

70-75

利用正态分布计算年最大波高

侯瑞科

(大连中心海洋站)

p731.22

提 要

本文提出用正态分布计算年最大波高,这种方法与当前使用的计算方法主要不同是,它具有较完善的子样检验方法,并且在一定的概率下计算出年最大波高。

一、引 言

在海洋工程设计中,对有严重破坏性的自然灾害要进行充分的估计。例如,在海上修建码头、栈桥和防波堤时,为了防止海浪冲坏码头、栈桥和防波堤这类事故的发生,设计时,要按照工程设计要求确定若干年一遇的最大波高。例如,50年一遇或100年一遇,甚至更长时间一遇的最大波高,也就是一定概率下的年最大波高。

若把每年最大波高视作随机变量 ξ ,则能把 ξ 的分布函数

$$F(x) = P(\xi < x) \quad (1)$$

找出来,就可以求出年最大波高不超过 x 的概率,也可以求出 N 年一遇的年最大波高

$$P(\xi \geq x) = 1 - F(x) = \frac{1}{N+1} \quad (2)$$

其中 $F(x)$ 表示年最大波高分布函数。

二、年最大波高的分布

海洋上的年最大波高,主要是由台风、气旋和冷空气引起,例如,黄海海域几乎每年都受到台风影响,根据海洋局出版的潮汐观测资料查出,大连老虎滩海区最大波高极值是8m,发生在1974年8月30日08时,由7416号台风引起的。台风引起的最大波高可以看作是由许多随机因子的影响所致,因此,可以用正态分布拟合海上年最大波高。

为了确定年最大波高总体 ξ 的分布函数 $F(x)$,这里引入渐近分布的估计理论,若年最大波高总体 ξ 的分布函数 $F(x)$ 满足^[1]

$$\lim_{x \rightarrow \infty} \frac{d}{dx} \left(\frac{1 - F(x)}{F(x)} \right) = 0$$

本文于1990年4月收到修改稿。

则 ξ^* 以 $A(x)$ 类型分布为渐近分布:

$$F_*(x) = A\left(\frac{x-b_*}{a_*}\right) = \exp\left[-e^{-\frac{x-b_*}{a_*}}\right] \quad (3)$$

因而求 ξ^* 的渐近分布问题便归结为由 ξ^* 的观察值来估计它的渐近分布中的参数 a_* 和 b_* 的问题。为了估计 a_* 和 b_* , 由(3)式得:

$$y = -\ln(-\ln F_*(x)) = \frac{x-b_*}{a_*} \quad (4)$$

若 $x_1 < x_2 < \dots < x_m$ 是经过变序按大小排列的 ξ^* 的 m 个观察值, 我们希望找到(4)式中 $x = x_i$ 时 y 的相应值 y_i , 有了 $\{(x_i, y_i)\}$ 便可通过最小二乘法来估计 a_* 和 b_* 了。因为 x_i 的取值是

随机的, 因而我们来计算 $M(F_*(x_i))$, $M(F_*(x_i)) = \frac{i}{m+1}$, 所以 $F_*(x_i)$ 为 $\frac{i}{m+1}$

的无偏估计, 即 $F_*(x_i)$ 以 $\frac{i}{m+1}$ 为中心而摆动, 因而我们用 $y_i = -\ln\left[-\ln\left(\frac{i}{m+1}\right)\right]$

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

$$Q = \sum_{i=1}^m (x_i - a_* y_i - b_*)^2 = \min$$

求得 a_* 和 b_* 的估计为:

$$\begin{cases} \hat{a}_* = \frac{\sum_{i=1}^m (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^m (y_i - \bar{y})^2} \\ \hat{b}_* = \bar{x} - \hat{a}_* \bar{y} \end{cases} \quad (5)$$

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, 我们便根据 ξ^* 的子样, 估计了 ξ^* 的渐近分布中的参数 a_* , b_* , 也就找到了 ξ^* 的渐近分布的估计函数

$$\hat{F}_*(x) = \exp\left[-e^{-\frac{x-\hat{b}_*}{\hat{a}_*}}\right] \quad (6)$$

有了 $\hat{F}_*(x)$, 便可预测今后 N 年最大波高了。通常, 在预测今后 N 年最大波高时, 把过去的年最大波高资料结合考虑, 因而实际上是求预测 $m+N$ 年中可能发生的最大波高, 也就是求

$$\hat{F}_*(x) = \frac{N+m}{N+m+1} \text{ 的解。}$$

三、年最大波高分布函数的检验

在上面通过渐近分布估计理论, 找到了年最大波高总体的渐近分布函数 $\hat{F}_*(x)$, 但这

种分布是否正态分布呢?在数理统计中,我们可以通过年最大波高的子样对总所服从的分布类型进行推断,其方法有图估计法或皮尔逊 χ^2 检验法。图估计法比较直观简便,只要把观测序列一点一点绘在一张正态概率纸上,若点的分布在纸上呈一直线,说明子样来自正态总体,反之则拒绝。

皮尔逊 χ^2 检验方法是将被测数据作为子样,并根据子样的分布特征,检验子样是否来自正态总体,这种方法效率较高,也较常用,具体作法是:

对于分布函数的类型已知,而参数未知的 χ^2 检验法:

1. 假设 $H_0: \xi \sim F(x)$, 已知 $F(x)$ 的类型, 但有 l 个参数 $\lambda_1, \lambda_2, \dots, \lambda_l$ 。
2. 用最大似然法确定 l 个参数的估计量 $\hat{\lambda}_1, \hat{\lambda}_2, \dots, \hat{\lambda}_l$ 。
3. 根据样本值的范围,把实轴分成 K 个不相交区间 $S_i = (C_i, C_{i+1})$; $S_i = (C_i, C_{i+1})$, ($i=1, 2, \dots, K$)。使样本在每个子区间 S_i 中出现的个数 $\mu_i \geq 5$, 计算理论频率:

$$\gamma_i = F(C_{i+1}) - F(C_i) = P(C_i \leq \xi < C_{i+1})$$

4. 选取统计量:
$$\chi^2 = \sum_{i=1}^K \frac{(\mu_i - n\gamma_i)^2}{n\gamma_i} \sim \chi^2(K-l-1)$$
, 式中 n 为样本实验或

观察次数。

5. 给出置信度 α , 确定否定域 $\chi^2 \geq \chi_{\alpha}^2(K-l-1)$ 。
6. 计算统计量 χ^2 , 并作出判断: 当 $\chi^2 \geq \chi_{\alpha}^2(K-l-1)$ 时, 拒绝原假设 H_0 ; 当 $\chi^2 < \chi_{\alpha}^2(K-l-1)$ 时, 接受原假设 H_0 。

通过上述两种检验方法, 可以知道年最大波高总体分布函数 $F(x)$, 并在一定概率下计算年最大波高。

四、实例计算

已知大连老虎滩海洋站1963年~1987年逐年最大波高资料 $\{x_i\}$, 见表1。

表1 老虎滩站1963~1987年年最大波高

时间 t (年)	1963	64	65	66	67	68	69	70	71	72	73	74	75
波高 (m)	2.6	3.4	4.0	3.3	4.5	3.2	4.6	3.1	3.2	4.0	4.2	8.0	2.6
时间 t (年)	1976	77	78	79	80	81	82	83	84	85	86	87	
波高 (m)	3.1	4.6	3.8	3.4	2.9	2.1	2.8	2.5	3.0	4.0	2.1	1.9	

计算25年、30年、50年、100年、200年、500年和1000年一遇的最大波高值。

首先验证年最大波高的子样是来自正态总体, 试分别用图估计法或皮尔逊 χ^2 检验法进行检验, 其作法如下:

1. 图估计法: 以大连老虎滩海洋站25年的年最大波高为子样, 作正态分布检验,

(1) 作试验数据的分组累计频率表如表2:

表2 分组累计频率表

组 别	(1.90~2.45)	(2.46~3.00)	(3.01~3.55)	(3.56~4.10)	(4.11~4.65)	(>4.65)
频 数 (次)	3	5	8	4	4	1
累计频数 (次)	3	8	16	20	24	25
累计频率 (%)	12	32	64	80	96	100

(2) 以分组累计频率表中分组的右端点为横坐标, 累计频率为纵坐标, 在正态概率纸上描点, 得到这些点近似地在一条直线上, 所以可认为年最大波高是服从正态分布的。

2. 用 ξ 表示年最大波高, 试用皮尔逊 χ^2 检验 ξ 服从正态分布, 其中 $\alpha=0.05$, 其检验方法是:

假设 $H_0: \xi \sim N(\mu, \delta^2)$, 这里 μ, δ^2 是未知参数, 由 μ, δ^2 的最大似然估计为:

$$\hat{\mu} = \frac{1}{25} \sum_{i=1}^{25} x_i = 3.5, \quad \hat{\delta}^2 = \frac{1}{25} \sum_{i=1}^{25} (x_i - \bar{x})^2 = 1.11$$

在实轴上取4个实数 $t_1, t_2, \dots, t_4$ 为: 2.51, 3.08, 3.65, 4.22, 将实轴分成5个左开右闭区间, 用唱票的办法求得数据落入各区间的频数 μ_i , 计算出 γ_i 及中间结果 $n\gamma_i, (\mu_i - n\gamma_i)^2, \frac{(\mu_i - n\gamma_i)^2}{n\gamma_i}$, 填在如表3:

表3 皮尔迪 χ^2 检验有关数据

$$(\hat{\mu} = 3.5, \hat{\delta} = 1.05, n = 25)$$

界 限	(- < 2.51)	(2.52 ~ 3.08)	(3.09 ~ 3.65)	(3.66 ~ 4.22)	(> 4.22)
μ_i	4	5	7	5	4
γ_i	0.1736	0.171	0.2111	0.1992	0.2451
$n\gamma_i$	4.340	4.275	5.2775	4.980	6.1275
$(\mu_i - n\gamma_i)^2$	0.1156	0.5256	2.9870	0.0004	4.5263
$\frac{(\mu_i - n\gamma_i)^2}{n\gamma_i}$	0.0266	0.1229	0.5622	0.0001	0.7387

表中 μ_i 表示每个小区间中出现的年最大波高值的个数, γ_i 表示理论频率, $\gamma_i = F(C_{i+1}) - F(C_i) = P(C_i < \xi \leq C_{i+1})$, n 表示样本实验次数。

由上表计算出统计量 χ^2 为:

$$\chi^2 = \sum_{i=1}^5 \frac{(\mu_i - n\gamma_i)^2}{n\gamma_i} = 1.45$$

在给定信度 $\alpha=0.05$ 下, 查 χ^2 分布表得 $\chi_{0.05}^2(5-2-1) = 5.991$, 因为

$x^2 = 1.45 < x_{0.05}^2(5-2-1) = 5.991$, 所以接受 H_0 , 可以认为大连老虎滩海区年最大波高是服从正态分布的, 其分布函数的类型为:

$$F(x) = \exp[-\exp(-y)] = \exp\left[-\exp\left(-\frac{x-b}{a}\right)\right] \quad (7)$$

其中:

$$y = \frac{x-b}{a} \quad (8)$$

为了估计参数 a 和 b , 由 ξ_n^* 的观察值 x_i 找出对应的 y_i 值, 将 (7) 式改为:

$$y_i = -\ln\left[-\ln F(x_i)\right] = \frac{x_i - b}{a},$$

其中

$$F(x_i) = \frac{n-i+1}{n+1},$$

将 x_i 与 y_i 的对应关系列表, 见表4:

表4 x_i 与 y_i 对应关系表

i	x_i	$F(x_i) = \frac{n-i+1}{n+1}$	y_i	i	x_i	$F(x_i) = \frac{n-i+1}{n+1}$	y_i	i	x_i	$F(x_i) = \frac{n-i+1}{n+1}$	y_i
1	8.0	0.96	3.22	10	3.4	0.62	0.71	19	2.8	0.27	-0.27
2	4.6	0.92	2.53	11	3.4	0.58	0.60	20	2.6	0.23	-0.39
3	4.6	0.88	2.12	12	3.3	0.54	0.48	21	2.6	0.19	-0.50
4	4.6	0.85	1.77	13	3.2	0.50	0.37	22	2.6	0.15	-0.63
5	4.2	0.81	1.56	14	3.2	0.46	0.26	23	2.1	0.12	-0.77
6	4.0	0.77	1.35	15	3.1	0.42	0.15	24	2.1	0.08	-0.94
7	4.0	0.73	1.17	16	3.1	0.38	0.04	25	1.9	0.04	-1.18
8	4.0	0.69	0.99	17	3.0	0.35	-0.06				
9	3.8	0.65	0.87	18	2.9	0.31	-0.17				

用表4, 可以利用最小二乘法原理找 x_i 与 y_i 的线性关系确定参数 a 和 b 为:

$$\begin{cases} a = \frac{\sum_{i=1}^n x_i y_i - n \bar{x} \bar{y}}{\sum_{i=1}^n y_i^2 - n \bar{y}^2} = 1.01843 \\ b = \bar{x} - a \bar{y} = 2.94023 \end{cases}$$

因此, 由 (8) 式得到计算 N 年一遇的最大波高公式为:

$$x = 2.94023 + 1.01843y \quad (9)$$

由 $F(x) = \frac{N+m}{N+m+1}$ 查表^[2], 求出对应的 y 值, 将 y 值代入 (9) 式中, 即可以求出今后若干年中可能出现的最大波高。

现在利用公式(9)计算大连老虎滩海区今后若干年中可能出现的最大波高值如表5:

表5 最大波高可能再现年

t 年	25	30	50	100	200	500	1000
X_{max} (m)	5.2	7.0	7.3	7.9	8.5	9.3	10.0

五、小 结

目前,关于年最大波高的统计方法有多种,各种方法计算的结果都不一样,因此很难比较各种方法的好坏。例如,用龚贝尔(GUMBEL)极值分布方法计算年最大波高值,其计算方法虽然简便,但计算出的结果比较粗造。因为它没有较完善的子样检验方法,而只要当样本 n 很大时,就可以作正态分布计算。如果从计算方法的合理性来看,正态分布有以下特点:

只要对变量作变换,就可用熟知的正态分布性质加以讨论,并给计算过程带来方便。

用统计方法计算年最大波高,存在随机误差,因此,必须在给定的概率下来计算最大波高。这对海洋工程设计部门是有参考价值的。

参 考 文 献

- 〔1〕 上海复旦大学数学系,概率论与数理统计,上海科学技术出版社,第二版,1978年3月。
- 〔2〕 自学函授概率论与数理统计,北京工业学院出版社,1986年12月。