·计算机应用·

多元数据正态性检验的 SAS/IML 程序设计*

郑州大学公共卫生学院(450052)　李海超　李颖琰△　王爱英

在均值向量检验和协方差阵的检验以及一些多元统计方法中都是假定样本来自多元正态总体,因此,所作统计推断的结论是否正确,在一定程度上取决于实际总体与正态总体接近的程度如何。所以,探讨多元数据的正态性检验问题具有重要的现实意义。本文利用 χ^2 统计量的 Q－Q 图检验法[1]的原理,并结合 SAS/IML[2]软件来建立对多元数据的正态性检验的方法。

χ^2 统计量的 Q－Q 图检验法

1.χ^2 统计量的 Q－Q 图检验法的理论[1,3]

设 $X(a)=(X,\cdots,X_{(ap)})'(a=1,\cdots,n)$ 为来自 p 元总体 X 的随机样本。检验:$H_0:X\sim N_P(\mu,\sum)$,$H_1:X$ 不服从 $N_P(\mu,\sum)$。

由于对多元正态总体 $(X-\mu)'\sum^{-1}(X-\mu)\sim\chi^2(p)$,所以在 H_0 下,将样品 X 到总体中心 μ 的马氏距离 $D^2(X,\mu)$ 记为 D^2,则有

$$D^2=(X-\mu)'\sum^{-1}(X-\mu)\sim\chi^2(p)$$

以下构造的检验方法就是检验统计量 D^2 是否有 $D\sim\lambda^2(p)$ 成立。直观的想法是:由样品 $X_{(a)}$ 计算 $D_a^2(a=1,\cdots,n)$,对 D_a^2 排序:

$$D_{(1)}^2\leqslant D_{(2)}^2\leqslant\cdots\leqslant D_{(n)}^2$$

统计量 D^2 的经验分布函数取为

$$F_n(D_{(t)}^2)=\frac{t-0.5}{n}=p_t\approx H(D_{(t)}^2\mid P),$$

其中 $H(D_{(t)}^2\mid P)$ 表示 $\chi^2(p)$ 的分布函数在 $D_{(t)}^2$ 的值。

利用分位数进行求解。所谓分位数是随机变量的重要数字特征,在求分位数时经常要用到分布函数的反函数。对 $0<p<1$,称满足不等式 $P(X\leqslant x)\geqslant p$,$P(X\geqslant x)\geqslant 1-p$ 的 x 值为随机变量 X 的 p 阶分位数。如果 X 是连续型的,那么 p 阶分位数就是满足方程 $F(x)=p$ 的 x 值。由于这个定义中,p 阶分位数存在唯一性问题,因此采用如下定义:x 的分布函数为 $F(x)$,对 $0<p<1$,定义 x 的 p 阶分位数为 $x_p=\inf\{x:F(x)\geqslant p\}$。所以 $x_p=F^{-1}(p)$ 就是分布函数的反函数,且只存在唯一的 p 分位数,即 $F(x)$ 的左侧分位数。

由经验分布得到样本的 p_t 分位数 $D_{(t)}^2=F_n^{-1}(p_t)$,同时设 χ^2 分布的 p_t 分位数为 χ_t^2,若假设检验 H_0 成立,应有

$$D_{(t)}^2\approx\chi_t^2$$

绘制点 $(D_{(t)}^2,\chi_t^2)$ 的散点图,这些点应散布在一条过原点且斜率为 1 的直线上,如果存在明显偏离,则可以拒绝无效假设。这种检验法其实就是分布 χ^2 的 Q－Q 图检验法。如果不利用分位数,直接用概率散点 $(p_t,H(D_{(t)}^2\mid p))$ 绘图,当 X 为正态总体时,这些点也应散布在一条过原点且斜率为 1 的直线上,这就是 χ^2 分布的 P－P 图检验法。

这里解释一下经验分布函数。设 $(x_1,x_2,\cdots,x_n)$ 是总体 X 的一组样本观察值,将它们按大小顺序排列为 $x_{(1)}\leqslant x_{(2)}\leqslant\cdots\leqslant x_{(n)}$,$x$ 为任意实数,称函数

$$F_{(n)}(x)=\begin{cases}0, & x<x_{(1)}\\ k/n, & x_{(k)}\leqslant x<x_{(k+1)}\\ 1, & x\geqslant x_{(n)}\end{cases}$$

为经验分布函数。经验分布函数的图形是一条阶梯曲线,若观察值不重复则阶梯的每一个跃度都是 $1/n$,若有重复,则按 $1/n$ 的倍数跳跃上升。对任意的实数 x,$F_n(x)$ 的值等于样本观察值 $x_1,x_2,\cdots,x_n$ 中不超过 x 的频率,由频率与概率的关系,$F_n(x)$ 可以作为总体 X 的分布函数 $F(x)$ 的一个近似,随 n 的增大,近似程度越好。

2.SAS/IML 程序

SAS/IML 程序如下[2,4]:

```
Proc iml;
Start invcov;/ * 定义样本协差阵的逆矩阵的模块 * /
n1 = nrow(matrix);
xmean = matrix[:,];
Sum = matrix[+,];
Sscp = matrix' * matrix - sum' * sum/n1;
s1 = sscp/n1;
s = inv(s1);
Finish invcov;
```

*:英国政府对外技术援助部赠款"非典"项目(HNIFD008)

△通讯作者:李颖琰

```
Start msjl;/ * 定义样品点到样本中心的马氏距离的模块 * /
do i=1 to n1;
x=(matrix[i, ]-xmean) * s * (matrix[i, ]-xmean)';
xx=xx//x;
end;
create mydata var{data};
append from xx;
close mydata;
sort mydata by data;
use mydata;
read all var{data} into xxx;
/ * xxx 为马氏距离矩阵 * /
Finish msjl;
Start fws;/ * 定义分位数的模块 * /
do i=1 to n1;
g=(i-0.5)/n1;
y=cinv(g,p);/ * p 为变量的个数,由用户录入 * /
yyy=yyy||y;/ * yyy 为 χ² 分位数矩阵 * /
end;
finish fws;
start plot;/ * 定义绘制散点图的模块 * /
call gstart;
xbox={0 10 10 0};
ybox={0 0 10 10};
call gwindow{0 0,10 12};
call gopen;
call gpoly(xbox,ybox);
call gpoint(xxx,yyy);
call gshow;
finish plot;
use tetsdata;
read all var{var1 var2…} into matrix;
/ * var1 var2…为要检验的 p 元变量 * /
run invcov;
run msjl;
run fws;
run plot;
quit;/ * 退出 IML 程序 * /
```

另外,可以根据需要,还可以分变量做 Q-Q 图(以理论分布的分位数为横轴,以变量观测值为纵坐标绘制的散点图),用以对经过排序的变量值和指定的理论分布的分位数进行比较,从而判断数据是否符合所指定的理论分布。但这个过程在 SAS 系统中有现成的程序,这里略去。

举 例

采用文献[5]中第 45 页的例子:20 名健康成年女性的出汗($x1$),钠的含量($x2$)和钾的含量($x3$)的数据。

具体检验步骤[1]:

(1)由 20 个 3 维样品点计算样本均值 $\bar{x}$ 和样本协方差阵 S

$$S=\frac{1}{n-1}\sum_{a=1}^{n}(x_{(a)}-\bar{x})(x_{(a)}-\bar{x})'$$

(2) 计算样品点 $x_{(t)}$ 到 $\bar{x}$ 的马氏距离:

$$D^2=(x-\bar{x})'S^{-1}(x-\bar{x})\quad(t=1,\cdots,20)$$

(3) 对马氏距离 D_t^2 按从小到大的次序排序:

$$D_{(1)}^2\leqslant D_{(2)}^2\leqslant\cdots\leqslant D_{(n)}^2$$

(4) 计算 $p_t=\frac{t-0.5}{n}(t=1,\cdots,20)$ 及 χ^2 分布的 p_t 分位数 χ_t^2。

(5) 以马氏距离为横坐标,χ_t^2 分位数为纵坐标作平面坐标系,用 20 个点$(D_{(t)}^2,\chi_t^2)$绘制散点图,见图 1。

(6) 考察这 20 个点是否散布在一条通过原点,斜率为 1 的直线上。

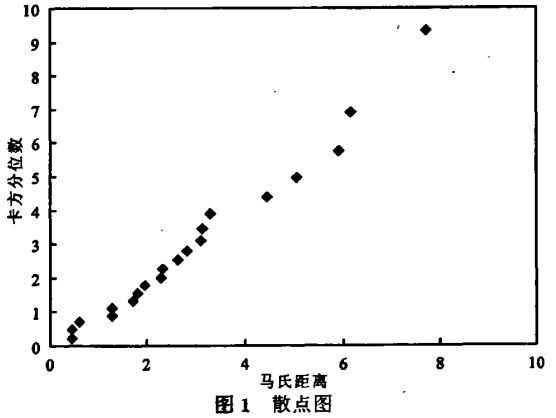

图 1 散点图

从图 1 可以看出,20 个散点有正态趋势但有变异。由于这种方法是建立在大样本基础上的,所以利用小样本进行检验,误差较明显。

参 考 文 献

1. 高惠璇主编.应用多元统计分析.北京:北京大学出版社,2005,100-101.
2. SAS Institute Inc. SAS/IML User's Guide, Version8, NC: SAS Institute Inc, 1999, 846.
3. 罗俊明主编.概率论论与数理统计.郑州:郑州大学出版社,2002,196-197.
4. 高惠璇,等.SAS 系统 Base SAS 软件使用手册.北京:中国统计出版社,1997,160-166.
5. 于秀林,任雪松.多元统计分析.北京:中国统计出版社,2006,32-48.

多元数据正态性检验的SAS/IML程序设计

作者： 李海超， 李颖琰， 王爱英
作者单位： 郑州大学公共卫生学院,450052
刊名： 中国卫生统计 ISTIC PKU
英文刊名： CHINESE JOURNAL OF HEALTH STATISTICS
年，卷(期)： 2007，24(6)
被引用次数： 0次

参考文献(5条)

1. 高惠璇 应用多元统计分析 2005

2. SAS Institute Inc SAS/IML User's Guide,Version8 1999

3. 罗俊明 概率论论与数理统计 2002

4. 高惠璇 SAS系统Base SAS软件使用手册 1997

5. 于秀林. 任雪松 多元统计分析 2006