

分布拟合优度检验方法综述

费鹤良

一、引言

由于理论和实际的需要,对于分布拟合优度检验方法的研究,一直是数理统计中一个引人注意的问题。特别在六十和七十年代,这方面的工作有了很大的发展。除了早为人们所熟知的 χ^2 检验以外,发展较快的有建立在经验分布函数基础上的各种 EDF 统计量。有时出于实际的需要,经常要在某两个分布之间作出选择,这就使人想起采用极大似然比统计量来进行检验,目前已有可供实际使用的检验用表。随着电子计算技术的发展,研究分布形状的一些统计量,例如子样偏、峰度统计量用来作用分布检验的方法有了新的发展,已能直接用于正态性检验的 $\sqrt{b_1}$ 、 b_2 和通用(Omnibus)检验的百分位点用表。而对于一些特殊的分布,例如正态和指数分布的检验提出了不少方法,许多人对于这些检验方法的一致性和效作了不少工作。下面就有关上述检验方法作一些评述。

二、拟合优度检验

设随机变量 X , 分布函数 $F(x)$ 。随机变量 X 的 n 个独立观察值 x_1, x_2, \dots, x_n 。根据子样观察值检验假设

$$H_0: F(x) = F_0(x; \theta) \quad (2.1)$$

$F_0(x; \theta)$ 为指定的一个分布, θ 为向量。

对原假设 H_0 , 可分两种情况:

- (1) 简单假设: 若 $F_0(x; \theta)$ 是完全确定的。
- (2) 复合假设: 若 $F_0(x; \theta)$ 形式上确定, 但含有未知参数 θ , 要由子样作出估计。

拟合优度检验根据所得子样是完全子样, 还是截尾(包括定数和定时截尾)子样的情况, 围绕着简单假设和复合假设建立相应的统计量, 寻出这些统计量的分布或渐近分布。由于计算技术的发展, 对于难于寻出精确分布的情形, 可以用 Monte-Carlo 方法给出统计量的百分位点, 以便于显著性检验。

从实际应用和理论的角度来看, 一个好的检验应该满足一些条件, Stephens^[1] (1978年)曾提出四点原则:

- (1) 从使用者来说, 希望能直接用原始资料, 如果可能, 应对 X 的值不要作麻烦的变换。
- (2) 当用某统计量作检验, 发生显著的情况时, 能从所用的统计量中对原始资料作出解释, 即在原始资料中有那些不规则的地方。
- (3) 检验应是一致的和无偏的。

本文于1981年6月6日收到

(4) 检验在一大类对立假设中是有效的。

三、EDF 统计量

在分布检验中,较早的检验方法是 K. Pearson 1900 年提出的 χ^2 检验, χ^2 检验特别适用于离散分布的检验。对于连续分布,由于需要对资料分组而失去一些信息。 χ^2 检验的优点在于可以适用于检验任何分布,毫无困难地用于分布中含有未知参数的情形,也可用于截尾子样的情形^{[2][3]}。

χ^2 检验满足原则 (2)、(2) 计算也比较方便。但对数据的使用信息不多,从效来说,比其它检验方法差些。

1. EDF 统计量

设 $F(x; \theta)$ 为连续分布函数, $F_n(x)$ 为经验分布函数,即 $F_n(x) = \frac{n_x}{n}$, n_x 为子样观察值中 $\leq x$ 的个数。经验分布函数统计量是建立在 $F_n(x)$ 与 $F(x; \theta)$ 之间的差异基础上的一个统计量。首先提出这种统计量的是柯尔莫哥洛夫,以后是斯米尔诺夫,常称为 $K-S$ 检验。

2. $K-S$ 检验

设 $F(x; \theta)$ 为连续分布, $F_n(x)$ 为经验分布,检验的 $K-S$ 统计量为:

$$D_n = \sup_{-\infty < x < +\infty} |F_n(x) - F(x; \theta)| \quad (3.1)$$

当 θ 为已知时,对有限的 n 可以求出 D_n 的精确分布。当 $n \rightarrow \infty$ 时, D_n 有极限分布:

$$\lim_{n \rightarrow \infty} p \left(D_n < \frac{\lambda}{\sqrt{n}} \right) = \sum_{K=-\infty}^{+\infty} (-1)^K e^{-2K^2 \lambda^2} \quad (3.2)$$

检验时所需的临界值表可参见[4]。

参数 θ 未知时,对于完全样本,可用子样观察值对 θ 作出估计。David 和 Johnson^[5] 指出在分布中仅含有未知位置和尺度参数的情况下,可以用极大似然估计来代替未知参数,能构造出进行 $K-S$ 检验的附表。以后 Liefors 在 1967 年和 1969 年用随机模拟方法给出了含有未知均值和方差时,正态分布的 $K-S$ 检验的临界值表^[6] 以及含有未知均值时指数分布的 $K-S$ 检验的临界值表^[7]。并说明了 $K-S$ 检验的效比 χ^2 检验大,见表 1。

对于截尾子样,当参数 θ 已知时, Barr 和 Davidson^[8] 1973 年给出定数和定时截尾时 $K-S$ 检验的统计量为:

$$\begin{aligned} D_r &= \sup_{F_n(x) \leq a} |F_n(x) - F(x; \theta)| \\ &= \sup_{x \leq x_{(r)}} |F_n(x) - F(x; \theta)| \end{aligned} \quad (3.3)$$

$$\begin{aligned} D_u &= \sup_{F(x; \theta) \leq a} |F_n(x) - F(x; \theta)| \\ &= \sup_{x \leq x_u} |F_n(x) - F(x; \theta)| \end{aligned} \quad (3.4)$$

其中 $x_{(r)}$ 为到有 r 个产品失效的时间, x_u 为预先规定的定时截尾时间,有关检验用的临界值

表可见^{[8],[13]}。

Koziol 和 Byar^[9] 给出了统计量:

$$D_{n,T} = \sup_{x < T} |F_n(x) - F(x; \theta)| \tag{3.5}$$

的渐近分布

$$\begin{aligned} \lim_{n \rightarrow \infty} P(\sqrt{n} D_{n,T} \leq d) \\ = \sum_{i=-\infty}^{\infty} (-1)^i \exp(-2i^2 d^2) P\left[\left| X - 2id \left(\frac{1-T}{T} \right)^{\frac{1}{2}} \right| < (T-T^2)^{-\frac{1}{2}} \right] \end{aligned} \tag{3.6}$$

其中 X 为 $(0, 1)$ 上均匀分布的随机变量, T 为截尾时间。检验的临界值表可见^{[9],[13]}。

表1 用 K-S 检验和 χ^2 检验时拒绝正态的概率

真实分布	K-S 检验		χ^2 检验	
	$\alpha=0.05$	$\alpha=0.10$	$\alpha=0.06$	$\alpha=0.12$
正态	0.06	0.10	0.06	0.12
$\chi^2(3)$ 分布	0.44	0.55	0.20	0.27
$t(3)$ 分布	0.50	0.58	0.40	0.52
指数	0.61	0.72	0.29	0.41
均匀	0.12	0.22	0.10	0.18

3、Cramer-Von Mises 型统计量

Cramer-Von Mises 型统计量是建立在经验分布与被检验分布 $F(x; \theta)$ 之差的均方积分的基础上, 为了要求在尾部 $F_n(x)$ 与 $F(x; \theta)$ 之间差异要小, 通常可以进行加权。由于使用的权函数不一样, 可得到各种统计量。

1. Cramer-Von Mises 统计量:

$$W_n^2 = n \int_{-\infty}^{\infty} [F_n(x) - F(x; \theta)]^2 dF(x; \theta) \tag{3.6}$$

2. Anderson-Darling (1954 年) 统计量:

$$A_n^2 = n \int_{-\infty}^{\infty} \frac{[F_n(x) - F(x; \theta)]^2}{F(x; \theta)[1 - F(x; \theta)]} dF(x; \theta) \tag{3.7}$$

3. Watson U^2 统计量:

Watson^[13] (1961 年) 提出统计量为:

$$U_n^2 = n \int_{-\infty}^{\infty} \left\{ F_n(x) - \int_{-\infty}^{\infty} (F_n(y) - F(y)) dF(y) \right\}^2 dF(x; \theta) \tag{3.8}$$

并给出了 U_n^2 的渐近分布:

$$P(U^2 > v) = \sum_{m=1}^{\infty} (-1)^{m-1} 2e^{-2mn^2v} \tag{3.9}$$

Stephens^[14] (1963年) 推导了 U_1^2 , U_2^2 , U_3^2 的精确分布, 以及对 $n > 3$ 时 U_n^2 精确分布下尾部的表示式:

$$P_n(z) = \frac{(n-1)! \sqrt{n} \left\{ \pi - \left(z \frac{1}{12n} \right) \right\}^{\frac{1}{2}(n-1)}}{F\left(\frac{1}{2}(n+1)\right)}$$

$$\frac{1}{12n} \leq z \leq \frac{1}{12n} + \frac{1}{2n^2} \quad (3.10)$$

上述统计量在参数已知的情况下与原分布 $F(x; \theta)$ 独立。在实际计算时, 作变换 $z_i = F(x_{(i)}; \theta)$, $i = 1, 2, \dots, n$ 。其中 $x_{(i)}$ 为将子样观察值由小到大排列时第 i 个值, 即 $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ 。上述统计量的计算公式为:

$$W_n^2 = \sum_{i=1}^n \left[z_i - \frac{2i-1}{2n} \right]^2 + \frac{1}{12n} \quad (3.11)$$

$$A_n^2 = - \left\{ \sum_{i=1}^n (2i-1) [\ln z_i + \ln(1 - z_{n+1-i})] \right\} / n^{-n} \quad (3.12)$$

$$U_n^2 = W_n^2 - n \left[\sum_{i=1}^n z_i \ln z_i - \frac{1}{2} \right]^2 \quad (3.13)$$

对于参数已知和完全样本的情况, Durbin 和 Knott (1972) 给出了 W_n^2 的分布和渐近分布^{[10][11][12]}。

统计量 A^2 , 在 θ 已知时, Anderson 和 Darling^{[12][17][15][16]} 给出了渐近分布和分位点表, 以后 Lewis (1961) 证明了对有限子样 A^2 的分布非常快的趋向于渐近分布, 因此实际上只要渐近分布就行了 ($n \geq 5$), 下面给出分位点表。

表 2 A^2 检验的上尾部百分位点 ($n \geq 5$)

显著性水平 α	0.15	0.10	0.05	0.025	0.01
临 界 值	1.610	1.933	2.492	3.070	3.853

当 θ 中含有未知分量时, $z_i = F(x_{(i)}; \hat{\theta})$ (其中 θ 为 $\hat{\theta}$ 的估计) 不是均匀分布, 统计量 W_n^2 , A_n^2 和 U_n^2 的分布将依赖于 n 和未知参数。如果 θ 的未知分量仅是位置和尺度参数时, 统计量 W_n^2 , A_n^2 和 U_n^2 的分布仅依赖于所检验的分布, 而不依赖于未知参数。Stephens^[18] (1976) 对 W_n^2 , U_n^2 和 A_n^2 讨论了它们的渐近分布。对于一些特殊的分布如正态分布^[15], 指数分布^[19], 未知尺度参数的 γ -分布^[20] 和含有未知位置和尺度参数的极值分布^[21] 能用 EDF 统计量作检验。下面给出 A^2 的检验用表。

I. 正态分布:

情况 1. σ^2 已知, μ 未知, 用子样均值 \bar{x} 估计 μ ;

情况 2. μ 已知, σ^2 未知, 用 $S_1^2 = \frac{1}{n} \sum (x_i - \mu)^2$ 估计 σ^2 ;

情况 3. μ, σ^2 未知, 用 \bar{x} 估计 μ , 用 $S = \frac{1}{n-1} \sum (x_i - \bar{x})^2$ 估计 σ^2 。

对于上述情况首先计算:

$$w_i = \frac{x_{(i)} - \bar{x}}{\sigma} \quad (\text{情况 1})$$

$$w_n = \frac{x_{(i)} - \mu}{s_1} \quad (\text{情况 2})$$

$$w_i = \frac{x_i - \bar{x}}{s} \quad (\text{情况 3})$$

于是 z_i 是标准正态分布对 w_i 的累积概率, 可由标准正态分布分位点表查得。然后由公式 (3.12) 计算 A_n^2 。再根据子样大小 n , 修正 A_n^2 , 得到 ($n \geq 5$)。

$$A^* = A_n^2(1.0 + 0.75/n + 2.25/n^2) \quad (3.14)$$

A^* 的检验用表^[17] 见下表 3。

表 3 A^* 临界值表

显著性水平 α		0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.0025
正态分布	情况 1	0.644		0.782	0.894	1.087	1.285	1.551	1.756	1.964
	情况 2	1.072		1.423	1.743	2.308	2.893	3.702	4.324	4.954
	情况 3	0.472	0.509	0.561	0.631	0.752	0.873	1.035	1.159	1.283
指数分布		0.736	0.816	0.916	1.062	1.321	1.591	1.959	2.244	2.534

II、指数分布:

$$F(x) = 1 - \exp\left(-\frac{x}{\beta}\right) \quad x > 0$$

当 β 未知时, 用极大似然估计 $\hat{\beta} = \bar{x}$, $z_i = 1 - \exp\left(-\frac{x_{(i)}}{\bar{x}}\right)$ $i = 1, 2, \dots, n$ 。用公式 (3.12) 计算 A_n^2 , 修正 A_n^2 得 A^* 为

$$A^* = A_n^2(1.0 + 0.3/n) \quad (3.15)$$

检验时用的临界值表见表 3。

III、极值分布:

$$F(x; \theta) = \exp\left[-\exp\left(-\frac{(x-\alpha)}{\beta}\right)\right] \quad -\infty < x < \infty \quad (3.16)$$

情况 1: β 已知, α 未知;

情况 2: α 已知, β 未知;

情况 3: α, β 未知。

α, β 用极大似然估计 $\hat{\alpha}, \hat{\beta}$ 。由此可求出 $z_i = F(x_{(i)}; \hat{\theta})$, 由 (3.12) 计算 A_n^2 , 并进行修正求出 A^* , A^* 与检验的临界值表见表 4^[17]。

表 4 A^* 界临界值表

修正 A^*		0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.025
情况 1	$A^* = A_n^2(1 + 0.3/n)$	0.736	0.816	0.916	1.002	1.321	1.591	1.959	2.244	2.534
情况 2	无	1.060			1.725	2.277	2.854	3.640		
情况 3	$A^* = A_n^2(1 + 0.2/5n)$	0.474			0.637	0.757	0.877	1.038		

IV、威布尔分布:

$$F(x; \theta) = 1 - \exp \left[- \left(\frac{x - \alpha}{\beta} \right)^\gamma \right] \quad x > \alpha \quad (3.17)$$

当 α 已知时, 令 $\gamma = -\ln(x - \alpha)$, 于是

$$F(y) = \exp \left[- \exp \left(- \frac{y - \alpha'}{\beta'} \right) \right] \quad y > \alpha'$$

$\alpha' = -\ln\beta, \beta' = \frac{1}{\gamma}$, 上述分布 $F(y)$ 是极值分布。

只要令 $y_{(i)} = -\ln(x_{(n+1-i)} - \alpha)$, $i = 1, 2, \dots, n$

$y_{(i)}$ 是极值分布次序统计量(从小到大)。用 III 作检验。

对于参数 θ 已知, 而子样是截尾的情形, Pettitt 和 Stephens^[22] (1976年) 讨论了 W^2 、 A^2 和 U^2 统计量, 并对这方面的工作作了总结, 给出了可供检验用的图表, 讨论了它们的渐近分布以及有关渐近分布的上、下百分位点。

对定时截尾, 到 α 截止, 即:

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(R)} \leq \alpha, \quad F(\alpha; \theta) = p$$

$z_i = F(x_{(i)}; \theta)$ 于是检验的统计量为:

$${}_p W_n^2 = \sum_{i=1}^R \left(z_i - \frac{2i-1}{2n} \right)^2 - \frac{R(4R^2-1)}{12n^2} + n \left[p \frac{R^2}{n^2} - p^2 \frac{R}{n} + \frac{p^3}{3} \right] \quad (3.18)$$

$${}_p A_n^2 = \sum_{i=1}^R \left(\frac{2i-1}{n} \right) [\ln(1-z_i) - \ln z_i] - 2 \sum_{i=1}^R \ln(1-z_i)$$

$$+ n \left[\frac{2R}{n} - \left(\frac{R}{n} \right)^2 - 1 \right] \ln(1-p) + \frac{R^2}{n} \ln p - pn \quad (3.19)$$

$${}_p U_n^2 = {}_p W_n^2 - np \left[\frac{R}{n} - \frac{p}{2} - \frac{1}{pn} \sum_{i=1}^R z_i \right]^2 \quad (3.20)$$

对于定数截尾, 只要将 $Z_r = F(x_{(r)}; \theta)$ 代替 p 和 $\gamma - 1 = R$ 将可。

EDF 统计量在一大类对立假设范围内, 比 χ^2 检验更有效, 具有一开始提出的四个要求。Stephens^[1] 用随机模拟方法比较了统计量的势, 在一类对立假设中, EDF 统计量中, $K-S$ 检验比 A^2 、 W^2 、 U^2 的效要差些。

四、用似然比检验区分分布

我们用 EDF 统计量或其它检验分布的方法, 有时会出现同时接受几个分布, 或者我们希望在两种分布中间选择一个较合适的分布。对于这种情况可以用似然比法, 对于具有未知位置和尺度参数的情况, 以及区分对数正态和威布尔分布的情况^{[23][24]} 都已有可供实际需要的检验用表。

1. 似然比检验

设随机变量 X , 子样 x_1, x_2, \dots, x_n 。 X 的密度函数为 $f_0(x; a, b)$ 或 $f_1(x; a, b)$ 。根据子样决定 X 的分布, 假设

$$H_0: X \sim f_0(x; a, b)$$

$$H_1: X \sim f_1(x; a, b)$$

极大似然比

$$RML = \frac{\max_{a,b} \prod_{i=1}^n f_1(x_i; a, b)}{\max_{a,b} \prod_{i=1}^n f_0(x_i; a, b)} \quad (4.1)$$

设 a, b 为未知的位置和尺度参数, 即

$$f_i(x; a, b) = \frac{1}{b} g_i\left(\frac{x-a}{b}\right), \quad i=0, 1 \quad (4.2)$$

则 RML 的分布与未知参数 a, b 独立。^[23] 给出了区分正态和柯希, 正态和指数, 正态和双指数分布的检验的临界值表和效。

2. 区分正态分布和指数分布

$H_0: X \sim N(a, b)$, 即密度函数为:

$$f_0(x; a, b) = \frac{1}{\sqrt{2\pi b}} e^{-\frac{(x-a)^2}{2b^2}}$$

$H_1: X \sim$ 指数分布, 即密度函数为:

$$f_1(x; a, b) = \frac{1}{b} e^{-(x-a)/b}$$

由子样 x_1, x_2, \dots, x_n , 获得正态分布 a, b 的极大似然估计为:

$$\hat{a}_0 = \frac{1}{n} \sum_{i=1}^n x_i = \bar{x}, \quad \hat{b}_0 = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$$

指数分布参数 a, b 的极大似然估计为:

$$\hat{a}_1 = \min_{1 \leq i \leq n} x_i, \quad \hat{b}_1 = \frac{1}{n} \sum_{i=1}^n (x_i - \min x_i)$$

于是

$$\begin{aligned} RML &= \frac{\prod_{i=1}^n \left(\frac{1}{\hat{b}_1}\right) e^{-\sum_{i=1}^n (x_i - \hat{a}_1)/\hat{b}_1}}{\left(\frac{1}{\sqrt{2\pi} \hat{b}_0}\right)^n e^{-\sum_{i=1}^n (x_i - \hat{a}_0)^2/2\hat{b}_0^2}} \\ &= \left(\frac{\hat{b}_0 \sqrt{2\pi}}{\hat{b}_1}\right)^n e^{-\frac{\sum_{i=1}^n (x_i - \hat{a}_1)}{\hat{b}_1}} = (\sqrt{2\pi/e})^n \left(\frac{\hat{b}_0}{\hat{b}_1}\right)^n \end{aligned}$$

可用统计量

$$D = \frac{\hat{b}_0}{\hat{b}_1} = \sqrt{\frac{\sum_{i=1}^n (x_i - \bar{x})^2}{\sum_{i=1}^n (x_i - \min x_i)}} \quad (4.3)$$

D 的临界值和效见下表5。

表5 H_0 : 正态, H_1 : 指数, $D \geq D_c$ 拒绝 H_0

n	$\alpha=0.01$		$\alpha=0.05$		$\alpha=0.10$	
	D_c	效	D_c	效	D_c	效
10	1.01	0.39	0.87	0.65	0.80	0.77
15	0.88	0.65	0.77	0.80	0.72	0.93
20	0.80	0.86	0.71	0.96	0.67	0.98
25	0.76	0.94	0.68	0.99	0.64	0.99
30	0.72	0.98	0.65	1.00	0.61	1.00

3. 区分正态分布和双边指数分布

 $H_0: X \sim N(a, b)$ $H_1: X \sim$ 双边指数分布, 即密度函数为:

$$f_1(x; a, b) = \frac{1}{2b} \exp\{-|x-a|/b\} \quad -\infty < x < \infty$$

极大似然比检验的统计量为 G :

$$G = \frac{\hat{b}_0}{\hat{b}_1} = \sqrt{n} \frac{\sum (x_i - \bar{x})^2 / \sum |x_i - x_{med}|}{\hat{b}_1} \quad (4.4)$$

 G 的临界值和效见表6。表6 H_0 : 正态, H_1 : 双边指数, $G \geq G_c$ 拒绝 H_0

n	$\alpha=0.01$		$\alpha=0.05$		$\alpha=0.10$	
	G_c	效	G_c	效	G_c	效
21	1.47	0.21	1.39	0.41	1.36	0.53
31	1.42	0.33	1.37	0.54	1.34	0.65
41	1.41	0.41	1.36	0.64	1.33	0.75
51	1.39	0.52	1.34	0.73	1.32	0.82
61	1.31	0.61	1.33	0.80	1.31	0.88

4. 区分对数正态分布和威布尔分布^[24] $H_0: X \sim$ 对数正态分布, 即

$$f_0(x; \mu, \sigma) = \frac{1}{x\sqrt{2\pi}\sigma} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}}, \quad x > 0$$

 $H_1: X \sim$ 威布尔分布, 即

$$f_1(x; n, \eta) = \frac{m}{\eta} \left(\frac{x}{\eta}\right)^{m-1} e^{-(x/\eta)^m} \quad x > 0$$

 μ, σ 的极大似然估计为:

$$\mu = \frac{1}{n} \sum \ln x_i, \hat{\sigma}^2 = \frac{1}{n} \sum (\ln x_i - \hat{\mu})^2$$

用叠代法求得威布尔分布参数 m 、 η 的极大似然估计为 \hat{m} 、 $\hat{\eta}$ 。

似然比函数：

$$RML = (2\pi e \hat{\sigma}^2)^{n/2} \prod_{i=1}^n \frac{1}{x_i} f_1(x_i; \hat{m}, \hat{\eta}) \tag{4.5}$$

Dumonceaux 和 Antle^[23] 给出了统计量

$$D = (RML)^{1/n} = (2\pi e \hat{\sigma}^2)^{1/2} \left[\prod_{i=1}^n \frac{1}{x_i} f_1(x_i; \hat{m}, \hat{\eta}) \right]^{1/n}$$

的临界值，见表 7。

表 7 H_0 : 对数正态 H_1 : 威布尔 $D \geq D_c$ 拒绝 H_0

n	$\alpha=0.01$		$\alpha=0.05$		$\alpha=0.10$		$\alpha=0.20$	
	D_c	效	D_c	效	D_c	效	D_c	效
20	1.144	0.22	1.082	0.48	1.038	0.61	1.015	0.75
30	1.095	0.39	1.044	0.63	1.020	0.75	0.993	0.86
40	1.070	0.53	1.028	0.76	1.007	0.85	0.984	0.93
50	1.054	0.63	1.014	0.83	0.998	0.91	0.976	0.96

上述统计量可用于区分正态分布和极值分布，只要将数据作变换，令 $z_i = \ln x_i$ 即可。

五、子样偏度 $\sqrt{b_1}$ 和峰度统计量 b_2

1. 分布函数及其偏峰度

设分布函数 $F(x)$ ，密度函数 $f(x)$ ，其 S 价中心矩 μ_3 ，于是分布的偏度和峰度为：

$$\sqrt{\beta_1} = \mu_3 / \mu_2^{3/2} \text{ 和 } \beta_2 = \mu_4 / \mu_2^2 \tag{5.1}$$

$\sqrt{\beta_1}$ 和 β_2 能反映密度函数的形态，左右对称的分布密度函数 $\sqrt{\beta_1} = 0$ ，而 β_2 反映出分布密度曲线在靠近中心处的平坦程度。对正态分布来说， $\sqrt{\beta_1} = 0$ ， $\beta_2 = 3$ 。

由于 μ_3 及 $\sqrt{\beta_1}$ 、 β_2 能反映 $f(x)$ 的形状，或能确定分布函数的形式。K. Pearson 分布系密度函数 $f(x)$ 满足下列微分方程^{[25][26]}。

$$\frac{1}{y} \frac{dy}{dx} = \frac{-(x+c_1)}{c_0+c_1x+c_2x^2} \tag{5.2}$$

其中参数 c_0 、 c_1 、 c_2 是前四阶中心矩 μ_r 的函数，并可以用 β_1 、 β_2 来表示。因此，对于 Pearson 分布系来说，知道 β_1 、 β_2 便可确定其概率密度函数。

Johnson 等人^[26](1963) 在 $\sqrt{\beta_1}$ 和 β_2 组成的直角坐标系上，给出了 Pearson 分布系的 $\sqrt{\beta_1}$ 、 β_2 所在的区域，并根据给定的 $\sqrt{\beta_1}$ 和 β_2 计算了 Pearson 分布系的分位点表。Hahn 和 Shapiro^[27](1967) 在 β_1 和 β_2 组成的坐标系上给出了正态分布， t 分布， β 分布，均匀分布， γ -分布，指数分布和对数正态分布的 β_1 和 β_2 所在的区域。

对于威布尔分布来说， β_1 、 β_2 仅是形状参数 m 的函数

$$\sqrt{\beta_1} = \frac{\gamma_3 - 3\gamma_2\gamma_1 + 2\gamma_1^3}{(\gamma_2 - \gamma_1^2)^{3/2}} \quad (5.3)$$

$$\beta_2 = \frac{\gamma_4 - 4\gamma_3\gamma_1 + 6\gamma_2\gamma_1^2 - 3\gamma_1^4}{(\gamma_2 - \gamma_1^2)^2} \quad (5.4)$$

其中 $\gamma_k = \gamma \left(1 + \frac{K}{m}\right)$ 。

Rousu^[28] (1973) 列出了 $m = 0.2(0.2)9(1)12, 15, 25$ 时对应的 β_1 和 β_2 的数值, 并给出在 β_1 、 β_2 所成的直角坐标系上威布尔分布的 β_1 、 β_2 所在的区域。

威布尔分布形状参数 $m = 3.60235$ 时, $\beta_1 = 0$ 。当 $m > 3.60235$ 有正偏度^[29]。Dubey^[30] (1966) 研究了正态分布和威布尔分布的关系, 当 m 在 3.25 至 3.61 内时, 正态分布可以作为威布尔分布的很好的近似, 在 $m = 3.60232$ 时标准正态分布和标准威布尔分布的百分位点十分接近, 在随机变量取值变化在 $-3.0(0.1)3$ 中, $\text{sup}|F_W(u) - F_N(u)| = 0.0078$ 。

对于分布函数的 $\sqrt{\beta_1}$ 、 β_2 的研究, 为分布的检验和选择提供了不少信息。例如, 我们可由一组数据, 先计算子样的偏度和峰度, 看其接近那一个分布的偏、峰度, 然后再用其它方法(如 x^2 , EDF 统计量等)作检验。

2. 子样 $\sqrt{b_1}$ 和 b_2 以及正在态分布检验中的应用

子样偏, 峰度 $\sqrt{b_1}$, b_2 。对正态分布来说, 关于 $\sqrt{b_1}$ 、 b_2 的分布早有研究, Fisher^[31] (1930) 证明了, 当 $\gamma \geq 3$ 时, $m_3/m_1^{3/2}$ 与 m_2 独立 (m_2 为子样 γ 阶中心矩)。所以 $E b_1 = E m_3^2 / E m_2^2$, 根据这个性质, 可以求出 $\sqrt{b_1}$ 、 b_2 的各阶矩, Fisher 求出了 $\sqrt{b_1}$ 的前六阶矩和 b_2 的前四阶矩, 不久 Hsu 和 Lawley^[32] 又求出了 b_2 的五阶和六阶矩。

根据前四阶矩的数值, 可以找到 $\sqrt{b_1}$ 、 b_2 的近似分布。Pearson^[33] 化了很长一段时期研究适宜于来自正态分布的子样偏度 $\sqrt{b_1}$ 和峰度 b_2 的分布, 使分布的前四阶矩与 $\sqrt{b_1}$ 、 b_2 的前四阶矩相等或十分接近, 研究结果表明, 由于 $\sqrt{b_1}$ 是对称分布, 而 b_2 是单边的, 此两种分布类型不同: 对 $\sqrt{b_1}$ 用皮尔逊 VII 型或 t 分布是适宜的; 对 b_2 用皮尔逊 VI 型或 IV 型是适宜的。并给出了 $\sqrt{b_1}$ 、 b_2 的可接受的 1%、5% 分位点。

Pearson^[35] 提出利用 Johnson^[36] 变换, 对 $\sqrt{b_1}$ 作等价标准正态变换:

$$X(\sqrt{b_1}) = \delta \sinh^{-1} \left(\frac{\sqrt{b_1}}{\lambda_1} \right) \quad (5.5)$$

$X(\sqrt{b_1}) \sim N(0, 1)$, δ 、 λ 为参数。^[37] 给出了当子样来自正态分布时说明 δ , $1/\lambda$ 与 n 有关; 并给出了 $n = 8(1)50(2)100(5)250(10)500(20)1000$ 时, δ 和 $1/\lambda$ 的数值表。Dagostino^[34] 提供了 $\sinh^{-1}x$ 的表, 这样可用正态分布的分位点对 $\sqrt{b_1}$ 作检验。

当子样来自正态时, Dagostino 和 Pearson^[37] (1973) 还给出了 b_2 分布的经验结果, 给出了可供检验正态性的 b_2 的图表。

3. 通用(Omnibus)检验

1973 年 Dagostino 和 Pearson 提出当 $n \geq 20$ 时, 对 $\sqrt{b_1}$ 、 b_2 建立统计量:

$$Y^2 = X^2(\sqrt{b_1}) + X^2(b_2) \quad (5.6)$$

在此子样来自正态总体, $X(\sqrt{b_1})$ 、 $X(b_2)$ 为由 Johnson 变换所得到的等价标准正态变量。于是 $X^2(\sqrt{b_1})$ 、 $X^2(b_2)$ 为自由度 1 的 x^2 变量, 而 Y^2 为自由度 2 的 x^2 变量。但后一假设是错误的, 因为 $X^2(\sqrt{b_1})$ 和 $X^2(b_2)$ 并不独立。他们在 1974 年^[38] 修正了这个错误。

1975 年 Bowman 和 Shenton^[39] 修正了 Y^2 , 对 $\sqrt{b_1}$ 、 b_2 作等价标准化正态变换, $X(\sqrt{b_1})$

用(5.5)式。对 b_2 分两种情况,当 $n \geq 25$ 时用

$$X_s(b_2) = \gamma_2 + \delta_2 \sinh^{-1} \left(\frac{b_2 - \xi}{\lambda_2} \right) \quad (5.7)$$

$n < 25$ 时用

$$X_e(b_2) = \gamma_2 + \delta_2 \log \left(\frac{b_2 - \xi}{\xi + \lambda_2 - b_2} \right) \quad (5.8)$$

随机模拟计算表明 $X(\sqrt{b_1})$ 和 $X(b_2)$ 的正态性是好的(见表8)。从表8可见,从前四阶矩可知 $X(b_1)$ 的正态是好的,对 b_2 来说 $X_e(b_2)$ 适宜小子样, $X_s(b_2)$ 对大子样较适用。从它们的分布的百分位点来说也与标准正态分布的百分位点较接近(见表9)。

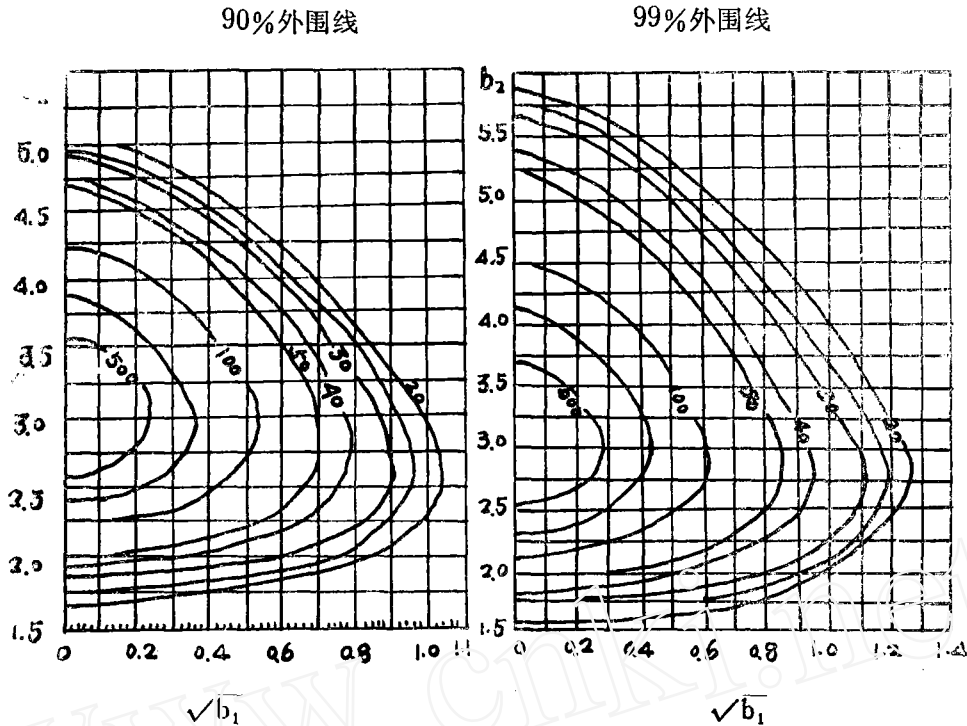
表 8

n	变 量	μ_1	σ	$\sqrt{\beta_1}$	β_2
20	$\sqrt{b_1}$	0.000	0.472	-0.001	3.54
	$X(\sqrt{b_1})$	0.001	0.998	-0.002	5.03
	b_2	2.708	0.767	1.840	9.52
	$X_s(b_2)$	-0.018	1.033	-0.257	4.00
	$X_e(b_2)$	-0.008	1.006	0.040	3.22
50	$\sqrt{b_1}$	0.000	0.327	0.002	3.57
	$X(\sqrt{b_1})$	0.001	1.001	0.001	3.06
	b_2	2.880	0.609	1.673	8.00
	$X_s(b_2)$	-0.042	1.011	0.055	2.89
	$X_e(b_2)$	-0.020	1.011	0.057	3.06

表 9

n	变 量	1%	5%	10%	90%	95%	99%
20	$X(\sqrt{b_1})$	-2.326	-1.644	-1.278	1.271	1.636	2.328
	$X_e(b_2)$	-2.316	-1.651	-1.289	1.262	1.618	2.333
	标准正态 X	-2.326	-1.645	-1.282	1.282	1.645	2.326
50	$X(b_1)$	-2.329	-1.636	-1.280	1.282	1.647	2.345
	$X_s(b_2)$	-2.305	-1.707	-1.358	1.251	1.623	2.327
	$X_e(b_2)$	-2.311	-1.671	-1.310	1.283	1.645	2.355

Bowman 和 Shenton 用(5.6)式,在不假设 $X^2(\sqrt{b_1})$ 和 $X^2(b_2)$ 独立的条件下,用随机模拟方法给出 Y^2 分布的 1%, 5%, 10%, 90%, 95%, 99% 的分位点。并将它们对应到 $(\sqrt{b_1}, b_2)$ 平面上,求得相应 90%, 95%, 99% 的外围线,而可直接用 $\sqrt{b_1}, b_2$ 的值作通用检验。



4. 非正态子样的通用检验

对非正态子样统计量 $\sqrt{b_1}$ 、 b_2 的分布, Bowman 和 Shenton^[40] 作了讨论。研究了某些分布的 $\sqrt{b_1}$ 、 b_2 的矩, 并建议用统计量

$$Y^2 = (X_1^2 - 2RX_1X_2 + X_2^2) / (1 - R^2) \quad (5.9)$$

作通用检验, 其中 X_1 、 X_2 为 $\sqrt{b_1}$ 、 b_2 的 Johnson Su 等价近似标准正态变量, R 为 X_1 与 X_2 的相关系数。他们计算了一些分布, 如皮尔逊 I 型, $\sqrt{\beta_1} = 2/7$, $\beta_2 = 33/14$, $n = 100$ 时 Y^2 的外围线。还计算了混合正态和均匀分布的情况, 详见[40]。

有关分布检验的方法很多, 例如 Shapiro 和 Wilk 在 1965 年提出的回归方法^{[41][42]}, 将被检验分布的次序统计量与子样次序统计量之间的回归来作分布检验, 详见[41]、[42]。由于查阅资料有限, 以及有关分布检验的方法和理论内容很多, 限于水平, 不能加以全部综合。

参 考 文 献

- [1] Stephens, M. A (1978), Goodness of Fit Tests with special Reference to Tests for exponentiality. AD-AO 60824.
- [2] Mihalko, D. P et al. (1977). Chi-square tests of fit for type II censored data. Mimeograph Series No. 505. Department of statistics. Purdue University and Biometrics 34 (1978). Abstract 2616, 163.
- [3] Lamborn, K (1969). On Chi-squared goodness of fit tests for sampling from more than one population with possibly censored data. Stanford University Technical Report No. 21, (PHS 2 Tol GMOOO 25-11).
- [4] 第四机械工业部标准化研究所(1979)“可靠性试验用表”, 国防工业出版社。
- [5] David, F. N et al (1948). The probability integral transformation when parameters are estimated from the sample. Biometrika 35. 182-190.
- [6] Lilliefors, H. W (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown: J. Am. Statist. Assoc 62. 399.
- [7] Lilliefors, H. W (1969). On the Kolmogorov-Smirnov test for the exponential distribution with mean

- unknown: J. Am. Statist. Assoc, 64, 387—389.
- [8] Barr, D. R. et al (1973). Kolmogorov Test for censored samples. *Technometrics* 15, 739—757.
- [9] Koziol, J. A et al (1975). Percentage points of the asymptotic distributions of one and two sample K-S statistics for truncated or censored data, *Technometrics* 17, 507—510.
- [10] Durbin, J. et al (1972). Components of the Cramer-Von Mises Statistics, I. *J. R. statist. Soc. B.* 34, 290—307.
- [11] Durbin, J. Knott, M. et al (1975), Components of Cramer-Von Mises Statistics, II. *J. R. Statist, Soc. B.* 37, 216—237.
- [12] Anderson, T. W. et. al (1952). Asymptotic theory of certain goodness-of-fit criteria based on stochastic processes. *Ann. Math. Statist.* 23, 193—212.
- [13] Watson, G. S(1961). Goodness-of-fit tests on a circle. *Biometrics* 48, 109—114.
- [14] Stephens, M. A(1963). The distribution of the goodness-of-fit statistic U_N^2 , *I. Biometrika* 50, 303—313.
- [15] Anderson, T. W et al (1954). A test of goodness of fit. *J. Am. Statist. Assoc* 49 765—769.
- [16] Lewis, P. A. W(1961). Distribution of the Anderson-Darling Statistic. *Ann Math. Statist* 32, 1118—1124.
- [17] Stephens, M. A (1980). The Anderson-Darling Statistic, AD-A079807.
- [18] Stephens, M. A(1976a). Asymptotic results for goodness-of-fit statistics with unknown parameters. *Ann, statist.* 4, 357—369.
- [19] Durbin, J(1975). Kolmogorov-Smirnov tests when parameters to tests of exponentiality and tests on spacings. *Biometrika* 62, 5—22.
- [20] Pettitt, A. N et al (1976). EDF tests for the gamma distribution with application to testing for equal variances. submitted to *J. Amer statist, Assoc.*
- [21] Stephens, M. A (1976b) [Goodness-of-fit for the extreme value distribution. Stanford University Technical Report, Department of statistics, Stanford University. To appear in *Biometrika*.
- [22] Pettitt, A. N et al (1976). Cramer-Von Mises type statistics for the goodness-of-fit of censored data simple hypotheses. AD-A025732.
- [23] Dumonceaux, R, Antle, et al (1973). likelihood ratio test for discrimination between two models with unknown location and scale parameters. *Technometrics* 15, 19—27.
- [24] Dumonceaux, R et al (1973). Discrimination between the Log-Normal and the weibull distributions. *Technometrics* 15, 923—926.
- [25] H. 克拉美, 统计学数学方法, 上海科技出版社, 1966.
- [26] Johnson, N. L, et al (1963). Table of percentage points of pearson curves on curves, for given $\sqrt{\beta_1}$ and β_2 , expressed in standard measure. *Biometrika* 50, 459.
- [27] Hahn, G. J et al (1967). *Statistical Models in Engineering*. Wiley Sons. Inc, New York.
- [28] Rousu, D. N (1973). Weibull skewness and Kurtosis as a function of the shape parameter. *Technometrics* 15, 927—930.
- [29] Cohen, A. C (1973). The reflected Weibull distributions. *Technometrics* 15, 867—873.
- [30] Dubey, S. D (1966). Normal and Weibull distributions. *Naval Research Logistics Quarterly* 14, 69—79.
- [31] Fisher, R. A (1930). The moments of the distribution for normal samples of measures of departures from normality. *Proc. Royal society, Series A*, 130, 17—28.
- [32] Hsu, C. T et al (1939). The derivation of the fifth and sixth moments of b_2 in samples from a normal population. *Biometrika* 31.
- [33] Pearson, E. S (1963). Some problems arising in approximating to probability distribution, Using moments. *Biometrika* 50, 95—111.
- [34] Dagostino, R. B (1970). Transformations to normality of the null distribution of g_1 . *Biometrika* 57, 679—681.
- [35] Pearson, E. S (1965). Tables of percentage points of $\sqrt{b_1}$ and b_2 in normal samples: a rounding off. *Biometrika* 52, 282—285.
- [36] Johnson, N. L (1965). Tables to facilitate fitting S frequency curves. *Biometrika* 52, 547—558.
- [37] Dagostino, R. B et al (1973). Tests for departures from normality Empirical results for the distribution of b_2 and $\sqrt{b_1}$. *Biometrika* 60, 612—622.
- [38] Dagostino, R. B et al (1974). Correction and amendment "Test for departures from normality empirical results for the distribution of b_2 and $\sqrt{b_1}$ ". *Biometrika* 61, 647.

- [39] Bowman, K. O. et al (1975). Omnibus test contours for departures from normality based on $\sqrt{b_1}$ and b_2 . *Biometrika* 62, 243—250.
- [40] Bowman, K. O et al (1976). Remarks on the distribution of skewness and kurtosis statistics in general. *ConF-751113-6*.
- [41] Shapiro, S. S et al (1965). An analysis of variance test for normality, *Biometrika* 42, 591—611.
- [42] Shapiro, S. S et al (1972). An analysis of variance test for the exponential distribution. *Technometrics* 14, 355—370.

www.cnki.net