研究论文

常用分布函数编程算法与新统计检验法

何于班　　黄　梅　　陆如华

（北京应用气象研究所，北京　100029）　（武陵大学）　（国家气象中心）

提　要

研制了 F 分布、χ^2 分布和 t 分布的编程算法，并在此基础上提出了新的统计检验法，免除了繁琐的手工查算，解决了自动分析中自由度随机变动时，传统检验法不能准确检验的难题，从而实现了统计分析自动化，精确筛选因子建立预报方程。这一工作是对传统检验方法的重大改进，适合广大气象台站应用。

关键词： 常用分布函数　编程算法　新的统计检验法

引　言

常用的统计检验有 F 检验、χ^2 检验和 t 检验。由于 F 分布、χ^2 分布和 t 分布计算比较困难，通常都靠查算表得出临界值供检验使用。如今计算机广泛使用，计算统计量，即使样本容量很大，也非常迅速。但随后的检验却是手工查表进行，这不仅使统计分析不能自动化，而且还存在以下两方面问题：

其一是 F 分布、χ^2 分布和 t 分布的查算表只列举有限几个检验水平（概率值）和自由度，不能适应各种情况，查表时往往需要插值，尤其 F 分布有两个自由度，插值更繁，若取邻近值代替，又影响精度。

其二是利用计算机做统计分析时，自由度可能随时改变，查表无法适应这种情况，从而严重影响分析结论的准确性。

为此，完全有必要寻找能解决以上问题的 F 检验、χ^2 检验和 t 检验的编程计算方法，笔者推导出三组有效的编程计算公式。不仅实现了分布函数自动计算，而且还提出了新的统计检验方法，免除手工操作，完全实现自动化的精确统计分析。由于篇幅所限，本文只给出 F 检验的编程算法。

1　F 分布函数的编程算法

F 分布密度函数如下：

$$f(x)=\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}m^{\frac{m}{2}}n^{\frac{n}{2}}x^{\frac{m}{2}-1}\cdot$$

$$\begin{cases}(mx+n)^{-\frac{m+n}{2}} & x>0\\ f(x)=0 & x\leqslant 0\end{cases}$$

分布函数：

$$F(m,n,x)=\int_0^x\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}m^{\frac{m}{2}}n^{\frac{n}{2}}t^{\frac{m}{2}-1}(mt+n)^{-\frac{m+n}{2}}\mathrm{d}t \qquad 0\leqslant x<\infty$$

当 $x<0$ 时，$F(m,n,x)$ 为 0。

将上面的积分加以整理，得出下式：

$$F(m,n,x)=\frac{\Gamma(\frac{m+n}{2})}{\Gamma(\frac{m}{2})\Gamma(\frac{n}{2})}\int_0^x(\frac{mt}{n})^{\frac{m}{2}-1}(\frac{mt}{n}+1)^{-\frac{m+n}{2}}\mathrm{d}(\frac{mt}{n})$$

令 $$M=\frac{m}{2}-1,N=\frac{n}{2}-1,\frac{V}{1-V}=\frac{mt}{n}$$

则有 $$F(m,n,x)=\frac{\Gamma(M+N+2)}{\Gamma(M+1)\Gamma(N+1)}\int_0^y V^M(1-V)^N\mathrm{d}V$$

记 $$G(M,N,y)=\frac{\Gamma(M+N+2)}{\Gamma(M+1)\Gamma(N+1)}\int_0^y V^M(1-V)^N\mathrm{d}V \tag{1}$$

其中 $$y=(\frac{mx}{n})/(1+\frac{mx}{n}),\ x=(\frac{ny}{m})/(1-y) \tag{2}$$

对应于 x 取值范围〔0,∞〕，y 取值范围是〔0,1〕。

m、n 是自然数，可能取的值是 1,2,3……，对应的 M、N 可能取的值是 $-\frac{1}{2}$,0,$\frac{1}{2}$……。

计算中还要用到下列 Γ 函数计算式：

$$\Gamma(K+1)=K\,\Gamma(K)$$

$$\Gamma(2)=\Gamma(1)=1$$

$$\Gamma(\frac{1}{2})=\sqrt{\pi}$$

当 M、N 分别取值 $-\frac{1}{2}$ 或 0 时，$G(M,N,y)$ 值如下：

$$G(-\frac{1}{2},-\frac{1}{2},y)=\frac{\Gamma(1)}{\Gamma(\frac{1}{2})\Gamma(\frac{1}{2})}\int_0^y V^{-\frac{1}{2}}(1-V)^{-\frac{1}{2}}\mathrm{d}V=\frac{2}{\pi}\mathrm{arctg}\sqrt{\frac{y}{1-y}}$$

$$G(-\frac{1}{2},0,y)=\frac{\Gamma(\frac{3}{2})}{\Gamma(\frac{1}{2})\Gamma(1)}\int_0^y V^{-\frac{1}{2}}\mathrm{d}V=\sqrt{y}$$

$$G(0,-\frac{1}{2},y)=\frac{\Gamma(\frac{3}{2})}{\Gamma(1)\Gamma(\frac{1}{2})}\int_0^y(1-V)^{-\frac{1}{2}}\mathrm{d}V=1-\sqrt{1-y}$$

$$G(0,0,y)=\frac{\Gamma(2)}{\Gamma(1)\Gamma(1)}\int_0^y\mathrm{d}V=y \tag{3}$$

当 M(或 N) 值 $\geqslant\frac{1}{2}$ 时，则利用递推方法建立 $G(M,N,y)$ 的计算式。这里需要引用两个不定积分公式：

$$\int V^{M+1}(1-V)^N\mathrm{d}V=\frac{M+1}{M+N+2}\int V^M(1-V)^N\mathrm{d}V-\frac{V^{M+1}(1-V)^{N+1}}{M+N+2} \tag{4}$$

$$\int V^M(1-V)^{N+1}\mathrm{d}V=\frac{N+1}{M+N+2}\int V^M(1-V)^N\mathrm{d}V+\frac{V^{M+1}(1-V)^{N+1}}{M+N+2} \tag{5}$$

这两个公式都可以把被积函数的方次 M(N) 降低。利用式(4)，

$$G(M+1,N,y)=\frac{\Gamma(M+N+3)}{\Gamma(M+2)\Gamma(N+1)}\int_0^y V^{M+1}(1-V)^N\mathrm{d}V$$

$$=\frac{\Gamma(M+N+3)}{\Gamma(M+2)\Gamma(N+1)}\left[\frac{M+1}{M+N+2}\int_0^y V^M(1-V)^N\mathrm{d}V-\frac{y^{M+1}(1-y)^{N+1}}{M+N+2}\right]$$

$$= \frac{\Gamma(M+N+2)}{\Gamma(M+1)\Gamma(N+1)}\int_0^y V^M(1-V)^N dV - \frac{\Gamma(M+N+3)}{\Gamma(M+2)\Gamma(N+1)}\frac{y^{M+1}(1-y)^{N+1}}{M+N+2}$$

再定义函数 $$H(M,N,y) = \frac{\Gamma(M+N+2)}{\Gamma(M+1)\Gamma(N+1)}\frac{y^M(1-y)^N}{M+N+1} \tag{6}$$

则有 $$G(M+1,N,y) = G(M,N,y) - H(M+1,N,y)(1-y) \tag{7}$$

利用式(5)，

$$G(M,N+1,y) = \frac{\Gamma(M+N+3)}{\Gamma(M+1)\Gamma(N+2)}\int_0^y V^M(1-V)^{N+1}dV$$

$$= \frac{\Gamma(M+N+2)}{\Gamma(M+1)\Gamma(N+1)}\int_0^y V^M(1-V)^N dV + \frac{\Gamma(M+N+3)}{\Gamma(M+1)\Gamma(N+2)}\frac{y^{M+1}(1-y)^{N+1}}{M+N+2}$$

于是有 $$G(M,N+1,y) = G(M,N,y) + H(M,N+1,y)y \tag{8}$$

如式(3)所示，当 M、N 值为 $-\frac{1}{2}$ 或 0 时，函数 G 很容易计算，我们记为 $G(M_s,N_s,y)$。当 M、$N \geqslant \frac{1}{2}$ 时，则利用式(7) 式(8) 递推。

$$G(M_s+1,N_s,y) = G(M_s,N_s,y) - H(M_s+1,N_s,y)(1-y)$$

$$G(M_s+2,N_s,y) = G(M_s+1,N_s,y) - H(M_s+2,N_s,y)(1-y)$$

$$= G(M_s,N_s,y) - H(M_s+1,N_s,y)(1-y) - H(M_s+2,N_s,y)(1-y)$$

……

$$G(M,N_s,y) = G(M_s,N_s,y) - \sum_{I=M_s+1}^{M} H(I,N_s,y)(1-y)$$

类似地递推可得

$$G(M,N,y) = G(M,N_s,y) + \sum_{J=N_s+1}^{N} H(M,J,y)y$$

最后有

$$G(M,N,y) = G(M_s,N_s,y) - \sum_{I=M_s+1}^{M} H(I,N_s,y)(1-y) + \sum_{J=N_s+1}^{N} H(M,J,y)y \tag{9}$$

其中 $M > M_s, N > N_s$。从 M、N 与 m、n 的关系可知，当 m 为奇数时，$M_s = -\frac{1}{2}$；m 为偶数时，$M_s = 0$。当 n 为奇数时，$N_s = -\frac{1}{2}$；n 为偶数时，$N_s = 0$。

用式(6) 直接计算 H 函数并不方便，而且当 M、N 较大时，计算 Γ 函数会产生溢出，导致计算失败。因此要用递推式计算 H 函数。按照式(6)

$$H(M+1,N,y) = \frac{\Gamma(M+N+3)}{\Gamma(M+2)\Gamma(N+1)}\frac{y^{M+1}(1-y)^N}{M+N+2}$$

$$= \frac{\Gamma(M+N+2)}{\Gamma(M+1)\Gamma(N+1)}\frac{y^M(1-y)^N}{(M+N+1)}\left(\frac{M+N+1}{M+1}\right)y$$

所以 $$H(M+1,N,y) = H(M,N,y)\left(\frac{M+N+1}{M+1}\right)y \tag{10}$$

同理可得

$$H(M,N+1,y)=H(M,N,y)(\frac{M+N+1}{N+1})(1-y) \tag{11}$$

式(9)与式(10)、(11)结合，就是 $G(M,N,y)$ 的编程计算式，也就是 $F(m,n,x)$ 的编程计算式。

具体步骤如下：

(1) 计算 M、N：$M=\frac{m}{2}-1$；

$$N=\frac{n}{2}-1$$

(2) 求 M_s、N_s：$M_s=-(M \bmod 2)/2$；

$$N_s=-(N \bmod 2)/2$$

(3) 计算 y：$y=\frac{mx}{n}/(1+\frac{mx}{n})$

(4) 用式(3) 计算 $G(M_s,N_s,y)$

(5) 用式(9)、(10)和(11)计算 $G(M,N,y)$。

2 检验临界值的计算

做统计检验时，往往不是给定 x(积分上限) 求分布函数值(概率)，而是设定检验水平 α，针对一定自由度求临界值 x_α。例如作 F 检验时，设定 α 后，令

$$F(m,n,x_\alpha)=1-\alpha$$

求 x_α 值，这也视为求方程 $F(m,n,x_\alpha)-(1-\alpha)=0$ 的根。

求根较好的算法是牛顿插值法，它往往迭代次数较少。但是在迭代过程中可能出现超常值，导致非法运算，程序无法运行下去。因此我们建议采用二分法求取临界值 x_α。以 F 检验为例，根据式(1)，要找出 x_α，满足 $F(m,n,x_\alpha)=1-\alpha$，也就是要找出 y，使 $G(m,n,y)=1-\alpha$。

其步骤如下：

(1) 确定临界值 x_α 要求的精度 E(例如 $E=0.001$)。

(2) 确定 y 的上界 yu 和下界 yl。

从式(2) 知，$yu=1$，$yl=0$。

(3) 令 $y_1=(yl+yu)/2$

按式(2) 求出与 y_1 对应的 x_1。

(4) 计算出 $G(M,N,y_1)=\beta$。

(5) 若 $\beta\geqslant 1-\alpha$，则将 y_1 值赋给 yu；

若 $\beta<1-\alpha$，则将 y_1 值赋给 yl。

(6) 令 $y_2=(yl+yu)/2$。

按式(2) 求出与 y_2 对应的 x_2。

(7) 若 $|x_2-x_1|<E$，则停止计算，x_2 就是要找的临界值 x_α(将 x_2 值给 x_α)；

若 $|x_2-x_1|>E$，则将 y_2 值赋给 y_1，返回(4) 继续计算。

以上就是二分法求临界值的步骤。χ^2 检验和 t 检验同样可以这样求临界值。上机计算表明，二分法不但稳妥，而且速度也很快。

3 统计检验新方法及其在气象上的应用

实现了分布函数编程计算，就可以研制新统计检验方法。

传统的统计检验方法是设定检验水平 α 之后，根据自由度查表得出临界值 x_α，随后对比统计量 x_s 是否超过 x_α 来作出检验决断。这种传统检验方法存在着前面已指出的问题，现在我们能立即算出分布函数值，就可以解决这个问题，不用查表确定临界值 x_α 就可以做统计检验。我们仍以 F 检验为例来说明统计检验新方法原理。

设定 α 之后，则存在一个 x_α，使得：$F(m,n,x_\alpha)=1-\alpha$

我们根据统计量 x_s，立即计算出分布函数值 $F(m,n,x_s)$。由于分布函数都是自变量 (x) 的增函数，如果：$F(m,n,x_s)>1-\alpha$，

则必定 $x_s\geqslant x_\alpha$；如果：$F(m,n,x_s)\leqslant 1-\alpha$，则必定 $x_s<x_\alpha$。

因此 $F(m,n,x_s)$ 是否超过 $(1-\alpha)$ 完全等价于 x_s 是否超过 x_α。如果以往是 $x_s>x_\alpha$，拒绝原假设；$x_s\leqslant x_\alpha$，接受原假设。现在就改为：$F(m,n,x_s)>1-\alpha$，拒绝原假设；$F(m,$

$n, x_s) \leqslant 1-\alpha$，接受原假设。

这就是新的统计检验法。它省去了计算临界值 x_α 的大量计算，也不必根据自由度手工查表得到，可以广泛用于气象统计问题，尤其用于逐步筛选回归，既方便又精确。

用逐步筛选回归建立预报方程时，是逐个选入因子，而且已选入的因子还有可能被剔除。选入某因子或剔除某因子都由 F 检验决定。这时统计量 x_s 如下：

$$x_s = \frac{V}{Q/(N-K-1)}$$

其中 V 是该因子的方差贡献，Q 为残差，N 为样本容量，K 为已选入方程的因子个数。作检验要将 x_s 与 x_α 比较。x_α 应满足

$$F(1, N-K-1, x_\alpha) = 1-\alpha$$

α 是给定的，N 是已知的，但是 K 是变化的。在逐步筛选回归程序运行过程中，不可能随着 K 的变化随时手工查出 x_α 值，来满足程序运行的需要。因此传统的做法是给定一个固定的 x_α 值，当 $x_s > x_\alpha$ 时，选入该因子；当 $x_s \leqslant x_\alpha$ 时，剔除该因子，这样就会误选入（剔除）因子，使预报方程质量下降。用新的检验法，就不会出现这一问题。在逐步筛选回归程序运行过程中，当计算出 x_s 后，立即计算出 $F(1, N-K-1, x_s)$ 值。

若 $F(1, N-K-1, x_s) > 1-\alpha$，则选入该因子；

若 $F(1, N-K-1, x_s) \leqslant 1-\alpha$，则剔除（不选入）该因子。

这样就实现了精确地自动筛选因子建立预报方程。

最后还要说明一点，根据递推关系，例如式(7)、(8)、(10)和(11)，可以运用有递归调用功能的语言（如C语言、QBASIC语言）编制函数递归调用程序作计算，我们也做过试验，在自由度(m, n)较小时是可行的。当自由度增大后，就可能出现堆栈溢出，导致计算失败。因此，采用本文各种求和公式和递推公式编程是稳妥的方案。

本文所述各种算法的推导过程虽然较复杂，但最后的编程公式并不复杂，我们上机编程计算证明，程序量较小，计算速度快，统计分析精度高。

The Programming Algorithm of Normal Distribution Function and the New Statistical Verification Methods

He Yuban (Beijing Institute of Applied Meteorology, 100029)
Huang Mei (Wuling College)
Lu Ruhua (National Meteorological Center)

Abstract

The programming algorithm of the F distribution functions was described. The new statistical verification method introduced not only advoids the trivial table look up but also overcomes the inaccuracy by changing the degree of freedom randomly in the automatical analysis for the traditional verification method. So the regression equations developed by screening predictors accurately and the statistical verification can be automatized. The traditional statistical verification method is improved by this new statistical verification method which is adoptive for meteorological stations in China.

Key Words: normal distribution function programming algorithm new statistical verification