**20. PROC UNIVARIATE过程及正态性检验**

**一、用UNIVARIATE 过程步生成单变量的简单统计分析**

语法：

PROC UNIVARIATE <可选参数> data = 数据集;

VAR *变量列表*;

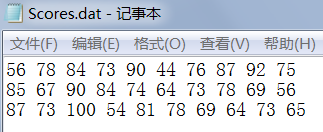
说明：（1）输出包括五个部分：（见**例1**）

1. 矩统计量；
2. 基本的位置和分散程度统计量；
3. 关于均值=0的三种检验：t检验、符号检验、符号秩检验；
4. 各个重要的分位数；
5. 观测数据的五个最低值和五个最高值；

（2）VAR指定要做统计分析的各个变量（默认是全部）；

（3）加上可选参数比如 NORMAL 将多输出一个正态分布检验。

**例1** 某班学生某科考试成绩的数据（C:\MyRawData\Scores.dat）:



读入数据，做简单的统计分析。

代码：

**data** class;

infile 'c:\MyRawData\Scores.dat';

input Score @@;

**run**;

**proc** **univariate** NORMAL data = class;

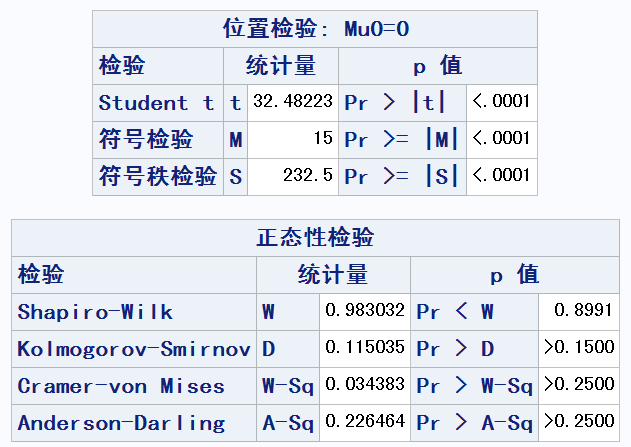
var Score;

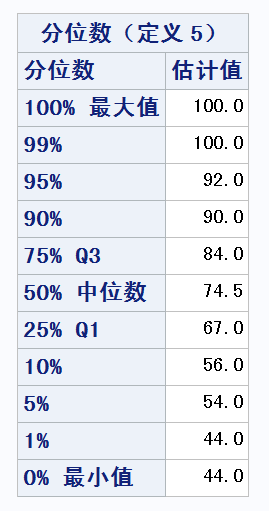
title;

**run**;

运行结果：









**二、用UNIVARIATE 过程步绘制统计图形**

用UNIVARIATE 过程步还可以绘制统计图形，方便对数据做出解释。

语法：

PROC UNIVARIATE <可选参数> data = 数据集;

VAR *变量列表*;

PLOT-绘图命令 *变量列表 </可选项>;*

PLOT-绘图命令有：

CDFPLOT——绘制经验分布图；

HISTOGRAM——绘制直方图，并拟合分布曲线；

PROBPLOT——绘制概率图；

PPPLOT——绘制 PP 图；

QQPLOT——绘制QQ图；

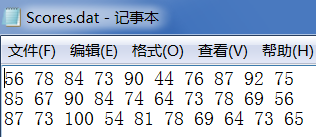
说明：

（1）若要再画上一条其它标准分布作对比，可以加上可选项“\分布名”，包括BETA, EXPONENTIAL, GAMMA, LOGNORMAL, NORMAL, WEIBULL；（PPPLOT, PROBPLOT, QQPLOT默认是NORMAL）；

（2）绘制概率图，是为了验证数据是否符合某种概率分布，若图形接近直线，说明符合。

注：Q-Q图与P-P图判断正态分布不具准确性，一般不太用。

**例2** 关于分数的数据（C:\MyRawData\Scores.dat）：



读入数据，绘制直方图并附上正态分布曲线，另绘制概率图正态分布。

代码：

**data** class;

infile 'c:\MyRawData\Scores.dat';

input Score @@;

**run**;

**proc** **univariate** data = class;

var Score;

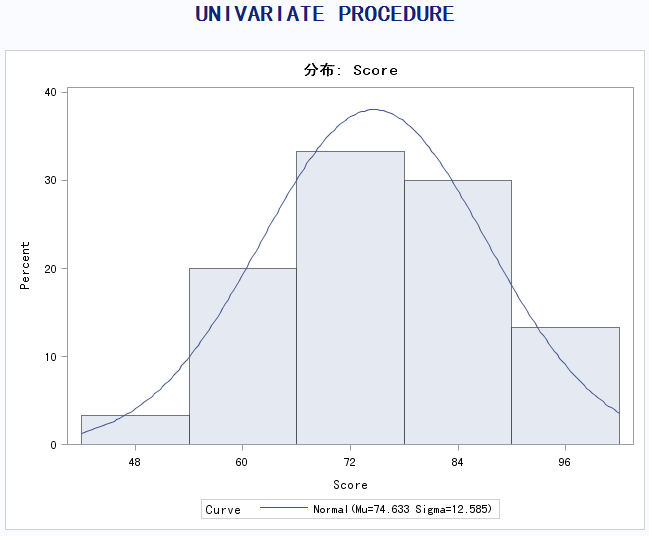
HISTOGRAM Score / NORMAL;

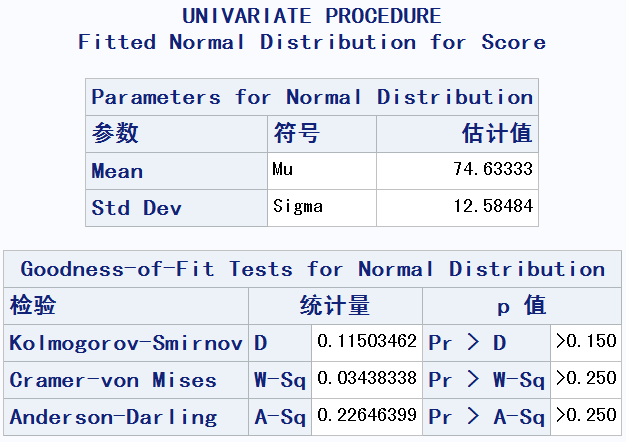
PROBPLOT Score;

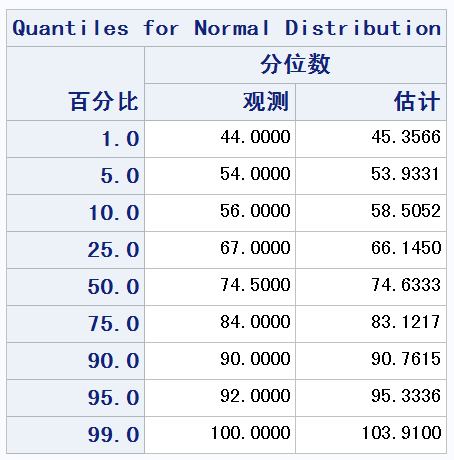
title;

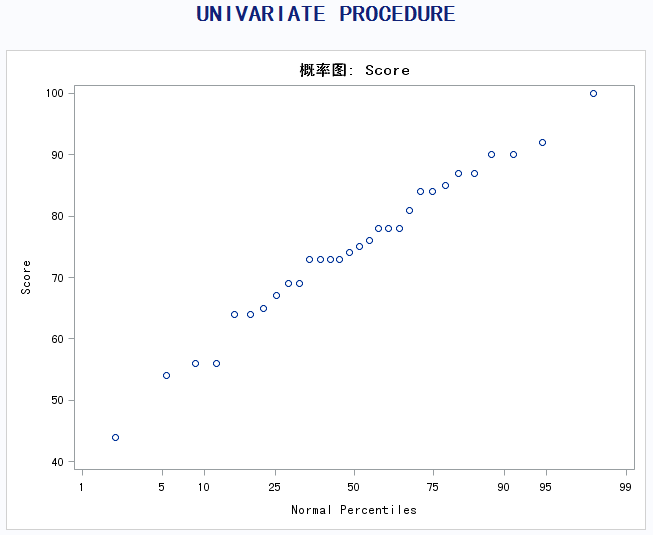
**run**;

运行结果：（类似**例1**五个部分的表格结果略）









注：概率图近似是直线，说明数据近似符合正态分布（PROBPLOT检验的默认分布是NORMAL正态分布）。

**三、正态性检验**

许多计量资料的分析方法要求数据分布是正态或近似正态，因此对原始独立测定数据进行正态性检验是十分必要的。正态性检验主要有三类方法：

1. 计算综合统计量

如动差法、夏皮罗-威尔克Shapiro-Wilk 法(W检验) 、达戈斯提诺D Agostino 法(D检验) 、Shapiro-Francia 法(W检验) .

2. 正态分布的拟合优度检验

如皮尔逊χ2 检验 、对数似然比检验 、柯尔莫哥洛夫Kolmogorov-Smirov 法检验 .

3. 图示法(正态概率图Normal Probability plot)

如分位数图(Quantile Quantileplot ,简称QQ图) 、百分位数(Percent Percent plot ,简称PP图) 和稳定化概率图(Stabilized Probability plot ,简称SP图) 等.

**SAS规则：**

当样本含量n ≤2000 时,结果以Shapiro - Wilk (W 检验) 为准,当样本含量n>2000 时,结果以Kolmogorov - Smirnov (D 检验) 为准。

**SAS过程**

正态分布检验的一般格式如下：

proc univariate data=数据集 normal;

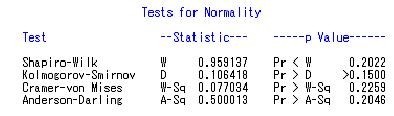
var 变量;

histogram 变量;

probplot 变量;

run;

 在检验中，原假设H0是服从正态分布，若检验结果的p值小于0.05水平，则拒绝零假设，否则接受零假设。

[](http://photo.blog.sina.com.cn/showpic.html#blogid=5fc375650100hdnu&url=http://s5.sinaimg.cn/orignal/5fc37565t73476f9edaf4)

在检验中，我们的零假设是变量服从正态分布，如果test for normality检验结果的p值小于0.05水平，则拒绝零假设，否则接受零假设。