

第 2 讲 SAS 数据

要进行数据分析，必须先弄清楚利用 SAS 软件进行统计分析的对象是什么。本讲简要介绍了 SAS 数据库和数据集以及如何在 SAS 软件中创建数据库和数据集。并介绍了导入外部数据的方法。本讲的重点和难点均在于如何在 SAS 中创建数据库和数据集等。

2.1 SAS 数据库和数据集

2.1.1 临时数据库和永久数据库

SAS 数据集存储在被称为 SAS 数据库的文件夹中，称为逻辑库(library)，根据存储方式的不同，分为永久数据库和临时数据库。

临时数据库只有 1 个，名为 WORK。它在每次启动 SAS 系统后自动生成，关闭 SAS 时，临时库中的数据被自动删除。临时数据库被认为是默认的数据库，在程序中引用该库中的数据可以省略库名。

永久数据库可有多个，且库中的数据被保存起来，再次启动系统时可以继续使用。SASUSER、SASHELP 是 SAS 自带的永久库，每次启动时都会自动显示库名，并显示其中的数据集。SASUSER 库保存与用户个人设置有关的文件，退出 SAS 时文件不会被删除。SASHELP 库保存与 SAS 帮助系统、例子有关的文件，也是永久数据库。

除了用工具栏中的“添加新库”按钮创建新的逻辑库，创建新库可以通过在资源管理器窗口右键单击来命名，或使用 LIBNAME 语句命名。下面举例讲解如何利用前两种方法创建永久数据库，假设这个数据库库名叫“mybase”。通过“创建新库”按钮创建永久性数据库的步骤已在 1.3.2.1 中介绍过，此处不再赘述。

【例 2.1.1】 利用资源管理器窗口创建

- (1) 在资源管理器窗口中，双击逻辑库(library)图标；
- (2) 在空白处单击右键，单击“新建”；
- (3) 在“新建逻辑库”窗口中，输入库名称“mybase”，选择“启动时启用”，以便每次启动 SAS 系统时，mybase 数据库能自动加载；

(4) 选择对应于这个数据库的目录(可以是计算机中的任何目录,也可创建新目录);

(5) 单击“确定”,新库建成,mybase 即出现在当前数据库列表中。

【例 2.1.2】 在程序编辑窗口用 LIBNAME 语句创建

可用 LIBNAME 语句指定永久库的库名,格式为: LIBNAME 库名“文件夹路径”。如指定“D:\SASDATA”为新库,名为“mybase”,可提交以下语句: LIBNAME mybase ‘D:\SASDATA’。注意:库名可以随意指定。

2.1.2 临时数据集和永久数据集

什么是数据集?数据集由若干个观测组成,观测的集合即数据集。

SAS 数据集分为两类,一类是临时数据集,另一类是永久数据集。临时数据集仅在当前会话期间有效,一旦退出 SAS,临时数据集就被删除。永久数据集是指存储在 SAS 外部存储介质上的数据集。

数据集的性质以数据集的名称来标识,每一个 SAS 数据集都有一个两级名称,第一级是库名,它指明该数据集所在的存储位置,第二级是数据集名,标识特定的数据集。两级名中间用“.”隔开,即“库名.数据集名”,如“work.example”,“mybase.example”。

SAS 的临时数据集全部存入库名 WORK 所对应的临时目录下,系统指定 WORK 作为临时数据集的第一级名称,通常这一级名称可以省略。而永久数据集须指定其存储位置,不能省略第一级名称,所以永久数据集必须由两级名称来标识。

2.2 创建 SAS 数据集

创建 SAS 数据集可以通过菜单法和编程法两种方法实现。

2.2.1 用菜单法创建数据集

【例 2.2.1】 利用表编辑器创建 SAS 数据集

利用表编辑器(Viewtable)可以直接创建 SAS 数据集,方法如图 2-2-1 所示:选择菜单栏中的“工具”→“表编辑器”,将打开如图 2-2-2 所示的新建空数据表文件。

【例 2.2.2】 利用资源管理器创建 SAS 数据集

打开资源管理器的逻辑库文件夹下的 WORK 子目录,在其右键弹出式菜单中单击“新建”菜单项,如图 2-2-3 所示,将打开如图 2-2-4 所示的窗口,在其中选择新建“表”。单击“确定”后将同样打开如图 2-2-2 所示的新建空数据表文件。

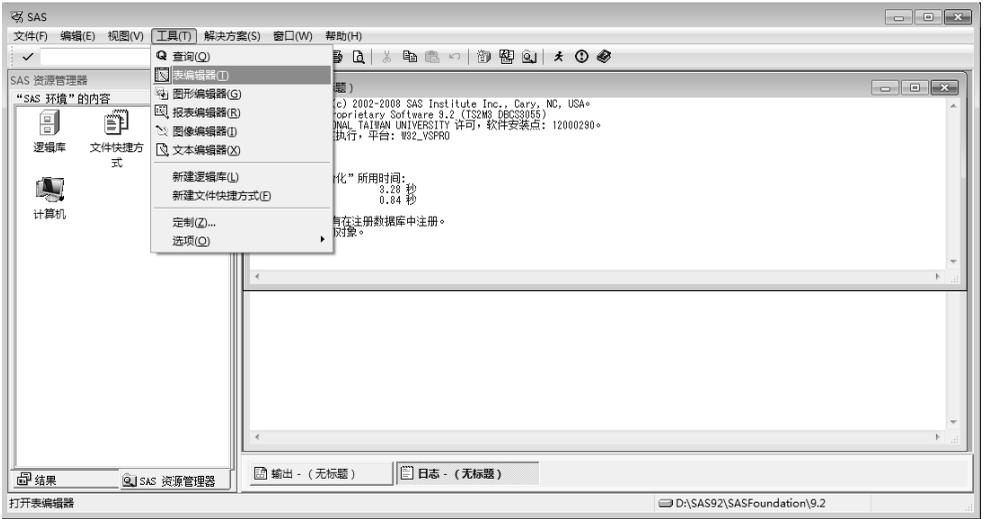


图 2-2-1 表编辑器中数据表的创建

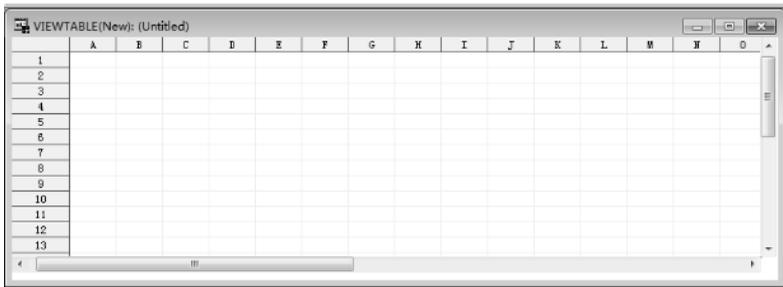


图 2-2-2 新建空数据表文件



图 2-2-3 资源管理器中数据表的创建



图 2-2-4 数据表的新建

【例 2.2.3】 利用 INSIGHT 创建 SAS 数据集

在 SAS 菜单中选择“解决方案”→“分析”→“交互式数据分析(interactive data analysis)”，如图 2-2-5 所示，打开“SAS/INSIGHT”对话框，如图 2-2-6 所示。选中逻辑库，可新建或打开已有的数据集。

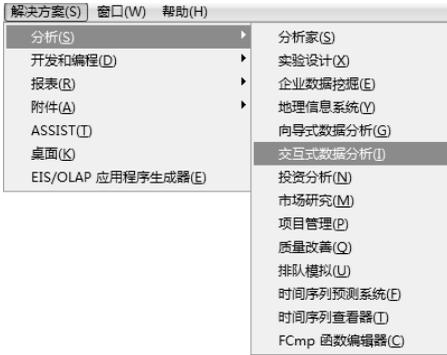


图 2-2-5 INSIGHT 中数据表的创建



图 2-2-6 数据表的新建

以上三种方法都可以创建一个空白“表”，在表中输入数据，再选择菜单栏中的“文件”→“保存”，在弹出的对话框中选择逻辑库，并给数据集命名，即可保存为一个 SAS 数据集。

2.2.2 用编程法创建数据集

利用表编辑器创建数据集比较方便，但表编辑器仅能显示和编辑数据集，无法对数据进行预处理，SAS/INSIGHT 对数据预处理的功能也有限。要想使数据更符合统计分析的需要，必须掌握传统的创建数据集的方法，即用编程法创建数据集。

SAS 程序由多个基本语句构成，而 SAS 程序按照功能的不同可以分为数据步和过程步。数据步主要用于数据文件的创建和管理，过程步指进行统计分析的程序。此处主要介绍 SAS 数据步。

在 SAS 系统中，数据步用于创建和管理数据集，SAS 数据步一般包括如下语句：

```
DATA 数据集名;           /*创建数据集并命名*/
INPUT 变量列表 <@@>;    /*列出数据集的变量名*/
LABEL 变量标签;         /*定义数据集中变量的标签*/
CARDS;                  /*数据区的开始标识*/
数据
;                        /*数据区的开始标识*/
RUN;                    /*数据步程序提交运行*/
```

上述语句中的 DATA、INPUT、LABEL、CARDS 为 SAS 语句的关键词。

其中，DATA 语句用于指定数据集名。如果用户省略逻辑库名，创建的数据集存储在 WORK 临时逻辑库中，SAS 软件关闭后将不会存在；若数据集指定在永久逻辑库，此时数据集名应为“逻辑库名.数据集名”。

INPUT 语句用于顺序列出输入数据的变量名，各个变量之间通过空格间隔，默认情况下输入的变量为数值型。如果输入的数据为字符串型，需要在变量名后加“\$”符号。INPUT 语句中的“@@”表示按照 INPUT 定义的变量顺序依次连续读入数据，无论数据分为多少行，遇到“;”时则停止数据读入。

CARDS 语句用于标识数据的开始，且输入的数据各列应与 INPUT 语句定义的变量顺序一致。数据在 CARDS 语句后开始输入，各个数据之间至少通过一个空格间隔。如果输入的数据中有缺失值，需要使用“.”标识。最后，所有数据输入完毕后，“;”分号不可以忘记。

RUN 语句用于向 SAS 系统提交数据步的程序。

【例 2.2.4】 用 CARDS 或 DATALINES 语句创建数据集
在程序编辑器中输入如下程序：

```
data score;
input x y;
cards;
34 56
78 90
35 67
89 10
23 65
77 45
;
run;
proc print;
run;
```

运行上述程序，可得到如下数据集。

Obs	x	y
1	34	56
2	78	90
3	35	90
4	89	10
5	23	65
6	77	45

用“datalines”替换“cards”可得同样结果。“Proc print; Run;”用来预览所创建的数据集。可省略。在临时数据库“work”中双击“score”亦可预览所创建的数据集。

启动 SAS 后，系统会自动创建一个临时数据存储区，用来临时存储运行 SAS 时创建或调用的 SAS 数据集。临时数据库的库名为 WORK，在 SAS 启动后自动生成，结束 SAS 后，库中的文件都会被删除。对数据库的库名不需要标注，即 SAS 程序中数据集 WORK.score 与 score 所表示的含义完全相同。即下面两个程序产生的结果完全相同：

```
data score;
input x y @@;
cards;
34 56 78 90 35 67 89 10 23 65 77 45
;
run;

data work.score;
input x y @@;
cards;
34 56 78 90 35 67 89 10 23 65 77 45
;
run;
```

2.3 导入外部数据

2.3.1 外部数据

对于 SAS 来说，数据存储按存储地址可分为两类：一类是保存在 SAS 系统外部的数据文件；另一类是存储在 SAS 逻辑库中的数据文件。

SAS 支持目前流行的大部分格式的数据文件和数据库，包括 SPSS、文本、数据库、JMP、Excel、Lotus、Access 以及 CSV 等格式的数据文件。因此，不论用户的数据以什么格式存储在计算机里，SAS 基本都可以帮助用户将数据导入到 SAS 中，进行进一步的分析处理。

2.3.2 外部数据的导入

1. 利用菜单法导入外部数据

SAS 系统支持直接录入数据，也支持外部数据的导入。但在实际的使用中，用户的数据量可能比较大，直接录入比较烦琐，同时，一般的数据文件为了便于查看

处理,多以 Excel 数据文件格式存储。对于 Excel、SPSS 以及文本格式的数据,Import 是最快捷易学的数据导入方法。使用 Import 向导工具,在导入数据的同时还可输出相应程序,初学者可借此学习和 Import 过程有关的程序。这种导入方法相当于对数据进行了复制和转化。下面通过一个实例具体演示如何在 SAS 中导入 Excel 数据。

【例 2.3.1】 利用菜单法导入外部文本文件数据

(1)通过菜单中的“文件”→“导入数据”,可以打开如图 2-3-1 所示的“数据导入向导”对话框。

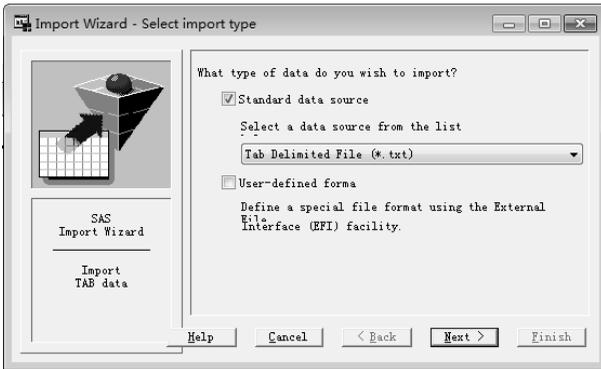


图 2-3-1 选择导入方式窗口

(2)从图 2-3-1 所示界面的下拉菜单中选择“Tab Delimited File (*.txt)”,单击“Next”按钮,会弹出“Select file”对话框,如图 2-3-2 所示,单击“Browse”按钮打开要导入的 txt 文件,之后单击“OK”按钮,进入选择库和文件名窗口,如图 2-3-3 所示。这里选择 WORK 临时库,输入数据集名“score”。

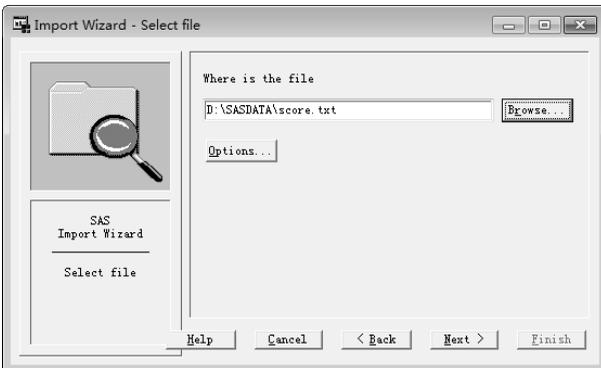


图 2-3-2 “Connect to MS Excel”对话框

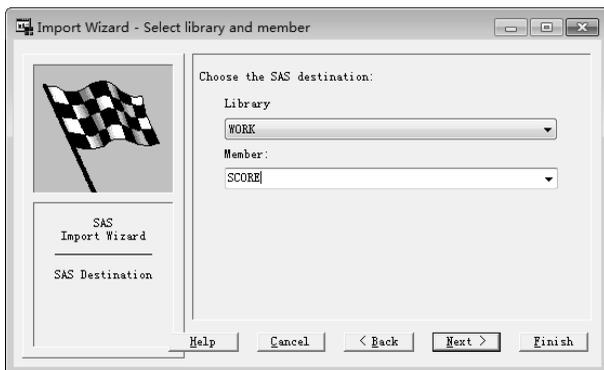


图 2-3-3 选择库和文件名窗口

导入向导可以自动生成使用 Import 过程导入数据的 SAS 程序代码。“Browse”按钮指定存储位置，即可保存程序到对应文件，如图 2-3-4 所示。



图 2-3-4 生成 SAS 程序窗口

数据文件导入完成后，用户可以在临时数据库中打开数据集文件查看其中的数据。

当 Excel 和 SAS 的版本有冲突时，可通过将 Excel 数据文件转化成.csv 数据文件或文本数据文件，完成数据的导入。

2. 用 INFILE 语句导入外部文件上的数据

在 SAS 系统中，可以用下面的语句导入数据集。

```
DATA 数据集名;           /*创建数据集并命名*/
INFILE '文件路径' [选项] /*从指定的路径下读取数据文件*/
INPUT 变量列表;         /*列出数据集的变量名*/
RUN;                     /*数据步程序提交运行*/
```

其中，INFILE 语句用于在指定的文件路径下读取数据文件，此处的文件路径应

为完整的文件路径，同时选项中可以设置 SAS 读取数据的相关参数，例如读取的数据行数等。INPUT 语句用于顺序列出读取的变量名。

【例 2.3.2】 用 INFILE 语句导入外部数据

```
data score ;  
infile 'D:\SASDATA\score.txt';  
input x y;  
put x= y= ;  
run;
```

执行上述程序将从外部文件输入数据，并在日志窗口显示如下数据。

```
x=34 y=56  
x=78 y= 90  
x=35 y= 67  
x=89 y=10  
x=23 y=65  
x=77 y= 45
```

对于初学者来说，建议只阅读关于使用向导导入数据的内容即可，这部分内容直观易懂。可以在深入了解 SAS 后，再返回学习其他几种略复杂的访问外部数据的方法。

2.4 本讲小结

SAS 中的一切统计分析只能对数据集中的数据进行分析。通过本讲的学习，读者将对 SAS 数据库和数据集、如何在 SAS 中快速创建内部数据文件以及对于外部数据文件的导入操作有基本的了解。在后续的讲节中，本书会循序渐进地为用户介绍 SAS 软件的具体功能。