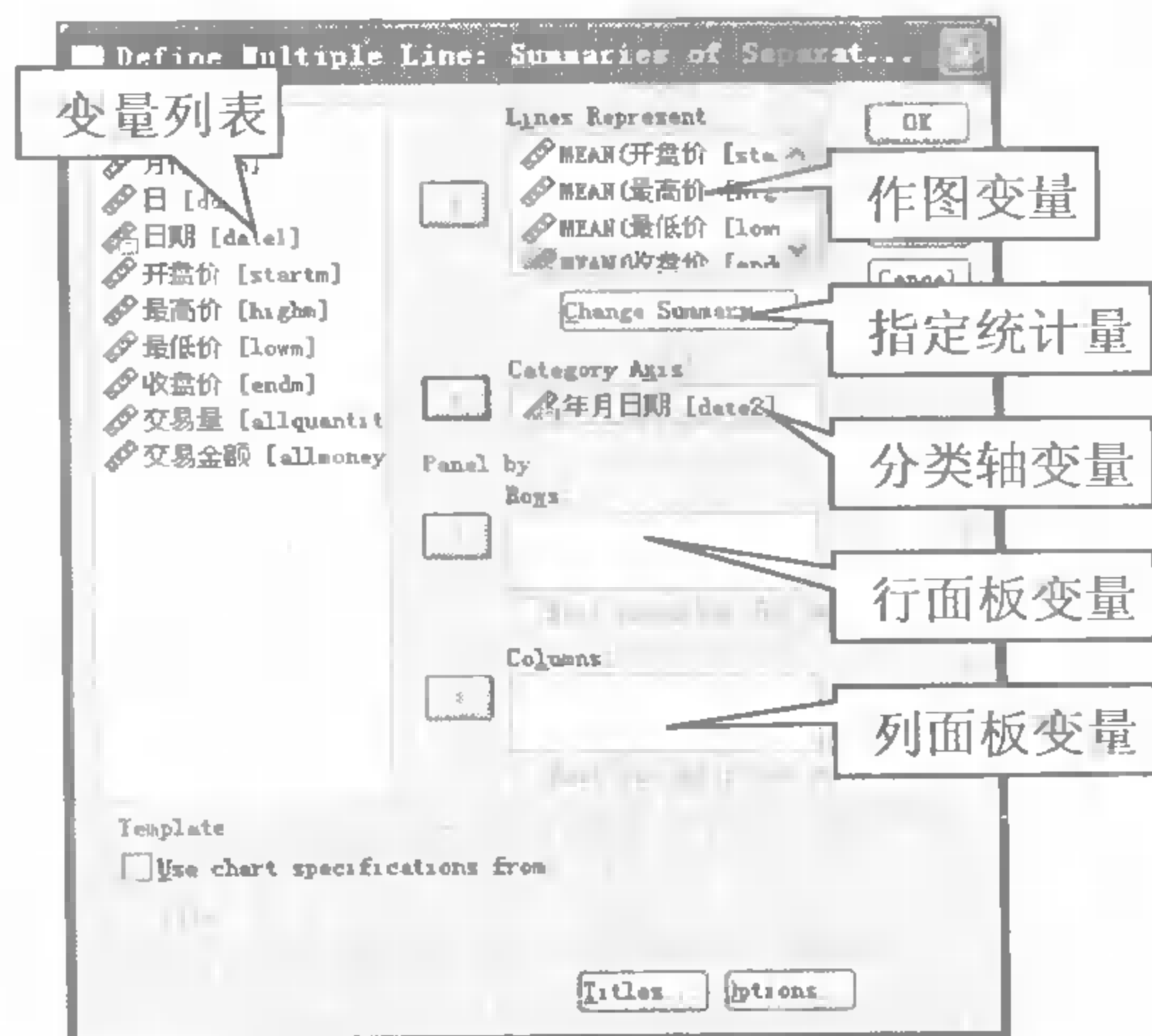


图 19-40 多变量分类线型图的设置

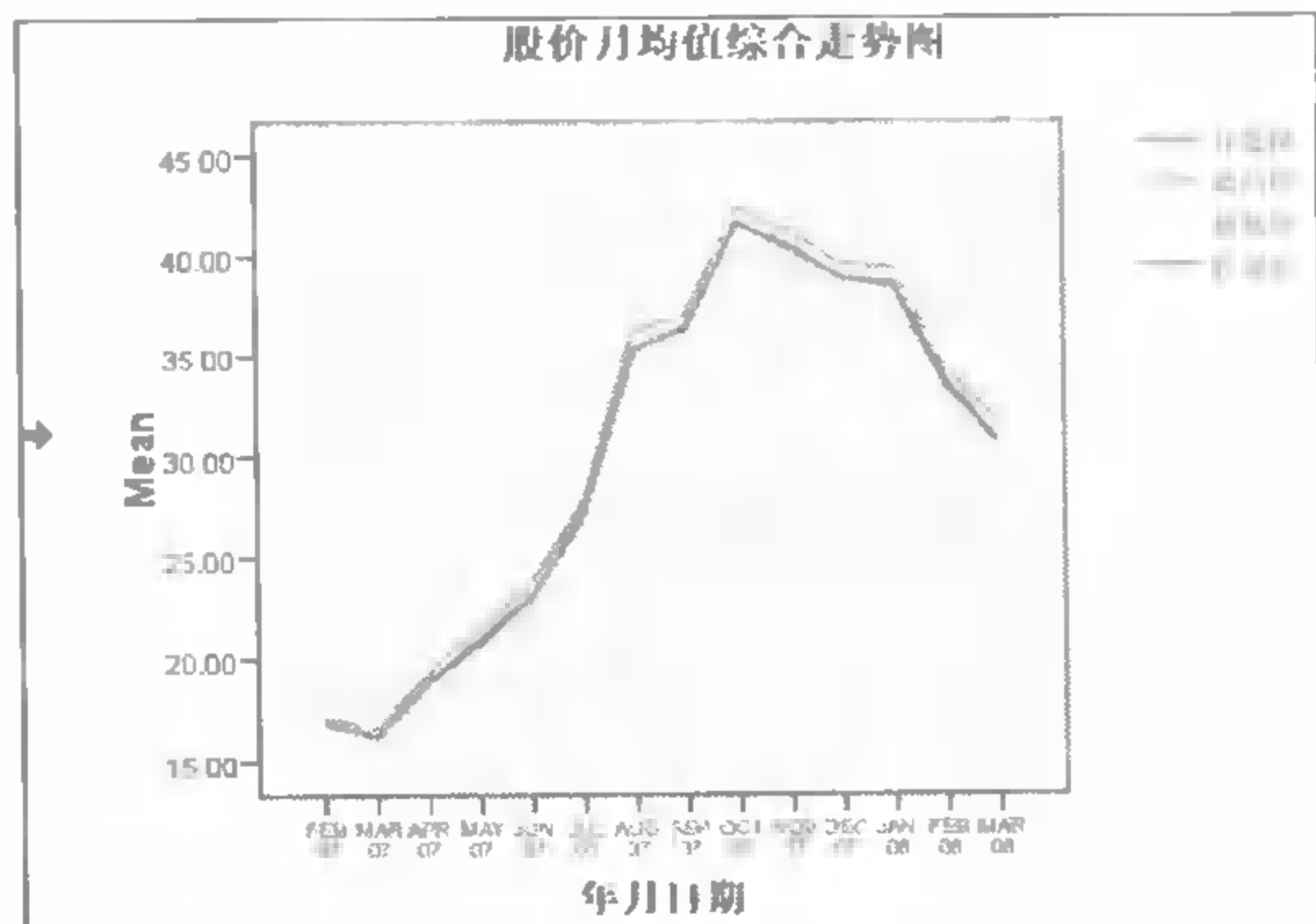


图 19-41 多变量分类线型图的输出



19.4 面积图

面积图通过面积的变化描绘连续型变量的分布形状或者变化趋势，直观地看，它相当于在线形图中用某种颜色填充线条和 X 轴之间的面积区域。面积图经常用作某个汇总变量随时间变化而连续变化的图形。

19.4.1 数据和问题描述

本节使用面积图来研究在经济发展过程中政府支出的变化趋势。所用数据文件为“政府支出 0.sav”，数据格式如图 19-42 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	year	Numeric	4	0	年份	None	None	8	Right	Scale
2	per	Numeric	3	0	时间段	1 50-52	None	8	Right	Ordinal
3	eco	Numeric	6	2	经济	None	None	8	Right	Scale
4	soc	Numeric	6	2	社会文教	None	None	8	Right	Scale
5	def	Numeric	6	2	国防	None	None	8	Right	Scale
6	adm	Numeric	6	2	管理	None	None	8	Right	Scale
7	deb	Numeric	6	2	债务	None	None	8	Right	Scale
8	oth	Numeric	6	2	其他	None	None	8	Right	Scale
9	tot	Numeric	7	2	总支出	None	None	8	Right	Scale

图 19-42 横向的政府支出数据格式

19.4.2 用图形构建器作面积图

依次单击菜单“Graphs→Chart Builder”打开图形构建器，如图 19-43 所示。单击 Gallery 标签，在 Choose from 列表单击选中 Area，就在其右侧显示预设的面积图图标。

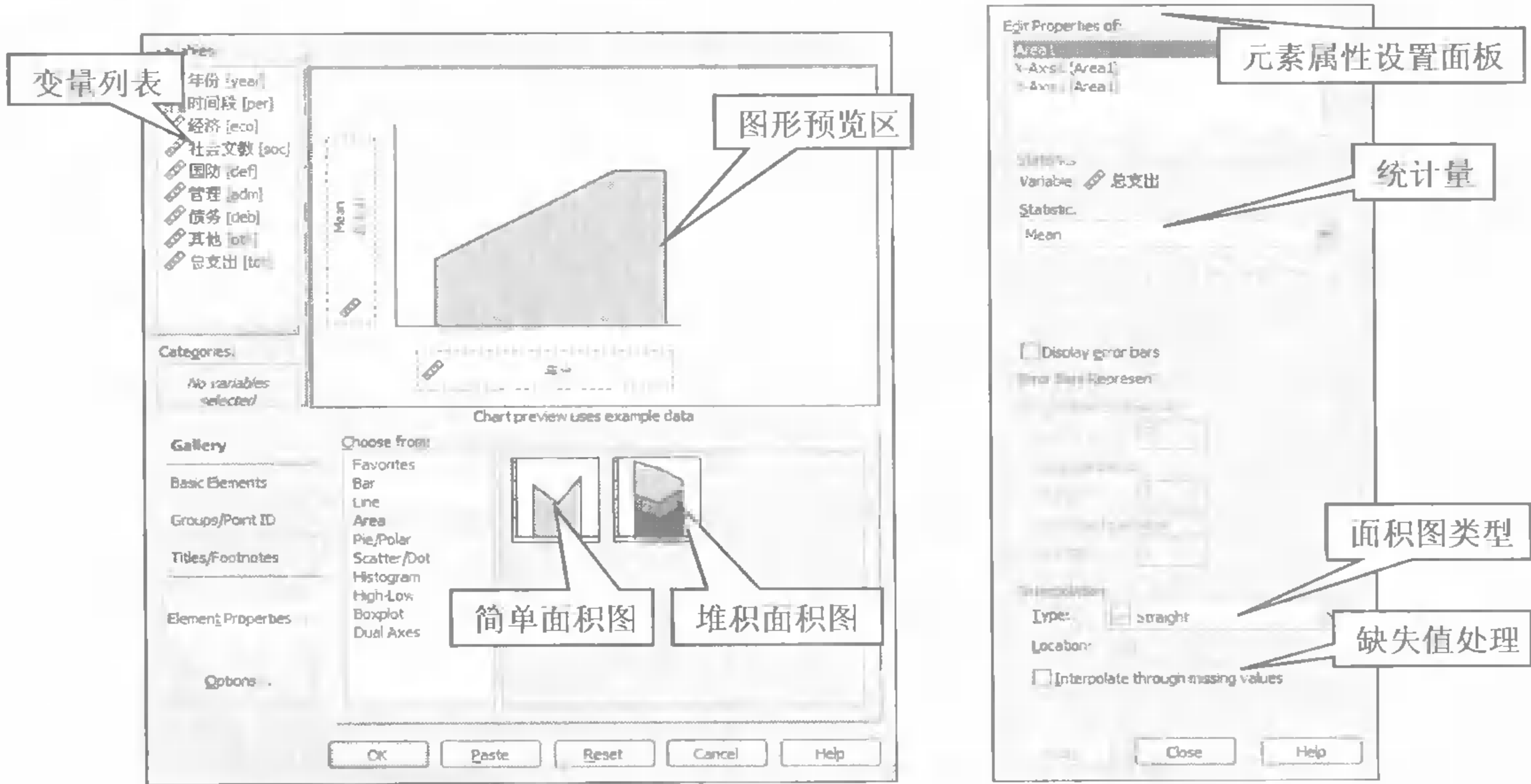



图 19-43 创建简单面积图的设置界面

1. 简单面积图

下面先来作政府总支出随年份变化而变化的面积走势图。

(1) 参数设置。在图 19-43 中, 双击预置图标 (Simple Area), 就会在图形预览区给出简单面积图的预览, 同时自动弹出元素属性设置面板; 把预置图标拖动至图形预览区, 可以达到相同的效果。

从变量列表中把年份、总支出两个变量, 分别拖动至预览区的 X-Axis、Y-Axis 两个虚线框中, 将其分别作为简单面积图的 X 坐标轴和 Y 坐标轴。

(2) 输出图形。在图 19-43 中, 单击 OK 按钮运行, SPSS Viewer 窗口的输出图形如图 19-44 所示。可见, 总支出呈现出波动上升的趋势。

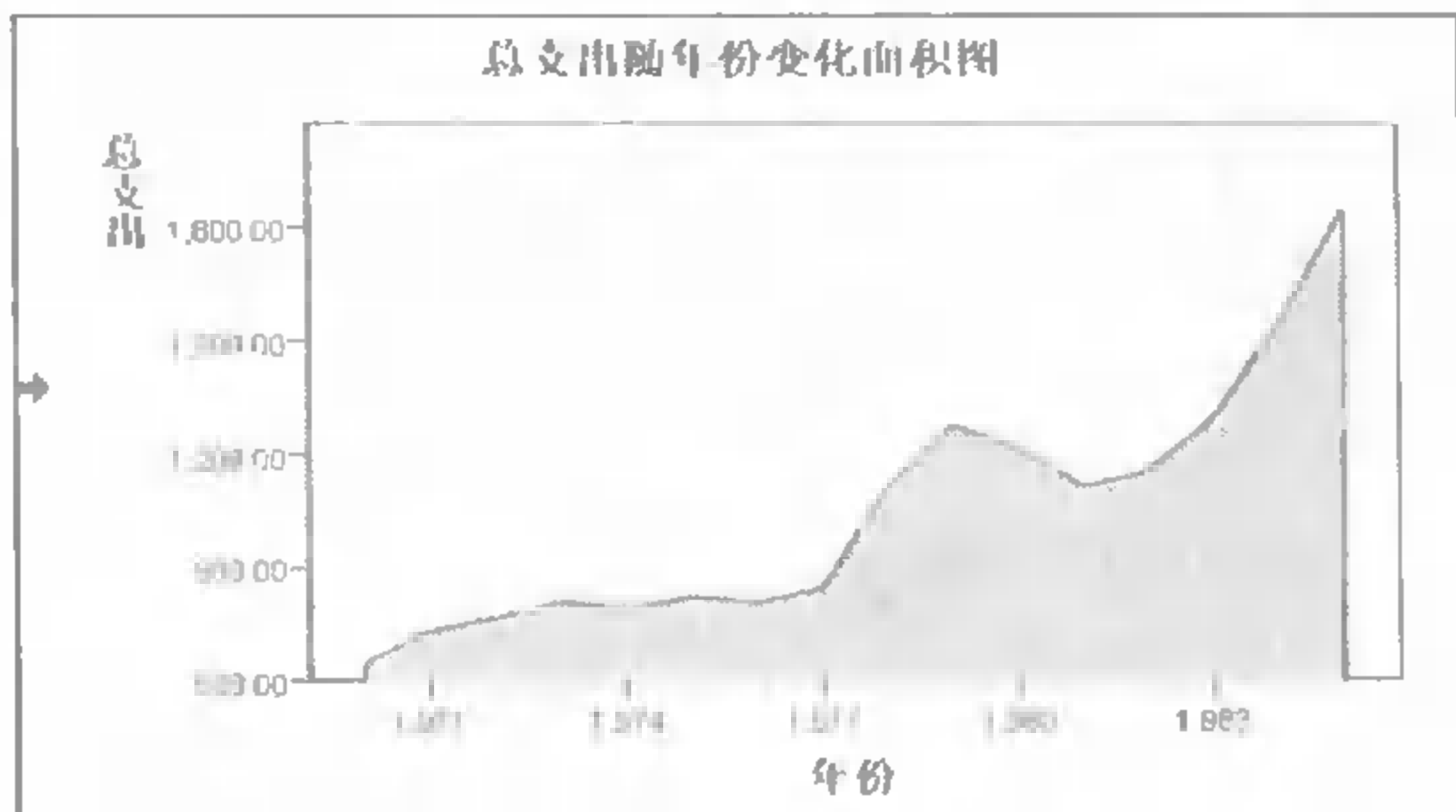


图 19-44 总支出随年份变化面积图


## 2. 堆积面积图

堆积面积图, 它对 X 轴的每个取值再按照某个指标进一步细分, 并作关于所得子类别的子图, X 轴上对应于同一点的多个子图逐次在 Y 轴方向堆积。

图 19-42 所示是横向格式的数据, 它不适用于作堆积面积图。随后使用如图 19-45 所示的纵向数据作图, 所用数据文件为“政府支出 1.asv”。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	year	Numeric	8	2	年份	None	None	8	Right	Scale
2	def	Numeric	8	2	支出数额	None	None	8	Right	Scale
3	type	String	8		支出类型	{def 国防}	None	7	Left	Nominal

图 19-45 纵向的政府支出数据格式

(1) 参数设置。在图 19-43 中, 双击预置图标 (Multiple Area), 就会在图形预览区给出堆积面积图的预览, 同时自动弹出元素属性设置面板; 把预置图标拖动至图形预览区, 可以达到相同的效果。

从变量列表中把年份、支出数额、支出类型三个变量, 分别拖动至预览区的 X-Axis、Y-Axis 和 Set color 三个虚线框中, 将其分别作为堆积面积图的 X 坐标轴、Y 坐标轴和子分类变量。指定了作图变量的图形预览区如图 19-46 所示。

(2) 输出图形。在图 19-46 中, 单击 OK 按钮 (未显示) 运行, SPSS Viewer 窗口的输出图形如图 19-47 所示。此图是把图 19-44 中的总支出按照支出类型“堆积”而成, 直观地显示了支出的构成。

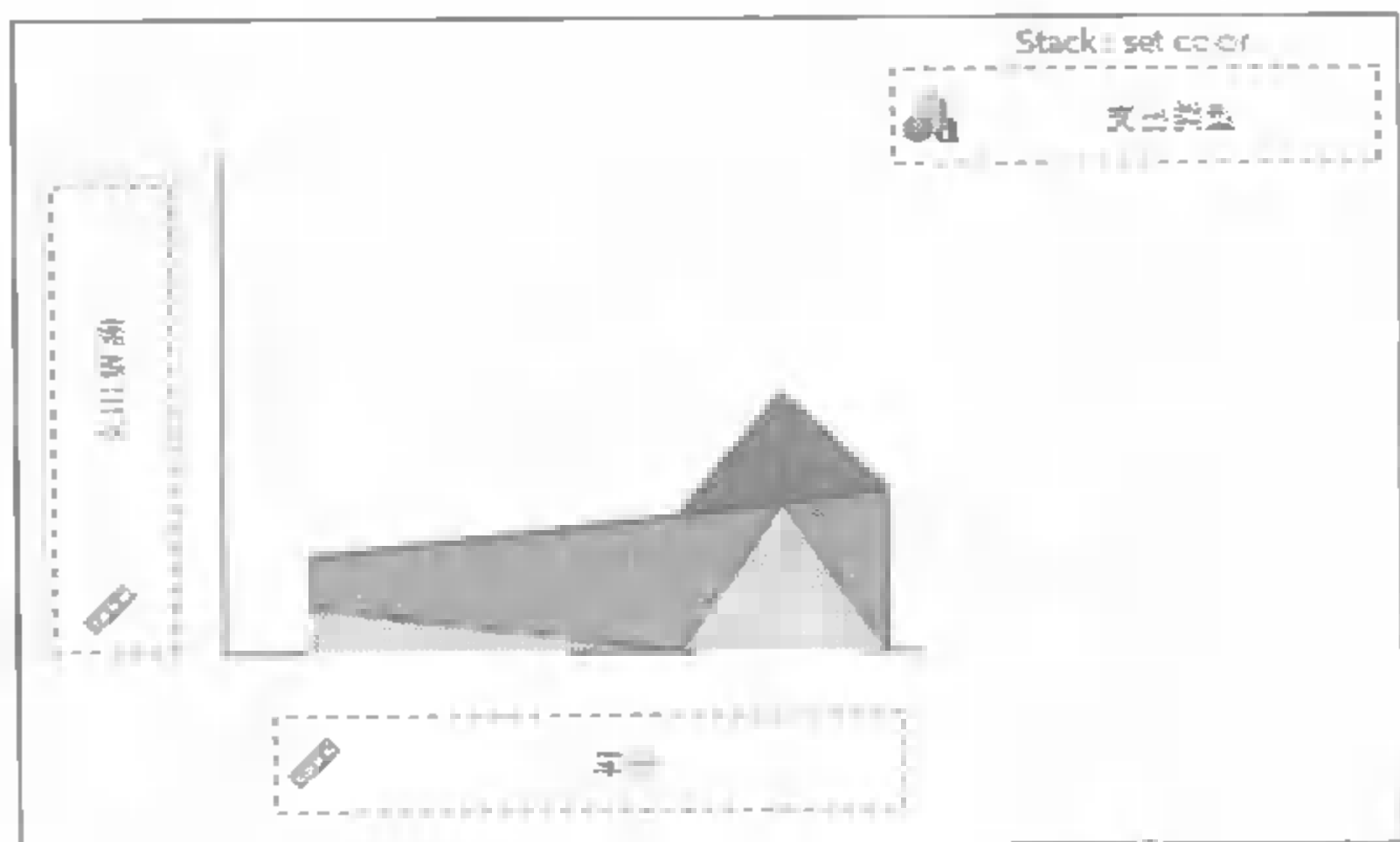


图 19-46 堆积面积图的设置预览

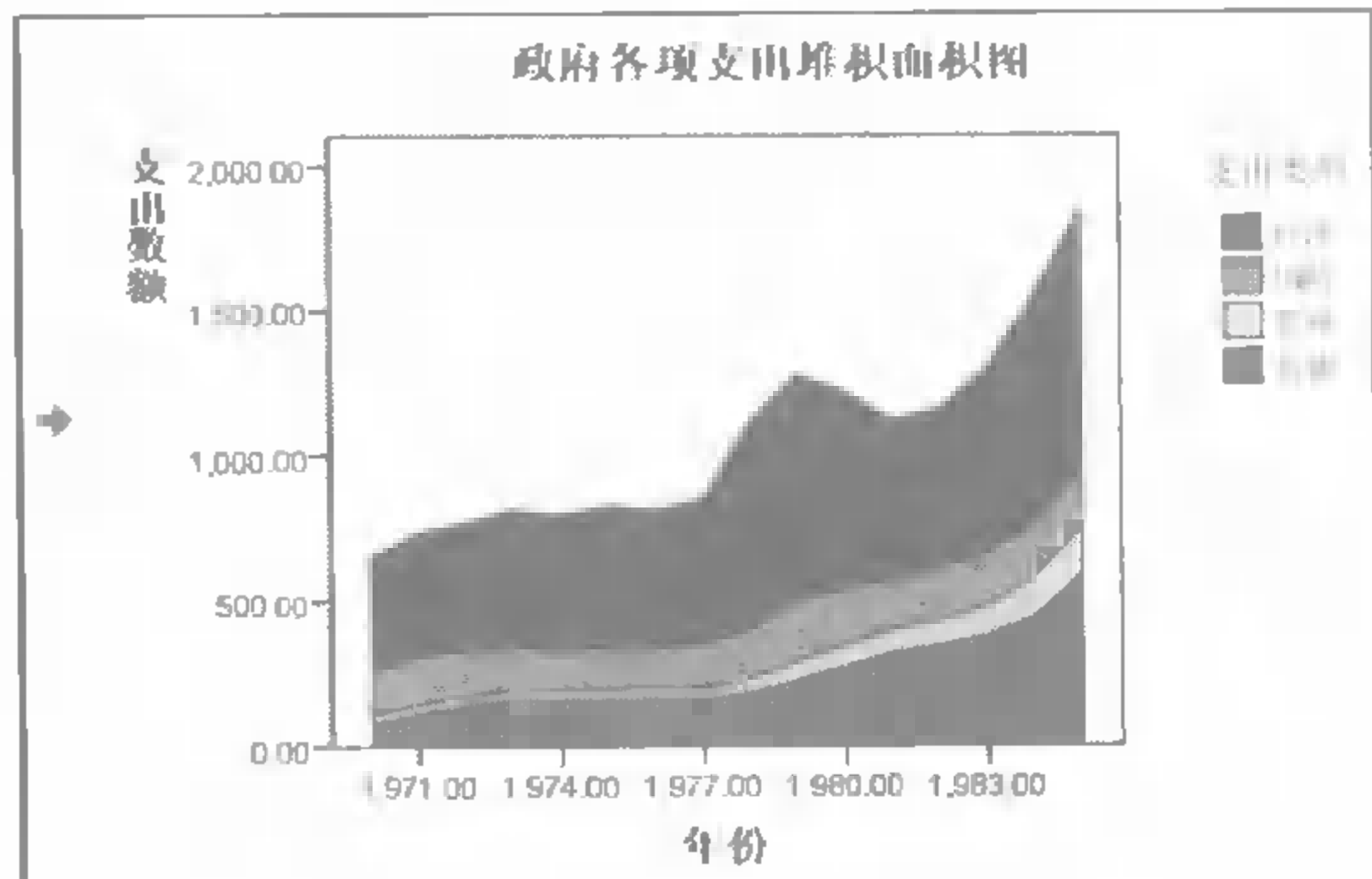


图 19-47 堆积面积图的输出

## 19.4.3 交互式面积图

依次单击菜单“Graphs→Interactive→Area...”，打开建立交互式面积图的操作界面，如图 19-48 所示。此界面的设置方法与图 19-20 相似，请参考第 19.2.3 节的介绍。

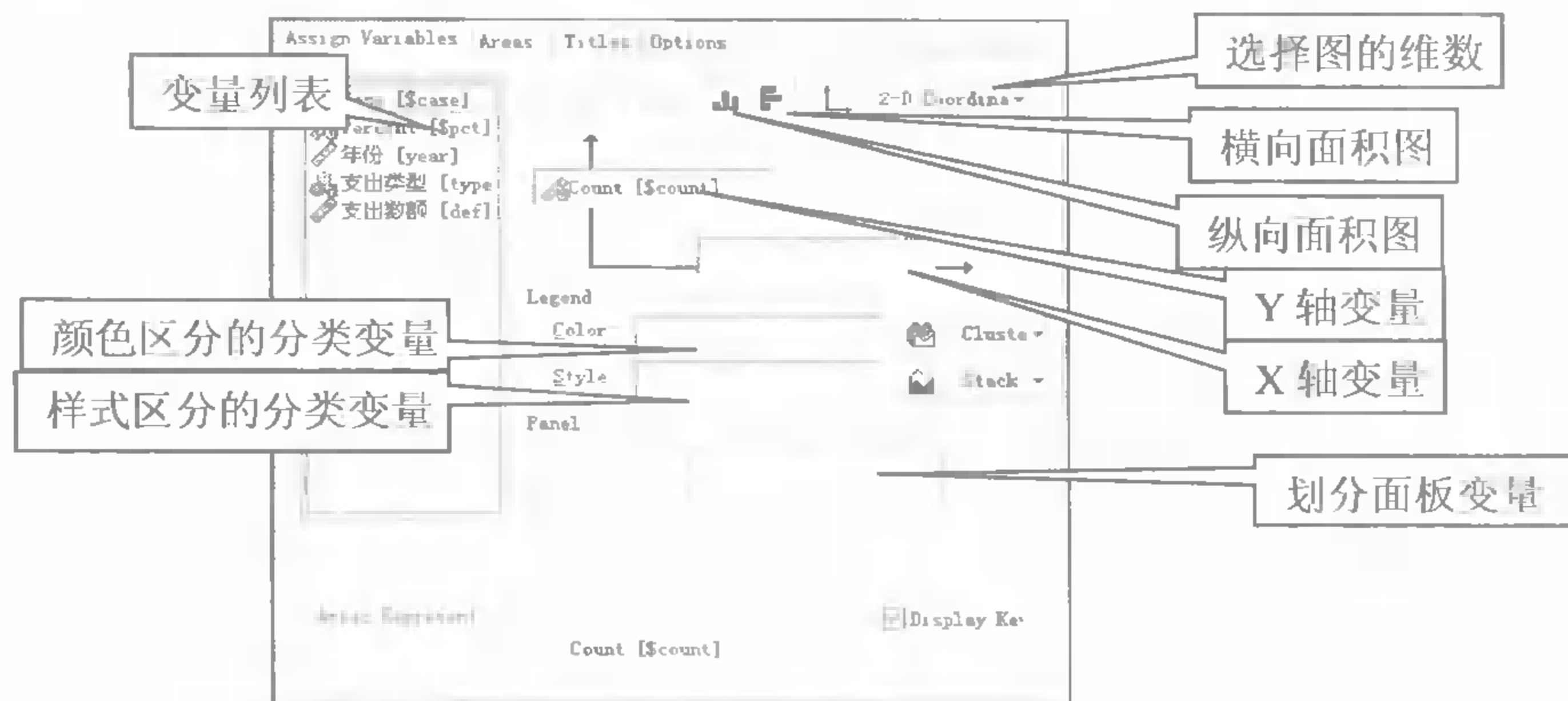


图 19-48 交互式面积图设置

打开文件“政府支出 1.sav”，在图 19-48 中：从变量列表中把年份、支出数额、支出类型，分别拖动至 X 轴变量、Y 轴变量和 Color 选框里，将其分别作为面积图的 X 坐标轴、Y 坐标轴和子分类变量；单击 Color 选框右侧的 Cluster 下拉列表选中 Stack 选项。单击确定按钮（未显示）运行，输出的图形与图 19-47 基本相同。

## 19.4.4 用对话框创建面积图

打开文件“政府支出 0.sav”，本节仍使用横向数据作关于支出数额的堆积面积图。

依次单击菜单“Graphs→Legacy Dialogs→Area...”，打开利用对话框创建面积图的选择界面，如图 19-49 所示。Data in Chart Are 子设置栏，用于指定作图的数据对象，关于它们的具体示例如图 19-29 所示。对两种面积图的设置方法都和第 19.2.4 节条形图的操作类似。

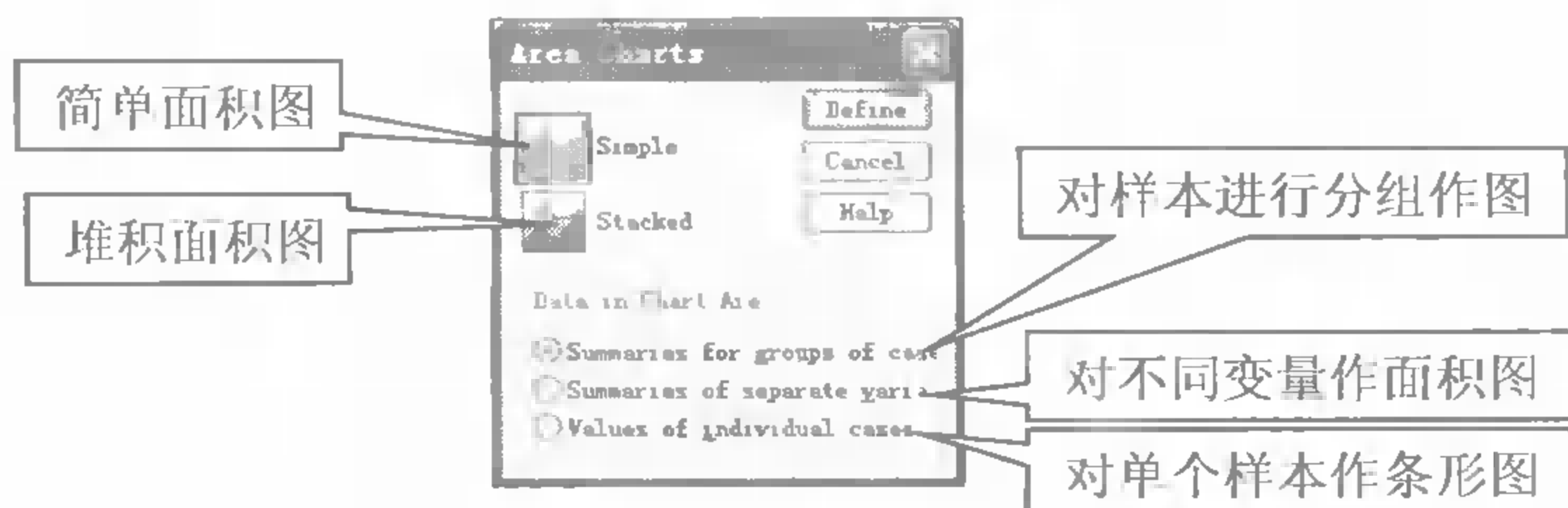




图 19-49 创建面积图的类型选择对话框

在图 19-49 中，单击选中 Stacked 图标，单击选中 Summaries of separate variable 单选框；单击 Define 按钮进入作堆积面积图的设置界面，如图 19-50 所示。

在变量列表选中从经济至其他的6个变量，单击从上至下第一个  按钮，将其作为作图变量选入 Areas Represent 列表框；在变量列表单击选中年份变量，单击从上至下第二个  按钮，将其作为 X 轴分类变量选入 Category Axis 选框。单击 OK 按钮运行，SPSS Viewer 窗口

的输出图形如图19-51所示。

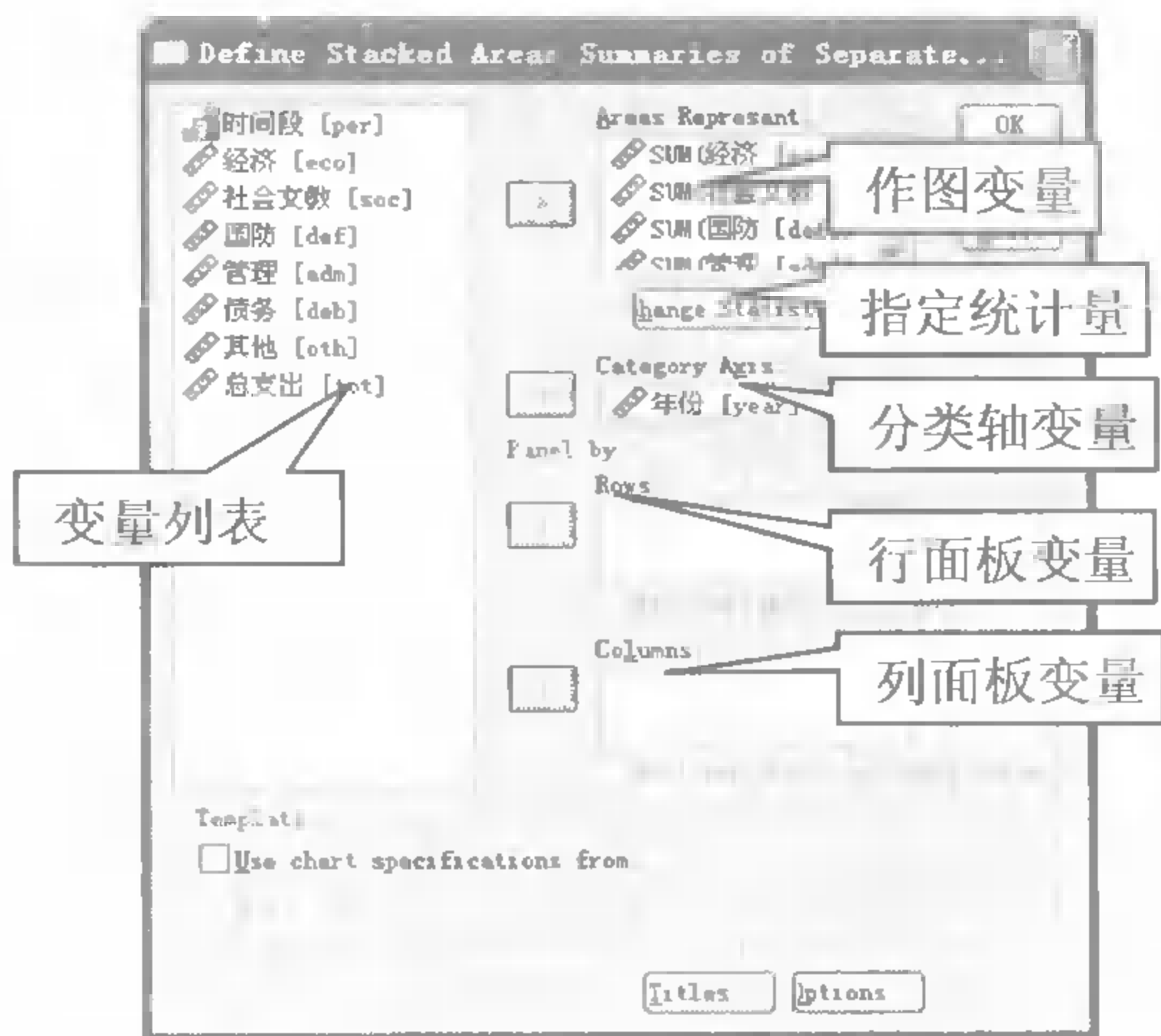


图 19-50 多变量堆积面积图的设置

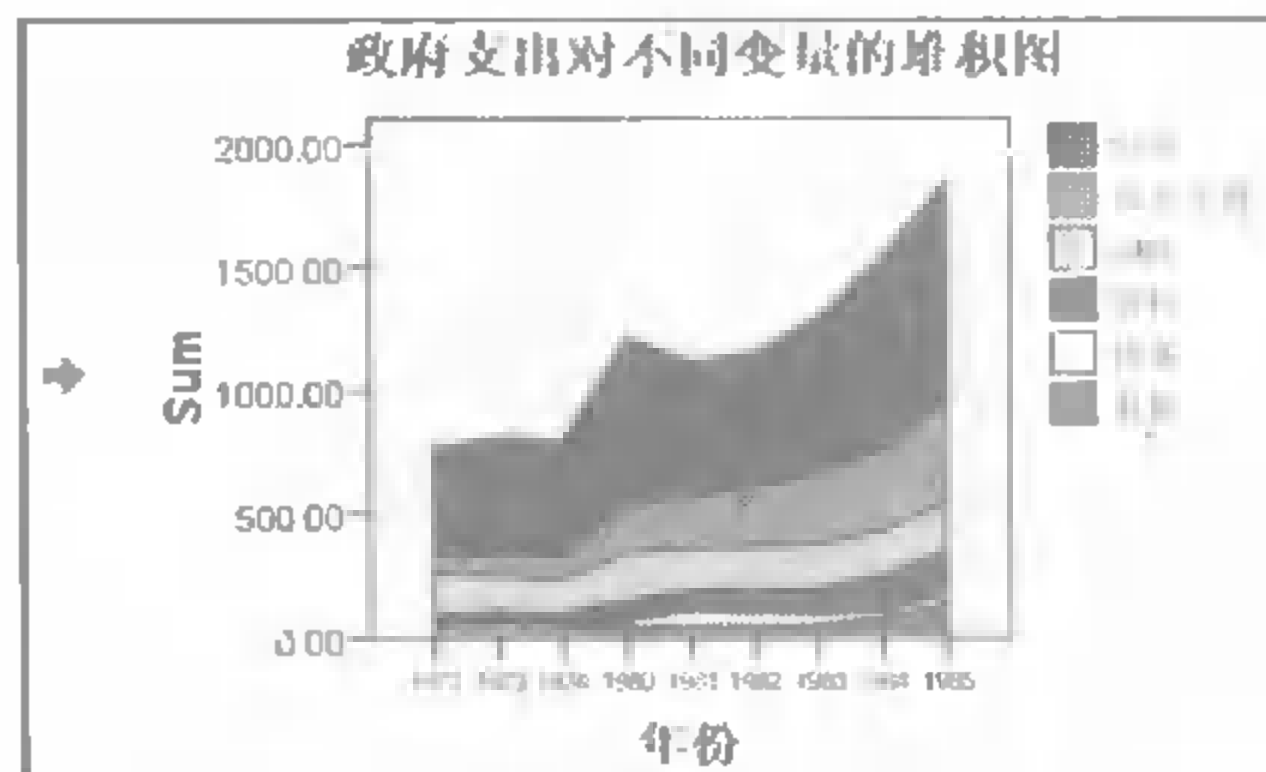


图 19-51 多变量堆积面积图的输出

## 19.5 饼图

**功能简述：**饼图也称圆图，它使用一个圆圈及其划分来表现百分比的构成。用户根据圆中各个扇形的面积大小，判断某一部分样本在全部样本中所占比例的多少。


### 19.5.1 数据和问题描述

本节使用饼图来统计美国大选的投票数据样本，观察投票人数的分布特点和各类人群的投票喜好。所用数据文件为“92 年美总统选举数据.sav”，数据格式如图 19-52 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	pres92	String	8	0	候选人	None	None	8	Left	Nominal
2	age	Numeric	2	0	投票人年龄	None	0 98 99	8	Right	Scale
3	agecat	Numeric	8	2	年龄段	1 00 不满35	None	8	Right	Scale
4	educ	Numeric	2	0	受教育年限	None	97 98 99	8	Right	Scale
5	degree	Numeric	1	0	受教育程度	0 11 high scho	7 8 9	8	Right	Scale
6	sex	String	6	0	性别	{male 男}	None	6	Left	Nominal

图 19-52 选举数据的变量格式

### 19.5.2 用图形构建器作饼图

依次单击菜单“Graphs→Chart Builder”打开图形构建器，如图 19-53 所示。单击 Gallery 标签；在 Choose from 列表单击选中 Pie/polar，双击其右侧的预置图标 (Pie Chart)。

单击 Groups/Point ID 标签，打开如图 19-54 所示的设置选项；勾选 Columns Panel variable 复选框后，在图形预览区新增一个 Panel 虚线框，用于选入划分面板的分类变量；单击 Gallery 标签返回图 19-53 所示的操作界面。

在图 19-53 中，从变量列表中把候选人、性别两个变量，分别拖动至预览区的 Slice by、Panel 两个虚线框中。单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-55 所示，可见男性选 Clinton 的居多，女性选 Bush 的比例要比男性大一些。



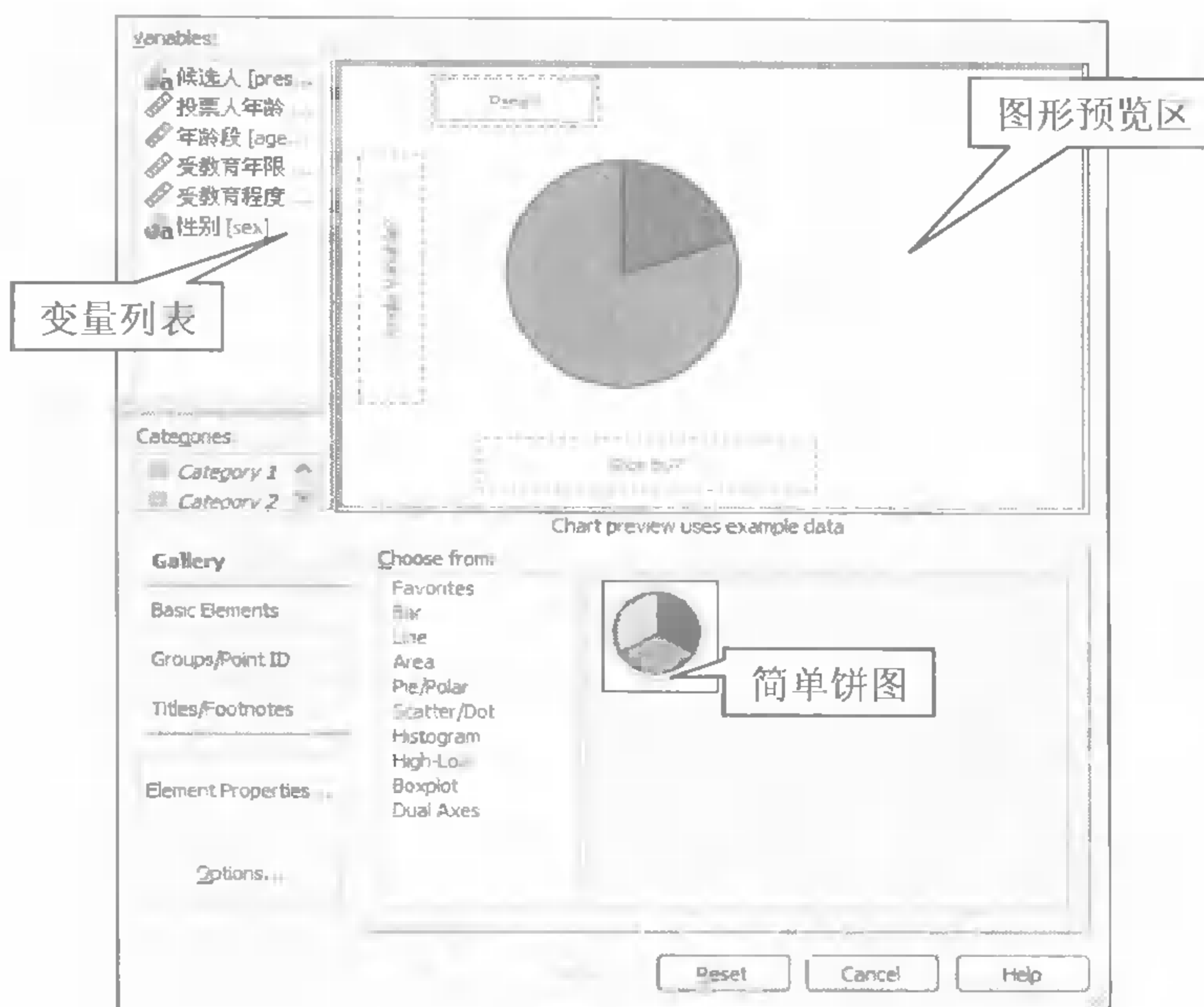


图 19-53 创建简单饼图的设置界面

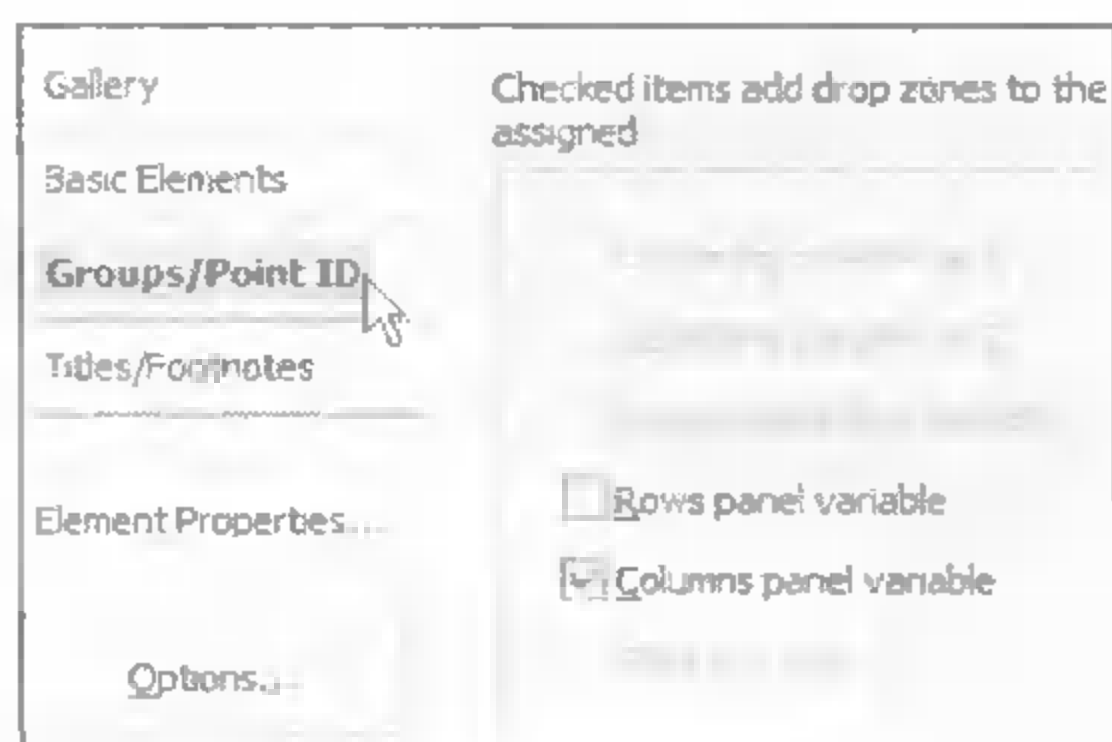


图 19-54 Groups/Point ID 设置

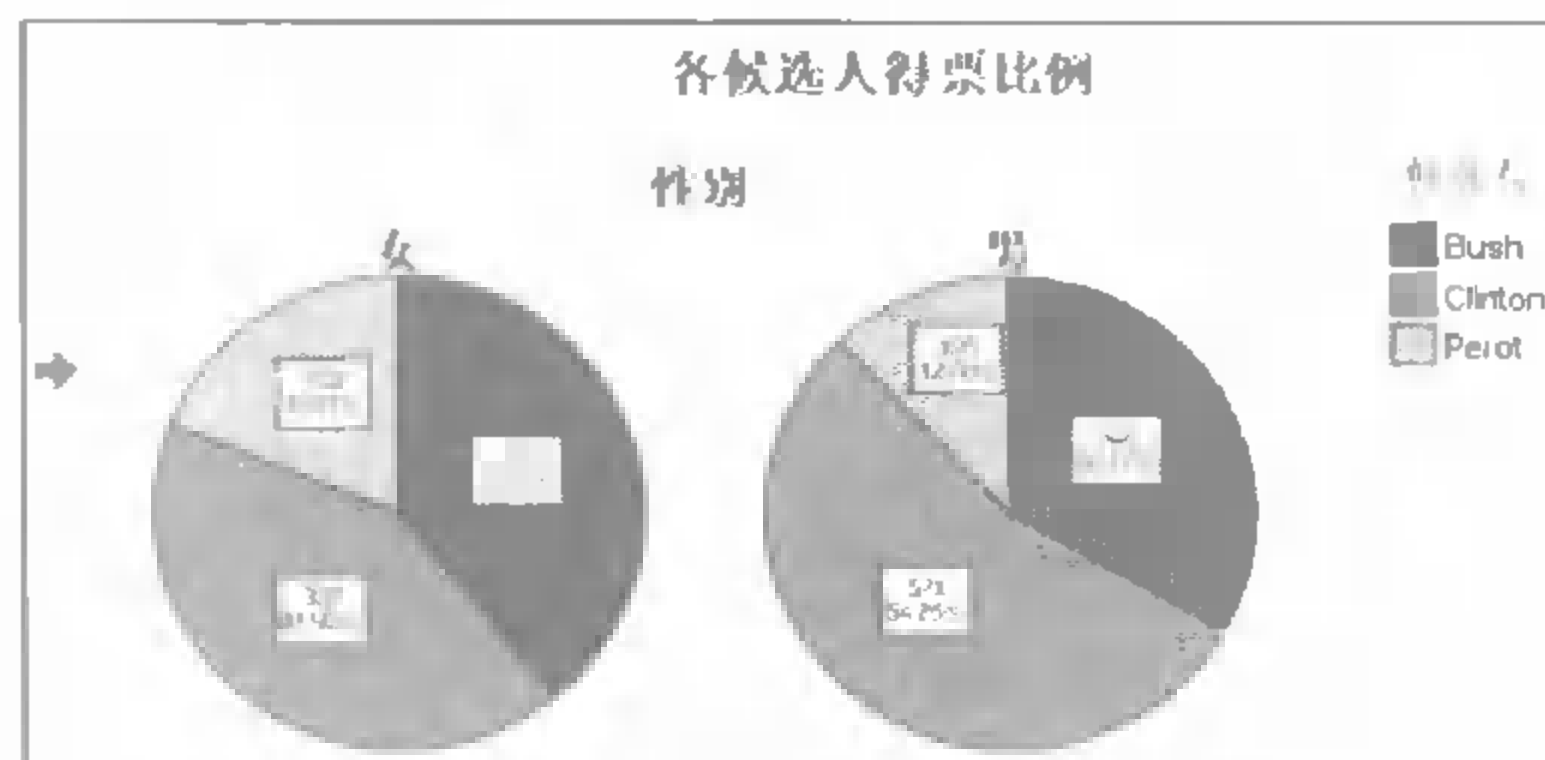


图 19-55 按性别分类对各候选人的投票比例

### 19.5.3 交互式饼图

依次单击菜单“Graphs→Interactive→Pie”，弹出如下 3 个子菜单：Simple（简单饼图）、Clustered（分类饼图）、Plotted（分组饼图）。

#### 1. Plotted Pie 的示例

依次单击菜单“Graphs→Interactive→Pie→Plotted”，打开建立交互式饼图的操作界面，如图 19-56 所示。此界面的设置方法与图 19-20 相似，请参考第 19.2.3 节的介绍。

从变量列表中把年龄段、性别、候选人，分别拖动至 X 轴变量、Y 轴变量和 Slice 选框里，将其分别作为饼图的 X 坐标轴、Y 坐标轴和划分扇形的变量。单击确定按钮运行，SPSS Viewer 窗口的输出图形如图 19-57 所示。

#### 2. Clustered Pie 的示例

依次单击菜单“Graphs→Interactive→Pie→Clustered”，打开建立交互式饼图的操作界面，与图 19-56 相仿。从变量列表中把候选人、性别，分别拖动至 Slice 选框和 Cluster 选框里，将其分别作为饼图的划分扇形变量和子分类变量。单击确定按钮运行，SPSS Viewer 窗口的

输出图形如图 19-58 所示。

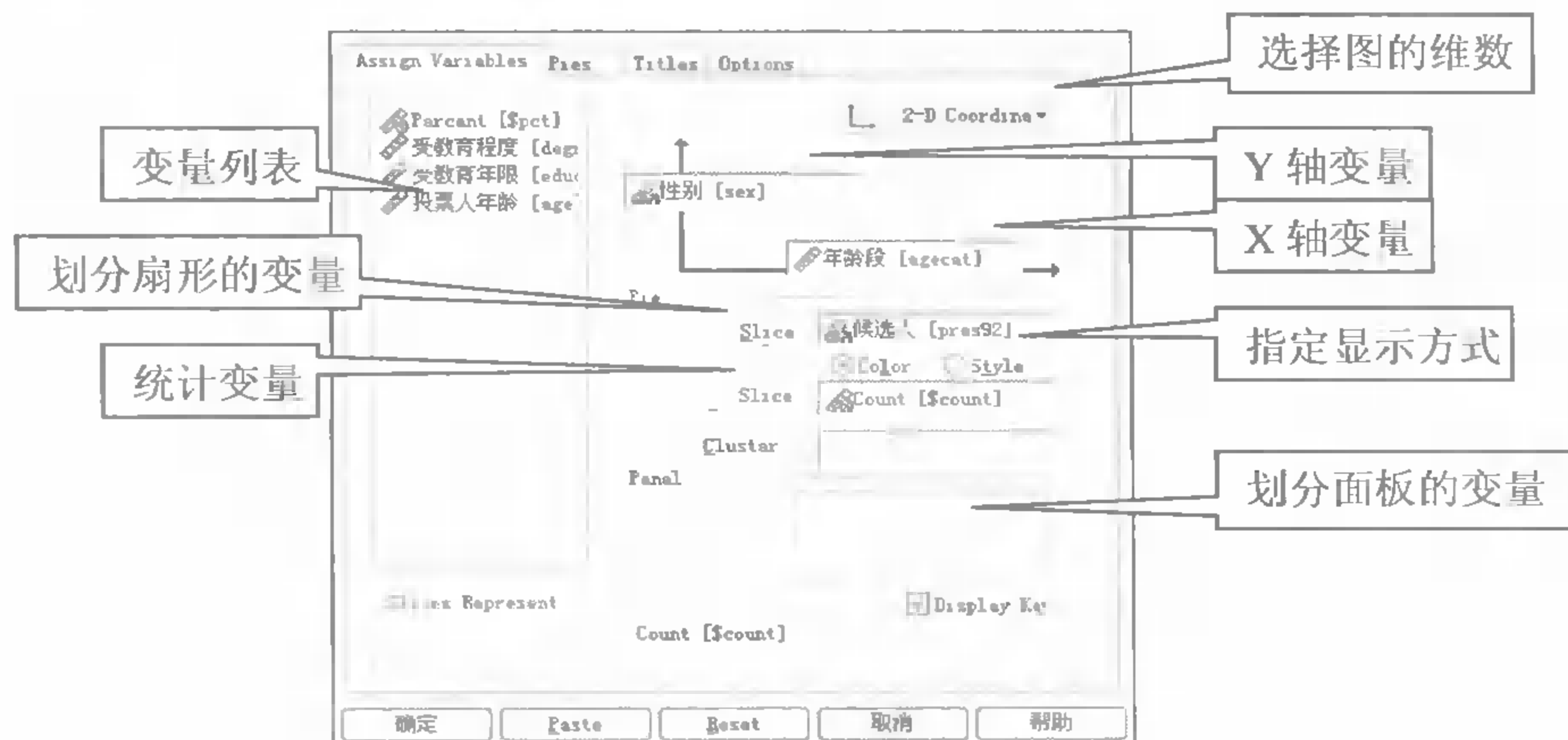


图 19-56 交互式面积图设置

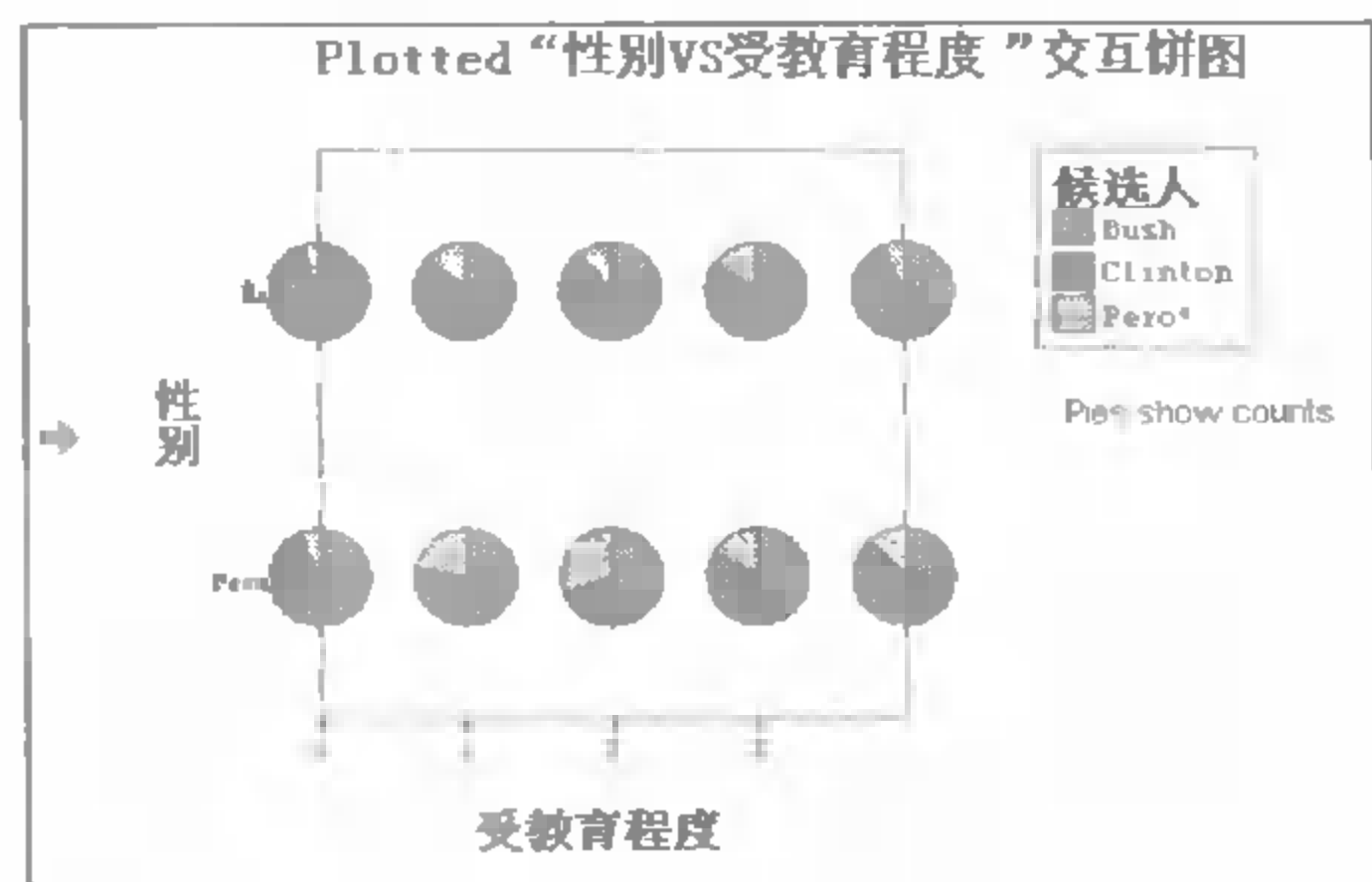


图 19-57 Plotted 交互饼图的输出

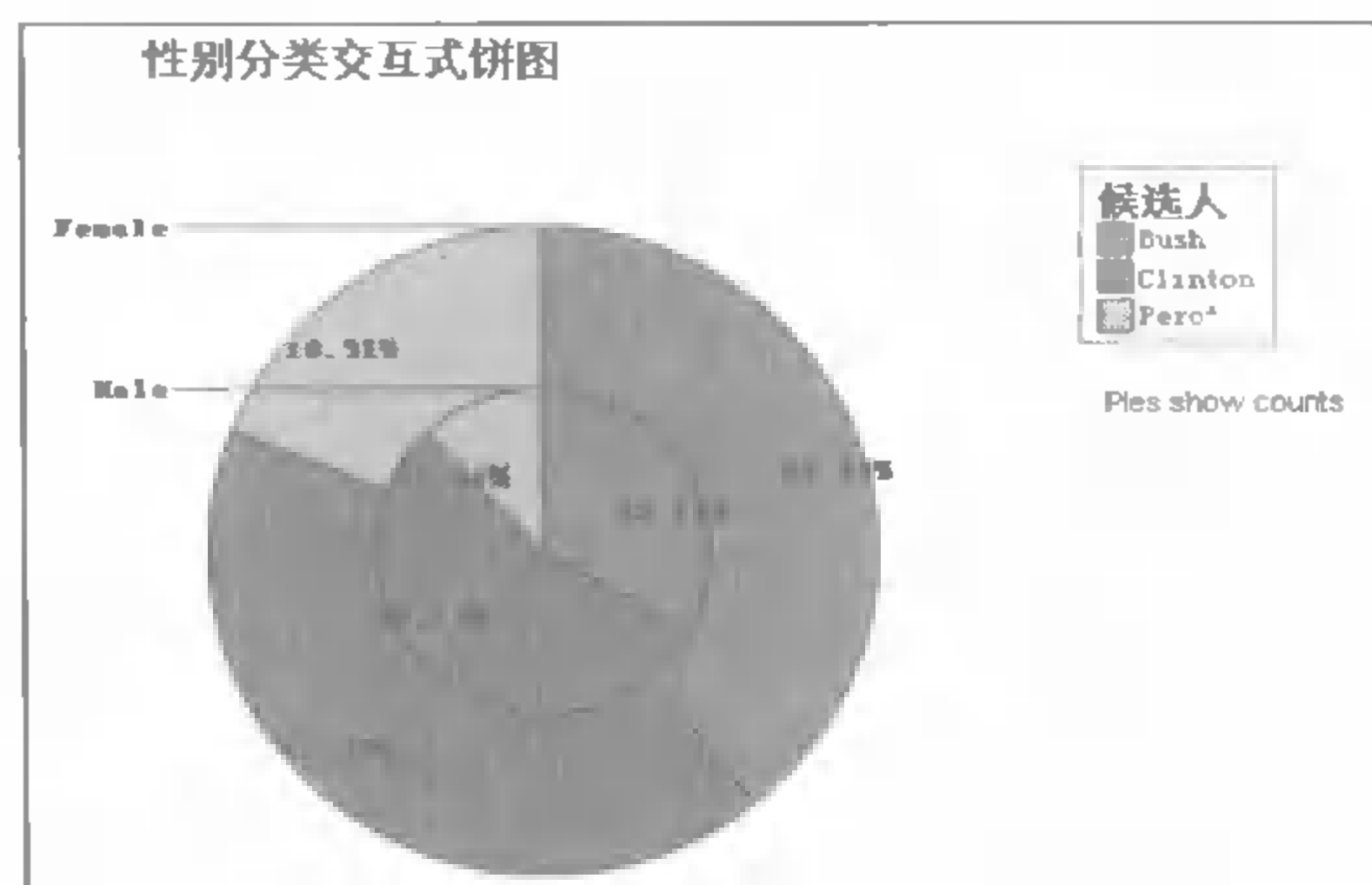


图 19-58 Plotted 交互饼图的输出

#### 19.5.4 用对话框创建饼图

依次单击菜单“Graphs→Legacy Dialogs→Pie...”，打开利用对话框创建饼图的类型选择界面，如图 19-59 所示。Data in Chart Are 子设置栏，用于指定作图的数据对象，关于它们的具体示例如图 19-29 所示。

对样本进行分组作图的情况，与第 19.2.4 节作条形图的设置相仿（如图 19-30 所示）。

对不同变量作饼图的情况，与第 19.3.4 节作线形图的设置相仿（如图 19-40 所示）。

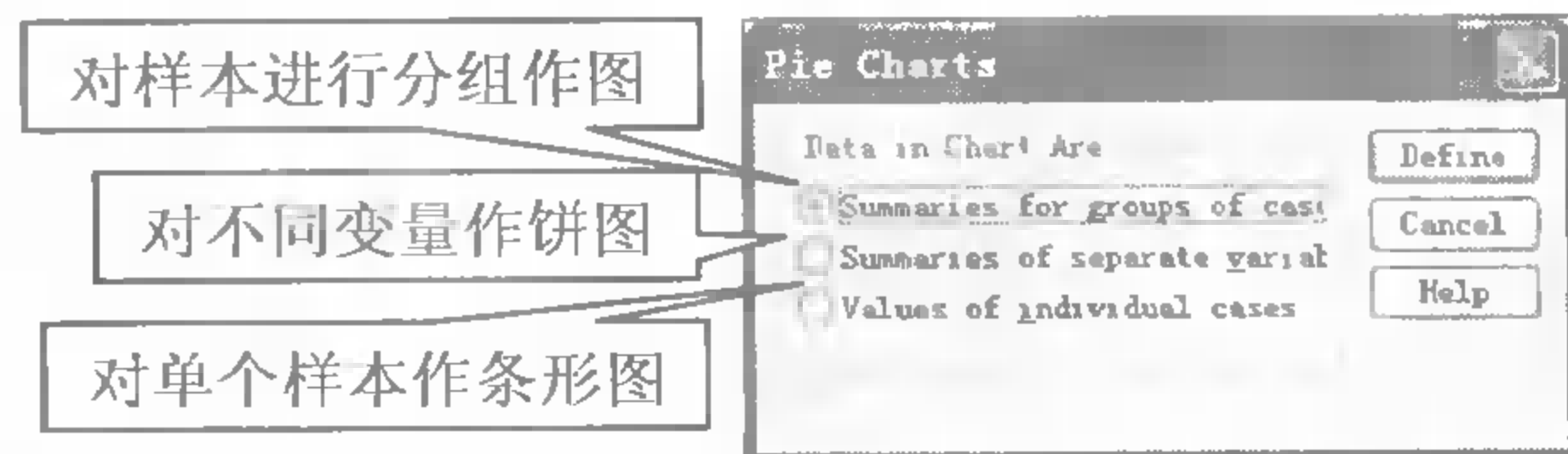


图 19-59 创建饼图的类型选择对话框

### 19.6 高低图

高低图适用于表现某种形式的数据区域，例如：测量值的范围（最小值～最大值）、95%

的置信区间（最低限～最高限）等。

### 19.6.1 数据和问题描述

本节用高低图来描绘股票价格的变化。所用数据文件有如下 3 个：“地产股票价格.sav”、“北京地区股票价格.sav”和“工商业股票价格.sav”，它们的数据格式分别如图 19-60、图 19-61 和图 19-62 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	cat	Numeric	3	0	券种	{1 兴业房}	None	8	Right	Nominal
2	hlc	Numeric	5	0	价格类别	{1 最高}	None	8	Right	Nominal
3	date	Date			日期	None	None	14	Right	Scale
4	value	Numeric	5	2	价格	None	None	8	Right	Scale

图 19-60 地产股票价格的数据格式

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	week	Date			Week	None	None	8	Right	Scale
2	date	Date			Date	None	None	8	Right	Scale
3	bl_hi	Numeric	5	2	Beilu	None	None	8	Right	Scale
4	bl_lo	Numeric	5	2	High and Low	None	None	8	Right	Scale
5	bl_cl	Numeric	5	2	Close	None	None	8	Right	Scale
6	wfj_hi	Numeric	5	2	Wangfujing	None	None	8	Right	Scale


图 19-61 北京地区股票价格的数据格式

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	date	Date			日期	None	None	10	Right	Scale
2	cat	Numeric	3	0	股票名称	{1 济南轻骑}	None	12	Right	Nominal
3	high	Numeric	5	2	最高价	None	None	8	Right	Scale
4	low	Numeric	5	2	最低价	None	None	8	Right	Scale
5	close	Numeric	5	2	收盘价	None	None	8	Right	Scale
6	group	Numeric	8	0	券种	{1 商业}	None	8	Right	Nominal

图 19-62 工商业股票价格的数据格式

### 19.6.2 用图形构建器作高低图

打开文件“地产股票价格.sav”，下面用高低图来描绘几种地产股票的最高价、最低价、收盘价的均值随日期而变化的趋势。

依次单击菜单“Graphs→Chart Builder”打开图形构建器，如图 19-63 所示。单击 Gallery 标签；在 Choose from 列表单击选中 Scatter/Dot，双击其右侧的预置图标（Drop-Line）后，在图形预览区给出高低图的预览，同时自动弹出元素属性设置面板。

在图 19-63 中，从变量列表里把日期、价格、价格类别三个变量，分别拖动至预览区的 X-Axis、Y-Axis 和 Set color 三个虚线框中，将其分别作为高低图的 X 坐标轴、Y 坐标轴和区分高低值的分类变量。

在图 19-63 中的元素属性设置面板里，单击选中 Edit 列表框里的 Y-Axis1；在 Axis Label 后面输入“平均价格”作为 Y 轴标签；单击 Minimum 复选框取消选中，在其后面的 Custom 输入框键入“10”作为 Y 轴的最小刻度；单击 Apply 按钮应用设置。

在图 19-63 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-64 所示，它同时描绘了地产股票的三个价格的走势规律。

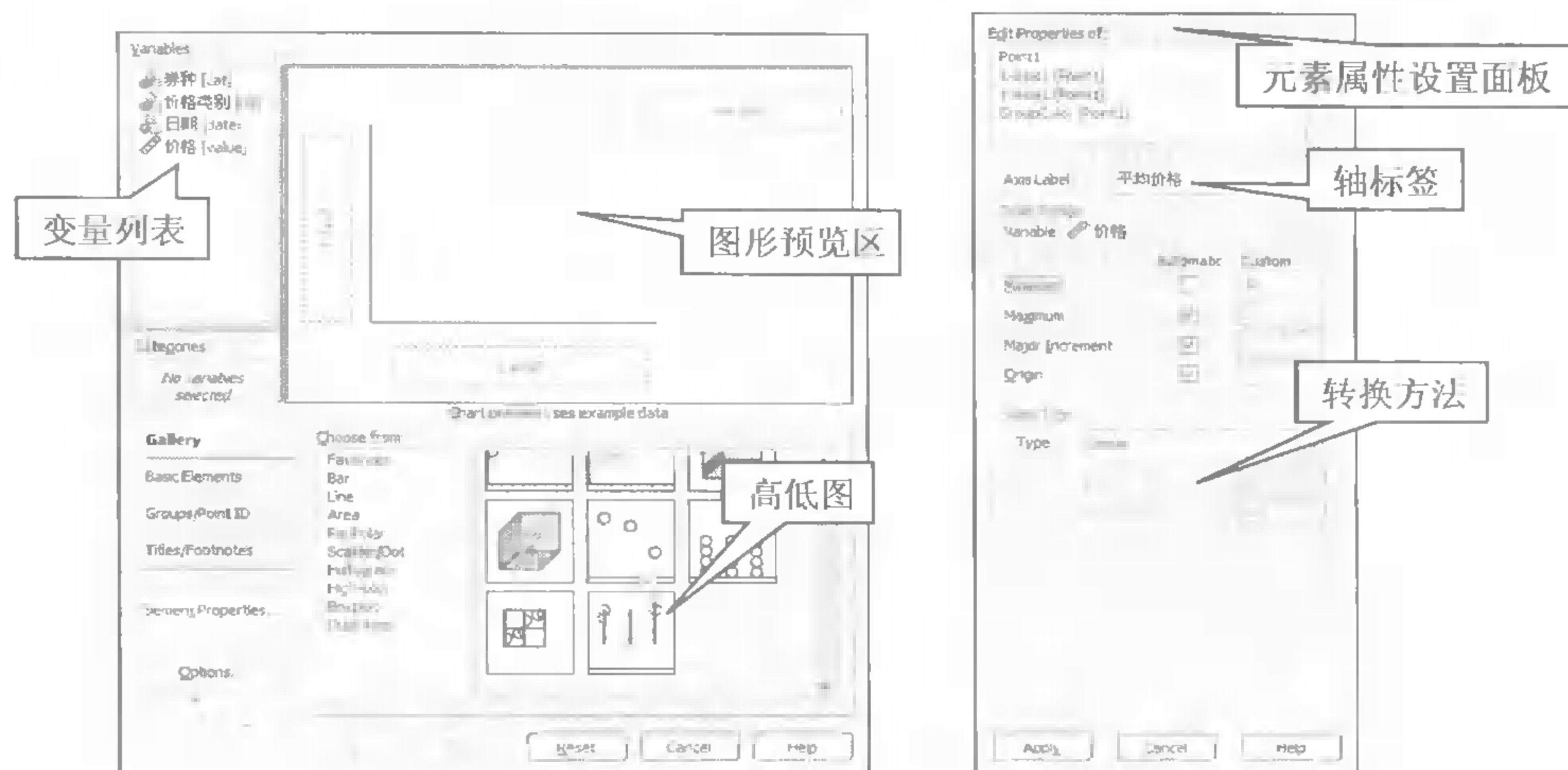


图 19-63 创建高低图的设置界面

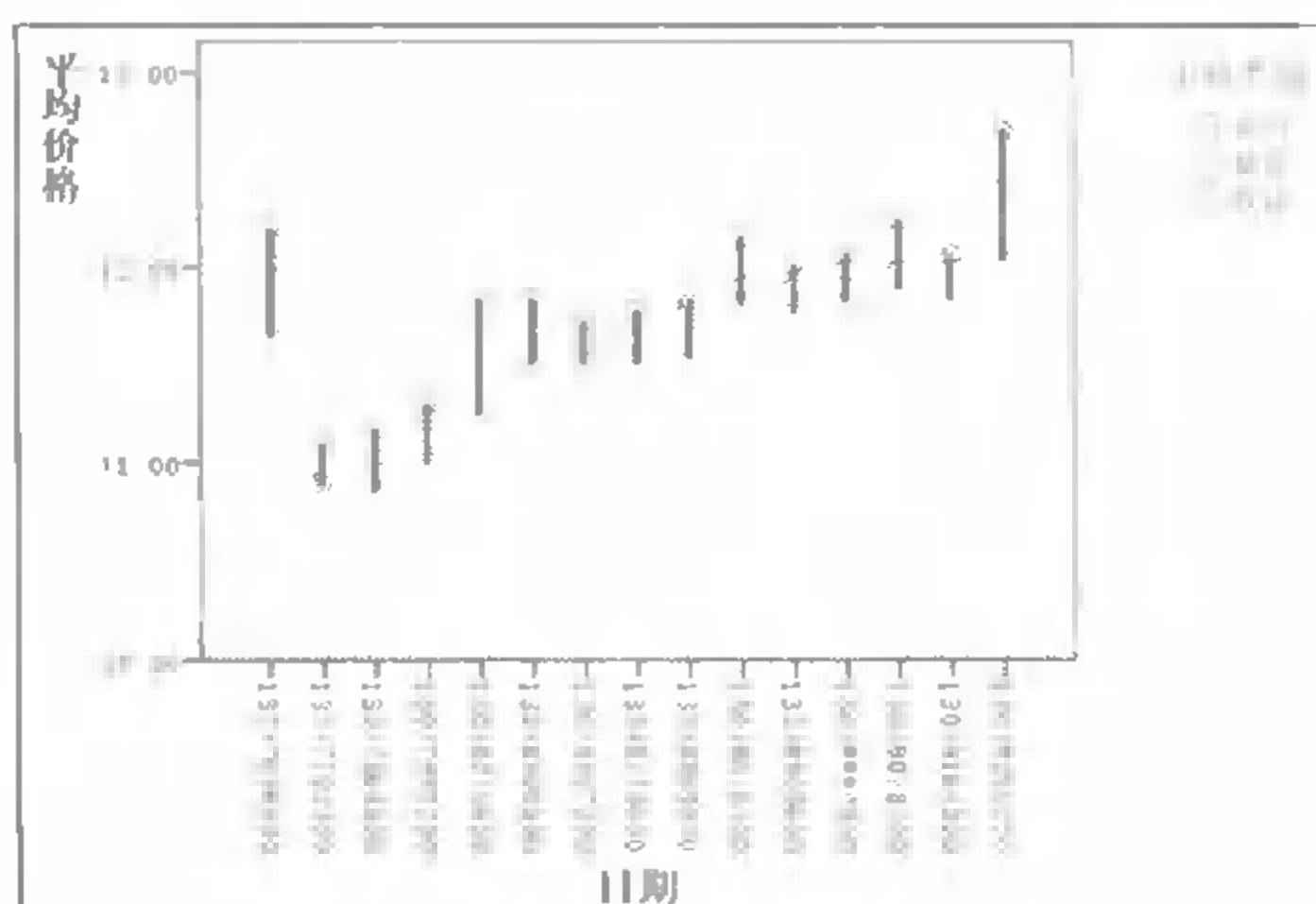


图 19-64 地产股票均价的走势高低图

### 19.6.3 交互式高低图

本节仍使用文件“地产股票价格.sav”来作高低图。

依次单击菜单“Graphs→Interactive→Drop Line...”，打开建立交互式高低图的操作界面，如图 19-65 所示。此界面的设置方法与图 19-36 相似，请参考第 19.3.3 节的介绍。

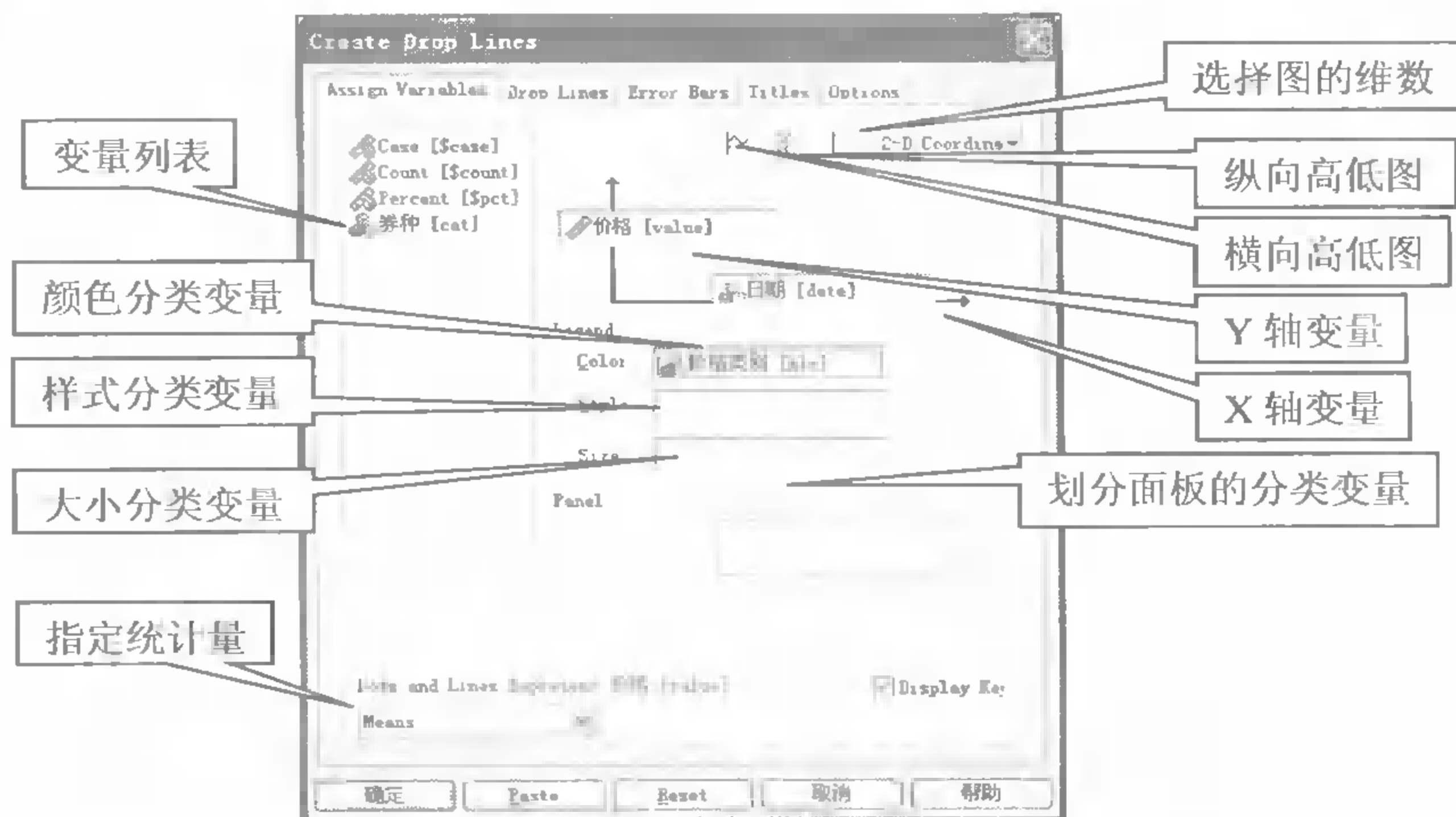


图 19-65 交互式高低图的设置



在图 19-65 中，从变量列表里把日期、价格、价格类别三个变量，分别拖动至 X 轴变量、Y 轴变量和 Color 三个选框里，将其分别作为高低图的 X 坐标轴、Y 坐标轴和区分高低值的分类变量。单击确定按钮运行，SPSS Viewer 窗口的输出图形与图 19-64 基本相同。

#### 19.6.4 用对话框创建高低图

依次单击菜单“Graphs→Legacy Dialogs→High-Low...”，打开利用对话框创建高低图的类型选择界面，如图 19-66 所示。Data in Chart Are 子设置栏，用于指定作图的数据对象，关于它们的具体示例如图 19-29 所示。

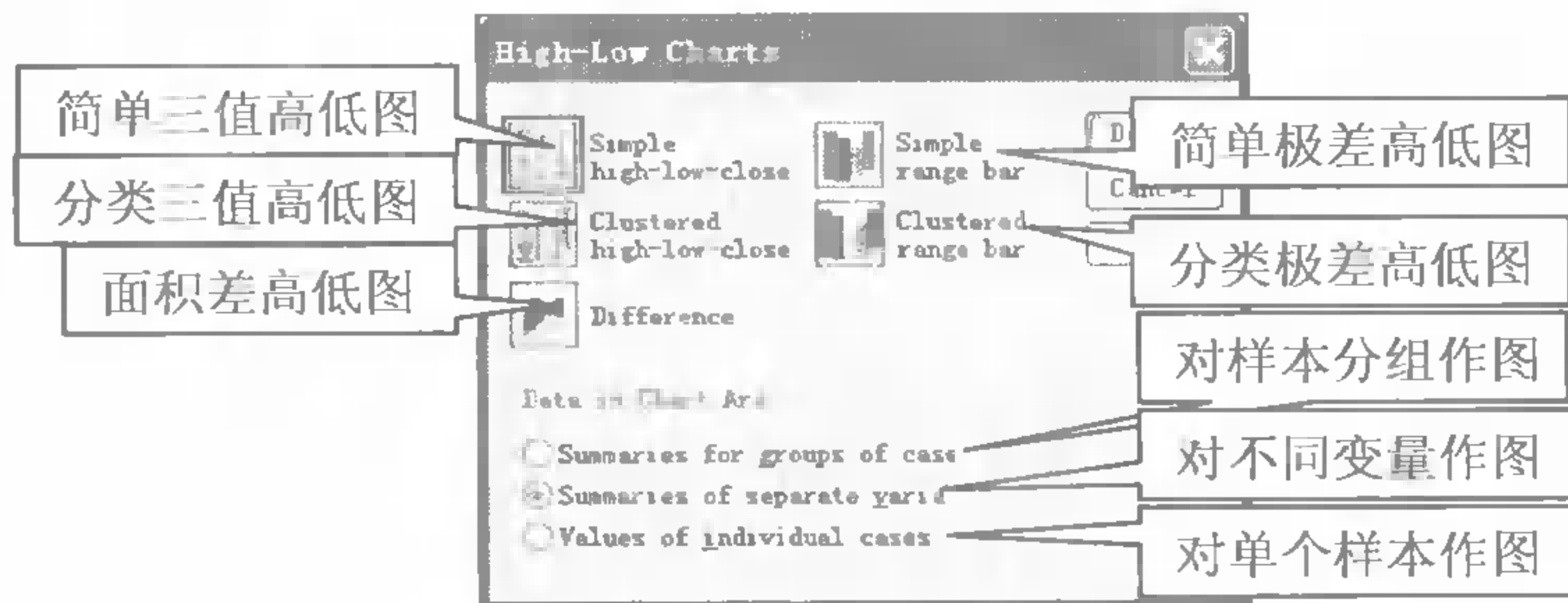


图 19-66 建立高低图选择对话框

##### 1. 对样本分组作图 (Summaries for Groups of Cases)

(1) 简单三值高低图。打开文件“地产股票价格.sav”，先来作关于地产股票价格的简单三值高低图。

在图 19-66 中，单击选中 Simple high-low-close 图标，单击选中 Summaries for Groups of Cases 单选框；单击 Define 按钮进入作简单三值高低图的设置界面，如图 19-67 所示。

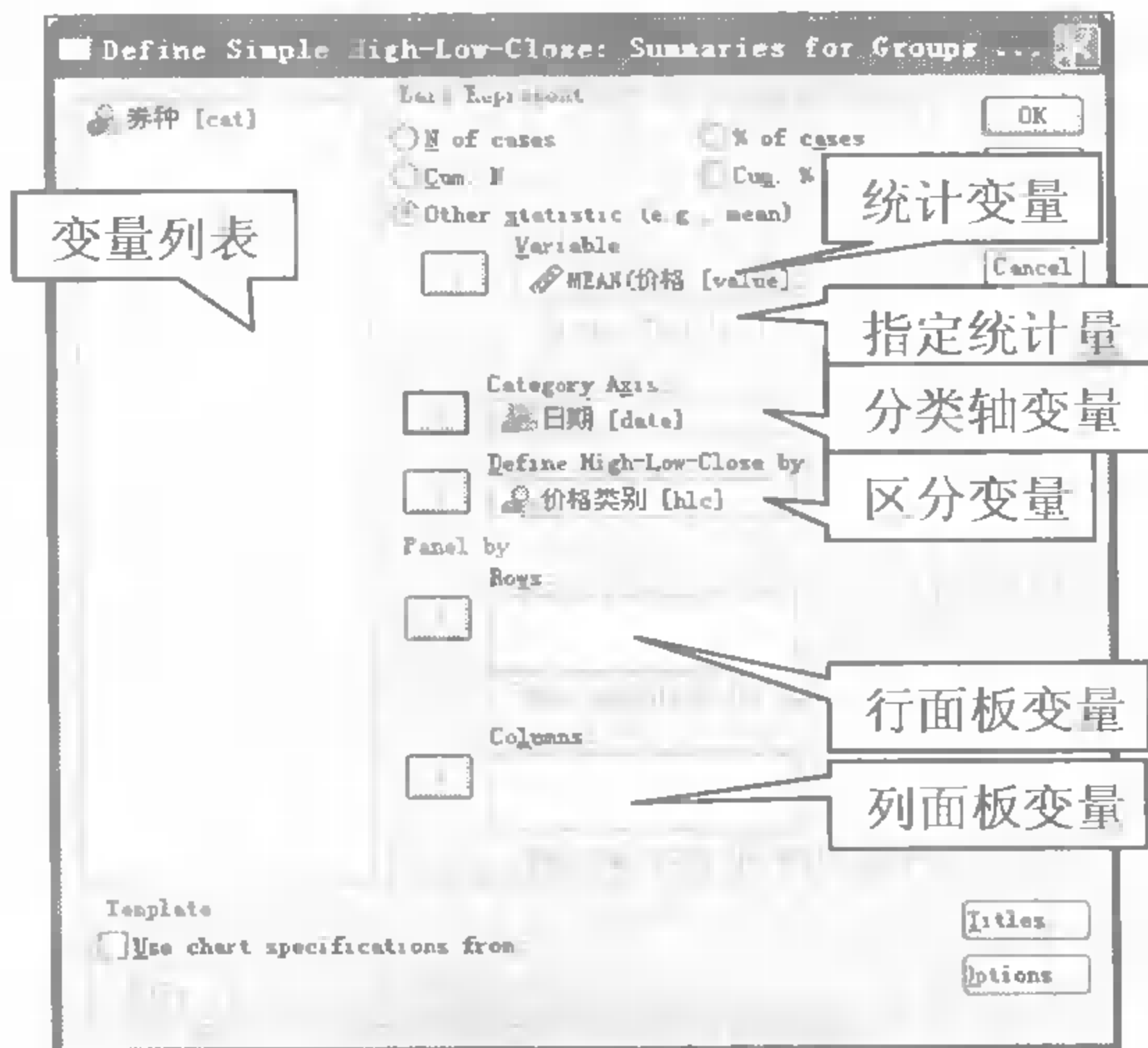





图 19-67 简单三值高低图的设置

在图 19-67 里：单击选中 Other 单选框；在变量列表单击选中价格变量，单击从上至下第一个  按钮，将其作为统计变量选入 Variable 选框；在变量列表单击选中日期变量，单击从上至下第二个  按钮，将其作为 X 轴分类变量选入 Category Axis 选框；在变量列表单击选中价格类别变量，单击从上至下第三个  按钮，将其作为区分高低值的分类变量选入 Define

high-low-close by 选框。单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-68 所示。

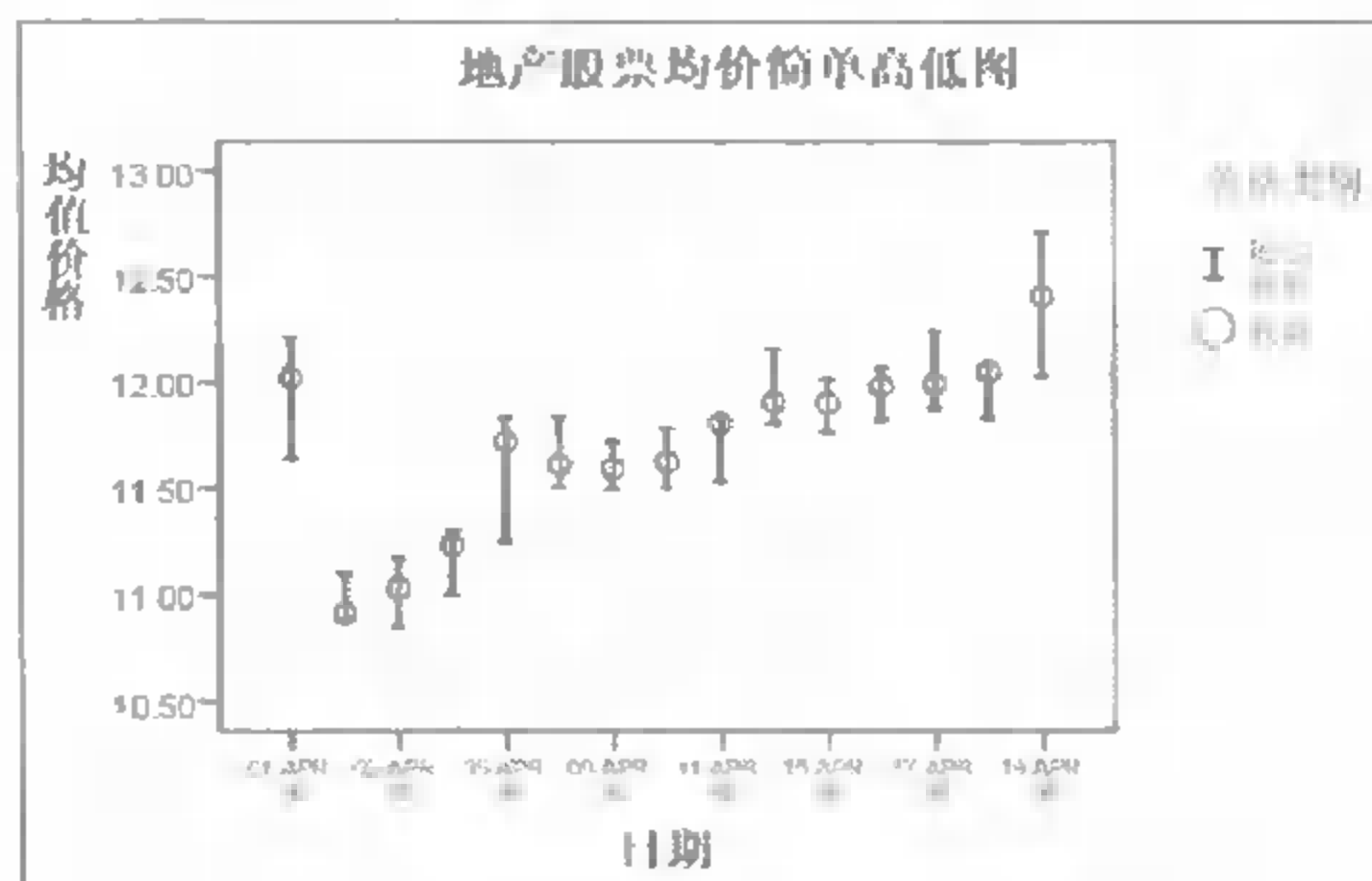





图 19-68 简单三值高低图的输出

(2) 简单极差高低图。打开文件“工商业股票价格.sav”，作关于工商业股票价格的简单极差高低图。

在图 19-66 中单击选中 Simple range bar 图标，单击选中 Summaries for Groups of Cases 单选框；单击 Define 按钮进入作简单极差高低图的设置界面，它与图 19-67 基本相同。

单击选中 Other 单选框；在变量列表中单击选中收盘价变量，单击从上至下第一个  按钮，将其作为统计变量选入 Variable 选框；在变量列表中单击选中日期变量，单击从上至下第二个  按钮，将其作为 X 轴分类变量选入 Category Axis 选框；在变量列表中单击选中币种变量，单击从上至下第三个  按钮，将其作为区分高低值的分类变量选入 Define 2 Groups By 选框。单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-69 所示。

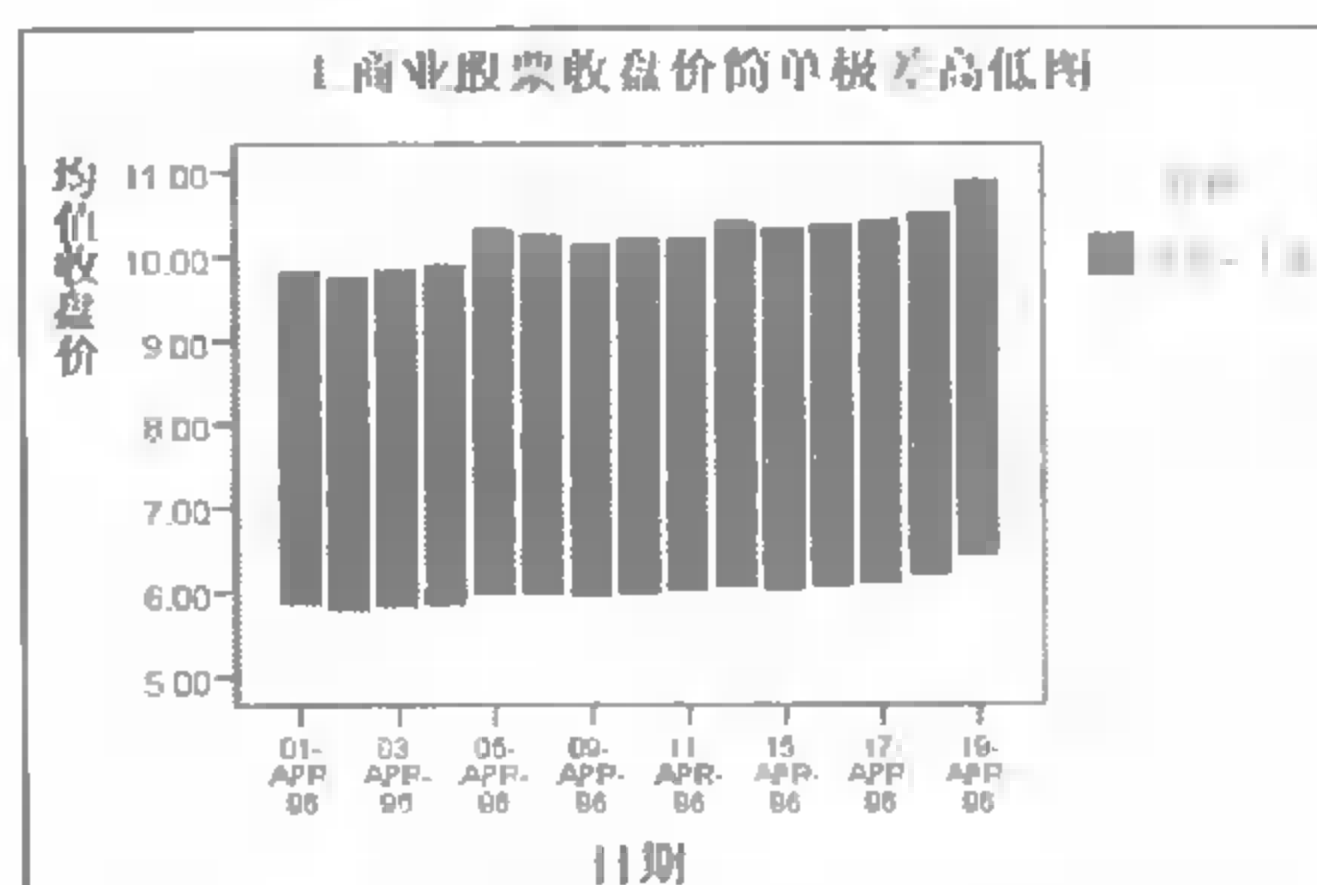


图 19-69 简单极差高低图的输出

(3) 面积差高低图。打开文件“工商业股票价格.sav”，作关于工商业股票价格的面积差高低图。

在图 19-66 中单击选中 Difference Area 图标，单击选中 Summaries for Groups of Cases 单选框；单击 Define 按钮进入作面积差高低图的设置界面，它与图 19-67 基本相同。

采用和作简单极差高低图时完全相同的设置方法，输出图形如图 19-70 所示。

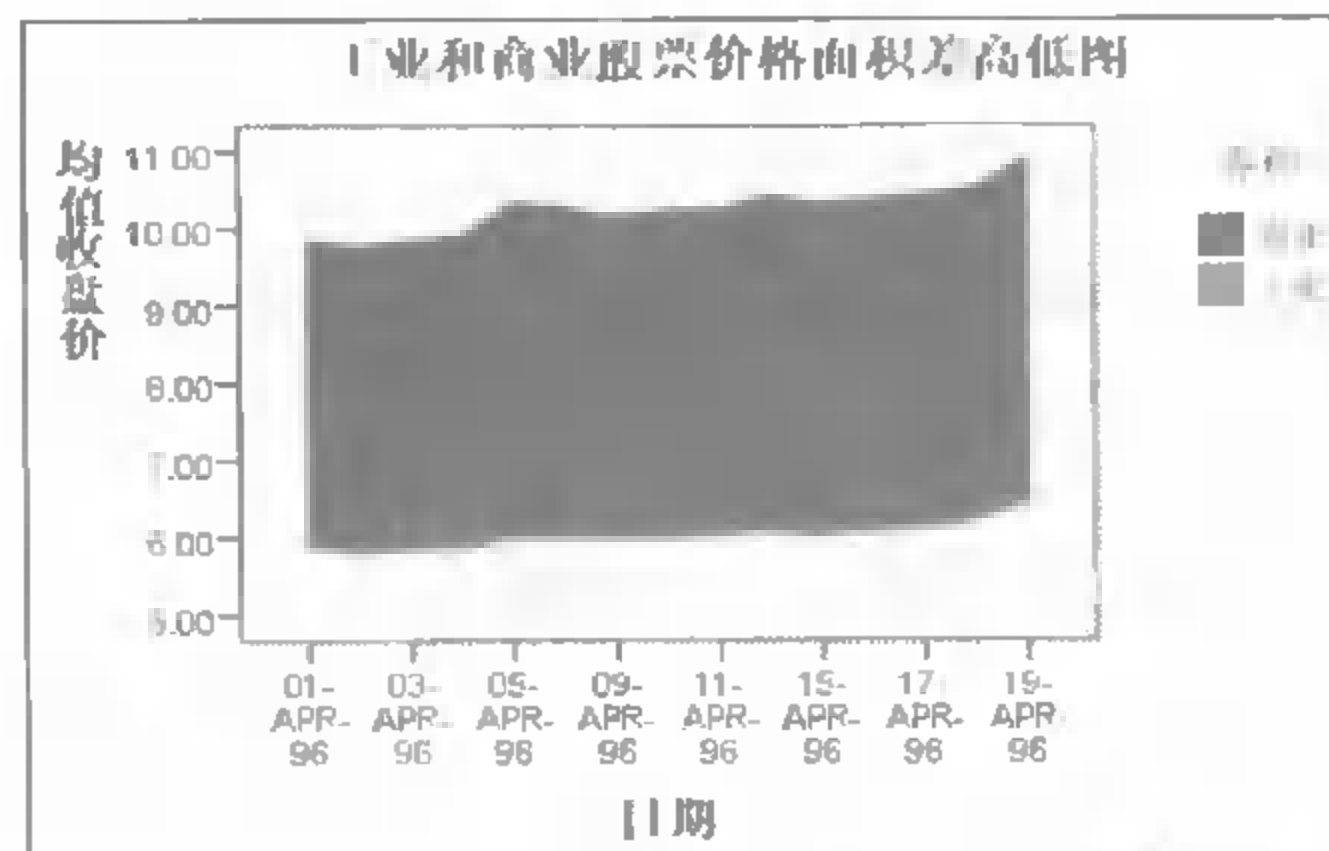


图 19-70 面积差高低图的输出

(4) 分类三值高低图。打开文件“工商业股票价格.sav”，作关于工商业股票价格的分类三值高低图。

在图 19-66 中单击选中 Clustered high-low-close 图标，单击选中 Summaries for Groups of Cases 单选框；单击 Define 按钮进入作分类三值高低图的设置界面，它与图 19-67 非常相似。

从变量列表里把最高价、最低价、收盘价、日期、券种这 5 个变量，分别选入 High、Low、Close、Category Axis 和 Define Cluster By 这 5 个选框，对 3 种价格的统计量都采用默认的 Mean（均值）。单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-71 所示。

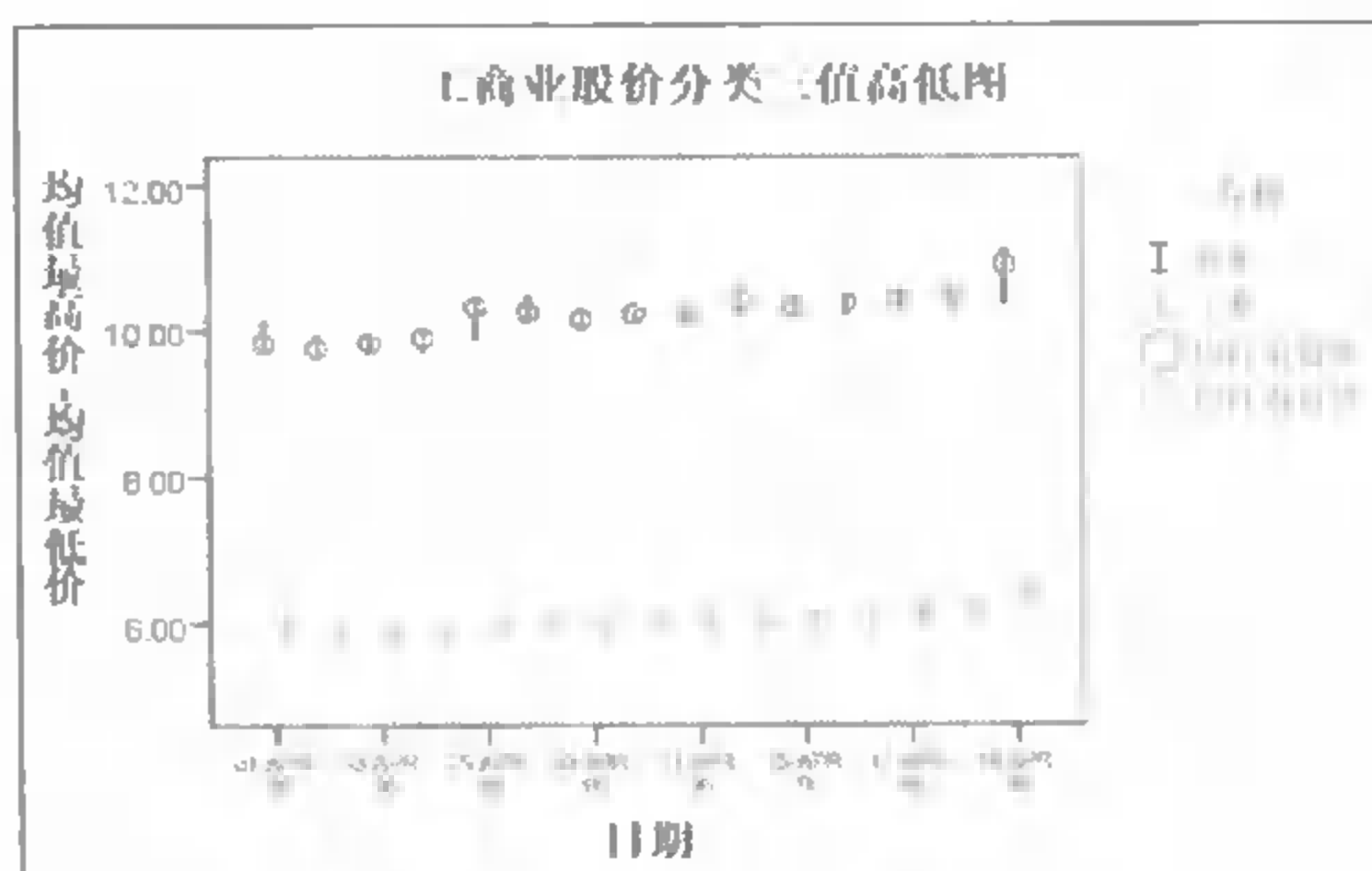


图 19-71 分类三值高低图的输出

(5) 分类极差高低图。打开文件“工商业股票价格.sav”，作关于工商业股票价格的分类极差高低图。

在图 19-66 中单击选中 Clustered range bar 图标，单击选中 Summaries for Groups of Cases 单选框；单击 Define 按钮进入作分类极差高低图的设置界面，它与图 19-67 非常相似。

从变量列表里把最高价、最低价、日期、券种这 4 变量，分别选入 1st、2nd、Category Axis 和 Define Cluster By 这 4 个选框，对两种价格的统计量都采用默认的 Mean（均值）。单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-72 所示。

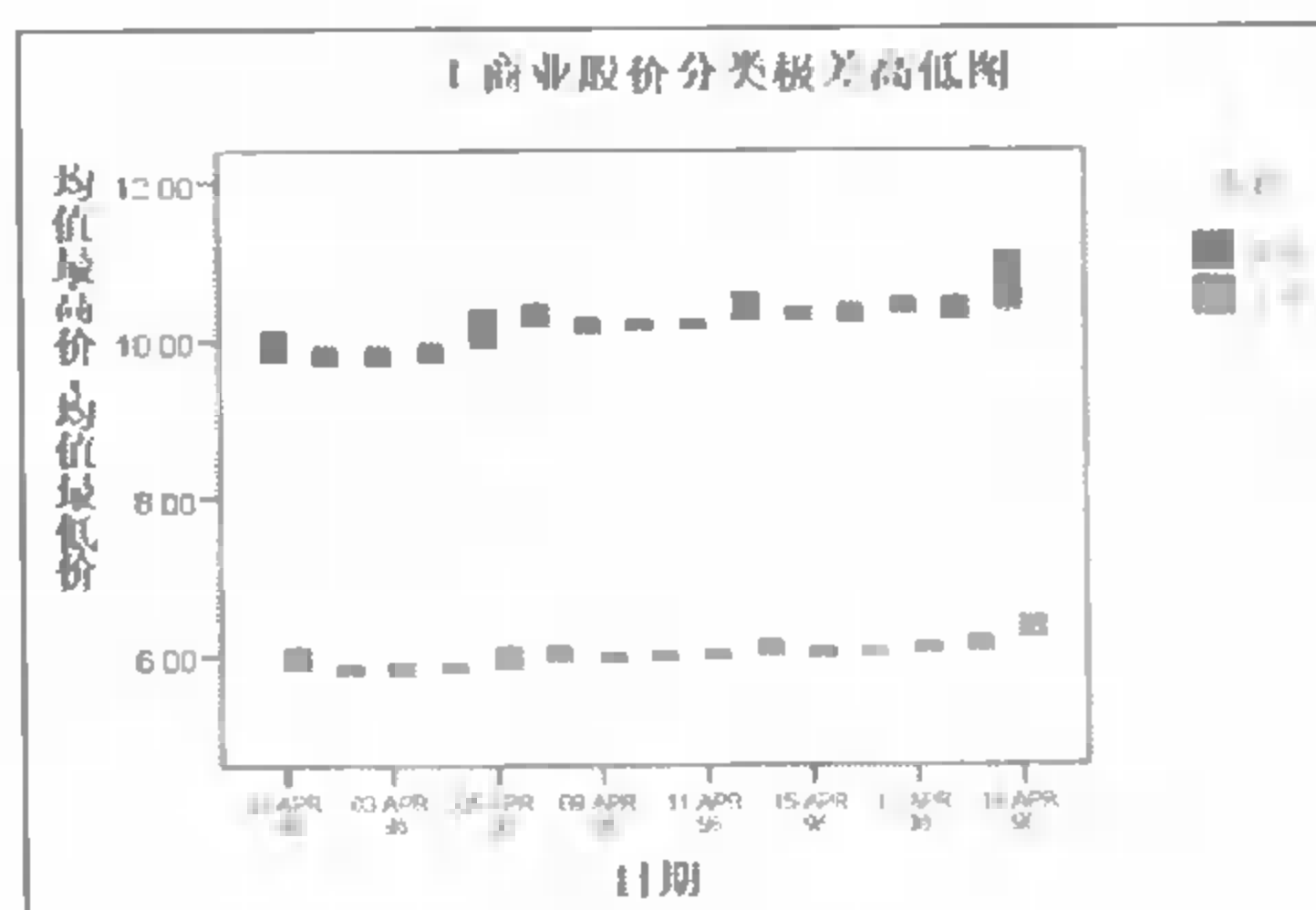


图 19-72 分类极差高低图的输出

## 2. 对变量分组作图 (Summaries of Separate Variables)

(1) 简单三值高低图。打开文件“北京地区股票价格.sav”，作关于北京地区股票价格的简单三值高低图。

在图 19-66 中单击选中 Simple high-low-close 图标, 单击选中 Summaries of Separate Variables 单选框; 单击 Define 按钮进入作图的设置界面, 它与图 19-67 非常相似。

从变量列表里把王府井的最高价 (wfj\_hi)、最低价 (wfj\_lo)、收盘价 (wfj\_cl) 和日期 (date) 这 4 个变量分别选入 High、Low、Close 和 Category Axis 这 4 个选框, 对 3 种价格的统计量都采用默认的 Mean (均值)。单击 OK 按钮运行, SPSS Viewer 窗口的输出图形如图 19-73 所示。

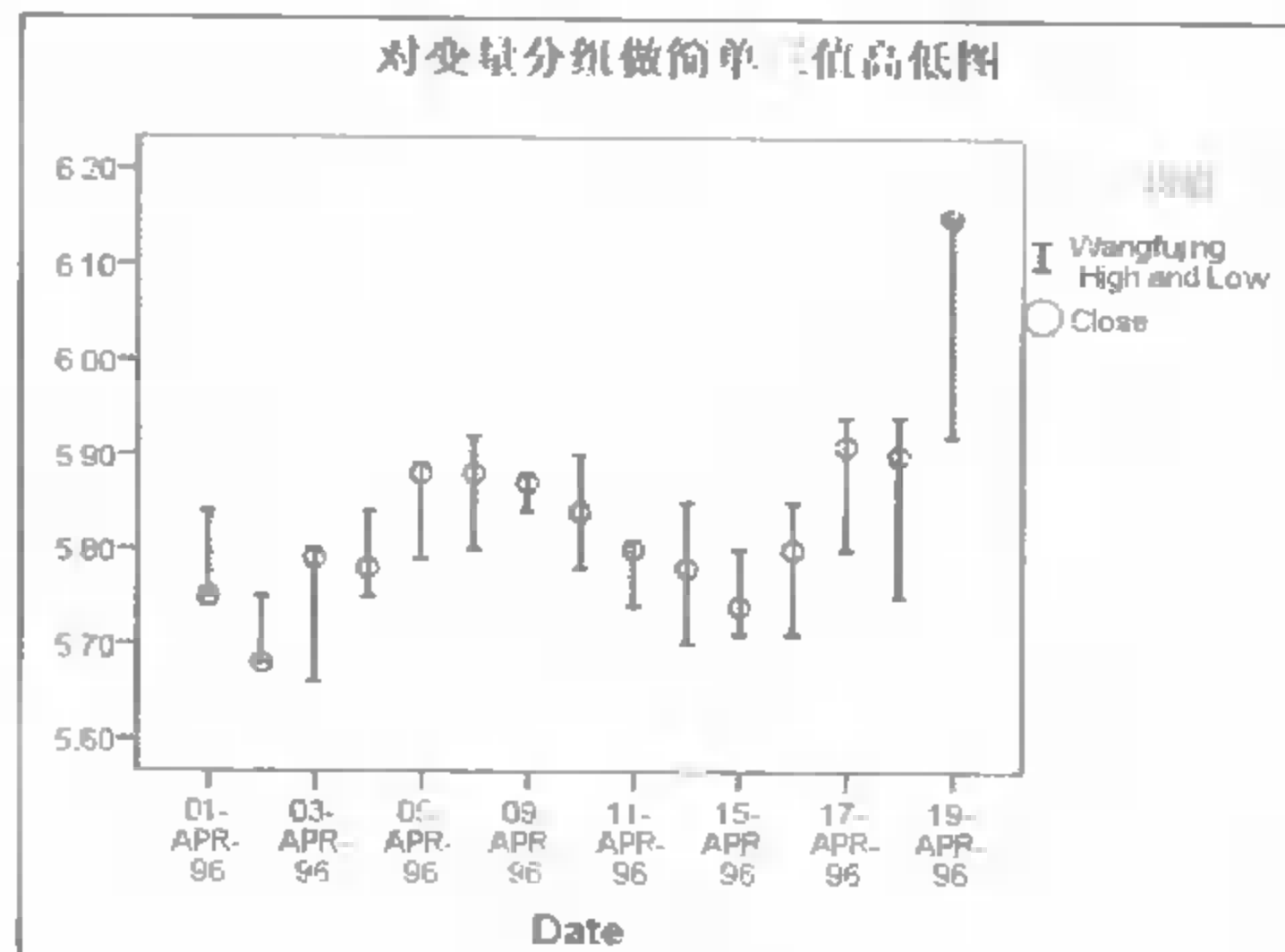


图 19-73 简单三值高低图的输出

简单极差高低图和面积差高低图的设置方法与输出结果, 均与简单三值高低图相似。

(2) 分类极差高低图。打开文件“北京地区股票价格.sav”, 作关于北京地区股票价格的分类极差高低图。

在图 19-66 中单击选中 Clustered range bar 图标, 单击选中 Summaries of Separate Variables 单选框; 单击 Define 按钮进入作图的设置界面, 它与图 19-67 非常相似。

从变量列表里把王府井的最高价 (wfj\_hi)、最低价 (wfj\_lo) 分别选入 1st、2nd 选框; 单击 Next 按钮清空 1st、2nd 选框, 设置下一对代表高低值的变量, 从变量列表里把北人的最高价 (br\_hi)、最低价 (br\_lo) 分别选入 1st、2nd 选框; 从变量列表里把日期 (date) 变量选入 Category Axis 选框。单击 OK 按钮运行, SPSS Viewer 窗口的输出图形如图 19-74 所示。

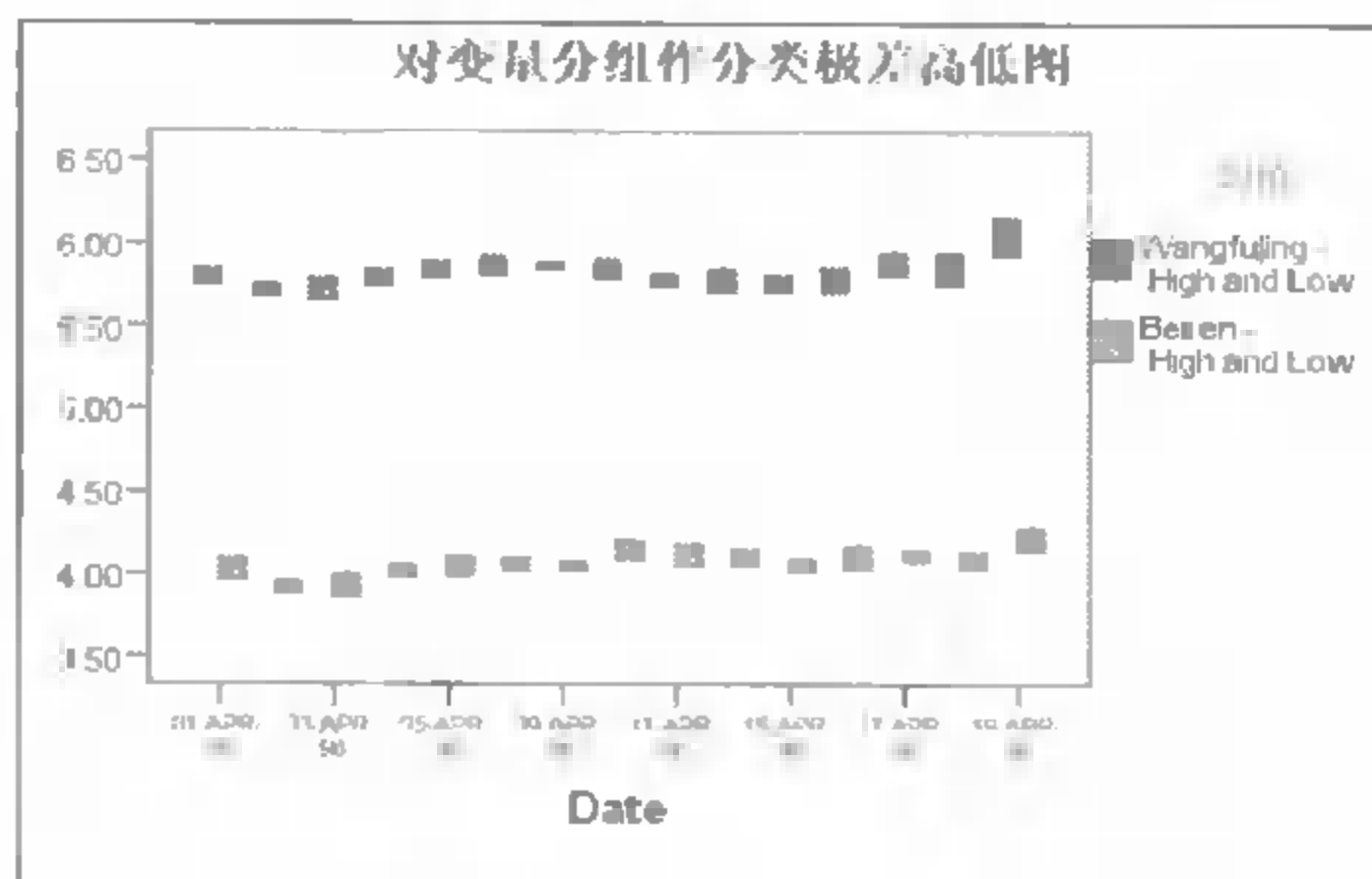


图 19-74 分类极差高低图的输出

分类三值高低图的设置方法与输出结果, 均与分类极差高低图相似。

### 3. 对观测记录作图 (Values of Individual Cases)

(1) 面积差高低图。打开文件“北京地区股票价格.sav”, 作关于北京地区股票价格的面积差高低图。

在图 19-66 中, 单击选中 Difference Area 图标, 单击选中 Values of Individual Cases 单选框; 单击 Define 按钮进入作面积差高低图的设置界面, 它与图 19-67 非常相似。



单击选中 Variable 单选框；从变量列表里把王府井的收盘价 (wfj\_cl)、北旅的收盘价 (bl\_cl)、日期 (date) 分别选入 1st、2nd、Variable 选框。单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-75 所示。

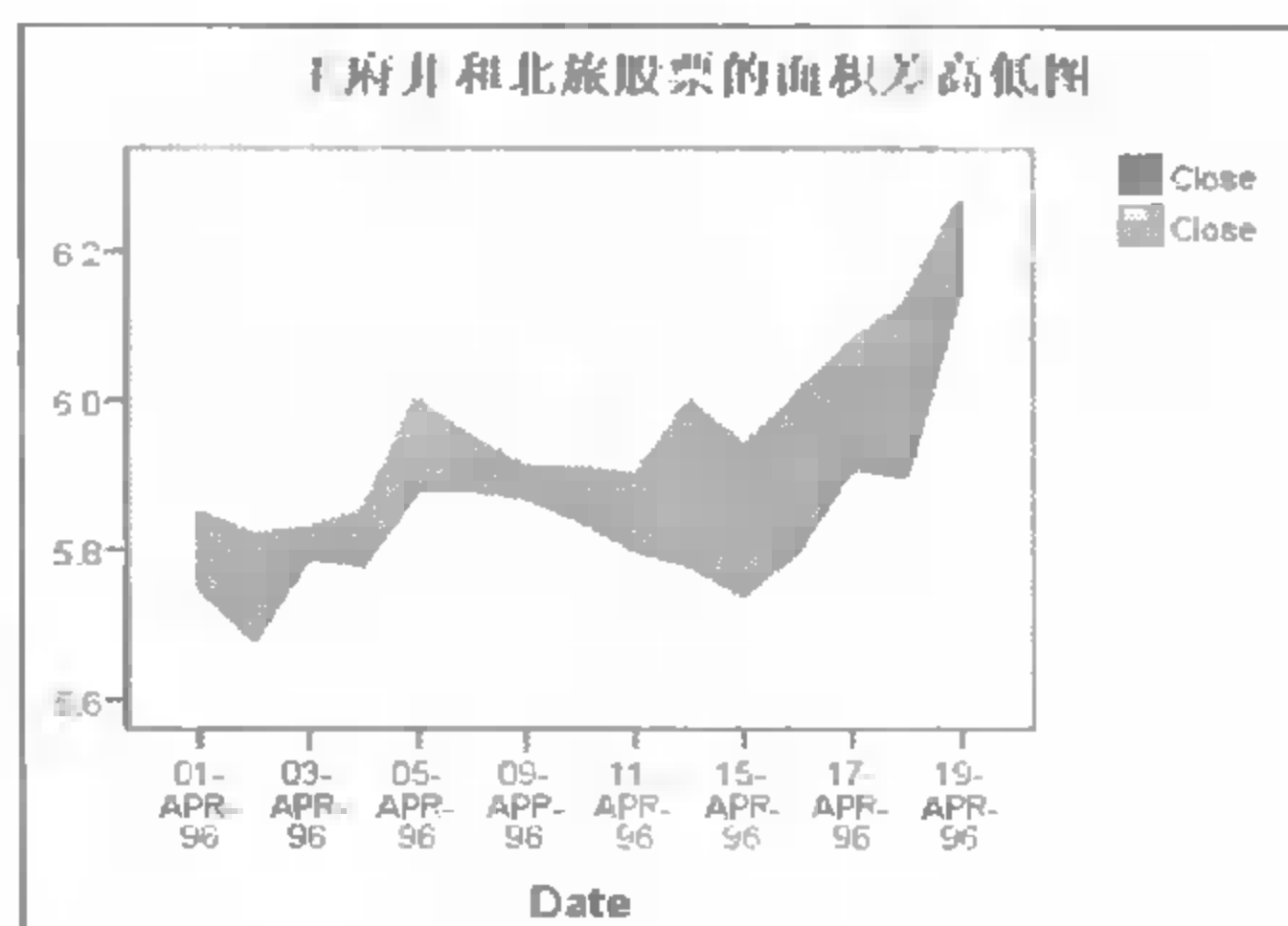


图 19-75 面积差高低图的输出

简单三值高低图和简单极差高低图的设置方法与输出结果，均与面积差高低图相似。

(2) 分类三值高低图。打开文件“北京地区股票价格.sav”，作关于北京地区股票价格的分类三值高低图。

在图 19-66 中单击选中 Clustered high-low-close 图标，单击选中 Values of Individual Cases 单选框；单击 Define 按钮进入作分类三值高低图的设置界面，它与图 19-67 非常相似。

从变量列表里把王府井的最高价 (wfj\_hi)、最低价 (wfj\_lo)、收盘价 (wfj\_cl) 分别选入 High、Low、Close 选框；单击 Next 按钮清空 High、Low、Close 选框，设置下一组代表高低值的变量，从变量列表里把北旅的最高价 (bl\_hi)、最低价 (bl\_lo)、收盘价 (bl\_cl) 分别选入 High、Low、Close 选框；单击选中 Variable 单选框，从变量列表里把日期 (date) 选入 Variable 选框。单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-76 所示。

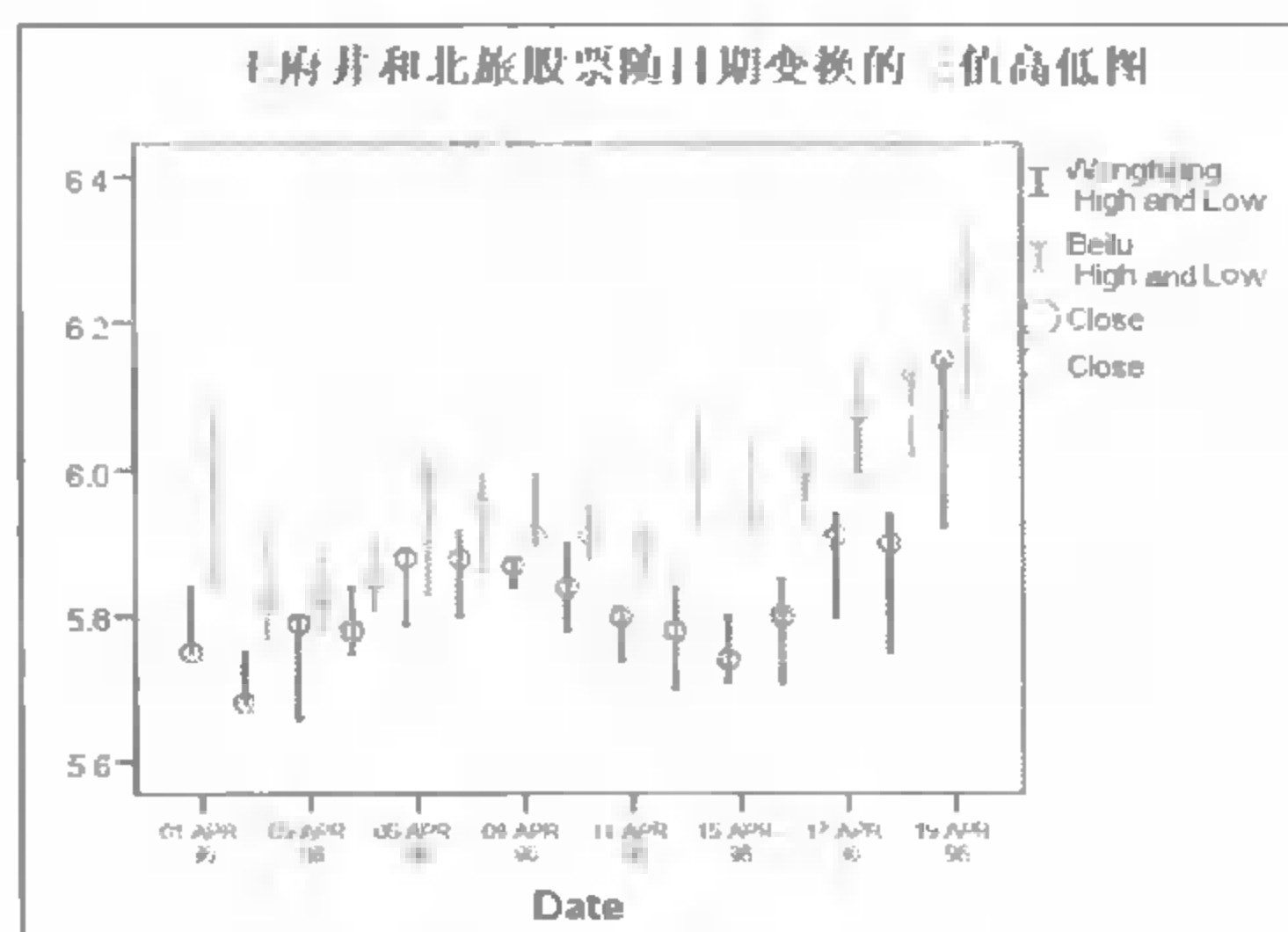


图 19-76 分类三值高低图的输出

分类极差高低图的设置方法与输出结果，均与分类三值高低图相似。

## 19.7 帕累托图

帕累托图用直条长短表现不同分组的绝对数大小，同时用线段的逐渐上升趋势表现不同组成部分的百分比逐步接近 100.00% 的过程。它又称为排列图或主次因素图，可用于区分影响某个现象的主要因素和次要因素，这使得它的应用非常广泛。

### 19.7.1 数据和问题描述

本节使用帕累托图来描绘农村生活的各种支出所占比重的变化规律。所用数据文件为“80 年代农村生活经济指标.sav”，数据格式如图 19-77 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	region	String	8		地区	None	None	8	Left	Nominal
2	food	Numeric	3	2	食品	None	None	6	Right	Scale
3	clothes	Numeric	3	2	衣着	None	None	6	Right	Scale
4	fuel	Numeric	3	2	燃料	None	None	6	Right	Scale
5	house	Numeric	3	2	住房	None	None	6	Right	Scale
6	danece	Numeric	3	2	生活用品	None	None	6	Right	Scale
7	culture	Numeric	3	2	文化支出	None	None	6	Right	Scale

图 19-77 农村生活经济指标数据

### 19.7.2 用对话框创建帕累托图

依次单击菜单“Analyze→Quality Control→Pareto Charts...”打开利用对话框创建帕累托图的类型选择界面，如图 19-78 所示。Data in Chart Are 子设置栏用于指定作图的数据对象，关于它们的具体示例如图 19-29 所示。

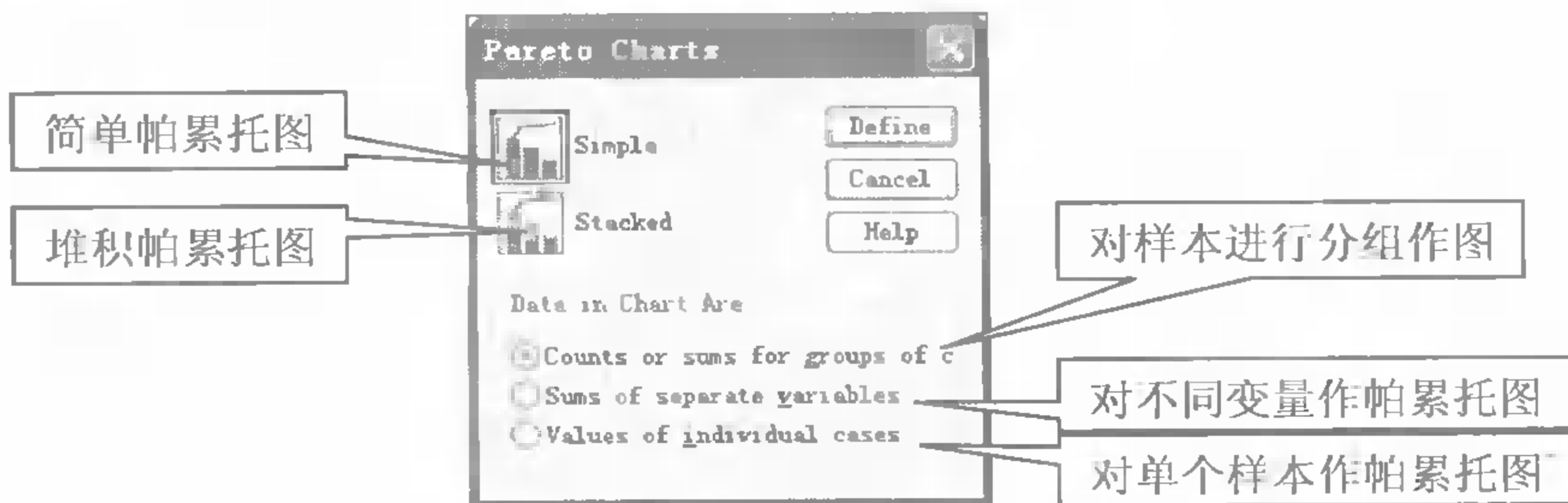


图 19-78 创建帕累托图的类型选择对话框

#### 1. 简单帕累托图

在图 19-78 中单击选中 Simple 图标，单击选中 Sums of Separate Variables 单选框；单击 Define 按钮进入作简单帕累托图的设置界面，如图 19-79 所示。

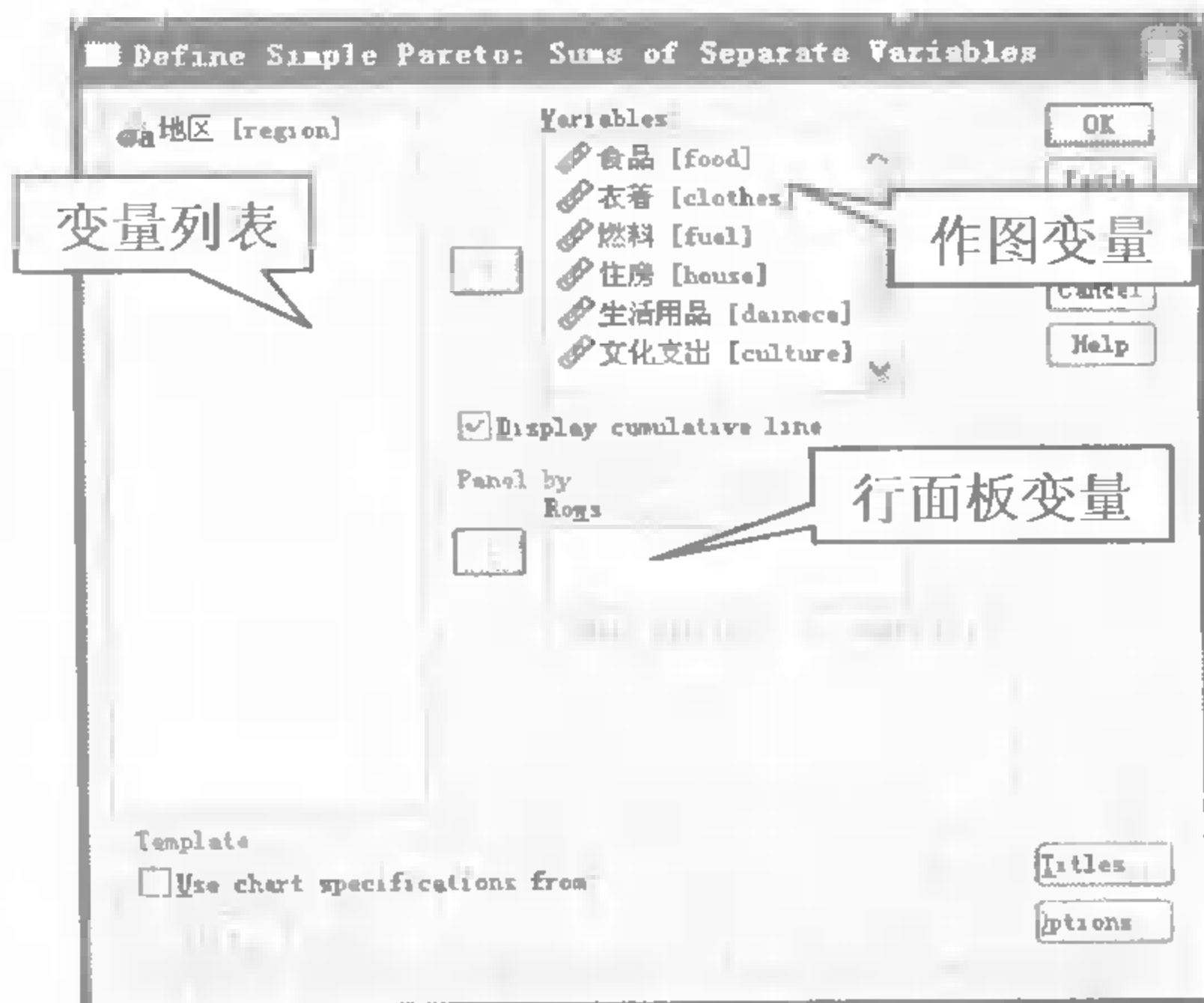



图 19-79 简单帕累托图的设置

在图 19-79 里,在变量列表中选中从食品至文化支出的 6 个变量,单击从上至下第一个  按钮,将其作为作图变量选入 Variables 列表框;默认 Display 复选框呈勾选状态。单击 OK 按钮运行,SPSS Viewer 窗口的输出图形如图 19-80 所示。

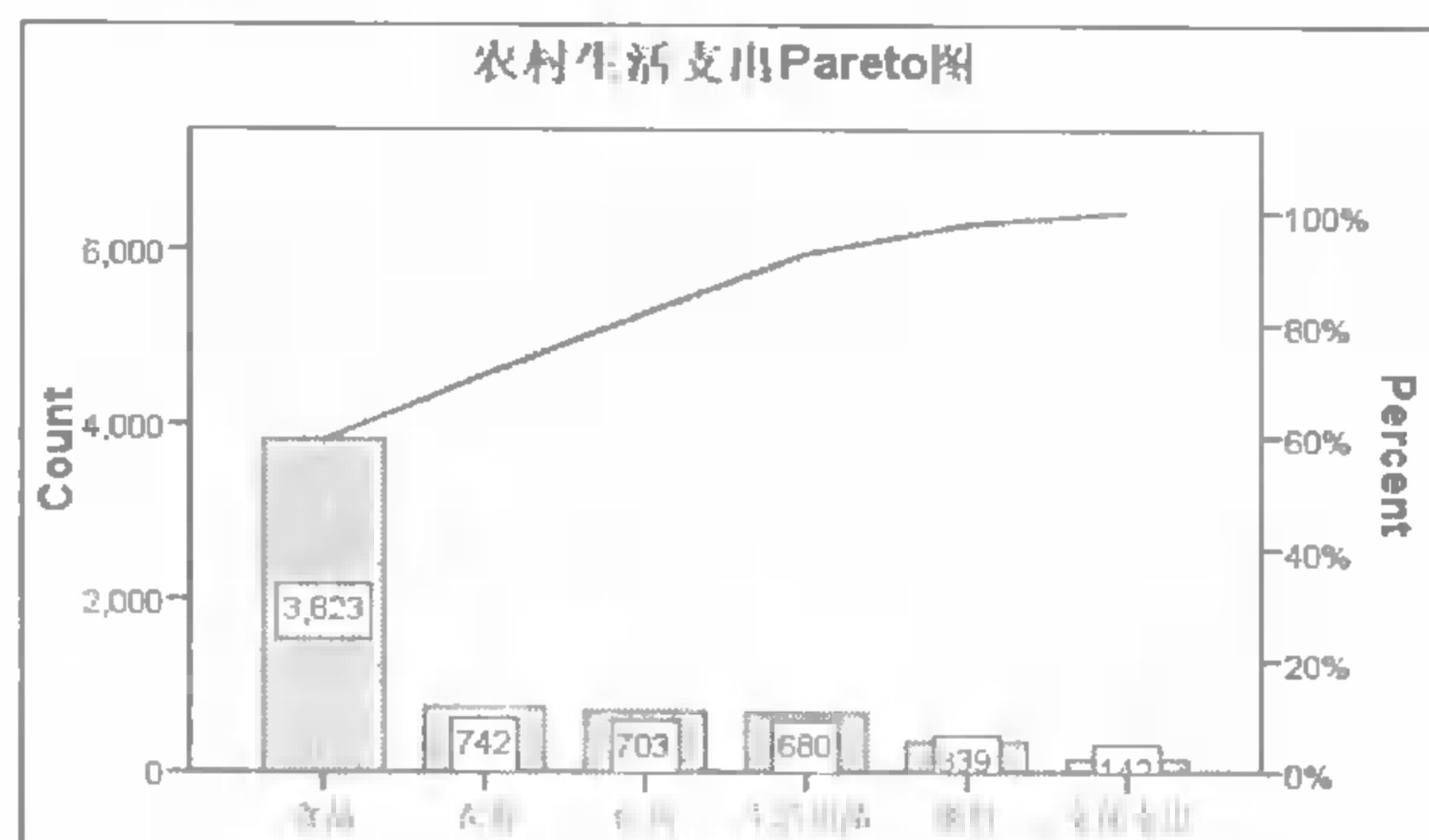




图 19-80 简单帕累托图的输出

## 2. 堆积帕累托图

在图 19-78 中单击选中 Stacked 图标,单击选中 Sums of Separate Variables 单选框;单击 Define 按钮进入作堆积帕累托图的设置界面,它与图 19-79 非常相似。

在变量列表中选中从食品至文化支出的 6 个变量,单击从上至下第一个  按钮,将其作为作图变量选入 Variables 列表框;在变量列表中单击选中地区变量,单击从上至下第二个  按钮,将其作为 X 轴分类变量选入 Category Axis 选框;默认 Display 复选框呈勾选状态。单击 OK 按钮运行,SPSS Viewer 窗口的输出图形如图 19-81 所示。

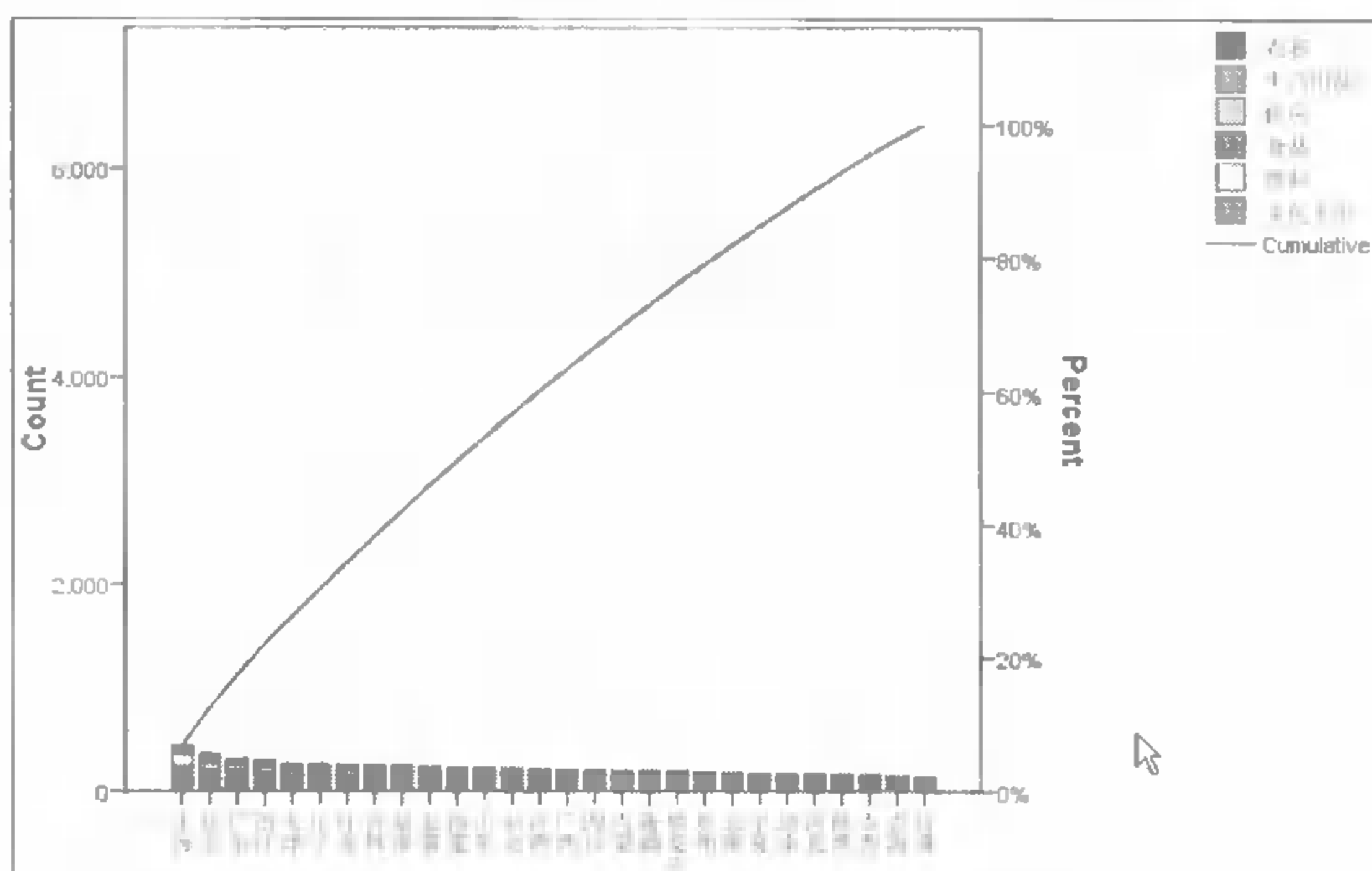


图 19-81 堆积帕累托图的输出

## 19.8 控制图

控制图是进行质量控制的常用工具,可用于引起用户对工作过程中发生的一些变化趋势的注意,以便分析原因、采取解决对策。控制图始于对产品质量的控制,现已推广到了生产领域以外的许多方面,例如经济学、医学等。

19.8.1 数据和问题描述

某研究人员收集了关于一种电解工序的效率数据，本节先利用这部分电解数据作控制图，观察有关效率的变化情况。所用数据文件“电解工序数据（1、2、3）.sav”，数据格式如图 19-82 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	date	Date	8	0	日期	None	None	14	Right	Scale
2	no	Numeric	2	0	班次	{1 上午}	None	3	Right	Scale
3	cat	Numeric	3	0	电解效率	None	None	5	Right	Ordinal
4	eec	Numeric	5	1				8	Right	Scale

“电解工序数据 1.sav”

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	no	Numeric	8	0	日期	None	None	8	Right	Scale
2	value	Numeric	8	1	电解效率	None	None	8	Right	Nominal

“电解工序数据 2.sav”

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	case	Numeric	8	0	样本分组	None	None	8	Right	Scale
2	products	String	8		合格与否	{no, 不合格}	None	8	Left	Nominal

“电解工序数据 3.sav”

图 19-82 电解工序数据的格式

19.8.2 用对话框创建控制图

依次单击菜单“Analyze→Quality Control→Control Charts...”打开利用对话框创建控制图的类型选择界面，如图 19-83 所示。

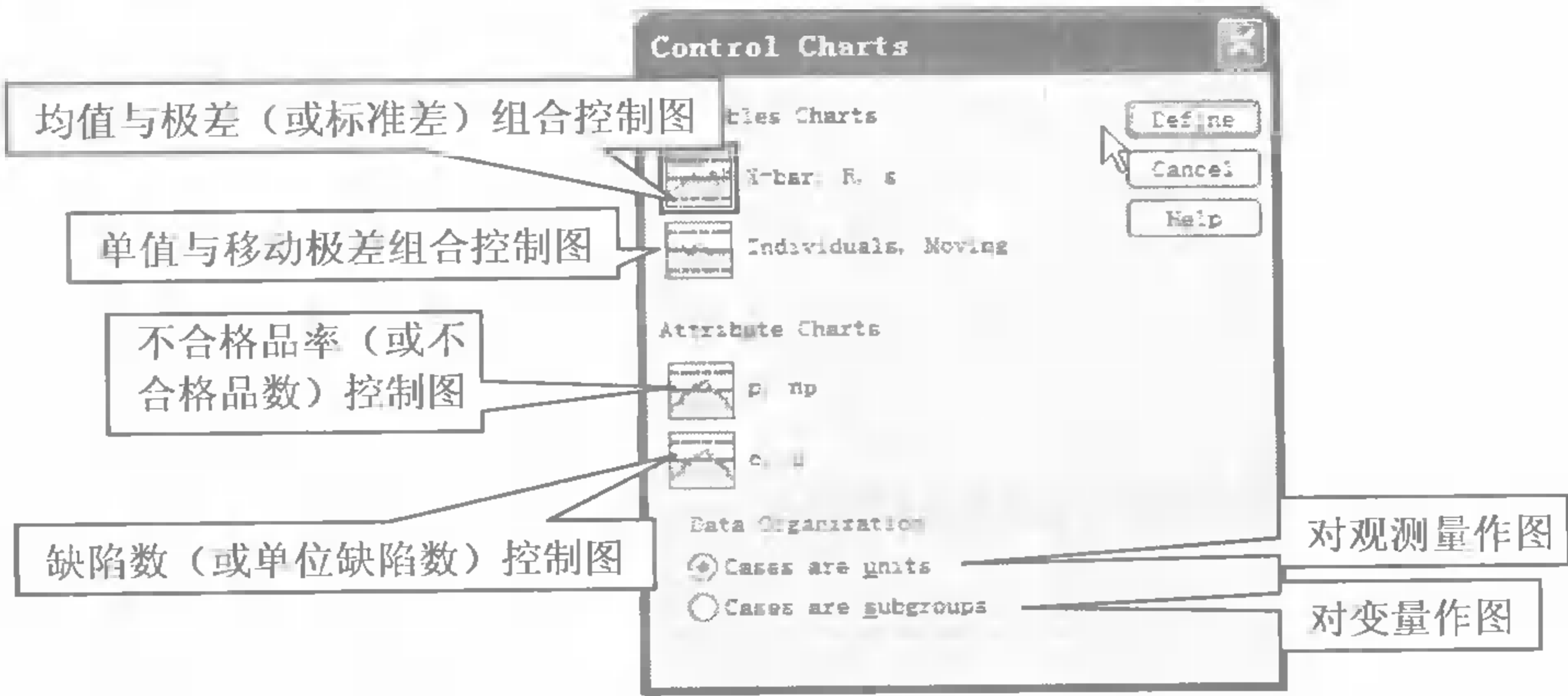


图 19-83 控制图选择界面

1. 对观测量作均值与极差（或标准差）组合控制图

打开文件“电解工序数据 1.sav”。在图 19-83 中单击选中“X-Bar,R,s”图标，单击选中 Cases are units 单选框；单击 Define 按钮进入作图界面，如图 19-84 所示。



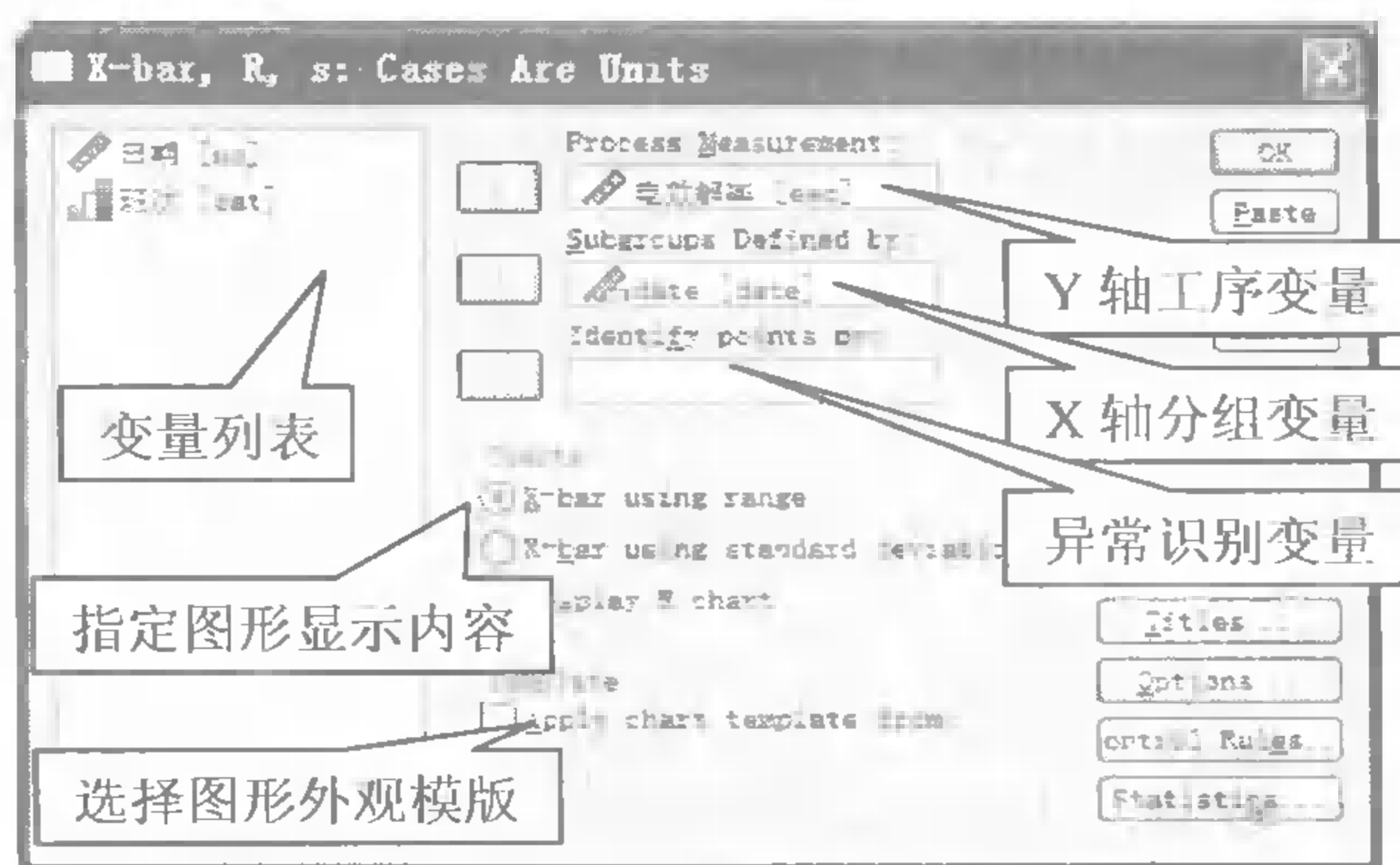


图 19-84 对观测量作控制图的主设置界面

(1) 指定作图变量。在变量列表中单击选中电解效率变量，单击从上至下第一个 ☐ 按钮，将其作为工序变量选入 Process Measurement 选框；在变量列表中单击选中 date（日期）变量，单击从上至下第二个 ☐ 按钮，将其作为分组变量选入 Subgroups 选框。

对其他选项的含义解释如下。

- X-Bar using range 单选项，表示输出均值 - 极值图。
- X-Bar using standard deviation 单选项，表示输出均值 - 标准差图。
- Display R (s) Chart 复选框，表示显示极值或标准差本身的控制图。
- Template 栏，指定图形模版，勾选 Apply 复选框后，单击 File 按钮指定文件路径。

(2) Options 选项设置。在图 19-84 中单击 Options 按钮，弹出如图 19-85 所示的子设置对话框。

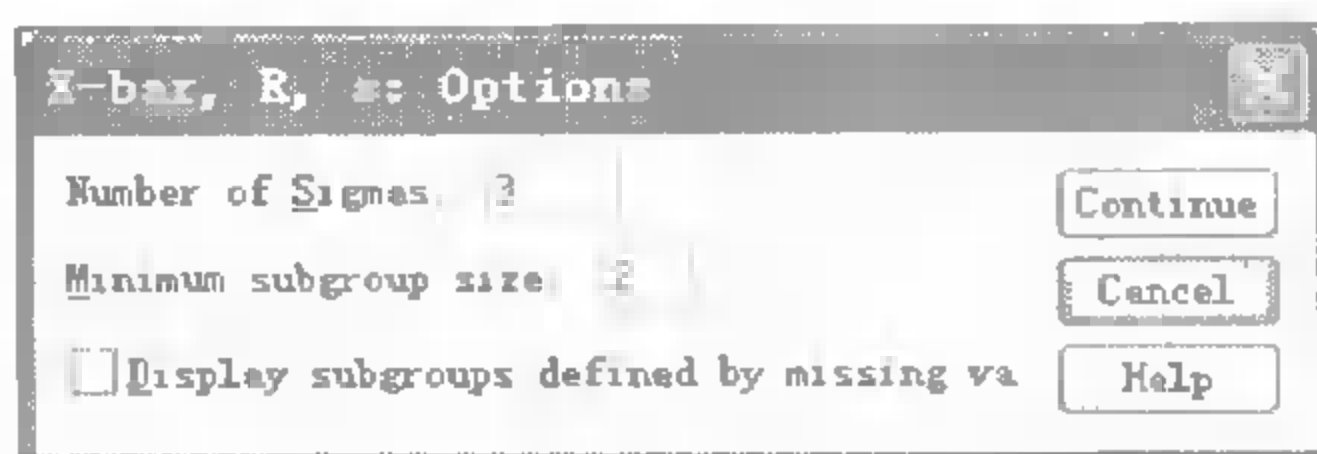


图 19-85 Options 参数设置

- Number 输入框，指定偏离均值的标准差的范围，默认为 3 倍标准差。
- Minimum 输入框，指定每个分组中最少需要的样本数，默认值为 2。
- Display 复选框，勾选它表示把缺失值作为一个单独的分组显示于图中。

(3) 统计量设置。在图 19-84 中单击 Statistics 按钮，弹出如图 19-86 所示的子设置对话框。

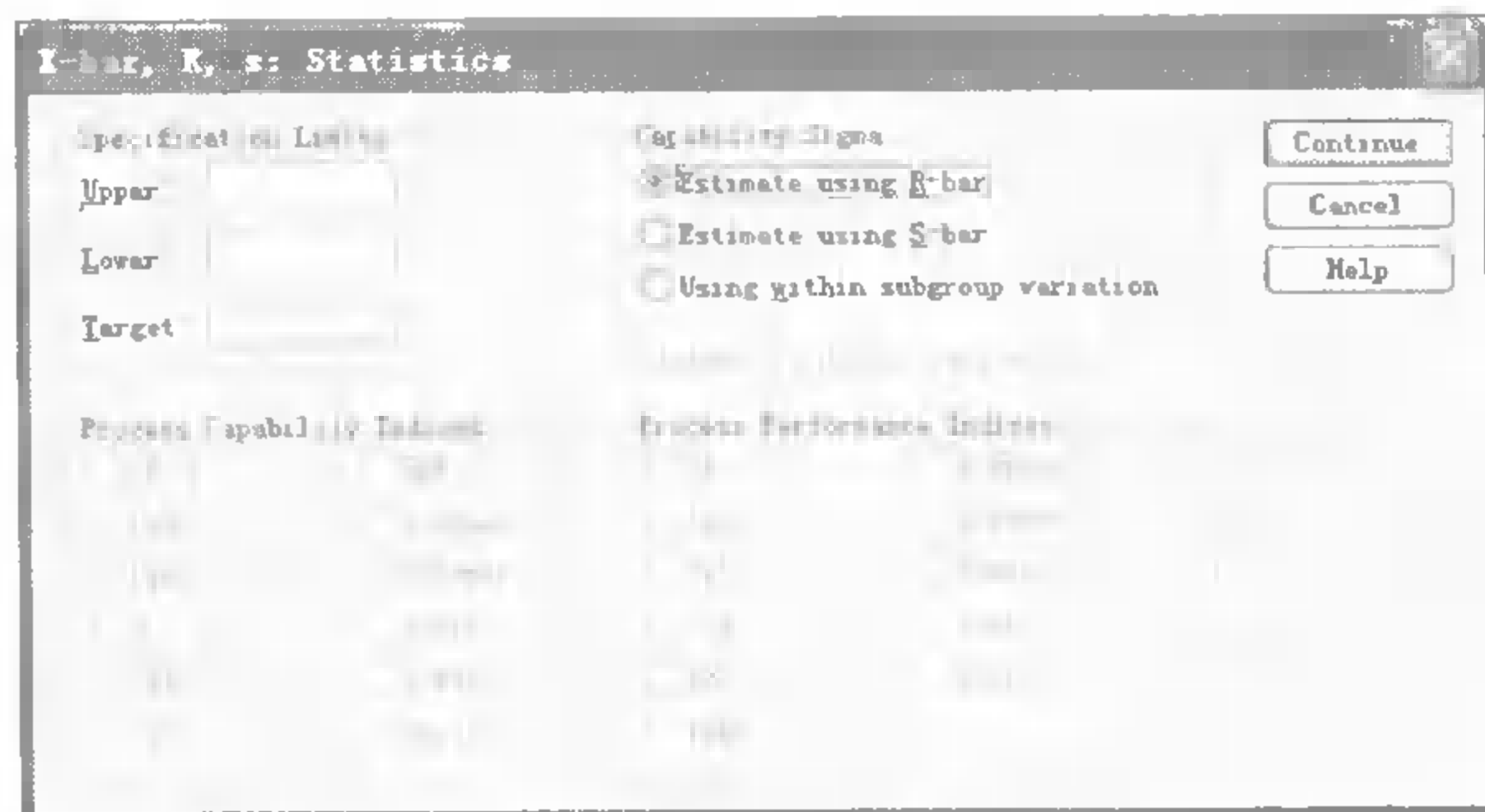


图 19-86 控制图的统计量设置

- Specification Limits 栏, 指定上限(Upper)、下限(Lower)和一个固定目标值(Target)。
- Capability Sigma 栏, 指定计算 capability indices 时的标准差范围。
- Actual % outside specification 复选框, 指定超出控制范围的样本比例。
- Process Capability Indices 栏, 指定衡量工序能力的统计量, 它们大都是基于 capability sigma 计算得到的。
- Process Performance Indices 栏, 指定衡量工序性能的统计量, 它们大都是基于工序的标准差计算得到的。

(4) 控制规则的设置。在图 19-84 中单击 Control Rules 按钮, 弹出如图 19-87 所示的子设置对话框。如果某个点违背了此处指定的规则, 它将在图中用区别于正常点的形状和颜色表示。

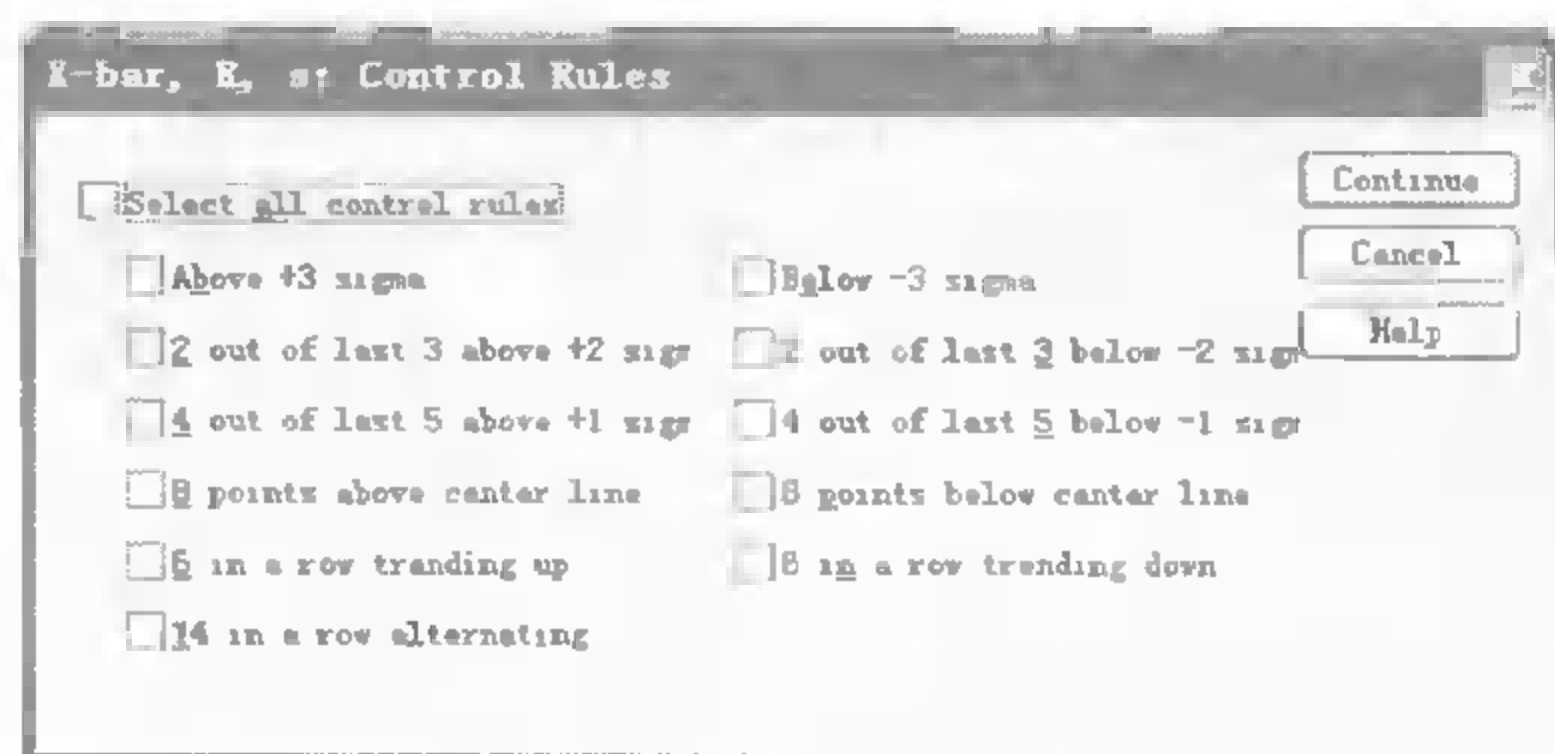


图 19-87 控制图的控制规则设置

(5) 输出图形。在图 19-84 中, 单击 OK 按钮运行, SPSS Viewer 窗口的输出图形如图 19-88 所示。它描绘了随着日期的变化, 电解效率均值 (Mean) 和全距 (Range) 的变动范围。

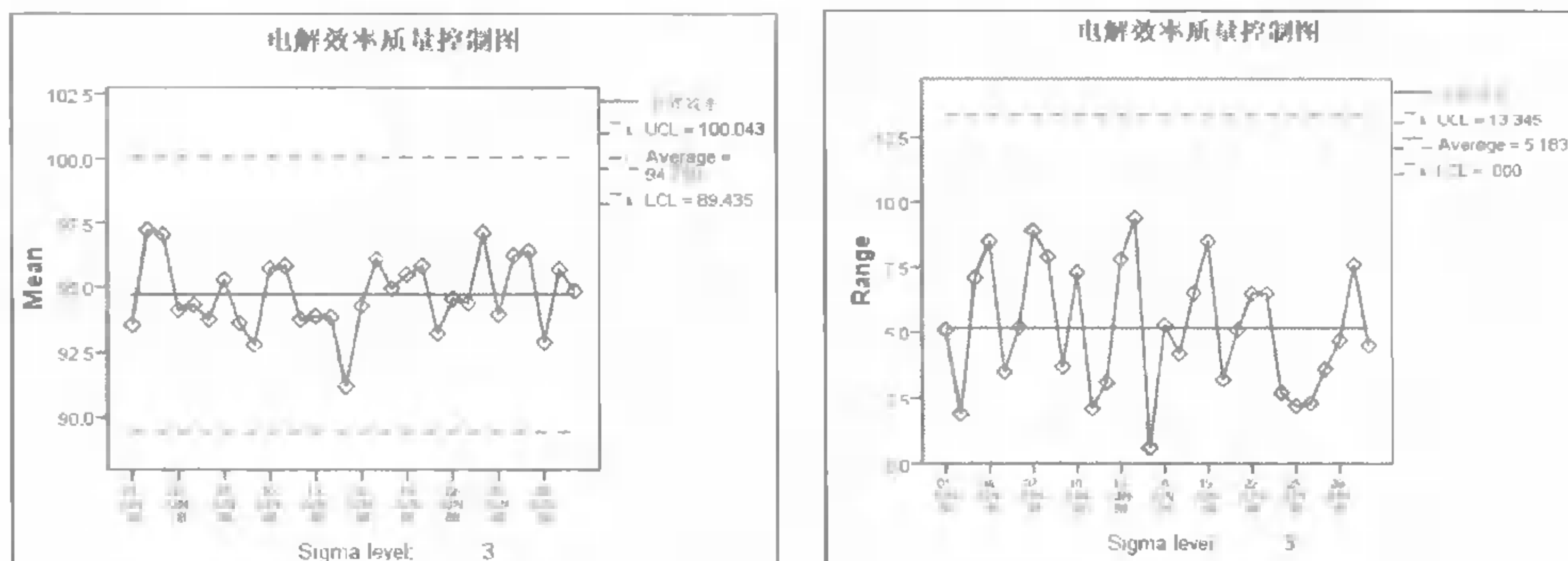




图 19-88 电解效率的均值-极差控制图

## 2. 对观测量作单值与移动极差组合控制图

打开文件“电解工序数据 2.sav”。在图 19-83 中单击选中“Individuals,Moving”图标, 单击选中 Cases are units 单选框; 单击 Define 按钮进入作图界面, 它与图 19-84 非常相似。

在变量列表中单击选中电解效率变量, 单击从上至下第一个  按钮, 将其作为工序变量选入 Process Measurement 选框; 在变量列表中单击选中 date (日期) 变量, 单击从上至下第二个  按钮, 将其作为分组变量选入 Subgroups 选框; 默认选中 Individuals and moving 单选框。

单击 OK 按钮运行, SPSS Viewer 窗口的输出图形如图 19-89 所示, 它描绘了随着日期的

变化, 电解效率、移动极差的变动趋势和变动范围。所谓移动极差, 就是由后一项与前一项的差所组成的新序列, 它反映了前后变化的波动大小。

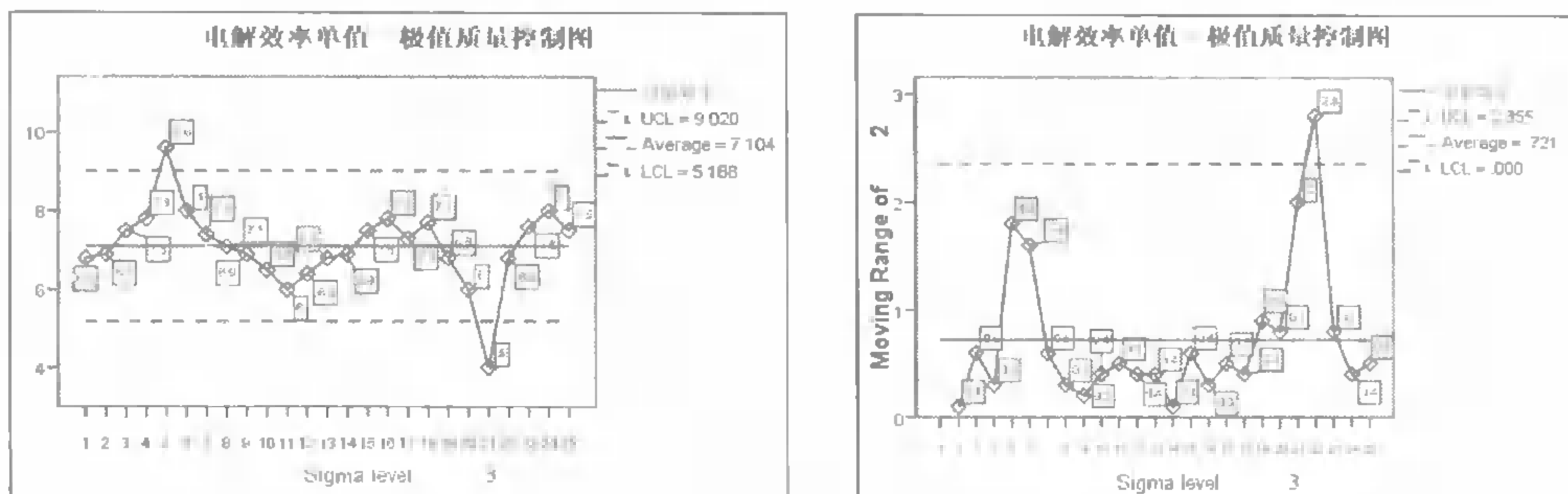


图 19-89 电解效率的单值 - 极值控制图

### 3. 对观测量作不合格品控制图

打开文件“电解工序数据 3.sav”。在图 19-83 中单击选中“p,np”图标, 单击选中 Cases are units 单选框; 单击 Define 按钮进入作图界面, 如图 19-90 所示。

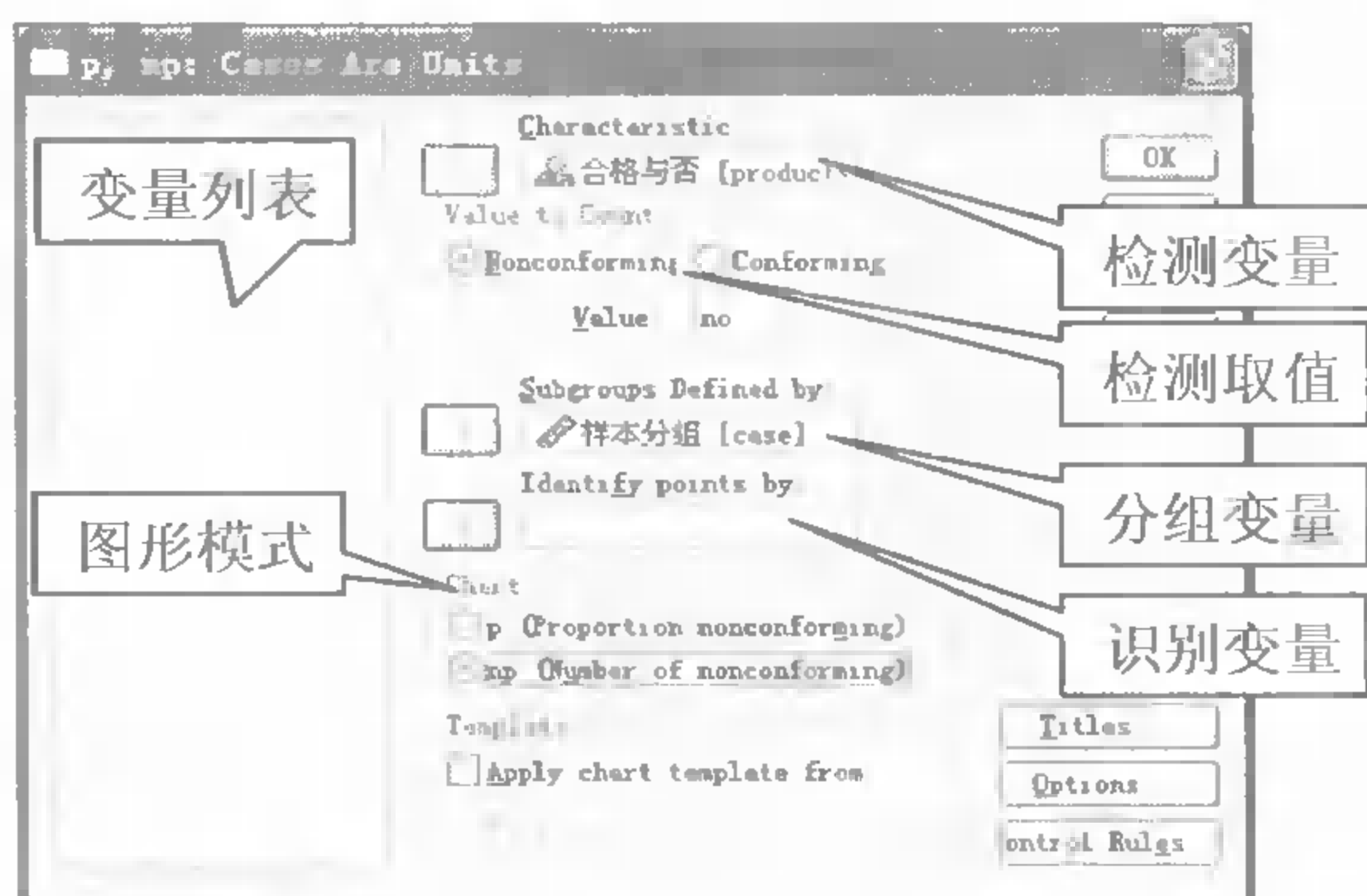




图 19-90 不合格品控制图的设置

(1) 指定作图变量。在变量列表中单击选中合格与否变量, 单击从上至下第一个  按钮, 将其作为检验变量选入 Characteristic 选框; 在变量列表中单击选中样本分组变量, 单击从上至下第二个  按钮, 将其作为分组变量选入 Subgroups Defined by 选框; 单击选中 Nonconforming 选项, 在 Value 后输入“no”作为对不合格品的取值; 单击选中 np 单选框。

对其他选项的含义解释如下。

- Characteristic 选框, 用于选入待检验的变量, 可以是数值型的或字符型的。
- Value to Count 栏, 指定变量的计数方式, 有两个可选项: Nonconforming, 统计不合格产品数; Conforming, 统计合格产品数。Value 输入框, 指定要统计的变量取值。
- Subgroups Defined by 选框, 用于选入分组变量。
- Chart 栏, 指定图形模式, 有两个可选项: p(Proportion of nonconforming) 表示作不合格品率控制图; np(Number of nonconforming) 表示作不合格品数控制图。

(2) 输出图形。在图 19-90 中, 单击 OK 按钮运行, SPSS Viewer 窗口的输出图形如图 19-91 所示。它描绘了随着样本分组的变化, 不合格品个数的变动范围。

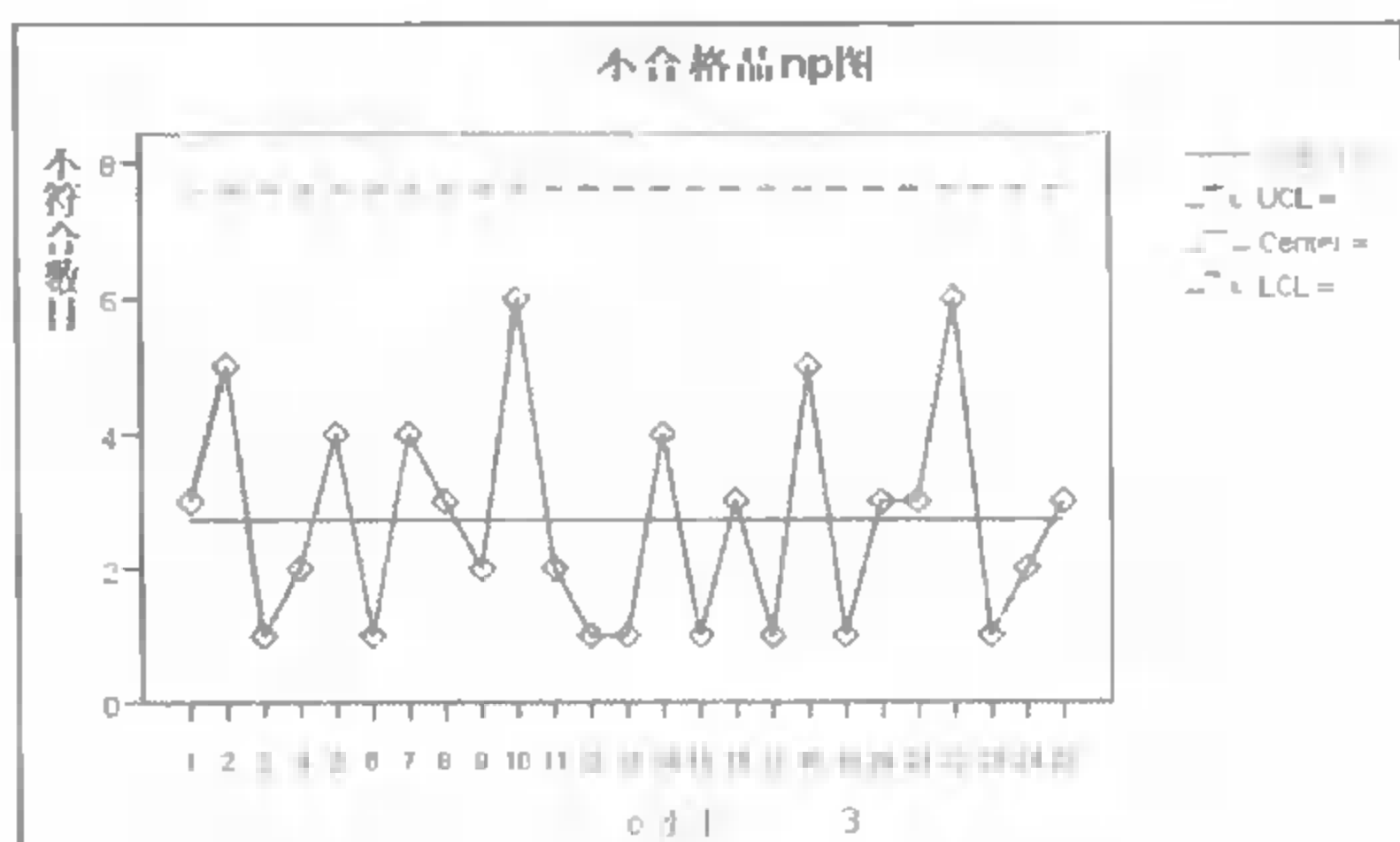


图 19-91 不合格品控制图的输出



#### 4. 对观测量作缺陷控制图

下面利用某种手术缺陷次数的样本作缺陷控制图，所用数据文件为“手术缺陷记录数据.sav”，数据格式如图 19-92 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	week	Numeric	8	0	时间（星期）	None	None	8	Right	Scale
2	aes	Numeric	8	0	手术缺陷次数	None	None	8	Right	Scale

图 19-92 手术缺陷记录的数据格式

在图 19-83 中单击选中“c,u”图标，单击选中 Cases are units 单选框；单击 Define 按钮进入作图界面，它与图 19-90 非常相似。

在变量列表中单击选中手术缺陷次数变量，单击从上至下第一个  按钮，将其作为检验变量选入 Characteristic 选框；在变量列表中单击选中时间（星期）变量，单击从上至下第二个  按钮，将其作为分组变量选入 Subgroups Defined by 选框；单击选中 u 单选框，表示作单位缺陷数的控制图。当不同分组中的记录个数不同时，建议用 c、u 图取代 p、np 图。

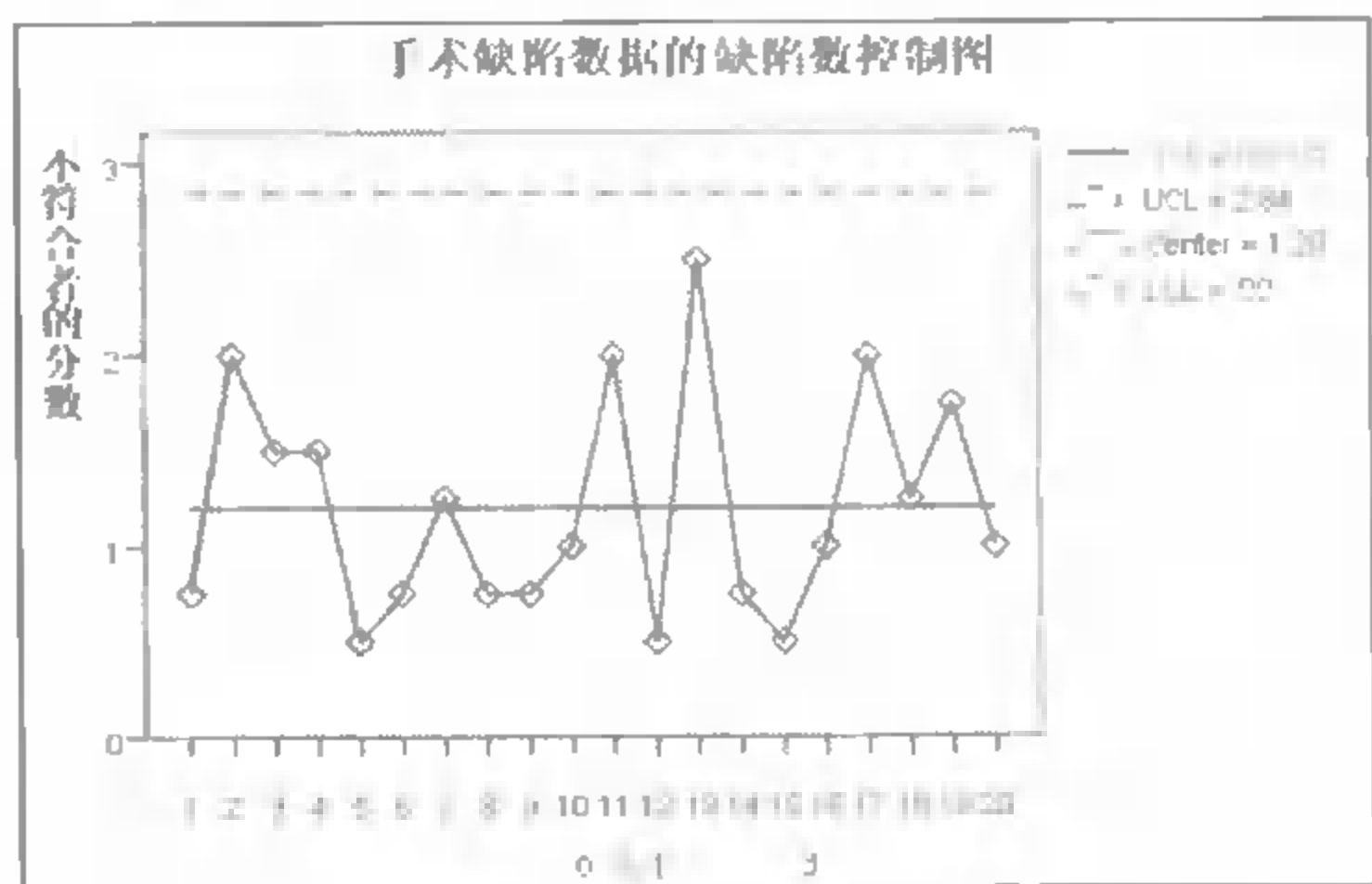


图 19-93 手术缺陷数控制图

单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-93 所示。它描绘了随着时间的变化，平均手术缺陷次数的变动范围。

#### 5. 对变量作均值与极差（或标准差）组合控制图



下面利用对某种钢板厚度的测量数据作均值与极差组合控制图，所用数据文件为“钢板厚度测量数据.sav”，数据格式如图 19-94 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	case	Numeric	5	0	序号	None	None	8	Right	Scale
2	t1	Numeric	4	2	测厚度度1	None	None	8	Right	Scale
3	t2	Numeric	4	2	测厚度度2	None	None	8	Right	Scale
4	t3	Numeric	4	2	测厚度度3	None	None	8	Right	Scale
5	t4	Numeric	4	2	测厚度度4	None	None	8	Right	Scale
6	t5	Numeric	5	2	测厚度度5	None	None	8	Right	Scale

图 19-94 五次测量的钢板厚度数据格式

在图 19-83 中单击选中“X-Bar,R,s”图标，单击选中 Cases are subgroups 单选框；单击 Define 按钮进入作图界面，它与图 19-84 非常相似。



在变量列表中选中测量厚度 1~5 的五个变量，单击从上至下第一个  按钮，将其作为统计变量选入 Samples 列表框；在变量列表中单击选中序号变量，单击从上至下第二个  按钮，将其作为分组变量选入 Subgroups Labeled by 选框。单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-95 所示，它描绘随着了序号（时间）的变化，钢板厚度均值（Mean）和全距（Range）的变动范围。

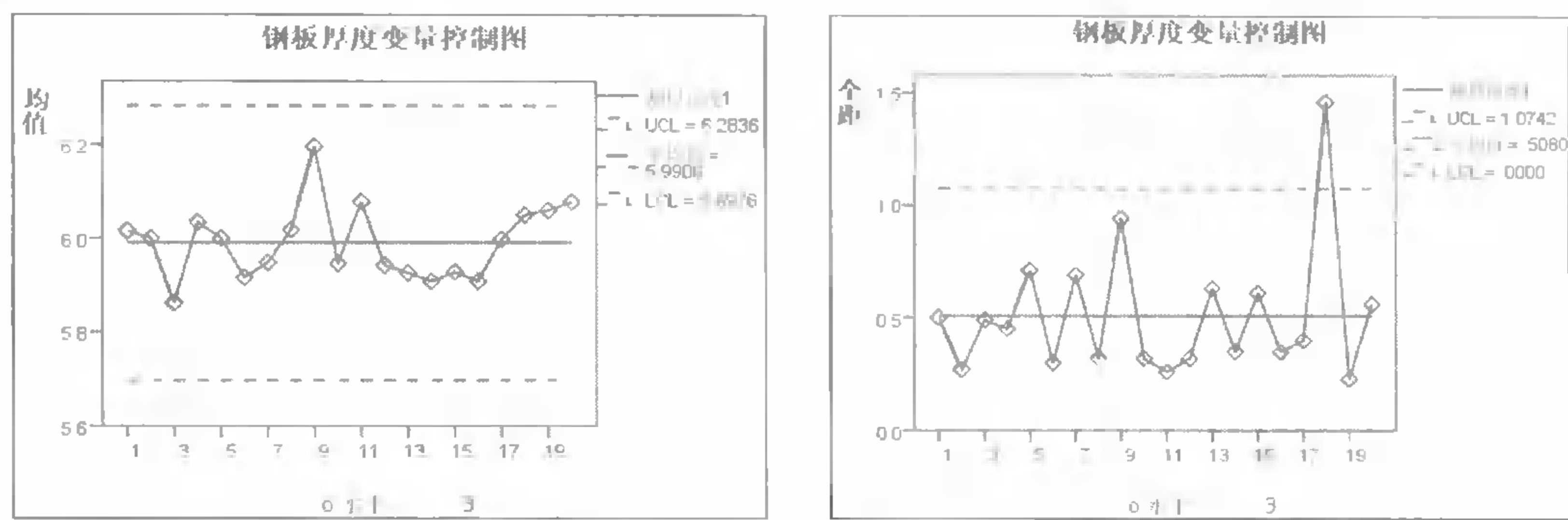


图 19-95 对变量作均值—极差控制图的输出




6. 对变量作不合格品控制图

下面利用对某种小螺丝质量的检测数据作不合格品控制图，所用数据文件为“小螺丝质量检测数据.sav”，数据格式如图 19-96 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	no	Numeric	2	0	检测序号	None	None	8	Right	Scale
2	sam	Numeric	5	0	样本量	None	None	8	Right	Scale
3	unq	Numeric	5	0	不合格数	None	None	8	Right	Nominal

图 19-96 小螺丝质量检测数据格式

在图 19-83 中单击选中“p,np”图标，单击选中 Cases are subgroups 单选框；单击 Define 按钮进入作图界面，如图 19-97 所示。

(1) 指定作图变量。在变量列表中单击选中不合格数变量，单击从上至下第一个  按钮，将其作为检测变量选入 Number Nonconforming 选框；在变量列表中单击选中检测序号变量，单击从上至下第二个  按钮，将其作为分组变量选入 Subgroups Labeled by 选框；单击选中 Variable 单选框，在变量列表单击选中样本量变量，单击从上至下第四个  按钮，将其选入 Variable 选框。

对各设置选项的含义解释如下。

- Number Nonconforming 选框，用于选入待检验的记录变量，必须是数值型的。
- Subgroups Defined by 选框，用于选入分组变量。
- Sample Size 栏，设置分组样本量的大小，有两个选择：Constant 输入框，指定一个代表所有分组样本大小的常数；Variable 选项，指定一个代表分组样本大小的变量。
- Chart 栏，指定图形模式，有两个可选项：p(Proportion of nonconforming)，表示作不合格品率控制图；np(Number of nonconforming)，表示作不合格品数控制图。

(2) 输出图形。在图 19-97 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-98 所示，它描绘了随着检测序号的变化，不合格小螺丝所占比例的变动趋势和变

动范围。

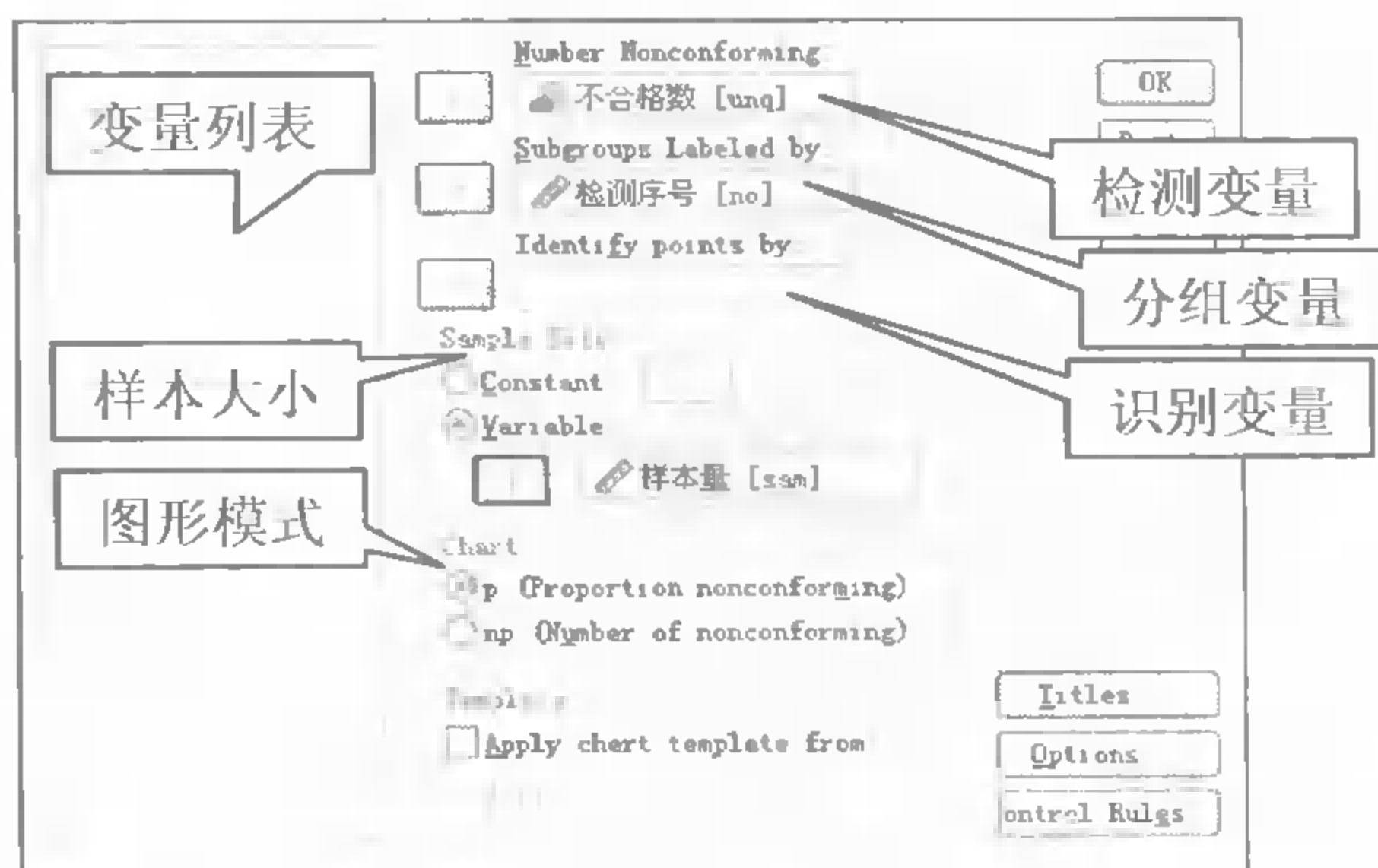


图 19-97 对变量作不合格品控制图的设置

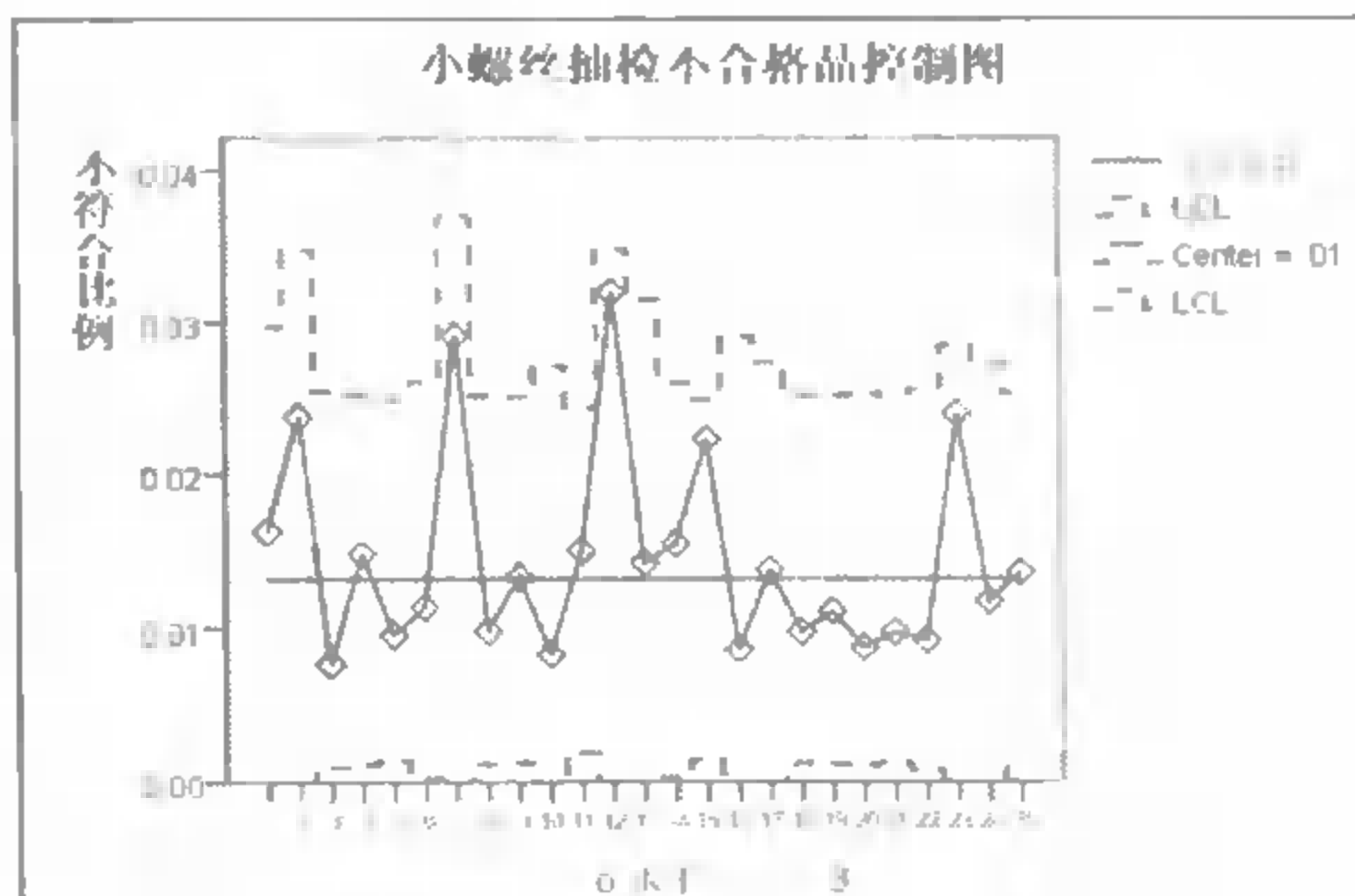


图 19-98 对变量作不合格品控制图的输出

## 19.9 箱图

箱图 (Boxplots) 又称为箱线图, 是一种用于描绘数据分布形式的统计图形。箱图包含了指定变量的最小值、1/4 分位数、中位数、3/4 分位数、最大值 5 个统计量, 它们在图形中从下至上依次显示。

### 19.9.1 数据和问题描述

本节利用箱图来描绘不同年龄 (10~13 岁)、不同性别的儿童的体重和身高特征。所用数据文件为“儿童身高体重数据.sav”, 数据格式如图 19-99 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	no	String	4		编号	None	None	4	Right	Nominal
2	gend	String	4		性别	0 男	None	4	Right	Nominal
3	age	Numeric	2	0	年龄	None	None	3	Right	Ordinal
4	high	Numeric	4	2	身高	None	None	8	Right	Scale
5	weight	Numeric	2	0	体重	None	None	6	Right	Scale


图 19-99 儿童身体特征数据格式

### 19.9.2 用图形构建器作箱图

依次单击菜单“Graphs→Chart Builder”打开图形构建器, 如图 19-100 所示。单击 Gallery 标签, 在 Choose from 列表单击选中 Boxplot, 就在其右侧显示预设的箱图图标。

#### 1. 简单箱图

下面先来作儿童体重按照年龄分布的简单箱图。

(1) 参数设置。在图 19-100 中双击预置图标  (Simple Boxplot) 后, 在图形预览区给出简单箱图的预览, 同时自动弹出元素属性设置面板; 把预置图标拖动至图形预览区, 可以达到相同的效果。

从变量列表中把年龄、体重两个变量分别拖动至预览区的 X-Axis、Y-Axis 两个虚线框中, 将其分别作为简单箱图的 X 坐标轴和 Y 坐标轴。

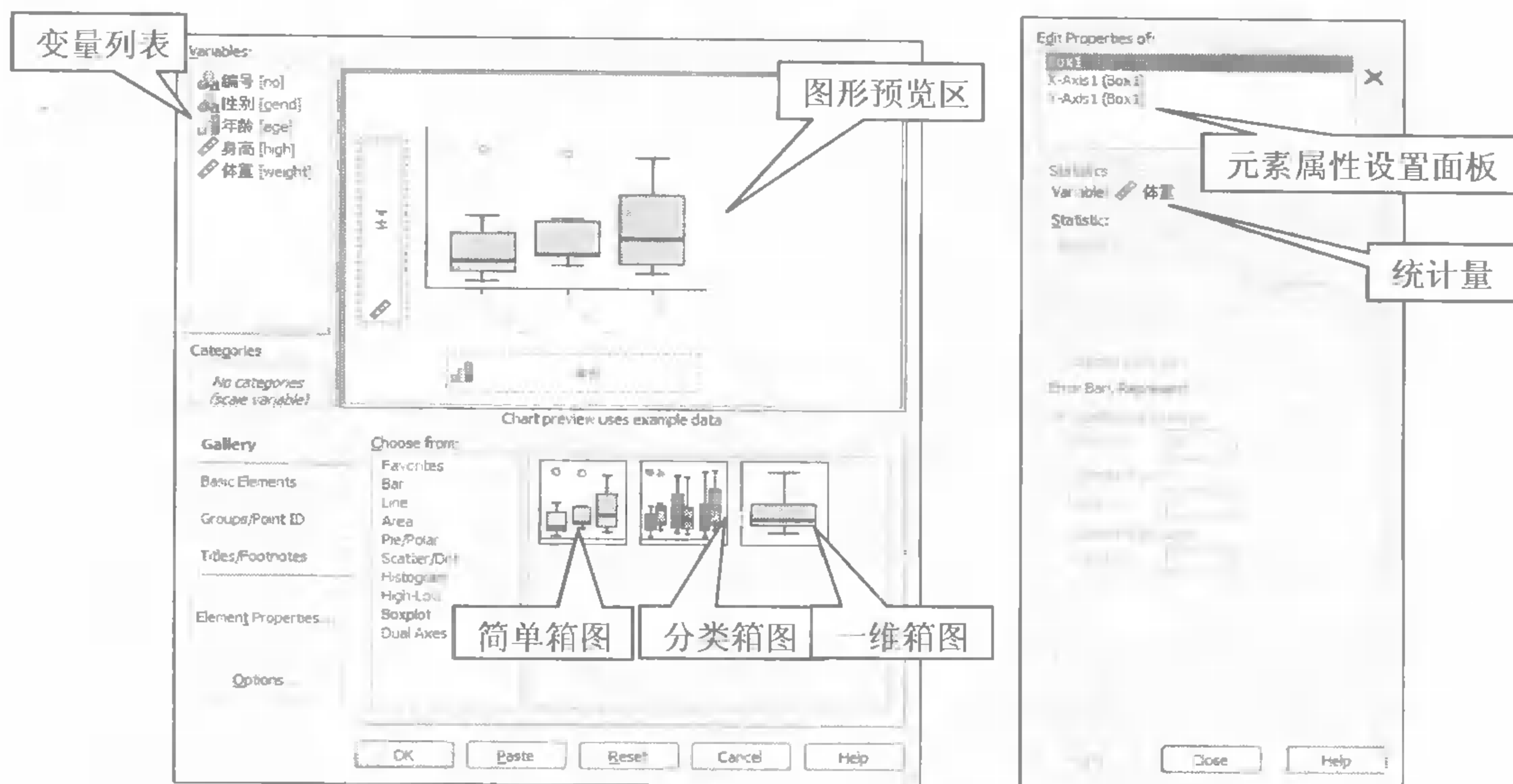


图 19-100 创建简单箱图的设置界面

(2) 输出图形。在图 19-100 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-101 所示，各图形元素的含义如图中标识。

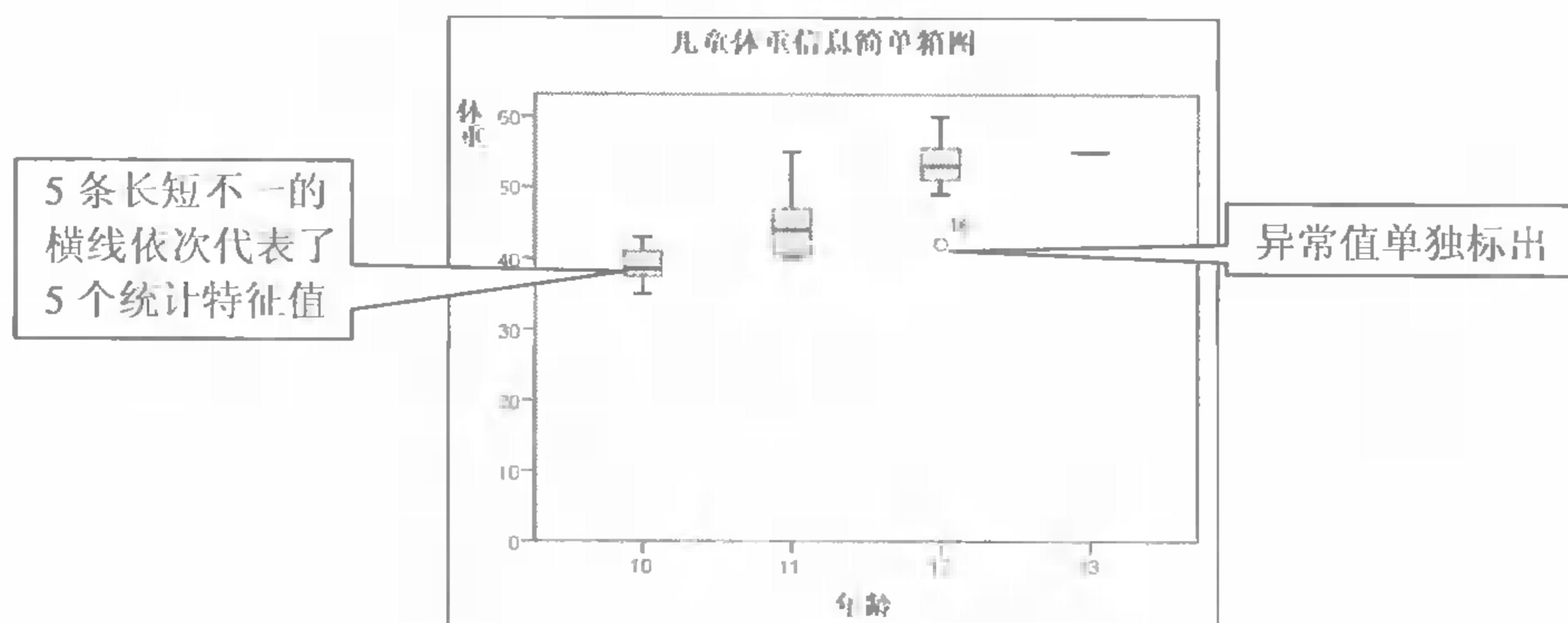



图 19-101 儿童体重信息简单箱图

## 2. 分类箱图

分类箱图需要指定一个分类变量，对它的每个取值都分别做箱形子图，以便观察和比较在不同类别下 Y 轴变量的分布特点。

在图 19-100 中双击预置图标  (Clustered Boxplot) 后，在图形预览区给出分类箱图的预览，同时自动弹出元素属性设置面板；把预置图标拖动至图形预览区，可以达到相同的效果。

从变量列表中把年龄、体重、性别三个变量分别拖动至预览区的 X-Axis、Y-Axis 和 Set color 三个虚线框中，将其分别作为分类箱图的 X 坐标轴、Y 坐标轴和子分类变量。单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-102 所示，它与图 19-101 意义相似，只是在每个年龄上又按照性别进行了区分统计，便于比较性别间的差异。

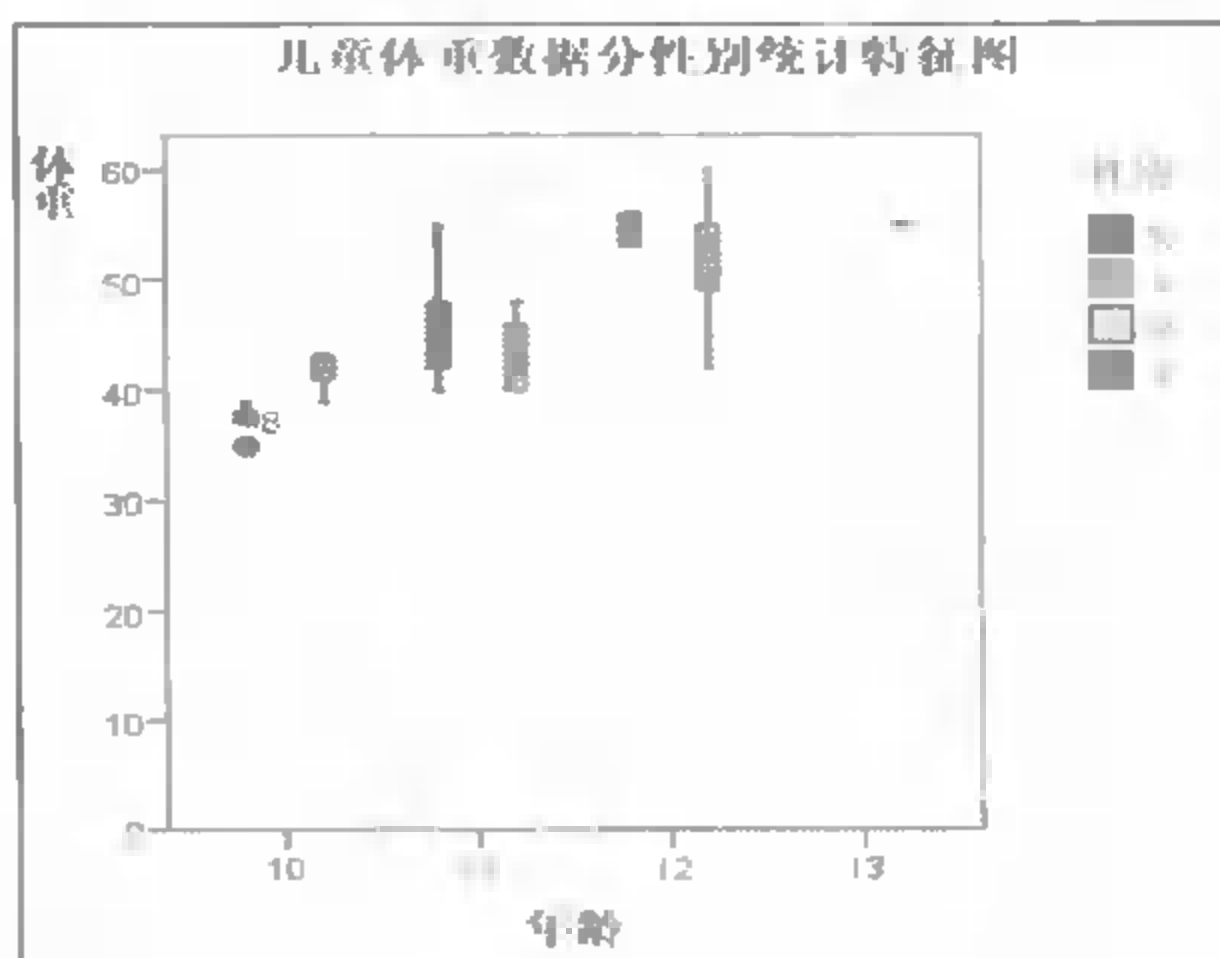


图 19-102 儿童体重信息分性别统计箱图

### 19.9.3 交互式箱图

依次单击菜单“Graphs→Interactive→Boxplot...”打开建立交互式箱图的操作界面，如图 19-103 所示。此界面的设置方法与图 19-36 相似，请参考第 19.3.3 节的介绍。

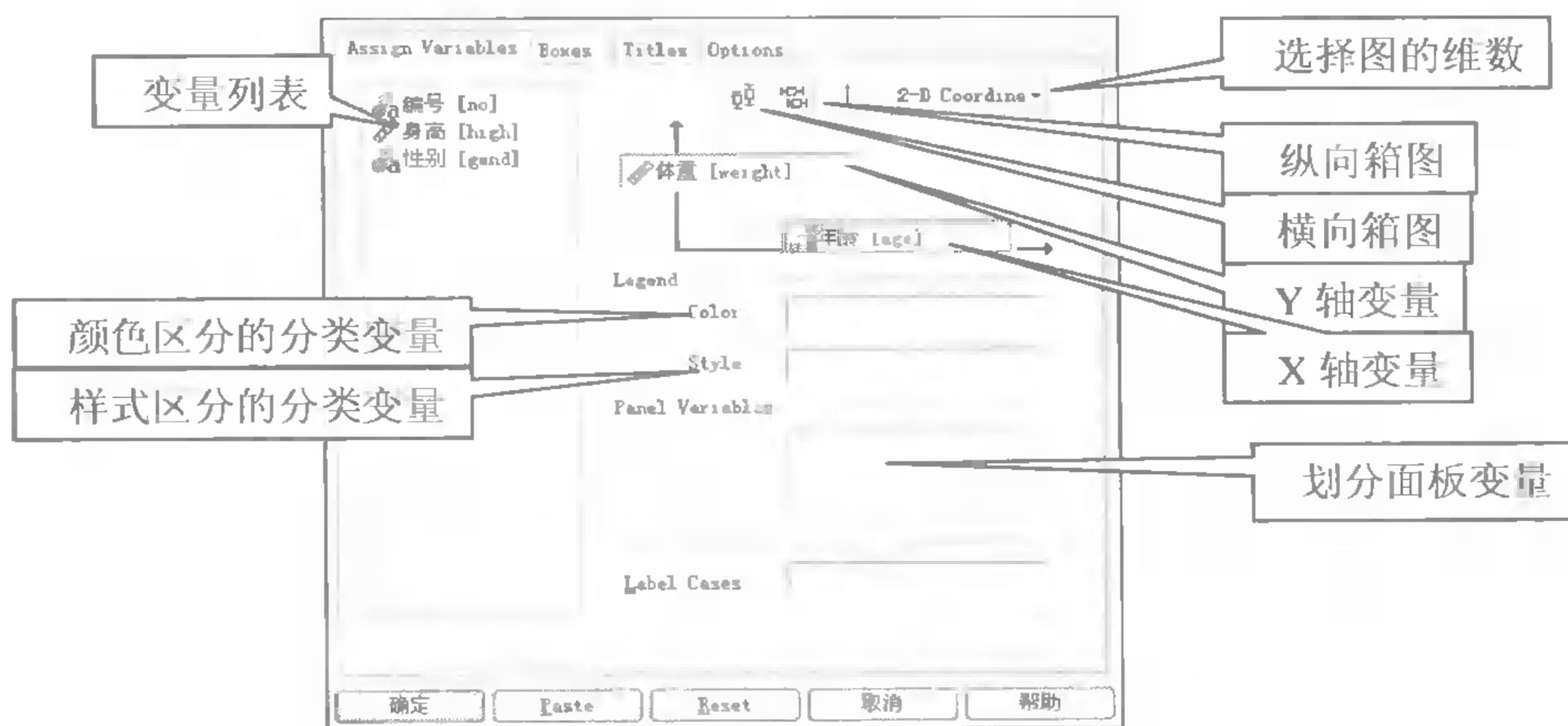


图 19-103 作交互式箱图的设置界面

#### 1. 简单箱图

在图 19-103 里，从变量列表中把年龄、体重，分别拖动至 X 轴变量、Y 轴变量所示的选框，将其分别作为简单箱图的 X、Y 坐标轴。单击确定按钮运行，SPSS Viewer 窗口的输出图形如图 19-104 所示，它与图 19-101 十分相似。

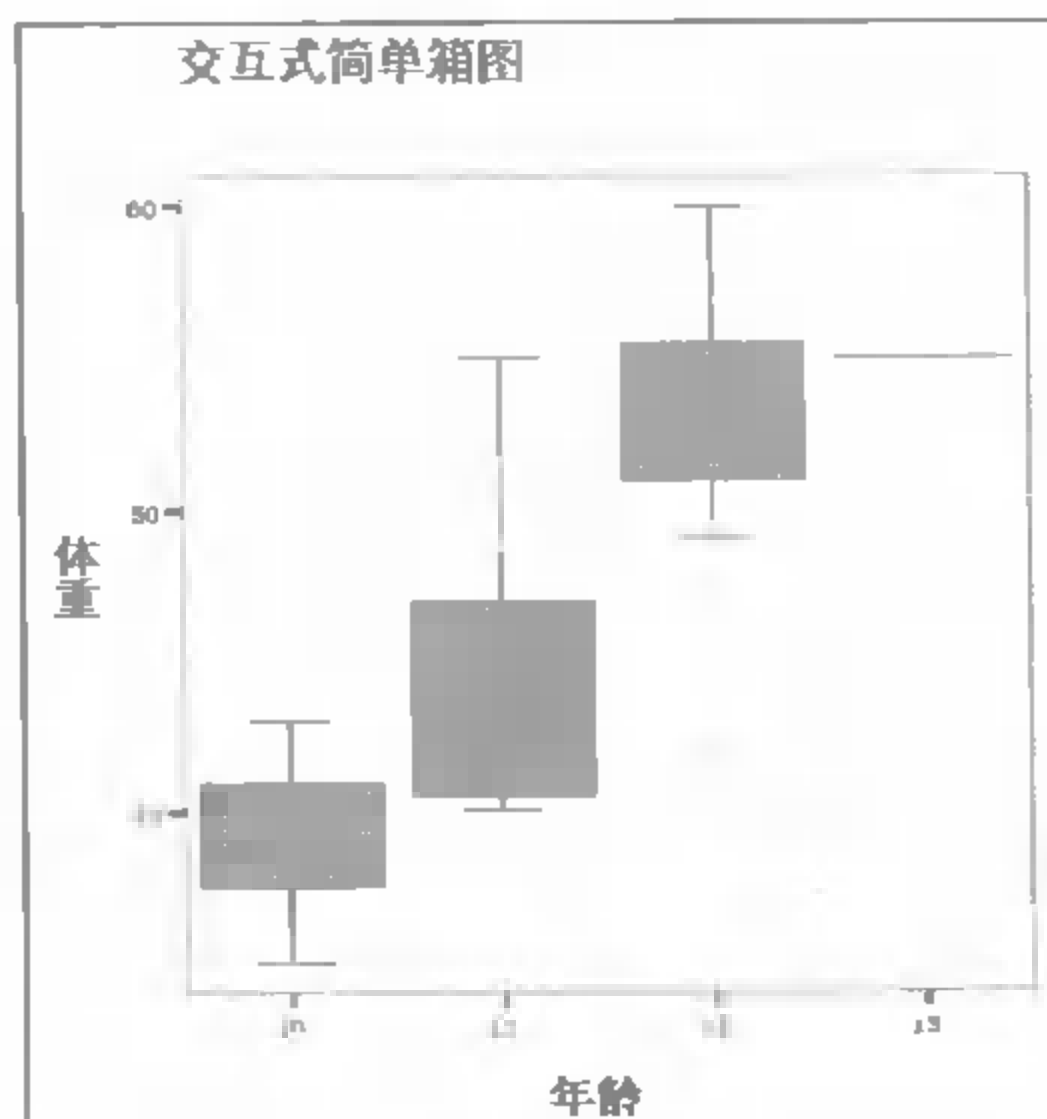


图 19-104 交互式简单箱图



## 2. 分类箱图

在图 19-103 里, 从变量列表中把年龄、体重、性别, 分别拖动至 X 轴变量、Y 轴变量和 Color 选框, 将其分别作为分类箱图的 X 坐标轴、Y 坐标轴和子分类变量。单击确定按钮运行, SPSS Viewer 窗口的输出图形如图 19-105 所示, 它与图 19-102 十分相似。

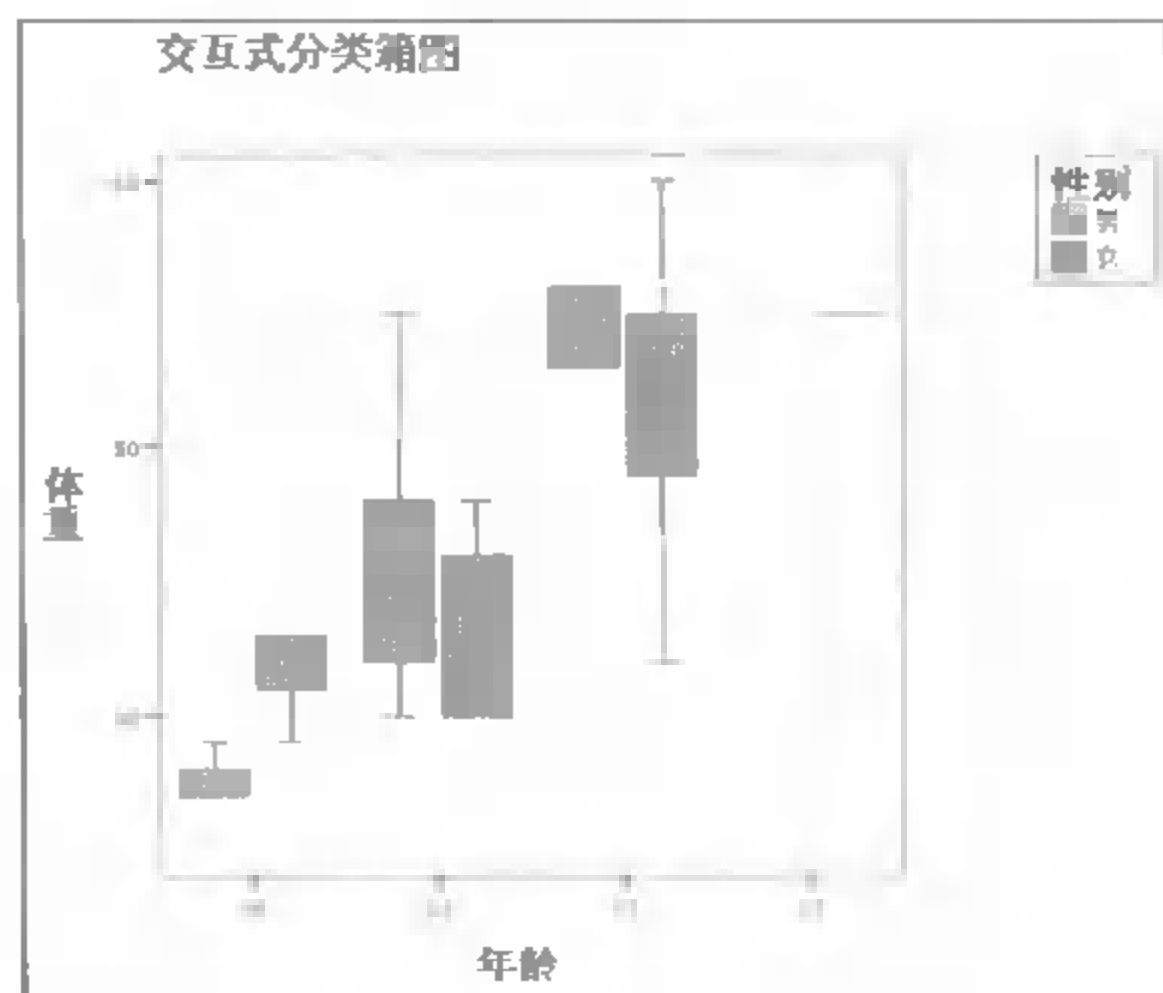


图 19-105 交互式分类箱图

### 19.9.4 用对话框创建箱图

依次单击菜单“Graphs→Legacy Dialogs→Boxplot...”打开利用对话框创建箱图的类型选择界面, 如图 19-106 所示。

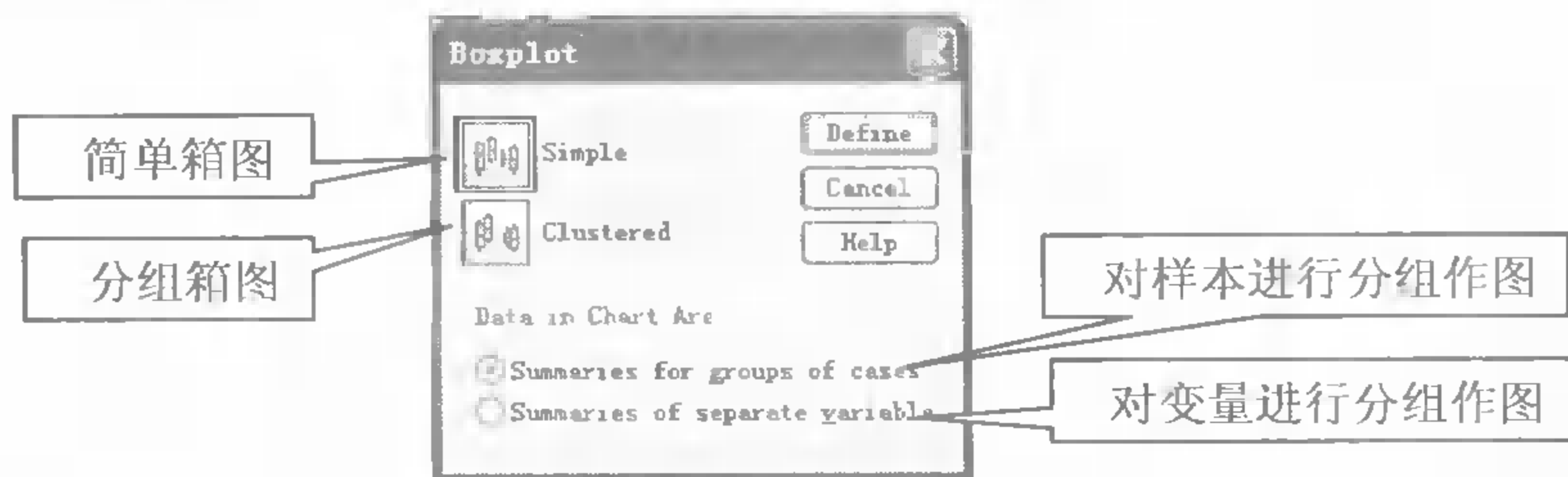


图 19-106 创建箱图类型选择界面

对样本进行分组作图的情况, 与第 19.2.4 节作条形图的设置相仿 (如图 19-30 所示)。

对变量进行分组作图的情况, 与第 19.3.4 节作线形图的设置相仿 (如图 19-40 所示)。

下面以对变量的分组箱图为例, 介绍此窗口的使用方法。

在图 19-106 中单击选中 Clustered 图标, 单击选中 Summaries of separate variable 单选框; 单击 Define 按钮进入作图界面, 如图 19-107 所示。

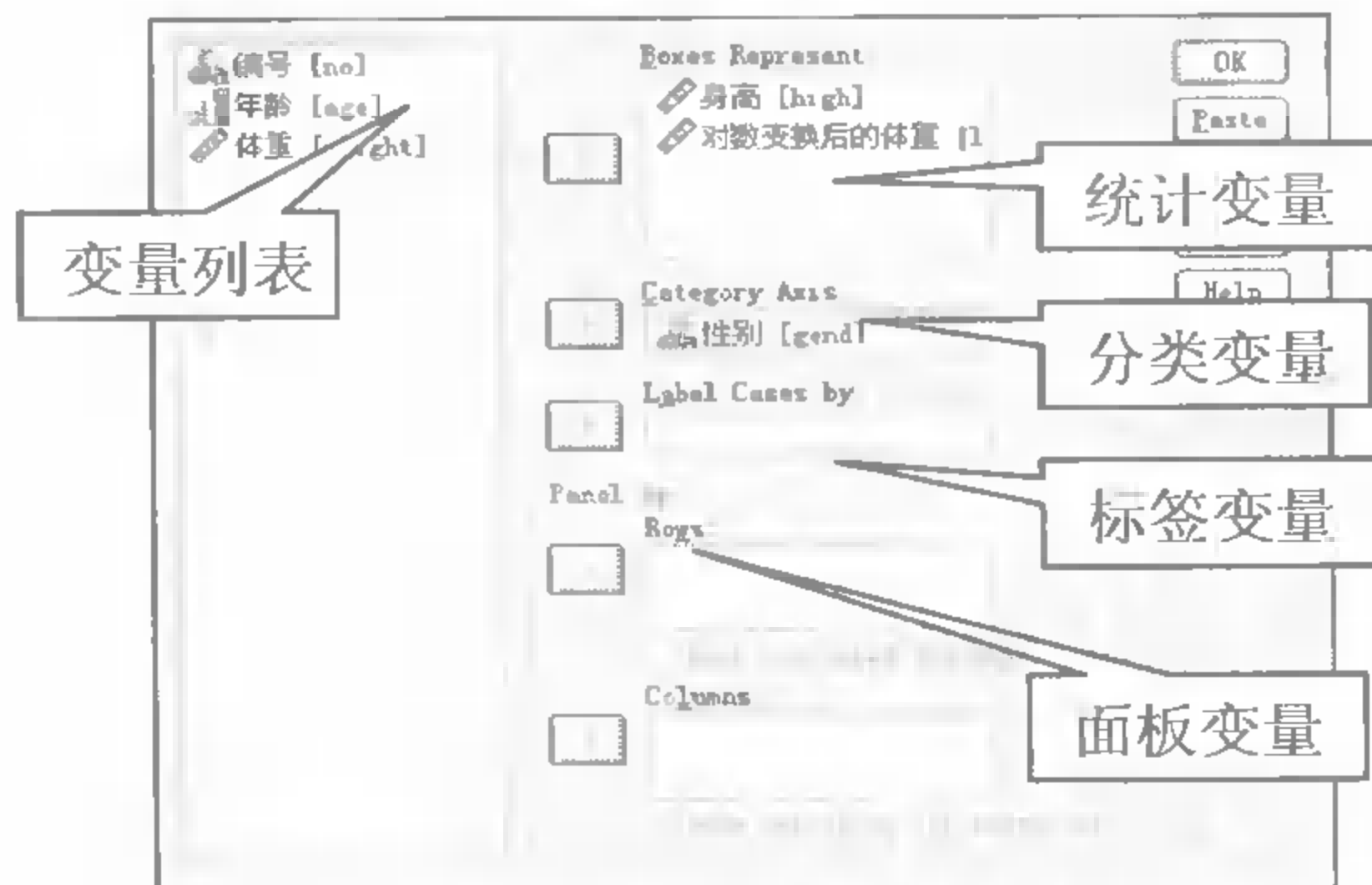




图 19-107 对话框箱图设置面板

在变量列表中选中身高和对数变换后的体重，单击从上至下第一个  按钮，将其作为统计变量选入 Boxes 列表框；在变量列表中单击选中性别，单击从上至下第二个  按钮，将其作为分类变量选入 Category Axis 选框。单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-108 所示，给出了关于两个变量在不同性别里的 5 个统计量的信息。

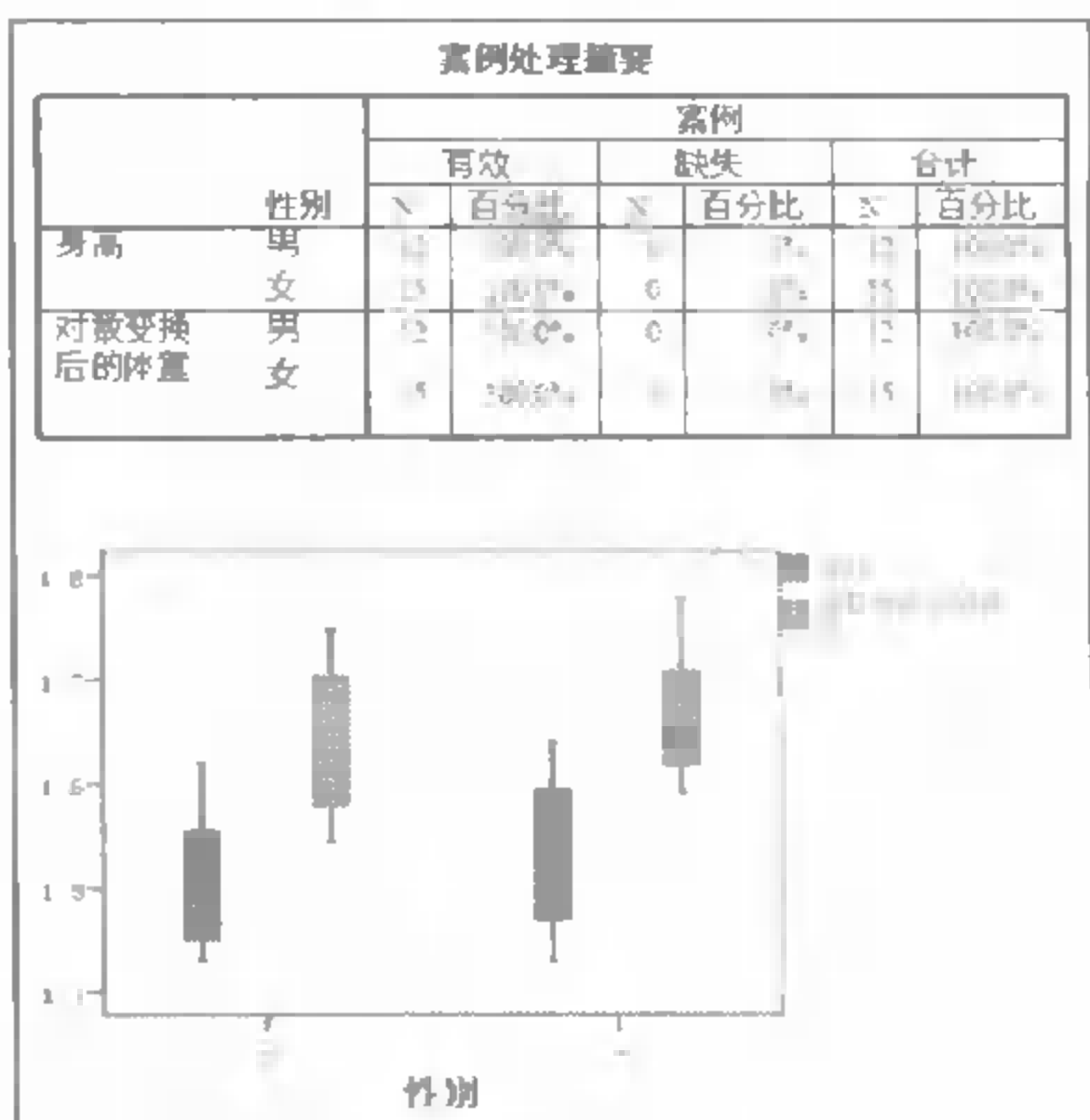


图 19-108 对话框箱图的输出结果

## 19.10 误差条图

误差条图 (Error Bar Charts) 是一种描绘数据离散情况的统计图形，它主要表现变量的均值及其置信区间、标准差或标准误等统计量。在误差条图里，用小方块表示平均值，方块的两端表示置信区间、标准差或标准误，还可以指定向单向或双向延伸误差。

误差条图可以伴随着其他图形的建立过程而输出，例如：条形图、线图等。

### 19.10.1 数据和问题描述

本节利用误差条图来描绘不同年龄 (10-13 岁)、不同性别的儿童的体重和身高特征。所用数据文件为“儿童身高体重数据.sav”，数据格式如图 19-99 所示。

### 19.10.2 交互式误差条图

依次单击菜单“Graphs→Interactive→Error Bar...”打开建立交互式误差条图的操作界面，如图 19-109 所示。此界面的设置方法与图 19-103 相似，请参考第 19.9.3 节的介绍。

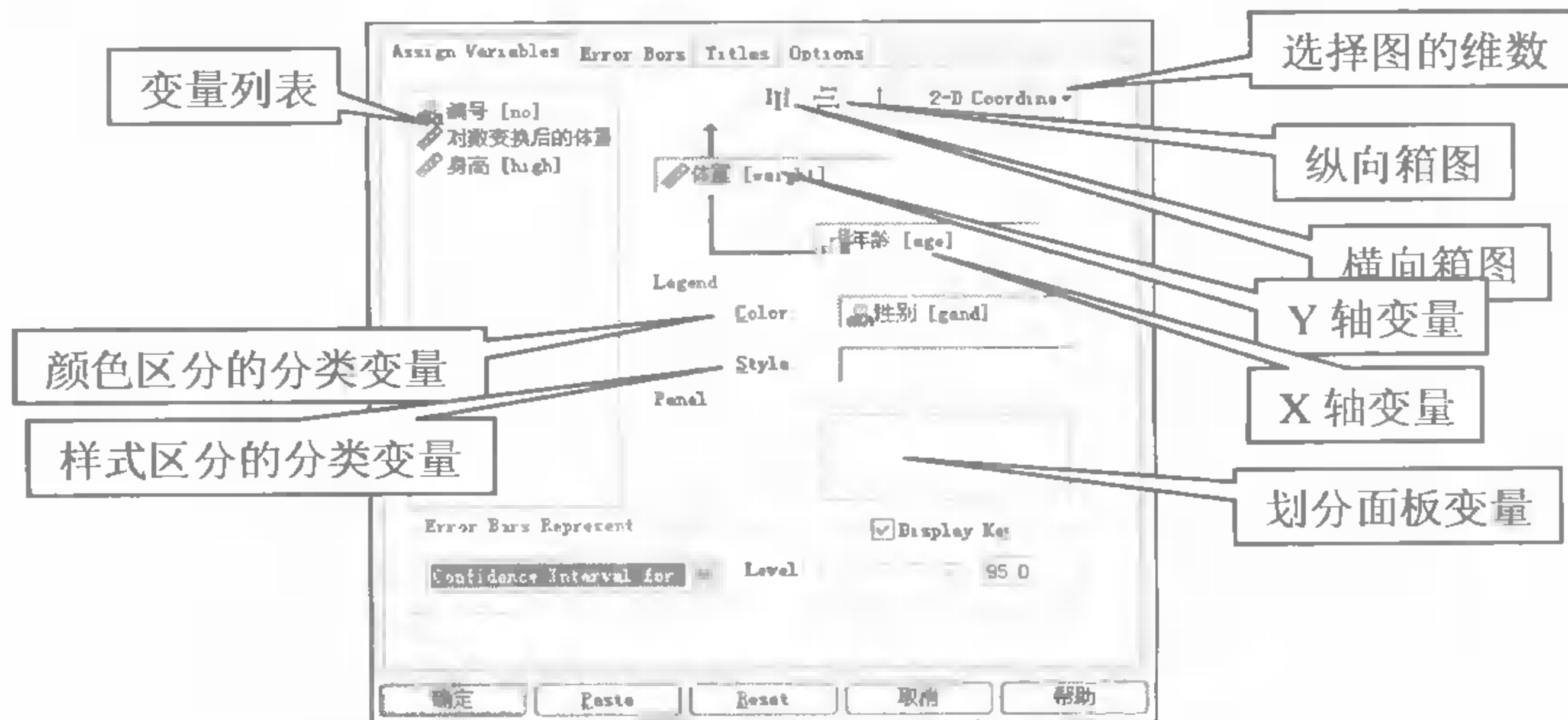


图 19-109 交互式箱图的设置界面

## 1. 设置选项

如图 19-109 所示, Error Bars Represent 子设置栏用于指定误差条图的显示内容和参数, 可选项有如下 3 个。

(1) Confidence Interval for Mean 选项, 表示显示均值的置信区间, 选中后需在 Level 输入框指定置信水平, 默认值为 95 (%)。

(2) Standard Deviations 选项, 表示显示均值的标准差, 选中后需在 Multiplier 输入框指定一个 0~6 的倍数, 默认值为 2。

(3) Standard Errors of Mean 选项, 表示显示均值的标准误, 选中后需在 Multiplier 输入框指定一个 0~6 的倍数, 默认值为 2。

## 2. 输出图形

在图 19-109 里, 从变量列表中把体重、年龄、性别, 分别拖动至 X 轴变量、Y 轴变量和 Color 选框, 将其分别作为误差条图的 X 坐标轴、Y 坐标轴和子分类变量。

在 Error Bars Represent 栏保留默认的设置, 表示图中将显示均值 95% 的置信区间。

单击确定按钮运行, SPSS Viewer 窗口的输出图形如图 19-110 所示, 可见, 误差条图的图形格式与箱图比较相似, 只是小方块和横线代表的统计量不同而已。

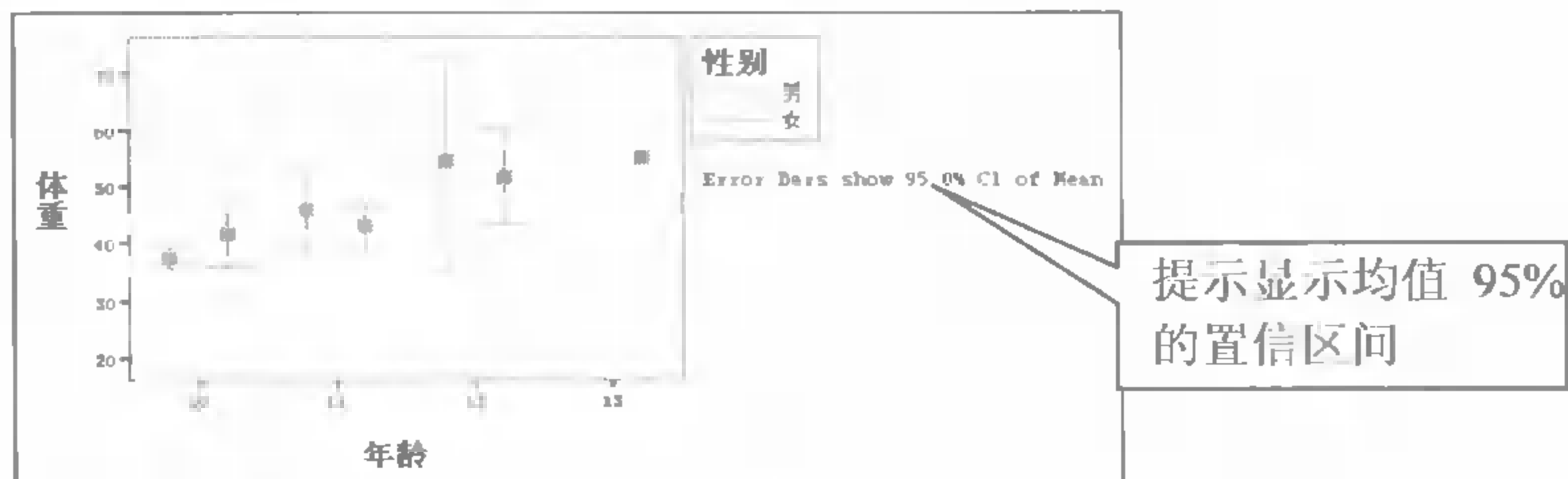


图 19-110 交互式分类误差条图的输出

### 19.10.3 用对话框创建误差条图

依次单击菜单“Graphs→Legacy Dialogs→Error Bar...”, 打开利用对话框创建误差条图的类型选择界面, 如图 19-111 所示, 它与图 19-106 所示的箱图选择界面相似。

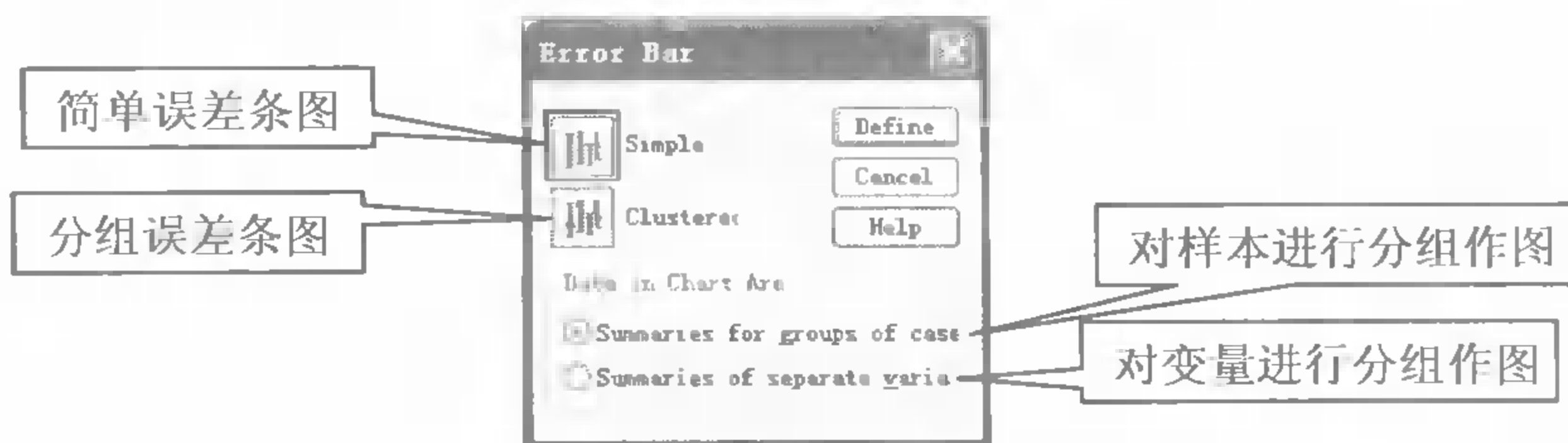


图 19-111 创建误差条图的类型选择界面

下面以对样本作简单误差条图为例, 介绍此窗口的使用方法。

在图 19-111 中单击选中 Simple 图标, 单击选中 Summaries for groups of cases 单选框; 单击 Define 按钮进入作图界面, 如图 19-112 所示。其中: Bars Represent 栏的选项内容和设置方法都和图 19-109 中的 Error Bars Represent 栏相同。

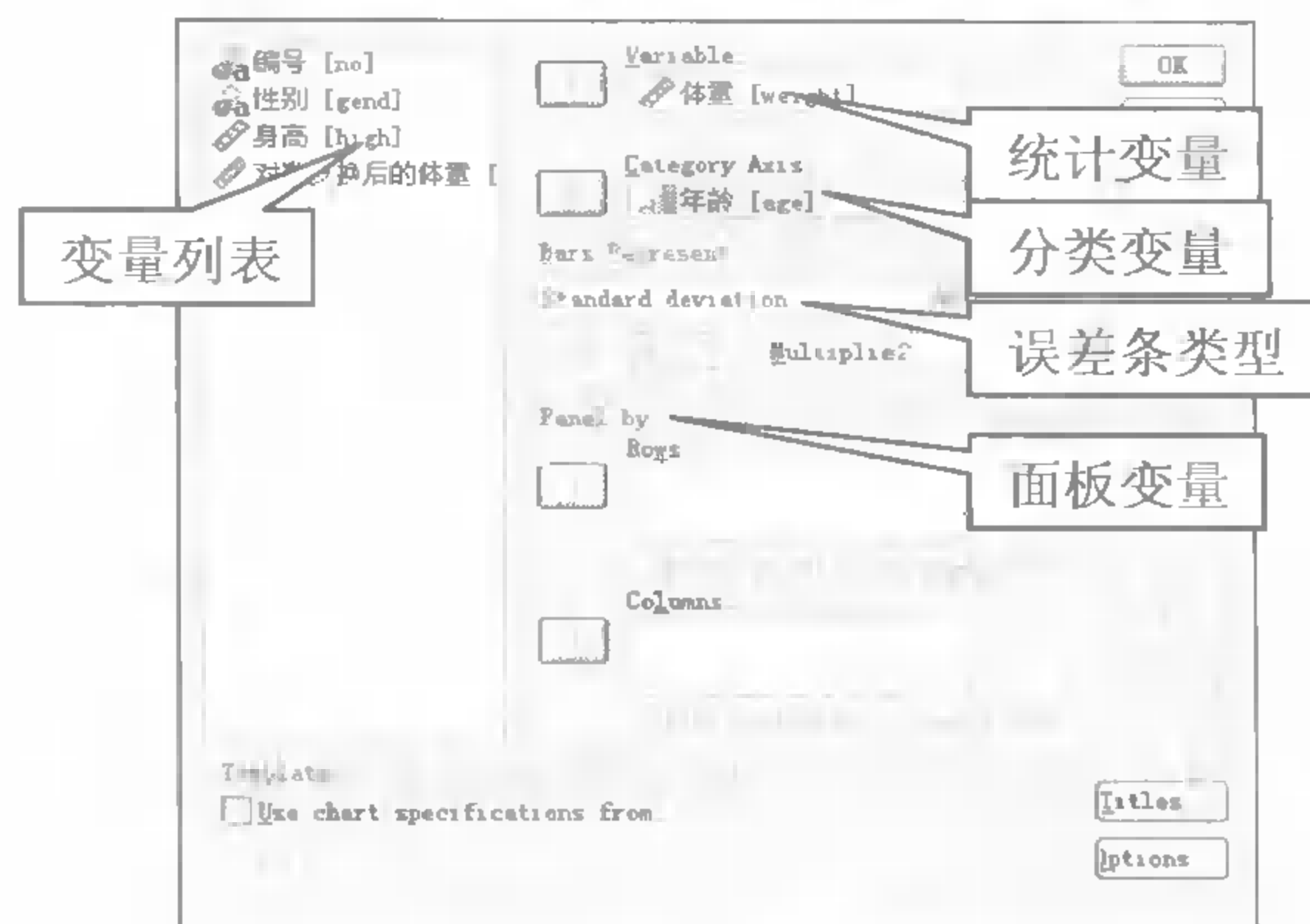




图 19-112 对话框箱图的设置面板

在变量列表中选中体重变量，单击从上至下第一个  按钮，将其作为统计变量选入 Variable 选框；在变量列表中单击选中年龄，单击从上至下第二个  按钮，将其作为分类变量选入 Category Axis 选框；单击 Bars Represent 下拉列表选中 Standard Deviations 选项。单击 OK 按钮运行，SPSS Viewer 窗口的输出如图 19-113 所示，可见它与箱图比较相似，其中的小圆圈和上下横线分别代表了均值、均值+2 倍标准差和均值-2 倍标准差。

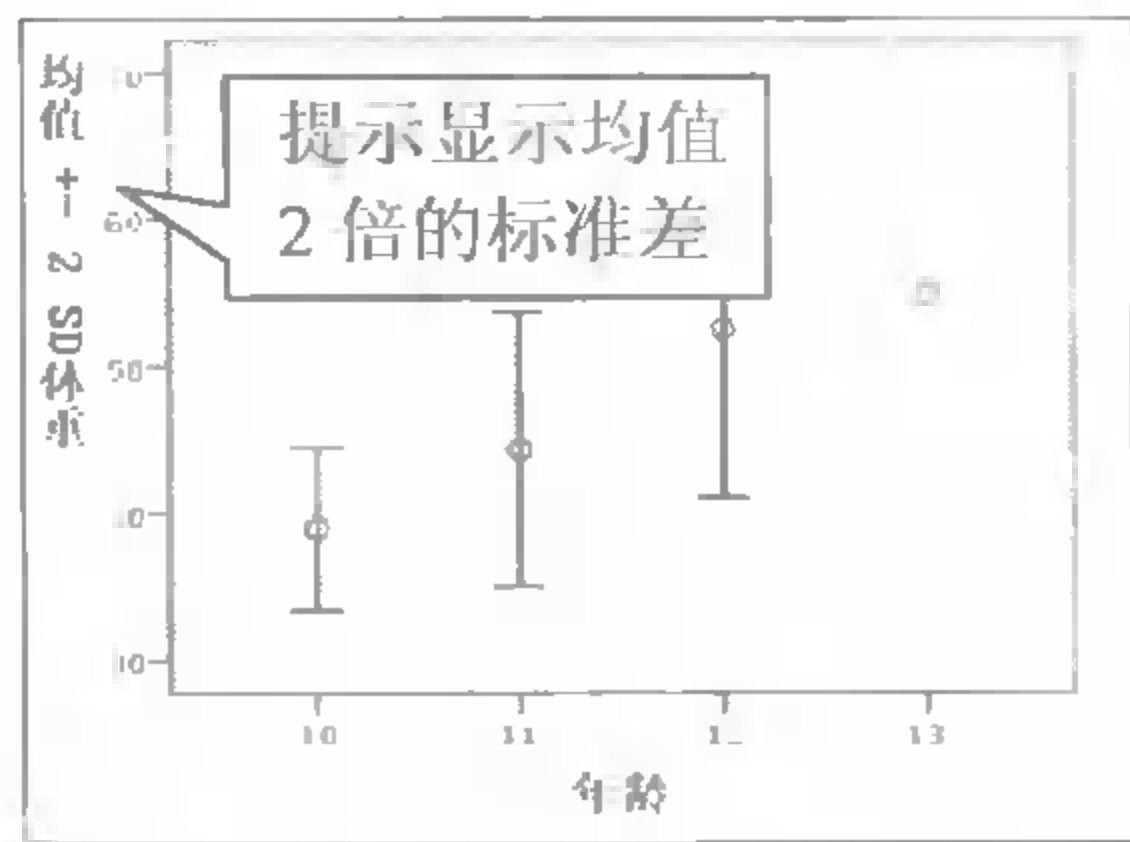


图 19-113 对话框箱图的输出结果

## 19.11 散点图

散点图适用于描绘测量数据的原始分布状况，它以点的分布反映变量之间的相互关系，用户可以从点的位置来判断测量值的高低、大小、变化趋势或变化范围。

### 19.11.1 数据和问题描述

本节利用散点图来描绘政府对教育的投资额度与当地经济发展水平之间的相互关系。所用数据文件为“教育投资与经济增长.sav”，数据格式如所图 19-114 示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	地区号	Numeric	2	0		None	None	8	Right	Scale
2	教育投资	Numeric	6	2	教育投资(万元)	None	None	8	Right	Scale
3	学生增长率	Numeric	6	2	学生增长(%)	None	None	8	Right	Scale
4	经济增率	Numeric	6	2	经济增长(%)	None	None	9	Right	Ordinal
5	地区类别	String	1	0		{a 经济增长	None	8	Left	Nominal

图 19-114 教育投资和经济增长指标的数据格式



## 19.11.2 用图形构建器作高低图

依次单击菜单“Graphs→Chart Builder”打开图形构建器，如图 19-115 所示。单击 Gallery 标签，在 Choose from 列表中单击选中 Scatter/Dot，就在其右侧显示预设的散点图图标。

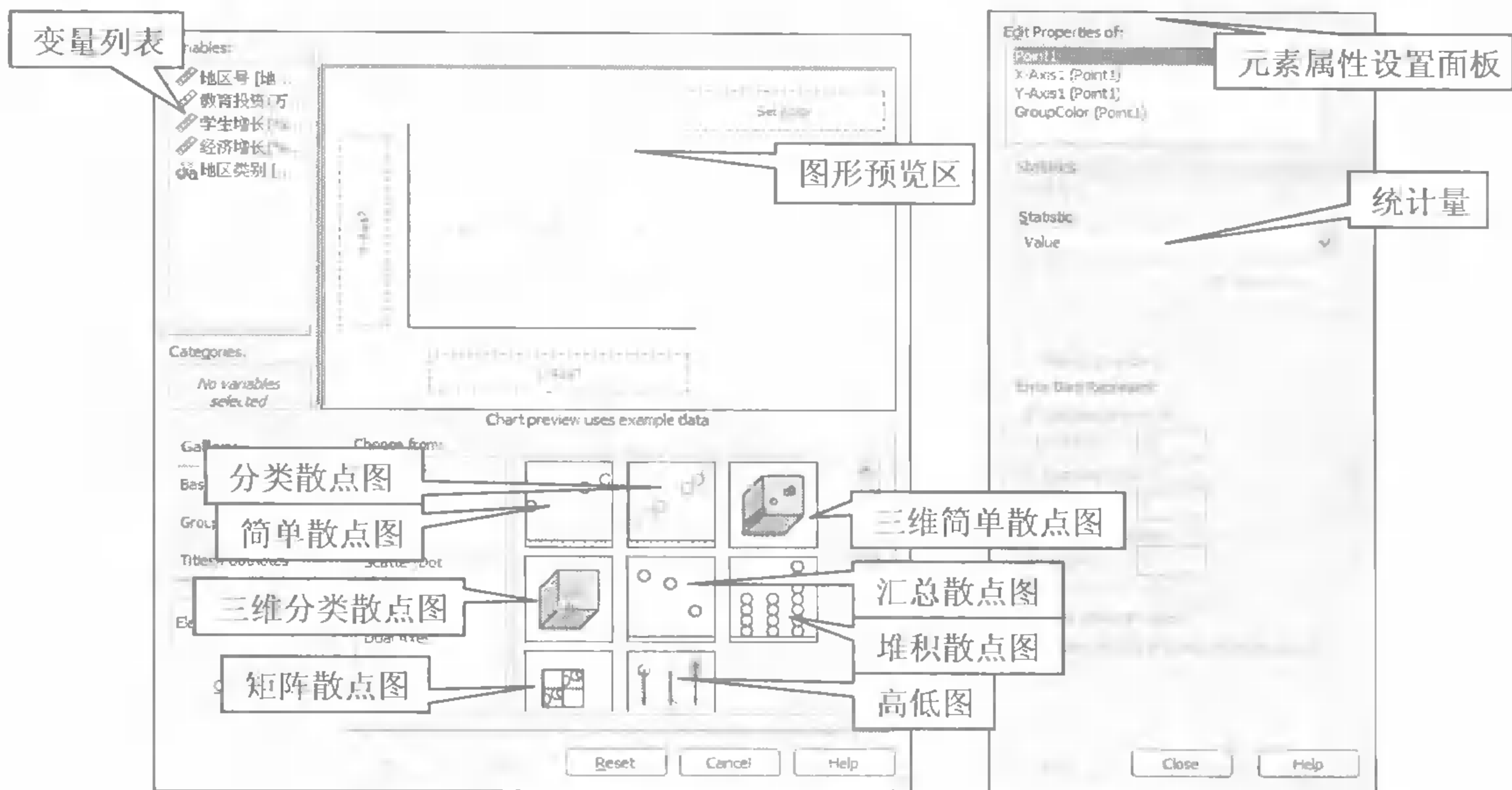



图 19-115 创建散点图的设置界面

## 1. 分类散点图

在图 19-115 中双击预置图标  (Grouped Scatter) 后，在图形预览区给出分类散点图的预览，同时自动弹出元素属性设置面板；把预置图标拖动至图形预览区，可以达到相同的效果。

从变量列表中把教育投资、经济增长、地区类别三个变量，分别拖动至预览区的 X-Axis、Y-Axis 和 Set color 三个虚线框中，将其分别作为分类散点图的 X 坐标轴、Y 坐标轴和子分类变量。单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-116 所示，可见随着教育投资的增加，不同地区的经济增长都有比较明显的加速趋势。

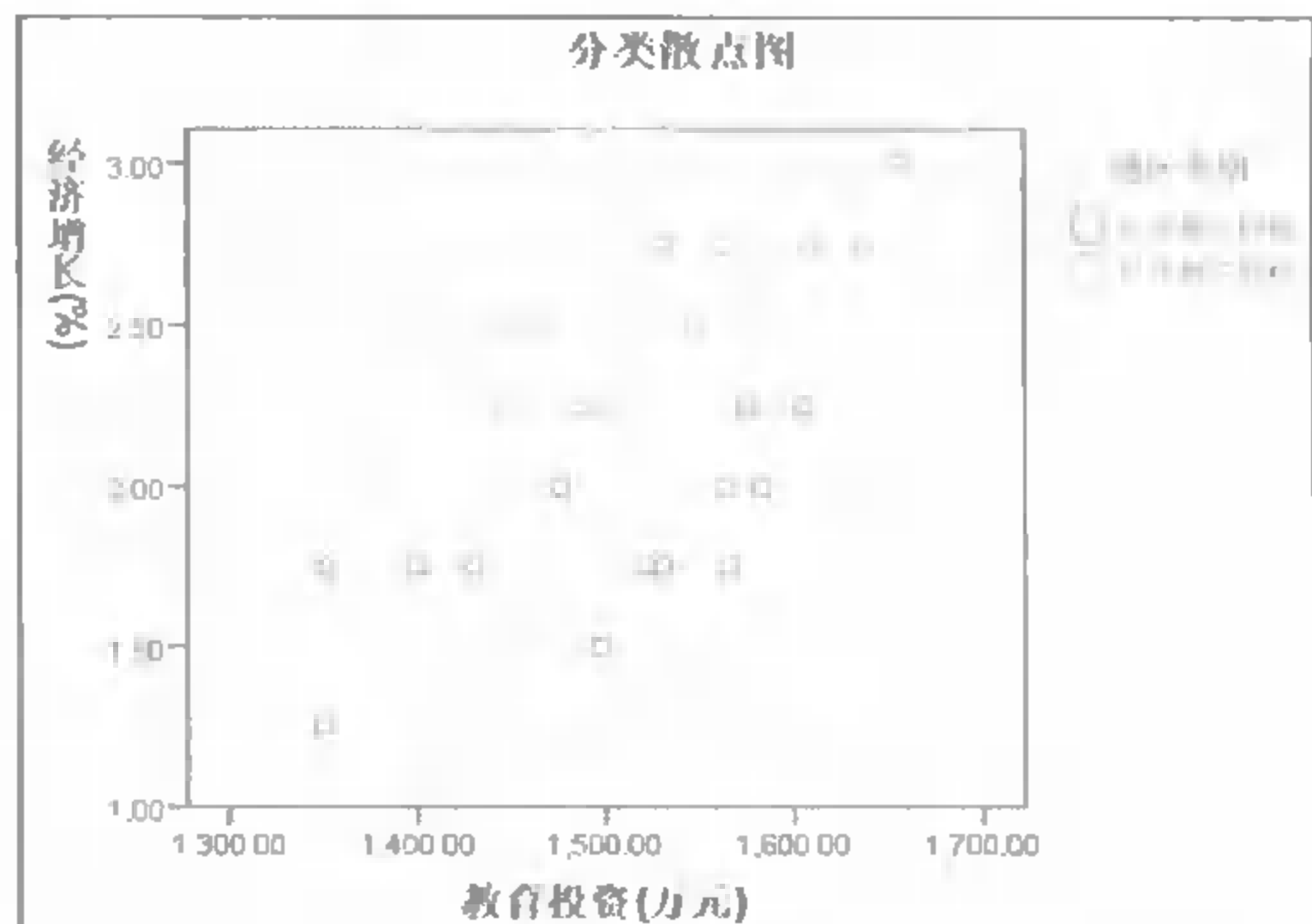


图 19-116 分类散点图设置

简单散点图、三维散点图的参数选项和设置方法，都与分类散点图的设置相仿。

汇总散点图的参数选项和设置方法，也与分类散点图的设置相仿；只是需要在元素属性设置面板里，把元素 Ponit1 的统计量 (Statistic) 设置为均值 (Mean)、求和 (Sum) 等。

## 2. 堆积散点图

在图 19-115 中，双击预置图标  (Simple Dot Plot) 后，在图形预览区给出堆积散点图

的预览，同时自动弹出元素属性设置面板。元素 Ponit1 的属性里多了如图 19-117 所示的 3 个选项，用于指定图中堆积的点的方向和对称性。

从变量列表中把经济增长变量，拖动至预览区的 X-Axis 虚线框中，将其作为堆积散点图的 X 坐标轴；在图 19-117 中单击选中 Symmetric 图标。单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-118 所示，可见经济增长率在 2%左右的地区比较多，较低或较高的经济增长率出现都比较少。

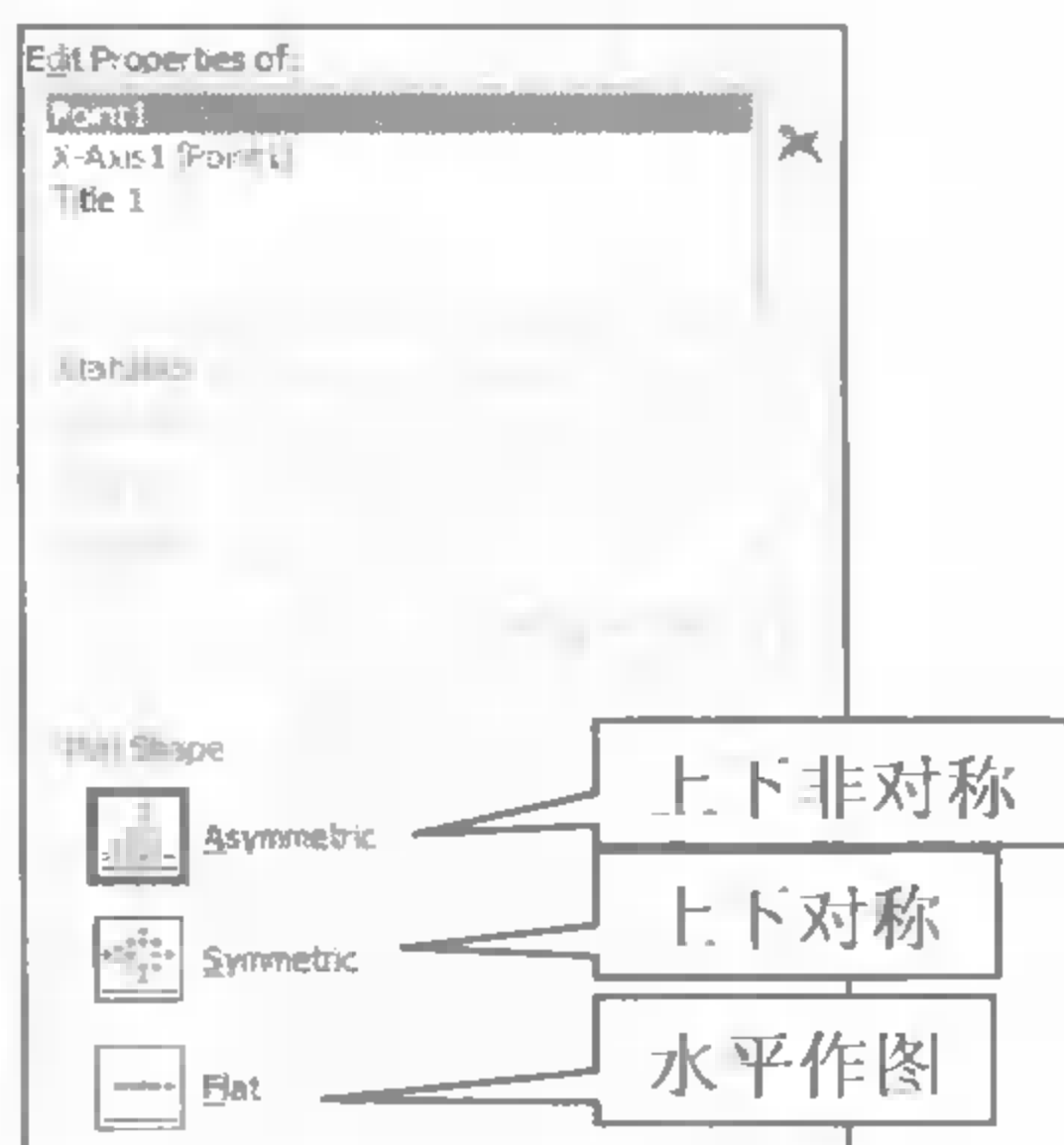


图 19-117 堆积散点图的设置

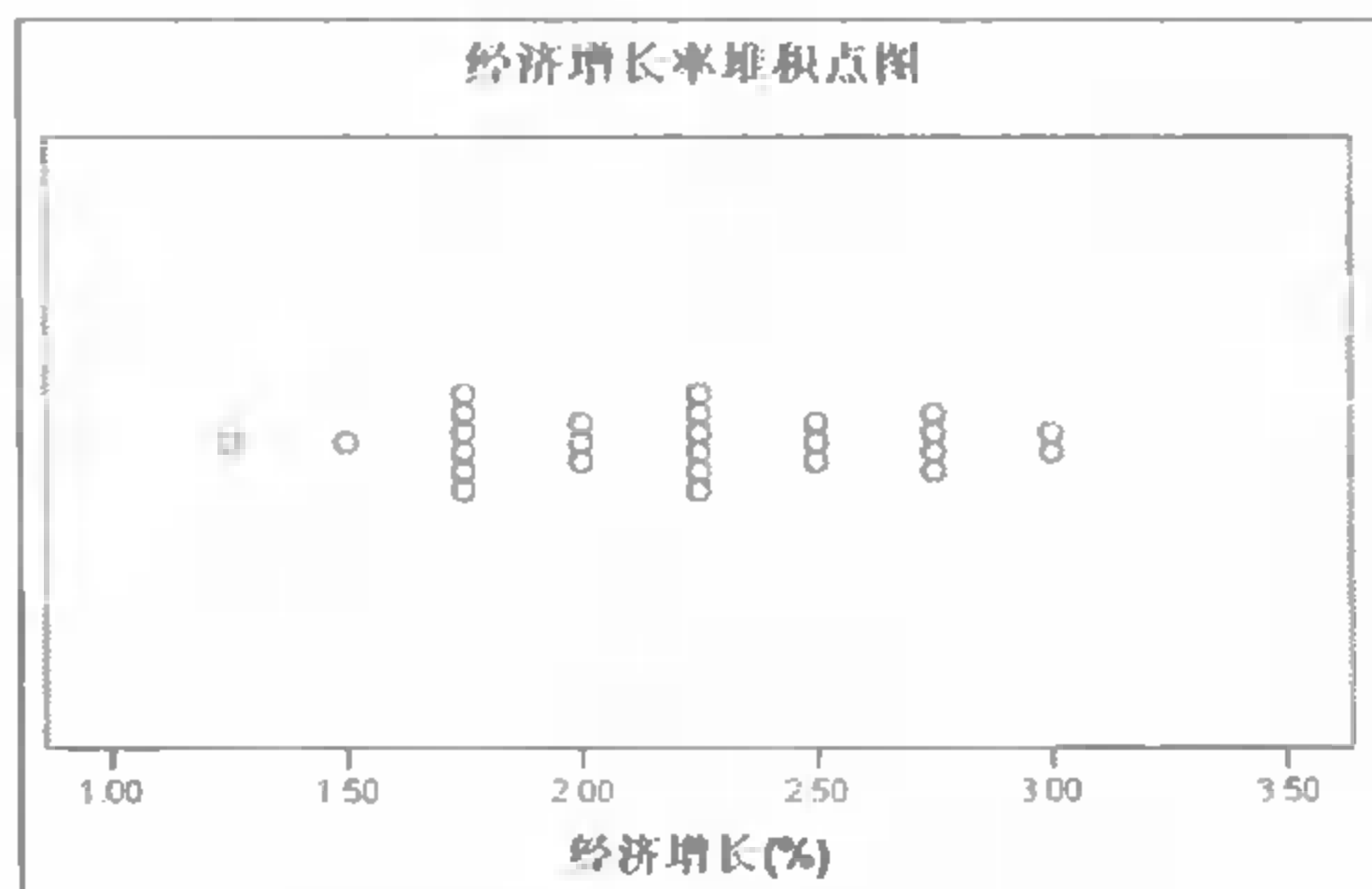


图 19-118 堆积散点图的输出

### 3. 矩阵散点图

矩阵散点图对多个变量的两两组合分别做散点图，并把所有子图形同时显示在一个以变量名为元素的矩阵中。

在图 19-115 中双击预置图标 (Scatter Plot Matrix) 后，在图形预览区给出矩阵散点图的预览，如图 19-119 所示。此时只有一个 Scattermatrix 虚线框供选入变量，但它允许同时选入多个变量，变量的显示顺序可以在元素属性设置面板里更改。

从变量列表中依次把教育投资、学生增长、经济增长 3 个变量，拖动至预览区的 Scattermatrix 虚线框中。单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 19-120 所示，可见矩阵散点图能同时描绘多个变量之间的关系。

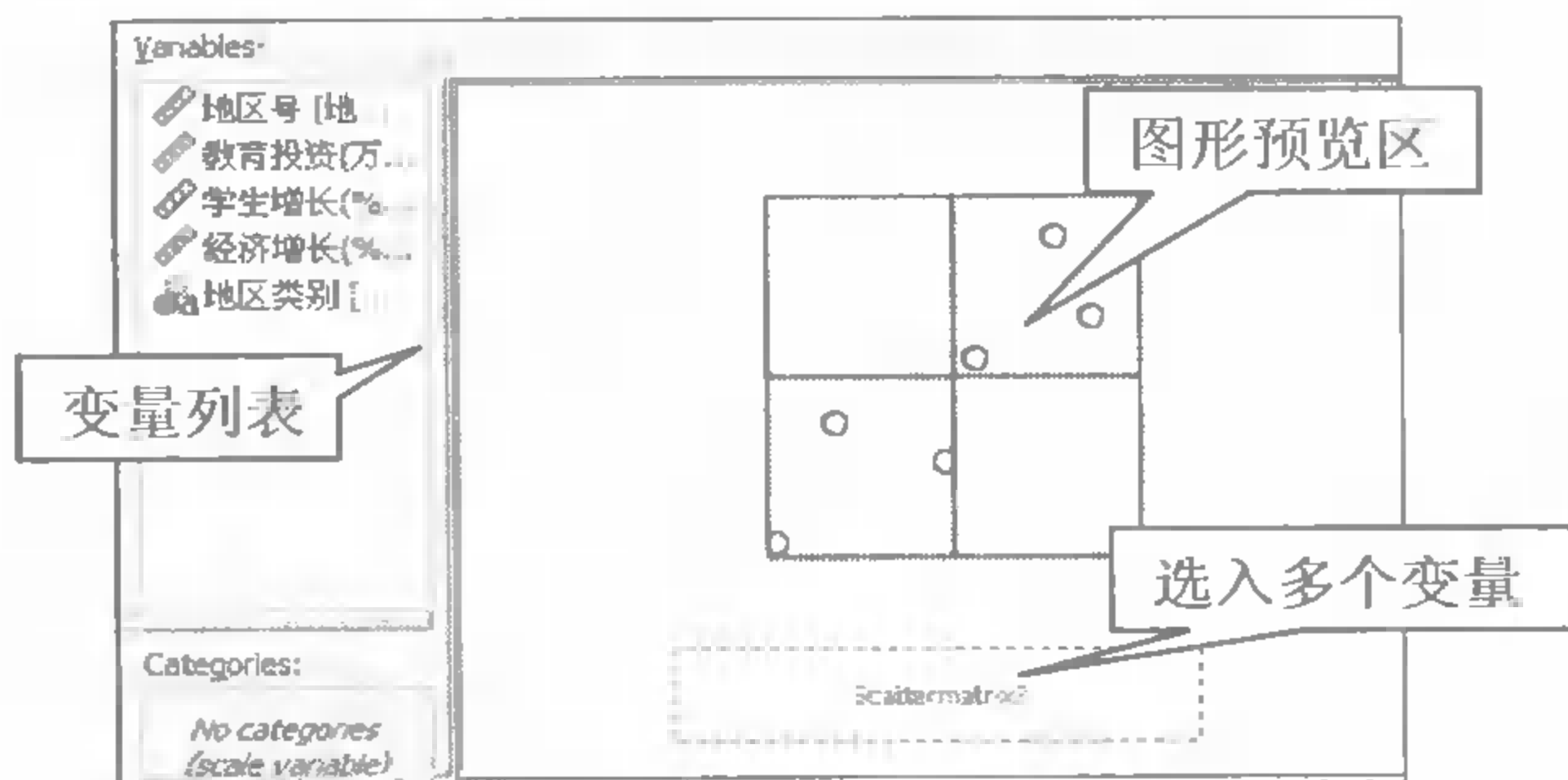


图 19-119 矩阵散点图的预览

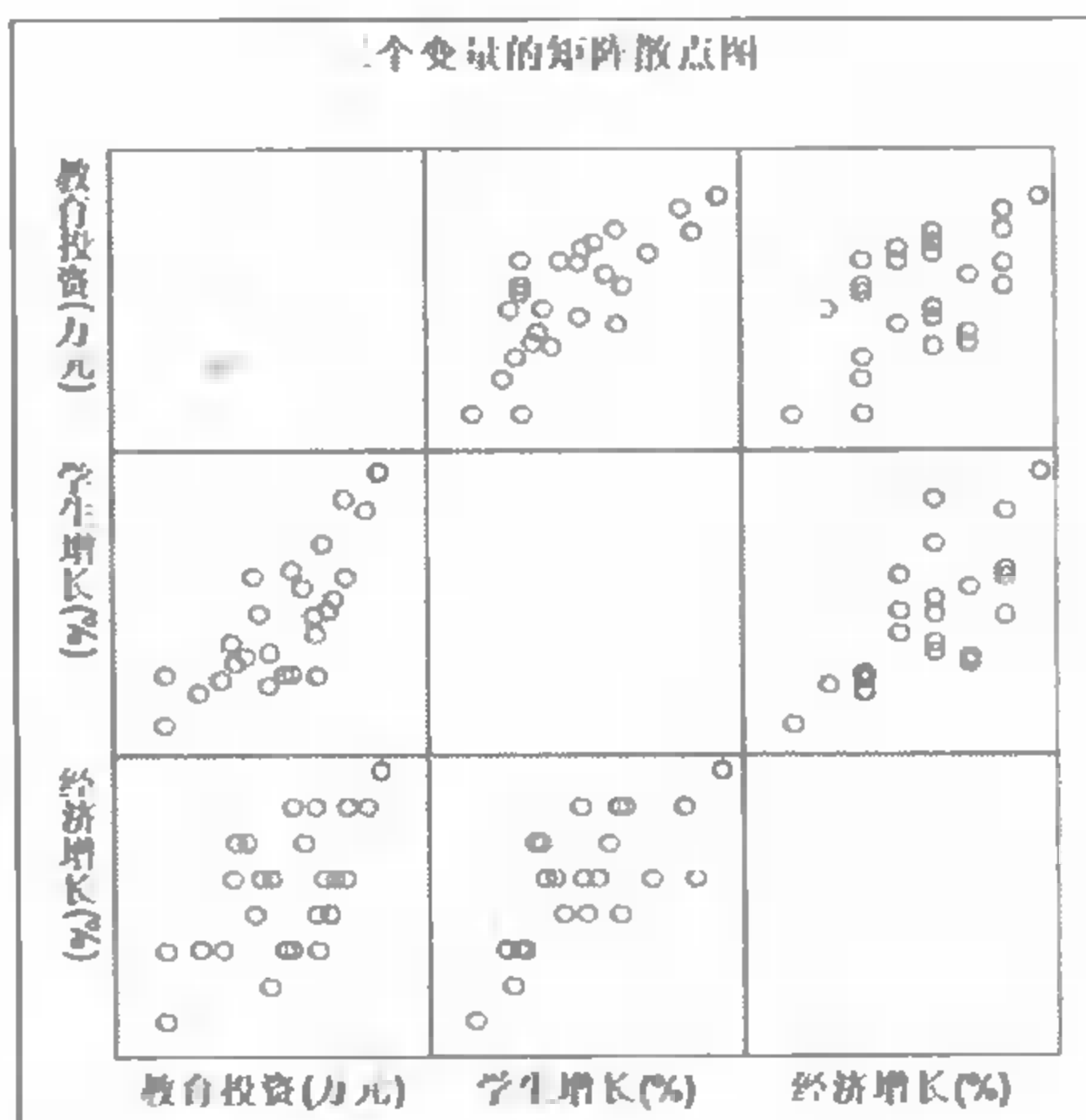


图 19-120 矩阵散点图的输出

#### 4. 重叠散点图

对于有 Y 轴选项的散点图，都可以做重叠散点图，它是由多个 Y 轴变量对应单个 X 轴变量的复合图形，下面以分类散点图为例，介绍如何做重叠散点图。

在图 19-115 中双击预置图标  (Grouped Scatter) 后，在图形预览区给出分类散点图的预览，同时自动弹出元素属性设置面板。

从变量列表中把教育投资，拖动至预览区的 X-Axis 虚线框中；然后，在变量列表中同时选中学生增长和经济增长 2 个变量，再把它们同时拖动到预览区的 Y-Axis 虚线框中，如图 19-121 所示，如此就指定了 2 个 Y 轴变量；此时 Set color 虚线框自动显示“variable”（不可更改），表示以变量名区分不同的散点系列。

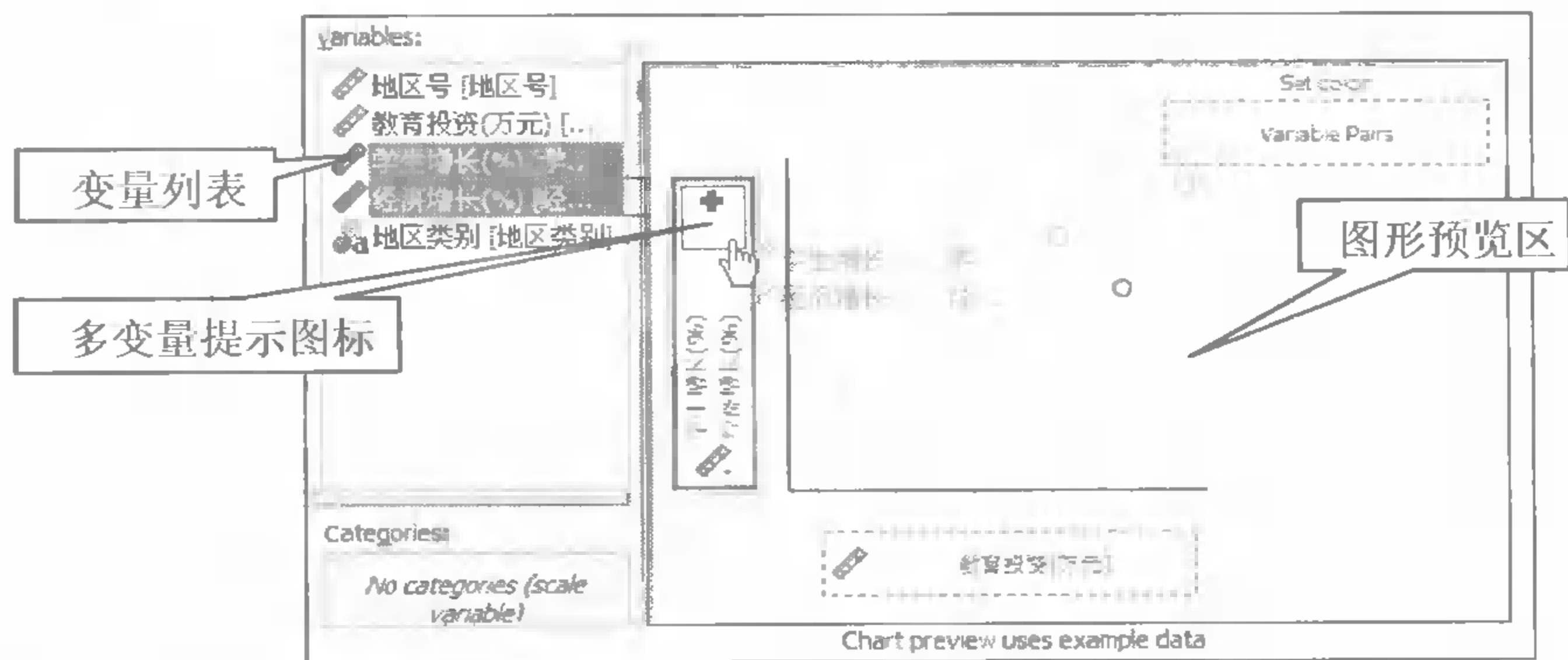


图 19-121 重叠散点图的设置预览区

在图 19-121 中，单击 OK 按钮（未显示）运行，SPSS Viewer 窗口的输出图形如图 19-122 所示。它把经济增长、学生增长分别对教育投资的散点图显示在一起，用颜色区分开来，观察可见，学生的增长水平普遍比经济增长水平快很多。

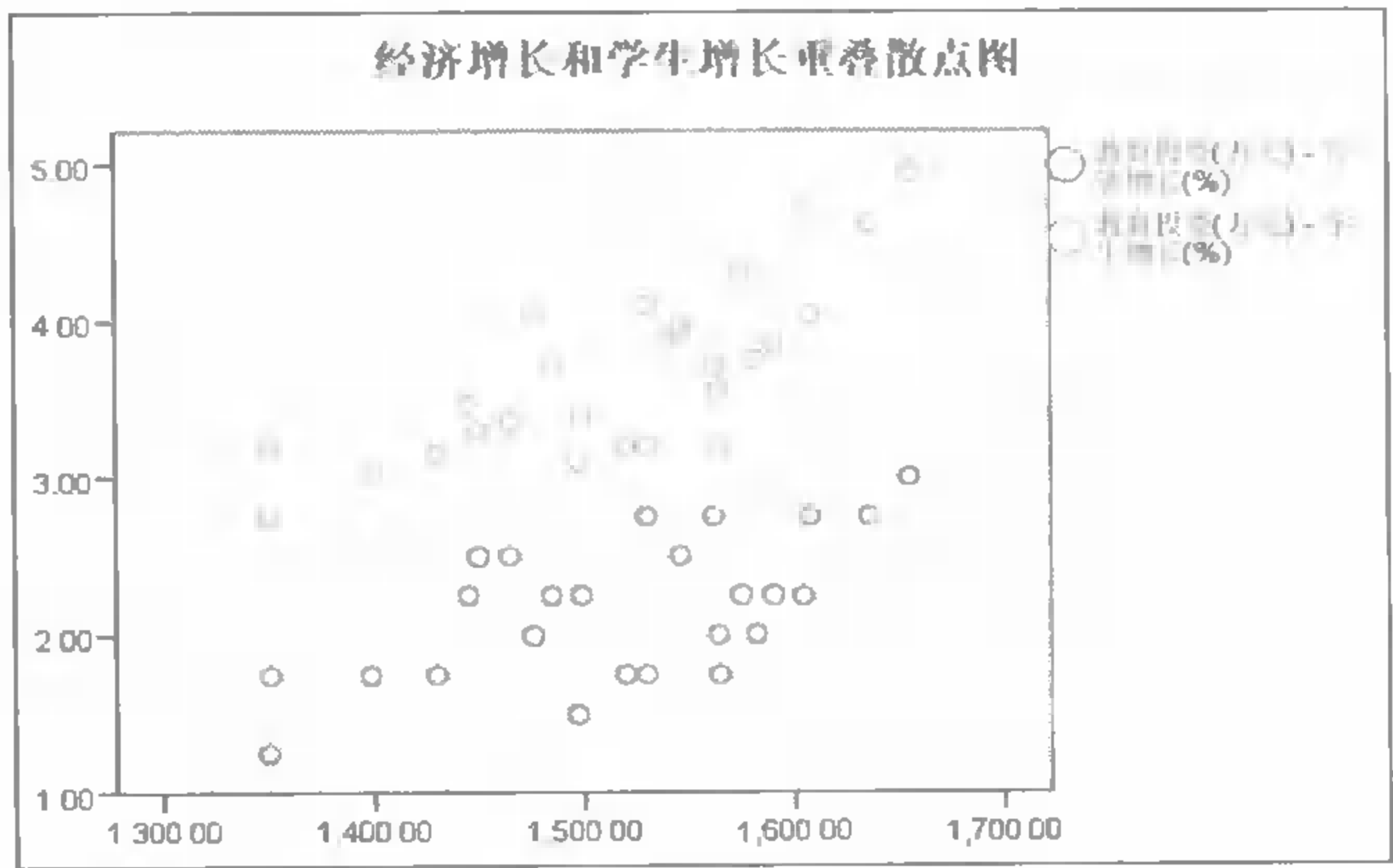


图 19-122 重叠散点图输出

#### 19.11.3 交互式散点图

依次单击菜单“Graphs→Interactive→Scatterplot...”，打开建立交互式散点图的操作界面，如图 19-123 所示。

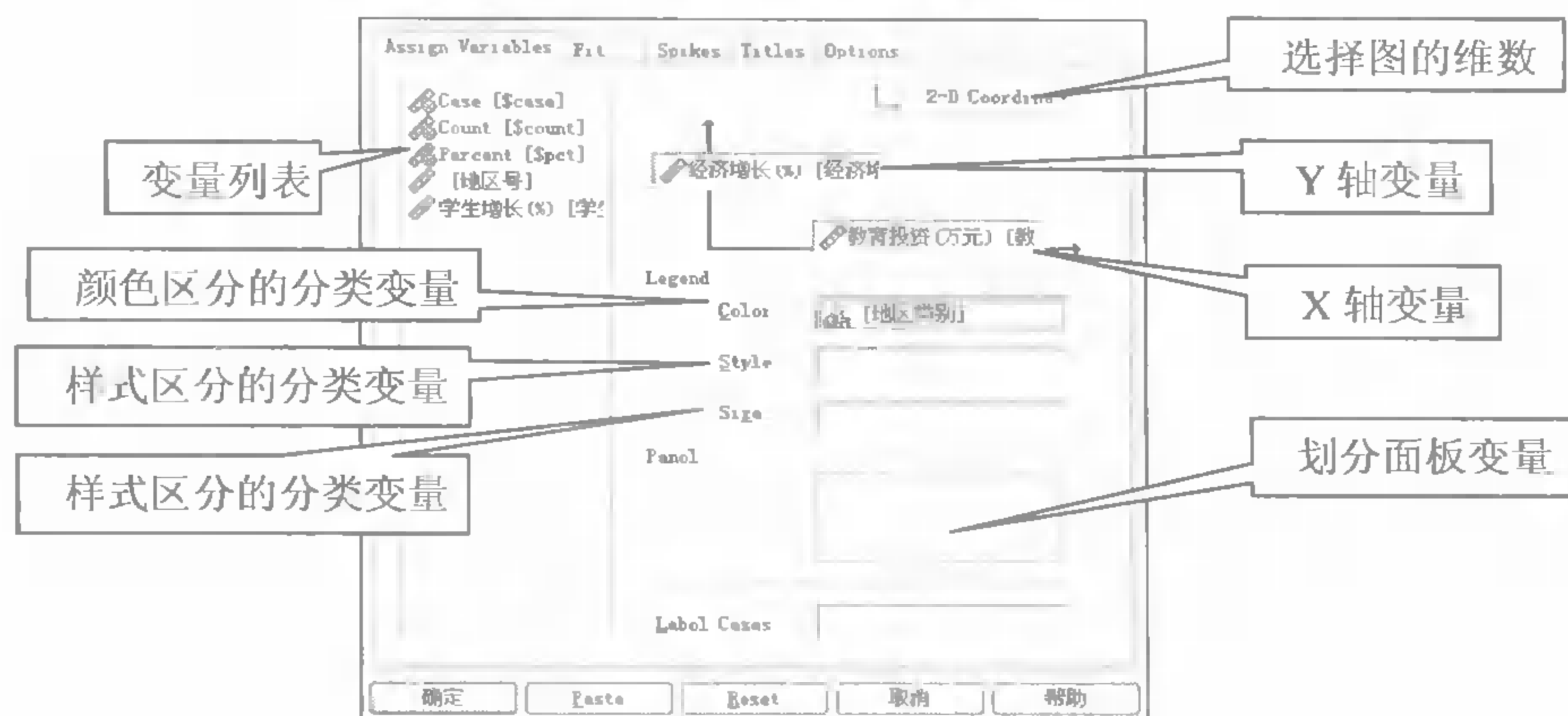


图 19-123 交互式散点图的变量设置

### 1. 变量设置

在图 19-123 中，从变量列表中把经济增长、教育投资、地区类别，分别拖动至 Y 轴变量、X 轴变量和 Color 选框，将其分别作为分类箱图的 Y 坐标轴、X 坐标轴和子分类变量。

### 2. 拟和方法的参数选项

在图 19-123 中，单击 Fit 标签，弹出如图 19-124 所示的选项设置界面，在此设置对散点做拟和曲线的方法和参数。Method 下拉列表给出了 4 个可选项。



图 19-124 交互式散点图的拟和选项设置

(1) None，表示不做拟和，默认选项。

(2) Regression，表示做回归拟和，选中后需要指定以下参数。

● Include constant in equality 复选框，选中表示在回归方程包含常数项。

● Prediction Lines 栏，设置关于置信区间的显示选项：Mean 复选框，表示显示均值的置信区间；Individual 复选框，表示显示散点的置信区间；Confidence Interval 输入框，指定置信区间的置信水平，默认为 95 (%)。

● Fit lines for 栏，设置回归的方式：Total 复选框，表示对所有散点生成一条回归线；Subgroups 复选框，表示为分组变量的每个取值分别生成一条回归线。

(3) Mean，表示做均值拟和，参数设置参考 Regression 方法。

(4) Smoother，表示做光滑拟和，选中后需要指定以下参数。

● Kernel 下拉列表，用于指定局部线性回归的核。



- Bandwidth, 设置拟和曲线分别在 X1、X2 轴上的带宽。
- Use same bandwidth for all smoother 复选框, 表示所有拟和曲线均采用相同的 带宽。

### 3. 散点连线的参数选项

在图 19-123 中, 单击 Spike 标签, 弹出如图 19-125 所示的设置界面, 在此设置关于散点连线的参数。连线是指把散点云团投射到某个点、某个坐标轴或者某个面, 它可以帮助用户识别不同坐标轴上的数值, 通过连线的长度还可以方便地比较各点的距离。

(1) Spikes 列表框, 给出了如下 8 种连线方式。

- Origin 复选框, 表示从散点云团投射到每个原始数据点的连线。
- Corner 复选框, 表示从散点云团投射到某个焦点的连线。
- Total Centroid 复选框, 表示从散点云团投射到全部数据中心点的连线。
- Subgroup Centroid 复选框, 表示从细分组的散点云团投射到该分组中心点的连线。
- X1 Axis 复选框, 表示从散点云团投射到 X1 坐标轴的连线。
- Y Axis 复选框, 表示从散点云团投射到 Y 坐标轴的连线。
- Floor 复选框, 表示从散点云团投射到坐标轴平面的连线。
- Fit Line 复选框, 表示从散点云团投射到拟和曲线或拟和曲面的连线。

(2) Color spikes by color legend 复选框, 表示把分组连线的颜色与分类变量的图例匹配。

(3) Style spikes by Style legend 复选框, 表示把分组连线的样式与分类变量的样式匹配。

### 4. 输出图形

在图 19-124 中, 单击 Method 下拉列表并选中 Regression 选项; 在图 19-125 中, 勾选 Fit Line 复选框, 勾选 Color spikes by color legend 复选框; 单击 Assign 标签返回图 19-123 所示的变量设置界面。

在图 19-123 中, 单击确定按钮运行, SPSS Viewer 窗口的输出图形如图 19-126 所示, 它同时显示了散点分布、拟和曲线、Spikes 连线 3 个成分, 比较全面地反应了原始数据的特征。



图 19-125 交互式散点图的连线选项设置

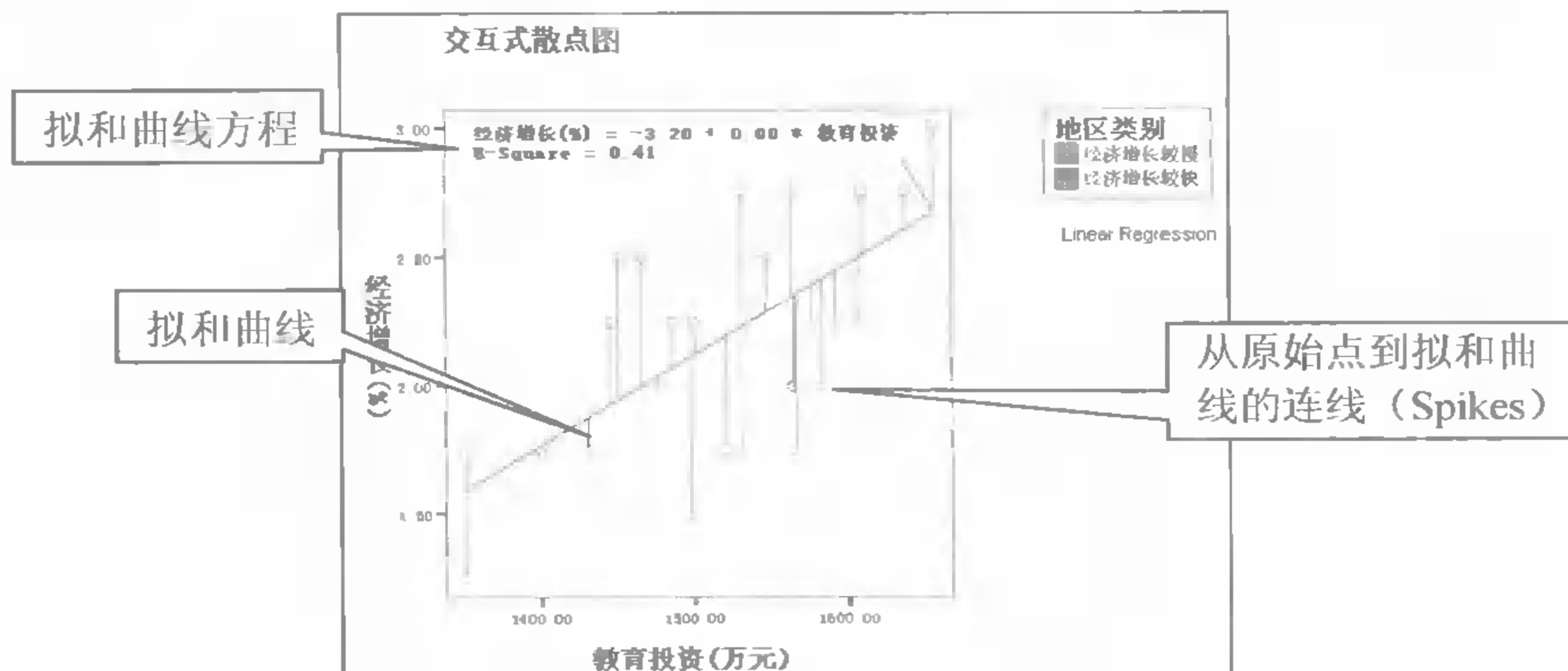


图 19-126 经济增长对教育投资的交互式散点图

### 19.11.4 用对话框创建散点图

依次单击菜单“Graphs→Legacy Dialogs→Scatter/Dot...”，打开利用对话框创建散点图的类型选择界面，如图 19-127 所示。

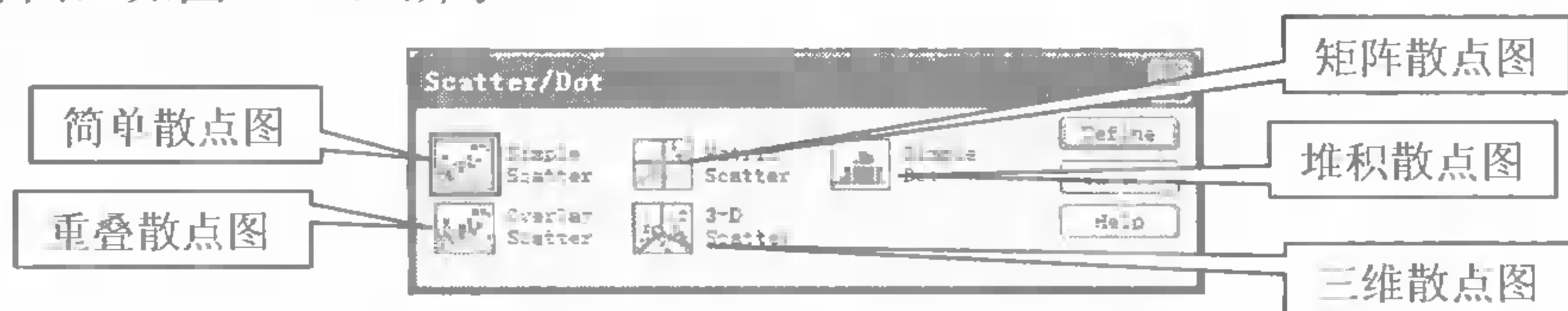


图 19-127 建立散点图的类型选择界面

简单散点图、三维散点图、堆积散点图的操作界面，均与图 19-30 或图 19-40 相似，它们的参数选项也大都在第 19.11.2 节的图形构建器中有所对应。

下面以重叠散点图为例，介绍如何用对话框建立散点图。本节仍然使用图 19-114 所示的“教育投资与经济增长.sav”文件来作图。

#### 1. 参数设置

在图 19-127 中，单击选中 Overlay Scatter 图标；单击 Define 按钮进入作重叠散点图的界面，如图 19-128 所示。

在变量列表同时选中经济增长和教育投资，单击从上至下第一个  按钮，将其作为一对作图变量选入 Y-X pairs 列表框，单击 Swap Pairs 按钮更改它们的顺序；用同样的方法选入变量对“学生增长-教育投资”。

#### 2. 输出图形

在图 19-128 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出如图 19-129 所示，可见它与用 Chart Builder 所作的重叠散点图基本相同。

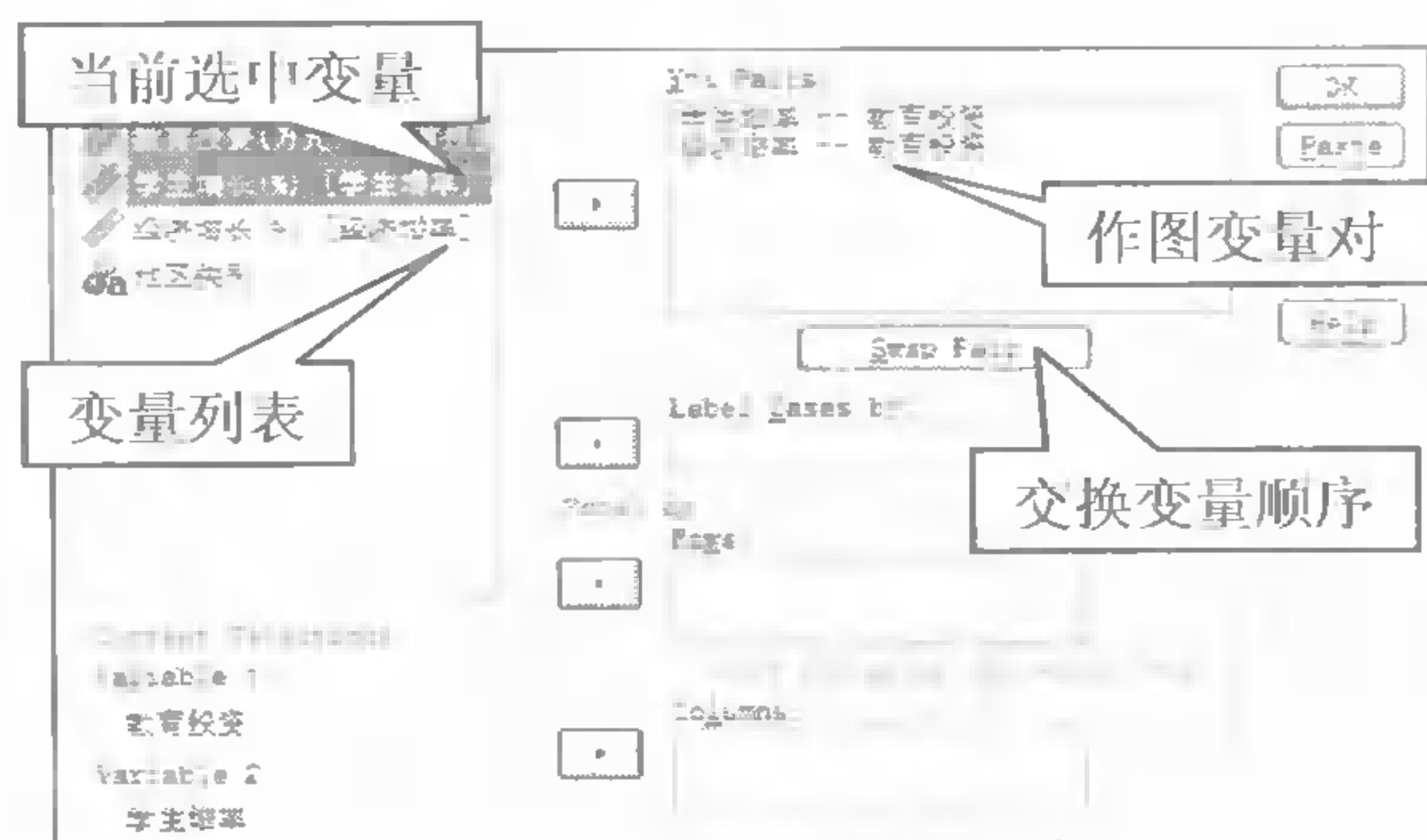


图 19-128 重叠散点图的参数设置

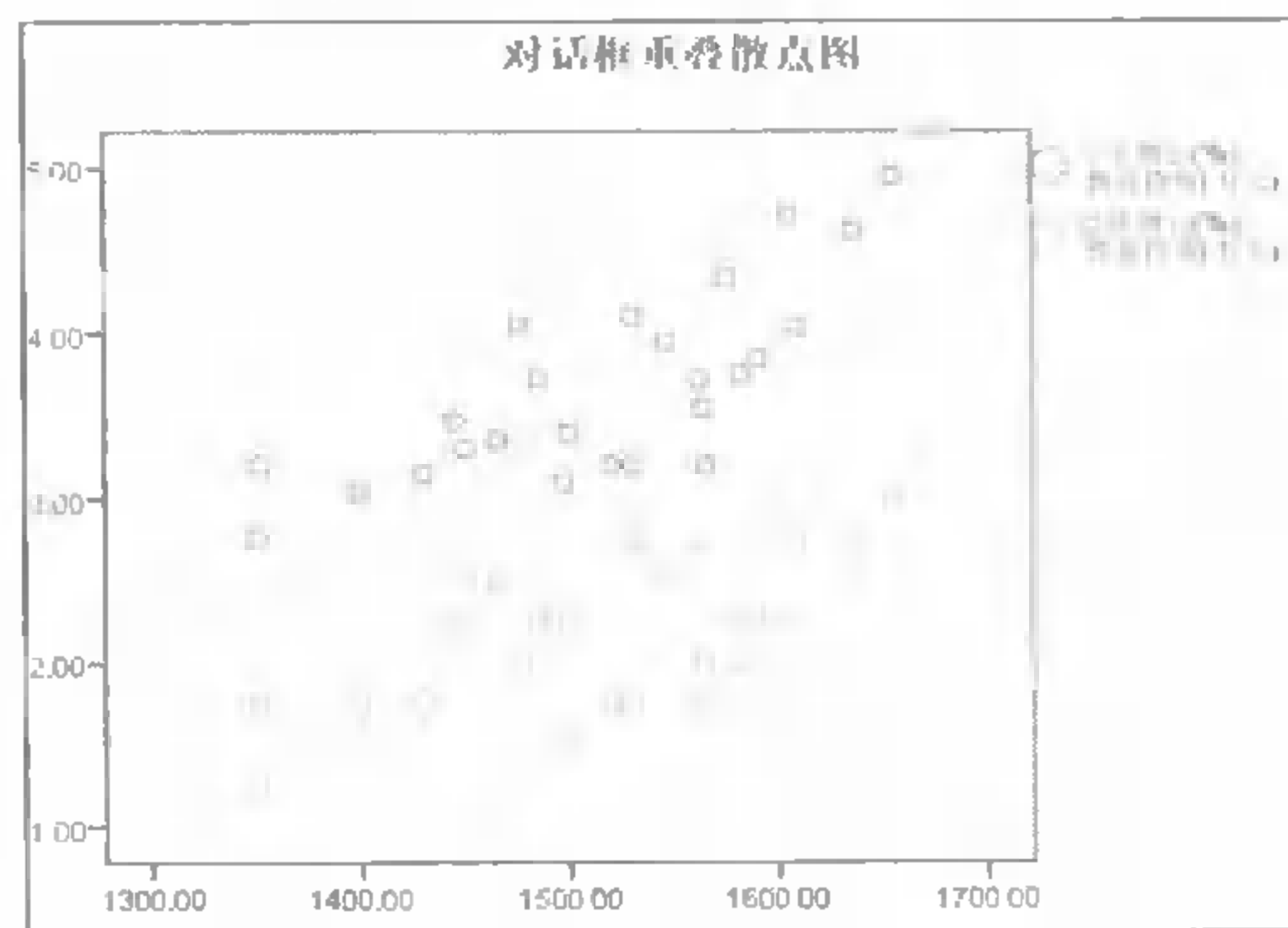


图 19-129 重叠散点图的输出

## 19.12 直方图

功能简述：直方图（Histogram）一种用无间隔的直条的长短，表现连续性变量的取值（或频数）分布特点的统计图形，图中每个条的高度都代表了相应组别的频数。

### 19.12.1 数据和问题描述

直方图与条形图的图形比较相似,本节对第 19.2 节曾使用的关于车祸次数的数据进行分析,通过创建直方图观察不同年龄段、不同性别的人发生车祸的次数分布有何特点。

所用数据文件为“autoaccident.sav”,数据格式如图 19-9 所示。

### 19.12.2 用图形构建器作直方图

依次单击菜单“Graphs→Chart Builder”打开图形构建器,如图 19-130 所示,单击 Gallery 标签;在 Choose from 列表框单击选中 Histogram,在其右侧列出预设的直方图图标。

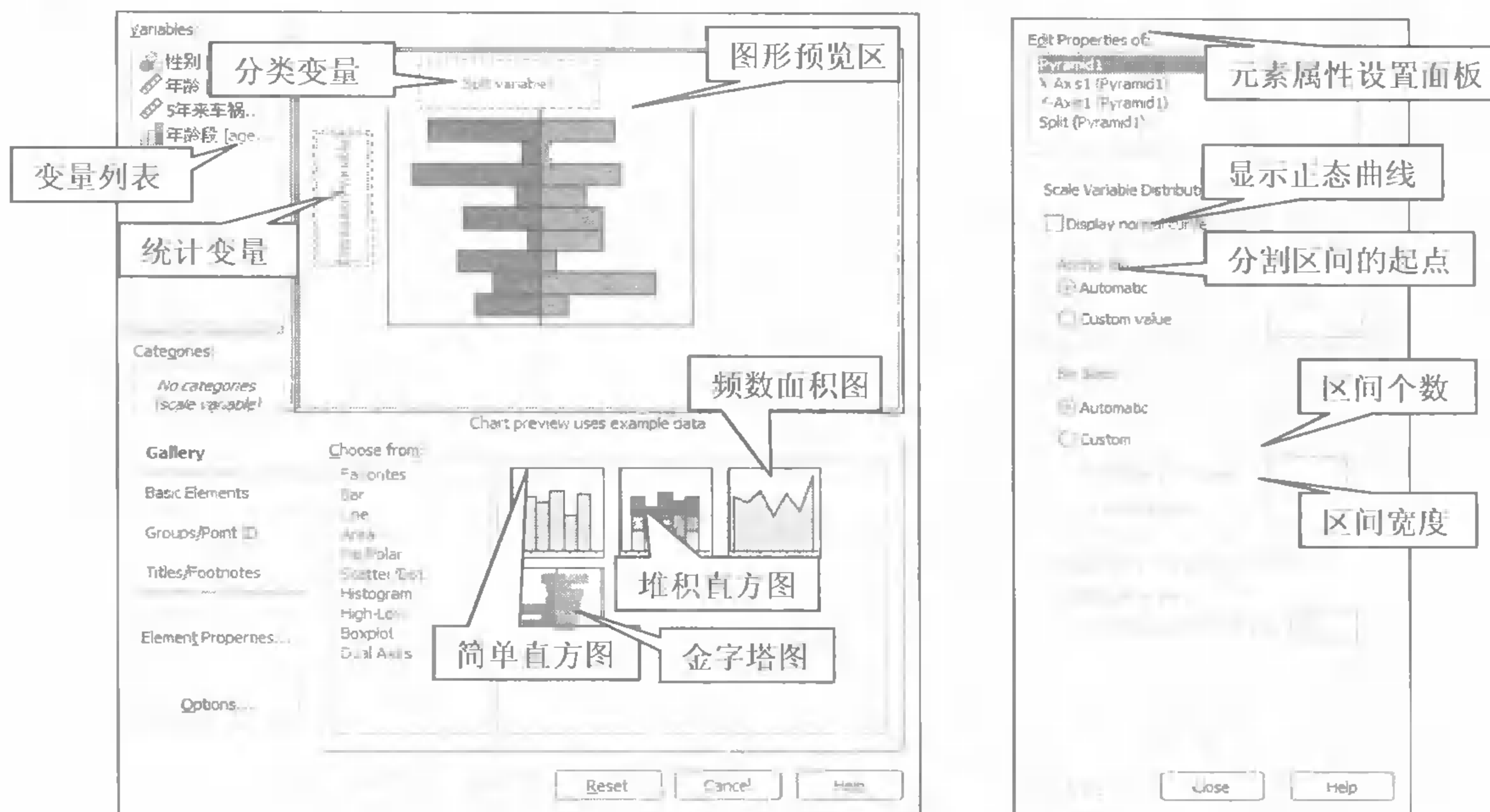


图 19-130 创建直方图的设置界面

简单直方图、堆积直方图和频数面积图的参数选项和设置方法,均与第 19.2.2 节介绍的条形图的设置相仿,输出的直方图也与相应的条形图相似。下面以金字塔图为例,来介绍如何通过图形构建器来建立直方图。

#### 1. 参数设置

在图 19-130 中双击预置图标 (Population Pyramid) 后,在图形预览区给出金字塔图的预览,同时自动弹出元素属性设置面板。

从变量列表中把车祸次数、性别变量,分别拖动至预览区的 Distribution Variable、Split Variable 虚线框中,将其分别作为金字塔图的统计变量和子分类变量。

#### 2. 输出图形

在图 19-130 中,单击 OK 按钮运行,SPSS Viewer 窗口的输出图形如图 19-131 所示。它非常直观地显示了不同性别的人 5 年来发生车祸次数的分布情况,可见随着车祸次数增加其出现频率呈指数衰减,鉴于这种特别的图形形状,称之为金字塔图。

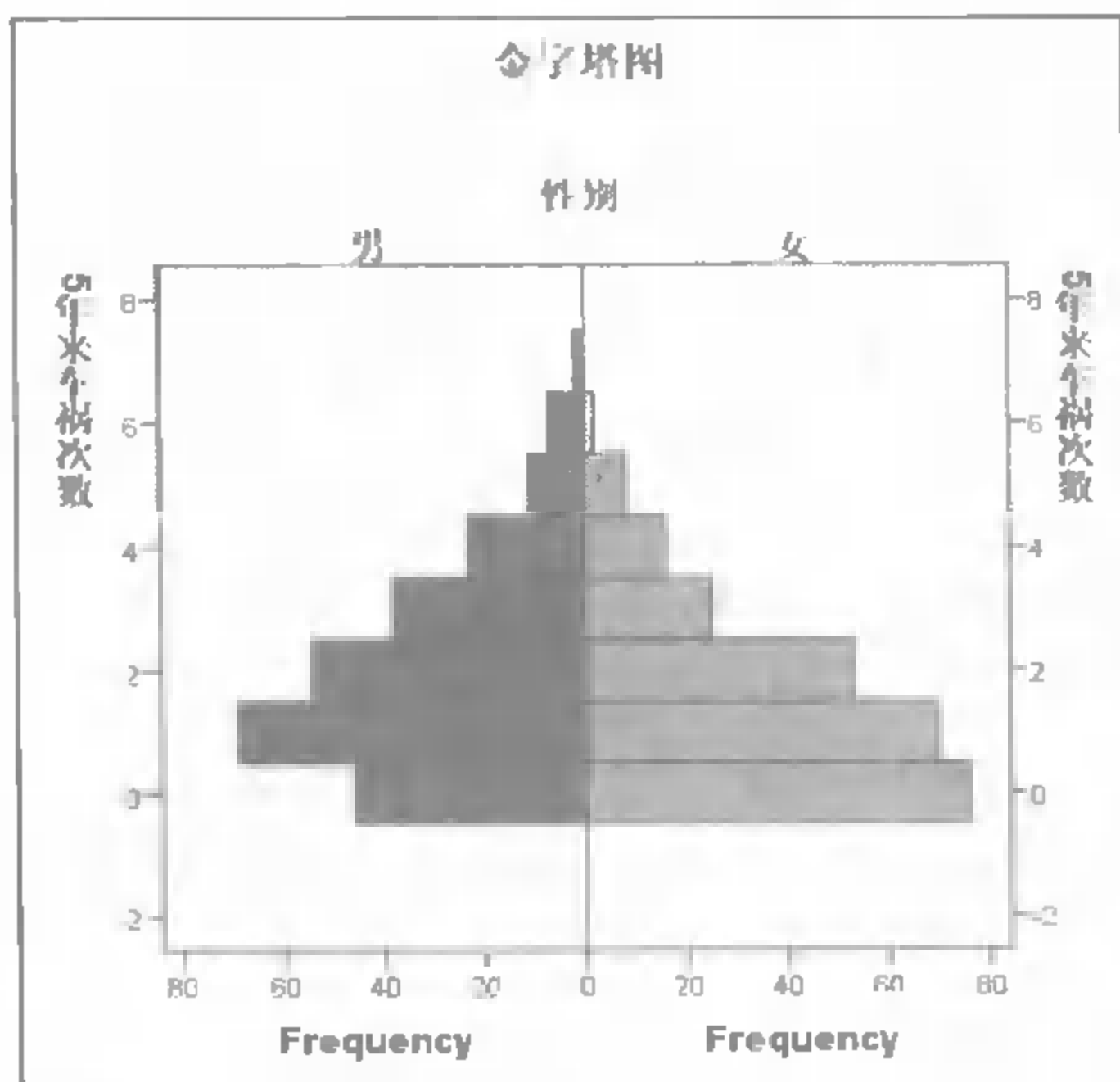


图 19-131 人口金字塔图输出结果

## 19.13 P-P 概率图

P-P 概率图主要用于验证样本数据是否服从某个指定的分布，它以样本的累计概率为横轴，以指定理论分布的累计概率为纵轴绘制散点图。如果样本来自于指定理论分布的总体，则所有散点应分散于从原点指向右上角的直线附近。

### 19.13.1 数据和问题描述

#### 1. 数据文件

本节使用 P-P 概率图来验证某高校学生单月支出的分布是否为正态的。所用数据文件为“高校学生月支出分布.sav”，数据格式如图 19-132 所示。

	Name	Type	Width	Decim	Label	Values	Missing	Column	Align	Measure
1	level	Numeric	8	2	支出水平	None	None	8	Right	Scale
2	freq	Numeric	3	0	支出频数	None	None	8	Right	Scale

图 19-132 某高校学生月支出分布

#### 2. 数据属性

此数据集为带有加权变量的样本类型，下面先来查看关于样本的加权信息。

依次单击菜单“Data→Weight Cases...”，打开为观测指定加权变量的设置对话框，如图 19-133 所示，Current 栏显示当前的加权变量为 freq（支出频数）；单击 Cancel 按钮返回 Data Editor 窗口。

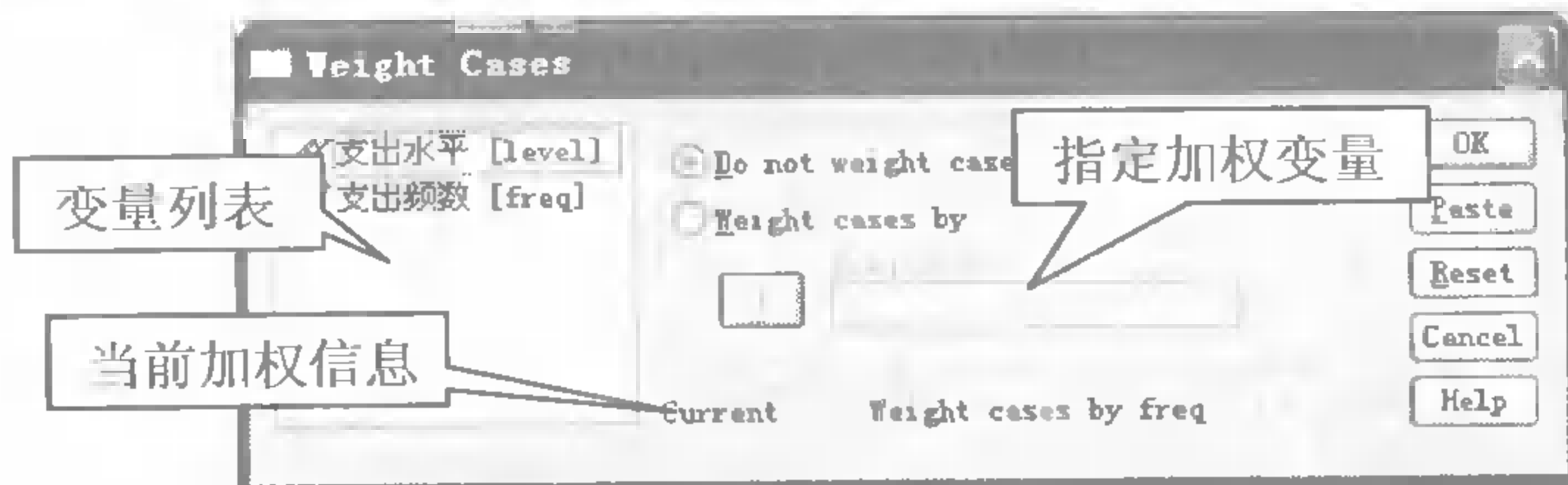


图 19-133 数据加权设置面板



## 19.13.2 用对话框创建帕 P-P 概率图

依次单击菜单“Analyze→Descriptive Statistics→P-P Plots...”打开用对话框创建 P-P 概率图的设置界面,如图 19-134 所示。

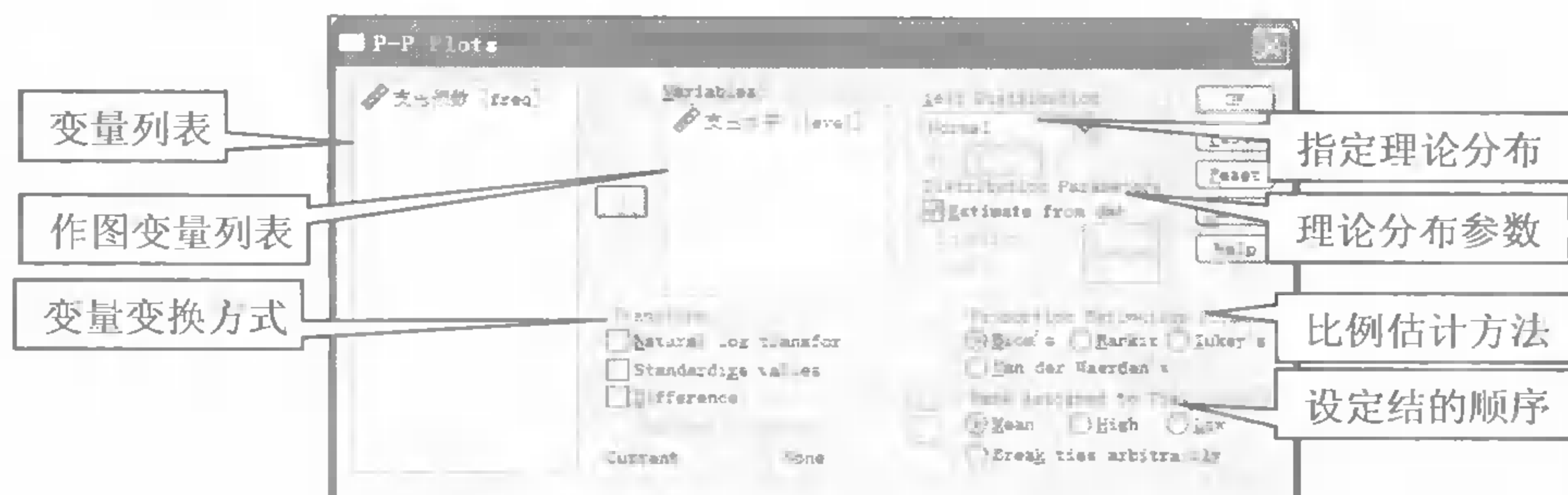



图 19-134 P-P 概率图设置面板

## 1. 参数设置

在变量列表中单击选中支出水平变量,单击  按钮,将其作为作图变量选入 Variables 列表框;在 Test Distribution 下拉列表保留默认的 Normal 选项,表示验证变量是否服从正态分布。下面对各参数选项的含义加以简单介绍。

(1) Transform 栏,指定对作图变量的变换方式,可选项有如下 4 种。

- Natural Log transform 复选框,表示作自然对数转换。
- Standardize values 复选框,表示作标准化转换,即转换为均值为 0 方差为 1 的样本。
- Difference 复选框,表示作差分转换,计算相邻两个变量值的差作为新的变量,后面的输入框用于指定差分的阶数。
- Seasonally difference 复选框,表示作季节差分转换,计算序列中恒定间距的两个取值的差作为新的变量,后面的输入框用于指定季节周期; Current 行显示当前周期。

(2) Test Distribution 下拉列表,用于指定待检验的理论分布,SPSS 给出了如下的可选项。

Beta (贝塔分布)、Chi-square (卡方分布)、Exponential (指数分布)、Gamma (伽马分布)、Half Normal (半正态分布)、Laplace (拉普拉斯分布)、Logistic (逻辑斯谛分布)、Lognormal (对数正态分布)、Normal (正态分布)、Students t (t 分布)、Uniform (均匀分布)、Weibull (威布尔分布)。df 输入框,用于设置关于指定分布的自由度。

(3) Estimate from data 复选框,选中表示从样本估计理论分布的有关参数,建议选中。

(4) Proportion Estimation Formula 栏,指定计算样本累计比率的估计方法,可选项有 4 个。

- Blom 单选框,使用公式  $\frac{r-3/8}{n+1/4}$  推算,其中: n 为观测个数, r 为样本的秩 (1~n)。
- Tukey 单选框,使用公式  $\frac{r-1/3}{n+1/3}$  推算。
- Rankit 单选框,使用公式  $\frac{r-1/2}{n}$  推算。
- Van der Waerden 单选框,使用公式  $\frac{r}{n+1}$  推算。

(5) Rank Assigned to Ties 栏,指定对结的处理方式,可选项有 4 个。

结 (tie)，指具有相同取值的多个观测组成的集合，此处设置结在所处样本中的秩次的计算方式：Mean，表示取平均秩序；High，表示取最高秩序；Low，表示取最低秩序；Break tie arbitrarily，表示取任意秩序。

2. 输出结果

在图 19-134 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如下。

(1) 模型摘要信息。如图 19-135 所示，“模型描述”表格给出了关于模型的摘要信息，包括检验变量、理论分布、参数来源。“个案处理摘要”表格给出了关于样本数据的使用情况。

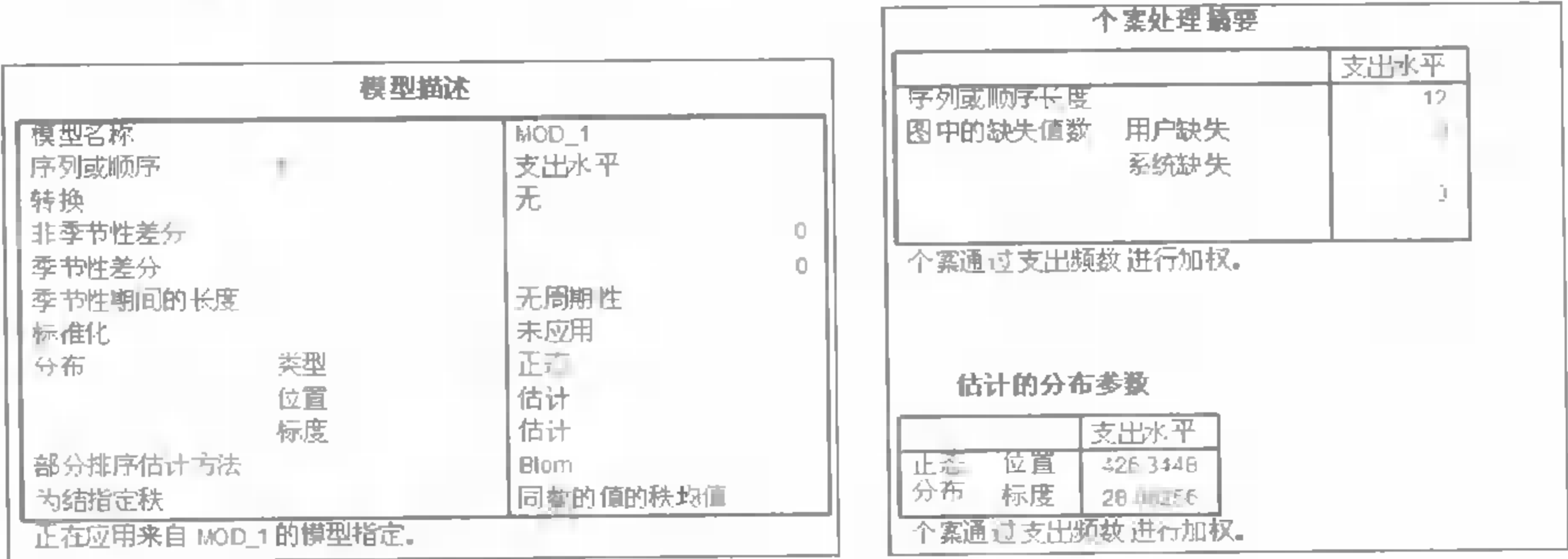


图 19-135 摘要信息和对理论模型的参数估计

(2) 理论分布的参数估计。如图 19-135 所示，“估计的分布参数”表格给出了用样本估计的正态分布的参数。

(3) P-P 图。如图 19-136 所示，“正态 P-P 图”是关于支出水平的 P-P 概率分布图，由于所有散点均分布于 45° 对角线的附近，故可以认为样本是服从正态分布的。“趋降正态 P-P 图”是剔除了趋势的 P-P 概率分布图，横轴为样本的累计概率，纵轴为样本累计概率和理论累计概率的差值，由于所有散点都均匀分布在过零点的横轴附近，故可认为样本服从待检验的理论分布。

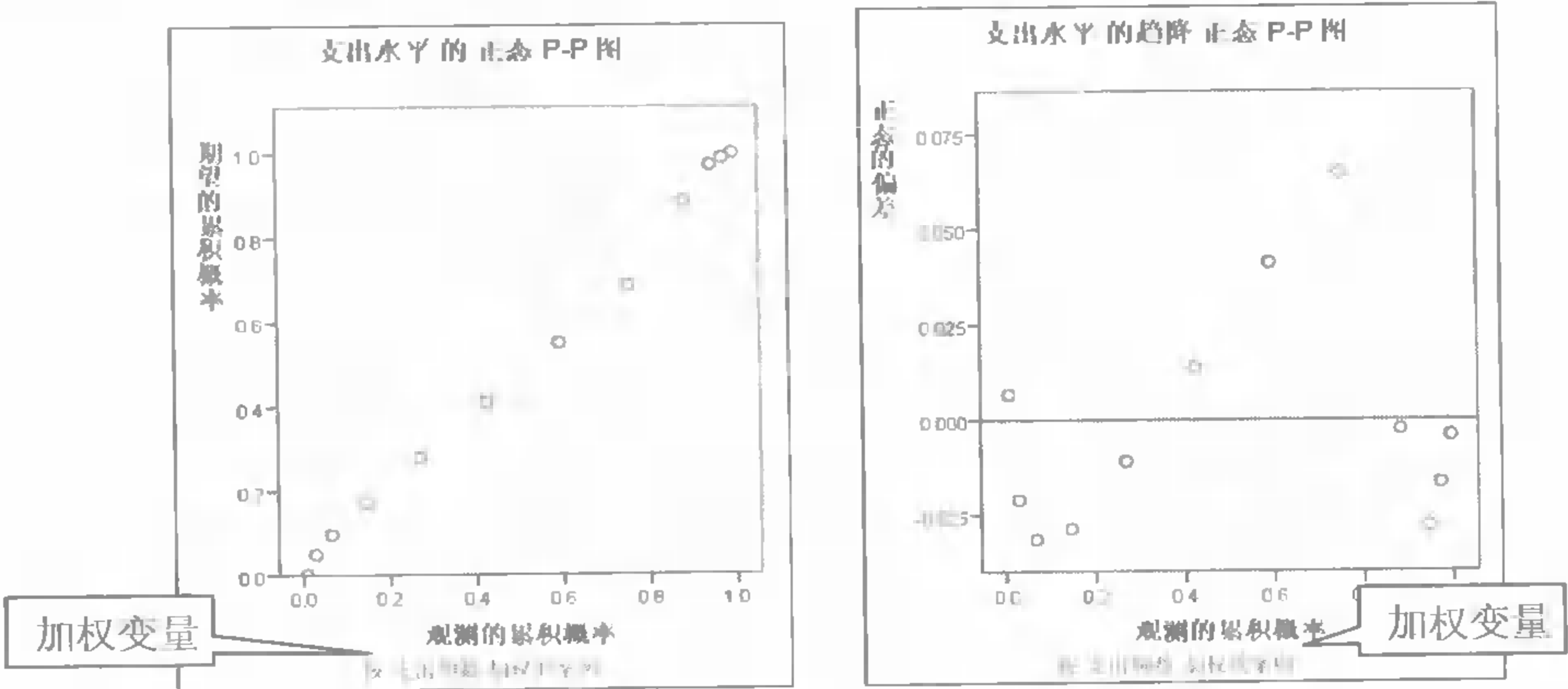


图 19-136 关于高校学生月支出的 P-P 概率图

19.14 Q-Q 概率图

功能简述：Q-Q 概率图主要用于验证样本数据是否服从某个指定的分布，原理与 P-P 图

相似，它以样本的分位数为横轴，以指定理论分布的分位数为纵轴绘制散点图。如果样本来自于指定理论分布的总体，则所有散点应分散于从原点指向右上角的直线附近。

### 19.14.1 数据和问题描述

本节使用 Q-Q 概率图来验证某高校学生的考试成绩是否服从于正态分布。所用数据文件为“考试成绩.sav”，数据格式如图 19-137 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	no	Numeric	7	0	学号	None	None	7	Right	Scale
2	name	String	5		姓名	None	None	5	Left	Nominal
3	x1	Numeric	3	0	系统分析	None	None	8	Right	Scale
4	x2	Numeric	3	1	货币银行	None	None	8	Right	Scale
5	x3	Numeric	3	0	国际贸易	None	None	8	Right	Scale
6	x4	Numeric	3	0	多媒体	None	None	8	Right	Scale
7	x5	Numeric	3	0	程序设计	None	None	8	Right	Scale

图 19-137 考试成绩数据文件各变量含义

### 19.14.2 用对话框创建 Q-Q 概率图

依次单击菜单“Analyze→Descriptive Statistics→Q-Q Plots...”打开用对话框创建 Q-Q 概率图的设置界面，如图 19-138 所示，它和图 19-134 所示的 P-P 图设置界面完全相同。

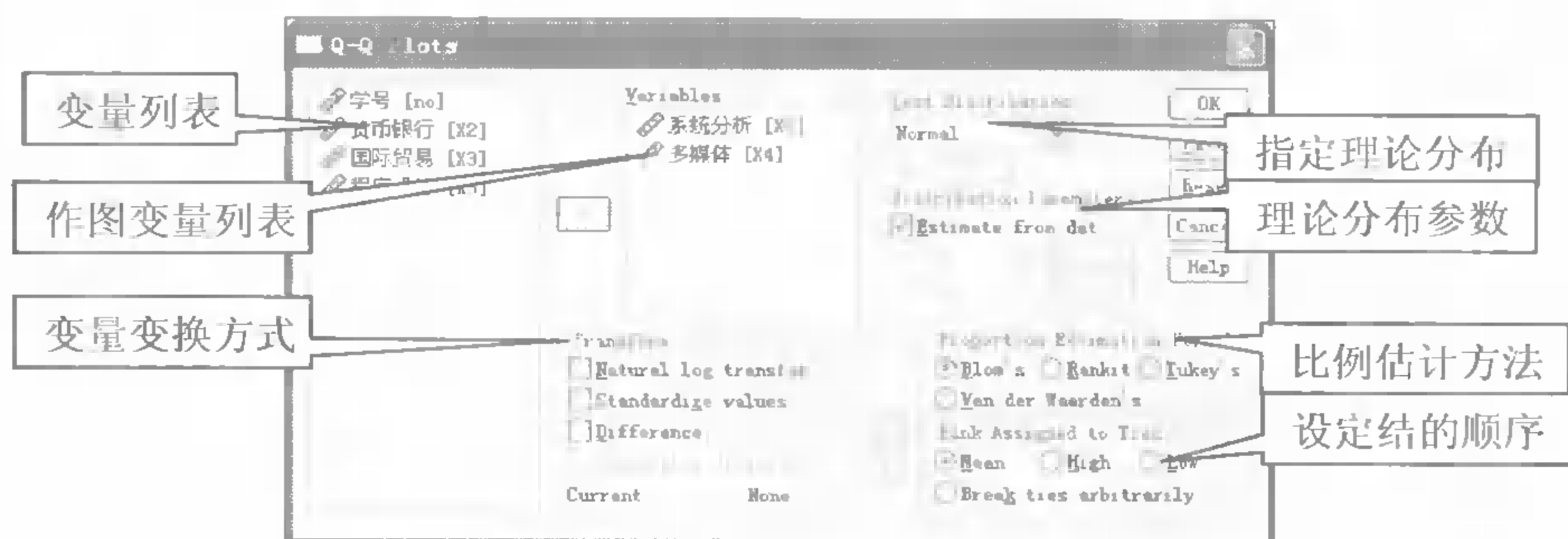


图 19-138 Q-Q 概率图设置面板

在变量列表中选中系统分析和多媒体两个变量，单击  按钮，将其作为作图变量选入 Variables 列表框。单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 19-139 和图 19-140 所示。

模型描述		个案处理摘要	
模型名称	MOD_3	系统分析	多媒体
序列或顺序	系统分析	序列或顺序长度	25
	多媒体	图中的缺失值数	0
转换	无	用户缺失	0
非季节性差分		系统缺失	0
季节性差分		个案未进行加权。	
季节性期间的长度			
标准化	无周期性		
分布	未应用		
类型	正态		
位置	估计		
标度	估计		
部分排序估计方法	Biom		
为结指定秩	同数的值的秩均值		
正在应用来自 MOD_3 的模型指定。			

估计的分布参数		系统分析	多媒体
正态分布	位置	82.16	89.20
	标度	7.255	6.442
个案未进行加权。			

图 19-139 摘要信息和参数估计信息

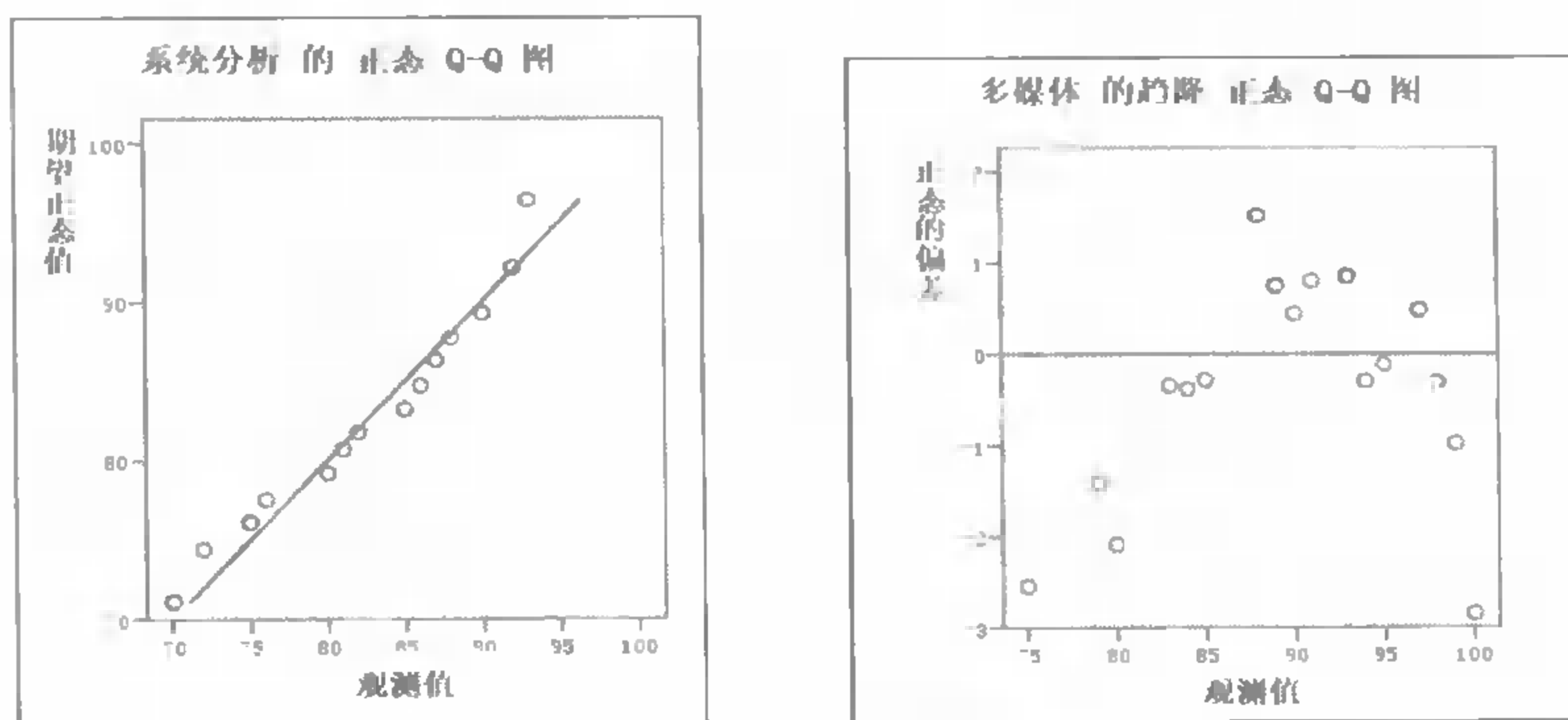


图 19-140 两个 Q-Q 概率分布图

如图 19-140 所示,“系统分析的正态 Q-Q 图”中所有散点都分布在 45 度对角线的附近,故可以认为系统分析课程的成绩是服从正态分布的。“多媒体的趋降正态 Q-Q 图”为剔除了趋势的 Q-Q 概率分布图,横轴为样本分位数,纵轴为样本分位数和理论分位数的差值,图中散点较均匀地分布在过零点的横轴附近,故可认为多媒体课程的成绩也是服从正态分布的。

## 19.15 时间序列图

时间序列 (Time Series) 图是反映测量指标随时间变化的统计图形,例如股市的日线变化、降雨量多年来的变化和改革开放后 GDP 的逐年变化等都可以用序列图描绘。SPSS 的时间序列分析模块提供了 4 种形式的时间序列图:普通序列图、自相关序列图、偏相关序列图和互相关序列图。

利用时间序列图,可以从动态角度认识事物的本质,例如研究几个时间序列之间的差别、识别时间序列的周期性和预测序列未来的走势等。

### 19.15.1 普通序列图

普通序列图 (Sequence Charts) 就是对数据集中的观测记录按照其当前顺序作图。它要求数据为时间序列数据或者已经按照有意义的顺序排列好了。

#### 1. 数据和问题描述

(1) 数据文件。本节所用数据文件为“航班人数变化数据.sav”,数据格式如图 19-141 所示,它记录的是关于某航班的单月旅客人数。其中 YEAR\_、MONTH\_、DATE\_ 为时间变量。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	YEAR_	Numeric	8	0	YEAR not periodic	None	None	10	Right	Ordinal
2	MONTH_	Numeric	2	0	MONTH period 12	None	None	8	Right	Ordinal
3	DATE_	String	8		Date Format MMM YYYY	None	None	10	Left	Nominal
4	pas	Numeric	8	0	旅客人数	None	None	8	Right	Scale
5	pas_1	Numeric	9	0	人数季节差分	None	None	11	Right	Scale

图 19-141 某航班月旅客人数数据

(2) 查看数据集的时间变量。依次单击菜单“Data→Define Dates...”打开建立时间变量的对话框,如图 19-142 所示,单击 Cancel 按钮返回 Data Editor 窗口。在此,Current Dates 栏显示当前的时间格式为“Years(1949) months(1:12)”,表示观测是从 1949 年 1 月开始逐月



记录的, 序列的周期为 12。

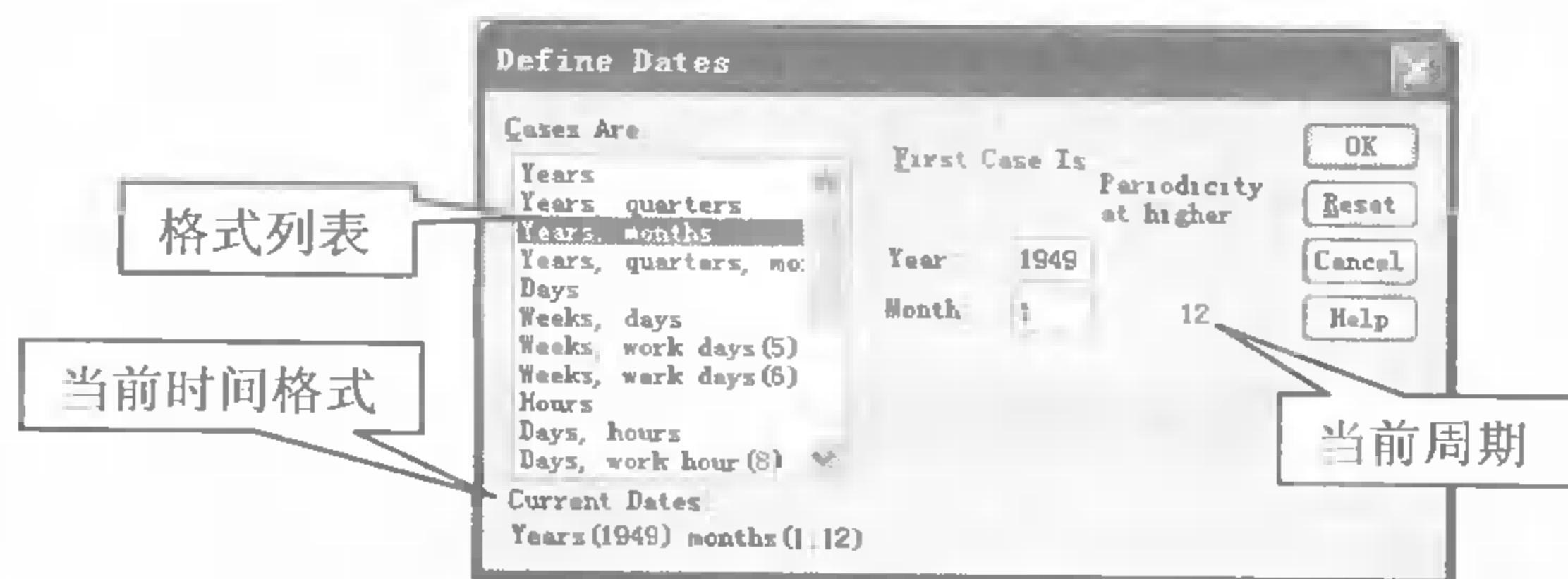


图 19-142 建立时间变量的对话框

(3) 对初始序列的季节差分。 $pas\_1$  (人数季节差分变量) 是由初始序列 ( $pas$ ) 经过一阶季节差分计算得到的, 计算公式为  $pas\_1 = SDIFF(pas\ 1)$ 。 $pas\_1$  是消除了季节影响的序列, 它能更真实地反映旅客人数的变化趋势。

## 2. 参数设置

依次单击菜单 “Analyze→Time Series→Sequence Charts...” 打开建立普通序列图的设置界面, 如图 19-143 所示。

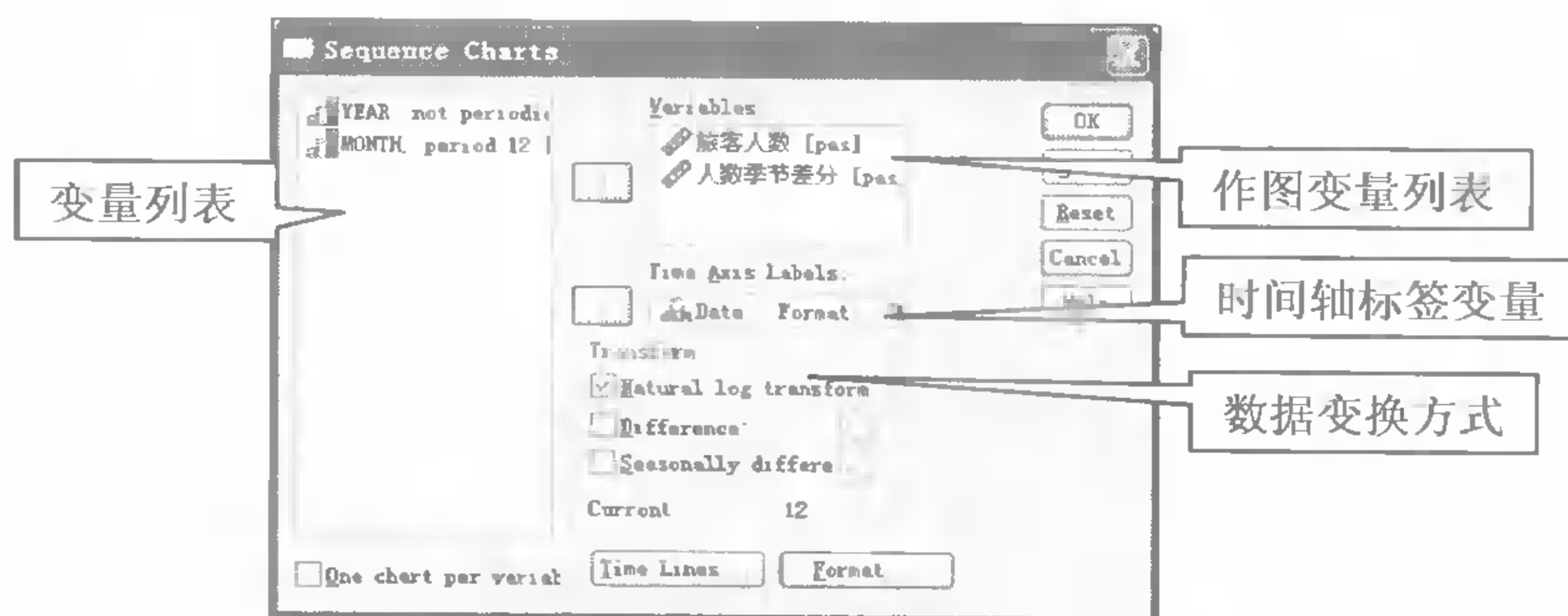


图 19-143 Sequence Charts 设置面板

(1) 指定作图变量和变换方式。在变量列表中选中旅客人数和人数季节差分两个变量, 单击从上至下第一个 按钮, 将其作为作图变量选入 Variables 列表框; 在变量列表中单击选中 Date 变量, 单击从上至下第二个 按钮, 将其作为时间轴变量选入 Time 选框; 勾选 Natural Log transform 复选框。

下面介绍各设置选项的具体含义。

① Variables 列表框, 用于选入作图变量。如果选入了多个变量, 将对其分别作图。

② Time Axis Labels list 选框, 用于选入时间轴分类变量, 可以是数值型、短字符型或长字符型的, 它用于在输出图形里标识时间轴。

③ One chart per variable 复选框, 如果指定了多个作图变量, 勾选此项表示对每个变量单独输出一图, 否则所有变量都将显示在同一个图形里。

④ Transform 栏, 指定对作图变量的变换方式, Current 行显示当前序列的周期。


● Natural Log transform 复选框, 表示作自然对数变换。

● Difference 复选框, 表示作差分变换, 后面的输入框用于指定差分的阶数。

● Seasonally difference 复选框, 表示作季节差分变换, 后面的输入框指定差分的阶数。

对于每个作图变量, 都是先依次作完指定的所有变换 (可能为多个) 后, 再对得到的最

终序列作图。

(2) Time Lines 选项的设置。在图 19-143 中单击 Time Lines 按钮,弹出如图 19-144 所示的子对话框,在此设置关于时间轴参照线的选项。单击选中 Lines at each Change of 单选框;在参照变量列表单击选中 YEAR 变量,单击  按钮,将其作为参照变量选入 Reference 选框;单击 Continue 按钮返回主界面。

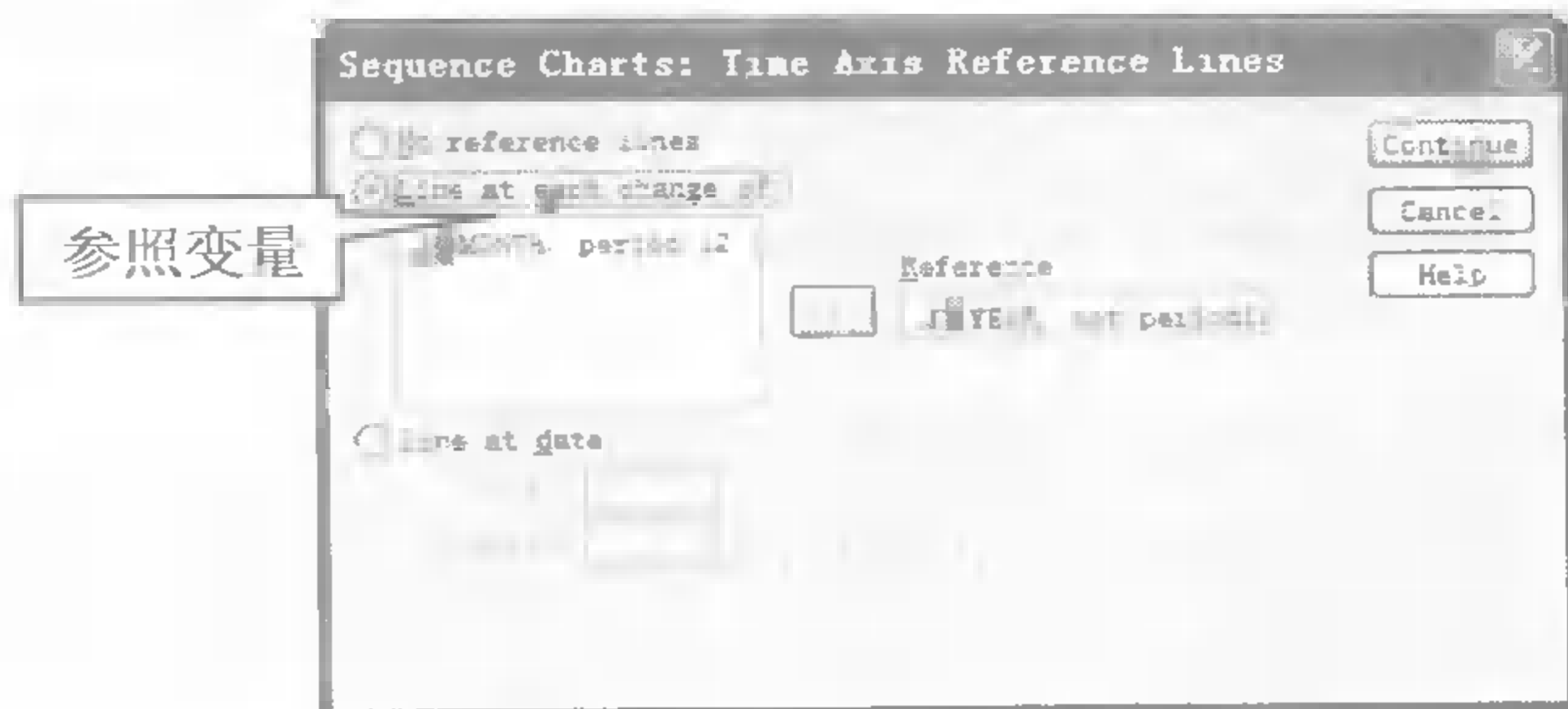


图 19-144 Time Lines 子设置对话框

下面介绍各设置选项的具体含义。

- ① No reference lines 单选项,表示无时间参照线,是默认选项。
- ② Lines at each change of 单选项,选中后把指定参照变量从下面的列表框选入右侧的 Reference 选框里。在输出图形里将按照这个变量的取值变化来定义参照线。
- ③ Line at date 单选项,表示只显示指定日期的参照线。Year、Month 输入框分别用于指定参照线的年和月。如果时间轴不是时间变量,此处只显示一个输入框,用于指定参照线取值。

(3) Format 选项的设置。在图 19-143 中单击 Format 按钮,弹出如图 19-145 所示的子对话框,在此设置关于图形显示方式的选项。单击 Continue 按钮返回主界面。

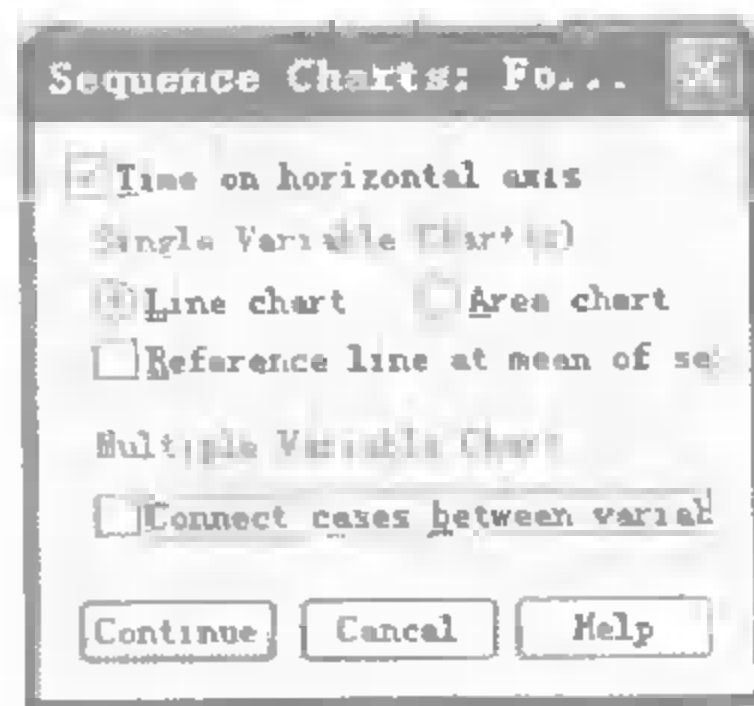


图 19-145 Format 子设置对话框

- ① Time on horizontal axis 复选框,勾选表示把时间轴作为横轴,否则时间轴将显示在纵轴上,默认为选中状态。
- ② Single Variable Chart(s)栏,设置关于简单图形(只含单个变量)的显示选项。
  - ① Line Chart 单选框表示作线形图;Area Chart 单选框表示作面积图。
  - ② Reference line at mean of series 复选框,表示在图中显示序列均值的参照线。
- ③ Multiple Variables Chart 栏,设置关于复合图形(含有多个变量)的显示选项。勾选 Connect cases between variables 复选框,表示把相同时间点上不同序列的取值用线段连接起来。

### 3. 输出结果

在图 19-143 中单击 OK 按钮运行,SPSS Viewer 窗口的输出结果如图 19-146 和图 19-147 所示。

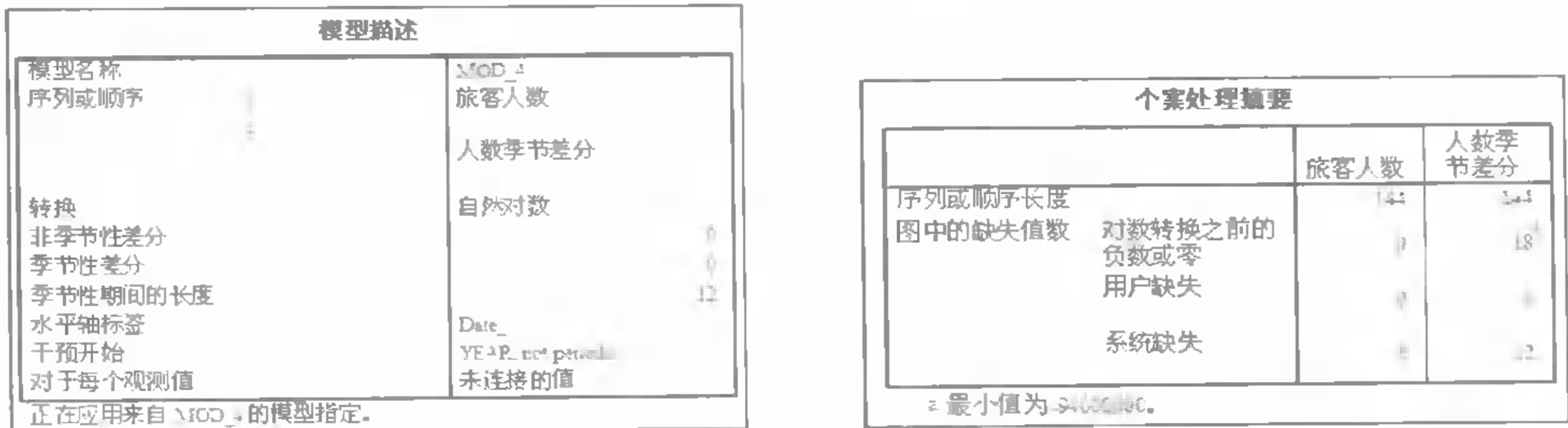


图 19-146 模型和个案摘要信息

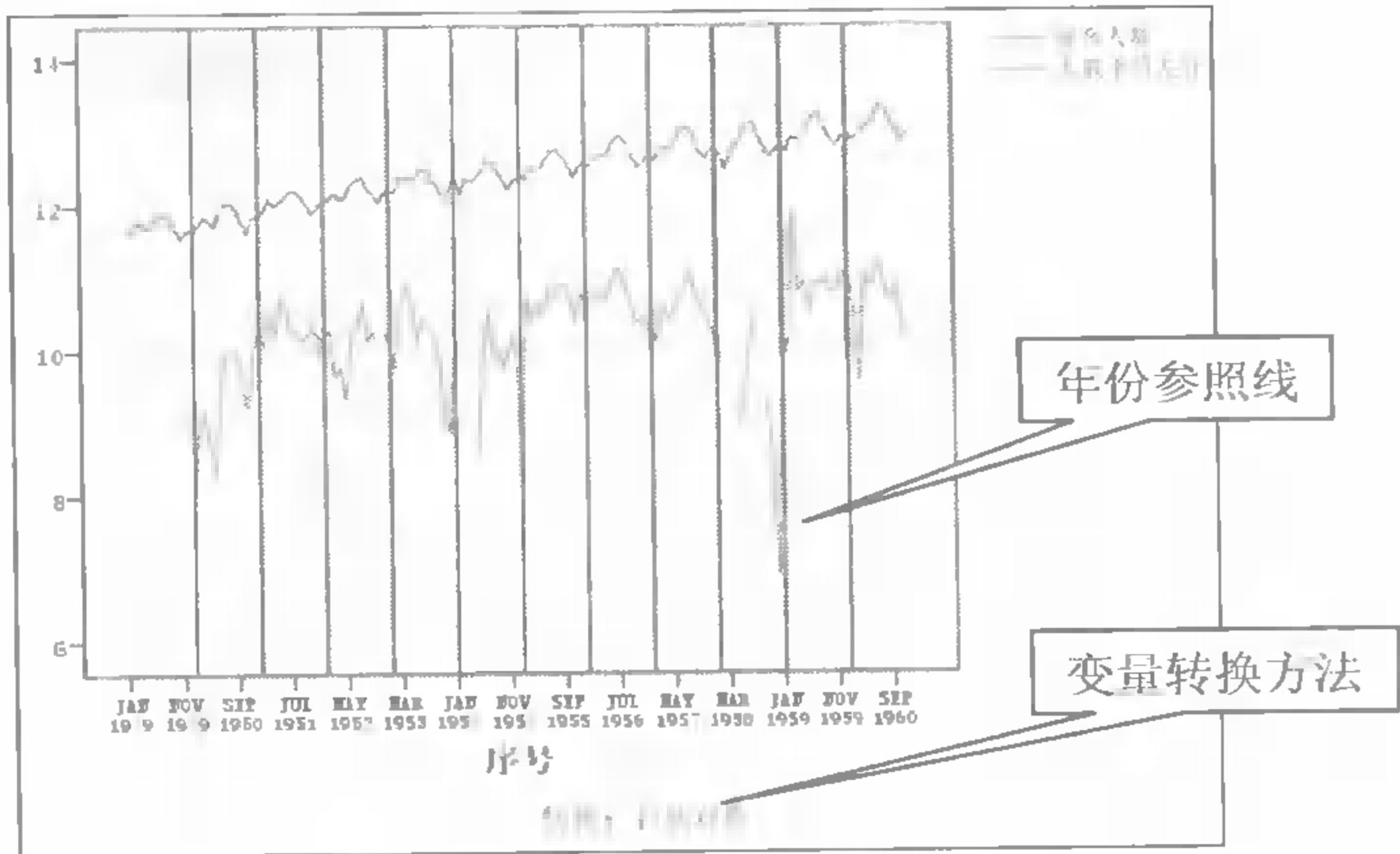


图 19-147 普通序列图的输出结果

- (1) 模型和个案摘要信息。如图 19-146 所示，“模型描述”表格给出了关于模型的各种参数设置信息；“个案处理摘要”表格给出了有关数据集的使用信息，可见在对差分变量作对数转换时有 18 个非法数值。
- (2) 普通序列图。如图 19-147 所示，是关于原始序列和季节差分后序列的图形。可见去除季节因素的影响后，序列的波动性明显放大了。

### 19.15.2 自相关序列图

自相关（autocorrelation）序列图和偏相关（partial autocorrelation）序列图是分别用于描绘时间序列的自相关函数（ACF）、偏相关函数（PACF）的图形。

#### 1. 有关概念


ACF 计算的是原始序列  $x_t$  和延迟序列  $x_{t-n}$  之间的相关系数， $n=1,2,\dots$ 。

PACF 计算的是原始序列  $x'_t = x_t - P_{\{x_{t-1}, \dots, x_{t-(n-1)}\}} x_t$  和延迟序列  $x'_{t-n} = x_{t-n} - P_{\{x_{t-1}, \dots, x_{t-(n-1)}\}} x_{t-n}$  之间的相关系数，其中  $P_{\{x_{t-1}, \dots, x_{t-(n-1)}\}} x_t$  表示  $x_t$  在  $\{x_{t-1}, \dots, x_{t-(n-1)}\}$  上的投影（最优线性组合），而  $x'_t$  就表示  $x_t$  消除了间隔数据  $\{x_{t-1}, \dots, x_{t-(n-1)}\}$  影响之后的序列。

#### 2. 参数设置

本节继续对如图 19-141 所示的航班月旅客人数数据集进行作图分析，观察旅客人数的自

相关序列图和偏相关序列图。

依次单击菜单“Analyze→Time Series→Autocorrelations...”打开建立相关序列图的设置界面，如图 19-148 中的主界面所示。在变量列表中单击选中旅客人数变量，单击  按钮，将其作为作图变量选入 Variables 列表框；勾选 Natural Log transform 复选框；勾选 Seasonally difference 复选框，在后面的输入框键入“1”；单击 Options 按钮弹出右侧的 Options 子对话框，单击 Continue 按钮返回主界面。

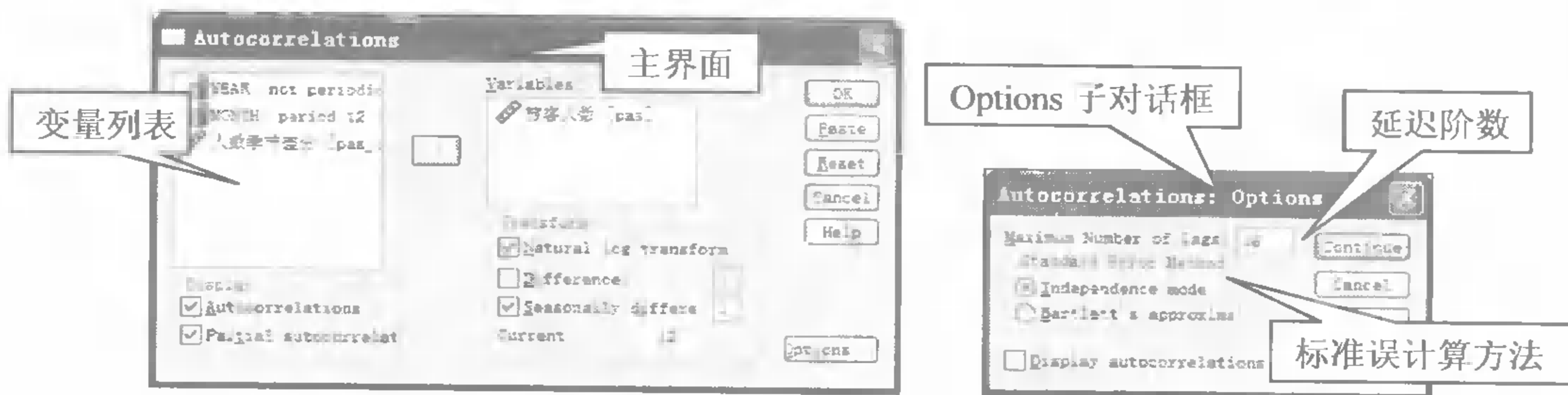


图 19-148 Autocorrelations 作图的参数设置

(1) Display 栏，选择要输出的图形类型，可选项有两个。Autocorrelation 复选框表示自相关序列图；Partial autocorrelation 复选框表示偏相关序列图。

(2) Transform 栏，指定对作图变量的转换方式，与图 19-143 中的 Transform 设置相同。

(3) Options 子对话框的设置

① Maximum number of lags 输入框，指定自（偏）相关函数的最大延迟阶数，默认值为 16。

② Standard Error Method 栏，指定计算标准误差的方法，只适用于自相关系数（ACF）。

● Independence model 单选框，此方法假设数据为白噪声序列。

● Bartlett's approximation 单选框，此方法适用于 k-1 阶的滑动平均序列，由此计算的标准误差会随着延迟阶数增加而递增。

③ Display autocorrelations at period 复选框，勾选它表示只输出延迟阶数为序列周期长度时的自（偏）相关序列。

### 3. 输出图形

在图 19-148 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 19-149～图 19-151 所示。

模型描述	
模型名称	MOD_5
序列名	旅客人数
转换	自然对数
非季节性差分	
季节性差分	
季节性期间的长度	
最大滞后数	16
为计算自相关的标准误差而假定的过程	独立性（白噪音） <sup>a</sup>
显示并绘图	所有滞后
正在应用来自 MOD_5 的模型指定。	
<sup>a</sup> 不适用于计算偏自相关的标准误差。	

个案处理摘要		
序列长度		144
缺失值数	对数转换之前的负数或零	0
	用户缺失	0
	系统缺失	0
有效值数		144
由于差分而丢失的值数		12
差分后的可计算第一滞后数		131

图 19-149 模型和个案摘要信息



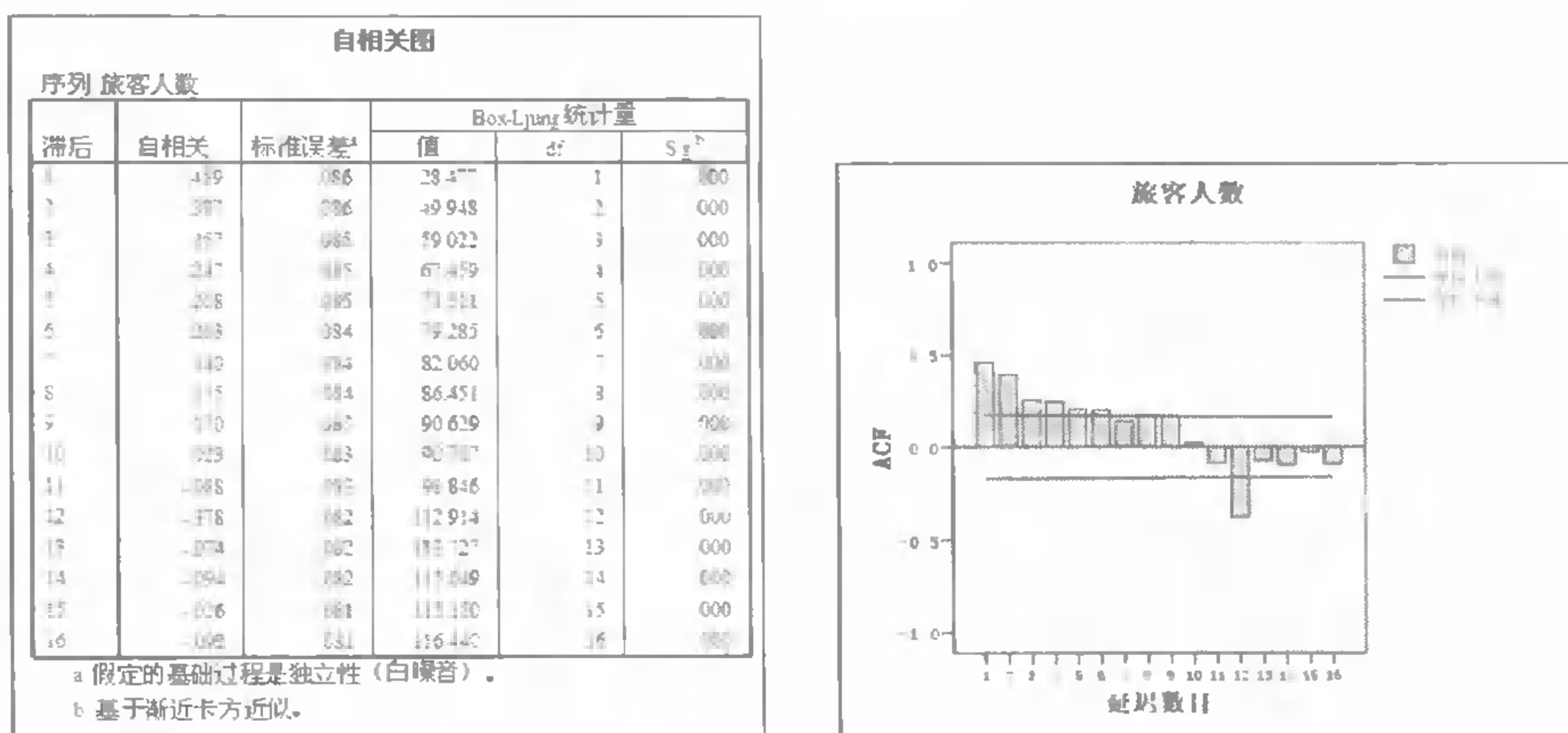


图 19-150 自相关序列图

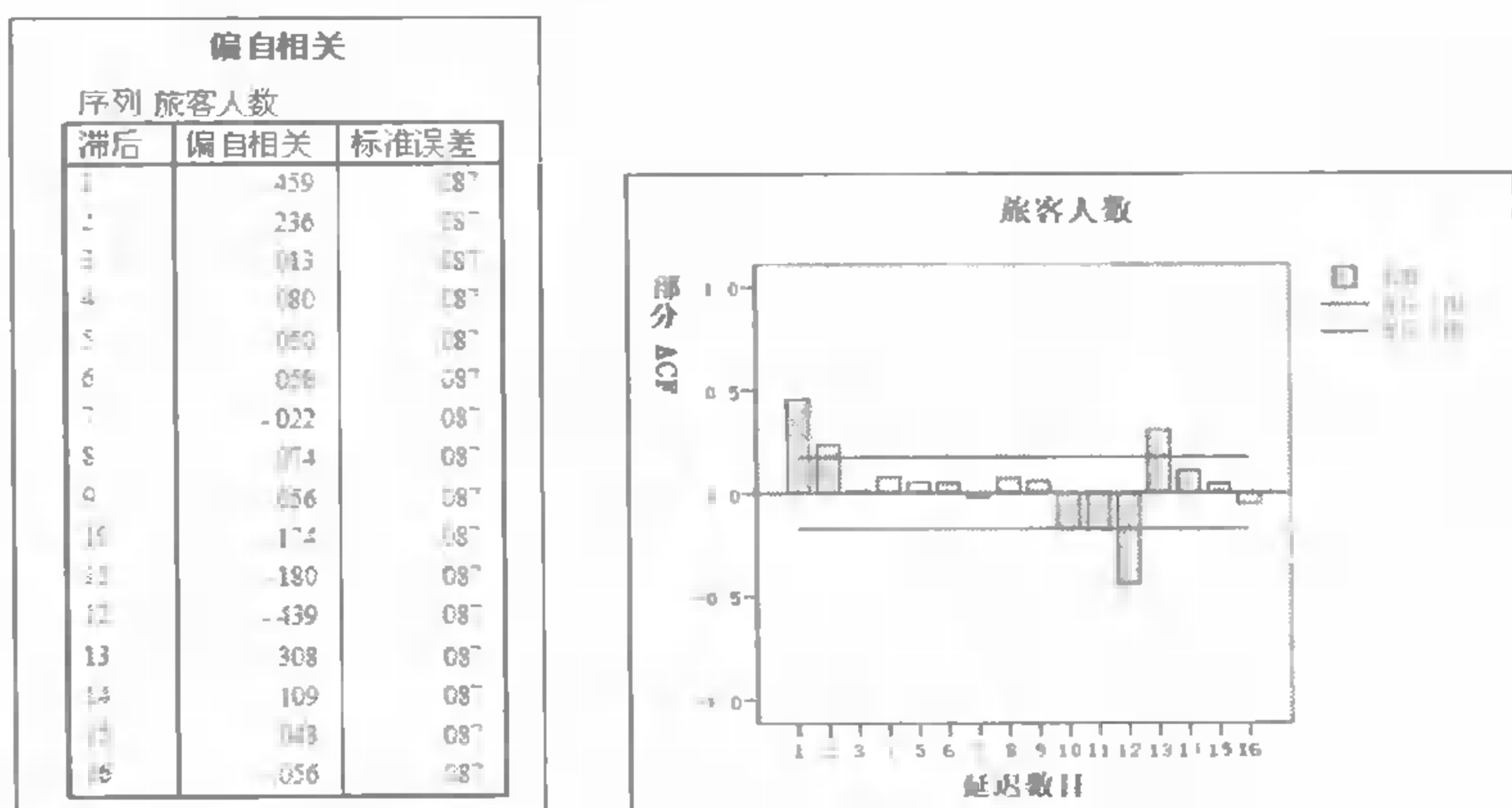


图 19-151 偏相关序列图

(1) 模型和个案摘要信息。如图 19-149 所示,“模型描述”表格给出了关于模型的各种参数设置信息;“个案处理摘要”表格给出了有关数据集的使用信息。

(2) 自相关序列图。如图 19-150 所示,旅客人数的自相关函数在一个周期内(12 个月)有较为明显的拖尾现象;而且延迟阶数为序列周期(12)时的自相关值较为异常。

(3) 偏相关序列图。如图 19-151 所示,旅客人数的偏相关函数在一个周期内(12 个月)有较为明显的截尾现象;而且延迟阶数为序列周期(12)时的偏相关值较为异常。

### 19.15.3 互相关序列图

互相关函数表示两个时间序列之间的相关函数,用于表现不同序列之间的相关关系。SPSS 的 Cross-correlations 过程可以输出正、负和零延迟的互相关系数,它只适用于时间序列数据。


#### 1. 数据和问题描述

本节利用互相关序列图来分析 1982~1993 年的人均收入与人均消费之间的互相关关系。所用数据文件为“收入与消费互相关分析.sav”,数据格式如图 19-152 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	year	String	6		年份	None	None	6	Left	Nominal
2	x	Numeric	6	2	人均国民收入 (元)	None	None	7	Right	Scale
3	y	Numeric	6	0	人均消费 (元)	None	None	6	Right	Scale

图 19-152 1982~1993 年人均收入与消费的数据格式

## 2. 参数设置

依次单击菜单“Analyze→Time Series→Cross-correlations...”打开建立互相关序列图的设置界面,如图 19-153 中的主界面所示。在变量列表中选中人均国民收入和人均消费,单击  按钮,将其作为作图变量选入 Variables 列表框;单击 Options 按钮弹出右侧的 Options 子对话框,单击 Continue 按钮返回主界面。

在此,主界面中 Transform 栏和 Options 子对话框的参数选项,都与图 19-148 所示作自相关序列图时的有关设置内容相同。本例均采用默认选项。

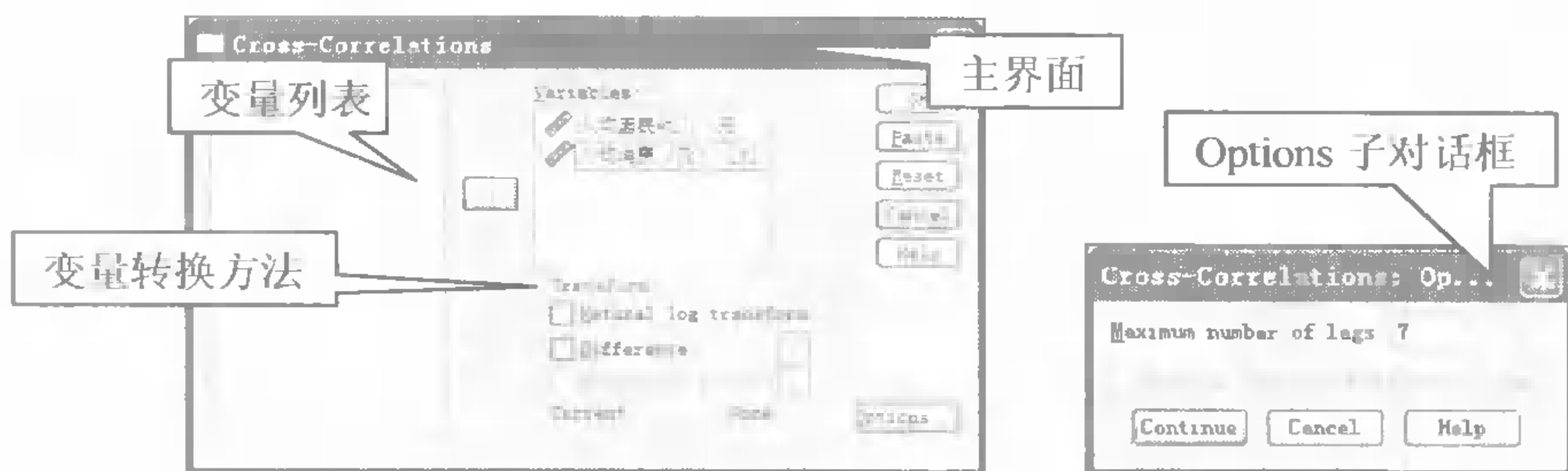


图 19-153 Cross-correlations 作图的参数设置

## 3. 输出结果

在图 19-153 中,单击 OK 按钮运行,SPSS Viewer 窗口的输出结果如下。

如图 19-154 所示,“模型描述”表格给出了关于模型的各种参数设置信息;“个案处理摘要”表格给出了有关数据集的使用信息。

模型描述		
模型名称	MOD_1	
序列名	人均国民收入 (元)	
	人均消费 (元)	
转换	无	
非季节性差分		
季节性差分		
季节性期间的长度	无周期性	
滞后范围	从	至
显示并绘图	所有滞后	
正在应用来自 MOD_1 的模型指定。		

个案处理摘要		
序列长度		15
由于以下原因排除的个案数	用户缺失值	0
	系统缺失值	0
有效个案数		15
差分后可计算的零阶相关数		11

图 19-154 模型和个案摘要信息

如图 19-155 所示,观察互相关序列图可知,两个序列在零延迟时相关性最强(接近 1),随着延迟阶数的增加它们的相关性呈指数递减。

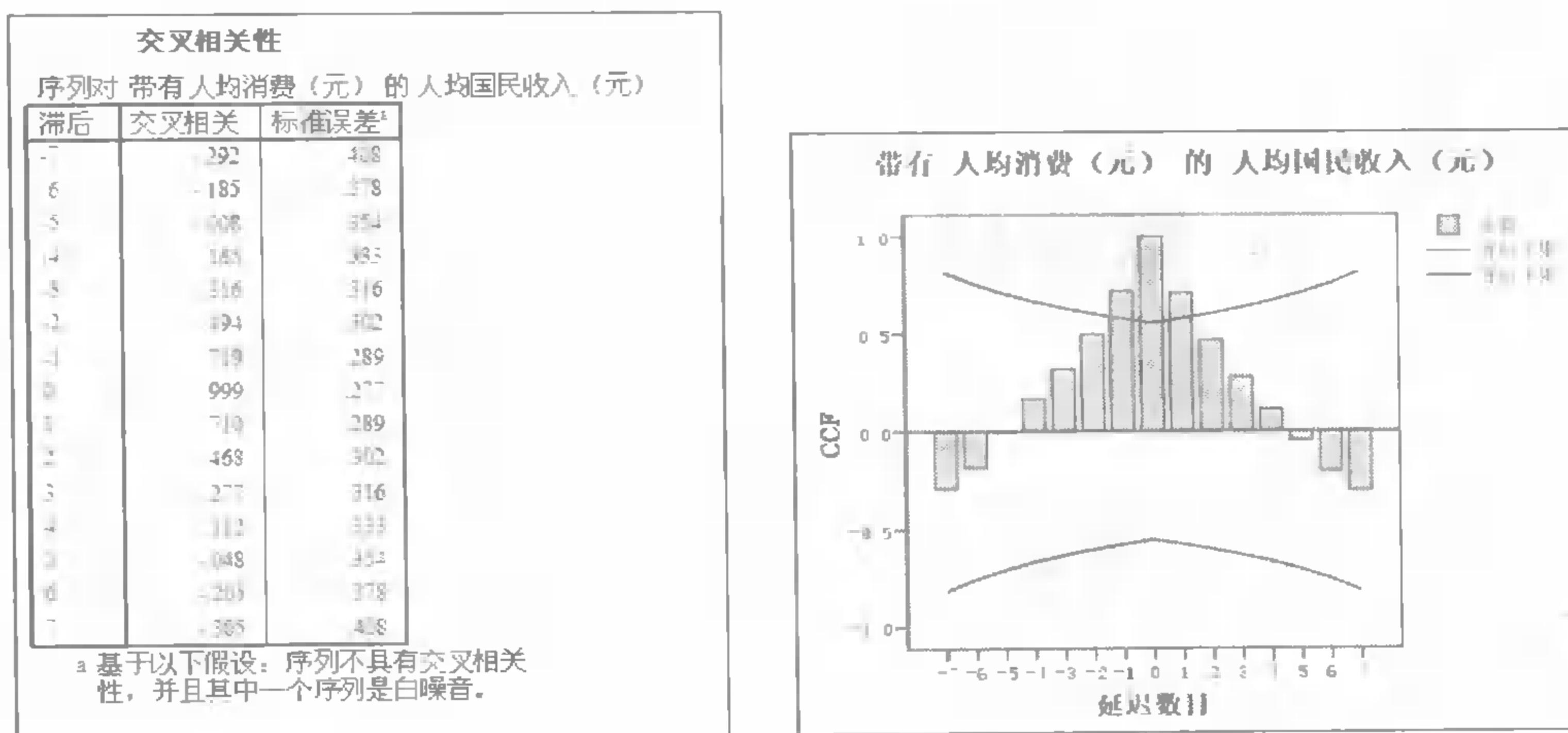


图 19-155 互相关序列图

## 19.16 双轴线图

有时需要把可以分开作的图综合显示在同一个图形中，以便比较不同对象之间的相互关系，这种图形都是复合图形，例如要在同一个图里显示不同年龄人士的体重信息和身高信息。这时就可能遇到随后的问题，不同作图对象的度量单位不同（如身高和体重）或者数量级不同（比如非比例数值和比例数值），通常的图形无法同时显示这些不一致的变量信息。双轴线图可以用来解决这个问题，它在一个图里给出两个纵坐标轴，分别用来刻画不同的变量（有度量单位或数量级冲突的变量）。

### 19.16.1 数据和问题描述

本节使用双轴线图来描绘 10~13 岁儿童的身高和体重特征，所用数据文件为“儿童身高体重数据.sav”，数据格式如图 19-156 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	no	String	4		编号	None	None	4	Right	Nominal
2	gend	String	4		性别	{0 男}	None	4	Right	Nominal
3	age	Numeric	2	0	年龄	None	None	3	Right	Ordinal
4	high	Numeric	4	2	身高	None	None	8	Right	Scale
5	weight	Numeric	2	0	体重	None	None	6	Right	Scale

图 19-156 儿童体重和身高数据

### 19.16.2 用图形构建器作双轴线图

依次单击菜单“Graphs→Chart Builder”打开图形构建器，如图 19-157 所示，单击 Gallery 标签；在 Choose from 列表框中单击选中 Dual Axes，在其右侧列出预设的双轴线图图标。其中横轴为分类变量的情况，适用于条形图和折线图相结合的方式；横轴为连续变量的情况，适用于散点图相结合的方式；这两种图形的设置方式相似，本节以前者为例进行介绍。

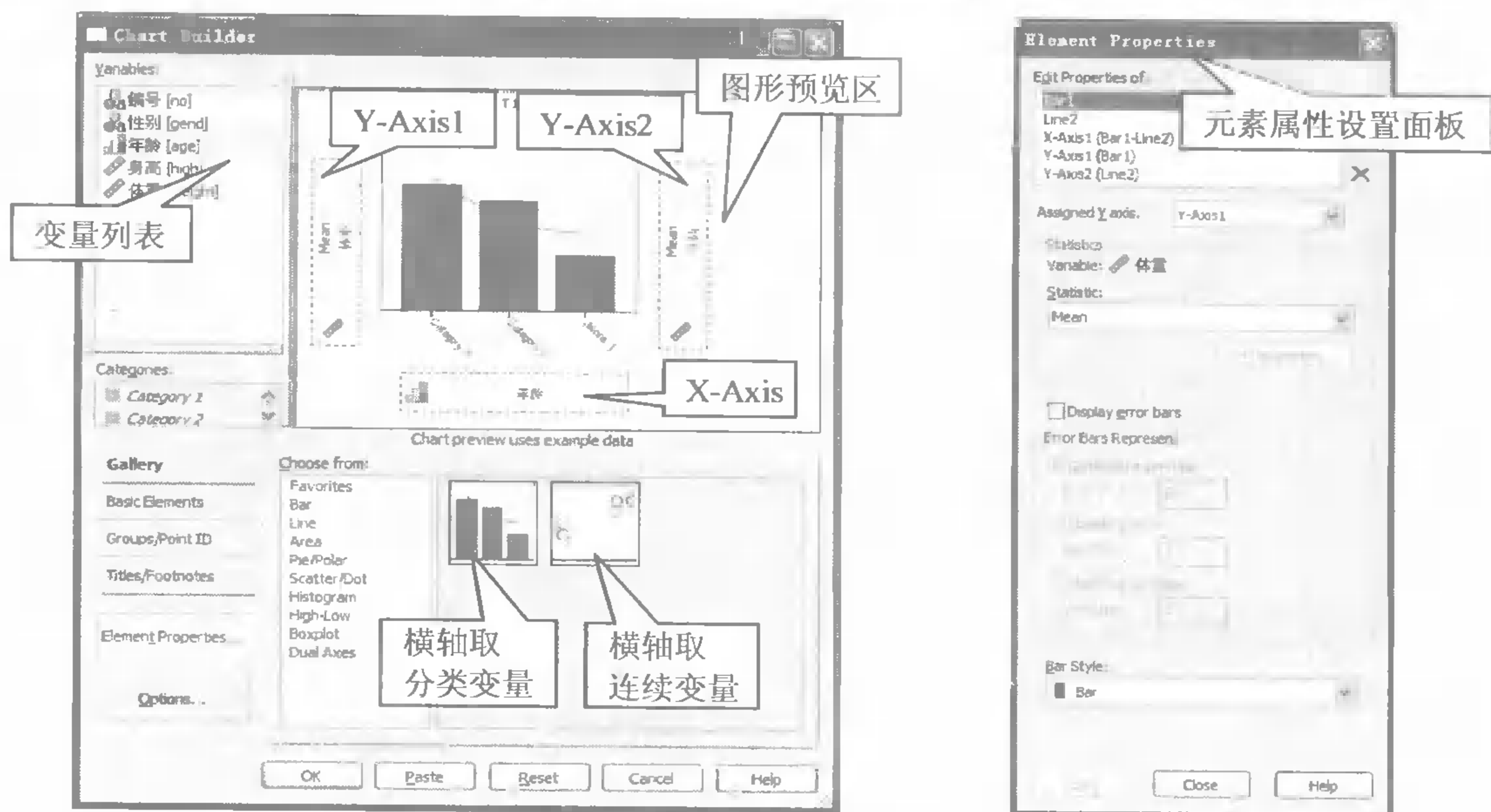
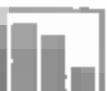


图 19-157 双轴线图设置界面

## 1. 参数设置

在图 19-157 中双击预置图标  (Categorical X Axis) 后, 在图形预览区给出双轴线图的预览; 同时自动弹出元素属性设置面板, 其 Edit 列表给出了如下 5 个元素: Bar1、Line2、X-Axis1(Bar1-Line2)、Y-Axis1(Bar1)和 Y-Axis2(Line2)。

从变量列表中把年龄、体重、身高 3 个变量分别拖动至预览区的 X-Axis、Y-Axis1 和 Y-Axis2 虚线框中, 将其分别作为双轴线图的 X 坐标轴、第 1 个 Y 坐标轴 (条形图) 和第 2 个 Y 坐标轴 (线形图)。

## 2. 输出结果

在图 19-157 中, 单击 OK 按钮运行, SPSS Viewer 窗口的输出图形如图 19-158 所示。

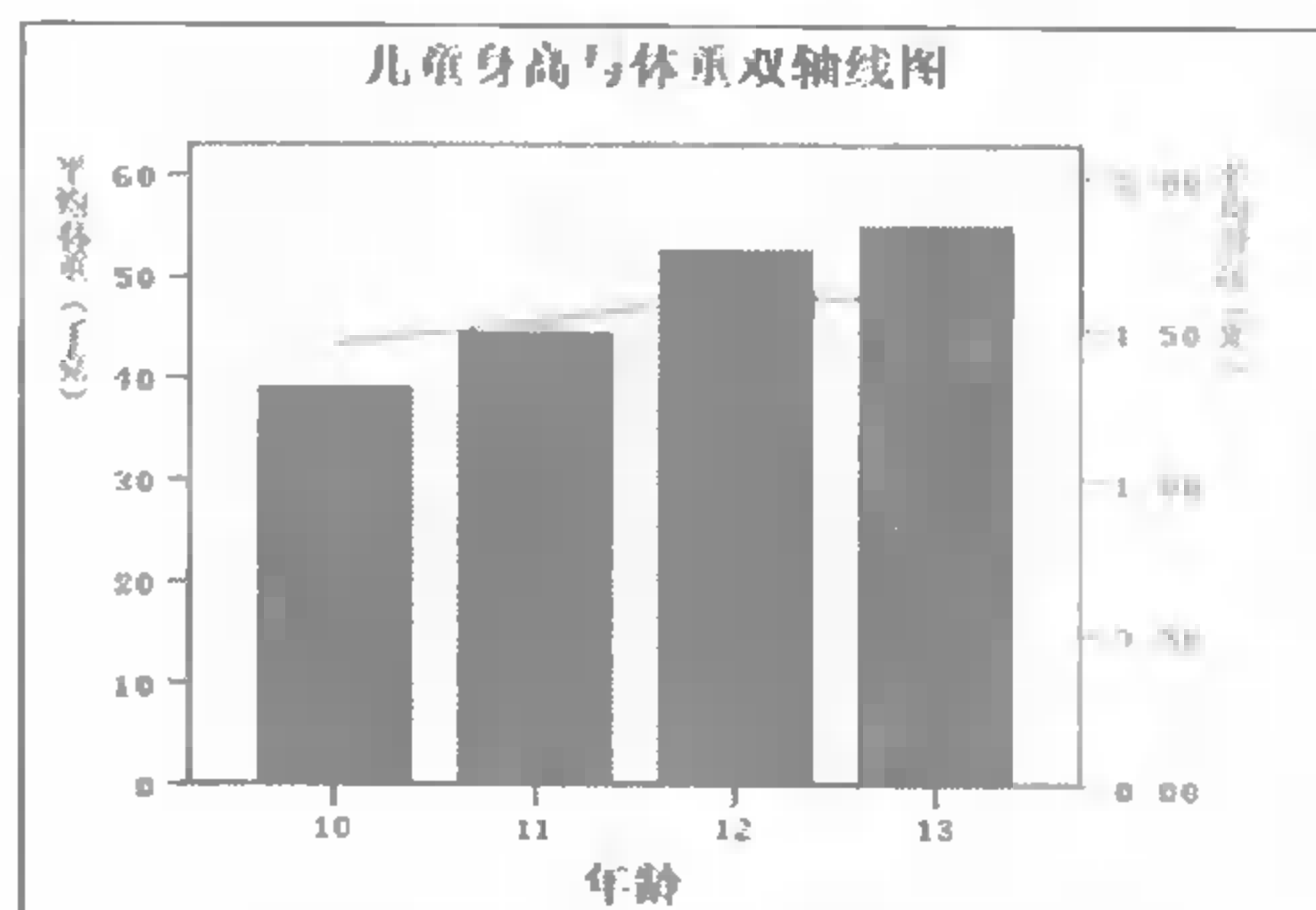


图 19-158 双轴线图的输出结果

条形图表示的是体重信息, 采用左侧的 Y 轴; 折线图表示的是身高信息, 采用右侧的 Y 轴。通过此图还可以观察体重和身高之间的某些联系, 例如随着年龄从 10~13 岁依次增加, 体重的变化要比身高的变化稍快一些。



# 第 20 章 上市公司财务危机预警分析

财务危机 (Financial crisis) 又称财务困境 (Financial distress), 是指企业由于营销、决策或不可抗拒因素的影响, 使经营循环和财务循环无法正常持续或陷于停滞的状态, 具体表现包括持续性亏损、无偿付能力、违约和破产等。

财务危机将给投资者、债权人以及银行等金融机构带来风险, 所以他们都希望在投资决策时就能得到关于财务危机的警示。财务危机给企业和社会带来了严重的影响, 适时、准确地对企业财务危机进行预测分析是市场竞争机制的客观要求。因此利用相关信息构建有效的财务危机预警模型, 获得上市公司财务状况恶化的预警信号, 对于投资者、债权人、经营者以及监管者等诸多方面都具有重要的现实意义。

## 20.1 财务危机预警的应用简介

企业的财务危机预警是日常管理中非常重要的一个环节, 企业内部的诸多问题一旦日积月累地汇集在一起, 就会使管理者逐步失去解决问题的能力, 就容易由财务危机最终导致企业的破产。财务危机预警通过大量历史样本数据, 总结规律, 建立合适的数学模型对企业是否发生财务危机进行预测。随着数学、计算机科学、统计分析软件等工具的发展, 已经有许多不同的方法, 从不同的角度建立预测财务危机模型, 并且能够通过样本检验来证明模型的预测能力。通过 SPSS 软件, 可以方便地进行多种预警模型的建立和检验。

### 20.1.1 财务危机的定量定义方法

对于上市公司的财务危机预警, 国外学者一般以破产为标准展开研究, 例如 Atiman、Beaver、Ohlson 等。但考虑到中国的实际情况, 即破产机制不健全, 目前国内学者大都将特别处理 (ST) 的上市公司作为存在财务危机的上市公司, 例如陈静、李华中等。本章就采用将“财务危机”定义为“因财务状况异常而被特别处理 (ST)”的定量定义方法。

我国证券交易所股票上市规则中, 规定了两种异常状况作为 ST 的判断标准, 一种是财务状况异常, 主要包括以下 3 种情况。

- (1) 最近两个会计年度审计结果显示的净利润均为负值。
- (2) 最近一个会计年度审计结果显示其每股净资产低于股票面值。
- (3) 注册会计师对最近一个会计年度的财务报告出具无法表示意见或否定意见的审计报告。

另一种异常状况统称为其它状况异常, 主要包括以下 5 种情况。

- (1) 因自然灾害、重大事故等原因导致经营设施遭受损失, 生产经营活动基本中止, 在

3 个月以内不能恢复的。

(2) 涉及负有赔偿责任的诉讼或仲裁案件, 依照法院或仲裁机构判决的赔偿金额累计超过公司最近经审计净资产 50% 的。

(3) 主要银行账号被冻结, 影响公司正常经营活动的。

(4) 人民法院受理公司破产案件, 可能依法宣告破产的。

(5) 主要债务人被法院宣告进入破产程序, 而公司相应债权未能计提足额坏账准备致使公司面临重大财务风险的。

以上所提 ST 的各种情况里, 当注册会计师对财务报告出具无法表示意见或否定意见的审计结论时, 有理由相信其会计信息是失真的; 而其它状况异常具有较大的不确定性, 难以从财务角度进行测算。因此建议在一般情况下, 选择样本时要剔除其他状况异常和财务状况异常中第三种情况的公司。

### 20.1.2 财务危机预警的模型选择

自 21 世纪以来, 财务危机预测在欧美得到广泛发展, 诸多学者从定性和定量角度分别进行了研究, 因此涌现出很多方法。当前财务危机预警模型主要采用定量分析的方法, 例如: 单变量检验 (t 检验等)、多元判别分析、logistic 回归方法、因子分析、时间序列方法等。

本章通过 SPSS, 采用 logistic 回归和判别分析两种方法进行财务危机预警建模, 对上市公司样本给出综合评判。同时在对两种方法加以对比之余, 提出了一些改进和完善的意见。

## 20.2 数据描述

本章所用到的关于上市公司的所有数据, 均来自 CCER (中国经济研究服务中心) 更为详尽的数据描述 (包括指标的计算方式), 均参考 <http://www.ccerdata.com/> 网址的说明。

### 20.2.1 数据说明

鉴于国内上市公司当年的财务报告是在次年 1~3 月公布, 是否被特别处理 (ST) 一般在次年 4 月份决定。也就是说, ST 和 \*ST 公司在第  $t$  年被实行特别处理, 说明它在第  $t-1$  和  $t-2$  年连续两年连续出现亏损, 或者第  $t-1$  年的每股净资产低于股票面值。而第  $t-1$  年的财务数据一般不用来做预测, 事实上在第  $t-1$  年的时候已经知道了该公司会不会在第  $t$  年被 ST, 所以使用第  $t-1$  年的信息预测第  $t$  年的财务困境在实际中是没有意义的; 通常选取第  $t-2$  年及其以前各年的财务数据来做预测及分析更有实际意义; 而有些研究者认为第  $t-2$  年的数据也不应该加以利用, 因为它在一定程度上也能决定第  $t$  年该公司是否被 ST 的走向。此外, 何沛俐、章早立通过利用时序样本实证研究发现, 在第  $t-4$  年时, 财务困境企业与正常企业之间的差异是不明显的。综合看来, 公司财务危机的有效预测期往往在第  $t-2$  年和第  $t-3$  年之中加以选择。

本章的研究样本取自 CCER 中的一般上市公司财务数据库, 直接数据源为 2004-2007 年随机抽取的部分公司的财务数据, 其中 06 或 07 年首次被 ST 的公司标记为发生财务危机, 而准备用 04 年的财务数据对 06 或 07 年是否发生财务危机进行预警。这样定义的一个特点是, 04 年相对于 06 年为第  $t-2$  期的数据, 而 04 年相对于 07 年为第  $t-3$  期的数据, 综合考虑了更多情形下的有效预测数据, 能使模型更加稳健。

本章案例中的公司是随机选择的, 根据研究的具体问题不同, 有必要以不同的标准来选

择特定类别的公司。对于有特殊需求的样本筛选，请参考第 20.2.3 节中的补充建议。

20.2.2 指标选择

选择哪些指标作为建模的变量，对模型的预测能力及其可靠性都会产生较大的影响。为了全面而客观地描述上市公司的经营情况，参考了其他文献的研究经验，本章节我们选取了能够反映上市公司如下 4 个方面的一些财务指标：偿债能力、盈利能力、成长能力和营运能力。具体指标类型和名称如表 20-1 所示。

表 20-1 反映上市公司经营业绩的财务指标

指标类型	包含的指标名称
偿债能力	流动比率、速动比率、存货流动负债比率、现金流动负债比率、资产充足率、债务资本比率、债务资产比率
盈利能力	净资产收益率、资产收益率、净利润率
成长能力	总资产增长率、营业利润增长率
营运能力	存货周转率、应收账款周转率、流动资产周转率、资产周转率、固定资产周转率

此外，在 CCER 的 Web 页面下载数据时，还选择了股票代码、股票简称、行业分类、总资产、所有者权益合计等指标选项，具体的下载设置如图 20-1 所示。

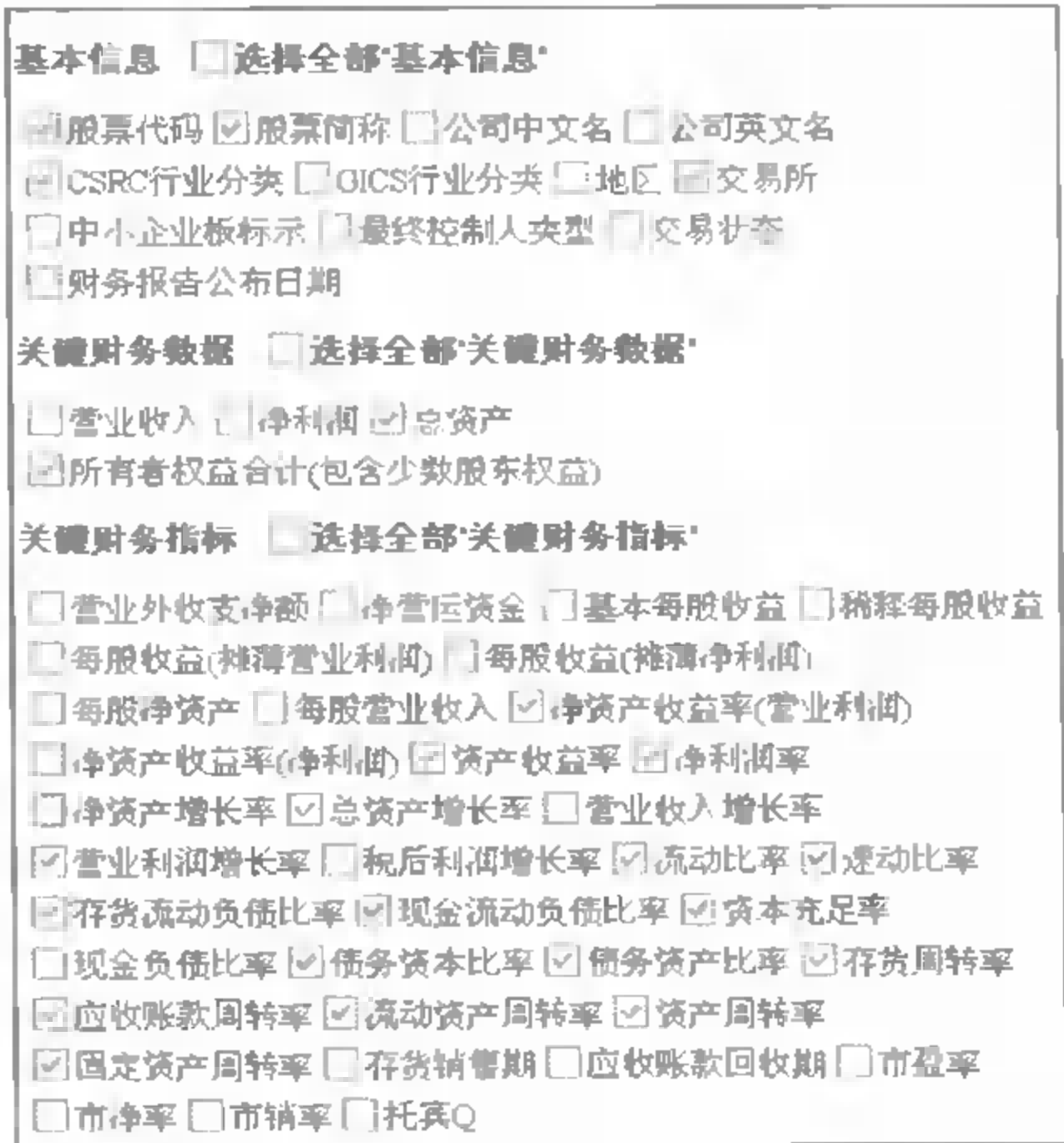


图 20-1 CCER 里财务危机预警的指标选择

20.2.3 补充说明

参考其他研究者使用的不同的样本选择方法，本节对样本和指标的选择做一些必要的补充说明。

(1) 样本行业分类。不同行业有其各自的特点，其公司在财务数据上也就可能有不一样的特征，所以有些研究者倾向于先对公司进行行业分类，对不同行业的公司分别建立其财务预警模型；也有另外一些研究者先在各行业按某个比例分别选取一定数目的公司，再对这些包含了指定行业的公司样本进行建模。

(2) 样本配对方法。选择了发生财务危机的 ST 公司作为目标公司，还应选择未被 ST 的公司进行配对，这样才能建立有效的预测模型。一般要根据同行业（按证监会行业代码分类）、同规模的原则查找与 ST 公司对应的配对样本，并以同行业为第一选择标准。目标样本和配

对样本应选取同期的财务指标。

关于配对比例的问题，不同的研究者采取了不同的方式。例如闫哲（2007）和孔丽娜（2006）就采取了不同的样本配对比例，前者 ST 和非 ST 样本的比例接近 1:10，比较接近于我国上市公司实际被 ST 的公司占总体上市公司的比例；后者采取了 1:1 的配对比例，即抽取的危机样本和正常样本个数相等。笔者建议，根据实际问题的特殊情况，灵活选择配对比例，不一定局限于这两种情况，本章的样本为随机选择，故没有特别指定配对的原则。

（3）CCER 的数据筛选。CCER 中新增加了股票筛选功能，其 Web 操作界面如图 20-2 所示，在这里可以设置综合的查询条件，以筛选需要的数据。例如：可以先筛选所有被 ST 的公司的财务信息，然后根据它们的行业分类，再筛选部分同行业的未被 ST 的配对公司的财务信息。

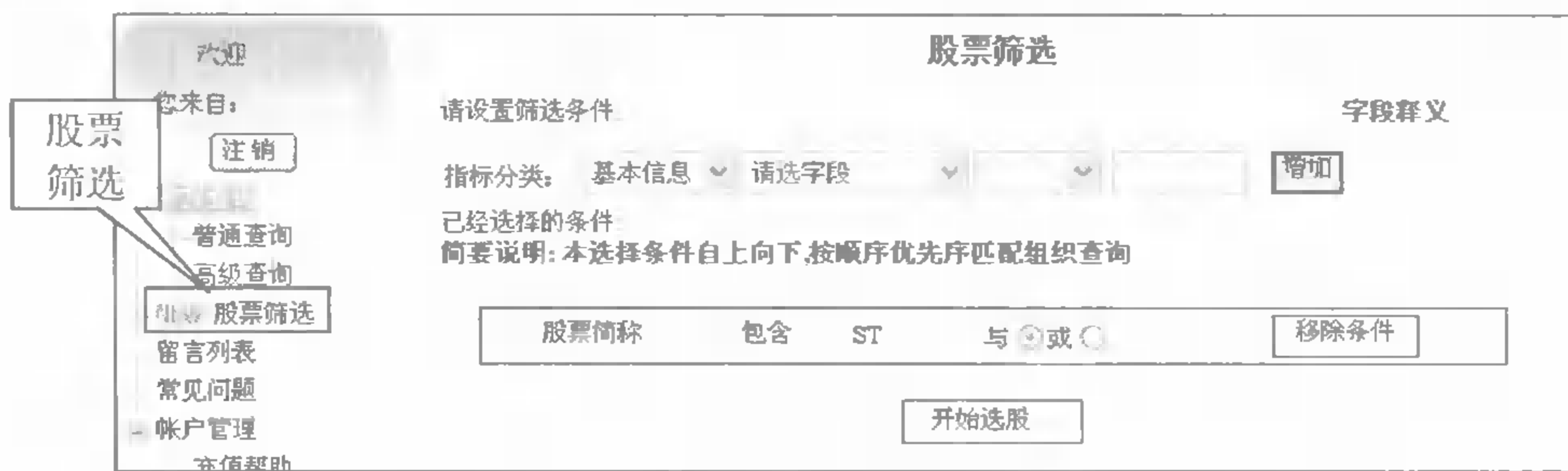


图 20-2 CCER 的股票筛选功能

## 20.3 分析方法概述

### 20.3.1 判别分析

判别分析方法是英国统计学家 Fisher 最先建立的一种统计方法。在财务危机预测研究中，该方法使用多个变量进行判定分析，是多元统计分析中用于判别样本所属类型的一种统计方法。判别分析模型主要解决的问题是，在已知某些研究对象的分类情况后，再利用这些已知类别的样本生成一种判别标准，用以确定新的样本属于已知类别中的哪一类。两分类判别分析模型的思想，是通过将多维数据投影到某个方向上，投影的原则是将类与类尽可能地分开，然后再选择合适的判别规则，将待判的样品进行分类判别。

1968 年，美国学者 Edward I. Altman 率先将多元线性判别方法引入财务预警领域，建立了 Z-Score 五变量模型，此后多变量分析方法被各国学者广泛采用，比较典型的有：Edmister (1972) 提出的专门针对小企业的财务预警模型，英国的 Taffler (1977) 的多变量模式，日本开发银行的多变量预测模型等。

多元判别分析模型与单变量模型相比有一定的优越性，从而得到广泛的应用，但是该方法本身也存在一些问题，使得该模型的应用受到较为严格的条件限制。例如：要求财务变量服从独立的多元正态分布假设；要求各类别具有相同的协方差。

本章将采用多元判别分析模型，对多个财务比率指标进行分析。

### 20.3.2 logistic 回归方法

Martin (1977) 首次运用多元 Logistic 模型进行银行破产预测，后来 Ohison (1980) 选



取了 9 个财务变量,又运用该模型来预测企业的财务危机。Logistic 模型在 20 世纪 80 年代以后得到了较为广泛的应用,一些学者对其进行了改进和扩展,如 Charitou 和 Trigeorgis(2000)采用 Logistic 回归方法并结合期权定价(B~S)模型中的相关变量构建了财务危机判别模型,对 1983 年到 1994 年期间的 139 家美国企业进行了对比检验,结果发现到期债务面值、企业资产的当期市价、企业价值变化的标准差等期权变量在预测破产方面作用显著。国内也有很多的学者使用 Logistic 回归模型来研究财务危机预警问题,如:吴世农和卢贤义(2001)、何沛俐和章早立(2002)等。

多元逻辑回归模型主要应用于二元取值的应变量,例如成功和失败的问题。它使用最大似然法估计参数值,最后得到应变量取某个值的概率。Logistic 回归模型的目标,是提供把观测对象归为某个类别的条件概率,由此衡量企业发生财务危机的概率有多大,而不仅仅是判断企业是否会发生财务危机。

Logistic 回归模型克服了多元判别分析的一些缺点,例如不要求多元正态分布和同协方差作为假设前提。但是该模型同样也存在如下一些问题:样本的数量不宜太少,否则容易引起参数估计的有偏性;模型对中间区域(概率在 0.5 附近)的判断性较为模糊,导致判别结果不稳定;由于对参数的估计运用最大似然估计法,计算和分析程序相对复杂,不过用 SPSS 软件可以轻松实现。

假设  $X_i = (x_{1i}, x_{2i}, \dots, x_{ki})$  是反映第  $i$  个公司财务状况的  $k$  变量,  $\alpha$  和  $\beta$  为待估参数,第  $i$  个公司破产的概率  $p_i$  为:  $P(X_i, \beta) = F(\alpha + \beta X_i) = \frac{1}{1 + e^{-(\alpha + \beta X_i)}}$ ; Logistic 回归模型的一般形式是:

$\text{Log}(\frac{p_i}{1 - p_i}) = \alpha + \sum_{k=1}^K \beta_k x_{ki}$ ; 利用极大似然估计法估计出此式中的参数  $\alpha$ 、 $\beta$ ,再计算公司破产的概率  $P(X_i, \beta)$ ,从而判定公司财务情况。当  $P$  值大于 0.5,表明企业发生财务危机的概率比较大,可以判定企业为财务危机类型;当  $P$  值低于 0.5,表明企业财务正常的概率比较大,可以判定企业为财务正常。

## 20.4 SPSS 建模过程和结论分析

本节先对采集的数据进行转换和筛选,再利用筛选后的数据分别建立关于财务危机预警的判别分析模型和 logistic 回归模型。

### 20.4.1 SPSS 数据筛选操作

本节所用数据均源自 CCER 的 Web 页面下载,初始下载的数据文件为“04-07 初始数据数据.xls”,下面先把这部分数据转化为 SPSS 格式的文件。

#### 1. 数据导入

依次单击菜单“File→Open→Data...”执行文件打开操作,其界面如图 20-3 所示,在文件类型下拉列表选中“Excel(\*.xls)项;然后,在上面的文件列表选中“04-07 初始数据数据.xls”;单击打开按钮,弹出如图 20-4 所示的文件打开确认窗口。在图 20-4 中,保留默认设置,单击 OK 按钮,即可新建一个包含导入数据的 SPSS 格式数据文件,将此文件另存为“初始财务数据.sav”。

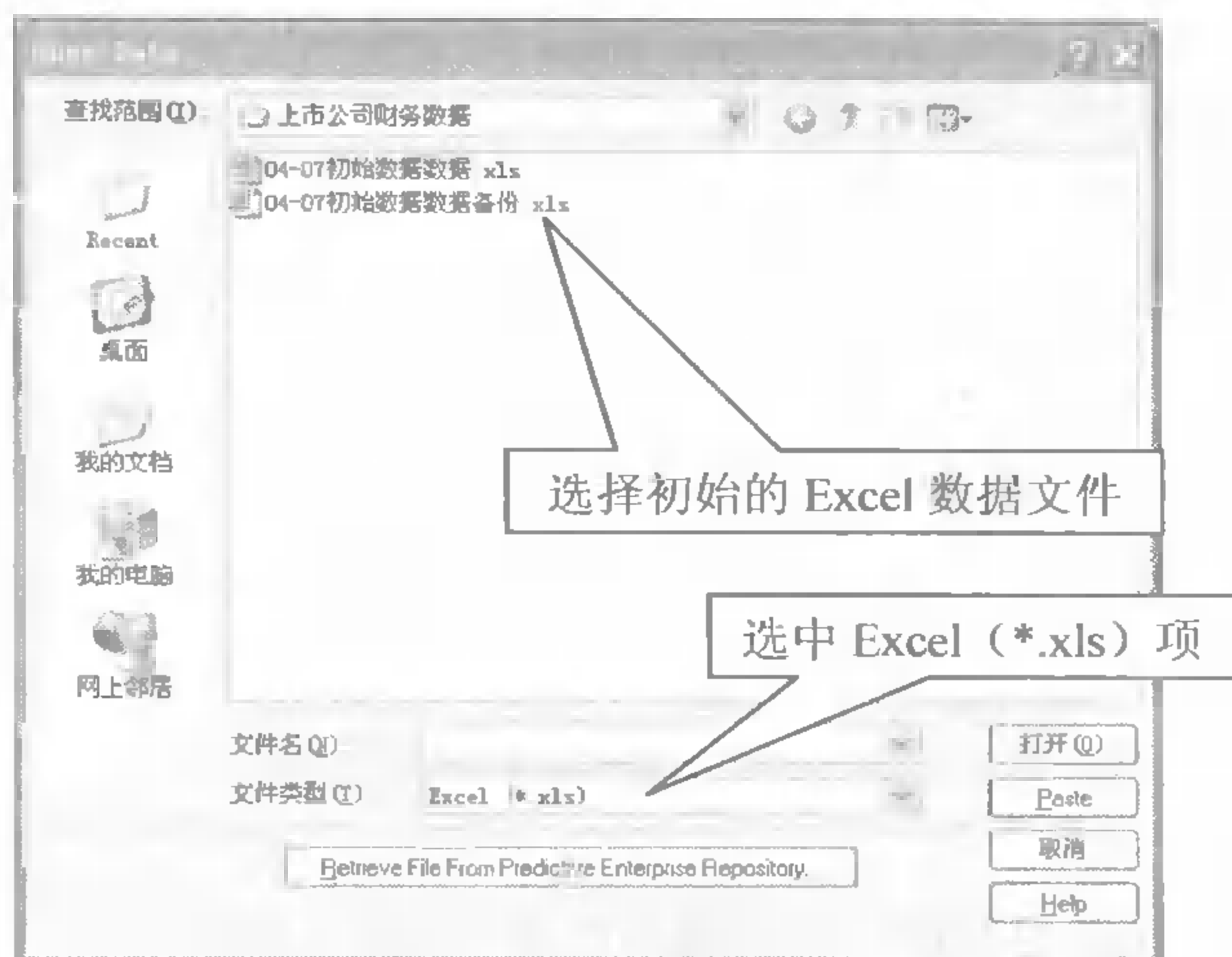


图 20-3 文件导入界面



图 20-4 文件导入操作

这只是数据导入的快捷方法，另外依次单击菜单“File→Open Database→New Query...”，通过建立对 Excel 数据库的连接和搜索，能够实现更复杂的数据导入操作。

打开“初始财务数据.sav”文件，在变量视图中把变量标签都设置为与对应的变量名相同的内容，然后把变量名更改为类似“x1”的格式，更改后的变量视图如图 20-5 所示，未在其中的变量将其删除。这些变量的其他设置都采用默认选项。

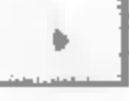

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	x1	Numeric	8	0	股票代码	None	None	6	Right	Scale
2	x2	String	255	1	股票简称	None	None	7	Left	Nominal
3	x3	Numeric	4	0	年度	None	None	7	Right	Scale
4	x4	Numeric	8	2	总资产	None	None	9	Right	Scale
5	x5	Numeric	8	2	所有者权益合计	None	None	8	Right	Scale
6	x6	Numeric	8	2	净资产收益率	None	None	8	Right	Scale
7	x7	Numeric	8	2	资产收益率	None	None	8	Right	Scale
8	x8	Numeric	8	2	净利润率	None	None	8	Right	Scale
9	x9	Numeric	8	2	总资产增长率	None	None	8	Right	Scale
10	x10	Numeric	8	2	营业利润增长率	None	None	8	Right	Scale
11	x11	Numeric	8	2	流动比率	None	None	8	Right	Scale
12	x12	Numeric	8	2	速动比率	None	None	8	Right	Scale
13	x13	Numeric	8	2	存货流动负债比	None	None	8	Right	Scale
14	x14	Numeric	8	2	现金流动负债比	None	None	8	Right	Scale
15	x15	Numeric	8	2	资本充足率	None	None	8	Right	Scale
16	x16	Numeric	8	2	债务资本比率	None	None	8	Right	Scale
17	x17	Numeric	8	2	债务资产比率	None	None	8	Right	Scale
18	x18	Numeric	8	2	存货周转率	None	None	8	Right	Scale
19	x19	Numeric	8	2	应收账款周转率	None	None	8	Right	Scale
20	x20	Numeric	8	2	流动资产周转率	None	None	8	Right	Scale
21	x21	Numeric	8	2	资产周转率	None	None	8	Right	Scale
22	x22	Numeric	8	2	固定资产周转率	None	None	6	Right	Scale

图 20-5 初始财务数据的变量视图

## 2. 数据转置

本小节通过 SPSS 的转置操作，改变数据的显示方式。

打开“初始财务数据.sav”文件，依次单击菜单“Data→Restructure...”进行数据转置的操作，主设置界面如图 20-6 所示，单击选中 Restructure selected cases into variables 单选项，表示把记录行转置为变量列；单击下一步，进入如图 20-7 所示的变量选择界面。

在变量列表选中股票代码变量（x1），单击上面的  按钮将其选入 Identifier Variables 列表，作为标识变量；在变量列表选中年度变量（x3），单击下面的  按钮将其选入 Index Variables 列表，作为索引变量。如此将按照“x1.x3”的格式对数据进行转置，这里的“.”表

示取索引值操作。

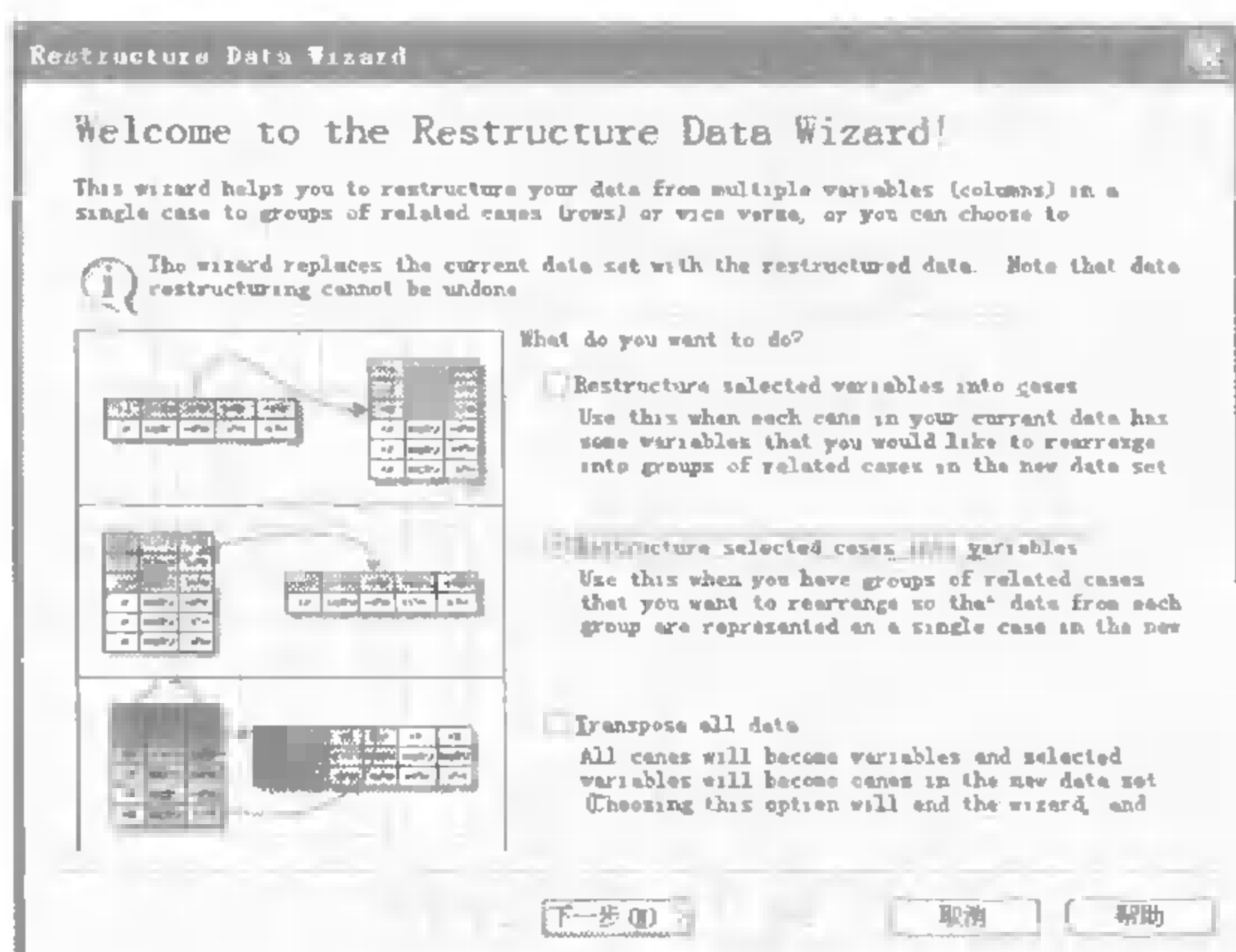


图 20-6 数据转置第 1 步

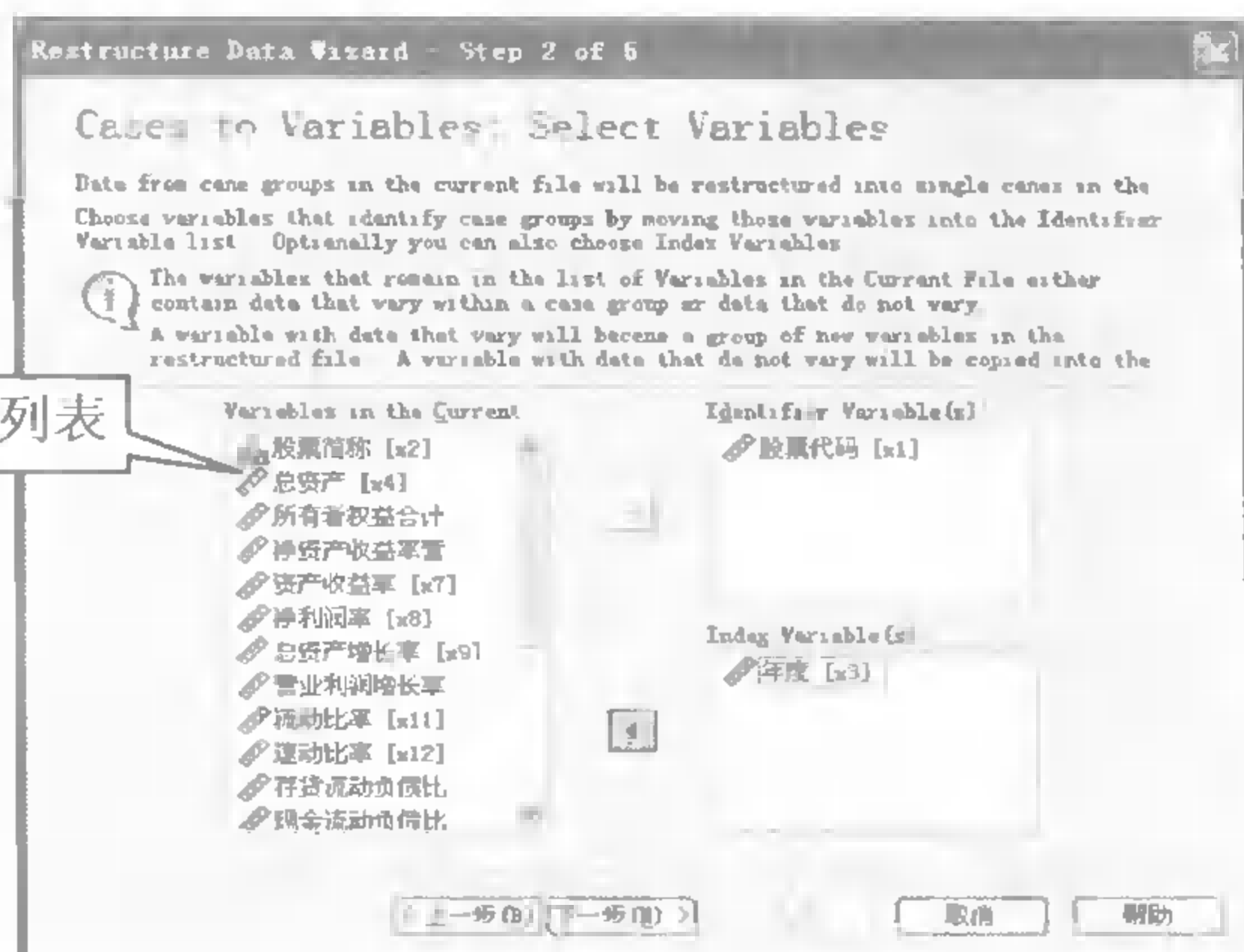


图 20-7 数据转置第 2 步

单击下一步按钮，随后的两个设置界面都是关于转置后记录行或变量列显示顺序的选项，都采用默认设置；一路单击下一步按钮，在最后一个界面单击完成按钮运行转置操作。

转置后的数据格式如图 20-8 所示，将其另存为“转置后的财务数据.sav”文件。由于同一股票代码的股票名称在不同年份可能不一样，转置后的财务数据里将只保留了它在 2004 年的名称作为标识。


	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	x1	Numeric	8	0	股票代码	None	None	6	Right	Scale
2	x2.2004	String	255		x2.2004: 股票简称	None	None	7	Left	Nominal
3	x2.2005	String	255		x2.2005: 股票简称	None	None	7	Left	Nominal
4	x2.2006	String	255		x2.2006: 股票简称	None	None	7	Left	Nominal
5	x2.2007	String	255		x2.2007: 股票简称	None	None	7	Left	Nominal
6	x4.2004	Numeric	8	2	x4.2004: 总资产	None	None	9	Right	Scale
7	x4.2005	Numeric	8	2	x4.2005: 总资产	None	None	9	Right	Scale
8	x4.2006	Numeric	8	2	x4.2006: 总资产	None	None	9	Right	Scale
9	x4.2007	Numeric	8	2	x4.2007: 总资产	None	None	9	Right	Scale
10	x5.2004	Numeric	8	2	x5.2004: 所有者权益	None	None	8	Right	Scale
11	x5.2005	Numeric	8	2	x5.2005: 所有者权益	None	None	8	Right	Scale
12	x5.2006	Numeric	8	2	x5.2006: 所有者权益	None	None	8	Right	Scale
13	x5.2007	Numeric	8	2	x5.2007: 所有者权益	None	None	8	Right	Scale
14	x6.2004	Numeric	8	2	x6.2004: 净资产收益	None	None	8	Right	Scale
15	x6.2005	Numeric	8	2	x6.2005: 净资产收益	None	None	8	Right	Scale
16	x6.2006	Numeric	8	2	x6.2006: 净资产收益	None	None	8	Right	Scale

图 20-8 转置后的数据格式

### 3. 数据筛选

经过前面两步的铺垫，下面可以来建立标识 06 或 07 年是否被 ST 的变量。

打开“转置后的财务数据.sav”文件，依次单击菜单“File→Open→Syntax...”执行文件打开操作，打开命令文件“财务数据过滤.sps”。Syntax 窗口的显示内容如图 20-9 所示。

在图 20-9 中，单击工具栏里的执行按钮运行程序后，在当前数据集生成代表是否被 ST 的变量 stfilter，stfilter=0 代表 04-07 年都没有被 ST，stfilter=1 代表 06 或 07 年首次被 ST。另外，删除了 stfilter=-1 的记录，因为这部分记录不符合本章对数据的要求，例如其中可能会包括 05 年就被 ST 的公司。

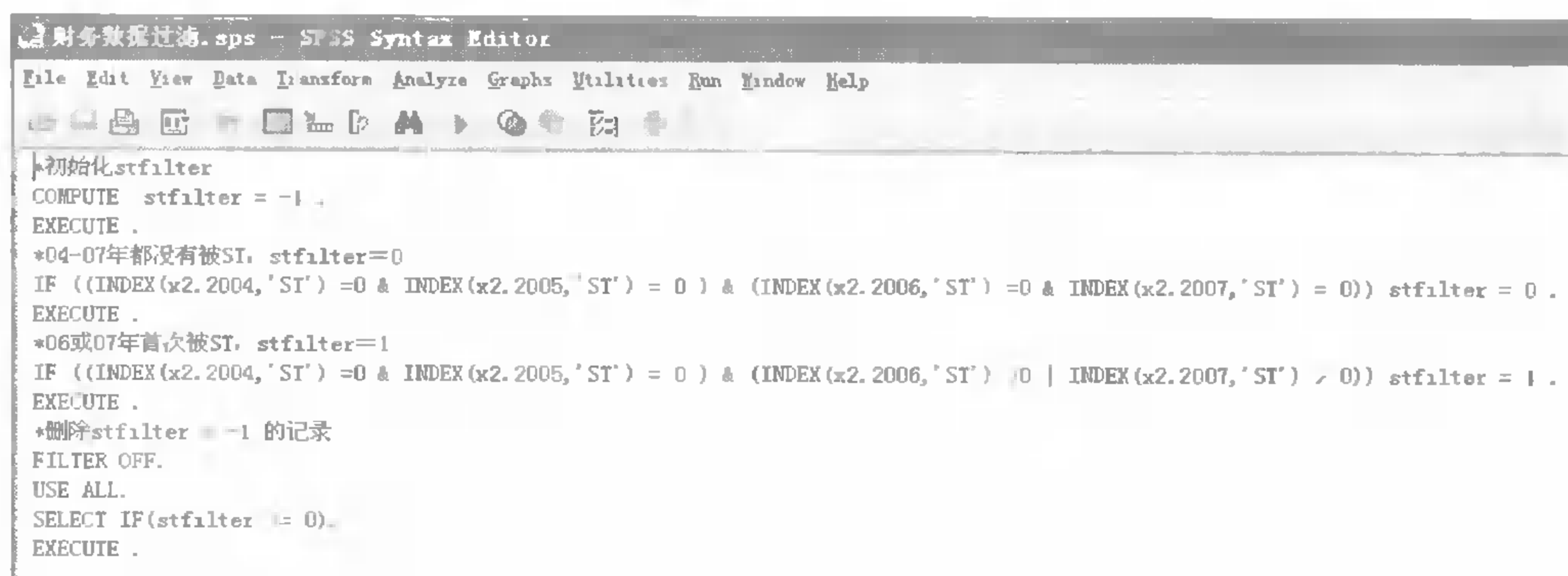


图 20-9 财务数据过滤的 Syntax 文件内容

另外，文件“财务数据过滤.sps”所实现的功能，完全可以通过 Computing Variables 过程和 Select Cases 过程实现，事实上文件里的代码都可以在这两个过程的设置面板里单击 Paste 按钮获得。


经过如上操作，将转置后的数据保存至“过滤后的财务数据.sav”文件，且只保留如下变量：股票代码、股票名称、2004 年的财务数据（格式为\*.2004）和 stfilter 变量，其他变量删除。如此得到的数据格式就成为随后建模时的输入数据，如图 20-10 所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	x1	Numeric	8	0	股票代码	None	None	5	Right	Scale
2	x2	String	255	1	股票简称	None	None	7	Left	Nominal
3	x3	Numeric	8	2	总资产	None	None	9	Right	Scale
4	x4	Numeric	8	2	所有者权益合计	None	None	8	Right	Scale
5	x5	Numeric	8	2	净资产收益率	None	None	8	Right	Scale
6	x6	Numeric	8	2	资产收益率	None	None	8	Right	Scale
7	x7	Numeric	8	2	净利润率	None	None	8	Right	Scale
8	x8	Numeric	8	2	总资产增长率	None	None	8	Right	Scale
9	x9	Numeric	8	2	营业利润增长率	None	None	8	Right	Scale
10	x10	Numeric	8	2	流动比率	None	None	8	Right	Scale
11	x11	Numeric	8	2	速动比率	None	None	8	Right	Scale
12	x12	Numeric	8	2	存货流动负债比	None	None	8	Right	Scale
13	x13	Numeric	8	2	现金流动负债比	None	None	8	Right	Scale
14	x14	Numeric	8	2	资本充足率	None	None	8	Right	Scale
15	x15	Numeric	8	2	债务资本比率	None	None	8	Right	Scale
16	x16	Numeric	8	2	债务资产比率	None	None	8	Right	Scale
17	x17	Numeric	8	2	存货周转率	None	None	8	Right	Scale
18	x18	Numeric	8	2	应收账款周转率	None	None	8	Right	Scale
19	x19	Numeric	8	2	流动资产周转率	None	None	8	Right	Scale
20	x20	Numeric	8	2	资产周转率	None	None	8	Right	Scale
21	x21	Numeric	8	2	固定资产周转率	None	None	6	Right	Scale
22	stfilter	Numeric	8	0	06/07是否ST	{0, 06/07未ST}	None	10	Right	Scale

图 20-10 过滤后的财务数据文件格式

#### 4. 数据概览

首先，观察一下新生成的 stfilter 变量的基本统计信息。

依次单击菜单“Analyze→Descriptive Statistics→Frequencies...”执行频数分析过程，如图 20-11 所示，在左侧的变量列表选中 06/07 是否 ST (stfilter) 变量，单击  按钮将其选入 Variable(s) 分析变量列表。单击 OK 按钮运行，SPSS Viewer 窗口的输出表格如图 20-12 所示。

观察图 20-12，本例的有效数据的 185 例，其中“目标值”06 或 07 年首次被 ST 的公司有 26 例，占有所有公司的 14.1%；由于我们没有指定特别的样本配对规则，所以认为这样的比例是比较符合实际的，适用于随后的分析。



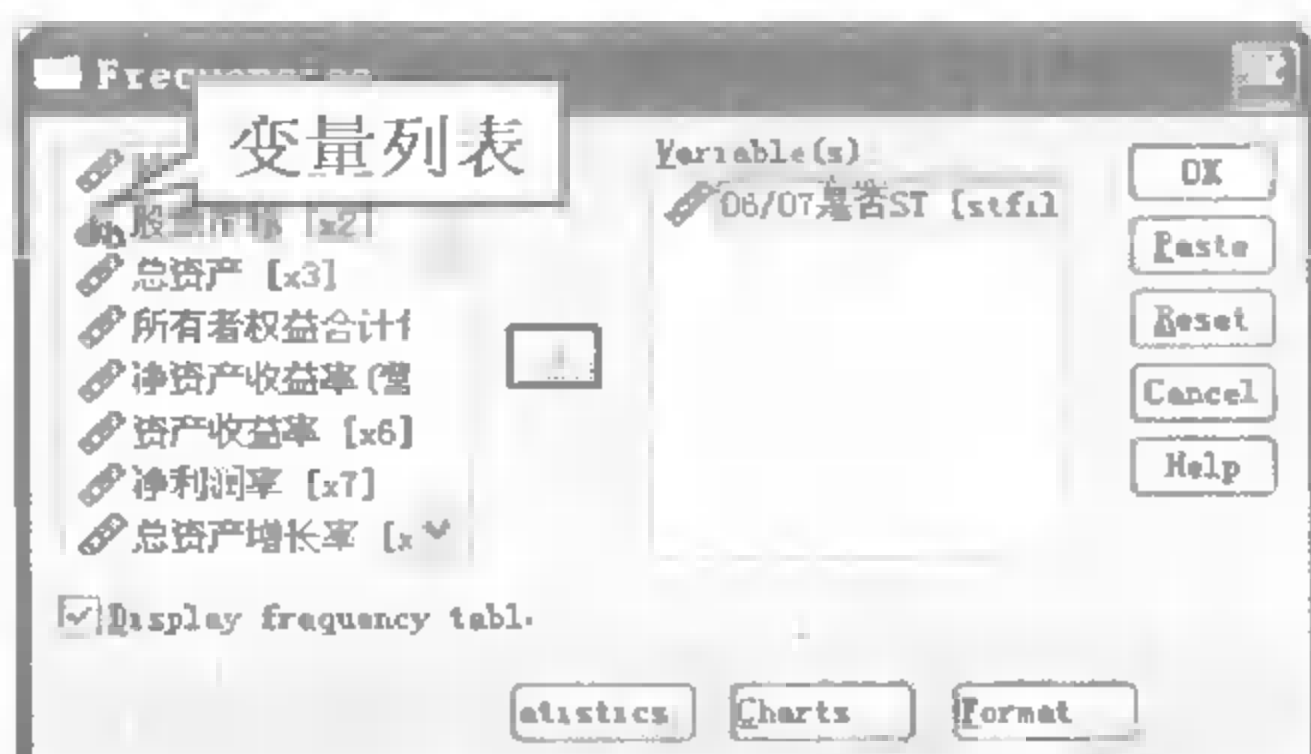




图 20-11 频率分析参数设置

06/07是否ST					
有效	06/07是否ST	频率	百分比	有效百分比	累积百分比
	06/07未ST	159	85.9	85.9	85.9
	06/07首ST	26	14.1	14.1	100.0
	合计	185	100.0	100.0	

图 20-12 stfilter 进行的频率统计输出

## 20.4.2 SPSS 判别分析建模与分析

### 1. SPSS 参数设置

打开文件“过滤后的财务数据.sav”，依次单击菜单“Analyze→Classify→Discriminant...”执行判别分析过程，其主界面如图 20-13 所示。在变量列表选中 06/07 是否 ST (stfilter) 变量，单击从上至下第一个  按钮将其选入 Grouping Variable 选框，作为目标分类变量；在变量列表选中从总资产至固定资产周转率的所有变量，单击从上至下第二个  按钮将其选入 Independents 列表，作为自变量；单击选中 Use stepwise method 单选项，使用逐步判别法。

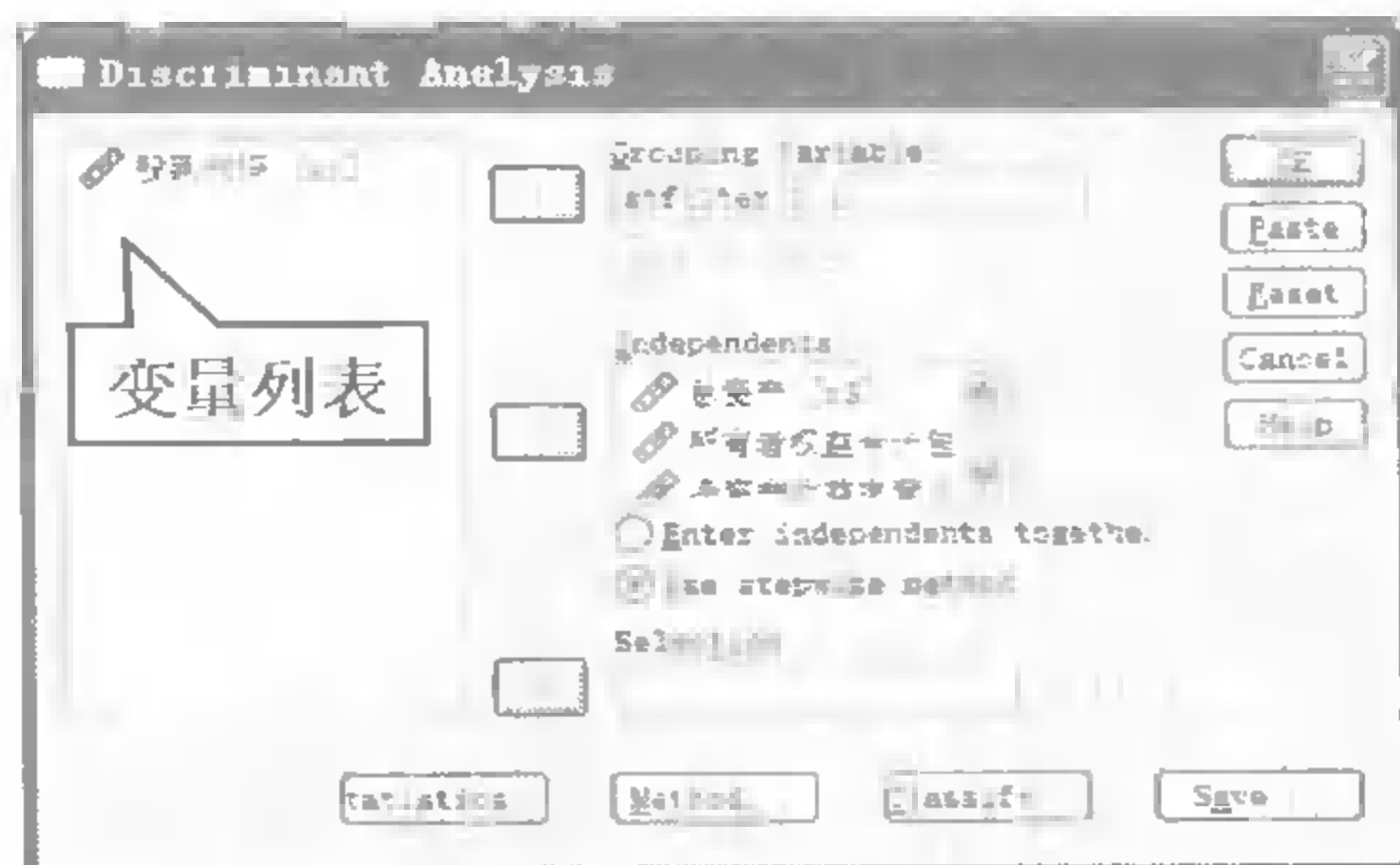


图 20-13 判别分析的主设置界面

在图 20-13 中的 Grouping Variable 选框，单击选中 stfilter 变量，然后单击 Define Range 按钮，弹出如图 20-14 所示的取值范围定义对话框，分别在 Minimum、Maximum 后面输入最小值 0、最大值 1，单击 Continue 按钮返回主面板。

在图 20-15 中，单击 Statistics 按钮，弹出如图 20-15 所示的统计量设置对话框，依次勾选如下 3 个复选框：Box's M (协方差检验)、Fisher's (判别系数)、Unstandardized (未标准化判别系数)。单击 Continue 按钮返回主面板。

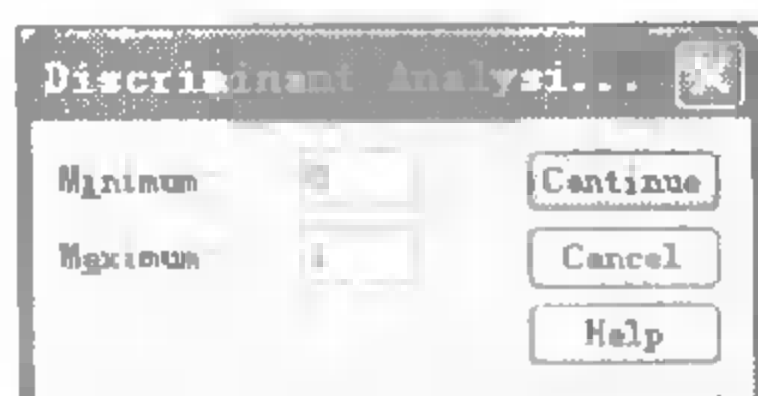


图 20-14 取值定义对话框

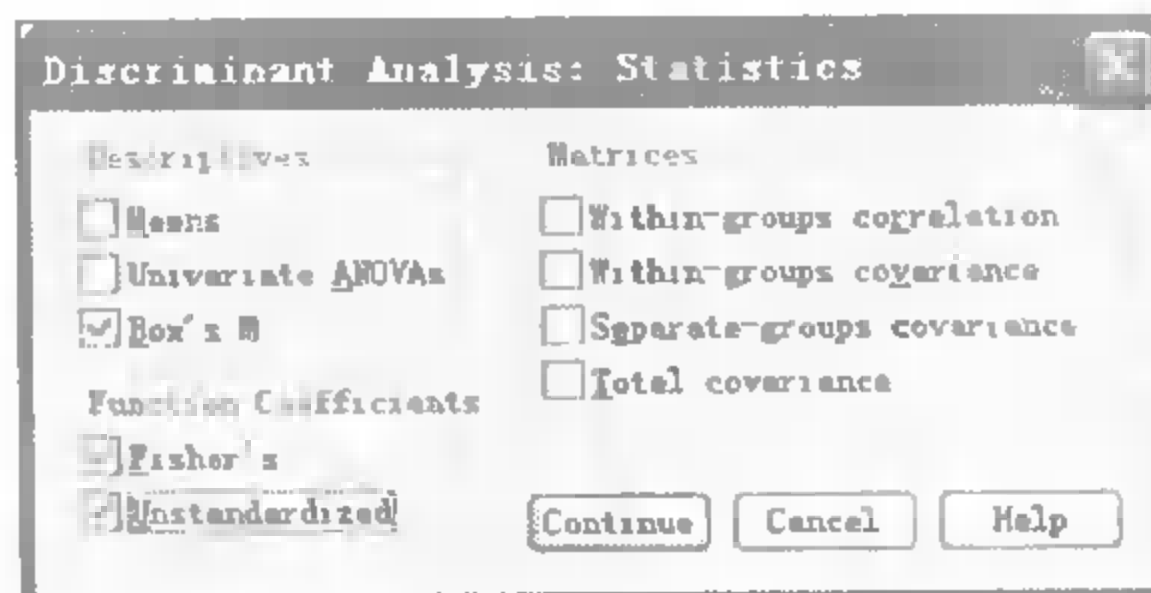


图 20-15 判别分析的统计量设置对话框

在图 20-13 中，单击 Method 按钮，弹出如图 20-16 所示的逐步方法设置对话框，依次勾选如下 3 个关于逐步判别方法的选项：Wilks' lambda 单选框 (判别统计量)、Use Probability of F

单选框（变量筛选准则）、Summary of steps 复选框（输出选项）。单击 Continue 按钮返回主面板。

在图 20-13 中，单击 Classify 按钮，弹出如图 20-17 所示的分类选项设置对话框，依次勾选关于分类规则的如下 5 个参数选项：Compute from groups sizes 单选框（由样本计算先验概率）、Within-groups 单选框（合并的类内协方差）、Summary table 复选框（分类结果总结表）。单击 Continue 按钮返回主面板。

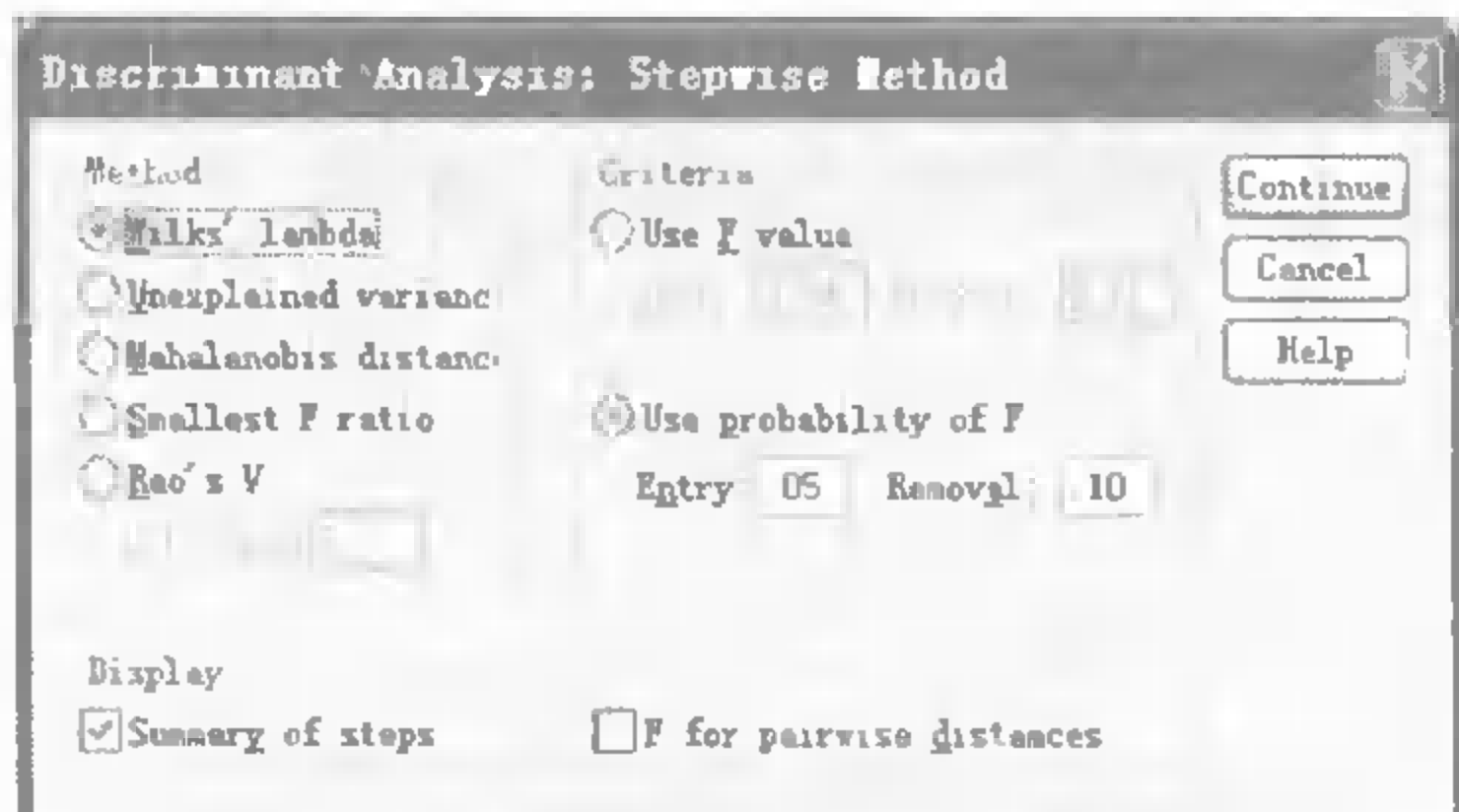


图 20-16 判别分析的逐步方法设置

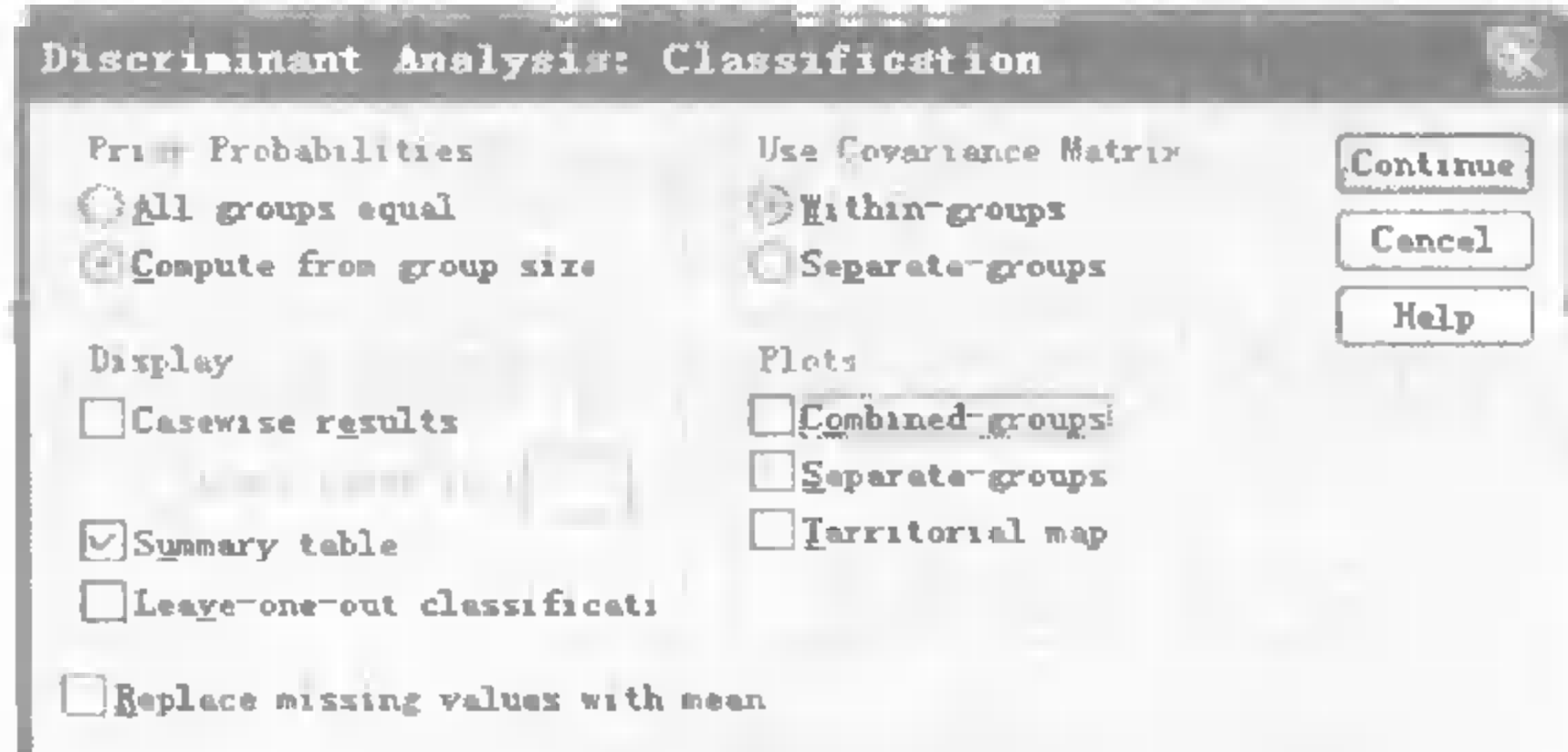


图 20-17 判别分析的分类选项设置

## 2. 初始模型的结果分析

在图 20-13 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 20-18 所示。

“分类结果”表显示，初始判别模型的分类准确率达到了 92.4%，分类效果还是不错的，保留这个结果以备和后面的改进模型作比较。

“检验结果”表里，Box's M 检验的显著性值 Sig 远小于 0.01，所以认为协方差相等的假设不成立，所以建议在图 20-17 所示的分类选项设置对话框里，采用 Separate-groups 选项进行分析。

检验结果				
Box's M		996.781		
F	近似	61.493		
	df1	15		
	df2	7601.207		
	Sig	.000		
对相等总体协方差矩阵的零假设进行检验。				

分类结果 <sup>a</sup>				
		预测组成员		合计
		06/07未ST	06/07首ST	
初始	计数	158	1	159
		13	13	26
%	06/07未ST	99.4	.6	100.0
	06/07首ST	50.0	50.0	100.0
a. 已对初始分组案例中的 92.4% 个进行了正确分类。				

图 20-18 初始模型的协方差检验结果和最终分类结果

## 3. 改进模型的参数设置和结果分析

在图 20-17 中，单击选中 Separate-groups 单选框，指定使用每个类别的协方差矩阵进行分类，如图 20-19 所示，单击 Continue 按钮返回主面板。其他设置同前。

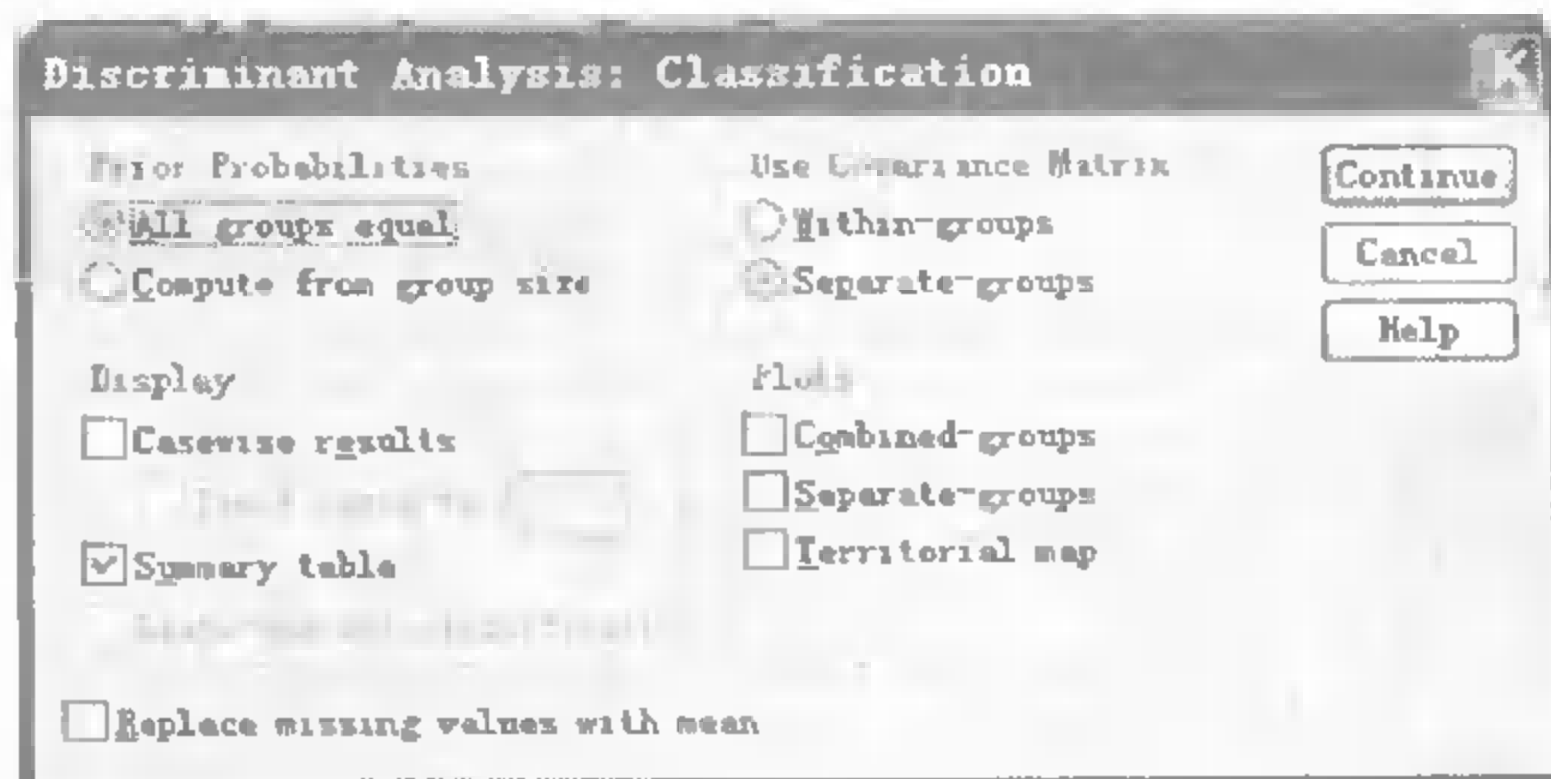


图 20-19 改进后的参数设置

回到图 20-13 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 20-20～图 20-26 所示。

警告		
选项“SEPARATE”意味着分类使用规范判别式函数的组协方差矩阵，而不是初始变量的组协方差矩阵。如果函数的数量比变量少，将有所不同。		
分析案例处理摘要		
未加权案例	N	百分比
有效	185	100.0
排除的		
缺失或越界组代码	0	0
至少一个缺失判别变量	0	0
缺失或越界组代码还有至少一个缺失判别变量	0	0
合计	0	0
合计	185	100.0

图 20-20 判别模型的处理摘要信息

输入的/删除的变量= b, c, d									
步骤	输入的	Wilks 的 Lambda							
		统计量	df1	df2	df3	统计量	df1	df2	Sig.
1	资产收益率	.666	1	1	183.000	.91647	1	183.000	.000
2	速动比率	.637	2	1	183.000	.51903	2	182.000	.000
3	现金流动负债比率	.601	3	1	183.000	.40005	3	181.000	.000
4	存货流动负债比率	.587	4	1	183.000	.31638	4	180.000	.000
5	总资产	.572	5	1	183.000	.26740	5	179.000	.000

在每个步骤中，输入了最小化整体 Wilks 的 Lambda 的变量。

- a. 步骤的最大数目是 38。
- b. 要输入的 F 的最大显著水平是 .05。
- c. 要删除的 F 的最小显著水平是 .10。
- d. F 级、容差或 VIF 不足以进行进一步计算。

图 20-21 判别分析的变量选择过程

特征值					Wilks 的 Lambda				
函数	特征值	方差的 %	累积 %	正则相关性	函数检验	Wilks 的 Lambda	卡方	df	Sig.
1	747 <sup>a</sup>	100.0	100.0	.654	1	.572	100.694	5	.000

a. 分析中使用了前 1 个规范判别式函数。

图 20-22 特征值输出和 Wilks' Lambda 检验结果

标准化的规范判别式函数系数		结构矩阵	
	函数		函数
	1		1
总资产	-.256	资产收益率	.819
资产收益率	.982	净利润率 <sup>a</sup>	.543
速动比率	1.083	净资产收益率营业利润 <sup>a</sup>	.515
存货流动负债比率	.246	债务资本比率 <sup>a</sup>	-.337
现金流动负债比率	.810	存货流动负债比率	.221
		流动资产周转率 <sup>a</sup>	.193
		债务资产比率 <sup>a</sup>	-.129
		存货周转率 <sup>a</sup>	.110
		现金流动负债比率	.097
		应收账款周转率 <sup>a</sup>	.097
		资本充足率 <sup>a</sup>	.085
		总资产增长率 <sup>a</sup>	-.084
		流动比率 <sup>a</sup>	-.082
		所有者权益合计包含少数股东权益	.081
		固定资产周转率 <sup>a</sup>	.079
		速动比率	-.076
		总资产	.075
		营业利润增长率 <sup>a</sup>	.083
		资产周转率 <sup>a</sup>	-.020

判别变量和标准化规范判别式函数之间的汇总组间相关性

按函数内相关性的绝对大小排序的变量。

- a. 该变量不在分析中使用。

图 20-23 标准化系数和结构矩阵

规范判别式函数系数		组质心处的函数	
	函数		函数
	1		1
总资产	.000	06/07是否ST	
资产收益率	17.977	06/07未ST	348
速动比率	-1.122	06/07首ST	-2.126
存货流动负债比率	.035	在组均值处评估的非标准化规范判别式函数	
现金流动负债比率	1.068		
(常量)	.424		
非标准化系数			

图 20-24 未标准化的判别系数和类别质心函数

组的先验概率				分类函数系数		
06/07 是否ST	先验	用于分析的案例			06/07 是否ST	
		未加权的	已加权的		06/07 未ST	06/07 首ST
06/07 未ST	859	159	159 000	总资产	1.98E-010	3.42E-010
06/07 首ST	141	26	26 000	资产收益率	4.842	-39.820
合计	1 000	185	185 000	速动比率	2.784	5.560
				存货流动负债比率	.021	-.065
				现金流动负债比率	-2.392	-5.034
				(常量)	-1.383	-6.440

Fisher 的线性判别式函数

图 20-25 先验概率和 Fisher 判别函数系数

分类结果 <sup>a</sup>					
		06/07 是否ST	预测组成员		合计
			06/07 未ST	06/07 首ST	
初始	计数	06/07 未ST	156	3	159
		06/07 首ST	10	16	26
	%	06/07 未ST	98.1	1.9	100.0
		06/07 首ST	38.5	61.5	100.0

a. 已对初始分组案例中的 93.0% 个进行了正确分类。

图 20-26 判别结果总结表

(1) 警告信息和摘要信息。如图 20-20 所示，警告栏提示用户当前分类使用的是每个类别的协方差矩阵 (Separate-groups)，而不是合并的类内协方差矩阵 (Within-groups)。

“分类案例处理摘要”表格给出参与分析的数据信息，有效案例为 185 例，无缺失数据。

(2) 协方差检验结果。改进后判别模型的 Box's M 检验结果，与图 20-18 中所示的结果相同，即认为协方差相等的假设不成立。

(3) 变量选择过程的输出。如图 20-21 所示，给出了变量筛选的过程，在第 1 步加入了资产收益率变量，在第 5 步加入了总资产变量，并且每一步的 Wilks' Lambda 检验都很显著 (Sig 值均远小于 0.01)，这说明每一步加入的变量对正确判断分类都是有显著作用的。

(4) 典型判别函数的检验。如图 20-21 所示，由于只有一个典型判别函数，所以它解释了所有的变异，并记录在了“特征值”表格里。而 Wilks' Lambda 检验的 Sig 远小于 0.01，表示这个判别函数的判别作用是显著成立的。

(5) 标准化的典型判别系数。如图 20-23 所示，“系数”表格输出的是判别函数中各个变量的标准化系数，由此可以判断各函数主要受哪些变量的影响。“结构矩阵”表格给出的是判别变量和标准化判别函数之间的相关性数据，同样可以用来判断判别函数受哪些变量的影响较大。综合这两个表格的数据，认为此判别函数与资产收益率、存货流动负债比率、总资产的相关性较大。

(6) 未标准化的典型判别系数和类别质心函数。如图 20-24 所示，“函数系数”表格给出了非标准化的判别函数系数，利用它可以直接通过原始变量进行计算；而标准化的判别系数在使用时，需要先将原始变量标准化，不太方便。

“组质心处的函数”表格给出的是各个类别的重心在平面上的坐标，由于本例只有一个判别函数，所以每个类别只有一个输出值，可以将其看作是直线上的坐标。根据典型判别函数 (标准化的或未标准化的)，可以计算出每个观测的平面坐标，再计算出它们和各类别的重心的距离，就可以判断其类别归属了。

(7) Fisher 判别系数。使用典型判别系数 (标准化的或未标准化的) 时，对每个观测先要计算出平面 (或直线) 坐标值，然后比较与类别重心的距离，再进行判别归类。



相比而言,使用 Fisher 判别函数就要简单得多,对每个观测直接利用 Fisher 判别函数计算其属于各类的得分,并把此观测归入得分最高的一个类别即可。Fisher<sup>\*</sup>判别函数的输出图 20-25 所示。


(8) 最终判别的结果总结表。如图 20-26 所示,“分类结果”表格给出了典型判别函数的判别效果。首先,此判别模型对所有案例的分类准确率达到了 93%,比图 20-21 所示初始模型的 92.4%的分类准确率有所提高,由此说明使用 Separate-groups 选项还是较为合理的;其次,93%的判断准确率也是比较高的,说明此判别分析模型能很好的用来预测上市公司的财务预警问题。

具体看来,原始数据里未 ST 的 159 家公司,经过模型判别有 156 家(98.1%)仍判定为未 ST 的;原始数据里的首 ST 的 26 家公司,经过模型判别有 16 家(61.5%)仍判定为首 ST 的,有 10 家首 ST 公司的财务状况预测错误。从后验概率的角度看,预测出 19 家财务危机预警的上市公司里,有 16 家(84.2%)是真的发生了财务危机。

### 20.4.3 logistic 回归建模与分析

#### 1. 变量标准化处理

如图 20-10 所示,在初始的 19 个自变量里,除了总资产、所有者权益合计两个变量是取值较大的观测变量外,其他的都是取值较小的比率变量,故而考虑对总资产、所有者权益合计这两个变量进行标准化处理。

打开文件“过滤后的财务数据.sav”,依次单击菜单“Analyze→Descriptive Statistics→Descriptives...”执行描述性统计分析,其设置界面如图 20-27 所示。在变量列表选中总资产、所有者权益合计,单击  按钮将其选入 Variable(s) 分析变量列表;勾选 Save 复选框;单击 OK 运行后,当前数据集增加两个名为 Zx3、Zx4 的变量,分别存储变量 x3、x4 标准化后的数据,随后的回归分析将不再使用变量 x3 和 x4。

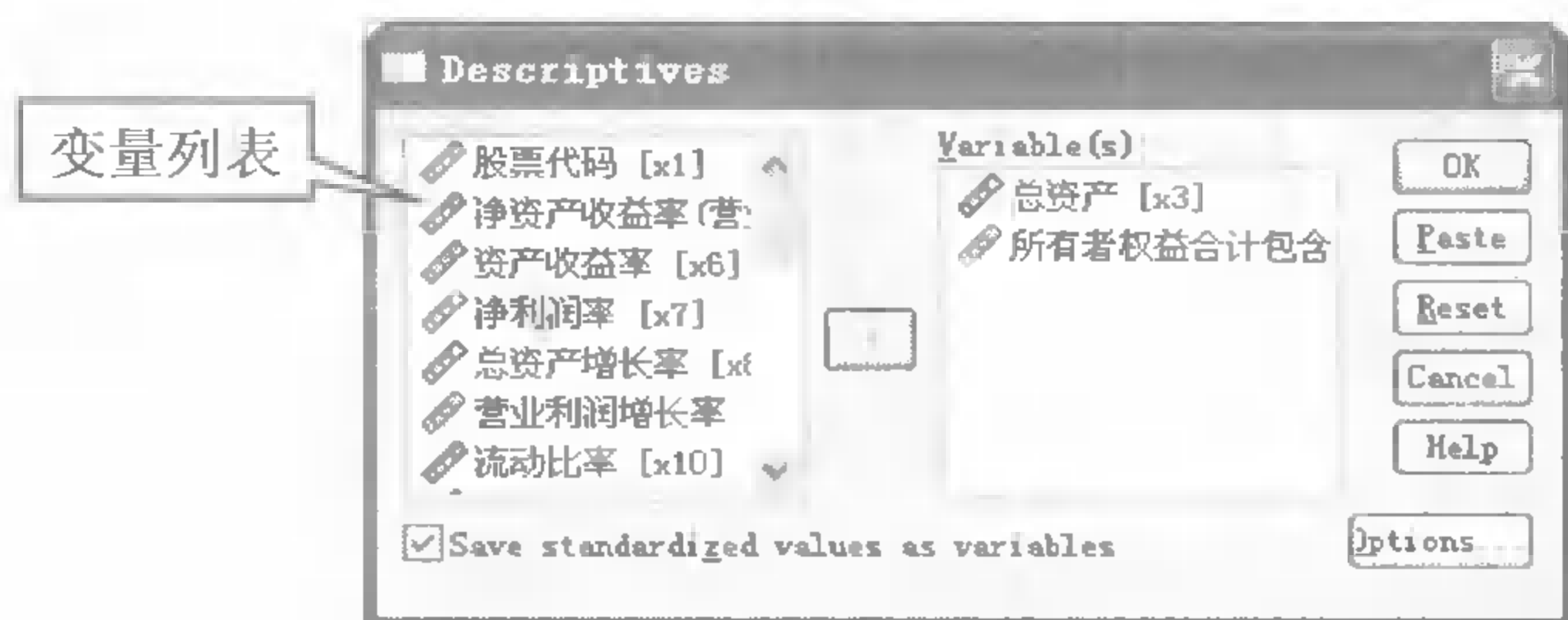


图 20-27 标准化处理

#### 2. SPSS 二元 Logistic 回归分析的参数设置

本例的目标变量只有 2 个取值,故选用二元逻辑回归模型进行分析。由于本例中的变量很多,而采用向后逐步法可能在最终回归方程保留更多的变量,读者可以自行选择向前法进行建模,通过比较会发现向后法的预测效果更好一点。

依次单击菜单“Analyze→Regression→Binary Logistic...”执行二元逻辑回归过程,主设置界面如图 20-28 所示。

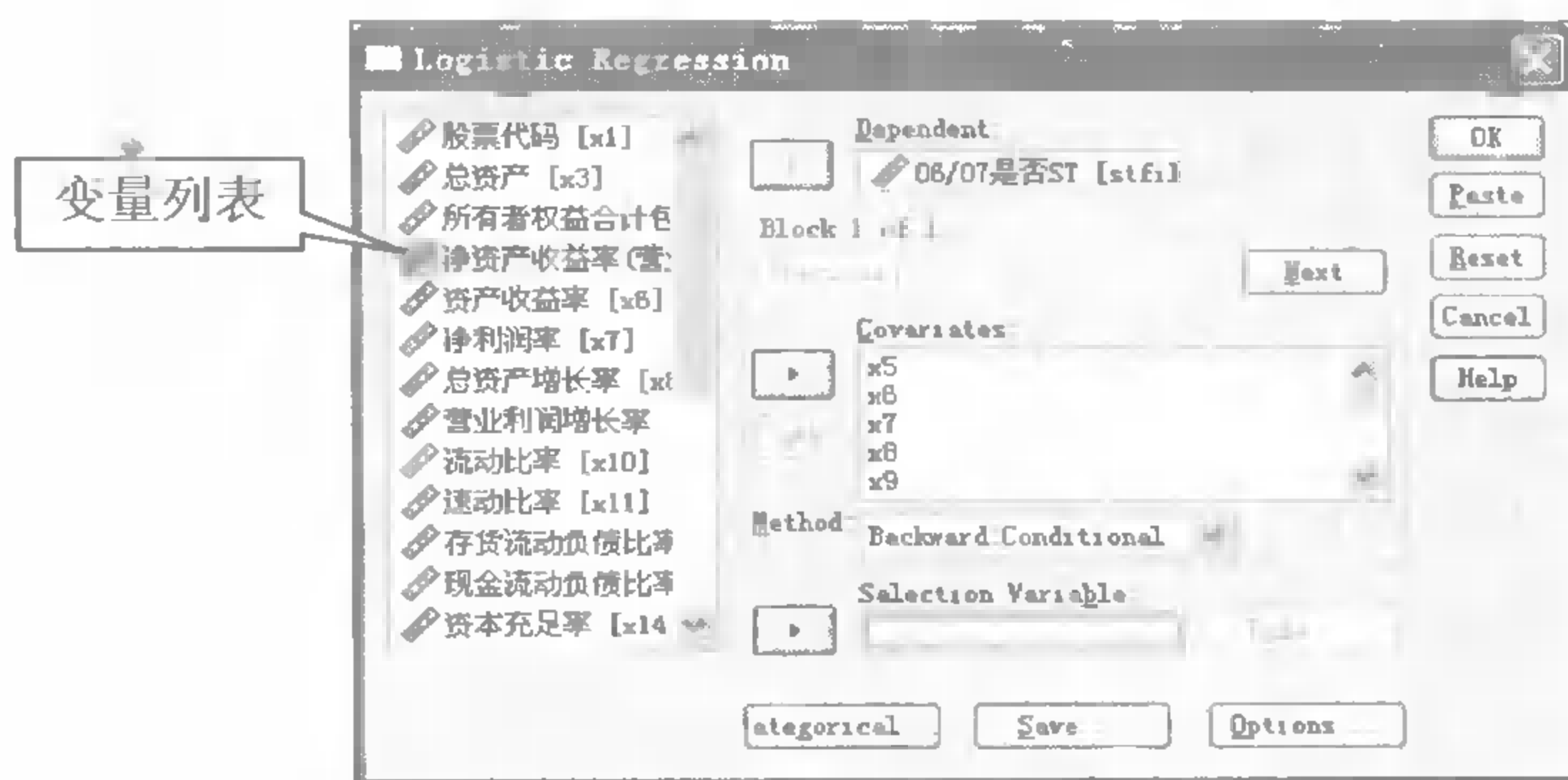


图 20-28 二元逻辑回归的参数设置

如图 20-28 所示, 在变量列表选中 06/07 是否 ST (stfilter) 变量, 单击从上至下第一个 按钮, 将其选入 Dependent 因变量选框; 在变量列表选中从 x5 至 x9 的所有变量 (不包括 x3、x4), 单击从上至下第二个 按钮, 将其选入 Independents 自变量列表; 单击 Method 栏后的下拉菜单, 选中 Backward: Conditional 向后逐步法。

在图 20-28 中, 单击 Options 按钮, 弹出如图 20-29 所示的选项设置面板, 在此选择输出统计量和输出图形。依次勾选如下选项: Classification plots 复选框、Hosmer-Lemeshow goodness-of-fit statistic 复选框、At last step 单选框、Include constant in model 复选框。单击 Continue 按钮返回主面板。

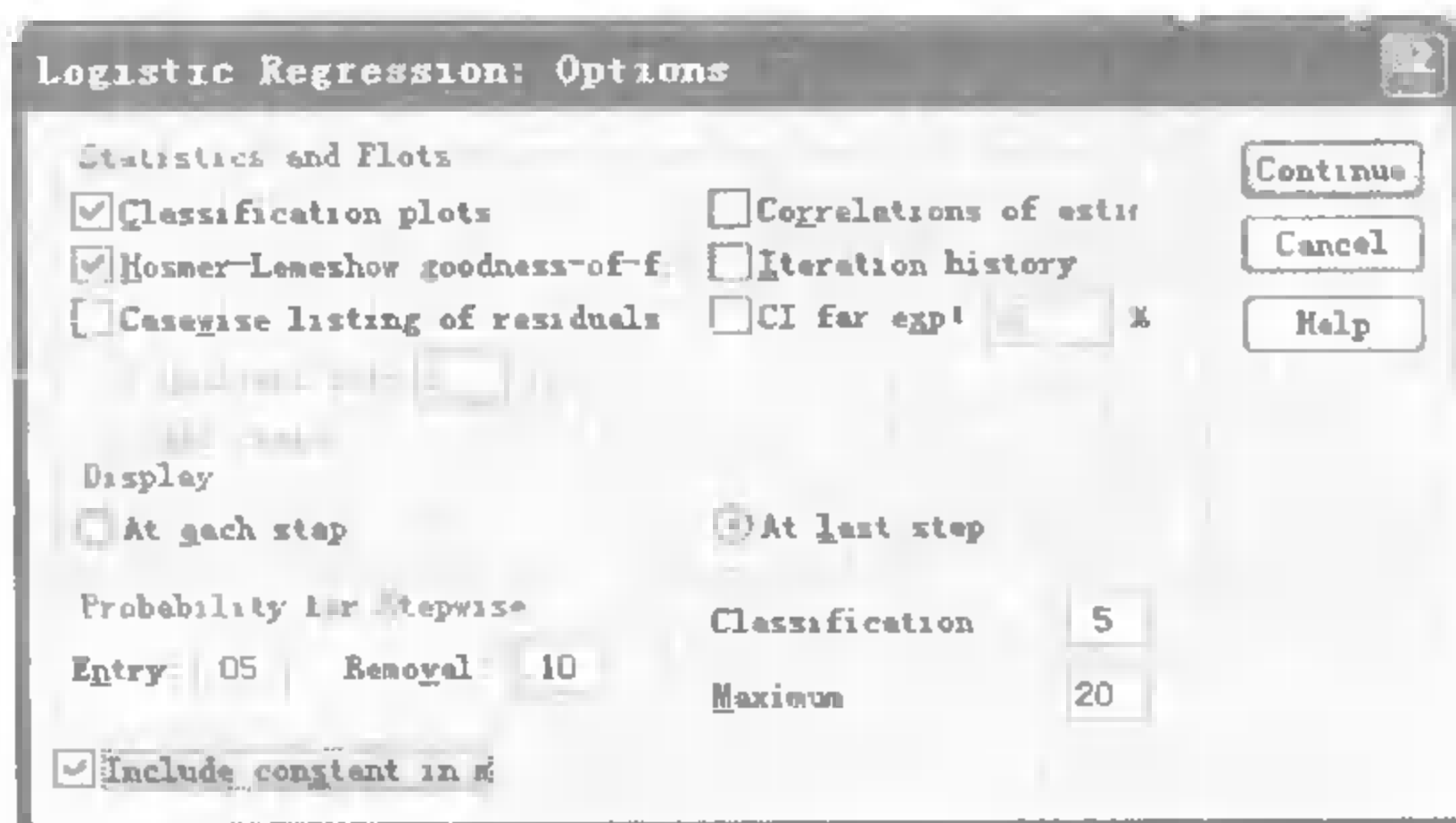


图 20-29 二元逻辑回归的 Options 选项设置

### 3. 结果分析

在图 20-28 中单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 20-30~图 20-34 所示。

案例处理摘要		
未加权的案例	N	百分比
已选定的案例 包括在分析中	185	100.0
缺失案例	0	0
总计	185	100.0
未选定的案例	0	0
总计	185	100.0

a. 如果权重有效, 请参见分类表以获得案例总数。

因变量编码	
初始值	内部值
06/07未ST	0
06/07首ST	1

图 20-30 数据摘要

Hosmer 和 Lemeshow 检验			
步骤	卡方	df	显著性
1	1.291	8	.996
2	1.295	8	.996
3	1.221	8	.996
4	1.209	8	.997
5	1.337	8	.995
6	1.364	8	.995
7	1.265	8	.996
8	1.207	8	.997
9	1.309	8	.995
10	1.171	8	.997
11	1.214	8	.996
12	5.463	8	.707
13	1.764	8	.987
14	4.006	8	.857

Hosmer 和 Lemeshow 检验的随机性表						
步骤		06/07是否ST = 06/07未ST		06/07是否ST = 06/07首ST		总计
		观察值	期望值	观察值	期望值	
14	1	19	19.000	0	.000	19
	2	19	18.998	0	.004	19
	3	19	18.969	0	.031	19
	4	19	18.901	0	.099	19
	5	19	18.728	0	.272	19
	6	18	18.401	1	.599	19
	7	16	17.624	3	1.376	19
	8	18	16.759	1	2.241	19
	9	12	11.274	7	7.726	19
	10	0	.347	14	13.653	14

图 20-31 Hosmer-Lemeshow 检验表

方程中的变量						
步骤	变量	B	S.E.	Wald	df	显著性
14	x6	-100.192	26.387	14.417	1	.000
	x7	5.779	2.370	5.943	1	.015
	x12	-4.570	1.981	5.322	1	.021
	x14	-14.880	6.615	5.060	1	.024
	x16	-17.276	6.397	7.293	1	.007
	x20	4.615	2.079	4.930	1	.026
	常量	12.813	6.233	4.226	1	.040
		Exp(B)				
		.000				
		323.341				
		.010				
		.000				
		.000				
		101.027				
		366870.55				

a. 在步骤 1 中输入的变量: x5, x6, x7, x8, x9, x10, x11, x12, x13, x14, x15, x16, x17, x18, x19, x20, x21, Zx3, Zx4

不在方程中的变量 <sup>a</sup>				
步骤	变量	得分	df	显著性
14	x5	.036	1	.849
	x6	.097	1	.755
	x9	.071	1	.791
	x10	.268	1	.604
	x11	.268	1	.604
	x13	1.368	1	.242
	x15	.036	1	.850
	x17	.253	1	.615
	x18	.585	1	.444
	x19	.936	1	.333
	x21	3.205	1	.073
	Zx3	.769	1	.381
	Zx4	.993	1	.319

图 20-32 在与不在方程中的变量

步骤摘要 <sup>a, b</sup>								
步骤	改进			模型			更正类 %	变量
	卡方	df	显著性	卡方	df	显著性		
2	.008	1	.929	99.028	18	.000	95.7%	OUT x17
3	.047	1	.828	98.981	17	.000	95.1%	OUT x15
4	.054	1	.816	98.927	16	.000	95.1%	OUT x8
5	.061	1	.776	98.846	15	.000	95.1%	OUT Zx3
6	.132	1	.716	98.714	14	.000	95.7%	OUT x18
7	.083	1	.773	98.631	13	.000	95.1%	OUT x5
8	.229	1	.633	98.402	12	.000	95.1%	OUT x9
9	.260	1	.597	98.122	11	.000	95.1%	OUT x11
10	.070	1	.792	98.053	10	.000	95.1%	OUT x10
11	.492	1	.483	97.561	9	.000	95.1%	OUT x21
12	1.231	1	.267	96.330	8	.000	95.1%	OUT x19
13	1.730	1	.188	94.599	7	.000	94.6%	OUT Zx4
14	1.640	1	.200	92.959	6	.000	95.1%	OUT x13

a. 无法从当前模型中删除更多变量或向其添加更多变量。  
b. 结束块 1

图 20-33 逐步回归的步骤摘要

分类表 <sup>a</sup>					
观察值			预测值		
			06/07是否ST		百分比校正
			06/07未ST	06/07首ST	
步骤 14	06/07是	06/07未ST	157	2	98.7
	否ST	06/07首ST	7	19	73.1
总百分比					95.1

a. 切割值为 .500

图 20-34 二元 Logistic 回归模型的预测分类

① 数据摘要信息。如图 20-30 所示,“案例处理摘要”表格给出了原始数据中用于分析的案例、缺失案例的统计信息,显示所有 185 个案例都用来建模,没有缺失信息。“因变量编码”表格给出了目标变量的编码取值,06/07 首 ST 的取值为 1。

② Hosmer-Lemeshow 检验结果。如图 20-31 所示,“Hosmer 和 Lemeshow 检验”表检验的零假设为:模型能够很好的拟合数据。从最终模型的显著性检验 Sig=0.857>0.5 来看,不能否定零假设,即认为模型能够很好的拟合数据。

“Hosmer 和 Lemeshow 检验的随机性”表格,根据目标变量的预测概率把结果分为个数大致相等的 10 个组,“总计”列中是每组的测量数,由于预测值相等的观测被分在一起,所以各组的观测数不一定相同。表中各行的观测值和预测值都大致相同,所以模型拟合效果不错,此表直观地反应了模型的预测效果,比较可信。

③ 进入模型的变量。如图 20-32 所示,给出了最终模型(第 14 步)中包含的和未包含的变量的统计量信息:方程中的变量显著性都小于 0.05,而不在方程中的变量显著性都大于 0.05,这说明最终回归方程中的各自变量对方程的贡献都是显著的。

“方程中的变量”表格还给出了最终模型的系数估计值,由“B”列的系数可得二元 Logistic

模型： $p=1/1+e^{-z}$  ( $z=12.813-100.192x_6+5.779x_7-4.57x_{12}-14.88x_{14}-17.276x_{16}+4.615x_{20}$ )。由此可见，B 列的系数是线性的，用它来检验系数的显著性比较方便。但在 Logistic 回归里，Exp (B) 列的系数更易于解释，它反应了自变量变动 1 个单位而引起的发生比 Odds 的变化率，可见变量  $x_7$  (净利润率) 和  $x_{20}$  (资产周转率) 对 Odds 比的影响最大。

利用此处得到的最终模型，就可以对上市公司的财务情况进行预测，当预测概率大于 0.5 时，就推断其两年后会发生财务危机，反之推断其两年后不会发生财务危机。

④ 逐步回归的步骤总结。如图 20-33 所示，给出了逐步回归各步骤的摘要信息，对于向后逐步法，最后一列显示了各步移出的变量，并且显示了每一步改进后模型的显著性卡方检验。从最终模型的卡方检验显著性值 (图中蓝色线框标识) 远小于 0 看，模型整体的线性关系是显著成立的。

⑤ 预测结果的总结。如图 20-34 所示，利用二元 Logistic 回归的整体预测准确率达到了 95.1%，与判别分析方法 93% 的准确率相比，Logistic 回归模型更适合用来预测上市公司的财务情况。

具体看来，原始数据里未 ST 的 159 家公司，经过模型判别有 157 家 (98.7%) 仍判定为未 ST 的；原始数据里首 ST 的 26 家公司，经过模型判别有 19 家 (73.1%) 仍判定为首 ST 的，有 7 家首 ST 公司的财务状况预测错误。从后验概率的角度看，预测出 21 家财务危机预警的上市公司里，有 19 家 (90.5%) 是真的发生了财务危机。

## 20.5 进一步的分析与应用

前面对各模型的输出结果做了较为详细的说明，总体看来两个模型的预测效果都不错，而且 Logistic 回归模型更好一点，下面从应用方面再作更进一步的分析。

### 20.5.1 分类结果的应用分析

如图 20-35 所示，将两个模型的预测效果表放在一起进行观察，下面计算预测的后验概率。

判别分析结果					Logistic 回归结果				
分类结果 <sup>a</sup>					分类表 <sup>a</sup>				
		预测组成员		合计			预测值		百分比校正
初始	计数	06/07是否ST	06/07未ST		观察值	06/07是否ST	06/07未ST	06/07首ST	
	06/07未ST		156	3	159				
	06/07首ST		10	16	26				
%	06/07未ST		98.1	1.9	100.0				
	06/07首ST		38.5	61.5	100.0				
* 已对初始分组案例中的 93.0% 个进行了正确分类。					步骤 14	06/07是否ST	06/07未ST	06/07首ST	百分比校正
							157	2	98.7
							7	19	73.1
									95.1
					a. 切割值为 .500				

图 20-35 两种模型的预测结果深入分析

对于判别分析结果，预测某公司 2 年后首 ST，该公司在 2 年后真被 ST 的后验概率为： $16/(3+16)=84.2\%$ ；而对所有 26 家首 ST 公司，预测出了其中的 61.5%，仍有 10 家漏网。

对于 Logistic 回归结果，预测某公司 2 年后首 ST，该公司在 2 年后真被 ST 的后验概率为： $19/(2+19)=90.5\%$ ；而对所有 26 家首 ST 公司，预测出了其中的 73.1%，只有 7 家漏网。

采用 Logistic 模型进行预测时，如果推断某公司将出现财务危机，那么这个推断的可信度能高达 90% 以上；而且能够找出多于 2/3 的确实会发生财务危机的公司。由此可见，Logistic 模型对于实际预测的应用是很有价值的。



## 20.5.2 建模方法的改进

本章的两个模型都是直接使用所有的原始数据进行建模，这样对于所得模型的稳定性和可推广性不能给出很好的建议。故而建议把原始数据分为训练集和验证集两个样本，利用训练集的样本估计模型系数，再利用由此得到的模型对训练集和验证集分别进行预测归类，然后选择对验证集能够进行较好分类（准确率高、或感兴趣的后验概率高）的结果作为最终模型；这是因为验证集的数据没有用于建模，对它有较好预测效果的模型更加稳定和实用。

关于对样本进行随机抽样，以及利用抽样结果把样本进行划分的操作，请参考第 8.4.3 节第 2 步的实例。如果已经有了对样本进行划分的筛选变量，就可以直接在图 20-13 或图 20-28 中的主设置面板，通过在 Selection 栏设置筛选变量及其取值来进行训练集的选择，取筛选变量的指定取值的案例将作为训练集，其他案例作为验证集。

## 20.6 建议和推广

### 20.6.1 时间序列研究

鉴于上市公司发生财务危机是一个随时间逐步演变的过程，所以采用时间序列模型进行研究具有一定的合理性，而且可以充分利用更多的历史数据。已有研究也表明，用时间序列对上市公司进行远期财务危机预警模型是可行的，通过对公司自上市以来的财务数据建立自回归模型，用其预测后  $k$  年的财务数据，可以达到远期预警的效果，以满足投资者、贷款者规避风险的要求，帮助其作出决策；而自回归函数的系数反映了不同会计年份在财务危机预警模型中的影响程度，符合越近的会计年度影响越大的实证结果。

### 20.6.2 数据的有效预警期

关于财务危机的有效预警期，周咏梅（2008）以 2000~2002 年的制造业上市公司为样本，选取了 54 家处于财务危机的公司和 54 家财务状况正常的公司，采用多元线性回归方法对公司特别处理前五年的数据分别进行回归。发现不同年份回归模型不完全一致，在前五年的模型中，资产利润率比较显著；在前三年和前四年的模型中，股东权益/总资产也开始显著化；在前二年和前一年的模型中，有关主营业务的指标亦开始显著。这说明上市公司财务危机确实存在着不同的阶段，在不同阶段各财务指标的信息含量也不尽相同；同时也说明了上市公司的财务报表具有预测信息含量，这种信息含量随着被特别处理的年份越远而愈小。

韩德宗（2005）指出，财务危机的有效预警期以第  $t-4$  年为起点，而在第  $t-5$  年时，ST 公司和非 ST 公司之间的差异是不明显的；这与本章所采用的  $t-2$ 、 $t-3$  年为有效预警期有一定的差别。

建议研究者在选取数据的有效预警期时，要根据实际数据的特点和所研究问题的侧重点等因素加以综合考虑，并在第  $t-1$  年至第  $t-5$  年之间进行选择。

### 20.6.3 指标的简化方法

财务危机预警模型中，用到的财务指标（及其他指标）较多，有的研究者甚至用到了数

十上百种的指标，因此进行适当的指标约简是有必要的，这样可以更加清晰明了的理清各指标之间的关系，以及真正在模型中起关键作用的指标。

根据前人的经验，对财务预警建模的步骤作以简单概括。

(1) 数据的描述性分析，了解数据基本特点，如均值、方差等。

(2) 各指标的单变量分析，如方差分析、t 检验等，此步骤可以约简部分在两个类别间差异不显著的指标。

(3) 指标的降维约简，如主成分分析、因子分析等，此步骤可以把指标进行有意义的归类和约简，并由此降低建模的复杂性，以及增强结果解释的清晰性和合理性。

(4) 用约简后的指标拟合预警模型，模型可以是判别分析、Logistic 回归、神经网络等多种情况。

(5) 对模型进行评价，然后用模型进行预测等应用。

# 第 21 章 影响汇率的因素分析

汇率是在商品交易和货币运动越出国界时产生的，又是一国货币价值在国际上的表现。因为一国货币汇率受制于经济、政治、军事、甚至心理因素的影响，这些因素彼此之间既相互联系，又相互制约，而且在不同的时间，各因素产生作用的强度也会出现交替变化。在某一时起主要作用的因素，在另一时期可能不起作用；即便在同一时期，同一因素对不同国家和地区货币汇率的影响也很不同。所以很难准确地找出究竟有哪些因素影响着一国货币汇率的变化，在此引用美联储主席格林斯潘的话以反映汇率变化的不确定性：“在过去 50 多年中，我一直试图预测汇率的变化趋势，但我不得不承认自己在这方面的能力有限”。

本章仅选择经济统计中的几个基本指标，来说明其对我国汇率变动的影响。

## 21.1 汇率影响因素的简介

在开放经济中，汇率是一种重要的资源配置价格。汇率的失衡或错估，不仅会破坏经济的外部均衡，而且会给国内宏观经济稳定和可持续的经济增长带来一系列不利影响。另外，汇率的变化还能对人们的日常生活和企业的生产销售产生较大的影响。所以对影响汇率的因素进行分析和探讨，对于指导汇率政策的制定、预测汇率变化趋势、优化投资策略，以及研究与汇率有关的生产消费等问题都有重要的应用价值。

汇率就是两种不同货币之间的比价，反映一个国家货币的对外价值。一个国家汇率的变动要受到许多因素的影响，包括经济因素、政治因素、心理因素等。下面选择几个比较重要的因素，来简单分析一下它们对汇率变动的影响。

(1) 国际收支。国际收支状况是一个国家汇率变动的直接原因，一个国家国际收支发生顺差，就会引起外国对该国货币需求的增长与外国货币供应的增加，顺差国的汇率就上升；反之，一个国家国际收支逆差，它的货币汇率就下降。例如：近年来中国对美国的贸易顺差，使得人民币对美元的汇率连连攀升，而汇率的升高，又会反过来制约贸易顺差的进一步扩大。

(2) 经济增长率。在不考虑其他因素的条件下，两国经济增长的差异，往往构成汇率变动的基础，因为它会影响对外贸易和外汇市场交易活动的变化。一般而言，经济增长加速，会导致国内需求水平提高，从而引起更多的进口，由此造成本国货币汇率向下的压力。

(3) 通货膨胀。一国货币价值的总水平是影响汇率变动的一个重要因素，它影响着—个国家商品、劳务在世界市场上的竞争能力。由于通货膨胀，国内物价上涨，一般会引起出口商品的减少和进口的增加。这些变化通过影响外汇市场上的供求关系和该国货币在国际上的信用地位，导致汇率下跌。例如：1974~1975 年，美国国内通货膨胀率从 11%降为

9%，同时美元汇率保持上升趋势；1977～1978 年，美国通货膨胀率上升，立即引起美元汇率的下跌。

(4) 财政赤字。政府的财政赤字常常用作预测汇率变化的重要指标。如果一个国家的财政预算出现巨额赤字，则意味着政府支出过度，一方面可能引起通货膨胀的上升，另一方面可使国家收支恶化，两者都会导致汇率的自动下浮。

(5) 利率。利率下降，国内资本流出；利率上升，国外资本流入。这种由两地利差引起的套利活动是国际资金流动的一种方式。资本流动将引起外汇市场供求变化，从而对汇率产生影响。在通常情况下，一个国家利率提高、信用紧缩，将导致该国货币升值；反之引起货币的贬值。

(6) 外汇储备。中央银行的外汇储备表明一个国家干预外汇市场和维持汇价的能力，所以它对稳定汇率有一定的作用。英国政府从 1932 年起就用部分外汇储备设立外汇平准基金，当英镑汇率下跌时，就卖出外汇买入英镑，促使英镑汇率上升；当英镑汇率过高时，就买入外汇，卖出英镑，使英镑汇率下跌。后来，美国、加拿大、瑞士等国纷纷效仿，设立外汇平准基金，这一做法一直沿用到现在。需要指出的是，只有一个国家拥有足够的外汇储备时，才能有效地干预外汇市场，影响汇率短期变动的方向与幅度。

(7) 心理预测。现实生活中，投资者往往根据自身对未来汇率的主观评价来决定资本转移的数量和方向，这对外汇市场有很大的影响，往往起到加大外汇波动幅度的作用。例如 1995 年日元对美元的汇率一度持续上升，其中市场交易者的心理预期起到了很大作用。

本章选取了一些能够反映如上部分经济特征的指标，包括 GDP、通货膨胀率、利率、净出口规模 and 外汇储备等，定量地研究这些因素对汇率的影响。

## 21.2 数据描述

本章选取 1985～2000 年的有关数据，分析人民币汇率及其影响因素的相关性。所用数据参考自“人民币汇率研究”（陈璠，CENET 网刊，2005）、“汇率决定模型与中国汇率走势分析”（孙煜，复旦大学《经济学人》，2004）和“人民币汇率的影响因素与走势分析”（徐晨，对外经济贸易大学在职人员以同等学力申请硕士学位论文，2002），其中通货膨胀率、一年期名义利率、美元利率和汇率 4 个指标的数据摘自《中国统计年鉴》（2001 年，中国统计出版社）；2000 年的部分数据来自国家统计局官方网站。

由于影响汇率的因素多种多样，此处只选择了几个能反映国民经济发展不同方面的指标，分析它们和汇率变化的影响关系。要精确把握汇率的变化趋势，还需要考虑政治和政策因素、国际环境因素和人民的心理因素等诸多方面。

表 21-1 为本章采用的原始数据。由于数据的度量单位不统一，包括比率、亿元、亿美元，故而建议在进行相关分析前，对数据做适当的标准化处理。

表 21-1

1985～2000 年的汇率及其影响因素数据

年份	通货膨胀率	一年期名义利率	美元利率	汇率	GDP (亿元)	净出口 (亿美元)	居民总储蓄 (亿元)	居民消费 (亿元)	外商直接投资(亿美元)	实使外资 (亿美元)	外汇储备 (亿美元)	外债规模 (亿美元)
1985	8.8	7.2	8	2.94	8 964.4	-149	3 042.7	4 625.7	63.33	19.56	26.44	158.3
1986	6	7.2	6.5	3.45	10 202.2	-119.7	4 163.36	5 214.1	33.3	22.44	20.72	214.8
1987	7.3	8.64	7	3.72	11 962.5	-37.7	4 642.08	6 011.5	37.09	23.14	29.23	302



续表

年 份	通货 膨胀率	一年期 名义利率	美元 利率	汇 率	GDP (亿元)	净出口 (亿美元)	居民总储蓄 (亿元)	居民消费 (亿元)	外商直接投 资(亿美元)	实使外资 (亿美元)	外汇储备 (亿美元)	外债规模 (亿美元)
1988	18.5	11.34	7.5	3.72	14928.3	-77.5	7 099.94	7 694.1	52.97	31.94	33.72	400
1989	17.8	10.08	9	3.76	16909.2	-66	6 996.74	8 588	56	33.93	55.5	413
1990	2.1	8.64	8	4.78	18 547.9	87.4	11 078.36	9 180.9	65.96	34.87	110.93	525.5
1991	2.9	7.56	5.5	5.32	21 617.8	80.5	8 478.02	10 377.7	119.77	43.66	217.12	605.6
1992	5.4	9.18	4	5.51	26 638.1	43.5	13 580.77	12 537.3	581.24	110.08	194.43	693.2
1993	13.2	10.98	3.5	5.75	34 634.4	-122.2	19 211.61	15 774.6	1 114.36	275.15	211.99	835.7
1994	21.7	10.98	5	8.61	4 662.3	54	26 949.28	20 925.8	826.8	337.67	516.2	928.1
1995	14.8	10.98	6	8.35	58 260.5	167	39 620.83	27 082.7	912.82	375.21	735.97	1 065.9
1996	6.1	7.47	5	8.32	67 800	122.2	34 055.32	32 322.9	732.76	417.26	50.29	1 162.8
1997	0.8	7.47	5.25	8.29	74 462.6	404.2	28 936.78	35 035.6	510.03	452.57	398.9	1 309.6
1998	-1.7	4.59	5.47	8.28	79 552.8	435.7	24 281.39	37 093.5	521.02	454.63	449.59	1 460.4
1999	-3	5.33	5.33	8.28	82 054	292.3	19 214.44	39 510.2	412.23	403.19	546.75	1 518.3
2000	-1.5	6.46	6.46	8.28	89 404	241.1	30 340.47	43 000.25	623.8	407.15	655.74	1 457.3

### 21.3 分析方法概述

#### 21.3.1 探索性分析

探索性数据分析是 20 世纪后半叶在西方最先兴起的统计分析方法，1977 年美国统计学家 Tukey 出版了《探索性数据分析》一书，成为该领域的第 1 个正式出版物。

探索性数据分析可以作为我们认识数据和问题的有力工具，同时也是正式建立统计分析模型之前的铺垫，是科学的统计分析的一个重要环节。许多研究者往往一开始就应用经验对数据进行建模分析，而忽略了对统计数据的探索研究这一过程，这样容易使其分析只停留在问题的表面而不能够深入。

探索性数据分析有如下 3 个方面的特点。

(1) 研究直接从原始数据入手，让数据说话。许多统计方法要先假定数据服从某种分布，然后用适应该种分布的模型进行拟合、分析和预测。但客观实际的数据并不总是满足理论上的分布，因而这些方法具有极大的局限性。探索性数据分析完全从客观数据出发，而不是从某种假定出发，到实际数据中去探索内在的数量规律性。

(2) 从实际出发，不以某种理论为根据。多数的统计模型都是以概率论为理论基础的，并对估计、检验和预测等方法能够给出精确度的度量方法和度量值。探索性数据分析在研究数据的内在特征、数量间的关系和变化时，所用方法会尽可能服从于数据特点和研究目的，并且更重视数据特征的稳健性，而相对放松对概率论理论和精确度的刻意追求。

(3) 分析工具简单直观，更易于普及。专业统计方法运用的数学理论越来越深，这样就使应用的人们越来越害怕统计。探索性数据分析运用简单直观的茎叶图、箱线图、残差图和字母值、数据变换、中位数平滑等方法，使具有初等数学知识的人就可以进行分析。

### 21.3.2 多元回归分析

多元回归分析 (Multiple Regression Analysis) 是研究单个因变量和多个自变量之间的相关关系的多元统计分析方法。比较基础且常用的是多元线性回归分析 (Multiple Linear Regression Analysis), 许多非线性回归 (Non-linear Regression) 和多项式回归 (Polynomial Regression) 都可以化为多元线性回归来解决, 因而多元线性回归分析有着广泛的应用。

#### 1. 多元线性回归的数学模型

假定因变量  $y$  与自变量  $x_1, x_2, \dots, x_m$  之间存在着线性关系, 其理论的数学模型为  $y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \dots + \beta_m x_{mj} + \varepsilon_j$ , ( $j=1, 2, \dots, n$ ); 式中  $\varepsilon_j$  为相互独立且都服从  $N(0, \sigma^2)$  的随机变量。

根据理论模型, 可以建立  $y$  对  $x_1, x_2, \dots, x_m$  的  $m$  元线性回归方程如下:  $\hat{y} = b_0 + b_1 x_1 + b_2 x_2 + \dots + b_m x_m$ ; 其中  $b_0, b_1, b_2, \dots, b_m$  为  $\beta_0, \beta_1, \beta_2, \dots, \beta_m$  的最小二乘估计值, 即  $b_0, b_1, b_2, \dots, b_m$  使实际观测值  $y$  与回归估计值  $\hat{y}$  的偏差平方和  $Q = \sum_{j=1}^n (y_j - \hat{y}_j)^2$  达到最小。

通过样本数据估计出回归方程的系数  $b_0, b_1, b_2, \dots, b_m$  后, 还需要进行多元线性模型的拟合优度检验和回归系数的显著性检验。利用  $t$  检验和  $F$  检验, 可以判断所建立的线性回归模型以及自变量与因变量之间的线性关系是否显著地成立。

#### 2. 利用多元线性回归进行相关性分析的基本步骤

运用多元回归进行预测分析或相关因素分析时, 基本思路是利用统计数据建立多元线性回归方程, 然后检验回归系数的显著性, 通过对各个因素进行逻辑检验和相关性检验, 决定各个因素的取舍, 逐步筛选出对因变量最有影响的因素。这个过程与逐步回归类似, 下面是分析步骤的简单总结。

① 利用统计数据建立多元线性回归预测模型。

② 对多元回归方程进行分析检验, 剔除不相关变量。

③ 重新建立多元线性回归预测模型, 再进行检验, 重复直至回归模型中没有不符合逻辑的变量, 且所有自变量都对因变量有显著影响为止。

## 21.4 SPSS 建模过程和结论分析

本节首先建立包含影响汇率因素的数据文件, 然后通过探索性数据分析方法研究各因素对汇率的影响程度, 接着建立关于汇率影响因素的多元回归分析模型。

### 21.4.1 数据准备

#### 1. 数据录入

本章所用原始数据如表 21-1 所示, 将其录入 (或直接复制) 到 SPSS 数据窗口, 格式如图 21-1 所示, 将其存为文件 “影响汇率的数据.sav”。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	x1	Numeric	8	0	年份	None	None	5	Right	Scale
2	x2	Numeric	8	2	汇率	None	None	5	Right	Scale
3	x3	Numeric	8	2	通货膨胀率	None	None	6	Right	Scale
4	x4	Numeric	8	2	一年期名义利率	None	None	5	Right	Scale
5	x5	Numeric	8	2	美元利率	None	None	5	Right	Scale
6	x6	Numeric	8	2	GDP(亿元)	None	None	7	Right	Scale
7	x7	Numeric	8	2	净出口(亿美元)	None	None	7	Right	Scale
8	x8	Numeric	8	2	居民总储蓄(亿元)	None	None	7	Right	Scale
9	x9	Numeric	8	2	居民消费(亿元)	None	None	6	Right	Scale
10	x10	Numeric	8	2	直接投资	None	None	6	Right	Scale
11	x11	Numeric	8	2	外资金额(亿美元)	None	None	6	Right	Scale
12	x12	Numeric	8	2	实际使用外资金额(亿美元)	None	None	6	Right	Scale
13	x13	Numeric	8	2	外汇储备(亿美元)	None	None	6	Right	Scale

图 21-1 影响汇率的数据格式

## 2. 变量的标准化处理

在初始的 12 个自变量里, 变量的取值单位有比率、亿元、亿美元, 度量方式不统一, 所以有必要先对它们进行标准化处理。

打开文件“影响汇率的数据.sav”, 依次单击菜单“Analyze→Descriptive Statistics→Descriptives...”执行描述性统计分析过程, 其设置界面如图 21-2 所示。

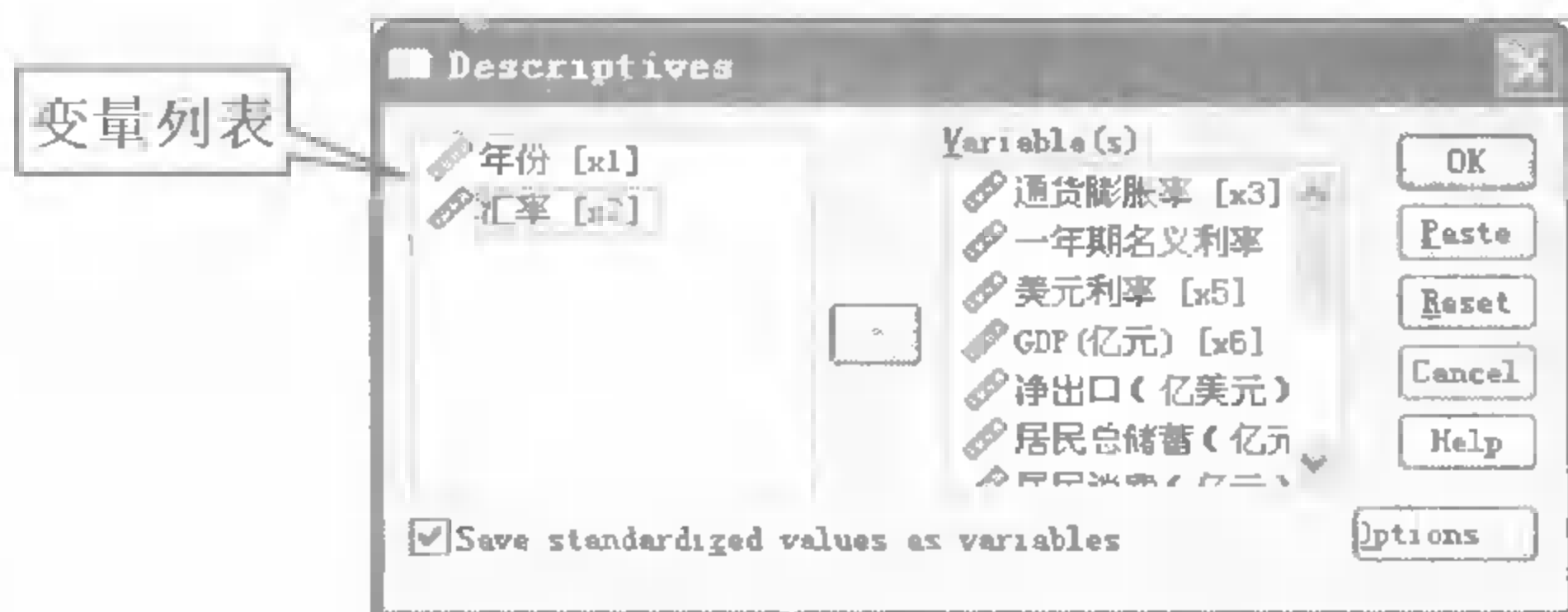



图 21-2 标准化处理

在变量列表中选中 x3~x13 的所有变量, 单击  按钮将其选入 Variable(s)列表作为分析变量; 勾选底部的 Save 复选框, 保存标准化后的数据。单击 OK 按钮运行后, 在当前数据集中增加名为 Zx3~Zx13 的新变量, 分别是变量 x3~x13 标准化后的数据, 随后的回归分析就使用 Zx3~Zx13 作为自变量, 而不再使用 x3~x13 了。

### 21.4.2 探索性分析

#### 1. 描述性统计分析

依次单击菜单“Analyze→Descriptive Statistics→Descriptives...”执行描述性统计分析过程, 设置界面如图 21-2 所示, 单击取消底部的 Save 复选框; 单击 Options 按钮, 弹出如图 21-3 所示的选项设置面板, 依次单击选中如下统计量: Mean (均值)、Std (标准差)、Minimum (最小值)、Maximum (最大值)、Kurtosis (峰度) 和 Skewness (偏度); 单击 Continue 按钮返回主面板。

在图 21-2 中, 单击 OK 按钮运行, SPSS Viewer 窗口的输出表格如图 21-4 所示。



图 21-3 描述性统计分析的 Options 设置

描述统计量									
	N	极小值	极大值	均值	标准差	偏度		峰度	
	统计量	统计量	统计量	统计量	统计量	统计量	标准误	统计量	标准误
汇率	16	2.94	8.61	6.0850	2.19050	-.072	.564	-1.864	1.091
通货膨胀率	16	-3.00	21.70	7.4500	7.72839	.458	.564	-.910	1.091
一年期名义利率	16	4.58	11.34	8.3813	2.08444	-.096	.564	-.886	1.091
美元利率	16	3.50	9.00	6.0944	1.51353	.247	.564	-.466	1.091
GDP (亿元)	16	4862.30	89404.00	38787.583	30656.516	.549	.564	-1.513	1.091
净出口 (亿美元)	16	-148.00	435.70	84.7375	183.82747	.606	.564	-.580	1.091
居民总储蓄 (亿元)	16	3042.70	39620.83	17805.756	11858.736	.394	.564	-1.187	1.091
居民消费 (亿元)	16	4825.70	43000.25	19685.928	13771.323	.505	.564	-1.453	1.091
“直接投资	16	33.30	1114.36	416.4675	363.32503	.412	.564	-1.113	1.091
外资金额 (亿美元)	16	19.56	454.63	215.1531	186.66141	.110	.564	-2.060	1.091
实际使用外资金额 (亿美元)	16	20.72	735.97	265.8450	247.44543	.662	.564	-1.018	1.091
外汇储备 (亿美元)	16	158.30	1518.30	815.6563	468.79037	.184	.564	-1.406	1.091
有效的 N (列表状态)	16								

图 21-4 描述性统计量输出

通过观察每个变量的描述性统计信息，可以了解这个变量的极值情况（最大值和最小值）、取值波动情况（标准差）以及分布情况（峰度和偏度）。从各变量的取值范围看，相差的数量级很大，故进行标准化处理很有必要；从峰度、偏度的取值看（都接近 0），各变量都没有很过分地偏离正态分布。

## 2. 其他检验和分析的简介

如果对单个变量的分布情况感兴趣，可以通过直方图、P-P 图等手段验证它们和某种分布的相似程度。对时间序列变量（如汇率），还可以作出它的趋势图加以观察和分析。

通过相关分析或一元回归分析，能够研究单个因素与汇率之间的相互影响关系。



另外，对这些变量还可以实行特定的转换（比如对数转换），观察转换后的数据特点，看它是否更加适用于描述这个案例。

研究者对哪些变量更感兴趣，可以自行对它进行更多的研究。随后，我们将采用多元回归分析来对汇率和这些因素之间的关系加以深入研究和描述。

### 21.4.3 多元回归分析

本案例的自变量较多且它们之间可能存在着共线性问题，故应采用逐步回归分析法。经过试验发现，向前逐步法只能在最终模型保留 1 个变量，而向后逐步法能在最终模型保留多个变量，可见向后逐步回归法更能充分利用本例的数据，下面就采用向后回归法进行分析。

#### 1. SPSS 参数设置

依次单击菜单“Analyze → Regression → Linear...”执行多元线性回归分析过程，其主界面如图 21-5 所示。在变量列表中选中汇率（x2）变量，单击从上至下第一个  按钮将其选入 Dependent 选框作为因变量；在变量列表中选中 Zx3~Zx13 的所有变量，单击从上至下第二个  按钮将其选入 Independent(s)列表作为自变量；单击展开 Method 后的下拉列表，选中 Backward 选项，表示采用向后逐步回归法。

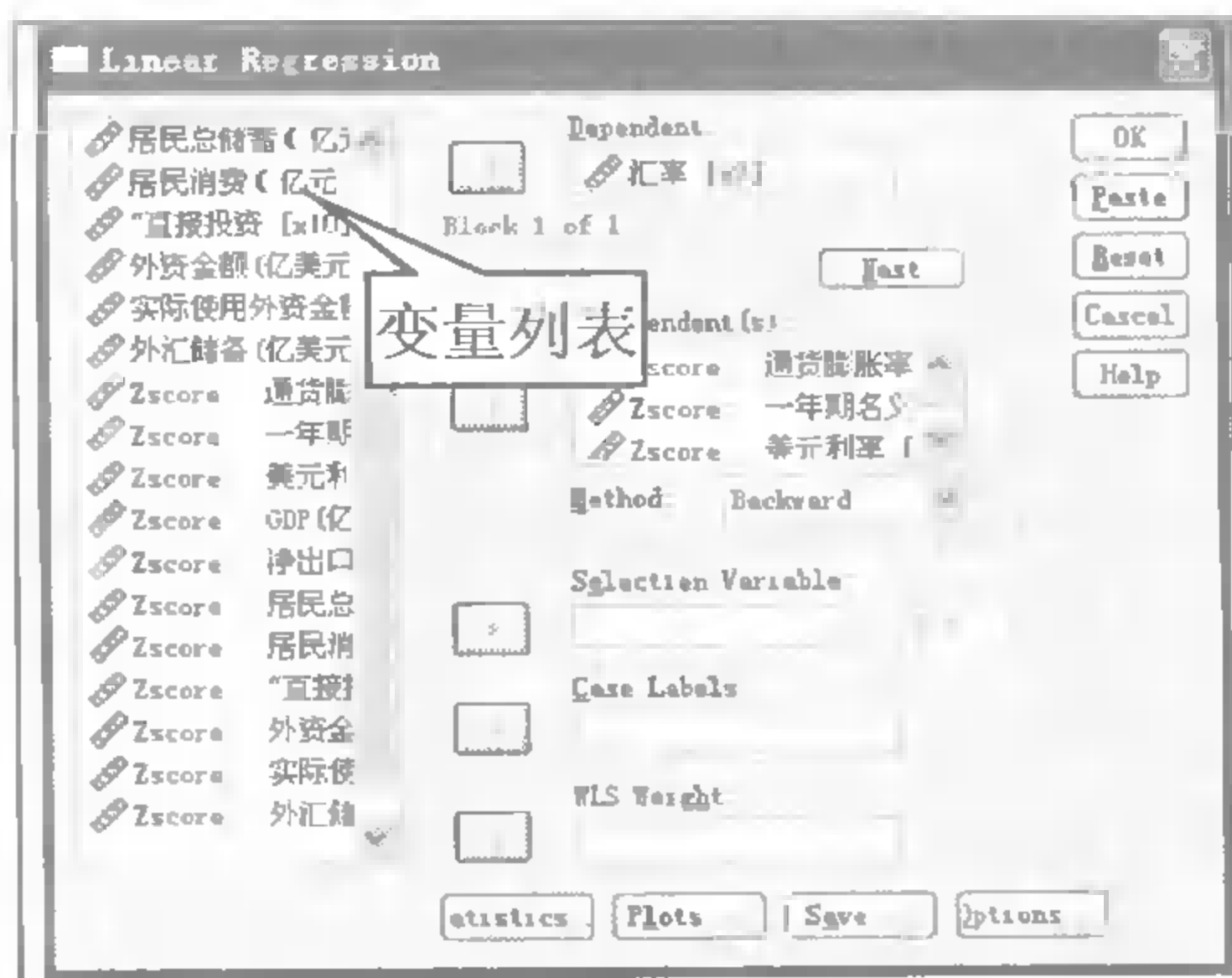




图 21-5 多元回归分析的主设置面板



在图 21-5 中单击 Statistics 按钮, 弹出如图 21-6 所示的统计量设置子面板, 依次勾选如下复选框: Estimates、Model fit、Collinearity diagnostics 和 Durbin-Watson; 单击 Continue 按钮返回主面板。

在图 21-5 中单击 Plots 按钮, 弹出如图 21-7 所示的作图设置子面板, 在变量列表中选中 “\*ADJPRED”, 单击从上至下第一个  按钮将其选入 Y 轴选框; 在变量列表中选中 “\*DEPENDNT”, 单击从上至下第二个  按钮将其选入 X 轴选框; 勾选底部的 Normal probability plots 复选框; 单击 Continue 按钮返回主面板。

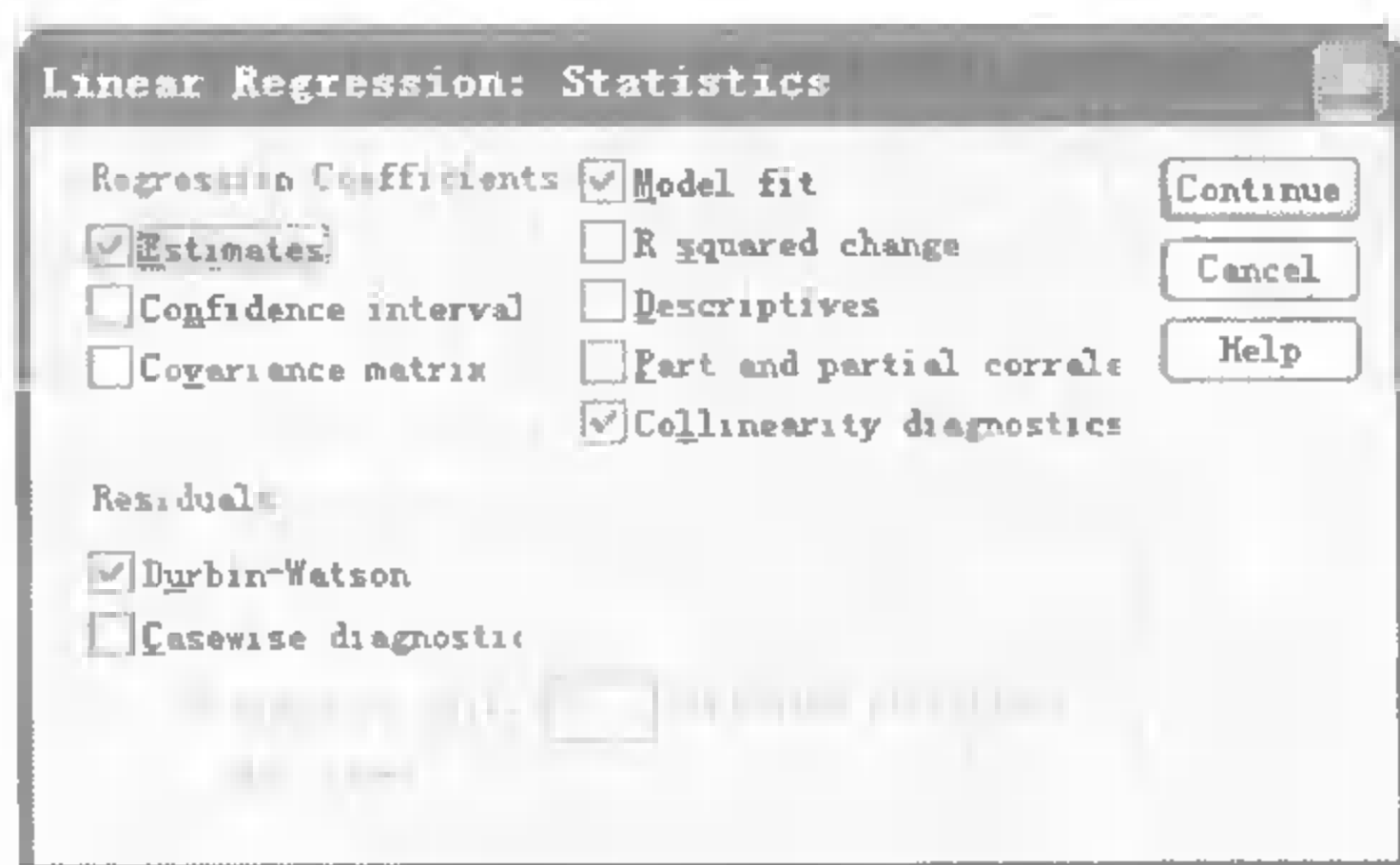


图 21-6 多元回归分析的统计量设置



图 21-7 多元回归分析的作图设置

在图 21-5 中单击 Save 按钮, 弹出如图 21-8 所示的保存设置子面板, 依次勾选如下复选框: Predicted Values 栏中的 Unstandardized; Residuals 栏中的 Unstandardized、Standardized 和 Studentized; 单击 Continue 按钮返回主面板。

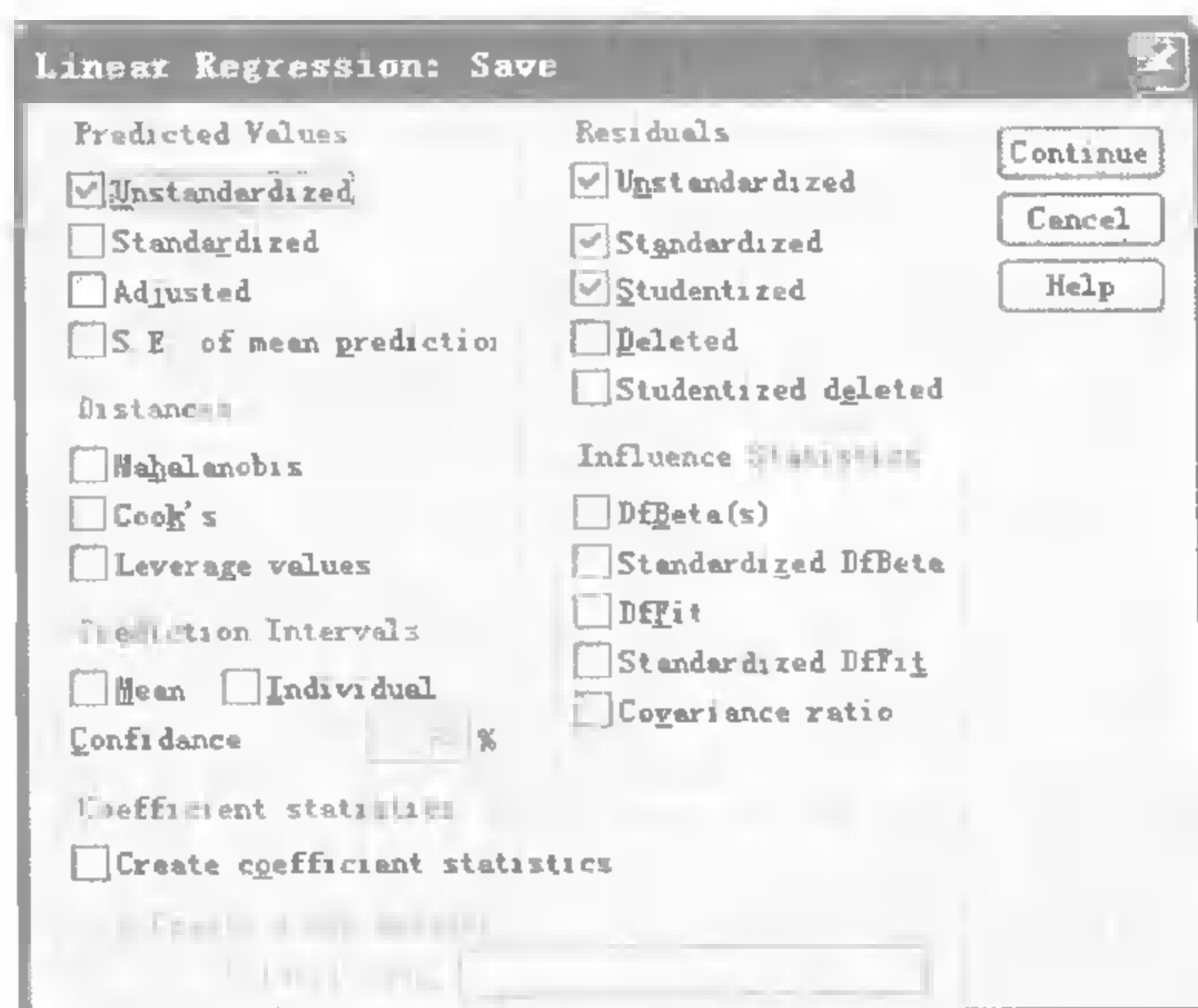


图 21-8 多元回归分析的保存设置

## 2. 结果分析

在图 21-5 中单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 21-9~图 21-12 所示。

(1) 模型摘要信息和方差分析结果。如图 21-9 所示, 模型摘要表给出逐步回归的各模型的拟合情况, 最终模型 (模型 6) 的 R 方和调整 R 方统计量都达到 0.99 以上, 即模型几乎解释了总变异的全部, 说明模型的整体拟合效果不错。Durbin-Watson 统计量距离 2 (与 0、4 相比) 相对较近, 故可初步认为残差序列不存在一阶的自相关性。

模型摘要 <sup>a</sup>					
模型	R	R 方	调整的 R 方	估计的标准差	Durbin-Watson
1	.999 <sup>a</sup>	.999	.995	15320	
2	.999 <sup>b</sup>	.999	.996	13708	
3	.999 <sup>c</sup>	.999	.997	12801	
4	.999 <sup>d</sup>	.998	.997	12510	
5	.999 <sup>e</sup>	.999	.998	13390	
6	.999 <sup>f</sup>	.999	.996	13597	2.943

ANOVA <sup>a</sup>						
模型		平方和	df	均方	F	显著性
1	回归	71.881	11	6.535	278.415	.000 <sup>a</sup>
	残差	.094	4	.023		
	合计	71.975	15			
2	回归	71.881	10	7.188	382.537	.000 <sup>b</sup>
	残差	.094	5	.019		
	合计	71.975	15			
6	回归	71.808	6	11.968	647.355	.000 <sup>f</sup>
	残差	.166	9	.018		
	合计	71.975	15			

图 21-9 模型摘要和方差分析的输出

ANOVA 表给出模型的方差分析结果,从模型 6 的 F 值检验 Sig 值远小于 0.01 看,最终模型的整体线性关系是显著成立的。

(2) 模型的参数估计。如图 21-10 所示,此处只给出了关于最终模型(模型 6)的变量信息。

系数 <sup>a</sup>								
模型		非标准化系数		标准化系数	t	显著性	共线性统计量	
		B	标准误差	Beta			容差	VIF
6	(常量)	6.085	.034		179.011	.000		
	Zscore 一年期名义利率	-.233	.065	-.107	-3.591	.006	.292	3.426
	Zscore 美元利率	-.217	.068	-.099	-3.180	.011	.265	3.779
	Zscore GDP(亿元)	-1.304	.110	-.595	-11.804	.000	.101	9.901
	Zscore 居民总储蓄(亿元)	1.474	.099	.673	14.872	.000	.126	7.966
	Zscore “直接投资	-.438	.108	-.200	-4.055	.003	.106	9.467
	Zscore 外汇储备(亿美元)	2.109	.122	.963	17.315	.000	.083	12.033

a. 因变量: 汇率

已排除的变量 <sup>f</sup>								
模型		Beta In	t	显著性	偏相关	容差	VIF	最小容差
6	Zscore 通货膨胀率	.030 <sup>e</sup>	.689	.510	.237	.144	6.939	.079
	Zscore 居民消费(亿元)	.134 <sup>e</sup>	.971	.360	.325	.014	73.869	.014
	Zscore 外资金额(亿美元) <sup>g</sup>	.027 <sup>e</sup>	.324	.754	.114	.040	24.709	.040
	Zscore 净出口(亿美元)	-.057 <sup>e</sup>	-1.093	.306	-.361	.093	10.728	.044
	Zscore 实际使用外资金额(亿美元)	.033 <sup>e</sup>	1.131	.291	.371	.286	3.495	.060

图 21-10 模型的变量信息

“系数”表包含的是进入模型的变量,主要描述模型的参数估计值(未标准化的和标准化的),以及每个变量的系数估计值的显著性检验和共线性检验。结果模型中所有变量系数的 t 检验 Sig 值都接近或小于 0.01,说明这些系数都显著地不为 0,即这些变量对最终模型的贡献都是显著的。最右一列的共线性统计量,外汇储备的 VIF 值较大(大于 10),故认为它与其他变量间可能存在共线性问题,我们将在下一节处理和改进这一点。

“已排除的变量”表给出的是所有未进入最终模型的变量检验信息,由 t 检验的 Sig 值都大于 0.1 看,这些变量对模型的贡献都不显著,所以它们都不包含在最终方程里。

(3) 残差的诊断和分析。如图 21-11 所示,残差统计表给出了预测值(Predicted Value)、标准化预测值(Std. Predicted value)、残差(Residual)和标准化残差(Std Residual)等的最小值(Minimum)、最大值(Maximum)、均数(Mean)和标准差(Std Deviation)等统计量。

残差统计量 <sup>a</sup>					
	极小值	极大值	均值	标准差	N
预测值	2.8711	8.5750	6.0850	2.18797	16
标准预测值	-1.469	1.138	.000	1.000	16
预测值的标准误差	.056	.130	.088	.020	16
调整的预测值	2.7934	8.5779	6.0625	2.15329	16
残差	-1.8897	2.1840	.00000	1.0532	16
标准残差	-1.390	1.606	.000	.775	16
学生化残差	-1.745	2.217	.042	1.037	16
已删除的残差	-.29794	4.1589	.02247	2.1342	16
学生化的已删除残差	-.2.023	3.101	.075	1.213	16
Mahal. 距离	1.612	12.760	5.625	2.980	16
Cook 的距离	.001	1.149	.190	.307	16
居中杠杆值	.107	.851	.375	.189	16

a. 因变量: 汇率

图 21-11 残差诊断表

在图 21-12 中, 标准化残差的 P-P 图通过比较样本残差分布与假设的正态分布是否相同来检验残差是否服从正态, 所有残差点都分布在对角的直线附近, 说明残差的正态性假设基本成立。

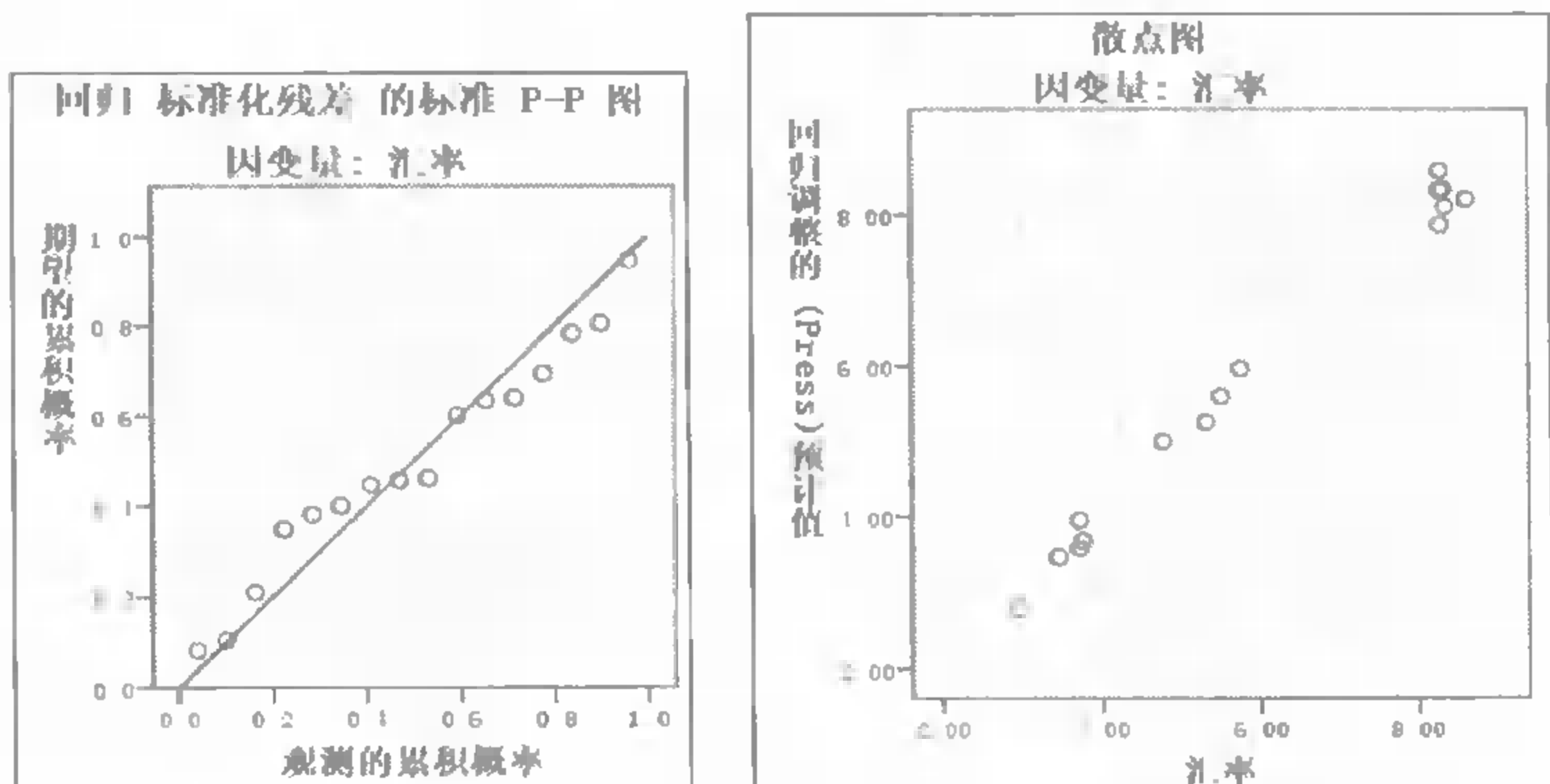


图 21-12 残差分析图和预测检验图

(4) 预测检验图。在图 21-12 中, 汇率的散点图是以汇率观测值为横轴、汇率的调整预测值为纵轴所作的图形, 所有散点都分布在对角线的附近, 说明预测值和观测值非常接近, 预测效果不错。

(5) 预测效果图。分析结束后, 在当前数据集自动生成名为 PRE\_1 的变量, 记录了预测的汇率值, 随后就用观测值和预测值共同作图, 来直观观察预测的效果。

依次单击菜单“Analyze→Time Series→Sequence Charts...”执行时间序列作图过程, 其主界面如图 21-13 所示。在变量列表中选中汇率 (x2) 和预测 1 (PRE\_1) 变量, 单击从上至下第一个 按钮将其选入 Variables 列表作为 Y 轴变量; 在变量列表中选中年份 (x1) 变量, 单击从上至下第二个 按钮将其选入 Time 选框作为 X 轴变量。

在图 21-13 中, 单击 OK 按钮运行, SPSS Viewer 窗口的输出图形如图 21-14 所示。图 21-14 中的线形图以时间为横轴, 在一个图形中描绘汇率观测值 (蓝线) 和汇率预测值 (绿线) 的变化趋势, 图中直观显示出两条线的接近程度很高, 从而判断预测值对观测值的拟合效果很好。

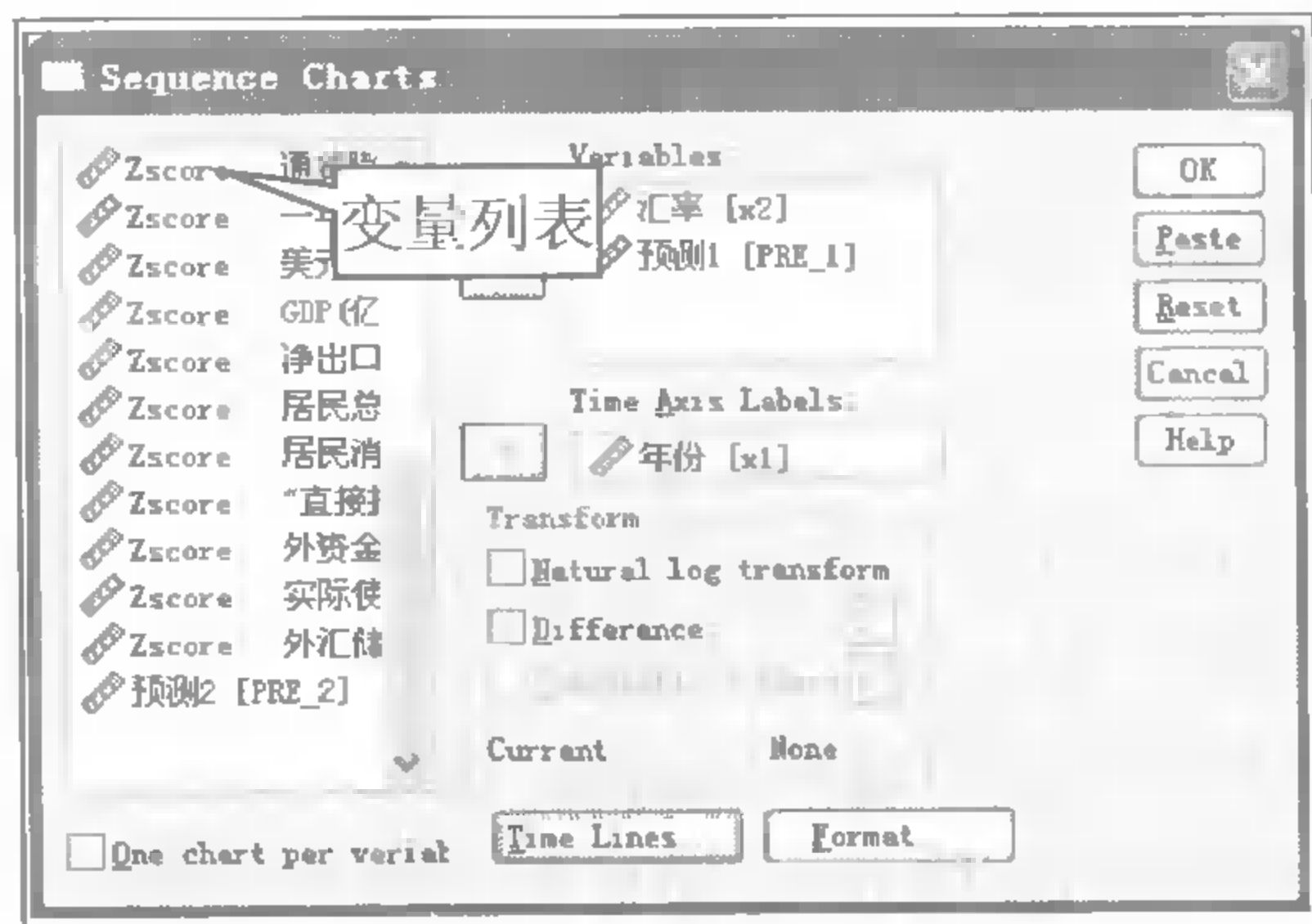


图 21-13 时序作图的参数设置

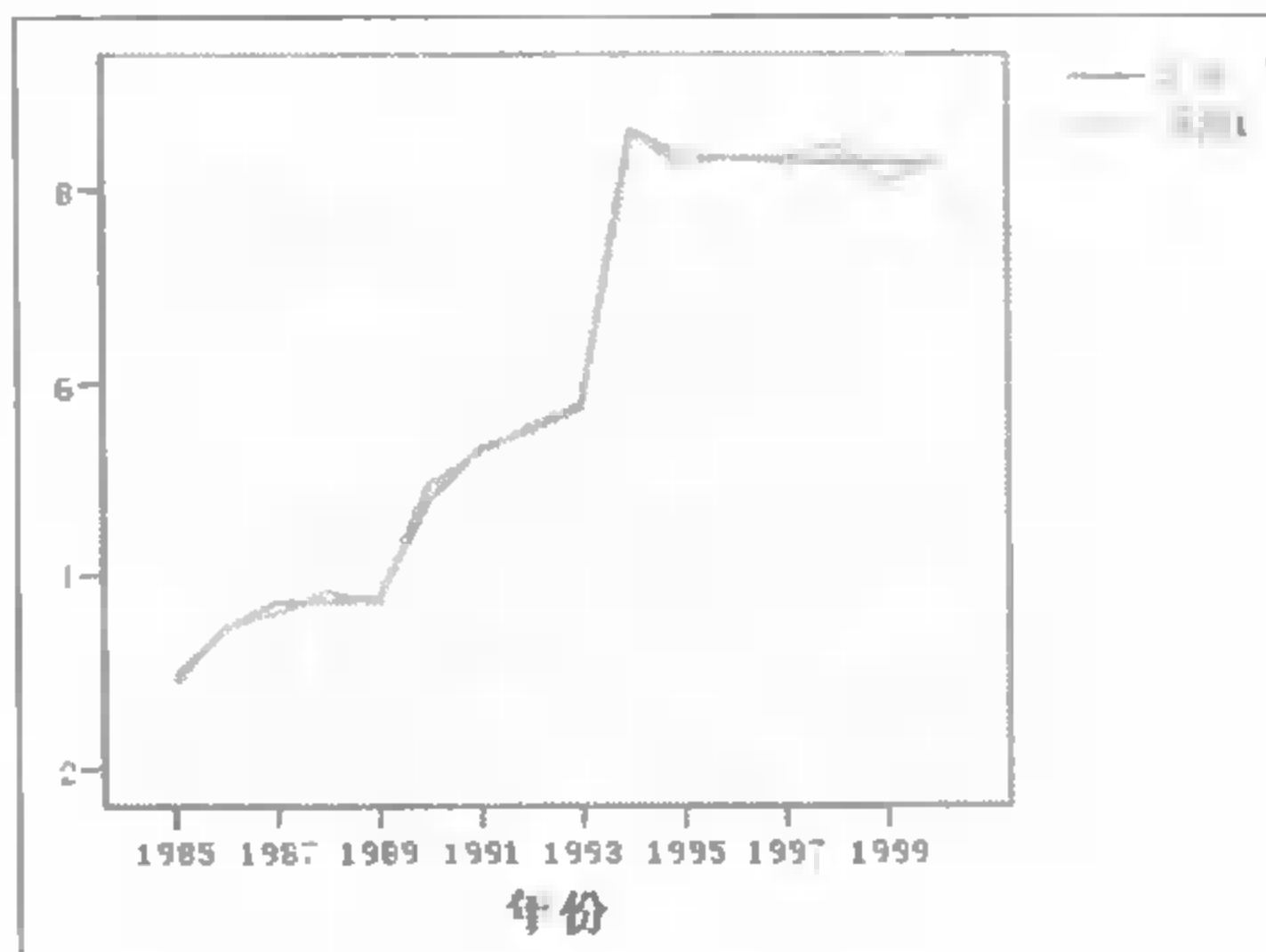


图 21-14 预测效果图

## 21.5 进一步的分析与应用

在前节建立的多元回归模型中, 最终模型中的自变量“外汇储备 (标准化)”共线性诊

断 VIF 统计量较大, 本节先把这个变量剔除后重新建立第二个多元回归模型, 在对结果分析后又提出新的改进建议, 最终得到第三个比较好的结果模型。

### 21.5.1 剔除存在共线性的外汇储备变量

如图 21-5 所示, 在 Independent(s)列表只保留一年期名义利率、美元利率、GDP、居民总储蓄和直接投资这 5 个变量, 其他变量选中后单击其左侧的 ☐ 按钮选出自变量列表; 其他设置不变, 单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 21-15 所示。

模型摘要 <sup>b</sup>					
模型	R	R 方	调整的 R 方	估计的标准差	Durbin-Watson
1	.960 <sup>a</sup>	.921	.881	75561	1.834
a. 预测变量 (常量), Zscore "直接投资, Zscore 一年期名义利率, Zscore 美元利率, Zscore GDP (亿元), Zscore 居民总储蓄 (亿元)。					
b. 因变量 汇率					

ANOVA <sup>b</sup>						
模型		平方和	df	均方	F	显著性
1	回归	66.265	5	13.253	23.212	.000 <sup>a</sup>
	残差	5.709	10	.571		
	合计	71.975	15			
a. 预测变量 (常量), Zscore "直接投资, Zscore 一年期名义利率, Zscore 美元利率, Zscore GDP (亿元), Zscore 居民总储蓄 (亿元)。						
b. 因变量 汇率						

系数 <sup>a</sup>								
模型		非标准化系数		标准化系数	t	显著性	共线性统计量	
		B	标准误差	Beta			容差	VIF
1	(常量)	6.085	.169		32.212	.000		
	Zscore 一年期名义利率	-.389	.358	-.178	-1.068	.302	.298	3.360
	Zscore 美元利率	-.515	.367	-.235	-1.402	.191	.283	3.540
	Zscore GDP (亿元)	-.037	.460	-.017	-.081	.937	.180	5.559
	Zscore 居民总储蓄 (亿元)	2.095	.513	.956	4.061	.002	.144	6.923
	Zscore "直接投资"	-.382	.600	-.174	-.637	.533	.106	9.459
a. 因变量 汇率								

图 21-15 第二个多元回归模型

R 方大于 90%, 调整 R 方也大于 85%, 模型线性关系成立的 F 检验显著性值 Sig 也远小于 0.01, 说明模型整体的拟合优度不错。系数表里的 VIF 统计量也都小于 10, 说明共线性也有所消除。

但是多个变量系数的 t 检验不再显著, 说明模型还有改进的余地。

### 21.5.2 回归模型的进一步改进

第 20.4 节的最终模型包含如下自变量: 一年期名义利率、美元利率、GDP、居民总储蓄、直接投资和外汇储备。其中直接投资的 VIF 统计量较大 (9.5), 且它和 GDP 的关系密切, 考虑到 GDP 包含的内容更全面, 所以建议从模型中剔除直接投资再次建模。剔除了直接投资的模型结果显示 (具体输出略), 只有美元利率的 t 检验 Sig 值太大 (0.9), 且外汇储备的 VIF 统计量仍很大 (12), 考虑到美元利率和外汇储备之间可能存在的相关关系, 下一步建议把 t 检验不显著的美元利率再次剔除。

如图 21-5 所示, 在 Independent(s)列表只保留 1 年期名义利率、GDP、居民总储蓄和外汇储备这 4 个自变量, 其他变量选中后单击其左侧的 ☐ 按钮选出自变量列表; 其他设置不变, 单击 OK 按钮运行, SPSS Viewer 窗口的输出结果如图 21-16 所示。



模型摘要 <sup>b</sup>					
模型	R	R 方	调整的 R 方	估计的标准差	Durbin-Watson
1	.997 <sup>a</sup>	.993	.991	20692	2.556

a. 预测变量 (常量), Zscore\_ 外汇储备(亿美元), Zscore\_ 一年期名义利率, Zscore\_ 居民总储蓄 (亿元), Zscore\_ GDP(亿元).

b. 因变量: 汇率

ANOVA <sup>b</sup>						
模型		平方和	df	均方	F	显著性
1	回归	71.504	4	17.876	417.496	.000 <sup>a</sup>
	残差	.471	11	.043		
	合计	71.975	15			

a. 预测变量 (常量), Zscore\_ 外汇储备(亿美元), Zscore\_ 一年期名义利率, Zscore\_ 居民总储蓄 (亿元), Zscore\_ GDP(亿元).

b. 因变量: 汇率

系数 <sup>a</sup>								
模型		非标准化系数		标准化系数	t	显著性	共线性统计量	
		B	标准误	Beta			容差	VIF
1	(常量)	6.085	.052		117.629	.000		
	Zscore_ 一年期名义利率	-.357	.087	-.163	-4.092	.002	.375	2.668
	Zscore_ GDP(亿元)	-1.329	.180	-.607	-8.323	.000	.112	8.931
	Zscore_ 居民总储蓄 (亿元)	1.244	.123	.568	10.091	.000	.188	5.326
	Zscore_ 外汇储备(亿美元)	2.103	.168	.960	12.505	.000	.101	9.906

a. 因变量: 汇率

图 21-16 再次改进的多元回归模型

此模型的拟合优度较图 21-15 所示的模型要好许多, R 方和调整 R 方都大于 99%, 且模型线性关系成立的 F 检验显著性值 Sig 也远小于 0.01; 而且这些自变量的 VIF 共线性检验统计量也没有特别大的异常值了: 同时所有变量系数的 t 检验都十分显著 (Sig<0.01), 即这些自变量对模型的贡献都是显著的。

从图 21-16 中的系数表可得改进的汇率预测模型为  $\text{汇率} = 6.085 - 0.357 * \text{一年期名义利率} - 1.329 * \text{GDP} + 1.244 * \text{居民总储蓄} + 2.103 * \text{外汇储备}$ , 这里的自变量都为标准化后的。此方程给出了各因素对汇率的量化影响, 其中标准化后的 GDP 和外汇储备对汇率影响最大, 且 GDP 为反向作用, 外汇储备为正向作用。

采用此改进回归模型建立的残差 P-P 图和预测趋势图如图 21-17 所示。由此可见, 残差的分布基本为正态的; 预测曲线对观测曲线的拟合效果也很不错。

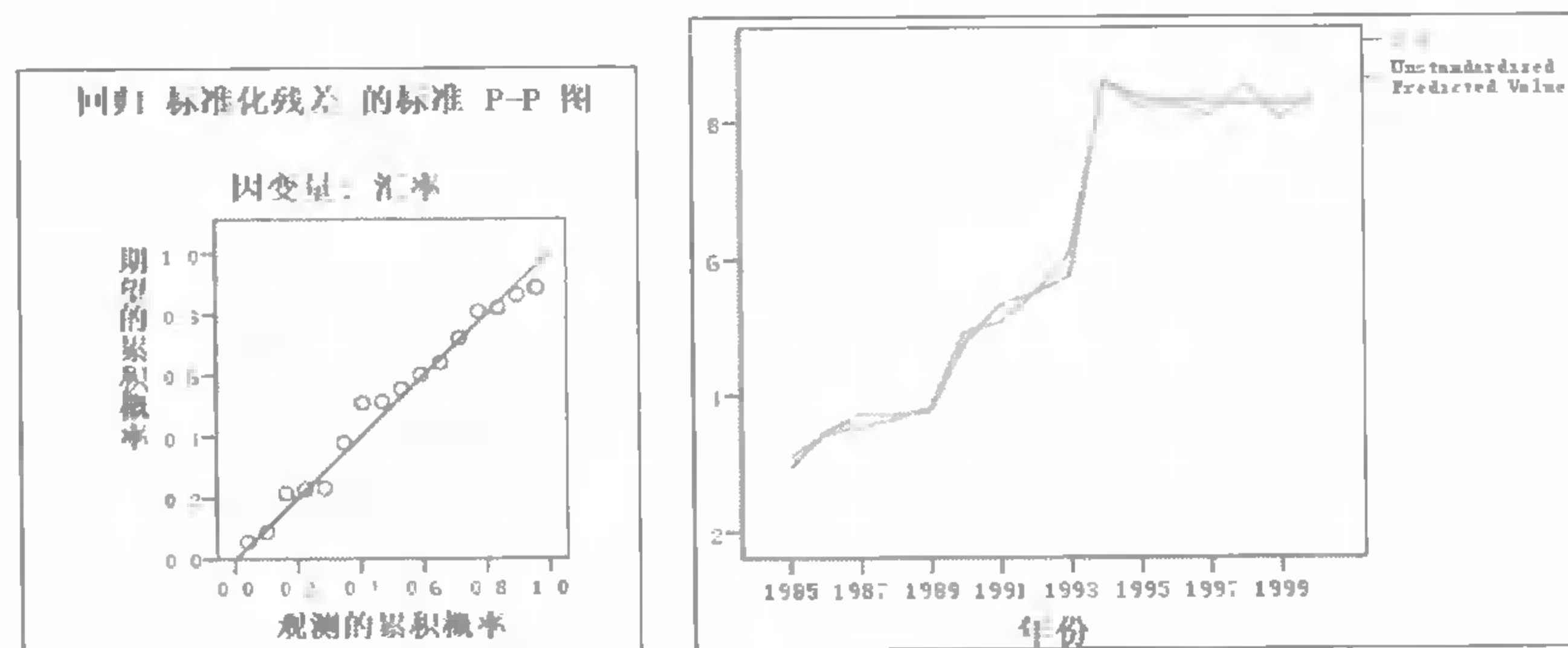


图 21-17 残差图和预测曲线图

### 21.5.3 两个回归模型的比较

采用与作图 21-14 时相同的方法, 把汇率的观测值、初始模型的预测值 (预测 1) 和最后改进模型的预测值 (预测 2) 放在一个图形中加以比较, 如图 21-18 所示。可见改进模型

以较少的变量就达到和初始模型不分伯仲的预测效果，所以建议使用改进后的模型进行其他分析和应用。

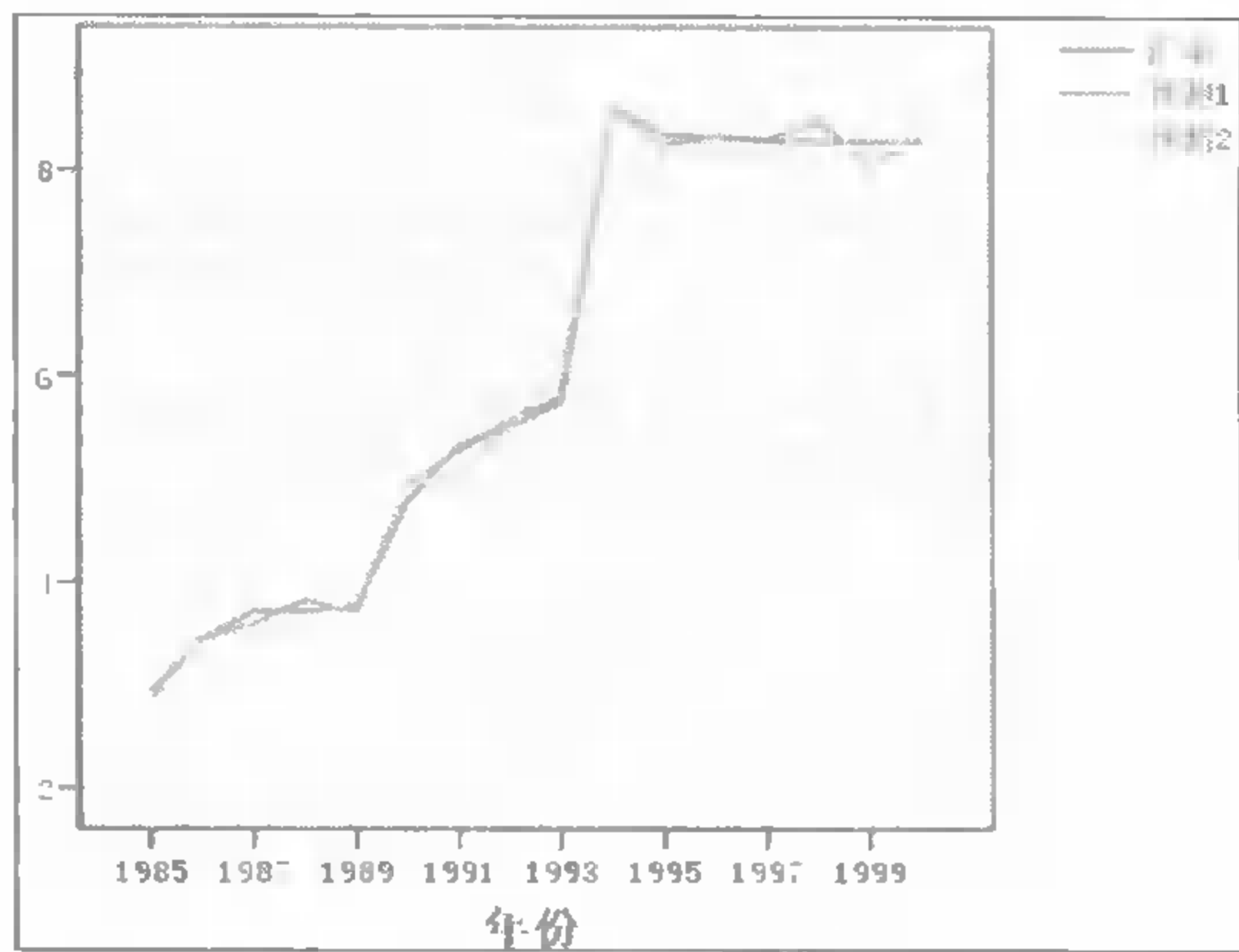


图 21-18 两个预测模型的直接比较

## 21.6 建议和推广

### 21.6.1 时间序列研究

汇率受诸多因素的影响，随着时间在无规则地波动，故许多研究者采用随机序列分析的方法对其进行分析。

“汇率的分析与预测”（高俊琦，《纺织高校基础科学学报》，2000）一文中用随机序列分析法及灰色系统理论分析了汇率的特征，建立出指数平滑模型与灰色模型 GM（1，1），用于预计汇率未来的发展趋势，并在时间层次上给出了汇率的预测值。

“人民币有效汇率走势与影响因素的实证分析”（叶丙南，《西安金融》，2006）一文建立了关于汇率的自回归模型，其结论说明人民币有效汇率一定程度上反映了国内主要经济因素的变化，尤其是货币供应量对有效汇率的影响更为显著，但有些因素变动对有效汇率影响还不稳定。

### 21.6.2 汇率影响因素的定性分析

许多研究者都多次提到影响汇率的因素是很多的，而且汇率作为经济运行中的一个重要元素，与某些经济指标之间的内在关联在经济学上本身就有合理的理论基础和解释，单纯从数据出发进行的建模甚至有可能违背基本的经济理论，所以从经济学的角度对影响汇率的因素进行分析也是很有必要的。

另外也有研究者认为，对汇率影响因素的分析应从浮动汇率入手，而不建议使用固定汇率，这样更能反映市场自身发展的规律，减少了人为因素。

# 第 22 章 因子分析在成绩综合评价中的应用

成绩可以是多方面的,包括在校大学生的考试成绩、高考生的入学成绩、公务员考试的笔试(面试)成绩、公司员工或政府官员的测评考核成绩等,本章就以学生的考试成绩为例,利用因子分析进行对考核对象的综合评价。

学生成绩能反映学生掌握知识和各种能力的程度,综合得分是评价一个学生学习好坏、评定奖学金和评先评优等工作中最最重要的一个指标,也是择优推荐就业很主要的参考因素。因此,合理的、公平的、科学的对学生成绩做出综合评价显得格外重要。目前在高校中,对大学生成绩综合评价的方法就有很多,如原始分数求和法、平均学分成绩法、主成分分析法和因子分析法等,每种方法都有其优缺点。本章介绍的就是采用因子分析进行大学生在校成绩综合评价的模型及其应用。

## 22.1 学生成绩的综合评价简介

学生成绩评定的含义通常是指学校根据一定的标准(即以教学大纲、教材的教学要求为标准),对教学过程中学生所产生(或者即将产生)的思想、学业、行动和个性等方面的变化(或者变化的发展趋势)做出恰如其份的评断。

学校教育的宗旨是为国家培养高素质的人才,学生成绩的评定要有一定的质量要求,需要定期考核学生在德、智、体等方面的发展与进步。它是教育领域必须解决的一个问题,一直受到社会学家、心理学家和教育研究者的关注,是当今世界教育十大变革内容之一,影响着儿童、年轻人、特别是在校学生的成长。良好的学生成绩评定系统可以促进同学们在生理、心理、文化等方面的进步,提高学习效率,可以培养他们积极的学习心态,树立正确的世界观、人生观、价值观和道德观。评价结果既能反馈教师的教学效果,起到诊断、调节和强化的作用;又能反馈学生的学业进展,起到激发学习积极性,增强自信心,萌发学习成功的感受等效应,促使学生整体素质的提高。反之,如果学生成绩评定系统落后或不全面,必将影响学生的学习生活,影响身心健康全面地发展。

在大中专院校中,经常遇到评定奖学金、择优分配和推荐研究生等问题,解决这些问题的关键是如何对学生在校期间的表现给予科学合理地评价,而评价的基础是学生在校期间通过多门课程的学习所获得的多方面的知识和能力。在现行的教学体制中,这些知识和能力具体表现在对课程的掌握上,即考试成绩。通过对学生在校期间各门课的成绩进行因子分析,可以找出评价学生获得知识和能力的主要因素,并据此对学生成绩的综合评价提供较合理的方法。

## 22.2 数据描述

本章来对某药科大学同一年级部分同学的综合成绩数据进行分析，原始数据文件为“某药科大学综合成绩表 35 科.xls”，数据来源于网络搜索。

学科成绩的原始数据记录了 102 名同学在校期间的 35 门课程的成绩，所有成绩都为百分制，且最小记分单位为 1。其中对于由于违纪、缺考等原因导致的课程无成绩情况，记为 0 分；对于有补考的情况，记最后一次补考的成绩。数据格式概览如表 22-1 所示。

本章的目的就是通过因子分析方法建立一个对学生的综合评价模型，一方面对它们的学习成绩进行打分排序；另一方面找出影响他们学习成绩的不同因素，并在此基础上提出改进教学的意见，提高学生的综合成绩和素质。

表 22-1 高校学生成绩概览

学号\课程	无机	哲学	思品	高数 I	体育 I	大学英语 I	有机 I	.....
05054001	93	85	95	75	100	90	94	.....
05054002	76	96	97	73	80	70	88	.....
05054004	60	84	90	61	80	73	78	.....
.....	.....	.....	.....	.....	.....	.....	.....	.....

## 22.3 分析方法概述

对学生成绩的综合评价，不同的研究者发展了多种数学模型，有平均学分绩法、主成分分析模型、专家调查法-主成分分析模型、因子分析模型等，“学生成绩排名的综合评价模型”（吴海英，《大学数学》，2006）一文，对这几种方法进行了比较，并指出因子的优点是体现出了每个学生在各个因子上的能力，由此能够对每个学生各方面的能力进行排名，然后对每个学生在各方面的综合能力进行排名，各因子的权重就是它的特征值所占的比率，这种方法的缺点是可能把学分并不高的课程看成了很重要的课程，这样就与综合评价的初衷有所违背。

在多指标综合评价方法中，传统方法对于不同因素权重的设置往往带有一定的主观随意性，将多元统计引入综合评价方法，如因子分析法可以克服人为确定权数的缺陷，使得综合评价结果更加客观合理。因子分析是多元统计的重要分析方法之一，其基本思想是根据相关性大小对变量进行分组，使得同组内的变量之间相关性较高，不同组的变量之间相关性较低，每组变量代表了一个基本结构，因子分析中将之称为公共因子。因子分析在教育学、社会学、心理学等领域都有广泛的应用价值。

### 22.3.1 应用因子分析进行成绩综合评价的步骤

假定因子分析模型为  $X = AF + \varepsilon$ ，其中  $X$  为  $p \times 1$  维向量； $A$  为  $p \times q$  维矩阵，称为因子载荷矩阵； $F$  为  $q \times 1$  维向量（ $q < p$ ），称为公共因子； $\varepsilon$  为  $p \times 1$  维向量，称为特殊因子。模型满足假设条件  $F: N(0, I_q)$ ， $\varepsilon: N(0, \Psi_{p \times q})$ ，其中  $\Psi$  为对角阵， $F$  与  $\varepsilon$  相互独立。



因子分析就是从  $X$  的  $n$  个样本数据出发, 来确定因子载荷矩阵  $A$ , 再根据因子载荷矩阵来确定公因子个数, 并结合各因子所包含的变量确定潜在变量的含义。

下面简单介绍主成分因子分析的一般步骤。

(1) 设有  $n$  个样本, 每个样本有  $p$  个变量, 原始数据记为  $X = (x_{ij})_{n \times p} = (X_1, X_2, \dots, X_p)$ , 先对样本  $X$  的列进行标准化处理。

(2) 计算相关系数矩阵  $R = (r_{ij})_{n \times n} = \frac{1}{n} X'X$ 。

(3) 计算  $R$  的特征值和特征向量, 并确定因子个数。

(4) 求因子载荷矩阵, 并进行因子旋转, 使各门课程在公共因子上的作用更加明显, 并易于解释。

(5) 计算每个学生的成绩在所有公共因子上的得分, 得到因子得分矩阵。

(6) 以各因子的贡献率为权重, 求和计算因子得分的综合得分, 根据综合得分排序。

### 22.3.2 应用因子分法进行成绩综合评价的注意事项

#### 1. 原始指标的转换处理

有时原始变量取值的量纲会有所不同, 这在成绩评价中既可以指评分标准的不同, 如百分制、5 分制等, 也可以指不同课程之间的差异, 如较难的课程得分低, 较易的课程得分高。此时, 若用原始指标直接求综合得分, 将很难给予比较合理的解释, 当原始指标变量数量级差异较明显时, 变量取值大的对综合指标公共因子的影响也大。因此在运用因子分析法时, 通常需要对原始指标进行无量纲化处理。

#### 2. 什么样的评价指标适合因子分析

因子分析方法在多元统计中属于降维思想中的一种, 其目的在于简化数据, 通过较少的公共因子反映复杂现象的基本结构。假如原始评价指标较少且意义明确, 已经能较好地反映评价对象, 这时就不建议使用因子分析了, 如果一定要运用, 反而会加大计算量, 意义也不大。

此外, 使用因子分析进行综合评价目的之一, 是为了避免评价指标之间的相关性所引起的权重偏倚。因此使用因子分析的一个前提条件是评价指标之间应有较强的相关关系, 如果指标之间的相关程度很小, 指标就不可能共享公共因子, 公共因子对于指标的综合解释能力也就会降低。一般来说, 可以先对指标的相关矩阵进行检验, 如果相关矩阵的大部分系数都显著低小于 0.3, 则不适合做因子分析。

#### 3. 结果选取多少个因子进行分析

因子分析的目的是寻求用较少的公共因子, 来解释协方差结构。选取的因子过多, 应用因子分析方法就失去原有的意义, 但选取的因子过少, 又可能造成原始信息量的大量损失。通常选择因子个数, 有以下 3 种准则。

(1) 以主成分的特征值为标准。原始评价指标标准化后, 由于每个指标的方差为 1, 假如主成分所对应的特征值小于 1, 意味着该主成分连一个指标的方差都无法解释, 所以应选取特征值大于或接近于 1 的主成分作为公共因子, 舍弃特征值远小于 1 的其它主

成分。

(2) 以主成分的方差累计贡献率为标准。方差累积贡献率反映了主成分保留原始信息量的多少。一般而言,主成分累积贡献率达到 85% 以上就可以很好地说明和解释原有问题,因此可以以此为标准选取累积贡献率达到 85% 以上的那些主成分作为公共因子。

(3) 根据分析问题的需要或专业理论来选取公共因子。在多维数据中,当维数大于 3 时便不能画出几何图形,但通过因子分析法选取主要的两个公共因子,可以画出正交因子得分图,以此反映评价对象在二维平面上的分布情况,从而直观地找出各评价对象在公共因子中的地位,进而还可以对评价对象进行分类处理。

(4) 是否需要进行因子旋转。建立因子分析模型的目的不仅是要找出主因子,更重要的是要知道每个主因子的意义,以便对实际问题进行分析。初始的公共因子是否具有明确意义,需要进一步分析因子载荷阵才能得出。如果从每个初始因子能较好地找出所代表的原始指标,就可以直接赋予这些因子合理的解释,并进行下一步的分析研究。但是如果因子载荷量较为平均,难以判别哪些指标与哪个因子的联系较为密切,也就是说无法从原始指标中寻求评价对象在各个因子上得分差异的原因,这时就需要进行因子旋转。

因子旋转的直观意义是经过旋转后,公共因子的贡献越分散越好,使每个指标仅在一个公共因子上有较大的载荷,而在其余公共因子上的载荷比较小。因子旋转的方法很多,如正交旋转、斜交旋转等,正交旋转又包括方差最大化旋转、四次方最大化旋转等。

## 22.4 SPSS 建模过程和结论分析

本节首先对搜集到的数据进行格式转换,并且观察各科成绩的基本统计信息,然后建立关于学生成绩综合评价的因子分析模型。

### 22.4.1 数据准备

分析数据为某药科大学同年级部分学生在某年份的 35 科得分成绩,原始数据为 Microsoft Excel 格式,下面把它导入到 SPSS 中。

#### 1. 数据导入

依次单击菜单“File→Open→Data...”执行文件打开操作,其界面如图 22-1 所示,在文件类型下拉列表选中“Excel (\*.xls)”项;然后在上面的文件列表选中“某药科大学综合成绩表 35 科.xls”;单击打开按钮,弹出如图 22-2 所示的文件打开确认窗口。在图 22-2 中,保留默认设置,单击 OK 按钮,即可新建一个包含导入数据的 SPSS 格式数据文件,将此文件另存为“大学综合成绩表 35 科.sav”。

这只是数据导入的快捷方法,另外依次单击菜单“File→Open Database→New Query...”,通过建立对 Excel 数据库的连接和搜索,能够实现更复杂的数据导入操作。

打开“大学综合成绩表 35 科.sav”文件,在变量视图中把变量标签都设置为与对应的变量名相同的内容,然后把变量名更改为类似“x1”的格式,这些变量的其它设置都采用默认选项。更改后变量名与变量标签的对应关系如表 22-2 所示。

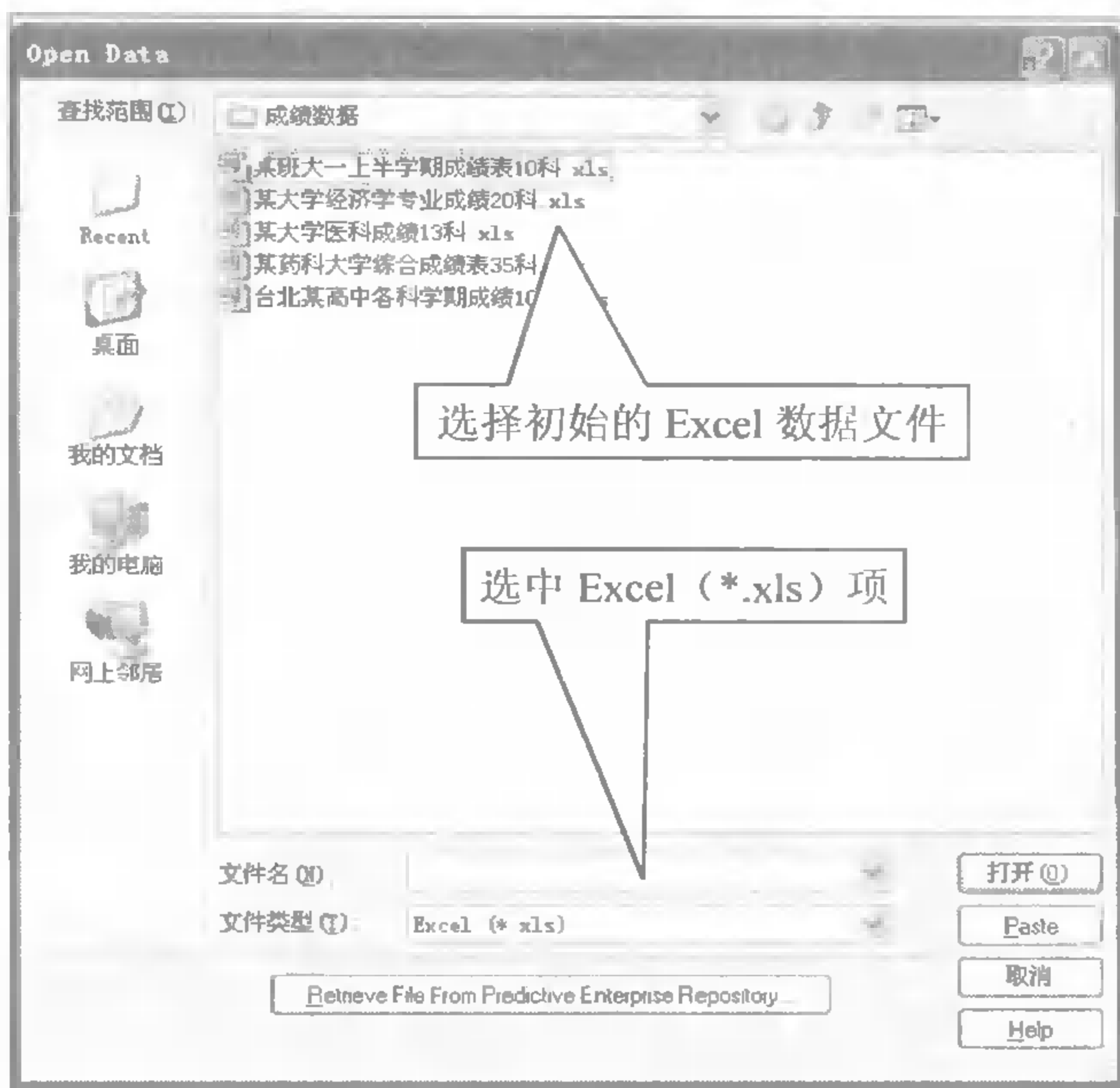


图 22-1 文件导入界面

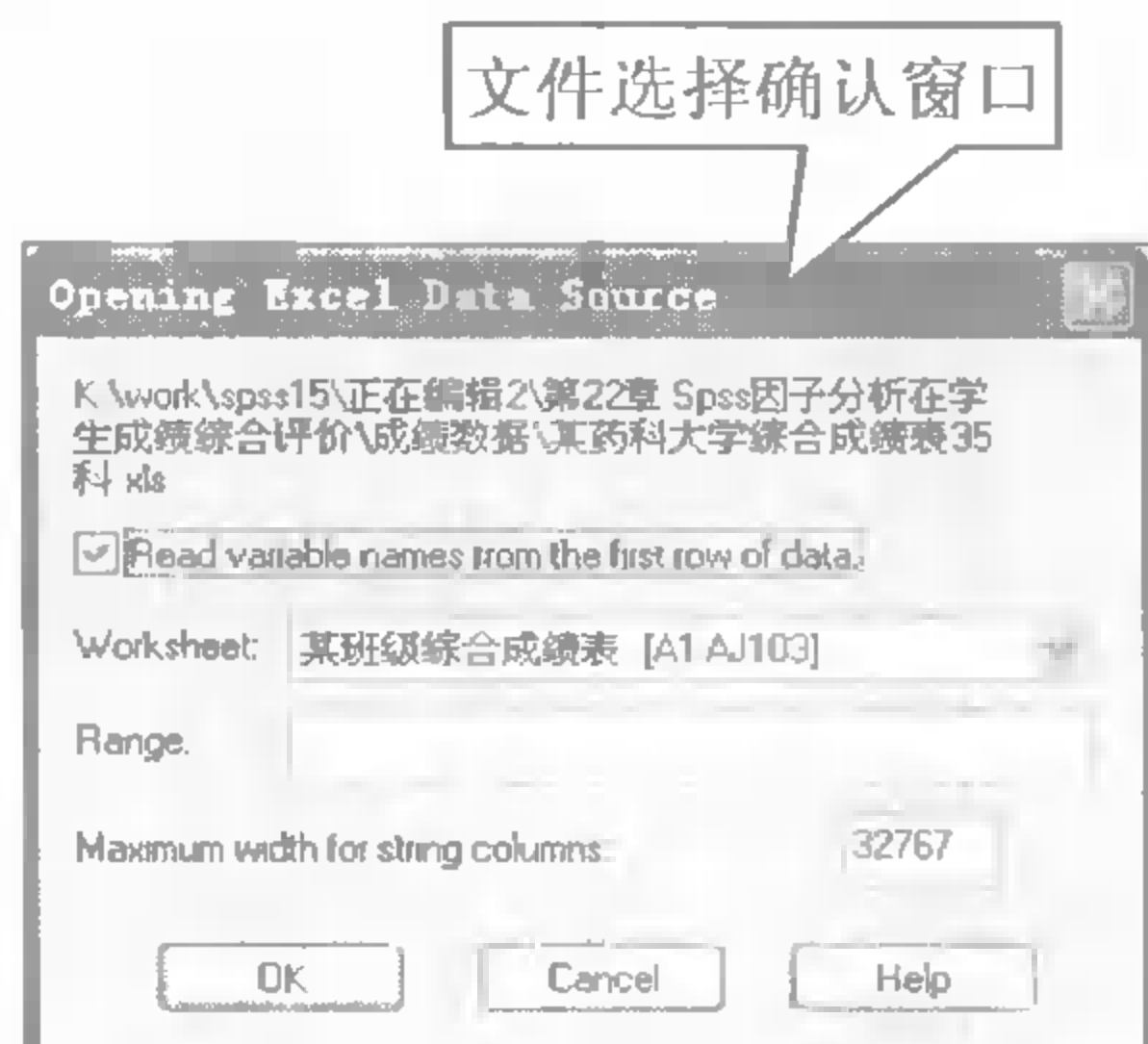


图 22-2 文件导入操作

表 22-2

变量名与变量标签的对应关系表格

变量名	变量标签	变量名	变量标签	变量名	变量标签	变量名	变量标签
x1	学号	x10	高数 II	x19	计算机基础与应用	x28	生物化学
x2	无机	x11	物理	x20	分化 II	x29	人体解剖生理学
x3	哲学	x12	体育 II	x21	物化 I	x30	新药开发
x4	思品	x13	大学英语 II	x22	体育 IV	x31	药剂学
x5	高数 I	x14	有机 II	x23	大学英语 IV	x32	药分 I
x6	体育 I	x15	分化 I	x24	计算机技术基础	x33	微生物
x7	大学英语 I	x16	数统	x25	世贸	x34	天然药化
x8	有机 I	x17	体育 III	x26	物化 II	x35	专业英语
x9	毛概	x18	大学英语 III	x27	药物化学	x36	药理学

## 2. 变量简要统计信息概览

本节来简单观察一下各成绩变量的基本统计信息，打开文件“大学综合成绩表 35 科.sav”。

依次单击菜单“Analyze→Descriptive Statistics→Descriptives...”执行描述性统计分析过程，其主设置界面如图 22-3 所示。在变量列表选中 x2~x36 的所有变量，单击  按钮将其选入 Variable(s) 列表作为分析变量。

在图 22-3 中单击 Options 按钮，弹出如图 22-4 所示的选项设置面板，依次单击选中如下统计量：Mean（均值）、Std（标准差）、Minimum（最小值）、Maximum（最大值）、Kurtosis（峰度）、Skewness（偏度）；单击 Continue 按钮返回主面板。

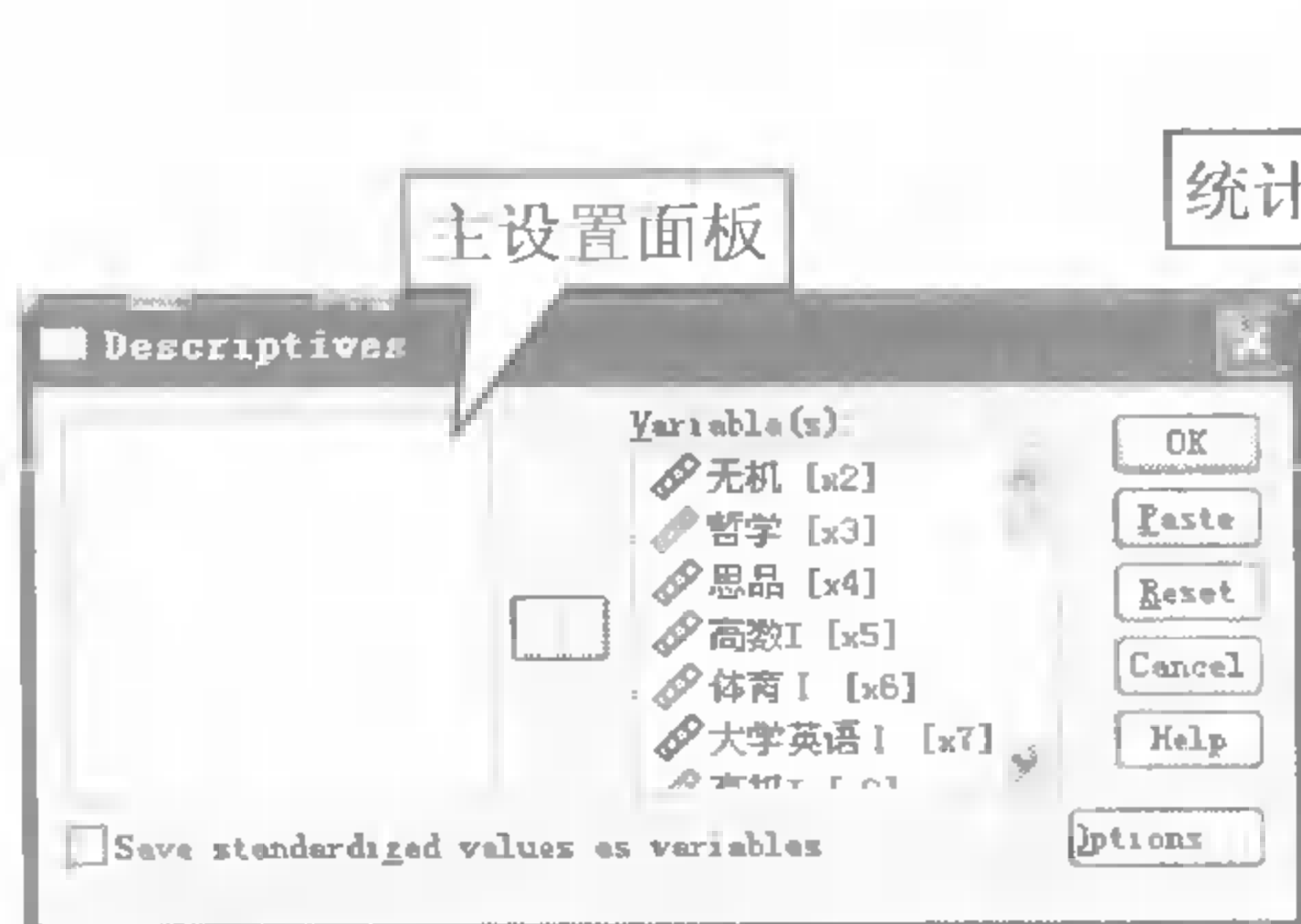


图 22-3 描述性统计分析的主界面

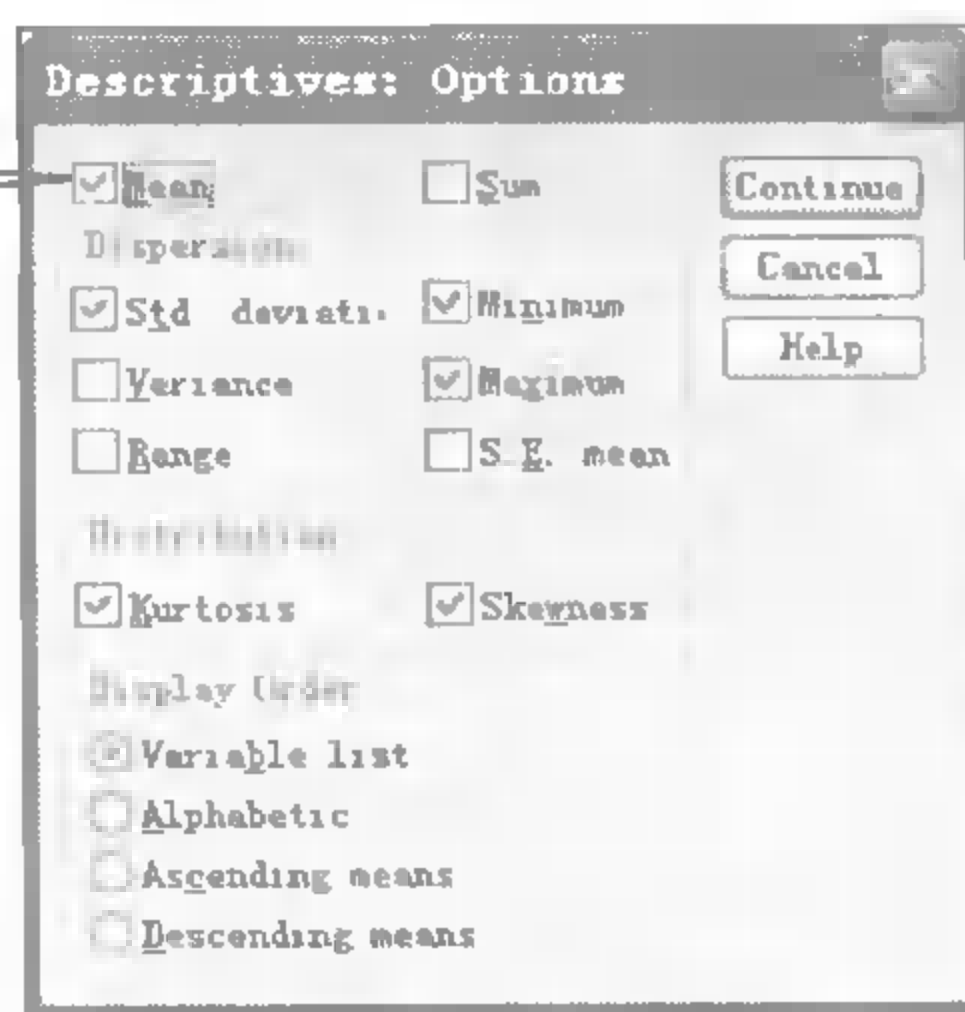


图 22-4 描述性统计分析的统计量选择

在图 22-3 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出表格如图 22-5 所示。

描述统计量									
	N	极小值	极大值	均值	标准差	偏度		峰度	
	统计量	统计量	统计量	统计量	统计量	统计量	标准误	统计量	标准误
无机	102	0	98	62.46	14.462	-1.534	.239	7.760	.474
哲学	102	0	96	74.97	13.771	-1.807	.238	7.659	.474
思品	102	70	98	90.75	6.560	-1.425	.239	1.672	.474
高数I	102	0	96	66.20	10.806	-1.592	.239	13.368	.474
体育I	102	0	100	78.42	24.212	-2.246	.239	5.182	.474
大学英语I	102	0	90	64.92	14.442	-1.709	.239	5.928	.474
有机I	102	1	95	71.24	14.975	-1.111	.239	4.380	.474
毛概	102	60	97	78.91	9.615	-.724	.239	-.416	.474
高数II	102	0	98	64.18	19.502	-1.169	.239	1.078	.474
物理	102	0	98	73.18	15.588	-.701	.239	3.072	.474
体育II	102	0	100	65.20	35.348	-1.092	.239	-.323	.474
大学英语II	102	23	83	61.88	7.417	-1.490	.239	12.554	.474
有机II	102	0	98	69.43	13.117	-.949	.239	6.496	.474
分化I	102	60	88	65.49	7.495	1.062	.239	-.151	.474
数统	102	80	90	86.82	7.624	1.271	.239	1.234	.474
体育III	102	0	100	67.33	31.968	-1.417	.239	.585	.474
大学英语III	102	0	90	70.45	14.176	-2.276	.239	10.192	.474
计算机基础与应用	102	60	98	76.74	10.719	-.203	.239	-1.176	.474
分化II	102	60	96	79.52	13.009	-.423	.239	-1.474	.474
物化I	102	60	97	76.38	13.221	-.122	.239	-1.575	.474
体育IV	102	0	100	63.96	27.784	-1.501	.239	1.262	.474
大学英语IV	102	60	88	67.25	8.194	.890	.239	-.323	.474
计算机技术基础	102	0	97	75.22	14.580	-1.271	.239	5.214	.474
世贸	102	60	98	89.55	7.468	-1.701	.239	8.581	.474
物化II	102	0	92	75.32	12.466	-2.279	.239	11.690	.474
药物化学	102	0	97	68.55	26.476	-1.365	.239	1.040	.474
生物化学	102	0	97	70.37	20.274	-1.511	.239	3.783	.474
人体解剖生理学	102	0	96	78.96	19.891	-.525	.239	7.612	.474
新药开发	102	0	99	83.12	17.374	-4.372	.239	18.960	.474
药剂学	102	0	88	65.12	14.953	-.026	.239	6.298	.474
药分I	102	0	95	69.71	16.584	-1.690	.239	4.038	.474
微生物	102	0	88	56.68	25.075	-1.066	.239	.000	.474
天然药化	102	0	95	73.68	23.365	-2.105	.239	3.877	.474
专业英语	102	0	96	86.22	11.354	-4.893	.239	38.232	.474
药理学	102	0	84	58.01	15.150	-.844	.239	1.110	.474
有效的 N (列表状态)	102								

图 22-5 学生成绩的描述性统计量输出


观察图 22-5 可知，处理后的数据有 102 例是有效的；思品、世贸、新药开发、专业英语四门课程的平均分较高；而药理学、微生物两门课程的平均分不满 60，值得引起教师的关注；体育、药物化学、微生物、天然药化四门课程的成绩波动（标准差）最为明显；另外对于单科成绩得分的数据分布，从偏度、峰度距离 0 较远看，它们基本都不服从正态分布。

## 22.4.2 SPSS 因子分析建模与分析

### 1. SPSS 参数设置

依次单击菜单“Analyze→Data Reduction→Factor...”执行因子分析过程，其主界面如



图 22-6 所示。在变量列表选中除了学号之外的所有变量，单击从上至下第一个  按钮将其选入 Variable(s) 列表作为分析变量。

在图 22-6 中，单击 Descriptives 按钮弹出如图 22-7 所示的统计量选择面板，依次勾选如下 3 个复选框：Coefficients、Significance levels 和 KMO and Bartlett's Test of Sphericity。单击 Continue 按钮返回主面板。

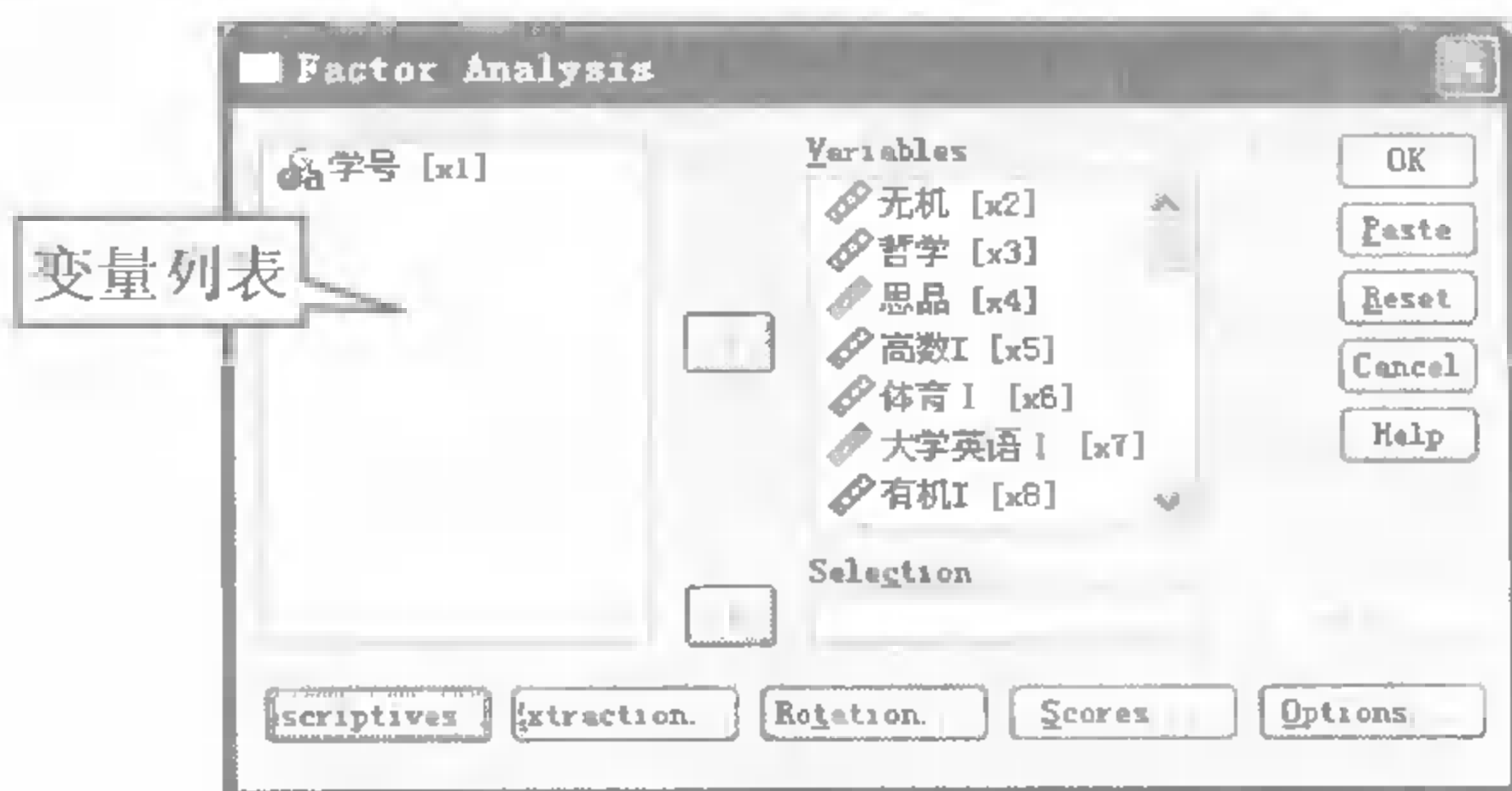


图 22-6 因子分析的主界面

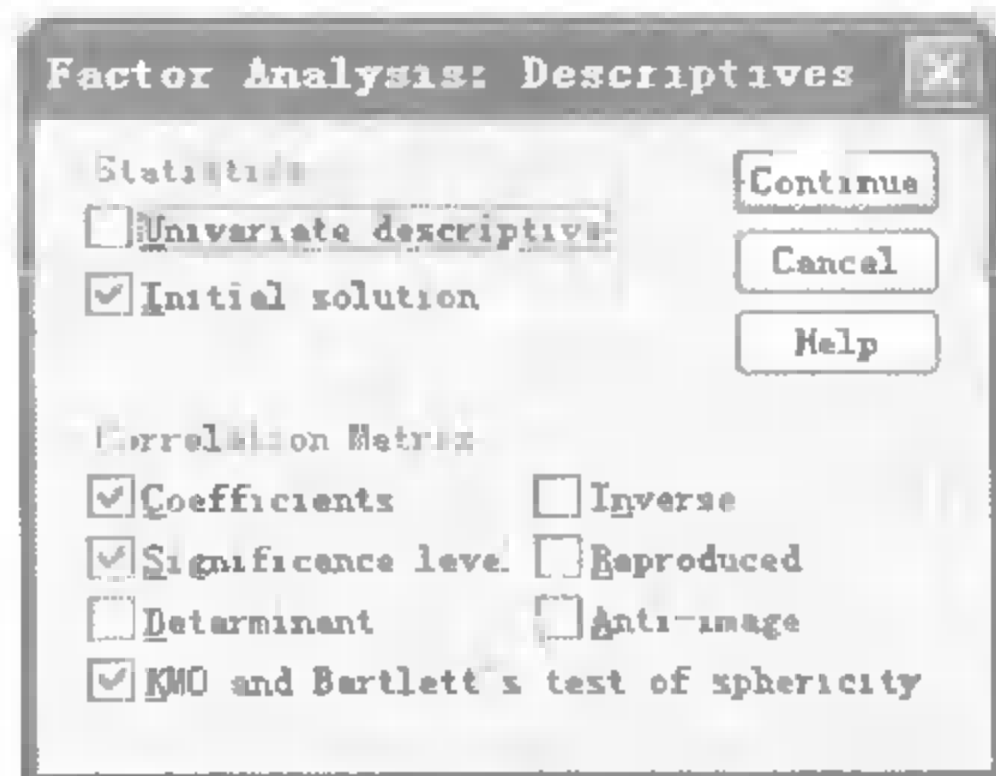


图 22-7 因子分析的统计量设置

在图 22-6 中单击 Extraction 按钮，弹出如图 22-8 所示的设置对话框。勾选 Scree plot 复选框；其他选项默认，即选择 Principal components 主成份法进行因子提取。单击 Continue 按钮返回主面板。

在图 22-6 中单击 Rotation 按钮，弹出如图 22-9 所示的旋转方法设置对话框，单击选中 Varimax 单选框，表示采用方差最大旋转法进行因子旋转。单击 Continue 按钮返回主面板。

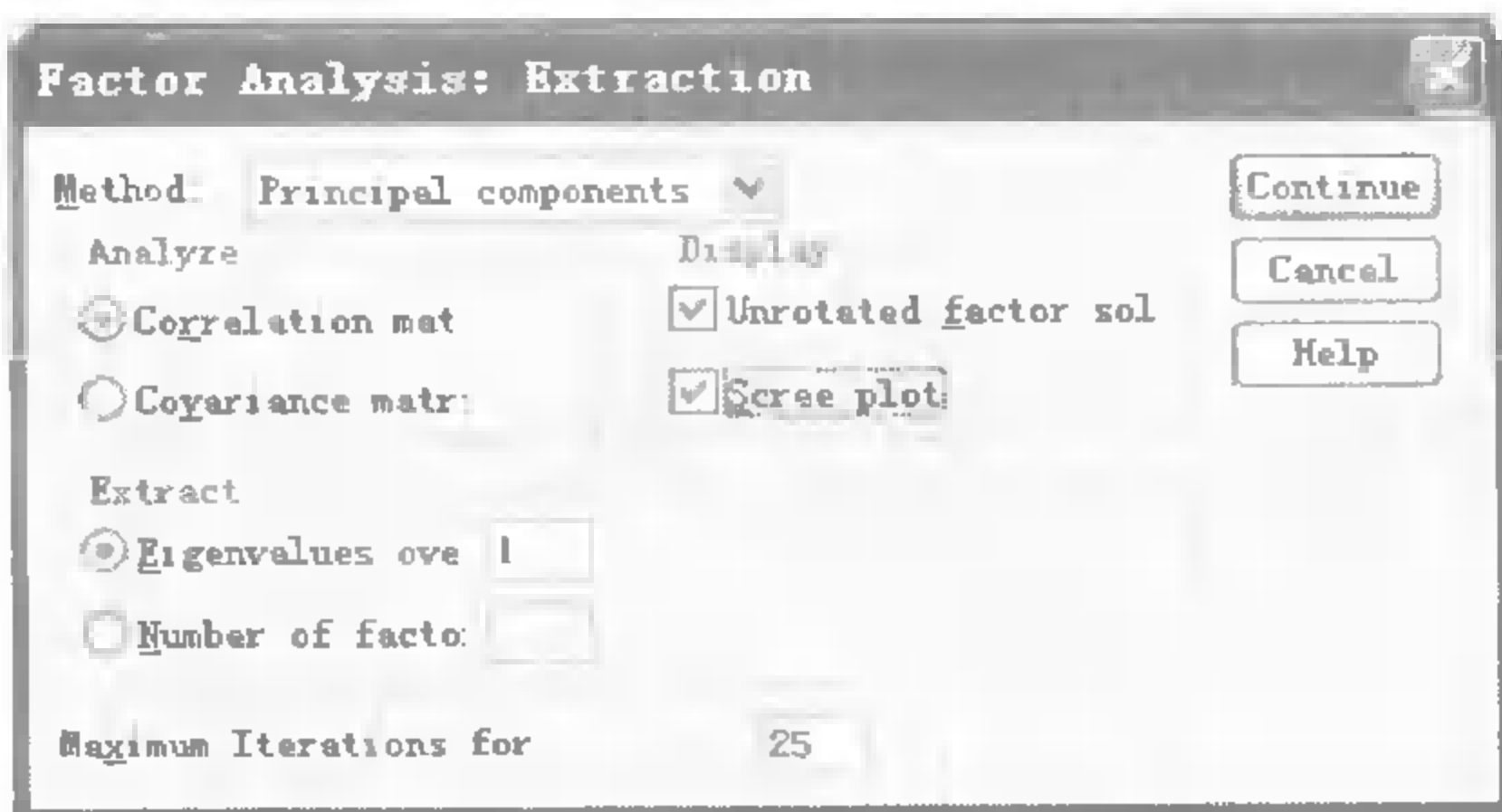


图 22-8 因子分析的提取方法设置

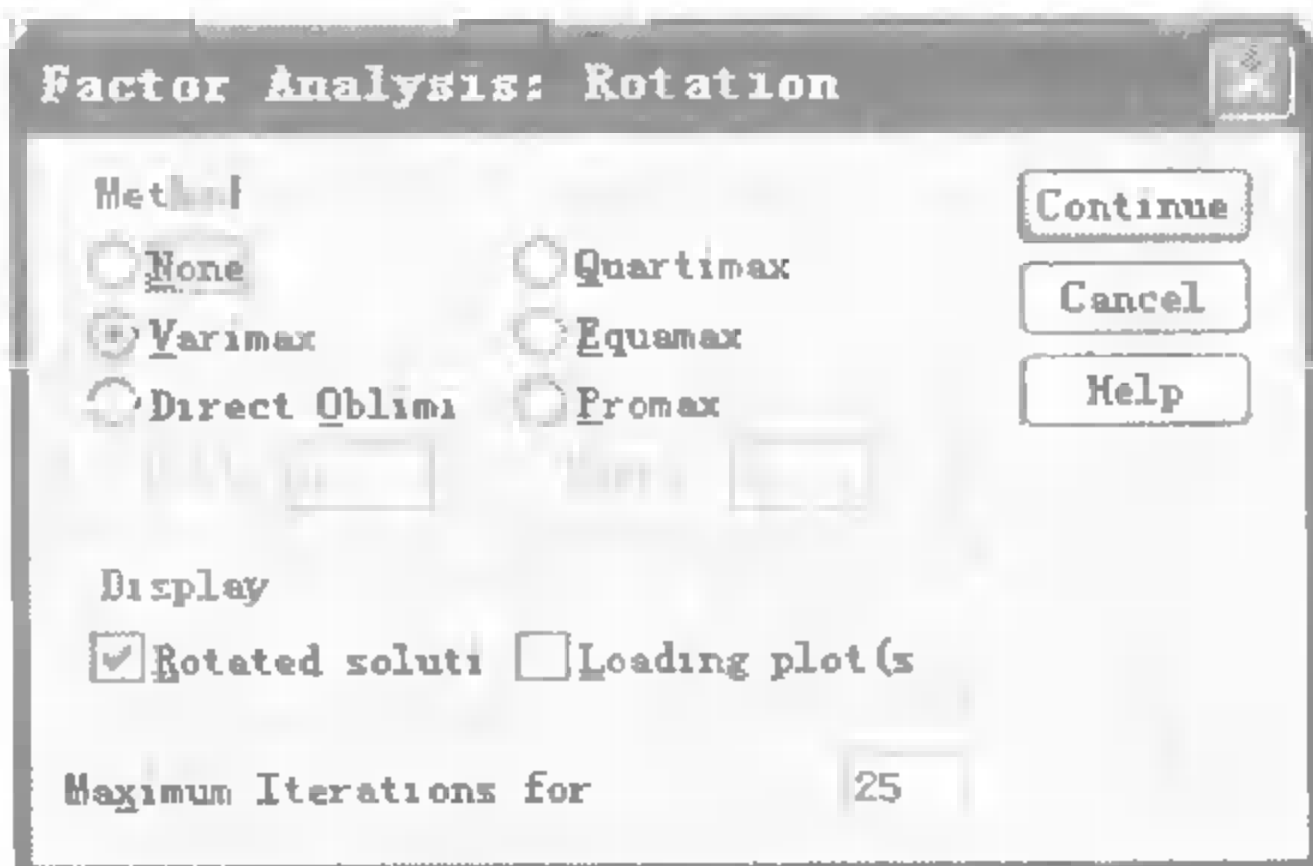


图 22-9 因子分析的旋转方法设置

在图 22-6 中单击 Scores 按钮，弹出如图 22-10 所示的因子得分设置对话框，勾选底部的 Display 复选框。单击 Continue 按钮返回主面板。

## 2. 因子分析的输出解释

在图 22-6 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如下。

(1) 初始变量的相关性检验。首先输出的是关于初始变量的相关系数矩阵，这些变量之间的相关系数在 0.5 左右的有很多，而且其对应的相关性检验 Sig 值大都小于 0.01，这说明这些变量之间存在着较为显著的相关性，进而说明有进行因子分析的必要。

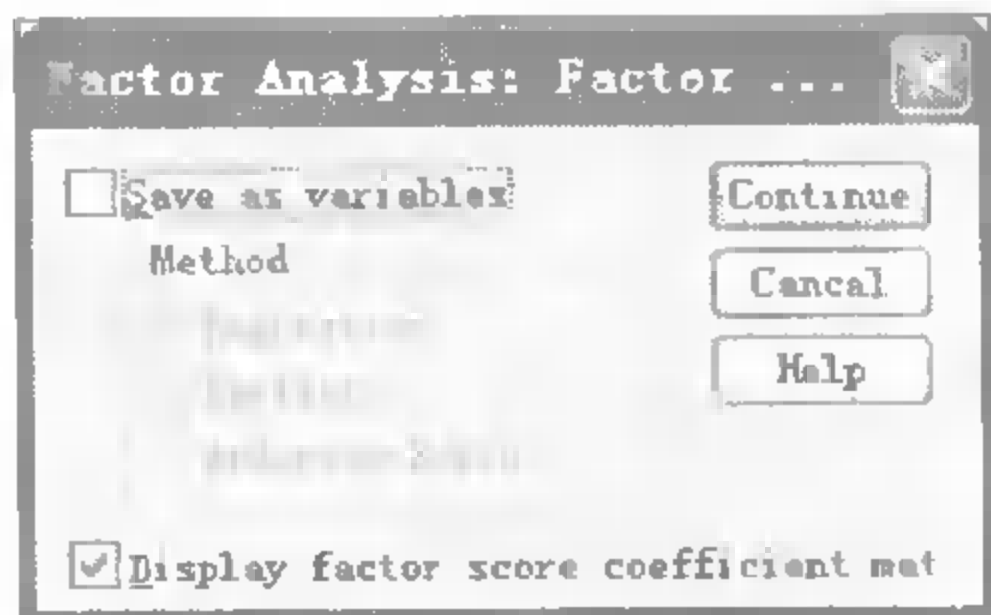


图 22-10 因子分析的因子得分设置

由于科目较多，输出的相关系数矩阵很大（35×35 矩阵），具体输出略。

KMO 和 Bartlett 的检验		
取样足够度的 Kaiser-Meyer-Olkin 度量。		857
Bartlett 的球形度检验	近似卡方	2160.400
	df	595
	Sig.	.000

图 22-11 KMO 检验和 Bartlett 球形检验输出

印证了作因子分析的必要性。

Bartlett 球形检验统计量的 Sig 值小于 0.01，由此否定相关矩阵为单位阵的零假设，即认为各变量之间存在显著的相关性，这与从相关矩阵得出的结论一致。

(3) 公因子提取的方差。如图 22-12 所示，给出了公因子对初始变量方差的提取情况，也就是常说的变量共同度。其中的“提取”一栏就是变量共同度的取值，代表了所有公因子能够解释的每个变量方差的比例，本例的方差提取多数都在 70% 左右，可见公因子对变量方差的解释效果可以接受。

(4) 方差解释表。如图 22-13 所示，方差解释表给出了每个因子所解释的总方差比例，以及所解释方差的累计和。观察初始特征值的“累积%”一列，前 8 各公因子的特征值都大于 1，且解释的累计方差达到了 68.558%，也就是说总体近 70% 的信息可以由这 8 个公共因子来解释，本例就取这前 8 个公因子进行分析。

公因子方差		
	初始	提取
无机	1.000	.746
哲学	1.000	.710
思品	1.000	.715
高数	1.000	.881
体育 I	1.000	.818
大学英语 I	1.000	.779
有机	1.000	.716
毛概	1.000	.758
高数 II	1.000	.871
物理	1.000	.809
体育 II	1.000	.781
大学英语 II	1.000	.811
有机 II	1.000	.875
分化	1.000	.588
英语	1.000	.487
体育 III	1.000	.371
大学英语 III	1.000	.597
计算机基础与应用	1.000	.582
分化 II	1.000	.718
物化	1.000	.777
体育 IV	1.000	.942
大学英语 IV	1.000	.381
计算机技术基础	1.000	.891
世国	1.000	.888
物化 II	1.000	.713
药物化学	1.000	.988
生物化学	1.000	.418
人体解剖生理学	1.000	.883
新药开发	1.000	.487
药理学	1.000	.882
药分 I	1.000	.718
微生物	1.000	.627
天然药化	1.000	.558
专业英语	1.000	.858
药理学	1.000	.854

图 22-12 公因子方差输出表

说明的总方差									
成分	初始特征值			提取平方和载入			旋转平方和载入		
	合计	方差的 %	累积 %	合计	方差的 %	累积 %	合计	方差的 %	累积 %
1	12.262	35.034	35.034	12.262	35.034	35.034	4.655	13.300	13.300
2	2.621	7.487	42.521	2.621	7.487	42.521	4.262	12.178	25.478
3	2.177	6.321	48.742	2.177	6.221	48.742	3.398	9.708	35.186
4	1.684	4.811	53.554	1.684	4.811	53.554	3.255	9.300	44.485
5	1.493	4.265	57.819	1.493	4.265	57.819	3.074	8.783	53.268
6	1.427	4.077	61.896	1.427	4.077	61.896	2.357	6.735	60.003
7	1.277	3.650	65.546	1.277	3.650	65.546	1.668	4.765	64.768
8	1.054	3.012	68.558	1.054	3.012	68.558	1.327	3.790	68.558
9	.905	2.586	71.144						
35	.075	.214	100.000						

提取方法：主成分分析。

图 22-13 方差解释结果

最后一栏“旋转平方和载入”表示经过因子旋转后得到的新公因子的方差贡献值、方差贡献率和累计方差贡献率，可以看到和未经旋转相比，每个因子的方差贡献值有变化，但累计方差贡献率不变。

(5) 方差解释表。如图 22-16 所示，是关于初始特征值（也就是方差贡献）的碎石图，它实际上就是根据图 22-13 中“初始特征值”一栏下的“合计”列所作的图形，特征值按照降序排列。从图中看第 1 个公因子的方差解释贡献最大，随后因子的方差贡献趋缓。

(6) 旋转前后的因子载荷矩阵。如图 22-14 所示，给出了旋转前后的因子载荷矩阵，

其中“成分矩阵”表是初始的未经旋转的载荷矩阵，“旋转成分矩阵”是经过旋转后的载荷矩阵。一般情况下，经过因子旋转后变量在因子上的载荷分布更加分散，因而比未旋转时容易解释。

成分矩阵 <sup>a</sup>								
	成分							
	1	2	3	4	5	6	7	8
无机	644	054	- 114	- 009	- 188	046	131	- 247
哲学	728	- 046	- 181	- 102	- 329	026	- 106	- 151
思品	504	133	- 125	- 384	- 085	- 009	026	291
高数I	402	- 458	- 044	142	119	088	516	002
体育I	286	- 176	- 043	002	- 087	705	- 053	182
大学英语I	701	064	- 233	307	- 305	043	- 155	125
有机I	746	- 240	- 139	168	127	- 182	013	156
毛概	771	- 121	- 228	- 107	- 023	093	- 177	129
高数II	579	- 353	- 028	- 094	374	- 095	153	- 169
物理	818	- 171	- 182	179	100	084	082	180
体育II	042	303	070	242	- 235	226	551	413
大学英语II	534	- 030	- 338	605	- 033	013	- 196	075
有机II	637	- 177	- 158	033	032	- 404	198	073
分化I	527	- 138	256	- 133	- 113	- 222	- 137	355
数统	507	020	- 408	- 042	228	- 020	371	- 203
体育III	287	- 282	727	259	- 055	154	085	- 199
大学英语III	617	- 046	- 172	- 215	- 318	156	- 212	- 220
计算机基础与应用	584	073	- 043	- 189	188	348	197	125
分化II	815	046	154	- 017	005	- 150	- 116	089
物化I	817	- 057	204	048	092	- 221	015	- 075
体育IV	321	- 294	727	296	094	151	011	063
大学英语IV	519	- 054	- 319	333	153	102	- 019	- 179
计算机技术基础	851	- 090	111	- 070	- 158	068	030	- 116
世贸	578	- 124	- 061	- 522	- 140	219	035	006
物化II	767	037	057	- 034	338	046	- 028	042
药物化学	760	063	169	- 152	- 058	- 068	026	132
生物化学	635	- 013	305	- 233	087	- 204	086	201
人体解剖生理学	679	047	238	- 098	- 238	- 090	- 096	- 223
新药开发	223	- 055	051	- 066	586	417	- 324	- 018
药剂学	610	505	046	- 067	090	082	127	- 132
药分I	668	403	093	245	157	114	040	- 039
微生物	480	439	276	- 016	287	091	- 186	049
天然药化	311	488	027	293	184	- 006	- 198	183
专业英语	037	756	147	- 097	156	058	146	- 053
药理学	396	591	- 062	079	- 122	- 061	117	- 250
提取方法 主成分分析法。								
旋转成分矩阵 <sup>a</sup>								
	成分							
	1	2	3	4	5	6	7	8
无机	135	554	226	222	326	101	- 048	031
哲学	279	711	316	073	163	028	- 015	- 054
思品	516	359	- 005	142	079	- 259	135	144
高数I	132	074	125	- 233	681	246	075	242
体育I	- 021	334	165	- 147	- 011	137	622	299
大学英语I	270	448	002	126	038	031	004	147
有机I	513	164	502	- 016	428	070	053	- 054
毛概	458	482	382	022	217	- 079	275	- 063
高数II	322	137	092	- 029	651	143	159	- 265
物理	499	251	531	045	383	044	144	- 058
体育II	- 027	124	019	182	- 002	039	- 048	827
大学英语II	068	123	874	064	142	034	043	018
有机II	504	173	303	006	492	- 027	- 230	- 030
分化I	717	165	122	- 036	- 046	165	- 002	008
数统	047	246	169	201	691	- 212	034	013
体育III	112	062	- 019	008	083	696	058	007
大学英语III	173	756	203	045	061	004	104	- 133
计算机基础与应用	284	291	040	241	357	- 006	451	199
分化II	602	334	303	281	177	201	048	108
物化I	545	273	272	259	366	318	- 031	- 166
体育IV	205	045	068	- 032	- 003	885	085	050
大学英语IV	- 037	201	552	129	396	022	145	- 098
计算机技术基础	421	593	229	165	299	296	092	- 010
世贸	379	617	142	- 020	208	- 059	278	027
物化II	423	218	240	343	405	142	309	- 178
药物化学	609	373	130	260	181	174	074	048
生物化学	703	167	- 037	199	220	187	038	- 003
人体解剖生理学	366	546	120	254	086	323	- 096	- 145
新药开发	046	079	076	136	105	066	726	- 326
药剂学	203	324	094	680	223	025	103	063
药分I	195	362	391	551	073	219	021	198
微生物	401	- 005	112	621	- 018	114	157	- 174
天然药化	177	- 124	398	538	- 127	- 025	128	025
专业英语	- 041	- 056	- 202	757	- 074	- 106	027	137
药理学	- 011	308	161	656	093	- 044	- 211	067
提取方法 主成分分析法。								

图 22-14 旋转前后的因子载荷

对一个变量来说,载荷绝对值较大的因子与它关系更为密切,也更能代表这个变量。按照这一观点,在图 22-14 中的旋转成分矩阵,特意把因子载荷较大的单元格用蓝色标识出来。

(7) 因子得分的系数矩阵。如图 22-15 所示,输出的是因子得分系数矩阵。对于每个因子,把系数和对应的课程名称相乘后再求和,可以得到最终的因子得分公式,利用它就能够对所有案例进行因子评分了。例如,因子 1 的得分公式为: $FAC1\_1 = -0.152 * \text{无机} - 0.065 * \text{哲学} + 0.232 * \text{思品} - 0.054 * \text{高数 I} + \dots + 0.121 * \text{微生物} + 0.069 * \text{天然药化} - 0.040 * \text{专业英语} - 0.154 * \text{药理学}$ 。

成分得分系数矩阵								
	成分							
	1	2	3	4	5	6	7	8
无机	-152	198	-023	046	102	043	-105	002
哲学	-065	265	029	-046	-064	-017	-082	-050
思品	232	045	-099	-032	-076	-204	062	148
高数I	-054	075	-045	-101	362	093	-006	227
体育I	-088	123	046	-110	-109	045	417	244
大学英语I	-002	079	266	-049	-160	-027	-040	110
有机I	136	-133	143	-077	080	-043	-028	008
毛概	086	069	079	-082	-073	-111	126	-015
高数II	-010	-072	-087	-014	299	027	029	-164
物理	109	-095	180	-065	017	-056	034	-001
体育II	064	-105	013	029	045	011	-002	637
大学英语II	-080	-079	407	-034	-066	-002	-001	011
有机II	152	-093	033	-054	173	-079	-224	020
分化I	350	-078	004	-106	184	-023	-032	063
数统	-145	014	-058	075	372	-106	-044	013
体育III	-096	010	-049	012	026	433	-001	001
大学英语III	-115	335	-013	-043	-109	-015	009	-120
计算机基础与应用	015	002	-105	042	116	-053	261	178
分化II	146	-021	032	030	-065	021	-034	-060
物化I	080	-047	000	048	076	092	-104	-108
体育IV	-022	-023	007	-025	-054	409	024	043
大学英语IV	-198	-004	196	034	139	014	046	-085
计算机技术基础	-024	156	-040	-006	019	095	-022	-001
世贸	053	237	-209	-074	-002	-084	131	047
物化II	025	-077	-013	092	107	012	140	-114
药物化学	172	012	-063	018	-042	005	-011	064
生物化学	277	-096	-131	009	012	-002	-028	046
人体解剖生理学	-022	192	-059	048	-067	126	-137	-127
新药开发	-052	-111	016	065	-005	011	483	232
药剂学	-075	050	-078	231	078	-002	018	018
药分I	-082	059	099	152	-056	090	-032	119
微生物	121	-139	-002	209	-084	010	080	-137
天然药化	069	-181	197	170	-141	-036	091	006
专业英语	-040	-038	-127	307	029	-032	027	059
药理学	-154	113	-004	240	049	003	-181	-009

提取方法 主成分分析法。

图 22-15 因子得分的系数矩阵

(8) 综合评分。如果要关心的是学生的综合能力,可对 8 个公因子的得分进行加权求和,权数就取其方差贡献值或方差贡献率,参看图 22-13 中“旋转平方和载入”一栏里的“合计”(方差值)、“方差的%”(方差贡献率)。本例采用方差贡献率作为加权变量,8 个旋转后公因子的方差贡献率依次为:13.3%、12.18%、9.71%、9.3%、8.78%、6.74%、4.77%、3.79%。

由此可得学生的综合得分计算公式如下:

$$zF = 13.30\% * FAC1\_1 + 12.18\% * FAC2\_1 + 9.71\%$$

$$* FAC3\_1 + 9.30\% * FAC4\_1 + 8.78\% * FAC5\_1 + 6.74\% * FAC6\_1 + 4.77\% * FAC7\_1 + 3.79\% * FAC8\_1, \text{其中 } FACn\_1 \text{ 表示第 } n \text{ 个因子的得分。}$$

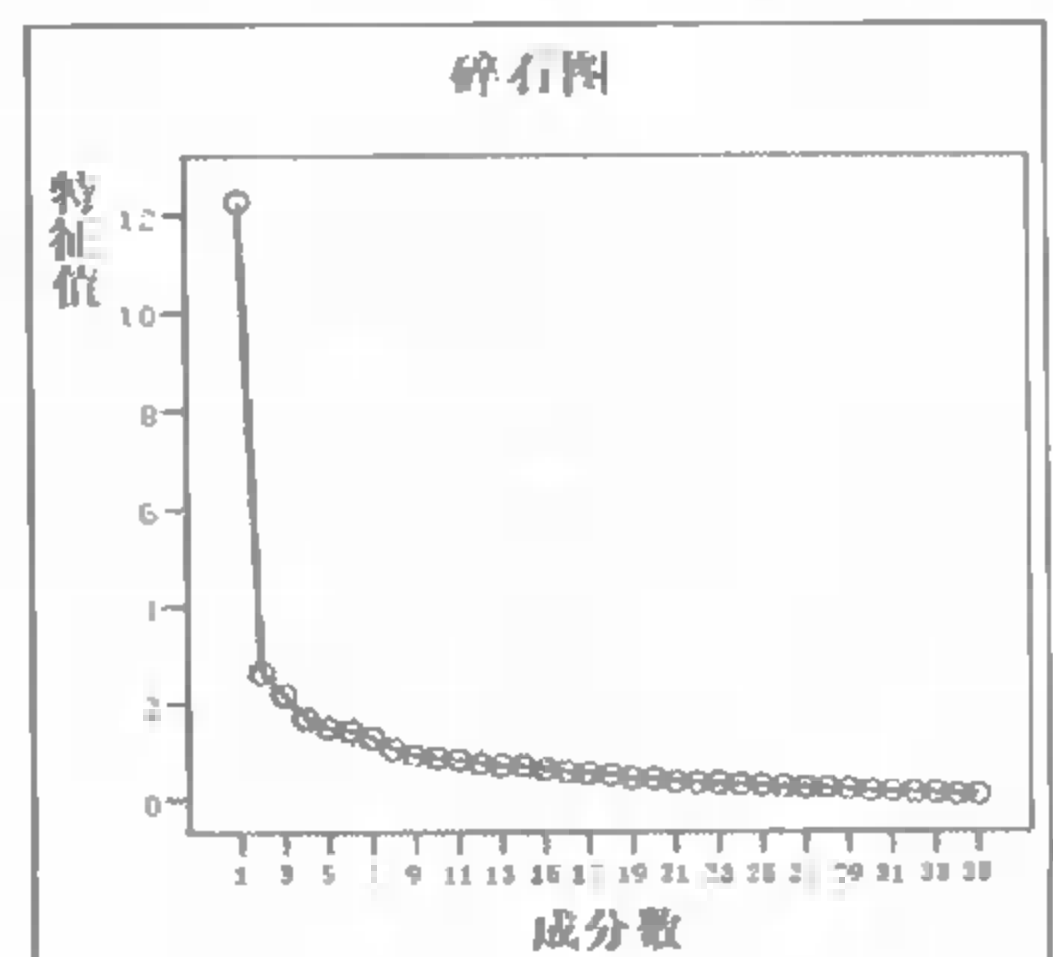


图 22-16 碎石图



## 22.5 进一步的分析与应用

通过前面的分析,利用因子得分系数和方差贡献率已经得到了能够进行综合评分的公式,由此可以对学生成绩进行更为科学的判断和排序。但有时我们不仅关心综合成绩,还要关心由成绩所反应出来的不同类型的学生的学习情况,下面就对各因子所反应的特征加以提取和详细分析。

根据图 22-14 中的旋转成分矩阵提供的信息,把在每个因子上载荷较大的变量提取出来,汇总成为表 22-3,下面以此表格为基础对各因子进行深入研究。

表 22-3 对各因子影响最大的变量

	主要影响变量
因子 1	思品、有机 1、有机 2、分化 1、分化 2、物化 1、物化 2、药物化学、生物化学
因子 2	无机、哲学、毛概、大学英语 3、计算机技术基础、世贸、人体解剖生理学
因子 3	大学英语 1、大学英语 2、大学英语 4、物理
因子 4	药剂学、药分 1、微生物、天然药化、专业英语、药理学
因子 5	高数 1、高数 2、数统
因子 6	体育 3、体育 4、
因子 7	体育 1、计算机应用与基础、新药开发
因子 8	体育 2

影响因子 1 的主要变量都是化学课,故可以把它总结为“专业基础水平”因子;影响因子 3 的主要变量都是英语课,故可以把它总结为“英语水平”因子;影响因子 4 的主要变量都是指定专业方向的课程,故可以把它总结为“专业高级水平”因子;影响因子 5 的主要变量都是数学课,故可以把它总结为“数学水平”因子。由因子模型诱导出的如上 4 个因子都比较清晰且易于解释。

因子 2 的影响变量主要是全校学生都要上的公共课,故建议把它总结为“基本学习水平”因子;因子 6、因子 7、因子 8 这 3 个因子的主要影响变量包括了所有的体育课程,可以考虑把它们 3 个合为一个因子,即“身体素质水平”因子。

至此,我们得到了 6 个方面的潜在因素:专业基础水平、专业高级水平、英语水平、数学水平、基本学习水平、身体素质水平,而且这些因素只有通过因子分析模型才能得到。结合前面得到的因子得分函数,可以计算每个学生在这 6 方面的得分,由此判断他的学习类型;还可以对所有学生按照这 6 个因素的得分进行分类,以观察他们整体学习情况的分布态势。

## 22.6 建议和推广

### 22.6.1 高中生的成绩综合评价

本章所举实例是对大学生在校成绩的综合评价,这里的方法同样适用于对高中生进行成绩综合评价。进入 21 世纪,家长们都更加关注孩子的学习问题,尤其高考的走俏更是极大地加重了人们对高中生成绩的关注。有研究显示,高中生的学习策略、自我监控学习行为与学

习成绩都显著地相关。

有对小学生的研究表明，除学生自身努力的程度外，其性别、入学年龄、家庭背景等因素对学习成绩也有明显影响。那么，当学生进入高中后，随着其生理及心理年龄的变化，这些因素的影响是否还存在呢？鉴于此，“基于 SPSS 软件分析影响高中生学习成绩的各因素”（戴凌霄，《信息技术与信息化》，2006）一文，进行了一次小规模的活动，并用 SPSS 对调查结果进行了分析，表明性别、是否是独生子女与学生的语文、英语成绩有较强的相关性。

### 22.6.2 对缺失数据的处理

对于由违纪、缺考而导致考生某课程没有成绩的情况，本章中的做法是把他的这门课程得分记为 0 分。除此之外，还有其它的处理方式，比如当考生数量较多时，可以直接从分析数据中删除此考生的记录；或者给这个考生的此门课程赋以较低的非零分（如百分制下的 10 分）；或者给这个考生的此门课程赋以接近平均成绩的得分。

某门课程没有成绩，并不能代表考生对此课程的知识一无所知，赋 0 可以加大此课程在区分学生能力方面的力度，但不一定在任何情况下都是恰当的。应该根据分析数据的特点选择一种切合实际的处理方式。

### 22.6.3 多种方法结合的综合评价模型

成绩综合评价有多种模型方案，而每种评价方案都有各自的优点和缺点，对于每个学生来说，按照不同方案的排名结果就可能不一样，在一种方案里排名较靠前，在另一种方案里排名可能就靠后，所以用任何一种排名方案就可能对某些学生来说有利，而对某些学生来说不利。

“学生成绩排名的综合评价模型”（吴海英，《大学数学》，2006）一文，提出一种新的排名方案，以尽可能的减少这种利和不利，从而做到对任何学生来说都比较公平。这个方案就是将多种排名方案所得到的结果进行主成分分析，取第一主成分的值来确定最终排名，这种方案虽说比较复杂，但是与其他单个的排名方案相比，是一种较科学、公平和合理的评价模型。而且按照这个综合方案对文献中的实例进行排名后，从结果也看出综合排名确实平衡了单个模型的排名结果，显示了它的公平性。

第

23

章

高等教育办学条件的聚类分析

高校的基本办学条件，作为高等教育的物质载体，其质量是保证高等教育教学质量的基础和重要前提，因此对它的研究具有十分重要的意义。

聚类分析是较为常用的数理统计方法，尤其在处理繁杂的大样本数据时，能快捷有效的把数据条理化，本章就利用了聚类分析法的这个特点，对事先没有任何了解的高校总体进行聚类。本章参考教发[2004]2号文件“普通高等学校基本办学条件指标（试行）”，将其中规定的五项基本办学条件指标作为研究变量，对河北省的100多所高校进行聚类分析，最后给出关于改善办学条件的某些建议。

### 23.1 数据描述

1977年，河北省只有高等学校22所，而且规模小，结构不合理，功能也不健全；到1998年，高等学校数量增加为46所，初步形成了以河北大学、河北工业大学为龙头，包括河北医科大学、河北师范大学、河北农业大学、河北经贸大学、河北科技大学在内的7所大学为骨干，50个重点学科为基础，带动所有高校共同发展的河北高等教育体系框架；1998年9月，燕山大学等3所普通高等学校划归河北省，进一步加强了其高等教育的实力。到1998年，河北省高等学校在校生达到144383人，比1977年增加了5.7倍；同时，为了适应经济发展的需要，不断地调整、优化高校专业设置，到1995年河北省高等院校本、专科专业的种类达到317种，专业点增加到537个。

从20世纪80年代中期开始，河北省便开始把高教工作重点放在提高教育质量上，采取多种有力措施，提高师资水平。尤其近几年来结合教育部对本科院校的资格评审工作和“211工程”，实施了“双重工程”，使河北省的高教质量有了质的飞跃。

本章就选取河北省的107所高校进行研究。

#### 23.1.1 关于基本办学条件指标合格与否的判定

本节给出教发[2004]2号文件“普通高等学校基本办学条件指标（试行）”中，关于基本办学条件指标合格与限制招生的评判标准。

表 23-1 基本办学条件指标：合格

学 校 类 别	本 科				
	生师比	具有研究生学位教师占 专任教师的比例（%）	生均教学行政用房 （平方米/生）	生均教学科研仪器 设备值（元/生）	生均图书 （册/生）
综合、师范、民族	18	30	14	5 000	100
工科、农、林院校	18	30	16	5 000	80

续表

学 校 类 别	本 科				
	生师比	具有研究生学位教师占 专任教师的比例 (%)	生均教学行政用房 (平方米/生)	生均教学科研仪器 设备值 (元/生)	生均图书 (册/生)
医学院校	16	30	16	5 000	80
语文、财经、政法	18	30	9	3 000	100
体育院校	11	30	22	4 000	70
艺术院校	11	30	18	4 000	80

学校类别	高职 (专科)				
	生师比	具有研究生学位教师占 专任教师的比例 (%)	生均教学行政用房 (平方米/生)	生均教学科研仪器 设备值 (元/生)	生均图书 (册/生)
综合、师范、民族	18	15	14	4 000	80
工科、农、林院校	18	15	16	4 000	60
医学院校	16	15	16	4 000	60
语文、财经、政法	18	15	9	3 000	80
体育院校	13	15	22	3 000	50
艺术院校	13	15	18	3 000	60

备注:

聘请校外教师经折算后计入教师总数,原则上聘请校外教师不超过专任教师总数的四分之一。

凡生师比指标不高于表中数值,且其它指标不低于表中数值的学校为合格学校。

表 23-2

基本办学条件指标: 限制招生

学 校 类 别	本 科				
	生师比	具有研究生学位教师占 专任教师的比例 (%)	生均教学行政用 房 (平方米/生)	生均教学科研仪器 设备值 (元/生)	生均图书 (册/生)
综合、师范、民族	22	10	8	3 000	50
工科、农、林院校	22	10	9	3 000	40
医学院校	23	10	5	2 000	50
语文、财经、政法	17	10	13	2 000	35
体育院校	17	10	11	2 000	40
艺术院校	22	10	8	3 000	50

学 校 类 别	高职 (专科)				
	生师比	具有研究生学位教师占 专任教师的比例 (%)	生均教学行政用 房 (平方米/生)	生均教学科研仪器 设备值 (元/生)	生均图书 (册/生)
综合、师范、民族	22	5	8	2 500	45
工科、农、林院校	22	5	9	2 500	35
医学院校	23	5	5	2 000	45
语文、财经、政法	17	5	13	2 000	30
体育院校	17	5	11	2 000	35
艺术院校	22	5	8	2 500	45

备注:

1. 生师比指标高于表中数值或其它某一项指标低于表中数值,即该项指标未达到规定要求。
2. 凡有一项指标未达到规定要求的学校,即被确定为限制招生(黄牌)学校。
3. 凡两项或两项以上指标未达到规定要求的学校,即被确定为暂停招生(红牌)学校。
4. 凡连续三年被确定为“黄”牌的学校,第三年即被确定为暂停招生(红牌)学校。



### 23.1.2 指标选取

参考上一节中“普通高等学校基本办学条件指标(试行)”(教发[2004]2号)给出的条件,选取能反应基本办学条件的5个指标:生师比、生均教学行政用房、生均图书数、研究生教师的比例、生均教学科研仪器设备值,其计算方法如下。

(1) 生师比=折合在校生数/教师总数;该指标反应了某学校学生和教师数量是否均衡,取值在一定范围内越小越好。

(2) 生均教学行政用房=(教学及辅助用房面积+行政办公用房面积)/全日制在校生数;该指标能反应某学校的学生能否有足够大的空间学习和娱乐。

(3) 生均图书=图书总数/折合在校生数;大学是饱览群书的大好时光,广泛的阅读对扩大在校学生视野极有好处,该指标反映的正是学生有无这样的途径及该途径的大小。

(4) 具有研究生学位教师占专任教师的比例=具有研究生学位的专任教师数/专任教师数;这是反映教师质量水平的一项指标。

### 23.1.3 数据格式

本章选取了河北省107所高校2005年的指定指标进行研究,其中本科院校43所,专科院校64所。所用数据文件为“河北高校办学条件数据.SAV”,数据格式如图23-1所示。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	name	String	30		学校名称			17	Left	Nominal
2	type	String	8		学校类型	None	None	7	Left	Nominal
3	x1	Numeric	8	2	生师比	None	None	8	Right	Scale
4	x2	Numeric	8	2	生均教学行政用房	None	None	8	Right	Scale
5	x3	Numeric	8	4	研究生教师比例	None	None	8	Right	Scale
6	x4	Numeric	8	2	生均图书	None	None	8	Right	Scale
7	x5	Numeric	8	2	生均仪器设备值	None	None	8	Right	Scale

图 23-1 河北高校办学条件的数据格式

## 23.2 聚类分析法简述

聚类分析法又称集群分析法,它是研究样品或指标分类问题的一种多元统计方法。聚类方法的内容十分丰富,包括系统聚类法、有序样品聚类法、动态聚类法、模糊聚类法、图论聚类法、聚类预报法等。下面简略介绍一下系统聚类法的基本原理。

为了将样品或指标进行分类,需要研究样品之间的关系,目前用得最多的方法有两个:一种方法是用相似系数,性质越接近的对象,他们相似系数的绝对值越接近1,而彼此无关的样品,他们的相似系数接近0;另一种方法是将每一个样品看作 $p$ 维空间的一个点,在此 $p$ 维空间定义距离,距离相近的点归为相同的类,距离较远的点归为不同的类,距离的远近是一个相对概念,需要根据具体情况具体对待。

在实际问题中,遇到的指标往往既有定量的也有定性的,本章涉及的数据均是定量指标,下面给出一些关于定量指标的距离和相似系数的计算方法。

设有 $n$ 个样品,每个样品测得 $p$ 项指标,由原始数据构成的矩阵为:  $X = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1p} \\ x_{21} & x_{22} & \cdots & x_{2p} \\ \vdots & \vdots & \cdots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{np} \end{bmatrix}$ ,

其中  $x_{ij}$  是第  $i$  个样品的第  $j$  个指标的观测数据。令  $d_{ij}$  表示变量  $X_i = (x_{i1}, \dots, x_{ni})'$  与变量  $X_j = (x_{1j}, \dots, x_{nj})'$  之间距离, 下面是两种常用的关于距离的定义。

● 明氏距离:  $d_{ij}(q) = \left( \sum_{a=1}^n |x_{ai} - x_{aj}|^q \right)^{1/q}$ , 当  $q=2$  时,  $d_{ij}(2)$  就是欧氏距离。

● 马氏距离:  $d_{ij}(M) = (X_{(i)} - X_{(j)})' S^{-1} (X_{(i)} - X_{(j)})$ , 其中  $S^{-1}$  为样本协差阵的逆矩阵。

马氏距离的特点是既排除了各指标之间的相关性, 而且不受各指标量纲的影响; 另外可以证明, 将原数据做线性变换后, 马氏距离保持不变。

本章中用到的是系统聚类法, 它又可细分为最短距离法、最长距离法、中间距离法、重心法、类平均法、可变类平均法、离差平方和法、最大似然谱系聚类、密度估计法等。系统聚类法的聚类原则取决于样品间距离(或相似系数)和类间距离的定义, 类间距离的定义不同就产生不同的系统聚类方法。关于系统聚类法的详细操作步骤, 请参考第 12 章的具体介绍。

### 23.3 SPSS 建模过程和结论分析

由于各学校办学层次的不同, 专科与本科之间进行聚类 and 比较并无大的现实意义, 所以本节将对 43 所本科院校、64 所专科院校分别进行生均指标的聚类分析。

#### 23.3.1 对专科院校进行聚类的设置操作

打开文件“河北高校办学条件数据.SAV”, 数据格式如图 23-1 所示。由于先只对专科院校进行分析, 而当前文件包含了所有类型的学校数据, 故在分析之前需要对数据进行“筛选”。

##### 1. 数据筛选

依次点击菜单“Data→Select Cases...”打开数据筛选的主设置界面, 如图 23-2 所示。在 Select 栏单击选中 If condition is satisfied 单选框; 在 Output 栏单击选中 Filter out unselected cases 单选框, 表示不满足前面设置条件的观测将在随后的分析中被剔除掉, 但是不会从当前数据集中将其删除。

在图 23-2 中, 点击 If condition is satisfied 选项下的 If 按钮, 弹出如图 23-3 所示的条件设置子界面, 在条件编辑框输入: type=“专”, 即表示选择学校类型为“专”的观测记录, 点击 Continue 返回主界面。

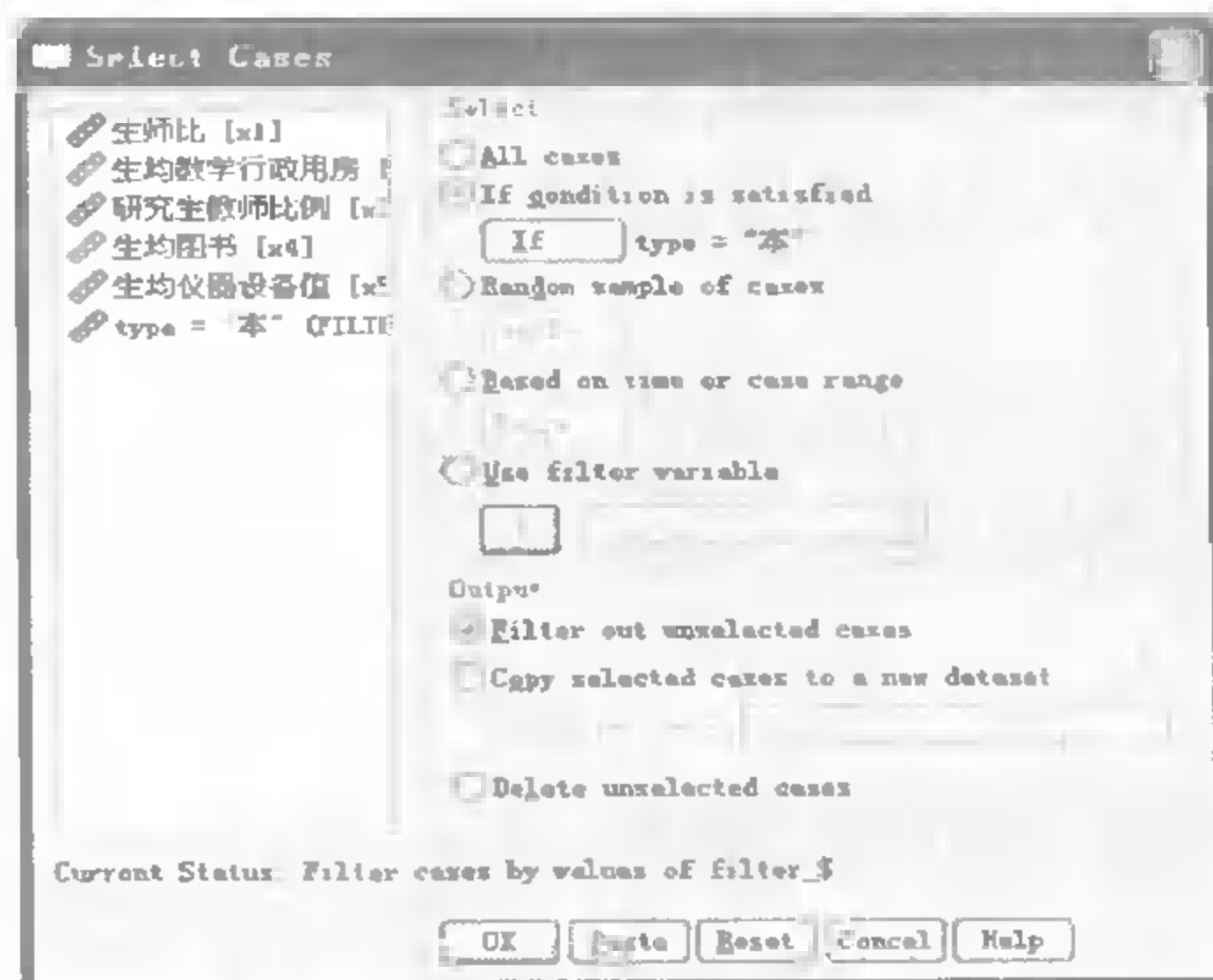


图 23-2 数据筛选的主设置界面

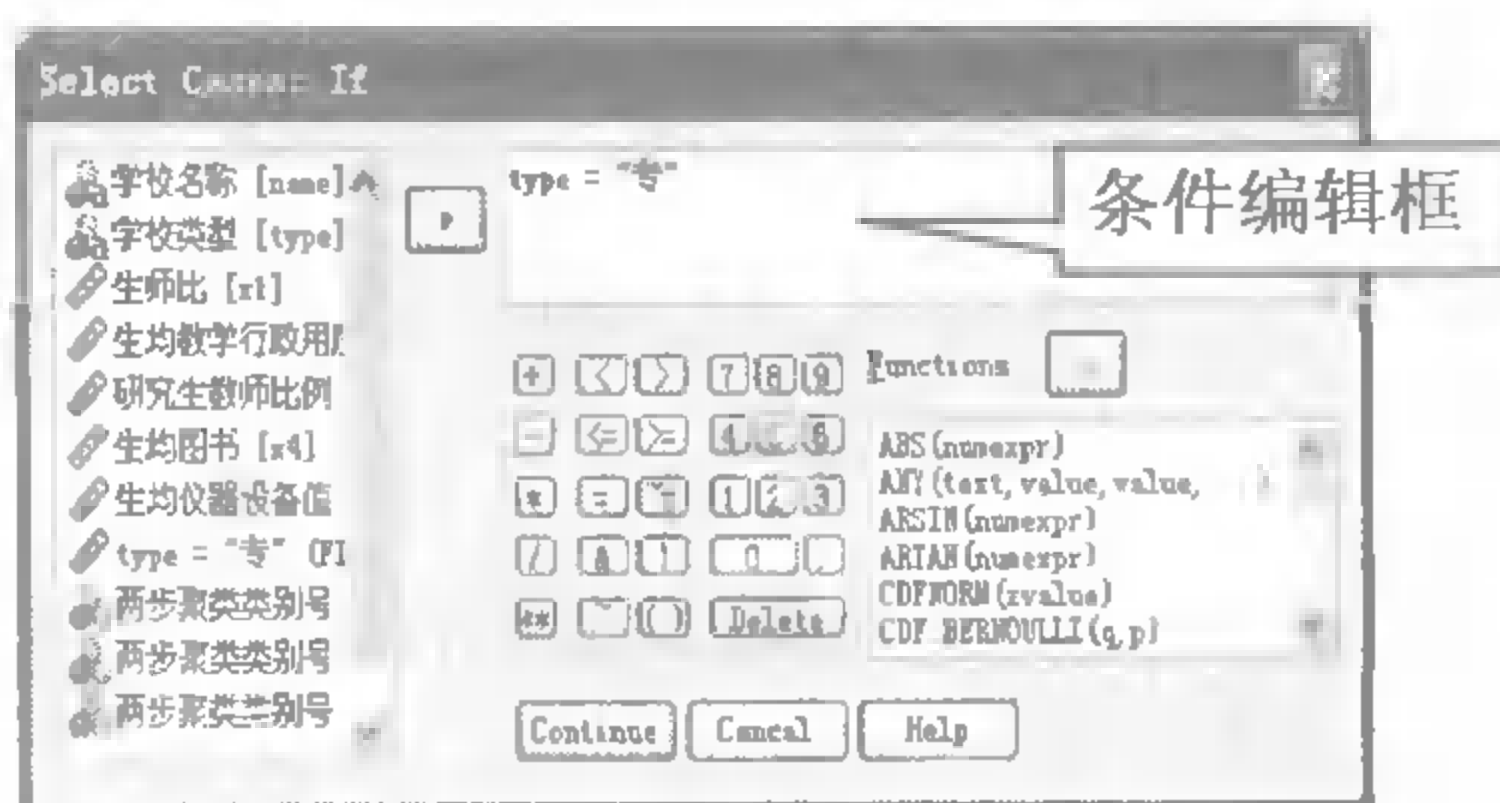


图 23-3 数据筛选的条件设置子界面

在图 23-2 中点击 OK 按钮运行后,在当前数据集生成一个名为 filter\_\$ 的过滤变量,对所有的本科院校取值为 0,专科院校取值为 1;随后的分析中将只使用 filter\_\$=1 的专科院校。

## 2. 二阶段聚类的参数设置


依次单击菜单 Analyze→Classify→TwoStep Cluster..., 弹出二阶段聚类过程的主设置面板,如图 23-4 所示。在变量列表选中从生师比至生均仪器设备值的 5 个变量,单击 Continuous Variables 栏左侧的  按钮,将其作为聚类变量选入;在 Distance Measure 栏,单击选中 Euclidean 单选框,以欧氏距离作为度量方式;其他设置选项采用默认方式,如图 23-5 所示。



图 23-4 二阶段聚类的主界面 1

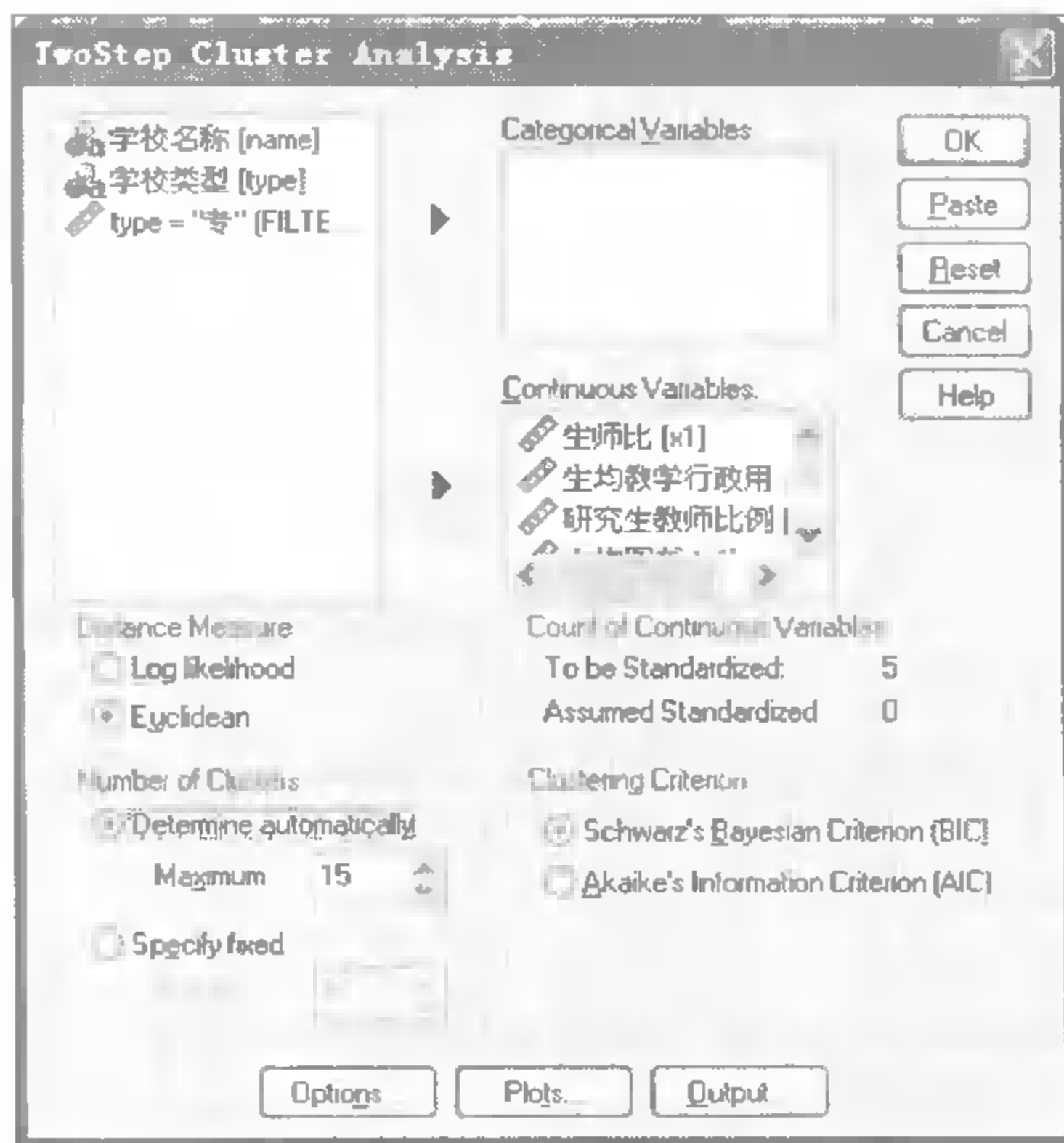


图 23-5 二阶段聚类的主界面 2

在图 23-5 中,单击 Options 按钮,弹出如图 23-6 所示的选项设置子界面。默认情况下,5 个连续变量都选入了 To be Standardized 列表,表示对它们进行标准化处理,由于欧氏距离对度量单位的依赖性,此处就采取默认方式,单击 Continue 按钮返回主界面。

在图 23-5 中,单击 Output 按钮,弹出如图 23-7 所示的输出设置子界面,勾选 Cluster pie chart 复选框,输出聚类结果饼图。单击 Continue 按钮返回主界面。

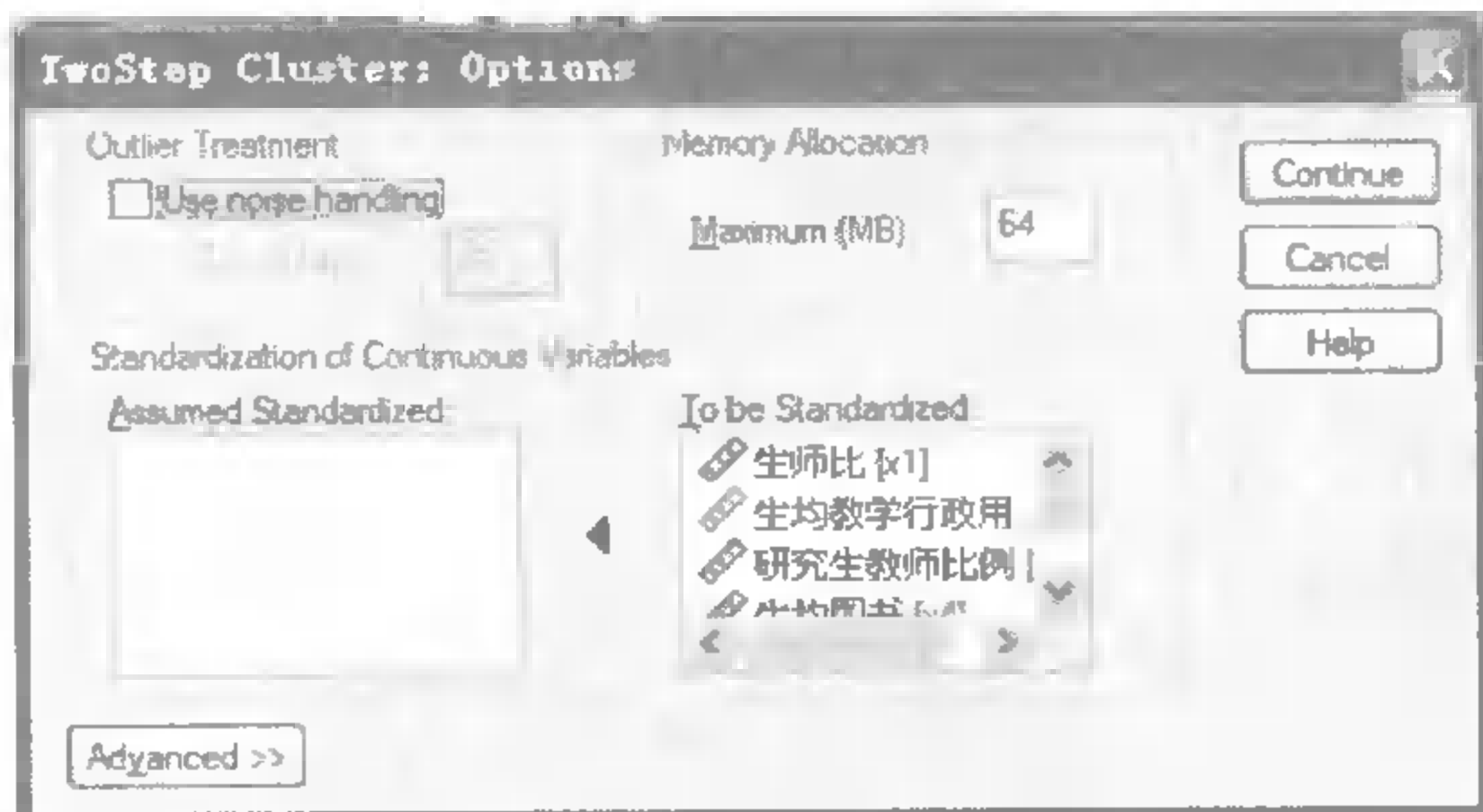


图 23-6 二阶段聚类的选项设置

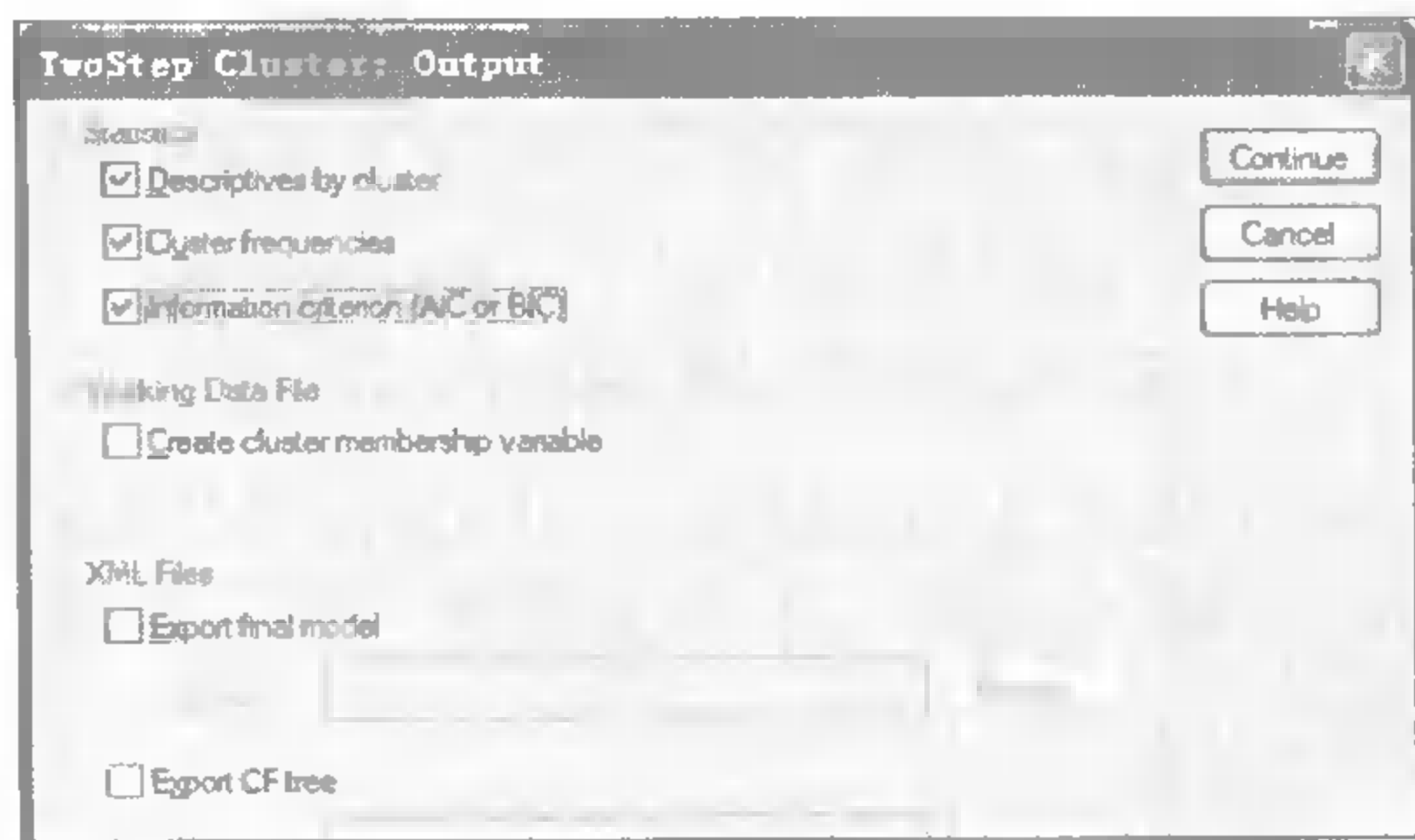


图 23-7 二阶段聚类的输出设置

在图 23-5 中单击 Plots 按钮,弹出如图 23-8 所示的作图设置子界面,勾选 Information criterion (AIC or BIC)复选框,输出 BIC 判别准则。单击 Continue 按钮返回主界面。

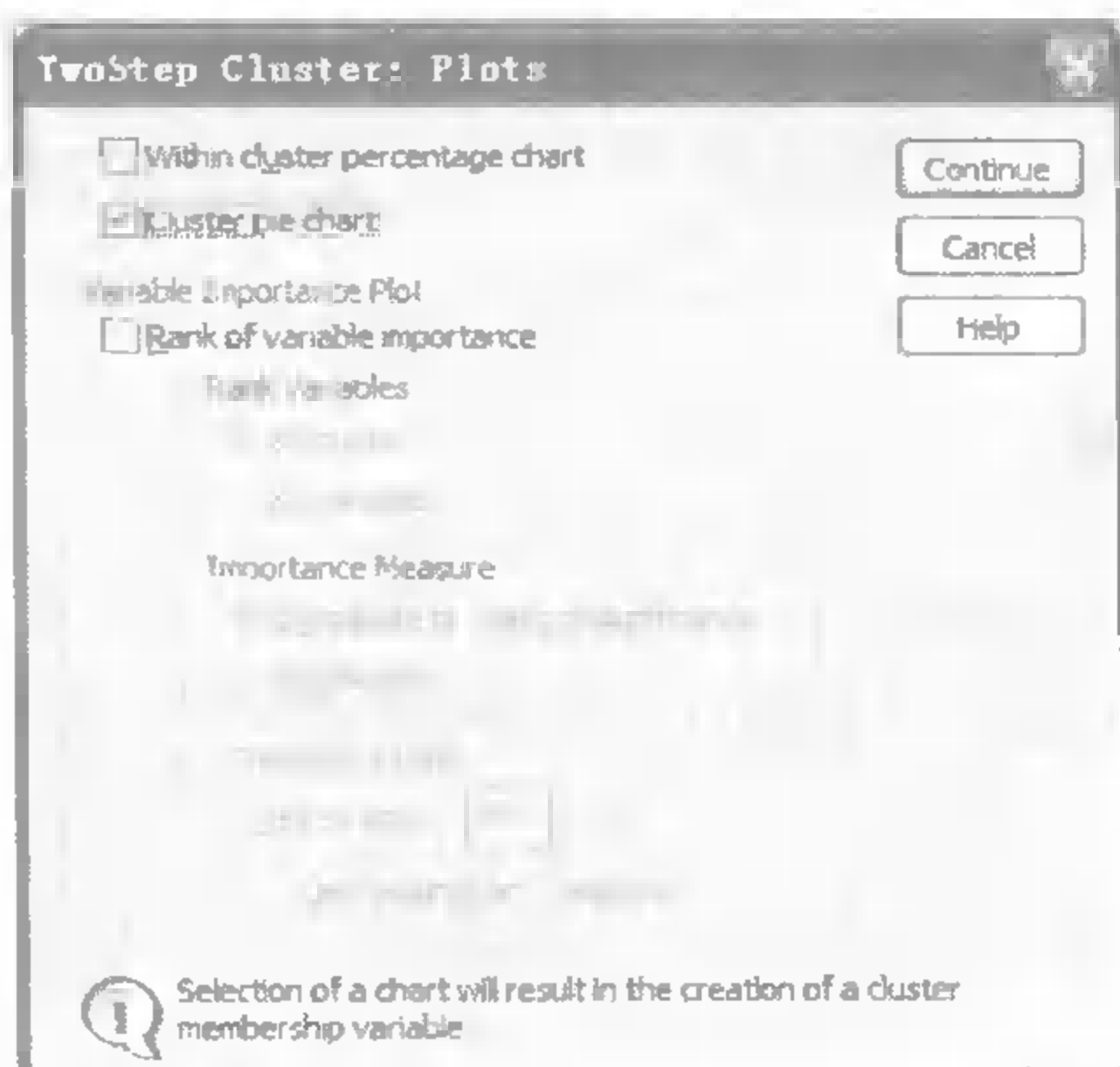


图 23-8 二阶段聚类的作图设置

### 3. 结果分析和改进

(1) 聚类个数的确定。在图 23-5 中单击 OK 按钮运行，SPSS Viewer 窗口输出的自动聚类结果如图 23-9 所示。

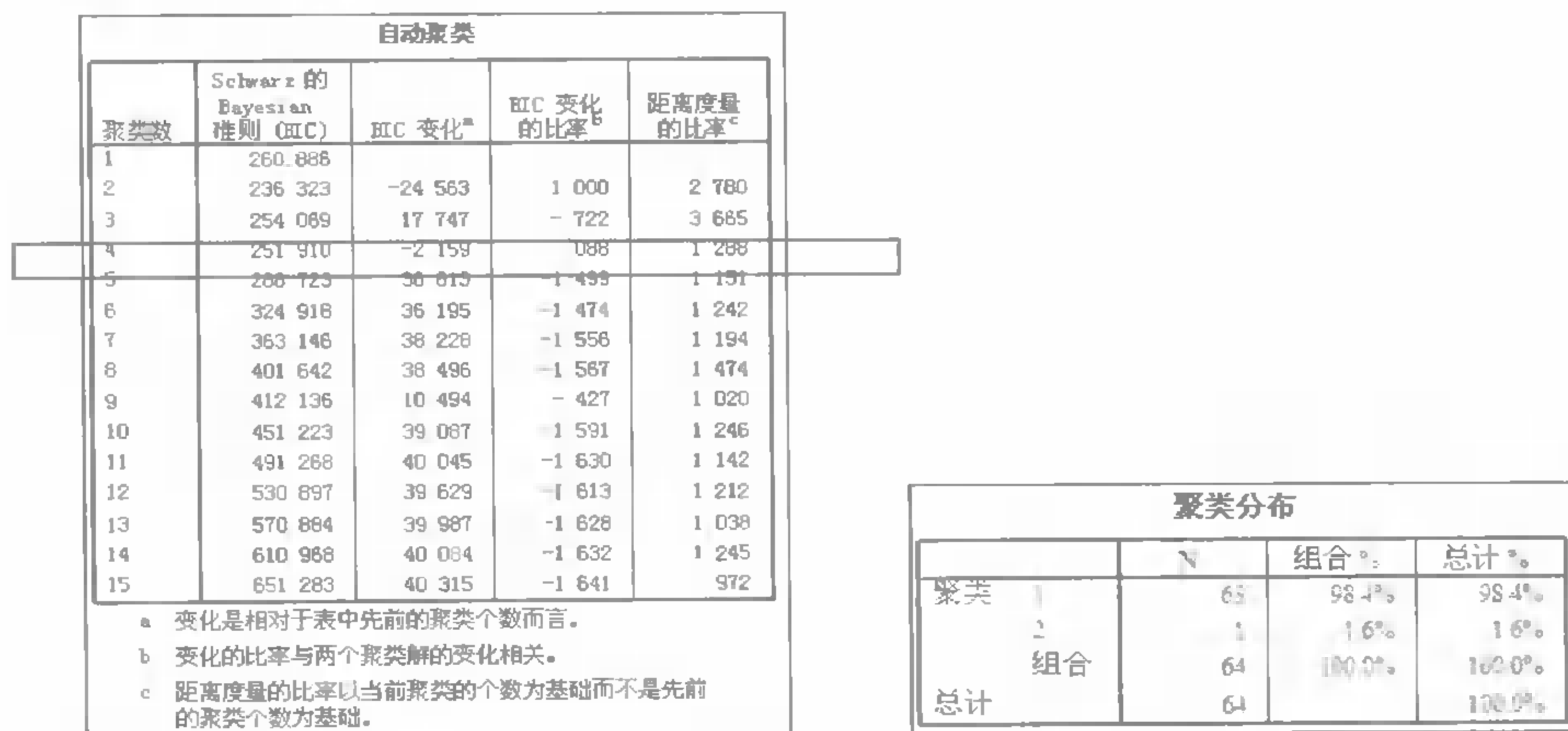


图 23-9 对专科院校的自动聚类结果

从“聚类分布”表格看，由系统自行确定的聚类个数为 2，而且其中一个类别的案例个数为 1，这个结论应用价值不大。

“自动聚类”表格中，观察 BIC 的几个判别准则，综合考虑 BIC 较小、两个比率较大的选取原则，建议设置聚为 4 个结果类别再进行一次聚类分析。

(2) 改进参数后的结果分析。在图 23-5 中，单击选中 Number of Clusters 栏中的 Specify fixed 单选框，在 Number 输入框键入 4 作为聚类个数；其他设置不变，如图 23-10 所示。

在图 23-10 中单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如图 23-11 和图 23-12 所示。

① 聚类结果分布。如图 23-11 所示，给出了聚类结果的分布统计表格和分布饼图，可见第 1、2 类的学校个

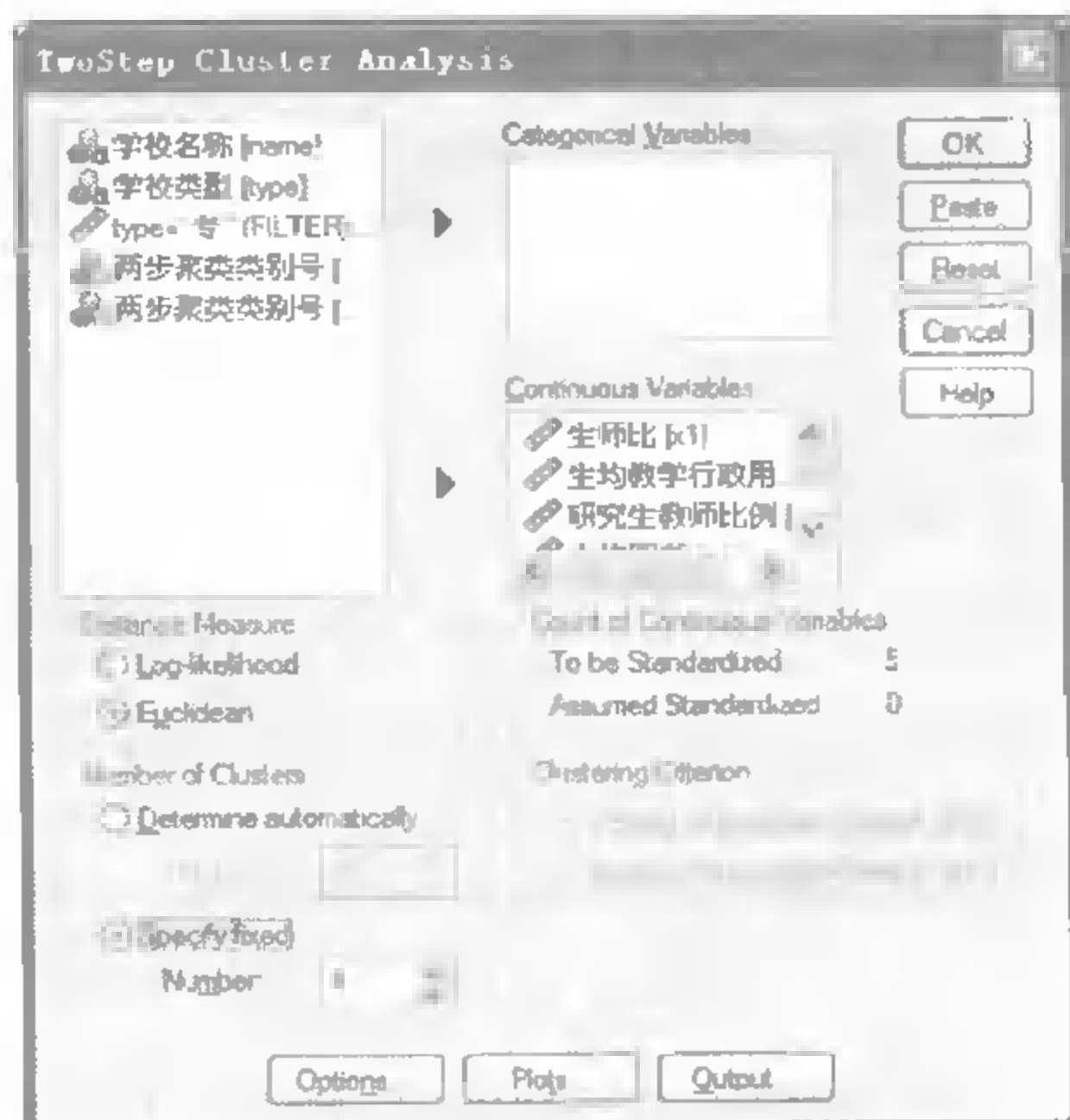


图 23-10 二阶段聚类的参数更改设置



数较多, 分别为 54 个、8 个, 第 3、4 类各有 1 个学校。

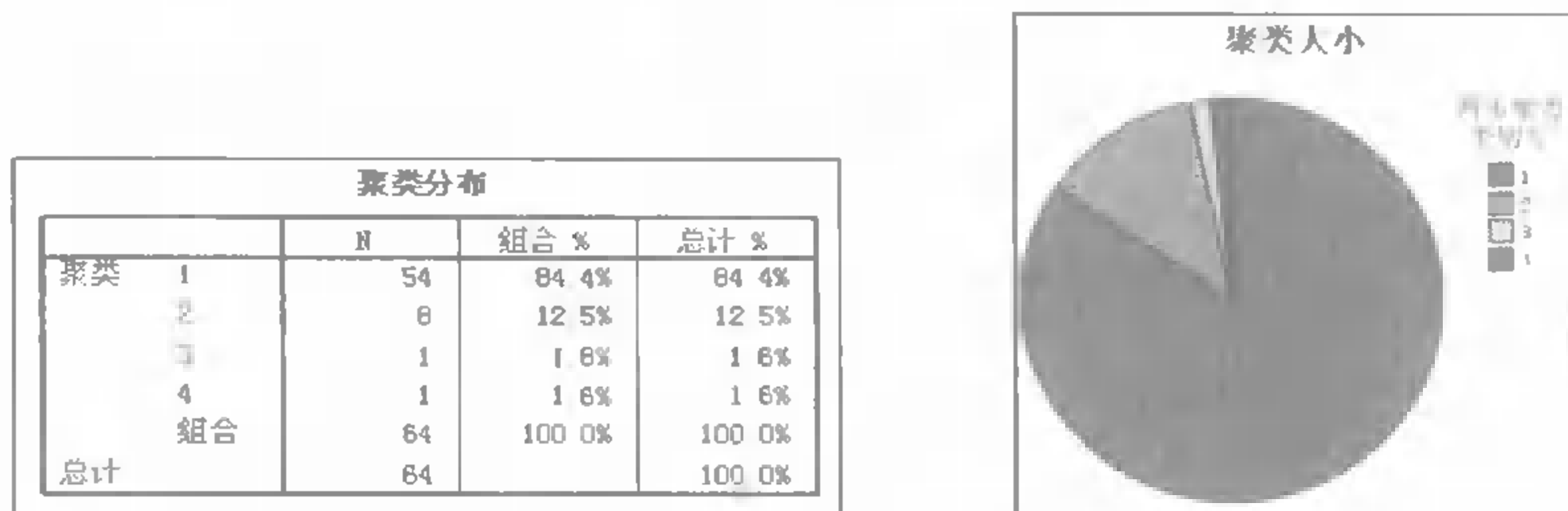


图 23-11 聚类结果的分布

② 聚类结果的描述性统计特征。如图 23-12 所示, 给出了各结果类别的质心统计信息, 结合原始数据做如下分析和建议。

质心										
聚类	生师比		生均教学行政用房		研究生教师比例		生均图书		生均仪器设备值	
	均值	标准差	均值	标准差	均值	标准差	均值	标准差	均值	标准差
1	14.5074	5.23963	16.2909	10.58977	106522	0972775	73.6380	38.67020	5066.9937	3294.2498
2	4.5825	1.76579	64.3775	26.16228	135713	1630772	237.4450	90.48818	15137.096	6224.8373
3	6300		208.2200		080800		559.7000		76358.210	
4	4.1100		77.8100		1.000000		60.6100		17003.370	
组合	12.8875	6.16911	26.2619	31.74599	123417	1530969	101.5050	92.40693	7657.4377	10345.055

图 23-12 聚类结果各类别的特征数据

第 3 类所包含的学校是石家庄科技信息职业学院, 它的生师比、生均用房、生均图书、生均仪器都比其它学校高出很多, 说明此学校的硬件设施非常好; 而有研究生学历的教师比重却明显偏低 (6%), 这一点就限制了它不能成为合格学校 (请参考表 23-1 中的指标要求)。建议此学校在坚实的硬件办学条件基础上, 积极引进人才, 促进学校建设的协调发展。

第 4 类所包含的学校是保定虎振职业技术学院, 它的情况与第 3 类的石家庄科技信息职业学院正好相反, 教师的综合水平较高, 但是反映基本办学条件的硬件措施暂时还跟不上步伐。

第 2 类包含的学校有: 石家庄东方美术职业学院、廊坊职业技术学院、唐山科技职业技术学院、石家庄工商职业学院、石家庄联合技术职业学院、石家庄外语翻译职业学院、河北司法警官职业学院, 它们的 5 个生均指标都不错, 基本办学条件是值得认可的, 是这一批学校里能够为学生提供较好环境、较高水平教育的学校。

第 1 类包含了大部分 (84.4%) 的学校, 但是其各项生均指标都不能或是只能刚刚达到合格水平 (请参考表 23-1 中的指标要求), 回顾近年的招生形式, 不断的扩招政策是引起基本办学条件降低的重要因素。

通过分析, 我们发现多数学校的生均教学水平都有低端化的倾向, 这应该引起有关部门和学校领导的重视, 如何探索适合于本学校的特色发展道路, 是关系学校自身和广大莘莘学子福祉的重要议题。

### 23.3.2 对本科院校的分析

#### 1. 数据筛选

下面只对本科院校进行分析, 在分析之前也需要对数据进行“筛选”。

依次单击菜单“Data→Select Cases...”打开数据筛选的主界面, 如图 23-2 所示, 在 Select

栏单击选中 If condition is satisfied 单选框；在 Output 栏单击选中 Filter out unselected cases 单选框；点击 If condition is satisfied 选项下的 If 按钮，弹出如图 23-13 所示的条件设置子界面，在条件编辑框输入：type=“本”，即表示选择学校类型为“专”的观测记录，点击 Continue 返回主界面。



图 23-13 对本科院校的筛选

在图 23-2 中点击 OK 按钮运行后，当前数据集的 filter\_ \$过滤变量，对所有的本科院校取值为 1，专科院校取值为 0；随后的分析中将只使用 filter\_ \$=1 的本科院校。

2. 参数设置

本节的分析步骤与对专科院校进行分析时一样，先用自动确定聚类个数的方法进行分析，并根据 BIC 准则判定合理的聚类个数；然后用新的聚类个数再进行一次分析。参数设置方法也与图 23-4~图 23-8 均相同。

3. 结果分析和改进

(1) 聚类个数的确定。在图 23-5 中单击 OK 按钮运行，SPSS Viewer 窗口输出的自动聚类结果如图 23-14 所示。

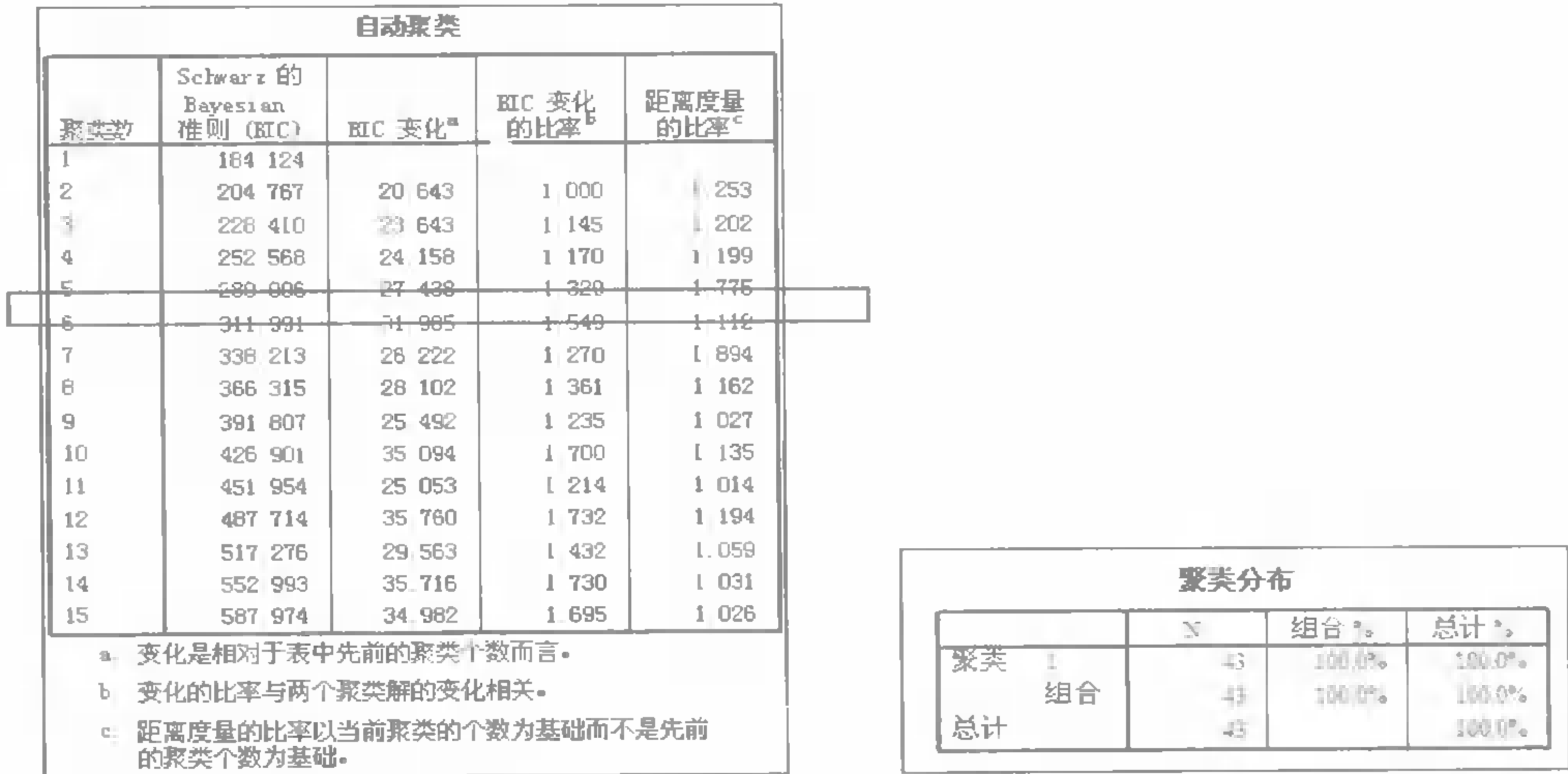


图 23-14 对本科院校的自动聚类过程

从“聚类分布”表格看，由系统自行确定的聚类个数为 1，这个结论没有价值。

“自动聚类”表格中，观察 BIC 的几个判别准则，综合考虑 BIC 较小、两个比率较大的选取原则，建议设置聚为 6 个结果类别再进行一次聚类分析。

(2) 改进参数后的结果分析。如图 23-10 所示, 在 Number of Clusters 栏中, Specify fixed 下的 Number 后键入 6 作为聚类个数; 其它设置同前, 点击 OK 按钮运行。SPSS Viewer 窗口的输出结果如下。

- 聚类结果分布。如图 23-15 所示, 给出了聚类结果的分布统计表格和分布饼图, 可见只有第 1 类的学校个数较多, 其他类别都只有 1 个或 2 个学校。

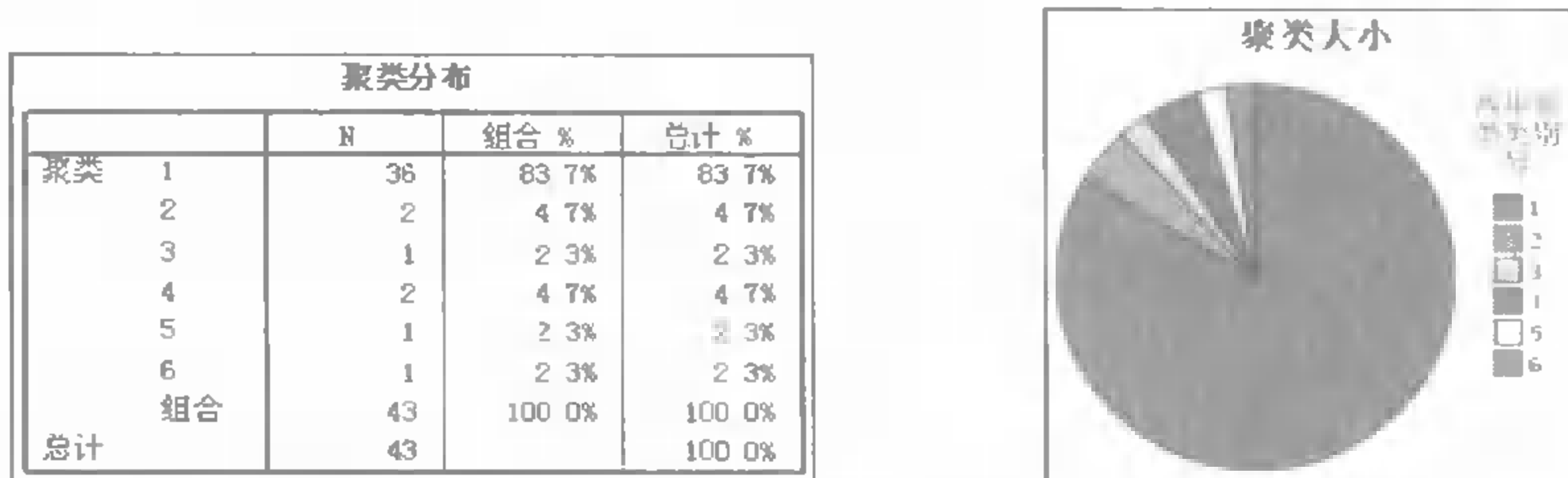


图 23-15 聚类结果的分布

- 聚类结果的描述性统计特征。如图 23-16 所示, 给出了各结果类别的质心统计信息, 结合原始数据加以分析, 会发现本科院校的聚类结果和专科院校有类似的结论。

聚类	生师比		生均教学行政用房		研究生教师比例		生均图书		生均仪器设备值	
	均值	标准差	均值	标准差	均值	标准差	均值	标准差	均值	标准差
1	18.0444	2.40445	10.9661	4.22784	314789	1271330	50.5564	22.14961	4144.9731	1509.9787
2	20.2550	1.68999	7.7300	1.92333	0.00000	0.000000	74.1400	45.07099	2168.5200	1157.2651
3	17.1400		43.5000		367300		125.0000		9523.8100	
4	17.6450	47376	10.0700	1.24451	191750	0.082731	37.3750	20.49903	11616.485	2770.4797
5	31.1100		3.0200		304300		9.0400		693.4800	
6	14.4500		40.2100		0.00000		54.4500		2882.7700	
组合	18.3279	3.07259	12.0258	7.84056	288081	1435532	51.8965	25.95144	4415.9326	2427.8760

图 23-16 聚类结果的特征数据

第 3 类的华北电力大学科技学院, 各指标的表现都最为突出, 是基本办学条件最好的学校。

第 5 类的河北工业大学城市学院, 其教师的综合水平较高, 但是反映基本办学条件的硬件措施的生均指标却较低, 改善办学环境是其当务之急。

第 2 类的河北经贸大学经济管理学院、河北大学工商学院和第 3 类的军械工程学院信息与管理分院, 有研究生学历的教师比例为 0, 对它们来说引进人才是提高教学质量的保证。

第 1 类包含了 83.7% 的学校, 其生均用房、生均图书、生均仪器都不高, 而研究生教师比例、生师比都很高, 这显然又是扩招带来的现象。

由此可见, 专科和本科院校都面临着由扩招衍生的一系列办学条件下降的问题, 这是隐藏在“招生繁荣”背后的严重社会问题; 只有相关的基本办学水平提高了, 才能正确配合招生政策的实施, 否则学校收取过多的学生, 却不能很好地为其服务, 不能使他们将来离开学校时达到一个较好的状态和水平, 这样学校的教育水平问题就转嫁给了不断加重的就业压力等一系列社会问题。鉴于此, 研究如何提高高校的教育水平问题刻不容缓。

## 23.4 建议和推广

本章使用聚类分析法对河北省的一百多所高校进行了聚类分析, 由于聚类方法的多样性可以预料, 使用不同的聚类变量得到的聚类结果会有所不同, 所以建议用户根据自己的研究兴趣选择不同的聚类变量, 从而使分析效果更加具有针对性。

另外, 也可以使用其它方法对这一问题进行详细研究, 例如层次分析法等。层次分析法

(Analytic Hierarchy Process, AHP) 是美国运筹学家 T. L. Saaty 于七十年代初提出的一种灵活、简便的多准则方法, 它把一个复杂的问题按一定原则分解为若干子问题, 对每个子问题作同样的处理, 由此就能得到按支配关系形成的多层次结构, 对同一层的各元素进行两两比较, 并用矩阵运算确定出该元素对上一层支配元素的相对重要性, 进而确定出每个子问题对总目标的重要性。

当前我国的经济正以令世人瞩目的速度向前发展, 与之相呼应的高等教育事业也从未像今天这样充满了蓬勃的活力。1999 年以来, 我国高等教育连续几年的大幅度扩大招生, 使高等院校学生数从 1997 年的 607 万增至 2001 年的 1 136 万, 扩大了约 1.9 倍, 其中普通高等院校的学生数翻了一番多, 高等教育毛入学率也由 8.84% 上升至约 14%。这几年的高等教育扩大招生, 一举扭转了我国高等教育规模偏小的局面, 并将迈进大众化高等教育阶段。那么今后跨入大众教育阶段的中国高等教育是否仍然保持这一扩招势头, 抑或有所降低? 国外多国数据显示, 高校的扩招速率与该国的经济增长速度基本上一致。近两年来, 我国的高校扩招幅度是经济增长速度的 2-3 倍。此速度是否适当、当前我国高等院校办学条件是否能够承受? 都是有待研究的问题。

从聚类结果中我们可以清晰的看到扩招对于高校基本办学条件的冲击。高等教育的扩招大大超过了高校师资队伍、教学仪器、图书、教学用房等基本办学条件的建设速度, 这过快的速度必然是以牺牲质量为代价的。通过聚类分析能够反应一些由高校扩招所带来的问题, 所以建议对高等教育基本办学条件的研究应周期进行, 从而辅助教育官员更好的决策。



# 第 24 章 试卷信度的检验与分析

试卷是考试运行的重要载体，其质量的高低不仅直接影响着考试的可靠度和准确度，往往还直接或间接地影响到学生的学习态度和学习行为。

本章运用教育测量学中关于信度的有关知识来进行试卷信度的检验分析，演示了如何利用 SPSS15.0 软件做试卷的信度分析以及如何解释信度检验的结果；通过本章的介绍，各种出题者都能够运用简单、便捷的分析软件和方法对试卷的信度进行定量分析，从而保证利用试卷进行测验的可靠性。

## 24.1 试卷信度检验的背景简介

试卷信度是教育测量及评价中一个重要概念，人们通过计算信度来考察此次试卷能否反映出被测试者的真实水平。试卷是由不同题型、不同难易度的试题组成的，难免会带有一定的主观随意性，并可能造成测试结果的偏差，所以如何确定考试的客观性和可靠性是一个十分重要的问题。影响考试信度的因素有很多，如考试的组织形式、试卷的信度评分是否客观等。要想从根本上提高试卷的信度，就要对影响试卷信度的因素进行深入的研究和讨论，从而更好的利用试卷这个检验工具。

### 24.1.1 测验内容的自身方面

试卷本身的某些因素会直接产生误差，例如试卷的难度越高，其试卷信度就会越低，因为作答时猜测的机率将引起测验的不稳定性，如果大量的题目都很难，需要进行猜测来解答，那么被测者的总分就接近于随机分布了。另外，试卷指导语的清晰度、试卷内容的取样多少以及试卷的长度等因素都会引起误差，从而影响试卷的信度。

### 24.1.2 施测过程

在考试的过程中，考试环境、考试时间以及一些不能预见的干扰等都会产生误差；不同考场采用不同的监考，由于主考官的主观意向，也可能会引起测试效果的误差；关于评分，有的试题意义含混，容易引起考生及评分者之间的歧义，有的考生故意以含糊不清、模棱两可的叙述掩盖知识缺陷，给分数评定带来困难，由此引起评分上的误差。

此外，评分者本人的某些特点也会产生评分误差，例如以下 3 点。

(1) 知识水平，评分者的知识水平会影响对问题的理解和评分标准的把握，尤其当考生答题与众不同具有创造性时，更需要评分者有较高的水平。

(2) 心理状况，评分者在不同的心理状态下，所掌握的评分宽严程度是有差异的。

(3) 评分者评分个性倾向的影响，有的评分宽松、就高不就低，有的扣分过严，也有的坚持中庸，有的追求卷面美观，也有的欣赏逻辑严谨等，都将影响评分的客观性。

24.1.3 被测试者的自身因素

被测者本身的特征也会造成测试分数的偏差，例如某考生和大多数考生的测试动机不相同，就会引起测试误差。应试经验也是引起误差的重要因素，面对测验的焦虑与心理都与测验的经验有关，过度的焦虑对于应试有不良的影响，从而给测验带来误差。当考生在生病或处于疲劳状态下进行的测验，所得的结果会与正常情况不同，有些测验考生可能由于疲劳引起暂时的注意力分散，从而不能作出符合自己日常行为或情感的反应。

综上所述，产生测验分数不可靠的原因有很多，它们都能导致测验信度的降低。

本章对已经实施完成的测试结果进行试卷信度的检验分析，由此判断测试实施的效果以及推断可能引起测试误差的因素。

24.2 数据描述

本章对某学院工商管理班高等数学考试的 30 分试卷进行试卷信度的检验分析。数据格式如表 24-1 所示，所用数据文件为“30 名学生的高数成绩表.sav”。

表 24-1 高等数学考试成绩

学号	第一题	第二题	第三题	第四题	第五题	第六题	总分
1	20	16	12	2	5	0	55
2	8	11	14	8	7	2	50
3	12	18	13	12	3	0	58
.....	.....	.....	.....	.....	.....	.....	.....
30	16	22	21	8	8	0	75

24.3 分析方法概述

信度（reliability）指可靠性或可靠的程度，试卷信度就是指试卷结果的可靠程度。

信度也可以用来指示实测值和真值相差的程度，实测值是对测验对象进行实际测验所获得的测定值，真值是测验对象真实水平的取值。如果实测值与真值相差较小，说明结果的信度较高，反之信度较低。为了能够真实、准确地反映测量对象的实际水平，必须重视对试卷信度的研究，从而正确地判断测量结果的价值。在实际工作中，既可以对测量信度的高低进行定性分析，也可以通过信度系数进行定量的分析，例如用克龙巴赫 $\alpha$ 系数度量结果的一致性程度，再用统计方法检验它是否达到了显著水平。

24.3.1 试卷信度的基本计算公式

试卷信度指的是考试结果的稳定性或可靠程度，即考试的结果是否真实、客观地反映了考生的实际水平。每个考生的考试分数( $X_i$ )通常包含两部分：真实分数( $X_{\infty}$ )和误差( $X_e$ )，即 $X_i = X_{\infty} + X_e$ ；当测量次数足够大时，误差的总和应该为零。对于考试所得分数的方差( $\sigma_e^2$ )，

也可以表示为真实分数的方差( $\sigma_x^2$ )与误差的方差( $\sigma_e^2$ )之和,即 $\sigma_t^2 = \sigma_\infty^2 + \sigma_e^2$ 。

信度在理论上被定义为在一组考试中真实分数方差与所得分数方差之比,即 $r_{tt} = \frac{\sigma_\infty^2}{\sigma_t^2} = 1 - \frac{\sigma_e^2}{\sigma_t^2}$ 。 $r_{tt}$ 称为信度系数,一般认为试卷信度在0.5至0.9以内是合理的。如果研究者的目的在于编制预测问卷或是测验某构思的先导性研究,信度系数在0.5至0.6就足够了;如果以发展测量工具为目的,信度系数应在0.70以上;若以基础研究为目的,信度系数最好在0.8以上;而在进行筛选、分组和是否接受特殊教育等问题的研究上,信度系数最好达到0.9以上。由此可见,试卷信度是衡量一个试卷(或称为量表)质量高低的重要指标,信度不合要求的试卷不建议使用。

### 24.3.2 试卷信度的估计方法

信度是反映测量中随机误差大小的指标,由于造成误差的方式和来源多种多样,所以信度的估计方法也是多种多样。在试卷信度的检验中一般采用的是同质性信度。

同质性信度(homogeneity reliability)也叫内部一致性系数,它是衡量测验内部所有题目间一致性程度的指标。这里题目间的一致性含义有两层意思,其一指所有题目测试的都是同一目标,其二是指所有题目的得分之间都具有较高的正相关。当一个测验具有较高的同质性信度时,说明测验主要测的是某一单个目标,实测结果就是该目标水平的反映;如果一个测验的同质性信度不高,就说明测验结果可能是几个目标的综合反映,这时的测验结果不好解释,这时一种好的解决办法就是把一个多目标的测验分解成多个单目标的分测验,再根据被试者在分测验上的得分对测试结果加以解释。需要注意的是,一些表面上看起来是测量同一目标的题目,如果其题目得分间不具有较高的正相关,则不能认为它们具有同质性,也就是说测量单一目标是同质性高的必要条件,而非充分条件。

下面简单介绍同质性信度几个常用的估计公式。

内部一致性系数的一种粗略估计方法是求测验的分半信度。因为分半方法很多,所得结果不太稳定,因此有人建议计算出所有可能的分半信度,并用其平均值来作为内部一致性估计,但这种办法太麻烦了,因为所有可能的分半信度个数简直是个天文数字。于是又提出了新的计算公式 $r_{tt} = K\bar{r}_{ij} / [1 + (K-1)\bar{r}_{ij}]$ ,其中K为一个测试的题目个数, $\bar{r}_{ij}$ 为所有题目间的相关系数的平均值。这个新公式也并不方便,因为对所有的题目求相关系数同样比较麻烦,不过由此导出了十分方便的库—理信度系数和克龙巴赫 $\alpha$ 系数。

#### 1. $KR_{20}$ 公式

$r_{tt} = [K / (K-1)][1 - (\sum p_i q_i) / S_x^2]$ ,其中K是题目数, $p_i$ 为答对第*i*道题的人数比例, $q_i$ 为答错第*i*道题的人数比例; $S_x^2$ 为测验总分的变异,公式 $S_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_j - \bar{x})^2$ , $x_j$ 表示每个测试者的测验总分数, $\bar{x}$ 为总得分的平均值。此公式是由库德和理查德逊于1937年提出的,它仅适用于(0、1)记分的对错是非题,也就是二分化计分的测验。

#### 2. $KR_{21}$ 公式

$r_{tt} = [K / (K-1)][1 - (K\bar{p}\bar{q}) / S_x^2]$ ,式中各指标含义与 $KR_{20}$ 相同,只是 $\bar{p}$ 与 $\bar{q}$ 分别表示题目

的平均通过率和失败率，此公式只有当所有题目的难度接近时才适用。

### 3. 克龙巴赫 $\alpha$ 系数


$\alpha = [K / (K - 1)] [1 - (\sum S_i^2) / S_x^2]$ ，其中  $S_i^2$  表示所有被试在第  $i$  道题上的分数变异，公式  $S_i^2 = \frac{1}{n-1} \sum_{j=1}^n (x_{ji} - \bar{x}_i)^2$ ， $x_{ji}$  表示第  $j$  个学生第  $i$  道题的得分， $\bar{x}_i$  表示第  $i$  道题的平均分数；其余指标的含义与  $KR_{20}$  相同。此公式是克龙巴赫在 1951 年提出的，它可以计算任何测验的内部一致性系数，而不要求测验题目必须是记分型的。实际上， $KR_{20}$  和  $KR_{21}$  只是  $\alpha$  系数的特例，比如在 (0、1) 记分时有  $\sum S_i^2 = \sum p_i q_i$ 。

$\alpha$  值越大必有测量信度越高，但  $\alpha$  值越小时却不能断定测量信度不高。

$\alpha$  值的计算一般按下述步骤进行：按一定要求抽取  $n$  个受试者的试卷，首先计算  $n$  个测验总分的方差  $S_x^2$ ；对于每一道题，求出这  $n$  个人在其上所得分数的方差  $S_i^2$  ( $i=1, 2, 3, \dots$ )；求出  $\sum S_i^2$  的值；按公式求  $\alpha$  系数的值。

## 24.4 SPSS 建模过程和结论分析

### 24.4.1 SPSS 信度分析的参数设置

打开文件“30 名学生的高数成绩表.sav”，依次单击菜单“Analyze→Scale→Reliability Analysis...”执行信度分析过程，其主设置界面如图 24-1 所示。在变量列表中选中从第 1 题得分 (s1) 至第 6 题得分 (s6) 的 6 个变量，单击  按钮将其作为分析变量选入 Items 列表；在 Model 栏后的下拉列表选采用默认的 Alpha 方法 (Cronbach  $\alpha$  系数)。

在图 24-1 中单击 Statistics 按钮，弹出如图 24-2 所示的统计量选择子设置界面。依次勾选如下复选框：Descriptives for 栏下的 Item、Scale、Scale if item deleted，Summaries 栏下的 Means、Variances、Covariance、Correlations，Inter-Item 栏的 Correlations；单击选中 ANOVA Table 栏下的 F test 单选框；勾选 Intraclass correlation coefficient 复选框，并保留此栏的默认方法 Two-Way Mixed 及其默认参数。单击 Continue 按钮返回主界面。

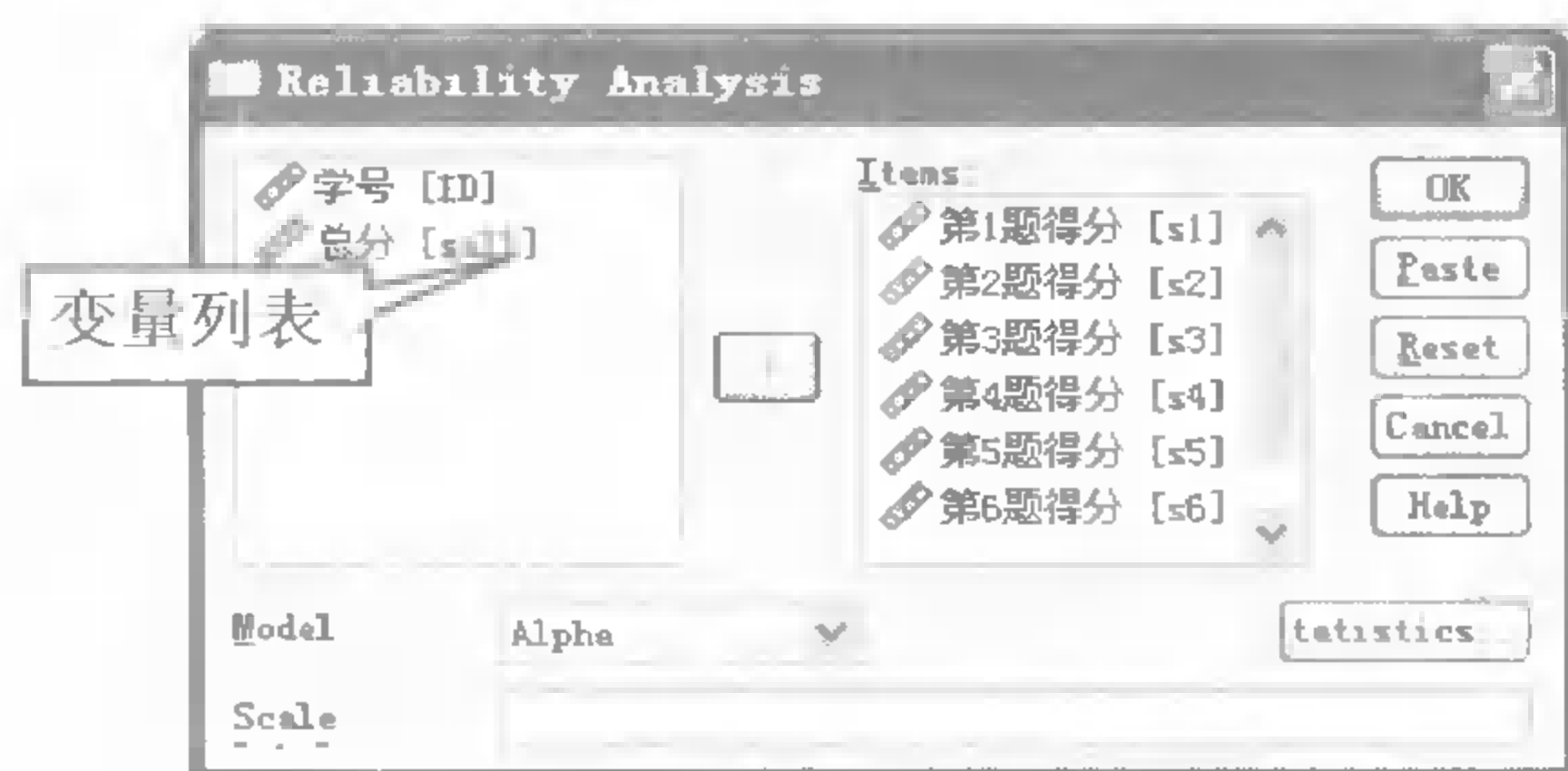


图 24-1 信度分析的主界面

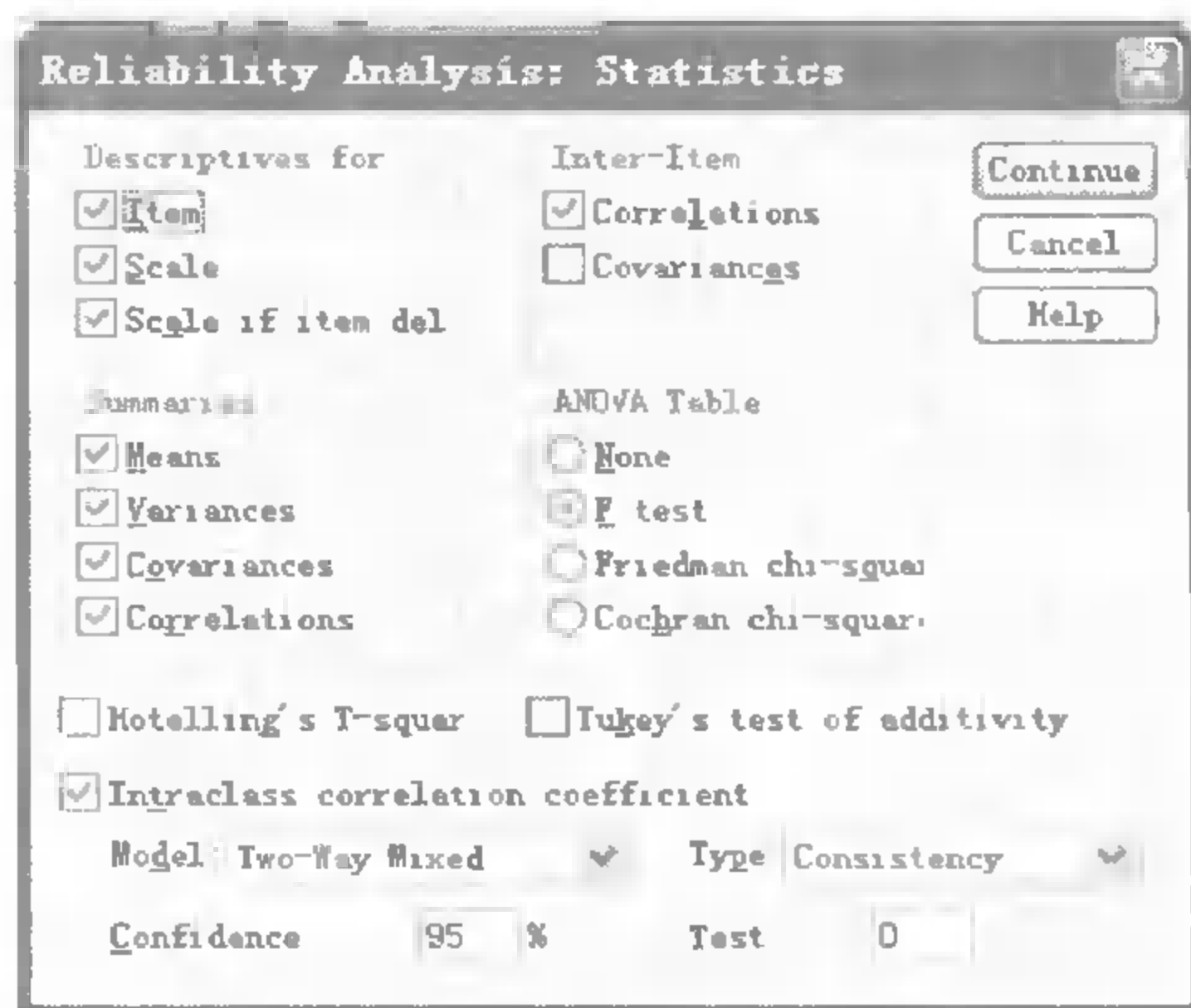


图 24-2 信度分析的参数设置



24.4.2 结果分析

在图 24-1 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出结果如下。

1. 可靠性统计量和项统计量输出

如图 24-3 所示，“摘要”表格给出了数据中有关缺失值的统计信息，本例的 30 个观测没有缺失，都用于分析。“可靠性统计量”表格给出了 Cronbach  $\alpha$  系数的计算结果，表中的 0.703 是对真实  $\alpha$  系数的估计（下界），由此判断利用此试卷所进行的测试结果可信度还是不错的。

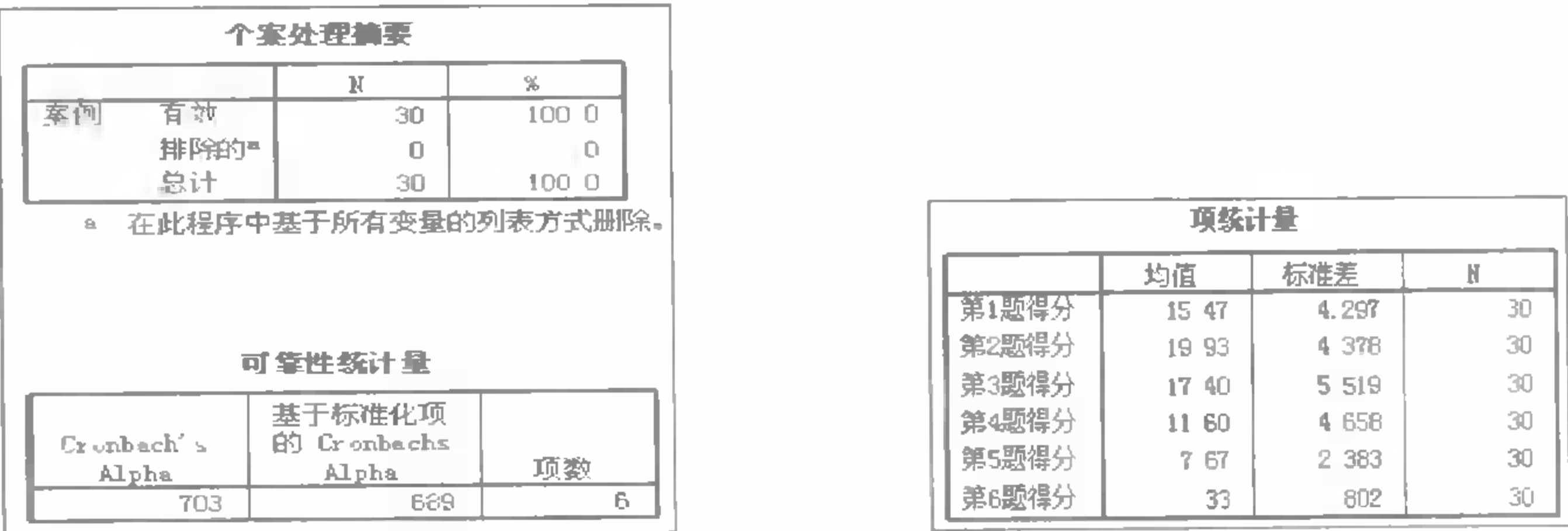


图 24-3 6 个题目得分的可靠性统计量和项统计量输出

“项统计量”表格给出了单个题目得分（变量）的基本统计量，可见第 2 题的平均得分最高，第 6 题的平均得分最低。

2. 各题目得分间的相关矩阵

如图 24-4 所示，“矩阵”表格给出的是各个问题得分间的相关矩阵。

各题目间的相关性不是很明显，对于提高试卷的信度是不利的，所以建议参考其它的分析结果（比如方差分析表、类内相关性等）后再下结论。

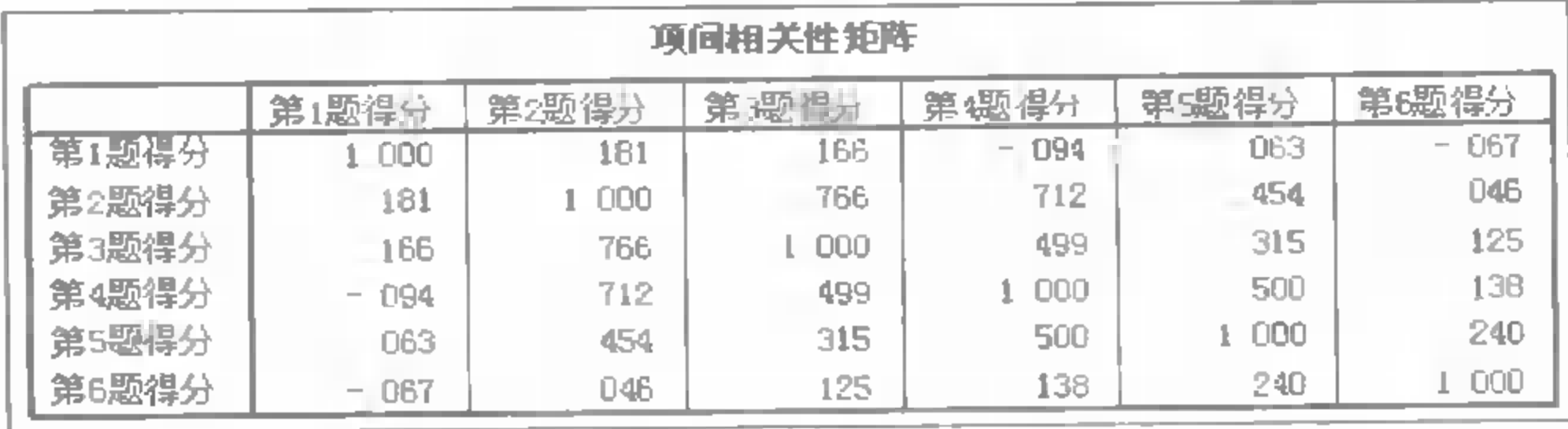


图 24-4 各题目得分间的相关矩阵

3. 其他输出

图 24-5 是信度分析过程的其他输出。

“标度统计量”表格给出的是 6 个题目总分的描述信息，30 个学生总分的平均分为 72.4。ANOVA 方差分析表给出了 6 个题目的方差分布及 F 检验信息。这里的 F 统计量指的是

$$F = \frac{Q_1^2 / (m-1)}{Q_2^2 / m(n-1)} = \frac{m(n-1)}{m} \times \frac{Q_1^2}{Q_2^2}, \text{ 其中 } Q_1^2 = \sum_{i=1}^6 \sum_{j=1}^{30} (\bar{x}_i - \bar{x})^2 \text{ 表示考试中由学生水平差异造成的}$$

误差， $Q_2^2 = \sum_{i=1}^6 \sum_{j=1}^{30} (x_{ij} - \bar{x}_i)^2$  表示考试中由偶然因素引起的随机误差； $x_{ij}$  表示第 i 号学生第 j

个题目的得分， $\bar{x}_i$  表示第  $i$  号学生的平均得分， $\bar{x}$  表示所有学生所有题目得分的平均值。由此可见， $F$  统计量取值越大，考试所反映的学生真实水平的差异就越可信，试卷的可靠性也就越好。本例的  $F=135.28$ ，且其显著性水平远小于 0.01，故而  $F$  统计量的取值是显著的大，由此推断本试卷的可信度不错。

信度统计量							
		均值	方差	标准差	项数		
		72.40	231.834	15.226	6		

ANOVA					
		平方和	df	均方	Sig.
人员之间		1120.533	29	38.639	
人员内部	项之间	7774.133	5	1554.827	.000
	残差	1666.533	145	11.493	
	总计	9440.667	150	62.938	
总计		10561.200	179	59.001	

总均值 = 12.07

类内相关系数							
	类内相关性 <sup>a</sup>	95% 置信区间		使用真值 0 的 F 检验			
		下限	上限	值	df1	df2	Sig.
单个测量	.282 <sup>b</sup>	.143	.469	3.362	29.0	145	.000
平均测量	.703 <sup>c</sup>	.501	.841	3.362	29.0	145	.000

双向混合效应模型，其中，人员影响是随机的而测量影响是固定的。

a. C 型类内相关系数使用一致性定义 - 从分母方差中排除之间测量方差。

b. 无论是否存在交互效果，估计器都相同。

c. 计算此估计时假设交互效果不存，否则就不可估计。

图 24-5 信度分析的部分输出

在“类内相关系数”表格里，平均测量的类内相关性的 95% 置信区间为 0.50~0.84，它用来衡量用平均分（总分）对学生进行评价的一致性程度，比如说好学生的总得分应该高于平均的总得分，此指标的取值越高试卷的可信度越高。由单个题目所得的类内相关性（0.28），要低于用平均分（总分）所得的类内相关度（0.70），这是符合常理的。

## 24.5 建议和推广

有了测试试卷信度的方法，如果测出试卷的信度不高，应该如何改进呢？下面给出几条常用的提高试卷信度的方法，仅供参考。

（1）适当增加试题的数量。由于试卷题量太小会降低测量的信度，因而提高测量信度的一个常用的方法就是增加一些与原测验中的题目具有较好同质性的题目，增大测验的长度。

但是有两点必须注意：新增的题目必须与试卷中原有的题目具有同质性，即测试目标保持一致；新增项目的数量必须适度，事实上增加测验长度的效果是符合报酬递减规律的，即测验的数量和时间过长可能引起被试者的疲劳和反感，从而降低测量信度。

（2）保持所有试题的难度程度接近正态分布。当测验中所有试题的难度接近正态分布并控制在中等难度水平时，被试团体的得分分布也会接近正态分布，以相关性为基础的信度值也会增大。

（3）努力提高测验试题的区分度。区分度是测验题目的质量指标，试题的区分度高低会直接影响测验的信度，努力提高测验中所有试题的区分度，可望获取较高的测验信度。

（4）监考和评分。主考官要严格执行实测规程；评分者严格按标准给分。

# 第 25 章 多因素试验的设计与分析

在科学研究和工农业生产中经常需要通过试验来寻找所要研究对象的变化规律，并通过对规律的研究达到各种试验目的，如提高产量、降低消耗、提高产品性能或质量等，特别是新产品试验，未知的东西很多，需要大量的试验来摸索工艺条件和配方。只有科学的、合理的试验设计才能用较少的试验次数、在较短的时间内达到预期的试验目的。

试验设计在日常生产生活中有着广泛的应用，其中以正交试验设计较为常用。在正交设计中，对单指标试验用直观分析法便可简便、快速地得出结果。而在多指标的试验中，不同指标的重要程度常常是不一致的，各因素对不同指标的影响程度也不完全相同，所以多指标试验的结果分析比较复杂，方法有综合评分法、综合平衡法、综合比较法和公式法等。

## 25.1 试验设计简介

试验是在人为控制条件下进行的一种实践活动，例如比较几种产品的好坏、几种生化流程的差异和几种药剂的疗效等。一个试验通常包括输入（因素或处理）、供试体、输出（因变量）和无法避免的随机干扰（也是一种输出）等要素，输入的变化通过供试体作用转换为输出的变化，这种变化规律正是做试验所要研究的。

试验可以比作是机器、方法、人以及其他资源的一种组合，它把一些输入转变为可观察的响应输出，关于试验的一些变量  $x_1, x_2, \dots, x_p$  是可控制的，另一些变量  $z_1, z_2, \dots, z_q$  是不可控制的，试验目的可分为如下 4 个方面。

- 确定哪些变量  $x$  对输出响应  $y$  的影响最大。
- 确定因素变量  $x$  取值多少时，使得响应变量  $y$  接近于所希望的指定值。
- 确定因素变量  $x$  取值多少时，使得响应变量  $y$  的变异性较小。
- 确定因素变量  $x$  取值多少时，使得不可控制变量  $z$  的效应最小。

所谓试验设计，就是为了更好地完成研究目的，在试验前审慎作出的一个试验实施方案，它要保证试验所得数据适合于用数理统计方法分析，并能得出有效而客观的结论。

### 25.1.1 试验设计的应用

试验设计在很多学科中都有广泛的应用，它是探讨系统或过程如何工作的一种有效途径。

试验设计在工程学领域是改进制造过程性能的重要手段。在新工序开发早期，应用试验设计方法大有裨益，它可以提高产量；减少变异性，使实验结果与额定值或目标值更为一致；减少开发时间；减少总成本等。试验设计在工程设计中的应用包括评价和比较基本的设计结

构；材料选择的评定：选择设计参数，使产品能在更广泛的条件下运行良好，即使产品是稳健的；确定影响产品性能的关键参数等。

### 25.1.2 试验设计问题的解决步骤

要想完整地解决一个试验设计问题，可以参考如下的步骤进行操作。

#### 1. 待研究问题的认识与描述

对于所要研究问题的认识和描述，主要包括如下 3 个方面。

① 因变量的选取。选定的因变量不仅要与试验目的相一致，而且其分布还应该与试验统计模型的假定一致或非常近似。还要区分因变量是定性的还是定量的，并决定测量因变量的方法，了解测量的精度。

② 因子（自变量）的选取。首先要将对因变量有影响的因素尽量罗列出来；然后根据试验目的将那些对因变量影响很小的因素作为误差因子，在试验中不必控制而任其随机变化；将那些对因变量影响较大而又不准备考察其影响的因素，在试验中对它们加以控制或制订一种试验计划保证在分析试验结果时可以消除它们的影响，就是不把它们作为试验问题中的因子；只有那些对因变量影响较大又希望通过试验对其加以研究的因素才当作试验问题中的因子。因子个数对试验次数有直接影响，选取因子时也要顾及试验次数是否承受得了。

③ 各因子的取值水平。确定各因子的取值水平，既要根据试验目的和实践经验，又要注意因子水平的个数对总试验次数的影响，综合考虑问题需要、经费、人力、时间的许可来确定各因子的水平。

因子是定性的还是定量的，如果是定量的，还要考虑在试验中如何控制它们的水平，以及控制的精度如何。因子是固定的还是随机的，如果因子的水平是按试验人的主观意图选定的，则称该因子是固定的；如果因子的水平是随机确定的，则称该因子是随机的，一个因子是随机还是固定的要由试验目的来确定。

#### 2. 试验计划的设计与实施

关于试验计划的设计与实施，包括如下一些方面需要考虑。

① 确定总试验次数，在包含尽可能多的信息、保证试验结果的统计精度等前提下，使试验次数尽量地少。

② 各因子诸水平的组合，明确试验计划中必须包含哪些水平组合，有时还要注意避免那些试验中不能实施的水平组合。

③ 试验顺序的安排，可以是随机化的方法，也可以人为指定试验顺序。

④ 如何建立统计分析模型。

⑤ 实施试验，在此过程中要严格监控试验计划的要求得到实现，并准确地记录试验结果。

#### 3. 试验结果的统计分析

① 搜集试验数据，并作适当整理和备份。

② 计算假设检验中的统计量以及模型中各参数的估计量。

③ 对统计分析的结果作出科学而符合实际的解释，提出合理建议。



## 25.2 数据描述

本章对一种湿地推土机的性能加以研究，通过试验设计的方法，希望找出 3 个影响因素各取什么水平时推土机能够达到最好的性能。

已知影响湿地推土机性能的因素有 3 个：接地压力、履带板型式和整机重心位置，它们各自的取值水平如表 25-1 所示。

表 25-1 因素水平表

水 平	A 履带板型式	B 接地压力（比压/kPa）	C 重心
1	无间隔	18	中点前 120mm
2	间隔大	21	中点
3	间隔小	23	中点后 120mm

试验要求记录如下 3 个指标来衡量推土机的性能：行走阻力（10N）、滑转率（%）和下陷深度（mm），并且这 3 个指标的取值越小越好。可见这是一个多因素多指标的试验。

该试验为 3 因素 3 水平试验，故选取  $L_9(3^3)$  正交表，试验方案及试验结果如表 25-2 所示。

表 25-2 试验方案和试验结果

试 验 号	因 素			指 标 值		
	A	B	C	行走阻力 (10N)	滑转率%	下陷深度 (mm)
1	1	1	1	638	4.1	8
2	1	2	2	632	3.3	10.7
3	1	3	3	816	9.1	10.6
4	2	1	2	861	5.5	10.3
5	2	2	3	838	9.4	15.5
6	2	3	1	773	6.5	14
7	3	1	3	627	2.3	10.6
8	3	2	1	615	4.4	11.8
9	3	3	2	632	5.8	12.5

## 25.3 分析方法概述

### 25.3.1 正交设计方法

正交试验设计简称正交设计，是利用标准化正交表科学地安排与分析多因素试验的方法，是最常用的试验设计方法之一。

在工业生产和科学研究的实践中所要考察的因素往往很多，而且因素水平数也常常多于两个，如果对每个因素的每个水平都相互搭配进行全面试验，要做的试验次数就会很多。例如对 3 因素 4 水平的试验，进行全面试验需要做  $4^3=64$  次试验，对 4 因素 4

水平的试验，全面试验的次数为  $4^4 = 256$  次，对 5 因素 4 水平的试验，全面试验的次数为  $4^5 = 1\,024$  次。随着因素数的增加，试验次数呈指数级递增，需要花费大量的人力、物力和时间。人们在长期的实践中发现，要得到理想的结果并不一定需要进行全面试验，在不影响试验效果的前提下，应该尽可能的减少试验次数，正交设计就是解决这个问题的有效方法。

正交表是一种特殊的规格化表格，它是正交设计中安排试验和分析试验结果的基本工具。一般的正交表记为  $L_n(m^k)$ ，其中“ $L$ ”表示正交表； $n$  表示行数，也就是要安排的试验次数； $k$  是表的列数，表示因素的个数； $m$  是各因素的水平数。例如表格 25-3 所示的就是  $L_9(3^4)$  正交表。

表 25-3

正交表样例

试 验 号	列 号			
	1	2	3	4
1	1	1	1	1
2	1	2	2	2
3	1	3	3	3
4	2	1	2	3
5	2	2	3	1
6	2	3	1	2
7	3	1	3	2
8	3	2	1	3
9	3	3	2	1

正交表根据各因素的水平数又可分为如下两类。

### 1. 等水平正交表

所谓等水平正交表，就是指各因素的水平数是相等的正交表，常见的有  $L_4(2^3)$ ,  $L_8(2^7)$ ,  $L_{16}(2^{15})$ ,  $\dots$ ,  $L_9(3^4)$ ,  $L_{27}(3^{13})$  等。它们的特点是表的任一系列中，不同的数字出现的次数相同；表中任意两列，把同一行的两个数字看成有序数字对时，所有可能的数字对（水平搭配）出现的次数相同。这两个性质合称为“正交性”，它能使试验点在试验范围内排列整齐、规律，并且散布均匀，即“整齐可比、均衡分散”。

### 2. 混合水平正交表

混合水平正交表是各因素的水平数不完全相同的正交表，常见的混合水平正交表  $L_8(4^1 \times 2^4)$ ,  $L_{12}(3^1 \times 2^4)$ ,  $L_{12}(6^1 \times 2^2)$ ,  $L_{16}(4^1 \times 2^{12})$  等。它的特点是表中任一系列中，不同的数字出现的次数相同；每两列中，同行两个数字组成不同水平搭配的出现次数也是相同的；但是，不同两列间所组成的水平搭配种类及出现次数是不完全相同的。由此可见，用混合水平的正交表安排试验时，每个因素各水平之间的搭配也是均衡的。

下面简单总结一下利用正交表安排试验并分析试验结果的步骤。

(1) 明确试验目的，确定要考核的试验目标。

- (2) 根据试验目的, 确定要考察的因素和各因素的水平。
- (3) 选择合适的正交表, 安排试验计划。
- (4) 明确试验方案, 进行试验, 测定各试验指标。
- (5) 对试验结果进行统计分析, 得出合理的结论。

### 25.3.2 综合评分方法

在实际生产和科学试验中, 整个试验结果的好坏往往不是一个指标能全面评判的, 所以多指标的试验设计是一类很常见的方法。由于不同指标的重要程度常常是不一致的, 各因素对不同指标的影响程度也不完全相同, 所以多指标试验的结果分析比较复杂。解决多指标试验的主要方法有综合平衡法和综合评分法。

综合平衡法是先对每个指标分别进行单指标的直观分析, 得到每个指标影响因素的主次顺序和最佳水平组合, 然后根据理论知识和实际经验, 对各指标的分析结果进行比较和分析, 使得各指标的最优生产条件相平衡, 得出较优的生产方案。

综合评分法是根据各个指标的重要程度, 对得出的试验结果进行分析, 给每个试验评出一个综合分数, 再根据这个总指标(分数)确定较好的试验方案, 此方法的关键是如何评分。随后介绍 3 种评分方法。

(1) 由研究者对每号试验结果的各个指标人为地加以权衡, 进行综合评价, 直接给出每个试验结果的综合分数。

(2) 先对每个试验的每个指标按一定的评分标准评出分数, 若各指标的重要性是一样的, 可以将同一个试验中各指标的分数总和作为该号试验的总分数。

(3) 先对每个试验的每个指标按一定的评分标准评出分数, 若各指标的重要性不相同, 先确定各指标相对重要性的权重, 然后求加权和作为该号试验的总分数。

目前应用较广泛的是综合加权评分法, 它根据各个试验指标在整个试验中的重要性来确定其权重, 根据权重把多个指标汇总成为一个综合指标, 从而将多指标的试验问题转化为单指标试验问题, 然后按单指标分析方法对方案进行综合选优。于是, 在多指标试验的综合加权评分法中确定各项试验指标的权重是首要也是关键的环节, 对各指标赋权的合理与否直接关系到结论的可靠性, 指标权重的确定方法主要有如下两种。

(1) 主观赋权法, 由分析人员根据各项试验指标的主观重视程度而赋权的一类方法, 常用的有专家调查法、循环打分法、二项系数法、层次分析法等, 这些都是基于对各项指标重要性的主观认知程度, 免不了带有一定程度的主观随意性, 优点是专家可以根据实际问题, 合理确定各指标权系数之间的排序, 防止出现指标系数和指标实际重要程度相悖的情况。

(2) 客观赋权法, 利用试验指标值所反映的客观信息确定权重的一种方法, 主要有变异值法(如标准差、方差或平均值等)、熵值法等, 缺点是有时确定的权系数与实际情况相悖, 优点是有效的克服了权系数确定时的主观随意性。

## 25.4 SPSS 建模过程和结论分析


本节先对衡量推土机性能的 3 个指标, 利用综合加权评分法求出一个综合指标, 然后通过这个综合指标来判断最佳的影响因素参数组合。

### 25.4.1 数据标准化

打开文件“湿地推土机的多因素试验设计.sav”，数据格式如图 25-1所示。由于 3 个测量指标的度量单位不一样，而且需要通过它们来计算一个综合指标，所以在进行下一步分析之前，要先对它们进行标准化处理。

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure
1	card	Numeric	8	0	试验编号	None	None	8	Right	Nominal
2	A	Numeric	8	0	履带板型式	{1, 无间隔}...	None	6	Right	Nominal
3	B	Numeric	8	0	接地压力	{1, 18kPa}...	None	6	Right	Nominal
4	C	Numeric	8	0	重心	{1, 中点前120m}	None	4	Right	Nominal
5	x1	Numeric	8	1	行走阻力(10N)	{1.0, Design}	None	7	Right	Scale
6	x2	Numeric	8	1	滑转率%	None	None	5	Right	Scale
7	x3	Numeric	8	1	下陷深度(mm)	None	None	5	Right	Scale

图 25-1 关于湿地推土机的试验数据

依次单击菜单“Analyze→Descriptive Statistics→Descriptives...”执行描述性统计分析过程，其设置界面如图 25-2 所示。在变量列表中选中 x1、x2、x3 变量，单击  按钮将其选入 Variable(s) 列表作为分析变量。单击 Options 按钮，弹出如图 25-3 所示的统计量选择界面，依次单击选中如下统计量：Mean（均值）、Std（标准差）、Minimum（最小值）、Maximum（最大值）、Range（区间）。单击 Continue 按钮返回主界面。

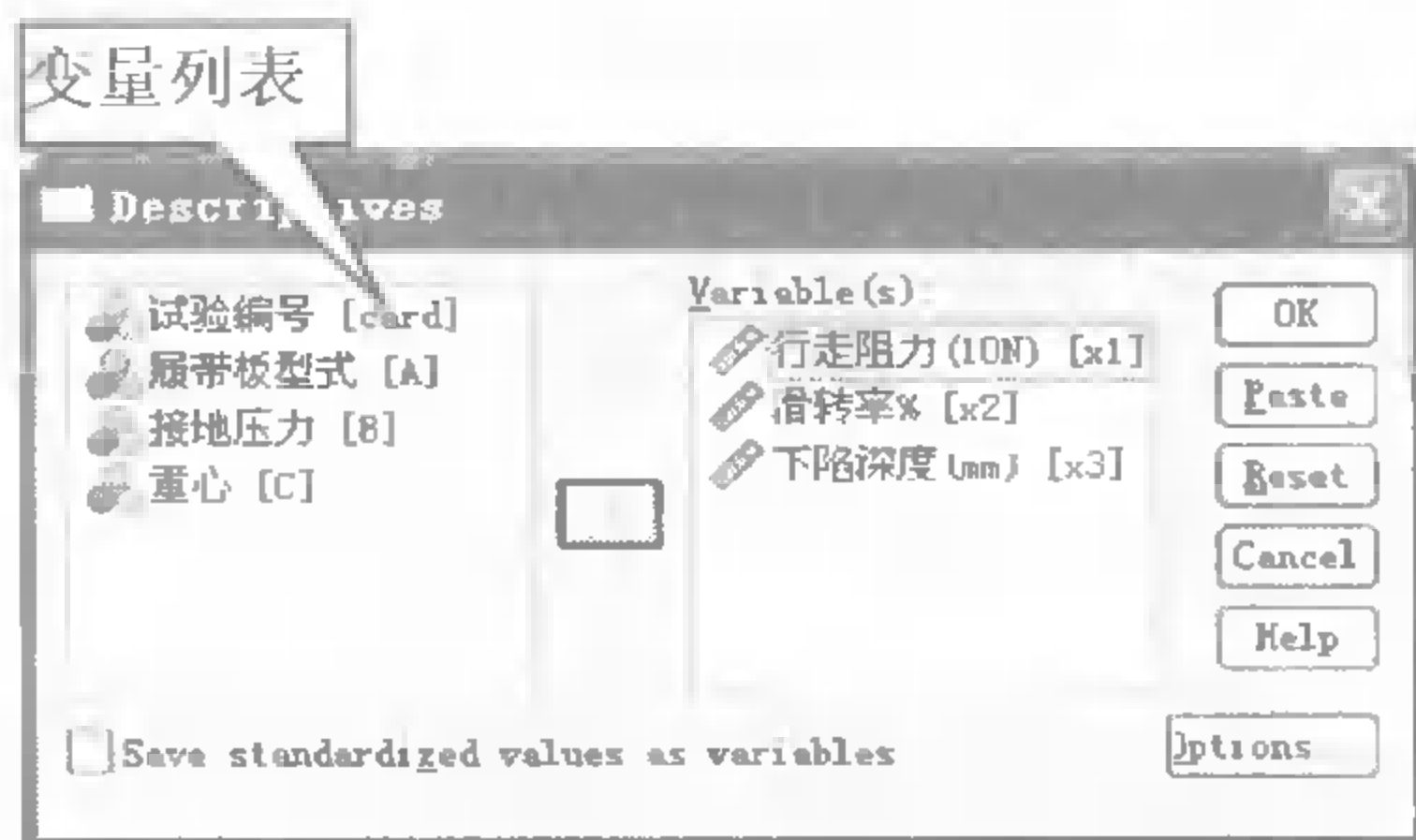


图 25-2 描述性统计分析的主界面



图 25-3 描述性统计分析的参数设置

在图 25-2 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出表格如图 25-4所示。

描述统计量						
	N	全距	极小值	极大值	均值	标准差
行走阻力(10N)	9	246.0	615.0	861.0	714.667	104.5586
滑转率%	9	7.1	2.3	9.4	5.600	2.4346
下陷深度(mm)	9	7.5	8.0	15.5	11.556	2.2154
有效的 N (列表状态)	9					

图 25-4 原始数据的描述性输出

下面就利用此处输出的均值和全距两个指标来对原始数据进行区间标准化，也就是用原始值减去平均值，再除以全距，这样就把原始数据归一化到区间-1~1了。由表中数据可得，对 3 个变量的标准化公式分别为  $Zx1 = (x1 - 714.67) / 246$ ； $Zx2 = (x2 - 5.6) / 7.1$ ； $Zx3 = (x3 - 11.56) / 7.5$ 。其中  $Zx1$ 、 $Zx2$ 、 $Zx3$  分别为  $x1$ 、 $x2$ 、 $x3$  标准化后的变量名称。

下面以变量 x1 为例，说明如何进行标准化处理。

依次单击菜单“Transform→Compute Variable...”执行计算新变量的过程，其设置界面如图 25-5所示，在 Target Variable 栏输入新变量名 Zx1，在右侧的 Numeric Expression 公式编辑框输入新变量的计算公式  $(x1 - 714.67) / 246$ 。单击 OK 按钮运行，在当前数据集就会生成对



$x_1$  标准化后的变量  $Zx_1$ 。采用同样的方法可以得到  $x_2$ 、 $x_3$  标准化后的变量  $Zx_2$ 、 $Zx_3$ 。

随后将只使用标准化后的  $Zx_1$ 、 $Zx_2$ 、 $Zx_3$  作为分析变量，而不再使用变量  $x_1$ 、 $x_2$ 、 $x_3$  了。

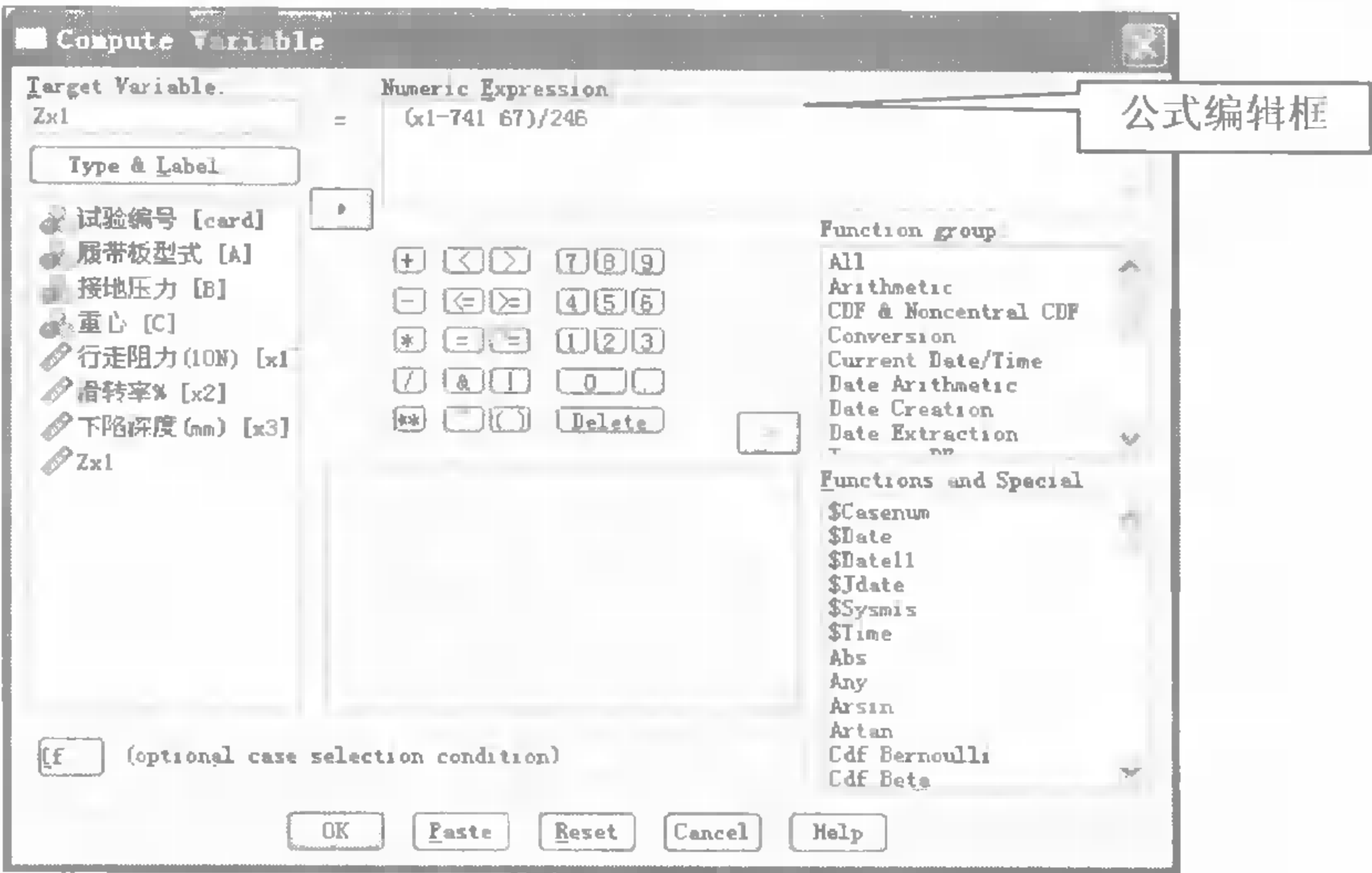


图 25-5 计算新变量的设置

### 25.4.2 性能指标权重的确定

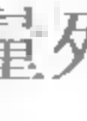
下面先采用两种方法分别确定 3 个指标的权重，然后综合考虑二者的结果，得到一个调整后的最优权重。

#### 1. 专家经验法

在该试验的 3 个指标中，根据实践经验采用专家测评法，得到关于 3 项指标的主观权重，分别为行走阻力 0.50，滑转率 0.30，下陷深度 0.20。

#### 2. 变异比重法

通过比较 3 个指标各自的变异（方差）水平，来反映它们所代表信息量的大小，并以此来推导其权重值。

依次单击菜单“Analyze→Descriptive Statistics→Descriptives...”执行描述性统计分析，其主界面如图 25-6 所示，在变量列表中选中  $Zx_1$ 、 $Zx_2$ 、 $Zx_3$  变量，单击  按钮将其选入 Variable(s) 列表作为分析变量。单击 OK 按钮运行，SPSS Viewer 窗口的输出表格如图 25-7 所示。

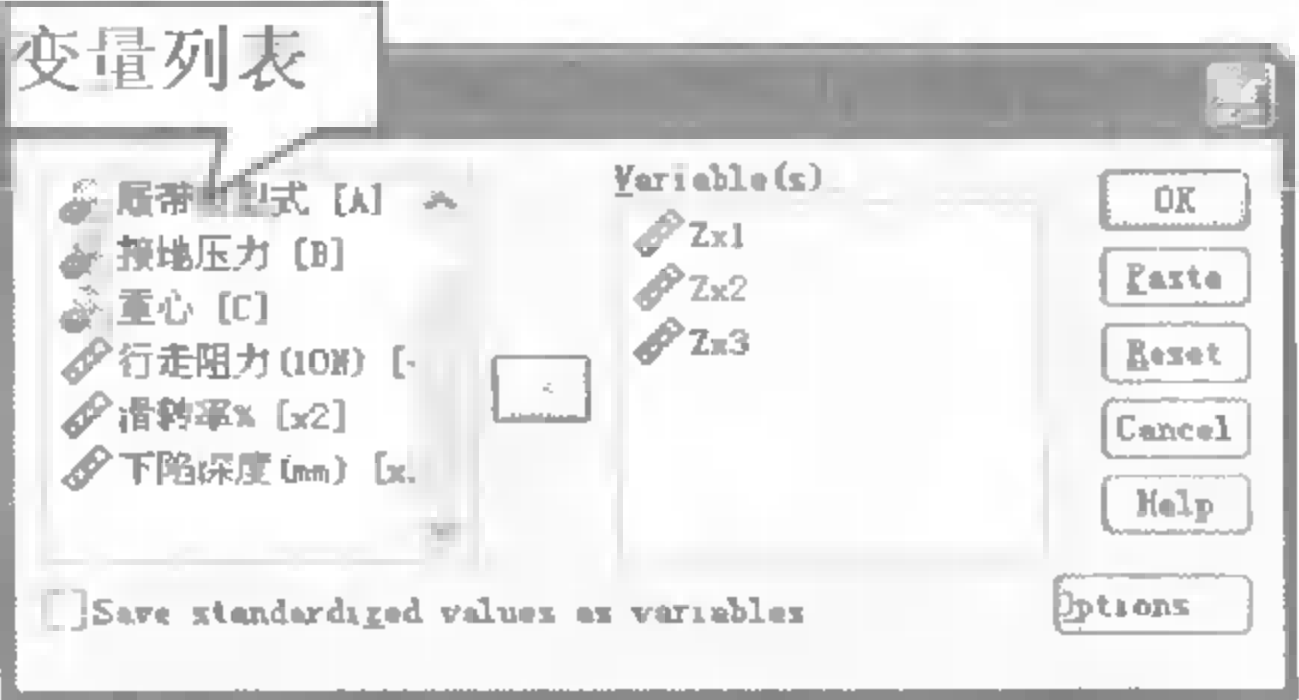


图 25-6 对标准化后的变量进行描述性分析

描述统计量						
	N	全距	极小值	极大值	均值	标准差
Zx1	9	1.00	-.51	.49	-.1098	.42503
Zx2	9	1.00	-.46	.54	.0000	.34291
Zx3	9	1.00	-.47	.53	-.0006	.29538
有效的 N (列表状态)	9					

图 25-7 标准化后的变量描述

图 25-7 显示， $Zx_1$ 、 $Zx_2$ 、 $Zx_3$  的标准差分别为 0.43、0.34、0.30，三者之和为  $0.43 + 0.34 + 0.30 = 1.07$ ，不等于 1。为使权重之和等于 1，由此推导各性能指标的权重分别如下：行走

阻力为  $0.43/1.07 = 0.40$ ，滑转率为  $0.34/1.07 = 0.32$ ，下陷深度为  $0.30/1.07 = 0.28$ 。

### 3. 综合权重

综合前面两种对权重的研究，取其平均值作为最终权重，行走阻力的权重为  $(0.50 + 0.40)/2 = 0.45$ ，滑转率的权重为  $(0.30 + 0.32)/2 = 0.31$ ，滑转率的权重为  $(0.20 + 0.28)/2 = 0.24$ 。


#### 25.4.3 利用权重求综合指标

有了权重，就可以求关于 3 个性能指标的综合评分了，计算公式为  $\text{Score} = 0.45 \times Zx1 + 0.31 \times Zx2 + 0.24 \times Zx3$ 。

依次单击菜单“Transform→Compute Variable...”执行计算新变量的过程，其主设置界面如图 25-5 所示，在 Target Variable 栏输入新变量名 Score，在右侧的 Numeric Expression 公式编辑框输入新变量的计算公式  $0.45 \times Zx1 + 0.31 \times Zx2 + 0.24 \times Zx3$ 。单击 OK 按钮运行，在当前数据集就会生成代表综合得分的变量 Score。

#### 25.4.4 对综合得分的进一步分析

有了综合得分 Score 后，通过对各因素不同水平的平均得分加以研究，就可以判断何种因素水平的组合会产生较好的性能了。下面以图形化的方式对这个问题加以研究。

依次单击菜单“Graphs→Chart Builder...”执行作图过程，其主设置界面如图 25-8 所示。在 Choose from 列表选中 Line（线形图），然后双击右侧的  图标（简单线形图），此时上面的绘图区会出现包括横、纵轴的预览图形；在左侧的变量列表，把履带板型式变量拖动至绘图区的横轴，把得分变量（Score）拖动至绘图区的纵轴。

在图 25-8 中，单击左下方的 Element Properties 按钮，弹出如图 25-9 所示的元素属性面板，在 Statistic 下拉列表（纵轴统计量）保留默认选项 Mean（均值）。

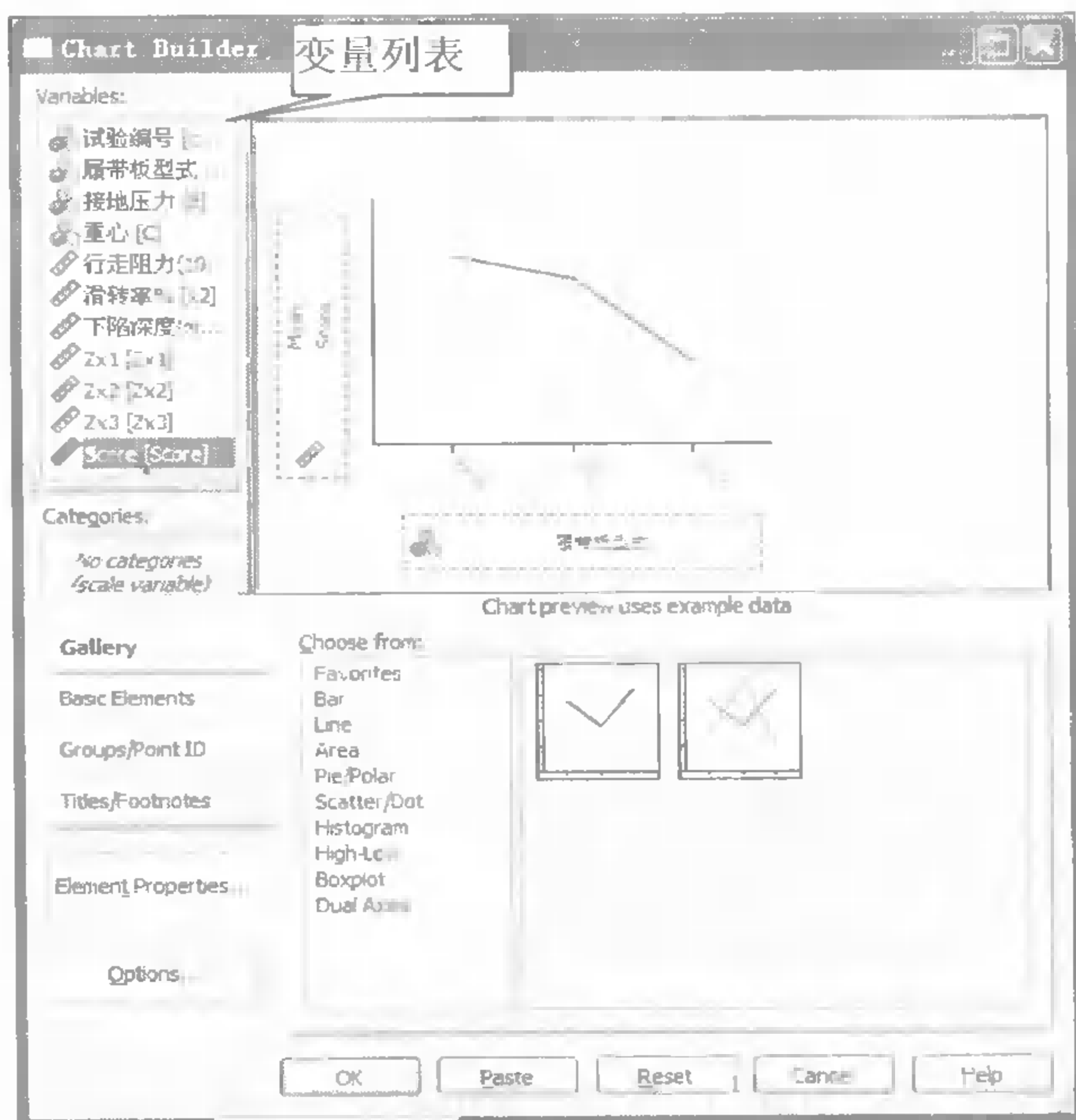


图 25-8 Chart Builder 的作图界面

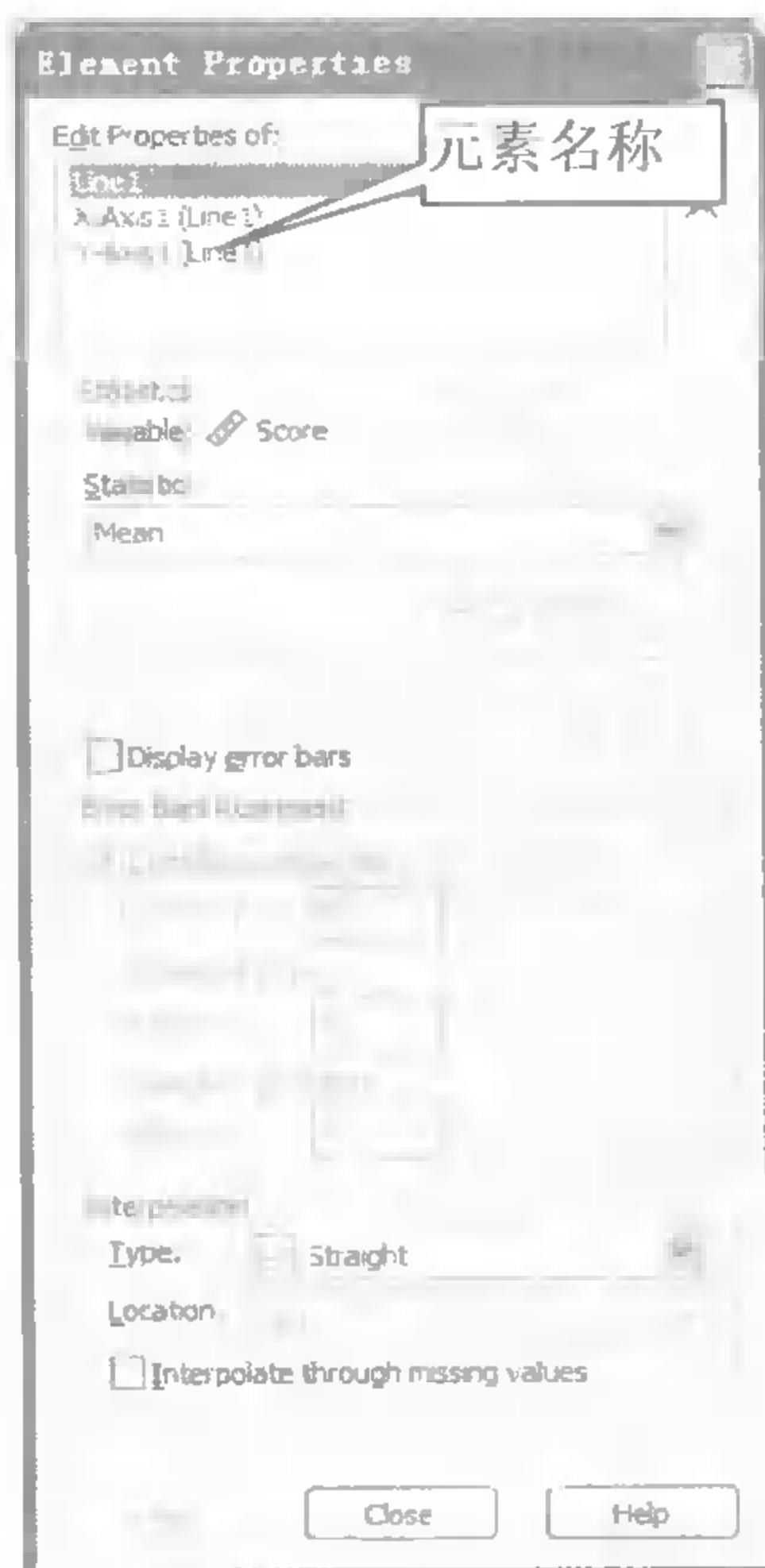


图 25-9 线形图的参数设置

在图 25-8 中，单击 OK 按钮运行，SPSS Viewer 窗口的输出图形如图 25-10 中的得分均值对履带板型式线形图所示。采用同样的方法绘制得分均值对接地压力的线形图和得分均值对重心的线形图，结果如图 25-10 中的另两个图形所示。

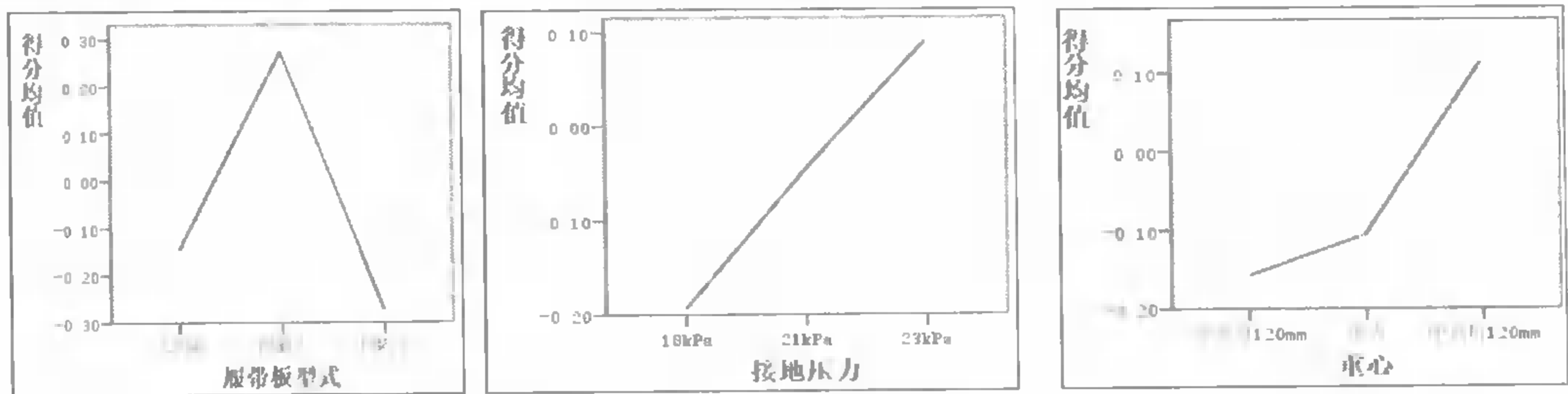


图 25-10 各因素不同水平的得分均值图

如图 25-10 所示，分别单独考虑 3 个影响因素时，履带板型式取间隔小、接地压力取 18kPa、重心取中心点前 120mm 时，湿地推土机的综合得分最小，也就是性能最好。以符号表示就是说因素组合的最佳水平为  $A_3B_1C_1$ 。而在已经进行的试验中，综合得分最小的是第 7 号试验  $score=-0.38$ ，其因素水平的组合为  $A_3B_1C_3$ ，可见由图 25-10 推断的最佳组合与已经试验过的最佳组合有一定差异，为了确定二者究竟哪个最好，建议进行第二批试验。

考虑到试验的批次不同，周围的环境也会不同，由此带来的问题是两个批次的试验结果之间不易比较，所以第二批试验对所关心的两个方案都进行了测试。为此所做第二批试验的结果数据如表 25-4 所示。

表 25-4 第二批试验数据

试验号	试验组合方案	指 标 值			综合加权评分值 $f_i$
		行走阻力 (10N)	滑转率%	下陷深度 (mm)	
10	$A_3B_1C_1$	659	2.3	10.2	15.6
11	$A_3B_1C_3$	773	3.0	11.4	46.3

从表 25-4 看，同一个批次的试验里， $A_3B_1C_1$  的各项得分都要小于  $A_3B_1C_3$  的得分，故推断该试验方案的最优因素组合为  $A_3B_1C_1$ ，即当湿地推土机履带板型式为间隔小、接地压力为 18kPa、重心距中点前 120mm 时，它的性能最好。

### 25.5 建议和推广

本章的整个分析过程都是直接建立在已知的试验结果基础上，试验数据如表 25-2 所示。事实上通过 SPSS 的正交设计过程，也可以实现关于试验因素的正交表设计。

依次单击菜单“Data→Orthogonal Design→Generate...”执行正交设计过程，其主设置界面如图 25-11 所示。

如图 25-11 所示，在前两个输入框分别输入因素变量的名称（如 A）和标签（如履带板型式），单击 Add 按钮加入下面的因素列表。在因素列表选中某个因素名称，单击 Define Values 按钮，弹出如图 25-12 所示的因素取值水平设置子对话框，在此设置所选因素变量的每个取值及其值标签，例如履带板型式可以取 1（对应的值标签为无间隔）。单击 Continue 按钮返回主界面。

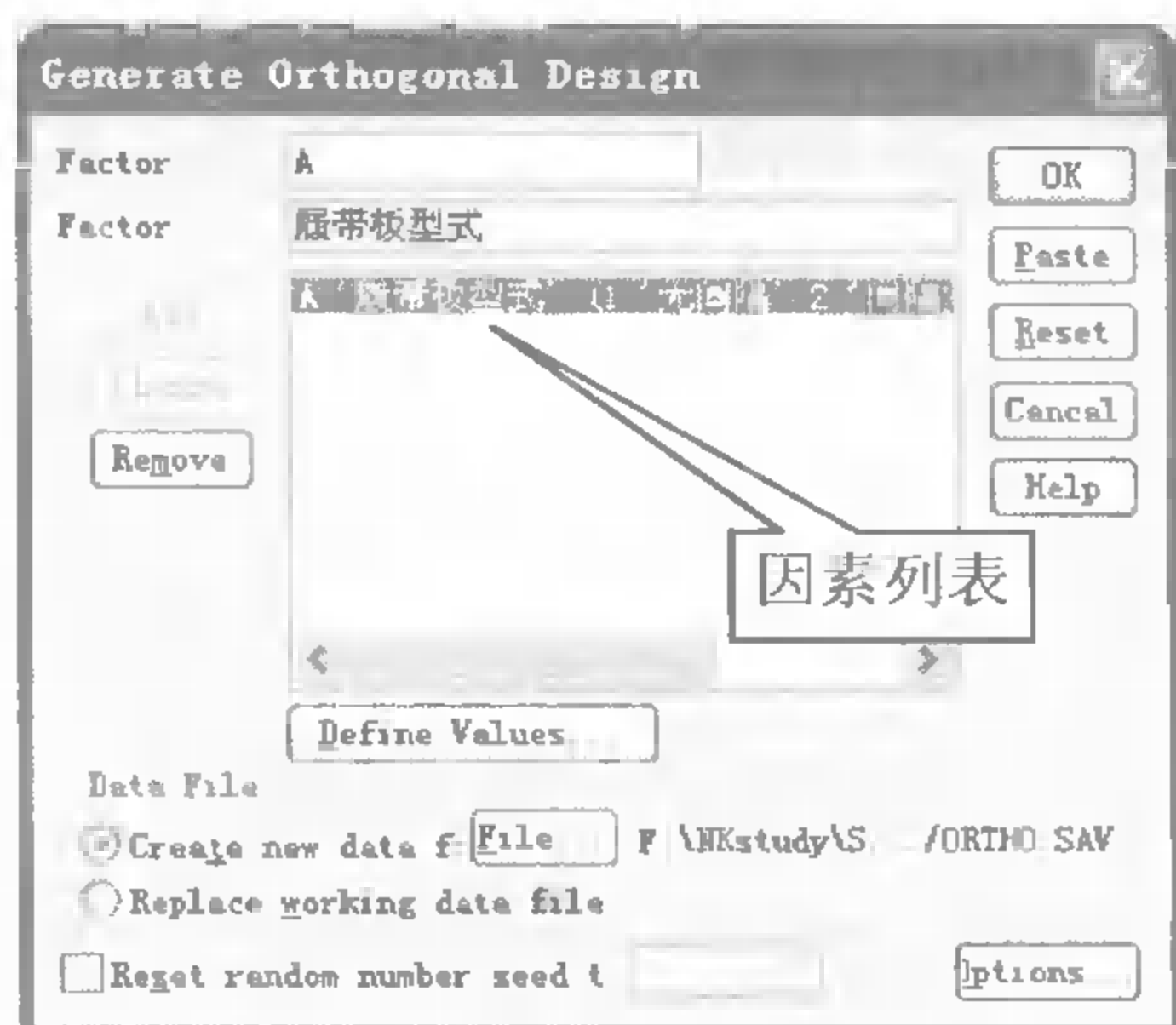


图 25-11 正交设计过程的主界面

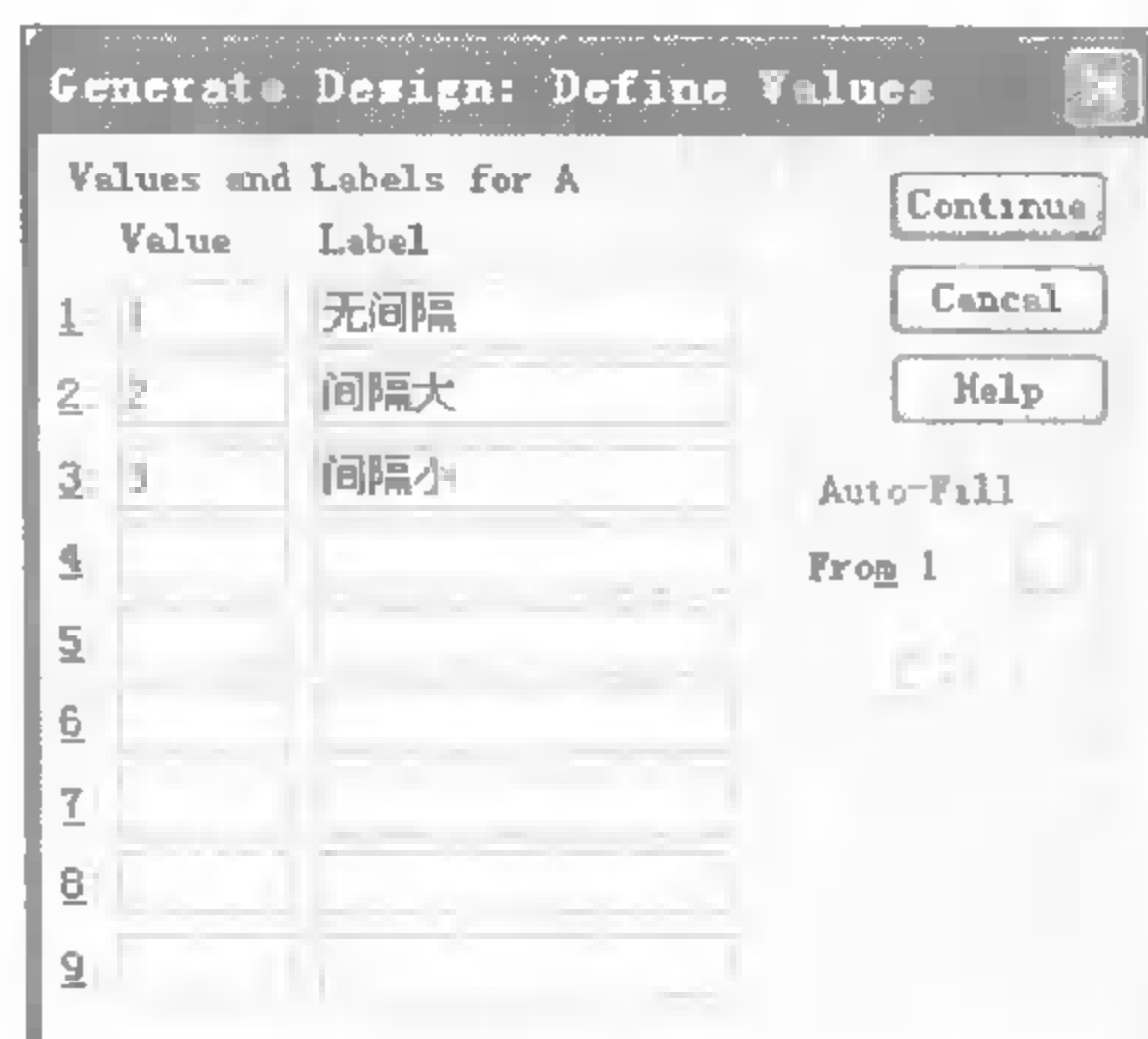


图 25-12 正交设计过程的参数设置

在图 25-11 中单击 OK 按钮运行，就可以按照指定的因素个数和水平个数，生成一个类似表 25-3 所示的正交表数据文件。

即使在因素个数和取值水平个数都相等的情况下，多次设计的正交表也不一定相同，这一点可以用如下的方法解决：在图 25-11 底部勾选 Reset random number seed 复选框，并在其后指定一个固定的随机种子，以后每次都使用这个随机种子就可以产生相同的正交表了。

用 SPSS 进行多因素试验设计与分析的一般步骤就是先按照本节介绍的方法设计好正交表；然后按照正交表给出的因素水平组合方案实施试验；每个试验完成后，把试验数据录入正交表中对应方案的行末；最后对试验数据进行统计和分析。