



## 第2章 SPSS统计分析前的准备

## 2.1 SPSS数据文件的建立

CONCEPT  
STRATE

SPSS数据文件的建立可以利用【File(文件)】菜单中的命令来实现。具体来说，SPSS提供了四种创建数据文件的方法：

- 新建数据文件；
- 直接打开已有数据文件；
- 使用数据库查询；
- 从文本向导导入数据文件。

## 2.1.1 新建数据文件

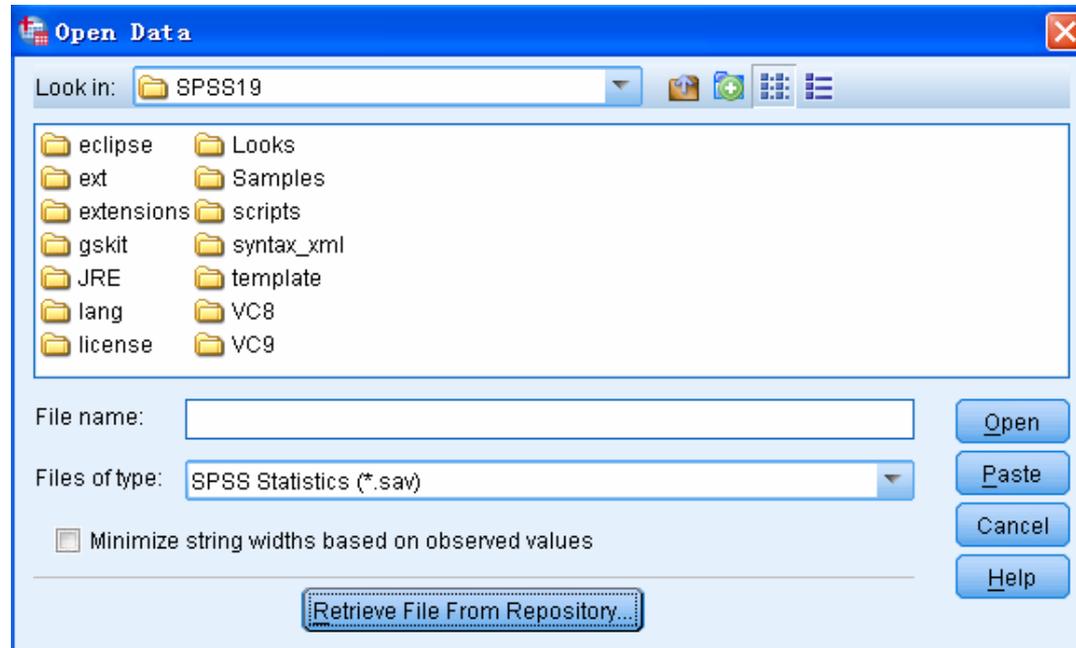


打开SPSS软件后，现在菜单栏中的【File(文件)】→【New(新建)】→【Data(数据)】命令，可以创建一个新的SPSS空数据文件。接着，用户可以进行直接录入数据等后续工作。

值得注意的是，SPSS19.0可以同时打开多个数据文件，用户可以在多个文件中进行转换操作，这比起低版本的SPSS来说，更方便用户使用。

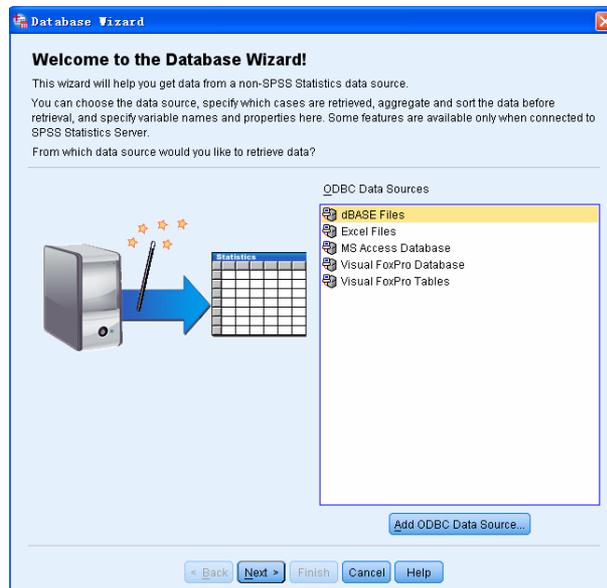
## 2.1.2 直接打开已有数据文件

- 打开SPSS软件后，现在菜单栏中的【File(文件)】→【Open(打开)】→【Data(数据)】命令，弹出【Open Data(打开数据)】对话框。选中需要打开的数据类型和文件名，双击打开该文件。



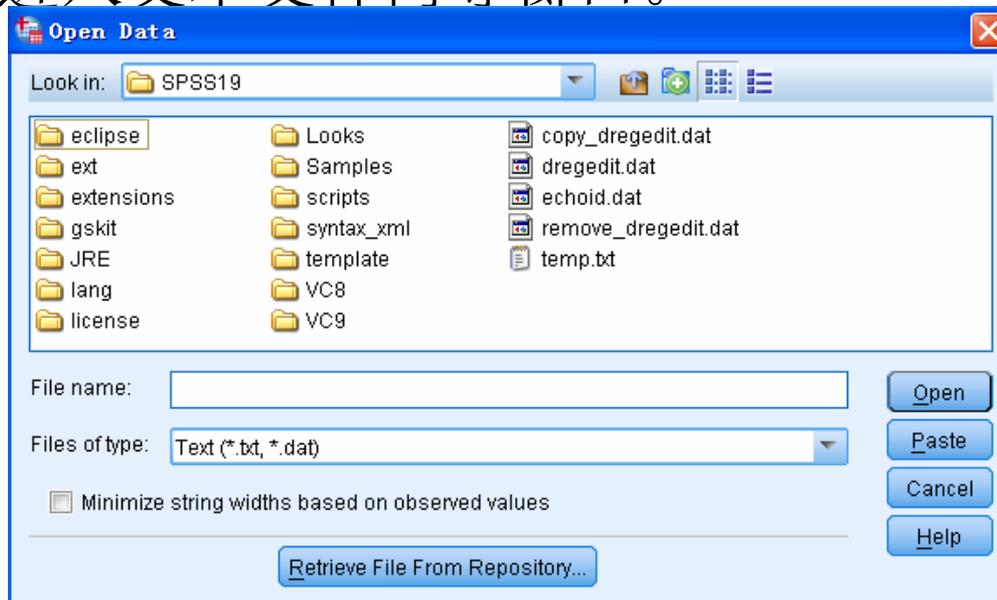
## 2.1.3 利用数据库导入数据

- 打开软件后，现在菜单栏中的【File(文件)】→【Open Database(打开数据库)】→【New Query(新建查询)】命令，弹出【Database Wizard(数据库向导)】对话框。通过这个数据库向导窗口，用户可以选择需要打开的文件类型，并按照窗口上的提示进行相关操作。



## 2.1.4 文本向导导入数据

- SPSS提供了专门读取文本文件的功能。打开软件后，现在菜单栏中的【File(文件)】→【Read Text Data(打开文本数据)】命令，弹出【Open Data(打开数据)】对话框。这里用户需要选择需要打开的文件名称，并且单击【Open(打开)】按钮进入文本文件向导窗口。





## 2.1.4 文本向导导入数据

**Text Import Wizard - Step 1 of 6**

Welcome to the text import wizard!  
This wizard will help you read data from your text file and specify information about the variables.

Does your text file match a predefined format?

Yes

No

Text file: C:\Program Files\SPSS19\temp.txt

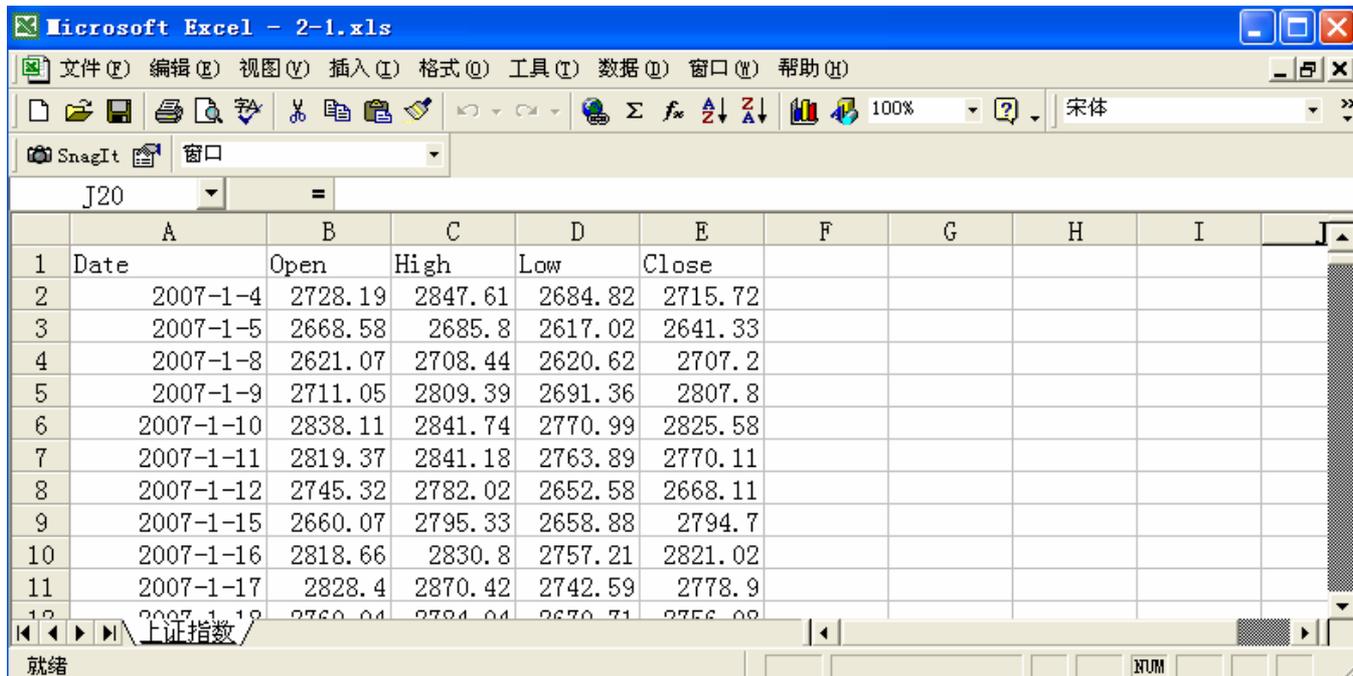
0 ..... 10 ..... 20 ..... 30 ..... 40 ..... 50 .....

1	DBQG4NOBE8KM2CR6GZWM83US94ILCFVVBJR9HEPF8WU7ONR4JD5KZ98G
---	--

< Back Next > Finish Cancel Help

## 2.1.5 实例分析：股票指数的导入

- 文件2-1.xls是上证指数从2007年1月4日至2008年10月16日的资料，包括了开盘价、当日最高价、当日最低价和收盘价等选项，请将该数据导入至SPSS中。



The screenshot shows a Microsoft Excel spreadsheet titled "2-1.xls" containing the following data:

	A	B	C	D	E	F	G	H	I	J
1	Date	Open	High	Low	Close					
2	2007-1-4	2728.19	2847.61	2684.82	2715.72					
3	2007-1-5	2668.58	2685.8	2617.02	2641.33					
4	2007-1-8	2621.07	2708.44	2620.62	2707.2					
5	2007-1-9	2711.05	2809.39	2691.36	2807.8					
6	2007-1-10	2838.11	2841.74	2770.99	2825.58					
7	2007-1-11	2819.37	2841.18	2763.89	2770.11					
8	2007-1-12	2745.32	2782.02	2652.58	2668.11					
9	2007-1-15	2660.07	2795.33	2658.88	2794.7					
10	2007-1-16	2818.66	2830.8	2757.21	2821.02					
11	2007-1-17	2828.4	2870.42	2742.59	2778.9					
12	2007-1-19	2760.04	2794.04	2670.71	2756.09					

## 2.1.5 实例分析：股票指数的导入

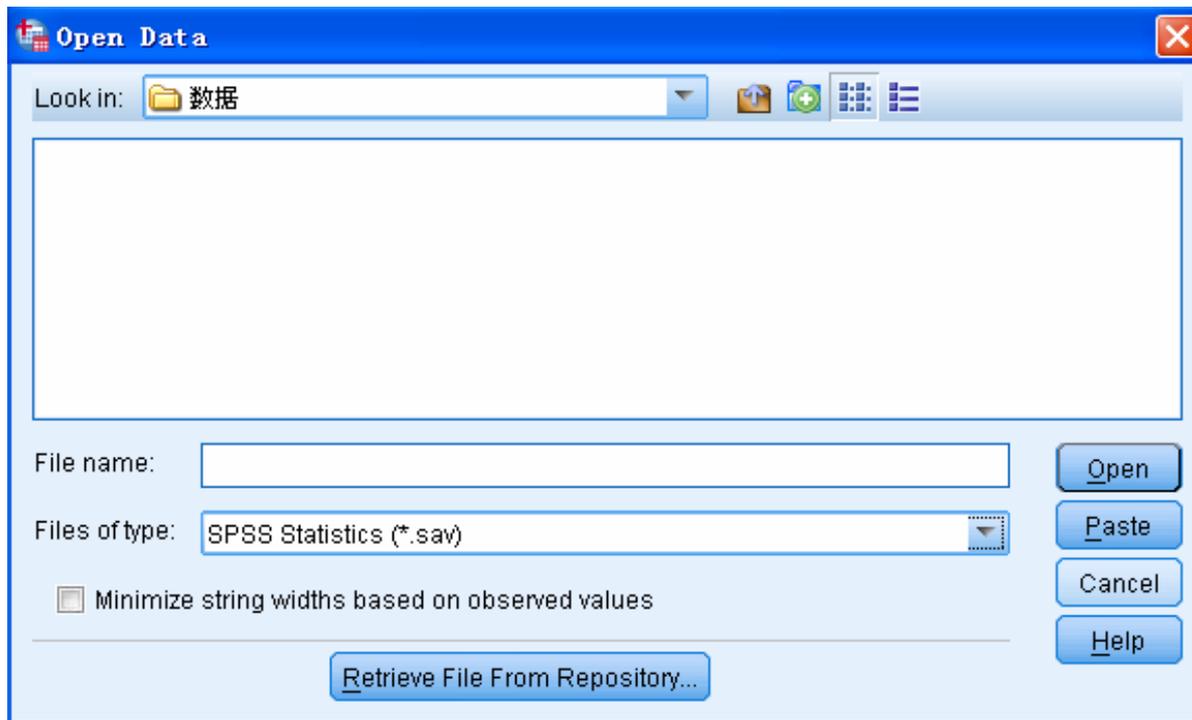
CONCEPT  
STRATE

- **Step01:** 选定对话框

打开SPSS软件，选择菜单栏中的【File(文件)】  
→ 【Open(打开)】 → 【Data(数据)】命令，弹出  
【Open Data(打开数据)】对话框。



## 2.1.5 实例分析：股票指数的导入



## 2.1.5 实例分析：股票指数的导入

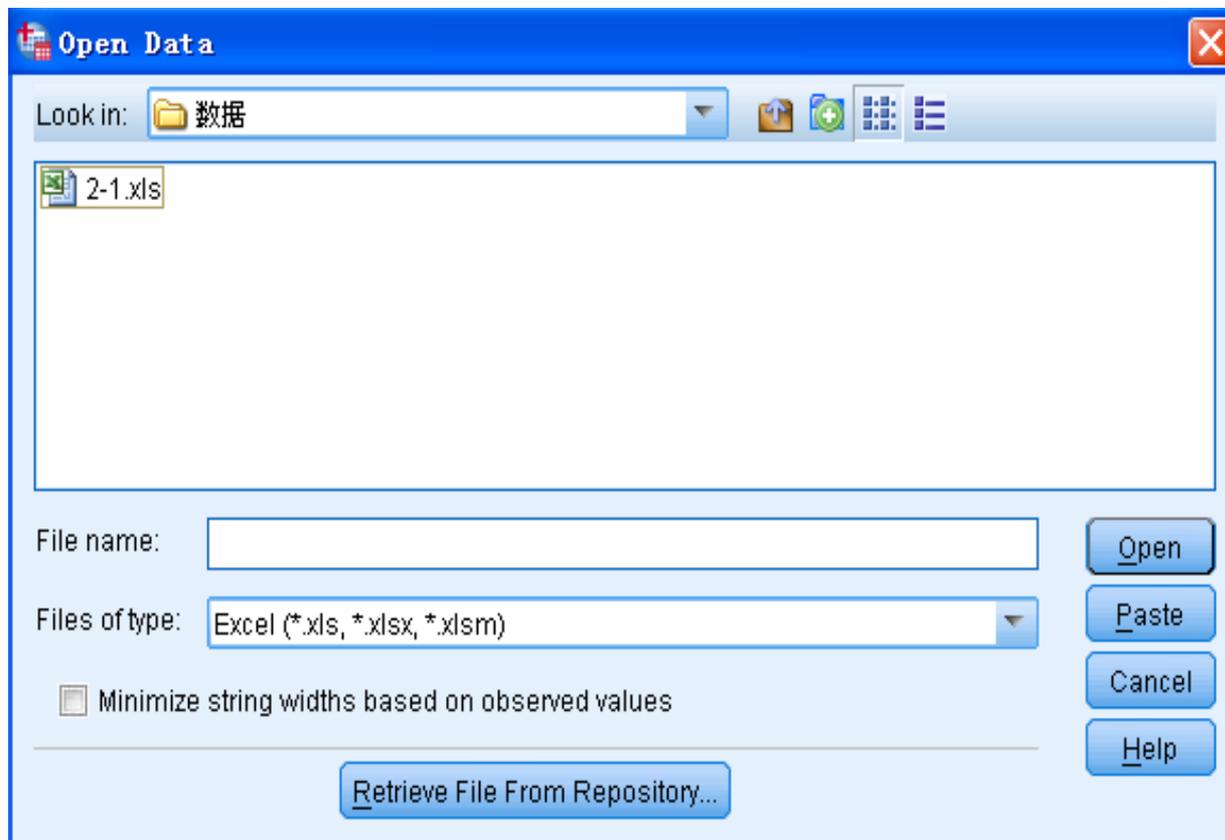
CONCEPT  
STRATE

- **Step02:** 选定打开文件类型

在【Files of type (文件类型)】下拉列表框中指定打开Excel文件类型。接着，选择2-1.xls文件。最后单击【Open (打开)】按钮。



## 2.1.5 实例分析：股票指数的导入



## 2.1.5 实例分析：股票指数的导入

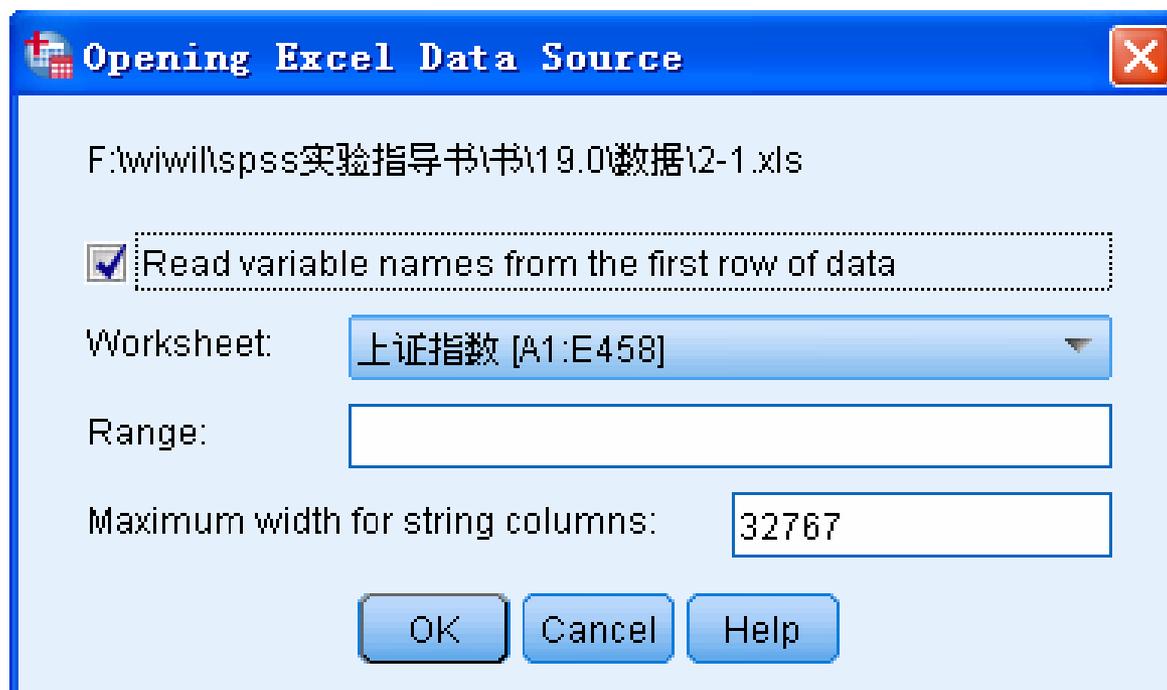
CONCEPT  
STRATE

- **Step03:** 设置变量名称

弹出的对话框中的【Read variable names from the first row of data(从第一行数据读取变量名)】复选框表示SPSS将Excel工作表的第一行设定为SPSS的变量名称，【Range(范围)】文本框表示选定Excel文件导入SPSS的数据范围。这里，保持系统默认选项。



## 2.1.5 实例分析：股票指数的导入



## 2.1.5 实例分析：股票指数的导入

CONCEPT  
STRATE

- **Step04: 完成操作**

最后，单击【OK(确定)】按钮，数据即可导入成功。此时，SPSS的数据浏览窗口中会出现相关的数据内容。



## 2.1.5 实例分析：股票指数的导入

	Date	Open	High	Low	Close
1	04-Jan-2007	2728.19	2847.61	2684.82	2715.72
2	05-Jan-2007	2668.58	2685.80	2617.02	2641.33
3	08-Jan-2007	2621.07	2708.44	2620.62	2707.20
4	09-Jan-2007	2711.05	2809.39	2691.36	2807.80
5	10-Jan-2007	2838.11	2841.74	2770.99	2825.58
6	11-Jan-2007	2819.37	2841.18	2763.89	2770.11
7	12-Jan-2007	2745.32	2782.02	2652.58	2668.11
8	15-Jan-2007	2660.07	2795.33	2658.88	2794.70
9	16-Jan-2007	2818.66	2830.80	2757.21	2821.02
10	17-Jan-2007	2828.40	2870.42	2742.59	2778.90
11	18-Jan-2007	2760.94	2784.04	2679.71	2756.98
12	19-Jan-2007	2761.89	2833.45	2761.89	2832.21
13	22-Jan-2007	2857.90	2934.65	2857.90	2933.19
14	23-Jan-2007	2964.69	2970.69	2851.92	2949.14
15	24-Jan-2007	2955.42	2994.28	2927.72	2975.13
16	25-Jan-2007	2946.50	2947.15	2853.82	2857.36
17	26-Jan-2007	2805.96	2905.98	2720.83	2882.56
18	29-Jan-2007	2897.25	2954.34	2885.86	2945.26
19	30-Jan-2007	2959.40	2980.51	2901.76	2930.56
20	31-Jan-2007	2926.07	2929.65	2766.75	2786.33

## 2.2 SPSS数据文件的属性

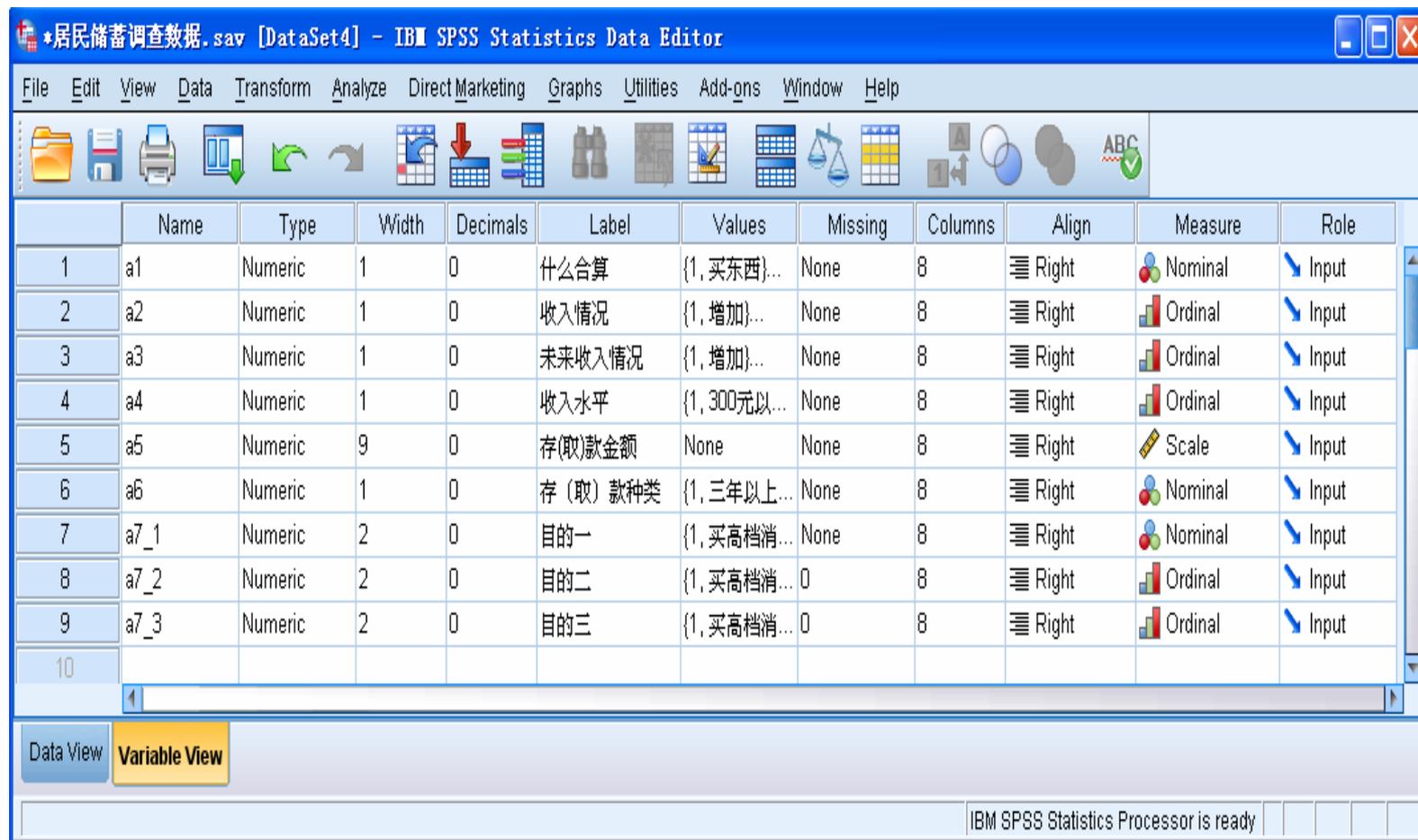
CONCEPT  
STRATE

一个完整的SPSS文件结构包括变量名称、变量类型、变量名标签、变量值标签等内容。用户可以在创建了数据文件后，单击数据浏览窗口左下方的【Variable View(变量视图)】选项卡，进入数据结构定义窗口。用户可以在该窗口中设定或修改文件的各种属性。

**注意：**SPSS数据文件中的一列数据称为一个变量，每个变量都应有一个变量名。SPSS数据文件中的一行数据称为一条个案或观测量（Case）。

## 2.2 SPSS数据文件的属性

CONCEPT  
RATE



\*居民储蓄调查数据.sav [DataSet4] - IBM SPSS Statistics Data Editor

File Edit View Data Transform Analyze Direct Marketing Graphs Utilities Add-ons Window Help

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	a1	Numeric	1	0	什么合算	{1,买东西}...	None	8	Right	Nominal	Input
2	a2	Numeric	1	0	收入情况	{1,增加}...	None	8	Right	Ordinal	Input
3	a3	Numeric	1	0	未来收入情况	{1,增加}...	None	8	Right	Ordinal	Input
4	a4	Numeric	1	0	收入水平	{1,300元以上}	None	8	Right	Ordinal	Input
5	a5	Numeric	9	0	存(取)款金额	None	None	8	Right	Scale	Input
6	a6	Numeric	1	0	存(取)款种类	{1,三年以上}	None	8	Right	Nominal	Input
7	a7_1	Numeric	2	0	目的—	{1,买高档消}	None	8	Right	Nominal	Input
8	a7_2	Numeric	2	0	目的二	{1,买高档消}	0	8	Right	Ordinal	Input
9	a7_3	Numeric	2	0	目的三	{1,买高档消}	0	8	Right	Ordinal	Input
10											

Data View Variable View

IBM SPSS Statistics Processor is ready

## 2.2.1 变量名：Name栏

CONCEPT  
STRATE

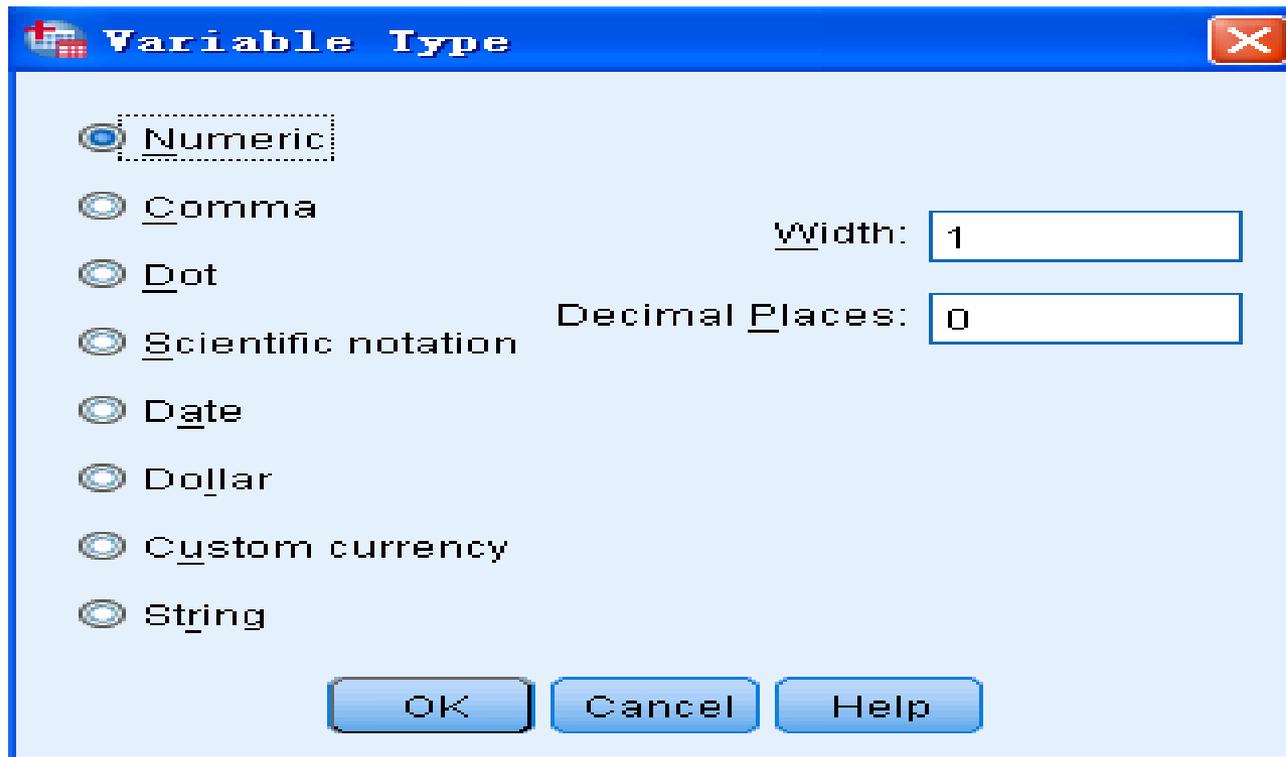
变量名（Name）是变量存取的唯一标志。在定义SPSS数据属性时应首先给出每列变量的变量名。变量命名应遵循下列基本规则：

- SPSS 变量长度不能超过64个字符（32个汉字）；
- 首字母必须是字母或汉字；
- 变量名的结尾不能是圆点、句号或下划线；
- 变量名必须是唯一的；
- 变量名不区分大小写；
- SPSS的保留字不能作为变量名，例如ALL、NE、EQ和AND等；
- 如果用户不指定变量名，SPSS软件会以“VAR”开头来命名变量，后面跟5个数字，如VAR00001、VAR00019等。

注意：为了方便记忆，用户所取的变量名最好与其代表的数据含义相对应。

## 2.2.2 变量类型：Type栏

- 变量类型是指每个变量取值的类型。SPSS提供了三种基本数据类型：数值型、字符型和日期型。



## 2.2.3 变量格式宽度：With栏

CONCEPT  
TRATE

变量格式宽度With是指在数据窗口中变量列所占的单元格的列宽度，一般用户采用系统默认选项即可。值得注意的是，如果变量宽度大于变量格式宽度，此时数据窗口中显示变量名的字符数不够，变量名将被截去尾部作不完全显示。被截去的部分用“\*”号代替。



## 2.2.4 变量小数位数：Decimals栏

【Decimals Places】文本框可以设置变量的小数位数，系统默认为两位。



## 2.2.5 变量名标签：Label栏

变量名标签（Label）是对变量名含义的进一步解释说明，它可以增强变量名的可视性和统计分析结果的可读性。用户有时在处理大规模数据时，变量数目繁多，此时对每个变量的含义加以标注，有利于用户弄清每个变量代表的实际含义。变量名标签可用中文，总长度可达120个字符。同时该属性可以省略，但建议最好给出变量名的标签。

## 2.2.5 变量名标签: Label栏

CONCEPT  
TRATE

	Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
1	x1	String	40	0	公司名称	None	None	40	≡ Left	 Nominal	 Input
2	x2	Numeric	8	2	总人数	None	None	8	≡ Right	 Scale	 Input
3	x3	Numeric	8	2	男员工数	None	None	8	≡ Right	 Scale	 Input
4	x4	Numeric	8	2	女员工数	None	None	8	≡ Right	 Scale	 Input
5	x5	Numeric	8	2	博士人数	None	None	8	≡ Right	 Scale	 Input
6	x6	Numeric	8	2	硕士人数	None	None	8	≡ Right	 Scale	 Input
7	x7	Numeric	8	2	学士人数	None	None	8	≡ Right	 Scale	 Input
8	x8	Numeric	8	2	大专人数	None	None	8	≡ Right	 Scale	 Input
9	x9	Numeric	8	2	中专以下人数	None	None	8	≡ Right	 Scale	 Input
10	x10	Numeric	8	2	35岁以下人数	None	None	8	≡ Right	 Scale	 Input
11	x11	Numeric	8	2	36~45岁人数	None	None	8	≡ Right	 Scale	 Input
12	x12	Numeric	8	2	46岁以上人数	None	None	8	≡ Right	 Scale	 Input
13	x13	Numeric	8	2	公司类别	{1.00, 全国...	None	8	≡ Right	 Scale	 Input
14	x10_1	Numeric	8	2	年轻人比例	None	None	8	≡ Right	 Scale	 Input
15	x5678_1	Numeric	8	2	受高等教育比例	None	None	8	≡ Right	 Scale	 Input

## 2.2.6 变量值标签: Values栏

CONCEPT  
RATE

- 变量值标签 (Values) 是对变量的可能的取值的含义进行进一步说明。变量值标签特别对于数值型变量表示非数值型变量时尤其有用。
- 定义和修改变量值标签，可以双击要修改值的单元格，在弹出的对话框的【Values (值)】文本框中输入变量值，在【Label (标签)】文本框中输入变量值标签，然后单击【Add (添加)】按钮将对应关系选入下边的白框中。同时，可以单击【Change (改变)】和【Remove (移动)】按钮对已有的标签值进行修改和剔除。最后单击【OK (确定)】按钮返回主界面。



## 2.2.6 变量值标签: Values栏

**Value Labels** [X]

Value Labels

Value:  Spelling...

Label:

1 = "300元以下"  
2 = "300~800元"  
3 = "800~1500元"  
4 = "1500元以上"

## 2.2.7 变量缺失值：Missing栏

在统计分析中，收集到的数据可能会出现这样的情况：一是数据中出现明显的错误和不合理的情形；另一种是有些数据项的数据漏填了。

双击【Missing(缺失)】栏，在弹出的对话框中可以选择三种缺失值定义方式。

The screenshot shows the 'Missing Values' dialog box with the following options and fields:

- No missing values
- Discrete missing values
  - Three empty text boxes for defining discrete values.
- Range plus one optional discrete missing value
  - Low: [ ] High: [ ]
  - Discrete value: [ ]

Buttons: OK, Cancel, Help

## 2.2.8 变量列宽：Columns栏

CONCEPT  
STRATE

- **【Columns (列)】** 栏主要用于定义列宽，单击其向上和向下的箭头按钮选定列宽度。系统默认宽度等于8。

## 2.2.9 变量对齐方式：Align栏

CONCEPT  
RATE

- **【Align(对齐)】** 栏主要用于定义变量对齐方式，用户可以选择**Left**（左对齐）、**Right**（右对齐）和**Center**（居中对齐）。系统默认变量右对齐。



## 2.2.10 变量测度水平：Measure栏

- **【Measure (测度)】** 栏主要用于定义变量的测度水平，用户可以选择Scale（定距型数据）Ordinal（定序型数据）和Nominal（定类型数据）。

Missing	Columns	Align	Measure
None	8	≡ Right	 Nominal
None	8	≡ Right	 Scale
None	8	≡ Right	 Ordinal
None	8	≡ Right	 Nominal



## 2.2.11 变量角色：Role栏

- 【Role(角色)】栏主要用于定义变量在后续统计分析中的功能作用，用户可以选择Input、Target和Both等类型的角色。

Columns	Align	Measure	Role
8	≡ Right	Nominal	Input
8	≡ Right	Ordinal	Input
8	≡ Right	Ordinal	Target
8	≡ Right	Ordinal	Both
8	≡ Right	Scale	None
8	≡ Right	Nominal	Partition
8	≡ Right	Nominal	Split
8	≡ Right	Ordinal	Input

## 2.2.11 实例分析：员工满意度 调查表的数据属性设计

CONCEPT  
STRATE

- 1. 实例内容

为了提高员工的工作积极性，完善公司各方面管理制度，并达到有的放矢的目的，某公司决定对本公司员工进行不记名调查，希望了解员工对公司的满意情况。请根据该公司设计的员工满意度调查题目（行政人事管理部分）的特点，设计该调查表数据在SPSS的数据属性。

## 2. 实例操作

CONCEPT  
STRATE

具体步骤如下。

- **Step01:** 打开SPSS中的Data View窗口，录入或导入原始调查数据。
- **Step02:** 选择菜单栏中的【File(文件)】→【Save(保存)】命令，保存数据文件，以免丢失。
- **Step03:** 单击SPSS中的【Variable View(变量视图)】选项卡，按窗口提示进行数据属性的定义，如变量名称、标签、标签值等。

### 3. 实例结果

CONCEPT  
TRATE

Name	Type	Width	Decimals	Label	Values	Missing	Columns	Align	Measure	Role
A1	Numeric	8	0	公司招聘程序是...	{1, 很合理}...	None	8	≡ Right	 Ordinal	 Input
A2	Numeric	8	0	最应作为考核的...	{1, 任务完成...	None	8	≡ Right	 Scale	 Input
A3	Numeric	8	0	福利政策是否完...	{1, 是}...	None	8	≡ Right	 Nominal	 Input
A4	String	12	0	自己最需要哪些...	None	None	8	≡ Left	 Scale	 Input
A5	Numeric	8	0	有没有发展前途	{1, 有}...	None	8	≡ Right	 Ordinal	 Input
A6	Numeric	8	0	除薪酬外，最看...	{1, 提高自己...	None	8	≡ Right	 Scale	 Input
A7	Numeric	8	0	目前的工作	{1, 很合适, ...	None	8	≡ Right	 Scale	 Input
A8	Numeric	8	0	当前人事管理最...	{1, 招聘}...	None	8	≡ Right	 Scale	 Input

## 2.3 SPSS数据文件的整理

CONCEPT  
STRATE

- 通常情况下，刚刚建立的数据文件并不能立即进行统计分析，这是因为收集到的数据还是原始数据，还不能直接利用分析。此时，需要对原始数据进行进一步的加工、整理，使之更加科学、系统和合理。这项工作在数据分析中称之为统计整理。
- **【Data(数据)】** 菜单中的命令主要用于实现数据文件的整理功能。

## 2.3.1 观测量排序： 地区生产总值分析

CONCEPT  
STRATE

### SPSS操作详解

- **Step01:** 打开观测量排序对话框

打开SPSS软件，选择菜单栏中的【File(文件)】→【Data(数据)】→【Sort Cases(排序个案)】命令，弹出【Sort Cases(排序个案)】对话框。



# 1. SPSS操作详解

CONCEPT  
STRATE

- **Step02:** 选择排序变量

在左侧的候选变量列表框中选择主排序变量，单击向右箭头按钮，将其移动至【Sort by (排序依据)】列表框中。

- **Step03:** 选择排序类型

在【Sort Order (排列顺序)】选项组中可以选择变量排列方案。

- **Step04:** 单击【OK】按钮，此时操作结束。

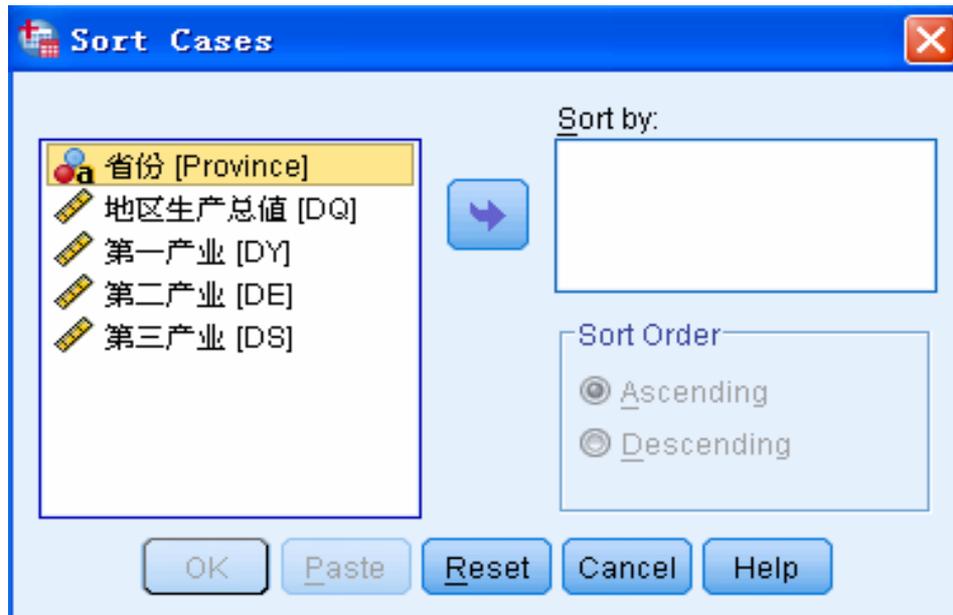
## 2.实例内容：地区生产总值分析

CONCEPT  
TRATE

地区生产总值是指某地区在一定时间内的国内生产总值，它可以作为衡量该地区经济发展的重要综合指标。随书光盘中的数据2-3. sav列出了2005年我国部分省份的地区生产总值及第一产业、第二产业和第三产业的生产总值，请根据这些数据分析不同省份经济发展状况的差异性。

- **Step01:** 选定对话框

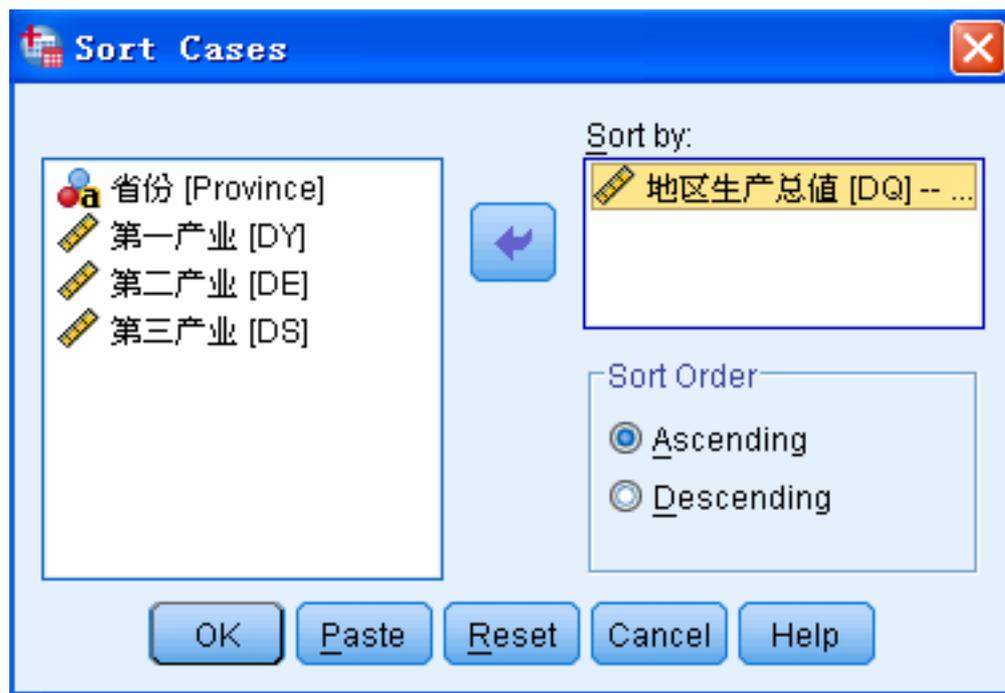
打开SPSS软件，选择菜单栏中的【Data(数据)】  
→【Sort Cases(排序个案)】命令，弹出【Sort Cases(排序个案)】对话框。





- Step02: 选择排序变量

在左侧的候选变量列表框中选择主排序变量DQ，单击向右箭头按钮，将变量选择进入【Sort by (排序依据)】列表框中。



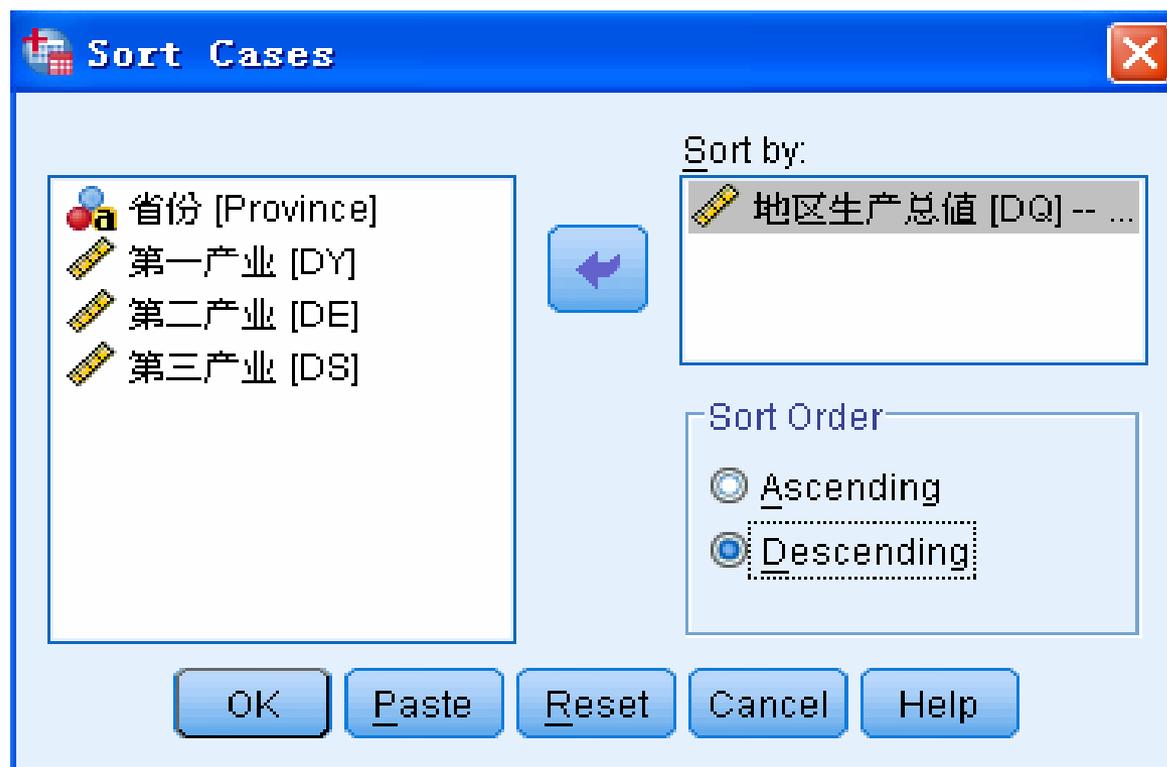


- Step03: 选择排序类型

为了表示不同省份生产总值的差异，按照从高到低的排列顺序，这里点选【**Descending (降序)**】单选钮，表示观测值按照降序进行排序。

## Step03: 选择排序类型

CONCEPT  
STRATE





- **Step04: 完成操作**

最后，单击【OK(确定)】按钮，操作完成。此时，SPSS的数据浏览窗口中观测量的顺序发生改变。

	Province	DQ	DY	DE	DS
1	广东	22366.54	1428.27	11339.93	9598.34
2	山东	18516.87	1963.51	10628.62	5924.74
3	浙江	13437.85	892.83	7166.15	5378.87
4	河南	10587.42	1892.01	5514.14	3181.27
5	河北	10096.11	1503.07	5232.50	3360.54
6	上海	9154.18	80.34	4452.92	4620.92
7	辽宁	8009.01	882.41	3953.28	3173.32
8	四川	7385.11	1481.14	3067.23	2836.74
9	北京	6886.31	97.99	2026.51	4761.81
10	湖南	6511.34	1274.15	2596.71	2640.48
11	黑龙江	5511.50	684.60	2971.68	1855.22
12	广西	4075.75	912.50	1510.68	1652.57
13	江西	4056.76	727.37	1917.47	1411.92
14	内蒙古	3895.55	589.56	1773.21	1532.78
15	天津	3697.62	112.38	2051.17	1534.07
16	云南	3472.89	669.81	1432.76	1370.32
17	重庆	3070.49	463.40	1259.12	1347.97
18	贵州	1979.06	368.94	826.63	783.49
19	甘肃	1933.98	308.06	838.56	787.36
20	海南	894.57	300.75	220.07	373.75
21	宁夏	606.10	72.08	281.23	252.79
22	西藏	251.21	48.04	63.52	139.65

## 2.3.2数据的转置： 国家财政分项目收入

CONCEPT  
STRATE

### 1.操作详解

- **Step01:** 打开转置对话框

打开SPSS软件，选择菜单栏中的【File(文件)】→ Data(数据)】→【Transpose(转置)】命令，弹出【Transpose(转置)】对话框。



- **Step02:** 选择转置变量

在左侧的候选变量列表框中选择需要进行转置的变量，单击向右箭头按钮，将其移动至【Variable(s)(变量)】列表框中。

- **Step03:** 新变量命名

从左侧的候选变量列表框中可以选择一个变量，应用它的值作为转置后新变量的名称。此时，选择该变量进入【Name Variable(名称变量)】列表框内即可。如果用户不选择变量命名，则系统将自动给转置后的新变量赋予Var001、Var002...的变量名。

- **Step04:** 单击【OK】按钮，操作结束。

注意：数据文件转置后，数据属性的定义都会丢失，因此用户要慎重选择本功能。

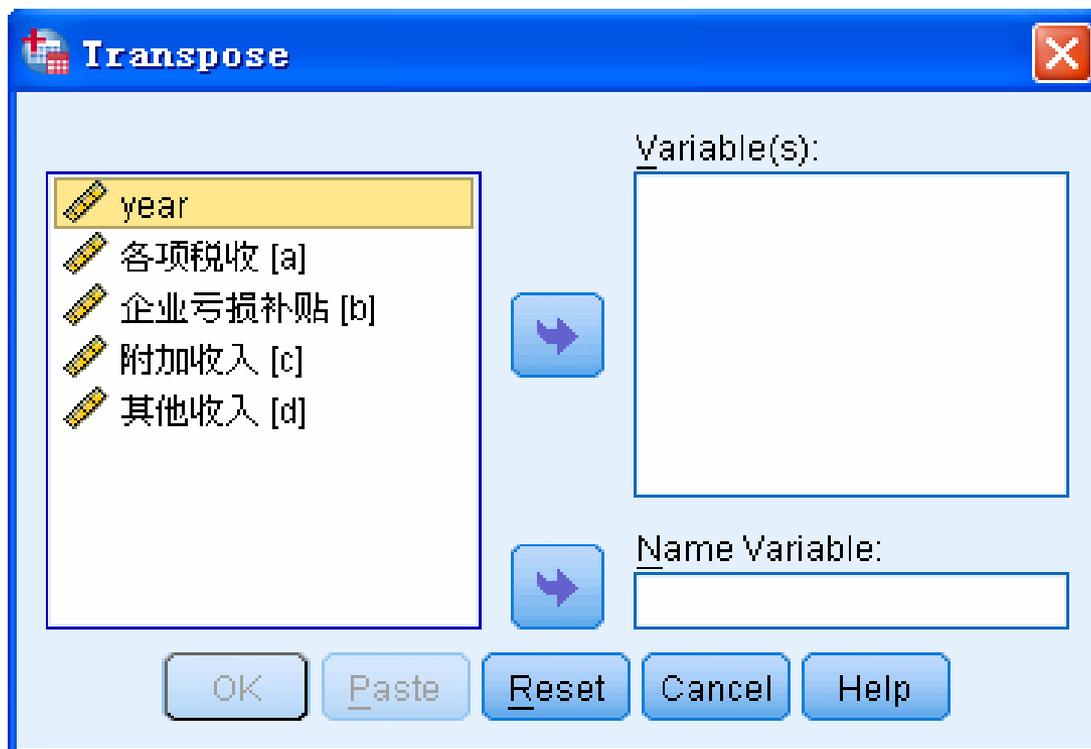
## 2. 实例内容：国家财政分项目收入数据 (2-4\_sav)

CONCEPT  
TRATE

	year	a	b	c	d
1	1991	2990.17	-510.24	28.01	240.10
2	1992	3296.91	-444.96	31.72	265.15
3	1993	4255.30	-411.29	44.23	191.04
4	1994	5126.88	-366.22	64.20	280.18
5	1995	6038.04	-327.77	83.40	396.19
6	1996	6909.82	-337.40	96.04	724.66
7	1997	8234.04	-368.49	103.29	682.30
8	1998	9262.80	-333.49	113.34	833.30
9	1999	10682.58	-290.03	126.10	925.43
10	2000	12581.51	-278.78	147.52	944.98
11	2001	15301.38	-300.04	166.60	1218.10
12	2002	17636.45	-259.60	198.05	1328.74
13	2003	20017.31	-226.38	232.39	1691.93
14	2004	24165.68	-217.93	300.40	2148.32
15	2005	28778.54	-193.26	356.18	2707.83



# Step01: 选定对话框



## Step02: 选择转置变量

CONCEPT  
STRATE





# Step03: 新变量命名



# Step04: 完成操作



	CASE_LBL	K_1991	K_1992	K_1993
1	a	2990.17	3296.91	4255.30
2	b	-510.24	-444.96	-411.29
3	c	28.01	31.72	44.23
4				
5				

## 2.3.3 文件合并：固定资产投资

CONCEPT  
STRATE

- **【data (数据)】** → **【Merge Files (合并文件)】** 菜单中有两个命令选项：**【Add Cases (添加个案)】** 和 **【Add Variables (添加变量)】**。

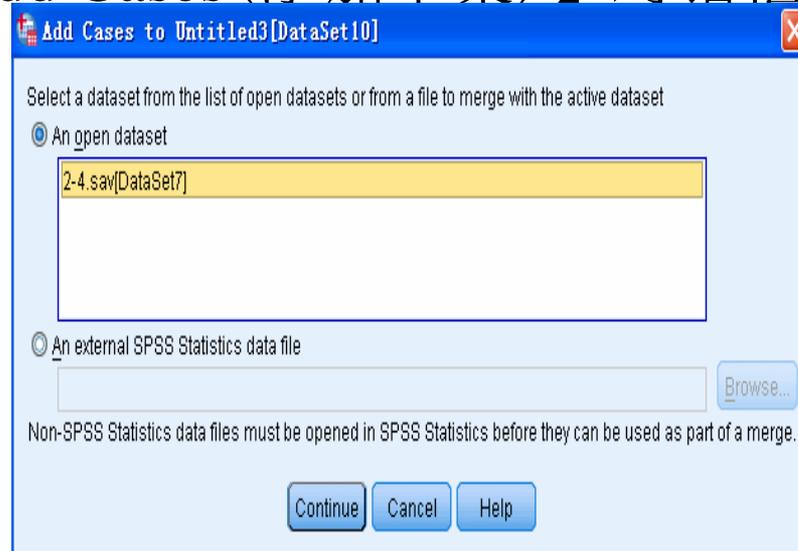
# 1. 观测量合并的SPSS操作详解

CONCEPT  
STRATE

观测量合并要求两个数据文件至少应具有一对属性相同的变量，即使它们的变量名不同。具体步骤如下。

**Step01:** 打开观测量合并对话框

选择菜单栏中的【File(文件)】→【Data(数据)】→【Merge Files(合并文件)】→【Add Cases(添加个案)】命令，弹出【Add Cases(添加个案)】对话框



- **Step02:** 选择合并文件

点选【An external SPSS Statistics data file (外部SPSS Statistics数据文件)】单选钮，同时单击【Browse】按钮，选中需要合并的文件，并指定文件路径，然后单击【Continue】按钮。

- **Step03:** 选择合并方法。
- **Step04:** 单击【OK】按钮，操作结束。

## 2.变量合并的SPSS操作详解

CONCEPT  
RATE

变量合并要求两个数据文件必须具有一个共同的关键变量（Key Variable），而且这两个文件中的关键变量还具有一定数量的相同的观测量数值。

- **Step01:** 打开变量合并对话框。
- **Step02:** 选择合并文件。
- **Step03:** 选择合并方法。
- **Step04:** 单击【OK】按钮，操作结束。

### 3. 实例内容： 固定资产 投资文件的合并



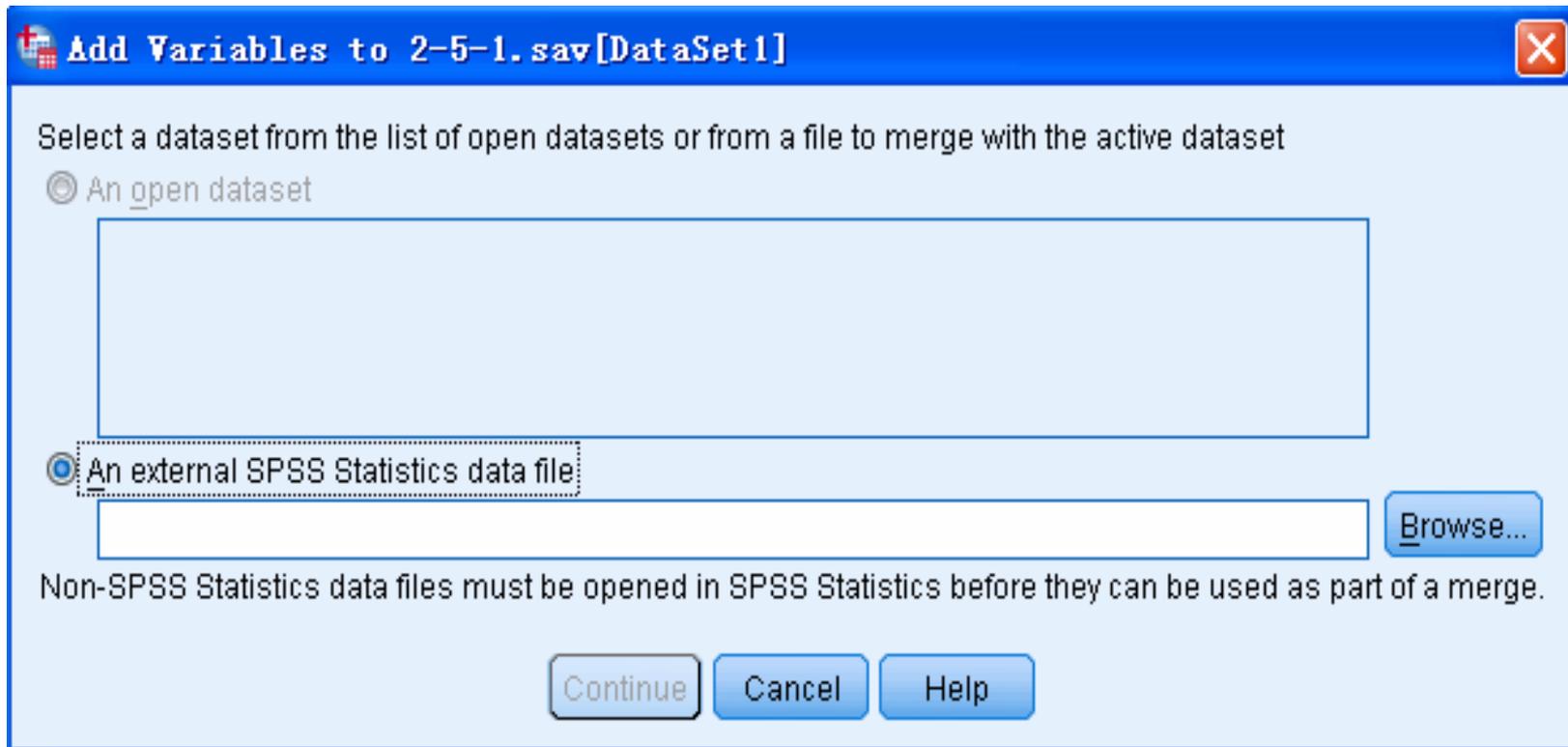
已知2-5-1.sav、2-5-2.sav和2-5-3.sav中的数据是北京、天津、河北等省市在2005年部分行业的固定资产投资额（亿元）数据，请完成以下问题。

问题一：将2-5-1.sav和2-5-2.sav的数据文件纵向合并。

问题二：将2-5-2.sav和2-5-3.sav的数据文件横向合并。

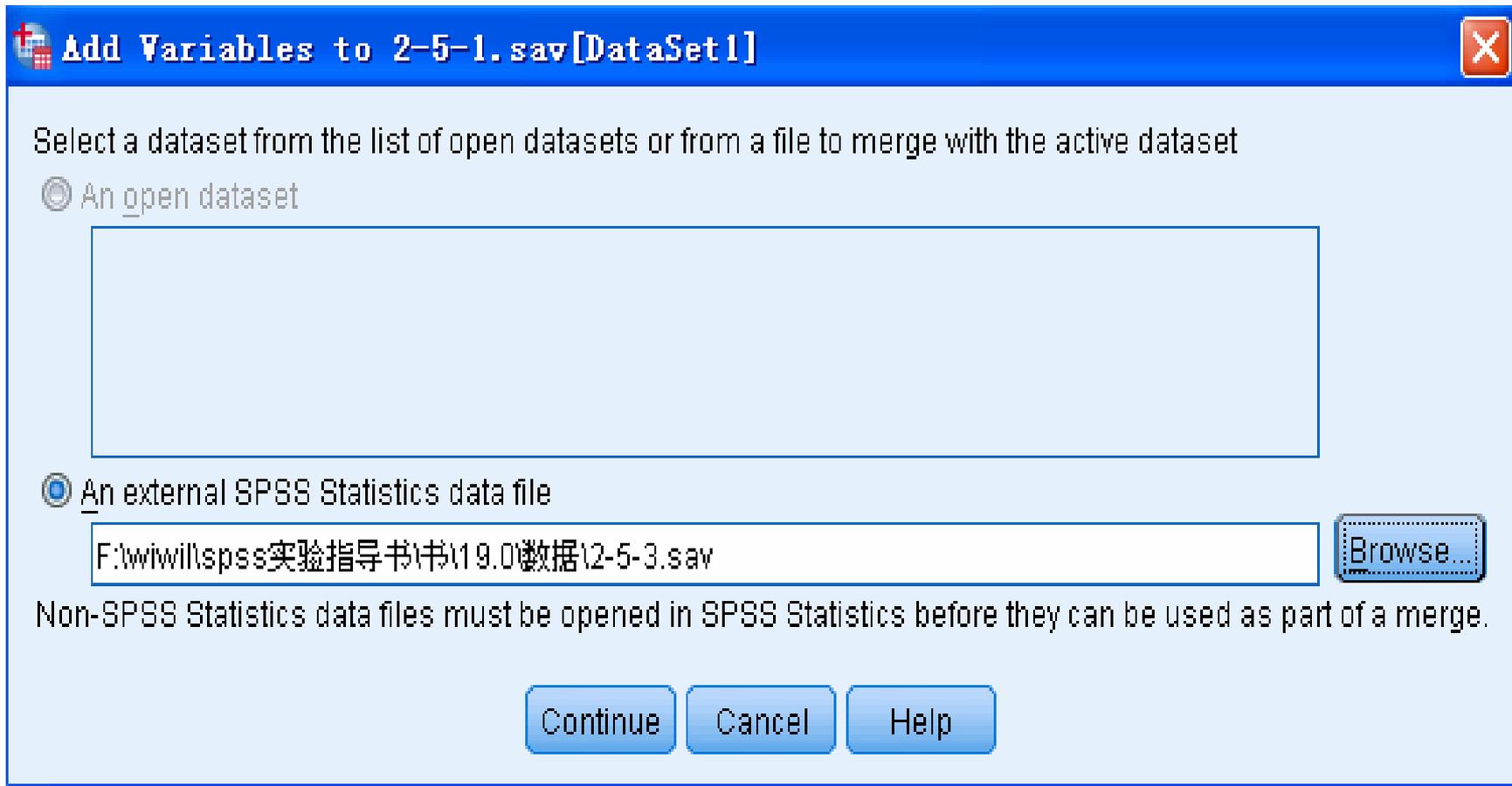


# Step01: 打开对话框（问题一）



## Step02: 选择合并文件

CONCEPT  
TRATE



# Step03: 选择合并方法

CONCEPT  
TRATE

Add Cases From F:\wiwil\spss实验指导书\书\19.0\... ✕

Unpaired Variables:

Variables in New Active Dataset:

地区\*  
采矿业  
制造业  
建筑业  
批发和零售业  
金融业(+)

Indicate case source as variable:  
source01

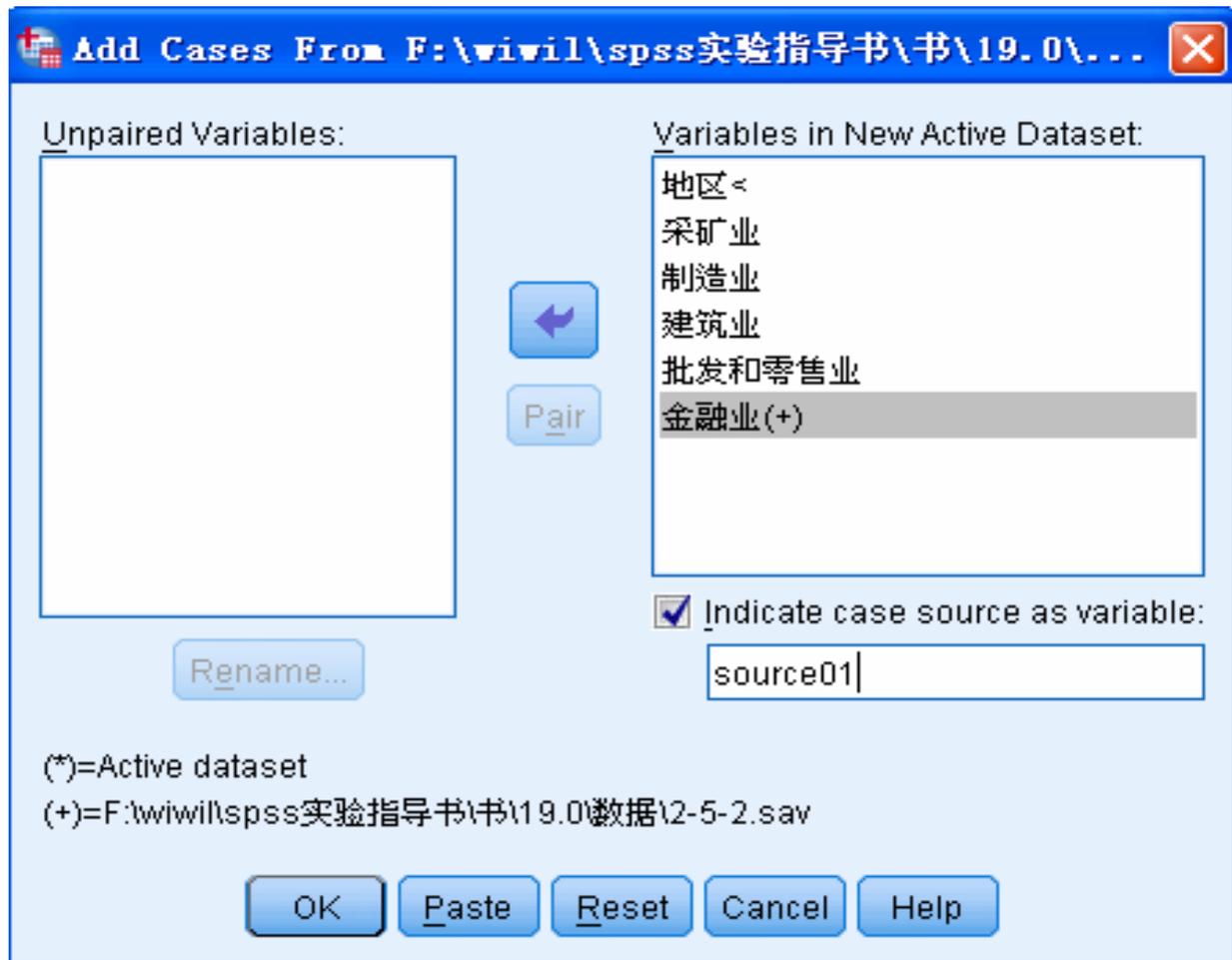
Rename...

(\*)=Active dataset  
(+)=F:\wiwil\spss实验指导书\书\19.0\数据\2-5-2.sav

OK Paste Reset Cancel Help

# Step04: 建立指示变量

CONCEPT  
TRATE



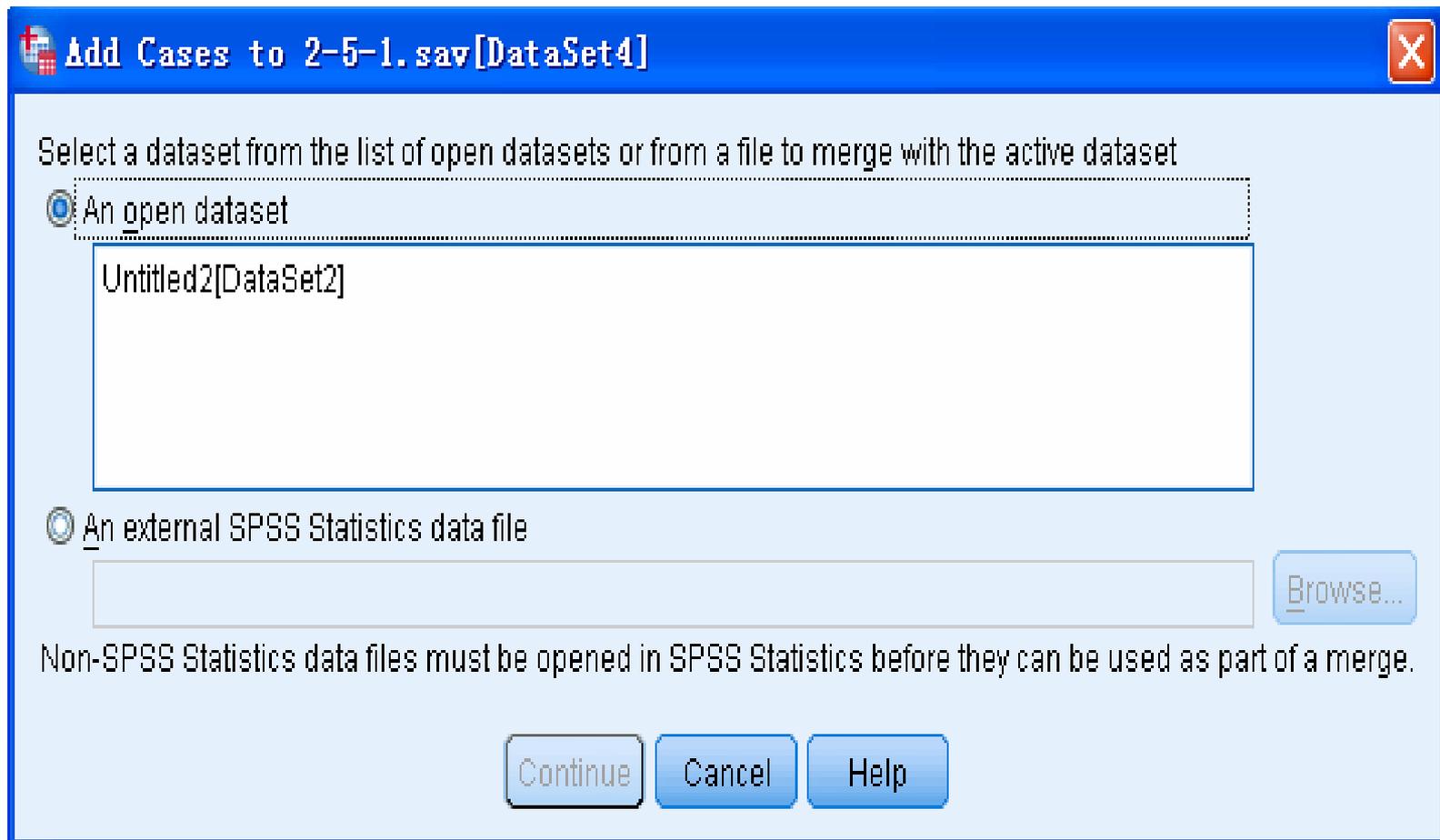
# Step05: 完成操作

CONCEPT  
RATE

	地区	采矿业	制造业	建筑业	批发和零售业	金融业	source01
1	北京	4.40	261.90	26.40	27.70	.	0
2	天津	124.30	417.80	16.30	27.40	.	0
3	河北	134.10	1486.40	22.40	159.00	.	0
4	山西	295.70	547.20	4.30	26.40	.	0
5	内蒙古	257.60	572.90	14.40	60.70	.	0
6	辽宁	205.50	1564.00	67.30	124.10	.	0
7	吉林	89.20	650.90	17.30	46.70	.	0
8	黑龙江	216.90	344.60	20.00	36.50	.	0
9	上海	2.30	873.80	7.90	34.30	.	0
10	江苏	31.90	3560.90	58.50	125.20	.	0
11	浙江	8.90	2256.70	39.70	59.00	5.00	1
12	安徽	131.00	593.10	42.00	41.60	1.90	1
13	福建	21.40	607.20	39.10	23.70	6.50	1
14	江西	42.40	612.60	12.70	35.80	3.10	1
15	山东	329.80	4435.80	273.40	245.10	5.40	1
16	河南	277.40	1315.80	26.10	114.20	4.50	1
17	湖北	48.00	744.80	98.30	63.60	2.70	1
18	湖南	70.70	617.60	39.30	116.00	4.60	1
19	广东	19.90	2292.00	121.60	77.20	9.30	1
20	广西	26.30	359.20	8.80	25.00	2.30	1

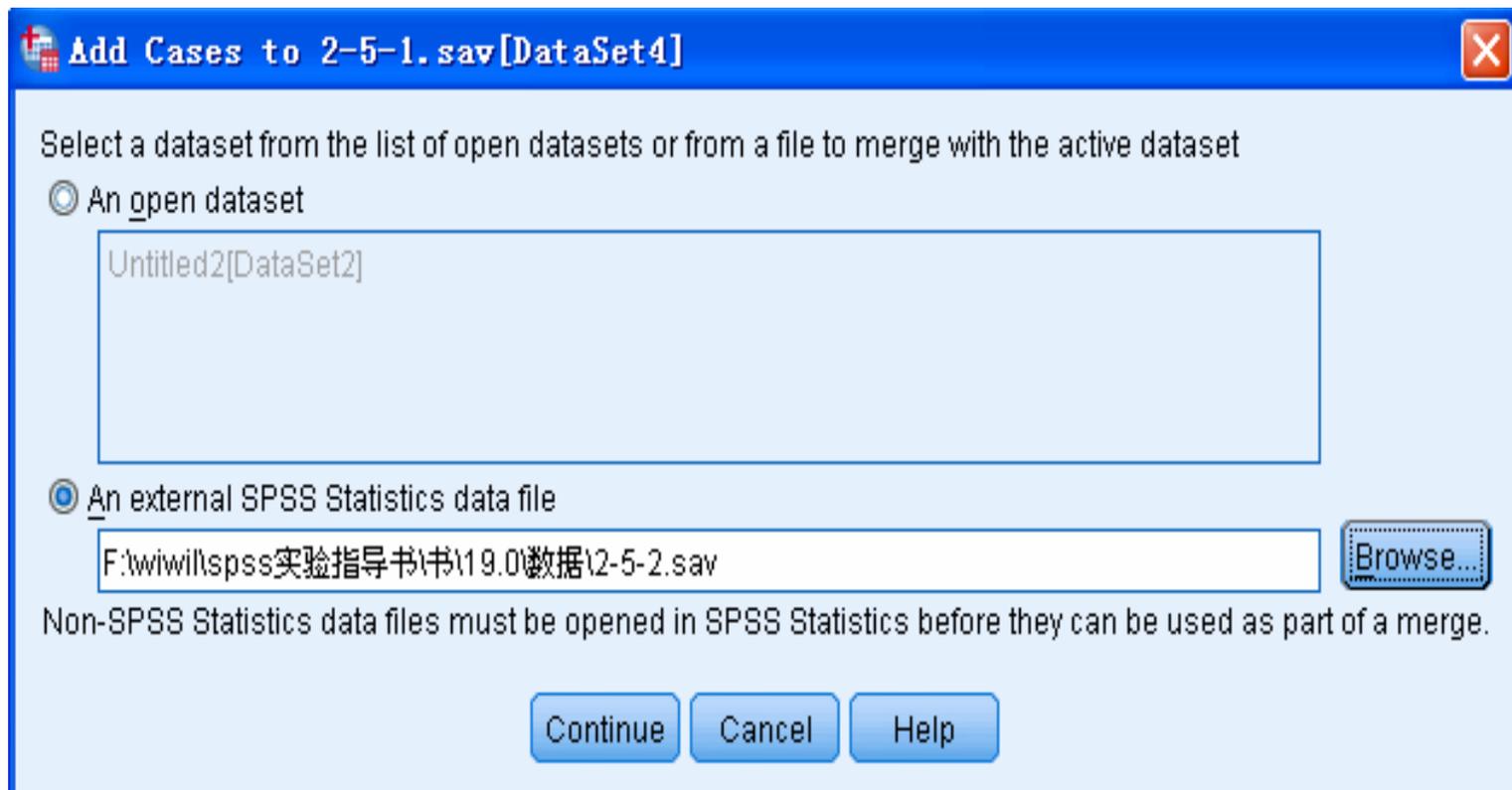


# Step01: 打开对话框（问题二）



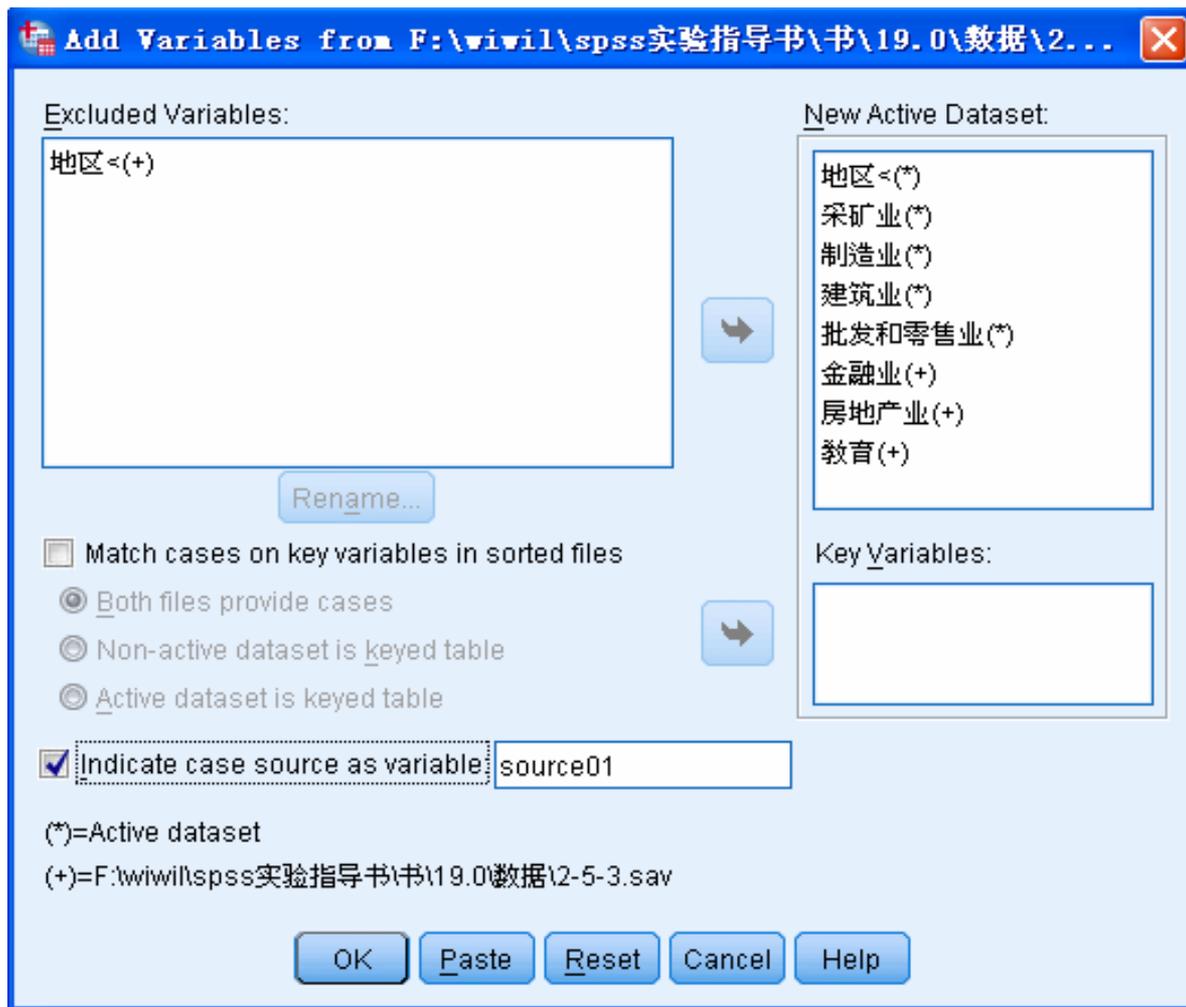
## Step02: 选择合并文件

CONCEPT  
TRATE





# Step03: 建立指示变量



# Step04: 完成操作

CONCEPT  
STRATE

地区	采矿业	制造业	建筑业	批发和零售业	金融业	房地产业	教育	source01
北京	4.40	261.90	26.40	27.70	1.60	1549.80	63.30	1
天津	124.30	417.80	16.30	27.40	.60	370.10	40.50	1
河北	134.10	1486.40	22.40	159.00	4.60	618.40	115.20	1
山西	295.70	547.20	4.30	26.40	1.70	233.70	36.40	1
内蒙古	257.60	572.90	14.40	60.70	1.90	191.90	33.60	1
辽宁	205.50	1564.00	67.30	124.10	17.40	977.60	92.80	1
吉林	89.20	650.90	17.30	46.70	4.60	230.90	43.10	1
黑龙江	216.90	344.60	20.00	36.50	.20	315.40	48.30	1
上海	2.30	873.80	7.90	34.30	.	.	.	0
江苏	31.90	3560.90	58.50	125.20	.	.	.	0

## 2.3.4 数据分类汇总：城乡居民储蓄存款

CONCEPT  
STRATE

对数据进行分类汇总就是按指定的分类变量值对所有的观测量进行分组，对每组观测量的变量求描述统计量，并生成分组数据文件。例如，将一个工厂的数据资料，按照该工厂的各个部门进行分组，并统计各个部门的人员年龄均值、方差等，这些工作就属于数据分类汇总的范畴。

# 1.数据分类汇总的SPSS操作详解

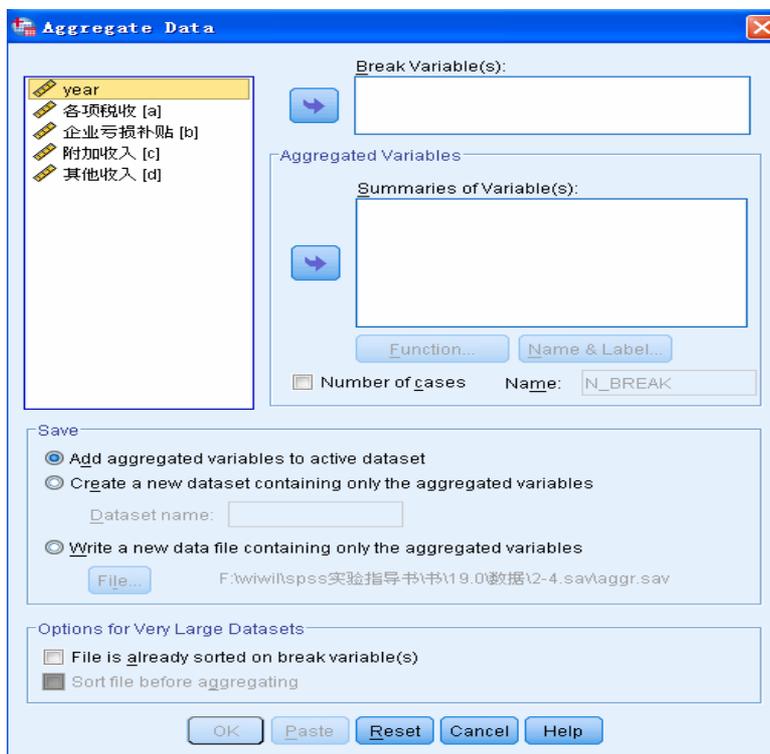
CONCEPT  
STRATE

- **分类变量 (Break Variable)**：它是对样本数据进行分类的变量，如按性别、民族、行业性质等进行分类汇总。这种变量可以为数值型或字符型变量。
- **汇总变量 (Aggregate Variable)**：利用源数据文件中的变量和相应的汇总函数产生的新变量名称及其表达式。这种变量必须要求为数值型变量。例如，按性别统计年收入的平均值，此时“性别”是汇总变量，“每人年收入”是源变量，“不同性别的年收入均值”就属于汇总变量。

# Step01: 打开数据汇总对话框



- 打开SPSS软件，选择菜单栏中的【File(文件)】→【Data(数据)】→【Aggregate(分类汇总)】命令，弹出【Aggregate Data(汇总数据)】对话框。



- **Step02:** 选择分类变量

在左侧的候选变量列表框中选择一个或多个变量作为分类变量，将其移入【Break Variable(s)(分组变量)】列表框中。

- **Step03:** 选择汇总变量

在左侧的候选变量列表框中选择一个或多个变量作为汇总变量，将其移入【Summaries of Variable(s)(变量摘要)】列表框中。

## Step04: 选择汇总函数

- 在【Summaries of Variable(s)(变量摘要)】列表框中可以选择相应汇总变量，此时可以单击下方的【Function】按钮，打开如下图所示的对话框。

**Aggregate Data: Aggregate Function**

Summary Statistics	Specific Values	Number of cases
<input checked="" type="radio"/> Mean	<input type="radio"/> First	<input type="radio"/> Weighted
<input type="radio"/> Median	<input type="radio"/> Last	<input type="radio"/> Weighted missing
<input type="radio"/> Sum	<input type="radio"/> Minimum	<input type="radio"/> Unweighted
<input type="radio"/> Standard Deviation	<input type="radio"/> Maximum	<input type="radio"/> Unweighted missing

**Percentages**

Above

Below Value:

Inside

Outside Low:  High:

**Fractions**

Above

Below Value:

Inside

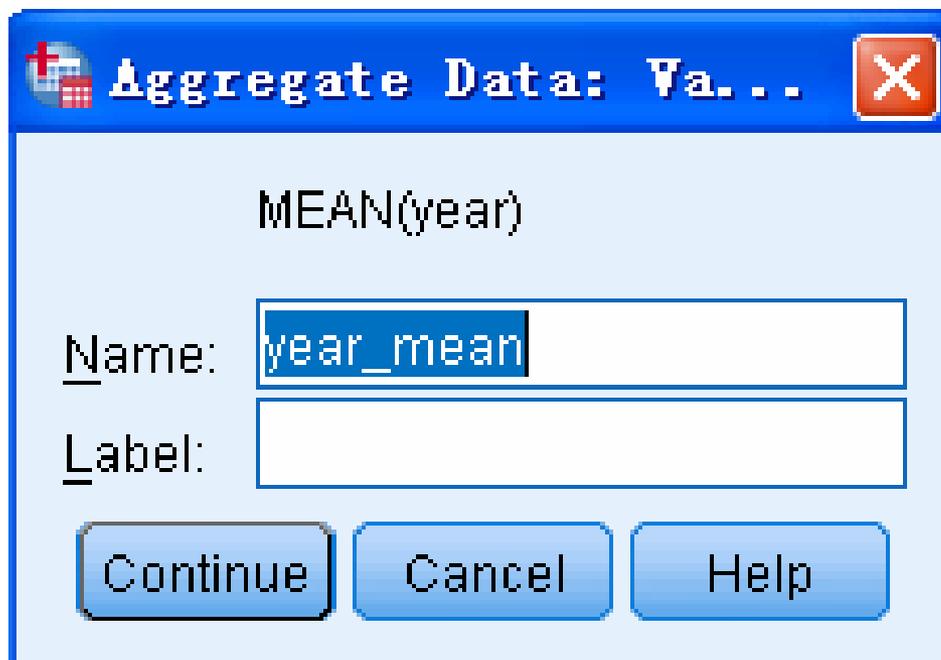
Outside Low:  High:

Continue Cancel Help

## Step05: 添加变量标签

CONCEPT  
STRATE

在【Summaries of Variable(s) (变量摘要)】列表框中可以选择相应汇总变量，此时可以单击下方的【Name and Label】按钮，弹出如下图所示的对话框。



- **Step06:** 选择汇总结果保存方式  
在【save(保存)】选项组中可以选择汇总结果的保存方式。
- **Step07:** 大规模数据的排序选择  
勾选【Options for Very Large Datasets (适用于大型数据集的选项)】复选框，可以对含有大规模数据的数据文件进行汇总之前的排序工作，这样能使得后续操作更有效率。
- **Step08:** 完成上述操作后，单击【OK】按钮，操作结束。



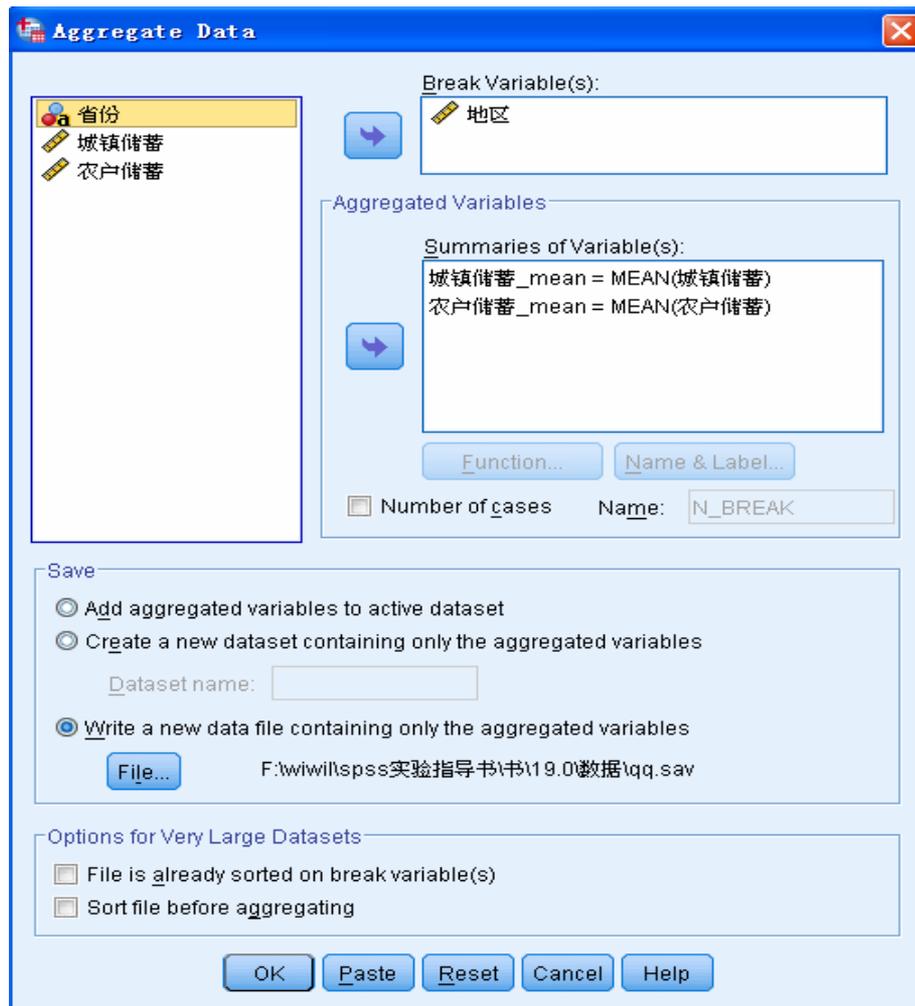
## 2. 实例内容：城乡居民人民币储蓄存款

- 下图是我国部分省份2004年度城乡居民的人民币储蓄存款金额（年底余额, 单位: 亿元）。

	省份	城镇储蓄	农户储蓄	地区
1	辽宁	5339.0	709.6	1
2	吉林	2114.6	291.0	1
3	黑龙江	3268.9	316.6	1
4	江苏	7622.0	1241.1	2
5	浙江	5558.0	1806.0	2
6	安徽	2355.7	616.7	2
7	河南	4338.8	1268.5	3
8	湖北	3340.7	525.0	3
9	湖南	2698.0	785.3	3
10	四川	3999.8	1019.6	4
11	贵州	911.0	183.5	4
12	云南	1679.2	372.9	4
13	陕西	2459.2	489.2	5
14	甘肃	1193.2	191.7	5
15	青海	278.1	21.2	5



# Step01: 打开对话框



- **Step02:** 选择分类变量

从对话框左侧的候选变量列表框中选择“省份”变量作为分类变量，将其移入【Break Variable(s)(分组变量)】列表框中。

- **Step03:** 选择汇总变量

从对话框左侧的候选变量列表框中选择“城镇储蓄”和“农户储蓄”作为汇总变量，将其移入【Summaries of Variable(s)(变量摘要)】列表框中。由于这里主要是比较存款金额的高低水平，因此选择系统默认的平均值函数。



- **Step04:** 选择汇总结果保存方式

在【Save(保存)】选项组中点选【Write a new data file containing only the aggregated variables】单选按钮，其目的是新建aggr.sav的外部数据文件保存汇总结果。

- **Step05:** 单击【OK(确定)】按钮完成操作。

地区	城镇储蓄_mean	农户储蓄_mean
1	3574.17	439.07
2	5178.57	1221.27
3	3459.17	659.60
4	2196.67	525.33
5	1310.17	234.03

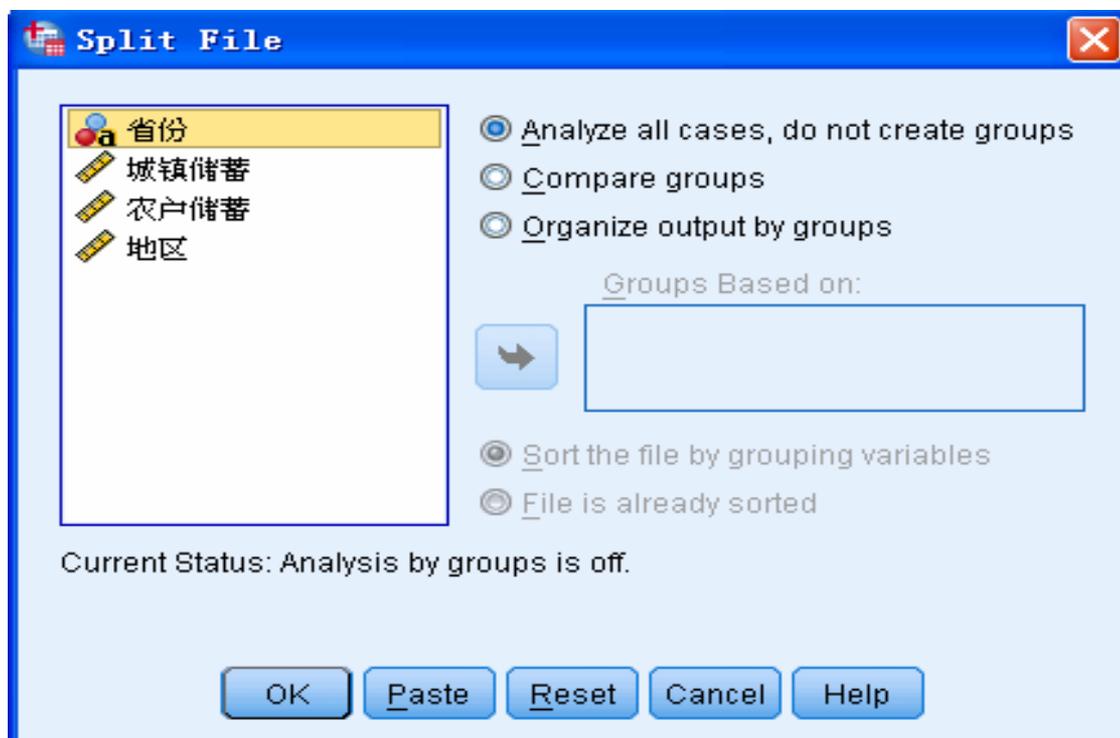
## 2.3.5 数据文件的拆分Split

CONCEPT  
STRATE

### 1.数据分类汇总的SPSS操作详解

Step01: 打开数据拆分对话框

- 打开SPSS软件，选择菜单栏中的【File(文件)】→【Data(数据)】→【Split File(拆分文件)】命令，弹出【Split File(拆分文件)】对话框。



- **Step02:** 选择数据拆分方式。
- **Step03:** 选择拆分变量 。
- **Step04:** 单击 **【OK】** 按钮，操作结束。

**注意：**拆分后的文件在保存之后，下次调用该文件时，拆分结果仍然有效。当不需要分组时，可以按上述操作，点选 **【Analyze all cases, do not create groups (分析所有个案，不创建组)】** 单选钮。



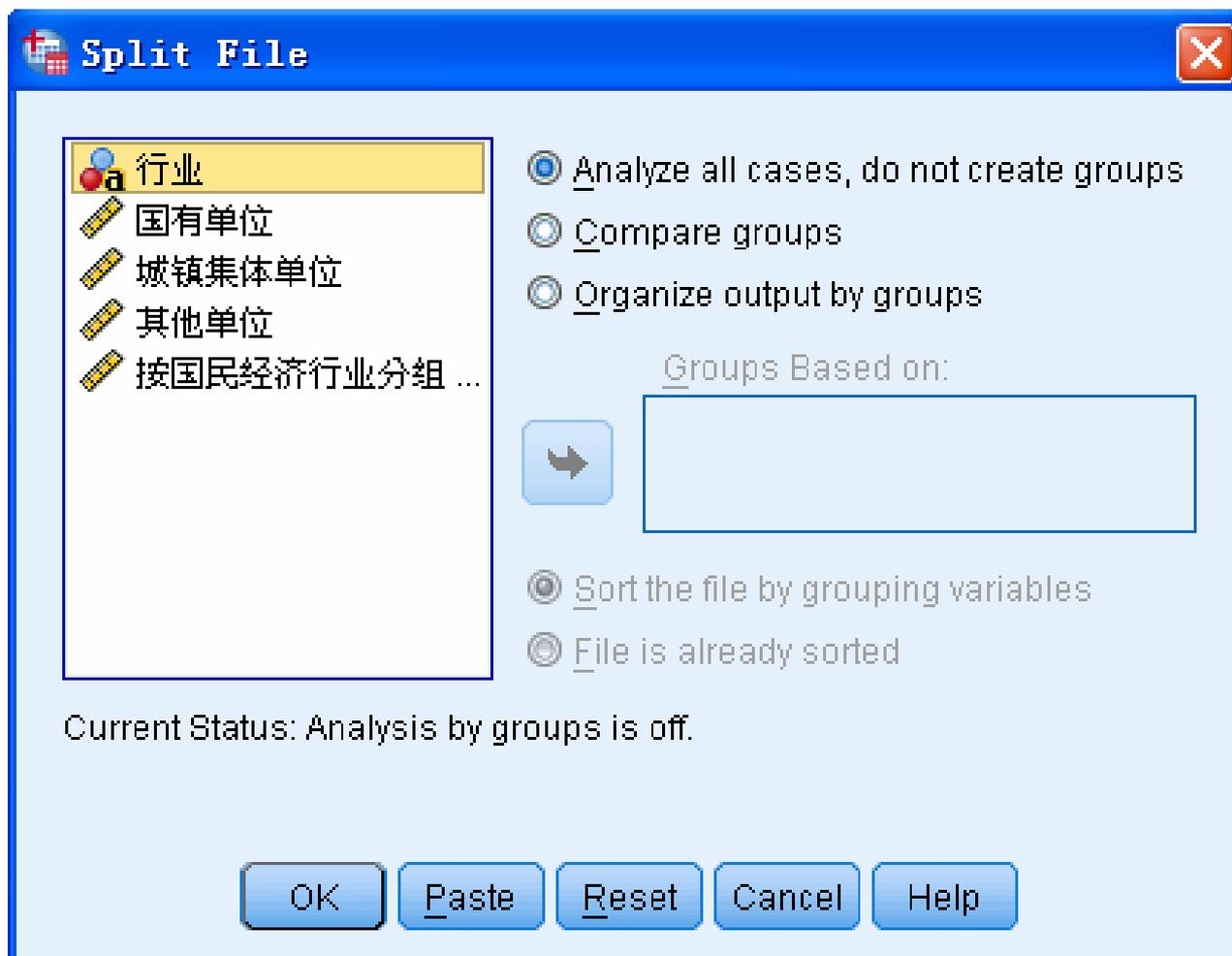
## 2. 实例内容：分行业职工平均工资

下图是2005年我国部分按细行业划分的职工平均工资，请根据不同的行业类型，对原始数据进行拆分，数据详见2-7.sav。

	行业	国有单位	城镇集体单位	其他单位	分类
1	畜牧业	7051.00	8170.00	14125.00	1
2	林业	7218.00	7620.00	8757.00	1
3	农业	7499.00	6671.00	8622.00	1
4	纺织业	9316.00	7802.00	11020.00	3
5	食品制造业	10130.00	8841.00	14346.00	3
6	渔业	10277.00	9865.00	12722.00	1
7	饮料制造业	11198.00	7852.00	14313.00	3
8	道路运输业	14308.00	9963.00	15847.00	4
9	城市公共交通运输业	16696.00	11562.00	17520.00	4
10	煤炭开采和洗选业	18821.00	11917.00	20654.00	2
11	黑色金属矿采选业	19983.00	9139.00	17125.00	2
12	铁路运输业	24437.00	9008.00	30085.00	4
13	石油和天然气开采业	30265.00	8364.00	31919.00	2
14	烟草制品业	45656.00	21250.00	27552.00	3

# Step01: 打开对话框

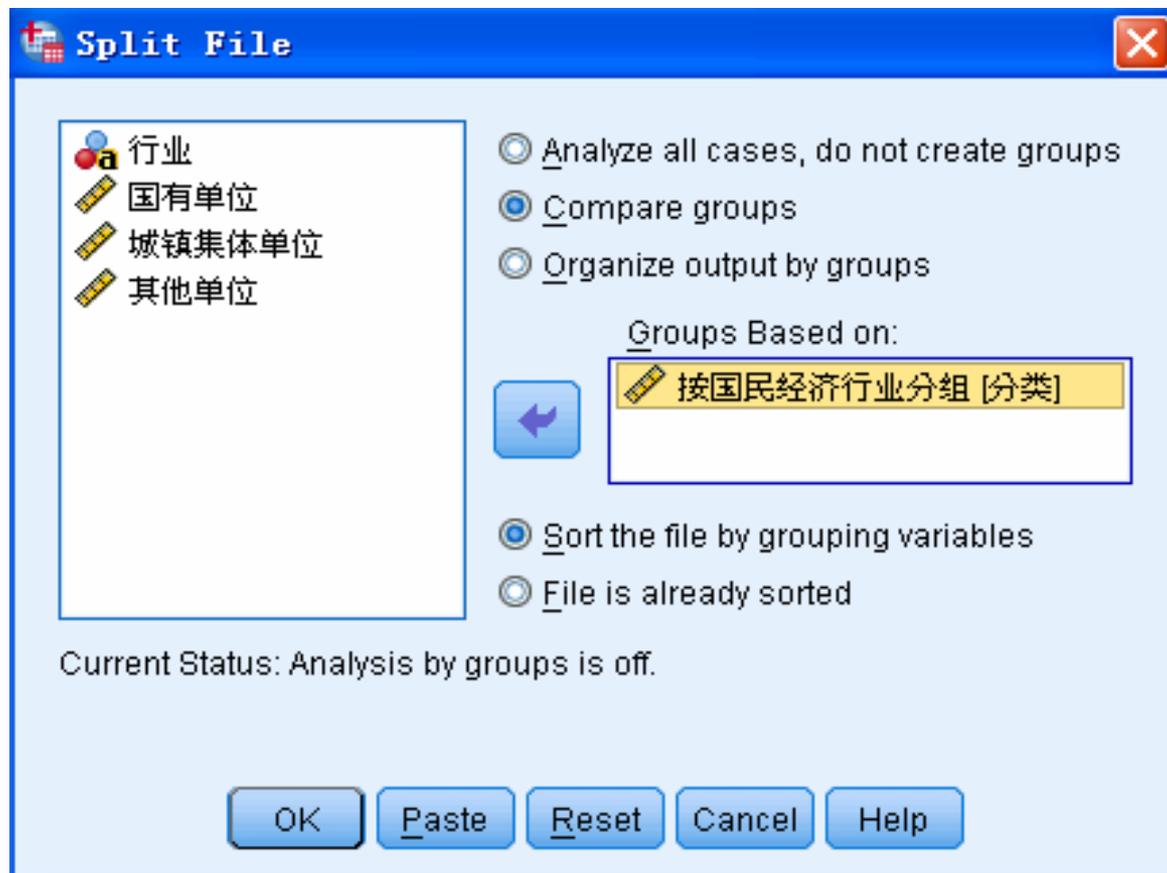
CONCEPT  
STRATE



# Step02: 选择数据拆分方式

## Step03: 选择拆分变量

CONCEPT  
STRATE



# Step04: 完成操作

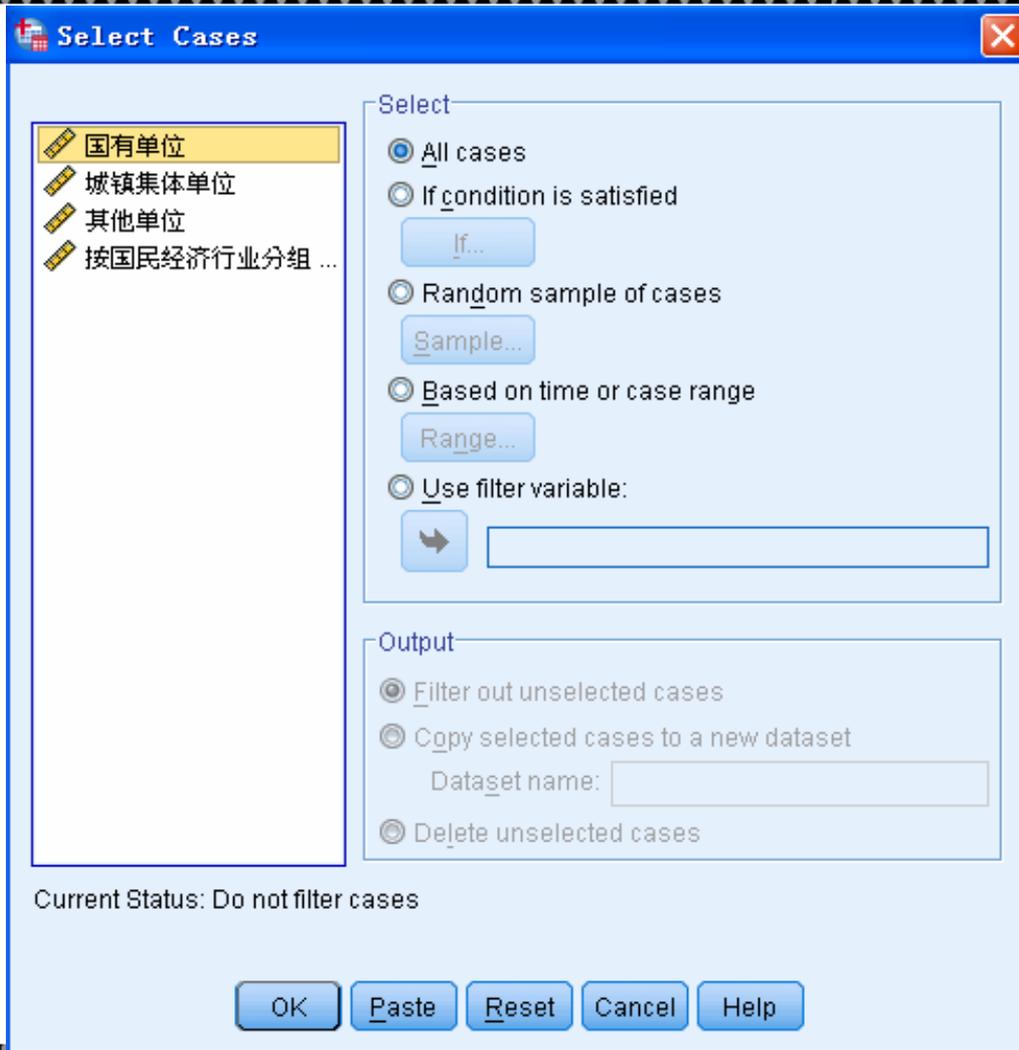
CONCEPT  
STRATE

	行业	国有单位	城镇集体单位	其他单位	分类
1	畜牧业	7051.00	8170.00	14125.00	1
2	林业	7218.00	7620.00	8757.00	1
3	农业	7499.00	6671.00	8622.00	1
4	渔业	10277.00	9865.00	12722.00	1
5	煤炭开采和洗选业	18821.00	11917.00	20654.00	2
6	黑色金属矿采选业	19983.00	9139.00	17125.00	2
7	石油和天然气开采业	30265.00	8364.00	31919.00	2
8	纺织业	9316.00	7802.00	11020.00	3
9	食品制造业	10130.00	8841.00	14346.00	3
10	饮料制造业	11198.00	7852.00	14313.00	3
11	烟草制品业	45656.00	21250.00	27552.00	3
12	道路运输业	14308.00	9963.00	15847.00	4
13	城市公共交通运输业	16696.00	11562.00	17520.00	4
14	铁路运输业	24437.00	9008.00	30085.00	4

## 2.3.6 选择数据：城市设施水平

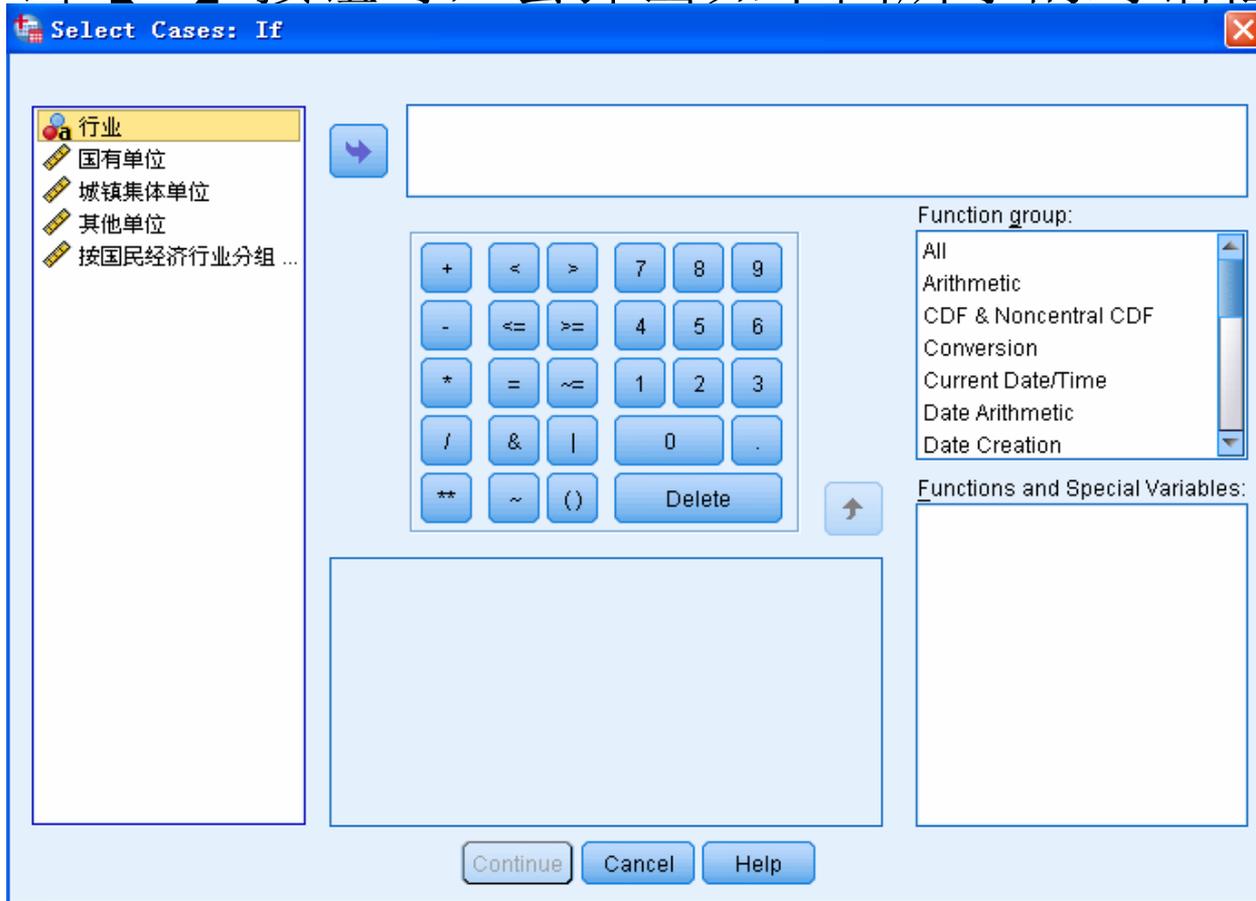
CONCEPT  
STRATE

- **1.操作详解**
- **Step01:** 打开数据选择对话框  
打开SPSS软件，在菜单栏中选择【File(文件)】→【Data(数据)】→【Select Cases(选择个案)】命令，弹出【Select Cases(选择个案)】对话框。

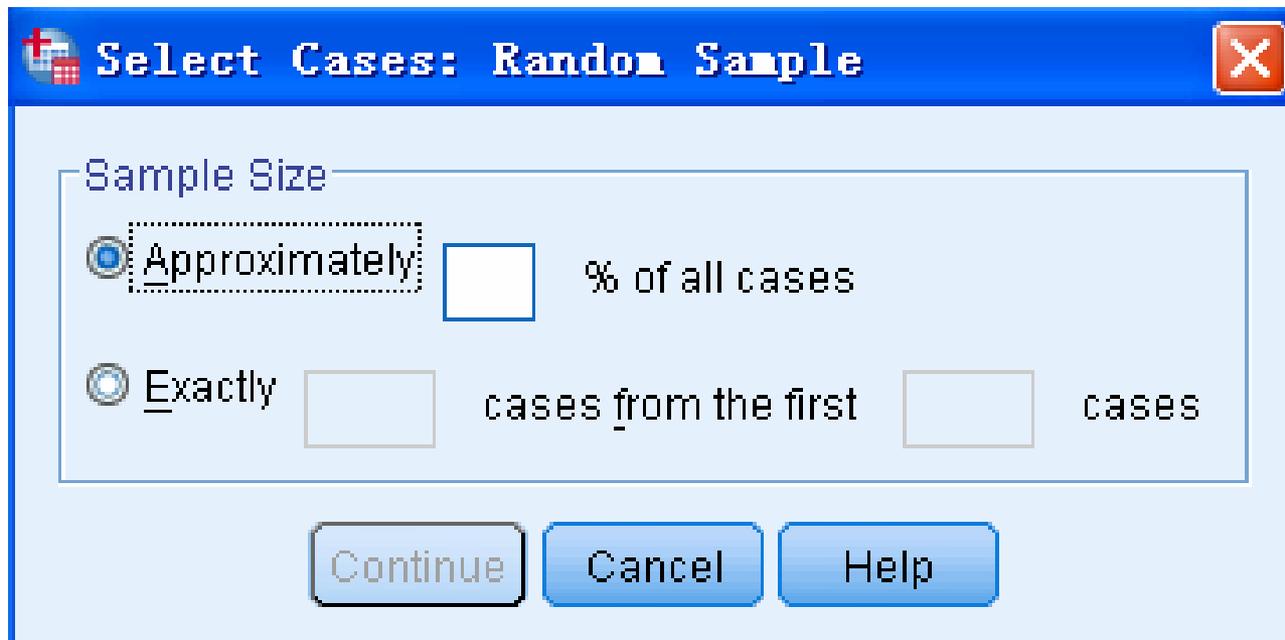


## Step02: 选择数据选择方式

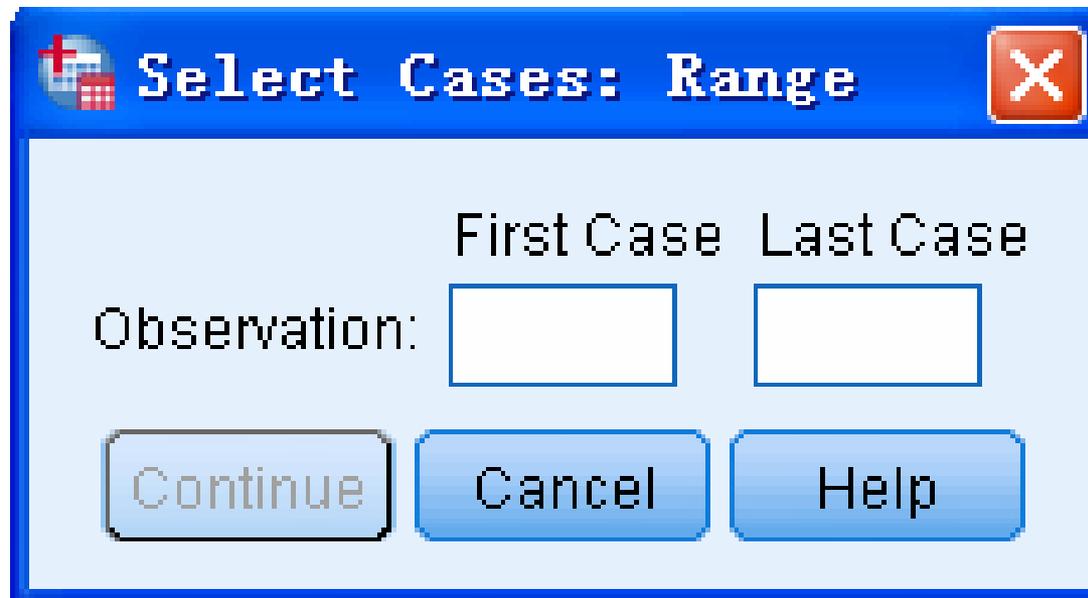
- 单击【If】按钮时，会弹出如下图所示的对话框。



- 单击【Sample】按钮，弹出如下图所示的对话框。



- 单击【Range】按钮，弹出如下图所示的对话框。





- **Step03:** 选择输出方式

在【**Select Cases (选择个案)**】对话框的【**Output (输出)**】选项组中可以选择变量的输出方式。

- **Step04:** 单击【**OK**】按钮，操作结束。

## 2. 实例内容：城市设施水平

CONCEPT  
RATE

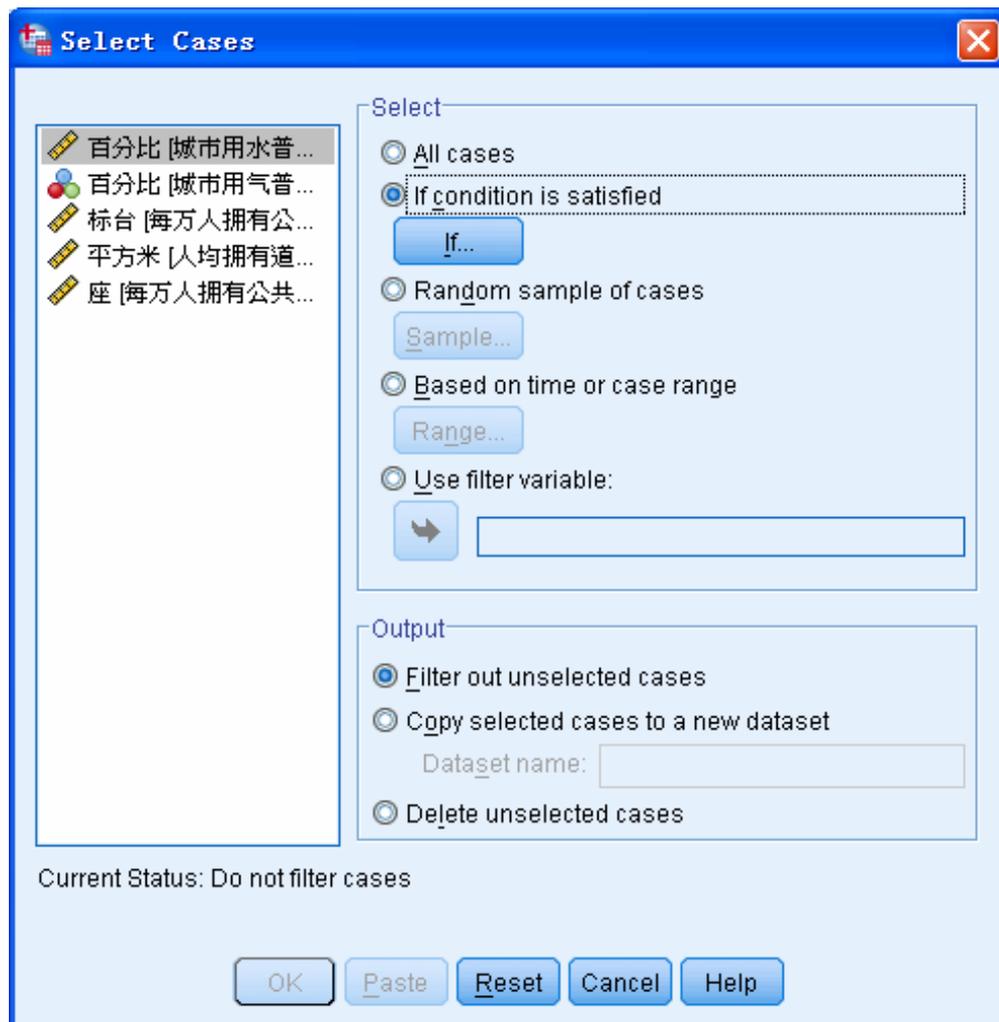
数据文件2-8.sav中是2006年我国部分地区城市设施水平指标，包括城市用水普及率、城市燃气普及率等。请根据这些原始数据，按照以下条件选择数据。

条件一：选择城市用水普及率和城市燃气普及率都大于90%的地区。

条件二：随机选取10个地区。



# (条件一) Step01: 打开对话框



## Step02: 设置数据选择方式

CONCEPT  
STRATE

- 点选【If condition is satisfied(如果条件满足)】选项，表示选择满足题目要求条件的观测量。同时单击【If】按钮，弹出条件选择对话框。

# Step03: 设置选择条件





# Step04: 完成操作

	地区	城市用水普及率	城市用气普及率	每万人拥有公交车台数	人均拥有道路面积	人均公共绿地面积	每万人拥有公共厕所	filter_\$
1	北京	123.36	113.84	22.19	7.40	10.68	3.82	1
2	天津	100.26	99.22	14.23	13.98	6.59	3.40	1
3	河北	92.01	86.96	8.05	12.38	7.87	3.97	0
4	山西	89.58	74.56	5.73	9.06	6.63	4.45	0
5	内蒙古	80.67	71.03	6.08	10.34	9.39	6.35	0
6	辽宁	92.14	87.95	9.28	8.51	7.93	4.07	0
7	吉林	80.53	75.03	7.65	8.51	7.34	4.92	0
8	黑龙江	79.20	70.72	8.72	8.47	7.29	7.74	0
9	上海	100.00	105.25	12.52	11.84	7.33	1.54	1
10	江苏	81.99	80.20	8.61	15.45	9.60	3.40	0
11	浙江	70.96	70.13	9.33	12.21	6.99	2.80	0
12	安徽	89.93	76.09	7.70	12.28	7.28	2.22	0
13	福建	78.37	76.73	9.04	9.63	7.51	1.69	0
14	江西	91.34	77.26	8.06	9.61	7.74	1.95	0
15	山东	97.17	94.47	10.50	18.14	12.77	1.97	1
16	河南	87.16	63.23	7.09	10.00	7.93	3.20	0
17	湖北	91.45	83.69	10.55	12.06	8.34	2.55	0
18	湖南	90.27	75.95	8.98	10.01	6.99	2.35	0
19	广东	76.60	71.02	5.74	9.65	9.25	1.13	0
20	广西	79.85	72.95	7.41	10.77	7.58	1.79	0
21	海南	80.40	70.90	7.85	14.21	10.85	1.29	0
22	重庆	81.38	75.84	9.29	8.14	6.45	2.64	0
23	四川	80.83	71.82	8.24	9.46	7.74	2.41	0
24	贵州	84.24	60.41	5.81	5.35	5.49	2.22	0
25	云南	74.46	57.37	9.69	7.47	6.47	1.90	0
26	西藏	48.63	48.63	15.47	16.44	9.21	6.44	0
27	陕西	85.66	71.24	9.10	9.07	5.89	1.92	0
28	甘肃	88.66	57.06	6.08	11.34	6.95	1.89	0

## 条件二

- 条件二属于随机选择的问题，因此需要点选【Random samples of cases (随机个案样本)】单选钮，同时在弹出的【Select Cases: Random Sample(选择个案：随机样本)】对话框的“Exactly \_cases form the first\_cases”文本框中分别输入10和31，表示从31个观测量中选择10个观测量。最后，单击【Continue】按钮返回主对话框，随机选取的样本结果如下页所示。

**Select Cases: Random Sample**

Sample Size

Approximately  % of all cases

Exactly  cases from the first  cases

Continue Cancel Help

	地区	城市用水普及率	城市用气普及率	每万人拥有公交车台数	人均拥有道路面积	人均公共绿地面积	每万人拥有公共厕所	filter_\$
1	北京	123.36	113.84	22.19	7.40	10.68	3.82	1
<del>2</del>	<del>天津</del>	<del>100.26</del>	<del>99.22</del>	<del>14.23</del>	<del>13.98</del>	<del>6.59</del>	<del>3.40</del>	<del>0</del>
<del>3</del>	<del>河北</del>	<del>92.01</del>	<del>86.96</del>	<del>8.05</del>	<del>12.38</del>	<del>7.87</del>	<del>3.97</del>	<del>0</del>
4	山西	89.58	74.56	5.73	9.06	6.63	4.45	1
5	内蒙古	80.67	71.03	6.08	10.34	9.39	6.35	1
<del>6</del>	<del>辽宁</del>	<del>92.14</del>	<del>87.95</del>	<del>9.28</del>	<del>8.51</del>	<del>7.93</del>	<del>4.07</del>	<del>0</del>
7	吉林	80.53	75.03	7.65	8.51	7.34	4.92	1
<del>8</del>	<del>黑龙江</del>	<del>79.20</del>	<del>70.72</del>	<del>8.72</del>	<del>8.47</del>	<del>7.29</del>	<del>7.74</del>	<del>0</del>
9	上海	100.00	105.25	12.52	11.84	7.33	1.54	1
10	江苏	81.99	80.20	8.61	15.45	9.60	3.40	1
11	浙江	70.96	70.13	9.33	12.21	6.99	2.80	1
12	安徽	89.93	76.09	7.70	12.28	7.28	2.22	1
<del>13</del>	<del>福建</del>	<del>78.37</del>	<del>76.73</del>	<del>9.04</del>	<del>9.63</del>	<del>7.51</del>	<del>1.69</del>	<del>0</del>
14	江西	91.34	77.26	8.06	9.61	7.74	1.95	1
<del>15</del>	<del>山东</del>	<del>97.17</del>	<del>94.47</del>	<del>10.50</del>	<del>18.14</del>	<del>12.77</del>	<del>1.97</del>	<del>0</del>
<del>16</del>	<del>河南</del>	<del>87.16</del>	<del>63.23</del>	<del>7.09</del>	<del>10.00</del>	<del>7.93</del>	<del>3.20</del>	<del>0</del>
<del>17</del>	<del>湖北</del>	<del>91.45</del>	<del>83.69</del>	<del>10.55</del>	<del>12.06</del>	<del>8.34</del>	<del>2.55</del>	<del>0</del>
18	湖南	90.27	75.95	8.98	10.01	6.99	2.35	1
<del>19</del>	<del>广东</del>	<del>76.60</del>	<del>71.02</del>	<del>5.74</del>	<del>9.65</del>	<del>9.25</del>	<del>1.13</del>	<del>0</del>
20	广西	79.85	72.95	7.41	10.77	7.58	1.79	1
21	海南	80.40	70.90	7.85	14.21	10.85	1.29	1
<del>22</del>	<del>重庆</del>	<del>81.38</del>	<del>75.84</del>	<del>9.29</del>	<del>8.14</del>	<del>6.45</del>	<del>2.64</del>	<del>0</del>
23	四川	80.83	71.82	8.24	9.46	7.74	2.41	1
24	贵州	84.24	60.41	5.81	5.35	5.49	2.22	1

## 2.3.7 数据加权:蔬菜的平均价格

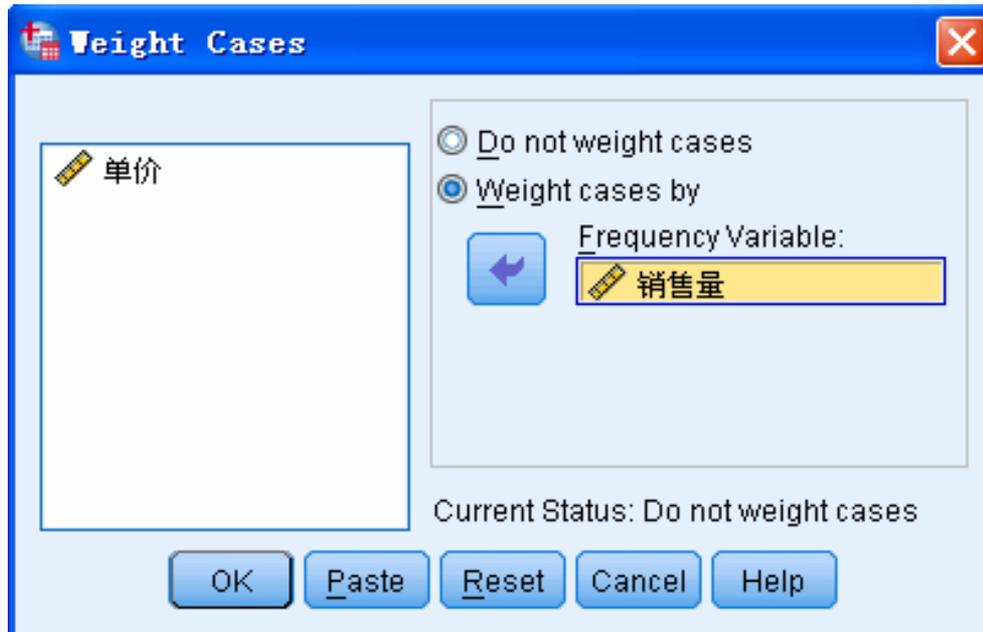
CONCEPT  
STRATE

- 权重是数据分析中的一个重要概念，它是一个相对的概念。权重的大小描述了该指标在整体评价中的相对重要程度。在数据处理中，常需要对数据进行加权处理。
- 在记录有大量数据的文件中，可能同一观测量值会反复出现，如性别、民族等。如果在建立数据文件时能定义一个频数变量，也称为权重变量，用它来代表相同观测量出现的次数，这样后续的统计分析工作就会极大的简化。

# 1. 数据加权的SPSS操作详解

- **Step01:** 打开数据加权对话框

打开SPSS软件，选择菜单栏中的【File(文件)】→【Data(数据)】→【Weight cases(加权个案)】命令，弹出【Weight cases(加权个案)】对话框。





## 2. 实例内容：蔬菜的平均价格

- 某经销商希望掌握菜市场的蔬菜销销售的平均价格，收集数据见数据文件2-9.sav。现请利用这些数据，求出这些蔬菜的平均价格。

	蔬菜	单价	销售量
1	萝卜	0.90	536.00
2	青菜	1.60	120.00
3	蘑菇	3.60	45.00
4	韭菜	2.00	60.00
5	花菜	1.80	40.00
6	豆腐	1.20	100.00
7	大白菜	0.60	300.00
8	油菜	1.50	150.00
9	土豆	0.90	200.00
10	西红柿	2.50	400.00

- Step1: 由于经销商要求掌握蔬菜的平均价格，如果仅仅只用蔬菜的单价进行简单的算术平均是很不合理的，这是因为不同蔬菜的销售量不同，所以要考虑销售量对平均价格的影响。因此，我们以蔬菜的销售量为权重计算各种蔬菜销售的平均价格更为合适。

这里选择“销售量”变量作为权重变量，将其放入【Frequencies Variable (频率变量)】列表框中，此时就可以进行后续的求平均值工作了。

- **Step02:** 选择变量是否加权，用户首先选择是否对观测量进行加权。
  - **Do not weight cases:** 不对观测量加权，系统默认项。
  - **Weight cases by:** 对观测量加权，同时从左侧的候选变量列表框中选择权重变量移入【**Frequency Variable (频率变量)**】列表框中。
- **Step03:** 单击【**OK**】按钮，操作结束。

## 2.4 SPSS数据的计算和变换

CONCEPT  
RATE

- 在数据分析中，经常要根据一些已知的数据变量计算新的变量。例如，根据历年的产量数据资料计算产量的发展速度，根据人口数据计算人口出生率、死亡率等。不仅如此，还需要进行不同类型变量之间的转换，如将数值型变量转化为字符型变量。这些工作都需要利用【Transform (转换)】菜单中的相关命令。

## 2.4.1 变量计算：国内生产总值的产业结构

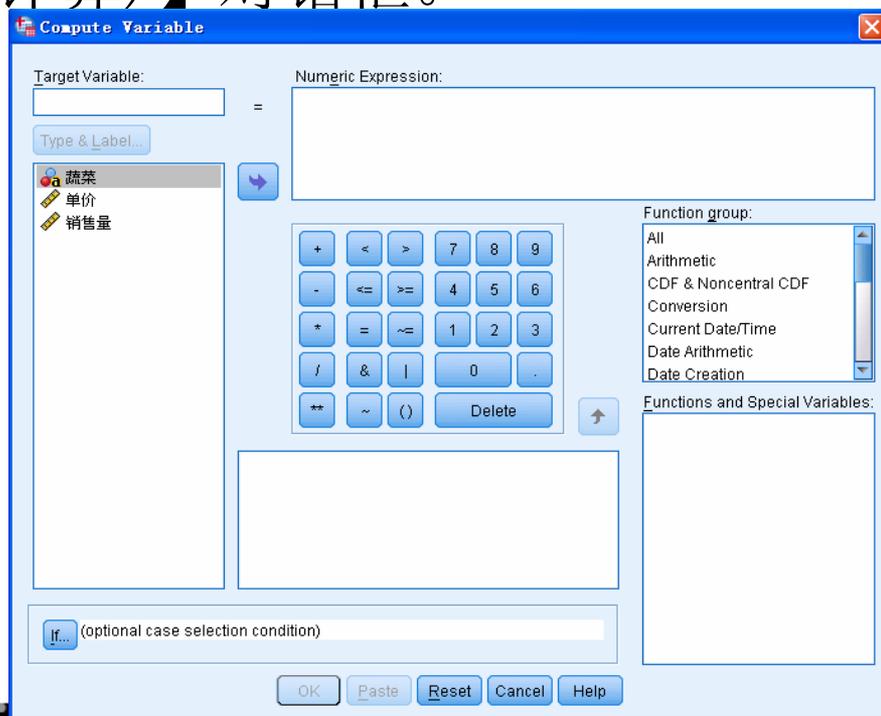
CONCEPT  
RATE

- 变量计算是数据分析中的重要内容之一。有些时候，收集到的原始数据并不能直接提供给我们许多有用的信息，此时，我们需要将原始数据进行计算变换，生成有用的新的变量。例如，根据职工的基本工资、各类保险、公积金等，计算职工的实际月收入；根据购房客户的贷款总额和按揭方案评价客户的潜在风险等。

# 1. SPSS操作详解

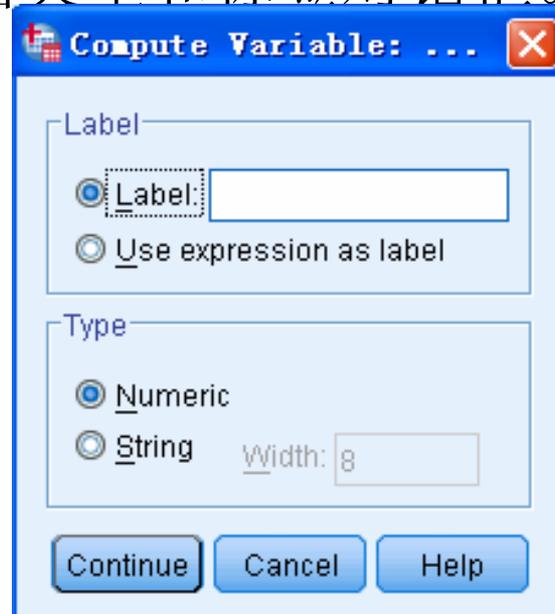
- **Step01:** 打开变量计算对话框

打开SPSS软件，选择菜单栏中的【File(文件)】→【Transform转换】→【Compute(计算)】命令，弹出【Compute(计算)】对话框。



## Step02: 定义新变量及其类型

- 在【Target Variable (目标变量)】文本框中用户需要定义目标函数名，它可以是一个新变量名，也可以是已经定义的变量名。单击下方的【Type&Label】按钮，弹出类型和标签对话框。



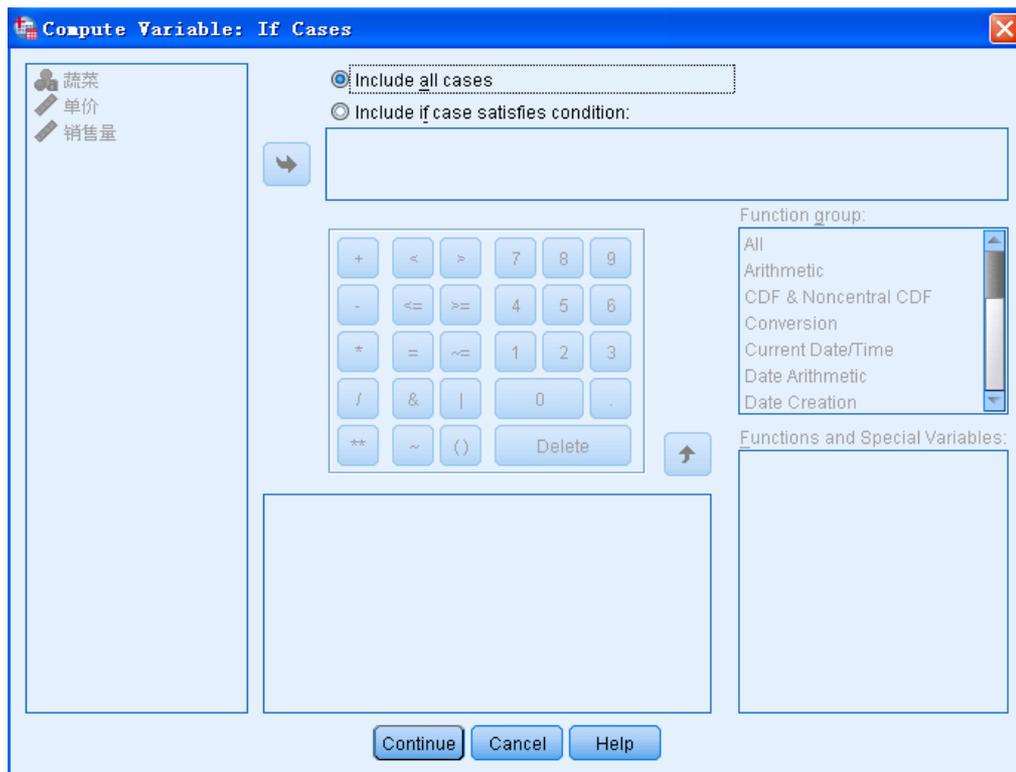
## Step03: 输入计算表达式

CONCEPT  
TRATE

- 可以使用计算器板或键盘将计算表达式输入到【**Nu  
meric Expression (数值表达式)**】文本中。如果用户需要调用函数，可以从右侧的【**Function(函数)**】列表中选择，系统提供了数学函数、逻辑函数、日期函数等。



- Step04: 条件样本选择  
单击【If】按钮，弹出的对话框如下图所示。



**Step05：结束操作**

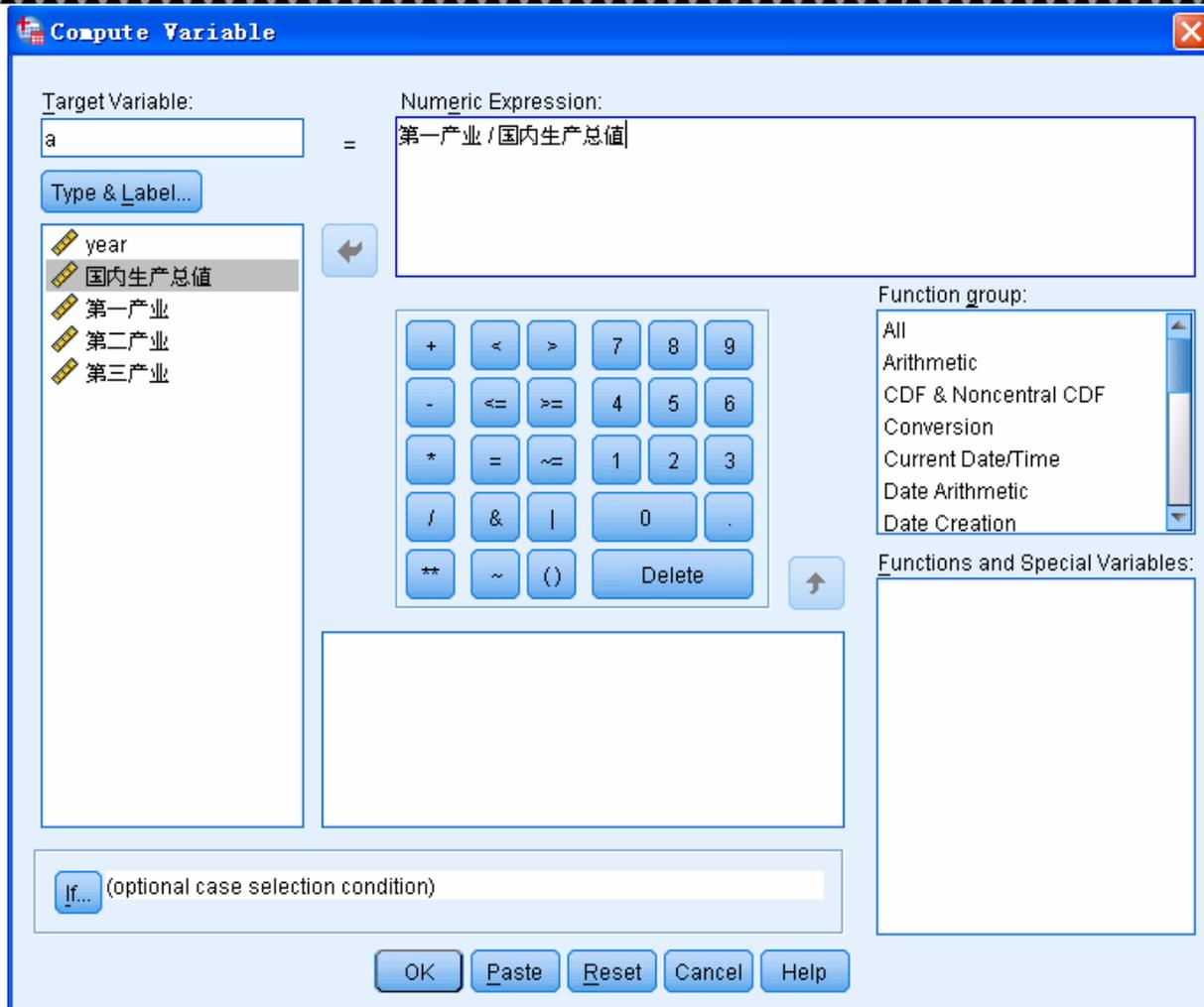
单击【OK】按钮，此时操作结束。

## 2.实例内容：国内生产总值的产业构成

CONCEPT  
TRATE

- 数据文件2-10.sav为我国1978-2005年国内生产总值、第一产业国内生产总值、第二产业国内生产总值和第三产业国内生产总值，请分析不同产业所占国内生产总值的变动情况。

# Step01: 打开对话框



- **Step02:** 定义第一产业比重变量

在【Target Variable(目标变量)】文本框中定义目标函数名为“a”，它表示第一产业生产总值所占总产值的比重。

- **Step03:** 计算第一产业生产总值所占比重

在【Numeric Expression(数值表达式)】文本框中输入计算表达式“a=第一产业/国内生产总值”。



## Step04: 完成操作

- 单击【OK(确定)】按钮，操作完成。此时，原数据文件新增加了“a”变量。

year	国内生产总值	第一产业	第二产业	第三产业	a
1978	3645.2	1018.4	1745.2	881.6	.28
1979	4062.6	1258.9	1913.5	890.2	.31
1980	4545.6	1359.4	2192.0	994.2	.30
1981	4891.6	1545.6	2255.5	1090.5	.32
1982	5323.4	1761.6	2383.0	1178.8	.33
1983	5962.7	1960.8	2646.2	1355.7	.33
1984	7208.1	2295.5	3105.7	1806.9	.32
1985	9016.0	2541.6	3866.6	2607.8	.28
1986	10275.2	2763.9	4492.7	3018.6	.27
1987	12058.6	3204.3	5251.6	3602.7	.27
1988	15042.8	3831.0	6587.2	4624.6	.25
1989	16992.3	4228.0	7278.0	5486.3	.25
1990	18667.8	5017.0	7717.4	5933.4	.27
1991	21781.5	5288.6	9102.2	7390.7	.24
1992	26923.5	5800.0	11699.5	9424.0	.22

## 2.4.2 变量重新赋值：空气质量等级划分

CONCEPT  
STRATE

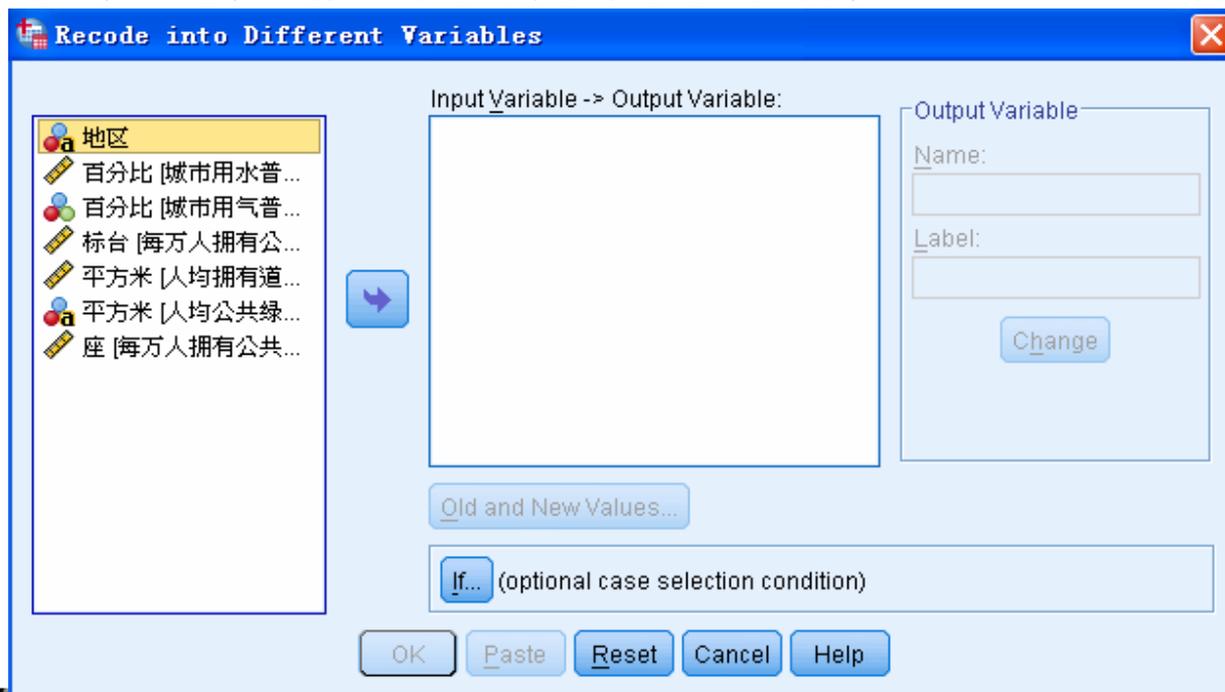
SPSS的【Transform(转换)】菜单中有【Recode into Same Variable(重新编码为相同变量)】和【Recode into Different Variable(重新编码为不同变量)】两个命令可以实现重新赋值功能，它们分别表示重新赋值到同一变量或不同变量。

下面以【Recode into Different Variable(重新编码为不同变量)】命令为例说明重新赋值功能。

# 1. SPSS操作详解

- **Step01:** 打开重新赋值对话框

选择菜单栏中的【File(文件)】→【Transform(转换)】→【Recode into Different Variable(重新编码为不同变量)】命令，弹出如下图所示的对话框。





## Step02: 选择重新赋值变量和输出变量

在候选变量列表框中选择要重新赋值的变量，将其移入【Input Variable->Output Variable (输入变量->输出变量)】列表框中，同时在【Output Variable (输出变量)】选项组中填写输出变量的名称【Name (名称)】及标签【Label (标签)】，单击【Change】按钮进行赋值转换。

## Step03: 设置重新赋值规则

CONCEPT  
STRATE

- **【Old and New Value】** 按钮被激活，单击此按钮，弹出如下图所示的对话框。

Recode into Different Variables: Old and New Values

Old Value

Value:

System-missing

System- or user-missing

Range:

through

Range, LOWEST through value:

Range, value through HIGHEST:

All other values

New Value

Value:

System-missing

Copy old value(s)

Old --> New:

Add

Change

Remove

Output variables are strings Width: 8

Convert numeric strings to numbers (5'>5)

Continue Cancel Help

- **Step04:** 选择样本赋值

如果用户不是对所有的候选变量进行赋值，而是选择其中符合某些条件的变量值进行赋值操作，此时需要单击【If】按钮进行操作。按照具体要求指定观察量的选择条件进行操作。

- **Step05:** 最后单击【OK】按钮，此时操作结束。

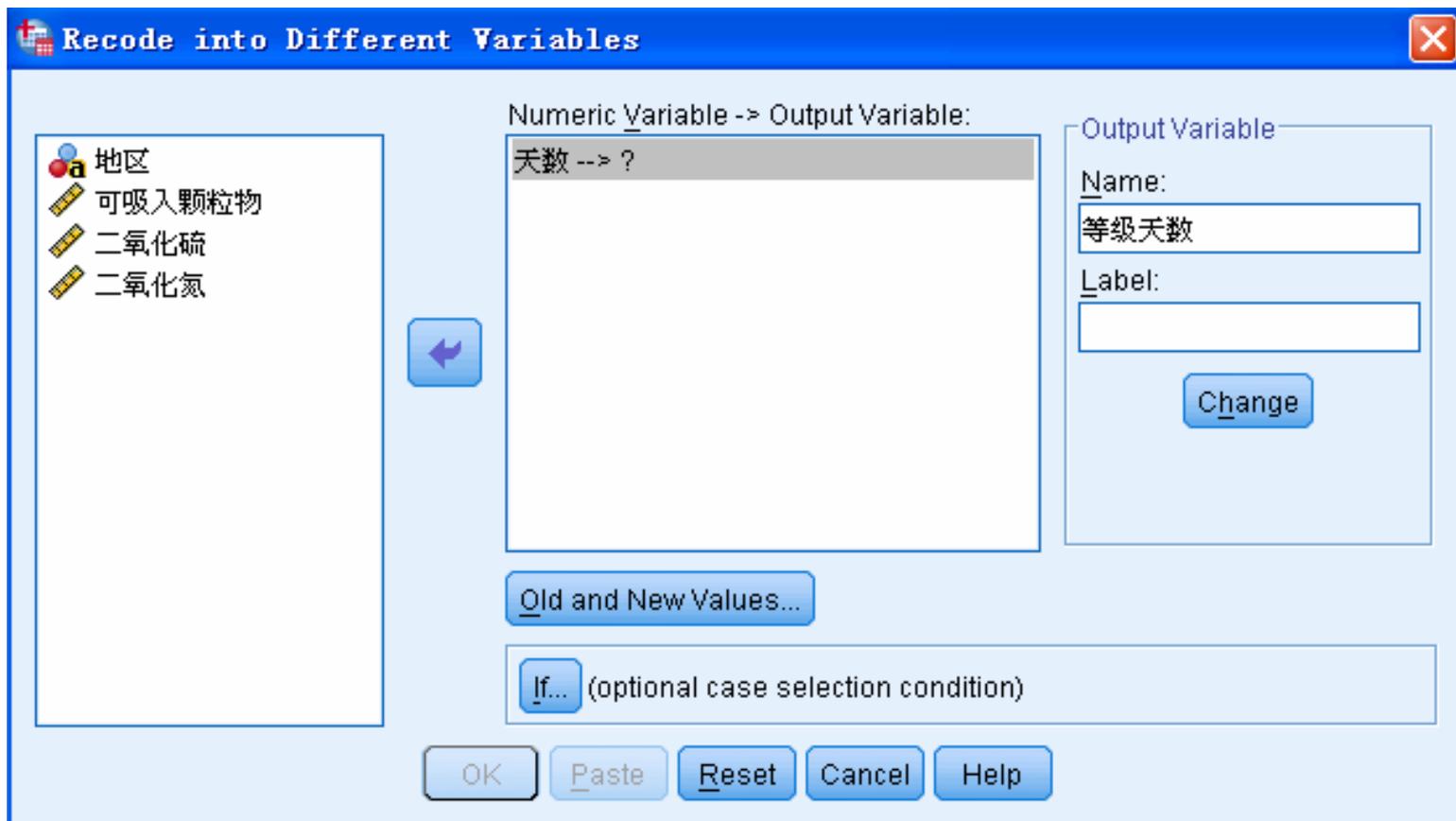
## 2.实例内容：空气质量等级的划分

- 下图是我国部分城市2005年空气质量的指标数据（见数据文件2-11.sav），请利用这个规则对不同城市的空气质量等级进行划分。

	地区	可吸入颗粒物	二氧化硫	二氧化氮	天数
1	北 京	.141	.050	.066	234
2	天 津	.106	.076	.047	298
3	石 家 庄	.132	.054	.041	283
4	太 原	.139	.077	.020	245
5	呼 和 浩 特	.097	.050	.041	312
6	沈 阳	.118	.054	.036	317
7	长 春	.099	.026	.035	340
8	哈 尔 滨	.104	.042	.056	301
9	上 海	.088	.061	.061	322
10	南 京	.110	.052	.054	304
11	杭 州	.112	.060	.058	301
12	合 肥	.095	.018	.025	329
13	福 州	.072	.016	.042	349
14	南 昌	.089	.050	.031	339
15	济 南	.128	.060	.024	262
16	郑 州	.109	.059	.039	300
17	武 汉	.119	.054	.050	271
18	长 沙	.122	.081	.036	245

# Step01: 打开对话框

CONCEPT  
TRATE





## Step02: 选择重新赋值变量和输出变量

- 在左侧的候选变量列表框中选择“天数”变量进入【Input Variable->Output Variable (输入变量->输出变量)】列表框，同时在【Output Variable (输出变量)】文本框中，填写输出赋值变量名称“等级天数”，同时单击【Change】按钮进行赋值转换。进行上述操作后，单击【Old and New Value】按钮。

# Step03: 设置赋值规则

CONCEPT  
STRATE

**Recode into Different Variables: Old and New Values**

**Old Value**

Value:

System-missing

System- or user-missing

Range:

through

Range, LOWEST through value:

Range, value through HIGHEST:

All other values

**New Value**

Value:

System-missing

Copy old value(s)

**Old --> New:**

320 thru Highest --> '一级天气'  
300 thru 320 --> '二级天气'  
280 thru 300 --> '三级天气'  
Lowest thru 280 --> '四级天气'

Output variables are strings    Width: 8

Convert numeric strings to numbers ('5' -> 5)

# Step04: 完成操作

CONCEPT  
STRATE

地区	可吸入颗粒物	二氧化硫	二氧化氮	天数	等级天数
北 京	.141	.050	.066	234	四级天气
天 津	.106	.076	.047	298	三级天气
石 家 庄	.132	.054	.041	283	三级天气
太 原	.139	.077	.020	245	四级天气
呼 和 浩 特	.097	.050	.041	312	二级天气
沈 阳	.118	.054	.036	317	二级天气
长 春	.099	.026	.035	340	一级天气
哈 尔 滨	.104	.042	.056	301	二级天气
上 海	.088	.061	.061	322	一级天气
南 京	.110	.052	.054	304	二级天气
杭 州	.112	.060	.058	301	二级天气
合 肥	.095	.018	.025	329	一级天气
福 州	.072	.016	.042	349	一级天气
南 昌	.089	.050	.031	339	一级天气
济 南	.128	.060	.024	262	四级天气
郑 州	.109	.059	.039	300	二级天气
武 汉	.119	.054	.050	271	四级天气

## 2.4.3 变量值计数： 消费价格指数的上涨项目

CONCEPT  
TRATE

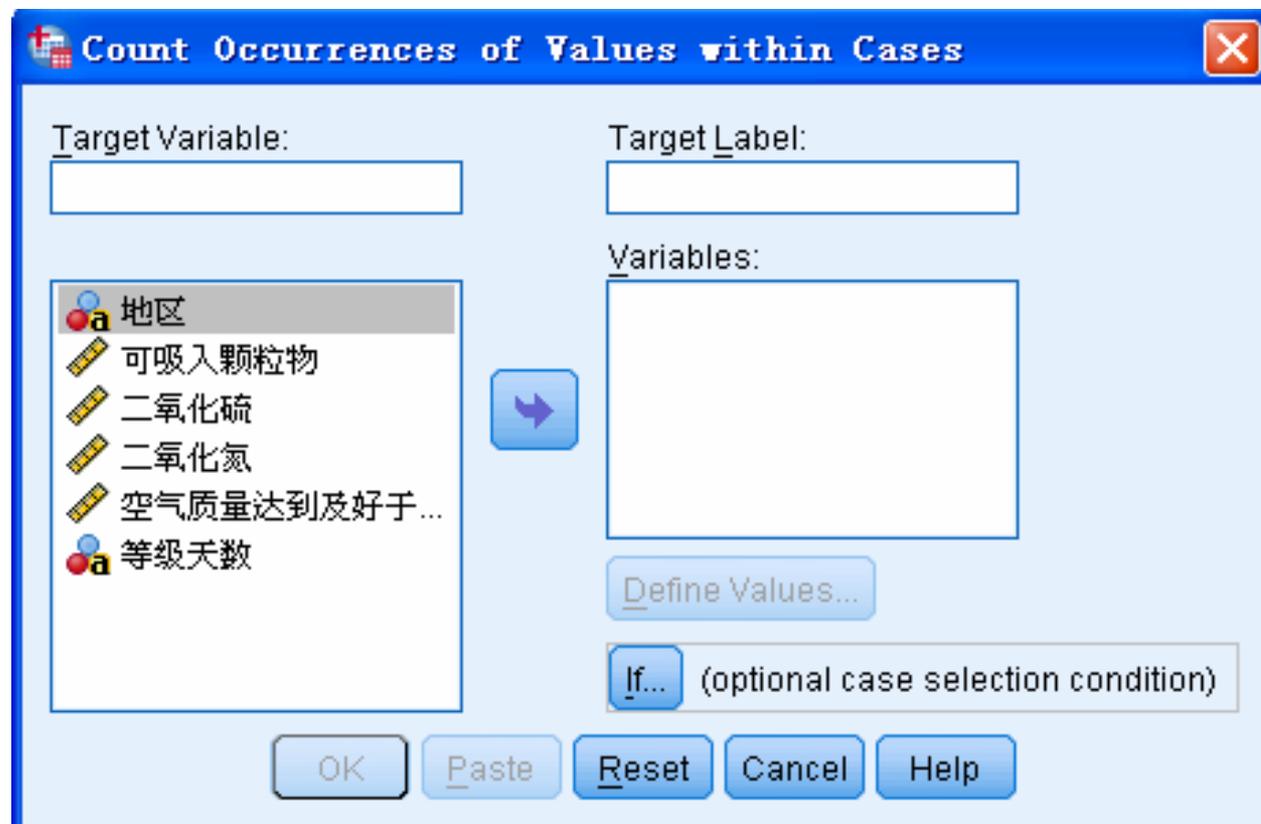
- 数据分析中，常常需要计算一些变量在同一个观测量中满足要求的特定变量值出现的次数。例如在进行产品市场调查时，要了解在所有的调查客户中有多少人使用过该产品，这就可以采用变量值计数功能来实现。

# 1.SPSS操作详解

CONCEPT  
STRATE

- **Step01:** 打开重新赋值对话框

打开SPSS软件，选择菜单栏中的【File(文件)】→【Transform(转换)】→【Count Values within Cases(对个案内的值计数)】命令，弹出【Count Occurrences of Values within Cases(计算个案内值的出现次数)】对话框。



- **Step02:** 输入目标计数变量

在【Target Variable (目标变量)】文本框中输入需要计数的变量名称，同时在【Target Label (目标标签)】文本框中填写计数变量的标签，便于注释说明。

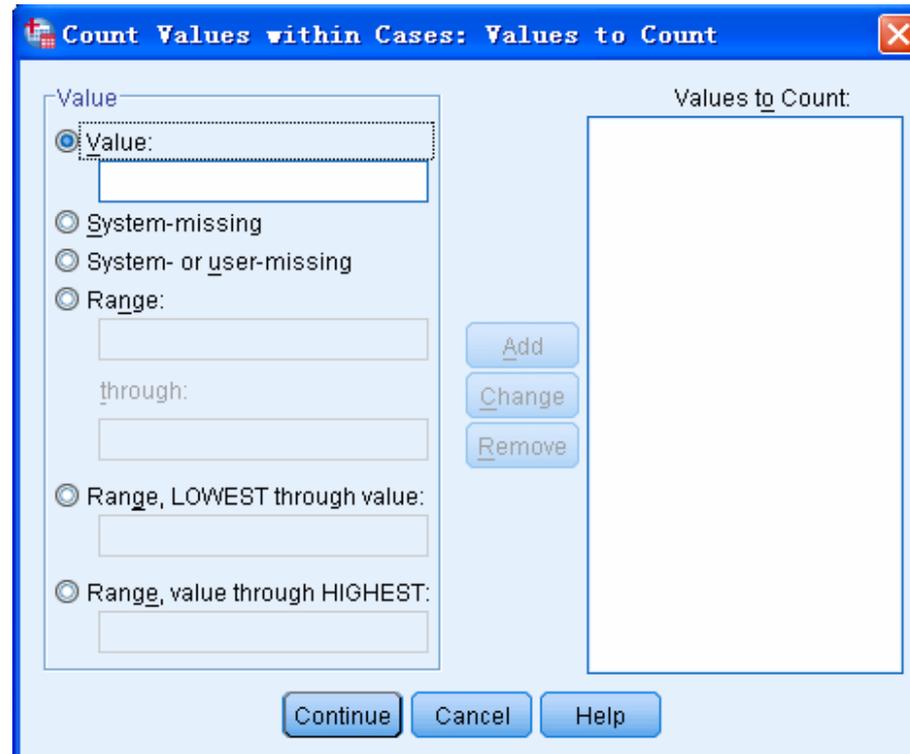
- **Step03:** 选择计数变量

在左侧的候选变量列表框中选择计数变量，将其移入右侧的【Variables (变量)】列表框中。需要注意，凡移入该列表框的变量必须具有相同的类型，当移入变量为数值型变量时，该栏标题改为“Number Variables”；当移入变量为字符型变量时，标题改为“String Variables”。

## Step04: 设置计数规则

CONCEPT  
TRATE

- 进行上述操作后，【Define Values】按钮被激活，单击此按钮，弹出如下图所示的对话框。





- **Step05:** 选择样本计数

如果用户不是对所有的候选变量进行计数，而是选择其中符合某些条件的变量值才进行计数操作，此时需要单击【If】按钮，按照具体要求指定观察量的选择条件进行操作。

**Step06:** 最后单击【OK】按钮，此时操作结束。

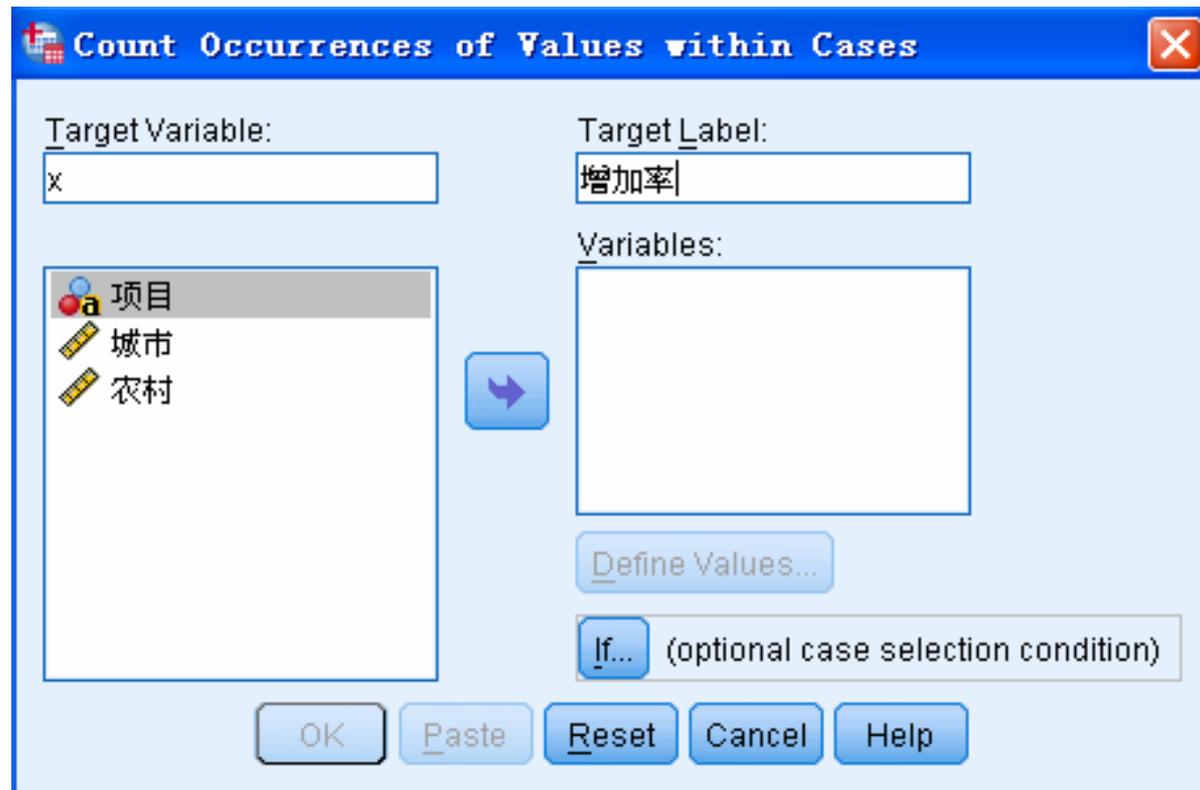


## 2.实例内容：消费价格指数的上涨项目

- 我国城市和农村居民消费价格分类指数数据见数据文件2-12.sav。由于不同产品的价格涨跌不同，请找出城市和农村居民消费价格指数都较去年上涨超过1%的项目。



# Step01: 打开对话框



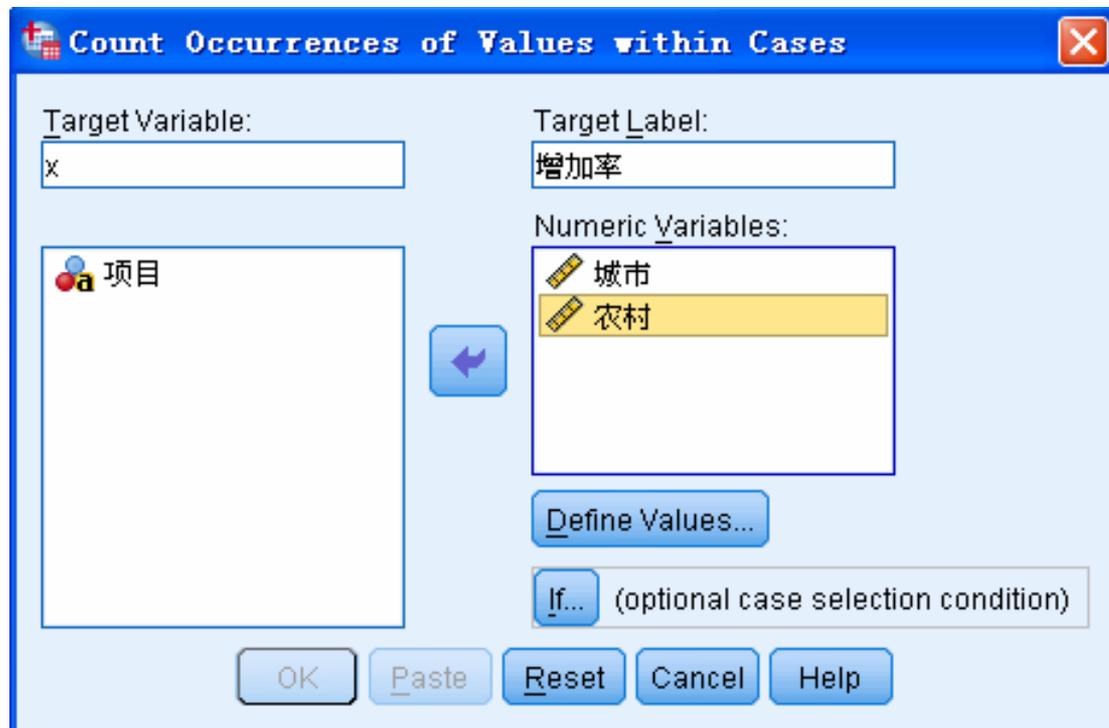
## Step02: 输入目标计算变量

CONCEPT  
RATE

- 在【Target Variable(目标变量)】文本框中输入需要计数的变量名称“x”，同时在【Target Label(目标标签)】文本框中填写标签“增加率”。



## Step03: 选择计数变量





# Step04: 设置计数规则

**Count Values within Cases: Values to Count**

Value

Value:  
[ ]

System-missing

System- or user-missing

Range:  
[ ]  
through:  
[ ]

Range, LOWEST through value:  
[ ]

Range, value through HIGHEST:  
101

Values to Count:  
101 thru Highest

Add  
Change  
Remove

Continue Cancel Help

# Step05: 完成操作



项目	城市	农村	x
粮食	101.5	101.3	2.00
油脂	94.0	94.4	.00
肉禽及其制品	102.1	103.1	2.00
蛋	104.2	105.4	2.00
水产品	106.1	105.5	2.00
菜	110.0	106.8	2.00
调味品	101.2	101.5	2.00
糖	102.9	105.6	2.00
茶及饮料	100.1	100.2	.00
干鲜瓜果	102.8	100.6	1.00
糕点饼干面包	100.8	101.2	1.00
奶及奶制品	100.7	101.9	1.00
烟草	100.4	100.4	.00
酒	100.4	100.8	.00
服装	97.9	98.7	.00

## 2.4.4 观测量求秩： 对外直接投资净额

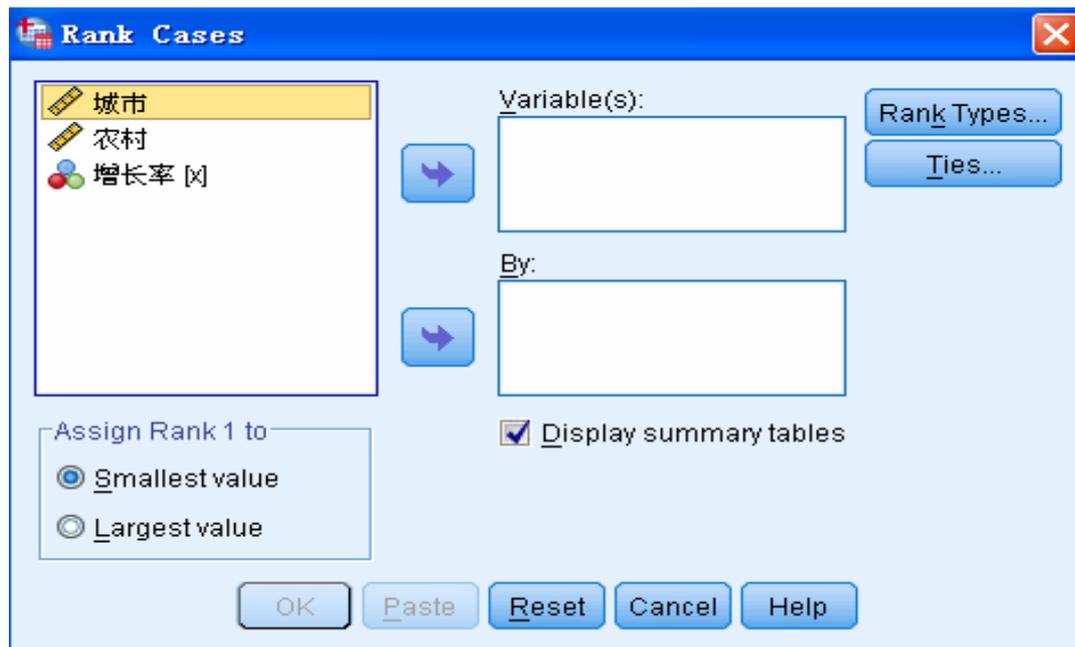
CONCEPT  
RATE

- “秩”（Rank）是数据整理中的重要概念，前面讲解的观测量排序是按照大小顺序重新排列观测量，而观测量求秩是指对观测量排序后指定的“名次”。例如，观测量的值依次为3、5、-2、0、7，它们按小到大排列后为-2、0、3、5、7，各观测量的秩等于3、4、1、2、5。

# 1. SPSS操作详解

- **Step01:** 打开观测量求秩对话框

打开SPSS软件，选择菜单栏中的【File(文件)】→【Transform(转换)】→【Rank Cases(个案排秩)】命令，弹出【Rank Cases(个案排秩)】对话框。



- **Step02: 选择求秩变量**

在左侧的候选列表框中选择求秩变量，将其移入【Variable(s)(变量)】，此时系统会产生一个新的秩变量，它是在该变量的前面添加“r”而构成。

- **Step03: 选择求秩顺序**

【Assign Rank 1 to(将秩1指定给)】选项组用于指定求秩顺序。

- **Step04: 选择分组变量**

在左侧的候选变量列表框中选择分组变量，将其移入【By(排序标准)】列表框，此时SPSS会按所选的分组变量来求秩，如果不设定本选项，将对所有的观测测量排秩。



- **Step05:** 选择汇总表输出

勾选【**Display summary tables (显示摘要表)**】复选框，系统将在输出窗口中显示概况原变量和新变量的摘要信息表。

- **Step06:** 秩类型选择

单击【**Rank Types**】按钮，在弹出的对话框中可以选择秩的类型。

**Rank Cases: Types** ✕

Rank
  Fractional rank as %

Savage score
  Sum of case weights

Fractional rank
  Ntiles:

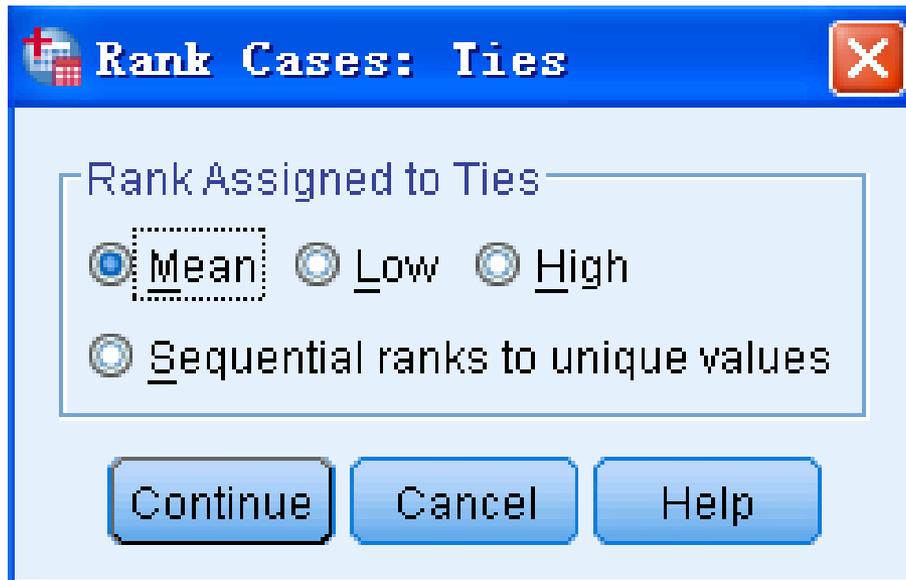
Proportion estimates
  Normal scores

Proportion Estimation Formula

Blom
  Tukey
  Rankit
  Van der Waerden

- **Step07:** Ties（结）类型选择

单击【Ties】按钮，在弹出的对话框中用户可以选择结类型。



Step08: 最后单击主对话框中的【OK】按钮，此时操作结束。



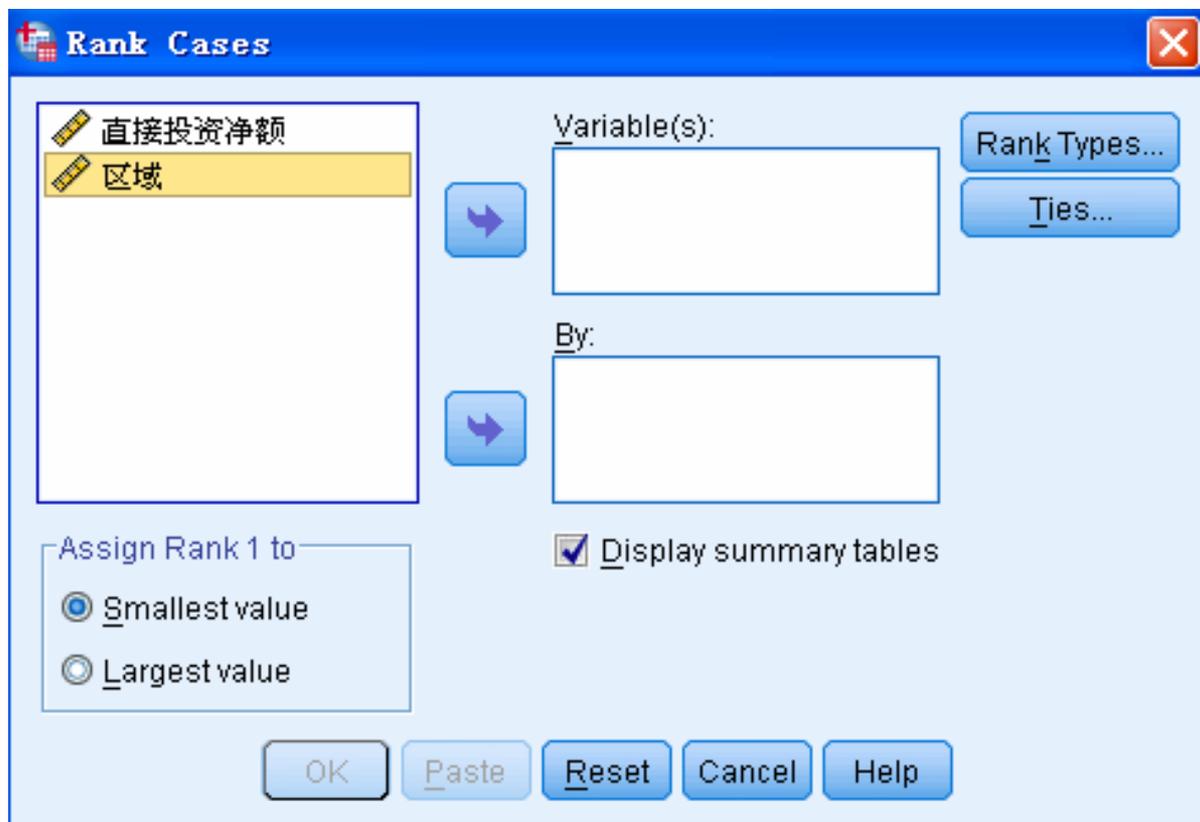
## 2.实例内容： 对外直接投资净额分析

- 2005年我国对主要国家（地区）对外直接投资金额（非金融类）的原始数据见数据文件2-13.sav，请按照区域类型不同对投资净额排秩。

	国家和地区	直接投资净额	区域
1	香港	341970.00	1
2	印度尼西亚	1184.00	1
3	日本	1717.00	1
4	澳门	834.00	1
5	新加坡	2033.00	1
6	韩国	58882.00	1
7	泰国	477.00	1
8	越南	2077.00	1
9	阿尔及利亚	8487.00	2
10	苏丹	9113.00	2
11	几内亚	1634.00	2
12	马达加斯加	14.00	2
13	尼日利亚	5330.00	2
14	南非	4747.00	2
15	英国	2478.00	3
16	德国	12874.00	3
17	法国	609.00	3
18	俄罗斯	20333.00	3
19	巴哈马	2295.00	4
20	开曼群岛	516275.00	4
21	墨西哥	355.00	4
22	英属维尔京群	122608.00	4



# Step01: 打开对话框





## Step02: 选择求秩变量

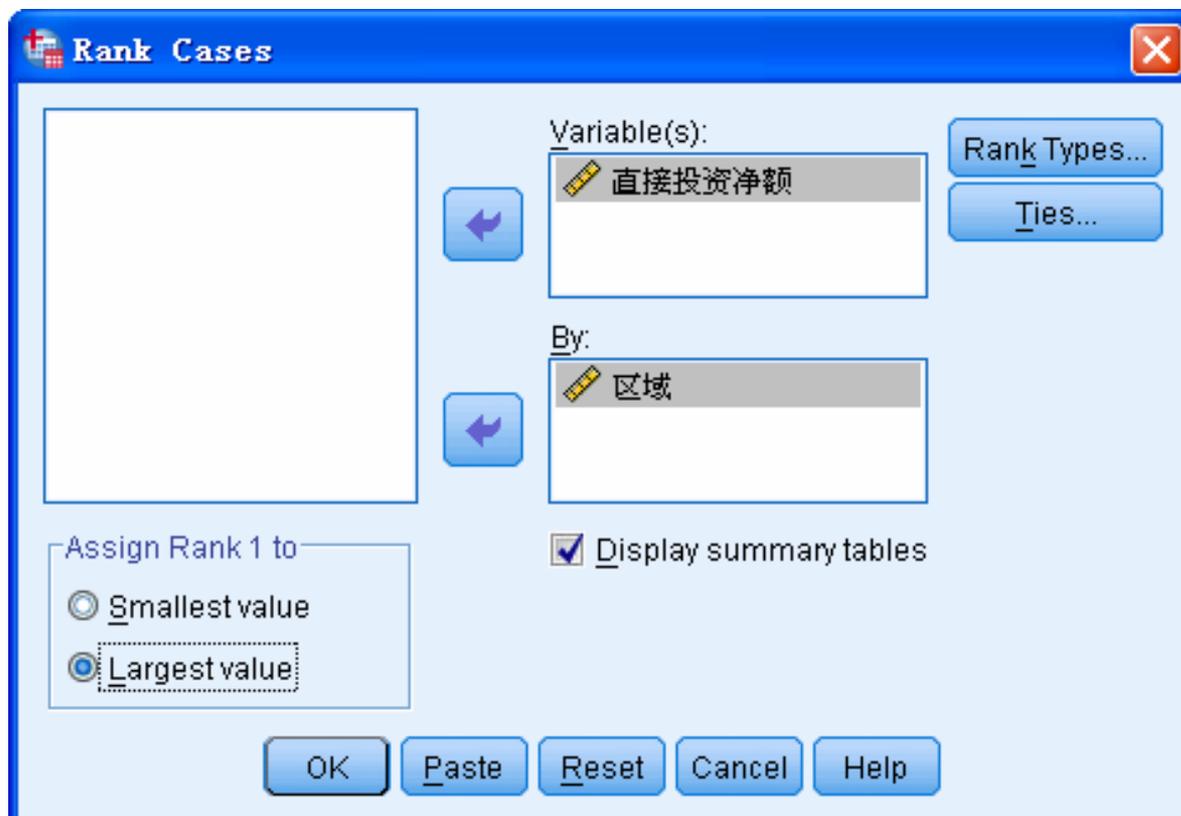


## Step03: 选择分组变量

CONCEPT  
STRATE

- 在左侧的候选变量列表框中选择分组变量“区域”，将其移入【By(排序标准)】列表框，此时SPSS会按它进行分组求秩。

# Step04: 选择求秩顺序



# Step05: 完成操作

CONCEPT  
STRATE

	国家和地区	直接投资净额	区域	R直接投
1	香港	341970.00	1	1.000
2	印度尼西亚	1184.00	1	6.000
3	日本	1717.00	1	5.000
4	澳门	834.00	1	7.000
5	新加坡	2033.00	1	4.000
6	韩国	58882.00	1	2.000
7	泰国	477.00	1	8.000
8	越南	2077.00	1	3.000
9	阿尔及利亚	8487.00	2	2.000
10	苏丹	9113.00	2	1.000
11	几内亚	1634.00	2	5.000
12	马达加斯加	14.00	2	6.000
13	尼日利亚	5330.00	2	3.000
14	南非	4747.00	2	4.000
15	英国	2478.00	3	3.000
16	德国	12874.00	3	2.000
17	法国	609.00	3	4.000
18	俄罗斯	20333.00	3	1.000
19	巴哈马	2295.00	4	3.000
20	开曼群岛	516275.00	4	1.000
21	墨西哥	355.00	4	4.000
22	英属维尔京群	122608.00	4	2.000



第3章

SPSS描述性统计  
分析

统计分析的目的是研究总体的数量特征。为实现上述分析，往往采用两种方式实现：第一，数值计算，即计算常用的基本统计量的值，通过数值来准确反映数据的基本统计特征；第二，图形绘制，即绘制常见的基本统计图形，通过图形来直观展现数据的分布特点。通常，这两种方式都是混合使用的。

SPSS 的许多模块均可完成描述性分析，但专门为该目的而设计的几个模块则集中在【Descriptive Statistics】菜单中。最常用的是列在最前面的四个过程。

- Frequencies: 产生频数表。
- Descriptives: 进行基本的统计描述分析。
- Explore: 探索性分析。
- Crosstabs: 列联表分析。

# 3.1 SPSS在频数分析中的应用

CONCEPT  
STRATE

## 3.1.1 频数分析的基本原理

### 1. 使用目的

频数分布表是描述性统计中最常用的方法之一。它主要能够了解变量取值的状况，对把握数据分布特征非常有用。例如，了解某班学生考试的学习成绩、了解某地区居民的收入水平等都可以借助于频数分析。

## 2. 软件使用方法

Frequencies 过程就是专门为产生频数表而设计的。它不仅可以直接产生详细的频数表，还可以按要求给出某百分位点的数值以及常用的条图、饼图等统计图。同时，SPSS 的频数分析还可以进行分位数、描述集中趋势的基本统计量等计算功能。这些统计量的具体分析会在以后章节中讲解。

### 3. Bootstrap方法

- (1) 采用重抽样技术从原始样本中抽取一定数量（自己给定）的样本，此过程允许重复抽样。
- (2) 根据抽出的样本计算给定的统计量 $T$ 。
- (3) 重复上述 $N$ 次（一般大于1000），得到 $N$ 个统计量 $T$ 。
- (4) 计算上述 $N$ 个统计量 $T$ 的样本值，最终得到统计量的估计值。

## 3.1.2 频数分析的SPSS操作详解

CONCEPT  
STRATE

Step01: 打开主窗口

选择菜单栏中的【Analyze(分析)】→【Descriptive Statistics(描述性统计)】→【Frequencies(频率)】命令，弹出【Frequencies(频率)】对话框，这是频数分析的主操作窗口。

# Step01: 打开主窗口

CONCEPT  
STRATE



## Step02: 选择分析变量

CONCEPT  
STRATE

在【Frequencies (频率)】对话框的左侧的候选变量列表框中，选取一个或多个待分析变量，将它们移入右侧的【Variable(s) (变量)】列表框中。

## Step03: 输出频数分析表

CONCEPT  
STRATE

勾选【Display frequency tables (显示频率表格)】复选框，输出频数分析表。

## Step04: 其他基本统计分析

CONCEPT  
STRATE

- 在对话框中还可以单击【Statistics（统计量）】和【Chars（图表）】等按钮。这些选项提供了丰富的统计输出结果。

单击【Statistics】按钮，在弹出的对话框中可以设置输出各类基本统计量结果。

**Frequencies: Statistics** [X]

**Percentile Values**

Quartiles

Cut points for:  equal groups

Percentile(s):

**Central Tendency**

Mean

Median

Mode

Sum

Values are group midpoints

**Dispersion**

Std. deviation     Minimum

Variance         Maximum

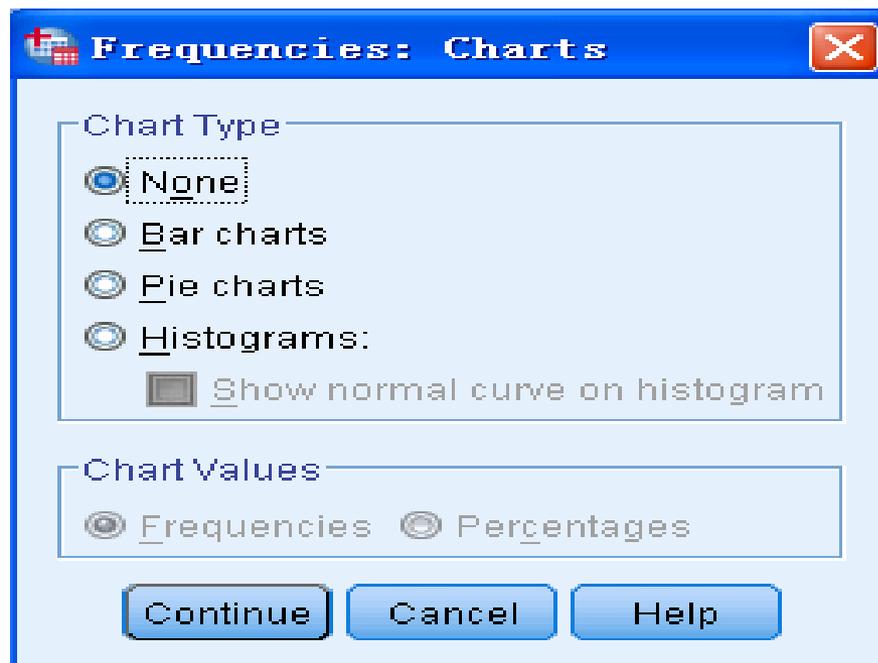
Range             S.E. mean

**Distribution**

Skewness

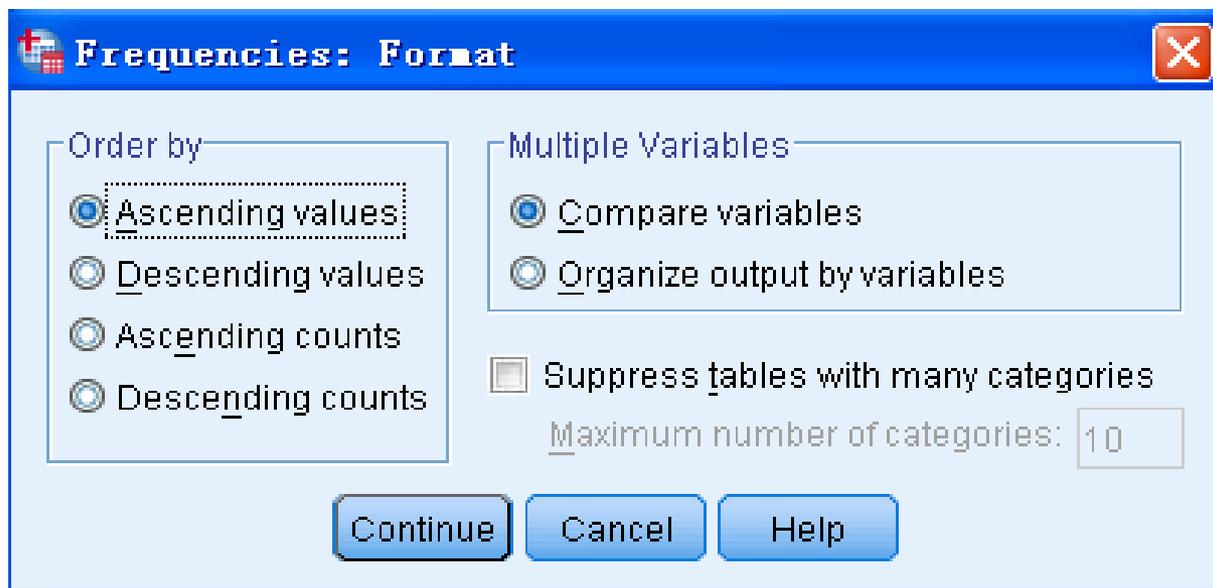
Kurtosis

单击【Charts】按钮，在弹出的对话框中设置输出图形结果。



## Step05: 输出格式选择

单击【Format】按钮，在弹出的对话框中设置频数表输出的格式。



# Step06: 相关统计量的Bootstrap估计

CONCEPT  
STRATE

单击【Bootstrap】按钮，在弹出的对话框中可以进行如下统计量的Bootstrap估计。

- 支持均值、标准差、方差、中位数、偏度、峰度和百分位数的Bootstrap估计。
- 支持百分比的Bootstrap估计。

**Bootstrap** ✕

**Perform bootstrapping**

Number of samples:

**Set seed for Mersenne Twister**

Seed:

**Confidence Intervals**

Level(%):

Percentile

Bias corrected accelerated (BCa)

**Sampling**

Simple

Stratified

Variables:

- 地区
- 可吸入颗粒物
- 二氧化硫
- 二氧化氮
- 空气质量达到及好于...

➔

Strata Variables:

## Step07: 完成操作



单击【OK】按钮，结束操作，SPSS软件自动输出结果。

# 3.1.3 实例图文分析：产品的销售量

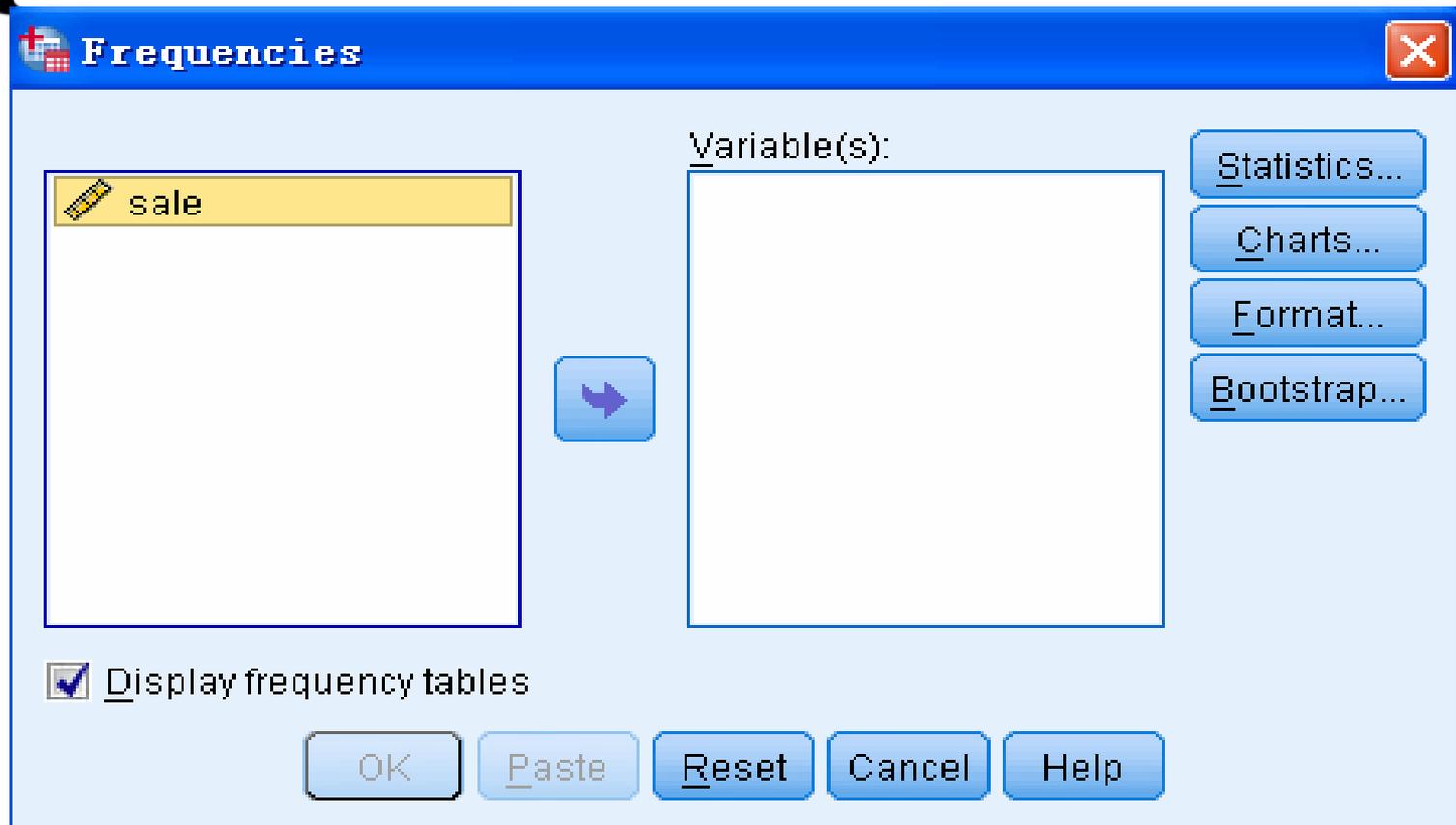
CONCEPT  
STRATE

假设某公司每周大约卖出2000万件产品，但市场的需求不稳定，该公司的生产经理想更好的掌握近期该产品的分布情况。假设下面给出的销售数字（单位：百万）代表近期公司该产品每周的销售数据。利用频数分析你能得到什么有助于生产及销售的的信息？

24	18	18	26	24	23	16	18	21	20	21	24	19
19	14	22	21	26	27							
15	19	17	20	20	19	22	23	16	23	21	15	19
21	20	22	15	24	19							

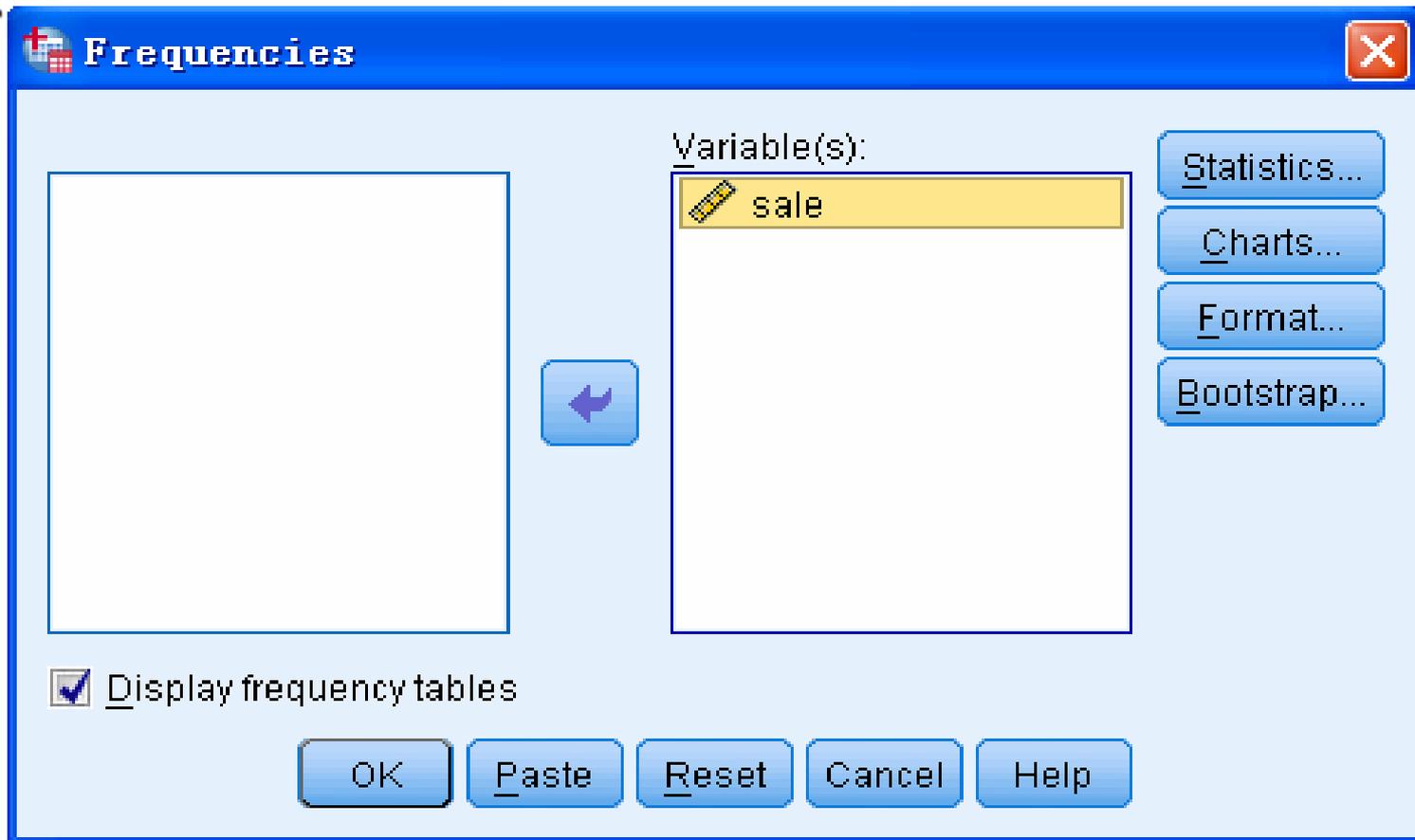
# Step01: 打开对话框

CONCEPT  
TRATE



# Step02: 选择分析变量

CONCEPT  
TRATE



# Step03: 选择输出统计量

**Frequencies: Statistics**

**Percentile Values**

- Quartiles**
- Cut points for:** 10 **equal groups**
- Percentile(s):**

**Central Tendency**

- Mean**
- Median**
- Mode**
- Sum**

**Values are group midpoints**

**Dispersion**

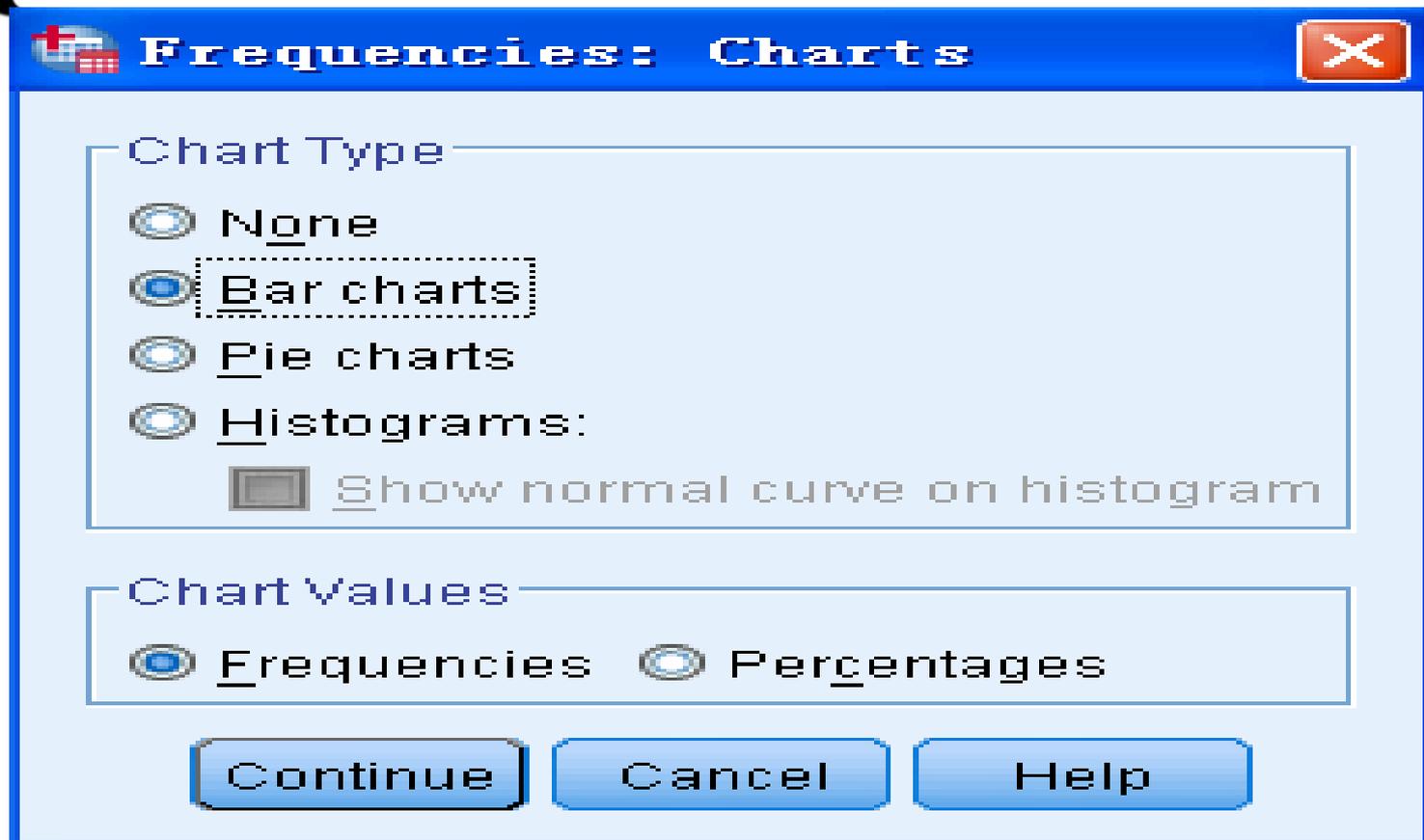
- Std. deviation**
- Variance**
- Range**
- Minimum**
- Maximum**
- S.E. mean**

**Distribution**

- Skewness**
- Kurtosis**

**Buttons:** Continue, Cancel, Help

# Step04: 选择输出图形类型





# Step05: 完成操作

## (1) 基本统计结果输出

### 频数分析基本统计结果

<b>N</b>	<b>Valid</b>	<b>38</b>
	<b>Missing</b>	<b>0</b>
<b>Percentiles</b>	<b>25</b>	<b>18.00</b>
	<b>50</b>	<b>20.00</b>
	<b>75</b>	<b>23.00</b>

## (2) 频数分析表输出

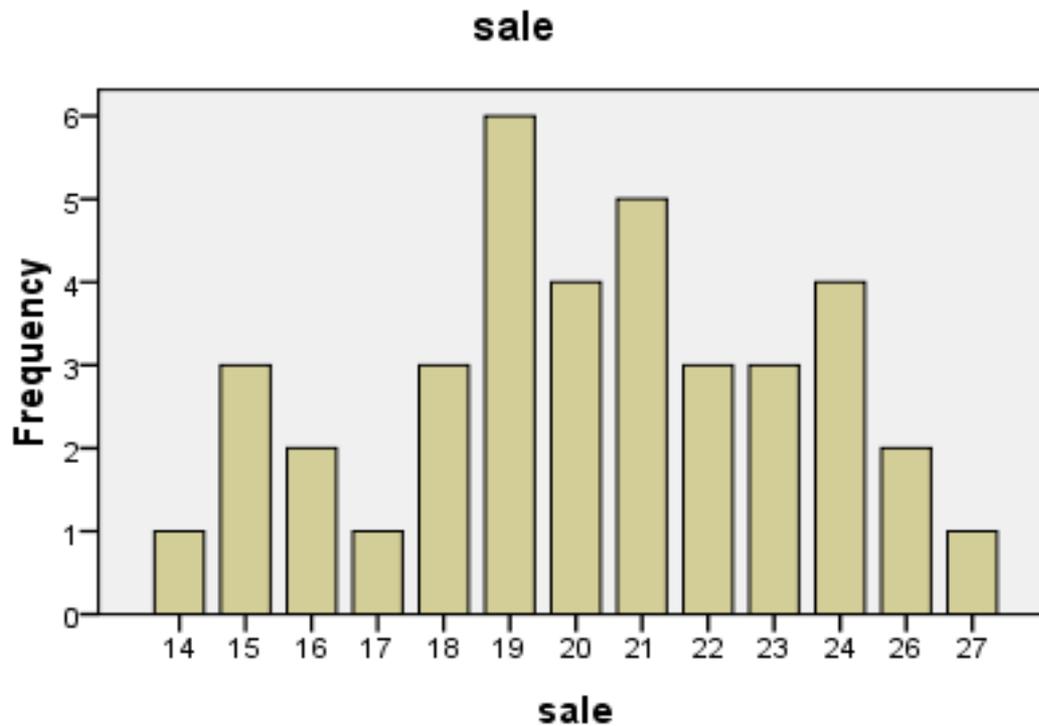


表3-2 频数分析表

### 频数分析表

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	14	1	2.6	2.6	2.6
	15	3	7.9	7.9	10.5
	16	2	5.3	5.3	15.8
	17	1	2.6	2.6	18.4
	18	3	7.9	7.9	26.3
	19	6	15.8	15.8	42.1
	20	4	10.5	10.5	52.6
	21	5	13.2	13.2	65.8
	22	3	7.9	7.9	73.7
	23	3	7.9	7.9	81.6
	24	4	10.5	10.5	92.1
	26	2	5.3	5.3	97.4
	27	1	2.6	2.6	100.0
	Total	38	100.0	100.0	

### (3) 直方图



## 3.2 SPSS在描述统计分析中的应用

CONCEPT  
STRATE

### 3.2.1 描述统计分析的基本原理

1. 使用目的
2. 刻画集中趋势的描述统计量
3. 刻画离散程度的描述统计量
4. 刻画分布形态的描述统计量

## 3.2.2 描述统计分析的SPSS操作详解

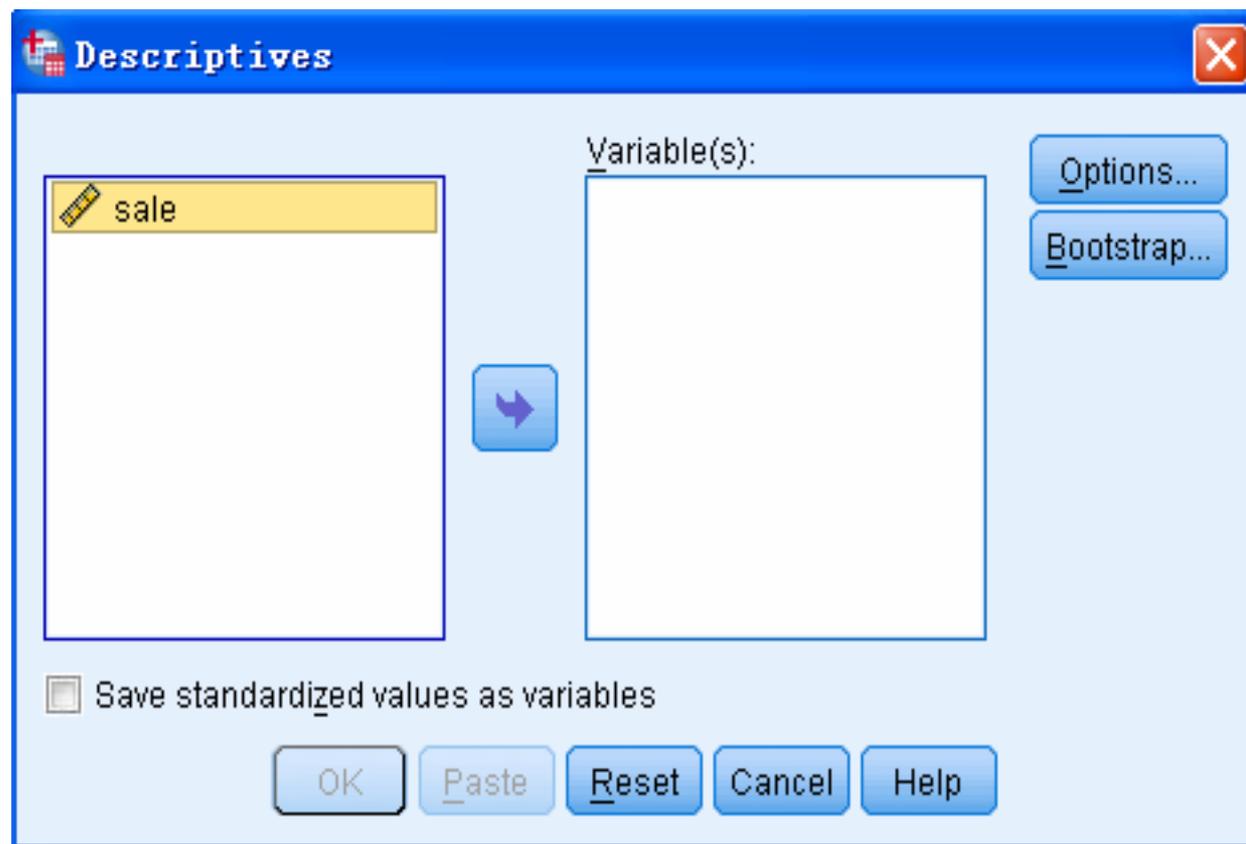
CONCEPT  
RATE

Descriptives 过程是连续资料统计描述应用最多的一个过程，它可对变量进行描述性统计分析计算，并列出一系列相应的统计指标。这和其他过程相比并无不同。但该过程还有个特殊功能，就是可将原始数据转换成标准化值，并以变量的形式保存。

# Step01: 打开主窗口



选择菜单栏中的【Analyze(分析)】→  
【Descriptive Statistics(描述性统计)】  
→【Descriptives(描述)】命令，弹出  
【Descriptives(描述)】对话框，该对话框  
是描述性统计分析的主操作窗口。



## Step02: 选择分析变量

CONCEPT  
TRATE

在左侧的候选变量列表框中选取一个或多个待分析变量，将它们移入右侧的【Variable(s) (变量)】列表框中。



## Step03: 计算基本描述性统计量

单击【Options】按钮，弹出【Options (选择)】对话框，该对话框用于指定输出的描述性统计量。这些统计量的含义是：均数 (Mean)、总和 (Sum)、标准差 (Std. deviation)、方差 (Variance)、全距 (Range)、最小值 (Minimum)、最大值 (Maximum)、标准误差 (S. E. mean)、偏度系数 (Skewness) 和峰度系数 (Kurtosis)。

 **Descriptives: Options** 

Mean  Sum

Dispersion

Std. deviation  Minimum  
 Variance  Maximum  
 Range  S.E. mean

Distribution

Kurtosis  Skewness

Display Order

Variable list  
 Alphabetic  
 Ascending means  
 Descending means

## Step04: 保存标准化变量

CONCEPT  
STRATE

勾选【Save standardized values as variables (保存标准化变量值)】复选框。

## Step05: 相关统计量的Bootstrap估计

CONCEPT  
STRATE

单击【Bootstrap】按钮，弹出【Bootstrap】对话框，可以进行均值、标准差、方差、偏度和峰度的Bootstrap估计。

## Step06: 完成操作



单击【OK】按钮，结束操作，SPSS软件自动输出结果。

### 3.2.3 实例图文分析：奥斯卡获奖者的年龄

CONCEPT  
STRATE

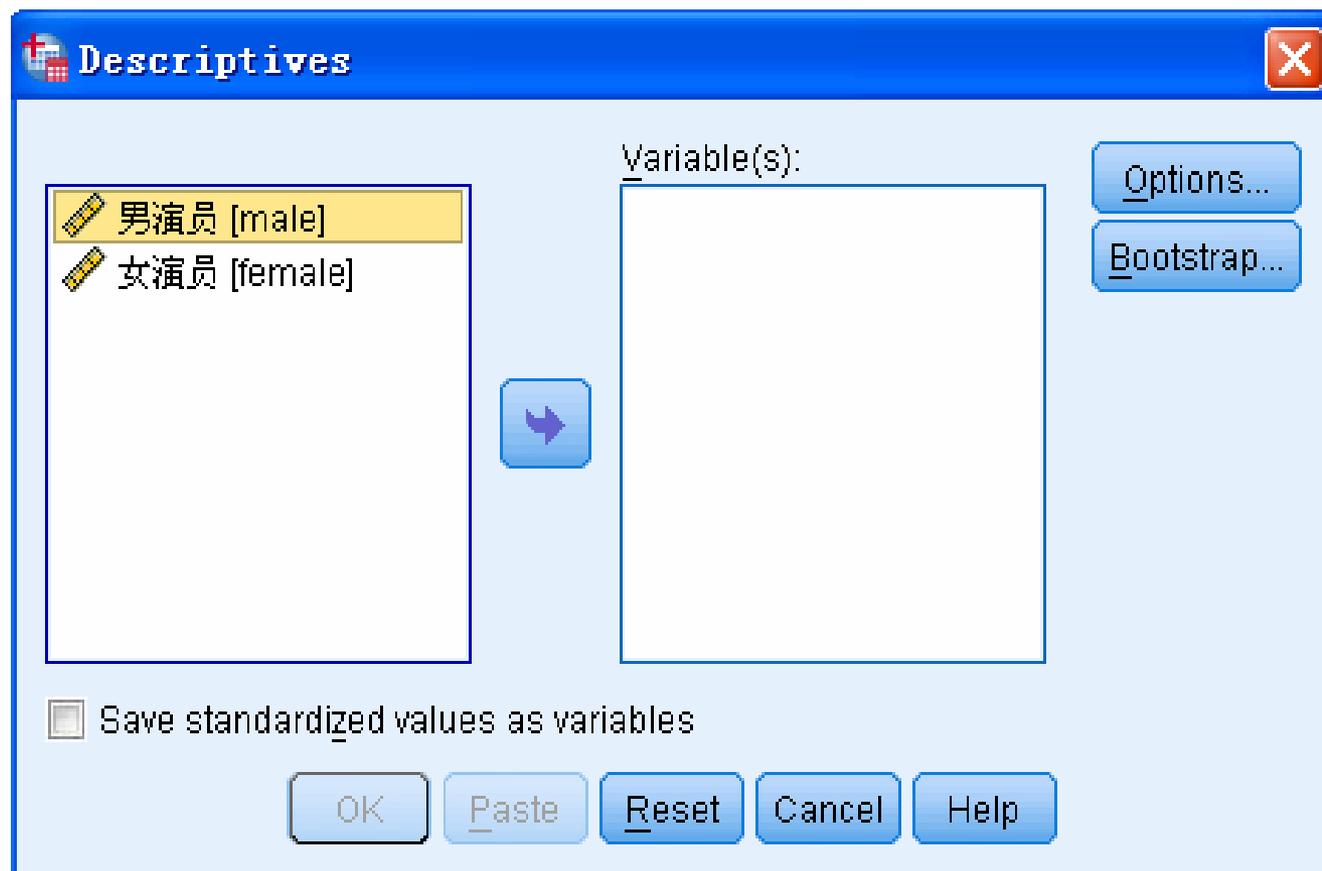
请你分析不同性别演员获得奥斯卡奖的年龄差异性。

男演员：32 37 36 32 51 53 33  
61 35 45 55 39 76 37 42 40 32  
60 38 56 48 48 40 43 62 43 42  
44 41 56 39 46 31 47 45 60

女演员：50 44 35 80 26 28 41  
21 61 38 49 33 74 30 33 41 31  
35 41 42 37 26 34 34 35 26 61  
60 34 24 30 37 31 27 39 34

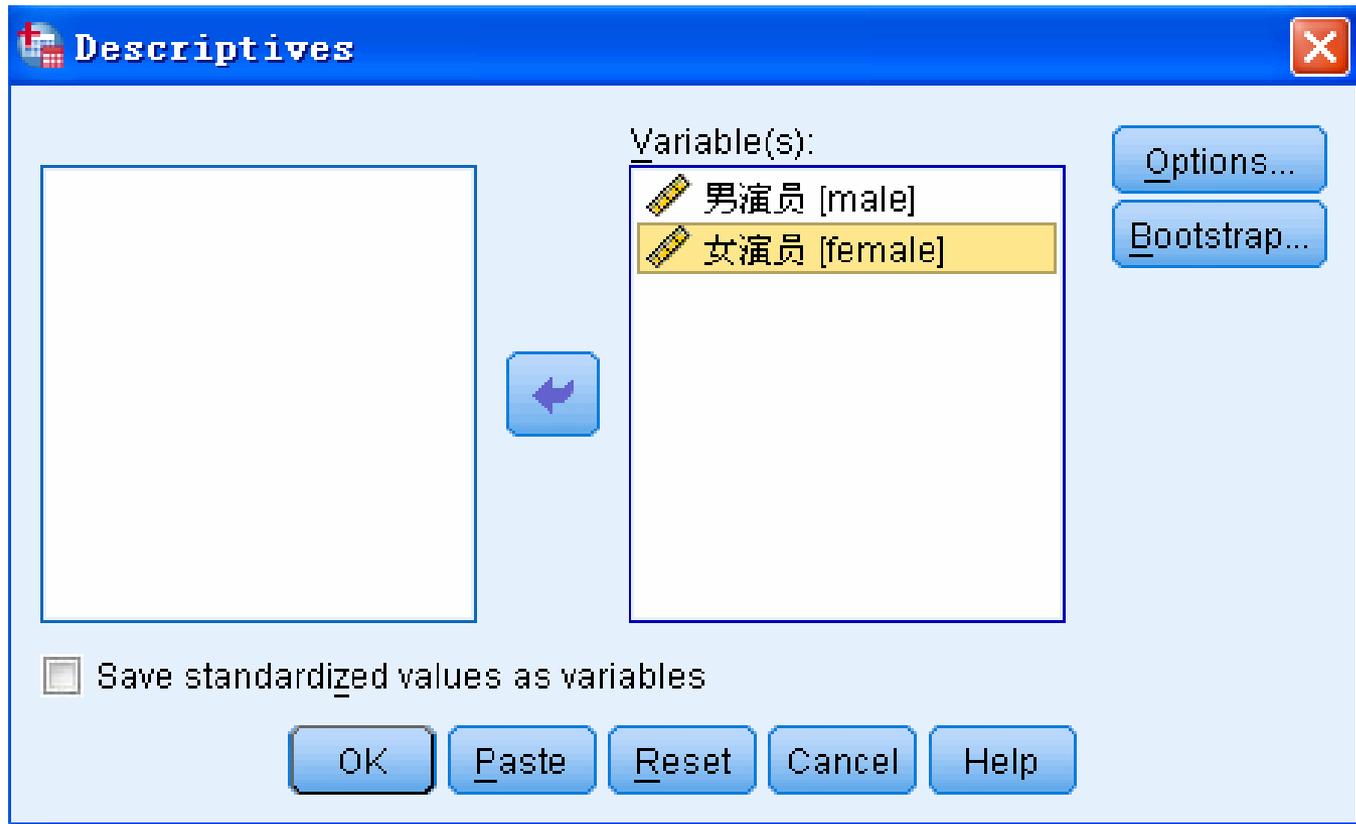
# Step01: 打开对话框

CONCEPT  
STRATE



## Step02: 选择分析变量

CONCEPT  
STRATE





## Step03: 选择输出描述性统计量

**Descriptives: Options**

Mean       Sum

Dispersion

Std. deviation       Minimum

Variance       Maximum

Range       S.E. mean

Distribution

Kurtosis       Skewness

Display Order

Variable list

Alphabetic

Ascending means

Descending means

Continue      Cancel      Help

## Step04: 完成操作



单击【OK】按钮，操作完成。



# 实例结果及分析

表 3-3 描述性统计分析结果

		男演员	女演员	Valid N (listwise)
Statistic	N	36	36	36
Statistic	Range	45	59	
Statistic	Minimum	31	21	
Statistic	Maximum	76	80	
Statistic	Mean	45.14	38.94	
Statistic	Std. Deviation	10.406	13.546	
Statistic	Skewness	.898	1.503	
Std. Error		.393	.393	
Statistic	Kurtosis	.704	2.111	
Std. Error		.768	.768	

## 3.3 SPSS在探索性分析中的应用

CONCEPT  
STRATE

### 3.3.1 探索性分析的基本原理

#### 1. 使用目的

探索性数据分析（Exploratory Data Analysis, 简称EDA）的基本思想是从数据本身出发，不拘泥于模型的假设而采用非常灵活的方法来探讨数据分布的大致情况，也可以为进一步结合模型的研究提供线索，为传统的统计推断提供良好的基础和减少盲目性。

## 2. 主要内容

一般来说，进行探索性分析主要考察以下内容。

- (1) 检查数据是否有错。过大或过小的数据均可能是异常值、影响点或错误值。要检查这样的数据，并分析原因，然后决定是否从分析中剔除这些数据。
- (2) 获得数据分布特征。很多统计方法模型对数据的分布有要求，如方差分析就需要数据服从正态分布。
- (3) 对数据的初步观察，发现一些内在规律。

## 3.3.2 探索性分析的SPSS操作详解

CONCEPT  
STRATE

SPSS中的Explore过程用于计算指定变量的探索性统计量和有关的图形。它既可以对观测量整体分析，也可以进行分组分析。从这个过程可以获得箱线图、茎叶图、直方图、各种正态检验图、频数表、方差齐性检验等结果，以及对非正态或正态非齐性数据进行变换，并表明和检验连续变量的数值分布情况。

# Step01 打开主窗口



选择菜单栏中的【Analyze(分析)】→  
【Descriptive Statistics(描述性统计)】  
→【Explore(探索)】命令，弹出【Explore  
(探索)】对话框，该对话框是探索性分析的  
主操作窗口。

**Explore** [Close]

男演员 [male]  
 女演员 [female]

Dependent List:

Factor List:

Label Cases by:

Statistics...  
Plots...  
Options...  
Bootstrap...

Display  
 Both  
  Statistics  
  Plots

## Step02 选择分析变量

CONCEPT  
STRATE

在【Explore(探索)】对话框左侧的【候选变量】清单中，选取一个或多个待分析变量，将它们移入右侧的【Dependent List (因变量列表)】列表框中，表示要进行探索性分析的变量。

## Step03 选取分组变量



在【Explore(探索)】对话框的候选变量列表框中，可以选取一个或多个分组变量，将它们移入右侧的【Factor List (因子列表)】列表框中。分组变量的选择可以将数据按该变量中的观测值进行分组分析。如果选择的分组变量不止一个，那么会以分组变量的不同取值进行组合分组。



## Step04 选择标签值

从候选变量列表框中选择一个变量作为标识变量，并将其移入【Label Cases by (标注个案)】列表框中。选择标识变量的作用在于，若系统在数据探索时发现异常值，便可利用标识变量加以标记，便于用户找这些异常值。如果不选择它，系统默认以id变量作为标识变量。

## Step05 选择输出类型

CONCEPT  
TRATE

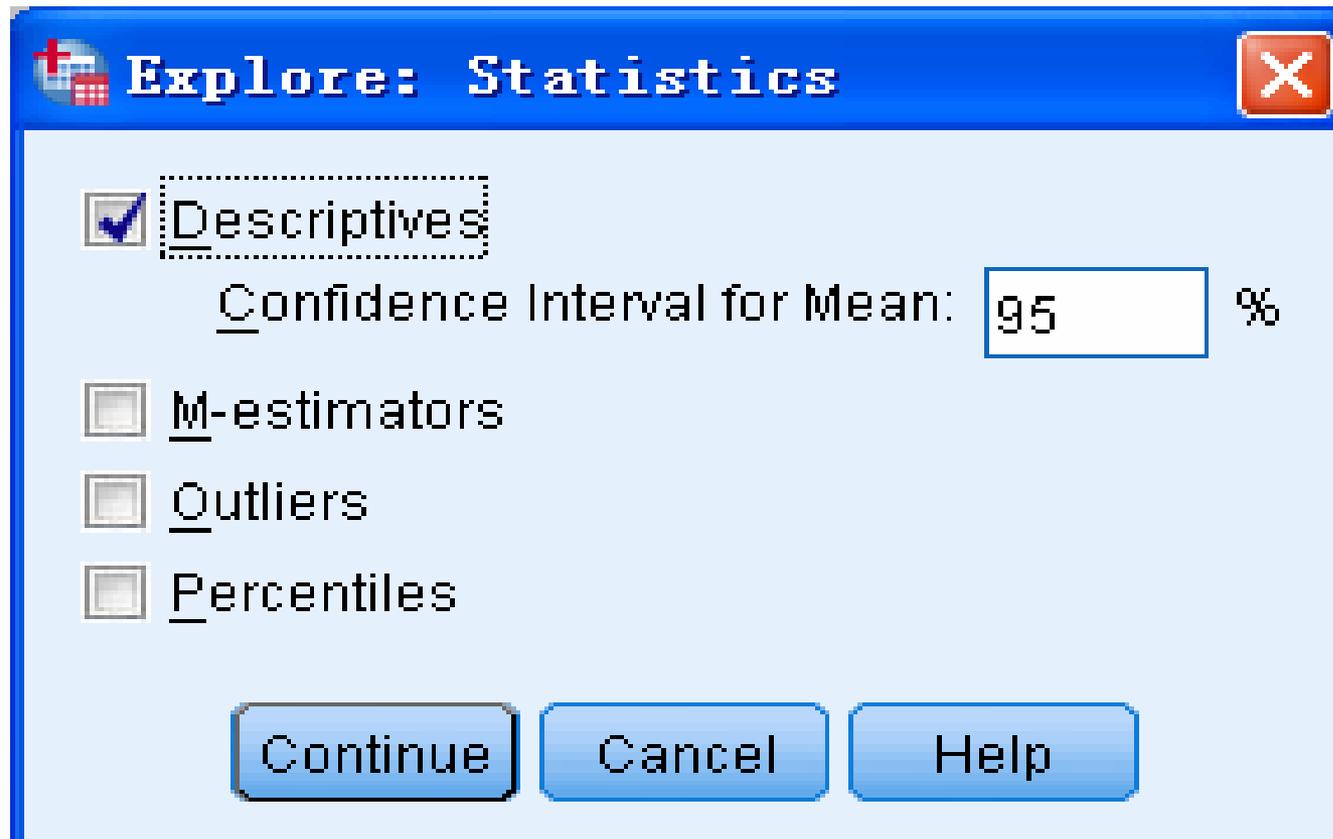
在【Explore(探索)】对话框下面的【Display】选项组中可以选择输出项。

- Both: 输出图形以及描述性统计量。
- Statistics: 只输出描述统计量。选择此项后激活【Statistics】功能按钮。
- Plots: 只输出图形。选择此项后激活【Plots】功能按钮。

## Step06 描述性统计量结果输出

CONCEPT  
STRATE

在【Explore(探索)】对话框中还可以单击【Statistics】按钮，弹出【Explore: Statistics】对话框，该对话框中提供了各类基本描述性统计输出结果。

A dialog box titled "Explore: Statistics" with a blue header bar. The header bar contains a small icon of a building with a red cross on the left and a red close button with a white 'X' on the right. The main area of the dialog is light blue and contains several options. The first option is "Descriptives", which is checked with a blue checkmark in a small square box. Below it is the text "Confidence Interval for Mean:" followed by a text input field containing the number "95" and a percent sign "%". Below this are three more options, each with an unchecked checkbox: "M-estimators", "Outliers", and "Percentiles". At the bottom of the dialog are three buttons: "Continue", "Cancel", and "Help", all with a light blue gradient and rounded corners.

### Explore: Statistics

Descriptives

Confidence Interval for Mean:  %

M-estimators

Outliers

Percentiles

Continue Cancel Help

## Step07 统计图形结果输出

CONCEPT  
STRATE

在【Explore(探索)】对话框中还可以单击【Plots】按钮，弹出【Explore: Plots】对话框。该对话框中提供了图形输出的类型。

**Explore: Plots** ✕

**Boxplots**

- Factor levels together
- Dependents together
- None

**Descriptive**

- Stem-and-leaf
- Histogram

Normality plots with tests

**Spread vs Level with Levene Test**

- None
- Power estimation
- Transformed Power: Natural log
- Untransformed

## Step08 选择缺失值的处理方式

在【Explore(探索)】对话框中还可以单击【Options】按钮，在弹出的对话框中确定对待缺失值的方式。



## Step09 相关统计量的Bootstrap估计

CONCEPT  
STRATE

单击【Bootstrap】按钮，弹出【Bootstrap】对话框，可以进行如下统计量的Bootstrap估计。

- 支持均值、5% 切尾均值、标准差、方差、中位数、偏度、峰度和内距的Bootstrap估计。
- M 估计量表支持Huber 的M 估计量、Tukey 的双权重、Hampel 的M 估计量和Andrew的Wave 的Bootstrap 估计。
- 百分位数表支持百分位数的Bootstrap 估计

## Step10 : 操作完成



单击【OK】按钮，结束操作，SPSS软件自动输出结果。

### 3.3.3 实例图文分析：中国南北城市的温度差异

CONCEPT  
STRATE

表 3-4 2002 年我国主要城市的年平均温度

中

城市	年平均温度	城市	年平均温度	城市	年平均温度
北京	13.1	杭州	17.4	海口	25.0
天津	13.2	合肥	17.2	桂林	19.3
石家庄	14.4	福州	20.9	重庆	18.7
太原	10.9	南昌	18.3	成都	17.4
呼和浩特	8.0	济南	15.0	贵阳	14.6
沈阳	9.2	青岛	13.4	昆明	16.1
大连	11.8	郑州	15.4	拉萨	8.5
长春	6.8	武汉	17.9	西安	15.4
哈尔滨	5.4	长沙	17.7	兰州	11.0
上海	17.5	广州	22.9	西宁	6.1
南京	16.6	南宁	21.7	银川	10.0

## Step01: 打开对话框



打开数据文件3-3. sav，其中增加变量“地域”表示所在城市的区域位置，“1”表示南方城市，“2”表示北方城市。选择菜单栏中的【Analyze(分析)】→【Descriptive Statistics(描述性统计)】→【Explore(探索)】命令，弹出【Explore(探索)】对话框。

**Explore** [Close]

**城市 [主要城市]**  
年平均气温 [年平均...]  
地域 [地域]

Dependent List:  
Factor List:  
Label Cases by:

Statistics...  
Plots...  
Options...  
Bootstrap...

Display:  
 Both  Statistics  Plots

OK Paste Reset Cancel Help

## Step02: 选择分析变量

CONCEPT  
STRATE

在候选变量列表框中将变量“年平均温度”添加至【Dependent List (因变量列表)】列表框中，表示它是进行探索性分析的变量。

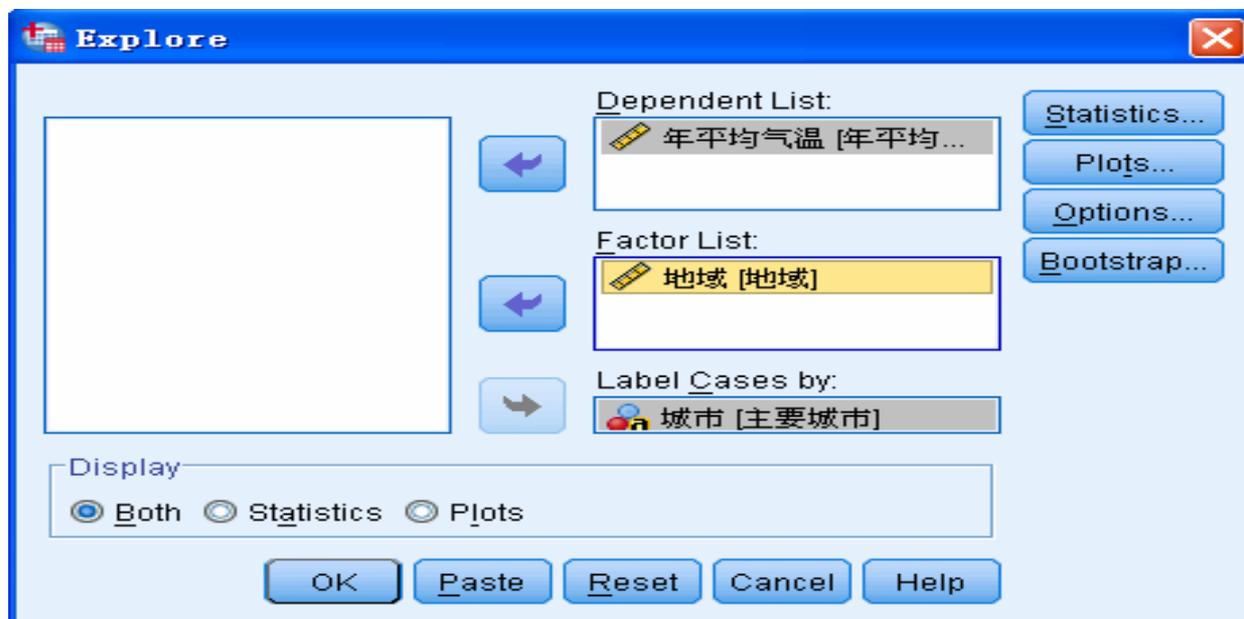
## Step03: 选择分组变量

CONCEPT  
STRATE

将变量“地域”添加至【Factor List (因子列表)】列表框中，表示根据地域位置不同来进行数据分析。

## Step04: 选择标签值

选择变量“城市”移入【Label Cases by (标注个案)】列表框作为标识变量。





## Step05: 选择输出描述性统计量

单击【Statistics】按钮，在弹出的对话框中勾选【M-estimators (M估计值)】复选框，分析样本数据的稳健性。其他选项保持SPSS默认状态。单击【Continue】按钮，返回【Explore (探索)】对话框。



## Explore: Statistics



Descriptives

Confidence Interval for Mean:  %

M-estimators

Outliers

Percentiles

Continue

Cancel

Help

## Step06: 完成操作



最后，单击【OK】按钮，操作完成。

# 3 实例结果及分析



## (1) 基本统计信息汇总

### 基本统计信息

	地域	Cases					
		Valid		Missing		Total	
		N	Percent	N	Percent	N	Percent
年平均气温	南方	16	100.0%	0	.0%	16	100.0%
	北方	17	100.0%	0	.0%	17	100.0%

## (2) 描述性统计表

表 3-6 描述性统计表

	地域		Statistic	Std. Error	
年平均 气温	南方	Mean	18.7000	.67200	
		95% Confidence Interval for Mean	Lower Bound	17.2677	
			Upper Bound	20.1323	
		5% Trimmed Mean	18.5778		
		Median	17.8000		
		Variance	7.225		
		Std. Deviation	2.68800		
		Minimum	14.60		
		Maximum	25.00		
		Range	10.40		
		Interquartile Range	3.25		
	Skewness	1.001	.564		
	Kurtosis	.782	1.091		
北方	Mean	11.0353	.80078		

95% Confidence Interval for Mean	Lower Bound	9.3377	
	Upper Bound	12.7329	
5% Trimmed Mean		11.1059	
Median		11.0000	
Variance		10.901	
Std. Deviation		3.30169	
Minimum		5.40	
Maximum		15.40	
Range		10.00	
Interquartile Range		5.65	
Skewness		-.251	.550
Kurtosis		-1.178	1.063

□

### (3) M估计量

M估计量结果表

	地域	Huber's M- Estimator <sup>a</sup>	Tukey's Biweight <sup>b</sup>	Hampel's M-Estimat or <sup>c</sup>	Andrews' Wave <sup>d</sup>
年平均 气温	南方	18.0694	17.7208	17.9776	17.7182
	北方	11.2075	11.1706	11.1741	11.1696

a. 权数取值为 1.339.

b. 权数取值为4.685.

c. 权数取值分别为1.700, 3.400, and 8.500

d. 权数取值为 $1.340 \cdot \pi$ .

## (4) 茎叶图



### 探索性分析的茎叶图

年平均气温 Stem-and-Leaf Plot for  
地域= 南方

Frequency	Stem & Leaf
1.00	1 . 4
11.00	1 . 66777777889
3.00	2 . 012
1.00	Extremes (>=25)

Stem width: 10.0

Each leaf: 1 case(s)

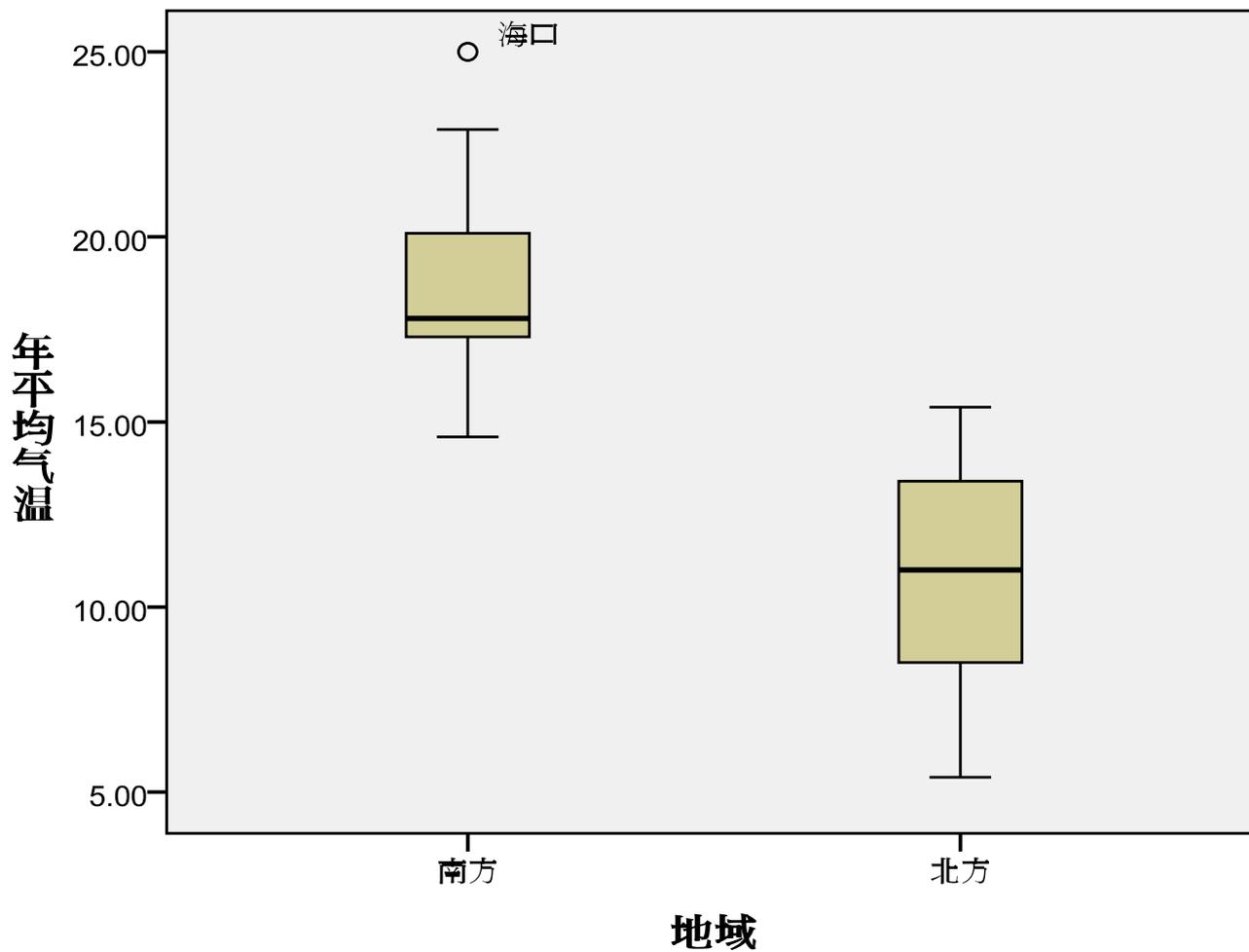
年平均气温 Stem-and-Leaf Plot for  
地域= 北方

Frequency	Stem & Leaf
6.00	0 . 566889
8.00	1 . 00113334
3.00	1 . 555

Stem width: 10.0

Each leaf: 1 case(s)

# (5) 箱图



## 3.4 SPSS在列联表分析中的应用

CONCEPT  
RATE

### 3.4.1 列联表分析的基本原理

#### 1. 使用目的

列联表是指一个频率对应两个变量的表（一个变量用来对行分类，第二个变量用来对列分类）。列联表非常重要，它经常被用来分析调查结果。它有两个基本任务：第一，根据收集到的样本数据产生二维或多维交叉列联表；第二，在列联表基础上，对两两变量间是否存在一定的相关性进行分析。

## 2. 交叉列联表

表 3-9 二维  $r \times c$  列联表

	$B_1$	$B_2$	...	$B_c$	合计
$A_1$	$n_{11}$	$n_{12}$	...	$n_{1c}$	$n_{1\cdot}$
$A_2$	$n_{21}$	$n_{22}$	...	$n_{2c}$	$n_{2\cdot}$
$\vdots$	$\vdots$	$\vdots$		$\vdots$	$\vdots$
$A_r$	$n_{r1}$	$n_{r2}$	...	$n_{rc}$	$n_{r\cdot}$
合计	$n_{\cdot 1}$	$n_{\cdot 2}$	...	$n_{\cdot c}$	$n$

其中,  $n_{i\cdot} = \sum_j n_{ij}$ ,  $n_{\cdot j} = \sum_i n_{ij}$ 。

### 3. 行列变量间关系的分析

列联表的频数分布不可能用来直接确定行、列变量之间的关系及关系的强弱。令人感兴趣的二维列联表的检验问题是行、列变量的独立性检验。

独立性检验指的是对列联表中行变量和列变量无关这个零假设进行的检验，即检验行、列变量之间是否彼此独立。常用的衡量变量间相关程度的统计量是简单相关系数，但在交叉列联表分析中，由于行、列变量往往不是连续等距变量，不符合计算简单相关系数的前提要求。

所以，一般采用的检验方法是卡方（ $\chi^2$ ）检验，它的计算公式为：

$$\chi^2 = \sum \frac{(f_0 - f_e)^2}{f_e}$$

其中， $f_0$ 表示实际观察频数， $f_e$ 表示期望频数。

## 3.4.2 列联表分析的SPSS操作详解

CONCEPT  
STRATE

### Step01 打开主窗口

选择菜单栏中的【Analyze(分析)】→【Descriptive Statistics(描述性统计)】→【Crosstabs(列联表)】命令，弹出【Crosstabs(列联表)】对话框，这是列联表分析的主操作窗口。



**Crosstabs**

城市 [主要城市]  
年平均气温 [年平均气温]  
地域 [地域]

Row(s):

Column(s):

Layer 1 of 1

Previous Next

Display layer variables in table layers

Display clustered bar charts

Suppress tables

OK Paste Reset Cancel Help

Exact...  
Statistics...  
Cells...  
Format...  
Bootstrap...

## Step02 选择行、列变量

CONCEPT  
STRATE

在【Crosstabs (列联表)】对话框左侧的候选变量列表框中，选取一个或多个待分析变量，将它们移入右侧的【Row(s) (行)】列表框中，作为列联表的行变量。同理，选择若干候选变量移入右侧的【Column(s) (列)】列表框中，作为列联表的列变量。

## Step03 选择层变量

CONCEPT  
STRATE

如果要进行三维或多维列联表分析，可以根据需要选择控制变量进入【Layer(层)】列表框中。该变量决定列联表的层。如果要增加另外一个控制变量，首先单击【Next】按钮，再选入一个变量。单击【Previous】按钮，可以重新选择以前确定的变量。

# Step04 列联表输出格式的选择

CONCEPT  
STRATE

在【Crosstabs (列联表)】对话框下面有两个复选框，用来选择列联表的输出格式。

- Display clustered bar charts: 显示各变量交叉分组下频数分布条形图。
- Suppress tables: 只输出统计量，而不输出列联表。

## Step05 行、列变量相关程度的度量

CONCEPT  
STRATE

在【Crosstabs (列联表)】对话框中单击【Statistics】按钮，在弹出的对话框中可以根据数据类型选择不同的独立性检验方法和相关度量。在对话框中选择输出统计量，完成后单击【Continue】按钮，返回主对话框。

**Crosstabs: Statistics** ✕

Chi-square  Correlations

**Nominal**

Contingency coefficient

Phi and Cramer's V

Lambda

Uncertainty coefficient

**Ordinal**

Gamma

Somers' d

Kendall's tau-b

Kendall's tau-c

**Nominal by Interval**

Eta

Kappa

Risk

McNemar

Cochran's and Mantel-Haenszel statistics

Test common odds ratio equals:

Continue
Cancel
Help

## Step06 选择列联表单元格的输出类型

CONCEPT  
STRATE

在【Crosstabs (列联表)】对话框中单击【Cells】按钮，在弹出的对话框中可以选择显示在列联表单元格中的统计量，包括观测数量、百分比和残差。在对话框中选择相应选项，完成后单击【Continue】按钮，返回主对话框。

**Crosstabs: Cell Display** [X]

**Counts**

Observed

Expected

Hide small counts  
Less than

**z-test**

Compare column proportions

Adjust p-values (Bonferroni method)

**Percentages**

Row

Column

Total

**Residuals**

Unstandardized

Standardized

Adjusted standardized

**Noninteger Weights**

Round cell counts       Round case weights

Truncate cell counts       Truncate case weights

No adjustments

Continue    Cancel    Help

# Step07 选择列联表单元格的输出排列顺序

CONCEPT  
TRATE

在【Crosstabs (列联表)】对话框中单击【Format】按钮，在弹出的对话框中可以选择各单元格的输出排列顺序。



## Step08 相关统计量的Bootstrap估计

CONCEPT  
RATE

单击【Bootstrap】按钮，在弹出的对话框中可以进行统计量的Bootstrap估计。

## Step09 完成操作



单击【OK】按钮，结束操作，SPSS软件自动输出结果。

# 3.4.3 实例图文分析：大学生身体素质调查



CONCEPT  
RATE

## 1. 实例内容

在一次上海大学生身体素质的实际调查中，选择了部分大专院校的学生进行实际问卷调查，收集的数据见3-4.sav。调查内容主要包括：性别、出生日期、身高、体重、血型、教育背景、学科、男女身高级别和男女体重级别等内容。请根据调查数据分析下面问题：

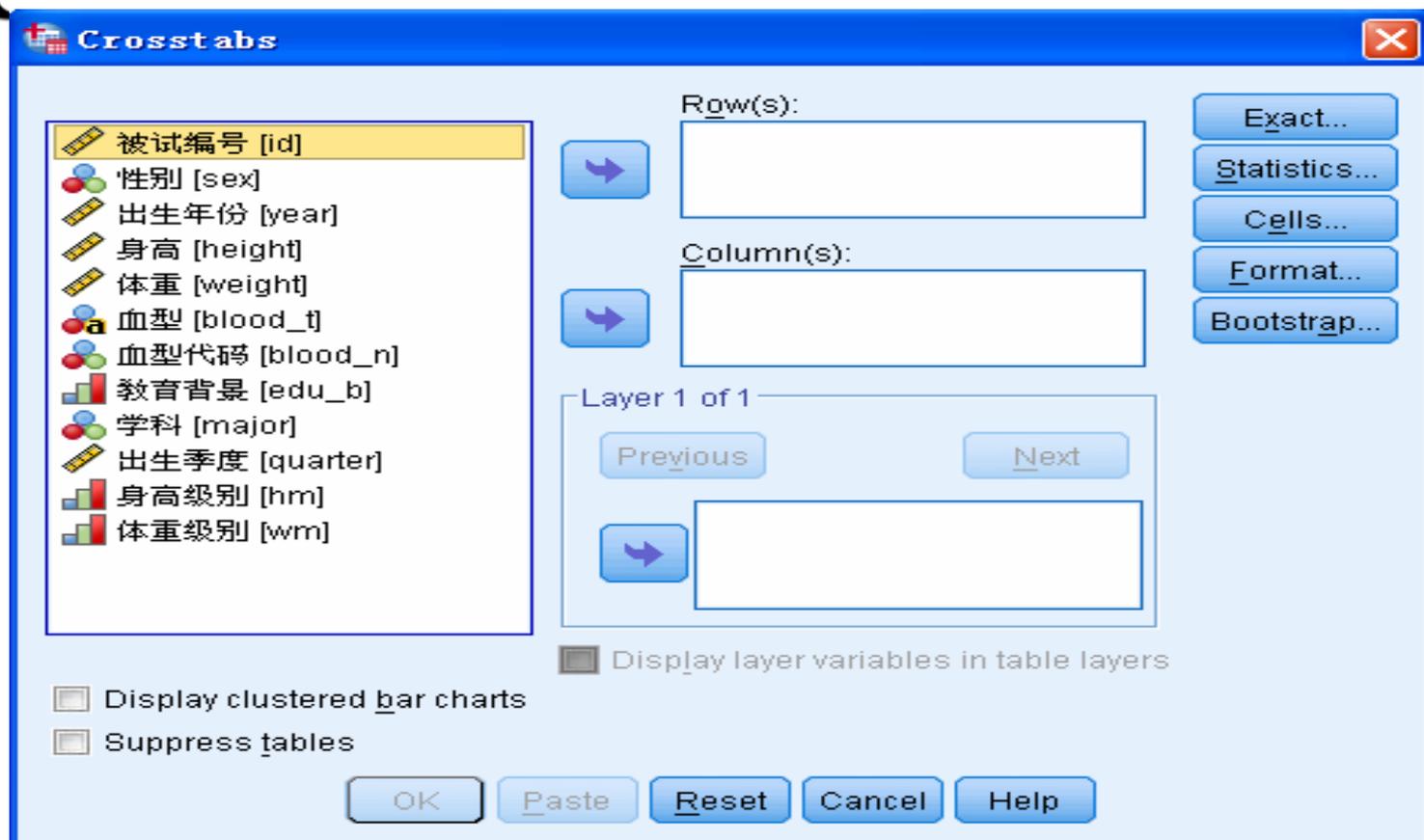
- (1) 进行“性别”和“体重级别”双因素交叉作用下的列联表分析，并研究“性别”对“体重级别”有无显著性影响。
- (2) 进行“教育背景”和“身高级别”双因素交叉作用下的列联表分析，并研究“教育背景”对“身高级别”有无显著性影响。

## Step01: 打开对话框



打开数据文件3-4. sav。选择菜单栏中的 **【Analyze(分析)】** → **【Descriptive Statistics(描述性统计)】** → **【Crosstabs(列联表)】** 命令，弹出 **【Crosstabs(列联表)】** 对话框。

CONCEPT  
STRATE



## Step02: 选择行、列变量

CONCEPT  
STRATE

- 在候选变量列表框中将变量“性别（sex）”添加至【Row(s) (行)】列表框中，表示它是交叉列联表中的行变量；将变量“体重级别（wm）”添加至【Column(列)】列表框中，表示它是交叉列联表中的列变量。

**Crosstabs** [Close]

- 被试编号 [id]
- 出生年份 [year]
- 身高 [height]
- 体重 [weight]
- 血型 [blood\_t]
- 血型代码 [blood\_n]
- 教育背景 [edu\_b]
- 学科 [major]
- 出生季度 [quarter]
- 身高级别 [hm]

Row(s):

性别 [sex]

---

Column(s):

体重级别 [wm]

Exact...

Statistics...

Cells...

Format...

Bootstrap...

Layer 1 of 1

Previous
Next

Display layer variables in table layers

Display clustered bar charts

Suppress tables

OK
Paste
Reset
Cancel
Help

## Step03: 独立性检验

CONCEPT  
STRATE

单击【Statistics】按钮，弹出【Crosstabs: Statistics(交叉表: 统计量)】对话框，勾选【Chi-square(卡方)】复选框，利用卡方检验来检验“性别”和“体重级别”的独立性。单击【Continue】按钮，返回【Crosstabs(列联表)】对话框。

**Crosstabs: Statistics** ✕

Chi-square  Correlations

**Nominal**

- Contingency coefficient
- Phi and Cramer's V
- Lambda
- Uncertainty coefficient

**Ordinal**

- Gamma
- Somers' d
- Kendall's tau-b
- Kendall's tau-c

**Nominal by Interval**

- Eta

- Kappa
- Risk
- McNemar

Cochran's and Mantel-Haenszel statistics

Test common odds ratio equals:

Continue
Cancel
Help

## Step04: 选择列联表输出格式

CONCEPT  
TRATE

由于要进行“性别”和“体重级别”的频数分析，因此单击【Cell】按钮，弹出【Crosstabs: Cell Display】对话框，勾选【Percentages】选项组中的【Row(行)】、【Column(列)】和【Total(总数)】复选框。单击【Continue】按钮，返回【Crosstabs(列联表)】对话框。

**Crosstabs: Cell Display** ✕

**Counts**

Observed

Expected

Hide small counts  
Less than

**z-test**

Compare column proportions

Ajust p-values (Bonferroni method)

**Percentages**

Row

Column

Total

**Residuals**

Unstandardized

Standardized

Ajusted standardized

**Noninteger Weights**

Round cell counts       Round case weights

Truncate cell counts       Truncate case weights

No addjustments

## Step05: 输出分布条形图



勾选【Display clustered bar charts (显示复式条形图)】复选框，表示利用条形图来反映不同性别之间的体重级别差异。

**Crosstabs** [Close]

- 被试编号 [id]
- 出生年份 [year]
- 身高 [height]
- 体重 [weight]
- 血型 [blood\_t]
- 血型代码 [blood\_n]
- 教育背景 [edu\_b]
- 学科 [major]
- 出生季度 [quarter]
- 身高级别 [hm]

Row(s):

性别 [sex]

---

Column(s):

体重级别 [wm]

Exact...

Statistics...

Cells...

Format...

Bootstrap...

Layer 1 of 1

Previous
Next

Display layer variables in table layers

Display clustered bar charts

Suppress tables

OK
Paste
Reset
Cancel
Help

## Step06: 完成操作



最后，单击【OK】按钮，操作完成。

# 实例结果及分析



## (1) 基本统计信息汇总

### 基本统计信息汇总

	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
性别 * 体重 级别	214	100.0 %	0	0.0%	214	100.0 %



## (2) “性别”和“体重级别”的列联表

			体重级别			Total
			轻 (60_)	中等 (60—70)	重 (70+)	
性别	男	Count	17	35	17	69
		% within 性别	24.6%	50.7%	24.6%	100.0%
		% within 体重级别	11.1%	85.4%	85.0%	32.2%
		% of Total	7.9%	16.4%	7.9%	32.2%
	女	Count	136	6	3	145
		% within 性别	93.8%	4.1%	2.1%	100.0%
		% within 体重级别	88.9%	14.6%	15.0%	67.8%
		% of Total	63.6%	2.8%	1.4%	67.8%
Total	Count	153	41	20	214	
	% within 性别	71.5%	19.2%	9.3%	100.0%	
	% within 体重级别	100.0%	100.0%	100.0%	100.0%	
	% of Total	71.5%	19.2%	9.3%	100.0%	

- (3) “性别”和“体重级别”的独立性检验

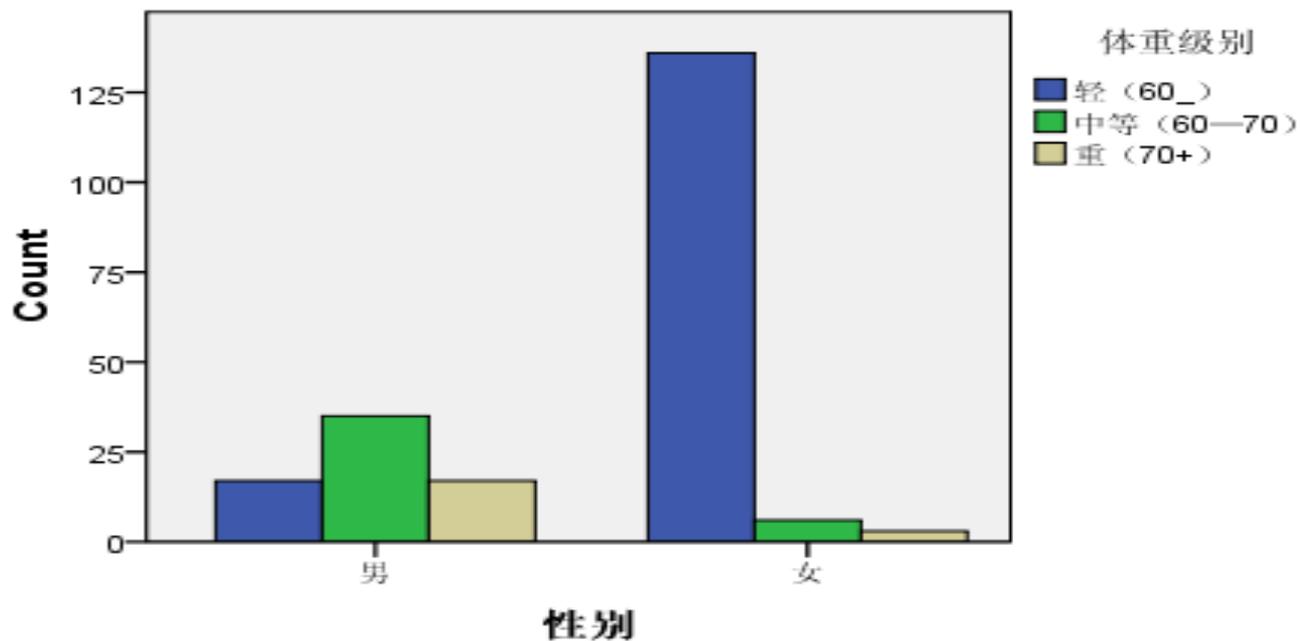
### 卡方检验结果

	Value	<i>df</i>	Asymp. Sig. (2-sided)
Pearson Chi-Square	109.715 <sup>a</sup>	2	0.000
Likelihood Ratio	111.290	2	0.000
Linear-by-Linear Association	92.739	1	0.000
N of Valid Cases	214		



- (4) 体重级别条形图

Bar Chart



## 3.5 SPSS在比率分析中的应用

CONCEPT  
RATE

### 3.5.1 比率分析的基本原理

比率分析生成比率变量，并对该比率变量计算基本描述性统计量（如均值、中位数、标准差、全距等），进而刻画出比率变量的集中趋势和离散程度。除此之外，SPSS 19.0还提供了其他对比描述指标，大致也属于集中趋势描述指标和离散程度描述指标的范畴。

## 3.5.2 比率分析的SPSS操作详解

CONCEPT  
RATE

Step01: 打开主窗口

选择菜单栏中的【Analyze(分析)】→【Descriptive Statistics(描述性统计)】→【Ratio(比率)】命令，弹出【Ratio(比率)】对话框，这是比率分析的主操作窗口。

### Ratio Statistics

 城市 [主要城市]  
 年平均气温 [年平均气温]  
 地域 [地域]

Display results  
 Save results to external file

File... Statistics...

OK Paste Reset Cancel Help

Numerator:  
Denominator:  
Group Variable:

Sort by group variable  
 Ascending order  
 Descending order

## Step02: 选择分子变量

CONCEPT  
RATE

在左侧的候选变量列表框中选取一个分析变量作为比率分析的分子，将它移入右侧的【Numerator(分子)】列表框中。

## Step03: 选择分母变量

CONCEPT  
RATE

在【Ratio Statistics(比率统计量)】对话框左侧的候选变量列表框中选取一个分析变量作为比率分析的分母，将它移入右侧的【Denominator(分母)】列表框中。

## Step04: 选择分组变量

CONCEPT  
RATE

在【Ratio Statistics(比率统计量)】对话框左侧的候选变量列表框中选取一个变量作为分组变量，将它移入右侧的【Group Variable(组变量)】列表框中。

## Step05: 结果显示选择

CONCEPT  
RATE

在【Ratio Statistics(比率统计量)】对话框中，用户可以选择比率分析的结果输出类型。

- Display result: 系统默认选项，选择是否显示结果。
- Save results to external file: 选择是否将分析结果保存至外部文件。同时，外部文件的保存路径需要单击【File】按钮来选择。



## Step06: 选择描述性统计量输出

单击【Statistics】按钮，弹出的【Ratio Statistics: Statistics】对话框主要用于输出各类基本统计量结果。

**Ratio Statistics: Statistics** [X]

**Central Tendency**

- Median
- Mean
- Weighted Mean
- Confidence intervals:  
Level (%):

**Dispersion**

- AAD
- COD
- PRD
- Median Centered COV
- Mean Centered COV
- Standard deviation
- Range
- Minimum
- Maximum

**Concentration Index**

**Between Proportions**

Low Proportion:

High Proportion:

Pairs:

**Within Percentage of Median**

Percentage of median:

Percentages:

## Step07 完成操作



单击【OK】按钮，结束操作，SPSS软件自动输出结果。

# 3.5.3 实例图文分析：城乡消费水平区域对比



## 1. 实例内容

	省份	农村居民	城镇居民	区域
1	北京	6635.00	16683.00	1
2	天津	4360.00	11394.00	1
3	河北	2449.00	7927.00	1
4	山西	2146.00	7104.00	1
5	内蒙古	2460.00	7103.00	1
6	辽宁	3267.00	8688.00	2
7	吉林	2467.00	7556.00	2
8	黑龙江	2419.00	6958.00	2
9	上海	9157.00	19573.00	3
10	江苏	4207.00	10199.00	3
11	浙江	5476.00	14097.00	3
12	安徽	2177.00	7136.00	3
13	山东	3078.00	9453.00	3
14	河南	2372.00	8145.00	4
15	湖北	2503.00	8051.00	4
16	湖南	2855.00	8477.00	4
17	重庆	2251.00	7959.00	5
18	四川	2432.00	7577.00	5
19	贵州	1563.00	7498.00	5
20	云南	1913.00	8285.00	5
21	西藏	1532.00	9040.00	5
22	陕西	2024.00	8234.00	6
23	甘肃	1812.00	7410.00	6
24	青海	1941.00	6947.00	6
25	宁夏	2231.00	7495.00	6
26	新疆	1884.00	7311.00	6

城乡居民消费水平

## Step01: 打开对话框



打开SPSS软件，选择菜单栏中的【Analyze(分析)】→【Descriptive Statistics(描述性统计)】→【Ratio(比率)】命令，弹出【Ratio Statistics(比率统计量)】对话框。

### Ratio Statistics

省份  
 农村居民  
 城镇居民  
 区域

Display results  
 Save results to external file

File... Statistics...

OK Paste Reset Cancel Help

Numerator: \_\_\_\_\_  
Denominator: \_\_\_\_\_  
Group Variable: \_\_\_\_\_

Sort by group variable  
 Ascending order  
 Descending order

## Step02: 选择分子变量

CONCEPT  
RATE

在【Ratio Statistics(比率统计量)】对话框左侧的候选变量列表框中，选取变量“城镇居民”作为比率分析的分子，将它移入右侧的【Numerator(分子)】列表框中。

## Step03: 选择分母变量

CONCEPT  
RATE

在【Ratio Statistics(比率统计量)】对话框左侧的候选变量列表框中，选取变量“农村居民”作为比率分析的分母，将它移入右侧的【Denominator(分母)】列表框中。

## Step04: 选择分组变量

CONCEPT  
RATE

在【Ratio Statistics(比率统计量)】对话框左侧的候选变量列表框中，选取变量“区域”作为分组变量，将它移入右侧的【Group Variable(组变量)】列表框中。

## Step05: 选择输出统计量

CONCEPT  
STRATE

单击【Statistics】按钮，在弹出的对话框中除了保留系统默认的输出统计量外，再勾选【Media(中位数)】、【Mean(均值)】和【ADD】复选框。最后单击【Continue】按钮，返回【Ratio Statistics(比率统计量)】对话框。

**Ratio Statistics: Statistics** [X]

**Central Tendency**

Median

Mean

Weighted Mean

---

Confidence intervals:  
Level (%):

**Dispersion**

AAD

Standard deviation

COD

Range

PRD

Minimum

Median Centered COV

Maximum

Mean Centered COV

**Concentration Index**

**Between Proportions**

Low Proportion:

High Proportion:

Pairs:

**Within Percentage of Median**

Percentage of median:

Percentages:



## 2. 实例结果及分析

### (1) 样本统计结果输出

样本统计结果表

		Count	Percent
区域	华北	5	19.2%
	东北	3	11.5%
	华东	5	19.2%
	华中	3	11.5%
	西南	5	19.2%
	西北	5	19.2%
Overall		26	100.0%
Excluded		0	
Total		26	



## (2) 比率分析结果表

### 比率分析结果表

Group	Mean	Median	Average Absolute Deviation	Price Related Differential	Coefficient of Dispersion	Coefficient of Variation
						Median Centered
华北	2.912	2.887	0.284	1.047	0.098	12.4%
东北	2.866	2.876	0.135	1.007	0.047	7.0%
华东	2.697	2.574	0.357	1.075	0.139	19.0%
华中	3.207	3.217	0.155	1.005	0.048	7.2%
西南	4.336	4.331	0.809	1.041	0.187	25.3%
西北	3.795	3.881	0.244	1.004	0.063	8.6%
Overall	3.343	3.227	0.565	1.098	0.175	25.4%



## 第4章SPSS的均值比较过程

SPSS主要有以下模块实现均值比较过程。

- One-Sample T Test: 单样本 t 检验。
- Independent-Sample T Test: 两个独立样本均值的 t 检验。
- Paired-Sample T Test: 两个配对样本均值的 t 检。

## 4.1 SPSS在单样本t检验的应用

CONCEPT  
RATE

- 1. 使用目的

单样本t检验的目的是利用来自某总体的样本数据，推断该总体的均值是否与指定的检验值之间存在明显的差异。它是对**总体均值的假设检验**。



## 2. 基本原理

单样本t检验作为假设检验的一种方法，其基本步骤和假设检验相同。其零假设为 $H_0$ ：总体均值与指定检验值之间不存在显著差异。该方法采用t检验方法，按照下式计算t统计量。

$$t = \frac{\overline{D}}{S / \sqrt{n}}$$

式中，D是样本均值与检验值之差；因为总体方差未知，故用样本方差S代替总体方差；n为样本数。

### 3. 概率P值

如果概率P值小于或等于显著性水平，则拒绝零假设；

如果概率P值大于显著性水平，则接受零假设。

## 4. 软件使用方法

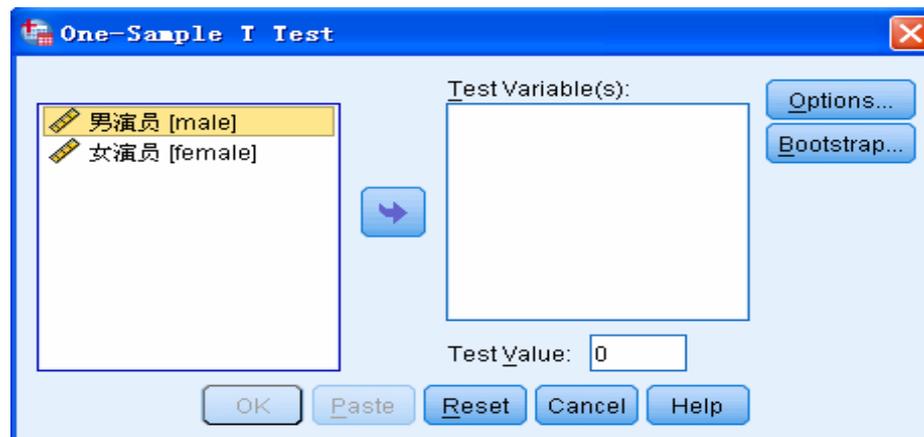
- (1) 在SPSS中，软件将自动计算t值，由于该统计量服从n-1个自由度的t分布，SPSS将根据t分布表给出t值对应的相伴概率P值。
- (2) 如果相伴概率P值小于或等于给定的显著性水平，则拒绝 $H_0$ ，认为总体均值与检验值之间存在显著差异。
- (3) 相反，相伴概率值大于给定的显著性水平，则不应拒绝 $H_0$ ，可以认为总体均值与检验值之间不存在显著差异。

## 4.1.2 单样本t检验的SPSS操作详解

CONCEPT  
STRATE

Step01: 打开单样本  $t$  检验对话框。

选择菜单栏中的【Analyze（分析）】→【Compare Means（比较均值）】→【One-Sample T Test（单样本T检验）】命令，弹出【One-Sample T Test（单样本T检验）】对话框。



## Step02: 选择检验变量。

在该对话框左侧的候选变量列表框中选择一个或几个变量，将其移入【Test Variable(s)（检验变量）】列表框中。其中，左侧候选变量列表框中显示的是可以进行  $t$  检验的变量。

Step03: 选择样本检验值。

在【Test Value (检验值)】文本框中输入检验值，相当于假设检验问题中提出的零假设 $H_0$ ：

$$\mu = \mu_0。$$

## Step04: 其他选项设置。

单击【Options】按钮，弹出【One-Sample T Test: Options(单样本T检验: 选择)】对话框。该对话框用于指定输出内容和关于缺失值的处理方法，其中各选项的含义如下。

Confidence Interval: 该文本框用于设置在指定水平下，样本均值与指定的检验值之差的置信区间，默认值为95%。

【Missing Values（缺失值）】选项组：用于设置缺失值的处理方式，它有以下两种处理方式。

- Exclude cases analysis by analysis: 点选该单选钮，表示当分析计算涉及到含有缺失值的变量时，删除该变量上是缺失值的观测量。
- Exclude cases listwise: 点选该单选钮，表示删除所有含缺失值的观测量后再进行分析。

## Step05: 相关统计量的Bootstrap估计

单击【Bootstrap】按钮，在弹出的对话框中可以进行如下统计量的Bootstrap估计。

- 支持均值和标准差的Bootstrap 估计。
- 支持平均值差值的Bootstrap 估计和显著性检验。

Step06: 单击【OK】按钮结束操作，SPSS软件自动输出结果。



## 2 实例操作

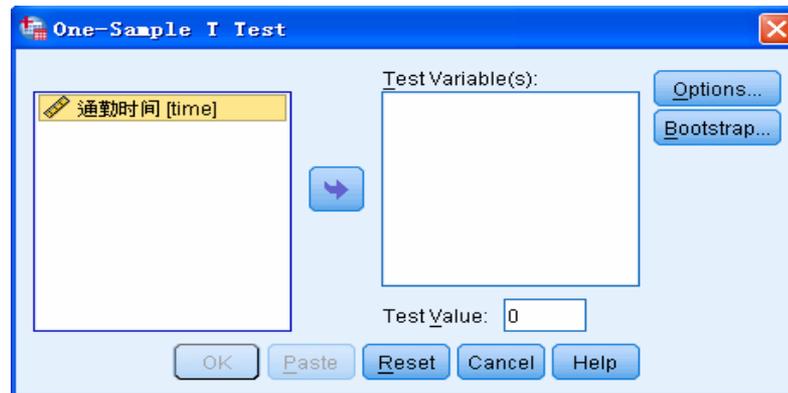


现在该名研究者要检验他所在城市的平均通勤时间和全国其他城市平均水平是否一致。由于题目中已给出了其他城市通勤时间的平均水平为19分钟，因此，这里就是要检验该城市通勤时间是否等于19分钟，即进行如下假设检验：

$$H_0 : t = 19; \quad H_1 : t \neq 19$$

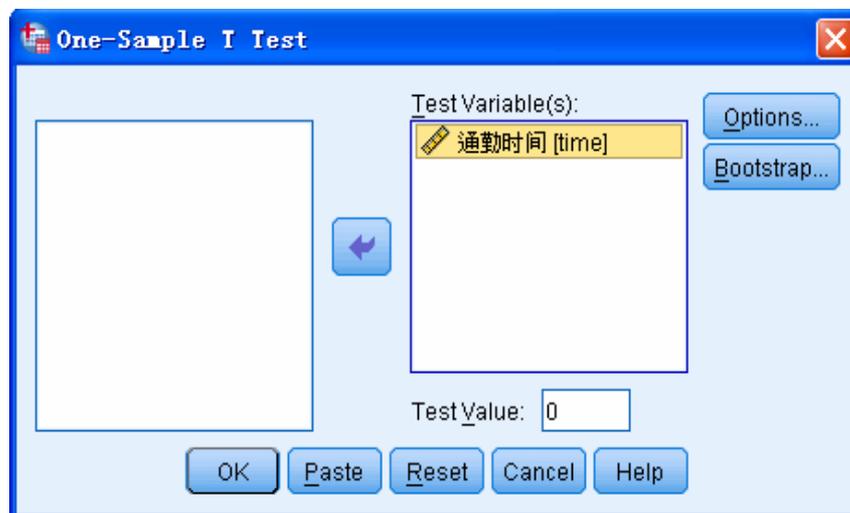
# Step01: 打开对话框

打开数据文件4-1.sav, 选择菜单栏中的【Analyze (分析)】→【Compare Means (比较均值)】→【One-Sample T Test (单样本T检验)】命令, 弹出【One-Sample T Test (单样本T检验)】对话框。



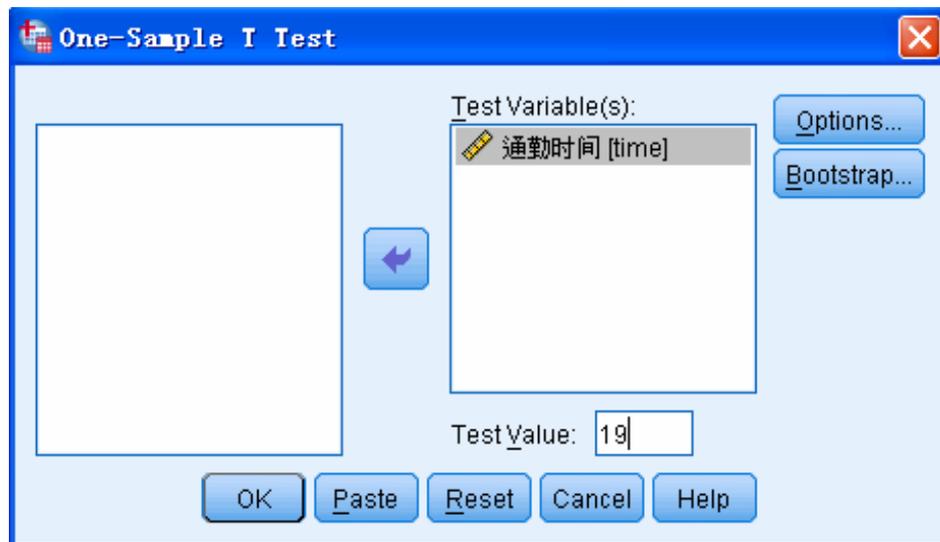
## Step02: 选择检验变量

在候选变量列表框中选择“time”变量，将其添加至【Test Variables（检验变量）】列表框中。



## Step03: 选择样本检验值

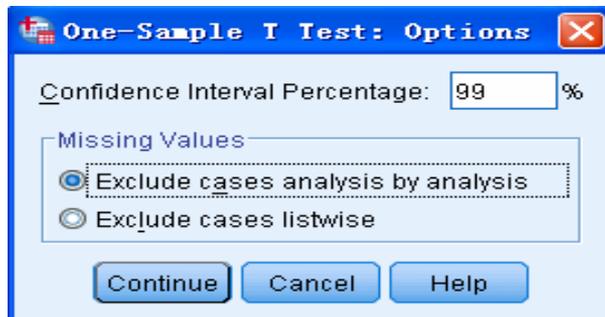
在【Test Value（检验值）】文本框中输入检验值“19”。



## Step04: 设置显著性水平

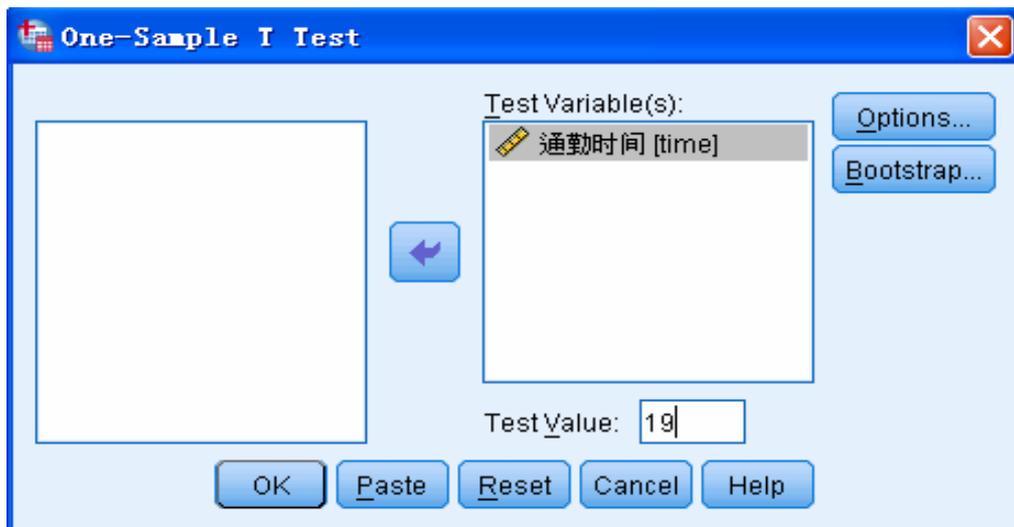
单击【Options】按钮，在弹出的对话框的【Confidence Interval Percentage（置信区间百分比）】文本框中将系统默认的95%修改为 99%，其目的是调整显著性水平。单击【Continue】按钮返回主对话框。

提示：如果不选择Options按钮，表示默认系统选项参数设置。



## Step05: 结束操作

单击OK按钮，完成操作。此时，软件输出结果出现在结果浏览窗口中。





### 3. 实例结果及分析

#### (1) 描述性统计分析表

	N	Mean	Std. Deviation	Std. Error Mean
通勤时间	26	19.5385	3.75479	.73638



## (2) 单样本t检验结果

	Test Value = 19					
	t	df	Sig. (2-tailed)	Mean Difference	99% Confidence Interval of the Difference	
					Lower	Upper
通勤时间	.731	25	.471	.53846	-1.5141	2.5911

## 4.1.4 实例进阶分析：机票的折扣费

CONCEPT  
STRATE

### 1. 实例内容

1995年2月，某个航班往返机票的平均折扣费是258美元（《今日美国》，1995年3月30日）。随机抽取了在3月份中15个往返机票的折扣费作为一个简单随机样本，结果得到下面的数据：

310 260 265 255 300 310 230

250 265 280 290 240 285 250 260

请你检验3月份往返机票的折扣费是否有所增加？

## 2 实例操作

CONCEPT  
TRATE

由于3月份机票的平均折扣费是258美元，而现在调查抽取了15个数据，可以计算得到它们的样本均值（Mean）等于270美元。从数值大小看到明显折扣费用增加了。但是，这种数值的增加是由实际情况变动还是抽样误差造成的，则可以通过单样本的t检验来验证。这里建立如下假设检验：

$$H_0: price = 258;$$

$$H_1: price \neq 258$$



由于单样本t检验要求样本数据服从正态分布，因此进行单样本的K-S检验，得到检验分析表。从检验结果看到，统计量Z等于0.697，相伴概率P等于0.716，远大于显著性水平，因此接受零假设，认为该数据服从正态分布，可以利用单样本t检验方法。具体操作步骤如下。

		机票折扣费
N		15
Normal Parameters <sup>a</sup>	Mean	270.00
	Std. Deviation	24.785
Most Extreme Differences	Absolute	.180
	Positive	.180
	Negative	-.087
Kolmogorov-Smirnov Z		.697
Asymp. Sig. (2-tailed)		.716

# Step01

CONCEPT  
STRATE

打开数据文件4-2. sav，选择菜单栏中的【Analyze（分析）】→【Compare Means（比较均值）】→【One-Sample T Test（单样本T检验）】命令，弹出【One-Sample T Test（单样本T检验）】对话框。

## Step02

CONCEPT  
STRATE

在候选变量列表框中选择“pirce”变量，将其添加至【Test Variables（检验变量）】列表框中。

## Step03



在【Test Value（检验值）】文本框中输入检验值“258”。

## Step04

ONCEPT  
TRATE

单击【OK】按钮，完成操作。



### 3. 实例结果及分析

下表所示为单样本t检验的分析结果，表格中各项的含义前面已经详细讲解了。由于这里双侧概率P值0.082略大于显著性水平0.05，因此接受零假设，认为3月份往返机票的折扣费没有变化。

单样本t检验分析结果

	Test Value = 258					
	t	df	Sig. (2-tailed)	Mean Difference	95% Confidence Interval of the Difference	
					Lower	Upper
机票折扣费	1.875	14	.082	12.000	-1.73	25.73

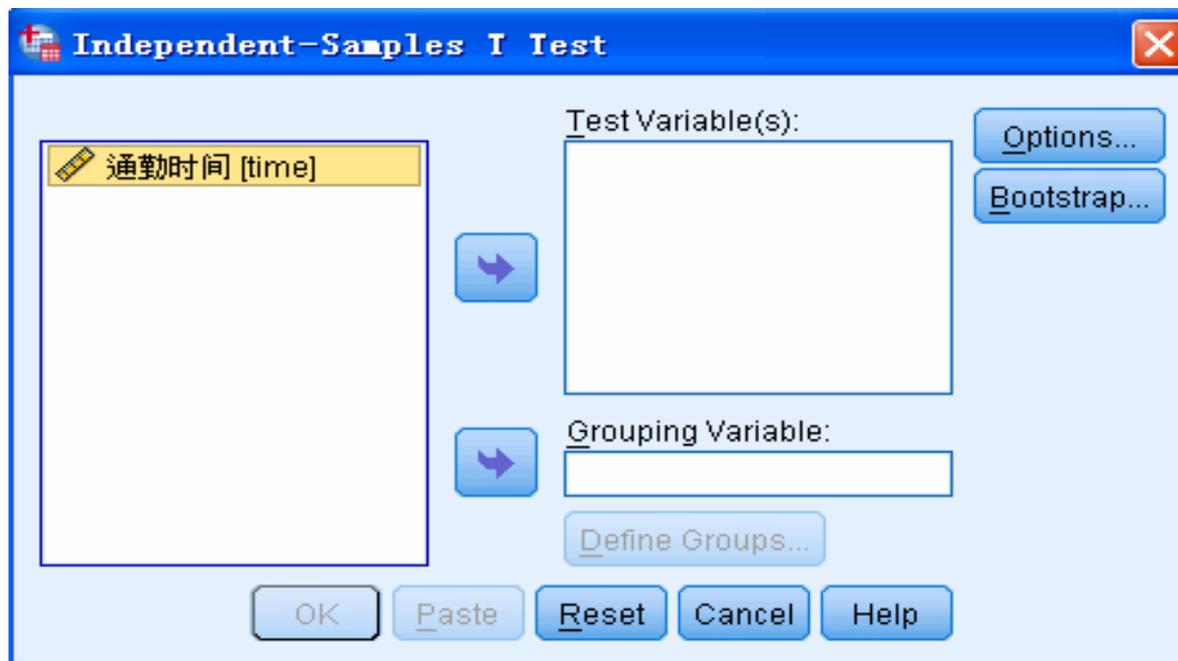
## 4.2 SPSS在两独立样本t检验的应用

CONCEPT  
STRATE

### 4.2.2 两独立样本t检验的SPSS操作步骤

Step01: 打开两独立样本  $t$  检验对话框。

选择菜单栏中的【Analyze（分析）】→【Compare Means（比较均值）】→【Independent-Samples T Test（独立样本T检验）】命令，弹出【Independent-Samples T Test（独立样本T检验）】对话框。



## Step02: 选择检验变量

CONCEPT  
TRATE

- 在左侧的候选变量列表框中选择检验变量，将其移入【Test Variable(s) (检验变量)】列表框中，这里需要选入待检验的变量。

## Step03: 选择分组变量

CONCEPT  
STRATE

在左侧的候选变量列表框中选择分组变量，将其移入【Grouping Variable(分组变量)】文本框中，目的是区分检验变量的不同组别。

## Step04 定义组别名称



单击【Define Groups】按钮，弹出【Define Groups（定义组）】对话框，此时需要定义进行  $t$  检验的比较组别名称。

该对话框中各选项的含义如下。

Use specified values: 分别输入两个对应不同总体的变量值。

Cut point: 用于定义分割点值。在该文本框中输入一个数字，大于等于该数值的对应一个总体，小于该值的对应另一个总体。

在该对话框中设置完成后，单击【Continue】按钮，返回【Independent-Samples T Test（独立样本T检验）】对话框。



## Step05: 相关统计量的Bootstrap估计

CONCEPT  
STRATE

单击【Bootstrap】按钮，在弹出的对话框中可以进行如下统计量的Bootstrap估计。

- 支持均值和标准差的Bootstrap 估计。
- 支持平均值差值的Bootstrap 估计和显著性检验。

## Step06

CONCEPT  
STRATE

单击【OK】按钮，结束操作，SPSS软件自动输出相关结果。

## 4.2.3 实例图文分析：机场等级分数比较

CONCEPT  
STRATE

### 1. 实例内容

国际航空运输协会 (The International Air Transport Association) 对商务旅游人员进行了一项调查，以便确定多个国际机场的等级分数。最高可能分数是10分，分数越高说明其等级也越高。假设有一个由50名商务旅行人员组成的简单随机样本，要求这些人给迈阿密机场打分。另外有一个由50名商务旅行人员组成的样本，要求这些人给洛杉矶机场打分。这两个组人员打出的等级分数如表4-5所示。请你判断迈阿密机场和洛杉矶机场的等级评分是否相同？



表4-5 两组人员打出的等级分数

城市	等级分数
迈阿密	6 4 6 8 7 7 6 3 3 8 10 4 8 7 8 7 5 9 5 8 4 3 8 5 5 4 4 4 8 4 5 6 2 5 9 9 8 4 8 9 9 5 9 7 8 3 10 8 9 6
洛杉矶	10 9 6 7 8 7 9 8 10 7 8 5 7 3 5 6 8 7 10 8 4 7 8 6 9 9 5 3 1 8 9 6 8 5 4 6 10 9 8 3 2 7 9 5 3 10 3 5 10 8

↵

## 2 实例操作



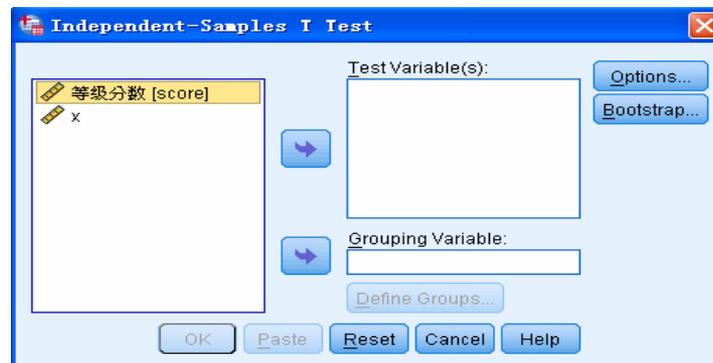
本案例中共有两组商务旅行人员分别对迈阿密和洛杉矶机场打分。由于这两组人员构成不同，因此由这两组人员组成的样本可以看作是相互独立的。现在要比较这两个机场的平均得分是否相同，也就是要检验这两个独立样本的均值是否相同，因此可以采用两独立样本t检验的方法。于是建立如下假设检验：

$H_0$  : 迈阿密机场和洛杉矶机场的等级得分相同。

$H_1$  : 迈阿密机场和洛杉矶机场的等级得分不同。

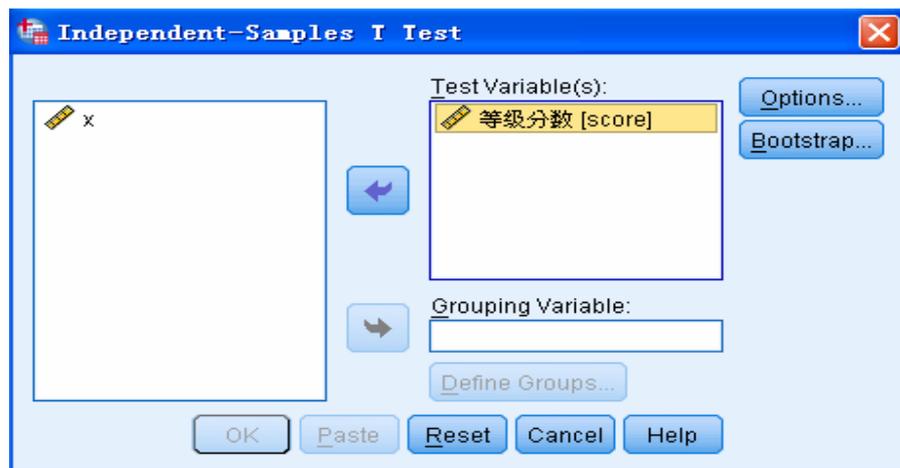
# Step01: 打开对话框

选择菜单栏中的【Analyze（分析）】→【Compare Means（比较均值）】→【Independent-Sample T Test（独立样本T检验）】命令，弹出【Independent-Sample T Test（独立样本T检验）】对话框，。这里变量score表示两个机场的得分；变量x是不同机场的标志变量，1表示迈阿密机场，2表示洛杉矶机场。



## Step02: 选择检验变量

在左侧的候选变量列表框中选择检验变量“score”，将其添加至右侧的【Test Variable(s) (检验变量)】列表框中，表示需要对其进行独立样本的T检验。

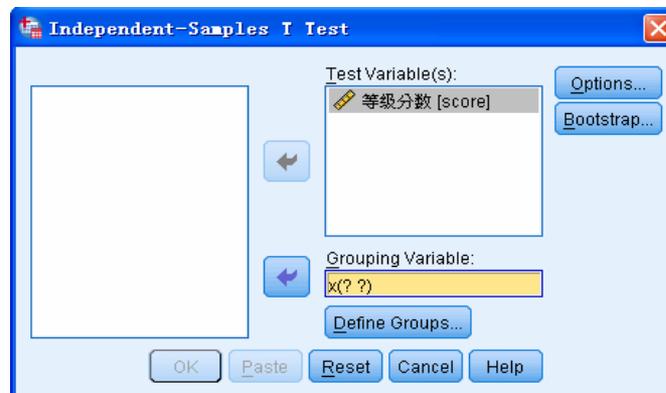


## Step03: 选择分组变量

CONCEPT  
STRATE

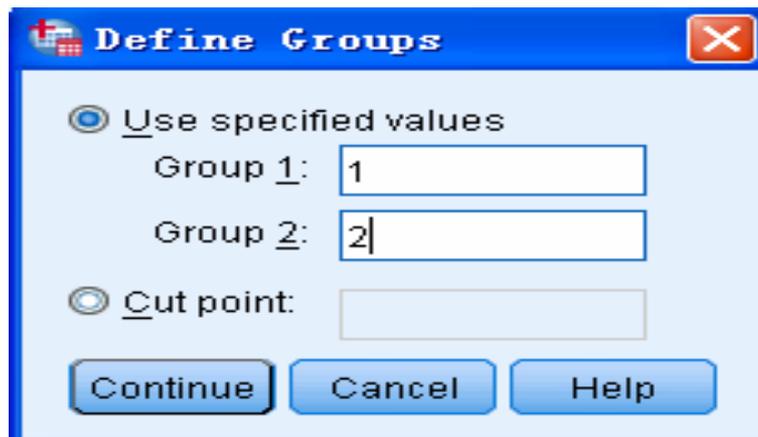
在左侧的候选变量列表框中选择分组变量“x”，将其添加至【Grouping Variable(s) (组变量)】文本框中。接着单击【Define Groups】按钮，弹出【Define Group (定义组)】对话框。

提示：如果不单击【Options】按钮，表示默认系统选项参数设置。



## Step04: 定义组别名称

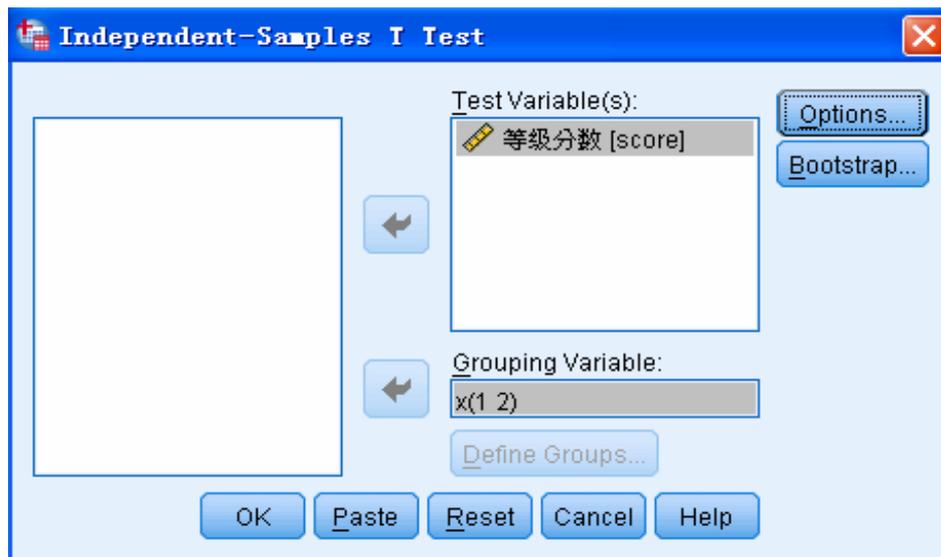
点选【Use specified values (使用指定值)】单选钮，在【Group1 (组1)】文本框中输入“1”，在【Group2 (组2)】文本框中输入“2”。输入完成后，单击【Continue】按钮返回。





## Step05: 完成操作

单击【OK】按钮，完成操作。此时，软件输出结果出现在结果浏览窗口中。





### 3. 实例结果及分析

#### (1) 基本统计信息汇总表

	x <sup>2</sup>	N <sup>2</sup>	Mean <sup>2</sup>	Std. Deviation <sup>2</sup>	Std. Error Mean <sup>2</sup>
等级分数 <sup>2</sup>	迈阿密 <sup>2</sup>	50	6.34	2.163	.306
	洛杉矶 <sup>2</sup>	50	6.76	2.378	.336

## (2) 独立两样本的t检验分析结果

### ①两总体方差是否相等的F检验

这里，该检验的F统计量的观察值为0.086，对应的概率P值为0.770。由于系统默认显著性水平 $\alpha$ 为0.05，而概率P值显然大于0.05，因此认为两总体的方差无显著性差异。

## ②两总体均值的检验

在SPSS中进行两独立样本t检验时，应首先对F检验作判断。如果方差相等，观察分析结果中Equal variances assumed列的t检验相伴概率值；如果方差不相等，观察Equal variances not assumed列的t检验相伴概率值。本案例的第一步分析中，由于两总体方差无显著差异，因此应看第一列（Equal variance assumed）的t检验结果。具体来说，t统计量的观测值为-0.924，对应的双尾概率P值为0.358，大于显著性水平0.05，因此认为两总体的均值不存在显著差异，即迈阿密机场和洛杉矶机场的等级得分相同。这个结论说明商务人员认为两个机场在服务水平质量等方面是没有差异的。

表 4-7 独立两样本的 t 检验分析结果

	Levene's Test for Equality of		t-test for Equality of Means						
	Variances		t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
	F	Sig.						Lower	Upper
Equal variances assumed	.086	.77	-.924	98	.358	-.42	.455	-1.322	.482
Equal variances not assumed			-.924	97.131	.358	-.42	.455	-1.322	.482

## 4.2.4 实例进阶分析：考试中的惊惶失措

CONCEPT  
STRATE

- 1. 实例内容

许多学生都有一次考试中因为第一道题目特别难而惊惶失措的不愉快经历。人们对考试题目的安排进行了研究，以弄清它对焦虑的影响。表4-8所示的分数是对“测验焦虑”的度量，有充分的证据支持考试题目的安排对分数有影响这一假设吗？

表 4-8 “测验焦虑”值列表<sup>4</sup>

方式 <sup>4</sup>	“测验焦虑”值 <sup>4</sup>									
问题从易到 难安排 <sup>4</sup>	24.64	39.29	16.32	32.83	28.02	33.31	20.60	21.13	26.69	28.90 <sup>4</sup>
	26.43	24.23	7.10	32.86	21.06	28.89	28.71	31.73	30.02	21.96 <sup>4</sup>
	24.49	38.81	27.85	30.29	30.72 <sup>4</sup>					
问题从难到 易安排 <sup>4</sup>	33.62	34.02	26.63	30.26	35.91	26.68	29.49	35.32	27.24	32.34 <sup>4</sup>
	29.34	33.53	27.62	42.91	30.20	32.54 <sup>4</sup>				

## 2 实例操作

CONCEPT  
STRATE

- 表4-8列出了两种考试方式下不同学生的焦虑测量值，其值越大，说明学生考试时越焦虑。现在要研究考试题目对分数的影响性，即比较这两种考试形式对学生有无显著的焦虑差异性。考虑到选取的学生不同，因此可以利用两独立样本的t检验，建立假设检验如下。

$H_0$ : 两种考试方式下学生的平均焦虑测量值相同。

$H_1$ : 两种考试方式下学生的平均焦虑测量值不同。

# Step01

CONCEPT  
STRATE

建立数据文件4-4. sav。这里变量anxiety表示两个机场的得分；变量x表示不同的考试方式，1表示问题从易到难安排，2表示各问题从难到易安排。

## Step02

CONCEPT  
STRATE

选择菜单栏中的【Analyze（分析）】→【Compare Means（比较均值）】→【Independent-Sample T Test（独立样本T检验）】命令，弹出【Independent-Sample T Test（独立样本T检验）】对话框。

## Step03

CONCEPT  
STRATE

- 在左侧的候选变量列表框中选择检验变量 anxiety，将其添加至【Test Variable(s) (检验变量)】列表框中。

## Step04



选择分组变量x，将其添加至【Grouping Variable(s) (分组变量)】文本框中。

## Step05

CONCEPT  
STRATE

单击【Define Groups】按钮，弹出【Define Group（定义组）】对话框。点选【Use specified values】单选钮，在【Group1（组1）】文本框中输入“1”，在【Group2（组2）】文本框中输入“2”。输入完成后，单击【Continue】按钮，关闭【Define Group（定义组）】对话框。

## Step06



单击【OK】按钮，结束操作。



### 3. 实例结果及分析

#### (1) 基本统计信息汇总表

	不同考试形式	N	Mean	Std. Deviation	Std. Error Mean
焦虑测量值	问题从易到难安排	25	27.0752	6.86988	1.37398
	问题从难到易安排	16	31.7281	4.26015	1.06504

## ②两总体均值的检验

在首先进行的方差相等假设检验中，F统计量等于1.986，对应的概率P值为0.167，大于显著性水平0.05，因此认为两组数据的方差是相等的。于是接着观察“Equal variance assumed”列所对应的t检验结果。由于t统计量对应的双尾概率P值为0.020，小于显著性水平0.05，因此认为两总体的均值存在着统计意义下的显著性差异。所以，问题“从易到难”和“从难到易”两种方式的题目设置安排，对学生考试产生了显著的焦虑影响，其平均焦虑值从27.0752上升至31.7281。所以，出题人在设置试卷考试难度的分配时，要予以充分的考虑。

表 4-10 独立 t 检验结果

		Levene's Test for Equality of Variances		t-test for Equality of Means						
		F	Sig.	t	df	Sig. (2-tailed)	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference	
									Lower	Upper
焦虑 测量 值	Equal variances assumed	1.986	.167	-2.421	39	.02	-4.65292	1.92157	-8.53966	-0.76619
	Equal variances not assumed			-2.677	38.986	.011	-4.65292	1.73842	-8.16926	-1.13659

## 4.3 SPSS在两配对样本t检验的应用

CONCEPT  
STRATE

### 4.3.1 两配对样本t检验的基本原理

#### 1. 使用目的

前一节中考虑的是独立样本情形下的总体均值相等的检验问题。但在现实中，总体或样本之间不仅仅表现为独立的关系，很多情况下，总体之间存在着一定的相关性。当分析这些相关总体之间的均值关系时，就涉及到两配对样本的t检验。

## 2. 基本原理

两配对样本t检验的目的是利用来自两个总体的配对样本，推断两个总体的均值是否存在显著差异。它和独立样本t检验的差别就在于要求样本是配对的。由于配对样本在抽样时不是相互独立的，而是相互关联的，因此在进行分析时必须考虑到这种相关性，否则会浪费大量的统计信息，因此对于符合配对情况的统计问题，要首先考虑两配对样本t检验。配对样本主要包括下列一些情况。

- (1) 同一实验对象处理前后的数据。例如对患肝病的病人实施某种药物治疗后，检验病人在服药前后的差异性。
- (2) 同一实验对象两个部位的数据。例如研究汽车左右轮胎耐磨性有无显著差异。
- (3) 同一样品用两种方法检验的结果。例如对人造纤维在60度和80度的水中分别作实验，检验温度对这种材料缩水率的影响性。
- (4) 配对的两个实验对象分别接受不同处理后的数据。例如对双胞胎兄弟实施不同的教育方案，检验他们在学习能力上的差异性。

### 3. 使用条件

进行配对样本检验时，通常要满足以下三个要求。

- (1) 两组样本的样本容量要相同；
- (2) 两组样本的观察值顺序不能随意调换，要保持一一对应关系；
- (3) 样本来自的总体要服从正态分布。

两配对样本t检验的基本思路是求出每对数据的差值：如果配对样本没有差异，则差值的总体均值应该等于零，从该总体中抽取的样本均值也应该在零值附近波动；反之，如果配对样本有差异，差值的均值就该远离零值。这样，通过检验该差值样本的均值是否等于零，就可以判断这两组配对样本有无差异性。

该检验对应的假设检验如下。

$H_0$ ：两总体均值之间不存在显著差异。

$H_1$ ：两总体均值之间存在显著性差异。

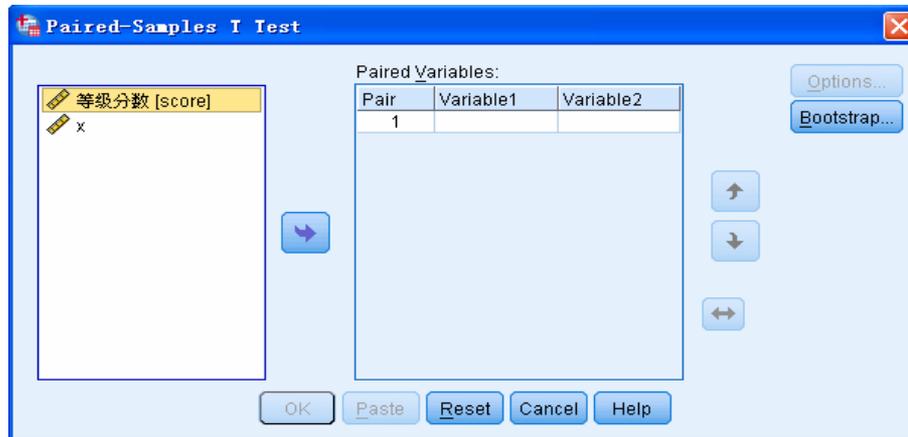
检验中所采用的统计量和单样本t检验完全相同

## 4.3.2 两配对样本t检验的SPSS操作详解

CONCEPT  
TRATE

Step01: 打开两配对样本  $t$  检验对话框

选择菜单栏中的【Analyze（分析）】→【Compare Means（比较均值）】→【Paired-Samples T Test（配对样本T检验）】命令，弹出【Paired-Samples T Test（配对样本T检验）】对话框。



## Step02: 选择配对变量

CONCEPT  
STRATE

在【Paired-Sample T Test（配对样本T检验）】对话框左侧的候选变量列表框中选择一对或几对变量，将其移入【Paired Variables（成对变量）】列表框中，这表示系统将对移入的成对变量进行配对检验。



## Step03: 其他选项选择

单击【Options】按钮，弹出【Paired-Samples T Test: Options (配对样本T检验: 选择)】对话框。该对话框用于指定输出内容和关于缺失值的处理方法，其中各选项的含义如下。

**Confidence Interval:** 用于设置在指定水平下样本均值与指定的检验值之差的置信区间，默认值为95%。

**【Missing Values (缺失值)】选项组:** 用于设置缺失值的处理方式，它有以下两种处理方式。

**Exclude cases analysis by analysis:** 点选该单选钮，表示当分析计算涉及到含有缺失值的变量时，删除该变量上是缺失值的观测量。

**Exclude cases listwise:** 点选该单选钮，表示删除所有含缺失值的观测量后再进行分析。

## Step04 相关统计量的Bootstrap估计

CONCEPT  
STRATE

单击【Bootstrap】按钮，在弹出的对话框中可以进行如下统计量的Bootstrap估计。

- 支持均值和标准差的Bootstrap 估计。
- 支持相关性的Bootstrap 估计。
- 检验表支持均值的Bootstrap 估计。

## Step05

CONCEPT  
STRATE

单击图【OK】按钮，结束操作，SPSS软件自动输出结果。

## 4.3.3 实例图文分析：看电视和读书的时间

CONCEPT  
STRATE

### 1. 实例内容

“每月读书俱乐部”的成员进行了一项调查，以确信其成员用于看电视的时间是否比读书的时间多。假定抽取了15个人组成的样本，得到了下列有关他们每周观看电视的小时数和每周读书时间的小时数的数据，见表4-11所示。你能够得到结论：“每月读书俱乐部”的成员每周观看电视的时间比读书的时间更多吗？

表 4-11 每周观看电视和读书时间

被调查者	看电视小时数	读书小时数
1	10	6
2	14	16
3	16	8
4	18	10
5	15	10
6	14	8
7	10	14
8	12	14
9	4	7
10	8	8
11	16	5
12	5	10
13	8	3
14	19	10
15	11	6

## 2. 实例操作



由于读书俱乐部的成员每人在每周可能既要看电视也要读书，因此要分析看电视和读书时间差异性，其实就是进行如下假设检验。

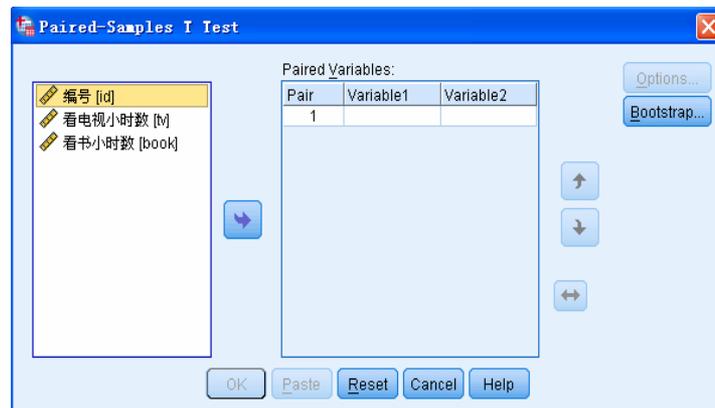
$H_0$  : 俱乐部成员看电视和读书所消耗的时间相同。

$H_1$  : 俱乐部成员看电视和读书所消耗的时间不同。

由于抽样数据中，样本都进行了看电视和读书两个方面的时间调查，它们的活动主体都是同一个人，因此，数据类型属于配对样本的类型，故利用配对样本t检验来分析。具体操作步骤如下。

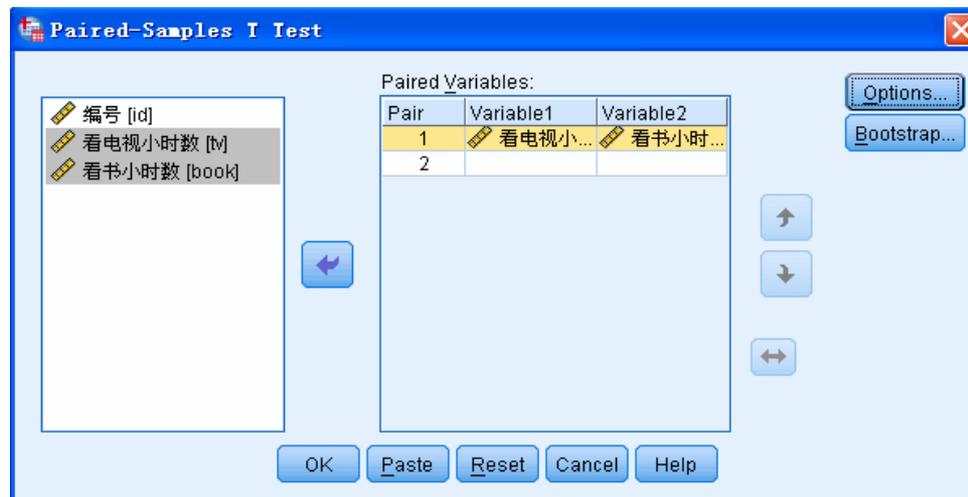
# Step01: 打开对话框

打开数据文件4-5.sav，选择菜单栏中的【Analyze（分析）】→【Compare Means（比较均值）】→【Paired -Sample T Test（配对样本T检验）】命令，弹出【Paired -Sample T Test（配对样本T检验）】对话框。这里变量“tv”表示成员每周看电视的时间；变量“book”表示成员每周读书的时间。



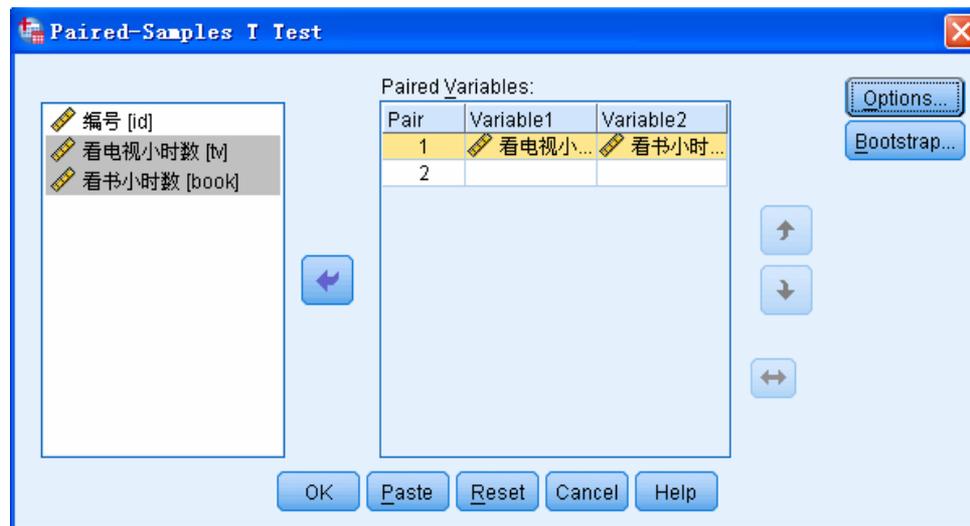
## Step02: 选择配对变量

在左侧的候选变量列表框中依次选择检验变量“tv”和变量“book”，将其添加至【Paired Variable(s)（成对变量）】列表框中。这表示进行“tv”和“book”的配对t检验。



## Step03: 完成操作

单击【OK】按钮，完成操作。此时，软件输出结果出现在结果浏览窗口中。





### 3. 实例结果及分析

#### (1) 基本统计信息汇总表

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	看电视小时数	12.00	15	4.536	1.171
	看书小时数	9.00	15	3.586	.926

## (2) 相关性分析

表4-13是进行两配对变量之间简单相关性分析结果输出表。表中第三列表示样本容量，第四列表示看电视时间和看书时间的简单相关系数，第五列表示概率P值。从结果来看，“tv”和“book”变量的相关系数等于0.193，呈简单正相关关系；同时相伴概率P值0.490大于显著性水平0.05说明这两组样本**相关性显著**。

表 4-13 两配对样本相关性检验结果

		N	Correlation	Sig.
Pair 1	看电视小时数 & 看书小时数	15	.193	.490



### (3) 两配对样本t检验结果表

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std.Deviation	Std.Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 看电视小时数-看书小时数	3	5.21	1.345	.115	5.885	2.23	14	.043

## 4.3.4 实例进阶分析：亚洲金融危机的影响

CONCEPT  
STRATE

### 1. 实例内容

在1997年，亚洲许多国家爆发了大规模的金融危机，致使许多国家的经济发展停滞不前。投资商预言：亚洲经济的低迷对1997年第四季度美国公司的收益造成负面影响。下面的样本数据表4-15显示了部分美国公司在1996年第四季度和1997年第四季度的每股收益（《华尔街日报》，1998年1月28日）。你能根据数据判断投资商的预言吗？

表 4-15 部分美国公司收益

公司	1996 年收益	1997 年收益
Atlantic Richfield	1.16	1.17
Balchem Corp.	0.16	0.13
Black&Decker Corp.	0.97	1.02
Dial Corp.	0.18	0.23
DSC Communications	0.15	-0.32
Eastman Chemical	0.77	0.36
Excel Communications	0.28	-0.14
Federal Signal	0.40	0.29
Ford Motor Company	0.97	1.45
GTE Corp.	0.81	0.73
ITT Industries	0.59	0.60
Kimberly-Clark	0.61	-0.27
Minnesota Mining&Mfr.	0.91	0.89
Procter&Gamble	0.63	0.71

## 2. 实例操作

CONCEPT  
STRATE

表4-15列出了美国公司在亚洲金融危机爆发前后第四季度的每股收益。如果亚洲金融危机对美国公司产生显著影响，那么这两组数据的均值就应该存在显著差异性。由于每组数据是同一公司在1996年和1997年第四季度的收益，因此本案例也属于两配对样本的t检验问题。因此，进行如下假设检验。

$H_0$ ：美国公司在1996年和1997年第四季度的收益没有显著差异，即亚洲金融危机对美国公司收益没有造成影响。

$H_1$ ：美国公司在1996年和1997年第四季度的收益存在显著差异，即亚洲金融危机对美国公司收益造成明显影响。

具体操作步骤如下。

# Step01

CONCEPT  
STRATE

打开数据文件4-6. sav。这里变量“x”表示1996年美国公司的收益；变量“y”表示1997年美国公司的收益。

## Step02

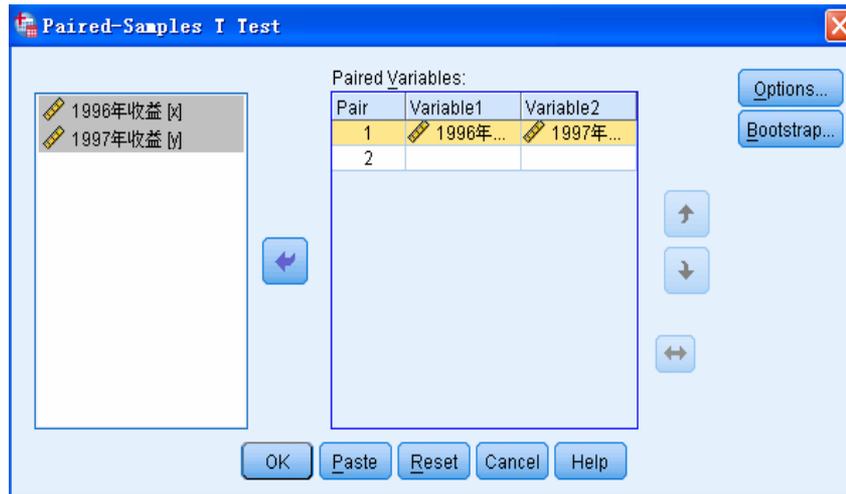
CONCEPT  
STRATE

选择菜单栏中的【Analyze（分析）】→【Compare Means（比较均值）】→【Paired -Sample T Test（配对样本T检验）】命令，弹出【Paired -Sample T Test（配对样本T检验）】对话框。

# Step03

CONCEPT  
TRATE

在左侧的候选变量列表框中依次选择检验变量“x”和变量“y”，将其添加至【Paired Variable(s)（成对变量）】列表框中，进行“x”和“y”变量的配对t检验。



## Step04

CONCEPT  
STRATE

单击【Paired -Sample T Test (配对样本T检验)】对话框中的【OK】按钮，结束操作



### 3. 实例结果及分析

#### (1) 基本统计信息汇总表

		Mean	N	Std. Deviation	Std. Error Mean
Pair 1	1996 年收益	.6136	14	.33626	.08987
	1997 年收益	.4893	14	.54262	.14502

## (2) 相关性分析

表4-17是1996年收益和1997年收益的简单相关性分析结果输出表。从结果来看，“x”和“y”变量的相关系数等于0.825，呈高度正相关关系；同时相伴概率P值0.000进一步说明这两组样本相关性显著。



表 4-17 两配对样本相关性检验结果

		N	Correlation	Sig.
Pair 1	1996年收益 & 1997年收益	14	.825	.000



### (3) 两配对样本t检验结果表

表 4-18 两配对样本 t 检验结果表

	Paired Differences					t	df	Sig. (2-tailed)
	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval of the Difference				
				Lower	Upper			
Pair 1 1996 年收益-1997 年收益	.12429	.32644	.08725	-.0642	.31277	1.425	13	.178



# 第5章 SPSS 的方差分析

# 5.1 方差分述析概



## 5.1.1 方差分析的概念

在第4章中我们讨论了如何对一个总体及两个总体的均值进行检验，如我们要确定两种销售方式的效果是否相同，可以对零假设进行检验。但有时销售方式有很多种，这就是多个总体均值是否相等的假设检验问题了，所采用的方法是方差分析。

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



表5-1 某公司产品销售方式所对应的销售量

序号 销售方式	1	2	3	4	5	水平 均值
方式一	77	86	81	88	83	83
方式二	95	92	78	96	89	90
方式三	71	76	68	81	74	74
方式四	80	84	79	70	82	79
总均值						81.5

方差分析中有以下几个重要概念。

(1) 因素 (Factor) :是指所要研究的变量,它可能对因变量产生影响。如果方差分析只针对一个因素进行,称为单因素方差分析。如果同时针对多个因素进行,称为多因素方差分析。

(2) 水平 (Level) :水平指因素的具体表现,如销售的四种方式就是因素的不同取值等级。

(3) 单元 (Cell) :指因素水平之间的组合。

(4) 元素 (Element) :指用于测量因变量的最小单位。一个单元里可以只有一个元素,也可以有多个元素。

(5) 交互作用 (Interaction) :如果一个因素的效应大小在另一个因素不同水平下明显不同,则称两因素间存在交互作用。

## 5.1.2 方差分析的基本思想

CONCEPT  
STRATE

在表5-1中，要研究不同推销方式的效果，其实就归结为一个检验问题，设为第 $i$  ( $i=1, 2, 3, 4$ )种推销方式的平均销售量，即检验原假设是否成立。从数值上观察，四个均值都不相等，方式二的销售量明显较大。

从表5-1可以看到，20个数据各不相同，这种差异可能是由以下两方面的原因引起的。

一是推销方式的影响，不同的方式会使人们产生不同消费冲动和购买欲望，从而产生不同的购买行动。这种由不同水平造成的差异，称之为系统性差异。

二是随机因素的影响。同一种推销方式在不同的工作日销量也会不同，因为来商店的人群数量不一，经济收入不一，当班服务员态度不一，这种由随机因素造成的差异，我们称之为**随机性差异**。

两个方面产生的差异用两个方差来计量：

一是变量之间的总体差异，即水平之间的方差。

二是水平内部的方差。前者既包括系统性差异，也包括随机性差异；后者仅包括随机性差异。

## 5.1.3 方差分析的基本假设

CONCEPT  
RATE

(1) 各样本的**独立性**。即各组观察数据，是从相互独立的总体中抽取的。

(2) 要求所有观察值都是从正态总体中抽取，且方差相等。在实际应用中能够严格满足这些假定条件的客观现象是很少的，在社会经济现象中更是如此。但一般应近似地符合上述要求。

水平之间的方差（也称为**组间方差**）与水平内部的方差（也称**组内方差**）之间的比值是一个服从F分布的统计量

$$F = \text{水平间方差} / \text{水平内方差} = \text{组间方差} / \text{组内方差}$$

## 5.2 SPSS在单因素方差分析中的应用

单因素方差分析也叫**一维方差分析**，它用来研究一个因素的不同水平是否对观测变量产生了显著影响，即检验由单一因素影响的一个（或几个相互独立的）因变量由因素各水平分组的均值之间的差异是否具有统计意义。

### 1. 使用条件

应用方差分析时，数据应当满足以下几个条件：

- ◆ 在各个水平之下观察对象是独立随机抽样，即独立性；
- ◆ 各个水平的因变量**服从正态分布**，即正态性；
- ◆ 各个水平下的总体具有**相同的方差**，即方差齐；

## 2. 基本原理

方差分析认为：

$SST$ （总的离差平方和）=  $SSA$ （组间离差平方和）+  $SSE$ （组内离差平方和）

如果在总的离差平方和中，组间离差平方和所占比例较大，说明观测变量的变动主要是由因素的不同水平引起的，可以主要由因素的变动来解释，系统性差异给观测变量带来了显著影响；反之，如果组间离差平方和所占比例很小，说明观测变量的变动主要由随机变量因素引起的。

SPSS将自动计算检验统计量和相伴概率P值，若P值小于等于显著性水平 $\alpha$ ，则拒绝原假设，认为因素的不同水平对观测变量产生显著影响；反之，接受零假设，认为因素的不同水平没有对观测变量产生显著影响。

### 3. 多重比较检验问题

多重比较是通过对总体均值之间的配对比较来进一步检验到底哪些均值之间存在差异。

#### 4. 各组均值的精细比较

多重比较检验只能分析两两均值之间的差异性，但是有些时候需要比较多个均值之间的差异性。具体操作是将其转化为研究这两组总的均值是否存在显著差异，即与是否有显著差异。这种比较是对各均值的某一线性组合结构进行判断，即上述检验可以等价改写为对进行统计推断。这种事先指定均值的线性组合，再对该线性组合进行检验的分析方法就是各组均值的精细比较。显然，可以根据实际问题，提出若干种检验问题。

## 5.2.2 单因素方差分析的SPSS操作详解

CONCEPT  
STRATE

Step01: 打开主操作窗口

选择菜单栏中的【Analyze (分析)】→【Compare Means (比较均值)】→【One-Way ANOVA (单因素ANOVA)】命令，弹出【One-Way ANOVA (单因素ANOVA)】对话框，这是单因素方差分析的主操作窗口。

Step02: 选择因变量

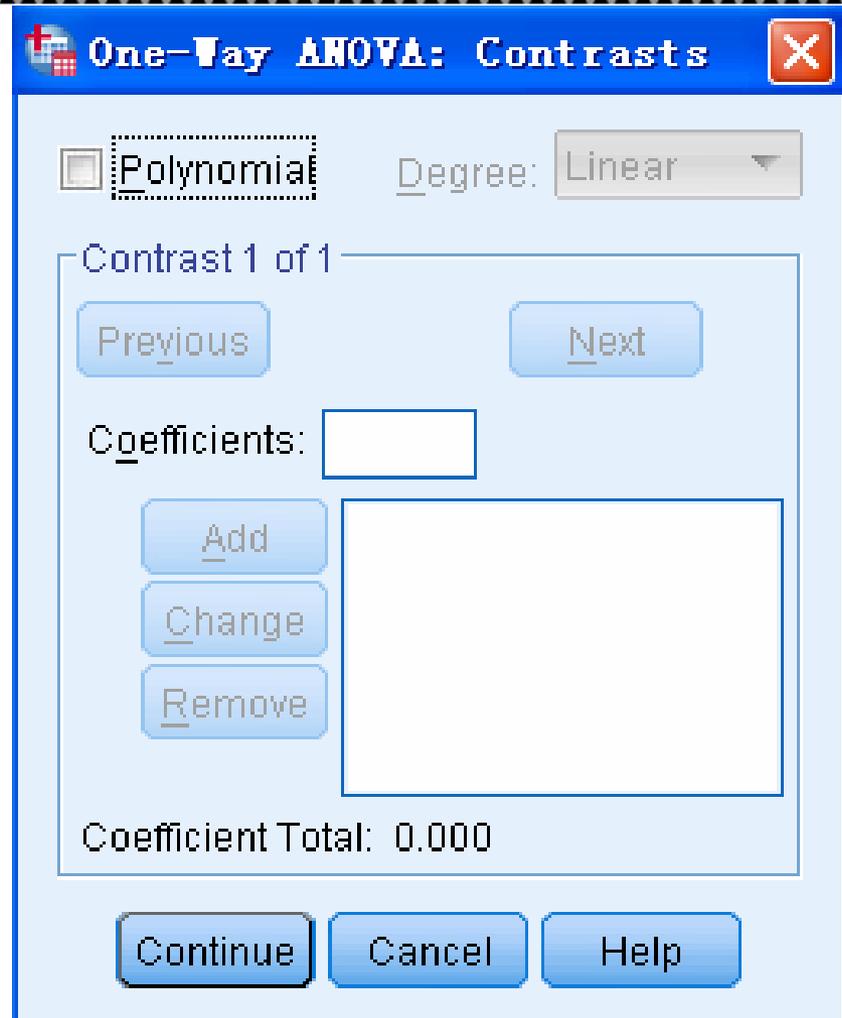
在【One-Way ANOVA (单因素ANOVA)】对话框的候选变量列表框中选择一个或几个变量，将其添加至【Dependent List (因变量列表)】列表框中，选择的变量就是要进行方差分析的观测变量（因变量）。

### Step03: 选择因素变量

在【One-Way ANOVA (单因素ANOVA)】对话框的候选变量列表框中选择一个变量，将其添加至【Factor (因子)】列表框中，选择的变量就是要进行方差分析的因素变量。

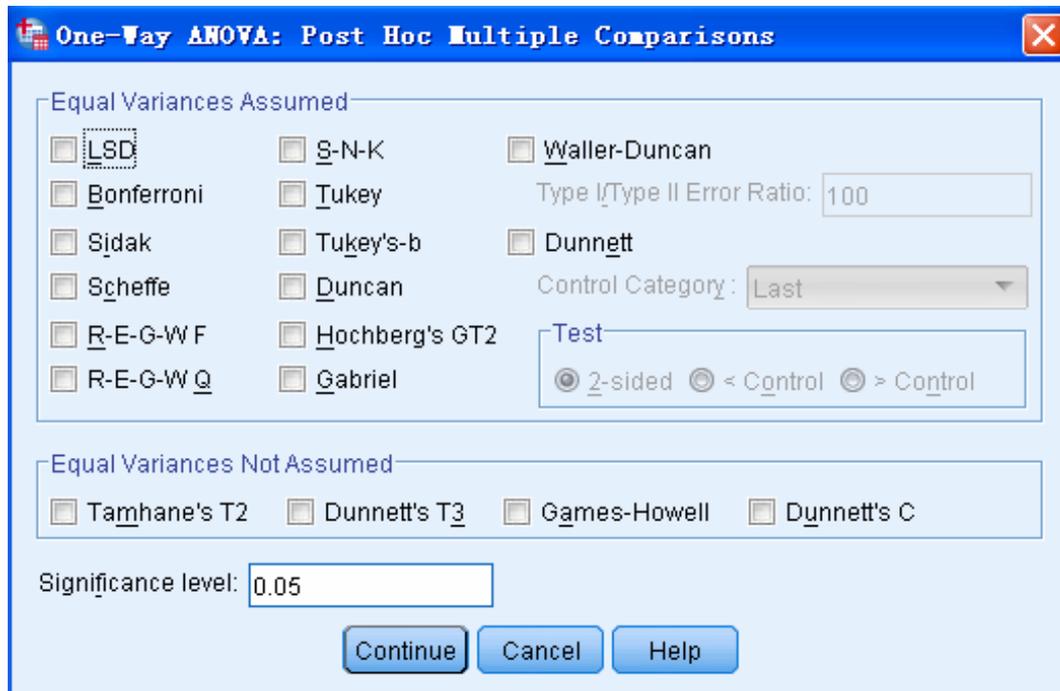
### Step04: 均值精细比较

单击【Contrasts】按钮，弹出如右图所示的【Contrasts (对比)】对话框。



## Step05: 均值多重比较

单击【Post Hoc】按钮，弹出如下图所示的【Post Hoc Multiple Comparisons (两两比较)】对话框，该对话框用于设置均值的多重比较检验。



(1) 方差齐性 (Equal Variances Assumed) 时, 有如下方法供选择。

- **LSD (Least-significant difference)**: 最小显著差数法, 用t检验完成各组均值间的配对比较。
- **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**: 计算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**: 用F分布对所有可能的组合进行同时进行的配对比较。此法可用于检查组均值的所有线性组合, 但不是公正的配对比较。
- **R-E-G-W F**: 基于F检验的Ryan-Einot-Gabriel-Welsch多重比较检验。

- R-E-G-W Q: 基于Student Range分布的Ryan-Einot-Gabriel-Welsch range test多重配对比较。
- S-N-K: 用Student Range分布进行所有各组均值间的配对比较。
- Tukey: 用Student-Range统计量进行所有组间均值的配对比较, 用所有配对比较误差率作为实验误差率。
- Tukey's-b: 用Student Range分布进行组间均值的配对比较, 其精确值为前两种检验相应值的平均值。
- Duncan: 指定一系列的Range值, 逐步进行计算比较得出结论。
- Hochberg's GT2: 用正态最大系数进行多重比较。
- Gabriel: 用正态标准系数进行配对比较, 在单元数较大时, 这种方法较自由。

- **Waller-Dunca:** 用t统计量进行多重比较检验，使用**贝叶斯逼近**的多重比较检验法。
- **Dunnett:** 多重配对比较的t检验法，用于一组处理对一个控制类均值的比较。默认的控制类是最后一组。
  - (2) 方差不具有齐性 (**Equal Varance not assumed**) 时，有如下方法供选择。
    - **Tamhane's T2:** 基于t检验进行配对比较。
    - **Dunnett's T3:** 基于Student**最大模**的成对比较法。
    - **Games-Howell:** Games-Howell比较，该方法较灵活。
    - **Dunnett's C:** 基于Student**极值**的成对比较法。
  - (3) **Significance:** 确定各种检验的显著性水平，系统默认值为**0.05**，可由用户重新设定。

## Step06: 其他选项输出

单击【Options】按钮，在弹出的对话框中进行如下设置。

(1) 【Statistics(统计量)】复选框:选择输出统计量。

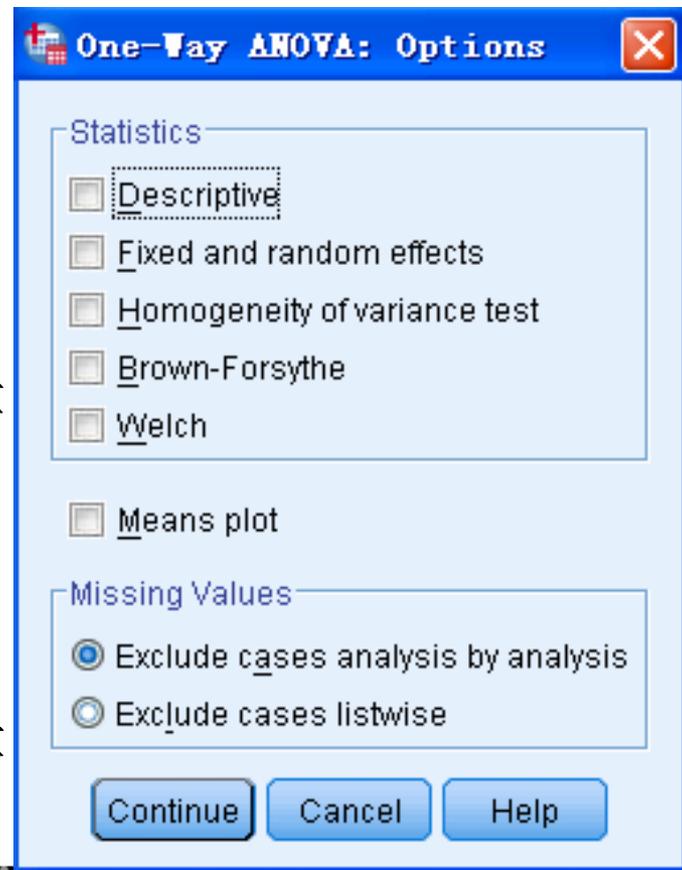
● Descriptive: 要求输出描述统计量。选择此项输出观测值容量、均值、标准差、标准误、最小值、最大值、各组中每个因变量的95%置信区间。

● Fixed and random effects: 显示固定和随机描述统计量。

● Homogeneity-of-variance: 计算Levene统计量进行方差齐性检验。

● Brown-Forsythe: 计算检验组均值相等假设的**布朗检验**。在方差齐性假设不成立时，这个统计量比F统计量更优越。

● Welch: 计算检验组均值相等假设的Welch统计量，在**不具备方差齐性假设**时，也是一个比F统计量更优越的统计量。



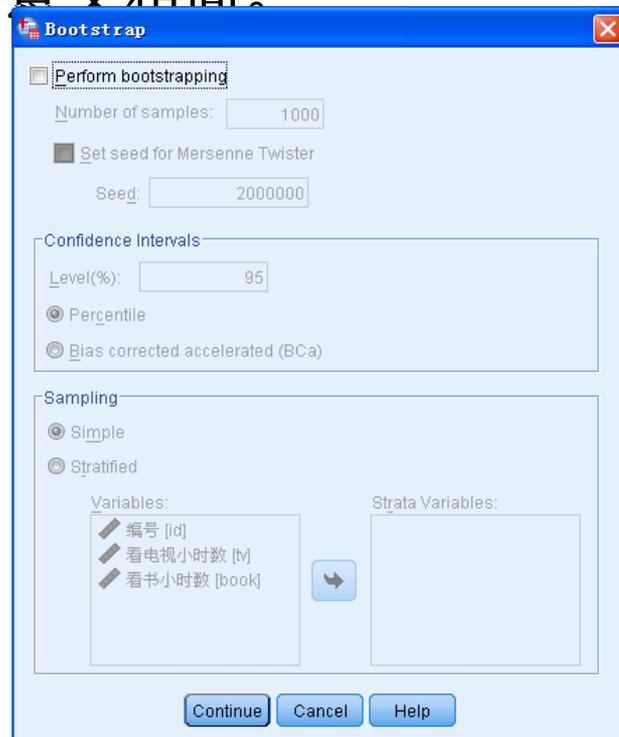
(2) Means plot: 均值折线图。根据各组均值变化描绘出因变量的分布情况。

(3) 【Missing Values (缺失值)】选项组中提供了缺失值处理方法, 该选项和均值比较过程中的缺失值选项意义相同。

Step07: 相关统计量的Bootstrap估计。

单击【Bootstrap】按钮, 弹出如右图所示的对话框。

- 描述统计表支持均值和标准差的bootstrap 估计。
- 多重比较表支持平均值差值的bootstrap 估计。
- 对比检验表支持对比值的bootstrap 估计和显著性检验。



## 5.2.3 实例图文分析：信息来源与传播

CONCEPT  
STRATE

### 1. 实例内容

某机构的各个级别的管理人员需要足够的信息来完成各自的任務。最近，一项研究调查了信息来源对信息传播的影响。在这项特定的研究中，信息来源是上级、同级和下级。在每种情况下，对信息传播进行测度：数值越高，说明信息传播越广。检验信息来源是否对信息传播有显著影响？你的结论是什么？

### 2. 实例操作

由于不同的信息来源可能导致信息传播测度不同。本案例中，信息来源是因素，“上级、同级和下级”是因素的三种不同水平，信息传播测度是因变量（观测变量）。由于这里有三个水平，因此不能采用两样本的均值检验过程，故考虑采用单因素方差分析法。

进行如下假设检验：

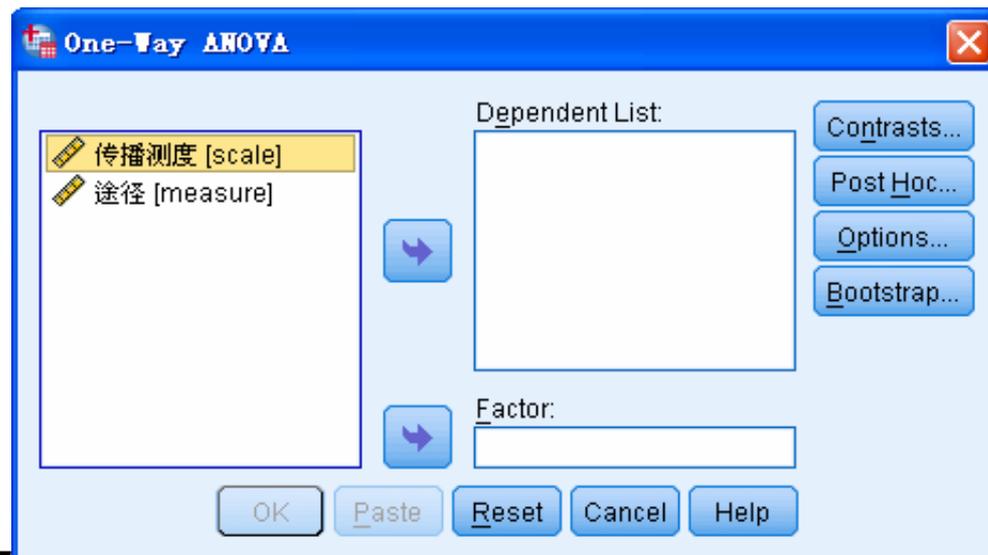
$H_0$ ：三种不同信息来源对信息传播测度平均值没有显著性影响；

$H_1$ ：三种不同信息来源对信息传播测度平均值存在显著性影响。

## Step01: 打开对话框

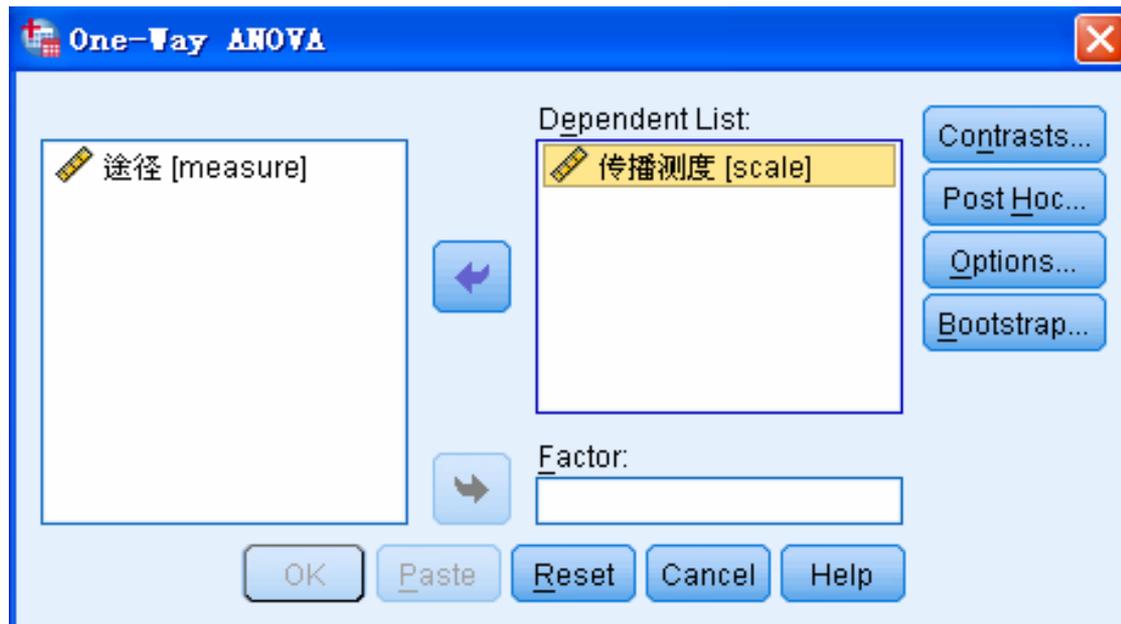
打开数据文件5-1.sav, 选择菜单栏中的【Analyze (分析)】→【Compare Means (比较均值)】→【One-Way ANOVA (单因素ANOVA)】命令, 弹出【One-Way ANOVA (单因素ANOVA)】对话框。

提示: 在使用前, 请注意数据是否符合方差分析的前提条件。



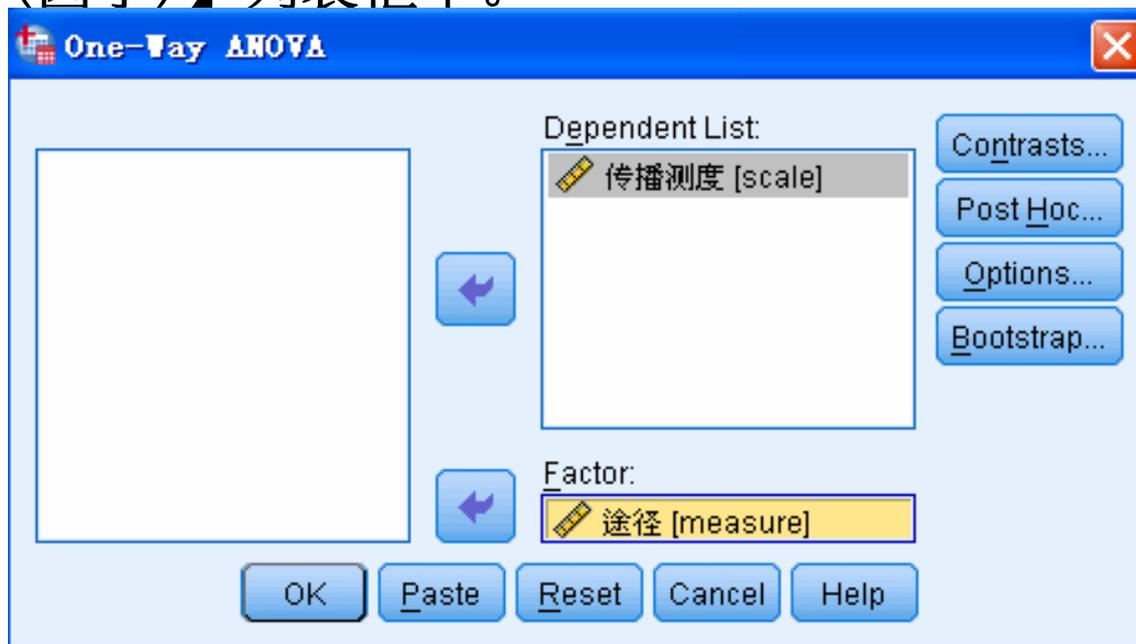
## Step02: 选择因变量

在候选变量列表框中选择“scale”变量作为因变量，将其添加至【Dependent List (因变量列表)】列表框中。



### Step03: 选择因素变量

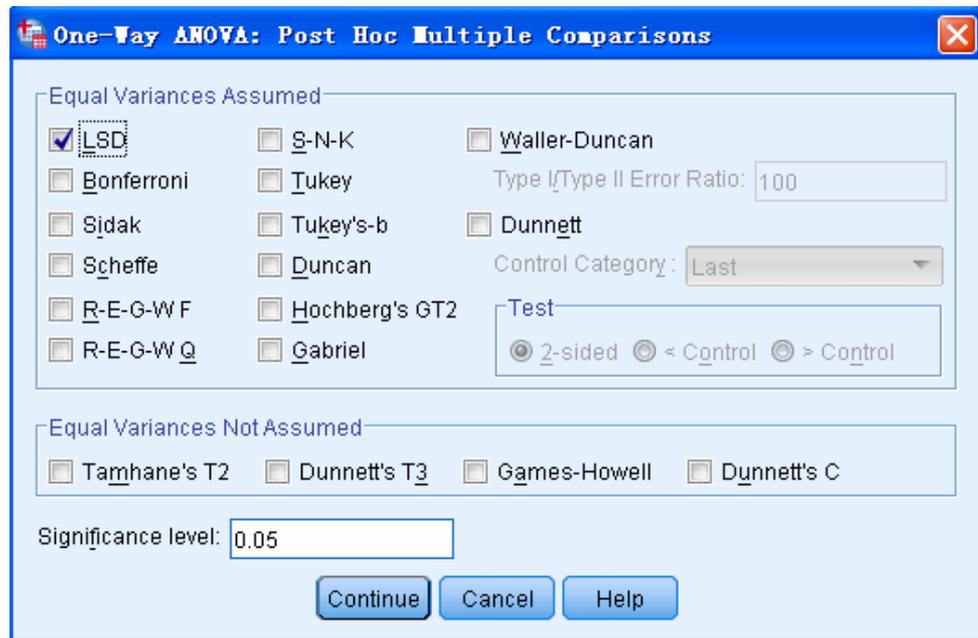
在候选变量列表框中选择“source”变量作为水平值，将其添加至【Factor(因子)】列表框中。



## Step04: 选择均值多重比较方法

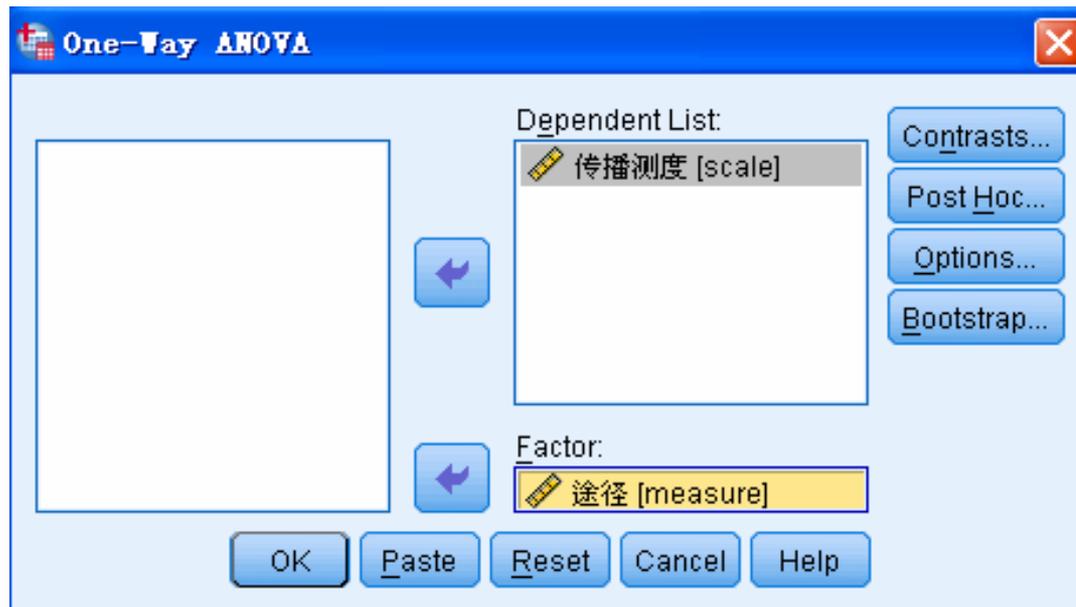
单击【Options】按钮，在弹出的对话框中勾选【Homogeneity of variance】复选框，表示输出方差齐性检验表。再单击【Continue】按钮返回主对话框。

提示：根据数据特点及您的实验要求，选择不同的均值多重比较方法。



## Step05: 完成操作

最后，单击【OK(确定)】按钮，操作完成。





### 3. 实例结果及分析

#### (1) 方差齐性检验

SPSS的结果报告中首先列出了方差分析检验结果。由于这里采用的是Levene检验法，故表格首先显示Levene统计量等于**0.055**。由于概率P值**0.946**明显大于显著性水平，故认为这三组数据的方差是相同的，满足方差分析的前提条件。

#### (2) 单因素方差分析表

表 5-4 方差分析检验表

传播测度	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.000	2	.500	.269	.767
Within Groups	39.000	21	1.857		
Total	40.000	23			

## 5.2.4 实例进阶分析：股票基金的费用比率

CONCEPT  
RATE

### 1. 实例内容

Money杂志报告了股票和债券基金的收益和费用比率。10种中等规模的资本股票基金、10种小额资本股票基金、10种混合型股票基金和10种专项股票基金的费用比率的数据见表5-5所示（单位：%）。

- (1) 请检验这4种类型股票基金之间的平均费用比率的差异性。
- (2) 混合型股票基金的费用比率是其他三种类型基金费用比率的平均水平吗？

表 5-5 股票基金费用比率<sup>①</sup>

中等规模资本 股票基金 <sup>②</sup>	小额资本 <sup>③</sup> 股票基金 <sup>④</sup>	混合型 <sup>⑤</sup> 股票基金 <sup>⑥</sup>	专项 <sup>⑦</sup> 股票基金 <sup>⑧</sup>
1.2 <sup>⑨</sup>	2.0 <sup>⑩</sup>	2.0 <sup>⑪</sup>	1.6 <sup>⑫</sup>
1.1 <sup>⑬</sup>	1.2 <sup>⑭</sup>	2.7 <sup>⑮</sup>	2.7 <sup>⑯</sup>
1.0 <sup>⑰</sup>	1.7 <sup>⑱</sup>	1.8 <sup>⑲</sup>	2.6 <sup>⑳</sup>
1.9 <sup>㉑</sup>	1.4 <sup>㉒</sup>	1.5 <sup>㉓</sup>	2.5 <sup>㉔</sup>
1.3 <sup>㉕</sup>	1.5 <sup>㉖</sup>	2.5 <sup>㉗</sup>	1.9 <sup>㉘</sup>
1.8 <sup>㉙</sup>	2.3 <sup>㉚</sup>	1.0 <sup>㉛</sup>	1.5 <sup>㉜</sup>
1.4 <sup>㉝</sup>	1.9 <sup>㉞</sup>	0.9 <sup>㉟</sup>	1.6 <sup>㊱</sup>
1.7 <sup>㊲</sup>	1.1 <sup>㊳</sup>	1.9 <sup>㊴</sup>	2.7 <sup>㊵</sup>
1.0 <sup>㊶</sup>	1.2 <sup>㊷</sup>	1.4 <sup>㊸</sup>	2.2 <sup>㊹</sup>
2.0 <sup>㊺</sup>	1.3 <sup>㊻</sup>	0.3 <sup>㊼</sup>	0.7 <sup>㊽</sup>

## 2. 实例操作

CONCEPT  
RATE

Step01: 打开或建立数据文件5-2. sav, 选择菜单栏中的【Analyze (分析)】→【Compare Means (比较均值)】→【One-Way ANOVA (单因素ANOVA)】命令, 弹出【One-Way ANOVA (单因素ANOVA)】对话框。这里“rate”变量表示基金的费用比率; “fund”变量表示基金的类型, 其中, “1”表示中等规模的资本股票基金, “2”表示小额资本股票基金, “3”表示混合型股票基金, “4”表示专项股票基金。

Step02: 在【候选变量】列表框中选择“rate”变量作为因变量，将其添加至【Dependent List (因变量列表)】列表框中。

Step03: 在【候选变量】列表框中选择“fund”变量作为水平值，将其添加至【Factor (因子)】列表框中。

Step04: 单击【Contrasts】按钮，弹出【One-Way ANOVA: Contrasts(单因素ANOVA: 对比)】对话框。勾选【Polynomial(多项式)】复选框，激活【Degree(度)】下拉菜单，默认选择【Linear(线性)】选项，表示要进行均值的精细比较。接着在【Coefficients(系数)】文本框中依次输入线性多项式的系数“1”、“1”、“-3”和“1”，并单击【Add(添加)】按钮确认设置。再单击【Continue】按钮，返回主对话框。

Step05: 单击【Post Hoc】按钮，弹出【Post Hoc(两两比较)】对话框。由于这里已计划好对这4组均值进行两两比较，则在其对话框中勾选【LSD】复选框。单击【Continue】按钮，返回主对话框。

Step06: 单击【Options】按钮，在弹出的对话框中勾选【Descriptive(描述性)】复选框表示输出描述性统计量；勾选【Homogeneity-of-variance(方差同质性)】复选框表示输出方差齐性检验表；勾选【Mean plot(均值图)】复选框表示输出各水平的均值折线图。再单击【Continue】按钮，返回主对话框。

Step07: 单击【One-Way ANOVA(单因素ANOVA)】对话框中的【OK】按钮，完成操作。



### 3. 实例结果及分析

(1) 描述性统计表SPSS的结果报告中首先输出了描述性统计量，如表5-6所示。首先，中等规模的资本股票基金的平均费用比率（1.440）最低，而专项股票基金的平均费用比率（2.000）最高，但各类型基金的平均值差距不大。其次，从标准差大小来看，中等规模的资本股票基金（0.3806）最低，而混合型股票基金（0.7379）最高。最后，表5-6还列出了各种类型基金的最大值、最小值及95%水平的置信区间。

表 5-6 描述性统计量

Case	N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
					Lower Bound	Upper Bound		
1	10	1.440	.3806	.1204	1.168	1.712	1.0	2.0
2	10	1.560	.4006	.1267	1.273	1.847	1.1	2.3
3	10	1.600	.7379	.2333	1.072	2.128	.3	2.7
4	10	2.000	.6583	.2082	1.529	2.471	.7	2.7
Total	40	1.650	.5844	.0924	1.463	1.837	.3	2.7

## (2) 方差齐性检验

表5-7是方差齐性检验结果表。表中显示Levene统计量等于2.086。由于概率P值0.119大于显著性水平0.05，故认为这四种类型基金费用比率的方差是相同的，满足方差分析的前提条件。

表 5-7 方差齐性检验结果表

	Levene Statistic	df1	df2	Sig.
基金的费用比率	2.086	3	36	.119

### (3) 单因素方差分析表

表5-7为单因素方差分析表。可以看到，费用比率总的离差平方总和为13.320；不同基金的组间离差为1.772；组内离差为11.548；它们的方差比分别为0.591和0.321，相除得F统计量的观测值为1.841，对应的概率P值为0.157。这里显著性水平为0.05，由于P值大于显著性水平0.05，所以接受零假设，认为不同类型基金的费用比率没有显著性差异。

表 5-7 单因素方差分析表

费用比率	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1.772	3	.591	1.841	.157
Within Groups	11.548	36	.321		
Total	13.320	39			

#### (4) 多重比较检验结果

表5-8显示了两两基金之间费用比率均值比较结果。表中的星号表示在显著性水平0.05的条件下，相应的两组均值存在显著性差异。表中第四列Mean Difference表示两两不同基金费用比率差值的均值。第六列是进行t检验的概率P值，可以通过比较P值大小来判断两两基金之间的费用比率是否有显著差异。从结果来看，只有第一种和第四种基金费用比率的概率P值（0.033）小于显著性水平。因此这四种基金中，只有它们之间的费用比率存在显著性差异，其他基金的费用比率之间都没有显著差异。

表 5-8 多重比较检验结果

	(I) 基金类型	(J) 基金类型	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
						Lower Bound	Upper Bound
LSD	1	2	-.1200	.2533	.639	-.634	.394
		3	-.1600	.2533	.532	-.674	.354
		4	-.5600*	.2533	.033	-1.074	-.046
	2	1	.1200	.2533	.639	-.394	.634
		3	-.0400	.2533	.875	-.554	.474
		4	-.4400	.2533	.091	-.954	.074
	3	1	.1600	.2533	.532	-.354	.674
		2	.0400	.2533	.875	-.474	.554
		4	-.4000	.2533	.123	-.914	.114
	4	1	.5600*	.2533	.033	.046	1.074
		2	.4400	.2533	.091	-.074	.954
		3	.4000	.2533	.123	-.114	.914

注：表中带“\*”的表示，均值差在 0.05 的显著性水平上有显著差异。

## (5) 方差分析的精细比较

案例中第二问要比较第三类基金的费用比率和其他基金之间的关系，其实就是要进行**均值之间的多项式比较**。表5-9首先列出了均值线性组合的系数，其实就是软件操作中第四步输入的数值。接着表5-10列出了多项式比较结果。SPSS分别给出了方差齐性和方差不齐性的检验统计量和概率P值。本案例中不管方差齐性还是不齐性，其概率P值都显著大于0.05，这说明了零假设成立，即混合型股票基金的费用比率是其他三种类型基金费用比率的平均水平。

表 5-9 多项式系数结果

Contrast	基金类型			
	1	2	3	4
1	1	1	-3	1



表 5-10 精细比较检验结果

		Contrast	Value of Contrast	Std. Error	t	df	Sig. (2-tailed)
费用 比率	Assume equal variances	1	.200	.6204	.322	36	.749
	Does not assume equal variances	1	.200	.7509	.266	11.803	.795

## (6) 均值折线图

图5-11显示了这四类基金费用比率的均值折线图。从图中明显看到，第四类基金的费用比率均值明显高于其他类型的基金。

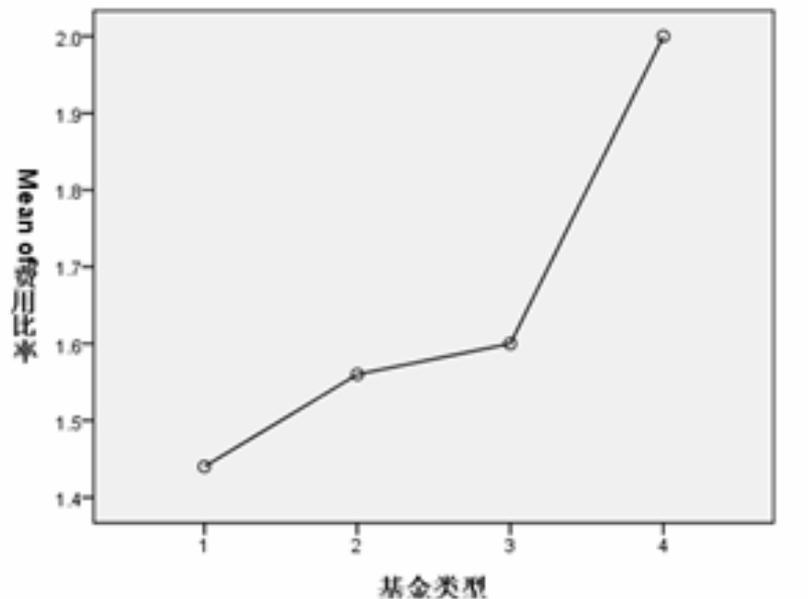


图5-11 均值折线图

## 5.3 SPSS在多因素方差分析中的应用

CONCEPT  
STRATE

### 5.3.1 多因素方差分析的基本原理

#### 1. 方法概述

多因素方差分析是对一个**独立变量**是否受**一个或多个因素**或变量影响而进行的方差分析。它不仅能够分析多个因素对观测变量的独立影响，更能够分析多个因素的交互作用能否对观测变量产生显著影响。例如，对稻谷产量进行分析时，不仅单纯考虑耕地深度和施肥量都会影响产量，但同时深耕和适当的施肥可能使产量成倍增加，这时，耕地深度和施肥量就可能存在交互作用。

$$Q_{\text{总}} = Q_{\text{控1}} + Q_{\text{控2}} + Q_{\text{控1控2}} + Q_{\text{随}}$$



## 2. 基本原理

由于多因素方差分析中观察变量不仅要受到多个因素独立作用的影响，而且因素其交互作用和一些随机因素都会对变量产生影响。因此观测变量值的波动要受到多个控制变量独立作用、控制变量交互作用及随机因素等三方面的影响。以两个因素为例，可以表示为：

$$Q_{\text{总}} = Q_{\text{控1}} + Q_{\text{控2}} + Q_{\text{控1控2}} + Q_{\text{随}}$$

其中， $Q$ 表示各部分对应的离差平方和。多因素方差分析比较

$Q_{\text{控1}}$ 、 $Q_{\text{控2}}$ 、 $Q_{\text{控1控2}}$ 、 $Q_{\text{随}}$  占  $Q_{\text{总}}$  的比例，以此推断不同因素以及因素之间的交互作用是否给观测变量带来显著影响。

### 3. 软件使用方法

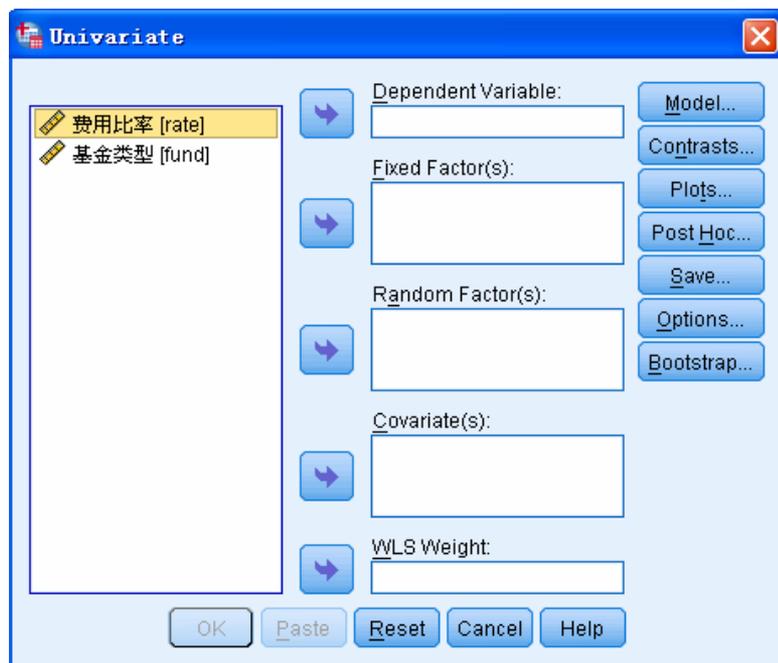
多因素方差分析仍然采用**F检验**，其零假设是 $H_0$ ：各因素不同水平下观测变量的均值无显著差异。SPSS将自动计算F值，并依据F分布表给出相应的概率P值。我们可以根据相伴概率P值和显著性水平 $\alpha$ 的大小关系来判断各因素的不同水平对观测变量是否产生了显著性影响。

## 5.3.2 多因素方差分析的SPSS操作详解

CONCEPT  
RATE

Step01: 打开主对话框

选择菜单栏中的【Analyze（分析）】→【General Linear Model（一般线性模型）】→【Univariate（单变量）】命令，弹出【Univariate（单变量）】对话框，这是多因素方差分析的主操作窗口。



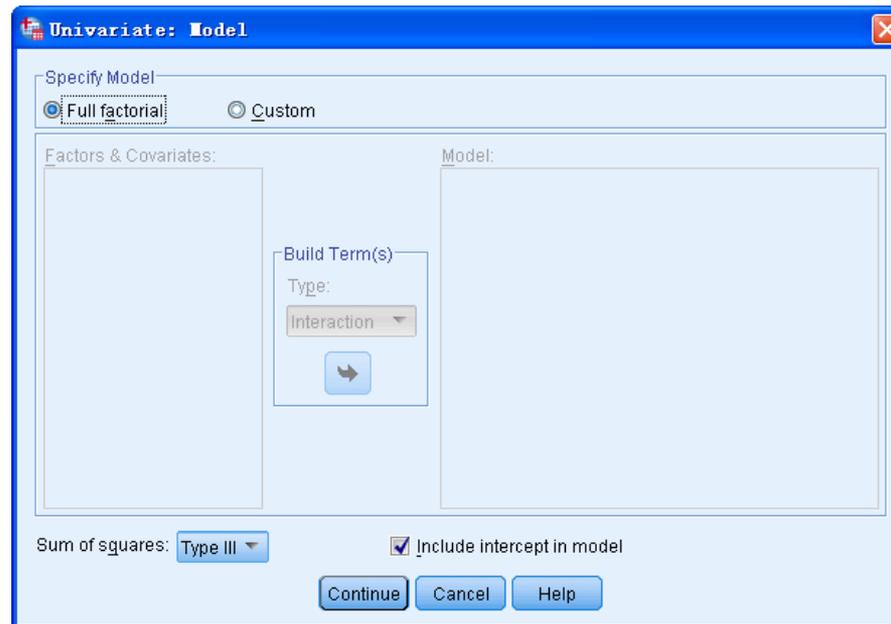
## Step02: 选择分析变量

在【Univariate(单变量)】对话框的候选变量列表框中，选择相应变量进行右侧的列表框中，其目的是设置分析变量。

- 选择**观测变量**（因变量）：添加至【Dependent Variable（因变量）】列表框中。
- 选择**因素变量**：添加至【Fixed Variable(s)（固定因子）】列表框中。
- 选择随机变量：添加至【Random Variable(s)（随机因子）】列表框中。
- 选择协变量：添加至【Covariate(s)（协变量）】列表框中。
- 选择权重变量：添加至【WLS Weight（WLS权重）】列表框中。

### Step03: 模型选择

单击【Model】按钮，弹出【Univariate: Model(单变量: 模型)】对话框，该对话框用于选择分析模型。



### (1) Full Factorial选项

系统默认选项。该项选择建立**全因素模型**，包括所有因素变量的**主效应**和所有的**交互效应**。例如有三个因素变量，全模型包括三个因素变量的主效应、两两的交互效应和三个因素的交互效应。选择该项后无需进行进一步的操作，即可单击【Continue】按钮返回主对话框。

### (2) Custom选项

建立用户自定义的方差分析模型。点择【Custom(设定)】单选钮后，【Factors & Covariates(因子与协变量)】、【Model(模型)】和【Build Term(s)(构建项)】选项被激活。在【Factors & Covariates(因子与协变量)】列表框中自动列出可以作为因素变量的变量名。

在【Build Term(s) (构建项)】选项组的下拉列表框中，可以选择模型的形式。

- Interaction: 选中此项可以指定任意的交互效应。
- Main effects: 选中此项可以指定主效应。
- All 2-way: 指定所有2维交互效应。
- All 3-way: 指定所有3维交互效应。
- All 4-way: 指定所有4维交互效应。
- All 5-way: 指定所有5维交互效应。
- Type I项: 一般适用于平衡的ANOVA模型。
- Type II项: 一般适用于平衡的ANOVA模型、主因子效应模型、回归模型和嵌套设计。

- Type III项：系统默认的平方和分解法。适用于平衡的ANOVA模型和非平衡的ANOVA模型。凡适用Type I和Type II的模型均可以用该法。
  - Type IV项：一般适用于Type I和Type II方法的模型、有缺失值的平衡或不平衡模型。
- (3) 【Include intercept in model(在模型中包含截距)】复选框：系统默认选项，通常截距包括在模型中。如果能假设数据通过原点，可以不包括截距，即不选择此项。

## Step04: 选择比较方法

单击【Contrasts】按钮，弹出【Univariate: Contrasts (单变量: 对比)】对话框。在【Factors (因子)】列表框中显示出所有在主对话框中选中的因素变量。因素变量名后的括号中是当前的比较方法。在该框中选择想要改变比较方法的因子，即鼠标单击选中的因子。这一操作使【Change Contrast (更改对比)】复选栏中的各项被激活。



展开【Contrast(对比)】参数框的下拉菜单,可得到各类比较方法。

- None: 不进行均数比较。
- Deviation: **偏差比较法**。除被忽略的水平外,比较预测变量或因素变量的每个水平的效应。可以点选【Last(最后一个)】(最后一个水平)或【First(第一个)】(第一个水平)作为忽略的水平。
- Simple: **简单比较法**。除去作为参考的水平外,对预测变量或因素变量的每一水平都与参考水平进行比较。选择【Last(最后一个)】或【First(第一个)】作为参考水平。
- Difference: **差值比较法**。对预测变量或因素每一水平的效应,除**第一水平**以外,都与其前面各水平的平均效应进行比较。与Helmert比较法相反。
- Helmert: **Helmert法**。对预测变量或因素的效应,除**最后一个水平**以外,都与后面的各水平的平均效应相比较。
- Repeated: **重复比较法**。对预测变量或因素的效应,除第一水平以外,对每一水平都与它前面的水平进行比较。
- Polynomial: **多项式比较**。比较线性、二次、三次等效应,常用于估计多项式趋势。

## Step05: 选择轮廓图

单击【Plot】按钮，弹出【Profile Plots(轮廓图)】对话框，在该对话框中设置均值轮廓图。

从【Factors(因子)】列表框中选择一个因素变量移入【Horizontal Axis(水平轴)】列表框（水平轴）定义轮廓图的横坐标。选择另一个因素变量移入【Separate Lines(单图)】列表框定义轮廓图的区分线。如果需要的话再从【Factors(因子)】列表框中选择一个因素变量移入【Separate Plots(多图)】列表框定义轮廓图的区分图

以上选择确定以后，单击【Add】按钮加以确定。需要对加入图清单框的选择结果进行修正，可单击【Change和Remove】按钮。



**Univariate: Profile Plots** [Close]

**F**actors: [Empty list box]

[Right Arrow] **H**orizontal Axis: [Text box]

[Right Arrow] **S**eparate Lines: [Text box]

[Right Arrow] **S**eparate Plots: [Text box]

**P**lots: [Add] [Change] [Remove]

[Empty list box]

[Continue] [Cancel] [Help]

A dialog box titled "Univariate: Profile Plots" with a close button in the top right. It features a "Factors:" list box on the left. To its right are three rows, each with a right-pointing arrow button and a text box: "Horizontal Axis:", "Separate Lines:", and "Separate Plots:". Below these are three buttons: "Add", "Change", and "Remove". At the bottom is a large empty list box for "Plots:". At the very bottom are three buttons: "Continue", "Cancel", and "Help".

## Step06: 选择多重比较

单击【Post Hoc】按钮，弹出【Post Hoc Multiple Comparisons for Observed Means(单变量: 观测均值的两两比较)】对话框。该对话框用于对均值作Post Hoc多重比较检验。从【Factor(s) (因子)】框选择相关变量使被选变量进入【Post Hoc test for(两两比较检验)】框。不难发现，这个对话框与单因素方差分析模型的Post Hoc多重比较检验对话框大致相同，各选项意义也一致。



**Univariate: Post Hoc Multiple Comparisons for Ob...** [X]

Factor(s): Post Hoc Tests for:

[Empty Box] ➔ [Empty Box]

Equal Variances Assumed

<input type="checkbox"/> LSD	<input type="checkbox"/> S-N-K	<input type="checkbox"/> Waller-Duncan
<input type="checkbox"/> Bonferroni	<input type="checkbox"/> Tukey	Type I/Type II Error Ratio: 100
<input type="checkbox"/> Sidak	<input type="checkbox"/> Tukey's-b	<input type="checkbox"/> Dunnett
<input type="checkbox"/> Scheffe	<input type="checkbox"/> Duncan	Control Category: Last
<input type="checkbox"/> R-E-G-W-F	<input type="checkbox"/> Hochberg's GT2	Test
<input type="checkbox"/> R-E-G-W-Q	<input type="checkbox"/> Gabriel	<input checked="" type="radio"/> 2-sided <input type="radio"/> < Control <input type="radio"/> > Control

Equal Variances Not Assumed

<input type="checkbox"/> Tamhane's T2	<input type="checkbox"/> Dunnett's T3	<input type="checkbox"/> Games-Howell	<input type="checkbox"/> Dunnett's C
---------------------------------------	---------------------------------------	---------------------------------------	--------------------------------------

[Continue] [Cancel] [Help]

## Step07: 预测值保存

单击【Save】按钮，弹出【Save(保存)】对话框。通过在对话框中的选择，可以将所计算的预测值、残差和检测值作为新的变量保存在编辑数据文件中。以便于在其他统计分析中使用这些值。

① Predicted Values : 预测值。

● Unstandardized: 非标准化预测值。

● Weighted: 加权预测值。如果在主对话框中选择了WLS变量，选中该复选框，将保存加权非标准化预测值。

● Standard error: 预测值标准误。

② Diagnostics: 诊断值。

● Cook's distance: Cook 距离。

● Leverage values: 非中心化 Leverage 值。

### ③ Residuals: 残差。

- Unstandardized: 非标准化残差值, 即观测值与预测值之差。
- Weighted: 加权非标准化残差。如果在主对话框中选择了WLS变量, 选中该复选框, 将保存加权非标准化残差。
- Standardized: 标准化残差, 又称Pearson残差。
- Studentized: 学生氏残差。
- Deleted: 剔除自变量值与校正预测值之差。

最后可以勾选【Coefficient statistics(系数统计)】复选框, 将参数协方差矩阵保存到一个新文件中。单击【File】按钮, 打开相应的对话框将文件保存。

**Univariate: Save** [X]

**Predicted Values**

- Unstandardized
- Weighted
- Standard error

**Diagnostics**

- Cook's distance
- Leverage values

**Residuals**

- Unstandardized
- Weighted
- Standardized
- Studentized
- Deleted

**Coefficient Statistics**

- Create coefficient statistics
  - Create a new dataset  
Dataset name:
  - Write a new data file

## Step08: 其他选项选择

单击【Options】按钮，弹出【Options(选项)】对话框。各选项含义如下。

① 【Estimated Marginal Means (估计边际均值)】：估测边际均值设置。

在【Factor(s) and Factor Interactions (因子和因子交互)】列表框中列出【Model(模型)】对话框中指定的效应项，在该框中选定因素变量的各种效应项。可以将其移入到【Display Means for(显示均值)】列表框中。

在【Display Means for(显示均值)】列表框中有主效应时，点选激活此框下面的【Compare main effects(比较主效应)】复选框，对主效应的边际均值进行组间的配对比较。

在【Confidence interval adjustment(置信区间调节)】参数框中，可以进行多重组间比较。打开下拉菜单，共有三个选项：LSD(n one)、Bonferroni和Sidak方法。

② 在【Display(输出)】列表框中指定要求输出的统计量。

● Descriptive statistics: 输出描述统计量。

● Estimates of effect size: 效应量的估计。

● Observed power: 功效检验或势检验。

● Parameter estimates: 各因素变量的模型参数估计、标准误、t检验的t值、显著性概率和95%的置信区间。

● Contrast coefficient matrix: 显示对照系数矩阵。

● Homogeneity test: 方差齐次性检验。

● Spread vs. level plot: 绘制观测量均值对标准差和方差的图形。

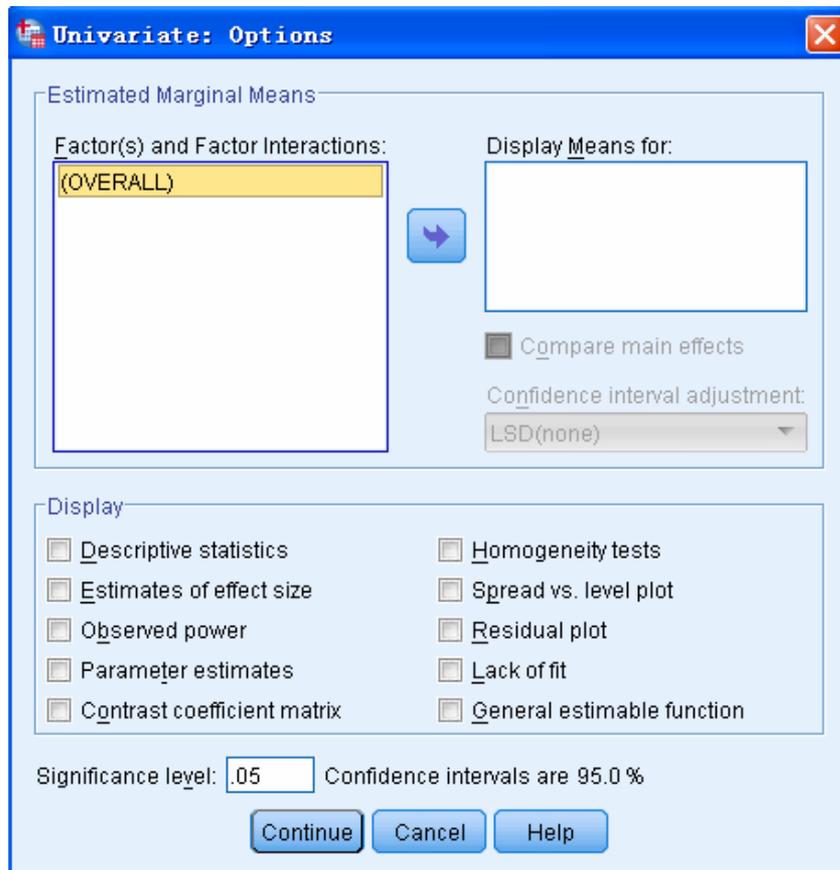
● Residual plot: 绘制因变量的观察值对于预测值和标准化残差的散点图。

● Lack of fit: 拟合度不足检验。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: 可以根据一般估计函数自定义假设检验。

CONCEPT  
STRATE

③ 【Significance level(显著性水平)】 文本框： 改变Confidence intervals(置信区间)内多重比较的显著性水平。



Step09 : 相关统计量的Bootstrap估计。

单击【Bootstrap】按钮，在弹出的对话框中可以进行如下统计量的Bootstrap估计。

- 描述统计表支持均值和标准差的Bootstrap 估计。
- 参数估计值表支持系数、B 的Bootstrap 估计和显著性检验。
- 对比结果表支持差值的Bootstrap 估计和显著性检验。
- 估计值表支持均值的Bootstrap 估计。
- 成对比较表支持平均值差值的Bootstrap 估计。
- 多重比较表支持平均值差值的Bootstrap 估计。

Step10: 单击【OK】按钮，结束操作，SPSS软件自动输出结果。

## 5.3.3 实例图文分析：薪金的区别

CONCEPT  
STRATE

### 1 实例内容

假设某一杂志的记者要考察职业为财务管理、计算机程序和药剂师的男女雇员其每周的薪金之间是否有显著性差异。从每种职业中分别选取了5名男性和5名女性组成样本，并且记录下来样本中每个人的周薪金（单位：美元）。所得数据见表5-11所示。请你分析职业和性别对薪金有无显著影响。



表 5-11 不同职业性别每周薪金

每周薪金	职业	性别	每周薪金	职业	性别
872	财务管理	男	884	计算机程序员	女
859	财务管理	男	765	计算机程序员	女
1028	财务管理	男	685	计算机程序员	女
1117	财务管理	男	700	计算机程序员	女
1019	财务管理	男	671	计算机程序员	女
519	财务管理	女	1105	药剂师	男
702	财务管理	女	1144	药剂师	男
805	财务管理	女	1085	药剂师	男
558	财务管理	女	903	药剂师	男
591	财务管理	女	998	药剂师	男
747	计算机程序员	男	813	药剂师	女
766	计算机程序员	男	985	药剂师	女
901	计算机程序员	男	1006	药剂师	女
690	计算机程序员	男	1034	药剂师	女
881	计算机程序员	男	817	药剂师	女

## 2 实例操作

由于薪金水平的高低和所从事的职业、性别等因素都有关系。因此这里要考虑两个因素水平下的薪金差异问题，即建立双因素的方差分析模型。本案例中，职业和性别是两个影响因素，而每周薪金是因变量。同时，我们也要考虑职业和性别这两个因素之间有无交互作用。具体操作步骤如下。

Step01: 打开对话框

打开数据文件5-3. sav，选择菜单栏中的【Analyze（分析）】→【General Linear Model（一般线性模型）】→【Univariate（单变量）】命令，弹出【Univariate（单变量）】对话框。这里“wage”变量表示每月薪金；“job”变量表示职业的类型；“sex”变量表示性别。

提示：在使用前，请注意数据是否符合方差分析的前提条件。



**Univariate** [Close]

每周薪金 [wage]  
职业 [job]  
性别 [sex]

Dependent Variable: [ ] Model...  
Fixed Factor(s): [ ] Contrasts...  
Random Factor(s): [ ] Plots...  
Covariate(s): [ ] Post Hoc...  
WLS Weight: [ ] Save...  
Options...  
Bootstrap...

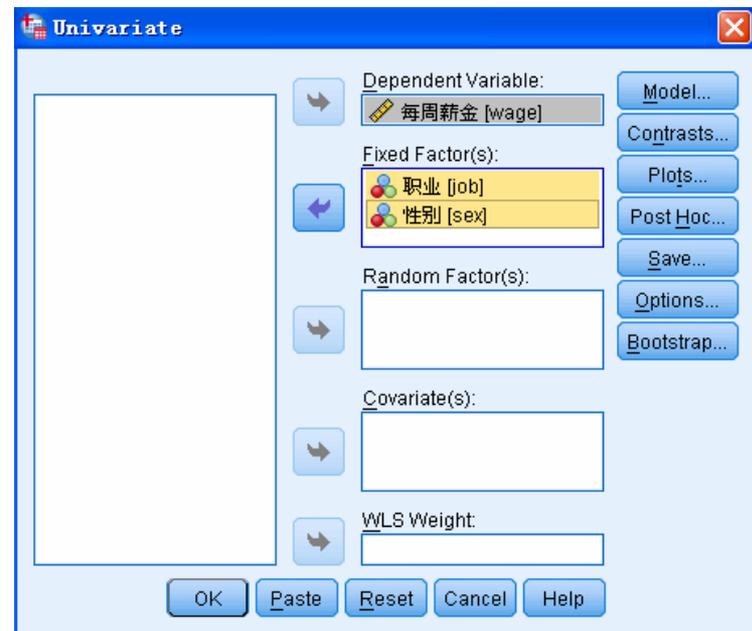
OK Paste Reset Cancel Help

## Step02: 选择观测变量

在候选变量列表框中选择“wage”变量作为因变量，将其添加至【Dependent Variable(因变量)】列表框中。

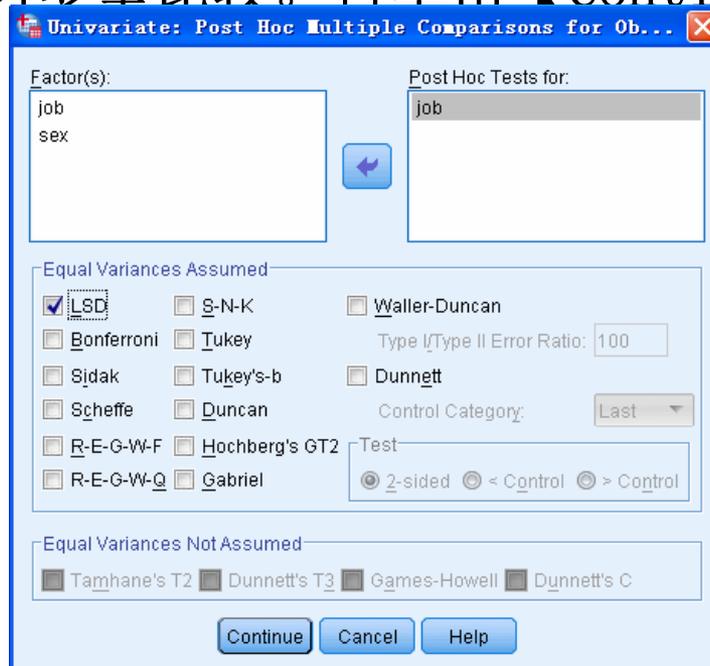
## Step03: 选择因素变量

选择“job”和“sex”变量作为因素变量，将它们添加至【Fixed Factor(s)(固定因子)】列表框中。



## Step04: 选择多重比较

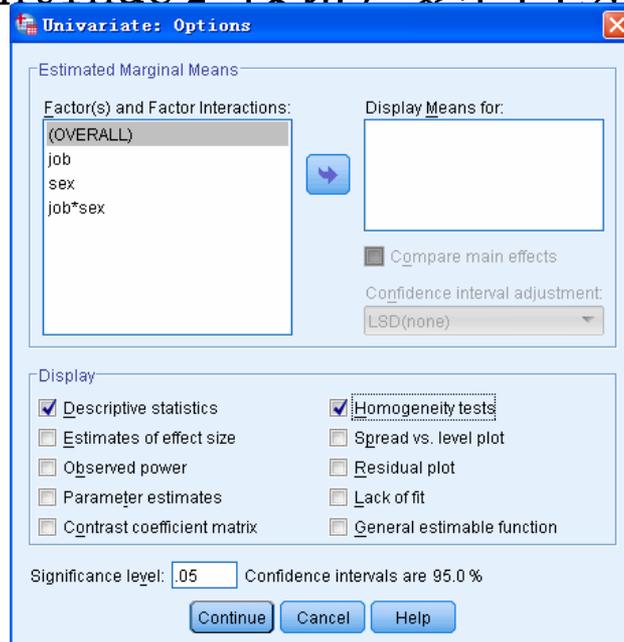
单击【Post Hoc】按钮，弹出【Post Hoc(两两比较)】对话框。在【Factors(因子)】列表框中选择“job”变量至【Post Hoc Test for(两两比较检验)】列表框，并且勾选【LSD】选项。这里表示要进行职业变量的两两多重比较。再单击【Continue】按钮，返回主对话框。



## Step05: 其他选项选择

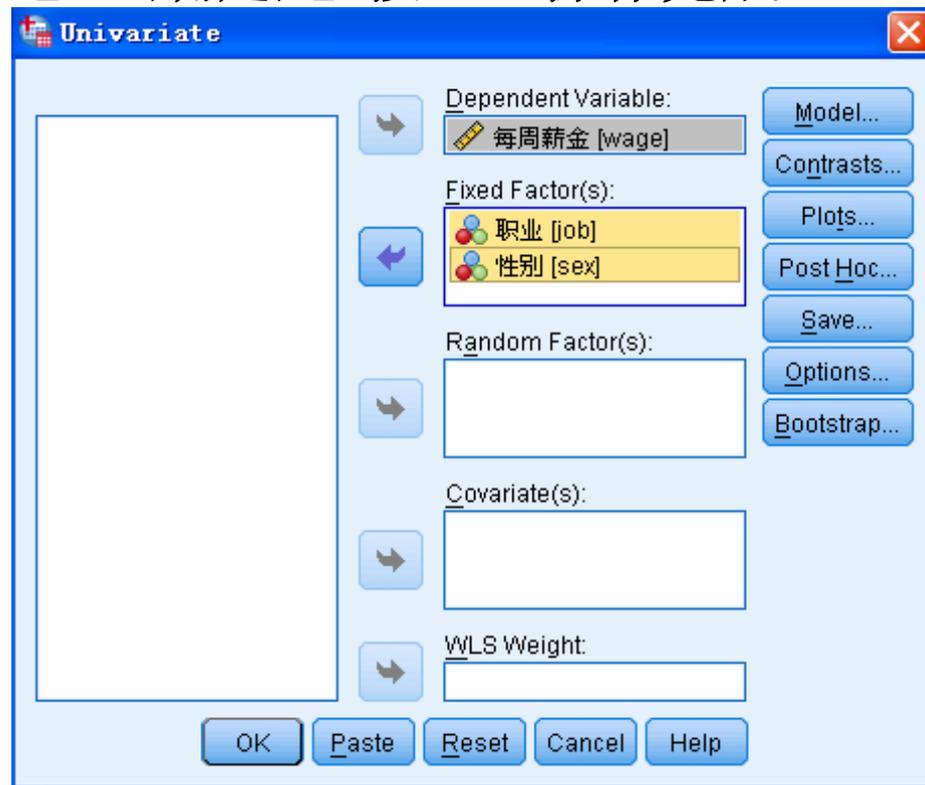
单击【Options】按钮，弹出【Options(选项)】对话框。勾选【Descriptive(描述性统计量)】复选框表示输出描述性统计量；勾选【Homogeneity-of-variance(方差同质性检验)】复选框表示输出方差齐性检验表。再单击【Continue】按钮，返回主对话框。

提示：根据数据特点及您的实验要求，选择不同的均值多重比较方法。



Step06: 完成操作

最后，单击【OK(确定)】按钮，操作完成。



### 3. 实例结果及分析

#### (1) 描述性统计分析表

表5-12和表5-13是对样本数据的基本描述结果。表5-12列出了各种水平下的样本个数。表5-13列出了不同职业、性别每周薪金的样本均值和标准差。从数值大小比较看，不少职业和性别之间每周薪金差异较大，说明有进一步采用方差分析的必要。

表 5-12 描述性统计分析表

		N
职业	1	10
	2	10
	3	10
性别	1	15
	2	15

表 5-13 描述性统计分析表

职业	性别	Mean	Std. Deviation	N
1	1	979.00	110.560	5
	2	635.00	116.951	5
	Total	807.00	210.672	10
2	1	797.00	90.529	5
	2	741.00	87.667	5
	Total	769.00	89.047	10
3	1	1047.00	96.636	5
	2	931.00	107.320	5
	Total	989.00	114.049	10
Total	1	941.00	142.956	15
	2	769.00	159.562	15
	Total	855.00	172.650	30

## (2) 方差齐性检验

SPSS的结果报告接着列出了方差齐性检验结果表5-14。由于这里采用的是Levene检验法，故表格首先显示Levene统计量等于0.383。由于概率P值0.856明显大于显著性水平，故认为样本数据的方差是相同的，满足方差分析的前提条件。

表 5-14 方差齐性检验表

	F	df1	df2	Sig.
每周薪金	.383	5	24	.856

### (3) 双因素方差分析检验表

在表5-15中，第一行的Corrected Model是对所用方差分析模型的检验，其原假设为模型中所有的影响因素均无作用，即职业、性别及两者的交互作用等对每周薪金都无显著影响。该检验的P值远小于0.05，因此所用模型有统计学意义，以上所提到的因素中至少有一个是有显著差异的，但具体是哪些则需要阅读后面的分析结果。

第二行是对模型中常数项是否等于0进行的检验，虽然根据概率P值判断它显著不等于零，但它在分析中没有实际意义，忽略即可。第三、四行分别是对职业、性别的影响效应进行的检验，其零假设分别是：职业或性别对薪金没有显著性差异。但这两行对应的相伴概率P都接近0，显然小于显著性水平0.05。可见，两者分别对薪金有显著性影响。

第五行是对职业和性别的交叉作用进行检验，可见P为0.011，小于显著性水平，表示交互作用对观测变量每周薪金有显著性影响作用。

从上面方差分析结果看到，职业、性别及其两者的交互项都直接影响了每周薪金的高低，存在统计学意义下的显著差异。

表 5-15 双因素方差分析检验表

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	613880.000 <sup>a</sup>	5	122776.000	11.761	.000
Intercept	2.193E7	1	2.193E7	2100.714	.000
job	276560.000	2	138280.000	13.246	.000
sex	221880.000	1	221880.000	21.254	.000
job * sex	115440.000	2	57720.000	5.529	.011
Error	250552.000	24	10439.667		
Total	2.280E7	30			
Corrected Total	864432.000	29			

a.  $R^2 = .710$  (调整  $R^2 = .650$ )

#### (4) 多重比较检验结果

表5-16显示了不同职业之间每周薪金均值比较结果。表中的星号表示在显著性水平0.05的条件下，相应的两组均值存在显著性差异。可以通过比较表中概率P值大小来判断职业之间的薪金水平是否有显著差异。从结果来看，药剂师和其他两个职业的每周薪金存在显著性差异。该职业的平均薪金要明显高于财务管理和计算机程序员职业。

表 5-16 多重比较检验结果

(I) 职业	(J) 职业	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
1	2	38.00	45.694	.414	-56.31	132.31
	3	-182.00*	45.694	.001	-276.31	-87.69
2	1	-38.00	45.694	.414	-132.31	56.31
	3	-220.00*	45.694	.000	-314.31	-125.69
3	1	182.00*	45.694	.001	87.69	276.31
	2	220.00*	45.694	.000	125.69	314.31

注：表中带“\*”的表示，均值差在0.05的显著性水平上有显著差异。

## 5.4 SPSS在协方差分析中的应用

CONCEPT  
STRATE

### 5.4.1 协方差分析的基本原理

#### 1、方法概述

无论是单因素方差分析还是多因素方差分析，它们都有一些人为可以控制的因变量。但在实际问题中，有些随机因素是很难人为控制的，但它们又会对结果产生显著的影响。如果忽略这些因素的影响，则有可能得到不正确的结论。

利用协方差分析就可以完成这样的功能。协方差分析是将那些很难控制的因素作为协变量。在排除协变量影响的条件下，分析因素变量对观察变量的影响，从而更加准确地对因素变量进行评价。这种方法要求协变量应是连续数值型变量，多个协变量间互相独立，且与因素变量之间也没有交互影响。

## 2、基本原理

在协方差分析中，将观察变量总的离差平方和分解为由因变量引起的、由因变量的交互作用引起的、由协变量引起的和由其他随机因素引起的。以双因素协方差分析为例，观察变量总的离差平方和可以分解为：

$$Q_{\text{总}} = Q_{\text{协}} + Q_{\text{控1}} + Q_{\text{控2}} + Q_{\text{控1控2}} + Q_{\text{随}}$$

也可以理解为：  
为：

$$Q_{\text{总}} - Q_{\text{协}} = Q_{\text{控1}} + Q_{\text{控2}} + Q_{\text{控1控2}} + Q_{\text{随}}$$

。即在扣除了协变量对观察变量的影响后，分析因变量对观察变量的影响。协方差分析也采用F检验法，处理计算思路和多因素方差分析相似。

## 5.4.2 协方差分析的SPSS操作详解

CONCEPT  
STRATE

### 1、确定是否存在协变量

采用协方差分析时，首先就应该明确是否存在某些因素对因变量造成影响，特别是一些难以人为控制的因素，例如年龄、身高和体重等等，它们的不同水平可能对因变量产生较为显著的影响。此时可以绘制图形，观察协变量和因变量之间有无关联性。若从图形可以判断两者有显著关系，则可以引入协方差分析。但这也是一种辅助判断方法，只有通过协方差检验结果才能更清晰说明这种协变量的存在性。

### 2、“Univariate”过程中引入协变量

由于协方差分析也是采用【General Linear Model(一般线性模型)】中的【Univariate(单变量)】命令，因此它的基本操作和多因素方差分析的SPSS操作是相同的，这里就不再重复了。只是特别的，需要将确定好的协变量引入到图5-12的【Covariate(s)(对比)】列表框即可。而【Univariate(单变量)】对话框中的各类辅助选项的用法也和多因素方差分析相同。

# 5.4.3 实例图文分析：人体的血清胆固醇



## 1 实例内容

某医生欲了解成年人体重正常者与超重者的血清胆固醇是否不同。而胆固醇含量可能与年龄有关系，具体资料数据见表5-17所示。请建立模型分析体重对人体胆固醇含量的影响，同时也要兼顾年龄的因素

表 5-17 血清胆固醇含量

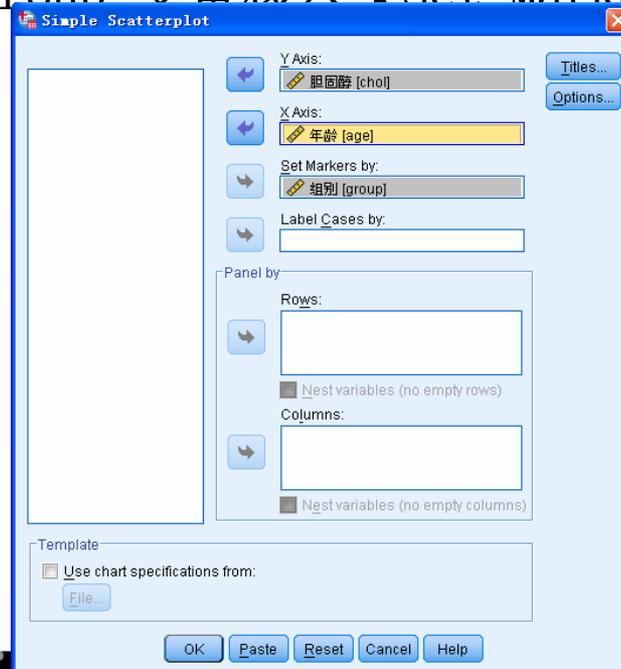
正常组		超重组	
年龄	胆固醇	年龄	胆固醇
48	3.5	58	7.3
33	4.6	41	4.7
51	5.8	71	8.4
43	5.8	76	8.8
44	4.9	49	5.1
63	8.7	33	4.9
49	3.6	54	6.7
42	5.5	65	6.4
40	4.9	39	6.0
47	5.1	52	7.5
41	4.1	45	6.4
41	4.6	58	6.8
56	5.1	67	9.2

## 2 实例操作

案例中需要分析体重对人体的血清胆固醇有无直接影响，同时体重这个因素分为正常组和超重组两个水平，因此可以考虑单因素方差分析模型。但如果仅分析体重的影响作用，而不考虑实验对象年龄的差异，那么得出的结论可能是不准确的。这是因为年龄的大小在一定程度上会影响人体的血清胆固醇含量的高低。因此，为了更准确描述体重对人体的血清胆固醇的影响，就应该尽量排除年龄因素对分析结果的影响。所以将年龄作为协变量引入模型，考虑建立协方差分析模型。在打开或建立数据文件5-4. sav后，具体操作步骤如下。

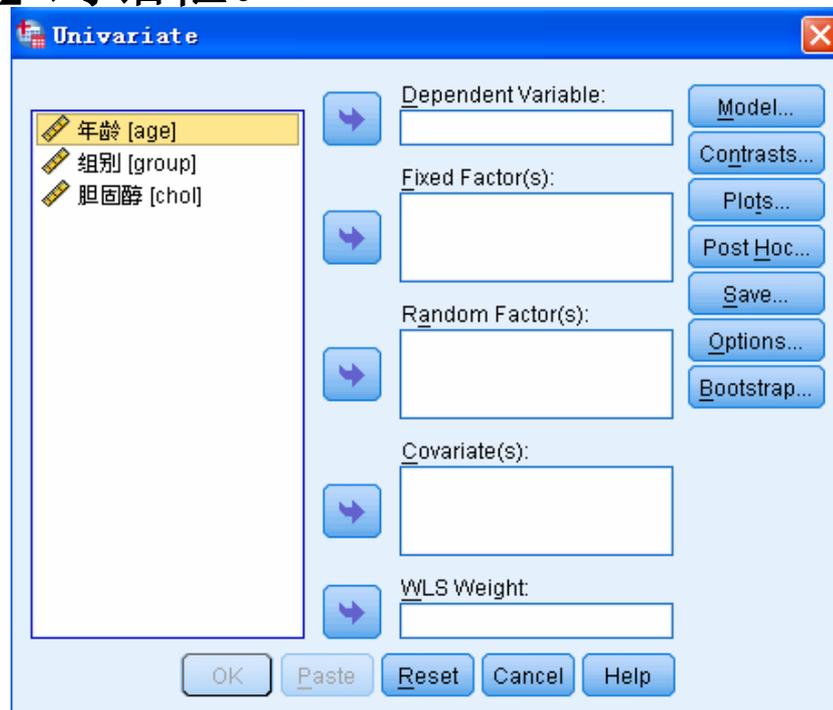
## Step01: 选择观测变量

选择菜单栏中的【Graphs(图形)】→【Legacy Dialogs(旧对话框)】→【Scatter/Dot(散点图/点图)】→【Simple/ Scatter(简单分布)】命令，弹出【Simple Scatterplot(简单分布图)】对话框。在候选变量列表框中选择“chol”变量移入【Y Axis(Y轴)】列表框中，选择“age”变量移入【X Axis(X轴)】列表框中，选择“group”变量移入【Set Markers by(设置标签)】列表框中。



## Step02: 打开对话框

选择菜单栏中的【Analyze（分析）】→【General Linear Model（一般线性模型）】→【Univariate（单变量）】命令，弹出【Univariate（单变量）】对话框。

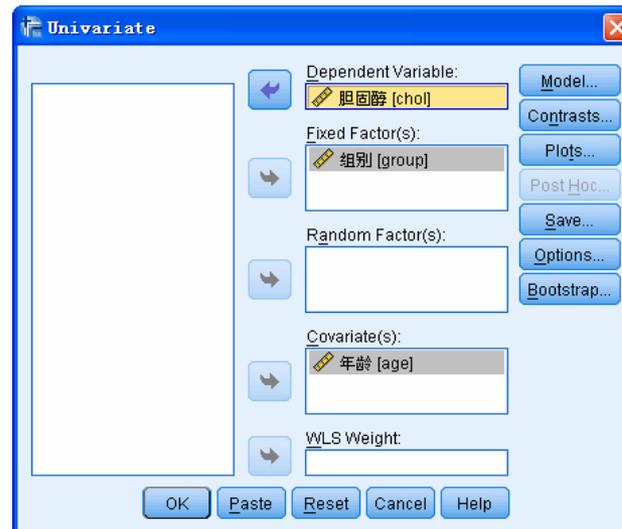


### Step03: 选择分析比较

在候选变量列表框中选择“chol”变量作为因变量，将其添加至【Dependent Variable(因变量)】列表框中。

选择“group”作为因素变量，将其添加至【Fixed Variable(s)(固定变量)】列表框中。

选择“age”作为协变量，将其添加至【Covariate(s)(对比)】列表框中。



### Step05: 其他选项选择

单击【Options】按钮，弹出【Options(选项)】对话框。勾选【Descriptive(描述性统计量)】复选框表示输出描述性统计量；勾选【Homogeneity-of-variance(方差同质性检验)】复选框表示输出方差齐性检验表。再单击【Continue按钮】，返回主对话框。

提示：根据数据特点及您的实验要求，选择不同的均值多重比较方法。

### Step06: 完成操作

最后，单击【OK(确定)】按钮，操作完成。



**Univariate: Options**

Estimated Marginal Means

Factor(s) and Factor Interactions:  
(OVERALL)  
group

Display Means for:

Compare main effects

Confidence interval adjustment:  
LSD(none)

Display

Descriptive statistics  
 Estimates of effect size  
 Observed power  
 Parameter estimates  
 Contrast coefficient matrix

Homogeneity tests  
 Spread vs. level plot  
 Residual plot  
 Lack of fit  
 General estimable function

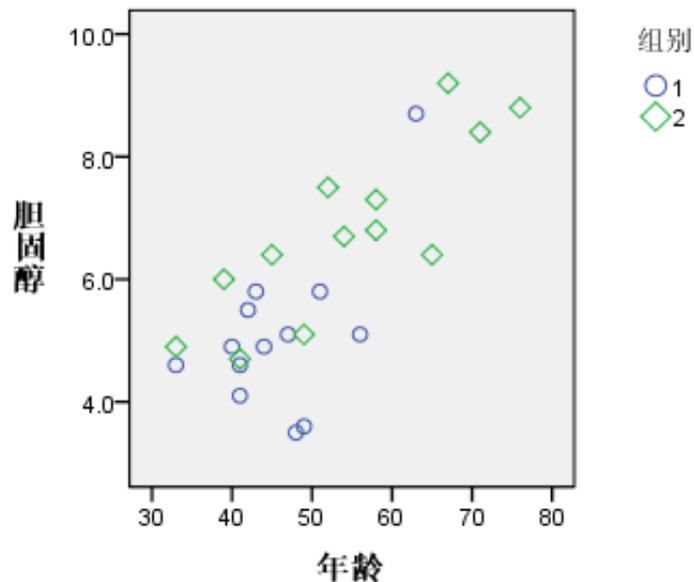
Significance level: .05    Confidence intervals are 95.0 %

Continue    Cancel    Help

### 3 实例结果及分析

#### (1) 散点图

散点图中，年龄为X轴，胆固醇为Y轴，体重组别作为分组标记，作出的散点图如下图所示。从中看到，实验对象的年龄和体内血清胆固醇含量呈较为明显的线性关系，且不同组别的斜率都基本相同。因此，可以将年龄变量作为协变量参与协方差分析。



## (2) 描述性统计分析表

表5-18和表5-19是对样本数据的基本描述结果。表5-18列出了两个组别的样本个数。表5-19列出了不同体重级别人群胆固醇含量的样本均值和标准差。从数值大小比较看，这两组人群胆固醇含量有一定的差异性，可以进一步采用方差分析。

表 5-18 不同组别样本容量

组别	N
1	13
2	13

表 5-19 描述性统计分析表

组别	Mean	Std. Deviation	N
1	5.092	1.3067	13
2	6.785	1.4416	13
Total	5.938	1.6005	26

### (3) 方差齐性检验

SPSS的结果报告接着列出了方差齐性检验结果表5-20。表格首先显示Levene统计量等于0.818。由于概率P值0.375明显大于显著性水平0.05，故认为两组样本数据的方差是相同的，满足方差分析的前提条件。

表 5-20 方差齐性检验表

	F	df1	df2	Sig.
胆固醇	.818	1	24	.375

#### (4) 协方差检验结果

表5-21列出了协方差检验结果，表5-21中包括各变差分解的情况、自由度、均方、F统计量值和概率P值。同时为了说明协方差模型的有效性，表5-22列出了只考虑体重级别的胆固醇单因素方差分析结果。

表 5-21 协方差检验结果表

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	42.995 <sup>a</sup>	2	21.498	23.493	.000
Intercept	1.527	1	1.527	1.668	.209
age	24.380	1	24.380	26.642	.000
group	4.458	1	4.458	4.872	.038
Error	21.047	23	.915		
Total	980.940	26			
Corrected Total	64.042	25			

a.  $R^2 = .671$  (调整  $R^2 = .643$ )

表 5-22 单因素方差分析结果表

Source	Type III Sum of Squares	df	Mean Square	F	Sig.
Corrected Model	18.615 <sup>a</sup>	1	18.615	9.835	.004
Intercept	916.898	1	916.898	484.425	.000
group	18.615	1	18.615	9.835	.004
Error	45.426	24	1.893		
Total	980.940	26			
Corrected Total	64.042	25			

a.  $R^2 = .291$  (调整  $R^2 = .261$ )

对比表5-21和表5-22，两种方差分析结果中，因变量的总变量（Corrected Total）都是64.042。同时单因素方差模型中，随机因素的可解释变差等于45.426。但是在协方差模型中，随机因素的可解释变差降低为21.047，这是由于扣除了年龄的影响造成的。这进一步说明了年龄变量对因变量的影响。不仅如此，体重级别可解释的变差由原来的18.615减少为4.458。这也是由于扣除了年龄因素的影响造成的。

综合起来，年龄因素对人体内胆固醇含量有显著的影响；同时，在排除了年龄因素的影响后，不同体重级别对胆固醇含量也存在显著的差异。可以通过表5-19看到：超重组的胆固醇含量要高于正常组的胆固醇含量。



## 第7章 SPSS的相关分析

# 7.1 相关分析概述

CONCEPT  
STRATE

## 7.1.1 相关的基本概念

### 1. 函数关系和相关关系

函数关系是指事物或现象之间存在着严格的依存关系，其主要特征是它的确定性，即对一个变量的每一个值，另一个变量都具有惟一确定的值与之相对应。变量之间的函数关系通常可以用函数式 $Y=f(x)$ 确切地表示出来。例如，圆的周长 $C$ 对于半径 $r$ 的依存关系就是函数关系： $C=2\pi r$ 。

相关关系反映出变量之间虽然相互影响，具有依存关系，但彼此之间是不能一一对应的。例如，学生成绩与其智力因素、各科学习成绩之间的关系、教育投资额与经济发展水平的关系、社会环境与人民健康的关系等等，都反映出客观现象中存在的相关关系。

# 7.1 相关分析概述

CONCEPT  
STRATE

## 2. 相关关系的类型

- (1) 根据相关程度的不同，相关关系可分为完全相关、不完全相关和无相关。
- (2) 根据变量值变动方向的**趋势**，相关关系可分为正相关和负相关。
- (3) 根据变量关系的**形态**，相关关系可分为直线相关和曲线相关。
- (4) 根据研究变量的**多少**，可分为单相关、复相关。

# 7.1.2 相关分析



CONCEPT  
RATE

## 1. 相关分析的作用

- (1) 判断变量之间有无联系
- (2) 确定选择相关关系的表现形式及相关分析方法
- (3) 把握相关关系的方向与密切程度
- (4) 相关分析不但可以描述变量之间的关系状况，而且用来进行预测。
- (5) 相关分析还可以用来评价测量量具的**信度**、**效度**以及项目的**区分度**等。

# 7.1.2 相关分析

CONCEPT  
STRATE

## 2. 相关系数

相关系数是在**直线相关**条件下，说明两个变量之间相关程度以及相关方向的统计分析指标。相关系数一般可以通过计算得到。作为**样本**相关系数，常用字母**r**表示；作为**总体**相关系数，常用字母 **$\rho$** 表示。

相关系数的数值范围是介于 -1 与 +1 之间（即  **$-1 \leq r \leq 1$** ），常用小数形式表示，一般要取小数点后**两位**数字来表示，以便比较精确地描述其相关程度。

两个变量之间的相关程度用相关系数r的**绝对值**表示，其**绝对值越接近1**，表明两个变量的**相关程度越高**；其**绝对值越接近于0**，表明两个变量**相关程度越低**。如果其**绝对值等于零**，则表示两个变量**完全直线相关**。如果其**绝对值为零**，则表示两个变量**完全不相关**（不是直线相关）。

## 7.1.2 相关分析

CONCEPT  
RATE

### 3. 相关系数

变量相关的方向通过相关系数 $r$ 所具有的符号来表示，“+”号表示正相关，即 $0 \leq r \leq 1$ 。“-”表示负相关，即 $0 \geq r \geq -1$ 。在使用相关系数时应该注意下面的几个问题。

- (1) 相关系数只是一个比率值，并不具备与相关变量相同的测量单位。
- (2) 相关系数 $r$ 受变量取值区间大小及样本数目多少的影响比较大。
- (3) 来自于不同群体且不同质的事物的相关系数不能进行比较。
- (4) 对于不同类型的数据，计算相关系数的方法也不相同。

## 7.2 SPSS在简单相关分析中的应用

CONCEPT  
STRATE

### 7.2.1 简单相关分析的基本原理

简单相关分析是研究两个变量之间关联程度的统计方法。它主要是通过计算简单相关系数来反映变量之间关系的强弱。一般它有图形和数值两种表示方式。

#### 1、相关图

在统计中制作相关图，可以直观地判断事物现象之间大致上呈现何种关系的形式。相关图是相关分析的重要方法。利用直角坐标系第一象限，把第一个变量置于横轴上，第二个变量置于纵轴上，而将两个变量对应的变量值用坐标点形式描绘出来，用以表明相关点分布状况的图形，这就是相关图

# 7.2 SPSS在简单相关分析中的应用

CONCEPT  
TRATE

## 2、相关系数

虽然相关图能够展现变量之间的数量关系，但这也只是种直观判断方法。因此，可以计算变量之间的相关系数。对不同类型的变量应当采取不同的相关系数来度量，常用的相关系数主要有：

### 皮尔逊（Pearson）相关系数

常称为积差相关系数，适用于研究连续变量之间的相关程度。例如，收入和储蓄存款、身高和体重等变量间的线性相关关系。注意Pearson相关系数适用于线性相关的情形，对于曲线相关等更为复杂的情形，系数的大小并不能代表其相关性的强弱。它的计算公式为：

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \sum_{i=1}^n (y_i - \bar{y})^2}}$$

利用相关系数r的大小可以判断变量间相关关系的密切程度，具体见表所示。

## 7.2 SPSS在简单相关分析中的应用

CONCEPT  
STRATE

### 7.2.1 简单相关分析的基本原理

中

相关系数的值	直线相关程度
$ r =0$	完全不相关
$0 <  r  \leq 0.3$	微弱相关
$0.3 <  r  \leq 0.5$	低度相关
$0.5 <  r  \leq 0.8$	显著相关
$0.8 <  r  \leq 1$	高度相关
$ r =1$	完全相关

□

## 7.2 SPSS在简单相关分析中的应用

CONCEPT  
STRATE

对Pearson简单相关系数的统计检验是计算t统计量，t统计量服从n-2个自由度的t分布。SPSS会自动计算r统计量和t值，并依据t分布表给出其对应的相伴概率值。

Spearman等级相关系数用来度量顺序水准变量间的线性相关关系。它是利用两变量的秩次大小作线性相关分析，适用条件为：

- ① 两个变量的变量值是以等级次序表示的资料；
- ② 一个变量的变量值是等级数据，另一个变量的变量值是等距或比率数据，且其两总体不要求是正态分布，样本容量n不一定大于30。

## 7.2 SPSS在简单相关分析中的应用

CONCEPT  
STRATE

从斯皮尔曼等级相关适用条件中可以看出，**等级相关**的应用范围要比**积差相关**广泛，它的突出优点是对数据的**总体分布**、**样本大小**都不做要求。但缺点是计算**精度不高**。斯皮尔曼等级相关系数常用符号  $r_R$  来表示。其基本公式为：

$$r_R = 1 - \frac{6 \sum D^2}{n(n^2 - 1)}$$

式中：D是两个变量每对数据等级之差，n是两列变量值的**对数**。

Spearman相关系数计算公式可以完全套用Pearson相关系数的计算公式，但公式中的x和y用它们的**秩次**代替即可。

## 7.2 SPSS在简单相关分析中的应用

CONCEPT  
STRATE

### Kendall's 等级相关系数

它是用于反映分类变量相关性的指标，适用于两个变量均为有序分类的情况。这种指标采用非参数检验方法测度变量间的相关关系。它利用变量的秩计算一致对数目和非一致对数目。显然，如果两变量具有较强的正相关，则一致对数目 $U$ 应较大；但若两变量相关性较弱，则一致对数目 $U$ 和非一致对数目 $V$ 应大致相等。故按照此思想，可得其定义为：

SPSS将自动计算它的相关系数、检验统计量和对应的概率 $P$ 值。

## 7.2 SPSS在简单相关分析中的应用

CONCEPT  
STRATE

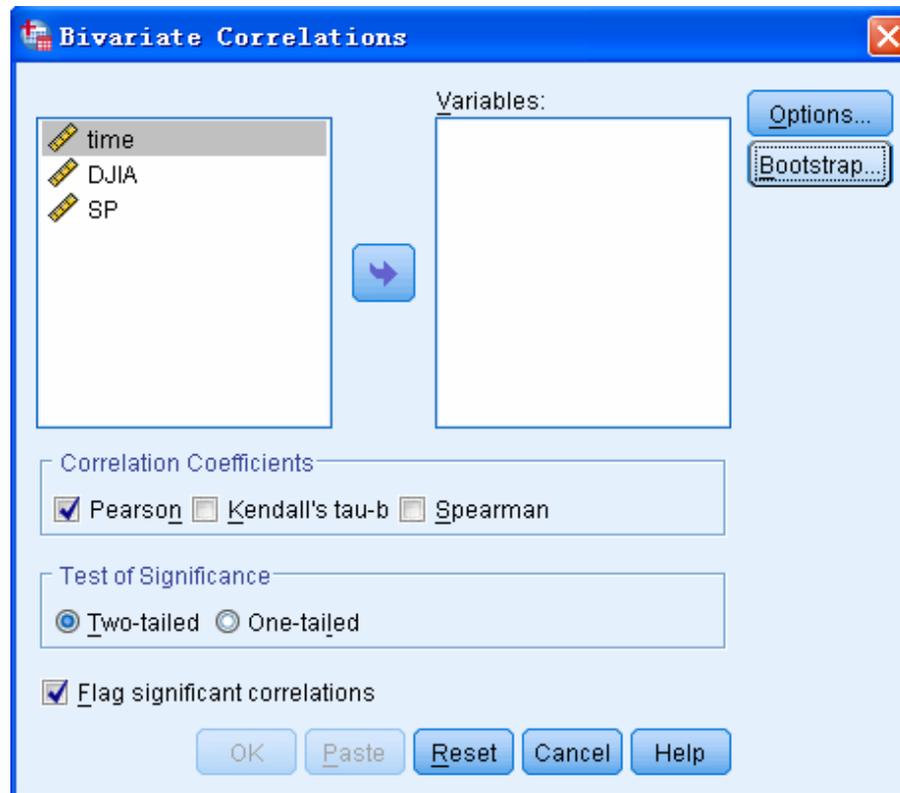
### 7.2.2 简单相关分析的SPSS操作详解

Step01: 打开主菜单

选择菜单栏中的【Analyze(分析)】→【Correlate(相关)】→【Bivariate(双变量)】命令，弹出【Bivariate Correlations(双变量相关)】对话框，如图7-1所示，这是简单相关检验的主操作窗口。

## 7.2 SPSS在简单相关分析中的应用

CONCEPT  
STRATE



## 7.2 SPSS在简单相关分析中的应用

### Step02: 选择检验变量

在【Bivariate Correlations(双变量相关)】对话框左侧的候选变量列表框中选择两个或两个以上变量将其添加至【Variables(变量)】列表框中，表示需要进行简单相关分析的变量。

### Step03: 选择相关系数类型

图中的【Correlation Coefficients(相关系数)】选项组中可以选择计算简单相关系数的类型。

- Pearson: 系统默认项，即积差相关系数，计算连续变量或是等间距测度的变量间的相关分析。
- Kendall: 等级相关，计算分类变量间的秩相关。
- Spearman: 等级相关，斯皮尔曼相关系数。

对于非等间距测度的连续变量，因为分布不明可以使用等级相关分析，也可以使用Pearson 相关分析；对于完全等级的离散变量必须使用等级相关分析相关性。当资料不服从双变量正态分布或总体分布型未知，或原始数据是用等级表示时，宜用Spearman 或Kendall相关。

## 7.2 SPSS在简单相关分析中的应用

### Step04: 假设检验类型选择

在图中的【Test of Significance(显著性检验)】选项组中可以选择输出的假设检验类型，对应有两个单选项。

- Two tailed: 系统默认项。**双尾检验**，当**事先不知道相关方向**（正相关还是负相关）时选择此项。
- One tailed: **单尾检验**，如果**事先知道相关方向**可以选择此项。

同时，可以勾选【Flag significant Correlations(标记显著性相关)】复选框。它表示选择此项后，输出结果中对在显著性水平0.05下显著相关的相关系数用一个星号“\*”加以标记；对在显著性水平0.01下显著相关的相关系数用两个星号“\*\*”标记。

## 7.2 SPSS在简单相关分析中的应用

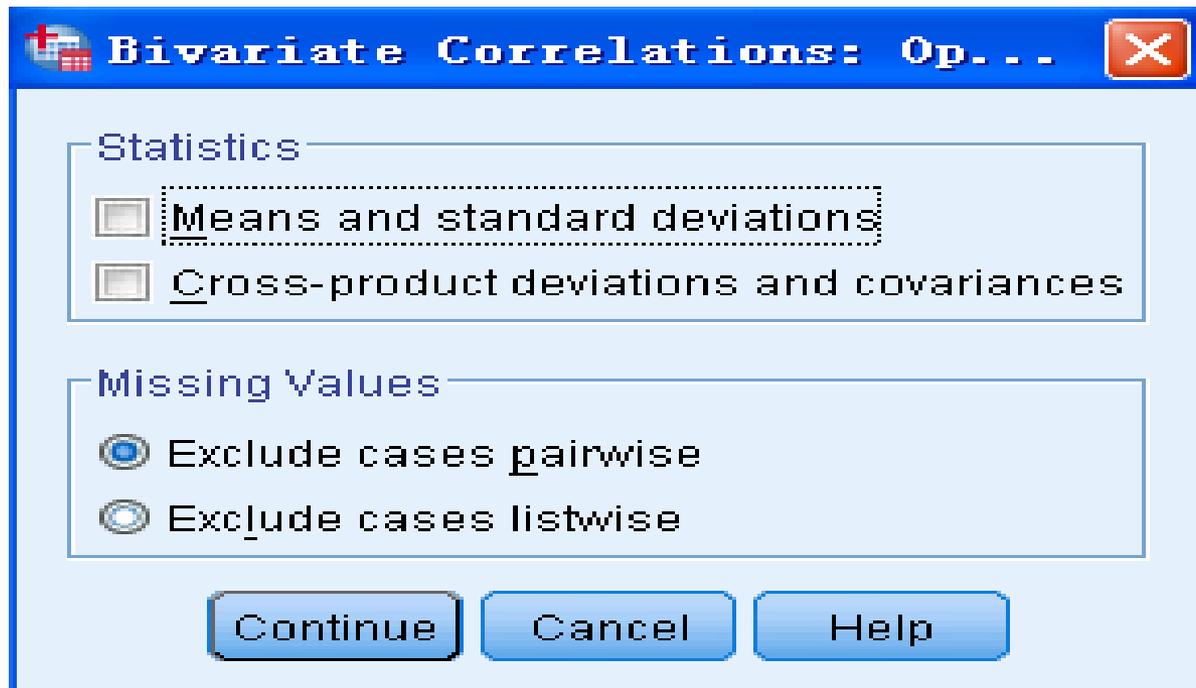
### Step05: 其他选项选择

单击【Options (选项)】按钮，弹出的对话框用于指定输出内容和关于缺失值的处理方法，主要包括以下选项。

- ① Statistics: 选择输出统计量。
  - Means and standard deviations: 将输出选中的各变量的观测值数目、均值和标准差。
  - Cross-product deviations and covariances: 输出反映选中的每一对变量之间的叉积离差矩阵和协方差矩阵。
- ② Missing Values: 用于设置缺失值的处理方式。它有两种处理方式:
  - Exclude cases pairwise: 系统默认项。剔除当前分析的两个变量值是缺失的个案。
  - Exclude cases listwise: 表示剔除所有含缺失值的个案后再进行分析。

## 7.2 SPSS在简单相关分析中的应用

CONCEPT  
STRATE



## 7.2 SPSS在简单相关分析中的应用

CONCEPT  
STRATE

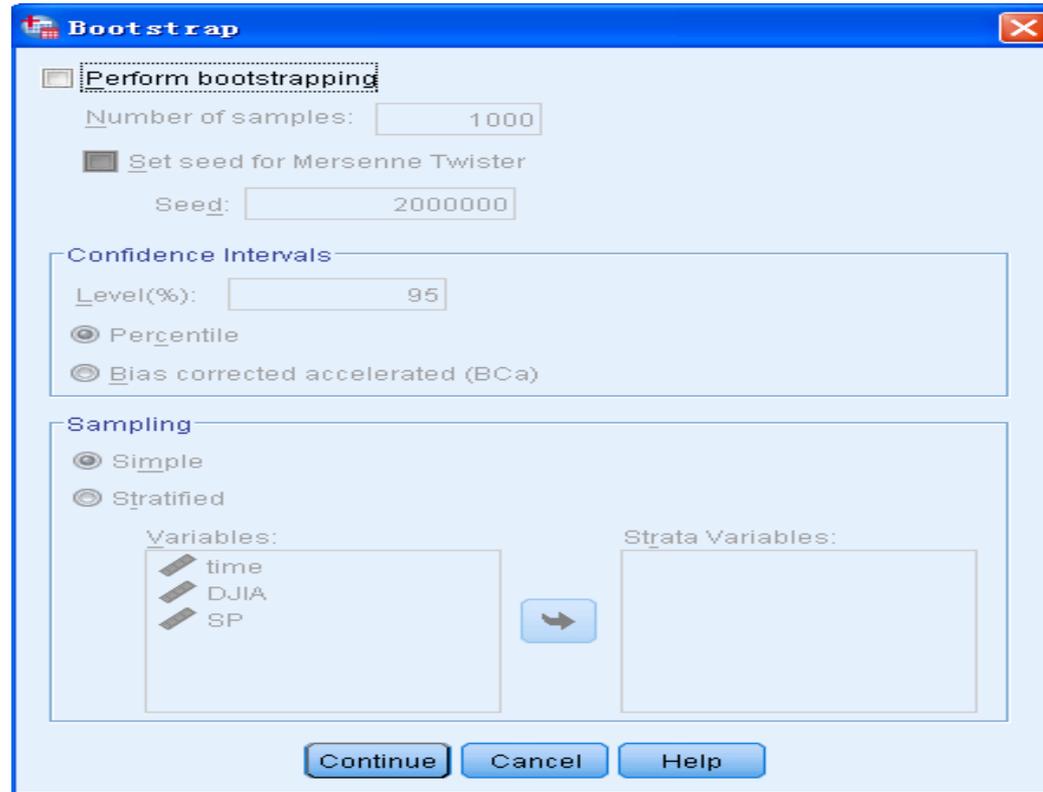
Step06: 相关统计量的Bootstrap估计

单击【Bootstrap】按钮，在弹出的对话框中可以进行如下统计量的Bootstrap估计。

- 描述统计表支持均值和标准差的Bootstrap 估计。
- 相关性表支持相关性的Bootstrap 估计。

## 7.2 SPSS在简单相关分析中的应用

CONCEPT  
STRATE



Step07: 单击【OK】按钮，结束操作，SPSS软件自动输出结果。

## 7.2 SPSS在简单相关分析中的应用

CONCEPT  
STRATE

### 7.2.3 实例分析：股票指数之间的联系

#### 1. 实例内容

道琼斯工业平均指数（DJIA）和标准普尔指数500（S&P 500）都被用做股市全面动态的测度。DJIA是基于30种股票的价格动态；S&P 500是由500种股票组成的指数。有人说S&P 500是股票市场功能的一种更好的测度，因为它基于更多的股票。表7-2显示了DJIA和S&P 500在1997年10周内的收盘价。请计算它们之间的样本相关系数。不仅如此，样本相关系数告诉我们DJIA和S&P 500之间的关系是怎样的？

## 7.2 SPSS在简单相关分析中的应用

CONCEPT  
STRATE

### 2. 实例操作

表给出了道琼斯工业平均指数和标准普尔指数在同一时间点的数值。由于这些数值都是连续型变量，同时根据两个股票指数的散点图，可见它们呈显著的线性相关，因此可以采用Pearson相关系数来测度它们之间的相关性。但为了比较，我们也计算了这两组变量的Kendall和Spearman相关系数。

## 7.2 SPSS在简单相关分析中的应用

CONCEPT  
TRATE

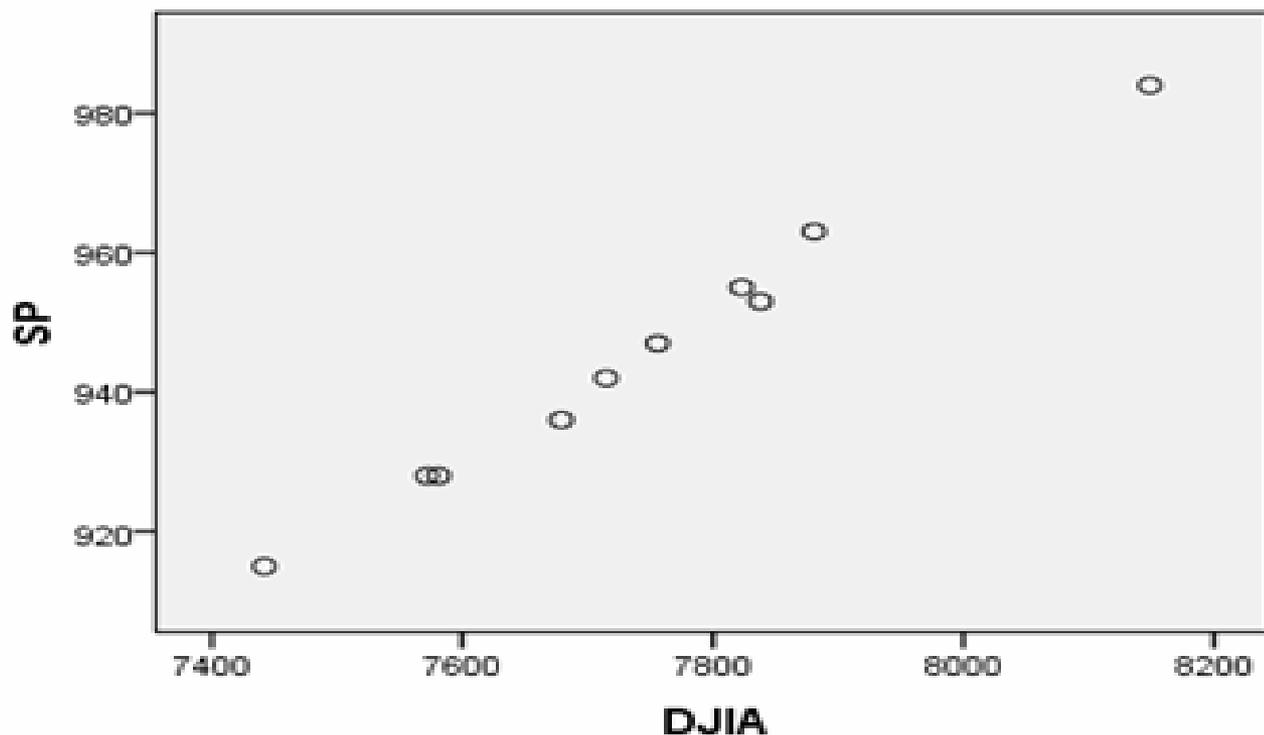


图 7-4 股票指数散点图

## 7.2 SPSS在简单相关分析中的应用

### 3. 实例结果及分析

#### (1) 描述性统计分析表

执行完上面的操作后，首先给出的是当前样本进行描述性统计的结果，如表7-3所示。可以看到样本容量都等于10，道琼斯工业平均指数和标准普尔指数的平均均值分别为7743.60和945.10，两者差距显著。同时，两者的方差差距也很明显。

表 7-3 描述性统计分析表

	Mean	Std. Deviation	N
DJIA	7743.60	197.326	10
SP	945.10	19.947	10

## 7.2 SPSS在简单相关分析中的应用

CONCEPT  
STRATE

### (2) Pearson相关系数表

接着SPSS列出了道琼斯工业平均指数和标准普尔指数的Pearson相关系数表7-4。可以看到，两种指数的Pearson系数值高达0.995，非常接近1；同时相伴概率P值明显小于显著性水平0.01，这也进一步说明两者高度正线性相关。

表 7-4 Pearson 相关系数表

		DJIA	SP
DJIA	Pearson Correlation	1	.995**
	Sig. (2-tailed)		.000
	N	10	10
SP	Pearson Correlation	.995**	1
	Sig. (2-tailed)	.000	
	N	10	10

注：“\*\*”表示相关系数在0.01的显著性水平（双尾）上显著相关。

## 7.2 SPSS在简单相关分析中的应用

CONCEPT  
STRATE

### (3) 非参数相关系数表

表7-5列出了两种股票指数的Kendall和Spearman相关系数，分别等于0.994和0.985；同时它们的概率P值也远小于显著性水平。但本案例中，Spearman相关系数和Kendall相关系数都小于Pearson相关系数，显然这是由于在秩变换或数据按有序分类处理时损失信息所导致的。

所以，通过以上分析看到，道琼斯工业平均指数和标准普尔指数具有高度正相关性，一个指数的上涨或下跌时，另一个指数也会伴随着上涨或下跌。

## 7.2 SPSS在简单相关分析中的应用

CONCEPT  
RATE

表 7-5 非参数相关系数表

			DJIA	SP
Kendall's tau_b	DJIA	Correlation Coefficient	1.000	.944**
		Sig. (2-tailed)	.	.000
		N	10	10
	SP	Correlation Coefficient	.944**	1.000
		Sig. (2-tailed)	.000	.
		N	10	10
Spearman's rho	DJIA	Correlation Coefficient	1.000	.985**
		Sig. (2-tailed)	.	.000
		N	10	10
	SP	Correlation Coefficient	.985**	1.000
		Sig. (2-tailed)	.000	.
		N	10	10

注：“\*\*\*”表示相关系数在 0.01 的显著性水平（双尾）上显著相关。

# 7.3 SPSS在偏相关分析中的应用

CONCEPT  
STRATE

## 7.3.1 偏相关分析的基本原理

### 1. 方法概述

简单相关分析计算两个变量之间的相互关系，分析两个变量间线性关系的程度。但是现实中，事物之间的联系可能存在于多个主体之间，因此往往因为第三个变量的作用使得相关系数不能真实地反映两个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 相关计算其相关系数，可以得出肺活量、身高和体重均存在较强的线性相关性质。但实际上呢，对体重相同的人而言，身高值越大其肺活量也不一定越大。因为身高与体重有着线性关系，肺活量与体重有着线性关系，因此得出了身高与肺活量之间存在较强的线性关系的错误结论。偏相关分析就是在研究两个变量之间的线性相关关系时控制可能对其产生影响的变量。

## 7.3 SPSS在偏相关分析中的应用

CONCEPT  
STRATE

### 2. 基本原理

偏相关分析是在相关分析的基础上考虑了两个因素以外的各种作用，或者说在扣除了其他因素的作用大小以后，重新来测度这两个因素间的关联程度。这种方法的目的就在于消除其他变量关联性的传递效应。

偏相关系数在计算时可以首先分别计算三个因素之间的相关系数，然后通过这三个简单相关系数来计算偏相关系数，公式如下：

$$r_{12(3)} = \frac{r_{12} - r_{13}r_{23}}{\sqrt{1-r_{13}^2}\sqrt{1-r_{23}^2}}$$

上式就是在控制了第三个因素的影响所计算的第一、第二个因素之间的偏相关系数。当考虑一个以上的控制因素时的公式类推。

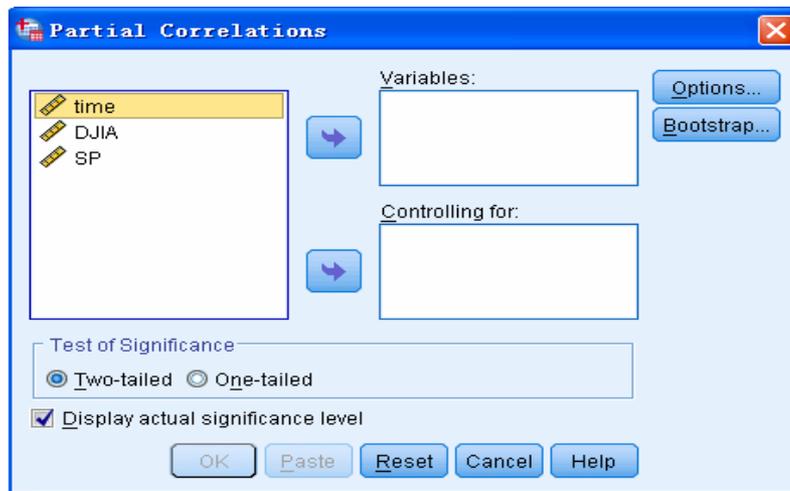


## 7.3 SPSS在偏相关分析中的应用

### 7.3.2 偏相关分析的SPSS操作详解

Step01: 打开主菜单

选择菜单栏中的【Analyze(分析)】→【Correlate(相关)】→【Partial(偏相关)】命令，弹出【Partial Correlations(偏相关)】对话框，如图7-9所示，这是偏相关检验的主操作窗口。





## 7.3 SPSS在偏相关分析中的应用

### Step02: 选择检验变量

在【Bivariate Correlations(偏相关)】对话框左侧的候选变量列表框中选择两个或两个以上变量，将其添加至【Variables(变量)】列表框中，表示需要进行偏相关分析的变量。

### Step03: 选择控制变量

在【Bivariate Correlations(偏相关)】对话框左侧的候选变量列表框中至少选择一个变量，将其添加至【Controlling for(控制)】列表框中，表示在进行偏相关分析时需要控制的变量。注意如果不选入控制变量，则进行的是简单相关分析。

### Step04: 假设检验类型选择

在【Test of Significance(显著性检验)】选项组中可以选择输出的假设检验类型，对应有以下两个选项。

- Two tailed: 系统默认项。双尾检验，当事先不知道相关方向（正相关还是负相关）时选择此项。
- One tailed: 单尾检验，如果事先知道相关方向可以选择此项。

同时，可以勾选【Flag significant Correlations】复选框。它表示选择此项后，输出结果中对在显著性水平0.05下显著相关的相关系数用一个星号“\*”加以标记；对在显著性水平0.01下显著相关的相关系数用两个星号“\*\*”标记。



## 7.3 SPSS在偏相关分析中的应用

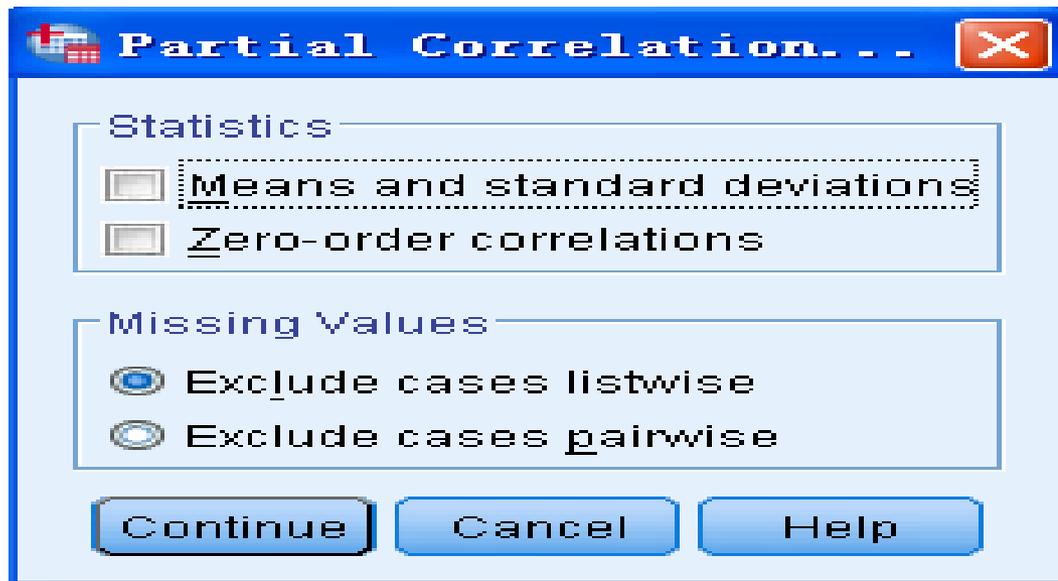
### Step05: 其他选项选择

单击【Options】按钮，弹出的对话框用于指定输出内容和关于缺失值的处理方法，主要包括以下选项。

- ① Statistics: 选择输出统计量。
  - Means and standard deviations: 将输出选中的各变量的观测值数目、均值和标准差。
  - Zero-order correlation: 显示零阶相关矩阵，即Pearson 相关矩阵。
- ② MissingValues: 用于设置缺失值的处理方式。它有两种处理方式：
  - Exclude cases pairwise: 系统默认项。剔除当前分析的两个变量值是缺失的个案。
  - Exclude cases listwise: 表示剔除所有含缺失值的个案后再进行分析。



## 7.3 SPSS在偏相关分析中的应用



## 7.3 SPSS在偏相关分析中的应用

CONCEPT  
STRATE

Step06: 相关统计量的Bootstrap估计

单击【Bootstrap】按钮，在弹出的对话框中可以进行如下统计量的Bootstrap估计。

- 描述统计表支持均值和标准差的Bootstrap 估计。
- 相关性表支持相关性的Bootstrap 估计。

Step07: 单击【OK】按钮，结束操作，SPSS软件自动输出结果。



## 7.3 SPSS在偏相关分析中的应用

### 7.3.3 实例分析：股票市场和债券市场

#### 1. 实例内容

在我国的金融市场中，股票市场和债券市场都是其中的重要组成部分。研究它们之间的关系有利于我们弄清楚金融市场之间的关联特征。但是我国债券市场主要由银行间债券市场和证券交易所债券市场组成，并且它们处于相对分割状态，在投资主体、交易方式等方面存在显著差异。数据文件7-2.sav列出了近几年我国股票市场、交易所国债市场和银行间国债市场的综合指数，请利用相关分析研究这三个市场的关联特征



## 7.3 SPSS在偏相关分析中的应用

### 2. 实例操作

由于这里要研究三个金融市场之间的关系，因此首先可以利用7.2节的简单相关分析来初步探讨它们之间的联系。表7-6计算了这三个市场之间的Pearson相关系数。从表中数据看到，三个市场间的价格相关系数较高，其中交易所和银行间国债市场相关系数高达0.922，而它们和股市的相关系数相对较低，分别是0.411和0.419，从数值大小看到这两个子市场和股市的关联性差异不明显。

但是，就相关系数本身而言，它未必是两事物间线性关系强弱的真实体现，往往有夸大的趋势，因为它在计算时都没有考虑第三方的影响，这就有可能导致对事物的解释出现偏差。这里，股市、银行间国债市场和交易所国债市场之间肯定是相互关联的，两个市场间的关系强弱肯定要受到第三方的影响制约，市场间的关系强弱可能存在传递效应。基于这种考虑，这里要引入偏相关系数测度市场间的关系。



## 7.3 SPSS在偏相关分析中的应用

### 3. 实例结果及分析

#### (1) 描述性统计分析表

执行完上述操作后，首先给出的是当前样本进行描述性统计的结果表7-7。可以看到样本容量都等于1321，三个市场综合指数的样本均值和样本方差都有一定的差距。

表 7-7 描述性统计分析表

	Mean	Std. Deviation	N
股票指数	1.6305E3	613.14785	1321
交易所国债指数	1.0625E2	5.51477	1321
银行间国债指数	1.0765E2	5.15824	1321



## 7.3 SPSS在偏相关分析中的应用

### (2) 偏相关系数表

表7-8~表7-10列出了三个市场之间的偏相关系数。在控制了股市指数后，银行间和交易所市场间的相关系数没有发生太大变化，仍然高达0.906，说明了这两个市场的关系密切且股市对两市波动影响较小。而银行间国债市场、交易所国债市场与股市的偏相关系数却发生了显著变化：银行间市场和股市的Pearson相关系数为0.419，而在控制了交易所指数后，它们之间的偏相关系数下降为0.114；同理，交易所国债市场和股市的相关系数也由0.411下降到0.070。这说明了第三方市场对剩余两个市场确实存在显著影响，通过简单相关系数还无法深入刻画市场之间的关系。这里引入偏相关系数是比较适合的。

## 7.3 SPSS在偏相关分析中的应用

表 7-8 股市和交易所国债市场的偏相关系数

+

Control Variables		股票指数	交易所国债指数
银行间国债指数	股票指数	Correlation	.070
		Significance (2-tailed)	.011
		df	1318
交易所国债指数	交易所国债指数	Correlation	.070
		Significance (2-tailed)	.011
		df	1318

□

## 7.3 SPSS在偏相关分析中的应用

表 7-9 股市和银行间国债市场的偏相关系数

Control Variables		股票指数	银行间国债指数
交易所国债指数	股票指数	Correlation	1.000
		Significance (2-tailed)	.000
		df	0
银行间国债指数	银行间国债指数	Correlation	.114
		Significance (2-tailed)	.000
		df	1318

□



## 7.3 SPSS在偏相关分析中的应用

表 7-10 交易所和银行间国债市场的偏相关系数

+

Control Variables+			交易所国债指数+	银行间国债指数+
股票指数+	交易所国债指数+	Correlation+	1.000	.906+
		Significance (2-tailed)+	.	.000+
		df+	0	1318+
	银行间国债指数+	Correlation+	.906	1.000+
		Significance (2-tailed)+	.000	.+
		df+	1318	0+

□

## 7.4 SPSS在距离分析中的应用

CONCEPT  
STRATE

### 7.4.1 距离分析的基本原理

简单相关分析和偏相关分析有一个共同点，那就是对分析的数据背景应当有一定程度的了解。但在实际中有时会遇到一种情况，在分析前对数据所代表的专业背景知识尚不充分，本身就属于探索性的研究。这时就需要先对各个指标或者案例的差异性、相似程度进行考察，以先对数据有一个初步了解，然后再根据结果考虑如何进行深入分析。



## 7.4 SPSS在距离分析中的应用

距离分析是对观测量之间或变量之间相似或不相似的程度的一种测度，是计算一对变量之间或一对观测量之间的广义的距离。根据变量的不同类型，可以有许多距离、相似程度测量指标供用户选择。但由于本模块只是一个预分析过程，因此距离分析并不会给出常用的P值，而只能给出各变量/记录间的距离大小，以供用户自行判断相似性。

调用距离分析过程可对变量内部各观察单位间的数值进行距离相关分析，以考察相互间的接近程度；也可对变量间进行距离相关分析，常用于考察预测值对实际值的拟合程度，也可用于考察变量的相似程度。在距离分析中，主要利用变量间的相似性测度（Similarities）和不相似性测度（Dissimilarities）度量研究对象之间的关系。



## 7.4 SPSS在距离分析中的应用

### 7.4.2 距离分析的SPSS操作详解

Step01: 打开对话框

选择菜单栏中的【Analyze(分析)】→【Correlate(相关)】→【Distances(距离)】命令，弹出【Distances(距离)】对话框，这是距离分析的主操作窗口。





## 7.4 SPSS在距离分析中的应用

### Step02: 选择检验变量

在【Distances(距离)】对话框左侧的候选变量列表框中选择两个或两个以上变量，将其添加至【Variables(变量)】列表框中，表示需要进行距离分析的变量。同时可以选择一个字符型标示变量移入【Label Cases(标注个案)】列表框中，在输出中将用这个标示变量值对各个观测量加以标记。缺省时，输出中用观测量的序号来标记。

### Step03: 选择分析类型

在【Compute Distances(计算距离)】选项组中可以选择计算何种类型的距离。

- Between cases: 系统默认项。表示作变量内部观察值之间的距离相关分析。
- Between variables: 表示作变量之间的距离相关分析。



## 7.4 SPSS在距离分析中的应用

### Step04: 测度类型选择

在【Measure(度量标准)】选项组中可以选择分析时采用的距离类型。

- Dissimilarities: 系统默认项。不相似性测距，系统默认采用**欧式距离测度**观测值或变量之间的不相似性。
- Similarities: 相似性测距。系统默认使用Pearson相关系数测度观测值或变量之间的相似性。

### Step05: 完成操作

单击【OK】按钮，结束操作，SPSS软件自动输出结果。

上述第四步中除了采用系统默认的距离测度类型外，还可以根据用户的需要自己选择测度类型，由于这里专业性很强，而且实际中使用很少，下面只做些简单的介绍。

在【Distances(距离)】对话框中，选择【Dissimilarities(不相似性)】距离类型后，单击【Measure】按钮，弹出下图所示的对话框。

## 7.4 SPSS在距离分析中的应用

**Distances: Dissimilarity Measures**

**Measure**

**Interval**  
Measure: Euclidean distance  
Power: 2 Root: 2

**Counts**  
Measure: Chi-square measure

**Binary**  
Measure: Euclidean distance  
Present: 1 Absent: 0

**Transform Values**

Standardize: None  
 By variable  
 By case

**Transform Measures**

Absolute values  
 Change sign  
 Rescale to 0-1 range

Continue Cancel Help



## 7.4 SPSS在距离分析中的应用

选择【Similarities(相似性)】时各种数据类型可用的测距方法有以下几种。

### ① Interval: 计量资料。

- Pearson correlation: 以Pearson相关系数为距离。
- Cosine: 以变量矢量的余弦值为距离, 介于-1至+1之间。

### ② Binary: 二分类变量。

- Russell and Rao: 以二分点乘积为配对系数。
- Simple matching: 以配对数与总对数的比例为配对系数。
- Jaccard: 相似比例, 分子与分母中的配对数与非配对数给予相同的权重。
- Dice: Dice配对系数, 分子与分母中的配对数给予加倍的权重。
- Rogers and Tanimoto: Rogers and Tanimoto配对系数, 分母为配对数, 分子为非配对数, 非配对数给予加倍的权重。
- Sokal and Sneath 1: Sokal and Sneath I型配对系数, 分母为配对数, 分子为非配对数, 配对数给予加倍的权重。
- Sokal and Sneath 2: Sokal and Sneath II型配对系数, 分子与分母均为非配对数, 但分子给予加倍的权重。



## 7.4 SPSS在距离分析中的应用

- Sokal and Sneath 3: Sokal and Sneath III型配对系数, 分母为配对数, 分子为非配对数, 分子与分母的权重相同。
- Kulczynski 1: Kulczynski I型配对系数, 分母为总数与配对数之差, 分子为非配对数, 分子与分母的权重相同。
- Kulczynski 2: Kulczynski平均条件概率。
- Sokal and Sneath 4: Sokal and Sneath条件概率。
- Hamann: Hamann概率。
- Lambda: Goodman-Kruskai相似测量的 $\lambda$ 值。
- Anderberg 's D: 以一个变量状态预测另一个变量状态。
- Yule 's Y: Yule综合系数, 属于 $2 \times 2$ 四格表的列联比例函数。
- Yule 's Q: Goodman-Kruskal  $\gamma$ 值, 属于 $2 \times 2$ 四格表的列联比例函数。
- Ochiai: Ochiai二分余弦测量。
- Sokal and Sneath 5: Sokal and Sneath V型相似测量。
- Phi 4 point correlation: Pearson相关系数的平方值。
- Dispersion: Dispersion相似测量。

## 7.4 SPSS在距离分析中的应用

CONCEPT  
STRATE

进行标准化的方法在【Standized(标准化)】后面的下拉列表中。单击矩形框右面的箭头按钮展开下拉列表，可选择的标准方法如下。

- None：不作数据转换，系统默认项。
  - Z-Scores：作标准Z分值转换，此时均值等于0，标准差等于1。
  - Range -1 to 1：作-1至+1之间的标准化转换。
  - Range 0 to 1：作0至1之间的标准化转换。
  - Maximum magnitude of 1：作最大值等于1的标准转换。
  - Mean of 1：作均数单位转换。
  - Standard deviation of 1：作标准差单位转换。
- 【Transform Values(转换值)】复选项：选择测度转换方法。在距离测度计算完成后，才进行对测度的转换。共有3个转换方法可以选择。每种转换方法给出一种转换结果。3种转换方法可以同时选择。



## 7.4 SPSS在距离分析中的应用

- Absolute values: 对距离取绝对值。当符号表明的是相关的方向，且仅对相关的数值感兴趣时使用这种转换。
- Change sign: 改变符号。把相似性测度值转换成不相似性测度值或相反。
- Rescale to 0~1 range: 重新调整测度值到范围0~1转换法。对已经按有意义的方法标准化的测度，一般不再使用此方法进行转换。



## 7.4 SPSS在距离分析中的应用

### 7.4.3 实例分析：价格指数的相关性

#### 1. 实例内容

价格指数是用来反映不同时期商品价格水平的变化方向、趋势和程度的经济指标，它属于经济指数的一种，通常以报告期和基期相对比的相对数来表示。价格指数是研究价格动态变化的一种工具，它为制定、调整 and 检查各项经济政策，特别是价格政策提供依据。表7-11列出了我国1991年—2005年间居民消费价格指数、城市居民消费价格指数、农村居民消费价格指数、商品销售价格指数、工业品出厂价格指数、原材料等购进价格指数和固定资产投资价格指数。请研究这些价格指数之间的关系。



## 7.4 SPSS在距离分析中的应用

### 2. 实例操作

本案例要讨论居民消费价格指数等七类价格指数之间关联特征。由于这些价格指数的构成复杂，因此可以采用距离分析来探讨它们之间的关系。由于都属于连续型数据，这里可以选择不相似性测距中的欧式距离来测度。



## 7.4 SPSS在距离分析中的应用

### 3. 实例结果及分析

#### (1) 基本统计汇总表

表7-12是对个案的基本统计汇总分析。本案例的样本数目等于15，没有缺失数据。

表 7-12 基本统计汇总表

Cases					
Valid		Missing		Total	
N	Percent	N	Percent	N	Percent
15	100.0%	0	.0%	15	100.0%



## 7.4 SPSS在距离分析中的应用

### (2) 距离矩阵

表7-13是根据欧式距离计算出的各个价格指数之间的距离。如果距离数值越小，说明两个价格指数越相关；反之。可以看到，居民消费价格指数和城市居民消费价格指数、农村居民消费价格指数和商品销售价格指数的距离都较小，说明它们都反映了社会总体或某方面价格水平的高低；但是它和工业品出厂价格指数、原材料等购进价格指数和固定资产投资价格指数的距离都较大，说明这些价格指数反映的类型有较大差别。其余指数之间的关系可以类似分析。



## 7.4 SPSS在距离分析中的应用

表 7-13 距离矩阵

	Euclidean Distance						
	居民消费价格指数	城市居民消费价格指数	农村居民消费价格指数	商品销售价格指数	工业品出厂价格指数	原材料等购进价格指数	固定资产投资价格指数
居民消费价格指数	.000	3.416	2.757	6.261	14.090	26.101	24.449
城市居民消费价格指数	3.416	.000	6.125	8.208	13.809	24.943	23.260
农村居民消费价格指数	2.757	6.125	.000	6.033	14.713	27.006	25.597
商品销售价格指数	6.261	8.208	6.033	.000	14.192	27.727	23.874
工业品出厂价格指数	14.090	13.809	14.713	14.192	.000	14.419	17.344
原材料等购进价格指数	26.101	24.943	27.006	27.727	14.419	.000	19.302
固定资产投资价格指数	24.449	23.260	25.597	23.874	17.344	19.302	.000

注：此为不相似测度矩阵。



## 第8章

# SPSS的回归分析

# 8.1 SPSS 在一元线性回归分析中的应用

CONCEPT  
STRATE

## 8.1.1 一元线性回归的基本原理

### 1. 方法概述

线性回归模型侧重考察变量之间的数量变化规律，并通过线性表达式，即线性回归方程，来描述其关系，进而确定一个或几个变量的变化对另一个变量的影响程度，为预测提供科学依据。

一般线性回归的基本步骤如下。

- ① 确定回归方程中的自变量和因变量。
- ② 从收集到的样本数据出发确定自变量和因变量之间的数学关系式，即确定回归方程。
- ③ 建立回归方程，在一定统计拟合准则下估计出模型中的各个参数，得到一个确定的回归方程。
- ④ 对回归方程进行各种统计检验。
- ⑤ 利用回归方程进行预测。

# 8.1 SPSS 在一元线性回归分析中的应用

CONCEPT  
STRATE

## 2、基本原理

当自变量和因变量之间呈现显著的线性关系时，则应采用线性回归的方法，建立因变量关于自变量的线性回归模型。根据自变量的个数，线性回归模型可分为一元线性回归模型和多元线性回归模型

一元线性回归模型是在不考虑其他影响因素的条件下，或是在认为其他影响因素确定的情况下，分析某一个因素（自变量）是如何影响因变量的。一元线性回归的经验模型是：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x$$

式中， $\hat{\beta}_0$  表示回归直线在纵轴上的截距，是回归系数，它表示当自变量变动一个单位所引起的因变量的平均变动值。



## 8.1 SPSS 在一元线性回归分析中的应用

### 3. 统计检验

在求解出了回归模型的参数后，一般不能立即将结果付诸于实际问题的分析和预测，通常要进行各种统计检验，例如拟合优度检验、回归方程和回归系数的显著性检验和残差分析等。这些内容，我们将结合案例来具体讲解。

## 8.1.2 一元线性回归的SPSS操作详解

CONCEPT  
STRATE

Step01: 打开对话框

选择菜单栏中的【Analyze(分析)】→【Regression(回归)】→【Linear(线性)】命令，弹出【Linear Regression(线性回归)】对话框，这是线性回归分析的主操作窗口。

Step02: 选择因变量

在【Linear Regression(线性回归)】对话框左侧的候选变量列表框中选择一个变量，将其添加至【Dependent(因变量)】列表框中，即选择该变量作为一元线性回归的因变量。

## 8.1.2 一元线性回归的SPSS操作详解

CONCEPT  
STRATE

### Step03: 选择自变量

在【Linear Regression (线性回归)】对话框左侧的候选变量列表框中选择一个变量，将其添加至【Independent(s) (自变量)】列表框中，即选择该变量作为一元线性回归的自变量。

## 8.1.2 一元线性回归的SPSS操作详解

CONCEPT  
STRATE

Step04: 选择回归模型中自变量的进入方式

在【Method（方法）】选项组中可以选择自变量的进入方式，一共有五种方法。可单击【Independent(s)（自变量）】列表框上方的【Next】按钮，选定的这一组自变量将被系统自动保存于一个自变量块（Block）中。接下来选择另一组自变量，单击【Next】按钮将它们保存于第二个自变量块中。重复上述操作，可以保存若干个自变量块。若需要输出以哪一组变量为自变量的回归方程，可以通过单击【Previous】按钮和【Next】按钮来选择。

## 8.1.2 一元线性回归的SPSS操作详解

CONCEPT  
STRATE

### Step05: 样本的筛选

从主对话框的候选变量列表框中选择一个变量，将其移至【Selection Variable (选择变量)】列表框中，这表示要按照这个变量的标准来筛选样本进行回归分析。具体操作可以在Rule窗口中实现。

### Step06: 选择个案标签

从候选变量列表框中选择一个变量进入【Case Labels (个案标签)】列表框中，它的取值将作为每条记录的标签。这表示在指定作图时，以哪个变量作为各样本数据点的标志变量。

### Step07: 选择加权二乘法变量

从候选变量列表框中选择一个变量进入【WLS Weigh (WLS权重)】列表框中，表示选入权重变量进行权重最小二乘法的回归分析。

Step08: 单击【OK】按钮，结束操作，SPSS软件自动输出结果。

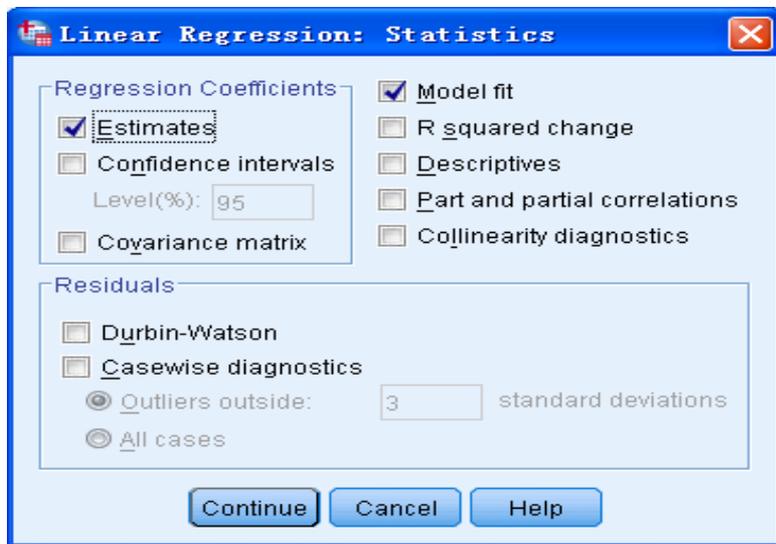


## 8.1.2 一元线性回归的SPSS操作详解

执行完上述操作后，可以输出一元线性回归的基本结果报告了。但是线性回归主对话框中还包括了其他功能选项。下面列出了它们的具体使用功能。

(1) **【Statistics (统计量)】**：选择输出需要的描述统计量，如图8-2所示。

其中，**【Regression Coefficients (回归系数)】**复选框组用于定义回归系数的输出情况，**【Residuals (残差)】**复选框组用于选择输出残差诊断的信息。



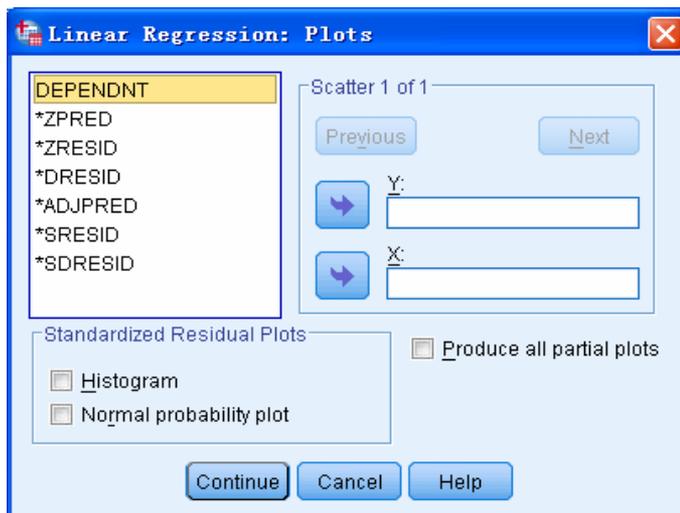
## 8.1.2 一元线性回归的SPSS操作详解

CONCEPT  
STRATE

- Estimates: 可输出回归系数B及其标准误, 回归系数的t检验值和概率p值, 还有标准化的回归系数beta。
- Confidence intervals: 每个回归系数的95%置信区间。
- Covariance matrix: 方差-协方差矩阵。
- Model fit: 模型拟合过程中进入、退出的变量的列表; 以及一些有关拟合优度的检验统计量, 例如R、R<sup>2</sup>和调整的R<sup>2</sup>、估计值的标准误及方差分析表。
- R squared change: 显示每个自变量进入方程后R<sup>2</sup>、F值和p值的改变情况。
- Descriptives: 显示自变量和因变量的有效数目、均值、标准差等, 同时还给出一个自变量间的相关系数矩阵。
- Part and partial correlations: 显示自变量间的相关、部分相关和偏相关系数。
- Collinearity diagnostics: 多重共线性分析, 输出各个自变量的特征根、方差膨胀因子、容忍度等。
- Durbin-Watson: 残差序列相关性检验。
- Casewise diagnostic: 对标准化残差进行诊断, 判断有无奇异值(Outliers)。

## 8.1.2 一元线性回归的SPSS操作详解

(2) 【Plots（绘制）】：用于选择需要绘制的回归分析诊断或预测图。



## 8.1.2 一元线性回归的SPSS操作详解

CONCEPT  
STRATE

用户可以根据上图从中选择部分变量作为X（横坐标）和Y（纵坐标）。  
同时还可以通过单击Next按钮来重复操作过程。绘制更多的图形。

- DEPENDENT: 因变量。
- \*ZPRED: 标准化预测值。
- \*ZRESID: 标准化残差。
- \*DRESID: 剔除的残差。
- ADJPRED: 调整后的预测值。
- SRESID: 学生化残差。
- SDRESID: 学生化剔除残差。

## 8.1.2 一元线性回归的SPSS操作详解

CONCEPT  
STRATE

选择【Standardized Residual Plots (标准化残差图)】选项, 可以选择输出标准化残差图, 其中包括以下选项。

- Histogram: 标准化残差的直方图。
- Normal probability plot: 标准化残差的正态概率图(P-P 图), 将标准化残差与正态分布进行比较。
- Produce all partial plots: 每一个自变量对于因变量残差的散点图。

(3) 【Save(保存)】: 将预测值、残差或其他诊断结果值作为新变量保存于当前工作文件或新文件。

【Predicted Values (预测值)】为预测栏, 用于选择输出回归模型的预测值。

- Unstandardized: 未标准化的预测值。
- Standardized: 标准化的预测值。
- Adjusted: 经调整的预测值。
- S. E. of mean predictions: 预测值的标准误差。

## 8.1.2 一元线性回归的SPSS操作详解

CONCEPT  
STRATE

【Residuals（残差）】为残差栏，包含以下选项。

- Unstandardized: 未标准化残差。
- Standardized: 标准化残差。
- Studentized: 学生化残差。
- Deleted: 剔除残差。
- Studentized Deleted: 学生化剔除残差。

【Distances（距离）】为距离栏，包含以下选项。

- Mahalanobis: 马氏距离。
- Cook's: 库克距离。
- Leverage values: 杠杆值。

## 8.1.2 一元线性回归的SPSS操作详解

CONCEPT  
STRATE

【Influence Statistics（影响统计量）】反映剔除了某个自变量后**回归系数**的变化情况。

- DfBeta(s)：由排除一个特定的观测值所引起的回归系数的变化。
- Standardized Dfbeta(s)：标准化的DfBeta值。
- DfFit：拟合值之差，由排除一个特定的观测值所引起的预测值的变化。
- Standardized DfFit：标准化的DfFit值。
- Covariance ratio：带有一个特定的剔除观测值的协方差（）阵与带有全部观测量的协方差矩阵的比率。

【Prediction intervals（预测区间）】为预测区间栏。

- Mean：均值预测区间的上下限。
- Individual：因变量单个观测量的预测区间。
- Confidence interval（置信区间）：默认值为95%，所键入的值必须在0~100之间。

## 8.1.2 一元线性回归的SPSS操作详解

CONCEPT  
STRATE

- (4) 【Options (选项)】：改变用于进行逐步回归 (Stepwise methods) 时的内部数值的设定以及对缺失值的处理方式。
- 【Stepping Method Criteria (步进方法标准)】为逐步回归标准选择项。
  - Use probability of F: 如果一个变量的F显著性水平值小于所设定的进入值 (Entry value)，那么这个变量将会被选入方程式中；如果它的F显著性水平值大于所设定的剔除值 (Removal value)，那么这个变量将会被剔除。
  - Use F value: 如果一个变量的F值大于所设定的进入值 (Entry value)，那么这个变量将会被选入方程式中；如果它的F值小于剔除值，那么那么这个变量将会被剔除。
  - Include constant in equation: 选择此项表示在回归方程式中包含常数项。
  - 【Missing value treatments (缺失值)】为缺失值处理方式选择项。
  - Exclude cases listwise: 系统默认项，表示剔除所有含缺失值的个案后再进行分析。
  - Exclude cases pariwise: 剔除当前分析的两个变量值是缺失的个案。
  - Replace with mean: 利用变量的平均数代替缺失值。

## 8.1.2 一元线性回归的SPSS操作详解

CONCEPT  
STRATE

- (5) **【Bootstrap】**：可以进行如下统计量的Bootstrap估计。
- 描述统计表支持均值和标准差的Bootstrap 估计。
  - 相关性表支持相关性的Bootstrap 估计。
  - 模型概要表支持Durbin-Watson 的Bootstrap 估计。
  - 系数表支持系数、B 的Bootstrap 估计和显著性检验。
  - 相关系数表支持相关性的Bootstrap 估计。
  - 残差统计表支持均值和标准差的Bootstrap 估计。

# 8.1 SPSS在一元线性回归分析中的应用

CONCEPT  
STRATE

## 8.1.3 实例分析：广告支出与销售量

### 1. 实例内容

表8-1中的数据是7大名牌饮料的广告支出（百万美元）与箱销售量（百万）的数据。请利用回归分析来分析广告支出与箱销售量的关系。

表 8-1 广告支出与箱销售量

品牌	广告支出（百万美元）	箱销售量（百万）
Coca-Cola Classic	131.3	1929.2
Pepsi-Cola	92.4	1384.6
Diet Coke	60.4	811.4
Sprite	55.7	541.5
Dr. Pepper	40.2	536.9
Mountain Dew	29.0	535.6
7-UP	11.6	219.5

# 8.1 SPSS在一元线性回归分析中的应用

CONCEPT  
TRATE

## 2. 实例操作

现在厂商要研究投入的广告支出与箱销售量之间的关系，则可以建立回归模型来探讨它们之间的关系，即

箱销售量=f（广告支出）

首先绘制了这两组变量的散点图8-6，图形显示它们呈线性关系，则可以建立一元线性回归模型如下：

$$sale_i = \beta_0 + \beta_1 \times expenditure_i + \varepsilon_i$$

## 8.1

# SPSS在一元线性回归分析中的应用

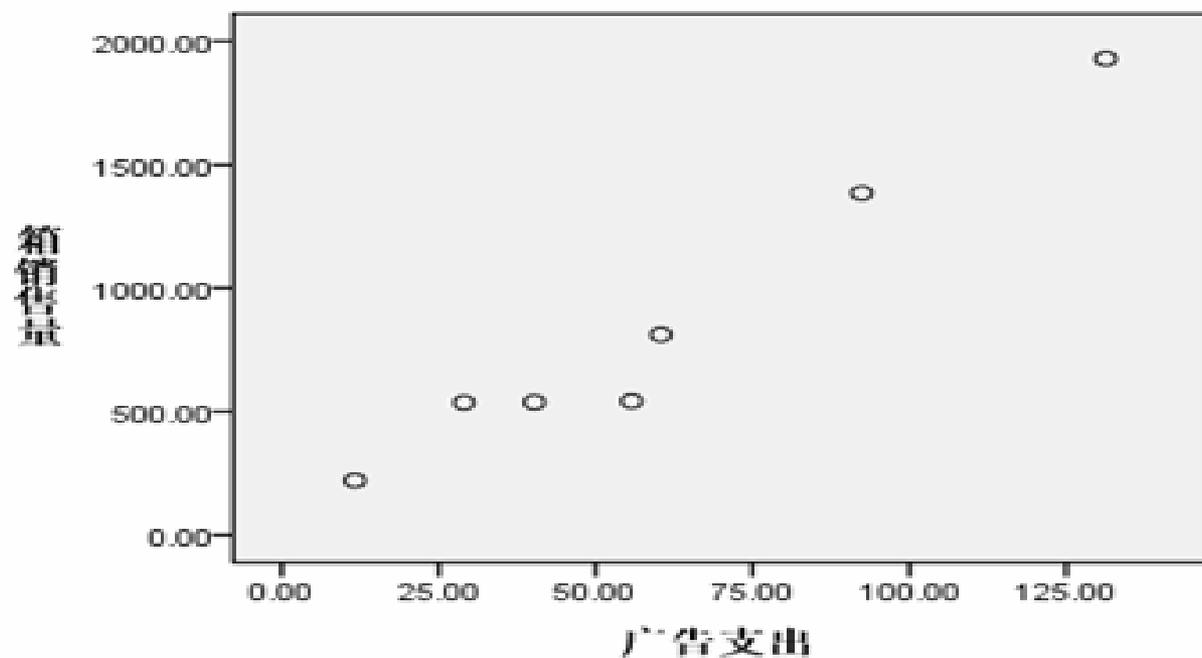


图 8-6 广告支出与箱销售量的散点图

# 8.1 SPSS在一元线性回归分析中的应用

CONCEPT  
STRATE

## 3. 实例结果及分析

### (1) 自变量进入方式

执行完上面的操作后，首先给出的是自变量进入方式表8-2。可以看到回归模型的选入变量是广告支出（expenditure），采用的自变量进入方式是强行进入法，也就是将所有的自变量都放入模型中。

表 8-2 自变量进入方式

Model	Variables Entered	Variables Removed	Method
1	广告支出		Enter

- a. 所有变量均进入方程
- b. 因变量: 箱销售量

## 8.1 SPSS在一元线性回归分析中的应用

### (2) 模型摘要

表8-3是对模型的简单汇总，其实就是对方程拟合情况的描述。通过这张表可以知道相关系数的取值（R），相关系数的平方即可决系数（R Square），校正后的可决系数（adjusted R Square）和回归系数的标准误（Std. Error of the Estimate）。注意这里的相关系数大小和前面相关分析中计算出的结果完全相同。可决系数R Square的取值介于0和1之间，它的含义就是自变量所能解释的方差在总方差中所占的百分比，取值越大说明模型的效果越好。本案例计算的回归模型中可决系数R<sup>2</sup>等于0.957，模型拟合效果较好。

表 8-3 模型摘要

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.978 <sup>a</sup>	.957	.948	136.21405 <sup>a</sup>

a. Predictors: (Constant), 广告支出

## 8.2 SPSS 在多元线性回归分析中的应用

CONCEPT  
STRATE

### 8.2.1 多元线性回归的基本原理

#### 1. 方法概述

在回归分析中，如果有两个或两个以上的自变量，就称为多元回归。

#### 2. 基本原理

多元线性回归模型是指有多个自变量的线性回归模型，它用于揭示因变量与多个自变量之间的线性关系。多元线性回归方程的经验模型是：

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \cdots + \hat{\beta}_k x_k$$

$$x_1, \cdots, x_k \quad \hat{\beta}_i (i = 1, \cdots, k)$$

上式中，假设该线性方程有k个自变量。  
是回归方程的**偏回归系数**。 $\hat{\beta}_i$ 表示在其他自变量保持不变的情况下，自变量  $x_i$  变动一个单位所引起的因变量的平均变动单位。



## 8.2 SPSS在多元线性回归分析中的应用

### 8.2.2 多元线性回归的SPSS操作详解

由于多元线性回归模型是一元回归模型的推广，因此两者在SPSS软件中的操作步骤是非常相似的。选择菜单栏中的【Analyze（分析）】→【Regression（回归）】→【Linear（线性）】命令，弹出【Linear Regression（线性回归）】对话框。这既是一元线性回归也是多元线性回归的主操作窗口。因此，读者可以参考8.1.2节的操作步骤。只不过由于多元回归模型涉及到多个自变量，因此在图8-1中要在【Linear Regression（线性回归）】对话框左侧的候选变量列表框中选择多个变量，将其添加至【Independent(s)（自变量）】列表框中，即选择这些变量作为多元线性回归的自变量。



## 8.2 SPSS在多元线性回归分析中的应用

### 8.2.3 实例分析：电视广告和报纸广告

#### 1. 实例内容

娱乐时光影剧院公司的老板希望了解公司投放的电视广告费用和报纸广告费用对公司收入的影响。以往8周的样本数据如表8-6所示（单位：千美元）。请建立模型分析这两种广告形式对公司营业收入的影响。

表 8-6 费用和收入

每周营业总收入	96	90	95	92	95	94	94	94
电视广告费用	5.0	2.0	4.0	2.5	3.0	3.5	2.5	3.0
报纸广告费用	1.5	2.0	1.5	2.5	3.3	2.3	4.2	2.5

↙

## 8.2 SPSS在多元线性回归分析中的应用

CONCEPT  
RATE

### 2. 实例操作

本案例要分析电视广告和报纸广告对公司收入的影响，则可以建立二元回归模型来探讨它们之间的关系，即

每周营业总收入=f（电视广告费用，报纸广告费用）

可以通过比较电视广告和报纸广告变量的系数大小来研究这两种广告形式对收入的影响程度高低。但是，是否收入和广告费用呈线性关系，则首先要绘制散点图来判断。通过三维散点图8-9看到，这三个变量之间呈明显的线性增长关系，因此可以建立营业收入的二元影响回归模型如下：



## 8.2 SPSS在多元线性回归分析中的应用

$$\text{income}_i = \beta_0 + \beta_1 \times \text{tv}_i + \beta_2 \times \text{newspaper}_i + \varepsilon_i$$

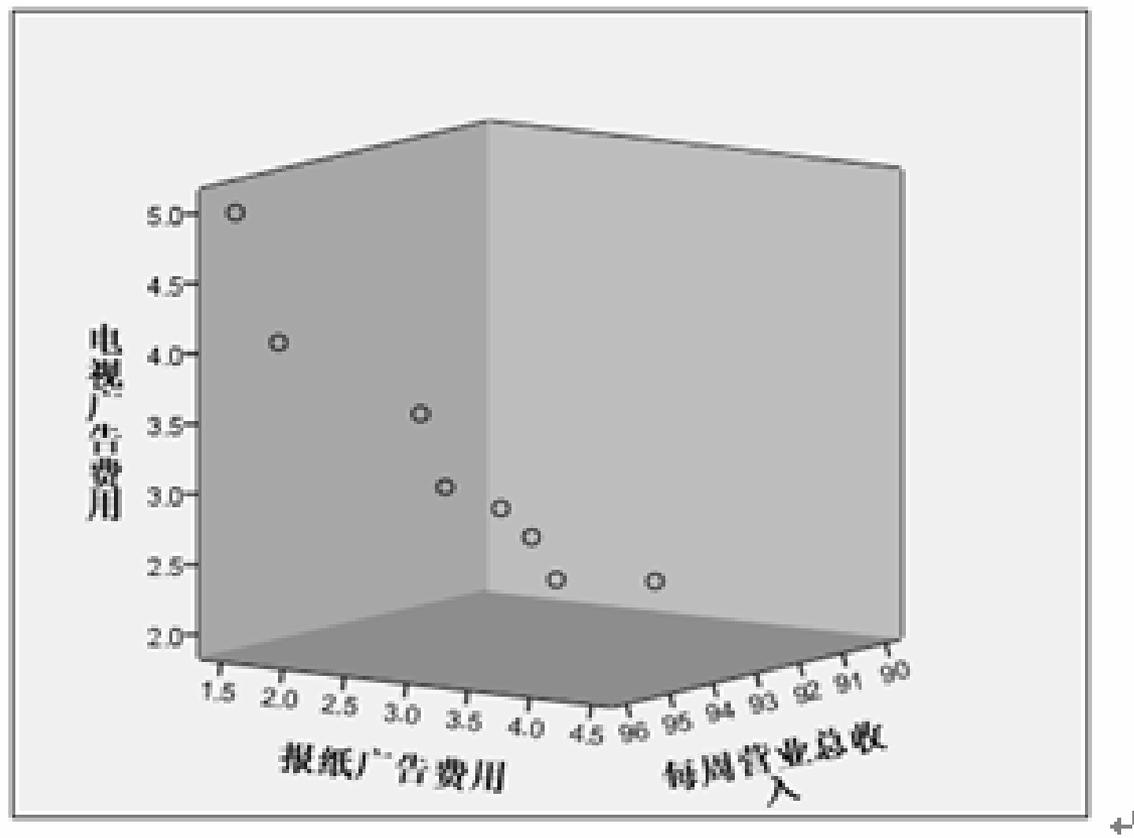


图 8-9 散点图



## 8.2 SPSS在多元线性回归分析中的应用

### 3. 实例结果及分析

#### (1) 自变量进入方式

执行完上面的操作后，首先给出的是自变量进入方式表8-7。由于这里的自变量进入方式采用的是系统默认，即强行进入法，可以看到回归模型的选入变量是报纸广告费用和电视广告费用。

表 8-7 自变量进入方式

Model	Variables Entered	Variables Removed	Method
1	报纸广告费用, 电视广告费用 <sup>a</sup>		. Enter

a. All requested variables entered



## 8.2 SPSS在多元线性回归分析中的应用

### (2) 模型摘要

表8-8给出了衡量该回归方程优劣的统计量。R为复相关系数，它表示模型中所有自变量（tv、newspaper）与因变量income之间的线性回归关系的密切程度大小。它的取值介于0和1之间；R越大说明线性回归关系越密切。可决系数R<sup>2</sup>等于复相关系数的平方，这里等于0.919。调整的R<sup>2</sup>为我们要重点关注的统计量；它的值越大，模型拟合效果得越好；表8-8中调整的R<sup>2</sup>为0.887。最后给出的是剩余标准差（Std. Error of the Estimate），它是残差的标准差，其大小反映了建立的模型预测因变量的精度。剩余标准差越小，说明建立的模型效果越好。

表 8-8 模型摘要

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.959	.919	.887	.643



## 8.2 SPSS在多元线性回归分析中的应用

### (3) 方差分析表

表8-9是对回归模型进行方差分析的检验结果。可以看到方差分析结果中F统计量等于28.378，概率P值0.002小于显著性水平0.05，所以该模型是有统计学意义的，即两种广告支出费用和每周营业收入之间的线性关系是显著的。

表 8-9 方差分析表

Model	Sum of Squares	df	Mean Square	F	Sig.	
1	Regression	23.435	2	11.718	28.378	.002
	Residual	2.065	5	.413		
	Total	25.500	7			

## 8.2 SPSS在多元线性回归分析中的应用

CONCEPT  
STRATE

### (4) 回归系数表

表8-10给出了回归模型的常数项 (Constant)、电视广告费用和报纸广告费用的偏相关系数，它们分别等于83.230、2.290和1.301。于是得到回归方程如下：

每周营业总收入 =  $83.230 + 2.290 \times \text{电视广告费用} + 1.301 \times \text{报纸广告费用}$

其中常数项表示当自变量取值全为0时，因变量的取值大小，即没有这两种广告投入时电影院的营业收入。同时比较电视广告和报纸广告系数看到，电视广告对电影院的收入影响要大于报纸广告的影响。

表8-10还给出了模型对tv和income变量的偏回归系数是否等于0的t检验结果。t值分别等于7.532和4.057，概率P值都小于显著性水平0.05，因此认为偏相关系数  $\beta_1$ 、 $\beta_2$  显著不等于0。同时，SPSS在输出一组偏回归系数的同时，也输出了各自的标准化偏回归系数 (Standardized Coefficients)。

## 8.2 SPSS在多元线性回归分析中的应用

表 8-10 回归系数表

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	83.230	1.574		52.882	.000
	电视广告费用	2.290	.304	1.153	7.532	.001
	报纸广告费用	1.301	.321	.621	4.057	.010

## 8.3 SPSS在曲线拟合中的应用

CONCEPT  
STRATE

### 8.3.1 曲线拟合的基本原理

#### 1. 方法概述

实际中，变量之间的关系往往不是简单的线性关系，而呈现为某种**曲线或非线性**的关系。此时，就要选择相应的曲线去反映实际变量的变动情况。为了决定选择的曲线类型，常用的方法是根据数据资料绘制出**散点图**，通过图形的变化趋势特征并结合专业知识和经验分析来确定曲线的类型，即**变量之间的函数关系**。

在确定了变量间的函数关系后，需要估计函数关系中的未知参数，并对拟合效果进行显著性检验。虽然这里选择的是曲线方程，在方程形式上是非线性的，但可以**采用变量变换的方法**将这些曲线方程转化为线性方程来估计参数。



## 8.3 SPSS在曲线拟合中的应用

### 2、常用曲线估计模型

SPSS的【Curve Estimation（曲线估计）】选项就是用来解决上述问题的。它提供了11种常用的曲线估计回归模型。

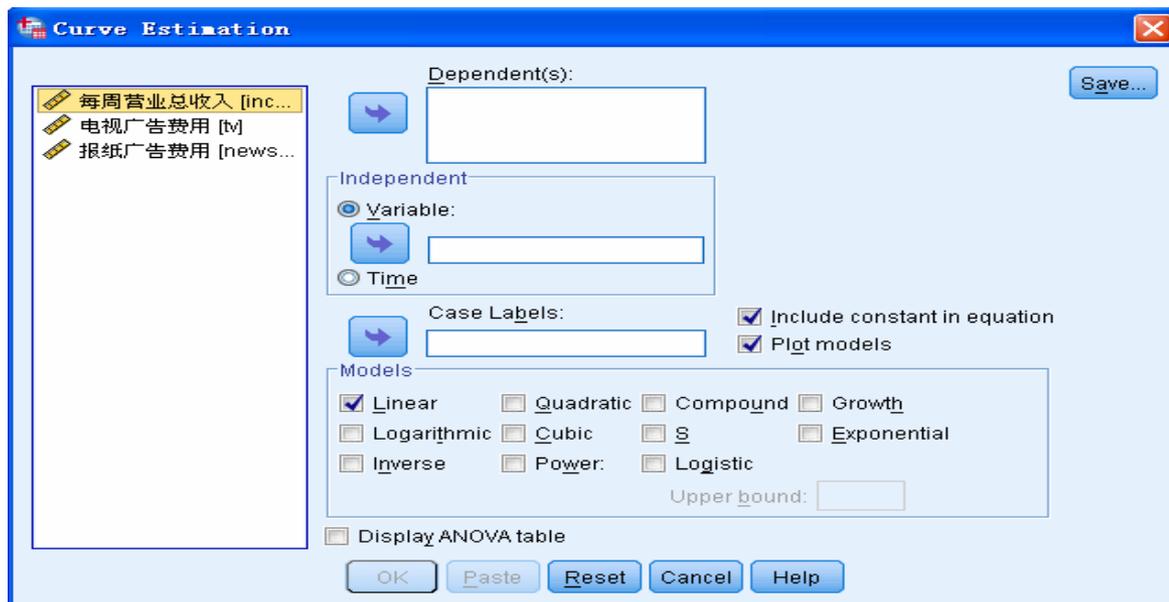


## 8.3 SPSS在曲线拟合中的应用

### 8.3.2 曲线拟合的SPSS操作详解

Step01: 打开对话框

选择菜单栏中的【Analyze（分析）】→【Regression（回归）】→【Curve Estimation（曲线估计）】命令，弹出【Curve Estimation（曲线估计）】对话框，这是曲线拟合的主操作窗口。





## 8.3 SPSS在曲线拟合中的应用

### Step02: 选择因变量

在【Curve Estimation (曲线估计)】对话框左侧的候选变量列表框中选择一个变量，将其添加至【Dependent(s) (因变量)】列表框中，即选择该变量作为曲线估计的因变量。

### Step03: 选择自变量

在【Curve Estimation (曲线估计)】对话框左侧的候选变量列表框中选择一个数值型变量，将其添加至【Independent (自变量)】栏中的【Variable (变量)】列表框中，即选择该变量作为曲线估计的自变量。如果自变量是时间变量或序列ID，可以选择它移入【Time (时间)】框中，此时自变量之间的长度是均匀的。

## 8.3 SPSS在曲线拟合中的应用

CONCEPT  
STRATE

### Step04: 选择个案标签

从候选变量列表框中选择一个变量进入【Case Labels (个案标签)】列表框中，它的取值将作为每条记录的标签。这表示在指定作图时，以哪个变量作为各样本数据点的标志变量。

### Step05: 选择曲线拟合模型

在【Models (模型)】复选框中共有11种候选曲线模型可以选择，用户可以选择多种候选模型进行拟合优度比较。

### Step06: 选择预测值和残差输出

单击【Save】按钮，弹出对话框。



## 8.3 SPSS在曲线拟合中的应用

【Save Variables (保存变量)】选项组中的选项是将预测值、残差或其他诊断结果值作为新变量保存于当前工作文件中。

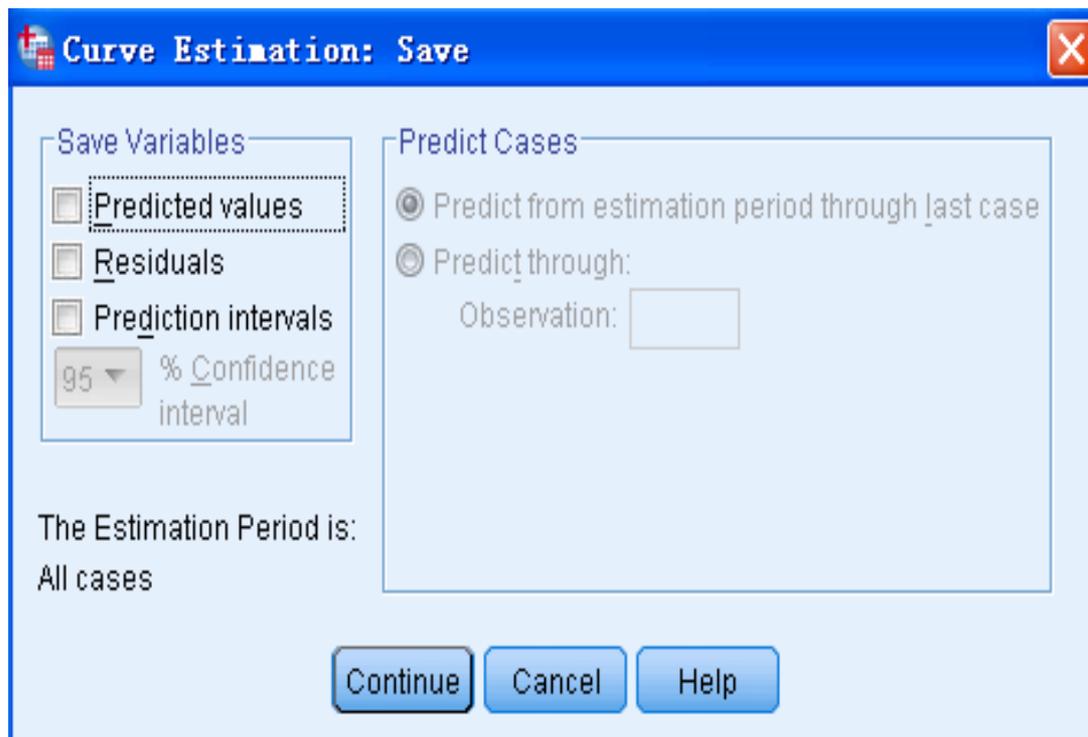
- Predicted Values: 输出回归模型的预测值。
- Residuals: 输出回归模型的残差。
- Predicted Intervals: 预测区间的上下限。
- Confidence Interval: 选择预测区间的置信概率。

【Predict Case (预测个案)】选项组是以时间序列为自变量时的预测值输出。

- Predict from estimation period through last case: 计算样本中数据的预测值。
- Predict through: 预测时间序列中最后一个观测值之后的值。选择该项后, 在下面的【Observation (观测值)】文本框中指定一个预测周期。



## 8.3 SPSS在曲线拟合中的应用



## 8.3 SPSS在曲线拟合中的应用

CONCEPT  
STRATE

Step07: 其他选项输出

在图中还有三个选项可供选择, 用户可根据自己的需要勾选这些选项。

- Display ANOVA Table: 结果中显示方差分析表。
- Include constant in equation: 系统默认值, 曲线方程中包含常数项。
- Plot models: 系统默认值; 绘制曲线拟合图。

Step08: 单击【OK】按钮, 结束操作, SPSS软件自动输出结果。



## 8.3 SPSS在曲线拟合中的应用

### 8.3.3 实例分析：空置率和租金率

#### 1. 实例内容

某管理咨询公司采集了市场上办公用房的空置率和租金率的数据。对于13个选取的销售地区，表8-13是这些地区的中心商业区的综合空置率（%）和平均租金率（元/平方米）的统计数据。请尝试分析空置率对平均租金率的影响。



## 8.3 SPSS在曲线拟合中的应用

### 2. 实例操作

本案例要分析空置率对平均租金率的影响，因此首先绘制它们之间的散点图8-18。从图形看到，随着空置率的增加，平均租金率呈显著的下降趋势。但是这种下降趋势并不是线性的，而表现为非线性的关系。故可以考虑采用曲线拟合的方法。

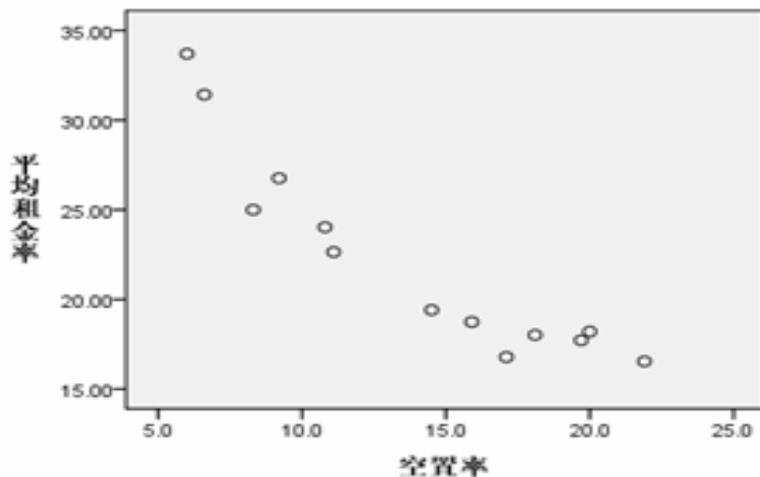


图 8-18 散点图



## 8.3 SPSS在曲线拟合中的应用

### 3. 实例结果及分析

#### (1) 模型描述

表8-14是SPSS对曲线拟合结果的初步描述统计，例如自变量和因变量、估计方程的类型等。

表 8-14 模型描述

Model Name		MOD_3
Dependent Variable	1	平均租金率
Equation	1	Linear
	2	Inverse
	3	Exponential
Independent Variable		空置率
Constant		Included
Variable Whose Values Label Observations in Plots		Unspecified



## 8.3 SPSS在曲线拟合中的应用

### (2) 模型汇总及参数估计

表8-15给出了样本数据分别进行三种曲线方程拟合的检验统计量和相应方程中的参数估计值。

对于直线拟合，它的可决系数 $R^2$ 为0.858，F统计量等于66.335，概率P值小于显著性水平0.05，说明该模型有统计学意义；并且直线拟合方程为：

$$z_j = 35.536 - 0.966 \times k z_j$$

对于逆函数方程和指数方程拟合来说，它对应的可决系数 $R^2$ 分别为0.972和0.900，模型也显著有效；具体估计方程分别为：

$$z_j = 10.208 + 139.250 / k z_j$$

$$z_j = 38.484 \times e^{-0.042 \times k z_j}$$

虽然上述模型都有显著的统计学意义，但从可决系数的大小可以清晰看到逆函数方程较其他两种曲线方程拟合效果更好，因此选择逆函数方程来描述空置率和租金率的关系。



## 8.3 SPSS在曲线拟合中的应用

表 8-15 模型汇总及参数估计

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Linear	.858	66.335	1	11	.000	35.536	-.966
Inverse	.972	378.015	1	11	.000	10.208	139.250
Exponential	.900	98.487	1	11	.000	38.484	-.042

□



## 8.3 SPSS在曲线拟合中的应用

### (3) 拟合曲线图

最后给出的是实际数据的散点图和三种估计曲线方程的预测图。从图8-22也进一步说明逆函数曲线方程的拟合效果最好。

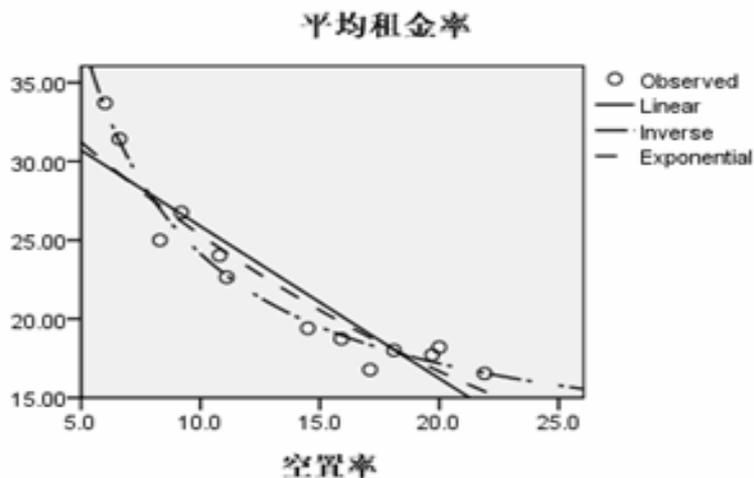


图 8-22 拟合曲线图

## 8.4 SPSS 在非线性回归分析中的应用

CONCEPT  
STRATE

### 8.4.1 非线性回归分析的基本原理

非线性回归分析是探讨因变量和一组自变量之间的非线性相关模型的统计方法。线性回归模型要求变量之间必须是线性关系，**曲线估计只能处理能够通过变量变换化为线性关系的非线性问题**，因此这些方法都有一定的局限性。相反的，非线性回归可以**估计因变量和自变量之间具有任意关系的模型**，用户根据自身需要可随意设定估计方程的具体形式。因此，本方法在实际应用中有很大的实用价值。

## 8.4 SPSS 在非线性回归分析中的应用

CONCEPT  
STRATE

非线性回归模型一般可以表示为如下形式：

$$y_i = \hat{y}_i + e_i = f(x, \theta) + e_i$$

其中  $f(x, \theta)$  为期望函数，该模型的结构和线性回归模型非常相似，所不同的是期望函数可能为任意形式，甚至在有的情况下没有显式关系式， $f(x, \theta)$  方程中参数的估计是通过迭代方法获得的。

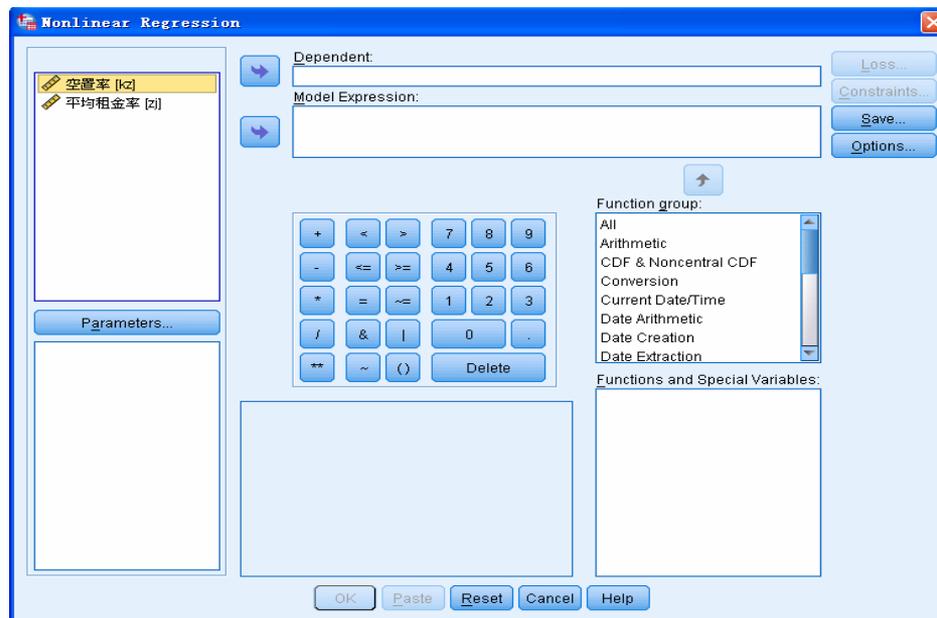
## 8.4 SPSS 在非线性回归分析中的应用

CONCEPT  
STRATE

### 8.4.2 非线性回归分析的SPSS操作详解

Step01: 打开对话框

选择菜单栏中的【Analyze（分析）】→【Regression（回归）】→【Nonlinear（非线性）】命令，弹出【Nonlinear Regression（非线性回归）】对话框，这是非线性回归的主操作窗口。



## 8.4 SPSS 在非线性回归分析中的应用

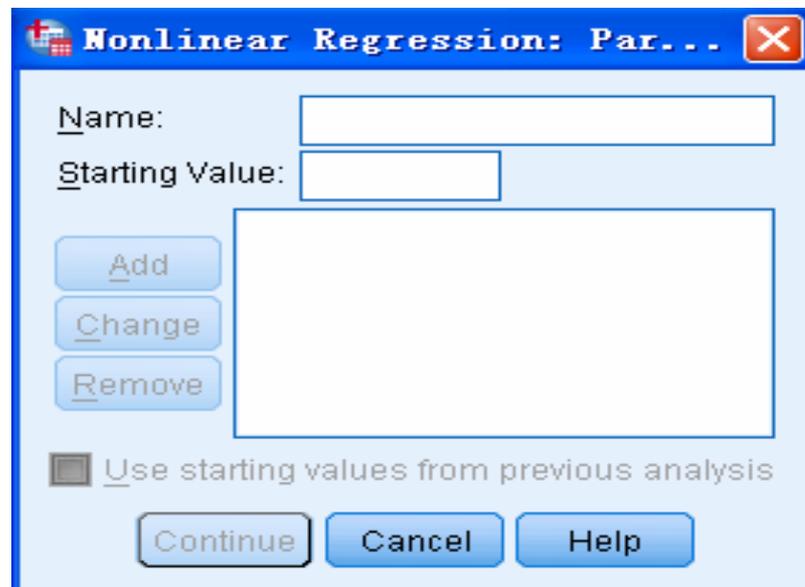
CONCEPT  
STRATE

### Step02: 选择因变量

在【Nonlinear Regression (非线性回归)】对话框左侧的候选变量列表框中选择一个变量，将其添加至【Dependent (自变量)】列表框中，即选择该变量作为非线性回归分析的因变量。

### Step03: 设置参数变量和初始值

单击【Parameters (参数)】按钮，将打开如下图所示的对话框，该对话框用于设置参数的初始值。



## 8.4 SPSS 在非线性回归分析中的应用

CONCEPT  
STRATE

- **【Name（名称）】** 文本框：用于输入参数名称。
- **【Starting Value（初始值）】** 文本框：用于输入参数的初始值。

当输入完参数名和初始值后，单击**【Add】**按钮，则定义的变量及其初始值将显示在下方的参数框中，参数的初始值可根据给定模型中参数定义范围情况而定。如果需要修改已经定义了的参数变量，则先将其选中，然后在**【Name（名称）】**和**【Starting Value（初始值）】**文本框里进行修改，完成后单击**【Change】**按钮确认修改。如果要删除已经定义了的参数变量，先用将其选中，然后单击**【Remove】**按钮删除。如果勾选**【Use starting values from previous analysis（使用上一分析的起始值）】**复选框，表示使用前一次分析确定的初始值；当算法的收敛速度减慢时，可选择它继续进行搜索。完成后单击**【Continue】**按钮返回主程序窗口。

## 8.4 SPSS 在非线性回归分析中的应用

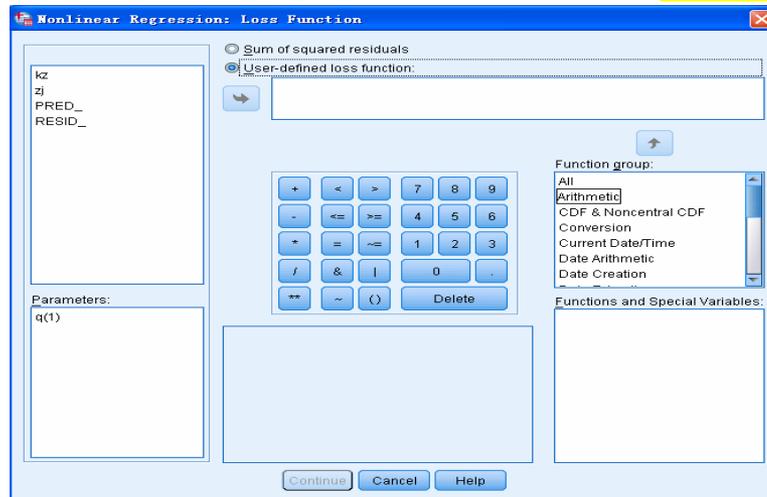
CONCEPT  
STRATE

### Step04: 输入回归方程

在【Model Expression (模型表达式)】文本框中输入需要拟合的方程式，该方程中包含自变量、参数变量和常数等。自变量从左侧的候选变量列表框中选择，参数变量从左侧的【Parameters (参数)】列表框里选入。同时，拟合方程模型中的函数可以从【Function (函数组)】列表框里选入；方程模型的运算符号可以用鼠标从窗口“数字符号”显示区中点击输入。

### Step05: 迭代条件选择

单击【Loss】按钮，将打开如下图所示的对话框。该对话框用来选择损失函数来确定参数的迭代算法。



## 8.4 SPSS 在非线性回归分析中的应用

CONCEPT  
STRATE

- Sum of squared residuals: 系统默认项，基于残差平方和最小化的迭代算法。
- User-defined loss function: 自定义选项，设置其他统计量为迭代条件。在下面文本输入框中输入相应的统计量的表达式，这里称为损失函数。

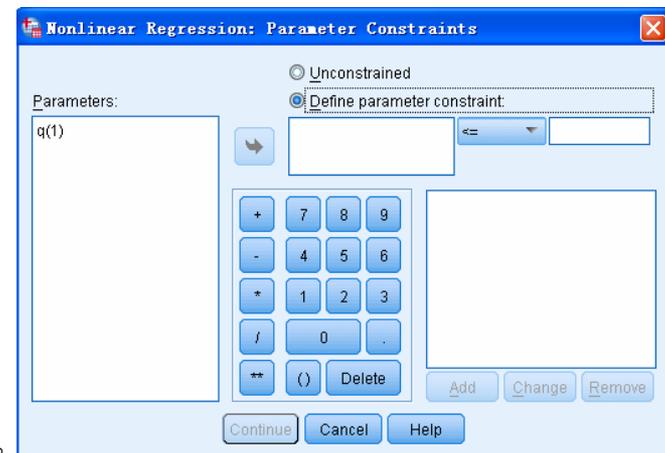
左侧的候选变量列表框中，“RESID\_”代表所选变量的残差；“PRED\_”代表预测值。可以从左下角的【Parameters（参数）】列表框中选择已定义的参数进入损失函数。

## 8.4 SPSS 在非线性回归分析中的应用

CONCEPT  
TRATE

### Step06: 参数取值范围选择

单击【Constraints】按钮，将打开如下图所示的对话框。该对话框用来设置回归方程中参数的取值范围。



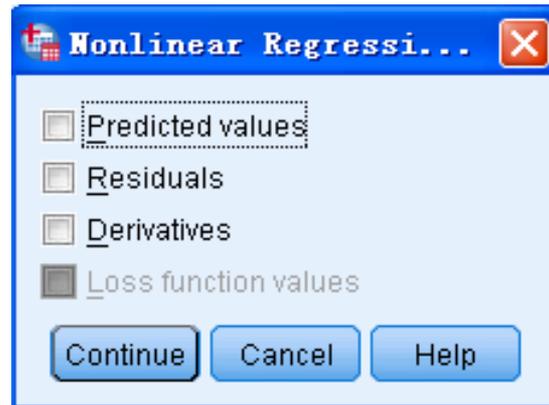
- Unconstrained: 无约束条件，系统默认项。
- Define parameter constraint: 可对选定的参数变量设置取值范围。参数的取值范围用不等式“=, <=, >=”来定义。例如这里限制参数“b”的迭代范围是“b<=5”。

## 8.4 SPSS 在非线性回归分析中的应用

CONCEPT  
STRATE

Step07: 选择预测值和残差等输出

单击【Save】按钮，弹出如下图所示的对话框。它表示要保存到数据文件中的统计量。



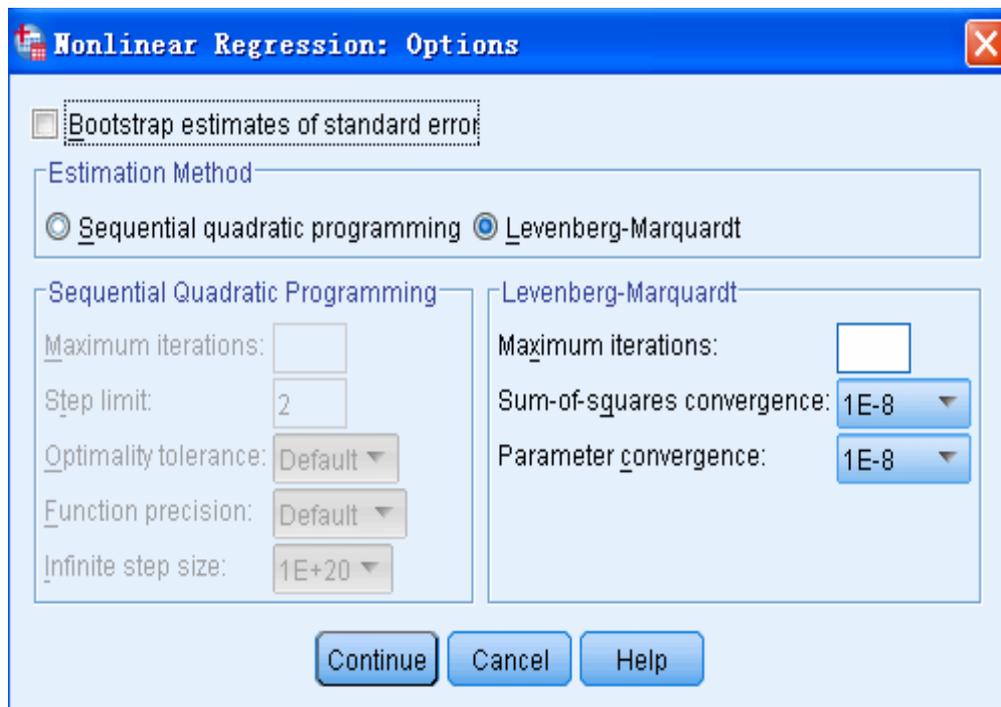
- Predicted Values: 输出回归模型的预测值。
- Residuals: 输出回归模型的残差。
- Derivatives: 模型各个参数的一阶导数值。
- Loss function values: 损失函数值。



## 8.4 SPSS 在非线性回归分析中的应用

Step08: 迭代方法选择

单击【Options】按钮，弹出如下图所示的对话框。它用于选择各类迭代算法。





## 8.4 SPSS 在非线性回归分析中的应用

Bootstrap estimates of standard error: 采用**样本重复法**计算标准误。样本重复法需要**顺序二次规划算法**的支持。当选中该项时，SPSS将自动选中【Sequential quadratic Programming (序列二次编程)】项。

【Estimation Method】框中列出了参数的两种估计方法：

- Sequential Quadratic Programming: 顺序二次规划算法。该方法要求输入的参数如下。
  - “Maximum”：**最大迭代步数**。
  - “Step limit”：**最大步长**。
  - “Optimality”：目标函数的迭代误差限。
  - “Function precision”：**函数精度**，应比目标函数的迭代误差限小。
  - “Infinite step size”：当一次迭代中参数值的变化大于设置值，则迭代停止。
- Levenberg-Marquardt: 系统缺省设置，**列文博格-麦夸尔迭代法**。该法要求输入的参数如下。
  - “Maximum iterations”：最大迭代步数。
  - “Sum-of-squares convergence”：在一步迭代中目标函数**残差平方和**的变化比例小于设置的值时，迭代停止。
  - “Parameter convergence”：在一步迭代中参数的变化比例**小于**设置值时，迭代停止。

Step09: 单击【OK】按钮，结束操作，SPSS软件自动输出结果。



## 8.4 SPSS 在非线性回归分析中的应用

### 8.4.3 实例分析：股票价格的预测

#### 1. 实例内容

假定数据文件8-4中是三个公司股票在15个月期间的股市收盘价。一家投资公司希望建立一个回归模型用股票B和股票C的价格来预测股票A的价格。请建立回归模型分析。

## 8.4 SPSS在非线性回归分析中的应用

### 2. 实例操作

本案例要利用股票B和股票C的价格来预测股票A的价格，因此选择股票B和股票C为自变量，股票A为因变量来建立回归方程：

$$y = f(x_1, x_2) + \varepsilon$$

其中， $y$ 、 $x_1$ 和 $x_2$ 分别表示股票A、股票B和股票C的价格。

## 8.4 SPSS在非线性回归分析中的应用

CONCEPT  
STRATE

接着利用散点矩阵图来判断三个变量之间的关系。散点矩阵图8-29分为9个子图，它们分别描述了三只股票中两两股票价格之间的变化。可以看到，股票A的价格和其他两只股票的价格都存在显著线性关系，这是否表示只需要建立一个二元线性模型即可呢？观察自变量股票B和股票C之间散点图看到，这两只股票的价格也存在显著的影响关系，这说明了这两个因变量之间可能存在交叉影响。于是，建立如下非线性回归方程：

$$y = a + bx_1 + cx_2 + dx_1x_2 + \varepsilon$$



## 8.4 SPSS在非线性回归分析中的应用

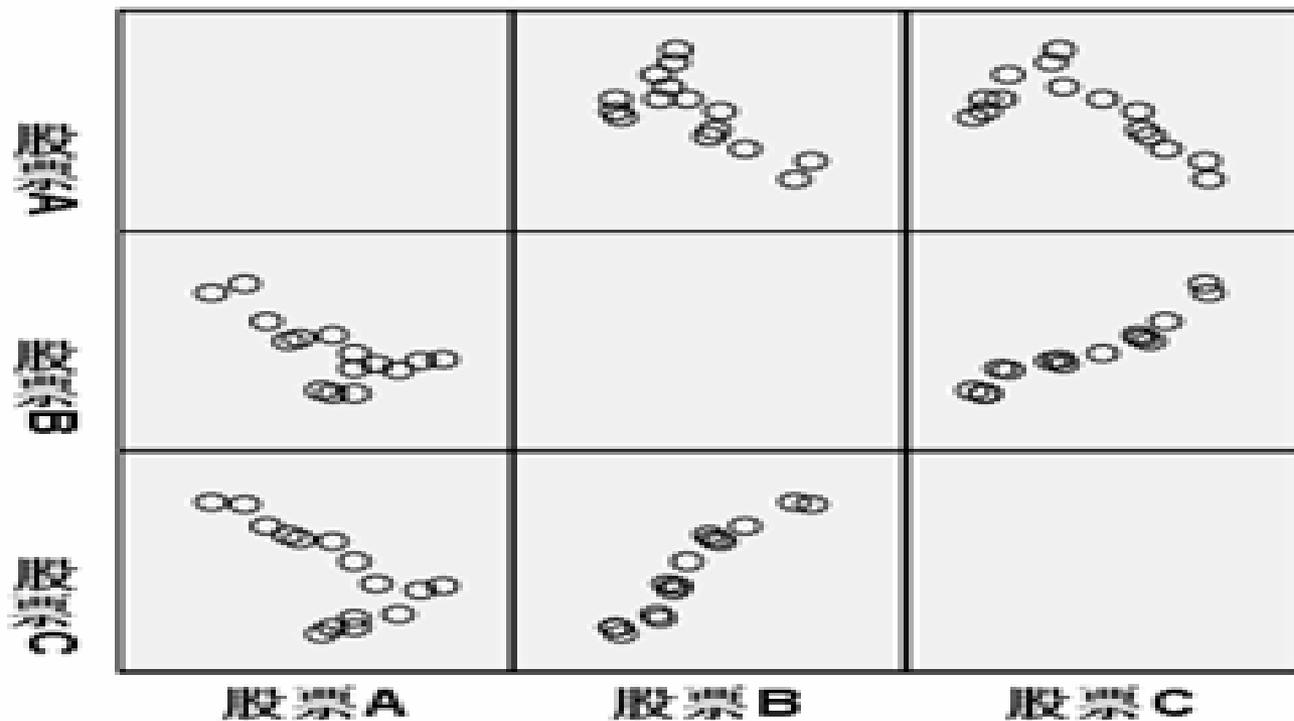


图 8-29 散点矩阵图



## 8.4 SPSS在非线形回归分析中的应用

### 3 实例结果及分析

#### (1) 迭代过程表

表8-17是回归方程参数估计的迭代过程记录。这里只进行了两次迭代就达到了精度要求。观察残差平方和“Residual Sum of Squares”的变化，可见随着迭代的进行，残差变得越来越小。但这一过程不是无限进行下去的，当进行了两步迭代后，残差以及各参数的估计值均稳定下去了，模型达到收敛标准。

表 8-17 迭代过程表

Iteration Number	Residual Sum of Squares	Parameter			
		a	b	c	d
1.0	3.861E8	1.000	1.000	1.000	1.000
1.1	93.087	12.046	.879	.220	-.010
2.0	93.087	12.046	.879	.220	-.010

## 8.4 SPSS在非线性回归分析中的应用

CONCEPT  
STRATE

### (2) 参数估计值

表8-18列出了回归模型中四个参数的迭代估计值、标准误差和95%的置信区间。于是，得到股票A关于股票B和C的预测回归模型为：

$$y = 12.046 + 0.879 \cdot x_1 + 0.220 \cdot x_2 - 0.010 \cdot x_1 x_2 + \epsilon$$

可以看到，股票B和股票C都和股票A的价格变动方向相同，而且股票B对股票A的影响更大。股票B、C的交互项会影响股票A下跌，但这种影响不太明显。



## 8.4 SPSS在非线性回归分析中的应用

表 8-18 参数估计值

Parameter	Estimate	Std. Error	95% Confidence Interval	
			Lower Bound	Upper Bound
a	12.046	9.312	-8.450	32.543
b	.879	.262	.302	1.455
c	.220	.144	-.095	.536
d	-.010	.002	-.015	-.005



## 8.4 SPSS在非线性回归分析中的应用

### (3) 参数的相关系数矩阵

表8-19是模型中四个估计参数的相关系数矩阵。对于较复杂的模型，参数间的相关系数可用来辅助进行模型的改进，本案例无太多价值。

表 8-19 参数的相关系数矩阵

	a	b	c	d
a	1.000	-.881	-.466	.966
b	-.881	1.000	.019	-.883
c	-.466	.019	1.000	-.470
d	.966	-.883	-.470	1.000

## 8.4 SPSS在非线性回归分析中的应用

CONCEPT  
STRATE

### (4) 方差分析表

表8-20是非线性回归分析的方差分析表。Uncorrected Total为未修正的总误差平方和，其值等于23368.000，自由度等于15；它被分解成回归平方和23274.913和残差平方和93.087，自由度分别是4和11。Corrected Total是经修正的总误差平方和，其值等于474.933，自由度是14；表的最后一列是均方。

表8-20最后一行公式： $R^2=1-\text{残差平方和}/\text{修正平方和}=0.804$ ，这个结果说明了这个非线性回归模型的拟合效果，总体来看还是不错的。



## 8.4 SPSS在非线性回归分析中的应用

表 8-20 方差分析表<sup>a</sup>

Source <sup>a</sup>	Sum of Squares	df <sup>a</sup>	Mean Squares <sup>a</sup>
Regression <sup>a</sup>	23274.913	4	5818.728 <sup>a</sup>
Residual <sup>a</sup>	93.087	11	8.462 <sup>a</sup>
Uncorrected Total <sup>a</sup>	23368.000	15 <sup>a</sup>	
Corrected Total <sup>a</sup>	474.933	14 <sup>a</sup>	

a. R squared =  $1 - (\text{Residual Sum of Squares}) / (\text{Corrected Sum of Squares}) = .804$ .



## 8.4 SPSS在非线性回归分析中的应用

### (5) 线性回归和非线性回归的股票预测图

图8-35显示了原始数据、线性回归模型、非线性回归模型三者的比较。其中，“股票A”是实际曲线，“Predicted Values”是本案例建立的非线性回归方程的预测曲线，“Unstandardized Predicted Values”是不考虑股票B、C交互项的二元线性模型的预测曲线。可以明显看到，非线性回归的预测效果要好于二元线性回归的预测效果，说明了这里我们引入股票B、C交互项的合理性。

## 8.4 SPSS在非线性回归分析中的应用

CONCEPT  
RATE

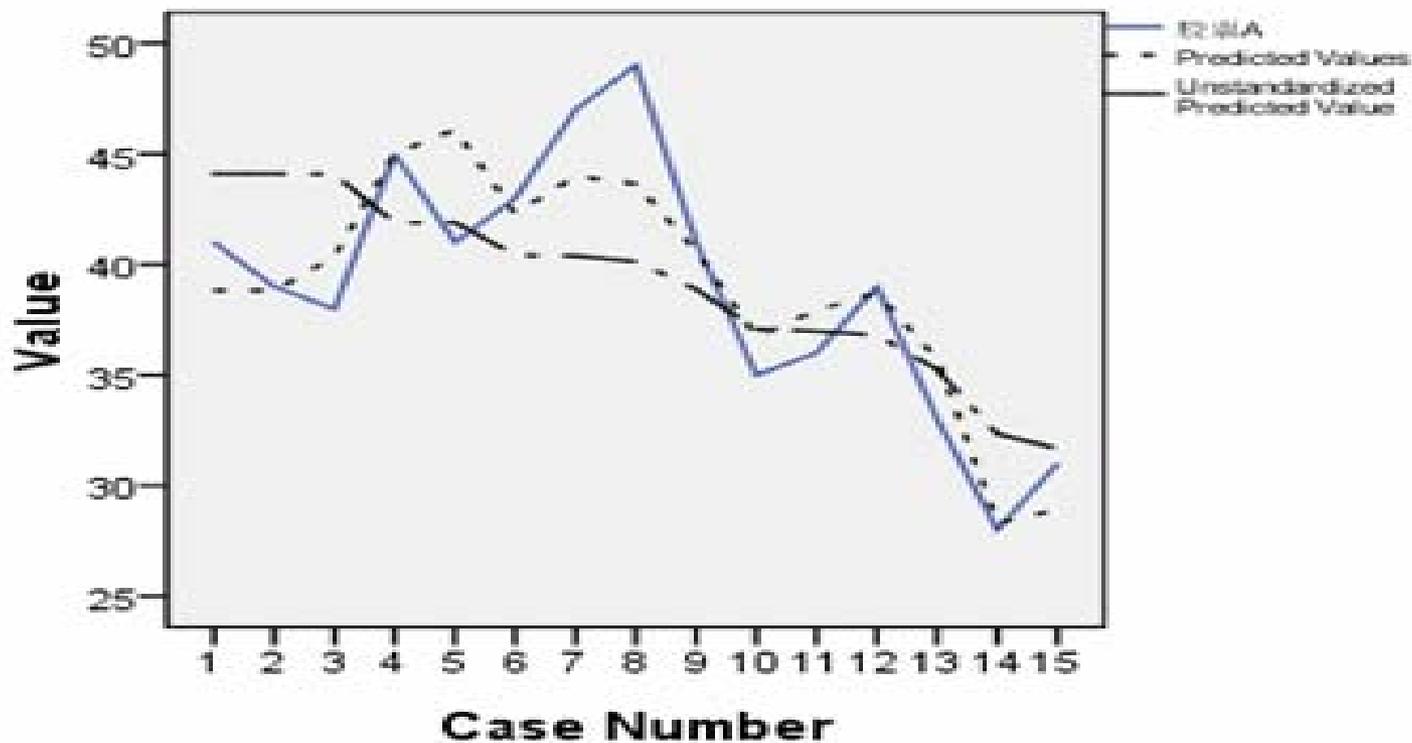


图 8-35 股票预测图+



## 第9章

# SPSS的多元统计 分析

# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

## 9.1.1 因子分析的基本原理

### 1、方法概述

人们在研究实际问题时，往往希望尽可能多的收集相关变量，以期对问题有比较全面、完整的把握和认识。

为解决这些问题，最简单和最直接的解决方案是减少变量数目，但这必然又会导致信息丢失或不完整等问题。为此，人们希望探索一种有效的解决方法，它既能减少参与数据分析的变量个数，同时也不会造成统计信息的大量浪费和丢失。

因子分析就是在尽可能不损失信息或者少损失信息的情况下，将多个变量减少为少数几个因子的方法。这几个因子可以高度概括大量数据中的信息，这样，既减少了变量个数，又同样能再现变量之间的内在联系。

# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

## 2、基本原理

通常针对变量作因子分析，称为R型因子分析；另一种对样品作因子分析，称为Q型因子分析，这两种分析方法有许多相似之处。

R型因子分析数学模型是：

设原有p个变量  $x_1, \dots, x_p \dots$  且每个变量（或经标准化处理后）的均值为0，标准差为1。现将每个原有变量用k（ $k < p$ ）个因子  $f_1, f_2, \dots, f_k$  的线性组合来表示，即有：

$$\begin{cases} x_1 = a_{11}f_1 + a_{12}f_2 + \dots + a_{1k}f_k + \varepsilon_1 \\ x_2 = a_{21}f_1 + a_{22}f_2 + \dots + a_{2k}f_k + \varepsilon_2 \\ \dots\dots\dots \\ x_p = a_{p1}f_1 + a_{p2}f_2 + \dots + a_{pk}f_k + \varepsilon_p \end{cases}$$

上式就是因子分析的的数学模型，也可以用矩阵的形式表示为  $X = AF + \varepsilon$

## 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

其中， $X$ 是可实测的随机向量。 $F$ 称为因子，由于它们出现在每个原有变量的线性表达式中，因此又称为公共因子。 $A$ 称为因子载荷矩阵， $a_{ij}(i=1,2,\dots,p; j=1,2,\dots,k)$ 称为因子载荷。 $\varepsilon$ 称为特殊因子，表示了原有变量不能被因子解释的部分，其均值为0

因子分析的基本思想是通过对变量的相关系数矩阵内部结构的分析，从中找出少数几个能控制原始变量的随机变量  $f_i(i=1,2,\dots,k)$  选取公共因子的原则是使其尽可能多的包含原始变量中的信息，建立模型  $X = AF + \varepsilon$ ，忽略  $\varepsilon$ ，以 $F$ 代替 $X$ ，用它再现原始变量 $X$ 的信息，达到简化变量降低维数的目的。

# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

## 3、基本步骤

由于实际中数据背景、特点均不相同，故采用因子分析步骤上可能略有差异，但是一个较完整的因子分析主要包括如下几个过程：

### (1) 确认待分析的原变量是否适合作因子分析

因子分析的主要任务是将原有变量的信息重叠部分提取和综合成因子，进而最终实现减少变量个数的目的。故它要求原始变量之间应存在较强的相关关系。进行因子分析前，通常可以采取计算相关系数矩阵、巴特利特球度检验和KMO检验等方法来检验候选数据是否适合采用因子分析。

### (2) 构造因子变量

将原有变量综合成少数几个因子是因子分析的核心内容。它的关键是根据样本数据求解因子载荷阵。因子载荷阵的求解方法有基于主成分模型的主成分分析法、基于因子分析模型的主轴因子法、极大似然法等。

# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

## (3) 利用旋转方法使因子变量更具有可解释性

将原有变量综合为少数几个因子后，如果因子的实际含义不清，则不利于后续分析。为解决这个问题，可通过因子旋转的方式使一个变量只在尽可能少的因子上有比较高的载荷，这样使提取出的因子具有更好的解释性。

## (4) 计算因子变量得分

实际中，当因子确定以后，便可计算各因子在每个样本上的具体数值，这些数值称为因子得分。于是，在以后的分析中就可以利用因子得分对样本进行分类或评价等研究，进而实现了降维和简化问题的目标。

# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

根据上述步骤，可以得到进行因子分析的详细计算过程如下。

- ①将原始数据标准化，以消除变量间在数量级和量纲上的不同。
- ②求标准化数据的相关矩阵。
- ③求相关矩阵的特征值和特征向量。
- ④计算方差贡献率与累积方差贡献率。
- ⑤确定因子：设 $F_1, F_2, \dots, F_p$ 为 $p$ 个因子，其中前 $m$ 个因子包含的数据信息总量（即其累积贡献率）不低于85%时，可取前 $m$ 个因子来反映原评价指标。
- ⑥因子旋转：若所得的 $m$ 个因子无法确定或其实际意义不是很明显，这时需将因子进行旋转以获得较为明显的实际含义。
- ⑦用原指标的线性组合来求各因子得分。
- ⑧综合得分：通常以各因子的方差贡献率为权，由各因子的线性组合得到综合评价指标函数。

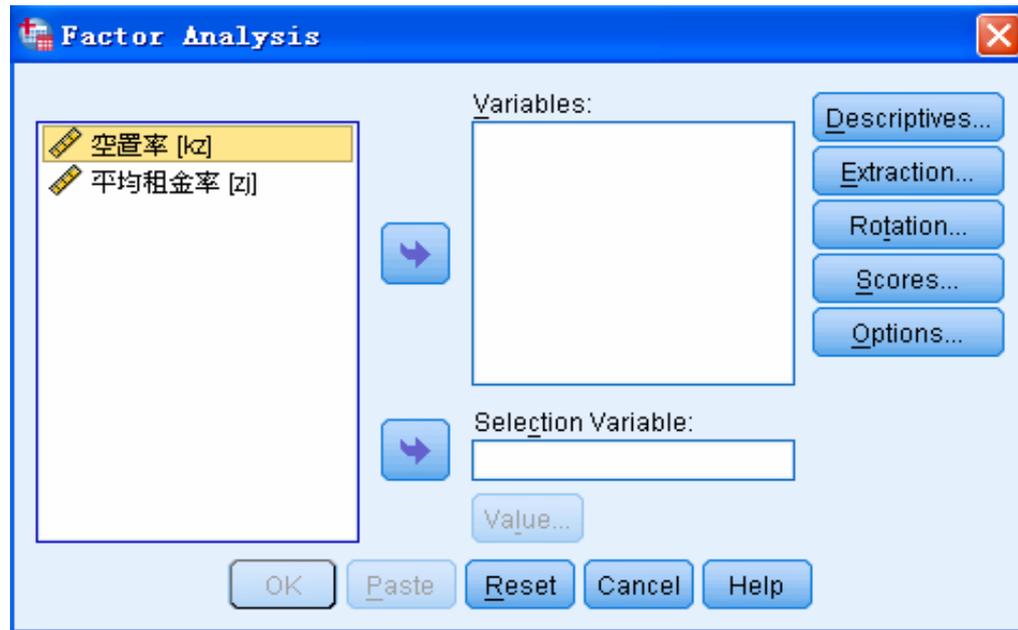
# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

## 9.1.2 因子分析的SPSS操作详解

Step01: 打开对话框

选择菜单栏中的【Analyze（分析）】→【Data Reduction（降维）】→【Factor（因子）】命令，弹出【Factor Analysis（因子分析）】对话框，这是因子分析的主操作窗口。



# 9.1 SPSS在因子分析中的应用



CONCEPT  
RATE

## Step02: 选择因子分析变量

在【Factor Analysis（因子分析）】对话框左侧的候选变量列表框中选择进行因子分析的变量，将其添加至【Variables（变量）】列表框中。如果要选择参与因子分析的样本，则需要将条件变量添加至【Selection Variable（选择变量）】列表框中，并单击【Value】按钮输入变量值，只有满足条件的样本数据才能进行后续的因子分析。

## Step03: 选择描述性统计量

单击【Descriptives】按钮，在弹出的对话框中可以选择输出描述性统计量及相关矩阵等内容。

# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

具体选项含义如下：

① 【Statistics (统计量)】选项组

- Univariate descriptives: 单变量描述统计量，即输出参与分析的各原始变量的均值、标准差等。
- Initial solution: 初始分析结果，系统默认项。输出各个分析变量的初始共同度、特征值以及解释方差的百分比等。

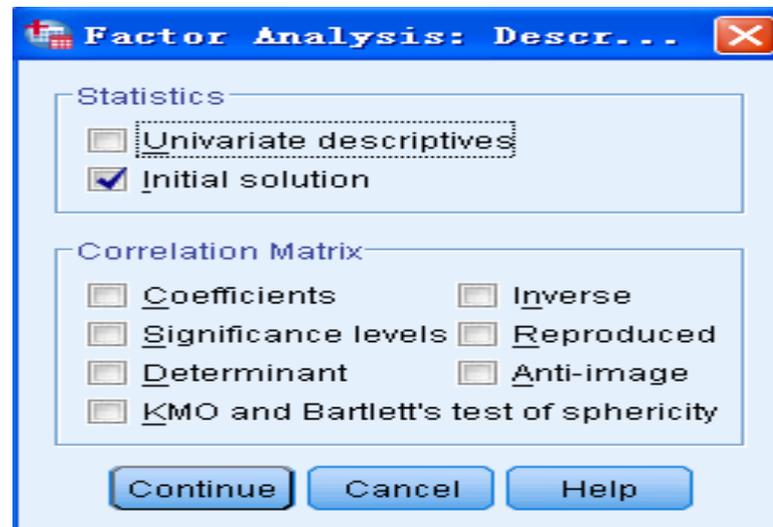
② 【Correlation Matrix (相关矩阵)】选项组

- Coefficients: 原始分析变量间的相关系数矩阵。
- Significance levels: 显著性水平。输出每个相关系数相对于相关系数为0的单尾假设检验的概率水平。
- 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以及残差阵。
- Anti-image: 象相关阵。包括偏相关系数的负数以及偏协方差的负数。在一个好的因子模型中，除对角线上的系数较大外，远离对角线的元素应该比较小。

## 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

KMO and Bartlett's test of sphericity: KMO 和 Bartlett 检验。前者输出抽样充足度的Kaisex-Meyer-Olkin 测度,用于检验变量间的偏相关是否很小。后者Bartlett 球度方法检验相关系数阵是否是单位阵。如果是单位阵,则表明因子模型不合适采用因子模型。



# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

## Step04: 选择因子提取方法

单击【 Extract (抽取) 】按钮，在弹出的对话框中可以选择提取因子的方法及相关选项。

- ① 在【Method (方法)】框下拉列表框中可以选择因子提取方法。
  - Principal components: 主成份分析法。该方法假设变量是因子的纯线性组合。第一成分有最大的方差，后续的成分其可解释的方差逐个递减。
  - Unweighted least square : 不加权最小二乘法。
  - Generalized least squares : 加权最小二乘法。
  - Maximum likelihood : 极大似然法。
  - Principal axis factoring : 主轴因子提取法。
  - Alpha factoring:  $\alpha$  因子提取法。
  - Image factoring: 映象因子提取法。

# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

## ② 【Analyze（分析）】选项组

- Correlation matrix: 相关系数矩阵, 系统默认项。
- Covariance matrix: 协方差矩阵。

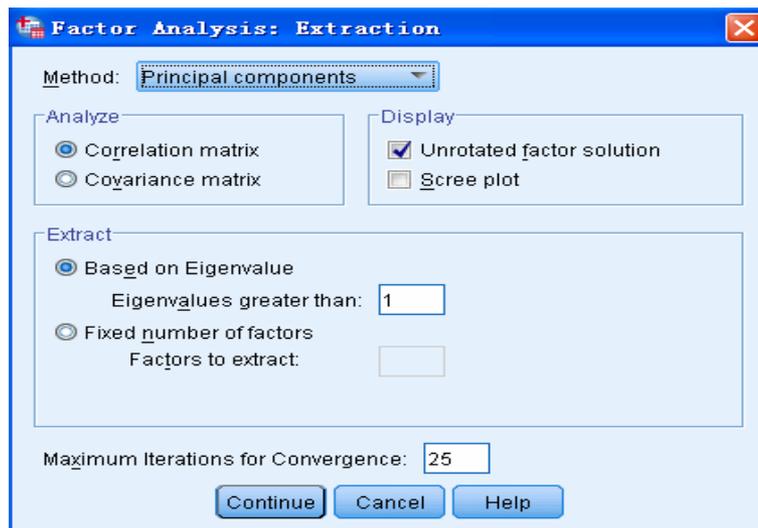
## ③ 【Display（输出）】选项组: 输出与因子提取有关的选项。

- Unrotated factor solution: 输出未经旋转的因子提取结果。此项为系统默认的输出方式。
- Scree plot: 输出因子的碎石图。它显示了按特征值大小排列的因子序号。它有助于确定保留多少个因子。典型的碎石图会有一个明显的拐点, 在该点之前是与大因子连接的陡峭的折线, 之后是与小因子相连的缓坡折线。

## 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

- ④ 【Extract（抽取）】选项组：输出与提取结果有关的选择项。由于理论上因子数目与原始变量数目相等，但因子分析的目的是用少量因子代替多个原始变量，选择提取多少个因子是由本栏来决定。
- Eigenvalues over: 指定提取的因子的特征值数目。在此项后面的矩形框中给出输入数值（系统默认值为1），即要求提取那些特征值大于1的因子。
- Number of f actors: 指定提取公因子的数目。用鼠标单击选择此项后，将指定其数目。
- ⑤ Maximum iterations for Convergence: 在对应的文本框中指定因子分析收敛的最大迭代次数。系统默认的最大迭代次数为25。

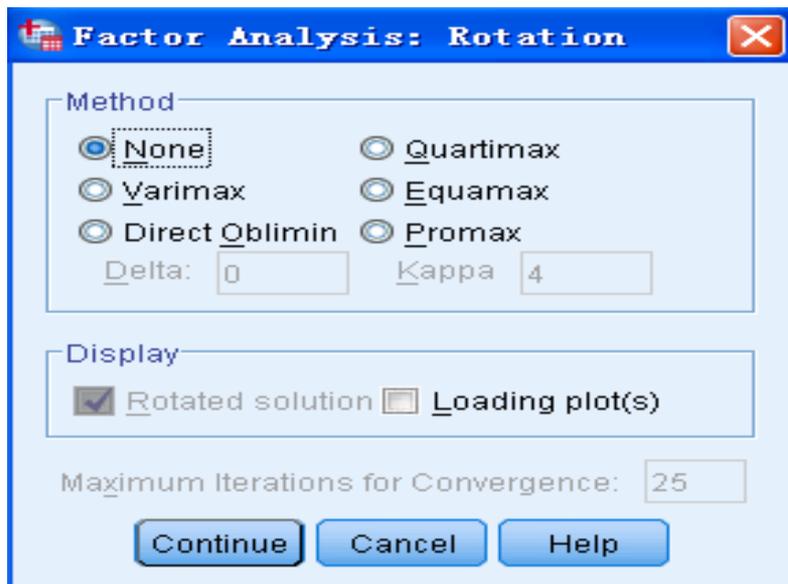


## 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

Step05: 选择因子旋转方法

单击【Rotation】按钮，在弹出的对话框可以选择因子旋转方法及  
相关选项。



# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

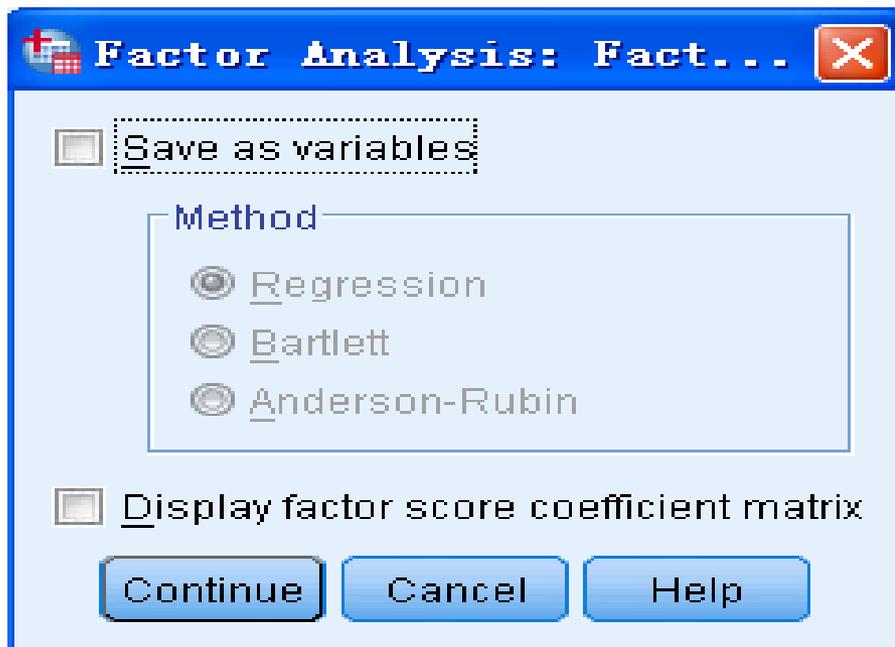
- ① **【Method（方法）】** 选项组选择旋转方法。
  - None: 不进行旋转, 此为系统默认的选择项。
  - Varimax: **方差最大旋转法**。这是一种**正交旋转**方法。它使每个因子具有最高载荷的变量数最小, 因此可以简化对因子的解释。
  - Direct Oblimin: **直接斜交旋转法**。指定此项可以在下面的“Delta”矩形框中键入  $\delta$  值, 该值应该在0~1 之间。系统默认的  $\delta$  值为0。
  - Quartma: 四次方最大**正交旋转法**。该旋转方法使每个变量中需要解释的因子数最少。
  - Equamax: 平均**正交旋转法**。
  - Promax: **斜交旋转方法**。允许因子彼此相关。它比直接斜交旋转更快, 因此适用于**大数据集**的因子分析。指定此项可以在下面的“Kappa”矩形框中键入“ $\kappa$ ”值, 默认为4 (此值最适合于分析)。
- ② **【Display（输出）】** 选项组: 选择有关输出显示。
  - Rotated solution: **旋转解**。在Method栏中指定旋转方法才能选择此项。
  - Loading plot(s): **因子载荷散点图**。指定此项将给出以前两因子为坐标轴的各变量的载荷散点图。
- ③ Maximum iterations for Convergence: 可以指定**旋转收敛**的**最大迭代次数**。系统默认值为25。可以在此项后面的文本框中输入指定值。

## 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

Step06: 选择因子得分

单击【Scores】按钮，在弹出的对话框中可以选择因子得分方法及相关选项。具体选项含义如下。



# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

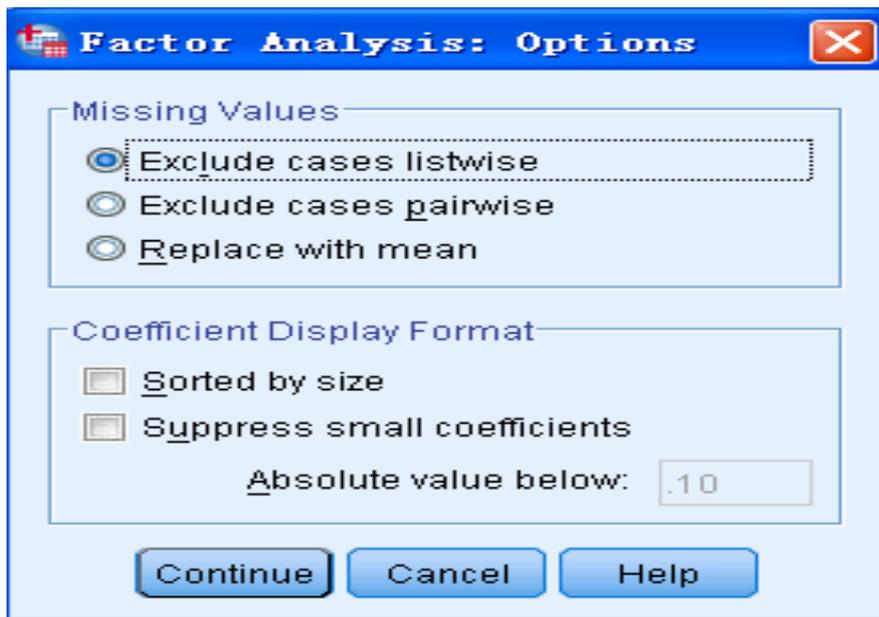
- ① 【Save as variables (保存为变量)】选项组：将因子得分作为新变量保存在数据文件中。
    - Save as variables: 将因子得分作为新变量保存在工作数据文件中。程序运行结束后，在数据窗中显示出新变量。
  - ② 【Method (方法)】选项组：指定计算因子得分的方法。
    - Regression: 回归法。选择此项，其因子得分的均值为0。方差等于估计的因子得分与实际因子得分值之间的复相关系数的平方。
    - Bartlett: 巴特利特法。选择此项，因子得分均值为0。超出变量范围各因子平方和被最小化。
    - Anderson-Rubin: 安德森-鲁宾法。选择此项，是为了保证因子的正交性。
- 本例选中“Regression”项。
- ③ 在输出窗中显示因子得分。
    - Display factor score coefficient matrix: 输出因子得分系数矩阵。

## 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

Step07: 其他选项输出

单击【Options】按钮，在弹出的对话框中可以选择一些附加输出项。具体选项含义如下。



# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

- ① **【Missing Values (缺失值)】** 选项组：选择处理缺失值方法。
  - Exclude cases listwise: 分析变量中带有缺失值的观测量都不参与后续分析。
  - Exclude cases pairwise: 成对剔除带有缺失值的观测量。
  - Replace with mean: 用该变量的均值代替工作变量的所有缺失值。
- ② **【Coefficient Display Format (系数显示格式)】** 选项组：选择载荷系数的显示格式。
  - Sorted by size: 将**载荷系数**按其大小排列构成矩阵，使在同一因子上具有较高载荷的变量排在一起。便于得出结论。
  - Suppress absolute values less than: 不显示那些**绝对值小于指定值**的载荷系数。选择此项后还需要在该项的参数框中键入0~1之间的数值作为临界值。系统默认的临界值为0.10。

Step08: 单击 **【OK】** 按钮，结束操作，SPSS软件自动输出结果。



# 9.1 SPSS在因子分析中的应用

## 9.1.3 实例分析：居民消费结构的变动

### 1. 实例内容

消费结构是指在消费过程中各项消费支出占居民总支出的比重。它是反映居民生活消费水平、生活质量变化状况以及内在过程合理化程度的重要标志。而消费结构的变动不仅是消费领域的重要问题，而且也关系到整个国民经济的发展。因为合理的消费结构及消费结构的升级和优化不仅反映了消费的层次和质量的提高，而且也为建立合理的产业结构和产品结构提供了重要的依据。

表9-1是某市居民生活费支出费用，具体分为食品、衣着、家庭设备用品及服务、医疗保健、交通通讯、文教娱乐及服务、居住和杂项商品与服务等8个部分。请利用因子分析探讨该市居民消费结构，为产业政策的制定和宏观经济的调控提供参考。



# 9.1 SPSS在因子分析中的应用

## 2. 实例操作

数据文件9-1.sav是某市居民在食品、衣着、医疗保健等八个方面的消费数据，这些指标之间存在着不同强弱的相关性。如果单独分析这些指标，无法能够分析居民消费结构的特点。因此，可以考虑采用因子分析，将这八个指标综合为少数几个因子，通过这些公共因子来反映居民消费结构的变动情况。



# 9.1 SPSS在因子分析中的应用

## 3. 实例结果及分析

### (1) 描述性统计表

下表显示了食品、衣着等这八个消费支出指标的描述统计量，例如均值、标准差等。这为后续的因子分析提供了一个直观的分析结果。可以看到，食品支出消费所占的比重最大，其均值等于39.4750%，其次是文化娱乐服务支出消费和交通通信支出消费。所有的消费支出中，医疗保健消费支出占的比重最低。



## 9.1 SPSS在因子分析中的应用

	Mean	Std. Deviation	Analysis N
食品	39.4750	2.29705	8
衣着	6.4875	.86592	8
家庭设备用品及服务	7.9125	2.87772	8
医疗保健	6.3625	1.54729	8
交通和通信	8.1750	2.61302	8
文化娱乐服务	14.4750	2.30016	8
居住	12.1625	2.91545	8
杂项商品与服务	2.9125	.52491	8



# 9.1 SPSS在因子分析中的应用

## (2) 因子分析共同度

下表是因子分析的**共同度**，显示了所有变量的共同度数据。第一列是因子分析初始解下的变量共同度。它表明，对原有八个变量如果采用主成分分析法提取所有八个特征根，那么原有变量的所有方差都可被解释，变量的共同度均为1（原有变量标准化后的方差为1）。

事实上，**因子个数**小于**原有变量的个数**才是因子分析的目的，所以不可能提取全部特征根。于是，第二列列出了按指定提取条件（这里为特征根大于1）提取特征根时的共同度。可以看到，所有变量的绝大部分信息（全部都大于83%）可被因子解释，这些变量信息丢失较少。因此本次因子提取的总体效果理想。



## 9.1 SPSS在因子分析中的应用

	Initial	Extraction
食品	1.000	.842
衣着	1.000	.842
家庭设备用品及服务	1.000	.976
医疗保健	1.000	.954
交通和通信	1.000	.925
文化娱乐服务	1.000	.953
居住	1.000	.978
杂项商品与服务	1.000	.947



# 9.1 SPSS在因子分析中的应用

## (3) 因子分析的总方差解释

接着Spss软件计算得到相关系数矩阵的**特征值**、**方差贡献率**及**累计方差贡献率**结果如表9-4所示。在下一页表中，第一列是因子编号，以后三列组成一组，组中数据项的含义依次是特征根、方差贡献率和累计贡献率。

第一组数据项（第二至第四列）描述了初始因子解的情况。可以看到，第一个因子的特征根值为4.316，解释了原有8个变量总方差的53.947%。前三个因子的累计方差贡献率为94.196%，并且只有它们的取值大于1。说明前3个公因子基本包含了全部变量的主要信息，因此选前3个因子为主因子即可。

同时，Extraction Sums of Squared Loadings和Rotation Sums of Squared Loadings部分列出了因子提取后和旋转后的因子方差解释情况。从表中看到，它们都支持选择3个公共因子。



# 9.1 SPSS在因子分析中的应用

## 因子分析的总方差解释

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.316	53.947	53.947	4.316	53.947	53.947	4.261	53.265	53.265
2	1.989	24.869	78.816	1.989	24.869	78.816	2.030	25.379	78.645
3	1.230	15.380	94.196	1.230	15.380	94.196	1.244	15.551	94.196
4	0.275	3.435	97.631						
5	0.122	1.524	99.155						
6	0.052	0.648	99.804						
7	0.016	0.196	100.000						
8	1.790E-17	2.237E-16	100.000						



## 9.1 SPSS在因子分析中的应用

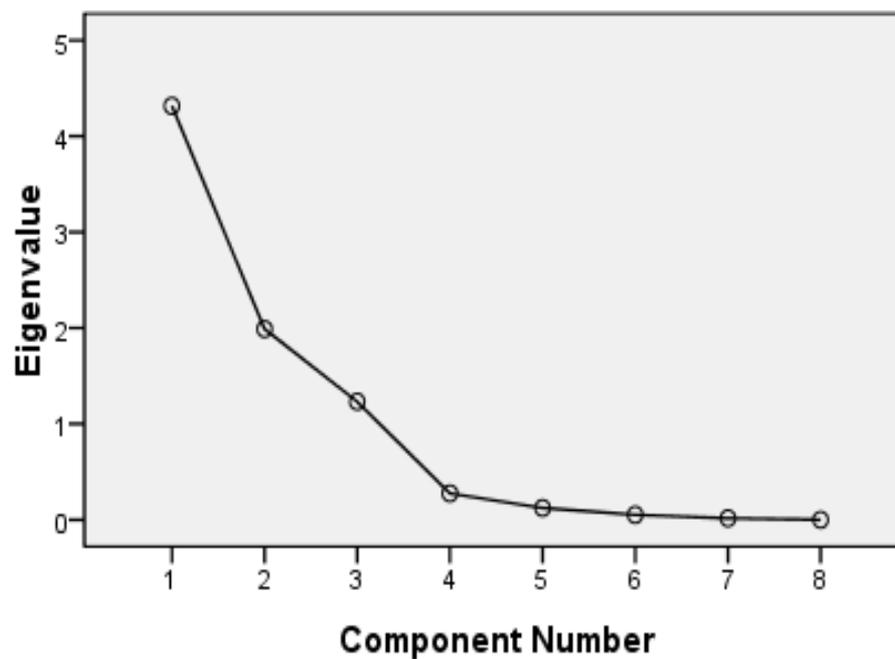
### (4) 因子碎石图

下图为因子分析的碎石图。横坐标为因子数目，纵坐标为特征根。可以看到，第一个因子的特征值很高，对解释原有变量的贡献最大；第三个以后的因子特征根都较小，取值都小于1，说明它们对解释原有变量的贡献很小，称为可被忽略的“高山脚下的碎石”，因此提取前三个因子是合适的。



# 9.1 SPSS在因子分析中的应用

Scree Plot





## 9.1 SPSS在因子分析中的应用

### (5) 旋转前的因子载荷矩阵

下表中显示了因子载荷矩阵，它是因子分析的**核心内容**。通过**载荷系数**大小可以分析不同公共因子所反映的主要指标的区别。从结果看，大部分因子解释性较好，但是仍有少部分指标解释能力较差，例如“食品”指标在三个因子的载荷系数区别不大。因此接着采用**因子旋转方法**使得因子载荷系数向0或1**两极分化**，使大的载荷更大，小的载荷更小。这样结果更具可解释性。



## 9.1 SPSS在因子分析中的应用

旋转前的因子载荷矩阵

	Component		
	1	2	3
医疗保健	0.967	0.102	0.093
文教娱乐及服务	0.962	0.144	-0.085
交通和通信	0.948	-0.082	0.140
家庭设备用品及服务	-0.833	0.503	-0.173
食品	-0.761	0.202	0.471
居住	0.008	-0.970	-0.190
衣着	0.527	0.826	-0.005
杂项商品与服务	0.081	-0.183	0.952



## 9.1 SPSS在因子分析中的应用

### (6) 旋转后的因子载荷矩阵

下表中显示了实施因子旋转后的载荷矩阵。可以看到，第一主因子在“交通和通信”和“医疗保健”等五个指标上具有较大的载荷系数，第二主因子在“居住”和“衣着”指标上系数较大，而第三主因子在“杂项商品与服务”上的系数最大。此时，各个因子的含义更加突出。



## 9.1 SPSS在因子分析中的应用

实施因子旋转后的载荷矩阵

	Component		
	1	2	3
交通和通讯	0.946	0.083	0.152
医疗保健	0.938	0.260	0.081
文教娱乐及服务	0.931	0.277	-0.101
家庭设备用品及服务	-0.895	0.343	-0.241
食品	-0.793	0.144	0.438
居住	0.159	-0.974	-0.058
衣着	0.396	0.889	-0.114
杂项商品与服务	0.086	-0.041	0.968



## 9.1

# SPSS在因子分析中的应用

可以看出第一个公因子主要反映了交通和通信、医疗保健、文化娱乐服务、家庭设备用品及服务 and 食品上有较大载荷，说明第一个公因子**综合**反映这几个方面的变动情况，可以将其命名为第一基本生活消费因子，即**享受性消费因子**。

第二个公因子在居住、衣着上的载荷系数较大，代表了这两个方面的变动趋势，可以将其命名为第二基本生活消费因子，即**发展性消费因子**。

第三个公因子在杂项商品与服务上的消费变动较大，因此可以将第三个公因子命名为第三基本生活消费因子，即**其他类型消费因子**。



## 9.1 SPSS在因子分析中的应用

### (7) 因子得分系数

下表中列出了采用回归法估计的因子得分系数。根据表中内容可写出以下因子得分函数：

因子 $F_1 = -0.198X_1 + 0.058X_2 - 0.226X_3 + 0.212X_4 + 0.221X_5 + 0.211X_6 + 0.079X_7 + 0.015X_8$ ;

因子 $F_2 = 0.123X_1 + 0.425X_2 + 0.200X_3 + 0.094X_4 + 0.008X_5 + 0.096X_6 - 0.498X_7 + 0.015X_8$ ;

因子 $F_3 = 0.365X_1 - 0.059X_2 - 0.174X_3 + 0.069X_4 + 0.119X_5 - 0.077X_6 - 0.088X_7 + 0.779X_8$ ;



## 9.1 SPSS在因子分析中的应用

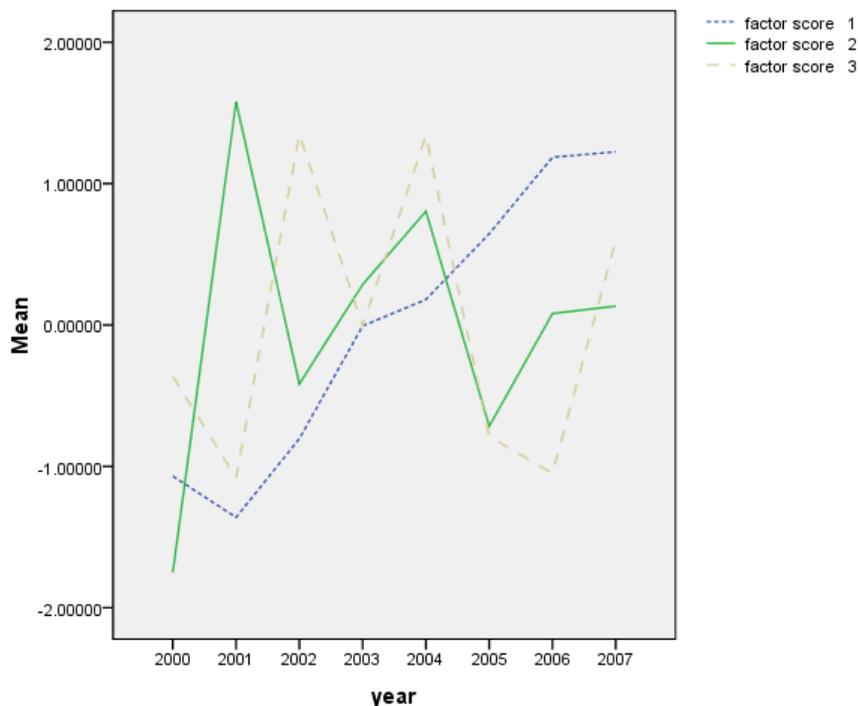
因子得分系数

	Component		
	1	2	3
食品	-0.198	0.123	0.365
衣着	0.058	0.425	-0.059
家庭设备用品及服务	-0.226	0.200	-0.174
医疗保健	0.212	0.094	0.069
交通和通讯	0.221	0.008	0.119
文教娱乐及服务	0.211	0.096	-0.077
居住	0.079	-0.498	-0.088
杂项商品与服务	0.015	0.015	0.779



## 9.1 SPSS在因子分析中的应用

不仅如此，原数据文件中增加了FAC1\_1、FAC2\_1和FAC3\_1三个变量，它们表示了三个因子在不同年份的得分值。为了进一步揭示因子的变动情况，绘制了如下图所示的因子变动趋势图。



## 9.2 SPSS在聚类分析中的应用

CONCEPT  
STRATE

### 9.2.1 聚类分析的基本原理

#### 1、方法概述

**聚类分析**又称**群分析**，它是研究（样品或指标）**分类问题**的一种**多元统计方法**，所谓**类**，通俗地说，就是指**相似元素的集合**。

#### 2、聚类分析的分类

根据分类对象的不同可分为**样品聚类**和**变量聚类**。

##### (1) 样品聚类

**样品聚类**在统计学中又称为**Q型聚类**。用SPSS的术语来说就是对事件(Cases)进行聚类，或是说对**观测量**进行聚类。它是根据被观测的对象的各种特征，即反映被观测对象的特征的**各变量值**进行分类。

- 由上图可以看出，在2000~2007年期间，第一公因子除了开始阶段有些下降外，此后每年都在逐步回升，并于2006年达到最高点。这主要是由于前几年国企改革和中国经济的软着陆，下岗职工大量增加，因此这段时间人们在享受性消费上的支出是减少的，而在其他基本生活消费上的支出增加。而随着经济的发展和收入的增加，享受性消费逐步增加，其他生活消费由于享受性消费的突然增加而减少后也会逐渐增加。第二公因子得分的起伏波动主要是由市民住房比重有升有降的变动引起的，根本原因还是和国家执行住房改革的力度密切相关，但由于住房改革政策的推行相对于其他政策而言较为缓慢，所以市民对住房消费存在一定的不确定性，这就造成了住房比重在总消费中的升降变化。第三公因子一直波动不已，这说明市民在杂项上的消费仍有较大的发展空间。



## 9.2 SPSS在聚类分析中的应用

### (2) 变量聚类

变量聚类在统计学又称为R型聚类。反映同一事物特点的变量有很多，我们往往根据所研究的问题选择部分变量对事物的某一方面进行研究。由于人类对客观事物的认识是有限的，往往难以找出彼此独立的有代表性的变量，而影响对问题的进一步认识和研究。例如在回归分析中，由于自变量的共线性导致偏回归系数不能真正反映自变量对因变量的影响等。因此往往先要进行变量聚类，找出彼此独立且有代表性的自变量，而又不丢失大部分信息。

值得提出的是将聚类分析和其它方法联合起来使用，如判别分析、主成分分析、回归分析等往往效果更好。



## 9.2 SPSS在聚类分析中的应用

### 3、距离和相似系数

为了将样品（或指标）进行分类，就需要研究样品之间关系。目前用得最多的方法有两个：一种方法是用相似系数，性质越接近的样品，它们的相似系数的绝对值越接近1，而彼此无关的样品，它们的相似系数的绝对值越接近于零。比较相似的样品归为一类，不怎么相似的样品归为不同的类。另一种方法是将一个样品看作P维空间的一个点，并在空间定义距离，距离越近的点归为一类，距离较远的点归为不同的类。但相似系数和距离有各种各样的定义，而这些定义与变量的类型关系极大。

$$d_{ij}(q) = \left( \sum_{a=1}^p |x_{ia} - x_{ja}|^q \right)^{1/q}$$

## 9.2 SPSS在聚类分析中的应用



常用的距离和相似系数定义如下：

### (1) 距离

如果把n个样品（X中的n个行）看成p维空间中n个点，则两个样品间相似程度可用p维空间中两点的距离来度量。令 $d_{ij}$ 表示样品 $X_i$ 与 $X_j$ 的距离。常用的距离有：

明氏（Minkowski）距离  $d_{ij}(q) = \left( \sum_{a=1}^p |x_{ia} - x_{ja}|^q \right)^{1/q}$

当 $q=1$ 时

$$d_{ij}(\infty) = \max_{1 \leq a \leq p} |x_{ia} - x_{ja}|$$

即绝对距离

当 $q=2$ 时

$$d_{ij}(2) = \left( \sum_{a=1}^p (x_{ia} - x_{ja})^2 \right)^{1/2}$$

即欧氏距离

当  $q=\infty$  时

$$d_{ij}(1) = \sum_{a=1}^p |x_{ia} - x_{ja}|$$

即切比雪夫距离



## 9.2 SPSS在聚类分析中的应用

马氏 (Mahalanobis) 距离

$$d_{ij}^2(M) = (X_i - X_j)' \Sigma^{-1} (X_i - X_j)$$

其中  $\Sigma$  表示指标的协差阵, 即:

$$\Sigma = (\sigma_{ij})_{p \times p}$$
$$\sigma_{ij} = \frac{1}{n-1} \sum_{a=1}^n (x_{ai} - \bar{x}_i)(x_{aj} - \bar{x}_j) \quad i, j = 1, \dots, p$$
$$\bar{x}_i = \frac{1}{n} \sum_{a=1}^n x_{ai} \quad \bar{x}_j = \frac{1}{n} \sum_{a=1}^n x_{aj}$$

马氏距离既排除了各指标之间相关性的干扰, 而且还不受各指标量纲的影响。除此之外, 它还有一些优点, 如可以证明, 将原数据作一线性交换后, 马氏距离仍不变等等。



## 9.2 SPSS在聚类分析中的应用

兰氏 (Canberra) 距离

它是由Lance和Williams最早提出的，故称兰氏距离。

$$d_{ij}(L) = \frac{1}{p} \sum_{a=1}^p \frac{|x_{ia} - x_{ja}|}{x_{ia} + x_{ja}} \quad i, j = 1, \dots, n$$

此距离仅适用于一切的情况，这个距离有助于克服各指标之间量纲的影响，**但没有考虑指标之间的相关性。**



## 9.2 SPSS在聚类分析中的应用

### (2) 相似系数

研究样品之间的关系，除了用距离表示外，还有相似系数，顾名思义，相似系数是描写样品之间相似程度的一个量，常用的相似系数有：

#### ● 夹角余弦

将任何两个样品  $X_i$  与  $X_j$  看成p维空间的两个向量，这两个向量的夹角余弦用  $\cos \theta_{ij}$  表示。则

$$\cos \theta_{ij} = \frac{\sum_{a=1}^p x_{ia} x_{ja}}{\sqrt{\sum_{a=1}^p x_{ia}^2 \cdot \sum_{a=1}^p x_{ja}^2}} \quad 1 \leq \cos \theta_{ij} \leq 1$$

当  $\cos \theta_{ij} = 1$ ，说明两个样品  $X_i$  与  $X_j$  完全相似；  
说明  $X_i$  与  $X_j$  相似密切；  
当  $\cos \theta_{ij} = 0$ ，说明  $X_j$  与  $X_i$  完全不一样；  
当  $\cos \theta_{ij}$  接近0，说明  $X_i$  与  $X_j$  差别大。  
当  $\cos \theta_{ij}$  接近1，说明  $X_j$  与  $X_i$  完全不一样。

## 9.2 SPSS在聚类分析中的应用

### ● 相关系数

通常所说相关系数，一般指变量间的相关系数，作为刻画样品间的相似关系也可类似给出定义，即第*i*个样品与第*j*个样品之间的相关系数定义为：

$$r_{ij} = \frac{\sum_{a=1}^p (x_{ia} - \bar{x}_i)(x_{ja} - \bar{x}_j)}{\sqrt{\sum_{a=1}^p (x_{ia} - \bar{x}_i)^2 \cdot \sum_{a=1}^p (x_{ja} - \bar{x}_j)^2}} \quad -1 \leq r_{ij} \leq 1$$

其中

$$\bar{x}_i = \frac{1}{p} \sum_{a=1}^p x_{ia}$$

$$\bar{x}_j = \frac{1}{p} \sum_{a=1}^p x_{ja}$$

聚类分析内容非常丰富，有系统聚类法、有序样品聚类法、动态聚类法、模糊聚类法等。本节主要介绍使用较多的快速聚类法和系统聚类法。

## 9.2 SPSS在聚类分析中的应用

### 9.2.2 快速聚类法的SPSS操作详解

**K-均值聚类法**又叫快速聚类法，可以用于**大量数据**进行聚类分析的情形。它是一种**非分层的**聚类方法。这种方法占用内存少、计算量、处理速度快，特别适合**大样本的聚类分析**。它的基本操作步骤如下：

- 1、指定**聚类数目k**，应由用户指定需要聚成多少类，最终也只能输出关于它的唯一解。这点不同于层次聚类。
- 2、确定k个**初始类的中心**。两种方式：一种是**用户指定方式**，二是根据数据本身结构的中心初步确定每个类别的**原始中心点**。
- 3、根据**距离最近**原则进行分类。逐一计算每一记录到各个中心点的距离，把各个记录按照距离最近的原则归入各个类别，并计算新形成类别的中心点
- 4、按照新的中心位置，重新计算每一记录距离新的类别中心点的距离，并重新进行归类。
- 5、重复步骤4，直到达到一定的**收敛标准**。

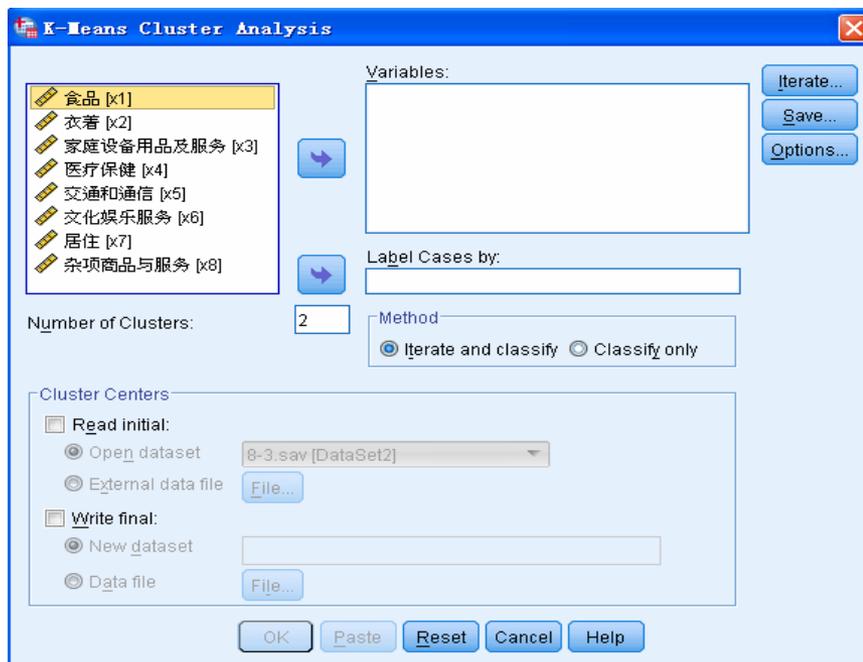
这种方法也常称为**逐步聚类分析**，即先把被聚对象进行初始分类，然后逐步调整，得到最终分类。



## 9.2 SPSS在聚类分析中的应用

Step01: 打开对话框

选择菜单栏中的【Analyze（分析）】→【Classify（分类）】→【K-Means Cluster（K均值聚类）】命令，弹出【K-Means Cluster Analysis（K均值聚类分析）】对话框，这是快速聚类分析的主操作窗口。





## 9.2 SPSS在聚类分析中的应用

### Step02: 选择聚类分析变量

在【K-Means Cluster Analysis (K均值聚类分析)】对话框左侧的候选变量列表框中选择进行聚类分析的变量，将其添加至【Variables (变量)】列表框中。同时可以选择一个标识变量移入【Label Cases by (个案标记依据)】列表框中。

### Step03: 确定分类个数

在【Number of Clusters (聚类数)】列表框中，可以输入确定的聚类分析数目，用户可以根据需要自行修改调整。系统默认的聚类数为2。

### Step04: 选择聚类方法

在【Method (方法)】下拉列表框中可以选择聚类方法。系统默认值选择【Iterative and classify (迭代与分类)】项。

- Iterate and classify: 选择初始类中心，在迭代过程中不断更新聚类中心。把观测量分派到与之最近的以类中心为标志的类中去。
- Classify only: 只使用初始类中心对观测量进行分类，聚类中心始终不变。



## 9.2 SPSS在聚类分析中的应用

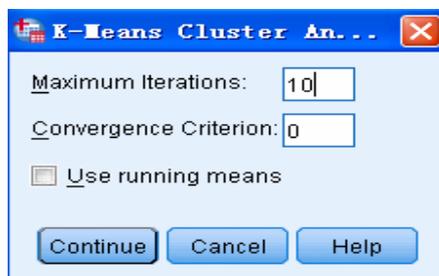
### Step05: 聚类中心的输入与输出

在主对话框中，【Cluster Centers（聚类中心）】选项组表示输入和输出聚类中心。用户可以指定外部文件或数据集作为初始聚类中心点，也可以将聚类分析的聚类中心结果输出到指定文件或数据集中。

- Read initial: 要求使用指定数据文件中的观测量或建立数据集作为初始类中心。
- Write final as File: 要求把聚类结果中的各类中心数据保存到指定的文件或数据集中。

## 9.2 SPSS在聚类分析中的应用

在主对话框中单击Iterate（迭代）按钮，打开设置迭代参数的对话框图，这里可以进一步选择迭代参数。



- Maximum Iterations: 输入K-Means 算法中的迭代次数。改变后面参数框中的数字，则改变迭代次数。当达到限定的迭代次数上限时，即使没有满足收敛判据，迭代也停止。系统默认值为10。选择范围为1-999。
- Convergence Criterion: 指定K-Means 算法中的收敛标准，输入一个不超过1的正数作为判定迭代收敛的标准。系统缺省的收敛标准是0.02，表示当两次迭代计算的最小的类中心的变化距离小于初始类中心距离的百分之2%时迭代停止。

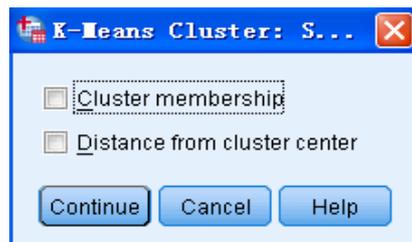
提示：如果设置了以上两个参数，只要在迭代过程中满足了一个参数，迭代就停止。

- Use running means: 使用移动平均。选中该复选框，限定在每个观测量被分配到一类后立刻计算新的类中心。如果不选择此项，则在完成了所有观测量的一个分配后再计算各类的类中心，这样可以节省迭代时间。

## 9.2 SPSS在聚类分析中的应用

### Step07: 输出聚类结果

在主对话框中单击【Save (保存)】按钮，弹出【Save New Variables (保存新变量)】对话框，它用于选择保存新变量。



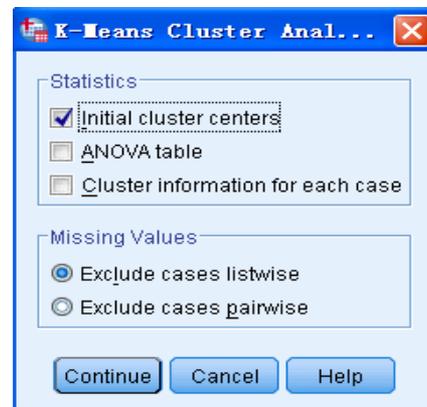
- Cluster membership: 在当前数据文件中建立一个名为“qc1\_1”新变量。其值表示聚类结果，即各观测量被分配到哪一类。它的取值为1、2、3…的序号。
- Distance from cluster center: 在当前数据文件中建立一个名为“qc1\_2”新变量。其值为各观测量与所属类中心之间的欧氏距离。



## 9.2 SPSS在聚类分析中的应用

### Step08: 其他选项输出

在主对话框中单击【Option (选项)】按钮，弹出【Option (选项)】对话框，它用于指定要计算的统计量和对带有缺失值的观测量的处理方式。具体见图：



① 【Statistics (统计量)】选项组：选择输出统计量。

- Initial cluster centers: 初始聚类中心。
- ANOVA table: 方差分析表。

● Cluster information for each case: 显示每个观测量的聚类信息。

② 【Missing Values (缺失值)】选项组：选择处理缺失值方法。

- Exclude cases listwise: 分析变量中带有缺失值的观测量都不参与后续分析。
- Exclude cases pairwise: 成对剔除带有缺失值的观测量。

Step09: 单击【OK】按钮，结束操作，SPSS软件自动输出结果。



## 9.2 SPSS在聚类分析中的应用

- 9.2.3 实例分析：全国环境污染程度分析

为了更深入地了解我国环境的污染程度状况，现利用2009年数据对全国31个省、自治区、直辖市进行聚类分析。



## 9.2 SPSS在聚类分析中的应用

现在要分析我国各个地区的环境污染程度，案例中选择了各地区“工业废气排放总量”、“工业废水排放总量”和“二氧化硫排放总量”三个指标来反映不同污染程度的环境状况，同时选择了北京等省市的数据加以研究。这个问题属于典型的多元分析问题，需要利用多个指标来分析各省市之间环境污染程度的差异。因此，可以考虑利用快速聚类分析来研究各省市之间的差异性，具体操作步骤如下。

- 打随书光盘中的数据文件9-2. sav, 选择菜单栏中的【Analyze (分析)】→【Classify (分类)】→【K-Means Cluster (K均值聚类)】命令, 弹出【K-Means Cluster Analysis (K均值聚类分析)】对话框。
- 在左侧的候选变量列表框中将 $X1$ 、 $X2$ 和 $X3$ 变量设定为聚类分析变量, 将其添加至【Variables (变量)】列表框中; 同时选择 $I$ 作为标识变量, 将其移入【Label Cases by (个案标记依据)】列表框中。
- 在【Number of Clusters (聚类数)】文本框中输入数值“3”, 表示将样品利用聚类分析分为三类, 如下图所示。



**K-Means Cluster Analysis**

Cluster Number of Cas...  
Distance of Case from ...

Variables:  
工业废气排放总量 [X1]  
工业废水排放总量 [X2]  
二氧化硫排放总量 [X3]

Label Cases by:  
省市 [Y]

Number of Clusters: 3

Method:  
 Iterate and classify  Classify only

Cluster Centers

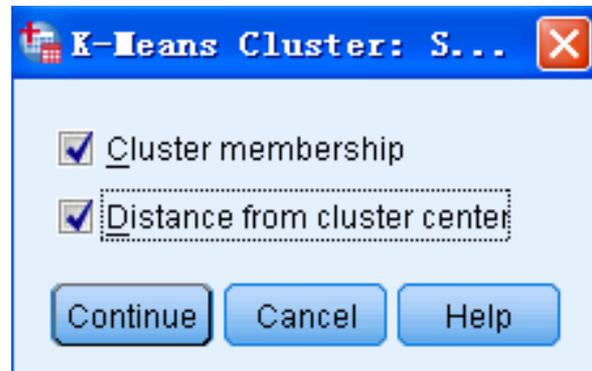
Read initial:  
 Open dataset  
 External data file File...

Write final:  
 New dataset  
 Data file File...

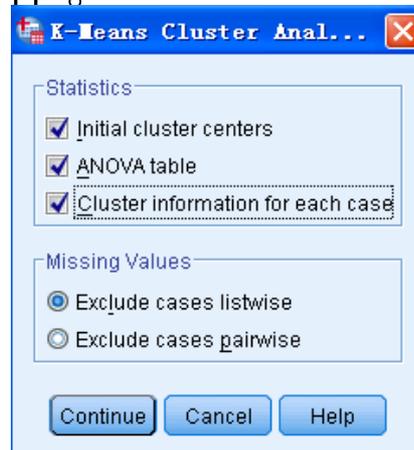
Iterate...  
Save...  
Options...

OK Paste Reset Cancel Help

- 单击【Save（保存）】按钮，弹出【K-Means Cluster Analysis: Save（K均值聚类分析：保存）】对话框；勾选【Cluster membership（聚类新成员）】和【Distance from cluster center（与聚类中心的距离）】复选框，表示输出样品的聚类类别及距离，其他选项保持系统默认设置，如下图所示，单击【Continue（继续）】按钮返回主对话框。



- 单击【Options (选项)】按钮，弹出【K-Means Cluster Analysis: Options (K均值聚类分析: 选项)】对话框；勾选【Statistics (统计量)】选项组中的复选框，其他选项保持系统默认设置，如下图所示，单击【Continue (继续)】按钮返回主对话框，单击【OK (确定)】按钮完成操作。





## 9.2 SPSS在聚类分析中的应用

### 实例结果及分析

#### (1) 快速聚类分析的初始中心

SPSS软件首先给出了进行快速聚类分析的**初始中心数据**。由于这里是要求将样品分为三类，因此软件给出了三个中心位置。但是，这些中心位置可能在后续的迭代计算中出现调整。

快速聚类分析的初始中心

	Cluster		
	1	2	3
工业废气排放总量	15	22186	27432
工业废水排放总量	942	140325	256160
二氧化硫排放总量	0.2	135.5	107.4



## 9.2 SPSS在聚类分析中的应用

### (2) 迭代历史表

下表显示了快速聚类分析的迭代过程。可以看到，第一次迭代的变化值最大，其后随之减少。最后第三次迭代时，聚类中心就不再变化了。这说明，本次快速聚类的迭代过程速度很快。

迭代历史表

Iteration	Change in Cluster Centers		
	1	2	3
1	29063.875	15957.005	26705.187
2	4706.401	3783.482	22208.692
3	0.000	0.000	0.000



## 9.2 SPSS在聚类分析中的应用

### (3) 聚类分析结果列表

通过快速聚类分析的最终结果列表可以看到整个样品被分为以下三大类。

- 第一类：北京、天津、山西、内蒙古等20个地区。这些地区工业废水、废气及二氧化硫的排放总量相对最低。
- 第二类：河北、福建、河南、湖北、湖南、广西和四川。它们的污染程度在所有省份中位居中等水平。
- 第三类：江苏、浙江、山东和广东。这些地区的工业废水、废气及二氧化硫排放总量是最高的，因此环境污染也最为严重。

表中最后一列显示了样品和所属类别中心的聚类，此表中的最后两列分别作为新变量保存于当前的工作文件中。



## 9.2 SPSS在聚类分析中的应用

### (4) 最终聚类分析中心表

如下表所示列出了最终聚类分析中心。可以看到，最后的中心位置较初始中心位置发生了较大的变化。

最终聚类分析中心

	Cluster		
	1	2	3
工业废气排放总量	9921	19079	26025
工业废水排放总量	33219	121194	207780
二氧化硫排放总量	56.0	93.0	110.9



## 9.2 SPSS在聚类分析中的应用

### (5) 最终聚类中心位置之间的距离

如下表所示为快速聚类分析最终确定的各类中心位置的距离表。从结果来看，第一类和第三类之间的距离最大，而第二类和第三类之间的距离最短，这些结果和实际情况是相符合的。

最终聚类中心位置之间的距离

Cluster	1	2	3
1		88449.975	175301.923
2	88449.975		86864.229
3	175301.923	86864.229	



## 9.2 SPSS在聚类分析中的应用

### (6) 方差分析表

如下表所示为方差分析表，显示了各个指标在不同类的均值比较情况。各数据项的含义依次是：**组间均方**、**组间自由度**、**组内均方**、**组内自由度**。可以看到，各个指标在不同类之间的差异是非常明显的，这进一步验证了聚类分析结果的有效性。

方差分析表

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
工业废气排放总量	5.458E8	2	86415059.434	28	6.316	0.005
工业废水排放总量	6.018E10	2	6.317E8	28	95.270	0.000
二氧化硫排放总量	7125.110	2	1510.247	28	4.718	0.017



## 9.2 SPSS在聚类分析中的应用

### (7) 聚类数目汇总

如下表所示是聚类数据汇总表，显示了聚类分析最终结果中各个类别的数目。其中第一类的数目最多，等于20；而第三类的数目最少，只有4个。

聚类数目汇总表

Cluster	1	20.000
	2	7.000
	3	4.000
Valid		31.000
Missing		0.000



## 9.2 SPSS在聚类分析中的应用

### 9.2.4 系统聚类法的SPSS操作详解

系统聚类法常称为**层次聚类法**、**分层聚类法**，也是聚类分析中使用广泛的一种方法。它有两种类型，一是对研究对象**本身**进行分类，称为**Q型聚类**；另一是对研究对象的**观察指标**进行分类，称为**R型聚类**。同时根据聚类过程不同，又分为分解法和凝聚法。

**分解法**：开始把**所有个体**（观测量或变量）都视为**同属一大类**，然后根据距离和相似性逐层分解，直到参与聚类的每个个体自成一类为止。

**凝聚法**：开始把参与聚类的**每个个体**（观测量或变量）视为**一类**，根据两类之间的距离或相似性逐步合并，直到合并为一个**大类**为止。



## 9.2 SPSS在聚类分析中的应用

SPSS中的系统聚类法采用的**凝聚法**，它的算法步骤具体如下。

- 1、首先**将数据各自作为一类**（这时有 $n$ 类），按照所定义的距离计算各数据点之间的距离，形成一个**距离阵**；
- 2、将**距离最近**的两条数据并为一个类别，从而成为 $n-1$ 个类别，计算新产生的类别与其他各个类别之间的距离或相似度，形成新的距离阵；
- 3、按照和第二步相同的原则，再将**距离最接近的两个类别合并**，这时如果类的个数仍然大于1，则继续重复这一步骤，**直到所有的数据都被合并成一个类别为止**。



## 9.2 SPSS在聚类分析中的应用

在系统聚类中，当每个类别有多于一个的数据点构成时，就会涉及如何定义两个类间的距离问题。根据距离公式不同，可能会得到不同的结果，这也就进一步构成了不同的系统聚类方法。常用的方法有如下几种。

- Between-groups linkage: 组间平均距离法。
- Within-groups linkage: 组内平均距离法。
- Nearest neighbor: 最短距离法。
- Furthest neighbor: 最远距离法。
- Centroid clustering: 重心法。
- Median clustering: 中间距离法。
- Ward's method: 离差平方和法。

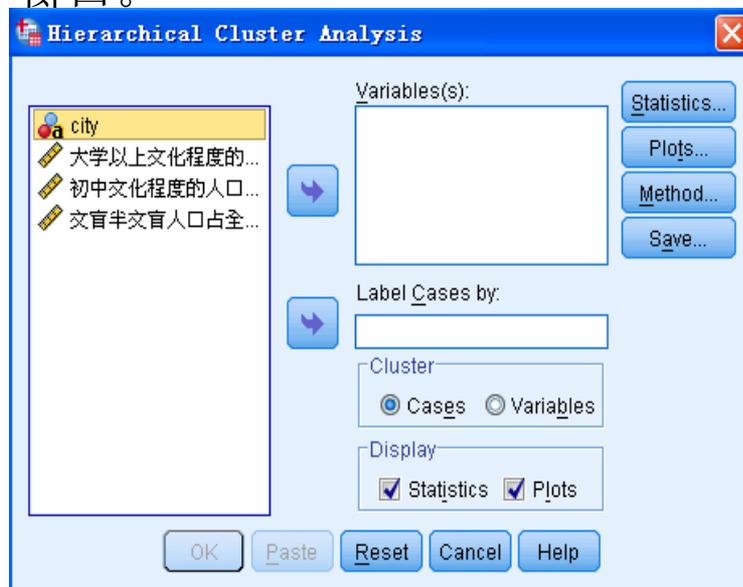


## 9.2 SPSS在聚类分析中的应用

SPSS具体操作步骤如下：

Step01：打开对话框

选择菜单栏中的【Analyze（分析）】→【Classify（分类）】→【Hierarchical Cluster（系统聚类）】命令，弹出【Hierarchical Cluster Cluster Analysis（系统聚类分析）】对话框，这是系统聚类分析的主操作窗口。





## 9.2 SPSS在聚类分析中的应用

### Step02: 选择聚类分析变量

在【Hierarchical Cluster Cluster Analysis（系统聚类分析）】对话框左侧的候选变量列表框中选择进行系统聚类分析的变量，将其添加至【Variable(s)（变量）】列表框中。同时可以选择一个标识变量移入【Label Cases by（标注个案）】列表框中。

### Step03: 选择聚类类型

在【Cluster（分群）】选项组中可以选择聚类类型。系统默认值是【Cases（个案0）】选项。

- Cases: 对观测量（样品）进行聚类，即Q型聚类。
- Variable: 对变量进行聚类，即R型聚类。

### Step04: 选择输出类型

在【Display（输出）】选项组中可以选择输出类型。系统默认值是【Statistics（统计量）】欧诺供给量和【Plots（图）】选项。

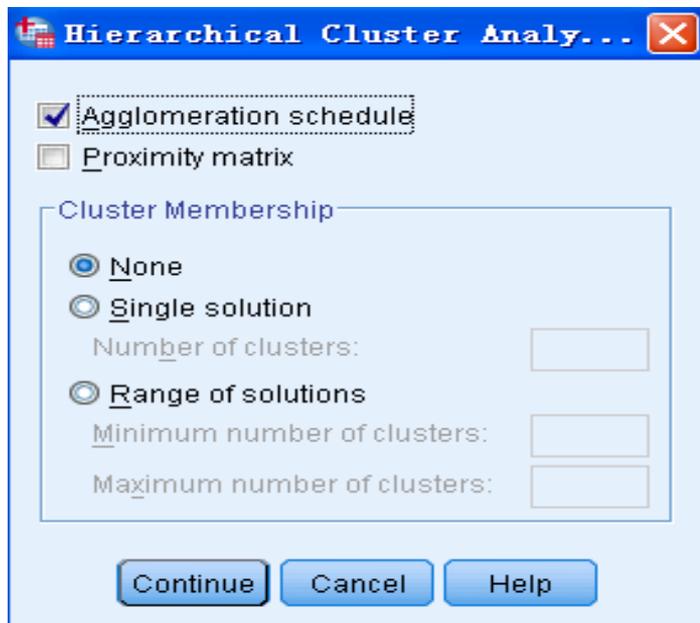
- Statistic: 输出主对话框【Statistics】按钮中设置的的统计量。
- Plots: 输出主对话框中【Plots（图）】按钮中聚类图形。



## 9.2 SPSS在聚类分析中的应用

Step05: 基本统计量输出选择

单击【Statistics】按钮，在弹出的对话框中可以选择进行系统聚类分析的基本统计量。具体选项含义如下。





## 9.2 SPSS在聚类分析中的应用

- ① **【Agglomeration schedule (合并进程表)】**：输出聚类过程表，系统默认选项。显示聚类过程中每一步合并的类或观测量，反映聚类过程中每一步样品或类的合并过程。
- ② **【Proximity matrix (相似性矩阵)】**：输出各类之间的距离矩阵。以矩阵形式给出各项之间的距离或相似性测度值。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 (方法)】**菜单中**【Measure (度量标准)】**栏中的选择。
- ③ **【Cluster Membership (聚类成员)】**栏可以选择**聚类数目**相关的输出项：
  - **【None (无)】**：不显示类成员表，它是系统默认选项。
  - **【Single solution (单一方案)】**：选择此项并在对应的**【Number of clusters (聚类数)】**参数框中指定分类数，这里要求分类数是一个大于1的整数。例如指输入数字“4”，则会在输出窗中显示聚为4类的分析结果。
  - **【Range of solutions (方案范围)】**：选择此选项并在下边的**【Minimum number of clusters (最小聚类数)】**和**【Maximum number of clusters (最大聚类数)】**参数框中输入最小聚类数目和最大聚类数目。它表示分别输出样品或变量的分类数从最小值到最大值的各种分类聚类表。输入的两个数值必须是**不等于1的正整数**，最大类数值不能大于参与聚类的样品数或变量总数。

## 9.2 SPSS在聚类分析中的应用

### Step06: 聚类统计图形输出选择

单击【Plots】按钮，弹出的对话框如下图所示。这里可以选择进行系统聚类分析的统计图形。可选择输出的统计图表有两种，一个是树形图，一个是冰柱图。具体选项含义如下。





## 9.2 SPSS在聚类分析中的应用

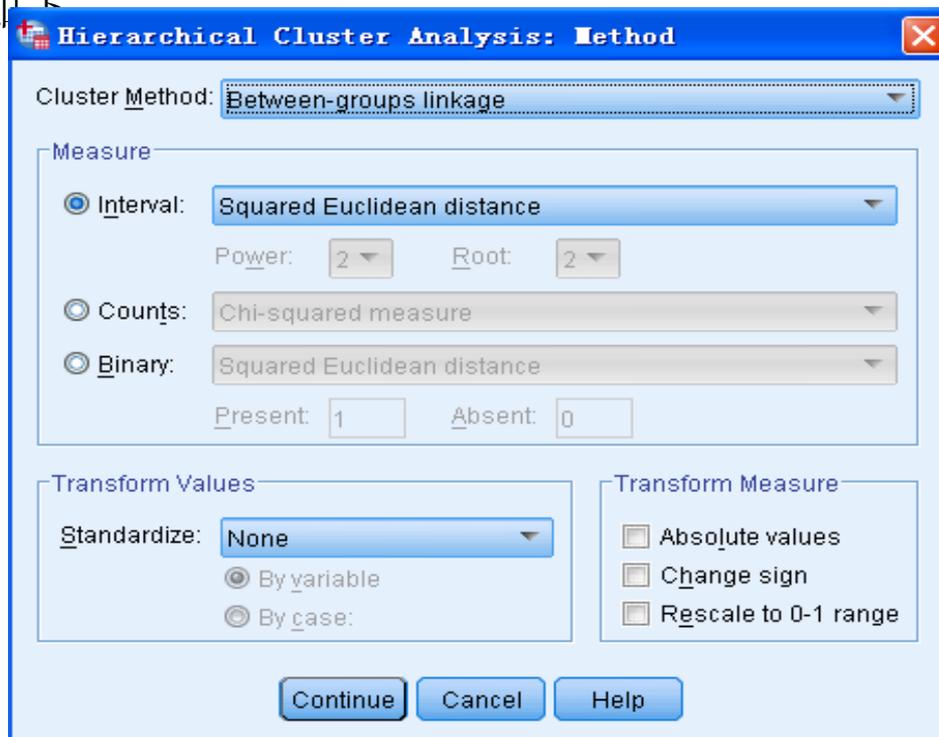
- ① 【Dendrogram (树状图)】：显示树形图。
  - ② 【Icicle (冰柱)】：显示冰柱图形。对于冰柱图的具体选项还可以进一步用以下选择项来确定。
    - All clusters: 显示全部聚类结果的冰柱图。可用此种图查看聚类的全过程。但如果参与聚类的个体很多会造成图形过大。
    - Specified range clusters: 限定显示的聚类范围。当选择此项时，在下面的【Start cluster (开始聚类)】、【Stop cluster (停止聚类)】和【By (排序标准)】后的参数框中输入要求显示聚类过程的开始聚类数、终止聚类数及步长。输入到参数框中的数字必须是正整数。例如，输入的结果是：3, 9, 2, 生成的冰柱图从第三步开始，显示第三、五、七、九步聚类的情况。
    - None: 不输出冰柱图。
- 同时，冰柱图显示方向可以在【Orientation (方向)】选项组中确定。
- Vertical: 纵向显示的冰柱图。
  - Horizontal: 横向显示的冰柱图。



## 9.2 SPSS在聚类分析中的应用

### Step07: 聚类方法选择

单击【Method（方法）】按钮，弹出的对话框如下图所示。在对话框中可以设定聚类方法、距离测度的方法、数值变换方法等内容。具体选项含义如下





## 9.2 SPSS在聚类分析中的应用

- ① 【Cluster Method (聚类方法)】下拉列表框：可以选择聚类方法，具体如下。
- Between-groups linkage: **组间**平均距离法。系统默认选项。合并两类的结果使所有的两类的平均距离最小。
  - Within-groups linkage: **组内**平均距离法。当两类合并为一类后，合并后的类中的所有项之间的平均距离最小。
  - Nearest neighbor: **最近**距离法。采用两类间最近点间的距离代表两类间的距离。
  - Furthest Neighbor: **最远**距离法。用两类之间最远点的距离代表两类之间的距离。
  - Centroid clustering: **重心法**。定义类与类之间的距离为两类中各样品的重心之间的距离。
  - Median clustering: **中位数法**。定义类与类之间的距离为两类中各样品的中位数之间的距离。
  - Ward's method: **最小离差平方和法**。聚类中使类内各样品的离差平方和最小，类间的离差平方和尽可能大。



## 9.2 SPSS在聚类分析中的应用

- ② 【Measure（度量标准）】选项组：可以选择距离测度方法，具体如下。
- 【Interval（区间）】参数框适合于等间隔测度的连续性变量。单击它的右侧框边向下箭头展开下拉菜单，在菜单中选择距离测度方法，具体如下。
- Euclidean distance: 欧氏距离。
  - Squared Euclidean distance: 欧氏距离平方。两项之间的距离是每个变量值之差的平方和。系统默认项。
  - Cosline: 余弦相似性测度，计算两个向量间夹角的余弦。
  - Pearson conelation: 皮尔逊相关系数。它是线性关系的测度，范围是-1~+1。
  - Chebychev: 切比雪夫距离。
  - Block: 曼哈顿（Manhattan）距离，两项之间的距离是每个变量值之差的绝对值总和。
  - Minkowski: 闵科夫斯基距离。
  - Customized: 自定义距离。
- 【Counts（计数）】参数框适合于计数变量（离散变量）。单击它右侧的向下箭头，展开下拉菜单的方法选择以下不相似性测度的方法。具体如下：
- Chi-square measure: 卡方测度。用卡方值测度不相似性。系统默认选项。
  - Phi-square measure: 两组频数之间的 $\Phi^2$ 测度。



## 9.2 SPSS在聚类分析中的应用

【Binary（二分数）】参数框适合于二值变量。首先应该明确，对二值变量，系统默认用1表示某特性出现(或发生)，用0表示某特性不出现(或不发生)。单击它的右侧框边向下箭头展开下拉菜单，在菜单中选择测度方法。具体如下：

- Euclidean distance: 二元变量欧氏距离。
- Squared Euclidean distance: 二元变量欧氏距离的平方。
- Size difference: 不对称指数。其值范围在0 ~ 1 之间。
- Pattern difference: 不相似性测度，范围为0 ~ 1。
- Variance: 方差不相似性测度。
- Dispersion: 离散测度，其范围为-1 ~ 1。
- Shape: 距离测度。范围无上下限。
- Simple matching: 简单匹配测度。
- Phi 4-point correlation: 皮尔逊相关系数二元变量模拟，其值范围为-1 ~ 1。
- Lambda: 其值是Goodman and Kruskal 的  $\lambda$  值，它是一种相似性测度。
- Anderberg' D: 安德伯格D系数。
- Dice: 戴斯匹配系数。
- Hamann: 哈曼匹配系数。



## 9.2 SPSS在聚类分析中的应用

- Jaccard: 杰卡得相似比。
- Kulczynski 1: 库尔津斯基匹配系数。
- Kulczynski 2: 库尔津斯基条件概率测度。
- Lance and Williams: 兰斯-威廉斯测度。
- Ochiai: 该指数是余弦相似性测度的二元形式。范围为0 ~ 1。
- Rogers and Tanimoto: 罗杰斯-谷本匹配系数。
- Russel and Rao: 它是内积(点积)的二元形式。对匹配与不匹配都给予相等的权重。
- Sokal and Sneath 1 ~ 5: 第一种~第五种索克尔-思尼斯匹配系数。
- Yule' s Y: 尤利Y综合系数。
- Yule' s Q: 尤利Q综合系数。。

从上述选项中选择一种测度方法。同时，还可以改变表示某事件发生与不发生的值。在【Present (存在)】和【Absent (不存在)】的参数框中键入用户自己定义的值。定义后，系统将忽略其他值。如果不进行自定义，那么，1代表某事件发生“Present”，0代表某事件不发生“Absent”。



## 9.2 SPSS在聚类分析中的应用

- ③ **【Transform Values (转换数)】** 选项组：可以选择数据标准化的方法。注意只有**等间隔**测度的数据（选择了Interval）或**计数数据**（选择了Counts）才可以进行标准化。具体如下：
- None：不进行标准化。系统默认值。
  - Z scores：数据标准化到Z 分数。标准化后变量均值为0，标准差为1。
  - Range -1 to 1：把数据标准化到-1 到+1 范围内。
  - Range 0 to 1：把数据标准化到0 到+1 范围内。
  - Maximum magnitude of 1：把数据标准化到最大值为1。表示各变量除以最大值。
  - Mean of 1：把数据标准化到均值为1。表示**各变量除以均值**。
  - Standard deviation of 1：把数据标准化到标准差为1。表示各变量除以标准差。

在选择了上述标准化方法后，要在选项组中点选**【By variable (对变量)】**或**【By case (对样品)】**单选钮实施标准化。



## 9.2 SPSS在聚类分析中的应用

④ 【Transform Measure】选项组：可以选择测度的转换方法，具体如下。

- Absolute Values: 把距离值取绝对值。
- Change sign: 把相似性值变为不相似性值或相反。
- Rescale to 0~1 range: 重新调整测度值到范围0~1。

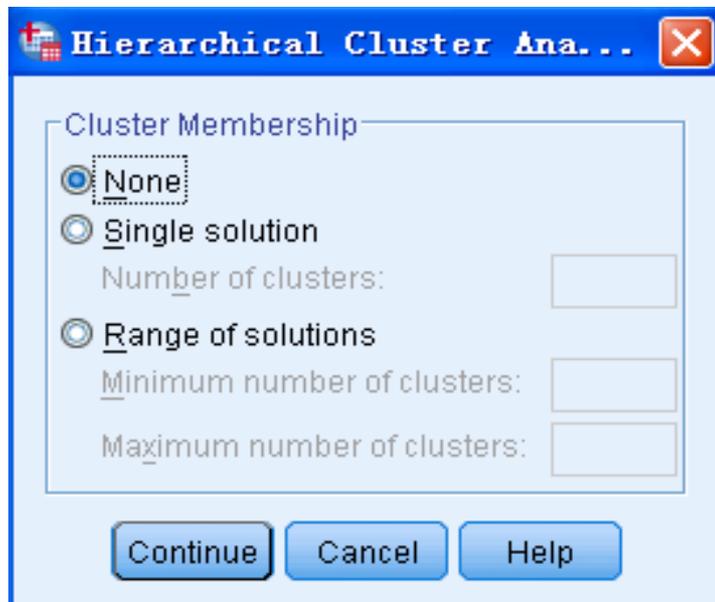
对于已经计算了相似性或不相似性测度的数据，一般不再使用此方法进行转换。如果使用的是已经存在的矩阵，可以选择此类选择项，对输入矩阵进行必要的转换。



## 9.2 SPSS在聚类分析中的应用

### Step08: 聚类结果保存选择

单击【Save】按钮，在弹出的对话框中可以将聚类结果用新变量保存在当前工作数据文件中。具体选项含义如下。





## 9.2 SPSS在聚类分析中的应用

- None: 不建立新变量。
- Single solution: 单个结果输出。生成一个新变量, 表明每个样品在聚类之后所属的类。在【Number of clusters (聚类数)】的矩形框中指定类数。
- Range of solutions: 选择此选项并在下边的【Minimum number of clusters (最小聚类数)】和【Maxmum number of clusters (最大聚类数)】文本框中输入最小聚类数目和最大聚类数目。它表示分别生成样品或变量的分类数从最小值到最大值的各种分类聚类变量。例如输入结果是“4”和“6”时, 它表示在聚类结束后在原变量后面增加了3个新变量分别表明分为4类时、分为5类时和分为6类时的聚类结果。即聚为4、5、6类时各样品分别属于哪一类。

Step09: 单击【OK】按钮, 结束操作, SPSS软件自动输出结果。



## 9.2 SPSS在聚类分析中的应用

### 9.2.5 实例分析：不同地区信息基础设施发展状况的评价

#### 1. 实例内容

要研究世界不同地区信息基础设施的发展状况，这里选取了发达地区、新兴工业化地区、拉美地区、亚洲地区中国家、转型地区等不同类型的20个国家的数据。描述信息基础设施的变量主要有六个。

- (1) Call—每千人拥有电话线数。
- (2) movecall—每千居民蜂窝移动电话数。
- (3) fee—高峰时期每三分钟国际电话的成本。
- (4) Computer—每千人拥有的计算机数。
- (5) mips—每千人中计算机功率（每秒百万指令）。
- (6) net—每千人互连网络户主数。



## 9.2 SPSS在聚类分析中的应用

### 2. 实例操作

现在要分析世界各个地区的信息基础设施的发展状况，案例中选择了“每千人拥有电话线数”、“每千户居民蜂窝移动电话数”等六个指标来反映不同国家信息设施的发展情况，同时选择了近二十个地区的数据加以研究。这个问题也属于典型的多元分析问题，需要利用多个指标来分析地区之间信息基础设施发展的差异。因此，可以利用系统聚类法。



## 9.2 SPSS在聚类分析中的应用

### 3 实例结果及分析

#### (1) 聚类过程表

SPSS软件首先给出了进行系统聚类分析的过程表。下表中的的第一列“Stage”列出了聚类过程的步骤号，第二列“Cluster 1”和第三列“Cluster 2”列出了某一步骤中哪些国家参与了合并。例如从结果中看出，在第一步中，第十个样品(Brazil)和第十二个样品(Mexico)首先被合并在一起。第四列“Coefficients”列出了每一步骤的聚类系数，这一数值表示被合并的两个类别之间的距离大小。第五列“Cluster 1”和第六列“Cluster 2”表示参与合并的国家(类别)是在第几步中第一次出现，0代表该记录是第一次出现在聚类过程中。第七列“Next Stage”表示在这一步骤中合并的类别，下一次将在第几步中与其他类再进行合并。

# 9.2 SPSS在聚类分析中的应用

聚类过程表

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	10	12	.107	0	0	4
2	8	9	.164	0	0	11
3	13	17	.278	0	0	7
4	10	14	.520	1	0	6
5	3	19	.675	0	0	14
6	10	15	1.055	4	0	10
7	13	18	1.099	3	0	15
8	7	20	1.249	0	0	12
9	4	6	1.343	0	0	17
10	10	11	1.421	6	0	13
11	2	8	1.809	0	2	16
12	5	7	1.880	0	8	14
13	10	16	2.247	10	0	15
14	3	5	2.359	5	12	16
15	10	13	3.878	13	7	18
16	2	3	4.719	11	14	18
17	1	4	6.407	0	9	19
18	2	10	11.117	16	15	19
19	1	2	25.049	17	18	0



## 9.2 SPSS在聚类分析中的应用

### (2) 聚类分析结果表

在系统聚类法的聚类结果中可以看到，聚类结果分为三大类。

第Ⅰ类：美国、瑞典、丹麦。

第Ⅱ类：日本、德国、瑞士、新加坡、中国台湾、韩国、法国、英国。

第Ⅲ类：巴西、墨西哥、波兰、匈牙利、智利、俄罗斯、泰国、印度、马来西亚。

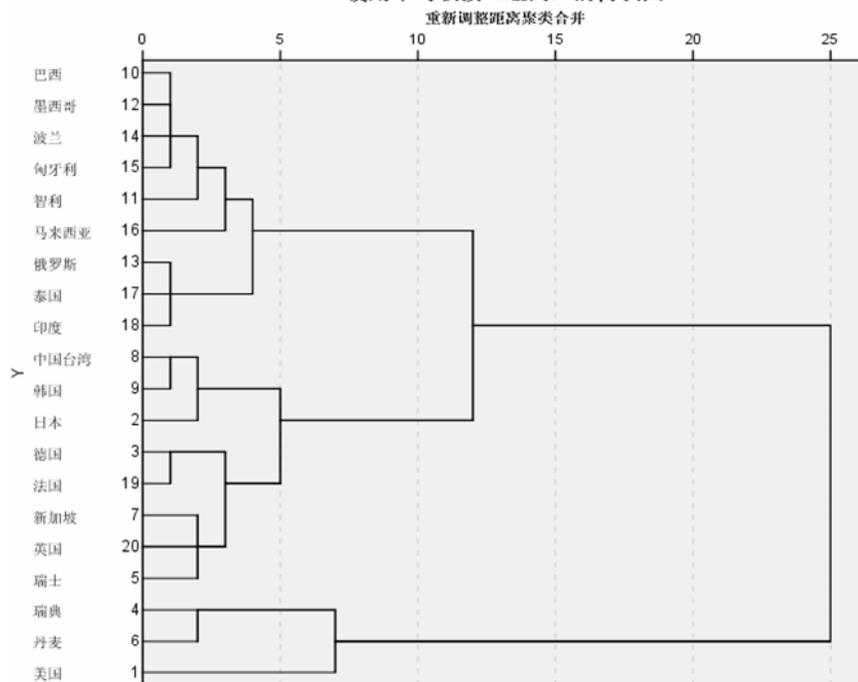


## 9.2 SPSS在聚类分析中的应用

### (3) 树形图

上述已给出了相关聚类结果，最后用树形图（Dendrogram）直观反映整个聚类过程和结果，如图9-37所示。从图中，可以明显看到每个样品从单独一类，逐次合并，一直到全部合并成一大类。

使用平均联接（组间）的树状图



# 9.3 SPSS在判别分析中的应用

CONCEPT  
STRATE

## 9.3.1 判别分析的基本原理

### 1、方法概述

判别分析是判别样品所属类型的一种统计方法，其应用之广可与回归分析媲美。

判别分析与聚类分析不同。判别分析是在**已知研究对象分成若干类型**（或组别）并已取得各种类型的一批已知样品的观测数据，在此基础上根据某些准则建立判别式，然后对未知类型的样品进行判别分类。

### 2、基本原理

判别分析内容很丰富，方法很多。判别分析按判别的组数来区分，有**两组判别分析**和**多组**判别分析；按区分不同总体的所用的数学模型来分，有线性判别和非线性判别；按判别时所处理的变量方法不同，有逐步判别和序贯判别等。

其中，**距离判别分析**是一种常见的判别分析方法。它的基本思想是：首先根据已知分类的数据，分别计算各类的重心即**分组（类）的均值**，判别准则是对任给的一次观测，若它与第*i*类的重心距离最近，就认为它来自第*i*类。



## 9.3 SPSS在判别分析中的应用

例如两个总体的距离判别法中，设有两个总体（或称两类） $G_1$ 、 $G_2$ ，从第一个总体中抽取 $n_1$ 个样品，从第二个总体中抽取 $n_2$ 个样品，每个样品测量 $p$ 个指标如下页表。

今任取一个样品，实测指标值为  $X = (x_1, \dots, x_p)'$ ，问 $X$ 应判归为哪一类？

首先计算 $X$ 到 $G_1$ 、 $G_2$ 总体的距离，分别记为  $D(X, G_1)$  和  $D(X, G_2)$ ，按距离最近准则判别归类，则可写成：

$$\begin{cases} X \in G_1, \text{当} D(X, G_1) < D(X, G_2) \\ X \in G_2, \text{当} D(X, G_1) > D(X, G_2) \\ \text{待判, 当} D(X, G_1) = D(X, G_2) \end{cases}$$

然后比较  $D(X, G_1)$  和  $D(X, G_2)$  大小，按距离最近准则判别归类。

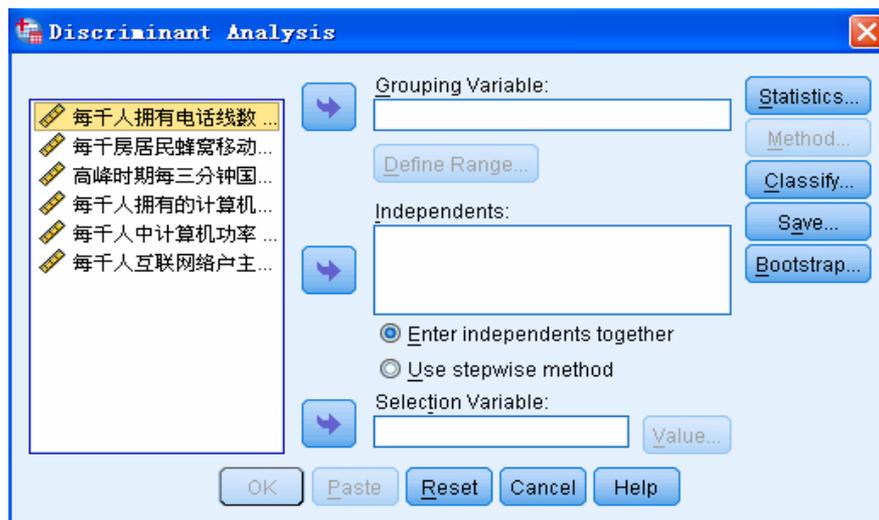


## 9.3 SPSS在判别分析中的应用

### 9.3.2 判别分析的SPSS操作详解

Step01: 打开对话框

选择菜单栏中的【Analyze（分析）】→【Classify（分类）】→【Discriminant（辨别）】命令，弹出【Discriminant Analysis（判别分析）】对话框，这是判别分析的主操作窗口。





## 9.3 SPSS在判别分析中的应用

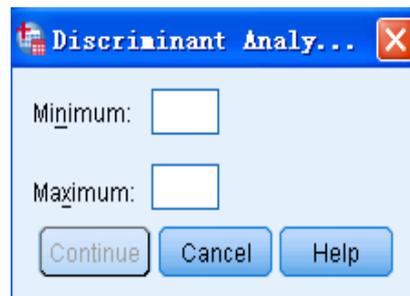
Step02: 选择判别分析变量

在【Discriminant Analysis (辨别分析)】对话框左侧的候选变量中选择进行判别分析的变量, 将其添加至【Independents (自变量)】列表框中, 将其作为自变量。

Step03: 指定分类变量及范围

在主对话框的候选变量中选择分类变量 (离散型变量) 移入【Grouping Variable (分组变量)】框中。此时它下面的【Define Range (定义范围)】按钮加亮, 单击该按钮, 屏幕弹出一个对话框, 提供指定该分类变量的数值范围。

- Minimum: 输入最小值。
- Maximum: 输入最大值。





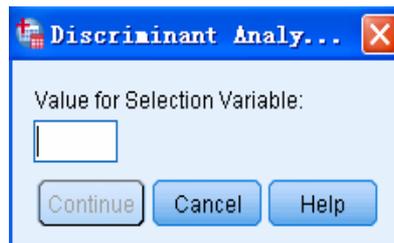
## 9.3 SPSS在判别分析中的应用

### Step04: 选择判别分析方法

在主对话框的【Independents（自变量）】列表框下面有两个按钮，它们提供了判别分析方法选择。

- Enter independent together: 建立所选择的所有变量的判别式。当认为所有自变量都能对观测量特性提供丰富的信息时使用该选择项。系统默认设置。
- Use stepwise method: 采用逐步判别法作判别分析。点选该项后，主菜单中的【Method（方法）】按钮加亮。可以进一步选择判别分析方法（见第 步）。

如果希望使用一部分观测量进行判别函数的推导，选择一个能够标记需选择的这部分观测量的变量将其移入【Selection Variables（选择变量）】框中；再单击其右侧的Value按钮，展开【Set Value（设置值）】对话框，键入能标记的变量值，如图所示。

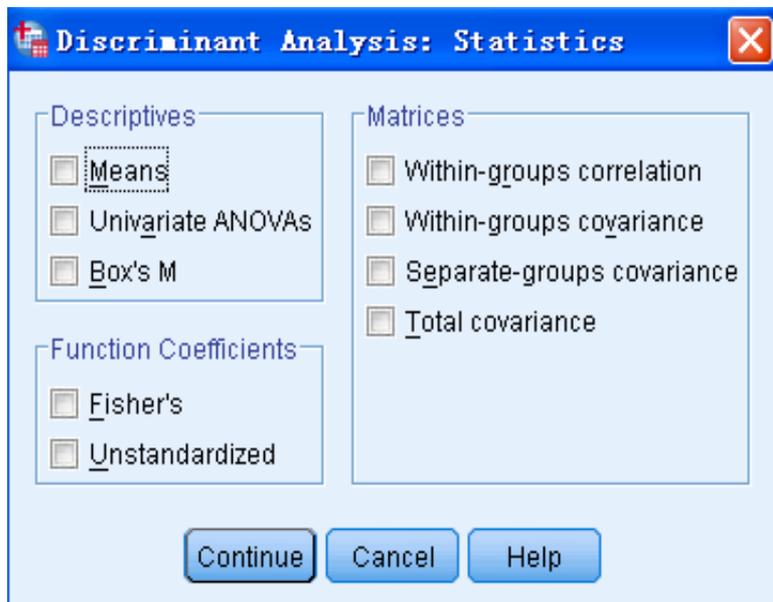




## 9.3 SPSS在判别分析中的应用

### Step05: 基本统计量输出选择

单击【Statistics】按钮，在弹出的对话框中可以选择进行判别分析的基本统计量输出。具体选项含义如下。





## 9.3 SPSS在判别分析中的应用

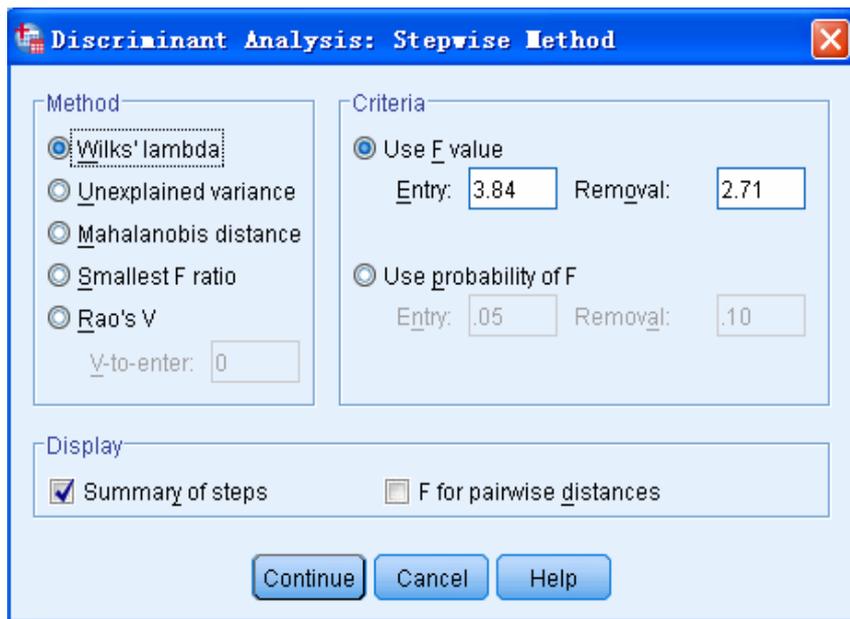
- ① **【Descriptives (描述性)】** 选项组：选择输出描述统计量。
  - Means：输出各类中各自变量的**均值**、**标准差**和各自变量总样本的均值、标准差。
  - Univariate ANOVAs：**单因素方差分析**。对各类中同一自变量进行均值检验，输出单因素方差分析结果。
  - Box' s M：对各类**协方差矩阵**相等的假设进行检验。
- ② **【Function coefficients (函数系数)】** 选项组：选择输出判别函数的系数。
  - Fisher' s：输出Fisher函数系数。对每一类给出一组系数，并给出该组中判别分数最大的观测量。
  - Unstandardized：未经标准化处理的判别函数系数。
- ③ **【Matrices (矩阵)】** 选项组：选择输出自变量的**系数矩阵**。
  - Within-groups correlation matrix：**类内相关矩阵**。
  - Within-groups covariance matrix：**类内协方差矩阵**
  - Separate-groups covariance matrices：对每一类**分别输出协方差矩阵**。
  - Total covariance matrix：总样本的协方差矩阵。



## 9.3 SPSS在判别分析中的应用

### Step06: 设置逐步判别分析选项

點選【Use stepwise method (使用步进式方法)】单选钮后, 就表示采用逐步判别法进行分析。接着单击主菜单中的【Statistics】按钮, 在弹出的对话框图中可以选择逐步判别分析的选项。具体选项含义如下。





## 9.3 SPSS在判别分析中的应用

- ① 【Method（方法）】选项组：选择变量进入判别函数的方式。
- Wilks' lambda: 每步都选择Wilks的 $\lambda$ 统计量最小的变量进入判别函数。
  - Unexplained variance: 每步都选择使类间不可解释的方差和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类间的Mahalanobis距离最大的变量进入判别函数。
  - Smallest F ratio: 每步都选择使任何两类间的“最小F值”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。选择此种方法后，应该在该项下面的【V-to-enter】文本框中输入这个增量的指定值。当某变量导致的V值增量大于指定值的变量时，该变量进入判别函数。



## 9.3 SPSS在判别分析中的应用

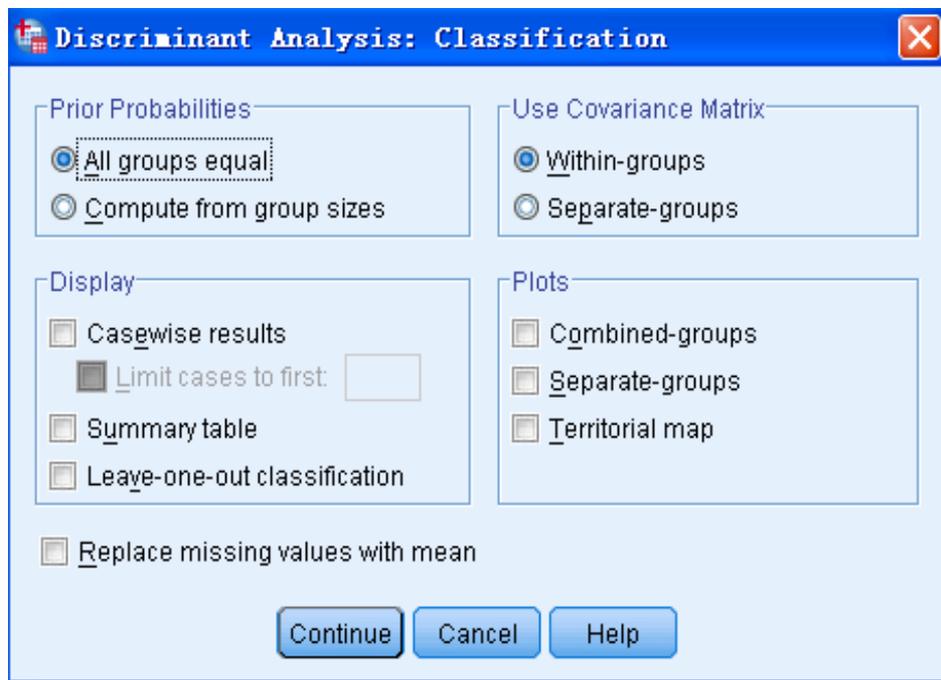
- ② **【Criteria (标准)】** 选项组：选择逐步判别停止的条件。
- **Use F value:** 使用F值，系统默认选项，当加入一个变量（或剔除一个变量）后，对在判别函数中的变量进行方差分析。当计算的F 值大于指定的Entry 值时，该变量保留在函数中。默认值是Entry 为3.84。当该变量使计算的F 值小于指定的Removal 值时，该变量从函数中剔除。默认值是Removal 为2.71。设置这两个值时应该要求Entry 值大于Removal 值。
  - **Use probability of F:** 使用F 检验的概率决定变量是否加入函数或被剔除。当计算的F 检验的概率小于指定的Entry 值时，该变量加入函数中。当该变量使计算的F 值的概率大于指定的Removal 值时，该变量从函数中剔除。
- ③ **【Display (输出)】** 栏选择逐步选择变量的过程和最后结果的显示：
- **Summary of steps:** 显示每步选择变量之后各变量的统计量结果。
  - **F for Pairwise distances:** 显示两类之间的F比值矩阵。



## 9.3 SPSS在判别分析中的应用

Step07: 设置分类参数与判别结果

单击【Classify】按钮，在弹出的对话框中可以设置判别分析的分类参数及结果。具体选项含义如下。





## 9.3 SPSS在判别分析中的应用

- ① **【Prior Probabilities (先验概率)】** 选项组：选择先验概率。
  - All groups equal: 各类先验概率相等，系统默认选项。若分为m类，则各类先验概率均为 $1/m$ 。
  - Compute from group sizes: 基于各类样本量占总样本量的比例计算先验概率。
- ② **【Use Covariance Matrix (使用协方差矩阵)】** 栏选择分类使用的协方差矩阵：
  - Within-groups: 使用合并组内协方差矩阵进行分类。
  - Separate-groups: 使用各组协方差矩阵进行分类。
- ③ **【Display (输出)】** 选项组：选择输出分类结果。
  - Casewise results: 输出每个观测量的判别分数、实际类、预测类（根据判别函数求得的分类结果）和后验概率等。选择此项后，下面的**【Limits cases to (将个案限制在前)】**项被激活，可以在它后面的文本框中输入观测量数n。选择此项则仅输出前n个观测量。
  - Summary table: 输出分类的小结表。
  - Leave-one-out classification: 输出对每一个观测量进行分类的结果，所依据的判别函数是由除该观测量以外的其他观测量导出的。



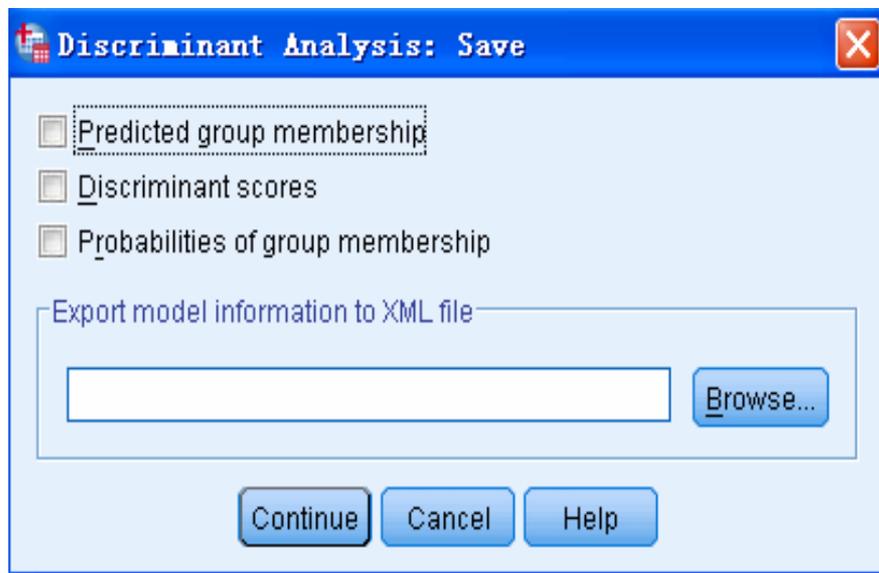
## 9.3 SPSS在判别分析中的应用

- ④ 【Plots (图)】选项组：选择输出统计图。
  - Combined-groups：生成全部类的散点图。该图是根据前两个判别函数值作的散点图。如果只有一个判别函数，就输出直方图。
  - Separate-groups：对每一类生成一张散点图。如果只有一个判别函数，就输出直方图。
  - Territorial map：生成根据判别函数值将观测量分到各类去的边界图。每一类占据一个区域。各类均值在各区中用星号标出。如果仅有一个判别函数，则不作此图。
- ⑤ 缺失值处理方式。
  - Replace missing value with mean：用该变量的均值代替缺失值。

## 9.3 SPSS在判别分析中的应用

### Step08: 结果保存设置

单击【Save】按钮，在弹出的对话框中可以设置判别分析的结果输出，具体选项含义如下。





## 9.3 SPSS在判别分析中的应用

- Predicted group membership: 建立新变量（系统默认变量名是dis\_1）保存预测观测量所属类的值。
- Discriminant score: 建立新变量保持判别分数。
- Probabilities of group membership: 建立新变量保存各个观测量属于各类的概率值。有m类，对一个观测量就会给出m个概率值，因此建立m个新变量。



## 9.3 SPSS在判别分析中的应用

### Step09 相关统计量的Bootstrap估计

单击【Bootstrap】按钮，在弹出的对话框中可以进行如下统计量的Bootstrap估计。

- 标准化典则判别函数系数表支持标准化系数的Bootstrap 估计。
- 典则判别函数系数表支持非标准化系数的Bootstrap 估计。
- 分类函数系数表支持系数的Bootstrap 估计。

Step10: 单击【OK】按钮，结束操作，SPSS软件自动输出结果。



## 9.3 SPSS在判别分析中的应用

### 9.3.3 实例分析：全国30个省市经济增长差异研究

#### 1. 实例内容

现要研究全国30个省市地区经济增长差异性，收集相关数据见数据文件9-3.sav。表中相关变量的含义分别是：x1—经济增长率（%）、x2—非国有化水平（%）、x3—开放度（%）、x4—市场化程度（%）。其中，辽宁、河北等省市归为一类，而黑龙江、吉林等省市归为另一类。请分析江苏、安徽和浙江的类别。



## 9.3 SPSS在判别分析中的应用

### 2. 实例操作

由于案例中已经将北京、上海、四川等省市按照经济增长特点分类，现在需要将另外三个待估省市：江苏、安徽和陕西分类。因此，可以利用判别分析来判别它们的归属。



## 9.3 SPSS在判别分析中的应用

### 3 实例结果及分析

#### (1) 判别分析概述表

SPSS软件首先给出了进行判别分析的概述表9-20。可以看到，参加分析的变量总数为30，有效观测量数为27，占90%；包含缺失值或分类变量范围之外的观测量数为3，占10%。

Unweighted Cases <sup>a</sup>		N <sup>a</sup>	Percent <sup>a</sup>
Valid <sup>a</sup>		27	90.0
Excluded <sup>a</sup>	Missing or out-of-range group codes <sup>a</sup>	3	10.0
	At least one missing discriminating variable <sup>a</sup>	0	.0
	Both missing or out-of-range group codes and at least one missing discriminating variable <sup>a</sup>	0	.0
	Total <sup>a</sup>	3	10.0
Total <sup>a</sup>		30	100.0



## 9.3 SPSS在判别分析中的应用

### (2) 分组统计表

下表给出了观测量按照类别不同进行的基本描述性统计量输出，其中包括均值 (Mean)、均方差 (Std. Deviation) 和有效观测量的个数等。可以从结果初步看到，不同类之间省市经济指标的差异比较明显，例如第一类省份的“非国有化水平”指标均值等于65.0282，而第二类却只有40.1081。

类别		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
1	经济增长率	15.7364	3.33175	11	11.000
	非国有化水平	65.0282	10.72709	11	11.000
	开放度	25.1491	21.26090	11	11.000
	市场化程度	74.3500	7.16398	11	11.000
2	经济增长率	11.5625	3.00397	16	16.000
	非国有化水平	40.1081	16.63743	16	16.000
	开放度	9.2281	5.94755	16	16.000
	市场化程度	58.1050	8.53527	16	16.000
Total	经济增长率	13.2630	3.72064	27	27.000
	非国有化水平	50.2607	18.96437	27	27.000
	开放度	15.7144	16.05658	27	27.000
	市场化程度	64.7233	11.31069	27	27.000



## 9.3 SPSS在判别分析中的应用

### (3) 类均值相等检验表

接着给出了不同类之间“经济增长率”等四个指标均值相等的检验结果如下表所示。从结果看到，它们的相伴概率P值都远小于显著性水平0.05，因此，可以认为两个类指标之间的均值存在显著差异，可以进行判别分析。

	Wilks' Lambda	F	df1	df2	Sig.
经济增长率	.684	11.524	1	25	.002
非国有化水平	.567	19.085	1	25	.000
开放度	.754	8.178	1	25	.008
市场化程度	.483	26.778	1	25	.000



## 9.3 SPSS在判别分析中的应用

### (4) 判别分析特征值表

下表为判别函数的特征值表。从表可见，本案例仅有一个判别函数用于分析，特征值 (Eigenvalue) 为1.479，方差百分比 (% of Variance) 为100%，方差累计百分比 (Cumulative %) 为100%，典型相关系数 (Canonical Correlation) 为0.771。

Function <sup>a</sup>	Eigenvalue <sup>a</sup>	% of Variance <sup>a</sup>	Cumulative % <sup>a</sup>	Canonical Correlation <sup>a</sup>
1 <sup>a</sup>	1.479	100.0	100.0	.772 <sup>a</sup>



## 9.3 SPSS在判别分析中的应用

### (5) Wilks' $\lambda$ 表

下表是对判别函数的显著性检验表。其中Wilks'  $\lambda$  值等于0.403，卡方统计量 (Chi-square) 等于20.878，自由度 (df) 等于4，相伴概率P值 (Sig.) 远小于显著性水平0.05，因此认为判别函数有效。

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.403	20.878	4	.000



## 9.3 SPSS在判别分析中的应用

### (6) 标准化判别函数系数

下表给出了标准化判别函数的系数，于是得到标准化判别函数如下：

Function=0.190\*经济增长率+0.242\*非国有化水平+0.360\*开放度+0.648\*市场化程度

根据判别系数看到，“市场化程度”变量对判别结果的影响是最大的，这是因为它的系数值最大，等于0.648；相反的，“经济增长率”变量对判别结果的影响最小。

	Function
	1
经济增长率	.190
非国有化水平	.242
开放度	.360
市场化程度	.648



## 9.3 SPSS在判别分析中的应用

### (7) 结构矩阵表

结构矩阵表如下表所示，是判别变量与标准化函数之间的合并类内相关系数，变量按照相关系数的绝对值大小排列，表面判别变量与判别函数之间的相关性，如变量“市场化程度”与判别函数关系最密切。

	Function
	1
市场化程度	.851
非国有化水平	.718
经济增长率	.558
开放度	.470



## 9.3 SPSS在判别分析中的应用

### (8) 非标准化判别函数系数

下表给出了非标准化判别函数系数，非标准判别函数为：  
Function=-7.263+0.060\*经济增长率+0.017\*非国有化水平+  
0.028\*开放度+0.081\*市场化程度  
根据这个判别函数代入各变量数值可以计算出判别值。

	Function
	1
经济增长率	.060
非国有化水平	.017
开放度	.025
市场化程度	.081
(Constant)	-7.263



## 9.3 SPSS在判别分析中的应用

### (9) 判别函数类心表

下表给出的是按照非标准判别函数计算的函数类心，即判别函数在各类均值处的判别分数值。可以看到，在两个类心处，判别分数值差异较大。

类别	Function
	1
1	1.411
2	-.970



## 9.3 SPSS在判别分析中的应用

### (10) 分类过程概述表

下表给出了分类过程概述情况。可以看到，共有30个观测量参与了分类过程，没有缺失变量存在。

Processed		30
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		30



## 9.3 SPSS在判别分析中的应用

### (11) 类先验概率表

下表给出了类先验概率表，按照先前的判别分析设置，先验概率都等于0.5。

类别	Prior	Cases Used in Analysis	
		Unweighted	Weighted
1	.500	11	11.000
2	.500	16	16.000
Total	1.000	27	27.000



## 9.3 SPSS在判别分析中的应用

### (12) 分类函数系数表

下表给出了Fisher线性判别函数的系数，因此可以建立各类线性判别模型。

类型一：

$$F1 = -54.567 + 1.812 * \text{经济增长率} - 0.337 * \text{非国有化水平} - 0.058 * \text{开放度} + 1.380 * \text{市场化程度}$$

类型二：

$$F2 = -36.746 + 1.669 * \text{经济增长率} - 0.377 * \text{非国有化水平} - 0.119 * \text{开放度} + 1.188 * \text{市场化程度}$$

将代判别的省市的各类经济指标代入上述两个判别函数进行计算，二者比较大小，如果 $F1 > F2$ ，对应的省市归入1类；否则，当 $F1 < F2$ ，对应的省市归入2类。



## 9.3 SPSS在判别分析中的应用

	类别	
	1	2
经济增长率	1.812	1.669
非国有化水平	-.337	-.377
开放度	-.058	-.119
市场化程度	1.380	1.188
(Constant)	-54.567	-36.746



## 9.3 SPSS在判别分析中的应用

### (13) 判别分析分类结果表

下表列出了最后判别分析的分类结果。可以看到，第一类的11个省市中，只有一个省市（广西省）判别错误，判别方法指出它应该归于第二类；同时，第二类中的16个省市全部判对。同时，数据文件中新增加变量“Dis\_1”列出了所有省市的判别结果。对于待判别省市来说，江苏和安徽被判属第一组，陕西被判属第二组，这与实际情况较吻合。



## 9.3 SPSS在判别分析中的应用

		类别	Predicted Group Membership		Total
			1	2	
Original	Count	1	10	1	11
		2	0	16	16
		Ungrouped cases	2	1	3
	%	1	90.9	9.1	100.0
		2	.0	100.0	100.0
		Ungrouped cases	66.7	33.3	100.0
a. 96.3% of original grouped cases correctly classified.					



## 第9章

# SPSS的多元统计 分析

# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

## 9.1.1 因子分析的基本原理

### 1、方法概述

人们在研究实际问题时，往往希望尽可能多的收集相关变量，以期对问题有比较全面、完整的把握和认识。

为解决这些问题，最简单和最直接的解决方案是减少变量数目，但这必然又会导致信息丢失或不完整等问题。为此，人们希望探索一种有效的解决方法，它既能减少参与数据分析的变量个数，同时也不会造成统计信息的大量浪费和丢失。

因子分析就是在尽可能不损失信息或者少损失信息的情况下，将多个变量减少为少数几个因子的方法。这几个因子可以高度概括大量数据中的信息，这样，既减少了变量个数，又同样能再现变量之间的内在联系。

# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

## 2、基本原理

通常针对变量作因子分析，称为R型因子分析；另一种对样品作因子分析，称为Q型因子分析，这两种分析方法有许多相似之处。

R型因子分析数学模型是：

设原有 $p$ 个变量  $x_1, \dots, x_p \dots$  且每个变量（或经标准化处理后）的均值为0，标准差为1。现将每个原有变量用 $k$ （ $k < p$ ）个因子  $f_1, f_2, \dots, f_k$  的线性组合来表示，即有：

$$\begin{cases} x_1 = a_{11}f_1 + a_{12}f_2 + \dots + a_{1k}f_k + \varepsilon_1 \\ x_2 = a_{21}f_1 + a_{22}f_2 + \dots + a_{2k}f_k + \varepsilon_2 \\ \dots\dots\dots \\ x_p = a_{p1}f_1 + a_{p2}f_2 + \dots + a_{pk}f_k + \varepsilon_p \end{cases}$$

上式就是因子分析的的数学模型，也可以用矩阵的形式表示为  $X = AF + \varepsilon$

# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

其中， $X$ 是可实测的随机向量。 $F$ 称为因子，由于它们出现在每个原有变量的线性表达式中，因此又称为公共因子。 $A$ 称为因子载荷矩阵， $a_{ij}(i=1,2,\dots,p; j=1,2,\dots,k)$ 称为因子载荷。 $\varepsilon$ 称为特殊因子，表示了原有变量不能被因子解释的部分，其均值为0

因子分析的基本思想是通过对变量的相关系数矩阵内部结构的分析，从中找出少数几个能控制原始变量的随机变量  $f_i(i=1,2,\dots,k)$  选取公共因子的原则是使其尽可能多的包含原始变量中的信息，建立模型  $X = A F + \varepsilon$ ，忽略  $\varepsilon$ ，以 $F$ 代替 $X$ ，用它再现原始变量 $X$ 的信息，达到简化变量降低维数的目的。

# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

## 3、基本步骤

由于实际中数据背景、特点均不相同，故采用因子分析步骤上可能略有差异，但是一个较完整的因子分析主要包括如下几个过程：

### (1) 确认待分析的原变量是否适合作因子分析

因子分析的主要任务是将原有变量的信息重叠部分提取和综合成因子，进而最终实现减少变量个数的目的。故它要求原始变量之间应存在较强的相关关系。进行因子分析前，通常可以采取计算相关系数矩阵、巴特利特球度检验和KMO检验等方法来检验候选数据是否适合采用因子分析。

### (2) 构造因子变量

将原有变量综合成少数几个因子是因子分析的核心内容。它的关键是根据样本数据求解因子载荷阵。因子载荷阵的求解方法有基于主成分模型的主成分分析法、基于因子分析模型的主轴因子法、极大似然法等。

# 9.1 SPSS在因子分析中的应用



CONCEPT  
RATE

## (3) 利用旋转方法使因子变量更具有可解释性

将原有变量综合为少数几个因子后，如果因子的实际含义不清，则不利于后续分析。为解决这个问题，可通过因子旋转的方式使一个变量只在尽可能少的因子上有比较高的载荷，这样使提取出的因子具有更好的解释性。

## (4) 计算因子变量得分

实际中，当因子确定以后，便可计算各因子在每个样本上的具体数值，这些数值称为因子得分。于是，在以后的分析中就可以利用因子得分对样本进行分类或评价等研究，进而实现了降维和简化问题的目标。

# 9.1 SPSS在因子分析中的应用

CONCEPT  
RATE

根据上述步骤，可以得到进行因子分析的详细计算过程如下。

- ①将原始数据标准化，以消除变量间在数量级和量纲上的不同。
- ②求标准化数据的相关矩阵。
- ③求相关矩阵的特征值和特征向量。
- ④计算方差贡献率与累积方差贡献率。
- ⑤确定因子：设 $F_1, F_2, \dots, F_p$ 为 $p$ 个因子，其中前 $m$ 个因子包含的数据信息总量（即其累积贡献率）不低于85%时，可取前 $m$ 个因子来反映原评价指标。
- ⑥因子旋转：若所得的 $m$ 个因子无法确定或其实际意义不是很明显，这时需将因子进行旋转以获得较为明显的实际含义。
- ⑦用原指标的线性组合来求各因子得分。
- ⑧综合得分：通常以各因子的方差贡献率为权，由各因子的线性组合得到综合评价指标函数。

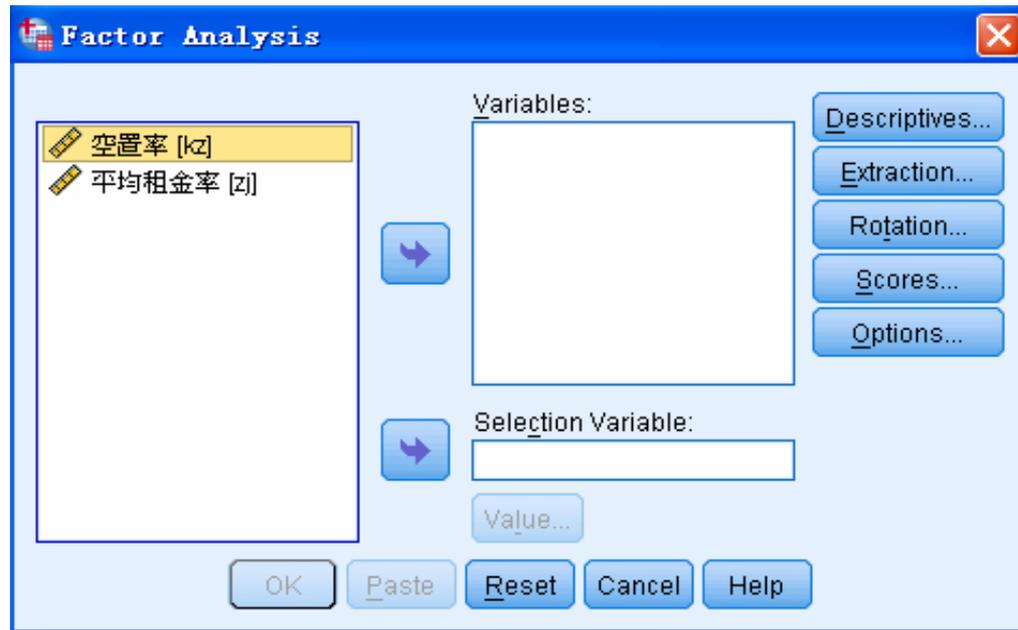
# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

## 9.1.2 因子分析的SPSS操作详解

Step01: 打开对话框

选择菜单栏中的【Analyze（分析）】→【Data Reduction（降维）】→【Factor（因子）】命令，弹出【Factor Analysis（因子分析）】对话框，这是因子分析的主操作窗口。



# 9.1 SPSS在因子分析中的应用



CONCEPT  
RATE

## Step02: 选择因子分析变量

在【Factor Analysis（因子分析）】对话框左侧的候选变量列表框中选择进行因子分析的变量，将其添加至【Variables（变量）】列表框中。如果要选择参与因子分析的样本，则需要将条件变量添加至【Selection Variable（选择变量）】列表框中，并单击【Value】按钮输入变量值，只有满足条件的样本数据才能进行后续的因子分析。

## Step03: 选择描述性统计量

单击【Descriptives】按钮，在弹出的对话框中可以选择输出描述性统计量及相关矩阵等内容。

# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

具体选项含义如下：

① 【Statistics (统计量)】选项组

- Univariate descriptives: 单变量描述统计量，即输出参与分析的各原始变量的均值、标准差等。
- Initial solution: 初始分析结果，系统默认项。输出各个分析变量的初始共同度、特征值以及解释方差的百分比等。

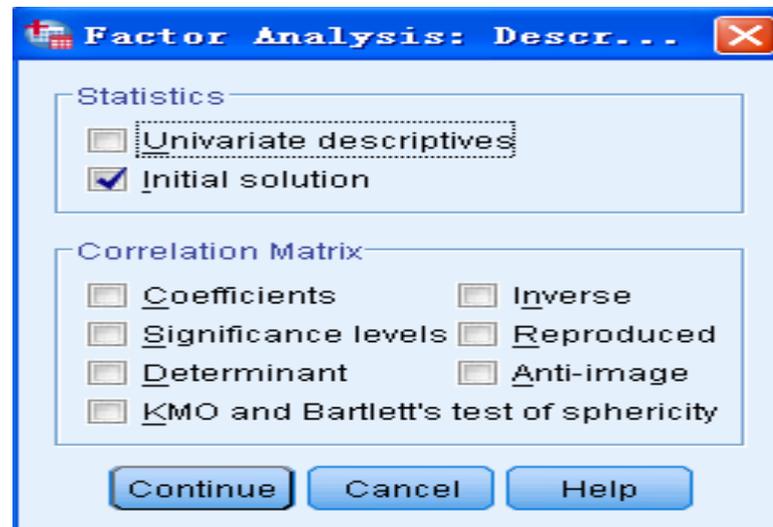
② 【Correlation Matrix (相关矩阵)】选项组

- Coefficients: 原始分析变量间的相关系数矩阵。
- Significance levels: 显著性水平。输出每个相关系数相对于相关系数为0的单尾假设检验的概率水平。
- 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以及残差阵。
- Anti-image: 象相关阵。包括偏相关系数的负数以及偏协方差的负数。在一个好的因子模型中，除对角线上的系数较大外，远离对角线的元素应该比较小。

## 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

KMO and Bartlett's test of sphericity: KMO 和 Bartlett 检验。前者输出抽样充足度的Kaisex-Meyer-Olkin 测度,用于检验变量间的偏相关是否很小。后者Bartlett 球度方法检验相关系数阵是否是单位阵。如果是单位阵,则表明因子模型不合适采用因子模型。



# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

## Step04: 选择因子提取方法

单击【 Extract (抽取) 】按钮，在弹出的对话框中可以选择提取因子的方法及相关选项。

- ① 在【Method (方法)】框下拉列表框中可以选择因子提取方法。
  - Principal components: 主成份分析法。该方法假设变量是因子的纯线性组合。第一成分有最大的方差，后续的成分其可解释的方差逐个递减。
  - Unweighted least square : 不加权最小二乘法。
  - Generalized least squares : 加权最小二乘法。
  - Maximum likelihood : 极大似然法。
  - Principal axis factoring : 主轴因子提取法。
  - Alpha factoring:  $\alpha$  因子提取法。
  - Image factoring: 映象因子提取法。

# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

## ② 【Analyze（分析）】选项组

- Correlation matrix: 相关系数矩阵, 系统默认项。
- Covariance matrix: 协方差矩阵。

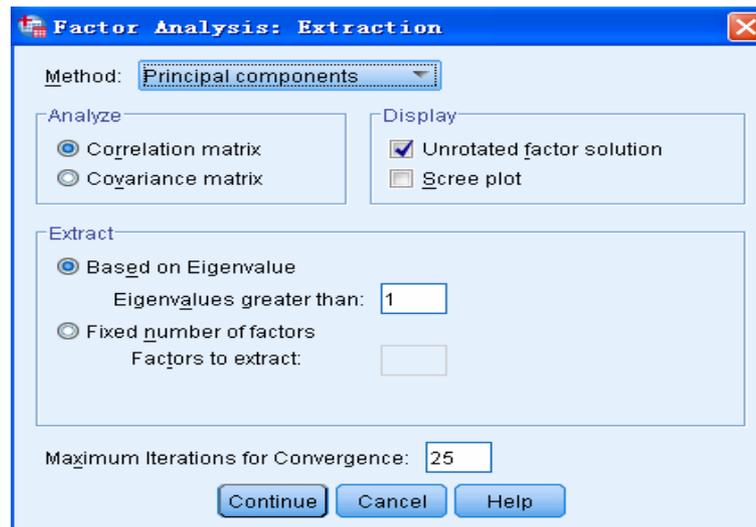
## ③ 【Display（输出）】选项组: 输出与因子提取有关的选项。

- Unrotated factor solution: 输出未经旋转的因子提取结果。此项为系统默认的输出方式。
- Scree plot: 输出因子的碎石图。它显示了按特征值大小排列的因子序号。它有助于确定保留多少个因子。典型的碎石图会有一个明显的拐点, 在该点之前是与大因子连接的陡峭的折线, 之后是与小因子相连的缓坡折线。

# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

- ④ **【Extract (抽取)】** 选项组：输出与提取结果有关的选择项。由于理论上因子数目与原始变量数目相等，但因子分析的目的是用少量因子代替多个原始变量，选择提取多少个因子是由本栏来决定。
- **Eigenvalues over:** 指定提取的因子的特征值数目。在此项后面的矩形框中给出输入数值（系统默认值为1），即要求提取那些特征值大于1的因子。
- **Number of factors:** 指定提取公因子的数目。用鼠标单击选择此项后，将指定其数目。
- ⑤ **Maximum iterations for Convergence:** 在对应的文本框中指定因子分析收敛的最大迭代次数。系统默认的最大迭代次数为25。

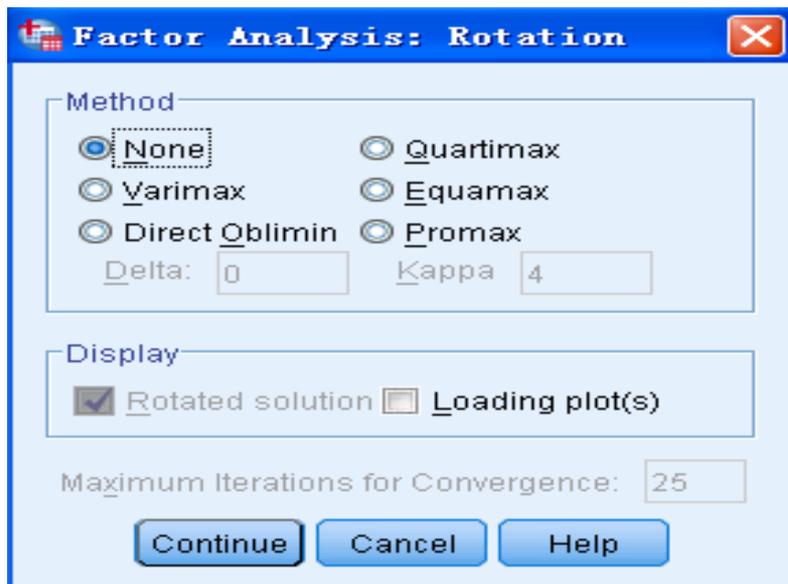


## 9.1 SPSS在因子分析中的应用



Step05: 选择因子旋转方法

单击【Rotation】按钮，在弹出的对话框可以选择因子旋转方法及  
相关选项。



# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

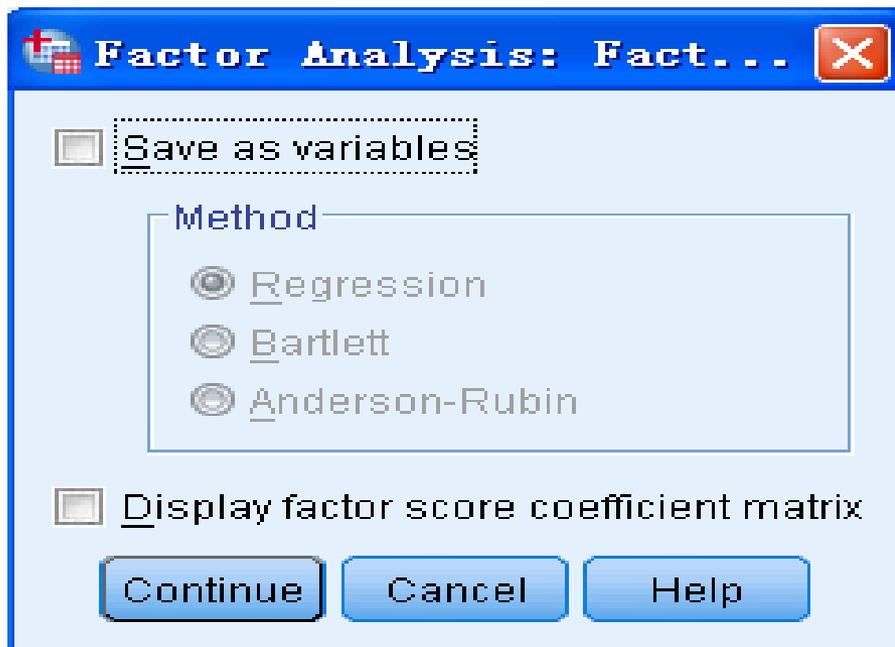
- ① **【Method（方法）】** 选项组选择旋转方法。
  - None: 不进行旋转, 此为系统默认的选择项。
  - Varimax: 方差最大旋转法。这是一种正交旋转方法。它使每个因子具有最高载荷的变量数最小, 因此可以简化对因子的解释。
  - Direct Oblimin: 直接斜交旋转法。指定此项可以在下面的“Delta”矩形框中键入  $\delta$  值, 该值应该在0~1 之间。系统默认的  $\delta$  值为0。
  - Quartma: 四次方最大正交旋转法。该旋转方法使每个变量中需要解释的因子数最少。
  - Equamax: 平均正交旋转法。
  - Promax: 斜交旋转方法。允许因子彼此相关。它比直接斜交旋转更快, 因此适用于大数据集的因子分析。指定此项可以在下面的“Kappa”矩形框中键入“ $\kappa$ ”值, 默认为4 (此值最适合于分析)。
- ② **【Display（输出）】** 选项组: 选择有关输出显示。
  - Rotated solution: 旋转解。在Method栏中指定旋转方法才能选择此项。
  - Loading plot(s): 因子载荷散点图。指定此项将给出以前两因子为坐标轴的各变量的载荷散点图。
- ③ **Maximum iterations for Convergence:** 可以指定旋转收敛的最大迭代次数。系统默认值为25。可以在此项后面的文本框中输入指定值。

## 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

Step06: 选择因子得分

单击【Scores】按钮，在弹出的对话框中可以选择因子得分方法及相关选项。具体选项含义如下。



# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

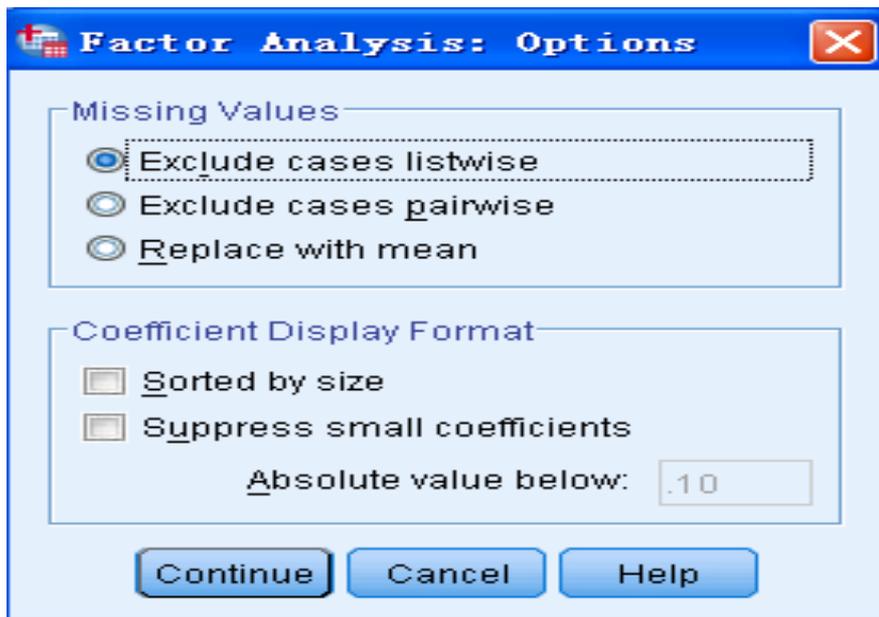
- ① 【Save as variables (保存为变量)】选项组：将因子得分作为新变量保存在数据文件中。
    - Save as variables: 将因子得分作为新变量保存在工作数据文件中。程序运行结束后，在数据窗中显示出新变量。
  - ② 【Method (方法)】选项组：指定计算因子得分的方法。
    - Regression: 回归法。选择此项，其因子得分的均值为0。方差等于估计的因子得分与实际因子得分值之间的复相关系数的平方。
    - Bartlett: 巴特利特法。选择此项，因子得分均值为0。超出变量范围各因子平方和被最小化。
    - Anderson-Rubin: 安德森-鲁宾法。选择此项，是为了保证因子的正交性。
- 本例选中“Regression”项。
- ③ 在输出窗中显示因子得分。
    - Display factor score coefficient matrix: 输出因子得分系数矩阵。

## 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

Step07: 其他选项输出

单击【Options】按钮，在弹出的对话框中可以选择一些附加输出项。具体选项含义如下。



# 9.1 SPSS在因子分析中的应用

CONCEPT  
STRATE

- ① **【MissingValues (缺失值)】** 选项组：选择处理缺失值方法。
  - Exclude cases listwise: 分析变量中带有缺失值的观测量都不参与后续分析。
  - Exclude cases pairwise: 成对剔除带有缺失值的观测量。
  - Replace with mean: 用该变量的均值代替工作变量的所有缺失值。
- ② **【Coefficient Display Format (系数显示格式)】** 选项组：选择载荷系数的显示格式。
  - Sorted by size: 将载荷系数按其大小排列构成矩阵，使在同一因子上具有较高载荷的变量排在一起。便于得出结论。
  - Suppress absolute values less than: 不显示那些绝对值小于指定值的载荷系数。选择此项后还需要在该项的参数框中键入0~1之间的数值作为临界值。系统默认的临界值为0.10。

Step08: 单击 **【OK】** 按钮，结束操作，SPSS软件自动输出结果。



# 9.1 SPSS在因子分析中的应用

## 9.1.3 实例分析：居民消费结构的变动

### 1. 实例内容

消费结构是指在消费过程中各项消费支出占居民总支出的比重。它是反映居民生活消费水平、生活质量变化状况以及内在过程合理化程度的重要标志。而消费结构的变动不仅是消费领域的重要问题，而且也关系到整个国民经济的发展。因为合理的消费结构及消费结构的升级和优化不仅反映了消费的层次和质量的提高，而且也建立了合理的产业结构和产品结构提供了重要的依据。

表9-1是某市居民生活费支出费用，具体分为食品、衣着、家庭设备用品及服务、医疗保健、交通通讯、文教娱乐及服务、居住和杂项商品与服务等8个部分。请利用因子分析探讨该市居民消费结构，为产业政策的制定和宏观经济的调控提供参考。



# 9.1 SPSS在因子分析中的应用

## 2. 实例操作

数据文件9-1.sav是某市居民在食品、衣着、医疗保健等八个方面的消费数据，这些指标之间存在着不同强弱的相关性。如果单独分析这些指标，无法能够分析居民消费结构的特点。因此，可以考虑采用因子分析，将这八个指标综合为少数几个因子，通过这些公共因子来反映居民消费结构的变动情况。



# 9.1 SPSS在因子分析中的应用

## 3. 实例结果及分析

### (1) 描述性统计表

下表显示了食品、衣着等这八个消费支出指标的描述统计量，例如均值、标准差等。这为后续的因子分析提供了一个直观的分析结果。可以看到，食品支出消费所占的比重最大，其均值等于39.4750%，其次是文化娱乐服务支出消费和交通通信支出消费。所有的消费支出中，医疗保健消费支出占的比重最低。



## 9.1 SPSS在因子分析中的应用

	Mean	Std. Deviation	Analysis N
食品	39.4750	2.29705	8
衣着	6.4875	.86592	8
家庭设备用品及服务	7.9125	2.87772	8
医疗保健	6.3625	1.54729	8
交通和通信	8.1750	2.61302	8
文化娱乐服务	14.4750	2.30016	8
居住	12.1625	2.91545	8
杂项商品与服务	2.9125	.52491	8



## 9.1 SPSS在因子分析中的应用

### (2) 因子分析共同度

下表是因子分析的共同度，显示了所有变量的共同度数据。第一列是因子分析初始解下的变量共同度。它表明，对原有八个变量如果采用主成分分析法提取所有八个特征根，那么原有变量的所有方差都可被解释，变量的共同度均为1（原有变量标准化后的方差为1）。

事实上，因子个数小于原有变量的个数才是因子分析的目的，所以不可能提取全部特征根。于是，第二列列出了按指定提取条件（这里为特征根大于1）提取特征根时的共同度。可以看到，所有变量的绝大部分信息（全部都大于83%）可被因子解释，这些变量信息丢失较少。因此本次因子提取的总体效果理想。



## 9.1 SPSS在因子分析中的应用

	Initial	Extraction
食品	1.000	.842
衣着	1.000	.842
家庭设备用品及服务	1.000	.976
医疗保健	1.000	.954
交通和通信	1.000	.925
文化娱乐服务	1.000	.953
居住	1.000	.978
杂项商品与服务	1.000	.947



# 9.1 SPSS在因子分析中的应用

## (3) 因子分析的总方差解释

接着Spss软件计算得到相关系数矩阵的特征值、方差贡献率及累计方差贡献率结果如表9-4所示。在下一页表中，第一列是因子编号，以后三列组成一组，组中数据项的含义依次是特征根、方差贡献率和累计贡献率。

第一组数据项（第二至第四列）描述了初始因子解的情况。可以看到，第一个因子的特征根值为4.316，解释了原有8个变量总方差的53.947%。前三个因子的累计方差贡献率为94.196%，并且只有它们的取值大于1。说明前3个公因子基本包含了全部变量的主要信息，因此选前3个因子为主因子即可。

同时，Extraction Sums of Squared Loadings和Rotation Sums of Squared Loadings部分列出了因子提取后和旋转后的因子方差解释情况。从表中看到，它们都支持选择3个公共因子。



# 9.1 SPSS在因子分析中的应用

## 因子分析的总方差解释

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	4.316	53.947	53.947	4.316	53.947	53.947	4.261	53.265	53.265
2	1.989	24.869	78.816	1.989	24.869	78.816	2.030	25.379	78.645
3	1.230	15.380	94.196	1.230	15.380	94.196	1.244	15.551	94.196
4	0.275	3.435	97.631						
5	0.122	1.524	99.155						
6	0.052	0.648	99.804						
7	0.016	0.196	100.000						
8	1.790E-17	2.237E-16	100.000						



## 9.1 SPSS在因子分析中的应用

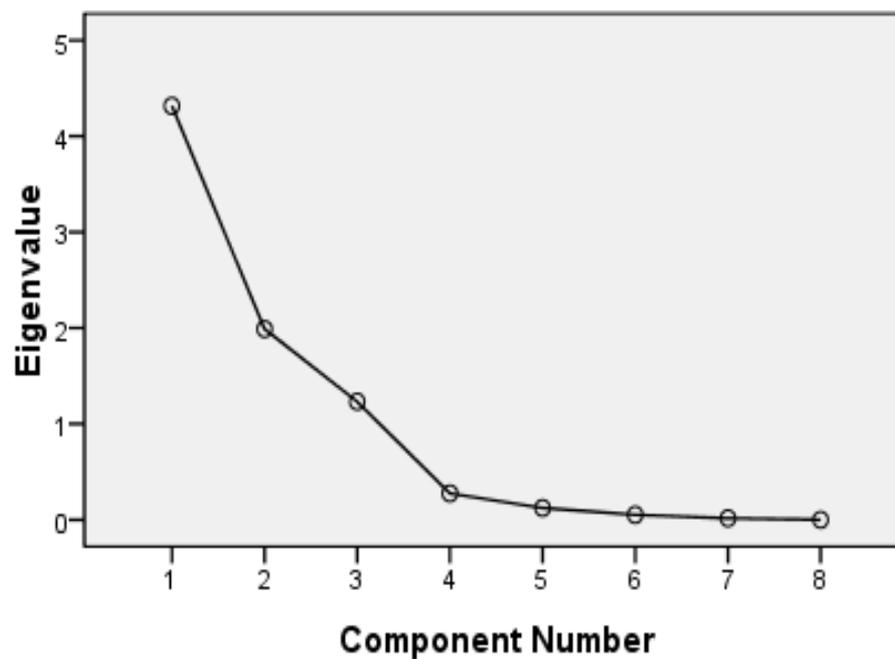
### (4) 因子碎石图

下图为因子分析的碎石图。横坐标为因子数目，纵坐标为特征根。可以看到，第一个因子的特征值很高，对解释原有变量的贡献最大；第三个以后的因子特征根都较小，取值都小于1，说明它们对解释原有变量的贡献很小，称为可被忽略的“高山脚下的碎石”，因此提取前三个因子是合适的。



# 9.1 SPSS在因子分析中的应用

Scree Plot





## 9.1 SPSS在因子分析中的应用

### (5) 旋转前的因子载荷矩阵

下表中显示了因子载荷矩阵，它是因子分析的核心内容。通过载荷系数大小可以分析不同公共因子所反映的主要指标的区别。从结果看，大部分因子解释性较好，但是仍有少部分指标解释能力较差，例如“食品”指标在三个因子的载荷系数区别不大。因此接着采用因子旋转方法使得因子载荷系数向0或1两极分化，使大的载荷更大，小的载荷更小。这样结果更具可解释性。



## 9.1 SPSS在因子分析中的应用

旋转前的因子载荷矩阵

	Component		
	1	2	3
医疗保健	0.967	0.102	0.093
文教娱乐及服务	0.962	0.144	-0.085
交通和通信	0.948	-0.082	0.140
家庭设备用品及服务	-0.833	0.503	-0.173
食品	-0.761	0.202	0.471
居住	0.008	-0.970	-0.190
衣着	0.527	0.826	-0.005
杂项商品与服务	0.081	-0.183	0.952



# 9.1 SPSS在因子分析中的应用

## (6) 旋转后的因子载荷矩阵

下表中显示了实施因子旋转后的载荷矩阵。可以看到，第一主因子在“交通和通信”和“医疗保健”等五个指标上具有较大的载荷系数，第二主因子在“居住”和“衣着”指标上系数较大，而第三主因子在“杂项商品与服务”上的系数最大。此时，各个因子的含义更加突出。



## 9.1 SPSS在因子分析中的应用

实施因子旋转后的载荷矩阵

	Component		
	1	2	3
交通和通讯	0.946	0.083	0.152
医疗保健	0.938	0.260	0.081
文教娱乐及服务	0.931	0.277	-0.101
家庭设备用品及服务	-0.895	0.343	-0.241
食品	-0.793	0.144	0.438
居住	0.159	-0.974	-0.058
衣着	0.396	0.889	-0.114
杂项商品与服务	0.086	-0.041	0.968



## 9.1 SPSS在因子分析中的应用

可以看出第一个公因子主要反映了交通和通信、医疗保健、文化娱乐服务、家庭设备用品及服务、食品上有较大载荷，说明第一个公因子综合反映这几个方面的变动情况，可以将其命名为第一基本生活消费因子，即享受性消费因子。

第二个公因子在居住、衣着上的载荷系数较大，代表了这两个方面的变动趋势，可以将其命名为第二基本生活消费因子，即发展性消费因子。

第三个公因子在杂项商品与服务上的消费变动较大，因此可以将第三个公因子命名为第三基本生活消费因子，即其他类型消费因子。



## 9.1 SPSS在因子分析中的应用

### (7) 因子得分系数

下表中列出了采用回归法估计的因子得分系数。根据表中内容可写出以下因子得分函数：

因子 $F_1 = -0.198X_1 + 0.058X_2 - 0.226X_3 + 0.212X_4 + 0.221X_5 + 0.211X_6 + 0.079X_7 + 0.015X_8$ ;

因子 $F_2 = 0.123X_1 + 0.425X_2 + 0.200X_3 + 0.094X_4 + 0.008X_5 + 0.096X_6 - 0.498X_7 + 0.015X_8$ ;

因子 $F_3 = 0.365X_1 - 0.059X_2 - 0.174X_3 + 0.069X_4 + 0.119X_5 - 0.077X_6 - 0.088X_7 + 0.779X_8$ ;



## 9.1 SPSS在因子分析中的应用

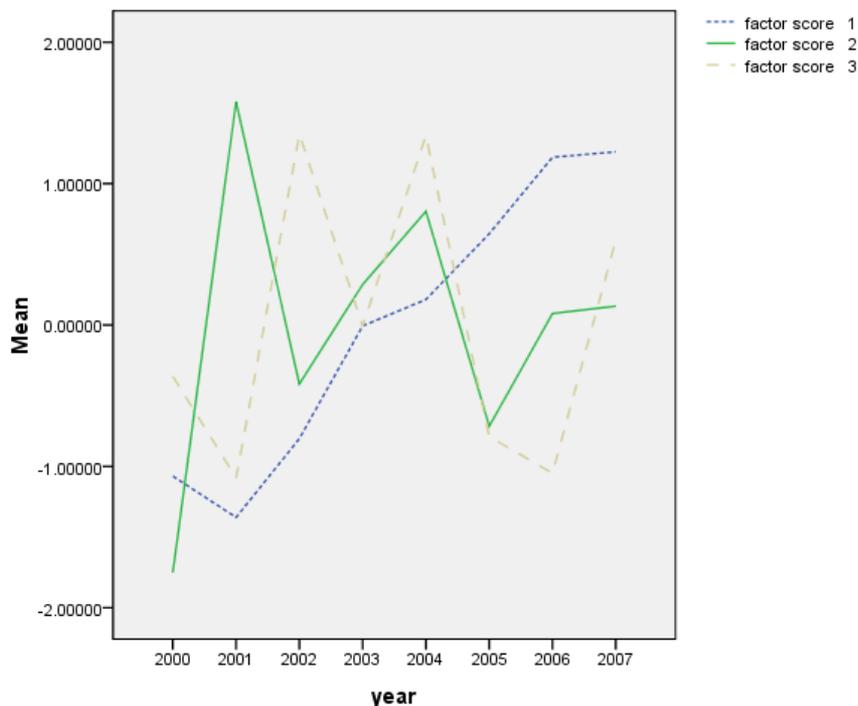
因子得分系数

	Component		
	1	2	3
食品	-0.198	0.123	0.365
衣着	0.058	0.425	-0.059
家庭设备用品及服务	-0.226	0.200	-0.174
医疗保健	0.212	0.094	0.069
交通和通讯	0.221	0.008	0.119
文教娱乐及服务	0.211	0.096	-0.077
居住	0.079	-0.498	-0.088
杂项商品与服务	0.015	0.015	0.779



## 9.1 SPSS在因子分析中的应用

不仅如此，原数据文件中增加了FAC1\_1、FAC2\_1和FAC3\_1三个变量，它们表示了三个因子在不同年份的得分值。为了进一步揭示因子的变动情况，绘制了如下图所示的因子变动趋势图。



## 9.2 SPSS在聚类分析中的应用

CONCEPT  
STRATE

### 9.2.1 聚类分析的基本原理

#### 1、方法概述

聚类分析又称群分析，它是研究（样品或指标）分类问题的一种多元统计方法，所谓类，通俗地说，就是指相似元素的集合。

#### 2、聚类分析的分类

根据分类对象的不同可分为样品聚类和变量聚类。

##### (1) 样品聚类

样品聚类在统计学中又称为Q型聚类。用SPSS的术语来说就是对事件(Cases)进行聚类，或是说对观测量进行聚类。它是根据被观测的对象的各种特征，即反映被观测对象的特征的各项变量值进行分类。

- 由上图可以看出，在2000~2007年期间，第一公因子除了开始阶段有些下降外，此后每年都在逐步回升，并于2006年达到最高点。这主要是由于前几年国企改革和中国经济的软着陆，下岗职工大量增加，因此这段时间人们在享受性消费上的支出是减少的，而在其他基本生活消费上的支出增加。而随着经济的发展和收入的增加，享受性消费逐步增加，其他生活消费由于享受性消费的突然增加而减少后也会逐渐增加。第二公因子得分的起伏波动主要是由市民住房比重有升有降的变动引起的，根本原因还是和国家执行住房改革的力度密切相关，但由于住房改革政策的推行相对于其他政策而言较为缓慢，所以市民对住房消费存在一定的不确定性，这就造成了住房比重在总消费中的升降变化。第三公因子一直波动不已，这说明市民在杂项上的消费仍有较大的发展空间。



## 9.2 SPSS在聚类分析中的应用

### (2) 变量聚类

变量聚类在统计学又称为R型聚类。反映同一事物特点的变量有很多，我们往往根据所研究的问题选择部分变量对事物的某一方面进行研究。由于人类对客观事物的认识是有限的，往往难以找出彼此独立的有代表性的变量，而影响对问题的进一步认识和研究。例如在回归分析中，由于自变量的共线性导致偏回归系数不能真正反映自变量对因变量的影响等。因此往往先要进行变量聚类，找出彼此独立且有代表性的自变量，而又不丢失大部分信息。

值得提出的是将聚类分析和其它方法联合起来使用，如判别分析、主成分分析、回归分析等往往效果更好。



## 9.2 SPSS在聚类分析中的应用

### 3、距离和相似系数

为了将样品（或指标）进行分类，就需要研究样品之间关系。目前用得最多的方法有两个：一种方法是用相似系数，性质越接近的样品，它们的相似系数的绝对值越接近1，而彼此无关系的样品，它们的相似系数的绝对值越接近于零。比较相似的样品归为一类，不怎么相似的样品归为不同的类。另一种方法是将一个样品看作P维空间的一个点，并在空间定义距离，距离越近的点归为一类，距离较远的点归为不同的类。但相似系数和距离有各种各样的定义，而这些定义与变量的类型关系极大。

$$d_{ij}(q) = \left( \sum_{a=1}^p |x_{ia} - x_{ja}|^q \right)^{1/q}$$



## 9.2 SPSS在聚类分析中的应用

常用的距离和相似系数定义如下：

### (1) 距离

如果把n个样品（X中的n个行）看成p维空间中n个点，则两个样品间相似程度可用p维空间中两点的距离来度量。令 $d_{ij}$ 表示样品 $X_i$ 与 $X_j$ 的距离。常用的距离有：

明氏（Minkowski）距离  $d_{ij}(q) = \left( \sum_{a=1}^p |x_{ia} - x_{ja}|^q \right)^{1/q}$

当 $q=1$ 时

$$d_{ij}(\infty) = \max_{1 \leq a \leq p} |x_{ia} - x_{ja}|$$

即绝对距离

当 $q=2$ 时

$$d_{ij}(2) = \left( \sum_{a=1}^p (x_{ia} - x_{ja})^2 \right)^{1/2}$$

即欧氏距离

当  $q=\infty$  时

$$d_{ij}(1) = \sum_{a=1}^p |x_{ia} - x_{ja}|$$

即切比雪夫距离



## 9.2 SPSS在聚类分析中的应用

马氏 (Mahalanobis) 距离

$$d_{ij}^2(M) = (X_i - X_j)' \Sigma^{-1} (X_i - X_j)$$

其中  $\Sigma$  表示指标的协差阵, 即:

$$\Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \dots & \sigma_{1p} \\ \sigma_{21} & \sigma_{22} & \dots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \dots & \sigma_{pp} \end{pmatrix} \quad p \times p$$
$$\sigma_{ij} = \frac{1}{n-1} \sum_{a=1}^n (x_{ai} - \bar{x}_i)(x_{aj} - \bar{x}_j) \quad i, j = 1, \dots, p$$
$$\bar{x}_i = \frac{1}{n} \sum_{a=1}^n x_{ai} \quad \bar{x}_j = \frac{1}{n} \sum_{a=1}^n x_{aj}$$

马氏距离既排除了各指标之间相关性的干扰, 而且还不受各指标量纲的影响。除此之外, 它还有一些优点, 如可以证明, 将原数据作一线性交换后, 马氏距离仍不变等等。



## 9.2 SPSS在聚类分析中的应用

兰氏 (Canberra) 距离

它是由Lance和Williams最早提出的，故称兰氏距离。

$$d_{ij}(L) = \frac{1}{p} \sum_{a=1}^p \frac{|x_{ia} - x_{ja}|}{x_{ia} + x_{ja}} \quad i, j = 1, \dots, n$$

此距离仅适用于一切的情况，这个距离有助于克服各指标之间量纲的影响，但没有考虑指标之间的相关性。



## 9.2 SPSS在聚类分析中的应用

### (2) 相似系数

研究样品之间的关系，除了用距离表示外，还有相似系数，顾名思义，相似系数是描写样品之间相似程度的一个量，常用的相似系数有：

#### ● 夹角余弦

将任何两个样品  $X_i$  与  $X_j$  看成p维空间的两个向量，这两个向量的夹角余弦用  $\cos\theta_{ij}$  表示。则

$$\cos\theta_{ij} = \frac{\sum_{a=1}^p x_{ia}x_{ja}}{\sqrt{\sum_{a=1}^p x_{ia}^2 \cdot \sum_{a=1}^p x_{ja}^2}} \quad 1 \leq \cos\theta_{ij} \leq 1$$

当  $\cos\theta_{ij} = 1$ ，说明两个样品  $X_i$  与  $X_j$  完全相似； $\cos\theta_{ij}$  接近1，说明  $X_i$  与  $X_j$  相似密切； $\cos\theta_{ij} = 0$ ，说明  $X_j$  与  $X_i$  完全不一样； $\cos\theta_{ij}$  接近0，说明  $X_i$  与  $X_j$  差别大。



## 9.2 SPSS在聚类分析中的应用

### ● 相关系数

通常所说相关系数，一般指变量间的相关系数，作为刻画样品间的相似关系也可类似给出定义，即第*i*个样品与第*j*个样品之间的相关系数定义为：

$$r_{ij} = \frac{\sum_{a=1}^p (x_{ia} - \bar{x}_i)(x_{ja} - \bar{x}_j)}{\sqrt{\sum_{a=1}^p (x_{ia} - \bar{x}_i)^2 \cdot \sum_{a=1}^p (x_{ja} - \bar{x}_j)^2}} \quad -1 \leq r_{ij} \leq 1$$

其中

$$\bar{x}_i = \frac{1}{p} \sum_{a=1}^p x_{ia}$$

$$\bar{x}_j = \frac{1}{p} \sum_{a=1}^p x_{ja}$$

聚类分析内容非常丰富，有系统聚类法、有序样品聚类法、动态聚类法、模糊聚类法等。本节主要介绍使用较多的快速聚类法和系统聚类法。



## 9.2 SPSS在聚类分析中的应用

### 9.2.2 快速聚类法的SPSS操作详解

K-均值聚类法又叫快速聚类法，可以用于大量数据进行聚类分析的情形。它是一种非分层的聚类方法。这种方法占用内存少、计算量、处理速度快，特别适合大样本的聚类分析。它的基本操作步骤如下：

- 1、指定聚类数目 $k$ ，应由用户指定需要聚成多少类，最终也只能输出关于它的唯一解。这点不同于层次聚类。
- 2、确定 $k$ 个初始类的中心。两种方式：一种为用户指定方式，二是根据数据本身结构的中心初步确定每个类别的原始中心点。
- 3、根据距离最近原则进行分类。逐一计算每一记录到各个中心点的距离，把各个记录按照距离最近的原则归入各个类别，并计算新形成类别的中心点
- 4、按照新的中心位置，重新计算每一记录距离新的类别中心点的距离，并重新进行归类。
- 5、重复步骤4，直达到达到一定的收敛标准。

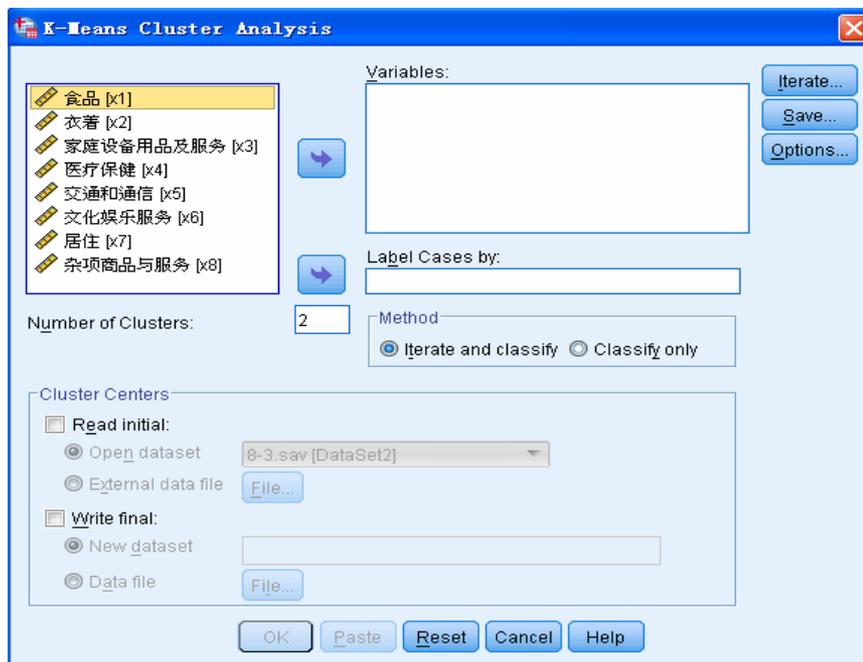
这种方法也常称为逐步聚类分析，即先把被聚对象进行初始分类，然后逐步调整，得到最终分类。



## 9.2 SPSS在聚类分析中的应用

Step01: 打开对话框

选择菜单栏中的【Analyze（分析）】→【Classify（分类）】→【K-Means Cluster（K均值聚类）】命令，弹出【K-Means Cluster Analysis（K均值聚类分析）】对话框，这是快速聚类分析的主操作窗口。





## 9.2 SPSS在聚类分析中的应用

### Step02: 选择聚类分析变量

在【K-Means Cluster Analysis (K均值聚类分析)】对话框左侧的候选变量列表框中选择进行聚类分析的变量, 将其添加至【Variables (变量)】列表框中。同时可以选择一个标识变量移入【Label Cases by (个案标记依据)】列表框中。

### Step03: 确定分类个数

在【Number of Clusters (聚类数)】列表框中, 可以输入确定的聚类分析数目, 用户可以根据需要自行修改调整。系统默认的聚类数为2。

### Step04: 选择聚类方法

在【Method (方法)】下拉列表框中可以选择聚类方法。系统默认值选择【Iterative and classify (迭代与分类)】项。

- Iterate and classify: 选择初始类中心, 在迭代过程中不断更新聚类中心。把观测量分派到与之最近的以类中心为标志的类中去。
- Classify only: 只使用初始类中心对观测量进行分类, 聚类中心始终不变。



## 9.2 SPSS在聚类分析中的应用

### Step05: 聚类中心的输入与输出

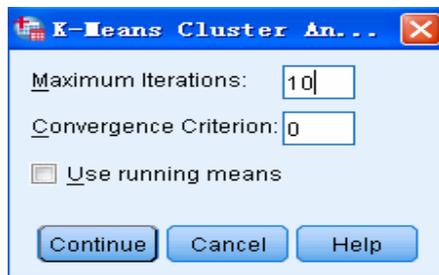
在主对话框中，【Cluster Centers（聚类中心）】选项组表示输入和输出聚类中心。用户可以指定外部文件或数据集作为初始聚类中心点，也可以将聚类分析的聚类中心结果输出到指定文件或数据集中。

- Read initial: 要求使用指定数据文件中的观测量或建立数据集作为初始类中心。
- Write final as File: 要求把聚类结果中的各类中心数据保存到指定的文件或数据集中。



## 9.2 SPSS在聚类分析中的应用

在主对话框中单击Iterate（迭代）按钮，打开设置迭代参数的对话框图，这里可以进一步选择迭代参数。



- **Maximum Iterations:** 输入K-Means 算法中的迭代次数。改变后面参数框中的数字，则改变迭代次数。当达到限定的迭代次数上限时，即使没有满足收敛判据，迭代也停止。系统默认值为10。选择范围为1-999。
- **Convergence Criterion:** 指定K-Means 算法中的收敛标准，输入一个不超过1的正数作为判定迭代收敛的标准。系统缺省的收敛标准是0.02，表示当两次迭代计算的最小的类中心的变化距离小于初始类中心距离的百分之2%时迭代停止。

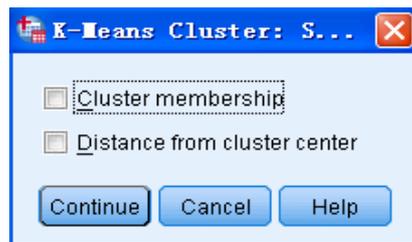
提示：如果设置了以上两个参数，只要在迭代过程中满足了一个参数，迭代就停止。

- **Use running means:** 使用移动平均。选中该复选框，限定在每个观测量被分配到一类后立刻计算新的类中心。如果不选择此项，则在完成了所有观测量的一个分配后再计算各类的类中心，这样可以节省迭代时间。

## 9.2 SPSS在聚类分析中的应用

### Step07: 输出聚类结果

在主对话框中单击【Save (保存)】按钮，弹出【Save New Variables (保存新变量)】对话框，它用于选择保存新变量。



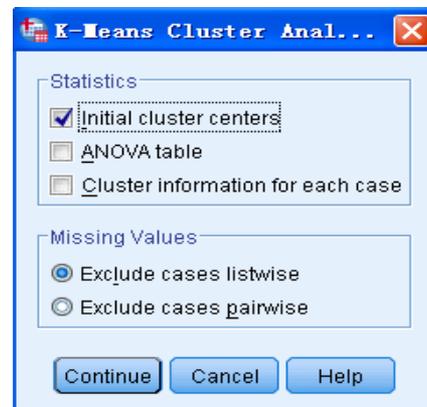
- Cluster membership: 在当前数据文件中建立一个名为“qcl\_1”新变量。其值表示聚类结果，即各观测量被分配到哪一类。它的取值为1、2、3…的序号。
- Distance from cluster center: 在当前数据文件中建立一个名为“qcl\_2”新变量。其值为各观测量与所属类中心之间的欧氏距离。



## 9.2 SPSS在聚类分析中的应用

### Step08: 其他选项输出

在主对话框中单击【Option (选项)】按钮，弹出【Option (选项)】对话框，它用于指定要计算的统计量和对带有缺失值的观测量的处理方式。具体见图：



① 【Statistics (统计量)】选项组：选择输出统计量。

- Initial cluster centers: 初始聚类中心。
- ANOVA table: 方差分析表。
- Cluster information for each case: 显示每个观测量的聚类信息。

② 【Missing Values (缺失值)】选项组：选择处理缺失值方法。

- Exclude cases listwise: 分析变量中带有缺失值的观测量都不参与后续分析。
- Exclude cases pairwise: 成对剔除带有缺失值的观测量。

Step09: 单击【OK】按钮，结束操作，SPSS软件自动输出结果。



## 9.2 SPSS在聚类分析中的应用

- 9.2.3 实例分析：全国环境污染程度分析

为了更深入地了解我国环境的污染程度状况，现利用2009年数据对全国31个省、自治区、直辖市进行聚类分析。



## 9.2 SPSS在聚类分析中的应用

现在要分析我国各个地区的环境污染程度，案例中选择了各地区“工业废气排放总量”、“工业废水排放总量”和“二氧化硫排放总量”三个指标来反映不同污染程度的环境状况，同时选择了北京等省市的数据加以研究。这个问题属于典型的多元分析问题，需要利用多个指标来分析各省市之间环境污染程度的差异。因此，可以考虑利用快速聚类分析来研究各省市之间的差异性，具体操作步骤如下。

- 打随书光盘中的数据文件9-2. sav, 选择菜单栏中的【Analyze (分析)】→【Classify (分类)】→【K-Means Cluster (K均值聚类)】命令, 弹出【K-Means Cluster Analysis (K均值聚类分析)】对话框。
- 在左侧的候选变量列表框中将 $X1$ 、 $X2$ 和 $X3$ 变量设定为聚类分析变量, 将其添加至【Variables (变量)】列表框中; 同时选择 $I$ 作为标识变量, 将其移入【Label Cases by (个案标记依据)】列表框中。
- 在【Number of Clusters (聚类数)】文本框中输入数值“3”, 表示将样品利用聚类分析分为三类, 如下图所示。



**K-Means Cluster Analysis**

Cluster Number of Cas...  
Distance of Case from ...

Variables:  
工业废气排放总量 [X1]  
工业废水排放总量 [X2]  
二氧化硫排放总量 [X3]

Iterate...  
Save...  
Options...

Label Cases by:  
省市 [Y]

Number of Clusters: 3

Method  
 Iterate and classify  Classify only

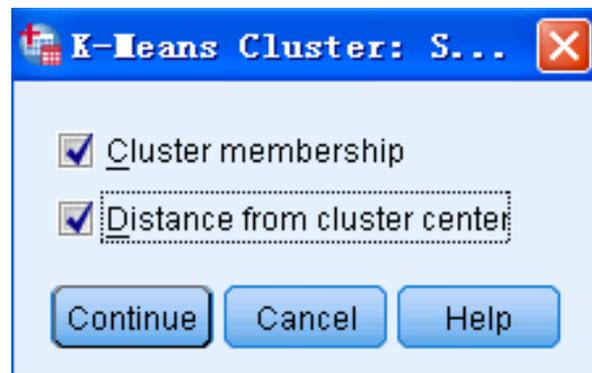
Cluster Centers

Read initial:  
 Open dataset  
 External data file File...

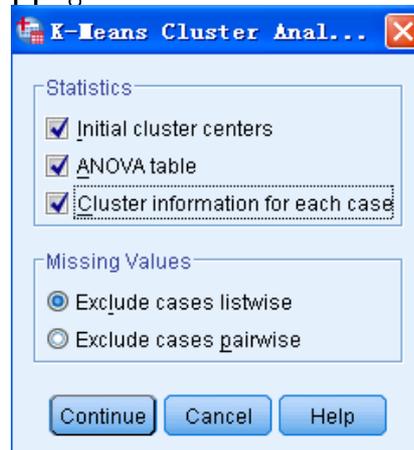
Write final:  
 New dataset  
 Data file File...

OK Paste Reset Cancel Help

- 单击【Save（保存）】按钮，弹出【K-Means Cluster Analysis: Save（K均值聚类分析：保存）】对话框；勾选【Cluster membership（聚类新成员）】和【Distance from cluster center（与聚类中心的距离）】复选框，表示输出样品的聚类类别及距离，其他选项保持系统默认设置，如下图所示，单击【Continue（继续）】按钮返回主对话框。



- 单击【Options（选项）】按钮，弹出【K-Means Cluster Analysis: Options（K均值聚类分析：选项）】对话框；勾选【Statistics（统计量）】选项组中的复选框，其他选项保持系统默认设置，如下图所示，单击【Continue（继续）】按钮返回主对话框，单击【OK（确定）】按钮完成操作。





## 9.2 SPSS在聚类分析中的应用

### 实例结果及分析

#### (1) 快速聚类分析的初始中心

SPSS软件首先给出了进行快速聚类分析的初始中心数据。由于这里是要求将样品分为三类，因此软件给出了三个中心位置。但是，这些中心位置可能在后续的迭代计算中出现调整。

快速聚类分析的初始中心

	Cluster		
	1	2	3
工业废气排放总量	15	22186	27432
工业废水排放总量	942	140325	256160
二氧化硫排放总量	0.2	135.5	107.4



## 9.2 SPSS在聚类分析中的应用

### (2) 迭代历史表

下表显示了快速聚类分析的迭代过程。可以看到，第一次迭代的变化值最大，其后随之减少。最后第三次迭代时，聚类中心就不再变化了。这说明，本次快速聚类的迭代过程速度很快。

迭代历史表

Iteration	Change in Cluster Centers		
	1	2	3
1	29063.875	15957.005	26705.187
2	4706.401	3783.482	22208.692
3	0.000	0.000	0.000



## 9.2 SPSS在聚类分析中的应用

### (3) 聚类分析结果列表

通过快速聚类分析的最终结果列表可以看到整个样品被分为以下三大类。

- 第一类：北京、天津、山西、内蒙古等20个地区。这些地区工业废水、废气及二氧化硫的排放总量相对最低。
- 第二类：河北、福建、河南、湖北、湖南、广西和四川。它们的污染程度在所有省份中位居中等水平。
- 第三类：江苏、浙江、山东和广东。这些地区的工业废水、废气及二氧化硫排放总量是最高的，因此环境污染也最为严重。

表中最后一列显示了样品和所属类别中心的聚类，此表中的最后两列分别作为新变量保存于当前的工作文件中。



## 9.2 SPSS在聚类分析中的应用

### (4) 最终聚类分析中心表

如下表所示列出了最终聚类分析中心。可以看到，最后的中心位置较初始中心位置发生了较大的变化。

最终聚类分析中心

	Cluster		
	1	2	3
工业废气排放总量	9921	19079	26025
工业废水排放总量	33219	121194	207780
二氧化硫排放总量	56.0	93.0	110.9



## 9.2 SPSS在聚类分析中的应用

### (5) 最终聚类中心位置之间的距离

如下表所示为快速聚类分析最终确定的各类中心位置的距离表。从结果来看，第一类和第三类之间的距离最大，而第二类和第三类之间的距离最短，这些结果和实际情况是相符合的。

最终聚类中心位置之间的距离

Cluster	1	2	3
1		88449.975	175301.923
2	88449.975		86864.229
3	175301.923	86864.229	



## 9.2 SPSS在聚类分析中的应用

### (6) 方差分析表

如下表所示为方差分析表，显示了各个指标在不同类的均值比较情况。各数据项的含义依次是：组间均方、组间自由度、组内均方、组内自由度。可以看到，各个指标在不同类之间的差异是非常明显的，这进一步验证了聚类分析结果的有效性。

方差分析表

	Cluster		Error		F	Sig.
	Mean Square	df	Mean Square	df		
工业废气排放总量	5.458E8	2	86415059.434	28	6.316	0.005
工业废水排放总量	6.018E10	2	6.317E8	28	95.270	0.000
二氧化硫排放总量	7125.110	2	1510.247	28	4.718	0.017



## 9.2 SPSS在聚类分析中的应用

### (7) 聚类数目汇总

如下表所示是聚类数据汇总表，显示了聚类分析最终结果中各个类别的数目。其中第一类的数目最多，等于20；而第三类的数目最少，只有4个。

聚类数目汇总表

Cluster	1	20.000
	2	7.000
	3	4.000
Valid		31.000
Missing		0.000



## 9.2 SPSS在聚类分析中的应用

### 9.2.4 系统聚类法的SPSS操作详解

系统聚类法常称为层次聚类法、分层聚类法，也是聚类分析中使用广泛的一种方法。它有两种类型，一是对研究对象本身进行分类，称为Q型聚类；另一是对研究对象的观察指标进行分类，称为R型聚类。同时根据聚类过程不同，又分为分解法和凝聚法。

**分解法：**开始把所有个体(观测量或变量)都视为同属一大类，然后根据距离和相似性逐层分解，直到参与聚类的每个个体自成一类为止。

**凝聚法：**开始把参与聚类的每个个体(观测量或变量)视为一类，根据两类之间的距离或相似性逐步合并，直到合并为一个大大类为止。



## 9.2 SPSS在聚类分析中的应用

SPSS中的系统聚类法采用的凝聚法，它的算法步骤具体如下。

- 1、首先将数据各自作为一类（这时有 $n$ 类），按照所定义的距离计算各数据点之间的距离，形成一个距离阵；
- 2、将距离最近的两条数据并为一个类别，从而成为 $n-1$ 个类别，计算新产生的类别与其他各个类别之间的距离或相似度，形成新的距离阵；
- 3、按照和第二步相同的原则，再将距离最接近的两个类别合并，这时如果类的个数仍然大于1，则继续重复这一步骤，直到所有的数据都被合并成一个类别为止。



## 9.2 SPSS在聚类分析中的应用

在系统聚类中，当每个类别有多于一个的数据点构成时，就会涉及如何定义两个类间的距离问题。根据距离公式不同，可能会得到不同的结果，这也就进一步构成了不同的系统聚类方法。常用的方法有如下几种。

- Between-groups linkage: 组间平均距离法。
- Within-groups linkage: 组内平均距离法。
- Nearest neighbor: 最短距离法。
- Furthest neighbor: 最远距离法。
- Centroid clustering: 重心法。
- Median clustering: 中间距离法。
- Ward's method: 离差平方和法。

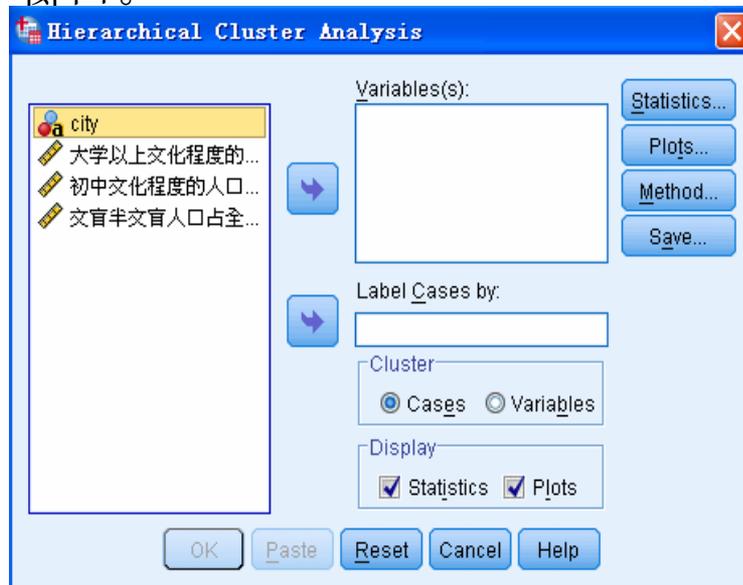


## 9.2 SPSS在聚类分析中的应用

SPSS具体操作步骤如下：

Step01：打开对话框

选择菜单栏中的【Analyze（分析）】→【Classify（分类）】→【Hierarchical Cluster（系统聚类）】命令，弹出【Hierarchical Cluster Cluster Analysis（系统聚类分析）】对话框，这是系统聚类分析的主操作窗口。





## 9.2 SPSS在聚类分析中的应用

### Step02: 选择聚类分析变量

在【Hierarchical Cluster Cluster Analysis (系统聚类分析)】对话框左侧的候选变量列表框中选择进行系统聚类分析的变量，将其添加至【Variable(s) (变量)】列表框中。同时可以选择一个标识变量移入【Label Cases by (标注个案)】列表框中。

### Step03: 选择聚类类型

在【Cluster (分群)】选项组中可以选择聚类类型。系统默认值是【Cases (个案0)】选项。

- Cases: 对观测量 (样品) 进行聚类, 即Q型聚类。
- Variable: 对变量进行聚类, 即R型聚类。

### Step04: 选择输出类型

在【Display (输出)】选项组中可以选择输出类型。系统默认值是【Statistics (统计量)】欧诺供给量和【Plots (图)】选项。

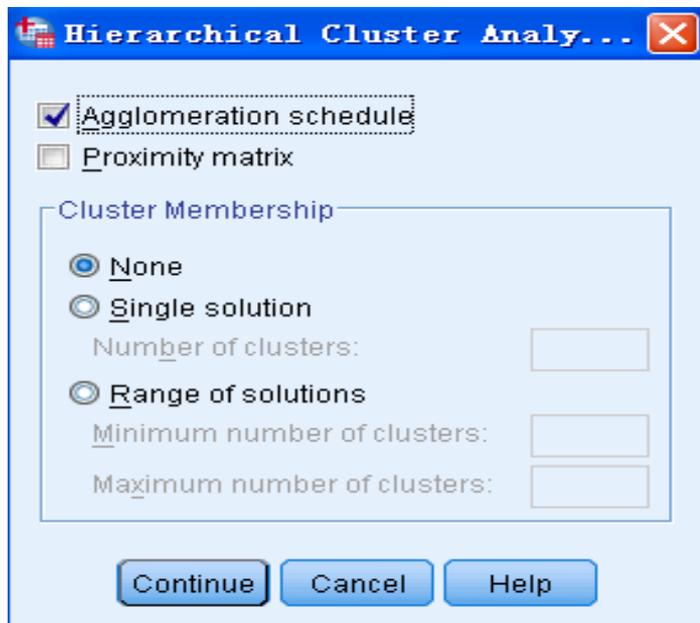
- Statistic: 输出主对话框【Statistics】按钮中设置的的统计量。
- Plots: 输出主对话框中【Plots (图)】按钮中聚类图形。



## 9.2 SPSS在聚类分析中的应用

Step05: 基本统计量输出选择

单击【Statistics】按钮，在弹出的对话框中可以选择进行系统聚类分析的基本统计量。具体选项含义如下。





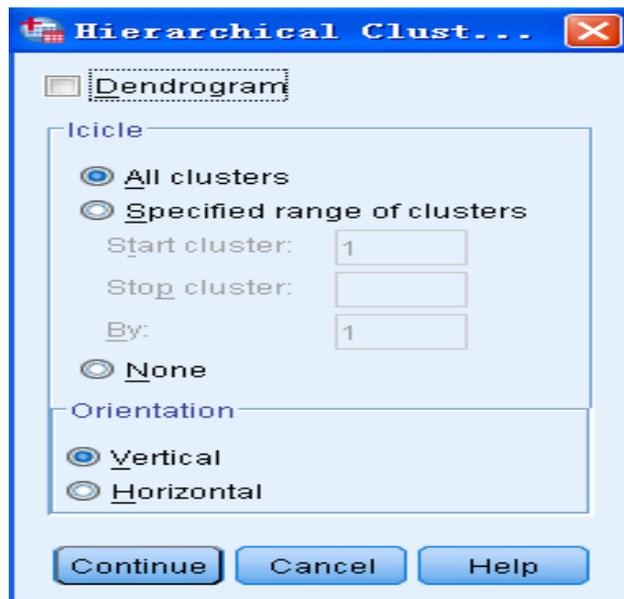
## 9.2 SPSS在聚类分析中的应用

- ① **【Agglomeration schedule (合并进程表)】**：输出聚类过程表，系统默认选项。显示聚类过程中每一步合并的类或观测量，反映聚类过程中每一步样品或类的合并过程。
- ② **【Proximity matrix (相似性矩阵)】**：输出各类之间的距离矩阵。以矩阵形式给出各项之间的距离或相似性测度值。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 (方法)】**菜单中**【Measure (度量标准)】**栏中的选择。
- ③ **【Cluster Membership (聚类成员)】** 栏可以选择聚类数目相关的输出项：
  - **【None (无)】**：不显示类成员表，它是系统默认选项。
  - **【Single solution (单一方案)】**：选择此项并在对应的**【Number of clusters (聚类数)】**参数框中指定分类数，这里要求分类数是一个大于1的整数。例如指输入数字“4”，则会在输出窗中显示聚为4类的分析结果。
  - **【Range of solutions (方案范围)】**：选择此选项并在下边的**【Minimum number of clusters (最小聚类数)】**和**【Maximum number of clusters (最大聚类数)】**参数框中输入最小聚类数目和最大聚类数目。它表示分别输出样品或变量的分类数从最小值到最大值的各种分类聚类表。输入的两个数值必须是不等于1的正整数，最大类数值不能大于参与聚类的样品数或变量总数。

## 9.2 SPSS在聚类分析中的应用

### Step06: 聚类统计图形输出选择

单击【Plots】按钮，弹出的对话框如下图所示。这里可以选择进行系统聚类分析的统计图形。可选择输出的统计图表有两种，一个是树形图，一个是冰柱图。具体选项含义如下。





## 9.2 SPSS在聚类分析中的应用

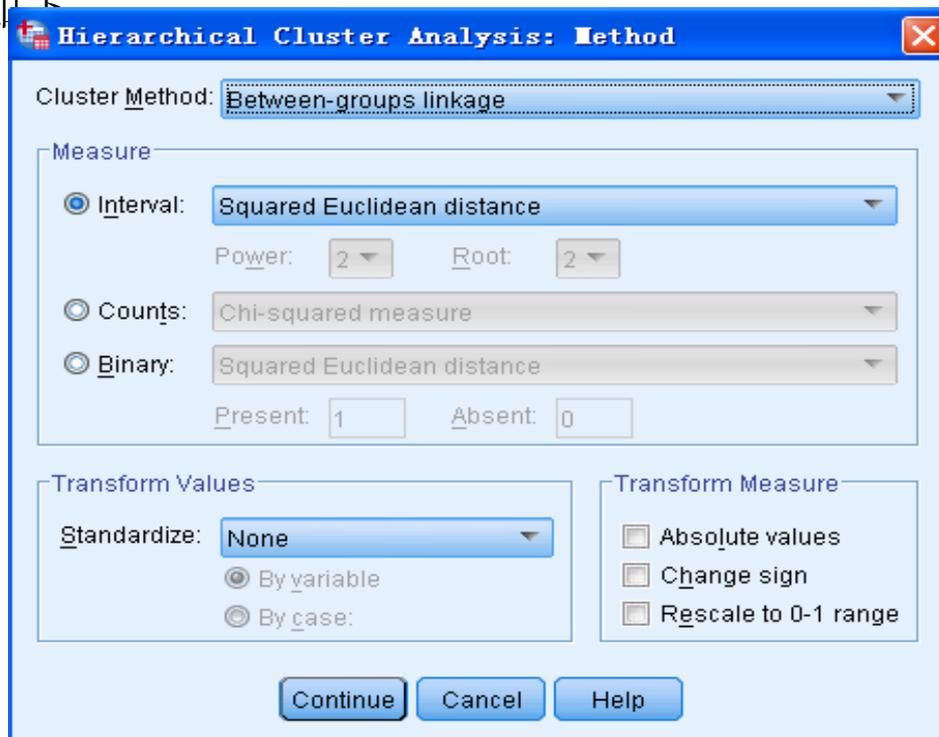
- ① **【Dendrogram（树状图）】**：显示树形图。
  - ② **【Icicle（冰柱）】**：显示冰柱图形。对于冰柱图的具体选项还可以进一步用以下选择项来确定。
    - **All clusters**：显示全部聚类结果的冰柱图。可用此种图查看聚类的全过程。但如果参与聚类的个体很多会造成图形过大。
    - **Specified range clusters**：限定显示的聚类范围。当选择此项时，在下面的**【Start cluster（开始聚类）】**、**【Stop cluster（停止聚类）】**和**【By（排序标准）】**后的参数框中输入要求显示聚类过程的开始聚类数、终止聚类数及步长。输入到参数框中的数字必须是正整数。例如，输入的结果是：3, 9, 2，生成的冰柱图从第三步开始，显示第三、五、七、九步聚类的情况。
    - **None**：不输出冰柱图。
- 同时，冰柱图显示方向可以在**【Orientation（方向）】**选项组中确定。
- **Vertical**：纵向显示的冰柱图。
  - **Horizontal**：横向显示的冰柱图。



## 9.2 SPSS在聚类分析中的应用

### Step07: 聚类方法选择

单击【Method（方法）】按钮，弹出的对话框如下图所示。在对话框中可以设定聚类方法、距离测度的方法、数值变换方法等内容。具体选项含义如下





## 9.2 SPSS在聚类分析中的应用

- ① 【Cluster Method（聚类方法）】下拉列表框：可以选择聚类方法，具体如下。
- Between-groups linkage: 组间平均距离法。系统默认选项。合并两类的结果使所有的两类的平均距离最小。
  - Within-groups linkage: 组内平均距离法。当两类合并为一类后，合并后的类中的所有项之间的平均距离最小。
  - Nearest neighbor: 最近距离法。采用两类间最近点间的距离代表两类间的距离。
  - Furthest Neighbor: 最远距离法。用两类之间最远点的距离代表两类之间的距离。
  - Centroid clustering: 重心法。定义类与类之间的距离为两类中各样品的重心之间的距离。
  - Median clustering: 中位数法。定义类与类之间的距离为两类中各样品的中位数之间的距离。
  - Ward's method: 最小离差平方和法。聚类中使类内各样品的离差平方和最小，类间的离差平方和尽可能大。



## 9.2 SPSS在聚类分析中的应用

② **【Measure（度量标准）】**选项组：可以选择距离测度方法，具体如下。

**【Interval（区间）】**参数框适合于等间隔测度的连续性变量。单击它的右侧框边向下箭头展开下拉菜单，在菜单中选择距离测度方法，具体如下。

- Euclidean distance: 欧氏距离。
- Squared Euclidean distance: 欧氏距离平方。两项之间的距离是每个变量值之差的平方和。系统默认项。
- Cosline: 余弦相似性测度，计算两个向量间夹角的余弦。
- Pearson conelation: 皮尔逊相关系数。它是线性关系的测度，范围是-1~+1。
- Chebychev: 切比雪夫距离。
- Block: 曼哈顿（Manhattan）距离，两项之间的距离是每个变量值之差的绝对值总和。
- Minkowski: 闵科夫斯基距离。
- Customized: 自定义距离。

**【Counts（计数）】**参数框适合于计数变量(离散变量)。单击它右侧的向下箭头，展开下拉菜单的方法选择以下不相似性测度的方法。具体如下：

- Chi-square measure: 卡方测度。用卡方值测度不相似性。系统默认选项。
- Phi-square measure: 两组频数之间的 $\Phi^2$ 测度。



## 9.2 SPSS在聚类分析中的应用

【Binary (二分数)】参数框适合于二值变量。首先应该明确,对二值变量,系统默认用1表示某特性出现(或发生),用0表示某特性不出现(或不发生)。单击它的右侧框边向下箭头展开下拉菜单,在菜单中选择测度方法。具体如下:

- Euclidean distance: 二元变量欧氏距离。
- Squared Euclidean distance: 二元变量欧氏距离的平方。
- Size difference: 不对称指数。其值范围在0 ~ 1 之间。
- Pattern difference: 不相似性测度, 范围为0 ~ 1。
- Variance: 方差不相似性测度。
- Dispersion: 离散测度, 其范围为-1 ~ 1。
- Shape: 距离测度。范围无上下限。
- Simple matching: 简单匹配测度。
- Phi 4-point correlation: 皮尔逊相关系数二元变量模拟, 其值范围为-1 ~ 1。
- Lambda: 其值是Goodman and Kruskal 的  $\lambda$  值, 它是一种相似性测度。
- Anderberg' D: 安德伯格D系数。
- Dice: 戴斯匹配系数。
- Hamann: 哈曼匹配系数。



## 9.2 SPSS在聚类分析中的应用

- Jaccard: 杰卡得相似比。
- Kulczynski 1: 库尔津斯基匹配系数。
- Kulczynski 2: 库尔津斯基条件概率测度。
- Lance and Williams: 兰斯-威廉斯测度。
- Ochiai: 该指数是余弦相似性测度的二元形式。范围为0 ~ 1。
- Rogers and Tanimoto: 罗杰斯-谷本匹配系数。
- Russel and Rao: 它是内积(点积)的二元形式。对匹配与不匹配都给予相等的权重。
- Sokal and Sneath 1 ~ 5: 第一种~第五种索克尔-思尼斯匹配系数。
- Yule' s Y: 尤利Y综合系数。
- Yule' s Q: 尤利Q综合系数。。

从上述选项中选择一种测度方法。同时，还可以改变表示某事件发生与不发生的值。在【Present (存在)】和【Absent (不存在)】的参数框中键入用户自己定义的值。定义后，系统将忽略其他值。如果不进行自定义，那么，1代表某事件发生“Present”，0代表某事件不发生“Absent”。



## 9.2 SPSS在聚类分析中的应用

- ③ **【Transform Values (转换数)】**选项组：可以选择数据标准化的方法。注意只有等间隔测度的数据（选择了Interval）或计数数据（选择了Counts）才可以进行标准化。具体如下：
- None：不进行标准化。系统默认值。
  - Z scores：数据标准化到Z 分数。标准化后变量均值为0，标准差为1。
  - Range -1 to 1：把数据标准化到-1 到+1 范围内。
  - Range 0 to 1：把数据标准化到0 到+1 范围内。
  - Maximum magnitude of 1：把数据标准化到最大值为1。表示各变量除以最大值。
  - Mean of 1：把数据标准化到均值为1。表示各变量除以均值。
  - Standard deviation of 1：把数据标准化到标准差为1。表示各变量除以标准差。

在选择了上述标准化方法后，要在选项组中点选**【By variable (对变量)】**或**【By case (对样品)】**单选钮实施标准化。



## 9.2 SPSS在聚类分析中的应用

④ **【Transform Measure】** 选项组：可以选择测度的转换方法，具体如下。

- Absolute Values: 把距离值取绝对值。
- Change sign: 把相似性值变为不相似性值或相反。
- Rescale to 0~1 range: 重新调整测度值到范围0~1。

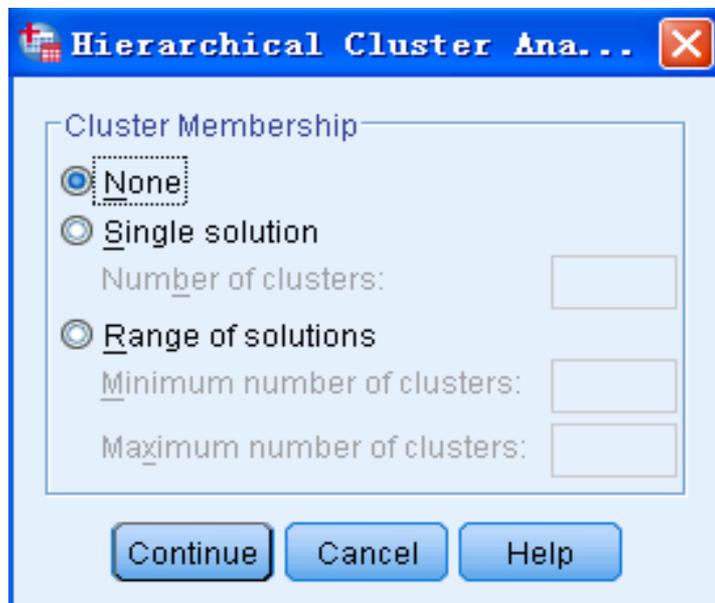
对于已经计算了相似性或不相似性测度的数据，一般不再使用此方法进行转换。如果使用的是已经存在的矩阵，可以选择此类选择项，对输入矩阵进行必要的转换。



## 9.2 SPSS在聚类分析中的应用

### Step08: 聚类结果保存选择

单击【Save】按钮，在弹出的对话框中可以将聚类结果用新变量保存在当前工作数据文件中。具体选项含义如下。





## 9.2 SPSS在聚类分析中的应用

- None: 不建立新变量。
- Single solution: 单个结果输出。生成一个新变量, 表明每个样品在聚类之后所属的类。在【Number of clusters (聚类数)】的矩形框中指定类数。
- Range of solutions: 选择此选项并在下边的【Minimum number of clusters (最小聚类数)】和【Maximum number of clusters (最大聚类数)】文本框中输入最小聚类数目和最大聚类数目。它表示分别生成样品或变量的分类数从最小值到最大值的各种分类聚类变量。例如输入结果是“4”和“6”时, 它表示在聚类结束后在原变量后面增加了3个新变量分别表明分为4类时、分为5类时和分为6类时的聚类结果。即聚为4、5、6类时各样品分别属于哪一类。

Step09: 单击【OK】按钮, 结束操作, SPSS软件自动输出结果。



## 9.2 SPSS在聚类分析中的应用

### 9.2.5 实例分析：不同地区信息基础设施发展状况的评价

#### 1. 实例内容

要研究世界不同地区信息基础设施的发展状况，这里选取了发达地区、新兴工业化地区、拉美地区、亚洲地区中国家、转型地区等不同类型的20个国家的数据。描述信息基础设施的变量主要有六个。

- (1) Call—每千人拥有电话线数。
- (2) movecall—每千居民蜂窝移动电话数。
- (3) fee—高峰时期每三分钟国际电话的成本。
- (4) Computer—每千人拥有的计算机数。
- (5) mips—每千人中计算机功率（每秒百万指令）。
- (6) net—每千人互连网络户主数。



## 9.2 SPSS在聚类分析中的应用

### 2. 实例操作

现在要分析世界各个地区的信息基础设施的发展状况，案例中选择了“每千人拥有电话线数”、“每千户居民蜂窝移动电话数”等六个指标来反映不同国家信息设施的发展情况，同时选择了近二十个地区的数据加以研究。这个问题也属于典型的多元分析问题，需要利用多个指标来分析地区之间信息基础设施发展的差异。因此，可以利用系统聚类法。



## 9.2 SPSS在聚类分析中的应用

### 3 实例结果及分析

#### (1) 聚类过程表

SPSS软件首先给出了进行系统聚类分析的过程表。下表中的的第一列“Stage”列出了聚类过程的步骤号，第二列“Cluster 1”和第三列“Cluster 2”列出了某一步骤中哪些国家参与了合并。例如从结果中看出，在第一步中，第十个样品(Brazil)和第十二个样品(Mexico)首先被合并在一起。第四列“Coefficients”列出了每一步骤的聚类系数，这一数值表示被合并的两个类别之间的距离大小。第五列“Cluster 1”和第六列“Cluster 2”表示参与合并的国家(类别)是在第几步中第一次出现，0代表该记录是第一次出现在聚类过程中。第七列“Next Stage”表示在这一步骤中合并的类别，下一次将在第几步中与其他类再进行合并。

# 9.2 SPSS在聚类分析中的应用

聚类过程表

Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	10	12	.107	0	0	4
2	8	9	.164	0	0	11
3	13	17	.278	0	0	7
4	10	14	.520	1	0	6
5	3	19	.675	0	0	14
6	10	15	1.055	4	0	10
7	13	18	1.099	3	0	15
8	7	20	1.249	0	0	12
9	4	6	1.343	0	0	17
10	10	11	1.421	6	0	13
11	2	8	1.809	0	2	16
12	5	7	1.880	0	8	14
13	10	16	2.247	10	0	15
14	3	5	2.359	5	12	16
15	10	13	3.878	13	7	18
16	2	3	4.719	11	14	18
17	1	4	6.407	0	9	19
18	2	10	11.117	16	15	19
19	1	2	25.049	17	18	0



## 9.2 SPSS在聚类分析中的应用

### (2) 聚类分析结果表

在系统聚类法的聚类结果中可以看到，聚类结果分为三大类。

第Ⅰ类：美国、瑞典、丹麦。

第Ⅱ类：日本、德国、瑞士、新加坡、中国台湾、韩国、法国、英国。

第Ⅲ类：巴西、墨西哥、波兰、匈牙利、智利、俄罗斯、泰国、印度、马来西亚。

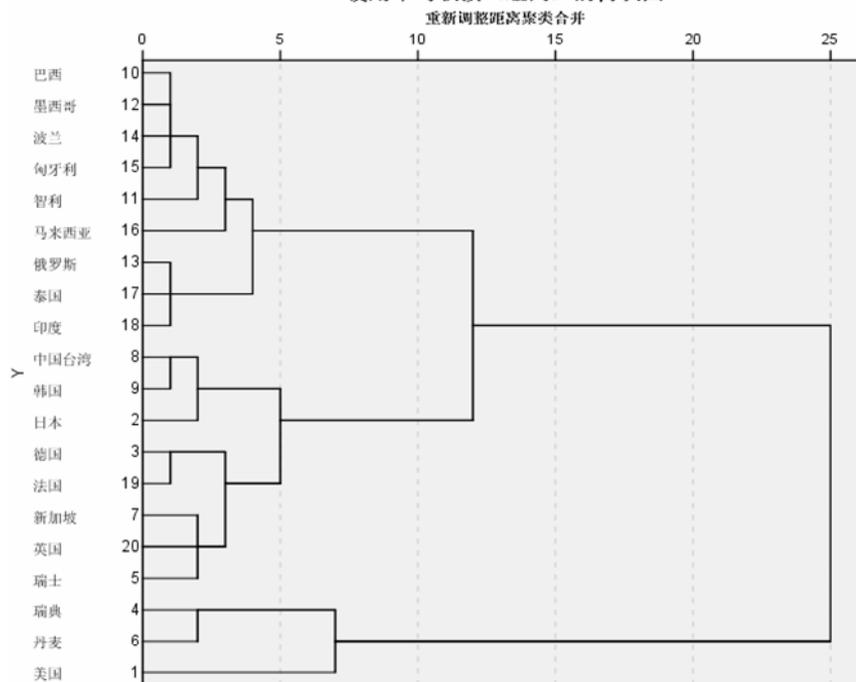


## 9.2 SPSS在聚类分析中的应用

### (3) 树形图

上述已给出了相关聚类结果，最后用树形图（Dendrogram）直观反映整个聚类过程和结果，如图9-37所示。从图中，可以明显看到每个样品从单独一类，逐次合并，一直到全部合并成一大类。

使用平均联接（组间）的树状图



# 9.3 SPSS在判别分析中的应用

CONCEPT  
STRATE

## 9.3.1 判别分析的基本原理

### 1、方法概述

判别分析是判别样品所属类型的一种统计方法，其应用之广可与回归分析媲美。

判别分析与聚类分析不同。判别分析是在已知研究对象分成若干类型（或组别）并已取得各种类型的一批已知样品的观测数据，在此基础上根据某些准则建立判别式，然后对未知类型的样品进行判别分类。

### 2、基本原理

判别分析内容很丰富，方法很多。判别分析按判别的组数来区分，有两组判别分析和多组判别分析；按区分不同总体的所用的数学模型来分，有线性判别和非线性判别；按判别时所处理的变量方法不同，有逐步判别和序贯判别等。

其中，距离判别分析是一种常见的判别分析方法。它的基本思想是：首先根据已知分类的数据，分别计算各类的重心即分组（类）的均值，判别准则是对任给的一次观测，若它与第 $i$ 类的重心距离最近，就认为它来自第 $i$ 类。



## 9.3 SPSS在判别分析中的应用

例如两个总体的距离判别法中，设有两个总体（或称两类） $G_1$ 、 $G_2$ ，从第一个总体中抽取 $n_1$ 个样品，从第二个总体中抽取 $n_2$ 个样品，每个样品测量 $p$ 个指标如下页表。

今任取一个样品，实测指标值为  $X = (x_1, \dots, x_p)'$ ，问 $X$ 应判归为哪一类？

首先计算 $X$ 到 $G_1$ 、 $G_2$ 总体的距离，分别记为  $D(X, G_1)$  和  $D(X, G_2)$ ，按距离最近准则判别归类，则可写成：

$$\begin{cases} X \in G_1, \text{当} D(X, G_1) < D(X, G_2) \\ X \in G_2, \text{当} D(X, G_1) > D(X, G_2) \\ \text{待判, 当} D(X, G_1) = D(X, G_2) \end{cases}$$

然后比较  $D(X, G_1)$  和  $D(X, G_2)$  大小，按距离最近准则判别归类。

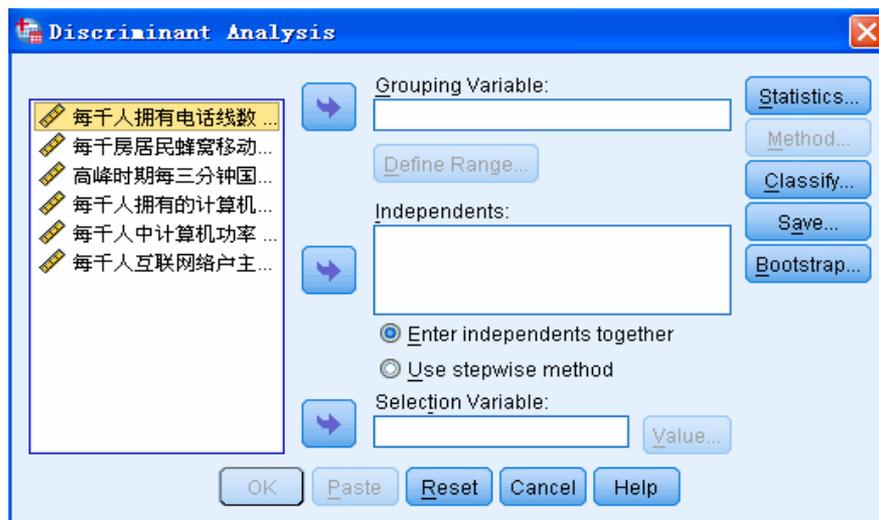


## 9.3 SPSS在判别分析中的应用

### 9.3.2 判别分析的SPSS操作详解

Step01: 打开对话框

选择菜单栏中的【Analyze（分析）】→【Classify（分类）】→【Discriminant（辨别）】命令，弹出【Discriminant Analysis（判别分析）】对话框，这是判别分析的主操作窗口。





## 9.3 SPSS在判别分析中的应用

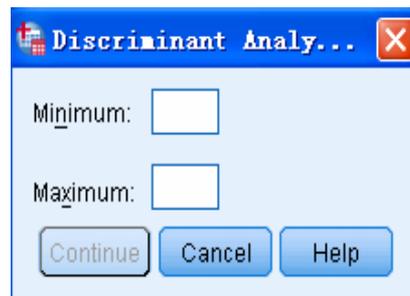
Step02: 选择判别分析变量

在【Discriminant Analysis (辨别分析)】对话框左侧的候选变量中选择进行判别分析的变量，将其添加至【Independents (自变量)】列表框中，将其作为自变量。

Step03: 指定分类变量及范围

在主对话框的候选变量中选择分类变量（离散型变量）移入【Grouping Variable (分组变量)】框中。此时它下面的【Define Range (定义范围)】按钮加亮，单击该按钮，屏幕弹出一个对话框，提供指定该分类变量的数值范围。

- Minimum: 输入最小值。
- Maximum: 输入最大值。





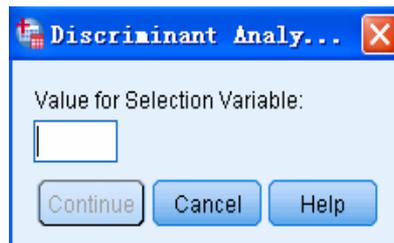
## 9.3 SPSS在判别分析中的应用

### Step04: 选择判别分析方法

在主对话框的【Independents（自变量）】列表框下面有两个按钮，它们提供了判别分析方法选择。

- Enter independent together: 建立所选择的所有变量的判别式。当认为所有自变量都能对观测量特性提供丰富的信息时使用该选择项。系统默认设置。
- Use stepwise method: 采用逐步判别法作判别分析。点选该项后，主菜单中的【Method（方法）】按钮加亮。可以进一步选择判别分析方法（见第 步）。

如果希望使用一部分观测量进行判别函数的推导，选择一个能够标记需选择的这部分观测量的变量将其移入【Selection Variables（选择变量）】框中；再单击其右侧的Value按钮，展开【Set Value（设置值）】对话框，键入能标记的变量值，如图所示。

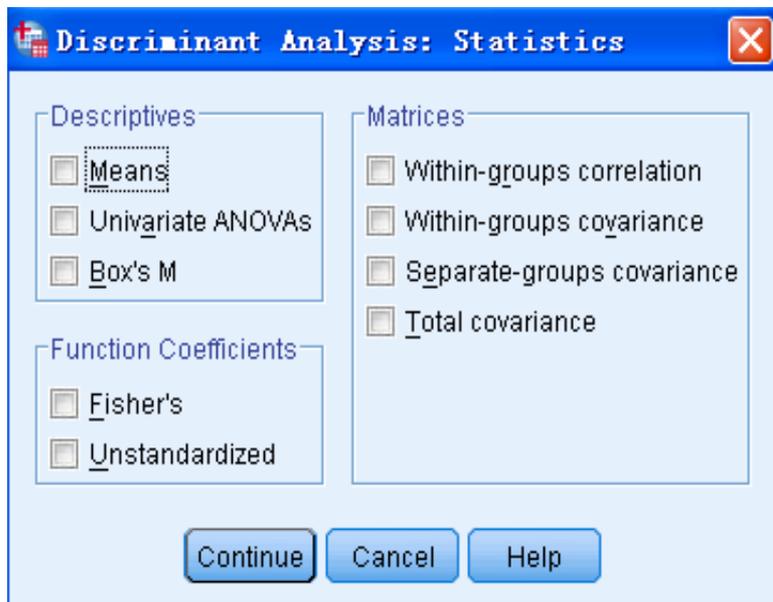




## 9.3 SPSS在判别分析中的应用

### Step05: 基本统计量输出选择

单击【Statistics】按钮，在弹出的对话框中可以选择进行判别分析的基本统计量输出。具体选项含义如下。





## 9.3 SPSS在判别分析中的应用

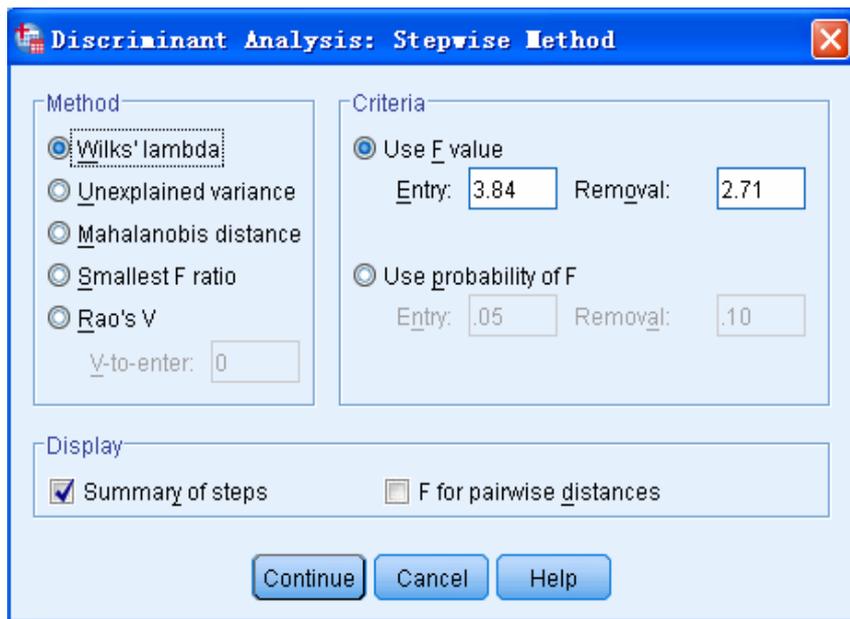
- ① **【Descriptives (描述性)】** 选项组：选择输出描述统计量。
  - Means：输出各类中各自变量的均值、标准差和各自变量总样本的均值、标准差。
  - Univariate ANOVAs：单因素方差分析。对各类中同一自变量进行均值检验，输出单因素方差分析结果。
  - Box' s M：对各类协方差矩阵相等的假设进行检验。
- ② **【Function coefficients (函数系数)】** 选项组：选择输出判别函数的系数。
  - Fisher' s：输出Fisher函数系数。对每一类给出一组系数，并给出该组中判别分数最大的观测量。
  - Unstandardized：未经标准化处理的判别函数系数。
- ③ **【Matrices (矩阵)】** 选项组：选择输出自变量的系数矩阵。
  - Within-groups correlation matrix：类内相关矩阵。
  - Within-groups covariance matrix：类内协方差矩阵
  - Separate-groups covariance matrices：对每一类分别输出协方差矩阵。
  - Total covariance matrix：总样本的协方差矩阵。



## 9.3 SPSS在判别分析中的应用

### Step06: 设置逐步判别分析选项

點選【Use stepwise method (使用步进式方法)】单选钮后, 就表示采用逐步判别法进行分析。接着单击主菜单中的【Statistics】按钮, 在弹出的对话框图中可以选择逐步判别分析的选项。具体选项含义如下。





## 9.3 SPSS在判别分析中的应用

- ① 【Method（方法）】选项组：选择变量进入判别函数的方式。
- Wilks' lambda: 每步都选择Wilks的 $\lambda$ 统计量最小的变量进入判别函数。
  - Unexplained variance: 每步都选择使类间不可解释的方差和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类间的Mahalanobis距离最大的变量进入判别函数。
  - Smallest F ratio: 每步都选择使任何两类间的“最小F值”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。选择此种方法后，应该在该项下面的【V-to-enter】文本框中输入这个增量的指定值。当某变量导致的V值增量大于指定值的变量时，该变量进入判别函数。



## 9.3 SPSS在判别分析中的应用

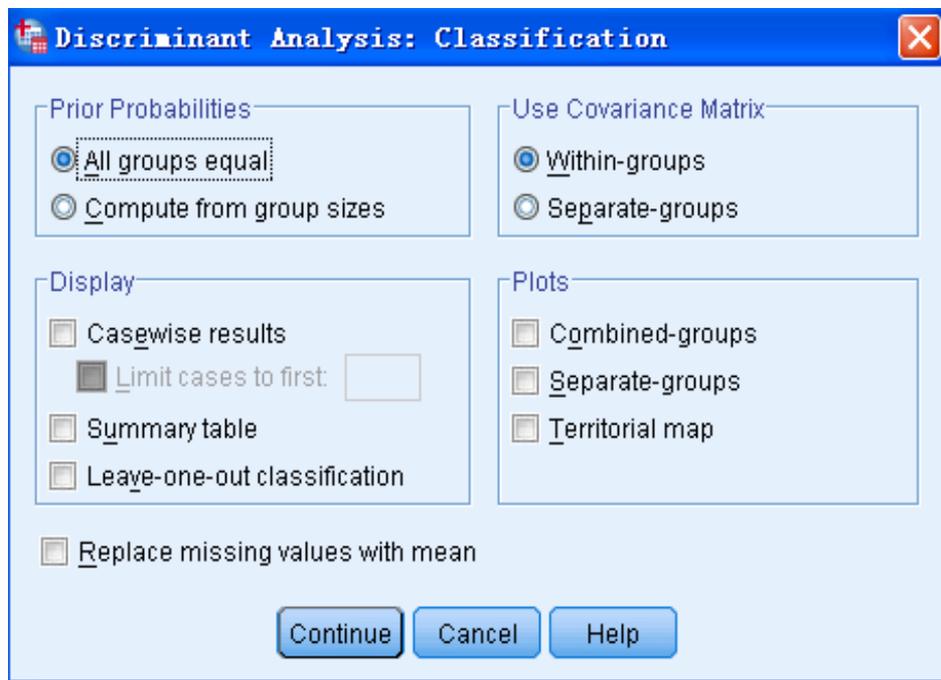
- ② **【Criteria (标准)】** 选项组：选择逐步判别停止的条件。
- **Use F value:** 使用F值，系统默认选项，当加入一个变量（或剔除一个变量）后，对在判别函数中的变量进行方差分析。当计算的F 值大于指定的Entry 值时，该变量保留在函数中。默认值是Entry 为3.84。当该变量使计算的F 值小于指定的Removal 值时，该变量从函数中剔除。默认值是Removal 为2.71。设置这两个值时应该要求Entry 值大于Removal 值。
  - **Use probability of F:** 使用F 检验的概率决定变量是否加入函数或被剔除。当计算的F 检验的概率小于指定的Entry 值时，该变量加入函数中。当该变量使计算的F 值的概率大于指定的Removal 值时，该变量从函数中剔除。
- ③ **【Display (输出)】** 栏选择逐步选择变量的过程和最后结果的显示：
- **Summary of steps:** 显示每步选择变量之后各变量的统计量结果。
  - **F for Pairwise distances:** 显示两类之间的F比值矩阵。



## 9.3 SPSS在判别分析中的应用

Step07: 设置分类参数与判别结果

单击【Classify】按钮，在弹出的对话框中可以设置判别分析的分类参数及结果。具体选项含义如下。





## 9.3 SPSS在判别分析中的应用

- ① **【Prior Probabilities (先验概率)】** 选项组：选择先验概率。
  - All groups equal: 各类先验概率相等，系统默认选项。若分为m类，则各类先验概率均为 $1/m$ 。
  - Compute from group sizes: 基于各类样本量占总样本量的比例计算先验概率。
- ② **【Use Covariance Matrix (使用协方差矩阵)】** 栏选择分类使用的协方差矩阵：
  - Within-groups: 使用合并组内协方差矩阵进行分类。
  - Separate-groups: 使用各组协方差矩阵进行分类。
- ③ **【Display (输出)】** 选项组：选择输出分类结果。
  - Casewise results: 输出每个观测量的判别分数、实际类、预测类（根据判别函数求得的分类结果）和后验概率等。选择此项后，下面的**【Limits cases to (将个案限制在前)】**项被激活，可以在它后面的文本框中输入观测量数n。选择此项则仅输出前n个观测量。
  - Summary table: 输出分类的小结表。
  - Leave-one-out classification: 输出对每一个观测量进行分类的结果，所依据的判别函数是由除该观测量以外的其他观测量导出的。



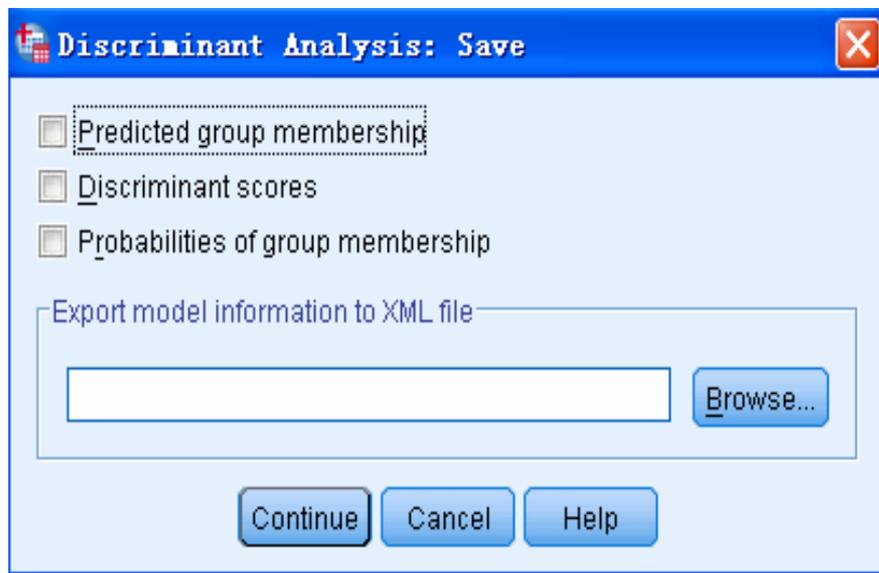
## 9.3 SPSS在判别分析中的应用

- ④ 【Plots (图)】选项组：选择输出统计图。
  - Combined-groups：生成全部类的散点图。该图是根据前两个判别函数值作的散点图。如果只有一个判别函数，就输出直方图。
  - Separate-groups：对每一类生成一张散点图。如果只有一个判别函数，就输出直方图。
  - Territorial map：生成根据判别函数值将观测量分到各类去的边界图。每一类占据一个区域。各类均值在各区中用星号标出。如果仅有一个判别函数，则不作此图。
- ⑤ 缺失值处理方式。
  - Replace missing value with mean：用该变量的均值代替缺失值。

## 9.3 SPSS在判别分析中的应用

### Step08: 结果保存设置

单击【Save】按钮，在弹出的对话框中可以设置判别分析的结果输出，具体选项含义如下。





## 9.3 SPSS在判别分析中的应用

- Predicted group membership: 建立新变量（系统默认变量名是dis\_1）保存预测观测量所属类的值。
- Discriminant score: 建立新变量保持判别分数。
- Probabilities of group membership: 建立新变量保存各个观测量属于各类的概率值。有m类，对一个观测量就会给出m个概率值，因此建立m个新变量。



## 9.3 SPSS在判别分析中的应用

### Step09 相关统计量的Bootstrap估计

单击【Bootstrap】按钮，在弹出的对话框中可以进行如下统计量的Bootstrap估计。

- 标准化典则判别函数系数表支持标准化系数的Bootstrap 估计。
- 典则判别函数系数表支持非标准化系数的Bootstrap 估计。
- 分类函数系数表支持系数的Bootstrap 估计。

Step10: 单击【OK】按钮，结束操作，SPSS软件自动输出结果。



## 9.3 SPSS在判别分析中的应用

### 9.3.3 实例分析：全国30个省市经济增长差异研究

#### 1. 实例内容

现要研究全国30个省市地区经济增长差异性，收集相关数据见数据文件9-3.sav。表中相关变量的含义分别是：x1—经济增长率（%）、x2—非国有化水平（%）、x3—开放度（%）、x4—市场化程度（%）。其中，辽宁、河北等省市归为一类，而黑龙江、吉林等省市归为另一类。请分析江苏、安徽和浙江的类别。



## 9.3 SPSS在判别分析中的应用

### 2. 实例操作

由于案例中已经将北京、上海、四川等省市按照经济增长特点分类，现在需要将另外三个待估省市：江苏、安徽和陕西分类。因此，可以利用判别分析来判别它们的归属。



## 9.3 SPSS在判别分析中的应用

### 3 实例结果及分析

#### (1) 判别分析概述表

SPSS软件首先给出了进行判别分析的概述表9-20。可以看到，参加分析的变量总数为30，有效观测量数为27，占90%；包含缺失值或分类变量范围之外的观测量数为3，占10%。

Unweighted Cases <sup>a</sup>		N <sup>a</sup>	Percent <sup>a</sup>
Valid <sup>a</sup>		27	90.0
Excluded <sup>a</sup>	Missing or out-of-range group codes <sup>a</sup>	3	10.0
	At least one missing discriminating variable <sup>a</sup>	0	.0
	Both missing or out-of-range group codes and at least one missing discriminating variable <sup>a</sup>	0	.0
	Total <sup>a</sup>	3	10.0
Total <sup>a</sup>		30	100.0



## 9.3 SPSS在判别分析中的应用

### (2) 分组统计表

下表给出了观测量按照类别不同进行的基本描述性统计量输出，其中包括均值 (Mean)、均方差 (Std. Deviation) 和有效观测量的个数等。可以从结果初步看到，不同类之间省市经济指标的差异比较明显，例如第一类省份的“非国有化水平”指标均值等于65.0282，而第二类却只有40.1081。

类别		Mean	Std. Deviation	Valid N (listwise)	
				Unweighted	Weighted
1	经济增长率	15.7364	3.33175	11	11.000
	非国有化水平	65.0282	10.72709	11	11.000
	开放度	25.1491	21.26090	11	11.000
	市场化程度	74.3500	7.16398	11	11.000
2	经济增长率	11.5625	3.00397	16	16.000
	非国有化水平	40.1081	16.63743	16	16.000
	开放度	9.2281	5.94755	16	16.000
	市场化程度	58.1050	8.53527	16	16.000
Total	经济增长率	13.2630	3.72064	27	27.000
	非国有化水平	50.2607	18.96437	27	27.000
	开放度	15.7144	16.05658	27	27.000
	市场化程度	64.7233	11.31069	27	27.000



## 9.3 SPSS在判别分析中的应用

### (3) 类均值相等检验表

接着给出了不同类之间“经济增长率”等四个指标均值相等的检验结果如下表所示。从结果看到，它们的相伴概率P值都远小于显著性水平0.05，因此，可以认为两个类指标之间的均值存在显著差异，可以进行判别分析。

	Wilks' Lambda	F	df1	df2	Sig.
经济增长率	.684	11.524	1	25	.002
非国有化水平	.567	19.085	1	25	.000
开放度	.754	8.178	1	25	.008
市场化程度	.483	26.778	1	25	.000



## 9.3 SPSS在判别分析中的应用

### (4) 判别分析特征值表

下表为判别函数的特征值表。从表可见，本案例仅有一个判别函数用于分析，特征值 (Eigenvalue) 为1.479，方差百分比 (% of Variance) 为100%，方差累计百分比 (Cumulative %) 为100%，典型相关系数 (Canonical Correlation) 为0.771。

Function <sup>↗</sup>	Eigenvalue <sup>↗</sup>	% of Variance <sup>↗</sup>	Cumulative % <sup>↗</sup>	Canonical Correlation <sup>↗ ↘</sup>
1 <sup>↗</sup>	1.479	100.0	100.0	.772 <sup>↗ ↘</sup>



## 9.3 SPSS在判别分析中的应用

### (5) Wilks' $\lambda$ 表

下表是对判别函数的显著性检验表。其中Wilks'  $\lambda$  值等于0.403，卡方统计量 (Chi-square) 等于20.878，自由度 (df) 等于4，相伴概率P值 (Sig.) 远小于显著性水平0.05，因此认为判别函数有效。

Test of Function(s)	Wilks' Lambda	Chi-square	df	Sig.
1	.403	20.878	4	.000



## 9.3 SPSS在判别分析中的应用

### (6) 标准化判别函数系数

下表给出了标准化判别函数的系数，于是得到标准化判别函数如下：

Function=0.190\*经济增长率+0.242\*非国有化水平+0.360\*开放度+0.648\*市场化程度

根据判别系数看到，“市场化程度”变量对判别结果的影响是最大的，这是因为它的系数值最大，等于0.648；相反的，“经济增长率”变量对判别结果的影响最小。

	Function
	1
经济增长率	.190
非国有化水平	.242
开放度	.360
市场化程度	.648



## 9.3 SPSS在判别分析中的应用

### (7) 结构矩阵表

结构矩阵表如下表所示，是判别变量与标准化函数之间的合并类内相关系数，变量按照相关系数的绝对值大小排列，表面判别变量与判别函数之间的相关性，如变量“市场化程度”与判别函数关系最密切。

	Function
	1
市场化程度	.851
非国有化水平	.718
经济增长率	.558
开放度	.470



## 9.3 SPSS在判别分析中的应用

### (8) 非标准化判别函数系数

下表给出了非标准化判别函数系数，非标准判别函数为：  
Function=-7.263+0.060\*经济增长率+0.017\*非国有化水平+  
0.028\*开放度+0.081\*市场化程度  
根据这个判别函数代入各变量数值可以计算出判别值。

	Function
	1
经济增长率	.060
非国有化水平	.017
开放度	.025
市场化程度	.081
(Constant)	-7.263



## 9.3 SPSS在判别分析中的应用

### (9) 判别函数类心表

下表给出的是按照非标准判别函数计算的函数类心，即判别函数在各类均值处的判别分数值。可以看到，在两个类心处，判别分数值差异较大。

类别	Function
	1
1	1.411
2	-.970



## 9.3 SPSS在判别分析中的应用

### (10) 分类过程概述表

下表给出了分类过程概述情况。可以看到，共有30个观测量参与了分类过程，没有缺失变量存在。

Processed		30
Excluded	Missing or out-of-range group codes	0
	At least one missing discriminating variable	0
Used in Output		30



## 9.3 SPSS在判别分析中的应用

### (11) 类先验概率表

下表给出了类先验概率表，按照先前的判别分析设置，先验概率都等于0.5。

类别	Prior	Cases Used in Analysis	
		Unweighted	Weighted
1	.500	11	11.000
2	.500	16	16.000
Total	1.000	27	27.000



## 9.3 SPSS在判别分析中的应用

### (12) 分类函数系数表

下表给出了Fisher线性判别函数的系数，因此可以建立各类线性判别模型。

类型一：

$$F1 = -54.567 + 1.812 * \text{经济增长率} - 0.337 * \text{非国有化水平} - 0.058 * \text{开放度} + 1.380 * \text{市场化程度}$$

类型二：

$$F2 = -36.746 + 1.669 * \text{经济增长率} - 0.377 * \text{非国有化水平} - 0.119 * \text{开放度} + 1.188 * \text{市场化程度}$$

将代判别的省市的各类经济指标代入上述两个判别函数进行计算，二者比较大小，如果 $F1 > F2$ ，对应的省市归入1类；否则，当 $F1 < F2$ ，对应的省市归入2类。



## 9.3 SPSS在判别分析中的应用

	类别	
	1	2
经济增长率	1.812	1.669
非国有化水平	-.337	-.377
开放度	-.058	-.119
市场化程度	1.380	1.188
(Constant)	-54.567	-36.746



## 9.3 SPSS在判别分析中的应用

### (13) 判别分析分类结果表

下表列出了最后判别分析的分类结果。可以看到，第一类的11个省市中，只有一个省市（广西省）判别错误，判别方法指出它应该归于第二类；同时，第二类中的16个省市全部判对。同时，数据文件中新增加变量“Dis\_1”列出了所有省市的判别结果。对于待判别省市来说，江苏和安徽被判属第一组，陕西被判属第二组，这与实际情况较吻合。



## 9.3 SPSS在判别分析中的应用

		类别	Predicted Group Membership		Total
			1	2	
Original	Count	1	10	1	11
		2	0	16	16
		Ungrouped cases	2	1	3
	%	1	90.9	9.1	100.0
		2	.0	100.0	100.0
		Ungrouped cases	66.7	33.3	100.0
a. 96.3% of original grouped cases correctly classified.					



## 第10章SPSS在调查问卷数据处理的应用

### 10.1 调查问卷数据处理概述

# 10.1.1 数据整理与转换



CONCEPT  
RATE

- 1、使用目的

调查问卷收集以后，需要先对调查问卷的结果进行一些整理，如对文字型的问题进行事前或事后编码，按变量分组、合并、加权、重新定义或计算新变量等，为最终的统计分析做准备。这些功能集中在Data和Transform菜单项中，下面将以了解高校毕业生就业意愿情况进行调查而获得的一份问卷为例，介绍一些常用的功能。



# 数据整理与转换

您的性别：男 女 您所学专业名称：\_\_\_\_\_ 年级：\_\_\_\_\_

1.你在班级里的学习成绩排名：

前10% 11%—30% 31%—70% 最后30%

2.您参加了今年的考研：

参加了 未参加（跳答一题）

3.您参加考研是否有本科毕业就业难方面的原因：

主要是 有一些 没有

4.本科毕业以后 您选择

参加工作 考研 边工作边考研

到国外 自主创业 暂时什么都不做



# 数据整理与转换

5.您一般通过哪些途径获取招聘信息？

- 招聘会      ■互联网      ■同学、朋友、熟人  
■报刊杂志    □职介机构    □其他

6.对您而言，选择职业时哪些因素影响较大（请选三项并排序）：

- 1单位类型及规模    □就业地区选择    □工资水平及福利  
        2有利于个人发展及晋升    3对工作本身的兴趣    工作稳定性  
        □工作的环境及舒适性    □父母意见  
        □学校老师影响            □其他

7. 您求职要求的工资底线        2000        元。

8. 你认为最理想的签约时间是        大四第一学期末        。

# 数据整理与转换

CONCEPT  
STRATE

- 2、基本原理

- (1)单项选择题的编码

- (2)多项选择题的编码

- (3)排序题的编码

- (4)开放式问题的编码

- (5)缺失值的编码

- (6)“不适用情况”的编码

- (7)数据转换

- 3、其他注意事项

用户缺失值与系统缺失值(System Missing)的含义不同。系统缺失值主要是指计算机默认的缺失方式，如果在输入数据时空缺了某些数据或输入了非法的字符，计算机就把其界定为缺失值，这时的数据标记为“.”，而用户界定的缺失值则不会在数据显示时出现“.”。

## 10.1.2 调查问卷数据的SPSS操作详解

CONCEPT  
STRATE

问卷调查数据的整理与转换的操作主要由以下几个模块来实现。

- (1) **【Transform→Compute Variable (转换→计算变量)】** 对原始数据进行四则运算等，进而派生出新的变量。
- (2) **【Transform→Recode into some Variable (转换→重新编码为相同变量)】** 和 **【Transform→Recode into Different Variable (转换→重新编码为不同变量)】**，重新编码数据，重新安排次序。
- (3) **【Transform→Count Occurrences of Value within Cases (转换→对个案内的值计数)】**，创建一个新变量用以计算某些变量共同发生的频次(即计数)。



## 10.2 调查问卷缺失值处理方法

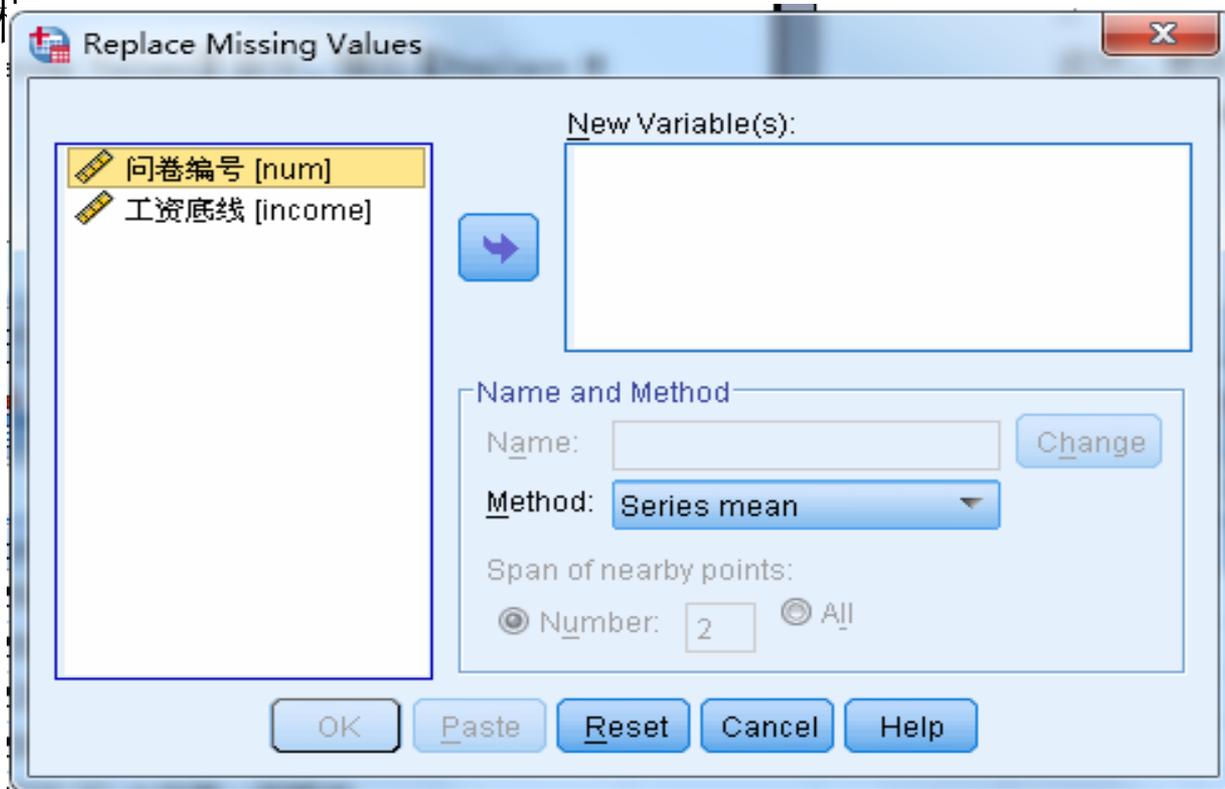
### 10.2.1 缺失值的类型与处理方法

缺失值的类型 : 完全随机缺失; 随机缺失 ; 完全非随机缺失

缺失值的处理方法 : 删除法和插补法

## 10.2.2 替换缺失值的SPSS操作详解

- **Step01:** 打开【Replace Missing Values (替换缺失值)】对话框  
选择菜单栏中的【Transform(转换)】→【Replace Missing Values(替换缺失值)】命令，弹出【Replace Missing Values(替换缺失值)】对话框



# 替换缺失值的SPSS操作详解

CONCEPT  
STRATE

## Step02: 选择检验变量

在该对话框左侧的候选变量列表框中选择一个或几个变量，将其移入【New Variable(s) (新变量)】列表框中，这时系统自动产生用于替代缺失值的新变量，用户也可在【name(名称)】框处自己定义替代缺失值的新变量名。

# 替换缺失值的SPSS操作详解

CONCEPT  
STRATE

## Step03 : 选择替换缺失值的方法

在【Method(方法)】下拉下箭头选择缺失值的替代方式。

●Series mean: 用该变量的所有非缺失值的均数做替代。

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

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

# 替换缺失值的SPSS操作详解

CONCEPT  
STRATE

**Linear interpolation:** 线性插值法填补缺失值。用该列数据缺失值前一个数据和后一个数据建立插值直线，然后用缺失点在线性插值函数的函数值填充该缺失值。

**Linear trend at point:** 缺失点处的线性趋势法。应用缺失值所在的整个序列建立线性回归方程，然后用该回归方程在缺失点的预测值填充缺失值

# 替换缺失值的SPSS操作详解

CONCEPT  
STRATE

## Step04：其他选项设置

当选择的替换缺失值的方法为【Mean of nearby points (临界点的均值)】或【Median of nearby points (临界点的中位数)】时，选项【Span of nearby points (临界点的跨度)】处于激活状态，可以选择取相邻点的跨度。

## Step05：单击【OK】按钮，结束操作，SPSS软件自动输出结果。

如果分析中没有用到含缺失值的变量，可以不用关心缺失值问题。在SPSS相关的分析过程中，选择“按对排除个案 (P)”，这时如果没有用到含缺失值的变量，缺失值对分析没有影响；如果选择“按列表排除个案 (L)”，含有缺失值的个案将不会用于分析，可能会造成信息损失。

# 10.2.3 实例图文分析：高校毕业生就业 意愿调查

CONCEPT  
STRATE

## 1. 实例内容

就业意愿描述的是大学生寻找工作之前的设想，这种设想与现实的匹配程度会影响其能否实现就业。为了深入了解毕业生的就业意向，了解大学生的就业意向和将来的就业形势，为进一步完善毕业生就业工作提供导向和决策依据，进行了毕业生就业意愿调查。假设有一个由 17 名毕业生的调查问卷组成的简单随机样本，其中对于工资底线这一题的回答存在缺失，要求对这些进行缺失值替换。

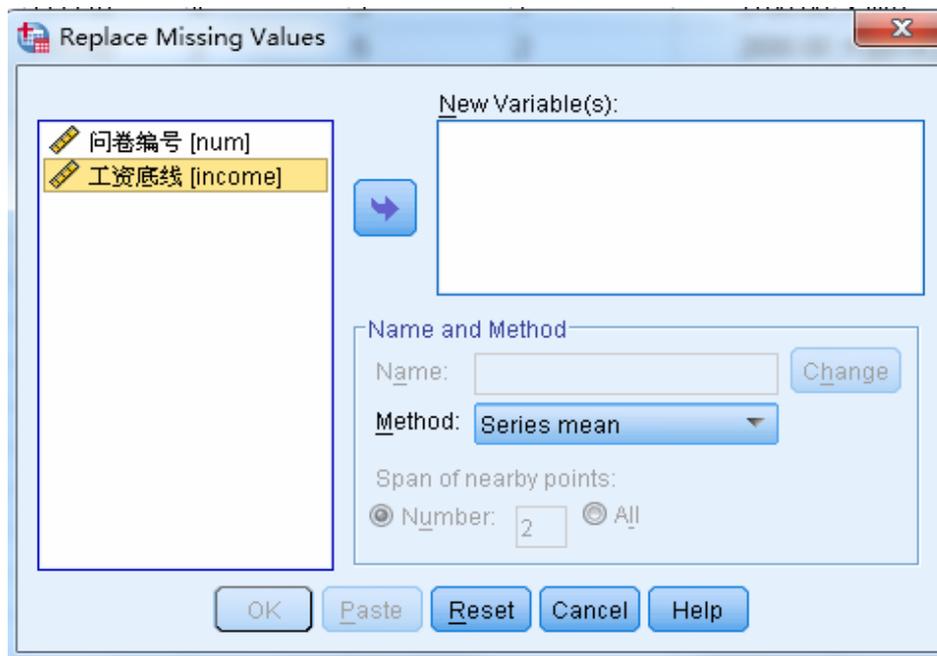
# 实例操作

CONCEPT  
STRATE

## Step01: 打开对话框

打开SPSS软件，选择菜单栏中的【Transform(转换)】→【Replace Missing Values(替换缺失值)】命令，弹出如下图所示的对话框。

# 实例操作



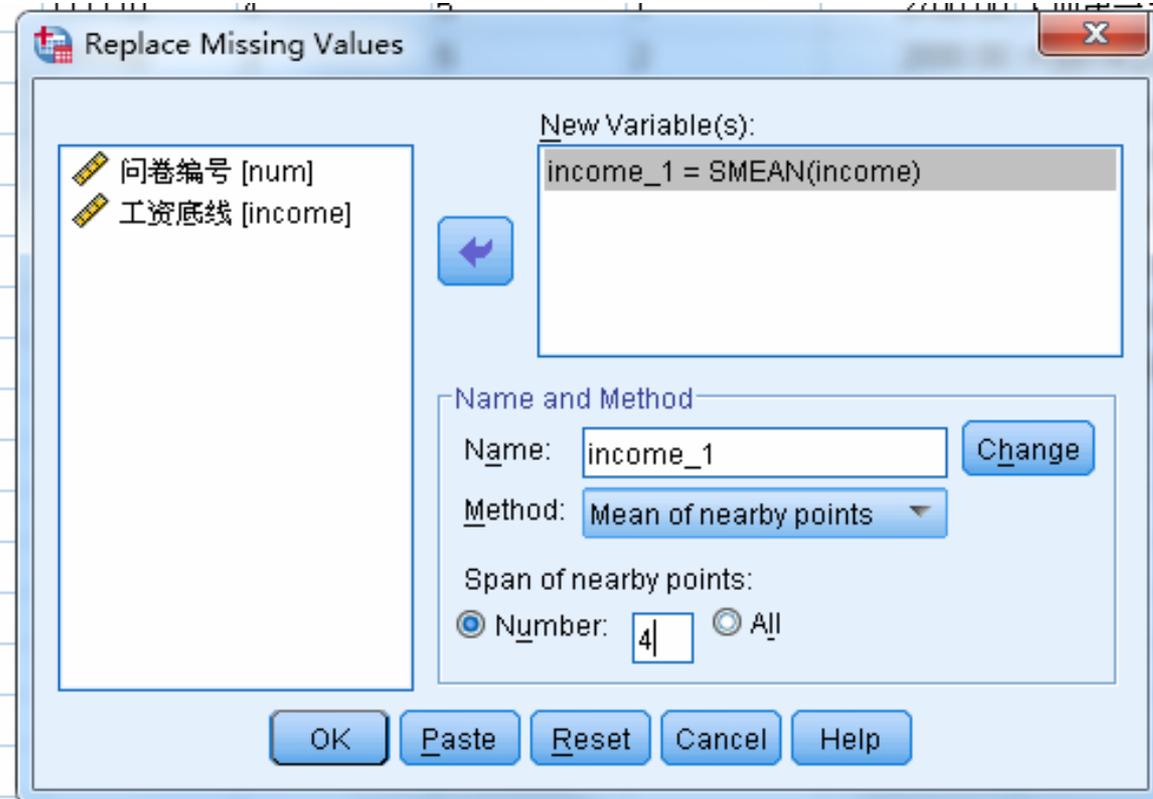
# 实例操作

CONCEPT  
STRATE

**Step02:** 在左侧的候选变量列表框中选择“工资底线”变量进入【New Variable(s) (新变量)】列表框，这时系统自动产生用于替代缺失值的新变量，用户也可在Name框处自己定义替代缺失值的新变量名。在【Method】下拉列表框中选择替换方法【Mean of nearby points (临界点的均值)】，并在【Span of nearby points (临界点的跨度)】文本框中输入“4”。

注意：进行缺失值替换时，只能对数字型变量进行缺失值替换。

# 实例操作



# 实例操作



## Step03: 完成操作

最后，单击【OK(确定)】按钮，操作完成。此时，原数据文件新增加了“income1”变量。

# 实例操作

CONCEPT  
STRATE

income	time	income_1
2500.00	大四第一学期末	2500.00
	大四第一学期末	2593.33
2200.00	大四5月以前	2200.00
2000.00	大四3月以前	2000.00
2500.00	大四第一学期末	2500.00
3000.00	大四第一学期末	3000.00
	大四4月以前	2593.33
1800.00	大四3月以前	1800.00
2300.00	大四第一学期末	2300.00
2700.00	大四第一学期末	2700.00
2000.00	大四5月以前	2000.00
2300.00	大四3月以前	2300.00
3000.00	大四第一学期末	3000.00
4000.00	大四5月以前	4000.00
2800.00	大四4月以前	2800.00
3000.00	大四5月以前	3000.00
2800.00	大四3月以前	2800.00

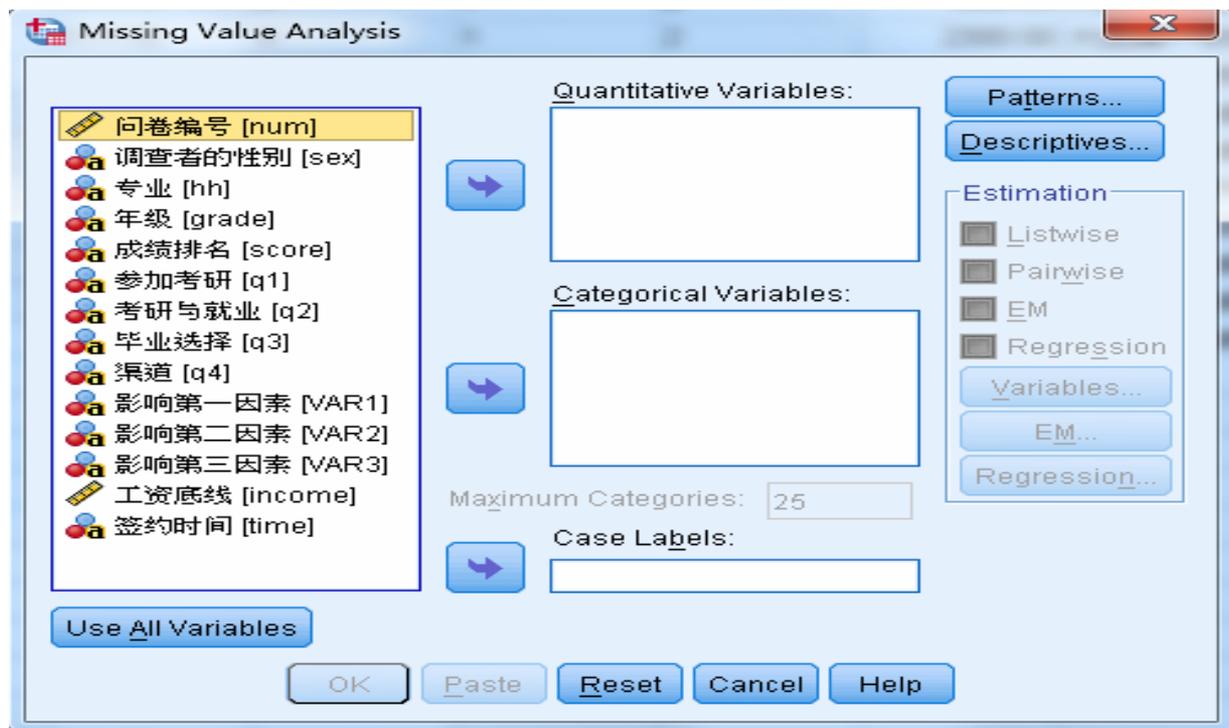
## 10.2.4 缺失值分析的SPSS操作详解

CONCEPT  
RATE

**Step01** : 打开【Missing Values Analysis (缺失值分析)】对话框  
选择菜单栏中的【Analyze (分析)】→【Missing Value Analysis (缺失值分析)】命令，弹出【Missing Value Analysis (缺失值分析)】对话框。



## 10.2.4 缺失值分析的SPSS操作详解





## 10.2.4 缺失值分析的SPSS操作详解

### Step02：选择检验变量

在该对话框左侧的候选变量列表框中选择一个或几个变量，将其移入【Quantitative Variables(定量变量)】或【categorical Variables(分类变量)】列表框中。定量变量是选择进入缺失值分析的变量。

### Step03：选择缺失值估计的方法

在【Estimation(估计)】列表框中选择缺失值的处理，从而对参数进行方式。

- Listwise: 分析时按列表排除个案，将缺失值排除在外，从而对变量进行分析。
- Pairwise: 按配对的方式对缺失值进行分析。
- EM: 用Expectationt Maxiumum方法对缺失值进行修补。
- Regression: 用线性回归的方法对对缺失值进行修补。



## 10.2.4 缺失值分析的SPSS操作详解

### Step04：其他选项设置

【Patterns (模式)】包含输出的模式、变量缺失的模式等五个部分。

(1) Display: 输出部分。

● Tabulated cases ,grouped by missing value patterns: 按照缺失值分组的表格模式。

● Cases with missing value ,sorted by missing value patterns: 按照缺失值排序的个案模式。

● All cases ,optionally sorted by selected variable: 按照选定变量指定顺序的所有个案。

(2) variables: 变量

● Missing Patterns for: 缺失模式。

(3) Additional information for: 附加信息。

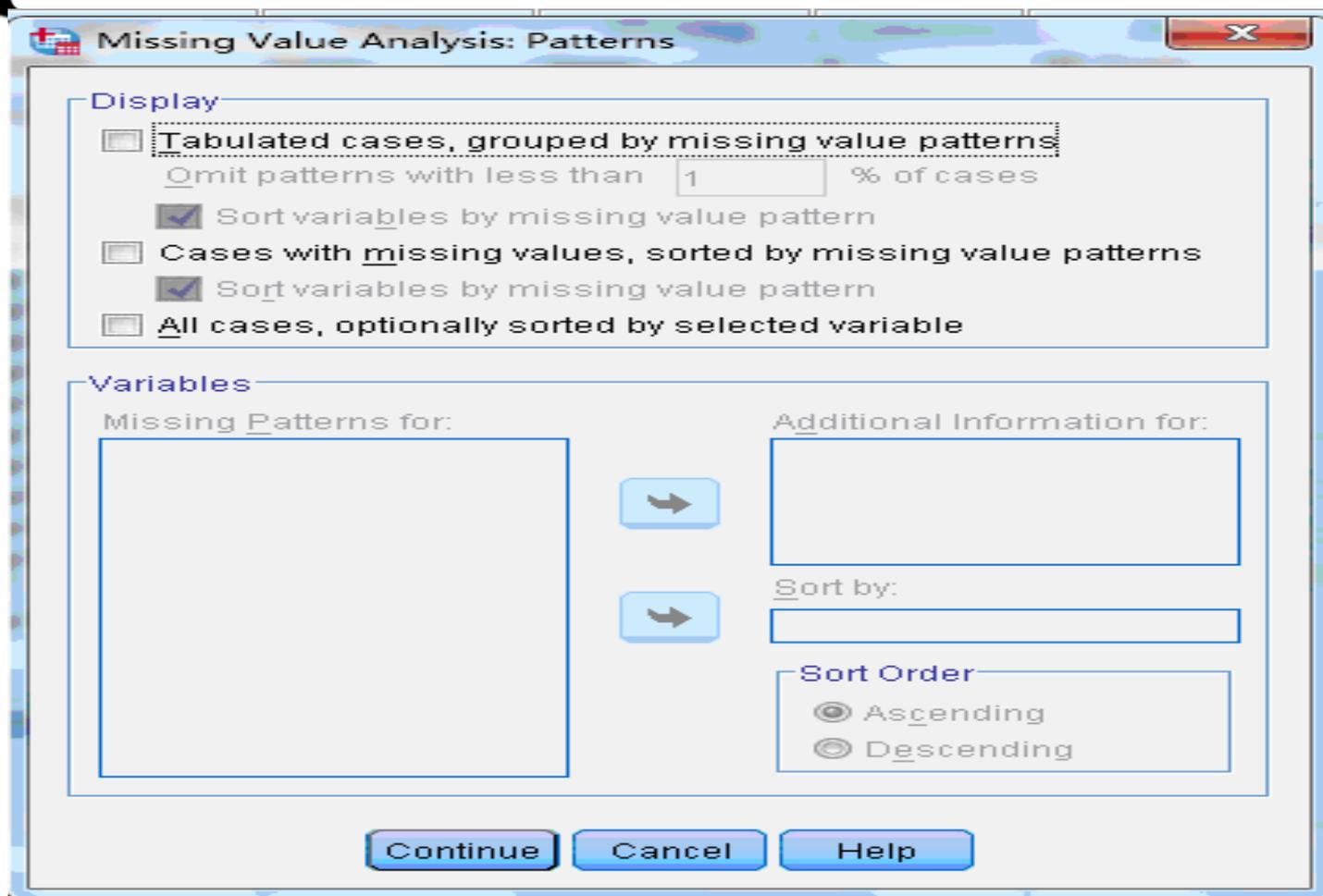
(4) Sort by: 排序依据。

(5) Sort Order : 排序顺序。

● Ascending : 升序。

● Descending: 降序。

## 10.2.4 缺失值分析的SPSS操作详解



## 10.2.4 缺失值分析的SPSS操作详解

**【Descriptives (描述)】** 主要对单变量统计量和指示变量统计量、忽略缺失值占总个案数的比例三部分。

(1)Univariate Statistics: 单变量统计量。

(2)Indicator variable Statistics: 指示变量统计量。

- Percent mismatch:百分比不匹配。

- t tests with groups formed by indicator variable: 使用有指示变量形成的分；组进行的T检验。

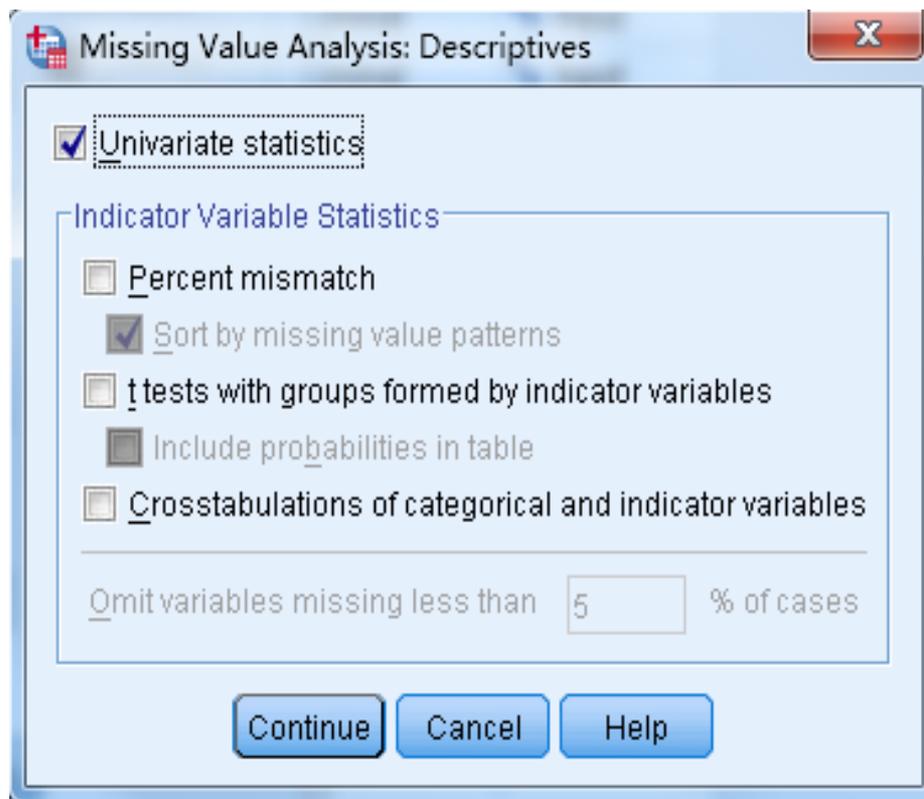
- Cross tabulations of categorical and indicator variable: 为分类变量和指示变量生成交叉表。

(3)Omit variables missing less than ()% of cases : 忽略缺失值占总个案数的比例小于的变量。

Step05 : 单击**【OK】**按钮，结束操作，SPSS软件自动输出结果。



## 10.2.4 缺失值分析的SPSS操作详解



## 10.2.5 实例图文分析： 一维正态随机数的缺失分析

CONCEPT  
STRATE

### 1. 实例内容

相关系数为-0.4的二维正态随机变量的2000个观测值，其边缘分布分别为均值为0.2，标准差为0.2的正态随机变量 $w_1$ ，和均值为0.3，标准差为0.1的正态随机变量 $w_2$ ，随机删除变量 $w_1$ 中的3%数据，随机删除变量 $w_2$ 中的5%数据，现在进行缺失值分析。

## 2 实例操作

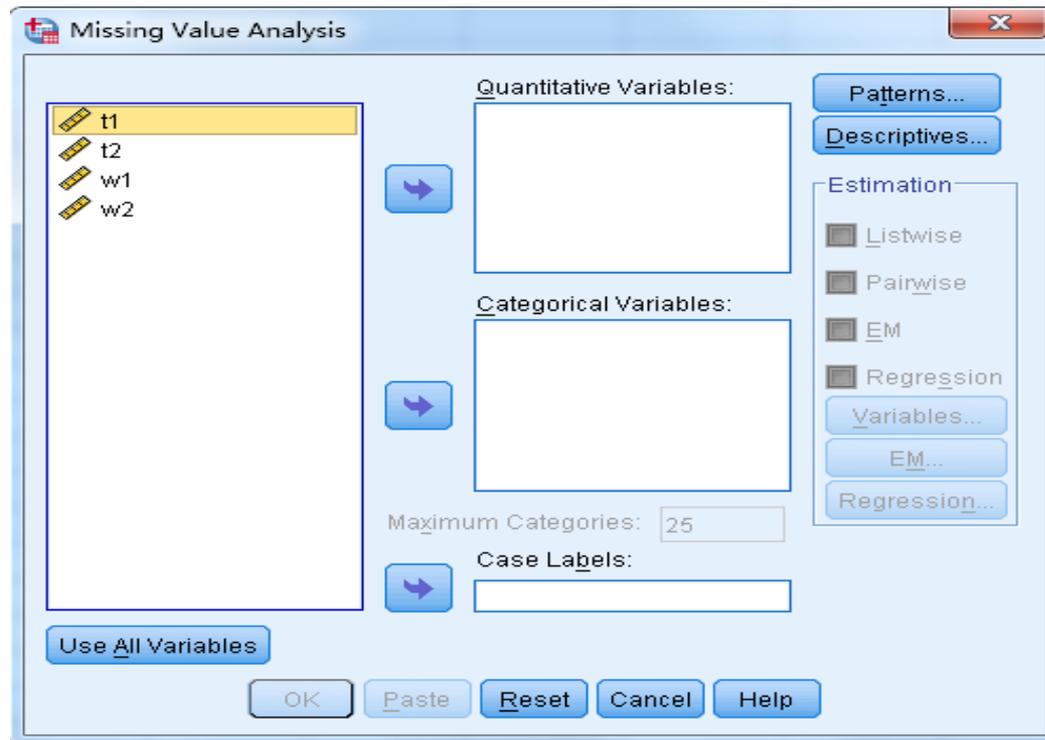


### Step01: 打开对话框

打开SPSS软件，选择菜单栏中的【Analyze(分析)】  
→【Missing Value Analysis(缺失值分析)】命令，弹出对话框。



## 2 实例操作



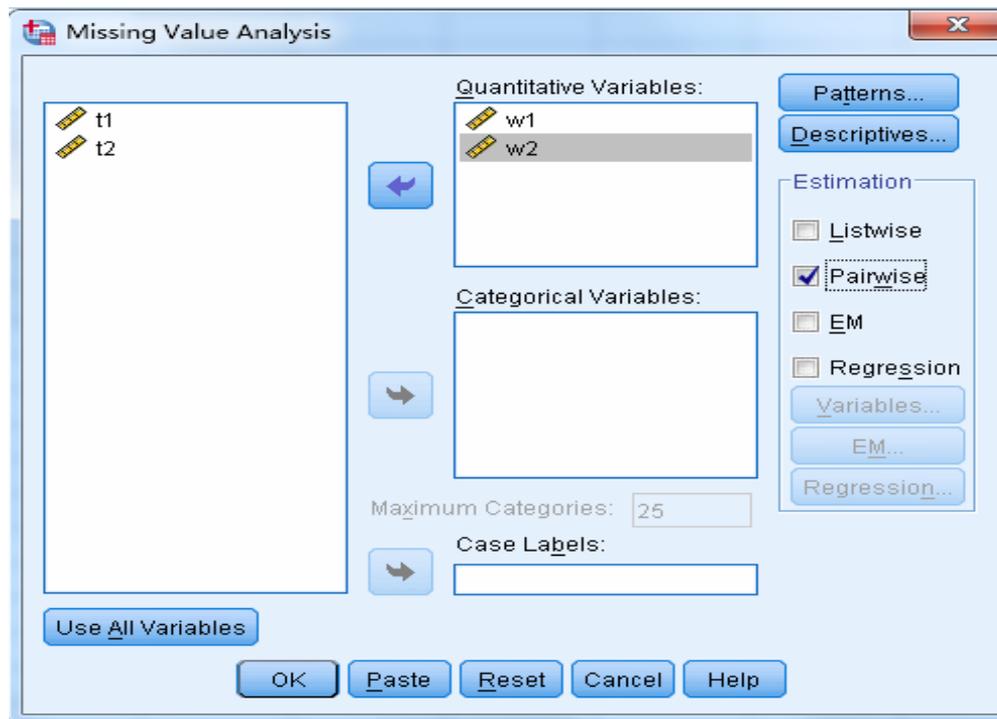
## 2 实例操作

CONCEPT  
STRATE

**Step02:** 在左侧的候选变量列表框中选择“w1”、“w2”变量进入【Quantitative Variables(定量变量)】列表框，在【Estimation(估计)】选项组中选择【Pairwise(成对)】复选框。



## 2 实例操作



## 2 实例操作

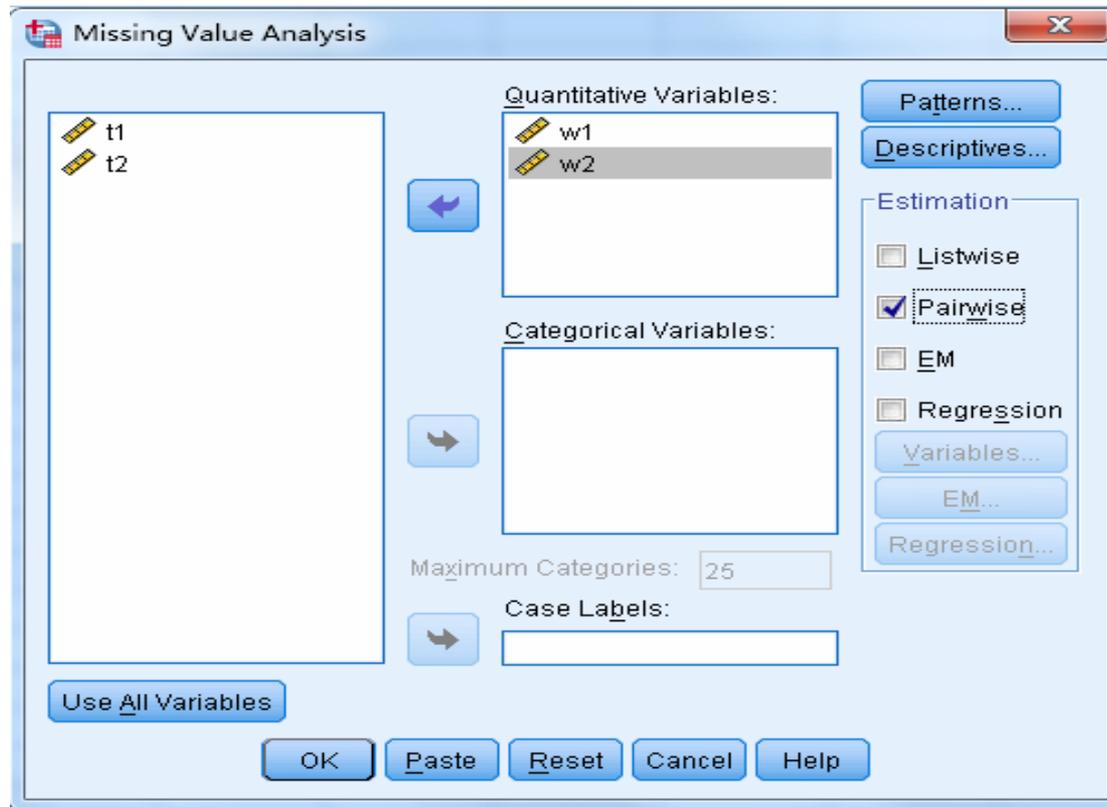


### Step03: 完成操作

最后，单击【OK(确定)】按钮，操作完成。此时，软件输出结果出现在结果浏览窗口中。



## 2 实例操作



# 3 实例结果及分析

CONCEPT  
STRATE

## (1) 单变量的基本统计信息汇总表

执行完上面操作后，在SPSS结果报告中首先给出的是两个变量的基本统计分析，见表10-3所示。变量w1数据个数为1940，缺失60个数据，缺失的百分比为3%，样本均值为0.20，标准差为0.19，比 $Q1-1.5*IQR$ 小的数据有5个，比 $Q3+1.5*IQR$ 大的数据有8个；



### 3 实例结果及分析

	N	Mean	Std. Deviation	Missing		No. of Extremes <sup>a</sup>	
				Count	Percent	Low	High
w1	1940	.20055337	.198140912	60	3.0	5	8
w2	1900	.30283098	.098391053	100	5.0	3	4

a. Number of cases outside the range (Q1 - 1.5\*IQR, Q3 + 1.5\*IQR).



### 3 实例结果及分析

#### (2) 配对分析结果

##### ① 两变量配对的频数

这里，变量w1数据个数为1940，变量w2数据个数为1900，变量w1和变量w2的配对数据个数为1842。

	w1	w2
w1	1940	
w2	1842	1900



### 3 实例结果及分析

#### ②两变量配对的均值

这里，变量w1的样本均值是0.20，变量w2的样本均值为0.30，变量w1的1940个数据在变量w2都不缺失的情况下的均值为0.20028339，变量w2的1900个数据在变量w1都不缺失的情况下的均值为0.30283098。

	w1	w2
w1	.20055337	.30266749
w2	.20028339	.30283098
Mean of quantitative variable when other variable is present.		



### 3 实例结果及分析

#### ③两变量配对的样本标准差

这里，变量w1的样本标准差是0.198140912，变量w2的样本标准差为0.098391053，变量w1的1940个数据在变量w2都不缺失的情况下的均值为0.199486851，变量w2的1900个数据在变量w1都不缺失的情况下的均值为0.098391053

	w1	w2
w1	.198140912	.098699395
w2	.199486851	.098391053

Standard deviation of quantitative variable when other variable is present.



### 3 实例结果及分析

- (4) 两变量配对的样本协方差

这里，变量w1的样本方差是0.039259821，变量w2的样本方差为0.009680799，配对的变量w1与变量w2的样本协方差为-0.007154109。

	w1	w2
w1	.03925982 1	
w2	<b>-.0071541 09</b>	.0096807 99



### 3 实例结果及分析

(5) 两变量配对的样本相关系数

配对的变量w1与变量w2的样本协方差为-0.363。

	w1	w2
w1	1	
w2	<b>-.363</b>	1

# 10.3 调查问卷的信度分析

CONCEPT  
STRATE

## 10.3.1 信度分析概述

### 1、使用目的

为了保证问卷具有较高的可靠性和有效性，在形成正式问卷之前，应当对问卷进行试测，并对试测结果进行信度和效度分析，根据分析结果筛选问卷题项，调整问卷结构，从而提高问卷的信度和效度。

信度分析是评价调查问卷是否具有稳定性和可靠性的有效的分析方法。

# 信度分析概述

## 2、基本原理

**重测信度法**是用同样的问卷对同一组被调查者间隔一定时间重复施测，计算两次施测结果的相关系数，适用于事实式问卷，如性别、出生年月等在两次施测中不应有任何差异。重测信度法属于稳定系数。

**复本信度法**是让同一组被调查者一次填答两份问卷复本，计算两个复本的相关系数。复本信度属于等值系数。

**折半信度法**是将调查项目分为两半，计算两半得分的相关系数，进而估计整个量表的信度。折半信度属于内在一致性系数，测量的是两半题项得分间的一致性。这种方法一般适用于态度、意见式问卷的信度分析。

**克朗巴哈信度系数法**是评价的量表中各题的得分之间一致性的，属于内在一致性系数。这种方法适用于态度、意见式问卷的信度分析，是目前最常用的信度系数，其公式为：

$$\alpha = \frac{k\bar{r}}{1 + (k-1)\bar{r}}$$

其中， $k$ 为调查问卷中题项的总数， $\bar{r}$ 为个项目相关系数的均值。



## 10.3.2 信度分析的SPSS操作详解

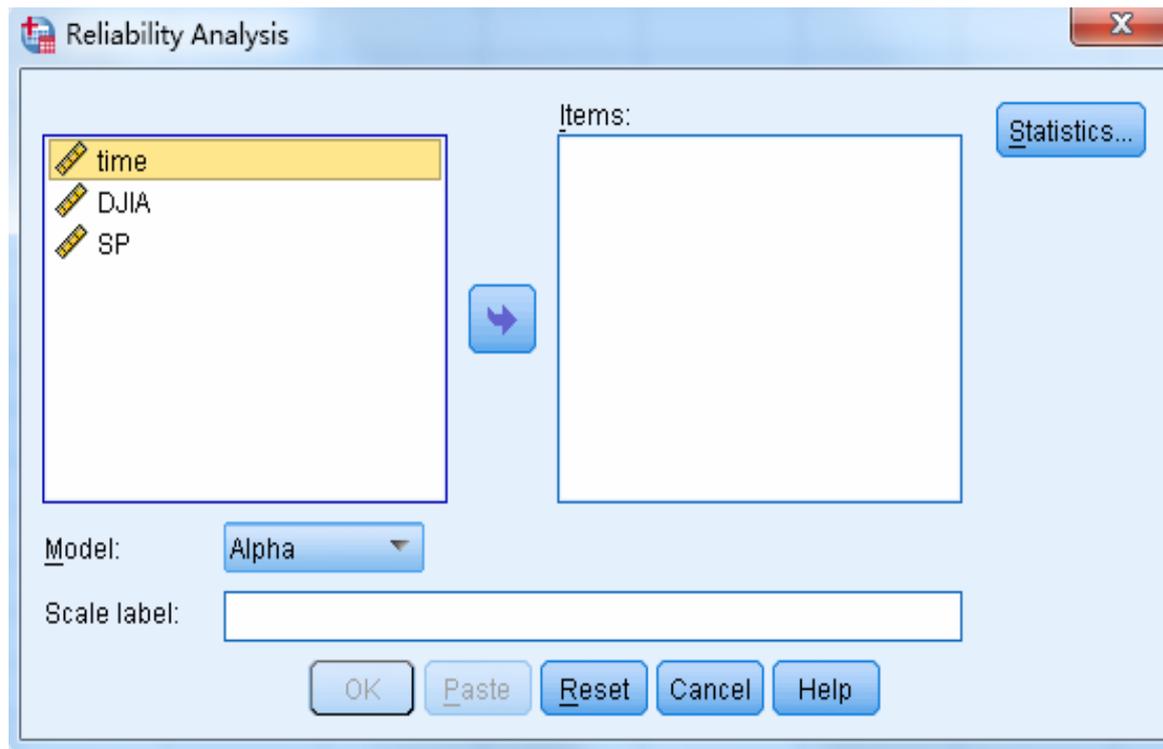
**Step01** : 打开【Reliability Analysis(可靠性分析)】对话框  
选择菜单栏中的【Analyze(分析)】→【Scale(度量)】→【Reliability Analysis(可靠性分析)】命令，弹出【Reliability Analysis(可靠性分析)】对话框。

**Step02** : 选择信度分析变量

在该对话框左侧的候选变量列表框中选择一个或几个变量，将其移入【items(项)】列表框中，选择进入信度分析的变量。

【Scale label(度量标签)】主要对信度分析的信度系数做一个标签。

# 信度分析的SPSS操作详解



# 信度分析的SPSS操作详解

CONCEPT  
STRATE

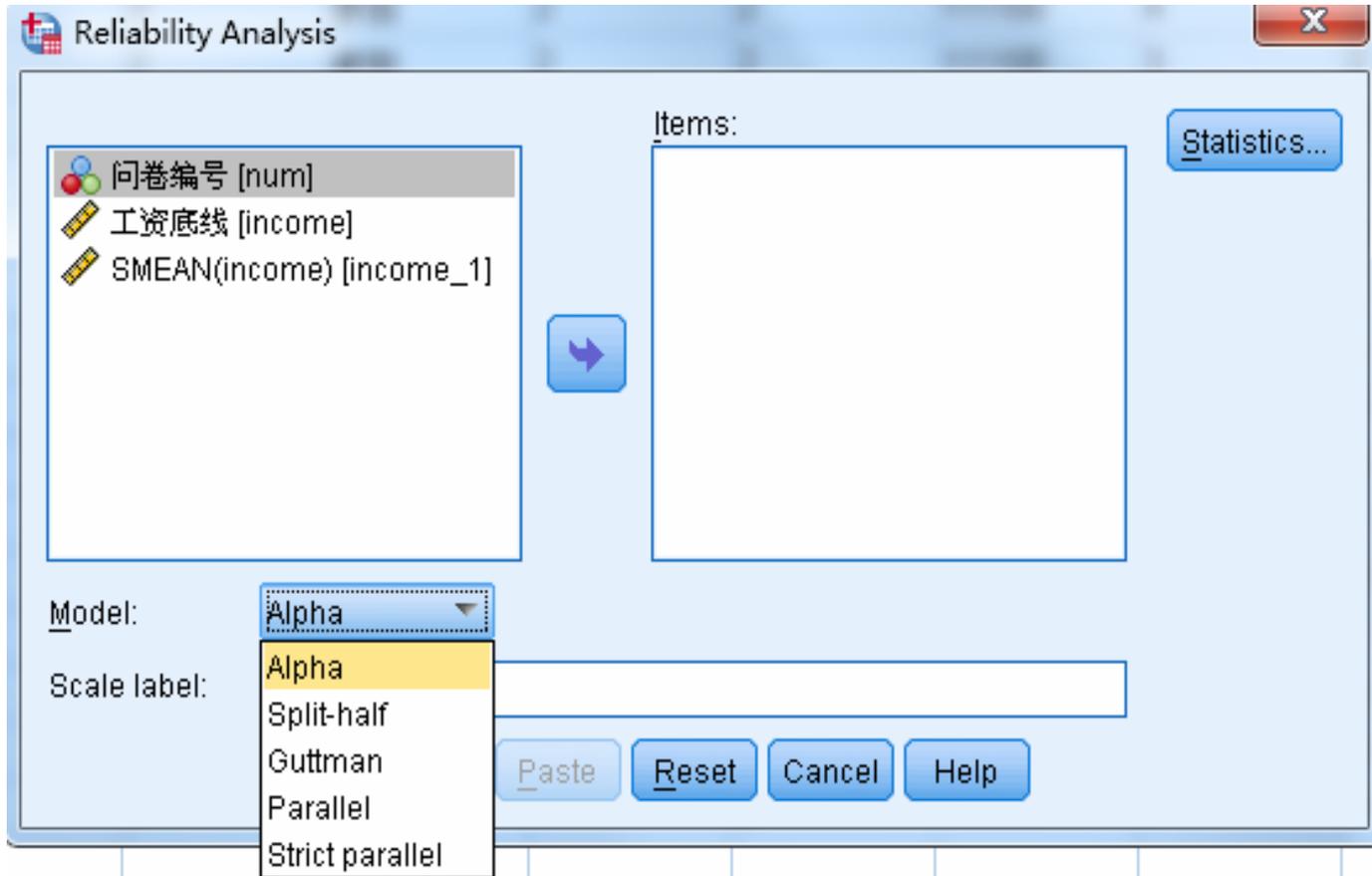
## Step03 : 选择信度分析的方法

在【Model(模型)】下拉列表框中选择信度分析的信度系数，从而对变量进行信度分析。

- Alpha : 克朗巴哈 (Cronbach) 信度系数法。
- Split-half: 折半信度系数。
- Guttman: Guttman最低下限真实信度法。
- Parallel: 各题目变异数同质时的最大概率 (maximum-likelihood) 信度。
- Strict parallel: 各题目平均数与变异数均同质时的最大概率信度。

# 信度分析的SPSS操作详解

CONCEPT  
STRATE



# 信度分析的SPSS操作详解

CONCEPT  
STRATE

## Step04 : 其他选项设置

【Statistics (统计量)】包含Hotelling的检验，Friedman等级变异数分析、Tukey的可加性检验等统计分析。

● Descriptives for: Item表示输出各评估项目的基本描述性统计，Scale表示输出各评估项目的总分的基本描述性统计，Scale if item deleted表示输出剔除某评项目后的均值、方差、协方差等基本统计量，从而对评估项目进行逐个评估。



# 信度分析的SPSS操作详解

Reliability Analysis: Statistics

Descriptives for

- Item
- Scale
- Scale if item deleted

Inter-Item

- Correlations
- Covariances

Summaries

- Means
- Variances
- Covariances
- Correlations

ANOVA Table

- None
- F test
- Friedman chi-square
- Cochran chi-square

- Hotelling's T-square
- Tukey's test of additivity
- Intraclass correlation coefficient

Model: Two-Way Mixed      Type: Consistency

Confidence interval: 95 %      Test value: 0

Continue    Cancel    Help

# 信度分析的SPSS操作详解

CONCEPT  
STRATE

● **【Inter-Item】** 选项组: Covariances、Correlations分别表示输出各评估项目的协方差系数矩阵和相关系数矩阵。

● **【Summaries:Means】** 选项组: 输出评估项目总分的平均分的基本描述性统计, Variance表示评估项目总分的样本方差的描述性统计, Covariances、Correlations分别输出评估项目总和的协方差矩阵、相关系数矩阵的描述性统计。

● **【ANOVA Tables】** 选项组: 提供了多种方法进行检验同一评估对象在评估项目上的得分是否具有一致性。None表示什么检验都不做, F Test表示进行反复测试的方差分析, 只适合于定距型的正态分布数据; Friedman chi-squared对配对样本的进行Friedman检验, 适合于非正态分布或定序型数据, Cochran chi-square表示进行多配对样本的Cochran检验, 适合于二值型数据。

Step05: 单击 **【OK】** 按钮, 结束操作, SPSS软件自动输出结果。

# 10.3.3 实例图文分析： 员工素质评估的信度分析



CONCEPT  
RATE

## 1. 实例内容

为评估某个公司员工的素质设计一套评价表格，其中包括的评价项目有：科学素质、文化素质、经济素质、道德素质，每个评估项目的满分25分，四个项目评估的总分100分，分数越高素质越高。为了研究评价体系的可信性，随机对30名员工进行了测试，现利用这些数据进行分析。

# 实例内容



姓名	科学素质	文化素质	经济素质	道德素质
小李	21	22	22	22
小张	20	21	22	23
小莫	20	21	22	22
小蔡	21	21	22	22
小毛	22	22	23	24
小华	22	22	23	23

注意：此表为部分数据

# 实例操作

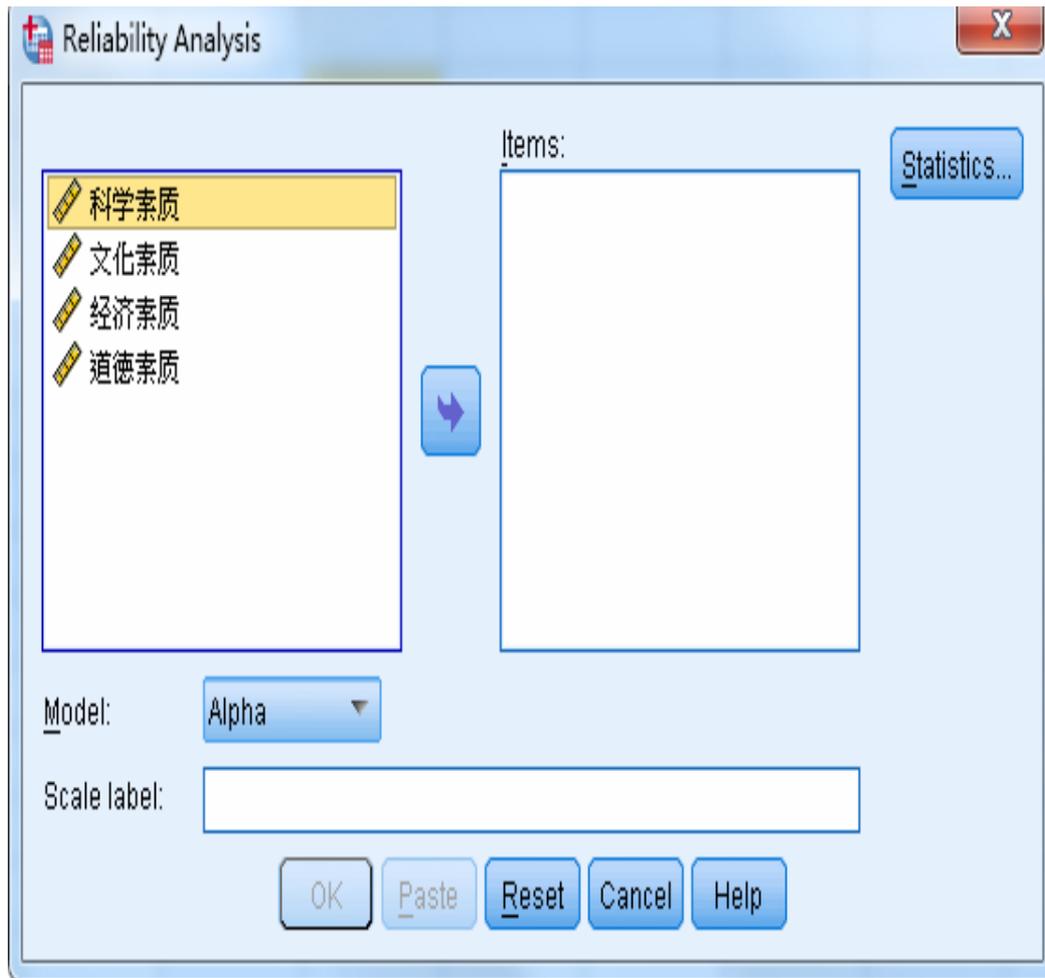
CONCEPT  
STRATE

## Step01: 打开对话框

打开SPSS软件，选择菜单栏中的【Analyze(分析)】→【Scale(度量)】→【Reliability Analysis(可靠性分析)】命令，弹出【Reliability Analysis(可靠性分析)】对话框。

# 实例操作

CONCEPT  
STRATE



# 实例操作

CONCEPT  
STRATE

**Step02:** 在左侧的候选变量列表框中选择“科学素质”、“文化素质”、“经济素质”、“道德素质”进入【items(项)】列表框，在【Model(模型)】下拉列表框中选择【Alpha()】选项，并单击【Statistics(统计量)】按钮进入【Statistics(统计量)】对话框。

# 实例操作

CONCEPT  
STRATE

Reliability Analysis: Statistics

Descriptives for

- Item
- Scale
- Scale if item deleted

Inter-Item

- Correlations
- Covariances

Summaries

- Means
- Variances
- Covariances
- Correlations

ANOVA Table

- None
- F test
- Friedman chi-square
- Cochran chi-square

Hotelling's T-square

Tukey's test of additivity

Intraclass correlation coefficient

Model: Two-Way Mixed

Type: Consistency

Confidence interval: 95 %

Test value: 0

Continue Cancel Help

# 实例操作

CONCEPT  
STRATE

**Step03:** 勾选【Scale if item deleted (如果项已删除则进行度量)】、【Correlations (相关性)】以及【Summaries (摘要)】复选框，然后单击【Continue (继续)】按钮，进入信度分析对话框。

# 实例操作

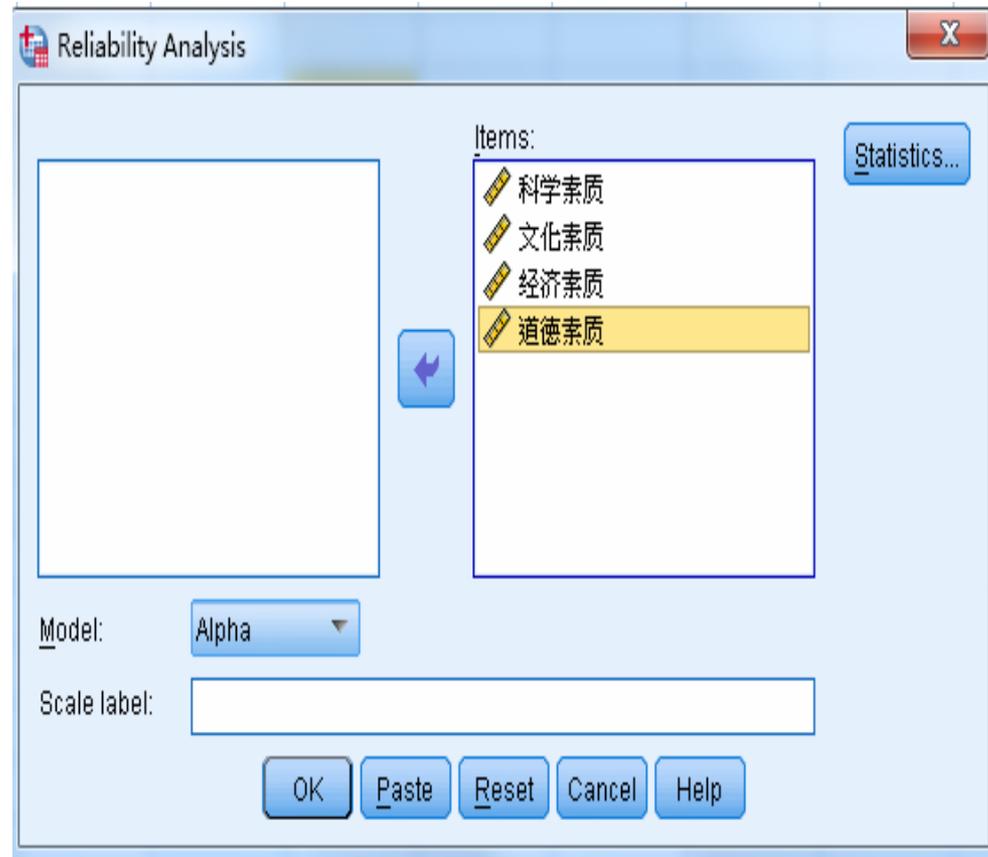


## Step04: 完成操作

最后，单击【OK(确定)】按钮，操作完成。此时，软件输出结果出现在结果浏览窗口中。

# 实例操作

CONCEPT  
STRATE



# 实例结果及分析

CONCEPT  
STRATE

## (1) 信度分析进行过程的摘要

执行完上面操作后，在SPSS结果报告中首先给出的是信度分析进行过程的摘要，见下表所示。信度分析的有效数据为30个，排除在外的数据个数为0，整个信度分析是基于所有数据来进行的。



# 实例结果及分析

		N	%
Cases	Valid	30	100.0
	Excluded <sup>a</sup>	0	.0
	Total	30	100.0

a. Listwise deletion based on all variables in the procedure.

# 实例结果及分析

CONCEPT  
RATE

## (2) 信度分析的信度系数计算的结果

在SPSS结果报告中给出克隆巴哈（Cronbach）信度系数的估计值为0.816，基于标准化评估项目（Based on Standardized Items）调整的克隆巴哈（Cronbach）信度系数为0.825，评估项目数为4个。由于信度系数在0.80~0.90之间，说明问卷调查中的题目具有较强的内在一致性。

# 实例结果及分析

CONCEPT  
RATE

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
.816	.825	4

# 实例结果及分析



## (3) 各个评估项目的相关系数矩阵

从相关系数矩阵可以看出，科学素质与文化素质之间相关系数为0.734，具有较强的正相关性；道德素质与经济素质之间相关系数为0.691，正相关性较强，文化素质与道德素质之间的相关系数为0.343，是相关性最低的两个项目。



# 实例结果及分析

	科学素质	文化素质	经济素质	道德素质
科学素质	1.000	.734	.691	.430
文化素质	.734	1.000	.487	.343
经济素质	.691	.487	1.000	.561
道德素质	.430	.343	.561	1.000

# 实例结果及分析

CONCEPT  
STRATE

## (4) 评估项目的描述性统计

下表的第一行显示了30名员工在4个评估项目上总分的均值为21.992，最大值为23.067，最小值21.000，全距2.067，样本均值的方差为0.812；第二行显示30名员工在4个评估项目上总分的样本方差为0.585，最大值为0.754，最小值0.437，全距0.317，样本方差的方差为0.026；可见，各个项目的平均分基本相当，各项评分的差异性比较平衡。

第三行显示4个评估项目协方差的均值为0.308，最大值为0.414，最小值0.202，全距0.211，样本方差的方差为0.006；第四行显示4个评估项目相关系数的均值为0.541，最大值为0.734，最小值0.343，全距0.391。可见，各个评项目的相关程度校稿，而且相关程度的差异较小。



# 实例结果及分析

	Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	21.992	21.000	23.067	2.067	1.098	.812	4
Item Variances	.585	.437	.754	.317	1.726	.026	4
Inter-Item Covariances	.308	.202	.414	.211	2.045	.006	4
Inter-Item Correlations	.541	.343	.734	.391	2.139	.021	4

# 实例结果及分析

CONCEPT  
STRATE

## (5) 剔除某个评估项目以后的结果

表10-12的第一列显示了剔除某个评估项目以后的剩余项目的总平均分，例如剔除了科学素质的剩余其他三项的总平均分为66.97，是第一列中最大的，这说明科学素质的得分影响比较大；第二列显示了剔除某个评估项目以后的剩余项目总分的样本方差，第三列是某评估项目与其余评估项目总分的简单相关系数，例如科学素质与剩余其他三项的总分之间的简单项系数为0.750，这再一次说明科学素质的地位比较重要；第四列是某评估与其余评估项目的复相关系数，反映了该评估项目与其余评估项目的总体相关程度；最后一列是剔除某个评估项目以后的剩余项目计算得到克朗巴哈（Cronbach）信度系数。



# 实例结果及分析

	Scale Mean if Item Deleted	Scale Variance if Item Deleted	Corrected Item-Total Correlation	Squared Multiple Correlation	Cronbach's Alpha if Item Deleted
科学素质	66.97	3.137	.750	.685	.710
文化素质	66.40	3.903	.622	.542	.777
经济素质	65.63	3.757	.718	.565	.740
道德素质	64.90	3.610	.506	.321	.840

## 10.3.4 实例进阶分析： 折半信度系数的分析

CONCEPT  
RATE

如果在实例操作的第二步中，在【Model(模型)】的下拉框中选择的列表框中选择【Split-half(半分)】，其他不变，那么，所进行的信度分析就是折半信度法，其结果会出现在折半项目的前提下所得到克朗巴哈（Cronbach）信度系数，和在折半项目的前提下得到评估项目的描述性统计，如下表所示。

### (1) 折半项目的前提下得信度分析结果

表10-14是在折半项目的前提下得信度分析结果。折半信度法将项目分成两部分，Part 1是关于科学素质与文化素质的，Part 2是关于经济素质与道德素质的，针对Part 1计算得到克朗巴哈（Cronbach）信度系数为0.837，针对Part 2计算得到克朗巴哈（Cronbach）信度系数为0.702，这说明想进一步改进调查问卷的质量，应针对经济素质与道德素质部分进行重新修订量表或增删题项。两部分总分的简单相关系数为0.583，说明两部分具有正相关性。由于两部分的项目是一样的，都是两个项目，一般都应采用Spearman-Brown修正方法对两部分总分的简单相关系数进行修正，修正的结果为0.736，两部分的Guttman Split-Half Coefficient为0.736，说明整个问卷是可行的一份问卷。



# 折半信度系数的分析

Cronbach's Alpha	Part 1	Value	.837
		N of Items	2 <sup>a</sup>
	Part 2	Value	.702
		N of Items	2 <sup>b</sup>
	Total N of Items		4
Correlation Between Forms			.583
Spearman-Brown Coefficient	Equal Length		.736
	Unequal Length		.736
Guttman Split-Half Coefficient			.736
a. The items are: 科学素质, 文化素质.			
b. The items are: 经济素质, 道德素质.			

# 折半信度系数的分析

CONCEPT  
STRATE

## (2) 折半项目的前提下得评估项目的描述性统计

下表显示了30名员工在科学素质，文化素质这两个项目上总分的均值为21.283，在经济素质，道德素质这两个项目上总分的均值为22.700。在科学素质，文化素质这2个评估项目协方差的均值为0.414，在经济素质，道德素质这两个项目上总分的均值为0.322。



# 折半信度系数的分析

		Mean	Minimum	Maximum	Range	Maximum / Minimum	Variance	N of Items
Item Means	Part 1	21.283	21.000	21.567	.567	1.027	.161	2 <sup>a</sup>
	Part 2	22.700	22.333	23.067	.733	1.033	.269	2 <sup>b</sup>
	Both Parts	21.992	21.000	23.067	2.067	1.098	.812	4
Item Variances	Part 1	.575	.461	.690	.229	1.496	.026	2 <sup>a</sup>
	Part 2	.595	.437	.754	.317	1.726	.050	2 <sup>b</sup>
	Both Parts	.585	.437	.754	.317	1.726	.026	4
Inter-Item Covariances	Part 1	.414	.414	.414	.000	1.000	.000	2 <sup>a</sup>



# 折半信度系数的分析

	Part 2	.322	.322	.322	.000	1.000	.000	2 <sup>b</sup>
	Both Parts	.308	.202	.414	.211	2.045	.006	4
Inter-Item Correlations	Part 1	.734	.734	.734	.000	1.000	.000	2 <sup>a</sup>
	Part 2	.561	.561	.561	.000	1.000	.000	2 <sup>b</sup>
	Both Parts	.541	.343	.734	.391	2.139	.021	4

a. The items are: 科学素质, 文化素质.

b. The items are: 经济素质, 道德素质.

# 10.4 调查问卷的多重响应分析

CONCEPT  
STRATE

## 10.4.1 多重响应分析概述

### 1、使用目的

多重响应（**Multiple Response**）是指对同一个问题被调查者可能有多个答案，它是调查研究中十分常见的数据形式。

### 2、基本原理

多重响应资料因其特殊性，不方便应用传统的多元统计分析方法进行研究，利用多重二分法和多重分类法两种数据转换方式可以极大的丰富对其建模的方法。

多重二分法的分类编码为**0**和**1**，即将每一个选项拆分为一个独立变量，如果选中的则录入**1**，没有选择的则录入为**0**。有多少个选项则拆分成多少个变量来，因此选项异常多的情况下此种方法有点麻烦。



## 10.4.2 多重响应分析的SPSS操作详解

**Step01 :** 打开 **【Define Multiple Response Sets (定义多元响应集)】** 对话框

选择菜单栏中的 **【Analyze (分析)】** → **【Multiple Responses (多元响应)】** → **【Define Multiple Response Sets (定义多元响应集)】** 命令，弹出 **【Define Multiple Response Sets (定义多元响应集)】** 对话框。

**Step02 :** 选择多重响应分析变量

在 **【Set Definition (定义集)】** 列表框列出所有的需要设置的变量，其中包括多选题的变量，将候选变量中选择一个或几个变量，将其移入 **【Variables in Set (集合中的变量)】** (集合中的变量) 列表框中，选择进入多重响应分析的变量。

# 多重响应分析的SPSS操作详解

CONCEPT  
TRATE

## Step03 : 设置多重响应集

然后在下方的【Variables Are Coded As (将变量编码为)】(将变量编码)中选择编码的方法。

【Dichotomies (二分法)】为多重二分法，【Counted Value (计数值)】输入需要统计的变量值，例如计数值输入“1”，意思是统计变量值为1的频率。

【Categories (类别)】为多重分类法，【Range (范围)】表示多重分类法的起点值，【Through (到)】表示多重分类法的终值。

【Label (标签)】为多重二分法或多重分类法的值标签的定义。

【name (名称)】为输入该多选题的题目名称。

在【name (名称)】中输入该多选题的题目名称，在【Label (标签)】中输入分类法的值标签的定义之后，点击【add (添加)】到【Multiple Responses Sets (多元响应集)】，点击【Close (关闭)】，就设置好多重响应集。



# 多重响应分析的SPSS操作详解

Define Multiple Response Sets

Set Definition

Variables in Set:

Multiple Response Sets:

Variables Are Coded As

Dichotomies Counted value:

Categories Range:  through

Name:

Label:

Note: Sets defined here are only available in the Multiple Response Frequencies and Crosstabs procedures. Use Define Multiple Response Sets on the Data menu for sets used elsewhere.

Close Help

Add

Change

Remove



# 多重响应分析的SPSS操作详解

## Step04：设置多重响应分析方法

点击【Close(关闭)】，设置好多重响应集，再选择菜单栏中的【Analyze(分析)】→【Multiple Responses(多元响应)】命令，可以看到，多出两个菜单选项。



# 多重响应分析的SPSS操作详解

The screenshot shows the SPSS software interface. The 'Analyze' menu is open, and the 'Multiple Response' option is highlighted. A sub-menu is displayed for 'Multiple Response', containing the following options:

- Define Variable Sets...
- Frequencies...
- Crosstabs...

The background shows a data editor window with a table containing the following data:

品牌11	var
1	
2	
3	
4	
8	
5	
6	
7	
1	
8	
2	
3	
4	
1	

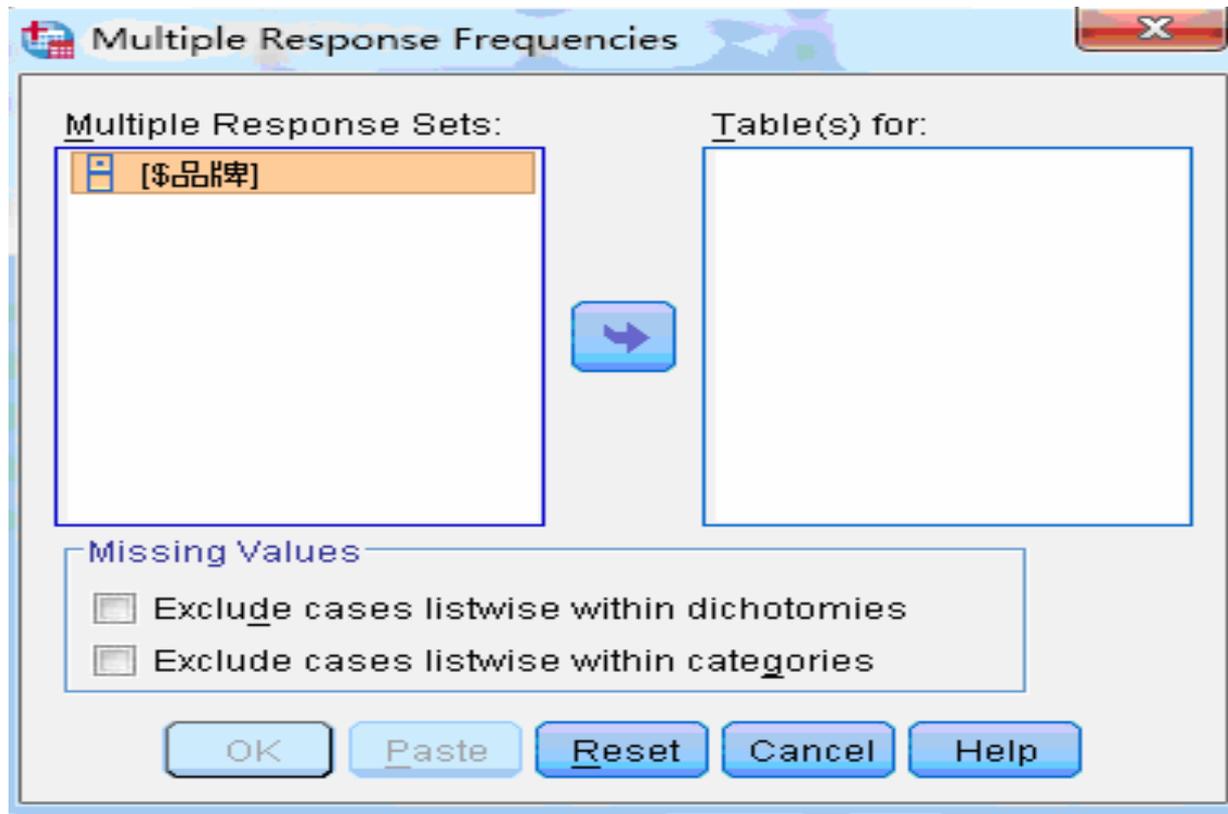
# 多重响应分析的SPSS操作详解

CONCEPT  
STRATE

选择菜单栏中的【Analyze(分析)】→【Multiple Responses(多元响应)】→【Frequencies(频率)】命令，弹出【Multiple Response Frequencies(多元响应频率)】对话框。



# 多重响应分析的SPSS操作详解



# 多重响应分析的SPSS操作详解

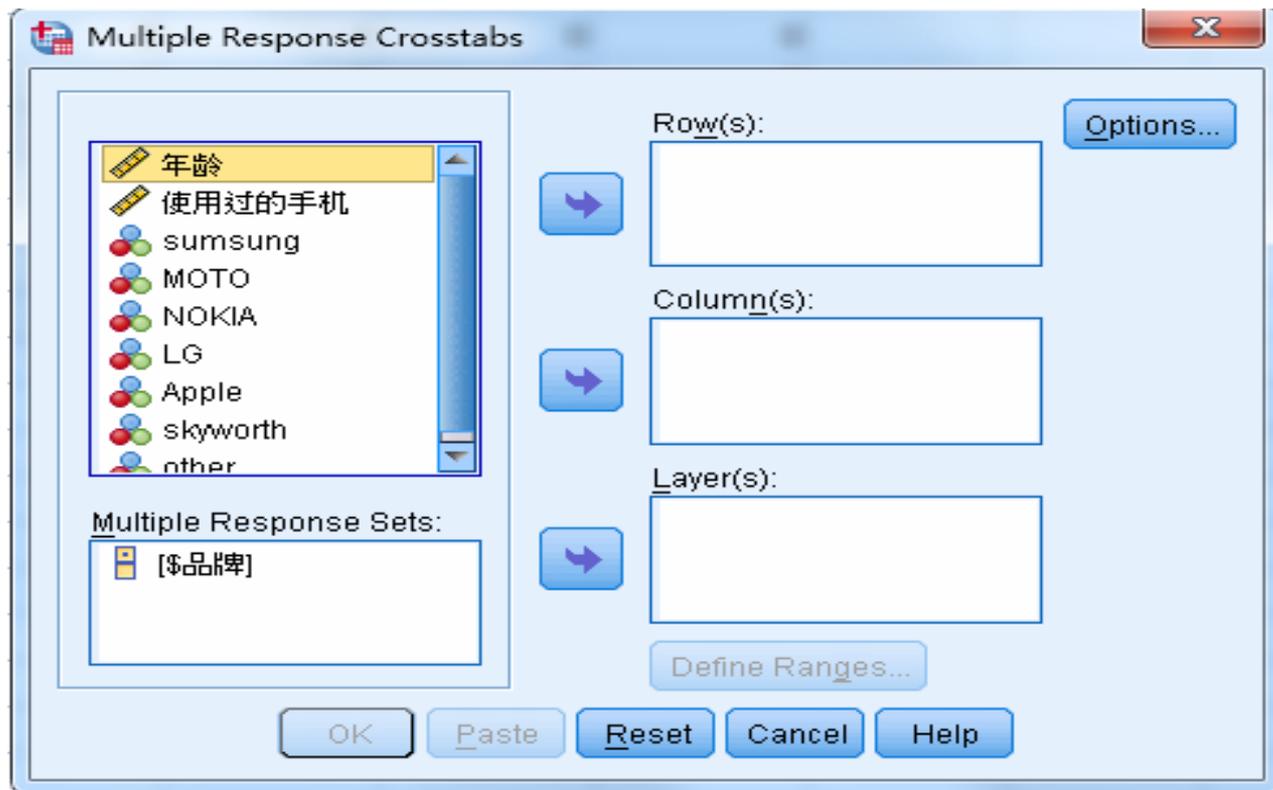
CONCEPT  
RATE

- **【Multiple Responses Sets (多元响应集)】**：显示设置好的多重响应集的名称。
- **【Table(s) for (列表为)】**：表示对选入的多重响应集进行列表分析。
- **【Missing Value (缺失值)】**：表示对缺失值的处理方法。Exclude cases listwise with in dichotomies表示对多重二分法的变量进行缺失值的处理，。Exclude cases listwise with in categories dichos表示对多重分类法的变量进行缺失值的处理。缺失值处理方法都是将缺失值排除在样本外进行频率分析。

选择菜单栏中的**【Analyze (分析)】** → **【Multiple Responses (多元响应)】** → **【Crosstabs (交叉表)】** 命令，进入**【Multiple Response Crosstabs (多元响应交叉表)】** 对话框。



# 多重响应分析的SPSS操作详解



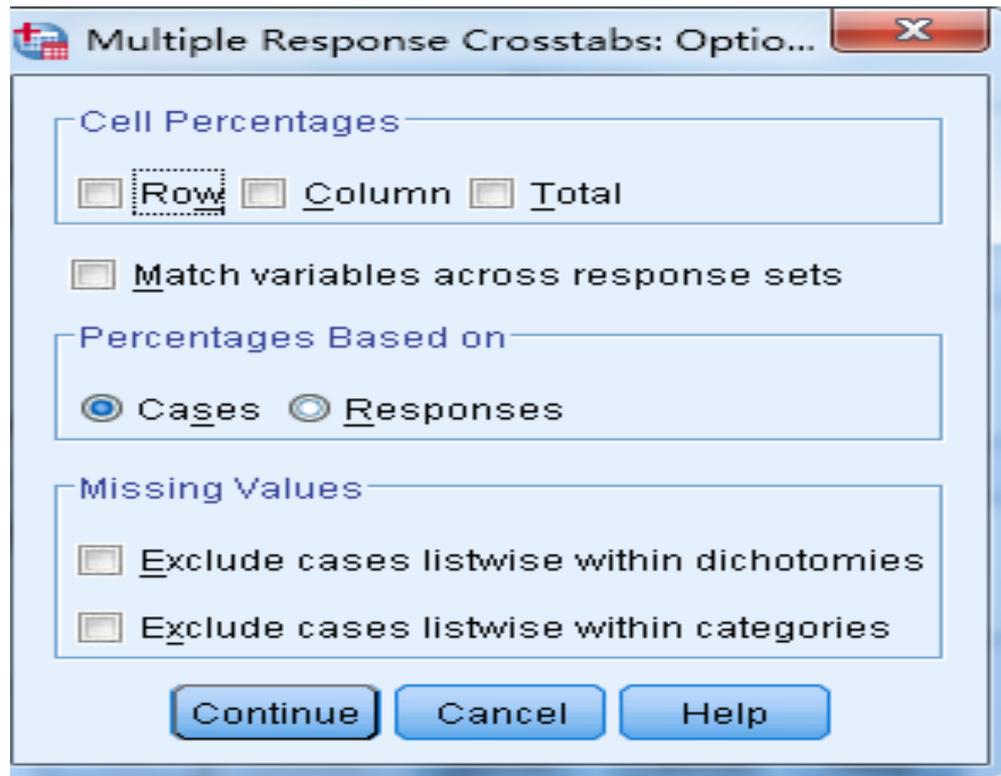
# 多重响应分析的SPSS操作详解

CONCEPT  
STRATE

- **【Multiple Responses Sets (多元响应集)】**：显示设置好的多重响应集的名称。
- **【Row(s)(行)】**：显示交叉分析的行变量。
- **【Column(s)(列)】**：显示交叉分析的列变量。
- **【Layer(s)(层)】**：显示交叉分析的分层变量。
- **【Define Ranges (定义范围)】**：定义行变量、或列表里、或层变量的取值范围。
- **【Options (选项)】**：交叉分析的一些选项，包括单元百分比（行的、列的、总的）、基于哪种百分比（基于个案的、基于响应的）、缺失值的处理（基于多重二分法的变量的、基于多重分类法的变量的）。



# 多重响应分析的SPSS操作详解



## 10.2.5 实例图文分析： 手机市场情况分析



### 1. 实例内容

为调查关于手机市场情况，设计了一份调查问卷，问卷内容包括性别，年龄，当前使用手机的品牌，在过去三年内曾经使用过的手机品牌等。随机对30名路人进行了测试，现利用这些数据进行多重相应分析分析。其中品牌1为三星，品牌2为摩托罗拉，品牌3为诺基亚，品牌4为LG，品牌5为苹果，品牌6为创维，和其他品牌，使用二分编码。定义7个变量，变量名分别为sumsung、MOTO、NOKIA、LG、Apple、Skyworth、other，值标签分别定义为0=“未选”，1=“选中”。定义了性别变量，值标签分别定义为0=“女”，1=“男”。

# 实例内容

CONCEPT  
TRATE

性别	年龄	品牌	使用过的手机	samsung	MOTO	NOKIA	LG	Apple	skyworth	other	性别变量
男	36	三星	24	1	0	0	0	0	0	0	1
女	33	摩托罗拉	54	0	1	0	0	0	0	0	0
男	23	诺基亚	21	0	0	1	0	0	0	0	1
女	28	LG	16	0	0	0	1	0	0	0	0
女	21	摩托罗拉	31	0	1	0	0	0	0	0	0
男	22	TCL	21	0	0	0	0	0	0	1	1
男	17	LG	23	0	0	0	1	0	0	0	1

# 实例操作

CONCEPT  
STRATE

## Step01: 打开对话框

打开SPSS软件，选择菜单栏中的【Analyze(分析)】→【Multiple Responses(多元响应)】→【Define Multiple Response Sets(定义多元响应集)】命令，弹出【Define Multiple Response Sets(定义多元响应集)】对话框。

# 实例操作



# 实例操作

CONCEPT  
STRATE

**Step02:** 在【Set Definition (定义集)】列表框中选择sumsung、MOTO、NOKIA、LG、Apple、Skyworth、other进入【Variables in Set】框，在【Variables Are Coded As (变量编码为)】选项组中选择编码的方法为【Dichotomies (二分法)】，并在【Counted Value (计数值)】文本框输入“1”，在【name (名称)】文本框中输入该多选题的题目名称为“品牌”。

# 实例操作



Define Multiple Response Sets

Set Definition

Variables in Set

- 年龄
- 使用过的手机
- 性别变量

- samsung
- MOTO
- NOKIA
- LG
- Apple
- skyworth
- other

Variables Are Coded As

Dichotomies Counted value: 1

Categories Range: through

Name: 品牌

Label:

Multiple Response Sets:

Add

Change

Remove

Note: Sets defined here are only available in the Multiple Response Frequencies and Crosstabs procedures. Use Define Multiple Response Sets on the Data menu for sets used elsewhere.

Close Help

# 实例操作

CONCEPT  
STRATE

**Step03:** 单击【Add(添加)】按钮将所选选项添加到【Multiple Responses Sets(多元响应集)】列表框，然后再单击【Close(关闭)】按钮，设置好多重响应集。

# 实例操作



Define Multiple Response Sets

Set Definition

Variables in Set:

Multiple Response Sets:

Age  
Used mobile phone  
samsung  
MOTO  
NOKIA  
LG  
Apple  
skyworth  
nther

Variables Are Coded As

Dichotomies Counted value:

Categories Range:  through

Name:

Label:

Add  
Change  
Remove

Note: Sets defined here are only available in the Multiple Response Frequencies and Crosstabs procedures. Use Define Multiple Response Sets on the Data menu for sets used elsewhere.

Close Help

\$品牌

# 实例操作

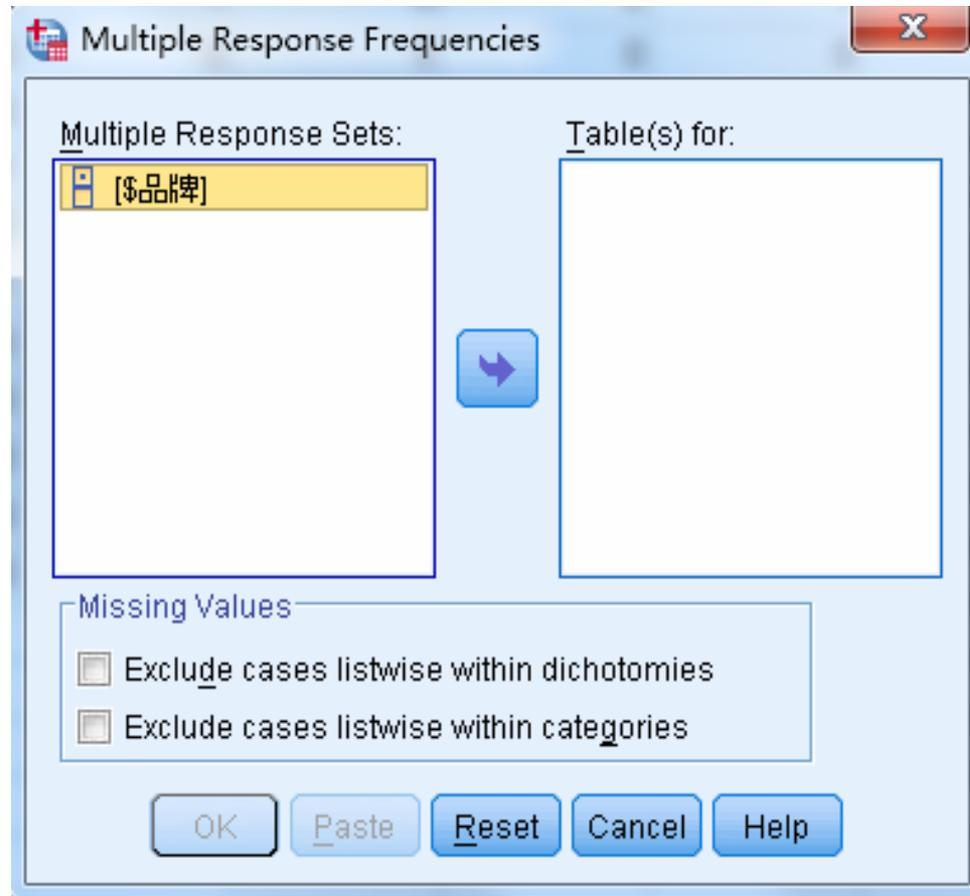
CONCEPT  
STRATE

**Step04:** 打开多重响应频数分析对话框:

选择菜单栏中的【Analyze(分析)】→【Multiple Responses(多元响应)】→【Frequencies(频率)】命令, 进入【multiple Response Frequencies(多元响应频率)】对话框。

# 实例操作

CONCEPT  
STRATE



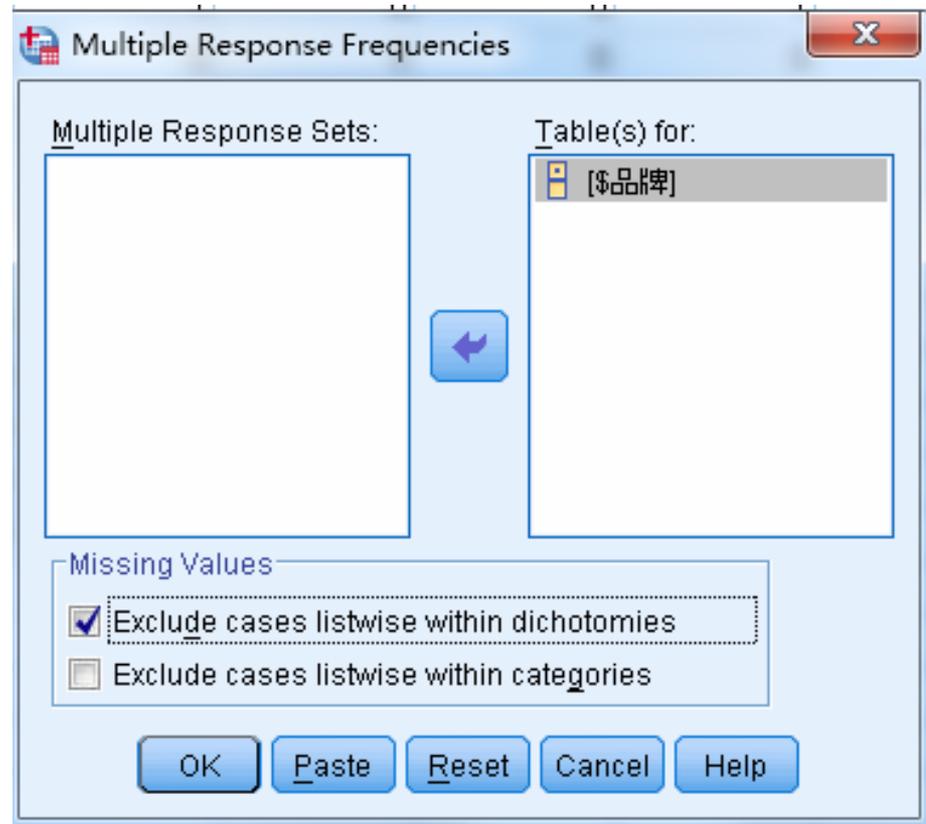
# 实例操作

CONCEPT  
STRATE

**Step05:** 多重响应频数分析的设置:

将【Multiple Responses Sets (多元响应集)】中的多重响应集“品牌”选入【Table(s) for (列表为)】，并在【Missing Value (缺失值)】选项组中勾选【Exclude cases listwise with in dichotomies】复选框。

# 实例操作



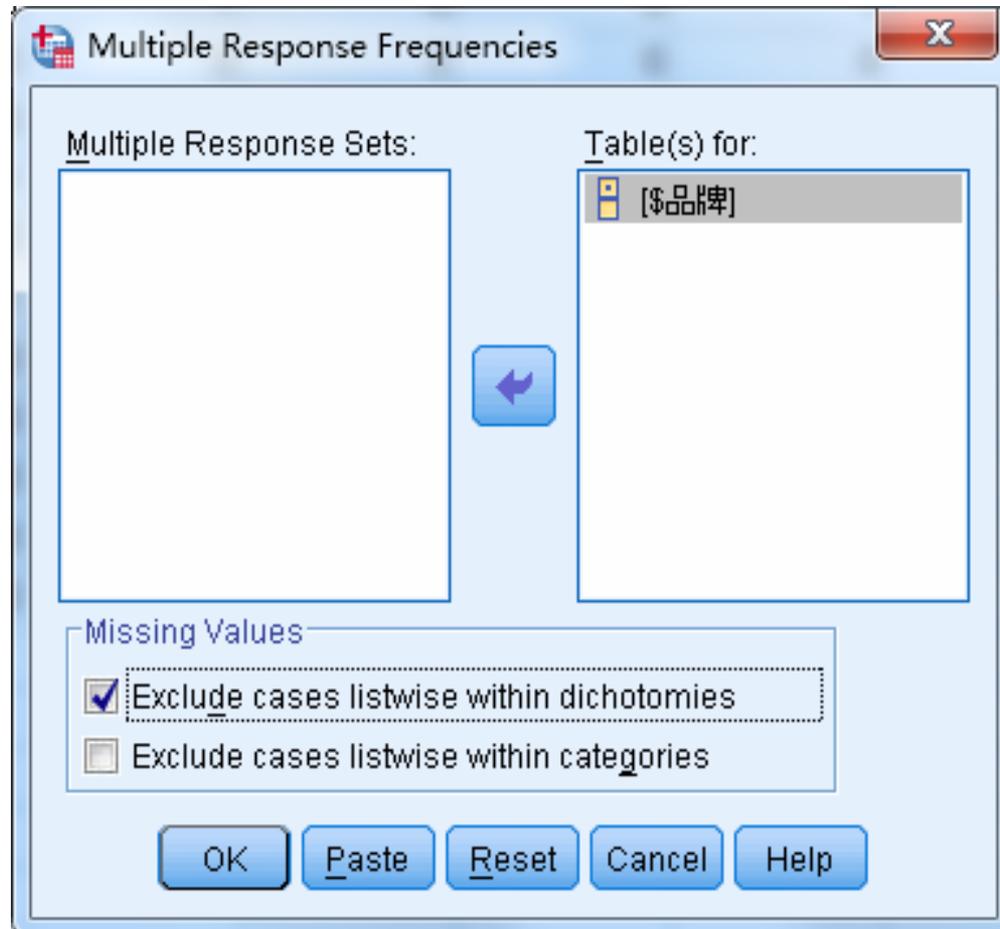
# 实例操作



## Step06: 完成操作

最后，单击【OK(确定)】按钮，操作完成。此时，软件输出结果出现在结果浏览窗口中。

# 实例操作



# 实例结果及分析



## (1) 案例的摘要

执行完上面操作后，在SPSS结果报告中首先给出的是案例的摘要，见下表所示。多重响应分析的样本数据为20个，缺失数据个数为0，整个多重响应分析是基于所有数据来进行的。



# 实例结果及分析

	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
\$品牌 <sup>a</sup>	20	100.0%	0	0.0%	20	100.0%

a. Dichotomy group tabulated at value 1.

# 实例结果及分析

CONCEPT  
STRATE

- (2) 多重响应频数分析的结果

在SPSS结果报告中给出各个品牌的频数与百分比。当前使用三星手机的人数为5个，百分比为25%，其市场占有比例与摩托罗拉的比例一致，当前使用苹果手机的人数为2个，其百分比为10%，这说明虽然苹果手机比较贵，但现在是比较时髦的。



# 实例结果及分析

		Responses		Percent of Cases
		N	Percent	
\$品牌 <sup>a</sup>	sumsung	5	25.0%	25.0%
	MOTO	5	25.0%	25.0%
	NOKIA	2	10.0%	10.0%
	LG	4	20.0%	20.0%
	Apple	2	10.0%	10.0%
	skyworth	1	5.0%	5.0%
	other	1	5.0%	5.0%
Total		20	100.0%	100.0%

a. Dichotomy group tabulated at value 1.

## 10.3.4 实例进阶分析： 多重响应交叉分析

CONCEPT  
STRATE

### 1. 实例操作

**Step01:** 打开多重响应交叉分析对话框

选择菜单栏中的【Analyze(分析)】→【Multiple Responses(多元响应)】→【Crosstabs(交叉表)】命令，进入【multiple Response Crosstabs(多元响应交叉表)】对话框。

# 实例操作

CONCEPT  
STRATE



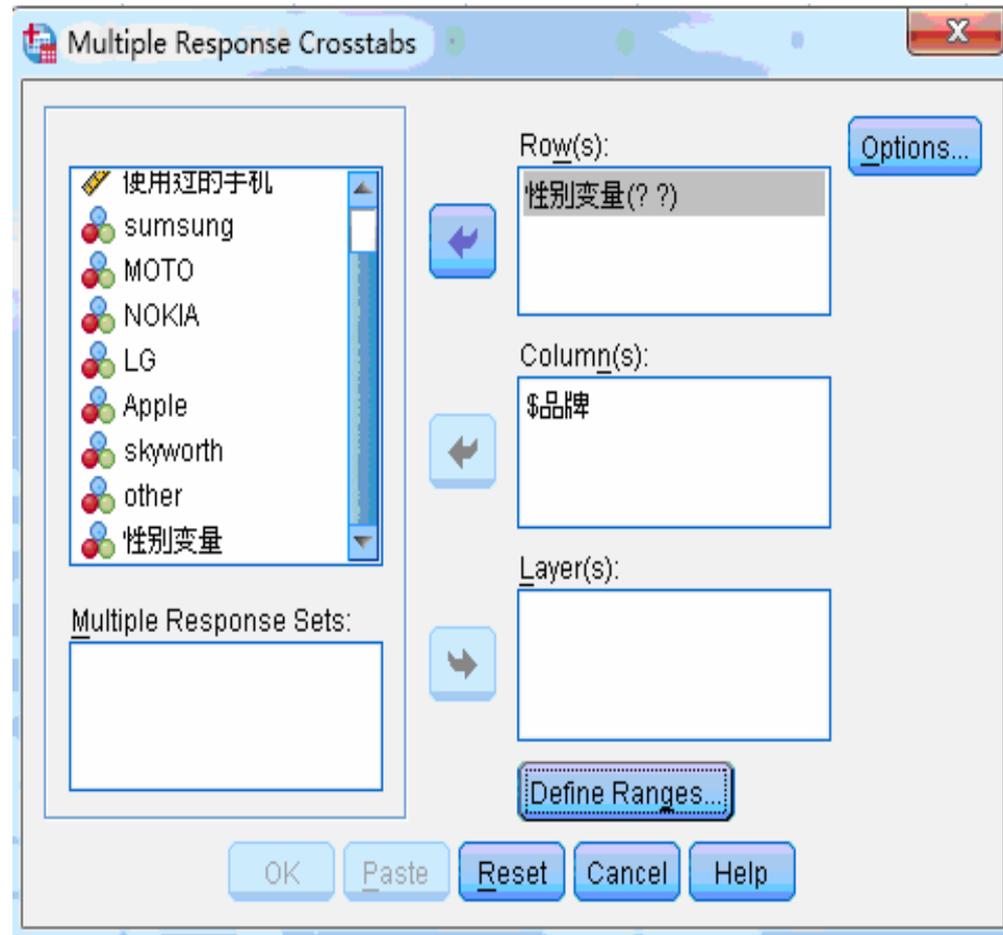
# 实例操作

CONCEPT  
STRATE

**Step02:** 将【Multiple Responses Sets (多元响应集)】中的多重响应集“品牌”选入【Column(s)(列)】列表框；将列表框中的性别变量选入【Row(s)(行)】列表框，以性别变量作为行变量, 并单击【Define Ranges (定义范围)】按钮进入定义行变量的取值范围设置。

# 实例操作

CONCEPT  
STRATE

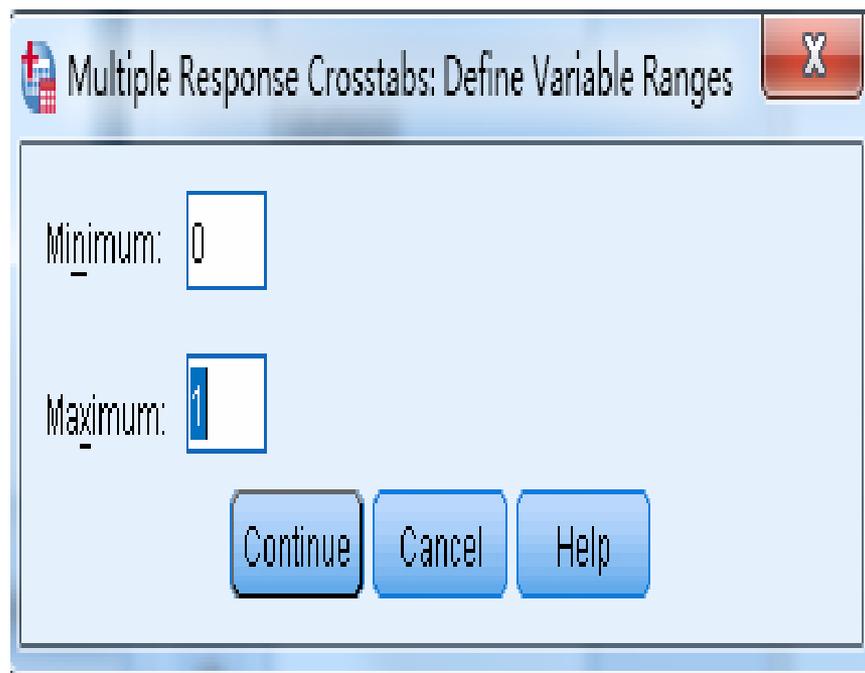


# 实例操作

CONCEPT  
STRATE

**Step03:** 设置范围的最小值为0，最大值为1，单击【Continue (继续)】按钮回到【Multiple Responses Crosstabs (多元响应交叉表表)】对话框。

# 实例操作



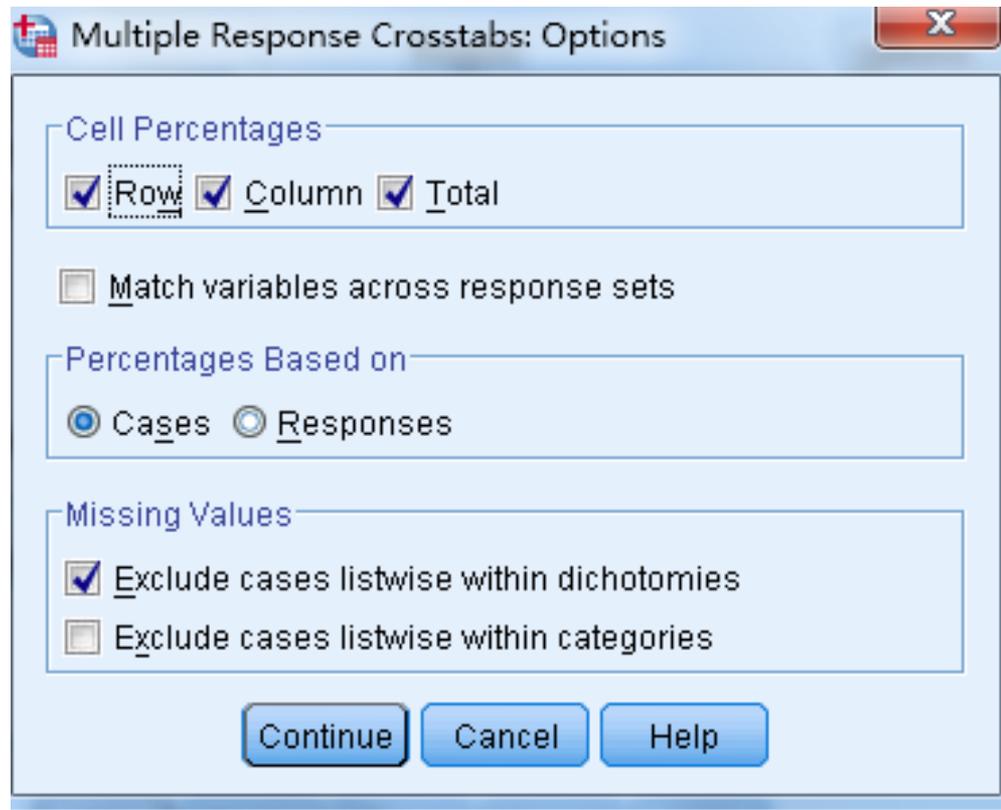
# 实例操作

CONCEPT  
STRATE

**Step04:** 单击【Options(选项)】按钮进入多重响应交叉分析的选项设置对话框。选项包括行的、列的、总的单元百分比选择基于个案的百分比，选择缺失值的处理是基于多重二分法的变量的，然后单击【Continue(继续)】按钮回到【Multiple Responses Crosstabs(多元响应交叉表)】对话框。

# 实例操作

CONCEPT  
STRATE



# 实例操作

CONCEPT  
STRATE

## Step05: 完成操作

最后，单击【OK(确定)】按钮，操作完成。此时，软件输出结果出现在结果浏览窗口中。

# 实例操作



# 实例结果及分析

CONCEPT  
STRATE

## (1) 案例的摘要

执行完上面操作后，在SPSS结果报告中首先给出的是案例的摘要。多重响应分析的样本数据为20个，缺失数据个数为0，整个多重响应分析是基于所有数据来进行的。

# 实例结果及分析



	Cases					
	Valid		Missing		Total	
	N	Percent	N	Percent	N	Percent
性别变量*\$品牌	20	100.0%	0	0.0%	20	100.0%

# 实例结果及分析

CONCEPT  
STRATE

## (2) 多重响应交叉分析的结果

在SPSS结果报告中给出不同性别下各个品牌的频数与百分比。女性中，当前使用三星手机的人数为3个，占女性的33.3%，即针对女性这个群体而言，大约1/3的使用三星手机，三星手机是女性比较喜欢一个手机品牌；针对三星品牌而言，有40% 的是女性使用者，60%的男性使用者，即女性对三星手机喜爱程度远远高于男性对三星手机喜爱程度高。

在所有的品牌中，对女性群体而言，最喜欢的手机是三星和摩托罗拉，大约1/3的使用三星手机，1/3的使用摩托罗拉。不同性别群体在手机的 brand 选择上差异比较大，男性对诺基亚手机非常喜欢，而女性比较喜欢三星。



# 实例结果及分析

			\$品牌 <sup>a</sup>						Total	
			samsung	MOTO	NOKI A	LG	Apple	skyworth		other
性别 变量	0	Count	3	3	0	2	1	0	0	9
		% within 性别变量	33.3%	33.3%	0.0%	22.2%	11.1%	0.0%	0.0%	
		% within \$品牌	60.0%	60.0%	0.0%	50.0%	50.0%	0.0%	0.0%	
		% of Total	15.0%	15.0%	0.0%	10.0%	5.0%	0.0%	0.0%	45.0%
	1	Count	2	2	2	2	1	1	1	11
		% within 性别变量	18.2%	18.2%	18.2%	18.2%	9.1%	9.1%	9.1%	
		% within \$品牌	40.0%	40.0%	100.0 %	50.0%	50.0%	100.0%	100.0%	
		% of Total	10.0%	10.0%	10.0%	10.0%	5.0%	5.0%	5.0%	55.0%
Total	Count	5	5	2	4	2	1	1	20	
	% of Total	25.0%	25.0%	10.0%	20.0%	10.0%	5.0%	5.0%	100.0%	

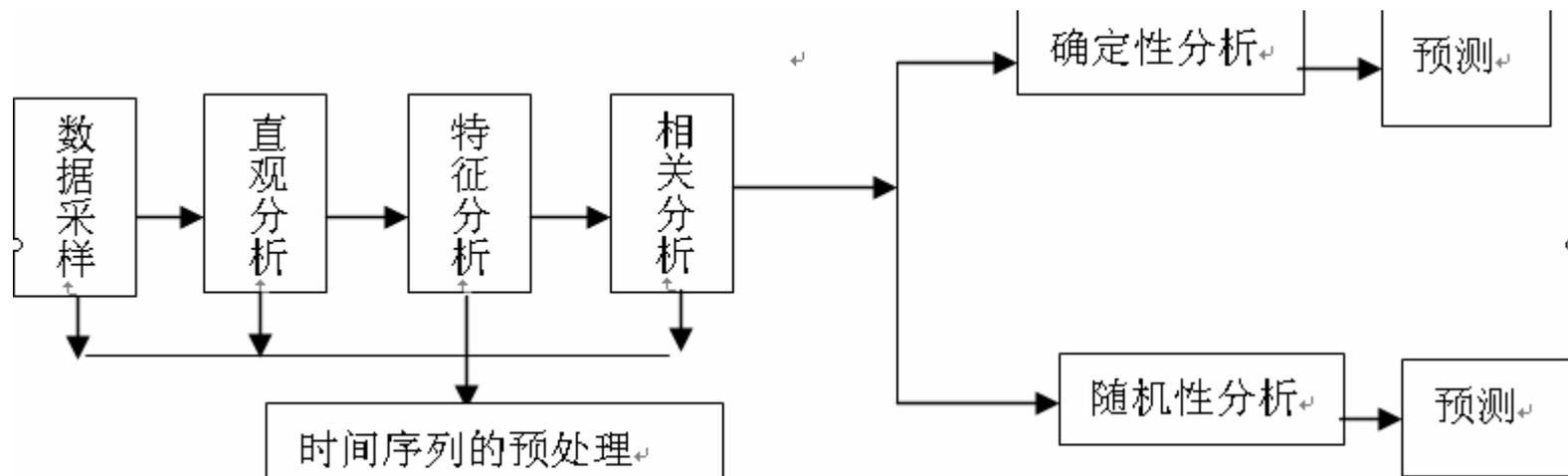
Percentages and totals are based on respondents.

a. Dichotomy group tabulated at value 1.



# 第11章 SPSS在时间序列 预测中的应用

时间序列分析(Time Series Analyze)是概率统计学科中应用性较强的一个分支,在金融经济、气象水文、信号处理、机械振动等众多领域有从所采用的数学工具和理论,时间序列分析分为时域分析和谱分析两大类分析方法  
预测的流程通常可以用下图来描述



# 11.1 时间序列的预处理

CONCEPT  
RATE

## 11.1.1 预处理的基本原理

### 1. 使用目的

通过预处理，一方面能够使序列的随“时间”变化的、“动态”的特征体现得更加明显，利用模型的选择；另一方面也使得数据满足与模型的要求。

### 2. 基本原理

#### (1) 数据采样

采样的方法通常有直接采样、累计采样等。

#### (2) 直观分析

时间序列的直观分析通常包括离群点的检验和处理、缺损值的补足、指标计算范围是否统一等一些比较简单的，可以采用比较简单手段处理的分析。

### (3)特征分析

所谓特征分析就是在对数据序列进行建模之前，通过从时间序列中计算出一些有代表性的特征参数，用以浓缩、简化数据信息，以利数据的深入处理，或通过概率直方图和正态性检验分析数据的统计特性。通常使用的特征参数有样本均值、样本方差、标准偏度系数、标准峰度系数等。

### (4)相关分析

所谓相关分析就是测定时间序列数据内部的相关程度，给出相应的定量度量，并分析其特征及变化规律。

理论上，自相关系数序列与时间序列具有相同的变化周期。所以，根据样本自相关系数序列随增长而衰减的特点或其周期变化的特点判断序列是否具有平稳性，识别序列的模型，从而建立相应的模型。

### 3. 其他注意事项

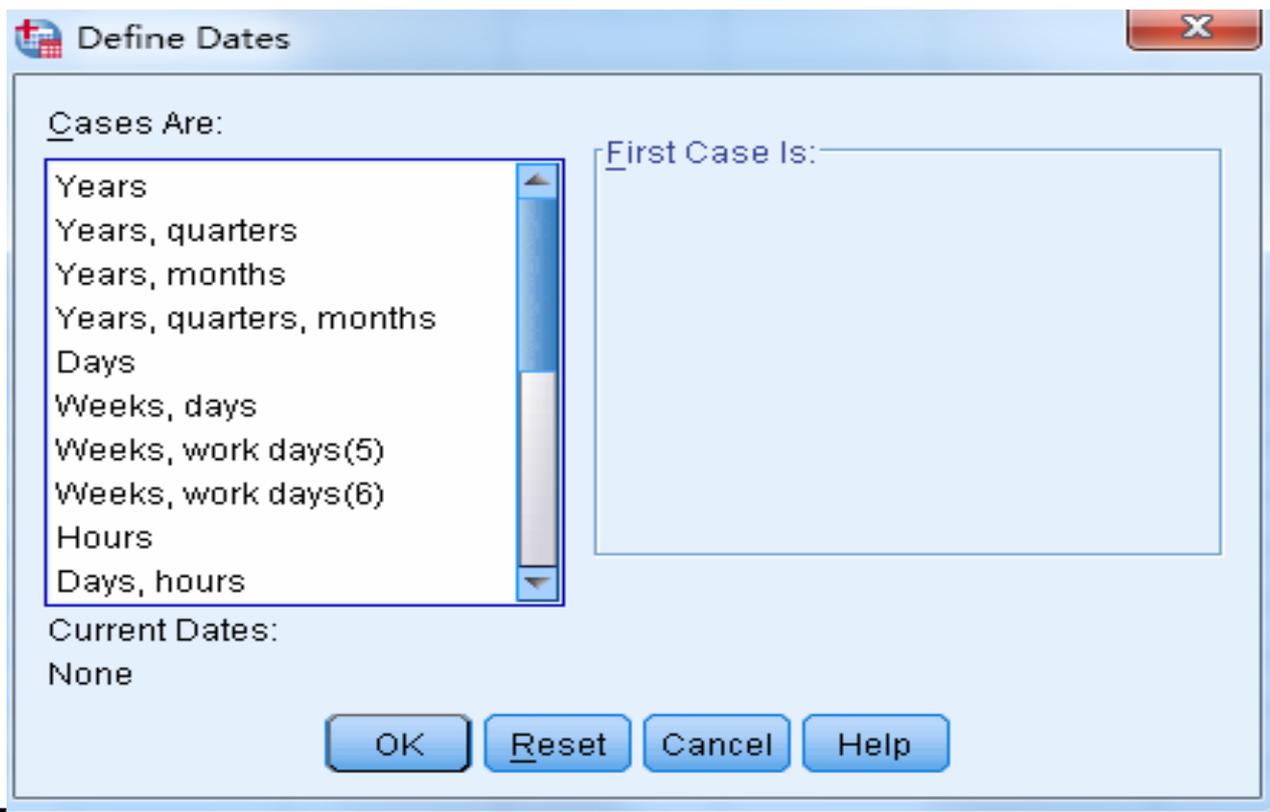
进行时间序列预处理的时候，常常需要对数据一些变换，例如，取对数，做一阶差分，做季节差分等。

## 11.1.2 时间序列预处理的SPSS操作详解

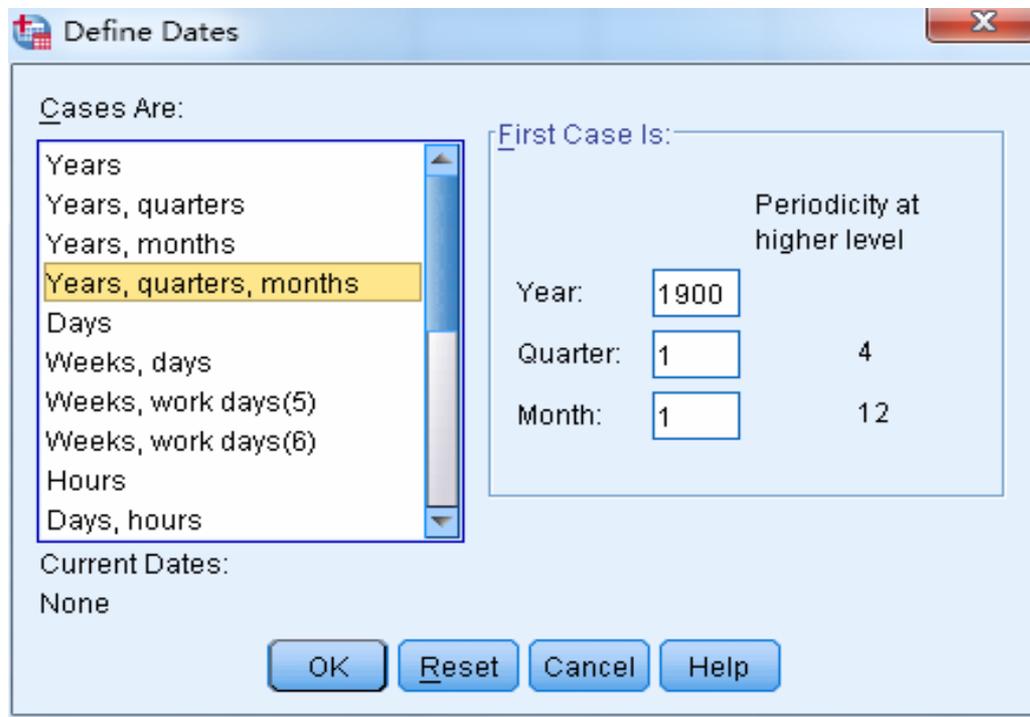
ONCEPT  
TRATE

Step01:数据准备

选择菜单栏中的【Data(数据)】→【Define Dates(定义日期)】命令，弹出【Define Dates(定义日期)】对话框。



如果选择月度数据或季度数据，将会出现【Periodicity at higher level(更高级别的周期)】。在其下方将显示数据的最大周期长度，月度数据默认周期长度为12，季度数据默认周期长度为4。



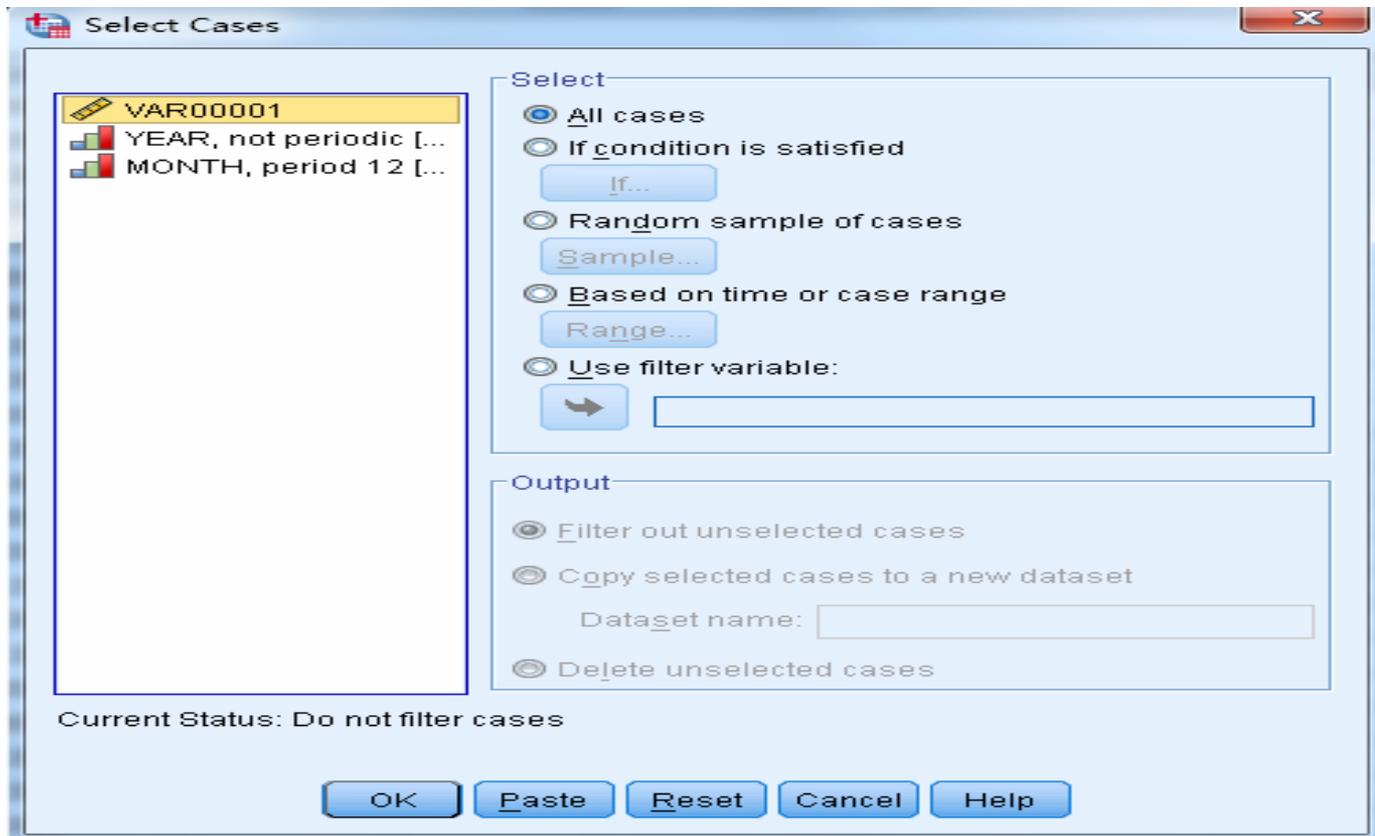


单击【OK(确认)】按钮，此时完成时间的定义，SPSS将在当前数据编辑窗口中自动生成标志时间的变量。

VAR00001	YEAR_	MONTH_	DATE_
2978.20	2000	1	JAN 2000
2822.10	2000	2	FEB 2000
2626.60	2000	3	MAR 2000
2571.50	2000	4	APR 2000
2636.90	2000	5	MAY 2000
2645.20	2000	6	JUN 2000
2596.90	2000	7	JUL 2000
2636.30	2000	8	AUG 2000
2854.30	2000	9	SEP 2000
3029.30	2000	10	OCT 2000
3107.80	2000	11	NOV 2000
3680.10	2000	12	DEC 2000
3127.20	2001	1	JAN 2001
3001.80	2001	2	FEB 2001
2876.10	2001	3	MAR 2001
2820.90	2001	4	APR 2001
2929.60	2001	5	MAY 2001
2908.70	2001	6	JUN 2001
2851.40	2001	7	JUL 2001
2889.40	2001	8	AUG 2001
3136.90	2001	9	SEP 2001
3347.30	2001	10	OCT 2001
3421.70	2001	11	NOV 2001

# Step02: 数据采样

选择菜单栏中的【Data(数据)】→【Select Cases(选择个案)】命令，弹出【Select Cases(选择个案)】对话框。

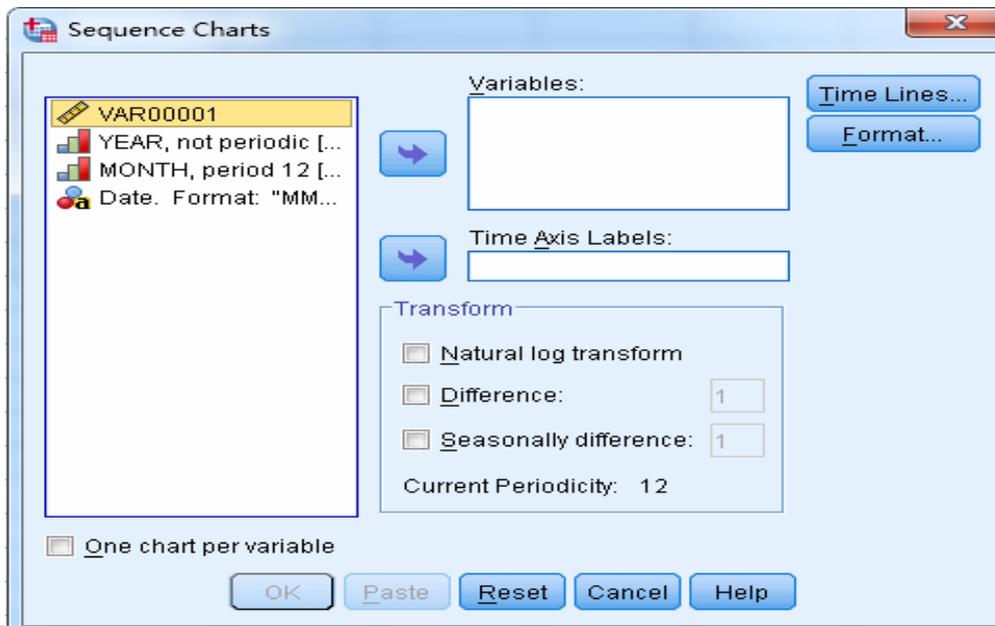


## Step03: 直观分析

CONCEPT  
STRATE

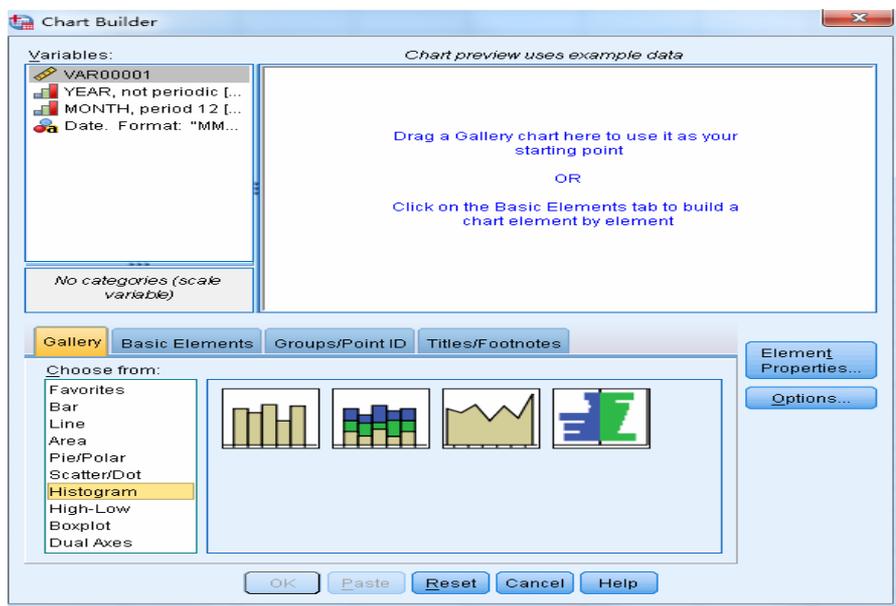
当数据准备好，为认识数据的变化规律，判断数据是否存在离群点和缺损值，最直接的观察方法是绘制序列的图像。

选择菜单栏中的【Data(数据)】→【Forecasting(预测)】→【Sequence Charts(序列图)】命令，弹出【Sequence Charts(序列图)】对话框。



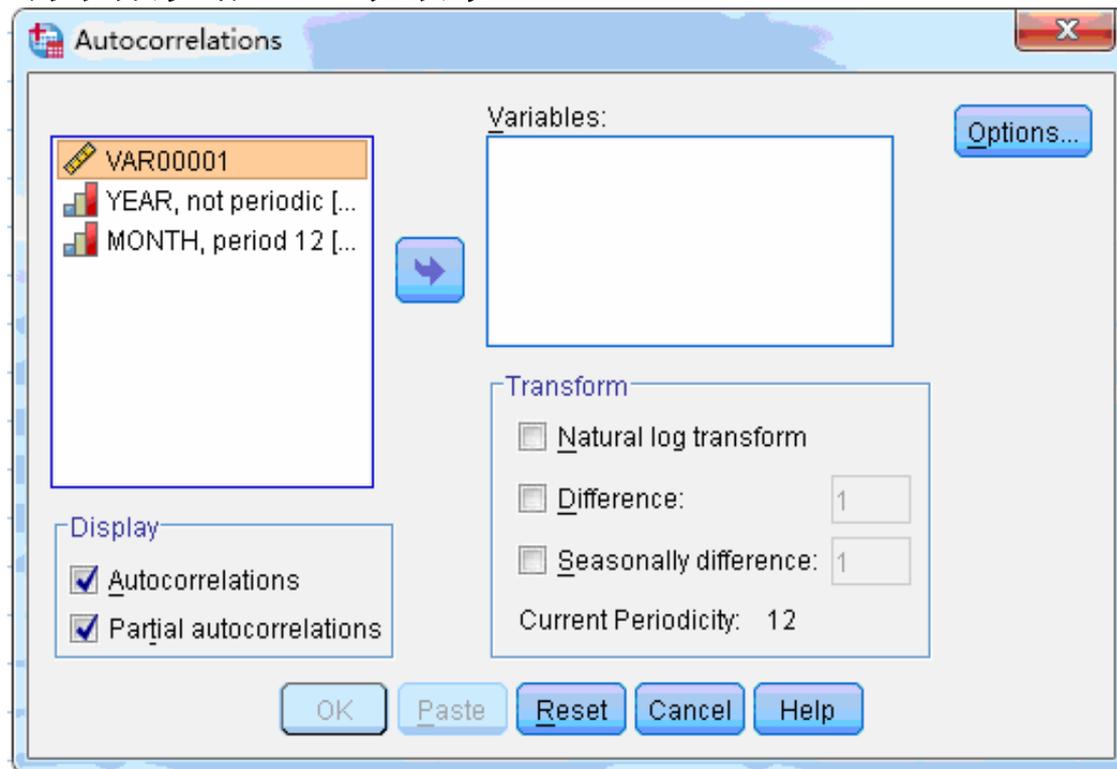
# Step04: 特征分析

选择菜单栏中的【Data(数据)】→【Graphs(图形)】→【Chart Builder(图表构建程序)】命令，弹出【Chart Builder(图表构建程序)】对话框。在【Gallery(库)】选项卡中选择【Histogram(直方图)】，并将直方图形拖入【Chart preview uses example data(图预览使用实例数据)】下方的白色区域，然后将所需要画直方图的变量拖入X轴，单击【OK(确认)】按钮就画出直方图了，图中将显示该变量的均值、方差、样本容量。



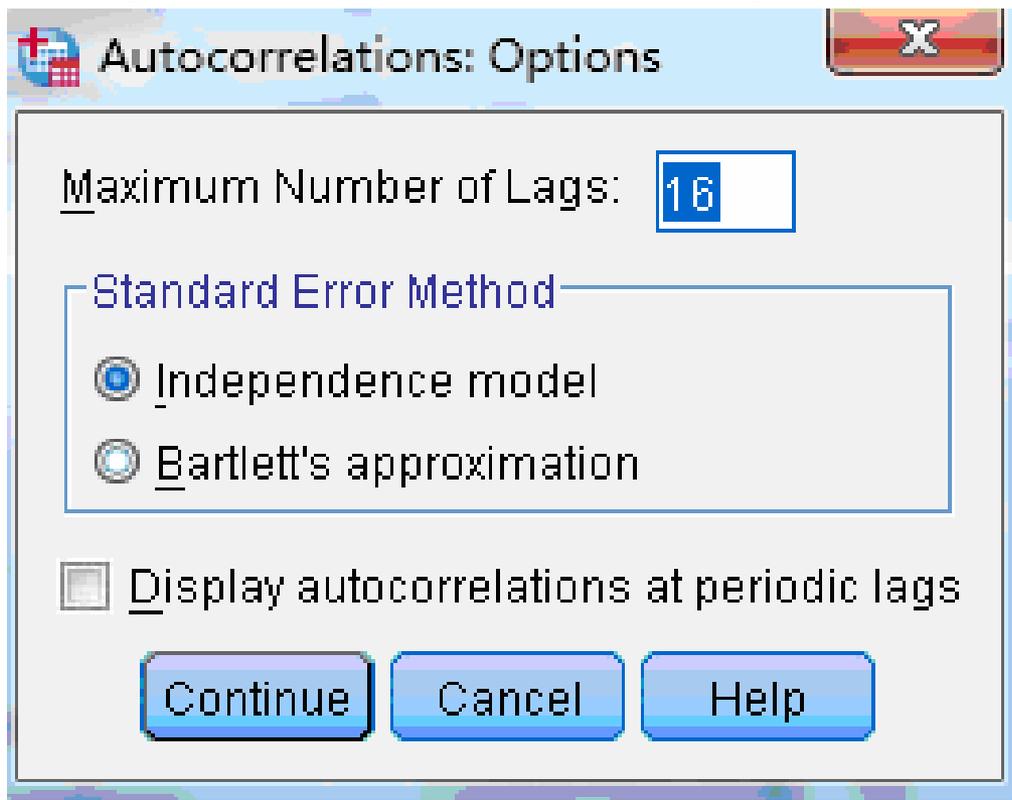
# Step05 : 相关分析

选择菜单栏中的【Analyze(分析)】→【Forecasting(预测)】→【Autocorrelations(自相关)】命令，弹出【Autocorrelations(自相关)】对话框。



在左侧的候选变量列表框中选择一个变量，将其移入【Variables (变量)】列表框中。

单击【Options】按钮，弹出【Options (选项)】对话框。



# 11.1.3 实例图文分析：社会商品零售总额的预处理

CONCEPT  
TRATE

## 1. 实例内容

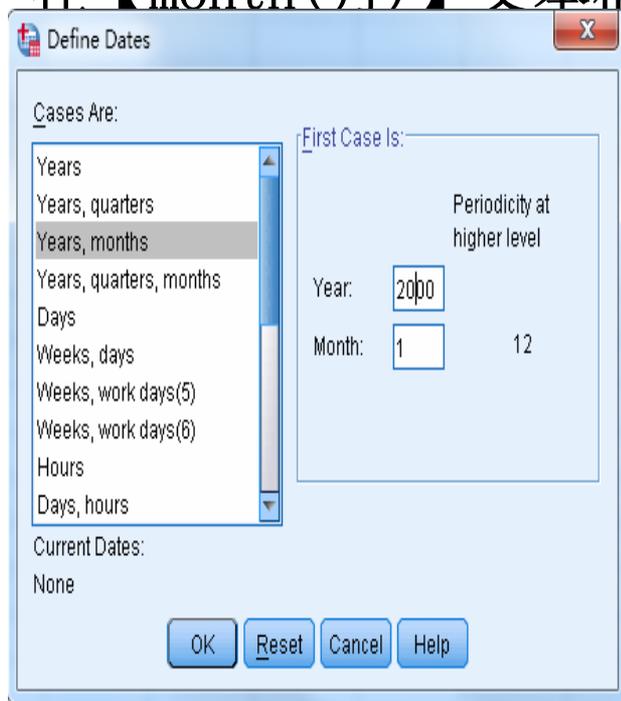
为了分析社会商品零售总额的变动趋势，收集了我国2000年1月到2010年5月社会商品零售总额的数据，现在对数据进行时间序列的预处理。

表 11-1 中国 2000 年 1 月到 2010 年 5 月社会商品零售总额的数据（部分）

2000M01	2978.2	2003M01	3907.4	2006M01	6641.6	2009M01	10756.6
2000M02	2822.1	2003M02	3706.4	2006M02	6001.9	2009M02	9323.8
2000M03	2626.6	2003M03	3494.8	2006M03	5796.7	2009M03	9317.6
2000M04	2571.5	2003M04	3406.9	2006M04	5774.6	2009M04	9343.2
2000M05	2636.9	2003M05	3463.3	2006M05	6175.6	2009M05	10028.4

## 2 实例操作

Step01: 数据准备输入社会商品零售总额的数据，然后选择菜单栏中的【Data(数据)】→【Define Dates(定义日期)】命令，弹出【Define Dates(定义日期)】对话框，选择【Years, month(年,月)】选项,并在【First Case is】选项组的【Year(年)】文本框中输入“2000”，在【month(月)】文本框中输入“1”。



## Step02: 标志时间的变量出现

单击【OK(确认)】按钮，此时完成时间的定义，SPSS将在当前数据编辑窗口中自动生成标志时间的变量，同时在输出窗口中将会出现一个简明的日志，说明时间标志变量及其格式和包含的周期等。

VAR00001	YEAR_	MONTH_	DATE_
2978.20	2000	1	JAN 2000
2822.10	2000	2	FEB 2000
2626.60	2000	3	MAR 2000
2571.50	2000	4	APR 2000
2636.90	2000	5	MAY 2000
2645.20	2000	6	JUN 2000
2596.90	2000	7	JUL 2000
2636.30	2000	8	AUG 2000
2854.30	2000	9	SEP 2000
3029.30	2000	10	OCT 2000
3107.80	2000	11	NOV 2000
3680.10	2000	12	DEC 2000
3127.20	2001	1	JAN 2001
3001.80	2001	2	FEB 2001
2876.10	2001	3	MAR 2001
2820.90	2001	4	APR 2001
2929.60	2001	5	MAY 2001
2908.70	2001	6	JUN 2001

### Step03 : 数据采样

选择菜单栏中的【Data(数据)】→【Select Cases(选择个案)】命令，弹出【Select Cases (选择个案)】对话框，点选【Based on time or case range(基于时间或个案全距)】单选钮，并单击【range(范围)】按钮，此时会出现新的对话框，在【First case(第一个个案)】选项组的【Year(年)】文本框中输入“2000”，在【month(月)】文本框中输入“1”，在【First case(最后个个案)】选项组的【Year(年)】文本框中输入“2009”，在【month(月)】文本框中输入“12”。单击【Continue(继续)】按钮，然后单击【Select Cases(选择个案)】对话框中的【OK(确认)】按钮，此时在输出窗口中将会出现一个简明的日志，说明此时只对2000年1月到2009年12月的数据做分析与建模。



Select Cases

VAR00001  
YEAR, not periodic [...]  
MONTH, period 12 [...]

Select

- All cases
- If condition is satisfied  
If...
- Random sample of cases  
Sample...
- Based on time or case range  
Range...
- Delete unselected cases

Select Cases: Range

	First Case	Last Case
Year:	2000	2009
Month:	1	12

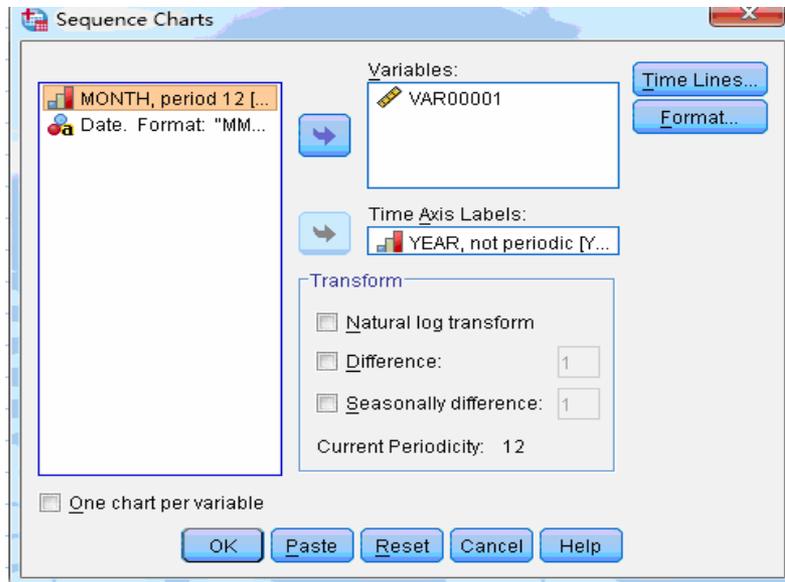
Continue Cancel Help

Current Status: Do not filter cases

OK Paste Reset Cancel Help

# Step04 : 直观分析

选择菜单栏中的【Data(数据)】→【Forecasting(预测)】→【Sequence Charts(序列图)】命令，弹出【Sequence Charts(序列图)】对话框，在该对话框左侧的候选变量列表框中选择【VAR00001】选项，将其移入【Variables(变量)】列表框中，选择【Year, not periodic】将其移入【Time Axis Labels(时间轴标签)】列表框，单击【OK(确认)】按钮即可生成线图。



# Step05 : 特征分析

选择菜单栏中的【Data(数据)】→【Graphs(图形)】→【Chart Builder(图表构建程序)】命令，弹出【Chart Builder(图表构建程序)】对话框。在【Gallery(库)】选项卡中选择【Histogram(直方图)】选项，并将直方图图形拖入【Chart preview uses example data(图预览使用实例数据)】下方的白色区域，然后将【VAR00001】拖入X轴，单击【OK(确认)】按钮即可生成直方图。

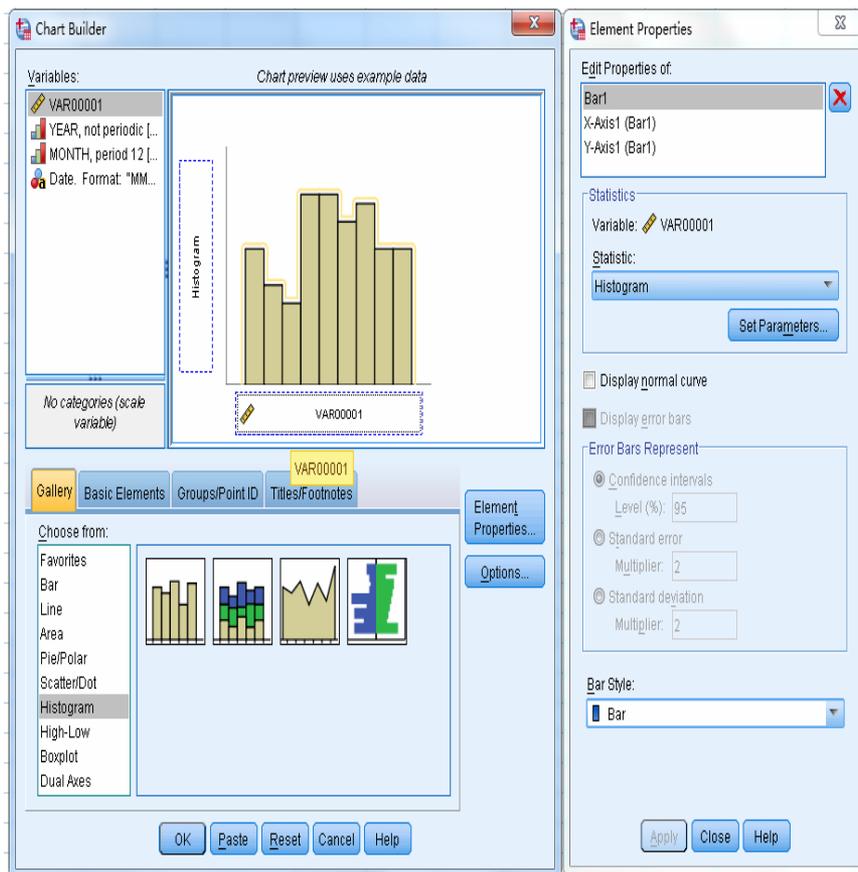
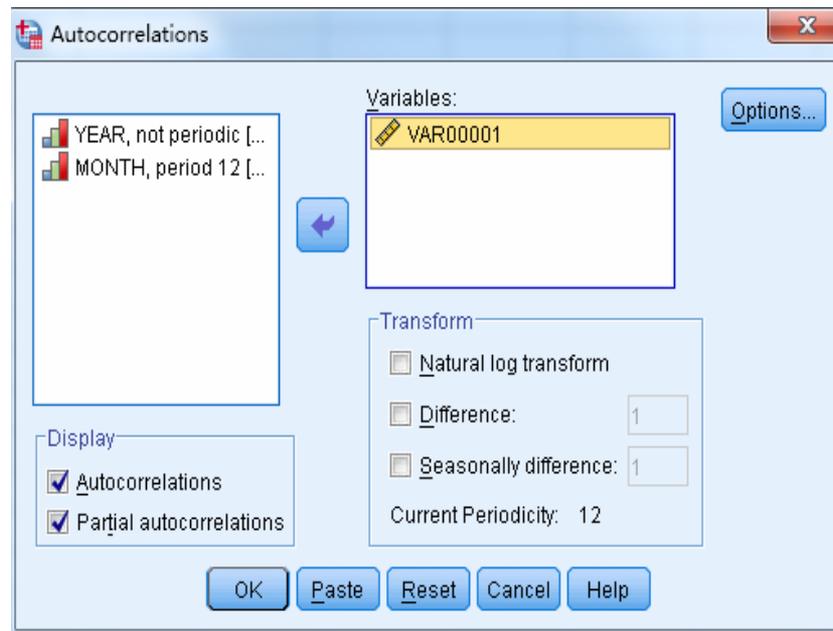


图11-13

# Step06 : 相关分析

CONCEPT  
STRATE

选择菜单栏中的【Analyze(分析)】→【Forecasting(预测)】→【Autocorrelations(自相关)】命令，弹出【Autocorrelations(自相关)】对话框。将【VAR00001】移入【Variables(变量)】列表框中，在【Display(显示)】选项组中勾选所以复选框，即展示自相关函数图、又偏相关函数图。单击【OK(确认)】按钮即可绘制自相关函数图和偏相关函数图。

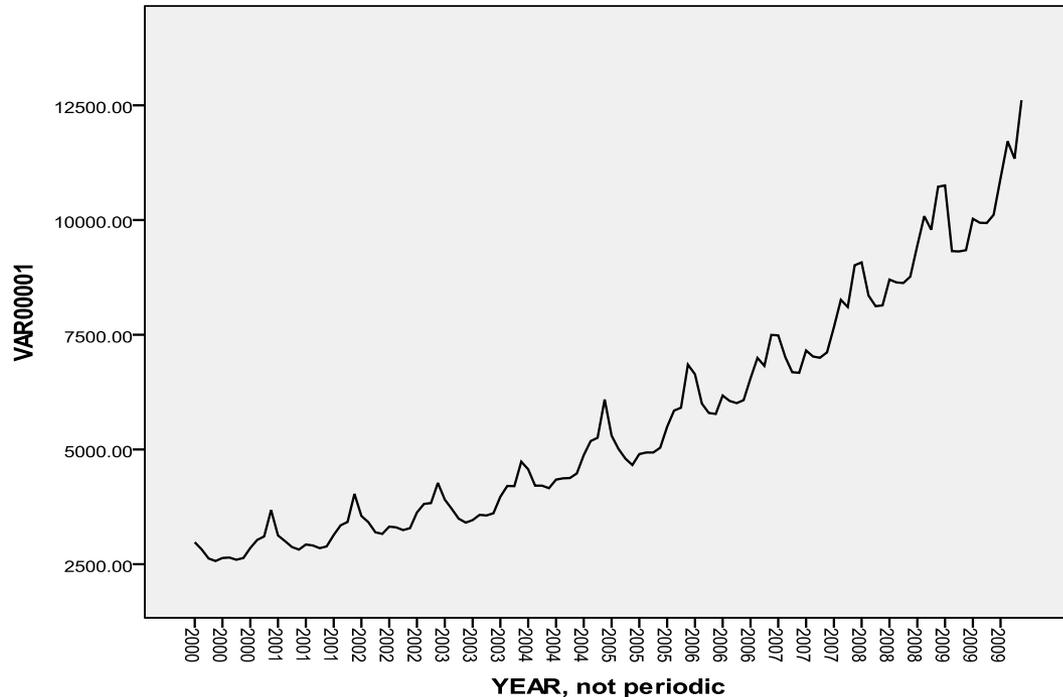


# 3 实例结果及分析



## (1) 直观分析的输出结果

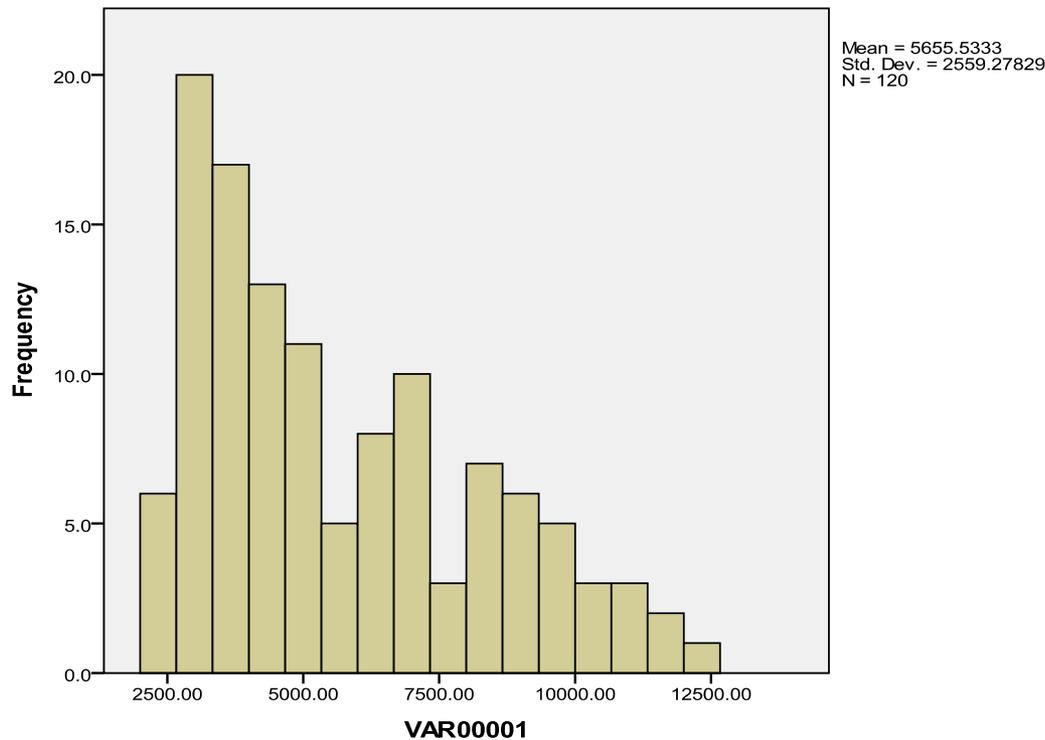
我国2000年1月到2009年12月社会商品零售总额的线图，从图上可以看出该序列有明显的趋势性或周期性这说明该序列，而且无离群点和缺失值。



## (2) 特征分析结果

CONCEPT  
STRATE

我国2000年1月到2009年12月社会商品零售总额的直方图，如图11-16所示。从图上可以看出该序列的样本均值为5655.5333，样本标准差为2559.27829，样本容量为120个。



## (3) 相关分析结果

CONCEPT  
TRATE

### (1) 样本自相关系数的值

在SPSS中给出了不同滞后期（Lag列）的样本自相关系数的值（Autocorrelation列），样本自相关系数的标准误差（Std Error列），以及Box-ljung Statistic的值、自由度（d f列）和相伴概率（Sig）。通过标准误差值以及Box-ljung Statistic的相伴概率都可以说该时间序列不是白噪声，是具有自相关性的时间序列，可以建立ARIMA等模型。Box-ljung Statistic的相伴概率是在近似认为Box-ljung Statistic服从卡方分布得到。



表 11-2 样本自相关系数的数据表

Series: VAR00001

Lag	Autocorrelatio n	Std. Error <sup>a</sup>	Box-Ljung Statistic <sup>b</sup>		
			Value	df	Sig.
1	.953	.090	111.695	1	.000
2	.916	.090	215.728	2	.000
3	.875	.089	311.599	3	.000
4	.841	.089	400.820	4	.000
5	.815	.089	485.316	5	.000
6	.789	.088	565.339	6	.000
7	.769	.088	641.894	7	.000
8	.747	.087	714.855	8	.000
9	.735	.087	786.041	9	.000
10	.722	.087	855.413	10	.000
11	.711	.086	923.394	11	.000
12	.689	.086	987.773	12	.000
13	.648	.085	1045.201	13	.000
14	.613	.085	1097.075	14	.000
15	.576	.085	1143.343	15	.000
16	.544	.084	1184.949	16	.000

a. The underlying process assumed is independence (white noise).

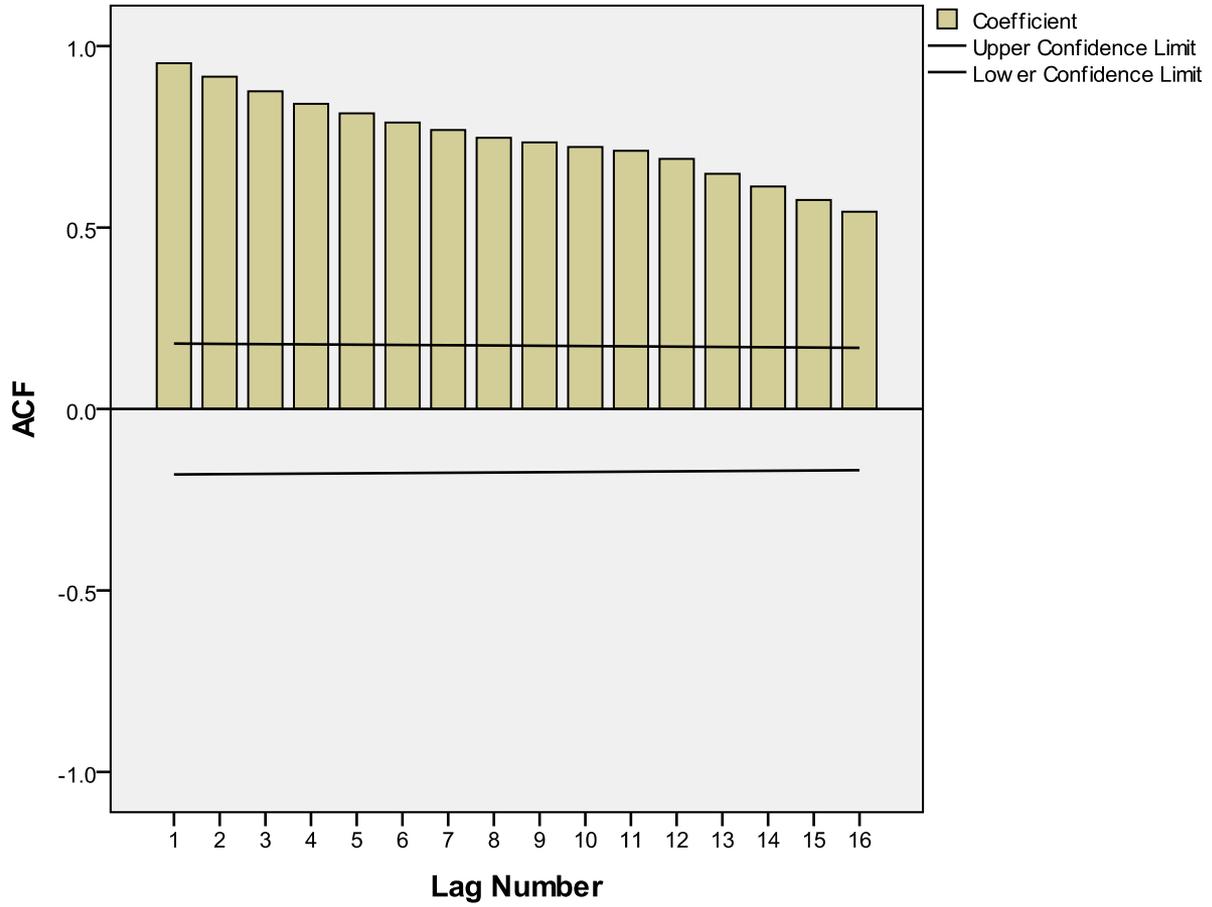
b. Based on the asymptotic chi-square approximation.

## (2) 样本自相关系数的图形

在SPSS中画出了样本自相关系数图。图中的横轴为滞后期（Lag Number），纵轴为样本自相关系数（ACF）。图中用条形形状来表示样本自相关系数，并画出了95%的置信上下限的线条。从下图可以看出该时间序列的自相关系数并不呈负指数收敛到零，其衰减速度比较慢，不是平稳时间序列。



### VAR00001



### (3) 样本偏相关系数的值

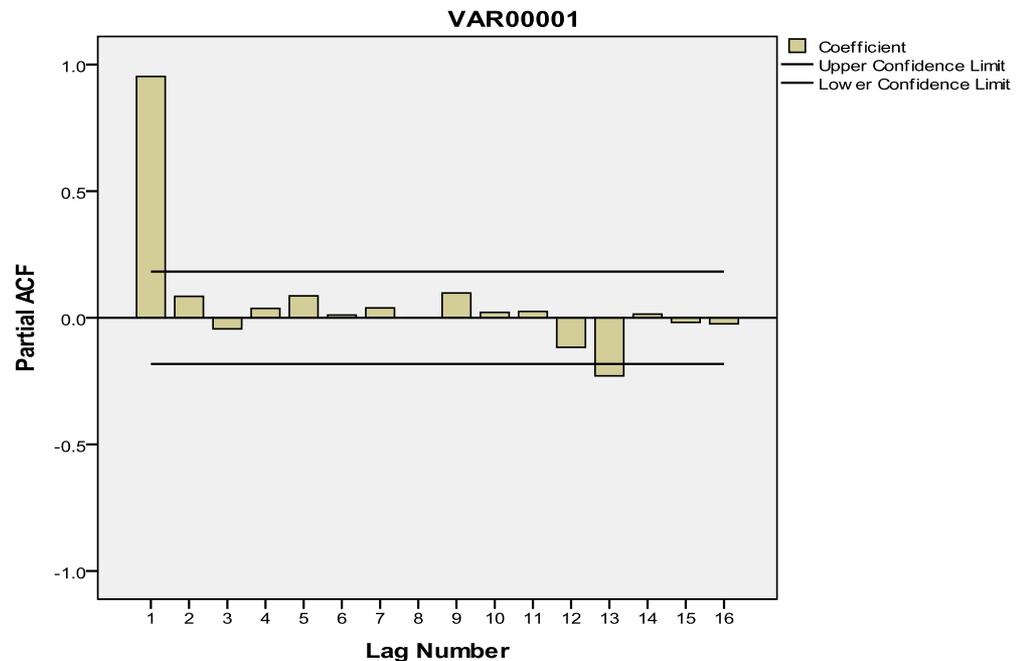
在SPSS中给出了不同滞后阶（Lag列）的样本偏相关系数的值（Partial Autocorrelations 列），样本偏相关系数的标准误差（Std Error列）。从表10-3样本偏相关系数的数据表可以看出该时间序列不是白噪声。

表 11-3 样本偏相关系数的数据表

Series:VAR00001		
Lag	Partial Autocorrelation	Std. Error
1	.953	.091
2	.085	.091
3	-.044	.091
4	.037	.091
5	.086	.091
6	.011	.091
7	.039	.091
8	.000	.091
9	.098	.091
10	.021	.091
11	.025	.091
12	-.116	.091
13	-.229	.091
14	.014	.091
15	-.018	.091
16	-.024	.091

#### (4) 样本偏相关系数的图形

图中的横轴为滞后期 (Lag Number)，纵轴为样本偏相关系数 (Partial ACF)。图中用条形形状来表示样本偏相关系数，并画出了95%的置信上下限的线条。从下图可以看出该时间序列的偏相关系数在一阶滞后期、12阶滞后期比较大，说明该时间序列具有周期性，不是平稳时间序列。



# 11.2 时间序列的确定性分析

CONCEPT  
STRATE

## 11.2.1 确定性分析的基本原理

### 1、使用目的

传统时间序列分析认为长期趋势变动、季节性变动、周期变动是依一定的规则而变化的，不规则变动因素在综合中可以消除。基于这种认识，形成了确定性时间序列分析。

通过确定性时间序列分析，一方面能够使序列的长期趋势变动特征、季节效应、周期变动体现得更加明显；另一方面能确立模型，从而成功捕捉数据的随“时间”变化的、“动态”的、“整体”的统计规律。因此，对时间序列进行确定分析，从而建立模型是非常必要的。

## 2、基本原理

### (1) 指数平滑法

指数平滑法有助于预测存在趋势和（或）季节的序列。指数平滑法分为两步来建模，第一步确定模型类型，确定模型是否需要包含趋势、季节性，创建最适当的指数平滑模型，第二步选择最适合选定模型的参数。

指数平滑模法一般分为无季节性模型、季节性模型。无季节性模型包括简单指数平滑法、布朗单参数线性指数平滑法等，季节性模型包括温特线性和季节性指数平滑法。

指数平滑法，又称指数加权平均法，实际是加权的移动平均法，它是选取各时期权重数值为递减指数数列的均值方法。

## (2) 季节分解法

季节分解的一般步骤如下：

第一步，确定季节分解的模型；

第二步，计算每一周期点（每季度，每月等等）的季节指数（乘法模型）或季节变差（加法模型）；

第三步，用时间序列的每一个观测值除以适当的季节指数（或减去季节变差），消除季节影响；

第三步，对消除了季节影响的时间序列进行适当的趋势性分析；

第四步，剔除趋势项，计算周期变动；

第五步，剔除周期变动，得到不规则变动因素；

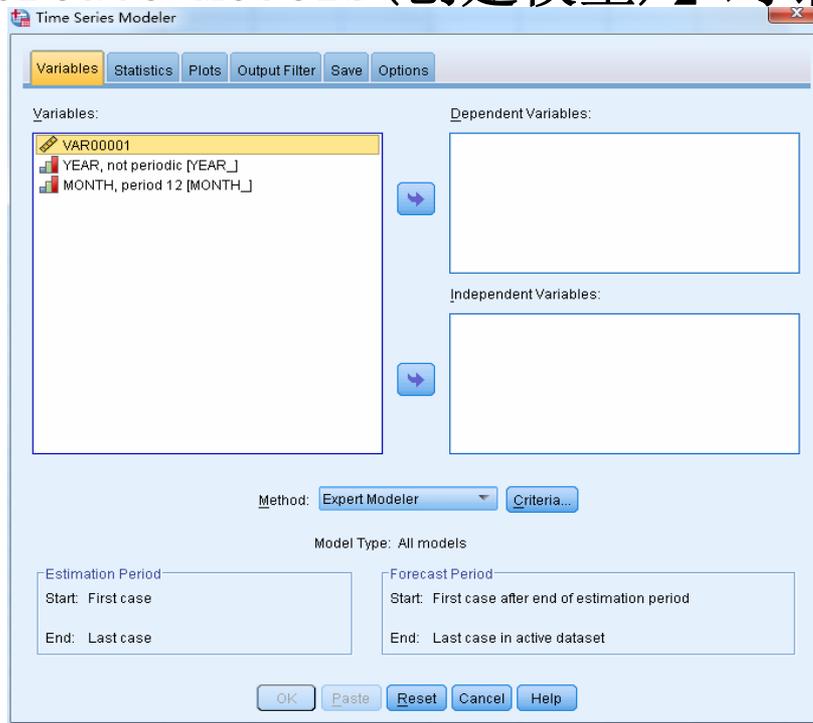
第六步，用预测值乘以季节指数（或加上季节变差），乘以周期变动，计算出最终的带季节影响的预测值。

## 11.2.2 指数平滑法的SPSS操作详解

CONCEPT  
STRATE

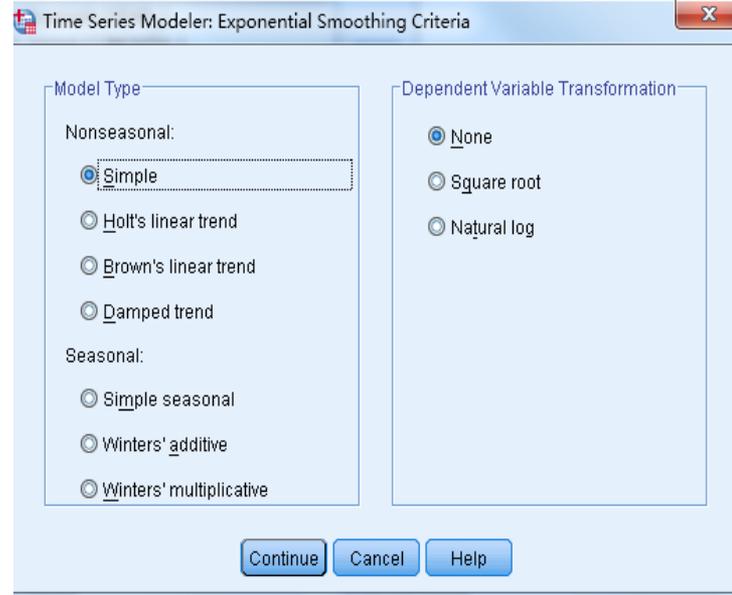
Step01 : 打开【Create Models(创建模型)】对话框

当时间序列的数据已经准备好以后，选择菜单栏中的【Analyze(分析)】→【Forecasting(预测)】→【Create Models(创建模型)】命令，弹出【Create Models(创建模型)】对话框。



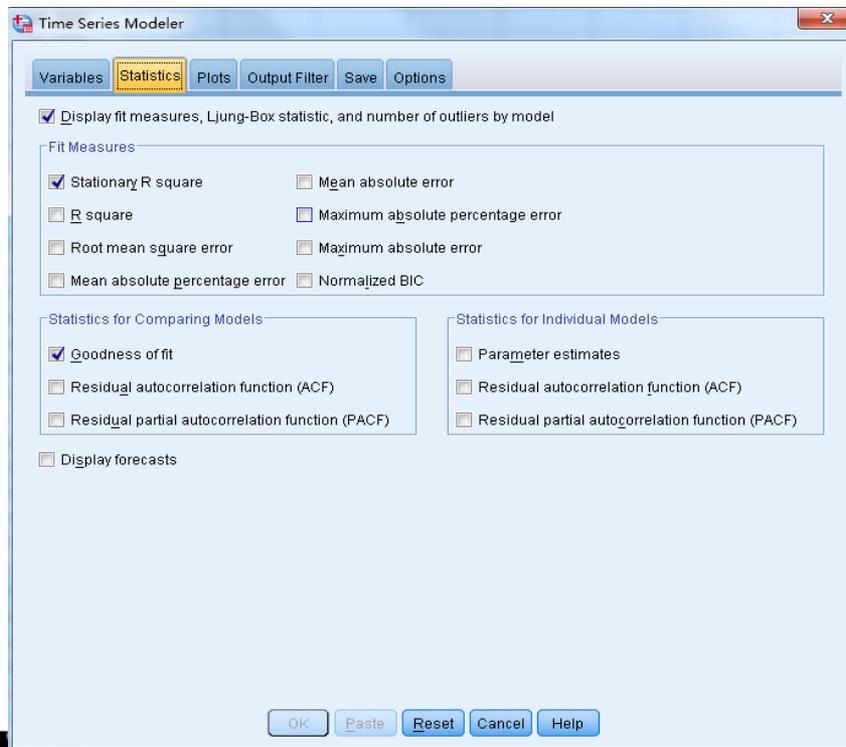
## Step02 : 指数平滑模型选择

在该对话框的左侧的【Variables(变量)】列表框中选择一个变量，将其移入【Dependent Variables(因变量)】列表框。在【Method(模型)】下拉列表框中选择建模方法，在【Method(模型)】下拉列表框中选择【Exponential Smoothing(指数平滑法)】选项，并单击【Criteria(条件)】按钮，弹出【Exponential Smoothing Criteria(指数平滑条件)】对话框。



## Step03 : 统计量的选择

在【Create Models(创建模型)】对话框的菜单中，选择【Statistics(统计量)】，弹出【Statistics(统计量)】对话框。

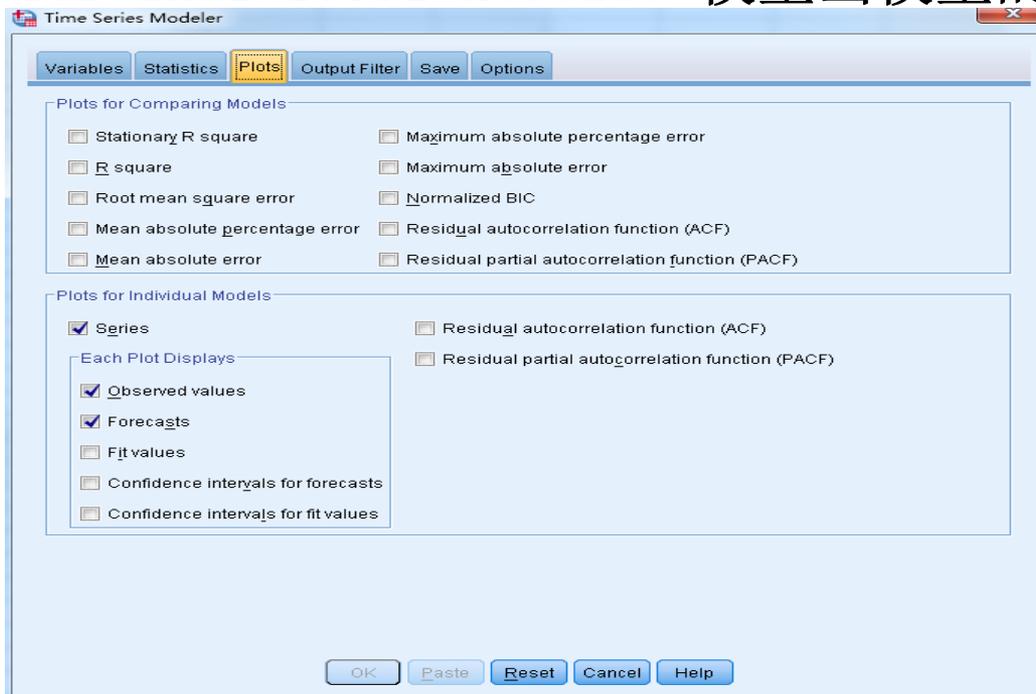


## Step04 : 图表的选择

【Plot (图表)】选项卡分成两部分。

① Plots for Comparing Models: 模型比较图。

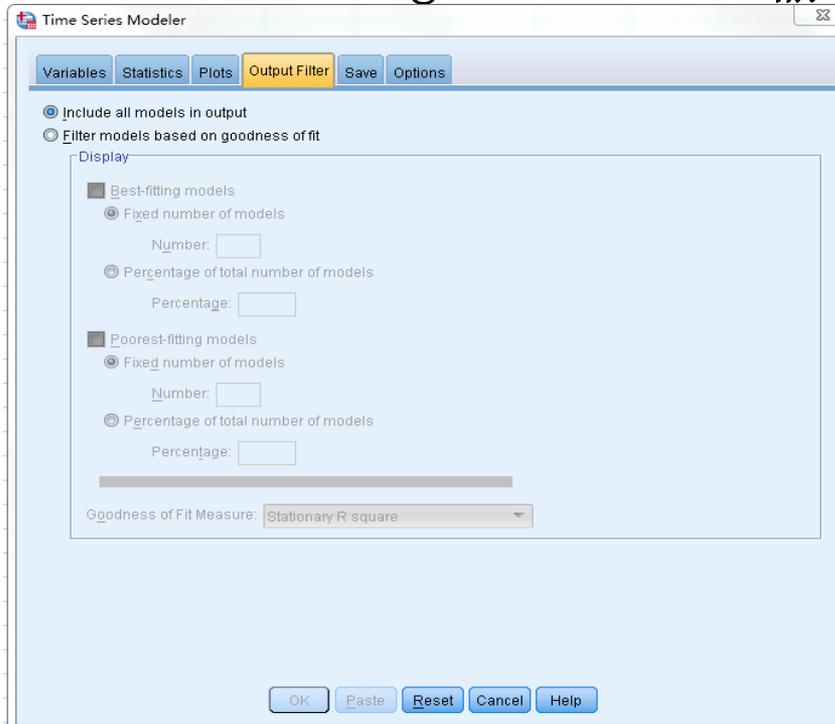
② Plots for individual Models: 模型当模型的图。



## Step05 : 输出的选择

【Output Filter (输出过滤)】选项卡中包括两部分。

① Include all models in output: 输出所有的模型, 系统默认选项。  
② Filter models based on goodness fit 输出基于拟合优度过滤的模型。



## Step06: 保存变量的选择

在【Save(保存)】选项卡中包括两部分。①Save Variables: 保存变量; ②Export Model File: 选择是否导出模型文件保存变量, 将模型文件保存在指定的目录中。

选择好以后, 在【Create Models(创建模型)】对话框的菜单中, 单击【Options(选项)】按钮, 弹出【Options(选项)】对话框。

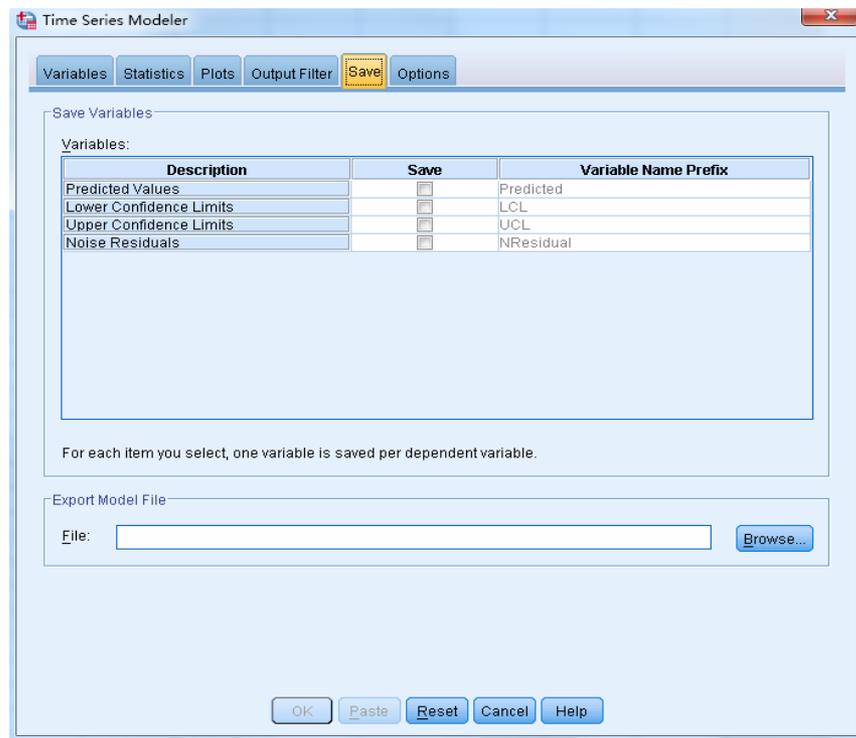
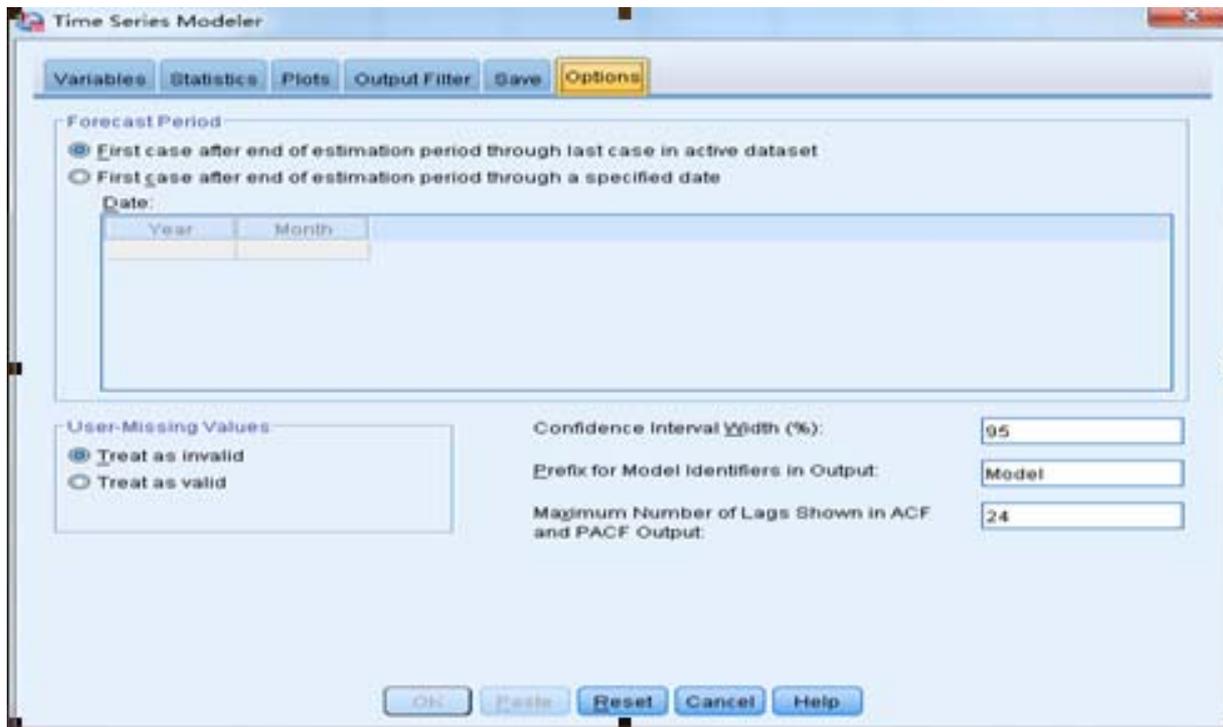


图11-24

CONCEPT  
STRATE

## Step07: 某些选项的选择



# 11.2.3 实例图文分析：进出口贸易总额的指数平滑建模

CONCEPT  
RATE

## 1. 实例内容

以我国1950-2005年进出口贸易总额年度数据为例，尝试建立指数平滑模型。

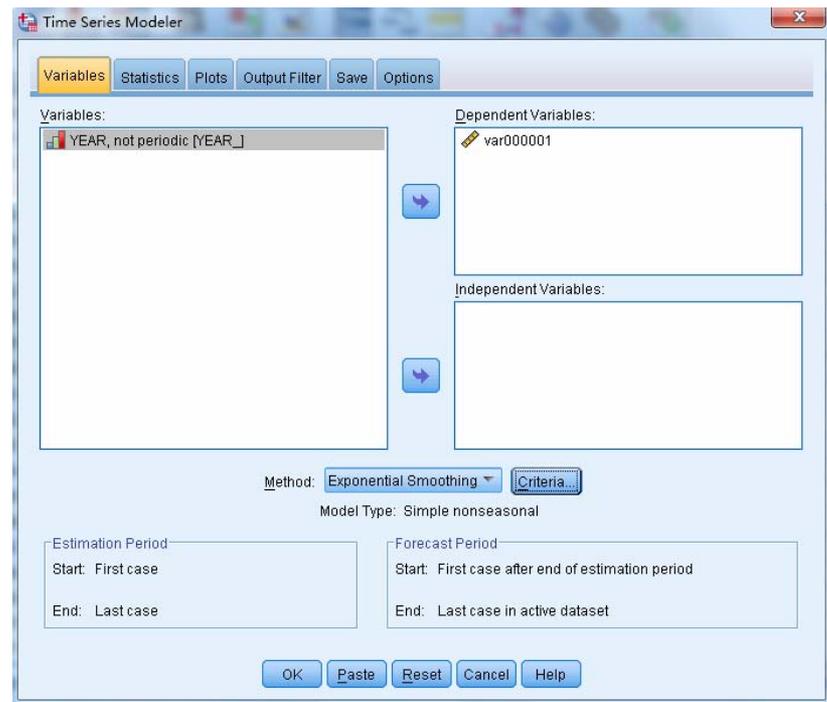
表 11-4 我国 1950-2005 年进出口贸易总额年度数据

41.5	59.5	64.6	80.9	84.7	109.8
108.7	104.5	128.7	149.3	128.4	90.7
80.9	85.7	97.5	118.4	127.1	112.2
108.5	107.7	112.9	120.9	146.9	220.5
292.2	290.4	264.1	272.5	355	454.6
570	735.3	771.3	860.1	1201	2066.7
2580.4	3084.2	3821.8	4155.9	5560.1	7225.8
9119.6	11271	20381.9	23499.9	24133.8	26967.2
26854.1	29896.3	39273.2	42183.6	51378.2	70483.5
95539.1	116921.8				

## 2. 实例操作

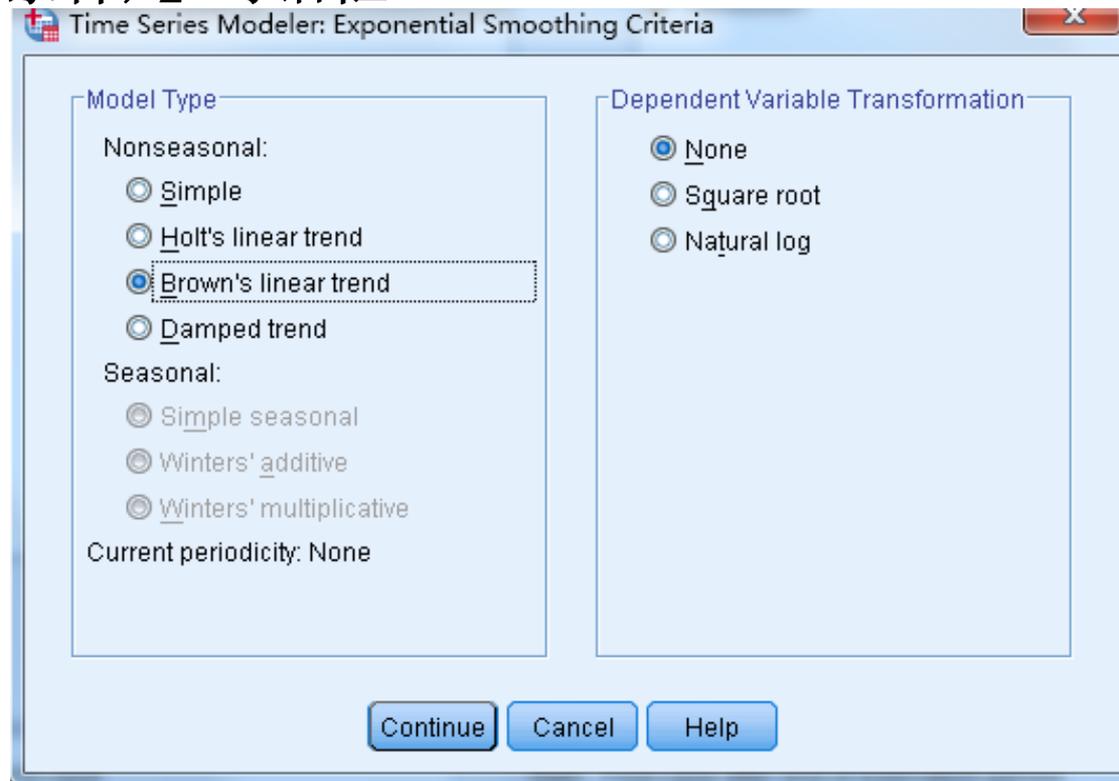
Step01: 打开【Create Models(创建模型)】对话框

选择菜单栏中的【Analyze(分析)】→【Forecasting(预测)】→【Create Models(创建模型)】命令，弹出【Create Models(创建模型)】对话框。将该对话框左侧的【VAR00001】变量移入【Dependent Variables(因变量)】列表。在【Method(模型)】下拉列表框中选择【Exponential Smoothing(指数平滑法)】选项。



CONCEPT  
STRATE

单击【Criteria(条件)】按钮，弹出【Exponential Smoothing Criteria(指数平滑条件)】对话框。

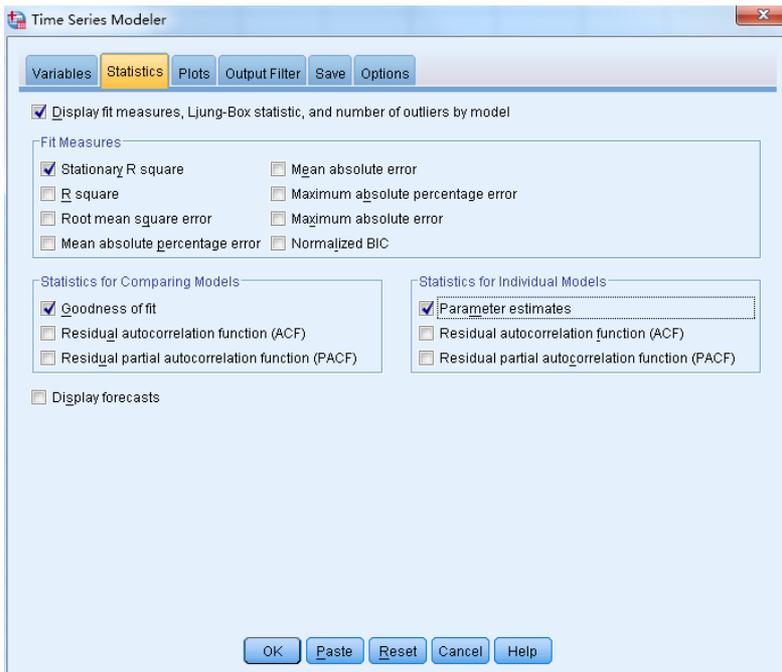


## Step02: 指数平滑模型选择

CONCEPT  
STRATE

由于数据具有明显的趋势性，所以选【Brown's linear trend (Brown线性趋势)】，点击【Continue(继续)】，返回到了【Create Models(创建模型)】对话框。

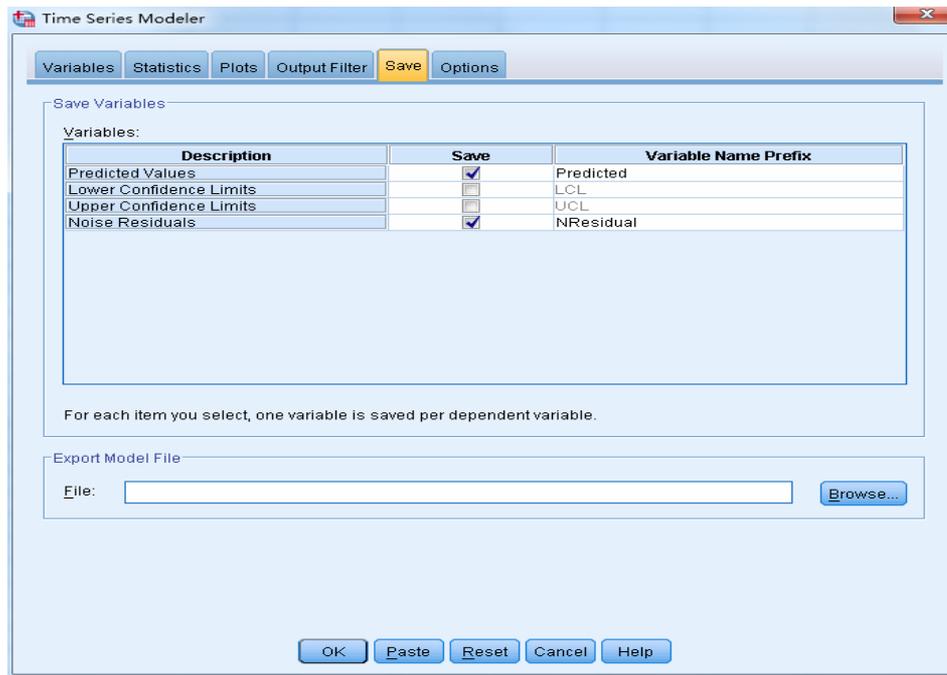
单击【Statistics(统计量)】选项卡，弹出如下图所示的界面。



# Step03 : 统计量的选择

CONCEPT  
STRATE

在【Statistics(统计量)】选项卡中，选择对展示模型拟合度量、jung -Box 统计量、被模型过滤掉的样本数据的个数的选项，选择显示模型参数的估计值，选择好以后，单击【Save(保存)】选项卡，对话框显示如下图所示。



# Step05 : 完成操作

CONCEPT  
RATE

选择好以后，单击【OK(确认)】输出结果，此时，SPSS将在当前数据编辑窗口中自动生成带前缀Predicted的预测值和带前缀NResidual的残差的值。

var000001	YEAR_	DATE_	Predicted_var 000001_Mod el_1	NResidual_va r000001_Mod el_1
41.50000	1950	1950	42.11159	-.61159
59.50000	1951	1951	58.48363	1.01637
64.60000	1952	1952	77.46526	-12.86526
80.90000	1953	1953	70.13779	10.76221
84.70000	1954	1954	96.83030	-12.13030
109.80000	1955	1955	88.91561	20.88439
108.70000	1956	1956	134.18630	-25.48630
104.50000	1957	1957	108.47272	-3.97272
128.70000	1958	1958	100.42773	28.27227
149.30000	1959	1959	151.93743	-2.63743
128.40000	1960	1960	169.99786	-41.59786
90.70000	1961	1961	108.91381	-18.21381
80.90000	1962	1962	53.60735	27.29265
85.70000	1963	1963	70.16662	15.53338
97.50000	1964	1964	89.97966	7.52034
118.40000	1965	1965	109.04876	9.35124
127.10000	1966	1966	138.98418	-11.88418
112.20000	1967	1967	136.20684	-24.00684
108.50000	1968	1968	98.11294	10.38706
107.70000	1969	1969	104.43984	3.26016
112.90000	1970	1970	106.79214	6.10786
120.90000	1971	1971	117.89324	3.00676
146.90000	1972	1972	128.79952	18.10048
220.50000	1973	1973	172.28535	48.21465



### 3 实例结果及分析

#### (1) 模型描述

该模型为Model\_1，模型的类型为Brown的线性趋势模型。

表 11-4 模型描述

			Model Type
Model ID	var000001	Model_1	Brown

#### (2) 模型拟合优度

对VAR00001建立Winters的乘积季节模型的拟合优度，包括了调整R-Square, 标准化的BIC等所有拟合优度的值。





### (3) 模型的统计量的结果

由于在【Statistics(统计量)】对话框中，选择了展示模型拟合度量、Ljung-Box统计量、被模型过滤掉的样本数据的个数的选项，所以，在输出结果中出现了调整R-Square, 标准化的BIC的值, Ljung-Box统计量的值。

从表10-5中可以看出Box-Ljung 统计量的相伴概率是0.524，可以接受残差序列是没有自相关性的。

表 11-5 模型的拟合优度

Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
var000001-Model_1	0	-.022	15.995	17	.524	0

由于在【Statistics(统计量)】对话框中，选择显示模型参数的估计值，所以，在输出结果中出现模型的参数估计的结果。从表10-6可以看出，水平指标的估计值是0.492，趋势指标的估计值是0.071，季节效应指标为0.849，T统计量的相伴概率都接受这些参数都是为非零的假设的。

表 11-6 参数的估计

Model			Estimate	SE	t	Sig.
var000001-	No	Alpha (Level	.983	.068	14.368	.000
Model_1	Transformation	and Trend)				

即模型为：

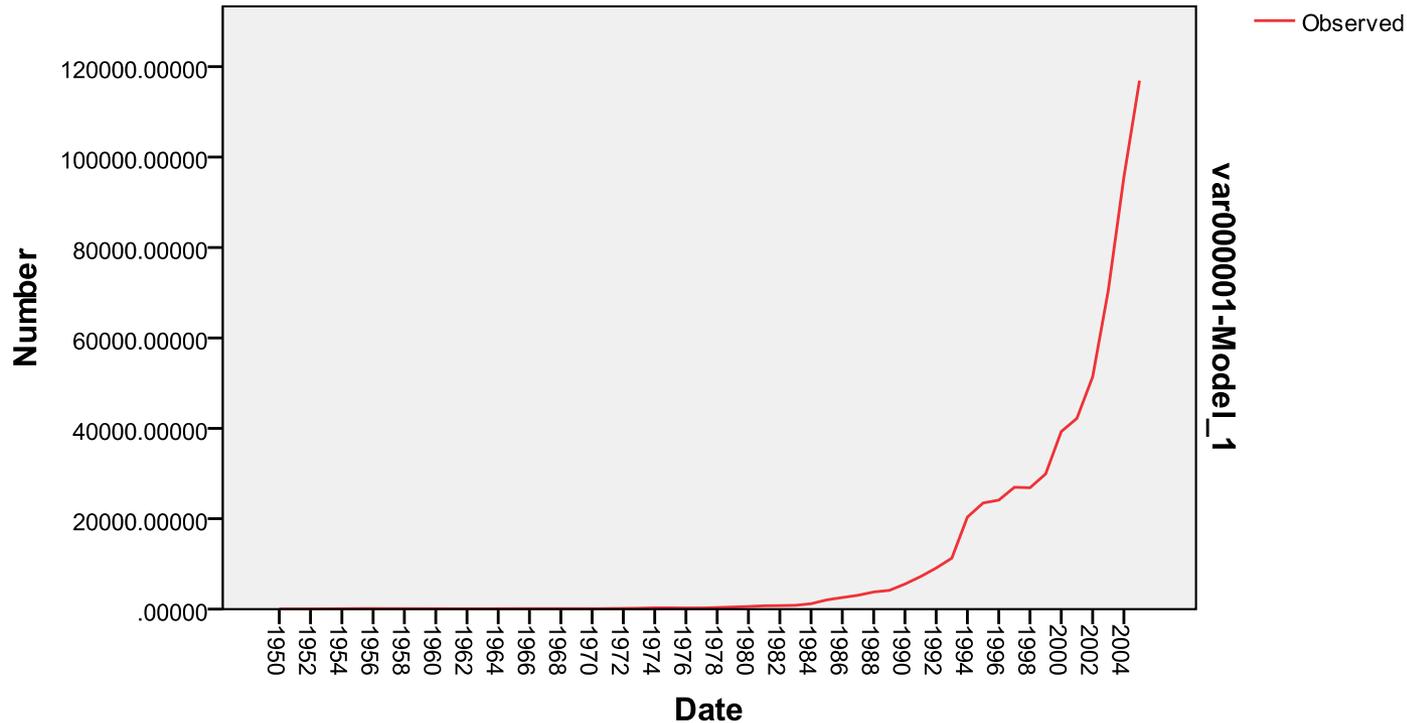
$$\hat{x}_t(l) = a_t + b_t l$$

式中  $a_t = 2S_t^r - S_t^m$ ，  $b_t = \frac{\alpha}{1-\alpha} (S_t^r - S_t^m)$ ，  $\alpha = 0.983$ 。

## (4) 模型的拟合图

CONCEPT  
RATE

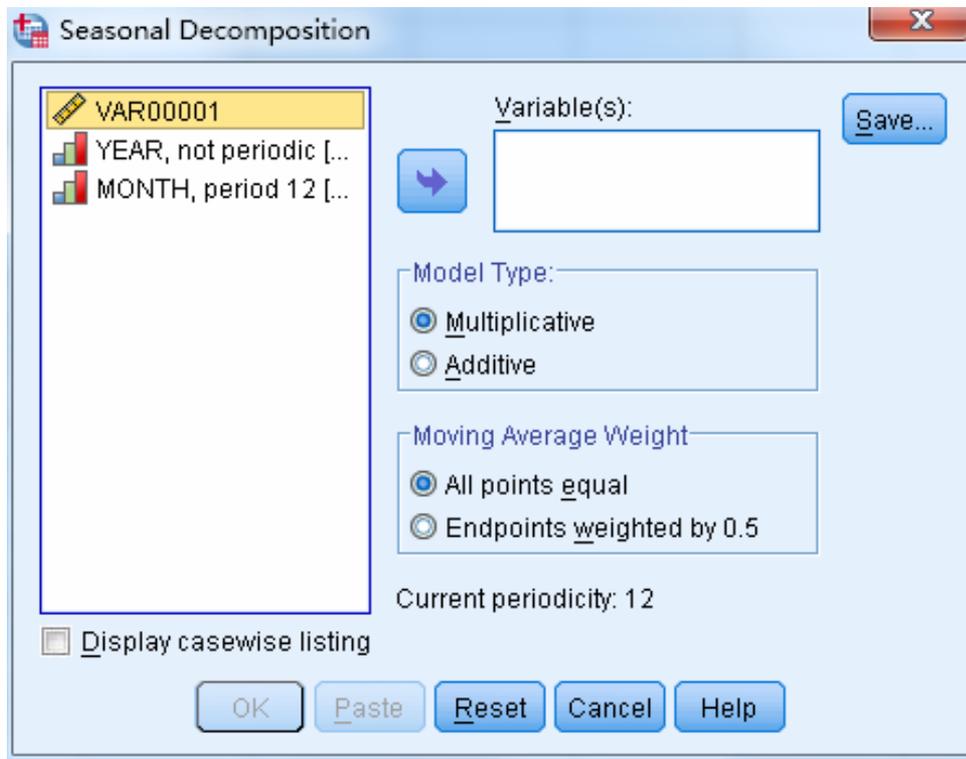
在获得了参数估计值和模型结构后，代入初值，便可以拟合数据，从而绘制图像。拟合数据以前缀为Predicted的变量Predicted—VAR000001— Model—1出现在SPSS的当前数据编辑窗口中。



## 11.2.4 季节分解的SPSS操作详解

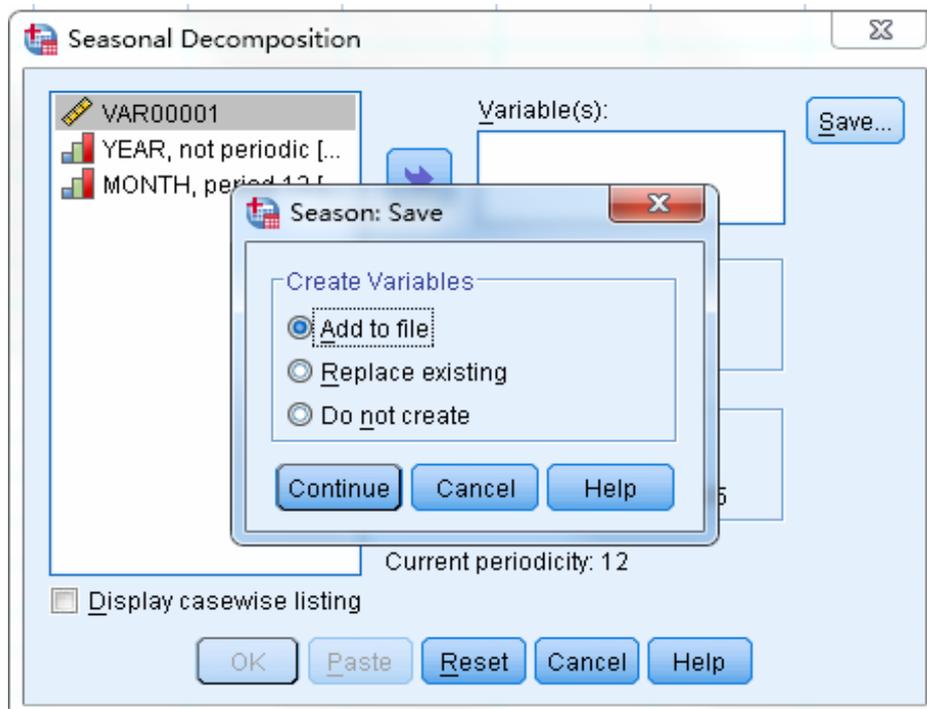
CONCEPT  
STRATE

Step01 : 选择菜单栏中的【Analyze(分析)】→【Forecasting(预测)】→【Seasonal Decomposition(周期性分解)】命令，弹出【Seasonal Decomposition(周期性分解)】对话框。



## Step02 : 季节分解模型的选择

在【Seasonal Decomposition(周期性分解)】对话框的左侧的候选变量列表框中选择一个变量，将其移入【Variables(变量)】列表框。在【Model Type(模型类型)】复选框中选择模型类型；单击【Save】按钮，弹出【Save(保存)】对话框。



CONCEPT  
RATE

### Step03 : 完成操作

如果不改变【Save(保存)】对话框中的默认选项，单击【Seasonal Decomposition(周期性分解)】对话框中的【OK(确认)】按钮，将进行季节分解。

# 11.2.3 实例图文分析：社会住宿与餐饮消费的季节分解

CONCEPT  
STRATE

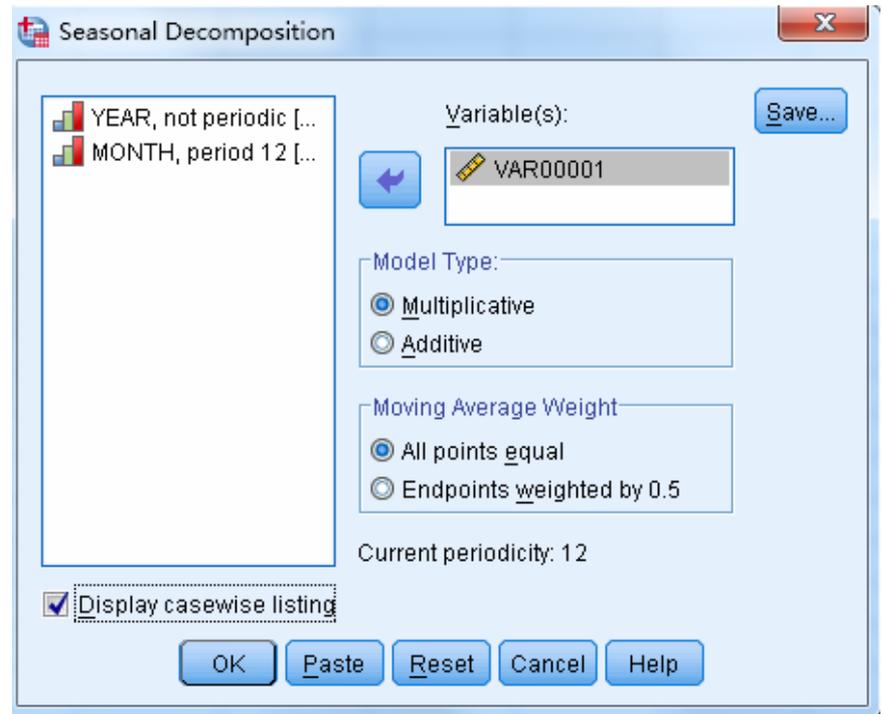
## 1. 实例内容

以我国1996年—2009年的社会住宿与餐饮消费（单位为亿元）的月度数据为例，尝试进行季节分解。

## 2. 实例操作

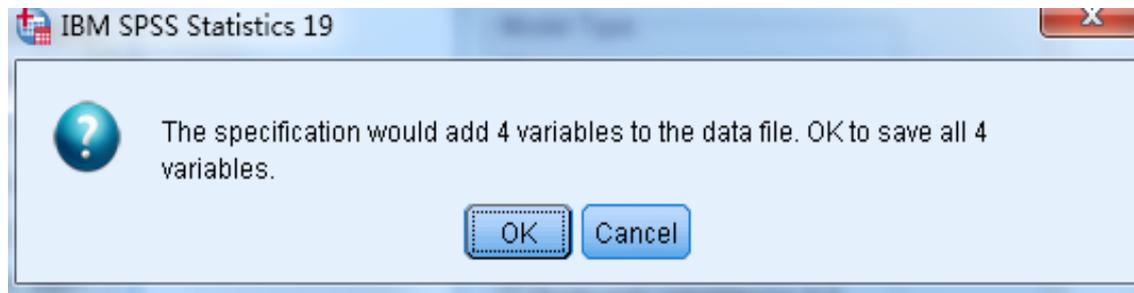
Step01: 打开【Seasonal Decomposition(周期性分解)】对话框

选择菜单栏中的【Analyze(分析)】→【Forecasting(预测)】→【Seasonal Decomposition(周期性分解)】命令，弹出【Seasonal Decomposition(周期性分解)】对话框。将该对话框左侧的【VAR00001】移入【Variables(变量)】列表框。在【Model Type(模型类型)】列表框中选择【Multiplicative】，在【Moving Average Weight(移动平均权重)】列表框中的选择【All points equal(所有点相等)】，并选择【Display casewise listing(显示对象删除列表)】显示对象删除列表。



## Step02: 完成操作

单击【Seasonal Decomposition(周期性分解)】对话框中的【OK(确认)】按钮，此时，SPSS将弹出一个对话框，提示在当前数据编辑窗口中将自动生成四个变量，再单击【OK(确认)】按钮，完成操作。





### Step03: 数据窗口的变化

单击【OK(确认)】按钮后，在当前数据编辑窗口将自动生成四个变量。第一个变量为不规则变动因素（前缀ERR），第二变量为季节调整后的变量（前缀SAS），第三变量为季节因子（前缀SAF），第四个变量为平滑后的趋势和循环波动变量（前缀STC）。

VAR00001	YEAR_	MONTH_	DATE_	ERR#1	SAS#1	SAF#1	STC#1
237.60	1996	1	JAN 1996	.95104	217.38318	1.09300	228.57476
234.90	1996	2	FEB 1996	1.00901	234.45771	1.00189	232.36391
242.60	1996	3	MAR 1996	1.02212	245.25084	.98919	239.94221
229.50	1996	4	APR 1996	1.01240	246.85159	.92971	243.82756
239.60	1996	5	MAY 1996	1.00032	243.72560	.98307	243.64744
230.10	1996	6	JUN 1996	.99522	241.48265	.95286	242.64326
228.70	1996	7	JUL 1996	.99395	239.73088	.95399	241.18931
236.80	1996	8	AUG 1996	1.02362	245.57679	.96426	239.91126
235.70	1996	9	SEP 1996	.98048	233.66662	1.00870	238.31791
241.90	1996	10	OCT 1996	.98450	234.19334	1.03291	237.88146
250.00	1996	11	NOV 1996	1.02733	244.59023	1.02212	238.08330
252.40	1996	12	DEC 1996	.99553	236.26259	1.06830	237.32438
266.20	1997	1	JAN 1997	.98592	234.40055	1.09300	237.74748
235.40	1997	2	FEB 1997	.97875	234.95677	1.00189	240.05686
246.80	1997	3	MAR 1997	1.01588	249.49673	.98919	245.59769
233.90	1997	4	APR 1997	1.00713	251.58425	.92971	249.80375
250.10	1997	5	MAY 1997	1.00469	254.40640	.98307	253.21770
238.90	1997	6	JUN 1997	.98228	250.71797	.95286	255.24100
249.60	1997	7	JUL 1997	1.02157	261.63895	.95399	256.11524
252.00	1997	8	AUG 1997	1.02874	261.34017	.96426	254.03832
243.70	1997	9	SEP 1997	.96365	241.59760	1.00870	250.71046
253.20	1997	10	OCT 1997	.97851	245.13333	1.03291	250.51779
262.70	1997	11	NOV 1997	1.02936	257.01542	1.02212	249.68566
278.50	1997	12	DEC 1997	1.05263	260.69387	1.06830	247.65941
243.60	1998	1	JAN 1998	.91778	222.87266	1.09300	242.84008
242.40	1998	2	FEB 1998	.98561	241.94359	1.00189	245.47702

### 3 实例结果及分析

#### (1) 模型描述

该模型为MOD\_1，模型的类型为Multiplicative模型，季节的周期长度为12，移动平均的方法是跨度为周期长度的等权重的中心移动平均。

表 11-7 模型描述

Model Name	MOD_1	
Model Type	Multiplicative	
Series Name	1	VAR00001
Length of Seasonal Period		12
Computing Method of Moving Averages	Span equal to the periodicity and all points weighted equally	
Applying the model specifications from MOD_1		

## (2) 季节分解表

由于选择【Display casewise listing(显示对象删除列表)】，所以，显示季节分解表。表中第一列为时间变量，第二列为原始数据。第三列为移动平均序列，第四列为原始数据除以移动平均序列的比值；第五列是季节因子，第六列是季节调整后的数据，第七列为平滑后的趋势和循环波动变量，第七列为不规则变动因素。



表 11-8 季节分解表 (部分)

DATE_	Original Series	Moving Average Series	Ratio of Original Series to Moving Average Series (%)	Seasonal Factor (%)	Seasonally Adjusted Series	Smoothed Trend-Cycle Series	Irregular (Error) Component
JAN 1996	237.600	.	.	109.3	217.383	228.575	.951
FEB 1996	234.900	.	.	100.2	234.458	232.364	1.009
MAR 1996	242.600	.	.	98.9	245.251	239.942	1.022
APR 1996	229.500	.	.	93.0	246.852	243.828	1.012
MAY 1996	239.600	.	.	98.3	243.726	243.647	1.000
JUN 1996	230.100	.	.	95.3	241.483	242.643	.995
JUL 1996	228.700	238.3167	96.0	95.4	239.731	241.189	.994
AUG 1996	236.800	239.8667	98.7	96.4	245.577	239.911	1.024
SEP 1996	235.700	239.9083	98.2	100.9	233.667	238.318	.980
OCT 1996	241.900	240.2583	100.7	103.3	234.193	237.881	.984
NOV 1996	250.000	240.6250	103.9	102.2	244.590	238.083	1.027
DEC 1996	252.400	241.5000	104.5	106.8	236.263	237.324	.996
JAN 1997	256.200	242.2333	105.8	109.3	234.401	237.747	.986
FEB 1997	235.400	243.9750	96.5	100.2	234.957	240.057	.979
MAR 1997	246.800	245.2417	100.6	98.9	249.497	245.598	1.016
APR 1997	233.900	245.9083	95.1	93.0	251.584	249.804	1.007
MAY 1997	250.100	246.8500	101.3	98.3	254.406	253.218	1.005

# 11.3 时间序列的随机性分析

CONCEPT  
RATE

## 11.3.1 随机性分析的原理

### 1. 使用目的

虽然长期趋势的分析，季节变动的分析和循环波动的分析控制着时间序列变动的基本样式，但毕竟不是时间序列变动的全貌，而且用随机过程理论和统计理论来考察长期趋势、季节性变动等许多因素的共同作用的时间序列更具有合理性和优越性。根据随机过程理论和统计理论，对时间序列进行分析，从而形成了时间序列的随机分析。

通过随机性时间序列分析，一方面能够建立比较精确地反映序列中所包含的动态依存关系的数学模型，并借以对系统的未来进行预报，另一方面，能够比较精确揭示系统动态结构和规律的统计方法。随机性时间序列分析大大丰富和发展了时间序列分析的理论和方法，成为时间序列分析的主流。

## 2、基本原理

时间序列的随机分析通常利用Box-Jenkins建模方法。利用Box-Jenkins方法建模的步骤为：

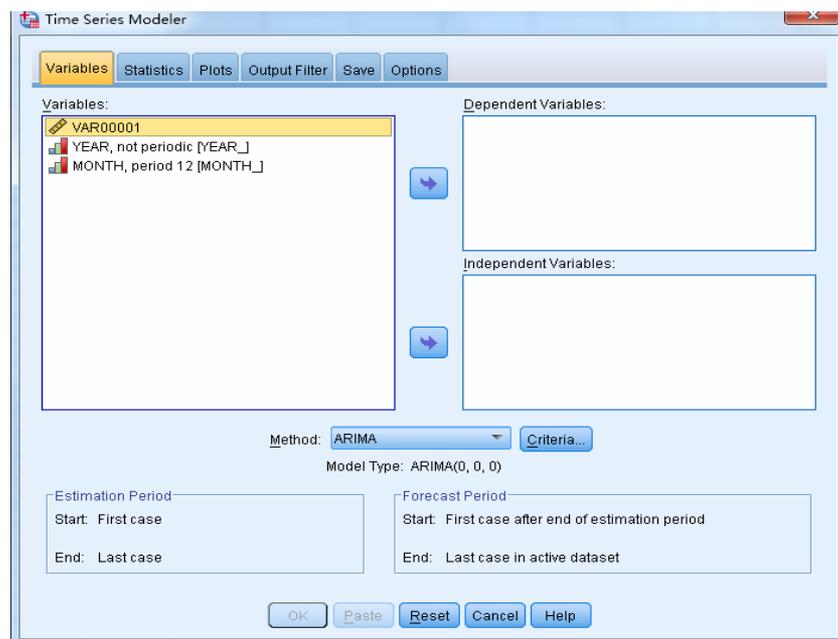
- (1) 计算观测序列的样本相关系数和样本偏相关系数。
- (2) 模式识别：检验序列是否为平稳非白噪声序列。如果序列是白噪声序列，建模结束；如果序列为非平稳序列，采用非平稳时间序列的建模方法，建立ARIMA模型或SARIMA模型；如果序列为平稳序列，建立ARMA模型。
- (3) 初步定阶和参数估计：模型识别后，框定所属模型的最高阶数；然后在已识别的类型中，从低阶到高阶对模型进行拟合及检验。
- (4) 拟合优度检验：利用定阶方法对不同的模型进行比较，以确定最适宜的模型。
- (5) 适应性检验：对选出的模型进行适应性检验和参数检验，进一步从选出的模型出发确定最适宜的模型。
- (6) 预测：利用所建立的模型，进行预测。

## 11.3.2 ARIMA模型的SPSS操作详解

CONCEPT  
STRATE

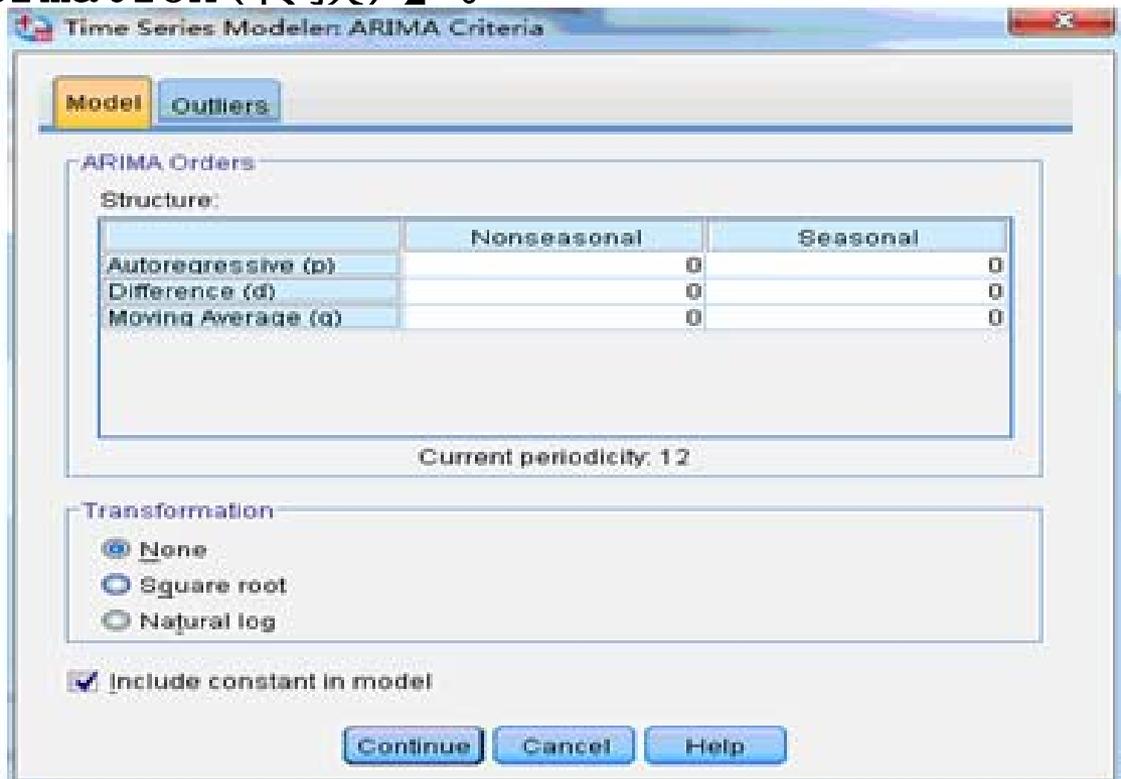
Step01：打开【Create Models(创建模型)】对话框

当时间序列的数据已经准备好以后，xz选择菜单栏中的【Analyze(分析)】→【Forecasting(预测)】→【Create Models(创建模型)】命令，弹出【Create Models(创建模型)】对话框。在该对话框左侧的【Variables(变量)】列表框中选择一个变量，将其移入【Dependent Variables(因变量)】列表框。在【Method(模型)】下拉列表框中选择【ARIMA】，然后选择【ARIMA】选项，并单击【Criteria(条件)】按钮，弹出【ARIMA Criteria(ARIMA条件)】对话框。



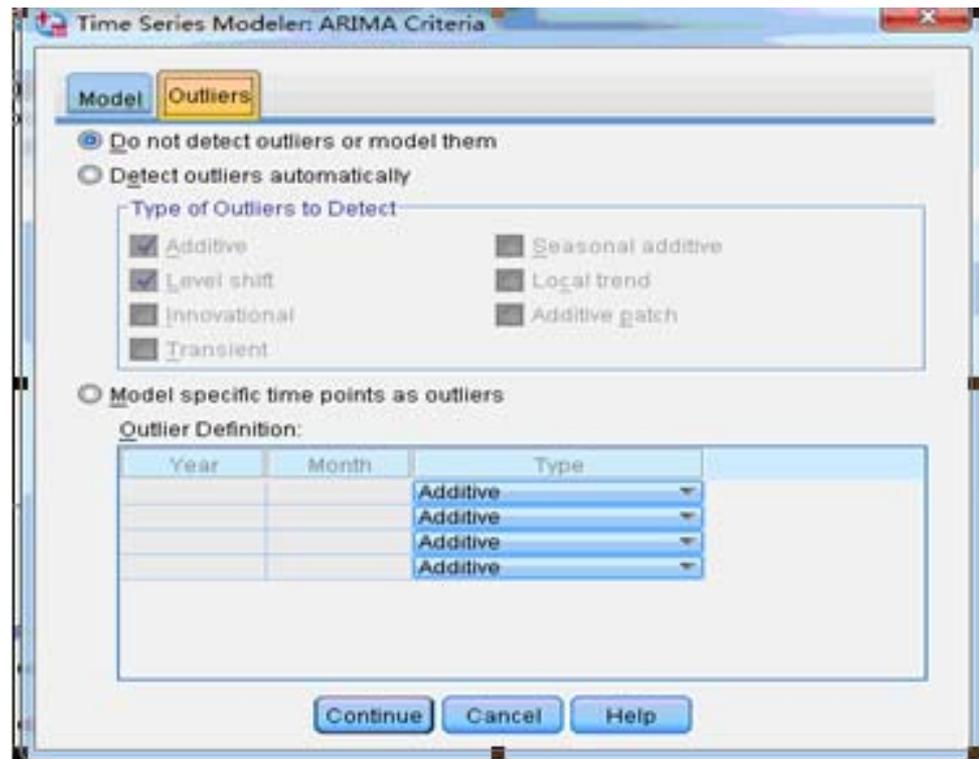
## Step02 : ARIMA模型选择

对话框中的第一部分为【ARIMA Order (ARIMA序列)】，第二部分为【Transformation (转换)】。



### Step03 : 离群值的处理

在【ARIMA Criteria(ARIMA条件)】对话框中单击【Outliers(离群值)】选项卡，弹出【Outliers(离群值)】对话框，这样可以选择对离群点的处理方式。





## Step04 : 完成操作

单击【Create Models(创建模型)】对话框中的【OK(确认)】按钮，将进行ARIMA模型建模。

# 11.3.3 实例图文分析：旅客周转量的ARIMA建模

CONCEPT  
STRATE

## 1. 实例内容

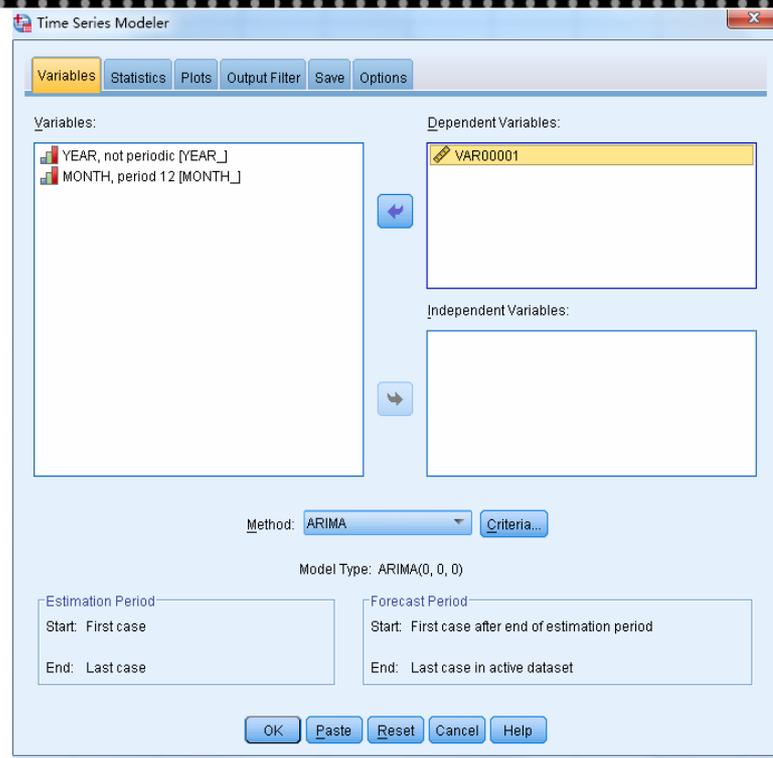
以我国2004年1月到2009年12月旅客周转量的数据为例，尝试建立ARIMA模型。

## 2. 实例操作

# Step01: 打开【Seasonal Decomposition(周期性分解)】对话框

CONCEPT  
STRATE

选择菜单栏中的【Analyze(分析)】→【Forecasting(预测)】→【Create Models(创建模型)】命令，弹出【Create Models(创建模型)】对话框。将该对话框左侧的【VAR00001】移入【Dependent Variables(因变量)】列表框。在【Method(模型)】下拉列表框中选择【ARIMA】，并选择【Criteria(条件)】选项，弹出【ARIMA Criteria(ARIMA条件)】对话框。



CONCEPT  
STRATE

## Step02: ARIMA模型选择

在【ARIMA Order (ARIMA序列)】选项组中输入阶数都为1，建立ARIMA (1, 1, 1) (1, 1, 1)模型，单击【Continue (继续)】按钮，返回【Create Models (创建模型)】对话框。

Time Series Modeler: ARIMA Criteria

Model Outliers

ARIMA Orders

Structure:

	Nonseasonal	Seasonal
Autoregressive (p)	1	1
Difference (d)	1	1
Moving Average (q)	1	1

Current periodicity: 12

Transformation

None  
 Square root  
 Natural log

Include constant in model

Continue Cancel Help

# Step03: 统计量的选择

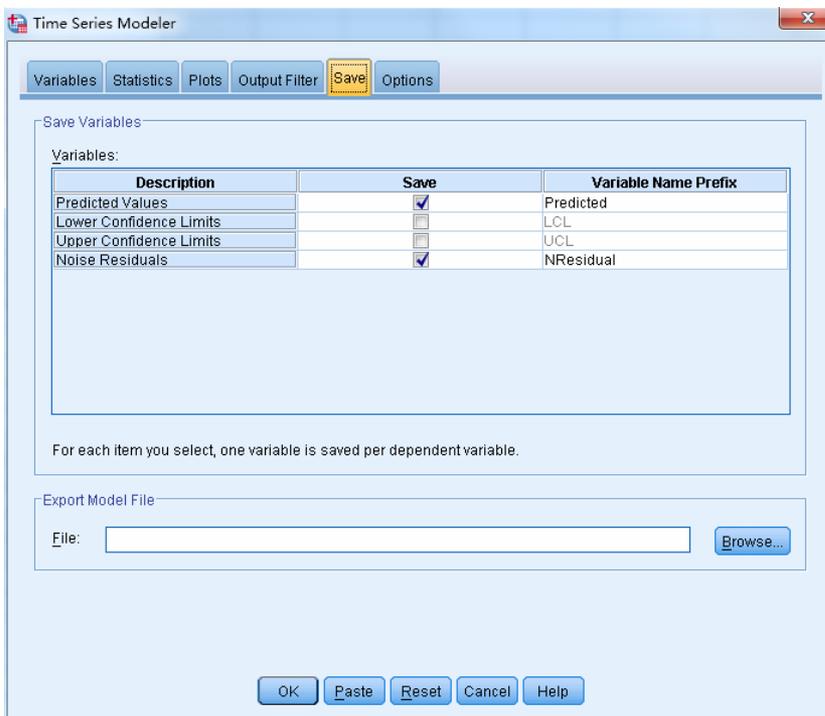
CONCEPT  
TRATE

单击【Create Models(创建模型)】中的【Statistics(统计量)】对话框中，选择展示模型拟合度量、Box-ljung 统计量、被模型过滤掉的样本数据的个数的选项，选择显示模型参数的估计值，选择好以后，单击【Save(保存)】选项卡。



## Step04: 保持变量的选择

在 【Save (保存)】 选项卡中选择保存预测值，保存残差的值。





## Step05: 完成操作

单击对话框中的【OK(确认)】按钮,将进行ARIMA模型建模,完成操作。此时,输出结果,同时在当前数据编辑窗口中自动生成带前缀Predicted的预测值和带前缀NResidual的残差的值。

VAR00001	YEAR_	MONTH_	DATE_	Predicted_VA R00001_Mod el_1	NResidual_V AR00001_Mo del_1
1359.630000	2004	7	JUL 2004	.	.
1418.500000	2004	8	AUG 2004	.	.
1344.360000	2004	9	SEP 2004	.	.
1390.590000	2004	10	OCT 2004	.	.
1249.700000	2004	11	NOV 2004	.	.
1205.430000	2004	12	DEC 2004	.	.
1429.600000	2005	1	JAN 2005	.	.
1793.890000	2005	2	FEB 2005	1377.636716	416.253284
1451.370000	2005	3	MAR 2005	1457.377090	-6.007090
1380.580000	2005	4	APR 2005	1435.143857	-54.563857
1439.820000	2005	5	MAY 2005	1481.610014	-41.790014
1357.860000	2005	6	JUN 2005	1384.033069	-26.173069
1494.980000	2005	7	JUL 2005	1495.988820	-1.008820
1529.310000	2005	8	AUG 2005	1555.585183	-26.275183
1414.060000	2005	9	SEP 2005	1476.927981	-62.867981
1514.570000	2005	10	OCT 2005	1513.069279	1.500721
1368.430000	2005	11	NOV 2005	1374.329977	-5.899977
1329.030000	2005	12	DEC 2005	1331.496685	-2.466685
1660.040000	2006	1	JAN 2006	1626.987021	33.052979
1734.930000	2006	2	FEB 2006	1819.628989	-84.698989
1562.360000	2006	3	MAR 2006	1547.773352	14.586648
1500.470000	2006	4	APR 2006	1506.564955	-6.094955
1564.110000	2006	5	MAY 2006	1566.018221	-1.908221
1498.750000	2006	6	JUN 2006	1481.342978	17.407022
1655.590000	2006	7	JUL 2006	1612.882839	42.707161
1697.240000	2006	8	AUG 2006	1664.762724	32.477276

# 3 实例结果及分析



## (1) 模型描述

该模型为Model\_1，模型的类型为ARIMA(1, 1, 1)(1, 1, 1)模型。

表 11-9 模型描述

			Model Type
Model ID	VAR00001	Model_1	ARIMA(1,1,1)(1,1,1)

## (2) 模型拟合优度

对VAR00001建立ARIMA(1, 1, 1)(1, 1, 1)模型的拟合优度，包括了调整R-Square, 标准化的BIC等所有拟合优度的值。



### (3) 模型的统计量的结果

由于在【Statistics(统计量)】对话框中，选择了展示模型拟合度量、Ljung-Box统计量、被模型过滤掉的样本数据的个数的选项，所以，在输出结果中出现了调整R-Square，标准化的BIC的值，Ljung-Box统计量的值。

从表11-11中可以看出标准BIC值为9.187。

表 11-11 模型的拟合优度

Model	Number of Predictors	Model Fit statistics		Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Normalized BIC	Statistics	DF	Sig.	
VAR00001-Model_1	0	.433	9.187	3.252	14	.999	0

由于在【Statistics(统计量)】对话框中，选择显示模型参数的估计值，所以，在输出结果中出现模型的参数估计的结果。

表 11-12 ARIMA 模型参数估计

				Estimate	SE	t	Sig.	
VAR0000	VAR0	No	Constant	.297	.812	.365	.716	
1-Model_	0001	Transforma- tion	AR	Lag 1	.029	.168	.175	.862
1	Difference		1					
	MA		Lag 1	.884	.104	8.499	.000	
	AR, Seasonal		Lag 1	.135	.366	.369	.714	
	Seasonal Difference		1					
	MA, Seasonal		Lag 1	.989	11.301	.088	.931	

从表11-12可以看出，AR(1)的参数估计值是0.029，T统计量的相伴概率为0.862，接受AR(1)为零的原假设。MA(1)的参数估计值是0.884，T统计量的相伴概率为8.499，拒绝MA(1)为零的原假设。SAR(1)的参数估计值是0.366，所以，该模型不是最优的模型，对数据的分析不是十分的恰当。



## (4) 模型的拟合图

在获得了参数估计值和模型结构后，代入初值，便可以拟合数据，从而绘制图像。拟合数据以前缀为Predicted的变量Predicted\_VAR000001\_Model\_1出现在SPSS的当前数据编辑窗口中。

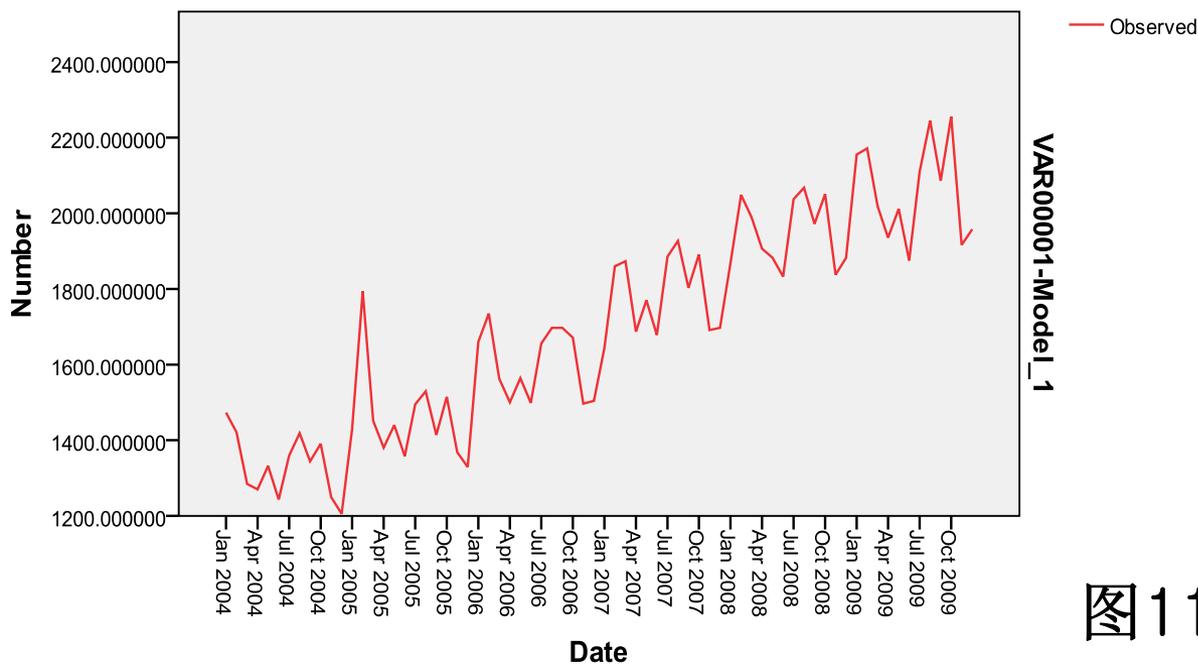


图11-45

## 11.3.4 实例进阶分析

CONCEPT  
STRATE

### 1. 进阶分析

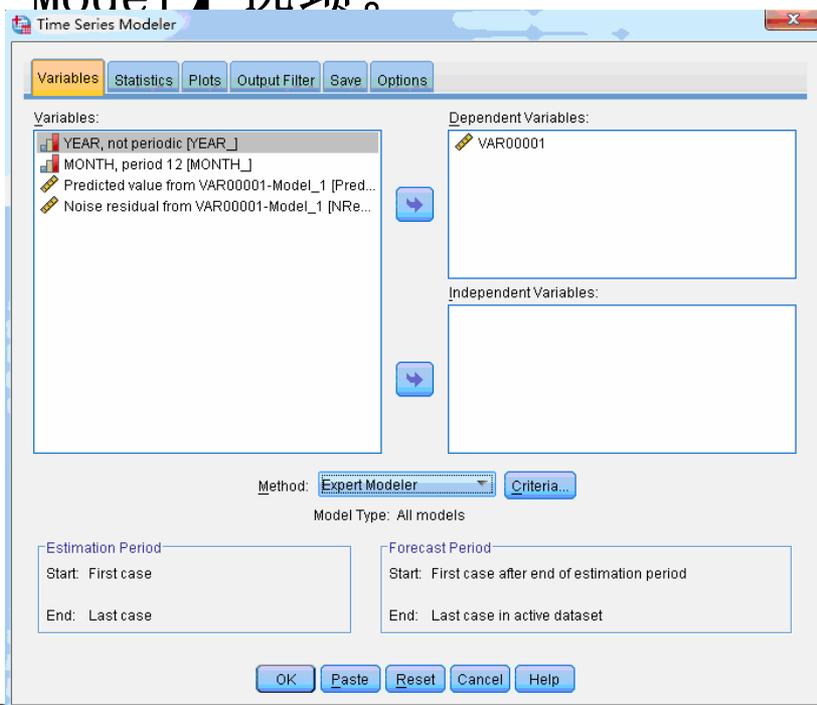
上面所建立的ARIMA(1, 1, 1)(1, 1, 1)模型并不是最佳的模型，所以，需要重新建模，可以利用专家建模器来完成。

### 2. 实例操作

# Step01: 打开【Seasonal Decomposition (周期性分解)】对话框

CONCEPT  
TRATE

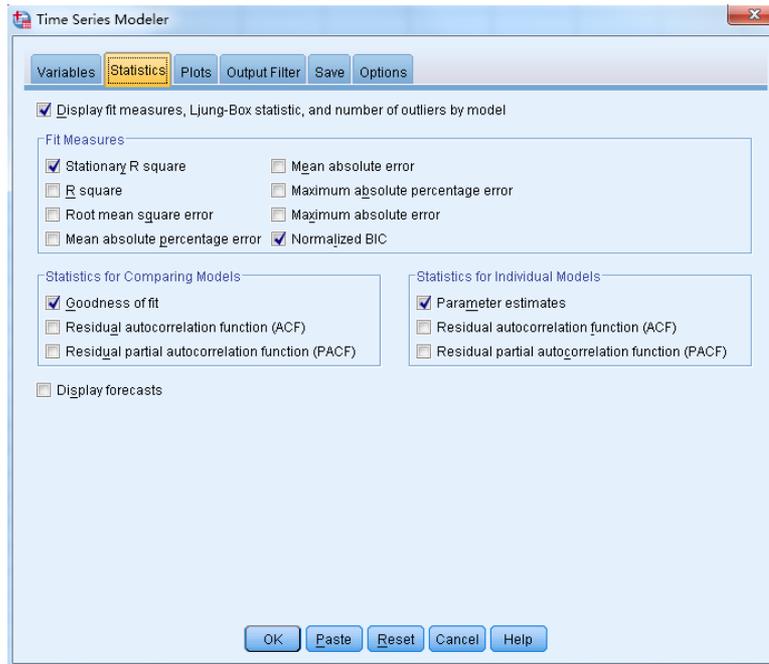
选择菜单栏中的【Analyze(分析)】→【Forecasting(预测)】→【Create Models(创建模型)】命令，弹出【Create Models(创建模型)】对话框。将该对话框左侧的【VAR00001】移入【Dependent Variables(因变量)】列表框。在【Method(模型)】下拉列表框中选择【Expert Model】选项。



# Step03: 统计量的选择

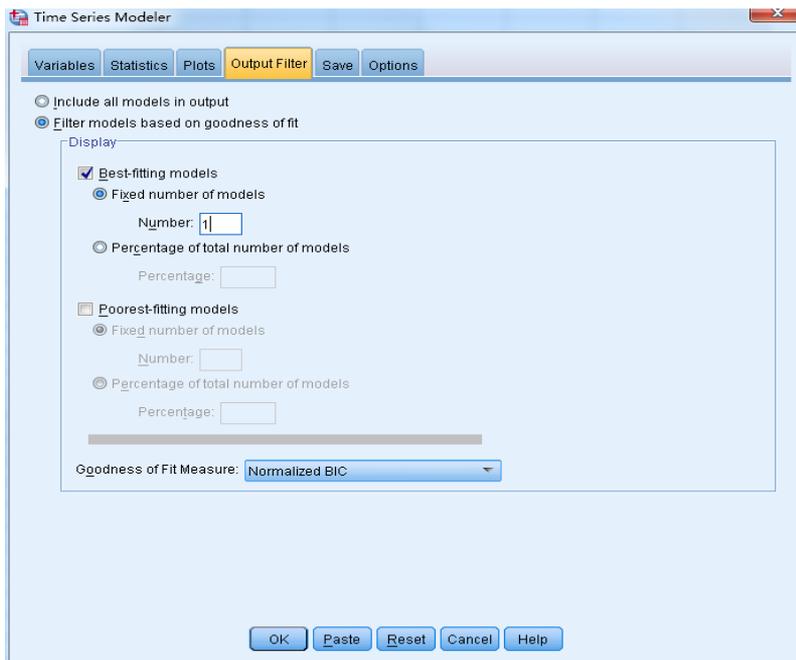
CONCEPT  
STRATE

单击【Statistics(统计量)】选项卡，选择展示模型拟合度量、Box-ljung 统计量、被模型过滤掉的样本数据的个数的选项，选择显示模型参数的估计值，选择好以后，单击【Output Filter(输出过滤)】选项卡。



CONCEPT  
STRATE

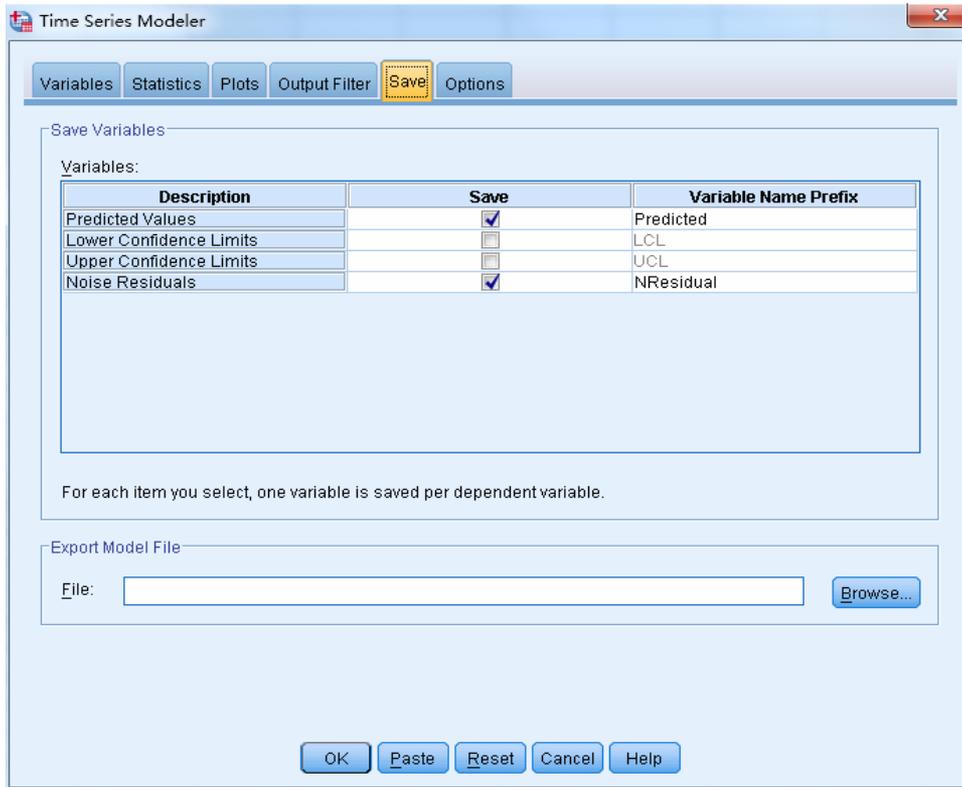
Step04: 在【Output Filter(输出过滤)】选项卡中选择【Filter models based on goodness fit】，输出拟合优度最好的那一个模型，选择拟合优度准则为标准BIC准则。选择好以后，单击【Save(保存)】选项卡。



CONCEPT  
STRATE

## Step05: 保持变量的选择

在 【Save (保存)】 选项卡中选择保存预测值，保存残差的值。





## Step06: 完成操作

单击【OK(确认)】，将进行ARIMA模型建模,完成操作。此时，输出结果，同时在当前数据编辑窗口中自动生成带前缀Predicted的预测值和带前缀NResidual的残差的值。

VAR00001	YEAR_	MONTH_	DATE_	Predicted_VA R00001_Mod el_1	NResidual_V AR00001_Mo del_1	Predicted_VAR00 001_Model_1_A	NResidual_VAR00001_Mod el_1_A
3029.30	2000	10	OCT 2000	.	.	.	.
3107.80	2000	11	NOV 2000	.	.	.	.
3680.10	2000	12	DEC 2000	.	.	.	.
3127.20	2001	1	JAN 2001	.	.	.	.
3001.80	2001	2	FEB 2001	2971.10	30.70	2964.40	.01
2876.10	2001	3	MAR 2001	2796.34	79.76	2783.13	.03
2820.90	2001	4	APR 2001	2792.09	28.81	2790.38	.01
2929.60	2001	5	MAY 2001	2875.73	53.87	2891.91	.01
2908.70	2001	6	JUN 2001	2918.07	-9.37	2927.75	-.01
2851.40	2001	7	JUL 2001	2863.87	-12.47	2866.10	.00
2889.40	2001	8	AUG 2001	2895.39	-5.99	2897.00	.00
3136.90	2001	9	SEP 2001	3109.60	27.30	3131.27	.00
3347.30	2001	10	OCT 2001	3301.84	45.46	3327.24	.01
3421.70	2001	11	NOV 2001	3409.06	12.64	3429.69	.00
4033.30	2001	12	DEC 2001	3989.35	43.95	4056.02	-.01
3552.20	2002	1	JAN 2002	3472.80	79.40	3438.38	.03
3416.10	2002	2	FEB 2002	3389.39	26.71	3361.57	.02
3197.40	2002	3	MAR 2002	3252.69	-55.29	3230.98	-.01
3163.30	2002	4	APR 2002	3162.56	.74	3147.32	.01
3320.50	2002	5	MAY 2002	3254.29	66.21	3260.43	.02
3302.80	2002	6	JUN 2002	3286.96	15.84	3294.11	.00



### 3 实例结果及分析

#### (1) 模型描述

该模型为 Model\_1，模型的类型为 Winters' 加法性模型。

表 11-13 模型描述

			Model Type
Model ID	VAR00001	Model_1	Winters' Additive

#### (2) 模型拟合优度

对VAR00001建立ARIMA(1, 1, 0)(0, 1, 1)模型的拟合优度，包括了调整R-Square, 标准化的BIC等所有拟合优度的值。



### (3) 模型的统计量的结果

由于在【Statistics(统计量)】对话框中，选择了展示模型拟合度量、Ljung-Box统计量、被模型过滤掉的样本数据的个数的选项，所以，在输出结果中出现了调整R-Square, 标准化的BIC的值, Ljung-Box统计量的值。

•从表11-15中可以看出标准BIC值为8.160，比 ARIMA(1, 1, 1)(1, 1, 1)模型的标准BIC值小一些。

表 11-15 模型的拟合优度

Model	Number of Predictors	Model Fit statistics		Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Normalized BIC	Statistics	DF	Sig.	
VAR00001-Model_1	0	.747	8.160	9.526	15	.848	0

a. Best-Fitting Models according to Normalized BIC (smaller values indicate better fit).

由于在【Statistics(统计量)】对话框中，选择显示模型参数的估计值，所以，在输出结果中出现模型的参数估计的结果。

表 11-16 ARIMA 模型参数估计<sup>a</sup>

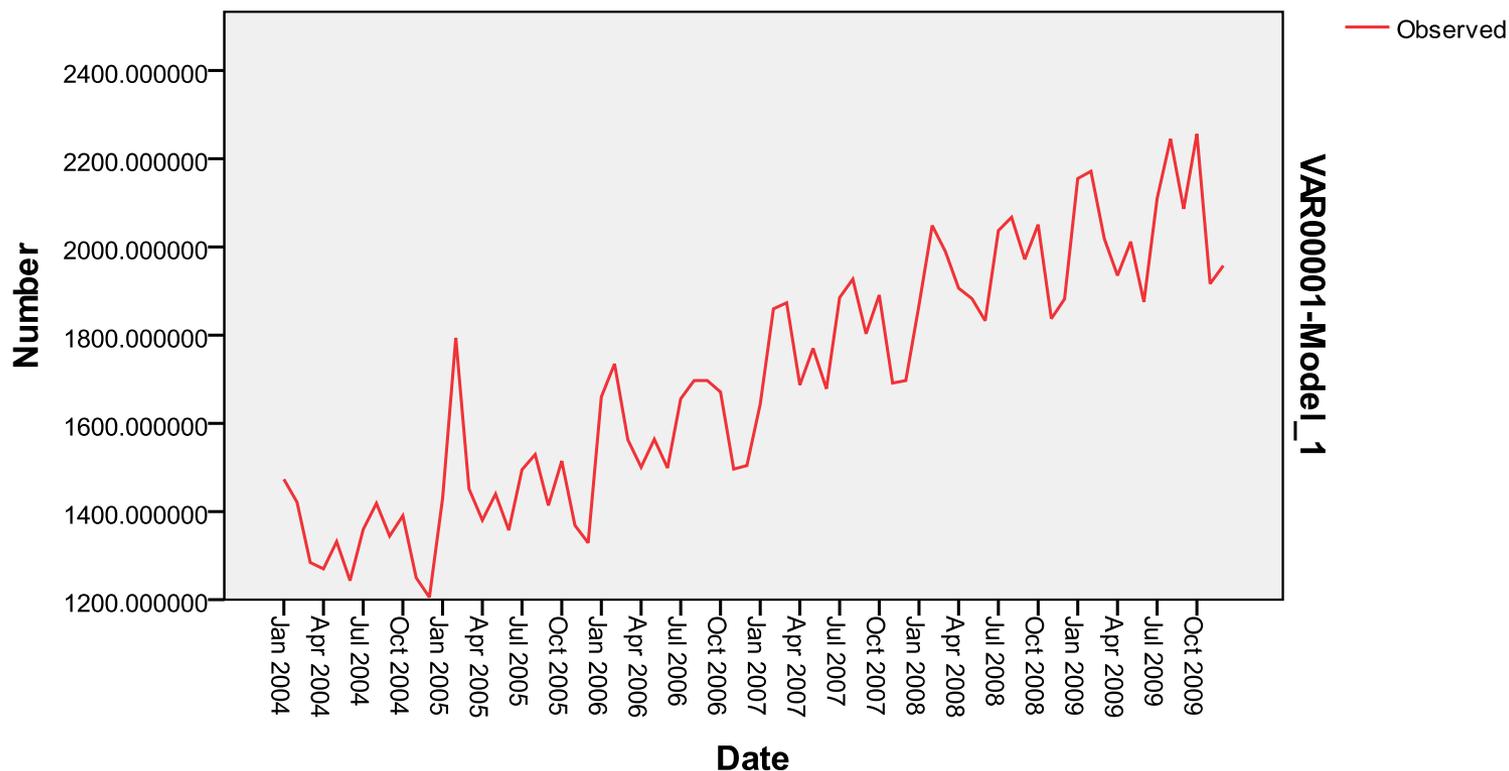
Model <sup>a</sup>			Estimate <sup>a</sup>	SE <sup>a</sup>	t <sup>a</sup>	Sig. <sup>a</sup>
VAR00	No	Alpha (Level) <sup>a</sup>	.091 <sup>a</sup>	.049 <sup>a</sup>	1.877 <sup>a</sup>	.065 <sup>a</sup>
001-Mo	Transformation <sup>a</sup>	Gamma (Trend) <sup>a</sup>	.000 <sup>a</sup>	.003 <sup>a</sup>	.131 <sup>a</sup>	.896 <sup>a</sup>
del_1 <sup>a</sup>		Delta (Season) <sup>a</sup>	.001 <sup>a</sup>	.095 <sup>a</sup>	.011 <sup>a</sup>	.992 <sup>a</sup>

<sup>a</sup> Best-Fitting Models according to Normalized BIC (smaller values indicate better fit).<sup>a</sup>

从表10-12可以看出，对原始数据建立的Winters' 加法模型，Alpha的参数估计值是0.091，该模型是最优的模型，对数据的分析比较的恰当。



# (4) 模型的拟合图





# 第12章 SPSS在市场调研中的应用

## 12.1 实例提出：绿色食品的认知研究

CONCEPT  
STRATE

二十一世纪以来，资源和环境问题日益受到人们的关注。在现代化进程中，人类过度的经济活动，给资源和环境带来很大压力。与此同时，农药残留和食品安全事件频繁发生，严重影响了人们的身体健康和生活质量。

数据10-1.sav是对应的调查数据。请利用这些资料分析以下问题：

问题一：请你将被调查者的基本信息作简要统计说明。

问题二：请分析性别、收入等因素对消费者在购买绿色食品顾虑上有无差异性。

## 12.2 实例的SPSS软件操作详解

### 1 问题一操作详解

问题一要求你将被调查者的基本信息作简要统计说明。由于问卷所给的调查信息中，被调查者的性别、年龄、受教育程度等都是调查者的基本信息。因此可以首先对这些变量进行描述性统计分析，绘制频数表和相关图形。

同时，可以采用列联表分析来研究不同基本信息之间的相互影响。

具体操作步骤如下：

Step01：打开数据文件

打开数据文件12-1.sav。同时单击数据浏览窗口的【Variable View】按钮，检查各个变量的数据结构定义是否合理，是否需要修改调整。

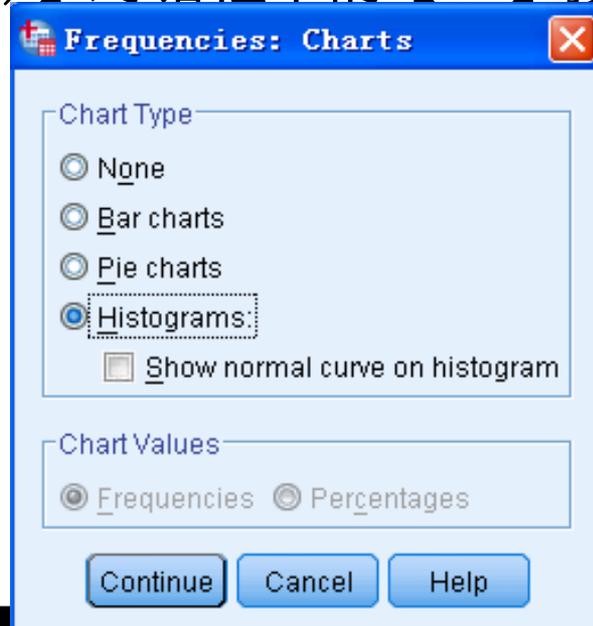
## Step02: 调查者基本信息的频数分析

选择菜单栏中的【Analyze（分析）】→【Descriptive Statistics（描述统计）】→【Frequencies（频率）】命令，弹出【Frequencies（频率）】对话框。选择A1—A8等8项指标作为分析对象，将其添加至右侧的【Frequencies（变量）】列表框中。



### Step03: 绘制直方图

单击【Charts】按钮，弹出【Frequencies:Charts(频率: 图表)】对话框。在图形类型【Chart Type(图表类型)】中，点选直方图【Histograms(直方图)】单选钮，并勾选其下的【With normal curve(显示正态曲线)】复选框。再单击【Continue】按钮，返回主菜单。最后单击【Frequencies(频率)】对话框中的【OK】按钮，完成本部分操作。



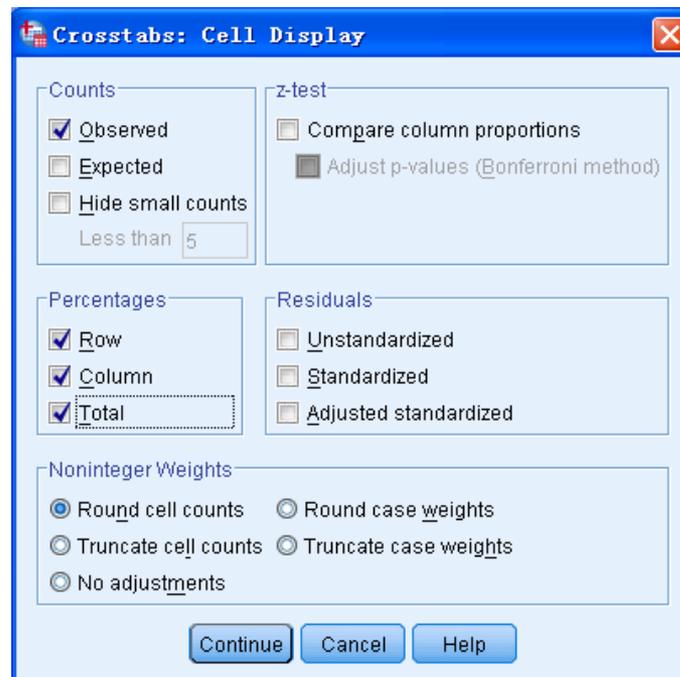
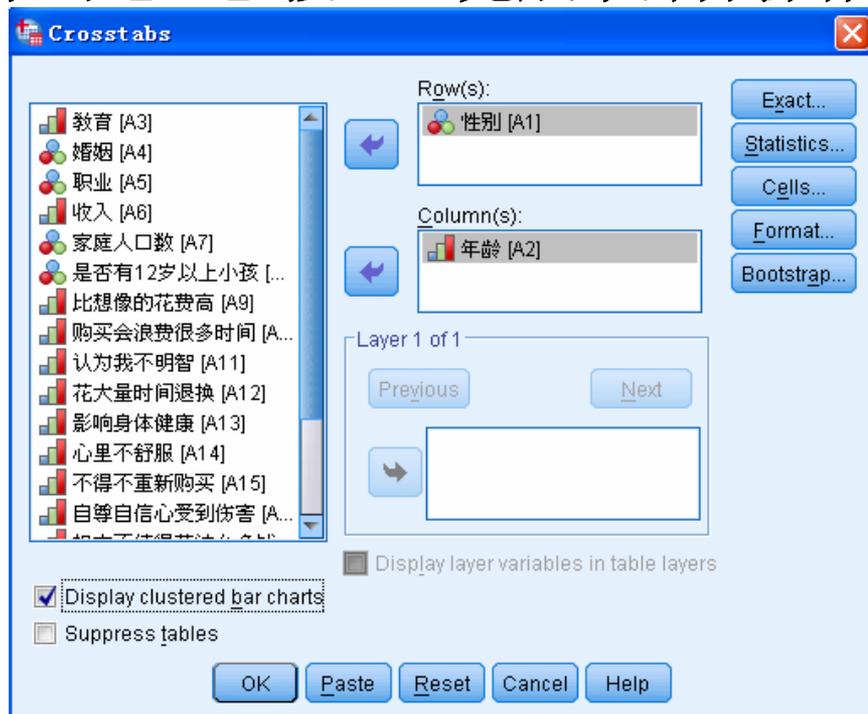
## Step04: 列联表分析

选择菜单栏中的【Analyze (分析)】→【Descriptive Statistics (描述统计)】→【Crosstabs (交叉表)】命令，弹出【Crosstabs (交叉表)】对话框。在候选变量列表框中选择A1 (性别) 变量添加至主对话框右侧的【Row(s) (行)】列表框中，选择A2 (年龄) 变量添加至右侧的【Column(s) (列)】列表框中 (这里只分析性别与年龄之间的关系，其他变量的关系可以类似求的)。勾选【Display clustered bar charts (显示复式条形图)】复选框，显示列联表柱状图。

单击【Cells】按钮，弹出【Crosstabs: Cell Display (交叉表: 单元显示)】对话框。在【Counts (计数)】选项组中勾选【Observed (观察值)】复选框；在【Percentages (百分比)】选项组中勾选【Row (列)】、【Column (行)】、【Total (总计)】复选框；在【Noninteger Weights (非整数权重)】选项组中点选【Round cell counts (四舍五入单元格计数)】单选钮。

再单击【Continue按钮】钮，返回主对话框。

单击【OK】按钮，完成本部分操作。



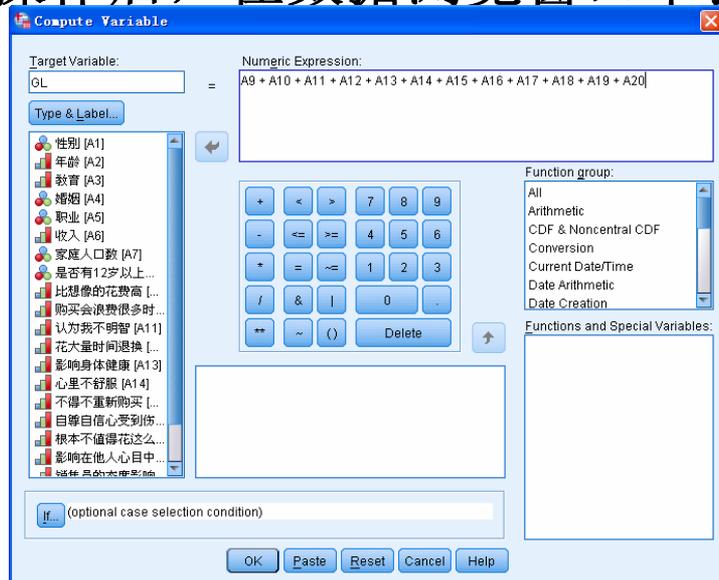
## 2 问题二操作详解

问题二要分析性别、教育程度等因素对消费者在购买绿色食品顾虑上有无差异性。表12-1中的第9问—第20问都是反映消费者的购买顾虑，每个题目取值越大说明消费者在该方面的顾虑越重。由于我们要综合考虑消费顾虑，于是将每个被调查者从第9问到第20问的得分相加，就可以得到综合顾虑值；然后通过单因素方差分析来分析性别、教育程度等因素对消费顾虑有无显著性影响。

具体操作步骤如下：

Step01：计算被调查者购买综合顾虑值

打开数据文件12-1.sav，接着选择菜单栏中的【File(文件)】→【Transform(转换)】→【Compute(计算)】命令。在弹出的【Compute(计算)】对话框的【Target Variable(目标变量)】文本框中，输入变量名GL表示要新建此变量来表示购买综合顾虑值。接着在【Numeric Expression(数学表达式)】文本框中输入综合顾虑值的计算公式。完成上述操作后，在数据浏览窗口中会新增变量“GL”



## Step02: 性别对购买顾虑的差异性研究

接着利用单因素方差分析来研究性别变量对消费者购买顾虑有无显著性差异。选择菜单栏中的【Analyze (分析)】→【Compare Means (比较均值)】→【One-Way ANOVA (单因素ANOVA)】命令，弹出【One-Way ANOVA (单因素ANOVA)】对话框。在候选变量列表框中选择“GL”变量作为因变量，将其添加至【Dependent List (因变量列表)】列表框中，同时也在【候选变量】列表框中选择“A1”变量作为水平值，将其添加至【Factor (因子)】列表框中。接着选择【Options (选项)】对话框中的【Homogeneity-of-variance】选项，表示输出方差齐性检验表。最后单击主对话框中的【OK】按钮，完成操作。



### Step03: 收入对购买顾虑的差异性研究

同样，也首先考虑利用单因素方差分析来研究收入程度对消费者购买顾虑有无显著性差异。但是在对其做方差齐性检验时，发现不同收入水平下方差不具备齐性的条件。于是可以考虑采用非参数检验中的多独立样本均值检验方法。

选择菜单栏中的【Analyze（分析）】→【Nonparametric Tests（非参数检验）】→【Legacy Dialogs（旧对话框）】→【K Independent Samples（k个相关样本）】命令，弹出【Tests for Several Independent Samples（多个关联样本检验）】对话框。在【候选变量】列表框中选择“GL”变量作为检验变量，将其添加至【Test Variable List（检验变量列表）】列表框中。选择分组变量A6（收入）将其添加至【Grouping Variable(s)（组变量）】列表框中。单击【Grouping Range】按钮，弹出相应对话框。在【Minimum（最小值）】文本框中输入1，在【Maximum（最大值）】文本框中输入5。最后单击主对话框中的【OK】按钮，完成操作。



## 12.3 实例的SPSS输出结果详解

CONCEPT  
TRATE

### 1 问题一输出结果详解

#### (1) 频数分析表

首先表12-2显示了性别、年龄等八项基本信息指标的基本统计情况，其中“Valid”列表示有效样本数目，“Missing”列表示缺失样本数目。例如，教育变量的有效样本数目为306，而仅有2个样本缺失。

表 12-2 基本统计表

	N	
	Valid	Missing
性别	308	0
年龄	308	0
教育	306	2
婚姻	306	2
职业	307	1
收入	296	12
家庭人口数	302	6
是否有 12 岁以上小孩	303	5

接着，软件输出了这八项指标的频数分析结果。从结果看到，所有调查者中57.8%为男性，其余为女性；所有调查者中21-30岁人群所占比重最大，达到了45.1%，而41-50岁的调查者最少，只有5.8%。

表 12-3 性别变量频数分析表

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	男	178	57.8	57.8	57.8
	女	130	42.2	42.2	100.0
	Total	308	100.0	100.0	

表 12-4 年龄变量频数分析表

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	20 岁以下	67	21.8	21.8	21.8
	21-30 岁	139	45.1	45.1	66.9
	31-40 岁	56	18.2	18.2	85.1
	41-50 岁	18	5.8	5.8	90.9
	51 岁及以上	28	9.1	9.1	100.0
	Total	308	100.0	100.0	

## (2) 直方图

图12-8和图12-9分别是性别和年龄变量的直方图。从图形的高低可以明显看到不同性别和年龄调查者数量的差异性。

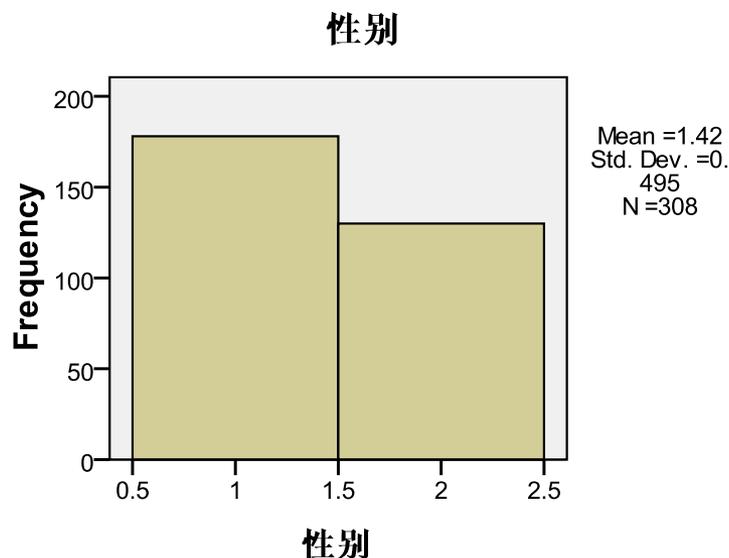


图12-8 性别变量直方图

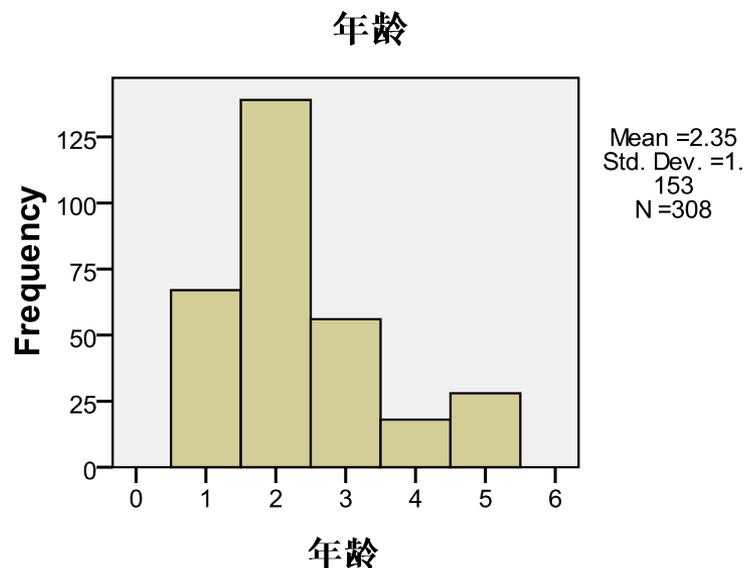


图12-9 年龄变量直方图

### (3) 列联表分析

表12-5是“性别”变量和“年龄”变量的列联表。行变量是“年龄”变量，列变量是“性别”变量。可以看到，总共178位男性调查者中，年龄在“20岁以下”的共有27人，“21-30岁”的有88人，依次类推。对比行分比、列百分比和合计百分比看到，男性中约一半的调查者年龄都介于21-30岁之间，而对于女性调查者来说，“20岁以下”和“21-30岁”所占比例最高，达到了30.8%和39.2%。

最后，从图12-10的条图也可以明显看到不同性别下各个年龄阶段的被调查人总数。

表 12-5 “性别\*年龄”列联表

		年龄					Total	
		20岁以下	21-30岁	31-40岁	41-50岁	51岁及以上		
性别	男	Count	27	88	33	12	18	178
		% within 性别	15.2%	49.4%	18.5%	6.7%	10.1%	100.0%
		% within 年龄	40.3%	63.3%	58.9%	66.7%	64.3%	57.8%
		% of Total	8.8%	28.6%	10.7%	3.9%	5.8%	57.8%
	女	Count	40	51	23	6	10	130
		% within 性别	30.8%	39.2%	17.7%	4.6%	7.7%	100.0%
		% within 年龄	59.7%	36.7%	41.1%	33.3%	35.7%	42.2%
		% of Total	13.0%	16.6%	7.5%	1.9%	3.2%	42.2%
Total		Count	67	139	56	18	28	308
		% within 性别	21.8%	45.1%	18.2%	5.8%	9.1%	100.0%
		% within 年龄	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%
		% of Total	21.8%	45.1%	18.2%	5.8%	9.1%	100.0%

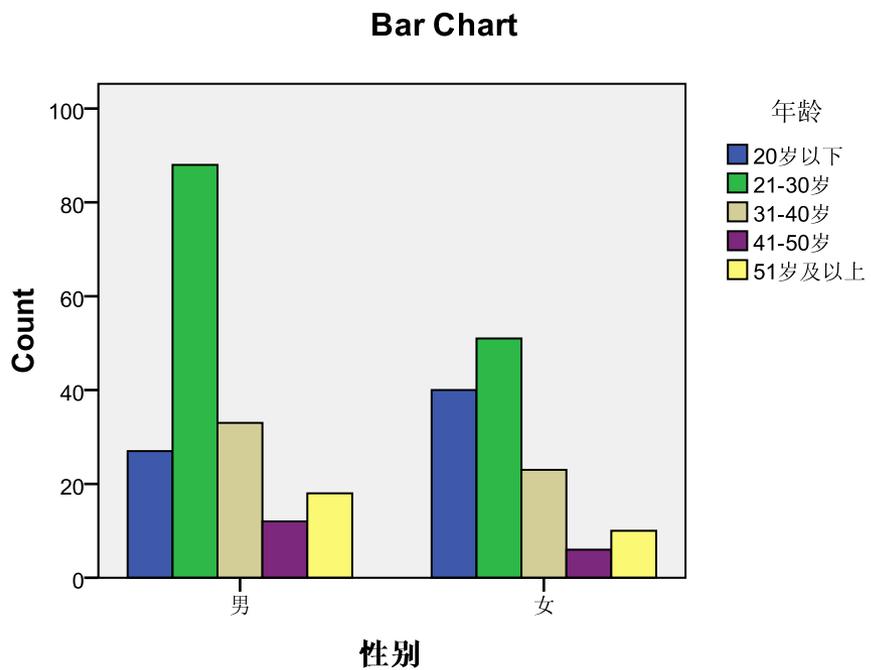


图12-10 性别和年龄条图

## 2. 问题二输出结果详解

### 一、性别因素对购买顾虑的差异性影响。

#### (1) 方差齐性检验

SPSS的结果报告中首先列出了方差齐性检验结果表12-6。由于这里采用的是Levene检验法，故表格首先显示Levene统计量等于0.006。由于概率P值0.937明显大于显著性水平，故认为不同性别下的购买顾虑值的方差是相同的，满足方差分析的前提条件。

表 12-6 方差齐性检验结果

Levene Statistic	df1	df2	Sig.
.006	1	295	.937

## (2) 单因素方差分析表

表12-7是方差分析表结果表。可以看到组间离差平方和为114.470，组内离差平方和为20190.076，总离差平方和为20304.545。方差分析F检验量等于1.673，F值对应的概率P值等于0.197。由于概率P值大于显著性水平，故认为性别因素对购买顾虑没有造成显著性影响。

表 12-7 方差分析检验表

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	114.470	1	114.470	1.673	.197
Within Groups	20190.076	295	68.441		
Total	20304.545	296			

## 二、收入因素对购买顾虑的差异性影响。

### (1) 方差齐性检验

表12-8是方差齐性检验结果表。表格显示Levene统计量等于2.495。由于概率P值0.043小于显著性水平0.05，故认为这五种收入水平下购买顾虑值的方差是不同的，故不能采用方差分析。

表 12-8 方差齐性检验结果表

Levene Statistic	df1	df2	Sig.
2.495	4	280	.043

## (2) 秩统计表

表12-9是多独立样本非参数检验的秩统计表。“800元以下”的平均秩为157.71，依次类推。比较平均秩大小看到，这五种收入水平的购买顾虑值差异较大。

表 12-9 秩统计表

	收入	N	Mean Rank
GL	800 元以下	47	157.71
	801-1500 元	117	137.53
	1501-3000 元	86	123.94
	3001-5000 元	26	178.54
	5001 元以上	9	216.67
	Total	285	

### (3) 非参数检验结果表

表12-10看到，Kruskal-Wallis H检验结果的Chi-Square 统计量等于18.669，自由度df等于4，近似相伴概率P值为0.001，小于显著性水平0.05。所以拒绝零假设，认为这五种收入水平下的购买顾虑值有显著差异。

表 12-10 非参数检验结果表<sup>+</sup>

<sup>+</sup>	GL <sup>+</sup>	<sup>+</sup>
Chi-Square <sup>+</sup>	18.669 <sup>+</sup>	
df <sup>+</sup>	4 <sup>+</sup>	
Asymp. Sig. <sup>+</sup>	.001 <sup>+</sup>	



# 第13章 SPSS在系统预测 中的应用

## 13.1 实例提出：汽车保有量的预测分析

CONCEPT  
STRATE

我国经济的快速发展为私人汽车提供了巨大的发展空间。据中国汽车工业协会估算，截止到2006年底，中国私人汽车保有量约为2650万辆，占全国汽车保有量的60%左右。在2006年，我国汽车销量为710多万辆，私人购买比例超过77%，中国已经成为仅次于美国的全球第二大新车市场。

据世界银行的研究，汽车保有量（尤其是私人汽车）与人均国民收入成正比。2003年，我国国内人均GDP首次突破1000美元，这预示着中国汽车开始进入家庭消费阶段。而事实表明，随着中国人均GDP的稳健增长，近年来，我国的家用汽车销量以两位数的增速急剧扩大。

汽车特别是用于消费的私人汽车保有量的多少，与经济发展程度、居民收入以及道路建设等有着密切的联系。随着私人汽车消费时代的到来，汽车保有量上升的一个重要因素就是国内汽车消费的快速增长。消费者购买力的增强和个体私营经济的快速发展，也带动了私人汽车的大发展。私人汽车保有量与一个国家或地区的社会经济发展的有关数据有着密切关系。附表13-1提供了我国某一经济发达地区的一些相关统计数据。

请根据附表中的相关数据分析影响该地区私人汽车保有量的因素，并预测到2008年该地区私人汽车保有量有多少？

表 13-1 1996-2008 年某地区相关的统计数据

年份	人均国内生产总值 (元)	全社会消费品零售总额 (亿元)	全社会固定资产投资总额 (亿元)	道路总长 (公里)	居民人均可支配收入 (元)	居民储蓄款余额 (亿元)	私人汽车保有量 (万辆)
1996	27000	297.35	327.53	737	16316	583.89	3.1
1997	30619	325.00	390.51	789	18600	707.67	3.6
1998	33282	423.00	474.63	894	19886	861.88	4.2
1999	33689	467.57	569.55	1015	20249	941.99	4.8
2000	41020	538.17	616.25	1198	21626	1082.6	6.7
2001	43344	832.04	686.37	1361	23544	1373.4	9.1
2002	46030	941.94	788.15	1710	24941	1756.5	13
2003	53887	1095.13	969.1	2100	25936	2199.5	18.9
2004	59271	1250.64	1092.6	2314	26596	2625.4	29
2005	64507	1437.67	1176.1	2500	28494	3229.4	51.1
2006	70597	1671.29	1273.7	2614	29628	3744.7	78.2
2007	79221	1905.03	1345	2897	30063	3792.6	96.1

## 12.2 实例的SPSS软件操作详解

CONCEPT  
TRATE

本实例要求分析人均国内生产总值、全社会消费品零售总额等因素对私人汽车保有量的影响情况，并做相应的预测分析。由于人均国内生产总值等指标都描述了国民经济的发展情况，因此它们和私人汽车保有量有着密切的关系。但是要全部考虑这些指标是非常困难的，这是因为指标数量众多，难以全部考虑。考虑到这些经济指标之间还存在着相关性，因此可以利用因子分析，利用降维的方法综合这些众多的经济指标为少量的公共因子，这样通过分析因子和私人汽车保有量的关系来预测私人汽车保有量的未来趋势。

具体操作步骤如下：

### Step01: 打开数据文件

打开数据文件13-1. sav。同时单击数据浏览窗口的【Variable View(变量视图)】按钮，检查各个变量的数据结构定义是否合理，是否需要修改调整。

### Step02: 因子分析

在候选变量列表框中选择X1、X2、...X6变量设定为因子分析变量，将其添加至【Variables(变量)】列表框中。单击【Extraction】按钮，勾选【Scree plot】复选框，其他选项保持系统默认，单击【Continue(继续)】按钮返回主对话框。

在主对话框中，单击【Score(尺度)】按钮，勾选【Save as variables(保存变量)】复选框，表示采用回归法计算因子得分并保持在原文件中；同时勾选【Display factor score coefficient matrix(显示因子得分系数矩阵)】复选框，表示输出因子得分系数矩阵。其他选项保持系统默认，单击【Continue(继续)】按钮返回。

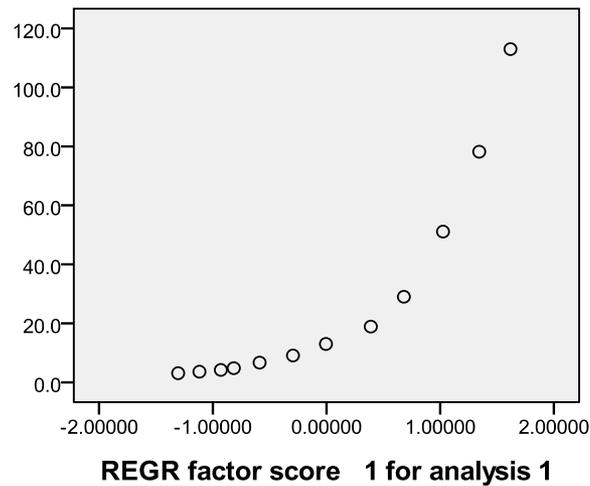
单击【OK】按钮，完成本步操作。



### Step03: 回归分析

在第二步因子分析中得到了不同年份的因子得分，这些因子得分充分反映了不同年份的综合经济发展值。于是可以考虑利用它来对私车保有量进行预测。

首先绘制综合经济发展值与私车保有量的散点图，判断两者之间是否存在相关关系。



选择菜单栏中的【Analyze(分析)】→【Regression(回归)】→【Curve Estimation(曲线估计)】命令，弹出【Curve Estimation(曲线估计)】对话框。在候选变量列表框中选择“y”变量设定为因变量，将其添加至【Dependent(s)(因变量)】列表框中。同时，选择“FAC1\_1”变量设定为自变量，将其添加至【Variable(变量)】列表框中。在【Model(模型)】复选框中勾选【Logistic】模型，并在【Upper bound(上限)】文本框中填入1000，表示上限值。注意，这里的上限值可以调整完善。最后单击【OK】按钮，完成操作。



**Curve Estimation** [Close]

**Dependent(s):** [Save...]  
私人汽车保有量 [Y]

**Independent**  
 Variable: REGR factor score ...  
 Time

**Case Labels:** [ ]  
 Include constant in equation  
 Plot models

**Models**  
 Linear     Quadratic     Compound     Growth  
 Logarithmic     Cubic     S     Exponential  
 Inverse     Power:     Logistic  
Upper bound: 1000

Display ANOVA table

[OK] [Paste] [Reset] [Cancel] [Help]

#### Step04: 预测私车保有量

在得到了私车保有量和综合经济发展值的模型后，要预测私车保有量在2010的数量，则首先需要估计综合经济发展值在2008的取值，利用线性回归模型得到经济发展值的预测值后，带入Logistic函数就可以估计出2008年该地区的私车保有量了。

## 12.3 实例的SPSS输出结果详解

CONCEPT  
RATE

### 1 因子分析结果

#### (1) 因子分析共同度

表13-2是因子分析的共同度，显示了所有变量的共同度数据。第一列是因子分析初始解下的变量共同度。它表明，对原有六个变量如果采用主成分分析法提取所有六个特征根，那么原有变量的所有方差都可被解释，变量的共同度均为1。第二列列出了按指定提取条件提取特征根时的共同度。可以看到，所有变量的绝大部分信息可被因子解释，这些变量信息丢失较少。

表 13-2 因子分析共同度

	Initial	Extraction
人均国内生产总值	1.000	.991
全社会消费品零售总额	1.000	.990
全社会固定资产投资总额	1.000	.994
道路总长	1.000	.991
居民人均可支配收入	1.000	.970
居民储蓄款余额	1.000	.984

## (2) 因子分析的总方差解释

接着计算得到相关系数矩阵的特征值、方差贡献率及累计方差贡献率结果如表13-3所示。在表13-3中，第一个因子的特征根值为5.920，解释了原有6个变量总方差的98.670%。因此只选取第一个因子为主因子即可。

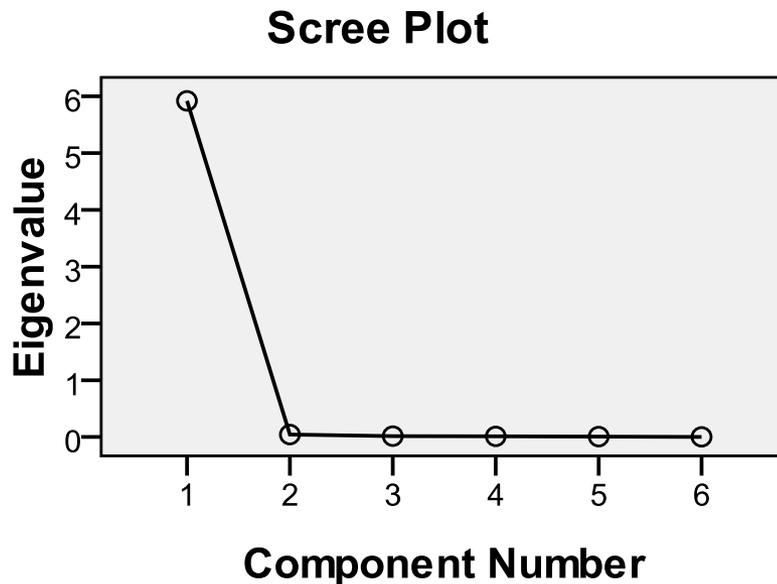
表 13-3 因子分析的总方差解释

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.920	98.670	98.670	5.920	98.670	98.670
2	.043	.719	99.389			
3	.016	.267	99.656			
4	.011	.186	99.842			
5	.007	.119	99.961			
6	.002	.039	100.000			

Extraction Method: Principal Component Analysis.

### (3) 因子碎石图

从碎石图看到，第一个特征值明显大于后面的特征值，说明提取第一个因子是合适的。



#### (4) 因子载荷矩阵

表13-4显示了因子载荷矩阵。通过载荷系数大小可以分析不同公共因子所反映的主要指标的区别。从结果看，第一主因子在这六个指标上的载荷值都很大，说明它综合反映了该地区综合经济发展值，故可以作为综合经济发展值看待。

表 13-4 因子载荷矩阵<sup>①</sup>

↕	Component <sup>①</sup>
	1 <sup>①</sup>
人均国内生产总值 <sup>①</sup>	.995
全社会消费品零售总额 <sup>①</sup>	.995
全社会固定资产投资总额 <sup>①</sup>	.997
道路总长 <sup>①</sup>	.996
居民人均可支配收入 <sup>①</sup>	.985
居民储蓄款余额 <sup>①</sup>	.992

## (5) 因子得分系数

表13-5列出了采用回归法估计的因子得分系数。同时在原数据浏览窗口中新增了变量“FAC1\_1”，它表示不同年份的综合经济发展值。

表 13-5 因子得分系数

	Component
	1
人均国内生产总值	.168
全社会消费品零售总额	.168
全社会固定资产投资总额	.168
道路总长	.168
居民人均可支配收入	.166
居民储蓄款余额	.168

## 2 曲线估计结果

### (1) 模型描述

表13-6是SPSS对曲线拟合结果的初步描述统计，例如自变量和因变量、估计方程的类型等。

表 13-6 模型描述

Model Name		MOD_15
Dependent Variable	1	私人汽车保有量
Equation	1	Logistic <sup>a,b</sup>
Independent Variable		REGR factor score 1 for analysis 1
Constant		Included
Variable Whose Values Label Observations in Plots		年份
a. The model requires all non-missing values to be positive.		
b. For all dependent variables, the theoretical upper bound is set to 1000.		

## (2) 模型汇总及参数估计

表13-7给出了样本数据进行Logistic回归的检验统计量和相应方程中的参数估计值。

模型的整体拟合优度值R<sup>2</sup>为0.994，F统计量等于1624.416，概率P值小于显著性水平0.05，说明该模型有统计学意义。得到估计方程如下：

$$y = \frac{1}{1/1000 + 0.072 \cdot 0.287^x}$$

表 13-7 模型汇总及参数估计

Equation	Model Summary					Parameter Estimates	
	R Square	F	df1	df2	Sig.	Constant	b1
Logistic	.995	2068.400	1	10	.000	.072	.287

### (3) 拟合曲线图

最后给出的是实际数据的散点图和Logistic回归方程的预测图，这进一步说明方程的拟合效果最好。

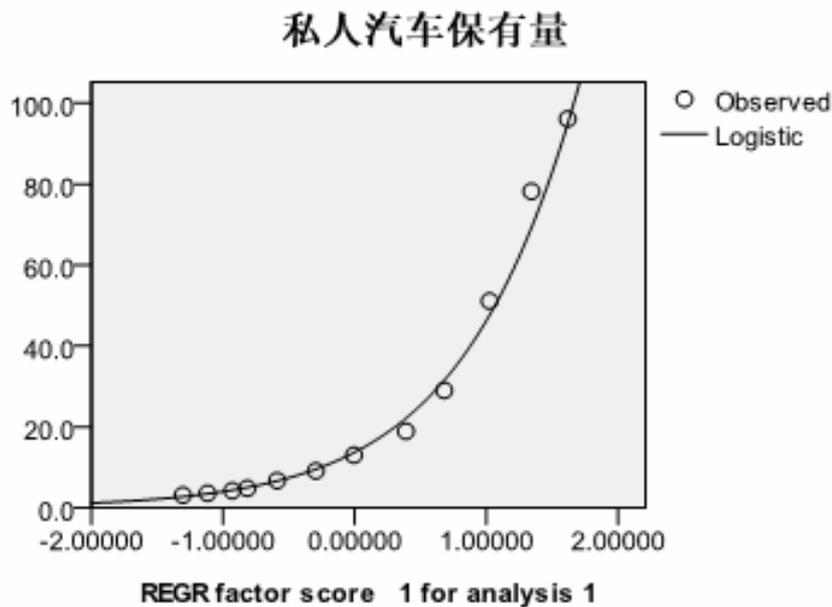


图 13-5 拟合图

### 3 预测私家车保有量

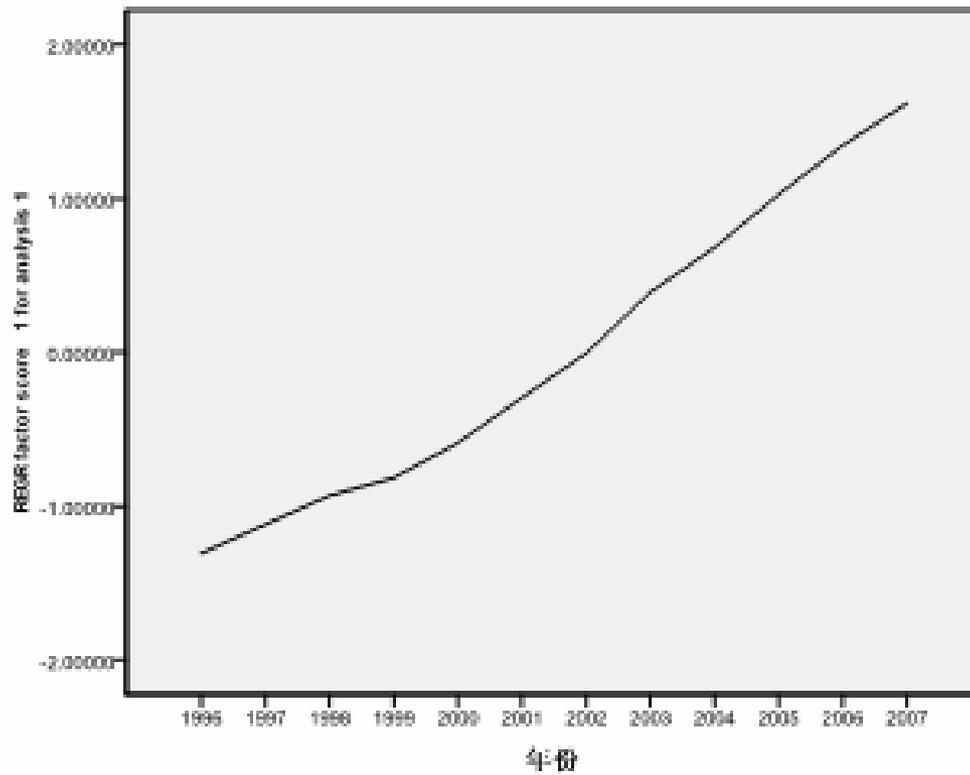
#### (1) 预测综合经济发展值

绘制综合经济发展变量“FAC1\_1”的时间序列图13-6看到，该变量的增长基本呈线性趋势，于是可以采用线性回归来估计该变量在2008年的取值。

采用曲线估计中的线性回归选项可以估计得到预测方程如下：

$$\text{FAC1\_1} = -1.788 + 0.275 * t$$

于是计算得到综合经济发展变量在2008的取值为1.787



## (2) 预测私车保有量

在计算得到综合经济发展变量的预测值后，带入Logistic回归模型，于是得到该地区在2008的私车保有量为：

$$y = \frac{1}{1/1000 + 0.072 \cdot 0.287^{1.787}} \approx 114.5 \text{万辆}$$



# 第14章 SPSS在社会学中的应用

# 14.1 实例提出：中国传统文化了解程度研究

CONCEPT  
STRATE

某大学研究机构对该校电气、管理、电信、外语、人文几个学院的同学进行了调查，各个学院发放的问卷数参照各个学院的人数比例。总共发放问卷250余份，回收有效问卷228份。

调查问卷设置了大学生对传统文化了解程度的题目，例如“佛教的来源是什么？”、“儒家的思想核心是什么？”、“《清明上河图》的作者是谁？”等。由于篇幅有限，数据文件14-1.sav给出了每位调查者对传统文化了解程度的总得分，同时也列出了被调查者性别、专业、年级等数据信息。请利用这些资料，分析以下问题：

- 1、分析大学生对中国传统文化了解程度得分，并按了解程度对得分进行合理的分类。
- 2、研究获取文化来源对大学生传统文化了解程度是否存在影响关系。

## 14.2 实例的SPSS软件操作详解

CONCEPT  
STRATE

### 1. 问题一操作详解

对于问题一，首先可以采用描述性统计对被调查者的文化了解程度进行分析，了解大学生整体的传统文化了解程度；接着可以利用百分位数对了解程度得分进行分类，将其分为“不了解”、“不太了解”、“一般了解”、“较了解”和“很了解”等五类。

具体操作步骤如下：

## Step01: 打开数据文件



打开数据文件14-1.sav。同时单击数据浏览窗口的【Variable View（变量视图）】按钮，检查各个变量的数据结构定义是否合理，是否需要修改调整。

## Step02: 频数分析



选择菜单栏中的【Analyze（分析）】→【Descriptive Statistics（描述统计）】→【Frequencies（频率）】命令，弹出【Frequencies（频率）】对话框。在此对话框左侧的候选变量列表框中选择“X9”变量，将其添加至【Variable(s)（变量）】列表框中，表示它是进行频数分析的变量。



单击【Statistics】按钮，在弹出的对话框的【Cut points for equal groups（割点相等组）】文本框中键入数字“5”，输出第20%、40%、60%和80%百分位数，即将数据按照题目要求分为等间隔的五类。接着，勾选【Std. Deviation（标准差）】、【Mean（均值）】等选项，表示输出了解程度得分的描述性统计量。再单击【Continue】按钮，返回【Frequencies（频率）】对话框。



单击【Charts】按钮，勾选【Histograms（直方图）】和【With normal curve（显示正态曲线）】复选框，即直方图中附带正态曲线。再单击【Continue】按钮，返回【Frequencies（频率）】对话框。

最后，单击【OK（确定）】按钮，操作完成。

## Step03: 了解程度分类

CONCEPT  
STRATE

在得到第20%、40%、60%和80%百分位数后，接着以它们为断点对得分数据进行分类，因此可以利用SPSS中的【Recode（编码）】功能来实现。

打开SPSS软件，在菜单栏中选择【File（文件）】→【Transform（转换）】→【Recode into Different Variable（重新编码为不同变量）】命令，弹出【Recode into Different Variable（重新编码为不同变量）】对话框。



**Recode into Different Variables**

Numeric Variable -> Output Variable:

X9 --> 等级

Output Variable

Name: 等级

Label:

Change

Old and New Values...

If... (optional case selection condition)

OK Paste Reset Cancel Help

- 专业 [X1]
- 年级 [X2]
- 性别 [X3]
- 中国传统文化在今天...
- 您的中国传统文化知...
- 您是否常看中国传统...
- 您是否常访问中国传...
- 您耐心的听完过一场...

在左侧的候选变量列表框中选择“X9”变量进入【Input Variable→Output Variable（输入变量→输出变量）】列表框，同时在【Output Variable（输出变量）】复选框中填写输出赋值变量名称“等级”。同时单击【Change】按钮进行赋值转换。

单击【Old and New Value按钮，弹出重编码规则设置对话框。接着按照等级转换赋值规则进行变量的重新赋值工作。设置完成后，单击【Continue（继续）】按钮返回主对话框

最后，单击【OK（确定）】按钮，操作完成。此时，原数据文件新增加了“天数”变量。

**Recode into Different Variables: Old and New Values**

**Old Value**

Value:

System-missing

System- or user-missing

Range:

  
 through  


Range, LOWEST through value:

Range, value through HIGHEST:

All other values

**New Value**

Value:

System-missing

Copy old value(s)

**Old --> New:**

Lowest thru 45.8 --> 1

45.8 thru 54 --> 2

54 thru 60.4 --> 3

60.4 thru 68 --> 4

68 thru Highest --> 5

Output variables are strings    Width:

Convert numeric strings to numbers ('5' -> 5)

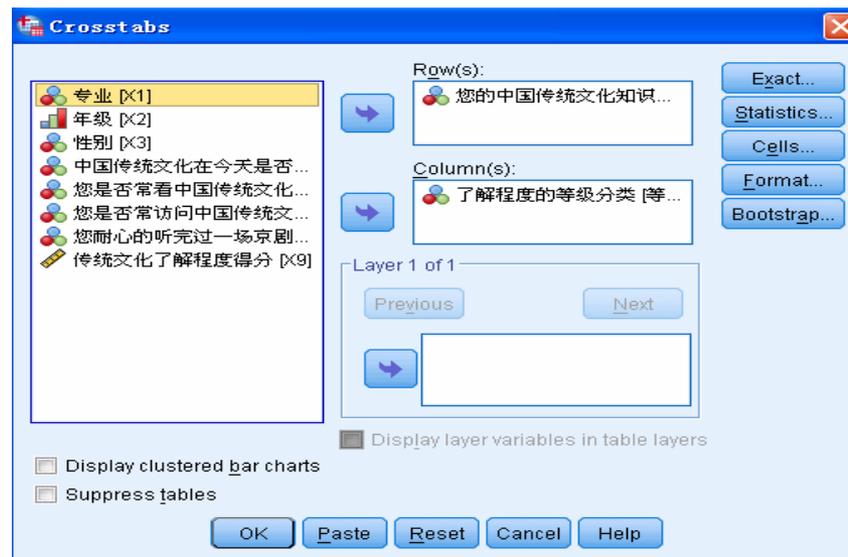


## 2. 问题二操作详解

对于问题二，大学生获取传统文化来源主要是从学校、家庭或自学等。因此本问题主要要分析不同学习途径对大学生传统文化了解程度是否存在显著性影响。由于文化来源途径和了解程度等级都是定性数据，因此可以考虑采用列联表分析中的行、列变量相关程度检验。

具体操作步骤如下：

选择菜单栏中的【Analyze（分析）】→【Descriptive Statistics（描述统计）】→【Crosstabs（交叉表）】命令，弹出【Crosstabs（交叉表）】对话框。



## 14.3 实例的SPSS输出结果详解

CONCEPT  
STRATE

### 1. 问题一结果

#### (1) 描述性统计量表

表14-1是被调查者对中国传统文化了解程度得分的描述性统计量输出表，其中包括了均值、中位数、方差等基本统计量。可以看到，大学生对传统文化了解程度得分均值等于57.18分，标准差为12.824，偏度为-0.116，峰度为-0.278等。

表 14-1 了解程度得分统计表

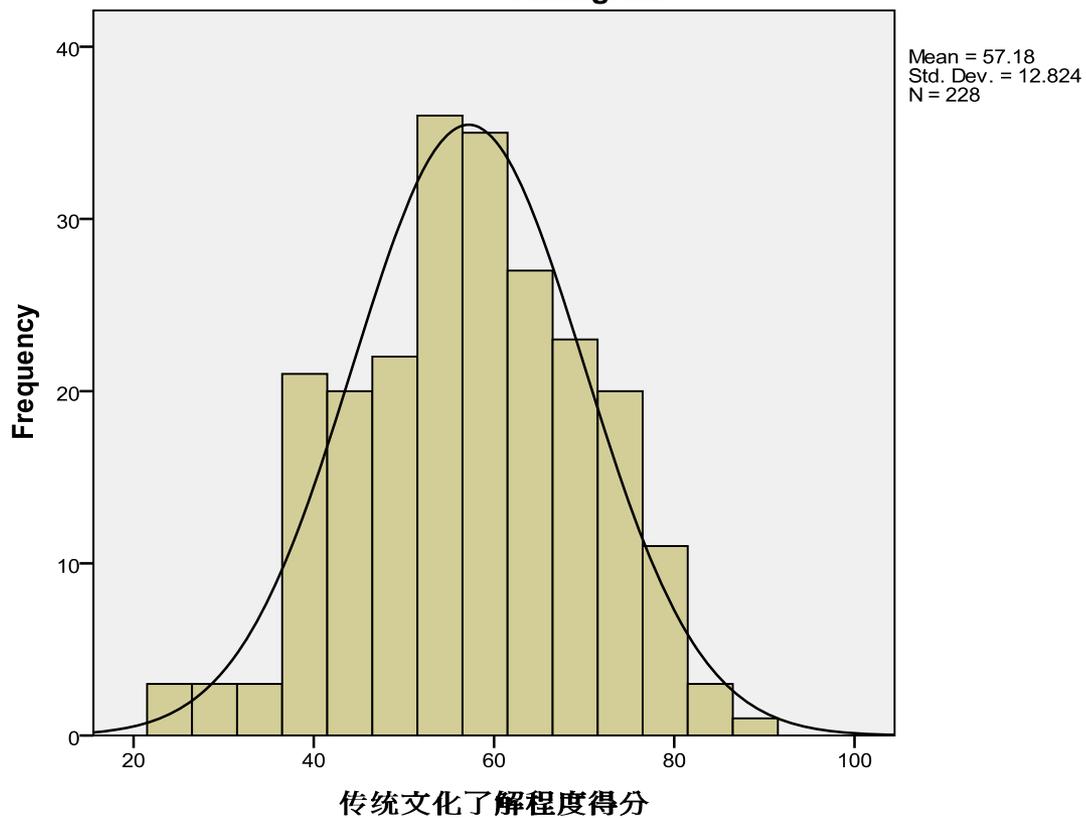
N	Valid	228
	Missing	0
Mean		57.18
Median		57.00
Mode		56
Std. Deviation		12.824
Variance		164.457
Skewness		-.116
Std. Error of Skewness		.161
Kurtosis		-.278
Std. Error of Kurtosis		.321
Range		66
Percentiles	20	45.80
	40	54.00
	60	60.40
	80	68.00

## (2) 直方图

SPSS输出结果也包括直方图。从图形特征看，数据呈稍微左偏分布。根据附带的正态分布曲线可见了解程度得分近似服从正态分布 $N(57.18, 12.824)$ ，说明大学生对中国传统文化的了解程度差异较大。



Histogram



## 2 问题二结果



### (1) 来源途径与了解程度等级的列联表

表14-2是来源途径与了解程度等级的列联表，表中数据列出了处于不同了解程度等级及来源途径的学生人数。可以看到，以“学校教育”为主要来源途径的学生大多数对传统文化了解程度位于“很不了解”和“不太了解”的等级，而采用“自学”方式来获取传统文化的学生对其了解程度都比较高，多数学生都“比较了解”或“很了解”传统文化。

表 14-2 来源途径与了解程度等级的列联表

		了解程度的等级分类					Total
		很不了解	不太了解	一般了解	比较了解	很了解	
您的中国传统文 化知识的主要来 源?	学校教育	25	24	23	15	10	97
	家庭教育	9	10	10	9	9	47
	自学	11	15	10	24	24	84
Total		45	49	43	48	43	228

## (2) 独立性检验

上面的列联表只是从数值大小的角度说明了不同来源途径的学生对传统文化了解程度差异很大，但究竟这种级别有无显著性差异，还是要借助于卡方检验。表14-3是“来源途径”对“了解程度等级”有无显著性影响的卡方检验结果。卡方检验的零假设是不同来源途径对传统文化了解程度没有显著性差异。系统默认显著性水平为0.05，由于卡方检验概率P值都小于0.05，则拒绝零假设，认为来源途径对学生了解中国传统文化程度有显著性差异。这表示应努力激发学生对传统文化的兴趣，只有建立在兴趣爱好的基础上，学生即使花费自己的工作学习时间，也会自学中国传统文化，提高自身文化修养水平。

表 14-3 Chi-Square Tests

	Value	df	Asymp. Sig. (2-sided)
Pearson Chi-Square	19.920 <sup>a</sup>	8	.011
Likelihood Ratio	20.347	8	.009
Linear-by-Linear Association	15.244	1	.000
N of Valid Cases	228		



# 第15章 SPSS在旅游业中的应用

## 15.1 实例提出：国内旅游收入影响因素

CONCEPT  
STRATE

考虑到国内旅游收入主要影响因素有收入水平、休假政策、交通状况等方面的影响。表15-1是依据文献资料，选取反映上述方面的统计指标，包括国内旅游收入、国内生产总值、旅游人均花费、城市居民旅游花费、农村居民旅游花费、公路线路里程、铁路线路里程。特别的为了分析1999年休假制度改革对旅游收入的影响，增加了虚拟变量“制度”来分析它对于旅游收入的影响。

表 15-1 我国旅游收入影响因素

年份	收入 (亿元)	GDP (亿元)	人均花费 (元)	城市(元)	农村(元)	公路里数 (万千米)	铁路里数 (万千米)	制度
1994	1023.5	48197.86	195.3	414.67	54.88	111.78	5.9	0
1995	1375.7	60793.73	218.7	464.02	61.47	115.7	6.2389	0
1996	1638.4	71176.59	256.2	534.1	70.45	118.58	6.49	0
1997	2112.7	78973.03	328.1	599.8	145.68	122.64	6.6	0
1998	2391.2	84402.28	345.0	607	197	127.85	6.64	0
1999	2831.9	89677.05	394.0	614.8	249.5	135.17	6.74	1
2000	3175.5	99214.55	426.6	678.6	226.6	140.27	6.87	1
2001	3522.4	109655.17	499.5	708.3	212.7	169.8	7.0058	1
2002	3878.4	120332.69	441.8	739.7	209.1	176.52	7.19	1
2003	3442.3	135822.76	395.7	684.9	200	180.98	7.3	1
2004	4710.7	159878.34	427.5	731.8	210.2	187.07	7.44	1
2005	5285.9	183217.4	436.1	737.1	227.6	334.52	7.54376	1
2006	6229.7	211923.5	446.9	766.4	221.9	345.7	7.7084	1
2007	7770.6	257305.6	482.6	906.9	222.5	358.37	7.79659	1

## 15.2 实例的SPSS软件操作详解

CONCEPT  
TRATE

本实例要分析国内旅游收入（Y）的影响因素，因此可以建立旅游收入与GDP、旅游人均花费、公路里程数等变量之间的回归模型。通过回归系数的大小来探讨这些因素对旅游收入的影响大小。但是根据相关性分析结果表15-2看到，自变量之间存在着高度的线性相关性。因此本实例直接利用回归分析模型来分析影响因素可能出现多重共线性的现象，造成部分回归系数不显著，因此首要需要考虑的是如何处理变量之间的多重共线性问题。

表 15-2 相关分析结果表

	国内生产总值	旅游人均花费	城市旅游花费	农村旅游花费	公路里数	铁路里数
国内生产总值	1	.727**	.901**	.615*	.951**	.938**
旅游人均花费	.727**	1	.924**	.916**	.624*	.861**
城市旅游花费	.901**	.924**	1	.812**	.787**	.953**
农村旅游花费	.615*	.916**	.812**	1	.518	.769**
公路里数	.951**	.624*	.787**	.518	1	.855**
铁路里数	.938**	.861**	.953**	.769**	.855**	1

因子分析方法是指用较少个数的公共因子的线性函数与特定因子之和来表达原解释变量的分量，以达到降低维数并能合理地解释原解释变量。本实例中，利用因子分析法中的主成分分析法消除经济因素变量的多重共线性问题，使得经济因素的解释变量在降低维度的同时消除多重共线性。通过分析因子和“制度”虚拟变量对国内旅游收入的影响来探讨旅游收入的影响因素。

具体操作步骤如下：

# Step01: 打开数据文件



打开或建立数据文件15-1. sav。同时单击数据浏览窗口的【Variable View (变量视图)】选项，检查各个变量的数据结构定义是否合理，是否需要修改调整。

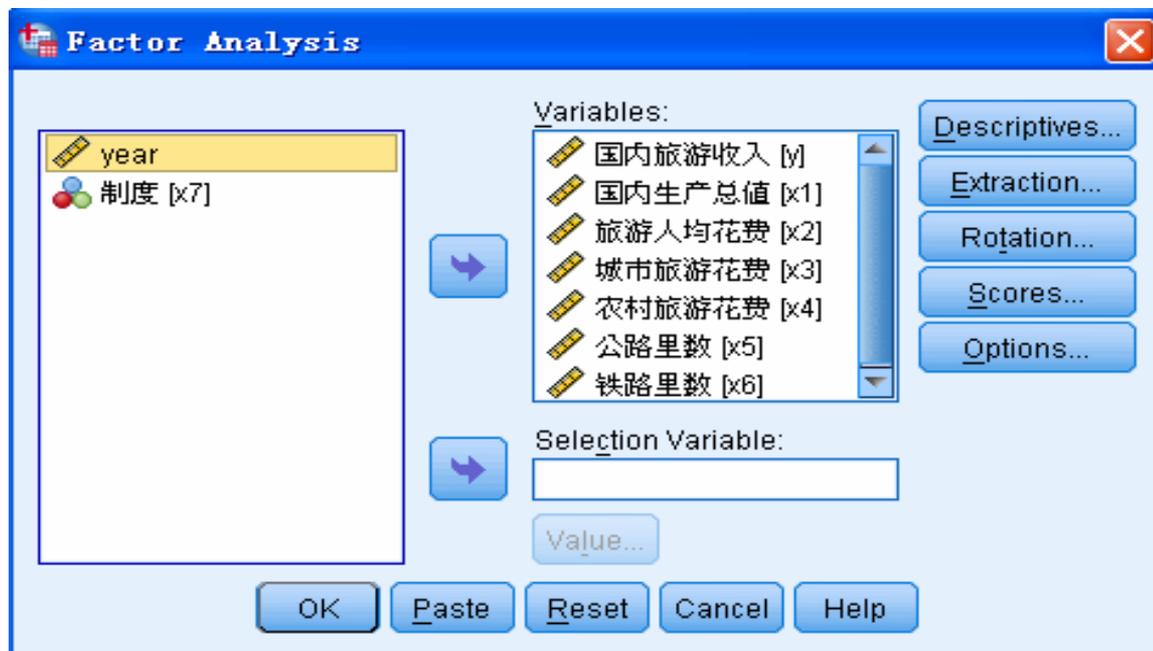
## Step02: 因子分析

CONCEPT  
STRATE

在候选变量列表框中选择X1、X2、...X6变量设定为因子分析变量，将其添加至【Variables（变量）】列表框中。单击【Descriptives】按钮，勾选【KMO and Bartlett's test of sphericity（KMO和Bartlett的球形检验）】复选框，表示进行因子分析适用性的巴特利特球度检验，其他选项保持系统默认，单击【Continue】按钮返回。

在主对话框中，单击【Score按】钮，勾选【Save as variables（保存为变量）】复选框，表示采用回归法计算因子得分并保持在原文件中。其他选项保持系统默认，单击【Continue】按钮返回主对话框。

单击【OK】按钮，完成本步操作。



## Step03: 回归分析

CONCEPT  
STRATE

在第二步因子分析中得到了所有旅游收入影响因素综合得分 $z$ ，这些因子得分充分反映了这些指标在不同年份的综合发展值。于是可以考虑利用它和制度虚拟变量来对国内旅游收入进行回归分析。具体模型如下：

$$y = \beta_0 + \beta_1 z + \beta_2 x_7$$

其中， $y$ 表示国内旅游收入， $z$ 表示综合旅游影响值， $x_7$ 表示虚拟变量。

选择菜单栏中的【Analyze（分析）】→【Regression（回归）】→【Linear（线性）】命令，弹出【Linear Regression（线性回归）】对话框，在左侧的候选变量列表框中选择“y”变量设定为因变量，将其添加至【Dependent（因变量）】列表框中。在左侧的候选变量列表框中选择“z”和“x7”变量设定为自变量，将其添加至【Independent(s)（自变量）】列表框中。最后，单击【OK（确定）】按钮，操作完成。



### Linear Regression

**Dependent:** 国内旅游收入 [y]

Block 1 of 1

**Independent(s):** Z, 制度 [x7]

Method: Enter

Selection Variable:  Rule...

Case Labels:

WLS Weight:

OK Paste Reset Cancel Help

Statistics... Plots... Save... Options... Bootstrap...

- year
- 国内生产总值 [x1]
- 旅游人均花费 [x2]
- 城市旅游花费 [x3]
- 农村旅游花费 [x4]
- 公路里数 [x5]
- 铁路里数 [x6]
- 制度 [x7]
- Z

## 12.3 实例的SPSS输出结果详解

CONCEPT  
STRATE

### (1) 巴特利特球度检验和KMO检验

首先表15-3显示了对数据进行因子分析适用性检验的结果。巴特利特球度检验统计量的观测值等于119.915,。如果显著性水平等于0.05,由于概率P值小于显著性水平,应拒绝原假设,认为相关系数矩阵与单位阵有显著差异。同时,KMO值为0.763,根据Kaiser给出的KMO度量标准可知原有变量适合进行因子分析。

表 15-3 KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy		.763
Bartlett's Test of Sphericity	Approx. Chi-Square	119.915
	df	15
	Sig.	.000

## (2) 因子分析共同度

表15-6是因子分析的共同度，显示了所有变量的共同度数据。如果对原有六个变量如果采用主成分分析法提取所有七个特征根，那么原有变量的所有方差都可被解释，变量的共同度均为1。接着，第二列列出了按指定提取条件提取特征根时的共同度。可以看到，所有变量的绝大部分信息可被因子解释，这些变量信息丢失较少。

表 15-6 因子分析共同度

	Initial	Extraction
国内生产总值	1.000	.872
旅游人均花费	1.000	.841
城市旅游花费	1.000	.956
农村旅游花费	1.000	.701
公路里数	1.000	.740
铁路里数	1.000	.956

### (3) 因子分析的总方差解释

接着Spss软件计算得到相关系数矩阵的特征值、方差贡献率及累计方差贡献率结果如表15-7所示。结果表明，由于数据的相关性较强，选择第一个因子为主因子即可，因为它解释了原有六个变量总方差的84.449%。

表 15-7 因子分析的总方差解释

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.067	84.449	84.449	5.067	84.449	84.449
2	.717	11.951	96.400			
3	.117	1.947	98.347			
4	.056	.936	99.282			
5	.034	.561	99.843			
6	.009	.157	100.000			

### (5) 因子载荷矩阵

表15-8显示了因子载荷矩阵。通过载荷系数大小可以看到不同公共因子所反映的主要指标的区别。从结果看，第一因子在所有变量的载荷系数都较大，基本都在0.80以上，说明它主要反映了旅游收入的综合影响因素。

表 15-8 因子载荷矩阵

	Component
	1
国内生产总值	.934
旅游人均花费	.917
城市旅游花费	.978
农村旅游花费	.837
公路里数	.860
铁路里数	.978

## (6) 因子得分系数

表15-9列出了采用回归法估计的因子得分系数。同时在原数据浏览窗口中新增了变量“FAC1\_1”，它表示不同年份的综合影响因素值。为了表述方便，将其改写为“Z”变量。

表 15-9 因子得分系数

	Component
	1
国内生产总值	.184
旅游人均花费	.181
城市旅游花费	.193
农村旅游花费	.165
公路里数	.170
铁路里数	.193



## 2 回归分析结果

### (1) 模型摘要

表15-10给出了衡量该回归方程优劣的统计量。调整的R<sup>2</sup>为0.928，说明拟合的线性回归模型反映了原始数据92.8%的信息，拟合效果较好。

表 15-10 模型摘要

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.969	.939	.928	517.16343

## (2) 方差分析表

表15-11是对回归模型进行方差分析的检验结果。可以看到方差分析结果中F统计量等于84.790，概率P值小于显著性水平0.05，所以该模型是有统计学意义的，即综合影响因素和制度变量是显著的。

表 15-11 方差分析表

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	45355649.962	2	22677824.981	84.790	.000
	Residual	2942038.202	11	267458.018		
	Total	48297688.164	13			

### (3) 回归系数表

表15-12给出了回归模型的参数估计结果，于是得到回归方程如下：

$$y = 4083.395 + 2209.809z - 864.292x_7$$

接着将表15-8的因子载荷系数带入到Z变量的表达式中，进入可以将上述回归模型改写为如下形式：

$$y = 4083.40 + 2063.96x_1 + 2026.40x_2 + 2161.19x_3 + 1849.61x_4 \\ + 1900.44x_5 + 2161.19x_6 - 864.29x_7$$

表 15-12 回归系数表

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	4083.395	361.165		11.306	.000
	Z	2209.809	258.091	1.146	8.562	.000
	制度	-864.292	519.043	-.223	-1.665	.124

将拟合预测值与实际值比较后得知，模型有很高的拟合优度，并且模型中各变量系数符号的经济意义合理。各项影响因素的增长，对旅游收入均存在正向作用。同时，由于制度变量的t统计量的概率P值等于0.124，大于显著性水平0.05，说明本模型中政策性因素不显著。

由方程系数可知，城市居民旅游花费和铁路里数的增加对于国内旅游收入的影响，相比较于其它影响因素最为显著。



## 第16章 SPSS在数据挖掘中的应用



## 16.1 实例提出：168信息点播业务

数据16-1.sav是某月陕西主要地区各类业务的流量数据，数据16-2.sav是该月每天各类业务的流量数据。请利用这些资料分析以下问题：

问题一：请分析在168信息点播服务方面陕西各地区（西安、宝鸡、咸阳等）总流量的差别。

问题二：请指出该月点播业务最好三项栏目，并分析它们之间的流量有无显著性差异。

问题三：请预测该月点播业务最好栏目的长期发展趋势。



## 16.2 实例的SPSS软件操作详解

### 问题一操作详解

问题一要求分析在168信息点播服务方面陕西各地区（西安、宝鸡、咸阳等）总流量的差别。由于各地区在股票点播、指数点播等业务上的流量数据差异较大，并没有统一的大小顺序关系，因此可以采用聚类分析研究陕西各地区的总流量差异。

# 问题一操作详解



## Step01: 打开数据文件及对话框

打开数据文件16-1sav, 选择菜单栏中的【Analyze(分析)】→【Classify(分类)】→【Hierarchical Cluster(系统聚类)】命令, 弹出【Hierarchical Cluster Cluster Analysis(系统聚类分析)】对话框。

## Step02: 选择聚类分析变量

在左侧的候选变量列表框中选择西安、宝鸡、榆林等十个地区变量设定为聚类分析变量, 将其添加至【Variables(变量)】列表框中。同时点选【Variable(变量)】单选钮, 表示选择聚类对象为指标变量。

## Step03: 输出聚类数目

在主对话框中单击【Statistics】按钮, 弹出相应对话框。点选【Single solution(单一方案)】单选钮, 并在【Number of clusters(聚类数)】文本框中键入数字“3”表示利用聚类分析将十个地区分为三类。其他选项保持系统默认, 单击【Continue】按钮返回主对话框。

# 问题一操作详解

CONCEPT  
STRATE

## Step04: 输出聚类图

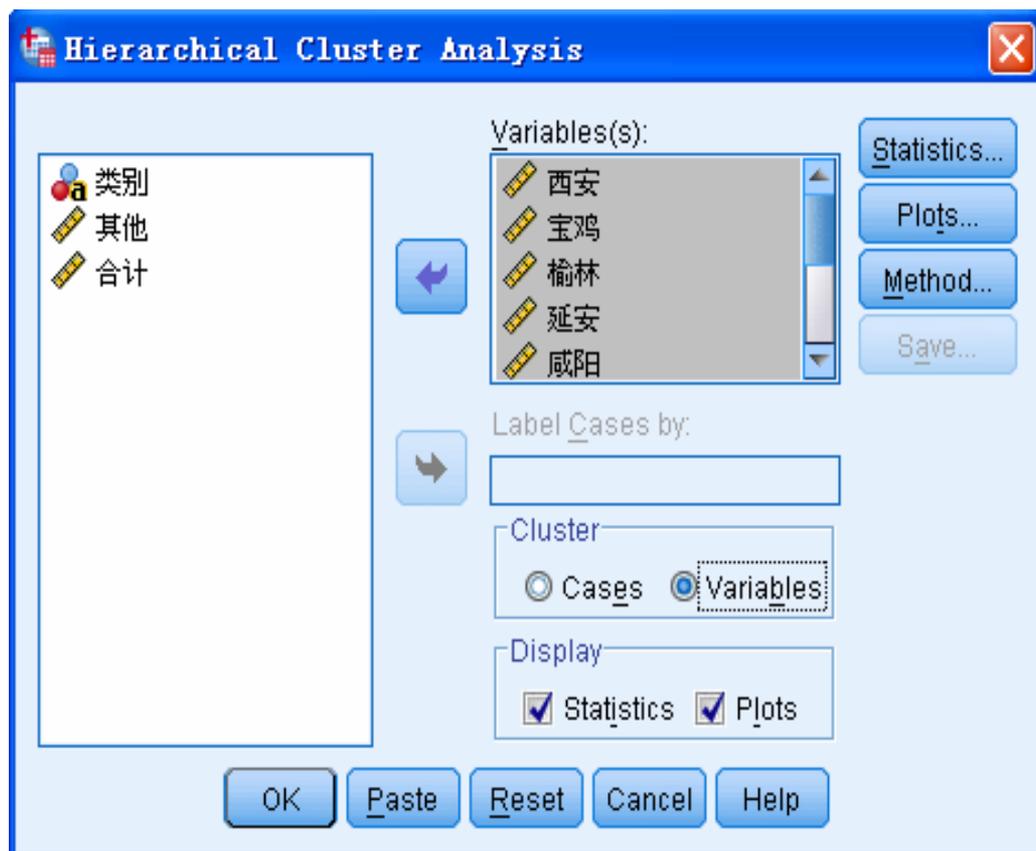
在主对话框中单击【Plots】按钮，弹出【Plots(绘制)】对话框。勾选【Dendrogram(树状图)】复选框，表示输出样品的聚类树形图。其他选项保持系统默认，单击【Continue】按钮返回主对话框。

## Step05: 聚类方法选择

在主对话框中单击【Method】按钮，弹出【Method(方法)】对话框。在【Transform Values(转换值)】选项组的【Standardize】下拉菜单中选择【Z scores(Z得分)】标准化方法。其他选项保持系统默认，单击【Continue按】钮返回主对话框。

## Step06: 单击【OK】按钮，完成操作。

# 问题一操作详解



# 问题二操作详解

CONCEPT  
STRATE

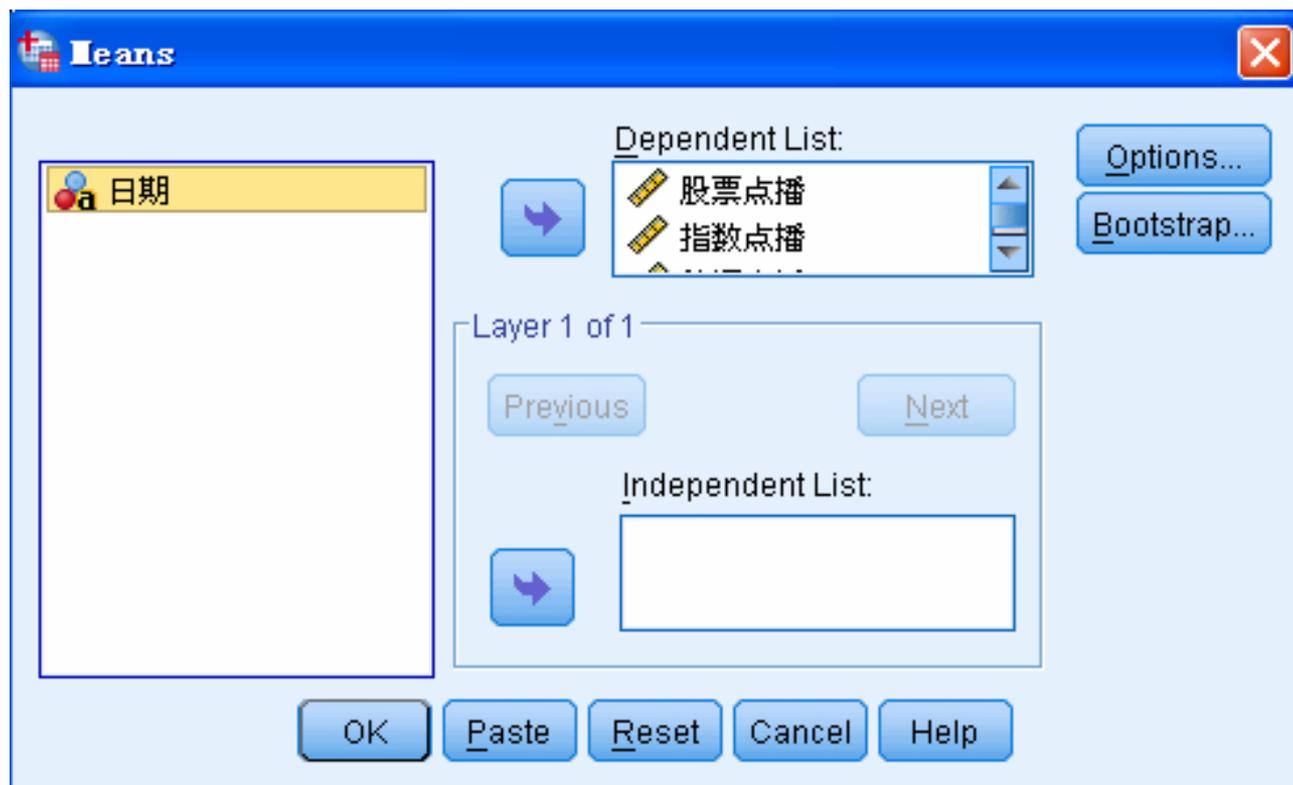
## Step01: 计算各项业务的日平均流量

打开数据文件16-2.sav，选择菜单栏中的【Analyze (分析)】→【Compare Means (比较均值)】→【Means (均值)】命令，弹出【Means (均值)】对话框。在左侧的候选变量列表框中选择“股票点播”、“指数点播”等业务。其他选项保持系统默认，单击【OK】按钮完成操作。

接着根据输出的业务流量统计数据表16-2.sav，可以确定日平均流量最大的三项业务“股票点播”、“每日运程”和“劲爆笑话”为点播业务最大的业务。

## 问题二操作详解

CONCEPT  
STRATE



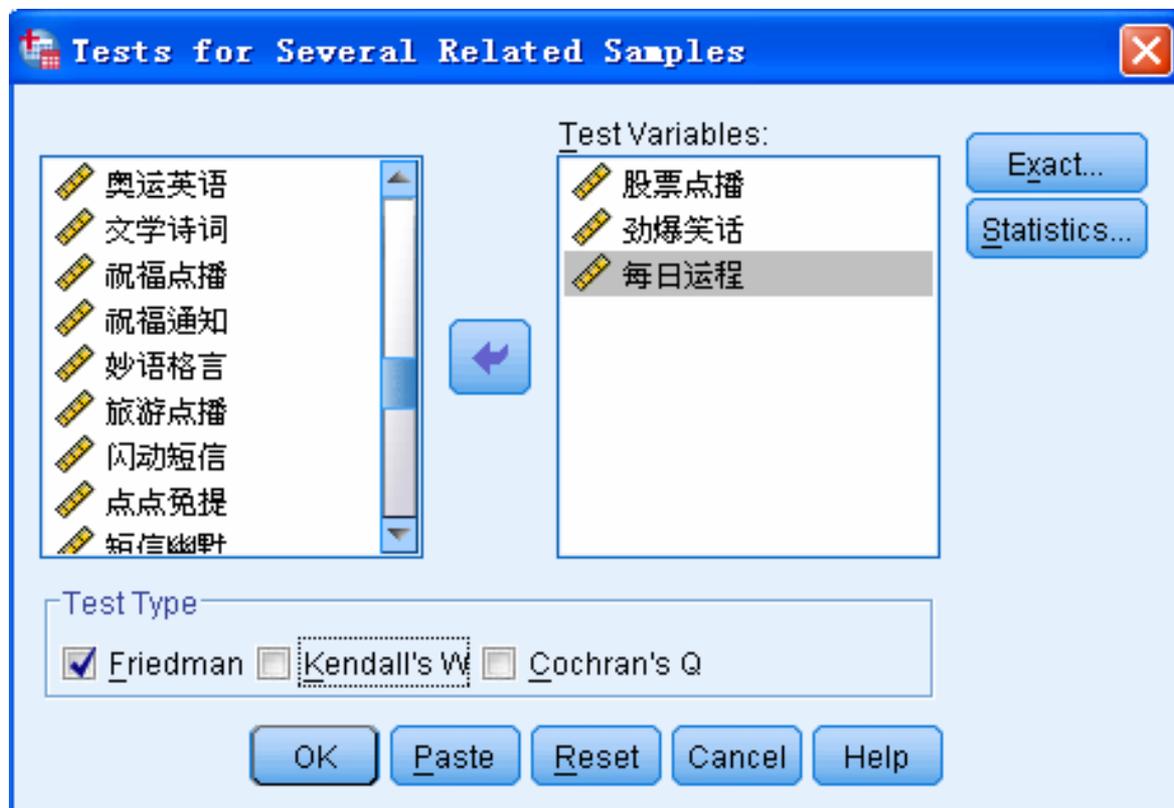
# 问题二操作详解

CONCEPT  
STRATE

## Step02：业务流量的差异性研究

选择菜单栏中的【Analyze(分析)】→【Nonparametric Tests(非参数检验)】→【Legacy Dialogs(旧对话框)】→【K Related Samples(K个相关样本)】命令，弹出【Tests for Several Related Samples(多个关联样本检验)】对话框。在候选变量列表框中同时选择“股票点播”、“每日运程”和“劲爆笑话”变量作为配对检验变量，将其同时添加至【Test Variable(s)(检验变量)】列表框中。在【Test Type(检验类型)】选项组中勾选【Friedman】复选框作为配对样本检验的方法。最后单击主对话框中的【OK】按钮，完成操作。

# 问题二操作详解



# 问题三操作详解

CONCEPT  
STRATE

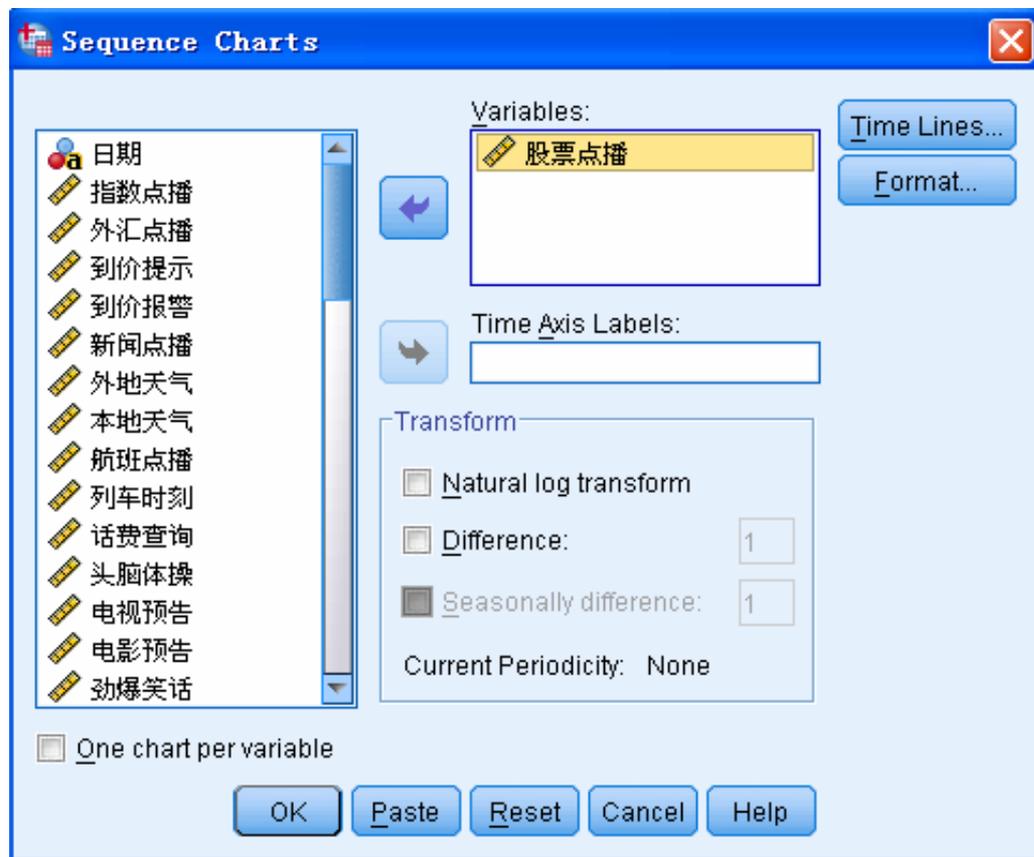
## 1. Step01: 绘制序列图

打开数据文件16-2.sav，选择菜单栏中的【Analyze(分析)】→【Forecasting(预测)】→【Sequence Charts(序列图)】命令，弹出【Sequence Charts(序列图)】对话框。在左侧的候选变量列表框中选择“股票点播”进入右侧的【Variables(变量)】列表框。其他选项保持系统默认，单击【OK】按钮完成操作。

根据序列图，观测到股票点播数据虽然平稳，但具有明显的周期性波动特征，因此可以利用ARMA模型来描述点播数据的波动性。



# 问题三操作详解



# 问题三操作详解

## Step02: 时间序列ARMA模型

选择菜单栏中的【Analyze(分析)】→【Forecasting(预测)】→【Create Models(创建模型)】命令，弹出【Time Series Modeler(时间序列建模器)】对话框。在左侧的候选变量列表框中选择“股票点播”进入右侧的【Dependent Variables(因变量)】列表框，表示对其进行ARMA模型分析。选择【Method(方法)】下拉菜单中的【ARIMA】选项，表示进行ARMA模型估计。接着单击【Criteria(条件)】按钮，弹出ARIMA模型阶数设定窗口。

观察序列图发现点播数据以7天为周期进行波动，反复进行ARMA模型滞后阶数的尝试后，最终选择AR(7)模型来描述股票点播流量的波动性。于是在【Time Series Modeler(时间序列建模器)】窗口【Autogressive(p)(自回归(p))】选项组的【Nonseasonal(非季节性)】文本框中填入数字“7”。在【Transformation(转换)】选项组中勾选【Natural log(自然对数)】单选钮。再单击

【Continue】按钮，返回主对话框。



# 问题三操作详解

The screenshot displays the 'Time Series Modeler' application window. At the top, there are tabs for 'Variables', 'Statistics', 'Plots', 'Output Filter', 'Save', and 'Options'. The 'Variables' tab is active, showing a list of variables on the left and two empty boxes for 'Dependent Variables' and 'Independent Variables' on the right. The 'Method' is set to 'ARIMA' and the 'Model Type' is 'ARIMA(7, 0, 0)'. Below these, there are sections for 'Estimation Period' and 'Forecast Period' with their respective start and end settings. At the bottom, there are buttons for 'OK', 'Paste', 'Reset', 'Cancel', and 'Help'.

**Time Series Modeler**

Variables: Statistics Plots Output Filter Save Options

Variables:

- 指数点播
- 外汇点播
- 到价提示
- 到价报警
- 新闻点播
- 外地天气
- 本地天气
- 航班点播
- 列车时刻
- 话费查询
- 头脑体操
- 电视预告
- 电影预告
- 劲爆笑话
- 趣味猜谜

Dependent Variables:

- 股票点播

Independent Variables:

Method: ARIMA Criteria...

Model Type: ARIMA(7, 0, 0)

Estimation Period: Start: First case, End: Last case

Forecast Period: Start: First case after end of estimation period, End: Last case in active dataset

OK Paste Reset Cancel Help

# 问题三操作详解



Time Series Modeler: ARIMA Criteria

Model Outliers

ARIMA Orders

Structure:

	Nonseasonal	Seasonal
Autoregressive (p)	7	0
Difference (d)	0	0
Moving Average (q)	0	0

Current periodicity: None

Transformation

None

Square root

Natural log

Include constant in model

Continue Cancel Help

# 问题三操作详解

CONCEPT  
STRATE

单击【Statistics】按钮，勾选其中的【Parameter estimates (参数估计)】复选框，表示输出模型参数估计结果和模型预测值；同时取消勾选【Goodness of fit (拟合优度)】复选框，其他选项保持系统默认。

单击【Plots】选项，勾选其中的【Residual autocorrelation function(ACF)(残差自相关函数)】和【Residual partial autocorrelation function(PACF)(残差部分自相关函数)】复选框，表示绘制残差的自相关图和偏相关图。不仅如此，勾选【Fit values (拟合值)】复选框输出模型的拟合效果图。其他选项保持系统默认。

最后，单击【OK】按钮完成操作。



## 16.3 实例的SPSS输出结果详解

### 问题一输出结果详解

#### (1) 聚类过程表

SPSS软件首先给出了进行系统聚类分析的过程表，它动态显示了所有地区的聚类过程。下表显示第二地区和第九个地区首先被合在一起，聚类系数等于2.356，它们将在第二步中与其他类再进行合并。其他结论可以依此类推。

# 问题一输出结果详解



Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	2	9	2.356	0	0	2
2	2	6	6.123	1	0	6
3	7	10	6.489	0	0	7
4	3	8	10.459	0	0	5
5	3	4	11.166	4	0	7
6	2	5	20.564	2	0	9
7	3	7	27.801	5	3	8
8	1	3	36.823	0	7	9
9	1	2	42.702	8	6	0

# 问题一输出结果详解

CONCEPT  
STRATE

## (2) 聚类分析结果表

下表显示了系统聚类法的聚类结果。可以看到聚类结果分为两大类：

第 I 类：西安；

第 II 类：宝鸡、咸阳、铜川、汉中；

第 III 类：榆林、延安、渭南、安康、商洛。

其中第 I 类地区西安是168信息各类点播业务流量最大的地区，第 III 类的五个地区在所有地区中是相对168信息点播业务流量最低，而第 II 类地区的点播业务流量是介于第 I 类和第 III 类之间，保持中游水平。

分析地区间的点播量的差异部分是由于地区特征的差异引起的，例如人口数量、经济发展状况（收入水平、手机拥有量、物价水平等），同时也与地区业务的宣传力度有密切联系。分析清楚这些原因后公司就可以采取相应的措施扩大业务。



# 问题一输出结果详解

Case	3 Clusters
西安	1
宝鸡	2
榆林	3
延安	3
咸阳	2
铜川	2
渭南	3
安康	3
汉中	2
商洛	3

# 问题一输出结果详解

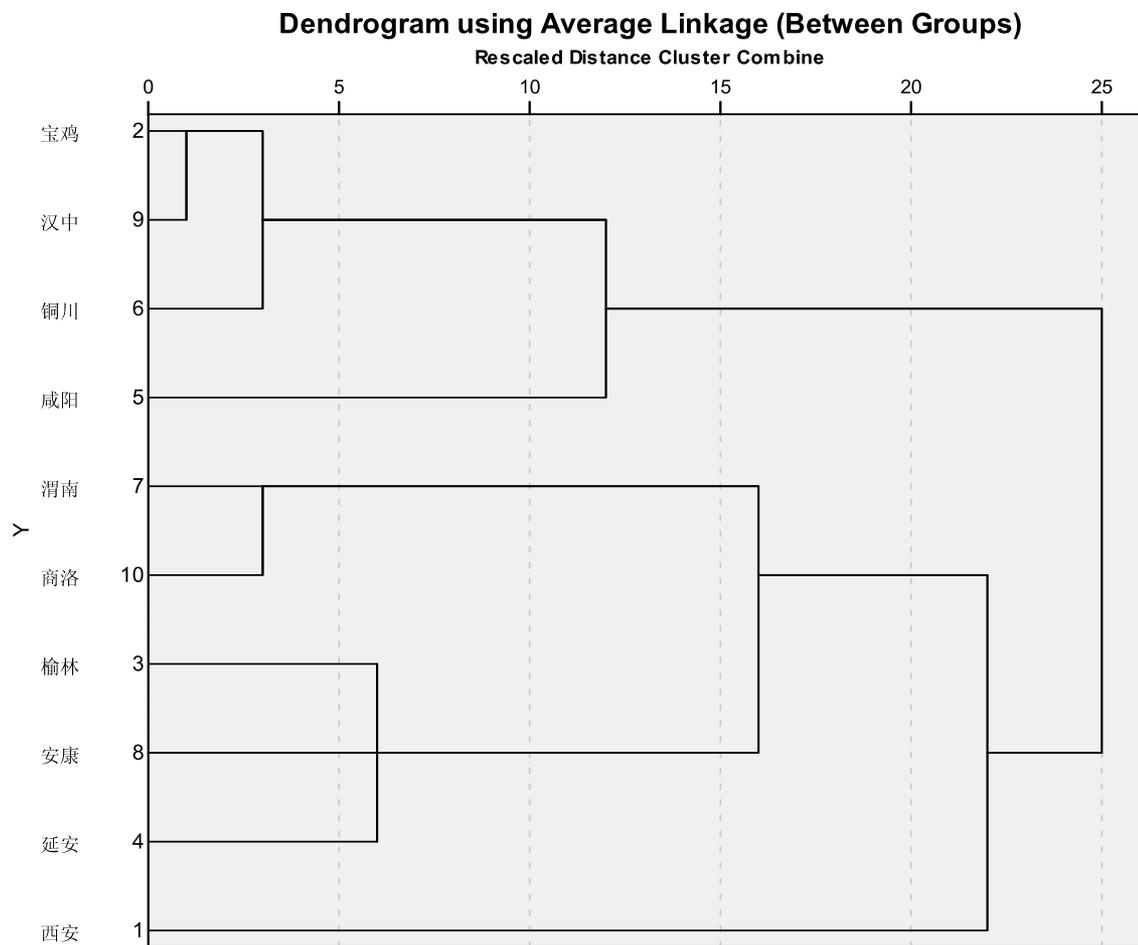
CONCEPT  
STRATE

## (3) 树形图

上表已给出了相关聚类结果，最后用树形图（Dendrogram）直观反映整个聚类过程和结果。



# 问题一输出结果详解



## 问题二输出结果详解

CONCEPT  
STRATE

- 下表（部分）是利用【Means(均值)】功能计算的各项业务在当月的平均点播量。表中具体给出了均值、统计数目及标准差等基本统计量。比较均值大小可以看到，“股票点播”、“每日运程”和“劲爆笑话”为点播量最大的业务，说明这些业务深受消费者欢迎，公司应努力增加在这些业务方面的内容更新及促销。而相反的，“商讯点播”、“区号邮编”等业务的点播量太低，因此公司可以考虑停止这些服务功能以节约成本。



## 问题二输出结果详解

	Mean	N	Std. Deviation
股票点播	7317.9677	31	4634.75391
指数点播	278.5484	31	164.77658
外汇点播	38.4194	31	14.17927
到价提示	11.6452	31	8.24439
到价报警	176.0645	31	125.84486
新闻点播	2040.2258	31	204.82427
外地天气	139.8387	31	32.26153
本地天气	185.1290	31	54.01280
航班点播	156.9355	31	52.17786
列车时刻	49.0645	31	16.98614
话费查询	2139.0645	31	3322.93176
头脑体操	124.6129	31	69.72311

# 问题二输出结果详解



## (2) 秩统计表

下表是多配对样本非参数检验的秩统计表。可以看到，“股票点播”变量的平均秩最大，等于2.42，说明它的点播量最大，排名更靠后；相反的，“劲爆笑话”变量的平均秩最小，等于1.35，说明它的点播量最小，排名更靠前。

	Mean Rank
股票点播	2.42
劲爆笑话	1.35
每日运程	2.23

## 问题二输出结果详解

CONCEPT  
STRATE

### (3) Friedman统计表

Friedman检验结果如下表所示，样本容量等于31，Chi-Square统计量等于19.935，自由度df等于2，近似相伴概率P值为0.000，远远小于显著性水平0.05。所以拒绝零假设，认为这三种业务的点播量存在显著差异。这说明虽然它们位居所有业务的前三位，但其点播量还是存在显著的差异。因此，公司需要分开对待它们各自的点播业务特点。

N	31
Chi-Square	19.935
df	2
Asymp. Sig.	.000

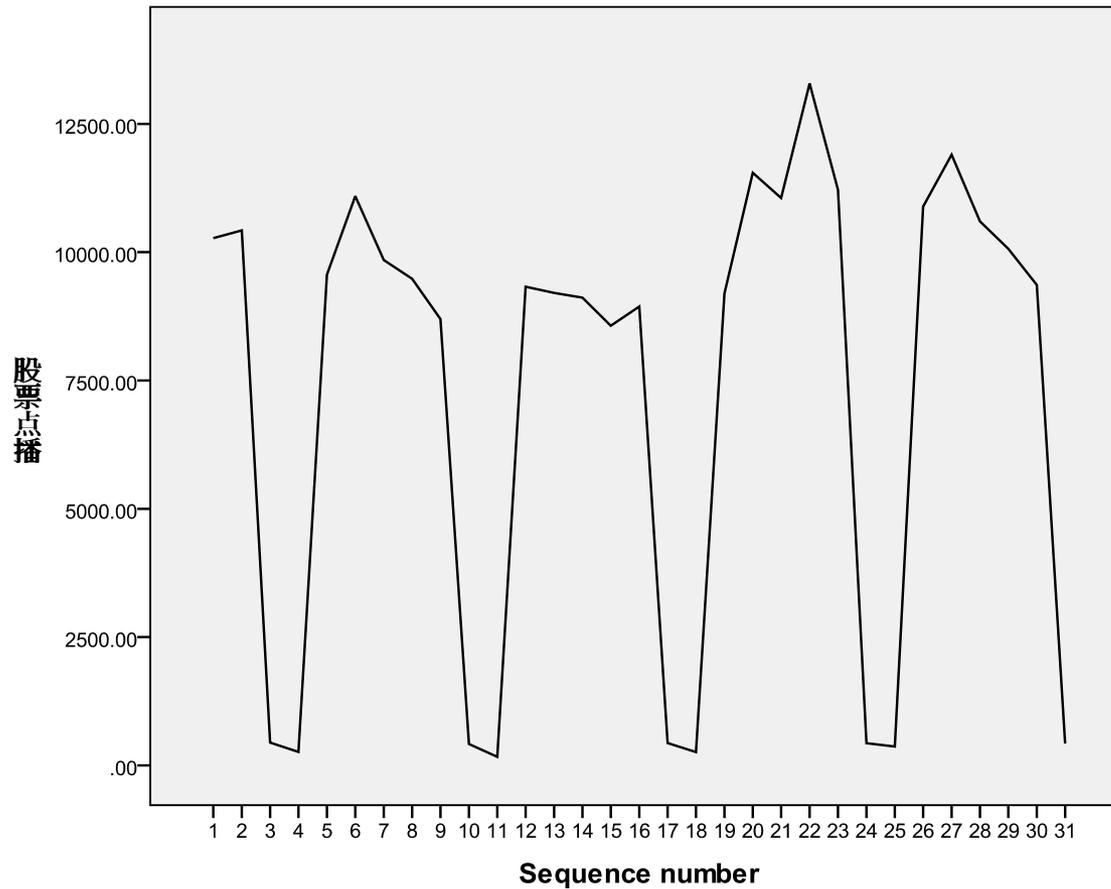
# 问题三输出结果详解

CONCEPT  
STRATE

## (1) 时间序列折线图

下图绘制了“股票点播”业务在该月每日点播量的时间序列图。可以看到，股票点播量是平稳的，但具有显著的周期性，在每个周末的点播量明显低于周内的点播量，这与股票周末休市有密切联系。于是考虑利用ARMA模型来刻画其波动性。

# 问题三输出结果详解



# 问题三输出结果详解

CONCEPT  
STRATE

## (2) 模型拟合优度检验表

下表给出了AR(7)模型的拟合优度值，可以看到拟合优度统计量 $R^2$ 等于0.880，说明模型的整体拟合效果较好。Ljung-Box Q统计量是对点播序列的线性相关性进行检验。从检验结果看，LB检验概率P值大于显著性水平0.05，说明序列基本不存在自相关性



# 问题三输出结果详解

Model	Number of Predictors	Model Fit statistics	Ljung-Box Q(18)			Number of Outliers
		Stationary R-squared	Statistics	DF	Sig.	
股票点播-Mode 1_1	0	.880	5.985	11	.874	0

# 问题三输出结果详解

CONCEPT  
STRATE

## (3) 模型参数估计值表

下表列出了AR(7)模型的参数估计值。可以看到除了滞后7阶（Lag 7）的系数显著外，其他滞后项系数都没有通过显著性检验，其t检验的概率P值都大于0.05。假设“每日股票点播量”记为 $X_t$ ，则最终拟合的模型为：

$$X_t = 8.268 + 0.916 X_{t-1}$$



# 问题三输出结果详解

			Estimate	SE	t	Sig.	
股票点播	Natural Log	Constant	8.268	.084	97.924	.000	
		AR	Lag 1	-.052	.075	-.697	.493
			Lag 2	-.064	.081	-.798	.433
			Lag 3	-.064	.081	-.786	.440
			Lag 4	-.047	.084	-.561	.580
			Lag 5	-.077	.080	-.965	.345
			Lag 6	-.028	.079	-.354	.727
			Lag 7	.916	.074	12.379	.000

# 问题三输出结果详解

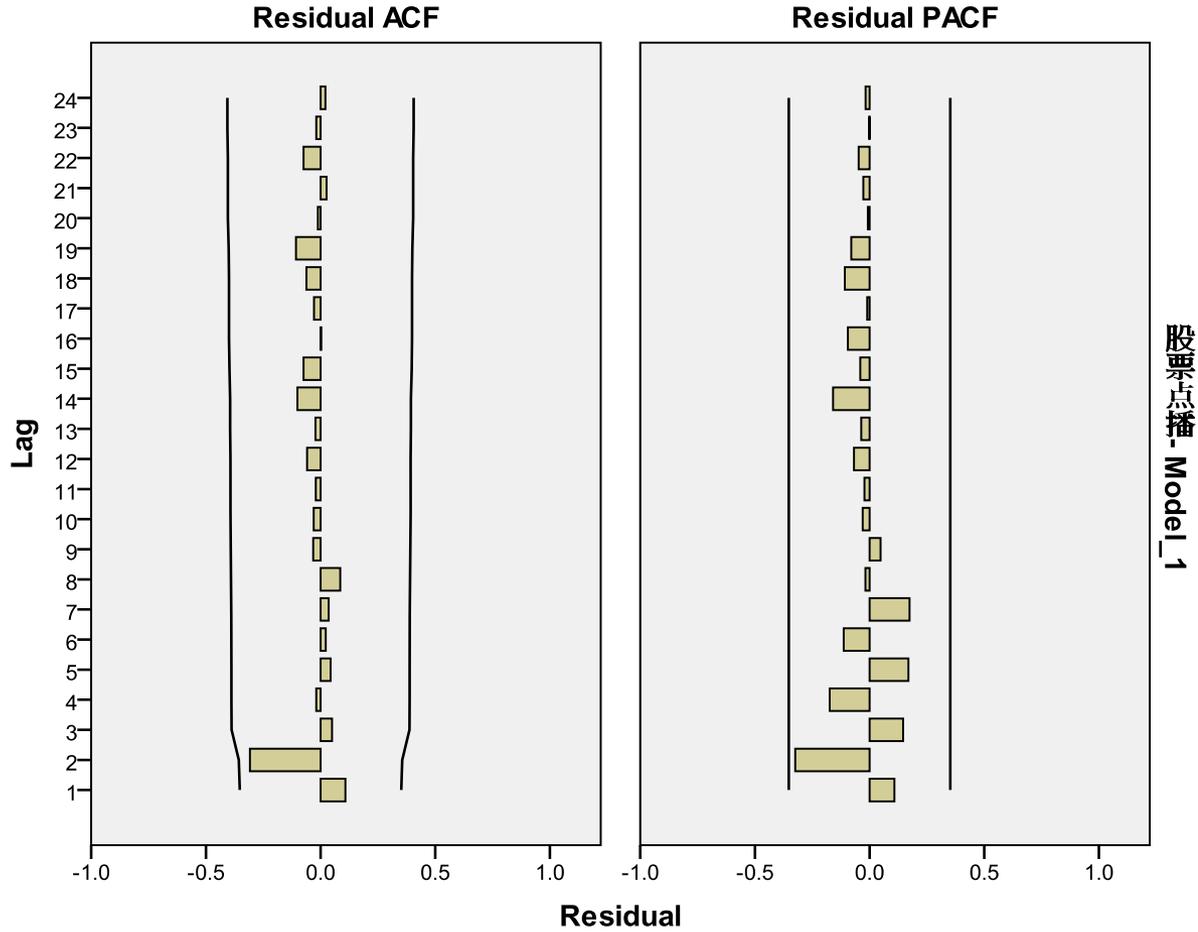
CONCEPT  
STRATE

## (4) 残差自相关和偏相关图

下图给出了不同阶数下拟合模型的残差的自相关和偏相关图。可以看到，两列相关系数都落在置信区间内，说明残差序列的各阶自相关函数值和偏相关函数值都显著等于0，符合白噪声的特征。这也进一步反映了AR(7)模型的合理性。



# 问题三输出结果详解



# 问题三输出结果详解

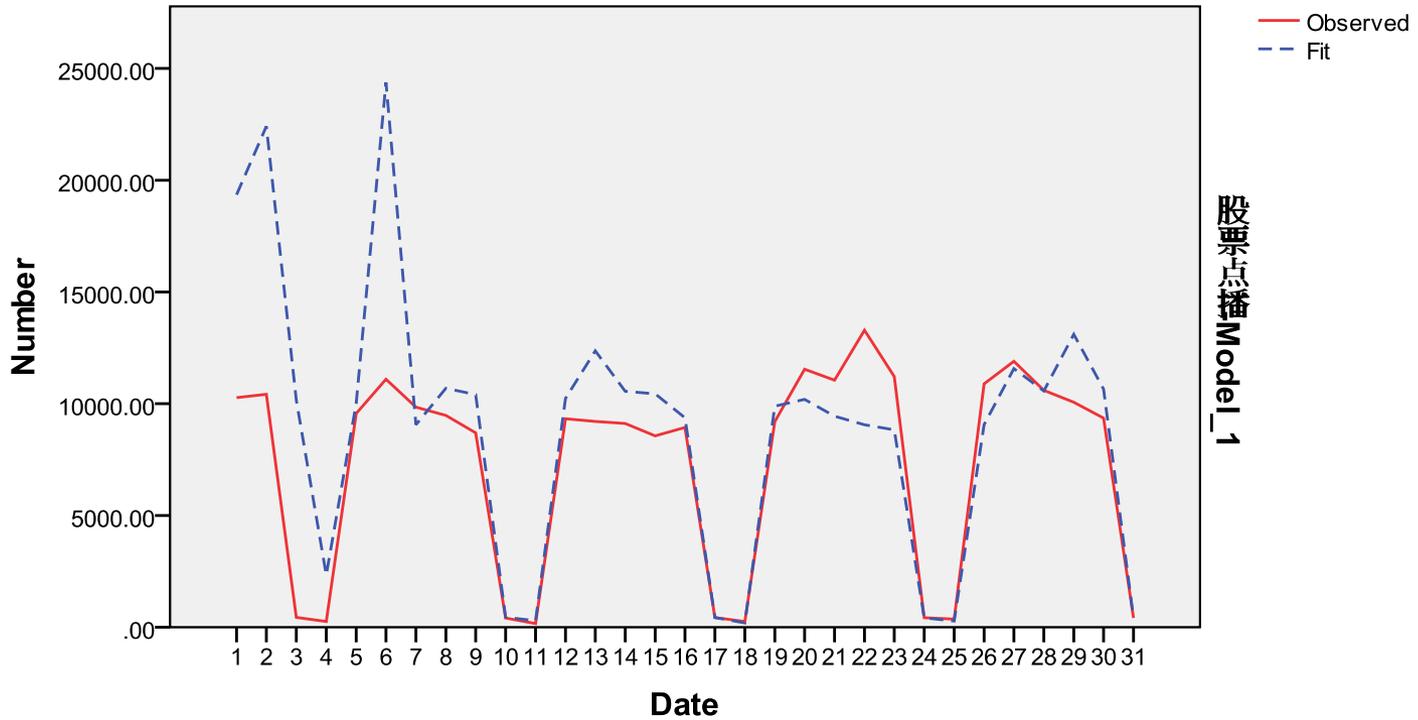
CONCEPT  
STRATE

## (5) 模型拟合效果图

最后，下图显示了本实例提出的AR(7)模型预测值与实际值的拟合效果图。从图形来看，除了在初始几天的模型拟合值偏高外，其他时间的模拟拟合效果都较好，这样可以利用该模型进行后续日期的预测。



# 问题三输出结果详解





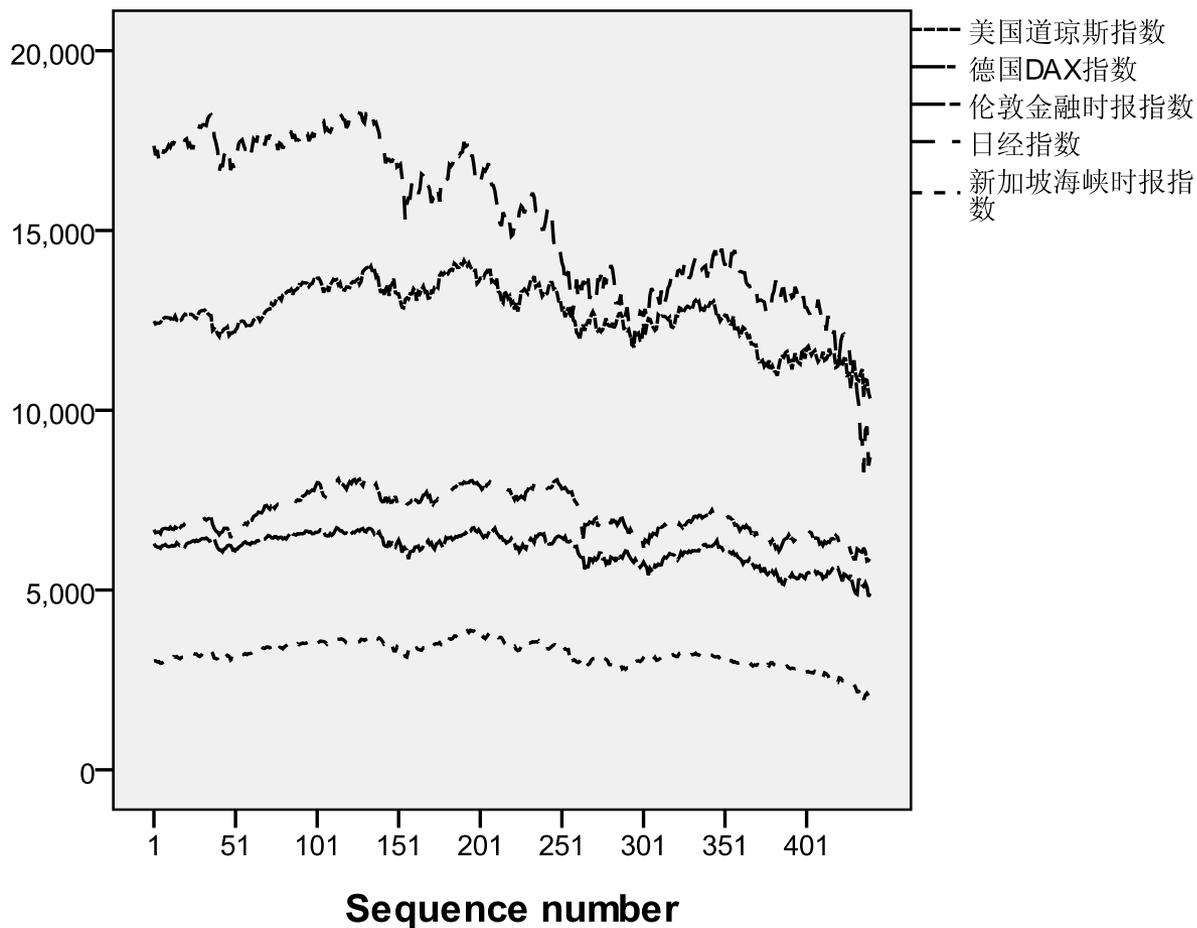
# 第17章 SPSS在金融市场中的应用

# 17.1 实例提出：美国金融危机下 全球股市的波动影响

CONCEPT  
STRATE

由于金融市场的传染效应，美国次贷危机已不仅仅影响到本国的股票市场，同时也影响了全球其他国家和地区的股票市场，例如，英国、日本和新加坡市场等。

下图表示了美国、英国、德国、日本、中国香港和新加坡等全球主要股票市场从2007年1月至2008年10月的股票价格日收盘指数。具体数据见17-1.sav所示。



不同国家股票指数走势图

# 三个问题

CONCEPT  
STRATE

请你利用这些数据，分析以下问题：

- 请建立美国股指波动的数学模型；
- 请分析美国股指波动对其他国家地区的股票市场造成的影响程度；
- 请分析不同国家地区股指波动的差异性。

## 17.2 实例的SPSS软件操作详解

CONCEPT  
RATE

### 问题一操作详解

问题一要建立美国道琼斯指数的波动模型，由于该指数主要随着时间的变动而变动，于是可以考虑建立该指数和时间之间的回归模型。首先从图形特点看，美股指数在研究日期内呈现明显的下降趋势，这反映了金融危机对其造成的显著影响。但是，指数的下跌并不是线性关系，而是表现为显著的非线性特征，于是可以考虑采用非线性回归模型进行数据的拟合分析。

# 具体操作步骤

CONCEPT  
STRATE

## Step01: 打开数据文件

打开数据文件17-1.sav。单击数据浏览窗口的【Variable View (变量视图)】按钮，检查各个变量的数据结构定义是否合理，是否需要修改调整。

## • Step02: 设置因变量和自变量

选择菜单栏中的【Analyze (分析)】→【Regression (回归)】→【Curve Estimation (曲线估计)】命令，弹出【Curve Estimation (曲线估计)】对话框。在候选变量列表框中选择“美国道琼斯指数”变量设定为因变量，将其添加至【Dependent(s) (因变量)】列表框中。同时点选【Time (时间)】按钮，表示设置自变量为时间变量。



**Curve Estimation** [Close]

**Dependent(s):** [Save...]  
美国道琼斯指数 [美国]

**Independent:**  
 Variable: [ ]  
 Time

**Case Labels:** [ ]  Include constant in equation  
 Plot models

**Models:**  
 Linear     Quadratic     Compound     Growth  
 Logarithmic     Cubic     S     Exponential  
 Inverse     Power:     Logistic  
Upper bound: [ ]

Display ANOVA table

[OK] [Paste] [Reset] [Cancel] [Help]



### Step03: 选择曲线拟合模型类型

从原始图像看到美股指数呈显著的非线性下跌趋势，于是在【Model(模型)】复选框中除了保留系统默认的【Linear(线性)】选项外，同时勾选【Exponential(指数分布)】和【Quadratic(二次项)】模型。这表示要对这三种模型进行曲线拟合，同时比较其拟合效果。

单击【OK】按钮，完成本部分操作。



## 问题二操作详解

具体操作步骤如下：

### **Step01:** 打开相关分析对话框

打开数据文件17-1.sav，选择菜单栏中的【Analyze(分析)】→【Correlate(相关)】→【Bivariate(双变量)】命令，弹出【Bivariate Correlations(双变量相关)】对话框。

### • **Step02:** 选择相关分析变量

在候选变量列表框中选择美国、日本、德国等五个国家股指变量，将其添加至【Variables(变量)】列表框中。这表示要分析两两国家之家股指的相关关系。



相关分析窗口

### Step03: 选择相关系数类型

在【Correlation Coefficients (相关系数)】选项组中勾选【Pearson (皮尔森)】、【Kendall (肯德尔)】和【Spearman】三种相关系数类型，表示结果窗口输出这三种类型的相关系数。

单击【OK】按钮，完成本部分操作。

# 问题三操作详解

CONCEPT  
STRATE

具体操作步骤如下：

**Step01:** 打开数据文件及对话框

打开数据文件17-1.sav，选择菜单栏中的【Analyze(分析)】→【Classify(分类)】→【Hierarchical Cluster(系统聚类)】命令，弹出【Hierarchical Cluster Analysis(系统聚类分析)】对话框。

**Step02:** 选择聚类分析变量

在候选变量列表框中选择美国、德国和日本等五个国家股指变量设定为聚类分析变量，将其添加至【Variables(变量)】列表框中。同时点选【Variable(变量)】单选钮。

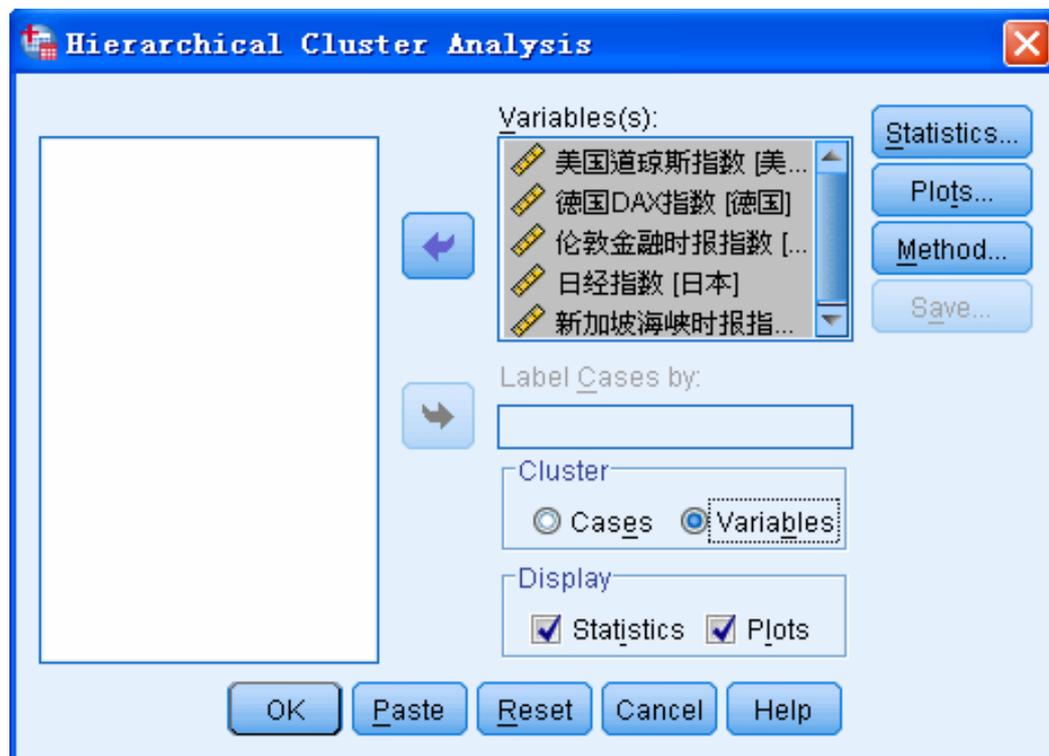
### Step03: 输出聚类图

在主对话框中单击【Plots (绘制)】按钮，弹出【Plots (绘制)】对话框。勾选【Dendrogram (冰柱)】复选框，表示输出样品的聚类树形图。其他选项保持系统默认，单击【Continue】按钮返回主对话框。

### Step04: 聚类方法选择

在主对话框中单击【Method (方法)】按钮，弹出【Method (方法)】对话框。选择【Transform Values (转换值)】→【Standardize (标准化)】下拉菜单的【Z scores (Z得分)】标准化方法。其他选项保持系统默认，单击【Continue】按钮返回主对话框。

### Step05: 单击【OK】按钮，完成操作。



## 聚类分析



## 17.3 实例的SPSS输出结果详解

### 问题一输出结果

#### (1) 模型汇总及参数估计

下表给出了样本数据分别进行三种曲线方程拟合的检验统计量和相应方程中的参数估计值。

从拟合优度值R Square看到，二次曲线的拟合效果相对较好，达到了76.3%，而线性模型和指数函数的拟合优度连50%都没有达到。

虽然上述三个模型都有显著的统计学意义，但从拟合优度值的大小可以看到二次曲线方程较其他两种曲线方程拟合效果更好，因此选择它来描述美股下跌的趋势。



		Equation		
		Linear	Quadratic	Exponential
Model Summary	R Square	.330	.763	.340
	F	215.314	702.740	225.819
	df1	1	2	1
	df2	438	437	438
	Sig.	.000	.000	.000
Parameter Estimates	Constant	13495.485	12292.738	13524.252
	b1	-3.668	12.659	.000
	b2		-.037	

## 模型汇总及参数估计

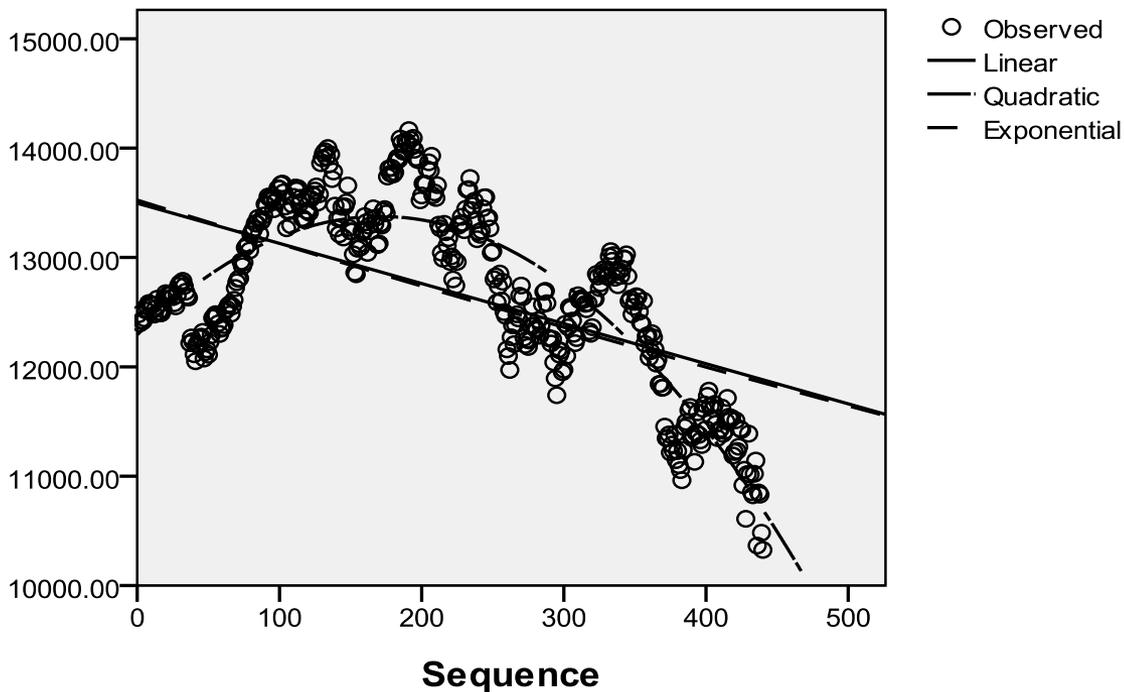
## (2) 拟合曲线图

最后给出的是实际数据的散点图和三种估计曲线方程的预测图，这也进一步说明二次函数曲线方程的拟合效果最好。

需要注意的是，虽然选择的二次函数曲线拟合效果最好，但是它的拟合优度值也只有76.3%，其值也偏低。这说明股市的波动情况复杂，在较长时间范围内，很难用单一的非线性函数加以刻画；相反的，在短期内，由于股市波动变动不大，用曲线拟合的方法能得到较好的结果。



## 美国道琼斯指数



拟合曲线图

# 问题二输出结果

CONCEPT  
STRATE

## (1) Pearson (皮尔森) 相关系数表

首先SPSS列出了道琼斯工业指数和德国DAX指数、伦敦金融时报指数等其他五类指数的Pearson (皮尔森) 相关系数表。从Pearson (皮尔森) 相关系数大小看到, 受美国股市影响强弱大小的其他国家股市分别为: 新加坡、德国、英国和日本。可若从系数值看到, 其他国家股市受美国股市影响都很大, 说明它们的协同运动特征很显著。

## (2) 非参数相关系数表

非参数相关系数表列出了这些股票指数的Kendall (肯德尔) 和Spearman相关系数, 它们系数值概率P值也远小于显著性水平。

# 问题三输出结果



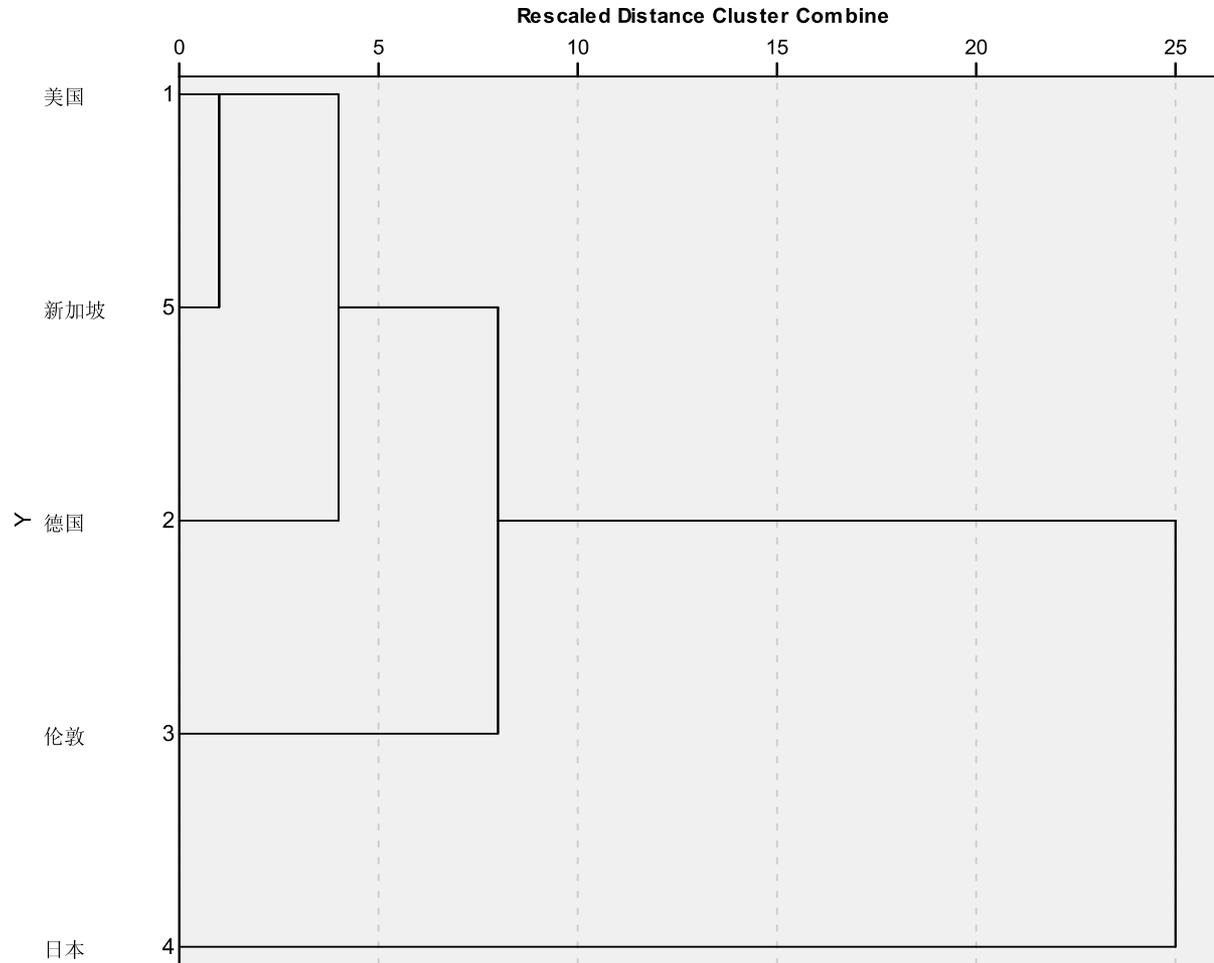
Stage	Cluster Combined		Coefficients	Stage Cluster First Appears		Next Stage
	Cluster 1	Cluster 2		Cluster 1	Cluster 2	
1	1	5	64.980	0	0	2
2	1	2	85.258	1	0	3
3	1	3	107.802	2	0	4
4	1	4	213.213	3	0	0

聚类过程表

# 树形图



Dendrogram using Average Linkage (Between Groups)





## 第18章 SPSS在心理学中的应用

# 18.1 实例提出：大学生心理问题研究



CONCEPT  
STRATE

大学生良好心理素质的培养与心理健康教育不仅关系到社会主义高等教育能否培养出身心健康、人格健全、全面发展、适应社会主义市场经济要求、能适应新世纪挑战的新型人才，而且关系到全民族素质的提高。

某大学对该校学生的心理健康状况进行了问卷调查分析。请利用这些资料和数据18-1.sav分析以下问题：

问题一：请你对调查问卷进行信度分析。

问题二：请综合评价大学生的心理健康状况。

问题三：请分析独生子女、系别对大学生的心理健康是否有显著影响。



## 18.2 实例的SPSS软件操作详解

### 1 问题一操作详解

问题一要求你对调查问卷进行信度分析，即对问卷的稳定性和可靠性进行有效分析。它反映了测量工具所得到的结果的一致性 or 稳定性，是被测特征真实程度的指标。因此可以利用SPSS中的信度分析功能来实现。

# 问题一的具体操作步骤



## Step01: 打开数据文件

打开数据文件18-1.sav。单击工具栏中的【Variable View(变量视图)】按钮，检查各个变量的数据结构定义是否合理，是否需要修改调整。

## Step02: 信度分析

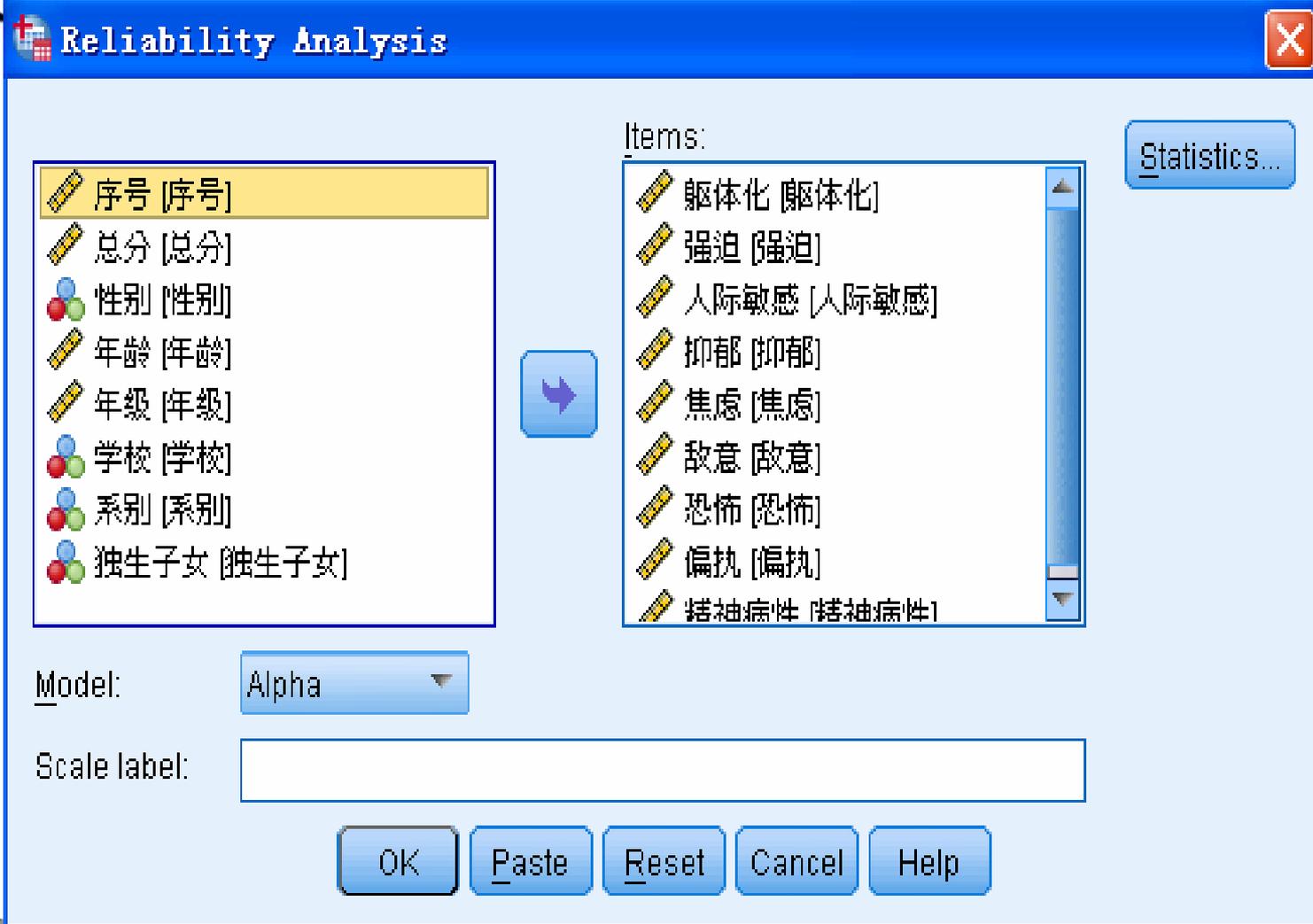
选择菜单栏中的【Analyze(分析)】→【Scale(度量)】→【Reliability Analysis(可靠性分析)】命令，弹出【Reliability Analysis(可靠性分析)】对话框。在左边的列表框中选择“躯体化”、“强迫”等九项因子作为分析对象，将其添加至右侧的【Items(项目)】列表框中。

单击【Statistics】按钮，弹出【Reliability Analysis:Statistics(可靠性分析:统计量)】对话框，并在【Descriptives for(描述性)】选项组中选择【Item(项)】选项，在【Inter-item(项之间)】选项组中选择【Correlations(相关性)】选项，再单击【Continue】按钮，返回主对话框。

最后单击【OK】按钮，完成本部分操作。

# 问题一的具体操作步骤

CONCEPT  
STRATE



# 问题二操作详解

CONCEPT  
STRATE

## Step01: 打开数据文件

打开数据文件18-1.sav，选择菜单栏中的【Analyze(分析)】→【Data Reduction(降维)】→【Factor(因子分析)】命令，弹出【Factor Analysis(因子分析)】对话框。

## Step02: 选择因子分析变量

在候选变量列表框中选择躯体化、强迫症状、人际关系敏感等九项因子设定为因子分析变量，将其添加至【Variables(变量)】列表框中，如图18-2所示。

## Step03: 选择因子旋转方法

在【Factor Analysis(因子分析)】对话框中，单击【Rotation】按钮，勾选【Varimax(旋转)】复选框，其他选项保持系统默认，单击【Continue】按钮返回主对话框。

## 问题二操作详解



### Step04: 选择因子得分

在【Factor Analysis(因子分析)】对话框中，单击【Score】按钮，勾选【Save as variables(保存为变量)】复选框，表示采用回归法计算因子得分并保持在原文件中；同时勾选【Display factor score coefficient matrix(显示因子得分系数矩阵)】复选框，表示输出因子得分系数矩阵。其他选项保持系统默认，单击【Continue】按钮返回主对话框。

### Step05: 其他选项选择

在【Factor Analysis(因子分析)】对话框中，单击【Options】按钮，勾选【Coefficient Display Format(系数显示格式)】选项组中的【Sorted by size(按大小排序)】复选框，表示将载荷系数按其大小排列构成矩阵。其他选项保持系统默认，单击【Continue】按钮返回主对话框。

### Step06: 单击【OK】按钮，完成操作。

# 问题二操作详解



# 问题三操作详解

CONCEPT  
STRATE

**Step01:** 对独生子女变量的影响性进行两独立样本t检验

选择菜单栏中的【Analyze(分析)】→【Compare Means(比较均值)】→【Independent-Sample T Test(单样本T检验)】命令，在弹出的对话框的候选变量列表框中选择检验变量“总分”，将其添加至【Test Variable(s)(检验变量)】列表框中。选择分组变量“独生子女”，将其添加至【Grouping Variable(s)(组变量)】文本框中。

接着，单击【Define Groups】按钮，弹出【Define Group(定义组)】对话框。点选【Use specified values(用特殊值)】单选钮，在【Group1(组1)】文本框中输入0，在【Group2(组2)】文本框中输入1。输入完成后，单击【Continue】按钮返回主对话框。

最后，单击【OK】按钮，完成操作。

# 问题三操作详解

CONCEPT  
TRATE

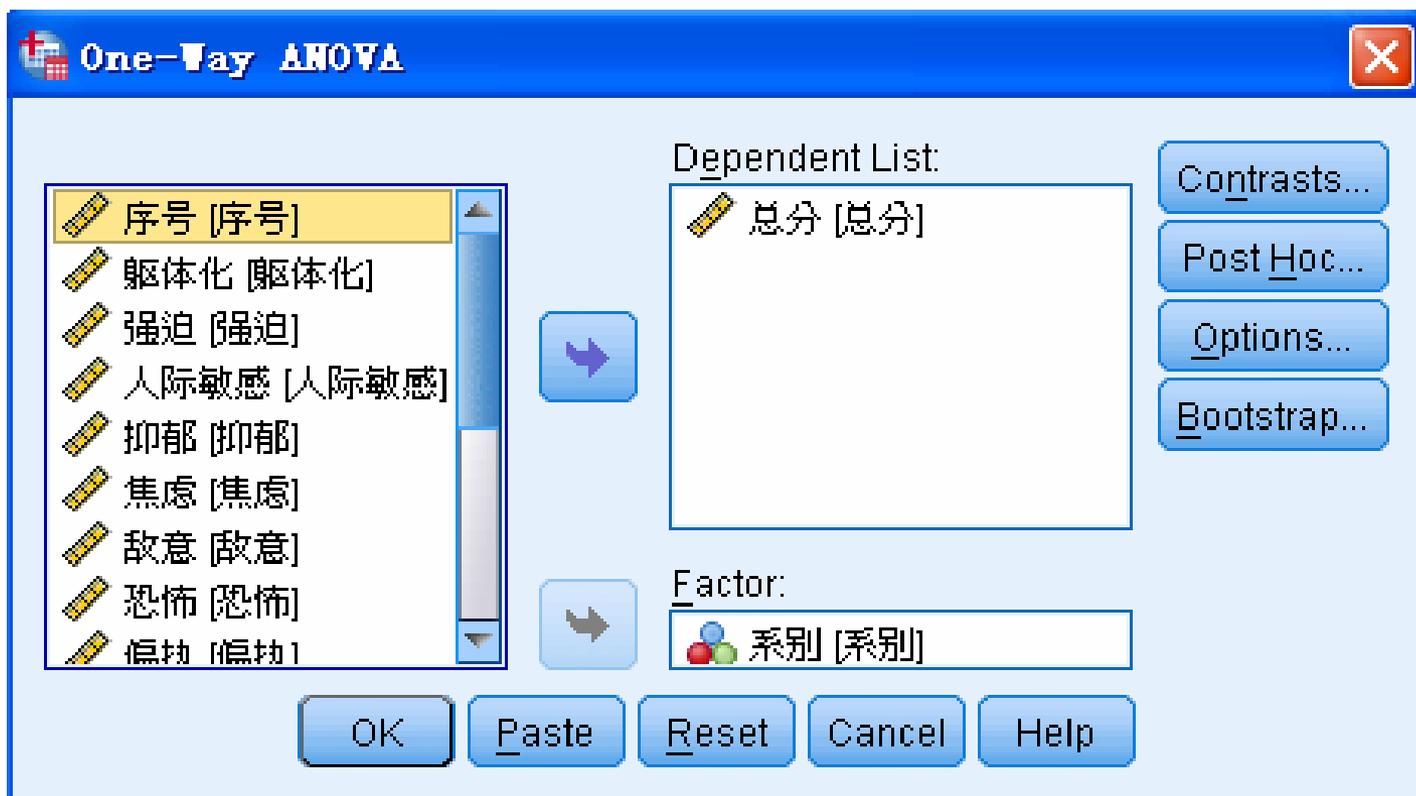
Step02: 对系别变量的影响性进行方差分析检验

选择菜单栏中的【Analyze(分析)】 → 【Compare Means (比较均值)】 → 【One-Way ANOVA(单因素ANOVA)】命令，弹出【One-Way ANOVA(单因素ANOVA)】对话框。

在候选变量列表框中选择“总分”变量作为因变量，将其添加至【Dependent List(因变量列表)】列表框中。在候选变量列表框中选择“系别”变量作为水平值，将其添加至【Factor(因子分析)】列表框中。

最后，单击【OK】按钮，完成操作。

# 问题三操作详解





## 18.3 实例的SPSS输出结果详解

### 问题一输出结果详解

#### (1) 评估因子的基本描述性统计量

下表所示是信度分析的评估因子的基本描述性统计量。表中给出了所有因子的均值、标准差以及参与分析的个案数。可以看到，“躯体化”、“抑郁”因子的平均评价得分最高，“人际敏感”和“精神病性”因子的平均得分最低。

# 问题一输出结果详解



	Mean	Std. Deviation	N
躯体化	43.82	36.924	305
强迫	38.61	11.906	305
人际敏感	35.02	13.689	305
抑郁	42.89	24.028	305
焦虑	36.34	15.782	305
敌意	37.65	12.731	305
恐怖	36.09	15.868	305
偏执	39.68	12.000	305
精神病性	35.89	15.407	305

# 问题一输出结果详解

CONCEPT  
STRATE

## (2) 评估因子的相关系数矩阵

表18-3所示是评估因子的相关系数矩阵。可以看到，除了躯体化和抑郁两个因子外，SCL-90其余各个因子的相关都在0.6以上，表明SCL-90在本研究中一定程度上具有较好的内容效度和结构效度。



# 问题一输出结果详解

	躯体化	强迫	人际敏感	抑郁	焦虑	敌意	恐怖	偏执	精神病性
躯体化	1.000	0.222	0.173	0.149	0.193	0.228	0.191	0.165	0.163
强迫	0.222	1.000	0.795	0.208	0.765	0.711	0.687	0.781	0.726
人际敏感	0.173	0.795	1.000	0.196	0.828	0.769	0.773	0.790	0.806
抑郁	0.149	0.208	0.196	1.000	0.242	0.231	0.196	0.193	0.244
焦虑	0.193	0.765	0.828	0.242	1.000	0.775	0.831	0.812	0.804
敌意	0.228	0.711	0.769	0.231	0.775	1.000	0.718	0.770	0.768
恐怖	0.191	0.687	0.77315	0.196	0.831	0.718	1.000	0.763	0.780
偏执	0.165	0.781	.790	0.193	0.812	0.770	0.763	1.000	0.750
精神病性	0.163	0.726	0.806	0.244	0.804	0.768	0.780	0.750	1.000

# 问题一输出结果详解

## (3) 信度分析的克隆巴哈 $\alpha$ 系数

克隆巴哈（Cronbach） $\alpha$  系数度量信度分析的一种重要方法。本实例中的系数值根据表18-4给出。表中不仅给出了克隆巴哈  $\alpha$  系数，还给出了评价因子的标准化  $\alpha$  系数。由于信度系数等于0.820，因此总体上该调查评估表的编制的内在信度是比较理想的。

Cronbach's Alpha	Cronbach's Alpha Based on Standardized Items	N of Items
0.820	0.911	9

# 问题二输出结果详解



## (1) 因子分析共同度

下表是因子分析的共同度，显示了所有变量的共同度数据。第二列列出了按指定提取条件提取特征根时的共同度。可以看到，所有变量的70%以上的信息可被因子解释，这些变量信息丢失较少。



## 问题二输出结果详解

	Initial	Extraction
躯体化	1.000	0.617
强迫	1.000	0.756
人际敏感	1.000	0.845
抑郁	1.000	0.532
焦虑	1.000	0.860
敌意	1.000	0.771
恐怖	1.000	0.783
偏执	1.000	0.817
精神病性	1.000	0.806

# 问题二输出结果详解

CONCEPT  
RATE

## (2) 因子分析的总方差解释

接着下表计算得到相关系数矩阵的特征值、方差贡献率及累计方差贡献率结果如表。根据特征值准则（取特征值大于等于1的主成分作为初始因子），应该选取两个因子。它们累积时解释了数据中总方差的75.5%。结果表明，第一个因子为主因子即可，因为它解释了原有六个变量总方差的84.449%。



## 问题二输出结果详解

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	5.757	63.971	63.971	5.757	63.971	63.971
2	1.030	11.446	75.416	1.030	11.446	75.416
3	.855	9.501	84.918			
4	.338	3.753	88.671			
5	.282	3.139	91.810			
6	.243	2.702	94.512			
7	.180	1.999	96.511			
8	.167	1.854	98.365			
9	.147	1.635	100.000			

## 问题二输出结果详解

CONCEPT  
TRATE

### (3) 旋转前因子载荷矩阵

下表显示了旋转前因子载荷矩阵。通过载荷系数大小可以看到不同公共因子所反映的主要指标的区别。通过载荷系数大小可以分析不同公共因子所反映的主要指标的区别。从结果看，大部分因子解释性较好，但是仍有少部分指标解释能力较差，例如躯体化因子，因此需要进行因子旋转。



## 问题二输出结果详解

	Component	
	1	2
焦虑	0.926	-0.055
人际敏感	0.914	-0.101
偏执	0.898	-0.105
精神病性	0.896	-0.063
恐怖	0.881	-0.075
敌意	0.878	-0.003
强迫	0.869	-0.028
躯体化	0.260	0.741
抑郁	0.292	0.668

## 问题二输出结果详解

CONCEPT  
STRATE

### (4) 旋转后的因子载荷矩阵

下表显示了实施因子旋转后的载荷矩阵。在进行因子旋转时采用的是正交旋转中的方差最大法，这便于对因子进行解释。可以看到，第一主因子除了在“躯体化”和“抑郁”等两个因子上载荷系数较小外，其他因子的载荷都较大，因此可以将它命名为态度公因子。相反的，第二主因子在“躯体化”和“抑郁”载荷上系数较大，可以将它命名为躯体化和抑郁因子。此时，各个因子的含义更加突出。

根据因子得分的大小顺序为焦虑>人际敏感>偏执>精神病性>恐怖>敌意>强迫>躯体化>抑郁。



## 问题二输出结果详解

	Component	
	1	2
焦虑	0.915	0.151
人际敏感	0.914	0.102
偏执	0.899	0.096
精神病性	0.888	0.136
恐怖	0.876	0.121
敌意	0.857	0.190
强迫	0.854	0.164
躯体化	0.090	0.780
抑郁	0.138	0.716



# 问题二输出结果详解

## (5) 因子得分系数

下表列出了经VARIMAX旋转后的因子值系数的回归估计值。因子值系数乘以对应变量的标准化值就是因子值。

	Component	
	1	2
躯体化	-0.115	0.712
强迫	0.153	0.007
人际敏感	0.177	-0.061
抑郁	-0.093	0.644
焦虑	0.169	-0.016
敌意	0.149	0.030
恐怖	0.165	-0.038
偏执	0.175	-0.065
精神病性	0.165	-0.026



# 问题三输出结果详解

## (1) 基本统计信息汇总表

被调查者中独生子女和非独生子女人数分别为139和154。他们心理健康状况总分的均值分别为12.531和15.511，标准差等于1.063和1.250。虽然他们的数值有一定差异，但还需要进行统计检验分析这种差异的统计学意义。

	独生子女	N	Mean	Std. Deviation	Std. Error Mean
总分	否	139	58.28	12.531	1.063
	是	154	58.23	15.511	1.250

# 问题三输出结果详解

CONCEPT  
RATE

## (2) 两总体均值的检验

在首先进行的方差相等假设检验中，F统计量等于0.295，对应的概率P值为0.588，大于显著性水平0.05，因此认为两组数据的方差是相等的。于是接着观察“Equal variance assumed”列所对应的t检验结果。由于t统计量对应的双尾概率P值为0.978，大于显著性水平0.05，因此认为两总体的均值不存在着显著差异。即大学生是否是独生子女对心理健康没有显著性影响。



# 问题三输出结果详解

		总分		
		Equal variances assumed	Equal variances not assumed	
Levene's Test for Equality of Variances	F	0.295		
	Sig.	0.588		
t-test for Equality of Means	t	0.028	0.029	
	df	291	287.563	
	Sig. (2-tailed)	0.978	0.977	
	Mean Difference	0.047	0.047	
	Std. Error Difference	1.658	1.641	
	95% Confidence Interval of the Difference	Lower	-3.217	-3.182
		Upper	3.311	3.276

# 问题三输出结果详解

CONCEPT  
RATE

## (3) 方差分析表

下表显示了方差分析表结果表。可以看到，心理健康得分总的离差平方总和为58987.532；不同系别的组间离差为1291.059；组内离差为57696.472；方差分析对应的F统计量的观测值为1.645，对应的概率P值为0.163。这里显著性水平为0.05，由于P值大于显著性水平0.05，所以接受零假设，认为不同系别的大学生心理健康没有显著性差异。



# 问题三输出结果详解

	Sum of Squares	df	Mean Square	F	Sig.
Between Groups	1291.059	4	322.765	1.645	0.163
Within Groups	57696.472	294	196.247		
Total	58987.532	298			