第五章　SPSS

§5.1　SPSS/PC+ 导引

§5.1.1　简介

SPSS 由美国斯坦福大学1965 年开始研究并于1970 年推出。SPSS- X 用于IBM CMS、MVS/TSO、UNIX 和DEC VAX/VMS 系统，允许用户以批处理方式运行。微机版SPSS/PC+V2.0，由Chicago 为基础的SPSS 公司于1987 年推出。SPSS 广泛用于商务、政府部门、教学与科研单位进行调查分析、市场研究、产品检测、人事管理与决策、卫生服务分析以及统计质量控制等。

本章主要介绍SPSS/PC+，其功能概要如下：

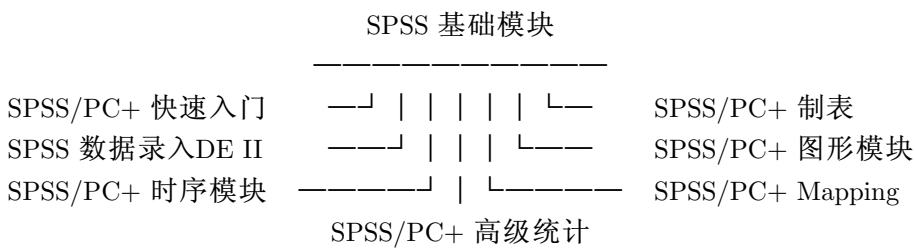

图 5.1　SPSS/PC+ 功能示意图

它支持的数据格式有ASCII、dBASE II-IV、Lotus 1- 2- 3 、symphoney 、mutiplan、及SPSS-X 传输格式。SPSS/PC+提供与其它图形软件的接口。其graph -in-the-box 允许用户在SPSS/PC+ 环境下产生、浏览和修正图形；SPSS/PC+ Mapping 支持Ashton-Tate MAP-MASTER。

SPSS/PC+ 高级统计拥有多元方差分析(MANOVA)、判别分析、因子分析、聚类分析、对数线性模型、非线性回归、logistic回归分析以及可靠性分析(、对应分析等。SPSS/PC+的多分类分析(Multiple Classification analysis, MCA) 一般统计分析软件不专门具备。

时间序列分析是利用它的Trend部分，其功能有指数平滑、曲线拟合、特定形式的回归、ARIMA 模型(Box-Jenkins)、谱分析。

§5.1.2　SPSS/PC+ 工作方式

在DOS 系统下，设软件存放于目录SPSS，打入命令：CD SPSS ¡Enter¿，这时可用三种方式运行SPSS/PC+。

㈠SPSS/PC+ 菜单方式

执行命令：C:\SPSS>SPSSPC <Enter>

屏幕显示：

这时打入菜单上的英文，则光标移至相应的项目，<Enter>键选择后，进入下一层菜单，再打入子菜单英文名，用<Enter>继续选择。

利用光标键移动时，每项下有相应的解释，右箭头由上级菜单向下级菜单推进，到达待选命令或选项时，以回车选定项目，如果不满意可以使用Alt-D 删除；左箭头令光标返回上级菜单。在菜单方式下，使用Alt-E 键进行编辑态，就可以编辑程序或插入外部程序(F3)，打入SPSS/PC+的关键字后再用Esc 键则系统立即调出键入命令有关信息。打入F10，系统提示下一步的运行方式，同第7章将介绍的SYSTAT 一样，SPSS/PC+ 可以从光标位置开始运行。

orientation	入门
read or write data	读写数据
modify data or files	修改数据或文件
graph data	数据绘图
analyze data	分析数据
session control& inf	运行控制和信息
run DOS or other pgms	运行DOS或其他程序
extended menus	扩展菜单
SPSS/PC+ options	选项
FINISH	完成
	F1=帮助Alt-E=编辑Alt-M=菜单开/关

图 5.2 SPSS/PC+ 主菜单

㈡行命令方式

在菜单方式下，使用Alt/SHIFT-F10 进入对话方式，系统使用提示SPSS/PC:，此时每输入一条命令，都被立即执行，执行结束后仍然返回提示下。每行命令以圆点(.) 结束，当一行写不完时，可用<Enter> 在下一行继续输入。这种运行方式适于程序的调试。

㈢批处理方式

在MS-DOS 系统下相应的批处理方式是：SPSSPC ¡程序文件名¿ ¡Enter¿。除非特别指定(SET LOG/LISTING)，程序的运行情况和结果分别存于SPSS. LOG 和SPSS.LIS 中。

现运行系统提供的基础模块检验程序BASETEST.INC：使用命令为：

C:\SPSS>SPSSPC BASETEST.INC <Enter>

同时处理多个程序，可把它依次写在SPSSPC 后面即可，程序之间用空格分开。

在SPSS/PC+ 提示下使用INC ”程序名”. ¡Enter¿，即交互下的批处理方式。也可以用@程序名.

SPSS/PC+ 通过执行命令FINISH、STOP 或BYE、EXIT 返回DOS 系统。使用DOS ¡DOS 命令¿ 运行DOS 命令或仅仅使用DOS. 命令时进入DOS 内核，返回外壳SPSS/PC+ 仍然用EXIT 命令。

在行命令方式下，使用REVIEW SCRATCH. 或REVIEW. 命令返回系统菜单，使用REVIEW LOG, REVIEW LISTING, REVIEW BOTH, REVIEW FILENAME 也都合法。SPSS/PC+ 菜单方式下，有关的功能键列表如5.3(a)-(b)：

编辑态的的几种功能可单独使用功能键或通过Ctrl、Alt 与功能键或字母的组合来完成，有些功能与不同的编辑状态有关，如F3仅当插入态能用。SPSS/PC+ 可以标记系统命令，这种做法很有特色，另外，SPSS/PC+可以对输出结果的小数位进行四舍五入。做法是先做标记，然后打Ctrl-F7 系统就要提示需要的小数位数。SPSS/PC+关键字的在线帮助可经Alt-G来完成，可在使用Alt-E并移动光标至命令字处，此时使用Alt-G则调出相应的字汇解释，Alt键与一些字母组合的功能如下：

信息	F1	Review 帮助和菜单，变量和文件列表，专用词汇表
窗口	F2	切换，大小，缩放
输入文件	F3	插入文件，编辑其它文件
行	F4	插入，删除，恢复
搜寻替换	F5	文本查找，替换文本
跳转	F6	区域，输出页，错误行，最未执行行
定义区域	F7	标记/取消行标记，矩形标记或命令标记
区域操作	F8	拷贝，移动，删除，数据操作，拷贝专用词项
输出文件	F9	写标记区或文件，文件删除
运行	F10	从光标或标记区运行，退至命令行提示下

图 5.3 review 功能键

ENTER	剪贴选择以及在菜单下移一个水平
TAB 或→	暂时剪贴选择以及菜单下移一个水平
ESC 或←	最未一个暂存剪贴上移一个水平
Alt-ESC	到主菜单(同Ctrl-ESC)
Alt-K	删除所有剪贴暂存区
Alt-T	进入录入窗口
Alt-E	编辑态切换
Alt-M	关闭/启用菜单
Alt-V	进入变量窗口
Alt-C	自光处标运行

图 5.4 主要菜单命令

Alt-B 向前插入一行
Alt-D 删除一行，不论是否为编辑态
Alt-F 磁盘文件列表
Alt-G 专用词汇表
Alt-H 开启/关闭帮助窗口
Alt-I 向后插入一行
Alt-P 块写文件
Alt-R REVIEW 帮助
Alt-S 窗口切换
Alt-U 删除恢复(UNDELETE)
Alt-W 写文件
Alt-X 标准菜单与扩展菜单切换，后者专门存放一些不太常用的信息
Alt-Z 窗口缩放切换，放大时可进行全屏幕编辑
如对系统命令比较熟悉，则进入SPSS/PC+后直接用Alt-Z进行全屏幕编辑。

§5.1.3 系统装卸

SPSS/PC+ 的“运行控制和信息”项下的SPSS MANAGER 命令完成。STATUS 用于显示当前的安装情况，INSTALL 指定安装，REMOVE 指定删除。

§5.2 SPSS/PC+ 语言

§5.2.1 语言要素

1. 表达式。常用操作符：
算术运算符+、-、*、/、** 分别对应加、减、乘、除、乘方等运算。逻辑操作符=、<>, <=, <, >, >=运算符，及Fortran中的EQ、GT等。关系运算符：all、by、and、not、or、to、with。

2. 函数。
ABS(绝对值) RND(四舍五入) TRUNC(取整)
MOD10(对10 取模) SQRT(平方根) LG(常用对数)
LN(自然对数) SIN(正弦) COS(余弦)
ARTAN(反正切) UNIFORM(0～x 间的均匀分布随机函数)
NORMAL(均值为零、标准差为x 的正态分布随机数)
LAG(函数取前一个变量的值赋给命名量)
YRMODA 是一个时间函数，把年月日转成天数。

在SPSS for Windows中函数的种类大大增多。

3. 语句。SPSS/PC+ 命令由关键字和说明部分组成，命令关键字告诉系统进行哪一种操作，说明部分也就是命令参数(命令对象和选择项)，即：

命令关键字＋命令对象＋命令选项＋命令结束符(.)

命令对象明确对变量表、表达式或文件进行操作。命令选项一般应予使用，分必选项和可选项两种。

如：GET /FILE='NEW.SYS' /KEEP=AGE. 关键字是GET，其余为说明部分。

约定文件名、变量名不超过8个字符，不能有空格，首字母必须是字母，文件名应用引号括起来。定义变量名时可以使用关键字TO，如定义变量X1,..., X10 可用X1 TO X10 表示。命令说明部分的关键字一般是保留关键字，如TO、LT 、NOT、LOWEST、THRU、HIGHEST等。变量值可以是数值型或字符型，数值型值可以是整数或小数，字符型必须用引号括起来。标号是对变量或变量值的说明。两个元素或变量之间一般用空格或逗号分隔，函数符号与自变量用圆括号分隔、子命令关键字与参数用等号分隔、子命令之间用斜杠分隔。

SPSS/PC+ 命令大致归类：系统命令，包括安装、参数设定、显示、列表等命令；数据定义命令，包括变量的产生、修改、分组调整等命令；文件管理、合并等命令；统计分析命令。频数表、列联表、方差分析、多元分析等；其它命令。SPSS/PC+命令可以(组合)简写：如：COM(COMPUTE)、REC(RECODE)、DATLIS(DATA LIST)、详见文件SPSSV4.TBL。

§5.2.2 数据和文件管理

SPSS/PC+ 的文件有活动文件(active file)、数据文件(data file)、系统文件(system file) 、引用文件(include file)、列表文件(list file)与工作文件(working file)几种。

活动文件是系统运行时所使用的文件，包括数据和数据结构，它们在退出系统后消失。数据文件一般是ASCII 文件，由DATA LIST 读取。系统文件是SPSS/PC + 内部文件，可以使用GET、JOIN、SAVE、AGGREGATE 命令进行操作，扩展名一般取为.SYS。引用文件使用扩展名.INC 及.LOG。由INCLUDE 命令引用。列表文件使用扩展名.LIS，存放执行的结果，如SPSS.LIS。工作文件由SPSS/PC+ 运行时使用的暂存文件，使用扩展名.SY1、.SY2。

活动数据文件经命令DATA LIST、IMPORT 或GET 生成，可使用ALL 通配文件中的所有变量。SPSS/PC+ 最多可用变量数为200，变量名不超于8个字符。

SPSS/PC+ 对数据的管理命令可大致分为数据的定义命令、数据转换命令、记录操作、变量操作(MODIFY VARS) 和文件操作命令。在文件读写和转换中，许多参数是相同的，/KEEP 表示保留变量, /DROP 表示删除变量，/RENAME() 中含有重新命名的变量名表，有关命令及格式可详见第16章。

§5.2.3 运行控制

SPSS/PC+ 的系统设置命令为SET，有关的设置内容：如SET LISTING='CHINA. LIS'. 设置运行结果存于文件CHINA.LIS 中。当前的设置状态可用命令SHOW 来显示。运行其它软件系统命令使用DOS 和EXECUTE，如：DOS DIR。

与SAS 软件件相比，SPSS/PC+ 没有严格区分出数据步和过程步，故它的运行控制、数据管理、数据分析功能可在同一程序中灵活出现，只有少数例外。

【例5.1】现对系统检验程序BASETEST.INC的数据处理过程进行简单说明，所分析的数据是某公司雇员情况的一个调查。使用命令INC '\basetest.inc' 运行。

```
SHOW.
DATA LIST /MOHIRED YRHIRED 12-15 DEPT79 TO DEPT82 SEX 16-20
      /SALARY79 TO SALARY82 6-25 AGE 54-55 RAISE80 TO RAISE82 56-70
      /JOBCAT 6 EMPNAME 25-48 (A).
DISPLAY.
```

```
MISSING VALUES DEPT79 TO SALARY82 AGE (0) RAISE80 TO RAISE82 (-999)
               JOBCAT (9).
VAR LABELS AGE 'Age in years'.
VALUE LABELS SEX 1 'Male' 2 'Female'/
             JOBCAT 1 'Stk Clk' 2 'Admin' 3 'Sales' 4 'Mgr'.
COMPUTE GRPAGE = AGE.
RECODE GRPAGE (low THRU 25=1) (26 THRU 30=2) (31  THRU  39= 3)
              (40 THRU 49=4) (50 THRU HI=5).
VALUE LABELS GRPAGE 1 'Low - 25' 2 '26 - 30' 3 '31 - 39'
                    4 '40 - 49' 5 '50-High'.
DISPLAY GRPAGE JOBCAT.
FREQUENCIES  VARIABLES=AGE GRPAGE /FORMAT=LIMIT(10) /HBAR NORMAL
             INCREMENT(4).
DES SALARY79 TO SALARY82.
 CROSSTABS DEPT82 BY GRPAGE BY SEX/ CELLS = COLUMN NONE
            /STATISTICS = CHISQ GAMMA.
SORT CASES BY EMPNAME.
PLOT HORIZONTAL='Raise in 1982' MIN(0)/VERTICAL=MIN(0)
                  /SYMBOLS=' '/PLOT RAISE81 WITH RAISE82.
PROCESS IF (GRPAGE = 5).
LIST VARIABLES = EMPNAME SALARY80 RAISE80.
SAVE FILE='TEST.SYS'.
SORT CASES BY grpage.
TRANSLATE to DBASE4.DBF/ type=DB4/map /REPLACE.

REPORT /VARIABLES salary79 to salary82 (label)
    /BREAK grpage '年龄分组' (LABEL)
    /SUMMARY MEAN '平均值' /summary STDEV '标准差'
    /summary MINIMUM '最小值' /summary MAXIMUM '最大值'
    /summary KURTOSIS '峰度'
    /title='SPSS/PC+ BASETEST.INC 运行结果'.
```

SHOW命令显示运行环境，用DATA LIST 命令准备数据，用DISPLAY 命令显示数据结构。使用MISSING VALUE 进行缺失值定义。利用RECODE命令对年龄分组。用FREQUENCIES显示数据的频数分布。利用DESCRIPTIVES命令获得描述数据的综合统计量。使用CROSSTABS命令计算列联表统计量，从而与SAS的PROC FREQ过程对应，与SAS PROC TABULATE 对应的命令是REPORT。

使用DATA LIST命令创建的原始数据，必须放在BEGIN DATA/END DATA. 命令之间，这与SAS的"CARDS;"类似，所有与DATA LIST 有关的变量定义和说明都应放在BEGIN DATA之前。其它操作有排序、用PLOT命令图示，使用PROCESS IF 命令选择部分数据处理。结果存贮。用REPORT 命令产生报表。

利用系统提供的TRANSLATE命令，把系统文件转换成.DBF格式，为了保持数据的格式，使用FORMAT命令，如：FORMAT engcc (comma7.1).

最后，以FINISH程序结束。

§5.3 描述统计

指统计指标的计算、频数表和直方图、列联表分析，对应命令DESCTIPTIVES、FREQUENCIES和CROSSTABLES 命令，此处略作介绍。

§5.3.1 DESCRIPTIVES

格式：DESCRIPTIVE VARIABLES=变量表/STATISTICS=N /OPTION=N.

意义：显示不带频数表的描述统计量，其中的STATISTICS 和OPTION 子命令在许多命令中出现，可借助其菜单提示选择，故在下面介绍中，一般不再列出。

/VARIABLES 指示分析变量名。分析变量后加括号，可引入记录Z-值的新变量。

/STATISTICS 后加号码指示输出的统计量，省略时，SPSS/PC+ 输出均值、标准差、最小值和最大值。指示了/STATISTICS 时，仅仅得到所要求的统计量，其内容如下：

1 均数(MEAN)

2 标准误(SEMEAN)

5 标准差(STDDEV)

6 方差(VARIANCE)

7 峰度(KURTOSIS)及其标准误(SEKURT)

8 偏度(SKEWNESS)及其标准误(SESKEW)

9 全距(RANGE)

10 最小值(MINIMUM)

11 最大值(MAXIMUM)

12 观测值之和(SUM)

13 默认(DEFAULTS)统计量：均值、标准差、最小值和最大值选择13 和其它项的组合得到默认统计量和其它统计量。

ALL 上述所有统计量。

/OPTION 指示缺失值的处理方法，默认情况下使用所有有效的记录。/OPTION 的几种选项含义说明如下：

1 包含了用户所指定缺失值的记录也参加计算，命令仅当使用了MISSING 命令后方有效。

5 使用"listwise"方法排除含有缺失值的记录，在DESCRIPTIVE 命令中的任何一个变量的值出现缺失时，这个记录就废弃不用。

3 对于/VARIABLES 中所有变量在活动文件中增加Z-值，新变量取名为Z 和原始变量名中的头七个字符。

用例：

DESCRIPTIVES /VARIABLES score1 score2 score3.

DESCRIPTIVES /VARIABLES age var1 to var5 income /OPTIONS 3 5 /STATISTICS ALL.

一般说来，缺失值处理方法中LISTWISE 指示仅仅使用各个分析变量均有效的哪些记录；PAIRWISE 删除一对变量有缺失值的哪些记录；INCLUDE 括入含有缺失值的哪些记录；MEANSUBSTITUTION 用均值代替含有缺失的资料。

§5.3.2 FREQUENCIES

显示频数表、统计量、条图和直方图，其格式为：

FREQUENCIES /VARIABLES=变量列表

/FORMAT=LIMIT(N)—ONEPAGE表格输出格式

/HISTOGRAM=INCREMENT NORMAL... 直方图

/BARCHART 条图

/HBAR=INCREMENT NORMAL... 直条图

/STATISTICS 统计量选择.

其它子命令有：/GROUPED、/PERCENTILES、/NTILES、/STATISTICS、/MISSING =INCLUDE。

/FORMAT 子命令中的LIMIT(N)表示当分组数超于N时不显示表格，ONEPAGE指示将一个大的表压缩到一责内显示。/HISTOGRAM 子命令中的INCREMENT(N) 表示纵轴的间隔尺度，NORMAL表示根据本变量的均值与标准差画正态曲线。

STATISTICS 可指示均值、中位数、标准差、偏度、极差、峰度、最大值、最小值、标准误、众数、方差、偏度标准误、峰度标准误及总和，默认内容为均值、标准差、标准误，ALL将显示所有统计量。

例：FREQUENCIES /VARIABLES sex race dept.

FREQUENCIES /VARIABLES systolic diastol hemoglob

/STATISTICS MEAN SEMEAN MEDIAN MINIMUM MAXIMUM.

FREQUENCIES /VARIABLES height weight /FORMAT NOTABLE

/HISTOGRAM NORMAL.

§5.3.3 CROSSTABS

用交叉制表的方式显示变量的分布，进行关联度量，其格式为：

CROSSTABS 变量表BY 变量表...

/TABLES=变量表/OPTIONS=选项表

/FORMAT=格式定义

/CELLS=格子统计量

/STATISTICS=列联表统计量表

/MISSING=TABLE—INCLUDE—REPORT 指定缺失值处理方式

/WRITE=NONE—CELLS—ALL 结果文件.

其中的BY 可最多有十层。

格子统计量内容有：COUNT、ROW、COLUMN、TOTAL、EXPECTED、RESID、SRESID、ASRESID、ALL、NONE

列联表统计量有χ^2值、列联表系数、λ 统计量、Kendall相关等，其关键字为：CHISQ、PHI、CC、LAMBDA、UC、B
、CORR 、KAPPA、RISK、ALL、NONE。

/FORMAT 子命令选项有：AVALUE、DVALUE、LABELS、NOLABELS、NOVALLABS
、INDEX、NOINDEX、TABLES、NOTABLES、BOX、NOBOX、

CROSSTABS /TABLES= vote BY sex /STATISTICS= CHI LAMBDA.

CROSSTABS /TABLES= educatn test BY sex agegroup race.

§5.3.4 PLOT

统计做图，其命令格式为：

PLOT /FORMAT /TITLE ' ' /VERTICAL /HORIZONTAL /VSIZE /HSIZE /MISSING
~/PLOT. 如：

PLOT /PLOT Y WITH X.

PLOT /FORMAT REGRESSION /PLOT sales WITH advertis.

PLOT /PLOT income WITH age BY sex.

经GRAPH 绘图程序制记扇形图、直条图、线图、直方图和散点图并把它们传递到绘图软件中，默认是Harvard Graphics。

§5.3.5 其它命令

使用以下命令进行描述报告和报表(reports and tables)：

1. LIST 给出数据列表。

2. REPORT 产生综合统计量的报告和观察列表，它有/FORMAT、/MISSING 、/TITLE、/FOOTNOTE、/BREAK
' '几个子命令。

若指定/FORMAT ，则它应在程序中应首先出现，它有几个函数：AUTOMATIC 提供格式化选项的默认值，LIST 指示按记录列表。其它关键字控制报告各部分间的留空，大部分的扩充菜单中出现。

/VARIABLES 是必选项，对“报告变量”进行命名，每个变量的报告中定义一列。可用(VALUE)、(LABEL)、(DUMMY) 等选项指示显示的内容。注意：含有缺失值的记录在记录列表中出现，但综合统计量计算时不被包括。

/MISSING 子命令指示用户包括定义的缺失值、若超过某个指定的界值，则记录彻底被剔除。

/TITLE 子命令给输出的每页显示一个标题。可以指定LEFT、CENTER或RIGHT 使之左齐、居中、右齐或不指定。

/FOOTNOTE 子命令则用于指定每页的脚注，其选项与/TITLE 相同。

/BREAK 子命令指定一个或多个分组变量，其选项有(NOBREAK)、(TOTAL) 、(VALUE)、(LABEL)
等。

/SUMMARY 子命令指定统计量的名称，当指定/FORMAT LIST时不使用。

/OUTFILE 子命令指定结果到其它的文件中。

3. EXAMINE 提供了茎叶图、盒式图、位置的稳健估计、正态性检验以及其它描述统计量和图形，分组分析也是可能的。

其较为重要的子命令是/MESTIMATOR，可以计算M-估计量，也即位置的稳健极大似然估计。SPSS/PC+ 计算的有四个：Huber's M-估计量、Andrew's wave 估计、Hampel's M-估计、Tukey's biweight 估计量。

例：EXAMINE /VARIABLES=engsize, cost.

EXAMINE /VARIABLES=mipergal BY prototyp,prototyp BY pistons.

EXAMINE /VARIABLES=yield weevils BY field

| /COMPARE=GROUPS/PLOT=SPREADLEVEL(.5).

其图形分析包括：STEMLEAF、BOXPLOT、NPPLOT、SPREADLEVEL、HISTOGRAM、ALL、NONE，其义自显。

4. TABLES 产生高质的stub-and-banner 表，SPSS/PC+程序TBLTEST.INC 是一个用例，此处不介绍。

5. PRINT TABLES, 把TABLES 的输出在各种打印机打出。

命令使产生的表格得到尽可能高的质量的打印效果，子命令/DEVICE 指定输出设备名，使用/PORTRAIT、/LANDSCAPE、/PICA、/ELITE、COMPRESSED 定义打印特性。支持的设备如：GIBM(IBM 图形打印机)、OTHER、PIBM、HPLASER、FXEPSON、RXEPSON、LQEPSON、93OKIDATA、92OKI、TEKTRONIX。

【例5.2】不同学历的调查对象对总统能力的评价[1]，变量EDUC 表示调查对象受教育的程度，取值1,2,3,4对应高中以下、高中、大学、研究生；变量RATING 表示对总统能力的打分，取值1,2,3,4对应差、尚可、好、很好。

```
data list free/educ rating count.
TITLE 'The Performance of President'.
variable labels educ 'Education' rating 'Rating scores'.
value labels educ 1 'Less than HS' 2 'HS degree'
                  3 'College' 4 'Post graduate' /
           rating 1 'Poor' 2 'Fair' 3 'Good' 4 'Excellent'.
begin data.
1 1 4   1 2 9   1 3 10  1 4 7
2 1 5   2 2 8   2 3 22  2 4 14
3 1 20  3 2 31  3 3 11  3 4 12
4 1 9   4 2 6   4 3 8   4 4 4
end data.
WEIGHT BY count.
CROSSTABS /TABLES= educ BY rating /STATISTICS BTAU CTAU CORR GAMMA.
```

Statistic	Value	ASE1	T-value	Approximate Significance
Kendall's Tau-b	-.20098	.06158	-3.26920	
Kendall's Tau-c	-.19416	.05939	-3.26920	
Gamma	-.27112	.08218	-3.26919	
Pearson's R	-.21816	.07180	-2.98249	.00326
Spearman Correlation	-.23747	.07206	-3.26149	.00133

本例数据属于有序的情形，针对H_0: ρ =0即相关为零的假设，首先检查一致(concordance, C)与不一致(discordance, D)的观察对子，本例C= 3171 ，D= 5530，C-D 反映了两变量关联的方

向，gamma(ɣ)=(C-D)/(C+D)=(3171- 5530) /(3171+5530)=-0.27，然后计算统计量z=ɣ/SE(ɣ)=-0.27/0.082=-3.29，有显著意义。由第二章的介绍，Kendall 相关系数的含义与此相仿。

SPSS/PC+ 的CROSSTABS 语句提供了Kappa统计量，它用于比较两种评判方法的一致性，因此当列联表的行列数相同时方可得到。

设有30名受试者，其行为分为无问题、内向、外向，评判两次，现考察评判者的一致性。一致的数目是(15+3+3)=21，占21/30x100题”的方向偏。第一次有：16/30=0.53，第二次有：20/30=0.67。若两次评判是独立的，就有0.53x0.67=0.36，即应有30x0.36=10.67个“无问题”。

		JUDGE2			
	Count				
	Exp Val				
					Row
		1.00	2.00	3.00	Total
JUDGE1					
	1.00	15	2	3	20
		10.7	4.0	5.3	66.7%
	2.00	1	3	2	6
		3.2	1.2	1.6	20.0%
	3.00	0	1	3	4
		2.1	.8	1.1	13.3%
	Column	16	6	8	30
	Total	53.3%	20.0%	26.7%	100.0%

Kappa修正的思想是：假设评判独立，则对角元可以用通常的期望卡方。公式为：Kappa=(ΣO-ΣE)/[N-ΣE), 其中O表示观察数目，E是理论数目。现在ΣO=21，ΣE=10.67+1.2+1.07=12.94，Kappa=0.4。修正是分子分母各除去只是偶然引起的一致，Cohen已经造了界值表，若一致性足够低，判断就值得怀疑。SPSS/PC+ 程序如下：

```
data list free /judge1 judge2 count.
begin data
1 1 15   2 3  2
1 2  2   3 1  0
1 3  3   3 2  1
2 1  1   3 3  3
2 2  3
end data.
weight by count.
CROSSTABS /table=judge1 by judge2 /STATISTICS=CHISQ KAPPA
          /CELLS= COUNT EXPECTED.
```

计算结果：Kappa=.47266, ASE1=.13615, T-value=3.68100。

进一步的讨论可以参考：Cohen, J.(1960). A Coefficient of agreement for nominal scales. Educational and Psychological Meansurement, 100,37 -46.

§5.4 统计检验

§5.4.1 t-TEST

1.成组t-检验

T-TEST GROUPS=分类变量(k1, k2) /VARIABLES=被检验变量名(K) /OPTIONS=N.

仅指示一个K时将按照K 分成两类，指示K1, K2 则是按K1 和K2 分成两类。OPTION=1 时括入缺失值的记录，OPTION=2 删除缺失值记录，OPTION=3 不显示变量说明。

2.配对t-检验

T-TEST PAIRS=变量1 [WITH 变量2] [/PAIRS=] 变量...] [/OPTION=N].

OPTION=1,2,3 时含义与T-TEST 相仿，OPTION=4 时表示WITH 前后的量一对一比较。

3. 样本与总体的检验

现进行总体均值(POPM) 为200 的检验，则可使用语句：

COMPUTE POPM=200.

T-TEST PAIRS=A M.

§5.4.2 MEANS

MEANS 计算均值与标准差、变量总和、方差，进行单因素方差分析，格式为：

MEAN 变量表[BY 变量表] [/STATISTICS N] [/OPTION N].

在STATISTICS=1 时进行完全随机的方差分析。结果中的eta 统计量主要适于因变量为等级变量而自变量为连续性的资料。eta 平方值反映了通过已知自变量值所解释的因变量总变异的比例大小。

§5.4.3 ONEWAY

进行单因子的方差分析，其格式为：

ONEWAY 变量BY 变量[/RANGE=范围] [/OPTIONS=选项] [/STATISTICS] [/CONTRAST] [/POLYNOMIAL].

/OPTIONS 指示缺失值的处理方式、变量标签等。

/STATISTICS 指示各组的描述统计量、固定效应和随机效应统计量以及方差齐性检验。

/POLYNOMIAL 子命令是将组间平方和分解成多项式。

例：ONEWAY Y BY X(1,2) /POLYNOMIAL=2.

将平方和分解成2次多项式。

均值间的两两比较方法有SNK, Scheffe, LSD, Duncan, BTurkey, Turkey, MODLSD。POLYNOMIAL 表示平方和对应的多项式的次数

方差分析命令ANOVA 的用法与ONEWAY 类似。

【例5.3】五种方法(method)对延迟黄油变质(spoilage)的作用[1]，前两种方法属一类，后两种方法属一类，最后一种作为对照。记五种方法下，效果的均值分别为 μ 1, μ 2, μ 3, μ 4, μ 5，现比较方法的作用以及两类方法效果一样吗？H_{01}:(μ 1+ μ 2)/2=(μ 3+ μ 4)/2, 即 μ 1+ μ 2+ μ 2+ μ 4=0,

第二个检验是前四种方法与最后的对照方法进行比较, 即H_{02}:(μ 1+ μ 2+ μ 3+ μ 4)= μ 5 或0.25(μ 1+ μ 2+ μ 3+ μ 4)- μ 5=0。

现使用单因素方差分析和Kruskal-Wallis 非参检验。

```
set width=80.
title 'Comparison of five methods to retard spoilage of magarine'.
data list free/method spoilage.
begin data.
1 28   2 30   3  7   4  23   5  52
1 37   2 19   3 16   4  23   5  42
1 43   2 20   3 23   4  30   5  38
1 31   2 18   3 11   4  20   5  54
end data.
oneway spoilage by method(1,5)
       /range=LSD /range=Tukey /range=Duncan
       /contrast=.5 .5 -.5 -.5 0
       /contrast=.25 .25 .25 .25 -1.
npar tests k-w spoilage by method(1,5).
finish
```

其中NPAR TESTS K-W 是用非参数方法进行比较，运行结果如下：

```
                    Analysis of Variance
                         Sum of         Mean          F      F
        Source   D.F.    Squares       Squares      Ratio   Prob.
Between Groups    4     2526.5000      631.6250   15.7578  .0000
Within Groups    15      601.2500       40.0833
Total            19     3127.7500
```

对比系数和方差估计：

```
Contrast  1    .5    .5   -.5   -.5    .0
Contrast  2    .3    .3    .3    .3  -1.0
                            Pooled Variance Estimate
                 Value   S. Error   T Value    D.F.       T Prob.
Contrast  1     9.1250     3.1656     2.883    15.0          .011
Contrast  2   -22.8125     3.5392    -6.446    15.0          .000
                            Separate Variance Estimate
                 Value   S. Error   T Value    D.F.       T Prob.
Contrast  1     9.1250     2.9660     3.077    10.8          .011
Contrast  2   -22.8125     4.1371    -5.514     3.9          .006
```

两两比较结果，程序给出在0.05水平下三种检验的界值：

```
 LSD    3.01   3.01   3.01   3.01
```

```
HSD    4.37   4.37   4.37   4.37
Duncan 3.01   3.16   3.26   3.31
```

第J个均值Mean(J) 与第I个均值Mean(I)的差是与下面的量比较：4.4768 * Range * Sqrt(1/N(I) + 1/N(J)) 星号(*)表示两组在0.05水平上有显著差异。

```
                        LSD         HSD          Duncan
Mean      Group      3 2 4 1 5   3 2 4 1 5    3 2 4 1 5
14.2500   Grp 3
21.7500   Grp 2
24.0000   Grp 4      *
34.7500   Grp 1      * * *       *            * * *
46.5000   Grp 5      * * * *     * * *        * * * *
```

均值为一致的子集(Homogeneous Subsets)划为一组，则最高与最低均值的差不超于相应样本下的最短距离。三种检验的结果用组号表示是：

LSD法　子集一：3,2；　子集二：2,4；　子集三：1；　子集四：5。

HSD法　子集一：3,2,4；　子集二：2,4,1；　子集三：1,5。

Duncan法　子集一：3,2,4；　子集二：1；　子集三：5。

可见的确HSD法不容易出现显著。

下面是Kruskal-Wallis 检验结果：

```
Mean Rank    Cases
    14.50        4   METHOD =    1
     7.00        4   METHOD =    2
     3.75        4   METHOD =    3
     9.25        4   METHOD =    4
    18.00        4   METHOD =    5
                --
                20   Total

                                          Corrected for Ties
   CASES    Chi-Square  Significance    Chi-Square  Significance
      20       15.0429          .0046       15.1110          .0045
```

方差分析与非参检验证明五种保存方法之间的差别有统计意义，非参检验的效率要较通常的F-检验略低。

【例5.4】活产数与初婚年龄及教育程度的关系分析。

	文化程度 (例数)		
初婚年龄	高	中	低
15-19	4.17 (518)	3.65 (888)	3.27 (24)
20-22	3.70 (231)	2.95 (643)	2.88 (322)
23-24	3.60 (21)	2.12 (300)	2.68 (309)
25-34	3.15 (10)	2.68 (134)	2.45 (476)

其分析程序如下：

```
set more off length=100.
```

```
data list free /agegrp educat children count.
title '活产数与初婚年龄及文化程度的关系'.
var labels agegrp   '初婚年龄' educat '文化程度'.
val labels agegrp 1 '15-19' 2 '20-22'
                  3 '23-24' 4 '25-34'/
          educat 1 '低'  2 '中'  3 '高'.
begin data.
1 1  4.17  518  1 2  3.65  888  1 3  3.27  24
2 1  3.70  231  2 2  2.95  643  2 3  2.88  322
3 1  3.60  21   3 2  2.12  300  3 3  2.68  309
4 1  3.15  10   4 2  2.68  134  4 3  2.45  476
end data.
weight by count.
anova children by agegrp(1,4) educat(1,3) /statistics 1.
```

计算结果：

活产数与初婚年龄及文化程度的关系

* * * A N A L Y S I S O F V A R I A N C E * * *

CHILDREN

BY AGEGRP 初婚年龄

EDUCAT 文化程度

Source of Variation	Sum of Squares	DF	Mean Square
Main Effects	1463.284	5	292.657
AGEGRP	615.731	3	205.244
EDUCAT	225.004	2	112.502
2-way Interactions	70.140	6	11.690
AGEGRP EDUCAT	70.140	6	11.690
Explained	1533.424	11	139.402
Residual	.000	3864	.000
Total	1533.424	3875	.396

CHILDREN

By AGEGRP 初婚年龄

EDUCAT 文化程度

```
Grand Mean =  3.162

                                          Adjusted for
                              Unadjusted Independents
Variable + Category      N    Dev'n  Eta Dev'n   Beta
AGEGRP
    1 15-19           1430      .67        .58
```

```
    2 20-22          1196     -.09       -.08
    3 23-24           630     -.72       -.62
    4 25-34           620     -.65       -.55
                                    .90          .77
EDUCAT
    1 低              780      .84        .50
    2 中             1965     -.04       -.14
    3 高             1131     -.51       -.10
                                    .74          .40
Multiple R Squared                               .954
Multiple R                                       .977
```

总均值(3.162) 加上分组变量和协变量调整值就是所得的调整结果，随着年龄的增加或教育程度的增加，平均子女数向下调整(原表23-24岁年龄组中、高文化的影响略为不同)。eta是自变量对因变量变异的解释程度，即自变量引起的方差占原方差的百分比；beta 是控制其它因素下的影响，其值越大则自变量对因变量的影响越大，本例年龄的影响要大一些。

§5.4.4 CORRELATIONS

计算Pearson 相关系数，进行相关分析，其格式为：

CORRELATION VARIABLES= 变量表1 [WITH 变量表2] [/OPTION N] [STATISTICS N]

/OPTION=4 把相关阵以及有效记录的个数写入结果文件，供其它程序使用，这时省略WITH 语句，/OPTION=3 显示双侧概率。

/STATISTICS 1,2 分别指明单变量的均值、标准差、样本例数，以及协方差。

回归分析使用REGRESSION 命令，其用法说明详见第五节。

§5.4.5 NPAR TESTS

使用NPAR TESTS命令进行非参数分析。能够实现第二节提到的所有分析方法。

【例5.5】现针对第2章中的例2.8，SPSS/PC+程序为：

```
data list free /class scores.
begin data.
1 2.87  2 3.23  3 2.25
1 2.16  2 3.45  3 3.13
1 3.14  2 2.76  3 2.44
1 2.51  2 3.77  3 3.27
1 1.80  2 2.97  3 2.81
1 3.01  2 3.53  3 1.36
1 2.16  2 3.01
end data.
npar tests k-w=scores by class(1,3).
finish
```

【例5.6】第2章例2.9 的实现程序如下：

```
data list free/subjects coffe judge.
begin data.
1  1 1  1 2 3  1 3 2
2  1 2  2 2 3  2 3 1
3  1 1  3 2 2  3 3 3
4  1 1  4 2 3  4 3 2
5  1 2  5 2 3  5 3 1
6  1 1  6 2 3  6 3 2
end data.
npar tests friedman =judge coffe subjects.
```

运行结果：

```
 Mean Rank    Cases
      7.64        7   CLASS =    1
     14.93        7   CLASS =    2
      8.67        6   CLASS =    3
                  --
                  20   Total

                                          Corrected for Ties
     CASES    Chi-Square  Significance    Chi-Square  Significance
       20        6.1313         .0466        6.1405         .0464
```

§5.4.6 其它

第一个例子是拟合优度检验，取20个均匀分布随机数，用K-S 检验与理论相符吗？程序如下：

```
data list free /i.
begin data.
 1  2  3  4  5  6  7  8  9 10
11 12 13 14 15 16 17 18 19 20
end data.
compute x=uniform(1).
list.
npar tests k-s(uniform,0,1)=x.
```

程序首先用UNIFORM函数产生20个随机数，数据中的变量i 是给这20个随机数计数的，原始数据可见LIST命令产生的列表，检验是非参的，在括号内指定分布类型，本例还包括均匀分布的区间，因为是0-1，故也可以省略。

```
      I        X             I        X
-------------------------------------
1.00      .49           11.00      .46
2.00      .75           12.00      .67
```

```
        3.00      .16          13.00      .22
        4.00      .85          14.00      .94
        5.00      .89          15.00      .11
        6.00      .20          16.00      .79
        7.00      .39          17.00      .33
        8.00      .68          18.00      .14
        9.00      .86          19.00      .66
       10.00      .69          20.00      .60
    - - - - - Kolmogorov - Smirnov Goodness of Fit Test
         Test Distribution  -  Uniform       Range:   .00 To 1.00
         Most Extreme Differences
  Absolute       Positive       Negative          K-S Z      2-tailed P
    .15892         .06366        -.15892           .711           .693
```

结果表明双尾的P值=0.693，20 个伪随机数的分布的确与理论相符合。NPAR TESTS 命令除了K-S 检验外，SPSS/PC+可以用/EXPECTED子命令指定期望值进行分布的拟合优度检验。

下面的例子采自Stevens J.(1992). Applied Multivariate Statistics for the Social Sciences, 2nd Ed. Lawrence Erlbaum Associates, Inc. 中的例子。资料是研究儿童今后五年阅读困难的可能性。儿童按单词识别(WI)、单词理解(WC)及段落理解(PC)各五种等级打分。有两组儿童，第一组26名，是低风险的，第二组有12 名儿童，是高风险的。分析时要看两者协方差阵是否相同，可以用SPSS/PC+的MANOVA 语句，其程序如下：

```
set length=200.
title  'Check for equal covariance matrices'.
data list free /wi wc pc treats.
begin data.
5.8   9.7  8.9 1 6.2   3.0 4.3 1 5.7 10.3 5.5 1 2.4 2.1 2.4 2
10.6 10.9 11.0 1 4.2   5.3 4.2 1 6.0  5.7 5.4 1 3.5 1.8 3.9 2
8.6   7.2  8.7 1 6.9   9.7 7.2 1 5.2  7.7 6.9 1 6.7 3.6 5.9 2
4.8   4.6  6.2 1 5.6   4.1 4.3 1 7.2  5.8 6.7 1 5.3 3.3 6.1 2
8.3  10.6  7.8 1 4.8   3.8 5.3 1 8.1  7.1 8.1 1 5.2 4.1 6.4 2
4.6   3.3  4.7 1 2.9   3.7 4.2 1 3.3  3.0 4.9 1 3.2 2.7 4.0 2
4.8   3.7  6.4 1 6.1   7.1 8.1 1 7.6  7.7 6.2 1 4.5 4.9 5.7 2
6.7   6.0  7.2 1 12.5 11.2 8.9 1 7.7  9.7 8.9 1 3.9 4.7 4.7 2
7.1   8.4  8.4 1 5.9   9.3 6.2 1                4.0 3.6 2.9 2
                                                5.7 5.5 6.2 2
                                                2.4 2.9 3.2 2
                                                2.7 2.6 4.1 2
end data.
list.
manova wi wc pc by treats(1,2)/
       print=cellinfo(means,cov,cor) homogeneity(cochran,boxm).
```

程序首先印出单变量的分组均值和标准差，其次是一元方差齐性检验，方差- 协方差阵和相

关阵。矩阵齐性的Box 检验:

```
Multivariate test for Homogeneity of Dispersion matrices
Boxs M =                          14.12135
F WITH (6,2993) DF =               2.08589, P =   .052 (Approx.)
Chi-Square with 6 DF =            12.54363, P =   .051 (Approx.)
```

可见P=0.052，是近似齐性的。下面是第 4 章重复测量分析的相应程序。

```
data list free/y1 y2 y3 group.
begin data.
223 242 248 1  53 102 104 2 206 199 237 3 202 229 232 4
 72  81  66 1  45  50  54 2 208 222 237 3 126 159 157 4
172 214 239 1  47  45  34 2 224 224 261 3  54  75  75 4
171 191 203 1 167 188 209 2 119 149 196 3 158 168 175 4
138 204 213 1 183 206 210 2 144 169 164 3 175 217 235 4
 22  24  24 1  91 154 152 2 170 202 181 3 147 183 181 4
             115 133 136 2  93 122 145 3 105 107  92 4
              32  97  86 2 237 243 281 3 213 263 260 4
              38  37  40 2 208 235 249 3 258 248 257 4
              66 131 148 2 187 199 205 3 257 269 270 4
             210 221 251 2  95 102  96 3
             167 172 212 2  46  67  28 3
              23  18  30 2  95 137  99 3
             234 260 269 2  59  76 101 3
                           186 198 201 3
end data.
sort by group.
report vars=y1 to y3 /break=group
       /summary=mean /summary=STDEV.
manova y1 to y3 by group(1,4) /transform=repeated
       /rename=average dif2and1 dif3and2
       /print=transform
       /analysis=(dif2and1 dif3and2/average)
       /design.
```

REPORT 命令求出各组的均值与标准差，REPORT命令在本章中已经使用。MANOVA 结果首先是转换矩阵的转置(从略)。第一部分是轮廓平行的检验。

```
EFFECT .. GROUP
Multivariate Tests of Significance (S = 2, M = 0, N = 19 )
Test Name         Value  Approx. F Hypoth. DF   Error DF  Sig. of F
Pillais          .05251     .36850       6.00      82.00       .897
Hotellings       .05394     .35061       6.00      78.00       .908
Wilks            .94817     .35956       6.00      80.00       .902
```

```
Roys             .02867
```

接受轮廓平行假设。第二部分是轮廓水平一致的检验，拒绝假设。

```
EFFECT .. CONSTANT
Multivariate Tests of Significance (S = 1, M = 0, N = 19 )
Test Name         Value  Approx. F Hypoth. DF   Error DF  Sig. of F
Pillais          .59687   29.61176       2.00      40.00       .000
Hotellings      1.48059   29.61176       2.00      40.00       .000
Wilks            .40313   29.61176       2.00      40.00       .000
Roys             .59687
```

第三部分是轮廓重合检验(变量AVERAGE)。

```
Tests of Significance for AVERAGE using UNIQUE sums of squares
Source of Variation          SS      DF        MS         F  Sig of F
WITHIN CELLS          206430.49      41   5034.89
CONSTANT              975937.26       1 975937.26    193.83      .000
GROUP                  24218.29       3   8072.76      1.60      .203
P>0.05
```

§5.5 多元统计分析

§5.5.1 回归及其残差分析

命令格式：

REGRESSION
~/VARIABLES 指定分析回归分析的变量
/DESCRIPTIVE 均值、标准差、相关矩阵等统计量。
/SELECT 选择分析所用的记录。
/MISSING 指定缺失值的处理办法。
/STATISTICS 计算统计量。
/CRITERIA 指定回归分析准则。
/REGWGT 指示回归中的权。
/ORIGIN 使回归线通过原点。
/NOORIGIN 关闭ORIGIN，使回归不通过原点。
~/DEPENDENT 指定回归分析的因变量。
~/METHOD 指示变量的筛选准则
/RESIDUALS 回归残差。
/CASEWISE 按记录给出的统计量。
/SCATTERPLOT 产生一个或几个散点图。
/PARTIALPLOT 偏回归残差图。
/SAVE 存贮残差分析结果。
命令规则：(1) 必选项为VARIABLES、DEPENDENT 和METHOD 子命令；(2) VARIABLES 只能使用一次且应置于程序开始，(3) DEPENDENT 子命令可使用多次，

对每个DEPENDENT 子命令，估计一个方程；(4) DEPENDENT必须紧接一个或多个METHOD 子命令；(5) MISSING、DESCRIPTIVE 和SELECT 子命令在满足(1)、(4)的条件下可在任意位置出现；(6) CRITERIA、STATISTICS 和ORIGIN 子命令在被替代之前对其后所有方程有效；(7) 所有子命令以斜杠分开。

主要子命令及其用法：

1. VARIABLES

指定参与分析的变量的变量名，默认值为/VARIABLES (COLLECT) , 即/DEPENDENT与/METHOD 子命令中所有的变量，使用了/VARIABLES (COLLECT) , 必须在ENTER之后给出一些自变量，如：REGRESSION /VARIABLES Y x1 to X10 /dependent y /method enter /method remove x8 x10.

2. DEPENDENT

指定因变量名。可多次使用，且每次必须在其后紧接一个以上的METHOD 子命令，所指定的因变量名必须是在VARIABLES 中已定义过的。

例：REGRESSION VARIABLES=X1 TO X5，Y /DEPENDENT=Y. 表示以Y 为因变量，X1 互X5 为自变量参与回归分析。

3. METHOD

指定一个变量选择方法，BACKWARD 、FORWARD、STEPWISE、ENTER 表示后退、前进、逐步及全部变量入选。

前进法(FORWARD) 的实施步骤：首先选取与因变量相关系数绝对值最大的变量作为备选变量。同时，还应当对该备选变量的系数为零的假设作F 检验以决定该变量是否的确应入选。REGRESSION提供了两种准则：

准则一(FIN)：一个变量入选则最小应达到的F统计量值(小于FIN 值不入选)。在REGRESSION中，其关键字为FIN，默认值为3.84。

准则二(PIN): 一个变量入选则最大不能超过超过的(该F统计量相应的) 概率值(大于PIN 值不入选，小于才能入选)，其关键字为PIN，默认值为0.05。

对准则一、二在程序中首先指定其一，若没有变量满足入选准则，则所有变量均不进入方程。

若第一个变量已进入方程，则前进法继续进行变量选择。此时，首先计算尚未进入方程的那些自变量与因变量的偏相关系数，绝对值大者作为下一个候选者，并考察是否满足指定的准则，若满足则入选。

前进法进行到到没有变量可进入方程为止(变量进入方程时，还必须考察其容许性)。

⑵后退法(BACKWARD)与前进法相反，后退法首先将所有变量包括进方程，然后逐个将不合要求者删除。REGRESSION 提供了两种变量删除准则供选择使用。

准则一(FOUT)：一个变量要留在方程中最小应达到的F 值(小于FOUT值则删除，大于则留下)，其关键字为FOUT，默认值为2.71。

准则二(POUT)：一个变量要留在方程中最大能具有的概率值(大于POUT值则删除，小于则留下),其关键字为POUT，默认值为0.10。

⑶逐步选择法(STEPWISE)是前进法后退法的组合，其实施步骤如下：

第一个变量入选方法与前进法相同，若无一变量满足入选准则，则终止；第二变量入选与前进法相同；每次有一个新的变量入选后，都应考察前面已入选变量是否满足删除准则而被删除；继续考察是否有方程之外的变量应入选；当既无变量入选又无变量可删除时，终止。

注意，为防止同一变量反复入选——删除，应保证PIN ¡POUT 或FIN¿FOUT。
4. STATISTICS
控制关于方程和自变量统计量的输出显示，当STATISTICS 子命令缺损时或未指定任何关键字果的输出，包括R、ANOVA、COEFF 和OUTS。
ALL 输出除F，LINE 和END 外的全部统计量。
R 包括R^2，校正R^2 和这些估计的标准差。
ANOVA 方差分析表，包括回归平方和，残差平方和，F 及其概率等。
CHA R^2 的改变量。
BCOV 未标准化回归系数的方差——协方差矩阵。
XTX 矩阵X'X。
COND 条件数的界，包括已入选方程变量构成有矩阵X'X 的条件数的上下界。
COEFF 回归系数，包括回归系数及其标准误差、标准化回归系数、t 一值等。
OUTS 关于尚未入选变量的统计量。
ZPP 零阶、部分和偏相关。
CI 未标准化系数的95SES 标准化回归系数的近似标准误差。
TOL 容许性，包括已入选变量和未入选变量的容许性，以及下一个即将入选变量的容许性。
F 回归系数的F 值及其概率。
5. CRITERIA
控制建立回归方程时的统计准则，关键字：DEFAULT。当CRITERIA 子命令缺损时的默认值，当准则已被改变时，可用DEFAULT 恢复默认值。
PIN(值)，变量入选的F 概率，默认取值0.05。
FIN(值)，变量入选的F 值，其默认值为3.84，PIN 和FIN 只须指定一个。
POUT(值)，变量删除的概率值，其默认值为0.01。
FOUT(值)，变量删除的F 值，默认值为2.71，POUT 和FOUT 只能指定一个。
TOLERANCE (值)，容许性，默认值为0.0001。
MAXSTEPS(N) 最大步数，其默认值:
后退法或前进法: 满足PIN/POUT或FIN/FOUT的变量个数。
逐步选择法: 自变量个数的两倍。
例: REGRESSION VARIANLES=X1 TO X5，Y
/CRITERIA=PIN(.1) POUT(.15) TOL(.001)
/DEPENDENT=Y /METHOD=BACKWARD.
表示变量进入和删除标准都比默认值松。
6. ORIGIN 和NOORIGIN
控制是否对数据作中心化(即方程中是否包括常数项) 。必须放在其修饰的DEPENDENT和METHOD 之前，其默认值为NOORIGIN–表示方程中包括常数项。
例：REGRESSION VAR=V1 TO V3, Y, Z /DEPENDENT=Y /METHOD=FORWARD
/ORIGIN /DEP=Z /METHOD=FORWARD。
表示第一个和第二个回归方程分别不作中心化变换和作中心化变换。
7. DESCRIPTIVES
输出显示变量的描述统计量，关键字为：
NONE 不作任何输出，也是DESCRIPTIVES 省略时的对应项。

DEFAULTS 输出MEAN,STDDEN 和CORR。
MEAN 变量均值。
STDDEN 变量标准差。
VARIANCE 变量方差。
CORR 相关矩阵。
SIG 相关系数的单边概率。
BADCORR 当某些系数不能计算时，输出相关矩阵。
COV 协方差矩阵。
XPROD 对均值离差的交叉积。
N 用于计算相关系数的观测个数。
ALL 显示所有描述统计量。
8. MISSING
指示缺失值的处理办法，关键字为：
LISTISE 默认，删除在/VARIABLES 中出现缺失时的任何记录。
PAIRWISE 分别计算相关系数，但这样做有时会出现一些不太可能的结果。
MEANSUBSITUTION 缺失值用变量均值代替，这样会影响相关系数和预测值。
INCLUDE 指定缺失值为有效。
程序用例：

```
REGRESSION VARIABLES=X1 TO X5, Y /DEPENDENT=Y /METHOD=ENTER X1
X2 X3.
REGRESSION VARIABLES=X1 TO X5, Y /DEPENDENT=Y /METHOD=STEPWISE.
```

第一行程序为以变量X1,X2 和X3 为自变量的回归分析。其余所有子命令均取其默认值,程序将输出，方差分析表，回归系数等有关统计量；第二行程序为逐步回归，采用逐步选择法选择变量。
在输出结果中，B 所在列为回归系数，SEB 所在列为回归系数的标准误差，Beta 所在列为标准化回归系数，T 所在列为相应的t 值，SigT 为t 的双边概率。
残差分析：(1) 作为可选项的子命令RESIDUALS、CASEWISE、SCATTERPLOT 和PARTIALPLOT必须紧接着某个方程的最后一个METHOD 子命令，当拟合多个方程时, 可对每一方程作一次残差分析。(2) 残差子命令可以以任意顺序设置。(3) 残差子命令只影响它们紧接着的方程。(4) 采用矩阵输入时，不得使用残差命令。
分析中，REGRESSION 计算12 个临时变量如PRED、RESID，等。
主要子命令简述如下：
1. RESIDUAL
控制异常点信息的显示和标记，对临时变量输出Durbin-Watson 统计量，直方图和正态概率图。
关键字：DEFAULTS 也是RESIDUALS 不选任何关键字时的默认值，在各量后的括号内示意出来，包括SIZE(LARGE)DURBIN，NORMPROB(2RESID) ，HISTOGRAM (2RESID)，OUTLIER(2RESID)。
SIZE() 指示图的尺寸，取值为LARGE 或SMALL。
HISTOGRAM() 标准化临时变量的直方图。
NORMPROB() 标准化值的正态概率图(P-P 图)。
OUTLIER() 指定的临时变量最为显著的10个异常点。

Durbin-Watson 检验统计量。
ID(变量名) 异常点图上的观测个体的标识。
用例：/RESID=DEFAULT 表示对变量ZRESID 作正态概率图、直方图，给出Durbin-Watson 统计量和异常点图。
2. CASEWISE
对所指定的临时变量产生残差逐点图，关键字为：
DEFAULTS 关于标准化残差绝对值大于3.0 时的记录的图示，如果显示宽度足够，标准化因变量及其预测值随图列出。DEFAULT 包括OUTLIERS(3) , PLOT(ZRESID), DEPENDENT, PRED和RESID。
PLOT(临时变量) 绘制除标准化残差以外的记录图示，可供的选择有删除残差、学生化残差和学生化删除残差。
OUTLIERS(值) 给定记录图示新的异常值界值，默认值为3.0。
ALL 对所有的记录绘图，并不仅仅是残差超于界值的记录。
DEPENDENT 因变量。
PRED 预测值。
RESID 残差。
ZPRED 标准预测值。
ADJPRED 调整的预测值。
ZRESID 标准残差。
DRESID 删除残差，其计算方法是因变量减去其预测值，预测值由除去本记录以外的记录算得。
SRESID 学生化残差，即因变量减去其预测值，并除以预测值的标准差，标准差依自变量而变。
SDRESID 学生化删除残差，即除去本记录后回归方程的学生化残差。
SEPRED 预测值标准误。
MAHALANOBIS 反映观察影响回归的程度，用观察与自变量平均值的距离来度量。
COOK 反映观察对回归的影响，用观察不参加回归时所有残差的改变来表示。
LEVER 杠杆值，也能反映记录对回归的影响，与Mahalanobis' 距离相关。
DFBETA 第i 个观察删除后回归系数的改变情况。
SDBETA 标准化的DFBETA。
DFFIT 第i 个记录删除后模型拟合上变化情况。
SDFIT 标准化DFFIT。
COVRATIO 第i 个记录删除后协方差行列式的改变。
MCIN 因变量平均响应的预测下界LMCIN 和上界UMCIN。
ICIN 单个观察的预测区间下界LICIN 和上界UCIN。
3. SCATTERPLOT
指定一对变量并输出其散点图。对于每对变量名，前者为纵坐标，后者为横坐标，对临时变量名前应加上* 号，所有散点图中的变量均应为标准化的(所以指定*RESID与指定*ZRESID 一样)
例：/SCATTERPLOT (*RES,*PRE) (*RES, V1) 产生两个散点图,一个是残差——预测值图,一个是残差——变量V1 图。
4. PARTIALPLOT

产生偏残差图。若PARTIAL 使用时不作任何变量指定，则对方程中每个自变量均产生一个偏残差图，也可在PARTIALPLOT 之后指定欲作图的自变量，则此时只对指定变量作偏残差图。

5. SAVE

存贮生成的临时变量，由紧随关键字后的括号指定生成的变量名，如：RESID() 和ZRESID()。另外，FITS() 用于存贮DFFIT，SDFIT,DFBETA,SDBETA 和COVRATIO。

典型程序：

```
REGRESSION VARIABLES V1 TO V3 Y
/DEPENDENT=Y METHOD=ENTER RESIDUALS
/SCATTERPLOT (*RESID, *2PRED) /PARTIALPLOT.
```

【例5.7 】第 4 章的汽车数据分析

```
* NOTE: A transportation data.
DATA LIST FREE/ X1 X2 Y.
* FORMATS Y (F6.3).
BEGIN DATA.
 1300        .45       .066    948          2       .005
 1444         .5       .076   1440        2.4       .011
  736        1.5       .001   1080          3       .003
 1652         .4        .17   1844          1        .14
 1736         .8       .156   1116        2.8       .039
 1754         .8        .12   1656       1.45       .059
 1200        1.8        .04   1536        1.5       .087
 1500         .6        .12    960        1.5       .039
 1200        1.7         .1   1784         .9       .222
 1476        .65       .129   1496        .65       .145
 1820         .4       .135   1060       1.83       .029
 1436          2       .099
END DATA.
LIST.
regression /variables y x1 x2 /dependent y /method=enter
           /CASEWISE DEPENDENT PRED RESID ZPRED DFBETA COVRATIO.
compute ty=(y**0.6-1.0)/0.6.
regression /variables ty x1 x2 /dependent ty /method=enter
           /CASEWISE DEPENDENT PRED RESID ZPRED DFBETA COVRATIO.
```

§5.5.2 对数线性模型

命令格式：

```
HILOGLINEAR
~variable list 指定分析变量。
/PRINT 指定打印结果。
```

/PLOT 残差图。
/MAXORDER 限定最高交互项。
/CRITERIA 改变收敛准则和最大迭代次数。
/METHOD BACKWARD 改变默认的向前法。
/MISSING INCLUDE 使MISSING VALUE 引入的缺失值纳入分析。
/DESIGN 指定模型。
命令规则: (1) 程序的必选项为定义具有至少两个变量的变量列表，每个变量之后给出其最小和最大取值，其余子命令均为可选项。(2) 变量必须在程序最开始定义。(3) METHOD、PRINT、PLOT、CRITERIA、MAXORDER 和CWEIGHT 子命令必须置于它们要修饰的DESIGN 子命令之前。(4) 可指定多个METHOD 子命令，但每个仅影响下一个DESIGN 子命令。(5) 子命令之间应以斜杠分开。
下面将主要子命令简述如下：
1. VARIABLE LIST (变量列表)
定义参与分析的变量。变量必须取整数值。
例：HILOGLINEAR V1(1,2) V2(1,3) V3(1,4) 表示分析由变量V1、V2 和V3 构成的2 x 3 x 4 列联表。
2. METHOD
对其后的DESIGN 子命令，指定采用后退删除进行模型选择。METHOD 缺损时，所有变量均进入模型。然后把P值小于0.05的去掉。关键字为BACKWARD，只影响下一个DESIGN。
3. MAXORDER
控制在其后的DESIGN 中模型的最高阶次，例：
HILOGLINEAR v1 v2 v3 (1,3) /MAXORDER=2 表示对于变量V1、V2 和V3 构成的列联表，拟合的模型的最高项为两两交互项。
4. CRITERIEA
对其后的DESIGN，改变迭代拟合模型选择的迭代停止规则。关键字为：
CONVERGE(n) 收敛准则，其默认值为0.25，当被拟合的频数的改变量小于指定值时停止迭代。
ITERATE(n) 迭代最大次数，其默认值为20。
P(prob) 模型的卡方概率，默认值为0.05，仅当指定BACKWARD 方法时有效。
MAXSTEPS(n) 最大步数，默认值为10，仅当指定BACKWARD 方法时才有效。
DEFAULT 用来把CRITERIA 中的关键字的参数改变为其默认值。
5. CWEIGHT
含义：对一个模型指定各格点的权重，通常被用于指定列联表中的结构零。
用法：有下列三种方法指定权重。
. 指定一个变量名，以该变量的取值为格点权重。
. 直接提供一个格点权重矩阵，权重按变量列的顺序由左至右取值。
. 在方法2 中，可使用n*CW 表示权重CW 重复n 次。
例HILOGLINEAR V1(1,2) V2(1, 3) /CWEIGTH= CELLWGT 表示权重由变量CELLWGT 的取值给出。
HILOGLINEAR V1(1,2) V2(1,3) /CWEIGHT=(1 1 1 1 0 1 1 1 0) 或等价地使用/CWEIGHT=(0 3*1 0 3*1 0) 表示了对角元结构零，即：

	V2		
V1	0	1	1
	1	0	1
	1	1	0

6. PRINT

对其后的DESIGN 控制输出显示，关键字为：

FREQ 频数，给出观测和期望格点频数。

RESID 残差，给出原有和标准化残差。

ESTIM 饱和模型的参数估计(对其它模型该选择无效)。

ASSOCIATION 饱和模型效应的偏相关。

DEFAULT PRINT 缺损时的默认显示，包括显示FREQ、RESID 和所有模型以及饱和模型的ESTIM。

ALL 全部显示。

/PRINT 影响到后面的模型输出。

7. PLOT

含义：对其后的DESIGN，给出残差图，关键字为：

RESID 观测和期望频数的标准化残差。

NORMPLOT 调整后残差的正态概率图。

NONE 不做任何图。

DEFAULT PLOT 缺损时的默认图形显示，包括DESIGN 和NORMPLOT。

ALL 给出全部图形显示。

8. DESIGN

其缺损值时计算包括变量列表中所有变量在内的饱和模型，使用DESIGN 指定该饱和模型不同的生成类。将最高阶效应项列出(使用变量名和* 号表示交互效应项)。一个DESIGN 只估计一个模型，可多次使用DESIGN 子命令。

例：HILOGLINEAR V1(1,2) V2(1,2) V3(1,3) /DESIGN=V1*V2, V3

表示将对变量V1、V2 和V3 建立一个2 x 2 x 3 列联表。按照DESIGN 子命令将产生一个包括全部主效应和包括V1 和V2 交互项的模型。

【例5.8】下面是例6.2的(DV,DP,VP)模型的程序：

```
data list free/ p v d count.
value labels p 1 'yes'   2 'no'/
             d 1 'white' 2 'black'/
             v 1 'white' 2 'black'.
begin data
1 1 1 19   1 1 2 0  1 2 1 11  1 2 2 6
2 1 1 132  2 1 2 9  2 2 1 52  2 2 2 97
end data.
weight by count.
hiloglinear p(1,2) d(1,2) v(1,2) /design d*v d*p v*p.
```

部分结果：观察频数、期望频数及残差：

```
Factor    Code         OBS count  EXP count  Residual  Std Resid
  P       yes
   D       white
    V       white          19.0       18.7       .33        .08
    V       black          11.0       11.3      -.32       -.09
   D       black
    V       white            .0         .3      -.33       -.57
    V       black           6.0        5.7       .32        .13
  P       no
   D       white
    V       white         132.0      132.3      -.32       -.03
    V       black          52.0       51.7       .30        .04
   D       black
    V       white           9.0        8.7       .32        .11
    V       black          97.0       97.3      -.30       -.03
```

拟合优度统计量：

```
Likelihood ratio chi square =      .70080    DF = 1  P =  .403
         Pearson chi square =      .37446    DF = 1  P =  .541
```

LOGLINEAR 的句法与HILOGLINEAR类似，注意非层次模型没有层次模型那样的包含关系。其子命令简介如下：BY 指示模型中的主效应变量。WITH 指示模型所用的非表格形式的协变量。/CONTRAST () 指示对照的方法，括号内为对照的因素名。DEVIATION 与总效应比较。DIFFERECE 是各水平与其前水平的均值比较。HELMERT 是各水平与后面水的平均比较。SIMPLE 用每效应最末的水平作为标准。REPEATED 指示相邻水平间的比较。PLOYNOMIAL 指示为多项式，在平衡设计中是正交多项式。SPECIAL 和BASE SPECIAL用户自定义的对照。/CRITERIA 收敛控制。CONVERT()是收敛的精度，默认值为0.001。ITERATRE()为最大迭代次数，默认为20。DELTA()指示迭时每格子加上的值，默认值为0.5 。DEFAULT为默认值。/PLOT 结果的图示，残差、去趋势正态图。/PRINT—NOPRINT指示ESTIM、COR、RESID、FREQ、FREQ、DES设计的主效应。James Stevens 的例：数据是关于电视网络好坏的调查。年份为1959、1971，对象为白人或黑人，结果有“好”、“一般”、“差”三种。

```
data list free/year color response freq.
value labels year     1 '1959'  2 '1971'  /
             color    1 'black' 2 'white' /
             response 1 'good'  2 'fair'  3 'poor'.
begin data.
1     1     1    81    2     1     1   224
1     1     2    23    2     1     2   144
1     1     3     4    2     1     3    24
1     2     1   325    2     2     1   600
```

```
1     2     2   253    2     2     2   636
1     2     3    54    2     2     3   158
end data.
weight by freq.
loglinear response(1,3)  by color(1,2) year(1,2)/
          criteria=delta(0)/
          print=default estim/
          contrast(response)=special(1 1 1 1 -0.5 0.5 0 1 -1)/
          design.
```

结果：

```
Analysis of Dispersion
                                  Dispersion
  Source of Variation       Entropy  Concentration      DF
     Due to Model            32.854         24.316
     Due to Residual       2338.172       1438.487
     Total                 2371.026       1462.803    5050
Measures of Association
         Entropy =   .013857
   Concentration =   .016623
Estimates for Parameters
RESPONSE
 Parameter      Coeff.   Std. Err.     Z-Value Lower 95 CI Upper 95 CI
       1  1.2829846562      .08466    15.15509     1.11706     1.44891
       2  1.4512383652      .10809    13.42659     1.23939     1.66309
RESPONSE BY COLOR
       3   .4526490203      .08466     5.34686      .28672      .61858
       4   .3018183041      .10809     2.79237      .08997      .51367
RESPONSE BY YEAR
       5   .2951119842      .08466     3.48597      .12918      .46104
       6   .1612112793      .10809     1.49150     -.05064      .37306
RESPONSE BY COLOR BY YEAR
       7   .1028092054      .08466     1.21442     -.06312      .26874
       8   .0271094119      .10809      .25081     -.18474      .23896
```

§5.5.3 LOGISTIC 回归

命令格式：

LOGISTIC REGRESSION
~/VARIABLES 回归因变量和自变量。
/CATEGORICAL 指定名义的或有序的自变量。
/CONTRAST 在/CATEGORICAL 子命令指定为分类变量的对比类型。
/METHOD 选择变量的方法。

/SELECT 选择部分记录进行分析。
/ORIGIN 强迫回归线过原点。
/PRINT 选项DEFAULT 打印变量的分类表及统计量。
/CRITERIA 决定估计何时停止。
/CLASSPLOT 每步因变量实际值与预测值的分类图示。
/MISSING INCLUDE 指定包括缺值。
/CASEWISE 按照记录列出预测值、残差和其它暂存量。
/ID 对记录列表时，标识记录的变量。
/SAVE 对活动数据集增加预测变量。
/EXTERNAL 分析时将结果存放在外部暂存文件，节省内存的使用。
该命令的用法与线性回归命令REGRESSION类似，主要子命令说明如下：
1. VARIABLES
指定模型中的变量，为必选项。
2. CONTRAST
指定因变量对比类型，关键字为：
DEVIATION 即各效应与变量最后分类的偏差，亦是默认值。
DIFFERENCE 因素的每个水平与其前面的水平的平均效应比较。
HELMERT 因素每个水平与其后面水平的平均效应相比较。
SIMPLE 因素每水平与省略的或“参考”的水平相比，比较不是正交的。
REPEATED 是因素相邻各水平之间的比较，除第一个水平外，因变量的每个分类与其前面的分类相比较。
POLYNOMIAL 第一自由度包含因子水平间的线性效应，第二自由度包含二次效应，等等。
INDICATOR 指示分类成员的出现或不出现。
SPECIAL 用户指定，紧跟的可以是一个(k-1)xk方阵，k是因变量的水平。
3. CRITERIA
控制收敛的准则。关键字为：
BCON() 回归系数的改变, ITERATE 迭代次数，LCON()对数似然的改变，PIN()，POUT()，EPS()是进出的概率值和redundancy 检查。
4. METHOD
建模方法，关键字为：
FSTEP() 括号内指示WALD或LR，向前法逐步回归。删除变量使用Wald 统计量或似然比统计量。BSTEP() 用法与FSTEP类似，可以指示WALD或LR，向后法删除变量。
5. SELECT
对子集分析，变量间可使用关系运算符：EQ,NE,LT,LE,GT,GE。
6. PRINT
关键字为：default 对每个/METHOD 子命令，显示分类表、进入方程变量的统计量或虽虽然未进入方程但在/METHOD 或/VARIABLES 子命令指定过的变量。summary 与DEFAULT效果相同，只是在最后一步给出结果。corr 给出估计量间的近似相关。iter 每迭代步上参数估计值。all 给出所有的输出。
7. CRITERIA

控制收敛的准则，其关键字为：
BCON() 回归系数B 的改变量。ITERATE() 最大迭代次数，默认值为20 。LCON() 对数似然下降的百分比，默认值为0.0001。PIN() 变量入选计分统计量的概率，默认值为0.05。POUT() 变量删除的概率值，默认值为0.1。EPS() 冗余检验(redundancy checking)的精度，取值范围为大于10E-12到小于等于0.05，默认为10E-8，该检验避免变量的线性组合入选方程。
8. CASEWISE
PRED 预测概率。PGROUP 预测分组残差。观察组编码为0 到1, 该值等于观察组别减去在第二组的预测概率。RESID 即残差。DEV 为离真度。LRESID 为logit 残差。SRESID 为学生化残差。ZRESID 标准化残差。LEVER 杠杆。COOK 为COOK氏距离。DFBETA 为删除该记录后回归系数的改变。OUTLIER() 是控制标准化残差SRESID 大于某个数值时方显示。
9. SAVE
其关键字与CASEWISE 类似，每个关键字后可紧随一括号指出存贮量新的名称。

【例5.9】第 4 章的LOGISTIC 回归分析用例

```
data list free / age sex DM SD CHD freq.
variable labels
         age  'AGE IN YEARS'
         sex  'SEX (0=FEMALE, 1=MALE)'
         DM   'DIABETES MELLITUS (0=NO, 1=YES)'
         SD   'SEX*DIABETES  (INTERACTION TERM) '
         CHD  'CORONARY HEART DISEASE (0=NO, 1=YES) '
         freq 'NUMBER OF OBSERVATIONS'.
begin data.
50 1 0 0 0 6434  50 0 0 0 0 8519  60 1 0 0 0 4298  60 0 0 0 0 6199
50 1 0 0 1  124  50 0 0 0 1   45  60 1 0 0 1  179  60 0 0 0 1  116
50 1 1 1 0  193  50 0 1 0 0  159  60 1 1 1 0  218  60 0 1 0 0  228
50 1 1 1 1    6  50 0 1 0 1    5  60 1 1 1 1   13  60 0 1 0 1   10
end data.
TITLE 'LOGISTIC REGRESSION example from Chapter four'.
weight by freq.
logistic regression /variables CHD with age sex DM SD
                    /METHOD ENTER.
```

§5.5.4 因子分析

命令格式：

FACTOR
~/VARIABLE 列出FACTOR 命令所要求的因子分析的所有变量。
/MISSING 缺失值的处理方法。
/ANLYSIS 指示部分变量分析。

/PRINT 输出显示控制。
/PLOT 抽取因子的图示。
/FORMAT 因子矩阵的显示格式。
/DIAGONAL 指定相关阵对角元的初始共因子方差估计。/CRITERIA 指定因子提取和旋转的准则。
/EXTRACTION 因子抽取方法。
/ROTATION 因子旋转方法。
/SAVE 存贮因子得分到活动文件。
规则：(1) 只有VARIABLES 子命令为必选项；(2) VARIABLES, MISSING 和WIDTH 为全局子命令，对整个FACTOR 程序有效并且只能使用一次，VARIABLE和MISSING必须最先设置。
子命令用法如下：
1. VARIABLES
含义：指定参与因子分析的变量。
用例：FACTOR VARIABLES = V1 TO V10 指示变量V1～V10 参与分析。
2. ANALYSIS
含义：指定VARIABLE 变量集内的一个子集作为分析之用。
用法：可建立多个ANALYSIS 模块进行多次分析，在ANALYSIS 之后列出该模块的变量名，每一个ANALYSIS 作为一个模块开始，待下一个ANALYSIS 出现或FACTOR的结束作为其结束。
例：FACTOR VARIABLES =V1 TO V5 /ANALYSIS=V1 TO V3 /ANALYSIS=V3 TO V5。
指定V1～V3 和V4～V5 分别作为模块进行因子分析。
3. FORMAT
含义：重新设置因子载荷矩阵，关键字为：
SORT 以因子载荷递减方式排列因子载荷矩阵；BLANK(n) 删除因子载荷矩阵中绝对值小于指定值的系数；DEFAULT 取消SORT 和BLANK(n) 的设置。
例：FACTOR VARIABLES=V1 TO V5 /FORMAT=SORT BLANK(0.2) 表示因子载荷矩阵中以载荷递减顺序排列，且删除所有绝对值小于0.2 的载荷。
4. PRINT
控制一个分析模块中的统计输出显示，关键字为：
UNIVARIATE 显示有效观测个体数、均值、标准差；INITIAL 输出每个变量的初始公共因子方差，每个因子的相关矩阵特征根、方差百分比；CORRELATION 输出显示相关矩阵；SIG 输出相关系数的显著性水平；DET 输出相关阵的行列式值；INV 输出相关矩阵的逆矩阵；AIC 输出反象相关阵；KMO 输出Kaiser- Meyer -Olkin 指数和Bartlett 检验；EXTRACTION 输出因子载荷矩阵，每个因子的特征根和方差百分比；ROTATION 输出旋转矩阵的因子载荷矩阵和变换矩阵；FSCORE 输出因子得分系数矩阵(用回归方法获得的) ；ALL 输出所有得到的统计量；DEFAULT 相当于输出INITIAL, EXTRACTION 和ROTATION 所包含的显示内容。
例：FACTOR VARS=V1 TO V6 /PRINT=DET FSCORE 表示既要输出DEFAULT 的内容，还要输出因子得分系数。
5. PLOT

控制图形显示，关键字：
EIGEN scree 图，以降序方式输出特征根分布图；ROTATION 对每个旋转给出图示空间中变量位置分布图，括号内指定图中因子轴对应的因子。如：
FACTOR ... /PLOT ROTATION (1,2) (1,3) (2,3) ... 其中的数字表示图轴所使用的因子。
6. CRITERIA
指定因子提取和旋转的准则，关键字：
FACTOR (因子数) 提取的因子个数，其默认值是特征根大于等于MINEIGEN 的个数；MINEIGEN(值) 控制因子提取的最小特征值(只提取其对应的特征值大于给定值的因子)，其默认值为1；ECONVERGE(值) 采用迭代法提取因子时的迭代收敛准则，其默认值为0.001；ITERATE(迭代次数) 因子提取或旋转求解过程中的迭代次数，其默认值为25；RCONVERGE(值) 旋转迭代的收敛准则，其默认值为0 .0001；Kaiser 旋转时的Kaiser 正规化，这也是其默认值，由NONKAISER 废弃；DELTA(值) 斜交旋转的 δ，仅当指示了/ROTATE OBLIMIN 以后使用，其默认值为0。；DEFAULT 将所有准则恢复为默认值。
例：FACTOR VARS=V1 TO V6 /CRITERIA=FACTORS(3) 表示提取三个因子。
7. EXTRACTION
指定因子提取的方法，关键字：
PC 主成分方法(默认方法)；PAF 主轴因子法；ALPHA α 方法；IMAGE 象因子法；ULS 不加权最小二乘法；GLS 广义最小二乘法；ML 极大似然法。
DIAGONAL 子命令指定/EXTRACTION PAF 中相关阵对角元的初始共因子方差估计，默认初始共因子方差估计值为复相关平方和(SMC)。
8. ROTATION
指定旋转方法，关键字：
VARIMAX 方差极大旋转法(是默认值) ; EQUMAX 使用equamax 旋转; QUARTIMAX 使用quartimax 旋转法；OBLIMIN 斜交旋转；NOROTATE 不作旋转。
例：FACTOR VARS= V1 TO V5 /EXTRACTION= GLS /ROTATION /ROTATION OBLIMIN 表示采用加权最二乘法提取因子，第一次旋转用VARIMAX 法，第二次使用斜交旋转。
9. SAVE
指定因子得分的计算方法，并将因子得分以新变量形式存贮到当前文件，关键字(计算因子得分的方法)：
REG 回归方法；BART 使用Bartlett 方法；AR 使用Anderson-Rubin 方法；DEFAULT 默认值(回归方法)，在关键字后面，紧接着置于括号内的要存放的因子得分个数和根名字，其中所指定的数字不能超过所能获得的因子个数，也可用ALL 代替该数字。必须紧随/ROTATE, 存贮多次是允许的。
例：FACTOR VARS=V1 TO V12 /SAVE REG(ALL,FACT) 表示建立FACT1, FACT2, ... 等名字存放因子得分。
10. MISSING
DEFAULT 和LISTWISE 是等价的, PAIRWISE, MEANSUB, INCLUDE 分别表示用变量对判定、用均值代替、包含有缺失值的记录。
典型程序及结果说明：

FACTOR VARIABLES= V1 TO V5.

使用因子分析唯一的变量选项VARIABLES，其余子命令内容采用默认值，即主成分方法提取因子、方差极大法旋转。

【例5.10】Linden 数据因子分析，采自Johnson, R.A.。数据是关于二次世界大战以来奥林匹克十项全能得分，共160组数据，对每项得分施以标准化，对样本相关矩阵做主成份和极大似然因子分析。十项运动为：百米(x1)、跳远(x2)、铅球(x3)、跳高(x4)、400米栏(x5)、110米栏(x6)、铁饼(x7)、撑杆跳(x8)、标枪(x9)、1500米(x10)，样本相关矩阵为：

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10
x1	1.00									
x2	0.59	1.00								
x3	0.35	0.42	1.00							
x4	0.34	0.51	0.38	1.00						
x5	0.63	0.49	0.19	0.29	1.00					
x6	0.40	0.52	0.36	0.46	0.34	1.00				
x7	0.28	0.31	0.73	0.27	0.17	0.32	1.00			
x8	0.20	0.36	0.24	0.39	0.23	0.33	0.24	1.00		
x9	0.11	0.21	0.44	0.17	0.13	0.18	0.34	0.24	1.00	
x10	-0.07	0.09	-0.08	0.18	0.39	0.00	-0.02	0.17	-0.00	1.00

相关矩阵的前四个特征值分别为3.78, 1.52, 1.11 和0.91，累积频率为73 .3%，取个3或4个主成分，程序及结果如下：

```
DATA LIST MATRIX FREE/ X1 TO X10.
N 160.
BEGIN DATA.
1.00
0.59  1.00
...
-0.07  0.09 -0.08  0.18  0.39  0.00 -0.02  0.17 -0.00 1.00
END DATA.
FACTOR READ=COR TRIANGLE /VARIABLES=X1 to X10/CRITERIA FACTORS (4)
      /ROTATION VARIMAX.
FACTOR READ=COR TRIANGLE /VARIABLES=X1 to X10/CRITERIA FACTORS (4)
      /EXTRACTION ML /ROTATION VARIMAX.
```

未旋转时的结果：

	主成分法				极大似然法			
	因子1	因子2	因子3	因子4	因子1	因子2	因子3	因子4
x1	.69052	.21701 -	.52025	.20603	-.07027	.34879	.82887	-.16853
x2	.78854	.18360 -	.19260	-.09249	.08966	.43140	.59312	.27456
x3	.70187	-.53462	.04699	.17534	-.08079	.99618	-.00394	-.00075
x4	.67366	.13401	.13875	-.39590	.17969	.39761	.33440	.44513
x5	.61965	.55112	-.08376	.41873	.38983	.22492	.67031	-.13721
x6	.68689	.04206	-.16102	-.34462	-.00028	.36337	.42341	.38776
x7	.62121	-.52112	.10946	.23437	-.02058	.73125	.02676	.01819
x8	.53848	.08698	.41090	-.43955	.16980	.25601	.22761	.39371
x9	.43405	-.43903	.37191	.23451	-.00035	.44169	-.0115	.09714
x10	.14660	.59611	.65812	.27866	.99999	.00080	-.00001	.00000

旋转后的结果：

	主成分法				极大似然法			
	因子1	因子2	因子3	因子4	因子1	因子2	因子3	因子4
X1	.88383	.13651	.15619	-.11324	.16675	.85723	.24576	-.13773
X2	.63130	.19420	.51465	-.00557	.23951	.47650	.58033	.01101
X3	.24462	.82467	.22272	-.14791	.96530	.15373	.20015	-.05852
X4	.23934	.15046	.74966	.07647	.24192	.17289	.63175	.11320
X5	.79687	.07452	.10159	.46816	.05489	.70923	.23635	.32988
X6	.40381	.15319	.63466	-.17019	.20509	.26105	.58863	-.07061
X7	.18583	.81365	.14698	-.07890	.69726	.13288	.17967	-.00937
X8	-.03626	.17578	.76179	.21688	.13709	.07797	.51264	.11624
X9	-.04775	.73493	.10988	.14135	.41667	.01854	.17521	.00213
X10	.04467	-.04090	.11167	.93353	-.05520	.05572	.11333	.99045

极大似然法，Chi-square Statistic:10.5626, D.F.:11, P=.4806

公共因子方差：

主成分法：.83702,.70115,.81140,.64776,.87005,.61828,.72438,. 65957, .57446,.88761；极大似然法：.84202,.62132,.99892,.50037,.67069, . 46166, .53655,.30115,.20477,.99999。

由此可以得到特殊因子方差估计。

使用两种方法所得结论很不相同，主成分解中，除1500长跑外，所有项目在第一因子上有较大的正载荷，这个因子可称为一般运动能力，但是其余因子不能很好解释。而在极大似然解中，1500米跑是唯一在第一因子上有较大载荷的变量，所以这个因子可称为长跑耐力因子，因子2似乎是臂力因素(铅球与铁饼有较大载荷)，因子3是短跑速度(100米和400米跑载荷较大)。

进行旋转后，容易看出，铅球、铁饼和标枪都在同一因子上有较大载荷，因子可称为爆发性臂力。跳高、100 米栏和撑杆跳在某种程度上是包括跳远都在另一因子上有较大载荷，可称为爆发性腿力，100米和400米跑在一定程度上包括跳远都在第三个因子上有较大载荷，可称之为短跑速度。最后，1500米长跑在第四因子上有较大载荷，400米跑有中等载荷，可称为长跑耐力因子。用因子分析法得到的结论与田径运动中传统分类基本一致。

§5.5.5 判别分析

命令格式：

DSCRIMINANT
~/GROUPS 指示分组变量。
~/VARIABLES 指示参加判别的分析变量。
/SELECT 对具有指定的某一变量的指定值的子样进行判别分析。
/ANALYSIS 对VARIABLES 中的不同变量进行不同的判别分析。
/METHOD 提供变量筛选方法。
/TOLERANCE 改变变量进入的容许性。
/FUNCTIONS 限制判别函数的数目。
/PRIORS 指定先验概率。
/SAVE 存贮新变量，包括分组记号，判别得分，分组概率。
/OPTIONS 选择输出和控制。
/STATISTICS 输出统计量。
规则：(1) 必选项为GROUPS 和VARIABLES，其余为可选项；(2) GROUPS 、VARIABLES 和SELECT 应该顺序放在其余子命令之前；(3) 每一个ANALYSIS 命令指定其单独分析中所使用的预测变量，且这些变量应为VARIABLES 中变量的子集；(4) 所有其他命令可以以任何顺序安排，且仅影响其紧接着的ANALYSIS；(5) OPTIONS 和STATISTICS 控制输出选择；(6) 子命令以斜杠分开。
各子命令简要介绍如下：
1. GROUPS= 分类变量名(min, max)
指定分类变量名称及其取范围，min 和max 是变量的最小值最大值。
2. VARAIABLES
指定(用以对观测个体进行分类的) 预测变量名。
例：DSCRIMINANT GROUPS=AB(1,3) /VARIABLES= V1 V2 V3.
表示分类变量名为AB，它只有三个取值1, 2, 3，故构成三类判别。参与判别的预测变量为V1、V2 和V3。
3. ANALYSIS
用法：(i) 可对VARIABLES 中的不同变量指定不同的判别分析；(ii) 在逐步判别分析中来控制变量的入选方式；(iii) 其默认值对应于将VARIABLES 中所有变量进行分析；如：
DSCRIMINANT GROUPS=G(0,1) /VARIABLES= X1 TO X9
/ANALYSIS= X6 TO X9 /ANALYSIS=ALL.
第一次分析将只使用变量X6, X7 X8, X9 进行判别分析，第二次将使用X1 到X9 的全部变量进行判别分析。
4. INCLUSION LEVELS
在判别分析中控制变量进入或删除顺序，用法：①在ANALYSIS 中各变量之后附上一个0 到99 之间的整数表示其入选水平，默认值为1。变量入选按入选水平高低而先后不同。②具有偶数水平的变量按组同时进入模型，而具有水平为1 的变量则单独进入模型；③只有水平为1 的变量才可能删除；④ 0 水平变量永远不入选，但参与入选准则计算；⑤不论水平高低，不满足TOLERANCE (容许值)，则该变量不入

选。

常用的入选方法有：①DIRECT (全部入选，即逐步判别) ANALYSIS=ALL(2)表示将全部变量同时进入方程；②STEPWISE (逐步选择法) ANALYSIS=ALL(1)表示按逐步选择法增删方程中的变量；③FORWARD (前进法) ANALYSIS=ALL(3) 表示按向前法进入变量(不做变量删除)；④BACKWARD (后退法) ANALYSIS=ALL(2) ALL(1) ⑤表示按后退法选择变量(首先将全部变量选入方程，然合将满足删除原则的变量删除)。如：

DSCRIMINANT GROUPS=G(1,2) /VARIABLES= V1 TO V3

/ANALYSIS=V1 TO V3 (2) V4 V5(1) /METHOD=WILKS.

表示当V1、V2、V3 满足容许限时，同时也进入方程V4 和V5 按照逐步选择法进入。

5. SELECT

确定训练样本，待判样本(或验证样本)。用法：(i) SELECT 变量是数值型的，不必在VARIABLES 中；(ii) 若使用OPTION 9，则只对未选择的个体判别(即对训练样本不再判别)，例：

DSC GRO=A(1,2) /VAR=V1 TO V5 /SEL =V0(1) /OPT=9.

表示带有变量V0=1 的个体构成训练样本，其每个体构成待判样本，且只对V0¡¿1 的个体进行判别。

6. METHOD

指定变量选择的准则。关键字为：

DIRECT (默认值) 对所有通过容许限的变量同时进入；WILKS 表示Wilks λ 极小者进入，MAHAL 使两组间Mahalanobis 距离最大者进入，MAXMINF 使组间最小F 比最大者进入，MINRESID 对所有组对未解释变异的和极小的变量。，RAO 使Rao 的V 统计量最大者进入。

7. FUNCTIONS

确定判别函数的个数，其默认值为给出所有线性判别函数。该命令有三个参数：nf 函数的最大个数；cp 累积待征值的百分比。sig 函数的显著水平(默认为1.0) 只需使用一个参数即可控制函数的个数，三个参数应以nf、cp、sig 的顺序出现，如：

DSC GRO=A(1,4) /VAR = 1 TO 9 /FUNCTIONS=3,100,0.9

注意此时为 4 类判判，预测变量为 9 个，nf 的默认值为 9 个，所以该例前面两项指定为默认值，第三项说明若显著怍水平大于0.9，将产生相应的先概率。

8. PRIORS

含义：为各母体指定相应的先验概率。

关键字为：EQUAL 表示各母体有相等先验概率，为PRIORS 的默认值；SIZE 表示采用样本中各母体个体比例为先验概率的估计，如：

DSC GRO=A(1,3) /VAR= V1 TO V10 /PRIORS =.1 .5 .3. 表示母体1, 2,3 分别具有先验概率0.1, 0.5 和0.3。

9. SAVE

含义：存贮每个体的判别结果信息。

关键字：CLASS 后给出一个变量名，用来存贮每个个体的分类信息；SCORES 后指定一个根名字来存贮判别得分；PROBS 后指定一个根名字存贮各个体归属母体的概率，如：

DSC GRO=A(1,2) /VAR= V1 TO V10 /SAVE CLASS=P SCORES=Q PROBS=R

对上述的两类判别，在对每个个体判别之后，将在每个个体的信息中附加如下变量。P (个体被判定所属母体)；Q(判别得分)；R1(属于母体1 的概率)；R2 (属于母体2 的概率)。

10. 判别结果输出

/OPTIONS 取值为：

6 模式矩阵(pattern matrix) 的VARIMAX 旋转。

7 结构矩阵(structure matrix) 的VARIMAX 旋转。

9 仅对/SELECT 未选的记录分类。

10 仅对分组变量范围以外的记录分类

11 使用各组的协方差阵分类而非合并组内协方差阵

1 使用用户缺损值。

8 分类过程中用均值替换缺损值。

/STATISTICS 取值为：

1 判别变量的总均值及各组均值

2 判别变量的总均值及各组均值和标准差

10 区域图，使用头两个判别函数为轴做图，组均值用星号表示。

13 分类结果表，显示正确分类的比例。

14 每个记录的分类信息。

15 所有组的散点图，组号做记号，头两个判别函数为图轴。

16 每组散点图或直方图。

常用输出有：STATISTICS 10 区域图，该图以前两个判别函数为坐标轴，将各类判别的区域显示出来；STATISTICS 11 给出非标准化判别函数；STATISTICS 13 给出分类结果表；STATISTICS 14 给出每个个体的分类信息。

非逐步判别：DSC GRO=G(1,2) /VAR=V1 TO V3 /STATISTICS=11,13,14.

说明：G(1,2) 表示分类变量为G 的两类判别，参与判别的变量为V1, V2 和V3，显示结果见(图1-图3)。

逐步判别：DSC GRO=G(1,2) /VAR=V1 TO V3 /METHOD=WILKS /STATISTICS=11 , 13,14。

说明：采用Wilks 的 λ 准则作为变量选择准则，变量选择方法的默认值表示采用逐步选择法。

结果1: 非标准化典型判别函数系数

Unstandardized canonical discriminant function coefficients

	FUNC
V1	.4253184
V2	-.1196977
V3	.5354328
(CONSTANT)	.3255612

由此，非标准化典型判别函数为：

D=0.3255612+0.4253184 V1 - 0.1196977 V2 + 0.5354528 V3

结果2、分类结果表(Classificaiton results)

实际类别	观察个体数	判别结果	
G1	40	37	3
G2	30	2	28

在由40 个体组成的G1 中，有37 个被正确判归G1，3 个误判属于G2，由30 个个体组成的G2 中，有28 个正确判归G2，2 个误判给G1。

结果3. 每个个体的详细判别信息

个体序号	缺失值	训练样本个体标识	实际类别		最可能组 P(D|G)	P(G|D)	次可能组	判别得分
1		yes	1	1	0.4328	0.9811	2 0.0189	2.3015
2		yes	1**	2	0.4882	0.9773	1 0.0227	0.0366
.		.	.	.	.	.	. .	.
.		.	.	.	.	.	. .	.
.		.	.	.	.	.	. .	.

上表给出的是每个观测个体的判别信息，判别并非直接根据判别得分｜大小或正负进行，而是按最大后验概率(highest probability) P(D|G) 对应的类别进行，如：对个体1，实际类别为1，最大后验概率P(G|D)=0.9811所对应的类别为1，故判为类别1；对个体2，实际类别为1，最大后验概率P(G| D) = 0. 9773 所对应的类别为2，故判为属于类别2 (此时为错判，以** 标记)。

§5.5.6 聚类分析

命令格式：

CLUSTER

~variable list 指示参与聚类的变量。

/PRINT 控制打印输出。

/PLOT 显示成类过程图。

/MEASURE 度量准则。

/SAVE 保存成类结果。

/METHOD 聚类方法。

/MISSING 缺失值处理方法。

/ID 指示一个聚类成员表的标志变量，默认情况下使用记录号。

说明：CLUSTER 为系统聚类法程序，当所有可选项均缺损时，观测个体之间距离采用平方欧氏距离，各类之间采用类间平均法。输出显示包括：用于分析的观察个体数量，并类过程表和纵向逐步并类图。将具有缺损观测分量的观测个体略去不参与分析。

规则：(1) 程序必选项为指定一个变量列表，其余的为可选项；(2) 变量列表必须首先在其它子命令之前给出；(3) 变量表和子命令可各被指定一次；(4) 对同一矩阵可使用多种聚类方法。

下面将各种子命令简述如下：

1. VARIABLE LIST (变量列表) 指定每观测个体所包含的变量指标。为必选项且必须在其他子命令前定义。例：CLUSTER V1 V2 V3 表示共三个变量指标V1、V2 和V3 参与分析。CLUSTER ALL 在当前文件中用户定义的所有变量均参与分析。

2. MEASURE 子命令指定观测之间个体距离的度量方式(缺损或DEFAULT 时为SEUCLID)。关键字为：SEUCLID(平方欧氏距离)、EUCLID(欧氏距离)、COSINE(变量间的夹角余弦) 、BLOCK(城区距离)、CHEBYCHEV(Chebyshev 距离)、POWER(p,r) 中的(p,r)取为，(2, 1)、(2,2)、(1,1) 表示平方欧氏距离、欧氏距离和城区距离。例：CLUSTER ALL /MEASURE=BLOCK。表示采用绝对距离作为观测个体这间距离的度量方式。

3. METHOD 子命令指定一个或多个类间的距离度量方法，只能使用一次该命令，但可指定多个方法，缺损为BAVERAGE。关键字为：

BAVERAGE 类间平均距离法，WAVERAGE(类内平均法)，SINGLE(最短距离法)，COMPLETE(最长距离法)，CENTROID(重心法)，MEDIAN(中间距离法)，WARD 法。例：CLUSTER V1 V2 V3 /METHOD=BAVERAGE CENTROID 表示对变量V1 V2 和V3，分别采用类间平均法和重心法进行聚类。

4. PRINT 子命令

控制除图形之外的其它聚类结果显示，缺损值为SCHEDULE。关键字为：

SCHEDULE 显示并类过程信息(agglomeration schedule)。CLUSTER() 类别成员表，min 和max 分别指在聚类解中最小和最大的类别数，在括号内可以打入一个数，表示聚成的类数；打入两个数则是成类数的范围。DISTANCE 距离或相似系数矩阵，其类型因度量而定。DEFAULT 同SCHEDULE。NONE 取消上述选择，当希望用SAVE 存贮时，可用该选项。

例：CLUSTER V1 V2 V3 /PRINT=CLUSTER(3,5) 将显示聚为3、4 和5 类时各观测个体类别归属。

5. PLOT 子命令控制图形输出。该命令缺损值为VICICLE。关键字为：

VICICLE(min,max,inc) 纵向逐步并类图(或称为冰棱图)。范围的指定是可选项。若指定范围时应采用整数。min 与max 是开始显示和结束显示的聚类解的个数，inc 为增量，其缺损值为1。

HICICLE(min,max,inc) 水平逐步聚类图，用法仿上。若同时指定VICICLE 和HICICLE，则最后指定者有效。

DENDROGRAM 树形图，以合并距离按比例缩放为坐标尺。

NONE 无图形输出。

例：CLUSTER V1 V2 V3 /PLOT=HICICLE(4,10,2) 表示产生4 类，6 类，8 类和10 类水平逐步聚类图。

6. SAVE 子命令

对所指定的聚类解的水平，将类别成员(即所包含的观测个体) 以新的变量存放到当前文件。该命令的唯一指定是关键字CLUSTER，并在其后按括号内的数字表示聚类解的个数，或用(min,max) 指定为聚类解的范围，与/PRINT CLUSTER() 对应。注意它们为SAVE 子命令中的必选项。

例：CLUSTER V1 V2 V3 /METHOD=BAVERAGE(CLSUMEM) /SAVE=CLUSTER(4,5) 表示首先产生两个新变量CLUSMEM4 和CLUSMEM5，分别包含了当为为4 类和5 类时各观测个体的类别。

对每个要存贮其聚类信息的聚类METHOD，应在其后指定根名(rootname)。因此，当使用SAVE 时，METHOD 成为必选项。

7. MISSING 子命令

控制对带有缺损观测分量的个体的处理，该命令的缺损值为LISTWISE。关键字为：

LISTWISE 略去所有带有缺损值分量的个体。

INCLUDE 将带有用户缺损分量的个体包括进来参加分析。

8. ID 子命令

含义：在类别成员表，逐步分类图和树形图中命名一个字符串变量作为每个个体的标识符。该命令缺损时，以个体序号为标识符。

此外，还有直接按读写数据矩阵的WRITE 和READ 子命令。

程序范例：

CLUSTER V1 V2 V3 /PLOT=DENDROGRAM, VICICLE /PRINT= CLUSTER(2, 4) , SCHEDULE.

程序说明：⑴参与分析的变量为V1、V2 和V3。⑵聚类采用平方欧氏距离和类间平均值。⑶显示树形图和纵向合并图。⑷显示在分为2 类、3 类和4 类时的类别成员。

输出结果说明：

⑴类别成员表(cluster membership of cases) 第一列(CASE) 为按序号排列的观测个体；以后各列分别为聚成不同类时，各个体的所属类别。

⑵聚类过程表(Agglomeration schedule for clustering) 第一列为聚类步(stage)，第二、三列(clusters combined)给出每步所合并的为哪两类；第四列(coefficient) 为该两类的距离；第五、六列表示该两类分别在哪一步形成(stage cluster 1st appears)；第七列(next stage) 表示现在形成的类在哪一步又被合并。

⑶纵向逐步并类图(纵向冰棱图vertical icicle plot) 横向表示各个体的序号(case number)，纵向表示分成几类(number of clusters)，各类之间用空格分开。

⑷树形图(DENDROGRAM) 横坐标表示各类之间距离按比例缩放后的尺度数，纵坐标为个体标号。

【例5.11】从21个工厂抽取同类产品，每个产品测两个指标，欲将各厂的质量情况进行分类[2]，其程序如下：title '杨维权，刘兰亭，林鸿洲:《多元统计分析》，第218页'.

```
set /more off /LENGTH 35.
data list free /x1 x2.
begin data.
 0 6  0  5  2 5  2 3  4 4 4 3 5  1 6  2  6  1  7  0
-4 3 -2  2 -3 2 -3 0 -5 2 1 1 0 -1 0 -2 -1 -1 -1 -3 -3 -5
end data.
list.
CLUSTER X1 X2 /METHOD SINGLE COMPLETE CENTROID MEDIAN WARD
              /PLOT DENDROGRAM VICICLE.
```

LIST 命令给出数据列表，CLUSTER 命令使用不同的聚类方法对样品进行聚类，输出结果包括聚类过程(Agglomeration Schedule)、树形图(Dendrogram)和垂直冰棱图(Vertical Icicle Plot)，详细输出结果从略。

§5.5.7 生存分析

SPSS 和生存分析命令有三组，即KM、Survival和COXREG。用例：

```
KM survt BY treat
   /STATUS=cens event(0)
   /TEST=logrank,breslow, tarone.
```

```
SURVIVAL /TABLES=studytim BY drug(1,3)
         /STATUS=died(0)
         /INTERVALS= THRU 30 BY 3.9
         /PLOT(logsurv).
COXREG survt WITH treat age
   /STATUS=cens event(0)
   /CATEGORICAL=treat
   /PRINT=all.
```

附：SPSS/PC+ 4.0 运行菜单

由于SPSS/PC+ 强大的菜单提示，实际上没有必要将其完整的语法都列出来，少数不明确的地方可以注明。在菜单方式下，用右光标键→进入下一级菜单，用回车键选定即可，若不选定，则用左光标键←返回上级菜单。现将SPSS/PC+ 4.0 的运行控制菜单列表如下：

★入门是SPSS/PC+ 总的说明。
★读写数据
DE 数据录入工具。
GET /FILE ' ' 读取SPSS/PC+ 系统文件。
SAVE /OUTFILE ' ' 存贮SPSS/PC+ 系统文件，默认文件名为SPSS.SYS。
TRANSLATE FROM " 转入外部文件。
TRANSLATE TO " 转贮外部文件。
对电子报表，/FIELDNAMES 和/RANGE 指示区域名。
DATA LIST [FILE "] FIXED/、TABLE/、FREE/ 读取列表数据并为系统设定活动文件。变量表包括变量名和格式(N—A)，FREE 表示自由格式读取数据，N 表示小数点位置，A 表示字符类型变量。
BEGIN DATA 列表数据开始。
标号与格式化
VARIABLE LABELS 变量标号。
VARIATE LEBELS 变量名'说明' [/...] 指示变量的标号。
VALUE LABELS 变量名值'说明' 值'说明' ... 指示变量各取值的标号。
ADD VALUE LABELS like VALUE LABELS
FORMATS 格式化变量，常见格式如：(F) (COMMA) (DOLLAR)。
MISSING VALUES 变量名(缺失值) ... 指示系统的缺失值(SYSMIS)。
DATE (Trend 选项) 产生一个规则间隔的时序资料。
IMPORT /FILE ' ' 读入SPSSX 格式文件。
EXPORT 生成SPSS/PC+ 格式文件，选项与IMPORT 类似。
MODIFY VARS 修改变量的属性。
WRITE 将活动文件写于ASCII 文件，/CASES和/VARIABLES指示记录数和变量。
★修改数据或文件
1. 修改数据的值
COMPUTE 命令生成SPSS/PC+ 新变量。

CREATE (Trend 选项) leads, lags, 差分，移动平均，及类似的运算。
IF (条件) 变量=值条件语句。
RECODE 变量表(取值列表=新值) [...]，对数据重新编码。
COUNT 新计数变量=旧变量(值)，依条件计数。
RMV (Trend 选项) 去掉时序资料中的缺失值。
AUTORECODE 将变量的值重编码为连续的整数。
RANK 产生秩次，正态性得分，Savage 分和分位点。
2. 选择或数据加权
SELECT IF 变量条件 值. 做为以后处理的永久性选择条件。
PROCESS IF 变量条件 值. 为下一个统计命令进行有条件的选择。
N 数. 选择前N 个数量的记录进行处理。
SAMPLE 样本比例 样本大小[FROM 文件大小]. 随机选择一定比例的量。
WEIGHT BY 变量. 指示分析所用的加权变量。
PREDICT (Trend 选项) 指示Trends 命令的预测范围。
USE (Trend 选项) 指示命令分析的记录。

3. 文件操作
SORT CASE [BY] 变量名[A/D] [变量名]. 对活动文件的记录进行排序。
JOIN 合并两个或多个SPSS/PC+ 文件。
AGGREGATE 把亚组合并成单一记录。
FLIP 将活动文件的行列转置。
★数据作图
1. PLOT 产生两变量的散点图，并且可以同时计算回归统计量。
2. GRAPH 计算统计量并把它们传给绘图软件包(Harvard Graphics等)。
3. CASEPLOT Trends 选项。时间序列的示意图，图中时间轴是垂直的，与TSPLOT 不同，结合其他绘图软件可以产生高分辨图形。
4. TSPLOT Trends 选项，产生一个或多个时间序列的图示，时间轴是水平的。
5. NPPLOT 是Trends 中的选项，产生一个或多个时间序列的正态概率图。
6. FASTGRAF 调Graph-in-the-Box 至内存，并且返回至REVIEW。
7. MAP 是Mapping 中的选项，计算统计量并且传递到由MapInfo 而来的SPSS/PC+ Map (或Ashton-Tate 的Map-Master) 显示地图。
★分析数据
1. 描述统计(DESCRIPTIVE STATISTICS)
FREQUENCIES 产生频数表、综合统计量、直条图和直方图。
DESCRIPTIVES 产生描述统计量。
CROSSTABS 使用交叉表形式显示两个变量的分布，可以计算列联表统计量。MEANS 显示分组的均值。
EXAMINE 进行探索性数据分析。
2. 报告和制表(REPORTS and TABLES)
LIST VARIABLES=变量表/CASE=FROM 值TO 值BY 值数据快速、简单的列表。
REPORT 产生综合统计量报告，和记录列表。
TABLES 产生高质量的表格。

PRINT TABLES 在多种打印机上进行表格打印。
3. 相关与回归(CORRELATION and REGRESSION)
CORRELATIONS pearson 相关计算。
REGRESSION 多元线性模型的估计、假设检验和残差分析。
Trends 中的选项有
CURVFIT 趋势回归模型。
AREG 出现一阶相关误差时的回归分析。
WLS 加权最小二乘回归。
2SLS 两阶段最小二乘回归。
Advanced Statistics 中的选项
NLR 非线性回归。
LOGISTIC REGRESSION LOGISTIC 回归分析
PROBIT probit 或logit 分析。
4. 均数比较(COMPARING GROUP MEANS)
T-TEST 两组均值相等的检验。
ANOVA 多因素方差协方差分析。

ONEWAY 单因素方差分析，进行两两比较。
5. 高级统计(ADVANCED STATISTICS)
MANOVA 处理包括协变和重复测量量在内的多元方差分析。
6. 分类与聚类(CLASSIFICATION and CLUSTERING)
FACTOR 因子分析。
QUICK CLUSTER 当类数已知时的高效聚类分析。
CLUSTER 一般的系统聚类分析。
Advanced Statistics 中的选项。
DSCRIMINANT 判别分析。
7. 时间序列分析(TIME SERIES)
EXSMOOTH 指数平滑模型
SEASON 季节模型
ACF 自相关函数
PACF 偏自相关函数
CCF 互相关函数
ARIMA Box-Jenkins ARIMA 模型分析
FIT 评价模型的拟合情况
SPECTRA 周期的谱分析
X11ARIMA Census Method II X-11 季节调节模型
8. 分类数据分析(CATEGORIES)
ANACOR 对应分析(correspondence analysis)
HOMALS 使用交错最小二乘(ALS) 的一致性分析(HOMogeneity analysis)。过程对名义尺度的分类数据进行分析，把观察分成相一致的子集。
PRINCALS 使用ALS 的主成分分析(PRINcipal Components analysis)。过程对一组变量进行分析，确定它们变动的维数。与普通主成分分析不同，命令不要求变量用区

间尺度测量，只假设变量间的关系是线性的。
OVERALS 使用ALS 技术进行两个或多个变量集的非线性典型相关分析。与普通典型相关不同，OVERALS 并不需要变量用区间尺度测量，也不假定变量间的关系为线性。
OTHOPLAN 为conjoint 分析准备正交的设计。设计能用较少的几种选择实现实验对象到各因子水平组合的分配，它为PLANCARDS 和CONJOINT 准备设计文件(”plan file”)。
PLANCARDS 打印CONJOINT 的设计内容或给实验对象的几种选择。
CONJOINT conjoint 分析研究的结果，它既使用ORTHOPLAN 的设计文件，又使用数据文件，数据文件包含了实验对象对几种选择所排的秩次或打分。
9. 其它(OTHER)
NPAR TESTS 非参检验
一个样本的Binomial, chi-square, Kolmogorov-Smirnov 和runs 检验。两个样本的McNemar, sign, Wilcoxon 检验。
K 个相关样本的Cochran, Friedman, Kendall 检验。
两个独立样本的Man-Whitney, Kolmogorov-Smirnov, Wald- Wolfowitz 和Moses 检验。
k 个相关样本的Kruskal-Wallis 和median 检验。
RELIABILITY 在可加性的尺度上进行item analysis，计算一系列常用的可靠性指标，如Cronbach’s alpha。RELIABILITY 并不是把这些标度直接施于分析数据，若分析的结果很好，使用COMPUTE 命令产生包括这个标度的新的变量值。
HILOGLINEAR n-维交叉表层次log-linear 分析，检验模型中所有效应的显著性，估计饱和模型的参数，进行模型的选择。
LOGLINEAR 进行log-linear 和logit 分析，采用Newton-Raphson 算法估计饱和及非饱和模型，检验模型中所指定的效应，使用极大似然方法估计参数。
SURVIVAL 利用寿命表，图示及有关统计量，考察两个事件时间的长度，记录可以分组分析和比较。时间间隔可以使用SPSS/PC+ 的日期转换函数YRMODA。
★运行控制及信息
1. Set 命令改变由show 命令所报告的所有设置。

⑴菜单控制
/AUTOMENU ON/OFF 控制菜单的自动出现。
/HELPWINDOWS ON/OFF 控制菜单旁边的帮助窗口。
/MENUS STANDARD/EXTENDED 设置窗口为标准/扩展。
⑵输出控制
控制屏幕、打印机和文件的输出。
/SCREEN ON/OFF 开启/关闭屏幕输出。
/PRINTER OFF/ON 关闭/开启打印机，开启时运行速度减慢。
/LENGTH 页长，默认为屏幕24 行和打印机59 行。
/WIDTH 页宽，默认为79 字符。
/EJECT 打印机或输出文件中的回车控制，当/SCREEN ON 时关闭，关闭时以折线分页。
/INCLUDE ON/OFF 显示INCLUDE命令文件中的命令。

/ECHO ON/OFF 将命令复制到结果文件。
/LISTING 改变默认的输出文件SPSS.LIS。
/LOG 拷贝执行的命令到磁盘文件，默认文件是SPSS.LOG。
/RESULTS ” 矩阵(CORRELATIONS, FACTOR, MANOVA, 等产生) 和WRITE 输出的文件名，默认为SPSS.PRC。
⑶操作控制
/RUNVIEW ON/(OFF 或MANUAL), 返回REVIEW，可以在SPSSPROF.INI 中改动。
/PROMPT ” 行命令方式的提示，默认为SPSS/PC:
/CPROMPT ” 行命令方式下的续行提示。
/MORE ON/OFF 输出时的暂停。
/BEEP ON/OFF 系统振铃控制。
/COLOR ON/OFF 开启/关闭颜色。
/RCOLOR() REVIEW 的颜色，用括号中的三个整数表示。
/VIEWLENGTH 屏幕显示行数，默认为25。
/ERRORBREAK ON/OFF 终止一组命令的执行。
⑷工作文件控制
/COMPRESS ON/OFF 指示文件是否被压缩。
⑸其它
/SEED 随机数的种子，默认用系统时钟。
/BLANK 默认是数值变量中的空格设为系统缺失值，可设如/BLANK -99999。
⑹categories plots
调整ANACOR, HOMALS, PRINCALS, OVERALS 几个命令中图轴的刻度。
/CPI 横轴的每英寸字符数，默认为10。
/LPI 纵轴的每英寸字符数，默认为6。
2. SHOW 报告当前的设置情况。
3. DISPLAY 显示活动文件的变量名和标号
4. SYSFILE INFOR 用于检查非活动系统文件的内容。
5. SPSS MANAGER 内容包括STATUS, INSTALL, REMOVE，如：

SPSS MANAGER INSTALL REGRESSION /FROM ’d:\SPSSBACK’.
6. TITLE 和SUBTITLE 指示输出标题，COMMENT 或* 指示程序与注释。
7. TIME SERIES UTILITIES 包括：
TSETS, 设置DEFAULT, /PRINT, /NEWVAR, /GRAPHICS, /GOUT, /GINVOKE, /MX-AUTO, /MXCROSS, /MXNEWVARS, /MXPREDICT, /MISSING。
TSHOW 显示当前的设置。
MODEL NAME, TDISPLAY, SAVE MODEL, and READ MODEL
save model / OUTFILE=” /KEEP /DROP /TYPE
read model /FILE ” /KEEP /DROP /TYPE
VERIFY 检查日期与记录中的一致性，变量由/VARIABLES 指定。
8. 图形设置(graphics setup)
设置与SPSS/PC+ 相连的绘图软件，包括：HARVARD, CHART, CMASTER, 3GTALK, 4GTALK, DA，如：

GSET PACKAGE HARVARD
GSHOW 显示GSET 参数的当前值。
★运行DOS 或其它程序
使用DOS 或EXECUTE 命令运行DOS 或其它命令，如：
DOS DIR.
EXECUTE '\FORMAT.COM ' 'A:'.
★扩展菜单(extended menus)
在菜单与帮助系统中的某些菜单包括了一些看不到的内容，需要借助于扩展菜单，它们是一些较高级的或不太常用的特色，使用Alt-X 或者SET 命令进行标准菜单与扩展菜单的切换，在REVIEW 屏幕的右下角状态行上显示当前值。
★SPSS/PC+ 选项(SPSS/PC+ options)
SPSS/PC+ 软件是一系列功能的组合，通过安装特定的程序实现这些功能。
★退出
用FINISH 退出SPSS/PC+ 至DOS 系统，注意数据、程序和结果的保存。
系统菜单用例，现欲使用数据录入工具DE建立系统文件，进入SPSS/PC+ 后，选择read or write data，打右箭头出现DE，因光标恰好在DE位置，打¡Enter¿，则编辑窗内出现关键字DE，用Alt-C(或F10, 择run from cursor) 即进入DE。一般的命令可先使用Alt-E，然后打入关键字，打ESC，则提示窗口出现有关的子命令，然后用左右箭头进行选定。