

# 水文随机变量二维分布及其应用

芮孝芳

(河海大学水文水资源学院,江苏 南京 210098)

**摘要:**基于概率论对具有任意分布函数的两水文随机变量之间的关系进行了识别。对服从二维正态分布的水文随机变量之间的关系特点进行了分析。论述了寻求一般二维分布的 Copula 函数法和形变函数法的理论基础,指出 Copula 函数法的实质是通过一个连接函数将两个具有相关关系的变量的边际分布耦合成二维分布,而形变函数法的实质则是抓住二维分布的条件均值和条件方差两个主要数字特征,通过形变函数来构建二维分布。最后对二维分布在水文资料插补展延、概率水文预报、非一致性样本频率分析、水工程风险率确定、设计洪水、干支流洪水及洪潮遭遇组合、地貌瞬时单位线等水文学问题中的应用进行了剖析与展望。

**关键词:**二维分布函数; Copula 函数;形变函数;资料插补展延;概率水文预报;非一致性样本频率分析;水工程风险率;设计洪水;地貌瞬时单位线

**中图分类号:**TV11 **文献标志码:**A **文章编号:**1006-7647(2019)05-0036-07

**Binary distribution function of hydrology random variable and it's applications//RUI Xiaofang( College of Hydrology and Water Resources, Hohai University, Nanjin 210098, China)**

**Abstract:** The correlation between two hydrology random variables with arbitrary distribution function is distinguished. The characteristics of correlation between two hydrology random variables with the normal distribution function are analysed. The fundamentals of Copula function and deformation function determined binary distribution function between two hydrology random variables are discussed. The natures of Copula function method and deformation function method are pointed out. Finally, applications of binary distribution function in interpolation and extension of data, probability hydrology forecasting, frequency analysis of inconsistency sample, risk probability of water engineering, design flood, geomorphologic instantaneous unit hydrograph etc. are dissected and prospected.

**Key words:** binary distribution function; Copula function; deformation function; interpolation and extension of data; probability hydrology forecasting; frequency analysis of inconsistency sample; risk probability of water engineering; design flood; geomorphologic instantaneous unit hydrograph

作为随机变量的水文要素或水文特征值之间,由于物理原因,其中有一些或多或少存在着一定的因果联系,揭示并应用这些联系来处理或解决一些水文学问题,是水文学的重要研究内容之一。笔者第一次在概率统计指导下接触这一学术领域始于我国著名水文学家刘光文先生的学术讲座。笔者已经保存了 56 年的听课笔记清楚地记录着,那是 1963 年 5 月 20 日下午,刘光文先生作了题为“二元机率分配及相关的基本概念”的学术讲座。从“二元机率分配的基本概念”,到“变数之间的关系”,刘光文先生作了缜密而富有启发的讲解,令人耳目一新,令笔者至今记忆犹新。在日后漫长的岁月中,这一讲座所涉及的内容及透视出的科学思想无时无刻不在

笔者的学术生涯中起着指导性作用。刘光文先生这一学术讲座开启了我国水文学界研究水文随机变量二维分布及其应用的先河。本文试图根据半个多世纪以来这一领域的发展和笔者的思考与实践,从概念、理论到实际应用,进一步探索二维分布在处理或解决水文学问题中的思路和方法,以期引起研究二维分布及其应用的兴趣,踏踏实实,满怀信心,走守正创新之路。

## 1 两个随机变量之间关系的数学描述

水文随机变量之间可能存在函数关系或相关关系,也可能相互独立。两个随机变量中,若一个随机变量  $X$  的每个现实  $x$ ,都只与另一个随机变量  $Y$  的

一个现实  $y$  对应,则称这两个随机变量之间为函数关系,又称确定性关系。根据物理意义,两水文随机变量的函数关系属于因果函数关系。水文随机变量随时间、空间的变化虽然也是一种函数关系,但不是因果函数关系,而是数量函数关系。两个随机变量中,若对应一个随机变量  $X$  的每个现实  $x$ ,另一个随机变量  $Y$  将以不同的概率取不同的值,或者说,对应于随机变量  $X$  的每一个实现  $x$ ,随机变量  $Y$  将有不同的条件分布,则称这两个随机变量之间为相关关系。两个随机变量中,若对应一个随机变量  $X$  的每个现实  $x$ ,另一个随机变量  $Y$  将有完全相同的条件分布,则称这两个随机变量之间为独立关系。

两个随机变量的二维分布函数就是它们之间关系的最完整描述。因为,若两个随机变量  $X$  与  $Y$  为函数关系,则由随机变量函数的分布函数理论知,只要已知其中一个的分布函数,另一个的分布函数就可以推导出来。这表明这时二维分布实际上已退化为一维分布了。若两个随机变量  $X$  和  $Y$  相互独立,则由概率论知,二维分布将等于这两个随机变量的分布函数之乘积:

$$f(x,y) = f_x(x) \cdot f_y(y) \quad (1)$$

$$F(x,y) = F_x(x) \cdot F_y(y) \quad (2)$$

若两个随机变量  $X$  与  $Y$  为相关关系,则由概率论知,二维分布将为边缘分布与条件分布之乘积:

$$f(x,y) = f_x(x) \cdot f_y(y|x) = f_y(y) \cdot f_x(x|y) \quad (3)$$

或

$$F(x,y) = F_x(x) \cdot F_y(y|x) = F_y(y) \cdot F_x(x|y) \quad (4)$$

式中: $f(x,y)$  和  $F(x,y)$  分别两个随机变量  $X$  和  $Y$  的二维密度函数和二维分布函数; $f_x(x)$  和  $F_x(x)$  分别为随机变量  $X$  的密度函数和分布函数,或称二维分布关于  $X$  的边缘密度函数和边缘分布函数; $f_y(y)$  和  $F_y(y)$  分别为随机变量  $Y$  的密度函数和分布函数,或称二维分布关于  $Y$  的边缘密度函数和边缘分布函数; $f_y(y|x)$  和  $F_y(y|x)$  分别为  $X$  发生条件下  $Y$  的条件密度函数和条件分布函数; $f_x(x|y)$  和  $F_x(x|y)$  分别为  $Y$  发生条件下的  $X$  的条件密度函数和条件分布函数。

命题“两个随机变量之二维分布是它们之间关系的最完整描述”的科学性之所以毋庸置疑是因为,如果两个随变量为函数关系,那么其二维分布必退化为一维分布,反之,如果一个二维分布可表达为一维分布,那么这两个随机变量必为函数关系;如果两个随机变量相互独立,那么必满足式(1)或式(2),反之,如果二维分布可表达成式(1)或式(2),那么这两个随机变量必相互独立;如果两个随机变

量之间只具有一定的相关关系,那么必满足式(3)或式(4),反之,如果二维分布可表达成式(2)或式(3),那么这两个随机变量之间必定只具有一定的相关关系。

因此,所谓两个随机变量之间的数学描述,实际上就是构建两个随机变量的二维分布函数。

## 2 二维正态分布的两个随机变量之间的关系

二维正态分布是迄今为止,为数不多的能给出解析数学表达式的二维分布,其密度函数为<sup>[1-3]</sup>:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-r^2}} \exp \left\{ -\frac{1}{2(1-r^2)} \left[ \frac{(x-\bar{x})^2}{\sigma_x^2} - \frac{2r(x-\bar{x})(y-\bar{y})}{\sigma_x\sigma_y} + \frac{(y-\bar{y})^2}{\sigma_y^2} \right] \right\} \quad (5)$$

式中: $\bar{x}, \bar{y}$  分别为随机变量  $X, Y$  的均值,又称数学期望; $\sigma_x, \sigma_y$  分别为随机变量  $X, Y$  的均方差; $r$  为随机变量  $X$  与  $Y$  的 Pearson 相关系数; $\pi$  为圆周率常数。

由式(5)可得,二维正态分布的两个边缘分布均为二维正态分布,分别为

$$f_x(x) = \frac{1}{\sigma_x\sqrt{2\pi}} \exp \left( -\frac{x-\bar{x}}{\sigma_x} \right)^2 \quad (6)$$

和 
$$f_y(y) = \frac{1}{\sigma_y\sqrt{2\pi}} \exp \left( -\frac{y-\bar{y}}{\sigma_y} \right)^2 \quad (7)$$

两个条件分布也都是二维正态分布,分别为

$$f_y(y|x) = \frac{f(x,y)}{f_x(x)} = \frac{1}{\sigma_y\sqrt{2\pi}\sqrt{1-r^2}} \exp \left\{ -\frac{1}{\sigma_y\sqrt{1-r^2}} \left[ y-\bar{y}-r\frac{\sigma_y}{\sigma_x}(x-\bar{x}) \right]^2 \right\} \quad (8)$$

$$f_x(x|y) = \frac{f(x,y)}{f_y(y)} = \frac{1}{\sigma_x\sqrt{2\pi}\sqrt{1-r^2}} \exp \left\{ -\frac{1}{\sigma_x\sqrt{1-r^2}} \left[ x-\bar{x}-r\frac{\sigma_x}{\sigma_y}(y-\bar{y}) \right]^2 \right\} \quad (9)$$

由以上两点并非充分条件,因为反之并不一定成立。

由式(5)还可以发现,当  $r=0$  时,式(5)和式(8)、式(9)将分别变为

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y} \exp \left\{ -\frac{1}{2} \left[ \frac{(x-\bar{x})^2}{\sigma_x^2} + \frac{(y-\bar{y})^2}{\sigma_y^2} \right] \right\} = \frac{1}{\sqrt{2\pi}\sigma_x} \exp \left( -\frac{x-\bar{x}}{\sigma_x} \right)^2 \frac{1}{\sqrt{2\pi}\sigma_y} \exp \left( -\frac{y-\bar{y}}{\sigma_y} \right)^2 = f_x(x)f_y(y) \quad (10)$$

$$f_y(y|x) = \frac{1}{\sigma_y\sqrt{2\pi}} \exp \left[ -\frac{(y-\bar{y})^2}{\sigma_y^2} \right] = f_y(y) \quad (11)$$

$$f_x(x|y) = \frac{1}{\sigma_x\sqrt{2\pi}} \exp \left[ -\frac{(x-\bar{x})^2}{\sigma_x^2} \right] = f_x(x) \quad (12)$$

这就说明,对于二维正态分布,两个随机变量  $X$  与  $Y$  的 Pearson 相关系数  $r=0$  是与式(10)~(12)完全等价的,也就是说, $r=0$  是服从二维正态分布的两个随机变量相互独立的必要和充分条件。这个“完全等价”对不服从二维正态分布的两个随机变量则是不成立的。

进一步考察二维正态分布的条件分布,还会有新的发现。事实上,由式(8)可知, $Y$  倚  $X$  的条件均值和条件均方差分别为

$$\bar{y}_{y/x}(x) = \bar{y} + r \frac{\sigma_y}{\sigma_x}(x - \bar{x}) \quad (13)$$

$$\sigma_{y/x} = \sigma_y \sqrt{1 - r^2} \quad (14)$$

由式(9)可知  $X$  倚  $Y$  的条件均值和条件方差分别为

$$\bar{x}_{x/y}(y) = \bar{x} + r \frac{\sigma_x}{\sigma_y}(y - \bar{y}) \quad (15)$$

$$\sigma_{x/y} = \sigma_x \sqrt{1 - r^2} \quad (16)$$

从式(13)~(16)不难发现:①无论是  $Y$  倚  $X$  的条件均值,还是  $X$  倚  $Y$  的条件均值并非不变,而是相应的自变量的线性函数,这就是说,对于二维正态分布,两条条件均值的轨迹线即回归线分别为以  $r \frac{\sigma_y}{\sigma_x}$  和以  $r \frac{\sigma_x}{\sigma_y}$  为斜率的直线;②无论是  $Y$  倚  $X$  的条件均方差,还是  $X$  倚  $Y$  的均方差,都是常数,其值仅随 Pearson 相关系数  $r$  而变:当  $r=0$  即  $X$  与  $Y$  相互独立时,条件方差等于边际分布的方差;当  $r=1$  即  $X$  与  $Y$  为因果函数关系时,条件方差为 0,回归线就成为  $X$  与  $Y$  的因果函数关系表达式。

由上述可以推论或证明,服从二维正态分布的两个随机变量的相关关系具有如下特点:①斜率分别为  $r \frac{\sigma_y}{\sigma_x}$  和  $r \frac{\sigma_x}{\sigma_y}$  的两条回归直线围绕点  $(\bar{x}, \bar{y})$  的转动与  $r$  有关(图 1),当  $r=0$  时,两条回归直线通过点  $(\bar{x}, \bar{y})$  而分别平行于坐标轴  $x$  和坐标轴  $y$ 。②条件均方差不随自变量而变,但与  $r$  有关(图 2),当  $r=0$  时,条件方差即为相应的边际分布的方差;当  $r=1$ ,条件方差为 0,表明对其中一个随机变量的每个现

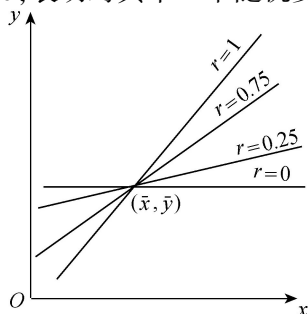


图 1 二维正态分布不同  $r$  的  $Y$  倚  $X$  的回归线

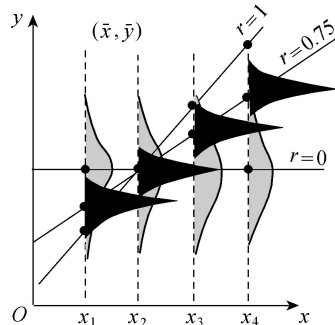


图 2 二维正态分布不同  $r$  的条件密度函数

实,另一个随机变量只能有一个值与此对应,这正是两个随机变量是因果函数关系的表征。③正态相关是一种回归为相关性主要部分的相关,对于这种相关,只要设法消除回归的影响,两个随机变量就可转换成互相独立的变量。④正态回归一定是线性回归,反之,若两个随机变量的回归关系为线性关系,则它们必服从二维正态分布。⑤最小二乘法是确定两个随机变量回归方程的常用方法。这种方法的数理统计学解释是对于两个随机变量中的一个随机变量的每个现实,另一个随机变量必将以其均值与之一一对应。因此,当两个随机变量服从二维正态分布时,最小二乘法得到的回归线必然是线性的。⑥用最小二乘法确定服从二维正态分布的两个随机变量的回归方程时,若采用随机变量  $Y$  的离差平方和最小作为目标,则得  $Y$  倚  $X$  的回归方程;若采用随机  $X$  的离差平方和最小作目标,则  $X$  倚  $Y$  的回归方程。但若采用  $Y$  的离差与  $X$  的离差之乘积和最小作为目标,则将得到  $Y$  与  $X$  的同频率相关方程

$$y_p = \bar{y} + \frac{\sigma_y}{\sigma_x}(x_p - \bar{x}) \quad (17)$$

式中: $y_p$  和  $x_p$  分别为概率为  $p$  时,由两个随机变量分布函数得到的值。可以证明,同频率相关方程的斜率  $\frac{\sigma_y}{\sigma_x}$  就是  $Y$  倚  $X$  和  $X$  倚  $Y$  两条回归线斜率之商的平方根:

$$\frac{\sigma_y}{\sigma_x} = \sqrt{\frac{r \frac{\sigma_y}{\sigma_x}}{r \frac{\sigma_x}{\sigma_y}}} \quad (18)$$

式(17)的意义在于,若两个随机变量服从二维正态分布,则其同频率相关线是一条通过点  $(\bar{x}, \bar{y})$ 、斜率为  $\sigma_y/\sigma_x$  的直线,也就是式(13)和式(15)在  $r=1$  时的结果。这就表明,两个均服从正态分布的随机变量,当为函数关系时,其一定是斜率为  $\frac{\sigma_y}{\sigma_x}$  的线性关系。

### 3 Copula 函数理论和方法

若两个随机变量均服从正态分布,则其二维分布即为式(5)。若两个随机变量中只有一个为正态分布或者两个均不为正态分布,则其二维分布就不能用式(5)表达,在这种情况下,将如何寻找其二维分布呢?显然,寻求两个随机变量的二维分布函数一般要比寻求一维随机变量分布函数困难得多,正因为如此,在水文学中二维分布的研究相对薄弱。本节和下一节仅对确定任意两个随机变量二维分布的 Copula 函数和形变函数的理论和方法进行讨论。

Copula 函数的起源可追溯到 1959 年<sup>[4]</sup>,是年,Sklan 指出:可以将任意一个  $n$  维分布函数分解为  $n$  个边际分布和一个 Copula 函数,其中边际分布描述每个随机变量的一维分布函数,Copula 函数则描述这些随机变量之间的相关性。因此,Copula 函数是一个将多个随机变量的一维分布“连接”成为多维分布的函数,顾名思义,可将 Copula 函数译作“连接函数”。Sklan 这一基本思想是以定理的形式公布于世的,以构建二维分布为例就是:令  $H$  为具有边际分布  $F$  和  $G$  的两个随机变量的二维分布,那么将存一个 Copula 函数  $C$ ,使得

$$H(x, y) = C[F(x), G(y)] \quad (19)$$

在式(19)中,若  $F$  和  $G$  是连续的,则 Copula 函数  $C$  将是唯一的。根据这一定理,可以得到如下推论:若  $H$  为具有边际分布为  $F$  和  $G$  的两个随机变量的二维分布函数, $C$  为其 Copula 函数, $F^{-1}$  和  $G^{-1}$  分别为  $F$  和  $G$  的反函数,则对于  $C$  的定义域  $I^2$  即  $[0, 1]^2$  内的任意  $(u, v)$ ,有

$$C(u, v) = H[F^{-1}(u), G^{-1}(v)] \quad (20)$$

上述 Sklan 定理及其推论显然表明,在两个随机变量的二维分布未知时,将可以通过边际分布和 Copula 函数来构建,而在二维分布已知时又可以利用边际分布的反函数求出相应的 Copula 函数。笔者认为,Copula 函数理论提出的意义不仅在于可以通过寻找 Copula 函数,继而构建出二维分布,而且在于能够揭示出隐含在二维分布中过去未曾被发现的 Copula 函数及其所描述的相关性质。

现有的文献表明,根据生成元的不同,Copula 函数可分为椭圆型、Archimede 型、二次型、极值型等类型<sup>[5]</sup>。其中 Archimede 型 Copula 函数,由于构造方便,使用容易,已得到较为广泛的应用,它又有 3 种具体型式:

a. Gumbel-Hougaard Copula 函数,公式为

$$C(u, v) = u + v + \exp \left\{ - \left[ (-\ln(1-u))^{\theta} + \right. \right.$$

$$\left. \left. (-\ln(1-v))^{\theta} \right]^{1/\theta} \right\} - 1 \quad (21)$$

式中: $u=F(x); v=G(y); \theta$  为 Copuar 参数,  $\theta \geq 1$ 。当  $\theta=1$  时,  $u$  与  $v$  相互独立,当  $\theta \rightarrow \infty$  时,  $u$  与  $v$  为函数关系。由于两个随机变量均为较大值时变化敏感,故式(21)能较好地描述具有上尾相关特性的两个随机变量之间的相关性。

b. Clayton Copular 函数,公式为

$$C(u, v) = u + v + [(1-u)^{-\theta} + (1-v)^{-\theta}]^{-1/\theta} - 1 \quad (22)$$

式中:符号意义同前述。当  $\theta \rightarrow 0$  时,  $u$  与  $v$  相应独立;当  $\theta \rightarrow \infty$  时,  $u$  与  $v$  为函数关系。由于两个随机变量均为较小值时变化敏感,故式(22)能较好地描述具有下尾相关特性的两个随机变量之间的相关性。

c. Frank Copula 函数,公式为

$$C(u, v) = u + v - \frac{1}{\theta} \ln \left[ 1 + \frac{(e^{-\theta(1-u)} - 1)(e^{-\theta(1-v)} - 1)}{(e^{-\theta} - 1)} \right] - 1 \quad (23)$$

式中:符号意义同前述。当  $\theta > 0$  时,  $u$  与  $v$  为正相关;当  $\theta < 0$  时,  $u$  与  $v$  为负相关;当  $\theta \rightarrow 0$  时,  $u$  与  $v$  相互独立。由于两个随机变量无论较大值还是较小值变化均不敏感,故式(23)难以快速捕捉到两者相关性的尾部变化。

以上 3 种常用的 Copula 函数中均包含有参数  $\theta$ ,在数学上现已研究出了一些确定  $\theta$  值的途径和方法,其中以根据 Kendall 秩次相关系数  $\tau$  与  $\theta$  之间的关系确定  $\theta$  值最为常见。对于 Archimede 型 Copula 函数,其参数  $\theta$  与 Kendall 秩相关系数  $\tau$  之间的关系列于表 1。

表 1 Archimede 型 Copula 函数的参数  $\theta$  与 Kendall 秩相关系数  $\tau$  的关系

Copula 函数名称	$\theta$ 与 $\tau$ 的关系
Gumbel-Hougaard	$\tau = 1 - \frac{1}{\theta}$
Clayton	$\tau = \frac{\theta}{\theta + 2}$
Frank	$\tau = 1 + \frac{4}{\theta} \left( \frac{1}{\theta} \int_0^{\theta} \frac{t}{e^t - 1} dt - 1 \right)$

Copula 函数理论和方法在快速发展的金融业刺激下,已有了长足的进展,水文学中使用它仅仅是近十多年来的事。

### 4 形变函数理论和方法

早在 1923 年, Narumi 就指出,在估计二维样本的联合分布即二维分布时应当考虑二维分布函数的两个最重要的数字特征:两个随机变量的回归线和条件方差<sup>[6]</sup>。前者描写了倚变量条件均值随另一

随机变化的每个现实的变化;后者可看出倚变量的条件方差随另一随机变量的每个现实的变化。嗣后,1934年别伦斯谦、1954年萨尔马诺夫、1951年可历克赛也夫<sup>[7]</sup>分别根据二维分布这两个重要数字特征先后提出了刚性相关、弹性相关和挠曲相关等概念,从而丰富了 Narumi 的学术思想。

刚性相关是指倚变量的条件均值随另一个随机变量的每个现实而变,而条件均方差则保持不变的相关。这里回归线可为线性,也可为非线性。刚性相关的两个随机变量相关散点图如图3所示。弹性相关是指倚变量的条件均值不随另一个随机变量的每个现实而变,为常数,但倚变量的条件方差却随另一个随机变量的每个现实而变,并在引进一个变形函数后则不随另一个随机变量的每个现实而变的相关。弹性相关的两个随机变量散点图如图4所示。挠曲相关是指虽然倚变量的条件均值和条件均方差随另一个随机变量的每个现实而变,但在引进一个变形函数后可以使条件均值和条件均方差都不再随另一个随机变量的每个现实而变的相关。挠曲相关的两个随机变量相关散点图如图5所示。不难看出,刚性相关和弹性相关都是挠曲相关的特例。这3种相关虽不能盖全,但由于抓住了二维分布中条件均值和条件均方差两个最主要的数字特征的变化特点,已能适用于许多情况了,因此,若能解决这3种相关的二维分布构建问题,则就能基本上满足水文学中构建二维分布的需要了。

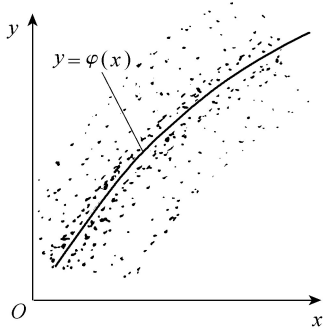


图3 刚性相关散点分布

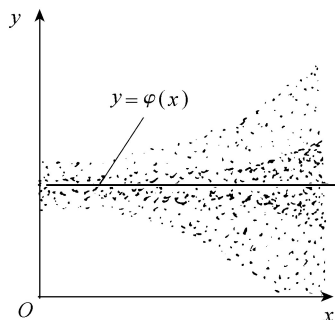


图4 弹性相关散点分布

利用形变函数构建两个随机变量二维分布的基

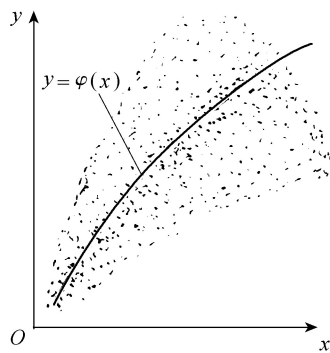


图5 挠曲相关散点分布

本思想是;首先根据两个随机变量  $X$  与  $Y$  的相关散点图的点据分布特点识别相关类型;然后将  $X$  和  $Y$  的现实  $x$  和  $y$  经由形变函数变换成新的变量  $u$  和  $v$ , 以达到消除原随机变量  $X$  与  $Y$  之间的相关性的目的。因为两个新随机变量  $U$  和  $V$  相互独立,故可得  $U$  和  $V$  的二维密度函数和分布函数分别为  $f(u, v)$  和  $F(u, v)$ ;最后通过变换再由求得的  $f(u, v)$  和  $F(u, v)$  分别得到原随机变量  $X$  与  $Y$  的二维密度函数和分布函数  $f(x, y)$  和  $F(x, y)$ 。

对于刚性相关,通过下列变换就可将原随机变量  $X$  和  $Y$  转变成两个相互独立的新随机变量  $U$  和  $V$ :

$$u = x \quad (24)$$

$$v = y - \varphi(x) = y \left( 1 - \frac{\varphi(x)}{y} \right) \quad (25)$$

式中: $\varphi(x)$ 为  $Y$  倚  $X$  的回归方程; $(1-\varphi(x)/y)$ 为刚性形变函数。

对于弹性相关,通过下列变换就可将原随机变量  $X$  和  $Y$  转变成两个相互独立的新随机变量  $U$  和  $V$

$$u = x \quad (26)$$

$$v = y\lambda(x) \quad (27)$$

式中: $\lambda(x)$ 为弹性形变函数。

一般地,对于挠曲相关,则通过变换:

$$u = x \quad (28)$$

$$v = \lambda(x) [y - \varphi(x)] = y\lambda(x) \left[ 1 - \frac{\varphi(x)}{y} \right] \quad (29)$$

就可将原随机  $X$  和  $Y$  转变成两个相互独立的新随机变量  $U$  和  $V$ 。式(29)中之  $\lambda(x) [1-\varphi(x)/y]$  称为挠曲形变函数。

由上述可知,根据形变函数理论构建刚性相关、弹性相关和挠曲相关的二维分布需要解决的问题有:寻找合适的弹性形变函数、检验新随机变量  $U$  和  $V$  的独立性、导出原随机变量与新随机变量的二维分布函数或密度函数的数学关系等。寻找合适的形变函数,至今尚无理论方法,一般只能根据相关散

点图的点据分布特点,用经验试错法确定。本文仅就后两个问题做进一步讨论。

在概率论中,检验两个随机变量之间独立性的最严格方法是它们的二维分布函数等于两个边际分布函数的乘积,或者是它们的二维密度函数等于两个边际密度函数的乘积。若对两个具有相关关系的随机变量  $X$  和  $Y$  已经获得了  $n$  个二维现实:  $(x_1, x_1)$ 、 $(x_2, x_2)$ 、 $(x_3, x_3)$ 、 $\dots$ 、 $(x_n, x_n)$ , 这  $n$  个二维现实实际上就是一个来自其总体的二维样本。按照数理统计理论,利用这个二维样本可以对总体的二维分布函数和两个边际分布作出估计,事实上有  $F(x_i, y_i) = P\{x \geq x_i \cap y \geq y_i\}$ ,  $F(x_i) = P\{X \geq x_i\}$ ,  $F(y_i) = P\{Y \geq y_i\}$  ( $i=1, 2, \dots, n$ )。因此,如果

$$P\{X \geq x_i \cap Y \geq y_i\} = P\{X \geq x_i\} \cdot P\{Y \geq y_i\} \quad (30)$$

那么  $X$  与  $Y$  将是相互独立的。图 6 是利用式(30)检验两个随机变量独立性的一个实例,图中点据“ $\times$ ”为原变量的计算结果,点据“ $\bullet$ ”则为由形变函数转换成新变量的计算结果。不难看出,对于两个具有相关性的随机变量,引入适当的形变函数可使它们的相关性减弱,甚至消除,从而使两个新随机变量相互独立。

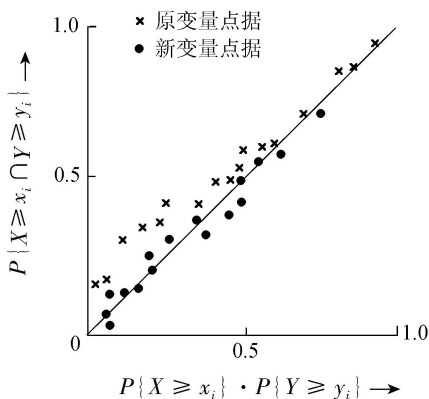


图 6 独立性检验

为了导出原随机变量与新随变量二维分布函数之间的数学关系,只需利用重积分知识,即有

$$F(x, y) = \int_x^\infty \int_y^\infty f(x, y) dx dy = \int_u^\infty \int_v^\infty f(u, v) |J| du dv \quad (31)$$

式中: $J$ 为雅可比行列式,其计算式为

$$J = \frac{\partial(x, y)}{\partial(u, v)} = \begin{vmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{vmatrix} \quad (32)$$

例如:对于刚性相关,可以通过式(24)和式(25)表达的变换来使新随机变量相互独立的,因此其雅可比行列式应为

$$J = \begin{vmatrix} 1 & 0 \\ \varphi'(x) & 1 \end{vmatrix} = 1 \quad (33)$$

这就表明,对于刚性相关,式(31)变为

$$F(x, y) = \int_x^\infty \int_y^\infty f(x, y) dx dy = \int_u^\infty \int_v^\infty f(u, v) du dv$$

因为已证明  $U$  与  $V$  相互独立,故上式变为

$$F(x, y) = \int_u^\infty \int_v^\infty f(u) \cdot f(v) du dv = \int_u^\infty f(u) du \int_v^\infty f(v) dv = F(u) \cdot F(v) \quad (34)$$

同理可得弹性相关和挠曲相关情况下原随机变量与新随机变量二维分布函数之间的数学关系。

## 5 在水文学中的应用

在求解众多的水文学科学和应用问题时,常常会遇到求解边际分布、条件分布、复杂事件概率、随机变数函数的分布等问题。求边际分布指的是在具有相关性的两个变量中,由一个变量  $X$  的分布函数推求另一个变量  $Y$  的分布函数。由概率论知,这个问题可表达为

$$F_y(y) = \int_0^1 F_x(x | y) dF_x(x) \quad (35)$$

求条件分布指的是:在具有相关性的两个随机变量中,当其中一个  $X$  取现实  $x$  时求另一个  $Y$  的分布函数。由概率论知,这个问题可表达为

$$F_y(y | x) = \frac{F(x, y)}{F_x(x)} \quad (36)$$

求复杂事件概率指的是推求包括有两个或两个以上随机变量的复杂随机事件的概率。由概率论知,构成复杂事件有“或”和“交”两种基本类型。因此,若复杂事件仅涉及两个随机变量,则其“或”和“交”的概率分别为

$$\begin{aligned} P\{X \geq x \cup Y \geq y\} &= P\{X \geq x\} + P\{Y \geq y\} - \\ P\{X \geq x \cap Y \geq y\} &= F_x(x) + F_y(y) - F(x, y) \end{aligned} \quad (37)$$

$$P\{X \geq x \cap Y \geq y\} = F(x, y) \quad (38)$$

求随机变量函数的分布函数指的是,当随机变量  $Z$  是另外一些随机变量  $X_1, X_2, \dots$  的函数即  $Z = g(X_1, X_2, \dots)$  时,通过  $X_1, X_2, \dots$  的联合分布函数推求  $Z$  的分布函数。由概率论知,有

$$F_z(z) = \int_{\Omega: g(X_1, X_2, \dots) \geq z} f(x_1, x_2, \dots) dx_1 dx_2 \dots \quad (39)$$

若  $Z$  仅是两上随机变量  $X$  和  $Y$  的函数即  $Z = g(X, Y)$ , 则式(39)变为

$$F_z(z) = \int_{\Omega: g(X, Y) \geq z} f(x, y) dx dy \quad (40)$$

式中: $\Omega$ 为积分域。

由式(35)~(40)容易看出,无论是求边际分布和条件分布,还是求复杂事件概率和随机变量函数的分布,都要涉及二维或多维分布问题。

在水文学中,资料的插补展延属于求边际分布问题<sup>[6]</sup>。概率水文预报属于求条件分布问题<sup>[7]</sup>。非一致性样本频率分析,有的属于求复杂事件概率问题,有的则属于求随机变量函数的分布函数问题。水工程风险率<sup>[8]</sup>、设计洪水<sup>[9-10]</sup>、干支流洪水和洪与潮遭遇组合<sup>[9,11]</sup>、地貌瞬时单位线<sup>[12-14]</sup>等一般均属于求随机变量函数的分布函数问题。因其中大多数问题可在现有的文献中找到,故本文仅对二维分布在非一致性样本频率分析中的应用作具体讨论。

用数理统计理论和方法确定随机变量分布函数的思路,是通过分析样本的统计规律来推断总体的统计规律。因此前提必然是样本必须来自同一总体,如果样本不完全来自同一整体,那么这个样本就是非一致性样本,这种不一致性样本不加区别地放在一个样本中显然是不能反映总体的统计规律的。从物理成因可知,一个样本之所以不一致,可能是形成机理上的差异,也可是受到了外因,如人类活动的干扰。由后一个原因导致的样本不一致性及其改正方法已有许多文献讨论过<sup>[15]</sup>,而由前一个原因导致的样本不一致性及改正,笔者发现有一种错误的观点正在流行<sup>[16]</sup>。这种错误观点认为若样本中有来自不同总体的两种信息,则其总体分布函数  $F(z)$  是这两种信息所对应的分布函数  $F_1(x)$  和  $F_2(y)$  分别以  $\alpha$  和  $(1-\alpha)$  为权重的加权平均即  $F(z) = \alpha F_1(x) + (1-\alpha) F_2(y)$ 。现以某站降雨频率分析为例来说明其错误所在。由分析得知该站年最大一日雨量可能出现在梅雨季,也可能出现在台风季。也就是说,该站年最大一日雨量可能是由梅雨和台风两种天气系统形成的。如果不分形成机理而将所得年最大一日雨量系列作为样本,那么这个样本将不具备一致性。在这种情况下,正确的思维应是先分别从梅雨季和台风季中各选取最大一日雨量样本,在求得这两个样本的分布函数  $F_1(x)$  和  $F_2(y)$  后,再按式(4)求得该站年最大一日雨量的分布函数  $F(z)$ 。因为

$$\{Z \geq z\} = \{X \geq z \cup Y \geq z\}$$

$$\text{所以 } P\{Z \geq z\} = P\{X \geq z\} + P\{Y \geq z\} - P\{X \geq z \cap Y \geq z\}$$

$$\text{即 } F(z) = F_1(z) + F_2(z) - F(z, z) \quad (41)$$

如果欲求该站年降雨量分布函数,那么由于年降雨量  $Z$  为梅雨季雨量  $X$  和台风季雨量  $Y$  之和,即  $Z = X + Y$ ,而梅雨雨量和台风雨量的形成机理不同,正确的思维应是先分别建立梅雨季雨量样本和台风季雨量样本,在求得这两个样本的分布函数  $F(x)$  和

$F(y)$ 后,再按下式求得年降雨量的分布函数:

$$F(z) = \iint_{\Omega: X+Y \geq z} f(x, y) dx dy \quad (42)$$

## 6 结 语

水文现象是十分复杂的,这不仅表现为其形成机理和时空变化十分复杂,而且表现为变量之间的关系十分复杂。对有些水文问题的解决,一维分布理论和方法已不能适应,而有待引入多维分布理论和方法。多维密度函数或多维分布函数是多维变量之间关系的最完整描述。揭示水文现象有关变量之间的关系,寻求多维分布函数,用于解决有关水文学问题已成为水文学的重要研究内容。

近一个世纪以来,无论是数学,还是水文学,对二维分布的研究都有了一些进步。在数学上提出了由两个边际分布,通过寻找连结函数构建二维分布的 Copula 函数理论和方法。在水文学上则发展了根据两个随机变量相关散点图的特点,通过引入形变函数构建二维分布的形变函数理论和方法。这两种理论和方法,各有千秋,如能深入研究,也许会碰撞出一些新的火花。

迄今为止,二维分布在资料插补展延、概率水文预报、非一致性频率分析、水工程风险率、设计洪水、干支流洪水及洪与潮遭遇组合、地貌瞬时单位线等水文学问题中得到了应用。笔者将二维分布处理以上问题归纳为三类:一是直接应用二维分布性质,如资料系列插补展延、概率水文预报等问题;二是通过分析事件而应用二维分布,如水工程风险率等问题;三是通过建立功能函数而应用二维分布,如设计洪水、干支流洪水和洪与潮遭遇组合、地貌瞬时单位线等。当然也有一些水文学问题涉及以上三类中之二,例如非一致性样本频率分析。正确应用二维分布的性质,正确分析事件之关系,以及正确选择和建立功能函数,就成为二维分布由理论通向应用的桥梁。

在水文观测年限不长,水文资料还不够丰富时,二维分布的使用必然受到很大的限制,因此在半个世纪前谈论二维分布在水文学中应用似乎过于超前,但现在面临的是信息爆炸时代,不失时机地将二维分布的研究提上议事日程,也许是当代水文学者的历史责任。

## 参考文献:

- [1] 复旦大学数学系. 概率论与数理统计[M]. 上海:上海科学技术出版社,1960.
- [2] 金光炎. 水文统计理论与实践[M]. 南京:东南大学出版社,2012.

(下转第 65 页)