《统计计算与SAS软件》

实验9 描述性统计量的计算、图示

1907402030 熊雄

2021年12月9日

1 实验目的

数据的简单加工并掌握$means$、$univariate$、$freq$过程。

2 实验内容

以$sashelp.class$数据集为例。

1. 用$means$过程计算$weight$的均值，标准差，极差，四分位极差，方差，偏度，峰度，并将这些统计量保存到数据集$resultmeans$中；
2. 用$univariate$过程计算$height$的均值，标准差，极差，中位数，众数，方差，偏度，峰度，并将这些统计量的值保存到数据集$resultuni$中；
3. 对变量$height$画正态概率图（Q-Q图）、茎叶图、盒形图；
4. 用$freq$过程统计男，女生的频数，并考虑性别与年龄是否独立?
5. 将性别变量的值加标签：M表示成“男”，F表示成“女”；
6. 给变量加中文标签：$name$的标签为“姓名”，sex的标签为“性别”，age的标签为“年龄”，$height$的标签为“身高”，$weight$的标签为“体重”；
7. 将$weight$分成三组，<85为一组，用“轻”表示，85-100为一组，用“中”表示，>100为一组，表示为“重”。并考虑sex与这种分组间是否独立；
8. 画$height$和$weight$两变量的散点图，数据点用“红色三角形”表示；
9. 画$height$的直方图、年龄的饼图（可以设置值，填充颜色、类型等）；
10. 按性别分组，计算不同年龄的平均身高。

3 代码实现

3.1 利用*MEAN*过程分析*weight*

以下代码实现了用*means*过程计算*weight*的均值(mean)，标准差(std)，极差(range)，四分位极差(qrange)，方差(variance/var)，偏度(skewness/skew)，峰度(kurtosis/kurt)，并将这些统计量保存到数据集*resultmeans*中。

在 *SAS* 中提交如下代码:

```
PROC MEANS  data = sashelp.class mean std range qrange var skew
kurt;
var weight;
label mean = '均值'
    std = '标准差'
    range = '极差'
    qrange = '四分位数极差'
    var = '方差'
    skew = '偏度'
    kurt = '峰度';
output out = work.resultmeans
    mean = mean
    std = std
    range = range
    qrange = qrange
    var = var
    skew = skew
    kurt = kurt; /*将输出结果保存在数据集resultmeans*/
RUN;
```

可以得到如下输出结果:

SAS 系统

MEANS PROCEDURE

分析变量: Weight 体重（磅）						
均值	标准差	极差	四分位数极差	方差	偏度	峰度
100.0263158	22.7739335	99.5000000	28.5000000	518.6520468	0.1833510	0.6833648

3.2 利用*UNIVARIATE*过程分析*weight*

以下代码实现了用*univariate*过程计算*height*的均值，标准差，极差，中位数，众数，方差，偏度，峰度，并将这些统计量的值保存到数据集*resultuni*中。

在 *SAS* 中提交如下代码：

```
PROC UNIVARIATE data = sashelp.class;
    var height;
    output out = work.resultuni;
run;
```

可以得到如下输出结果：

SAS 系统

UNIVARIATE PROCEDURE
变量: Height (身高（英寸）)

矩			
N	19	权重总和	19
均值	62.3368421	观测总和	1184.4
标准差	5.12707525	方差	26.2869006
偏度	-0.2596696	峰度	-0.1389692
未校平方和	74304.92	校正平方和	473.164211
变异系数	8.22479143	标准误差均值	1.17623173

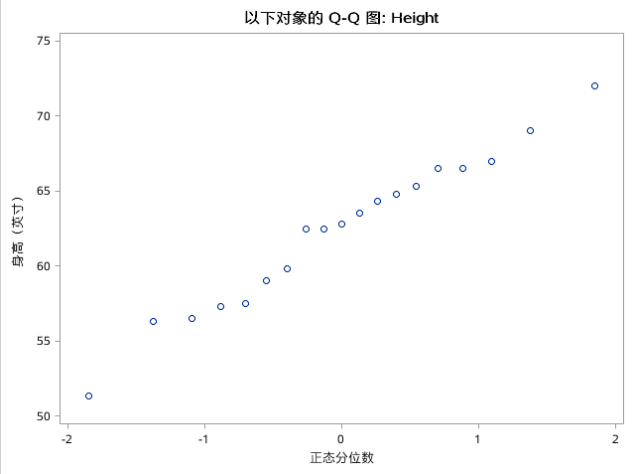
3.3 对变量*height*画图

3.3.1 对变量*height*画正态概率图（Q-Q图）

在 *SAS* 中提交如下代码：

```
PROC UNIVARIATE data = sashelp.class;
    QQplot height;
run;
```

可以得到如下Q-Q图：

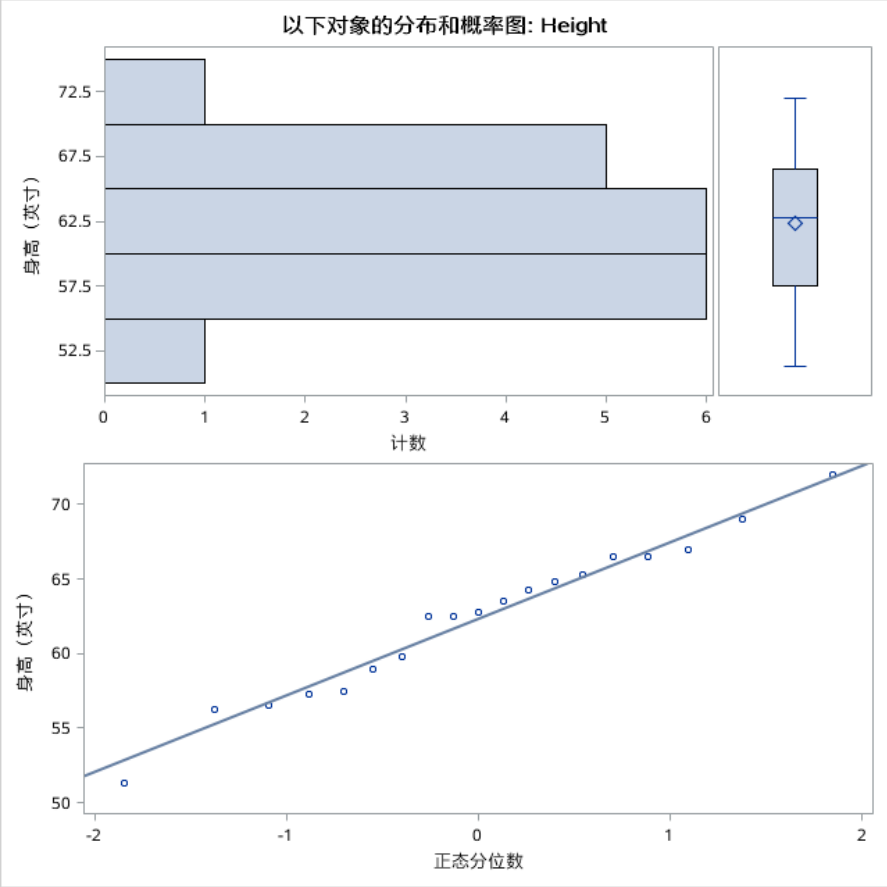


3.3.2 对变量*height*画茎叶图与盒形图

使用PLOT选项UNIVARIATE 过程将产生三种图形：平行条状图 (Horizontal Bar Chart)、盒状图(Box Plot)、正态分布拟合图 (Normal Probability Plot)。在 *SAS* 中提交如下代码：

```
ods select plots;
PROC UNIVARIATE data = sashelp.class plot;
    var height;
RUN;
```

可以得到如下图形：



3.4 利用*freq*过程统计男，女生的频数，并考虑性别与年龄是否独立

在 *SAS* 中提交如下代码:

```
PROC FREQ data = sashelp.class;
    table sex;
    table sex * age /chisq;
RUN;
```

可以得到如下结果:

SAS 系统

FREQ 过程

性别				
Sex	频数	百分比	累积频数	累积百分比
男	10	52.63	10	52.63
女	9	47.37	19	100.00

频数 百分比 行百分比 列百分比	表 - Sex * Age						
	Age(年龄)						
Sex(性别)	11	12	13	14	15	16	合计
男	1 5.26 10.00 50.00	3 15.79 30.00 60.00	1 5.26 10.00 33.33	2 10.53 20.00 50.00	2 10.53 20.00 50.00	1 5.26 10.00 100.00	10 52.63
女	1 5.26 11.11 50.00	2 10.53 22.22 40.00	2 10.53 22.22 66.67	2 10.53 22.22 50.00	2 10.53 22.22 50.00	0 0.00 0.00 0.00	9 47.37
合计	2 10.53	5 26.32	3 15.79	4 21.05	4 21.05	1 5.26	19 100.00

表“Age-Sex”的统计量

统计量	自由度	值	概率
卡方	5	1.4848	0.9148
似然比卡方检验	5	1.8748	0.8662
Mantel-Haenszel 卡方	1	0.0672	0.7955
Phi 系数		0.2795	
列联系数		0.2692	
Cramer V		0.2795	
WARNING: 100% 的单元格的期望计数小于 5。卡方可能不是有效的检验。			

样本大小 = 19

因此性别与年龄并不是相互独立的。

3.5 将性别变量的值加标签

利用*FORMAT*过程将性别变量的值加标签：*M*表示成“男”，*F*表示成“女”。在*SAS* 中提交如下代码即可：

```
PROC FORMAT;
    value $setsex 'M' = '男' 'F' = '女';
RUN;
DATA sss;
    set sashelp.class;
    format sex setsex.;
run;
```

打开数据集*work.sss*，可以看到性别已经变成中文了：

VIEWTABLE: Work.Sss

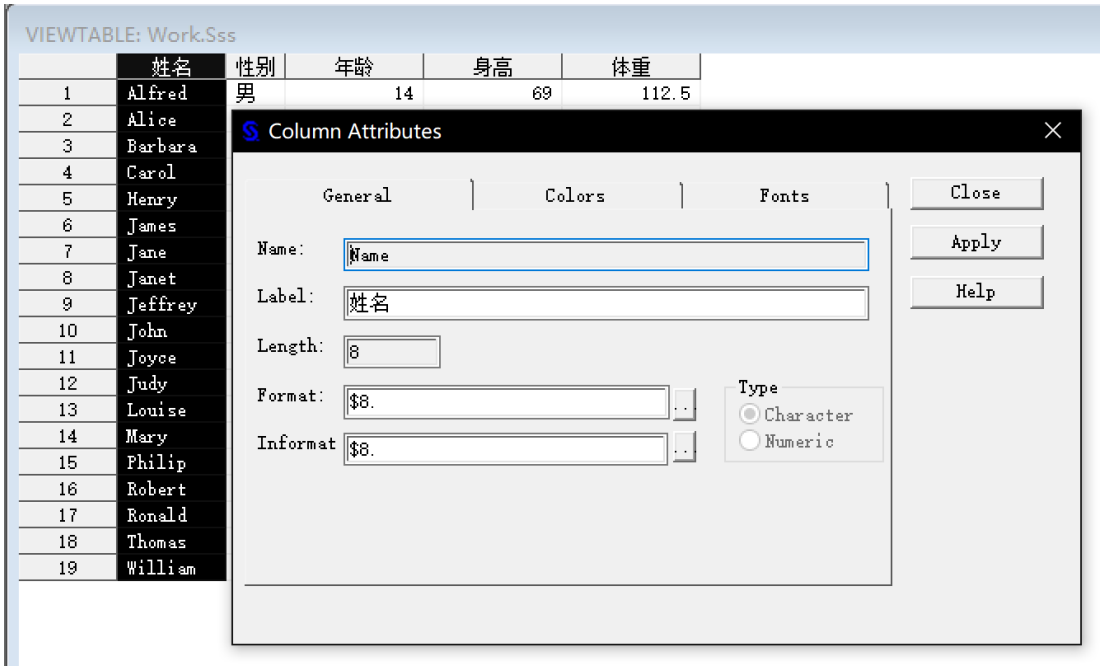
	Name	Sex	Age	Height	Weight
1	Alfred	男	14	69	112.5
2	Alice	女	13	56.5	84
3	Barbara	女	13	65.3	98
4	Carol	女	14	62.8	102.5
5	Henry	男	14	63.5	102.5
6	James	男	12	57.3	83
7	Jane	女	12	59.8	84.5
8	Janet	女	15	62.5	112.5
9	Jeffrey	男	13	62.5	84
10	John	男	12	59	99.5
11	Joyce	女	11	51.3	50.5
12	Judy	女	14	64.3	90
13	Louise	女	12	56.3	77
14	Mary	女	15	66.5	112
15	Philip	男	16	72	150
16	Robert	男	12	64.8	128
17	Ronald	男	15	67	133
18	Thomas	男	11	57.5	85
19	William	男	15	66.5	112

3.6 给变量加中文标签

在*DATA*步中给变量加中文标签：*name*的标签为“姓名”，*sex*的标签为“性别”，*age*的标签为“年龄”，*height*的标签为“身高”，*weight*的标签为“体重”。在 *SAS* 中提交如下代码：

```
DATA sss;
set sss;
label name = '姓名'
    sex = '性别'
    age = '年龄'
    height = '身高'
    weight = '体重';
run;
```

打开数据集*work.sss*，点开变量就可以看到Label了：



3.7 将*weight*分组

先利用*FORMAT*过程将*weight*分成三组的格式，<85为一组，用“轻”表示，85-100为一组，用“中”表示，>100为一组，表示为“重”，再利用*FREQ*过程实现分组。

```
PROC FORMAT;
VALUE setweight LOW - 85 = '轻'
    85 - 100 = '中'
    100 - HIGH = '重';
RUN;

PROC FREQ data = work.sss;
    table weight;
    format weight setweight.;
    table sex * weight / chisq;
RUN;
```

得到输出如下：

SAS 系统

FREQ 过程

体重				
Weight	频数	百分比	累积频数	累积百分比
轻	7	36.84	7	36.84
中	3	15.79	10	52.63
重	9	47.37	19	100.00

频数 百分比 行百分比 列百分比	表 - Sex * Weight			
	Weight(体重)			
Sex(性别)	轻	中	重	合计
F	4 21.05 44.44 57.14	2 10.53 22.22 66.67	3 15.79 33.33 33.33	9 47.37
M	3 15.79 30.00 42.86	1 5.26 10.00 33.33	6 31.58 60.00 66.67	10 52.63
合计	7 36.84	3 15.79	9 47.37	19 100.00

表“Weight-Sex”的统计量

统计量	自由度	值	概率
卡方	2	1.4275	0.4898
似然比卡方检验	2	1.4499	0.4844
Mantel-Haenszel 卡方	1	0.6384	0.4243
Phi 系数		0.2741	
列联系数		0.2644	
Cramer V		0.2741	
WARNING: 100% 的单元格的期望计数小于 5。卡方可能不是有效的检验。			

样本大小 = 19

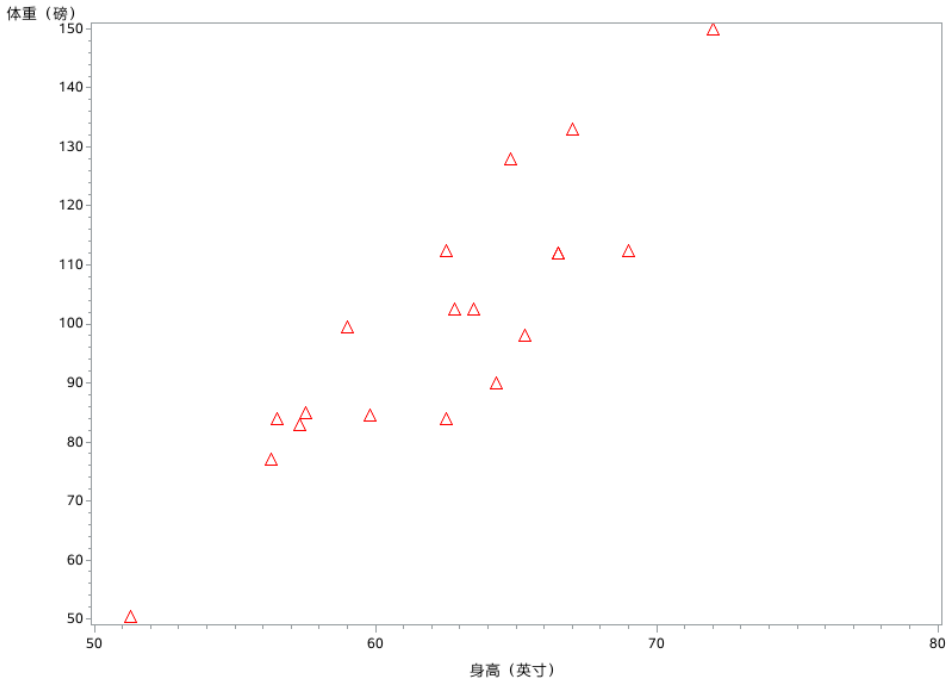
由“表‘Weight-Sex’的统计量”可以看到*sex*与这种分组间并不独立。

3.8 画*height*和*weight*两变量的散点图

画*height*和*weight*两变量的散点图，数据点用“红色三角形”表示；

```
proc gplot data = sashelp.class;
    symbol V = TRIANGLE, CV = RED, H = 2;
    plot weight * height;
run;
```

得到如下图像：



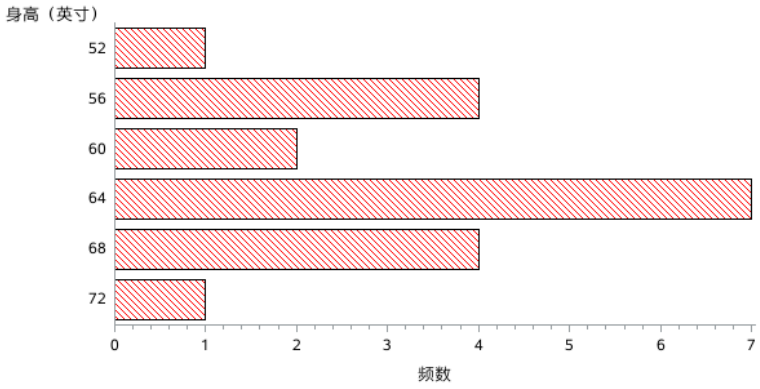
3.9 画*height*的直方图、年龄的饼图

3.9.1 *height*的直方图

输入以下*SAS*代码：

```
GOPTIONS RESET = ALL;
proc gchart data = sashelp.class;
    HBAR height / levels = 6 CFRAME = WHITE NOFRAME;
    PATTERN V = L2, C = red;
run;
```

提交后得到如下直方图：



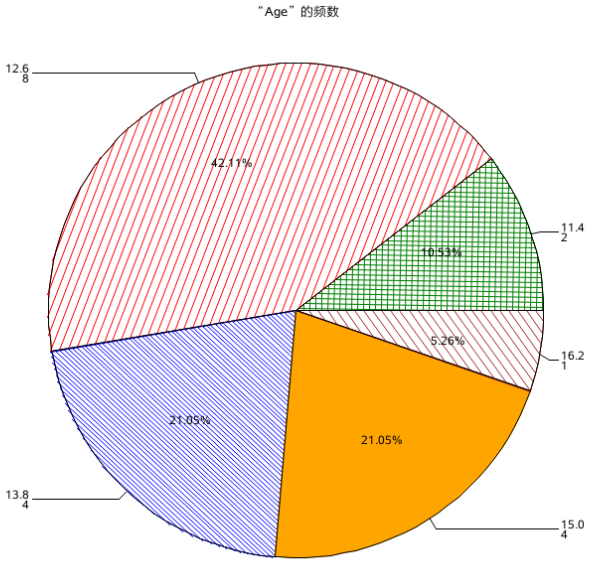
	FREQ.	CUM. FREQ.	PCT.	CUM. PCT.
52	1	1	5.26	5.26
56	4	5	21.05	26.32
60	2	7	10.53	36.84
64	7	14	36.84	73.68
68	4	18	21.05	94.74
72	1	19	5.26	100.00

3.9.2 *age*的饼图

输入以下*SAS*代码:

```
GOPTIONS RESET = ALL;
proc gchart data = sashelp.class;
    PIE age / slice = arrow percent = inside value = arrow;
    PATTERN1 v = p3x70, c = green;
    PATTERN2 v = p2n45, c = red;
    PATTERN3 v = p5, c = blue;
    PATTERN4 v = ps, c = orange;
    PATTERN5 v = p2n45, c = brown;
run;
```

提交后得到如下饼图:



3.10 按性别分组，计算不同年龄的平均身高

输入以下*SAS*代码:

```
PROC MEANS data = sashelp.class mean ;
    class age sex;
    var height;
RUN;
```

提交后得到如下表格，即按性别分组后，不同年龄的平均身高:

SAS 系统

MEANS PROCEDURE

分析变量: Height 身高			
年龄	性别	观测的个数	均值
11	女	1	51.3000000
	男	1	57.5000000
12	女	2	58.0500000
	男	3	60.3666667
13	女	2	60.9000000
	男	1	62.5000000
14	女	2	63.5500000
	男	2	66.2500000
15	女	2	64.5000000
	男	2	66.7500000
16	男	1	72.0000000

4 反思与总结

4.1 统计分析的几个常用过程

1. *MEANS*: 分类计算变量的常用统计量;
2. *UNIVARIATE*: 计算单变量的统计量和分布的拟合检验（正态检验 histogram /normal lognormal）;
3. *FREQ*: 计算变量取值的频数。

4.2 *FREQ*输出频数表格

```
PROC FREQ  DATA = 数据集名;
    TABLES  变量*变量  变量*变量 . . ./
              nocol（不显示列统计）  norow（不显示行）  nocum  nofreq

              nopercent  missing  list
              out = 数据集 outpct ;
         WETGHT  变量名; #按照某个变量加权
         BY  变量名; #按照某个分类变量统计
RUN;
```

4.3 定义输出格式的*FORMAT*过程（对连续性变量分组）

```
PROC  FORMAT  library=saslib;
    VALUE  格式名
    范围1 = 格式化值1 ….
    范围n = 格式化值n ;
 RUN;
```

注1：对字符型数据定义格式，格式名要加前缀$。

注2：若范围超过一个值，离散值可以用，或-用于连续区间，此外<表示到某个临界值范围但不包括该值。

```
例:    'A' = 'Asia'
    1,3,5,7,9 = 'Odd'
    50000 - HIGH = 'Not Affordable'
    13 -< 20 = 'Teenager'
    0 <- HIGH = 'Positive Non Zero'
    OTHER = 'Bad Data'
```

4.4 *MEANS*—描述统计分析

4.4.1 *MEANS*过程步

```
PROC MEANS  DATA = 数据集名 maxdec = 位数 fw = 域宽 noprint 输出统计量名列;
    VAR  变量名列;
    CLASS  变量名列;
    BY  变量名列;
    ID  变量名;
    OUTPUT  OUT= 数据集名 记入数据集统计量名列;
RUN;
```

4.4.2 *MEANS*过程常用选项：

```
MAX     the maximum value
MIN     the minimum value
MEAN    the mean
MEDIAN  the median
MODE    the mode(new in sas9.2)
N   number of non-missing values
NMISS   number of missing values
RANGE   the range
STDDEV  the standard deviation
SUM the sum
CLM two-sided confidence limits
UCLM    upper confidence limits
LCLM    lower confidence limits
CSS   corrected sum of squares 离差平方和
USS   uncorrected sum of squares 平方和
ALPHA    confidence level
CV  coefficient of variation
SKEWNESS    skewness
KURTOSIS    kurtosis
Q1(P25)     25% quantile
Q3(P75)     75% quantile
P1,P5,P10,P90,P95,P99
STDERR      standard error of mean
VAR     variance
PROBT         probability  for  Student's t
```

```
T       Student's t
MAXDEC    number of   decimal places
FW  field width
```

4.5 *UNIVARIATE*过程——单变量特征的概括描述

```
PROC  UNIVARIATE  DATA = 数据集名  noprint ;
    VAR 变量名列 ;
    HISTOGRAM  变量名列 /  midpoints=中点列
    normal(mu=均值  sigma=标准差  图象选项)
    exp(theta= 阈值  图象选项)
    lognormal(theta=阈值  图象选项) … … ;
    INSET  统计量关键名= '显示名'  格式 ;
    QQPLOT  变量名列 /  square . . .  ; %检验是否正态/直线
    PROBPLOT  变量名列 /   . . .  ;
    BY 变量名列 ;
    ID  变量名 ;
RUN ;
```

4.6 上述三种描述性过程的output区别

1. *FREQ*

   在table后（选项）（一般输出频数 频数百分比）。

2. *MEANS*（与var并列）

   output out=数据集 需要输出的统计量（不设定的话 默认输出n min max mean 并且在一列）。

   （1）自己选择要输出的统计量后就会输出成不同的列;

   （2）第一行会输出一行汇总数据（一般无用）。

3. *UNIVARIATE*（与var并列）

   output out=数据集 需要输出的统计量（不设定的话 只输出var的值）。