

2変数MEP分布とその特性に関する研究

Study on Bivariate MEP Distribution and its Characteristics

信州大学工学部 正員 寒川典昭

信州大学工学部 正員 荒木正夫

信州大学大学院 学生員 寺島彰

1. まえがき

相関のある2変数を同時に統計解析するためには2変数分布が必要であるが、従来の研究は主として、2変数正規分布と2変数ガンマ分布に立脚したものであった。2変数ガンマ分布は、対称形しか評価できない2変数正規分布の欠点を是正し、広く非対称分布に対する適用性をもち、応用面において有効な分布とされているが、水文量はしばしば小標本であるため、いかなる母集団を採用すべきかという問題に多くの不確定さが存在する。

著者らは、このような問題に対処するために、不確定さの尺度であるエントロピーの概念を導入した1変数MEP分布を検討してきた。本研究は、これを2変数に拡張し、モーメントによる制約のもとで相互エントロピーを最大にする分布、すなわち、2変数MEP分布を開発し、その特性を検討するものである。これは、最大エントロピー原理に基づいた分布であり、制約条件の与え方によって種々の確率分布を持ち得るため、ピアソン系の分布等と同レベルのものであるとされる。又、エントロピー最大という目的関数により、できるだけ客観的な立場から分布形を求めていることは、自然現象の評価において妥当な立場を与えている。以下、2でMEP分布からの既存の分布の導出と、確率分布としてのこの分布の位置付けを、3でMEP分布のパラメータ同定法を、4で適合度の検討と感度分析を、5で実測データへの適用を述べることにする。

2. 既存分布の導出とMEP分布の位置付け

2.1 既存分布の導出

確率変数が x 、 y の同時確率密度関数を $p(x, y)$ とすると、相互エントロピー $H(X, Y)$ は次式で与えられる。

$$H(X, Y) = - \iint p(x, y) \ln p(x, y) dx dy \quad (1)$$

一方、確率密度関数が具備すべき条件と任意関数 $g_r(x, y)$ の期待値は、以下のように表現される。

$$\iint p(x, y) dx dy = 1 \quad (2)$$

$$\iint g_r(x, y) p(x, y) dx dy = E[g_r(x, y)], \quad r=1, 2, \dots, M \quad (3)$$

ここで、 $E[\cdot]$ は期待値記号である。

さて、(2)、(3)式を制約条件として(1)式を最大にする $p(x, y)$ を求めることは、母集団から得られる情報は $g_r(x, y)$ の期待値の形で取り入れ、それ以外はできるだけ一様となる $p(x, y)$ を求めていることになる。換言すると、このことは、情報の与え方には主観が入るが、それ以外はできるだけ客観的な分布、すなわち、相互エントロピーを最大にする分布を求めていることになる。我々は、水文事象のようなデータ数の少ない自然現象の説明にこのような確率分布を採用することは、自然であり妥当であると考える。

そこで、上記の問題をラグランジュの未定乗数法で解く。ラグランジュ関数は、

$$L = H(X, Y) + (\lambda_0 - 1) \{1 - \iint p(x, y) dx dy\} + \sum_{r=1}^M \lambda_r \{E[g_r(x, y)] - \iint g_r(x, y) p(x, y) dx dy\} \quad (4)$$

となり、上式を $p(x, y)$ に関して偏微分して零とおくことにより、最大エントロピー推定値は次のようになる。

$$p(x, y) = \exp \left\{ -\lambda_0 - \sum_{r=1}^M \lambda_r g_r(x, y) \right\} \quad (5)$$

ただし、 λ_r はラグランジュ乗数である。一般に、統計学においてよく知られている分布形のいくつかは、適切な $g_r(x, y)$ を与えることによって(5)式で表現される。

a) 2変数一様分布

$x[0, x_0], y[0, y_0]$ とし、 $g_r(x, y) = 0$ とおくと(5)式は、

$$p(x, y) = 1 / (x_0 y_0) \quad (6)$$

となり、2変数一様分布が得られる。ここで、 x_0, y_0 は任意定数である。

b) 2変数正規分布

$x(-\infty, \infty), y(-\infty, \infty)$ において(3)式を

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)^2 p(x, y) dx dy = \sigma_x^2 \quad (7)$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y - \mu_y)^2 p(x, y) dx dy = \sigma_y^2 \quad (8)$$

$$\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (x - \mu_x)(y - \mu_y) p(x, y) dx dy = \rho \sigma_x \sigma_y \quad (9)$$

とおくと(5)式は、

$$p(x, y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2(1-\rho^2)}\left\{\frac{(x-\mu_x)^2}{\sigma_x^2} - 2\rho\frac{(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y} + \frac{(y-\mu_y)^2}{\sigma_y^2}\right\}\right] \quad (10)$$

となり、 x, y について、平均 μ_x, μ_y 分散 σ_x^2, σ_y^2 相関係数 ρ の2変数正規分布が得られる。

c) 2変数指数分布(相関が零の場合)

$x[0, \infty), y[0, \infty)$ において(3)式を

$$\int_0^{\infty} \int_0^{\infty} x p(x, y) dx dy = \sigma_x \quad (11)$$

$$\int_0^{\infty} \int_0^{\infty} y p(x, y) dx dy = \sigma_y \quad (12)$$

とおくと(5)式は、

$$p(x, y) = \frac{1}{\sigma_x \sigma_y} \exp\left(-\frac{x}{\sigma_x} - \frac{y}{\sigma_y}\right) \quad (13)$$

となり、 x, y について平均 σ_x, σ_y 相関係数零の2変数指数分布が得られる。

2.2 M E P分布の位置付け

ピアソンは、統計学に出てくる確率密度関数の多くは微分方程式

$$\frac{1}{p(x)} \frac{dp(x)}{dx} = \frac{a+x}{b_0+b_1x+b_2x^2}$$

の解であることを示した。ここで $p(x)$ は確率密度関数であり、 a, b_0, b_1, b_2 は定数である。 $p(x)$ は定数相互間の関係により13の形に分類され、それらはピアソン系の分布と呼ばれている。⁷⁾一方、我々は、(2), (3)式の制約の下で(1)式を最大にする分布を2変数M E P分布と名付けた。これは $g_r(x, y)$ の与え方によって種々の分布形をとることができ、2.1では3つの分布の導出を示した。両者を比較すると上述の微分方程式は(1)式に、定数相互間の関係は $g_r(x, y)$ に相当すると考えられる。このことは、M E P分布はピアソン系の分布と同レベルで議論されるべきものであり、 $g_r(x, y)$ を指定した分布がピアソン k 型分布($k=1, 2, \dots, 12$)に対応されるべきものであることを示している。

3. 理論式の算定

3.1 M E P分布の算定

$x[0, \infty), y[0, \infty)$ とし、我々は母集団に関して知り得る情報、すなわち(3)式を

$$\int_0^{\infty} \int_0^{\infty} x^a p(x, y) dx dy = a \mu_x, \quad a=1, 2, \dots, N_a \quad (14)$$

$$\int_0^{\infty} \int_0^{\infty} y^b p(x, y) dx dy = b \mu_y, \quad b=1, 2, \dots, N_b \quad (15)$$

$$\int_0^{\infty} \int_0^{\infty} x^c y^d p(x, y) dx dy = c d \mu_{xy}, \quad c=1, 2, \dots, N_c, d=1, 2, \dots, N_d \quad (16)$$

とおくと、(5)式は

$$\hat{p}(x, y) = \exp\left(-\alpha - \sum_{a=1}^{N_a} \beta_a x^a - \sum_{b=1}^{N_b} \gamma_b y^b - \sum_{c=1}^{N_c} \sum_{d=1}^{N_d} \delta_{cd} x^c y^d\right) \quad (17)$$

となる。ここで、 $\alpha, \beta_a, \gamma_b, \delta_{cd}$ はラグランジュ乗数である。いま、 $\beta_a, \gamma_b, \delta_{cd}$ の近似値を、 $\beta_a^0, \gamma_b^0, \delta_{cd}^0$ 残差を $\epsilon_a^{\beta}, \epsilon_b^{\gamma}, \epsilon_{cd}^{\delta}$ とおくと、

$$\beta_a = \beta_a^0 + \epsilon_a^{\beta}, \quad \gamma_b = \gamma_b^0 + \epsilon_b^{\gamma}, \quad \delta_{cd} = \delta_{cd}^0 + \epsilon_{cd}^{\delta} \quad (18)$$

となり、これらを(14)~(16)式に代入して線形化すると次の連立方程式が得られる。

$$\sum_{a=1}^{Na} (z_{i+a,0} - i^{\mu} x z_{a,0}) \epsilon_a^{\beta} + \sum_{b=1}^{Nb} (z_{i,b} - i^{\mu} x z_{0,b}) \epsilon_b^{\gamma} + \sum_{c=1}^{Nc} \sum_{d=1}^{Nd} (z_{i+c,d} - i^{\mu} x z_{c,d}) \epsilon_{cd}^{\delta} = z_{i,0} - i^{\mu} x z_{0,0} \quad i=1,2,\dots,Na \quad (19)$$

$$\sum_{a=1}^{Na} (z_{a,j} - j^{\mu} y z_{a,0}) \epsilon_a^{\beta} + \sum_{b=1}^{Nb} (z_{0,j+b} - j^{\mu} y z_{0,b}) \epsilon_b^{\gamma} + \sum_{c=1}^{Nc} \sum_{d=1}^{Nd} (z_{c,j+d} - j^{\mu} y z_{c,d}) \epsilon_{cd}^{\delta} = z_{0,j} - j^{\mu} y z_{0,0} \quad j=1,2,\dots,Nb \quad (20)$$

$$\sum_{a=1}^{Na} (z_{k+a,1} - k l^{\mu} x y z_{a,0}) \epsilon_a^{\beta} + \sum_{b=1}^{Nb} (z_{k,1+b} - k l^{\mu} x y z_{0,b}) \epsilon_b^{\gamma} + \sum_{c=1}^{Nc} \sum_{d=1}^{Nd} (z_{k+c,1+d} - k l^{\mu} x y z_{c,d}) \epsilon_{cd}^{\delta} = z_{k,1} - k l^{\mu} x y z_{0,0} \quad k=1,2,\dots,Nc, \quad l=1,2,\dots,Nd \quad (21)$$

ここで、

$$z_{i,j} = \int_0^{\infty} \int_0^{\infty} x^i y^j \exp \left(- \sum_{a=1}^{Na} \beta_a^0 x^a - \sum_{b=1}^{Nb} \gamma_b^0 y^b - \sum_{c=1}^{Nc} \sum_{d=1}^{Nd} \delta_{cd}^0 x^c y^d \right) dx dy$$

(19)~(21)式は、未知数の数と方程式の数が一致しており、解くことが可能である。具体的には次のようにしてラグランジュ乗数を評価する。

1) $\beta_a^0 = 0, \gamma_b^0 = 0, \delta_{cd}^0 = 0$ と仮定する。2) (19)~(21)の連立方程式を $\epsilon_a^{\beta}, \epsilon_b^{\gamma}, \epsilon_{cd}^{\delta}$ に関して解く。3) 求めた $\epsilon_a^{\beta}, \epsilon_b^{\gamma}, \epsilon_{cd}^{\delta}$ を(18)式に代入して $\beta_a, \gamma_b, \delta_{cd}$ を計算し、これを次の近似値とする。4) $\epsilon_a^{\beta}, \epsilon_b^{\gamma}, \epsilon_{cd}^{\delta}$ が微小となったときの $\beta_a, \gamma_b, \delta_{cd}$ を解とみなし、 α は次式から決定する。

$$\alpha = \ln \left(\int_0^{\infty} \int_0^{\infty} \exp \left(- \sum_{a=1}^{Na} \beta_a^0 x^a - \sum_{b=1}^{Nb} \gamma_b^0 y^b - \sum_{c=1}^{Nc} \sum_{d=1}^{Nd} \delta_{cd}^0 x^c y^d \right) dx dy \right) \quad (22)$$

3.2 モーメントの基準化

確率変数 X, Y の1次モーメントが1となるような変数変換、すなわち

$$u = x / 1^{\mu_x}, \quad v = y / 1^{\mu_y} \quad (23)$$

を考えると、 u, v の確率密度関数 $q(u, v)$ は

$$q(u, v) = J \left(\frac{x, y}{u, v} \right) p(x, y) \quad (24)$$

となる。ここで、 $J(\cdot)$ はヤコビアンであり、

$$J \left(\frac{x, y}{u, v} \right) = 1^{\mu_x} \cdot 1^{\mu_y} \quad (25)$$

となる。この関係を(14)~(16)式に代入して整理すると、

$$a^{\mu_x} = a^{\mu_x} / (1^{\mu_x})^a, \quad b^{\mu_y} = b^{\mu_y} / (1^{\mu_y})^b, \quad cd^{\mu_{xy}} = cd^{\mu_{xy}} / \{ (1^{\mu_x})^c (1^{\mu_y})^d \} \quad (26)$$

を得る。ここで

$$a^{\mu_x} = \int_0^{\infty} \int_0^{\infty} u^a q(u, v) du dv, \quad b^{\mu_y} = \int_0^{\infty} \int_0^{\infty} v^b q(u, v) du dv, \quad cd^{\mu_{xy}} = \int_0^{\infty} \int_0^{\infty} u^c v^d q(u, v) du dv$$

であり、これらを基準化されたモーメントと呼ぶ。以下では、モーメントはすべて基準化されている。

4. 適合度の検討と感度分析

4.1 2変数ガンマ分布に対する適合度

ここでは1例として、次式で与えられる2変数ガンマ分布を母集団と仮定する。

$$p(x, y) = \frac{1}{\Gamma(v) (\sigma_1 \sigma_2)^{\frac{v+1}{2}} (1-\rho)^{\frac{v-1}{2}}} \exp \left\{ - \frac{x}{\sigma_1 (1-\rho)} - \frac{y}{\sigma_2 (1-\rho)} \right\} (xy)^{\frac{v-1}{2}} I_{v-1} \left(\frac{2\sqrt{\rho}}{1-\rho} \sqrt{xy} \right) \quad (27)$$

2G(2,0.5)でも、2M(4,4,1,1)から全体的によく一致してきているが、ピーク付近の再現度は2G(4,0.5)にやや劣る。2G(1,0.2)では、2M(3,3,1,1)で十分よく母集団に一致している。図-1を等確率線として表現したのが図-2であり、上述の考察がより明確にされる。

4.2 モーメントの安定性

一般に実測値から得られるモーメントは、母集団のモーメントに一致せず、モーメントの次数が大きくなるに従つて真値からのずれが大きくなることが予想される。当然それはM-E-P分布にひずみを生じさせる。そこで、モーメン

表-1 適合度検定におけるラグランジュ剰数

(a) $2G(2, 0, 5)$

	$2M(2, 2, 1, 1)$	$2M(4, 4, 1, 1)$	$2M(4, 4, 2, 2)$
α	0.57084	1.6852	1.2734
$\beta_1 (=y_1)$	0.14688	-2.0921	-1.5278
$\beta_2 (=y_2)$	0.77881	2.8953	3.2772
$\beta_3 (=y_3)$	—	-0.70465	-0.89304
$\beta_4 (=y_4)$	—	0.06828	0.077565
δ_{11}	-1.1866	-1.0480	-2.9656
δ_{12}	—	—	0.40641
δ_{21}	—	—	0.40641
δ_{22}	—	—	-0.067129

(b) $2G(4, 0.5)$

	2M(2,2,1,1)	2M(4,4,1,1)	2M(4,4,2,2)
α	2.1283	5.3627	4.5707
$\beta_1 (= \gamma_1)$	-1.8801	-8.4525	-7.2592
$\beta_2 (= \gamma_2)$	2.3276	9.7708	10.475
$\beta_3 (= \gamma_3)$	—	-3.1215	-3.6144
$\beta_4 (= \gamma_4)$	—	0.39584	0.42859
δ_{11}	-2.6116	-2.3235	-6.3267
δ_{12}	—	—	1.0551
δ_{21}	—	—	1.0551
δ_{22}	—	—	-0.22084

(c) $2G(1, 0, 2)$

	$2M(3, 3, 1, 1)$	$2M(4, 4, 1, 1)$
α	-0.14418	-0.13749
$\beta_1 (= \gamma_1)$	1.1441	1.1308
$\beta_2 (= \gamma_2)$	-0.0029132	0.0069515
$\beta_3 (= \gamma_3)$	0.0026707	0.00050528
$\beta_4 (= \gamma_4)$	—	0.00013316
δ_{11}	-0.15045	-0.15042
δ_{12}	—	—
δ_{21}	—	—
δ_{22}	—	—

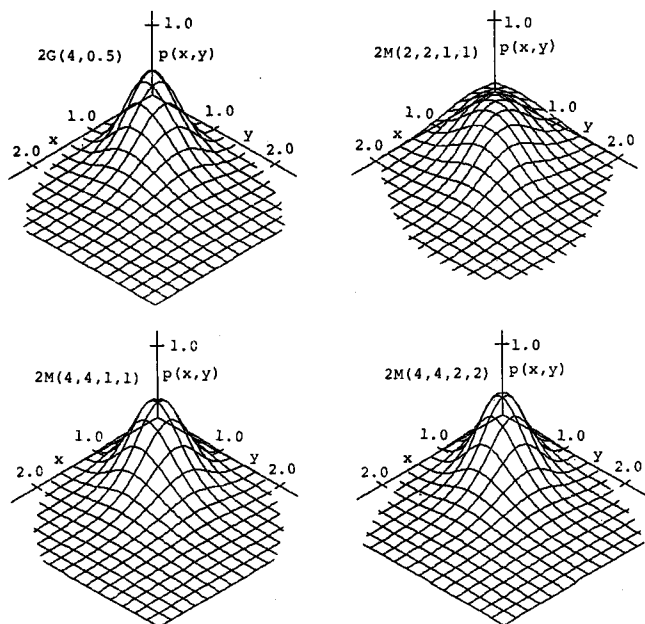
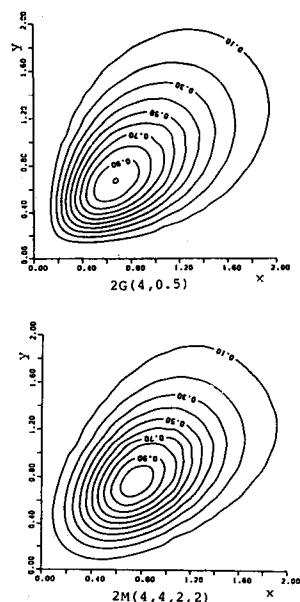


図-1 母集団と適合度分布の比較



圖一 2 等確率線圖

トがもつ誤差を検討するために、 $2G(2,0.5)$ 、 $2G(4,0.5)$ 、 $2G(1,0.2)$ を母集団として個数 $n(n=10, 20, \dots, 150)$ のデータを50個発生させ、 $1^m_x, 2^m_x, 3^m_x, 4^m_x, 11^m_{xy}, 12^m_{xy}, 21^m_{xy}, 22^m_{xy}$ の平均と標準偏差を求めた。

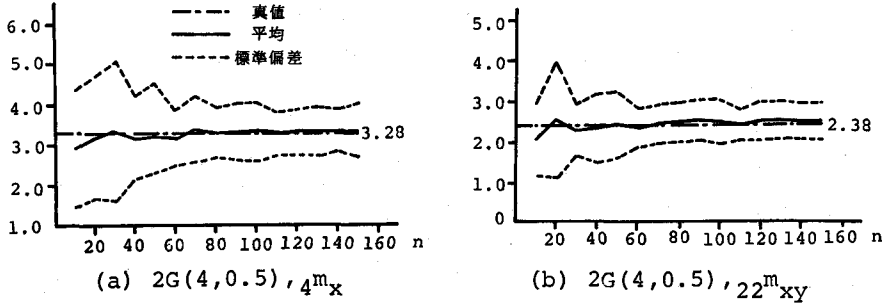


図-3 データ数の増加に伴うモーメントの安定性

図-3はその1例で、4.1で取り扱った $2G(4,0.5)$ の最高次数モーメントについて示したものである。どの母集団の場合も、モーメントの次数が小さい程、真値に対する相対的な変動が小さく、データ数の増加とともに平均値が母集団のモーメントに近づき、標準偏差が小さくなっていく様子がうかがえる。

4.3 感度分析

(1) データ数とMEP分布

ここでは、データ数、すなわちモーメントのもつ誤差がMEP分布に及ぼす影響を分析するために、 $2G(4,0.5)$ を母集団としてデータを発生させ、 $2M(2,2,1,1)$ 、 $2M(4,4,2,2)$ を同定した。

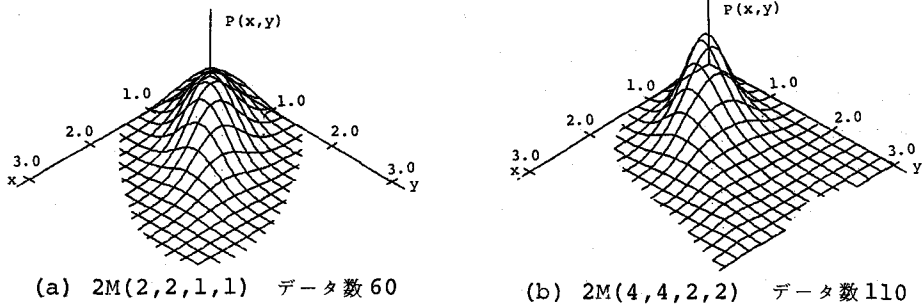


図-4 データより求めたMEP分布

図-4は、データ数60個の $2M(2,2,1,1)$ とデータ数110個の $2M(4,4,2,2)$ である。 $2M(2,2,1,1)$ ではデータ数が60個付近から、真値のモーメントを用いた分布に近づき、80個ではほぼ一致している。60個以下でも図-4(a)と大きく異なることはない。一方、 $2M(4,4,2,2)$ ではデータ数が110個付近から真値のモーメントを用いた分布に近づき、これ以上の数では真値のモーメントを用いた分布とよく一致している。又、このことは、母集団との一致度が高いことを示している。しかし、データ数が少ないところでは、高次モーメントのもつ誤差が大きくなり、MEP分布が存在しない場合もあった。

(2) 高次モーメントの変動とMEP分布

ここでは、モーメントの変動がMEP分布に及ぼす影響を分析するために、高次モーメントをMEP分布が存在する区間内で次のように変動させ、その感度を分析した。

$$2G(4,0.5), 2M(2,2,1,1) : 1.10 \leq 2^m_x \leq 19.0 \text{ (真値 } 1.25), 0.9 \leq 11^m_{xy} \leq 1.249 \text{ (1.125)} \quad (28a)$$

$$2G(4,0.5), 2M(4,4,2,2) : 3.22 \leq 4^m_x \leq 3.795 \text{ (3.2813)}, 2.35 \leq 22^m_{xy} \leq 2.495 \text{ (2.3828)} \quad (28b)$$

$$2G(1,0.2), 2M(3,3,1,1) : 4.25 \leq 3^m_x \leq 10.0 \text{ (6.0)}, 1.01 \leq 11^m_{xy} \leq 1.95 \text{ (1.2)} \quad (28c)$$

$$2G(1,0.2), 2M(4,4,1,1) : 20.5 \leq 4^m_x \leq 24.39 \text{ (24.0)}, 0.4 \leq 11^m_{xy} \leq 1.3 \text{ (1.2)} \quad (28d)$$

ただし、変動させるモーメント以外の値はすべて真値に固定している。計算の結果、次の区間内のMEP分

布は真値のモーメントを使ったMEP分布と形状的によく一致していた。

$$2G(4,0.5), 2M(2,2,1,1): 1.20 \leq m_x \leq 1.35, 1.05 \leq m_{xy} \leq 1.175 \quad (29a)$$

$$2G(4,0.5), 2M(4,4,2,2): 3.22 \leq m_x \leq 3.795, 2.35 \leq m_{xy} \leq 2.43 \quad (29b)$$

$$2G(1.0,2), 2M(3,3,1,1): 6.0 \leq m_x \leq 10.0, 1.01 \leq m_{xy} \leq 1.75 \quad (29c)$$

$$2G(1,0.2), 2M(4,4,1,1): 23.0 \leq m_x \leq 24.39, 0.8 \leq m_{xy} \leq 1.3 \quad (29d)$$

5. 実測データへの適用

適用例として、小市(犀川流域)と杭瀬下(千曲川流域)の29年分の年最大時間流量を採用した。小市、杭瀬下の下流で犀川と千曲川は合流している。ここでの適用は、データ数が少ないため $2M(2,2,1,1)$ にかぎった。図-5は計算結果であり、 $P(x,y)$ 軸は、1つのメッシュがもつ確率をメッシュ面積で除した値である。 $2M(2,2,1,1)$ は実測データの全体的な傾向をとらえているが、さらに整備されたデータへの適用を検討することが必要であろう。

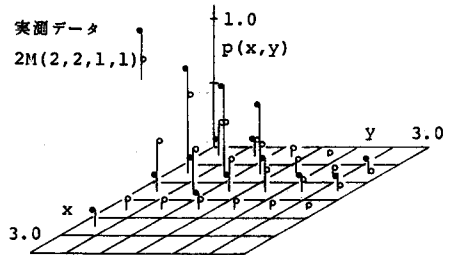


図-5 年最大日流量のMEP分布
(x:小市、y:杭瀬下)

6. あとがき

本研究では、水文量の2変数統計解析の基礎となる2変数MEP分布を開発し、MEP分布の位置付けを明確にしながらその諸特性を分析した。以下、結果を要約する。

1) 制約条件の与え方により種々のタイプのMEP分布を導出できることより、MEP分布はピアソン系の分布等と同レベルで議論されるべきであることが明らかにされた。

2) 統計モーメントを制約条件としたMEP分布のパラメタは、それを制約条件に代入して線形化した連立方程式を繰り返し計算することにより評価された。

3) ガンマ分布を母集団とした適合度結果では、 $2M(4,4,1,1)$ 程度までで十分よい再現が得られた。

4) モーメントの変動に伴うMEP分布の安定性では、数例であるが、データ数の増加及び高次モーメントの変動に対するMEP分布の適合性が明示された。

5) 実測データへの適用例として、相関のある2地点の年最大時間流量についてMEP分布が評価された。尚、本研究を行なうにあたり、計算等で信州大学大学院生佐藤健次君、学部学生刑部博君の援助を得た。記して謝辞を表する。

参 考 文 献

- 1) 長沢敏夫：複合確率と河川工事計画におけるその応用について、建設省直制工事技術報告、第14回、pp.791~796、昭和35年。
- 2) 石原安雄、長尾正志：計画高水流量の2変数確率論的研究、京都大学防災研究所年報、第6号、pp.181~188、昭和38年7月。
- 3) 栗林宗人：多次元確率と合流問題、建設省直制工事技術報告、第20回、pp.695~698、昭和41年。
- 4) 長尾正志、角屋 睦：二変数ガンマ分布とその適用に関する研究(1)、(2)、(3)、(4)、(5)、京都大学防災研究所年報、第13号B、第14号B、第14号B、第14号B、第15号B、pp.105~133、pp.43~56、pp.57~75、pp.77~85、pp.333~347。
- 5) 寒川典昭、荒木正夫：水文事象の頻度分析へのMEP導入について、土木学会論文報告集、第335号、pp.89~95、昭和58年7月。
- 6) たとえば、大泉充郎、野口正一、本多波雄：情報理論、オーム社、p78、昭和37年。
- 7) Elderton&Johnson: Systems of frequency curves, Cambridge press, p45, 1960.