

极值分布及其在气象上的应用

陈玉华 李正洪

(四川省气象学校)

在气象统计中,常用各种气象要素的极端值来说明气候特征。在大型工程的设计中,我们需要对一些有严重破坏性的自然灾害进行充分估计。例如,在建设高大建筑物时,必须考虑大风的破坏作用,设计时要选取适当的最大风速为依据;建造桥梁时,必须考虑到保用期内该河流段的可能最高水位,这些极值不仅有理论上的意义,而且也具有十分重要的现实意义,如果考虑不周,就可能造成很大的损失。但在设计时又不能盲目地加大安全系数,以免造成不必要的浪费。

本文就是为解决类似问题,就极值分布理论作些介绍,并用它来预测成都地区今后50年、100年的一日最大降水量。

一、极值分布

设 $x_1, x_2, \dots, x_n$ 为随机变量 X 的一个简单随机样本,则其中的最大值

$$x_{m \cdot n} = \max(x_1, x_2, \dots, x_n)$$

及最小值

$$x_{m \cdot n} = \min(x_1, x_2, \dots, x_n)$$

都是随机变量,它们的概率分布就称为极值分布。由于 $x_1, x_2, \dots, x_n$ 是相互独立且与 X 具有相同的分布,因此, $x_{m \cdot n}$ 的分布函数为

$$\begin{aligned} F_{m \cdot n}^*(x) &= p(x_{m \cdot n}^* < x) = p(x_1^* < x) p(x_2^* < x) \dots p(x_n^* < x) \\ &= F_1(x) F_2(x) \dots F_n(x) \\ &= [F(x)]^n \end{aligned} \quad (1)$$

而最小值的分布函数为

$$\begin{aligned} F_1^*(x) &= p(x_{m \cdot n}^* < x) \\ &= 1 - p(x_{m \cdot n}^* \geq x) \\ &= 1 - p(x_1^* \geq x, x_2^* \geq x, \dots, x_n^* \geq x) \\ &= 1 - \{[1 - p(x_1^* < x)][1 - p(x_2^* < x)] \dots [1 - p(x_n^* < x)]\} \\ &= 1 - \{[1 - F(x)][1 - F(x)] \dots [1 - F(x)]\} \\ &= 1 - [1 - F(x)]^n \end{aligned} \quad (2)$$

式中的 $x_1^*, x_2^*, \dots, x_n^*$ 是将 $x_1, x_2, \dots, x_n$ 按从小到大顺序排列的值,即

$$x_{m \cdot n} = x_1^* < x_2^* < \dots < x_n^* = x_{m \cdot n},$$

$F(x)$ 为随机变量 X 的分布函数,称为 X 的原始分布。本文只就极大值的分布进行讨论。

如果能把 x_n^* 的分布函数 $F_n^*(x)$ 找出来,就可以求 x_n^* 不超过某一特定值 A 的概率

$$F_n^*(A) = p(x_n^* < A)$$

也可以求出某个气象要素 N 年一遇的极大值 A_N 的概率

$$p(x_n^* \geq A_N) = 1 - F_n^*(A_N) = \frac{1}{N+1} \quad (3)$$

因此,我们的问题已化成了求极大值 x_n^* 的分布函数 $F_n^*(x)$ 的问题。如果 X 的原始分布 $F(x)$ 已知,可求出 $F_n^*(x)$, 问题就解决了。但是,一般说来, $F(x)$ 并不完全知道,虽然如此,我们可以设法寻找当 n 很大(一般都能满足这一条件)时 x_n^* 的渐近分布,从而进行推断。

数学家格涅坚科已证明, 当原始分布为指数型分布 (正态分布、皮尔森Ⅲ型分布、指数分布) 时, 样本极大值的分布以

$$F_n^*(x) = \exp \left[-e^{\frac{(x-b_n)}{a_n}} \right] \quad (4)$$

($-\infty < x < +\infty$), ($a_n > 0$)

为渐近分布。其中的 a_n 和 b_n 为分布的参数。如果我们能用实测样本把 a_n 和 b_n 估计出来, 问题就解决了。

二、参数估计

如何估计分布的未知参数 a_n 和 b_n 呢? 下面介绍三种估计方法。

1. 最小二乘法

对(4)式取两次对数, 并记

$$y = -\ln [-\ln F_n^*(x)] = \frac{x-b_n}{a_n} \quad (5)$$

若 $x_1 < x_2 < \dots < x_n$ 是经过变序后按从小到大排列的 x_n^* 的 m 个观测值, 我们希望找到(5)式中当 $x = x_i$ 时 ($i = 1, 2, \dots, m$), y 的相应值 y_i , 有了 $\{x_i, y_i\}$ 就可以用最小二乘法来估计 a_n 和 b_n 了。因 x_i 的取值是随机的, 且当 X 为连续型随机变量时, 可以证明

$$E[F_n^*(x)] = \frac{i}{m+1} \quad (6)$$

所以 $F_n^*(x)$ 是 $\frac{i}{m+1}$ 的无偏估计量, 即

表1 a_n 和 b_n 计算表 (x_i 的单位: mm, 为成都地区日最大降水量)

序号	x_i	dx_i	dx_i^2	$dx_i dy_i$	$F(x_i) = \frac{i}{m+1}$	y_i	dy_i
1	82.3	-96.1	9235.21	163.23546	0.04	-1.1690	-1.6986
2	99.5	-78.9	6225.21	114.88629	0.08	-0.9265	-1.4561
3	102.6	-75.8	5745.64	97.10738	0.12	-0.7515	-1.2811
4	113.6	-64.8	4199.04	73.56744	0.16	-0.6057	-1.1353
5	113.7	-64.7	4186.09	65.05585	0.20	-0.4759	-1.0055
6	131.3	-47.1	2218.41	41.69763	0.24	-0.3557	-0.8853
7	140.5	-37.9	1436.41	29.21711	0.28	-0.2413	-0.7709
8	145.9	-32.5	1056.25	21.45325	0.32	-0.1305	-0.6601
9	151.4	-27.0	729.00	14.87700	0.36	-0.0214	-0.5510
10	153.7	-24.7	610.09	10.92234	0.40	0.0874	-0.4422
11	164.8	-13.6	184.96	4.51928	0.44	0.1973	-0.3323
12	168.0	-10.4	108.16	2.29112	0.48	0.3093	-0.2203
13	172.9	-5.5	30.25	0.57640	0.52	0.4248	-0.1048
14	173.5	-4.9	24.01	-0.07546	0.56	0.5450	0.0154
15	174.3	-4.1	16.81	-0.58261	0.60	0.6717	0.1421
16	193.9	15.5	240.25	4.29660	0.64	0.8058	0.2772
17	194.1	15.7	246.49	6.64424	0.68	0.9528	0.4232
18	203.2	24.8	615.04	14.47328	0.72	1.1132	0.5836
19	212.1	33.7	1135.69	25.72658	0.76	1.2930	0.7634
20	213.4	35.0	1225.00	33.96050	0.80	1.4999	0.9703
21	224.9	46.5	2162.25	56.59515	0.84	1.7467	1.2171
22	243.6	65.2	4251.04	99.58648	0.88	2.0570	1.5274
23	299.7	121.3	14713.69	237.10511	0.92	2.4843	1.9547
24	409.7	231.3	53499.69	617.31657	0.96	3.1985	2.6689
合计	4282.6		114094.68	1734.45300		12.7102	
平均	178.4					0.5296	

$F_n^*(x)$ 以 $\frac{i}{m+1}$ 为中心摆动, 因而用 y_i 和 x_i 对应起来, 由

$(x_1, y_1), (x_2, y_2), \dots, (x_m, y_m),$

用最小二乘法, 即由

$$\sum_{i=1}^m (x_i - a_n y_i - b_n)^2 = \text{最小}$$

求得 a_n 和 b_n 的估计值 (表 1)

$$\begin{cases} \hat{a}_n = \frac{\sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^m (y_i - \bar{y})^2} \\ \hat{b}_n = \bar{x} - \hat{a}_n \bar{y} \end{cases} \quad (7)$$

其中, $\bar{x} = \frac{1}{m} \sum_{i=1}^m x_i$, $\bar{y} = \frac{1}{m} \sum_{i=1}^m y_i$ 。这样,

根据 x_n^* 的样本就估计了 x_n^* 的渐近分布 $\hat{F}_n^*(x)$ 。从而, 便可以预测某个随机变量今后 N 年一遇的极大值了。一般地, 在预测今后 N 年一遇的极大值时, 把过去 m 年的资料结合考虑。因此, 实际上是预测 $m+N$ 年内可能发生的最大值。

2. 矩法

由(4)式知密度函数

$$f(x) = \frac{1}{a_n} e x p \left[-\frac{(x-b_n)}{a_n} - e^{-\frac{(x-b_n)}{a_n}} \right] \quad (8)$$

通过积分可得如下关系

$$\begin{cases} a_n = \frac{\sqrt{6}}{\pi} \sigma_{x_n^*} \\ b_n = E(x_n^*) - \gamma \frac{\sqrt{6}}{\pi} \sigma_{x_n^*} \end{cases} \quad (9)$$

其中, $\gamma = 0.57722$, 为欧拉常数, x_n^* 表示极大值, $E(x_n^*)$ 为 x_n^* 的数学期望, $\sigma_{x_n^*}$ 为 x_n^*

的标准差。以 $\bar{x}$ 估计 $E(x_n^*)$, S_x 估计 $\sigma_{x_n^*}$,

从而得 a_n 和 b_n 的估值

$$\begin{cases} \hat{a}_n = 0.7797 S_x \\ \hat{b}_n = \bar{x} - 0.450 S_x \end{cases} \quad (10)$$

因为超过保证率函数

$$\begin{aligned} p(x_n^* \geq x) &= 1 - F_n^*(x) \\ &= 1 - e x p \left[-e^{-\frac{(x-b_n)}{a_n}} \right] \end{aligned} \quad (11)$$

所以有

$$\begin{aligned} x &= \bar{x} - 0.450 S_x + 0.7797 S_x \\ &\quad \left\{ -\ln \left[-\ln \left(1 - \frac{i}{m+1} \right) \right] \right\} \end{aligned} \quad (12)$$

式中 $\frac{i}{m+1}$ 为资料按从大到小排列后的超过保证率。

3. 耿贝尔法

令(4)式中 $\frac{1}{a_n}(x-b_n) = Y$, 则

$$p(y) = p(Y \geq y) = 1 - e^{-e^{-y}} \quad (13)$$

$$E(y) = \frac{1}{a_n} [E(x) - b_n]$$

$$D(y) = \frac{1}{a_n^2} D(x)$$

由此得

$$\begin{cases} a_n = \frac{\sigma_x}{\sigma_y} \\ b_n = E(x) - a_n E(y) = E(x) - \frac{\sigma_x}{\sigma_y} E(y) \end{cases} \quad (14)$$

设由实测的 m 个样本极大值, 按其从大到小的顺序排列时, 可得平均值 $\bar{x}$ 和标准差 S_x , 将它们作为 $E(x)$ 和 σ_x 的估计值。为了计算 $E(y)$ 和 σ_y 的估计值, 需要确定每一个 x_i 所对应的 y_i 值。我们由经验保证率

$$p = \frac{i}{m+1}, \text{ 得}$$

$$p = 1 - F_n^*(x) \approx \frac{i}{m+1} \quad (15)$$

(15)式是将 m 个极大值由大到小排列的结果。

因为 $y = \frac{x-b_n}{a_n}$, 则有

$$\frac{i}{m+1} = 1 - e^{-e^{-y}}$$

表2 样本容量m与 $\bar{y}(1)$

m	0	1	2	3	4	5	6	7	8	9
10	0.4952	0.4996	0.5033	0.5070	0.5100	0.5123	0.5157	0.5181	0.5202	0.5220
20	0.5236	0.5252	0.5268	0.5283	0.5296	0.5309	0.5320	0.5332	0.5343	0.5353
30	0.5362	0.5371	0.5383	0.5388	0.5396	0.5402	0.5410	0.5418	0.5424	0.5430
40	0.5436	0.5442	0.5448	0.5453	0.5458	0.5463	0.5463	0.5473	0.5477	0.5481
50	0.5485	0.5489	0.5493	0.5497	0.5501	0.5504	0.5508	0.5511	0.5515	0.5518
60	0.5521	0.5524	0.5527	0.5530	0.5533	0.5535	0.5538	0.5540	0.5543	0.5546
70	0.5548	0.5550	0.5552	0.5555	0.5557	0.5559	0.5561	0.5563	0.5565	0.5567
80	0.5569	0.5570	0.5572	0.5574	0.5576	0.5578	0.5580	0.5581	0.5583	0.5585
90	0.5586	0.5587	0.5589	0.5591	0.5592	0.5593	0.5595	0.5596	0.5598	0.5599
100	0.5600									

表3 样本容量m与 $S_y(1)$

m	0	1	2	3	4	5	6	7	8	9
10	0.9496	0.9676	0.9333	0.9971	1.0095	1.0206	1.0316	1.0411	1.0493	1.0565
20	1.0623	1.0696	1.0754	1.0811	1.0834	1.0915	1.0961	1.1004	1.1047	1.1086
30	1.1124	1.1159	1.1193	1.1226	1.1255	1.1285	1.1313	1.1339	1.1363	1.1388
40	1.1413	1.1436	1.1458	1.1480	1.1499	1.1519	1.1538	1.1557	1.1574	1.1590
50	1.1607	1.1623	1.1638	1.1658	1.1667	1.1681	1.1696	1.1708	1.1721	1.1734
60	1.1747	1.1759	1.1770	1.1782	1.1793	1.1803	1.1814	1.1824	1.1834	1.1844
70	1.1854	1.1863	1.1873	1.1881	1.1890	1.1898	1.1906	1.1915	1.1923	1.1930
80	1.1938	1.1945	1.1953	1.1959	1.1967	1.1973	1.1980	1.1987	1.1994	1.2001
90	1.2007	1.2013	1.2020	1.2026	1.2032	1.2038	1.2044	1.2049	1.2055	1.2060
100	1.2065									

易得 $y_i = -\ln \left[-\ln \left(1 - \frac{i}{m+1} \right) \right] \quad (16)$

$(i = 1, 2, \dots, m)$

由这些 y_i 可求得 $\bar{y}$ 及 S_y 作为 $E(y)$ 及 σ_y 的估计值, 则得

$$\begin{cases} \hat{a}_n = \frac{S_y}{S_y} \\ \hat{b}_n = \bar{x} - \frac{S_y}{S_y} \bar{y} \end{cases} \quad (17)$$

由(16)式知 $\bar{y}$ 与 S_y 只与m有关, 在应用时查表2和表3就行了, 无需计算。

由(13)式得

$$x_p = \hat{b}_n - \hat{a}_n \ln[-\ln(1-p)] \quad (18)$$

其中的 p 为超过 x_p 的保证率。由(18)式就可预测 $1/p$ 年一遇的极大值了。

三、拟合分布曲线的适线法

用观测资料拟合理论分布, 包括选择理论分布模型, 计算模型中的参数以及检验所选理论分布模型的正确性等内容。本文采用实际工作中常用的图解法来拟合理论分布, 可以减少工作量。

因为 $x_p = b_n - a_n \ln[-\ln(1-p)]$

将(9)代入得

$$x_p = E(x_n^*) - \frac{\sqrt{6}}{\pi} \sigma_{x_n^*} \{0.57722 + \ln[-\ln(1-p)]\} \quad (19)$$

$$\text{令 } \Phi_p = -\frac{\sqrt{6}}{\pi} \{0.57722 + \ln[-\ln(1-p)]\}, \text{ 则}$$

$$x_p = E(x_n^*) [\Phi_p C_v + 1] \quad (20)$$

其中, $C_v = \frac{\sigma_{x_n^*}}{E(x_n^*)}$ 为极大值的变差系数, Φ_p

称为耿贝尔分布的离均系数, 它只与保证率 P 有关, 计算时可查表(4)。将用矩法所得 $E(x_n^*)$, $\sigma_{x_n^*}$ 的估计值 $\bar{x}$, S_x 代入(20)式, 得各种 P 值所对应的 x_p 值, 从而得到 x_n^* 的理论分布。将它们点绘在包伟尔为此分布而设计的概率格纸上应为一-直线。所以, 在点绘保证率曲线时, 只需要计算两个点就行了。

由(20)式可看到, 当 $x_p = E(x_n^*)$ 时, $\Phi_p = 0$, 则 $p = 42.96\%$ 。这就是说总体平均值固定在 $p = 42.96\%$ 的地方。因此, 实际上只要计算一个点就行了。连接两点所得的直线就是所求的理论分布直线。

画出理论分布曲线后, 将实测的经验保证率(按表1中的 $1 - \frac{i}{m+1}$ 计算得 p , 与 x_i 对应)点在同一图上, 比较拟合程度, 若满意即告结束, 若拟合不好, 可适当调整 C_v , 重新计算理论分布, 直至满意为止。

表4 Φ_p 与 p 的关系表(1)

$p(\%)$	Φ_p	$p(\%)$	Φ_p
0.001	8.53	30	0.35
0.01	6.73	40	0.07
0.05	5.48	50	-0.16
0.1	4.94	60	-0.38
0.5	3.68	70	-0.59
1	3.14	80	-0.82
2	2.59	90	-1.10
3	2.27	95	-1.31
5	1.87	97	-1.43
10	1.30	99	-1.64
20	0.72	99.9	-1.96

若无包伟尔概率纸, 也可以用普通的概率纸来拟合理论分布, 横坐标用保证率 p 表示, 纵坐标用 x_p 来表示, 点绘理论分布曲线。本例就是如此。

四、成都地区50年和100年一遇的一日最大可能降水量

由于一日最大降水量受多种因素的影响, 而且单点站的资料具有一定的局限性, 不能完全代表一地的情况。为更符合实际, 我们选用地理条件相似, 天气过程基本一致的成都地区13个站的1963年至1986年共24年的历年一日最大降水量资料作为成都地区历年一日最大降水量资料, 来预测成都地区50年和100年一遇的一日最大可能降水量。

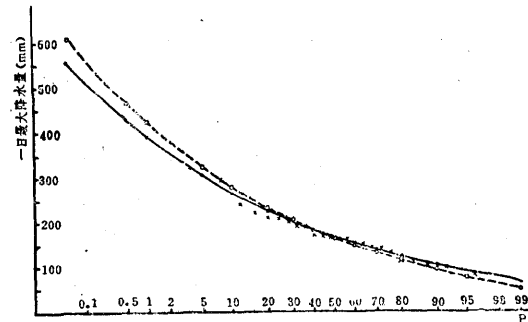
用三种不同的方法预测的结果见表5。

表5 成都地区一日最大降水量

	最小二乘法	矩法	耿贝尔法
$\hat{a}_n(\text{mm})$	61.226	53.76	63.465
$\hat{b}_n(\text{mm})$	145.07	147.44	144.83
$x_{50}(\text{mm})$	409.9	379.1	418.4
$x_{100}(\text{mm})$	441.3	406.8	451.0

表5中的 x_{50} 和 x_{100} 分别代表成都地区今后50年和100年的一日最大降水量。从表5可见, 用最小二乘法和耿贝尔法估计参数 a_n 和 b_n 而计算的一日最大降水量的值比较接近, 而用矩法所得的预测值偏小。在实际工作中, 一般都用第一种方法或者第三种方法估计参数 a_n 和 b_n 。而不用矩法, 因为参数 a_n 和 b_n 不能简单地表示成仅为 x 的矩的函数。

那么, 成都地区历年一日最大降水量是否真的服从我们所假定的指数分布呢? 为此, 我们进行了适线法检验(附图), 方法如



附图 适线法检验图

(三)所述。不同的只是用的普通概率纸,没有用包伟尔概率纸。按(20)式计算理论保证率,并将 p 为横坐标,对应的 x_p 为纵坐标)单位为mm)点在概率纸上,用光滑的曲线连接。就是理论分布曲线。图中实线为 $C_v = 0.386$,断线经适当调整后取 $C_v = 0.44$ 而画的理论分布曲线。然后按表(1)中 $p =$

$1 - F_n^*(x)$ 的实测经验保证率 $1 - \frac{i}{m+1}$ 点

在同一图上(图中带“×”者)。从图上可见拟合较好,特别是 $C_v = 0.44$ 的线拟合更好些,这就说明了成都地区历年一日最大降水量是服从所假定分布的。由此图可预测成都地区任意多年一遇的一日最大降水量。

另外,若用成都站单点资料,预测得成

都50年一遇的一日最大降水量等于240.8mm,100年一遇的一日最大降水量等于257.4mm。成都今后50年一遇的极大瞬时风速等于 $30.1\text{m}\cdot\text{s}^{-1}$ 。由重庆沙坪坝1957—1980年的降水量资料,用极值分布理论预测得沙坪坝今后50年一遇的一日最大降水量等于270.9mm

总之,我们认为用单点站资料来预测有较大的局限性,应该用地理条件相似,天气过程也基本一致的相邻数点的资料综合考虑来进行预测,才能代表那个地区的情况,使其更符合实际,更有实用价值。

参 考 文 献

- [1]屠其璞等编,气象应用概率统计学,气象出版社,1984.