数据统计分析软件SPSS入门

主要内容

- 1-SPSS概述
- 2-SPSS数据管理
- 3-SPSS的统计分析功能

主要内容

- 1-SPSS概述
- 2-SPSS数据管理
- 3-SPSS的统计分析功能

1-SPSS概述

- 1.1-SPSS简介
- 1.2-窗口介绍
- 1.3-数据库的构建

1-SPSS概述

□1.1-SPSS简介

□1.2-窗口介绍

□1.3-数据库的构建

1.1-SPSS简介

- SPSS是世界最早的统计分析软件，广泛应用于通信、医疗、银行、证券、保险、制造、商业、市场研究、科研、教育等许多领域和行业。
- SPSS的基本功能包括数据管理、统计分析、图表分析、输出管理等，具体的内容包括描述统计、总体的均值比较、相关分析、回归模型分析、聚类分析、时间序列分析、非参数检验等多个大类。

1.1-SPSS简介

□SPSS的特点

- 包括了各种成熟的统计方法与模型，为统计分析用户提供了全方位的统计学算法
- 提供了各种数据准备与数据整理技术
- 自由灵活的表格功能
- 各种常用的统计学图形
- 界面友好，操作简单，容易上手

1.2-窗口介绍

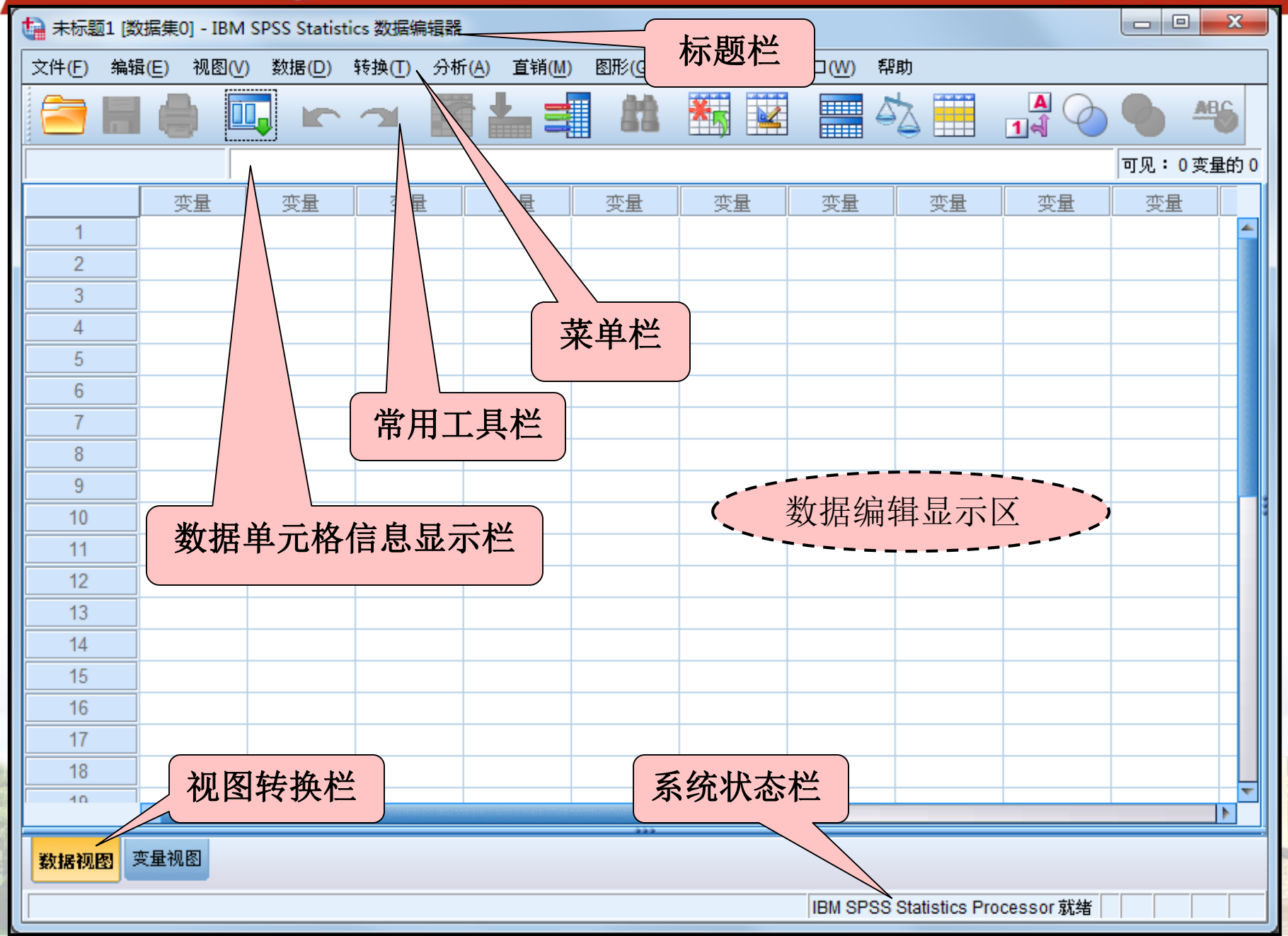
口启动

打开一个已存在的（历史记录）数据源

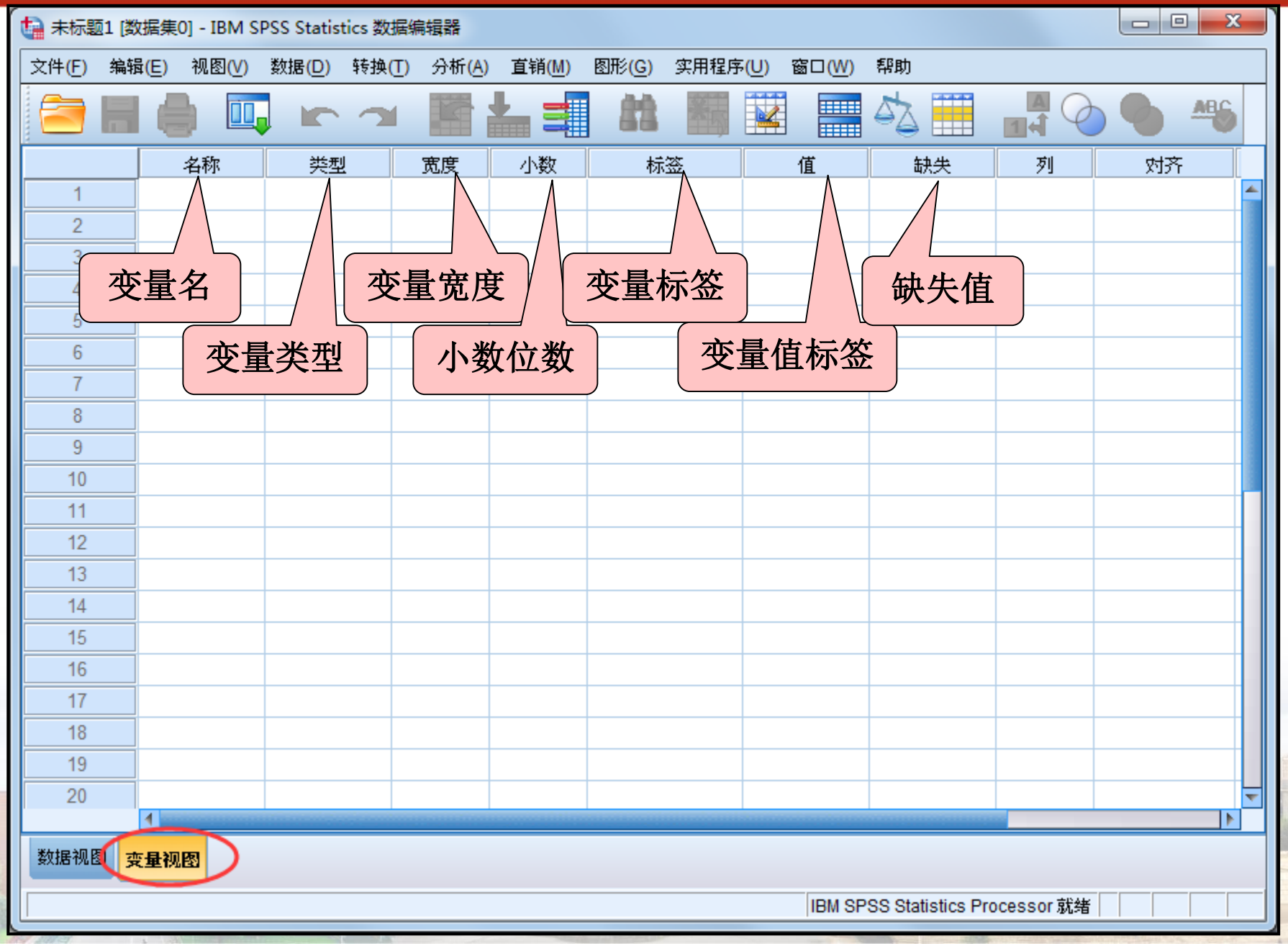
1.2-窗口介绍

- 数据窗口
- 变量窗口
- 结果输出窗口
- 图表编辑窗口
- 程序编辑窗口

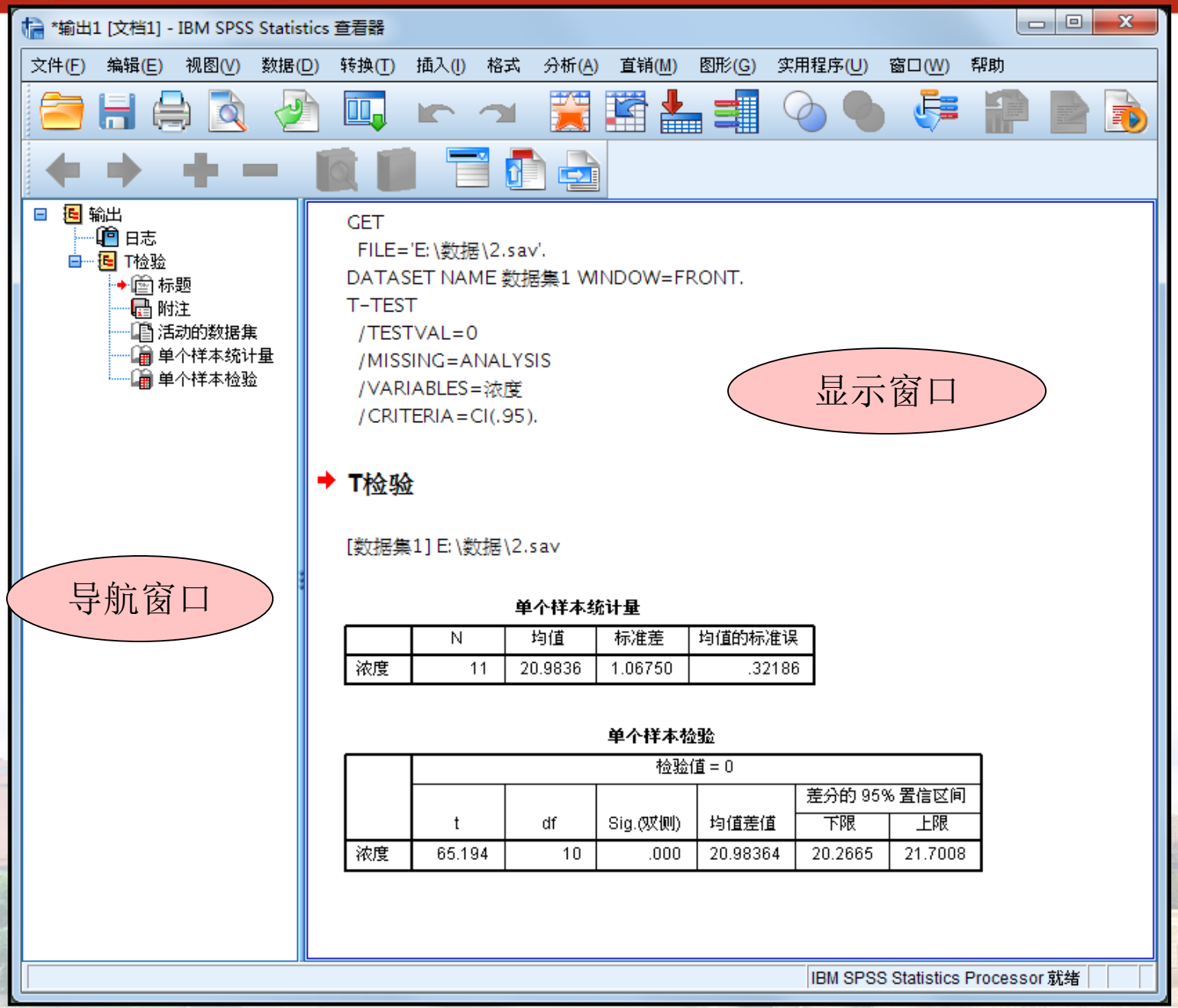
数据窗口


未标题1 [数据集0] - IBM SPSS Statistics 数据编辑器
标题栏
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 口(W) 帮助
菜单栏
常用工具栏
可见：0 变量的 0
变量
数据单元格信息显示栏
数据编辑显示区
视图转换栏
系统状态栏
数据视图
变量视图
IBM SPSS Statistics Processor 就绪

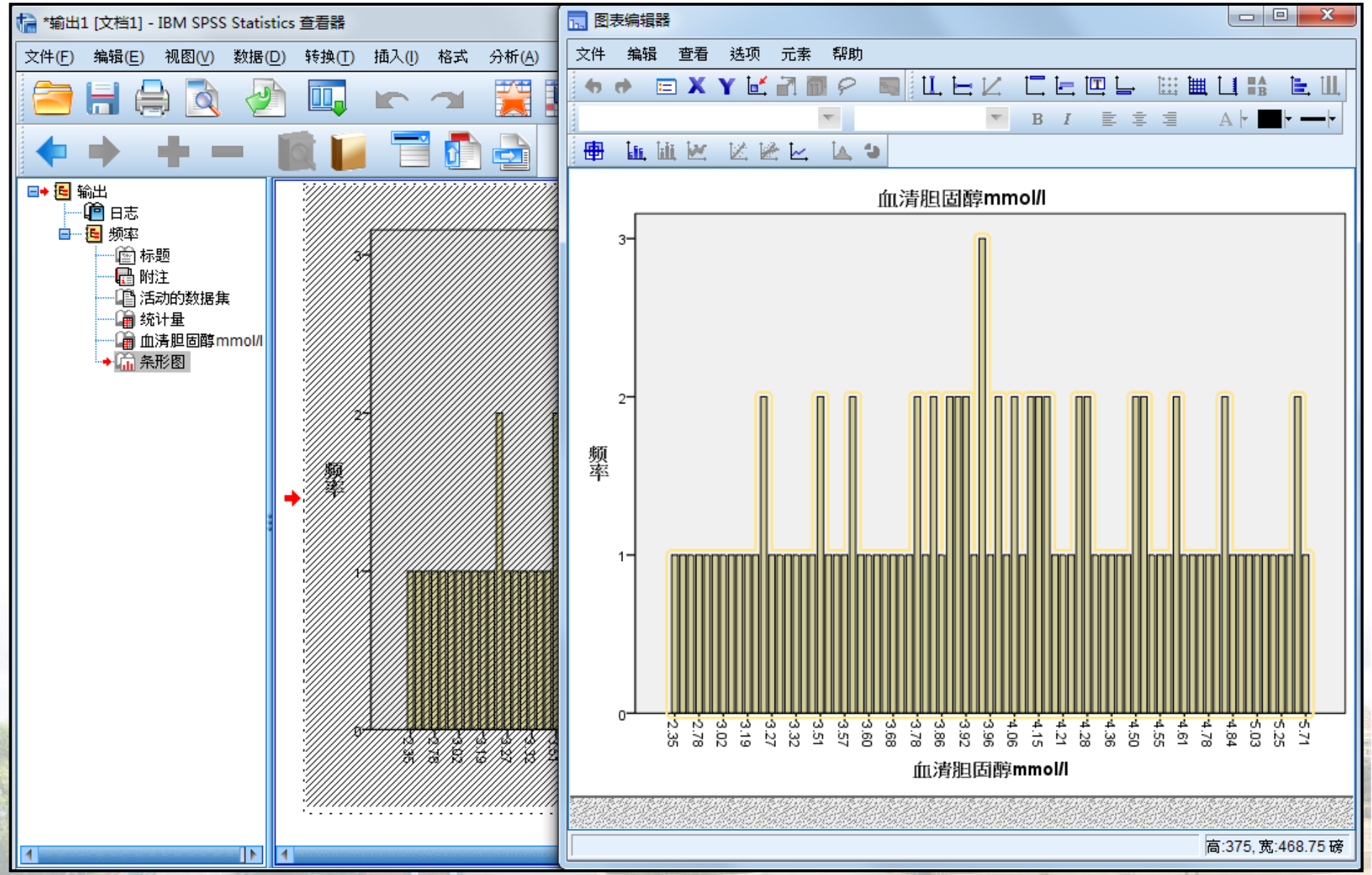
变量窗口


未标题1 [数据集0] - IBM SPSS Statistics 数据编辑器
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W) 帮助
名称
类型
宽度
小数
标签
值
缺失
列
对齐
变量名
变量类型
变量宽度
小数位数
变量标签
变量值标签
缺失值
数据视图
变量视图
IBM SPSS Statistics Processor 就绪

结果输出窗口

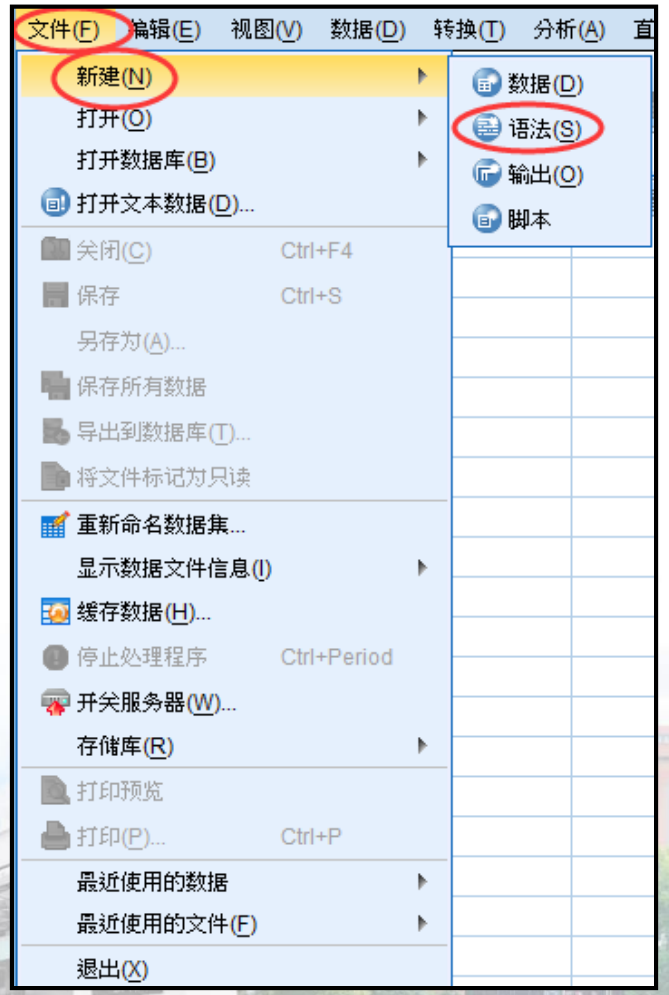

*输出1 [文档1] - IBM SPSS Statistics 查看器
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 插入(I) 格式 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W) 帮助
输出
日志
T检验
标题
附注
活动的数据集
单个样本统计量
单个样本检验
导航窗口
GET
FILE='E:\数据\2.sav'.
DATASET NAME 数据集1 WINDOW=FRONT.
T-TEST
/TESTVAL=0
/MISSING=ANALYSIS
/VARIABLES=浓度
/CRITERIA=CI(.95).
显示窗口
T检验
[数据集1] E:\数据\2.sav
单个样本统计量
N 均值 标准差 均值的标准误
浓度 11 20.9836 1.06750 .32186
单个样本检验
检验值 = 0
t df Sig.(双侧) 均值差值 差分的 95% 置信区间 下限 上限
浓度 65.194 10 .000 20.98364 20.2665 21.7008
IBM SPSS Statistics Processor 就绪

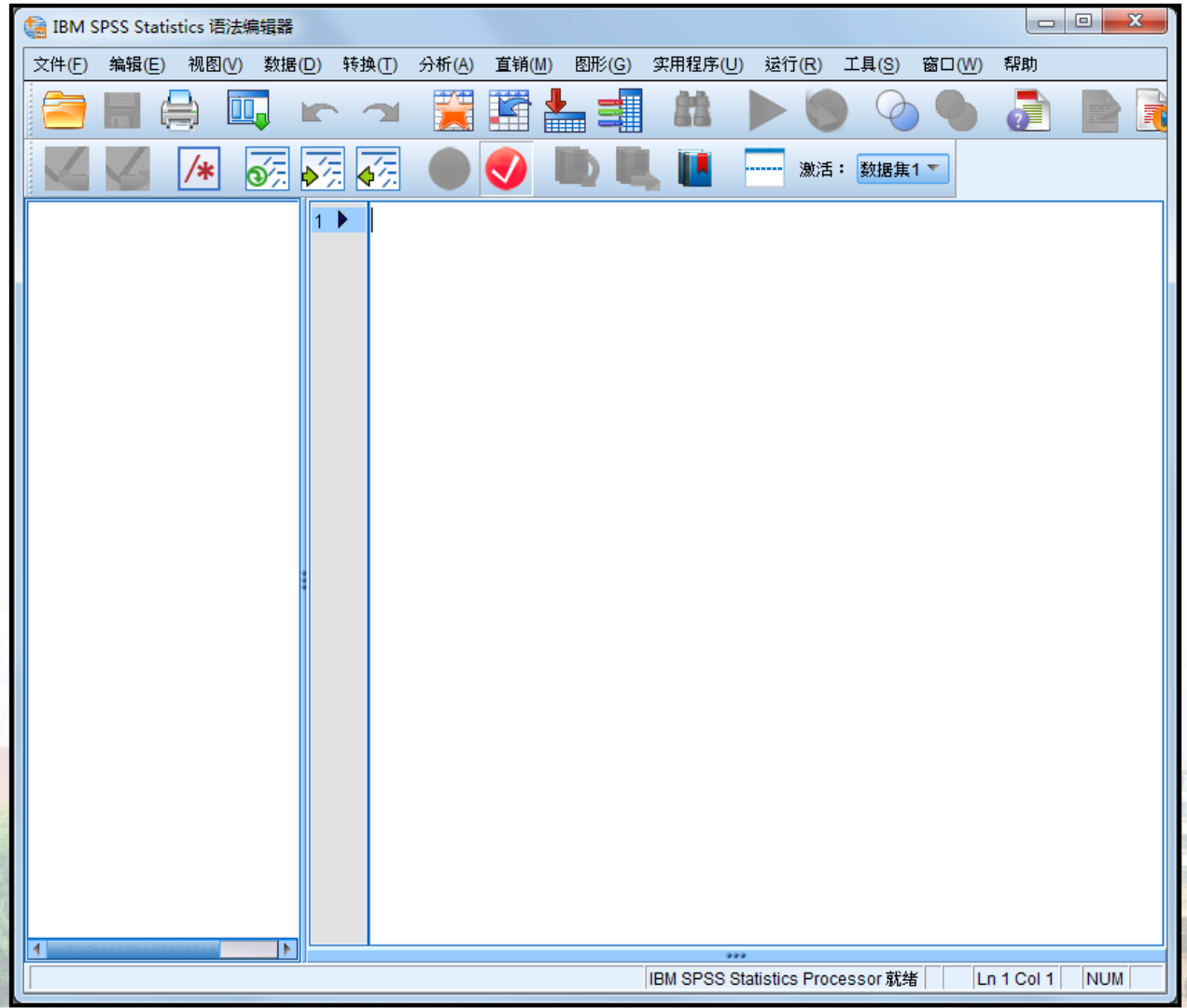
图表编辑窗口


*输出1 [文档1] - IBM SPSS Statistics 查看器
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 插入(I) 格式 分析(A)
输出
日志
频率
标题
附注
活动的数据集
统计量
血清胆固醇mmol/l
条形图
频率
图表编辑器
文件 编辑 查看 选项 元素 帮助
血清胆固醇mmol/l
频率
血清胆固醇mmol/l
高:375, 宽:468.75 磅

程序编辑窗口

口“文件”—“新建”—“语法”可打开程序编辑窗口





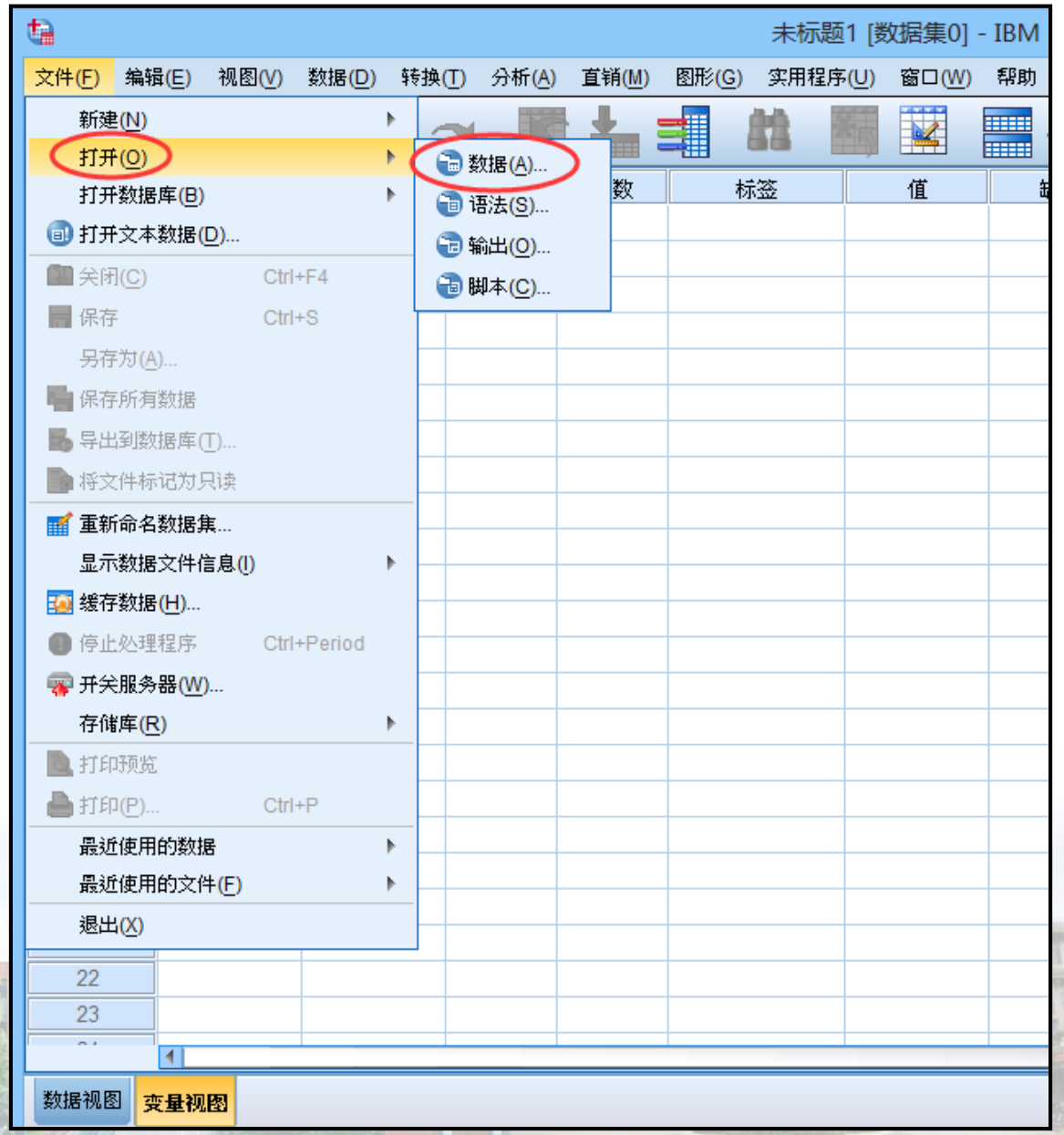
1.3-数据库的构建

- 现有文件导入
- 直接录入数据

现有文件导入

口直接导入SPSS或其它形式的数据库

- 例1：导入例1.sav、例2.txt


未标题1 [数据集0] - IBM
文件(F)
编辑(E)
视图(V)
数据(D)
转换(T)
分析(A)
直销(M)
图形(G)
实用程序(U)
窗口(W)
帮助
新建(N)
打开(O)
打开数据库(B)...
打开文本数据(D)...
关闭(C) Ctrl+F4
保存 Ctrl+S
另存为(A)...
保存所有数据
导出到数据库(T)...
将文件标记为只读
重新命名数据集...
显示数据文件信息(I)
缓存数据(H)...
停止处理程序 Ctrl+Period
开关服务器(W)...
存储库(R)
打印预览
打印(P)... Ctrl+P
最近使用的数据
最近使用的文件(F)
退出(X)
数据(A)...
语法(S)...
输出(O)...
脚本(C)...
标签
值
22
23
数据视图
变量视图

主要内容

- 1-SPSS概述
- 2-SPSS数据管理
- 3-SPSS的统计分析功能

2-SPSS数据管理

□2.1-数据的整理

2.1-数据的整理

- 数据合并
- 数据拆分
- 数据排序

数据合并

口将若干小的数据文件合并成一个大的数据文件

口纵向合并

- 几个数据集中的数据纵向堆叠，组成一个新的数据集，新的数据集中的记录数是原来的几个数据集中记录数的总和
- 合并条件：（1）待合并的SPSS数据文件，其内容合并是有实际意义的；（2）不同数据文件中，数据含义相同的列，最好起相同的名字，变量类型和变量长度也尽量相同

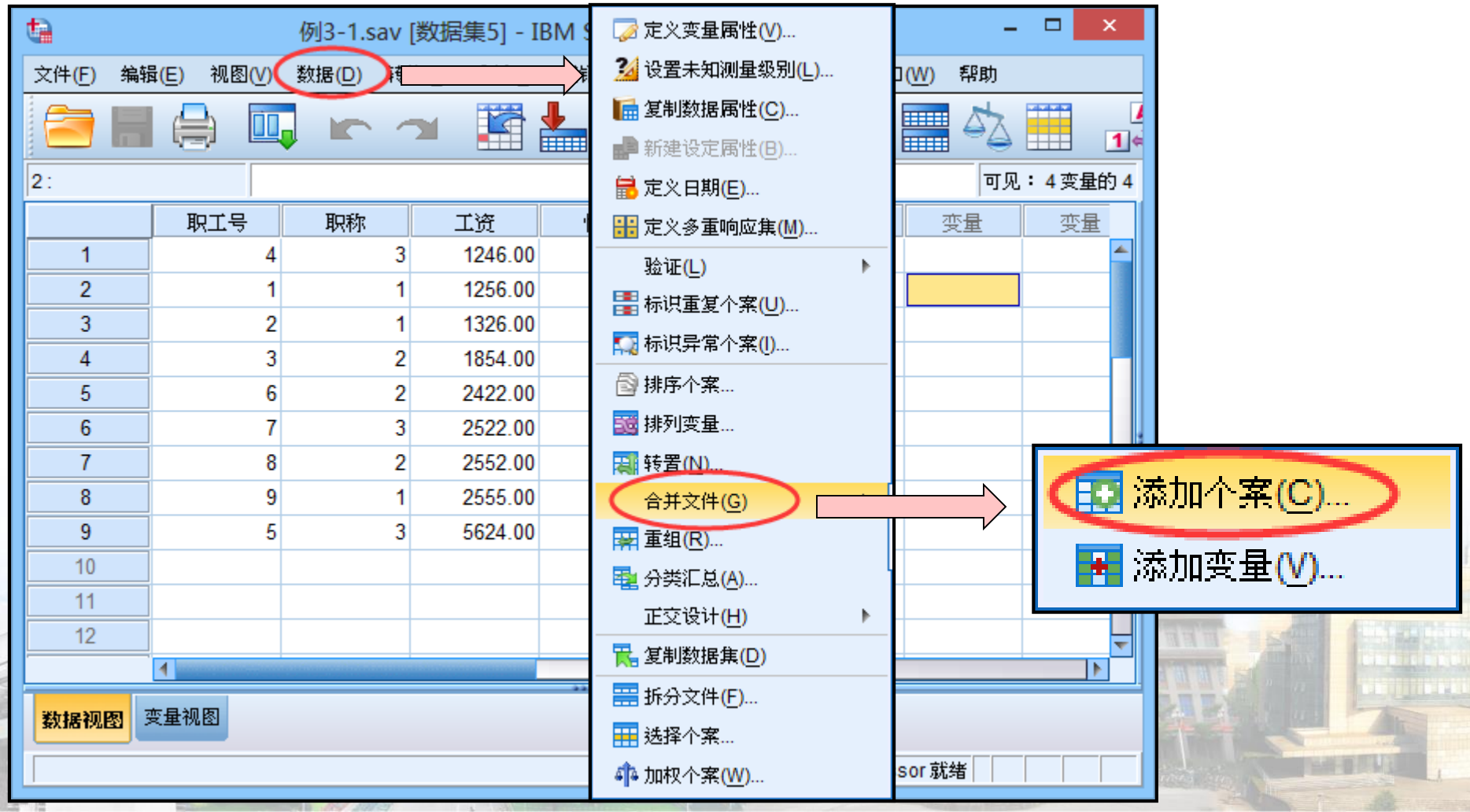
数据合并

口横向合并

- 按照记录的次序，或者某个关键变量的数值，将不同数据集中不同变量合并为一个数据集，新数据集的变量数是所有原数据集中不重名变量的总和，实质就是将两个数据文件的记录，按照记录对应，一一进行左右对接，合并的两个数据文件的变量不同，但具有相同个案例数。
- 合并条件：（1）如果不是按照记录号对应的规则进行合并，则两个数据文件必须至少有一个变量名相同的公共变量，这个变量时两个数据文件横向合并的依据，称为关键变量。（2）如果是使用关键变量进行合并的对应，则两个数据文件都必须事先按关键变量进行升序排列。（3）为方便SPSS数据文件的合并，在不同数据文件中，数据含义不相同的列，变量名不应取相同的名称。

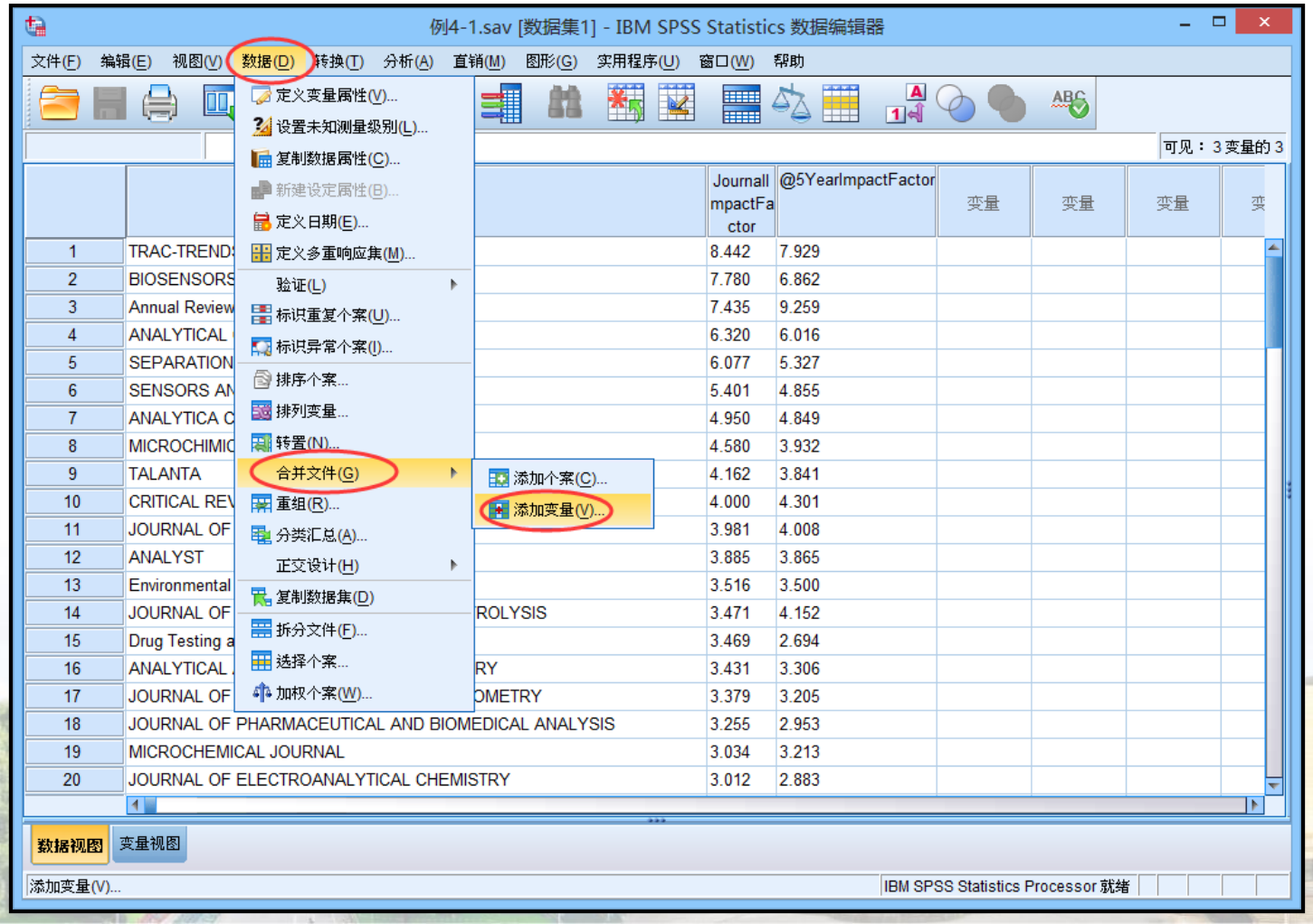
数据合并

口例3：将例3-1.sav与例3-2.sav数据进行合并（纵向合并）



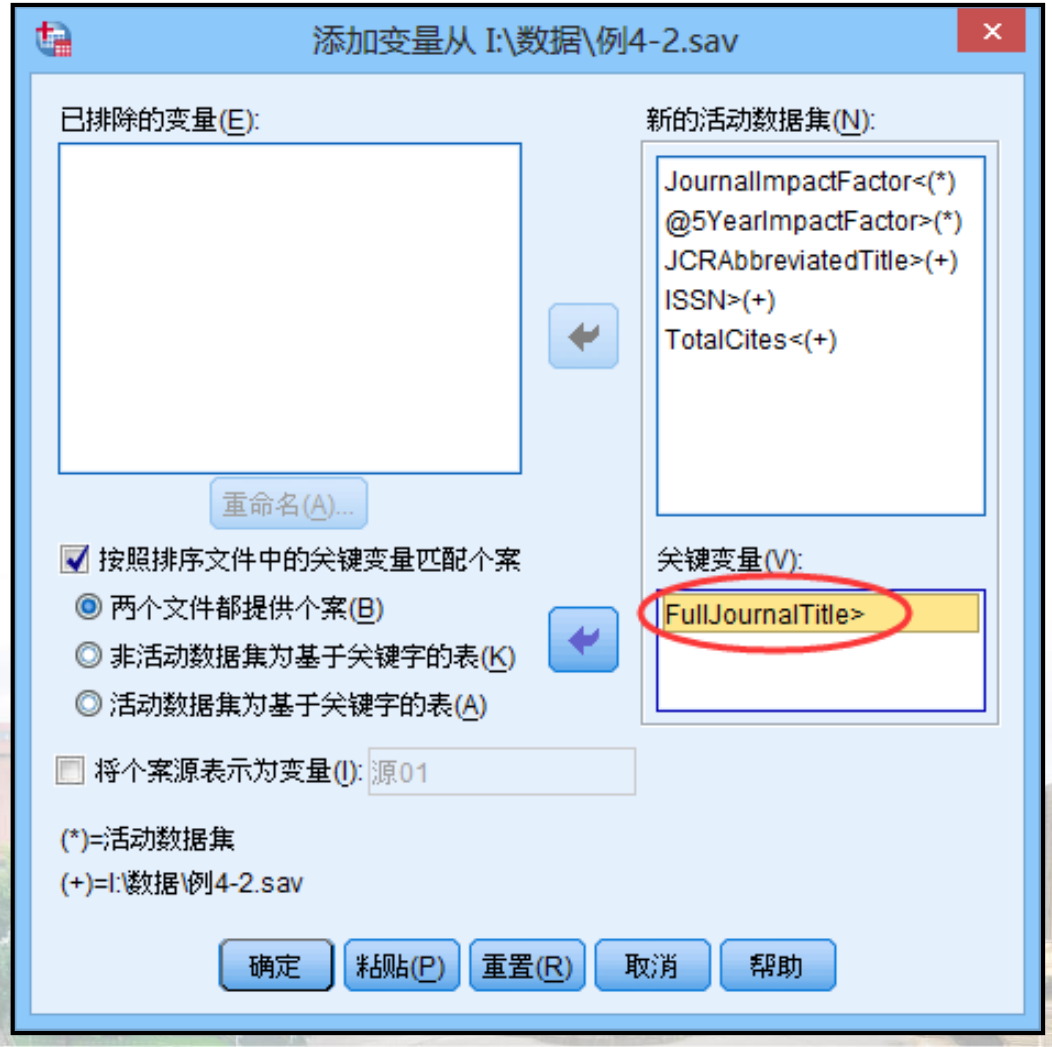
数据合并

口例4：将例4-1.sav与例4-2.sav数据进行合并（横向合并）



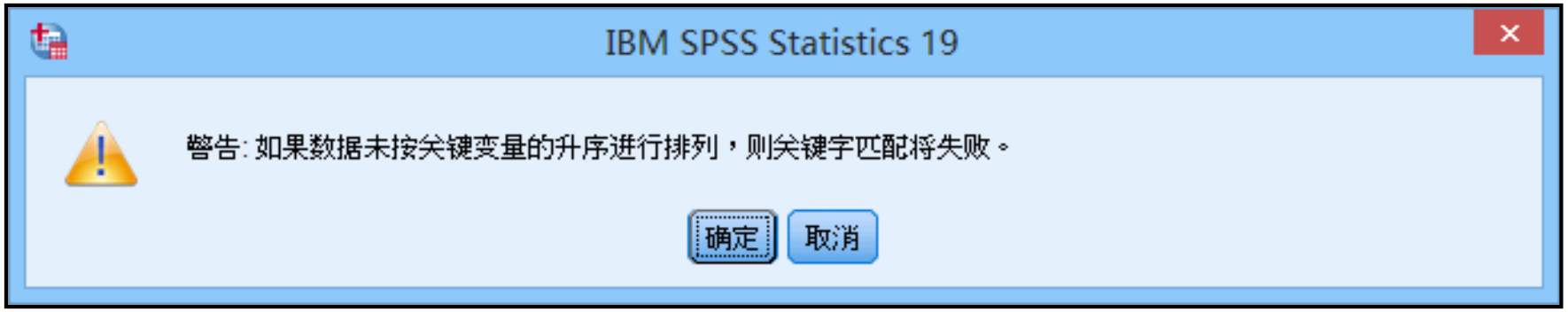
数据合并

口两个文件的数据变量中“期刊名称”是统一的，因此可以将“期刊名称” 作为关键变量



数据合并

口合并前还需要将关键变量进行升序排序



口来看看在Excel中如何处理这类问题

- VLOOKUP(lookup_value, table_array, col_index_num, range_lookup)

lookup_value	要查找的值
table_array	要查找的区域
col_index_num	返回数据在查找区域的第几列数
range_lookup	精确匹配/近似匹配

剪贴板 | 蓝牙 | 字体 | 对齐方式 | 数字 | 样式 | 单元格 | 编辑

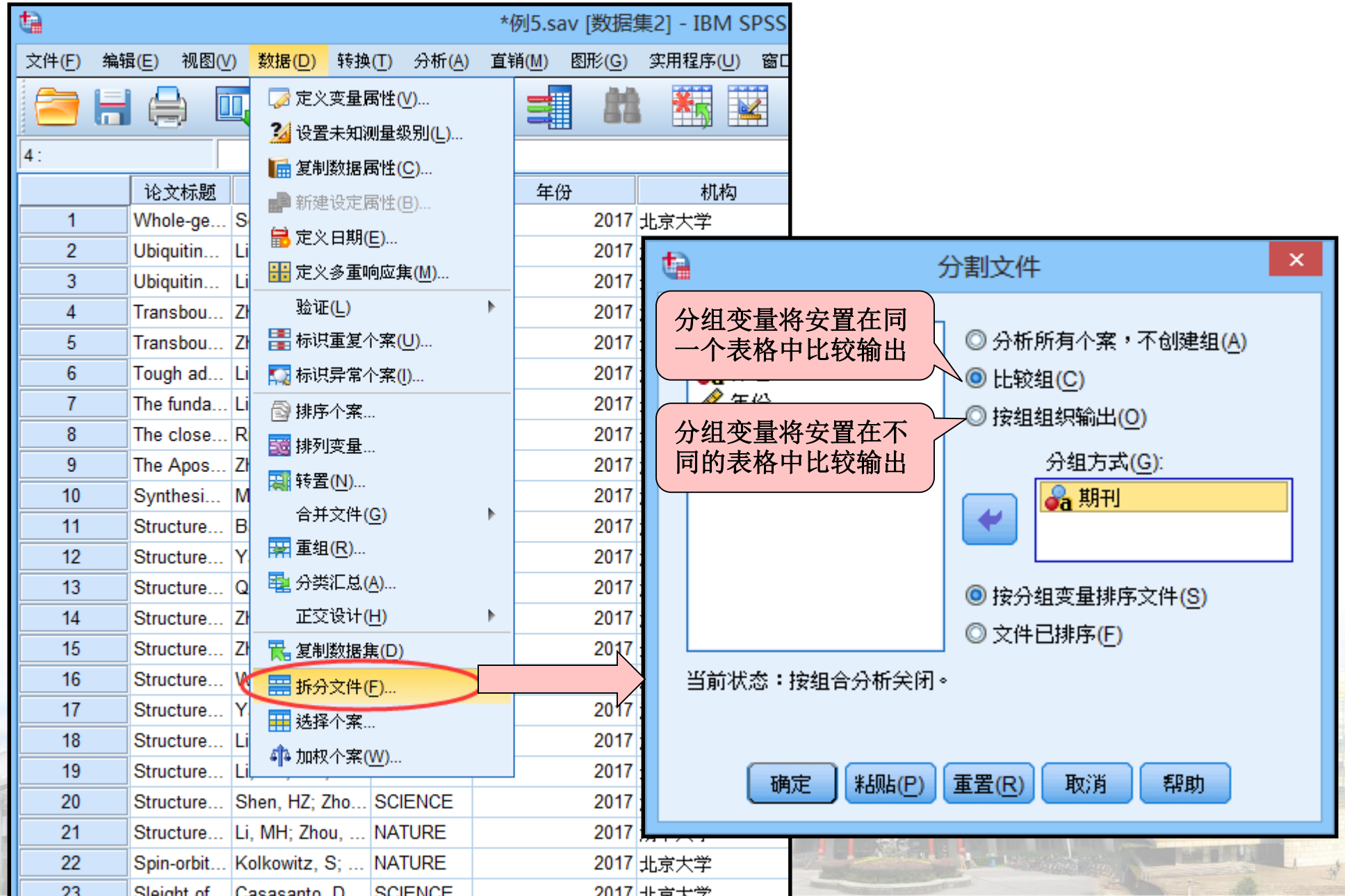
D2 fx =VLOOKUP(A2,'例4-2'!$A$1:$D$77,2,FALSE)

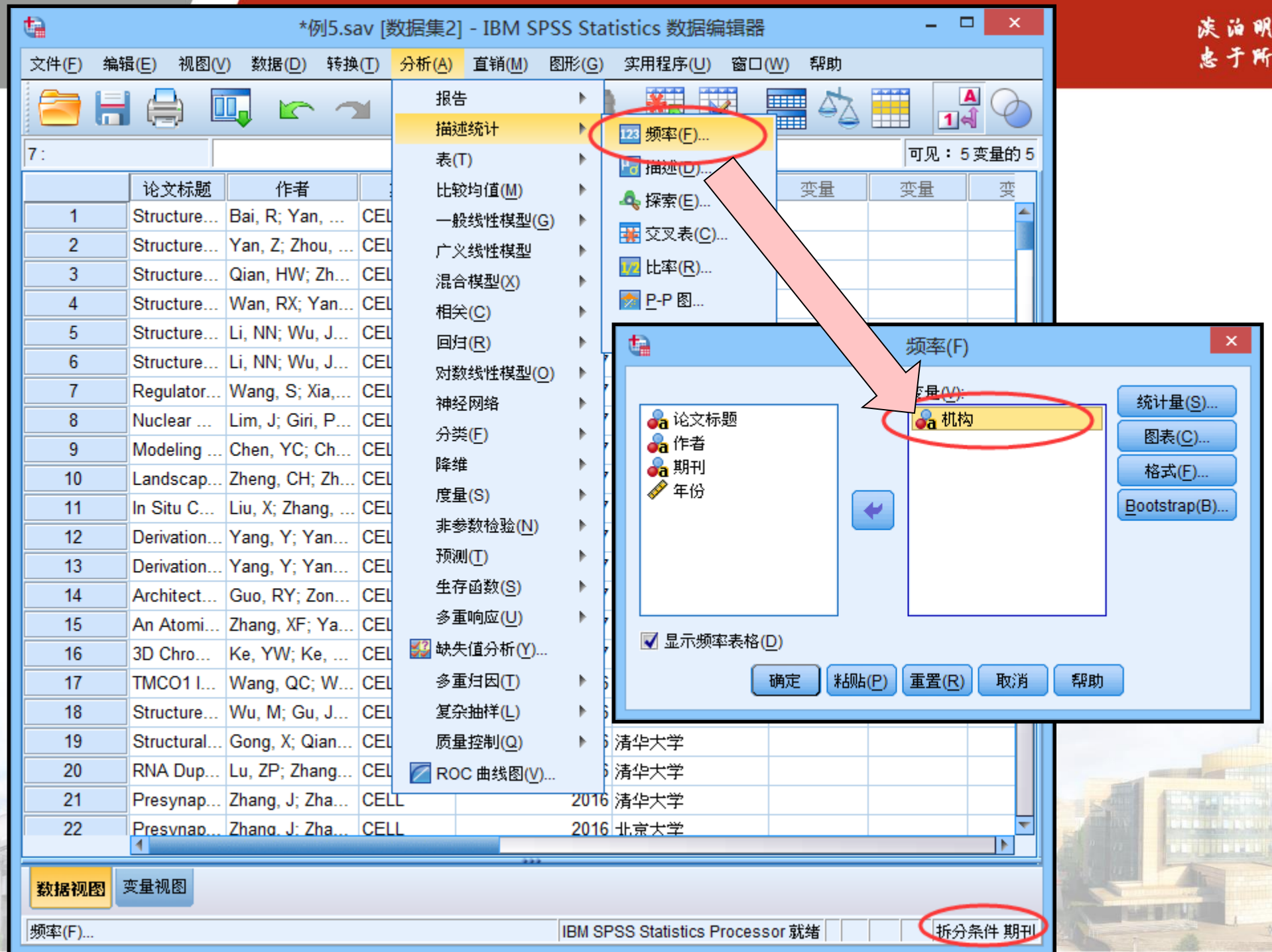
	A	B	C	D	E	F
1	期刊名称	期刊影响	期刊五年影响因子	期刊缩写	ISSN	总被引次数
2	TRAC-TRENDS IN ANALYTICAL CHE	8.44	7.93	TRAC-TREND ANAL CHEM	0165-9936	12038
3	BIOSENSORS & BIOELECTRONICS	7.78	6.86			
4	Annual Review of Analytical Chemist	7.44	9.26			
5	ANALYTICAL CHEMISTRY	6.32	6.02			
6	SEPARATION AND PURIFICATION R	6.08	5.33			
7	SENSORS AND ACTUATORS B-CHEI	5.4	4.86			

数据拆分

口其功能为按某一个变量值进行分组，数据仍在同一个文件中。但是，以后进行统计分析时，将根据拆分结果分别进行统计。

- 例5：对例5.sav文件数据：2010-2017清华大学、北京大学和四川大学在CNS上发表的论文信息（与实际数据略有删减）进行拆分
- 目的是分别查看三本期刊中三所高校在发表论文数量上的分布


*例5.sav [数据集2] - IBM SPSS
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U)
定义变量属性(V)...
设置未知测量级别(L)...
复制数据属性(C)...
新建设定属性(B)...
定义日期(E)...
定义多重响应集(M)...
验证(L)
标识重复个案(U)...
标识异常个案(I)...
排序个案...
排列变量...
转置(N)...
合并文件(G)
重组(R)...
分类汇总(A)...
正交设计(H)
复制数据集(D)
拆分文件(F)...
选择个案...
加权个案(W)...
论文标题
年份
机构
分割文件
分组变量将安置在同一个表格中比较输出
分组变量将安置在不同的表格中比较输出
分析所有个案，不创建组(A)
比较组(C)
按组组织输出(O)
分组方式(G):
期刊
按分组变量排序文件(S)
文件已排序(F)
当前状态：按组合分析关闭。
确定
粘贴(P)
重置(R)
取消
帮助


*例5.sav [数据集2] - IBM SPSS Statistics 数据编辑器
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W) 帮助
报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
相关(C)
回归(R)
对数线性模型(O)
神经网络
分类(F)
降维
度量(S)
非参数检验(N)
预测(T)
生存函数(S)
多重响应(U)
缺失值分析(Y)...
多重归因(T)
复杂抽样(L)
质量控制(Q)
ROC 曲线图(V)...
频率(F)...
描述(D)...
探索(E)...
交叉表(C)...
比率(R)...
P-P 图...
可见：5 变量的 5
论文标题 作者
1 Structure... Bai, R; Yan, ...
2 Structure... Yan, Z; Zhou, ...
3 Structure... Qian, HW; Zh...
4 Structure... Wan, RX; Yan...
5 Structure... Li, NN; Wu, J...
6 Structure... Li, NN; Wu, J...
7 Regulator... Wang, S; Xia,...
8 Nuclear ... Lim, J; Giri, P...
9 Modeling ... Chen, YC; Ch...
10 Landscap... Zheng, CH; Zh...
11 In Situ C... Liu, X; Zhang, ...
12 Derivation... Yang, Y; Yan...
13 Derivation... Yang, Y; Yan...
14 Architect... Guo, RY; Zon...
15 An Atomi... Zhang, XF; Ya...
16 3D Chro... Ke, YW; Ke, ...
17 TMCO1 I... Wang, QC; W...
18 Structure... Wu, M; Gu, J...
19 Structural... Gong, X; Qian...
20 RNA Dup... Lu, ZP; Zhang...
21 Presynap... Zhang, J; Zha... CELL 2016 清华大学
22 Presynap... Zhang, J; Zha... CELL 2016 北京大学
数据视图 变量视图
频率(F)...
IBM SPSS Statistics Processor 就绪
拆分条件 期刊
频率(F)
论文标题
作者
期刊
年份
机构
统计量(S)...
图表(C)...
格式(F)...
Bootstrap(B)...
显示频率表格(D)
确定 粘贴(P) 重置(R) 取消 帮助

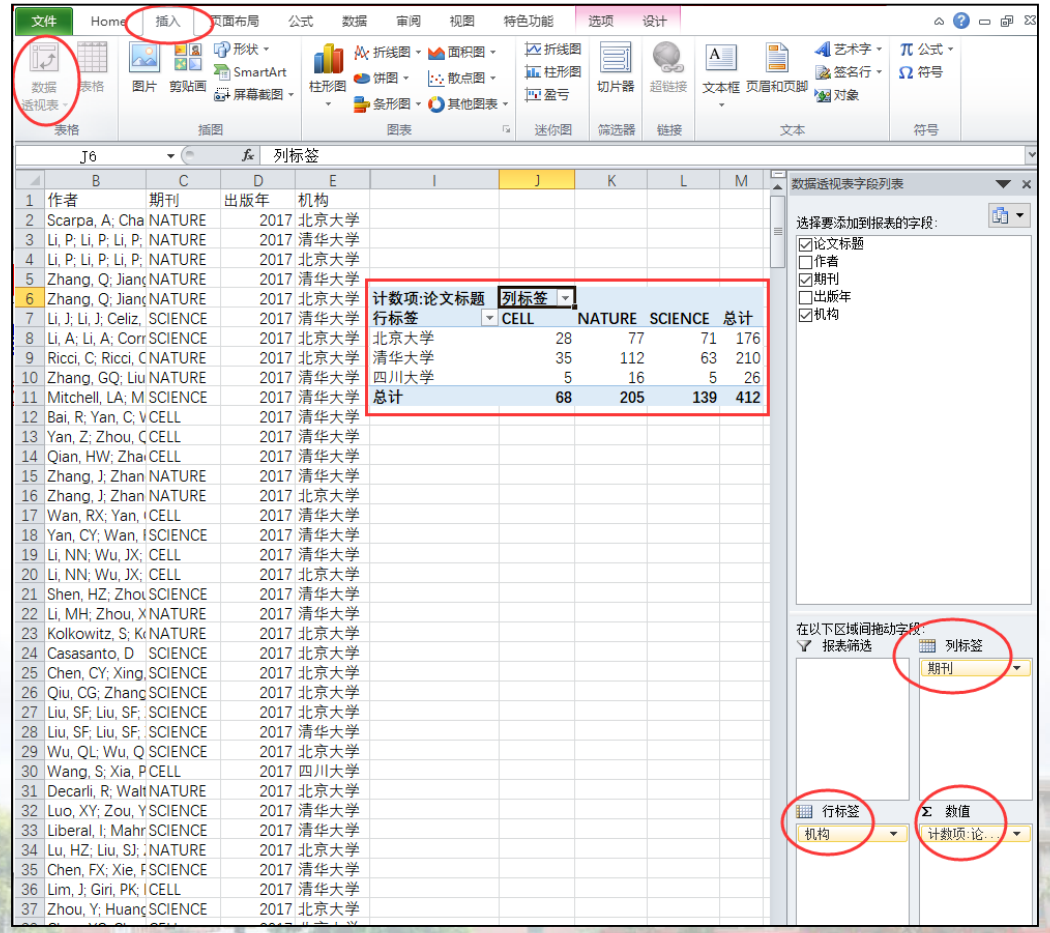
口输出结

机构

期刊			频率	百分比	有效百分比	累积百分比
CELL	有效	北京大学	28	41.2	41.2	41.2
		清华大学	35	51.5	51.5	92.6
		四川大学	5	7.4	7.4	100.0
		合计	68	100.0	100.0	
NATURE	有效	北京大学	77	37.6	37.6	37.6
		清华大学	112	54.6	54.6	92.2
		四川大学	16	7.8	7.8	100.0
		合计	205	100.0	100.0	
SCIENCE	有效	北京大学	71	51.1	51.1	51.1
		清华大学	63	45.3	45.3	96.4
		四川大学	5	3.6	3.6	100.0
		合计	139	100.0	100.0	

口来看看在Excel中如何处理这类问题

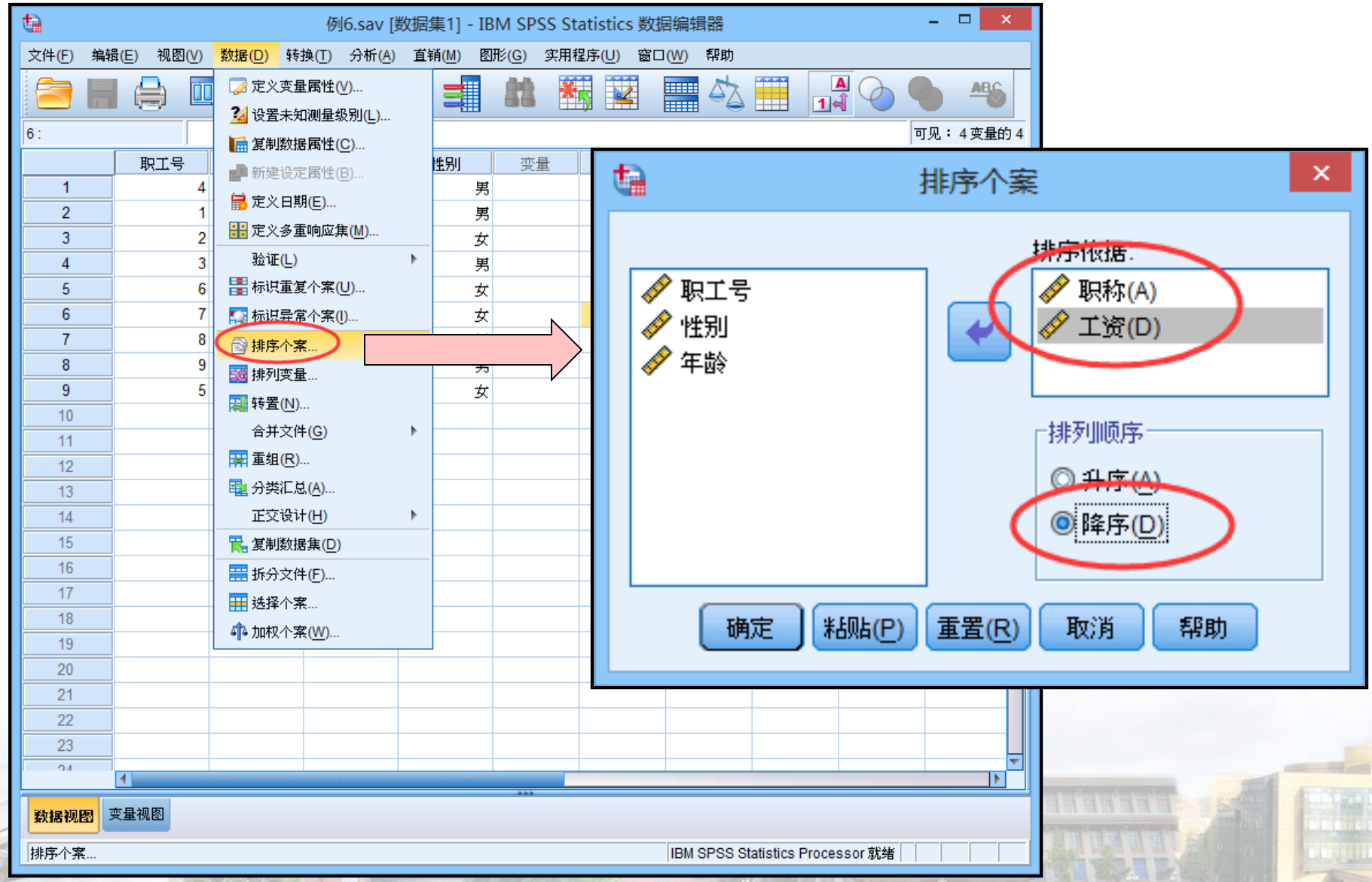
- 透视表



数据排序

口将数据按指定的某一个或多个变量值的升序或降序重新排列，所指定的变量称为排序变量。

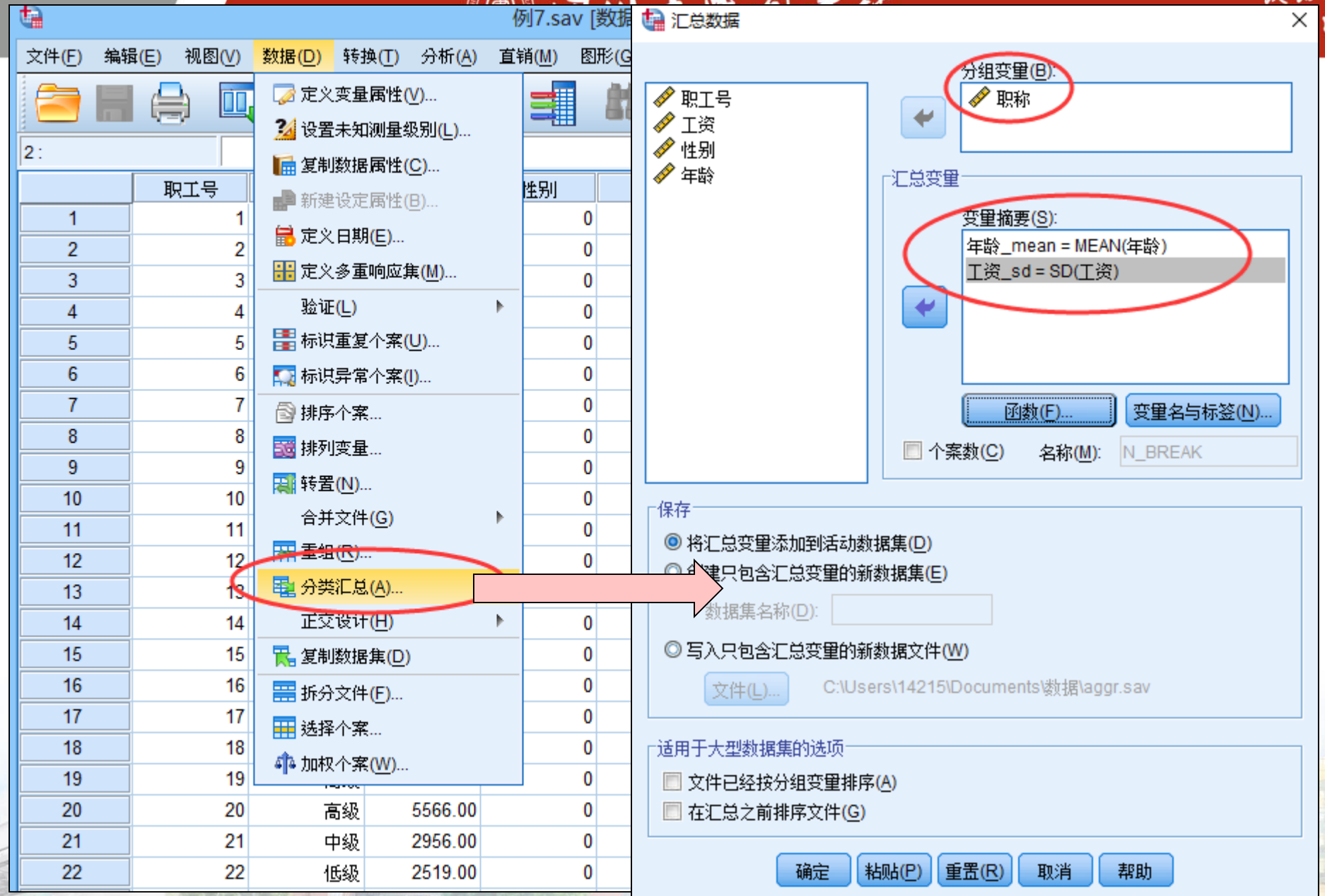
- 单值排序：排序变量只有一个。
- 多重排序：排序变量有多个，多重排序的第一个排序变量称为主排序变量，其它排序变量依次称为第二排序变量、第三排序变量等。
- 例6：打开例6. sav数据文件。要求职称按升序排序，工资按降序排序。


例6.sav [数据集1] - IBM SPSS Statistics 数据编辑器
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W) 帮助
定义变量属性(V)...
设置未知测量级别(L)...
复制数据属性(C)...
新建设定属性(B)...
定义日期(E)...
定义多重响应集(M)...
验证(L)
标识重复个案(U)...
标识异常个案(I)...
排序个案...
排列变量...
转置(N)...
合并文件(G)
重组(R)...
分类汇总(A)...
正交设计(H)
复制数据集(D)
拆分文件(F)...
选择个案...
加权个案(W)...
可见：4 变量的 4
职工号
性别
变量
数据视图
变量视图
排序个案...
IBM SPSS Statistics Processor 就绪
排序个案
职工号
性别
年龄
排序依据:
职称(A)
工资(D)
排列顺序
升序(A)
降序(D)
确定
粘贴(P)
重置(R)
取消
帮助

数据的分类汇总

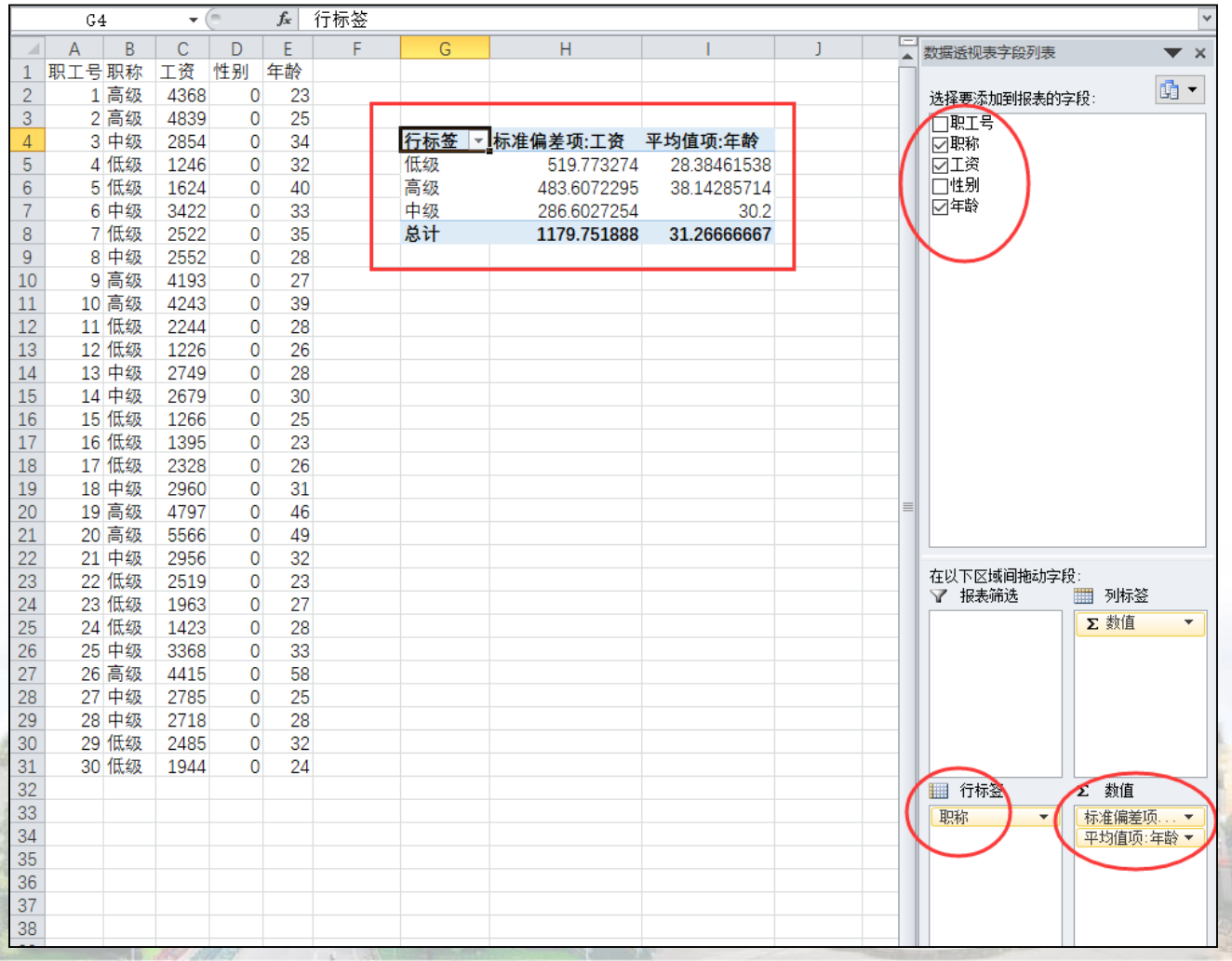
口按指定的分类变量对观测值进行分组，对每组记录的各变量求指定的描述统计量。

- 例7：将例7. sav中数据，以职称作为分组变量，对职工年龄的均值和工资的标准差进行汇总。


例7.sav [数据
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G
定义变量属性(V)...
设置未知测量级别(L)...
复制数据属性(C)...
新建设定属性(B)...
定义日期(E)...
定义多重响应集(M)...
验证(L)
标识重复个案(U)...
标识异常个案(I)...
排序个案...
排列变量...
转置(N)...
合并文件(G)
重组(R)...
分类汇总(A)...
正交设计(H)
复制数据集(D)
拆分文件(F)...
选择个案...
加权个案(W)...
2 :
职工号
性别
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22
0
20 高级 5566.00 0
21 中级 2956.00 0
22 低级 2519.00 0
汇总数据
职工号
工资
性别
年龄
分组变量(B):
职称
汇总变量
变量摘要(S):
年龄_mean = MEAN(年龄)
工资_sd = SD(工资)
函数(F)...
变量名与标签(N)...
个案数(C)
名称(M):
N_BREAK
保存
将汇总变量添加到活动数据集(D)
创建只包含汇总变量的新数据集(E)
数据集名称(D):
写入只包含汇总变量的新数据文件(W)
文件(L)...
C:\Users\14215\Documents\数据\aggr.sav
适用于大型数据集的选项
文件已经按分组变量排序(A)
在汇总之前排序文件(G)
确定
粘贴(P)
重置(R)
取消
帮助

口来看看在Excel中如何处理这类问题

- 透视表

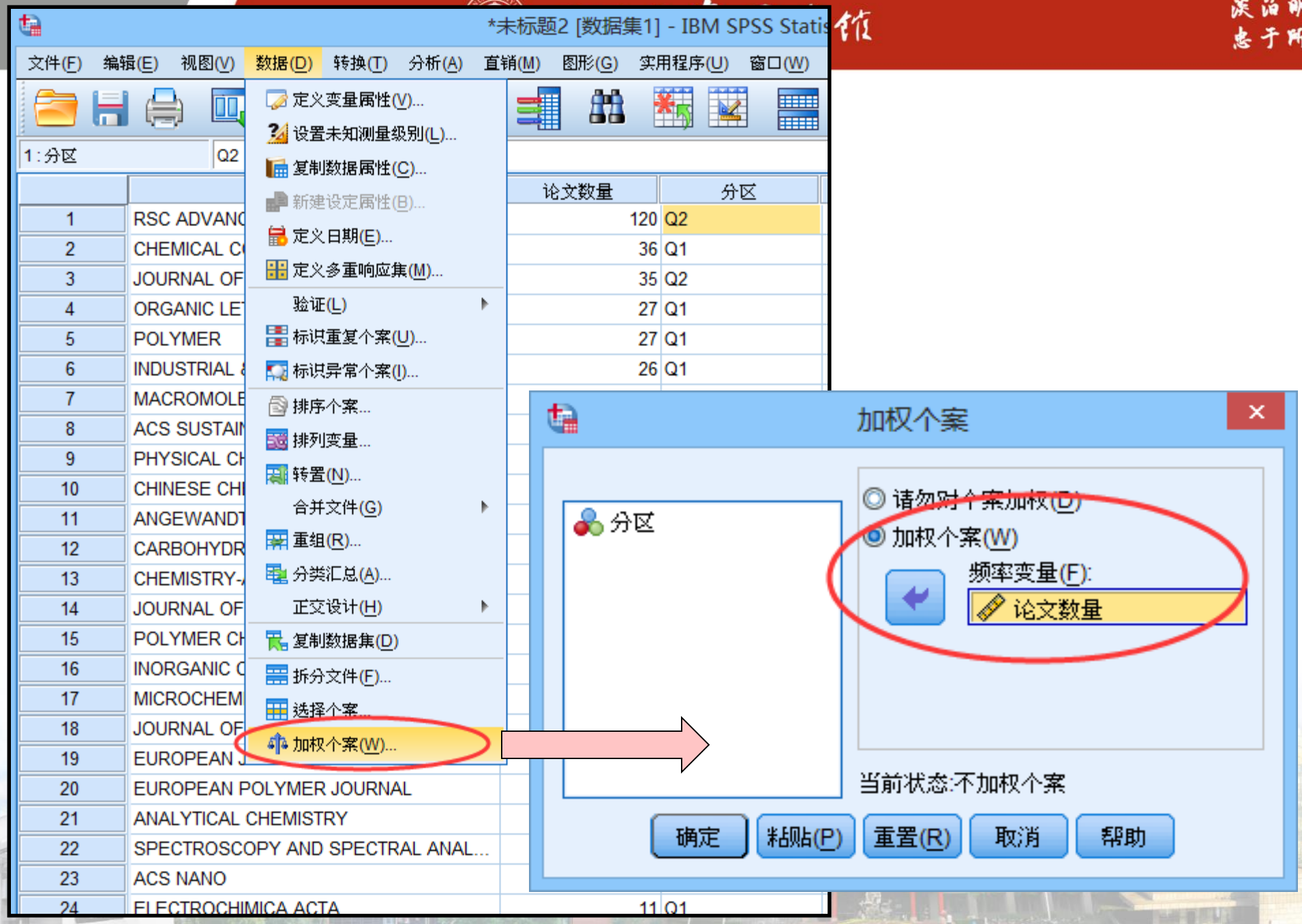


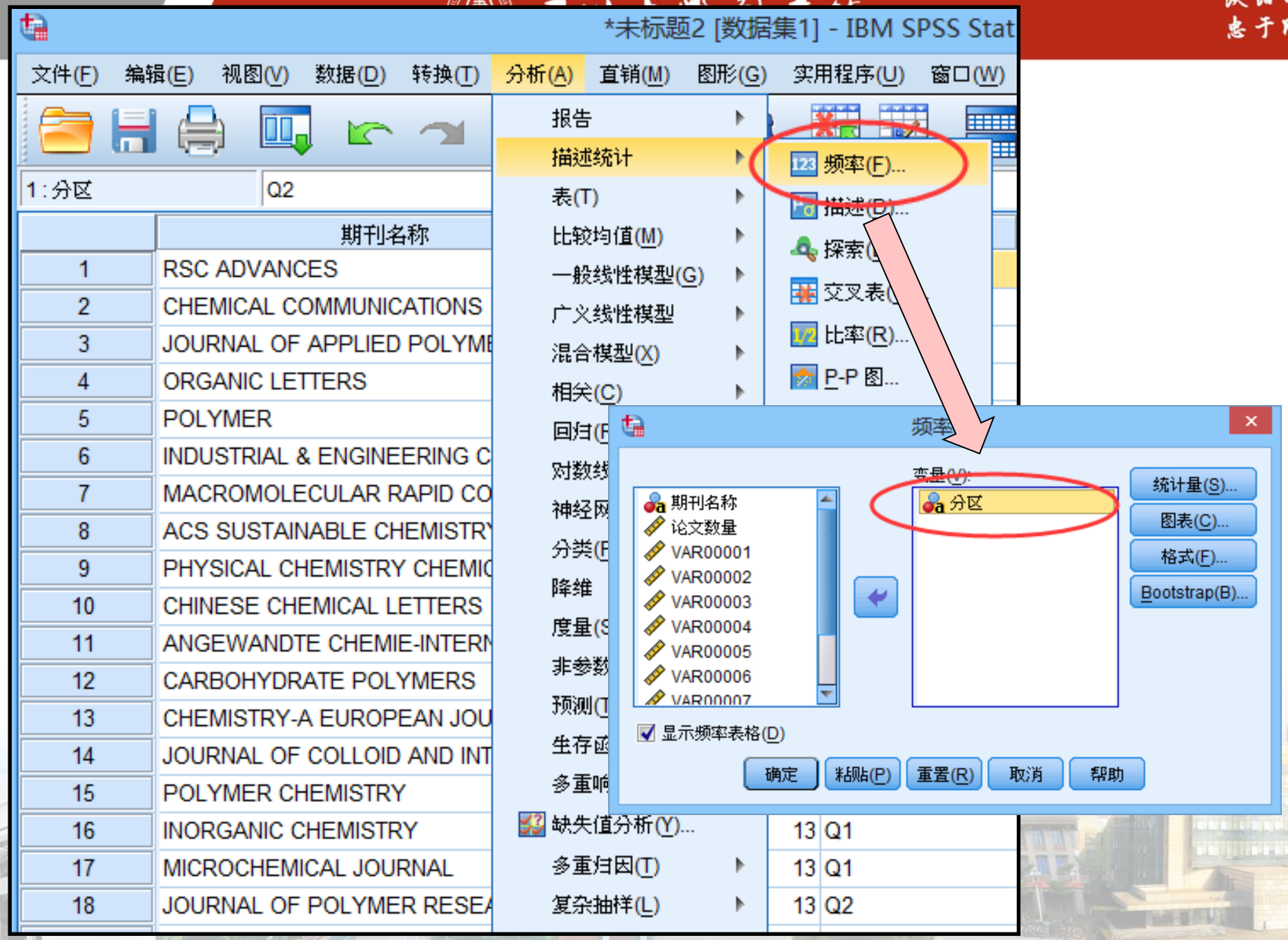
职工号	职称	工资	性别	年龄
1	高级	4368	0	23
2	高级	4839	0	25
3	中级	2854	0	34
4	低级	1246	0	32
5	低级	1624	0	40
6	中级	3422	0	33
7	低级	2522	0	35
8	中级	2552	0	28
9	高级	4193	0	27
10	高级	4243	0	39
11	低级	2244	0	28
12	低级	1226	0	26
13	中级	2749	0	28
14	中级	2679	0	30
15	低级	1266	0	25
16	低级	1395	0	23
17	低级	2328	0	26
18	中级	2960	0	31
19	高级	4797	0	46
20	高级	5566	0	49
21	中级	2956	0	32
22	低级	2519	0	23
23	低级	1963	0	27
24	低级	1423	0	28
25	中级	3368	0	33
26	高级	4415	0	58
27	中级	2785	0	25
28	中级	2718	0	28
29	低级	2485	0	32
30	低级	1944	0	24

行标签	标准偏差项:工资	平均值项:年龄
低级	519.773274	28.38461538
高级	483.6072295	38.14285714
中级	286.6027254	30.2
总计	**1179.751888**	**31.26666667**

数据的加权

- 加权是为了为告知统计软件你这一行数据代表的并不是单个值，而是表示实际样本很多个，有相应的“频数” 之和那么多的样本数。
  - 例8：对例8.sav数据进行加权，分析我校在2017年化学学科发表的期刊中，1-4区期刊的占比。


*未标题2 [数据集1] - IBM SPSS Statis
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W)
定义变量属性(V)...
设置未知测量级别(L)...
复制数据属性(C)...
新建设定属性(B)...
定义日期(E)...
定义多重响应集(M)...
验证(L)
标识重复个案(U)...
标识异常个案(I)...
排序个案...
排列变量...
转置(N)...
合并文件(G)
重组(R)...
分类汇总(A)...
正交设计(H)
复制数据集(D)
拆分文件(F)...
选择个案...
加权个案(W)...
1 : 分区
Q2
论文数量
分区
1 RSC ADVANC 120 Q2
2 CHEMICAL C 36 Q1
3 JOURNAL OF 35 Q2
4 ORGANIC LE 27 Q1
5 POLYMER 27 Q1
6 INDUSTRIAL & 26 Q1
7 MACROMOLE
8 ACS SUSTAI
9 PHYSICAL CH
10 CHINESE CHI
11 ANGEWANDT
12 CARBOHYDR
13 CHEMISTRY-A
14 JOURNAL OF
15 POLYMER CH
16 INORGANIC C
17 MICROCHEM
18 JOURNAL OF
19 EUROPEAN J
20 EUROPEAN POLYMER JOURNAL
21 ANALYTICAL CHEMISTRY
22 SPECTROSCOPY AND SPECTRAL ANAL...
23 ACS NANO
24 ELECTROCHIMICA ACTA 11 Q1
加权个案
分区
请勿对个案加权(D)
加权个案(W)
频率变量(F):
论文数量
当前状态:不加权个案
确定
粘贴(P)
重置(R)
取消
帮助


*未标题2 [数据集1] - IBM SPSS Stat
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W)
1 : 分区
Q2
期刊名称
1 RSC ADVANCES
2 CHEMICAL COMMUNICATIONS
3 JOURNAL OF APPLIED POLYME
4 ORGANIC LETTERS
5 POLYMER
6 INDUSTRIAL & ENGINEERING C
7 MACROMOLECULAR RAPID CO
8 ACS SUSTAINABLE CHEMISTRY
9 PHYSICAL CHEMISTRY CHEMIC
10 CHINESE CHEMICAL LETTERS
11 ANGEWANDTE CHEMIE-INTERN
12 CARBOHYDRATE POLYMERS
13 CHEMISTRY-A EUROPEAN JOU
14 JOURNAL OF COLLOID AND INT
15 POLYMER CHEMISTRY
16 INORGANIC CHEMISTRY
17 MICROCHEMICAL JOURNAL
18 JOURNAL OF POLYMER RESEA
报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
相关(C)
回归(R)
对数线
神经网
分类(F)
降维
度量(S)
非参数
预测(T)
生存函
多重响
缺失值分析(Y)...
多重归因(T)
复杂抽样(L)
频率(F)...
描述(D)...
探索(E)...
交叉表(C)...
比率(R)...
P-P 图...
频率
变量(V):
期刊名称
论文数量
VAR00001
VAR00002
VAR00003
VAR00004
VAR00005
VAR00006
VAR00007
分区
统计量(S)...
图表(C)...
格式(F)...
Bootstrap(B)...
显示频率表格(D)
确定
粘贴(P)
重置(R)
取消
帮助
13 Q1
13 Q1
13 Q2

口结果

分区

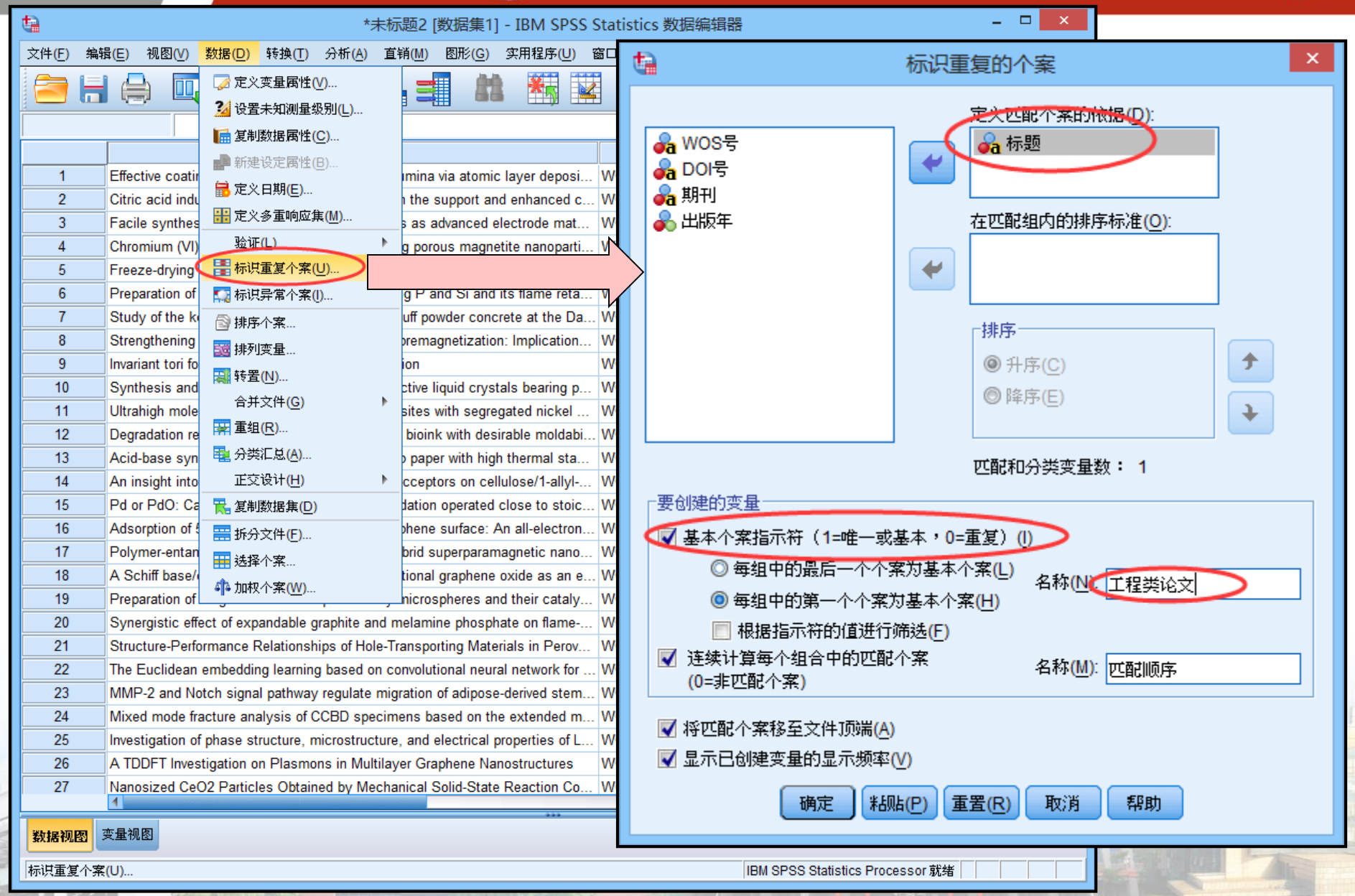
		频率	百分比	有效百分比	累积百分比
有效	#N/A	7	.6	.6	.6
	Q1	607	49.3	49.3	49.9
	Q2	398	32.3	32.3	82.2
	Q3	135	11.0	11.0	93.2
	Q4	84	6.8	6.8	100.0
	合计	1231	100.0	100.0	

重复数据的查找

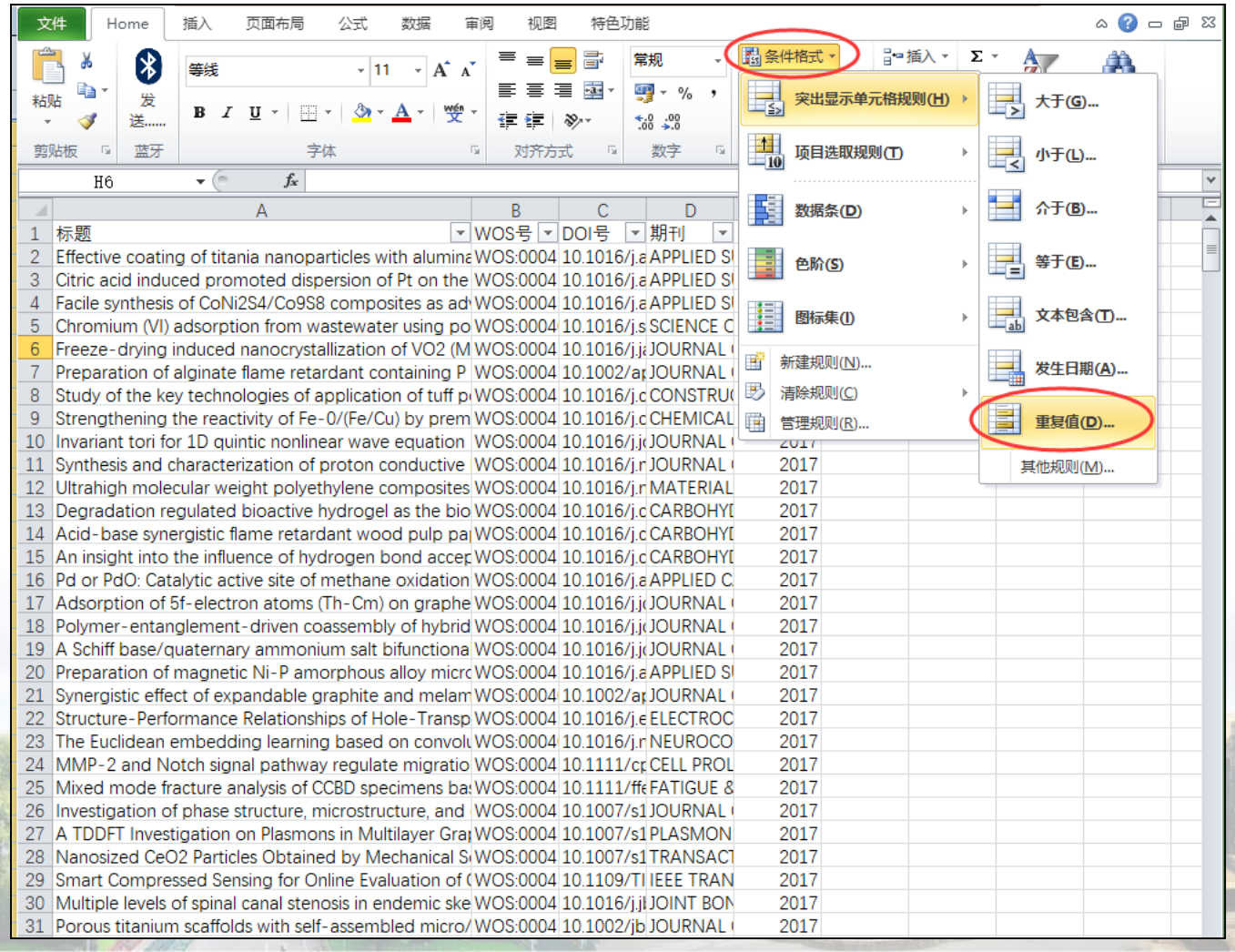
口定位重复的个体。适用于数据双录入后的数据检索。

- 例9：查找例9.sav中的重复数据。
- 数据中第1-500条个案是从WOS中下载的500篇论文，包括标题、WOS号、DOI号等信息。第501-805条个案是所有工程类论文的列表。
- 目的是找出第1-500条个案中属于工程类的论文。

四川大学图书馆
淡泊明志 宁静致远
志于所学 继续求学



口来看看在Excel中如何处理这类问题



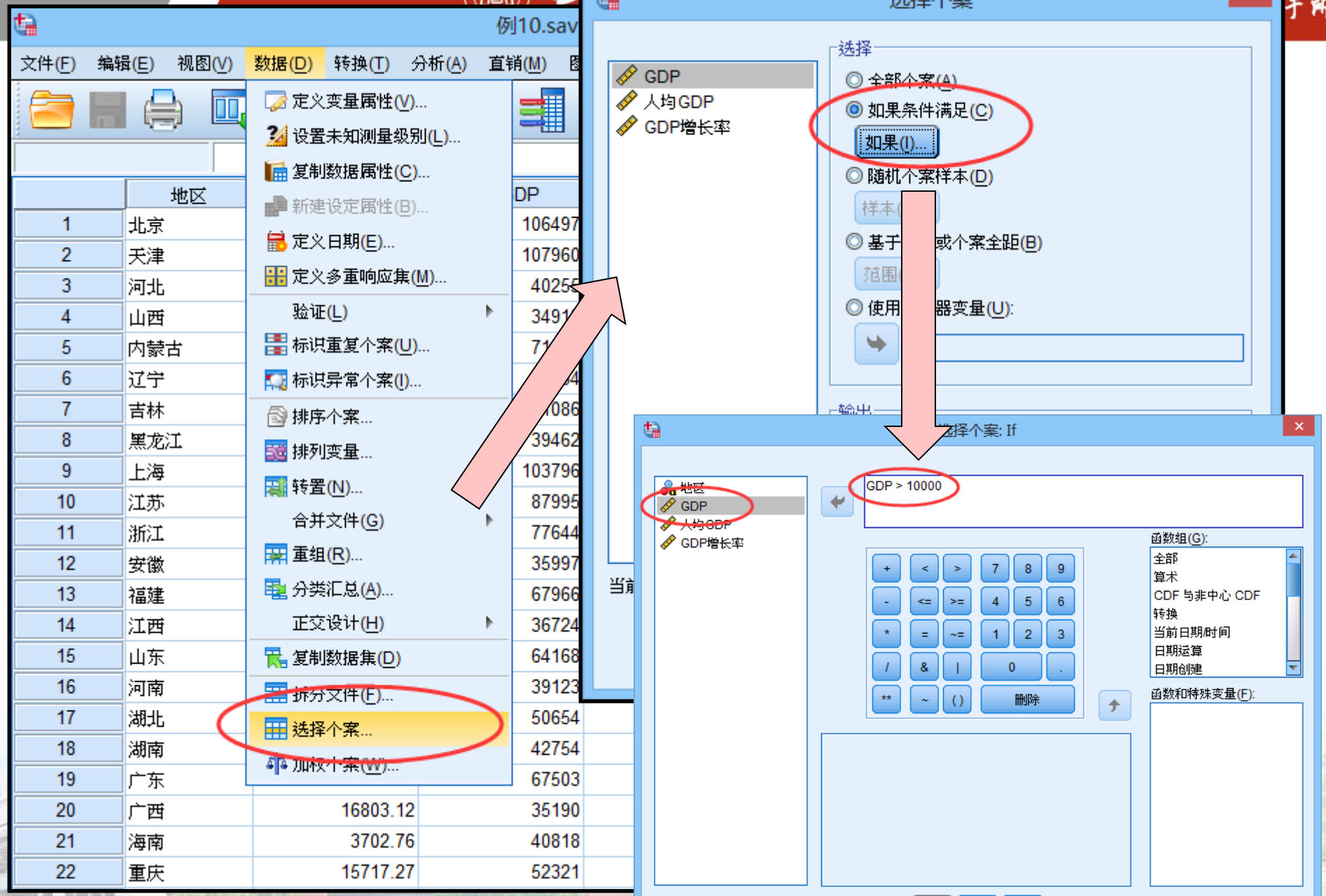
个案的选择

口根据不同的要求，从所有个案中筛选出特定的个案。可以通过给数据表设置选择条件或者过滤条件来满足这一要求。

- 按条件选择：给出一个条件表达式，选取符合该表达式的个案
- 按数据范围选择：选择一定的数据范围内的全部个案，要求给出数据范围的上、下界个案编号
- 随机选择：对数据编辑窗口中的所有个案进行随机筛选
- 过滤变量选择：选择制定的一个已存在的变量作为个案选取的标准

口例10-1：选择GDP大于10000亿元的地区

口例10-2：选择GDP增长率在6%-9%之间的地区


例10.sav
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M)
定义变量属性(V)...
设置未知测量级别(L)...
复制数据属性(C)...
新建设定属性(B)...
定义日期(E)...
定义多重响应集(M)...
验证(L)
标识重复个案(U)...
标识异常个案(I)...
排序个案...
排列变量...
转置(N)...
合并文件(G)
重组(R)...
分类汇总(A)...
正交设计(H)
复制数据集(D)
拆分文件(F)...
选择个案...
加权个案(W)...
地区
DP
1 北京 106497
2 天津 107960
3 河北 4025
4 山西 3491
5 内蒙古 71
6 辽宁 4
7 吉林 086
8 黑龙江 39462
9 上海 103796
10 江苏 87995
11 浙江 77644
12 安徽 35997
13 福建 67966
14 江西 36724
15 山东 64168
16 河南 39123
17 湖北 50654
18 湖南 42754
19 广东 67503
20 广西 16803.12 35190
21 海南 3702.76 40818
22 重庆 15717.27 52321
选择个案
GDP
人均GDP
GDP增长率
选择
全部个案(A)
如果条件满足(C)
如果(I)...
随机个案样本(D)
样本(
基于
或个案全距(B)
范围(
使用
器变量(U):
输出
当前
选择个案: If
地区
GDP
人均GDP
GDP增长率
GDP > 10000
+ < > 7 8 9
- <= >= 4 5 6
* = ~= 1 2 3
/ & | 0 .
** ~ ( ) 删除
函数组(G):
全部
算术
CDF 与非中心 CDF
转换
当前日期/时间
日期运算
日期创建
函数和特殊变量(F):
继续
取消
帮助

口来看看在Excel中如何处理这类问题

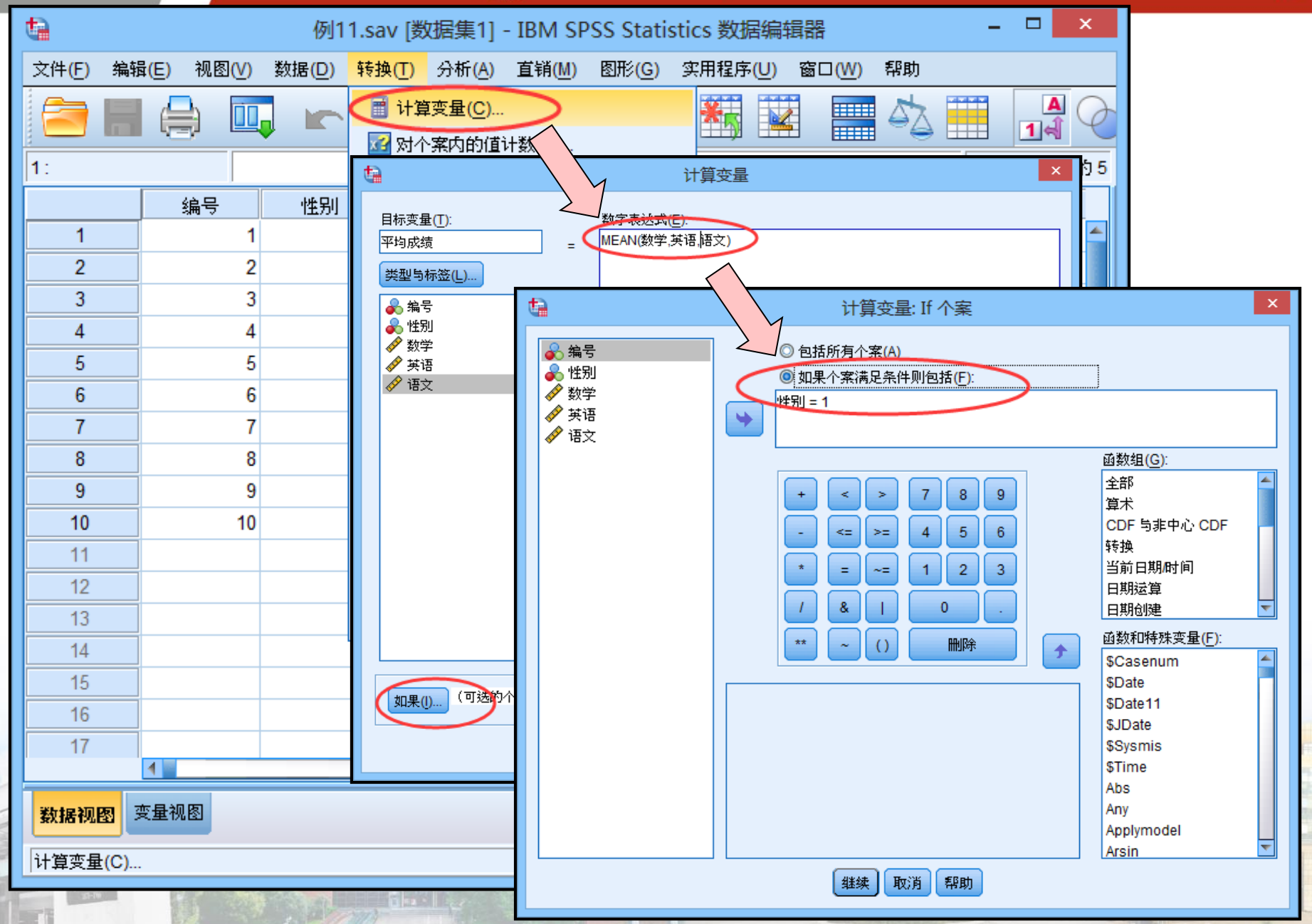
- “筛选”功能或者“if”函数

计算新变量

口使用SPSS算数表达式及函数，对所有记录或满足SPSS条件表达式的记录，计算出一个新结果，并将结果存入一个指定的变量中。

- 例11：计算例11. sav文件数据中男生的平均成绩。

四川大学图书馆
淡泊明志 宁静致远
志于所学 继续求学



口来看看在Excel中如何处理这类问题

- “if”函数

剪贴板 | 蓝牙 | 字体 | 对齐方式 | 数字

F2 | fx | =IF(B2="男",AVERAGE(C2,D2,E2),0)

	A	B	C	D	E	F	G
1	编号	性别	数学	英语	语文	平均成绩	
2	1	男	86	85	90	87	
3	2	女	85	95	89		
4	3	女	89	67	86		
5	4	男	87	85	83		
6	5	女	84	81	84		
7	6	男	73	52	79		
8	7	男	90	86	88		
9	8	女	89	76	86		
10	9	男	94	83	84		
11	10	男	75	80	83		
12							
13							

变量值的重新编码

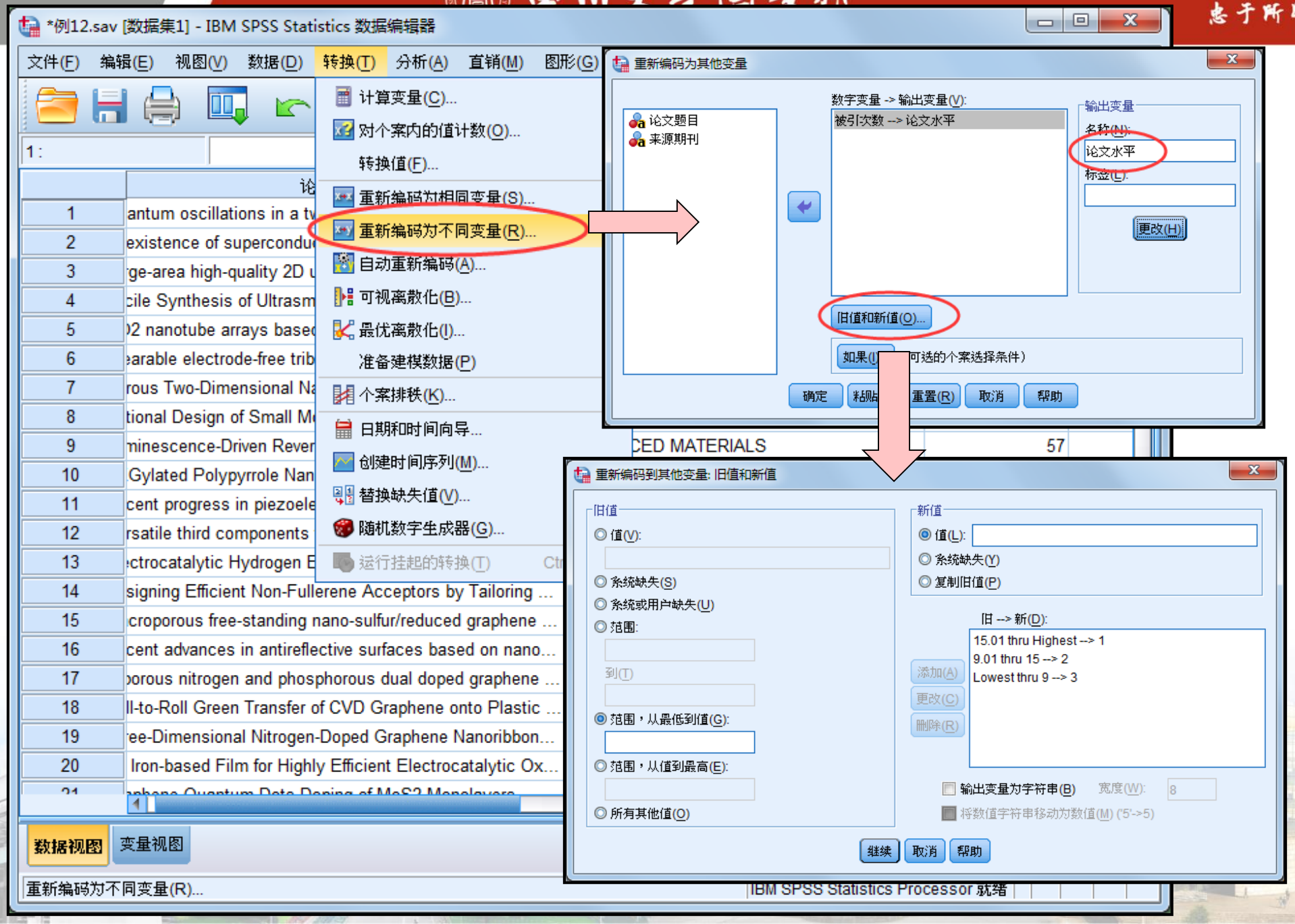
口数据分析中，将连续变量转换为等级变量，或者将分类变量不同的变量等级进行合并，例如把同学的成绩分为优、良、中、差四个等级。

- 重新编码为相同变量：对原始变量的取值进行修改，用新编码直接取代原变量的取值。
- 重新编码为不同变量：将新编码存入新的变量，根据原始变量的取值生成一个新变量来表示分组情况。

口例12：将例12.sav中论文按被引次数分成三个等级。

- 论文发表年份：2015；所在学科：材料科学；
- 引用次数基准线：
  - 15<citation：前10%，高水平论文
  - 9<citation≤15：10%-50%，优秀论文
  - Citation≤9：50%-，一般论文

四川大学图书馆
淡泊明志 宁静致远
志于所学 继续求学



3-SPSS的统计分析功能

- 3.1-统计描述分析
- 3.2-T检验
- 3.3-方差分析
- 3.4-线性回归与相关
- 3.5-聚类分析
- 3.6-因子分析

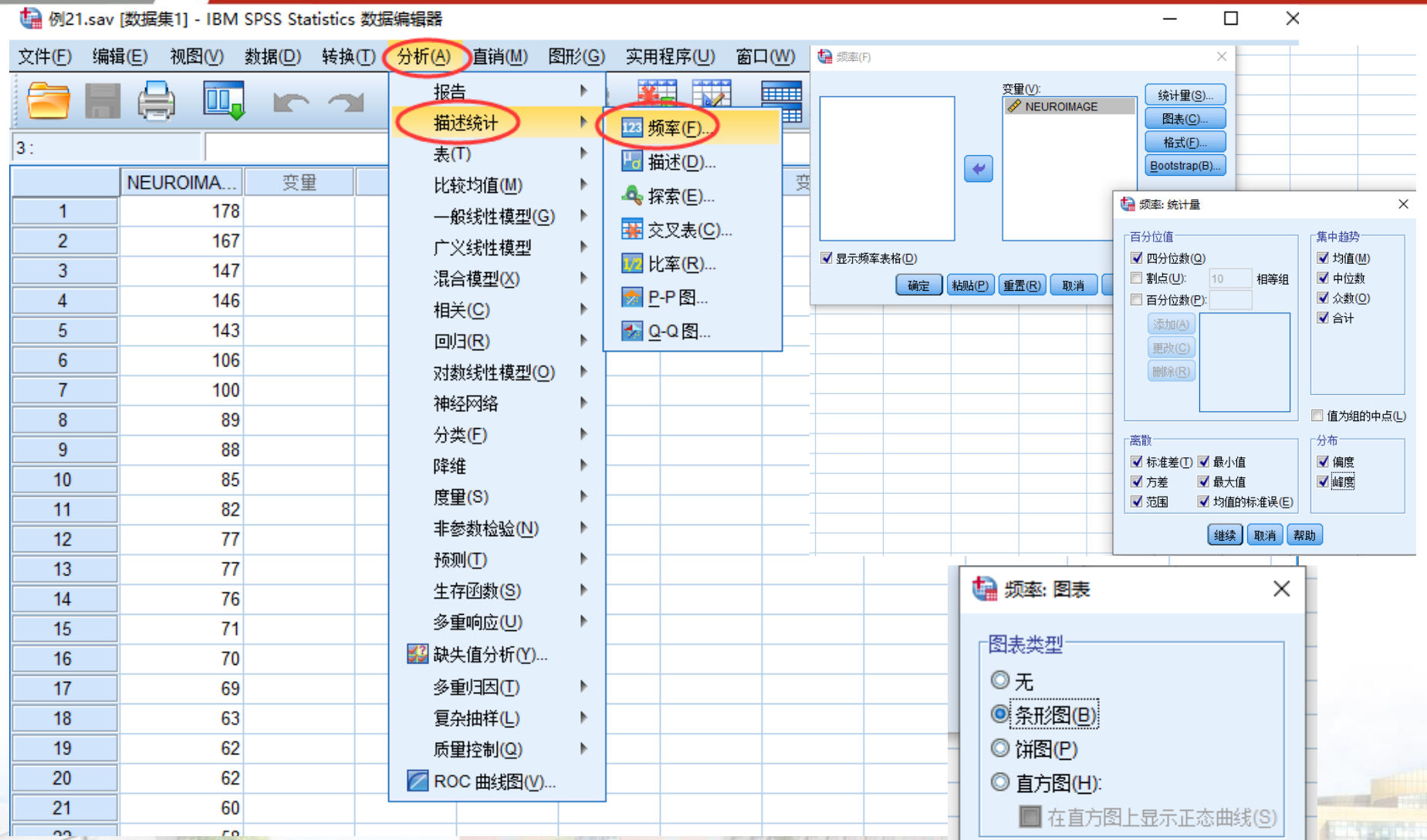
3.1-统计描述分析

口统计描述分析是为了对总体特征有比较准确的把握。

- 频数分布分析
- 描述性统计分析

- 频数分布分析主要通过频数分布表、条图、直方图以及集中趋势和离散趋势各种统计量，描述数据的分布特征。
- 例21：了解全球范围内“神经图像学”论文被引次数的分布特征

四川大学图书馆
淡泊明志 宁静致远
志于所学 继续求学

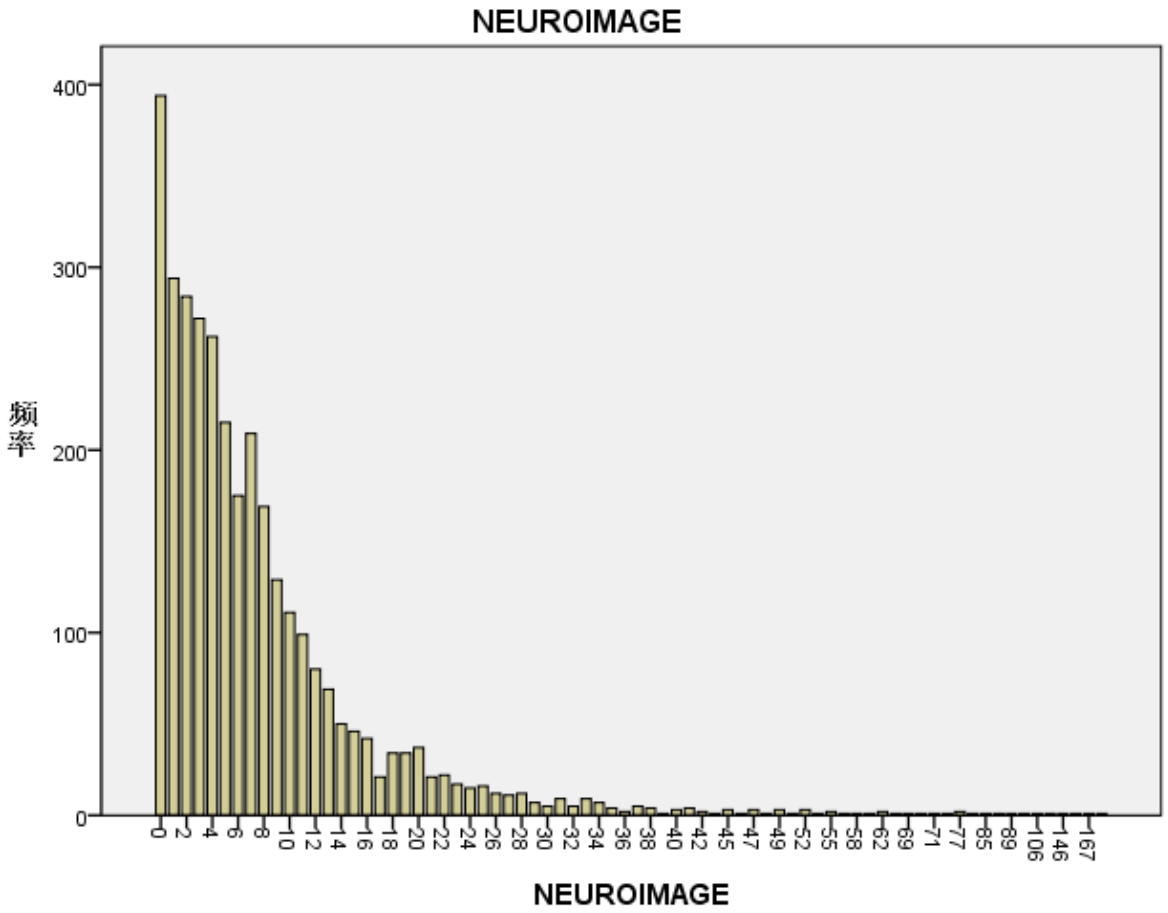


	NEUROIMA...	变量
1	178	
2	167	
3	147	
4	146	
5	143	
6	106	
7	100	
8	89	
9	88	
10	85	
11	82	
12	77	
13	77	
14	76	
15	71	
16	70	
17	69	
18	63	
19	62	
20	62	
21	60	

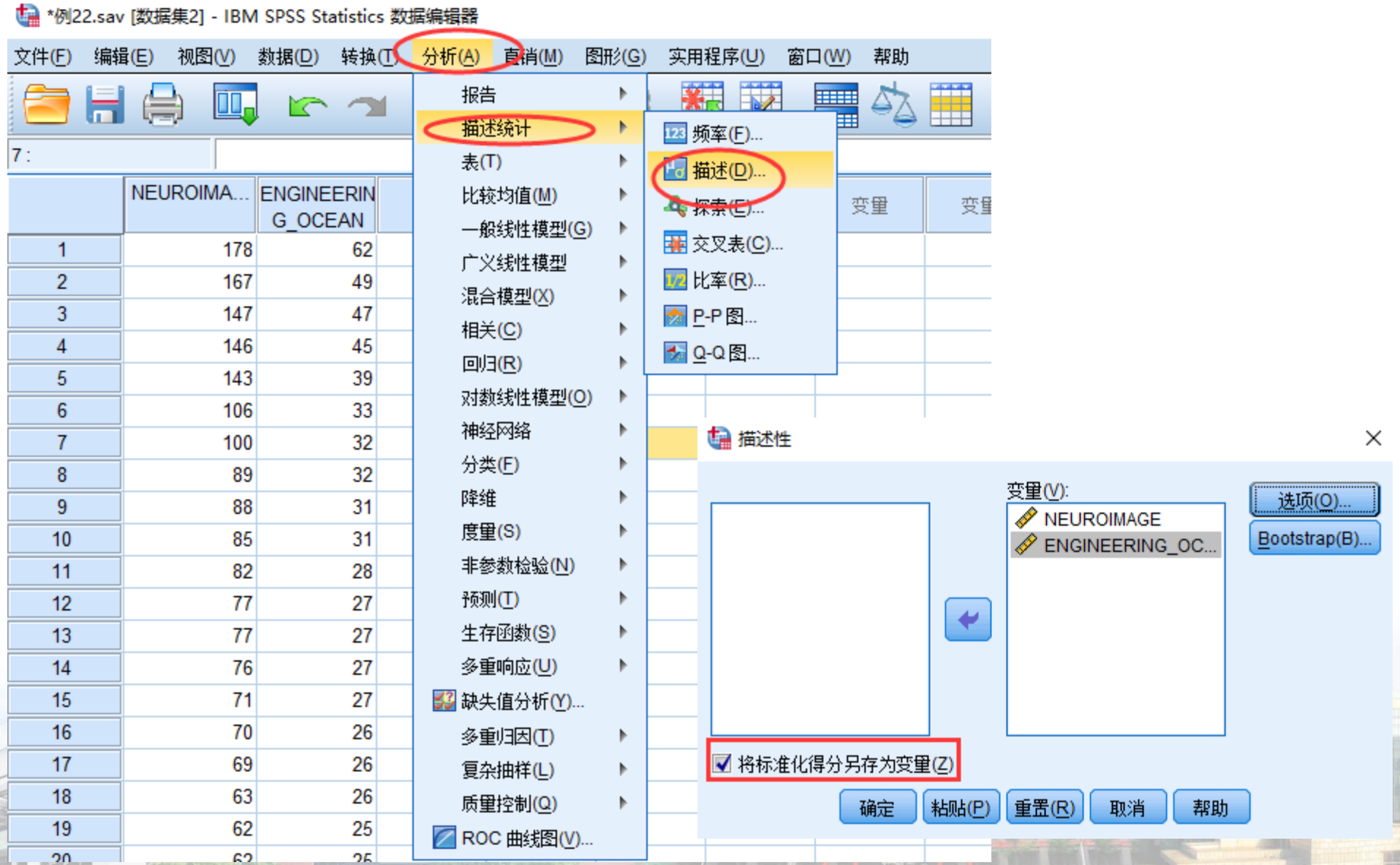
统计量

NEUROIMAGE

N	有效	3261
	缺失	2571
均值		7.87
均值的标准误		.193
中值		5.00
众数		0
标准差		11.024
方差		121.531
偏度		5.953
偏度的标准误		.043
峰度		62.138
峰度的标准误		.086
全距		178
极小值		0
极大值		178
和		25668
百分位数	25	2.00
	50	5.00
	75	10.00



- 描述性统计分析
- 主要用以计算描述集中趋势和离散趋势的各种统计量，另外还有一个重要功能是对变量做标准化变化，即Z变换。
- 例22：分析全球范围内“神经图像学”和“海洋工程”2个学科论文被引次数的统计特征以及比较任意两篇论文的被引次数


*例22.sav [数据集2] - IBM SPSS Statistics 数据编辑器
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W) 帮助
7 :
NEUROIMA...
ENGINEERING_OCEAN
1 178 62
2 167 49
3 147 47
4 146 45
5 143 39
6 106 33
7 100 32
8 89 32
9 88 31
10 85 31
11 82 28
12 77 27
13 77 27
14 76 27
15 71 27
16 70 26
17 69 26
18 63 26
19 62 25
报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
相关(C)
回归(R)
对数线性模型(O)
神经网络
分类(F)
降维
度量(S)
非参数检验(N)
预测(T)
生存函数(S)
多重响应(U)
缺失值分析(Y)...
多重归因(T)
复杂抽样(L)
质量控制(Q)
ROC 曲线图(V)...
频率(F)...
描述(D)...
探索(E)...
交叉表(C)...
比率(R)...
P-P 图...
Q-Q 图...
变量
描述性
变量(V):
NEUROIMAGE
ENGINEERING_OC...
选项(O)...
Bootstrap(B)...
将标准化得分另存为变量(Z)
确定
粘贴(P)
重置(R)
取消
帮助

口显然，从发文量、被引次数、离散趋势等各方面看，不同学科的引文存在着明显的差异，如何进行跨学科的横向对比，是科研评价领域一个非常重要的研究问题

描述统计量

	N	全距	极小值	极大值	均值		标准差	方差	偏度		峰度	
	统计量	统计量	统计量	统计量	统计量	标准误	统计量	统计量	统计量	标准误	统计量	标准误
NEUROIMAGE	3261	178	0	178	7.87	.193	11.024	121.531	5.953	.043	62.138	.086
ENGINEERING_OCEAN	2571	62	0	62	2.98	.094	4.784	22.886	3.725	.048	23.955	.097
有效的 N（列表状态）	2571											

3.2-T检验

- T检验是检验样本的均值和给定的均值是否存在显著性差异。T检验分为三类：
- 单样本检验（单个总体，方差未知，均值的检验）
- 独立样本检验（两个总体，方差未知但相等，均值是否相等的检验）
- 配对样本检验

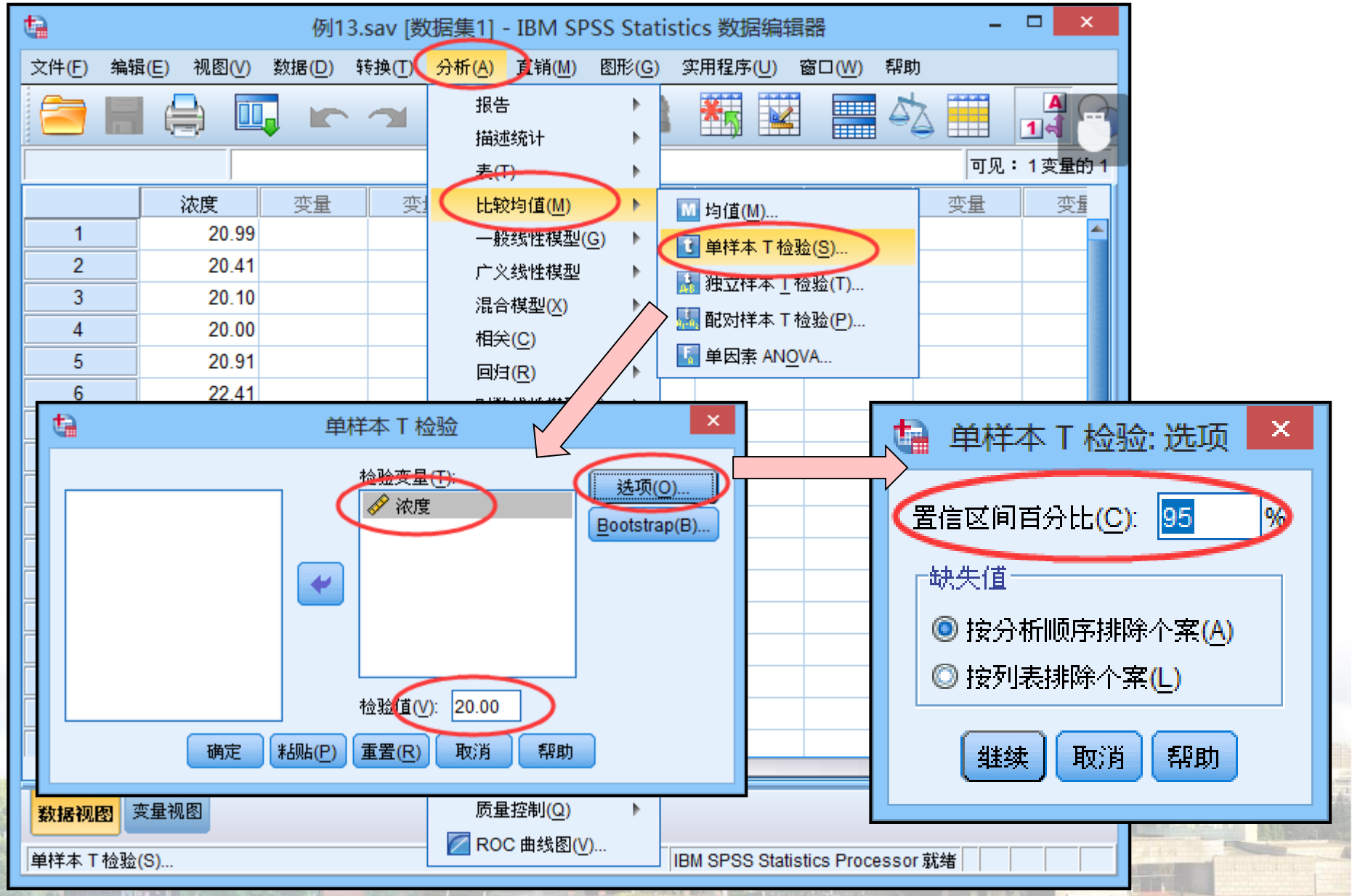
口总体分布为正态分布$N(\mu,\sigma^2)$时，需要检验$H_0:\mu=\mu_0$。

- 检验统计量$t=\frac{\bar{x}-\mu}{s}\sqrt{n}$
- SPSS将自动把样本均值μ_0、样本方差、样本数带入上式，计算出t统计量的观测值和对应的概率P值。
- 如果概率P值小于显著性水平α，则拒绝原假设，认为总体均值与检验值之间存在显著差异；反之则接受原假设。

单样本检验

口例13：某药物在某种溶剂中溶解后的标准浓度为20.00mg/L，现采用某种方法，测量该药物溶解度11次，测量后得到的结果见例13.sav，问：用该方法测量所得结果是否与标准浓度值有所不同?

单样本检验


例13.sav [数据集1] - IBM SPSS Statistics 数据编辑器
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W) 帮助
报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
相关(C)
回归(R)
均值(M)...
单样本 T 检验(S)...
独立样本 T 检验(T)...
配对样本 T 检验(P)...
单因素 ANOVA...
浓度
1 20.99
2 20.41
3 20.10
4 20.00
5 20.91
6 22.41
单样本 T 检验
检验变量(T):
浓度
选项(O)...
Bootstrap(B)...
检验值(V): 20.00
确定 粘贴(P) 重置(R) 取消 帮助
单样本 T 检验: 选项
置信区间百分比(C): 95 %
缺失值
按分析顺序排除个案(A)
按列表排除个案(L)
继续 取消 帮助
数据视图 变量视图
质量控制(Q)
ROC 曲线图(V)...
单样本 T 检验(S)...
IBM SPSS Statistics Processor 就绪

单样本检验

口结果显示

- P值=0.012<0.05，因此认为测量所得结果与标准浓度值有差异

T检验

[数据集1] I:\数据\例13.sav

单个样本统计量

	N	均值	标准差	均值的标准误
浓度	11	20.9836	1.06750	.32186

单个样本检验

	检验值 = 20.00					
					差分的 95% 置信区间	
	t	df	Sig.(双侧)	均值差值	下限	上限
浓度	3.056	10	.012	.98364	.2665	1.7008

独立样本检验

口两个独立样本符合正态分布，且满足方差齐性。

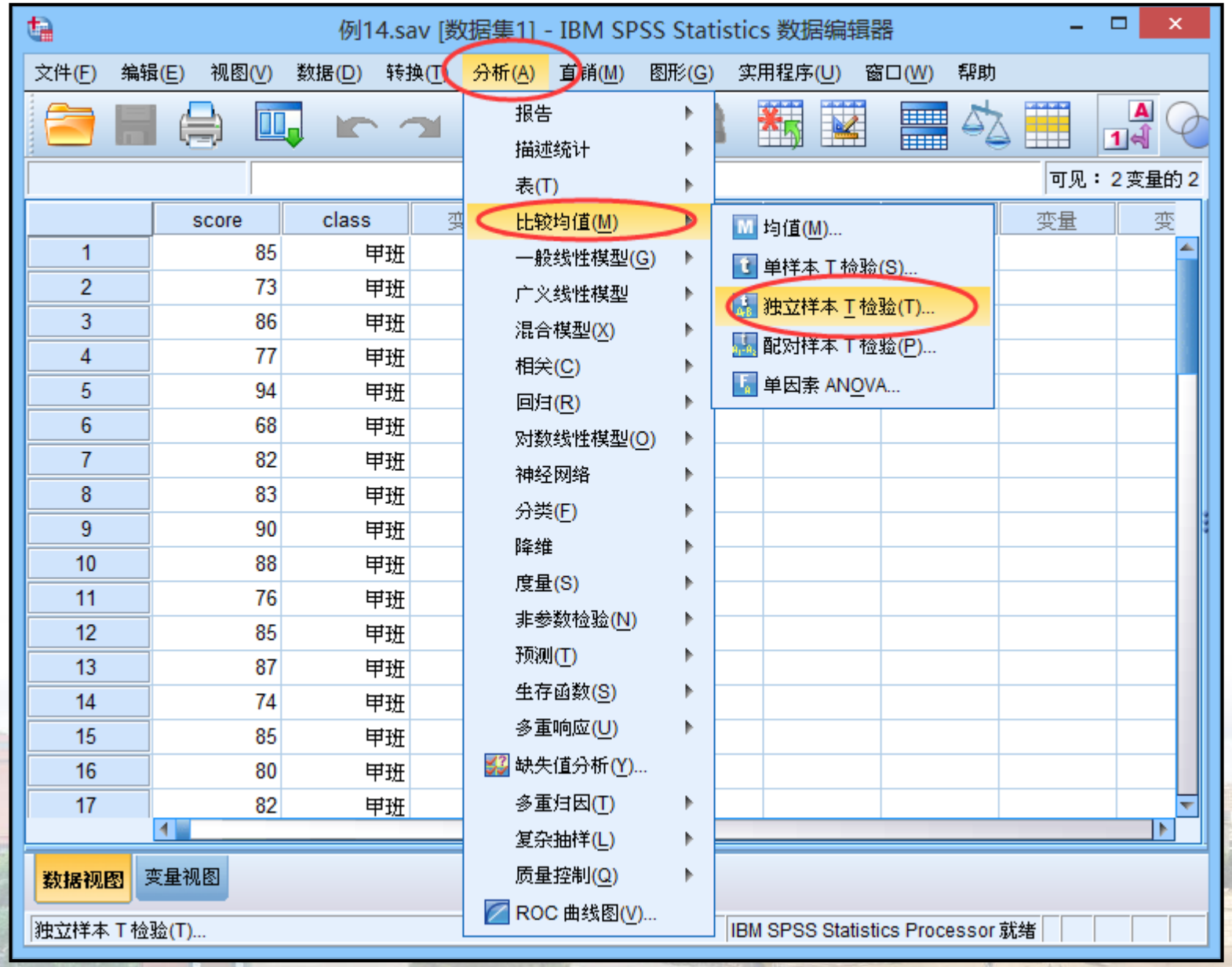
需要检验$H_0: \mu_1 - \mu_2 = 0$。

- 选取检验统计量为t统计量，$t = \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$
- 计算F统计量和t统计量的观测值以及相应的概率P值。
- 利用F检验判断两总体的方差是否相等。
- 利用t检验判断两总体的均值是否存在显著差异。

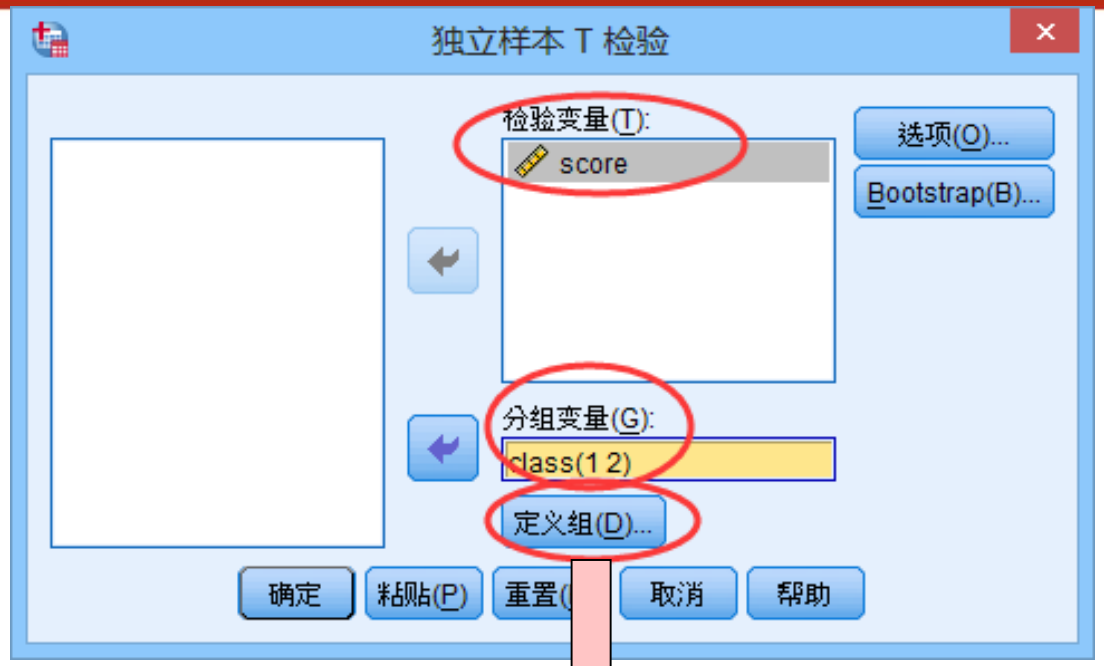
独立样本检验

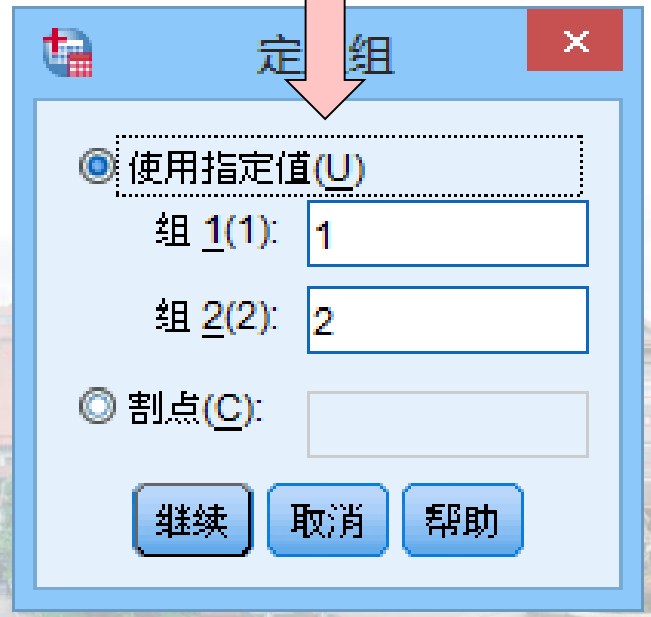
口例14：现希望评价两位老师的教学质量，试比较其分别任教的两班考试后的成绩是否存在差异。数据见例14.sav。

独立样本检验


例14.sav [数据集1] - IBM SPSS Statistics 数据编辑器
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W) 帮助
可见：2 变量的 2
score class
1 85 甲班
2 73 甲班
3 86 甲班
4 77 甲班
5 94 甲班
6 68 甲班
7 82 甲班
8 83 甲班
9 90 甲班
10 88 甲班
11 76 甲班
12 85 甲班
13 87 甲班
14 74 甲班
15 85 甲班
16 80 甲班
17 82 甲班
报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
相关(C)
回归(R)
对数线性模型(O)
神经网络
分类(F)
降维
度量(S)
非参数检验(N)
预测(T)
生存函数(S)
多重响应(U)
缺失值分析(Y)...
多重归因(T)
复杂抽样(L)
质量控制(Q)
ROC 曲线图(V)...
均值(M)...
单样本 T 检验(S)...
独立样本 T 检验(T)...
配对样本 T 检验(P)...
单因素 ANOVA...
变量
数据视图
变量视图
独立样本 T 检验(T)...
IBM SPSS Statistics Processor 就绪

独立样本检验





*“使用指定值”*表示用实现定义好的变量值表示不同的分组，在本例中，甲班的值为1，乙班的值为2；

*“割点”*表示分组变量为连续变量时，输入一个数字，大于等于该数值的为一个总体，对应一组样本，小于该值的为另一总体，对应另一组样本

独立样本检验

口结果显示

- F检验的P值为0.397>0.05，故方差齐性。
- 不同组间独立样本t检验P值为0.004<0.05，因此认为甲、乙两班的成绩存在差异。

T检验

[数据集1] I:\数据\例14.sav

组统计量

	class	N	均值	标准差	均值的标准误
score	甲班	20	83.30	6.906	1.544
	乙班	20	75.45	9.179	2.053

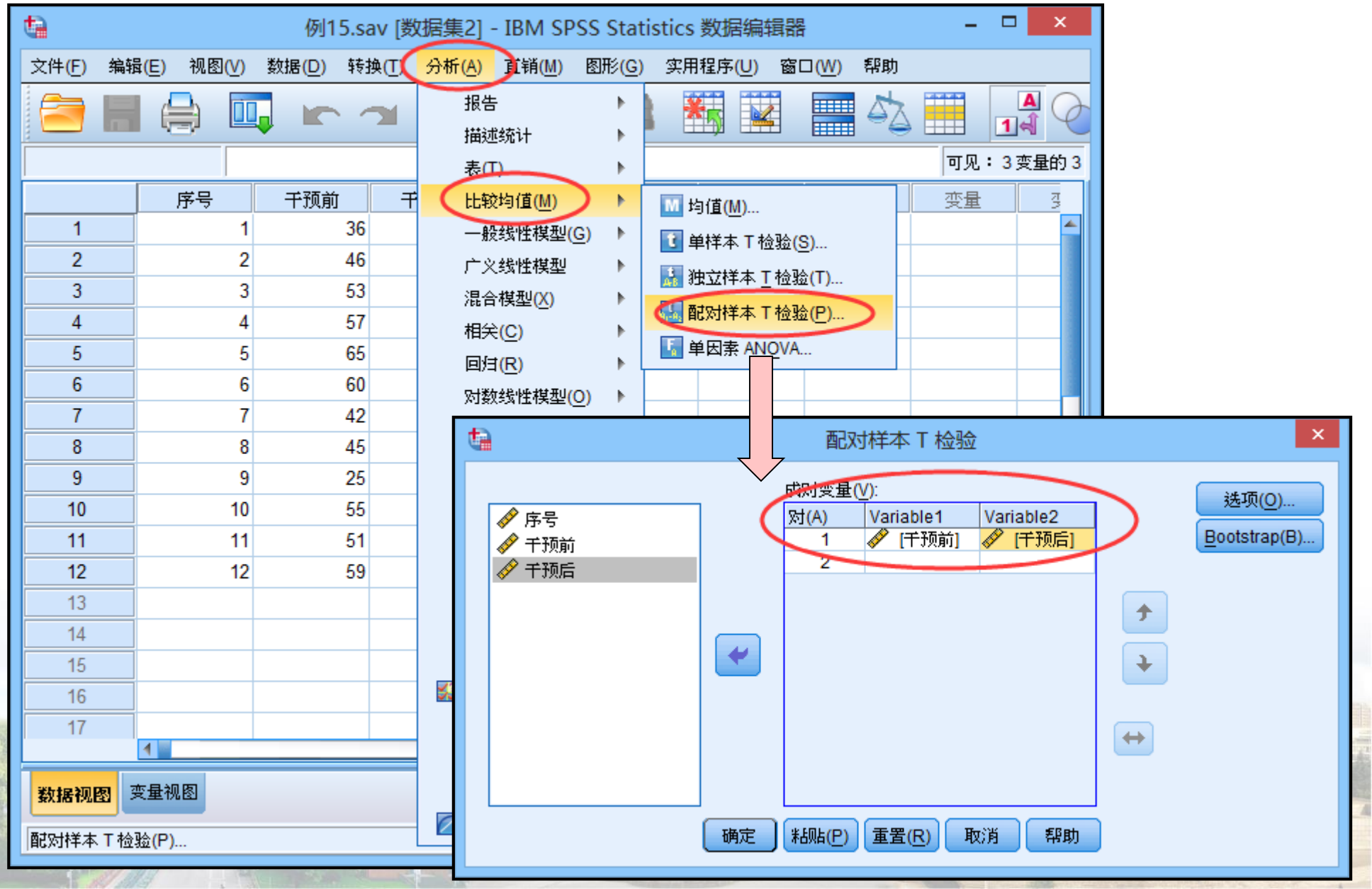
独立样本检验

		方差方程的 Levene 检验		均值方程的 t 检验						
									差分的 95% 置信区间	
		F	Sig.	t	df	Sig.(双侧)	均值差值	标准误差值	下限	上限
score	假设方差相等	.733	.397	3.056	38	.004	7.850	2.569	2.650	13.050
	假设方差不相等			3.056	35.290	.004	7.850	2.569	2.637	13.063

配对样本检验

- 利用来自两个不同总体的配对样本，推断两个总体的均值是否有差异。
- 对两组样本分别计算每对观察值的差值得到差值样本，然后利用差值样本，通过对其均值是否显著为0的检验来推断两总体均值的差是否显著为0。
- 例15：某地区随机抽取12名贫血儿童的家庭，实行健康教育干预三个月，干预前后儿童的血红蛋白（%）测量结果见例15.sav，试问干预前后该地区贫血儿童血红蛋白（%）平均水平有无变化?

配对样本检验


例15.sav [数据集2] - IBM SPSS Statistics 数据编辑器
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W) 帮助
报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
相关(C)
回归(R)
对数线性模型(O)
均值(M)...
单样本 T 检验(S)...
独立样本 T 检验(T)...
配对样本 T 检验(P)...
单因素 ANOVA...
可见：3 变量的 3
序号 千预前
1 1 36
2 2 46
3 3 53
4 4 57
5 5 65
6 6 60
7 7 42
8 8 45
9 9 25
10 10 55
11 11 51
12 12 59
数据视图 变量视图
配对样本 T 检验(P)...
配对样本 T 检验
序号
千预前
千预后
成对变量(V):
对(A) Variable1 Variable2
1 [千预前] [千预后]
2
选项(O)...
Bootstrap(B)...
确定 粘贴(P) 重置(R) 取消 帮助

配对样本检验

口结果显示

- 统计量P值为0.007<0.05，因此认为干预前后该地区贫血儿童血红蛋白（%）水平有变化。

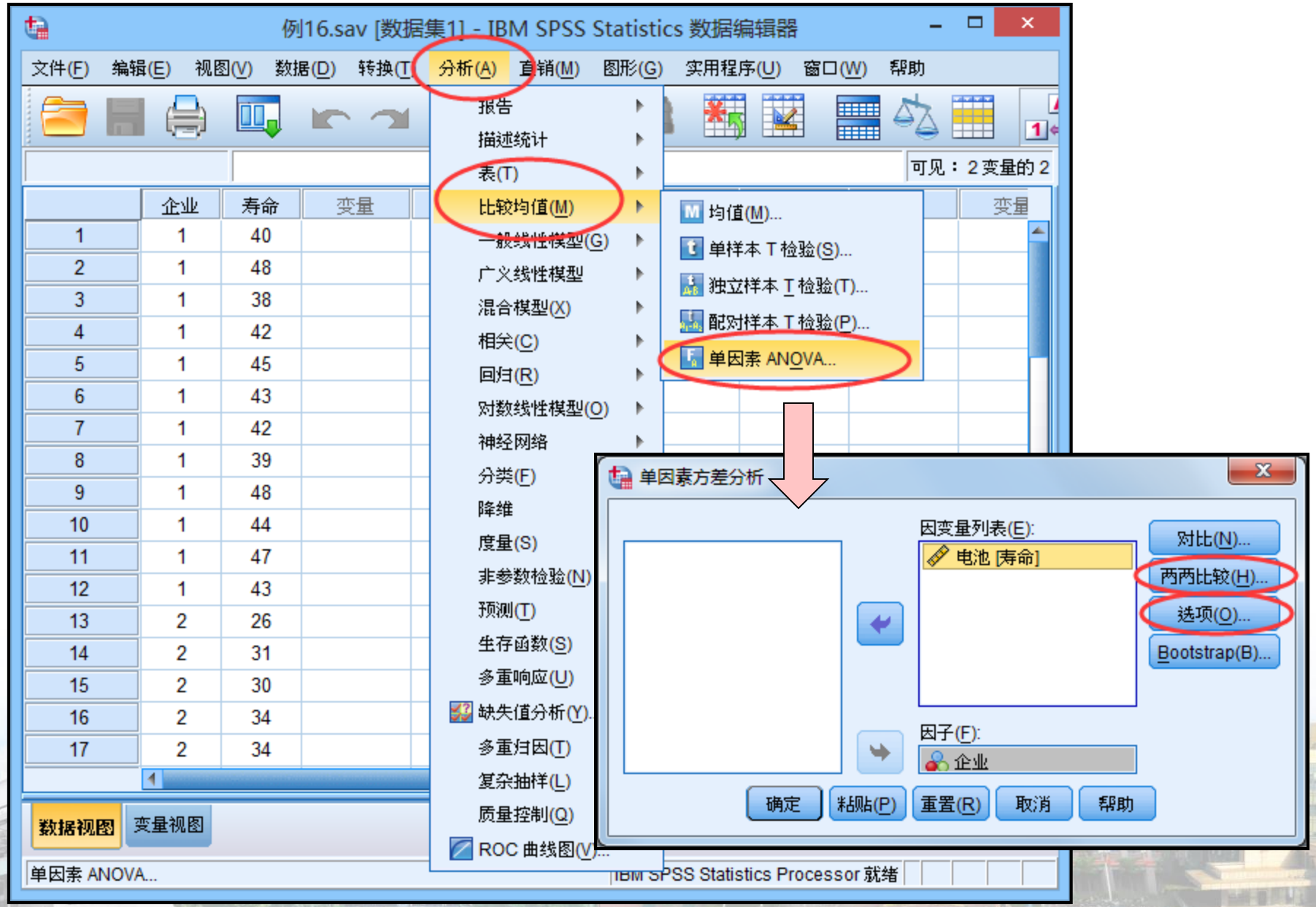
成对样本检验

		成对差分					t	df	Sig.(双侧)
		均值	标准差	均值的标准误	差分的 95% 置信区间 下限	差分的 95% 置信区间 上限			
对 1	干预前 - 干预后	-10.667	11.179	3.227	-17.769	-3.564	-3.305	11	.007

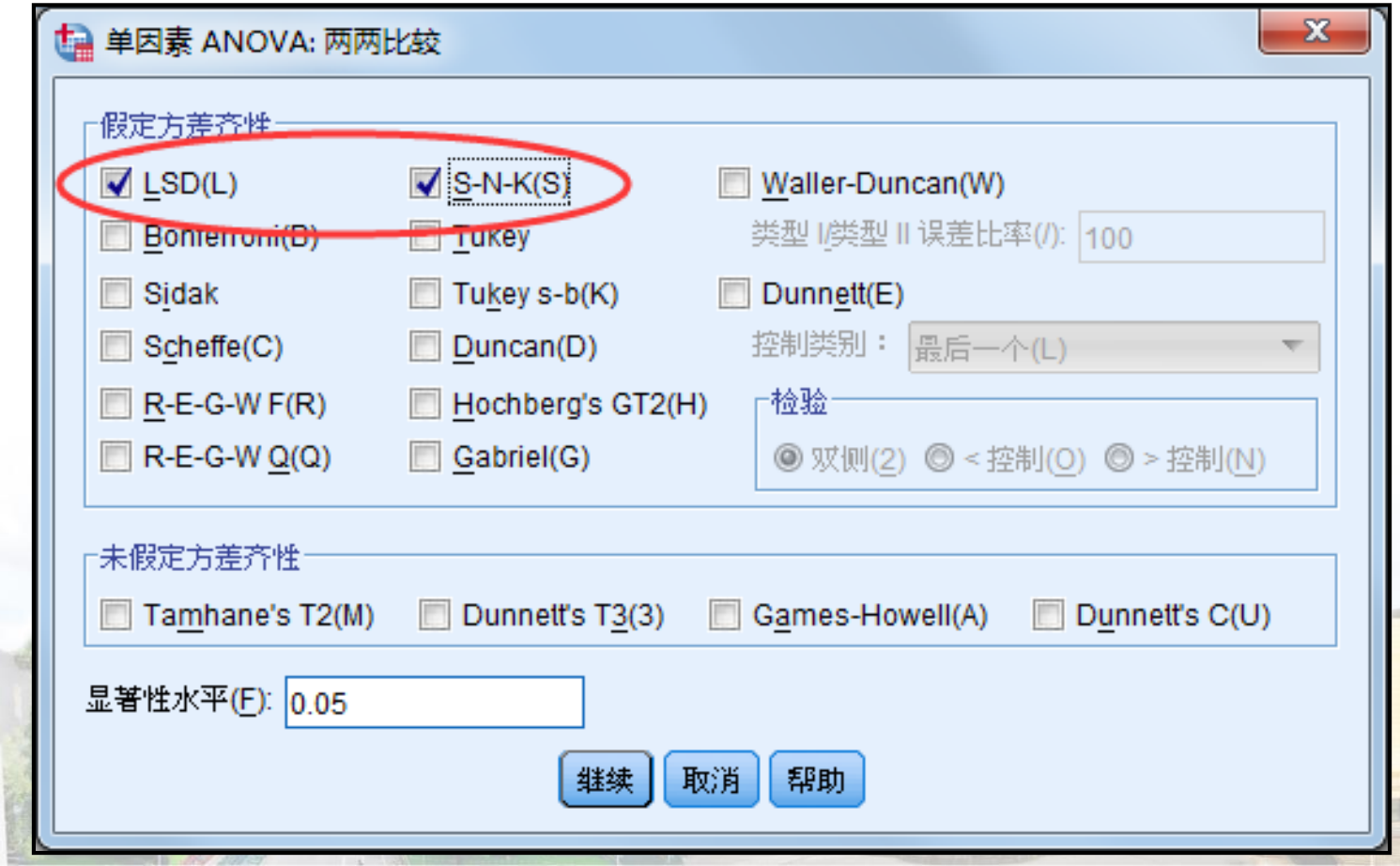
3.3-方差分析

- 当比较两组资料均值是否相等时，可以采用t检验。当组数大于2组时，如果仍然采用t检验，这不但复杂，而且有很大的可能性导致错误结论。这时应该采用方差分析。
- 方差分析的应用条件如下：独立；正态；方差齐性。

口例16：比较3个不同电池生产企业生产电池的寿命，见例16.sav

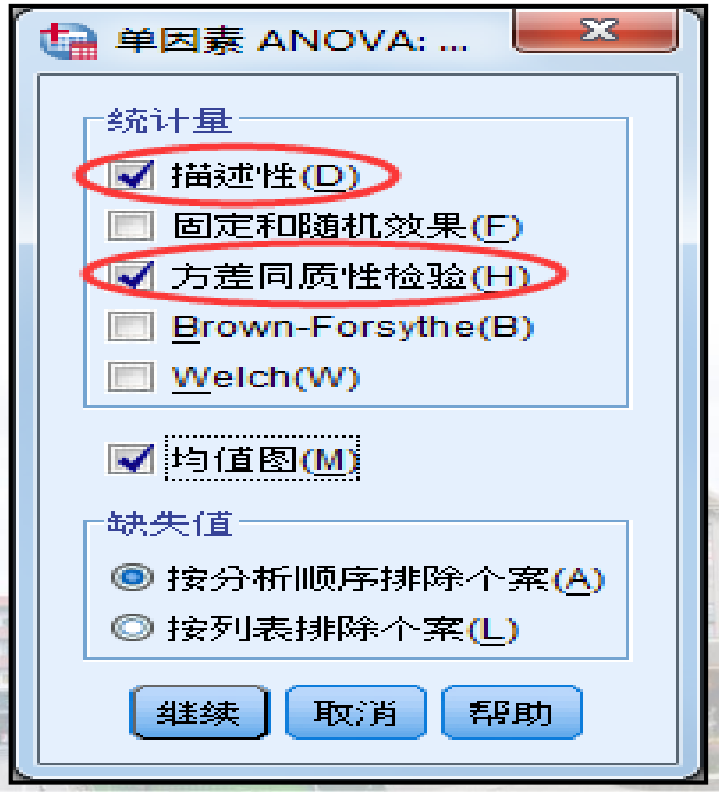

例16.sav [数据集1] - IBM SPSS Statistics 数据编辑器
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W) 帮助
报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
相关(C)
回归(R)
对数线性模型(O)
神经网络
分类(F)
降维
度量(S)
非参数检验(N)
预测(T)
生存函数(S)
多重响应(U)
缺失值分析(Y)...
多重归因(T)
复杂抽样(L)
质量控制(Q)
ROC 曲线图(V)...
均值(M)...
单样本 T 检验(S)...
独立样本 T 检验(T)...
配对样本 T 检验(P)...
单因素 ANOVA...
可见：2 变量的 2
企业 寿命 变量
1 1 40
2 1 48
3 1 38
4 1 42
5 1 45
6 1 43
7 1 42
8 1 39
9 1 48
10 1 44
11 1 47
12 1 43
13 2 26
14 2 31
15 2 30
16 2 34
17 2 34
数据视图 变量视图
单因素 ANOVA...
IBM SPSS Statistics Processor 就绪
单因素方差分析
因变量列表(E):
电池 [寿命]
因子(F):
企业
对比(N)...
两两比较(H)...
选项(O)...
Bootstrap(B)...
确定 粘贴(P) 重置(R) 取消 帮助

口两两比较：如果结果显示不同企业生产的电池寿命存在差异，那么通过“两两比较”可以获知是哪两个厂家的电池有差异。



口选项

- 描述性：显示每个因变量的个数、均值、标准差等信息
- 方差同质性检验：计算Levene方差齐性检验



口结果显示-1

描述

电池

	N	均值	标准差	标准误	均值的 95% 置信区间		极小值	极大值
					下限	上限		
1	12	43.25	3.334	.962	41.13	45.37	38	48
2	12	32.33	3.576	1.032	30.06	34.61	26	37
3	12	43.83	3.881	1.120	41.37	46.30	39	50
总数	36	39.81	6.405	1.067	37.64	41.97	26	50

口结果显示-2

- 方差齐性检验：显著性=0. 680>0. 05，各组方差齐性

方差齐性检验

电池

Levene 统计量	df1	df2	显著性
.390	2	33	.680

口结果显示-3

- 显著性=0. 000<0. 05，表示三个厂家生产的电池寿命存在差异

ANOVA

电池

	平方和	df	均方	F	显著性
组间	1007.056	2	503.528	38.771	.000
组内	428.583	33	12.987		
总数	1435.639	35			

口结果显示-4（LSD法结果）

- 企业1与企业2显著性=0. 000<0. 05，存在差异
- 企业1与企业3显著性=0. 694>0. 05，无差异
- 企业2与企业3显著性=0. 000<0. 05，存在差异

多重比较

因变量:电池

	(I) 企业	(J) 企业	均值差 (I-J)	标准误	显著性	95% 置信区间	
						下限	上限
LSD	1	2	10.917*	1.471	.000	7.92	13.91
		3	-.583	1.471	.694	-3.58	2.41
	2	1	-10.917*	1.471	.000	-13.91	-7.92
		3	-11.500*	1.471	.000	-14.49	-8.51
	3	1	.583	1.471	.694	-2.41	3.58
		2	11.500*	1.471	.000	8.51	14.49

*. 均值差的显著性水平为 0.05。

口结果显示-5（SNK法结果）

- 企业2是一个类别，企业1与企业3是一个类别

同类子集

电池

	企业	N	alpha = 0.05 的子集	
			1	2
Student-Newman-Keuls[a]	2	12	32.33	
	1	12		43.25
	3	12		43.83
	显著性		1.000	.694

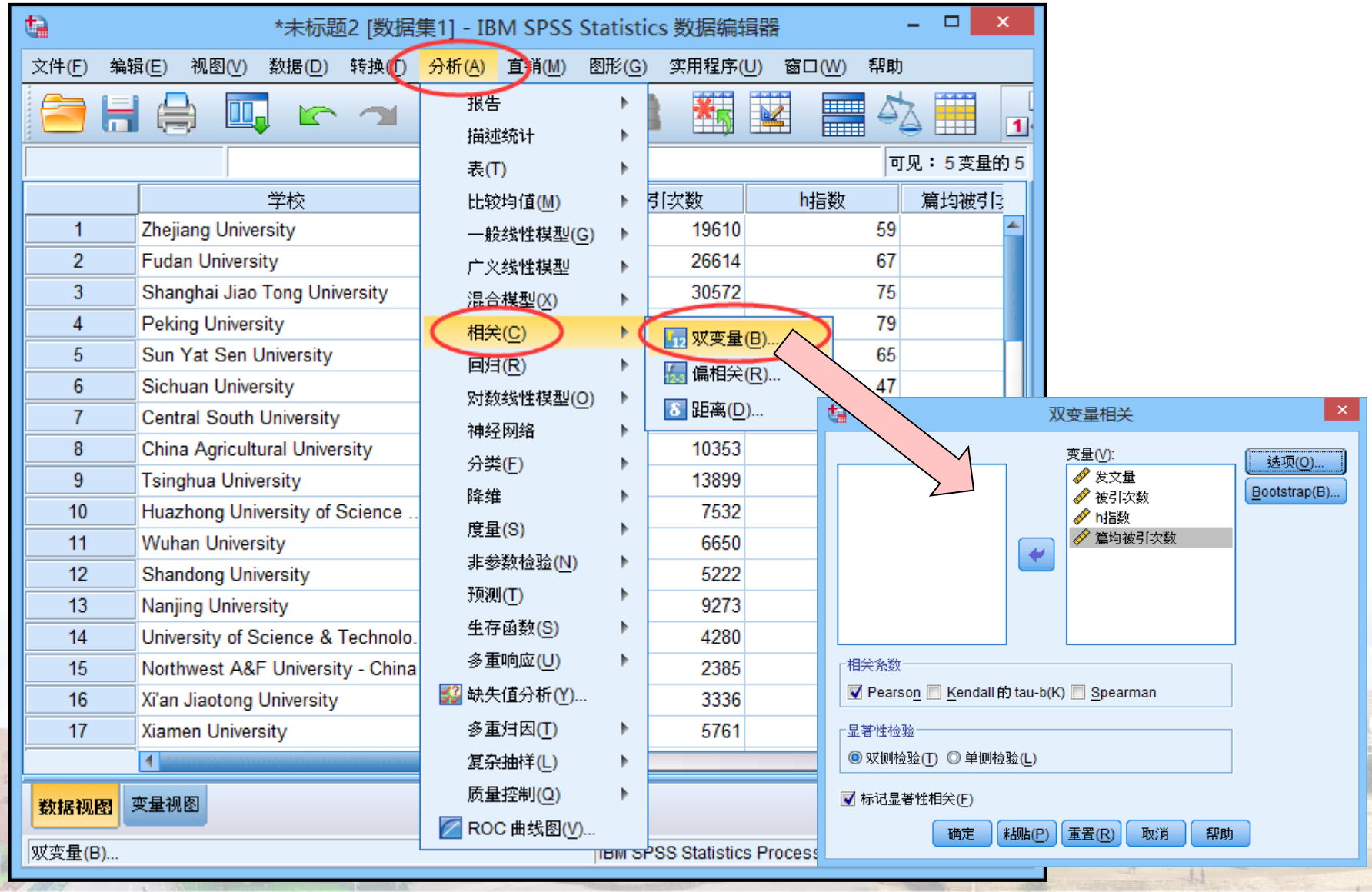
将显示同类子集中的组均值。

a. 将使用调和均值样本大小 = 12.000。

3.4-线性回归与相关

- 线性相关系数：Pearson相关系数
- 取值范围-1≤r≤1
- 绝对值越接近1，表示两变量间的相关关系的密切程度越高
- 例17：分析发文量、被引次数、h指数、篇均被引次数4个指标之间的相关性。相关数据见例17.sav。

3.3-线性回归与相关


*未标题2 [数据集1] - IBM SPSS Statistics 数据编辑器
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W) 帮助
可见：5 变量的 5
学校
h指数
1 Zhejiang University 19610 59
2 Fudan University 26614 67
3 Shanghai Jiao Tong University 30572 75
4 Peking University 79
5 Sun Yat Sen University 65
6 Sichuan University 47
7 Central South University
8 China Agricultural University 10353
9 Tsinghua University 13899
10 Huazhong University of Science 7532
11 Wuhan University 6650
12 Shandong University 5222
13 Nanjing University 9273
14 University of Science & Technolo. 4280
15 Northwest A&F University - China 2385
16 Xi'an Jiaotong University 3336
17 Xiamen University 5761
报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
相关(C)
回归(R)
对数线性模型(O)
神经网络
分类(F)
降维
度量(S)
非参数检验(N)
预测(T)
生存函数(S)
多重响应(U)
缺失值分析(Y)...
多重归因(T)
复杂抽样(L)
质量控制(Q)
ROC 曲线图(V)...
双变量(B)...
偏相关(R)...
距离(D)...
数据视图
变量视图
双变量(B)...
双变量相关
变量(V):
发文量
被引次数
h指数
篇均被引次数
选项(O)...
Bootstrap(B)...
相关系数
Pearson
Kendall 的 tau-b(K)
Spearman
显著性检验
双侧检验(T)
单侧检验(L)
标记显著性相关(F)
确定
粘贴(P)
重置(R)
取消
帮助

3.3-线性回归与相关

口结果显示

相关性

[数据集1]

相关性

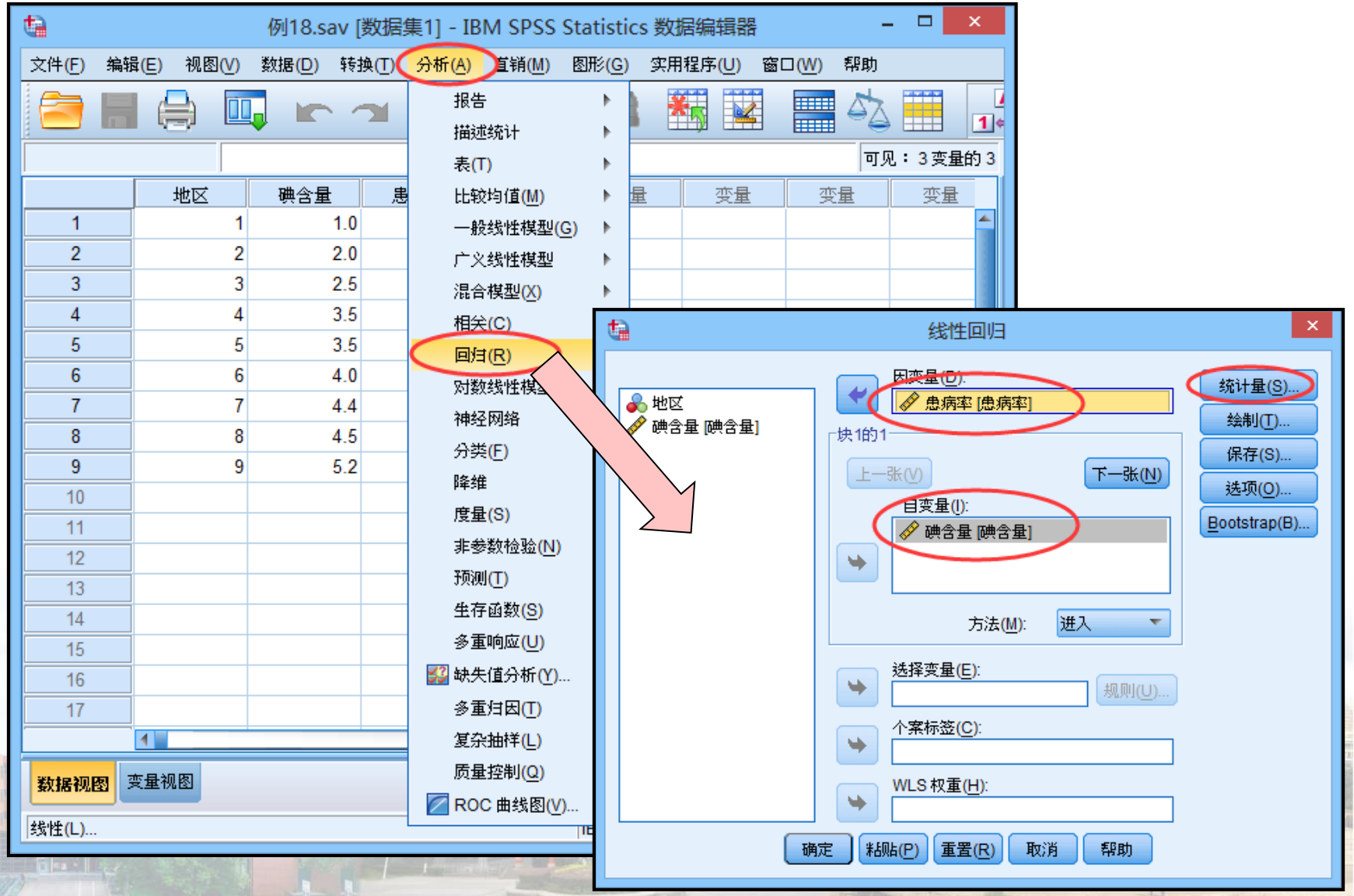
		发文量	被引次数	h指数	篇均被引次数
发文量	Pearson 相关性	1	.947**	.913**	-.104
	显著性（双侧）		.000	.000	.530
	N	39	39	39	39
被引次数	Pearson 相关性	.947**	1	.926**	-.032
	显著性（双侧）	.000		.000	.847
	N	39	39	39	39
h指数	Pearson 相关性	.913**	.926**	1	-.147
	显著性（双侧）	.000	.000		.372
	N	39	39	39	39
篇均被引次数	Pearson 相关性	-.104	-.032	-.147	1
	显著性（双侧）	.530	.847	.372	
	N	39	39	39	39

**. 在 .01 水平（双侧）上显著相关。

3.3-线性回归与相关

- 线性回归是分析两个定量变量间数量依存关系的统计分析方法。
- 回归分析主要包括三方面内容
  - 提供建立有相关关系的变量之间的数学关系式
  - 判别影响变量的众多变量中哪些影响是显著的
  - 利用所得到的经验公式进行预测和控制
- 例18：对某省9个地区水质的碘含量及其甲状腺肿的患病率作调查得到一组数据，见例18.sav，试进行回归分析

3.3-线性回归与相关


例18.sav [数据集1] - IBM SPSS Statistics 数据编辑器
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W) 帮助
报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
相关(C)
回归(R)
对数线性模型
神经网络
分类(F)
降维
度量(S)
非参数检验(N)
预测(T)
生存函数(S)
多重响应(U)
缺失值分析(Y)...
多重归因(T)
复杂抽样(L)
质量控制(Q)
ROC 曲线图(V)...
线性(L)...
地区 碘含量 变量
数据视图 变量视图
线性回归
因变量(D):
患病率 [患病率]
地区
碘含量 [碘含量]
块 1的1
上一张(V)
下一张(N)
自变量(I):
碘含量 [碘含量]
方法(M): 进入
选择变量(E):
规则(U)...
个案标签(C):
WLS 权重(H):
统计量(S)...
绘制(T)...
保存(S)...
选项(O)...
Bootstrap(B)...
确定 粘贴(P) 重置(R) 取消 帮助

3.3-线性回归与相关

模型汇总

模型	R	R 方	调整 R 方	标准 估计的误差
1	.971[a]	.943	.934	1.5747

a. 预测变量: (常量), 碘含量。

1、线性回归出来的相关系数为R=0.971。
2、方程拟合优度R方是0.943，调整后的R方为0.934
R方是对回归方程拟合情况的描述，R方是方程中变量X对Y的解释程度，越接近1，表明方程中X对Y的解释能力越强，拟合度越好。

3.3-线性回归与相关

Anova[b]

模型		平方和	df	均方	F	Sig.
1	回归	285.504	1	285.504	115.136	.000[a]
	残差	17.358	7	2.480		
	总计	302.862	8			

a. 预测变量: (常量), 碘含量。
b. 因变量: 患病率

在确认线性回归之前，必须判断变量的关系是否满足一元线性模型，即转换由**Y=a+bX=e，e服从正态分布，检验假设H0：b=0；H1：b≠0。**
F统计量P值=0<0.05，说明模型整体是显著的，具有统计学意义

3.3-线性回归与相关

系数[a]

模型		非标准化系数		标准系数	t	Sig.
		B	标准 误差	试用版		
1	(常量)	17.484	1.507		11.600	.000
	碘含量	4.459	.416	.971	10.730	.000

a. 因变量: 患病率

给出了常数项和回归系数。
患病率=17.484+4.459*碘含量

3.5-聚类分析

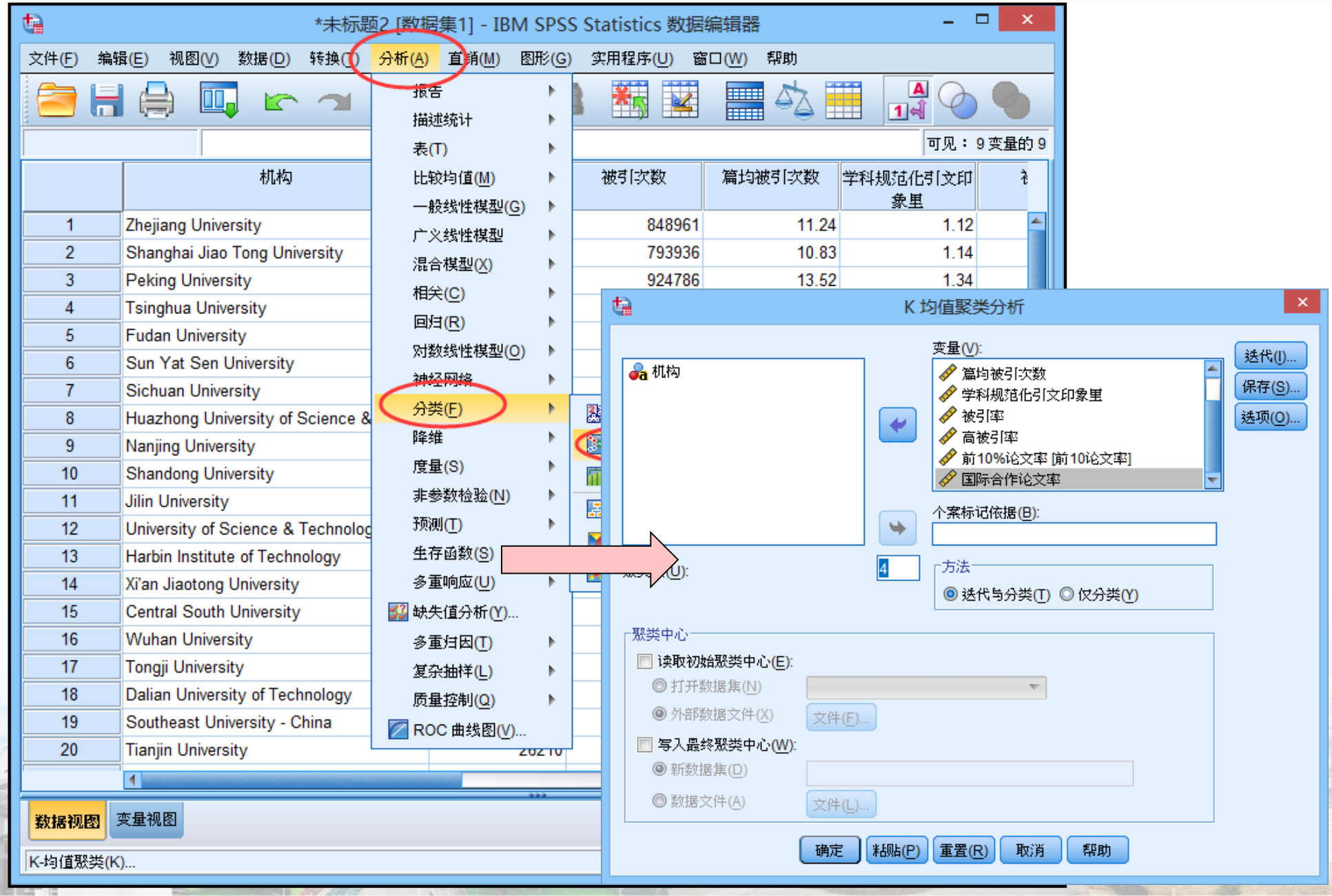
口聚类是根据某些数量特征将观察对象进行分类的一种数理统计方法

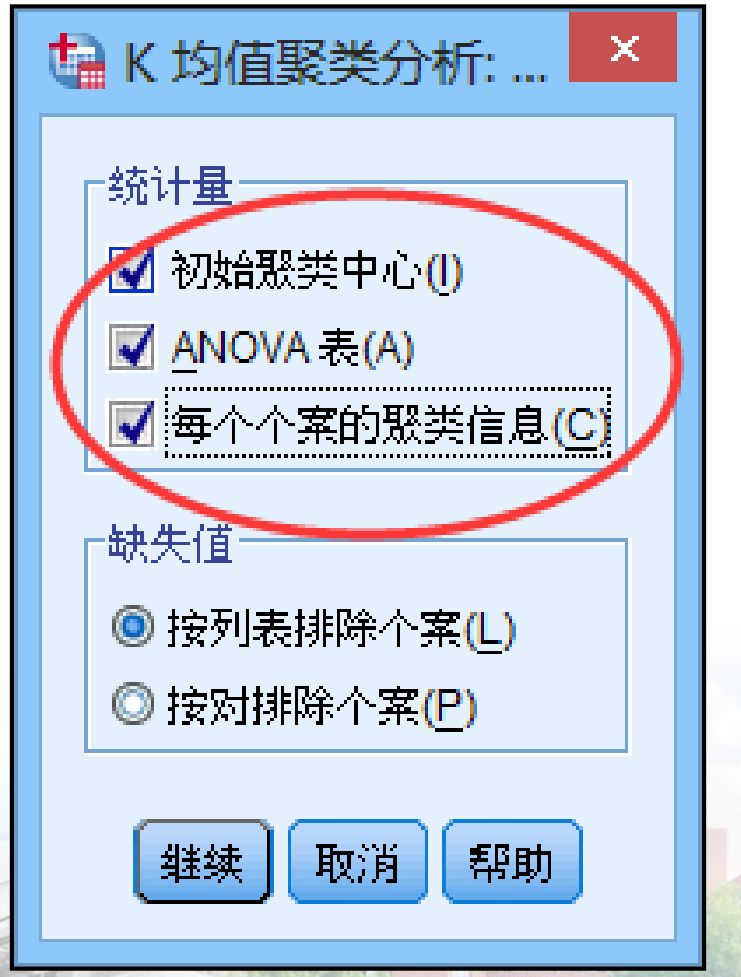
。

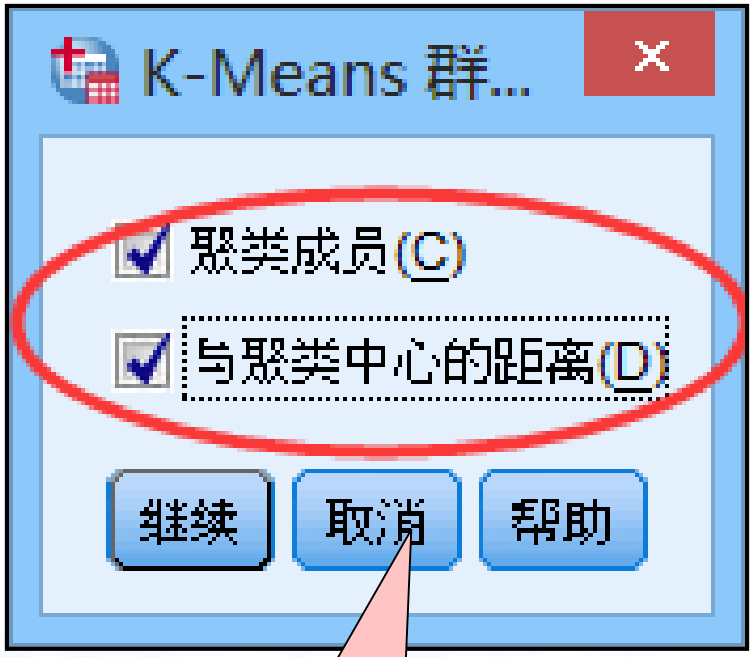
- 系统聚类法：首先将n个样品看成n类，然后将性质最接近的两类合并为一类，得到n-1类，然后再从这些类中找出性质最接近的两个类合并为n-2类，重复上述步骤，一直到所有样品聚为一类。整个过程可以绘成聚类图或树状图。
- 动态分类法：首先将样品粗糙分为n类，然后根据某种最优准则进行调整至不能调整为止。
- ……

3.4-聚类分析

- K-中心聚类：快速高效，特别是大量数据时，准确性高一些，但是需要指定聚类的类别数量。
- 例20：根据30所大学的各指标数据（例20.sav），将其分为4类。


*未标题2 [数据集1] - IBM SPSS Statistics 数据编辑器
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W) 帮助
可见：9 变量的 9
机构
被引次数
篇均被引次数
学科规范化引文印象里
1 Zhejiang University 848961 11.24 1.12
2 Shanghai Jiao Tong University 793936 10.83 1.14
3 Peking University 924786 13.52 1.34
4 Tsinghua University
5 Fudan University
6 Sun Yat Sen University
7 Sichuan University
8 Huazhong University of Science &
9 Nanjing University
10 Shandong University
11 Jilin University
12 University of Science & Technolog
13 Harbin Institute of Technology
14 Xi'an Jiaotong University
15 Central South University
16 Wuhan University
17 Tongji University
18 Dalian University of Technology
19 Southeast University - China
20 Tianjin University
报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
相关(C)
回归(R)
对数线性模型(O)
神经网络
分类(F)
降维
度量(S)
非参数检验(N)
预测(T)
生存函数(S)
多重响应(U)
缺失值分析(Y)...
多重归因(T)
复杂抽样(L)
质量控制(Q)
ROC 曲线图(V)...
数据视图
变量视图
K-均值聚类(K)...
K 均值聚类分析
机构
变量(V):
篇均被引次数
学科规范化引文印象里
被引率
高被引率
前10%论文率 [前10论文率]
国际合作论文率
迭代(I)...
保存(S)...
选项(O)...
个案标记依据(B):
4
方法
迭代与分类(T)
仅分类(Y)
聚类中心
读取初始聚类中心(E):
打开数据集(N)
外部数据文件(X)
文件(F)...
写入最终聚类中心(W):
新数据集(D)
数据文件(A)
文件(L)...
确定
粘贴(P)
重置(R)
取消
帮助





在数据中新加两列表示样本所属的类别以及与聚类中心的距离

口选取了其中4个样本作为初始聚类中心

初始聚类中心

	聚类			
	1	2	3	4
论文篇数	17248	21605	68393	37635
被引次数	127523	322748	924786	553011
篇均被引次数	7.39	14.94	13.52	14.69
学科规范化引文印象里	1.01	1.34	1.34	1.41
被引率	71.72	81.22	76.43	80.04
高被引率	.93	1.71	1.71	1.97
前10%论文率	9.49	14.84	13.13	15.33
国际合作论文率	24.15	22.47	34.89	30.32

重庆大学	南开大学	北京大学	中科大

口最终聚类中心

最终聚类中心

	聚类			
	1	2	3	4
论文篇数	18874	31416	70506	44832
被引次数	166265	314972	860243	575029
篇均被引次数	8.80	10.31	12.27	12.92
学科规范化引文印象里	1.12	1.14	1.25	1.28
被引率	74.25	75.89	77.06	77.07
高被引率	1.15	1.15	1.46	1.43
前10%论文率	10.06	10.92	12.10	12.95
国际合作论文率	27.65	23.61	29.87	28.27

口聚类结果

聚类成员

案例号	聚类	距离
1	3	12345.772
2	3	66366.096
3	3	64577.828
4	3	14238.975
5	4	97716.315
6	4	47894.580
7	2	84661.599
8	2	96682.683
9	4	27674.896
10	2	100442.996
11	2	60920.591
12	4	23163.764
13	2	33152.465
14	2	4773.266
15	2	34247.604
16	2	23737.988
17	2	63562.157
18	2	22757.203
19	2	53529.304
20	2	72385.791
21	2	31806.703

口需要注意

- 输入的是一个M*N的矩阵，但表示的是每个个案在每个指标上的具体表现。

- 共引矩阵表示的是每一个个案与其它个案的关系，与上述矩阵有本质区别，因此无法利用spss对共引矩阵进行聚类分析。

机构	论文篇数	被引次数	篇均被引	学科规范化引文影响力	被引率	高被引...	前10论文	国际合作
Zhejiang University	75520	848961	11.24	1.12	77.16	1.11	10.78	26.27
Shanghai Jiao Tong U...	73312	793936	10.83	1.14	75.60	1.08	10.33	27.76
Peking University	68393	924786	13.52	1.34	76.43	1.71	13.13	34.89
Tsinghua University	64799	873288	13.48	1.40	79.04	1.92	14.15	30.55
Fudan University	52493	672444	12.81	1.25	75.21	1.31	12.17	29.35
Sun Yat Sen University	46977	527182	11.22	1.20	73.28	1.07	11.46	26.27
Sichuan University	45649	398429	8.73	.91	73.52	.68	7.92	18.64
Huazhong University ...	42728	410991	9.62	1.17	75.37	1.09	11.05	24.97
Nanjing University	42221	547477	12.97	1.26	79.73	1.39	12.85	27.14
Shandong University	41459	414912	10.01	1.04	76.33	.82	9.59	23.46
Jilin University	39048	375413	9.61	.96	73.41	.76	9.24	20.21
University of Science ...	37635	553011	14.69	1.41	80.04	1.97	15.33	30.32
Harbin Institute of Tec...	36374	347752	9.56	1.14	75.85	1.35	10.28	23.96
Xi'an Jiaotong University	36141	315652	8.73	1.06	73.53	1.04	9.40	25.53
Central South University	33846	280811	8.30	1.07	73.30	1.00	9.09	22.51
Wuhan University	31297	338710	10.82	1.19	75.45	1.20	11.57	24.56
Tongji University	29312	251445	8.58	1.09	73.70	.92	10.12	28.05
Dalian University of Te...	26889	292670	10.88	1.10	77.66	1.06	10.67	23.57
Southeast University -...	26782	261644	9.77	1.18	73.28	1.38	11.50	24.79
Tianjin University	26210	242774	9.26	1.10	75.07	1.08	10.40	21.22
South China Universit...	23808	284089	11.93	1.49	78.13	1.64	14.16	21.57
Beihang University	22277	170316	7.65	1.06	72.64	1.17	9.13	24.46
Nankai University	21605	322748	14.94	1.34	81.22	1.71	14.84	22.47
Xiamen University	21563	258461	11.99	1.23	77.12	1.53	12.76	30.97
Lanzhou University	19951	243056	12.18	1.16	81.30	1.16	12.16	21.34
Beijing Normal Univer...	19793	212707	10.75	1.22	77.74	1.36	11.18	34.62
China Agricultural Uni...	19050	206523	10.84	1.19	78.29	1.11	11.18	31.80
University of Electroni...	18137	130941	7.22	1.12	71.63	1.05	9.16	28.73
Chongqing University	17248	127523	7.39	1.01	71.72	.93	9.49	24.15
Beijing Institute of Tec...	16736	149577	8.94	1.10	73.49	1.28	10.24	22.13

表 6.3　文献共引矩阵（片段）

	A1	A2	A3	A4	A5	A6	A7	A8	A9
A1	19	0	2	1	0	0	0	0	0
A2	0	14	1	0	0	0	0	0	0
A3	2	1	20	6	0	0	4	1	1
A4	1	0	6	14	0	0	4	1	1
A5	0	0	0	0	19	0	0	0	1
A6	0	0	0	0	0	14	0	4	0
A7	0	0	4	4	0	0	15	0	0
A8	0	0	1	1	0	4	0	14	0
A9	0	0	1	1	1	0	0	0	13

3.5-因子分析

口做一个形象的比喻，在观看电影时或观看以后，我们能够说出电影是否精彩，这是判别分析；并且我们会迅速地将对电影的印象形成两类：精彩或不精彩，把现在看的这部归入到某一类中，这是聚类分析；我们之所以可以认为这部电影精彩，是因为它具有精彩的影视作品所具有的一些共同特点：演员的演技、画面制作精良、讲述的故事有趣，等等。这种从研究对象中寻找公共因子的办法就是因子分析。

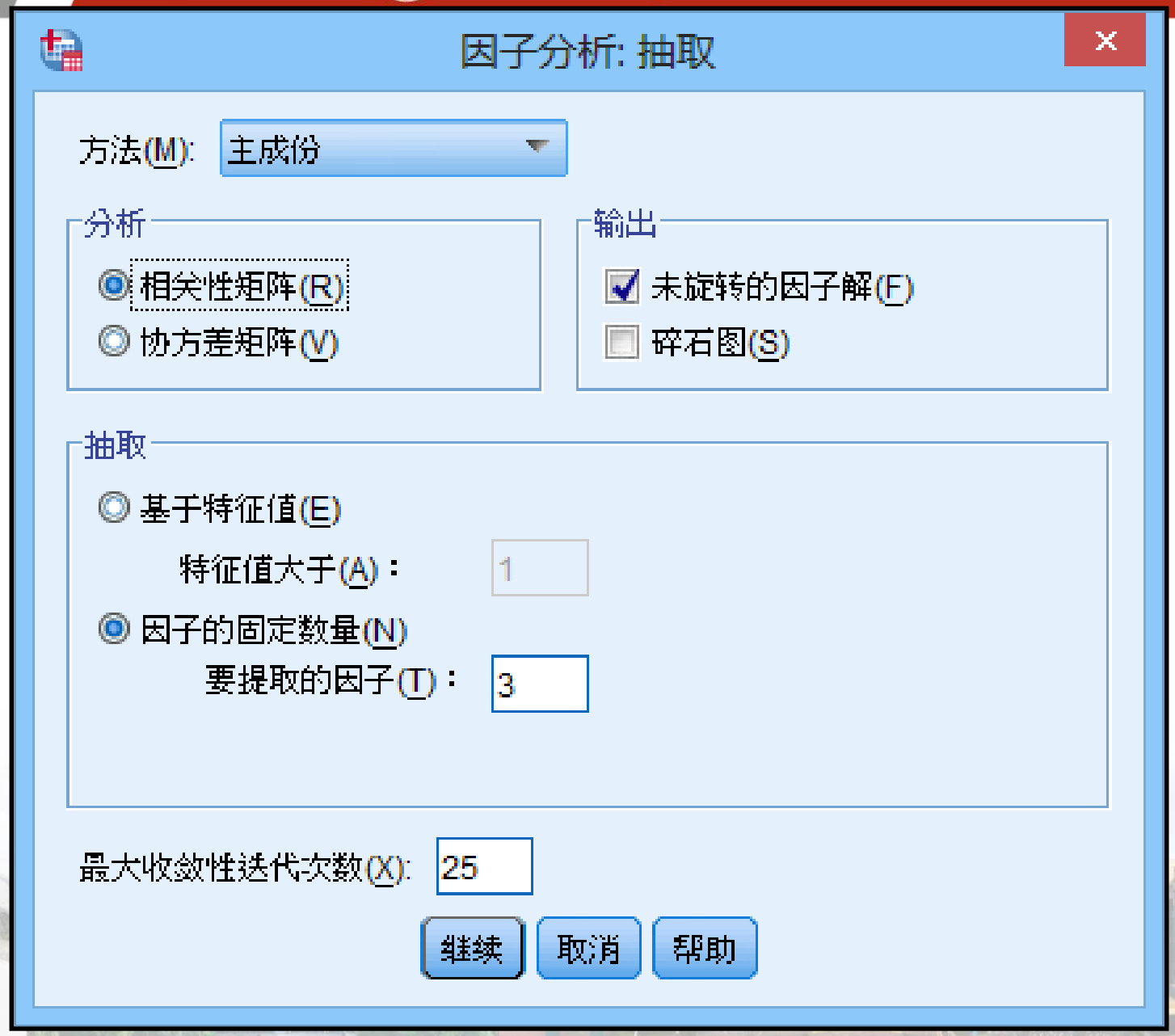
3.5-因子分析

- 简单地说，因子分析就是将原始变量分解成几个公共因子，在每个公共因子上有载荷的体现。如果一些原始变量在同一个公共因子上都具有较高的载荷，那么则说明这些原始变量有共同的内在（公共因子）。
- 例19：根据例19.sav提供的指标，设计一个具有3个一级指标的指标评价体系。

3.5-因子分析


例19.sav [数据集1] - IBM SPSS Statistics 数据编辑器
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W) 帮助
可见：8 变量的 8
学校
Shanghai Jiao Tong University
Zhejiang University
Peking University
Tsinghua University
Fudan University
Sun Yat Sen University
Sichuan University
Shandong University
Huazhong University of Scie...
Nanjing University
Jilin University
Xi'an Jiaotong University
Central South University
Harbin Institute of Technology
University of Science & Tec...
Tongji University
Wuhan University
热点论文率
高被引论文率
报告
描述统计
表(T)
比较均值(M)
一般线性模型(G)
广义线性模型
混合模型(X)
相关(C)
回归(R)
对数线性模型(O)
神经网络
分类(F)
降维
度量(S)
非参数检验(N)
预测(T)
生存函数(S)
多重响应(U)
缺失值分析(Y)...
多重归因(T)
复杂抽样(L)
质量控制(Q)
ROC 曲线图(V)...
因子分析(F)...
对应分析(C)...
最优尺度(O)...
数据视图
变量视图
因子分析(F)...
IBM SPSS Statistics Processor 就绪

四川大学图书馆
淡泊明志 宁静致远
志于所学 继续求学



3.5-因子分析

口结果显示

成份矩阵[a]

	成份		
	1	2	3
论文数量	.508	.854	.087
被引次数	.662	.743	.034
篇均被引次数	.943	-.198	-.069
被引率	.714	-.160	-.539
热点论文率	.573	-.230	.721
高被引论文率	.837	-.365	.224
前10%论文率	.916	-.204	-.238

总结

- 理解数理统计的基本工具方法是关键
- 对输入数据的要求和对处理结果的分析解释是使用SPSS的主要工作
- 建议在掌握SPSS的同时学会一门程序语言

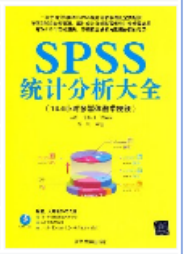
口 课件部分内容及数据来自《SPSS统计分析大全》（武松，潘发明等编著，清华大学出版社）

- 索书号C819/1348，文理馆
- 随书光盘（包括教学视频和源数据）可以下载。

第 1 条记录(共 1 条)

标准格式 | SFX

系统号- 图书	000956148
ISBN	978-7-302-34789-7 : CNY69.80 (含光盘)
	978-7-89414-781-3 (光盘)
作品语种	chi
题名	SPSS统计分析大全 / 武松, 潘发明等编著
出版发行	北京 : 清华大学出版社, 2014
随书光盘	随书光盘下载
内容简介	《SPSS统计分析大全》由浅入深，全面、系统地介绍了SPSS19.0的应用。《SPSS统计分析大全》涉及面广，从软件基本操作到高级统计分析技术，几乎涉及SPSS目前的绝大部分应用范畴。书中提供了大量应用案例，供读者实战演练。另外，《SPSS统计分析大全》配1张DVD光盘，收录了作者为本书录制的16小时配套高清教学视频及书中所有案例的数据文件。《SPSS统计分析大全》共30章，分为3篇。第1篇为 (更多)
目录:	第1篇 SPSS软件基础篇 第1章 SPSS19.0概述 1.1 SPSS19.0简介



感谢大家参与！