臨床実験のためのストッピング・ルール

伏見正則

1. はじめに

本稿では，臨床実験によって2種類の治療法のうちのどちらがより有効であるかを判定するための逐次実験法について述べる．

逐次実験法あるいは逐次検定法は，サンプル数一定の方式に比べて，平均的に見て，多くの場合にずっと少ないサンプル数で結論を下せる方式であるが，実施上のわずらわしさのために，工場での実験や検査等では実際にはほとんど用いられていないようである．しかしながら臨床実験では，患者はふつう逐次にしかやってこないので，観測を一斉に行なうことは不可能であり，逐次実験の手法を使うのがごく自然である．また臨床実験では多数の症例は集めにくいことが多いので，この点でも，平均的に見て少ないサンプル数で結論を下せる逐次実験方式のほうが好ましい．

もうひとつ重要なのは倫理的な配慮であり，この観点からすれば，たとえ研究のためであっても，劣っている治療法はなるべく使わないことが望ましい．もちろん，実験を開始する時点では，二つの治療法の優劣は判定し難いのであるから，実際には劣っているほうの治療法を使わざるをえないが，最終的には劣っている治療法を施された患者の数のほうがすぐれている治療法を施された患者の数より少ないほうが望ましいのである．

逐次解析の手法を統一的に論じたのはWald[7]であるが，臨床実験の分野への応用を目的として1960年頃までに発表された手法は Armitage [1] に紹介されている．この書物の中で扱われている手法では，いずれもつぎつぎにやってくる患者を2人ずつ対にして，そのうちの1人に一つの治療法をランダムに選んで施し，残りの1人にもう一つの治療法を施すという割りつけ規則（サンプリング・ルール）を採用している．（これを vector-at-a-time サンプリング・ルールとよび，以後VTと略記する．）したがって，実験が終了するまでに2種類の治療法を施される患者の数は完全に同じであり，劣っている治療法を施される患者の数をできるだけ少なくするという点はとくに考慮されていない．本稿では，この点を考慮に入れて最近10年間ぐらいの間に考案された手法のうちで比較的単純なものをいくつか紹介する．より一般的な議論に関しては，本特集の城島・浅野両氏の解説をご覧願いたい．

2. 仮定と要請

本稿で述べるサンプリング・ルールおよびストッピング・ルールは，つぎの仮定が成り立つ場合に適用できる．――治療が有効であったか否かが治療開始後比較的早くわかること，もっと正確にいえば，つぎに実験に加えられる患者の治療開始までに判定できること．この仮定が成り立つ場合には，治療の結果はベルヌーイ試行とみなす（成功は1，失敗は0で表わす）ことができ，問題は2種類の治療法が成功する確率 p_1, p_2（ともに0

と1の間の未知のパラメタ）のうちのいずれが大きいかを判定する問題に帰着する．

さて，好ましい逐次検定方式というのは，(イ)正しい判定を下す確率が大きく，(ロ)サンプル数の期待値が小さいものでなくてはならない．しかし，この二つの要請は一般に相反する性格のものであるから，p_1 と p_2 の値がほとんど等しい場合にも正しい判断を下すことを要求すると，サンプル数がきわめて大きくならざるを得ない．ところが，p_1 と p_2 が近い場合には，どちらの治療法がより有効であると判定しても，事実上あまりさしさわりはないと考えられる．そこで，p_1 と p_2 の差がある程度以上のときのみ，(イ)の条件を満たすことを要請する．すなわち，

$$|p_1-p_2|\geq\Delta^* \text{ ならば } \mathrm{P\{CS\}}\geq P^* \qquad (1)$$

であることを要求する．ここに，P{CS} は正しい判定を下す確率であり，Δ^* は0と1，P^* は1/2と1の間の適当な数であって，ともに実験に先立って研究者が指定する．

(ロ)のサンプル数については，全サンプル数の期待値 $E(N)$ と，劣っている治療法を施される患者数の期待値 $E(N_B)$ とが考えられるが，先に述べた倫理的な理由により，後者を最も重要視することにする．

3. サンプリング・ルール

劣った治療法の使用回数を減らすことを目的として考案されたサンプリング・ルールの代表的なものは，play-the-winner(今後PWと略記する)ルールとよばれるものである．これは，ある患者に施した治療法が有効であった場合には，つぎの患者にも同じ治療法を施し，無効であった場合にはもう一つの治療法を施すというものである．このルール自身を考案したのは Robbins[5]であるが，臨床実験への導入を検討したのは Zelen[8]であり，その後多くの人々によって研究された．本稿では，このサンプリング・ルールをもっぱら取り上げることとし，旧来のVTサンプリング・ルールと比較してみることにする．

4. ストッピング・ルール

Sobel & Weiss [6] は，2種類の治療法が有効であった回数の差があらかじめ定めた数に達したなら実験を止めるというストッピング・ルールを提案した．

$$\boldsymbol{R}_1: \quad |S_1-S_2|=r \qquad (2)$$

ここに，r は確率に関する要請(1)を満たす最小の整数にとる．たとえば，$P^*=0.95$，$\Delta^*=0.2$ の場合には $r=11$ ととればよい．また，一般に P^* があまり小さくない場合には，最適な r は近似的には次式によって与えられる．

$$r_0=[\log\{2(1-P^*)\}/\log(1-\Delta^*)]+1 \qquad (3)$$

ただし，[ ]はガウス記号．$r=11$ の場合のサンプル数の期待値は，表1の第1列に示すとおりである．また，VT サンプリング・ルールに対して(2)式の型のストッピング・ルールを用いる場合は，前記の P^*, Δ^* に対して $r=4$ で十分であり，このときのサンプル数の期待値は表1の第2列のようになる．なお，最終判定は，いずれのサンプリング・ルールの場合にも，有効であった回数の多いほうをすぐれていると判定するものであることは言うまでもない．

表1の第1および2列を比較してみれば，p_1, p_2 が比較的大きい場合には PW ルールが VT ルールよりすぐれているが，p_1, p_2 が小さい場合にはむしろ劣っていることがわかる．そこで，p_1, p_2 が大きい場合の PW ルールの良さは保存したまま，小さい場合にもサンプル数を減少させるようにストッピング・ルールを修正することが試みられた．

Fushimi [2] は，

$$\boldsymbol{R}_2: \quad |S_1-S_2|=r \qquad (4\mathrm{a})$$
$$\text{または} \quad F_1+F_2=s \qquad (4\mathrm{b})$$

が初めて成立した時に実験を終了することを提案した．ここに，F_1, F_2 は各治療法が有効でなかった回数であり，r, s は(1)が満たされるようにあ

表 1　$P^*=0.95$, $\Delta^*=0.2$ の場合のサンプル数の期待値

$E\{N_B|\Delta=0.2\}$

$\bar{p}$	R_1(PW)	R_1(VT)	R_2	R_3	$R_4(2)$	$R_4(3)$	$R_4(4)$	R_5
.1	44.5	20.0	20.9	13.1	33.5	31.5	29.5	19.8
.2	39.2	19.8	22.4	14.0	30.1	28.4	26.7	19.7
.3	34.0	19.2	23.1	14.1	26.6	25.2	23.5	19.1
.4	28.6	18.7	22.5	15.3	22.7	21.6	20.2	18.0
.5	23.1	18.5	20.5	16.8	18.8	18.0	16.9	16.1
.6	17.5	18.7	16.9	15.5	14.7	14.4	13.7	13.6
.7	12.2	19.2	12.2	11.7	10.6	10.5	10.5	10.5
.8	7.1	19.8	7.1	7.1	6.5	6.7	6.8	6.8
.9	2.3	20.0	2.3	2.4	2.3	2.3	2.4	2.4
				$E\{N\|\Delta=0.2\}$				
.1	100.0	40.0	47.0	28.9	74.7	70.2	65.8	44.2
.2	89.5	39.6	51.1	31.6	68.2	64.3	60.3	44.5
.3	78.9	38.4	53.5	32.1	61.3	57.9	54.1	44.0
.4	68.1	37.4	53.4	36.5	53.7	50.9	47.5	42.2
.5	56.9	37.0	50.2	41.0	45.8	43.8	41.1	39.2
.6	45.7	37.4	43.6	40.2	37.8	36.8	35.1	34.7
.7	34.7	38.4	34.2	34.0	29.6	29.7	29.2	29.2
.8	24.2	39.6	23.9	23.9	21.4	22.2	22.6	22.6
.9	14.2	40.0	14.2	14.4	13.3	14.0	14.7	14.7
				$E\{N\|\Delta=0\}$				
.0	∞	∞	42.0	∞	∞	∞	∞	44.0
.1	1100.0	177.8	46.7	65.5	593.2	511.4	429.4	48.5
.2	495.0	100.0	52.4	56.2	278.6	241.8	204.8	52.1
.3	293.3	76.2	59.0	52.4	173.8	152.2	130.3	54.5
.4	192.5	66.7	65.2	56.3	121.2	107.7	93.5	55.5
.5	132.0	64.0	68.9	73.3	89.1	81.0	71.7	54.6
.6	91.7	66.7	67.5	80.8	66.7	62.8	57.3	51.3
.7	62.9	76.2	57.9	60.4	49.2	48.7	46.5	45.2
.8	41.3	100.0	41.1	41.6	34.6	36.0	36.5	36.2
.9	24.4	177.8	24.5	24.3	21.9	23.6	25.2	25.1
1.0	11.0	∞	11.0	11.2	11.0	12.0	13.0	13.0

(注)　$\bar{p}=(p_1+p_2)/2$, $\Delta=|p_1-p_2|$.
R_1(PW)は $r=11$, R_1(VT)は $r=4$. R_2 は $r=11$, $s=42$. R_3 は $r=11$, $t=4.2$.
R_4 はいずれも $r=8$ で, (　)内は u の値. R_5 は $r=8$, $u=4$, $s=44$.

らかじめ定めておく整数である. 最終判定は前述のものと同じであるが, $S_1=S_2$ の場合には, どちらか一方をランダムに選べばよい.

ルール R_1 を用いた場合には, $|p_1-p_2|\geq\Delta^*$ の範囲内における P{CS}の最小値は, $|p_1-p_2|=\Delta^*$ 上で $p_1=1$ あるいは $p_2=1$ にきわめて近いところにある. そしてこの近傍の (p_1, p_2) に対しては, ルール R_2 を用いたとしても, ほとんど確実に (4a) の条件のほうが最初に成立して実験を終了することになり, 条件 (4b) を追加したことによる min P{CS}の変化は事実上無視し得る程度に小さい. したがって, R_2 の r は R_1 の r と同じにとればたいていの場合十分であり, このとき条件 (4b)があることによって R_2 は R_1 より一様に良

表 2　R_2 のパラメタの値

P^*	Δ^*			
	.1		.2	
	r	s	r	s
.90	17	98	8	26
.95	23	182	11	42
.99	38	322	18	82

表 3　R_3 のパラメタ t の値

P^*	Δ^*	
	.1	.2
.90	6.5	3.1
.95	9.0	4.2
.99	14.0	6.7

くなる（すなわち，任意の(p_1, p_2)に対して，R_2の$E(N_B)$, $E(N)$がR_1のそれらより小さくなる）というメリットがある．$P^*=0.95$, $\Delta^*=0.2$に対しては，$r=11$, $s=42$ととればよく，この場合のサンプル数の期待値は表1の第3列に示すとおりである．R_1に比べていちじるしく改良されていることが読みとれる．なお，いくつかのP^*, Δ^*に対するr, s（簡単のために偶数に限った）の値を表2に示しておく．

p_1, p_2が小さい場合のルールR_1の欠点を除くために，Nordbrock[4]はつぎのストッピング・ルールを考案した．

$$|S_1-S_2|=r \qquad (5\text{a})$$

R_3：　または

$$\left|\frac{S_1}{S_1+F_1}-\frac{S_2}{S_2+F_2}\right|\geq\frac{t}{F_1+F_2} \qquad (5\text{b})$$

ここに，rおよびtは確率に関する要請(1)を満たすように定める．この方式は，p_1, p_2が小さい場合には治療が有効でなかった回数F_1+F_2が急速に増大するので，(5a)よりも(5b)の条件のほうが先に満たされることによって実験が終了し，サンプル数が小さくなることをねらったものである．

R_3においても，R_2のところで述べたのと同様に，たいていの場合(5a)のrはR_1のrと同じもので十分であり，したがってR_3はR_1よりも一様に良いことになる．tの値は表3に示しておいた．$P^*=0.95$, $\Delta^*=0.2$に対しては，$r=11$, $t=4.2$とすればよく，このときのサンプル数は表1の第4列に示すようになる．

R_2とR_3を比較すると，$\Delta=0.2$の場合には$E\{N_B\}$, $E\{N\}$ともほぼ一様にR_3のほうが小さい．しかし，$\Delta=0$（したがってまたその近く）での$E\{N\}$および$E\{N_B\}$は一部分を除いてR_2のほうが小さい．

ところで，VTルールが2種類の治療法を常に対にして使うという意味で，“対称な”サンプリング・ルールであるのに対して，PWルールは“非対称な”サンプリング・ルールであるといえる．したがってPWルールとともに用いるストッピング・ルールもまた，これに見合った非対称性をもつのが自然なように思われる．すなわち，最初の患者に施すためにランダムに選ばれた治療法をⅠ，他の治療法をⅡとすると，(PWルールを用いて何人かの患者を治療した後で)Ⅱの治療法が有効でなかった時点における$S_Ⅰ-S_Ⅱ$は$p_Ⅰ-p_Ⅱ$に対する“公平な”尺度であるが，Ⅰの治療法を使用している間の$S_Ⅰ-S_Ⅱ$は公平な尺度ではなく，$p_Ⅰ$のほうを過大に評価していると考えられる．それ故，Ⅰを使用している間の$S_Ⅰ-S_Ⅱ$に対する限界値は，Ⅱを使用している間の$S_Ⅱ-S_Ⅰ$に対する限界値より大きく取るのがよいと思われる．そして，そのような型のストッピング・ルールは，PWルールとWaldの逐次確率比検定(Sequential Probability Ratio Test)との関係を考えると，つぎのようにしてごく自然に導き出せる[3]．

PWルールを用いた場合，(ストッピング・ルールを無視すると)同一の治療法が続けて成功する回数は幾何分布をする．すなわち，

$$S_Ⅰ-S_Ⅱ=x_1-y_1+x_2-y_2+\cdots$$

で，$x_1, x_2, \cdots$および$y_1, y_2, \cdots$は二つの独立な幾何分布

$$P\{X=k\}=p_1{}^k(1-p_1),\ P\{Y=k\}=p_{\text{II}}{}^k(1-p_{\text{II}})$$
$$(k=0,1,2,\cdots)$$

からの独立な標本とみなせる．一方，p_1, p_2 が既知であると仮定して，二つの単純仮説

$$H_0: P\{X=k\}=p_1{}^k(1-p_1),\quad P\{Y=k\}=p_2{}^k(1-p_2)$$
$$H_1: P\{X=k\}=p_2{}^k(1-p_2),\quad P\{Y=k\}=p_1{}^k(1-p_1)$$
$$(k=0,1,2,\cdots)$$

に対する逐次確率比検定を考えよう．X, Y についての n 個ずつの標本 $x_1, y_1, x_2, y_2, \cdots, x_n, y_n$ を観測した時点における対数尤度比は，

$$\log\left[\prod_{j=1}^{n}\{p_2{}^{x_j}(1-p_2)p_1{}^{y_j}(1-p_1)\}/\prod_{j=1}^{n}\{p_1{}^{x_j}(1-p_1)p_2{}^{y_j}(1-p_2)\}\right]=\sum_{j=1}^{n}(x_j-y_j)\log(p_2/p_1)$$

であり，$(2n+1)$番目の標本 x_{n+1} まで観測した時点における対数尤度比は，

$$\{\sum_{j=1}^{n}(x_j-y_j)+x_{n+1}\}\log(p_2/p_1)+\log\{(1-p_2)/(1-p_1)\}$$

である．逐次確率比検定では，対数尤度比が二つの定数の間にある限りサンプリングを続けるのであるから，結局，

$$-a<\sum_{j=1}^{n}(x_j-y_j)<b \qquad (6\text{a})$$

または，

$$-a+c<\sum_{j=1}^{n}(x_j-y_j)+x_{n+1}<b+c \qquad (6\text{b})$$

が成り立つ限りサンプリングを続けることになる．ここに a, b, c は正の定数である．そして，逐次確率比検定は，H_0, H_1 の下では，あらゆる検定方式の中でサンプル数の期待値が最小な方式であるから，PW ルールに対して(6)式の形のストッピング・ルールを用いれば，劣った治療法を施される患者の数の期待値 $E(N_B)$ も小さくなるであろうと期待される．なお，われわれの問題においては，2種類の治療法を対等に扱っているのであるから，(6)式において $a=b$ ととるのが自然である．

以上の考察によりつぎのいずれかの条件が最初に成立した時点で実験を終了するというストッピング・ルールを提案する．

R_4:

Ⅰを使用して成功した時点で，
$$S_{\text{I}}-S_{\text{II}}\geq r+u \qquad (7\text{a})$$
Ⅰを使用して失敗した時点で，
$$S_{\text{I}}-S_{\text{II}}\leq -r+u \qquad (7\text{b})$$
Ⅱを使用して成功した時点で，
$$S_{\text{I}}-S_{\text{II}}\leq -r \qquad (7\text{c})$$
Ⅱを使用して失敗した時点で，
$$S_{\text{I}}-S_{\text{II}}\geq r \qquad (7\text{d})$$

ここに，r と u は整数で，確率に関する要請(1)を満たすようにあらかじめ定めておく．また，最終判定のルールは R_1 と同じである．

P^*, Δ^* を定めたとき，確率の要請を満たす r, u は一般に一意的には定まらない．たとえば，$P^*=0.95$, $\Delta^*=0.2$ の場合には，$(r,u)=(9,2)$, $(9,3)$, $(9,4)$のいずれも要請を満たす．そして，表1に示すこれらのサンプル数を相互に比較してみると，どれも他より一様に良いということはない．そこで，もしいずれか一つを“最適な”ものとして選ぼうとするならば，なんらかの基準を導入しなければならないが，ここでは，他の著者達もよく用いているミニマックス基準を用いることにする．すなわち，$|p_1-p_2|\geq\Delta^*$ の範囲内で $E\{N_B\}$ の最大値を最小にする (r, u) を選ぶことにする．いくつかの P^*, Δ^* に対してこれを示したものが表4である．

表1に示す R_4 のサンプル数を見ると，p_1, p_2 の小さいところでは，R_1(PW)よりはかなり良いものの，R_1(VT)に比べるとなお劣っている．そこで，R_1(PW) の欠点を，F_1+F_2 が大きくなった時サンプリングを打切ることによって除いたのと同じ工夫をここでもしてみよう．

R_5:
$$R_4 \text{ の条件のいずれか または } F_1+F_2=s \qquad (8)$$

表 4　R_4 のパラメタの値

P^*	Δ^*			
	.1		.2	
	r	u	r	u
.90	15	5	6	2
.95	21	6	9	4
.99	35	7	15	4

が初めて成立した時点でサンプリングを終了する．sは，R_2 の s と同じ値にとって事実上さしつかえない．$P^*=0.95$，$\Delta^*=0.2$に対して，$r=8$，$u=4$，$s=44$($s=42$ とすると，$\min \mathrm{P}\{\mathrm{CS}\}$ が0.95をわずかに下まわる．)とした場合のサンプル数を表1に示す．これはもちろん $R_4(u=4)$ に比べて一様に良くなっている．そしてまた，R_2 に比べてもほぼ一様に良くなっていると見てよいであろう．最後に R_5 と R_3 との比較であるが，$\Delta=0.2$ では，p_1，p_2 が大きいところでは R_5，小さいところでは R_3 のほうがすぐれている．また p_1 と p_2 がほぼ等しいところでは，R_5 のほうがほぼ一様に良いと言える．

5. おわりに

臨床実験のためのストッピング・ルールとしては，きわめて多数のものが発表されているが，本稿では紙数の制限のため，比較的単純でしかも割合よいものをいくつか選んで解説した．また最適なパラメタの選び方等の数式に関する詳細もいっさい省略したが，とくに興味をもたれる読者は参考文献を参照していただきたい．

本稿で述べたストッピング・ルールの中では，R_2, R_3, R_5 が比較的良い方法であるといえよう．しかし，いずれもあらゆる p_1, p_2 について他より一様に良いというわけではない．したがって，このような方式を実際に使おうとすると，いったいどれを使ったらよいかという疑問がとうぜん出てくるであろう．これに対するひとつの解答は，事前情報の利用である．すなわち，臨床実験開始に先立って行なわれた動物実験の結果やその他の知識により，われわれは p_1, p_2 の値について多分おおよその見当をつけられるであろう．そこで，そのような範囲内の p_1, p_2 に対して一番良い方式を選ぶことにすれば，実際上は良いであろう．これはいわばベイジアン的な考え方である．(本特集の竹内氏の稿，314〜5ページ参照) もちろん，何らかの意味で（厳密に）最適な方式を求めるためには，p_1, p_2 に関する事前情報を確率分布の形で表現して議論をしなければならない．しかし，そのようにして求められる最適方式は，理論的には興味があっても，おそらく複雑で実用には向かないであろうから，ここでは論じないことにしよう．

参考文献

[1] Armitage, P.: *Sequential Medical Trials*. Blackwell Scientific Publications, Oxford, 1960. (佐久間 昭訳：「医学における逐次実験法」，東京大学出版会，1967.)

[2] Fushimi, M.: An Improved Version of a Sobel-Weiss Play-the-Winner Procedure for Selecting the Better of Two Binomial Populations. *Biometrika*, Vol. 60(1973), 517-523.

[3] Fushimi, M.: in preparation.

[4] Nordbrock, E.: An Improved Play-the-Winner Sampling Procedure for Selecting the Better of Two Binomial Populations. *J. Amer. Stat. Assoc.*, Vol. 71(1976), 137-139.

[5] Robbins, H.: A Sequential Decision Problem with a Finite Memory. *Proc. National Academy of Sciences*, Vol. 42(1956), 920-923.

[6] Sobel, M. & Weiss, G. H.: Play-the-Winner Sampling for Selecting the Better of Two Binomial Populations. *Biometrika*, Vol. 57(1970), 357-365.

[7] Wald, A.: *Sequential Analysis*. John Wiley and Sons, New York, 1947.

[8] Zelen, M.: Play-the-Winner Rule and the Controlled Clinical Trial. *J. Amer. Stat. Assoc.*, Vol. 64(1969), 131-146.

(ふしみ・まさのり　東京大学工学部)