第七章　SYSTAT

§7.1　SYSTAT 应用概要

1983年Leland Wilkinson 就已经拥有微机上SYSTAT版本。历经CP/M、MS-DOS、VAX/VMS, UNIX, DATA General, NCR Tower, IBM PC兼容机及Apple Macintosh 系统。SYSTAT 3.0 和4.1 分别是SYSTAT 公司1986 年和1989 年推出的产品。它们的特征完全相仿。SYSTAT 最显著的特点是模块化功能，目前SYSTAT 也有Windows下的产品，如SYSTAT for Windows 5.0基本上保持原有的模块化特征。

SYSTAT 3.0 系统由12 个相对独立的功能模块组成，这些模块可分成数据处理模块和统计模块。SYSTAT 的数据管理模块是DATA，它用于SYSTAT 的数据预处理并把外部数据文件如ASCII、dBASE、Lotus 格式，用于其系统文件。DATA 模块以外的模块是用做统计分析的。在统计处理过程中，有一些统计模块也能产生存储计算结果的SYSTAT 数据文件，对同一个SYSTAT 文件的数据，可以在多个统计模块中使用，进行不同的统计处理。

SYSTAT 其它的产品有Probit、Logit和Score等。PROBIT 使用累积正态分布估计二分类变量的反应函数，它是一个极大似然程序，可以自动产生哑变量和MGLH 中其它的特征。LOGIT 对二分类数据或多分类数据进行多项logit 模型分析，可以处理更大的模型和数据。SCORE 提供一些检验综合统计量，可靠性系数以及item analysis、Rasch 模型、多项选择或两极尺度(bipolar scales) 问题。REPORT WRITER 是为科学和商务报告而设计的，有格式输出、标题居中、综合统计量以及打印机字型控制。另外，有每页的边界、页长、行宽控制。数字和字符串可以分别定格式而打印。STAT/TRANSFER 模块提供了一种方便的方法，能够在SYSTAT, LOTUS, SPSS/PC，和STATA 之间转换数据，具体操作不过是简单的菜单选择。LAZERTE EDITOR 是一个高速数据编辑器，需要8087 或80287 数学协处理器。Macintosh 版包括下拉式菜单、窗口、以及剪贴板等与其它Macintosh 软件的接口。Mainframe 版: DEC VAX 11/780、MicroVAX 和Hewlett-Packard 9000 大型机上有相应的产品，IBM 大型机版本于1986 年秋问世。

§7.1.1　运行

应注意检查CONFIG.SYS 文件和一个引导文件(名为SYSTAT)，另外还在AUTOEXEC.BAT 文件中设SYSTAT 程序的路径说明，如：PATH C:\DOS;C:\SYSTAT;D:\wp51 在硬盘上运行SYSTAT 很简单，在DOS 引导之后，于根目录下键入：C:\ >SYSTAT 这时可看到由“#”字符组成的SYSTAT 字样出现，按回车后即进入SYSTAT 菜单。只要键入该模块编号或名称后按回车即可。例如要调用STATS模块则键入：

> 3 或

> STATS

这时屏幕清屏，于屏幕底部的左则再次出现SYSTAT 提示符“>”，即表明已进入程序模块。在菜单系统中可以使用HELP 和SUBMIT 命令寻求帮助和运行命令文件。从菜单或程序模块返回DOS 系统，用QUIT 命令。

SYSTAT的各个模块可以分别运行，如：

A>DATA

进入数据管理模块。注意单个模块的运行应有DATA.DEF 文件存在。

程序调入内存后，屏幕上出现由“#”字符组成的SYSTAT 字样，随后下方出现箭头“>”提示符，此时说明已进入SYSTAT 系统，即可开始工作。从一个程序模块转入另一个程序模块时，一般要先退出当前模块(用QUIT 命令)，然后再调入另一模块。如果所使用的程序不在同一张软盘上，则必须更换B 驱动器上的软盘。

在SYSTAT 的执行菜单上追加其它程序也是可能的，这时应编辑文件SYSTAT. DEF. 指定的内容包括命令名、有效文件名、一行或多行的帮助信息。命令名前导以@，必须用大写，文件名应从第十列开始，下面是一个例子。

@WORDPERF \WP\WP.EXE

命令WORDPERF 启用WordPerfect 进行文字处理，使用它来生成SYSTAT 的命令文件或其它文本。文件名应是DOS 的可执行程序。

SYSTAT 作图使用的是IBM 扩展图形字符集。一般的打印机不能打印这个图形集。因此，要从打印机上输出图形，必须改变隐含图形设置。即恢复使用标准ASCII 字符集。则需将DATA 盘中的名为DATA.DEF 文件与名为GENERIC.DBF 文件名互换。

SYSCROLL 可以浏览系统进行过的操作。程序允许在内存中保存最多九屏的输出，由于未采用直接视频显示，所以运行速度较慢，也不与DOS 的其它命令冲突。使用DOS 通常的办法运行SYSCROLL. EXE ，默认保存四屏内容，可以使用SYSCROLL 2 或SYSCROLL 9 等等来调节屏数。

程序驻留后，使用Ctrl- 健来激活SYSCROLL，其时功能键有：

PgUp - 上滚一屏　Up arrow - 上卷一行　Home - 窗口顶部

PgDn - 下翻一屏　Dn arrow - 下卷一行　End - 窗口的底

再次使用Ctrl- 则返回运行的程序。

§7.2 SYSTAT 命令和模块

§7.2.1 SYSTAT 命令

进入SYSTAT 模块后，显示器屏幕上会出现SYSTAT 字样，并显示'>'。这时，就可以打入该模块的命令，命令计算机进行相应的操作。若命令较长，一行打不下，可在这行结束处打一个逗号，表明命令未输入完，然后转下一行继续输入。

㈠. 命令特点

SYSTAT 的命令有冷热之分。打入热命令，计算机立即执行，并给出执行结果。而打入冷命令时，计算机并不立即执行。实际上打入冷命令，只是作了某种选择，或指定了某个条件。一般地说，冷命令的次序可以任意，但必须在打入一个热命令之前，把所需要的冷命令全部打完。对于SYSTAT 的命令，计算机只辨认其前两个字母。因此，输入命令时，可以只打入命令的前两个字母。

在SYSTAT 系统中，所有模块都能用USE 命令读取SYSTAT 文件中数据，而用SAVE 命令来存写数据，但只有数据模块能够读取来自其他途径的数据。

㈡. 公用命令

在SYSTAT 中，有些命令可以在所有的模块中使用，这些命令称为公用命令。常用的公用命令有：

1. BY：指定一个或多个分组变量，命令格式为：

   BY <变量> [,<变量>,<...>] . 如：

BY AGE (指定数值变量AGE 为分组变量)

BY SEX,NAME$ (指定变量SEX 和变量NAME$ 为分组变量)

BY 命令后面的变量必须是经过排序的。重新用USE 命令打开文件或打入一个后面无变量的BY 命令都可取消以前指定的分组变量。

2. FORMAT：规定输出数据小数点后的位数，命令格式为：

FORMAT=5 (规定保留5位小数)

FORMAT=3 (规定保留3位小数，即默认值)

FORMAT 命令后面的数字不得小于0，不得大于9。

3. HELP：输出帮助信息，命令格式为：

HELP [<命令>]，如：

HELP (显示有关模块的帮助信息)

HELP BY (显示有关BY命令的帮助信息)

4. OUTPUT：指定输出装置，命令格式为：

OUTPUT * (指定显式器为输出装置)

OUTPUT @ (指定打印机为输出装置)

OUTPUT <文件>，如：

OUTPUT RESULT (建立ASCII 码的磁盘文件RESULT.DAT，存储输出结果)

一旦打入OUTPUT 命令，此后的输出结果将从指定装置输出。

5. QUIT：终止SYSTAT 的运行。命令格式为：

QUIT

6. SAVE：建立一个新的SYSTAT 数据文件，格式为：

SAVE <文件>，如：

SAVE MANOVA (建立名为MANOVA.SYS 的数据文件)

不需输入文件的扩展名'.SYS'，SAVE 命令会自动地加上。

7. SELECT：规定一项或多项选择数据的标准，命令格式为：

SELECT <变量>=<数值><字符变量>=<字符串>< ... >

SELECT STATE$='NY'

SELECT REGION=4 STATUS=2

打入后面无选择标准的SELECT 命令可取消以前所作的规定。

8. SUBMIT：从扩展名为'.CMD'的命令文件中取出命令并执行。命令格式

SUBMIT <文件>，如：

SUBMIT MYFILE

打入这个命令后，计算机就寻找文件MYFILE.CMD，并依次执行文件中的所有命令。

9. USE：打开一个已存在的SYSTAT数据文件，命令格式为：

   USE <文件>，如：

   USE OLDFILE (打开文件OLDFILE.SYS，准备读取数据)

   执行USE命令将显示被打开的数据文件的所有变量名称。

10. WEIGHT：指定一个权数变量，命令格式为：

    WEIGHT=<变量>，如：

    WEIGHT=NUMBER (指定变量NUMBER 为权数变量)：

只有QUIT 命令是热命令，其余均为冷命令。

㈡. SYSTAT 文件和变量

SYSTAT 系统有四种文件：SYSTAT 数据文件，ASCII 码的字符文件，SYSTAT 命令文件和SYSTAT 的临时文件。四种文件的扩展名分别为.SYS .DAT .CMD 和.TMP。SYSTAT 中的文件名由1–8个字母或数字构成，但必须是字母打头。为了说明文件盘所在的驱动器，可以在字母和数字组成的名字前面加上驱动器符。在SYSTAT 中不需输入文件的扩展名，计算机会自动加上。

SYSTAT 中的变量有两类：数值型变量和字符型变量。数值变量其取值均为数值，而字符变量的取值均为字符。数值变量名的构成与文件名相同，由1–8个字母或数字组成，不同之处是在变量名中可以使用字符'_'。字符变量名是在数值变量名之后再加上一个字符'$'。

在SYSTAT 中，无论数值变量或是字符变量都可以加上最多两位数的下标，并且在一些命令中还可以指定下标的范围。

§7.2.2 数据和统计模块

㈠DATA 模块

DATA 模块是SYSTAT 软件包中唯一的数据模块。它能够接受来自键盘，ASCII 码文件和已存在的SYSTAT 文件的数据，经过整理加工，生成新的SYSTAT 数据文件。

1. DATA 模块的命令

   (a) APPEND：将两个具有相同变量的SYSTAT 文件串连，串连生成数据文件例数为两文件数据例数之和，命令格式：
   APPEND <文件> <文件>，如：
   APPEND FILE1 FILE2 把文件FILE2.SYS 追加到文件FILE1.SYS。

   (b) DELETE：删除当前例数据。

   (c) DROP：删除指定的变量，命令格式：
   DROP <变量> [,<变量>,< ... >]，如：
   DROP SEX,NAME$ (删除数值变量SEX 和字符变量NAME$)

   (d) EDIT：进入全屏幕编辑器，命令格式：
   EDIT [<文件>]，如：
   EDIT (进入编辑器，编辑一组新数据)
   EDIT AFILE (进入编辑器，编辑文件AFILE.SYS 的数据)

(e) GET：打开一个扩展名为.DAT 的ASCII 码文件，命令格式：

GET <文件>，如：

GET ASCFILE (打开ASCII 码文件ASCFILE.DAT)

(f) IF：规定一个比较条件，并判断其是否满足，命令格式：

IF <条件> THEN <命令>，如：

IF CASE=4 THEN LET AGE=39

(g) INPUT：规定输入数据的变量个数，以及各变量的名称和性质，命令格式：

INPUT <变量>[,<变量>, < ... >]，如：

INPUT AGE NAME$

(h) LET:计算表达式值，赋给变量，命令格式：

LET <变量>=<表达式>，如：

LET LAGE=LOG(AGE) (计算变量AGE 的对数赋给变量LAGE)

(i) LIST:输出文件中所有或部分变量的数据，命令格式：

LIST [<变量>, < ... >]，如：

LIST (输出文件中所有变量的数据)

LIST AGE,NAME$ (输出文件中变量AGE 和NAME$ 的内容)

(j) LRECL:规定各类输入数据的读取长度，命令格式：

LRECL=<数值>，如：

LRECL=256 (对每例数据，读前256 个子符，后面的不读)

此命令的缺省值为80，即每例数据只读前80 个子符。

(k) PUT:建立一个扩展名为.DAT 的ASCII 码文件，命令格式：

PUT <文件>，如：

PUT ASCFIL (在磁盘上建立文件ASCFIL.DAT)

(l) RUN: 执行此命令之前的冷命令所规定的任务，命令格式：

RUN

(m) SORT:将文件中所有或部分变量的数据，按其值的大小，从小到大排序。命令格式：

SORT[<变量>, < ... >]，如：

SORT (将文件中所有变量的数据排序)

SORT AGE (将文件中变量AGE的数据排序)

(n) USE: 打开SYSTAT 文件，指定可读取数据的变量；并连两个SYSTAT 文件中的所有或部分变量，命令格式：

USE <文件>[<变量>, < ... >][<文件>[<变量>, < ... >]]，如：

USE DATAFILE (打开文件DATAFILE.SYS)

USE DATAFILE(AGE,NAME$) (指定文件DATAFILE.SYS 中AGE，NAME$ 为可读取数据的变量)

USE FILE1,FILE2 (并连文件FILE1.SYS和文件FILE2.SYS的所有变量)

USE FILE1(AGE) FILE(NAME) (将文件FILE1 中的变量AGE 和文件FILE2 .SYS 中的变量NAME 并连起来)

(o) NEW: 取消在此之前的所有命令，并清除数据空间，命令格式：

NEW

以上15 条命令中，APPEND,RUN 和NEW 命令是热命令，其余均为冷命令。

2. 从键盘输入数据

从键盘输入的数据，有两种方法：

(a) 在DATA 模块中直接输入。

在操作系统状态下，进入DATA 模块。然后打入：

SAVE MYFILE

INPUT AGE NAME$

RUN

这时屏幕显示：

INPUT DATA ONE CASE AT A TIME AFTER PROMPT ARROW

现在可以输入数据。注意按例输入，一行输入一例，每行末尾应按一下回车键。不要打'>'。下面是屏幕上看到的输入内容：

```
>33 YANGHONG
>22 WANGWEI
>24 LIMING
>36 ZHANGJIE
>26 YUANPING
>37 LIZHIQIANG
>42 WANGHONG
```

数据输入完时，应在下一个'>'出现之后，打入'~'。这时应看到：

```
     7 CASES AND 2 VARIABLES PROCESSED
SYSTAT FILE CREATED.
WORKSPACE CLEAR FOR CREATING NEW DATASET
```

至此，一个内含2 个变量7 例数据的SYSTAT文件就被建立在磁盘上，其名字为MYFILE.SYS。如果想看一下输入的数据，输入命令：

USE MYFILE

屏幕上显示：

SYSTAT FILE VARIABLES AVAILABLE TO YOU ARE:

AGE NAME$

因为只是简单地看一下数据，所以只需再打入如下命令：

LIST

RUN

这时屏幕上出现：

```
                    AGE          NAME$
CASE   1         33.000         YANGHONG
CASE   2         22.000          WANGWEI
CASE   3         24.000           LIMING
CASE   4         36.000         ZHANGJIE
CASE   5         26.000         YUANPING
CASE   6         37.000       LIZHIQIANG
CASE   7         42.000         WANGHONG
```

前面打入的命令，RUN 是热命令，其余均为冷命令。

(b) 利用全屏幕编辑器输入。

在DATA 模块中有一个全屏幕编辑器，利用它可以很方便地建立SYSTAT 文件。

在DATA 模块中，打入EDIT 命令即可进入全屏幕编辑器。进入编辑器，屏幕上会出现一个数据表格，光标在变量名行上。首先按下面顺序输入所有变量名：

'AGE <Enter> 'NAME$ <Enter>

然后按<Home>键使光标转到数据区左上角。再开始输入数据，按例输入，每输入一个数据就应按一下回车键。实际输入顺序为：

```
33    'YANGHONG   <Enter>
22    'WANGWEI    <Enter>
24    'LIMING     <Enter>
36    'ZHANGJIE   <Enter>
26    'YUANPING   <Enter>
37    'LIZHIQIANG <Enter>
42    'WANGHONG   <Enter>
```

数据输入完了应按<Esc>键，使光标转到命令行，然后打入SAVE MYFILE 将输入的数据存入文件MYFILE.SYS。最后打入QUIT 命令退出全屏幕编辑器。

3. ASCII 码文件数据的转换

为了实现与其他软件的数据交换，达到数据共享的目的，DATA 模块提供了ASCII 码文件与SYSTAT 文件相互转换的功能。

(a) ASCII 码文件转换成SYSTAT 文件。首先，检查待转换的磁盘文件是否为ASCII 码文件。此外，还应保证文件扩展名为.DAT 。然后，在操作系统状态下打数据模块的名字，进入DATA 模块，打入如下命令：

```
GET ASCFILE
INPUT AGE NAME$
SAVE MYFILE
RUN
```

这样，一个名为ASCFILE.DAT 的ASCII 码文件就被转换成SYSTAT 文件MYFILE.SYS。若ASCII 码文件有些数据的列数超过80 列(比如最多是236 列)，则应在热命令RUN 之前输入：LRECL=236

(b) SYSTAT 文件转换成ASCII 码文件。

这一转换的方法很简单，只需在DATA 模块中打入如下命令：

```
USE MYFILE
PUT ASCFILE
RUN
```

就可将SYSTAT 文件MYFILE.SYS 转换成ASCII 码文件ASCFILE.DAT。

4. SYSTAT 文件的再加工

在实际统计分析过程中，常常需要对已存在的SYSTAT 文件的数据重新整理，经过取舍组合，加工成新的SYSTAT文件。

(a) 对一个SYSTAT 文件的再加工。

在DATA 模块中可对一个SYSTAT 文件中的数据，实施排序，转换，删除变量或数据等操作。

①排序：在DATA 模块中打入如下命令：

```
USE MYFILE
SORT AGE
SAVE AGESORT
RUN
```

就可将文件MYFILE.SYS 的数据，按变量AGE 数值大小，从小到大排序并存入新的SYSTAT 文件AGESORT.SYS。

②转换：利用转换命令LET 可将文件中的数值变量X 转换成它的某种函数f(X)，函数f(X) 指用运算符将变量和标准函数连接起来的表达式。下面的命令：

```
USE DATAFILE
LET SAGE=SQR(AGE)+.5
SAVE AGESQR
RUN
```

将文件DATAFILE.SYS 包含的变量AGE，求其平方根加上0.5 作为新的变量SAGE，并存入新的SYSTAT文件AGESQR.SYS。

SYSTAT 中的标准函数及运算符有：

	标准函数		运算符
SQR(X)	平方根函数	+	加号
LOG(X)	自然对数函数	-	减号
EXP(X)	指数函数	*	乘号
ABS(X)	绝对值函数	/	除号
SIN(X)	正弦函数	^	乘方号
COS(X)	余弦函数	<	小于号
TAN(X)	正切函数	=	等于号
ASN(X)	反正弦函数	>	大于号
ACS(X)	反余弦函数	<>	不等于号
ATN(X)	反正切函数	<=	小于或等于号
INT(X)	取整函数	=>	等于或大于号

③删除变量：如：

```
USE DATAFILE
SAVE AGEFILE
DROP NAME$
RUN
```

即可删除文件DATAFILE. SYS 中的字符变量NAME$，并将结果存入文件AGEFILE.SYS 中。

④删除部分数据：利用条件命令IF 规定删除数据的条件，然后用删除命令DELETE 把符合条件的数据删除掉。输入命令：

```
USE DATAFILE
SAVE AFILE
IF CASE>5 THEN DELETE
RUN
```

其执行结果：删掉了文件DATAFILE.SYS 中的第6，7 例数据，并把未被删除的数据存入文件AFILE.SYS 中。

(b) 对两个SYSTAT 文件的再加工。

在DATA 模块中，可以将两个SYSTAT 文件的数据并连或串连。

①合并文件：将两个文件的变量并列，合成一个包括所有变量的新文件。若两文件的变量不相同，则新文件的变量个数是两个文件变量个数之和。并连文件的命令如下：

```
USE FILE1 FILE2
SAVE ALLFILE
RUN
```

此命令序列可将文件FILE1.SYS 和FILE2.SYS 的数据并连起来，存入新文件ALLFILE.SYS 中。

②追加文件：将两个具有相同变量的文件顺序衔接，形成一个新文件。新文件的数据例数是两个文件数据例数之和。顺接文件的命令如下：

```
SAVE ALLFILE
APPEND MANAGE1 MANAGE2
RUN
```

此命令序列可将文件MANAGE1.SYS 和MANAGE2.SYS 的数据顺接起来，存入文件ALLFILE0.SYS中。

5. 数据的修改

对于已建立的SYSTAT 文件中的错误数据，可以用下面两种方法修改：

(a) 利用全屏幕编辑器修改

在DATA 模块，打入跟有文件名的EDIT 命令，计算机把文件中的数据读入全屏幕编辑器。这时就可以使用光标移动键，把光标移到一个待修改数据的位置上，输入正确的数据。这样一个错误的数据就修改完了。依次重复上述步骤，直到文件中所有错误数据都修改完毕。打入SAVE 命令，把修改后的数据存入一个新文件。最后用QUIT 命令退出全屏幕编辑器。

(b) 利用条件和转换命令修改

在DATA 模块中，打开待修改的文件，用条件命令IF 和转换命令LET 组合使用，修改错误数据。比如，文件DATAFILE.SYS 中第4 例变量AGE 的值应是39，变量NAME$ 的'ZHANGJIE' 应为'GAODA'。欲以改正，打入如下命令

```
USE DATAFILE
IF CASE=4 THEN LET AGE=39
IF NAME$='ZHANGJIE' THEN LET NAME$='GAODA'
RUN
```

㈡GRAPH 模块

GRAPH 模块是一个统计模块，它主要功能是根据SYSTAT 文件中的数据，按照使用者的绘图命令绘制各种统计图，并通过指定输出装置输出。GRAPH 模块既可以绘制常用的统计图如：直方图，条图，散点图和概率图；也可以绘制一些较少使勇的新型统计图，如：茎叶图和盒式图。此外，若数据是多组的，还可使用BY命令指定分组变量，绘制各组的统计图。

BAR：绘制所有或部分指定的数值变量，字符变量的条图。命令格式：

BAR [<变量><变量>< ... >][/CUM LOW=¡数值¿ WIDTH=¡数值¿] . 如

BAR (对文件中所有变量各绘制一个条图)

BAR TYPE,MONTH$ (绘制指定变量TYPE, MONTH$的条图)

BAR TYPE/CUM (绘制数值变量TYPE的累计条图)

我们用一组关于医院工作质量的数据来说明绘制统计图的具体步骤。数据已经存入名为HOSPITAL.SYS 的数据文件中，它含六个变量：有效率(X1)，病死率(X2) ，平均住院日(X3)，病床周转率(X4)，病床使用率(X5)和分组变量(GROUP)。下面就是在DATA 模块中列出的这组数据，一共是12 例。

```
                X1       X2       X3       X4       X5     GROUP
CASE    1     94.270    2.020   15.990   17.750   84.190       1
```

CASE	2	94.060	2.080	14.230	16.480	82.680	1
CASE	3	95.080	1.570	13.240	20.090	81.680	2
CASE	4	94.480	2.010	15.360	16.390	80.160	2
CASE	5	94.740	1.850	15.810	17.770	83.510	2
CASE	6	94.940	1.810	16.580	16.960	83.570	2
CASE	7	95.250	1.820	16.800	16.630	83.900	2
CASE	8	93.430	2.410	15.860	17.240	81.570	1
CASE	9	94.120	2.000	16.000	16.120	82.240	1
CASE	10	93.360	2.080	16.120	17.150	83.360	1
CASE	11	94.090	2.120	16.240	16.030	83.340	1
CASE	12	96.000	1.800	16.120	16.220	82.420	2

实际上，在GRAPH 模块中，绘制统计图的方法非常简单。我们只要用USE 命令打开包含待绘图变量的文件，就可以根据统计分析的需要，打入不同的绘图命令，计算机将立即执行打入的每条绘图命令，对指定的变量绘制相应的统计图，并把结果从指定的输出装置上输出出来。

下面用HOSPITAL.SYS 文件中的数据绘制统计图：

进入GRAPH 模块，打入命令USE HOSPITAL，打开待处理的文件，输入绘图命令绘制相应的统计图。如要绘制变量X1的直方图，只需打入命令HISTOGRAM X1。

如果我们要进一步考察变量X1的分布是否为正态分布，可以绘制这个变量的正态概率图。即打入命令PPLOT X1，执行结果将出现在显示器上。

利用BAR 命令绘制条图。如绘制X1 的条图，命令为：

BAR X1/LOW=93,WIDTH=1 (X1的最小值取93, 组距取1)

其它绘图命令的使用方法与之类似。

㈢STATS 模块

STATS 模块是一个基本统计模块。它的主要功能是计算各种统计量，如：均值(MEAN)，标准差(SD)，偏度系数(SKEWNESS)，峰度系数(KURTOSIS)，极大值(MAX)，极小值(MIN)，全距(RANGE)，方差(VARIANCE)，标准误(SEM) 和数据和(SUM)。对于分组数据，STATS 模块不仅可以计算各组的统计量，还能对各组的均值作t-检验或方差分析。

1. STATISTICS：计算所有或部分指定变量的统计量。可输出的统计量有：均值，标准差，标准误，偏度系数，峰度系数，最大值，最小值，全矩，方差和总和。命令格式为：

   STATISTICS [<变量>< ... >] [/MEAN SD SEM SKEWNESS KURTOSIS MAX MIN RANGE VARIANCE SUM]，如：

   STATISTICS (计算文件中所有数值变量的均值，标准差，极大值和极小值)

   STATISTICS TREAT/MEAN SEM VARIANCE (计算变量TREAT 的均值，标准误，方差)

2. TTEST：对指定变量作配对t检验或分组t检验。格式如下

   TTEST <变量> [<变量>< ... >][* <变量>]，如：

   TTEST X1 X2 (将变量X1和X2配成对，作配对t检验)

   TTEST X*SEX (根据变量SEX分组，对变量X 作分组t检验)

TTEST X1 X2 X3 (将三个变量X1，X2，X3 两两配对，分别作配对t检验)

TTEST X Y*SEX (根据变量SEX 分组，分别对变量X，Y作分组t检验)

上面两条统计命令均为热命令。

3. PRINT：规定结果输出的等级。其命令格式如下：

PRINT=SHORT (规定仅仅输出基本的计算结果)

PRINT=LONG (规定除基本结果之外，还输出更详细的信息)

这是一条冷命令。除在STATS模块中可使用外，它还可以在后面介绍的几个模块中使用。

以下用文件HOSPITAL.SYS说明使用STATS 模块的统计命令。进入STATS 模块，打入命令FORMAT=5,规定输出结果保留五位小数；尔后就可以作如下的统计处理：

假若我们要计算文件HOSPITAL.SYS 中的三个变量X1，X2，X3 的均值，标准差，偏度和峰度系数，打入命令：

USE HOSPITAL

STATISTICS X1 X2 X3/MEAN SD SKEWNESS KURTOSIS

计算结果:

```
TOTAL OBSERVATIONS:   12

                          X1          X2          X3
 N OF CASES               12          12          12
 MEAN                 94.485       1.964      15.696
 STANDARD DEV          0.763       0.212       1.008
 SKEWNESS              0.310       0.218      -1.447
 KURTOSIS             -0.452       0.240       1.163
```

STATS模块可以作分组t-检验和配对t-检验。下面以医院工作质量为例，进行分组t-检验：

USE HOSPITAL

TTEST X1*GROUP

在TTEST 命令中分组的变量必须在星号后面。分组t-检验的结果如下：

```
INDEPENDENT SAMPLES T-TEST ON       X1     GROUPED BY    GROUP

     GROUP         N     MEAN           SD
       1.000       8       94.069        0.474
       2.000       4       95.318        0.472

  SEPARATE VARIANCES T =       4.312 DF =  10.0 PROB = .002
   POOLED VARIANCES T =       4.306 DF =    10 PROB = .002
```

现在，我们看一下两组医院工作质量X1,X2 X3的均值，标准差和它们的极值，可打入下面命令：

```
USE HOSPITAL
BY GROUP
STATISTICS X1 X2 X3
```

BY 命令指定以变量GROUP的值分组，前提条件是GROUP 已排序。

如果在命令STATISTICS 之前，还打入命令PRINT=LONG，除了得到所计算的统计量外，得到均数差别的显著性检验。

其中BARTLETT TEST 是两组间方差齐性检验，大样本时计算的是Bartlett 卡方值；若例数小于10 时，则改用计算小样本的近似F 值(APPROXIMATE F) 。OVERALL MEAN 是两组合并的均数。最后一行的T STATISTICS 是按POOLED T检验计算的统计量。

如果有一个文件是按一个或多个分组变量分组的，而且我们想再建立一个文件把各组的统计量存起来，那么只要将命令SAVE、BY和STATISTICS配合起来使用，就能够完成这个任务。下面的命令序列

```
USE HOSPITAL
BY GROUP
SAVE MEANHOS
STATISTICS X1, X2, X3/MEAN
SAVE SDHOS
STATISTICS X1, X2, X3/SD
```

就可将所计算的三个变量的各组均值存入文件MEANHOS.SYS，而把计算的标准差存入文件SDHOS.SYS。注意：由BY命令所规定的每一组在新文件中只是一例数据。

㈡TABLES 模块

TABLES 模块也是一个基本统计模块，它可以产生各种维数的表，并能用对数线性模型加以拟合，还可以对其拟合结果作卡方拟合检验。TABULATE: 产生一个一维的或多维的表。其格式为

TABULATE <变量> [* <变量> * < ... >] [/FREQUENCY PERCENT ROWPCT COLPCT LIST]，如：

TABULATE AGE (产生一个按变量AGE 分组的一维频数表)

TABULATE AGE*SEX (产生一个按变量AGE，SEX 交叉分组的二维频数表)

TABULATE AGE*SEX*STATE$ (产生一个按变量AGE，SEX 和字符变量 STATE$ 交叉分组的三维频数表)

TABULATE AGE，SEX*STATE$ (产生两个二维频数表)

TABULATE AGE*SEX/PERCENT (产生一个以总计数为分母的二维百分数表)

TABULATE AGE*SEX/ROWPCT (产生一个以行合计数为分母的二维百分数表)

TABULATE AGE*SEX/COLPCT (产生一个以列合计数为分母的二维百分数表)

TABULATE AGE/LIST (以清单形式列出频数表)

MODEL：指定一个对数线性模型，用以拟合前面TABULATE命令产生的表，并对其结果作拟合检验。

MODEL <变量> + <变量> + < ... > + <变量> * <变量> + < ... > [/FITTED DIFFERENCES RESIDUALS]，如：

表 7.1 护理工作评分比较资料

患者评分	科主任评分		
	差	一般	好
差	20	15	15
一般	30	30	20
好	25	30	15

MODEL AGE+SEX+AGE*SEX (规定一个二维模型，对指定的频数表拟合，并输出拟合值)
MODEL AGE*SEX (规定一个二维模型，与前面命令相同)
MODEL AGE+SEX (规定一个忽略交互作用的二维模型)
MODEL AGE*SEX+AGE*STATE+SEX*STATE (规定一个忽略变量AGE，SEX 和STATE的交互作用的三维模型)
MODEL AGE*SEX/RESIDUALS (规定一个二维模型，对指定的频数表拟合，并输出剩余值)
MODEL AGE*SEX/FITTED RESIDUALS (规定一个二维模型，对指定的频数表拟合，并输出拟合值和剩余值)

所谓n 维表就是依据n 个定性或等级变量将考察对象分组，数各组的考察对象个数而获得的频数表。

按一个定性或等级变量分组而得到的频数表称一维表，也就是通常意义上的频数表。按两个定性或等级变量交叉分组而得到的频数表称二维表，也就是通常所说的R x C 表。若依据分组的定性或等级变量多于两个，则称获得的频数表为多维表。

以下面的二维表为例说明在DATA 模块中建立产生n 维表的数据文件的方法。这是一个关于某医院评价护理人员工作质量的调查资料。在评价时，由患者和科主任同时给护士评分，评分结果分为好、一般和差三个等级。见表 7.1

首先，为两个等级变量取名，并将其值进行编码。即

科主任评分	GRADE1	GRADE1=1	差
		GRADE2=2	一般
		GRADE3=3	好
患者评分	GRADE2	GRADE2=1	差
		GRADE2=2	一般
		GRADE2=3	好

然后，进入DATA 模块，输入如下命令：

```
SAVE TABLE
INPUT GRADE1 GRADE2 FREQUEN
RUN
1 1 20
1 2 30
```

```
1 3 25
2 1 15
2 2 30
2 3 30
3 1 15
3 2 20
3 3 15
```

上述命令将在磁盘上建立一个名为TABLE.SYS的SYSTAT数据文件。注意文件中除两个等级变量GRADE1和GRADE2之外，还有一个数值变量FREQUEN，它是为存储二维表格内的频数而设立的。

在我们的例子中，各格子里的频数是按一种特殊的顺序输入的，也可以不这样做。事实上，在输入数据时，格子可以重复，TABULATE命令会自动地把相同格内的频数加在一起。

产生N 维表的具体步骤是：先用USE命令打开数据文件，然后打入WEIGHT命令指定频数变量，最后输入TABULATE命令产生所期望的N 维表。下面以前面建立的TABLE.SYS 文件为例，打入命令：

```
USE TABLE
WEIGHT=FREQUEN
TABULATE GRADE2*GRADE1
```

显示下面结果：

```
TABLE OF   GRADE2     (ROWS) BY    GRADE1    (COLUMNS)
FREQUENCIES                1          2          3     TOTAL
                    ---------------------------------
                  1 |     20         15         15   |     50
                  2 |     30         30         20   |     80
                  3 |     25         30         15   |     70
                    ---------------------------------
           TOTAL          75         75         50        200
```

如果TABULATE 命令加上相应的选择项，还可以得到以行合计，列合计或总合计为分母的百分数表。

对于二维表来说，如果在上面命令序列中的TABULATE命令之前，还打入了PRINT=LONG命令，那么，我们除了可以得到产生的表之外，还可以得到这个表的假设检验的结果。本例有：

```
TEST STATISTIC                      VALUE          DF          PROB
  PEARSON CHI-SQUARE                2.286           4          .683
  LIKELIHOOD RATIO CHI-SQUARE       2.305           4          .680
  MCNEMAR SYMMETRY CHI-SQUARE       9.500           4          .023

COEFFICIENT                         VALUE       ASYMPTOTIC STD ERROR
  PHI                               .1069
```

```
CRAMER V                          .0756
CONTINGENCY                       .1063
GOODMAN-KRUSKAL GAMMA            -.0203            .09709
KENDALL TAU-B                    -.0133            .06388
STUART TAU-C                     -.0131            .06283
COHEN KAPPA                      -.0093            .04845
SPEARMAN RHO                     -.0149            .07120
SOMERS D    (COLUMN DEPENDENT)   -.0134            .06395
LAMBDA      (COLUMN DEPENDENT)    .0400            .05813
UNCERTAINTY (COLUMN DEPENDENT)    .0053            .00697
```

一般表的假设检验则主要靠MODEL 命令进行的，即在产生表之后，打入MODEL 命令对指定模型进行假设检验，检验的结果将输出。我们还以前面二维表为例，说明一般表假设检验的步骤，命令序列为：

```
USE TABLE
WEIGHT=FREQUEN
TABULATE GRADE2*GRADE1
MODEL GRADE1+GRADE2
```

结果：

```
MODEL WAS FIT AFTER  2 ITERATIONS.
TEST OF FIT OF MODEL
DEGREES OF FREEDOM =   4
PEARSON CHI-SQUARE =       2.29    PROBABILITY =  .683
LIKELIHOOD RATIO CHI-SQUARE =       2.30    PROBABILITY =  .680
```

㈣CORR 模块

CORR 模块它是一个专门用来计算数据集合内变量间的相关或相似系数矩阵的统计模块。通过该模块的计算结果可以考察各变量的关联程度，为进一步的统计处理作准备。现介绍五个命令：

1. SSCP：计算文件中所有或指定数值变量的离均差平方和及各变量间的离均差叉积和，命令格式：

   SSCP [<变量>, <变量>, < ... >]，如：SSCP (计算文件中所有数值变量的离均差平方和及叉积和) SSCP HEIGHT，WEIGHT (计算变量HEIGHT和WEIGHT的离均差平方和及叉积和)

2. COVARIANCE：计算文件中所有或指定数值变量的方差及各变量间的协方差，命令格式：

   COVARIANCE [<变量>, <变量>, < ... >]，如：

   COVARIANCE (计算文件中所有变量的方差及协方差)

   COVARIANCE HEIGHT，WEIGHT (计算变量HEIGHT 和WEIGHT 的方差及协方差)

3. PEARSON：计算文件中所有或指定数指变量间的PEARSON 乘积矩阵相关系数，命令格式：

   PEARSON [<变量>,<变量>,< ... >]，如：

   PEARSON (计算文件中所有数值变量的相关系数)

   PEARSON HEIGHT，WEIGHT (计算指定变量HEIGHT 和WEIGHT 的相关系数)

4. SPEARMAN：将文件中所有或指定数值变量排序编秩，计算SPEARMAN 等级相关系数，命令格式：

   SPEARMAN [<变量>,<变量>,< ... >]，如：

   SPEARMAN (计算文件中所有数值变量的等级相关系数)

   SPEARMAN HEIGHT，WEIGHT (计算变量HEIGHT 和WEIGHT 的等级相关系数)

5. EUCLIDEAN：计算文件中所有或指定数值变量间的欧氏距离，并用样本含量N 相除得其平均距离，命令格式：

   EUCLIDEAN [<变量>,<变量>,< ... >]，如：

   EUCLIDEAN (计算文件中所有数值变量的平均距离)

   EUCLIDEAN HEIGHT，WEIGHT (计算变量HEIGHT 和WEIGHT 的平均距离)

在CORR 模块中，统计命令的使用很简单.只要打开数据文件，就可打入上述统计命令中的任意一条。计算各变量间的相应统计量。下面以计算PEARSON 相关系数为例进行讨论。以前面的医院工作质量数据为例，计算X1，X2 ，X3 之间的PEARSON相关系数，命令如下：

```
USE HOSPITAL
FORMAT=5
PEARSON X1 X2 X3
```

命令序列中FORMAT=5 是用来规定输出结果的小数位数，计算结果：

```
PEARSON CORRELATION MATRIX
                    X1          X2          X3
     X1            1.00000
     X2           -0.80204     1.00000
     X3            0.00503     0.27146     1.00000
NUMBER OF OBSERVATIONS:   12
```

如果打算在CORR 模块存储某个统计命令的计算结果的话，只需在相应统计命令之前打入形如：SAVE <文件名> 的命令。这样做就可以把后面统计命令的计算结果(如相关系数)，存储在以SAVE 命令中的¡文件名¿ 为名的SYSTAT 数据文件中，利用SYSTAT 中的其他模块，可以对这个文件的数据进行各种统计处理。下面的命令是计算文件HOSPITAL 中三个变量的相关系数，并将其存入指定的名为CORRHOS 文件中，命令如下：

```
USE HOSPITAL
SAVE CORRHOS
PEARSON X1 X2 X3
```

㈤MGLH 模块

MGLH模块是一个高级统计模块。它不仅能够估计各种单变量或多变量的一般线性模型的参数，而且能够对其参数的线性假设进行检验。因此，MGLH 模块的功能比大多数的回归程序要强得多。利用它可以很容易地进行简单回归(一个因变量对一个自变量的回归)，多元回归(一个因变量对多个自变量的回归) 和多因变量的多元回归(多个因变量对多个自变量的回归)；各种试验设计的单元或多元方差分析；以及其他一些多元统计分析方法，如多变量断面分析，线性判别分析，典型相关分析等等。这里介绍其中8 条命令的功能及格式。

1. CATEGORY：指定一个或多个数值变量为分组变量。并限定分组的个数。命令格式：

   CATEGORY <变量>=<数值> [<变量>=<数值>, < ... >]，如：

   CATEGORY GROUP=3 (指定变量GROUP为分组变量，并限定其组数为三组)

   CATEGORY GROUP=3 SEX=2 (指定变量GROUP 和SEX 为分组变量，并限定其组数分别为三组和两组)

2. MODEL：规定一个线性模型，命令格式：

   MODEL <变量> [, <变量>, < ... >] = [CONSTANT+]<变量> [+ <变量> + < ... >][+ <变量> ∗ <变量> + < ... >]，如：

   MODEL Y=CONSTANT+X (指定一个因变量Y，自变量X 的简单线性模型)

   MODEL Y=CONSTANT+X1+X2+X3 (指定一个因变量Y，自变量X1,X2,X3 的线性模型) MODEL Y1,Y2,Y3=CONSTANT+X1+X2+X3 (规定多个变量Y1,Y2,Y3 为因变量，变量X1，X2，X3 为自变量的简单线性模型)

3. ESTIMATE：对MODEL 命令所规定的线性模型作最小二乘估计。

4. STEP：对MODEL命令所规定的线性模型作逐步回归，命令格式：

   STEP [/ENTER=<数值>，REMOVE=<数值>]

   STEP (按标准阈值ENTER=.15,REMOVE=.15 作逐步回归)

   STEP /ENTER=.3，REMOVE=.3

   注意：引进变量(ENTER) 的阈值必须小于等于剔出变量(REMOVE) 的阈值。

5. HYPOTHESIS：进入假设检验程序，标志假设检验的开始，命令格式：HYPOTHESIS

6. EFFECT：指定线性模型中一个自变量或一个自变量组合，以便对其系数或系数组合进行假设检验，命令格式：

   EFFECT=X (规定对变量X的系数进行假设检验)

   EFFECT=X*GROUP (规定对变量X,GROUP的系数组合进行假设检验)

7. CONTRAST 规定一个多重比较假设，以便对EFFECT 命令所指定的系数或系数组合进行多重假设检验，命令格式：

   CONTRAST <数值>, <数值>, <数值> [, <数值>, <数值>, < ... >]. 如

   CONTRAST

   1,-1,0,0

   注：输入的数值之和必须为零。

表 7.2 工作人员能力和生产效率打分

能力(X)	生产效率(Y)	能力(X)	生产效率(Y)
41	32	38	29
35	20	38	33
34	35	46	36
40	24	36	23
33	27	32	22
42	28	43	38
37	31	42	26
42	33	30	20
30	26	41	30
43	41	45	30

8. TEST 按所规定假设进行假设检验，命令格式：TEST

以上命令，其中ESTIMATE,STEP,TEST命令是热命令，其余均为冷命令。

关于MGLH 模块的应用，我们将介绍如何使用MGLH 模块中的统计命令去进行回归分析。1. 简单回归

例：工作人员能力测定与其生产效率(表 7.2)

假设我们已经建立了包含这些数据的SYSTAT 文件，其文件名为LEVE.SYS 文件中的两个数值变量分别为X 和Y。则做简单回归的步骤为：首先进入MGLH 模块，打入命令：

```
USE LEVE
MODEL Y=CONSTANT+X
ESTIMATE
```

显示器显示下面的结果

```
DEP VAR:     Y   N:  20   MULTIPLE R: .609   SQUARED MULTIPLE R: .371
ADJUSTED SQUARED MULTIPLE R: .336  STANDARD ERROR OF ESTIMATE:  4.769

VARIABLE  COEFFICIENT  STD ERROR   STD COEF TOLERANCE   T   P(2 TAIL)
CONSTANT      1.016       8.710       0.000 1.0000000   0.117   0.908
       X      0.734       0.225       0.609 1.0000000   3.260   0.004

                        ANALYSIS OF VARIANCE
  SOURCE   SUM-OF-SQUARES     DF  MEAN-SQUARE     F-RATIO       P
REGRESSION        241.766      1      241.766      10.629       0.004
  RESIDUAL        409.434     18       22.746
-------------------------------------------------------------------------
```

2.多元回归

表 7.3 13个疾病观察点的发病水平及病因学因素

观察点编号	疾病学因素				发病水平
NO	X1	X2	X3	X4	Y
1	7	26	6	60	78.5
2	1	29	15	52	74.3
3	11	56	8	20	104.3
4	11	31	8	47	87.6
5	7	52	6	33	95.9
6	11	55	9	22	109.2
7	3	71	17	6	102.7
8	1	31	22	44	72.5
9	2	54	18	22	93.1
10	21	47	4	26	115.9
11	1	40	23	34	83.8
12	11	66	9	12	113.3
13	10	68	8	12	109.4

例：设调查了13 个疾病观察点的某病发病水平(Y) 及一组病因学观察指标(X1,X2,X3,X4)，资料见表 7.3。

并假设数据已存入MULREG.SYS 文件中。则多元回归的操作步骤为：进入MGLH 模块，然后键入如下命令

```
USE MULREG
MODE Y=CONSTANT+X1+X2+X3+X4
ESTIMATE
```

运算结果如下：

```
DEP VAR:  Y    N: 13   MULTIPLE R: .991   SQUARED MULTIPLE R:   .982
ADJUSTED SQUARED MULTIPLE R: .974  STANDARD ERROR OF ESTIMATE: 2.446

VARIABLE  COEFFICIENT  STD ERROR   STD COEF TOLERANCE   T   P(2 TAIL)
CONSTANT    62.405       70.071      0.000  1.0000000  0.891    0.399
      X1     1.551        0.745      0.607   .0259766  2.083    0.071
      X2     0.510        0.724      0.528   .0039305  0.705    0.501
      X3     0.102        0.755      0.043   .0213363  0.135    0.896
      X4    -0.144        0.709     -0.160   .0035397 -0.203    0.844
 REGRESSION       2667.899     4      666.975     111.479       0.000
   RESIDUAL         47.864     8        5.983
-------------------------------------------------------------------------
```

3.逐步回归

我们仍以上面的疾病病因学研究为例。则逐步回归的运算步骤为：进入MGLH 模块，然后使用命令：

```
USE MULREG
MODEL Y=CONSTANT+X1+X2+X3+X4
STEP/ENTER=.2,REMOVE=.2
```

输出结果:

```
STEPWISE REGRESSION WITH ALPHA-TO-ENTER= .200 AND ALPHA-TO-REMOVE= .200

  STEP=  1     ENTER       X4     R=  .821     RSQUARE=  .675
  STEP=  2     ENTER       X1     R=  .986     RSQUARE=  .972
  STEP=  3     ENTER       X2     R=  .991     RSQUARE=  .982
  STEP=  4    REMOVE       X4     R=  .989     RSQUARE=  .979

  THE SUBSET MODEL INCLUDES THE FOLLOWING PREDICTORS:
          CONSTANT
              X1
              X2
USE  THESE  PREDICTORS  IN  A  NEW  MODEL  SENTENCE  TO  ESTIMATE  THE
COEFFICIENTS.
```

STEP 命令仅仅输出了筛选步骤，并给出了每一步的复相关系数和决定系数。最后得到的是在给定的显著性水平下保留的预测因子(即自变量)，程序称之为子集(SUBSET)。本例只有X1和X2在0.2水平下有显著意义。要求这个回归方程只要用MODEL 语句加上这些预测因子，然后键入ESTIMATE 命令即可。

```
USE MULREG
MODEL Y=CONSTANT+X1+X2
ESTIMATE
```

输出结果:

```
  DEP VAR:   Y   N:  13   MULTIPLE R: .989  SQUARED MULTIPLE R:  .979
  ADJUSTED SQUARED MULTIPLE R: .974  STANDARD ERROR OF ESTIMATE:2.406

  VARIABLE  COEFFICIENT  STD ERROR   STD COEF TOLERANCE   T  P(2 TAIL)
  CONSTANT    52.577       2.286      0.000   1.0000000  22.998  0.000
        X1     1.468       0.121      0.574    .9477514  12.105  0.000
        X2     0.662       0.046      0.685    .9477514  14.442  0.000

                          ANALYSIS OF VARIANCE

     SOURCE   SUM-OF-SQUARES    DF  MEAN-SQUARE     F-RATIO       P
```

表 7.4 甘蓝叶中核黄素之浓度

高锰酸钾			高锰酸钾		
样本重量	处理	未处理	样本重量	处理	未处理
	27.2	39.5		24.6	38.6
0.25g	23.2	43.1	1.00g	24.2	39.5
	24.8	45.2		22.2	33.0

表 7.5 甘蓝叶中核黄素之浓度($\mu g/g$)

测定次数	经高锰酸钾处理		未经高锰酸钾处理	
	0.25g 样本	1.00g 样本	0.25g 样本	1.00g 样本
1	27.2	24.6	39.5	38.6
2	23.2	24.2	43.1	39.5
3	24.8	22.2	45.2	33.0

```
REGRESSION       2657.859     2     1328.929      229.504       0.000
  RESIDUAL         57.904    10        5.790
-------------------------------------------------------------------------
```

从输出的参数来看，取 x 1和 x 2两个因子来预测的效果较好。其回归方程为: Y=52.577+1.468X1+0.662X2

4. 析因设计的方差分析(1)2 x 2 析因设计的方差分析

2 x 2 设计是指有两个因素，每个因素有两个水平的实验设计，共有2 x 2=4 个组，各因素各水平均相遇一次。这是析因设最简单的一种行式。如以a1 表示a 因素1 水平，a2 表示a 因素2 水平，b1 表示b 因素1 水平，b2 表示b 因素2 水平，则2 x 2 设计的模型为：

a_1b_1	a_1b_2
a_2b_1	a_2b_2

表 7.4是甘蓝叶中核黄素含量($\mu g/g$)的荧光测定结果。所用的甘蓝叶有经过二氧化氢高锰酸钾处理的，也有未经处理的，甘蓝叶的样本有0. 25g 与1g两种。试问高锰酸钾处理与否样本量不同对甘蓝叶中核黄素含量的测定结果有无显著差别。

本题甘蓝叶的处理方法是一个因素，分处理与否两个水平；样本重量为另一个因素，分为0.25g 与1g 两个水平；每种组合都经三次测定。为了便于对应于2×2 析因设计模型建立数据文件，现将表 7.4整理成以下形式：

和两因素方差分析一样，建立数据文件对每个因素各设一个变量，水平取值用1,2,3,...表示。现设处理因素为TREAT，取值1 表示用高锰酸钾处理，2 表示未处理，重量因素为WEIGHT，取值 1 表示0.25g，2 表示1.00g。因考虑到测定次数可能引起的误差，我们把测定次数看作区组，用变量BLOCK 表示，取值1，2，3，count 为测量值，RESULT 作为测量变量。

设数据存于文件EXAM111，分析步骤如下：

```
C:\SYSTAT>MGLH
```

```
>USE EXAM111
>CATEGORY TREAT=2,WEIGHT=2,BLOCK=3
>MODE RESULT=CONSTANT+TREAT+WEIGHT+BLOCK+TREAT*WEIGHT
>ESTIMATE
```

这里MODEL 语句从形式上看是三因素方差分析，但我们主要研究的因素是TREAT 和WEIGHT，BLOCK只是为控制不同时间测量的影响而设计的变量。TREAT *WEIGHT 就是指定求两因素的交互作用。

输出结果：

```
DEP VAR:RESULT  N:  12  MULTIPLE R:  .970  SQUARED MULTIPLE R: .940

              ANALYSIS   OF   VARLANCE
 SOURCE  SUM-OF-SQUARES   DF   MEAN-SQUARE  F-RATIO      P
 TREAT         716.108     1      716.108    87.547     0.000
WEIGHT          36.401     1       36.401     4.450     0.079
 BLOCK           3.762     2        1.881     0.230     0.801
TREAT*
WEIGHT          13.021     1       13.021     1.592     0.254
ERROR           49.078     6        8.180
```

结果表明，只有TREAT 有极显著性差异。即样品用高锰酸钾处理与否对测定结果有极显著的影响。

对析因设计资料作方差分析时，一般来说，如果计算出来的交互作用没有显著意义的话，可以把这部分的差异与误差项合并，然后重新计算 F 值。用程序计算时，只要从MODEL语句中去掉交互项即可。下面是本题去掉交互项后的方差分析结果。

```
DEP VAR:RESULT  N:  12  MULTIPLE R:  .961  SQUARED MULTIPLE R: .924

              ANALYSIS   OF   VARLANCE
 SOURCE  SUM-OF-SQUARES   DF   MEAN-SQUARE  F-RATIO      P
 TREAT         716.108     1      716.108    80.722     0.000
WEIGHT          36.401     1       36.401     4.103     0.082
 BLOCK           3.762     2        1.881     0.212     0.814
ERROR           62.099     7        8.871
```

去掉交互项后，结论仍和前面一样。但是由于误差项的离均差平方和增大，各因素的F 值变小了。

⑵$3\times2\times2$ 析因设计的方差分析

$3\times2\times2$ 设计是指有三个因素，其中一个因素有3个水平，另两个因素各有2 个水平的实验设计。这些因素的水平在一个设计中相互组合一次。

表 7.6是一个钩端螺旋体的资料，血清种类有兔血清与胎盘血清两种，每种血清有5%与8% 两种浓度，所有基础液有三种，即缓冲剂、蒸馏水与自来水。试分析钩端螺旋体计数学清种类、血清浓度及基础液种类的关系。为了便于说明，令A表示基础液，B 表示血清种类，C 表示浓度，按各因素的水平数，构成$3\times2\times2$ 实验。试作三因素析因设计方差分析。

表 7.6 $3 \times 2 \times 2$ **析因实验结果(钩端螺旋体计数)**

基础液 (A)	血清种类(A)			
	兔血清		胎盘血清	
	血清浓度(C)		血清浓度(C)	
	5%	8%	5%	8%
缓冲剂	648	1144	830	578
	1246	1877	853	669
	1398	1671	441	643
	909	1845	1030	1002
蒸馏水	1763	1447	920	933
	1241	1883	709	1024
	1381	1896	848	1092
	2421	1962	574	742
自来水	508	1789	1126	685
	1026	1215	1176	546
	1026	1434	1280	595
	830	1651	1212	566

建立数据文件，各因素的变量名称和各水平的取值意义如下：

变量名	取值意义
基础液A	1－缓冲剂，2－蒸馏水，3－自来水
血清种类B	1－兔血清，2－胎盘血清
血清浓度C	1%～5%，2%～8%
测量值COUNT	

设数据文件存于EXAM112，分析步骤如下：

```
C:\SYSTAT>MGLH
 >USE EXAM112
 >CATEGORY A=3,B=2,C=2
 >MODE COUNT=CONSTANT+A+B+C+A*B+A*C+B*C+A*B*C
 >ESTIMATE
```

本题有三个因素，可以构成高阶交互项。一般这种检验都从高阶交互项开始，如果高阶交互项不显著，则把它从模型中去掉，然后再检验低阶交互项如果都不显著，则构造一个只有主效应的模型。上面MODEL 语句包含了所有交互项，故称“饱和”模型，输出结果：

```
DEP VAR:COUNT   N:  48  MULTIPLE R:  .872  SQUARED MULTIPLE R: .761

             ANALYSIS   OF   VARLANCE
 SOURCE  SUM-OF-SQUARES   DF   MEAN-SQUARE  F-RATIO      P
    A      692513.375     2   346256.688     4.978      0.012
    B     4142462.521     1  4142462.521    59.551      0.000
```

```
      C      248976.021     1   248976.021     3.579      0.067
     A*
      B      726923.042     2   363461.521     5.225      0.010
     A*
      C       99091.542     2    49545.771     0.712      0.497
     B*
      C     1111729.688     1  1111729.688    15.982      0.000
     A*
     B*
      C      946085.375     2   473042.688     6.800      0.003
  ERROR     2504233.750    36    69562.049
```

检验的结果表明，除主效应C和A*C交互项外，其他各项均有显著意义。现将A*C 从MODE语句中去掉，重新检验，结果如下:

```
DEP VAR:COUNT   N:  48  MULTIPLE R:  .867  SQUARED MULTIPLE R: .751

              ANALYSIS   OF   VARLANCE
 SOURCE  SUM-OF-SQUARES   DF   MEAN-SQUARE  F-RATIO      P

      A      692513.375     2   346256.688     5.054      0.011
      B     4142462.521     1  4142462.521    60.466      0.000
      C      248976.021     1   248976.021     3.634      0.064
     A*
      B      726923.042     2   363461.521     5.305      0.009
     B*
      C     1111729.688     1  1111729.688    16.228      0.000
     A*
     B*
      C      946085.375     2   473042.688     6.905      0.003
  ERROR     2603325.292    38    68508.560
```

去掉A*C交互项后结论不变。

由于本题三个因素ABC 的交互作用显著，主效应诸均数比较时，就有可能被交互作用所掩盖。在这种情况下，可将一个因素固定在一定水平上，用Duncan 多重极差检验来比较在该水平下处于另一因素的诸均数。

首先进入STATS 模块，用分组统计命令打印小组(共12 组) 的均数：

```
C:\SYSTAT>STATS
 >USE EXAM112
 >BY A,B,C
 >STATISTICS COUNT/MEAN
```

为了便于分析，现将各组均数从以上输出中整理于下表。

(注:括号中的数字对应于各因素变量的取值)

表 7.7 各小组均数

血清种类	浓度	基础液(A)		
(B)	(C)	缓冲剂(1)	蒸馏水(2)	自来水(3)
兔血清	(1) 5%(1)	1050.25	1701.50	847.50
	8%(2)	1634.25	1788.00	1522.25
胎盘血清	(2) 5%(1)	788.50	762.75	1198.50
	8%(2)	723.00	947.75	598.00

计算Duncan 多重极差检验的显著性界值(仍在STATS 模块下)。这一步运算需要用到前方差分析的结果，并通过键盘输入。命令格式如下：

DUNCAN/K=组数，MSE=误差的均方

ALPHA=显著性水平，DFE=误差的自由度

N=每组观察例数

本题的计算操作为:

>DUNCAN/K=3，MSE=68508.56

ALPHA=0.05; DFE=38; N=4

因本题最多只有3个组对比，所以K=3。结果输出:

```
DUNCAN  MULTIPLE  RANGE  TESTS
ORDERED MEANS DIFFER AT ALPHA=  .050 F THEY EXCEED FOLLOWING GAPS
GAP ORDER DIFFERENCE
      1      374.850
      2      393.968
THIS TEST ASSUMES THE COUNTS PER GROUP ARE EQUAL
```

结果中的差值(DIFFERENCE)即为显著性检验的极差值。这里我们给定ALPHA=0 .05，这就意为着，如果排序后的两均数之差大于相应间隔(GAP)的差值，则在此0 .05水平上拒绝无效假设。

下面计算每两均数的相差与间隔的差值比较确定显著性。加“*”号的表示P¡0. 05。

①在B x C 同一水平上比较A (三种基础液)

B x C 兔学清浓度5对比组	差值
蒸馏水与缓冲剂	651.25
蒸馏水与自来水	854.5*
缓冲剂与自来水	202.75
B x C 兔血清浓度8%	
对比组	差值
蒸馏水与缓冲剂	153.75
缓冲剂与自来水	112.00
蒸馏水与自来水	265.75
B x C 胎盘血清浓度5 %	
对比组	差值
自来水与缓冲剂	410.00
蒸馏水与蒸馏水	25.75
自来水与蒸馏水	435.75
B x C 胎盘血清浓度8%	
对比组	差值
蒸馏水与缓冲剂	224.75
缓冲剂与自来水	125.00
蒸馏水与自来水	349.75

②在A x C 同一水平上比较B (两种血清)

基础液	浓度	兔血清	胎盘血清	差值
缓冲剂	5 %	1050.25	788.50	261.75
缓冲剂	8 %	1634.25	723.00	911.25*
蒸馏水	5 %	1701.50	762.75	938.75*
蒸馏水	8 %	1788.00	947.75	840.25*
自来水	5 %	847.50	1198.50	-351.00
自来水	5 %	1522.25	598.00	924.25*

③在A x B 同一水平上比较C（两种血清浓度）

基础液	血清种类	浓度8%	浓度5%	差值
缓冲剂	兔血清	1634.25	1050.25	584.00*
蒸馏水	兔血清	1788.00	1701.50	85.50
自来水	兔血清	1522.25	847.50	674.75*
缓冲剂	胎盘血清	723.00	788.50	-65.50
蒸馏水	胎盘血清	947.75	762.75	185.00
自来水	胎盘血清	598.00	1198.50	-600.50*

从方差分析来看，主效应C (浓度间) 差别不显著，当在A x B 同一水平比较时，则有三对均数的差别有显著意义，说明前者由于交互作用显著掩盖了浓度均数间的显著性。通过上述比较不难得出结论：用兔血清浓度为8馏水为基础液时，钩端螺旋体计数较高。

⑶正交试验设计的方差分析

当研究的因素超过三个时，并且因素间又有可能存在交互作用，可用正交试验设计。正交试验是将各试验因素，各水平进行合理组合，均匀搭配，由此大大减少试验次数而又能得到较多的信息。正交试验设计的分析可采用方差分析，它把总变异的离均差平方和及其自

表 7.8 过氧乙酸稳定性试验的因素分析及水平

试验因素	水平	
	1	2
A：稳定剂	加磷酸0.3%	不加磷酸
B: 水溶温度	25~30℃	35 40℃
C: 浸泡口表	浸泡口表10支	不浸泡口表
D: 加盖与否	加盖	不加盖

表 7.9 过氧乙酸定性试验安排及其结果

试验号	不同因素的水平号		24 小时过氧乙酸		残存量(mg/3ml)	
	A	B	C	D	1	2
1	1	1	1	1	7.00	4.11
2	1	1	2	2	6.05	3.50
3	1	2	1	2	1.10	0.80
4	1	2	2	1	1.90	0.96
5	2	1	1	2	2.40	1.65
6	2	1	2	1	4.00	1.50
7	2	2	1	1	0.35	0.30
8	2	2	2	2	0.30	0.90

由度分为各因素的各水平间，因素间的交互作用及误差几部分，因此能经确地说明各因素诸水平间的差别，确切地判断各因素间的交互作用。

过氧乙酸是广泛应用的一种杀灭病毒性肝炎病毒的主要消毒剂，但其有效成份极不稳定，以致影响其消毒效果。现欲通过实验找出有关因素对其稳定性的影响，并指出哪个是主要的，哪个是次要的，哪个起交互作用，尔后选出各因素中的一个最佳水平，组成保持过氧乙酸稳定性的最优条件。

本例试验因素及其水平数如表 7.8：

本题有4 个因素，每个因素各有2 个水平，此外还要观察稳定剂与温度(A *B)、稳定剂与加盖与否(A*D)的交互作用。试验选用L8(27)正交表，试验安排及测定结果见下表。

用SYSTAT 程序作用多因素方差分析，交互作用只要在模型中指定就能自动计算，所以正交表中的交互列在建立数据文件时不必考虑，误差列也是如此，上面的正交表每列设一个变量。每个变量都按水平号取值，另设一个测量值变量(本题设为VAL)。

设数据存于文件EXAM13，分析步骤如下：

```
C:\SYSTAT>MGLH
 >USE EXAM113
 >CATEGORY A=2,B=2,C=2,D=2
 >MODE VAL=CONSTANT+A+B+C+D+A*B+A*D
 >ESTIMATE
```

结果输出:

```
DEP VAR      X    N:  16  MULTIPLE R:  .896  SQUARED MULTIPLE R: .803

              ANALYSIS   OF   VARLANCE
SOURCE  SUM-OF-SQUARES   DF   MEAN-SQUARE  F-RATIO      P
    A          12.205     1       12.285     9.708     0.011
    B          34.810     1       34.810    27.507     0.000
    C           0.123     1        0.123     0.097     0.762
   A*
    B           4.202     1        4.202     3.321     0.098
   A*
    D           0.164     1        0.164     0.130     0.726
ERROR          12.655    10        1.266
```

方差分析结果表明，稳定剂(A)与水浴温度(B)是影响过氧乙酸稳定性两个最主要因素，A与B及A与D的交互作用不显著。为了便于选择最优水平，我们可以在SYSTAT 模型下用分组统计命令计算出 4 个因素各水平的合计数，结果如下：

	A	B	C	D
水平1	25.42	30.21	17.71	20.12
水平2	11.40	6.61	19.11	16.70

数值大者说明过氧乙酸残存量多，更有利于过氧乙酸的稳定。通过比较不难看出，当A、B因素取 1 水平时效果较好。C、D 两因素的作用不显著，可以任选一水平，不过D 因素的1 水平从数值上还是明显大于2 水平，故选1 水平为宜。最后结论：加0.3件。

5. 协方差分析

是把直线回归与方差分析结合起来的一种统计方法。它利用回归的关系把与因变量Y值呈直线关系的自变量X值化成相等后，再进行方差分析。它比较的是调正均数间的差异。通过协方差分析，能够校正由于各组X值的不同所引起的偏差，更恰当地评价各种处理的优劣。

⑴完全随机设计的协方差分析

男性运动员和大学生的平均肺活量分别为4399cm，3667cm，经假设检验有差别。但我们已知肺活量与身高有一定关系(一般来说，肺活量随身高增大而增大)。本例中运动员的身高高于大学生，因此在比较肺活量时必须对身高作校正，这就需用协方差分析进行处理。

协方差分析的前提条件是，各组资料（样本）都来自方差相同的正态分布总体；各组的回归系数本身有显著性意义，但各个回归系数间无显著性差别，即斜率是齐性的。关于资料的方差齐性检验可用STATS 模块的STATISTICS 命令去完成。协方差分析的数据格式与方差分析基本相似，也要设置一个分组变量来标识对比的各组，所不同的是增加了一个自变量。本假设分组变量为GROUP；运动员取值为1，大学生取值为2。在作协方差分析之前，我们先检验一下两组的斜率是否相同。所谓斜率齐性检验就是对协变量X和分组变量GROUP的交互项作假设检验。

设数据存于文件EXAM121，分析步骤如下：

```
C:\SYSTAT>MGLH
 >USE EXMA121
```

表 7.10 20 运动员及大学生的身高(X,cm)与肺活量(Y,cm^3)

运动员		大学生		运动员		大学生	
X1	Y1	X2	Y2	X1	Y1	X2	Y2
184.9	4300	168.7	3450	169.0	4500	173.8	4150
167.9	3850	170.8	4100	188.0	4780	174.0	3450
171.0	4100	165.0	3800	176.7	3700	170.5	3250
171.0	4300	169.7	3300	179.0	5250	176.0	4100
188.0	4800	171.5	3450	183.0	4250	169.5	3650
179.0	4000	166.5	3250	180.5	4800	176.3	3950
177.0	5400	165.0	3600	179.0	5000	163.0	3500
179.5	4000	165.0	3200	178.0	3700	172.5	3900
187.0	4800	173.0	3950	164.0	3600	177.0	3450
187.0	4800	169.0	4000	174.0	4050	173.0	3850

```
>CATEGORY GROUP=2
>MODE Y=CONSTANT+GROUP+X+GROUP*X
>ESTIMATE
```

因为交互项GROUP*X 的作用不显著(P=0.859)，故认为两组的回归斜率无显著差异。从上面的结果可以看出，自变量X的作用有显著意义，这说明总体直线回归系数不为0。如果X无显著性，则作协方差分析就无意义了。在这种情况下可以不考虑X 的影响，直接作方差分析即可。

现在我们拟合一个上面资料的协方差分析模型：

```
>CAEGORY GROUP=2
>MODE Y=CONSTANT+GROUP+X
>ESTIMATE
```

如果熟悉前面的方差分析和线性回归模型就会发现，协方差模型正是两者的组合。

```
                ANALYSIS  OF  VARIANCE

SOURCE  SUM-OF-SQUARES  DF  MEAN-SQUARE   F-RAIO   P
 GROUP     1407847.095   1  1407847.095    9.220   0.004
     X     1630762.635   1  1630762.635   10.679   0.002
 ERROR     5649992.365  37   152702.496
```

结果表明，均衡了协变量的影响后，两组间的肺活量仍有极显著差别。

协方差分析是对修正均数的差别进行显著性检验，程序并没有输出修正均数。若要显示修正均数，必须在ESTIMATE命令之前打入一条存SYSTAT文件的命令，并加上选择项ADJUSTED,如SAVE MYFILE/ADJUSTED。这样程序就会在MYFILE 中产生一个名为ESTIMATE的变量用于存放修正均数。由于这个文件没有分组变量，所以用户要记住每一组在文件中的

表 7.11 三组大鼠的进食量(X，g) 与所增体重(Y，g)

窝别	(1)核黄素缺乏组		(2)限食量组		(3)不限食量组	
	X	Y	X	Y	X	Y
1	256.9	27.0	260.3	32.0	544.7	160.3
2	271.6	41.7	271.1	47.7	481.2	91.6
3	210.2	25.0	214.7	36.7	418.9	114.6
4	300.1	52.0	300.1	65.0	556.6	134.8
5	262.2	14.5	269.7	39.0	394.5	76.3
6	304.4	48.8	307.5	37.9	426.6	72.8
7	272.4	48.0	278.9	51.5	416.1	99.4
8	248.2	9.5	256.2	26.7	549.9	133.7
9	242.8	37.0	240.8	41.0	580.5	147.0
10	342.9	56.5	340.7	61.3	608.3	165.8
11	365.9	76.0	365.3	102.1	559.6	169.8
12	198.2	9.2	199.2	8.1	371.9	54.3

起始记录号。当然，也可以用下面的方法把原文件的分组变量与ESTIMATE变量连结后形成一个新文件,便于结果的阅读。

```
C:\SYSTAT>DATA
 >USE EXAM121(GROUP)MYFILE(ESTIMATE)
 >SAVE NEWFILE
 >RUN
```

新产生的NEWFILE 文件包含GROUP 和ESTIMATE 两个变量，用LIST 命令即可看到GROUP 的不同取值对应不同的修正均数为3805.836。由此作出结论，两组肺活量均数在消除身高因素的影响后仍有极显著差别，运动员的肺活量大于大学生。

⑵随机区组设计的协方差分析

如果实验中包含两因素，其中一个因素的记录具有依存关系(直线关系) 的成对(X，Y) 数值，也可用协方差分析。

在“核黄素缺乏对于蛋白质利用的影响之研究”中，将体重相近(30 38g),出生三周的大鼠36只,按照窝别、性别等条件分成12窝,每窝3只,随机分到三个不同饲料组进行喂养。观察记录列于表 7.11，观察黄素缺乏对体重增长的影响。

本题作协方差分析要设置两个分组变量。处理组变量设为GROUP，取值1 表示核黄素缺乏组，2 表示限食量组，3 表示不限食量组;窝别变量设为BLOCK，取值1 12，表示12 区组;进食量X 为协变量; 所增体重Y 为因变量。

设数据存于EXAM122，分析步骤如下：

先作斜率的齐性检验。

```
C:\SYSTAT>MGLH
 >USE EXAM122
 >CATEGORY GROUP=3,BLOCK=12
```

```
 >MODE Y=CONSTANT+GROUP+BLOCK+X+GROUP*X
 >ESTIMATE
```

结果输出:

```
DEP VAR:    Y    N:  36  MULIPLE R: .986  SQUARED MULIPLE R:  .971

               ANALYSIS  OF  VARIANCE
SOURCE  SUM-OF-SQUARES  DF  MEAN-SQUARE   F-RAIO   P
 GROUP         105.208   2       52.604    0.461   0.637
 BLOCK        3765.619  11      342.329    3.001   0.017
     X        2827.148   1     2827.148   24.788   0.000
GROUP*
     X          66.105   2       33.052    0.290   0.752
 ERROR        2167.034  19      114.054
```

交互项的F值的概率=0.752,说明各组回归的斜率无显著性差异。下面作协方差分析:

```
>MODEY=CONSTANT+GROUP+BLOCK+X
>SAVE MODY/ADJUSTED
>ESTIMATE
```

在配合模型时用SAVE命令指定存贮修正均数,文件名为MODY。结果输出;

```
DEP VAR:    Y    N:  36  MULIPLE R: .985  SQUARED MULIPLE R:  .971

               ANALYSIS  OF  VARIANCE
SOURCE  SUM-OF-SQUARES  DF  MEAN-SQUARE   F-RAIO   P
 GROUP         469.157   2      234.578    2.206   0.135
 BLOCK        3761.319  11      341.938    3.216   0.010
     X        6175.031   1     6175.031   58.069   0.000
 ERROR        2233.139  21      106.340
```

协方差分析结果表明，当消耗食物量相同时，各不同饲料组大鼠平均增重没有明显不同。本题如果不考虑进食量，仅作随机区组的方差分析，则处理间比较的概率小于0.01,结论恰恰相反。

由于协方差模型包含区组变量，所以MODY文件中同一处理组的各例的修正均数不尽相同。要求各处理组的修正均数，可在STATS模块下按GROUP分组计算ESTIMATE 均数,操作方法如下:

```
C:\SYSTAT>DATA
 >USE EXAM122(GROUP)MODY(ESTIMATE)
 >SAVE NEWFILE
 >RUN
C:\SYSTAT>STATS
 >USE NEWFILE
```

表 7.12 六组公鼠的食物消耗量(X,10cal)及所增体重(Y,g)

高蛋白						低蛋白					
牛	肉	谷	类	猪	肉	牛	肉	谷	类	猪	肉
X	Y	X	Y	X	Y	X	Y	X	Y	X	Y
108	73	99	98	194	94	165	90	124	107	140	49
136	102	117	74	198	79	164	76	95	95	177	82
138	118	90	56	196	96	161	90	116	97	189	73
159	104	141	111	198	98	159	64	112	80	142	86
146	81	106	95	210	102	175	86	123	98	216	81
141	107	112	88	196	102	135	51	110	74	200	97
175	100	110	82	230	108	132	72	137	74	255	106
149	87	117	77	222	91	190	90	105	67	173	70
174	117	111	86	220	120	145	95	135	89	153	61
176	111	122	92	228	105	142	78	126	58	160	82

```
>BY GROUP
>STATISTICS ESTIMATE/MEAN
```

⑶析因设计的协方差分析

在析因设计时，如果对比的各水平要考虑协变量的影响就应用析因设计的协方差分析。

将60只公鼠随机分成六组,分别饲以不同来源及成分的蛋白质，并记录食物消耗(X,10cal),所增体重(Y,g)于表 7.12，试作分析。

本题主要分析的因素有两个，蛋白质含量的高低和蛋白质的食物来源。前者用变量 A 来表示，取值1 表示高蛋白组，2 表示低蛋白组；后者用变量 B 表示，取值1 为牛肉，2 为谷类，3 为猪肉。分析的目的是要了解公鼠所增体重是否与蛋白质含量高低有关，是否与蛋白质的食物来源有关以及蛋白质含量高低与食物来源间对体重增加有无交互作用。分析时公鼠的食物消耗作为协变量考虑。

设数据存于文件EXAM123，分析步骤如下：

统计分析就是把析因设计的方差分析与协变量配合在一个模型里。

```
C:\SYSTAT>MGLH
 >USE EXAM123
 >CATEGORY A=2,B=3
 >MODE Y=CONSTANT+A+B+X+A*B
 >SAVE MODY/ADJUSTED
 >ESTIMATE
```

结果输出:

```
DEP VAR:     Y     N:  60  MULIPLE R: .685  SQUARED MULIPLE R:  .469

                    ANALYSIS  OF  VARIANCE
```

表 7.13 各小组修正均数

高蛋白组			低蛋白组		
牛肉	谷类	猪肉	牛肉	谷类	猪肉
101.55	100.77	80.21	78.42	96.72	69.55

```
SOURCE  SUM-OF-SQUARES  DF  MEAN-SQUARE   F-RAIO   P
     A        2343.463   1     2343.463   14.450   0.000
     B        1673.305   2      836.653    5.159   0.009
     X        2990.626   1     2990.626   18.441   0.000
    A*
     B         933.812   2      466.906    2.879   0.065
 ERROR        2233.139  21      106.340
```

从结果中看出，当均衡了进食量影响后，蛋白质含量高低间有极显著差别，蛋白质食物来源间也有极显著差别，而两者的交互作用无意义。由于交互作用的不显著，分析主效应就可以单独比较修正均数。按前述方法，在STATS 模块下统计出各小组的修正均数(分组命令用BY A,B)，整理于表 7.13。

总的来看，高蛋白组的体重增加大于低蛋白组。这样只要在高蛋白组中比较食物来源就能得出结论。小组间均数的两两比较可采用DUNCAN多重极差检验。

在STATS模块健入如下命令,求出显著界值:

```
C:\SYSTAT>STATS
>DUNCAN/K=3,MSE=162.177,ALPHA=0.05,DFE=53,N=10
```

结果输出：

```
DUNCAN MULTIPLE RANGE TESTS
ORDERED MEANS DIFFER AT ALPHA=0.050 IF THEY EXCEED FOLLOWING GAPS
GAP  ORDER       DIFFERENCE
       1              11.427
       2              12.018
THIS TEST ASSUMES THE COUNTS PER GROUP ARE EQUAL
```

经比较，在高蛋白组中，牛肉组与谷类组的差别无显著意义，而二者分别与猪肉组比较差别有显著意义。有此我们得到的结论是，当公鼠的摄入量（按热量计）相同时，进食高蛋白牛肉或谷类食物的体重增加较快。

⑷多元协方差分析在比较两组或多组因变量时，如果因变量与多个自变量间存在着一定的线性关系，就应考虑这些自变量的影响。多元协方差分析就是将各个自变量调整到相同的水平，再对因变量的均数作比较。

某地30名初生至三周岁儿童的身高、体重和体表面积，记录于下表。考虑男、女两组的体表面积与身高、体重的关系是否相同，能否合并为一个推算方程。

建立数据文件EXAM124，设性别变量为GROUP，取值1 为男性，2为女性。

首先进行斜率的齐性检验。因为有两个协变量，所以模型中要包含两个交互项。

表 7.14 30 名婴儿身高(X1,cm)体重(X2,kg)及体表面积(Y,cm2)

男			女		
X1	X2	Y	X1	X2	Y
54	3	2446.2	54	3	2117.3
50.5	2.25	1928.4	53	2.25	2200.2
51	2.5	2094.5	51.5	2.5	1906.2
56.5	3.5	2506.7	51	3	1850.3
52	3	2121.0	51	3	1632.5
76	9.5	3845.9	77	7.5	3934.0
80	9	4380.8	77	10	4180.4
74	9.5	4314.2	77	9.5	4246.1
80	9	4078.4	74	9	3358.8
76	8	4134.5	73	7.5	3809.7
96	13.5	5830.2	91	12	5358.4
97	14	6013.6	91	13	5601.7
99	16	6410.6	94	15	6074.9
92	11	5283.3	92	12	5299.4
94	15	6101.6	91	12.5	5291.5

```
C:\SYSTAT>MGLH
>USE EXAM124
>CATEGORY GROUP=2
>MODE Y=CONSTANT+GROUP+X1+X2+GROUP*X1+GROUP*X2
>ESTIMATE
```

结果输出:

```
DEP VAR:    Y    N:  30  MULIPLE R: .993  SQUARED MULIPLE R:  .986

                ANALYSIS  OF  VARIANCE
SOURCE  SUM-OF-SQUARES  DF  MEAN-SQUARE   F-RAIO   P
 GROUP       88053.569   1    88053.569    2.161   0.155
    X1      976354.460   1   976354.460   23.965   0.000
    X2      331960.960   1   331960.960    8.148   0.009
GROUP*
    X1       62455.886   1    62455.886    1.533   0.228
GROUP*
    X2       46356.884   1    46356.884    1.138   0.297
 ERROR      977776.305  24    40740.679
```

结果表明两组回归的斜率无显著差别。接着作二元协方差分析。

```
>MODE Y=CONSTANT+GROUP+X1+X2
```

```
>ESTIMATE
```

结果输出:

```
DEP VAR:    Y    N:  30  MULIPLE R: .992  SQUARED MULIPLE R:  .985
                ANALYSIS  OF  VARIANCE
SOURCE  SUM-OF-SQUARES  DF  MEAN-SQUARE   F-RAIO   P
 GROUP      139769.340   1   139769.340    3.411   0.076
    X1      938153.704   1   938153.704   22.895   0.000
    X2      368954.790   1   368954.790    9.004   0.006
 ERROR     1065399.759  26    40976.914
```

均衡了身高、体重后，男、女间的体表面积仍无显著差别，故认为两者可合并建立推算方程。

㈥FACTOR 模块

FACTOR 模块用于进行主成分分析和因子分析，它的基本命令格式为:

FACTOR <变量1>,<变量2>……

这个命令可输出相关矩阵、特征根、特征根的贡献率及因子负荷。

某单位研究儿童生长发育情况，测量了30 名三岁男童的六项基本体格指标：体重(X1)，身高(X2)，胸围(X3)，上臂围(X4)，三头肌(X5)，肩胛下角(X6)。设数据文件为EXAM161，试作主成分分析。

```
C:\SYSTAT>FACTOR
 >USE EXAM161
 >NUMBER=2
 >FACTOR X1,X2,X3,X4,X5,X6
```

上述一组操作命令中出现了二个FACTOR,第一个FACTOR是从SYSTAT进入FACTOR 模块；第二个FACTOR 后跟六个变量名，表示对该六个原指标作主成分分析。NUMBER=2 表示只取前二个主成分，即P=2。其输出结果如下：

```
MATRIX TO BE FACTORED
      X1       X2       X3       X4       X5       X6
 X1   1.000
 X2   0.609   1.000
 X3   0.716   0.487    1.000
 X4   0.771   0.401    0.451   1.000
 X5   0.312  -0.222    0.256   0.350    1.000
 X6   0.391   0.032    0.273   0.555    0.432    1.000

 TATENT ROOTS(EIGENVALUES)
    1          2          3          4          5          6
 3.086      1.432      0.654      0.427      0.276      0.126
 COMPONENT LOADINGS
                          1              2
```

表 7.15 30 名三岁男童六项体格指标测量结果

X1	X2	X3	X4	X5	X6
13.500	95.000	52.500	15.500	10.000	6.000
14.500	102.000	49.000	16.000	8.000	7.000
13.000	97.600	49.000	15.000	8.000	6.000
15.400	100.000	53.500	15.500	8.000	5.000
16.500	100.000	54.000	17.000	9.000	8.000
13.100	93.500	51.000	15.000	9.000	8.000
14.700	97.500	50.000	15.500	9.000	7.000
14.300	95.100	51.400	15.700	9.000	6.000
13.850	95.600	52.000	14.500	10.000	6.000
11.250	99.000	51.000	13.700	7.000	5.000
15.000	100.000	52.000	15.500	10.000	6.000
15.300	100.000	53.000	16.000	9.000	7.000
11.700	93.400	45.500	14.000	7.000	6.000
12.500	93.300	48.500	15.500	8.000	6.000
14.250	92.800	52.500	16.000	11.000	9.000
14.750	100.000	51.500	15.300	6.000	7.000
14.750	98.500	51.500	16.000	7.000	5.000
13.300	92.600	48.000	15.300	7.000	6.000
13.500	93.500	49.500	16.000	12.000	7.000
12.500	93.000	49.000	15.900	8.000	7.000
13.250	95.800	51.400	14.000	6.000	5.000
14.100	95.400	50.000	15.000	9.000	6.000
13.100	94.900	50.500	14.000	9.000	6.000
12.700	93.300	51.200	13.500	8.000	6.000
17.300	97.600	54.500	17.000	12.000	7.000
14.700	99.500	49.400	15.800	8.000	6.000
11.350	90.400	46.500	14.000	10.000	6.000
12.550	93.000	49.500	14.500	8.000	5.000
13.700	95.300	49.000	14.500	10.000	6.000
14.200	92.700	50.000	15.000	10.000	6.000

```
          X1              0.929             0.167
          X2              0.579             0.720
          X3              0.772             0.209
          X4              0.856            -0.100
          X5              0.441            -0.740
          X6              0.603            -0.533
 VARIANCE EXPLAINED BY COMPONENTS
                            1                 2
                        3.086             1.432
 PERCENT OF TOTAL VARIANCE EXPLAINED
                            1                 2
                       51.426            23.860
```

结果最前面的部分是PEARSON相关系数矩阵，接着是特征根(LATENT ROOTS),它体现了某一主成分对所有指标的总贡献。COMPONENT LOADINGS为主成分负荷量，它反映的是某一主成分对某个指标的贡献（即主成分所包含原来某个指标的信息量）。从结果中看出，第一主成分主要包含X1,X3,X4中的信息量; 而第二主成分主要包含X5,X2,X6的信息量;应该指出X6 在两个主成分中的作用基本是相等的。VARIANCE EXPLAINED BY COMPLAINED表示主成分的总贡献率。此处，二个主成分的贡献率已达到75.3需要可取三个或四个主成分，以便增加总贡献率，一般达到85

应该注意：要指定保留主成分的个数的方法，除了上面的NUMBER命令外，还有指定最小特征根的方式,其命令格式为:

EIGEN=<要保留的最小特征根>

不用这个命令，相当于EIGEN=0。

即保留所有特征根大于0的主成分。

如果既用了EIGRN 命令,又用了NUMBER 命令, 则根据两种标准所得的主成分数，哪个少就按哪个输出。

FACTOR 命令加上PLOT 可选项,则可打出指标Xi(i=1,2,⋯,m) 与主成分Zj (j=1,2,⋯,p) 之间的关系。其命令格式为：

FACTOR/PLOT 或

FACTOR <变量1>, <变量2>,... /PLOT

上面的例子加上PLOT 选择项，则输出散点图。图上每一个字母代表一个指标, 字母顺序与因子负荷的指标顺序相对应。很明显，FACTOR 1 主要反映了A,D, C 三个指标: FACTOR 2 主要反映了B，E 二个指标; 而对于F 指标，FACTOR1 和2 都有较大的反映。

FACTOR 模块还提供了一些可选择的命令:

⑴、因子负荷排序

如果在FACTOR 命令之前,键入SORT 命令, 就能使每个变量的负荷量按高到低顺序输出。这条命令仅仅改变负荷量的输出顺序，对结果的其他方面无任何影响。

⑵、因子旋转

最常用的是最大方差旋转。其命令格式为：

ROTATE=VARIMAX

用了这个命令后,先输出未旋转的结果,然后紧接着输出旋转后的结果。如果加上PLOT选择项,那么因子负荷图也是旋转后的。

经过最大方差旋转后,因子负荷量有所改变,也就是说,各指标在主成分中的作用大小改变了。在实际应用中,应根据指标的专业意义来决定是否需要旋转。理想的情况是,要使得每个主成分能突出反映所观察指标的某一部分的特征。

⑶、将部分结果存入SYSTAT 文件中

在每次键入FACTOR 命令之前,可以用SAVE 命令把算出的因子得分,负荷量或因子得分系数作为SYSTAT 文件存入盘中,以进一步分析。

SAVE <文件名>/SCORES(存因子得分) 或

SAVE <文件名>/LOADINGS(存负荷量) 或

SAVE <文件名>/COEF(存因子得分系数)

每次只能存一种结果。如果文件名后面不跟选择项,程序默认存入因子得分。若存入了得分,则文件中的变量名为FACTOR(i)(i=1,2,⋯,n)。存入的得分都经过了标准化,均值为0。如果计算用的是相关矩阵,则方差为 1 ,如果用的是协方差阵,而且未旋转,则因子得分不进行标准化,方差之和与原始数据算得的相同。

FACTOR 在默认状态下采用相关矩阵进行因子分解。但是,还允许用户选择协方差阵进行因子分解。

命令格式为:

TYPE=COVARIANCE

而不用TYPE 命令就相当于键入了

TYPE=CORRELATION (采用相关矩阵)

现在我们将选择的命令联合起来处理。拟选取三个主成分,按主成分各指标贡献的大小顺序输出,并作最大方差旋转,将因子得分存入SYSTAT 文件。操作如下:

```
C:\SYSTAT>FACTOR
 >USE EXAM161
 >SAVE SCORE
 >NUMBER=3
 >ROTATE=VARIMAX
 >SORT
 >FACTOR X1,X2,X3,X4,X5,X6/PLOT
```

取三个主要成份总的贡献率已达86.2 %。

在结果输出之后随即将因子得分存入名为SCORE 的文件中。这个文件中的变量名为FACTOR(1)和FACTOR(2)⋯。这些因子都经过了标准化,在必要时可利用因子得分作进一步的统计分析。例如可对SCORE 文件直接作聚类分析,把体格发育特征相似的儿童进行分类等等。

此外,取三个主成分,使用PLOT 选择项。所以,有了三个图(FACTOR 1 FACTOR 2,FACTOR 1 FACTOR 3,FACTOR 2 FACTOR 3)。其上的A、B、C、D、E、F 都是在这些平面上的投射,实质上表示了因子量的大小。

FACTOR 模块使用的注意事项:

FACTOR 命令对丢失数据的处理采用成行消除的原则。如果想用成对消除原则,在TYPE语句后面加上PAIRWISE 选择项:

TYPE=COVARIANCE/PAIRWISE 或TYPE=CORRELATION/PAIRWISE

程序在输出负荷时经过判断,如果主成分的负负荷大于正负荷,则取它的反方向, 这就避免了输出的因子带有很多负号。

㈦MDS 模块进行多维尺度变换。SYSTAT MDS 所使用的数据类型可以是协方差阵、相关阵、相似阵或不相似阵，建立时需要在DATA块中进行说明。现第四章的例子程序如下：

```
save CITY
NOTE 'INTERCITY FLYING MILEAGES'
TYPE= SIMILARITY
INPUT  ATLANTA,CHICAGO,DENVER,HOUSTON,LOSANGEL,MIAMI,NEWYORK,
       SANFRAN,SEATTLE,WASHDC
RUN
    0
  587    0
 1212  920    0
  701  940  879    0
 1936 1745  831 1374    0
  604 1188 1726  968 2339    0
  748  713 1631 1420 2451 1092    0
 2139 1858  949 1645  347 2594 2571    0
 2182 1737 1021 1891  959 2734 2408  678    0
  543  597 1494 1220 2300  923  205 2442 2329    0
RUN
SWITCHTO MDS
CHARSET GENERIC
METHOD=KRUSKAL
DIMENSION=2
SCALE
```

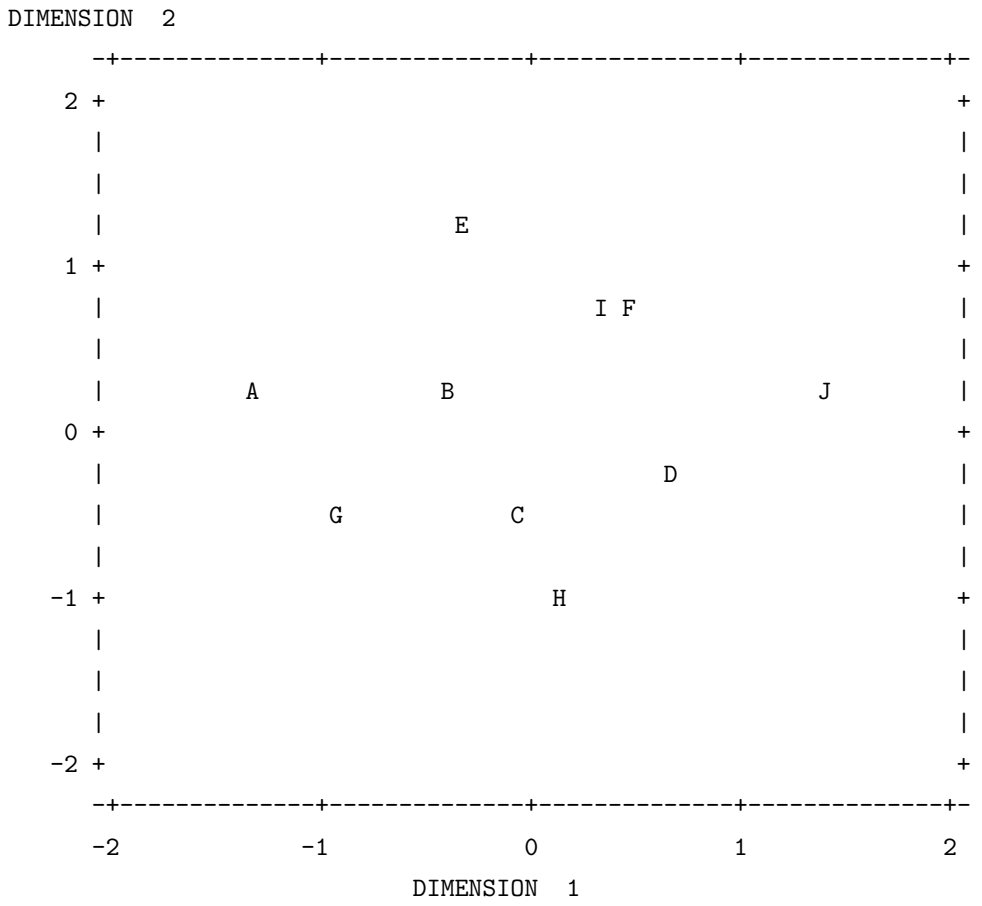
运行结果：

```
                          INTERCITY FLYING MILEAGES

 MONOTONIC MULTIDIMENSIONAL SCALING
 MINIMIZING KRUSKAL STRESS (FORM 1) IN 2 DIMENSIONS
 ITERATION  STRESS          ITERATION  STRESS
 ---------  ------          ---------  ------
     0       .419               8       .262
     1       .316               9       .260
     2       .297              10       .260
     3       .287              11       .260
     4       .280              12       .259
     5       .275              13       .259
     6       .269              14       .259
     7       .264
```

```
    STRESS OF FINAL CONFIGURATION IS:      .25909
                                  SHEPARD DIAGRAM
DISTANCES
      -+---------+---------+---------+---------+---------+---------+-
  3.0 +                                                             +
      |           +                                                 |
      |                                                             |
  2.5 +    +  +                                                     +
      |                                                             |
      |              +                            +                 |
  2.0 +              +            +                  +  +          +
      |            2         +                    +   ++  +        |
      |                 +          + +                              |
  1.5 +                       +            +                       +
      |                   ++              +   +       +   +         |
      |                   +                   +                     |
  1.0 +            + ++   +    2        2 2           +            +
      |                  22                                         |
      |                                                             |
  0.5 +                                                             +
      |                                                       +     |
      |                                                             |
  0.0 +                                                             +
      -+---------+---------+---------+---------+---------+---------+-
       0       500      1000      1500      2000      2500      3000
                              SIMILARITIES

    COORDINATES IN 2 DIMENSIONS
   VARIABLE    PLOT   DIMENSION
   --------    ----   ---------
                        1     2
    ATLANTA      A   -1.32   .24
    CHICAGO      B    -.39   .13
     DENVER      C    -.09  -.52
    HOUSTON      D     .69  -.47
   LOSANGEL      E    -.32  1.11
      MIAMI      F     .47   .51
    NEWYORK      G    -.95  -.54
    SANFRAN      H     .14 -1.23
    SEATTLE      I     .35   .63
     WASHDC      J    1.41   .14
```



§7.3 SYSTAT 4.1 简介

SYSTAT 4.1 模块共有17个，即DATA、EDIT、SSORT、MACPC、MACRO、GRAPH、STATS、TABLES、CORR、MGL、SYGRAPH，可通过执行其相应的.EXE文件而调用。文件SYSTAT.DEF 对各个模块进行了说明。

EDIT 模块允许交互式输入、编辑和转换数据。DATA 模块提供许多工具，用于产生和转换SYSTAT 数据文件。DATA 拥有文件管理工具、转换命令、BASIC 程序语言、Lotus, dBASE和DIF 数据文件转贮。SSORT 工具程序允许对SYSTAT文件快速排序，可以使用10个关键变量。MACPC 工具程序为Macintosh 和MS- DOS/PC- DOS 机间互换二进制SYSTAT 系统文件。MACRO 用于宏定义，即产生完成特定功能的程序。SYGRAPH 包可产生一系列2-维和3-维高分辨率图形，如散点图、矩阵图、平面和轮廓图、直方图、茎叶图、箱式图、Chernov脸谱图、概率分布和分位点图、圆图和条图。CLUSTER 模块提供原始数据或对称数据阵的聚类分析。相关分析模块用于计算对称相关或相似三角阵。FACTOR 进行主成分分析，进行旋转和计算因子得分。GRAPH 模块产生字符统计图形，如散点图、直方图、茎叶图、箱式图、概率分布图和分位点图。MDS 在1-5 维空间对相似或不相似阵进行非度量的多维尺度变换。MGLH 过程用于估计和检验一元或多变元线性模型。NONLIN 模块实现拟牛顿(Quasi-Newton)和单纯形法非线性估计，可对极大似然和相关方法指定损失函数。NPAR 模块进行非参数统计。SERIES 模块用于时序分析。三条基本的命令实现一系列时序分析模型，包括Box-Jenkins ARIMA, Fourier 分析以及线性和非线性滤波。STATS 计算综合统计量。TABLES 模

块用于产主多维列联表并拟合对数线性模型。SYSTAT 4.1支持数学协处理器。SYSTAT 4.1的SURVIVAL 和LOGIT 能够进行Cox 回归和多分类LOGIT 分析。基本的模块需要九张360 KB 软盘，有专门的安装程序INSTALL，其安装大致与第一节中介绍的一样。各模块的调用风格也保持了原来的特色。

SYSTAT 4.1 增强了3.0 的共用命令和分模块命令，这里介绍几个。

1. SWITCHTO 'module'[<file>/ECHO]

   切换到另一个SYSTAT 模块。原来的版本中，不同的分析要用不同的模块，当进行模块切换时，需要重新调用数据，提供这个命令以后，就不必这么做了。

2. DATA 模块命令IMPORT/EXPORT 命令可用于转贮外部文件的数据，用法详见第16章。

   CASELIST 命令在热命令RUN 运行之后进行记录的列表，格式：

   CASELIST [<变量1>, <变量2>, < ... >]

   用例：

   CASELIST (列出整个文件)

   CASELIST MURDER,ROBBERY (列出所有记录的MURDER 和ROBBERY)

   同样可以用REPEAT N 命令列出前面N 个记录。

3. FEDIT 命令启动SYSTAT 文件编辑，可对任何ASCII 文本文件，包括SYSTAT 命令文件和输出文件，在文件编辑时可使用块标记。

   语法：FEDIT <file> | * |#

   用例：

   FEDIT 文件名(编辑新文件或旧文件，永久保留改动)

   FEDIT * (浏览最近一次屏幕输出，可倒数256 行)

   FEDIT > (浏览命令记录文件，编辑和重新提交命令)

   FEDIT # (编辑当前SYSTAT 输出文件)

4. FPATH 命令指定一个自动前缀给SYSTAT 文件，用于和特定的目录或设备联系，有七种文件可以分别加上前缀。

   GET 指GET 命令的ASCII 输入文件(.DAT)

   OUTPUT 指PUT 和OUTPUT 命令的ASCII 输出文件(.DAT)

   SAVE 指SYSTAT 输出文件(.SYS)

   SUBMIT 指SYSTAT 命令文件(.CMD)

   USE 指所有SYSTAT 输入数据文件(.SYS)

   FEDIT 指所有由FEDIT 存取的文件

   TRANSFER 指DATA 模块中所有由IMPORT 或EXPORT 命令存取的文件

   语法：FPATH 'prefix' / GET OUTPUT SAVE SUBMIT USE FEDIT TRANSFER

   用例：

FPATH 'D:' / SAVE (指示所有SYSTAT 输出数据文件到D:)

FPATH '\MYDATA\'/USE GET SAVE (.DAT 和.SYS 文件在\MYDATA)

FPATH 'C:\USR\SYSTAT\' /SUBMIT (.CMD 文件在C: 的目录)

5. CHARSET 命令选择IBM 屏幕/打印机图形字符或通用字符。

语法：CHARSET GRAPHICS | GENERIC

用例：

CHARSET GRAPHICS

CHAR GENERIC (使用通用字符)