第十一章　Genstat

§11.1　Genstat 简介

Genstat 由Rothamsted Experimental Station六十年代开发。用于VAX/VMS 系统、许多大型机、小型机、Unix 工作站和IBM 兼容微机。与众多的统计软件包相比，它为通用的目的而设计。虽然使用它的指令可进行通常的统计分析，但它不仅仅是从有限种预先写好的程序的集合中挑选，而是一个灵活的命令语言，需要产出标准方法没有提供的结果、需要进行新方法的研究，则Genstat 是一个优良的选择。

Genstat 的统计功能分为四个部分，回归分析、实验设计分析、多元和聚类分析以及时间序列分析。回归分析包括一元和多元线性回归、非线性回归、广义线性模型如生物检测(bioasay) 的probit 和logit 分析以及列联表的对数线性模型。方差分析可以包括几乎所有的标准实验设计进行，如完全随机化正交设计、随机区组设计、裂区设计、拉丁方和希腊拉丁方(Graeco-Latin)、重复测量数据分析、平衡不完全区组设计以及平衡混杂因素的其它设计如Youden 方，可对这些设计进行协变量分析并处理缺失值。多元分析包括主成分分析、典型变量分析、因子旋转、主坐标分析以及Procrustes 旋转。使用矩阵、向量的操作可以进行对应分析和典型相关分析等。亦有多种系统聚类方法，可以产生最小支撑树。时间序列分析包括Box-Jenkins 的ARIMA 及其季节模型。序列之间的关系可以用传递函数模型进行研究，进行模型的选择与检查、估计与预测。计算富里叶转换，进行谱分析等。

没有任何计算机程序能完成用户所需要的一切，Genstat 可以使用Fortran 77 来增强自身的功能，进行与其它软件的数据交换等。OWN 语句对于所有版本的Genstat 都有效、PASS 指示使用户不必把它们直接连进来，有时软件并不包括此功能。

§11.2　Genstat 语言

Genstat 功能和语言特色类似GLIM(见第11章)，如下例：

```
VARIATE [NVALUES=10] X
READ X
24.3 25.6 57.3 43.8 45.3
46.5 47.9 97.0 77.5 64.3 :
CALCULATE Xbar=MEAN(X)
PRINT STRUCTURE=Xbar; DECIMALS=2
PRINT STRUCTURE=X; DECIMALS=1
STOP
```

第一句指示X 长度为10，即包含10 个值，接下去三条语句给X 赋值，在数据的末尾用冒号(:) 结束。接下去算得X 的均值，最后的PRINT 语句把数据打印出来。

与Fortran 等高级语言一样，Genstat 的说明(declaration) 用于指示一种数据结构的类型和标识，这种说明可以是显式的或是隐式的，上例中头三句是显式的，而CALCULATE 语句中计算一个均值是隐式说明了一种常数类型数据结构，隐式的定义常称为默认的或缺损的(default definitions)。Genstat 的语句有相同的语法，首先是指令的名称，其次是语句的选项(options)、参数(parameters)。选项放在方括号内，参数放在选项方括号的外面，参数项

之间用分号(;) 隔开。Genstat 的续行符号是反斜杠(\)，注释是用一对双引号括起来的字串。Genstat的字符集是ASCII 码的子集，与多数软件是相同的，需要特别指出的是一些特殊的用法，他们有：

& 表示重复最近一次的命令或过程及其相应的选项设置；* 表示缺失值；$ 表示某些结构中数据的子集，此时多后随一个由方括号括起来的数据表；！引入一个未命名数据结构，有的计算机上使用符号(|)，# 用于引入一个数据结构并嵌入到当前的程序中，在有的计算机上使用英镑符号。

Genstat 的字符组成了六种项(items)，它们是数、字串、标识(identifier) 系统专用字(system words)、缺失值、操作符。它的算术以及关系和逻辑运算符与Fortran 也类似，特别指出的是以下操作符：

== 与.EQ. 等价，/= 与.NE. 等价；字串相等使用.EQS.，不等使用.NES.；标识相等使用.IS.，不等使用.ISNT.；包含使用.IN.，不包含使用.NI.；.EOR. 表示异或(exclusive disjunction)；矩阵相乘使用*+。另外，Genstat 有一些公式操作符：加(+)、点积(.)、交叉积(*)、嵌套积(/)、删除(-)、交叉删除(-*)、嵌套删除(-/)、项目间连接(//)。

列表(链表, lists) 是一些项的集合，有数字列表、字符串列表及标识列表，列表中各项通常用逗号分开，可以省略符(...) 助记，如：

-2,-1.5...0.4 相当于-2, -1.5, -1, -0.5, 0，1 到10 的数表可记为1.. .10，2(A,B,C) 相当于A,A,B,B,C,C，而('a','b')2 相当于'a','b','a','b'。

Genstat 的宏也是一段Genstat 程序，定义之后，可以使用一对替换符号进行调用。设文件ALG.DAT 存放一段程序，用于迭代计算一个数的平方根，可用以下的程序调用。

```
SET [IPRINT=statements,macros]
SCALAR IDENTIFIER=X,Root; VALUE=48;
TEXT [NVALUES=3] Estsqrt
OPEN NAME='ALG.DAT'; CHANNEL=2
READ [CHANNEL=2] STRUCTURE=Estsqrt
##Estsqrt
##Estsqrt
##Estsqrt
PRINT [IPRINT=*] '3 Iterations to calculate sqrt(48) as',Root
STOP
```

文件ALG.DAT 的内容为：

```
'CALCULATE Previous=Root'
'        & Root=(X/Previous+Previous)/2'
'PRINT STRUCTURE=Root,Previous; DECIMALS=4':
```

Genstat 也有交互和批处理两种类型的工作方式，运行结束时使用STOP 命令返回操作系统。

Genstat 的工作控制：包括循环、选择、过程的形成。

Genstat 的程序是标准指令或过程组成的一系列语句。程序可用于检查不同的数据集，同时进行几种分析等。指令JOB/ENDJOB 把Genstat 程序分为功能上相互独立的工作。许多

环境在不同的工作之间保持不变，除非使用了JOB 语句。在一项工作结束后，所有过程和数据结构的值和标识均被删除。

用例：

```
1 JOB 'Example of ENDJOB messages'
2 PRINT 'This job just prints this message.'
3 ENDJOB
```

STOP 指令用于结束一个Genstat 程序。可见Genstat 的JOB/ENDJOB/STOP 与GLIM 软件中的SUBFILE/STOP 相应。

下面的循环用于求平方根：

```
FOR [NTIMES=3]
    CALCULATE Previous=Root
  & Root=(X/Previous + Previous)/
    PRINT root,Previous; DECIMALS=4
ENDFOR
```

其中的NTIMES 是循环的次数，除此以外，还可用COMPILE 指示语句为编译状态执行。语句允许更为复杂的循环指示变量，如：

```
FOR Ind=x1,x2,x3; Dir='descending','ascending'
    SORT [INDEX=Ind; DIRECTION=#Dir] x1,x2,x3
    PRINT x1,x2,x3
ENDFOR
```

相当于下列语句：

```
SORT [INDEX=x1; DIRECTION='descending'] x1,x2,x3
PRINT x1,x2,x3
SORT [INDEX=x2; DIRECTION='ascending'] x1,x2,x3
PRINT x1,x2,x3
SORT [INDEX=x3; DIRECTION='descending'] x1,x2,x3
PRINT x1,x2,x3
```

选择结构中，有块IF 语句，其格式为：IF ... ELSEIF ... ENDIF，用法与Fortran 相似。第二种格式为：CASE ... OR ... ELSE ... ENDCASE 与Foxbase+ 中的DO CASE 语句相类似。这些控制都用EXIT 指令退出，其选项为：NTIMES (控制结构的数目)、CONTROL (控制类型，for, if, case, procedure)、REPEAT (在FOR 中是否转到后续参数控制)。

指令PROCEDURE 用于指示一个Genstat 过程，指令OPTION 与PARAMETER 用于定义过程的选项和参数，两者均可以带名称选项，过程以ENDPROCEDURE 指令结束。可以使用用户自己定义的过程库，如：

```
OPEN 'graphicslib'; CHANNEL=2; FILETYPE=procedurelibrary;
```

存贮的方法是语句STORE，如：

```
STORE [CHANNEL=1;SUBFILE=Jacknife; PROCEDURE=yes] Jacknife
```

指令CATALOG 可用于显示一个库及其了文件的内容。
程序运行过程中，可以执行中断调试，用BREAK 和DEBUG 完成，如：

```
1 PROCEDURE 'polar'
2    PARAMETER 'x','y','r','theta'
3    CALCULATE R=SQRT(X*X+Y*Y)
4    CALCULATE THETA=ARCOS(X/R)
5    CALCULATE THETA=THETA+2*(3.14159-THETA)*(Y<0)
6 ENDPROCEDURE
7 SCALAR Xpos,Ypos; VALUES=3,4;
8 DEBUG
9 POLAR Xpos; Y=Ypos; R=Radius; THETA=Angle
10 ENDBREAK
11 PRINT R
12 ENDBREAK
13 PRINT THETA
14 ENDBREAK
15 CALCULATE Deg=THETA*180/3.14159
16 PRINT Deg
17 ENDDEBUG
18 PRINT Xpos,Ypos,Radius,Angle
```

下面列出Genstat 的几类函数，其中的x、y 可以是常量、变量、因素、表、矩阵、对角阵或对称阵，s 表示常量，f 表示因素，v 表示变量，t 表示表，d 表示哑元。

通用和数学函数

- ABS(x) 绝对值函数。
- ANG(p)或ANGULAR(p) 角度转换，对于$0 < p < 100, x = (180/\pi)\arcsin(\sqrt{p/100})$
- ARCCOS(x), $-1 \le x \le 1$ 反余弦函数，结果为孤度值。
- ARCSIN(x), $-1 \le x \le 1$ 反正弦函数。
- CIRCULATE(x;s) 把x 向左(s¡0)右循环移动s 个位置，s 的默认值为1。
- COS(x) 余弦函数。
- CUMULATE(x) 或CUM(x) 累积和。
- DIFFERENCE(x;s) s 阶差分。
- EXP(x) 自然指数。
- INTEGER(x) 取整函数。
- LOG(x) 自然对数。
- LOG10(x) 常用对数。
- MVREPLACE(x;y) 用y 中相应的值替换x 中的缺失值，当x 与y 均为缺失值给以警告信息。

- NEWLEVELS(f;x) 从因素f 构造一个变量，x 存放相应于各水平的值。
- REVERSE(x) 反转数据。
- ROUND(x) 四舍五入到最近的整数。
- SHIFT(x;s) 把x 的值向左移或右移s 个位置，移动将产生一些缺失值。
- SIN(x) 正弦函数。
- SORT(x;y) 按y 的值的升序对x 排序。
- SQRT(x) 平方根函数。

常量函数

- MAXIMUM(x) 或MAX(x) 最大值函数。
- MEAN(x) 均值函数。
- MEDIAN(x) 或MED(x) 中位数函数。
- MINIMUM(x) 或MIN(x) 最小值函数。
- NCOLUMNS(x) 矩阵的列的数目。
- NLEVELS(f) 因素的水平数。
- NMV(x) x 中未缺失值的数目。
- NOBSERVATIONS(x) 非缺失值的数目。
- NROWS(m) 矩阵的行数。
- NVALUES(x) 包括缺失值在内的x 的长度。
- SUM(x) 或TOTAL (x) 求和。
- VARIANCE(x) 或VAR(x) 方差函数。
  以上函数除了NMV(x) 和NVALUES(x) 以外，均不对缺失值操作。

变量函数

有VMAXIMA(p)、VMEDIANS(x)、VMINIMA(p)、VNMV(p)、VNOBSERVATIONS(p) 、VNVALUES(p)、VSUMS(p)
们与常量函数类似，只是各个函数前缀以V，只有一个指针形式的参量。

矩阵函数

- CORRMAT(x) 从对称阵x 中形成一个相关阵。
- CHOLESKI(x) 进行cholesky 分解。
- DETERMINANT(x) 或DET(x) 或D(x) 求对称阵行列式的值。
- INVERSE(x) 或INV(x) 或I(x) 对称阵求逆。
- LTPRODUCT(x;y) 即x 转置与y 的积。
- PRODUCT(x;y) x 与y 的积。
- QPRODUCT(x;y) y 关于x 的二次型，即x *+ y *+ TRANSPOSE(x)。
- RTPRODUCT(x;y) 即x 与y 转置的积。

- SOLUTION(x;y) 求齐次线性方程组的解。
- TRACE(x) 求矩阵的积。
- TRANSPOSE(x) 或T(x) 求方阵的转置。

表函数

这些函数由表形式的数据组成边缘值，它们有TMAXIMA(t)、TMEDIANS(x) 、TMEANS(t)、TMINIMA(p)、T
或TTOTAL(p)、TVARIANCES(p)。它们与常量函数类似，只是各个函数前缀以T，只有一个指针形式的参量。

哑元函数

UNSET(d) 用于检验哑元是否已定义，返回逻辑值0、1。

元素操作函数

- ELEMENTS(x;e1,e2) 指定一个元素的集合，e1 与e2 是表达式。
- EXPAND(x;s) 从x 的值形成一个0 和1 变量，s 是结果的长度。
- RESTRICTION(x) 相应于x 当前的限制构造一个值为1 的变量。

统计函数

- ANGULAR(%p) 角度转换，$\%p$ 是百分比，结果是$(180/\pi)arcsin(\sqrt{\%p/100})$。
- CED(p;s) 卡方分布变量，给定自由度为s，概率为p，返回相应的卡方值。
- CHISQ(x;s) 自由度为s 的χ^2分布下，小于x 的概率值。
- FED(p;s1;s2) 给定概率值、自由度时F-分布的变量。
- FRATIO(x;s1;s2) 或$FPROBABILITY(x;s1;s2)$ 小于x 的概率。
- LLBINOMIAL(x;n;p) 或LLB(x;n;p) 二项分布$B(n;p)$ 对数似然值。
  $\Sigma xlog(np/x)) + (n-x)log(n(1-p)/(n-x))$，样本量为n，比例为p。
- LLGAMMA(x;m;d) 或LLG(x;m;d) 伽马分布对数似然值。
  $\Sigma d(log(dx/m) - x/m) - LOGGAMMA(d)$，均值为m，index 为d。
- LLNORMAL(x;m;v) 正态分布对数似然值。
  $-0.5\Sigma LOG(v) + (x-m)(x-m)/v$，均值为m，方差为v。
- LLPOISON(x;m) 泊松分布对数似然值。
  $\Sigma xlog(m/x) + x - m, m$ 是样本大小。
- NED(p) 与概率p 相应的正态变量。
- NORMAL(x) 正态分布小于x 的概率。
- URAND(s1;s2) 产生均匀分布伪随机数，s1 是随机数的种子，s2 是长度。

下面是一个Genstat 进行奇异值分解(SVD) 的例子：它把一个矩阵表成左右两个正交阵与一个对角阵的乘积，即 m A n = m U p *+ p S p *+ p V n 。这一分解对于线性方程组的求解、求矩阵的广义逆等都很有意义。

```
1. matrix [rows=6;columns=4]A;
     values=(15,5,9,16,3,20,7,12,22,17,10,11,\
2  13,8,1,23,2,4,6,14,18,21,24,19)
3  SVD[print=left,singular,right]A
```

奇异值分解结果(Singular Value Decomposition)

奇异值矩阵=$diag$(65.30　17.75　14.29　10.82)

左右奇异值向量

$$\begin{pmatrix} .35066 & -.33717 & -.30338 & .26324 \\ .32462 & .30654 & .69925 & -.39495 \\ .45861 & .18847 & -.51086 & -.52922 \\ .37075 & -.71706 & .15091 & -.29662 \\ .20711 & -.27069 & .36641 & .39876 \\ .61629 & .41157 & -.03151 & .49765 \end{pmatrix} \quad \begin{pmatrix} .50011 & -.13783 & -.80932 & -.27549 \\ .50254 & .53368 & .40553 & -.54607 \\ .40479 & .48070 & -.09456 & .77210 \\ .57749 & -.68199 & .41425 & .17258 \end{pmatrix}$$

下面程序利用SVD 的结果得到Moore-Penrose 广义逆或伪逆：

```
4  MATRIX [ROWS=6;COLUMNS=4] Uda
5  & [ROWS=4;COLUMNS=4] Vda
6  DIAGONALMATRIX [ROWS=4] Sda
7  SVD A; LEFT=Uda; SINGULAR=Sda; RIGHT=Vda
8  CALCULATE [ZDZ=zero] Splus=Sda/Sda/Sda
9  & Aplus=Vda *+ Splus *+ TRANSPOSE(Uda)
10 & Aa,Aap=A,Aplus *+ Aplus, A *+ A, Aplus
11 PRINT A; FELDWIDTH=9; DECIMALS=3
12 & Aa; FIELDWIDTH=9; DECIMALS=3
13 & Aplus; FIELDWIDTH=9; DECIMALS=3
14 & Aap; FIELDWIDTH=9; DECIMALS=3
15 CALCULATE Asa, Asapa=A, Aplus *+ Aplus,A
16 PRINT Asa; FIELDWIDTH=9; DECIMALS=3
12 & Aspa; FIELDWIDTH=9; DECIMALS=3
```

其中的CALCULATE 语句是得到奇异值的逆，伪逆为：

$$\begin{pmatrix} 0.016 & -0.029 & 0.044 & 0.007 & -0.027 & -0.009 \\ -0.029 & 0.052 & 0.021 & 0.001 & -0.016 & -0.009 \\ 0.014 & -0.022 & -0.026 & -0.039 & 0.020 & 0.051 \\ 0.011 & 0.005 & -0.026 & 0.030 & 0.029 & -0.003 \end{pmatrix}$$

EDIT 指示提供了一系列行编辑命令。

§11.3　统计图形

Genstat 提供两种基本的绘图功能，一种是为行打印机准备的字符类型的，如直方图、散点图和线图及等值线图，分别用指令HISTOGRAM、GRAPH 和CONTOUR 来实现；另一种是为图形输出设备如图形打印机、绘图仪准备的高分辨图形。

第一类指令说明如下：

HISTOGRAM 的选项有：CHANNEL= 输出文件的通道号，默认值为当前文件；TITLE=总标题；LIMITS=分组组界；NGROUPS=在没有指定LIMITS 时，设定分组数；LABELS=各组标题；SCALE=每个星号所表示的单元数。参数有：DATA=绘直方图的数据；NOBSEVATIONS=存贮每组例数的一维表；GROUPS=存贮变量所指定的分组信息；SYMBOLS=每个直方图条的符号；DESCRIPTION=关键字脚注。

```
TEXT Title
READ [CHANNEL=2;SERIAL=yes;SETNVALUES=yes] Title,Data
HISTOGRAM [TITLE=Title] Data
```

GRAPH 的选项有：CHANNEL=功能同HISTOGRAM；TITLE=总标题；YTITLE=y 轴标题；XTITLE=x 轴标题；YLOWER=y 轴下界；YUPPER=y 轴上界；XLOWER=x 轴下界；XUPPER=x 轴上界；MULTIPLE=每个帧(FRAME) 中的图；JOIN= 连接点的次序(ascending,given)；EQUAL=边界设置(no, scale, lower, upper)；NROWS= 每个帧中的行数；NCOLUMNS=每个帧中的列数；YINTEGER=轴是否一致XINTEGER 是否一致；参数有Y=y 坐标；X=x 坐标；METHOD=每个图的类型(point, line, curve , text)；SYMBOLS=每个单元的标号；DESCRPITION=关键字脚注，如：

```
1 VARIATE [VALUES=-16,-7,9,16,7,-8,-12,-5,0,10,4,-4,-3,3,16] X
2   & [VALUES=0,-14,-12,5,0,14,0,12,0,-10,-9,5,6,-6,-1,5,16] Y
3 GRAPH Y;X
```

CONTOUR 的选项有：CHANNEL=输出文件号；INTERVAL=等值区间；TITLE= 总标题；YTITLE=y 轴标题；XTITLE=x 轴标题；YLOWER=y 轴下界；YUPPER=y 轴上界；XLOWER=x 轴下界；XUPPER=x 轴上界；YINTEGER=y 轴标号的一致；XINTEGER=x 轴标号的一致；LOWERCUTOFF=数组的最小值；UPPERCUTOFF=数组的最大值；参数有GRID=数据指针；DESCRIPTION=关键字脚注，如：

```
1 MATRIX [ROWS=5; COLUMNS=7] Xval,Yval; VALUES=!((1...7)5),!(7(1...5))
2 CALCULATE Z=(Xval-2.5)*(Xval-6)*Xval-10*(Yval-3)(Yval-3)
3 TEXT [VALUES='Z(x;y)=x*(xl-2.5)*(x-6)-10*(y-3)**2'] Top
4 TEXT [VALUES='X Values'] Botttom
5 TEXT [VALUES='Y Values'] Side
6 CONTOUR [TITLE=Top; YTITLE=Side; XTITLE=Bottom] Zval
```

第二类高分辨率图形的有四种即为DHISTOGRAM、DGRAPH、DCONTOUR 及DPIE。

高分辨率图形的绘制，与有关的设置是很有关系的。其设置有：AXES、PEN 、DEVICE、FRAME 分别用于定义图轴、笔、设备及窗口位置。

功　　能	AXIS 图轴	PEN 笔的特性	DEVICE 设备切换	FRAME 视窗位置
选项：				
EQUAL	图轴的等同 (no,scale,lower,upper)			
参数：				
NUMBER			指示设备号	
WINDOWS	窗口数目			窗口号码
YTITLE	y-轴标题			
XTITLE	x-轴标题			
YLOWER	y-轴下界			同左
YUPPER	y-轴上界			同左
XLOWER	x-轴下界			同左
XUPPER	x-轴上界			同左
YINTEGER	y-轴标题一致(yes/no)			
XINTEGER	x-轴标题一致(yes/no)			
YMARKS	y-轴标度			
XMARKS	x-轴标度			
YLABELS	y-轴标号			
XLABELS	x-轴标号			
YORIGIN	y-轴原点			
XORIGIN	x-轴原点			
STYLE	轴的类型 (none,x,y,xy,box,grid)			
NUMBER		笔的数目		
COLOUR		每笔所用的颜色		
LINESTYLE		线型		
METHOD		定点方法(point, line, monotonic,closed,open)		
SYMBOLS		点的记号		
JOIN		连点的次序(ascending,given)		
BRUSH		指示填充的区域号		

例：

```
AXIS WINDOW=3; YLOWER=10; YUPPER=0; XLOWER=0; XUPPER=10; \
     XMARK=Xval; XLABEL=!T(NORTH,EAST,SOUTH,WEST); \
     YTITLE='Y AXIS'; XTITLE='X AXIS'
FACTOR [LEVELS=4; VALUES=1...4] F1
PEN NUMBER=1,2; COLOUR=1; LINESTYLE=1,2; METHOD=line,monotonic; \
    SYMBOLS=F1,3; JOIN=given,ascending;
```

四种指令的选项与参数列如下表，选项为TITLE、WINDOW、KEYWINDOW 、SCREEN和参数PEN 和DESCRIPTION。TITLE 是横标题，WINDOW 与KEYWINDOW 取值范围为0

到8，后者为窗口准备一个指示关键字。CLEAR 取值为'clear' 或'keep'，指示绘图之前当前屏幕是否保存。PEN 指示绘图用的笔，可以通过上述专用的指令设置默认值。DESCRIPTION 指示在key 的位置显示一段文本。

功　　能	DHISTOGRAM 绘直方图	DGRAPH 散点图和线图	DCONTOUR 绘等值线图	DPIE 绘圆图
选项:				
TITLE	标题	同左	同左	同左
WINDOW	图形窗口	同左	同左	同左
KEYWINDOW	关键字窗口号	同左	同左	同左
LIMITS	组界变量			
NGROUPS	分组数			
LABELS	各组的标号			
APPEND	y/n 直方图连接			
SCREEN	c/k 清屏/保留	同左	同左	同左
INTERVAL			等值间隔	
LOWERCUTOFF			数组的下界	
UPPERCUTOFF			数组的上界	
参数:				
DATA	绘图数据			
NOBSERVATIONS	存贮每组数目的表			
GROUPS	从变量定义的因子			
PEN	每个直方图的笔号	笔号	笔号	笔号
DESCRIPTION	关键字附注	同左	同左	说明
Y		纵坐标		
X		横坐标		
YLOWER		纵条的下界		
YUPPER		纵条的上界		
XLOWER		横条的下界		
XUPPER		横条的上界		
GRID			数据指针	
SLICE				扇面大小

图形的输出：OPEN NAME='文件名'; CHANNEL=1; FILETYPE='graphics'; 将产生一个转贮文件(metafile)，最常见的是GHOST、GINO 和GKS。

下面程序产生一系列伪随机数，然后绘直方图。

```
CALCULATE Var[1...3]=URAND(1237,0,0;30)
        & Var[1...3]=10,11,12+NED(Var[1...3])*1,1.2,1.3
"Default histogram with single colour pen"
PEN 1...3; COLOUR=1
DHISTOGRAM [TITLE='Default'] Var[]; PEN=1...3
"Repeat setting the APPEND option and different brush styles"
```

```
PEN 1...3; BRUSH=4,5,9
DHISTOGRAM [TITLE='Appending & new brush style'; APPEND=YES] \
           Var[]; DESCRIPTION='First','Second','Last'
VARIATE [VALUES=6,9,12,15] Limits
AXES WINDOW=1; YUPPER=30
DHISTOGRAM [TITLE='YUPPER set & limits'; LIMITS=Limies] \
           DATA=Var[]; PEN=1,2,3; DESCRITION='First','Second','Last'
STOP
```

下面程序画一个三个扇面的的圆图：

```
PEN 1...3; COLOR=1; BRUSH=1,6,11
DPIE [TITLE='Pie Chart'] 2,4,8; PEN=1...3
```

§11.4 统计分析

§11.4.1 回归分析

MODEL 指令定义响应变量和模型类型（线性、广义线性或非线性模型）。

FIT 指令配合模型。

RDISPLAY 指令显示拟合情况。

RKEEP 指令保存结果。

TERMS 指令指示一个最完全的模型，用于后续的模型分析。

ADD、ROPS 和WITCH 指令增加或减少模型中的项。

TRY 指令显示单一变量改变对模型的影响。

STEP 指令根据均方误差的比值进行模型变量的筛选。

PREDICT 指令用于预测。

下面的程序利用Draper 与Smith (1981) 的材料，对每月用水量与温度、产量、开工日数以及雇员数四种生产指标的关系，进行多元回归分析。

```
UNIT [NVALUES=17]
OPEN 'WATER.DAT'; CHANNEL=2
READ [CHANNEL=2] Temp,Product,Opdays,Employ,Water
MODEL Water
FIT Temp, Product, Opdays, Employ
TERMS Temp, Product, Opdays, Employ
ADD [PRINT=estimates] Product, Employ, Temp
DROP [PRINT=estimates] Temp
SWITCH [PRINT=estimates,accumulated] Temp, Employ
```

```
FIT [PRINT=*] Temp, Product, Employ
STEP [PRINT=estimates,changes; INRATIO=4; OUTRATIO=4] \
     Temp, Product, Opdays, Employ
```

继续使用第三节的例子：相应的Genstat 程序是：

```
Variate [nvalues=8] dose,y,n
read dose,y,n
1.691 6  59 1.724 13 60 1.755 18 62 1.784 28 56
1.811 52 63 1.837 53 59 1.861 61 62 1.884 60 60
model [distribution=binomial]y;nbinomial=n
terms dose
fit [print=m,s,e,f] dose
rkeep vcov=v
print v
```

结果是logit(p)=-60.74+34.29 log(dose) ，还可以计算出半数有效量ED_{50}。

§11.4.2 实验设计

Genstat 主要针对平衡设计，待拟合的模型用BLOCKSTRUCTURE、COVARIATE 和TREATMENTSTRUCTURE 指示，分析则采用ANOVA 指令，利用ADISPLAY 进一步显示结果，分析结果用AKEEP保存。使用GET 指令可以得到当前的模型，并且可以使用SET 指令来改变模型，DECIMALS 参数可以指示结果的小数位数，RESTRICT 指令指示仅对部分单元进行分析，使用RANDOMIZE 指令对处理进行随机分配。

```
1  "3x2 Factorial Design (Snedecor and Cochran 1980)"
2  UNITS [NVALUES=60]
3  FACTOR [LEVELS=!T(beef,cereal,pork]; VALUES=(1...3)20] Source
4  READ Gain
5  TREATMENTSTRUCTURE Source*Amount
6  ANOVA [PRINT=aovtable] Gain
7  ADISPLAY [PRINT=information, covariates, missingvalues]
8  ADISPLAY [PRINT=means]
9  ADISPLAY [PRINT=effects]
10 ADISPLAY [PRINT=%cv]
```

其中TREATMENTSTRUCTURE 指示ANOVA 语句拟合的处理因素，Genstat 采用了点(.)操作方法。此外，还有指令ADISPLAY 显示其它的结果。在几个误差项出现时，可以使用BLOCKSTRUCTURE 定义区组因素。使用COVARIATE 指令可以进行协方差分析，并继续使用ADISPLAY 和AKEEP 指令。

§11.4.3 多元分析

指一些同时分析多个变量的统计方法。关联的数据有两类，第一类是n 个样本p 个变量的数据，第二类可能是一种对称矩阵，包含了所有样本对或变量对的关联信息。象相关表示

的是变量间的关联，这样一类分析称做R-型的，另外一种是对单元间的关联分析，即是Q-型的分析。

基于平方和及乘积的方法有主成分分析(PCP) 和典型变量分析(CVA)，它们者可以使用FACROTATE 指令进行varimx 或quartimax 旋转。

生态学中的关联有序化(ordination) 或多维尺度变换(multidimensional scaling) 在Genstat 中进行的主成分分析、过程库中的对应分析，都属于这一类。Genstat 提供了一个更为一般的方法，即主坐标分析(principal coordinates analysis)，该方法经PCO 指令完成。也能用ADDPOINTS 增加新点。下面是一个主成分分析用例：

```
UNITS [NVALUES=12]
POINTER [VALUES=Height, Length, Width, Weight] Dmat
READ [PRINT=errors] Dmat[]
LRV [PRINT=Dmat; COLUMNS=2] Latent
PCP [PRINT=loadings] Dmat; LRV=Latent
FACROT [PRINT=rotation,communities] Latent[1]
```

§11.4.4 聚类分析

Genstat 提供了系统聚类和非系统聚类两种方法，列表如下：

功　　能	HCLUSTER 系统聚类	CLUSTER 非系统聚类
选项：		
PRINT	输出类型 (dendrogram,amalgamations)	输出类型 (criterion,optimum, units,typical,initial)
METHOD	聚类准则 (singlelink,nearestneighbour, completelink,furthestneighbour, averagelink,mediansort, groupaverage)	
DATA		分析数据矩阵或指针
CRITERION		分类准则 (transfer, swop)
INTERCHANGE		组间允许的移动
START		初始分类
参数：		
SIMILARITY	对称相似矩阵	
GTHRESHOLD	分组界值	
GROUPS	存贮形成的组	
PERMUTATION	聚类图中单元的次序	
AMALGAMATION	存贮聚类过程的链表。	
NGROUPS		要分到的目标类数

用例：

```
POINTER [NVALUES=4] Y
VARIATE [NVALUES=30] Y[]
READ [SERIAL=yes] Y[]
FACTOR [LEVELS=2; NVALUES=30] Optimum[2]
     & [LEVELS=5] Optimum[5]
CLUSTER [PRINT=criterion,optimum,typical,units; DATA=Y; \
        CRITERION=predictive] NGROUPS=5,2; GROUPS=Optimum[5,2]
CLUSTER [PRINT=criterion; DATA=Y; CRITERION=predictive] NGROUPS=6,5
```

§11.4.5 时序分析

Box-Jenkins 时序分析(TSM) 包括识别、估计和检查几个部分，在Genstat 中，提供了在时域和频域上的分析方法，计算一系表征时间序列的样本统计量如自相关、富里叶变换、ARIMA 模型、周期图、传递函数模型。

CORRELATION 构造变量间的相关、变量的自相关以及变量间的互相关。

FOURIER 计算实值或复值序列的富里变换。

ESTIMATE 估计Box-Jenkins 模型。

TDISPLAY 允许对ESTIMATE 进一步显示。

TKEEP 保存ESTIMATE 的结果。

FORECAST 预测时间序列未来的值。

TRANSFERFUNCTION 指示输入、输出序列和传递函数，以进行模型估计。

FILTER 使用时序模型对时序数据进行滤波。

TSUMMARIZE 时序模型特征显示。

Box 与Jenkins (1970) 的数据进行自相关计算。

```
VARIATE [NVALUES=132] Apt
OPEN 'airline data'; CHANNEL=2
READ [CHANNEL=2] Apt
CACULATE Dlapt=DIFFERENCE(LOG(Apt))
CORRELATE [MAXLAG=50; GRAPH=autocorrelations] Dlapt
```

利用此数据建立季节ARIMA 模型，程序是：

```
OPEN 'airline.dat'; CHANNEL=2
UNITS [NVALUES=132]
READ [CHANNEL=2] Apt
VARIATE [VALUES=0,1,0,1,1,12] Ord
      & [VALUES=0,0,0.00143,0.34,0.54] Par
```

```
TSM Airpass; ORDERS=Ord; PARAMETERS=Par
ESTIMATE [MAXCYCLE=0; PRINT=model] Apt; TSM=Airpass
FORECAST [MAXLEAD=12; FORECAST=Fcst12]
```

现假设有六个新的数据，放于文件airline2.dat，可用下面的程序括进来：

```
OPEN 'airline2.dat'; CHANNEL=3
READ [CHANNEL=3; SETNVALUES=yes] New6
FORECAST [PRINT=sfe; ORIGIN=6; MAXLEAD=0; FORECAST=New6]
```

其中的ORIGIN 指示了新数据的数目，设定MAXLEAD=0 避免了计算新的预测值。FORECAST 选项指定包含新数据的变量名，此处为New6。也可以包含新的数据并且产生预测。

```
FORECAST [ORIGIN=6; MAXLEAD=6; FORECAST=New6fcst6]
```

对于抗肺炎球菌血清的例子[5]，Genstat 程序如下：

```
VARIATE [NVALUES=5] DOSE, Y,X
READ DOSE,Y,N
0.0028  35  40
0.0056  21  40
0.0112  9   40
0.0225  6   40
0.0450  1   40
CALCULATE LOGDOSE=LOG(DOSE)
MODEL [DISTRIBUTION=BINOMIAL]Y;NBINOMIAL=N
TERMS LOGDOSE
FIT [PRINT=M,S,E,F] LOGDOSE
RKEEP VCOV=V
PRINT V
```

第k个滞后的样本相关是$r_k = (1 - k/n) \times C_k/C_0$ 其中

$$C_k = \sum_{t=1}^{n-k} \{(y_t - \overline{y}) \times (y_{t+k} - \overline{y})\} n_k$$

，n_k 是求和中所含的项的数目，$\overline{y}$是通常的样本均值。AUTOCORRELATION 允许用户存贮样本自相关。TEST 参数提供自相关为零的假设检验，其定义为：

$$S = n \times \sum_{k=1}^{m} r_k^2$$

当n很大而m相对于n很小时，S服从$\chi^2(m)$分布。

Genstat从自相关计算偏自相关，第k个滞后取值为：

$$corr(y_t, y_{t-k} | y_{t-1}, y_{t-2}, ...y_{t-k+1})$$

并记作$\phi_{k,k}$，可以想象成自回归预测方程中的最后一项：

$$y_t = c + \phi_{k,1} \times y_{t-1} + ... + \phi_{k,k} \times y_{t-k} + e_{k,t}$$

其计算方法：

$$\begin{aligned}\phi_{k,k} &= (r_k - \phi_{k-1,1} \times r_{k-1} - ... - \phi_{k-1,k-1} \times r_1)/\nu_{k-1} \\ \phi_{k,j} &= \phi_{k-1,j} - \phi_{k,k} \times \phi_{k-1,k-j}, j = 1...k-1 \\ \nu_k &= \nu_{k-1}/(1-\phi_{k,k}^2)\end{aligned}$$

由$\nu_0 = 1$开始，$\nu_k = variance(e_{k,t})/variance(y_t)$。

互相关函数公式为：$r_k = (1-k/n) \times C_k/(s_x \times s_y)$ 其中

$$C_k = \sum_{t=1}^{n-k}\{(x_t - \overline{x}) \times (y_{t+k} - \overline{y})\}$$

序列x与y可以不等长。Genstat用Crosscorrelation 计算互相关并提供类似自相关的检验。

指令FOURIER进行富氏变换，如：FOURIER R; TRANSFORM=F将自相关$r_0, ..., r_n$进行变换并将结果存于F，这些值相应于角频率$\pi \times j/m$ 即周期为$2m/j, j = 0, ..., m$

$$f_j = r_0 + \sum_{k=1}^{n}\{2r_k \times cos(\pi \times j \times k)/m\}$$

一般地实序列变换式为：

$$(a_j + ib_j) = \sum_{t=0}^{N-1}\{(x_t + iy_t) \times \exp(i2\pi \times j \times t/N)\}$$

ARIMA 模型为：$\phi(B)\{\nabla^d y_t^\lambda - c\} = \theta(B)a_t$，其中$B$为后移算子，$\nabla$为差分算子。$\phi(B) = 1 - \phi_1 \times B - ... - \phi_p \times B^p$, $\theta(B) = 1 - \theta_1 \times B - ... - \theta_q \times B^q$, c是∇y_t的均值，λ是Box-Cox指数转换参数。季节型ARIMA模型为:

$$\phi(B)\Phi(B^s)\{\nabla^d \nabla_s^D y_t^\lambda - c\} = \theta(B)\Theta(B^s)a_t$$

$\Phi(B^s)$与$\Theta(B^s)$季节自回归与移动平均，∇_s^D是差分阶数D。

传递函数模型为:$y_t = \nu(B)x_t + i(B)a_t$，其中$x_t$与$y_t$ 分别为输入序列和输出序列，其第二项又可以看成噪声序列。