统计分析系统SAS软件

实用教程

惠大丰　姜长鉴　编著

OUTPUT
COMMAND===>

Source	DF	Sum of Squares	Mean Square	F Value	Pr > F
Model	2	196.64157	98.32088	54.022	0.0001
Error	8	14.56007	1.82001		
C Total	10	211.20182			

LOG
COMMAND===>

```
     6  CARDS;
    18  ;
NOTE: The data set WORK.NEW has 11 observations and 3 variables.
NOTE: The DATA statement used 2.00 seconds.
```

PROGRAM EDITOR
COMMAND===>

```
00019   PROG REG CORR;
00020   MODEL y=x1 x2/CLM CLI;
00021   RUN;
```

北京航空航天大学出版社

统计分析系统SAS软件
实用教程

惠大丰　姜长鉴　编著

北京航空航天大学出版社

图书在版编目(CIP)数据

统计分析系统SAS软件实用教程/惠大丰等编著.-北京:北京航空航天大学出版社,1996.5
ISBN 7-81012-639-3

Ⅰ.统… Ⅱ.惠… Ⅲ.生物统计-统计分析-统计程序-分析程序-计算机应用 Ⅳ.①C813-39②Q-332

中国版本图书馆CIP数据核字(96)第01693号

内 容 简 介

SAS是国际著名的统计分析软件,具有技术先进、功能强大和使用方便的特点。本书系统地讲述了微机DOS 6.04版本SAS的基本使用方法,并以农学及生物学试验为例,详细介绍了SAS用于各种试验数据分析的方法。

全书共6章2个附录。内容包括SAS概况和基本使用,各种常用生物统计方法的SAS编程,SAS在多变数分析和试验设计等方面的应用,SAS程序设计和常用过程,以及SAS语句、过程速查和结果输出中统计术语的英汉对照。

本书内容丰富,取材新颖,注重实用,详略得当,通俗易懂,可作为高等农业院校计算机统计课教材或相应专业的研究生教材,或供从事教学、科研和生产实践的科技工作者使用,亦可作为SAS软件统计分析系统部分的培训教材。

●书　　名:**统计分析系统SAS软件实用教程**
●编 著 者:**惠大丰　姜长鉴**
●责任编辑:韦秋虎
●责任校对:张韵秋
●出 版 者:北京航空航天大学出版社(100083)
●地　　址:北京学院路37号(010)62015720(发行科电话)
●印 刷 者:朝阳科普印刷厂
●发　　行:新华书店总店北京发行所
●经　　售:全国各地新华书店
●开　　本:787×1092　1/16
●印　　张:11.75
●字　　数:297千字
●印　　数:5000册
●版　　次:1996年9月第1版
●印　　次:1996年9月第1次印刷
●书　　号:ISBN 7-81012-639-3/F·043
●定　　价:14.60元

代　序

生物统计学是应用数理统计的原理和方法处理生物学中各种数量资料的科学，对于农学、医学、生命科学等领域的定量研究有着重要的作用。但我国在80年代以前，统计处理基本上是依赖于一些诸如机械或电子计算器之类的简单计算工具，效率低下，精度又差，一些复杂分析实际上不能完成。微型电子计算机的出现，为大批量实验数据的高速度分析奠定了基础。然而，将计算机技术和统计方法相结合，科学地编制各种统计方法的计算机程序软件，乃是一项巨大工程。

美国是国际上发展计算机最早的国家之一，他们在70～80年代就已组织力量开发出多种国际通用的统计分析软件。由SAS软件研究所开发的SAS("统计分析系统"的缩写)就是其中最为著名的一种。它几乎囊括了生物学中用过的所有统计方法，又容易学习，使用方便。持有者只需掌握SAS的基本操作，通过编写简短的SAS程序，即可完成数量资料的统计处理，而不需要再学习其它计算机语言。所以，对于从事各学科专业研究而又涉及数量资料的科技人员，SAS将是特别适用的。

姜长鉴和惠大丰同志都曾是我的学生；姜君以后(1987年)又在美国北卡罗来纳州立大学统计系获博士学位，现为扬州大学教授。他们都有较长期的统计实践，是掌握计算机统计学的新一代(这是相对于像我这样不太熟悉微机的老一代而言的)。此次他们结合自己的教学和科研实践，编写了《统计分析系统SAS软件实用教程》一书，将SAS软件的使用方法简单明了地介绍给读者，我认为是为统计学及其应用的普及、发展和提高做了一件很有意义的事。愿作推荐，以广流传。

莫惠栋

于扬州大学农学院

1996年2月

前　言

SAS是美国SAS软件研究所研制的一套大型集成应用软件系统，具有完备的数据存取、数据管理、数据分析和数据展现功能。尤其是创业产品——统计分析系统部分，由于其具有强大的数据分析能力，一直为业界著名软件，在数据处理和统计分析领域，被誉为国际上的标准软件，广泛应用于政府行政管理、科研、教育、生产和金融等不同领域，发挥着重要的作用。

虽然在我国SAS的广泛应用还是近几年的事，但是随着计算机应用的普及和信息事业的不断发展，越来越多的单位采用了SAS软件。尤其在教育、科研领域，SAS已成为进行科学研究和方法教学的得力工具。学习和掌握了SAS的使用，足以满足教学、科研和生产实践的需要。

然而系统地学习和掌握SAS，需要花费一定的时间和精力。而对大多数科技工作者而言，需要掌握的仅是如何利用SAS来解决自己的实际问题，他们迫切需要一本简明的SAS实用教程。同时，对于农学和生物学类的在校大学生、研究生，亦缺乏一本合适的教材。为此，我约请姜长鉴老师合编了这本书。尽管为了突出实用，本书不可能（也无必要）详细描述SAS的诸多性能，但读者仍能从中体会出SAS的卓越功能和使用方便的设计风格。

本书从生物领域中各种实际应用出发，针对科研和生产实践中需要解决的问题，讲述SAS编程方法。在应用举例中，主要采用莫惠栋教授所著《农业试验统计》（第二版）中的例题作为编程实例。这些例题大多为实际试验结果，为科技工作者喜闻乐见，因而便于临摹套用。具体分析原理和例题的详细解释可参阅《农业试验统计》[1]。书中述及的原理和方法对于自然科学的相关领域和社会科学同样适用。

本书系作者根据近几年SAS应用和教学、科研实践，从实用出发，参考国内、外资料编著而成。在撰写过程中，得到莫惠栋教授的指导和帮助。扬州大学农学院农学系和教务处给予了大力支持。SAS软件研究所上海办事处对本书的出版亦给予极大的关注和支持，在此一并表示衷心感谢。

限于作者水平，书中差错与不足在所难免，恳请各位读者和同行专家批评指正。

惠大丰
于扬州大学农学院
1996年1月

目　录

第一章　统计分析系统SAS概述

SAS是美国使用最为广泛的三大著名统计分析软件(SAS,SPSS和SYSTAT)之一,是目前国际上最为流行的一种大型统计分析系统,被誉为统计分析的标准软件。

SAS为"Statistical Analysis System"的缩写,意为统计分析系统。它于1966年开始研制,1976年由美国SAS软件研究所实现商品化。1985年推出SAS PC微机版本,1987年推出DOS下的SAS 6.03版,之后又推出6.04版。以后的版本均可在WINDOWS下运行,目前最高版本为SAS 6.11版。SAS集数据存取、管理、分析和展现于一体,为不同的应用领域提供了卓越的数据处理功能。它独特的"多硬件厂商结构"(MVA)支持多种硬件平台,在大、中、小与微型计算机和多种操作系统(如UNIX,MVS WINDOWS和DOS等)下皆可运行。SAS采用模块式设计,用户可根据需要选择不同的模块组合。它适用于具有不同水平与经验的用户,初学者可以较快掌握其基本操作,熟练者可用于完成各种复杂的数据处理。

目前SAS已在全球100多个国家和地区拥有29 000多个客户群,直接用户超过300万人。在我国,国家信息中心、国家统计局、卫生部、中国科学院、宝山钢铁集团公司和中国银行江苏省分行等都是SAS系统的大用户。SAS已被广泛应用于政府行政管理、科研、教育、生产和金融等不同领域,并且发挥着愈来愈重要的作用。

1.1　SAS的设计思想、功能和特点

1.1.1　SAS的设计思想

SAS的设计思想是为统计学家和科学工作者提供这样的一个工具,利用它可以完成包括从简单的描述性统计到复杂的多变数分析的各种运算,从而使人们从繁重的计算任务中解脱出来,有更多的时间和精力用于分析和解释计算的结果,而不必为如何获得这些结果花费过多的时间和精力。

1.1.2　SAS的功能

SAS是数据管理和分析软件包,能够完成各种统计分析、矩阵运算和绘图等。

SAS的各项功能由功能模块完成。其中BASE模块为必需模块,其它模块可任选。供选择的模块包括统计(STAT)、矩阵运算(IML)、绘图(GRAPH)和全屏幕操作(FSP)等20余个。

基础模块(BASE)具有以下功能:进行数据存储、调入、修改、追加、拷贝和文件处理;编写报告、打印图表;进行数据排序、分类等操作;完成一些基本统计数计算(如平均数和相关系数);与一些软件包(dBASE,LOTUS等)及大型机进行数据交换和通讯。BASE模块为SAS系统的核心模块。

统计模块(STAT)提供一些高度可靠、完整的统计分析过程。主要有方差分析(包括一元、多元的单因素及多因素试验设计的方差分析),线性相关和回归分析(包括一元与多元线性相关和回归分析,多项式回归分析),非线性回归分析,多变数分析(包括聚类分析、主成份分析、因子分析、典范相关分析)以及非参数测验等,共计26个过程。每个过程还提供多种不同算法

和选项，从而使SAS系统成为一个全面、细致、科学的统计分析方法集。STAT模块为SAS系统的核心和精华。

矩阵运算模块(IML)是一种交互式矩阵语言。可直接进行矩阵运算(加法、乘法、求逆、计算特征值和特征向量等)，适用于高级统计、工程运算和数学分析。

绘图模块(GRAPH)能在微机的绘图设备上绘制图形。可制作三维图形、地图和幻灯等。

全屏幕操作模块(FSP)为一交互式全屏幕软件。利用它可以建立、修改和浏览SAS数据集中的观察值，定义用户屏幕等。

1.1.3 SAS的特点

SAS把数据存取、管理、分析和展现有机地融为一体。主要特点如下。

1. 功能强大，统计方法齐、全、新

SAS提供了从基本统计数的计算到各种试验设计的方差分析、相关回归分析以及多变数分析的多种统计分析过程，几乎囊括了所有最新分析方法，其分析技术先进、可靠。分析方法的实现通过过程调用完成。许多过程同时提供了多种算法和选项。例如方差分析中的多重比较，提供了包括LSD,DUNCAN,TUKEY测验在内的10余种方法；回归分析提供了9种自变量选择的方法(如STEPWISE,BACKWARD,FORWARD,RSQUARE等)。回归模型中可以选择是否包括截距，还可以事先指定一些必须包括在模型中的自变量子组(SUBSET)等。对于中间计算结果，可以全部输出、不输出或选择输出，也可存储到文件中供后续分析过程调用。

2. 使用简便，操作灵活

SAS以一个通用的数据(DATA)步产生数据集，尔后以不同的过程调用完成各种数据分析。其编程语句简洁、短小，通常只需很少的几条语句即可完成一些复杂的运算，得到满意的结果。结果输出以简明的英文给出提示，统计术语规范易懂，具有初步英语和统计基础即可。使用者只要告诉SAS“做什么”，而不必告诉其“怎样做”。同时SAS的设计，使得任何SAS能够“猜”出的东西用户都不必告诉它(即无需设定)，并且能自动修正一些小的错误(例如将DATA语句的DATA拼写成DATE,SAS将假设为DATA继续运行，仅在LOG中给出注释说明)。对运行时的错误它能尽可能地给出错误原因及改正方法。因而SAS将统计的科学、严谨和准确与便于使用有机地结合起来，极大地方便了使用者。

3. 提供联机帮助功能。

使用过程中按下功能键F1，可随时获得帮助信息，得到简明的操作指导。

1.2 SAS的安装、启动和退出

1.2.1 SAS的安装

1. 硬件要求

适用于PC286,386,486等微机和大、中、小各种型号的计算机，支持单色和彩色监视器，能在多种操作系统下运行。DOS环境下的SAS称为微机SAS。本书主要介绍DOS下的SAS 6.03和6.04版。SAS要求至少512K内存(最好640K)，可使用扩充内存。对硬盘的存储空间要求为：基础模块(BASE)——5.3MB；统计模块(STAT)——4.1MB；矩阵运算模块(IML)——0.7MB；绘图模块(GRAPH)——7.2MB；全屏幕操作模块(FSP)——0.8MB。以上共计约18.1MB。

2. 软件要求

DOS 3.3 以上版本。

3. 安装方法

对于原装 SAS 软件，将标有 BASE INSTALLATION(安装盘)的磁盘插入 A 驱动器中，在 DOS 提示符 A:\>下输入命令 SASLOAD↙，按提示选择和插入磁盘即可。

对于备份方式存储的 SAS 软件，可以用 RESTORE 命令重新装入硬盘。即在 DOS 提示符 C:\>下输入命令 RESTORE A: C:/S↙，按提示依次插入 BACKUP 01,02,…，直至所有的备份盘都装入硬盘。在 MS DOS 6.0 以后版本下备份的软件，可利用 MSBACKUP 进行 RESTORE，具体操作可参阅相应的 DOS 使用手册。

4. 安装注意事项

1) 配置文件的修改：SAS 系统要求修改系统配置文件 CONFIG.SYS，将同时打开文件数设置为 FILES=30。因此，如果原 CONFIG.SYS 中的文件数未设置或小于 30，应进行修改，并重新启动机器使之生效。另外，SAS 系统亦有一配置文件，文件名为 CONFIG.SAS。一般情况下不需修改，但有时也要改变设定。例如，当安装的 SAS 系统目录不是 c:\sas 而是 f:\sas 时，应将 CONFIG.SAS 中的"-SET SASROOT c:\sas"改为"-SET SASROOT f:\sas"。详细情况可参阅 SAS 目录下的 CONFIG.HLP 文件。配置文件内容示例如下：

系统配置文件 CONFIG.SYS。

```
DOS 6.0 版本                          DOS 3.3 版本
device=c:\dos\himem.sys               files=30
devicehigh=c:\dos\emm386.exe          buffers=20
dos=high
files=50
buffers=40
```

SAS 配置文件 CONFIG.SAS。

```
-SET SASROOT c:\sas          /* 定义 SAS 系统所在目录，这里为 C 盘 */
                             /* SAS 子目录，必须为第一条语句    */
-SASHELP   ! sasroot\sashelp /*  help file 目录              */
-SASUSER   ! sasuser         /*  user profile 目录          */
-WORK      ! saswork         /*  work 目录                  */
-MSG       ! sasroot\sasmsg  /*  message file 目录          */
-DMS                         /* 进入显示管理系统状态        */
-EMS ALL                     /* 设定使用 EMS 的大小         */
-PATH   ! sasroot\sasexe\core /* SAS 可执行文件查找顺序     */
-PATH   ! sasroot\sasexe\base
-PATH   ! sasroot\sasexe\stat
-PATH   ! sasroot\sasexe\iml
-PATH   ! sasroot\sasexe\af
-PATH   ! sasroot\sasexe\fsp
-PATH   ! sasroot\sasexe\graph
……
```

2) 批处理文件的设置：一般应在自动批处理文件 AUTOEXEC.BAT 的 PATH 语句中，加入 SAS 系统所在的子目录(通常为 C:\SAS)。这样开机启动后，在任一目录下键入 SAS，皆可启动 SAS 系统。另外，SAS 启动时将自动执行 SAS 程序文件 AUTOEXEC.SAS，所以在此文件中可加入一些常用的设置命令。例如：OPTIONS NODATE PAGESIZE=60; 定义结果

输出不打印系统日期、每页输出 60 行。LIBNAME data "e:\sasdata" prog "e:\sasprog"； 定义库关联目录 data 为 e:\data，prog 为 e:\prog，这样 sas 启动时将自动完成这些设定。自启动文件内容示例如下：

自动批处理文件 AUTOEXEC. BAT。

```
ECHO OFF
PROMPT $P$G
PATH C:\;C:\SAS
```

SAS 自启动文件 AUTOEXEC. SAS。

```
FILENAME rlink "! sasroot\saslink\tso. scr";
OPTIONS NODATE PS=60 LS=78;
LIBNAME data "e:\sasdata" prog "e:\sasprog";
FILENAME dtmp "e:\sasdata\tmp. ssd" ptmp "e:\sasprog\tmp. sas";
```

3) SAS 的注册：SAS 系统必须注册。当机器系统时间不在软件的注册有效期内时，SAS 将不能正常启动。

1.2.2 SAS 的启动和退出

1. SAS 的启动

安装完成后，用 CD\SAS 命令进入 SAS 子目录，出现提示符 C:\SAS>，输入 SAS↙，即可启动系统，稍候，出现如图 1-1 所示屏幕。将整个屏幕分成 3 个窗口，即结果输出窗(OUTPUT)、运行记载窗(LOG)和程序编辑窗(PROGRAM EDITOR，简记为 PGM)。光标停留在程序编辑窗(PGM)内。

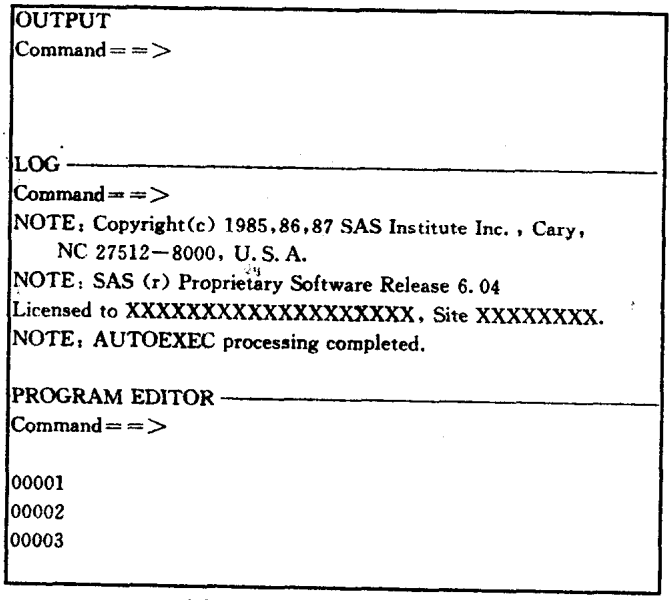

图 1-1 SAS 启动后的屏幕

SAS 启动时也可以带有选项。例 1：SAS-EMS ALL，将利用扩充内存(EMS)调入 SAS 的总控系统和大多数的 SAS 过程，这样可以将常规内存节省给程序、数据使用。当微机的扩充内存较大时，SAS 将使用多达 2M 的扩充内存，这样可以节省近 187K 的常规内存。例 2：SAS-NODATE-CENTER，启动后，结果输出时将不输出系统日期并居中显示。例 3：SAS-NODMS，以行交互方式启动，此时用户每输入一条语句，SAS 立即予以响应执行。此选项建议

不要使用。例 4:SAS myprog,启动后直接执行 SAS 程序 myprog. sas,结果存 myprog. log,运行过程存 myprog. lst。常用的启动方式为直接键入 SAS 启动。

2. SAS 程序的输入和运行

SAS 程序在程序编辑窗内行号的后面空一格开始输入。输入完成后,按功能键 F10 或在命令行输入 SUBMIT 命令运行。运行过程显示在运行记载窗(LOG)内,结果输出到结果输出窗(OUTPUT)内。

3. SAS 的退出

移光标至命令行 COMMAND= =>处,空一格输入 BYE 或 ENDSAS↙,即退出 SAS 系统,返回 DOS 状态。键入 X↙,可临时退至 DOS 环境,执行完 DOS 命令后,输入 EXIT 返回 SAS 系统。若临时执行一 DOS 命令,亦可输入 X "DOS 命令"; 这里引号可为双引号或单引号,并且后面引号可省略。例如:X "DIR *.SAS↙,将显示当前目录下的所有扩展名为.SAS 的文件。显示结束,按键返回 SAS。

1.3 SAS 的程序结构

先看一个简单的 SAS 程序,如图 1-2。

```
PROGRAM EDITOR
COMMAND= =>

00001  DATA new;①
00002  INPUT x y;
00003  CARDS;
00004  12  134
00005  11  156
00006  21  200
00007  32  345
00008  56  765
00009  20  198
00010  78  988
00011  46  480
00012  ;
00013  PROC MEANS;
00014  VAR x y;
00015  RUN;
00016
```

图 1-2 SAS 程序

这是一个完整的 SAS 程序。熟悉计算机语言的用户很容易看出,该段程序的功能是读入变量 x 和 y,并计算其平均数。

这个程序包括两个部分。第一部分准备数据,行号 1～12,称为数据步(DATA Step)。第二部分分析数据,行号 13～15,称为过程步(PROC Step)。

典型的 SAS 程序都包括这两个部分。一般情况下一个程序至少包括一个数据步和一个过

① SAS 对程序中的字母大小写不加区分。本书为区别 SAS 关键词(大写字母编排,例如 DATA),将用户定义的数据集名、变量名等用小写字母编排,例如 new,x,y。但在 SAS 的结果输出中,变量名、数据集名等皆用大写输出。

程步。

按功能键 F10(SUBMIT)运行该程序,PGM 窗口中的所有语句被清除,PGM 窗口下的状态行右下角有一"R"在闪烁,表明程序正在运行。运行结果输出到 OUTPUT 窗。运行过程记载在 LOG 窗中。按 F3 进入 LOG 窗,检查有无错误发生(错误一般用红色显示)。按 F4 进入 OUTPUT 窗,移动光标即可查看分析结果。其结果输出如下:

N Obs	Variable	N	Minimum	Maximum	Mean	Std Dev
8	X	8	11.000	78.000	34.500	23.737
	Y	8	134.000	988.000	408.250	315.989

结果表明,变量 x 和 y 的平均数(Mean)为 34.500 和 408.25,标准差(Std Dev)分别为 23.74 和 315.99。按 F3 进入 LOG 窗,显示如下:

```
NOTE: Copyright(c) 1985,86,87 SAS Institute Inc., Cary, NC 27512-8000, U.S.A.
NOTE: SAS (r) Proprietary Software Release 6.04
      Licensed to XXXXXXXXXXXXXXXXXXXXX, Site XXXXXXXX.
NOTE: AUTOEXEC processing completed.
   1  DATA new;
   2  INPUT x y;
   3  CARDS;
  12  ;
NOTE: The data set WORK.NEW has 8 observations and 2 variables.
NOTE: The DATA statement used 5.00 seconds.
  13  PROC MEANS;
  14  VAR x y;
  15  RUN;
NOTE: The PROCEDURE MEANS used 4.00 seconds.
```

最上面的信息为版权、注册号以及 SAS 自启动文件 AUTOEXEC.SAS 运行完成的注释。这些信息系首次启动 SAS 时给出,以后运行程序时将不再显示。接下来显示的为上述程序的运行过程,其行号为 PGM 窗中程序前面的行号,如果运行出错可以由此确定错误所在行。数据部分没有显示。1～12 为读数部分。NOTE 注释给出 DATA 语句建立了一个临时数据集(data set)WORK.NEW,内有 8 条观察值和 2 个变量,建立过程用了 5 秒钟。随后 13～15 行执行了 MEANS 过程,用时 4 秒。

从上面可以看出,SAS 功能的实现是通过由 SAS 语句组成的程序来完成的。每一条语句指定 SAS 执行特定的操作。

SAS 语句的第一个词一般为 SAS 关键词(或保留字),指定执行某一操作。语句的其它部分说明如何执行这一操作,描述操作所需信息。语句以语句结束符分号";"结束。例如 DATA new; DATA 语句表明建立一数据集,new 为数据集名,即建立数据集 new。PROC MEANS; PROC 表明运行一个 SAS 过程,MEANS 为过程名,即调用 MEANS 过程。

SAS 程序的输入格式相当自由。语句可从一行的任一列开始输入。一行中可以输入任意多条语句。一条语句允许占用多行。语句中间允许空行。SAS 语句对字母大小写不加区分,输入语句时可以用大写字母、小写字母或混用。但要注意每条语句结束必须有结束符";",语句中的各项之间至少应有一个空格分隔。

为易于阅读和理解,建议开始每行写一条语句。DATA,PROC,RUN 等语句从第 1 列开

始书写，其余可缩格写。

有关程序编辑的详细方法参见 1.7。

1.4 SAS 的数据(DATA)步

数据步用于为 SAS 的分析过程准备数据。一般情况下 SAS 将数据读入、存储在一临时数据集中。只有将数据组织到数据集中，SAS 过程才能进行分析。

临时数据集存储于临时工作目录中，在 SAS 运行期间该数据集一直有效，但退出 SAS 后该数据集被自动删除。相应的永久数据集是指建立在磁盘中的数据集，类似磁盘文件，一旦建立就一直有效。

SAS 的数据集类似于 DOS 系统的文件。数据集的命名有一定的规则，要求数据集名必须以字母或下划线开头，后面可以是字母或数字，但不允许有一些特殊符号(如“,”,“--”等)，可以使用下划线，总长度不超过 8 个字符。例如 mydata，addr95。数据集的扩展名为.SSD。SAS 数据集以 SAS 特有的方法存储数据，数据集内不仅包括数据，还包括变量名称、格式、长度等属性信息。

数据集全名为两段名(two-level)，即数据库目录关联名加数据集名，中间用“.”连接，形式为 libref.dataset。这里 libref 为库目录关联名，用于标注数据集的存储位置，代表磁盘的某一目录，由 LIBNAME 语句定义。例如：LIBNAME paper "e:\data"; paper 将代表 e:\data 目录。不使用 libref 以单名给出的数据集为临时数据集，例如 new，此时缺省的 libref 由系统自动赋予库目录关联 WORK，一般指向 SASWORK 子目录(由 CONFIG.SAS 文件设定)，这里数据集全名为 WORK.NEW，存储在 WORK 关联目录下。永久数据集通过两段名给出，例如：paper.mydata，代表 e:\data 目录中的永久数据集 mydata.ssd。

数据集的建立由 DATA 步来完成。DATA 步中有很多语句，可以完成各种复杂的数据集构建。其中可以使用函数、表达式、数组以及各种程序结构，其功能优于 FORTRAN 等程序设计语言(参阅第 5 章)。DATA 步常用的语句有 DATA，INFILE，INPUT 和 CARDS 等语句。DATA 语句指明开始创建 SAS 数据集，INFILE 语句指明外部数据文件，INPUT 语句描述输入数据，CARDS 表明数据行的开始。数据行的结束以语句结束符“;”判定。带有“;”号语句的上一物理行为数据末行，因而当数据输入完毕后，应加一“;”行(空语句)，以示数据行结束。当数据行数据中出现“;”时，用 CARDS4 代替 CARDS，用“;;;;”表示数据行结束。

数据集的建立依数据来源或读入方式一般分以下 4 种方法。

1. *直接输入法*

该方法直接在程序行中输入数据。例如：

```
DATA new;
INPUT x y name $;
CARDS;
12  134 A
11  156 B
21  200 C
32  345 D
56  765 E
20  198 F
```

```
78  988 G
46  480 H
;
```

DATA 语句指明建立一数据集。这里为临时数据集 new。它表明 DATA 步的开始，和其后的一系列语句一起构成一 DATA 步。以上为一完整的 DATA 步。

这里 DATA 后为单名 new，表明为临时数据集。DATA 后的数据集名也可省略，此时系统依次赋名为 DATA1，DATA2 等。

INPUT 语句描述 SAS 数据集中的每条记录(或称观测值)。本例表明每一条记录由 3 个变量构成，即变量 x，y 和 name。变量名由用户给出，命名要求与数据集的命名规则类似，即以字母或下划线开头，后面可带数字或字母，长度不超过 8 个字符，但不可含有一些特殊字符(如空格，“.”，“;”，“+”，“－”运算符等)。SAS 系统常用的变量类型有数值型、字符型和日期型。字符型变量在变量名后加一“$”表示，如变量 name。INPUT 语句的书写很重要，其书写方法亦有多种，本例为最简形式，详细使用方法可参阅 5.4。

CARDS 语句表明数据的开始，其后紧随数据行。

数据行　这部分为输入的数据。输入方法应与 INPUT 语句的描述相对应。本例每一条记录包括 x，y 和 name 三个变量的一次取值，相当于数据行中的一行，如 x＝12，y＝134，name＝“A”即为一条记录。数值数据缺省时一般用圆点“.”表示。数据行的后面应加一行空语句“;”结束。当数据行的下一物理行为一条带有“;”的完整语句时，空语句可省略。

这里空语句";"表明 DATA 步的结束，开始运行数据步。本例 DATA 步运行结束，将建立数据集 new，其中存储含有 3 个变量的 8 条记录。

本方法主要用于数据量不太大的情况。若数据量较大，或已经录入磁盘文件中，可使用下面的一些方法。

2. 读取外部数据文件法

该法借助于已经建立在磁盘上的外部数据文件，从中读取数据。假设有一数据文件 data.txt，其内容为：

```
12  134
11  156
21  200
32  345
56  765
20  198
78  988
46  480
```

建立数据集的程序如下：

```
DATA new;
INFILE "data.txt";
INPUT x y;
RUN;
```

DATA 和 INPUT 语句的意义同 1.。

INFILE 用于调用外部磁盘数据文件，这里为 data.txt。该语句必须在 DATA 和 INPUT 语句之间。DATA 步运行结束，将建立一临时数据集，内有 2 个变量、8 条记录。

RUN 语句开始运行上面的数据步。

如果运行下面的程序，将在磁盘 e:\data 下建立一永久数据集 mydata.ssd，其内容类似上述临时数据集 new。

```
LIBNAME paper "e:\data";
DATA paper.mydata;
INFILE "data.txt";
INPUT x y;
RUN;
```

3. 读取 dBASE 数据库文件法

SAS 系统可以从其它多种数据库管理系统(如 dBASE、FoxBASE 等)的数据库文件中直接读取数据。假设磁盘中存有一 FoxBASE 数据库文件 e:\data\name.dbf，那么：

```
FILENAME dbfile "e:\data\name.dbf";
PROC DBF DB3=dbfile OUT=new;
RUN;
```

将建立一个临时数据集 new，其变量等同于 FoxBASE 的字段，观察值(记录)等同于 FoxBASE 的记录。由于 SAS 系统目前尚不能将汉字作为变量名，因而当从数据库文件中读取数据时，数据库文件不能使用汉字作为字段名。

4. 利用永久数据集法

SAS 可以从已建立的永久数据集中读入数据，进行加工处理，并允许选择记录和变量。假设磁盘中存有永久数据集 e:\data\name.ssd，那么：

```
LIBNAME paper "e:\data";
DATA new;
SET paper.name;
RUN;
```

将建立一临时数据集 new，其内容类似 e:\data\name.ssd。

在磁盘上已建立永久数据集的情况下，如果不进行变量或观察值的选择，可以省略临时数据集的建立。因为大多数 SAS 过程可直接指明分析所用的 SAS 数据集。例如：

```
LIBNAME paper "e:\data";
PROC MEANS DATA=paper.mydata;
RUN;
```

将对 e:\data\mydata.ssd 中的数值变量计算平均数等。

以上为数据集建立的基本方法。一些特殊数据(如相关系数)的读取，参阅第 5 章 SAS 程序设计。

1.5 SAS 的过程(PROC)步

1.5.1 过程步的基本知识

过程步是 SAS 程序的另一个重要组成部分。它用于对已建立的数据集中的数据进行分析并给出处理结果。SAS 系统的数据分析通常由过程调用完成。一个过程步即是完成某些操作的一个程序模块。SAS 系统的过程步与数据步独立。通过数据步建立的 SAS 数据集可以被任一个过程步调用。因而调用不同过程，即可应用不同方法对数据进行处理。

前例中：

```
PROC MEANS; VAR x y; RUN;
```

即为一次完整的过程调用。它完成对变量 x 和 y 计算算术平均数和标准差等，并将结果输出到

OUTPUT 窗口中。

SAS 程序可以是 DATA 步和 PROC 步的任意组合。在比较简单的情况下，一个 SAS 程序是 DATA 步后跟一 PROC 步。实际上数据步和过程步可以以任意次序出现。例如 DATA+PROC+PROC 或 DATA+PROC+DATA+PROC 等。只有一个 DATA 步或在已经建立永久数据集的情况下，只有一个 PROC 步的 SAS 程序也是可能的。

一般情况下，SAS 过程步要求输入要处理的数据集名、变数名以及是否分组等。在不特别指明的情况下，SAS 系统将采用缺省设置对新近(最后一次)建立的数据集中的所有变量不分组进行处理。正是由于这种设计，使得 SAS 使用相当简便。很多情况下仅需指明过程名，再加一 RUN 语句即可。例如：PROC PRINT；RUN； 即完成打印过程 PRINT 的过程调用，输出整洁、美观的打印结果。

SAS 过程步中的通用语句有 PROC，VAR，BY，OUTPUT 等，其它语句依过程不同而异。多数情况下，PROC 及过程语句后可带选项。当使用多个选项时，选项之间亦应空格。例如：PROC MEANS N MEAN STD；这里 N MEAN STD 为 PROC MEANS 的三个选项。过程中语句的顺序大多可以任意，但有些过程对某些语句的出现有一定的次序要求，使用时应加以注意。

1.5.2 两个实用过程简介

1. PRINT 过程

读入 SAS 数据集中的数据，将变量排成易读的形式输出。例如：

```
PROC PRINT DATA=new;
VAR x1 x2 x5 x6 y;
RUN;
```

读入 SAS 数据集 new 中的数据，对其中的变量 x1，x2，x5，x6 和 y 排成易读的形式输出。

2. SORT 过程

用于对数据集中的数据按指定的变量进行排序，排序结果存入新数据集或存回原数据集中。例如：

```
DATA score;
INPUT class $ name $ math stat;
CARDS;
nx93  Zhang1 80 92
nx92  Li1    78 82
xm92  Zhao1  80 89
nx93  Zhang2 83 87
... ...
nx92  Wang3  70 96
xm93  Zhang8 92 65
;
PROC PRINT; RUN;
PROC SORT;
BY class;
RUN;
PROC PRINT; RUN;
```

这里 INPUT 语句指明每条记录含四个变量。变量 class 与 name 后的“$”号表明这两个变量为字符型变量。

BY 语句指明依变量 class 对数据进行排序分组，变量 class 取值相同的归于一组。

第一个 PRINT 过程对原数据进行打印输出，第二个 PRINT 过程按排序后的顺序打印输出。

许多 SAS 过程可以利用 BY 语句对数据分组进行处理。例如学生考试成绩分析，希望分班级 class 计算各门课的平均成绩，可以使用这样的程序：

```
PROC MEANS;
VAR math stat;
```