返回总目录

目　　录

第 **1** 章　**SAS** 系统内七种常用的描述性统计程序 4
1.1　七种常用的程序 4
1.2　PROC MEANS，SUMMARY，UNIVARIATE，TABULATE 5
1.3　PROC CHART，TABULATE 程序的异同 5
第 **2** 章　描述性统计分析：统计程序 **PROC MEANS** 7
第 **3** 章　产生描述性统计值的输出文件：统计程序 **PROC SUMMARY** 8
3.1　PROC SUMMARY 程序概述 8
3.2　如何撰写 PROC SUMMARY 程序 8
3.3　范　　例 14
3.4　注 意 事 项 15
第 **4** 章　描述性统计值的计算与绘图：统计程序 **PROC UNIVARIATE** 18
4.1　PROC UNIVARIATE 程序概述 18
4.2　如何撰写 PROC UNIVARIATE 程序 18
4.3　范　　例 22
4.4　注 意 事 项 24
第 **5** 章　统计值的图形表示：统计程序 **PROC CHART** 26
5.1　PROC CHART 程序概述 26
5.2　举 例 说 明 27
5.3　如何撰写 PROC　CHART 程序 39
第 **6** 章　统计表格的制作：统计程序 **PROC TABULATE** 44
6.1　PROC TABULATE 程序概述 44
6.2　举 例 说 明 44
6.3　表格制作的基本概念 46
6.4　如何撰写 PROC TABULATE 程序 49
6.5　注 意 事 项 55
6.6　范　　例 65
第 **7** 章　关系强度的测量：统计程序 **PROC CORR** 77
7.1　PROC CORR 程序概述 77
7.2　如何撰写 PROC CORR 程序 77
7.3　范　　例 82
7.4　注 意 事 项 89
第 **8** 章　一般制图：统计程序 **PROC PLOT** 91
8.1　PROC PLOT 程序概述 91
8.2　如何撰写 PROC PLOT 程序 91

8.3 如何在同一页的报表纸上多重绘图 97
8.4 范　　例 99

第一部分
描述性统计分析

第 1 章　SAS 系统内七种常用的描述性统计程序

1.1　七种常用的程序

本章的目的旨在介绍 SAS 系统内七种常用的描述性统计程序。顾名思义，这七种程序旨在形容样本（Sample）的平均数、标准差、偏度、峰度等统计值或为样本的数据绘图、制表格。因此，这些程序的执行与统计分配理论（Statistical Distribution Theory）或母群参数的估计值无关。

第 2 章至第 8 章，我们将依序介绍 PROC MEANS，PROC SUMMARY，PROC UNIVARIATE，PROC CHART，PROC TABULATE，PROC CORR 及 PROC PLOT 等程序。为了深入浅出地介绍、比较这七种描述性统计程序，我们设计了下表供读者参考，表中以数字 1~7 代替这七种程序(有关统计值的定义，请查所属的章节)：

1 = PROC MEANS
2 = PROC SUMMARY
3 = PROC UNIVARIATE
4 = PROC CHART
5 = PROC TABULATE
6 = PROC CORR
7 = PROC PLOT

表 1.1　SAS 系统内七种常用的描述性统计程序之比较

统计值 \ 统计程序	1	2	3	4	5	6	7
1　遗漏数据	X	X	X	X	X		
2　有效数据	X	X	X		X	X	
3　加权值的总和	X	X	X		X		
4　平均数	X	X	X	X	X	X	
5　总和	X	X	X	X	X	X	
6　最大值 / 最小值	X	X	X		X	X	
7　全距=最大值-最小值	X	X	X		X		
8　(未)矫正过的离差平方和	X	X	X		X		
9　变异数	X	X	X		X		
10　标准差	X	X	X		X	X	
11　标准误	X	X	X		X		
12　变异系数	X	X	X		X		
13　偏度 / 峰度	X		X				
14　t 检定值 / 统计显著度	X	X	X		X		
15　中数			X			X	
16　四分位数 / 众数			X				
17　皮尔逊的积差相关系数						X	
18　Spearman 的等级相关系数						X	
19　Kendall 的相关系数						X	
20　Hoeffding 的 D 相关系数						X	

续表

统计值 \ 统计程序	1	2	3	4	5	6	7
其它特征							
21　产生报表输出文件	是	否	是	是	是	是	是
22　产生 SAS 输出文件	是	是	是	否	否	是	是
23　含 CLASS 指令	否	是	否	否	是	否	否
24　含 BY 指令	是	是	是	是	是	是	是

1.2　PROC MEANS，SUMMARY，UNIVARIATE，TABULATE 程序的异同

从表 1.1 中，读者或许会认为 MEANS，SUMMARY，UNIVARIATE 与 TABULATE 等四个程序大同小异，可以彼此取代。实际上，它们相同之处的确远多于相异之处。以下详细讨论这四个程序的异同处：

同

平均计算上节表 1.1 中所列的 (1)~(12) 以及 (14) 等十六种统计值，而且均可借 BY 指令将输入资料文件分成几个小资料文件，以便分别计算各统计值。

异

一、PROC MEANS 与 UNIVARIATE 计算样本的偏度与峰度。然而，PROC SUMMARY 与 TABULATE 并不计算这两种统计值。

二、PROC UNIVARIATE 计算样本的四分位数与众数，然而 PROC MEANS，SUMMARY 与 TABULATE 不会计算这两种统计值。

三、PROC SUMMARY 不列出任何计算结果。然而，PROC MEANS，UNIVARIATE 与 TABULATE 会列出计算的结果。

四、PROC MEANS，SUMMARY 与 UNIVARIATE 将计算结果收集在一个 SAS 输出资料文件内，以供进一步的分析。但是，PROC TABULATE 无法产生任何输出资料文件，只将计算结果列出而已。

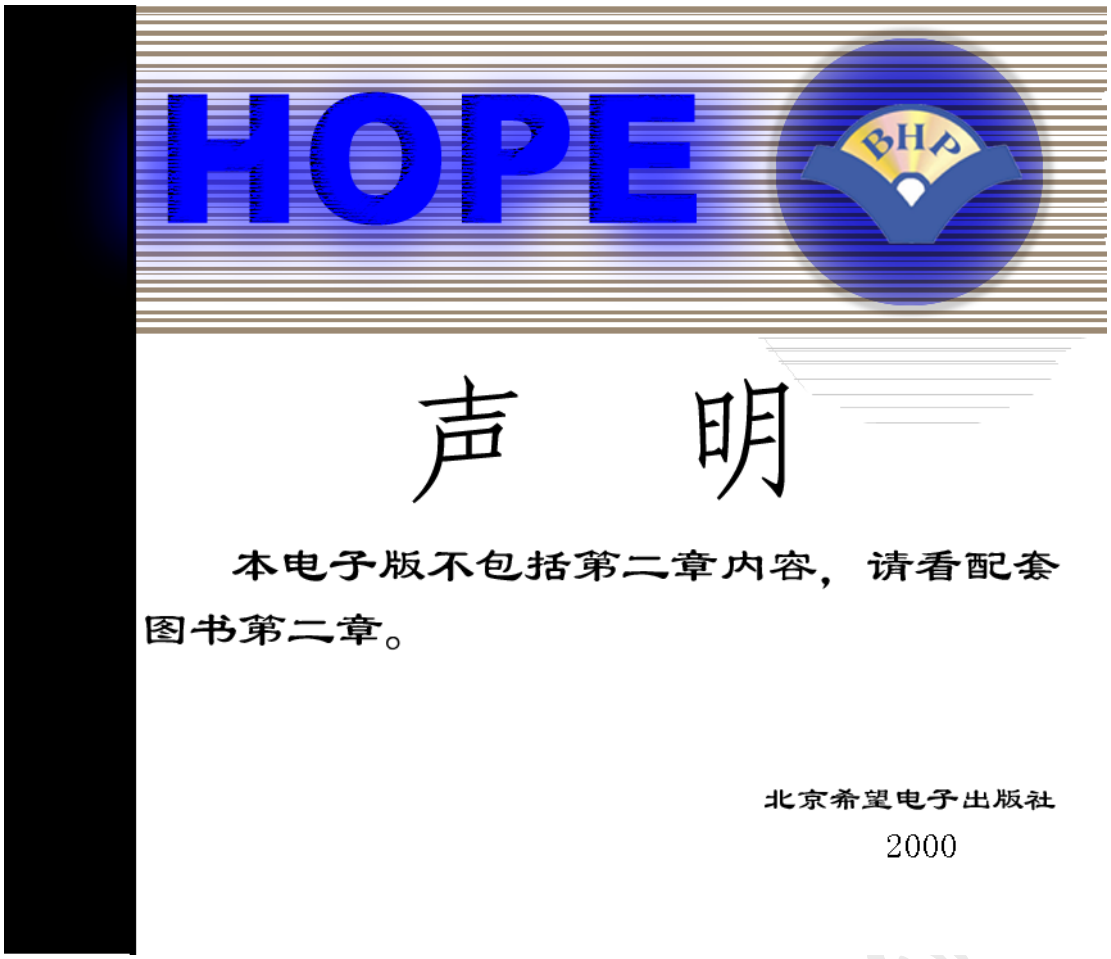
五、PROC SUMMARY 与 TABULATE 含 CLASS (分组) 的指令。然而，PROC MEANS 与 UNIVARIATE 的程序中不可包括 CLASS (分组) 的指令。

1.3　PROC CHART，TABULATE 程序的异同

CHART 与 TABULATE 两程序的功能都在于呈现统计值的大小。然而，PROC CHART 是用图形的方式呈现，PROC TABULATE 却以表格的方式呈现。此外，PROC CHART 在图形上所呈现出来的统计值只包括平均数、总和、(累积) 次数及 (累积) 百分比等。然而，PROC TABULATE 在表格内所能呈现的统计值则包括了上节表 1.1 中第 (1)~(12) 以及 (14) 等十六种统计值。

因此，当读者在斟酌选用那一种统计程序之前，必须首先明白这七种统计程序的异同以及描述、整理样本数据的最终目的。本章第 1.2 节所提供的表 1.1 应能助你选择一个较精简又完备的描述性统计分析程序!

第 2 章　描述性统计分析：统计程序 PROC MEANS



声　明

本电子版不包括第二章内容，请看配套图书第二章。

北京希望电子出版社
2000

第 3 章　产生描述性统计值的输出文件：统计程序 PROC SUMMARY

3.1　PROC SUMMARY 程序概述

本程序可用来计算数值变量的描述性统计值，如：平均数、标准差、峰度、偏度、最大值，及最小值等等，并将这些统计值储存在一个 SAS 的输出文件内，以供进一步的分析。

比较 SUMMARY 程序与 MEANS 程序

这两个程序有一个相同之处，两个相异之处。

同

两个程序都可以用来计算数值变量的描述性统计值。

异

(1) 输出文件不同

SUMMARY 程序只能产生含统计值的输出文件，而不能产生报表输出文件。MEANS 程序则可以同时产生两种输出文件。

(2) 执行分组的指令不同

虽然而两个程序都可以将输入的文件，按某个或某些变量的值，将观察体加以分组，然后对各组分别进行分析，但两程序用不同的指令来执行分组。在 MEANS 程序中，BY 指令是唯一可用来执行分组的指令。但在 SUMMARY 程序中，可使用如下的三种方法来执行分组：CLASS 指令，BY 指令，或同时使用 CLASS 及 BY 指令。

3.2　如何撰写 PROC SUMMARY 程序

PROC SUMMARY 含八道指令，它们的格式如下：

PROC SUMMARY	选项串;
VAR	变量名称串;
CLASS	变量名称串;
BY	变量名称串;
FREQ	变量名称;
WEIGHT	变量名称;
ID	变量名称串;
OUTPUT	OUT=统计值输出文件名称/统计值关键字符串;

其中，PROC SUMMARY，VAR 及 OUTPUT 指令是必须的，不可省略。除 PROC SUMMARY 指令必须为首外，其余各指令无先后顺序的限制。每一个 SUMMARY 程序中只可以有一个 OUTPUT 指令。

指令 #1　PROC SUMMARY 选项串；

有七个选项可供选择，分述如下：

(1) DATA= 输入文件名称

指明到底取用那一个文件来计算统计值。若省略此选项，则 SAS 程序会自动找出在此程序之前最后形成的文件，并为此文件内的变量计算统计值。

(2) MISSING

视遗漏数据的观察体为 CLASS 变量下的一个组别，并且将它们纳入计算过程。

(3) NWAY

只输出在系统变量 _TYPE_ 上具有最大值之分组的统计值。_TYPE_ 是系统变量，其值代表 CLASS 变量分类的情形。有关 _TYPE_ 的详细介绍，请参阅本章第 3.4 节的注意事项。

(4) IDMIN

若先前已界定 ID 指令，则输出文件中 ID 变量的值，取各分组中 ID 变量最小的值来代表。请参阅 ID 指令的介绍。

(5) DESCENDING

要求统计值输出文件内，各分组之统计值的呈现次序，是根据系统变量 _TYPE_ 的值由大而小排列。_TYPE_ 值最大的分组，其统计值会最先呈现在输出文件内；_TYPE_ 值最小的分组，其统计值将最后呈现。_TYPE_ 的最小值为 0，代表所有的观察体 (不分组) 的总体。换句话说，所有观察体的统计值将在最后呈现。若省略此选项，则分组之统计值的呈现次序，根据 _TYPE_ 值由小而大排列。另外，若同时选用 NWAY 及 DESCENDING 两选项，则 DESCENDING 选项的界定变成无效。

(6) ORDER=FREQ

ORDER=DATA

ORDER=INTERNAL (内设值)

ORDER=EXTERNAL 或 FORMATTED

此选项决定输出文件内各分组的呈现顺序。有四种可能的呈现顺序：

FREQ：依每个分组的观察体总数，作由大而小的顺序呈现。

DATA：依输入文件内分组的呈现顺序为分组的输出顺序。

INTERNAL：类别次序由英文字母先后决定。

EXTERNAL 或 FORMATTED：类别次序由外在格式决定。

(7) VARDEF=N

VARDEF=DF

VARDEF=WEIGHT(或 WGT)

VARDEF=WDF

此选项决定计算变异数所用的分母：

N：	观察体总数。
DF：	观察体总数减去 1，这是此选项的内设值。
WEIGHT(或 WGT)：	加权后的观察体总数。
WDF：	就是上述 WEIGHT 值减去 1。

指令 #2　VAR 变量名称串；

读者可在本指令中列举所有参与分析的数值变量之名称。

指令 #3　CLASS 变量名称串；

本指令列举一个或多个分组变量。比方说，有两个分组变量：SEX (下分男、女) 及 SCHOOL (下分重点中学和非重点中学) 可界定为：

```
CLASS SEX SCHOOL;
```

本指令会产生九种分组，分别是：

组别	SEX	SCHOOL
(1)	男女混合	重点和非重点混合
(2)	男	重点和非重点混合
(3)	女	重点和非重点混合
(4)	男女混合	重点
(5)	男女混合	非重点
(6)	男	重点
(7)	男	非重点
(8)	女	重点
(9)	女	非重点

PROC SUMMARY 将分别计算九种分组的统计值。请注意，若用 CLASS 指令分组，则不必先用 PROC SORT 将观察体按分组变量的值加以排列。

指令 #4　BY 变量名称串；

BY 指令与上述的 CLASS 指令相同之处，在于两者都是用来界定分组变量的。但 BY 指令与 CLASS 指令有两点相异之处：第一，用 BY 指令分组时，一定要用 PROC SORT 将观察体按分组变量排列。第二，虽用同样的分组变量，但 BY 指令所产生的组别和 CLASS 指令不同。比方说，用 BY 指令界定和上述相同的两个分组变量，SEX 及 SCHOOL：

```
BY SEX SCHOOL;
```

如此界定的分组有四个，它们分别是：

组别	性别	校别
(1)	男	重点

(2)　　男　　非重点
(3)　　女　　重点
(4)　　女　　非重点

看到上述结果，读者或许会问，若写：

```
CLASS   SEX;
BY      SCHOOL;
```

则会产生什么样的分组呢? 答案是六个分组。它们分别是：

组别　　SCHOOL　SEX
(1)　　重点　　男女混合
(2)　　重点　　男
(3)　　重点　　女
(4)　　非重点　男女混合
(5)　　非重点　男
(6)　　非重点　女

所以结论是：读者必须学会如何巧妙的使用 CLASS 或 BY 指令来产生自己实际需要的分组。

指令 #5　FREQ 变量名称;

这个变量必须是输入文件中的一个数值变量，其值代表输入文件内每一个观察体重复出现的次数。若此变量的值含小数，则只取其整数部分。若其值小于 1 或是一个遗漏数据，则观察体将被剔除于计算过程外。

指令 #6　WEIGHT 变量名称;

这个变量必须是输入文件中的一个数值变量，其值代表每一观察体的加权值。这个变量的值可以是任何正实数。若其值小于 0，则将以 0 取代。WEIGHT 变量主要用于计算加权平均数、加权标准差，及加权变异数。其计算公式请见第 4 章 PROC UNIVARIATE 的说明。

指令 #7　ID 变量名称串;

ID 指令所列举的变量叫识别变量，它将与各统计值同时出现在输出文件内。

若在 ID 指令上只列举一个识别变量，则各分组在此识别变量上的最大值，就是呈现于输出文件上识别变量的值。但如果在 PROC SUMMARY 指令中界定 IDMIN 选项，则取各分组在 ID 变量上的最小值为输出文件内识别变量的值。

若在本指令中列举多个识别变量，则输出时识别变量的值又如何决定呢? 比方说 ID 指令中列举 ID1 及 ID2 两个识别变量串。两个观察体，其中一个的 ID1=1，ID2=777，另一个的 ID1=2，ID2=222。 SAS 处理时，将 ID1 与 ID2 的值合并成 1777 及 2222。1777 与 2222 互比大小。然后根据是否选用 IDMIN 选项而决定取那一个组合为输出文件中 ID1 与 ID2 的值。若选用 IDMIN，则取 ID1=1，ID2=777。否则，取 ID1=2，ID2=222。

指令 #8　OUTPUT OUT= 统计值输出文件名称　统计值关键字符串；

本指令的界定，将产生一个含统计值的输出文件。在这个输出文件中，观察体个数相当于 CLASS 及 BY 指令所形成的分组组数。若没有分组(亦即未使用 CLASS 或 BY 指令)，则观察体个数等于 1。这个 SAS 文件所含的变量是多个分组 (或全部观察体) 在 VAR 指令上所列举之变量的统计值。

指令 #8 OUTPUT 的选项列举于下：

(1) OUT= 统计值输出文件名称

为 SUMMARY 程序所产生的输出文件命名。若省略此选项，则内设文件名是 DATAn，其中 n 由 1 开始，依输出文件产生的顺序每次累加 1 。若读者欲将输出文件储存为永久性磁盘文件，则必须使用二段式命名。若使用一段式命名，则输出文件只能在同一个 SAS 程序中应用。程序结束后，这个输出文件即告消失。

(2) 统计值关键字符串

首先，我们说明有哪些关键字，其次再谈如何界定这些关键字。SUMMARY 程序内，有十六个代表统计值的关键字。下表是这些关键字及其所代表的意义：

关键字	意义
N	分组内或所有观察体的有效观察体总数
NMISS	分组内或所有观察体中含遗漏数据的观察体个数
MEAN	平均数
STD	标准差
MIN	最小值
MAX	最大值
RANGE	最大值与最小值的差
SUM	变量值的总和
VAR	变异数
USS	未矫正的平方和
CSS	矫正后的平方和
CV	变异系数 (Coefficient of Variation)
STDERR	平均数的标准误
T	t 检定，用来检定母群之平均数等于 0 的虚无假设是否成立
PRT	上述 t 检定的显著性
SUMWGT	WEIGHT 变量的总和

这些关键字的界定方式有四种：

(1) 统计值关键字= 代表统计值的变量名称串

这种界定方式自动计算出 VAR 指令所列举之所有变量的统计值。因此，等号右边所列举的变量，必须与 VAR 指令中所列举的变量前后对应而且数目相同。以下例而言，S1M 与 S2M 分别代表变量 S1 与 S2 的平均数：

```
PROC    SUMMARY;
        CLASS  GROUP;
        VAR  S1 S2;
        OUTPUT  OUT= FILE MEAN= S1M S2M;
```

(2)　统计值关键字 (变量名称串)= 代表统计值的变量名称串

这种界定方式将产生 VAR 指令中所列举之部分变量的统计值。以下例而言，输出文件 WORK.FILE 中，包括变量 S1 与 S2 的平均数 S1M 与 S2M；但只包括 S2 的标准差 S2SD：

```
PROC    SUMMARY;
        CLASS GROUP;
        VAR S1 S2;
        OUTPUT OUT= FILE  MEAN= S1M S2M
               STD(S2)= S2SD;
```

(3)　统计值关键字=

这种表达方式在等号右边是空白，未界定任何代表统计值的变量名称。因此，将以原变量的名称来代表。以下例而言，在输出文件中，S1 代表原变量 S1 的平均数，S2 代表原变量 S2 的平均数：

```
PROC    SUMMARY;
        CLASS GROUP;
        VAR S1 S2;
        OUTPUT OUT= FILE  MEAN=;
```

须注意，若使用这种界定方式，则一个 VAR 变量，只能产生一种统计值，而不能同时产生其他种统计值。这是因为一个 VAR 变量的名称只能同时代表一种统计值。

(4)　统计值关键字 (变量名称串)=

这种表达方式在等号右边是空白，未界定任何代表统计值的变量名称。在等号左边的括号内列举部分 VAR 指令中所提到的变量。因此，这些变量的统计值，仍以变量的原名表示。以下例而言，输出文件中将包括变量 S1M 及 S2M (分别代表变量 S1 与 S2 的平均数)；此处还包括 S1 (代表原变量 S1 的最大值)：

```
PROC    SUMMARY;
        CLASS GROUP;
        VAR S1 S2;
        OUTPUT OUT= FILE MEAN= S1M S2M MAX(S1)=;
```

注意，一个在 VAR 指令中界定的变量名称只能代表一种统计值。因此，使用这种界定方式时，不要将同一个变量重复界定在不同的统计值关键字之后。

3.3　范　　例

例一：描述性统计值的输出

说明

(1) 文件 A，产生一个包含六个变量及八个观察体的 SAS 文件 WORK.A。

(2) SUMMARY 程序以 GRP 及 SEX 为分组变量，在各分组内计算 A，B，C 变量的统计值。观察体以 AREA 为识别变量。

(3) SUMMARY 程序产生一个叫 WORK.MSD 的输出文件。此文件将包括变量 MA，MB，MC (分别代表变量 A，B，C 的平均数)；S1，S2，S3 (分别代表变量 A，B，C 的标准差)；及 NA (代表变量 A 的观察体个数)。因为此例只界定 N=NA，所以只输出变量 A 的观察体个数。变量 B 与 C 的观察体个数不会被输出。因此，这个输出文件将包括十个读者自设的变量 (亦即 AREA，GRP，SEX，MA，MB，MC，S1，S2，S3，NA)，以及两个系统变量 (_TYPE_，_FREQ_)。

(4) WORK.MSD 文件包含九个观察体。它们分别是：

●_TYPE_=0 的观察体有一个

代表文件中所有观察体的统计值。由于文件中有两个观察体在分组变量 GRP 上有遗漏数据，并且 PROC SUMMARY 中未选用 MISSING 选项，所以实际参与分析的观察体只有六个 (_FREQ_=6)。

●_TYPE_=1 的观察体有两个

是由分组变量 SEX 所形成的两个分组 (SEX=1，GRP=. 及 SEX=2，GRP=.)。

●_TYPE_=2 的观察体有两个

是由分组变量 GRP 所形成的两个分组 (GRP=1，SEX=. 及 GRP=2，SEX=.)。

●_TYPE_=3 的观察体有四个

是由分组变量 SEX 与 GRP 交叉分类所形成的四个分组 (即 GRP=1，SEX=1 及 GRP=1，SEX=2 及 GRP=2，SEX=1 及 GRP=2，SEX=2)。

●因此，WORK.MSD 文件中将含 1+2+2+4=9 个观察体。

(5) WORK.MSD 文件中，每个观察体在识别变量 AREA 上的值是什么呢？由于在 PROC SUMMARY 中，未选用 IDMIN 选项；因此，它的值就是每个分组中，AREA 值最大者。对文件中所有观察体而言，AREA 值最大者是 555，所以对 _TYPE_ = 0 的观察体而言，其 AREA 值就是 555 。对文件中 GRP=1，SEX=2 的观察体而言，它们的 AREA 值分别是 222 及 100。因此 222 成为这一组别的 AREA 值。

程　序

```
DATA A;
     INPUT GRP A B C SEX AREA @@;
     CARDS;
```

```
1 80 90 70 1 111    . 70 60 80 2 200
1 70 80 70 2 222    2 60 80 70 1 444
1 70 60 60 2 100    2 55 65 70 1 555
. 90 70 90 1 333    2 90 80 70 2 300
;
PROC SUMMARY;
      CLASS GRP SEX; VAR A B C; ID AREA;
      OUTPUT OUT=MSD MEAN=MA MB MC STD=S1-S2 N=NA;
PROC PRINT;
RUN;
```

结　果

报表 3.1　描述性统计值的输出

SAS

OBS	AREA	GRP	SEX	_TYPE_	_FREQ_	MA	MB	MA	S1	S2	NA
1	555	.	.	0	6	70.8333	75.8333	68.3333	12.8128	11.1430	6
2	555	.	1	1	3	65.0000	78.3333	70.0000	13.2288	12.5831	3
3	300	.	2	1	3	76.6667	73.3333	66.6667	11.5470	11.5470	3
4	222	1	.	2	3	73.3333	76.6667	66.6667	5.7735	15.2753	3
5	555	2	.	2	3	68.3333	75.0000	70.0000	18.9297	8.6603	3
6	111	1	1	3	1	80.0000	90.0000	70.0000	.	.	1
7	222	1	2	3	2	70.0000	70.0000	65.0000	0.0000	14.1421	2
8	555	2	1	3	2	57.5000	72.5000	70.0000	3.5355	10.6066	2
9	300	2	2	3	1	90.0000	80.0000	70.0000	.	.	1

3.4　注 意 事 项

■遗漏数据的处理

一般而言，在 CLASS 变量上含遗漏数据的观察体将被剔除在分析过程之外。但若读者在 PROC SUMMARY 指令中选用 MISSING 选项，则这些含遗漏数据的观察体将自成一个分组，而且被纳入分析过程之中。

■分组的限制

CLASS 指令中所列举的分组变量，最多不可超过二十四个。而由分组变量所形成的交叉分类组别，最多不可超过 32767 个 (即 2^{15}-1)。

■输出文件所包含的变量

这里所谓的输出文件指包含统计分析值的 SAS 文件(注意：SUMMARY 程序不会产生报表输出文件)。SAS 输出文件所包含的变量如下：

*BY 指令中所列举的变量 (如果已界定了 BY 指令)。

*ID 指令中所列举的变量 (如果已界定了 ID 指令)。

*CLASS 指令中所列举的变量。
*系统变量 _TYPE_，此变量表示 CLASS 指令所界定之分组变量的分组情形。
*系统变量 _FREQ_，表示每个分组所包含的观察体个数。
*OUTPUT 指令内所列举的统计值。

■输出文件所包含的观察体个数

这个输出文件内的观察体以 CLASS 指令及 BY 指令所形成的分组为单位。一个分组就是一个观察体。分组的情形则由系统变量 _TYPE_ 来表示。

(1) 一个分组变量

若只有一个分组变量，如：

```
CLASS A;
```

则输出文件中包括两类观察体；一类是输入文件中所有观察体的统计值(_TYPE_=0)。另一类是分组变量 A 所形成的分组。每一个分组其实就是输出文件内的一个观察体，它们在_TYPE_ 上的值都是 1。

(2) 两个分组变量

若有两个分组变量，如：

```
CLASS B A;
```

则输出文件中除了包括上述 _TYPE_=0，_TYPE_=1 的观察体外，还包括 _TYPE_=2 及_TYPE_=3 的观察体。_TYPE_=2 的观察体是由分组变量 B 所形成的分组。而_TYPE_=3 的观察体是由分组变量 B 与 A 所形成的交叉分类组别。

(3) 三个分组变量

若再多加一个分组变量，如：

```
CLASS C B A;
```

则输出文件中，除了包含上述 _TYPE_=0，1，2，3 的观察体外，还有下列不同 _TYPE_ 值的观察体：

TYPE=4，表示 C 所形成的分组
TYPE=5，表示 A 与 C 所形成的交叉分类组别
TYPE=6，表示 B 与 C 所形成的交叉分类组别
TYPE=7，表示 A 与 B 与 C 所形成的交叉分类组别

当分组变量的数目逐渐增多时，分组所产生的观察体在系统变量 _TYPE_ 上的值，就如上述所介绍的过程一样继续累积上去。

下页的表格列举分组变量所产生的组别及其 _TYPE_ 变量的值，还有输出文件所包含的观察体个数：

表 3.1　指令 CLASS 所产生的组别及其 _TYPE_ 变量的值

分组变量			_TYPE_值	分组变量所界定的组别	具有相同 _TYPE_ 值的观察体个数	输出文件所包含的观察体个数
C	B	A				
0	0	0	0	文件中所有的观察体	1	
0	0	1	1	A	a	1+a （一个分组变量）
0	1	0	2	B	b	1+a+b+a*b
0	1	1	3	A*B	a*b	（两个分组变量）
1	0	0	4	C	c	1+a+b+a*b+c+a*c+b*c+
1	0	1	5	A*C	a*c	a*b*c
1	1	0	6	B*C	b*c	（三个分组变量）
1	1	1	7	A*B*C	a*b*c	
1 表示该分组存在。 0 表示该变量不存在。			A，B，C 是 CLASS 指令中所列举的分组变量		a，b，c 分别表示分组变量 A，B，C 的组别数	

第 4 章　描述性统计值的计算与绘图：统计程序 PROC UNIVARIATE

4.1　PROC UNIVARIATE 程序概述

统计程序 UNIVARIATE 与统计程序 MEANS，FREQ 及 SUMMARY 的功能大同小异；它们都可以计算数值变量的描述性统计值。但 UNIVARIATE 程序能够对变量的分配情形提供更多的信息。例如，

- 指出一个变量上的极端值。
- 计算四分位数 (Quartiles) 。
- 绘制分配图。
- 产生次数分配表。
- 检定资料是否呈现常态分配。
- 产生统计值输出文件，以供稍后的分析。

4.2　如何撰写 PROC UNIVARIATE 程序

PROC UNIVARIATE 含七道指令，它们的格式如下：

```
PROC   UNIVARIATE    选项串;
       VAR           变量名称串;
       BY            变量名称串;
       FREQ          变量名称;
       WEIGHT        变量名称;
       ID            变量名称串;
       OUTPUT        OUT=统计值输出文件名称/统计值关键字符串;
```

在一个 UNIVARIATE 程序中，可以多次使用 OUTPUT 指令，但其他六道指令只能出现一次。此外，PROC UNIVARIATE 指令后的六道指令可以按任何顺序出现。

指令 #1　PROC UNIVARIATE 选项串;

有下列七个选项可供选择：

(1) DATA=输入资料文件名称

指明到底对那一个资料文件进行分析。若省略此选项，则 SAS 会自动找出在本程序之前最后形成的资料文件，并对它进行分析。

(2) NOPRINT

若只要求产生统计值的输出文件，以供稍后的分析，而不想印出报表，可用此选项来抑止报表的产生。

(3) PLOT

读者可用此选项要求 UNIVARIATE 程序产生三种图形：

●茎叶图 (Stem-And-Leaf Plot) 或平行条状图 (Horizontal Bar Chart)

●盒状图 (Box Plot)

●常态概率图 (Normal Probability Plot)(参阅 Tukey，1977)

有关这三种图形的详细介绍及例子，请见范例及注意事项。

(4) FREQ

读者可用此选项要求 UNIVARIATE 程序产生一个次数分配表，这个表包括变量值的出现次数、百分比及累积百分比。

(5) NORMAL

此选项可用来要求 UNIVARIATE 程序检定输入资料是否呈现常态分配，并且输出其检定的结果。

(6) PCDLDEF={1/2/3/4/5}

UNIVARIATE 程序中有五种计算百分位数 (Percentiles)的方法。有关这五种方法的详细介绍，请参阅注意事项。此选项是用来决定某一种计算方法。如 PCDLDEF=1，表示用第一种计算方法，以此类推。若省略此选项，则 UNIVARIATE 程序会自动采用用第四种计算方法。

(7) VARDEF=N

VARDEF=DF

VARDEF=WEIGHT (或 WGT)

VARDEF=WDF

此选项决定计算变异数所用的分母：

N	：观察体总数。
DF	：观察体总数减去 1，这是本选项的内设值。
WEIGHT (或 WGT)	：加权后的观察体总数。
WDF	：上述 WEIGHT 值减去 1。

指令 #2　VAR 变量名称串；

此指令列举需要进行描述性统计分析的数值变量名称。若省略此指令，则 UNIVARIATE 程序将对输入资料文件中所有数值变量进行分析。若选用 OUTPUT 指令，则不可省略 VAR 指令。

指令 #3　BY 变量名称串；

UNIVARIATE 程序依据此指令所列举的变量，将资料文件分成几个小资料文件。然后就每个小资料文件，分别执行分析。选用此指令时，资料文件内的数据必须先按照 BY 变量串的值做由小到大的重新排列，这个步骤可借 PROC SORT 达成。

指令 #4　FREQ 变量名称；

这个变量必须是输入资料文件中的一个数值变量。其值代表观察体重复出现的次数。若此变量的值含小数，则取其整数部分。若其值小于 1，则此观察体将被剔除在计算过程之外。

指令 #5　WEIGHT 变量名称；

这个变量称为加权变量，其主要功用在于计算加权平均数、加权标准差及加权变异数。加权平均数的计算公式如下：

$$\overline{X}_w = \sum W_i X_i / \sum W_i$$

上述公式中，$\overline{X}_w$ 是加权平均数

W_i 是加权变量的值

X_i 是被加权的变量，即 VAR 指令中所列举的变量。

加权变异数的计算公式则是：

$$S_w^2 = \sum W_i (X_i - \overline{X}_w)^2 / d$$

上述公式中，S_w^2 是加权变异数，

d 是 PROC UNIVARIATE 指令中 VARDEF 选项的值。

请注意：若选用了 WEIGHT 指令，则 UNIVARIATE 程序将不计算偏度与峰度这两个统计值。这两个统计值将以遗漏数据 (.) 表之。

此外，WEIGHT 指令对四分位数、极端分数，及观察体总数的计算，并不发生任何作用。

指令 #6　ID 变量名称串；

这个变量称为识别变量，它对报表输出文件及统计值输出文件都有影响：

(1) 报表输出文件

报表输出文件在输出变量的五个最大值及五个最小值时，将会取第一个识别变量值的前八个文字或符号，一并输出。若无识别变量，则报表输出文件就只输出这些极端值，而不加上识别代号。

(2) 统计值输出文件

统计值输出文件内，将包含 ID 指令中所列举的所有变量串。由于此种输出文件可能包含所有观察体的统计值，因此文件内观察体个数只有一个。此种输出文件也可能包含分组 (因选用了 BY 指令) 的统计值。在这种情况下，文件内观察体个数等于组别数；而且输出文件内的识别变量值是每个组内第一个观察体在识别变量上的值。

指令 #7　OUTPUT OUT=统计值输出文件名称　统计值关键字符串；

本指令使 UNIVARIATE 程序产生统计值的输出资料文件。若省略此指令，则统计值

输出文件不会自动产生。

(1) OUT=统计值输出文件名称

读者可自设输出文件文件名。若将储存此输出文件为永久性的磁盘文件，必须使用包括文件型与文件名的二段式命名法。若只是暂时性的输出文件，则使用一段式命名即可。若省略此选项，则 SAS 将以内设的命名方式，自动给予 DATAn 的文件名 (如 DATA1，DATA2...) ，n 按输出文件产生的先后顺序，由 1 逐次累加而成。

(2) 统计值关键字符串

这些关键字代表输出的统计值，这些统计值在输出文件内的名称必须明白界定，否则不会输出。首先说明有那些关键字，然后谈如何来表达这些关键字。UNIVARIATE 程序内有二十六个统计值的关键字，下表是这些关键字及其所代表的意义：

关键字	意义
N	有效观察体个数
NMISS	含遗漏数据的观察体个数
NOBS	观察体总数
MEAN	平均数
SUM	变量值的总和
STD	标准差
VAR	变异数
SKEWNESS	偏度
KURTOSIS	峰度
SUMWT	所有观察体在 WEIGHT 变量上的总和
MAX	变量的最大值
MIN	变量的最小值
RANGE	最大值减去最小值所得的差
Q3	第三个四分位数
MEDIAN	中位数 (第 50 的百分位数)
Q1	第一个四分位数
QRANGE	Q3 减去 Q1 之差
P1	第 1 的百分位数
P5	第 5 的百分位数
P10	第 10 的百分位数
P90	第 90 的百分位数
P95	第 95 的百分位数
P99	第 99 的百分位数
MODE	众数，如果有不只一个众数，取最小值的那一个
SIGNRANK	等级符号检定法 (The Signed Rank Statistic；Lehmann，1975)
NORMAL	常态分配的检定 (Test Statistic for Normality) 若观察体个数少于 51，则采用 Shapiro-Wilk 的 W Statistic 的方法检定；否则，采用用 Kolomogorov 的 D Statistic 的方法检定

这些关键字的表达方式是：

统计值关键字 = 代表统计值的变量名称串

这些代表统计值的变量名称，必须根据 VAR 指令内所列举的变量顺序，对应地一一列举。未列举者，不予输出。请看下面这个例子：

```
PROC UNIVARIATE;
     VAR X Y;
     BY SEX;
     OUTPUT OUT=MSD MEAN=MX MY STD=SDX;
```

假如分组变量 SEX 的值是 1 或 2，则 UNIVARIATE 程序所产生的报表输出文件，将是两个 SEX 组在变量 X 与 Y 上的描述性统计值。而且因 OUTPUT 指令的界定，这个输出文件的文件名会是 WORK.MSD (暂时的文件) 。此文件包括四个变量，即：SEX，MX，MY 及 SDX。其中 SEX 是分组变量，MX 与 MY 分别是变量 X 与 Y 的平均数，SDX 是变量 X 的标准差。由于关键字 STD 后，只界定一个变量名称 (SDX) ，故此变量自动指 VAR 指令内所列举的第一个变量。至于变量 Y 的标准差，因未界定其相对应的变量名称，故无法输出。

4.3 范　　例

例一：如何检验变量的分配是否符合常态分配?

说　明

(1) DATA A; 阶段内，INPUT 指令读取 ID 及 CNT 两个变量，并界定位置指标 @@，表示一个资料卡可能包含数个观察体的资料。

(2) 在 UNIVARIATE 程序内；PROC UNIVARIATE 指令选用了三个选项：

FREQ　　选项产生变量的次数分配表。

NORMAL　选项检定样本是否呈现常态分配。

PLOT　　选项绘制茎叶图、盒状图，及常态概率图。

此外，VAR 指令界定分析的变量是 CNT。ID 指令界定识别变量是 ID。在输出极端分数时，这个识别变量的值也将一并输出。

程　序

```
DATA A;
     INPUT ID CNT @@;
     CARDS;
4005 20 4006 34
4205 53 4028 54 4523 55
4008 65 4009 65 4010 66 4011 67 4050 68 4055 69
4112 60 4222 61 4444 64 4422 65 4007 66 4114 66 4056 67
4118 70 4224 71 4312 71
```

```
    4048 72 4060 74 4023 74 4002 75 4213 78
    4125 88 4333 99 4001 120
    ;
    PROC UNIVARIATE FREQ NORMAL PLOT;
          VAR CNT;
          IDID;
    RUN;
```

结　果

报表可分三部分讨论：

第一部分包括描述性统计值及五个最大与最小的极端分数 。这是 PROC UNIVARIATE 的内设输出资料。

第二部分是由 PLOT 选项所产生的三种图形。

第三部分是由 FREQ 选项所产生的次数分配表。

报表 4.1　如何检验变量的分配是否符合常态分配

UNIVARIATE PROCEDURE

(第一部分)

Variable=CNT

Moments　　Quantiles(Def=5)　　Extremes

N	29	Sum Wgts	29	100% Max	120	99%	120	Lowest ID	Highest ID
Mean	67.48276	Sum	1957	75% Q3	72	95%	99	20(4005)	75(4002)
Std Dev	17.46143	Variance	304.9015	50% Med	67	90%	88	34(4006)	78(4213)
Skewness	0.247237	Kurtosis	3.904144	25% Q1	64	10%	53	53(4205)	88(4125)
USS	140601	CSS	8537.241	0% Min	20	5%	34	54(4028)	99(4333)
CV	25.87539	Std Mean	3.242506			1%	20	55(4523)	120(4001)
T:Mean=0	20.81192	Prob>|T|	0.0001	Range	100				
Sgn Rank	217.5	Prob>|S|	0.0001	Q3-Q1	8				
Num ^= 0	29			Mode	65				
W:Normal	0.876517	Prob<W	0.0025						

(第二部分)

```
Stem  Leaf                  #    Boxplot                Normal Probability Plot

 12   0                     1      *       130+                                  *
 10                                         |                                +++++++
  8   89                    2      0        |                            +++++*++*++
  6   014555666778901124458 21   +--+--+    |                ** **+*+**+***+*+* * *
  4   345                   3      |        |          ++**+*+++++
  2   04                    2      *        30++++++*+++*
      ----+----+----+----+-                 +----+----+----+----+----+----+----+----+----+

Multiply Stem.Leaf by 10**+1                    -2       -1        0       +1       +2
```

(第三部分)

Frequency Table

		Percents				Percents	
Value	Count	Cell	Cum	Value	Count	Cell	Cum
20	1	3.4	3.4	68	1	3.4	58.6
34	1	3.4	6.9	69	1	3.4	62.1
53	1	3.4	10.3	70	1	3.4	65.5
54	1	3.4	13.8	71	2	6.9	72.4
55	1	3.4	17.2	72	1	3.4	75.9
60	1	3.4	20.7	74	2	6.9	82.8
61	1	3.4	24.1	75	1	3.4	86.2
64	1	3.4	27.6	78	1	3.4	89.7
65	3	10.3	37.9	88	1	3.4	93.1
66	3	10.3	48.3	99	1	3.4	96.6
67	2	6.9	55.2	120	1	3.4	100.0

4.4 注 意 事 项

■ 遗漏数据的处理

处理的方法，依遗漏数据的性质而异。

(1) VAR 指令中的变量

若观察体在 VAR 指令之某个变量上含遗漏数据，则该观察体将被排除在这个变量的计算过程之外。 然而若在其他变量上无遗漏数据，仍会被纳入其他 VAR 变量的计算过程内。这些含遗漏数据的观察体个数，及它们占总观察体数的百分比，将被纳入报表输出文件。

(2) WEIGHT 指令中的变量

若在 WEIGHT 的加权变量上含遗漏数据，则观察体的加权变量值就是 0。这些观察体仍然而会被纳入百分位数的计算，以及极端分数的挑选。

(3) FREQ 指令中的变量

若在 FREQ 指令的次数变量上含遗漏数据，则观察体将被剔除在所有计算过程之外。

(4) BY 指令中的变量

若在 BY 的分组变量上含遗漏数据，则这些观察体在分析的过程中自成一个分组。

(5) ID 指令中的变量

若在 ID 的识别变量上有遗漏数据，则在需要识别观察体的地方，仍以遗漏值 (.) 呈现。

■ 百分位数的计算法

在 PROC UNIVARIATE 指令内的选项 PCTLDEF=，可用来选择百分位数的计算方法。UNIVARIATE 程序共提供五种计算法。这五种方法在 PCTLDEF= 选项中以 1 到 5 的整数值表示。

五种计算法在运算之前，都先将观察体按变量的值做由小而大的排序，排序数据以 X_1，X_2，X_3，...，X_n 表示，n 是有效观察体个数。欲求出第 t 个百分位数，y，则必须先找出百分位数在排序数据中的相对位置 j 和 g。j 和 g 的定义如下：

p=t/100

np=j+g

上面公式中，p 是以百分比表示的 t 值。np 的乘积表示第 t 个百分位数在排序数据中的相对位置；j 是 np 的整数值，g 是小数值。SAS 利用下面五种公式计算与 t 相对应的百分位数 y；其中第五种计算法是内设处理法。

第一种计算法：$y=(1-g)\ X_j+gX_{j+1}$

第二种计算法：$y=X_I$　其中，i 等于 np+1/2 的整数值

第三种计算法：$y=X_j$ (当 g=0 时)

$y=X_{j+1}$ (当 g>0 时)

第四种计算法：$y=(1-g)\ X_j+gX_{j+1}$　　公式中，(n+1) p=j+g

第五种计算法：$y=(X_j+X_{j+1})/2$ (当 g=0 时)

$y=X_{j+1}$　　(当 g>0 时)　公式中，np=j+g

■ PROC UNIVARIATE 指令中 PLOT 选项所绘制的三种图形

Tukey(1977) 曾大力鼓吹以纸笔的图示法来探究资料，以发掘资料的内部结构，作为进一步分析的参考。茎叶图及盒状图就是 Tukey 所提出的简易图示法。

(1) 茎叶图

茎叶图 (Stem-And-Leaf Plot)以分数为纵坐标，发生的次数为横坐标，将分数一一予以登录，来显示资料的分配情形。如果某一个分数间距 (Interval) 所包含的观察体个数超过 48，则不绘制茎叶图，而改绘平行条状图 (Horizontal Bar Chart)。

(2) 盒状图

盒状图 (Box Plot 或 Schematic Plot) 划出两条平行横线。下面的一条线指出第 25 的百分位数所在，上面的一条线则指出第 75 的百分位数所在。线中的加号(+)指出平均数所在。由第 75 的百分位数的平行线上划出一条垂直线，称为须(Whisker)。须的长度大约是第 25 与第 75 百分位数间距离的 1.5 倍。须之上，以星号 (*) 及 0 表示极端的分数。若极端分数值是第 25 与 75 百分位数之差的 3 倍 (或 3 倍以上)，则以星号 (*) 表示。若是在第 25 与 75 百分位数之差的 1.5 倍与 3 倍之间者，以 0 表示。

(3) 常态概率图

常态概率图 (Normal Probability Plot) 以标准常态分配的百分位数作为横轴，以实际观察值所求出的百分位数为纵轴。在图中，以星号 (*) 标示每个实际观察值，以加号 (+) 标示一条参考线，它是根据数据的平均数与标准差划出的。如果观察值呈现常态分配，则实际观察值应落在这线参考线上。图上的横坐标以下列公式求得：

$\Phi^{-1}[(r_i-3/8)/(n+1/4)]$　　其中，　Φ^{-1} 是标准常态分配的反函数

r_i 是资料值的排名 (Rank)

n 是有效观察体个数

第 5 章　统计值的图形表示：统计程序 PROC CHART

5.1　PROC CHART 程序概述

本章所介绍的统计程序 PROC CHART 适用于绘制下列的图形：横轴图、纵轴图、方形图、圆形图、与星形图。这些图形可以表示一个变量的描述性统计值或多个变量之间的关系。

请注意，若是想作集群分析的制图，则应该使用 PROC TREE。有关 PROC TREE 的介绍，请见第九部分第 46 章。

在使用本程序时，应该提供 SAS 三项资料：

1. 想制作那一种图形?
2. 想以那一种描述性统计值制图?
3. 在制图过程中，希望如何将变量的值加以归类?

(1) **图形的选择**

利用不同的指令可以要求 CHART 程序绘制不同的图形：

若希望一个	则须选用
●横轴图	HBAR　指令
●纵轴图	VBAR　指令
●方形图	BLOCK　指令
●圆形图	PIE　指令
●星形图	STAR　指令

(2) **描述性统计值的选择**

利用 TYPE= 选项可以指定 SAS 用某一种描述性统计值来制图：

若希望选用	则须指明
●次数制图	TYPE = FREQ
●百分比制图	TYPE = PCT
●累积次数制图	TYPE = CFREQ
●累积百分比制图	TYPE = CPCT
●总分制图	TYPE = SUM
●平均数制图	TYPE = MEAN

(3) **变量值的分类方法**

下列选项会决定变量值归类的方法：

若希望	则须选用
●将连续变量的值当作类别	DISCRETE
●并列图型的归类	GROUP=

- ●小组的数值归类　　SUBGROUP=
- ●以各区间的中点为连续变量或文字变量归类　　MIDPOINTS=
- ●以第二个连续变量的值（如：次数总和、或平均数）为纵轴的值　　SUMVAR=

一般而言，PROC CHART 可处理文字或数值变量。文字变量的名称及其值不可超过十六个英文字母的长度。若数值变量是连续性的，则 SAS 会自动决定区间的大小，但读者也可以自定区间的中点值。当一个数据恰巧落在两个相邻区间的边界上，则 SAS 自动将其归属在较高分的区间。若是类别变量或文字变量，则其数值自动决定其分类区间。

5.2 举 例 说 明

下面是十一个简单的图形例子。

例 1　次数的纵 / 横轴图

下面的纵轴图表示某一个皮鞋公司内男女职员的人数，其资料如下：

```
F 1 F 1 F 1 F 1 F 1 F 1 F 1
F 2 F 2 F 2 F 2 F 2 F 2 F 2
F 2 F 2 F 2 F 2 F 2 F 2 F 2
F 2 F 2 F 2 F 2 F 2 F 2 F 2
F 2 F 2 F 2 F 2 F 2 F 2 F 2
M 2 M 2 M 2 M 2 M 2 M 2 M 2
M 2 M 2 M 2 M 2 M 2 M 3 M 3
M 3 M 3 M 3 M 3 M 3 M 3 M 3
M 3 M 3 M 3 M 3 M 3 M 3 M 3
M 4 M 4 M 4 M 4 M 4 M 4 M 4
M 4 M 4 M 4 M 4 M 4 M 4 M 4
M 4 M 4 M 4 M 4 M 4 M 5 M 5
```

M=男性　　F=女性

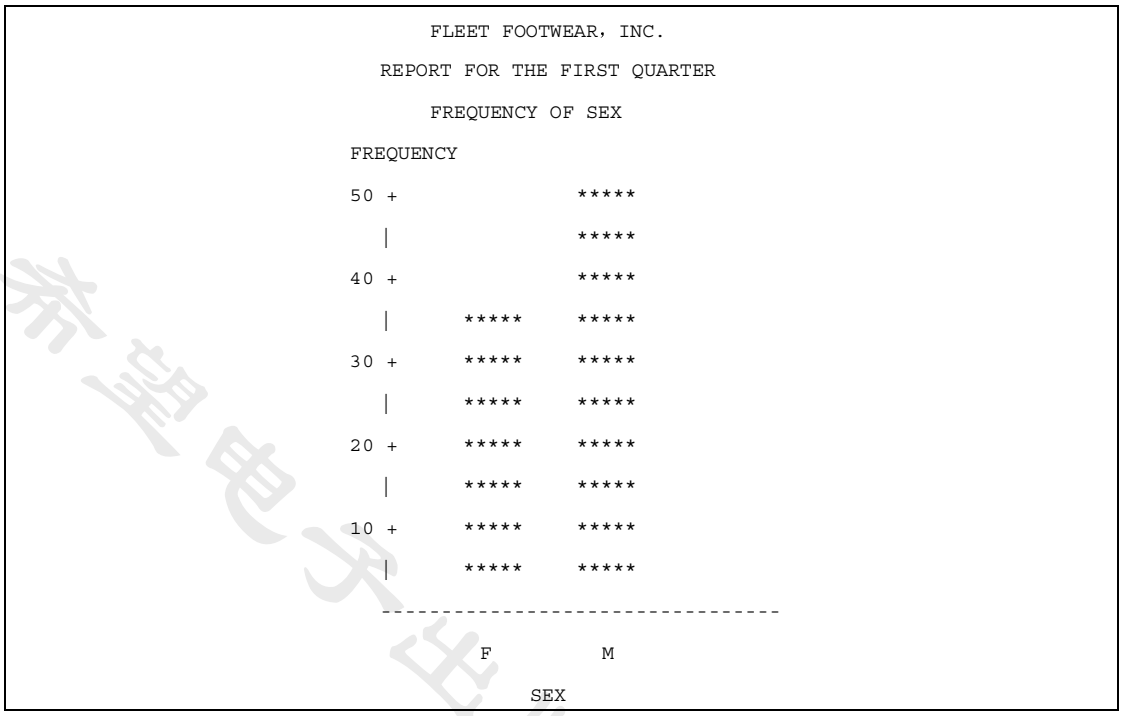
1=初中毕业　2=高中肄业　3=高中毕业

4=大学肄业　5=大学毕业

以下便是绘制纵轴图的程序：

```
OPTIONS PAGESIZE=25;
PROC CHART;
     VBAR SEX;
TITLE 'FLEET FOOTWEAR, INC.';
TITLE1 'REPORT FOR THE FIRST QUARTER';
```

结果如下图所示：

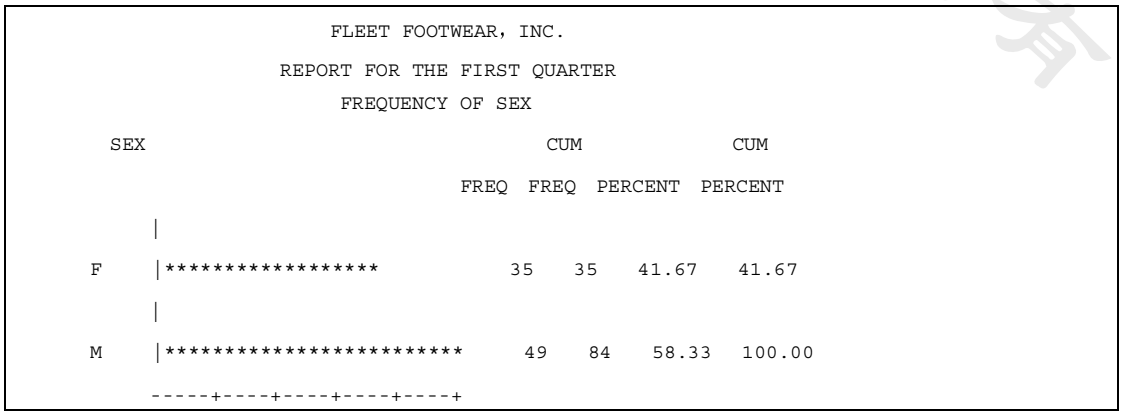


上图横轴的两个区间代表两种性别：女 (F) 及男 (M)。而纵轴则代表实际的人数。所以，此图表显示这个公司有三十五位女士，四十九位男士。

若惯于审视横的图形，则可利用 HBAR 指令，将上面的程序修改成：

```
OPTIONS PAGESIZE=25;
PROC CHART;
     HBAR SEX;
     TITLE 'FLEET FOOTWEAR, INC.';
     TITLE1 'REPORT FOR THE FIRST QUARTER';
```

结果如下图所示：



```
              10   20   30   40   50
                          FREQUENCY
```

例 2　百分比的纵轴图

如果要想以百分比来表示前述的资料，则可利用下列程序：

```
OPTIONS PAGESIZE=25;
PROC CHART;
     VBAR SEX / TYPE = PERCENT;
TITLE 'FLEET FOOTWEAR, INC.';
TITLE1 'REPORT FOR THE FIRST QUARTER';
```

其结果如下：

```
                    FLEET FOOTWEAR, INC.
                  REPORT FOR THE FIRST QUARTER
                       PERCENTAGE OF SEX
              PERCENTAGE
              60 +                 *****
                 |                 *****
              45 +        *****    *****
                 |        *****    *****
              30 +        *****    *****
                 |        *****    *****
              15 +        *****    *****
                 |        *****    *****
                 ---------------------------------
                           F         M


                             SEX
```

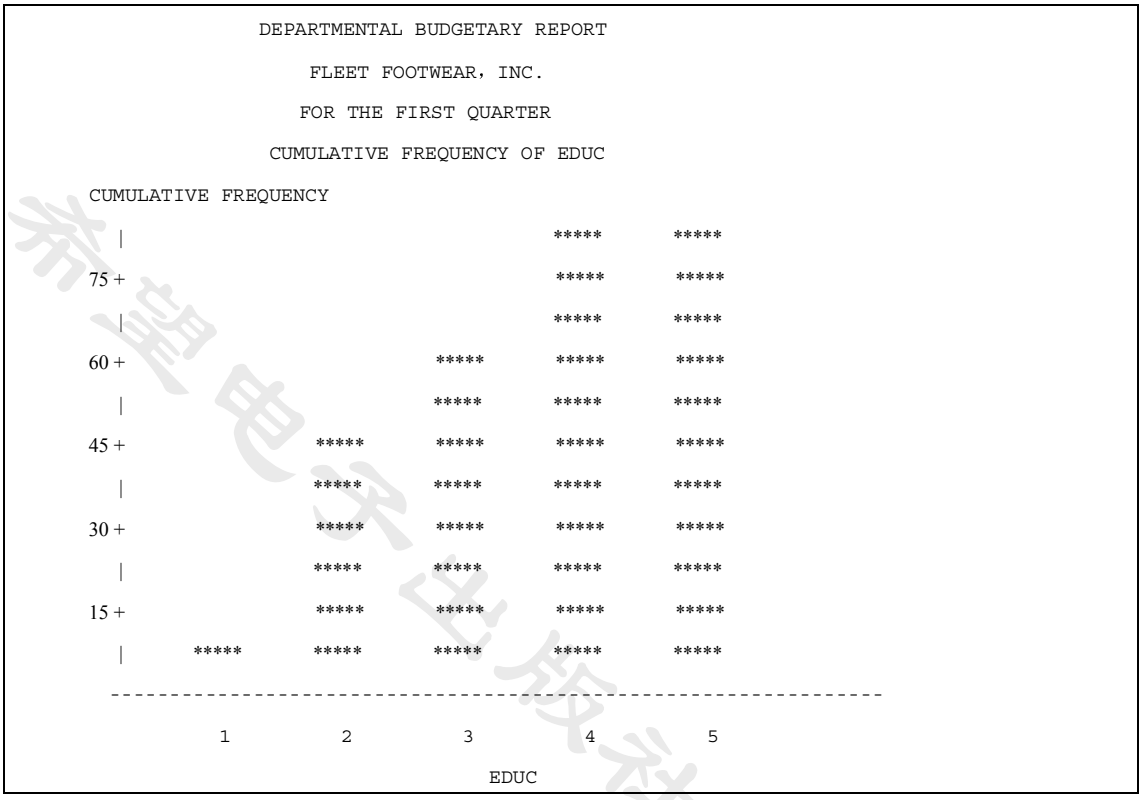
例 3　累积次数的纵轴图

累积次数图中，每一区间的次数是其本身的次数，加上所有在这个区间前的次数的和。当选用 TYPE = CFREQ 选项时，SAS 以累积次数来制图。在本例中，变量名称 EDUC 表示公司员工的教育程度。

```
OPTIONS PAGESIZE=25;
PROC CHART;
     VBAR EDUC / TYPE = CFREQ DISCRETE;
TITLE 'FLEET FOOTWEAR, INC.';
TITLE1 'DEPARTMENTAL BUDGETARY REPORT';
```

```
TITLE2 'FOR THE FIRST QUARTER';
```

结果如下图所示：



如上图所示，该公司员工中有六十三人具高中 (即 HS GRAD) 或高中以下的学历。

例 4　累积百分比的纵轴图

这一类图形上，每一分数区间的百分比，是其本身的百分比加上所有在这个区间前的百分比的和。选项 TYPE=CPERCENT 要求 SAS 以累积百分比来制图。现将同一公司员工的学历制成累积百分比的纵轴图：

```
OPTIONS PAGESIZE=25;
PROC CHART;
     VBAR EDUC /TYPE = CPERCENT DISCRETE;
TITLE 'FLEET FOOTWEAR, INC.';
TITLE1 'DEPARTMENTAL BUDGETARY REPORT';
TITLE2 'FOR THE FIRST QUARTER';
```

其结果如下：

```
                              FLEET FOOTWEAR, INC.
                          DEPARTMENTAL BUDGETARY REPORT
                             FOR THE FIRST QUARTER
                          CUMULATIVE PERCENTAGE OF EDUC


      CUMULATIVE PERCENTAGE

      100 +                                    *****      *****
          |                                    *****      *****
       80 +                          *****     *****      *****
          |                          *****     *****      *****
       60 +               *****      *****     *****      *****
          |               *****      *****     *****      *****
       40 +               *****      *****     *****      *****
          |               *****      *****     *****      *****
       20 +               *****      *****     *****      *****
          |      *****     *****      *****     *****      *****
          ---------------------------------------------------------
                    1          2          3          4          5
                                        EDUC
```

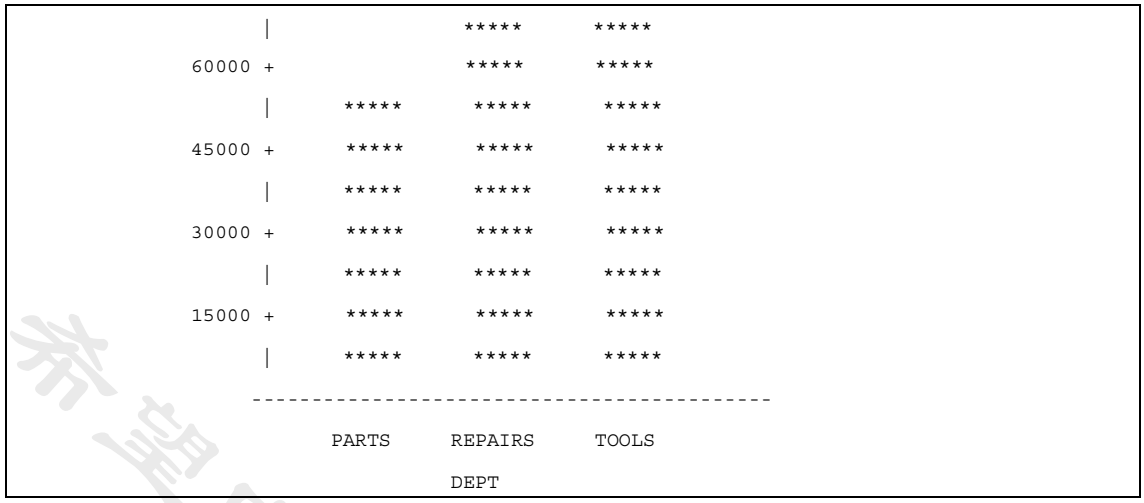
例 5　总和的纵轴图

此类图形显示横轴变量所形成的分组在纵轴变量上的总和 (利用选项 SUMVAR=)。一个公司内，三个部门 (DEPT) 的销售 (SALES) 业绩，可由下图表示：

```
OPTIONS PAGESIZE=25;
PROC CHART;
     VBAR DEPT / SUMVAR = SALES;
TITLE 'DEPARTMENT SALES FOR THE YEARS 1982-1984';
```

其结果如下：

```
                DEPARTMENT SALES FOR THE YEARS 1982-1984
                       SUM OF SALES BY DEPT
            SALES SUM
                   |                *****
             90000 +                *****
                   |                *****
             75000 +                *****
```

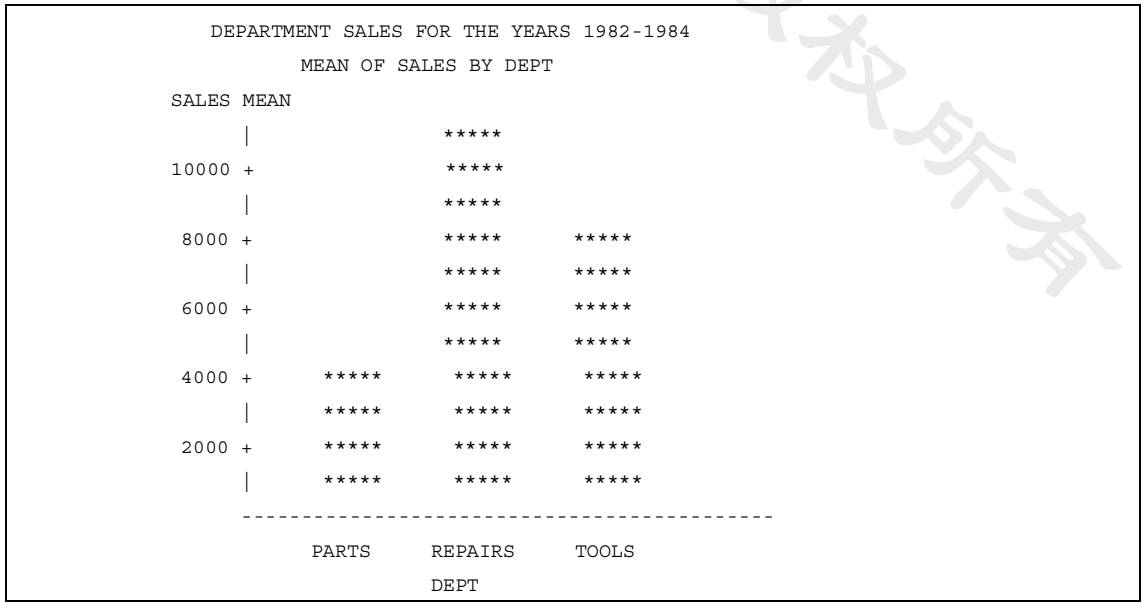


例 6　以第二变量（如：SALES）的平均值绘制纵轴图

若输入资料文件内，销售业绩已经是总额 (换句话说，资料文件内三个部门在变量 SALES 上的值已是销售总额) 且利用上述 SAS 指令制图，则所得的图形之纵轴，将以 "SALES VALUE" 做标号 (而不再是 "SALES SUM")。若想用第二变量的平均值制图，则可选用 TYPE=MEAN 选项。如下图的纵轴 (SALES MEAN) 就表示每一次交易的平均金额。

```
OPTIONS PAGESIZE=25;
PROC CHART;
     VBAR DEPT / TYPE = MEAN SUMVAR = SALES;
TITLE 'DEPARTMENT SALES FOR THE YEARS 1982-1984';
```

其结果如下：



例 7　纵轴图的符号化

若希望在纵轴图上再引进一个新变量，则可用 SUBGROUP= 选项来符号化图形。比方说，要在图形上表示某公司三个部门内男女职员的人数，则可用下列的指令：

```
OPTIONS PAGESIZE=30;
PROC CHART;
     VBAR DEPT / SUBGROUP = SEX;
TITLE 'DEPARTMENT SALES FOR THE YEARS 1982-1984';
```

产生如下的图形：

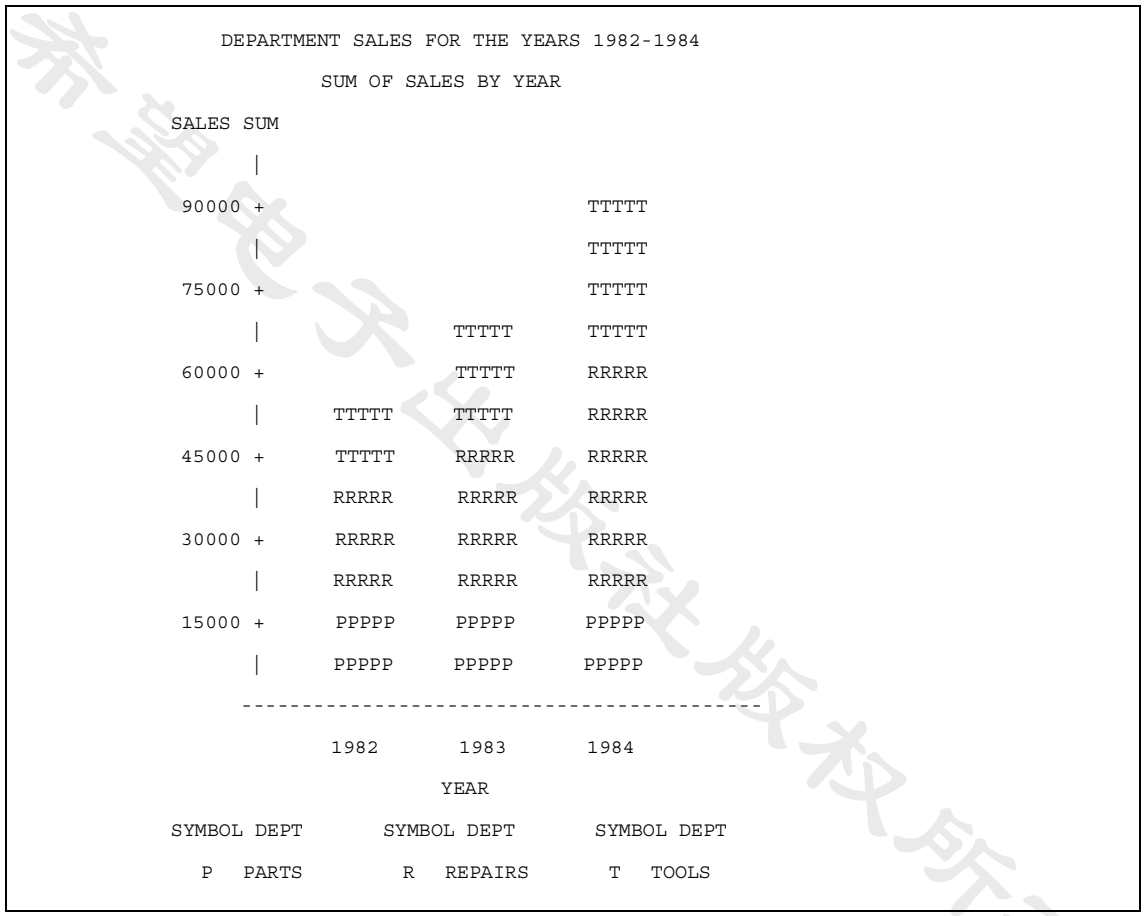
```
          DEPARTMENT SALES FOR THE YEARS 1982-1984
                  FREQUENCY OF DEPT
     FREQUENCY
     12 +      MMMMM
        |      MMMMM
     10 +      MMMMM
        |      MMMMM          MMMMM          MMMMM
      8 +      MMMMM          MMMMM          MMMMM
        |      MMMMM          MMMMM          MMMMM
      6 +      MMMMM          FFFFF          FFFFF
        |      MMMMM          FFFFF          FFFFF
      4 +      MMMMM          FFFFF          FFFFF
        |      FFFFF          FFFFF          FFFFF
      2 +      FFFFF          FFFFF          FFFFF
        |      FFFFF          FFFFF          FFFFF
        ------------------------------------------
             PARTS          REPAIRS          TOOLS
                         DEPT
               SYMBOL SEX     SYMBOL SEX
                  F   F          M   M
```

另外，也可以用符号图形表示三个变量的关系。请见下例。

三个变量分别是年度 (YEAR)，部门 (DEPT) 及销售额 (SALES)。制图的目的是为了表示 1982 到 1984 年间，一家公司内三个部门各年度的销售额，则可利用下页的 SAS 指令(DISCRETE 指令将数值变量视为类别变量)：

```
OPTIONS PAGESIZE=30;
PROC CHART;
     VBAR YEAR / SUBGROUP = DEPT
     SUMVAR = SALES DISCRETE;
TITLE 'DEPARTMENT SALES FOR THE YEARS 1982-1984';
```

获得如下的图形：

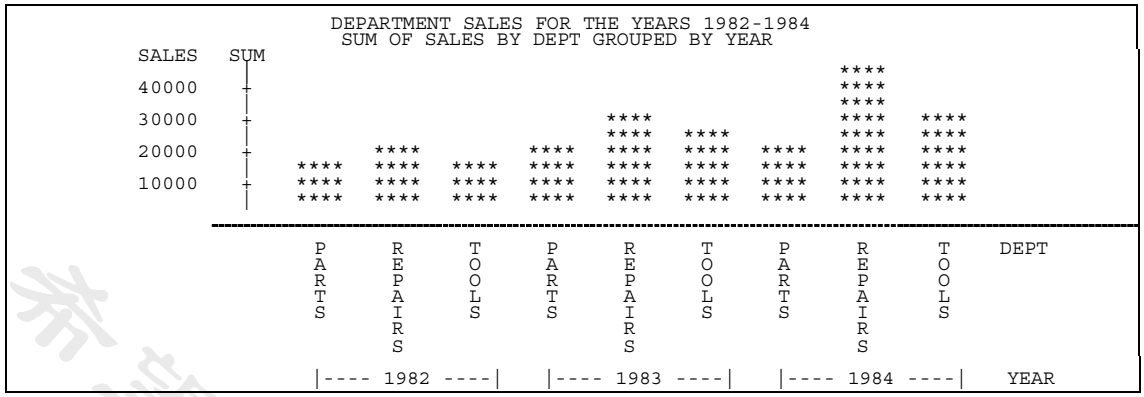


例 8　并列的图形表示法

若想比较过去三年来三个公司部门的销售成绩，则可考虑使用并列的表示法，以便比较。指令与图表如下所示：

```
OPTIONS PAGESIZE=30;
PROC CHART;
     VBAR DEPT / SUMVAR = SALES GROUP = YEAR;
TITLE 'DEPARTMENT SALES FOR THE YEARS 1982-1984';
```

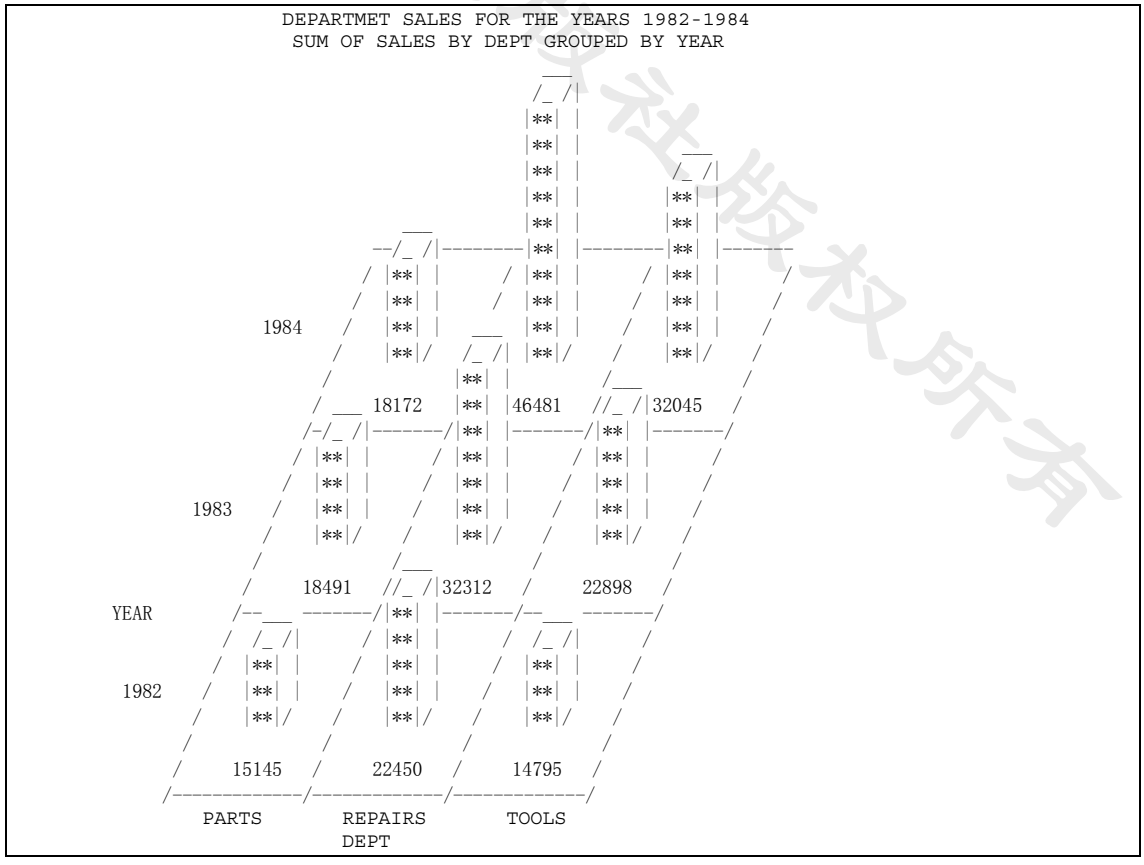
其结果如下：



例 9　方形图

方形图的外观，酷似大城市里的高楼大厦。大厦的平底是变量的值，其高度则代表该值出现的次数或百分比。请看下面指令与图形的示范(资料仍由例 8 的公司销售成绩而来)：

```
OPTIONS PAGESIZE=60;
PROC CHART;
     BLOCK DEPT/SUMVAR = SALES
                GROUP = YEAR DISCRETE;
TITLE 'DEPARTMENT SALES FOR THE YEARS 1982-1984';
```



说　明

零件部门 (PARTS) 在 1984 年的销售业绩是 18172。比较 (8) 与 (9) 这两个图形，它们都代表同样的资料数据，但看起来的视觉效果却不相同。

例 10　圆形图

圆形图利用圆的面积来表示次数或百分比。例如，利用圆的面积来比较例 8 公司的三个部门在 1982 到 1984 年间的销售总额。

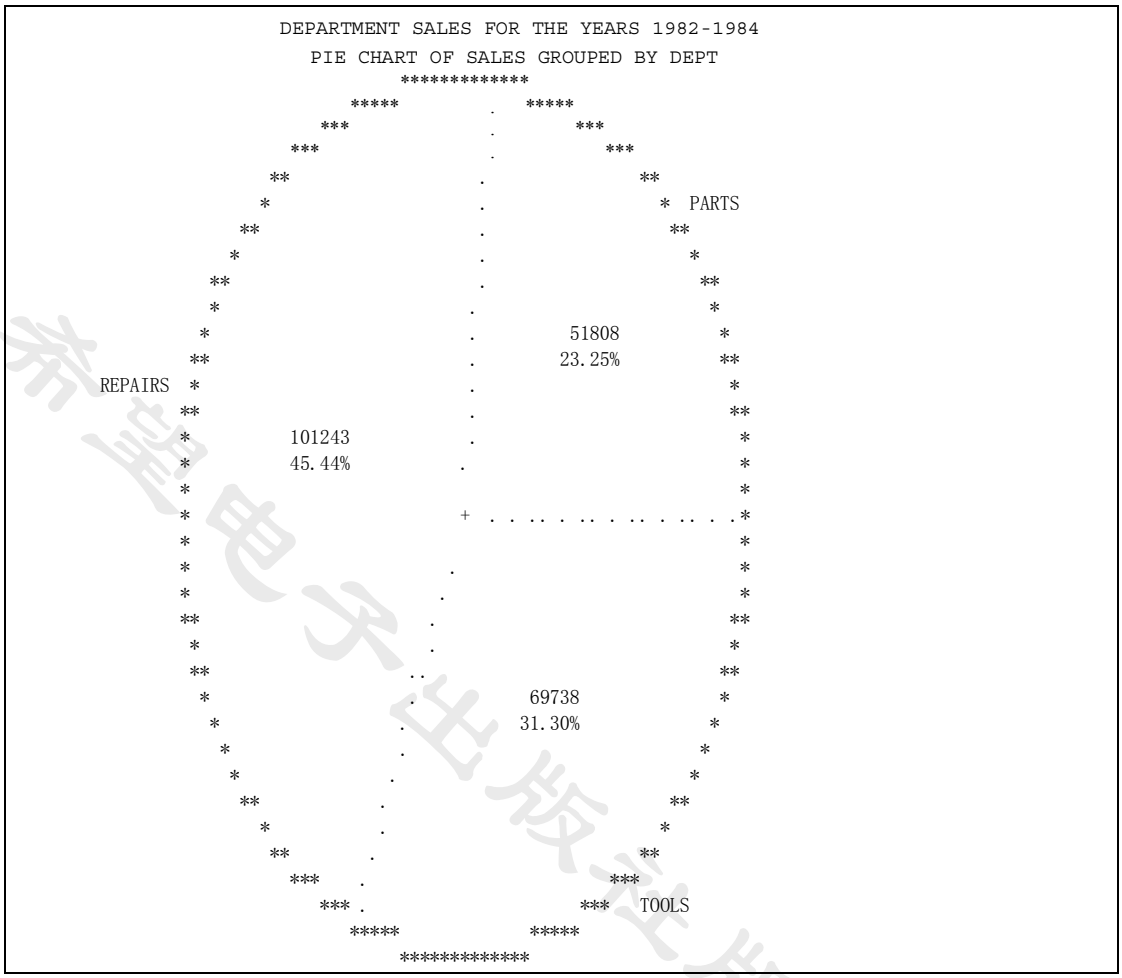
下面是该公司的销售资料：

DEPARTMENT SALES FOR THE YEARS 1982-1984

OBS	DEPT	YEAR	SALES
1	PARTS	1982	3500
2	PARTS	1983	2500
3	PARTS	1984	800
4	PARTS	1982	3651
5	PARTS	1983	5391
6	PARTS	1984	4500
7	PARTS	1982	2644
8	PARTS	1983	3500
9	PARTS	1984	3000
10	TOOLS	1982	5672
11	TOOLS	1983	6100
12	TOOLS	1984	7400
13	TOOLS	1982	1253
14	TOOLS	1983	4698
15	TOOLS	1984	9345
16	REPAIRS	1982	9050
17	REPAIRS	1983	12062
18	REPAIRS	1984	15931
19	REPAIRS	1982	8700
20	REPAIRS	1983	10310
21	REPAIRS	1984	14320
22	REPAIRS	1982	4700
23	REPAIRS	1983	9940
24	REPAIRS	1984	16230
25	PARTS	1982	5350
26	PARTS	1983	7100
27	PARTS	1984	9872
28	TOOLS	1982	7870
29	TOOLS	1983	12100
30	TOOLS	1984	15300

利用下列两行指令，可将该公司三个部门在 1982 到 1984 年间的销售业绩画在圆形图上：

```
OPTIONS PAGESIZE=60;
PROC CHART;
    PIE DEPT / SUMVAR=SALES;
    TITLE 'DEPARTMENT SALES FOR THE YEARS 1982-1984';
    TITLE1 'PIE CHART OF SALES GROUPED BY DEPT';
```



例 11　星形图

星形图最适于循环性的资料。如：一年十二个月或一天廿四小时内的变化。本例的数据是圣路易城市一年十二个月的均温。

```
OPTIONS PAGESIZE=60;
PROC FORMAT;
     VALUE MON 1='JAN' 2='FEB' 3='MAR' 4='APR' 5='MAY' 6='JUN'
               7='JUL' 8='AUG' 9='SEP' 10='OCT' 11='NOV' 12='DEC';
DATA MONTHLY;
     TITLE 'ST.LOUIS MONTHLY MEAN TEMPERATURES';
     DO MONTH=1 TO 12;
       INPUT TEMP @@;
       OUTPUT;
     END;
     FORMAT MONTH MON.;
```

```
    CARDS;
31.3 35.1
43.3 56.5
65.8 74.9
78.6 77.2
69.6 59.1
45.0 34.6
;
 PROC CHART;
    STAR MONTH / SUMVAR=TEMP DISCRETE;
 RUN;
```

图形的显示如下：

```
                ST.LOUIS MONTHLY MEAN TEMPERATURES

 CENTER = 0              SUM OF TEMP BY MONTH     OUTSIDE = 78.6

                          APR
                          56.5
              MAY         ************        MAR
               65.8   *****              *****   43.3
                   ***                       ***
                 ***                           ***
                **                               **
               *     .*...........*.               *
              **    ..  .         . .              **
             *   ....    .        .  ...             *
        JUN  **  ..        .      .    .             ** FEB
        74.9 * *.           .      .     .*            * 35.1
             * .. ...         .     .     ...           *
            ** .    ...        .    .    .. ..          **
            * .        ..       .   .    ..  ..         *
           ** .         ....    .   .   ..    ..        **
           * .             ..    .  .  ..   .. *        *
           *.                ...  . . ......   .        *
           ..                 ........          .       *
    JUL    *.....................    .........*           * JAN
    78.6   ..                  ........      .           *  31.3
           *.                ...  . . . ....   .         *
           *..             ...    .  .  .    .. .        *
           **.           ...      .   .   .     .*       **
            *..        ...       .    .   .    ..       *
            **.     ...          .    .    .  ..        **
             *.. ...             .    .     ...         *
              **.                .      .     .*        *
        AUG   * ..          .       .     ..           *  DEC
        77.2   *  ..       .        .    .             *  34.6
               **  ..     .         .  ..              **
                 *   ..  .          ...                *
                  **   .*...........*.                  **
                    ***                         ***
                       ***                   ***
                    SEP     *****          *****   NOV
                    69.6         ************        45
                                 OCT
                                 59.1
```

注：上图中的直线条，并不是 SAS 输出的一部分，而是编者自行附加上去的。这些线条的长短直接代表十二个月的气温高低。

5.3　如何撰写 PROC　CHART 程序

PROC CHART 含七道指令，它们的格式如下：

```
PROC CHART  选项串;
    BY      变量名称串;
    VBAR    变量名称串/选项串;
    HBAR    变量名称串/选项串;
    BLOCK   变量名称串/选项串;
    PIE     变量名称串/选项串;
    STAR    变量名称串/选项串;
```

在 PROC CHART 指令之后，读者可以依据自己的需要，同时选用多个图形指令。

指令 #1　PROC CHART 选项串;

此处有两个选项：

(1) DATA= 输入资料文件名称
指明到底根据那一个资料文件制图。若省略此选项，则 SAS 自动找出在此程序之前最后形成的 SAS 资料文件，并为其制图。

(2) LPI= P
P 值会决定圆形图与星形图的大小比例，内设值是 6。若自己需要调整 P 值，则其计算公式如下：
P = (打印机在一英吋空间内所印的列数/打印机在一英吋空间内所印的行数)*10
比方说：打印机在一英吋空间内可印八横列，十二纵行，则 P 值为 (8/12)*10=6.6667。

指令 #2　BY 变量名称串;

CHART 程序依据此指令所列举的变量，将观察体予以分组。然后对每一个小组分别制图。选用此指令前，资料文件内的数据，必须先按照 BY 变量的值由小到大重新排列，这个步骤可藉 PROC SORT 达成。下节将集中介绍指令 #3 到指令 #7 的变量名称串。

指令 #3-#7 制图形指令的变量名称串;

指示 CHART 程序为每一个列举的变量制图。图形指令中所列举的变量数目就是图形的个数。例如下面的界定：

```
PROC CHART;
     VBAR A B C;
```

将得到三个分别在 A，B，C 变量上的纵轴图。界定时变量之间以一个空格分开。

指令 #3-#7 制图形指令的选项串；

并不是所有的选项都适用于这五个图形指令 (即：VBAR，HBAR，BLOCK，PIE 及 STAR)。下面数据选项与图形的联用情形，将选项归类讨论：

下面七个选项可以和所有五个图形指令联用：

(1) MISSING

要求 CHART 程序将遗漏数据集合成一个组别而且绘入图形中。

(2) DISCRETE

若选用此选项，则 CHART 程序将资料文件内所有的数值变量视为类别变量。若省略此选项，则 CHART 程序就会视资料文件内的数值变量为连续变量。当 CHART 程序在处理连续变量时，若不指定区间的中点 (用 MIDPOINT= 选项指定)，或区间的个数(用 LEVELS=选项指定)，则 CHART 程序将自行以内设方式处理。

(3) TYPE = FREQ

TYPE = PERCENT (或 PCT)

TYPE = CFREQ

TYPE = CPERCENT (或 CPCT)

TYPE = SUM

TYPE = MEAN

此选项指示 CHART 程序用何种统计值来制图。FREQ 是次数，PERCENT 是百分比， CFREQ 是累积次数， CPERCENT 是累积百分比， SUM 是总和， 而 MEAN 是平均数。一般而言，内设值是 TYPE=FREQ。但若同时选用 SUMVAR= 选项，则选项的内设值是 TYPE=SUM。

(4) SUMVAR = 变量名称

要求 CHART 程序算出观察体在此一变量上的总和、平均数或次数。比方说，下面这个指令：

```
VBAR LOCATION / TYPE=MEAN SUMVAR = YIELD;
```

要求 CHART 程序算出 LOCATION 各组在 YIELD 变量上的平均数，然后以这些平均数对 LOCATION 各组制图。若选用 SUMVAR= 选项，同时也界定 TYPE=FREQ，PERCENT，CFREQ 或 CPERCENT，则 TYPE= 的界定将被视为无效，CHART 程序会自动以 TYPE=SUM 来加以处理。

(5) MIDPOINTS = 各分数区间的中点数

中点数的界定遵循下列四种模式：

SAS 将视此区间为对数的区间。

一、由小到大列举出所有的中点数，而且其间隔相等，如：

```
    MIDPOINTS = 10 20 30 40 50;
```

二、由小到大列举出所有的中点数，但其间隔不等，如：

```
MIDPOINTS = 10 100 1000 10000;
```

三、列出最小和最大的中点数值，及每两中点间的差距。如第一例可写成：

```
MIDPOINTS = 10 TO 50 BY 10;
```

四、文字变量也可用来为各分数区间命名。如：

```
MIDPOINTS = 'JAN' 'FEB' 'MAR';
```

若省略此选项，则图形上的数据自动按英文字母的顺序或由小到大的数字次序呈现。

(6) FREQ = 变量名称

此变量的值代表各观察体重复被使用的次数 (或加权比重)。 一般而言，这个变量的值应是整数。若 FREQ 变量的值出现小数的情况，则 CHART 程序将只取其整数值 (如：4.7 将被视为 4)。若 FREQ 变量的值是一个负数或遗漏数据时，则 CHART 程序将视之为 0。此选项与 SUMVAR= 选项合用时，所得的总和会自动再乘以 FREQ 变量的值。

(7) AXIS = 最小值，最大值　或

AXIS = 最大值

这个选项的值可用来控制纵轴的长短；若只界定一个值时，代表最大值。最小值则由 SAS 内设为 0。当这个选项与星形图联用时，最小值指的是星形图的中心点，而星形图的半径就是最大值与最小值的差。

下面七个选项只可以和 VBAR，HBAR 及 BLOCK 图形指令联用：

(1) GROUP = 变量名称串

指示 CHART 程序制作并列型的图表，如：

```
VBAR SEX / GROUP = DEPT;
```

上面的界定，使各部门内的男女人数分别绘制于图形内。

(2) SUBGROUP = 变量名称

要求 CHART 程序在图形上再标示 SUBGROUP 的次数。如：

```
VBAR DEPT / SUBGROUP = SEX;
```

此纵轴图是以部门 (DEPT) 的人数为主。但各部门人数的图表上，再以 M 表示男，F 表示女。M 及 F 是 SEX 变量的值。SUBGROUP 变量取其组名的第一个字母为代号，如：M 表示 MALE，F 表 FEMALE。若有两组的组名都以同一字母开头(如：America 和 Africa)，则另以 A，B，C 等做为代号。这些代号所代表的组名会在图的上端印出。SUBGROUP 变量下的遗漏数据自成一组，也将在图表上显示。

(3) LEVELS = 分数区间的个数

此选项须与连续的数值变量联用，以订出区间的个数。

(4) SYMBOL = 图形上的符号

此选项不可与 SUBGROUP = 选项联用。内设值是星号 (*)。读者用此选项来规定一个图形上的符号。如 SYMBOL = "A"，则图形上各组的长宽皆由 A 字构成。

若报表的印刷机有重叠印刷的功能，则读者可选择两个到三个符号，如：SYMBOL = "AOX" 则 AOX 会重叠印出，形成一特殊的符号。

(5) NOSYMBOL
要求 SAS 不印出纵轴图或横轴图下端有关代号的说明。

(6) NOZEROS
若一区间不包含任何观察体，则报表上不印出该区间。

(7) G100
与 GROUP= 选项联用，使次数或百分比在每分组内的总和等于 100 或 100%。

下面三个选项只可以和 VBAR，HBAR 图形指令联用：

(1) ASCENDING
按各区间所含次数的多少，由小而大，决定它们在坐标轴上的顺序。

(2) DESCENDING
与上述选项恰好相反。区间的顺序是按其次数由大到小排列在坐标轴上。

(3) REF = 一数值
要求在图表上划出一条参考线。这个数值须与 TYPE= 选项中的统计值种类相对应。如：当 TYPE = PERCENT 时，REF 的值应是一个介于 1 与 100 之间的百分比。

下面七个选项只适用于 HBAR 图形指令：

(1) NOSTAT
不印出任何描述性的统计数值。

(2) FREQ
印出各区间的次数或加权次数。

(3) CFREQ
印出各区间的累积次数。

(4) PERCENT
印出各区间的百分比。

(5) CPERCENT
印出各区间的累积百分比。

(6) SUM
印出各区间的总观察数。

(7) MEAN
印出各区间的平均数。

注：在横轴图的制图过程中，若没有选用 SUMVAR = 选项，则 CHART 程序自动印出各区间的 (累积) 次数及 (累积) 百分比。若选用 SUMVAR = 选项及 TYPE = MEAN，则自动印出各区间的次数与平均数。若选用 SUMVAR = 选项及 TYPE = SUM，则自动印出各区间的次数及总和。

下面这个选项只可以和 VBAR 图型指令联用：

(1) NOSPACE
使纵轴图上各区间之间不留空隙。若使用此选项后仍无法将整个纵轴图浓缩在报

表上，CHART 程序会自动将纵轴图改为横轴图输出。

总整理　制图形指令的选项名称串

注：若想在图形指令中选用任何选项，则必须以一删除号 (/) 分开变量名称及选项名称。

表 5.1　PROC CHART 图形指令的总整理

	VBAR	HBAR	BLOCK	PIE	STAR
●	MISSING	MISSING	MISSING	MISSING	MISSING
●	DISCRETE	DISCRETE	DISCRETE	DISCRETE	DISCRETE
●	TYPE=	TYPE=	TYPE=	TYPE=	TYPE=
●	SUMVAR=	SUMVAR=	SUMVAR=	SUMVAR=	SUMVAR=
●	MIDPOINTS=	MIDPOINTS=	MIDPOINTS=	MIDPOINTS=	MIDPOINTS=
●	FREQ=	FREQ=	FREQ=	FREQ=	FREQ=
●	AXIS=	AXIS=	AXIS=	AXIS=	AXIS=
●	GROUP=	GROUP=	GROUP=		
●	SUBGROUP=	SUBGROUP=	SUBGROUP=		
●	LEVELS=	LEVELS=	LEVELS=		
●	SYMBOL=	SYMBOL=	SYMBOL=		
●	NOSYMBOL	NOSYMBOL	NOSYMBOL		
●	NOZEROS	NOZEROS	NOZEROS		
●	G100	G100	G100		
●	ASCENDING	ASCENDING			
●	DESCENDING	DESCENDING			
●	REF=	REF=			
●	NOSPACE	NOSTAT			
		FREQ			
		CFREQ			
		PERCENT			
		CPERCENT			
		SUM			
		MEAN			

第 6 章　统计表格的制作：统计程序 PROC TABULATE

6.1　PROC TABULATE 程序概述

本程序旨在制作各式的统计表格 (而非统计的图形)。这些表格至多可分为三个向度，即表格的行 (Column)、列 (Row) 与页 (Page)。通常这三个向度由三个文字 (或数值) 变量来定义。表格内的统计值与其他程序如：PROC MEANS，FREQ，SUMMARY 所产生的统计值大同小异。所不同于这三种程序的是：PROC TABULATE 能制作出更美观的统计表格，指令的撰写最富弹性，所制作出来的表格易于命名与修饰。在本章下一节里，我们将举几个 TABULATE 程序的指令以及报表打印出来的表格作为例子。

6.2　举 例 说 明

本节所举的例子来自同一个资料文件。在这个资料文件里，所要分析的数据是几个城市的人口数(以 POP 表示) 、城市的大小 (以 CITYSIZE 表示)、与城市所属的区域 (以 REGION 表示)。城市的大小有三，即：大 (=L)、中 (=M) 和小 (=S)。城市所属的区域则分成四类，即：北中区 (=NC)、东区 (=NE)、南区 (=SO) 以及西区 (=WE)。

从下面我们举几个例子分别示范 TABULATE 程序的指令撰写。

例 1　各区域的人口分布

下面的指令要求 SAS 以长形表格列出四个不同区域的人口数，四个区域的代号是 NC，NE，SO 与 WE。

```
PROC TABULATE;
     CLASS REGION;
     VAR POP;
     TABLE REGION, POP*SUM;
```

这个分析的结果如下：

	POP
	SUM
REGION	
NC	4650000.00
NE	6666000.00
SO	6864000.00
WE	8376000.00

例 2　各地理区域内城市大小与人口分布的关系

将上面的分析作更详细的处理，以便考虑城市的大小问题，则指令必须更改如下：

```
PROC TABULATE;
     CLASS REGION CITYSIZE;
     VAR POP;
     TABLE REGION, CITYSIZE*POP*SUM;
```

其报表的结果如下：

	CITYSIZE		
	L	M	S
	POP	POP	POP
	SUM	SUM	SUM
REGION			
NC	3750000.00	750000.00	150000.00
NE	5022000.00	1422000.00	222000.00
SO	4488000.00	2088000.00	288000.00
WE	5592000.00	2592000.00	192000.00

例 3

前页的报表也可用一个长形的表格来表示，其指令如下：

```
TABLE REGION*CITYSIZE, POP*SUM;
```

报表的结果如下：

		POP
		SUM
REGION	CITYSIZE	
NC	L	3750000.00
	M	750000.00
	S	150000.00
NE	CITYSIZE	
	L	5022000.00
	M	1422000.00
	S	222000.00
SO	CITYSIZE	
	L	4488000.00
	M	2088000.00
	S	288000.00
WE	CITYSIZE	
	L	5592000.00
	M	2592000.00
	S	192000.00

例 4

若读者想列出各组的平均数与总人口数，则指令如下：

```
TABLE REGION*CITYSIZE, POP*(SUM MEAN);
```

报表的结果如下：

		POP	
		SUM	MEAN
REGION	CITYSIZE		
NC	L	3750000.00	625000.00
	M	750000.00	125000.00
	S	150000.00	25000.00
NE	CITYSIZE		
	L	5022000.00	837000.00
	M	1422000.00	237000.00
	S	222000.00	37000.00
SO	CITYSIZE		
	L	4488000.00	748000.00
	M	2088000.00	348000.00
	S	288000.00	48000.00
WE	CITYSIZE		
	L	5592000.00	932000.00
	M	2592000.00	432000.00
	S	192000.00	32000.00

6.3　表格制作的基本概念

从上面的四个举例看来，PROC TABULATE 的核心指令是 TABLE。若想有效地使用这个指令，读者必须首先了解这个指令所含的三个重要元素：

* 变量的本质 (分类的或被分析的变量)
* TABLE 指令的写法 (星号、括号、空格等的用途)
* 表格的排列 (页数、横轴与纵轴的安排)

下面分别解释这三个重要元素的意义：

* 变量的本质（分类的或被分析的变量）

分类的变量相当于变异数分析中的自变量，而被分析的变量就好比是因变量。因此，在上述的几个例子中，REGION 和 CITYSIZE 同属分类的变量，而 POP 就是被分析的变量。

* TABLE 指令的写法

对 PROC TABULATE 而言，每一个表格 (TABLE) 是许多细格 (CELLS) 的集合体。细格内的数值由下列的资料决定：

(1) 分类变量的组别 (LEVEL)
(2) 被分析变量的值
(3) 描述性统计值
(4) 表格的形式

比方说，希望在一个细格内以数字表示南方地区小城市的人口总数，则在撰写 PROC TABULATE 的指令时，读者可以将地区 (即 REGION='SO')、城市大小 (即 CITYSIZE='S') 视为分类的变量，而人口总数 (以 SUM 表示) 就是描述性统计值。

表格的形式大部分是靠标点符号如星号、括号、空格等来控制的。这些符号的运用与它们在实验设计中的使用完全相同。下面分别介绍这些符号的定义，并举例示范它们的撰写：

星号：以分类变量的所有排列组合为细格

所以，A*B 产生的组别是：

A=1			A=2		
B=1	B=2	B=3	B=1	B=2	B=3

B*A 产生的组别是：

B=1		B=2		B=3	
A=1	A=2	A=1	A=2	A=1	A=2

上述两种写法相当于实验设计里的交叉效果 (Crossed Effect)。在星号左边的变量组别算是表格的主分类 (Primary Classification)；而星号右边的变量组别算是表格的次分类 (Secondary Classification)。次分类的组别永远在主分类组别之下。

括号：要求 PROC TABULATE 先处理括号内的变量组别，

括号：再处理括号外的变量组别

所以，如果撰写：A*(B C)

则表格的形式将会是：

A=1						A=2					
B=1	B=2	B=3	C=1	C=2	C=3	B=1	B=2	B=3	C=1	C=2	C=3

从这个例子里，我们可知当星号与括号联用时，这个星号相当于四则运算的乘号；因此，它会遵循分配律的原则。

空格：分类型变量的组别并列在同一行上

所以，A　B　(B 的组别在 A 之后) 将产生如下的表格：

A=1	A=2	B=1	B=2	B=3

若将星号与空格混合使用，如：

A*B　C

则产生下面的表格：

A=1			A=2					
B=1	B=2	B=3	B=1	B=2	B=3	C=1	C=2	C=3

若将星号移到 B 与 C 之间，如：

A　B*C

则上面的表格会改变成：

		B=1			B=2			B=3		
A=1	A=2	C=1	C=2	C=3	C=1	C=2	C=3	C=1	C=2	C=3

＊ 表格的排列（页数、横轴与纵轴的安排）

页数、横轴与纵轴统称表格的三个向量 (Dimensions)，它们的别名也包括：

(1) WAFER 或 PAGE (即　页)

(2) STUB 或 SIDE　(即横轴)

(3) BANNER 或 TOP (即纵轴)

这三个向量在撰写时以逗号分开，其先后表示的顺序是：

页→横轴→纵轴

若读者在程序中只界定两个向量，则它们必须都是横轴与纵轴。若读者只界定一个向量，则它必须是纵轴。

前面 TABLE 指令所示范的例子都是以横轴的方式呈现的；若以纵轴的方式呈现，则指令 A B 会产生如下的表格：

A=1
A=2
B=1
B=2
B=3

同样地，A　B*C　的指令会产生下面的表格：

A=1	
A=2	
B=1	C=1
	C=2
	C=3
B=2	C=1
	C=2
	C=3
B=3	C=1
	C=2
	C=3

若表格的细格内，含统计值如总数 (SUM) 或平均数 (MEAN)，则读者可利用横轴向量表示各式的统计值，而以纵轴向量表示被分析的变量名称。请看下面的例子：

```
CLASS REGION CITYSIZE;
VAR POP;
TABLE REGION*CITYSIZE*(SUM MEAN), POP;
```

报表的结果如下：

			POP
REGION	CITYSIZE		
NC	L	SUM	3750000.00
		MEAN	625000.00
	M	SUM	750000.00
		MEAN	125000.00
	S	SUM	150000.00
		MEAN	25000.00
NE	CITYSIZE		
	L	SUM	5022000.00
		MEAN	837000.00
	M	SUM	1422000.00
		MEAN	237000.00
	S	SUM	222000.00
		MEAN	37000.00
SO	CITYSIZE		
	L	SUM	4488000.00
		MEAN	748000.00
	M	SUM	2088000.00
		MEAN	348000.00
	S	SUM	288000.00
		MEAN	48000.00
WE	CITYSIZE		
.			.
.			.
.			.

有时，读者或许希望按某个分类变量的组别 (如：男、女) 将制作的表格分别画在不同的页上，此时就得利用页的向量了。

在下面的例子里，我们为一种产品 (PRODUCT，下分 A100 与 A200 两种) 分别制作一个表格。每一个表格内均含该产品在各地区 (REGION)、各型城市 (CITYSIZE) 销售的量 (QUANTITY)与金额 (AMOUNT) 的总数。除此之外，销售的方式 (SALETYPE，下分 R 与 W 两种方式) 是一个主分类变量，所以它的类别将会出现在表的最上端。

程序的写法如下：

```
TABLE PRODUCT, REGION CITYSIZE, SALETYPE*(QUANTITY AMOUNT);
```

报表的结果分两页：一页是产品 A100 的销售成果表 (请看下表)，另一页则是产品 A200 的销售成果表 (请看次页的表)。

PRODUCT A100

	SALETYPE			
	R		W	
	QUANTITY	AMOUNT	QUANTITY	AMOUNT
	SUM	SUM	SUM	SUM
REGION				
NC	1250.00	31250.00	1250.00	25000.00
NE	1600.00	40000.00	1600.00	32000.00
SO	1880.00	47000.00	1880.00	37600.00
WE	1840.00	46000.00	1840.00	36800.00
CITYSIZE				
L	3190.00	79750.00	3190.00	63800.00
M	2440.00	61000.00	2440.00	48800.00
S	940.00	23500.00	940.00	18800.00

PRODUCT A200

	SALETYPE			
	R		W	
	QUANTITY	AMOUNT	QUANTITY	AMOUNT
	SUM	SUM	SUM	SUM
REGION				
NC	1295.00	32375.00	1295.00	25900.00
NE	1645.00	41125.00	1645.00	32900.00
SO	1925.00	48925.00	1925.00	38500.00
WE	1885.00	47125.00	1885.00	37700.00
CITYSIZE				
L	3250.00	81250.00	3250.00	65000.00
M	2500.00	63500.00	2500.00	50000.00
S	1000.00	24800.00	1000.00	20000.00

最后，请读者注意，当界定了被分析的变量 (如上例中的 PRODUCT) 而未提到任何统计值时，TABULATE 程序自动将总数 (SUM) 印出。然而，若已界定了表格的分类变量，而未提到任何被分析变量时，TABULATE 程序则自动印出分类变量的次数。

6.4　如何撰写 PROC TABULATE 程序

PROC TABULATE 含十道指令，它们的格式如下：

```
PROC TABULATE    选项串;
     CLASS       变量名称串;
     VAR         变量名称串;
     FREQ        变量名称;
     WEIGHT      变量名称;
     BY          变量名称串;
     FORMAT      变量名称串格式.;
     LABEL       变量=变量的标;
     TABLE       页向量的定义，横轴向量的定义，
                 纵轴向量的定义 /选项串;
     KEYLABEL    统计值之代号='代号的解释'...;
```

其中，PROC TABULATE，CLASS 与 TABLE 三道指令是必须的，不可省略。

指令 #1　PROC TABULATE 选项串;

有六个选项，分别说明如下：

(1) DATA=输入资料文件名称

指明到底为那一个资料文件的变量制作统计表格，若省略此选项，则 SAS 会自动找出在此程序之前最后形成的资料文件，为它制作表格。

(2) FORMAT=格式名称

此选项用来控制细格内统计值的有效位数。内设格式是 F12.2 (即每一统计值的表示法不超过十二位数，其中小数位数据两位，小数点占一位)。然而，若读者在 TABLE 指令中定义了其他 FORMAT 的格式，则此处的选项 FORMAT=格式的名称自动无效。

(3) ORDER=FREQ

ORDER=DATA

ORDER=INTERNAL (内设值)

ORDER=FORMATTED

界定分类变量下各类别的输出次序。当 ORDER=FREQ 时，次序先后依各类别次数多少而定。次数最多的那一个类别最先，次数第二多的那一个类别第二，依此类推。比方说，一组内有男一百二十人，女八十人，则此选项规定男生组是第一组，女生组是第二组。当 ORDER=DATA 时，类别次序就是它们在输入资料文件内出现的顺序。当 ORDER=INTERNAL 时，类别次序由英文字母先后决定。如性别的两个类别，男以 MALE 代表，女以 FEMALE 代表。此选项会定女生组为第一组，男生组为第二组。当 ORDER=FORMATTED 时，类别次序由外在格式决定。当省略此选项时，内设值是 ORDER=INTERNAL。另外，遗漏数据总是排在最前面。

(4) MISSING

要求 TABULATE 程序将含遗漏数据的观察体也包括在表格内；若省略此选项，则所有含一个 (或一个以上) 遗漏数据的观察体均将自表格中剔除。

(5) FORMCHAR (符号的位置，以 1 到 11 的数字表示) = '十一个画表格的符号'

TABULATE 利用这个指令来控制画统计表格所需的符号。这些符号最多可达十一个。下面简单地说明这些符号的位置、数字代号与内设的符号：

符号的位置	数字代号	内设的符号
纵轴	1	|
横轴	2	—
左上角	3	—
中上方	4	—
右上角	5	—
中心点的左边	6	|
中心点	7	+
中心点的右边	8	|
左下角	9	—
中下方	10	—
右下角	11	—

因此，如果读者想用星号来表示四个角时，可写

```
FORMCHAR( 3 5 9 11 ) = '****'
```

若读者将这个选项改写成：

```
FORMCHAR = '     '   (括号内含 11 个空白)
```

则报表上的表格就只含文字或统计值，不含任何的线条或符号。

若报表会在 IBM 公司的 6670 型打印机上打印，则下列 FORMCHAR 的定义最理想：

```
FORMCHAR = 'FABFACCCBCEB8FECABCBBB'X
```

若打印机上附带有 TN (文字) 的打印机 (Print Train)，则上述的定义可改成：

```
FORMCHAR = '4FBFACBFBC4F8F4FABBFBB'X
```

(6) DEPTH=介于 1 与 10 的整数

此选项界定表格中所接受的交叉分类型变量的个数 (亦即上限)。根据这个定义，A*B 的 DEPTH 等于 2。若读者已经将表格的一个向量撰写成

```
A* (B*N ALL)
```

则这个向量至多只可与另外三个分类变量产生排列组合，因为 A*(B*N) 本身已经包含了三个分类的变量。

指令 #2　CLASS 变量名称串；

这个指令所定义的变量串均是分类变量，在 TABLE 指令将会用到。大部分的分类变量是不连续变量，如男、女或重点、非重点学校。少数的分类变量是连续变量，如年龄或每月平均所得；在这些情况下，读者最好将这些连续变量转变成不连续变量，然后再藉 CLASS 指令宣告这些变量是分类变量。

指令 #3　VAR 变量名称串；

此指令界定被分析的变量串，这些变量串将会出现在 TABLE 指令中。另外，被分析变量必须是数值变量而非文字变量。

指令 #4　FREQ 变量名称；

FREQ 变量值代表资料文件内观察体重复出现的次数。若 FREQ 变量的值带有小数，则 TABULATE 程序只取整数的部分；若其值小于 1，则视为遗漏数据。

指令 #5　WEIGHT 变量名称；

这个变量的值代表各观察体的加权值 (一般而言，每个观察体应只代表一个数据点)。这个变量的值必须是正有理数，带小数点亦可。WEIGHT 变量的值是用来计算加权平均数、加权变异数、加权总和等统计值的。

指令 #6　BY 变量名称串；

SAS 依此指令所列举的变量将资料文件分成几个小的资料文件，然后针对每一个小的资料文件分别执行分析。当读者选用此指令时，资料文件内的数据必须先按照 BY 变量串的值重新做由小到大的排列。这个步骤可藉 PROC SORT 达成。

指令 #7　FORMAT 变量名称串格式；

指令 #8　LABEL 变量名称=变量的标；

这两个指令都是用来解释 TABLE 指令中所界定的分类变量或是被分析的变量。

FORMAT 指令专用来解释分类变量的组别，所用的方法视 FORMAT 的格式而定，如：$DOLLAR.。

LABEL 指令既可用来解释分类变量也可用来解释被分析变量的组别，如：INCOME= ' Monthly Income '。

指令 #9　TABLE 页向量的定义，横轴向量的定义，纵轴向量的定义 / 选项串；

删除号 (/) 前的定义是以逗号来分开的，它们分别与一个统计表格的页向量、横轴向量、纵轴向量的设计有关。若读者只界定两个向量的定义，则 TABULATE 程序自动以前者为横轴向量，后者为纵轴向量。若读者只定义一个向量，则它必须是纵轴向量。

不论向量的多少或先后的顺序，每一个向量的定义都可能牵涉到下面几个符号，兹将它们的意义简述如下：

向量的定义符号		符号的意义
逗号	[,]	分开相邻的两个向量定义
星号	[*]	分类变量的各类排列组合或次分组 (Subgroup)
空白	[　]	将各分类变量的组别并列
括号	[()]	界定主、次分类的打印
不等号	[<>]	界定分母项
等号	[=]	解释变量、统计值或 FORMAT 指令

下面举两个有关 TABLE 指令的例子：

例 1

```
PROC TABULATE DATA=EXAMPLE;
     CLASS SCHOOL;
     VAR STUDENT;
     TABLE SCHOOL, STUDENT*SUM;
```

根据上述程序，报表上会呈现出一个含横轴与纵轴的表格：横轴是各学校 (SCHOOL) 的名称，纵轴则是各学校学生人口的总和 (SUM)。 在此，星号 (*) 解释成次分组 (Subgroup) 而非分类变量的各种排列组合。

例 2

```
PROC TABULATE;
     CLASS REGION CITYSIZE;
     VAR POP;
     TABLE REGION*CITYSIZE*(SUM MEAN), POP;
```

报表上的表格仍然而只有两个向量 (即横轴与纵轴)。横轴的 DEPTH=3，因为它包含了 REGION，CITYSIZE 以及 (SUM MEAN) 并列的交叉排列组合。根据范例 2 的程序所得到的表格请见本章第 6.3 节。

TABULATE 程序所能计算的统计值有下面几种：

统计之代号	代号的定义
N	各细格内的有效观察体个数
NMISS	各细格内的无效观察体个数，亦即在分类变量上有遗漏数据的观察体个数
MEAN	平均数
STD	标差
MIN	最小值
MAX	最大值
RANGE	最大值与最小值的差距，又称全距
SUM	总和
USS	未矫正过的平方总和

CSS	矫正过的离差平方总和
TDERR	平均数的标误
CV	变异数(Coefficient of Variation)
T	用来检定母群之平均数是否等于 0 的单尾 t 检定
PRT	上述单尾 t 检定的统计显著程度
VAR	变量的变异数
SUMWGT	加权值的总和
PCTN	次数 (N) 的百分比
PCTSUM	总和 (SUM) 的百分比

若读者在 TABLE 指令中未界定任何被分析的变量，则可要求 TABULATE 程序计算分类变量的次数 (N) 或次数的百分比 (PCTN)。若读者在 TABLE 指令中已界定被分析的变量但没有要求任何统计值，则 TABULATE 程序会自动计算被分析变量的总和。最后，若读者在 TABLE 指令中既未提及任何被分析的变量，也未要求任何统计值，则 TABULATE 程序只在图表上印出分类变量下各细格的次数而已。

虽然而每一个 TABLE 指令只能规划一个统计的表格，然而，读者可在 TABULATE 程序中撰写不只一个 TABLE 的指令。

删除号 (/) 之后的选项有六个，分述如下：

(1) MISSTEXT='二十个字以内的字符串'

单引号内的字符串会出现在所有含遗漏数据的细格内，其目的在于解释或注明这些遗漏数据。字符串的总字数必须在二十个以内。

(2) FUZZ=极小的正整数

FUZZ 的值是用来鉴定被分析的变量值或统计值是否有效。比方说，读者设 FUZZ=.0001，则当 X 变量值或统计值的绝对值小于 .0001 时，SAS 自动将这个值视为 0。FUZZ 的内设值依电脑的机型而定，一般而言，它等于电脑的硬件系统可表示的最小正实数。

(3) RTSPACE(或 RTS)=正整数 n

代表列标题的打印长度。打印长度也包括了左右两边的边际。一般而言，这个选项的内设值等于每一横列长度的四分之一。每一横列的长度由指令 OPTIONS LINESIZE= 来控制。

(4) BOX=_PAGE_

BOX=变量名称或变量标签

BOX='字符串'

这个选项是用来为统计表格左上角的空格补白的。

当 BOX=_PAGE_ 时，空格内会印上页向量的标题。

当 BOX= 变量名称或变量标签时，左上角的空格即以此名称或标签补白。

当 BOX='字符串' 时，则单引号内的字符串会出现在左上角的空格内。无论用上述那一种方式补白，只要空格内的宽度够宽，则所有的字都可纳入补白。否则，TABULATE 程序自动按照空格的宽度将字符串对齐，请读者注意这个问题。

(5) ROW=CONSTANT(或 CONST)

ROW=FLOAT

界定标题中的空白部分在列交叉 (Row Crossing) 的表格设计下是否仍保留其空白的部分。此指令的内设值等于 CONSTANT (或 CONST)，即空白的部分照样打印在所有的标题上。然而，当 ROW=FLOAT 并且列标题中有空白 (如：N=' ')，则在列交叉的布局里，非空白标题部分将等分列标题所占的宽度。如此一来，空白的部分会被认为无效。

(6) CONDENSE

要求将两张 (或以上) 的统计表格打印在同一张报表纸上。只要报表纸的长度够长，TABULATE 程序会自动将几个表格同时印在同一页上以节省报表纸的空间。

指令 #10　KEYLABEL 统计值之代号='代号的解释' …;

这个指令的目的在于解释或注明代表各统计值之代号。这些代号及其解释的撰写规则如下：

代号：　必须是指令 TABLE 所提及的十八个统计值 (如：N，NMISS，...，PCTSUM 等) 或 ALL (表示所有的统计值，见第 6.5 节的说明)。

代号的解释：说明上述统计值之代号的字符串；必须在四十字以内，而且用单引号括住。

如果读者想解释一个以上的统计值之代号时，可依下列的形式来撰写：

```
KEYLABEL ALL='TOTAL $'
    MEAN='AVERAGE'
    PCTSUM='PERCENT OF SUM';
```

6.5　注 意 事 项

■ 内设的统计值

若读者只界定了被分析的变量而不提任何统计值的代号，则 TABULATE 程序会主动计算该变量的总和 (SUM)。若读者既未界定被分析的变量，亦未提及任何统计值的代号，则 TABULATE 程序仍会计算分类变量值交叉所造成的排列组合的次数 (N)。

■ 遗漏数据

遗漏数据可能出现在分类变量上或被分析变量上。当分类变量下有遗漏数据时，该观察体将从统计值的计算中剔除，除非读者在 PROC TABULATE 的指令中选用 MISSING 选项。MISSING 选项的作用是把遗漏数据归成一个单独的类别。若是造成遗漏数据的原因不同，MISSING 选项会将它们纳入不同的类别内。不论读者是否使用 MISSING 选项，遗漏数据只影响 NMISS 统计值的计算，并不影响任何其他描述性统计值的计算。当统计表格中有遗漏细格时，读者可利用 TABLE 指令中 MISSTEXT 选项

印出一个二十字以内的字符串来取代空白的细格。这种字符串的打印将有助于读者日后审视报表时比较容易辨认。

■ 列标题的宽度

列标题的宽度可藉 TABLE 指令中的 RTS 来控制。内设的宽度是一横列宽度(亦即 LINESIZE 的值) 的四分之一。这个内设的宽度一般而言已绰绰有余。

■ 特殊类别 ALL 的使用

ALL 的意义就是总结。当统计表格中含 A 与 B 两个分类变量的交叉而且 B 涵盖了 A 的组别，因而产生如下的镶嵌设计：

B=1		B=2		B=3	
A=1	A=2	A=1	A=2	A=1	A=2

则指令 B*(A ALL) 表示将自动总结每一个 B 分组 (如 B1，B2，B3) 下所有 A 分组的值。如此一来，统计表格的上标题会添加 ALL 一栏。在 ALL 栏内的值是各 B 分组下所有 A 分组数值的总结：

B=1			B=2			B=3		
A=1	A=2	ALL	A=1	A=2	ALL	A=1	A=2	ALL

若读者在指令中用到两次 ALL，如：(ALL B)*(ALL A)，则上述的标题将会改变成：

ALL			B=1			B=2			B=3		
ALL	A=1	A=2	ALL	A=1	A=2	ALL	A=1	A=2	ALL	A=1	A=2

这个标题显示：不但各 B 分组下所有 A 分组的总结会算出，而且各 A 分组下所有 B 分组的总结也将打印在 (ALL A1) 栏内或 (ALL A2) 栏内。最后，(ALL ALL) 栏则表示所有数据的总结。

根据上面简单的示范原则，让我们再回到本章第 6.2 节所举的例子：亦即三种产品 (以 A100，A200，A300 代表) 在四个地区 (以 NC，NE，SO，WE 代表) 中三个种类的城市 (以 L，M，S 代表) 销售的总数 (以 QUANTITY 表之) 或钱数 (以 SUM 表之) 的统计值。由于销售的方式 (SALETYPE) 下又分两种：R 或 W 方式，地区 (REGION) 下分四区，城市 (CITYSIZE) 下分三类别，我们必须在指令 TABLE 中引用 ALL 选项四次：其中三次与上述三个分类变量相对应，第四个 ALL 是将这三个分类变量的次数总结后再总结一次。如此，我们就可全盘了解产品销售的情形。其表格的设计指令如下：

```
TABLE ALL PRODUCT, REGION ALL CITYSIZE ALL,
   (SALETYPE ALL)*(QUANTITY*F=6. AMOUNT*F=10.2);
```

上述指令中 QUANTITY*F=6. 表示统计值 QUANTITY 以六位数字 (不含小数位数) 的 F 格式打印。同理，AMOUNT*F=10.2 表示统计值 AMOUNT 以十位数字 (内含两位小数位数以及一位小数点) 的 F 格式打印。

另外，前述的指令将造成四页 (Logical Pages) 的报表：其中第三页是三种不同产品的次总结，第四页的结果则是三种次总结的总整理。因此，报表的形式将如下所示：

ALL

	SALETYPE				ALL	
	R		W			
	QUANT-ITY	AMOUNT	QUANT-ITY	AMOUNT	QUANT-ITY	AMOUNT
	SUM	SUM	SUM	SUM	SUM	SUM
REGION						
NC	3810	95200.00	3810	76200.00	7620	171400.00
NE	4869	121725.00	4869	97380.00	9738	219105.00
SO	5706	143450.00	5706	114120.00	11412	357570.00
WE	5576	139400.00	5576	111520.00	11152	250920.00
ALL	19961	499775.00	19961	399220.00	39922	898995.00
CITYSIZE						
L	9651	241275.00	9651	193020.00	19302	434295.00
M	7400	185950.00	7400	148000.00	14800	333950.00
S	2910	72550.00	2910	58200.00	5820	130750.00
ALL	19961	499775.00	19961	399220.00	39922	898995.00

PRODUCT A100

	SALETYPE				ALL	
	R		W			
	QUANT-ITY	AMOUNT	QUANT-ITY	AMOUNT	QUANT-ITY	AMOUNT
	SUM	SUM	SUM	SUM	SUM	SUM
REGION						
NC	1250	31250.00	1250	25000.00	2500	56250.00
NE	1600	40000.00	1600	32000.00	3200	72000.00
SO	1880	47000.00	1880	37600.00	3760	84600.00
WE	1840	46000.00	1840	36800.00	3680	82800.00
ALL	6570	164250.00	6570	131400.00	13140	295650.00
CITYSIZE						
L	3190	79750.00	3190	63800.00	6380	143550.00
M	2440	61000.00	2440	48800.00	4880	109800.00
S	940	23500.00	940	18800.00	1880	42300.00
ALL	6570	164250.00	6570	131400.00	13140	295650.00

PRODUCTA200

	SALETYPE					
	R		W		ALL	
	QUANT-ITY	AMOUNT	QUANT-ITY	AMOUNT	QUANT-ITY	AMOUNT
	SUM	SUM	SUM	SUM	SUM	SUM
REGION						
NC	1295	32375.00	1295	25900.00	2590	58275.00
NE	1645	41125.00	1645	32900.00	3290	74025.00
SO	1925	48925.00	1925	38500.0	3850	87425.00
WE	1885	47125.00	1885	37700.00	3770	84825.00
ALL	6750	169550.00	6750	135000.00	13500	304550.00
CITYSIZE						
L	3250	81250.00	3250	65000.00	6500	146250.00
M	2500	63500.00	2500	50000.00	5000	113500.00
S	1000	24800.00	1000	20000.00	2000	44800.00
ALL	6750	169550.00	6750	135000.00	13500	304550.00

PRODUCTA300

	SALETYPE				ALL	
	R		W			
	QUANT-ITY	AMOUNT	QUANT-ITY	AMOUNT	QUANT-ITY	AMOUNT
	SUM	SUM	SUM	SUM	SUM	SUM
REGION						
NC	1265	31575.00	1265	25300.00	2530	56875.00
NE	1624	40600.00	1624	32480.00	3248	73080.00
SO	1901	47525.00	1901	38020.00	3802	85545.00

WE	1851	46275.00	1851	37020.00	3702	83295.00
ALL	6641	165975.00	6641	132820.00	13282	298795.00
CITYSIZE						
L	3211	80275.00	3211	64220.00	6422	144495.00
M	2460	61450.00	2460	49200.00	4920	110650.00
S	970	24250.00	970	19400.00	1940	43650.00
ALL	6641	165975.00	6641	132820.00	13282	298795.00

■ 利用 PCTN 与 PCTSUM 来计算百分比

PCTN 是次数 (N) 的百分比，而 PCTSUM 是总和 (SUM) 的百分比。次数 (N) 的定义比较宽广，泛指被分析变量的非遗漏数据或两个分类变量之排列组合的出现次数。总和 (SUM) 则只能指被分析变量的总和，与分类变量无关。

PCTN 或 PCTSUM 的撰写遵循下面的格式：

(PCTN 或 PCTSUM)<分母串>{=百分比的解释}

下页首先定义分母串的意义：

分母串的定义

分母的界定是为了让 TABULATE 指令知道是否该将 TABLE 指令中交叉的分类组别合并，或选择一个被分析变量的值为分母。若读者对前节所述的 ALL 选项有把握，则也可以轻易地掌握字母串的定义。

当 ALL 与单一的分类变量并用因而造成两个统计值时：第一个是该分类变量下各分组的统计值；第二个则是各组合并后所得到的统计值之总结。若这个统计值恰巧是 N 或 SUM，则 ALL 就代表总和。

百分比是某个总和除以适当的分母而得到的。所以分母串的定义旨在引导 TABULATE 程序将适当的分组合并，使其成为百分比的分母。因此，如果你的表格设计如下：

```
TABLE    A*(B  ALL);
```

则表格输出如下：

A					
1			2		
B			B		
1	2	ALL	1	2	ALL
N	N	N	N	N	N
10	30	40	20	40	60

则 TABLE A*B*PCTN<B>；的指令将导致 TABULATE 程序在 A1 与 A2 分组内，各以其次数总和 (即 40 与 60) 为分母以便单独求出 B1 与 B2 分组在 A1 或 A2 组内所占的百分比。根据上述的指令写法，得到下面的报表：

A			
1		2	
B		B	
1	2	1	2
PCTN	PCTN	PCTN	PCTN
25	75	33	67

若比较这两张报表，会发现分母的定义其实是由统计值 ALL 直接而的来。统计值 ALL 是针对被分析变量而言，与分类变量无关。

下面再举一例以示范如何用 X 的总和 (X SUM) 当分母以便将 Y 的总和 (Y SUM) 转变成百分比：

```
TABLE A*B, C*(X*SUM Y*(SUM PCTSUM <A*B*X>))/RTS=15;
```

从上述的指令中，我们可以得知有三种分类变量，分别是 A，B，C。A*B 的排列组合界定列向量的设计，C*(X*SUM Y*(SUM PCTSUM)) 等界定行向量的设计。至于 <A*B*X> 的撰写则要求 TABULATE 程序将 A*B 的所有分组合并以便计算 X 变量的总和 (SUM)。而这个 X 的总和也就是 Y 变量在 A*B*C 的分组下百分比 (PCTSUM) 的分母。请看下面依此 TABLE 指令所产生的统计图表：

		C					
		1			2		
		X	Y		X	Y	
		SUM	SUM	PCTSUM	SUM	SUM	PCTSUM
A	B						
1	1	59	123	57	59	100	27
	2	78	49	23	14	28	7
2	B						
	1	21	71	33	140	106	28
2	59	49	23	161	2	0	

请读者注意，当 C=1 时，Y 之 PCTSUM 的分母是 217(=59+78+21+59，X 的 SUM)。当 C=2 时，Y 之 PCTSUM 的分母等于 374(=59+14+140+161，X 的 SUM)。利用这两个不等的分母去除 Y 的 SUM 值，所得到的百分比分别列在两行 PCTSUM 内。由于这两个分母的定义与 Y 值无关，因此，PCTSUM 在各栏内相加的总和也不等于 100。

以列总和为分母的百分比

若读者有意以列总和当作列百分比的分母，则定义分母时应完全以行向量的分类变量为主。请看下面 TABLE 指令的示范：

```
TABLE A, B*(N PCTN<N>);
```

根据上述指令所设计的表格如下：

	B			
	1		2	
	N	PCTN	N	PCTN
A				
1	10	25	30	75
2	20	33	40	67

在这个表格里，当 A=1 时，PCTN 的分母是 40(=10+30)。当 A=2 时，PCTN 的分母是 60(=20+40)。因此，所求得的百分比是名正言顺的列百分比。

以行总和为分母的百分比

若读者有意以行总和当作行百分比的分母，则定义分母时应完全以列向量的分类变量为主。

请看下面 TABLE 指令的示范：

```
TABLE A, B(N PCTN<A>);
```

根据上述指令所设计的表格如下：

	B			
	1		2	
	N	PCTN	N	PCTN
A				
1	10	33	30	43
2	20	67	40	57

在这个表格里，当 B=1 时，PCTN 的分母是 30(=10+20)。当 B=2 时，PCTN 的分母是 70(=30+40)。因此，所求得的百分比是名正言顺的行百分比。

以其他总和为分母的百分比

若读者有意用页向量的总和当作百分比的分母，则定义分母时应完全以页向量所牵涉到的所有分类变量为主。下面的例子是以细格的总和作分母，以便计算百分比：

```
TABLE A, B*(N PCTN<A*B>);
```

根据上述指令所设计的表格如下：

	B			
	1		2	
	N	PCTN	N	PCTN
A				
1	10	10	30	30
2	20	20	40	40

在这个表格里，所有细格的百分比都是以 100(=10+20+30+40) 作分母的。因此，所求得的百分比都是细格百分比。

多重选择的百分比计算

在一些问卷调查 (如对顾客喜好的市场调查) 的设计里，一个问题可能有五种可选的答案，而且受试者可多重选择一到五个答案。在这种情况下，每一种答案的百分比应以受试者的人数为分母，而不应以实际圈选的总答案数为分母。让我们举一例来说明这种百分比的计算：

设 ID=受试者的识别 (学) 号

Q=问卷上的题号

V1-V5=每一题上的五种可能答案。这五个变量的值是二分的：

1 表示受试者已圈选该答案，

0 表示受试者未圈选该答案。

则下列的指令可以正确地计算出各题上 V1-V5 五种答案被受试者选中的百分比：

```
TABLE ID*N (V1 V2 V3 V4 V5 )*(N PCTN<ID>), Q;
```

若读者想更进一步说明报表上所得的统计值，不妨将上述指令改写成：

```
PROC TABULATE F=7.2;
     CLASS Q;
     VAR ID V1-V5;
     TABLE ID*N='RESPONSES'*F=7. (V1 V2 V3 V4 V5 )*(N='COUNT'*F=7.
     PCTN<ID>='PERCENT'),Q / RTS=20
```

```
LABEL ID='TOTAL'
      V1='RESPONSE 1'
      V2='RESPONSE 2'
      V3='RESPONSE 3'
      V4='RESPONSE 4'
      V5='RESPONSE 5'
       Q='QUESTION ';
```

根据这个程序以及虚设的数据所产生的统计分析表见下：

		QUESTIONS					
		1	2	3	4	5	6
TOTAL	RESPONSES	300	296	281	286	294	300
RESPONSE1	COUNT	122	116	108	129	109	110
	PERCENT	40.67	39.19	38.43	45.10	37.07	36.67
RESPONSE2	COUNT	114	123	111	101	112	113
	PERCENT	38.00	41.55	39.50	35.31	38.10	37.67
RESPONSE3	COUNT	133	129	107	100	118	128
	PERCENT	44.33	43.58	38.08	34.97	40.14	42.67
RESPONSE4	COUNT	121	122	127	114	131	114
	PERCENT	40.33	41.22	45.20	39.86	44.56	38.00
RESPONSE5	COUNT	129	122	111	101	103	123
	PERCENT	43.00	41.22	39.50	35.31	35.03	41.00

上述的表格类似 PROC ITEM (见第 13 章) 所产生的报表输出。 不同的是，TABULATE 程序所产生的表格只含百分比或总和，不作试题之信度或效度的分析，也不考虑每一种答案的正确性。因此，本章的表格适用在社会、民意、市场等调查资料上。 第 13 章的 ITEM 程序适用于成就测验或性向测验的试题分析上。

分母串定义的其他注意事项

若读者在关键字 PCTN 或 PCTSUM 之后未界定任何分母时，TABULATE 程序自动以整个资料文件的总和当作分母。若读者在原资料文件内有分组 (由 BY 指令所界定)，则总和的定义是指分组后的小资料文件内之总和。

分母的定义也可藉统计值 ALL 而达成，请看下面指令与报表的示范：

```
TABLE A*B ALL,X*(SUM PCTSUM<A*B ALL>)/RTS=16;
```

		X	
		SUM	PCTSUM
A	B		
1	1	160	35
	2	80	17
2	B		
	1	100	22
	2	120	26
ALL		460	100

在这个例子里，X 的百分比之分母是 460，也就是 A*B 的四个分组的总和。因此，TABLEA*B ALL 的目的是求出 460 这个值。然而，PCTSUM<A*B ALL> 的目的是定义 PCTSUM 的分母。请读者注意，A*B ALL 的部分必须在 TABLE 以及 PCTSUM 的撰写上完全相同。

另外，值得注意的是，虽然而 TABLE 之行、列、页的定义部分可以简写，比方说：A*B*(C D)即等于 (A*B*C A*B*D)；然而，这种简写的表示法不可应用在分母的定义上。所以，如果读者想以 B*C 以及 B*D 当分母，则必须将这两个同时包括在定义内。请看下面 TABLE 指令的示范：

```
TABLE A*B, (C D)*(N PCTN<B*C B*D>);
```

但如果读者想用 (A*B) 分组的和来取代 (B*C) 以及(B*D) 两个分母，则只须在分母定义部分界定一次即可 (见下面 TABLE 指令的改写)：

```
TABLE A*B, (C D)*(N PCTN<A*B>);
```

■ 表格的标题与形式设计

每一表格的行、列、页向量的标题内可含下列四种文字符串：

甲、分类变量的值

乙、TABLE 指令中界定的变量或关键字的文字解释

丙、INPUT 指令中界定的变量名称或标签 (Label)

丁、统计值 (含 ALL) 的代号或文字定义

有关 (甲) 分类变量值的字符串，读者可由输入资料文件内直接获得或藉 FORMAT 程序来界定。

关于 (乙) TABLE 指令中界定的文字解释，读者必须用单引号括住而且紧接在被解释的变量或关键字之后，如：ALL='STUDENTS PASSED'。

若属 (丙) 类的变量名称或标签，则读者必须在 DATA= 输入资料文件名称的选项上先行界定。

有关这方面的细节，读者可参阅本书附录 B。

(丁) 类的统计值代号或文字定义的处理与 (乙)类字符串几乎完全相同。值得注意的是：百分比的文字说明必须放在分母串的定义之后。详情请参阅上一节 '分母串的定义' 的说明。只要 TABULATE 程序中含 LABEL 或 KEYLABEL 指令及其界定的变量名称或标签，则在 TABLE 指令中可以不再重复已界定过的变量名称或标签。当一个标题 (或变量名称、标签)超过内定的长度时，TABULATE 程序会在适当的空白处或连接号 (-) 的地方对齐。若标题内无适当的空白或连接号，则 TABULATE 程序在对齐的标题部分添加连接号当作该字符串的字尾；然后从下一列的第一行起继续打印标题的剩余字符串。

因此，如果读者预先将连接号 (-) 加插在适当的音节间 (如：SUM-MARY)，则可保证打印出来的标题就是所要的样子。表格中细格的值亦可再经规划。规划的方式是藉助格式表达式。格式表达式的撰写形式如下：

```
FORMAT （ 或 F ） =格式名称
```

格式名称必须是事先在 PROC FORMAT 中已定义过的格式，然后，才能在 TABLE 指令中提及。若一个 TABLE 指令中含一个以上的格式表达式而且彼此冲突，则最后一次提及的格式表达式，或与行向量关系最密切的格式表达式，才是有效的格式表达式。

页向量上所界定的格式表达式对同一页上的列、行两向量均有效。同理，列向量上

所界定的格式表达式对同一列上的每一行均有效。最后，行向量上所界定的格式表达式对同一行内的每一细格均有效。

每一直行的宽度视列元素的宽度而定。内设的宽度等于 F=12.2，也就是十二位数 (内含一位的小数点以及两位小数位数)。

下页的例子示范表格的标题与形式的设计，数据来自本章第 6.3 节所提及的市场销售量的示范：

程　序

```
PROC FORMAT;
     VALUE $REGFMT "NC"='NORTH CENTRAL'
                   "NE"='NORTHEAST'
                   "SO"='SOUTH'
                   "WE"='WEST';
     VALUE $SIZEFMT "S"='UNDER 50000'
                   "M"='50000 TO 500000'
                   "L"='OVER 500000';
     VALUE $SALEFMT "R"='RETAIL'
                   "W"='WHOLESALE';
PROC TABULATE;
     CLASS PRODUCT REGION CITYSIZE SALETYPE;
     VAR QUANTITY AMOUNT;
     FORMAT REGION $REGFMT.;
     FORMAT CITYSIZE $SIZEFMT.;
     FORMAT SALETYPE $SALEFMT.;
     LABEL PRODUCT ='PRODUCT CODE'
           REGION='REGION OF COUNTRY'
           CITYSIZE='CITY SIZE'
           SALETYPE='TYPE OF SALE'
           AMOUNT='$ AMOUNT';
     TABLE (ALL PRODUCT)*F=8., REGION ALL='REGIONAL TOTAL',
           (SALETYPE ALL)*(QUANTITY AMOUNT);
     KEYLABEL SUM= 'OF SALES'
              ALL= 'TOTAL';
```

报　表

1

SALES FIGURES

TOTAL

	TYPE OF SALE					
	RETAIL		WHOLESALE		TOTAL	
	QUANTITY OF SALES	$AMOUNT OF SALES	QUANTITY OF SALES	$AMOUNT OF SALES	QUANTITY OF SALES	$AMOUNT OF SALES
REGION OF COUNTRY						
NORTH CENTRAL	3810	95200	3810	76200	7620	171400
NORTHEAST	4869	121725	4869	97380	9738	219105
SOUTH	5706	143450	5706	114120	11412	257570
WEST	5576	139400	5576	111520	11152	250920
REGIONAL TOTAL	19961	499775	19961	399220	39922	898995

2

PRODUCT CODE A100

	TYPE OF SALE				TOTAL	
	RETAIL		WHOLESALE			
	QUANTITY OF SALES	$AMOUNT OF SALES	QUANTITY OF SALES	$AMOUNT OF SALES	QUANTITY OF SALES	$AMOUNT OF SALES
REGION OF COUNTRY						
NORTH CENTRAL	1250	31250	1250	25000	2500	56250
NORTHEAST	1600	40000	1600	32000	3200	72000
SOUTH	1880	47000	1880	37600	3760	84600
WEST	1840	46000	1840	36800	3680	82800
REGIONAL TOTAL6570	164250	6570	131400	13140	295650	

3

PRODUCT CODE A200

	TYPE OF SALE				TOTAL	
	RETAIL		WHOLESALE			
	QUANTITY OF SALES	$AMOUNT OF SALES	QUANTITY OF SALES	$AMOUNT OF SALES	QUANTITY OF SALES	$AMOUNT OF SALES
REGION OF COUNTRY						
NORTH CENTRAL	1295	32375	1295	25900	2590	58275
NORTHEAST	1645	41125	1645	32900	3290	74025
SOUTH	1925	48925	1925	38500	3850	87425
WEST	1885	47125	1885	37700	3770	84825
REGIONAL TOTAL6750	169550	6750	135000	13500	304550	

4

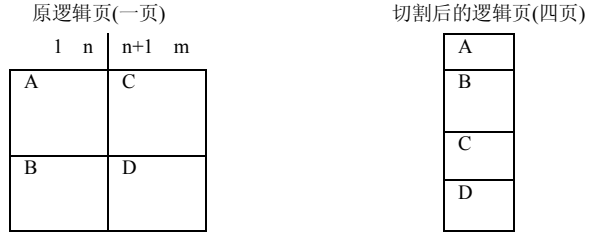
PRODUCT CODE A300

	TYPE OF SALE				TOTAL	
	RETAIL		WHOLESALE			
	QUANTITY OF SALES	$AMOUNT OF SALES	QUANTITY OF SALES	$AMOUNT OF SALES	QUANTITY OF SALES	$AMOUNT OF SALES
REGION OF COUNTRY						
NORTH CENTRAL	1265	31575	1265	25300	2530	56875
NORTHEAST	1624	40600	1624	32480	3248	73080
SOUTH	1901	47525	1901	38020	3802	85545
WEST	1851	46275	1851	37020	3702	83295
REGIONAL TOTAL	6641	165975	6641	132820	13282	298795

根据上述的报表，输出之表格共有四页 (见每一表格右上角的页数)，每一页的表格含各产品 A100，A200，A300 的销售情形 (见第二页至第四页的报表) 或三种销售情形的总结 (见第一页的报表)。这四页是指 SAS 软件所规划出来的逻辑页数 (Logical Pages)，与报表纸的大小不一定相符。

当逻辑页数远小于报表纸的大小时，读者可在 TABLE 指令中加入选项 CONDENSE 以便浓缩逻辑页数间的距离。

但当一页逻辑页数远大于报表纸的大小时，TABULATE 程序依下列示范的版面切割逻辑页：



切割的原则是将 (1 至 n 行) 的数列分成两页，如：A，B。然后再将 (n+1 至 m 行)

的数列照样分成两页，如：C，D。打印时，按 A，B，C，D 的顺序。

6.6 范　　例

例一：格局表达式的多种形式

本例的数据与产品 (PRODUCT)的销售状况有关。销售状况以 QUANTITY 和 AMOUNT 来衡量。可能影响销售状况的分类变量则包括销售地区 (REGION)、市场的大小 (CITYSIZE)、市场的人口数 (POP) 及销售方式 (SALETYPE)。

分析整理的步骤是首先将各种格局表达式在 PROC FORMAT 中加以定义；然后利用这些表达式制造出一个统计表格，内含五种统计值 (即 SUM，N，PCTSUM，PCTN，ALL 等) 以及它们的标签。标签的命名同时示范 TABULATE 程序如何将较长的标签 (或标题) 在空白处或连接号 (-) 处对齐。

程　序

```
 DATA SALES;
INPUT REGION $ CITYSIZE $ POP PRODUCT $ SALETYPE $ QUANTITY AMOUNT  @@;  CARDS;
NC S   25000 A100 R 150   3750.00  NC S   45000 A100 W 250   5750.00
NE S   37000 A100 R 200   5000.00  NE S   57000 A100 W 280   6000.00
SO S   48000 A100 R 410  10250.00  SO S   68000 A100 W 410  10250.00
WE S   32000 A100 R 180   4500.00  WE S   42000 A100 W 280   5500.00
NC M  125000 A100 R 350   8750.00  NC M  155000 A100 W 450   8950.00
NE M  237000 A100 R 600  15000.00  NE M  247000 A100 W 620  18000.00
SO M  348000 A100 R 710  17750.00  SO M  368000 A100 W 810  19750.00
WE M  432000 A100 R 780  19500.00  WE M  489000 A100 W 880  21500.00
NC L  625000 A100 R 750  18750.00  NC L  675000 A100 W 750  22750.00
NE L  837000 A100 R 800  20000.00  NE L  857000 A100 W 870  22500.00
SO L  857000 A100 R 850  21000.00  SO L  787000 A100 W 880  24000.00
WE L  900000 A100 R 900  20500.00  WE L  910000 A100 W 900  23500.00
;
PROC FORMAT;
     VALUE $REGFMT "NC"='NORTH CENTRAL'
                   "NE"='NORTHEAST'
                   "SO"='SOUTH'
                   "WE"='WEST';
     VALUE $SIZEFMT "S"='UNDER 50000'
                   "M"='50000 TO 500000'
                   "L"='OVER 500000';
     VALUE $SALEFMT "R"='RETAIL' "W"='WHOLESALE';
PROC FORMAT;
```

```
     PICTURE PCT LOW - <0=  '000.00%'  (PREFIX='-')
               0 - HIGH= '0000.00%';
PROC TABULATE;
     CLASS REGION SALETYPE; VAR AMOUNT;
     FORMAT REGION $REGFMT.; FORMAT CITYSIZE $SIZEFMT.;
     FORMAT SALETYPE $SALEFMT.;
     TABLE ALL*F=DOLLAR11. REGION*F=COMMA11., /*ROW DIMESION*/
     SALETYPE='REGIONAL SALES ANALYSIS'*   /*COLUMN DIMESION*/
          AMOUNT=' '*
          (SUM  PCTSUM<REGION ALL>*F=PCT.
          N*F=6. PCTN  <REGION ALL>*F=PCT.)
          /RTSPACE=12;   /*TABLE OPTIONS*/
     KEYLABEL SUM= 'REVENUE'
              N  = 'LOCA-TIONS'
              PCTSUM = 'PERCENT OF SALES'
              PCTN   = 'PERCENT OF LOCNS'
              ALL= 'TOTAL';  RUN;
```

结　果

报表 6.1　格局表达式的多种形式

	REGIONAL SALES ANALYSIS							
	RETAIL				WHOLESALE			
	REVENUE	PERCENT OF SALES	LOCA-TIONS	PERCENT OF LOCNS	REVENUE	PERCENT OF SALES	LOCA-TIONS	PERCENT OF LOCNS
TOTAL	$164,750	100.00%	12	100.00%	$188,450	100.00%	12	100.00%
REGION								
NORTH CENTRAL	31,250	18.96%	3	25.00%	37,450	19.87%	3	25.00%
NORTHEAST	40,000	24.27%	3	25.00%	46,500	24.67%	3	25.00%
SOUTH	49,000	29.74%	3	25.00%	54,000	28.65%	3	25.00%
WEST	44,500	27.01%	3	25.00%	50,500	26.79%	3	25.00%

例二：实验仪器的每月记录资料

本例的数据来自两间实验室 (以 LABORATORY 1 以及 LABORATORY 2 识别) 、两部仪器 (以 MACHINE 1 以及 MACHINE 2 分别表之) 在十二个月内的记录资料。以随机的方式读取资料，每月均为四次。分析时取这四次资料的平均率 (Mean)、全距 (Range) 以及标准差 (Standard Deviation)。

分析的过程含格局表达式的定义 (PROC FORMAT) 以及上述统计值的计算 (PROC TABULATE)。除此之外，PROC TABULATE 的指令 LABEL 与 KEYLABEL 示范如何更有效的设计表格的标题与形式。

报表含两页 (Logical Pages) 的表格：每一间实验室的记录资料占一页。

程　序

```
DATA READINGS;
     DO LAB_ID=1 TO 2;
        DO M_ID=1 TO 2;
           DO PART=1 TO 4;
              DO MONTHNUM=1 TO 12;
                 INPUT READING @@ ;  OUTPUT;
              END;
           END;
        END;
     END;
     CARDS;
60 36 59 46 53 35 53 39 49 42 61 42 61 42 92 97 91 82 83 16 83 79 94 24
56 62 33 44 42 34 99 44 74 70 47 36 86 75 30 54 38 32 33 74 93 68 60 82
91 79 67 63 34 40 90 81 39 70 50 72 51 41 83 40 64 36 65 25 38 77 92 15
20 57 52 46 41 78 22 30 88 31 11 19 32 83 60 36 79 96 43 22 75 49 93 88
21 84 35 80 79 27 85 63 69 61 68 16 26 81 39 51 76 38 32 47 43 52 61 58
76 81 41 34 40 49 70 43 50 90 97 48 82 70 71 61 87 65 20 91 70 61 52 71
54 30 83 31 44 38 77 53 79 23 98 21 84 44 29 51 90 52 94 59 46 42 74 38
47 72 93 77 88 69 45 30 62 83 68 31 49 82 90 24 81 76 36 61 56 87 62 53
;
PROC FORMAT;
     VALUE MONFMT  1='JANUARY'  7='JULY'
                   2='FEBRUARY'8='AUGUST'
                   3='MARCH'   9='SEPTEMBER'
                   4='APRI'    10='OCTOBER'
                   5='MAY'     11='NOVEMBER'
                   6='JUNE'    12='DECEMBER';
     VALUE MFMT1='MACHINE 1'
                   2='MACHINE 2';
PROC TABULATE ;
     CLASS LAB_ID M_ID MONTHNUM;
     VAR READING;
     FORMAT MONTHNUM MONFMT.;
     FORMAT M_ID MFMT.;
     LABEL LAB_ID='LABORATORY'
              M_ID='MACHINE LAB REPORT'
           MONTHNUM='MONTH';
    KEYLABEL ALL='SUMMARY'
```

```
                    N='FREQUENCY'
                  STD='STANDARD DEVIATION';
        TABLE LAB_ID*F=9.3,MONTHNUM ALL, M_ID*READING*
        (N*F=9. MEAN RANGE STD) / RTS=12;
    RUN;
```

结　果

报表 6.2　实验仪器的每月记录资料

LABORATORY1

	MACHINELABREPORT							
	MACHINE1				MACHINE2			
	READING				READING			
	FREQUENCY	MEAN	RANGE	STANDARD DEVIATION	FREQUENCY	MEAN	RANGE	STANDARD DEVIATION
MONTH								
JANUARY	4	65.750	30.000	13.672	4	48.500	71.000	31.075
FEBRUARY	4	53.750	39.000	18.007	4	65.000	42.000	19.664
MARCH	4	53.500	62.000	28.781	4	65.500	31.000	13.178
APRIL	4	60.250	53.000	24.878	4	46.250	27.000	11.899
MAY	4	56.000	53.000	24.180	4	54.500	45.000	20.761
JUNE	4	45.750	50.000	24.199	4	62.500	60.000	29.275
JULY	4	67.000	66.000	29.620	4	55.000	68.000	29.200
AUGUST	4	43.250	58.000	23.852	4	39.500	59.000	27.863
SEPTEMBER	4	74.750	44.000	18.839	4	60.000	50.000	25.390
OCTOBER	4	64.750	37.000	15.903	4	56.750	46.000	20.887
NOVEMBER	4	65.500	47.000	20.042	4	61.500	82.000	39.179
DECEMBER	4	46.000	58.000	25.140	4	48.500	73.000	36.991
SUMMARY	48	58.021	83.000	22.111	48	55.292	85.000	24.692

LABORATORY2

	MACHINELABREPORT							
	MACHINE1				MACHINE2			
	READING				READING			
	FREQUENCY	MEAN	RANGE	STANDARD DEVIATION	FREQUENCY	MEAN	RANGE	STANDARD DEVIATION
MONTH								
JANUARY	4	51.250	61.000	32.201	4	58.500	37.000	17.253
FEBRUARY	4	79.000	14.000	6.164	4	57.000	52.000	24.138
MARCH	4	46.500	36.000	16.523	4	73.750	64.000	30.126
APRIL	4	56.500	46.000	19.227	4	45.750	53.000	23.768
MAY	4	70.500	47.000	20.857	4	75.750	46.000	21.515
JUNE	4	44.750	38.000	16.215	4	58.750	38.000	17.115
JULY	4	51.750	65.000	30.750	4	63.000	58.000	27.142
AUGUST	4	61.000	48.000	21.787	4	50.750	31.000	14.245
SEPTEMBER	4	58.000	27.000	13.589	4	60.750	33.000	13.841
OCTOBER	4	66.000	38.000	16.553	4	58.750	64.000	31.330
NOVEMBER	4	69.500	45.000	19.468	4	75.500	36.000	15.780
DECEMBER	4	48.250	55.000	23.472	4	35.750	32.000	13.451
SUMMARY	48	58.583	81.000	21.050	48	59.500	77.000	22.243

例三：实验仪器的每月记录资料之比较

本例的数据来源与例二完全相同，只是分析的指令稍有不同。在本例中，我们希望将两部仪器的记录资料印成上下两列，以便于比较。因此，TABLE 指令将仪器 (M_ID) 定义成列向量的分类变量之一，如此就能将两部仪器的记录上下比较了。

另外，PROC FORMAT 也刻意将十二个月份的文字说明部分添加连接号(—)以使十二个月份都能适当地并列在同一页的十二行内。

程　序

```
DATA READINGS;
    DO LAB_ID=1 TO 2;
      DO M_ID=1 TO 2;
        DO PART=1 TO 4;
          DO MONTHNUM=1 TO 12;
            INPUT READING @@ ;  OUTPUT;
          END;
        END;
      END;
    END;
    CARDS;
60 36 59 46 53 35 53 39 49 42 61 42 61 42 92 97 91 82 83 16 83 79 94 24
56 62 33 44 42 34 99 44 74 70 47 36 86 75 30 54 38 32 33 74 93 68 60 82
91 79 67 63 34 40 90 81 39 70 50 72 51 41 83 40 64 36 65 25 38 77 92 15
20 57 52 46 41 78 22 30 88 31 11 19 32 83 60 36 79 96 43 22 75 49 93 88
21 84 35 80 79 27 85 63 69 61 68 16 26 81 39 51 76 38 32 47 43 52 61 58
76 81 41 34 40 49 70 43 50 90 97 48 82 70 71 61 87 65 20 91 70 61 52 71
54 30 83 31 44 38 77 53 79 23 98 21 84 44 29 51 90 52 94 59 46 42 74 38
47 72 93 77 88 69 45 30 62 83 68 31 49 82 90 24 81 76 36 61 56 87 62 53
;
PROC FORMAT;
     VALUE MONFMT  1='JANUARY'  7='JULY'
                   2='FEBRUARY'8='AUGUST'
                   3='MARCH'   9='SEPTEMBER'
                   4='APRIL'       10='OCTOBER'
                   5='MAY'     11='NOVEMBER'
                   6='JUNE'    12='DECEMBER';
     VALUE MFMT 1='1'
                2='2';
PROC TABULATE ;
     CLASS LAB_ID M_ID MONTHNUM;
```

```
        VAR READING;
        FORMAT MONTHNUM MONFMT.;
        FORMAT M_ID MFMT.;
        LABEL LAB_ID='LABORATORY'
                M_ID='MACHINE'
              MONTHNUM='MONTH';
        KEYLABEL ALL='SUMMARY'
                  N='NUMBER OF READINGS'
                STD='STANDARD DEVIATION';
        TABLE LAB_ID*F=6.3,(M_ID ALL)*(N*F=6. MEAN RANGE STD),
              READING*(MONTHNUM ALL) / RTS=22;
RUN;
```

结　果

报表 6.3　实验仪器的每月记录资料之比较

LABORATORY1

		READING												
		MONTH												SUMMARY
		JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER	
MACHINE														
1	NUMBER OF READINGS	4	4	4	4	4	4	4	4	4	4	4	4	48
	MEAN	65.750	53.750	53.500	60.250	56.000	45.750	67.000	43.250	74.750	64.750	65.500	46.000	58.021
	RANGE	30.000	39.000	62.000	53.000	53.000	50.000	66.000	58.000	44.000	37.000	47.000	58.000	83.000
	STANDARD DEVIATION	13.672	18.007	28.781	24.878	24.180	24.199	29.620	23.852	18.839	15.903	20.042	25.140	22.111
2	NUMBER OF READINGS	4	4	4	4	4	4	4	4	4	4	4	4	48
	MEAN	48.500	65.000	65.500	46.250	54.500	62.500	55.000	39.500	60.000	56.750	61.500	48.500	55.292
	RANGE	71.000	42.000	31.000	27.000	45.000	60.000	68.000	59.000	50.000	46.000	82.000	73.000	85.000
	STANDARD DEVIATION	31.075	19.664	13.178	11.899	20.761	29.275	29.200	27.863	25.390	20.887	39.179	36.991	24.692
SUMMARY	NUMBER OF READINGS	8	8	8	8	8	8	8	8	8	8	8	8	96
	MEAN	57.125	59.375	59.500	53.250	55.250	54.125	61.000	41.375	67.375	60.750	63.500	47.250	56.656
	RANGE	71.000	47.000	62.000	61.000	57.000	64.000	77.000	65.000	55.000	48.000	83.000	73.000	88.000
	STANDARD DEVIATION	24.062	18.462	21.693	19.543	20.879	26.427	27.974	24.095	22.148	17.710	28.889	29.310	23.354

LABORATORY2

		READING												
		MONTH												
	JANUARY	FEBRUARY	MARCH	APRIL	MAY	JUNE	JULY	AUGUST	SEPTEMBER	OCTOBER	NOVEMBER	DECEMBER	SUMMARY	
MACHINE														
1	NUMBEROFREADINGS	4	4	4	4	4	4	4	4	4	4	4	4	48
	MEAN	51.250	79.000	46.500	56.500	70.500	44.750	51.750	61.000	58.000	66.000	69.500	48.250	58.583
	RANGE	61.000	14.000	36.000	46.000	47.000	38.000	65.000	48.000	27.000	38.000	45.000	55.000	81.000
	STANDARDDEVIATION	32.201	6.164	16.523	19.227	20.857	16.215	30.750	21.787	13.589	16.553	19.468	23.472	21.050
2	NUMBEROFREADINGS	4	4	4	4	4	4	4	4	4	4	4	4	48
	MEAN	58.500	57.000	73.750	45.750	75.750	58.750	63.000	50.750	60.750	58.750	75.500	35.750	59.500
	RANGE	37.000	52.000	64.000	53.000	46.000	38.000	58.000	31.000	33.000	64.000	36.000	32.000	77.000
	STANDARDDEVIATION	17.253	24.138	30.126	23.768	21.515	17.115	27.142	14.245	13.841	31.330	15.780	13.451	22.243
SUMMARY	NUMBEROFREADINGS	8	8	8	8	8	8	8	8	8	8	8	8	96
	MEAN	54.875	68.000	60.125	51.125	73.125	51.750	57.375	55.875	59.375	62.375	72.5004	2.000	59.042
	RANGE	63.000	54.000	64.000	56.000	50.000	49.000	74.000	61.000	36.000	67.000	46.000	55.000	82.000
	STANDARDDEVIATION	24.228	20.107	26.798	20.822	19.817	17.153	27.516	17.900	12.783	23.519	16.716	18.928	21.546

例四：收支帐目平衡表

本例的数据来自跳蚤市场 (Flea Market) 在某年第一个季度里 (含一月、二月、三月) 的收支平衡。由于跳蚤市场内设五个部门，即会计 (ACCOUNTING)、人力 (HUMAN RESOURCES) 、管理(SYSTEMS)、生产 (PRODUCTION) 以及市场调查 (MARKETING) 等，收支的帐目先按各部门在三个月内的情形打印。然后，总结这三个月的支出 (QUARTER TOTAL)，并与预算额 (ALLOCATED FUNDS) 相比。相差的金额就是收支平衡的数目 (REMAINING FUNDS)：正值代表盈余，负值代表亏损。

此外，由于各部门内支出的帐号不只一个，因此，同一部门内所有支出帐号 (ACCOUNT) 下的支出也总加起来，称为 SUBTOTAL。这个总支出额横印在每一部门的最后一列。SUBTOTAL 的总和就是 TOTAL，印在表格的最下端。

程　序

```
PROC FORMAT;
     VALUE  DEPT1='ACCOUNTING'
```

```
                    2='HUMAN RESOURCES'
                    3='SYSTEMS'
                    4='PRODUCTION'
                    5='MARKETING';
DATA SUMMARY;
     INPUT DEPT ACCT JAN FEB MAR BUDGET;
     TOTAL=JANFEBMAR;LEFT=BUDGETTOTAL;
CARDS;
1  01345  12980  14009  17800  40000
1  01578   8000   7900   4500  40000
1  01674  11950  13534  17994  40000
2  02134  34520  26560  24399  80000
2  02403  10435  15494  10009  40000
3  04138  24850  22530  24399  70000
3  04279   9984  14209  13500  40000
3  04290  10948  14539  11459  40000
4  05139  12000  14532  12098  40000
4  05260  15893  14099   7304  40000
4  05370  11980  13900  11480  40000
4  05399  20435  19095  18053  60000
5  06120  23435  23543  19054  60000
5  06342  13049  15349  18943  50000
5  06401  20943  25943  19432  65000
;
PROC TABULATE ;
     TITLE1 '  '; TITLE2 '  ';
     TITLE3 'FLEEMARKET'; TITLE4 'DEPARTMENTAL BUDGETARY REPORT';
     TITLE5 'FOR THE FIRST QUARTER'; TITLE6 '  ';
     CLASS DEPT ACCT; VAR JAN FEB MAR TOTAL LEFT BUDGET; FORMAT DEPT DEPT.;
     LABEL TOTAL='QUARTER TOTAL'; LABEL BUDGET='ALLOCATED FUNDS';
     LABEL LEFT='REMAINING FUNDS'; LABEL JAN='JANUARY EXPENDITURES';
     LABEL FEB='FEBRUARY EXPENDITURES'; LABEL MAR='MARCH EXPENDITURES';
     TABLE /*ROW DIMENSIOM*/
           (DEPT=' '*(ACCT=' ' ALL='SUBTOTAL') ALL='TOTAL')
           *F=COMMA12.2,/*COLUMN DIMENSION*/
           ((JAN FEB MAR)*SUM='  ' TOTAL*SUM=' ' (BUDGET LEFT)*SUM=' ')/
           /*TABLE OPTIONS*/
           BOX='DEPARTMENT ACCOUNT';
RUN;
```

结　果

报表 6.4　收支帐目平衡表

FLEEMARKET

DEPARTMENTAL BUDGETARY REPORT

FOR THE FIRST QUARTER

DEPARTMENT ACCOUNT		JANUARY EXPENDITURES	FEBRUARY EXPENDITURES	MARCH EXPENDITURES	QUARTER TOTAL	ALLOCATED FUNDS	REMAINING FUNDS
ACCOUNTING	1345	12,980.00	14,009.00	17,800.00	44,789.00	40,000.00	4,789.00
	1578	8,000.00	7,900.00	4,500.00	20,400.00	40,000.00	19,600.00
	1674	11,950.00	13,534.00	17,994.00	43,478.00	40,000.00	3,478.00
	SUBTOTAL	32,930.00	35,443.00	40,294.00	108,667.00	120,000.00	11,333.00
HUMANRESOURCES	2134	34,520.00	26,560.00	24,399.00	85,479.00	80,000.00	5,479.00
	2403	10,435.00	15,494.00	10,009.00	35,938.00	40,000.00	4,062.00
	SUBTOTAL	44,955.00	42,054.00	34,408.00	121,417.00	120,000.00	1,417.00
SYSTEMS	4138	24,850.00	22,530.00	24,399.00	71,779.00	70,000.00	1,779.00
	4279	9,984.00	14,209.00	13,500.00	37,693.00	40,000.00	2,307.00
	4290	10,948.00	14,539.00	11,459.00	36,946.00	40,000.00	3,054.00
	SUBTOTAL	45,782.00	51,278.00	49,358.00	146,418.00	150,000.00	3,582.00
PRODUCTION	5139	12,000.00	14,532.00	12,098.00	38,630.00	40,000.00	1,370.00
	5260	15,893.00	14,099.00	7,304.00	37,296.00	40,000.00	2,704.00
	5370	11,980.00	13,900.00	11,480.00	37,360.00	40,000.00	2,640.00
	5399	20,435.00	19,095.00	18,053.00	57,583.00	60,000.00	2,417.00
	SUBTOTAL	60,308.00	61,626.00	48,935.00	170,869.00	180,000.00	9,131.00
MARKETING	6120	23,435.00	23,543.00	19,054.00	66,032.00	60,000.00	6,032.00
	6342	13,049.00	15,349.00	18,943.00	47,341.00	50,000.00	2,659.00
	6401	20,943.00	25,943.00	19,432.00	66,318.00	65,000.00	1,318.00
	SUBTOTAL	57,427.00	64,835.00	57,429.00	179,691.00	175,000.00	4,691.00
TOTAL		241,402.00	255,236.00	230,424.00	727,062.00	745,000.00	17,938.00

例五：年龄与病历的关系

本例的数据在于探讨年龄是否与某些病历有关联的问题。年龄原是一个连续变量，但在本例中分为四个年龄组 (从三十岁至五十岁间，每增五岁则归成一组)。病历也仅限于四种，各以 1 到 4 的整数代表：5

1 = 胸口疼痛

2 = 偶而感冒

3 = 心脏病突发

4 = 呼吸不顺

分析的结果含四种统计值：即各病历在所有年龄组的发生次数 (FREQUENCY)、同一个病历在不同年龄组出现的百分比 (% OF THIS EVENT)、不同病历在同一年龄组出现的百分比 (% OF AGEGROUP) 以及不同的病历在不等的年龄组出现的百分比 (% OF ALL EVENTS)。

程　序

```
PROC FORMAT;
     VALUE AGEFMT 1='3035'
```

```
                    2='3640'
                    3='4145'
                    4='4650';
      VALUE CDFMT   1='CHEST PAIN'
                    2='OCCATIONAL COLD'
                    3='HEART ATTACK'
                    4='SHORT OF BREATH';
    PROC TABULATE F=7.2 ;
    CLASS AGE CODE;
    VAR ID;
    TABLE
         /*ROW DIMESION*/
         (ALL CODE=' ')*
         (N*F=5.
            (PCTN<ID AGE>='% OF THIS EVENT'
             PCTN<CODE ALL>='% OF AGE GROUP'
             PCTN<ALL*AGE*ID ALL*ID CODE*AGE8ID CODE*ID>
                          ='% OF ALL EVENTS')*F=7.2),
         /*COLUMN DIMENSION*/
         ID AGE
         /*TABLE OPTIONS*/
         /RTS=33 MISSTEXT='NONE'
         BOX='MEDICAL EVENTS BY AGE GROUP';
    LABEL ID='ALL AGE GROUPS'
          AGE='AGE GROUPS';
    FORMAT AGE AGEFMT.;
    FORMAT CODE CDFMT.;
    KEYLABEL ALL='ALLEVENTS'
             N='FREQUENCY';
RUN;
```

结　果

报表 6.5　年龄与病历的关系

MEDICALEVENTSBYAGEGROUP		ALLAGEGROUPS	AGEGROUPS			
			30-35	36-40	41-45	46-50
ALLEVENTS	FREQUENCY	100	20	25	30	25
	%OFTHISEVENT	100.00	20.00	25.00	30.00	25.00
	%OFAGEGROUP	100.00	100.00	100.00	100.00	100.00
	%OFALLEVENTS	100.00	20.00	25.00	30.00	25.00
CHESTPAIN	FREQUENCY	30	6	4	12	8
	%OFTHISEVENT	100.00	20.00	13.33	40.00	26.67
	%OFAGEGROUP	30.00	30.00	16.00	40.00	32.00
	%OFALLEVENTS	30.00	6.00	4.00	12.00	8.00

HOSPITALVISIT	FREQUENCY	18	4	6	6	2
	%OFTHISEVENT	100.00	22.22	33.33	33.33	11.11
	%OFAGEGROUP	18.00	20.00	24.00	20.00	8.00
	%OFALLEVENTS	18.00	4.00	6.00	6.00	2.00
HEARTATTACK	FREQUENCY	33	5	10	8	10
	%OFTHISEVENT	100.00	15.15	30.30	24.24	30.30
	%OFAGEGROUP	33.00	25.00	40.00	26.67	40.00
	%OFALLEVENTS	33.00	5.00	10.00	8.00	10.00
SHORTOFBREATH	FREQUENCY	19	5	5	4	5
	%OFTHISEVENT	100.00	26.32	26.32	21.05	26.32
	%OFAGEGROUP	19.00	25.00	20.00	13.33	20.00
	%OFALLEVENTS	19.00	5.00	5.00	4.00	5.00

例六：含多选题的问卷调查资料

本例旨在示范如何应用 TABULATE 程序来分析多选题的答案。数据则取自于一个对 150 名上班族口味的问卷调查。口味 (TASTE) 的变量下分五类，即：

1='CHINESE'　(中国菜)
2='AMERICAN'　(美国菜)
3='JAPANESE'　(日本料理)
4='KOREAN'　(韩国口味)
5='MEXICAN'　(墨西哥餐)

当受试者选中一样口味后，还要再陈述其圈选的理由，如：

R1='FLAVOR'　(味道好)
R2='TEXTURE'　(咀嚼起来有劲)
R3='NUTRITIONAL VALUE'　(有营养)
R4='PRICE'　(价钱公道)
R5='AVAILABILITY'　(方便)

由于受试者可圈选的理由不只一个，所以在计算理由的百分比时，应采用受试者的总人数为分母，而非被选中的理由总数。

程　序

```
PROC FORMAT;
     VALUE TFMT 1='CHINESE'
                2='AMERICAN'
                3='JAPANESE'
                4='KOREAN'
                5='MEXICAN';
PROC TABULATE F=7.2 ;
     CLASS TASTE;
     VAR ID R1R5;
     TABLE ID*N='RESPONDENTS'*F=7.
           (R1 R2 R3 R4 R5)*(N='COUNT'*F=7. PCTN<ID>='PERCENT'),
           TASTE=' ' ALL='ALL TASTE'
           /RTS=27 BOX='TASTE SURVEY';
     LABEL ID='TOTAL'
           R1='FLAVOR'
```

```
            R2='TEXTURE'
            R3='NUTRITIONAL VALUE'
            R4='PRICE'
            R5='AVAILABILITY';
        FORMAT TASTE TFMT.;
        RUN;
```

结　果

报表 6.6　含多选题的问卷调查资料

TASTESURVEY		CHINESE	AMERICAN	JAPANESE	KOREAN	MEXICAN	ALLTASTES
TOTAL	RESPONDENTS	50	25	40	25	10	150
FLAVOR	COUNT	15	6	17	15	5	58
	PERCENT	30.00	24.00	42.50	60.00	50.00	38.67
TEXTURE	COUNT	12	9	28	8	9	66
	PERCENT	24.00	36.00	70.00	32.00	90.00	44.00
NUTRITIONAL VALUE	COUNT	23	6	14	16	4	63
	PERCENT	46.00	24.00	35.00	64.00	40.00	42.00
PRICE	COUNT	32	16	30	14	9	101
	PERCENT	64.00	64.00	75.00	56.00	90.00	67.33
AVAILABILITY	COUNT	35	15	22	18	797	
	PERCENT	70.00	60.00	55.00	72.00	70.00	64.67

第 7 章　关系强度的测量：统计程序 PROC CORR

7.1　PROC CORR 程序概述

CORR 程序可用来测量两个变量之间的关系强度。针对测量变量所用的尺度 (Measurement Scale) 不同，CORR 程序提供以下数种测量关系强度的方法：

■以等距尺度或比例尺度测量的参数统计方法，产生相关系数矩阵

- ●积差相关 (Pearson's Product Moment Correlation)
- ●加权的积差相关 (Weighted Product Moment Correlation)
- ●Cronbach 的阿尔法相关系数 (Alpha Coefficient)

■以等级尺度测量的无参数统计方法，产生关联系数矩阵

- ●等级相关 (Spearman's Rank-Order Correlation)
- ●Kendall's tau-b
- ●Hoeffding's Measure of Dependence，D

CORR 程序除了可执行上述的各种统计方法外，还可将运算所得的相关系数矩阵或关联系数矩阵，储存在一个 SAS 输出文件里，以供进一步的分析。

7.2　如何撰写 PROC CORR 程序

PROC CORR 含六道指令，它们的格式如下：

```
PROC CORR       选项串;
     VAR        变量名称串;
     WITH       变量名称串;
     PARTIAL    变量名称;
     WEIGHT     变量名称;
     FREQ       变量名称;
     BY         变量名称串;
```

PROC CORR 指令之后的指令，可随意排列顺序。

指令 #1　PROC CORR 选项串;

此处有十九个选项可供选择。

这些选项可分为四大类：第一类选项界定输出 / 输入资料文件名称，第二类选项界

定测量关系强度的方法，第三类选项界定输出资料的项目，第四类选项界定有关计算过程的各种事宜。若读者省略所有十九个选项，则 CORR 程序会自动计算积差相关的系数，以及其显著性检验的结果。 另外 CORR 程序也会自动印出一些基本的描述性统计值，如：平均数、标准差、最大及最小值等。

第一类选项　下列五个选项界定输出 / 输入资料文件的名称：

(1) DATA= 输入资料文件名称

指明到底对那一个资料文件执行相关分析。若省略此选项，则 SAS 会自动找出在此程序之前最后形成的 SAS 资料文件，对它执行分析。

(2) OUTP= 输出资料文件名称

这一个 TYPE= CORR 的输出资料文件，含有 PEARSON 积差相关系数之矩阵 、各变量的平均数、标准差、观察体个数。

选用 OUTP= 选项时，CORR 程序会自动进行积差相关分析。因此，可以不必再选用 PEARSON 选项，以免重复。

(3) OUTS= 输出资料文件名称

这一个 TYPE=CORR 的输出资料文件，含有 Spearman 等级相关系数之矩阵 、各变量的平均数、标准差、观察体个数。

选用 OUTS= 选项时，CORR 程序会自动进行等级相关分析。因此，可以不必再选用 SPEARMAN 选项，以免重复。

(4) OUTK= 输出资料文件名称

这一个 TYPE=CORR 的输出资料文件，含有 Kendall's tau-b、各变量的平均数、标准差、观察体个数。

选用 OUTK = 选项时，CORR 程序会自动计算 Kendall's tau-b 相关系数。因此，可以不必再选用 KENDALL 选项，以免重复。

(5) OUTH= 输出资料名称

这一个 TYPE=CORR 的输出资料文件，含有 Hoeffding's D 系数、各变量的平均数、标准差、观察体个数。

选用 OUTH = 选项时，CORR 程序会自动计算 Hoeffding's D 系数。因此，可以不必再选用 HOEFFDING 选项，以免重复。

第二类选项　下列四个选项界定测量关系强度的方法，内设值是 PEARSON：

(1) PEARSON

要求计算积差相关系数，这也是此类选项的内设值。如果读者想同时计算积差相关系数与其他种类的系数如：KENDALL，SPEARMAN 或 HOEFFDING 等，则必须选用 PEARSON 选项，不可省略。否则，CORR 程序不会自动算出积差相关系数。

(2) SPEARMAN

计算 SPEARMAN 等级相关系数。这种相关系数适用于以等级尺度测量的资料。SAS 在计算相关之前，先将原始分数转换成名次 (Rank)，再进行运算。等级相关系数介于 -1 与 +1 之间。系数的绝对值大小表示变量间关系强度的高低。

若读者选用 SPEARMAN 选项，则不可同时选用 WEIGHT 指令。

(3) KENDALL

计算 KENDALL 的 tau-b 符合系数。这种相关系数适用于以等级尺度测量的资料。SAS 在计算 tau-b 之前，会先将原始分数转换成名次 (Rank)，再加以运算。tau-b 的值在-1 与 +1 之间。其绝对值的大小表示变量间符合度的强弱。

若读者选用 KENDALL 选项，则不可同时选用 WEIGHT 指令。

(4) HOEFFDING

计算 HOEFFDING 的独立性统计值 (Measure of Independence 或 D)。SAS 执行时，将 HOEFFDING 的 D 值 (参阅 Hoeffding，1948；Hollander and Wolfe，1973) 乘以 30，使 D 值介于 -.5 与 +1 之间。当 D 值为正值且其值愈大时，表示变量间彼此独立性愈强，关连性愈低。

若读者选用 HOEFFDING 选项，则不可同时选用 WEIGHT 指令。

(5) ALPHA

计算阿尔法系数，此系数等于 VAR 指令中每一变量与其余变量之总和的积差相关。一般而言，阿尔法系数代表一个测验的信度，其计算的公式可根据标准化分数或未经标准化的原始分数，其结果可由 OUTP＝SAS 资料文件输出。

第三类选项　下列八个选项界定输出的项目：

(1) BEST=n

只印出每一变量与其他变量间最高的 n 个相关系数。

(2) NOSIMPLE

不印出变量的描述性统计值，如：平均数、标准差、中位数、最大和最小值等。

(3) NOPRINT

不印出任何报表。若读者只需要相关系数矩阵或关联系数矩阵的输出文件，而不要报表，则该选用此选项。

(4) NOCORR

在输出资料文件中，不包括相关系数。但是文件型仍然而是 CORR (TYPE=CORR)。欲将文件型改为 SSCP 或 COV，可在界定输出文件时，利用 TYPE 选项加以更改。例如：

```
PROC CORR NOCORR SSCP OUT=SSCPMAT(TYPE=SSCP);
```

(5) NOPROB

不印出相关系数的显著性检定结果。

(6) SSCP

印出变量之离差平方和 (Sum of Squares，以 SS 表示) 及变量间离差内乘积 (CrossProduct，以 CP 表示) 的联合矩阵 (SSCP)。因 SSCP 矩阵只与积差相关系数有关，因此 SSCP 选项不应同时与 SPEARMAN，KENDALL，或 HOEFFDING 选项同时选用。

若读者同时选用 SSCP 及 OUTP= 选项，则将产生一个包含 SSCP 矩阵 (_TYPE_=SSCP) 与积差相关系数矩阵 (_TYPE_=CORR) 的 SAS 资料文件。

(7) COV

印出变量间共变异数 (Covariance) 的矩阵。因为这个矩阵只与积差相关系数有关，故 COV 选项不应同时与 SPEARMAN，KENDALL 或 HOEFFDING 选项选用。若同时选用 COV 及 OUTP= 选项，则将产生一个包含共变异数矩阵与积差相关系数矩阵的 SAS 资料文件。 其中共变异数矩阵在系统变量 _TYPE_ 上的值是 COV。

(8) RANK

使每一变量与其他变量的相关系数，依其绝对值，由大而小印出。若省略此选项，则将依各变量界定的顺序而输出相关系数。

第四类选项　下列两个选项规定有关计算过程中的各种事宜：

(1) VARDEF=N

VARDEF=DF

VARDEF=WEIGHT(或 WGT)

VARDEF=WDF

界定变异数计算时所用的分母，有下列四种选择：

N：　观察体总数。

DF：　观察体总数减 1，这是此选项的内设值。

WEIGHT(或 WGT)：　加权后的观察体总数。

WDF：　上述 WEIGHT 值减 1。

(2) NOMISS

若某个观察体在计算的任何一个变量上有遗漏数据，它就被剔除在所有的计算过程之外。若省略此选项，则 CORR 程序会采用取一种比较宽容的方式，在这种方式之下，观察体只在计算的一对变量上有遗漏数据，才被删除。若在其它变量上无遗漏数据，则仍然而被纳入运算之中。

指令 #2　VAR 变量名称串；

读者可在本指令中列举被分析的变量。若省略此指令，则 CORR 程序自动对输入资料文件中所有数值变量进行分析。

指令 #3　WITH 变量名称串；

须与 VAR 指令联用。WITH 指令中所列举的 m 个变量，与 VAR 指令中所列举的 n 个变量，将联合产生 m*n 的矩阵。矩阵中，WITH 的变量是横列变量 (Row)，VAR 的变量是纵行变量(Column)。然而，若只选用 VAR 指令而省略 WITH 指令，所产生的则是 n*n 的正方对称矩阵。

程序	产生的积差相关矩阵
PROC CORR; VAR A B C ;	rAA rAB rAC rBA rBB rBC rCA rCB rCC

```
PROC CORR;                    rXA rXB rXC
     VAR A B C;               rYA rYB rYC
     WIHT X Y Z;              rZA rZB rZC
```

指令 #4　PARTIAL 变量名称串；

这个指令用来计算净相关系数，与选项 PEARSON，SPEARMAN，KENDALL 等联用。其目的在于将 PARTIAL 指令中提到的变量对 VAR 或 WITH 的变量之影响力排除。如此，所求得的相关系数就是净相关系数。当选用此指令时，含遗漏数据的观察体不列入计算中。

指令 #5　WEIGHT 变量名称；

本指令使一个观察体根据 WEIGHT 变量的值被重复多次使用。换句话说，一个观察体被当成数个观察体而进行分析，也因此而导出加权的积差相关系数。

本指令可与 PEARSON 选项合用，产生加权的积差相关系数。但本指令不可与 SPEARMAN，KENDALL 或 HOEFFDING 选项联用。

指令 #6　FREQ 变量名称；

FREQ 变量的值表示观察体重复出现的次数或加权值的大小。当 CORR 程序在检定相关系数的显著水准时，观察体的总数就是 FREQ 变量值的总和。

FREQ 变量与前述的 WEIGHT 变量，其作用有异有同。相同之处是加权的相关系数之计算一样。相异之处是系数显著性检验所用的自由度不同。WEIGHT 变量所用的自由度是根据原有观察体个数而得的，FREQ 变量所用的自由度则是根据加权后的观察体个数为准。此外，两者还有一个相异之处：WEIGHT 变量只适用于积差相关分析 (即 PEARSON)，而 FREQ 变量则可通用于其他相关系数的计算上 (如：SPEARMAN，KENDALL 或 HOEFFDING)。

指令 #7　BY 变量名称串；

这个指令的目的是将原输入资料文件按 BY 变量的值分成几个小资料文件。然后在每个小资料文件内分别进行相关系数的计算。当选用此指令时，资料内的数据必须先按照 BY 变量串的值，重新做由小到大的排列，这个步骤可借 PROC SORT 达成。

下举两例说明 BY 指令与 SORT 程序的关系。在第一个例子中，观察体已经按照 BY 指令中的 GRP 变量值作由小而大的排列，GRP=1 的观察体呈现于 GRP=2 的观察体之前。CORR 程序将分别就这两个分组 (GRP=1，GRP=2) 进行变量 V1 至 V3 的积差相关分析：

```
DATA A;
     INPUT GRP V1-V3 @@;
     CARDS;
1 80 90 70 1 70 60 60
1 80 80 70 1 90 70 90
```

```
2 70 60 80 2 55 65 70
2 60 80 70 2 90 80 70
;
PROC CORR;
    VAR V1-V3; BY GRP;
```

如果观察体没有依 BY 指令中变量的值作由小到大的排列，那么，应先用 PROC SORT 整理这个资料文件。请参考下例：

```
DATA A;
     INPUT GRP V1-V3 @@;
     CARDS;
1 80 90 70 2 70 60 80
1 80 80 70 2 60 80 70
1 70 60 60 2 55 65 70
1 90 70 90 2 90 80 70
;
PROC SORT;
     BY GRP;
PROC CORR;
     VAR V1-V3; BY GRP;
```

7.3 范　　例

例一：一个不精简的例子

这一程序的撰写并不算精简，因为 PEARSON 及 RANK 两选项是多余的。选项 OUTP= 本身已具有执行 PEARSON 分析的功能。故 PEARSON 选项是多写的。

而 RANK 选项只对报表输出文件有影响，对矩阵输出文件则无效。而 NOPRINT 选项则抑制报表不使其印出。故 RANK 选项也是多写的。

最后利用 PRINT 程序，将 COR 资料文件印出。

程　序

```
DATA A;
     INPUT AGE VAR1-VAR5;
     CARDS;
17.25       83  131     59  90  71
17.66       96  140     59  96  62
13.57       78  66      51  49  65
13.10       90  75      50  44  56
14.85       63  50      44  46  70
14.58       58  60      43  43  74
```

```
17.00        84  93  56  61  66
13.00        99  71  53  47  54
14.52        124 113 73  71  59
15.04        97  76  57  62  59
14.42        61  55  32  33  52
14.84        70  55  30  33  44
13.51        62  63  35  36  56
13.79        70  60  36  37  52
14.08        64  67  33  36  51
13.74        57  62  29  31  51
14.99        86  78  45  49  53
14.51        101 108 51  60  51
13.06        61  57  40  37  66
13.45        48  45  26  25  55
;
PROC CORR PEARSON RANK NOPRINT OUTP=COR;
     VAR AGE VAR1-VAR5;
PROC PRINT DATA=COR;
RUN;
```

结　果

报表 7.1　一个不精简的例子

OBS	_TYPE_	_NAME_	AGE	VAR1	VAR2	VAR3	VAR4	VAR5
1	MEAN		14.5480	77.6000	76.2500	45.1000	49.3000	58.3500
2	STD		1.3577	19.4135	26.9207	12.4980	19.0818	8.0608
3	N		20.0000	20.0000	20.0000	20.0000	20.0000	20.0000
4	CORR	AGE	1.0000	0.2879	0.7181	0.4613	0.7839	0.3641
5	CORR	VAR1	0.2879	1.0000	0.7162	0.8802	0.6945	-0.0213
6	CORR	VAR2	0.7181	0.7162	1.0000	0.7683	0.9420	0.2271
7	CORR	VAR3	0.4613	0.8802	0.7683	1.0000	0.8434	0.4463
8	CORR	VAR4	0.7839	0.6945	0.9420	0.8434	1.0000	0.4345
9	CORR	VAR5	0.3641	-0.0213	0.2271	0.4463	0.4345	1.0000

例二：使用四次不同的 CORR 程序来分析同一组资料

在第一个 CORR 程序中因 VAR 与 WITH 指令同时联用，我们将得到一个 1*5 的矩阵 (见报表 7.2a)。这个分析的结果应包含：

●VAR 与 WITH 指令内所列举的六个变量的描述性统计值。

●一个 1*5 的积差相关系数矩阵及其显著性的矩阵。

●一个 1*5 的 tau-b 系数矩阵及其显著性的矩阵。

在第二个 CORR 程序中使用相同的资料文件(A)，但将 WITH 指令中的变量合并于 VAR 指令里，于是将产生一个 6*6 的矩阵 (见报表 7.2b)。

在第三个 CORR 程序中，使用相同的资料文件 (A)。并在 PROC CORR 指令中选

用 RANK 选项，使积差相关矩阵依相关系数的大小，顺序印出 (见报表 7.2c)。读者若仔细比较报表的 6*6 矩阵与报表 7.26 的矩阵，将不难发现此处的矩阵将每一横列重新定义过 RANK 指令的效果。由此我们可以迅速地看出与 AGE 相关最高的是 VAR4 与 VAR1 最高的是 VAR3 等等。

第四个 CORR 程序旨在示范 PARTIAL 指令的功能。 在此， 我们将 AGE 的效果自 VAR1－VAR5 两相关中剔除。因此所得的相关系数是皮尔森的净相关系数(见报表 7.2d)。

程　序

```
DATA A;
        INPUT AGE VAR1-VAR5;
        CARDS;
 17.25   83 131 59  90  71
 17.66   96 140 59  96  62
 13.57   78  66 51  49  65
 13.10   90  75 50  44  56
 14.85   63  50 44  46  70
 14.58   58  60 43  43  74
 17.00   84  93 56  61  66
 13.00   99  71 53  47  54
 14.52   124 113 73 71  59
 15.04   97  76 57  62  59
 14.42   61  55 32  33  52
 14.84   70  55 30  33  44
 13.51   62  63 35  36  56
 13.79   70  60 36  37  52
 14.08   64  67 33  36  51
 13.74   57  62 29  31  51
 14.99   86  78 45  49  53
 14.51  101 108 51  60  51
 13.06   61  57 40  37  66
 13.45   48  45 26  25  55
 ;
 PROC CORR KENDALL PEARSON;
      VAR VAR1-VAR5;
      WITH AGE;
 PROC CORR KENDALL PEARSON ;
     VAR AGE VAR1-VAR5;
 PROC OCRR PEARSON RANK;
     VAR VAR1-VAR5;
     PARTIAL AGE;
RUN;
```

结　果

报表 7.2a　PROC CORR 程序中 VAR 与 WITH 指令的示例

```
                              CORRELATION ANALYSIS

1     'WITH'    Variables    AGE
                     :
5     'VAR'     Variables    VAR1     VAR2     VAR3     VAR4     VAR5
                     :

                              Simple Statistics

Vari            N        Mean     Std Dev      Median     Minimum      Maximum
able
AGE            20     14.5480     1.35771    14.46500    13.00000     17.66000
VAR1           20    77.60000    19.41351    74.00000    48.00000    124.00000
VAR2           20    76.25000    26.92069    66.50000    45.00000    140.00000
VAR3           20    45.10000    12.49800    44.50000    26.00000     73.00000
VAR4           20    49.30000    19.08182    45.00000    25.00000     96.00000
VAR5           20    58.35000     8.06079    56.00000    44.00000     74.00000
Pearson Correlation Coefficients / Prob > |R| under Ho:Rho=0 / N = 20
              VAR1         VAR2         VAR3         VAR4         VAR5
AGE        0.28789      0.71813      0.46133      0.78390      0.36411
            0.2184       0.0004       0.0406       0.0001       0.1145
Kendall Tau b Correlation Coefficients / Prob > |R| under Ho:Rho=0 / N = 20
              VAR1         VAR2         VAR3         VAR4         VAR5
AGE        0.21164      0.29630      0.32805      0.43619      0.16625
            0.1939       0.0689       0.0440       0.0077       0.3126
```

报表 7.2b　PROC CORR 程序中 VAR 指令的示例

```
                              CORRELATION ANALYSIS

6'VAR'Variables:    AGE      VAR1      VAR2      VAR3      VAR4      VAR5

                              Simple Statistics

Variable                N        Mean     StdDev      Median     Minimum      Maximum
AGE                    20    14.54800    1.35771    14.46500    13.00000     17.66000
VAR1                   20    77.60000   19.41351    74.00000    48.00000    124.00000
VAR2                   20    76.25000   26.92069    66.50000    45.00000    140.00000
VAR3                   20    45.10000   12.49800    44.50000    26.00000     73.00000
VAR4                   20    49.30000   19.08182    45.00000    25.00000     96.00000
VAR5                   20    58.35000    8.06079    56.00000    44.00000     74.00000

     Pearson Correlation Coefficients / Prob > |R| under Ho:Rho=0 / N = 20
                    AGE         VAR1         VAR2         VAR3         VAR4         VAR5

     AGE        1.00000      0.28789      0.71813      0.46133      0.78390      0.36411
                    0.0      0.2184       0.0004       0.0406       0.0001       0.1145

     VAR1       0.28789      1.00000      0.71622      0.88022      0.69453     -0.02126
                 0.2184          0.0       0.0004       0.0001       0.0007       0.9291

     VAR2       0.71813      0.71622      1.00000      0.76831      0.94204      0.22708
                 0.0004       0.0004          0.0       0.0001       0.0001       0.3357

     VAR3       0.46133      0.88022      0.76831      1.00000      0.84335      0.44631
                 0.0406       0.0001       0.0001          0.0       0.0001       0.0485

     VAR4       0.78390      0.69453      0.94204      0.84335      1.00000      0.43453
                 0.0001       0.0007       0.0001       0.0001          0.0       0.0556

     VAR5       0.36411     -0.02126      0.22708      0.44631      0.43453      1.00000
                 0.1145       0.9291       0.3357       0.0485       0.0556          0.0
```

Kendall Taub Correlation Coefficients / Prob > |R|under Ho：Rho=0 / N=20

	AGE	VAR1	VAR2	VAR3	VAR4	VAR5
AGE	1.00000	0.21164	0.29630	0.32805	0.43619	0.16625
	0.0	0.1939	0.0689	0.0440	0.0077	0.3126
VAR1	0.21164	1.00000	0.57447	0.61702	0.59894	-0.02696
	0.1939	0.0	0.0004	0.0002	0.0003	0.8705
VAR2	0.29630	0.57447	1.00000	0.62766	0.68985	0.13478
	0.0689	0.0004	0.0	0.0001	0.0001	0.4149
VAR3	0.32805	0.61702	0.62766	1.00000	0.90910	0.37200
	0.0440	0.0002	0.0001	0.0	0.0001	0.0244
VAR4	0.43619	0.59894	0.68985	0.90910	1.00000	0.35232
	0.0077	0.0003	0.0001	0.0001	0.0	0.0339
VAR5	0.16625	-0.02696	0.13478	0.37200	0.35232	1.00000
	0.3126	0.8705	0.4149	0.0244	0.0339	0.0

报表 7.2c　PROC CORR 程序中 RANK 指令的示例

CORRELATION ANALYSIS

6'VAR'Variables：　AGE　VAR1　VAR2　VAR3　VAR4　VAR5

Simple Statistics

Variable	N	Mean	StdDev	Sum	Minimum	Maximum
AGE	20	14.54800	1.35771	290.96000	13.00000	17.66000
VAR1	20	77.60000	19.41351	1552	48.00000	124.00000
VAR2	20	76.25000	26.92069	1525	45.00000	140.00000
VAR3	20	45.10000	12.49800	902.00000	26.00000	73.00000
VAR4	20	49.30000	19.08182	986.00000	25.00000	96.00000
VAR5	20	58.35000	8.06079	1167	44.00000	74.00000

PearsonCorrelationCoefficients/Prob>|R|underHo：Rho=0/N=20

AGE

AGE	VAR4	VAR2	VAR3	VAR5	VAR1
1.00000	0.78390	0.71813	0.46133	0.36411	0.28789
0.0	0.0001	0.0004	0.0406	0.1145	0.2184

VAR1

VAR1	VAR3	VAR2	VAR4	AGE	VAR5
1.00000	0.88022	0.71622	0.69453	0.28789	-0.02126
0.0	0.0001	0.0004	0.0007	0.2184	0.9291

VAR2

VAR2	VAR4	VAR3	AGE	VAR1	VAR5
1.00000	0.94204	0.76831	0.71813	0.71622	0.22708
0.0	0.0001	0.0001	0.0004	0.0004	0.3357

VAR3

VAR3	VAR1	VAR4	VAR2	AGE	VAR5
1.00000	0.88022	0.84335	0.76831	0.46133	0.44631
0.0	0.0001	0.0001	0.0001	0.0406	0.0485

VAR4

VAR4	VAR2	VAR3	AGE	VAR1	VAR5
1.00000	0.94204	0.84335	0.78390	0.69453	0.43453
0.0	0.0001	0.0001	0.0001	0.0007	0.0556

VAR5

VAR5	VAR3	VAR4	AGE	VAR2	VAR1
1.00000	0.44631	0.43453	0.36411	0.22708	-0.02126
0.0	0.0485	0.0556	0.1145	0.3357	0.9291

报表 7.2d　PROC CORR 程序中 PARTIAL 指令的示例

Correlation Analysis

1　'PARTIAL'　Variables:　AGE
5　'VAR'　Variables:　VAR1　VAR2　VAR3　VAR4　VAR5

Simple Statistics

Variable	N	Mean	Std Dev	Sum	Minimum	Maximum	Partial Variance 净变异数	Partial Std Dev 净标准差
AGE	20	14.5480	1.3577	290.9600	13.0000	17.6600		
VAR1	20	77.6000	19.4135	1552	48.0000	124.0000	364.8513	19.1011
VAR2	20	76.2500	26.9207	1525	45.0000	140.0000	370.4798	19.2479
VAR3	20	45.1000	12.4980	902.0000	26.0000	73.0000	129.7882	11.3925
VAR4	20	49.3000	19.0818	986.0000	25.0000	96.0000	148.1638	12.1723
VAR5	20	58.3500	8.0608	1167	44.0000	74.0000	59.4930	7.7132

Pearson Partial Correlation Coefficients / Prob > |R| under Ho: Partial Rho=0 / N = 20

VAR1

VAR1	VAR3	VAR4	VAR2	VAR5
1.00000	0.87965	0.78852	0.76447	-0.14136
0.0	0.0001	0.0001	0.0001	0.5638

VAR2

VAR2	VAR4	VAR1	VAR3	VAR5
1.00000	0.87738	0.76447	0.70779	-0.05308
0.0	0.0001	0.0001	0.0007	0.8291

VAR3

VAR3	VAR1	VAR4	VAR2	VAR5
1.00000	0.87965	0.87447	0.70779	0.33684
0.0	0.0001	0.0001	0.0007	0.1585

VAR4

VAR4	VAR2	VAR3	VAR1	VAR5
1.00000	0.87738	0.87447	0.78852	0.25784
0.0	0.0001	0.0001	0.0001	0.2865

VAR5

VAR5	VAR3	VAR4	VAR1	VAR2
1.00000	0.33684	0.25784	-0.14136	-0.05308
0.0	0.1585	0.2865	0.5638	0.8291

例三：阿尔法系数的计算

这个例子旨在示范如何利用 CORR 程序以便导出一个测验的阿尔法系数。在心理与教育测验的领域里，阿尔法系数被当作是信度的指标，其值的大小表示格测验值（或子测验）与总测试分数之间相关的程度。其计算公式如下：

$$\text{阿尔法系数}（\alpha）=（\frac{n}{n-1}）（1-\frac{\text{各测验题变异数的和}}{\text{总测验的变异数}}）$$

在此，n=测验题数。

下面的程序分两部分：第一部分界定一个 SAS 资料文件“ACHIEVE”，其变量值由磁盘上 DATA 子目录下的 ACHIEVE.DAT 引进。这个资料文件含六个子测验的分数，分别是：VOCAB（词汇），READING（阅读），SPELLING（拼音），CAPITAL（大小写），PUNC（标点符号），USAGE（文法）等。这些子测验均与语文的智力测验有关。因此，

程序的第二部分（亦即 CORR 程序部分）界定 ALPHA 选项以便计算这六个子测验与其总分和相关程度。另一选项 NOSIMPLE 的界定是为了免去任何有关这六个子测验的描述性统计值（如平均数、变异数、标准差等）以节省报表的空间。

程　序

```
OPTIONS LS=80 PAGENO=1 NODATE;
TITLE  'EXAMPLE 7.3 CRONBAH ALPHA';
[第一部分]
DATA ACHIEVE;
    INFILE 'A:\DATA\ACHIEVE.DAT';
    INPUT IV1 1 GRADE 2 IV2 3 SEX 4 ID 6-8 VOCAB 25-26 READING 27-28
        SPELLING 29-30 CAPITAL 31-32 PUNC 33-34 USAGE 35-36/;
[第二部分]
PROC CORR DATA=ACHIEVE ALPHA NOSIMPLE;
    VAR VOCAB READING SPELLING CAPITAL PUNC USAGE;

RUN;
```

结　果

报　表 7.3　阿尔法系数的计算

Example 7.3 Cronbach alpha　1

Correlation Analysis

6 'VAR' Variables: VOCAB READING SPELLING CAPITAL PUNC USAGE

Example 7.3 Cronbach alpha　2

[第一部分]

Correlation Analysis

Cronbach Coefficient Alpha

for RAW variables　: 0.901731

for STANDARDIZED variables: 0.903575

[第二部分]

	Raw Variables		Std. Variables	
Deleted Variable	Correlation with Total	Alpha	Correlation with Total	Alpha
VOCAB	0.658981	0.895441	0.665012	0.896787
READING	0.726066	0.886880	0.730951	0.887104
SPELLING	0.749084	0.881637	0.751982	0.883964
CAPITAL	0.774339	0.877696	0.763269	0.882268
PUNC	0.696889	0.890531	0.684252	0.893986
USAGE	0.815621	0.871280	0.818848	0.873814

[第三部分]

Pearson Correlation Coefficients / Prob > |R| under Ho: Rho=0 / N = 119

	VOCAB	READING	SPELLING	CAPITAL	PUNC	USAGE
VOCAB	1.00000	0.64250	0.62536	0.49198	0.38139	0.65616
	0.0	0.0001	0.0001	0.0001	0.0001	0.0001
READING	0.64250	1.00000	0.60435	0.57976	0.52107	0.68578
	0.0001	0.0	0.0001	0.0001	0.0001	0.0001
SPELLING	0.62536	0.60435	1.00000	0.62953	0.58077	0.66730
	0.0001	0.0001	0.0	0.0001	0.0001	0.0001
CAPITAL	0.49198	0.57976	0.62953	1.00000	0.75037	0.69504
	0.0001	0.0001	0.0001	0.0	0.0001	0.0001
PUNC	0.38139	0.52107	0.58077	0.75037	1.00000	0.63337
	0.0001	0.0001	0.0001	0.0001	0.0	0.0001
USAGE	0.65616	0.68578	0.66730	0.69504	0.63337	1.00000
	0.0001	0.0001	0.0001	0.0001	0.0001	0.0

报表的第一页只含 PROC CORR 的标题而无结果，这是因为在程序中界定 NOSIMPLE 选项所致。

第 2 页的分析结果可分三部分解释如下：

第一部分含两个阿尔法系数：第一个值（0.901731）是根据原始分数导出的，第二个值 (0.903575) 则根据标准化后的分数导出。二者的值都相当高，因此，我们可下结论说：由这六个子测验分数的总和所形成的语文智力测验显示极高的可信度。

报表的第二部分含各子测验与总测验的相关系数。就 USAGE（文法）而言，该子测验与总分的相关最高（r=0.815621 或 r=0.818848，根据标准化分数）。因此，若总测试中不包括此子测验，则其阿尔法信度简为 0.871280（根据原始分数）或 0.73814（根据标准化分数）。

报表的第三部分是一个 6×6 的皮尔森系数矩阵，其元素为六个子测验两两相关的系数以及其统计检定的显著程度。此矩阵在前面两个示范的例题中已解释过，故不再赘述。

注：阿尔法系数只有在一种情况下会达到最大值（1.00）:就是当测验题目（或子测验）间两两的相关系数是正的 100%。若某些题目（或子测验）间的相关是负的，则阿尔法系数的值可能会低于零。一般而言，一个测验的可信度最好在 0.8 以上。

7.4　注 意 事 项

■ 输出资料文件的进一步说明

透过 PROC CORR 指令内的 OUTP=，OUTS=，OUTK= 及 OUTH=选项，可分别产生积差相关系数，等级相关系数，tau-b，及 Hoeffding's D 相关系数的矩阵输出文件。各输出文件中除了包含上述的相关系数外，还有各变量的平均数，标准差及观察体个数等。

关于矩阵输出文件的内容，有下列几点说明：

(1) 输出文件内的变量

- ●系统变量 _TYPE_，指出资料的性质
- ●系统变量 _NAME_，指出横列变量的名称
- ●VAR 指令中所列举的变量，指纵行变量而言。

系统变量 _TYPE_ 可能有下列几种值：

- ●MEAN，　各变量的平均数
- ●STD，　各变量的标准差
- ●N，　参与各变量计算过程的有效观察体个数
- ●SUMWGT，　变量的加权值
- ●SSCP，　离差平方和以及离差内乘积的联合矩阵
- ●COV，　共变异数矩阵
- ●CORR，　相关系数矩阵

(2) 命名

若读者希望将输出文件储存成永久性的磁盘文件，则必须以二段式文件名命名。否则，一段式文件名即可。

第 8 章　一般制图：统计程序 PROC PLOT

8.1　PROC PLOT 程序概述

PROC PLOT 的功能是将观察体在两个变量上的值视为坐标值，然后用二维空间的图形描出这些观察体相对的位置。从这个图形里，我们可以检视两变量间相互的关系，图形上空白区域的意义或经过回归分析后预测值与实际值相近的程度等。

在实际制图的过程中，读者可自由选择：

甲、放大或缩小 X，Y 轴；

乙、在 X 或 Y 轴上画参考线；

丙、描点的符号；

丁、为文字变量的值绘图；

戊、将两个以上的图形重叠在一起；

己、把纵轴 (亦即 Y 轴) 颠倒过来，也就是把 Y 轴的值由小 (上) 至大 (下) 打印在坐标轴上；

庚、改变 X，Y 轴的起点值与终点值；

辛、将一个三维空间的图形以轮廓图 (Contour) 的方式呈现。

8.2　如何撰写 PROC PLOT 程序

PROC PLOT 含三道指令，它们的格式如下：

```
PROC  PLOT    选项串;
      BY      变量名称串;
      PLOT    图形指令串 / 选项串;
```

指令 #1 PROC PLOT 选项串;

此指令的选项有八，下面分别介绍之：

(1) DATA=输入资料文件

指明 SAS 利用哪一个资料文件的变量值制图。若省略此选项，则 SAS 会自动找出在此程序之前最后形成的 SAS 资料文件，取其变量值描点。

(2) UNIFORM

与指令 BY 合用。BY 指令的作用是将资料文件分成几个小的资料文件，然后在每一个小的资料文件内分别绘图。利用 UNIFORM 选项，读者可使各个图形的 X 轴与 Y 轴的长短、单位完全一致。如此的控制有利于读者比较从各小资料文件

里得到的图形之异同。

(3) NOLEGEND

要求 PLOT 程序在图形的上端不印出变量的名字或绘图的符号。

(4) NOMISS

要求 PLOT 程序将在 X 轴或 Y 轴变量上有遗漏数据的观察体自坐标轴的刻度中剔除。若不选用此选项，则观察体只要含至少一个有效数据，PLOT 程序仍会将其纳入两个坐标轴的刻度中。当读者选用 VAXIS 或 HAXIS 的选项时，此选项无效。

(5) FORMCHAR (符号的位置，以 1 到 11 的数字表示) = '十一个绘图符号'

PLOT 程序利用这个指令来控制绘图所需的符号。这些符号最多可以有十一个，它们所对应的位置也有十一个。下面简单地说明这些位置、数字代号及其内设的符号：

符号的位置	数字代号	内设的符号
纵轴	1	|
横轴	2	—
左上角	3	—
中上方	4	—
右上角	5	—
中心点的左边	6	|
中心点	7	+
中心点的右边	8	|
左下角	9	—
中下方	10	—
右下角	11	—

在 PLOT 程序中我们只须定义 1，2，3，5，7，9，11 等符号位置即可。如果读者想用星号来表示四个角时，可利用下列指令：

```
FORMCHAR(3 5 9 11)='****'
```

若将前述选项改写成：

```
FORMCHAR='          '   (引号内含十一个空白)
```

则报表上的图表就只含文字或统计值，不含任何的线条或符号。若是在 IBM 公司出品含文字码 (1 或 2 型) 之打印机上打印时，则下列的定义最理想：

```
FORMCHAR='B3C4DAC2BFC3C5B4C0C1D9'X
```

(6) VTOH=正实数

此正实数界定纵轴对横轴画图符号的比例。当读者界定此选项后，PLOT 程序会自动调整符号间的距离，使两轴坐标单位的长、宽几乎相等。若读者另外选用 VSPACE 或 HSPACE 选项，则此选项无效。

(7) VPRECENT(或 VPCT)=报表纸长度分割的百分比

此选项决定一张报表纸自上端至下端长度分割的比例 (以百分比表示)。比方说：

```
VPCT=33
```

则 PLOT 程序将利用每一页报表纸上端的三分之一来描绘个别的图形。若依下列的指令：

```
VPCT=50 25 25
```

则 PLOT 程序会在同一页纸上绘出不重叠的三个图，其中第一图占整页纸长度的一半，第二与第三图则各占四分之一。若指令改写成：

```
VPCT=75  0
```

则 PLOT 程序自动将第一个图印在报表纸上端四分之三的地方，第二个图则印在下一页纸上 (亦占全页的四分之三)。最后，若图形太长，须将几页纸合并印图，则可使百分比的值大于 100，如：

```
VPCT=200
```

如此 PLOT 程序会将两页的报表纸合并，以便容纳一个较长的图形。

(8) HPERCENT(或 HPCT)=报表纸宽度分割的百分比

此指令的写法与上述选项 VPERCENT= 完全相同，只是当百分比的值大于 100 时，PLOT 程序会将图形宽度分割，分别印在前后两页纸上。

指令 #2　BY 变量名称串；

此指令的目的在于将原资料文件分成几个小资料文件，然后在每个小资料文件内分别制图。当读者选用这道指令时，原资料文件的数据必须依照 BY 变量串的值由小到大重新排列过，这个步骤可借 PROC SORT 来达成。

指令 #3　PLOT 图形指令串 / 选项串；

这个指令直接控制报表上图形的呈现方式，有 "图形指令串" 与 "选项串" 两种控制语句。现分别解释如下：

图形指令串

图形指令串界定 X，Y 轴以及绘图的符号。它的语法不外乎下面三种格式：

图形指令串的格式	举　例
(1) Y 轴之变量名 * X 轴之变量名；	PROC PLOT; PLOT GRADE*IQ;
(2) Y 轴之变量名 * X 轴之变量名='符号'；	PROC PLOT; PLOT Y*X='+';
(3) Y 轴之变量名 * X 轴之变量名 =含符号之变量名称；	PROC PLOT; PLOT HEIGHT*WEIGHT=SEX;

若读者选第一种格式制图，则图形上的点以英文大写字母表示：A 代表一点，B 代表两点，...，Z 代表二十六个或二十六个以上的点的重叠。

若读者选第二种格式制图，则图形上所有的点均以同一个符号来表示。这个符号也就是读者在单引号内所界定的符号。

若读者选第三种格式制图，则图形上的点以含符号的变量值来表示。 比方说，在 (3) 的例子中，我们界定 SEX (性别) 为含符号的变量。若 SEX (性别) 变量下有 FEMALE (女生) 与 MALE (男生) 两个值，则图形上女生的数据会以 F 来表示，而男生的数据则以 M 代表。若男、女生的数据完全相同，然而只能用一个符号来代表二者，则排在前面的观察体决定共用的符号。 因此，如果男生的资料排在女生前面，那么男女重叠的点将以 M 来表示。

采用用第三种格式的另一个结果是：重叠点的个数无法以不同的符号再表示。这个限制是格式 (2) 与 (3) 共同的限制；然而，格式 (1) 则没有这个问题。

此外，PLOT 指令可以界定一个以上的图形，如：

```
PROC PLOT;
     PLOT A*B Y*X;
```

表示两个图形分别由 A*B 及 Y*X 来定义。

当读者需要变量间的各种排列组合以便绘图时，不妨采用用下列较精简的表示法：

PLOT (Y X)*(A B); 等于　PLOT Y*A Y*B X*A X*B;

PLOT Y*(A--C);　　等于　PLOT Y*A Y*B Y*C;

选项串

删除号(／)之后的选项分六类：第一类选项界定横轴与纵轴的长短以及单位，第二类选项界定参考线，第三类选项控制整个图形的大小，第四类选项与重叠的图形有关，第五类选项界定轮廓图的有关事宜，第六类选项调整图形在报表上的打印。下面就这六种类别分别介绍选项的名称与功能。

第一类选项　界定横轴与纵轴的长短以及单位：

(1) VAXIS=纵轴的单位

此选项界定纵轴的单位坐标，如：

```
PROC PLOT;
     PLOT Y*X / VAXIS=10 TO 100 BY 5;
```

根据这个写法，Y 轴上的坐标单位会是 10，15，20，...，100 等。坐标单位的值不一定要以等值累加，如：

```
VAXIS=10 100 1000 10000;
```

根据这样的界定，读者将会得到一个以 10 为底的对数函数图。除了以数字为坐标的单位之外，读者也可用文字定义纵轴，如：

```
VAXIS='01JAN89'D TO '01JAN90'D BY MONTH; 或者
VAXIS='01JAN89'D TO '01JAN90'D BY QTR;
```

其中 D 代表天 (DAY)。第一种的写法会产生十三个单位坐标的点。然而，第二种的写法只产生四个单位坐标的点 (每三个月为一季，故一年内有四个季节) 。

有关时间单位的界定，读者可参阅 INTCK 以及 INTNX 两个 SAS 内设函数的写法，或借 SASTUTOR 程序中 HELP SAS Functions 的指令学习其撰写规则（见附录 A）。

(2) HAXIS=横轴的单位

这个选项的写法与选项 VAXIS= 完全相同，故不另赘述。

(3) VZERO

要求纵轴的坐标以 0 开始。若读者已经用选项 VAXIS= 界定横轴的坐标单位或数据中含负的纵轴坐标，则 VZERO 选项会被忽略。

(4) HZERO

要求横轴的坐标以 0 开始。若读者已经用选项 HAXIS=界定横轴的坐标单位或数据中含负的横轴坐标，则 HZERO 选项会被忽略。

(5) VREVERSE

将纵轴的坐标单位颠倒过来，亦即将最小的值印在纵轴的最顶点，最大的值印在原点的位置。

第二类选项　界定参考线：

(1) VREF=纵轴上的坐标

PROC PLOT 根据这个选项所界定的 Y 坐标，画一条与 X 轴平行的参考线。

(2) VREFCHAR='参考线的符号'

用来画参考线的记号。这个符号可以是键盘上任何一个键所代表的符号，内设值是减号(-)。

(3) HREF=横轴上的坐标

PROC PLOT 根据这个选项所界定的 X 坐标，画一条与 Y 轴平行的参考线。

(4) HREFCHAR='参考线的符号'

用来画参考线的记号。这个符号可以是键盘上任何一个键所代表的符号，内设值是竖号（|）。

第三类选项　控制整个图形大小的选项：

(1) VPOS=图形的宽度，以正整数表示

这个选项的最大值必须比报表实际的宽度少八行。

(2) HPOS=图形的长度，以正整数表示

必须预留最顶端的三行当作图形的标题。

界定纵、横轴，行、列的选项：

(3) VSPACE=正整数(如 5)

界定纵轴上坐标单位间的列数 (Print Lines)，如五列。

(4) HSPACE=正整数(如 4)

界定横轴上坐标单位间的行数 (Print Positions)，如四行。

第四类选项　界定图形的重叠：

(1) OVERLAY

此选项要求将两个或两个以上的图形重叠在一起。这个重叠的图画会以第一个图

的变量来定义 X 轴与 Y 轴。当读者选择此选项时，最好在 SAS 程序的起首宣告图形的重叠打印。这个宣告的语句是：

```
OPTIONS=OVP;
```

如此宣告之后，重叠图上即使有一点代表不止一个图形的数据，它的绘图符号也将揉合原来图形的所有绘图符号。否则，这类数据将以第一个图所用的符号来表示。

第五类选项　界定轮廓图的有关事宜：

(1) CONTOUR=(1 到 10 的一个整数)

界定轮廓图深浅的层次，如：

```
PROC PLOT;
PLOT Y*X=Z / CONTOUR=10;
```

根据这个程序，轮廓图的深浅度由变量 Z 值的大小决定。选项 CONTOUR=10 则要求将 Z 值的大小分成十类，每类的数据以不同深浅度的符号描绘在轮廓图上。

(2) S1='代表最浅度的符号'

S2='代表次浅度的符号'

.

.

请看下面例子：

```
PROC PLOT;
     PLOT HT*WT=Z/CONTOUR=3
          S1='A'S2='+'S3='XOA';
```

这个例子所界定的轮廓图有三层深浅度 (其深浅度由 Z 变量的值来决定)。最浅一层的数据用 'A' 表示，次浅一层的数据用 '+' 表示，最深一层的数据则以 'XOA' 重复打印来表示。

第六类选项　调整图形在报表上的打印：

(1) BOX

要求 PLOT 程序将整个图形用直线加框，而非只画出 X 轴与 Y 轴的正实数之坐标轴线。

(2) VEXPAND

要求将图形的纵轴增长，以充分利用报表纸的空间。不过，当数据够多且范围够大时，此选项可能不产生任何作用。

一般而言，PLOT 程序会先计算数据里最小的五个数值间最小的差距 (称作 Delta)，然后以 Delta 值决定图形纵轴的列距。

值得注意的是，VEXPAND 选项的作用是扩大图形而非改变图形与报表纸大小的相对比例。

(3) HEXPAND

要求将图形的横轴加宽，作用与上述选项 VEXPAND 完全相同。

8.3　如何在同一页的报表纸上多重绘图

若将指令 PLOT 在同一个 PROC PLOT 下重复使用，则可得多重的图形。这些图形不会彼此重叠，而且 SAS 会使它们同时出现在同一页的报表纸上。有关这种多重绘图的指令撰写，请看例 1 与例 2 的示范。

例 1　将六个图形同时画在一张报表纸上

```
OPTIONS NODATE;
DATA A;
    INPUT X Y Z A B;
    CARDS;
0.0  -2.00  1  0.0  -2.00
0.5  -2.25  2  0.5  -1.25
1.0  -2.00  3  1.0   0.00
1.5  -1.25  4  1.5   1.75
2.0   0.00  5  2.0   4.00
2.5   1.75  1  2.5   6.75
3.0   4.00  2  3.0  10.00
3.5   6.75  3  3.5  13.75
4.0  10.00  4  4.0  18.00
4.5  13.75  5  4.5  22.75
;
PROC PLOT HPERCENT=50 VPERCENT=33;
    TITLE 'MULTIPLE PLOTS PER PAGE';
    PLOT Y*X;
    PLOT Y*X / HAXIS=0 TO 12 BY 2;
    PLOT Y*X='+' / HAXIS=0 TO 12 BY 2;
    PLOT Y*X=Z / HAXIS=0 TO 12 BY 2 BOX;
    PLOT Y*X=Z / HAXIS=0 TO 12 HREF=4 8;
    PLOT Y*X=Z A*B /OVERLAY;
RUN;
```

在 PC 上执行上述程序后，所得的结果如下：

```
                          MULTIPLE PLOTS PER PAGE


     A=1, B=2, etc.  Plot of Y*X.          A=1, B=2, etc.  Plot of Y*X.
 Y |                                   Y |
20 +                                  20 +
10 +                      A  A  A     10 +          AAA
 0 +  A  A  A  A  A  A  A              0 + AAA AAAA
   ---+-------+-------+--------+--       --+----+----+----+---+---+----+-
     0.0      1.5      3.0      4.5         0    2    4    6   8   10   12
                  X                                    X
```

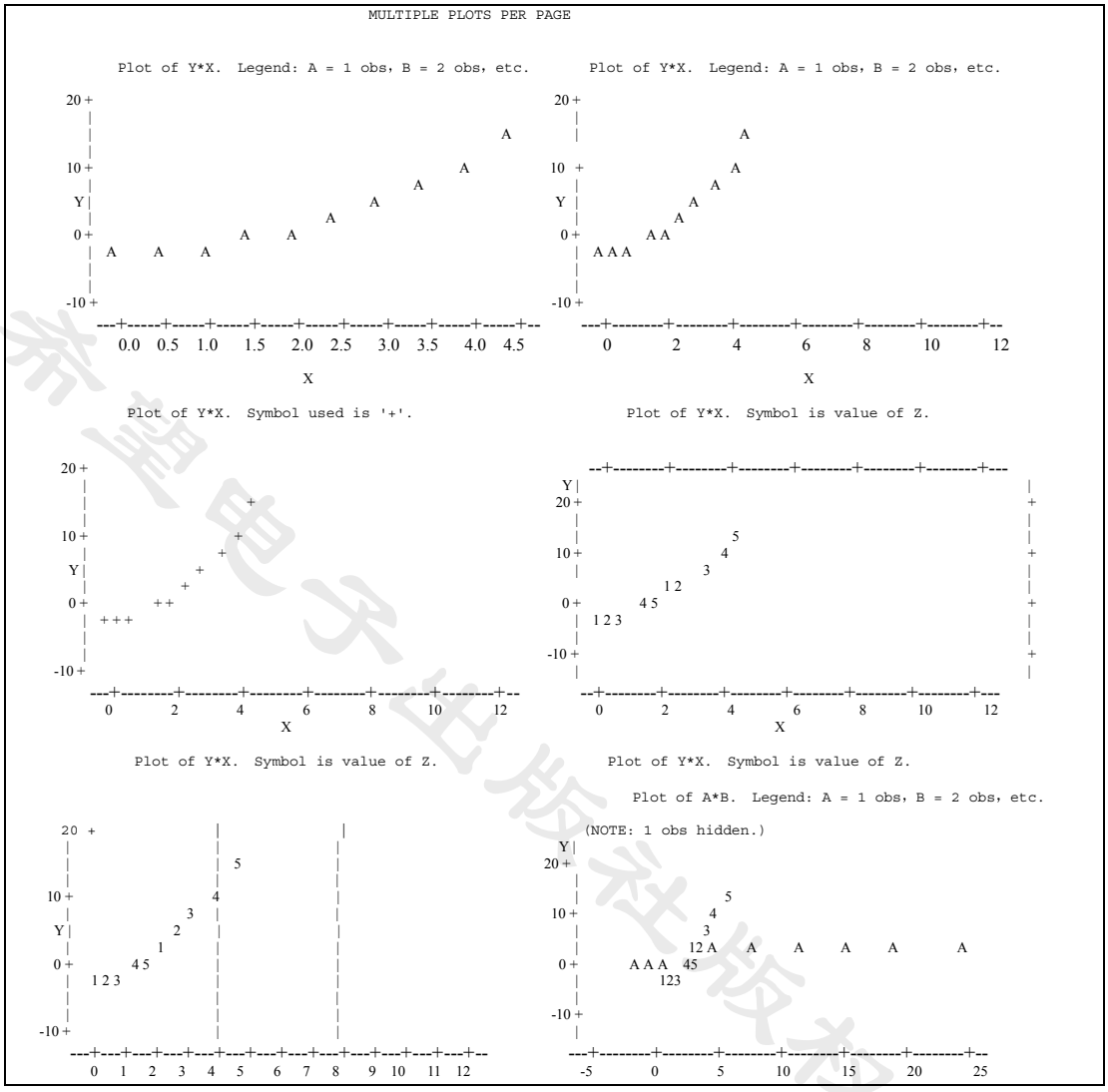
```
 Plot of Y*X.  Symbol used is '+'.    Plot of Y*X.  Symbol is value of Z.
 Y
20 +                                  Y -+----+----+----+----+----+----+--
10 +         +++                     10 +          345                    +
 0 + +++ ++++                         0 +123 4512                          +
   --+----+----+---+----+---+---+-      -+----+----+----+----+----+----+--
     0    2    4    6    8   10   12      0    2    4    6    8   10   12
                  X                                 X

 Plot of Y*X.  Symbol is value of Z.    Plot of Y*X.  Symbol is value of Z.
 Y                                    A=1, B=2, etc.  Plot of A*B.
10 +        345      |                   Y  (NOTE: 7 obs hidden.)
 0 +  1234512 |       |                  10 +          35
  ---+-+-+-+-+-+-+-+-+-+-+-+--           0 +     AA135A A  A  A  A   A
   0 1 2 3 4 5 6 7 8 9 1 1 1              -+-------+-------+------+-------+
                       0 1 2              -10     0      10      20      30
                  X                                 X
```

例 2　将例 1 的图形放大，采用用 LINESIZE=132 的宽度

```
OPTIONS PAGESIZE=66 LINESIZE=132;
DATA A;
    INPUT X Y Z A B;
    CARDS;
0.0   -2.00    1   0.0  -2.00
0.5   -2.25    2   0.5  -1.25
1.0   -2.00    3   1.0   0.00
1.5   -1.25    4   1.5   1.75
2.0    0.00    5   2.0   4.00
2.5    1.75    1   2.5   6.75
3.0    4.00    2   3.0  10.00
3.5    6.75    3   3.5  13.75
4.0   10.00    4   4.0  18.00
4.5   13.75    5   4.5  22.75
;
PROC PLOT HPERCENT=50 VPERCENT=33;
    TITLE 'MULTIPLE PLOTS PER PAGE';
    PLOT Y*X;
    PLOT Y*X / HAXIS=0 TO 12 BY 2;
    PLOT Y*X='+' / HAXIS=0 TO 12 BY 2;
    PLOT Y*X=Z / HAXIS=0 TO 12 BY 2 BOX;
    PLOT Y*X=Z / HAXIS=0 TO 12 HREF=4 8;
    PLOT Y*X=Z A*B /OVERLAY;
RUN;
```

在 PC 上执行上述程序后，所得的结果如下 (鉴于本书版面规格的限制，实际报表的宽度已被浓缩了约四分之一，长度则缩短约十分之一)：



8.4 范　　例

例一：心理系学生在电脑课上期中考与期末考成绩的绘图

在这个例子里，我们用最简单的 PLOT 指令来绘制某大学心理系十一位学生在电脑课上期中考 (MIDTERM) 与期末考 (FINAL) 成绩的相关图形。两次考试的成绩均以一百二十分为满分。图形显示，这两次考试的结果有正相关。

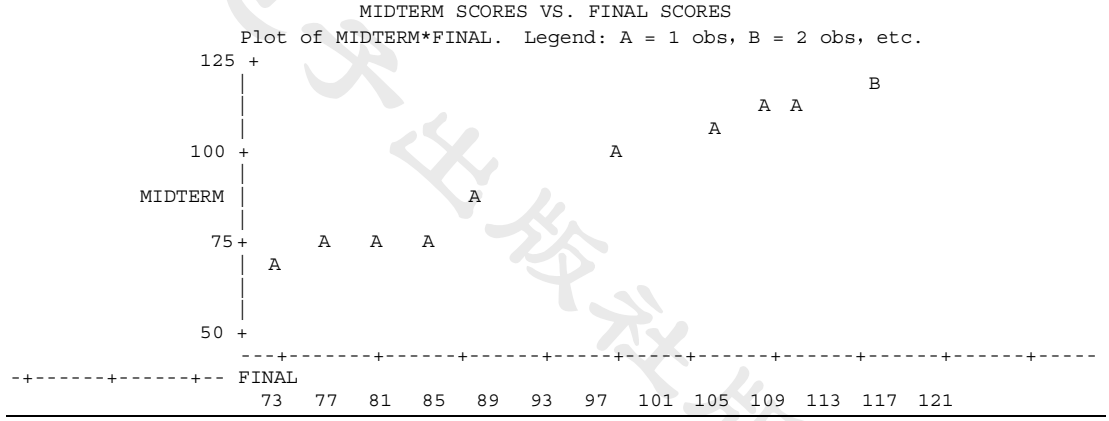
程　序

```
DATA SCORES;
    INPUT MIDTERM FINAL @@;
    CARDS;
```

```
 75  85   90  89
110 112   70  73
 75  81  118 121
105 108  103 100
112 115  118 121
 75  77
;
PROC PLOT;
    PLOT MIDTERM*FINAL;
    TITLE 'MIDTERM SCORES VS. FINAL SCORES';
RUN;
```

结 果

报表 8.1 心理系学生在电脑课上期中考与期末考成绩的绘图



例二：坐标轴起点的控制与绘图符号的选择

这个例子旨在示范如何界定绘图的符号以及坐标轴的单位。数据的来源是一个共变量分析 (ANCOVA)。其中，RESPONSE 代表因变量 (又称反应变量)，NUISANCE 代表共变量。TREAT 则是自变量，分 P，Q，R 三组。绘图的目的是为了检视是否 RESPONSE 与 NUISANCE 的线性关系在三个自变量组内均是一致的。

由于 RESPONSE 的值大部分介于 125 与 185 之间，所以本例的指令采用 VAXIS 选项来控制纵轴起、终点的坐标。

程 序

```
DATA COVAR;
    INPUT TREAT $ RESPONSE NUISANCE @@;
    CARDS;
P 125 3.1   Q 160 12.6
P 135 9.0   Q 165 17.1
P 144 14.9  R 154 3.2
```

```
P 153 20.2   R 164 10.5
Q 136 2.0    R 173 15.4
Q 152 8.5    R 183 20.7
;
PROC PLOT;
    PLOT RESPONSE*NUISANCE=TREAT / VAXIS=125 TO 185 BY 5;
    TITLE 'RESPONSE VS. NUISANCE';
    TITLE2 'THREE TREATMENTS';
RUN;
```

结　果

报表 8.2　坐标轴起点的控制与绘图符号的选择

```
                          RESPONSE VS. NUISANCE
                          THREE TREATMENTS

              Plot of RESPONSE*NUISANCE.  Symbol is value of TREAT.

RESPONSE
   185 +                                                  R
   180 +
   175 +                                 R
   170 +
   165 +                          R                Q
   160 +                                 Q
   155 +        R                                        P
   150 +                    Q
   145 +                                      P
   140 +
   135 +     Q                P
   130 +
   125 +        P
       --+----------+---------+----------+----------+----------+-
         0          5         10         15         20         25
                                  NUISANCE
```

例三：数学公式的图形表示

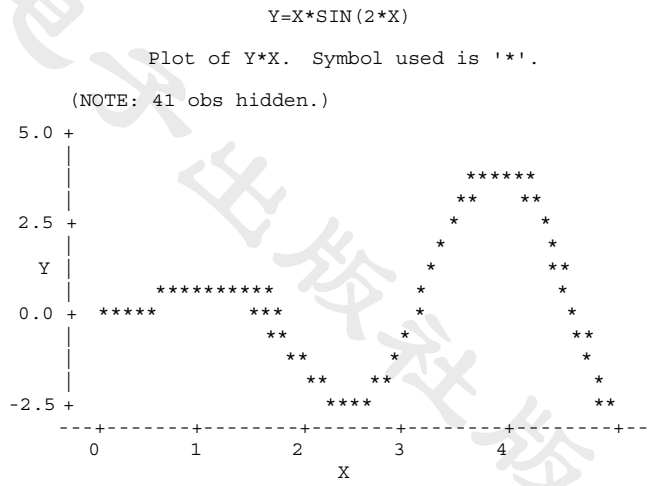
本例利用 SAS 内设的正弦函数产生一组 X,Y 的对应值。然后，将这些数据以 PROC PLOT 的指令再画出来。绘图时，采用 '*' 的符号而非内设的英文大写字母。

程　序

```
DATA PROGRAM;
    DO X=0 TO 5 BY .05;
       Y=X*SIN(2*X);
       OUTPUT;
    END;
PROC PLOT;
    PLOT Y*X='*';
    TITLE 'Y=X*SIN(2*X)';
RUN;
```

结　果

报表 8.3　数学公式的图形表示



例四：将回归分析后的预测值与实际值重叠地绘图

这个例子的数据是有关一组学生的身高 (HEIGHT) 与体重 (WEIGHT)。在绘图之前，先用 PROCREG 求出一组身高的预测值 (称作 PREDHT)。这些预测值是利用体重的值来预测的。因此，我们可将身高的实际值 (以 A，B 等表示) 与预测值 (以 * 表示)对同一组体重值重叠地画在一张图形上。

程　序

```
DATA HTWT;
    INPUT HEIGHT WEIGHT @@;
    CARDS;
69.0  112.5  56.5   84.0   65.3   98.0   62.8  102.5  63.5  102.5
57.3   83.0  59.8   84.5   62.5  112.5   62.5   84.0  59.0   99.5
51.3   50.5  64.3   90.0   56.3   77.0   66.5  112.0  72.0  150.0
64.8  128.0  67.0  133.0   57.5   85.0   66.5  112.0
;
PROC REG;
```

```
    MODEL HEIGHT=WEIGHT;
    OUTPUT OUT=BOTH P=PREDHT;
PROC PLOT DATA=BOTH;
    PLOT HEIGHT*WEIGHT PREDHT*WEIGHT='*' /OVERLAY;
    TITLE 'PREDICTED VS. ACTUAL';
```

结　果

报表 8.4　将回归分析后的预测值与实际值重叠地绘图

```
                          PREDICTED VS. ACTUAL

  Model: MODEL1
  Dependent Variable: HEIGHT

                       Analysis of Variance

                          Sum of        Mean
      Source        DF     Squares      Square     F Value      Prob>F
      Model          1   364.57626   364.57626      57.076      0.0001
      Error         17   108.58795     6.38753
      C Total       18   473.16421

         Root MSE      2.52736     R-square      0.7705
         Dep Mean     62.33684     Adj R-sq      0.7570
         C.V.          4.05435

                          PREDICTED VS. ACTUAL

                       Parameter Estimates

                    Parameter     Standard   T for H0:
      Variable  DF     Estimate        Error  Parameter=0    Prob > |T|
      INTERCEP   1    42.570142   2.67988870       15.885        0.0001
       WEIGHT    1     0.197615   0.02615725        7.555        0.0001

                          PREDICTED VS. ACTUAL

         Plot of HEIGHT*WEIGHT.  Legend: A = 1 obs, B = 2 obs, etc.
         Plot of PREDHT*WEIGHT.  Symbol used is '*'.
        (NOTE: 12 obs hidden.)

HEIGHT |
    80 +
       |
       |                                                      A
    70 +                                  A         *
       |                               A      B      * A
       |                          A   A   **B    A      A
    60 +                          A*  *    A
       |                        A   BA
       |      *
    50 +      A
       |
       --+--------+--------+---------+---------+---------+--------+-
        40        60        80        100       120       140       160

                                  WEIGHT
```

例五：轮廓图（Contour）的示范

本例的轮廓图代表一个三维空间的小山丘。山丘的高度由 Z 来表示，山的底部则以 X，Y 两变量代表。Z 值与 X，Y 之间的关系如下：

$$Z=46.2+0.09X-0.0005X^2+0.1Y-0.0005Y^2+0.0004XY$$

为了表示山丘的高低，我们采用用 CONTOUR 选项而且要求将高度的层次定为 10。所以在报表上，读者可以很清楚地看出有十个不同的绘图符号。这十个符号在图形上制造出由浅入深的视觉效果。愈浅的区域表示 Z 的值愈低或说山丘愈矮，反之亦然。

程 序

```
OPTIONS LS=72 PS=44;
DATA CONTOURS;
    FORMAT Z 5.1;
    DO X=0 TO 400 BY 5;
      DO Y=0 TO 350 BY 10;
        Z=46.2+.09*X-.0005*X**2+.1*Y-.0005*Y**2+.0004*X*Y;
        OUTPUT;
      END;
    END;
PROC PLOT;
    PLOT Y*X=Z / CONTOUR=10;
    TITLE 'CONTOUR PLOT OF X VS.Y';
    TITLE2 'CONTOURS ARE Z';
RUN;
```

结　果

报表 8.5　轮廓图（Contour）的示范

```
                CONTOUR PLOT OF X VS.Y
                CONTOURS ARE Z
                Contour plot of Y*X.
        Y |
          |
         350 +  ====+++OOOOXXXXXWWWWWWWWWWWWWWWWXXXXXOOOO
         340 +  ===+++OOOXXXXXWWWWWWWWWWWWWWWWWXXXXXOOO
         330 +  =+++OOOOXXXXWWWWWWWWWWWWWWWWWWXXXXXOO
         320 +  +++OOOOXXXXWWWWWWWW**********WWWWWWXXXXOO
         310 +  ++OOOXXXXWWWWWW***************WWWWWXXXXOO
         300 +  +OOOXXXXWWWWW*****************WWWWWXXXXO
         290 +  OOOXXXXWWWWW*******************WWWWXXXXO
         280 +  OOXXXXWWWW*********##**********WWWWXXXXO
         270 +  OXXXXWWWW******#########********WWWWXXXXO
         260 +  XXXXWWWW******###################WWWWXXXXO
         250 +  XXXWWWW******####################WWWWXXXXO
         240 +  XXWWWW*****#####################WWWWXXXXO
         230 +  XWWWW*****#####################WWWWXXXXO
         220 +  WWWW*****#####################WWWWXXXXO
         210 +  WWWW****#####################WWWWXXXXO
         200 +  WWW*****#####################WWWXXXOO
         190 +  WW*****#####################WWWXXXOO
         180 +  WW****#####################WWWWXXXOO
         170 +  W*****#####################WWWWXXXOO
         160 +  W*****#####################WWWXXXOOO
         150 +  *****#####################WWWWXXXOO+
         140 +  *****#####################WWWXXXOOO+
         130 +  *****####################WWWWXXXOO++
         120 +  *****####################WWWXXXOOO++
         110 +  *****###################WWWWXXXOO+++
         100 +  *****###################WWWWXXXOOO++=
          90 +  *****#################WWWWXXXOOO++==
          80 +  *****################WWWXXXXOO+++==
          70 +  ******##############WWWWXXXOOO++==-
          60 +  *******############WWWWXXXOOO++===-
          50 +  ********#########WWWWWXXXOOO+++==--
          40 +  W*****************WWWWWXXXOOO+++==--'
          30 +  WW***************WWWWWXXXOOO+++==--''
          20 +  WWW*************WWWWWXXXXOOO+++==---''
          10 +  WWWW***********WWWWWWXXXXOOO+++===--''.
         0 +  WWWWWW********WWWWWWWWXXXXOOO+++===--''..
          |
          ---+---------+---------+---------+---------+--
             0        100       200       300       400

                      X
Symbol          Z     Symbol          Z     Symbol          Z
.....    2.2 -  8.1     +++++   25.8 - 31.7     *****   49.4 - 55.4
'''''    8.1 - 14.0     OOOOO   31.7 - 37.6     #####   55.4 - 61.3
-----   14.0 - 19.9     XXXXX   37.6 - 43.5
=====   19.9 - 25.8     WWWWW   43.5 - 49.4
```

例六：如何把年、月、日标明在横坐标上?

本例的数据是某城市从 1982 年 1 月 4 日至 1982 年 12 月 22 日间随机取三十天当样

本，然后计算这三十天里，公用电话使用的次数 (以 CALLS 表示)。

为了明白电话使用的次数与月份的相关，我们将横坐标的单位定义为十二个月份的每个月的第一天。因此，同一个月份取样的结果均以该月份的首日为横坐标，以公用电话使用的次数为纵坐标。

程 序

```
DATA SAMPLE;
    INPUT DATE:DATE7. CALLS @@;
    LABEL DATE='DATE'
          CALLS='NUMBER OF CALLS';
    CARDS;
1APR82 134   11NOV82 294
2MAR82 289   2DEC82  511
3JUN82 184   22DEC82 413
4JAN82 179   13FEB82 488
5APR82 360   14MAR82 460
6MAY82 245   15APR82 356
7JUL82 280   16JUN82 480
8AUG82 494   17JUL82 388
9SEP82 309   17NOV82 328
11APR82 384  18AUG82 280
21MAR82 201  19SEP82 394
13JUN82 152  23NOV82 590
14JAN82 128  24FEB82 201
15APR82 350  25MAR82 183
15DEC82 150  26APR82 412
16MAY82 240  27MAY82 292
17JUL82 499  28JUN82 309
18AUG82 248  29JUL82 330
19SEP82 356  30AUG82 321
10OCT22 222
;
PROC PLOT;
    PLOT CALLS*DATE/HAXIS='1JAN82'D '1FEB82'D '1MAR82'D '1APR82'D
    '1MAY82'D '1JUN82'D '1JUL82'D '1AUG82'D '1SEP82'D '1OCT82'D
    '1NOV82'D '1DEC82'D '1JAN83'D;
    FORMAT DATE DATE7.;
    TITLE 'CALLS TO CITY EMERGENCY SERVICES NUMBER';
    TITLE2 'SAMPLE OF DAYS FOR 1982';
RUN;
```

结　果

报表 8.6　如何把年、月、日标明在横坐标上?

```
                CALLS TO CITY EMERGENCY SERVICES NUMBER
                      SAMPLE OF DAYS FOR 1982
      Plot of CALLS*DATE.  Legend: A = 1 obs, B = 2 obs, etc.
      |
      |
  600 +
      |                                                 A
      |
      |
      |
      |                                                 A
  500 +                              A  A
      |   A                 A
      |
N     |     A
U     |
M     |
B     |           A                                          A
E 400 +                                          A
R     |                 A                A
      |
O     |                AB                                A
F     |
      |                                  A                A
C     |                              A       AA
A 300 +                                                   A
L     |           A           A      A     A
L     |
S     |
      |                     AA                A
      |                                             A
      |
  200 +          A   A
      |  A             A          A
      |
      |                          A                                A
      |
      |    A           A
      |
  100 +
      |
      --+---+---+---+---+---+--- +---+---+---+---+---+---+---+---+-
        0   0   0   0   0   0   0   0   0   0   0   0   0   0   0   0
        1   1   1   1   1   1   1   1   1   1   1   1   1   1   1   1
        J   F   M   A   M   J   J   A   S   O   N   D   J   A   S   O
        A   E   A   P   A   U   U   U   E   C   O   E   A   N   E   C
        N   B   R   R   Y   N   L   G   P   T   V   C   N   N   B   R
        8   8   8   8   8   8   8   8   8   8   8   8   8   8   8   8
        2   2   2   2   2   2   2   2   2   2   2   2   3   3   3   3
                                    DATE
```