基于偏态分布的百分位估计公式的建立*

周 云[1)] 侯 威[2)†] 钱忠华[1)] 何文平[2)]

1)(扬州大学物理科学与技术学院,扬州 225009)

2)(国家气候中心,北京 100081)

(2010 年 10 月 25 日收到; 2011 年 2 月 28 日收到修改稿)

顺序统计量将累积概率与数据排序后的位置建立相关联系,可用于估计数据分布的累积概率. 鉴于不同气候要素概率分布存在着不同程度的偏态特征,基于偏态分布条件下的累积概率函数,通过理论推导和数值模拟建立了与偏态指数相关的位置参数的回归模型,从而给出了基于数据偏态特征的经验百分位估计公式. 利用 1980 年—2009年全球夏季逐日平均温度资料,进一步对比分析了偏态百分位估计方法与 Jenkinson 方法下得到的第 90 个百分位值所对应的温度排序后位置的差异.

关键词: 顺序统计量, 偏态分布, 百分位

PACS: 92. 60. Wc

1. 引 言

1977 年,Jenkinson[1] 运用经验排序法给出了基于数据排序后所处位置的累积概率分布的经验估计. 根据该经验估计,对于长度为 n 的时间序列,将这 n 个元素按升序排列,则发生小于或等于排序后第 m 个元素的概率为

$$p = \frac{m - 0.31}{n + 0.38}. \tag{1}$$

若时间序列长度 n 等于 100,那么第 90 个百分位值对应的数据排序后的位置为 90.65,介于元素 x_{90}(p = 89.35%) 与 x_{91}(p = 90.35%) 之间. (1)式在应用时无需时间序列的具体统计模型,且计算方便,近年来被广泛应用于各种极端事件百分位阈值的估计以及回归时间的确定等研究中[2—6]. 但值得注意的是,只有当所分析的数据服从或近似服从高斯分布时,利用(1)式才能得到较准确的结果[7,8]. 当某一要素所服从的概率分布函数已知时,则可给出在这一特定分布下累积概率的经验百分位值估计. 如 Goel 等[9] 给出了广义极值分布条件下的经验百分位估计式,该估计式不仅与数据排序后的位置有关,而且还与广义极值分布中的形状参数相关联.

大气系统是一个复杂的耗散的非线性系统[10—17],数学模型总是不可能与实际情况完全一致. 在研究温度时间序列时,通常以高斯分布作为其统计模型,但事实上对于不同的地区,温度序列所表现出的高斯分布特征存在着不同程度的偏态性质,若一概以高斯分布作为其统计模型,势必会在统计及相关的分析过程中带来较大偏差,而偏态分布函数可以用偏态指数来表征偏态分布的特征,从很大程度上可以避免这一问题[18]. 本文以偏态分布函数作为温度时间序列的统计模型,针对偏态分布条件下的累积概率分布函数,通过理论推导和数值模拟建立了温度序列偏态分布条件下的经验百分位估计公式,从而给出与偏态指数相关的经验百分位值估计. 并进一步对比分析了 1980 年—2009 年全球各格点夏季逐日平均温度资料的第 90 个百分位值所对应的温度顺序统计量位置在偏态百分位估计方法和 Jenkinson 估计方法[1] 下的异同.

2. 资料和方法

2.1. 资 料

本文使用的温度资料来自美国国家环境预报

* 国家自然科学基金(批准号:40930952,41005043)和国家科技支撑计划(批准号:2007BAC29B01)资助的课题。

† 通讯联系人. E-mail: hou_w@sohu.com

中心和美国国家大气研究中心联合发布的全球逐日平均温度再分析资料(地面资料),分辨率为 2.5° ×2.5°,选取 1980 年—2009 年夏季逐日平均温度为研究对象.

2.2. 偏态概率密度函数

采用偏态分布函数作为全球 1980 年—2009 年夏季逐日平均温度资料的统计模型,其概率密度函数为

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma(\mu - a)} \times \exp\left(- \frac{\left(\left(\frac{x-a}{\mu-a}\right)^{\lambda} - 1\right)^2}{2\sigma^2\lambda^2} \right)\left(\frac{x-a}{\mu-a}\right)^{\lambda-1}. \quad (2)$$

(2)式由经偏态数据在 Box-Cox 变换后所满足的正态假定条件推导而来,其中 a 为对温度时间序列的平移量,对每个格点上的平移量,本文取为与各格点 1980 年—2009 年夏季逐日平均温度最小值偏差为 0.1 ℃的值,以保证经过 Box-Cox 变换得到的数据皆为正值;μ 为原始温度时间序列的均值;σ 为经 Box-Cox 变换后新序列的标准差;λ 为 Box-Cox 变换中通过极大似然法求得的偏态指数,用它表征概率分布的偏态性质. 如图 1 所示,当 $\lambda < 1$ (图 1(a)),表明温度时间序列呈正偏分布,分布的右侧平缓,左侧陡峭;当 $\lambda > 1$ (图 1(b)),表明温度时间序列呈负偏分布,其分布特征与正偏分布相反. λ 值偏离 1 越大,表明分布的偏态性越显著.

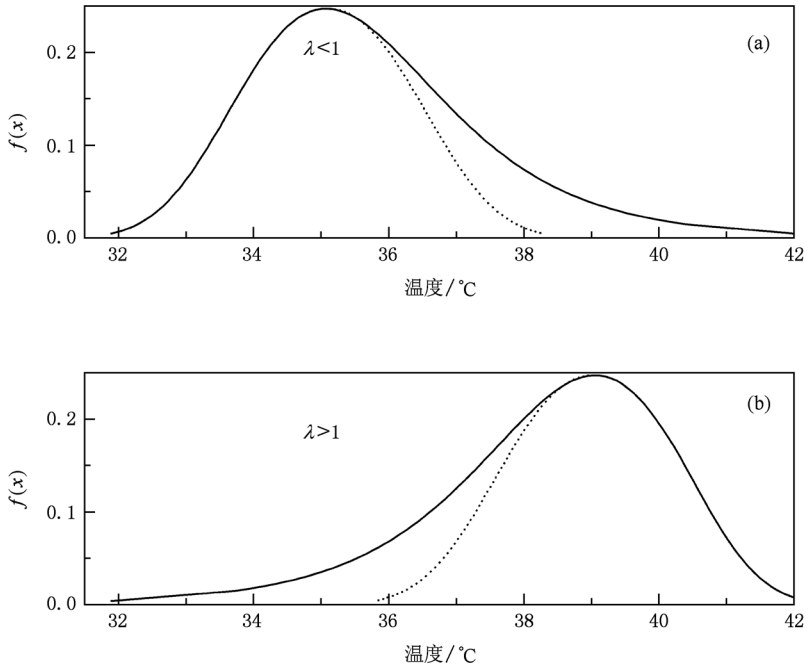

图 1 偏态分布示意图 (a)正偏分布型,(b)负偏分布型

当 $\lambda = 1$,表明温度时间序列呈高斯分布,左右两侧对称,此时(2)式即可转化为高斯分布函数

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma(\mu - a)}\exp\left(- \frac{(x-\mu)^2}{2\sigma^2(\mu-a)^2}\right). \quad (3)$$

图 2 为 1980 年—2009 年全球各格点夏季逐日平均温度的偏态指数分布情况. 从图 2 可以看出,全球大部分地区呈现负偏,即 $\lambda > 1$,表明夏季发生的最概然温度(偏态分布函数的概率密度最大处所对应的温度,即最有可能发生的温度事件)普遍高于温度序列的均值,夏季更倾向于发生高于均温的温度事件,而低于均温的温度事件的极端程度相对较高,且不容易发生. 其中在 60°S 附近负偏呈现比较明显的带状分布,该区域最概然温度与平均温度的偏离更为明显;而在极地以及赤道附近部分地区则呈现正偏分布,夏季更易发生低于均温的温度事件.

2.3. 偏态分布函数

偏态分布函数可由(1)式偏态概率密度函数积分得到,

$$F(x) = \int_a^x f(x)\,\mathrm{d}x$$

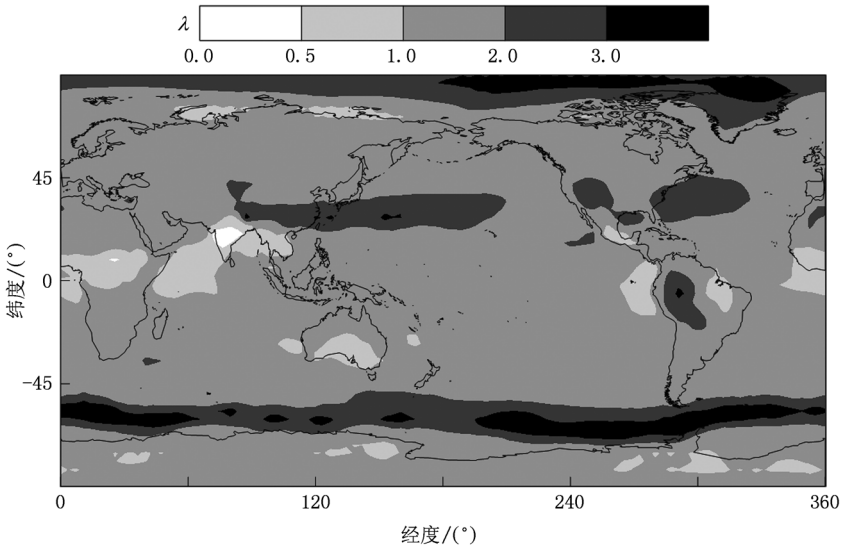

图2 1980年—2009年全球夏季逐日平均温度的偏态指数的空间分布

$$= \int_a^x \frac{1}{\sqrt{2\pi}\sigma(\mu - a)} \times \exp\left(- \frac{\left(\left(\frac{x-a}{\mu-a}\right)^{\lambda} - 1\right)^2}{2\lambda^2\sigma^2} \right)\left(\frac{x-a}{\mu-a}\right)^{\lambda-1} \mathrm{d}x$$

$$= \int_a^x \frac{1}{\sqrt{2\pi}\sigma\lambda} \exp\left(- \frac{\left(\left(\frac{x-a}{\mu-a}\right)^{\lambda} - 1\right)^2}{2\lambda^2\sigma^2} \right) \times \mathrm{d}\left(\frac{x-a}{\mu-a}\right)^{\lambda}. \quad (4)$$

令

$$y = \left(\frac{x-a}{\mu-a}\right)^{\lambda},$$

代入(4)式可得

$$F(x) = \int_0^y \frac{1}{\sqrt{2\pi}\lambda\sigma} \exp\left(- \frac{(y-1)^2}{2\sigma^2\lambda^2} \right) \mathrm{d}y. \quad (5)$$

根据指数函数二次幂的积分公式

$$F(x) = \frac{1}{\sigma\sqrt{2\pi}} \int \exp\left(- \frac{(x-\mu)^2}{2\sigma^2} \right) \mathrm{d}x = \frac{1}{2}\left(1 + \mathrm{erf}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right), \quad (6)$$

可得

$$F(x) = \frac{1}{2}\left(\mathrm{erf}\left(\frac{y-1}{\lambda\sigma\sqrt{2}}\right) + \mathrm{erf}\left(\frac{1}{\lambda\sigma\sqrt{2}}\right)\right). \quad (7)$$

这里 erf(·) 为误差函数,

$$\mathrm{erf}(x) \approx \frac{x}{|x|}\sqrt{1 - \exp\left(- x^2 \frac{4/\pi + \alpha x^2}{1 + \alpha x^2} \right)}. \quad (8)$$

可以由(8)式近似求解误差函数,其中 $\alpha \approx 0.14$.

3. 理论基础

3.1. 顺序统计量

Jenkinson[1]对累积概率给出形如 $P_m = \frac{m+a}{n-b}$ 的估计,这为样本中某一元素对应的百分位与其排序后的位置两者之间建立联系. 这里参数 a 和 b 即为对与百分位值相应的排序后第 m 个元素位置的微调,称为位置参数,需要基于顺序统计量的相关理论来推导并求解位置参数 a 和 b. 对样本的 n 个元素,按从小到大排序为 $x_1 \leqslant x_2 \leqslant \cdots \leqslant x_m \leqslant \cdots \leqslant x_n$, 则排序后的序列或其中的部分称为顺序统计量,它主要用于构造样本的经验分布,并估计经验分布的百分位值.

设样本序列的概率密度函数和分布函数分别为 $f(x)$ 和 $F(x)$, 用 100α 代表百分位 ($0 < \alpha < 1$),累积概率 $F(\xi_\alpha) = \alpha$. 若分布函数 F 已知,则任意 α 对应的百分位值 ξ_α 均可直接确定. 但实际应用中更容易获知样本的经验分布,如何用 x_m 来估计 ξ_α, 即对百分位值作经验估计,找出 p 与第 m 个顺序统计量 x_m 之间的关系,使得 $p = F(x_m)$, 是一个需要解决的问题.

设原始的 n 个顺序统计量之间相互独立,其对应的累积概率分别为 $y_1 = F(x_1)$, …, $y_n = F(x_n)$, 因分布函数 F 为连续单调递增函数,有

$$y_1 \leqslant \cdots \leqslant y_m \leqslant \cdots \leqslant y_n, \tag{9}$$

则 $y \in [0,1]$, 服从 0—1 之间的均匀分布[19], 即

$$\begin{aligned} f(y) &= 1, \\ F(y) &= y. \end{aligned} \tag{10}$$

y_m 即可视为位于 $[0,1]$ 之间且呈均匀分布的 n 个顺序统计量中的第 m 个元素. 将第 m 个顺序统计量出现 y_m 的概率密度函数表示为 $f_m(y_m)$. y_m 仅作为顺序统计量中的某一个元素发生 y_m 的概率为 $f(y_m)$; 假设在 y_m 之前, n 个样本量中有 $m-1$ 个元素值小于 y_m, 其发生概率为 $\frac{n!}{(m-1)!(n-m)!}F(y_m)^{m-1}$; 余下大于 y_m 的 $n-m$ 个元素的发生概率为 $(1-F(y_m))^{n-m}$. 因为假设了 n 个样本量之间相互独立, 故在 n 个顺序统计量中第 m 个元素值为 y_m 的发生概率为

$$\begin{aligned} f_m(y_m) &= \frac{n!}{(m-1)!(n-m)!} f(y_m) \\ &\quad \times F(y_m)^{m-1}(1-F(y_m))^{n-m} \\ &= \frac{n!}{(m-1)!(n-m)!} \\ &\quad \times (y_m)^{m-1}(1-y_m)^{n-m}. \end{aligned} \tag{11}$$

同理, 当已获知样本 $\{x\}$ 的具体分布模型, 在 n 个顺序统计量中出现 x_m 的概率可表示为

$$\begin{aligned} g_m(x_m) &= \frac{n!}{(m-1)!(n-m)!} f(x_m) \\ &\quad \times F(x_m)^{m-1}(1-F(x_m))^{n-m}. \end{aligned} \tag{12}$$

(12)式将依赖于 x 的分布情况.

3.2. 基于偏态分布的第 m 个顺序统计量的分布特征

采用以上介绍的方法, 计算得到 1980 年—2009 年夏季逐日平均温度序列的概率密度函数 $f(x)$ ((2)式)及其分布函数 $F(x)$ ((7)式), 代入(12)式后可得到基于偏态分布的第 m 个顺序统计量的概率密度分布 $g_m(x)$. 图 3 所示为基于高斯分布、正偏分布以及负偏分布下的第 m 个顺序统计量的概率密度分布, 横坐标为顺序统计量分布的累积概率 $F(x)$, 图中以样本量 n 取为 30, m 分别取 1, 2, 4, 10, 15, 20, 27, 29, 30.

由图 3(a)可见, 当序列具有高斯分布时, 第 15 个顺序统计量分布的最高概率密度高于其他顺序

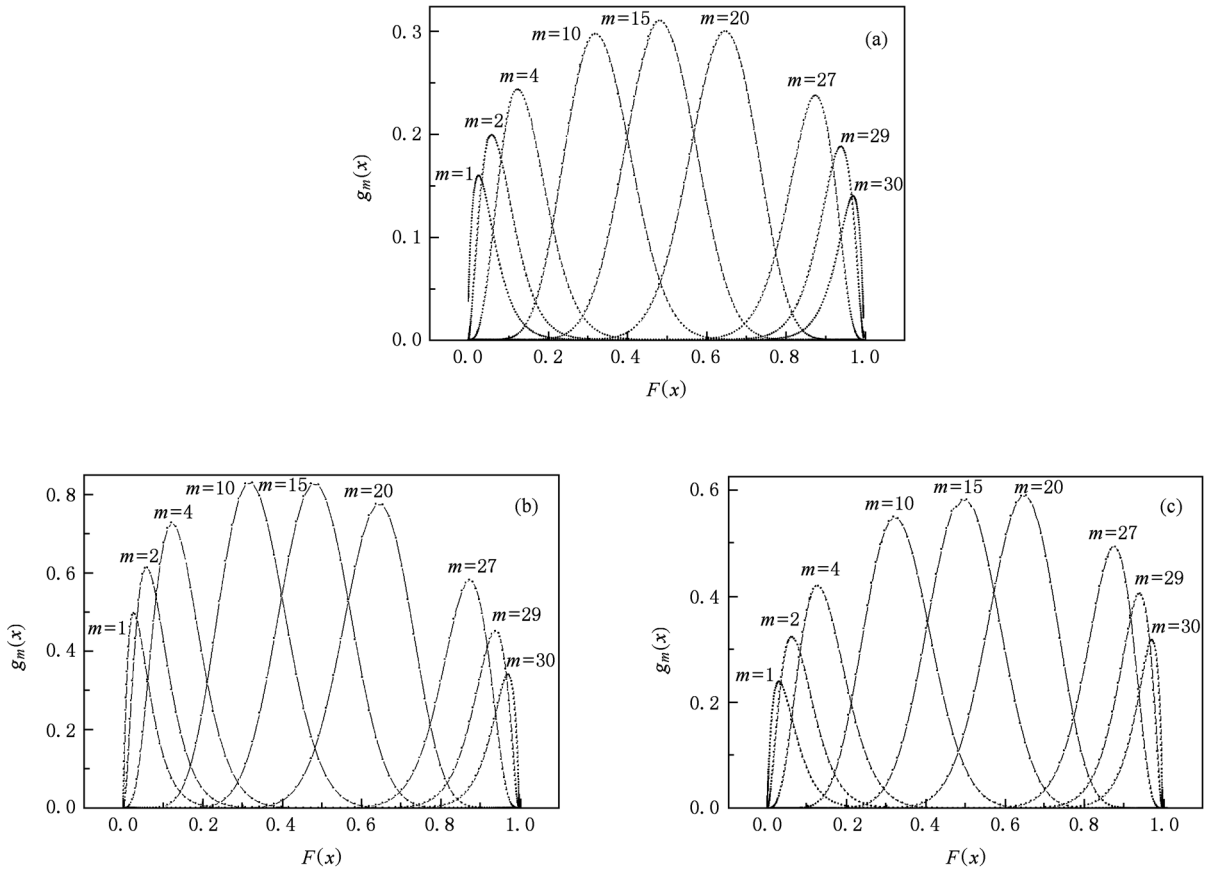

图 3 第 m 个顺序统计量的概率密度分布特征 (a)高斯分布, (b)正偏分布, (c)负偏分布

统计量,其他顺序统计量的分布关于第 15 个顺序统计量分布近似对称,即 $m = 10$ 与 $m = 20$, $m = 4$ 与 $m = 27$ 等等. 且 m 的值越大或越小,越可能出现概率递减. 图 3(b)为序列处于正偏分布时第 m 个顺序统计量的分布特征,此时第 15 个顺序统计量分布中的最高概率密度不再高于其他顺序统计量,而是出现在第 10 个顺序统计量中,其他顺序统计量也不再关于第 15 个顺序统计量对称,在左右对应的顺序统计量分布中左侧分布的最概然概率高于右侧. 而在图 3(c)中,当序列处于负偏分布时,其情况恰与图 3(b)相反,也就是左右对应的顺序统计量分布中右侧分布的最概然概率高于左侧.

4. 经验百分位公式的建立

从图 3 可知,当序列处于不同的分布型时,对该序列第 m 个顺序统计量的分布特征具有一定的影响. 而在序列所满足的偏态分布中,偏态指数是表征序列偏态程度的量,据此可以判定,序列中概率分布的偏态指数与第 m 个顺序统计量的分布形式相关联. 第 m 个顺序统计量的分布即代表着某序列第 m 个元素可能出现的值的分布,对第 m 个元素值进行估计,即可进一步对第 m 个元素所对应的累积概率进行估计,从而将排序后元素所对应的位置与其相应累积概率相联系. 而由于第 m 个顺序统计量在越偏离中间位置的顺序统计量时,其分布的偏态程度将越明显,若以均值作为估计值,受极端值的影响也将越显著,以致越偏离于最有可能发生的值. 为此,将第 m 个顺序统计量分布中最高概率密度所对应的值作为对第 m 个元素的估计值.

下面给出建立基于偏态分布的经验百分位公式的步骤.

步骤 I 选定样本量 n.

步骤 II 选定表征不同分布型的偏态指数 λ.

步骤 III 在某一分布型下,对于 n 个顺序统计量中第 m 个值为 x_m 的概率密度函数可由(13)式获得,以 x_m 分布函数中最高概率密度所对应的 x_m 值作为最有可能发生的第 m 个顺序统计量,记为 $x_{m(\max g_m)}$.

步骤 IV 由历史数据的分布函数求解 $F(x_{m(\max g_m)})$,作为对经验百分位 P_m 的估计.

步骤 V 将 $P_m = \dfrac{m + a}{n - b}$ 写成 $nP_m - m = a + bP_m$,令 $S_m = nP_m - m$,以 P_m(即 $F(x_{m(\max g_m)})$)为变量,S_m 为因变量(其中 m 在 1—n 之间取值)建立线性回归方程,得到位置参数 a 和 b.

步骤 VI 在已知样本量的情况下,选定不同的分布,即改变偏态指数,重复步骤 II—步骤 V,得到不同的位置参数 a 和 b.

步骤 VII 建立位置参数 a 和 b 关于偏态指数 λ 的回归方程.

步骤 VIII 选取不同的样本量,重复步骤 II—步骤 VII.

按照以上步骤,即可求得偏态分布的百分位公式.

图 4 为样本量 $n = 100$ 时,不同偏态指数 λ 下 S_m 随 P_m 的变化. 由图 4 可知,在 λ 取不同值时,S_m 随 P_m 变化的曲线满足不同的线性关系,可由线性拟合得到位置参数 a 和 b,且当序列的概率分布从正偏分布向负偏分布变化(即偏态指数 λ 值逐渐变大)时,位置参数 a 有增大的趋势,位置参数 b 则有减小的趋势.

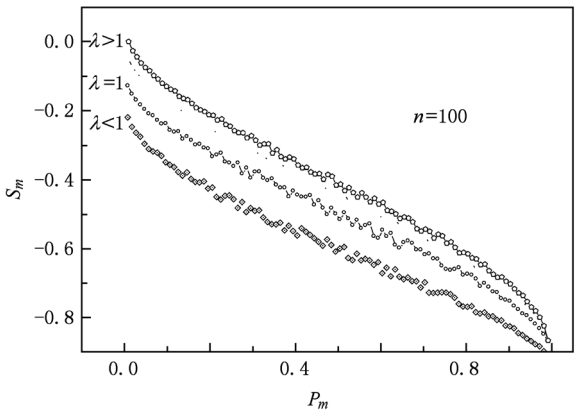

图 4 S_m 随 P_m 的变化

从图 4 中已经获知,位置参数 a 和 b 随偏态指数 λ 的变化具有不同的特征,那么它们之间究竟存在的是何种关系呢? 为此,同样以样本量 $n = 100$ 为例,分别建立位置参数 a 和 b 随 λ 变化的回归方程. 图 5 所示为位置参数 a 和 b 随偏态指数 λ 的变化,基于曲线拟合,位置参数 a 和 b 均可建立关于 λ 的指数回归方程,即

$$\begin{aligned} a &= a_0 + a_1 \exp(a_2 \lambda), \\ b &= b_0 + b_1 \exp(b_2 \lambda). \end{aligned} \tag{13}$$

对位置参数 a 和 b 拟合较好的 λ 范围主要集中在

0.5—3 之间,对于负偏以及正偏相对比较严重的情 况,拟合效果相对较差.

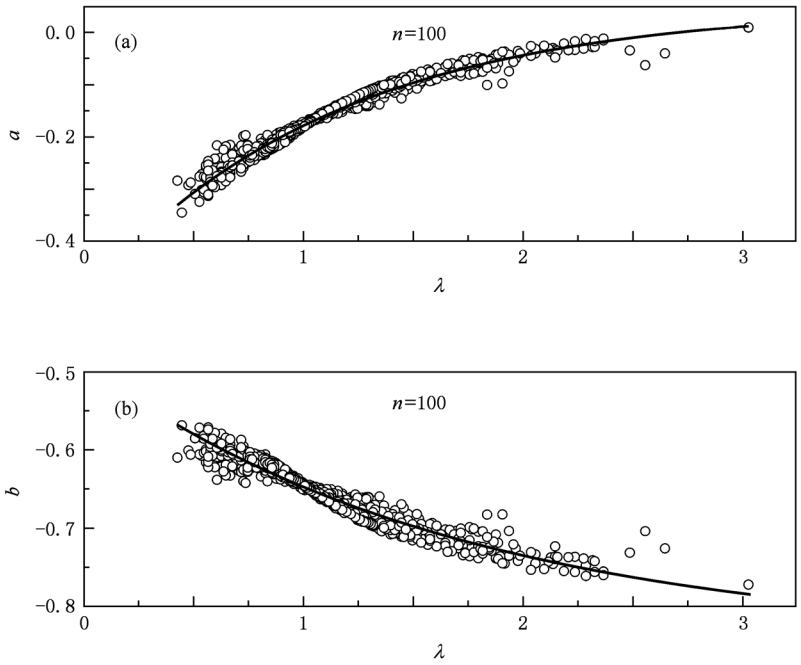

图 5 位置参数 a 和 b 随偏态指数 λ 的变化 (a)参数 a, (b)参数 b

表 1 不同样本量所得的各项系数

n	a_0	a_1	a_2	b_0	b_1	b_2
10	0.05856	−0.62674	−0.93664	−0.81255	0.39290	−0.77126
30	0.05669	−0.59363	−0.91352	−0.83076	0.37272	−0.68126
50	0.05481	−0.57900	−0.90317	−0.83461	0.36354	−0.65186
70	0.05183	−0.56880	−0.90050	−0.83887	0.35854	−0.62066
100	0.05057	−0.55665	−0.88833	−0.84378	0.35441	−0.59072
150	0.04979	−0.54735	−0.87807	−0.84958	0.35183	−0.56023

表 1 为选取不同的样本量时所得到的位置参数 a 和 b 关于偏态指数的回归方程的各项系数. 从表 1 可知,在不同的样本量下所得的系数虽有所不同,但随着样本量不断增加,它们的差异越来越小. 图 6 显示了各项系数随样本量 n 的变化关系. 经计算,在样本量很大的情况下,各项系数值可最终收敛于一个常数,分别为 $a_0 \approx 0.04793$, $a_1 \approx -0.54480$, $a_2 \approx -0.87394$, $b_0 \approx -0.84915$, $b_1 \approx 0.35186$, $b_2 \approx -0.55255$. 而在样本量 $n \geqslant 300$ 时,各项系数的误差均可控制在 1×10^{-3} 内. 保留上述各项系数小数点后两位小数,将其代入 $P_m = \dfrac{m + a}{n - b}$ 中,可得偏态分布下的百分位公式为

$$P_m = \frac{m - 0.54\exp(-0.87\lambda) + 0.05}{n - 0.35\exp(-0.55\lambda) + 0.85}. \quad (14)$$

至此,已建立与偏态分布相关的经验百分位公式.

值得一提的是,Jenkinson[1] 给出的不考虑具体分布形式的经验百分位公式,是以第 m 个顺序统计量 x_m 概率分布的中值作为第 m 个顺序统计量的估计值时得到的理论公式;而同样在不考虑具体分布形式下,若以第 m 个顺序统计量 x_m 概率分布的均值作为第 m 个顺序统计量的估计值时,所得的经验百分位公式为 $p = \dfrac{m}{n + 1}$. 在考虑偏态分布条件下,以第 m 个顺序统计量 x_m 概率分布中最概然概率对应的值作为第 m 个顺序统计量的估计值,以此得到基于偏态分布的经验百分位公式. 当偏态指数 $\lambda = 1$ (代表左右对称的高斯分布),代入(14)式,可得 $p = \dfrac{m - 0.175}{n + 0.65}$, 其位置参数 a 和 b 与以中值和平均

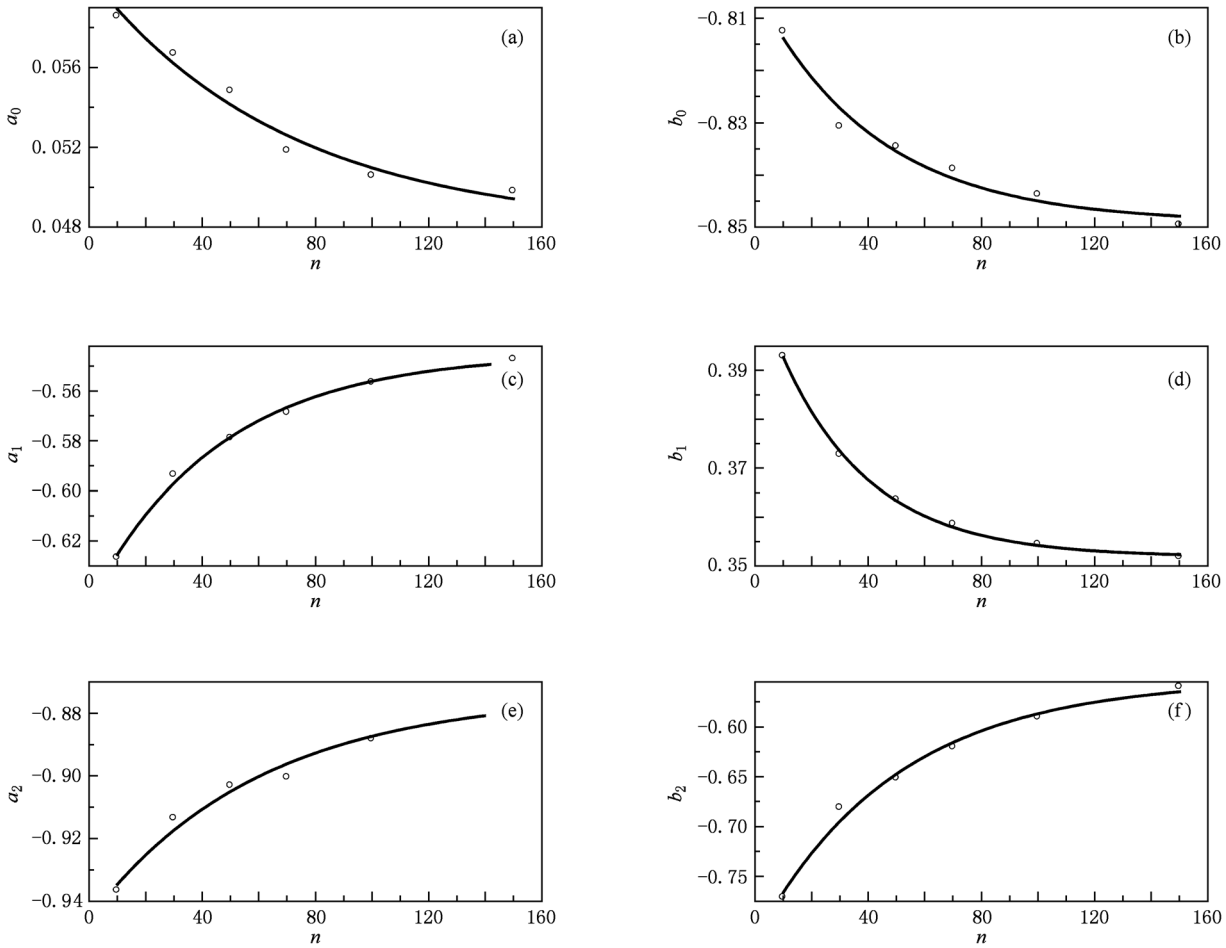

图 6 系数 $a_0, b_0, a_1, b_1, a_2, b_2$ 随样本量 n 的变化 (a)a_0, (b)b_0, (c)a_1, (d)b_1, (e)a_2, (f)b_2

值作为估算值时均有所不同. 由此可知,即使都是在对称分布下,若处理方法不同其位置参数也会有所不同.

5. 偏态与 Jenkinson 方法相同百分位下温度顺序统计量位置的比较

Jenkinson 百分位估计式((1)式)虽是在不考虑具体分布的情况下获得,但在一定的条件下有其适用性,是一种比较常用的百分位值估计方法. 在统计模型存在一定的偏态特征时,所得偏态分布下的经验百分位公式((14)式)可针对不同概率分布条件下的百分位值进行估计. 在一定的样本数据下对于某一特定的百分位,Jenkinson 百分位估计方法中对应的顺序统计位置也是确定的,但偏态百分位估计方法中随偏态程度的不同对应的顺序统计量位置也会有所不同. 在大气科学中常根据研究的需要,将温度资料的第 90 个、第 95 个、第 99 个等一些百分位值作为极端高温的阈值[5,6,20],在此方法下百分位值所对应的温度在排序后序列中所处的位置将决定阈值的大小,文中取温度序列的第 90 个百分位值进行研究. 在温度资料满足偏态分布的条件下,第 90 个百分位值所对应的温度在排序后序列中的位置将由于偏态程度的不同而有所差异.

利用(1)和(14)式,对 1980 年—2009 年全球各格点夏季逐日平均温度资料进行百分位估计,比较这两种方法得到第 90 个百分位值对应的温度在排序后序列中位置的差异,将此位置差异用 Δm 表示,其分布如图 7 所示. 因 1980 年—2009 年夏季日平均温度资料存在一定的偏态特征,故其在偏态分布条件与 Jenkinson 估计下的第 90 个百分位值所对应的排序后序列中位置差值的分布与图 2 中偏态指数的分布具有紧密的联系. 当偏态指数越小时,其对应的位置差值越大,在处于正偏的赤道附近部分地区,即亚洲大陆北部、非洲北部及其东部海域以及南美附近海域,位置差值在 0.1 乃至 0.15 以上. 在处于负偏的地

区,位置差值在 0.1 以内,在 60°S 附近以及北极地区,即负偏比较明显的地区,位置差值在 0.05 以内.

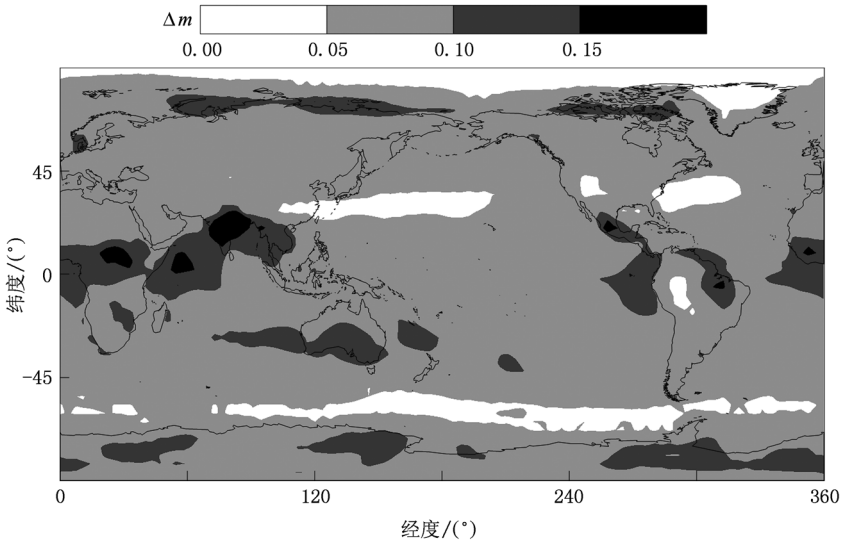

图 7 偏态与 Jenkinson 方法下第 90 个百分位值所对应的位置差异 Δm 的分布

6. 结 论

本文以偏态分布函数作为温度时间序列的统计模型,针对原始数据序列偏态分布条件下的累积概率函数,通过理论推导和数值模拟建立了偏态分布下的经验百分位公式. 对基于偏态分布的顺序统计量给出简单的经验百分位值估计. 通过相关推导和计算表明,偏态分布下的经验百分位公式与表征偏态程度的偏态指数相关,利用本文得到的百分位公式可以更精确地估计不同分布下的百分位值. 对全球各格点夏季逐日平均温度资料的应用比较表明,当正偏分布越明显,即偏态指数越小时,与 Jenkinson 公式对应的温度顺序统计量位置相差越大.

[1] Jenkinson A F 1977 *Synoptic Climatol. Branch Memo* **58** 41

[2] Horton E B, Folland C K, Parker D E 2001 *Climatic Change* **50** 267

[3] Bonsal B R, Zhang X, Vincent L A, Hogg W D 2001 *J. Climate* **14** 1959

[4] Feng G L, Gong Z Q, Zhi R 2008 *Acta Meteor. Sin.* **66** 892 (in Chinese) [封国林、龚志强、支 蓉 2008 气象学报 **66** 892]

[5] Gong Z Q, Wang X J, Zhi R, Feng G L 2009 *Acta Phys. Sin.* **58** 4342 (in Chinese) [龚志强、王晓娟、支 蓉、封国林 2009 物理学报 **58** 4342]

[6] Feng G L, Wang Q G, Hou W, Gong Z Q, Zhi R 2009 *Acta Phys. Sin.* **58** 2853 (in Chinese) [封国林、王启光、侯 威、龚志强、支 蓉 2009 物理学报 **58** 2853]

[7] Folland C K, Anderson C W 2002 *J. Climate* **15** 2954

[8] Makkonen L 2005 *J. Appl. Meteor. Climatol.* **45** 334

[9] Goel N K, De M 1993 *Stoch. Hydrol. Hydraul.* **7** 1

[10] Hou W, Yang P, Feng G L 2008 *Acta Phys. Sin.* **57** 3932 (in Chinese) [侯 威、杨 萍、封国林 2008 物理学报 **57** 3932]

[11] Zhang D Q, Qian Z H 2008 *Acta Phys. Sin.* **57** 4634 (in Chinese) [章大全、钱忠华 2008 物理学报 **57** 4634]

[12] Feng G L, Dong W J, Gong Z Q, Hou W, Wan S Q, Zhi R 2006 *Nonlinear Theory and Methods on Spatial-temporal Distribution of the Observational Data* (Beijing: China Metrological Press) (in Chinese) [封国林、董文杰、龚志强、侯 威、万仕全、支 蓉 2006 观测数据非线性时空分布理论和方法 (北京:气象出版社)]

[13] He W P, Feng G L, Dong W J, Li J P 2005 *Chin. Phys.* B **14** 21

[14] He W P, Feng G L, Wu Q, Wan S Q, Chou J F 2008 *Nonlin. Proc. Geophys.* **15** 601

[15] Feng G L, Yang J, Wan S Q, Hou W, Zhi R 2009 *Acta Meteor. Sin.* **67** 61 (in Chinese) [封国林、杨 杰、万仕全、侯 威、支 蓉 2009 气象学报 **67** 61]

[16] Feng G L, Gao X Q, Dong W J, Li J P 2008 *Chaos Solitons Fract.* **37** 487

[17] Feng G L, Gong Z Q, Zhi R, Zhang D Q 2008 *Chin. Phys.* B **17** 2745

[18] Qian Z H, Feng G L, Gong Z Q 2010 *Acta Phys. Sin.* **59** 7498 (in Chinese) [钱忠华、封国林、龚志强 2010 物理学报 **59** 7498]

[19] Chen X R 2009 *Advanced Mathematical Statistics* (Hefei: University of Science and Technology of China Press) p164 (in Chinese) [陈希孺 2009 高等数理统计学(合肥:中国科学技术大学出版社)第 164 页]

[20] Zhang L, Zhang D Q, Feng G L 2010 *Acta Phys. Sin.* **59** 5897 (in Chinese) [张 璐、章大全、封国林 2010 物理学报 **59** 5897]

Development of percentile estimation formula for skewed distribution*

Zhou Yun[1)] Hou Wei[2)†] Qian Zhong-Hua[1)] He Wen-Ping[2)]

1) (*College of Physics Science and Technology, Yangzhou University, Yangzhou* 225009, *China*)

2) (*National Climate Center, Beijing* 100081, *China*)

(Received 25 October 2010; revised manuscript received 28 February 2011)

Abstract

Order statistics establishes a relation between the position of the ranked data and corresponding cumulative probability, so it can be used to estimate the cumulative probability. Owing to the fact that different climatological data have different skewness degrees, in this paper, according to the cumulative probability function under the skewed distribution conditions, we perform theoretical analysis and numerical simulation to establish the position parameters of the regression model which are related to skewness index, then give an amperic percentile formula under the skewed distribution. By using the data about the summer temperature in global from 1980 to 2009, we compare the positions of ranked data corresponding to the 90th percentile, which are obtained by this formula and Jenkinson's formula.

Keywords: order statistics, skewed distribution, percentile

PACS: 92. 60. Wc

* Project supported by the National Natural Science Foundation of China (Grant Nos. 40930952, 41005043) and the State Key Program of Science and Technology of China (Grant No. 2007BAC29B01).

† Corresponding author. E-mail: hou_w@sohu.com