バルクライン生産費の標本推定について

三枝義清

一、問題

ここで採り上げる問題を要約すると、〃今ある年度の米販売農家について米の生産費の累積分布を考える。累積相対頻度 P×100% を与える生産費が所要の P×100% のバルクライン生産費であるとした時、その推計方法は如何（問題一）。更に米の生産費の累積分布を戸数ではなくて、農家毎の販売量で累積した場合、累積相対比率 R×100%を与える生産費の推計方法は如何（問題二）。〃

このように限定してしまえば、純然たる標本調査上の問題になつて平均値と同じ母集団分布の location parametre の推計に過ぎなくなる。然し以下で述べるように平均値の推定と比べて、いろいろの点で異つた処理がいるようである。先ず第一の問題を検討してから、三節で第二の問題に触れたい。

〔一・一〕 母集団分布（累積）を $F(x)$、その密度函数を $f(x)$ とした時、

$$P=\int_{-\infty}^{\eta} f(x)\,dx$$

を満足する η を母集団分布の P×100% point と云うことにして η_P と記す。P=0.5 にとれば $\eta_{.5}$ はメディアンになる。

今大いさnの独立標本を測定値xの大いさの順序に配列したものを、

$$x_1<x_2<\cdots\cdots<x_n$$

と記すとx_iの分布に関しては既にいろいろな結果が導出されている。必要な事項を簡単に要約しておくと、

(イ)　$u_r=F(x_r)$　の密度函数は、

$$\frac{n!}{(r-1)!(n-r)!}u_r^{r-1}(1-u_r)^{n-r}\qquad 但し\ n!=1.2\cdots\cdots n$$

即ち、母集団分布$F_{(x)}$には関係なく、

$$E(u_r)=\frac{r}{n+1},\quad \sigma^2(u_r)=\frac{r(n-r+1)}{(n+1)^2(n+2)}$$

なるベータ分布に従う。

(ロ)　x_rは一般には偏りをもつが、母集団分布の $\frac{r}{n+1}\times100\%$ point の推定値である。nが充分に大きければこの偏りは無視しうる。

且つ(イ)からx_rの平均平方誤差（M.S.E）は

$$M.S.E(x_r)\doteqdot\sigma^2(u_r)/f^2\{E(u_r)\}$$

である。

Pの大いさが与えられたとして、標本における P×100% point ($\hat{\eta}_P$)を次の如く定義すると、

$$\hat{\eta}_P=x_{[nP]+1},\ \cdots\cdots\cdots\quad\cdots\cdots\ nP\ が整数でない場合$$

$$\hat{\eta}_P=\frac{1}{2}\left\{x_{nP}+x_{nP+1}\right\}\cdots\cdots\cdots nPが整数の場合$$

但し，〔nP〕は nP よりも小なる整数のうちで，nP に最も近いもの

$\hat{\eta}_P$ は漸近的に、

平均，　η_P.　　標準偏差，　$\sqrt{\dfrac{P(1-P)}{n}}\Big/ f(\eta_P)$.

なる正規分布に従うことが知られている。

(II)　η_P の信頼区間は次の如く母集団分布 $F(x)$ に関係なく設定することが出来る。

今 $s>r$ とすれば

$$P_r\left\{x_r<\eta_P<x_s\right\}=P_r\left\{u_r<P<u_s\right\}$$

$$=I_{P(r,n-r+1)}-I_{P(s,n-r+1)}$$

但し，

$$I_{P(r,n-r+1)}=\frac{\Gamma(n+1)}{\Gamma(r)\Gamma(n-r+1)}\int_0^P u^{r-1}(1-u)^{n-r}du$$

で不完全ベータ表から計算できる。(1)

なおuの分布を用いなくても直接、二項分布を用いて求めることも出来る。(2)

n個の標本のうちその測定値xが η_P よりも小になるものの個数を n_P とすれば、n_P は確準変数で、

$$P_r(x_r<\eta_P<x_s)=P_r(r\leqq n_P<s)$$

$$=\sum_{i=r}^{s-1}\binom{n}{i}P^i(1-P)^{n-i}$$

となる。

標本の大いさnが大くなるとη_Pは漸近的に正規分布（平均 nP，標準誤差$\sqrt{nP(1-P)}$）に近接してくるから、次のようにしても信頼区間を設定することができる。[3]

次節以下の論議はすべてこの方法（Woodruff の方法とよぶことにする）によるから、詳しく述べる（第一図を参照）。

第一図の$\hat{F}_{(x)}$は標本から作られるところの累積分布。実際には$x_{i\,(i=1,2,\cdots n)}$でジャンプする段階函数であるが、見易くするために滑めらかにしてある。問題によつてPが決まるから、Pを中心にして$\pm k\sqrt{\dfrac{P(1-P)}{n}}$のところに、A、B線を横軸に平行に引く。$\hat{F}_{(x)}$との交点の横座標を$x_A$並び$x_B$とする。今

$$\hat{P}\equiv\hat{F}(\eta_P)=n_P/n$$

とすると、$\hat{P}$はn個の標本値のうちで、η_Pよりも小くなるものの割合である。第一図から知れるように、$\hat{P}$がAとBの間に落ちればη_Pはその時に限つて必ずx_Aとx_Bの間に入るのであるから、次の等式(1)が成立することになる。

$$P_r\left\{x_A<\eta_P<x_B\right\}=P_r\left\{A<\hat{P}<B\right\}\cdots\cdots\cdots\cdots\cdots\cdots(1)$$

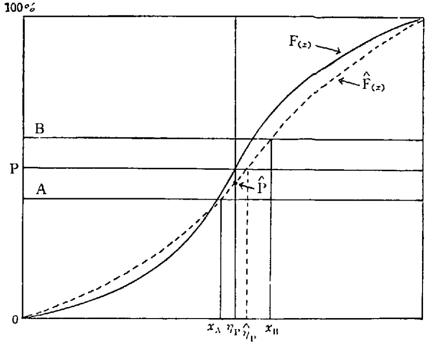
$$A=P-k\sqrt{\frac{P(1-P)}{n}}$$

$$B=P+k\sqrt{\frac{P(1-P)}{n}}$$

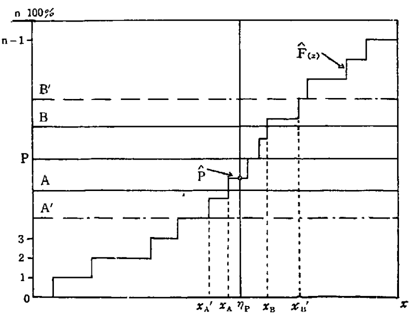
従つて、例えばk=1.96にとれば

$$P_1\{x_A < \eta_P < x_B\} \doteqdot 0.95 \cdots\cdots(1)'$$

となつて η_P に対する95%の信頼区間が設立できるわけである。なお、(1)′式の x_A、x_B は標本における P×100% point $\hat{\eta}_P$ の左右に 1 96$\sqrt{nP(1-P)}$ 個（実際には〔1 96$\sqrt{nP(1-P)}$〕個）だけ数えて得られる標本値をとることと同じことになつている。第二図は x_A、x_B の求め方をやや詳しく図式化したものである。説明は要しないと思うが、特にA、



第 1 図



第 2 図

B両線が段階函数 $\hat{F}_{(x)}$ の水平部分に一致した場合の横座標の読み方をA′、B′として区別してある。

註(1) S S. Wilks, Ordered Statistics, *Bul Amer, Math Soc* Vol. 54 (1948), 14.

(2) A. M. Mood, *Introduction to the Theory of Statistics*, 388～89.

(3) R. S. Woodruff, Confidence Intervals for Medians and other Position Measures, *Jour. Amer. Stat. Soc*

〔I・二〕 今(1)式に注目してみると、この式は $F_{(x)}$, 並び $\hat{F}_{(x)}$ が単調な非減少函数であれば成立するのであつて、母集団分布には無関係になつている。又前節では任意標本が単純な任意標本であるとしてきたが、(1)式にはこの限定は不必要である。確率標本であれば恒に成立する。単純な任意標本であれば $\hat{F}$ が第二図の如く $^1/n$ ずつのジャンプをもつ段階函数となり、$\hat{P}$ の分散 $V(\hat{P})$ が

$$P(1-P)/n \qquad \text{或いは} \quad \frac{(N-n)}{(N-1)n}P(1-P) \qquad \text{(反覆なしの抽出)}$$

となつて、既知になる。Nは母集団の大いさ。単純な抽出以外の抽出設計を用いた場合への拡張は次のように行つてゆけばよい。

(イ) $E(\hat{P})=P$ であるのが望ましいのであるから、$\hat{F}_{(x)}$ を構成するに当つて、その時の抽出設計に応じて標本の各抽出単位を適当に加重する。即ち確率標本を $x_1, x_2, \ldots\ldots x_j, \ldots\ldots x_n$（ここの j は大いさの順序でなくて、抽出における順序を表すだけである）とすれば

$$\hat{P}=\sum_{j=1}^{n}\omega_j\delta_j, \quad E(\hat{P})=P \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots (1)$$

$$\text{但し} \begin{cases} x_j < \eta_P & \text{であれば} \quad \delta_j=1. \\ x_j \geqq \eta_P & \quad〃 \qquad \delta_j=0. \end{cases}$$

なるようにウェイト $\omega_j (j=1,2, \ldots n)$ を定めて $\hat{F}(x)$ を集計せねばならぬ。この問題は恰も平均値を推計する際の推定式の構成に対応するものと考えればよい。

(ロ) 以上のように $\hat{F}(x)$ が構成されれば $V(\hat{P})$ の導出は容易である。これまでの標本調査論で得られている分散式がそのまま応用できる。

(ハ) 単純な任意抽出以外の抽出設計に拡張してゆく際の困難な問題は $V(\hat{P})$ の推計に関してである。単純な任意抽出の時の $V(\hat{P})$ はPが指定されれば確実に既知であつたが、次節に見られるようにその他の設計ではPを指定しただけでは未知なパラメターが分散式のなかに現れてくる。この点は平均の推定の場合も同じであつたが、大半の場合標本から不偏推定値を求めることが可能であつた。今の場合でも標本からの推定値で代用するわけであるが、このようにして得られた分散の推定値の期待値を導出することが困難になつてくる。

(ニ) $V(\hat{P})$ の推定値が求められれば後は単純な任意抽出の時と同じように、Pの周りに $\pm k\sqrt{V(\hat{P})}$ の巾をとつてA、B線を引き、$\hat{F}(x)$ から x_A, x_B を読めばよいのである。kを適当に選らべば所要の信頼区間が設定できよう。

以下これらの点を代表的な抽出設計について解説してゆく。

二、Woodruffの方法の応用

〔二・一〕 複雑な抽出設計への適用。今母集団の大いさがNでK個の層から成立つものとする。$N=\sum_{k=1}^{K} N_k$, k層か

ら大いさn_kの任意標本を抽出する(反覆なし)。$\sum_{k=1}^{K} n_k = n$. 標本値をx_{kj}と書く。$\hat{F}_{(x)}$をx_{kj}で$\frac{N_k}{N}\cdot\frac{1}{n_k}$だけジャンプする段階函数にとれば

$$\hat{P} = \sum_{k=1}^{K} \frac{N_k}{N} \cdot \frac{1}{n_k} \sum_{j=1}^{n_k} \delta_{kj} \cdots\cdots\cdots\cdots (2)$$

但し $\begin{cases} x_{kj} \geqq \eta_P & \text{ならば} \quad \delta_{kj}=0 \\ x_{kj} < \eta_P & \text{〃} \quad \delta_{kj}=1 \end{cases}$

$E(\hat{P})=P$ は明らか。$\hat{\eta}_P$ は〔一・一〕節の(ハ)によつて求められる。或いは第二図で$P=\hat{F}_{(x)}$なるような横座標を求めても同じことである。

$$V(\hat{P}) = \sum_{k=1}^{K} \left(\frac{N_k}{N}\right)^2 \frac{N_k - n_k}{(N_k-1)n_k} P_k(1-P_k) \cdots\cdots\cdots\cdots (3)$$

但し　$P_k = \sum_{j=1}^{N_k} \delta_{kj} \Big/ N_k$, 即ちk層でxの値が$\eta_P$よりも小なるものの割合

Pは既知であるが、$P_{k\,(k=1,2,\cdots K)}$は未知である。従つてη_Pの代りに標本におけるP×100% point $\hat{\eta}_P$を代用することとなる。従つて分散の推定式$\hat{V}(\hat{P})$は

$$\hat{V}(\hat{P}) = \sum_{k=1}^{K} \left(\frac{N_k}{N}\right)^2 \frac{N_k - n_k}{(N_k-1)n_k} \hat{P}_k(1-\hat{P}_k) \cdots\cdots\cdots\cdots (4)$$

但し$\hat{P}_k$はn_k個の標本値のうち$\hat{\eta}_P$よりも小なるものの割合

集落抽出の場合も同様である。m個の集落が単純に任意抽出されたものとする。各集落の大いさを$N_{l\,(l=1,2\cdots m)}$。$\hat{F}_{(x)}$

の集計は前述の単純任意抽出と同じで $1/n$ ずつのジャンプをもつ $\left(n=\sum_{i=1}^{m} N_i\right)$

$$\hat{P}=\sum_{i=1}^{m}\sum_{j=1}^{N_i}\delta_{ij}\Big/\sum_{i=1}^{m}N_i \quad\cdots\cdots(5)$$

δ_{ij} の定義は層化の時に準ずる

$\hat{P}$ は比推定で、一般には不偏でない。

$$E(\hat{P})\doteqdot P$$

$$V(\hat{P})\doteqdot\left(\frac{M-m}{Mm}\right)\cdot\frac{1}{M-1}\sum_{i=1}^{M}\left(\frac{N_i}{\bar{N}}\right)^2(P_i-P)^2 \quad\cdots\cdots(6)$$

M；母集団の集落総数

$P_i=\sum_{j=1}^{N_i}\delta_{ij}\Big/N_i$，$i$ 集落内の N_i 個の単位のうち，x_{ij} が η_P よりも小なるものの割合

$\bar{N}=N/M$

従つて

$$\hat{V}(\hat{P})\doteqdot\left(\frac{M-m}{Mm}\right)\cdot\frac{1}{m-1}\sum_{i=1}^{m}\left(\frac{N_i}{\bar{N}}\right)^2(\hat{P}_i-P)^2 \quad\cdots\cdots(7)$$

$\hat{P}_i$，i 集落内の単位のうち，$\hat{\eta}_P$ よりも小なるものの割合

$\hat{F}(x)$ の集計には母集団の各単位の選出確率が既知であればよいのであるが、実際の集計作業としては選出確率を均等にしておく方が便利である。例えば層化抽出の場合、$n_k \propto N_k$ にしておけば、$\hat{F}(x)$ の集計は加重なしで済まされるか

ら、標本がいずれの層から抽出されたかと云うことに拘ることなく観測値のみに注目するだけでよい。一方抽出設計が等確率でないのを無視して加重なしで $\hat{F}(x)$ を集計したものとすれば、この場合への上記の Woodruff の方法の適用は困難になる。$\hat{F}(x)$ の集計と云う点から云えば、抽出設計としては等確率と云う条件を犠牲にすることはできないだろう。従つて米の生産費調査としては副次抽出方法によるわけだが、第一次抽出単位（たとえば臨時農業基本調査の時の農業集落）を戸数に比例する確率で抽出、そして各集落から同数の農家を任意抽出する方式（所謂確率比例抽出法）によらざるをえないだろう。階層が入れば比例割当（各層の総対称農家数に比例）を採ることになる。然し実際問題として厳密に等確率と云う条件を実現することにはかなりの問題があろう。

今、抽出設計として層別確率比例抽出法をとつたものとすれば $V(\hat{P})$ 次の如くなろう（この際の抽出方法その他については註4を参照)。

$$V(\hat{P}) \fallingdotseq \frac{1}{m}\sum_{k=1}^{K}\frac{N_k}{N}\sigma_{1k}^2+\frac{1}{m\bar{n}}\sum_{k=1}^{K}\frac{N_k}{N}\sigma_{2k}^2 \quad \cdots\cdots\cdots\cdots(8)$$

N_k; k層の総対称農家数

$\sigma_{1k}^2=\sum_{i=1}^{M_k}\frac{N_{ki}}{N_k}(P_{ki}-P_k)^2$, ($m_k$) M_k は k層における（標本）集落数 $m_k \propto N_k \cdot m=\sum_{k=1}^{K} m_k$

$\sigma_{2k}^2=\sum_{i=1}^{M_k}\frac{N_{ki}}{N_k}P_{ki}(1-P_{ki})$

$\bar{n}$; 各集落からの調査農家戸数

P_{ki}; i 集落の対称農家のうち、$x_{kij} < \eta_P$ なるものの割合

$P_k=\sum_{i=1}^{M_k}\frac{N_{ki}}{N_k}P_{ki}$

この分散式から、平均生産費推定の場合の層化の効果が $\bar{x}_k$ についての層化変動に依存するのに対して、問題(1)の場合の推定では P_k についての層間変動に依存することが分かる。P_k の層間変動は先ず η_P の大いさで決るわけだから、Pをいくらにとるかによつて、いろいろに変るわけである。仮りに P=0.5 とおけば、米の生産費分布によるわけだが、η_P は平均生産費に近接したものになろう。この場合、通常の平均生産費推定のための層化の効果が同じ程度にメディアンの推定に及ぶかどうかは分らない。又Pが変ればそれに応じて実現される層化の効果も変化してくる。

この点から考えると η_P の推定については、ある層内の各集落が厳密に統計的な意味で同質——即ち、各集落は同一母集団からの random group——になつていることが望まれるのである。云い換えればバルクライン生産費推定の場合の層化の効果の判定は難しい。層化の問題を云々するためには、先ず事後的な吟味が計画されねばならない。

分散の推定式は

$$\hat{V}(\hat{P}) \doteqdot \frac{1}{m}\sum_{k=1}^{K}\frac{N_k}{N}\left\{\frac{1}{m_k-1}\sum_{i=1}^{m_k}(\hat{P}_{ki}-\hat{P}_k)^2\right\} \cdots\cdots\cdots\cdots\cdots\cdots (9)$$

註(4) 津村善郎著『標本調査法』一一〇頁。

〔II・II〕 $V(\hat{P})$ の推定値について。η_P の代りに $\hat{\eta}_P$ を代用した時の $\hat{V}(\hat{P})$ の期待値を求めることが困難であるので、この点、Woodruff が仮想母集団について抽出実験を行うことによつて代用効果をチェックしている。参考のために掲げておこう。[5]

(イ) 層別抽出の実験。1から60までの数が次のように五個の層に層別されている。$N_k=12$.

層1° 1, 3, 5, 7, 8, 9, 10, 13, 15, 46, 48, 50.

層二° 2, 4, 6, 11, 12, 16, 21, 29, 31, 32, 40, 43.
〃三° 17, 18, 22, 25, 26, 28, 34, 36, 41, 42, 52, 56.
〃四° 14, 20, 23, 27, 30, 33, 35, 45, 49, 51, 53, 59.
〃五° 19, 24, 37, 38, 39, 44, 47, 54, 55, 57, 58, 60.

反復なしで各層から大いさ6の任意標本を30組抽出した。P＝0.5、即ちメディアンの信頼区間を設定するため$V(n\hat{P})$を30組について求め、この平均を計算した。その結果は、

$V(n\hat{P})$	3.409	真の分散値
$V(n\hat{P})$	3.383	（母集団の$\eta_{.5}$を用いて推定したもの）
$V(n\hat{P})$	3.420	（標本の$\hat{\eta}_{.5}$を用いて推定したもの）
分散推値の M.S.E	.2063	（$\eta_{.5}$を用いた場合の推定値について）
〃　〃	.1804	（$\hat{\eta}_{.5}$を用いた場合の推定値について）
二つの推定値の間の相関係数	.86	

(ロ)、集落抽出法の実験。1から80までの数が次のようにグループされている。M＝20, N_1＝4.

x_{ij} \ i	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	6	5	4	7	1	61	42	38	51	17	2	16	35	46	27	18	48	54	32	47
	15	23	9	10	3	68	45	56	67	39	8	21	43	50	33	19	49	73	59	55
	20	24	14	12	28	75	60	57	70	40	11	25	44	65	52	22	69	76	72	58
	26	31	29	13	36	79	64	62	80	41	30	34	53	66	64	37	71	78	74	77

反覆なしで大いさ10の集落標本を30組抽出した。P=0.5として$V(n\hat{P})$を求めた。30組の平均は

$V(n\hat{P})$	16.84	
$\hat{V}(n\hat{P})$	16.91	（$\eta_{.5}$を用いた場合の30組の平均）
$\hat{V}(n\hat{P})$	16.37	（$\hat{\eta}_{.5}$を用いた場合の30組の平均）
分散推定値の M.S.E	4.262	（$\eta_{.5}$を用いた場合の推定値について）
〃 〃	4.910	（$\hat{\eta}_{.5}$を用いた場合の推定値について）
二つの推定値の間の相関係数	.66	

以上の抽出実験は限定されたものであるけれども、Woodruff は$\hat{\eta}_P$を用いることにより蒙る偏りは小で、二組の推定値の M.S.E が極めて接近したものであることを確認している。米生産費調査の場合には階層の数も増大するし又標本数も極めて大いのであるから、この代用の効果は問題にはならぬと思う。

註(5)　註(3)参照。

〔二・三〕 標本をグループにした場合。Woodruff の方法は大標本を前提にしたものである。このような場合には$\hat{F}_{(x)}$の集計は個々の観測値を用いるよりも、適当にグループにした資料を用いた方が便利である。実際バルクライン生産費の計算もこの操作をとつている。

今n個の標本が一定の巾ΔXをもつたクラスによつてグループされたものとしよう。あるクラス$(X, X+\Delta X]$において下限Xまでの累積相対頻度がH'、このクラスの相対頻度をU'とした時、

$$H'<P\leqq H'+U' \quad \cdots\cdots (9)$$

であれば、このクラスを P×100% class（標本における）と呼ぶことにしよう。この場合の標本の P×100% point $\hat{\eta}_P$ は普通 P×100% class における補間で計算することになる。即ち

$$\hat{\eta}_P=X+\left(\frac{P-H'}{U'}\right)\varDelta X=X+\theta'\varDelta X \quad \cdots\cdots (10)$$

θ' は補間の程度を示す項と云えよう。（なお第三図を参照）。今母集団分布も $\hat{F}_{(x)}$ と同じ巾 $\varDelta X$ をもつたクラスによつてグループされたものと想定して、母集団における P×100% class〔X, X+$\varDelta$X〕を

$$H<P\leqq H+U \quad \cdots\cdots (11)$$

で定める。H、およびUの定義は標本の場合に準ずる。η_P は(10)式に準じて定義されたものとする。即ち

$$\eta_P=X+\left(\frac{P-H}{U}\right)\varDelta X=X+\theta\varDelta X \quad \cdots\cdots (12)$$

以上のように諸量を定義すれば、グループされた場合の $\hat{P}$ は次式で表すことが出来る。

$$\hat{P}=\hat{P}_1+\theta\hat{P}_2 \quad \cdots\cdots (13)$$

$$=\hat{P}_1'-\theta^*\hat{P}_2 \quad \cdots\cdots (13)'$$

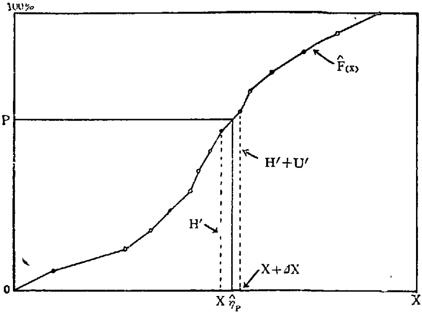
ここで

$\hat{P}_1$；標本のうち0から P×100% class の下限Xまでの値をとるものの割合

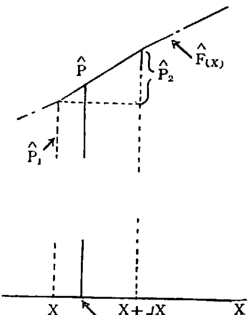
$\hat{P}_2$；〃　〃　Xから X+$\varDelta$X までの値をとるものの割合

$$\hat{P}_1' = \hat{P}_1 + \hat{P}_2, \quad \theta^* = 1-\theta$$

上式の成立することは第四図から明らかであろう。n個の標本の抽出設計が等確率でなければ、各クラスの相対頻度U′の集計はグループせぬ時と同様に適当に標本を加重して求めねばならない。たとえば層化抽出で各層への配分が比例的でなければ、クラス（X, X＋ΔX）の相対頻度U′は次のように集計せねばならぬ。



第 3 図



第 4 図

$$U'=\sum_{k}\frac{N_k}{N}\frac{1}{n_k}\sum_{j=1}^{nk}\delta_{kj}$$

但し　$X<x_{kj}\leqq X+\varDelta X$　ならば　$\delta_{kj}=1$

然らざれば　$\delta_{kj}=0$

各クラスのU′の集計を以上のように不偏に行つてゆくことにすれば

$$E(\hat{P}_1)=H,\ E(\hat{P}_2)=U,\ で\ E(\hat{P})=P$$

且つ $\hat{P}$ の分散 $V(\hat{P})$ は

$$V(\hat{P})=V(\hat{P}_1)+\theta^2V(\hat{P}_2)+2\theta COV(\hat{P}_1,\hat{P}_2)\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots(14)$$

となる。たとえばn個の標本が単純な任意標本（反覆なしの）であれば

$$V(\hat{P}_1)=\frac{(N-n)}{(N-1)n}H(1-H),\quad V(\hat{P}_2)=\frac{(N-n)}{(N-1)n}U(1-U),\quad C_0U(\hat{P}_1,\hat{P}_2)=-\frac{(N-n)}{(N-1)n}HU$$

であるから

$$V(\hat{P})=\frac{(N-n)}{(N-1)n}\left\{P(1-P)-U\theta(1-\theta)\right\}\cdots\cdots\cdots\cdots\cdots\cdots(15)$$

クラスの巾を充分小くとればグループなしの場合の分散に近づくのは当然である。H、U、θ 等は母集団における $P\times100\%$ class についての量であるから、仮りに単純な任意抽出であつても、グループなしの場合のように、既知にはならない。従つて $V(\hat{P})$ の計算には標本における $P\times100\%$ class を代用せねばならない。この場合の分散の推定式は

$$\hat{V}(\hat{P})=\frac{(N-n)}{(N-1)n}\left\{P(1-P)-U'\theta'(1-\theta')\right\}\quad\cdots\cdots(16)$$

実例（Woodruff の先掲論文より）　一九四五年センサス局で行われたラジオ聴取慣習調査によつて、第一表の結果られた。第一欄は夜間、聴取している放送局数（number of station）、第二欄は聴取局数についてのラジオ所有世帯の頻度分布。第三欄は累積相対頻度である。

第一表から明らかなように $\Delta X=1.0$。50% class, 即ちメディアン、クラスを求めると〔1.5, 2.5〕。従つて聴取局後のサンプル、メディアンは

$$\hat{\eta}_5=1.5+\left(\frac{50-44}{26}\right)\times1.0=1.7$$

第一表の $\sqrt{\hat{V}(\hat{P})}=0.02560$ を用いれば P=0.5 に対する

$$\begin{cases}1\sigma \text{ limit は } 47.44\%\sim52.56\%\\ 2\sigma \text{ limit は } 44.88\%\sim55.12\%\end{cases}$$

になる。$\hat{\eta}_P$ を求めた時と全く同様の操作を経れば、メディアンに対する

$$\begin{cases}1\sigma \text{ の信頼限界は } 1.632\sim1.829\\ 2\sigma \text{ の 〃 〃は } 1.534\sim1\ 928\end{cases}$$

第 1 表

聴取局数 (x)	U′	H′	分散の推定；
(1)	(2)	(3)	$V(\hat{P}_1)=0.0007070$
0	14	14	$V(\hat{P}_2)=0\ 0001708$
1	30	44	$C_0U(\hat{P}_1,\hat{P}_2)=-0.0001323$
2	26	70	$\theta'=0.2308$
3	19	89	$\hat{V}(\hat{P})=0.0006550$
4	8	97	$\sqrt{\hat{V}(\hat{P})}=0.02560$
5	2	99	

になる。

〔二・四〕代表的な抽出設計への拡張。〔二・一〕節と全く同様であるので、結果だけを列挙しておくと、層別抽出法の場合。

$$V(\hat{P})=\sum_{k=1}^{K}\left(\frac{N_k}{N}\right)^2\frac{(N_k-n_k)}{(N_k-1)n_k}\left\{(H_k+\theta U_k)(1-H_k-\theta U_k)-U_k\theta(1-\theta)\right\}\cdots\cdots\cdots\cdots(17)$$

但し　H_k；N_k のうち $P\times100\%$ class の下限 X よりも小なるものの割合

U_k；　〃　〃　$P\times100\%$ class $(X, X+\varDelta X]$ に含れるものの割合

集落抽出法の場合。

$$V(\hat{P})\fallingdotseq\frac{M-m}{M\,m}\cdot\frac{1}{M-1}\sum_{i=1}^{M}\left(\frac{N_i}{\bar{N}}\right)^2(H_i+\theta U_i-P)^2\cdots\cdots\cdots\cdots(18)$$

但し　H_i；集落 i のうち，$P\times100\%$ class の下限 X よりも小なるものの割合

U_i；集落 i のうち，$P\times100\%$ class $(X, X+\varDelta X]$ に含まれるものの割合

層別確率比例抽出法の場合 $(m_k\propto N_k)$。

$$V(\hat{P})\fallingdotseq\frac{1}{m}\sum_{k}\sum_{i}\frac{N_{ki}}{N}\left\{(H_{ki}-H_k)+\theta(U_{ki}-U_k)\right\}^2$$
$$+\frac{1}{m\bar{n}}\sum_{k}\sum_{i}\frac{N_{ki}}{N}\left\{(H_{ki}+\theta U_{ki})(1-H_{ki}-\theta U_{ki})-U_{ki}\theta(1-\theta)\right\}^2\cdots\cdots\cdots\cdots(19)$$

クラスの巾を充分小くとれば $H_{ki}+\theta U_{ki}\rightarrow P_{ki}$ となつて(8)式がえられることが分かる。

分散の推定式 $\hat{V}(\hat{P})$ は

$$\hat{V}(\hat{P}) \doteqdot \frac{1}{m}\sum_k \frac{N_k}{N}\cdot\frac{1}{m_k-1}\sum_{i=1}^{m_k}\left\{(\hat{H}_{ki}-\hat{H}_k)+\theta(\hat{U}_{ki}-\hat{U}_k)\right\}^2 \cdots\cdots (20)$$

$$=\frac{1}{m}\sum_k \frac{N_k}{N}\cdot\frac{1}{m_k-1}\sum_{i=1}^{m_k}\left\{(\hat{H}_{ki}'-\hat{H}_k')-\theta^*(\hat{U}_{ki}-\hat{U}_k)\right\}^2 \cdots\cdots (20)'$$

但し $\hat{H}_{ki}$ $(\hat{U}_{ki})$； k戸から抽出された集落 i の $\bar{n}$ 戸の調査農家のうち，標本における P×100 class の下限よりも小なるものの割合 $\{(X, X+\varDelta X)$ に含まれるものの割合$\}$

$\hat{H}_k=\frac{1}{m_k}\sum_{i=1}^{m_k}\hat{H}_{ki}$， $\hat{U}_k=\frac{1}{m_k}\sum_{i=1}^{m_k}\hat{U}_{ki}$　実際には θ の代りに θ' を代用することになる

$\hat{H}_{ki}'=\hat{H}_{ki}+\hat{U}_{ki}$， $\hat{H}_k'=\hat{H}_k+\hat{U}_k$

一般には P×100% class の U′ は小であろうから，分散の主要項は $(\hat{H}_{ki}-\hat{H}_k)$ の変動であろう。従つて第二項を無視して第一項の変動だけで近似することが考えられる。このような近似を行う場合には θ の値の小なる方が望ましいわけだから，もし $\theta>\theta^*$ であれば (20)′ 式の方を用いて近似式として次式によるべきであろう。

$$\hat{V}(\hat{P}) \doteqdot \frac{1}{m}\sum_k \frac{N_k}{N}\cdot\frac{1}{m_k-1}\sum_{i=1}^{m_k}(\hat{H}_{ki}'-\hat{H}_k')^2 \cdots\cdots (21)$$

実例　先掲のラジオ聴取慣習調査の例で，標本値が次の如く，グループされていたものとする。

累積頻度(H′)	10	20	35	55	70	100
聴取局数	0	1	2	3	4	5

この場合のメディアン，クラスは〔2 5, 3.5〕. 従つて

$$\theta=\frac{50-35}{20}=0.75,\quad \theta^*=0.25$$

$V(\hat{P})$ の推定を $V(\hat{P}_J)$ で近似しようとするならば $V(\hat{P}_J)$ は、標本世帯のうち聴取局が2乃至それ以下になるものの割合の分散としてではなくて、聴取局数が3乃至それ以下になるものの割合の分散として推定せねばならない。

三、販売量累積への拡張

変数yを各農家の販売量、xは前節同様、米生産費としよう。集計はグループされた資料によらざるをえないだろうから、この場合に就いてのみ検討することにする。

〔三・一〕 数量累積による R×100% バルクライン生産費の定義。二、三節の時と同じく、クラスの巾を ΔX とする。これは一定で、集計計画が決まれば確定するものとする。今あるクラス〔X, ΔX〕における相対頻度がUであったとする。販売量累積であるから、Uは次のようになる。Nは総対象農家数。

$$U=\sum_{j=1}^{N}\delta_j' y_j\Big/\sum_{j=1}^{N}y_j=\sum_{X<x_j\leqq X+\Delta X}y_j\Big/\sum_{j=1}^{N}y_j$$

但し　$X<x_j\leqq X+\Delta X$　であれば　$\delta_j'=1$

然らざれば　$\delta_j'=0$

各クラスに対して以上のようにして相対頻度が定まるから、クラス〔X, X+ΔX〕の下限までの累積頻度も定まる。Hと記す。若し

$$H < R \leqq H + U \quad \cdots\cdots (22)$$

であればクラス〔X, X＋⊿X〕を母集団における R×100% class と呼ぶことにする。二、三節の時と同様にして、R×100% のバルクライン生産費を次のように定義する。

$$\eta_R \equiv X + \left(\frac{R-H}{U}\right)\varDelta X \quad \cdots\cdots (23)$$

$$= X + \varphi \varDelta X$$

〔X, X＋⊿X〕は R×100% class である

従つて η_R はクラスの巾 ⊿X のとり方によつて異つてくるわけである。

今 n 個の標本が単純な任意抽出によつて選ばれた場合を考えると、標本におけるあるクラス〔X, X＋⊿X〕の相対頻度 U′ は

$$U' = \sum_{j=1}^{n} \delta_j' y_j \Big/ \sum_{j=1}^{n} y_j = \sum_{X < x_j \leqq \varDelta X} y_j \Big/ \sum_{j=1}^{n} y_j$$

で求められる。母集団の時と同じように、$H' < R \leqq H' + U'$（H′ はクラスの下限までの累積頻度）なる如く R×100% class を定めて、補間すれば η_R の推定値が求まる。

$$\hat{\eta}_R = X + \left(\frac{R-H'}{U'}\right)\varDelta X \quad \cdots\cdots (24)$$

$$= X + \varphi' \varDelta X$$

〔X, X＋⊿X〕は標本における R×100% class

〔三・二〕 $\hat{\eta}_R$の信頼区間。前節の要領で、標本における累積曲線がえられるから、η_Rの縦座標を$\hat{R}$とおけば二、三節の$\hat{P}$と全く同様に（第三並び四図参照）

$$\hat{R}=\hat{R}_1+\varphi\hat{R}_2 \quad \cdots\cdots(25)$$

但し　$\hat{R}_1=\sum_{j=1}^{n}\delta_j y_j \Big/ \sum_{j=1}^{n} y_j$，　$x_j\leqq X$ の時に限り $\delta_j=1$.　$(X, X+\Delta X)$ は母集団における R×100% class.

$\hat{R}_2=\sum_{j=1}^{n}\delta'_j y_j \Big/ \sum_{j=1}^{n} y_j$，　$X<x_j\leqq X+\Delta X$ の時に限り $\delta'_j=1$

であることが分かる。〔一・一〕節の(1)式

$$P_r\{x_A<\eta_R<x_B\}=P_R\{A<\hat{R}<B\}$$

が成立することは明らかである。然しここで注意すべき点は、$\hat{R}_1$, $\hat{R}_2$ が比推定であつて、一般には母集団におけるR×100% class のH、並びUの不偏推定にはならぬことである、従つて $E(\hat{R})\neq R$。然し我々の場合にはこの種の偏りは無視して差支えないであろう。$E(\hat{R})=H+\theta U=R$ とすれば、残る問題は $V(\hat{R})$ の導出とその推定だけである。$V(\hat{R})$ が求められれば戸数累積の時と全く同様な手順で Woodruff の方法を適用して、所要の信頼区間を設定できるわけである。$\hat{R}$の分散は一般的に次のように書ける。

$$V(\hat{R})=V(\hat{R}_1)+\varphi^2V(\hat{R}_2)+2\varphi C_oV(\hat{R}_1, \hat{R}_2) \quad \cdots\cdots(26)$$

これを戸数累積の場合の(15)式と比較しうる形に書改めると、

$$V(\hat{R}) \fallingdotseq \frac{N-n}{(N-1)n} \cdot (1+C_y^2) \cdot \left\{ (G_1+\varphi G_2)(1-G_1-\varphi G_2)+(G_1+\varphi G_2-R)^2-G_2\varphi(1-\varphi) \right\} \cdots \quad (27)$$

但し　C_y；yの変動係数

$G_1 \fallingdotseq \sum_{j=1}^{N} \delta_j y_j^2 \Big/ \sum_{j=1}^{N} y_j^2 \fallingdotseq \sum_{x_j<X} y_j^2 \Big/ \sum_{j=1}^{N} y_j^2$，$[X, X+\Delta X]$ は $R \times 100\%$ class

$G_2 = \sum_{j=1}^{N} \delta_j' y_j^2 \Big/ \sum_{j=1}^{N} y_j^2 = \sum_{X<x<X+\Delta X} y_j^2 \Big/ \sum_{j=1}^{N} y_j^2$

クラスの巾を充分に小くとれば

$$G_1+\varphi G_2 \fallingdotseq \sum_{x_j<\eta_R} y_j^2 \Big/ \sum_{j=1}^{N} y_j^2 \equiv G$$

既に述べたように$R=H+\varphi U$であるから、一般には$R \neq G$であろう。⒂式と対比してみると戸数累積の場合の$V(\hat{P})$の主要項が$P(1-P)$であるのに対して(27)式ではC_y^2の他にy^2についての累積割合$(G_1+\varphi G_2)(1-G_1-\varphi G_2) \fallingdotseq G(1-G)$が関係してくる。場合によっては〔〕内の第二項$(G_1+\varphi G_2-R)^2 \fallingdotseq (G-R)^2$も無視しえないだろう。Rの大きさが指定された場合Gの大きさが如何程度の大いさになるかは不明である。米生産費調査の例では恐らく

$$R(1-R) > G(1-G)$$

であろうが、然し全体として〔〕内の項がどの程度になるかは標本から推定してみない限り、現在の知識では類推しえない。

分散$V(\hat{R})$の推定式は

$$\hat{V}(\hat{R})=\frac{N-n}{(N-1)n}\cdot\frac{1}{\hat{y}^2}\cdot\frac{1}{n}\sum_{j=1}^{n}\left\{(1-R)^2\sum_{x_j\leqq X}y_j{}^2+(\varphi-R)^2\sum_{X<x_j\leqq X+\varDelta X}y_j{}^2+R^2\sum_{X+\varDelta X<x_j}y_j{}^2\right\}\cdots\cdots\cdots\cdots(28)$$

但し　$\hat{y}$; 販売量の標本平均

$(X, X+\varDelta X]$ は標本における $R\times100\%$ class である。 $\varphi=\dfrac{R-H'}{U'}$

集落抽出法の場合。U' の集計式は単純な任意抽出の時と同じで、分散式は

$$V(\hat{R})=\frac{M-m}{(M-1)m}\cdot\frac{1}{\bar{y}^2}\left\{\frac{1}{M}\sum_{i=1}^{M}y_i{}^2(H_i+\varphi U_i-R)^2\right\}\cdots\cdots\cdots\cdots(29)$$

但し　$\bar{y}$; 集落当りの平均販売量

y_i ; i 集落の販売総量

H_i ; i 集落の販売総量 y_i のうち，$x_{ij}\leqq X$ なる農家の販売量の占める割合

U_i ;　〃　〃　$X<x_{ij}\leqq X+\varDelta X$ なる農家の販売量の占める割合

$(X, X+\varDelta X]$ は母集団の $R\times100\%$ class

確率比例抽出法の場合。この場合、階層を入れるとすれば、比推定に二通りの推定式があるのに対応してある、クラス $(X, X+\varDelta X]$ の U' の集計式にも二通りが考えられる。

$$U'=\sum_k\frac{N_k}{N}\cdot\frac{1}{m_k\bar{n}}\sum_{i,j}\delta'_{kij}\,y_{kij}\Big/\sum_k\frac{N_k}{N}\cdot\frac{1}{m_k\bar{n}}\sum_{i,j}y_{kij}\cdots\cdots\cdots\cdots(30)$$

$$=\sum_{k,i,j}\delta'_{kij}\,y_{kij}\Big/\sum_{k,i,j}y_{kij}\quad(m_k\propto N_k)\cdots\cdots\cdots\cdots(30)'$$

$$=\sum_{X<x_{kij}\leqq X+\varDelta X}y_{kij}\Big/\sum_{k,i,j}y_{kij}$$

$$U' = \sum_{k}^{K} \left(\frac{Y_k}{Y} \right) \cdot \frac{\sum_{i,j} \delta'_{kij}\, y_{kij}}{\sum_{i,j} y_{kij}} \quad \cdots\cdots\cdots\cdots (31)$$

但し　N_k；k 層の総対称農家数．m_k；k 層からの標本集落数，各集落より $\bar{n}$ 個の農家を任意抽出する．標本農家の販売量を y_{kij} $(i=1,2,\cdots\cdots m_k,\ j=1,2,\cdots\cdots \bar{n})$

δ_{kij}，δ'_{kij} の定義は単純な任意抽出の時に準ずる。

(31)式の方は層毎に販売総量 $\left(Y_k,\ Y=\sum_{k=1}^{K} Y_k \right)$ を必要とするし，更に加重集計になるから(30)式 ($m_k \propto N_k$ として) を用いる方がよい。(30)'式の分散は，

$$V(\hat{R}) = \frac{1}{\bar{y}^2} \sum_{k=1}^{K} \left(\frac{N_k}{N} \right)^2 \left\{ \frac{1}{m_k} \sum_{i=1}^{m_k} \frac{N_{ki}}{N_k} (\bar{z}_{ki} - \bar{z}_k)^2 + \frac{1}{m_k \bar{n}} \sum_{i=1}^{M_k} \frac{N_{ki}}{N_k} \cdot \frac{(N_{ki} - \bar{n})}{(N_{ki} - 1)} \cdot \sigma_{ki} \right\} \cdots\cdots\cdots\cdots (32)$$

但し　$z_{kij} = y_{kij}(\delta_{kij} + \varphi\delta'_{kij} - R)$

$\bar{z}_{ki} = \frac{1}{N_{ki}} \sum_{j=1}^{N_{ki}} z_{kij}$（i 集落についての平均）．　$\bar{z}_k = \sum_{k} \frac{N_{ki}}{N_k} \bar{z}_{ki} = \sum_{i,j} z_{kij} \Big/ N_k$　（k 層の平均）

$\sigma_{ki}^2 = \frac{1}{N_{ki}} \sum_{j=1}^{N_{ki}} (z_{kij} - \bar{z}_{ki})^2$

(19)式と対比し易く書改めると，

$$\sigma^2_{ki} = \bar{y}^2_{ki}(1 + C^2_{ki}) \left\{ G_{ki}(1 - G_{ki}) + (G_{ki} - R)^2 - G_{2ki}\varphi(1 - \varphi) - \frac{(H_{ki} + \varphi U_{ki} - R)^2}{(1 + C^2_{ki})} \right\}$$

但し　C_{ki}；　k 層の集落内の y_{kij} 間の変動係数

$$G_{kl}=\left(\sum_{x_{klj}\leqq X} y^2_{klj}+\varphi\sum_{X<x_{klj}\leqq X+\varDelta X} y^2_{klj}\right)\Big/\sum_{j=1}^{N_{kl}} y^2_{klj}$$

$$G_{2kl}=\sum_{X<x_{klj}\leqq X+\varDelta X} y^2_{klj}\Big/\sum_{j=1}^{N_{kl}} y^2_{klj}$$

$$H_{kl}+\varphi U_{kl}\doteqdot\left(\sum_{x_{klj}\leqq X} y_{klj}+\varphi\sum_{X<x_{klj}\leqq X+\varDelta X} y_{klj}\right)\Big/\sum_{j=1}^{N_{kl}} y^2_{klj}$$

$(X, X+\varDelta X]$; $R\times 100$ class

$$(\bar{z}_{kl}-\bar{z}_k)^2=\left\{\bar{y}_{kl}(H_{kl}+\varphi U_{kl}-R)-\bar{y}_k(H_k+\varphi U_k-R)\right\}^2$$

$$\doteqdot\left[(\bar{y}_{kl}-\bar{y}_k)(H_{kl}+\varphi U_{kl}-R)+\bar{y}_k\left\{(H_{kl}-H_k)+\varphi(U_{kl}-U_k)\right\}\right]^2$$

(19)式の第一項と対比すると、集落の平均販売量 $\bar{y}_{kl}$ の集落間変動が追加されてくることが知れる。然し $\bar{y}_{kl}$ と H_{kl} とは一般に負の相関関係にあろうから、$\bar{y}_{kl}$ の集落間変動の追加的影響がどの程度になるかを知るには、上記の相関についての資料が必要である。

分散 $V(\hat{R})$ の推定式は

$$\hat{V}(\hat{R})=\frac{1}{\hat{y}^2}\sum_k\left(\frac{N_k}{N}\right)\cdot\frac{1}{m_k(m_k-1)}\left\{\sum_{l=1}^{m_k}(\hat{z}_{kl}-\hat{z}_k)^2\right\}\cdots\cdots\cdots\cdots(33)$$

但し $\hat{z}_{kl}=\dfrac{1}{\bar{n}}\sum_{j=1}^{\bar{n}} z_{klj}=\dfrac{1}{\bar{n}}\left\{\sum_{x_{klj}\leqq X} y_{klj}+\varphi\sum_{X<x_{klj}\leqq X+\varDelta X} y_{klj}-R\sum_{j=1}^{\bar{n}} y_{klj}\right\}$

$$\hat{z}_k=\sum_{i=1}^{m_k}\sum_{j=1}^{\bar{n}} z_{kij}\Big/m_k\bar{n}$$

$$\hat{y}=\sum_k\frac{N_k}{N}\cdot\frac{1}{m\bar{n}_k}\sum_{i,j} y_{kij}$$

〔X, X+ΔX〕は標本の R×100% class.　$\varphi=\frac{R-H'}{U'}$

なお〔二・四〕節で附記したように$\hat{z}_{ki}$の計算において｛ ｝の第一項を$\sum_{x_{kij}\leqq X+\Delta X} y_{kij}$で推計することにより、$\hat{z}_{ki}$を第一項のみで代置することも可能である。

〔追記〕

問題としてはあと補助変量の利用並びに抽出確率の最適化の問題が残つているわけであるが、バルクライン生産費の推定に関しては実際問題として論議の余地がないのではないかと思われる。抽出確率については、等確率と云う条件は外せないだろうし、補助変量の利用と云う点も、平均や総計の推定の場合の利用式をそのまま融通することは難かしいであろう。

（研究員）