第 35卷 增刊
1996年

中山大学学报 (自然科学版)
ACTA SCIENTIARUM NATURALIUM
UNIVERSITATIS SUNYATSENI

Vol. 35 Suppl.
1996

水文频率线型选优的最小信息熵准则*

黄克中 张金阳**
(中山大学城市与资源规划系, 广州 510275)

摘 要 根据信息熵的基本概念,提出了最小信息熵准则作为水文频率线型选优的判别标准,给出了推导频率分布信息熵的一般方法,及各种信息熵的计算公式,最后举出应用算例.

关键词 水文频率, 线型选择, 信息熵, 四参数普遍 Γ分布

线型选择和参数估计是水文频率计算中的两大问题. 迄今, 关于参数估计方法已有较多的研究; 对线型选择的研究却遇到较大的困难, 水文频率应服从何种分布, 在理论上尚缺乏充分论证. 世界各国多以样本资料为据, 经过大量的适线分析、总结经验, 选用符合本国大多数水文系列的线型[1~3]. 当然, 由于水文频率分布具有很大的地域性, 在一个幅员广大的国家里, 也可能同时使用其他线型[4]. 在我国, 自 80年代以来, 随着水文资料的增加和电子计算机的广泛使用, 对线型问题重新进行了检验和评价[5~10]. 其中有人建议使用通用性很强的四参数普遍 Γ分布 (简称四参数 Γ分布)[7~9].

我国水文界对线型的优选通常是以资料点据与理论频率曲线配合良好作为准则; 但由于“配合良好”的概念具有模糊性, 结果会因人而异; 因此, 人们继而将之定量化, 以水文变量的绝对离差和 (或离差平方和) 最小作为准则. 在此基础上, 为了反映资料点据与所配曲线在纵横坐标两个方向上的拟合情况, 文〔6〕补充了“频率绝对离差和最小”准则. 大量的适线工作表明, 有时不同线型的绝对离差和彼此相差不大, 但曲线外延后的水文变量却相差较大. 文〔11〕的研究证实了在绝对离差和准则下, 拟合优度与设计值精度之间是零相关; 而在离差平方和准则下, 两者相关较小. 值得注意的是, 文〔12〕曾根据信息熵理论, 将频率分布的理论熵与经验熵的绝对离差和作为准则, 利用信息熵进行拟合是一种新的尝试, 然而遗憾的是, 由于水文样本系列一般较短, 点据的间距又很不均匀, 导致在作频率直方图时所取的组间距具有较大的任意性, 从而使经验熵和理论熵都难以确定; 因此, 该准则至今未见付诸应用. 本文试图应用信息熵的基本概念, 提出一个线型优选准则, 给出了推导水文分布信息熵的一般方法及各种水文常用分布的信息熵计算公式, 并举出应用算例.

收稿日期: 1995-07-03
* 水利部、电力部水利水电科学基金资助项目
** 1991级自然地理专业研究生

1　线型优选的信息熵最小准则

设 x 代表某一随机系统的离散型随机变量，系统处于状态 x_i ($i=1,2,\cdots,n$)的概率为 $p(x_i)$，则系统的信息熵 H 表达为[13]

$$H=-\sum_{i=1}^{n}p(x_i)\ln p(x_i) \tag{1}$$

如果状态变量 x 是连续型，式 (1) 改写为

$$H=\int_{-\infty}^{\infty}f(x)\ln f(x)\,dx \tag{2}$$

式中，$f(x)$是概率密度函数.

为了给出由式 (1) 定义的信息熵的实际意义，设 $P(x_i|\hat{x_i})$ 是在已取得一个样本 $\hat{x_i}$ 以后，系统处于状态 x_i 下的 (后验) 概率. 样本所提供的信息量定义为 $\ln[p(x_i|\hat{x_i})/p(x_i)]$，如果样本的信息量愈大，则比值 $p(x_i|\hat{x_i})/p(x_i)$愈大；如果观测得到的样本资料完全没有误差，则 $p(x_i|\hat{x_i})=1$，所提供的信息量为 $-\ln p(x_i)$. 因为 $0\leqslant p(x_i)\leqslant 1$，故当 $p(x_i)$愈小，样本资料所提供的信息量愈大. 由此可见，由式 (1) 所定义的熵 H 是在随机试验系列进行之后样本资料所提供的平均信息量. 由于平均信息量的增加就是系统状态不确定性的减小，所以信息量的大小同时也是随机试验系列之前系统状态不确定性的量度. 确定系统所需的信息量愈大，系统的不确定性愈大；反之，确定系统所需的信息量愈小，系统的不确定性愈小.

将上述概念用于水文线型选优的思路如下：对考虑的线型，利用样本资料对参数进行估计，得到了概率密度分布曲线；显然，这时得到的分布仍然是一个随机系统，具有不确定性. 该系统在没有获得新信息之前，所具有的信息熵愈小，则不确定性愈小. 因此，可以选择几种可能适用的线型，利用相同的样本资料，分别对各线型的参数进行估计，并求出各自的信息熵值，其中信息熵最小的频率分布曲线即为最优的频率分布曲线. 这就是线型选优的最小信息熵准则.

2　信息熵的计算

为了应用最小信息熵准则，必须给出各种常用线型的信息熵计算式，它们可以由理论推导得到，文〔13〕对此给出了一些结果. 在此基础上，本文给出了通用的推导方法，并从四参数 Γ 分布出发，导得了该曲线族的信息熵通用计算式. 此外，也给出了别的常用分布的信息熵计算式.

推导信息熵的一般方法和过程如下：① 由已知的概率密度分布反求约束条件；② 应用最大熵原理，建立以拉格朗日乘子表达的概率密度函数，以及线型参数与拉格朗日乘子的关系；③ 通过各个拉格朗日乘子之间的偏导数关系，建立参数估计方程；④ 通过信息熵的基本定义式和以拉格朗日乘子表达的概率密度函数，建立信息熵的计算式.

由于四参数 Γ 分布包含了 P-Ⅲ型等多种常用水文分布[8,14]，所以这些同族分布的信息熵可以作为四参数 Γ 分布信息熵的特殊情形而得到. 由于对数正态分布和极值Ⅰ型分布的离均系数值分别与四参数 Γ 分布当 $b=10$ 和 $b=10$, $C_s=1.139$ 时的离均系数值近似相等[8,15]，所以当进行线型选优工作时，也可以近似地将这两种分布作为四参数 Γ 分布的

特殊情形而得到它们的信息熵; 当然, 也可以从它们各自的概率密度函数出发, 求得它们精确的信息熵计算式.

已有的研究[9,10]表明: 在四参数 Γ分布曲线族范围内的线型选优已转化为该分布的参数估计工作. 也就是说, 在该曲线族范围内的最优线型就是四参数 Γ分布. 因此, 在进行线型选优工作时, 应该选择四参数 Γ分布作为一种供比较的线型, 而不需要再选择与它同族的线型了; 余下就再选用一些与它非同族的线型以供比较, 其中最常用的是对数 P-Ⅲ型. 为了节省篇幅, 下面只以推导四参数 Γ分布的信息熵计算式为例, 来说明推导的方法和过程; 对其他分布, 将直接给出结果.

四参数 Γ分布的概率密度为

$$f(x) = \frac{}{}(x - c)^{(a/b)-1}\exp[-(x-c)^{1/b}]/[b\Gamma(a)] \qquad c \leqslant x < \infty \tag{3}$$

对(3)式取对数,有

$$\ln f(x) = \ln + (a/b - 1)\ln(x - c) - \ln[b\Gamma(\)] - (x-c)^{1/b} \tag{4}$$

式(4)两边乘以 $[-f(x)]$再积分,有

$$H[f(x)] = -\int_c^{\infty} f(x)\ln f(x)\,dx = -\{\ln - \ln[b\Gamma(\)]\}\int_c^{\infty} f(x)\,dx + \int_c^{\infty}(x-c)^{1/b}f(x)\,dx - (/b - 1)\int_c^{\infty}\ln(x-c)f(x)\,dx \tag{5}$$

通过考查上式,便得约束条件

$$\int_c^{\infty} f(x)\,dx = 1 \tag{6}$$

$$\int_c^{\infty}(x-c)^{1/b}f(x)\,dx = E[(x-c)^{1/b}] \tag{7}$$

$$\int_c^{\infty}\ln(x-c)f(x)\,dx = E[\ln(x-c)] \tag{8}$$

应用最大信息熵原理来求概率密度函数 $f(x)$的问题是一个条件泛函求极值问题,可利用拉格朗日方法.引入拉格朗日乘子 $(\lambda_0 - 1),\lambda_1,\lambda_2$,由式(3),(6~8)得无条件泛函 F 及 F 的变分 F为

$$F[f(x)] = -\int_c^{\infty}[\ln f(x) + (\lambda_0 - 1) + \lambda_1(x-c)^{1/b} + \lambda_2\ln(x-c)]f(x)\,dx$$

$$\delta F[f(x)] = -\int_c^{\infty}[\ln f(x) + \lambda_0 + \lambda_1(x-c)^{1/b} + \lambda_2\ln(x-c)]f(x)\,dx = 0$$

便得具信息熵意义的概率密度函数

$$f(x) = \exp[-\lambda_0 - \lambda_1(x-c)^{1/b} - \lambda_2\ln(x-c)] \tag{9}$$

将式(9)代入式(6)有

$$\lambda_0 = \ln\int_c^{\infty}(x-c)^{\lambda_2}\exp[-\lambda_1(x-c)^{1/b}]dx \tag{10}$$

对上式中的积分作变量置换

$$x = (y/\lambda_1)^b + c \tag{11}$$

并注意到 Γ函数的定义,积分后取对数得

$$\lambda_0 = \ln\{\Gamma[b(1-\lambda_2)]\} - \ln(1/b) - b(1-\lambda_2)\ln\lambda_1 \tag{12}$$

如果将式(12)代入式(9)有

$$f(x) = \exp\{-\ln[\Gamma(b(1-\lambda_2))] + \ln(1/b) + b(1-\lambda_2)\ln\lambda_1$$

$$-\lambda_1(x-c)^{1/b}-\lambda_2\ln(x-c)\}$$

$$=\{\Gamma[b(1-\lambda_2)]^{*}\ b\}^{-1}\lambda_1^{b(1-\lambda_2)}(x-c)^{-\lambda_2}\ \exp[-\lambda_1(x-c)^{1/b}] \tag{13}$$

令 $\lambda_1=U,\lambda_2=1-(a/b)$即 $1-\lambda_2=a/b$,有

$$f(x)=\{U^{T}/[b\Gamma(T)]\}(x-c)^{(a/b)-1}\exp[-U(x-c)^{1/b}] \tag{14}$$

式(14)与(3)完全一致,故推导无误.

对式(10),(12)分别取 $\partial/\partial\lambda_1$,经化简整理,有等式

$$\frac{\partial\lambda_0}{\partial\lambda_1}=b(\lambda_2-1)\ \lambda_1=-E[(x-c)^{1/b}]$$

即

$$T/U=E[(x-c)^{1/b}] \tag{15}$$

对式(10),(12)分别取 $\partial/\partial\lambda_2$,经化简整理,有

$$\frac{\partial\lambda_0}{\partial\lambda_2}=\frac{\partial\Gamma[b(1-\lambda_2)]}{\partial\lambda_2}+b\ln\lambda_2=-E[\ln(x-c)]$$

可改写为

$$b[\psi(a)-\ln U]=E[\ln(x-c)] \tag{16}$$

式中,$\psi(a)$为普西(Digamma)函数.

同理,对 λ_0 分别取 $\frac{\partial}{\partial\lambda_1^2}$ 和 $\frac{\partial}{\partial\lambda_2}$,可得如下两个方程

$$b(1-\lambda_2^2)\lambda_1^2=\mathrm{Var}[(x-c)^{1/b}]\qquad 即\ T/U^2=\mathrm{Var}[(x-c)^{1/b}] \tag{17}$$

$$b^2\frac{\mathrm{d}}{\mathrm{d}\Gamma}\psi(a)=\mathrm{Var}[\ln(x-c)]\qquad 即\ b^2\sum_{k=0}^{\infty}(T+k)^{-2}=\mathrm{Var}[\ln(x-c)] \tag{18}$$

式(15~18)即用于求解参数的普遍 Γ 熵法方程组[14].由式(9)有

$$\ln f(x)=-\lambda_0-\lambda_1(x-c)^{1/b}-\lambda_2\ln(x-c) \tag{19}$$

式(9)代入式(5)有

$$H=\lambda_0+\lambda_1E[(x-c)^{1/b}]+\lambda_2E[\ln(x-c)] \tag{20}$$

式(12)代入式(20)有

$$H=\ln\Gamma(T)+\ln b-T\ln U+UE[(x-c)^{1/b}]+[1-(T/b)]E[\ln(x-c)]$$

即

$$H=\ln[b\Gamma(T)/U^{T}]+UE[(x-c)^{1/b}]+[1-(T/b)]E[\ln(x-c)] \tag{21a}$$

以式(15)和(16)代入上式有

$$H=\ln[b\Gamma(T)/U^{T}]+T+(b-T)[\psi(T)-\ln U] \tag{21b}$$

式(21a)或(21b)为四参数 Γ 分布的信息熵计算式.

表1给出与四参数 Γ 分布同族(或近似同族)的5种常用分布的信息熵计算式和与它非同族的对数 P-Ⅲ分布的信息熵计算式.

表1　与四参数 Γ 分布同族或非同族的常用分布的信息熵

Tab.1　The information entropy formulas of the four-parameter gamma generalized distribution family (or non-family)

与四参数 Γ 分布的关系	线　型	概率密度函数 $f(x)$	熵法参数方程组	信息熵 H
同　族 $b=1$	P-Ⅲ分布	$U^{T}(x-c)^{T-1}\exp[-U(x-c)]/\Gamma(T)$ 参数　$T>0,U>0,0<c<x$	$T/U=E(x)-c$ $\psi(T)-\ln U=E[\ln(x-c)]$ $T/U^2=\mathrm{Var}(x-c)$	$\ln[\Gamma(T)/U^{T}]+T+(1-T)[\psi(T)-\ln U]$

(续表 1)

与四参数 Γ 分布的关系	线 型	概率密度函数 $f(x)$	熵法参数方程组	信息熵 H
同 族 $b=1, c=0$	Γ分布	$\alpha^{\beta}x^{\beta-1}\exp(-\alpha x)/\Gamma(\beta)$ 参数 $\beta>0, \alpha>0$	$\beta/\alpha=E(x)$ $\psi(\beta)-\ln\alpha=E(\ln x)$	$\ln[\Gamma(\beta)/\alpha^{\beta}]+\beta+(1-\beta)[\psi(\beta)-\ln\alpha]$
同 族 $\alpha=\beta/d^{1/b}$ $c=0$	克-P分布	$\beta x^{(\beta/b)-1}\exp[-\beta(x/d)^{1/b}]/[d^{\beta/b}b\Gamma(\beta)]$ 参数 $b>0, \beta>0, d>0$	$d^{1/b}=E(x^{1/b})$ $b=[E(\ln x)-\ln d]/[\psi(\beta)-\ln d]$ $d^{2/b}/\beta=\mathrm{Var}(x^{1/b})$	$\ln[bd^{\beta/b}\Gamma(\beta)/\beta^{\beta}]+\beta+(b-\beta)[\psi(\beta)-\ln(\beta/d^{1/b})]$
近似同族 $b=10$	对数正态分布	离均系数 H_p 与四参数 Γ 分布相近	$\beta/\alpha=E[x-c)^{1/10}]$ $10[\psi(\beta)-\ln\alpha]=E[\ln(x-c)]$ $\beta^2/\alpha=\mathrm{Var}[(x-c)^{1/10}]$	$\ln[10\Gamma(\beta)/\alpha^{\beta}]+\beta+(10-\beta)[\psi(\beta)-\ln\alpha]$
近似同族 $b=10$, $C_s=1.139$ (即 $\beta=740$)	极值 I型分布	离均系数 H_p 与四参数 Γ 分布相近	$740/\alpha=E[(x-c)^{1/10}]$ $10[\psi(740)-\ln\alpha]=E[\ln(x-c)]$	$\ln[10\Gamma(740)/\alpha^{740}]+740-730[\psi(740)-\ln\alpha]$
非同族	对数 P-III分布	$[y-c)/a]^{b-1}\exp[-(y-c)/a]/[a\Gamma(b)]$ 其中 $y=\ln x$ 参数 $a>0, b>0, 0<c<\ln x$	$ab=E(y)-c$ $\psi(b)+\ln a=E[\ln(y-c)]$ $a^2b=\mathrm{Var}(y)$	$\ln[a^b\Gamma(b)-(c/a)]+E(y)(a+1)/a-(b-1)E[\ln(y-c)]$

3 应用算例

从我国水文年鉴上摘录了 14个样本系列,水文变量包括年最大洪水流量和年平均流量,应用最小信息熵准则进行线型选优. 参加选优的线型是四参数 Γ 分布和对数 P-III分布. 此外,为了证实在四参数 Γ 分布族中,四参数 Γ 分布是最优的,也同时将 P-III分布和对数正态分布参加比较. 在计算中,为了避免受不同的估计参数方法的影响,现全部统一使用熵法 (估参方程见表 1). 参加优选的各种线型的信息熵值及选优结果见表 2,结果表明,各站均为四参数 Γ 分布最优. 从表 2中任选两站绘出几种线型的频率曲线 (见图 1和图 2),可以看到它们在拟合处延时的不同情形,其中关于四参数 Γ 分布频率曲线的求作方法见文献〔17, 18〕.

表 2 不同线型的信息熵

Tab. 2 The comparisons between the information entropy values of different distributions

河名	站名	变量	样本容量	信息熵 (单位: nat) 四参数 Γ 分布	对数 P-III分布	P-III分布	对数正态分布	最优线型
松花江	哈尔滨	年平均流量	78	7.537	7.574	7.537	7.571	四参数 Γ 及 P-III分布
黄河	循化	年最大流量	25	7.800	7.830	7.813	7.842	四参数 Γ 分布
渭河	华县	年平均流量	45	5.960	6.038	6.000	6.038	四参数 Γ 分布

(续表 2)

河名	站名	变量	样本容量	信息熵 (单位: nat)				最优线型
				四参数 Γ 分布	对数 P-Ⅲ 分布	P-Ⅲ 分布	对数正态分布	
长江	汉口	年平均流量	115	9.426	9.492	9.498	9.490	四参数 Γ 分布
漠阳江	双捷	年最大流量	26	7.882	7.997	7.966	8.007	四参数 Γ 分布
大通河	享堂	年平均流量	31	4.255	4.327	4.321	4.338	四参数 Γ 分布
郁江	南宁	年最大流量	33	8.920	9.003	8.984	9.012	四参数 Γ 分布
郁江	南宁	年平均流量	33	7.199	7.209	7.206	7.224	四参数 Γ 分布
北江	石角	年最大流量	31	9.362	9.539	9.469	9.539	四参数 Γ 分布
北江	石角	年平均流量	31	7.250	7.327	7.301	7.333	四参数 Γ 分布
东江	龙川	年最大流量	33	8.526	8.569	8.666	8.603	四参数 Γ 分布
东江	龙川	年平均流量	33	5.565	5.653	5.574	5.678	四参数 Γ 分布
西江	梧州	年最大流量	39	10.268	10.431	10.376	10.433	四参数 Γ 分布
西江	梧州	年平均流量	39	8.558	8.679	8.646	8.685	四参数 Γ 分布

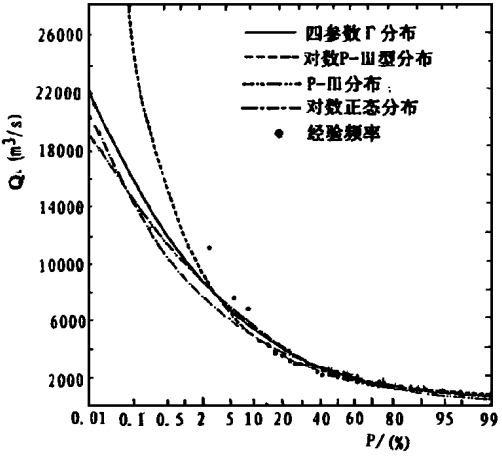

图 1　东江龙川站年最大流量不同理论频率曲线比较

Fig. 1　The comparisons between the hydrologic frequency curves of annual maximum discharge on Longchuan Station of Eastern River

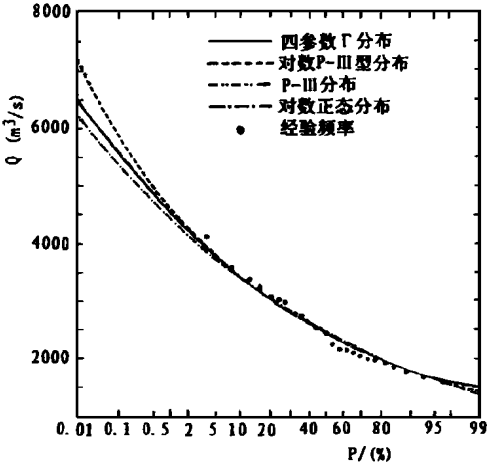

图 2　黄河循化站年最大流量不同理论频率曲线比较

Fig. 2　The comparisons between the hydrologic frequency curves of annual maximum discharge on Xunhua Station of Yellow River

本文由信息熵的基本概念出发,摒弃了传统的以资料点据与频率曲线拟合的概念,提出了最小信息熵准则作为水文频率线型选优的判别标准, 这是一个新的尝试. 最后还应强调, 使用本准则的基础仍然要求样本资料对总体有足够代表性; 同时, 应从当地的水文等自然地理条件对判别结果 (特别是外延情形) 的合理性进行分析.

参 考 文 献

1 苏联部长会议国家建设委员会. 设计河川水工建筑物时最大流量的计算规范. 周曾盛等译. 北京: 水利出版社, 1958

2 Interagency Advisory Committe on Water Data, et al, Guide for Datermining Flood Flow Frequency, Bulletin of the Hydrology Subcommittee. Editorial Corrections, 1982

3 水利电力部. 水利水电工程设计洪水计算规范 (试行). 北京: 水利出版社, 1980

4 水文统计分布综述. 国际水文科学协会讯息. 1990, 4

5 刘光文. 水文频率计算评议. 水文, 1986, 3

6 李松仕. 几种频率分布线型对我国洪水资料适应性的研究. 水文, 1984, 1

7 孙济良, 肖玉泉. 指数 Γ分布及其对洪水极值分布适应性的研究. 水利水电科学院研究论文集第 28集. 北京: 水电出版社, 1988

8 孙济良, 秦大庸. 水文频率分析通用模型研究. 水利学报, 1989, 4

9 孙济良. 论水文频率线优选及参数估计. 水力发电, 1992, 3

10 孙济良. 关于洪水频率分析中的线型问题. 水利水电技术, 1987, 6

11 丛树铮, 陈远芳. 适线法中拟合优度与设计值精度关系的分析. 河海大学学报, 1989, 4

12 Еусаласв И В. Выбор кривыф расиРелеления речно го стока и оченка их параметров методом минимакса ентролии. Метеороло ия и Гидропо ия, 1982, 6

13 Shannon C E. A mathematical theory of communication. July and ct, Bell System Techn J, 1948

14 黄克中. 一种通用水文频率分布的信息熵理论. 中山大学学报 (自然科学版), 1994, 增刊

15 李松仕. 指数 Γ分布及其在水文中的应用. 水利学报, 1990, 5

16 Singh V P, et al. Derivation of some frequancy disdributions using the principle of maximum entropy. Adv Water Resources, 1986, 9

17 水利水电科学研究院水资源所. 指数 Γ型分布曲线表. 1986

18 张金阳. 中山大学研究生毕业论文, 1994

Criterion of Minimum Entropy for Optimizing the Hydrologic Frequency Distribution

Huang Kezhong[*] *Zhang Jinyang*

Abstract On the basis of the information entropy concept, the criterion of minimum information entropy is prososed as a standard for optimizing the hydrologic frequency distribution in the paper. In view of the demand for calculating the entropy values, we also give the generalized derivation method of the information entropy and the information entropy formula of some useful hydrologic frequencys. Fourteen practical examples are given at last.

Keywords hydrologic frequency, option of distribution function, information entropy, four-parameter generalized gamma distribution

[*] Department of City and Resource Planning, Zhongshan University, Guangzhou 510275