

確率的大小とその応用

統計数理研 柳本 武美

日本 IBM 渋谷 政昭

確率的大小の概念はノ・パラメトリック検定の理論においてきわめて基本的であり、単調尤度比、正(負)の相関性、increasing hazard rate、対称性の検定(一変数および二変数)と密接な関係をもっている。

S を $\mathbb{R}^2$ の半平面 $\mathbb{R}_x = \{(x, y); x > y\}$ のボレル部分集合、 tS を直線 $x = y$ に関して S に対称な集合とする。 $\mathbb{R}_x$ のボレル集合のある族 $\mathcal{R} = \{S, S \subset \mathbb{R}_x\}$ が与えられたとき、確率ベクトル (X, Y) が

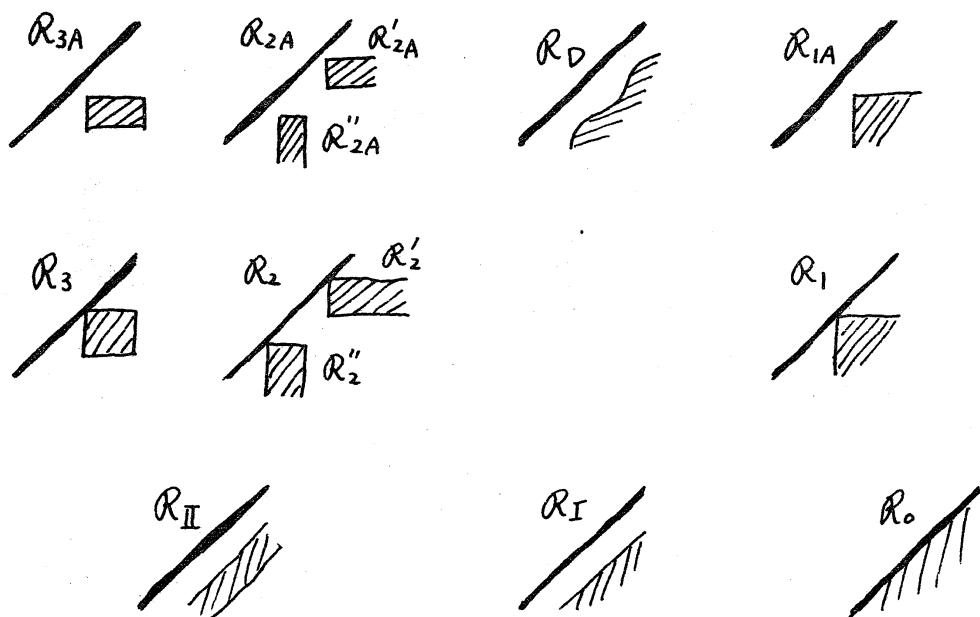
$$P\{(X, Y) \in S\} \geq P\{(X, Y) \in tS\} \quad \text{for all } S \in \mathcal{R}$$

を満たす時、 X が Y よりも $\mathcal{R}$ の意味で確率的に大きいと言ひ、

$$X \succ Y \quad (\mathcal{R})$$

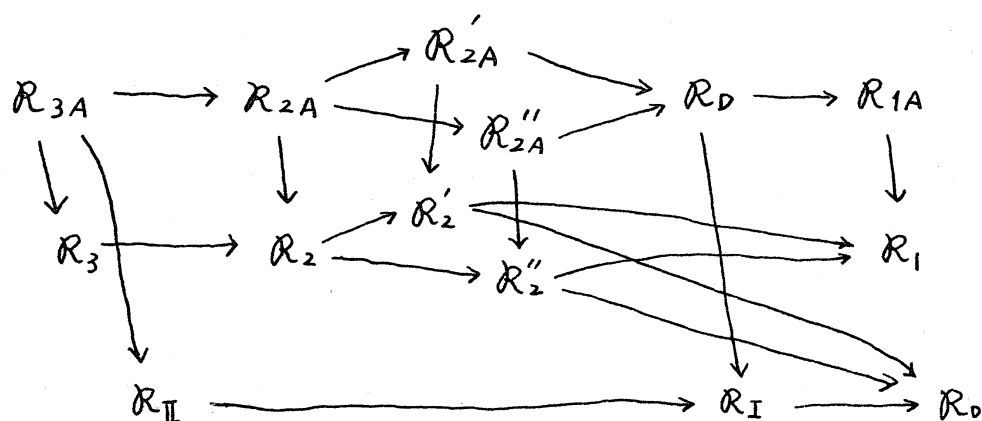
と書く。

$\mathcal{R}$ として意味のあるのは、次の図で典型的な集合 S が表わされるような族である。斜の太線は、 $x = y$ である。



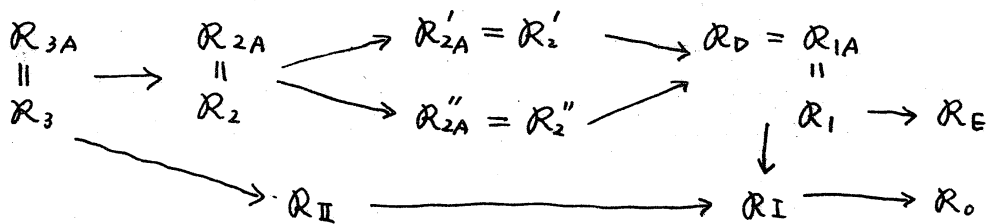
ただし R_D は $x < x'$, $y > y'$ のとき $(x, y) \in S$ なら $(x', y') \in S$ であるような本レール集合 S の族である。また、 $R_{2A} = R'_{2A} \cup R''_{2A}$, $R_2 = R'_2 \cup R''_2$ である。

このような集合族で定義される確率的大小について次のような包含関係が成立する。また矢印のない集合族内では、包含関係は成り立たない。



包含関係の不成立を確かめることは、比較的少数の反例で
 できる。たとえば右のような9点上の
 分布を考えると $X \succ Y (R_3)$ だが、
 $X \not\succ Y (R_I)$ であり、これから直ちに
 $X \succ Y (R_3, R_2, R'_2, R''_2, R_1, R_0)$ のい
 ずれも $X \succ Y (R_{3A}, R_{2A}, R'_{2A}, R''_{2A}, R_D, R_I, R_E)$ を含
 まぬことが言える。

$X \parallel Y$ のときは、上の諸定義のいくつかが同等となり、図
 式は ずっと簡単となる。



R_3 は単調尤度比, R_1 は「確率的に大」, R_2 は Pfanzagl が
 導入した概念である。上で新しく導入した R_E は X, Y
 の分布関数をそれぞれ $G(x), H(y)$ とするとき、
 $\int (H(t) - G(t)) dt \geq 0$ で $X \succ Y (R_E)$ を定義する。た
 だし、積分は $\pm \infty$ を許すものとし、 $+\infty$ は正、 $-\infty$ は負とみ
 なす。

2次元確率分布が正の相関性 ~~をもつとは~~ $\rho(i, j)$ をもつとは、図の
 ような任意の $a_1 < a_2 < a_3$; $b_1 < b_2 < b_3$ によって作られる

2次元区間 S_1, S_2, S_3, S_4 に対し常に

$$P(S_1) \cdot P(S_4) \geq P(S_2) \cdot P(S_3)$$

が成り立つことである。逆の不等号が常に成り立つならば負の相関性 $N(i, j)$ と呼ぶ。

図は必ずしもすべてこの場合を含まない。それは $p(3, 2'')$ など " で正の方向への無限の2次元区間を含むこと, $p(2', 1)$ など ' で負の方向への無限区間を含むことを示し, また $p(i, j)$ で $p(i, j')$ および $p(i, j'')$ を意味する便法をとっている。

$p(3, 3)$

	a_1	a_2	a_3
b_3			
		S_2	S_4
b_2			
		S_1	S_3
b_1			

$p(3, 2')$

	a_1	a_2	a_3
		S_2	S_4
b_2			
		S_1	S_3
b_1			

$p(3, 1)$

	a_1	a_2	a_3
		S_2	S_4
b_1			
		S_1	S_3

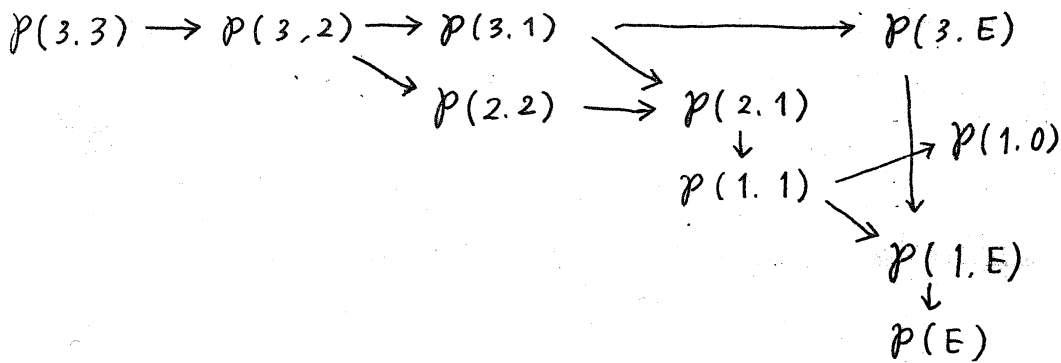
$p(2', 1)$

	a_1	a_2
	S_2	S_4
b_1		
	S_1	S_3

$p(1,1)$

	b_1
a_1	S_2
	S_4
	S_1
	S_3

包含関係も $p(i, j)$ が $p(i, j')$, $p(i, j'')$ 等を含み, これらがより弱い関係を含むことになるがそれを省略している。



$p(3,3)$ は positively likelihood ratio dependent とよばれているものである。 $p(3,1)$ は positively regression dependent とよばれているもの、 $p(1,1)$ は positively quadrant dependent とよばれているものである。

このような定義はまた条件付分布の確率的大小によって定義することもできる。たとえば $p(3,1)$ は

$$Y|_{X=x_2} > Y|_{X=x_1} \quad (R_1) \quad \forall x_2 > x_1$$

$$Y|_{X>x} > Y|_{X\leq x} \quad (R_3) \quad \forall x$$

と同等である。 $p(3,E)$ などは

$$Y|_{X=x_2} > Y|_{X=x_1} \quad (R_E) \quad \forall x_2 > x_1$$

で定義される。 $p(E)$ は

$$\int \{F(x,y) - F(x,\infty)F(\infty,y)\} dx dy > 0$$

で定義される。積分は $\pm\infty$ を許す。

上の包含関係により、正の相関性についてのいくつかの命題はより簡単に、より本質の明確な形で証明することができる。たとえば $X_{(1)} \leq \dots \leq X_{(n)}$ を任意の確率分布から

の順序統計量とすると $(X(r), X(s)) \in \rho(3, 3)$, したがって相関係数が正である, というもっとも弱い意味を含めた種々の正相関性をもつ。

一変量対称性の検定

一変量の対称性の検定は R^1 上の確率変数 X , その分布関数 $F(x)$, が点 μ に関して対称であるか否かを扱う。以下一般性を失う事なしに, $\mu = 0$ とする。

対立仮説として分布が “正に偏っている” を採用する。

R を一つの確率的大小とするととき,

$$1. X \succ -X \quad (R)$$

$$2. X \succ -X' \quad (R)$$

$$3. X \succ 0 \quad (R)$$

$$4. X|_{X \geq 0} \succ -X|_{X \leq 0} \quad (R)$$

により定義される。ただし X' は X と独立に同一分布に従う確率変数とする。

上を調べる事に依り 5 種類の “正の偏り” が定義される。

以下簡単のためにのみ密度関数 $f(x)$ が存在するとする。

$$1. X \succ 0 \quad (P_0) \iff F(0) \leq \frac{1}{2}$$

$$2. X \succ 0 \quad (P_1) \iff F(x) \leq 1 - F(-x)$$

$$3. X \succ 0 \quad (P_2) \iff f(-x) \leq f(x) \quad \text{ただし } x \geq 0$$

$$4. X \succ 0 \quad (P_3) \iff f(x)/f(x) \text{ は } x \geq 0 \text{ で増加.}$$

$$5. X \succ 0 \quad (P_4) \iff f(x)/f(-x) \text{ は増加.}$$

上の定義を拡張して、二つの確率変数 X, Y に於て、

" X の方が Y より正に偏っている"を定義出来、以下と同様の議論が成り立つ。しかし はんぱつ になるのでここからは省略する。

定義の間には次の包含関係が成り立つ。

$$(i) \quad P_4 \rightarrow P_2 \rightarrow P_1 \rightarrow P_0 \\ \quad \quad \quad \searrow \\ \quad \quad \quad P_3$$

$$(ii) \quad X \succ 0 \quad (P_2) \text{ かつ } (P_3) \Rightarrow X \succ 0 \quad (P_4)$$

$$(iii) \quad X \succ 0 \quad (P_0) \text{ かつ } (P_3) \Rightarrow X \succ 0 \quad (P_1)$$

手始めに次の性質を述べる。

$$(i) \quad X \succ 0 \quad (P), \quad -X \succ 0 \quad (P) \text{ であれば } X \text{ は対称である。}$$

但し、 P は P_4, P_3, P_2 及び P_1 のいずれかとする。

$$(ii) \quad f(x) \text{ が増加な奇関数とする。この時勝手な } P \text{ について}$$

$$X \succ 0 \quad (P) \text{ ならば } f(X) \succ 0 \quad (P) \text{ が成り立つ。}$$

次に $x_1, \dots, x_n$ を母集団、その連続な分布関数 $F(x)$ からの size n の標本とする。 $|x_i|$ の $\{|x_1|, \dots, |x_n|\}$ のランクを r_i とし、 $z_{r_i} = 1 \quad x_i > 0$, $z_{r_i} = 0 \quad x_i < 0$ とし、 $z = (z_1, \dots, z_n)$ を考える。それらを確率変数とみなす時大文字で

記す。 Z の集合 Π に半順序を導入する。

$$1. \quad Z = (z_1, \dots, z_n) \geq Z' = (z'_1, \dots, z'_n) \quad (O'_1)$$

$$\iff i < j \text{ があって } z_i = z'_j = 1, \quad z'_i = z_j = 0.$$

$$\text{かつ } z_k = z'_k \quad k \neq i, j$$

$Z \geq Z' \quad (O_1)$ は (O'_1) の連なりで結ばれる時。

$$2. \quad Z \geq Z' \quad (O_2) \iff z_i \geq z'_i \quad i = 1, 2, \dots, n$$

$$3. \quad Z \geq Z' \quad (O_3) \iff \sum_{i=1}^k z_i \geq \sum_{i=1}^k z'_i \quad k = 1, 2, \dots, n$$

これらの半順序の間に次の関係が存在する。

$$Z > Z' \quad (O_3) \iff Z \geq Z' \quad (O_1 \vee O_2)$$

但し右の式は Z と Z' が (O_1) と (O_2) の連なりで結ばれることを意味する。

母集団の分布と統計量の分布の間に次の命題が成り立つ。

$F > 0 \quad (P)$ は母集団と分布関数を同一視して得る。

$$(i) \quad F > 0 \quad (P_2) \quad Z \geq Z' \quad (O_2) \text{ ならば } P_F(Z = Z) \geq P_F(Z = Z')$$

$$(ii) \quad F > 0 \quad (P_3), \quad Z \geq Z' \quad (O_1) \text{ ならば } P_F(Z = Z) \geq P_F(Z = Z')$$

(i)(ii)をあわせて (iii)を得る。

$$(iii) \quad F > 0 \quad (P_k), \quad Z \geq Z' \quad (O_3) \text{ ならば } P_F(Z = Z) \geq P_F(Z = Z')$$

$$(iv) \quad T(Z) \text{ が非減少であるとは } Z \geq Z' \quad (O_3) \text{ ならば } T(Z)$$

$\geq T(z')$ が成り立つことを云う。 $F \succ 0 (p_1)$ で $T(z)$ が非減少であれば、任意の c に対して、 $P_F(T(z) \geq c) \geq P_{F_0}(T(z) \geq c)$ 。但し、 F_0 は対称な分布関数とする。

二変量対称性の検定

(X, Y) を R^2 上の確率変数、その分布関数を $F(x, y)$ 、周辺分布関数を $F_1(x)$ 、 $F_2(y)$ とする。帰無仮説として直線 $x = y$ に關して対称、対立仮説として $X \succ Y (R)$ を採用する。実際には、 $X \succ Y (R_S)$ 、 $X \succ Y (R_3)$ が重要である。更にこの場合、 $X \succ Y (R_4)$ を導入する。

$$\begin{aligned} X \succ Y (R_4) &\Leftrightarrow P_F(S(a_1, b_1; a_2, b_2)) P_F(t_S(a_2, b_2; a_3, b_3)) \\ &\geq P_F(t_S(a_1, b_1; a_2, b_2)) \cdot P_F(S(a_2, b_2; a_3, b_3)) \quad a_3 > a_2 > a_1 > b_1 > b_2 > b_3 \end{aligned}$$

R_I, R_{II} を用いる事も興味あるが、これは $X - Y (p_1)$ 、 $X - Y (p_2)$ と同等になり、一変量の対称性の検定の問題に帰着する。

他の統計の分野と次のつながりを持つ。

- (i) $X \perp Y$ の場合には結局、等 size の 2 標本問題になる。
- (ii) $y = -x$ に退化している場合、一変量対称性の検定と同じになる。
- (iii) $x > y$, $x < y$ に落ちる標本の個数が一定という条件を

つけると $(X, Y) \mid_{X > Y}$ と $(Y, X) \mid_{X < Y}$ の二変量二標本問題とみなされる。

確率の大小の不変性を調べる。

(i) $\mathcal{R}$ が $\mathcal{R}_1, \mathcal{R}_D, \mathcal{R}_2, \mathcal{R}_3, \mathcal{R}_4$ のいずれかとする。 $f(x)$ が単調増加ならば $X > Y (\mathcal{R})$ の時、 $f(X) > f(Y) (\mathcal{R})$ 。逆に、 $\mathcal{R}$ が $\mathcal{R}_1$ か $\mathcal{R}_D$ ならば、任意の $X > Y (\mathcal{R})$ なる (X, Y) に対して $f(X) > f(Y)$ ならば f は単調増加である。

(ii) $\mathcal{R}$ が $\mathcal{R}_D$ か $\mathcal{R}_3$ とする。 $r = (r_1(x, y), r_2(y, x))$ について $r_1(x, y)$ が x について単調増加、 y について単調減少であれば $X > Y (\mathcal{R})$ の時、 $r_1(X, Y) > r_1(Y, X) (\mathcal{R})$ 。逆に任意の $X > Y (\mathcal{R}_D)$ なる (X, Y) に対して $r_1(X, Y) > r_1(Y, X) (\mathcal{R}_D)$ が成立てば $r_1(x, y)$ は上の条件を満たす。

(iii) は関連して (ii) を与える。 $\mathcal{R}_D$ の例を得る。

(iii) $r = (r_1(x, y), r_2(y, x))$ で r_1, r_2 が (ii) の条件を満たす。
 $X > Y (\mathcal{R}_D)$ ならば $r_1(X, Y) > r_2(Y, X) (\mathcal{R}_D)$ 。特に (X, Y) が対称、 $f(x) \geq g(x)$ で f, g 共に増加ならば $f(X) > g(Y) (\mathcal{R}_D)$ 。

$\begin{pmatrix} x_1 \\ y_1 \end{pmatrix}, \dots, \begin{pmatrix} x_n \\ y_n \end{pmatrix}$ を母集団からの size n のサンプルとする。
 分布は連続であつ $x = y$ 上には正の確率を持たないとする。
 $\max\{x_1, y_1\} > \dots > \max\{x_n, y_n\}$ なるよう添字をつけかえて、
 $x'_i = \max\{x_i, y_i\}$, $y'_i = \min\{x_i, y_i\}$ と記す。 $z_i =$

$\operatorname{sgn}(x_i - y_i)$, r_i, s_i を x'_i, y'_i の $\{x_1, y_1, \dots, x_n, y_n\}$ のラン
クとする。 $m = (m_{ij})$ を $m_{ij} = c(x'_i - x'_j) c(y'_j - y'_i)$ で定
義する。 ここで $c(x) = 1$ for $x > 0$, $= 0$ otherwise とす
る。

(i) $P_F(T((\frac{x_1}{y_1}), \dots, (\frac{x_n}{y_n})) \geq c)$ があらゆる対称な F について
一定であれば, $P_F(T((\frac{x_1}{y_1}), \dots, (\frac{x_n}{y_n})) \geq c \mid ((x'_1, y'_1), \dots, (x'_n, y'_n)))$
 $= P_F(T(z_1, \dots, z_n) \geq c \mid ((x'_1, y'_1), \dots, (x'_n, y'_n))) = P_F(T \geq c)$ が
a.e. な $((x'_1, y'_1), \dots, (x'_n, y'_n))$ について成り立つ。 即ち, 条件
付き符号検定になる。

(ii) $T((\frac{x_1}{y_1}), \dots, (\frac{x_n}{y_n})) = T((\frac{f(x_1)}{f(y_1)}, \dots, (\frac{f(x_n)}{f(y_n)}))$ が 任意の狭義
単調増加な f について a.e. で成り立つ為の必要条件は T が,
 $z = (z_1, \dots, z_n)$ と $((r_1, s_1), \dots, (r_n, s_n))$ のみに依存すること
である。

(iii) $T((\frac{x_1}{y_1}), \dots, (\frac{x_n}{y_n})) = T((\frac{r(x_1, y_1)}{r(y_1, x_1)}, \dots, (\frac{r(x_n, y_n)}{r(y_n, x_n)}))$ が $r(x, y)$
が x について増加, y について減少, 更に $r = (\frac{r(x, y)}{r(y, x)})$ が
 y について減少である関数について a.e. に成り立つとする。
その為の必要条件は T が m と z にのみ依存することである。

不偏性と (ii) の不変性を仮定して条件付き符号順位統計量に
ついて, 統計的性質を与える。

(i) $X > Y$ ($\mathcal{R}_3$), $z \geq z'$ ($\mathcal{O}_2$) ならば $P_F(Z = z \mid ((r_1, s_1), \dots,$
 $\dots, (r_n, s_n))) \geq P_F(Z = z' \mid ((r_1, s_1), \dots, (r_n, s_n)))$.

(ii) $X \succ Y (R_q)$ $z \succeq z' (O'_1)$ $r_i > r'_i > s_j > s'_j$ ならば,

$$P_F(Z=z | ((r_i, s_i), \dots, (r_n, s_n))) \geq P_F(Z=z' | ((r_i, s_i), \dots, (r_n, s_n)))$$

(iii) T を用いて, 棄却域 R を $T \geq c(\alpha, ((r_i, s_i), \dots, (r_n, s_n)))$ で構成する。等号は level を用いて普通に決める。この時,

更に, $((x'_i), \dots, (x'_n)) \in R$ で $x' \geq x, y' \leq y$ であれば $((y'_i), \dots, (y'_i), \dots, (x'_n)) \in R$ なる条件を加える。 $X \succ Y (R_D)$

であれば, $P_F(R) \geq P_{F_0}(R) = \alpha$ 。 F_0 はある対称な分布関数とする。

(iv) T が次の条件を満たすとしよう。

(i) $z \succeq z' (O_2)$ ならば $T(z | ((r_i, s_i), \dots, (r_n, s_n))) \geq T(z' | ((r_i, s_i), \dots, (r_n, s_n)))$.

(ii) z^0 に対して $z_i^0 = 1$ ならば $r_i \geq r'_i > s'_i \geq s_i$; $z_i^0 = 0$ ならば $r'_i \geq r_i > s_i \geq s'_i$ が成立しているとする。この時, $T(z | ((r'_i, s'_i), \dots)) < T(z^0 | ((r'_i, s'_i), \dots))$ ならば $T(z | ((r_i, s_i), \dots)) < T(z^0 | ((r_i, s_i), \dots))$, 亦 $T(z | ((r_i, s_i), \dots)) > T(z^0 | ((r_i, s_i), \dots))$ ならば $T(z | ((r'_i, s'_i), \dots)) > T(z^0 | ((r_i, s_i), \dots))$.

更に元の分布 (X, Y) について, $(X, Y) \sim (f(U), g(V))$ であると仮定する。但し, (U, V) は適当な対称な確率変数, f, g は $f(x) \geq g(x)$ for $\forall x$ なる適当な実狭義増加関数とする。

$P_{(X, Y)}(T \geq c) \geq P_{(U, V)}(T \geq c)$ が成立立つ。