数据统计分析软件SPSS入门

主要内容

- 1-SPSS概述
- 2-数据管理
- 3-统计分析与统计图

主要内容

- 1-SPSS概述
- 2-数据管理
- 3-统计分析与统计图

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

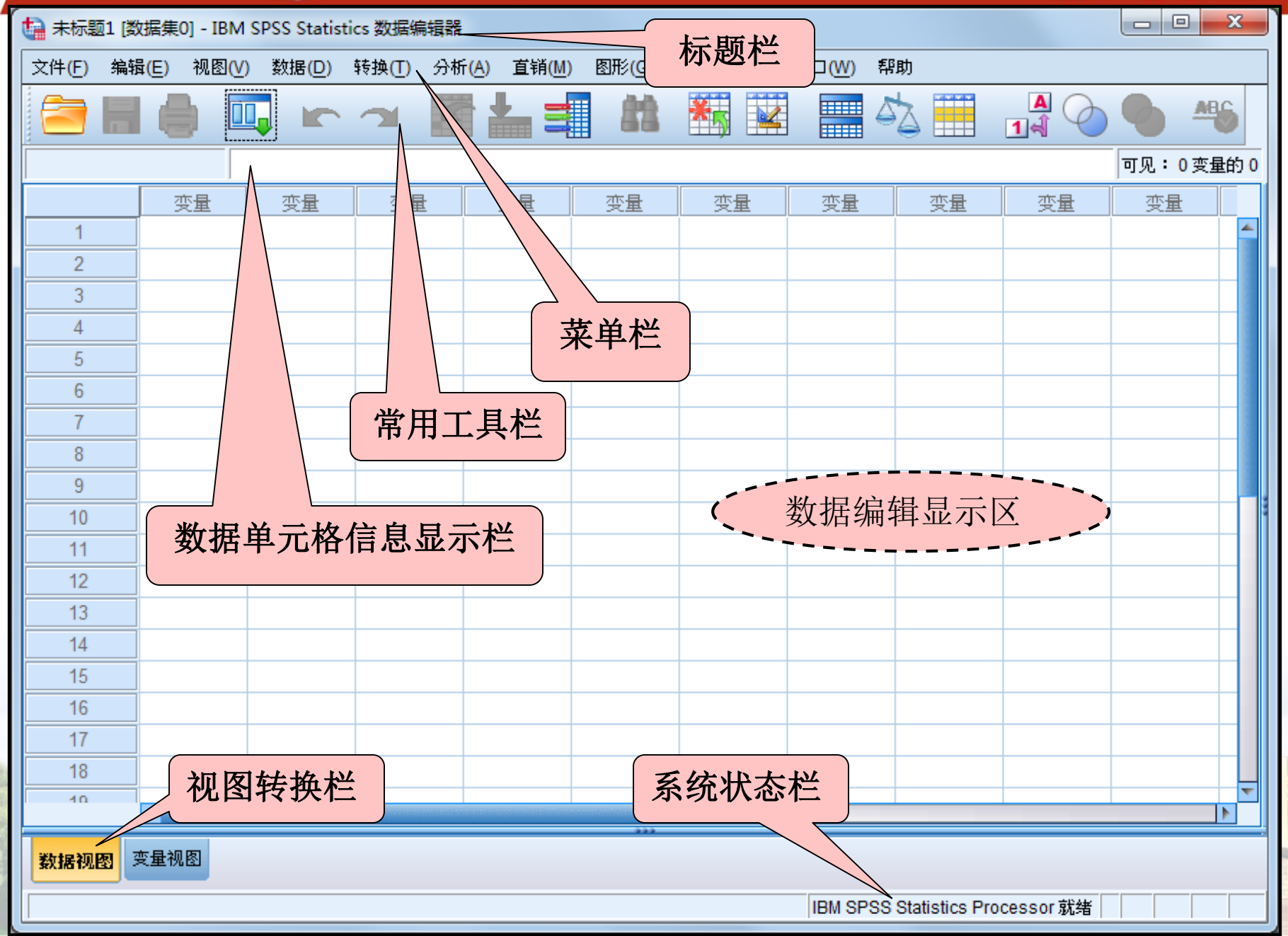
口启动

打开一个已存在的（历史记录）数据源

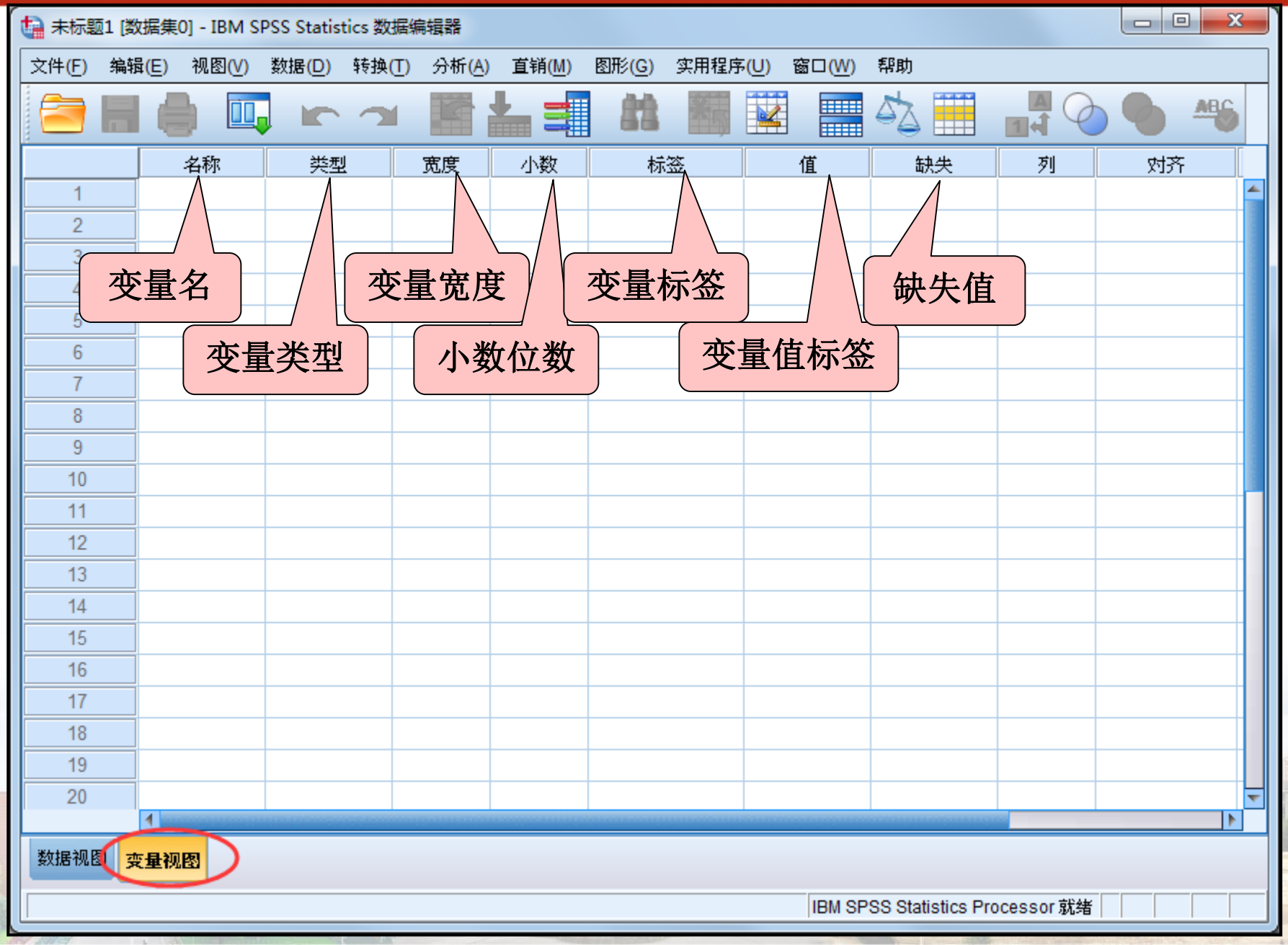
1.2-窗口介绍

- 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

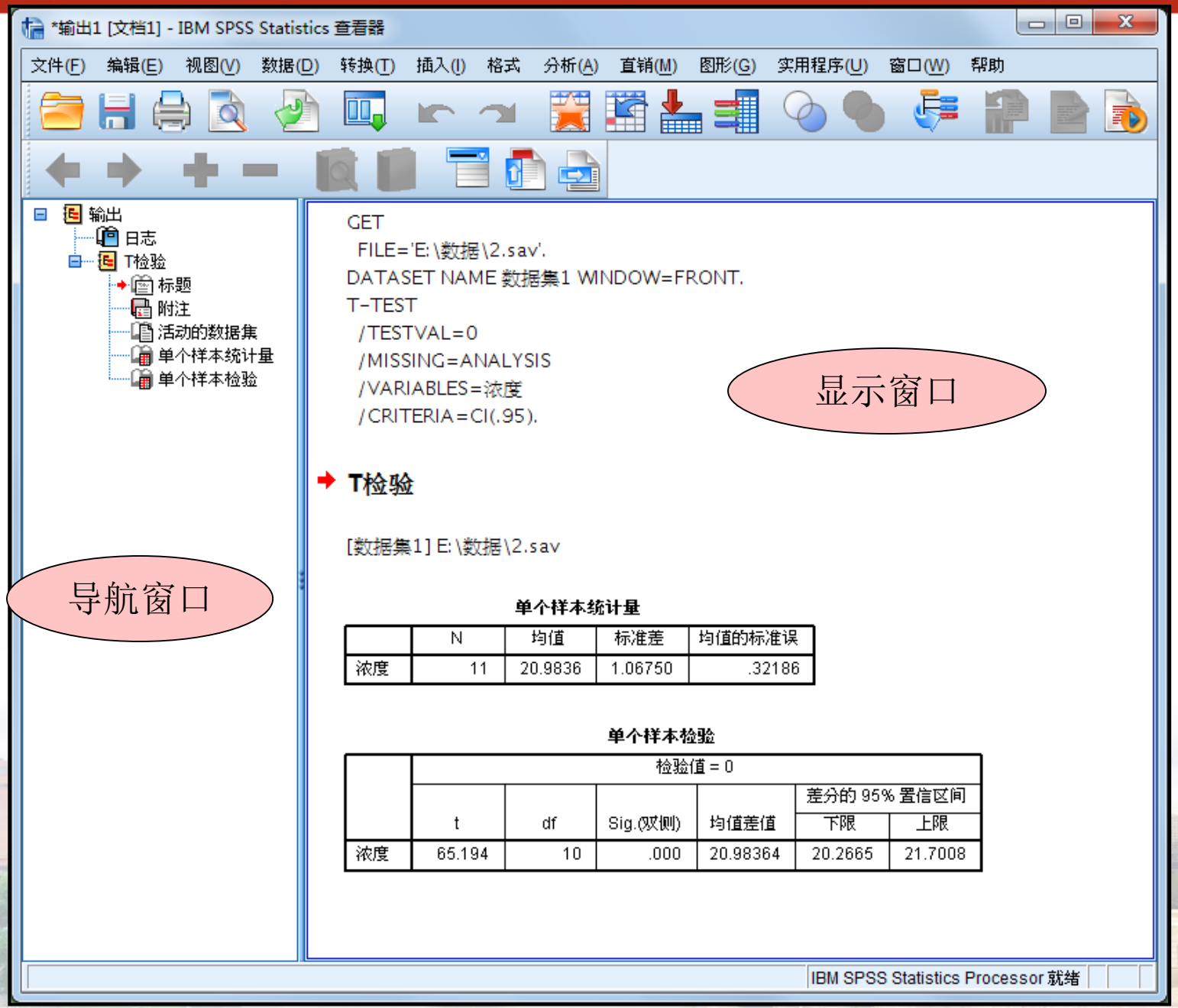
数据窗口


未标题1 [数据集0] - IBM SPSS Statistics 数据编辑器
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
可见：0 变量的 0
变量
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
数据视图
变量视图
IBM SPSS Statistics Processor 就绪
标题栏
菜单栏
常用工具栏
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

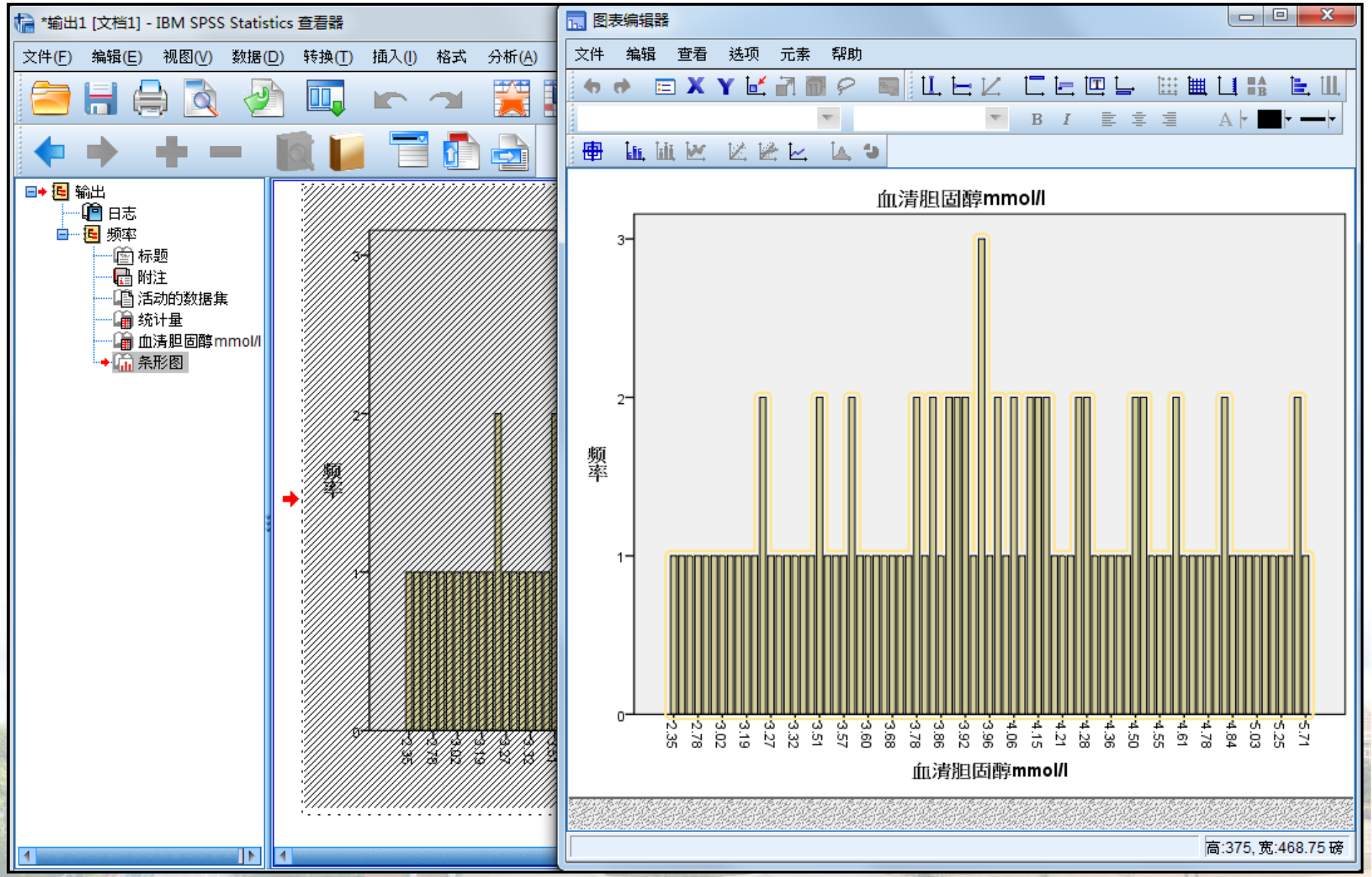
变量窗口


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

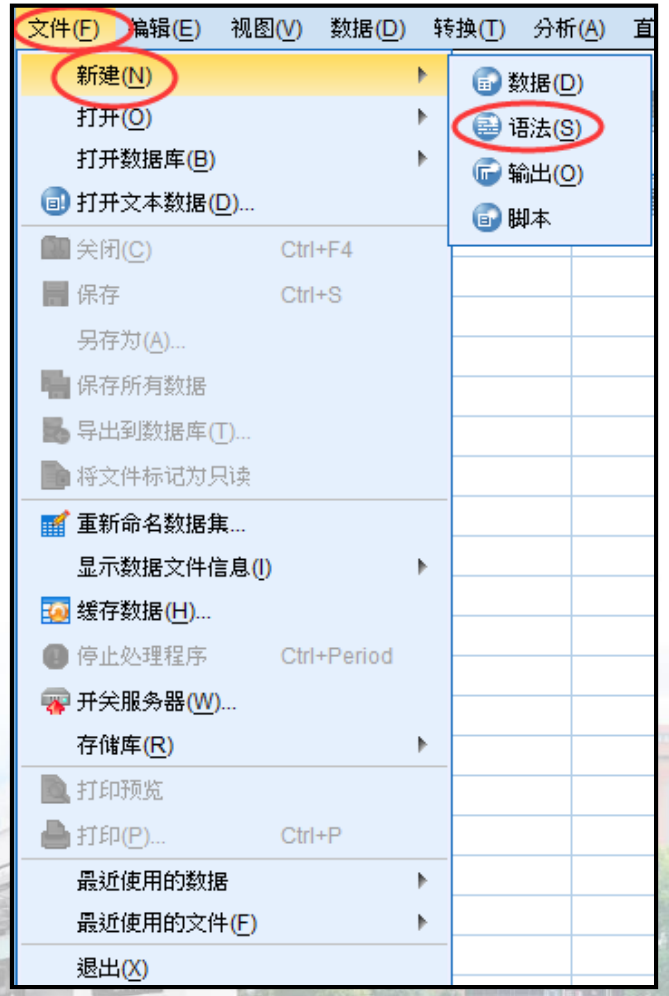

*输出1 [文档1] - IBM SPSS Statistics 查看器
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 插入(I) 格式 分析(A) 直销(M) 图形(G) 实用程序(U) 窗口(W) 帮助
输出
日志
T检验
标题
附注
活动的数据集
单个样本统计量
单个样本检验
导航窗口
GET
FILE='E:\数据\2.sav'.
DATASET NAME 数据集1 WINDOW=FRONT.
T-TEST
/TESTVAL=0
/MISSING=ANALYSIS
/VARIABLES=浓度
/CRITERIA=CI(.95).
显示窗口
T检验
[数据集1] E:\数据\2.sav
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

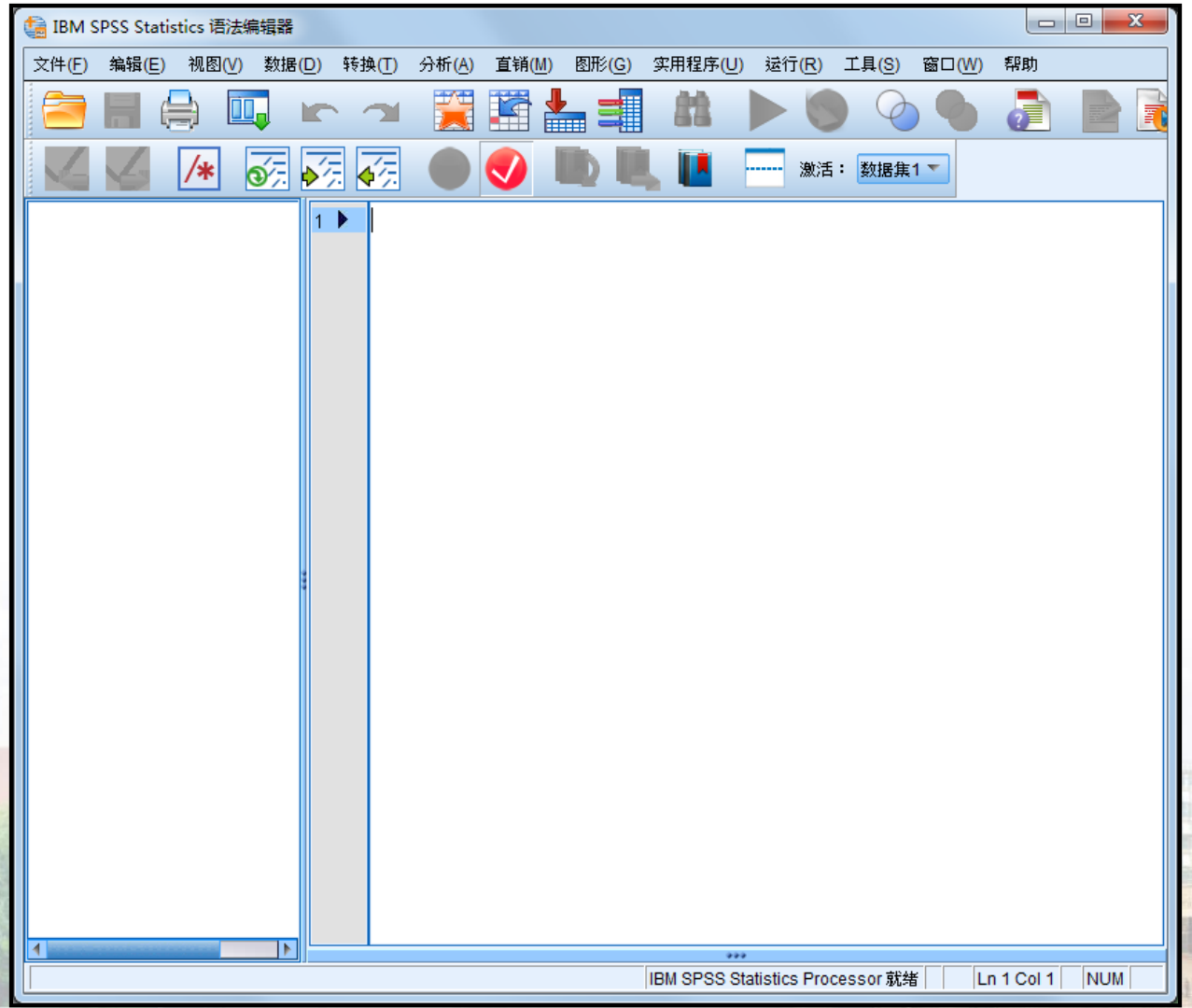
图表编辑窗口


*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

口关于度量标准的理解可参考：

- https://mp.weixin.qq.com/s?__biz=Mzg3MTY5NTE3MA==&mid=2247489403&idx=1&sn=fe56306b0299898a2bc10c298f9853a2&chksm=cefbc156f98c48401875d22f871cb989cacaf25154b4dd83be2fb756b1322c47ed51c668fde2&scene=21#wechat_redirect

口例2-2

- 按例2-2的数据，录入到SPSS中

主要内容

- 1-SPSS概述
- 2-数据管理
- 3-统计分析与统计图

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

口使用场景

- 有来自不同的网站的多个数据文件，格式可能一样、也可能不一样，需要合并成一个统一的数据文件

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

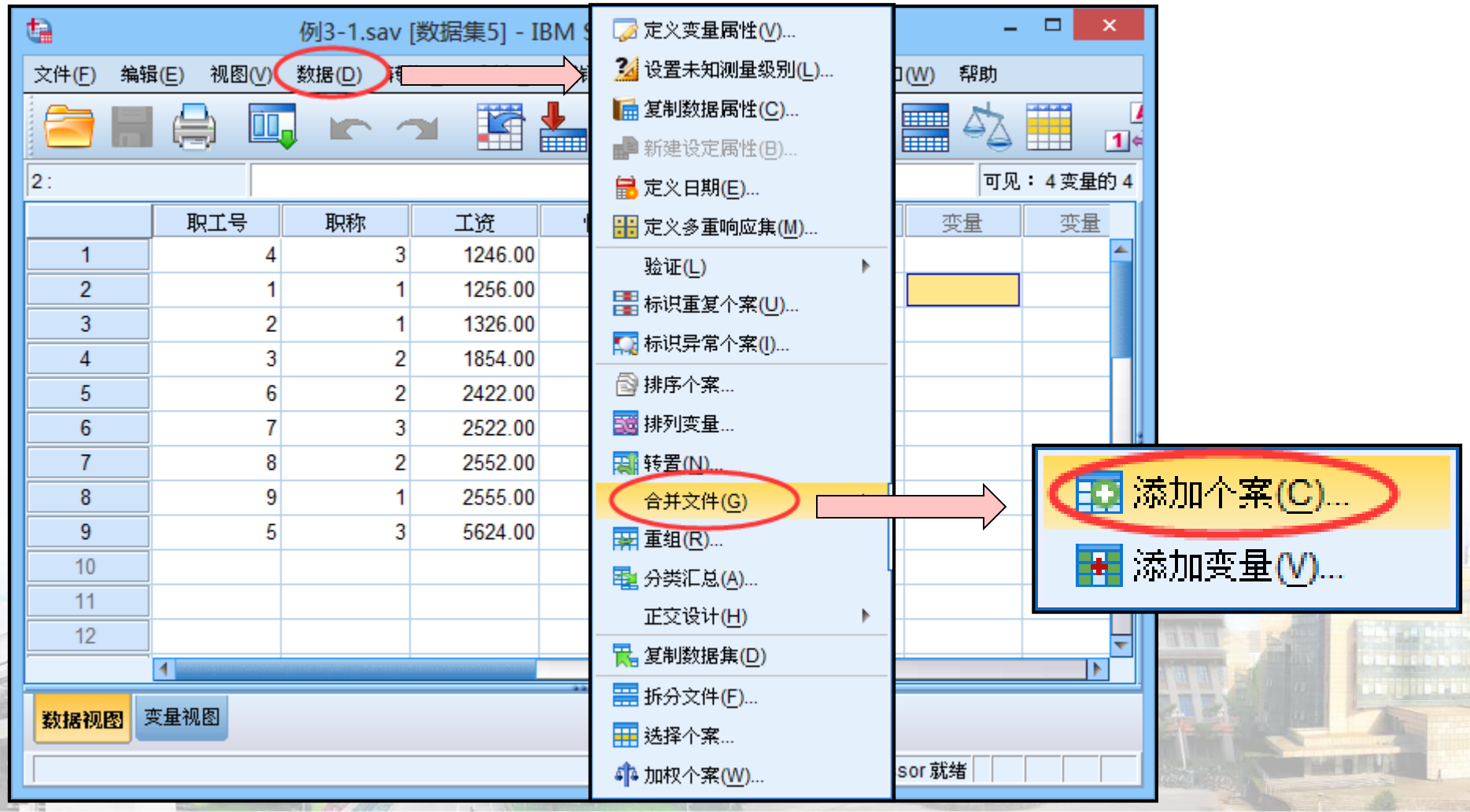
数据合并

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

口变量顺序是否需要一致?

- 将例3-1. sav与例3-3. sav数据进行合并

口变量数量是否需要一致?

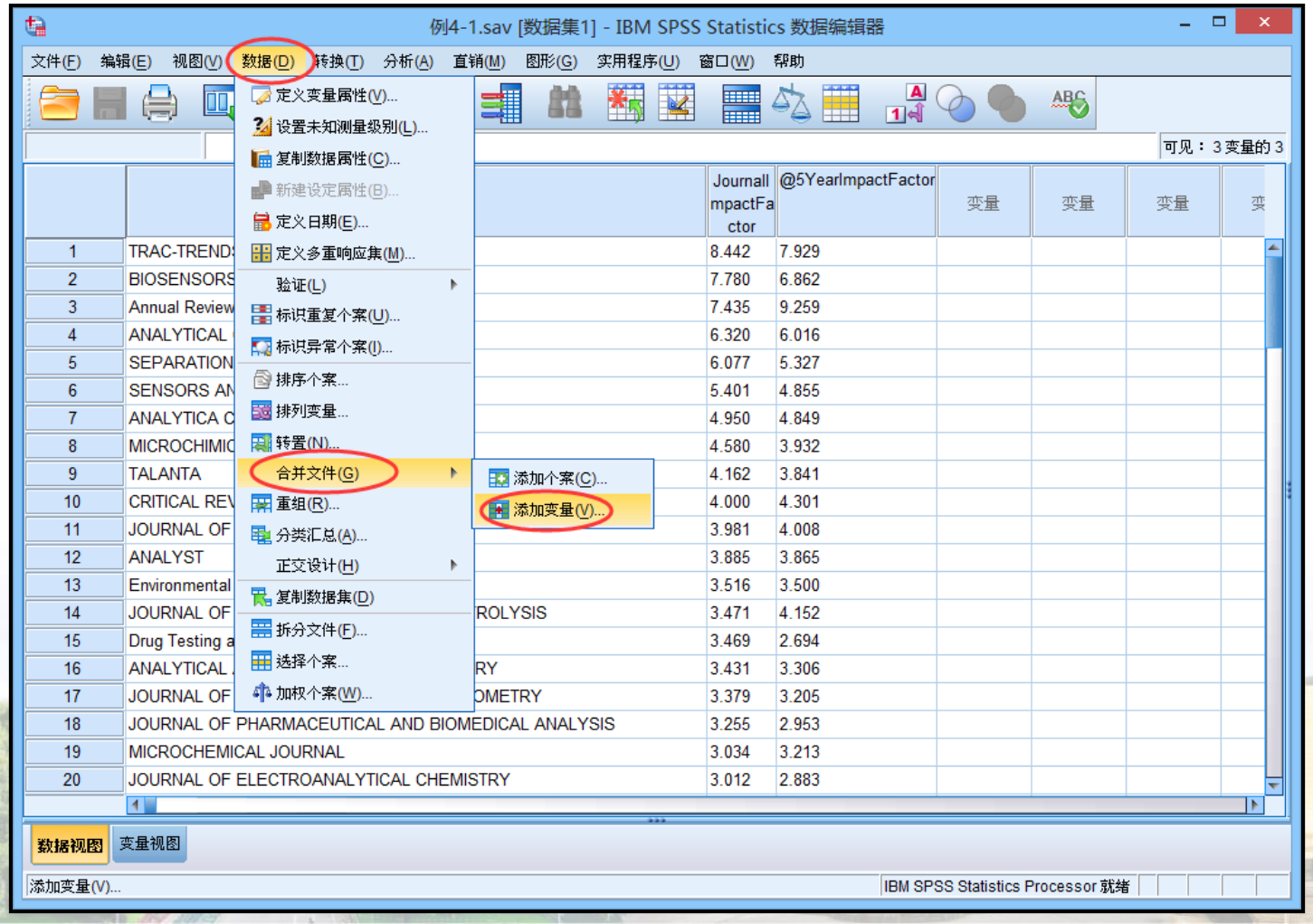
- 将例3-1. sav与例3-4. sav数据进行合并

口变量名称是否需要一致?

- 将例3-1. sav与例3-5. sav数据进行合并

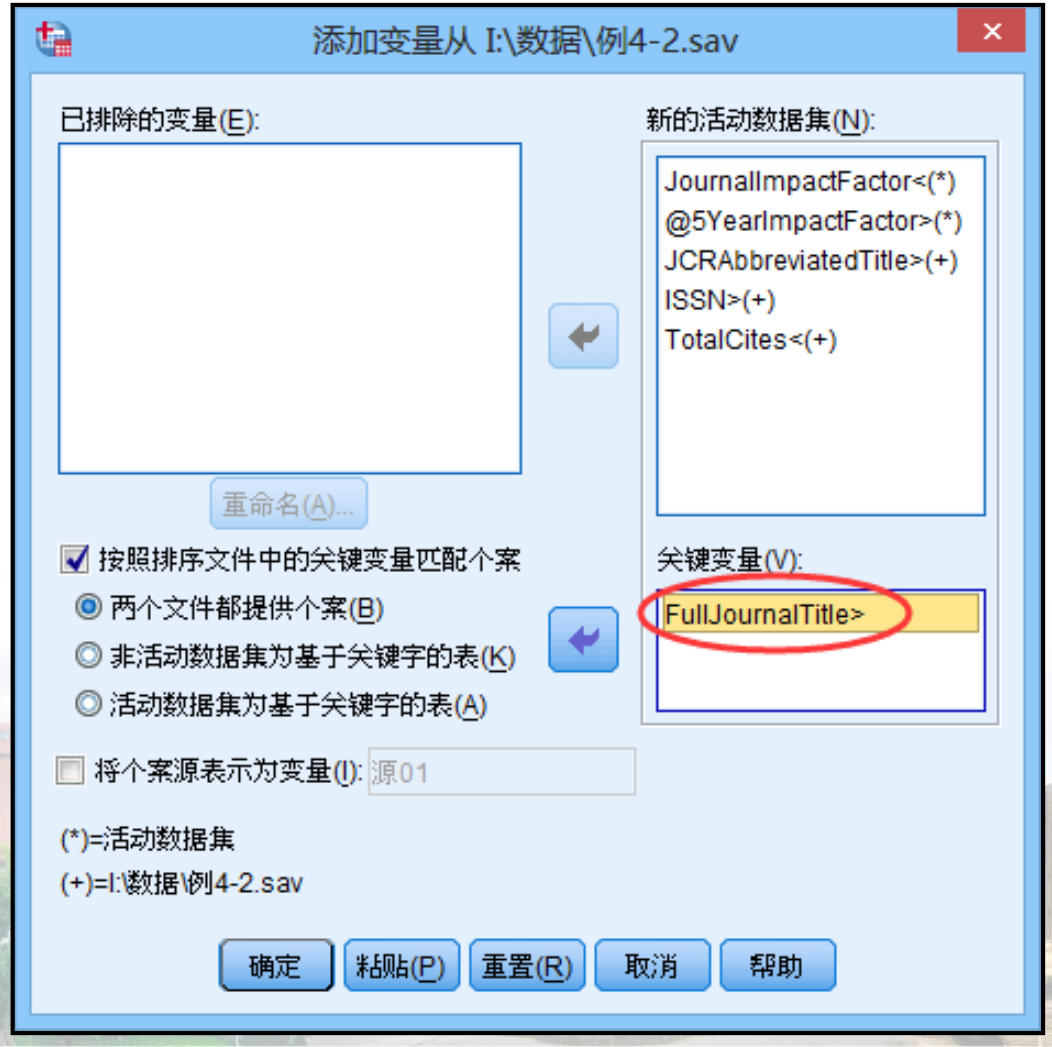
数据合并

口例4：将例4-1.sav与例4-2.sav数据进行合并（横向合并）



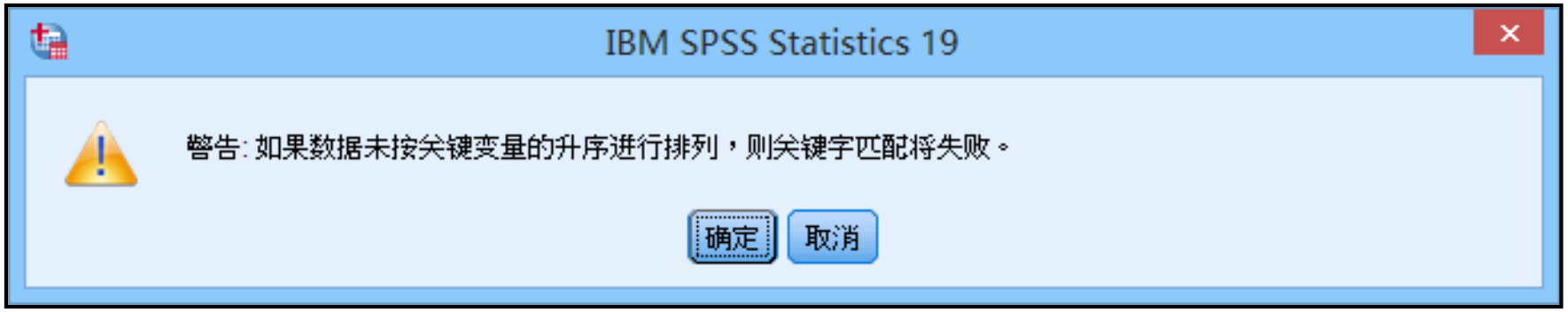
数据合并

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

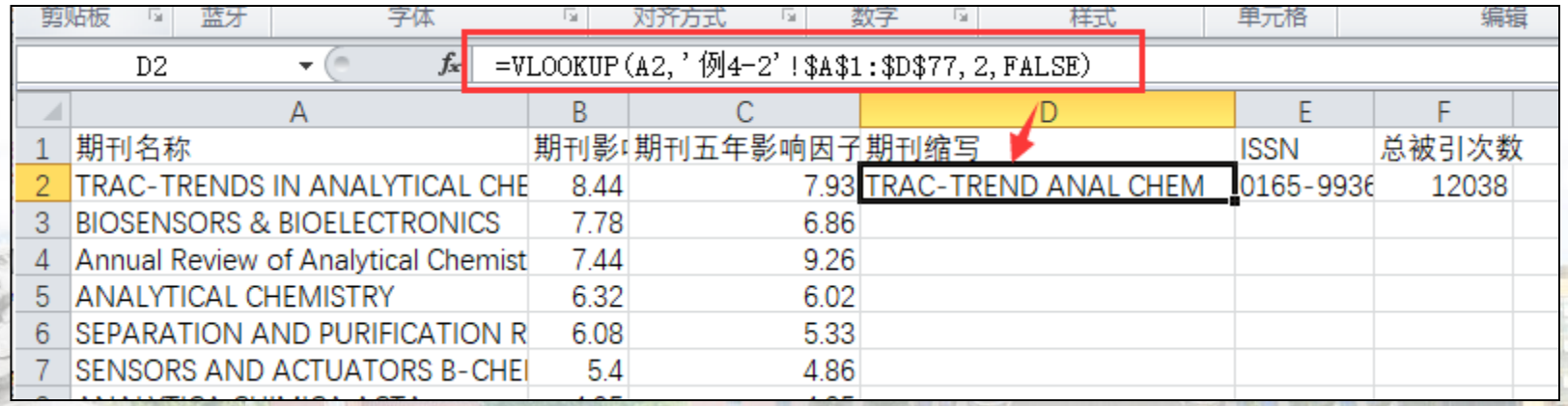
口合并前还需要将关键变量进行升序排序



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

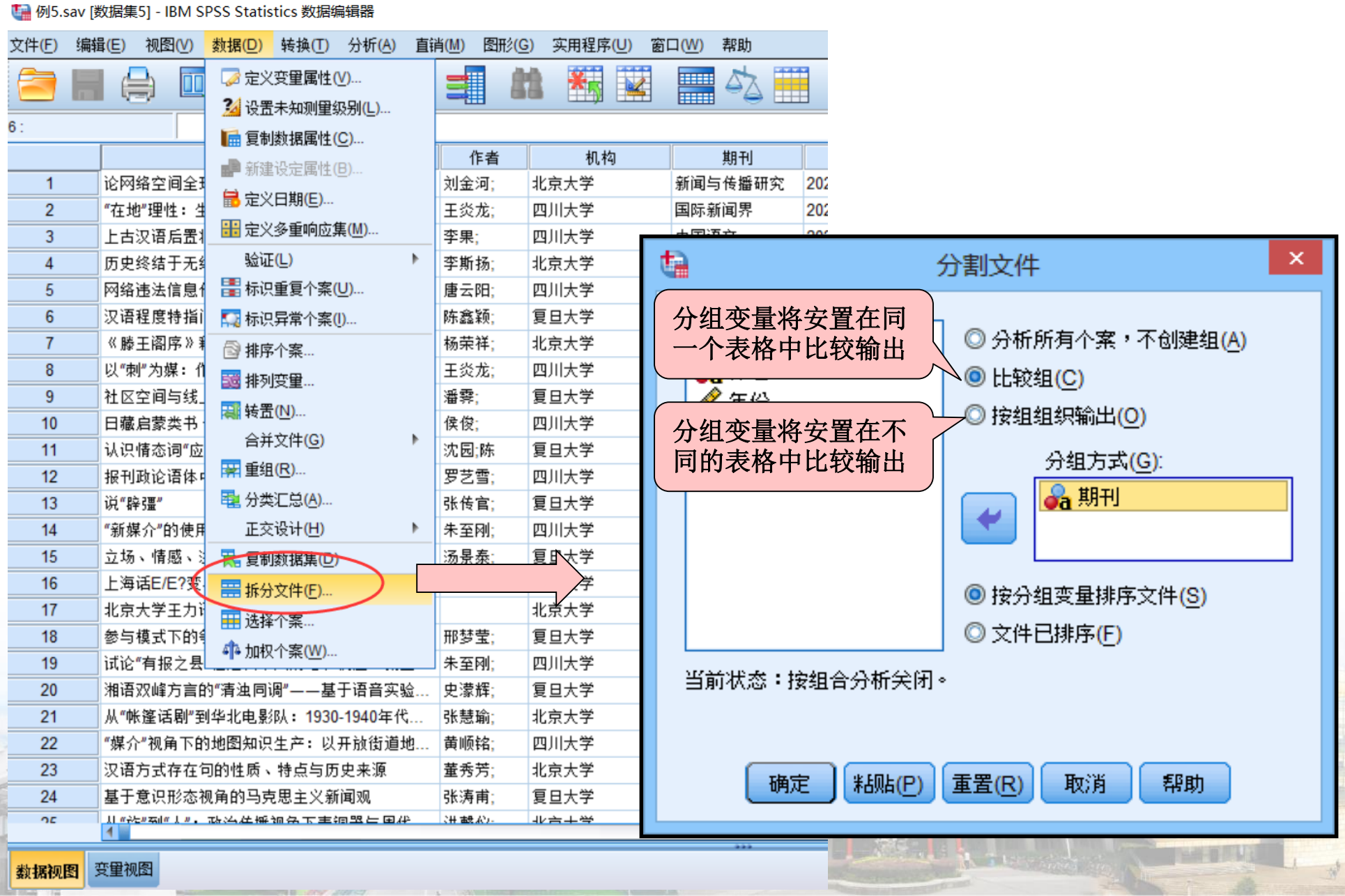


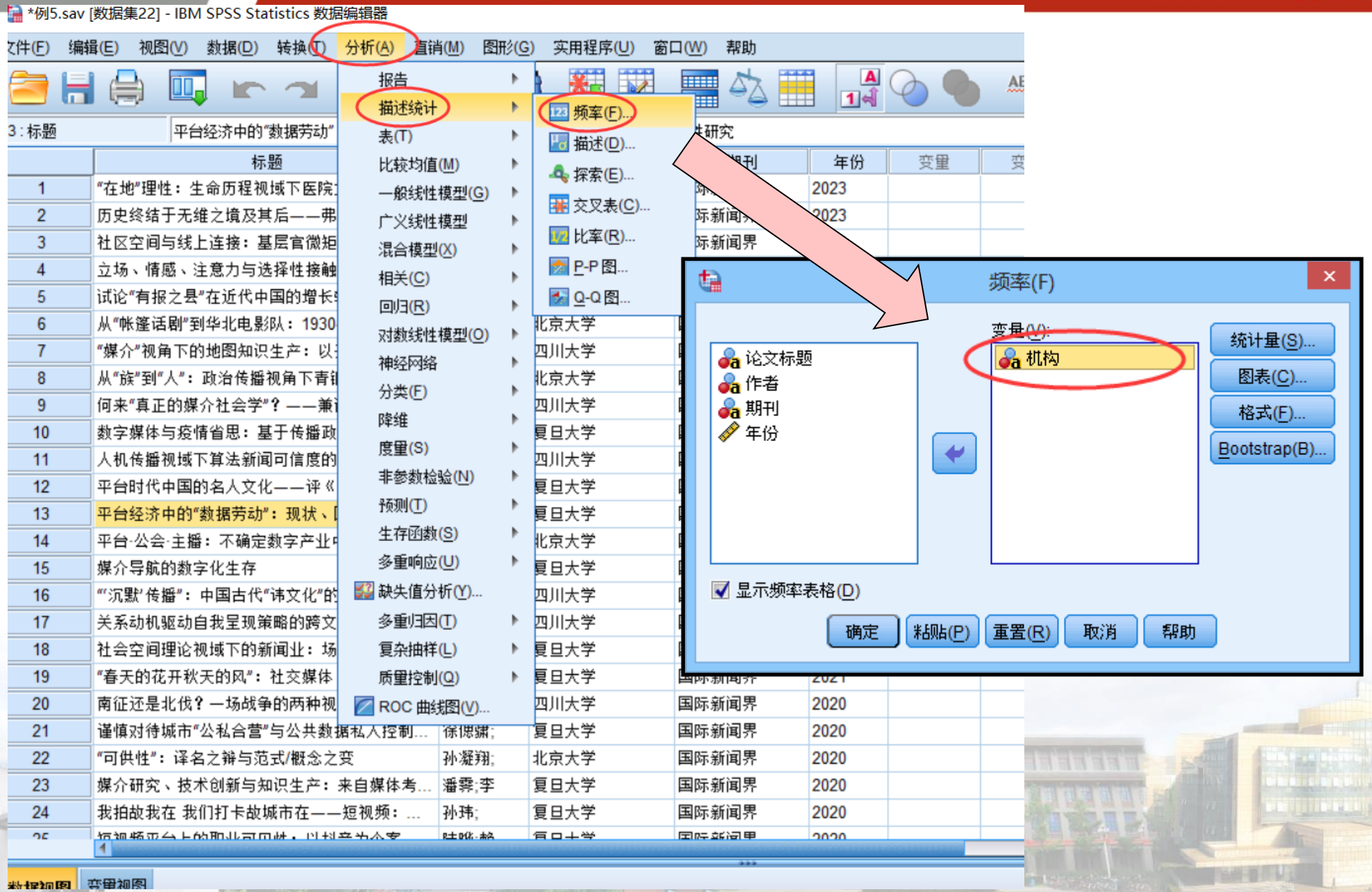
	A	B	C	D	E	F
1	期刊名称	期刊影响	期刊五年影响因子	期刊缩写	ISSN	总被引次数
2	TRAC-TRENDS IN ANALYTICAL CHE	8.44	7.93	TRAC-TREND ANAL CHEM	0165-9936	12038
3	BIOSENSORS & BIOELECTRONICS	7.78	6.86			
4	Annual Review of Analytical Chemist	7.44	9.26			
5	ANALYTICAL CHEMISTRY	6.32	6.02			
6	SEPARATION AND PURIFICATION R	6.08	5.33			
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：2018-2023北京大学、复旦大学和四川大学在中国语文、国际新闻界、新闻与传播研究上发表的论文信息（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5] - 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)...
作者
机构
期刊
刘金河;
北京大学
新闻与传播研究
王炎龙;
四川大学
国际新闻界
李果;
四川大学
李斯扬;
北京大学
唐云阳;
四川大学
陈鑫颖;
复旦大学
杨荣祥;
北京大学
王炎龙;
四川大学
潘霁;
复旦大学
侯俊;
四川大学
沈园;陈
复旦大学
罗艺雪;
四川大学
张传官;
复旦大学
朱至刚;
四川大学
汤景泰;
邢梦莹;
复旦大学
朱至刚;
四川大学
史濛辉;
复旦大学
张慧瑜;
北京大学
黄顺铭;
四川大学
董秀芳;
北京大学
张涛甫;
复旦大学
数据视图
变量视图
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22] - 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...
Q-Q 图...
3 : 标题
平台经济中的"数据劳动"
标题
年份
变量
数据视图
变量视图
频率(F)
论文标题
作者
期刊
年份
变量(V):
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

口输出结果

机构

期刊			频率	百分比	有效百分比	累积百分比
国际新闻界	有效	北京大学	9	22.0	22.0	22.0
		复旦大学	18	43.9	43.9	65.9
		四川大学	14	34.1	34.1	100.0
		合计	41	100.0	100.0	
新闻与传播研究	有效	北京大学	6	16.7	16.7	16.7
		复旦大学	18	50.0	50.0	66.7
		四川大学	12	33.3	33.3	100.0
		合计	36	100.0	100.0	
中国语文	有效	北京大学	18	41.9	41.9	41.9
		复旦大学	20	46.5	46.5	88.4
		四川大学	5	11.6	11.6	100.0
		合计	43	100.0	100.0	

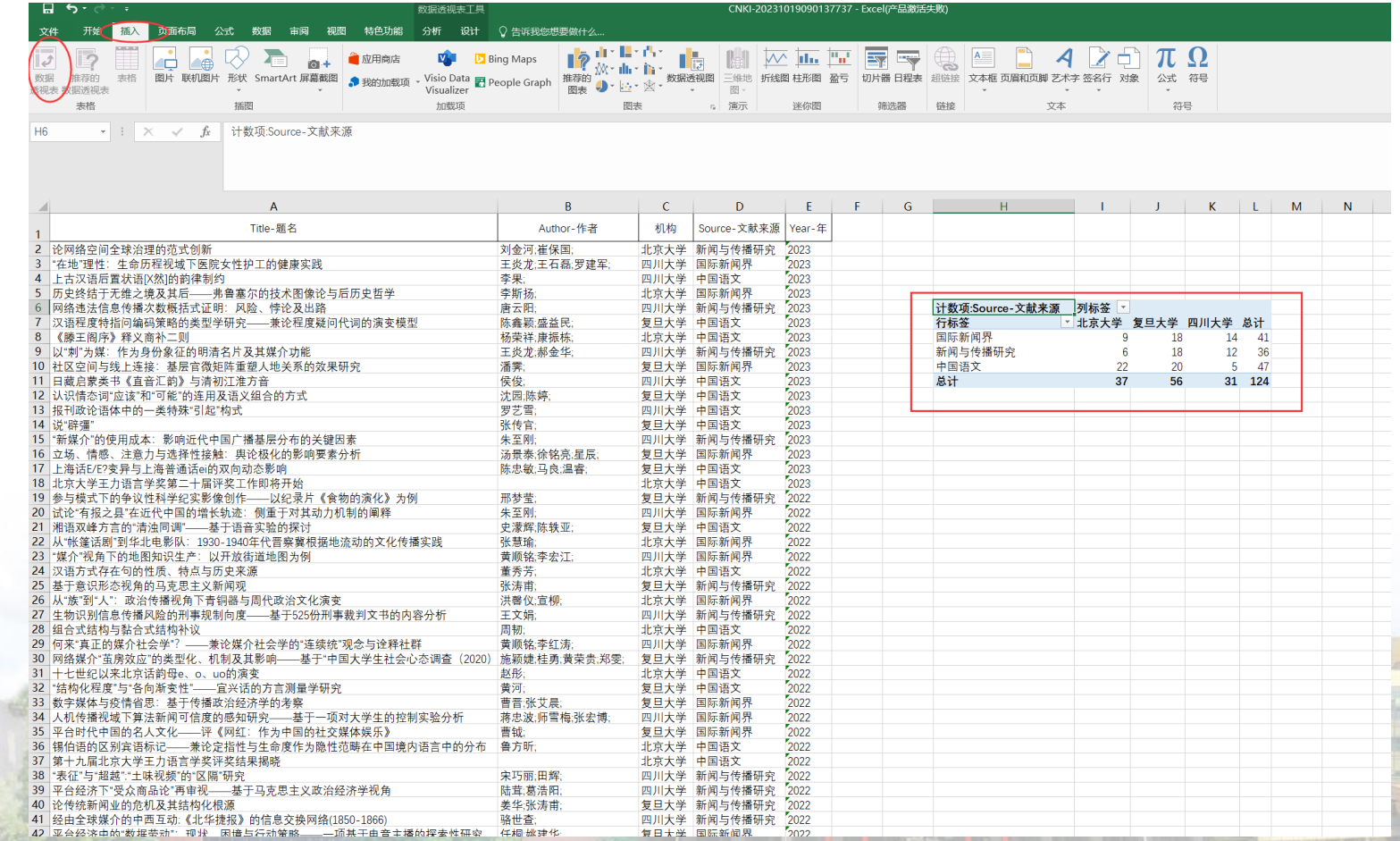
口也可以按机构、期刊进行拆分，再按年份进行频率统计，查看每个机构在各期刊上的年度发文趋势

四川大学	国际新闻界	有效	2018	6	21.4	21.4	21.4
			2019	4	14.3	14.3	35.7
			2020	4	14.3	14.3	50.0
			2021	4	14.3	14.3	64.3
			2022	8	28.6	28.6	92.9
			2023	2	7.1	7.1	100.0
			合计	28	100.0	100.0	
	新闻与传播研究	有效	2018	4	16.7	16.7	16.7
			2020	2	8.3	8.3	25.0
			2021	4	16.7	16.7	41.7
			2022	8	33.3	33.3	75.0
			2023	6	25.0	25.0	100.0
			合计	24	100.0	100.0	
	中国语文	有效	2020	2	20.0	20.0	20.0
			2021	2	20.0	20.0	40.0
			2023	6	60.0	60.0	100.0
			合计	10	100.0	100.0	

口按机构分组，查看每个机构下作者的出现频率

口来看看在Excel中如何处理这类问题

- 透视表



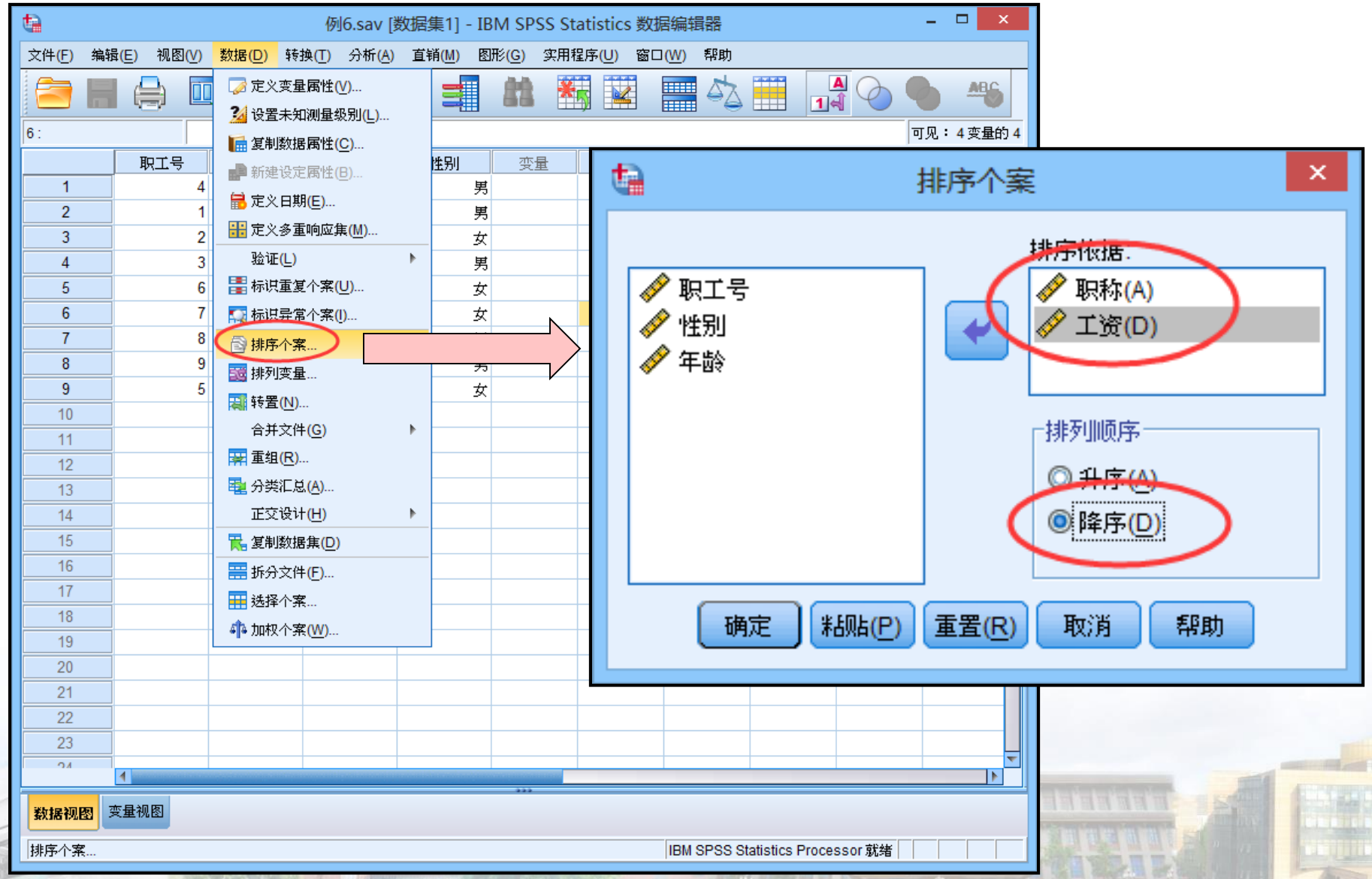
	A	B	C	D	E
1	Title-题名	Author-作者	机构	Source-文献来源	Year-年
2	论网络空间全球治理的范式创新	刘金河;崔保国;	北京大学	新闻与传播研究	2023
3	“在地”理性：生命历程视域下医院女性护工的健康实践	王炎龙;王石磊;罗建军;	四川大学	国际新闻界	2023
4	上古汉语后置状语[X然]的韵律制约	李果;	四川大学	中国语文	2023
5	历史终结于无维之境及其后——弗鲁塞尔的技术图像论与后历史哲学	李斯扬;	北京大学	国际新闻界	2023
6	网络违法信息传播次数概括式证明：风险、悖论及出路	唐云阳;	四川大学	新闻与传播研究	2023
7	汉语程度特指问编码策略的类型学研究——兼论程度疑问代词的演变模型	陈鑫颖;盛益民;	复旦大学	中国语文	2023
8	《滕王阁序》释义商补二则	杨荣祥;康振栋;	北京大学	中国语文	2023
9	以“刺”为媒：作为身份象征的明清名片及其媒介功能	王炎龙;郝金华;	四川大学	新闻与传播研究	2023
10	社区空间与线上连接：基层官微矩阵重塑人地关系的效果研究	潘霁;	复旦大学	国际新闻界	2023
11	日藏启蒙类书《直音汇韵》与清初江淮方音	侯俊;	四川大学	中国语文	2023
12	认识情态词“应该”和“可能”的连用及语义组合的方式	沈园;陈婷;	复旦大学	中国语文	2023
13	报刊政论语体中的一类特殊“引起”构式	罗艺雪;	四川大学	中国语文	2023
14	说“辟彊”	张传官;	复旦大学	中国语文	2023
15	“新媒介”的使用成本：影响近代中国广播基层分布的关键因素	朱至刚;	四川大学	新闻与传播研究	2023
16	立场、情感、注意力与选择性接触：舆论极化的影响要素分析	汤景泰;徐铭亮;星辰;	复旦大学	国际新闻界	2023
17	上海话E/E?变异与上海普通话ei的双向动态影响	陈忠敏;马良;温睿;	复旦大学	中国语文	2023
18	北京大学王力语言学奖第二十届评奖工作即将开始		北京大学	中国语文	2023
19	参与模式下的争议性科学纪实影像创作——以纪录片《食物的演化》为例	邢梦莹;	复旦大学	新闻与传播研究	2022
20	试论“有报之县”在近代中国的增长轨迹：侧重于对其动力机制的阐释	朱至刚;	四川大学	国际新闻界	2022
21	湘语双峰方言的“清浊同调”——基于语音实验的探讨	史濛辉;陈轶亚;	复旦大学	中国语文	2022
22	从“帐篷话剧”到华北电影队：1930-1940年代晋察冀根据地流动的文化传播实践	张慧瑜;	北京大学	国际新闻界	2022
23	“媒介”视角下的地图知识生产：以开放街道地图为例	黄顺铭;李宏江;	四川大学	国际新闻界	2022
24	汉语方式存在句的性质、特点与历史来源	董秀芳;	北京大学	中国语文	2022
25	基于意识形态视角的马克思主义新闻观	张涛甫;	复旦大学	新闻与传播研究	2022
26	从“族”到“人”：政治传播视角下青铜器与周代政治文化演变	洪磬仪;宣柳;	北京大学	国际新闻界	2022
27	生物识别信息传播风险的刑事规制向度——基于525份刑事裁判文书的内容分析	王文娟;	四川大学	新闻与传播研究	2022
28	组合式结构与黏合式结构补议	周韧;	北京大学	中国语文	2022
29	何来“真正的媒介社会学”？——兼论媒介社会学的“连续统”观念与诠释社群	黄顺铭;李红涛;	四川大学	国际新闻界	2022
30	网络媒介“茧房效应”的类型化、机制及其影响——基于“中国大学生社会心态调查（2020）”	施颖婕;桂勇;黄荣贵;郑雯;	复旦大学	新闻与传播研究	2022
31	十七世纪以来北京话韵母e、o、uo的演变	赵彤;	北京大学	中国语文	2022
32	“结构化程度”与“各向渐变性”——宜兴话的方言测量学研究	黄河;	复旦大学	中国语文	2022
33	数字媒体与疫情省思：基于传播政治经济学的考察	曹晋;张艾晨;	复旦大学	国际新闻界	2022
34	人机传播视域下算法新闻可信度的感知研究——基于一项对大学生的控制实验分析	蒋忠波;师雪梅;张宏博;	四川大学	国际新闻界	2022
35	平台时代中国的名人文化——评《网红：作为中国的社交媒体娱乐》	曹钺;	复旦大学	国际新闻界	2022
36	锡伯语的区别宾语标记——兼论定指性与生命度作为隐性范畴在中国境内语言中的分布	鲁方昕;	北京大学	中国语文	2022
37	第十九届北京大学王力语言学奖评奖结果揭晓		北京大学	中国语文	2022
38	“表征”与“超越”:“土味视频”的“区隔”研究	宋巧丽;田辉;	四川大学	新闻与传播研究	2022
39	平台经济下“受众商品论”再审视——基于马克思主义政治经济学视角	陆茸;葛浩阳;	四川大学	新闻与传播研究	2022
40	论传统新闻业的危机及其结构化根源	姜华;张涛甫;	复旦大学	新闻与传播研究	2022
41	经由全球媒介的中西互动:《北华捷报》的信息交换网络(1850-1866)	骆世查;	四川大学	新闻与传播研究	2022
42	平台经济中的“数据劳动”：现状、困境与行动策略——一项基于电竞主播的探索性研究	任相;姚建华;	复旦大学	国际新闻界	2022

计数项:Source-文献来源	列标签			
行标签	北京大学	复旦大学	四川大学	总计
国际新闻界	9	18	14	41
新闻与传播研究	6	18	12	36
中国语文	22	20	5	47
总计	37	56	31	124

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。

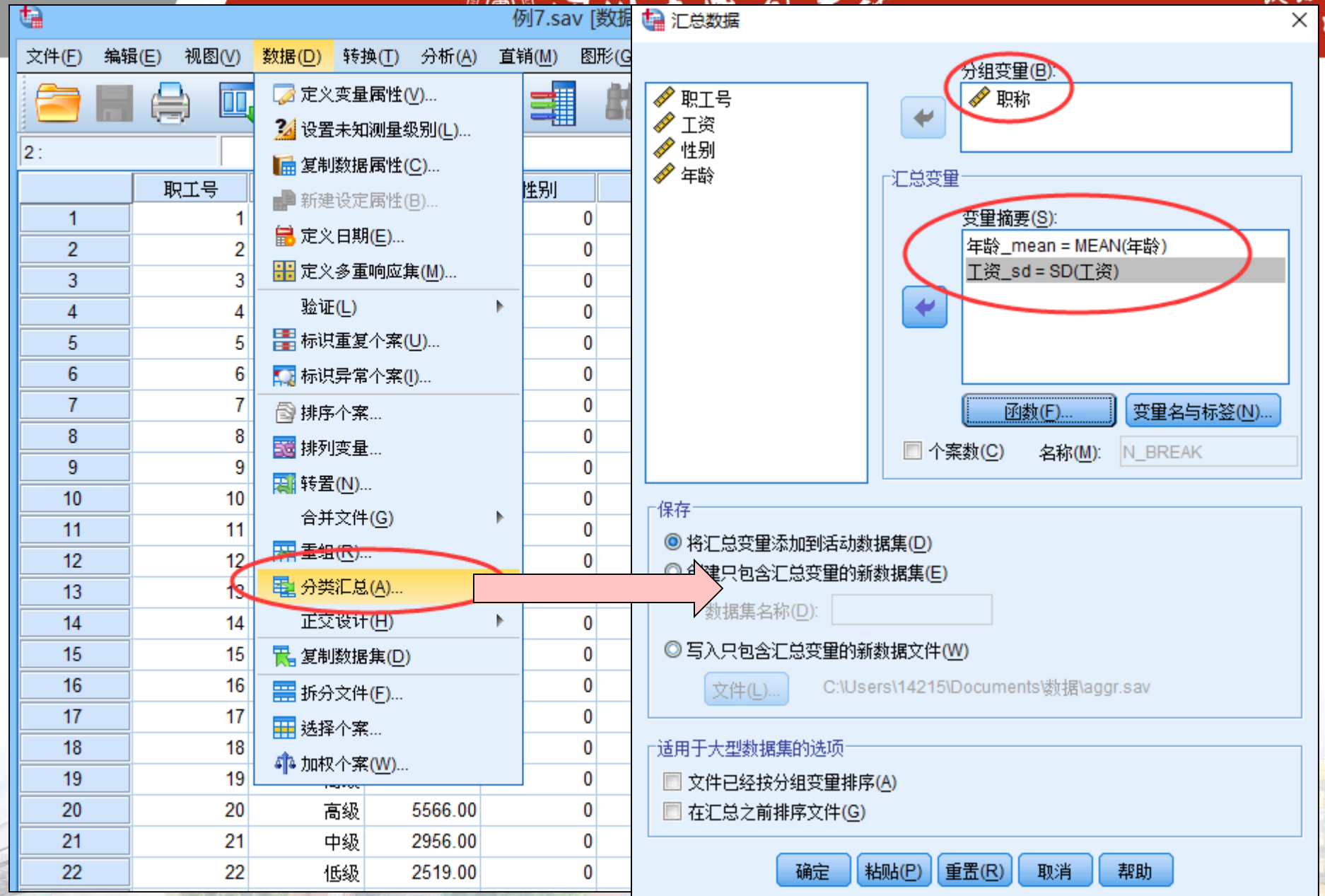
- 单值排序：排序变量只有一个。
- 多重排序：排序变量有多个，多重排序的第一个排序变量称为主排序变量，其它排序变量依次称为第二排序变量、第三排序变量等。
- 例6：打开例6. sav数据文件。要求职称按升序排序，工资按降序排序。


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。

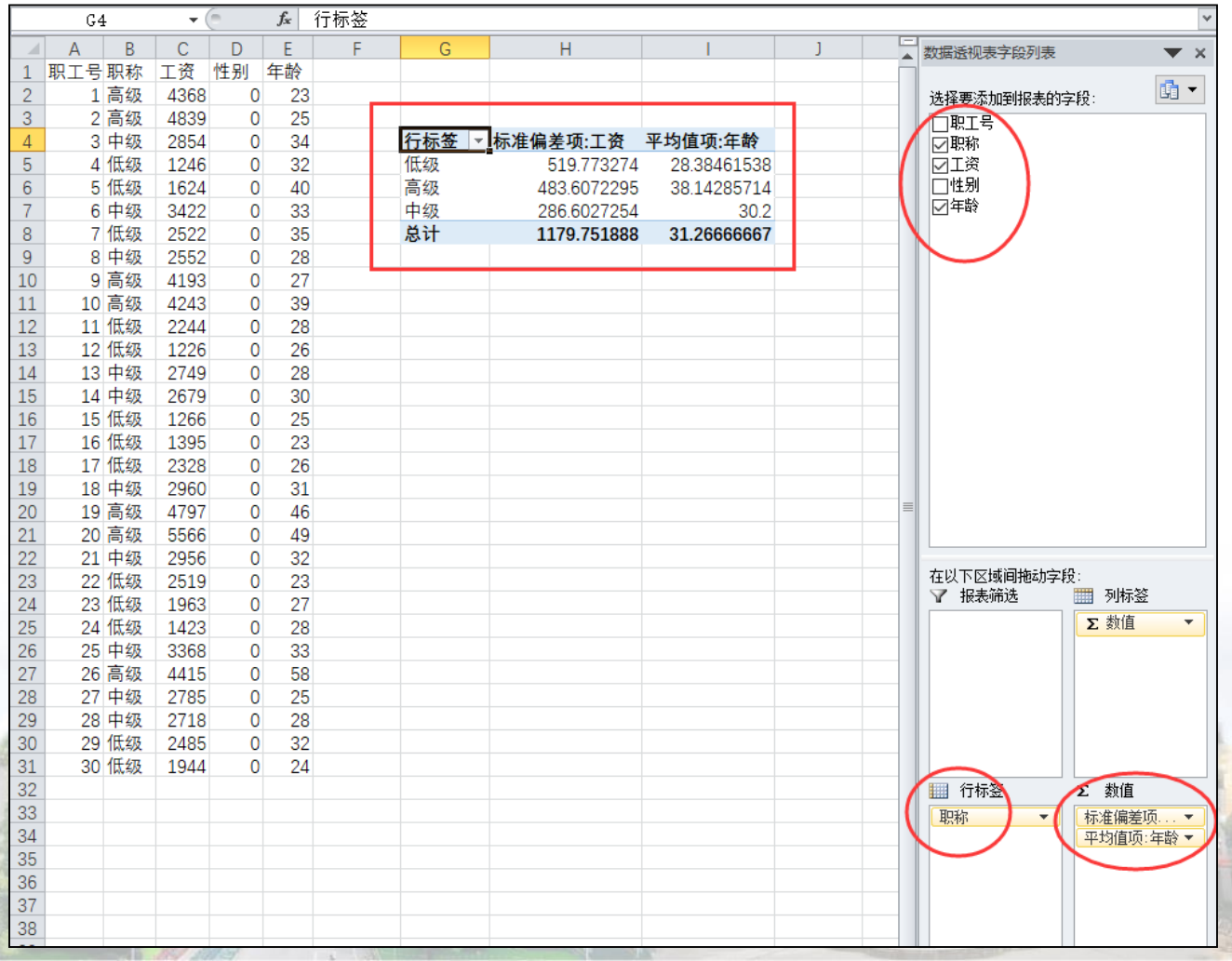

例7.sav [数据
文件(F) 编辑(E) 视图(V) 数据(D) 转换(T) 分析(A) 直销(M) 图形(G
定义变量属性(V)...
设置未知测量级别(L)...
复制数据属性(C)...
新建设定属性(B)...
定义日期(E)...
定义多重响应集(M)...
验证(L)
标识重复个案(U)...
标识异常个案(I)...
排序个案...
排列变量...
转置(N)...
合并文件(G)
重组(R)...
分类汇总(A)...
正交设计(H)
复制数据集(D)
拆分文件(F)...
选择个案...
加权个案(W)...
2 :
职工号
性别
1 1 0
2 2 0
3 3 0
4 4 0
5 5 0
6 6 0
7 7 0
8 8 0
9 9 0
10 10 0
11 11 0
12 12 0
13 13
14 14 0
15 15 0
16 16 0
17 17 0
18 18 0
19 19 0
20 20 高级 5566.00 0
21 21 中级 2956.00 0
22 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

口以性别为分组，分别计算不同组别的人数、最大年龄和工资总数

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，每个关键词在不同报纸上的频率分布。

例8.sav [数据集26] - 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

数据(D) 菜单：
- 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)...

	关键词	次	变量	变量	变量	变量
1	一带...					
2	共建					
3	发展					
4	习近平					
5	合作					
6	总书记					
7	中国					
8	高峰...					
9	推动					
10	国际...					
11	中国					
12	发展					
13	一带...					
14	国家					
15	记者					
16	亚运会					
17	共建					
18	世界					
19	杭州					
20	合作	25	光明...			
21	中国	66	经济...			
22	发展	55	经济...			
23	增长	43	经济...			
24	持续	42	经济...			

数据视图 变量视图

加权个案

请勿对个案加权(D)
加权个案(W)
频率变量(F):
次数
当前状态:不加权个案
确定 粘贴(P) 重置(R) 取消 帮助

口结果

来源

关键词			频率	百分比	有效百分比	累积百分比
持续	有效	经济日报	42	100.0	100.0	100.0
发布	有效	经济日报	35	100.0	100.0	100.0
发展	有效	光明日报	49	25.4	25.4	25.4
		经济日报	55	28.5	28.5	53.9
		人民日报	89	46.1	46.1	100.0
		合计	193	100.0	100.0	
高峰论坛	有效	人民日报	38	100.0	100.0	100.0
共建	有效	光明日报	26	19.1	19.1	19.1
		人民日报	110	80.9	80.9	100.0
		合计	136	100.0	100.0	
国际合作	有效	人民日报	34	100.0	100.0	100.0
国家	有效	光明日报	46	57.5	57.5	57.5
		经济日报	34	42.5	42.5	100.0
		合计	80	100.0	100.0	
杭州	有效	光明日报	26	100.0	100.0	100.0
合作	有效	光明日报	25	36.2	36.2	36.2
		人民日报	44	63.8	63.8	100.0
		合计	69	100.0	100.0	
记者	有效	光明日报	32	100.0	100.0	100.0
建设	有效	经济日报	31	100.0	100.0	100.0
全国	有效	经济日报	32	100.0	100.0	100.0
世界	有效	光明日报	26	100.0	100.0	100.0

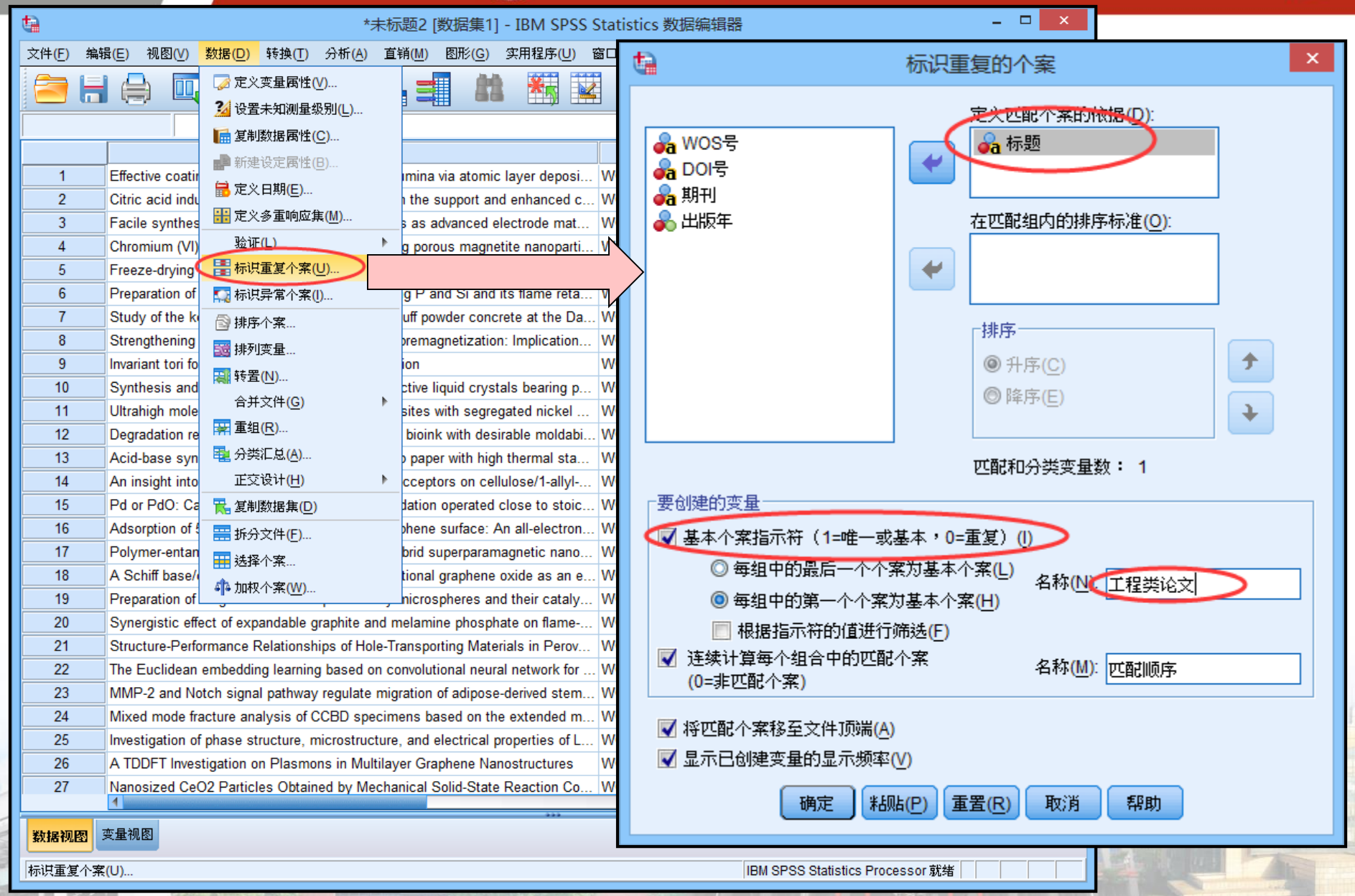
口例8-1选取了三个高校新闻学院的发文学科，现在希望对比这三个机构在不同学科下的发文情况

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



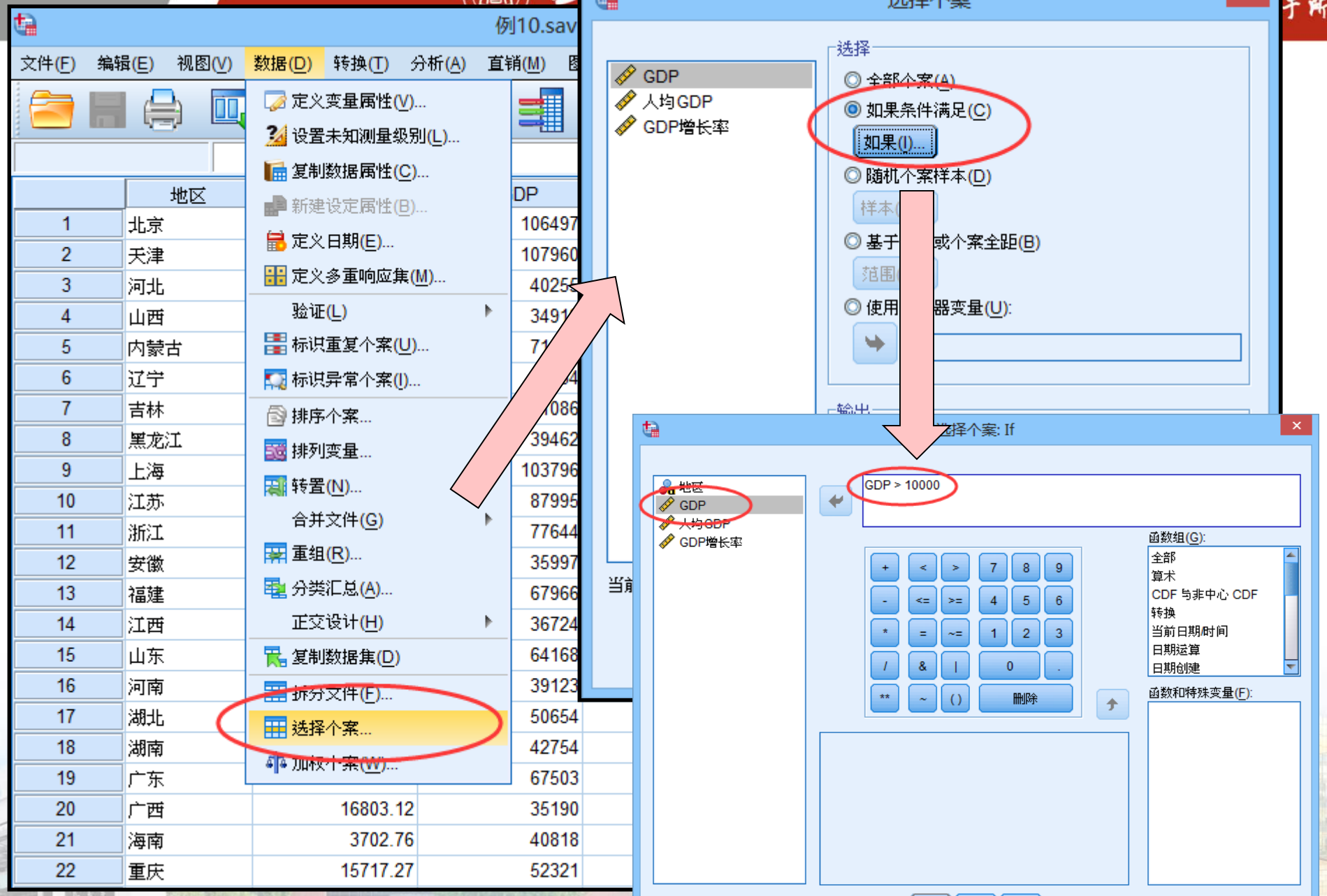
个案的选择

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
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

口选择人均GDP在30000—80000区间的地区

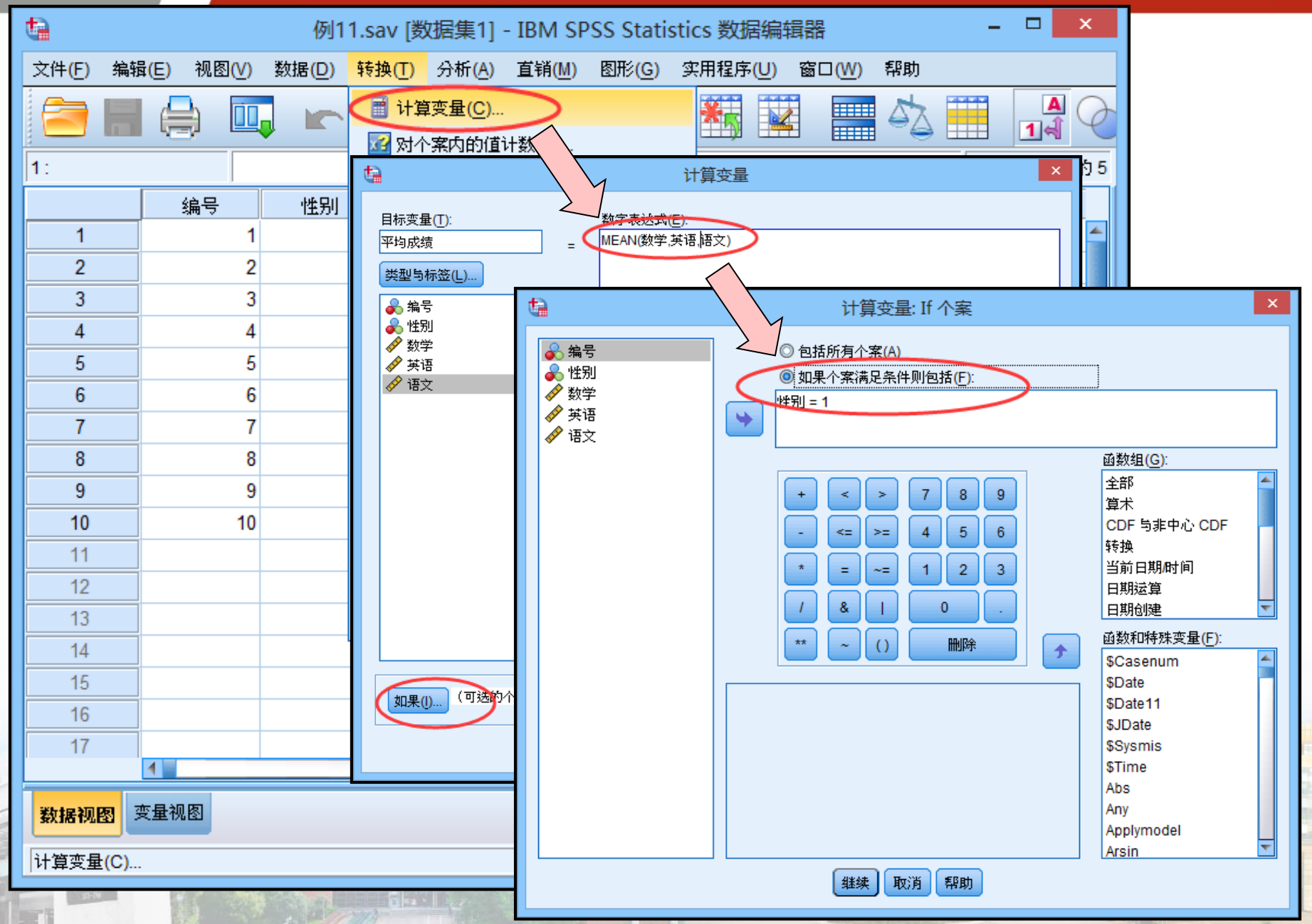
口选择GDP增速低于5%或高于8%的地区

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



口求数学成绩大于84分的同学的平均成绩

变量值的重新编码

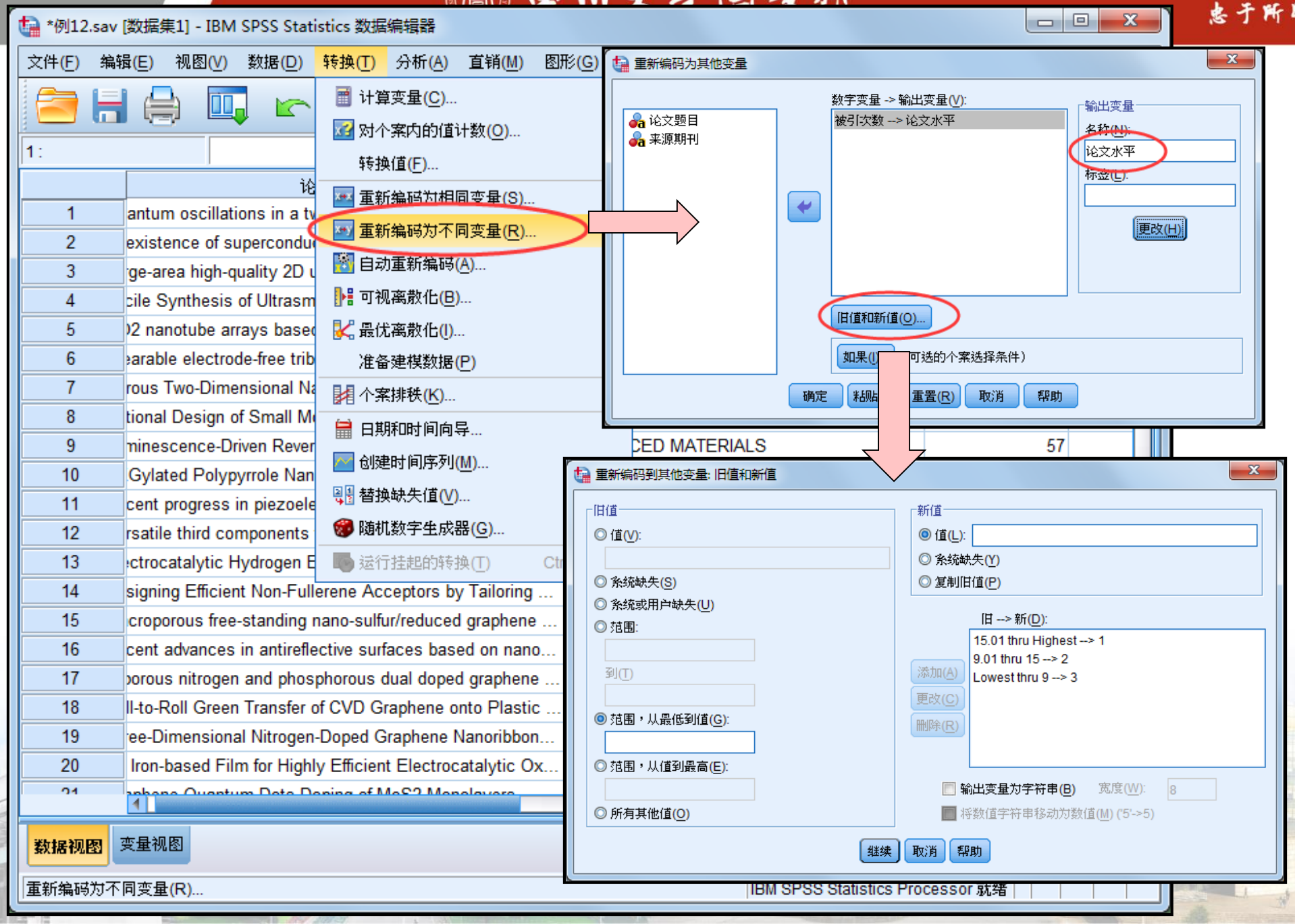
口数据分析中，将连续变量转换为等级变量，或者将分类变量不同的变量等级进行合并，例如把同学的成绩分为优、良、中、差四个等级。

- 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



口根据例12-1，按照下载量对新闻的阅读优先级进行区分

- 40以上：今天阅读
- 10-39：明天阅读
- 9以下：以后阅读

主要内容

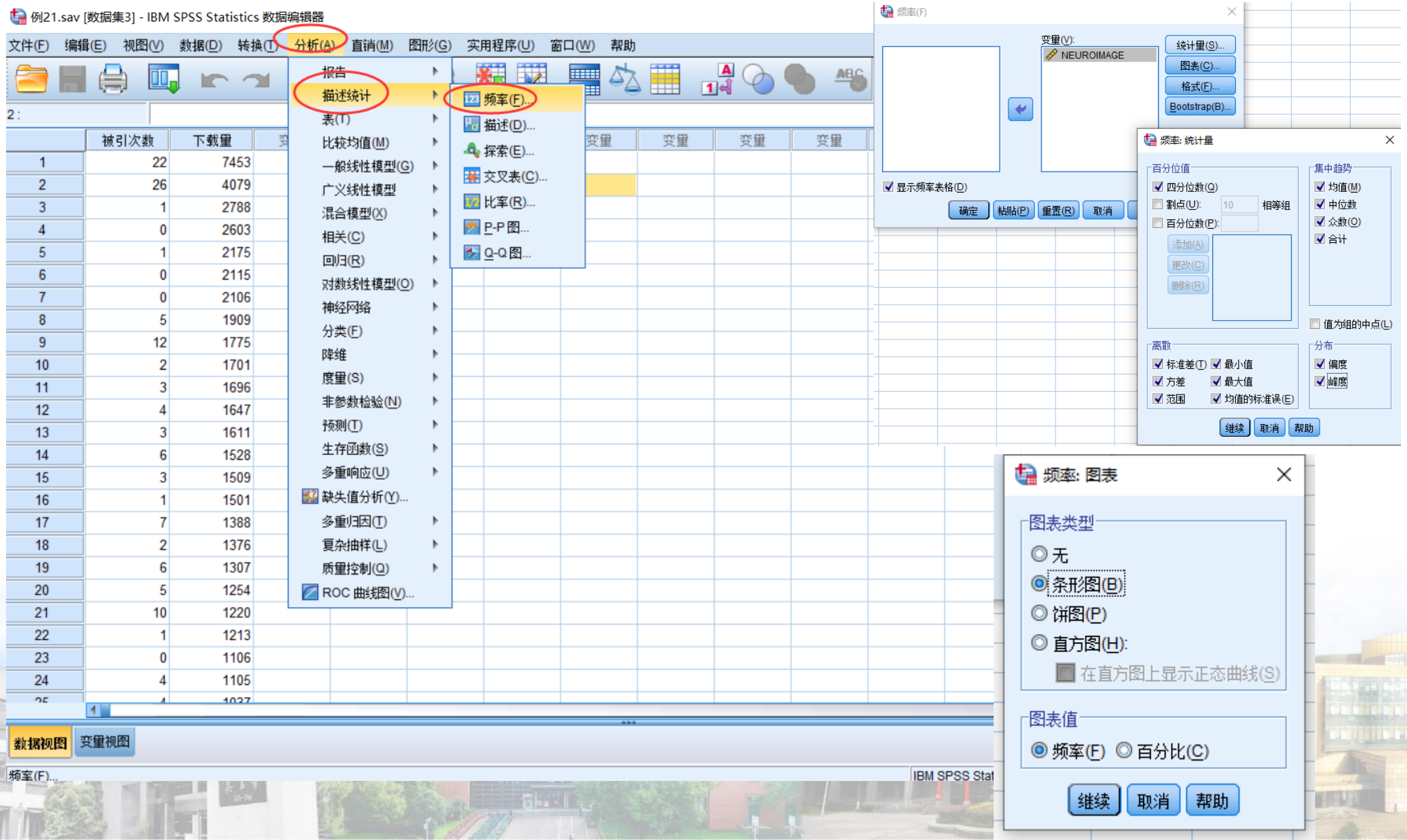
- 1-SPSS概述
- 2-数据管理
- 3-统计分析与统计图

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

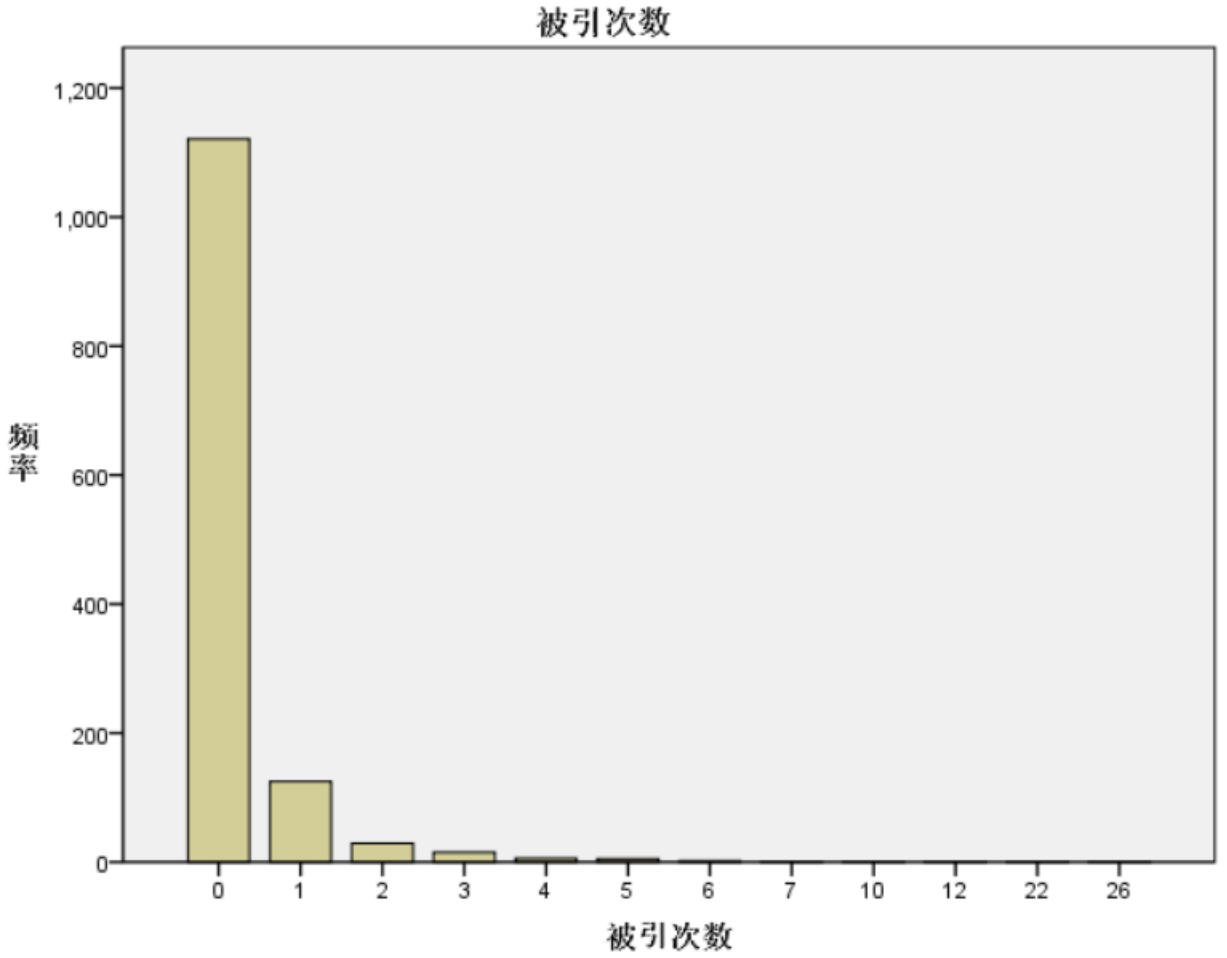
- 频数分布分析主要通过频数分布表、条图、直方图以及集中趋势和离散趋势各种统计量，描述数据的分布特征。
- 例21：了解 “新闻学”领域论文被引次数与下载量的分布特征


例21.sav [数据集3] - 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...
Q-Q 图...
被引次数 下载量
数据视图 变量视图
频率(F)
变量(V):
NEUROIMAGE
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
频率: 统计量
百分位值
四分位数(Q)
割点(U): 10 相等组
百分位数(P):
添加(A)
更改(C)
删除(R)
集中趋势
均值(M)
中位数
众数(O)
合计
值为组的中点(L)
离散
标准差(T) 最小值
方差 最大值
范围 均值的标准误(E)
分布
偏度
峰度
继续 取消 帮助
频率: 图表
图表类型
无
条形图(B)
饼图(P)
直方图(H):
在直方图上显示正态曲线(S)
图表值
频率(F) 百分比(C)
继续 取消 帮助

统计里

被引次数

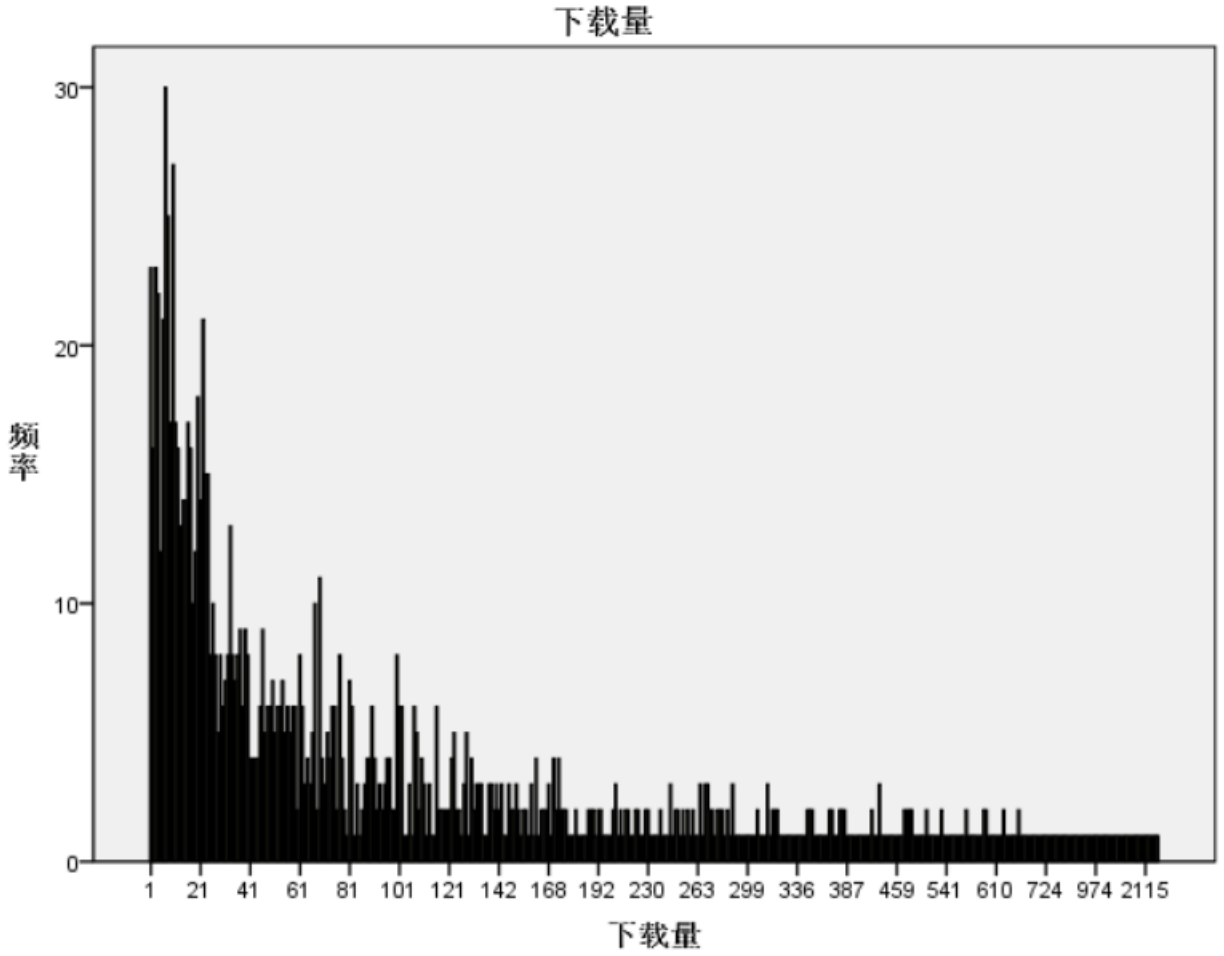
N	有效	1308
	缺失	30
均值		.28
中值		.00
众数		0
偏度		12.676
偏度的标准误		.068
峰度		221.339
峰度的标准误		.135



统计量

下载量

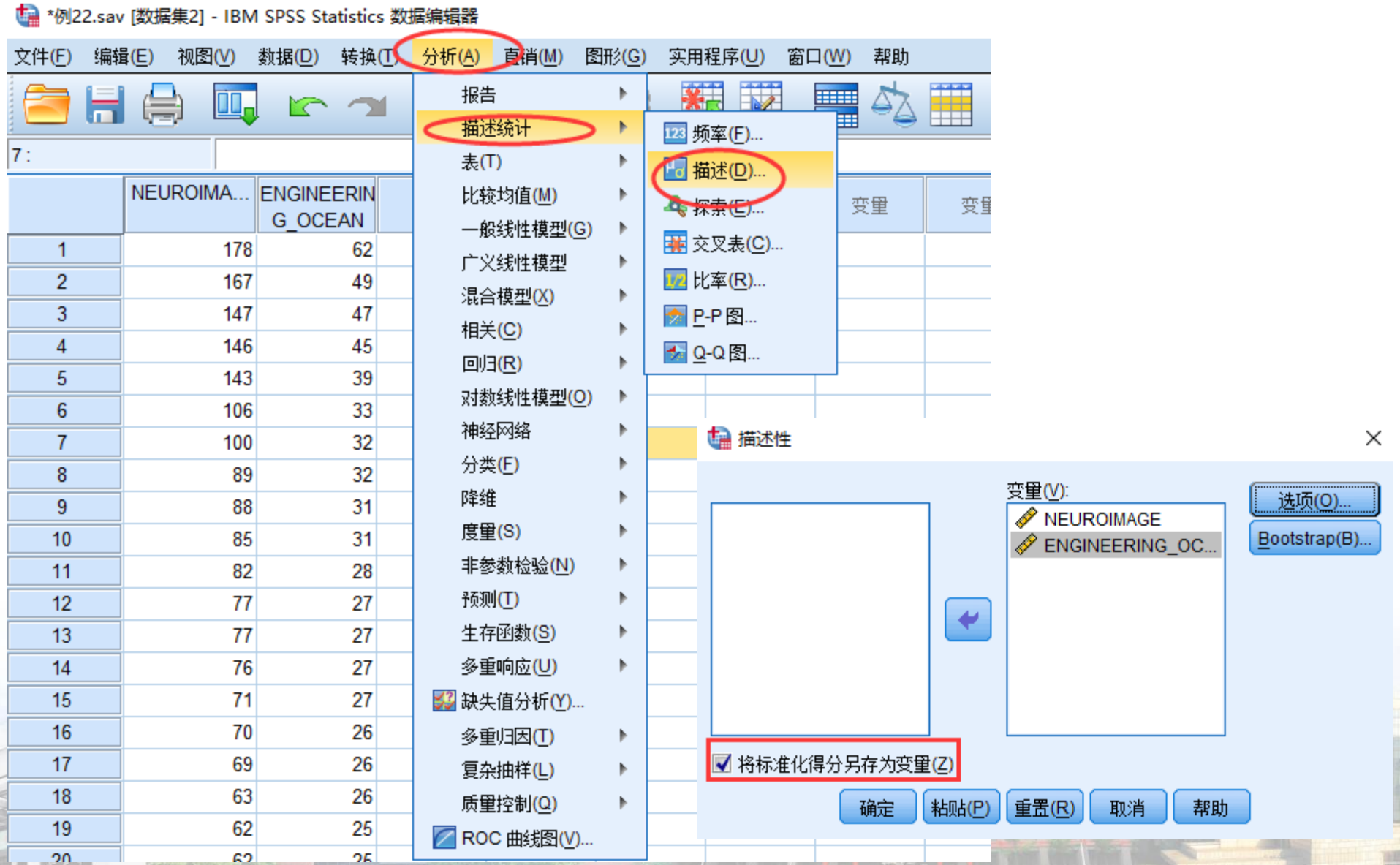
N	有效	1308
	缺失	30
均值		161.56
中值		58.00
众数		7
偏度		9.642
偏度的标准误		.068
峰度		157.537
峰度的标准误		.135



口描述性统计分析

- 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。

口例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
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...
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
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											

单样本检验

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。

- 认为均值相等，根据小概率原则，我们认为小于5%的小概率事件在单次抽样中基本不可能发生
- 可以构造出一个样本的t统计量，把根据抽样计算出的结果与临界值比较，如果超过了临界值（偏离0“太多”），则认为小概率事件发生了
- 如果均值为X，我们竟然在一次抽样中抽到了非常小概率的情况，那么说明均值很可能不是X

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。

口为什么要进行方差齐性检验

- 在总体方差未知的情况下，需要方差齐性的前提才能计算出t值

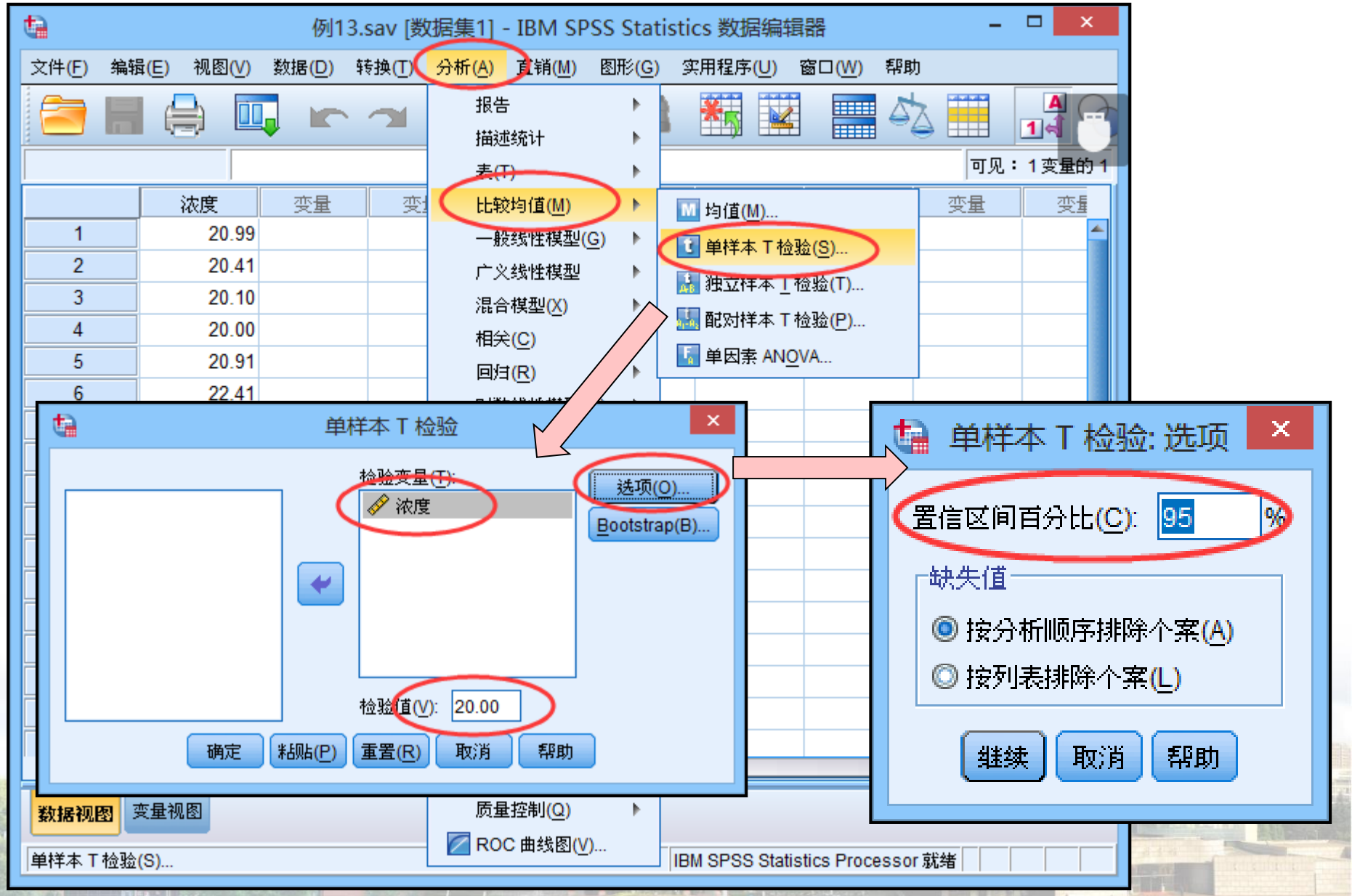
单样本检验与独立样本检验

口应用场景：

- 推断两个或更多总体之间水平的不同，例如已知某领域的热点新闻转载量的均值，现在挑选了50条该领域的新闻，希望检验一下这50条新闻是否可以代表热点新闻。

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

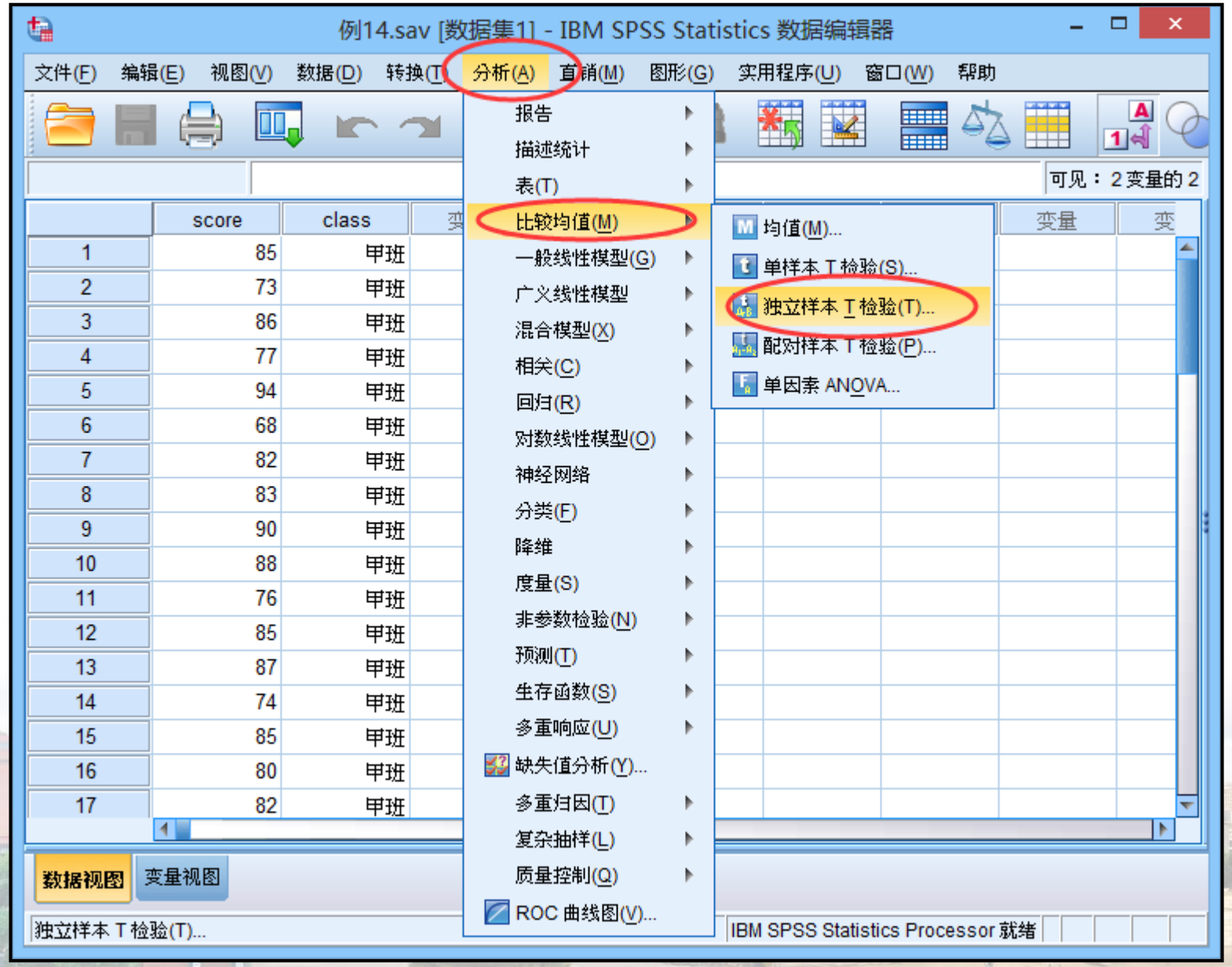
单个样本检验

	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

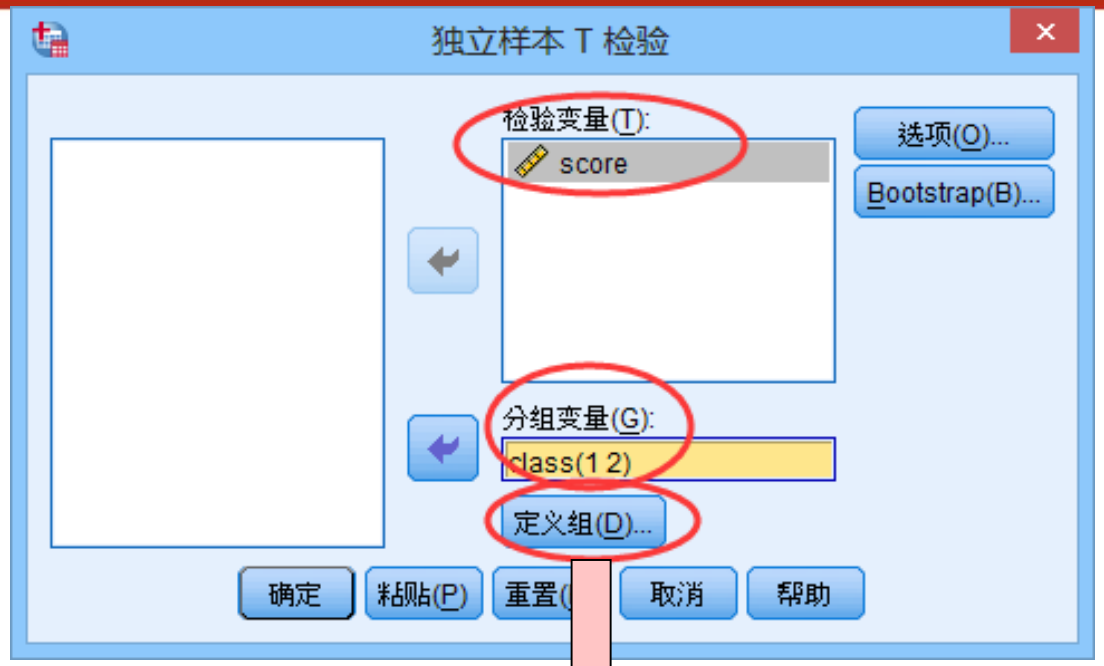
独立样本检验

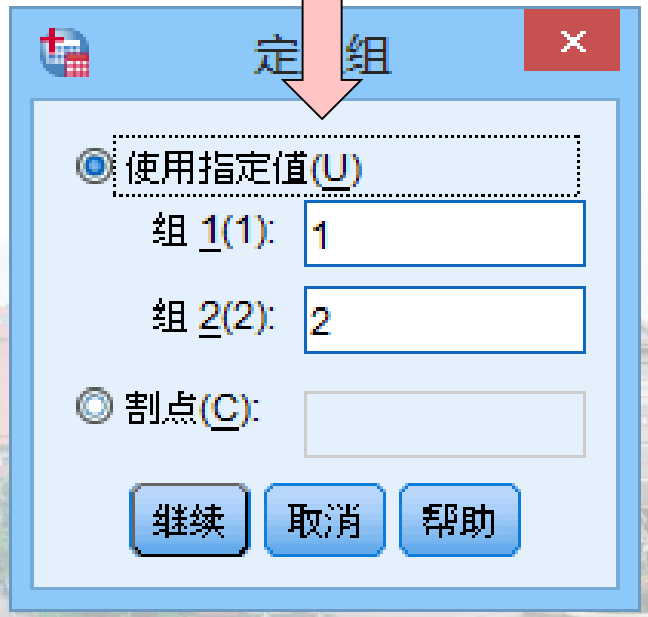
口例14：现希望评价两位老师的教学质量，试比较其分别任教的两班考试后的成绩是否存在差异。数据见例14.sav。

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

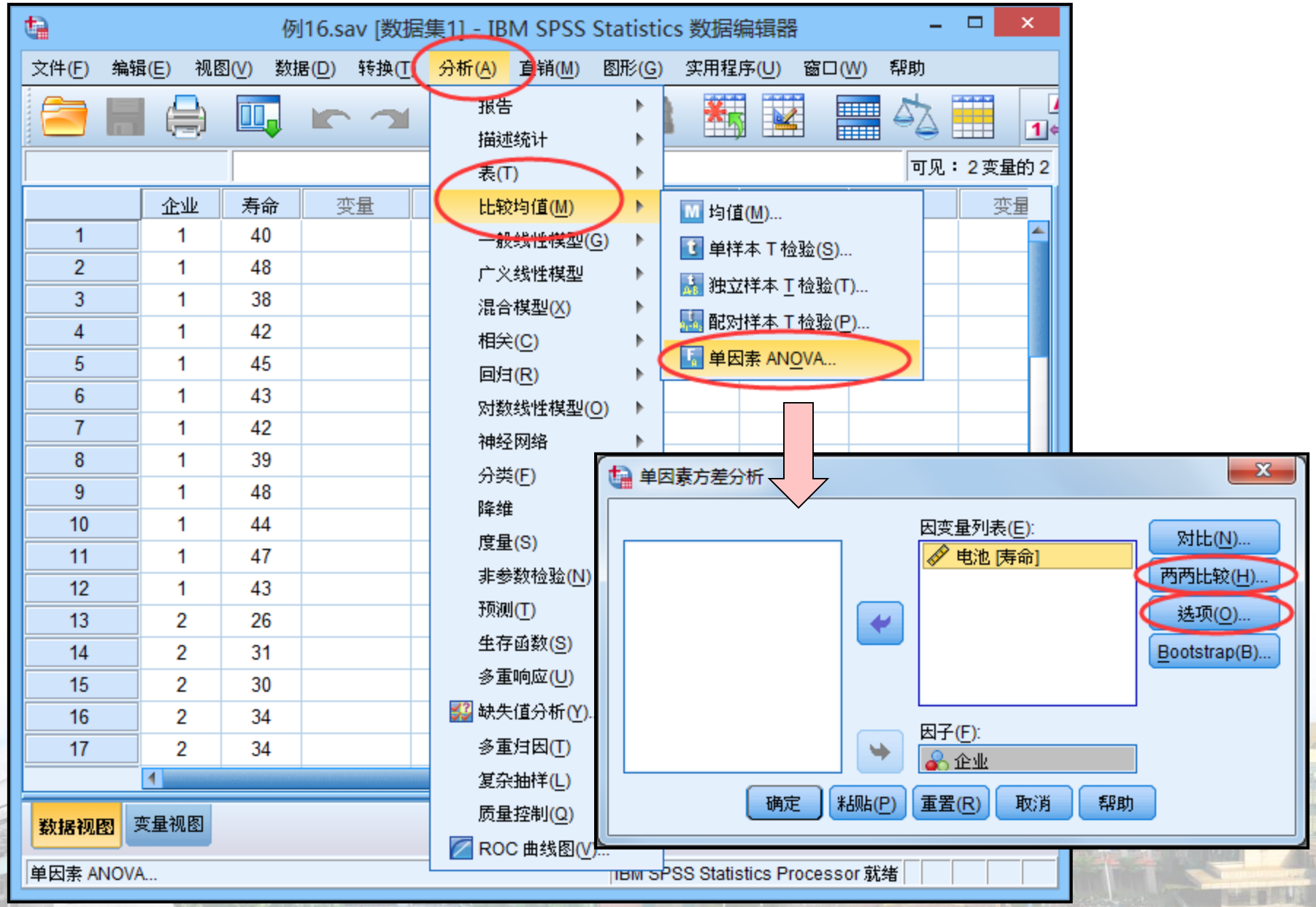
独立样本检验

		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

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。

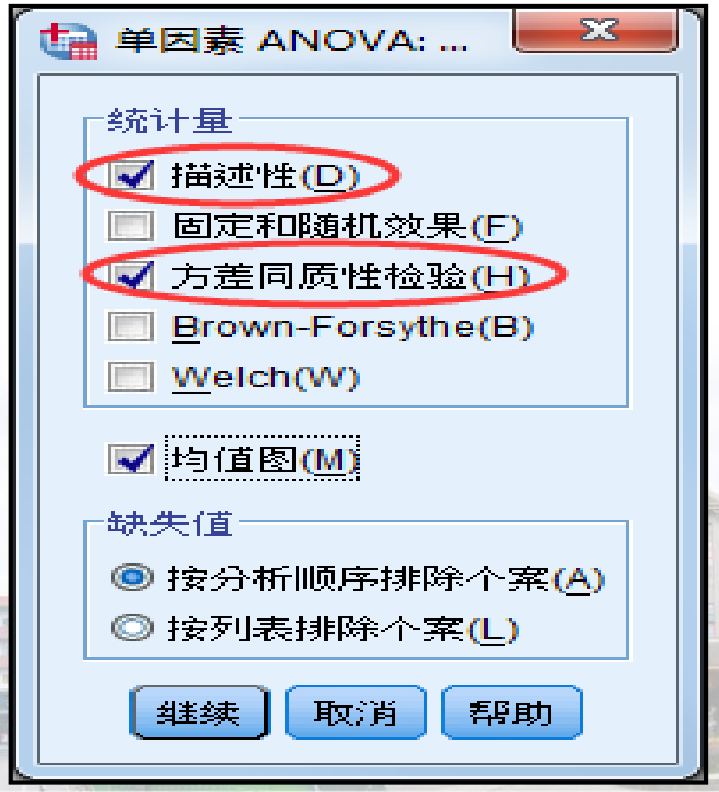
口例15：比较3个不同电池生产企业生产电池的寿命，见例16.sav


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

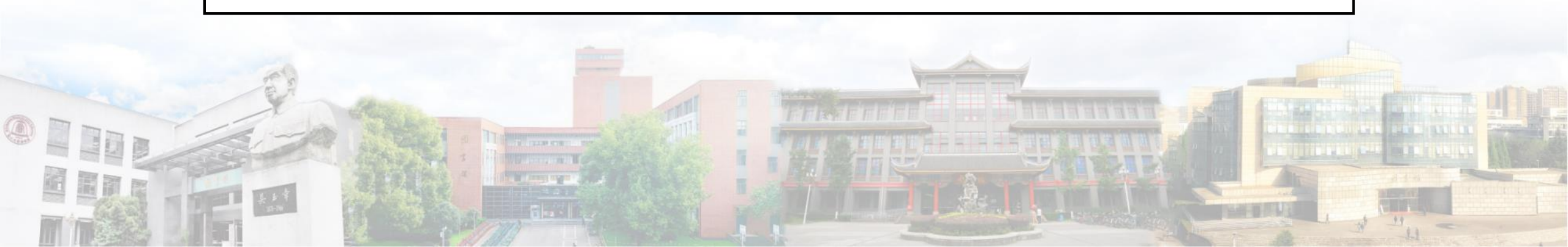


口结果显示-1

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。

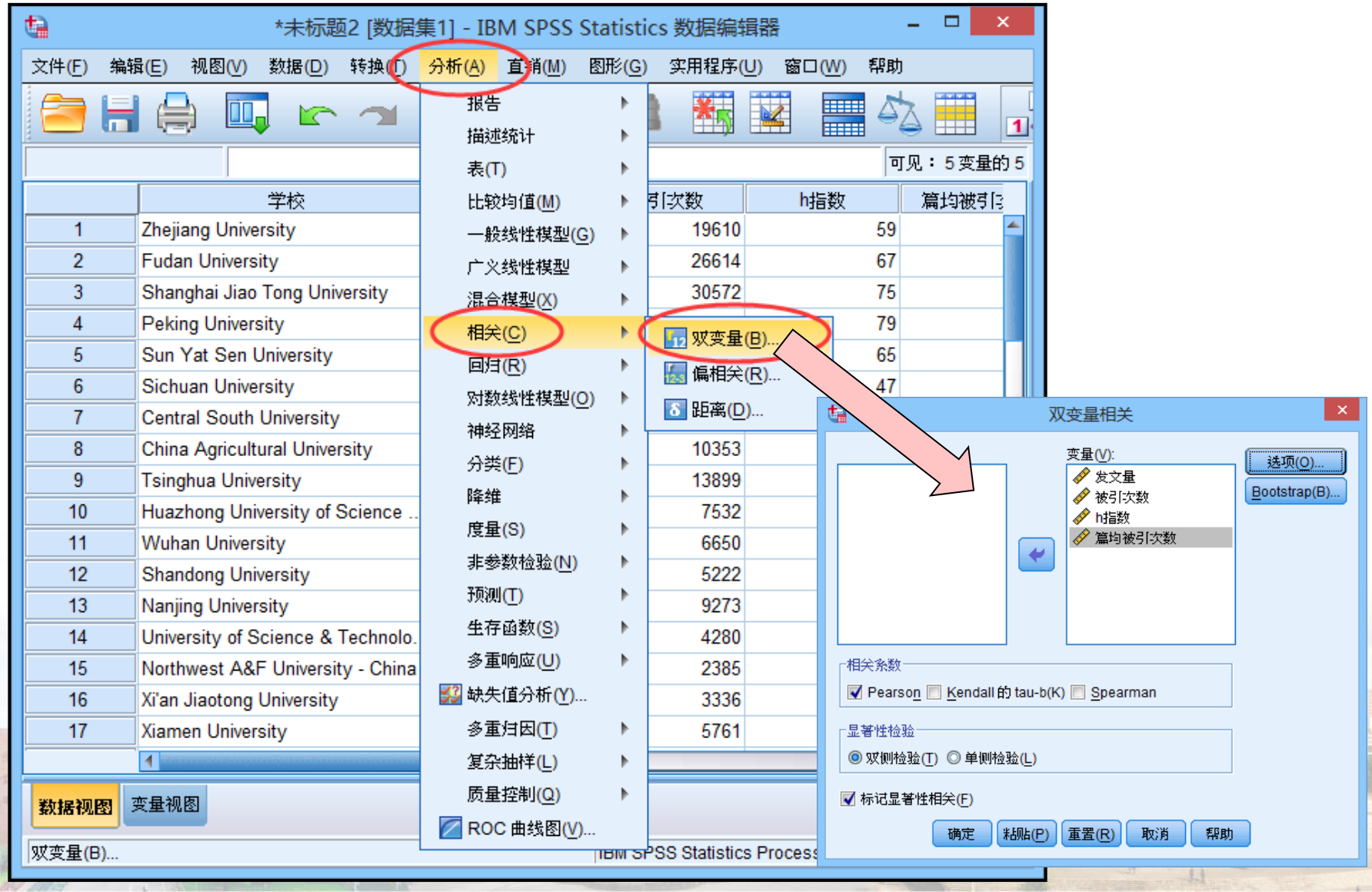
相关分析

口线性相关系数：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

口例16：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5.sav。

相关分析


*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

相关分析

口结果显示

相关性

[数据集1]

相关性

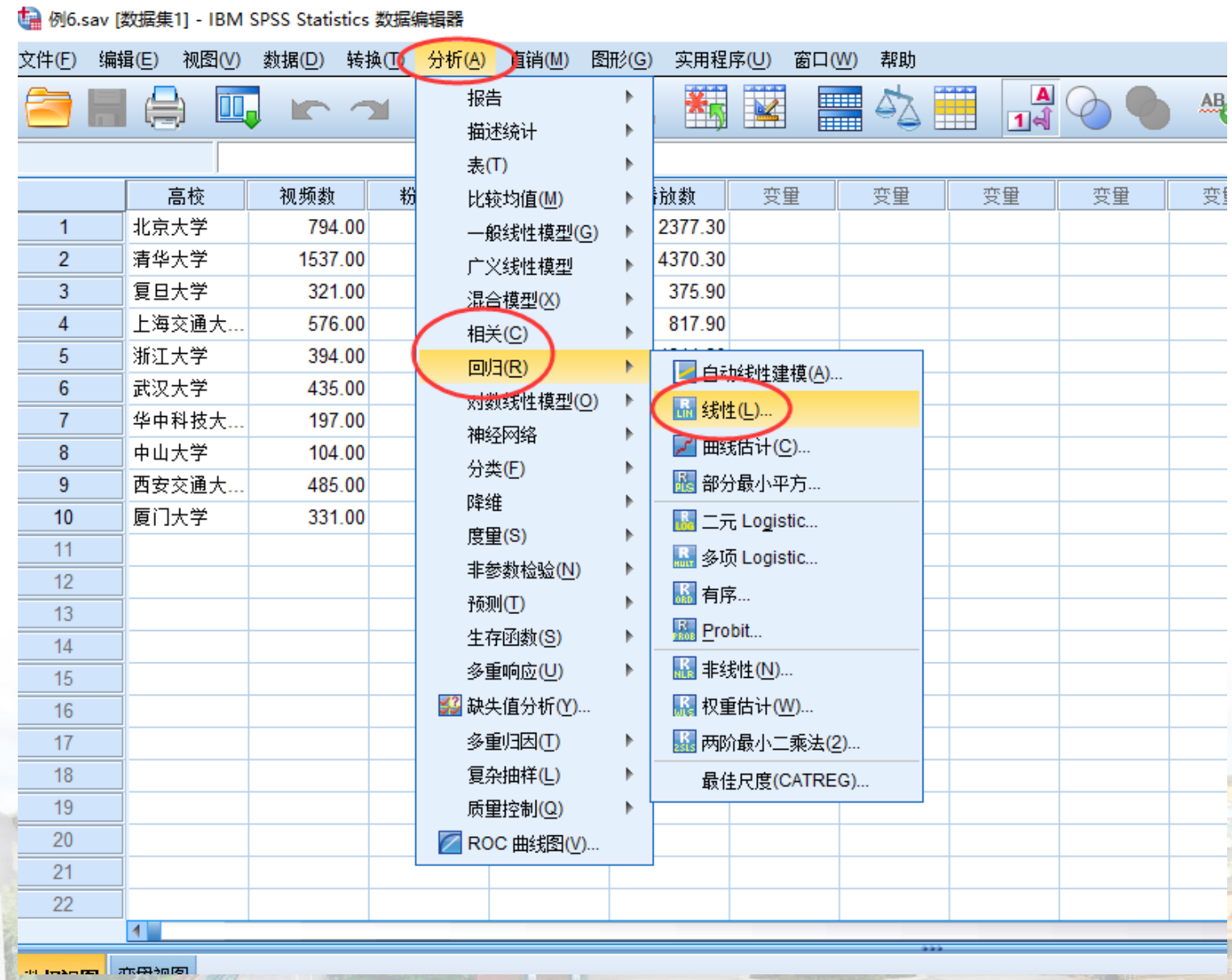
		发文量	被引次数	h指数	篇均被引次数
发文量	Pearson 相关性	1	.947**	.913**	-.104
	显著性（双侧）		.000	.000	.530
	N	39	39	39	39
被引次数	Pearson 相关性	.947**	1	.926**	-.032
	显著性（双侧）	.000		.000	.847
	N	39	39	39	39
h指数	Pearson 相关性	.913**	.926**	1	-.147
	显著性（双侧）	.000	.000		.372
	N	39	39	39	39
篇均被引次数	Pearson 相关性	-.104	-.032	-.147	1
	显著性（双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

**. 在 .01 水平（双侧）上显著相关。

线性回归

- 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17：10个高校在B站官方账号的获赞数和播放量数据见例6.sav，试进行回归分析

线性回归


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
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)...
自动线性建模(A)...
线性(L)...
曲线估计(C)...
部分最小平方...
二元 Logistic...
多项 Logistic...
有序...
Probit...
非线性(N)...
权重估计(W)...
两阶最小二乘法(2)...
最佳尺度(CATREG)...
高校 视频数 播放数 变量 变量 变量 变量 变量
1 北京大学 794.00 2377.30
2 清华大学 1537.00 4370.30
3 复旦大学 321.00 375.90
4 上海交通大... 576.00 817.90
5 浙江大学 394.00
6 武汉大学 435.00
7 华中科技大... 197.00
8 中山大学 104.00
9 西安交通大... 485.00
10 厦门大学 331.00
11
12
13
14
15
16
17
18
19
20
21
22

线性回归

模型汇总

模型	R	R 方	调整 R 方	标准 估计的误差	更改统计量				
					R 方更改	F 更改	df1	df2	Sig. F 更改
1	.990[a]	.981	.975	18.45098	.981	179.557	2	7	.000

a. 预测变量: (常量), 视频数, 粉丝数。

1、线性回归出来的相关系数为R=0.990。
2、方程拟合优度R方是0.981，调整后的R方为0.975
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。

线性回归

Anova[b]

模型		平方和	df	均方	F	Sig.
1	回归	122256.574	2	61128.287	179.557	.000[a]
	残差	2383.070	7	340.439		
	总计	124639.644	9			

a. 预测变量: (常量), 视频数, 粉丝数。
b. 因变量: 获赞数

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
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

线性回归

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)	-22.936	12.836		-1.787	.117
	粉丝数	.811	.215	.579	3.776	.007
	视频数	.123	.044	.427	2.785	.027

a. 因变量: 获赞数

- 给出了常数项和回归系数。
- 分析X的显著性（P值），如果呈现出显著性，则说明X对Y有影响关系。如果不显著，则应剔除该变量。
- 可以看到，2个解释变量对获赞数的显著性分析P值均小于0.05，说明X对Y均有显著性影响关系。
- 获赞数=-22.936+0.811*粉丝数+0.123*视频数

- 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

感谢大家参与！