⑲ あるNONPARAMETRIC TEST について

森　村　英　典

§ 1. Introduction.

non parametric test のうち，*two-sample test*（或いは，もつと一般に *k-sample test*）や独立性，若しくは任意性の検定等において主要な役割を果しているものに *run* の長さ又は総数による *test* がある。

これは，例えば *2-sample* で云えば，

a a a a a a a b b a b b a b b

というような列が得られたとしたとき，*a*, *b* が同一母集団から得られたものであれば，どの一つをとつても，同じ確率で現われるべきだと考え，長さ7以上の *run* の生ずる確率を求めると，5％以下になるので，同一母集団から取られたという仮説を5％の危険率で棄却する。　或いは又總数6個の *run* の生ずる確率が5％以下ならば，同様に，同じ仮説を棄却する訳だが，この場合，その確率は5％以上になるので5％の危険率では，この仮説は棄却出來ない。　又逆に，

a a a a a b b b b b a a a a

というような列であると，前と同様，*a*, *b* の総数は夫々7個

6個である。　然し，こゝでは一番長い *run* の長さは 6，*run* の総数は 3 であるから，5%の危険率で，前と同一の仮説は，長さによる検定では棄却されるが，総数による検定では，棄却されないこととなる。　この現象は，どちらの検定方式をとってみても，*run* の持っている2つの *information*（長さと数とに関する）の何れか一方を捨ててしまっていることに起因しているものと考えられる。　その両方を併せ考えるような検定方式であれば，これらの検定より能率のよいものとなることは当然想像される。　こゝで述べようとする検定方式は，2－*sample* の場合には *aa* というような *tie* の数によるもので，これは *run* の長さが長くなれば多くなり，数が減っても同じように多くなるから，この方向に一歩近ずいたものとも云えるように思われる。

所で，*k*－*sample test*，独立性，任意性の *test* は，同一の確率で，各 *sample* が得られることを仮説としての検定方式であるから，次節で述べる *entropy* の概念を用いて，この検定方式を作ることが出来る。　一般の *k*－*sample* の場合について，この検定方式が使われ得るが，その標本分布についてはまだ計算していないので，こゝでは 2－*sample* 及び 3－*sample* の場合について述べる。　更に又，*sample size* を大きくした時の漸近分布や数表等についても，なされるべきことが残されている。

尚，（　）は2項係数を，［　］は多項係を示す。

§ 2. Entropy.

Shannon [1] によれば，次々にある symbol を生み出して行く process（discrete information source と呼ばれている）の無秩序性を測る measure として，entropy H という量が用いられる：

$$H = -\sum_{i,j} p(i)\, p_i(j) \log p_i(j) \quad ,$$

こゝで i というのは，state と呼ばれて過去の履歴を表わし，j はその state にある時，次に生み出される symbol を示している。 つまり，$p(i)$，$p_i(j)$ は夫々 state i の起る確率，state i にあるという條件の下で symbol j の生ずる確率である。 若し，次々に生み出される symbol が独立であるならば，

$$p_i(j) = p(j)$$

となるから，H は

$$H_I = -\sum_i p(i) \log p(i)$$

という形を取る。 この H_I はすべての $P(i)$ が等しい時に最大となつて，symbol の種類が r 個であるとすれば，

$$H_M = \log r$$

となる。 一般に $H \leqq H_I$ が証明されるから，結局

$$H \leqq H_I \leqq H_M$$

の関係は常に成立している。 独立であれば H は H_I に，等しいのだから，標本からこれに対応する量を作れば，独立のときには

H_Iに近い価をとることが多くなる。 又，*nonparametric*の場合の所謂 *randomness* ということは，一様性と同義であるから，その検定のためには，HがH_Mに近いかどうかを調べればよい。 結局，独立性を検定するためには H_I-H をとつて，これらが0に近いかどうかで検定を行えばよいことになる。
この H_I-Hという量を書き直すと，

$$H_I-H=-\sum_i p(i)\log p(i)+\sum_{i,j}p(i)p_i(j)\log p_i(j)$$
$$=\sum_{i,j}p(i,j)(\log p_i(j)-\log p(j))$$
$$=\sum_{i,j}p(i,j)\log\frac{p_i(j)}{p(j)}$$

となるが，若し独立であれば $\log$の中は 1となるから，この値は常に0となる。 独立であるかないかを個々のi,j についてみても，全体として見るためには，それを平均化することになろう。この量は丁度，それを表現しているわけである。

然し，一般にiという *state* は起り得るすべての履歴と考えられるから，之を忠実に反映する量は作れる筈もないし，実際には，かなり短い長さの列について考えるわけだから，あまり長い履歴を考えることは，かえって不自然でもある。

一般にある長さの列が標本から得られた時，それからどの位の履歴を考えるのが *Optimum* かということも問題にされ得るかも知れないが，それが得られたとしても，長い履歴を考えると標本分布を求めることは著しく困難となるし，独立か否かという問題に対しては，ある *symbol* と他の *symbol* との間に，関連が考えられるかどうかの問題だから，理論的には単純マルコフとして $p_i(j)$ を考えれば十分であろう。

たゞ，この時には本当は独立であるのに，見かけ上，関連があ

るように見える場合が，2重，3重のマルコフと考えた時より多くなることが当然想像される。　云換えれば，2重，3重のマルコフと考えれば，検定はより精密となるであろう。

こゝでは，単純 Markov としたとき上記の量に対応する量を標本から作り，それを Criterion とする。　2-sample の時の criterion は前節でも触れたように tie の数字である。これらの，独立という仮説の下での標本分布を求めることを次節以下で取扱う。　これらの分布は run の場合より複雑ではあるが，3-sample 以上の場合には，実際の数値計算の上からは寧ろ簡単であろう。

§ 3. Lemma.

標本分布を求めることは，結局 aa というような tie が u 個出来る場合の数を求めることが基本になる。

それで先ず，次の lemma を用意する。

[Lemma.]　N 個の symbol をならべたとき，a という symbol（総数 n）が u 個の tie を含む場合の数は，$u<n-1$ ならば

$$\sum_{i=0}^{N-2n+u+1}\ \sum_{k=1}^{P_{n-u-1}^{(n-u-1+i)}} f(n,u,i;d_k)\cdot(N-2n+u+2-i)$$

で与えられる。　但し，$f(n,u,i;d_k)$ は；$(n-u-1+i)$ を $(n-u-1)$ 個の自然数の和として書表わした時（その表わし方の総数は $P_{n-u-1}(n-u-1+i)$ 個），その1つの書表わし方 d_k が j という数を t_j 個ずつ含んでいるものとすれば

$$f(n,u,i;d_k)=\begin{bmatrix} n-1 \\ u.\ t_j \end{bmatrix}$$

で与えられるような函数である。

t_j については，従って

$$\sum_j t_j = n-u-1, \quad \sum_j j t_j = n-u-1+i$$

である。　尚 $p_\mu(\lambda)$ は自然数λをμ個の自然数で書表わす場合の数で，分割数とよばれ，一般に~~母~~函数を用いて求めることが出来，λ，μ の小さい所では表が出来ている（例えば[2]）．

それで，これから，書き表わし方を check することが出来る。

尚又，$u = n-1$ のときは簡単に

$$N-n+1$$

で与えられる．

（証　明）　a 以外の symbol は・で表わすことにする。N個の列の中でaが最初に出た所から最後に出た所までを取り出して考えると，この列の長さは $2n-u-1+i$ である。

i は0以上の自然数で，例えば $n=5$，$u=1$ としたとき，

$$a.\ aa\cdot a\cdot a$$

のように，つまっていれば0であり

$$a\cdot\cdot aa\cdot a\cdot a$$

のようになれば　$i=1$ としたことになる。　このようなすべてのaを含んだ列が移動してN個の長さの列のどこかを占めるわけだから，i を fix したとき，この長さで，この排列の列が生ずる場合の数は，それをずらすことの出来る場合の数だけ，即ち

$$N-(2n-u-1+i)+1$$

だけである。　この長さの列の起り得る permutation の数は次のようにして求められる。　この中には・が

$$(2n-u-1+i)-n$$

あり，これの位置の変化によって各々の場合が生ずるのであるが，常に tie の数は u 個であることが要求されるから，・の分れ方はいつも，$n-u-1$ である。何となれば，a と a との間は全部で $n-1$ 個あるが，そのうち u 個の tie を作るために，・を入れることの出来ない個所 u 個を除いただけが・の分れ方の数になるからである。・の分れ方が同じもので，その位置により生じ得る permutation の数は，その分け方 d_k が j という数字を t_j 個ずつ含んでいて，a と a との間に各 t_j 回，j 個の・を入れた場合の総数だから，tie のある所は 0 個の・を入れると考え，結局 u 個の 0，及び t_j 個の j を $n-1$ 個の位置（各 a の間はこれだけあるから）に入れる順列の数：

$$f(n.u.i;d_k)=\begin{bmatrix} n-1 \\ u\ t_j \end{bmatrix}$$

が求める permutation の数である。これを・の分け方，つまり各 d_k について加え，更に起り得るすべての i について加え合わせれば，N 個の symbol をならべた時，u 個の a の tie の出来るすべての場合の数を得る。k の動き得る範囲は，1 から $p_{n-u-1}(n-u-1+i)$，i の動き得る範囲については，a を含む列の最大の長さは N だから，

$$2n-u-1+i=N$$

から $N-2n+u+1$ が i のとり得る最大値となる。よって lemma の前半は証明された。

$u=n-1$ とすると，

$$p_{n-u-1}(n-u-1+i)=p_0(0+i)$$

となってしまい，或る自然数を 0 個の自然数に分割するということは意味を持たなくなるから，今迄の考え方は成立たない。この場合は，すべての a が間に 1 つも・を入れずに並んでいる場合で

あるから，その列の長さはnであり，それをずらした$N-n+1$個の場合だけ生ずることになる。（証明終）.

§ 4. 2 − Sample の場合

Symbol の種類が2つしかない時は，前節の・は，すべてbというような symbol で埋められる。従って，この場合，総数Nとaの個数nを求めておけば，aaという tie の数uが与えられれば，ab，ba，bbとなる個数はすべて決定される。

それで，entropy H に対応して標本から作られる量は，uによって一意に定まるから，独立の仮説の下でのHの分布はuの分布に等しい。

一般に，2 − Sample であれば，H_I-H はuの函数として凸函数となり，極小値を $u=\frac{n}{2}$ のときに取る。独立性の検定をする時は H_I-H の大きい値が significant になるから，

$$P_r(H_I-H\geq\alpha)=2\,P_r(u\geq\beta)$$

の関係から $P_r(u\geq\beta)$ を求めればよいことになる。勿論こゝでの確率は独立の仮説の下におけるものであり，又αは$u=\beta$のときの H_I-H の値である。そして若し，この確率がε%以下ならば，ε%の危険率で独立という仮説を棄却するわけである。

two-sample test のように，a，b の両 symbol が分离しているかどうかを見る検定では，uの大きい方だけを significant と考えればよいから $P_r(u\geq\beta)$がεより大きいか小さいかを知ればよい。所で，$P_r(u\geq\beta)$は前節で得た数を起り得るすべての数で割ったものであるから，$\beta\geq\frac{n}{2}$ として

$$P_r(u \geq \beta) = \left\{ \sum_{u=\beta}^{n-2} \sum_{i=0}^{N-2n+u+1} \sum^{p_{n-u-1}(n-u-1+i)} f(n,u,i\,;d_k)\cdot(N-2n+u+2-i) + N-n+1 \right\} \Big/ \binom{N}{n}$$

となる。　この式はかなり面倒のようだが，実際に問題となるような箇所では次の例に示すように，割合簡単に数値計算される。

［例　1］　（鍋谷［3］170頁の例）

$b\,b\,a\,a\,b\,b\,a\,a\,a\,a\,a\,a\,a\,a\,b\,b\,b\,b\,b\,b$

のような列が得られた時の独立性の検定．

$N=20$，$n=10$，$u=8$　である。

$p_{10-8-1}(10-8-1+i) = p_1(1+i) \equiv 1$，

即ち1つの i に対して d_k は1通りしかない。　それで t_j も j のどこか1つだけ1で他の j については，すべて0。

よって

$$f(N,u,i\,;d_k) \equiv \begin{bmatrix} 10-1 \\ 8,\,1 \end{bmatrix} = \frac{9!}{8!} = 9$$

又，i の動き得る範囲は　$N-2n+u+1=9$．

lemma で求めた数を $Q(u)$ で示すと．

$$Q(8) = 9\sum_{i=0}^{9}(10-i) = 495$$

$$Q(9) = 20-10+1 = 11$$

$$\therefore\quad P_r(u \geq 8) = \frac{495+11}{\binom{20}{10}} = \frac{506}{184756} < 0.003$$

従って，この2倍をとって，1％の危険率でこの仮説は棄却される。　run の総数による検定でも，長さによる検定でも，共にこの仮説は棄却されるが，その危険率は5％であるから，これら

の検定より，よい精度のものであるように思える。

§ 5. 3 − sample の場合

symbol を a, b, c とする。 この場合は9種の Chain が出来るわけだが，a, b, c の夫々の総数と任意の3種の chain の数が決 れば，他はすべて決る。 今，次表の如くに，その個数をきめたとする。 （ ）の中は自然に決ってくる数を示す。但し，列の最後に来た symbol と最初の symbol とで出来る Chain （例えば，最初が a で最後が b ならば ba ）の数は実際の数に 1 を加えたものである。

2位 \ 1位	a	b	c	total
a	u	w	$(n-u-w)$	n
b	v	$(m-w-v)$	(w)	m
c	$(n-u-v)$	(v)	$(l-n+u)$	(l)
total	n	m	(l)	N

a, b, c の記号には特別の意味はないから，得られた列の最初の symbol を a，最後の symbol を b と仮定 する。最後が a のこともあるが，その時は僅か修正を施せばよいから除外する。従って w は実際の ba という Chain の数より 1 だけ多い。

先ず N 箇の位置に a を勝手にならべたとする。 そのとき aa が u 個生ずる場合を考える。 その総数は lemma で与えられている。 a の次に b が v 個はいるのだからその場合の数は，$\binom{n-u}{v}$，a の前に b が $w-1$ 個はいる場合の数は $\binom{n-u-1-t_1}{w-1-t_1}$ である。 何となれば，a と a との間が 1 つしか空いていないと，

こゝにbが1個はいることによって，abもbaも1個ずつに勘定されるわけだから，aの後にbを自由におくとすると，aの前にbを置く方は，$a\cdot a$ となっている場合の数を引いて考えなければならないからである。 $a\cdot a$ となる個数は d_k に含まれる1の数だから，t_1 である。 結局この場合は

$$P_r(u,v,w)=\left\{\sum_{i=0}^{N-2n+u+1}\ \sum^{P_{n-u-1}^{(n-u-1+i)}}(N-2n+u-i)\cdot f(n,u,i;d_k)\cdot\binom{n-u-1-t_1}{w-1-t_1}\right.$$

$$\left.\times\binom{n-u}{v}\right\}\Bigg/\left(1-\frac{n}{N}\right)\begin{bmatrix}N\\ n.m\end{bmatrix}$$

が，aa，ab，ba の Chain の数が夫々 u，v，wとなる確率を与える。

この場合は，2－sample のときのようにuのみを変数としてHが決まらないので，u自体を Criterionとしたようなわけには行かない。 H（原理的には H_I-H ）に対応する標本から作られる量それ自体を Criterion に採用した方が良い。

［例 2］

$$a\,b\,c\,a\,a\,b\,a\,a\,a\,a\,a\,c\,c\,b\,b\,b\,c\,c\,c\,b\,b$$

のような列が得られた時の 3－sample test． $n=8$，$m=7$，$l=6$ ； $u=5$，$v=2$，$w=2$ である。 この時のHは，

$$H(5.2.2)=-\left\{\frac{8}{21}\left(\frac{5}{8}\log\frac{5}{8}+\frac{2}{8}\log\frac{2}{8}+\frac{1}{8}\log\frac{1}{8}\right)+\frac{7}{21}\left(\frac{1}{6}\log\frac{1}{6}+\frac{3}{6}\log\frac{3}{6}\right.\right.$$

$$\left.\left.+\frac{2}{6}\log\frac{2}{6}\right)+\frac{6}{21}\left(\frac{1}{6}\log\frac{1}{6}+\frac{2}{6}\log\frac{2}{6}+\frac{3}{6}\log\frac{3}{6}\right)\right\}$$

$$=1.39806$$ （但し対数の底は2）

となり，これより小さなHを生ずる確率は 0.001 より小となる。

即ち，1％の危険率で3つの標本は同一母集団から取られたという仮説は棄却される。

run の総数による検定では5%の危険率でこの仮説は棄却される。この場合も run によるより精度が良さそうである。

参考文献

[1] C.E. Shannon; The Mathematical Theory of Communication, The University of Illinois Press: Urbana, 1949

[2] 伏見康治; 確率論及び統計論, 河出書房

[3] 成実. 遠藤. 鍋谷.; 統計解析の理論, 朝倉書店