ある種の Dependent Outlier を含む確率模型

——大相撲における星取表分布の統計解析——

統計数理研究所　鈴　木　義一郎

(1979年3月　受付)

Note on Some Stochastic Model with Dependent Outlier—Data Analysis of ŌZUMŌ Tournament

Giitiro Suzuki

(The Institute of Statistical Mathematics)

In this note, some analyses for the data of ŌZUMŌ (Japanese wrestling tournament) are discussed. Regarding these a new stochastic model is presented. Namely, some stochastic properties of compound binomial distribution with a dependent outlier are presented.

最近また人気をもり返した感のある大相撲は，日本の国技である．この大相撲のデータをうまく利用すれば，身近で面白い統計教育用教材が得られる．またこの種のデータを用いた報告事例としては，川上 [2] や杉原 [4] 等がある．

この小稿では，この大相撲のデータ解析事例の一端を紹介するとともに，附随して構築できたある種の確率モデルについての考察を行う．ここで提示されるモデルは，ある複合2項分布に従う変量に若干の従属性のある部分 (dependent outlier) が附加された変量の分布で説明できるものである．

1. 大相撲の勝数と複合2項分布

杉原 [4] の報告によれば，横綱と平幕力士とでは勝つ確率が異るだろうし，さらに同一力士が15日間いつも同じ力を発揮できるものでもない．従って15日間での勝ち星の数に単純な2項分布をあてはめてみても，うまくいかないことは明白である．実際 p の値が 0.5 の2項分布を906組のデータにあてはめた場合，15日間全勝した回数3は理論値 0.027 の100倍強にもなっている．全般に両端の観測値が理論値を上まわる傾向にあるために，6，7，8，9勝のケースが少なめにでてくる．表1より χ^2 の値を算出してみると 557 で，自由度が14の上側1％点 29.141 と比較してとんでもなく大きい．そこで彼は，各取組での勝率 p に正規分布を想定し，もっと適合度のよい分布をあてはめている．但し [4] では，グラフ表示だけで数値結果は報告されてない．

一般に2項分布

$$\binom{n}{x} p^x (1-x)^{n-x}$$

の p に，(0, 1) 上のある種の確率分布 $\varphi(p)$ を用いて

$$f(x) = \int_0^1 \binom{n}{x} p^x (1-p)^{n-x} \varphi(p)\, dp \tag{1}$$

で与えられる分布は，複合2項分布と呼ばれている（例えば [1] 参照）．φ のモーメントを

表 1　勝数別度数分布 ([4])

勝　数	回　数	理 論 値
0	0	0.027
1	3	0.415
2	13	2.903
3	27	12.58
4	48	37.74
5	81	83.03
6	128	138.38
7	129	177.90
8	169	177.90
9	106	138.38
10	104	83.03
11	50	37.74
12	27	12.58
13	11	2.903
14	7	0.415
15	3	0.027
計	906	905.95

(昭和26年10月～昭和29年3月の幕内と十両)

$$\mu = \int p\,\varphi(p)\,dp$$
$$\mu_l = \int (p-\mu)^l\,\varphi(p)\,dp \qquad (2)$$

と書けば (1) の表現は

$$f(x) = \binom{n}{x} \sum_{i=0}^{n-x} \binom{n-x}{i} (-1)^i \sum_{j=0}^{x+i} \binom{x+i}{j} \mu_{x+i-j}\,\mu^j \qquad (3)$$

のように表現できる．(3) 式の表現のほうが，$\varphi(p)$ の値域が (0, 1) 外であっても（例えば正規分布の場合のように）確率分布が定義されることがあるので，(1) 式よりは一般的である．

Ishii-Hayakawa [1] は，$\varphi(p)$ が

$$\varphi(p) = \frac{1}{B(a,b)}\,p^{a-1}(1-p)^{b-1} \qquad (4)$$
$$B(a,b) = \int_0^1 t^{a-1}(1-t)^{b-1}\,dt$$

なるベータ分布に従う場合を考えて，杉原 [3] のデータには $a=b=50$ としたときがよく適合することを，グラフ表示で与えている．なお φ が (4) で与えられたときの (1) 式の表現は

$$f(x) = \frac{B(x+a,\ n+b-x)}{(n+1)\,B(a,b)\,B(x+1,\ n+1-x)}$$

となり，さらに平均と分散は

$$\mu = \frac{na}{a+b}\quad,\quad \sigma^2 = \frac{nab\,(a+b+n)}{(a+b)^2\,(a+b+1)}$$

で与えられる．つまり通常の2項分布の場合より分散が大きめになる．この分布は，通例ベータ・2項分布と呼ばれるものである．

表2に最近3年間の結果を，幕内十両別に示してみた．いやに8勝力士の多いのが目につく．勝ち越し（8勝以上）力士の数を負け越した力士の数で割った値を，各場所毎にプロットしてみたのが図1である．平均勝ち数はほぼ7.5勝であるから，この種の値はほぼ1になるべきなのに，結果は勝ち越し力士のほうが圧倒的に多い．特に十両では，6割強の力士が勝ち越している勘定になる．

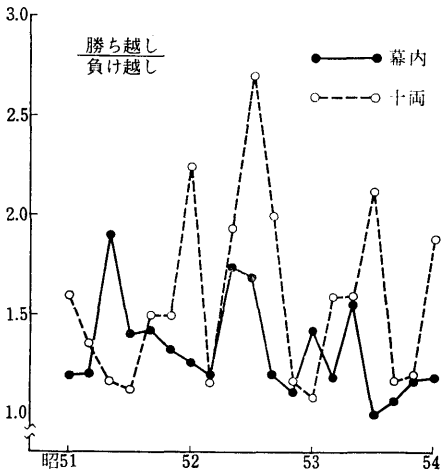

図 1

表2 大相撲星取表の分布 ([3])

	勝数	51年 1月	3	5	7	9	11	52年 1	3	5	7	9	11	53年 1	3	5	7	9	11
	0																		
	1						1									1			
	2		1				1	1		1		1		1					2
	3	1	1	1	3	3			1	1	3	2	3	1	1	3		1	
幕	4	5	2	1	2	1	2	1	4	1		1	3	4	2	2	2	2	2
	5	2	4	5	5	4	4	4	5	4	5	5	3	4	4	3	10	6	3
	6	3	6	2	3	4	5	8	4	2	4	3	4	4	5	4	3	2	5
	7	5	2	2	2	2	2	1	2	3	1	3	4	0	4	1	3	7	4
	8	9	7	10	11	8	9	8	5	11	12	6	5	8	9	11	6	9	6
	9	4	4	4	5	5	4	3	8	4	7	5	7	4	4	6	7	4	8
内	10		4	5	1	3	3	3	1	2	1	4	4	5	3		2	3	2
	11	2	1		2	2	2	2	2	2		1	1	1	1	2	1		1
	12	3	1		1	1		2	1	1						1		2	
	13	1	2	2			1	1	1	1	1		1	1	2				1
	14				1	1	1					1	1			2	1	1	
	15								1		1	1		1			1		1
	0																		
	1							1		1					1				
	2		1		1		0	1			1			1		2		2	
	3	1	2	2	2		0		2	1		1		1	2	1			
十	4	1		1		4	2				2				1	1		1	1
	5	2	1	3	2		1		2	2		3	5	1	1	2	4	3	2
	6	2	1	1	1	2	5	3	2		1	3	5	1	2	2	2	2	6
	7	4	6	5	2	4	3	3	6	8	3	1	2	8	3	2	2	4	4
	8	11	7	5	9	7	7	12	8	5	15	9	6	5	7	7	11	6	6
	9	2	6	5	3	6	5	5	3	6	2	4	7	3	6	5	1	3	3
両	10	2	1	2	4	2	2	1	2	1	1	2		3	1	2	5	4	2
	11	1	1	2	1		1			1	1	1	1	2	2	2		1	1
	12																		
	13																		1
	14								1										
	15																		

2. ある種の dependent outlier を含む確率モデル

8勝力士の数がなぜ多すぎて7勝力士が少ないか．この問題を考えるために，14日目までに丁度7勝している力士の千秋楽での勝敗を，最近3年間について調べてみた．表3の結果からもわかるように，7勝力士の千秋楽での勝率が良すぎる感じである．とにかく勝ち越せば給金が上がるし，次の場所での上位躍進が保証されている．8勝と7勝とではたったの1勝の差であっても，雲泥の相違がある．丁度7勝している力士にとっては，千秋楽での勝負に渾身の力を振りしぼる．結果として勝率が高くなるのは当然かもしれない．そこで次のようなモデルを想定してみることにしよう．

今，14日目までの勝星の数 Y の分布は $n=14$ の複合2項分布

$$g(y)=\int\binom{14}{y}p^y(1-p)^{14-y}\varphi(p)\,dp$$

表 3

	14日目で7勝している力士の数	千秋楽で勝った力士数
昭51年1月	4	3
3	6	5
5	4	4
7	9	8
9	3	2
11	7	6
52 1	3	3
3	0	0
5	5	5
7	7	6
9	6	4
11	1	1
53 1	5	5
3	4	4
5	5	5
7	2	1
9	6	4
11	6	3

に従うものとする．そして千秋楽での勝ち数 X の分布は

$$\Pr\{X=1\}=p$$
$$\Pr\{X=1\,|\,Y=7\}=p+\varepsilon$$

のように，Y の結果如何に従属するものと想定する．この2つの関係より

$$\Pr\{X=1\,|\,Y\neq 7\}=p-\frac{g(7)}{1-g(7)}\varepsilon$$

でなければならない．特に ε の値が0になる場合が，通常の独立な場合になる．つまりこの ε というパラメータは，X の Y に対する従属性の度合いを示すものである．そこでこの種の X は，Y に対するある種の "dependent outlier" と解釈することもできる．

さてこのような条件の下で，15日間での勝ち星の数

$$Z=Y+X$$

の分布 $h(z)$ はどうなるか．

$$\bar{g}(z)=[g(z)+g(z-1)]/2$$
$$\Delta(z)=g(z)-g(z-1)\qquad(g(-1)=g(15)=0)$$
$$\delta=g(7)/(1-g(7))$$

と置けば

$$h(7)=\bar{g}(7)+\delta[\Delta(7)-1]\varepsilon \tag{5}$$
$$h(8)=\bar{g}(8)+\delta[\Delta(8)+1]\varepsilon \tag{6}$$
$$h(z)=\bar{g}(z)+\delta\Delta(z)\varepsilon\quad(z\neq 7,8)$$

と表わされることが容易にわかる．さらに (5)，(6) の関係より

$$\varepsilon=\frac{h(8)-h(7)}{\delta[2(1-g(7))+g(8)+g(6)]} \tag{7}$$

という関係が得られる．$h(z)$ はデータから直接推定できるから，$g(y)$ の値が推定できさえすれば (7) より ε の値が（最良かどうかは別として）推定可能となる．

ところで

$$\begin{aligned}\mathrm{E}\{XY\}&=\sum y\Pr\{X=1,\ Y=y\}\\&=\sum yg(y)\Pr\{X=1\,|\,Y=y\}\\&=\left(p-\frac{g(7)}{1-g(7)}\varepsilon\right)\sum_{y\neq 7}yg(y)+(p+\varepsilon)7g(7)\\&=p\sum yg(y)+[7-\sum yg(y)]\delta\varepsilon\end{aligned}$$

という関係より

$$\sum yg(y)=7 \tag{8}$$

という関係が成り立てば

$$\mathrm{E}\{XY\}=\mathrm{E}\{X\}\,\mathrm{E}\{Y\}$$

つまり X と Y とは無相関になる（勿論独立ではない）．ところで $g(y)$ は14日間の勝数の分

布であったから，$g(y)$ の期待値は7と考えるのが自然で，従って (8) の関係が成立しているとみてよかろう．結局

$$\begin{aligned} \mathrm{V}\{Z\} &= \mathrm{V}\{Y\} + \mathrm{V}\{X\} \\ &= \mathrm{V}\{Y\} + p(1-p) \end{aligned}$$

$\mathrm{V}\{Z\}$ はデータより直接推定できるから，p の値を適当に想定すれば $\mathrm{V}\{Y\}$ が推定される．従って Y の分布 $g(y)$ が平均と分散によって特定できる場合，平均は既に7と与えられているので $g(y)$ が推定でき，かくて ε が推定できることになる．

さて表1のデータより平均，分散を算出してみると，それぞれ

7,598,　　5,916

である．g としてベータ・2項分布，p の値として 1/2 を考えると $\mathrm{V}\{Y\}$ の推定値 5.916−0.25 =5,666 より，g の分布としては n=14，a=b=10 のベータ・2項分布が想定できる．表4にこのようにして得られた理論値と観測値が対比されている．最後の列に示した数値は，15日間の勝数に直接ベータ・2項分布をあてはめた場合のものである．χ^2 の値でみる限りはそれほどの改良はみられないが，7勝と8勝のところではかなりあてはまりがよくなっている．このモデルでの χ^2 の値が大きくでているのは，14勝15勝あたりで単純なベータ・2項分布の場合よりほんの少し悪くなっているためである．

表5には，表2で与えられた最近の結果にこのモデルをあてはめたときの χ^2 値が示してある．各年度毎に幕内・十両別，3年間まとめた場合の幕内と十両別，そして幕内と十両を合わせた場合の計9通りである．各ケースの最初の行の数値は $\varphi(p)$ の分布のパラメータで，ベータ・2項分布の場合が a=b の値，正規・2項分布の場合が正規分布の分散の値である．2番目の行に2組ずつ並んでいる数値が χ^2 値で，左側が吾々のモデルを用いたケース，右側の数値が単純な複合2項分布をあてはめた場合のものである．特に幕内の場合が，ここで提示したモデルにより適合度がかなり改良されている（十両の場合はもっと別の観点からのモデル構築

表　4

z	$g(\cdot)$	$\bar{g}(\cdot)$	$\Delta(\cdot)$	理論値	観測値	ベータ・2項分布
0	0.0010	0.0005	0.0010	0.47	0	0.45
1	0.0061	0.0035	0.0051	3.31	3	3.13
2	0.0197	0.0129	0.0137	11.97	13	11.40
3	0.0451	0.0324	0.0254	29.90	27	28.82
4	0.0807	0.0629	0.0355	57.70	48	56.39
5	0.1189	0.0998	0.0382	91.17	81	90.22
6	0.1486	0.1337	0.0297	121.77	128	121.73
7	0.1598	0.1542	0.0112	119.72	129	140.86
8	0.1486	0.1542	−0.0112	159.72	169	140.86
9	0.1189	0.1337	−0.0297	120.56	106	121.73
10	0.0807	0.0998	−0.0382	89.62	104	90.22
11	0.0451	0.0629	−0.0355	56.27	50	56.39
12	0.0197	0.0324	−0.0254	28.87	27	28.82
13	0.0061	0.0129	−0.0137	11.42	11	11.40
14	0.0010	0.0035	−0.0051	3.10	7	3.13
15	—	0.0005	−0.0010	0.43	3	0.45
			カイ2乗値	29.95		33.68

表 5

		ベータ・2項分布		正規・2項分布
51	幕内	7.55	8	0.0155
	(207)	12.9 35.8	13.3 35.8	11.7 34.9
	十両	81.78	82	0.0015
	(153)	21.4 31.4	21.4 31.4	21.3 31.4
52	幕内	6.94	7	0.0168
	(206)	60.9 66.8	61.7 66.0	47.8 57.0
	十両	−2050.76	−2050	−0.0001
	(153)	67.5 92.1	67.5 92.1	67.5 92.1
53	幕内	7.02	7	0.0166
	(213)	70.2 72.6	69.9 70.9	56.6 62.5
	十両	21.55	22	0.0057
	(154)	23.5 31.3	23.7 31.3	23.4 31.2
51〜53	幕内	7.16	7	0.0163
	(626)	106.4 143.7	103.4 139.9	86.5 129.0
	十両	51.95	52	0.0024
	(460)	72.0 107.9	72.1 107.9	72.0 107.8
幕内と十両		11.49	11	0.0104
(1086)		169.2 232.0	162.3 228.7	155.7 221.6

が必要と思われる)．尚52年の十両の結果では，観測されたデータの分散が異常に小さすぎたために，パラメータの値が負になっている．ところが我々の場合は (1) 式の代りに (3) の表現を用いているために，このような場合でも形式的に導出したものが確率分布になっていて，適合度も他のケースとそう違ってはいない．

さらに非整数値パラメータのベータ・2項分布，整数値のベータ・2項分布，そして正規・2項分布の3つのケースについて適合度を比較してみると，それほどの差はみられない．そこで計算上の簡便さから考えて，整数値パラメータのベータ・2項分布を用いるのが妥当と思われる．この場合のモデルとデータとの適合具合いを示したのが図2，3である．十両の場合，あてはまりがそうよくないことが観察できる．

また同じ図に2組の数値が附記されているが，カッコの中の数値はデータ数，アンダーライン上の数値は ε の推定値である．総じて ε の値は，幕内の場合のほうが十両より大きめである また昭和26〜29年の場合の ε はそれほど大きくなく，そのため7勝力士の数もそう少なくはない．また最近3年間の幕内の場合のデータをみると，10〜12勝のところでモデルの線より下側に，13〜15勝のところで上側にでている．このことは，横綱陣がそろって大勝するのに反して大関以下がなかなか2ケタの勝星をあげられないといった昨今の風潮を，如実に物語っていると思われる．この傾向は昭和20年代の場合にはあてはまらないようである．

同じような計算を昭和35年からのデータにあてはめて，ベータ分布のパラメータの値と，ε の推定値とをプロットしたのが図4，図5である．図4をみると，幕内では全般に安定して小

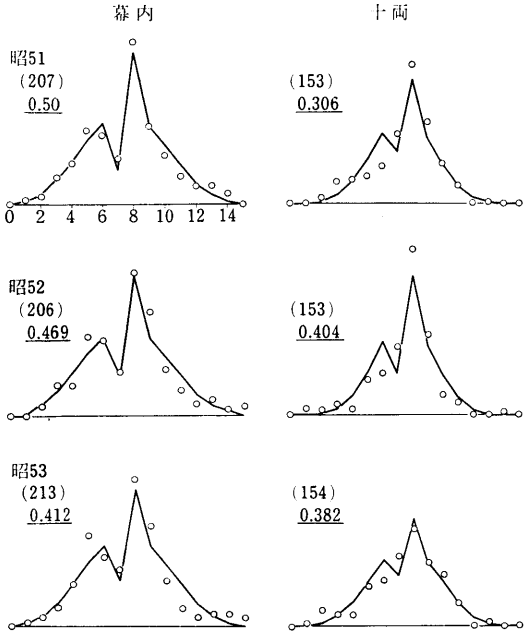

図 2

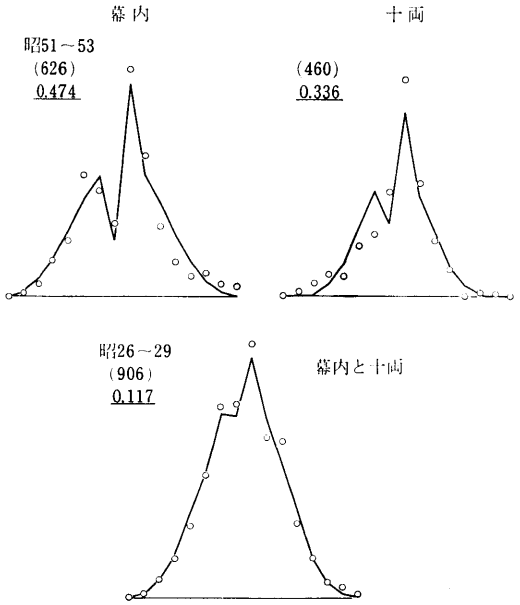

図 3

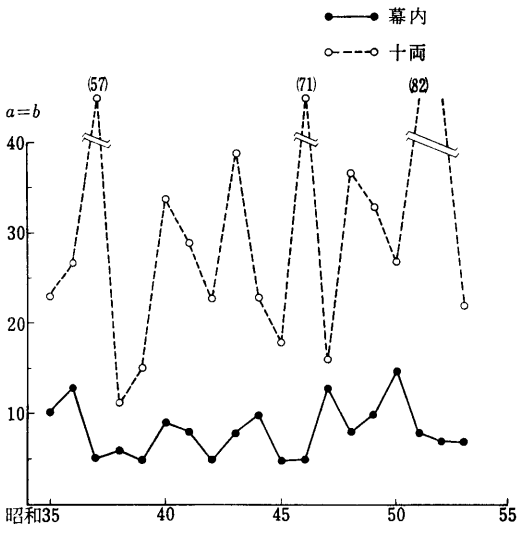

図 4

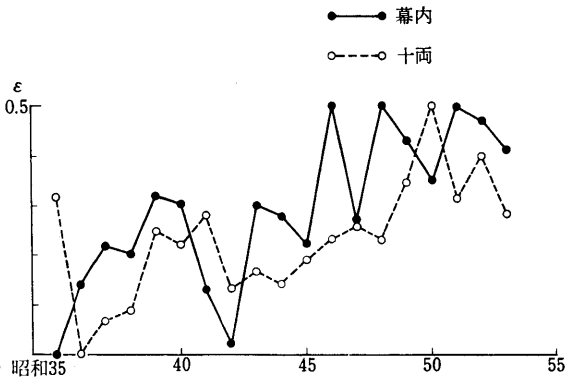

図 5

さく十両の場合が大きめである．この数値が大きいほど単純な2項分布に近いから，幕内では p の変動がはげしく十両ではどんぐりの背くらべ的であるといった事情が読みとれる．さらに図5をみると，最近では ε の値がかなり増加している．つまり，14日目で7勝している力士は千秋楽でほとんど勝ち越ししているといった傾向を示している．

3. 一般的考察

今，2つの離散型変量 X, Y を考え，Y の値域の部分集合 A が1つ固定されているとする．そして Y が与えられたときの X の条件は分布が，Y が集合 A に入るか入らないかにだけ依存して定まるものとする．

そこで

$$f_1(x) = \Pr\{X = x | Y \in A\}$$
$$f_2(x) = \Pr\{X = x | Y \bar{\in} A\}$$

と置く．X の条件付でない確率関数 $f(x)$ との間には

$$f(x) = \lambda f_1(x) + (1-\lambda) f_2(x)$$
$$\lambda = \Pr\{Y \in A\}$$

という関係が成立している．さらに

$$\mu = \sum x f(x) \quad , \quad \sigma^2 = \sum (x-\mu)^2 f(x)$$
$$\mu_i = \sum x f_i(x) \quad , \quad \sigma_i^2 = \sum (x-\mu_i)^2 f_i(x)$$

と置けば

$$\mu = \lambda \mu_1 + (1-\lambda)\mu_2$$
$$\sigma^2 = \lambda \sigma_1^2 + (1-\lambda)\sigma_2^2 + \lambda(1-\lambda)(\mu_1 - \mu_2)^2$$

となることも容易にわかる．

次に Y の分布を $g(y)$ として

$$\bar{g}(z) = \sum_y g(y) f(z-y)$$
$$\Delta(z) = \sum_y g(y)[f(z-y) - f_1(z-y)]$$
$$\gamma(z) = \sum_{y \in A} g(y)[f(z-y) - f_1(z-y)]$$

と置けば

$$Z = Y + X$$

の分布は

$$h(z) = \bar{g}(z) + \delta \Delta(z) - (1+\delta)\gamma(z) \quad \delta = \lambda/(1-\lambda)$$

と表わされる．つまりここで考えているタイプの従属した和 (dependent convolution) は，X と Y の周辺分布 f, g が固定されているとき集合 A と

$$d(x) = f(x) - f_1(x)$$

によって定まる性格のものであることがわかる．これを X と Y の **$[A, d]$ convolution** と呼び，記号で

$$Z = Y + X \quad [A, d]$$

と書くことにする．

例えば S_r を2項分布 $B(r, p)$ に従う確率変数，$A_r = \{r\}$，$\lambda_r = \Pr\{S_r \in A_r\}$，$X_{r+1}$ を $\{0, 1\}$ の値をとり

$$f_1(1) = p \quad , \quad f_2(1) = \bar{p}$$
（この場合 $d_r(1) = (p - \bar{p})(1 - \lambda_r)$ である）

なるディコトミー変量として

$$S_{r+1} = S_r + X_{r+1} \quad [A_r, d_r]$$

なる変量が定義できる．以下 $l = r+1, \cdots, n-1$ について

$$A_l = \{r, r+1, \cdots, l\}$$
$$\Pr\{S_l \in A_l\} = \lambda_l$$
$$S_{l+1} = S_l + X_{l+1}[A_l, d_l]$$
$$(X_{l+1}, d_l \text{ は } X_{r+1}, d_r \text{ と同じ})$$

より最終的に S_n という確率変数が定義できる．この S_n の分布は，r 回の成功後成功率 p を $\tilde{p}$ に変えて n 回まで試行したときの成功回数の分布と一致する．これは，筆者が文献 [5] で考えた変形2項分布 $MB(n, r; p, \tilde{p})$ に他ならない．

X と Y とがどんな場合に無相関になるか，次の命題が成立する．

X と Y とが無相関であるための必要十分条件は

$$\text{(i)} \quad \sum xd(x) = 0$$

または

$$\text{(ii)} \quad \sum_{y \in A} yg(y) = \lambda \sum yg(y)$$

の少くとも一方が成立することである．

何故なら

$$\begin{aligned}
\mathrm{E}\{XY\} &= \sum_x x \sum_y y \Pr\{X = x,\ Y = y\} \\
&= \sum_x f_1(x) \sum_{y \in A} yg(y) + \sum_x xf_2(x) \sum_{y \in A} yg(y) \\
&= \sum_x xf_1(x) \sum_y yg(y) + \sum_x [f_1(x) - f_2(x)] \sum_{y \in A} yg(y) \\
&= \sum_x f(x) \sum_y yg(y) + \delta \sum_x xd(x) \left[\sum yg(y) - \sum_{y \in A} yg(y)/\lambda\right]
\end{aligned}$$

であるから，最後の式の第2項が0となるときに限り X と Y とは無相関になることがわかる

特に X が0か1の値しかとらない場合には，

$$f(0) = q,\ f(1) = p \quad (p + q = 1)$$
$$d(0) = \varepsilon = -d(1)$$

だけで X のほうの分布は特定化できる．ε が0でない限り (i) の条件は成立しないから，(ii) の条件が成立するときに限り X と Y とは無相関になる．$g(y)$ の値域を $\{0, 1, \cdots, n\}$ とし，

$$g(-1) = g(n+1) = 0$$

と約束すれば，$0 \leqq z \leqq n+1$ に対して

$$\bar{g}(z) = qg(z) + pg(z-1)$$
$$\Delta(z) = \varepsilon[g(z) - g(z-1)]$$
$$\gamma(z) = \begin{cases} \Delta(z) & z \in A_{11} \\ \varepsilon g(z) & z \in A_{10} \\ -\varepsilon g(z-1) & z \in A_{01} \\ 0 & z \in A_{00} \end{cases}$$

ここで

$$A_{11} = \{z \mid z \in A,\ z-1 \in A\}$$
$$A_{10} = \{z \mid z \in A,\ z-1 \notin A\}$$
$$A_{01} = \{z \mid z \notin A,\ z-1 \in A\}$$
$$A_{00} = \{z \mid z \notin A,\ z-1 \notin A\}$$

さらに $A=\{n_0\}$ の場合を考えると

$$A_{11}=\phi,\ A_{10}=\{n_0\},\ A_{01}=\{n_0+1\}$$
$$A_{00}=\{1,\ 2,\ \cdots,\ n_0-1,\ n_0+2,\ \cdots,\ n\}$$

であるから

$$h(n_0)=\bar{g}(n_0)+\delta\Delta(n_0)-\delta\varepsilon$$
$$h(n_0+1)=\bar{g}(n_0+1)+\delta\Delta(n_0+1)+\delta\varepsilon$$
$$h(z)=\bar{g}(z)+\delta\Delta(z)\quad(z\neq n_0,\ n_0+1)$$
$$\delta=g(n_0)/(1-g(n_0))$$

と表現される．また X と Y とが無相関になるための (ii) の条件が

$$\sum yg(y)=n_0$$

であることも容易にわかる．特に $n=14$, $n_0=7$, さらに $p=0.5$ の場合が，第2節で考えた確率モデルになっている．

この種のより一般のモデルに対する諸性質については，いずれ稿を改めて記述する予定である．

参 考 文 献

[1] Ishii, G. and Hayakawa, R. (1960) On the compound binomial distribution, *Ann. Inst. Statist. Math.*, **12**, 69-80.

[2] 川上理一 (1959) 相撲星取表の分析，生物統計学雑誌 **6**, 137-141.

[3] 読売新聞社，縮刷版.

[4] 杉原雪夫 (1958) 生物統計学上における二項分布法則の適用に就て，生物統計学雑誌 **5**, 67-74.

[5] 鈴木義一郎 (1976) 変形二項分布モデル，統計数理研究所彙報 **24**, 41-46.