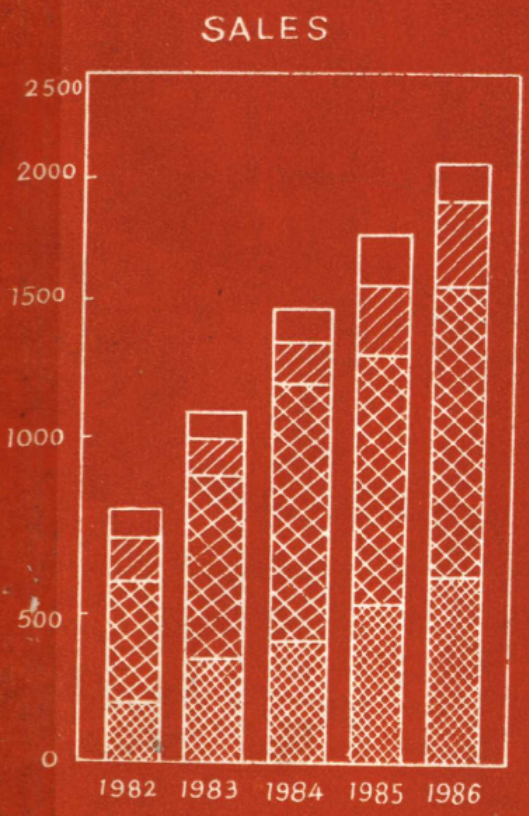
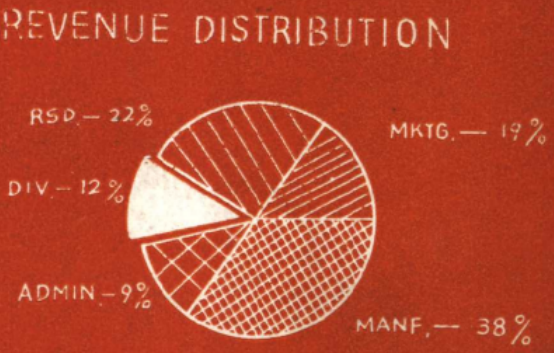
董大钧　张尔强　等译

SAS 过程指导





辽宁科学技术出版社

SAS 过程指导

〔美〕SAS研究所 著

董大钧 张尔強等译

于江红 孔英芳 何武校

辽宁科学技术出版社

（辽）新登字4号

内 容 简 介

SAS软件已是当今世界最流行的统计软件包之一。它对数据的处理和预处理功能极强。目前不仅广泛地应用于统计工作中，而且越来越多地应用于数据管理。

本书译自美国SAS研究所的《SAS Procedures Guide For Personal Compu ter》

本书不仅可作为使用 SAS软件包的参考手册， 而且也可作为大中专开设SAS 课程的教材。

SAS 过 程 指 导
SAS Guo cheng Zhidao
董大钧　张尔强等译

辽宁科学技术出版社出版发行　（沈阳市和平区北一马路108号）
沈阳市光华印刷厂印刷

开本：787×1092 1/16　　印张：25.25字数：570,000
1992年1月第1版　　1992年1月第1次印刷

责任编辑：枫　岚
封面设计：秀　中

印数：1—1,000
ISBN7—5381—1303—7/TP.17　　定价：12.00元

前　言

信息时代的今天，人们在工作实践中会获取到大量的信息。如何存贮、整理和分析处理它们是一件极重要的工作。

对数据分析工作大都是基于基本的统计原理进行的。国内外学者多年来为进行数据定量分析编制了许多统计软件包。SAS软件包则是诸多统计软件包中的佼佼者。它的使用简便，几乎能用极简单的命令去作你想作的一切数据整理和分析工作。

SAS (Statistical Analysis System) 是美国SAS研究所于1976年推出的用来分析数据和编写报告的软件系统。它对数据可以进行各种统计、多元分析并建立每次处理的报告，也可给出图形或进行预测。它具有很强的图形显示功能，甚至三维显示和地图输出。与其它几个世界上流行的统计软件包相比，SAS系统在数据管理方面具有独特之处。因此，除了广泛应用于统计工作之外，SAS在管理领域也得到越来越多的应用。

日本1986年开始安装SAS软件，目前在日本各大学、研究机构、政府机关和一些企业都已广泛使用SAS语言。

我国近年来应用SAS的人越来越多，为使更多的人了解SAS，我们翻译了本书。

该书的第一章至第六章由陈声权同志翻译，第七、九、十四、二十三、二十六章是车文与董大钧同志合译的。第八与第三十一章是董大钧同志翻译。第十至第十三章、第十五至第十八章是何晓源同志翻译。第十九至第二十二与第二十四章是邹怀军同志翻译，第二十五、二十八、二十九和第三十二与第三十五章由刘尚辉同志翻译。第二十七、三十、三十三、三十四章的内容由张尔强同志翻译。韩爱觉、关建海、耿杰同志为本书出版也作了大量的工作。

校对：孔英芳、于江红、何武。董大钧副教授最后又通审了全文。由于译者水平所限，文中内容难免存在错误，请读者批评指正

1991年9月

目　录

第一章 SAS基本统计过程

§1.1 介 绍

以下SAS过程计算单变量或二元变量统计值，例如，平均值、总和、标准差及相关：

UNIVARIATE 进行单变量统计，包括分位数及描绘分布图。

SUMMARY及MEANS 按观测值分组计算基本单变量统计值。分组是由CLASS语句中的变量所决定。统计结果可以被打印或输出到SAS数据集中。

TABULATE 打印基本统计的复杂表格。

CORR 用于数值变量，求变量间相关系数。

进行基本统计的其它过程包括：

CHART 画频数，均值，总和的条形图，立体直方图，饼图及星图。

FREQ 对分类变量计算频数分布，并做多维列联表。

§1.2 过 程 比 较

表1.1给出从每个过程可得到的各种统计量和一些其它重要特点的概述。

表1.1 SAS基本统计过程和它们的特点

统计量	MEANS	UNIVARIATE	SUMMARY	TABULATE	CORR
缺项值数	×	×	×	×	
非缺项值数	×	×	×	×	×
权重和	×	×	×	×	×
均值	×	×	×	×	×
和	×	×	×	×	×
最小值	×	×	×	×	×
最大值	×	×	×	×	×
全距	×	×	×	×	
未修正平方和	×	×	×	×	×
修正平方和	×	×	×	×	×

续表

统计量	MEANS	UNIVARIATE	SUMMARY	TABULATE	CORR
方差	×	×	×	×	×
标准差	×	×	×	×	×
标准误	×	×	×	×	
变异系数	×	×	×	×	
偏度		×			
峰度		×			
学生 t 值	×	×	×	×	
大于 t 值的概率	×	×	×	×	
中位数		×			
四分位数		×			
众数		×			
泊松相关系数					×
其它特性					
打印输出	yes	yes	yes	yes	yes
输出到SAS数据集	yes	yes	yes	no	yes
CLASS语句	yes	no	yes	yes	no
BY语句	yes	yes	yes	yes	yes

§1.3 效　率

SAS处理分位数，包括中位数时，对大的样本所需要的时间与 nlog (n) 成比例，所以，过程UNIVARIATE比其它基本统计过程需要更多的时间。UNIVARIATE也因为数据寄存于内存，需要更多的存贮空间。

如果需要分组计算每一组的统计值，可以在以上任何过程中，用BY语句指明如何分组。然而，按BY分组要对数据集排序，这对大数据集所需要的开销较大。SUMMARY, MEANS和 TABULATE 过程可以不通过排序而用分类进行统计，PROC SUMMARY和PROC MEANS可以建立输出数据集，而 PROC TABULATE 用分层次的列联表打印描述性统计表，但不产生输出数据集。

§1.4 关键字和公式

标准关键字集是指SAS过程中的单变量统计量，在SAS语句中，这些关键字用于要

求将统计值打印或存贮于输出数据集中。

以下符号用于对所有非缺项值的求和：

x_i：变量中第 i 个非缺项观测值。

w_i：如果使用WEIGHT语句，w_i表示与x_i相联系的权重，否则w_i为1。

n：非缺项观测值的数目。

$\bar{x}=\sum w_i * x_i / \sum w_i$

$d=n$　　当规定选择项VARDEF=N时；

$=n-1$　　当VARDEF=DF时；

$=\sum w_i$　　当VARDEF=WEIGHT或WGT时；

$=\sum w_i-1$　当VARDEF=WDF时，

$S^2=\sum w_i(x_i-\bar{x})^2/d$

$z_i=(x_i-\bar{x})/s$ 标准化变量。

下面给出每个统计的公式和标准关键字。在一些公式中，关键字用来标明相应的统计量。

N	非缺项观测值数目
NMISS	缺项观测值数目
MIN	最小值
MAX	最大值
RANGE	MAX−MIN，全距
SUM	$\sum w_i x_i$，加权和
SUMWGT	$\sum w_i$，权重和
MEAN	$\bar{X}$，算术平均值
USS	$\sum w_i x_i^2$，未修正的平方和
CSS	$\sum w_i (x_i-\bar{x})^2$，对于均值修正的平方和
VAR	s^2，方差
STD	s，标准差
STDERR	$s/\sqrt{n}$，平均值的标准误
CV	$100*s/\bar{x}$，变异系数（百分率）
SKEWNESS	$\sum z_i^3 * n/(n-1)(n-2)$，偏度
KURTOSIS	$\sum z_i^4 \cdot n(n+1)/(n-1)(n-2)(n-3)-3(n-1)^2/(n-2)(n-3)$，峰度(尾重测量)
T	$t=\bar{x}\sqrt{n}/s$，对Ho：总体均值= 0 的学生 t 值
PRT	自由度为 n−1的学生 t 值的双尾 p 值，即在获得的 t 的绝对值大于 样本中观测的 t 值无效假设下的概率
MEDIAN	当x_i按值排序，并且 n 为奇数时为中值；当 n 为偶数时为两中值的平均值
QUARTILE	x_i的上下四分位数的值
MODE	x_i的最频繁值

§1.5 单变量统计的数据要求

如果不能计算统计值，则该统计值设为缺项值，N和NMISS不要求任何非缺项观测值。SUM，MEAN，MAX，MIN，RANGE，USS和CSS要求至少有一个非缺项观测值。其它的统计要求如下：

- VAR，STD，STDERR，CV，T和PRT要求至少有两个观测值。
- SKEWNESS 要求至少有3个观测值。
- KURTOSIS 要求至少有4个观测值。
- SKEWNESS，KURTOSIS，T和PRT要求STD>0。
- 当使用WEIGHT语句时，不能计算SKEWNESS和KURTOSIS。
- CV 要求MEAN不为0。

§1.6 统计概念简介

该部分对一些必要的统计概念提供了简略介绍，以解释SAS过程对基本统计值的输出，详细的讨论请参照有关统计课本。

一、总体和参数

通常，有一个你感兴趣的明显定义的元素集，这个元素集叫做全域。与这些元素相联的值的集合叫做总体，统计总体是一个数值集合。例：我们把一所学校的全体学生当作全域来分析，可以有两个我们关心的总体：一个是身高值，一个是体重值。全域还可看作是一个公司制造的所有器具的集合，而总体可以是每个器具在用坏前的使用时间的长短。

总体可以用它的累积分布函数描述，该函数给出少于每个可能值的总体比例，离散总体亦能用概率(PROBABILITY)函数描述，该函数给出等于每个可能值的总体比例。连续总体可经常用密度(DENSITY)函数描述，它是累积分布函数衍生的。密度函数可以由直方图来逼近，直方图给出了一系列间隔值中每个小范围之内的总体构成比，象CHART过程所产生的那样。概率密度函数类似于带有无穷个无穷小间隔的直方图。

在技术术语中，当术语“分布”不带限定时，它通常指累积分布函数，在非正式写作中“分布”有时指的是密度函数。“分布”经常指抽象的总体，而不是具体的总体。同样，统计术语指许多类型的抽象分布，诸如，正态分布，指数分布，Cauchy分布等等。当使用象正态分布这样的术语时，无须特意指出是累计分布还是密度分布。

根据几个概括分布的重要特性的测量描述一个总体是方便的，从总体计算得到的测量被称作参数。人们定义了许多不同的参数来测量分布函数的不同方面。

最普遍使用的参数是（算术）均值，如果总体包含一确定数目的值，其均值为总体内所有值的和被总体内元素的数目除。对于不确定的总体，均值的概念是类似的，但需要较复杂的数学过程。

我们用E(x) 表示由 x 表示的总体的均值。例如身高，这里 E 表示期望值，我们亦可考虑由原始值导出函数的期望值。例如，如果 x 表示高度，那么E(x^2) 就是高度平方的期望值，即身高总体内每个值平方所构成的总体均值。

二、样本及统计

测量一个总体中的所有值经常是不可能的，一个测量值的集合称作一个样本(SAMPLE)，一个样本值的数学函数称作一个统计 (STATISTIC)。一个统计对应一个样本，就象一个参数对应一个总体一样，习惯上用罗马字母表示统计，而用希腊字母表示参数。例：总体均值经常写作μ。而样本均值写作 $\bar{x}$。称作统计的数学分支很大程度是关于样本统计行为的研究。

样本可用大量方法进行选择。大多SAS过程假设数据构成一个简单的随机样本。这意味着样本是按所有样本被选择的机会是均等的这样一种方法选择的。

来自一个样本的统计能被用于对一个总体的参数做推理或合理的揣测。例：如果拿一所中学的三十个学生作一个随机样本。那么这三十个学生的平均身高是这所中学的学生平均身高的合理揣测或估计值。其它的统计如标准误可以提供有关一种估计好坏程度的信息。

对任何总体参数，可使用许多统计对其进行估算。然而经常有一个特殊统计习惯用于估计一个所给的参数。例：样本均值通常是总体均值的估计值。在这种情况下，对参数和统计的求值公式是一样的，在其它情形，对于一个参数公式也许不同于最普遍使用估计值的公式。在所有应用中，不应假设最普遍使用的估计值是最好的估计值。

三、定位测量

定位测量包括均值，众数和中值，这些测量通常被认为是分布中心的描述，在下述定义中，注意如果整个样本每个观测值都加一个固定的量，那么这些定位测量将偏移同样固定的量。

1．均值 (MEAN)：

如上所述，总体的均值μ=E(x) 通常用样本均值来估计：

$$\bar{x}=\sum x_i/n$$

2．中值 (MEDIAN)：

总体的中值是中心值，位于总体的上下两半之间。当数据按降序或升序排列时，样本的中值是中间值。对偶数个观测值，两个中间值的中点通常认为是中值。

3．众数 (MODE)：

众数是总体密度最大处的值。一些密度有一个以上的局部峰值，被称做多众数。样本众数是样本中最经常出现的值。如果有几个最经常发生的样本值，频数相同，UNIVARIATE过程报告最低的这样的值。如果总体是连续的，那么所有样本值产生一次，样本众数无用。

四、分位数（QUANTILES）

分位数（包括百分位数，四分位数及中值）对分布的详细研究是有用的，对一个按大小排序的测量集，百分之P位表示测量值中有P%的小于该值，而（100－P）%大于该值。中值是百分之50的分位。由于按精确的期望百分位划分数据几乎不可能，所以有一个更精确的定义（看UNIVARIATE过程）。

一个分布的上四分位是75%的测量值在其下的值（即第75%分位），25%的测量值落在下四分位值之下。通过 UNIVARIATE 过程只可计算某些百分位值和四分位值。RANK过程能计算任何期望的百分位值。

分位数和定位测量举例

下面例子中数据是通过调用一个伪随机数函数人为产生的。UNIVARIATE过程计算大量的百分位和定位测量，并把值输出到一个SAS数据集，然后一个DATA 步使用SYMPUT子程序将统计值赋给宏变量，宏变量用于宏体%FORMAGEN中，为PROC FORMAT产生值标。结果格式用于PROC CHART显示图表的统计值。

下面程序产生输出1.1

```
TITLE 'EXAMPLE OF QUANTILES AND MEASURES OF LOCATION';
DATA RANDOM;
 DROP N;
  DO N=1 TO 1000;
    X=FLOOR(EXP(RANNOR(314159)*.8+1.8));
    OUTPUT;
  END;
PROC UNIVARIATE;
  VAR X;
  OUTPUT     OUT=LOCATION  MEAN=MEAN  MODE=MODE
          MEDIAN=MEDIAN  Q1=Q1  Q3=Q3  P5=P5  P10=P10
          P90=P90 P95=P95 MAX=MAX;
PROC PRINT;
DATA _NULL_;
  SET LOCATION;
  CALL SYMPUT ('MEAN', ROUND(MEAN, 1)) ;
  CALL SYMPUT ('MODE', MODE) ;
  CALL SYMPUT ('MEDIAN', ROUND(MEDIAN, 1)) ;
  CALL SYMPUT ('Q1', ROUND(Q1, 1)) ;
  CALL SYMPUT ('Q3', ROUND(Q3, 1)) ;
  CALL SYMPUT ('P5', ROUND(P5, 1)) ;
  CALL SYMPUT ('P10', ROUND(P10, 1)) ;
  CALL SYMPUT ('P90', ROUND(P90, 1)) ;
```

```
  CALL SYMPUT ('P95', ROUND(P95, 1)) ;
  CALL SYMPUT ('MAX', MIN(50, MAX)) ;
RUN;
%MACRO FORMGEN;
%DO I=1 %TO &MAX;
    %LET VALUE=&I;
    %IF &I=&P5      %THEN %LET VALUE=&VALUE P5;
    %IF &I=&P10     %THEN %LET VALUE=&VALUE P10;
    %IF &I=&Q1      %THEN %LET VALUE=%VALUE Q1;
    %IF &I=&MODE    %THEN %LET VALUE=&VALUE MODE;
    %IF &I=&MEDIAN  %THEN %LET VALUE=&VALUE MEDIAN;
    %IF &I=&MEAN    %THEN %LET VALUE=&VALUE MEAN;
    %IF &I=&Q3      %THEN %LET VALUE=&VALUE Q3;
    %IF &I=&P90     %THEN %LET VALUE=&VALUE P90;
    %IF &I=&P95     %THEN %LET VALUE=&VALUE P95;
    %IF &I=&MAX     %THEN %LET VALUE=)=&VALUE;
    &I=" &VALUE" ;
    %END;
%MEND;
PROC FORMAT PRINT;
VALUE STAT %FORMGEN;
OPTION PS=60;
PROC CHART DATA=RANDOM;
VBAR X/MIDPOINTS=1 TO &MAX BY 1;
FORMAT X STAT.;
FOOTNOTE  'P5  =5TH  PERCENTILE';
FOOTNOTE2 'P10 =10TH PERCENTILE';
FOOTNOTE3 'P90 =90TH PERCENTILE';
FOOTNOTE4 'P95 =95TH PERCENTILE';
FOOTNOTE5 'Q1  =1ST  QUANTILE';
FOOTNOTE6 'Q3  =3RD  QUANTILE';
RUN;
```

输出 1.1分位数和位置测量例

EXAMPLE OF QUANTILES AND MEASURES OF LOCATION

UNIVARIATE PROCEDURE

VARIABLE = X

Moments

N	1000	Sum Wgts	1000
Mean	7.605	Sum	7605
Std Dev	7.381698	Variance	54.48946
Skewness	2.730385	Kurtosis	11.18706
USS	112271	CSS	54434.98
CV	97.06375	Std Mean	0.233443
T:Mean=0	32.57939	Prob>|T|	0.0001
Sgn Rank	244777.5	Prob>|S|	0.0001
Num ^= 0	989		

P5 = 5TH PERCENTILE

EXAMPLE OF QUARTILES AND MEASURES OF LOCATION

UNIVARIATE PROCEDURE

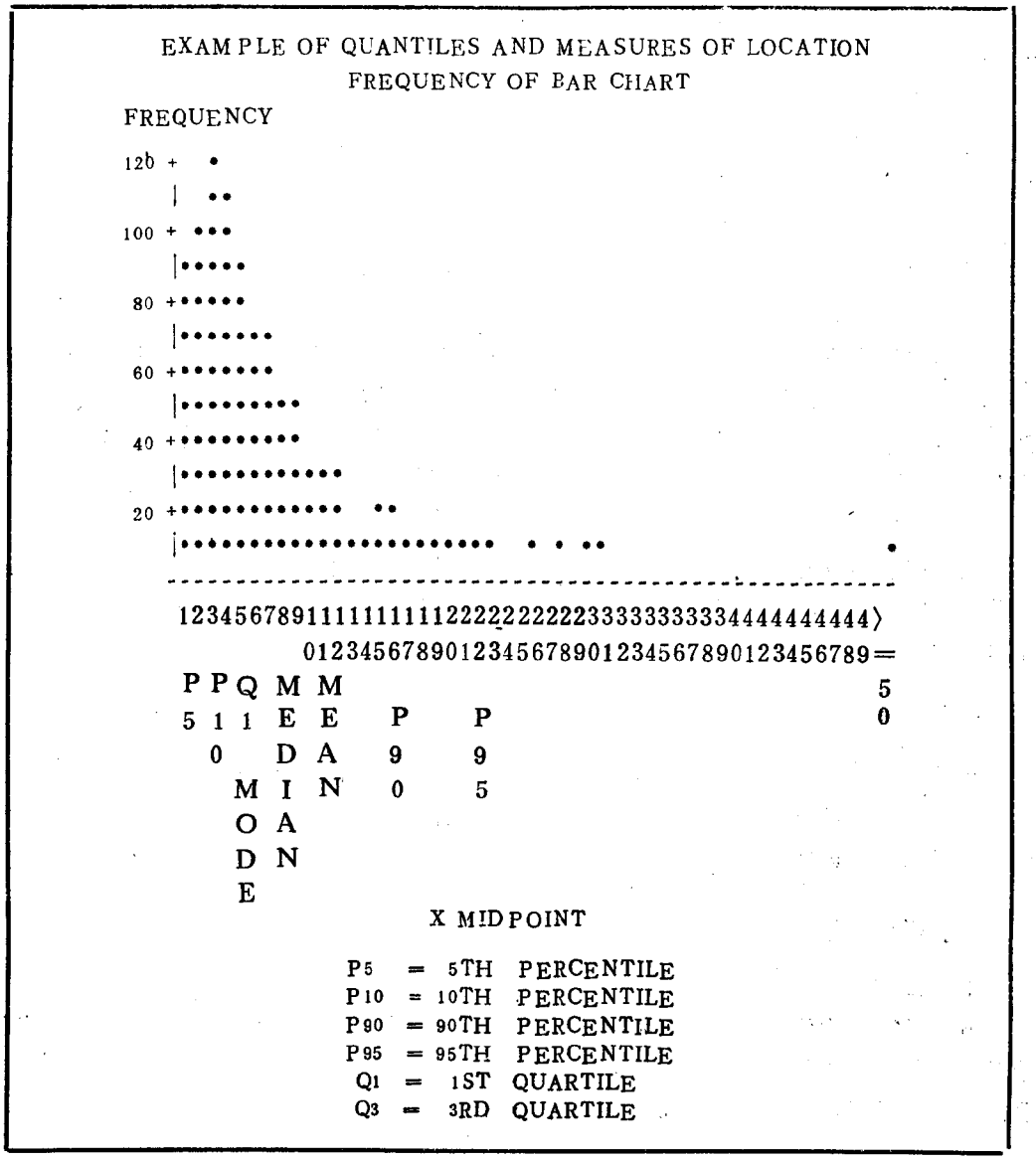
VARIABLE = X

Quantiles(Def=5)

100% Max	62	99%	37.5
75% Q3	9	95%	21.5
50% Med	5	90%	16
25% Q1	3	10%	2
0% Min	0	5%	1
		1%	0
Range	62		
Q3−Q1	6		
Mode	3		

Extremes

Lowest Obs	Highest Obs
0(941)	44(216)
0(756)	44(486)
0(402)	57(319)
0(-358)	61(951)
0(323)	62(147)



五、变异性测量

对研究总体分布很重要的另一组统计是测量值的变异性或扩展性。注意在下面给出的定义中，如果给整个样本中每个观测值添加一个固定量，那么这些统计值不发生变化。然而如果样本中每个观测值乘以一个常数，那么这些统计值将发生变化。

1．全距：

样本全距是样本中最大值与最小值之差。对许多总体而言，至少在统计理论中，全距是无穷的，这样，样本全距可能不会告诉你有关总体的多少信息。当样本含量增加时，样本全距趋向增加，如果所有样本值同乘一个常数则样本全距亦乘同一常数。

2．四分位数间距：

四分位数间距是上四分位数与下四分位数之差。如所有样本同乘一常数，则四分位数间距亦乘同一常数。

3．方差：

当弄清所要考虑的总体时，总体的方差通常用σ^2表示，它是值与总体均值之差的平方的期望值。

$$\sigma^2=E(x-\mu)^2$$

样本方差由S^3表示，计算如下：

$$S^2=(x_i-\bar{x})^2/(n-1)$$

观测值与均值之差称为离均差。方差是离均差平方的均值。若所有观测值离均值很近时，则方差很小，但不会小于0；当值较离散时，则方差较大。如果样本值同乘一常数，则样本方差乘以这个常数的平方。有时不以n－1作分母，VARDEF选择控制使用什么数作为除数。

4．标准差：

无论在总体或样本中，标准差是方差的平方根。对总体标准差通常用符号σ表示，对样本则用S表示，标准差的量纲同观测值一样。如果所有样本值同乘一常数，则样本标准差乘以同一常数。

六、变异系数

变异系数是无量纲的相对变异性测量，它被定义为标准差与均值的比率，表示为一百分比。只有变量按同一尺度测量时变异系数才有意义。如果所有样本值同乘一个常数，则样本变异系数保持不变。

七、形状测量

1．偏度：

方差是偏离均值总量的一种测量，由于方差公式对离均差进行平方，因此，正负偏差对方差的贡献是相同的。但是在许多分布中，正偏差在数量上可能比负偏差大，或者反之。偏度就是偏差在一个方向比另一个方向大的测量。例如上例中的数据是向右倾斜的。

总体偏度被定义为：

$E(x-\mu)^3/\sigma^3$因为偏差是立方的而不是平方的，所以偏差的符号被保持，立方偏差也强调了大偏差的作用。公式中包括一个除数σ^3，以便消除标度效应。这样，所有的值同乘一个常数不改变偏度。偏度可以解释为总体一侧尾部比另一侧偏重。

SAS计算样本偏度为：

$$n/(n-1)(n-2)\sum_{i=1}^{n}(x_i-\bar{x})^3/s^3$$

偏度可正可负且是无界的。

2．峰度：

一个总体的尾重影响了许多统计行为，因而进行尾重测量是很有用的。峰度即为这