计量地理课程实验
指导书

瓦哈甫·哈力克

新疆大学资源与环境科学学院

前言 5
1. 实验简介..... 5
1.1 前期准备阶段..... 5
1.2 基本操作阶段..... 5
1.3 技术提高阶段 5
2. 课程实验目的要求..... 5
3. 实验需求的基本要求和设备..... 5
3.1 对学生的基础素质要求..... 5
3.2 实验的基本设备..... 5
4. 实验方式与基本要求..... 6
5. 实验的考核与实验报告..... 6
6. 适用专业..... 6
7. 实验所需的软件简介..... 6
实验 1 SPSS软件的安装与使用 7
1.1 [试验目的与要求]..... 7
1.2[实验步骤]..... 7
1.2.1 SPSS的安装步骤：..... 7
1.2.2 数据的输入和保存：..... 7
1.2.3 定义变量..... 8
1.2.4 数据导入..... 9
1.2.5 保存数据..... 9
试验 2　相关分析 10
2.1[实验目的与要求]..... 10
2.2 [实验内容]..... 10
2.3[实验步骤]..... 10
2.3.1 相关分析定义 10
2.3.2 定义变量，建立数据文件并输入数据 10
2.3.3　输出结果表..... 12
试验 3 回归分析 13
3.1[试验目的]..... 13
3.2[实验内容]..... 13
3.3[实验步骤]..... 13
3.3.1 回归分析定义。..... 13
3.3.2. 定义变量，建立数据文件并输入数据。..... 13
3.3.3 输出结果表..... 16
试验 4 时间序列分析..... 18
4.1 [实验目的]..... 18
4.2 [实验步骤]..... 18
4.2.1 时间分析的定义：..... 18
4.2.2 时间序列分析中自回归分析实例..... 18
4..2.3 自回归分析过程..... 18
4.2.4 自回归分析实例..... 19
4.3 输出结果..... 22
4.4　时间序列分析中季节分解法实例..... 23

4.4.1 季节分解法概述 23
4.4.2 季节分解法分析过程 23
4.4.3 季节分解法分析实例 24
4.5 输出结果 25
实验 5 系统聚类分析（ Hierarchical Cluster过程） 27
5.1 实验目的 27
5.2 实验原理 27
5.3 实例操作 27
5.3.1 数据准备 28
5.3.2 统计分析 28
5.4 输出结果 30
实验 6 主成分分析 33
6.1 [目的要求] 33
6.2 [实验内容] 33
6.3 [实验步骤] 33
6.3.1 主成分分析的概念 33
6.3.2 定义变量，建立数据文件并输入数据 33
实验 7 描述统计分析与探索统计分析 38
7.1[目的要求] 38
7.2[实验内容] 38
7. 3[实验步骤] 38
7.3.1 最简单的描述统计分析过程与实例 38
7.3.2 最简单的描述统计分析过程 38
7.3.3 应用实例 38
7.3.4 输出结果 40
7.4 最简单的探索统计分析过程与实例 40
7.4.1 探索分析概述 40
7.4.2 探索分析过程 41
7.4.3 探索分析实例 42
实验八 8 Excel在统计分析中应用 50
8.1 实验说明 50
8.2 实验目的与要求 50
8.3 实验步骤 50
8.3.1 描述统计工具 50
8.3.2 直方图工具 52
8.3.3 绘制两轴折线图 54
实验 9 Excel中统计函数的应用 58
9.1 实验目的与意义 58
9.2 基本原理和方法 58
9.3 实验内容及步骤 58
9.3.1 掌握Excel及其主要函数应用 58
9.3.2 常用函数的应用并实例 59

前言

计量地理是一门实践性很强的课程，针对该课程，特编写了《计量地理实验指导书》。通过实验，同学们能够对理论知识有更深的理解，并能应用到实际工作中去。

计量地理实验是与理论课同步进行的课程实验，是非独立开设的实验课。本实验中，有许多内容仍属于理论课内容的延伸。

1．实验简介

实验的软件为 SPSS,Excel 上机实验是计量地理课程的重要环节，它贯穿于整个《计量地理》课程教学过程中。本课程的实验分为三个阶段，其主要内容和基本要求为：

1.1 前期准备阶段

计量地理课程实验的第一阶段为前期准备阶段。在这一阶段主要目的是理解有关计量地理学的统计软件。

1.2 基本操作阶段

该阶段的主要任务是掌握 SPSS 的最基本功能包括数据管理，统计分析，图表分析，输出管理等等基本操作，并能够针对简单的实际问题提出解决方法，得到需要的结果。

1.3 技术提高阶段

技术提高阶段的实验，要求学生不仅要把课本上的内容掌握好，同时还需要自学一些相关的知识，包括表的属性、函数的应用、数据更新、统计图形的输出、输出设计。将理论知识与实际问题相结合。

2．课程实验目的要求

实验的主要目标是：

通过上机操作，加深对计量地理实践知识的理解。要求学生在理解相关的统计学原理和 SPSS 原理的基础上，能够掌握其操作方法进行统计分析，并且能够根据软件运行结果解释，论证假设。

3．实验需求的基本要求和设备

3.1 对学生的基础素质要求

学习过计算机文化基础、计算机技术基础、计算机制图、数据库及相关的专业课程，并希望同学在上实验课前要对上述课程进行复习，熟练运用这些学科的知识，勤于思考，认真实验和记录，将其与该课程融合贯通。实验要作到理论联系实际，实事求是。

3.2 实验的基本设备

软件：操作系统为 Microsoft Windows XP

4．实验方式与基本要求

(1)第一次实验前，任课教师向学生讲清实验的整体要求及实验的目标任务；平时考核内容、期末考试办法、实验守则及实验室安全制度；讲清上机操作的基本方法。

(2)《计量地理》课程是以理论课为主、实验为辅的课程。每次实验前：应当先弄清相关的理论知识，再预习实验内容、方法和步骤，避免出现盲目上机的行为。

(3)实验 1 人 1 组，在规定的时间内，学生独立完成，出现问题时，教师引导学生独立分析、解决，不得包办代替。

(4)该课程实验是连续的整体，需要有延续性。机房应有安全措施或学生自己配备一些常用的存储设备。避免前面的实验数据、程序被清除、改动影响后面的实验操作。

(5)实验前清点学生人数，实验中按要求做好实验情况及结果记录，实验后认真填写实验记录。

(6)学生最好能自备计算机，课下能多做操作练习，以便能够熟悉和精通实验方法。如果能结合实际课题进行训练，会达到更好的效果。

5．实验的考核与实验报告

《计量地理》课程采用理论课和上机实验课综合评定成绩的方法计分，其中理论课占 70%，实验占 30%。上机实验采用平时实验和最后考核结合的方法评定成绩。

实验后，要认真填写实验报告，实验中所用数据环境和实验结果需要保存在作业盘上。任课教师对每个学生的上机实验结果和报告要批改。

6．适用专业

适用于资源与环境科学学院的地理信息系统、地理科学、资源环境与城乡规划管理及生态学等专业。

7．实验所需的软件简介

SPSS 原意为“Statistical Package for Social Science”，即“社会科学统计软件包”。2000 年 SPSS 公司将其英文全称改为“Statistical Product and Service Solutions”，意为“统计产品与服务解决方案”。

SPSS for windows 是在 SPSS/PC(for DOS)基础上发展起来的，是目前世界上流行的三大统计分析软件之一，除了适用于社会科学之外，还适用于自然科学各领域的统计分析。近年来，SPSS 为我国经济，工业，管理，医疗卫生，体育，心理，教育等领域的科研工作者广泛使用。

Microsoft Excel 是美国微软公司开发的 Windows 环境下的电子表格系统，它是目前应用最为广泛的办公室表格处理软件之一。自 Excel 诞生以来 Excel 历经了 Excel5.0、Excel95、Excel97 和 Excel2000 等不同版本。随着版本的不断提高，Excel 软件的强大的数据处理功能和操作的简易性逐渐走入了一个新的境界，整个系统的智能化程度也不断提高，它甚至可以在某些方面判断用户的下一步操作，使用户操作大为简化。Excel 具有强有力的数据库管理功能，丰富的宏命令和函数，强有力的决策支持工具，图表绘制功能，宏语言功能，样式功能，对象连接和潜入功能，连接和合并功能，这些特性，已使 Excel 称为现代办公软件重要的组成部分。

实验 1 SPSS 软件的安装与使用

1.1［试验目的与要求］

学会 SPSS 安装及进入，退出步骤，掌握 SPSS 界面主要菜单的功能
熟练掌握变量的定义，数据的输入，编辑，转换和保存

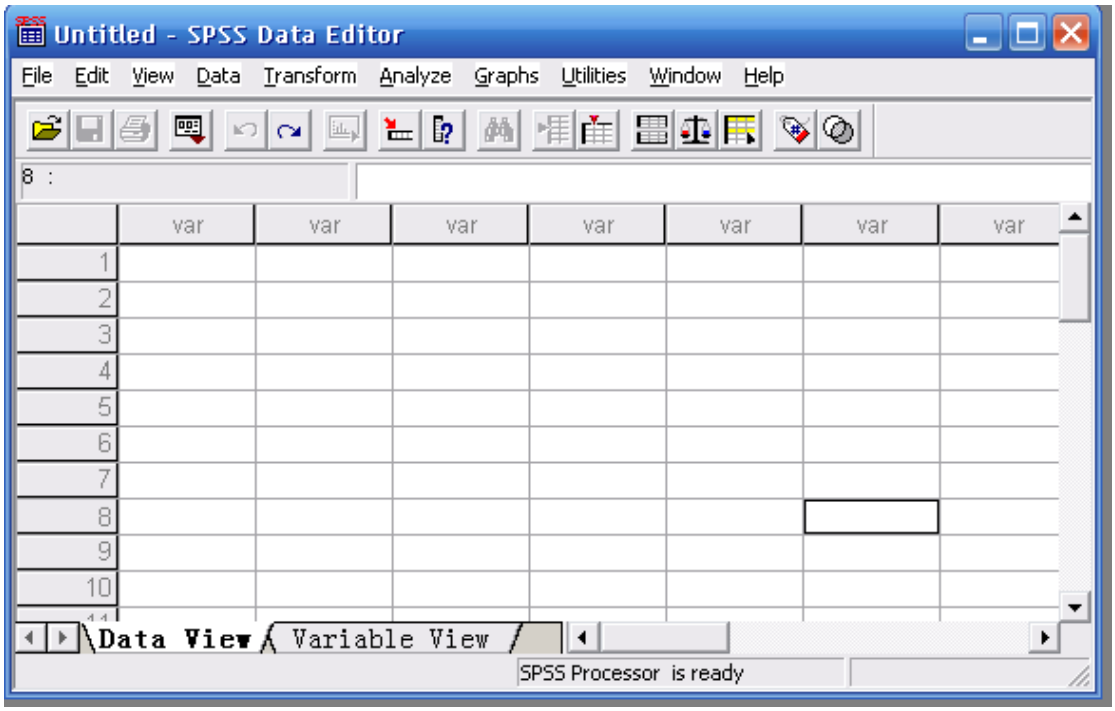
1.2[实验步骤]

1.2.1 SPSS 的安装步骤：

2.1.1.启动 Windows，找到 SPSS11.5，找到并双击即运行安装程序。
2.1.2 安装程序运行后，出现安装选项对话框，并点击 INSTALL SPSS，按提示安装。
2.1.3 指定安装的目标盘和安装文件的路径。
2.1.4 输入软件系列号码、用户姓名和单位名称。
在桌面上点击已安装好的软件图标进入 SPSS 界面。

1.2.2 数据的输入和保存：

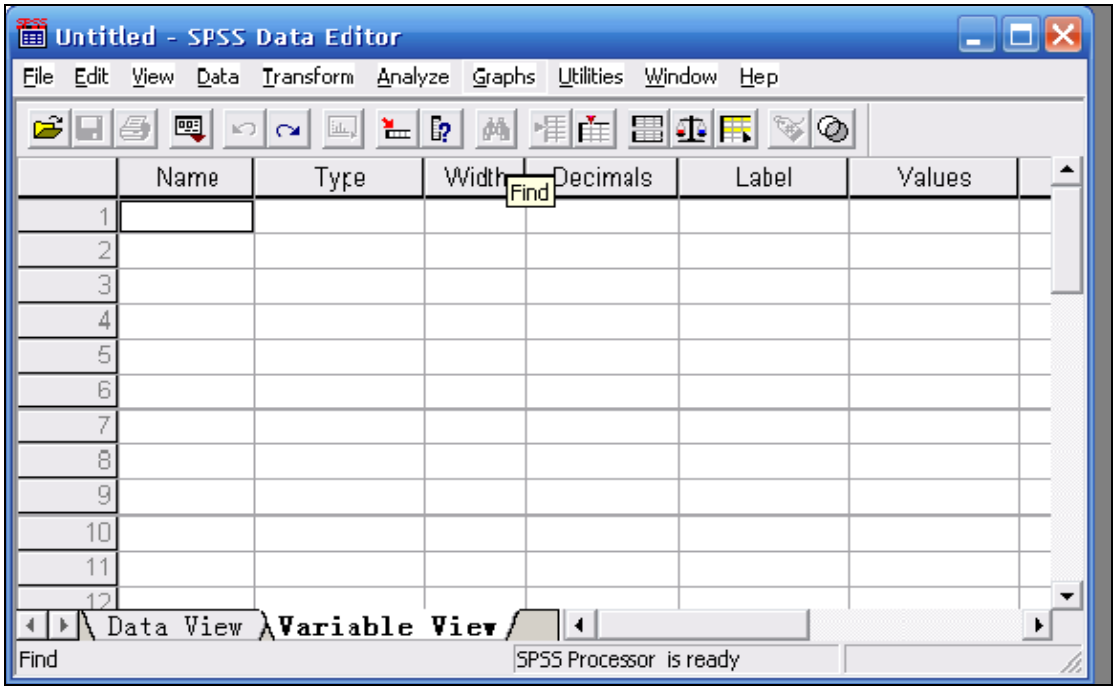
当打开 SPSS 后，展现在我们面前的界面如下：
具体包括文件操作，数据编辑，观察（视图-1.1），建立数据与数据整理，变量变换，统计分析，作图，实用程序，视窗控制和在线帮助 10 个部分。



将鼠标在上图中的各处停留，很快就会弹出相应部位的名称。
请注意窗口顶部显示为“SPSS Data Editor”，表明现在所看到的是 SPSS 的数据管理窗口。这是一个典型的 Windows 软件界面，有菜单栏、工具栏。特别的，工具栏下方的是数据栏，数据栏下方则是数据管理窗口的主界面。

1.2.3 定义变量

单击屏幕左下方的 Variable View（变量观察），得到：



建立数据文件的第 1 步是定义变量。在数据编辑窗左下角激活 Variable View(变量窗)，如图。定义变量有如下内容：变量名（Name），变量类型（Type），变量宽度（Width），保留小数位（Decimal），变量标签（Label），变量值标签（Values），缺失值（Missing），数据列宽（Columns），对齐方式（Align），度量类型（Measure）。

①变量名（Name）

在框中输入要定义的变量名称。若不定义，系统将依次默认为“var00001”、 “var00002”、“var00003”•••。

定义变量名应遵循以下原则：

• 变量名最长不超过 8 个字符（4 个汉字）

• 首字符必须是字母或汉字，不能以下划线“-”或圆点“.”结尾。

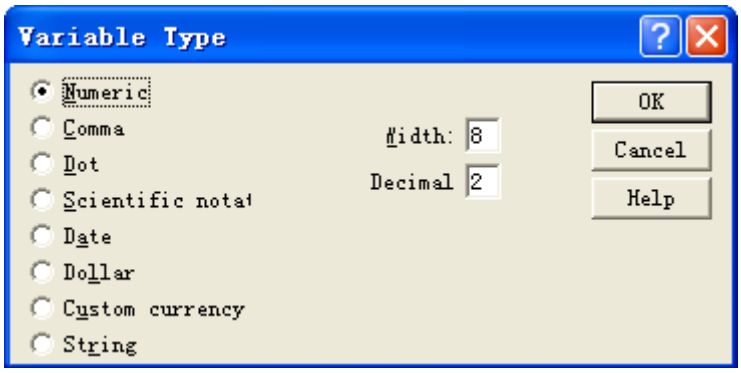
• 变量名中不能有空格或某些特殊符号，如“ !”、“ ？ ”和“ *”等。

• 变量名不能与 SPSS 的关键字相同，不能用 ALL、AND、BY、EQ、GE、GT、LE、LT、NE、NOT、OR、TO、WITH 等作变量名。

• 对变量名英文字母的大小写为作区分，一律显示小写字母。

②变量类型（Type），变量宽度（Width）及，保留小数位（Decimal）

当光标移到某个变量的变量类型单元格时，该单元格右方会显示一灰色按钮，单击该按钮，弹出 Variable Type（变量类型）对话框，有 8 种类型供选择。



• Width：变量宽度。数值型变量系统默认 8 位，小数点算作 1 一位。
• Decimal：保留小数位。数值型变量系统默认 2 位。
• Numeric：标准数值型变量。系统默认。
• Comma：逗号数值型变量。千进位用逗号分隔，小数与整数间用圆点分隔。
• Dot：圆点数值型变量。千进位用圆点分隔，小数与整数间用逗号分隔。
• Scientific notation：科学记数法。
• Date：日期型变量。对话框中列出 27 种日期型，既可以表示季、月、周、日，也可以表示时、分、秒及百分秒。“现以 1999 年 8 月 18 日” 为例，列举几种常用的输入与显示方式：

dd-mmm-yyy	18-AUG-1999
dd-mmm-yy	18-AUG-99
mm/dd/yyyy	08/18/1999
mm/dd/yy	08/18/99
dd.mm.yyyy	18.08.1999
dd.mm.yy	18.08.99

• Dollar：带美元符号的数值型变量。
• Custom currency：自定义变量。
• String：字符型变量。其标识作用明显，但不能参与运算。

一般说来，标准数值型变量、日期型变量和字符型变量 3 种类型最常用，系统默认标准数值型变量（Numeric）。

1.2.4 数据导入

选择菜单 File==>Open Data，将文件类型定义为已有的数据库，如 EXCEL（*.xls）再打开选取已有的数据，即将已有数据导入 SPSS 中。

1.2.5 保存数据

选择菜单 File==>Save，由于该数据从来没有被保存过，所以弹出 Save as 对话框，如下：单击保存类型列表框，可以看到 SPSS 所支持的各种数据类型，有 DBF、FoxPro、EXCEL、ACCESS 等，这里我们仍然将其存为 SPSS 自己的数据格式（*.sav 文件）。

试验 2　相关分析

2.1[实验目的与要求]

本试验主要是引导学生掌握利用 SPSS 软件进行相关分析的基本方法，包括简单相关分析，偏相关分析和其其它相关系数的计算。

2.2［实验内容]

2.2.1 两变量的相关分析（Bivariate 过程）

2.2 .2 偏相关分析(Partial 过程)

2.3[实验步骤]

2.3.1 相关分析定义

当分析两个变量之间是否存在相关关系时，可采用双变量相关分析（Bivariate）。

Bivariate 过程 此过程用于进行两个/多个变量间的参数/非参数相关分析，如果是多个变量，则给出两两相关的分析结果。这是 Correlate 子菜单中最为常用的一个过程，实际上我们对他的使用可能占到相关分析的 95%以上。下面的讲述也以该过程为主。

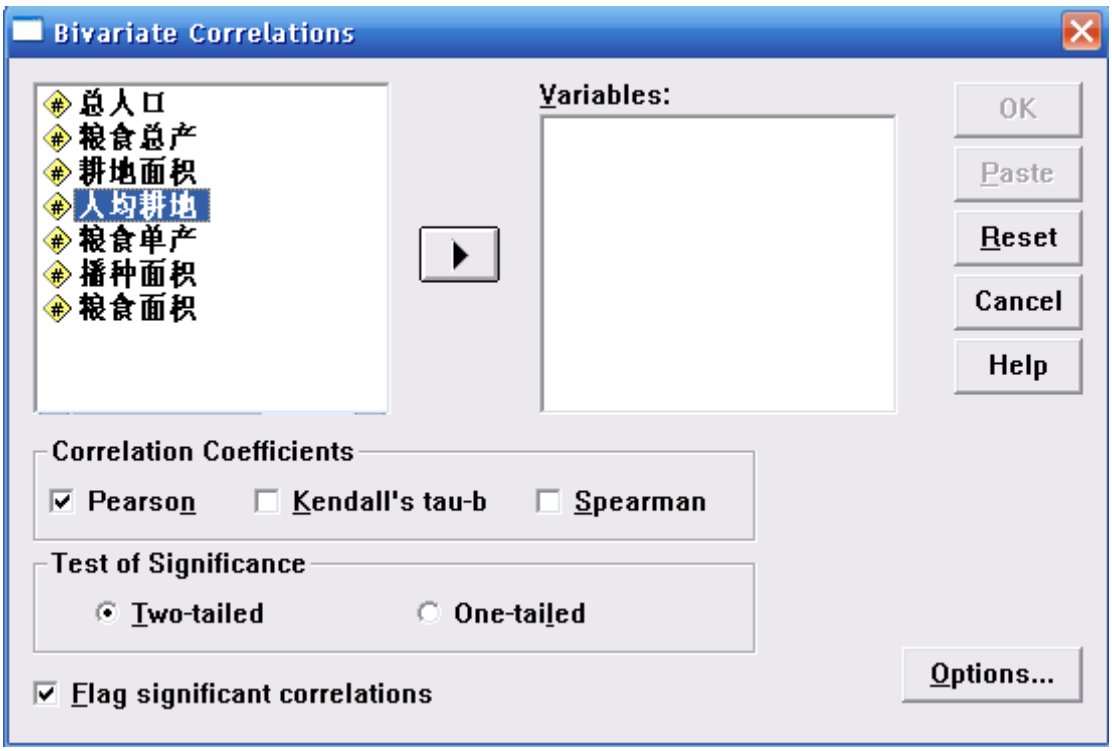
Partial 过程 如果需要进行相关分析的两个变量其取值均受到其他变量的影响，就可以利用偏相关分析对其他变量进行控制，输出控制其他变量影响后的相关系数，这种分析思想和协方差分析非常类似。Partial 过程就是专门进行偏相关分析的。

Distances 过程 调用此过程可对同一变量内部各观察单位间的数值或各个不同变量间进行距离相关分析，前者可用于检测观测值的接近程度，后者则常用于考察预测值对实际值的拟合优度。该过程在实际应用中用的非常少。

双变量相关分析中，对于双变量正态分布资料，可选择积矩相关系数（Person 相关系数）；对于非双变量正态分布资料，可选择等级相关系数（Spearman 相关系数）和 Kendall 相关系数等非参数方法。该过程还可以给出基本统计量均数和标准差等统计量。

2.3.2 定义变量，建立数据文件并输入数据

在主菜单中点击 Analyze==>Correlate==>Bivariate，进行双变量相关分析界面：



界面说明

【Variables 框】

用于选入需要进行相关分析的变量，至少需要选入两个。

【Correlation Coefficients（相关分析系数）复选框组】

用于选择需要计算的相关分析指标，有：

- Pearson 复选框 选择进行积距相关分析，即最常用的参数相关分析
- Kendall's tau-b 复选框 计算 Kendall's 等级相关系数
- Spearman 复选框 计算 Spearman 相关系数，即最常用的非参数相关分析（秩相关）

【Test of Significance（显著性检验）单选框组】

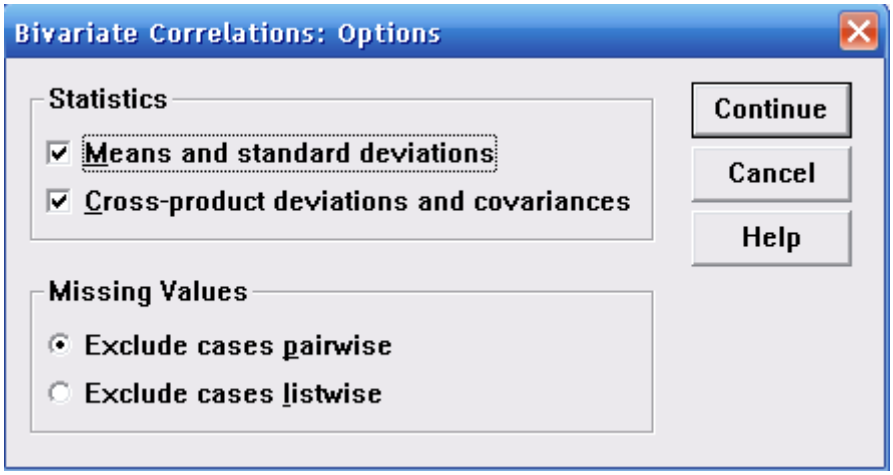
用于确定是进行相关系数的单侧（One-tailed）或双侧（Two-tailed）检验，一般选双侧检验。

【Flag significant correlations】标出有显著性意义的相关系数

用于确定是否在结果中用星号标记有统计学意义的相关系数，一般选中。此时 P<0.05 的系数值旁会标记一个星号，P<0.01 的则标记两个星号。

【Options 钮】

弹出 Options 对话框，选择需要计算的描述统计量和统计分析：



Statistics 复选框组 可选的描述统计量。它们是：

1. Means and standard deviations 每个变量的均数和标准差

2. Cross-product deviations and covariances 各对变量的交叉积和以及协方差阵

Missing Values 单选框组 定义分析中对缺失值的处理方法，可以是具体分析用到的两个变量有缺失值才去除该记录（Exclude cases pairwise），或只要该记录中进行相关分析的变量有缺失值（无论具体分析的两个变量是否缺失），则在所有分析中均将该记录去除(Excludes cases listwise)。默认为前者，以充分利用数据。

单击【Continue】==>【OK】,即得到结果，

2.3.3 输出结果表

Correlations

		人均耕地	耕地面积
人均耕地	Pearson Correlation	1	.887**
	Sig. (2-tailed)	.	.001
	Sum of Squares and Cross-products	.001	79.527
	Covariance	.000	8.836
	N	10	10
耕地面积	Pearson Correlation	.887**	1
	Sig. (2-tailed)	.001	.
	Sum of Squares and Cross-products	79.527	9029758
	Covariance	8.836	1003306
	N	10	10

**. Correlation is significant at the 0.01 level (2-tailed).

我们可以看到人均耕地和耕地面积的相关系数为 0. 887，且在 P<0.05 时经过双尾检验相关系数是显著的。

试验 3 回归分析

3.1[试验目的]

本试验主要是引导学生掌握利用 SPSS 软件进行回归分析的基本方法，包括一元线性回归分析，多元线性回归分析，包含虚拟变量的线性回归分析，曲线参数估计法，二值多元 Logistic 回归分析。特别是，学生应掌握在 SPSS 软件中进行多元线性回归方法和曲线参数的估计方法。

3.2[实验内容]

线性回归分析(Linear 过程)

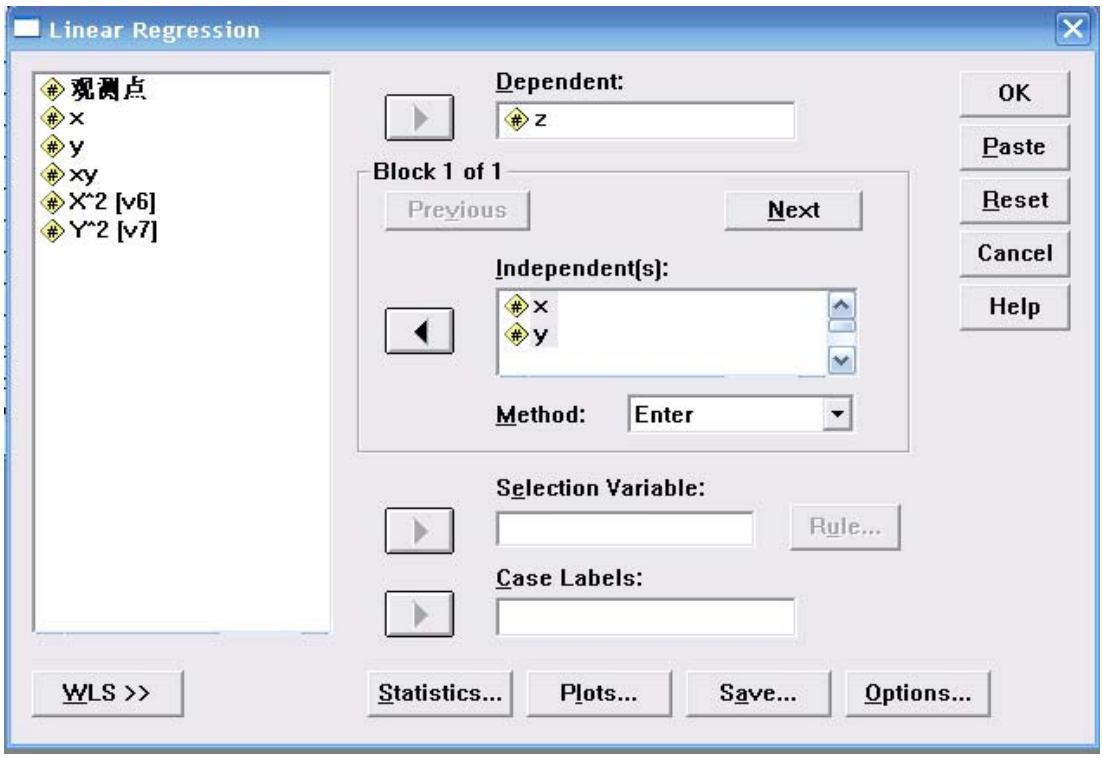
3.3[实验步骤]

3.3.1 回归分析定义。

回归分析（Regression）是研究一个自变量或多个自变量与一个因变量（Dependent）之间是不存在某种线性关系或非线性关系的一种统计学分析方法。而线性回归分析（Linear Regression）是研究一个或多个自变量（independent）与一个因变量之间是否存在某种线性关系的统计学方法。

3.3.2. 定义变量，建立数据文件并输入数据。

在菜单中选择 Analyze==>Regression==>liner，系统弹出线性回归对话框如下：



【Dependent 框】

用于选入回归分析的应变量。

【Block 按钮组】

由 Previous 和 Next 两个按钮组成，用于将下面 Independent 框中选入的自变量分组。由于多元回归分析中自变量的选入方式有前进、后退、逐步等方法，如果对不同的自变量选入的方法不同，则用该按钮组将自变量分组选入即可。【Independent 框】

用于选入回归分析的自变量。

【Method 下拉列表】

用于选择对自变量的选入方法，有 Enter（强行进入法）、Stepwise（逐步法）、Remove（强制剔除法）、Backward（向后法）、Forward（向前法）五种。该选项对当前 Independent 框中的所有变量均有效。

【Selection Variable 框】

选入一个筛选变量，并利用右侧的 Rules 钮建立一个选择条件，这样，只有满足该条件的记录才会进入回归分析。

【Case Labels 框】

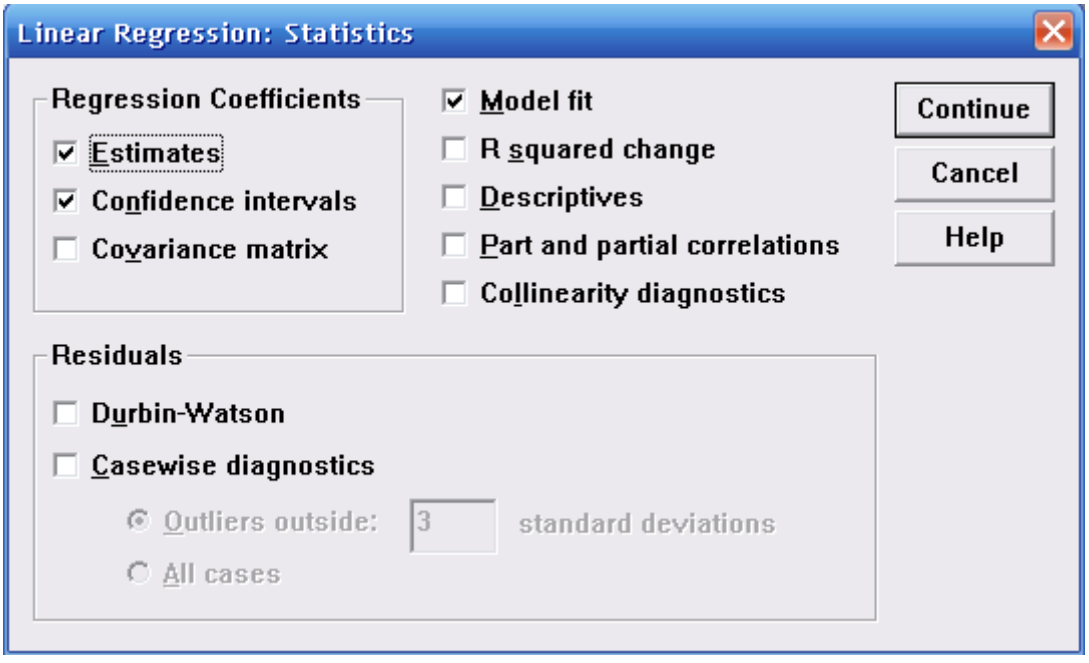
选择一个变量，他的取值将作为每条记录的标签。最典型的情况是使用记录 ID 号的变量。

【WLS>>钮】

可利用该按钮进行权重最小二乘法的回归分析。单击该按钮会扩展当前对话框，出现 WLS Weight 框，在该框内选入权重变量即可。

【Statistics 钮】

弹出 Statistics 对话框，用于选择所需要的描述统计量。有如下选项：



- Regression Coefficients 复选框组：定义回归系数的输出情况，选中 Estimates 可输出回归系数 B 及其标准误，t 值和 p 值，还有标准化的回归系数 beta；选中 Confidence intervals 则输出每个回归系数的 95%可信区间；选中 covariance matrix 则会输出各个自变量的相关矩阵和方差、协方差矩阵。以上选项默认只选中 Estimates。
- Residuals 复选框组：用于选择输出残差诊断的信息，可选的有 Durbin-Watson 残差序列相关性检验、超出规定的 n 倍标准误的残差列表。
- Model fit 复选框：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:，R，R2 和调整的 R2, 标准误及方差分析表。
- R squared change 复选框：显示模型拟合过程中 R2、F 值和 p 值的改变情况。
- Descriptives 复选框：提供一些变量描述，如有效例数、均数、标准差等，同时还给出一个自变量间的相关矩阵。
- Part and partial correlations 复选框：显示自变量间的相关、部分相关和偏相关系数。
- Collinearity diagnostics 复选框：给出一些用于共线性诊断的统计量，如特征根（Eigenvalues）、方差膨胀因子(VIF)等。

以上各项在默认情况下只有 Estimates 和 Model fit 复选框被选中。

【Plot 钮】

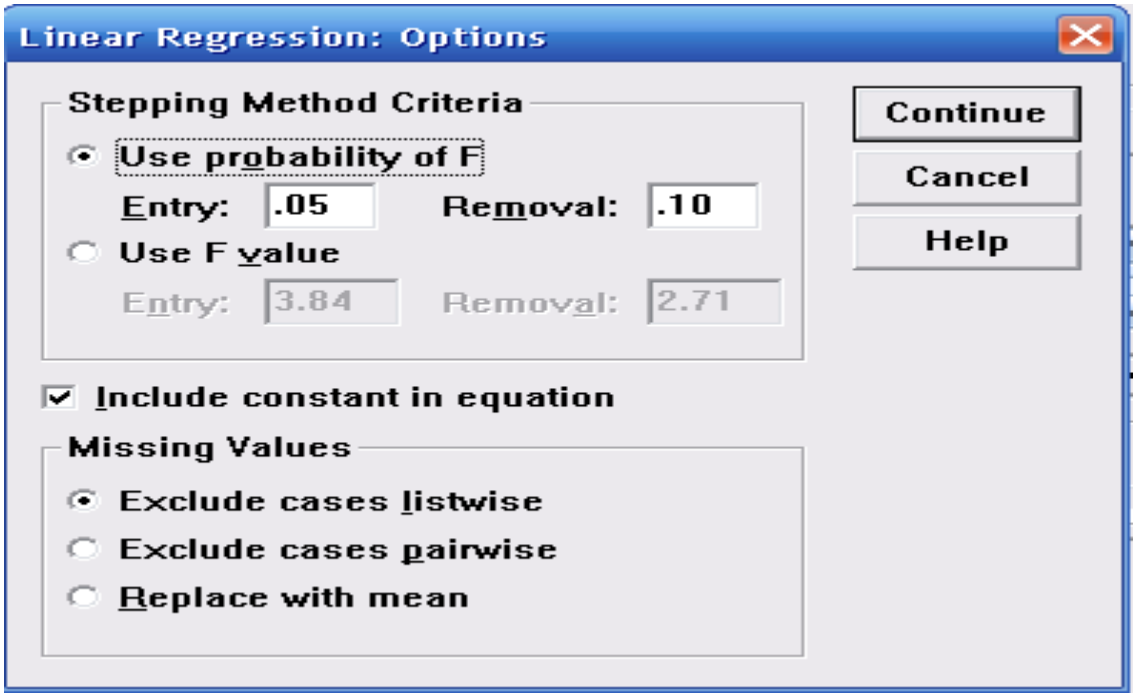
弹出 Plot 对话框，用于选择需要绘制的回归分析诊断或预测图。可绘制的有标准化残差的直方图和正态分布图，应变量、预测值和各自变量残差间两两的散点图等。

【Save 钮】

许多时候我们需要将回归分析的结果存储起来，然后用得到的残差、预测值等做进一步的分析，Save 钮就是用来存储中间结果的。可以存储的有：预测值系列、残差系列、距离（Distances）系列、预测值可信区间系列、波动统计量系列。下方的按钮可以让我们选择将这些新变量存储到一个新的 SPSS 数据文件或 XML 中。

【Options 钮】

设置回归分析的一些选项：



：

- Stepping Method Criteria 单选钮组：设置纳入和排除标准，可按 P 值或 F 值来设置。
- Include constant in equation 复选框：用于决定是否在模型中包括常数项，默认选中。
- Missing Values 单选钮组：用于选择对缺失值的处理方式，可以是不分析任一选入的变量有缺失值的记录（Exclude cases listwise）而无论该缺失变量最终是否进入模型；不分析具体进入某变量时有缺失值的记录（Exclude cases pairwise）；将缺失值用该变量的均数代替（Replace with mean）。

设置完成后，点击 OK，输出分析结果

3.3.3 输出结果表

Model Summary

Model	R	R Square	Adjusted R Square	Std. Error of the Estimate
1	.826[a]	.682	.629	8.25840

a. Predictors: (Constant), Y, X

ANOVA[b]

Model		Sum of Squares	df	Mean Square	F	Sig.
1	Regression	1754.275	2	877.137	12.861	.001[a]
	Residual	818.415	12	68.201		
	Total	2572.689	14			

a. Predictors: (Constant), Y, X

b. Dependent Variable: Z

Coefficients[a]

Model		Unstandardized Coefficients		Standardized Coefficients	t	Sig.
		B	Std. Error	Beta		
1	(Constant)	204.933	32.411		6.323	.000
	X	-1.316	.262	-.827	-5.028	.000
	Y	-.548	.397	-.227	-1.379	.193

a. Dependent Variable: Z

经过检验，得到回归系数，根据系数列出回归方程。B 分别为常数项，和 X、Y 的系数。

试验 4 时间序列分析

4.1［实验目的］

用 SPSS 统计软件学会建立时间序列新变量方法

4.2［实验步骤］

4.2.1 时间分析的定义：

时间序列，也叫时间数列或动态数列，是要素（变量）的数据按照时间顺序变动排列而形成的一种数列，它反映了要素（变量）随时间变化的发展过程。地理过程的时间序列分析，就是通过分析地理要素（变量）随时间变化的历史过程，揭示其发展变化规律，并对其未来状态进行预测。

4.2.2 时间序列分析中自回归分析实例

在描述实际中出现的某些问题时，一种非常有用的随机模型就是自回归模型（Autoregression）.在该模型中，过程的当前值被表示过程的有穷线性组合在加上一个重击 e_t.我们用$X_t,X_{t-1},X_{t-2},\ldots$，记在等间隔时间t,t-1,t-2,…上的过程值。此外，用$Z_t,Z_{t-1},Z_{t-2},\ldots$,记关于均值u的偏差，即$Z_t=X_t-u$。则：

$$Z_t=\phi_1 Z_{t-1}+\phi_2 Z_{t-2}+\ldots+\phi_p Z_{t-p}+e_t$$

便叫做为P阶自回归（AR）过程，当P=1时，称为一阶自回归模型。

4..2.3 自回归分析过程

1) 定义变量，建立数据文件并输入数据，至少要有一个变量。打开 Data 菜单中的 Define Dates 对话框，定义时间序列的周期。采用 Transform 菜单中的 Create Time Series 的方法，建立一个时间序列的新的变量。
2) 按 Analyze ⇒ Time series ⇒ Autoregression 顺序展开相应的对话框。
3) 选择一个因变量，将其移到 Dependent 框。选择一个或多个自变量移到 independent(s) 框。在 Media 栏中，从三种方法中选择一种预测方法。如果在回归方程中不需要包括常数项，可不选 Include constant in model 复选项。

4） 单击 Save 按钮展开保存对话框，在对话框中选择计算结果存放方式。

- 在 Create Variables 栏中给出
  - ✧ Add to file 选项，将新建变量存放在原数据文件中，是系统默认的。
  - ✧ Replace existing 选项，用新建变量数据替代数据文件中原先存在的计算结果。
  - ✧ Do not create 选项，在原数据文件中不建立新变量。
- % Confidence Interval 框，设定置信区间。可在 90，95 和 99 三个值中 选择其一，系统默认值是 95.

- The Estttimation Period 栏，此时显示给出多长时期的预测结果。

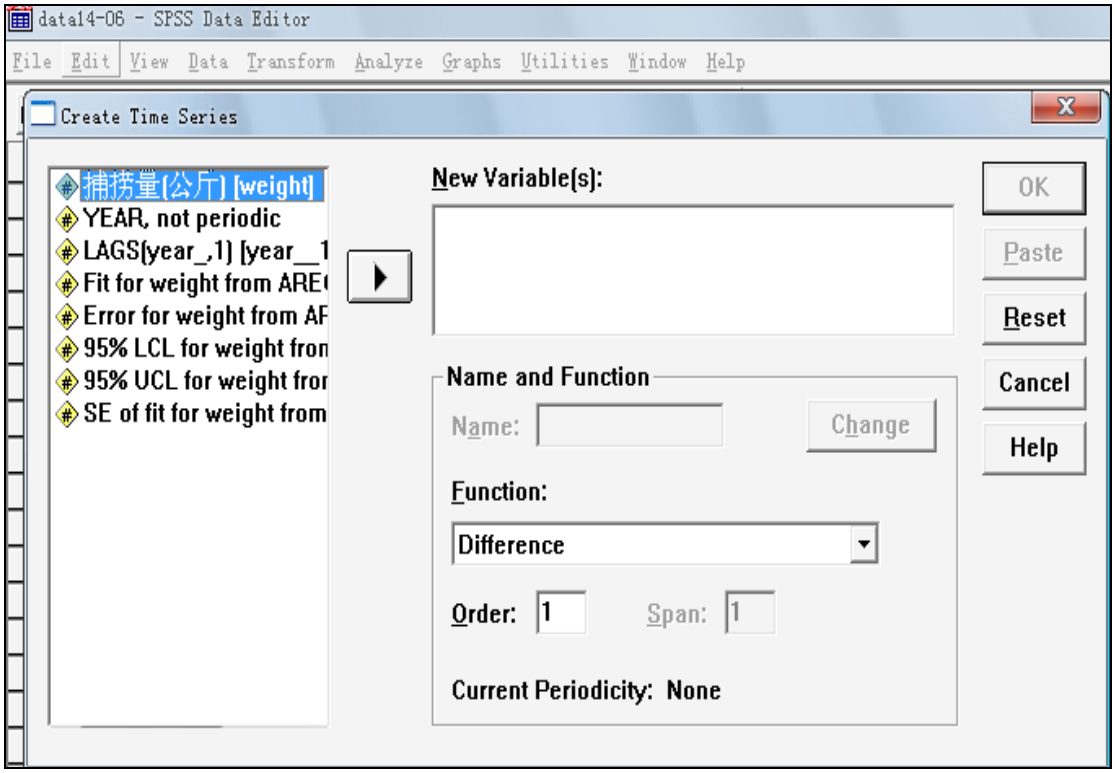
5） 单击 Options 按钮，展开选项对话框：

4.2.4 自回归分析实例

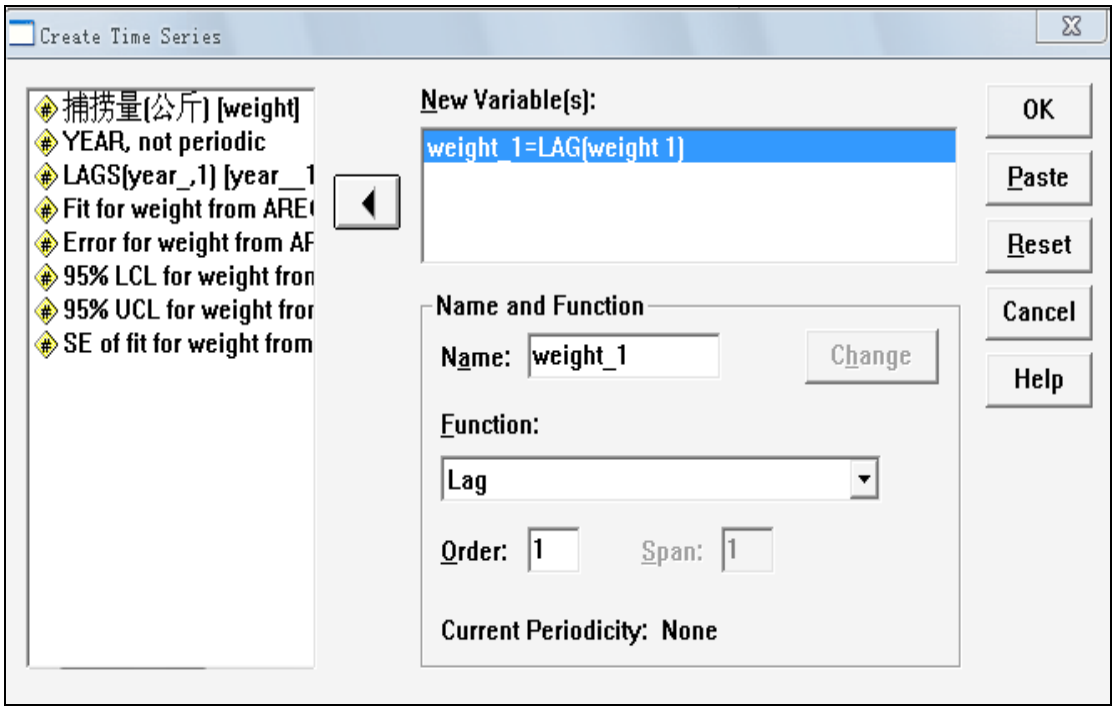
1）数据 Data 14-06,变量 weight 为某养鱼肠历年的年捕捞量。为提高经营管理水平，需建立自回归模型，预测 2002 年的捕捞量。

操作步骤：

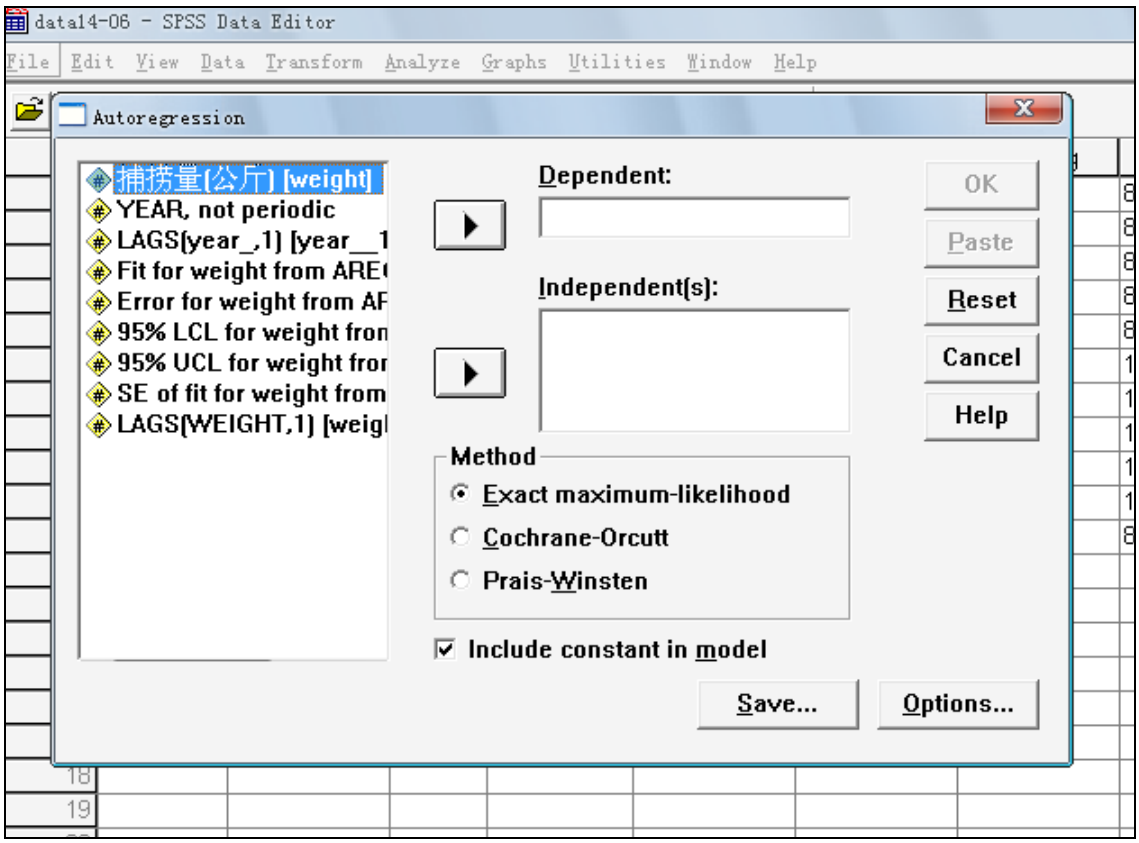
- 按 Transform ⇒Create Time Series 顺序展开 create Time Series 对话框，见图



- Function 框选择需要转换最初变量生成新变量的函数 Lag.
- 选择 weight 变量，将其移到 New Variable(s)框中。

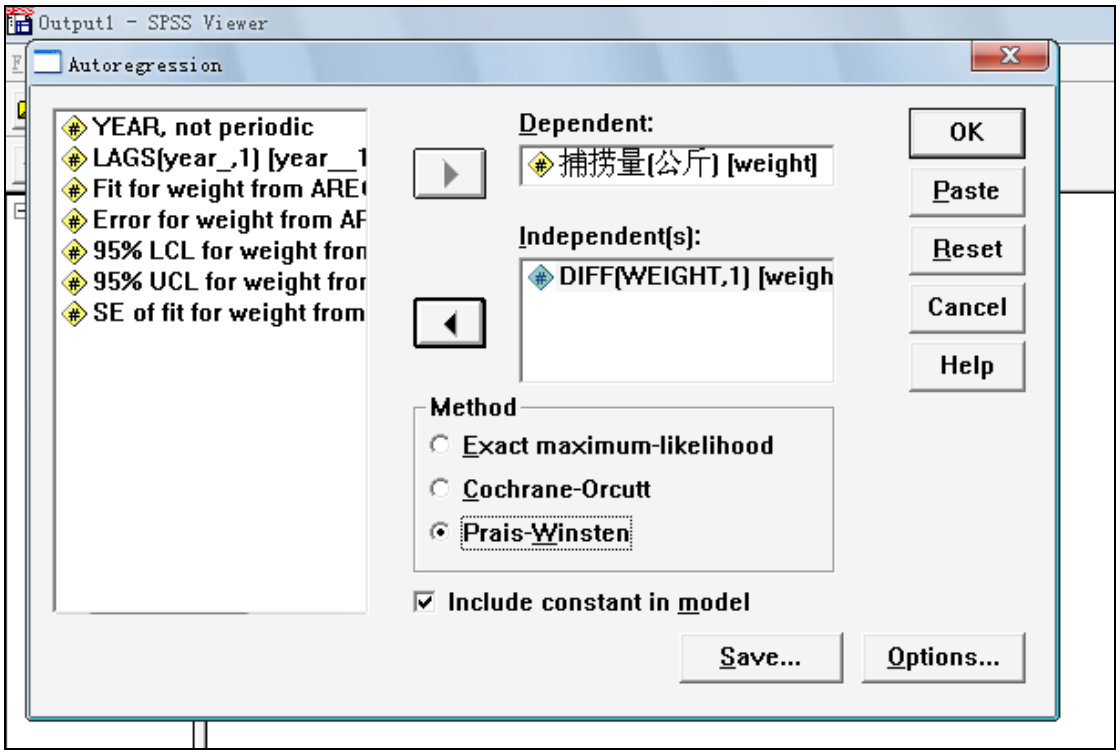


- 在 Name and Function 栏下的 Name 框中，采用系统定义的新变量名 wight-1.
- 单击 OK 按钮，系统运行，在数据窗口中生成 weight-1 滞后新变量。
- 按 Analyze⇒Time series⇒Autoregression 顺序展开如图所示的对话框。

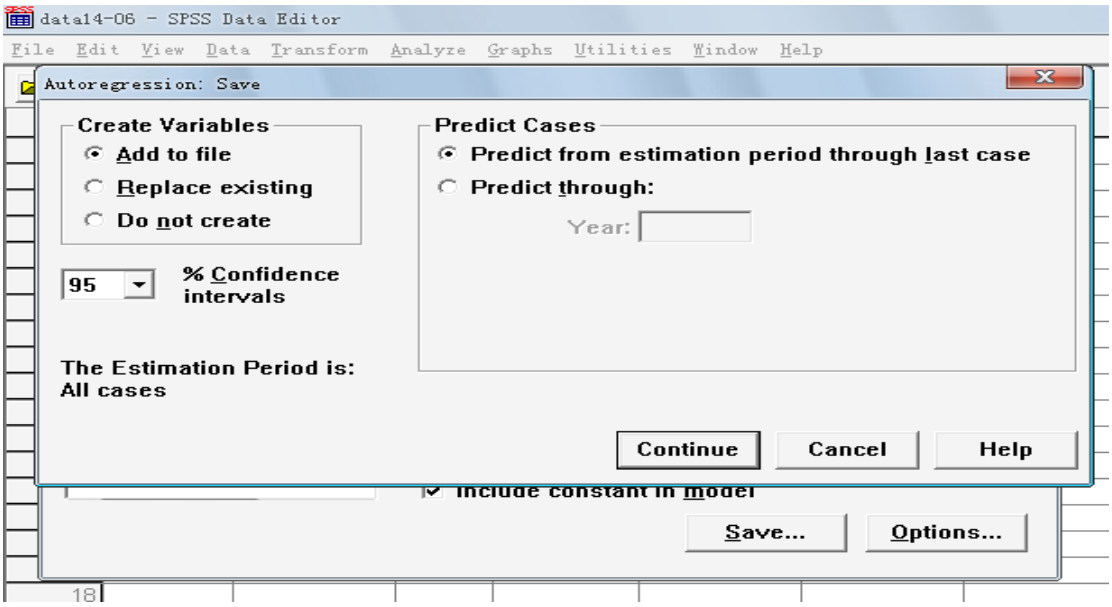


- 选择 Weight 作为因变量并将其移动 Dependent 框。选择通过 Lags(weight,1)转换生成的滞后变量 weight-1 作为自变量并将其移到 Independents 框，在 Method 栏中选择

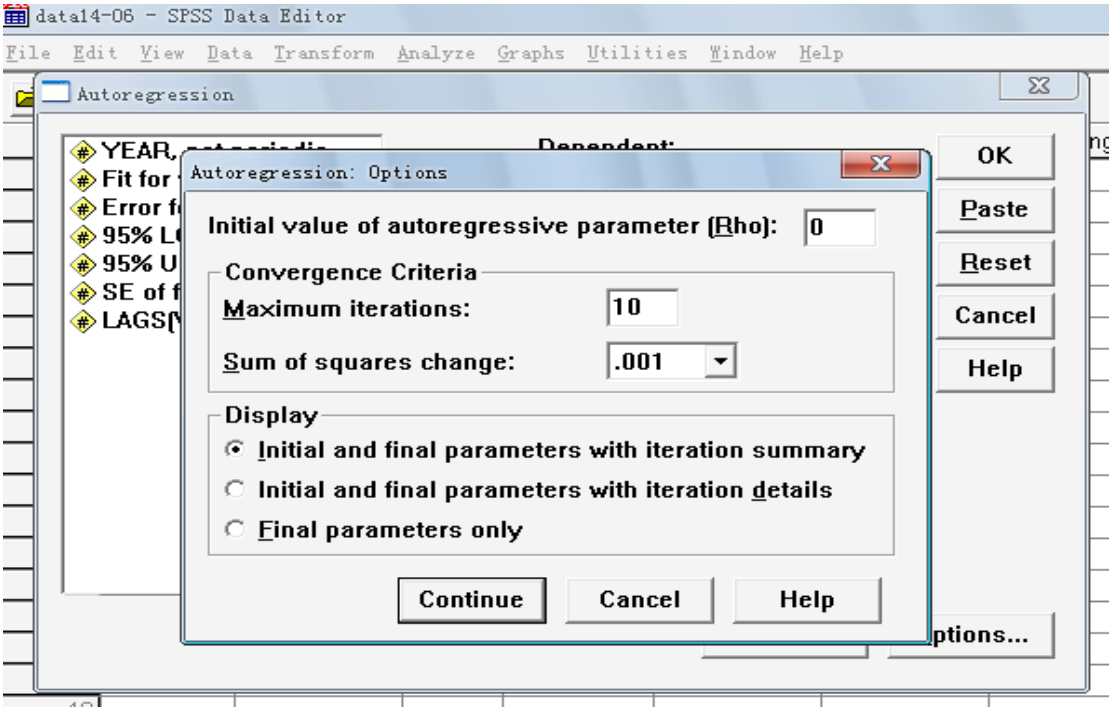
Prais-winsten 预测方法。



- 单击 Save 按钮展开保存对话框，如图所示。选择 Add to file 项。设定置信区间为系统默认值 95%。选择 Predict from estimation period through last case,给出全部年限的预测结果。



- 单击 Options 按钮，展开选择对话框，如图所示。在 Display 栏选择 Filnal Parameter only 项，只显示最终参数。其他各个选项均使用系统默认值。



- 单击 OK 按钮，执行运算。

4.3 输出结果

输出结果如图所示：

MODEL: MOD_2

Split group number: 1 Series length: 6

Number of cases skipped at beginning because of missing values: 1

Conclusion of estimation phase.

Estimation terminated at iteration number 0 because:

R-squared is one within working tolerance.

FINAL PARAMETERS:

Estimate of Autocorrelation Coefficient

Rho	0

Prais-Winsten Estimates

Multiple R	.99997569
R-Squared	.99995139
Adjusted R-Squared	.99993923
Standard Error	3.553549
Durbin-Watson	2.8177017

Analysis of Variance:

	DF	Sum of Squares	Mean Square
Regression	1	1039000.8	1039000.8
Residuals	4	50.5	12.6

Variables in the Equation:

	B	SEB	BETA	T	SIG T
WEIGHT_1	1.13461	.003956	.99997569	286.84392	.00000000
CONSTANT	-211.86179	13.054911	.	-16.22851	.00008436

The following new variables are being created:

Name	Label
FIT_3	Fit for WEIGHT from AREG, MOD_2
ERR_3	Error for WEIGHT from AREG, MOD_2
LCL_3	95% LCL for WEIGHT from AREG, MOD_2
UCL_3	95% UCL for WEIGHT from AREG, MOD_2
SEP_3	SE of fit for WEIGHT from AREG, MOD_2

在数据库中生成weight 的预测值Fit_1,预测值95%置信区间的下线LCL_1,95%置信区间的上限UCL_1 及SEP等 5 个新变量。从上面的表中方程式中的变量（Variables in the Equation）可知，所求的预测方程式为$Xt=1.13461X_{t-1}-221.86179$
因此，2002 的预测捕捞量为：$X_{2002}=1.13461\times4168-211.86179=4517.19$(公斤)

4.4 时间序列分析中季节分解法实例

4.4.1 季节分解法概述

时间序列的变化受多种因素的影响，一般可将这些因素分为一下四种：
1）.长期趋势因素（T）
长期趋势因素反映了某种现象在一个较长时间内的发展方向，可以在一个相当长的时间内表现出一种近似直线的持续向上。持续向下或平稳的趋势。长期趋势一旦形成，便会延续很长时间，因此对其进行预测研究具有特别重要的意义。
2）季节变动因素（S）
季节变动因素是某种现象受季节变动影响所形成的一种长度和幅度固定的周期波动。许多时间序列如销售量及温度等都显示出年周期的变化。
3）.周期变动因素（C）
周期变动因素也称循环变动因素，它是由于某些其他物理原因或经济原因的影响而显示出有固定周期的变化。例如，每日的温度变化，股票价格的变化等，具有明显的周期变动特征。
4）.不规则变动因素（I）
不规则变动因素又称随即变动，它是受各种偶然因素的影响所形成的不规则波动。

4.4.2 季节分解法分析过程

1）.进行季节分解的数据，要有一个至少4个完整季节数据的变量。打开Data菜单中的Define Dates 对话框，定义时间序列的周期。Seasonal Decomposition 程序仅当已完成对趋势选项的设置后有效、
2）按Analyze⇒ Time series⇒ Seasonal Decomposition顺序展开Seasonal Decomposition 对话框。本程序用来估计时间序列的乘性或加性季节因素。

3）指定需要季节分解处理的变量。从左侧变量列表中选中需要进行分析的变量，移到Variable(s)框中。各指定的时间序列必须包括4个完整的季节数据。

4）在Model 栏中，根据时间序列过程的特点，共有两种模型可供选用，即乘法模型Multiplicative和加法模型Additive.

5)在Moving Average Weight (移动平均的权重)栏中，允许用户指定在计算移动平均是如何对待时间序列

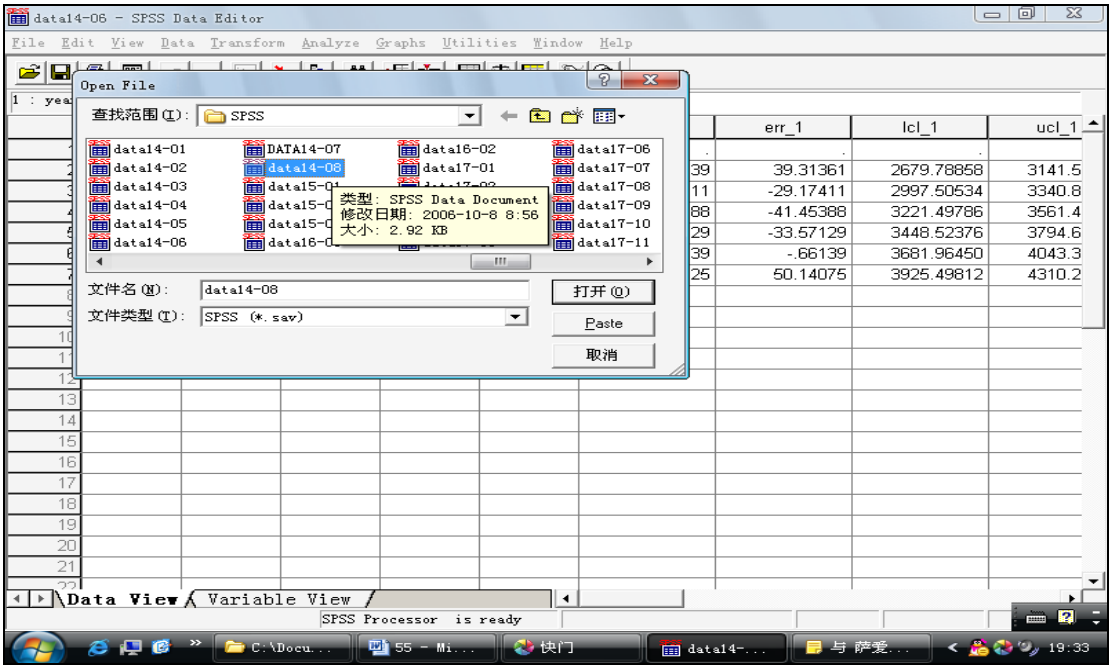
6）Display casewise listing ，在运算过程中对每个变量生成一行4个新序列值。

7）单击Save 按钮，展开保存对话框。

8）单击OK按钮，系统立即执行命令。单击Paste按钮，在Syntax窗口生成程序。

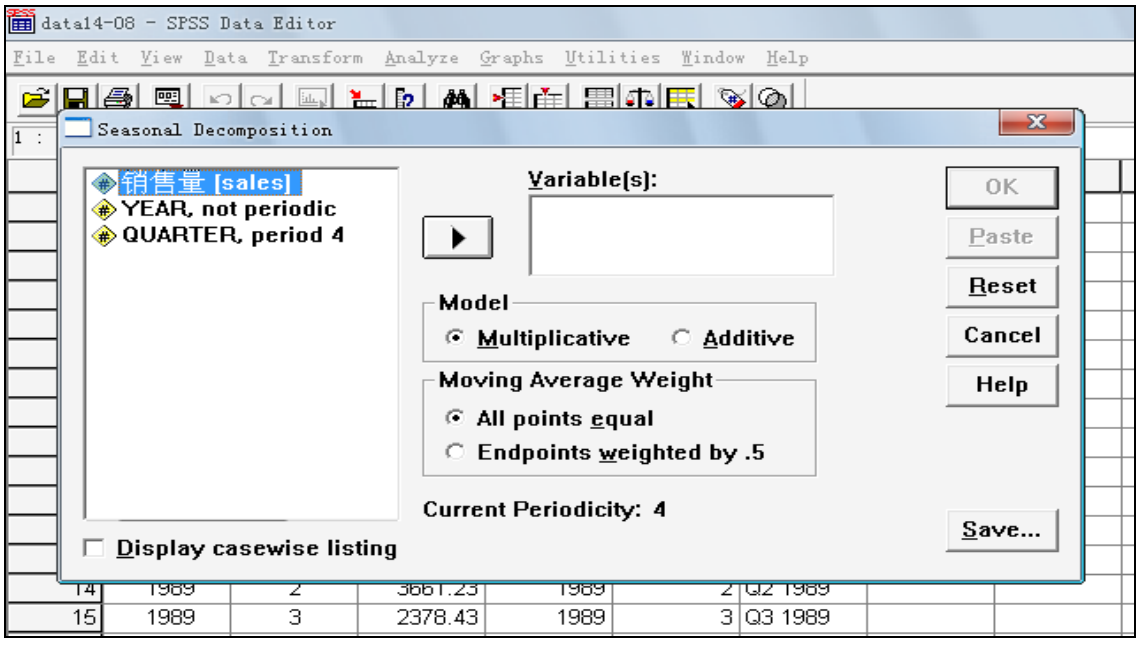
4.4.3 季节分解法分析实例

数据data14-8中变量Sales 为某公司1986-1997年间各季度某商品的销售数据用季节分解法对其进行统计学分析

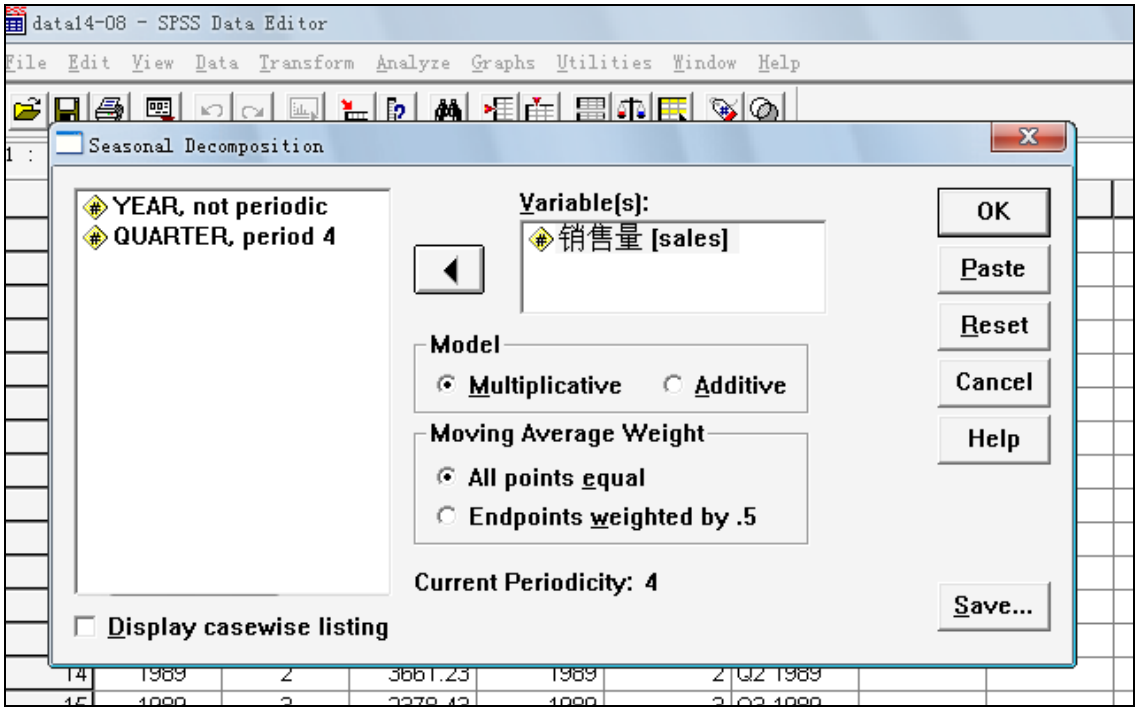


1）从 SPSS for windows 窗口进入打开 file/open/date 选择（data 14-08）数据

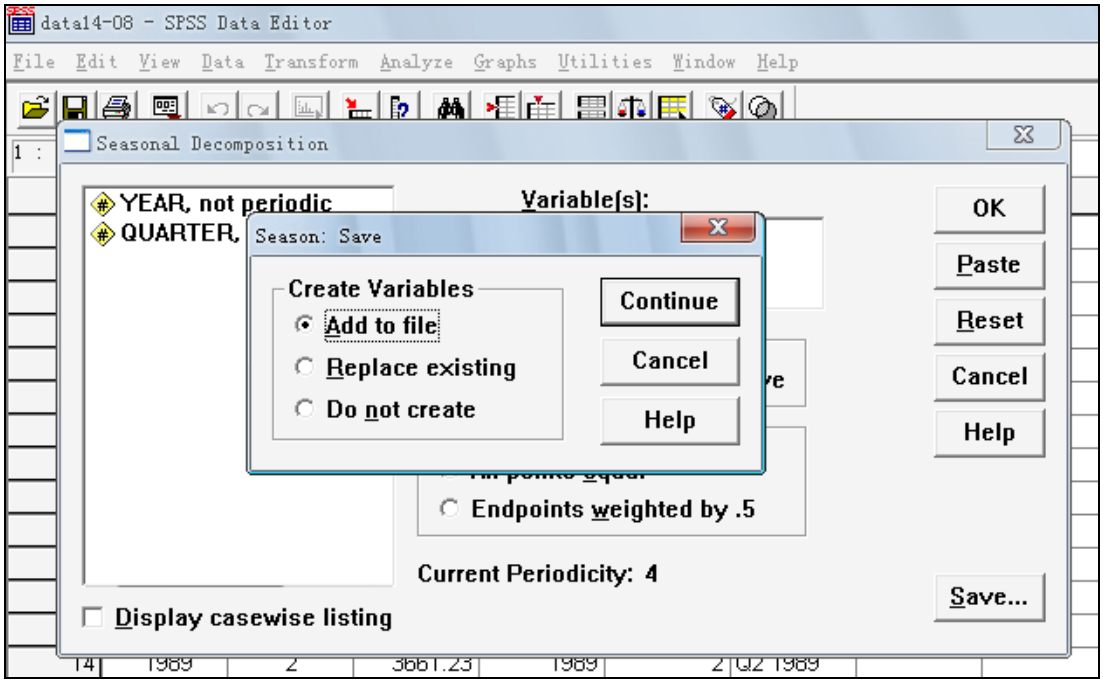
2）按按 Analyze⇒Time series⇒Seasonal Decomposition 顺序展开如图所示的对话框。



3）选择 Sales 变量进入 Variables 对话框。



- 在 Model 框中，选定 Multiplicative 项
- 在 Moving Average Weight 框中，选定 ALL points equal 项。
- 使用 Save 对话框中默认设置，见如图所示的对话框：



- 单击 OK 按钮，设置运算。

4.5 输出结果

在数据窗口中生成误差项 Err_1.长期趋势，季节变动指数，周期变动指数 4 列新数据。

MODEL: MOD_4.
Results of SEASON procedure for variable SALES
Multiplicative Model. Equal weighted MA method. Period = 4.

Period	Seasonal index (* 100)
1	111.818
2	109.198
3	75.774
4	103.210

The following new variables are being created:

Name	Label
ERR_1	Error for SALES from SEASON, MOD_4 MUL EQU 4
SAS_1	Seas adj ser for SALES from SEASON, MOD_4 MUL EQU 4
SAF_1	Seas factors for SALES from SEASON, MOD_4 MUL EQU 4
STC_1	Trend-cycle for SALES from SEASON, MOD_4 MUL EQU 4

实验 5 系统聚类分析

5.1 实验目的

掌握利用 SPSS 统计软件，对于所提供的分析样本进行系统聚类分析的主要原理与基本操作步骤。

5.2 实验原理

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

5.3 实例操作

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

编号	钙	镁	铁	锰	铜	血红蛋白
N0.	X1	X2	X3	X4	X5	X6
1	54.89	30.86	448.70	0.012	1.010	13.50
2	72.49	42.61	467.30	0.008	1.640	13.00
3	53.81	52.86	425.61	0.004	1.220	13.75
4	64.74	39.18	469.80	0.005	1.220	14.00
5	58.80	37.67	456.55	0.012	1.010	14.25
6	43.67	26.18	395.78	0.001	0.594	12.75
7	54.89	30.86	448.70	0.012	1.010	12.50
8	86.12	43.79	440.13	0.017	1.770	12.25
9	60.35	38.20	394.40	0.001	1.140	12.00
10	54.04	34.23	405.60	0.008	1.300	11.75
11	61.23	37.35	446.00	0.022	1.380	11.50
12	60.17	33.67	383.20	0.001	0.914	11.25
13	69.69	40.01	416.70	0.012	1.350	11.00
14	72.28	40.12	430.80	0.000	1.200	10.75
15	55.13	33.02	445.80	0.012	0.918	10.50
16	70.08	36.81	409.80	0.012	1.190	10.25
17	63.05	35.07	384.10	0.000	0.853	10.00
18	48.75	30.53	342.90	0.018	0.924	9.75
19	52.28	27.14	326.29	0.004	0.817	9.50
20	52.21	36.18	388.54	0.024	1.020	9.25
21	49.71	25.43	331.10	0.012	0.897	9.00
22	61.02	29.27	258.94	0.016	1.190	8.75
23	53.68	28.79	292.80	0.048	1.320	8.50
24	50.22	29.17	292.60	0.006	1.040	8.25
25	65.34	29.99	312.80	0.006	1.030	8.00
26	56.39	29.29	283.00	0.016	1.350	7.80
27	66.12	31.93	344.20	0.000	0.689	7.50

28	73.89	32.94	312.50	0.064	1.150	7.25
29	47.31	28.55	294.70	0.005	0.838	7.00

5.3.1 数据准备

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

5.3.2 统计分析

激活 Statistics 菜单选 Classify 中的 Hierarchical Cluster...项，弹出 Hierarchical Cluster Analysis 对话框（图 5.1）。从对话框左侧的变量列表中选 x1、x2、x3、x4、x5、x6，点击➢钮使之进入 Variable(s)框；在 Cluster 处选择聚类类型，其中 Cases 表示观察对象聚类，Variables 表示变量聚类，本例选择 Variables。

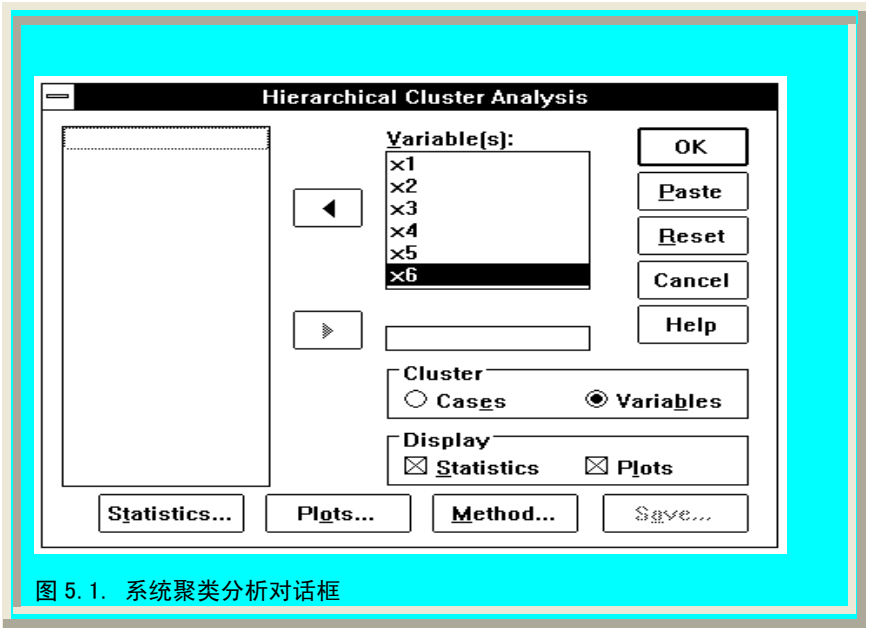


图 5.1. 系统聚类分析对话框

点击 Statistics...钮，弹出 Hierarchical Cluster Analysis: Statistics 对话框，选择 Distance matrix，要求显示距离矩阵，点击 Continue 钮返回 Hierarchical Cluster Analysis 对话框（图 5.2）。

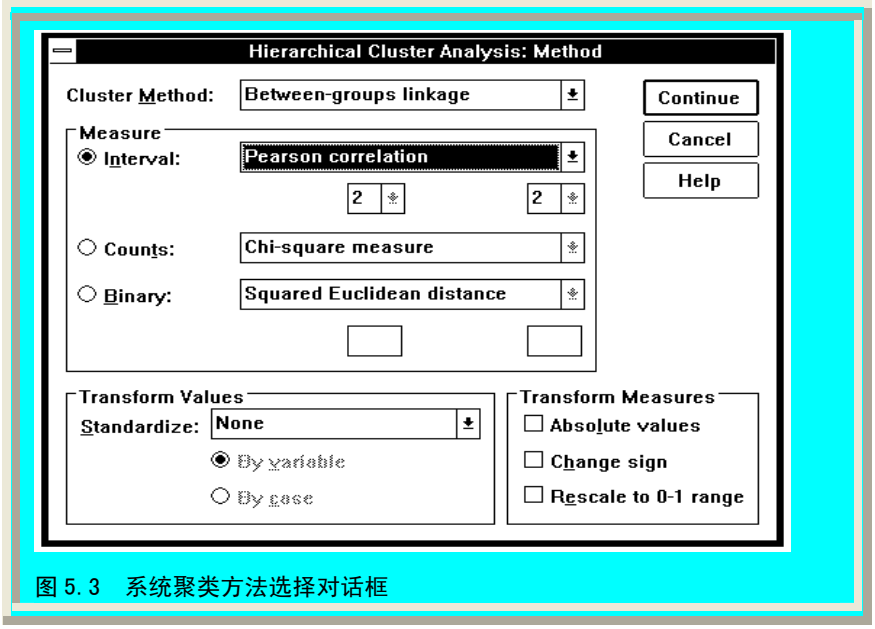


图 5.3　系统聚类方法选择对话框

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

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

Between-groups linkage：类间平均链锁法；

Within-groups linkage：类内平均链锁法；

Nearest neighbor：最近邻居法；

Furthest neighbor：最远邻居法；

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

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

Ward's method：离差平方和法，应与欧氏距离平方法一起使用。

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

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

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

Cosine：变量矢量的余弦，这是模型相似性的度量；

Pearson correlation：相关系数距离，适用于 R 型聚类；

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

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

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

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

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

5.4 输出结果

在结果输出窗口中将看到如下统计数据：

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

Data Information

29 unweighted cases accepted.

0 cases rejected because of missing value.

Correlation measure used.

Correlation Similarity Coefficient Matrix

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

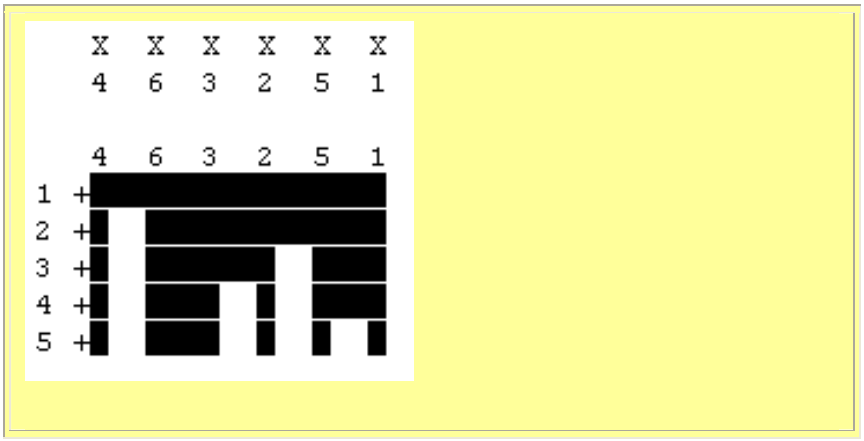
Agglomeration Schedule using Average Linkage (Between Groups)

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

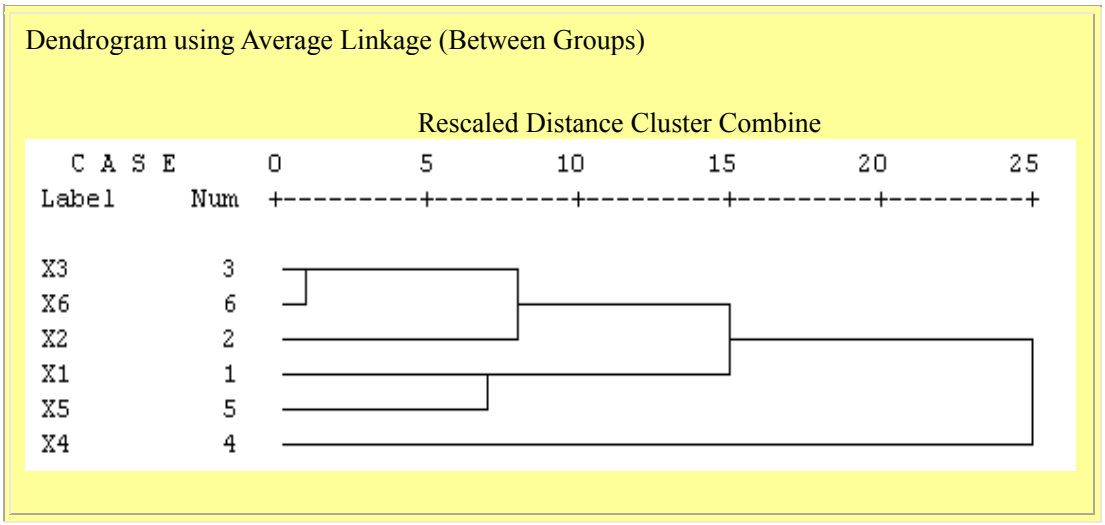
按类间平均链锁法，变量合并过程的冰柱图如下。先是 X3 与 X6 合并，接着 X1 与 X5 合并，然后 X3、X6 与 X2 合并，接着再与 X1、X5 合并，最后加上 X4，六个变量全部合并。

Vertical Icicle Plot using Average Linkage (Between Groups)

(Down) Number of Clusters (Across) Case Label and number



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



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

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

本例相关指数的均值依次为：

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

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

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

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

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

故选择镁（变量 X2）典型指标。

实验 6 主成分分析

6.1［目的要求］

本实验主要是引导学生掌握利用 SPSS 软件进行主成分分析的方法，并能对实验结果进行解释。

6.2［实验内容］

主成分分析（Factor 过程）

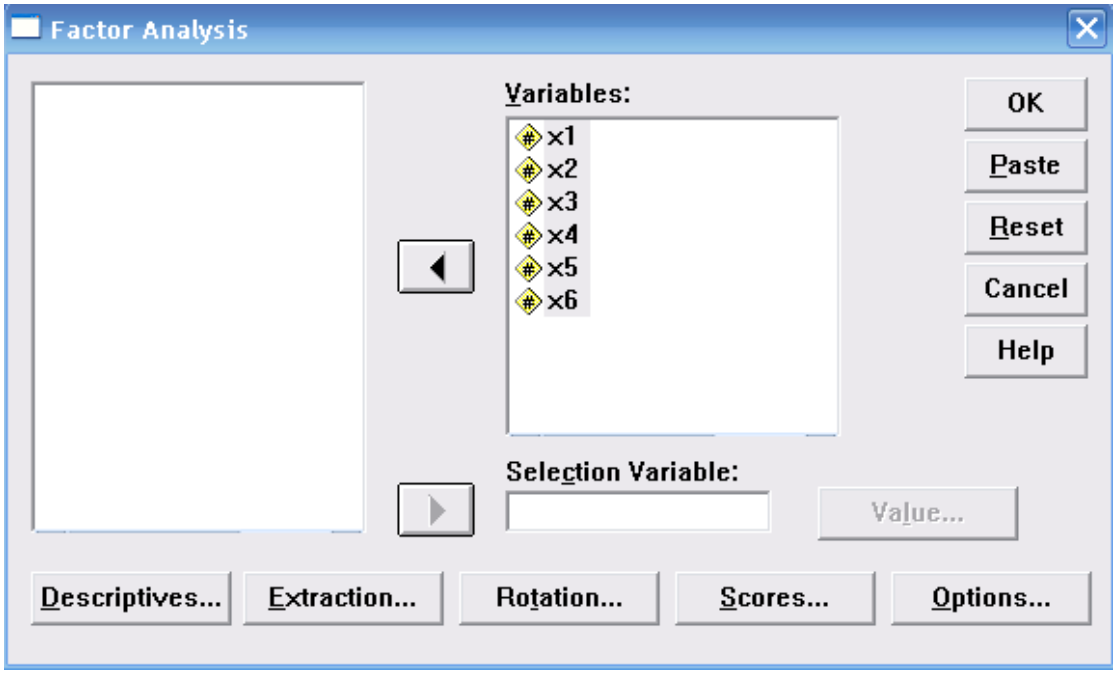
6.3［实验步骤］

6.3.1 主成分分析的概念

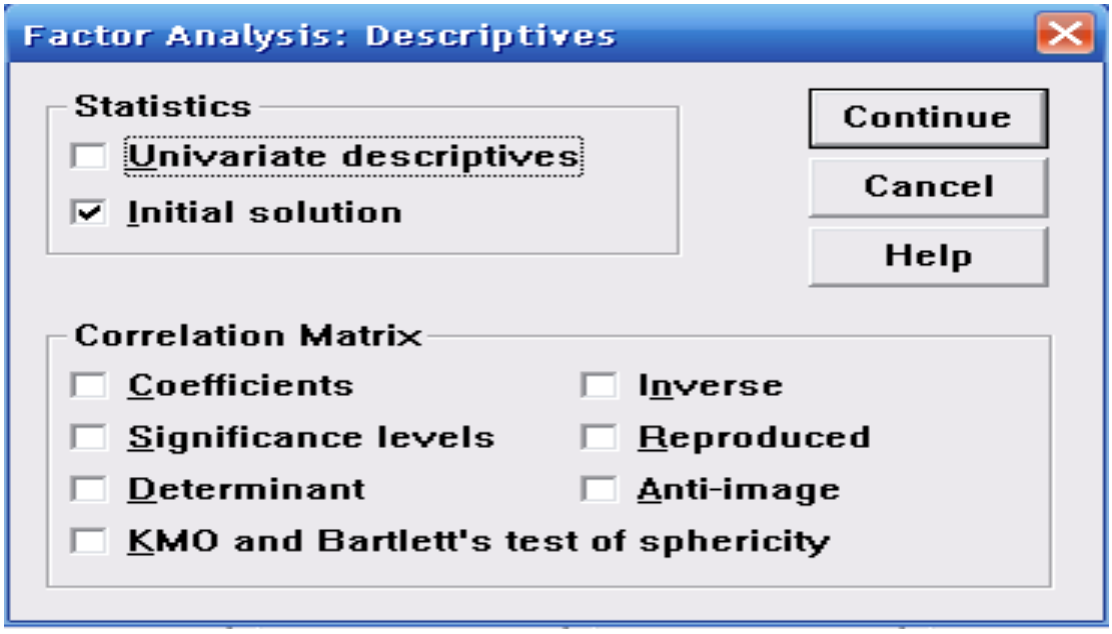
主要功能:多元分析处理的是多指标的问题。由于指标太多，使得分析的复杂性增加。观察指标的增加本来是为了使研究过程趋于完整，但反过来说，为使研究结果清晰明了而一味增加观察指标又让人陷入混乱不清。由于在实际工作中，指标间经常具备一定的相关性，故人们希望用较少的指标代替原来较多的指标，但依然能反映原有的全部信息，于是就产生了主成分分析、对应分析、典型相关分析和因子分析等方法。

6.3.2 定义变量，建立数据文件并输入数据

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

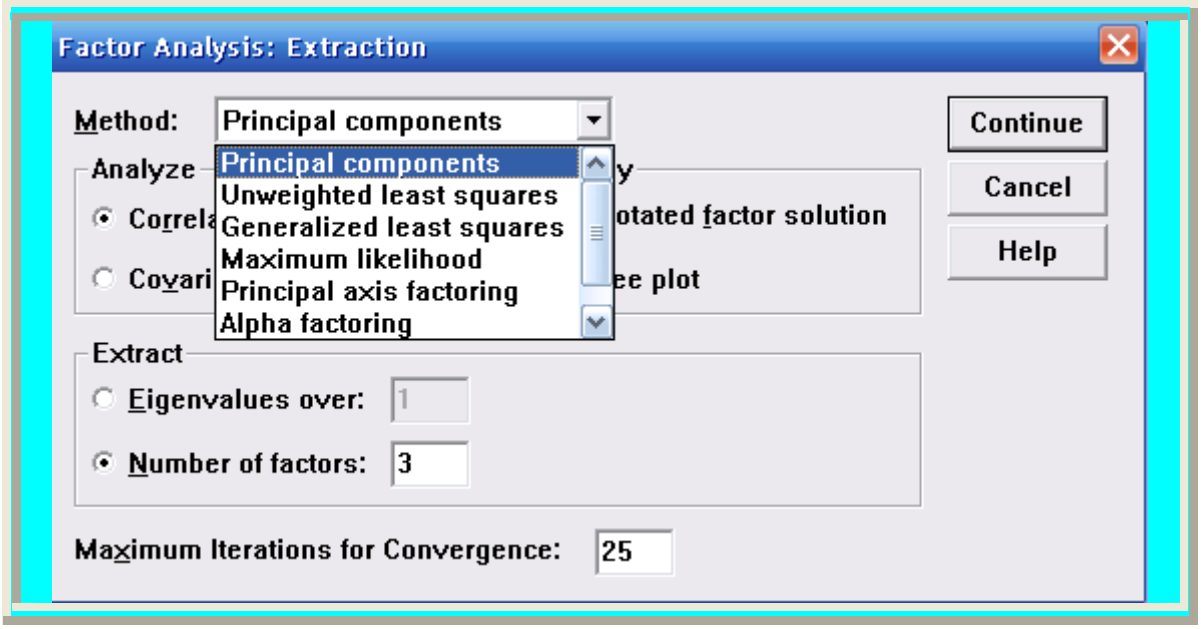


点击Descriptives钮，弹出Factor Analysis： Descriptives对话框（图11.3）：



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

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



Principal components：主成分分析法；

Unweighted least squares：未加权最小平方法；

Generalized least squares：综合最小平方法；

Maximum likelihood：极大似然估计法；

Principal axis factoring：主轴因子法；

Alpha factoring：α因子法；

Image factoring：多元回归法。

本例选用Principal components方法，之后点击Continue钮返回Factor Analysis对话框。

提取公因子时提取三个公因子，使其累积贡献率大于85%。

点击Rotation钮，弹出Factor Analysis:Rotation对话框（图11.5），系统有5种因子旋转方法可选：

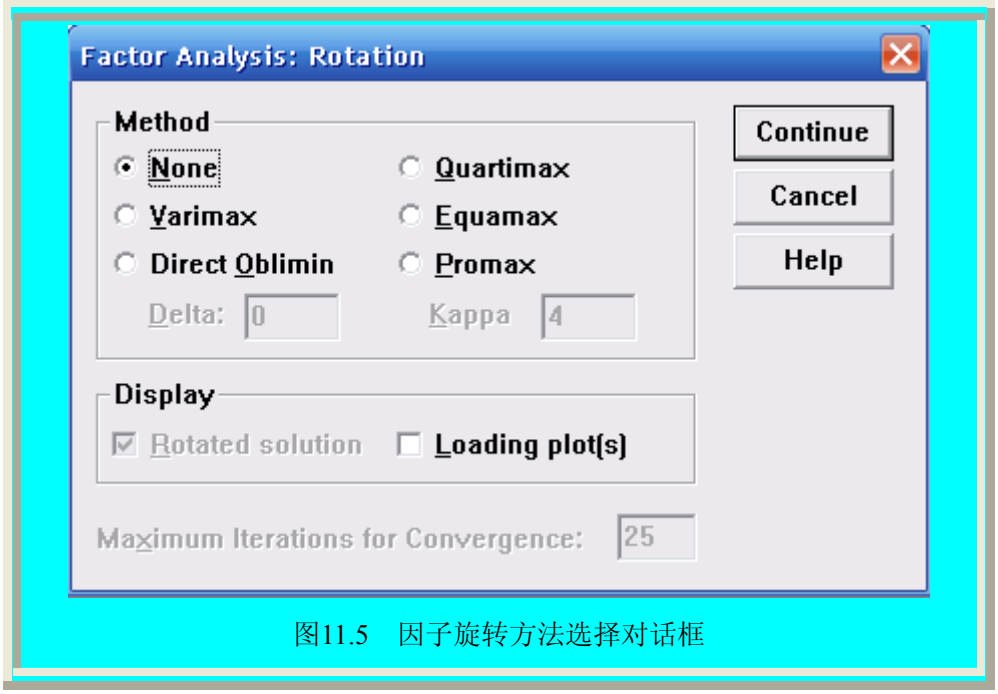


图11.5　因子旋转方法选择对话框

None：不作因子旋转；

Varimax：正交旋转；

Equamax：全体旋转，对变量和因子均作旋转；

Quartimax：四分旋转，对变量作旋转；

Direct Oblimin：斜交旋转。

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

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

图11.6　估计因子分方法对话框

在输出结果窗口中将看到如下统计数据：

Total Variance Explained

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	3.370	56.174	56.174	3.370	56.174	56.174
2	1.127	18.788	74.962	1.127	18.788	74.962
3	.823	13.711	88.672	.823	13.711	88.672
4	.491	8.188	96.860			
5	.145	2.415	99.275			
6	.043	.725	100.000			

Extraction Method: Principal Component Analysis.

Component Matrix[a]

	Component		
	1	2	3
X1	.913	-.161	-.217
X2	.923	-.121	-.079
X3	.891	-.050	-.243
X4	.552	.730	-.196
X5	.580	-.575	.422
X6	.499	.469	.703

Extraction Method: Principal Component Analysis.

a. 3 components extracted.

系统首先输出各变量的均数（Mean）与标准差（Std Dev），并显示观察单位进入分析；特征值和累积贡献率，接着输出相关系数矩阵（Correlation Matrix），从结果看出，提取三个公因子时其累计贡献率为 88.672%大于 85%，所以这三个主成分就能反映了，第一个公因子支配X_1，X_2，　X_3，X_5，　第二个公因子支配X_4，第三个公因子支配X_6。

实验 7　描述统计分析与探索统计分析

7.1[目的要求]

本实验主要是引导学生初步掌握利用 SPSS 软件进行基本统计分析。掌握利用 SPSS 软件进行基本统计量均值标准误，中位数，众数，方差和标准差，四分位数，十分位数和百分位数，频数，峰度，偏度的计算，进行标准化 2 分数及其线性转换，统计，统计图的显示。

7.2[实验内容]

7.2.1 描述性分析（Descriptives 过程）

7.2.2 探索分析（Explore 过程）

7. 3[实验步骤]

7.3.1 最简单的描述统计分析过程与实例

描述统计分析过程通过平均值，算术和，标准差，最大值，最小值，方差，范围和平均数的标准误等统计量对变量进行描述。通过 Z 分数探明异常观测量。描述统计分析过程使用于正态分布的尺度变量。描述统计分析过程有 Descriptives ,Explore.

7.3.2 最简单的描述统计分析过程

1）定义变量，建立数据文件并输入数据。

2)选择菜单 Analyze⇒Descriptive Statistics⇒Descriptives 顺序，打开如图所示的 Descriptives 对话框。

3）在左侧的源变量框中选择一个或多个变量作为待分析变量移入 Variable(s)框中。

4）选择 Save standardized values variables,对所选项的每一个变量进行标准化，产生相应的 Z 分值，作为新变量并保存在当前数据文件中。其变量名为相应变量名加前缀 Z。标准化的计算公式如下：$Zi=\frac{Xi-\bar{X}i}{S}$.

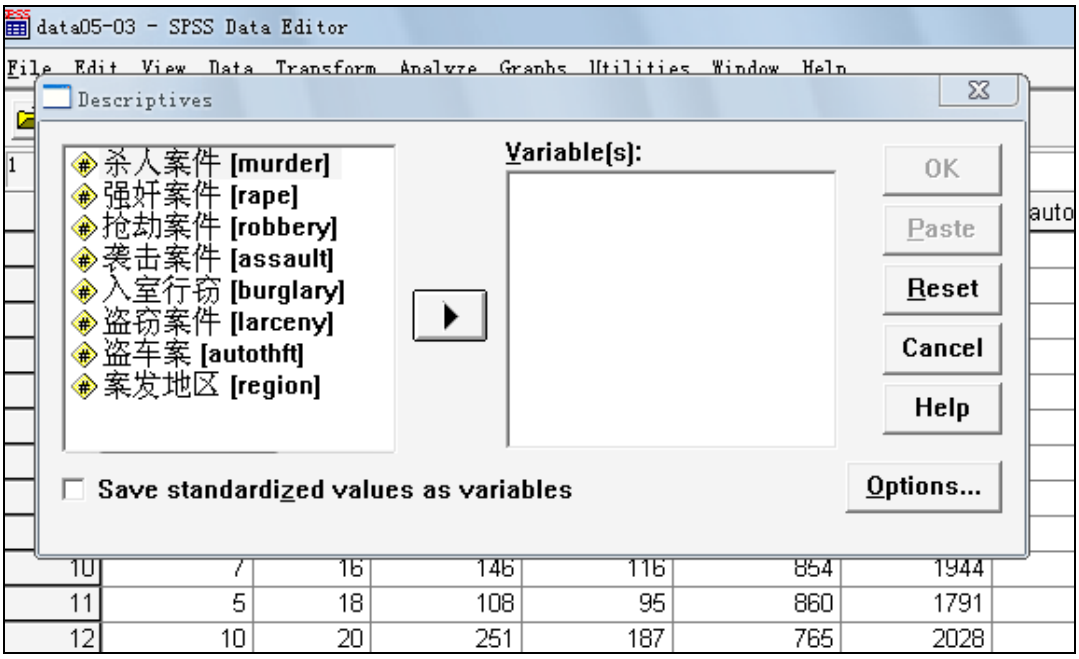
式中X_i为变量x的第i个观测值，$\bar{x}$为变量x的平均数，S为变量x的标准差。

5）单击 Options 按钮，打开对话框。在对话框中可以指定其他统计量与输出统计结果显示的顺序。

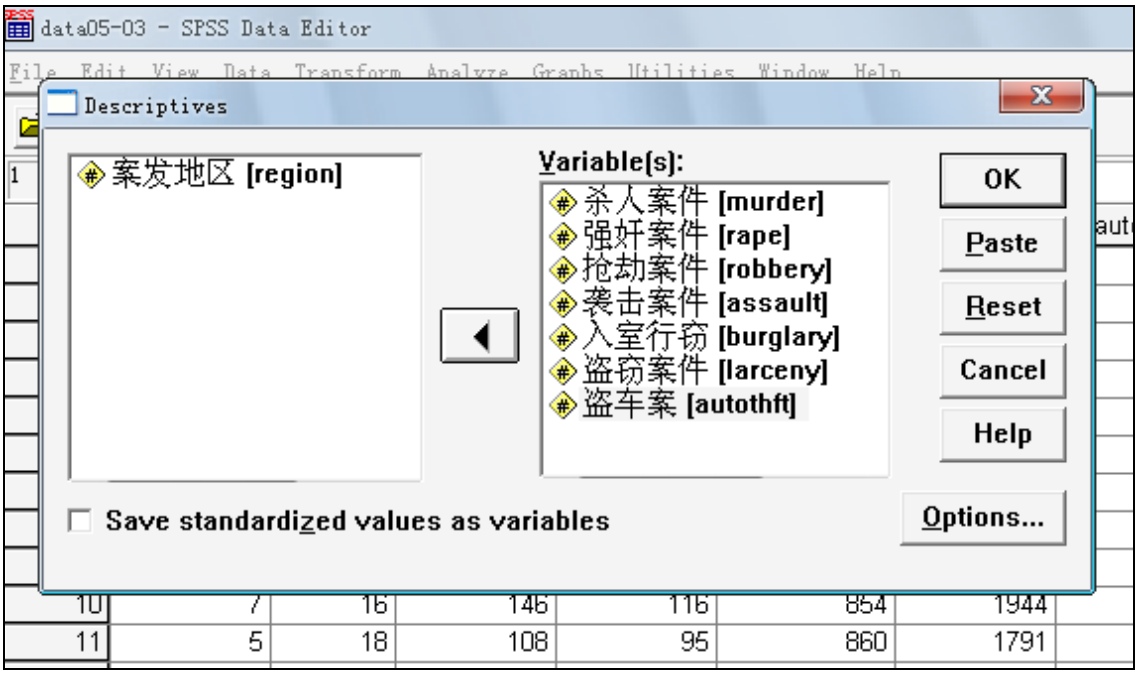
7.3.3 应用实例

数据 data05-03 是对 1985 年美国联邦调查局对五十个州各种犯罪情况调查的数据。变量：murder,rape,robbery,assault,burglar,larceny,autothft 的案件数进行描述分析。

1）打开数据文件，按 Analyze⇒Descriptive Statistics⇒Descriptives 顺序打开如图所示的对话框。

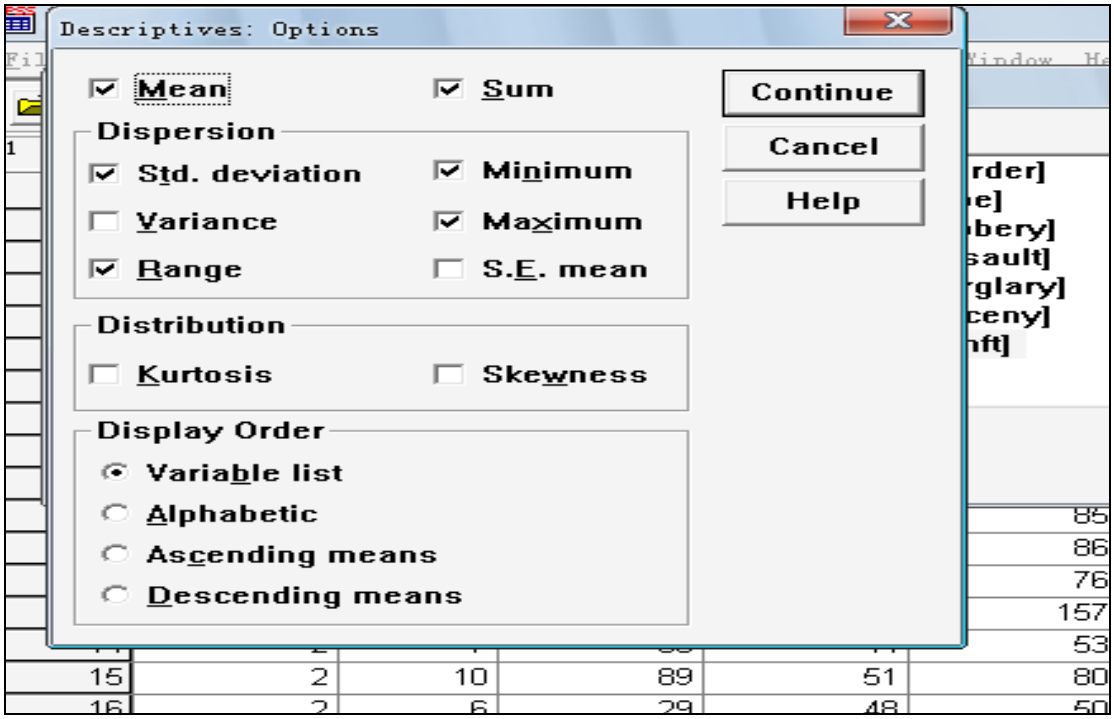


2）选择 murder,rape,robbery,assault,burglar,larceny,autothft 变量送入 Variable(s) 栏中。



3）选中 Save standardized values variables 要求计算变量的标准化值，并保存在当前数据文件中。

4）单击 options 按钮，打开对话框。选中 mean,sum,std.deviation,minimum,maximum,range 要求计算的描述统计量。



5）在主对话框中单击 OK 按钮，提交运行。输出结果表，此表中，从左至右分别为变量名称，观测值，观测量的频数，全距，最小值，最大值，和，均值以及标准差

7.3.4 输出结果

Descriptive Statistics

	N	Range	Minimum	Maximum	Sum	Mean	Std. Deviation
杀人案件	50	15	1	15	343	6.86	3.848
强奸案件	50	32	4	36	781	15.62	7.348
抢劫案件	50	437	7	443	5076	101.51	91.193
袭击案件	50	272	21	293	6771	135.42	68.170
入室行窃	50	1467	286	1753	46540	930.80	361.050
盗窃案件	50	2856	694	3550	97182	1943.64	709.829
盗车案	50	800	78	878	18393	367.86	199.610
Valid N (listwise)	50						

7.4 最简单的探索统计分析过程与实例

7.4.1 探索分析概述

探索分析过程提供对测的数据的考查。考查可以有一下两个方法：

1） 检查数据是否有错误

过大或过小的数据均有可能是异常值，影响点或是错误输入的数据。因为异常值和影响点往往对分析结果影响较大，不能真实地反映数据的总体特征。

2） 检查数据分布特征

许多分析方法对数据的分析有一定的要求，例如要求样本来自正态分布总体。从试验或实际测量得到的数据是否符合正态分布规律，决定了他们是否可以选用只对正态分布数据适用的分析方法。

7.4.2 探索分析过程

1） 输入数据。按 Analyze⇒Descriptive Statistics⇒Explore,打开对话框。

2） 从源变量框中，选择若干个数值型变量作为因变量送入 Dependent 框中。此时单击 OK 按钮即可获得默认的统计分析，这其中包括箱图，茎叶图以及基本的描述统计量。默认情况下缺失值将会被排除到分析过程之外。

3） 制定变量

在源变量框中选择一个或多个分组变量进入 Factor 框中。分组变量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，那么会以分组变量各取值进行组合分组。例如以相别 sex （f,m）,年龄段变量 age1(11,12,13)为指定的分组变量，则按组合分组为：（f,11）,(f,12)(f,13),(m,11),(m,12),(m,13),分组对数据进行分析。

4） 选择表示变量

在源变量表中指定一个变量作为观测量的表示变量，送入 Lable cases by 框中。当输出涉及各个观测量时，例如异常值的输出，使用该变量值标识各观测量。

5） Disply 栏，确定输出项

6） 选择描述统计量

单击 Statistics 按钮

- Decriptives 复选项，要求输出基本描述统计量：平均数，中位数，众数，5%的调整平均值，标准误，方差，标准差，最大值，最小值，范围，等距四分位数，峰度与偏度，及峰度与偏度的标准误。在 Confidence intervala for mean 框中设置均值的置信区间。在参数框中键入置信区间，选择的范围从 1%到 99%，常用的数值为 90%，95%，99%，95%为默认值。
- M-estimators 复选项，要求输出集中趋势最大似然比的稳健估计。
- Outliers 复选项，要求输出第 5，10，25，50，75，90 以及 95 百分位数。

7） 统计图形及其参数，展开 Plots 对话框。

- boxplots 栏，确定箱图选项。
- Descriptive 栏，选择描述图形。
- Normality plots with tests 复选项，输出输出正太概率与离散概率图。同时输出 Kolmogorov_smirnov 统计量的 Liliefors 检验的显著水平，计算，输出 Shapiro_wilk 统计量。

- Spread vs level with levene test 栏，对所有的散步/层次图来说，同时输出回归直线斜率以及方差齐性的 levene 检验结果。如果没有指定分组变量，此选项无效。

将观测量数据转换为散步一层次图，在散步一层次图上将显示回归斜线，Levene 稳健估计。如果选择了 Transformed 转换选项，将依据转换后的数据进行计算。

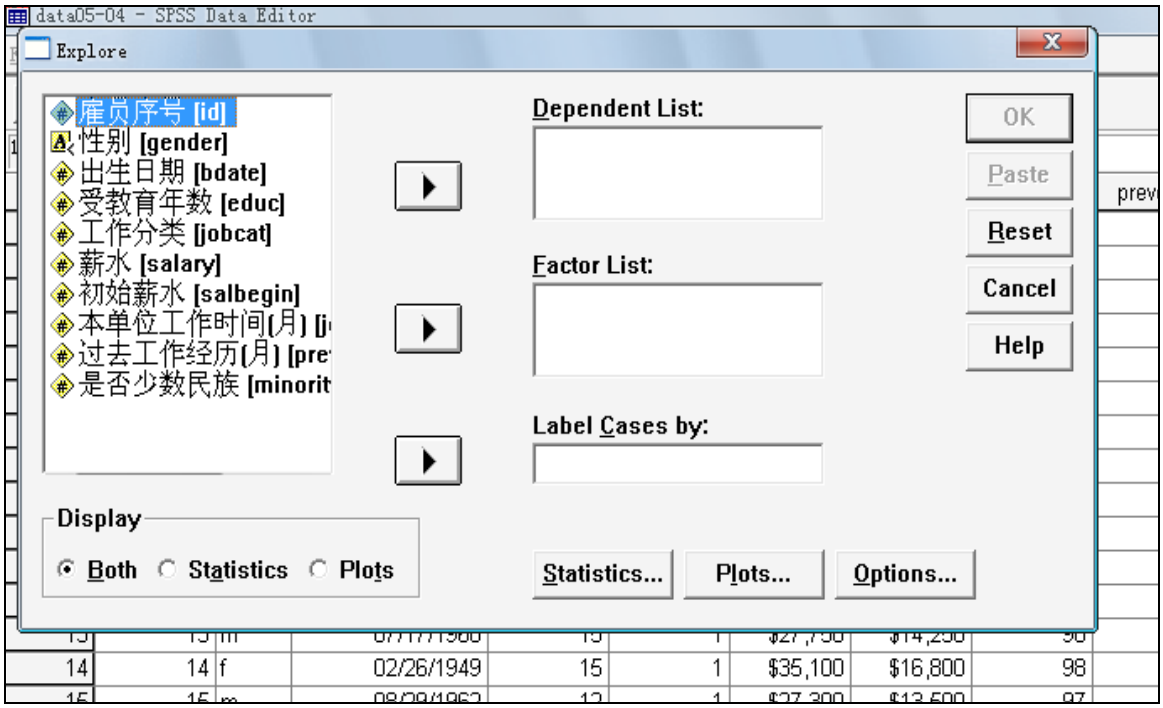
8）单击 Options 按钮，展开如图所示的对话框。在对话框中确定对待缺失值的方式。

7.4.3 探索分析实例

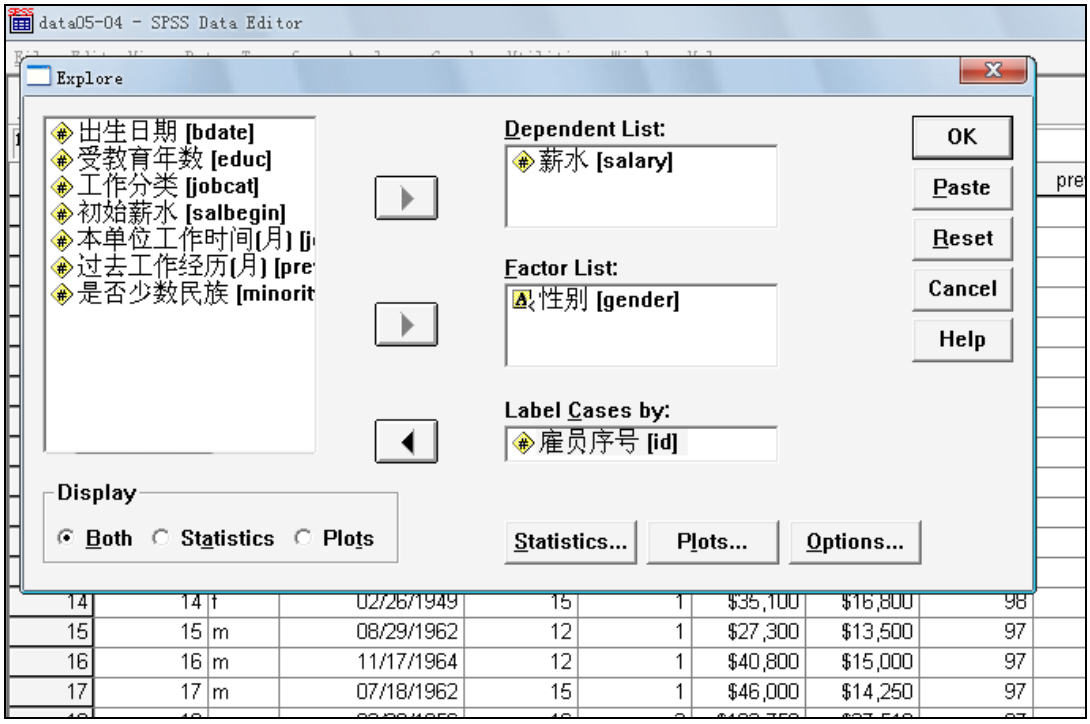
Data05-04 包括 1969~1971 年美国一家银行的 474 名雇员情况的数据包括变量：当前工资 salary ，受教育水平（年）edue，工作经历（月）prevexp,种族 minority(0:非少数民族，1：少数民族)，职务等级 jobcat 等。

1. 选择变量，指定选项。

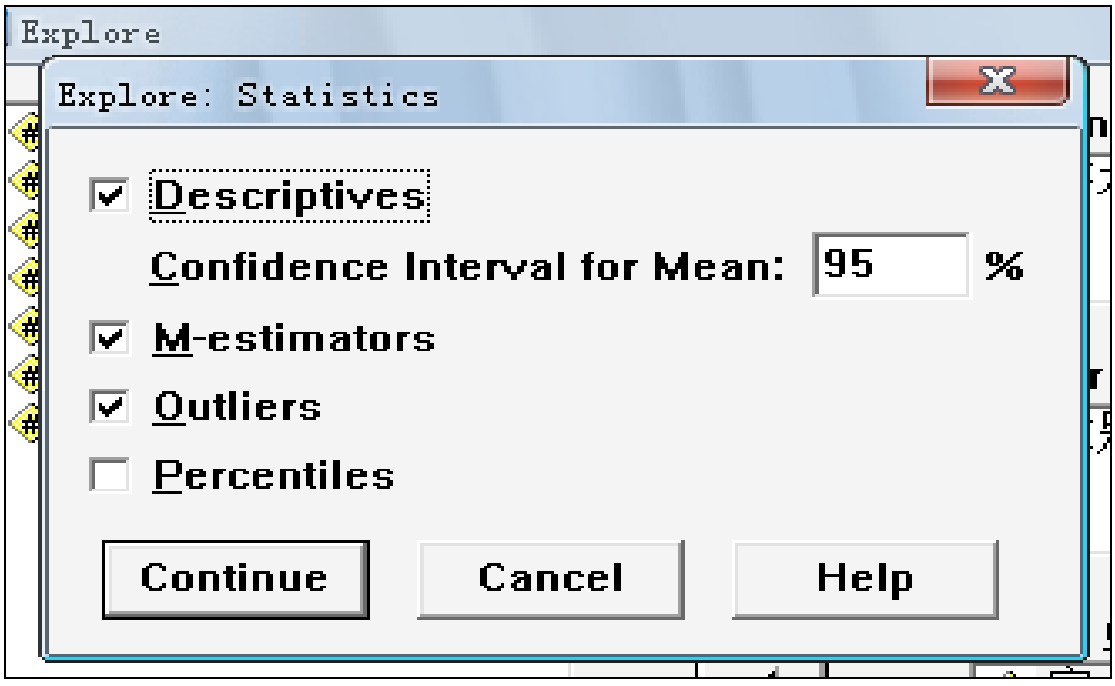
1） 打开 data05-04，按 Analyze ⇒Descriptive Statistics ⇒Explore 顺序打开主对话框。



2） 选择 salary 变量进入 Dependent 框，选择 gender 变量进入 Factor list 框，选择 id 变量进入 label cases 框。在 Display 栏中，选择 Both 项。

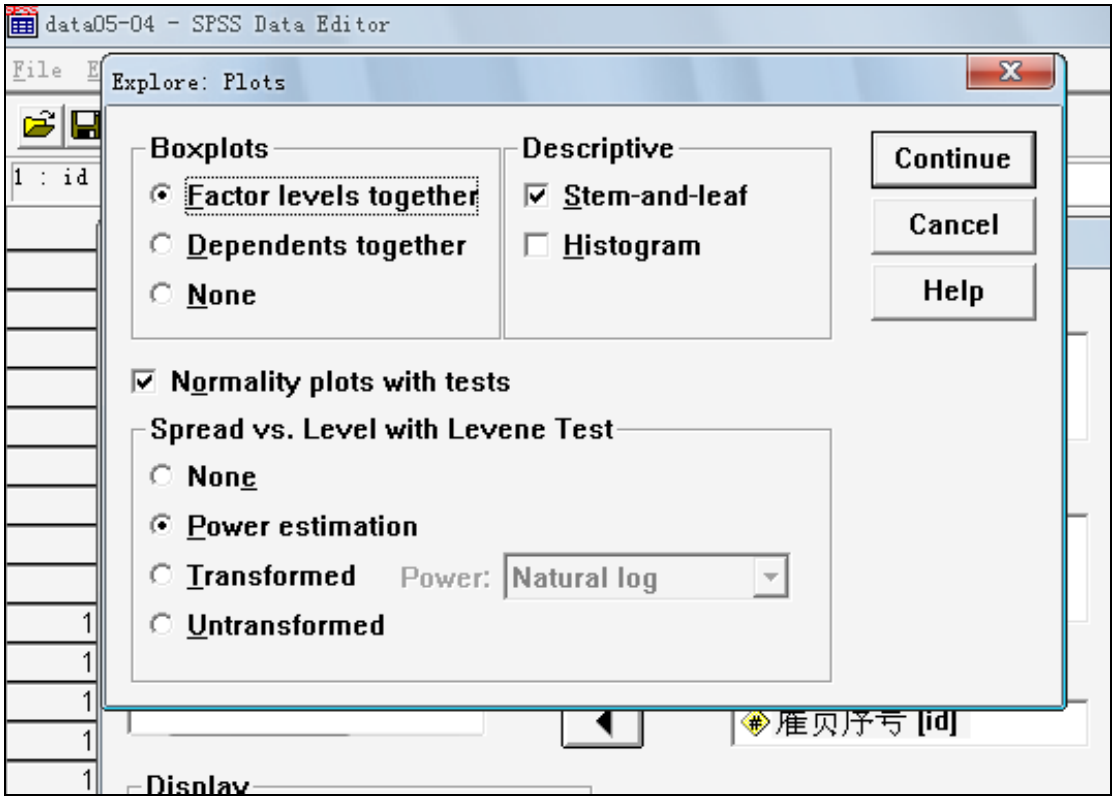


3） 单击 Statistics 按钮，打开 Statistics 对话框。选中 Descriptives,Outliers M-estimators.



4)单击 Plots 按钮，打开 plots 对话框。选择 Boxplots 栏中的 Factor levels together;

选择 Descriptive 栏内的 stem-and-leaf,选中 Normality plots with tests ;在 spread vs .level with levene test 栏中选择 power estimation.



5)单击 OK 按钮，提交运行。

6）部分输出结果见表 7-1 至表 7-6，图 7-1 至图 7-4.

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	性别	N	Percent	N	Percent	N	Percent
薪水	女	216	100.0%	0	.0%	216	100.0%
	男	258	100.0%	0	.0%	258	100.0%

（7-1 观测量摘要表）

表 7-1 为数据的一般统计量。说明具有合法值的女性（Female）观测值共有 216 个，具有合法的男性（Male）观测量共有 258 个，它们都没有缺失值。

表 7-2 的统计量为：因变量 salary；因素变量 gender；统计量；标准误；均值；均值的 95%置信区间；上限值；下限值；5%的调整平均值，即剔除 5%的最大与最小观测量后计算所得的均值；中位数；方差值；标准差；最小值；最大值；全距；四分位数间距；偏度，其数值为 1.863 说明数据呈现非正态分布状态；峰度值为 4.641，说明变量 salary 的分布要比标准正态峰高。

Descriptives

	性别			Statistic	Std. Error
薪水	女	Mean		$26031.9	$514.258
		95% Confidence Interval for Mean	Lower Bound	$25018.3	
			Upper Bound	$27045.6	
		5% Trimmed Mean		$25248.3	
		Median		$24300.0	
		Variance		5.7E+07	
		Std. Deviation		$7558.02	
		Minimum		$15,750	
		Maximum		$58,125	
		Range		$42,375	
		Interquartile Range		$7,012.50	
		Skewness		1.863	.166
		Kurtosis		4.641	.330
	男	Mean		$41441.8	$1213.97
		95% Confidence Interval for Mean	Lower Bound	$39051.2	
			Upper Bound	$43832.4	
		5% Trimmed Mean		$39445.9	
		Median		$32850.0	
		Variance		3.8E+08	
		Std. Deviation		$19499.2	
		Minimum		$19,650	
		Maximum		$135,000	
		Range		$115,350	
		Interquartile Range		$22675.0	
		Skewness		1.639	.152
		Kurtosis		2.780	.302

（表 7-2 Salary 的描述统计量）

表7-3下面的a,b,c,d分布表示四种M估计统计量的各自加权常数。与表7-2的均值比较，发现M估计值全部要比均值较小（Female=$26031.92,Male=$41,441.78）但是与中位数（Female=$24300,Male=$32850）十分接近，初步判定观测量数据可能呈现偏态分布。

M-Estimators

	性别	Huber's M-Estimator(a)	Tukey's Biweight(b)	Hampel's M-Estimator(c)	Andrews' Wave(d)
薪水	女	$24,607.10	$24,014.73	$24,421.16	$24,004.51
	男	$34,820.15	$31,779.76	$34,020.57	$31,732.27

（表7-3 M估计量）

a The weighting constant is 1.339.

b The weighting constant is 4.685.

c The weighting constants are 1.700, 3.400, and 8.500

d The weighting constant is 1.340*pi.

表7-4中Case number 是观测量号，Employee code 是职工编号。显示了按性别分组的各组中的5个量最大（最高薪水）和5个最小值（最低薪水）。

Extreme Values

	性别			Case Number	雇员序号	Value
薪水	女	Highes	1	371	371	$58,125
		t	2	348	348	$56,750
			3	468	468	$55,750
			4	240	240	$54,375
			5	72	72	$54,000
		Lowest	1	378	378	$15,750
			2	338	338	$15,900
			3	411	411	$16,200
			4	224	224	$16,200
			5	90	90	$16,200
	男	Highes	1	29	29	$135,000
		t	2	32	32	$110,625
			3	18	18	$103,750
			4	343	343	$103,500
			5	446	446	$100,000
		Lowest	1	192	192	$19,650
			2	372	372	$21,300
			3	258	258	$21,300
			4	22	22	$21,750
			5	65	65	$21,900

（表7-4 变量的极端值）

表7-5为检测数据是否为正态分布的统计量：自左至右分别为：Kolmogorov_smirnov统计量值，自由度，显示水平值，Shapiro_wilk检验的统计量，自由度，显著水平值。由于显著水平值均为Sig =0.000<0.05,所以拒绝数据的正态分布的假设。

Tests of Normality

		Kolmogorov-Smirnov(a)			Shapiro-Wilk		
	性别	Statistic	df	Sig.	Statistic	df	Sig.
薪水	女	.146	216	.000	.842	216	.000
	男	.208	258	.000	.813	258	.000

a Lilliefors Significance Correction

（表7-5 正态分布检测统计量）

表7-6为方差齐性检验结果。自左至右：levene统计量，自由度1，自由度2，显著水平值，自上至下：一句均值的结果，依据中位数所得的结果，依据中位数与调整后的自由度所得的统计量，依据调整均值所得的各个统计量。交叉点上的数值含义很明显。

依据各种各种集中趋势统计量所做检验的显著水平值全部为0.000，拒绝方差相等的零假设。也就是说各性别分组薪水方差不等。

Test of Homogeneity of Variance

		Levene Statistic	df1	df2	Sig.
薪水	Based on Mean	119.669	1	472	.000
	Based on Median	51.603	1	472	.000
	Based on Median and with adjusted df	51.603	1	310.594	.000
	Based on trimmed mean	95.446	1	472	.000

（表7-6方差齐性检验）

图 7-1（a）,(b)分别为男，女工资水平的茎叶图，可以推断男性工资集中在 25000~39000 之间，女性工资集中在 16000~29000 之间。男女之间的工资水平有较大差异

```
薪水  Stem-and-Leaf Plot for
GENDER=  男
 Frequency     Stem &  Leaf
      1.00      1 .  &
     18.00      2 .  11222344
     64.00      2 .  55555666666667777777888889999
     60.00      3 .  000000000000001111111233334 4
     22.00      3 .  5555667899
     16.00      4 .  000023&
     11.00      4 .  55678&
      9.00      5 .  0124&
     10.00      5 .  5569&
      8.00      6 .  001&
     14.00      6 .  56688&
      6.00      7 .  03&
      5.00      7 .  58
      4.00      8 .  &&
     10.00 Extremes    (>=86250)
 Stem width:      10000
 Each leaf:       2 case(s)
```

(a)

薪水 Stem-and-Leaf Plot for
GENDER= 女

Frequency	Stem &	Leaf
2.00	1 .	55
16.00	1 .	6666666666777777
14.00	1 .	88889999999999
31.00	2 .	0000000000000111111111111111111
35.00	2 .	22222222222222222222333333333333333
38.00	2 .	44444444444444444444444444555555555555
22.00	2 .	66666666666777777777777
17.00	2 .	88888899999999999
7.00	3 .	0001111
8.00	3 .	22233333
8.00	3 .	44444555
5.00	3 .	66777
2.00	3 .	88
11.00 Extremes		(>=40800)

Stem width: 10000
Each leaf: 1 case(s)

（b）

图 7-2 为正太概率图，其中的斜线是正态分布的标准线，围绕斜线的各点为预测值，如果观测数据为正态分布，这些点组成的直线与斜线重合。图中大量的点偏离了斜线，因此数据不是正态分布。图 7-3 中，点组成 V 形曲线，也可得出拒绝正态分布的结论

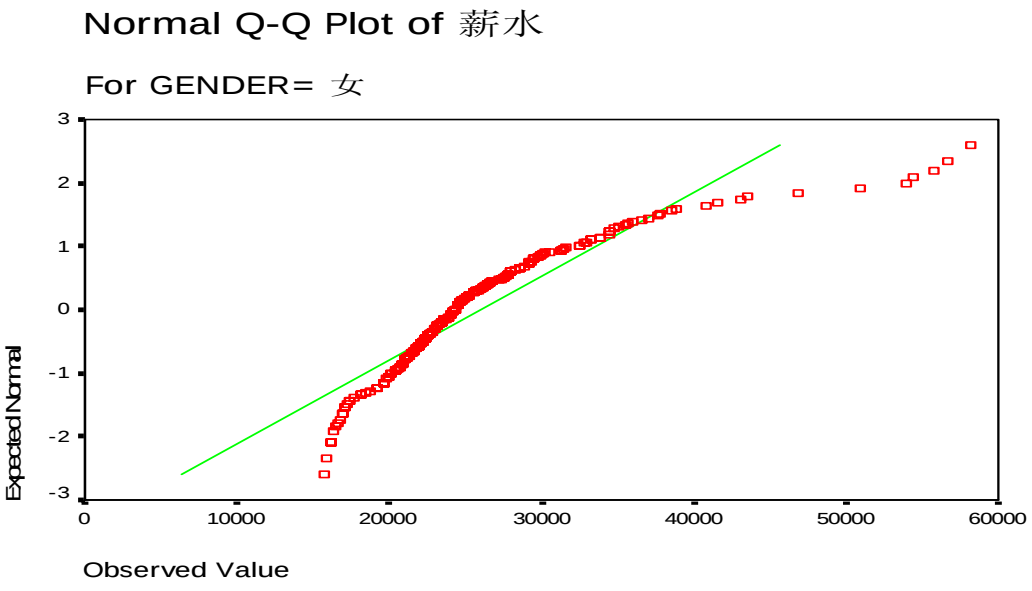


图 7-2 Male 足心谁的正态概率图

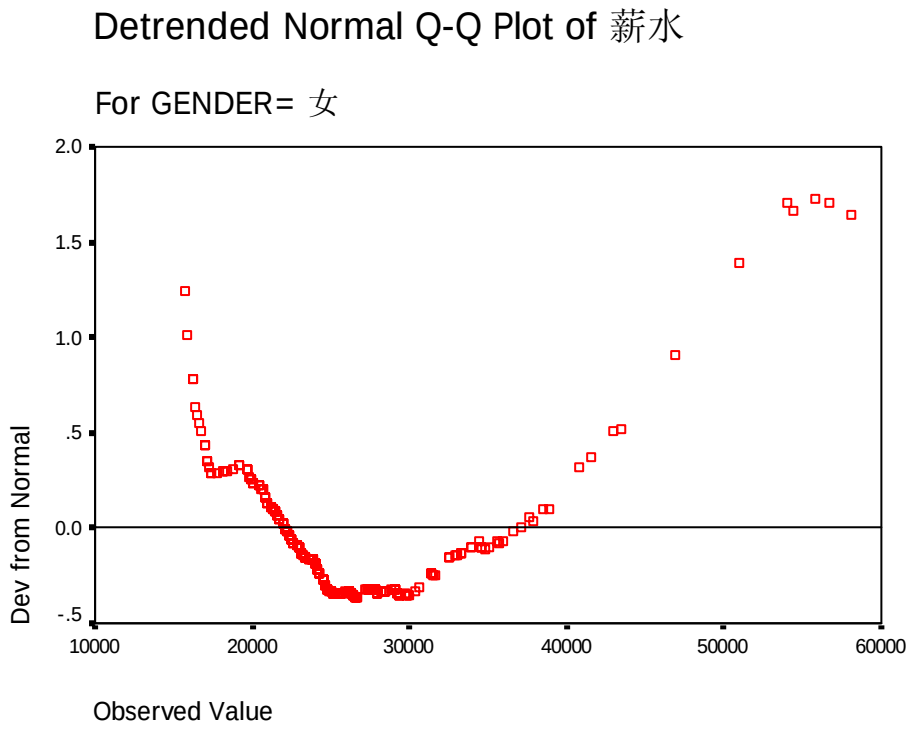
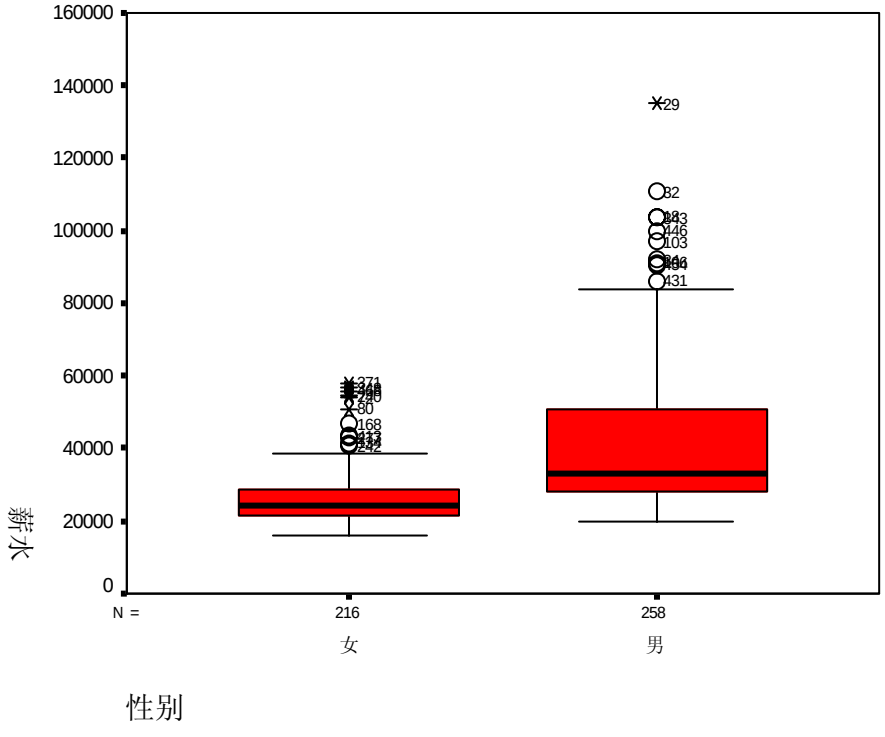


图 7-3 Male 组的薪水离散正太概率

图 7-4 为变量 gender 的两个分组 Female 与 Male 的薪水箱图。女性当前工资水平的全距较变量男性的小，但是两组变量中都存在不少异常量。如 Female 中的 134，242，277 号观测量与极值，如 Female 中的 371，348，468 号观测量等。



实验八 8 Excel 在统计分析中应用

8.1 实验说明

Microsoft Excel 是美国微软公司开发的 Windows 环境下的电子表格系统，它是目前应用最为广泛的办公室表格处理软件之一。自 Excel 诞生以来 Excel 历经了 Excel5.0、Excel95、Excel97 和 Excel2000 等不同版本。随着版本的不断提高，Excel 软件的强大的数据处理功能和操作的简易性逐渐走入了一个新的境界，整个系统的智能化程度也不断提高，它甚至可以在某些方面判断用户的下一步操作，使用户操作大为简化。Excel 具有强有力的数据库管理功能，丰富的宏命令和函数，强有力的决策支持工具，图表绘制功能，宏语言功能，样式功能，对象连接和潜入功能，连接和合并功能，这些特性，已使 Excel 称为现代办公软件重要的组成部分。

8.2 实验目的与要求

本实验重点介绍Excel在统计分析中的应用，包括Excel在描述统计中的应用以及Excel在推断统计中的应用，要求学生熟练掌握运用 Excel 进行统计分析的方法，并能过对分析结果进行解释。

8.3 实验步骤

8.3.1 描述统计工具

描述统计工具用于生成对输入区域中数据的单变量分析，提供数据趋中性和易变性等有关信息。通过描述统计工具可生成以下统计指标，按从上到下的顺序其中包括样本的平均值（X），标准误差（$nS/$），组中值(Medium)，众数(Mode)，样本标准差(S)，样本方差(S2)，峰度值，偏度值，极差(Max-Min)，最小值(Min)，最大值(Max)，样本总和，样本个数（n）和一定显著水平下总体均值的置信区间。下面介绍用表 2.6.1 所提供的数据，在 Excel 中实现描述统计分析的操作过程。

表 8.3.1　某班学生英语成绩

学号	成绩	学号	成绩	学号	成绩	学号	成绩
1	42	6	75	11	67	16	93
2	85	7	86	12	85	17	98
3	93	8	91	13	88	18	75
4	75	9	54	14	70	19	81
5	64	10	98	15	49	20	64

（一）操作步骤　在 Excel 中实现描述统计的一般步骤如下：

（1）用鼠标点击工作表中待分析数据的任一单元格。

（2）选择“工具”菜单的“数据分析”子菜单（见图 8.3.1）。如果“数据分析”命令没有出现在“工具”菜单上，则必须运行“安装”程序来加载“分析工具库”。安装完毕之后，必须通过“工具”菜单中的“加载宏”命令，在“加载宏”对话框中选择并启动它。

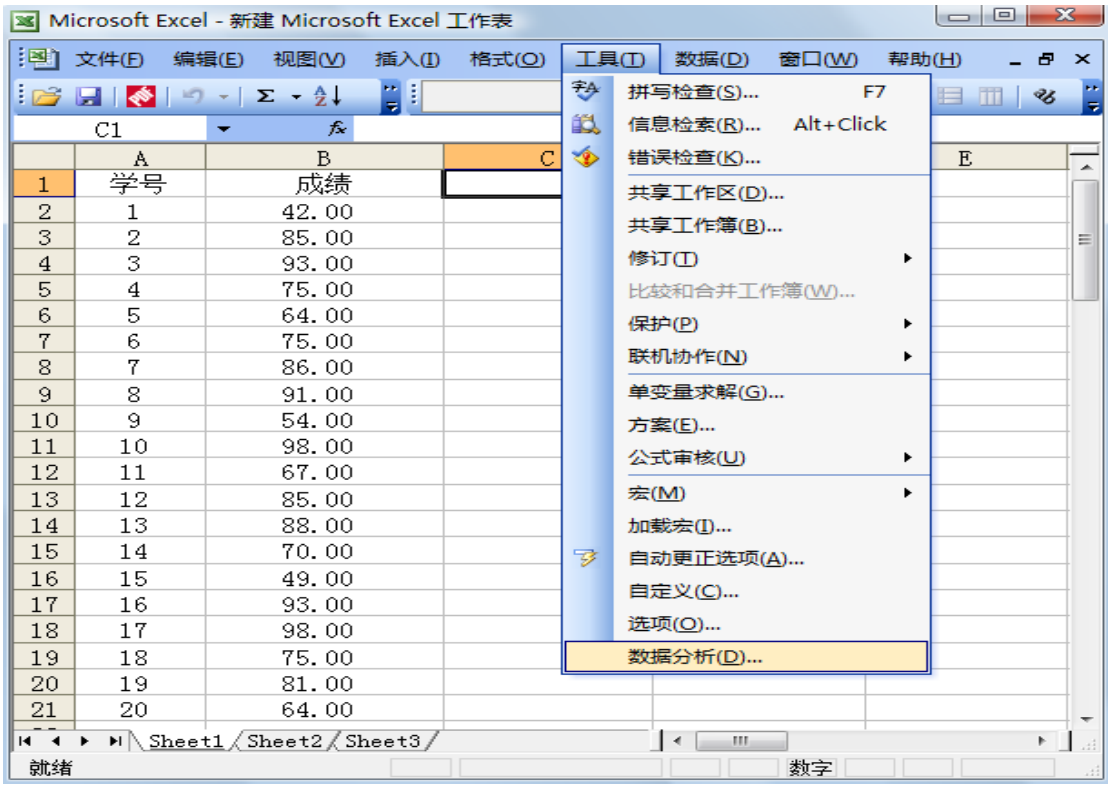
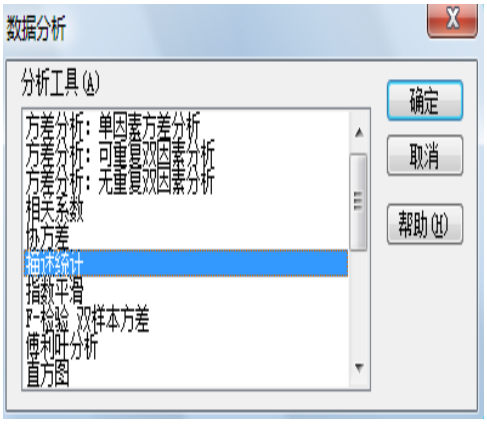


图 8.3.1

（3）用鼠标双击数据分析工具中的“描述统计”选项(见图 8.3.2)。

（4）出现“描述统计”对话框（见图 8.3.3），对话框内各选项的含义如下：



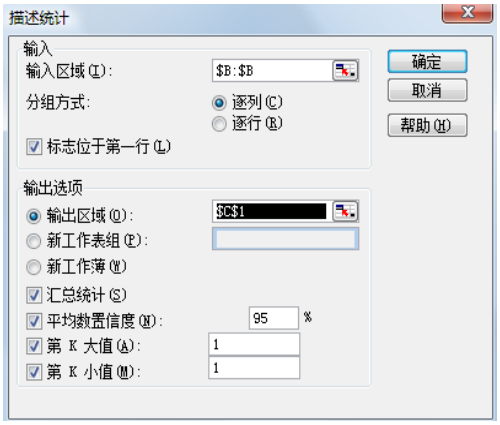
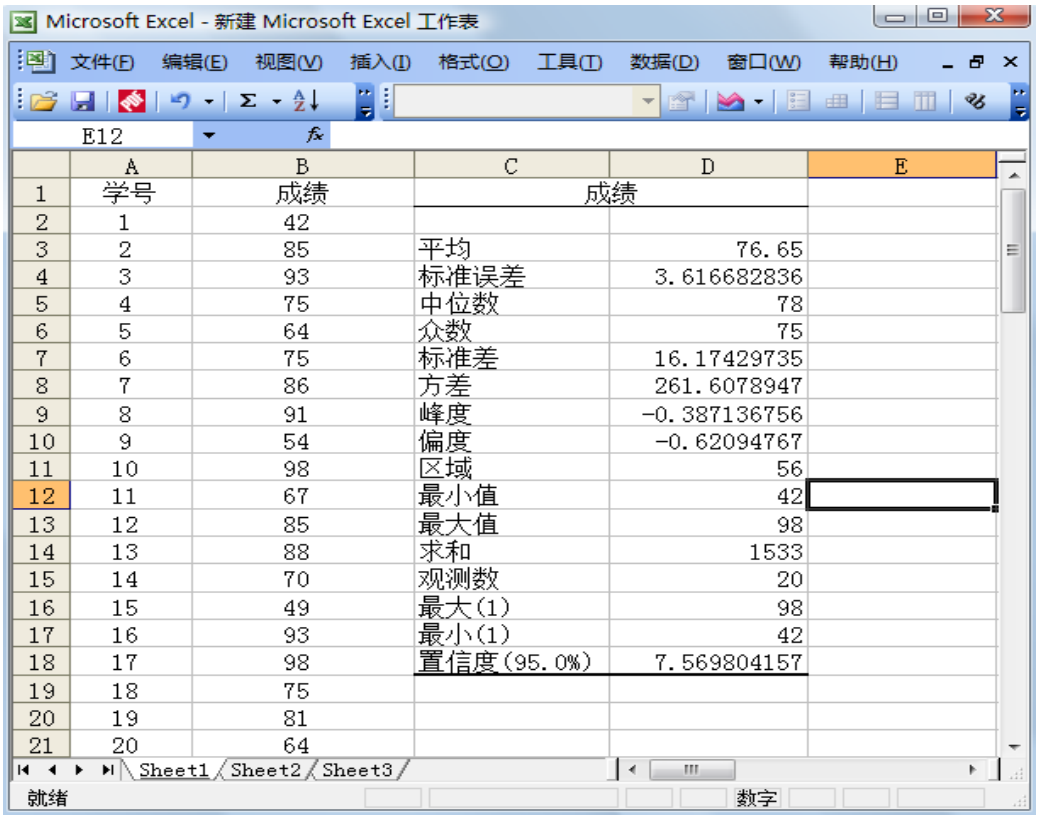


图 8.3.2　　图 8.3.3

输入区域：在此输入待分析数据区域的单元格范围。一般情况下 Excel 会自动根据当前单元格确定待分析数据区域；**分组方式：**如果需要指出输入区域中的数据是按行还是按列排列，则单击“行”或“列”；**标志位于第一行/列：**如果输入区域的第一行中包含标志项(变量名)，则选中“标志位于第一行”复选框；如果输入区域的第一列中包含标志项，则选中“标志位于第一列”；**复选框；**如果输入区域没有标志项，则不选任何复选框，Excel 将在输出表中生成适宜的数据标志；**均值置信度：**若需要输出由样本均值推断总体均值的置信区间，则选中此复选框，然后在右侧的编辑框中，输入所要使用的置信度。例如，置信度 95%可计算出的总体样本均值置信区间为 10，则表示：在 5%的显著水平下总体均值的置信区间为(X -10, X +10)；**第 K 个最大/小值：**如果需要在输出表的某一行中包含每个区域的数据的第 k 个最大/小值，则选中此复选框。然后在右侧的编辑框中，输入 k 的数值；**输出区域：**在此框中可填写输出结果表左上角单元格地址,用于控制输出结果的存放位置。整个输

出结果分为两列,左边一列包含统计标志项，右边一列包含统计值。根据所选择的“分组方式”选项的不同,Excel 将为输入表中的每一行或每一列生成一个两列的统计表;**新工作表：**单击此选项，可在当前工作簿中插入新工作表，并由新工作表的 A1 单元格开始存放计算结果。如果需要给新工作表命名，则在右侧编辑框中键入名称；**新工作簿：**单击此选项，可创建一新工作簿，并在新工作簿的新工作表中存放计算结果；**汇总统计：**指定输出表中生成下列统计结果，则选中此复选框；**这些统计结果有：**平均值、标准误差、中值、众数、标准偏差、方差、峰值、偏斜度、极差（全距）最小值、最大值、总和、样本个数。

（5）填写完“描述统计”对话框之后，按“确定”按扭即可（见图 8.3.4）。

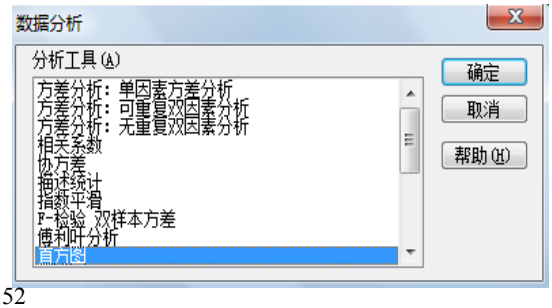


	A	B	C	D	E
1	学号	成绩	成绩		
2	1	42			
3	2	85	平均	76.65	
4	3	93	标准误差	3.616682836	
5	4	75	中位数	78	
6	5	64	众数	75	
7	6	75	标准差	16.17429735	
8	7	86	方差	261.6078947	
9	8	91	峰度	-0.387136756	
10	9	54	偏度	-0.62094767	
11	10	98	区域	56	
12	11	67	最小值	42	
13	12	85	最大值	98	
14	13	88	求和	1533	
15	14	70	观测数	20	
16	15	49	最大(1)	98	
17	16	93	最小(1)	42	
18	17	98	置信度(95.0%)	7.569804157	
19	18	75			
20	19	81			
21	20	64			

图 8.3.4

8.3.2 直方图工具

直方图工具，用于在给定工作表中数据单元格区域和接收区间的情况下，计算数据的个别和累积频率，可以统计有限集中某个数值元素的出现次数。例如，在一个有 50 名学生的班级里，可以通过直方图确定考试成绩的分布情况,它会给出考分出现在指定成绩区间的学生个数,而用户必须把存放分段区间的单元地址范围填写在在直方图工具对话框中的“接收区域”框中。完整的结果通常包括三列和一个频率分布图，第一列是数值的区间范围，第二列是数值分布的频数，第三列是频数分布的累积百分比。



（一）操作步骤

（1）用鼠标点击表中待分析数据的任一单元格。

（2）选择“工具”菜单的“数据分析”子菜单。

（3）用鼠标双击数据分析工具中的“直方图”选项（看图 8.3.5）。

（4）出现“直方图”对话框（见图 8.3.6），对话框内主要选项的含义如下：

输入区域：在此输入待分析数据区域的单元格范围；**接收区域(可选)：**在此输入接收区域的单元格范围，该区域应包含一组可选的用来计算频数的边界值。这些值应当按升序排列。只要存在的话，Excel 将统计在各个相邻边界直之间的数据出现的次数。如果省略此处的接收区域，Excel 将在数据组的最小值和最大值之间创建一组平滑分布的接收区间；**标志：**如果输入区域的第一行或第一列中包含标志项，则选中此复选框；如果输入区域没有标志项，则清除此该复选框，Excel 将在输出表中生成适宜的数据标志；**输出区域：**在此输入结果输出表的左上角单元格的地址。如果输出表将覆盖已有的数据，Excel 会自动确定输出区域的大小并显示信息；**柏拉图：**选中此复选框，可以在输出表中同时显示按降序排列频率数据。如果此复选框被清除，Excel 将只按升序来排列数据；**累积百分比：**选中此复选框，可以在输出结果中添加一列累积百分比数值，并同时在直方图表中添加累积百分比折线。如果清除此选项，则会省略以上结果；**图表输出：**选中此复选框，可以在输出表中同时生成一个嵌入式直方图表。

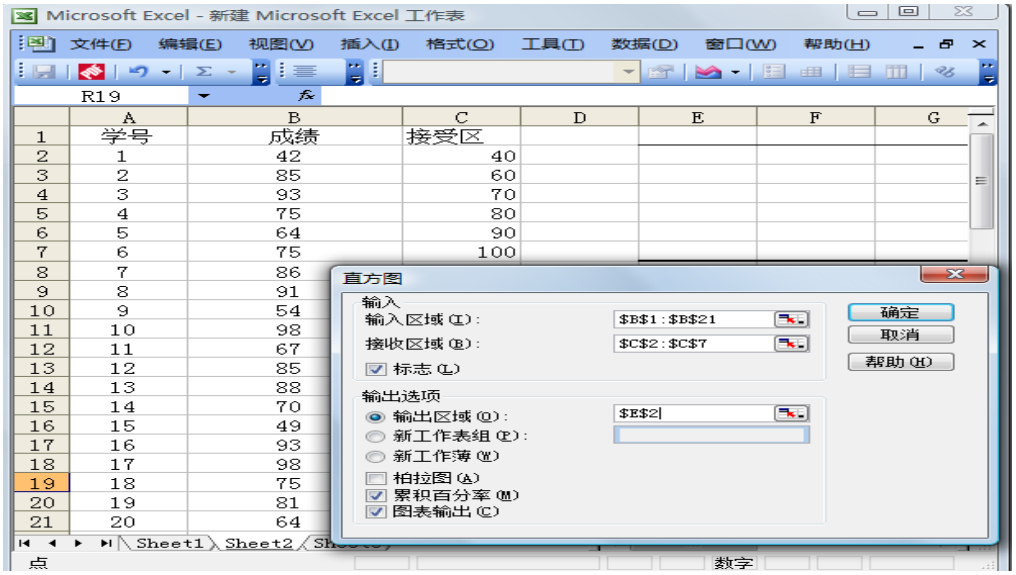
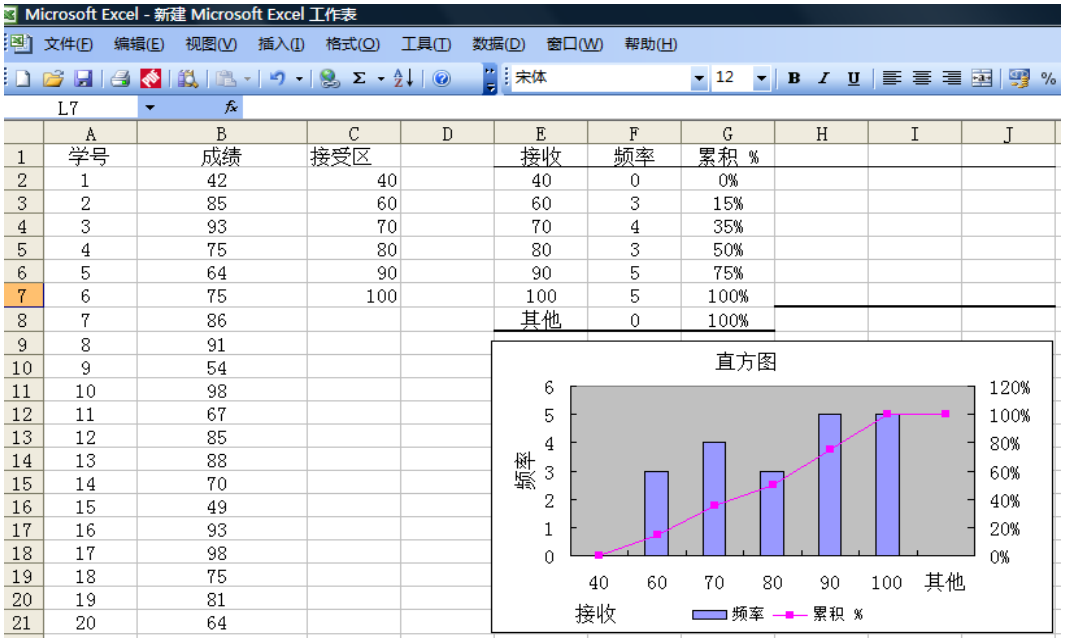


图 8.3.6

（5）按需要填写完“直方图”对话框之后，按“确定”按扭即可。



	A	B	C	D	E	F	G	H	I	J
1	学号	成绩	接受区		接收	频率	累积 %			
2	1	42	40		40	0	0%			
3	2	85	60		60	3	15%			
4	3	93	70		70	4	35%			
5	4	75	80		80	3	50%			
6	5	64	90		90	5	75%			
7	6	75	100		100	5	100%			
8	7	86			其他	0	100%			
9	8	91								
10	9	54								
11	10	98								
12	11	67								
13	12	85								
14	13	88								
15	14	70								
16	15	49								
17	16	93								
18	17	98								
19	18	75								
20	19	81								
21	20	64								

图 8.3.7

8.3.3 绘制两轴折线图

两轴折线图是曲型的组合图标，两条折线绘制在同一个坐标平面上。现在用表 8.2 中提供的数据练习两轴折线图的绘制过程。

表 8.3.2　于田县人口与耕地面积历年数据

年分	耕地面积（万 亩）	人口(万人)	年分	耕地面积(万 亩)	人口(万人)
1949	23.77	81269	1959	34.97	96502
1950	25.36	82326	1960	45.54	97130
1951	26.45	83561	1961	46.98	102176
1952	26.45	85149	1962	40.37	102685
1953	26.45	87022	1963	41.09	103829
1954	26.45	87842	1964	40.05	101652
1955	26.45	89224	1965	41.75	104637
1956	26.95	89923	1966	44.51	106401
1957	28.01	90772	1967	45.55	107638
1958	29.55	92246	1968	45.88	107638

(一)步骤

（1）用鼠标点击表中待分析数据的任一单元格。

（2）选中绘图数据范围后，选择“插入”菜单的“图标”子菜单（看图 8.3.8）。

（3）双击图标导向窗口中的“自定义类型”选项卡，双击“两轴折线图”选项（看图 8.3.9）。

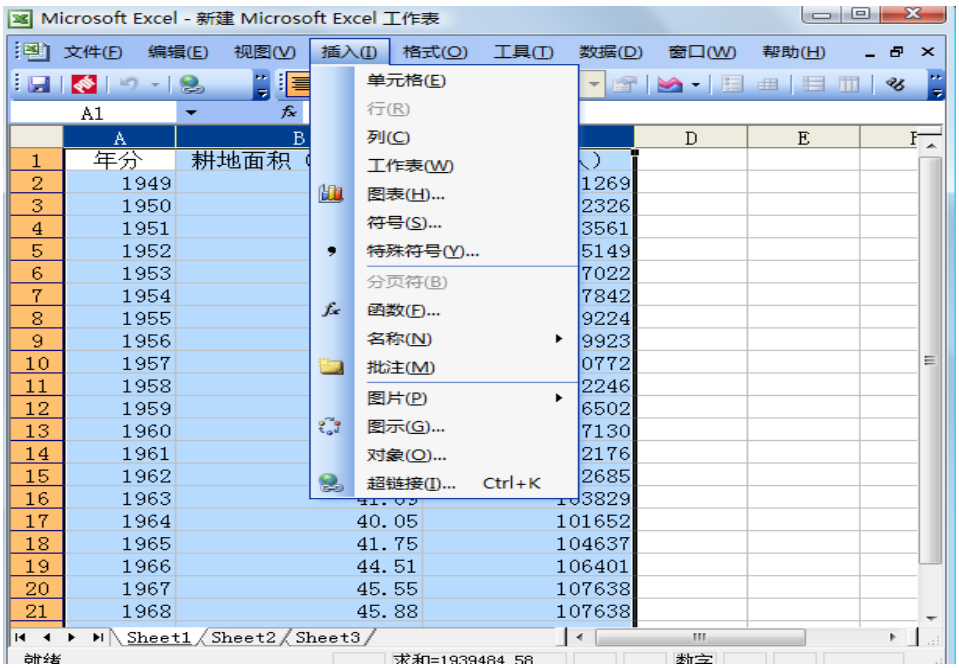


图 8.3.8

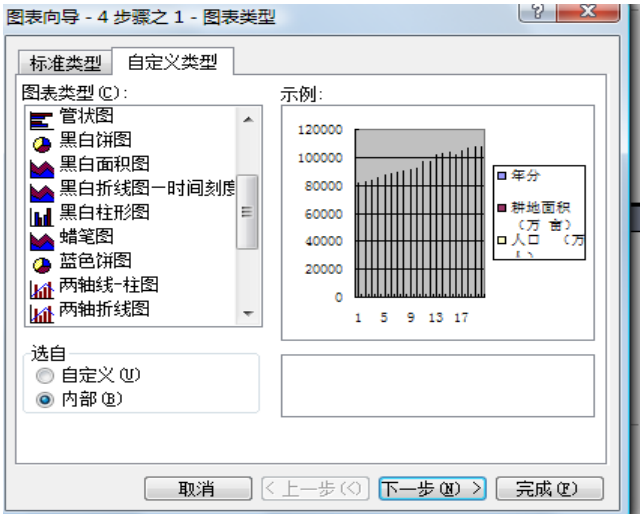


图 8.3.9

（4）按下一步，选择“系列”选项卡，删除原有的全部系列，重新添加自己要建立的系列并选择相应的数据范围（看 8.3.10）。

（5）按下一步，打开图标选项对话框，分别点开标题、坐标轴、网格线、图例、数据标志、数据表选项卡，进行图标输出设计（看图 8.3.11，8.3.12）。

（6）按下一步，打开“图标位置”对话框，根据图表输出的具体位置要求，选择图表位。本题中，我们把图标作为该工作簿中的一个对象来输出（看图 8.3.13）。

（7）点击完成的图表的绘制（看图 8.3.14）。

（8）鼠标指针分别点击在 X,Y 轴上，打开“坐标轴格式”对话框，对已做好的图标进行进一步修改（看图 8.3.15）。

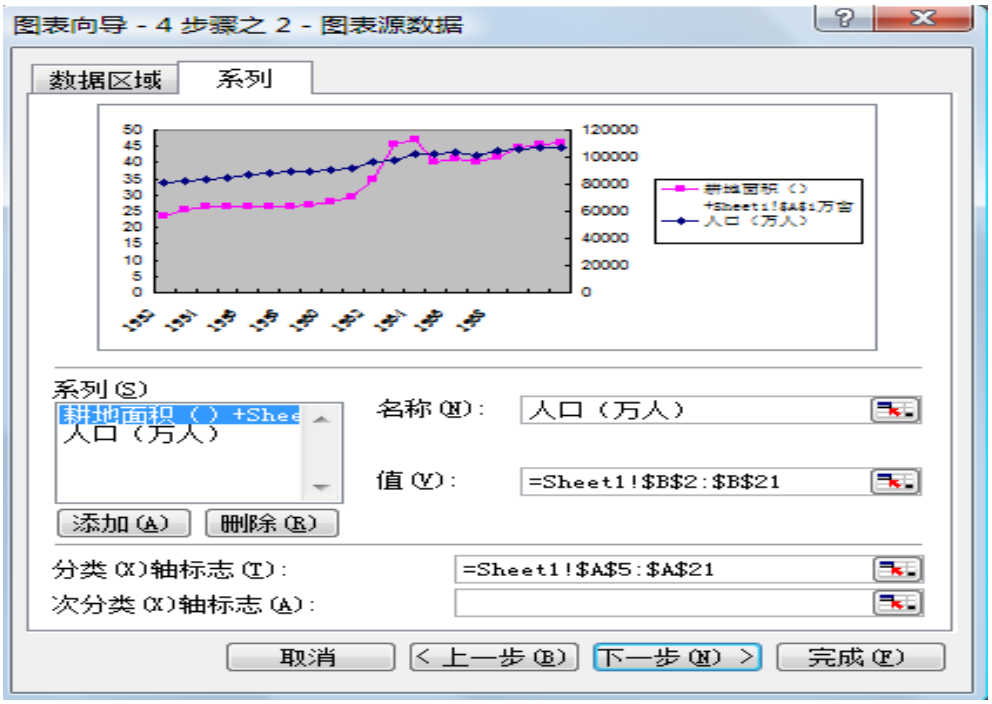


图 8.3.10

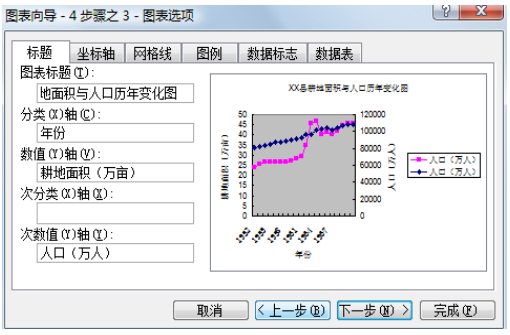


图 8.3.11

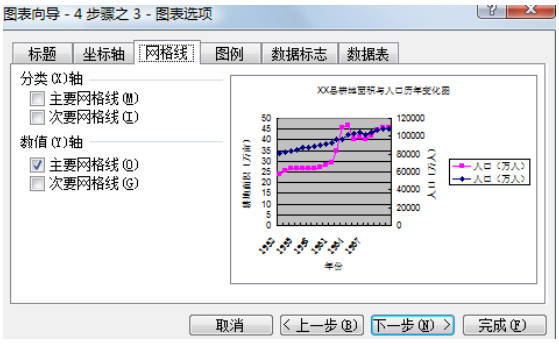


图 8.3.12

图 8.3.13

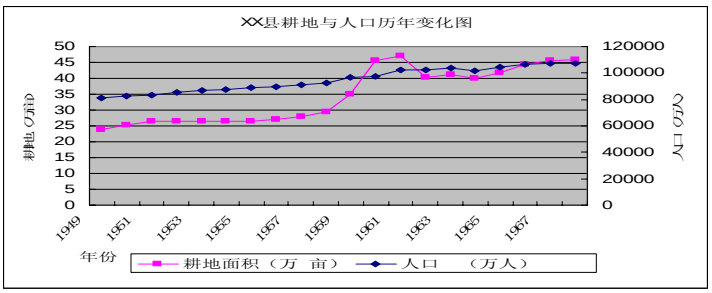


图 8.3.14

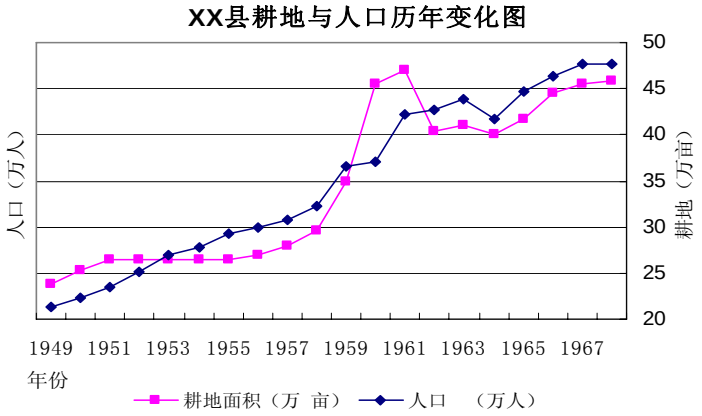

XX县耕地与人口历年变化图
人口（万人）
耕地（万亩）
50
45
40
35
30
25
20
1949 1951 1953 1955 1957 1959 1961 1963 1965 1967
年份
耕地面积（万 亩）
人口 （万人）

图 8.3.15

实验 9 Excel 中统计函数的应用

9.1 实验目的与意义

（1）通过实验室学生进一步熟悉 Excel 的使用技巧。
（2）学会在 Excel 中常用的函数的使用

9.2 基本原理和方法

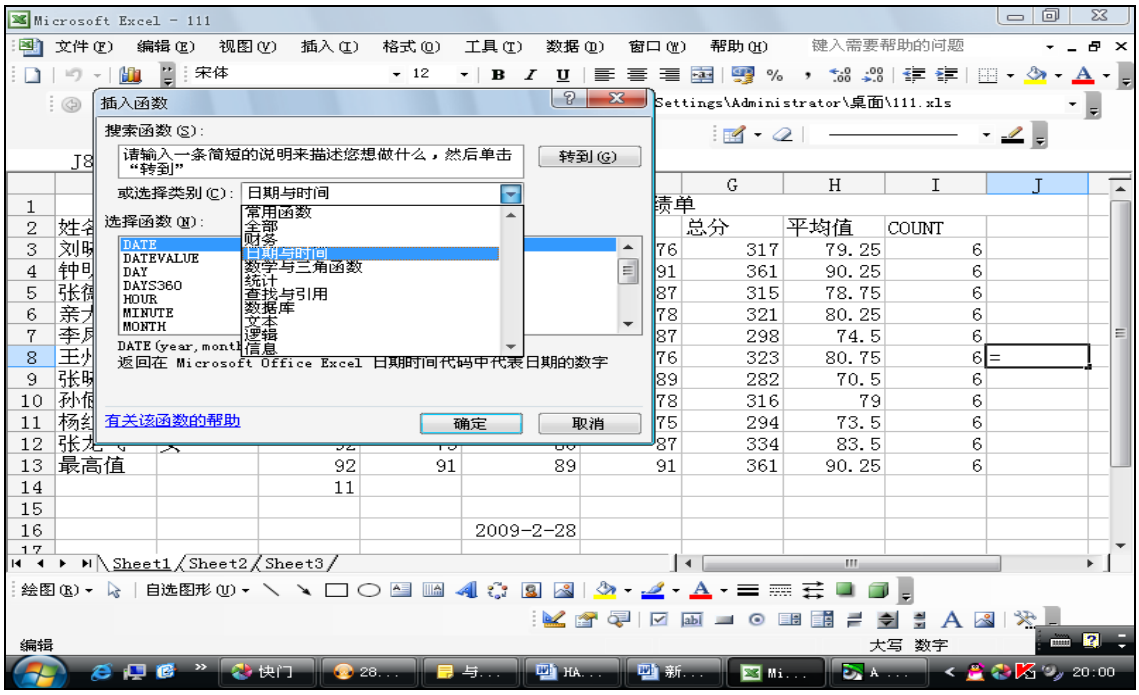
（1）Excel 的使用技巧。

（2）基于 Excel 的常用的函数。

9.3 实验内容及步骤

9.3.1 掌握 Excel 及其主要函数应用

Excel 函数一共有 11 类，分别是数据库函数、日期与时间函数、工程函数、财务函数、信息函数、逻辑函数、查询和引用函数、数学和三角函数、统计函数、文本函数以及用户自定义函数。打开 Excel 主窗口从菜单栏选择**插入**菜单从**插入**菜单中选择**函数子**菜单见到如图所示：



1）.数据库函数--当需要分析数据清单中的数值是否符合特定条件时，可以使用数据库工作表函数。

2）.日期与时间函数--通过日期与时间函数，可以在公式中分析和处理日期值和时间值。

3）.工程函数--工程工作表函数用于工程分析。这类函数中的大多数可分为三种类型：对复数进行处理的函数、在不同的数字系统（如十进制系统、十六进制系统、八进制系统和二进制系统）间进行数值转换的函数、在不同的度量系统中进行数值转换的函数。

4）.财务函数--财务函数可以进行一般的财务计算，如确定贷款的支付额、投资的未来值或净现值，以及债券或息票的价值。财务函数中常见的参数： 未来值 (fv)--在所有付款发生后的投资或贷款的价值。期间数 (nper)--投资的总支付期间数。付款 (pmt)--对于一项投资或贷款的定期支付数额。现值 (pv)--在投资期初的投资或贷款的价值。例如，贷款的现值为所借入的本金数额。利率 (rate)--投资或贷款的利率或贴现率。类型 (type)--付款期间内进行支付的间隔，如在月初或月末。

5）.信息函数--可以使用信息工作表函数确定存储在单元格中的数据的类型。信息函数包含一组称为 IS 的工作表函数，在单元格满足条件时返回 TRUE。例如，如果单元格包含一个偶数值，ISEVEN 工作表函数返回 TRUE。如果需要确定某个单元格区域中是否存在空白单元格，可以使用 COUNTBLANK 工作表函数对单元格区域中的空白单元格进行计数，或者使用 ISBLANK 工作表函数确定区域中的某个单元格是否为空。

6）.逻辑函数--使用逻辑函数可以进行真假值判断，或者进行复合检验。例如，可以使用 IF 函数确定条件为真还是假，并由此返回不同的数值。

7）.查询和引用函数--当需要在数据清单或表格中查找特定数值，或者需要查找某一单元格的引用时，可以使用查询和引用工作表函数。例如，如果需要在表格中查找与第一列中的值相匹配的数值，可以使用 VLOOKUP 工作表函数。如果需要确定数据清单中数值的位置，可以使用 MATCH 工作表函数。

8）.数学和三角函数--通过数学和三角函数，可以处理简单的计算，例如对数字取整、计算单元格区域中的数值总和或复杂计算。

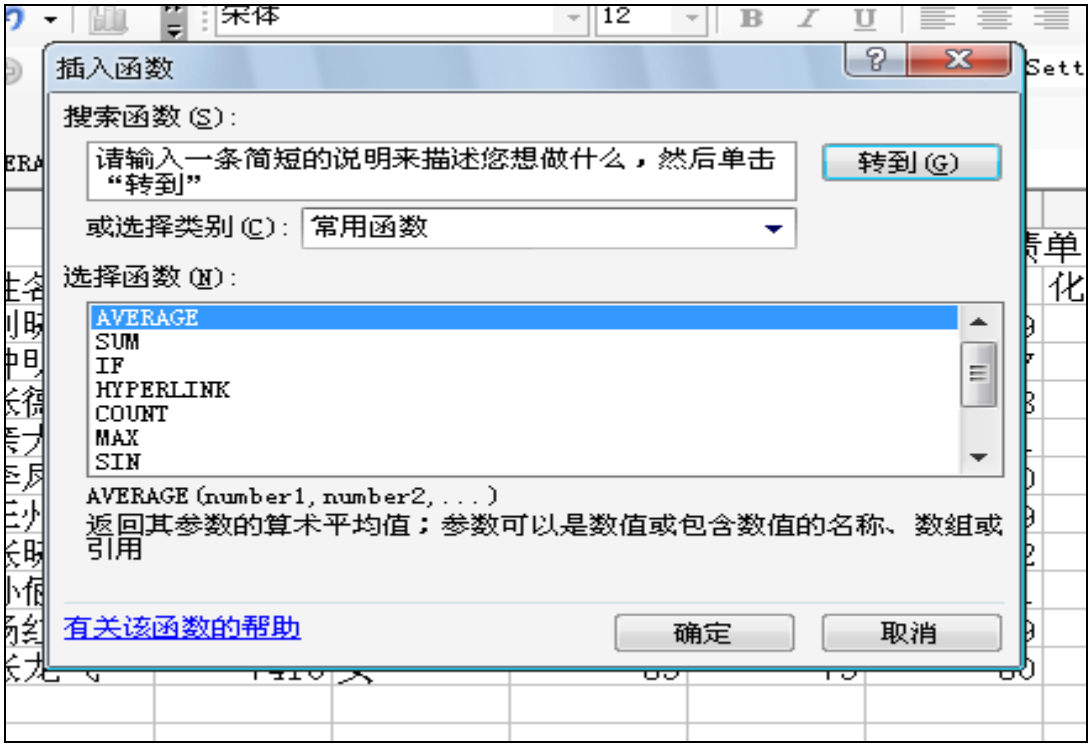
9）.统计函数--统计工作表函数用于对数据区域进行统计分析。例如，统计工作表函数可以提供由一组给定值绘制出的直线的相关信息，如直线的斜率和 y 轴截距，或构成直线的实际点数值。

10）.文本函数--通过文本函数，可以在公式中处理文字串。例如，可以改变大小写或确定文字串的长度。可以将日期插入文字串或连接在文字串上。下面的公式为一个示例，借以说明如何使用函数 TODAY 和函数 TEXT 来创建一条信息，该信息包含着当前日期并将日期以"dd-mm-yy"的格式表示。

11）.用户自定义函数--如果要在公式或计算中使用特别复杂的计算，而工作表函数又无法满足需要，则需要创建用户自定义函数。这些函数，称为用户自定义函数，可以通过使用 Visual Basic for Applications 来创建。

9.3.2 常用函数的应用并实例

Excel 的统计工作表函数用于对数据区域进行统计分析。在下面用常用的函数的实例应用来介绍 Excel 函数的实际应用。打开 Excel 主窗口从菜单栏选择【插入】菜单从【插入】菜单中选择【函数】子菜单。选择【函数】子菜单并弹出【插入函数】对话框。看到如图所示：



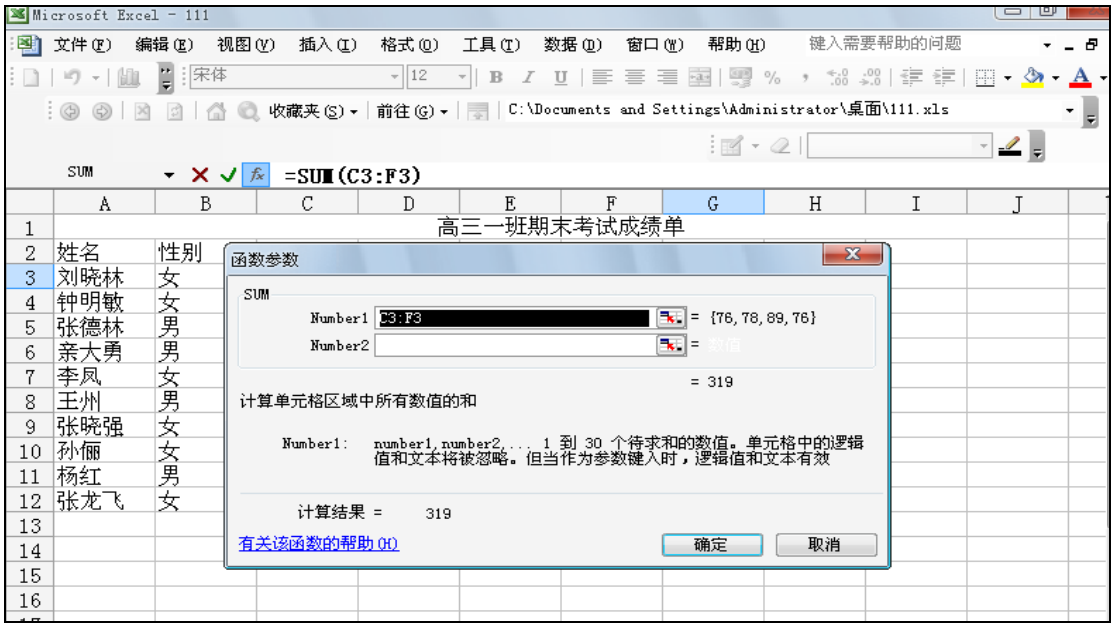
9.3.2.1 SUM 函数的应用

下面使用 SUM 函数来计算图 1 成绩单中每个同学的总分。

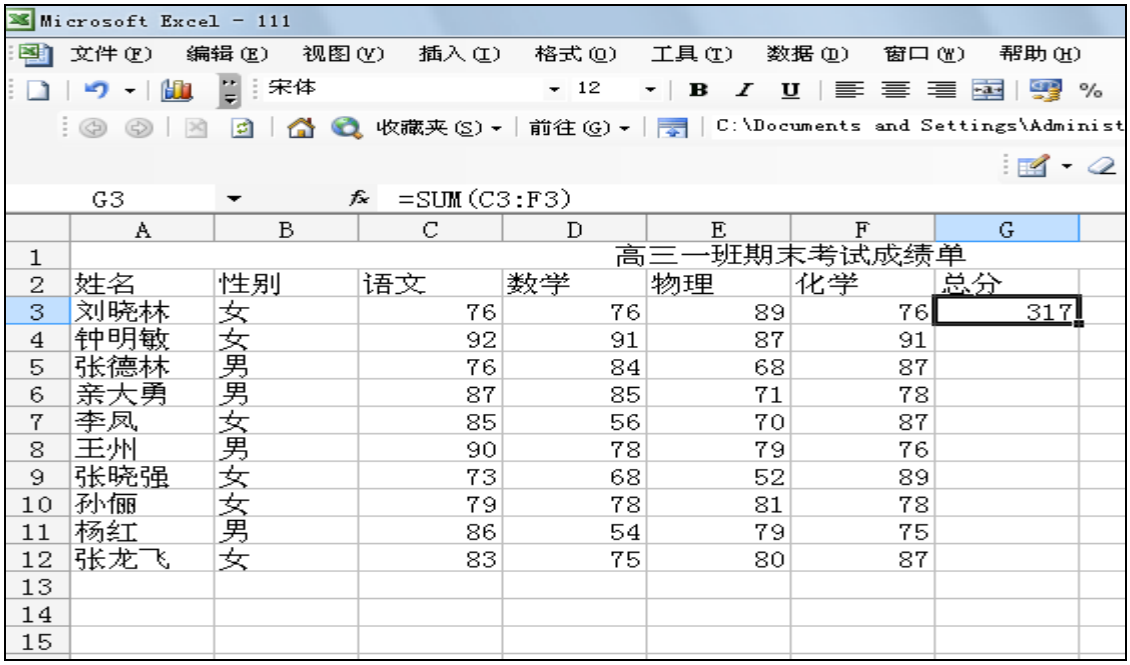
1) 单击 G3 单元格。

2）单击编辑栏中的“=”，弹出函数选项板。

3）单击编辑栏中函数名右边的下拉箭头，在列表中选择 SUM 函数，如图 2 所示。

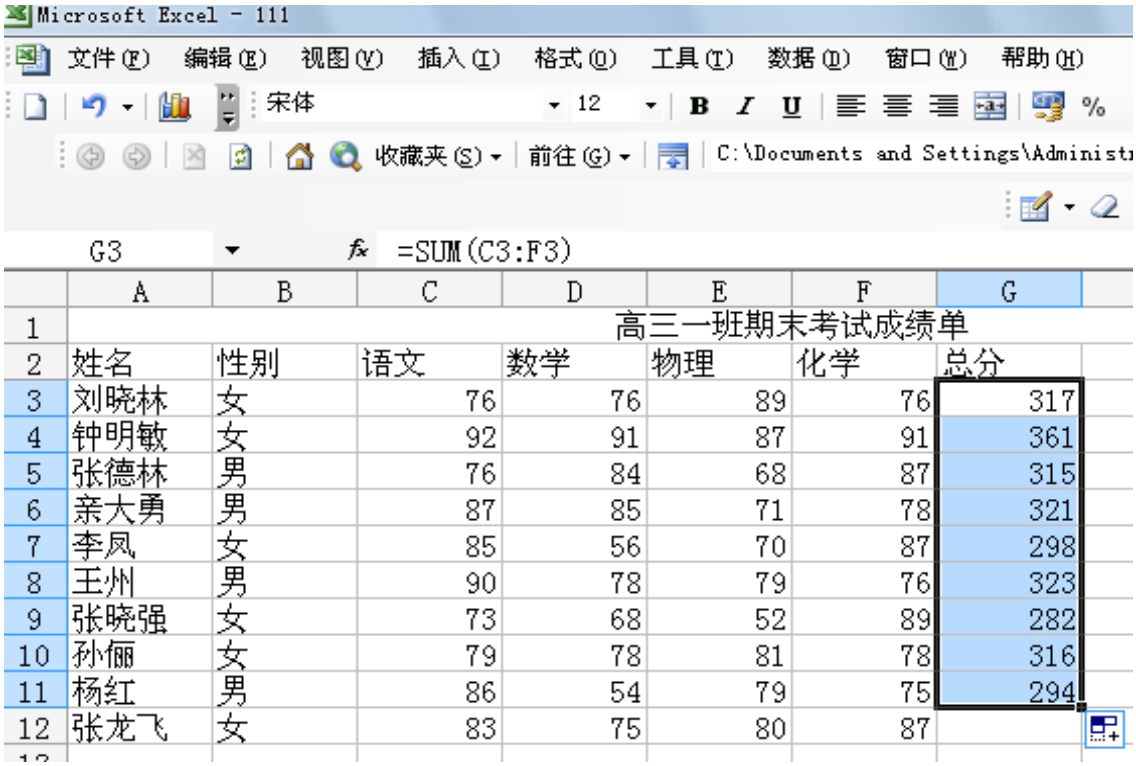


4）再单击工作表中参数输入栏右边的按钮，回到函数选项板。在选项板中单击【确定】按钮，G3 单元格中显示出刘晓林同学的总分。



	A	B	C	D	E	F	G
1	高三一班期末考试成绩单						
2	姓名	性别	语文	数学	物理	化学	总分
3	刘晓林	女	76	76	89	76	317
4	钟明敏	女	92	91	87	91	
5	张德林	男	76	84	68	87	
6	亲大勇	男	87	85	71	78	
7	李凤	女	85	56	70	87	
8	王州	男	90	78	79	76	
9	张晓强	女	73	68	52	89	
10	孙俪	女	79	78	81	78	
11	杨红	男	86	54	79	75	
12	张龙飞	女	83	75	80	87	
13							
14							
15							

5） 算完 G3 单元格的总成绩以后在该单元格右下角按鼠标左键这时在该单元格右下角出现“ ”符号，按鼠标左键并不放拉到 G12 可以得到刘晓林到赵飞龙 10 个学生的总成绩。



	A	B	C	D	E	F	G
1	高三一班期末考试成绩单						
2	姓名	性别	语文	数学	物理	化学	总分
3	刘晓林	女	76	76	89	76	317
4	钟明敏	女	92	91	87	91	361
5	张德林	男	76	84	68	87	315
6	亲大勇	男	87	85	71	78	321
7	李凤	女	85	56	70	87	298
8	王州	男	90	78	79	76	323
9	张晓强	女	73	68	52	89	282
10	孙俪	女	79	78	81	78	316
11	杨红	男	86	54	79	75	294
12	张龙飞	女	83	75	80	87	

9.3.2.2 用于求平均值的统计函数 AVERAGE

求参数的算术平均值函数 AVERAGE

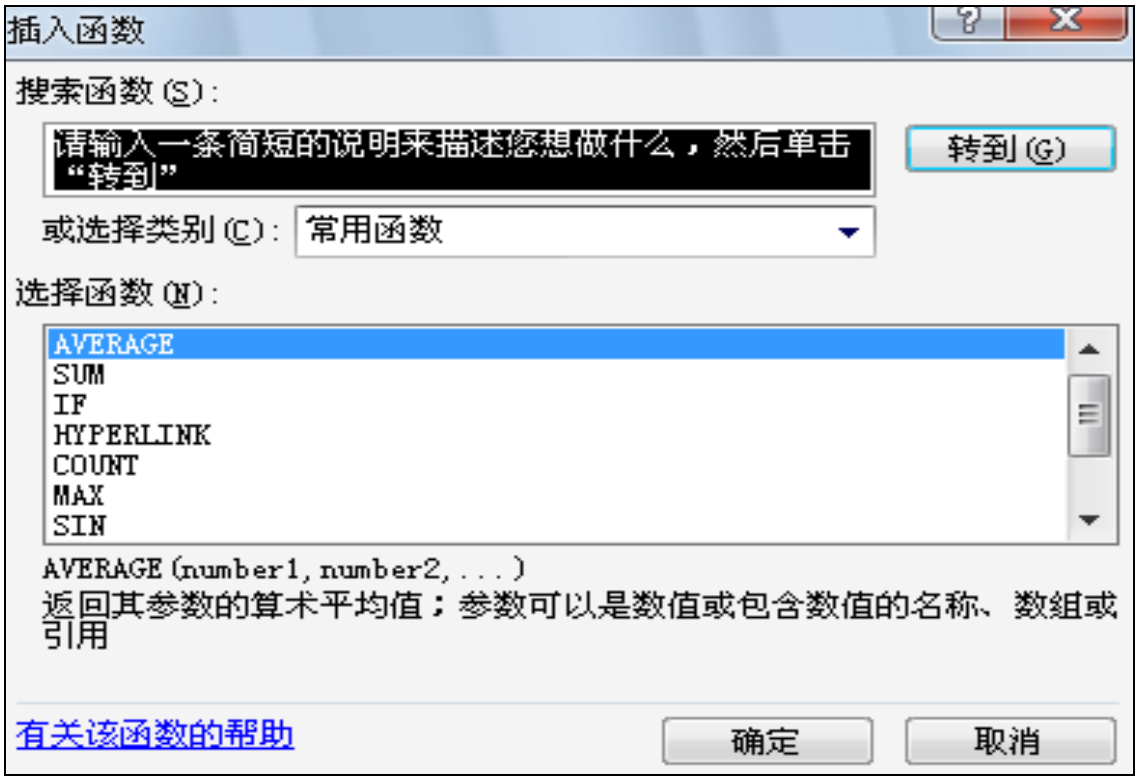
语法形式为 AVERAGE(number1,number2, ...)

其中 Number1, number2, ...为要计算平均值的 1～30 个参数。这些参数可以是数字，或者是涉及数字的名称、数组或引用。如果数组或单元格引用参数中有文字、逻辑值或空单元格，则忽略其值。但是，如果单元格包含零值则计算在内。

下面用 AVERAGE 函数来计算成绩单中的平均分。

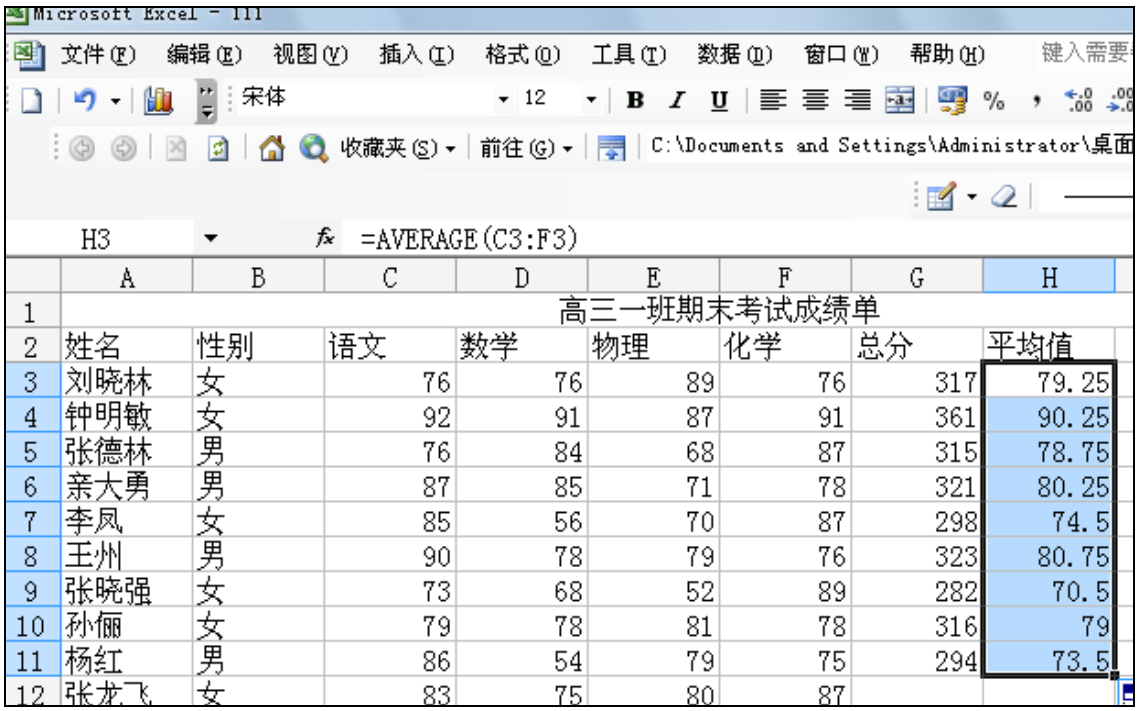
1）单击 H3 单元格。

2） 单击【插入】中的选择，从菜单栏选择**插入**菜单从**插入**菜单中选择**函数子**菜单并选择 AVERAGE 函数，如图所示。单击【确定】按钮



3）再单击工作表中参数输入栏右边的按钮，回到函数选项板。在选项板中单击【确定】按钮，G3 单元格中显示出刘晓林同学的平均值。

4)算完 G3 单元格的平均成绩以后在该单元格右下角按鼠标左键这时在该单元格右下角出现“ ”符号，按鼠标左键并不放拉到 G12 可以得到刘晓林到赵飞龙 10 个学生的平均成绩。



	A	B	C	D	E	F	G	H
1	高三一班期末考试成绩单							
2	姓名	性别	语文	数学	物理	化学	总分	平均值
3	刘晓林	女	76	76	89	76	317	79.25
4	钟明敏	女	92	91	87	91	361	90.25
5	张德林	男	76	84	68	87	315	78.75
6	亲大勇	男	87	85	71	78	321	80.25
7	李凤	女	85	56	70	87	298	74.5
8	王州	男	90	78	79	76	323	80.75
9	张晓强	女	73	68	52	89	282	70.5
10	孙俪	女	79	78	81	78	316	79
11	杨红	男	86	54	79	75	294	73.5
12	张龙飞	女	83	75	80	87		

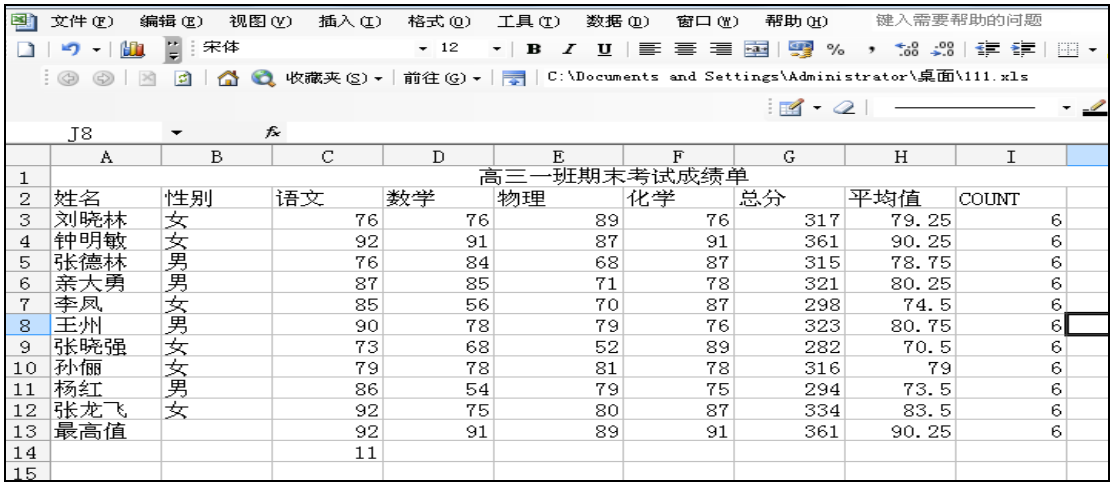
9.3.2.3 用于求单元格个数的统计函数COUNT

语法形式为 COUNT(value1,value2, ...)

其中 Value1, value2, ...为包含或引用各种类型数据的参数（1～30 个），但只有数字类型的数据才被计数。函数 COUNT 在计数时，将把数字、空值、逻辑值、日期或以文字代表的数计算进去；但是错误值或其他无法转化成数字的文字则被忽略。 如果参数是一个数组或引用，那么只统计数组或引用中的数字；数组中或引用的空单元格、逻辑值、文字或错误值都将忽略。如果要统计逻辑值、文字或错误值，应当使用函数 COUNTA。 举例说明 COUNT 函数的用途，示例中也列举了带 A 的函数 COUNTA 的用途。仍以上例为例，要计算一共有多少评委参与评分(用函数 COUNTA)，以及有几个评委给出了有效分数(用函数 COUNT)。

下面用 COUNT 函数来计算成绩单中的数值个数

1) 单击任意单元格。
2) 单击【插入】中的选择，从菜单栏选择**插入**菜单从**插入**菜单中选择**函数子**菜单并选择 count 函数。单击【确定】按钮。得到数值个数。



	A	B	C	D	E	F	G	H	I
1	高三一班期末考试成绩单								
2	姓名	性别	语文	数学	物理	化学	总分	平均值	COUNT
3	刘晓林	女	76	76	89	76	317	79.25	6
4	钟明敏	女	92	91	87	91	361	90.25	6
5	张德林	男	76	84	68	87	315	78.75	6
6	亲大勇	男	87	85	71	78	321	80.25	6
7	李凤	女	85	56	70	87	298	74.5	6
8	王州	男	90	78	79	76	323	80.75	6
9	张晓强	女	73	68	52	89	282	70.5	6
10	孙俪	女	79	78	81	78	316	79	6
11	杨红	男	86	54	79	75	294	73.5	6
12	张龙飞	女	92	75	80	87	334	83.5	6
13	最高值		92	91	89	91	361	90.25	6
14			11						
15									

一组用于求数据集的满足不同要求的数值的函数

9.3.2.4 求数据集的最大值 MAX 与最小值 MIN

这两个函数MAX、MIN就是用来求解数据集的极值（即最大值、最小值）。函数的用法非常简单。语法形式为 函数（number1,number2,...），其中Number1,number2,... 为需要找出最大数值的 1 到 30 个数值。如果要计算数组或引用中的空白单元格、逻辑值或文本将被忽略。因此如果逻辑值和文本不能忽略，请使用带A的函数MAXA或者MINA 来代替。

求数据集中第K个最大值LARGE与第k个最小值SMALL

这两个函数LARGE、SMALL与MAX、MIN非常想像，区别在于它们返回的不是极值，而是第K个值。语法形式为：函数(array,k)，其中Array为需要找到第 k 个最小值的数组或数字型数据区域。K为返回的数据在数组或数据区域里的位置（如果是LARGE为从大到小排，若为SMALL函数则从小到大排）。 说到这，大家可以想得到吧。如果K=1 或者K=n（假定数据集中有n个数据）的时候，是不是就可以返回数据集的最大值或者最小值了呢。

求数据集中的中位数MEDIAN

MEDIAN函数返回给定数值集合的中位数。所谓中位数是指在一组数据中居于中间的数，换句话说，在这组数据中，有一半的数据比它大，有一半的数据比它小。 语法形式为MEDIAN(number1,number2, ...)其中Number1, number2,...是需要找出中位数的 1 到 30 个数字参数。如果数组或引用参数中包含有文字、逻辑值或空白单元格，则忽略这些值，但是其值为零的单元格会计算在内。

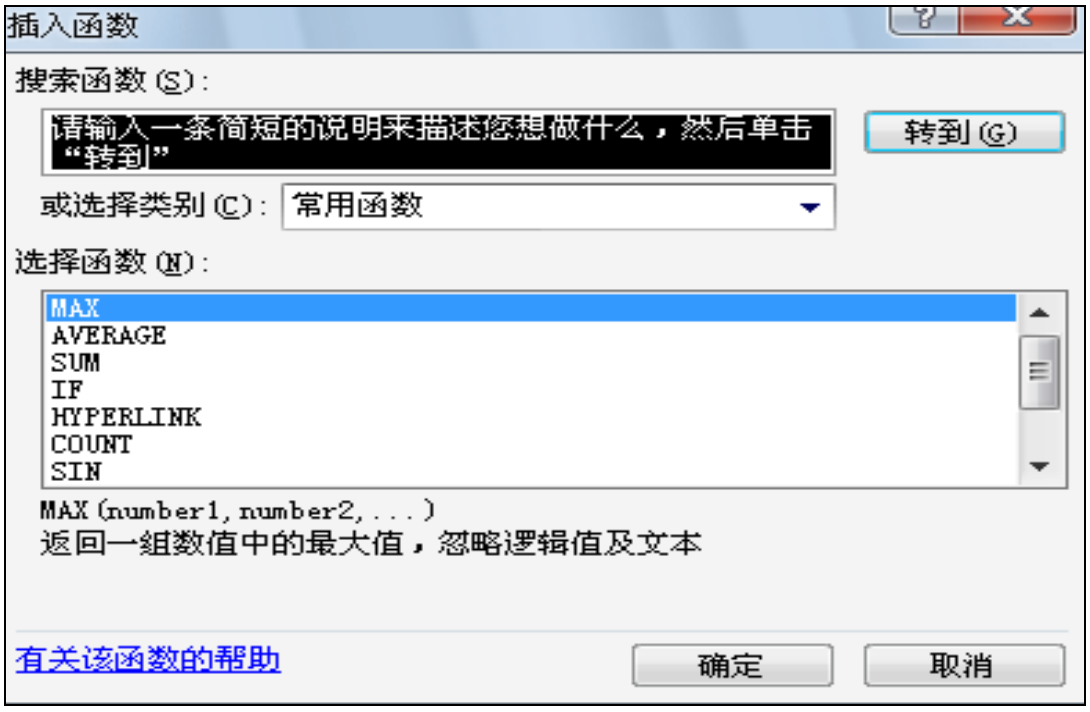
需要注意的是，如果参数集合中包含有偶数个数字，函数 MEDIAN 将返回位于中间的两个数的平均值。

求数据集中出现频率最多的数MODE

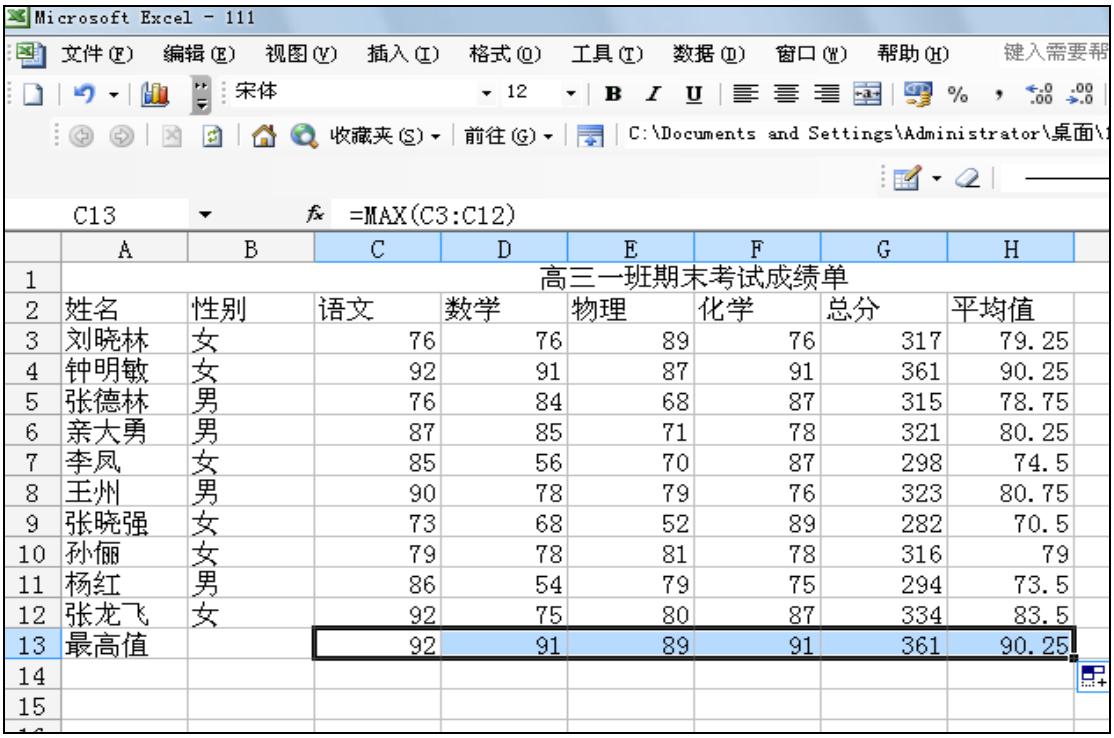
MODE函数用来返回在某一数组或数据区域中出现频率最多的数值。跟 MEDIAN 一样，MODE 也是一个位置测量函数。 语法形式为MODE(number1,number2, ...)其中Number1, number2, ... 是用于众数（众数指在一组数值中出现频率最高的数值）计算的 1 到 30 个参数，也可以使用单一数组（即对数组区域的引用）来代替由逗号分隔的参数。

下面用 MAX 函数来计算成绩单中每门功课，平均值和总成绩的最高分。

还是使用数组公式第一步和第二步跟上面的两个函数一样从菜单栏选择**插入**菜单从**插入**菜单中选择**函数子**菜单并选择 MAX 函数，如图所示。单击【确定】按钮



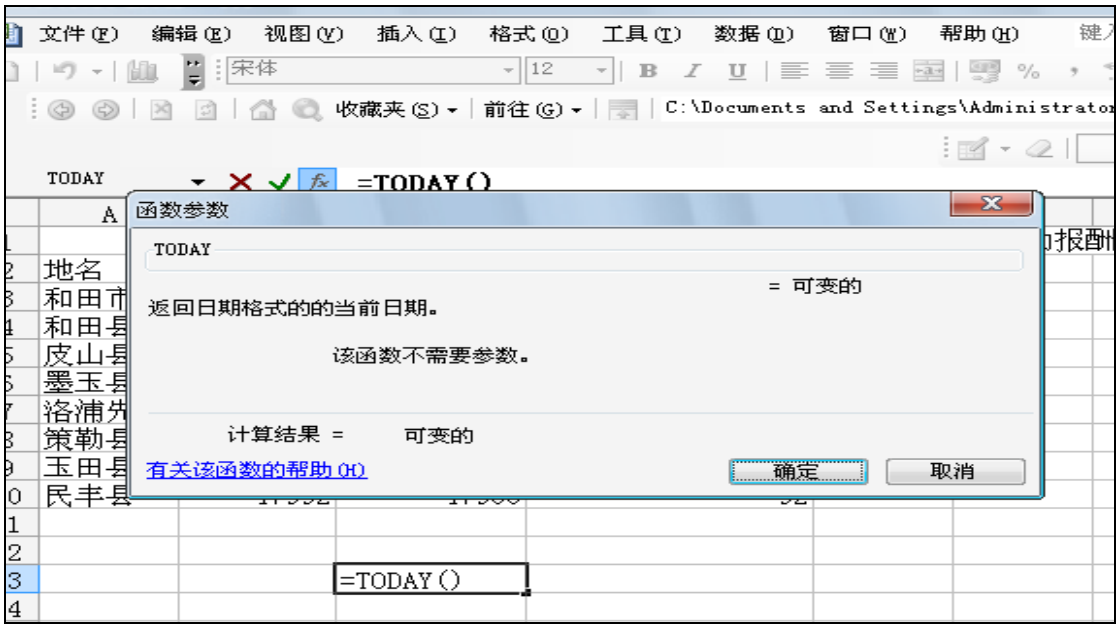
1) 再单击工作表中参数输入栏右边的按钮，回到函数选项板。在选项板中单击【确定】按钮，C13 单元格中显示出从刘晓林同学到张龙飞同学 10 个人里面在语文成绩上最高的成绩。
2) 算完C元格中最高成绩以后在C13单元格右下角按鼠标左键这时在该单元格右下角出现“  “符号，按鼠标左键并不放拉到 G13 可以得到刘晓林到赵飞龙 10 个学生的语文，数学，物理，化学，总成绩，平均成绩 的最高值。



9.3.2.5 TODAY 函数的应用

用户可以直接输入制表日期，也可以应用 TODAY 函数填充制表日期。

单击任意的一个单元格。从菜单栏选择**插入**菜单从**插入**菜单中选择**函数子**菜单并选择 TODAY 函数。单击【确定】按钮



9.3.2.6 函数 CRITBINOM：

1） 说明：函数 CRITBINOM 可称为 BINOMDIST 的逆向函数，它返回使累积二项式分布概率 P(X<=x)大于等于临界概率值的最小值。

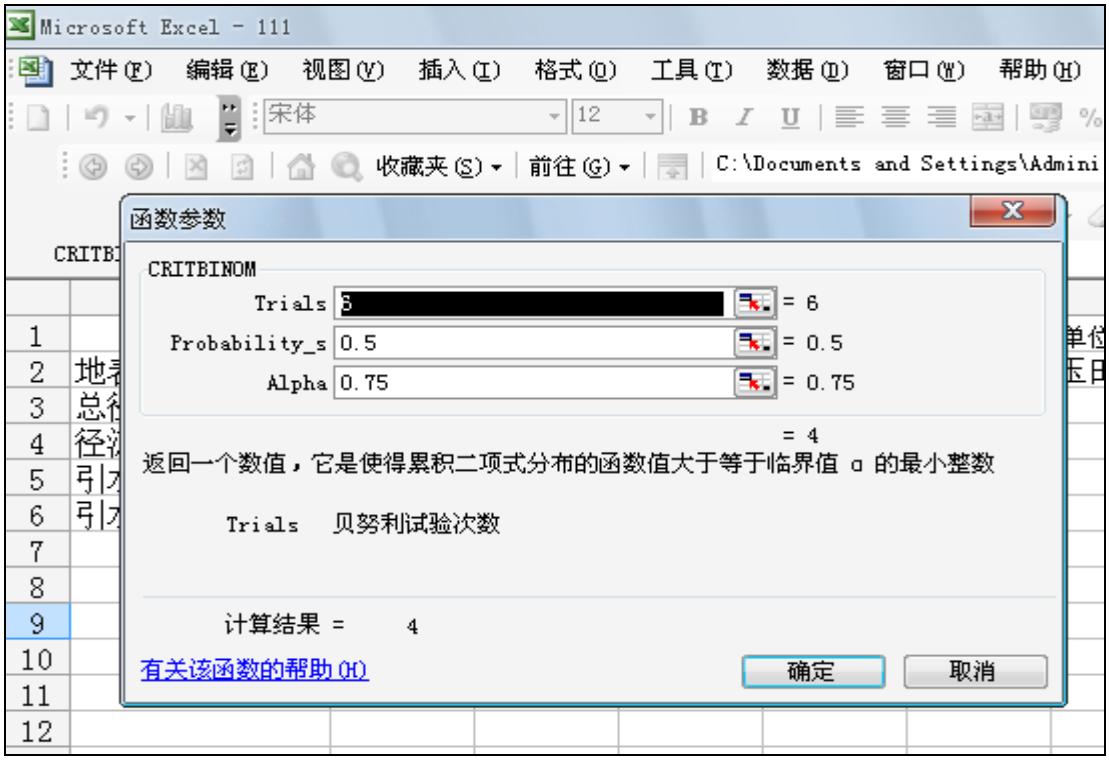
2） 语法：CRITBINOM(trials,probability_s,alpha)

Trials：贝努利实验次数。

Probability_s：一次试验中成功的概率。

Alpha： 临界概率。

下面实例来说明:先从【插入】中选择【函数】并函数子菜单中选择【CRITBINOM】(6,0.5,0.75)等于 4，



表明如果每次试验成功的概率为 0.5,那么 6 次试验中成功的次数小于或等于 4 的概率恰好超过或等于 0.75 。

9.3.2.7 函数 HYPGEOMDIST：

1） 说明：函数 HYPGEOMDIST 返回超几何分布。给定样本容量、总体容量和样本总体中成功的次数，函数 HYPGEOMDIST 返回样本取得给定成功次数的概率。使用函数 HYPGEOMDIST 可以解决有限总体的问题，其中每个观察值或者为成功或者为失败，且给定样本区间的所有子集有相等的发生概率。

2） 语法：HYPGEOMDIST(sample_s,number_sample,population_s,number_population)

Sample_s： 样本中成功的次数。

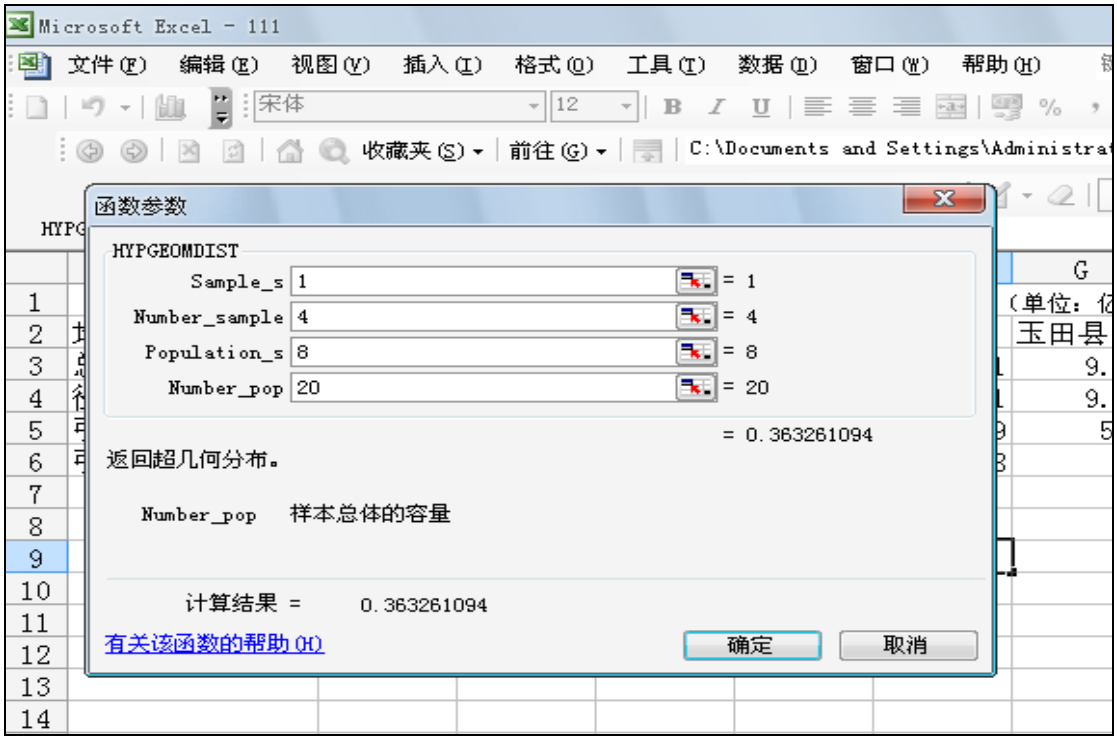
Number_sample：样本容量。

Population_s： 样本总体中成功的次数。

Number_population： 样本总体的容量。

3） 举例：容器里有 20 块巧克力，8 块是焦糖的，其余 12 块是果仁的。

如果从中随机选出 4 块:



从上面面函数计算式计算出只有一块是焦糖巧克力的概率：HYPGEOMDIST(1,4,8,20)=0.363261。

9.3.2.8 函数 NEGBINOMDIST：

1） 说明：函数 NEGBINOMDIST 返回负二项式分布。当每次试验成功概率固时，函数 NEGBINOMDIST 返回在到达指定次数成功之前，出现 n 次失败的概率。此函数与二项式分布相似，只是它的成功次数固定，试验总数为变量。与二项分布类似的是，试验次数被假设为自变量。

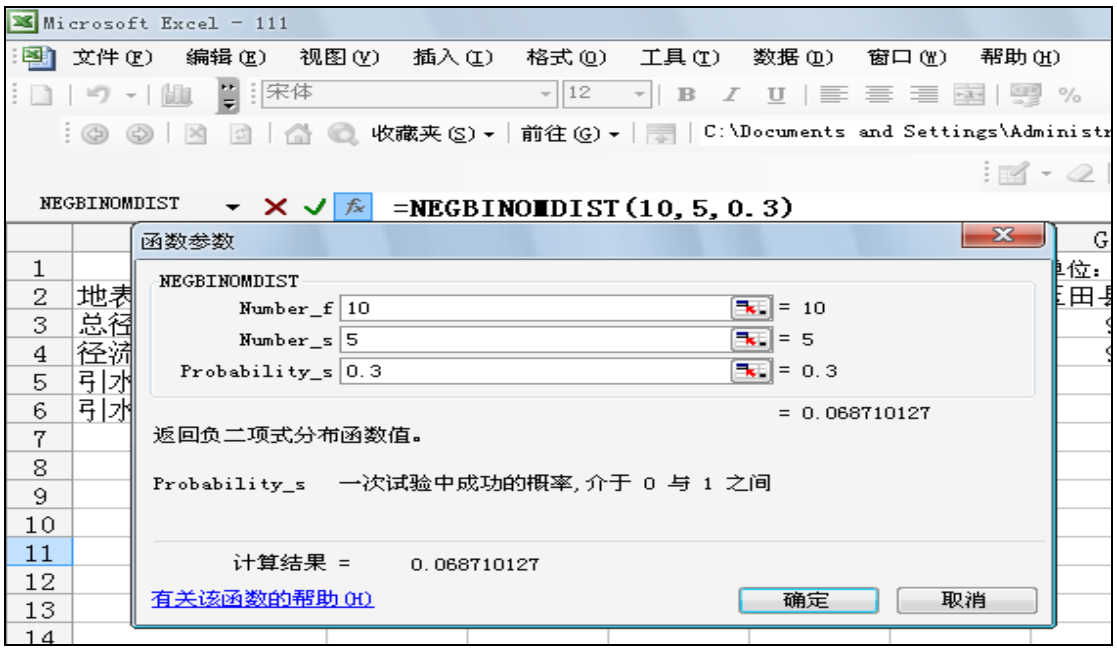
2） 语法：NEGBINOMDIST(number_f,number_s,probability_s)

Number_f：失败次数。

Number_s：成功的临界次数。

Probability_s：成功的概率。

3） 举例：例如，如果要找出 5 个反应敏捷的人，且已知具有这种特征的候选人的概率为 0.3。以下公式将计算出在找到 5 个合格候选人之前，需要面试 10 个候选人的概率：



从上面的实验可以知道在这次试验中的失败概率 0.06871

NEGBINOMDIST(10,5,0.3)= 0.06871

9.3.2.9 函数 POISSON:

1）说明：函数 POISSON 返回泊松分布。泊松分布通常用于预测一段时间内事件发生指定次数的概率，比如一分钟内通过收费站的轿车的数量为 n 的概率。

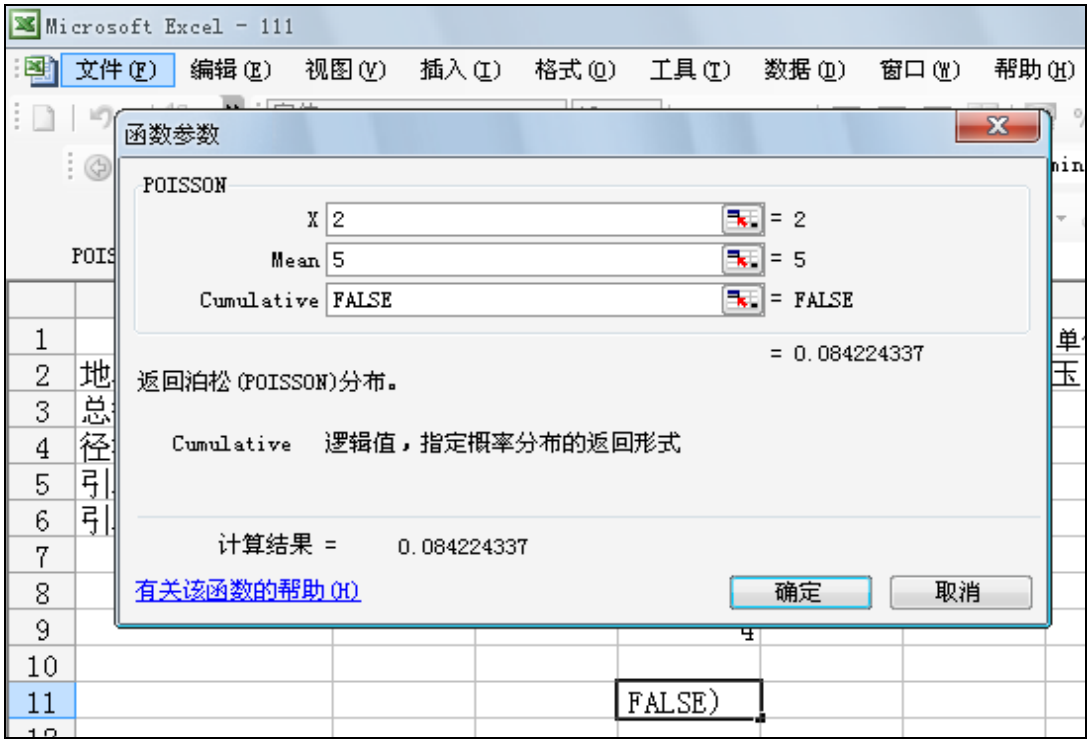
2） 语法：POISSON(x,mean,cumulative)

X： 事件数。

Mean： 期望值。

Cumulative： 为一逻辑值，确定所返回的概率分布形式。如果 cumulative 为 TRUE，函数 POISSON 返回累积分布函数，即，随机事件发生的次数在 0 和 x 之间（包含 0 和 1）；如果为 FALSE，则返回概率密度函数，即，随机事件发生的次数恰好为 x。

3）举例：



POISSON(2,5,FALSE)=0.084224 表明，若某一收费站每分种通过的轿车平均数量为 5 辆，那么某一分钟通只 2 辆的概率为 0.084224。

9.3.2.10 正态分布函数 NORMDIST：

1） 说明：正态分布在模拟现实世界过程和描述随机样本平均值的不确定度时有广泛的用途。函数 NORMDIST 返回给定平均值和标准偏差的正态分布的累积函数。同样可以用类似“七”中的方法，利用 NORMDIST 函数建立正态分布密度函数图，这里不再赘述。

2） 语法：NORMDIST(x,mean,standard_dev,cumulative)

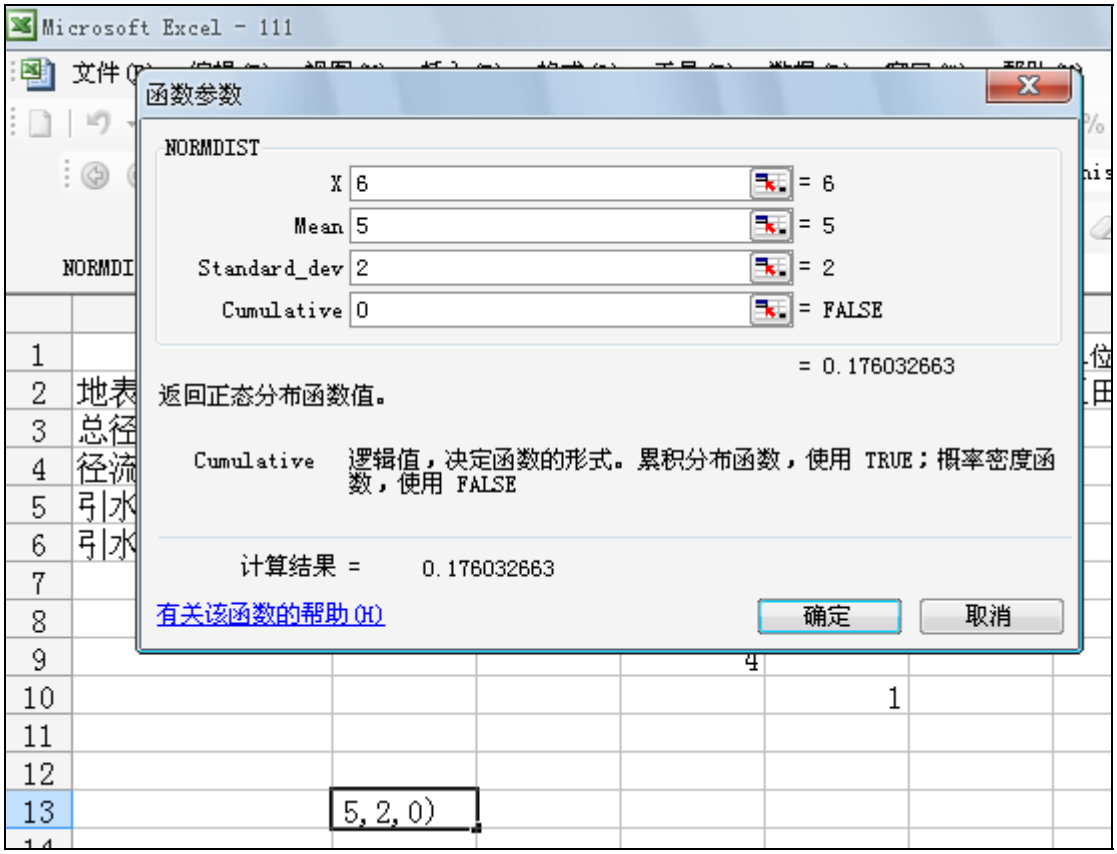
X： 为需要计算其分布的数值。

Mean： 分布的算术平均值。

Standard_dev：分布的标准偏差。

Cumulative： 为一逻辑值，指明函数的形式。如果 cumulative 为 TRUE，函数 NORMDIST 返回累积分布函数；如果为 FALSE，返回概率密度函数。

3） 举例：



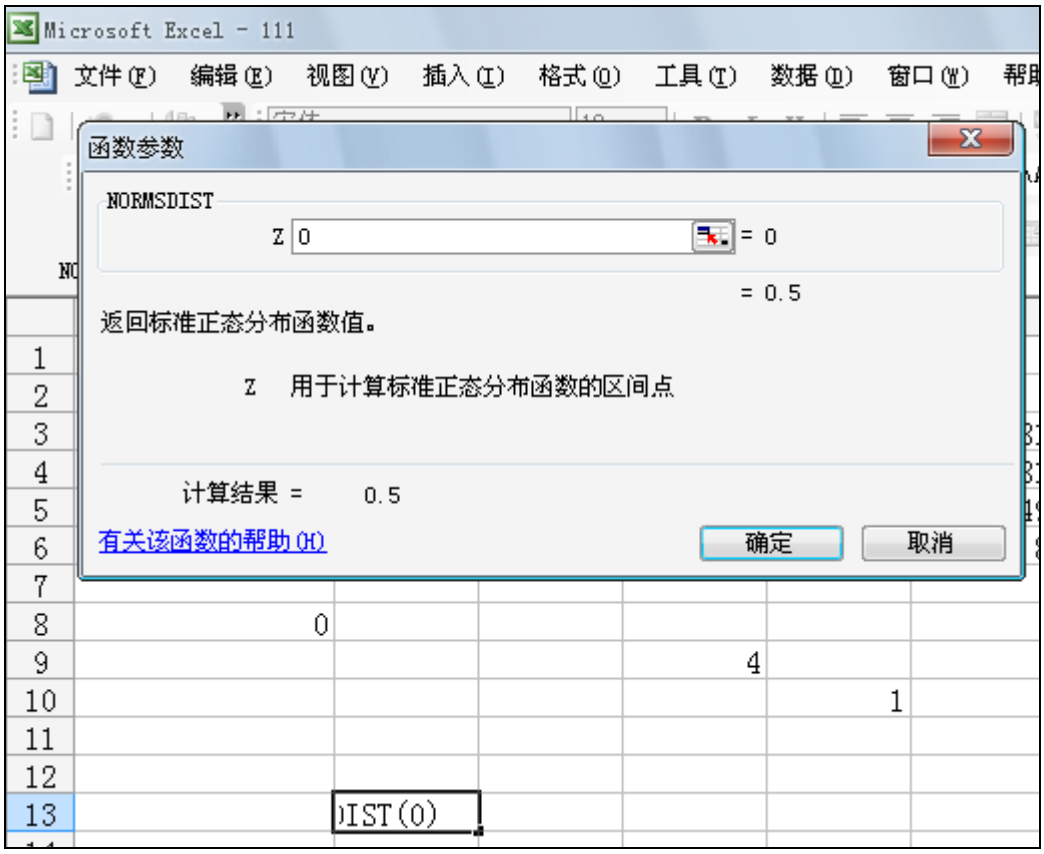
公式 NORMDIST(6,5,2,0)返回平均值为 5、标准差为 2 的正态函数当 X=6 时概率密度函数的数值，公式 NORMDIST（60，50，4，1）返回平均值为 50、标准差为 4 的正态分布函数当 X=60 时累积分布函数的数值。

9.3.2.11 函数 NORMSDIST：

1） 说明：函数 NORMSDIST 返回标准正态分布的累积函数。

2） 语法： NORMSDIST(Z),Z 为需要计算其分布的数值。

3）举例：



NORMSDIST(0)=0.5

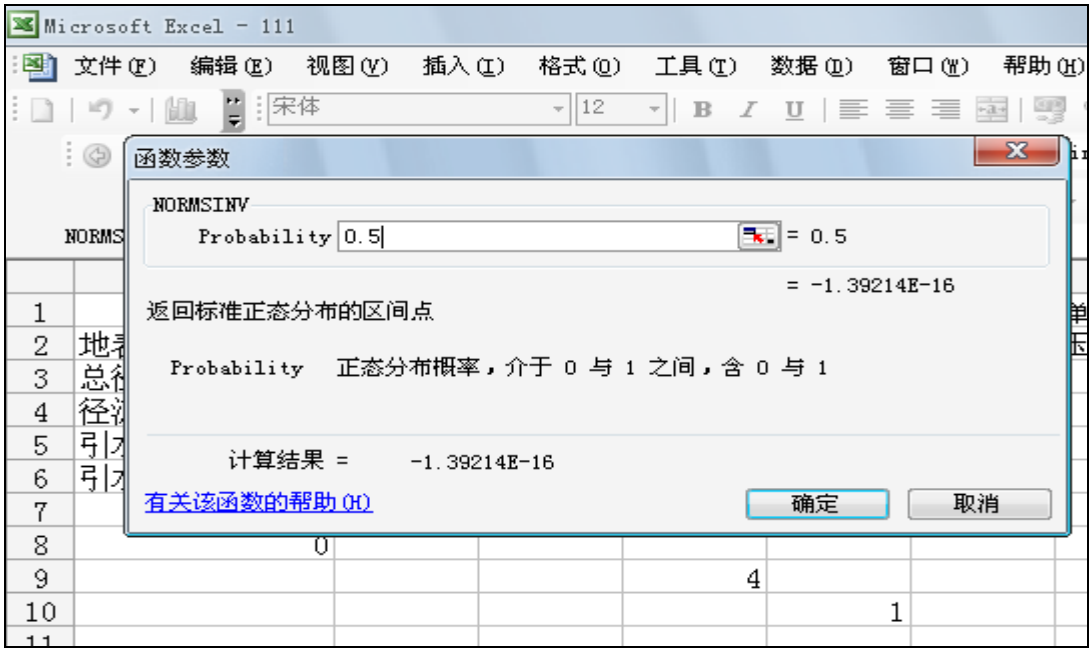
9.3.2.12 函数 NORMSINV:

1） 说明：函数 NORMSINV 返回标准正态分布累积函数的逆函数。

2） 语法：NORMSINV(probability)

Probability： 正态分布的概率值。

3)举例：



NORMSINV(0.5)=-1.39214E-16

9.3.2.13 分布函数 TDIST

1） 说明：函数 TDIST 返回 student 的 t 分布数值。T 分布用于小样本数据集合的假设检验。使用此函数可以代替 t 分布的临界值表。

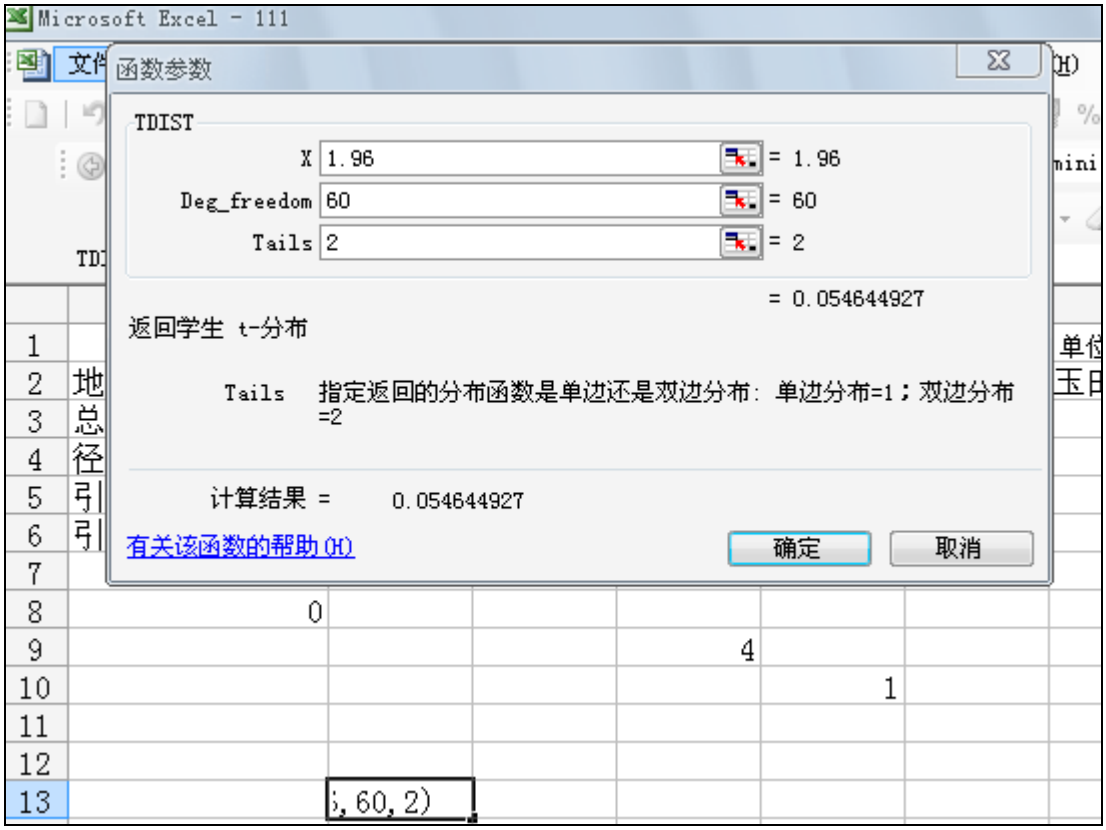
2） 语法：TDIST(x,degrees_freedom,tails)

X：为需要计算分布的数字。

Degrees_freedom：为表示自由度的整数。

Tails：指明返回的分布函数是单尾分布还是双尾分布。如果 tails = 1，函数 TDIST 返回单尾分布。如果 tails = 2，函数 TDIST 返回双尾分布。

3） 举例：



TDIST(1.96,60,2)=0.054645