

论我国季、月降水量的正态性和正态化

施能 陈辉

(南京气象学院)

提 要

本文讨论了我国35个测站的月、季降水量以及按区域平均降水量的正态性问题。同时,研究比较了各种降水量正态化的方法和效果。

一、前 言

日降水量的统计分布明显偏离正态。这是因为降水量不可能取小于零的值。有明显正偏差的降水的概率虽然不大,但数值明显超过负偏差,所以日降水量的众数小于众位数,是正偏态分布。对与日降水密切有关的月、季、年降水量来说,情况比较复杂,很难对一般情况作出肯定的回答。

由中心极限定理〔1〕可知,在大量独立因素作用下的某变量,如果每一个因素对变量的影响都很小时,该变量接近正态分布。但是,降水量通常受到少数重要因素,如环流、地形、地方性特点等的影响,可能不服从正态分布。由于月降水量是有限个被加项(30或31)按时间顺序求和,所以中心极限定理的条件未被满足。文献〔2〕估计了独立的被加项数

$$N_1 = N^2/2 \sum_{\tau=0}^{N-1} (N-\tau)R(\tau) \quad (1)$$

式中 N 为独立观测数(30或31), $R(\tau)$ 是标准化的相关函数。如果相关函数近似用 r^τ 表示, r 是相邻观测之间的相关系数,则当 r 为0.3—0.5时, $N_1=6-8$ 。

研究单点月、季降水量及区域平均降水量的正态性是个较为重要的问题。这涉及到能否使用线性模型的问题。此外, t 分布、 F 分布、 χ^2 分布均是正态随机变量作为自变量的函数的概率分布。所以,当讨论的要素不服从正态分布时,对该变量的各种统计检验(t , χ^2 , F 检验)理论上都不能进行,这也影响到统计方法的效果。

对年降水量的正态性早已有过研究〔3〕,但是对全国范围的单点或区域平均的季、月降水量的正态性检验工作尚未见到。本文打算回答这个问题,并且给出了若干变量正态

化方法的效果。

二、资料和方法

对变量是否服从正态分布的检验方法有皮尔逊(Pearson)判据(即 χ^2 检验法),柯尔莫科洛夫(Колмогоров)判据,以及偏度、峰度检验法。我们使用偏度、峰度检验法。

$$\text{偏度: } r_1 = \frac{1}{n} \sum_i (x_i - \bar{x})^3 / \sigma^3 \quad (2)$$

$$\text{峰度: } r_2 = \frac{1}{n} \sum_i (x_i - \bar{x})^4 / (\sigma^4 - 3)$$

$$\sigma = \left[\frac{1}{n} \sum_i (x_i - \bar{x})^2 \right]^{1/2}$$

对于标准正态分布 $r_1=0$, $r_2=0$ 。偏度、峰度检验法是计算统计量

$$G_1 = \frac{[n(n-1)]^{1/2} r_1}{n-2} \quad (3)$$

$$G_2 = \frac{n-1}{(n-2)(n-3)} [(n+1)r_2 + 6]$$

统计学证明〔4〕, $n \rightarrow \infty$ 时,统计量 G_1 、 G_2 服从正态分布,其参数为

$$M(G_1) = M(G_2) = 0$$

$$D(G_1) = \frac{6n(n-1)}{(n-2)(n+1)(n+3)} \quad (4)$$

$$D(G_2) = \frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}$$

所以当计算的 G_1 、 G_2 满足

$$|G_1| > 2.58 [6n(n-1)/(n-2)(n+1)(n+3)]^{1/2} \quad (5)$$

$$|G_2| > 2.58 [24n(n-1)^2/(n-3)(n-2)(n+3)(n+5)]^{1/2} \quad (5)'$$

时,认为变量不服从正态分布,信度0.01;反之,认为服从正态分布。

我们取全国范围35个测站的月降水量, 资料年代1951—1984年, $n=34$ 。由(4)式算得, $D(G_1)=0.1625, D(G_2)=0.6208$ 。所以 $|G_1|$ 、 $|G_2|$ 的临界值分别是1.04, 2.03。

三、月降水量的正态分布

我们在计算时发现, 大约有10%的不能同时满足(5)式和(5)'式, 为此我们作下列规定: (1) $|G_1|>1.04$, 且 $|G_2|>2.03$ 时, 称为不服从正态, 用双圆圈表示。

(2) 当 $|G_1|>2$, $|G_2|>4$ 时, 称为明显不服从正态, 用实心圆表示。(3) $|G_1|<1.04$, 且 $|G_2|<2.03$ 时称服从正态, 用空白表示。(4) $|G_1|<1.3$ 或 $|G_2|<2.6$, 且 $|G_1|+\frac{1}{2}|G_2|<2.07$ 称为基本服从正态分布, 用空心圆表示。

表1给出了35个站逐月、季降水量正态性情况。从表1可见, (1)从单站看, 正态性最好的站是上海站; 其次是长沙、济南、合肥、南昌、贵阳等; 正态性最差的站是吐鲁番, 其次是天津、石家庄、北京、呼和浩特、广州、太原等。(2)从各月看, 单站降水量正态性最好的是9月, 其次是8、3、6、7月, 冬季各月(12、1、2月)正态性

表1 我国35站月、季、年降水量正态性表

站	1	2	3	4	5	6	7	8	9	10	11	12	季	年
北京	●	●	●	●	●	●	●	●	●	●	●	●	●	●
天津	●	●	●	●	●	●	●	●	●	●	●	●	●	●
承德	●	●	●	●	●	●	●	●	●	●	●	●	●	●
石家庄	●	●	●	●	●	●	●	●	●	●	●	●	●	●
太原	●	●	●	●	●	●	●	●	●	●	●	●	●	●
呼和浩特	●	●	●	●	●	●	●	●	●	●	●	●	●	●
沈阳	●	●	●	●	●	●	●	●	●	●	●	●	●	●
大连	●	●	●	●	●	●	●	●	●	●	●	●	●	●
长春	●	●	●	●	●	●	●	●	●	●	●	●	●	●
齐齐哈尔	●	●	●	●	●	●	●	●	●	●	●	●	●	●
哈尔滨	●	●	●	●	●	●	●	●	●	●	●	●	●	●
鸡西	●	●	●	●	●	●	●	●	●	●	●	●	●	●
上海	○	○	○	○	○	○	○	○	○	○	○	○	○	○
徐州	○	○	○	○	○	○	○	○	○	○	○	○	○	○
南京	○	○	○	○	○	○	○	○	○	○	○	○	○	○
杭州	○	○	○	○	○	○	○	○	○	○	○	○	○	○
合肥	○	○	○	○	○	○	○	○	○	○	○	○	○	○
安庆	○	○	○	○	○	○	○	○	○	○	○	○	○	○
福州	○	○	○	○	○	○	○	○	○	○	○	○	○	○
南昌	○	○	○	○	○	○	○	○	○	○	○	○	○	○
赣州	○	○	○	○	○	○	○	○	○	○	○	○	○	○
济南	○	○	○	○	○	○	○	○	○	○	○	○	○	○
郑州	○	○	○	○	○	○	○	○	○	○	○	○	○	○
武汉	○	○	○	○	○	○	○	○	○	○	○	○	○	○
常德	○	○	○	○	○	○	○	○	○	○	○	○	○	○
长沙	○	○	○	○	○	○	○	○	○	○	○	○	○	○
广州	○	○	○	○	○	○	○	○	○	○	○	○	○	○
百色	○	○	○	○	○	○	○	○	○	○	○	○	○	○
南宁	○	○	○	○	○	○	○	○	○	○	○	○	○	○
成都	○	○	○	○	○	○	○	○	○	○	○	○	○	○
贵阳	○	○	○	○	○	○	○	○	○	○	○	○	○	○
昆明	○	○	○	○	○	○	○	○	○	○	○	○	○	○
西安	○	○	○	○	○	○	○	○	○	○	○	○	○	○
兰州	○	○	○	○	○	○	○	○	○	○	○	○	○	○
吐鲁番	●	●	●	●	●	●	●	●	●	●	●	●	●	●
正态站数	13	18	24	22	23	23	23	26	28	20	23	12	255	26

最差。图1给出了1、4、7、10月全国35站降水量 $|G_1|$ 值。

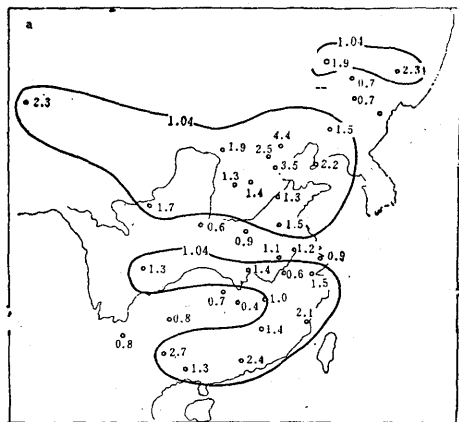


图1a 1月全国35站降水量 $|G_1|$ 值

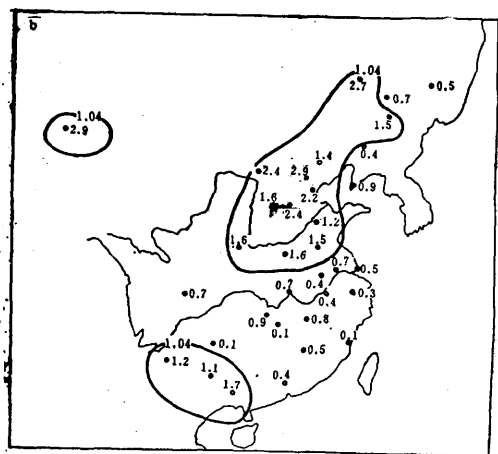


图1b 4月全国35站降水量 $|G_1|$ 值

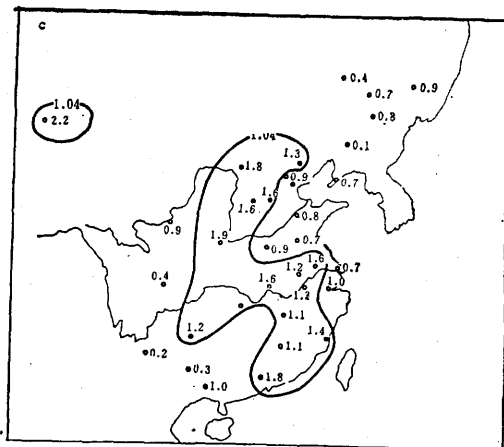
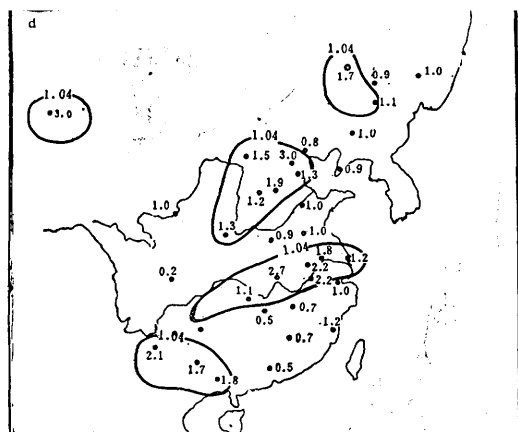


图1c 7月全国35站降水量 $|G_1|$ 值

图 1d 10月全国35站降水量|G₁|值

四、年、季降水量的正态性

由表 1 还可以看出, 一般来说, 求时间平均是有利于正态化的, 但是单站的季降水量并不总能满足正态分布, 特别是春季。另外, 在极个别情况下, 从不符合正态的月降水量求得的季降水量会更不符合正态, 例如, 齐齐哈尔和哈尔滨春季。

五、区域平均的月降水量的正态性

降水量的区域平均就是降水量的空间平均,这也有利降水量正态化。我们取如下单站平均作为区域平均降水。华北区:北京、天津、承德、石家庄、太原、呼和浩特、济南;华南区:广州、百色、南宁、赣州;西南区:成都、贵阳、昆明;东北区:沈阳、长春、齐齐哈尔、哈尔滨、鸡西;西北区:兰州、西安;长江中游区:安庆、南昌、武汉、常德、长沙;长江下游区:上海、南京、杭州、合肥。后两区组成长江流域区。

图 2 给出了区域平均降水量的 $|G_1|$ 、 $|G_2|$ 值, 根据 $|G_1|$ 、 $|G_2|$ 的值, 用判断单站降水量正态性标准判断的区域降水量正态性情况列于表 2。由表 2 可看出, 经过空间平均后, $|G_1|$ 、 $|G_2|$ 值有所降低, 但仍有些月份不服从正态分布, 特征与单点月降水量的正态性类似。

六、正态化变换的方法与效果

1. 利用变量本身信息的变换

这是取变量的某个函数作变换^[4],使变换后的变量趋向正态分布。常用的有立方根变换,即 $x' = x^{\frac{1}{3}}$; 4次方根变换,即 $x' = x^{\frac{1}{4}}$; 双曲正切变换,即 $x' = \text{th}[(x - \bar{x})/\bar{x}]$ 。为简单起见,我们只列出对表1中

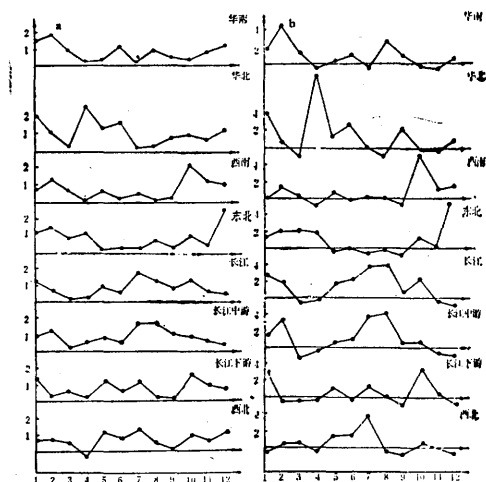


图2 各大区逐月平均月降水量的 $|G_1|$ (a)
和 $|G_2|$ (b)

表2 区域平均降水量的正态性表

区	月	1	2	3	4	5	6	7	8	9	10	11	12
华南区		●	●										○
华北区		●	○		●	○	●						
西南区			○								●		
东北区			●										●
长江流域		●						●	●		○		
长江中游			●					●	●				
长江下游		●									○		
西北区								○					

不服从正态分布的站月进行这三种变换后的平均绝对值 $|\bar{G}_1|$ 、 $|\bar{G}_2|$ (见表3)。从表3可看出:

(1) 平均来看,正态性最好的是9月,其次是夏季月份,最差的是冬季。这与表1的结果一致。

(2) 立方根变换明显地优于 4 次方根变换和双曲正切变换。根据对表 1 中 255 次变换统计,立方根变换的正态化效率达 95%,而 4 次方根变换和双曲正切变换则分别为 70%、71%。

(3) 资料分析表明, 4 次方根变换、双曲正切变换比较适合于明显不正态的站月。吐鲁番 4 月、12 月, 哈尔滨 12 月用立方根变换仍不能正态化, 而用 4 次方根、双曲正切变换则达到正态化。

2. 利用自然正交展开进行正态化

将 m 个测站的 n 次观测资料阵作自然正交展开

$${}_mX_n = {}_mV_{mm}Z_n \quad m = 35, \quad n = 34 \quad (6)$$

展开时, 先将X阵作标准化处理。这样

表3 变换前后的 $|\bar{G}_1|$ 、 $|\bar{G}_2|$

	月 份	1	2	3	4	5	6	7	8	9	10	11	12
$ \bar{G}_1 $	x	1.55	1.46	1.13	1.11	1.04	0.99	1.04	1.06	0.88	1.34	1.25	1.77
	$x^{\frac{1}{3}}$	0.46	0.38	0.46	0.37	0.45	0.32	0.26	0.33	0.24	0.46	0.29	0.33
	$x^{-\frac{1}{4}}$	1.46	1.50	1.60	0.50	0.79	0.50	0.48	0.70	0.47	1.80	1.49	1.60
	$\text{th} \frac{x-\bar{x}}{x}$	1.21	1.30	1.39	0.51	0.78	0.49	0.47	0.64	0.50	1.64	1.30	1.41
$ \bar{G}_2 $	x	3.42	3.21	2.36	2.32	2.21	1.41	1.81	2.16	1.24	2.55	1.99	4.41
	$x^{\frac{1}{3}}$	1.30	0.73	0.60	0.58	0.79	0.39	0.41	0.64	0.51	0.91	0.67	0.74
	$x^{-\frac{1}{4}}$	5.03	3.81	4.90	0.99	2.50	0.78	0.60	1.53	2.02	7.00	4.57	5.24
	$\text{th} \frac{x-\bar{x}}{x}$	4.30	3.21	4.30	0.96	2.24	0.74	0.57	1.48	1.75	5.76	3.47	4.01

mV_m 的列就是相关阵 (m 阶) 的特征向量, mZ_n 是时间系数阵, 这里也可称为 m 个主成份组成的矩阵。将相关阵的特征值 λ_i 从大到小排列, 取 k 个特征值, 则得到 X 阵的拟合场。

$$m\hat{X}_n = mV_{kk}Z_n \quad (7)$$

(1) 时间系数阵 mZ_n 的正态性

我们对12个月的35个时间系数分别用上述方法计算 G_1 、 G_2 , 作正态性检验。为简单起见, 给出前14个系数在12个月的 $|\bar{G}_1|$ 、 $|\bar{G}_2|$ 值, 统计 $|\bar{G}_1|$ 、 $|\bar{G}_2|$ 判断不满足正态

的月数 (见表4)。由表4可见, 绝大多数时间系数符合正态分布, 这就为我们用线性回归、线性判别将时间系数作为预报因子或预报对象找到了理论根据。事实上, 我们对全部 (12×35) 系数进行统计, 不服从正态分布的频率仅 $23/(35 \times 12) = 0.055$, 而原始降水资料不服从正态分布的频率 (据表1) 为 $(420-255)/35 \times 12 = 0.393$ 。但是, 遗憾的是不服从正态的时间系数近一半发生在序号偏小 (特征值偏大) 的时间系数中。所以, 对某些个别月份展开时, 不能保

表4 前14个时间系数的 $|\bar{G}_1|$ 、 $|\bar{G}_2|$, 以及12个月中不满足正态的月数

展 开 系 数	1	2	3	4	5	6	7	8	9	10	11	12	13	14
$ \bar{G}_1 $	0.89	0.48	0.57	0.66	0.35	0.40	0.66	0.41	0.30	0.37	0.36	0.41	0.34	0.39
$ \bar{G}_2 $	1.46	0.69	0.85	1.35	0.63	1.16	2.06	0.94	0.55	0.74	0.56	0.44	0.41	0.66
不正态月数	2	0	1	1	0	2	4	0	0	0	0	0	0	0

证所用的组合因子即时间系数服从正态分布。

表5给出了逐月14个系数的 $|\bar{G}_1|$ 、 $|\bar{G}_2|$ 以及35个时间系数中满足正态分布的个数。由表5可见, 时间系数的正态性随月份分布是不均匀的, 12月份的展开系数不服从正态分布的频率最大 (9/35), 9月、3月最小

(0/35)。此特征与表1的结果大致相符。所以, 可以推测, 自然正交展开后的时间系数比原始变量符合正态。但是所展开的月份资料明显不服从正态分布时, 展开后的时间系数也易不符合正态分布。

(2) 拟合场 $\hat{X}$ 的正态性

利用 (7) 式取 k 个特征向量 ($k \ll m$)

表5 35个时间系数的逐月 $|\bar{G}_1|$ 、 $|\bar{G}_2|$ 以及不正态的个数

月 份	1	2	3	4	5	6	7	8	9	10	11	12	合 计
$ \bar{G}_1 $	0.35	0.51	0.40	0.44	0.58	0.49	0.38	0.40	0.45	0.54	0.47	0.64	—
$ \bar{G}_2 $	0.92	1.07	0.68	0.60	1.37	0.79	0.99	0.52	0.61	0.81	0.72	1.63	—
不正态数	1	4	0	1	2	2	1	0	0	1	2	9	23

拟合X时,其精度可用 $\sum_{i=1}^k \lambda_i / \sum_{i=1}^{35} \lambda_i$ 表示。

这时的X场实质是对X场的一种平滑和滤

波。这样恢复的35个站的降水量能趋向正态分布。表6给出了逐月的 $|\bar{G}_1|$ 和 $|\bar{G}_2|$ 。表6最后一行给出了用10个特征向量模拟X场的逐月符合正态分布的站数。

表6 取K个特征向量拟合时的累积贡献率以及 $|\bar{G}_1|$ 、 $|\bar{G}_2|$

k		1	2	3	4	5	6	7	8	9	10	11	12
8	$ \bar{G}_1 $	0.71	0.73	0.48	0.61	0.66	0.78	0.48	0.41	0.41	0.48	0.45	0.85
	$ \bar{G}_2 $	1.01	1.47	0.82	1.10	1.44	1.51	0.81	0.77	0.87	0.74	0.62	1.85
	$\sum_{i=1}^8 \lambda_i / \sum_{i=1}^{35} \lambda_i$	0.813	0.814	0.748	0.757	0.722	0.717	0.691	0.699	0.702	0.780	0.797	0.801
10	$ \bar{G}_1 $	0.67	0.68	0.43	0.52	0.69	0.63	0.43	0.41	0.41	0.46	0.49	0.84
	$ \bar{G}_2 $	0.95	1.37	0.74	0.99	1.48	1.35	0.48	0.79	0.68	0.87	0.62	1.92
	$\sum_{i=1}^{10} \lambda_i / \sum_{i=1}^{35} \lambda_i$	0.863	0.868	0.822	0.826	0.795	0.793	0.765	0.773	0.774	0.842	0.829	0.859
14	$ \bar{G}_1 $	0.60	0.59	0.39	0.48	0.59	0.51	0.40	0.37	0.45	0.43	0.45	0.76
	$ \bar{G}_2 $	0.73	1.24	0.79	0.96	1.12	1.00	0.84	0.73	0.63	0.72	0.62	1.53
	$\sum_{i=1}^{14} \lambda_i / \sum_{i=1}^{35} \lambda_i$	0.932	0.938	0.912	0.912	0.897	0.880	0.875	0.874	0.878	0.922	0.913	0.937
10	正态站数	34	29	34	31	28	29	31	32	34	32	33	28

由表6可见:(1)随着拟合精度的提高, $|\bar{G}_1|$ 、 $|\bar{G}_2|$ 减小。这表示拟合场更接近正态化。(2)当取10个特征向量拟合时,正态化的站次达到375(占89.3%),比实际场X的服从正态分布的站次(255次;占60.7%)增加28.6%。所以,自然正交展开方法既有浓缩资料的优点,又有使气象场正态化的优点。

七、小结

1.我国的月降水量的正态性随地区、随月份不同。一般来说,9月、3月正态性较好,冬季月份正态性差。上海、长沙的正态性较好,吐鲁番及天津、石家庄等站正态性较差。没有一个站各月降水量均服从正态分布。

2.求空间、时间平均,有利于正态化。

但仍有些区域、季节降水量不服从正态分布。

3.立方根变换正态化效率较高。4次方根变换、双曲正切变换适用于明显不正态的资料。

4.自然正交展开既是一种浓缩资料的方法,也是一种有效的正态化方法。

参 考 资 料

- [1] H. 克拉美, 统计学数学方法, 上海科学技术出版社, 1966年。
- [2] Виленин С. Я., Статистические методы исследования систем автоматического регулирования, Советское радио, 1967.
- [3] 徐尔源, 论年降水量之常态性, Vol.21, 气象学报, 1950.
- [4] 施能, 概率统计方法在中期天气预报中的应用, 气象, 1979, 7.

On normality and normalization of monthly and seasonal precipitation in China

Shi Neng, Chen Hui

(Nanjing Meteorological Institute)

Abstract

Normality of monthly and seasonal precipitation is discussed based on the data of 35 stations. Some method of precipitation normalization are compared.