藉抽樣後輔助變值之假分層以推算平均值

Estimation of the Mean by Pseudo-post-stratification of an Auxiliary Variate

國立臺灣大學 魏應澤

一、緒論

利用輔助變值 z (Auxiliary Variate) 推算主要變值 y 之族群平均值 $\bar{Y}$ (Population Mean) 的方法有比率推算及迴歸推算兩種 (Ratio and Regression Estimation)。在簡單逢機抽樣 (Simple Random Sampling) 下此兩種推算之公式如次：

$$\text{比率推算：}\quad \hat{\bar{Y}}_R = \frac{\bar{y}}{\bar{z}}\bar{Z} = \hat{R}\bar{Z}$$

$$\text{迴歸推算：}\quad \bar{y}_{lr} = \bar{y} + b(\bar{Z}-\bar{z})$$

式中 $\bar{y}$ 及 $\bar{z}$ 爲樣品平均值 (Sample Mean)，$\bar{Z}$ 爲族群平均值，b 爲樣品迴歸係數。就一般而論，此兩種推算法具有偏估現象。所謂偏估 (Biased Estimation) 係指從所有可能樣品求得之推算值的總平均值不等於欲要推算之族群介量 (Parameter of the Population) 而言；亦即推算值之期望值 (Expected Value) 不等於欲要推算之族群介量。我們已知此兩種推算值之偏估量 (Bias) 爲 n^{-1} 級，n 係樣品大小 (參閱 Cochran W.G. 1963)。Mickey 氏 (1959) 曾導出有限族群 (Finite Population) 之無偏比率及迴歸推算法。Hartley 及 Ross 兩氏 (1954) 亦創造了一個無偏比率推算公式，該公式可說是 Mickey 氏所導出的一個特殊場合下之公式。

Mickey 氏曾注意到用迴歸推算式 $\bar{y}_{lr} = \bar{y} + b(\bar{Z}-\bar{z})$ 推算族群平均值 $\bar{Y}$ 時，假如式中之 b 爲任何固定常數，則勿論 y 與 z 之關係如何，$\bar{y}_{lr}$ 恒爲無偏推算式。因此，利用此事實，他將簡單逢機樣品 (Simple Random Sample) 按其被抽之順序聚首 m ($m<n$) 個單位爲一小樣品而剩餘之 $(n-m)$ 個單位則視爲從含有 $(N-m)$ 個單位之有限族群中以簡單逢機抽樣法抽出的樣品，並設立下列之推算式：

$$u_m = \frac{n\bar{y}-m\bar{y}_m}{n-m} - h(G_m)\left[\frac{n\bar{z}-m\bar{z}_m}{n-m} - \frac{N\bar{Z}-m\bar{z}_m}{N-m}\right]$$

式中 $\bar{y}$ 及 $\bar{z}$ 係 n 個單位之樣品平均值，$\bar{y}_m$ 及 $\bar{z}_m$ 係居首 m 個單位之小樣品平均值，G_m 表示樣品按其被抽的順序排列後居首 m 對觀測值之集合，而 $h(G_m)$ 爲此 m 對觀測值之函數。當我們固定 G_m 後 u_m 之附件期望值 (The Conditional Expected Value of u_m, given G_m) 如次：

$$E(u_m|G_m) = \frac{N\bar{Y}-m\bar{y}_m}{N-m}$$

因此，我們從上面之附件期望值經矯正偏估量後很容易獲得 $\bar{Y}$ 之無偏推算式，以 t_m 表示之，即

$$t_m = \frac{(N-m)u_m + m\bar{y}_m}{N}$$

$$= \frac{(N-m)n}{N(n-m)}\left[\bar{y}-h(G_m)(\bar{z}-\bar{Z})\right] - \frac{(N-n)m}{N(n-m)}\left[\bar{y}_m - h(G_m)(\bar{z}_m-\bar{Z})\right] \qquad (1)$$

故令 $h(G_1) = y_1/z_1$，則從 (1) 式得

$$t_1 = \frac{y_1}{z_1}\bar{Z} + \frac{(N-1)n}{N(n-1)}\left(\bar{y} - \frac{y_1}{z_1}\bar{z}\right)$$

爲 $\bar{Y}$ 之無偏推算式。因爲樣品中每一對觀測值都有可能在第一次被抽的緣故，我們可求得 n 個 t_i's，將其平均之即得精度 (Precision) 較高的 $\bar{Y}$ 之無偏推算式：

$$\bar{t}_1 = \bar{r}\,\bar{Z} + \frac{(N-1)n}{N(n-1)}(\bar{y} - \bar{r}\,\bar{z}) = \hat{\bar{Y}}_r \tag{2}$$

中 $$\bar{r} = \frac{1}{n}\sum_{i=1}^{n}\frac{y_i}{z_i} = \frac{1}{n}\sum_{i=1}^{n} r_i$$

上式 $\hat{\bar{Y}}_r$ 就是 Hartley 及 Ross 兩氏之無偏比率推算式。假如我們令 $h(G_m) = \bar{y}_m/\bar{z}_m$，$\hat{R}_m$，則可得另一種無偏推算式如下：

$$t_m = \hat{R}_m\bar{Z} + \frac{(N-m)n}{N(n-m)}(\bar{y} - \hat{R}_m\bar{z}) \tag{3}$$

若令 $h(G_m) = \sum_{i=1}^{n}(y_i - \bar{y}_m)(z_i - \bar{z}_m)/\sum_{i=1}^{n}(z_i - \bar{z}_m)^2 = b_m$，則可得 $\bar{Y}$ 之無偏廻歸推算式如次：

$$t_m = \bar{y} + b_m(\bar{Z} - \bar{z}) - \frac{(N-m)n}{N(n-m)}[\bar{y}_m - \bar{y} - b_m(\bar{z}_m - \bar{z})] \tag{4}$$

現在爲了求以上諸推算值之變方 (Variance) 的無偏推算式起見，我們按樣品被抽之順序聚首 $m_1, m_2, \cdots\cdots, m_{k+1}=n$ 個單位分別組成小樣品 $(0<m_1<m_2<\cdots\cdots<m_{k+1}=n)$。然後利用每兩個集合 G_{m_i} 及 $G_{m_{i+1}}$ $(i=1, 2, \cdots\cdots, k)$ 利用公式 (3) 或 (4) 獲得 k 個 $\bar{Y}$ 之無偏比率或廻歸推算式，以 $t(m_1, m_2)$, $t(m_2, m_3)$, $\cdots\cdots$, $t(m_k, n)$ 表示之。將此 k 個推算值平均之即可得精度較高的無偏推算值，以式示之如次：

$$\bar{t} = \frac{1}{k}\sum_{i=1}^{k} t(m_i, m_{i+1}) \tag{5}$$

其變方之無偏推算式如下：

$$\mathrm{var}(\bar{t}) = \frac{1}{k(k-1)}\sum_{i=1}^{k}[t(m_i, m_{i+1}) - \bar{t}]^2$$

Williams 氏 (1961, 1962) 也曾將樣品分爲等大之 k 個集合而導出 $\bar{Y}$ 之無偏比率及廻歸推算式。查以上所論之諸種無偏推算式在實際應用時計算手續非常繁雜且 k 之大小直接影響推算之精度 (Precision)，即 k 愈大則推算之精度愈高，但計算手續愈複雜。再者，當輔助變值與主要變值間之關係爲非直線時，其推算雖然無偏但精度不高。鑒於此，著者擬導出精度較高而簡單易算之推算方法。

二、順序介值之抽樣分布 (The Sampling Distribution of Order Statistics)

設某一有限族群 (Finite Population) 含有 N 對密切相關之變值 (z_{oi}, y_i)，$i=1, 2, \cdots\cdots, N$，其 N 個輔助變值 z_{oi} 係按其數值大小由小而大順序排列；即 $z_{o1}<z_{o2}<\cdots\cdots<z_{oN}$。倘若有等大之輔助變值存於族群中時，我們可用逢機 (Random) 方法決定其順序。今設以不歸還簡單逢機抽樣法 (Simple Random Sampling Without Replacement) 從該族群中抽出 n 對變值組成一樣品。將抽出的 n 對變值按輔助變值大小由小而大順序排列後所得之成對變值以 $(z_{(i)}, y_{(i)})$，$i=1, 2, \cdots\cdots, n$ 表示之：$z_{(i)}$ 係輔助變值在樣品中的第 i 個順序介值 (Order Statistic) 而 $y_{(i)}$ 係附隨於 $z_{(i)}$ 之主要變值。

我們先討論第 i 個順序介值 $z_{(i)}$ 之抽樣分布。在樣品大小 n 一定及 $z_{(i)} = z_{ox_i}$ 下，我們需要從小於 z_{ox_i} 之 (x_i-1) 個 z_{oj} $(j=1, 2, \cdots\cdots, x_i-1)$ 中逢機抽出 $(i-1)$ 個；又需要從大於 z_{ox_i} 之 $(N-x_i)$ 個 z_{ok} $(k=x_i+1, x_i+2, \cdots, N)$ 中逢機抽出 $(n-i)$ 個。因此，按照組合分析 (Combinatorial Analysis) 可得 $z_{(i)}$ 之機率密度函數 (Probability Density Function) 如下：

$$P[z_{(i)} = z_{ox_i}] = \frac{\binom{x_i-1}{i-1}\binom{N-x_i}{n-i}}{\binom{N}{n}} = P(x_i|N, n, i) \tag{6}$$

式中 $x_i = i, i+1, \cdots\cdots, N-n+i$。觀察 $P(x_i|N, n, i)$ 之內容，我們一方面可視其爲順序介值 $z_{(i)}$ 之

機率密度函度，該介值之變域爲 $z_{o1}, z_{o(i+1)}, \cdots\cdots, z_{o(N-n+1)}$ 另一方面可視其爲表示順序介值 $Z_{(i)}$ 在族群中排列地位之順位介值 (Rank Statistic) x_i 的機率密度函數，該介值之變域爲 i, i+1,……, N−n+i。因此，利用順位介值於推算主要變值 y 之族群介量 (Parameter) 時比較利用順序介值者容易處理。$P(x_i|N,n,i)$ 係屬於超幾何分布 (Hypergeometric Distribution) 之機率密度函數。

其次，我們欲求順位介值 x_i 之平均及變方。在決定分立變值 (Discrete Variate) 之動差 (Moment) 以前，先求其階乘動差 (Factorial Moment) 較爲方便，參看 Wilks, S. S. (1962)。第 r 級階乘動差爲

$$E\{x^{[r]}\}=E\{x(x-1)\cdots\cdots(x-r+1)\}$$

現在我們先求 $E\{(x_i+r-1)^{[r]}\}$，然後從這階乘動差求出所需要之動差。因爲 $P(x_i|N,n,i)$ 是順位介值 x_i 之機率密度函數，在 x_i 之變域上的所有 $P(x_i|N,n,i)$ 總計必等於 1，所以我們可得下列之關係式：

$$\sum_{x_i=i}^{N-n+i}\binom{x_i-1}{i-1}\binom{N-x_i}{n-i}=\binom{N}{n}$$

再者，我們可證明下列等式成立：

$$\binom{x_i-1}{i-1}=\binom{x_i+r-1}{i+r-1}\frac{(i+r-1)^{[r]}}{(x_i+r-1)^{[r]}}$$

因此，

$$E\{(x_i+r-1)^{[r]}\}=\sum_{x_i=i}^{N-n+i}(x_i+r-1)^{[r]}P(x_i|N,n,i)$$

$$=(i+r-1)^{[r]}\frac{\binom{N+r}{n+r}}{\binom{N}{n}} \tag{7}$$

令 r=1 及 2，我們可得 x_i 之平均及變方如下：

$$E(x_i)=\frac{N+1}{n+1}\,i \tag{8}$$

$$Var(x_i)=E(x_i^2)-\{E(x_i)\}^2=\frac{i(n-i+1)(N+1)(N-n)}{(n-1)^2(n+2)} \tag{9}$$

繼之，我們欲求兩順位介值 x_i 和 x_j $(x_i<x_j)$ 之變積 (Covariance)。倣照以前之組合分析方法，可得順序介值 $z_{(i)}$ 和 $z_{(j)}$ (i<j) 或即順位介值 x_i 和 x_j 之聯合機率密度函數。在一定的樣品大小 n，及 $z_{(i)}=z_{ox_i}$ 和 $z_{(j)}=z_{ox_j}$ 之場合下，我們可得

$$P[z_{(i)}=z_{ox_i},\ z_{(j)}=z_{ox_j}]=\frac{\binom{x_i-1}{i-1}\binom{x_j-x_i-1}{j-i-1}\binom{N-x_j}{n-j}}{\binom{N}{n}}$$

$$=P(x_i,x_j|N,n,i,j) \tag{10}$$

式中 $x_i=i, i+1,\cdots\cdots N-n+i$, $x_j=x_i+j-1,\cdots\cdots, N-n+j$。爲了獲得 x_i 和 x_j 之變積，我們先求下列 r 級階乘動差：

$$E\{(x_j-x_i+r-1)^{[r]}\}=\sum_{x_i=i}^{N-n+i}\sum_{x_j=x_i+j-1}^{N+n-j}(x_j-x_i+r-1)^{[r]}P(x_i\,x_j|N,n,i,j)$$

$$=\frac{(j-i+r-1)^{[r]}}{\binom{N}{n}}\sum_{x_i}\sum_{x_j}\binom{x_i-1}{j-1}\binom{x_j-x_i+r-1}{j-i+r-1}\binom{N-x_j}{n-j}$$

$$=(j-i+r-1)^{[r]}\frac{\binom{N+r}{n+r}}{\binom{N}{n}} \tag{11}$$

令上式之 r=1 及 2 ，我們可求得

$$E(x_j-x_i)=(j-i)\frac{N+1}{n+1} \tag{12}$$

$$Cov(x_i, x_j)=E(x_i x_j)-E(x_i)E(x_j)=\frac{i(n-j+1)(N+1)(N-n)}{(n+1)^2(n+2)} \tag{13}$$

察看 (12) 式，兩個順位介值之差的期望值 (Expected Value) 僅受該兩個順位介值在樣品中所處位置的距離遠近之影響。詳言之，因爲抽樣結果係按輔助變值之大小順序排列之故，x_i 係排在樣品中的第 i 個位置而 x_j 排在第 j 個位置 (i<j)，該兩介值相差之期望值等於兩介值之距離 (j−i) 與 (N+1)/(n+1) 之相乘積。所以，只要選擇距離相等之任何兩個順位介值 x_i 和 x_j ，我們可獲得該兩介值相差之期望值相等之結果。例如：$E(x_9-x_4)=E(x_7-x_2)$。

我們現在要將原來的逢機變值 (Random Variate) x_i 轉換爲新的逢機變值 $x_i'=x_i/N$ 而研究該轉換後的逢機變值之極限分布。假若族群大小 N 足够大時，x_i' 可視爲以不歸還抽樣法從在間隔 [0, 1] 上的矩形分布 (Rectangular Distribution) 之族群抽出 n 個逢機變值組成的樣品中第 i 個順序介值，因此，x_i' 的機率密度函數參看 Wilks, S. S. (1962) 如下：

$$P(x_i'=x)=\frac{\Gamma(n+1)}{\Gamma(i)\Gamma(n-i+1)}x^{i-1}(1-x)^{n-i} \quad 0\leq x\leq 1 \tag{14}$$

式中 $\Gamma(i)$ 爲 Gamma 函數。再者，p 個逢機變值 x'_{k_1}，$x'_{k_1+k_2}$，……，$x'_{k_1+\cdots+k_p}$ 的聯合機率密度函數如次：

$$P(x'_{k_1}=t_1, x'_{k_1+k_2}=t_2, \cdots\cdots, x'_{k_1+\cdots+k_p}=t_p)$$
$$=\frac{\Gamma(n+1)}{\Gamma(k_1)\Gamma(k_2)\cdots\Gamma(k_p)\Gamma(n-k_1-\cdots\cdots-k_p+1)}$$
$$\cdot t_1^{k_1-1}(t_2-t_1)^{k_2-1}\cdots(t_p-t_{p-1})^{k_p-1}(1-t_p)^{n-k_1-\cdots\cdots-k_p} \tag{15}$$

式中 $0\leq t_1<t_2<\cdots\cdots<t_p\leq 1$。

在求 x_i' 之變方及 x_i' 和 x_j' 之變積以前，我們先求 $x_i'^p x_j'^q$ 及 $x_i'^p$ 之期望值，式中 i<j, p 及 q 爲正整數。從 (15) 式得 x_i' 及 x_j' 之聯合機率密度函數如下：

$$P(x_i'=s, x_j'=t)=Ks^{i-1}(t-s)^{j-i-1}(1-t)^{n-j}, \quad 0\leq s<t\leq 1$$

式中　　$K=\Gamma(n+1)/\Gamma(i)\Gamma(j-i)\Gamma(n-j+1)$

因此，$x_i'^p x_j'^q$ 之期望值如次：

$$E(x_i'^q x_j'^q)=K\int_0^1\int_0^t s^{p+i-1}t^q(t-s)^{j-i-1}(1-t)^{n-j}\,ds\,dt$$

令 $s/t=u$ 及 $t=v$，則變數轉換後之 Jacobian 等於 $|J|=v$，故

$$E(x_i'^p x_j'^q)=K\int_0^1\int_0^1 u^{p+i-1}(1-u)^{j-i-1}v^{p+q+j-1}(1-v)^{n-j}\,du\,dv$$
$$=K\frac{\Gamma(p+i)\Gamma(j-i)}{\Gamma(p+j)}\cdot\frac{\Gamma(p+q+j)\Gamma(n-j+1)}{\Gamma(n+p+q+1)}$$
$$=\frac{\Gamma(n+1)\Gamma(p+i)\Gamma(p+q+j)}{\Gamma(i)\Gamma(p+j)\Gamma(n+p+q+1)} \tag{16}$$

同理，我們可求出 $x_i'^p$ 之期望值如下：

$$E(x_i'^p)=\frac{\Gamma(n+1)\Gamma(p+i)}{\Gamma(i)\Gamma(n+p+1)} \tag{17}$$

利用 (16) 及 (17) 兩式，我們可求出下列諸式：

$$E(x_i')=\frac{i}{n+1} \tag{18}$$

$$Var(x_i')=\frac{i(n-i+1)}{(n+1)^2(n+2)} \tag{19}$$

$$\mathrm{Cov}(x_i', x_j') = \frac{i(n-j+1)}{(n+1)^2(n+2)} \qquad (20)$$

假如我們利用 x_i 之機率密度函數及 x_i 和 x_j 之聯合機率密度函數來求 x_i' 之動差 (Moments) 時，從 (8)、(9) 及 (13) 三式可得如次：

$$E(x_i') = \frac{1}{N}E(x_i) = \frac{i}{n+1}\cdot\frac{N+1}{N}$$

$$\mathrm{Var}(x_i') = \frac{1}{N^2}\mathrm{Var}(x_i) = \frac{i(n-i+1)}{(n+1)^2(n+2)}\cdot\frac{(N+1)(N-n)}{N^2}$$

$$\mathrm{Cov}(x_i', x_j') = \frac{1}{N^2}\mathrm{Cov}(x_i, x_j) = \frac{i(n-j+1)}{(n+1)^2(n+2)}\cdot\frac{(N+1)(N-n)}{N^2}$$

當 N 很大時，(N+1)/N 及有限族群改正項 (Finite Population Correction Term) (N−n)/N 可視爲 1，因此上列三式分別變成 (18)，(19) 及 (20) 三式。由此可見，從超幾何分布 (Hypergeometric Distribution) 導出之 x_i' 的極限分布與從矩形分布 (Rectangular Distribution) 導出的結果相一致。

三、族群平均值的無偏推算式之誘導

根據上節所論，我們已知 x_i 順位介值 (Rank Statistics) 較 $z_{(i)}$ 順序介值 (Order Statistics) 容易處理，故我們欲利用成對變值 $(x_i, y_{(i)})$ 替代 $(z_{(i)}\ y_{(i)})$ 或取代 $(z_{o\,i}, y_{(i)})$ 於推算問題上。

族群平均值 $\bar{Y}$ 或總數 Y 之無偏推算式之獲得可先固定一組觀測值後應用分層推算原理求出未被固定的剩餘觀測值之族群平均值或總數之無偏推算式。爲使推算式之誘導簡便計，我們暫時假設樣品大小 n 等於 2m+1, m≥2，待後可知此種假設並不失去所誘導出來的推算式之一般性。我們首先固定處在偶數位之成對變值，即

$$\{(x_{2i}, y_{(2i)}) : i=1,2,\cdots\cdots, m\}$$

然而，當 $(x_2, y_{(2)})$ 固定時，$(x_1, y_{(1)})$ 可視爲從 (x_1-1) 對變值中逢機選出的樣品。當 $(x_2, y_{(2)})$ 及 $(x_4, y_{(4)})$ 固定時，$(x_3, y_{(3)})$ 可視爲從 (x_4-x_2-1) 對變值中逢機選出的樣品。一般言之，當 $(x_{2i-2}, y_{(2i-2)})$ 及 $(x_{2i}, y_{(2i)})$ 固定時，$(x_{2i-1}\ y_{(2i-1)})$ 可視爲從 $(x_{2i}-x_{2i-2}-1)$ 對變值中逢機選出的樣品。因此，未被固定的剩餘觀測值之族群總數的附件無偏推算式 (Conditionally Unbiased Estimator) 按分層推算原理可得如次：

$$Y_1^* = (x_2-1)y_{(1)} + (x_4-x_2-1)\ y_{(3)} + \cdots\cdots + (x_{2i}-x_{2i-2}-1)\ y_{(2i-1)} + \cdots\cdots + (N-x_{n-1})y_{(n)}$$

所以，整個族群總數之附件無偏推算式可將上式與被固定的一組觀測值相加而得之，即

$$\hat{Y}_1 = Y_1^* + \sum_{i=1}^{m} y_{(2i)}$$

同樣方法，我們固定處在奇數位之成對變值，即

$$\{(x_{2i+1}\ y_{(2i+1)})\ i=1,2,\cdots\cdots, m-1\}$$

然後依照上述理論可得未被固定的剩餘觀測值之族群總數的附件無偏推算式如下：

$$Y_2^* = \frac{1}{2}(x_3-1)(y_{(1)}+y_{(2)}) + (x_5-x_3-1)y_{(4)} + \cdots\cdots + (x_{2i+1}-x_{2i-1}-1)y_{(2i)} + \cdots\cdots$$

$$+ \frac{1}{2}(N-x_{n-2})(y_{(n-1)}+y_{(n)})$$

因此，整個族群總數之附件無偏推算式爲

$$\hat{Y}_2 = Y_2^* + \sum_{i=1}^{m-1} y_{(2i+1)}$$

茲將以上所得之兩個附件無偏推算式平均之，則得族群總數之無偏推算式如下：

$$\hat{Y}_u = \frac{1}{2}(\hat{Y}_1 + \hat{Y}_2)$$

而族群平均值之無偏推算式爲

$$ {}_2\hat{\bar{Y}}_u = \frac{1}{N}\,{}_2\hat{Y}_u = \frac{1}{2N}\sum_{i=1}^{n} w_i y_{(i)} \qquad (21) $$

式中 ${}_2\hat{\bar{Y}}_u$ 的左下角註號 2 係表示 ${}_2\hat{\bar{Y}}_u$ 由兩個附件無偏推算值平均而得，右下角註號 u 係 Unbiased 之縮寫，各 w_i 如下：

$$ w_1 = (x_2 - 1) + \frac{1}{2}(x_3 - 1) = \frac{1}{2}(2x_2 + x_3 - 3) $$

$$ w_2 = 1 + \frac{1}{2}(x_3 - 1) = \frac{1}{2}(x_3 + 1) $$

$$ w_3 = 1 + (x_4 - x_2 - 1) = x_4 - x_2 $$

$$ \vdots $$

$$ w_i = 1 + (x_{i+1} - x_{i-1} - 1) = x_{i+1} - x_{i-1} $$

$$ \vdots $$

$$ w_{n-2} = 1 + (x_{n-1} - x_{n-2} - 1) = x_{n-1} - x_{n-3} $$

$$ w_{n-1} = 1 + \frac{1}{2}(N - x_{n-2}) = \frac{1}{2}(N + 2 - x_{n-2}) $$

$$ w_n = (N - x_{n-1}) + \frac{1}{2}(N - x_{n-2}) = \frac{1}{2}(3N - x_{n-2} - 2x_{n-1}) $$

上列所有的 w_i's 總計等於 2N。這些 w_i's 可適用於 n 大於 3 的任何奇數或偶數之樣品。

查無偏推算式 ${}_2\hat{\bar{Y}}_u$ 係由二個附件無偏推算式 $\hat{\bar{Y}}_1^*$ 及 $\hat{\bar{Y}}_2^*$ 之平均，而 $\hat{\bar{Y}}_1^*$ 及 $\hat{\bar{Y}}_2^*$ 係分別固定一組成對變值後利用分層推算原理求得之兩推算式。我們利用被固定之一組成對變值的順位介值 x_i 做爲層界，將族群分成爲 m（或 m+1）層，各層之樣品數除了首尾兩層外都是相等，僅僅有一個。因此，我們可視爲抽樣後分層法 (Post-stratification)。但是在一般所謂之抽樣後分層法中各層的族群大小 (Population Size of Each Stratum) 係已知不變的，而各層樣品大小係逢機變數，因樣品而異。相反地，本研究所用之分層法中各層的的族群大小係逢機變數，而各層樣品大小固定不變。又本研究之分層目的在藉分層推算之原理以導出族群平均值 $\bar{Y}$ 之推算式並未完全達到分層之原來用意，故稱本研究之分層法爲抽樣後假分層法 (Pseudo-post-stratification)。

在 (21) 式中之各 w_i 因樣品之不同而異，蓋其爲逢機變值 x_i 之函數故也。w_i 的期望值及變方可利用 (8)，(9)，(13) 式很容易求得，其結果如次：

$$ E(w_1) = 2\left(\frac{N+1}{n+1}\right) + \frac{3}{2}\left(\frac{N-n}{n+1}\right) = E(w_n) $$

$$ E(w_2) = 2\left(\frac{N+1}{n+1}\right) - \frac{1}{2}\left(\frac{N-n}{n+1}\right) = E(w_{n-1}) $$

$$ E(w_i) = 2\left(\frac{N+1}{n+1}\right) \quad \text{for} \quad i = 3, \cdots\cdots, n-2 $$

$$ \mathrm{Var}(w_1) = \left[2(n-1) + \frac{11}{4}(n-2)\right]\frac{(N+1)(N-n)}{(n+1)^2(n+2)} = \mathrm{Var}(w_n) $$

$$ \mathrm{Var}(w_2) = \left[2(n-1) - \frac{1}{4}(5n-2)\right]\frac{(N+1)(N-n)}{(n+1)^2(n+2)} = \mathrm{Var}(w_{n-1}) $$

$$ \mathrm{Var}(w_i) = \frac{2(n-1)(N+1)(N-n)}{(n+1)^2(n+2)} \quad \text{for} \quad i = 3, \cdots\cdots, n-2 $$

從上面之結果可以看出在中間的 $w_i (i = 3, \cdots\cdots, n-2)$ 具有相等的期望值及變方，最首的 w_1 與最尾的 w_n 及首第二的 w_2 與倒數第二的 w_{n-1} 各具有相等的期望值及變方，成對稱現象。w_1 與 w_n 的期望值及變方爲最大，表示 $y_{(1)}$ 與 $y_{(n)}$ 的加權數 (Weight) w_1 及 w_n 就平均而言爲最大而其變異性亦最大。

上面所導出的無偏推算式係僅固定處在偶數及奇數位之成對變值求出兩個附件無偏推算式後將其平

均而得，每一假分層之樣品大小除首尾兩層外皆爲 1。我們若將假分層數減少而增加每分層之樣品大小時，可能提高推算之精度 (Precision)。我們知道樣品平均值之變方大小與樣品大小成反比，增加樣品大小可縮小樣品平均值之變方，而分層數之多寡影響層內之變異性，就一般而言層大則其層內變異性亦大。可見此兩種措施之效果成對立現象，蓋因本研究之假分層法的總樣品大小 n 爲固定的緣故，增加假分層數時每層內樣品大小必隨之減小。因此，我們需要考慮如何配合假分層數和每層樣品大小以增高推算之精度。這個問題相當於決定需要誘導之附件無偏推算式之最適數目，蓋因各層樣品大小爲 1 時我們可誘導二個附件無偏推算式，而各層樣品大小爲 k 時我們可誘導 k+1 個附件無偏推算式故也。

我們現在要按某一定之間隔固定成對變值誘導更多的附件無偏推算式。爲誘導上簡便計，我們暫假設樣品大小 n=km+k−1。首先，我們固定下列之一組成對變值：

$$\{(x_{i1+1},\ y_{(i1+1)}) : i=1,\ \cdots\cdots,\ m\}$$

利用分層推算原理，我們可得未被固定的剩餘觀測值之族群總數的附件無偏推算式如次：

$$Y_1^*=\frac{1}{k}(x_{k+1}-1)\sum_{j=1}^{k}y_{(j)}+\frac{1}{k-1}\sum_{i=2}^{m}(x_{i1+1}-x_{i1-k+1}-1)\sum_{j=1}^{k-1}y_{(i1-k+1+j)}+\frac{1}{k-2}(N-x_{km+1})\sum_{j=1}^{k-2}y_{(km+1+j)}$$

將上式與被固定的一組觀測值相加則得整個族群總數之附件無偏推算式如下：

$$\hat{Y}_1=Y_1^*+\sum_{i=1}^{m}y_{(i1+1)}$$

同樣的方法，我們固定下列之一組成對變值：

$$\{(x_{i1},\ y_{(i1)}) : i=1,\ \cdots\cdots,\ m\}$$

並利用分層推算原理則可得族群總數之另一附件無偏推算式如次：

$$\hat{Y}_2=Y_2^*+\sum_{i=1}^{m}y_{(i1)}$$

式中

$$Y_2^*=\frac{1}{k-1}(x_k-1)\sum_{j=1}^{k-1}y_{(j)}+\frac{1}{k-1}\sum_{i=2}^{m}(x_{i1}-x_{i1-k}-1)\sum_{j=1}^{k-1}y_{(i1-k+j)}+\frac{1}{k-1}(N-x_{km})\sum_{j=1}^{k-1}y_{(km+j)}$$

再次，我們固定下列之一組成對變值：

$$\{(x_{i1-1},\ y_{(i1-1)}) : i=1,\ \cdots\cdots,\ m\}$$

並利用分層推算原理則得族群總數之另一附件無偏推算式如下：

$$\hat{Y}_3=Y_3^*+\sum_{i=1}^{m}y_{(i1-1)}$$

式中

$$Y_3^*=\frac{1}{k-2}(x_{k-1}-1)\sum_{j=1}^{k-2}y_{(j)}+\frac{1}{k-1}\sum_{i=2}^{m}(x_{i1-1}-x_{i1-k-1}-1)\sum_{j=1}^{k-1}y_{(i1-k-1+j)}+\frac{1}{k}(N-x_{km-1})\sum_{j=1}^{k}y_{(km-1+j)}$$

我們再固定下列之另一組成對變值：

$$\{(x_{i1-2},\ y_{(i1-2)}) : i=1,\ \cdots\cdots,\ m+1\}$$

則可得

$$\hat{Y}_4=Y_4^*+\sum_{i=1}^{m+1}y_{(i1-2)}$$

式中

$$Y_4^*=\frac{1}{k-3}(x_{k-2}-1)\sum_{j=1}^{k-3}y_{(j)}+\frac{1}{k-1}\sum_{i=2}^{m+1}(x_{i1-2}-x_{i1-k-2}-1)\sum_{j=1}^{k-1}y_{(i1-k-2+j)}+(N-x_{n-1})y_{(n)}$$

繼續重覆上列步驟一直到各成對變值除第一及最後之兩成對變值外在誘導中都被固定一次爲止。最後一共可得 k 個附件無偏推算式，將其平均之則得族群總數之無偏推算式如下：

$$ {}_k\hat{\bar{Y}}_u = \frac{1}{k}\sum_{i=1}^{k}\hat{Y}_i $$

因此，族群平均值 $\bar{Y}$ 之無偏推算式爲

$$ {}_k\hat{\bar{Y}}_u = \frac{1}{N}\,{}_k\hat{Y}_u = \frac{1}{kN}\sum_{i=1}^{n} w_i^{(k)} y_{(i)} \qquad (22) $$

式中 k 表示附件無偏推算式之數目，各 $w_i^{(k)}$ 如次：

$$ w_1^{(k)} = \sum_{j=1}^{k}\frac{1}{j}(x_{j+1}-1) $$

$$ w_h^{(k)} = \sum_{j=h}^{k}\frac{1}{j}(x_{j+1}-1)+1+\frac{1}{k-1}\sum_{j=2}^{h-1}(x_{k+j}-x_j-1) \quad \text{for} \quad h=2,\cdots\cdots,k $$

$$ w_{ki+h}^{(k)} = \frac{1}{k-1}\sum_{j=1}^{k-1}(x_{ki+h+j}-x_{ki+h-j}) \quad \text{for} \quad i=1,\cdots\cdots,m-1;\ h=1,\cdots\cdots,k-1 $$

$$ w_{n+1-h}^{(k)} = \sum_{j=h}^{k}\frac{1}{j}(N-x_{n-j})+1+\frac{1}{k-1}\sum_{j=2}^{h-1}(x_{n+1-j}-x_{km-j}-1) \quad \text{for} \quad h=2,\cdots\cdots,k $$

$$ w_n^{(k)} = \sum_{j=1}^{k}\frac{1}{j}(N-x_{n-j}) $$

上列之 $w_i^{(k)}$'s 可適用於任何樣品大小 $n>k+1$，又其總計等於 kN；亦卽 n 個 $y_{(i)}$'s 之權數 (Weights) 的總計被 kN 除時等於 1，此爲達到無偏性的條件之一。

現在舉一數例來證明 ${}_2\hat{\bar{Y}}_u$ 之無偏性，並與直線廻歸推算值 $\bar{y}_{lr}$ 及 Hartley and Ross 兩氏之無偏比率推算值 $\hat{\bar{Y}}_r$ 比較之。設族群含有 8 個抽樣單位，每次以簡單逢機抽樣法 (Simple Random Sampling) 抽出 6 個組成樣品，則共有 28 種可能樣品。令族群之八個成對變值 (x_i, y_i) 爲

(1, 1) (2, 3) (3, 4) (4, 6) (5, 7) (6, 4) (7, 3) (8, 2)

x 及 y 之族群平均值分別爲 $\bar{X}=4.5$ 及 $\bar{Y}=3.75$。從這二十八種可能樣品計算所得的各種推算值列於下表以便比較。

表一：所有可能樣品的各種推算值

Sample number	${}_2\hat{\bar{Y}}_u$	$\bar{y}_{lr}$	$\hat{\bar{Y}}_r$	Sample number	${}_2\hat{\bar{Y}}_u$	$\bar{y}_{lr}$	$\hat{\bar{Y}}_r$
1	4.3125	4.9953	5.3925	16	3.7500	4.2286	4.3223
2	4.0000	4.3928	4.9759	17	3.5312	4.0000	3.9694
3	3.6875	3.9595	4.5778	18	3.6875	3.8250	3.6309
4	3.6250	3.6863	4.1839	19	3.3438	3.3206	3.0037
5	3.3750	3.3774	3.8125	20	3.4375	3.4412	3.0402
6	3.4375	3.1738	3.4502	21	3.4375	3.7865	3.2163
7	3.8438	3.8631	4.1664	22	4.1875	4.5000	4.4689
8	3.6250	3.5622	3.8031	23	3.9688	4.3643	4.1067
9	3.7812	3.3517	3.4507	24	4.1250	4.2516	3.7605
10	3.1250	2.8333	2.7808	25	3.7812	3.7917	3.1402
11	3.9062	4.1366	4.3353	26	3.9375	3.9534	3.1652
12	3.6878	3.8650	3.9724	27	4.0625	4.4286	3.3217
13	3.8438	3.6667	3.6221	28	4.3750	4.9619	3.3645
14	3.5000	3.1526	2.9686	Mean	3.75	3.8625	3.75
15	3.6250	3.2790	2.9973	MSE	0.0920	0.2966	0.4217

從上表可知所有可能樣品計算所得之 ${}_2\hat{\bar{Y}}_u$ 的平均值；亦即 ${}_2\hat{\bar{Y}}_u$ 之期望值等於族群平均值 $\bar{Y}=3.75$。因此，${}_2\hat{\bar{Y}}_u$ 的無偏性由該數例證實。直線迴歸推算值 $\bar{y}_{lr}$ 之偏估量爲 0.1125，而 Hartley and Ross 兩氏之比率推算值 $\hat{\bar{Y}}_r$ 亦是無偏的，但是其變方爲 0.4217 較 ${}_2\hat{\bar{Y}}_u$ 之變方 0.0920 大 4.6 倍。表中 MSE 係 Mean Squared Error 之縮寫，譯爲均方偏差，當推算值爲無偏時，均方偏差等於變方。本例中 y 依 x 而變之迴歸是非直線的，因此 Hartley and Ross 兩氏之無偏比率推算值的變方增大，同時直線迴歸推算值之偏估量加大。假如不利用輔助變值來推算時，主要變值之樣品平均值 $\bar{y}$ 的變方等於 0.1637。可見在此例中本研究所誘導的無偏推算式具有較高的精度。因爲本例中 y 依 x 而變之迴歸爲非直線的緣故，Hartley and Ross 兩氏之無偏比率推算值變方反而比不利用輔助變值之樣品平均值 $\bar{y}$ 的變方大 2.5 倍多，誠然徒勞無功。

四、直線數學模式下推算值之變方

在沒有假設任何模式 (Model) 之場合下，我們很難求出一簡單的公式來表示本研究所創立的推算值 ${}_k\hat{\bar{Y}}_u$ 之變方。我們將用下列之直線數學模式來求推算值 ${}_k\hat{\bar{Y}}_u$ 之變方：

$$y_{(i)}=A+Bx_i+e_i \tag{23}$$

式中 x_i 與 e_i 互相獨立，B 爲 y 依 x 而變之迴歸係數 (Regression Coefficient of y on x)，

$$A=\bar{Y}-\frac{1}{2}(N+1)B$$

$$E(e_i|x_i)=0 \quad \text{for all } i$$

$$E(e_i^2|x_i)=S_e^2 \quad \text{for all } i$$

$$E(e_ie_j|x_i, x_j)=0 \quad \text{for } i\neq j$$

推算值 ${}_k\hat{\bar{Y}}_u$ 之變方可分兩部份求之，即 ${}_k\hat{\bar{Y}}_u$ 之變方等於固定一組的 x_i's 之附件變方期望值加固定一組的 x_i's 之附件期望值的變方，參看 Hansen, M. H., et al (1953)，以式示之如下：

$$\mathrm{Var}({}_k\hat{\bar{Y}}_u)=E\mathrm{Var}[{}_k\hat{\bar{Y}}_u|x_i\text{'s}]+\mathrm{Var}[E({}_k\hat{\bar{Y}}_u|x_i\text{'s})] \tag{24}$$

現在分別求出上式等號右邊之兩成份。爲使符號簡化起見，我們將去除 $w_i^{(k)}$ 的右上角註號 (k)。第一成份爲

$$E\mathrm{Var}[{}_k\hat{\bar{Y}}_u|x_i\text{'s}]=\frac{S_e^2}{k^2N^2}\sum_{i=1}^{n}E(w_i^2)$$

在 (22) 式中之權數 w_i 於 $i=k+1, \cdots\cdots, n-k$ 時可寫成如次式：

$$w_i=\frac{1}{k-1}\sum_{j=1}^{k-1}(x_{i+j}-x_{i-j})$$

利用 (8)、(9) 及 (13) 三式，我們可求 w_i^2 之期望值如下：

$$\begin{aligned}E(w_i^2)&=\frac{1}{(k-1)^2}E\Big[\sum_{j=1}^{k-1}(x_{i+j}^2+x_{i-j}^2-2x_{i+j}x_{i-j})\\&\qquad+2\sum_{j<h}(x_{i+j}x_{i+h}-x_{i+j}x_{i-h}-x_{i-j}x_{i+h}+x_{i-j}x_{i-h})\Big]\\&=\frac{1}{(k-1)^2}\Big[\sum_{j=1}^{k-1}\Big\{2j(n-2j+1)C+4j^2\Big(\frac{N+1}{n+1}\Big)^2\Big\}\\&\qquad+2\sum_{j<h}\Big\{2j(n-2h+1)C+4hj\Big(\frac{N+1}{n+1}\Big)^2\Big\}\Big]\\&=\frac{k}{3(k-1)}\Big[(3n+5-4k)C+2(2k-1)\Big(\frac{N+1}{n+1}\Big)^2\\&\qquad+(k-2)(2n-3k+3)C+(k-2)(3k-1)\Big(\frac{N+1}{n+1}\Big)^2\Big]\\&=\frac{k}{3(k-1)}\Big[\{n(2k-1)-3k^2+5k-1\}C+3k(k-1)\Big(\frac{N+1}{n+1}\Big)^2\Big]\end{aligned}$$

式中　$C=(N+1)(N-n)/(n+1)^2(n+2)$

因爲最前及最後的 k 個 w_i^2's 之期望值不容易求出，故擬利用中間部份的 $w_i^2\,(i=k+1,\cdots\cdots,n-k)$ 之期望值做爲它們的近似值。因此，n 個 w_i^2's 之期望值總計如下：

$$\sum_{i=1}^{n} E(w_i^2)\fallingdotseq \frac{k}{3(k-1)}\left[\{(2k-1)n-3k^2+5k-1\}C+3k(k-1)\left(\frac{N+1}{n+1}\right)^2\right]$$

故固定一組的 x_i's 後 ${}_k\hat{Y}_u$ 之附件變方之期望值可得如次：

$$E\,Var[{}_k\hat{Y}_u|x_i's]=\frac{S_e^2}{n}\left[1+\frac{2k-1}{3k(k-1)}\right] \tag{25}$$

上式僅保留 n^{-1} 級之各項而已。

(24) 式等號右邊之第二成份可求如下：

$$Var[E({}_k\hat{Y}_u|x_i's)]=\frac{1}{k^2N^2}Var\left[\sum_{i=1}^{n} w_i(A+Bx_i)\right]$$

$$=\left[\frac{B}{kN}\right]^2 Var\left(\sum_{i=1}^{n} w_i x_i\right)$$

因爲上式中 x_i 的加權總計之變方很不容易求出，所以我們擬利用下列之近似法求之。設 x 及 y 分別爲 μ_x 及 μ_y 之無偏推算值，則其數學模式可分別訂立如次：$x=\mu_x+e_x$，$y=\mu_y+e_y$，式中 $E(e_x)=0$，$E(e_x^2)=Var(x)$，$E(e_y)=0$，$E(e_y^2)=Var(y)$，$E(e_xe_y)=Cov(x,y)$。然而，xy 之變方近似值如下：

$$\begin{aligned}Var(xy)&\fallingdotseq E(xy-\mu_y\mu_y)^2=E(\mu_xe_y+\mu_ye_x+e_xe_y)^2\\&\fallingdotseq\mu_x^2\,Var\,(y)+\mu_y^2\,Var\,(x)+2\mu_x\mu_y\,Cov\,(x,y)\end{aligned} \tag{26}$$

上式中忽略 $2\mu_xCov(x,y^2)+2\mu_yCov(x^2,y)+Cov(x^2,y^2)$，蓋其屬於第三及第四級項，於樣品大小 n 相當大時可忽視故也。同理，設 x, y, u 及 v 分別爲 μ_x, μ_y, μ_u 及 μ_v 之無偏推算值，則其數學模式分別可訂立如次：$x=\mu_x+e_x$，$y=\mu_y+e_y$，$u=\mu_u+e_u$ 及 $v=\mu_v+e_v$，式中 $E(e_t)=0$，$E(e_t^2)=Var(t)$；$t=x,y,u,v$ 及 $E(e_te_r)=Cov(t,r)$；$t\neq r$。然而，xy 與 uv 之變積近似值如次：

$$\begin{aligned}Cov(xy,uv)&\fallingdotseq E(xy-\mu_x\mu_y)(uv-\mu_u\mu_v)\\&\fallingdotseq\mu_x\mu_uCov(y,v)+\mu_x\mu_vCov(y,u)\\&\qquad+\mu_y\mu_uCov(x,v)+\mu_y\mu_vCov(x,u)\end{aligned} \tag{27}$$

式中忽略 $\mu_xE(e_ye_ue_v)$，$E(e_xe_ye_ue_v)$ 等等屬於第三及第四級之各項。

利用 (26) 及 (27) 兩式，經過冗長繁雜之演算後簡化而得 $\sum_{i=1}^{n} w_ix_i$ 之變方近似值如下：

$$\begin{aligned}Var\left(\sum_{i=1}^{n} w_ix_i\right)&\fallingdotseq\sum_{i=1}^{n} Var(w_ix_i)+2\sum_{i<j} Cov(w_ix_i,w_jx_j)\\&\fallingdotseq\frac{n^3(N+1)^3(N-n)}{24(k-1)(n+1)^4(n+2)}(k-2)(k-3)(k^2-3k-3)\end{aligned}$$

再者，$B=\rho S_y/S_x$，式中 ρ 係 x 與 y 之相關係數，又

$$\begin{aligned}S_x^2&=\frac{1}{N-1}\left[\sum_{i=1}^{N} x_i^2-\left(\sum_{i=1}^{N} x_i\right)^2/N\right]\\&=\frac{1}{N-1}\left[\frac{N(N+1)(2N+1)}{6}-\frac{N(N+1)^2}{4}\right]\\&=\frac{N(N+1)}{12}\end{aligned}$$

$$\therefore\quad B=\sqrt{12}\,\rho S_y/\sqrt{N(N+1)}$$

利用上列所得之結果，我們可獲得固定一組的 x_i's 後 ${}_k\hat{Y}_u$ 之附件期望值的變方近似值如下：

$$Var[E({}_k\hat{Y}_u|x_i's)]\fallingdotseq\frac{(k-2)(k-3)(k^2-3k-3)}{2k^2(k-1)n^2}\rho^2S_y^2 \tag{28}$$

將 (25) 及 (28) 兩式代入 (24) 式，即得 ${}_k\hat{Y}_u$ 之變方近似值如次：

$$\mathrm{Var}({}_k\hat{\bar{Y}}_u) \doteqdot \frac{S_e^2}{n}\left[1+\frac{2k-1}{3k(k-1)}\right]+\frac{(k-2)(k-3)(k^2-3k-3)}{2k^2(k-1)n^2}\rho^2 S_y^2 \qquad (29)$$

我們知道當 (23) 式所假設之直線數學模式的前提可成立時，直線廻歸推算值 $\bar{y}_{lr}$ 爲最適直線無偏推算值 (Teh Best Linear Unbiased Estimate)，其變方可參看 Cochran, W.G. (1963) 如次：

$$\mathrm{Var}(\bar{y}_{lr}) = \frac{S_e^2}{n}\left(1+\frac{1}{n}\right)$$

因此，顯然地可知在直線數學模式之前提下，本研究所誘導的 ${}_k\hat{\bar{Y}}_u$ 不如 $\bar{y}_{lr}$。但是，推算式 ${}_k\hat{\bar{Y}}_u$ 之重要價值在於直線廻歸推算式 $\bar{y}_{lr}$ 具有偏估之場合下仍爲無偏的；換言之，${}_k\hat{\bar{Y}}_u$ 在任何場合下都是無偏的，而且 ${}_k\hat{\bar{Y}}_u$ 之變方尙可比 Hartley and Ross 兩氏的無偏比率推算值 $\hat{\bar{Y}}_r$ 之變方爲小，可從前節之數例中察知之。

我們現在欲決定推算式中所含附件無偏推算式之最適當的數目。爲簡便計，我們將 (29) 式化成下列之簡單式子：

$$\mathrm{Var}({}_k\hat{\bar{Y}}_u) = \frac{S_e^2}{n}\left(1+\frac{2}{3k}\right)+\frac{k}{2n^2}\rho S_y^2$$

設 $|\rho|$ 卡 1 及 n 一定，則我們就上式對 k 偏微分並令其結果等於零卽得

$$k_{opt} = \left[\frac{4n(1-\rho^2)}{3\rho^2}\right]^{\frac{1}{3}} \qquad (30)$$

從上式可見，最適當的附件無偏推算式之數目因族群資料之成對變值間的相關程度及樣品大小而異。玆利用前節所舉之數例來求其最適當的附件無偏推算式之數目。由資料計算得 y 與 x 之相關係數 $\rho=0.7$ 而 $n=6$，故

$$k_{opt} = \left[\frac{4\times 6(1-0.49)}{3\times 0.49}\right]^{\frac{1}{3}} \doteqdot 3$$

現在利用 (22) 式令 $k=3$ 計算所有可能樣品之 ${}_3\hat{\bar{Y}}_u$ 並將其結果列於表二中。

表二、 所有可能樣品的推算值 ${}_3\hat{\bar{Y}}_u$

Sample number	${}_3\hat{\bar{Y}}_u$	Sample number	${}_3\hat{\bar{Y}}_u$	Sample number	${}_3\hat{\bar{Y}}_u$	Sample number	${}_3\hat{\bar{Y}}_u$
	4.3889	8	3.5347	15	3.4306	22	4.2917
2	4.1111	9	3.5069	16	3.8889	23	4.0903
3	3.8333	10	2.9306	17	3.6875	24	4.1181
4	3.6111	11	3.9583	18	3.7153	25	3.7153
5	3.3750	12	3.7569	19	3.2847	26	3.9097
6	3.3056	13	3.7847	20	3.4375	27	4.1944
7	3.7500	14	3.2778	21	3.6389	28	4.4722

上表中廿八個 ${}_3\hat{\bar{Y}}_u$ 值之平均值等於 3.75，可見 ${}_3\hat{\bar{Y}}_u$ 爲族群平均值 $\bar{Y}$ 之無偏推算值，其變方等於 0.1367，比 ${}_2\hat{\bar{Y}}_u$ 之變方 (0.0920) 大。查本資料之 x 與 y 間的關係並非直線而是一種二次曲線 (Quadratic Form) 關係，但 (30) 式之最適當的附件無偏推算式之數目 k_{opt} 係根據 x 與 y 間爲直線關係誘導而得，因此與實際資料之最適數目稍有差異。本資料之 k_{opt} 應爲 2 比利用 (30) 式求得者小 1，僅僅相差 1，所以 (30) 式可供決定最適附件無偏推算式數目之參考。

五、${}_k\hat{\bar{Y}}_u$ 與抽樣後分層推算比較效率

我們在第二節中討論逢機變值 $x'_i = x_i/N$ 之極限分布結果獲悉其在間隔 [0, 1] 上呈現矩形分布 (Rectangular Distribution)，所以我們擬在下列之直線數學模式下做推算效率之比較：

$$y_{(i)}=A+Bx_i'+e_i \tag{31}$$

式中 x_i' 與 e_i 互相獨立，且

$$E(e_i|x_i')=0 \text{ for all } i$$

$$E(e_i^2|x_i')=S_e^2 \text{ for all } i$$

$$E(e_iej|x_i', x_j)=0 \text{ for } i\neq j$$

$$A=\bar{Y}-\frac{1}{2}B,\ B=\sqrt{12}\rho S_y$$

假定樣品大小 n 充分地大足使各層都有樣品抽出，則抽樣後分層推算值 (Post-stratified Estimate) $\bar{y}_w$ 之變方參看 Cochran, W.G. (1963) 如下：

$$Var(\bar{y}_w)=\frac{1}{n}\sum_{h=1}^{L}W_hS_{yh}^2+\frac{1}{n^2}\sum_{h=1}^{L}(1-W_h)\ S_{yh}^2$$

若 $W_h=W$ 及 $S_{yh}^2=S_y^2$ $(h=1,\cdots,L)$，則 $\bar{y}_w$ 之變方可簡化如次：

$$Var(\bar{y}_w)=\frac{L}{n}S_y^2[W+\frac{1}{n}(1-W)]$$

式中之 W 及 S_y^2 在 (31) 式之前提下分別爲

$$W=1/L$$

$$S_y^2=S_e^2+\frac{1}{L^2}\rho^2S_y^2$$

茲假定分層數 L 等於 n/4，則 $\bar{y}_w$ 之變方爲

$$Var(\bar{y}_w)=\frac{5}{4n}S_e^2$$

式中忽略高於或等於 n^{-2} 級各項。

查本研究之無偏推算值 ${}_s\hat{Y}_u$ 所含各附件無偏推算式係將族群分爲 n/4 層後利用分層推算原理而得，所以擬將其與上述抽樣後分層推算值 $\bar{y}_w$ 比較效率。將 (29) 式中高於或等於 n^{-2} 級之各項忽視後得 ${}_s\hat{Y}_u$ 之變方如下：

$$Var({}_s\hat{Y}_u)=\frac{S_e^2}{n}(1+\frac{3}{20})$$

因此，${}_s\hat{Y}_u$ 對 $\bar{y}_w$ 之相對精度 (Relative Precision) 如次：

$$R.E.=\frac{Var(\bar{y}_w)}{Var({}_s\hat{Y}_u)}=\frac{25}{23}$$

可見本研究所誘導的無偏推算值 ${}_s\hat{Y}_u$ 之精度 (Precision) 稍比抽樣後分層推算值 $\bar{y}_w$ 者高些。假如我們採用具有最適當 k 的 ${}_k\hat{Y}_u$ 與 $\bar{y}_w$ 比較，則相對精度可能增高。

六、推算值之變方的估算

因爲推算值 ${}_k\hat{Y}_u$ 所含的 k 個附件無偏推算值係每一次利用固定不同的一組成對變值中的輔助變值將族群劃分爲 m (或 m+1) 層，然後應用分層推算原理而導出的緣故，我們仍可應用分層推算原理以估算 ${}_k\hat{Y}_u$ 之變方。茲設樣品大小 $n=k(m+1)$。在獲得 k 個附件無偏推算式中之一個時，我們曾經固定下列的一組成對變值：

$$\{(x_{ki}, y_{(ki)}): i=1,2,\cdots,m\}$$

我們將利用這些 x_{ki}'s 當做層界而訂立估算 ${}_k\hat{Y}_u$ 之變方的公式如下：

$$var\ ({}_k\hat{Y}_u)=\frac{1}{k^2N^2}\sum_{i=1}^{m+1}t_i^2s_{yi}^2 \tag{32}$$

式中　$t_i=\frac{1}{k-1}\sum_{j=1}^{k-1}(x_{k+j}-1)$

$$t_i=\frac{1}{k-1}\sum_{j=1}^{k-1}(x_{k\,i+j}-x_{k\,i-j}) \text{ for } i=2,\cdots,m$$

$$t_{m+1}=\frac{1}{k-1}\sum_{j=1}^{k-1}(N+1-x_{k\,m+j})$$

$$s_{y_i}^2=\frac{1}{k-1}\sum_{j=1}^{k}(y_{(k\,i-k+j)}-\bar{y}_i)^2$$

$$\bar{y}_i=\frac{1}{k}\sum_{j=1}^{k}y_{(k\,i-k+j)}$$

因爲第 i 層之樣品含有 k 個成對變值：$\{(x_{k\,i-k+j},\ y_{(k\,i-k+j)}):j=1,\cdots,k\}$，故該層層內 y 變值之變方的無偏推算值爲 $s_{y_i}^2$，因此該層樣品平均值 $\bar{y}_i$ 之變方的無偏推算值爲 $s_{y_i}^2/k$。再者，當我們固定 $\{(x_{ki},\ y_{(ki)}):i=1,\cdots,m\}$ 時，第 i 層之族群大小爲 $x_{k\,i+1}-x_{k\,i-k-1}$（實際上應爲 $x_{k\,i+1}-x_{k\,i-k-1}-1$），固定 $\{(x_{k\,i+1},\ y_{(k\,i+1)}):i=1,\cdots,m\}$ 時，第 i 層之族群大小爲 $x_{k\,i+2}-x_{k\,i-k}$（實際上應爲 $x_{k\,i+2}-x_{k\,i-k}-1$），依此類推，我們可得 $k-1$ 種第 i 層之族群大小，將其平均即得 t_i，因此某一附件無偏推算值 $\hat{Y}_i$ 之變方的推算式可利用分層推算原理而設立如下：

$$\Sigma\left(\frac{t_i}{N}\right)^2\frac{s_{y_i}^2}{k}$$

上式中忽略了有限族群改正項 (Finite Population Correction Term)。玆因 ${}_k\hat{Y}_u$ 係 k 個附件無偏推算值之平均，故其變方之推算式應以 k 除上式而得如 (32) 式。

現在我們擬在 (23) 式之直線數學模式的前提下，求出 $\text{var}({}_k\hat{Y}_u)$ 之期望值並與 (29) 式之 $\text{Var}({}_k\hat{Y}_u)$ 比較之，以察其適否可當做 $\text{Var}({}_k\hat{Y}_u)$ 之推算式。在 n 個 x_i's 固定之下，我們很容易證明次式成立：

$$E(s_{y_i}^2|x_i\text{'s})=S_e^2+B^2s_{x_i}^2$$

式中

$$s_{x_i}^2=\frac{1}{k-1}\sum_{j=1}^{k}(x_{k\,i-k+j}-\bar{x}_i)^2$$

$$\bar{x}_i=\frac{1}{k}\sum_{j=1}^{k}x_{k\,i-k+j}$$

因此，在 n 個 x_i's 固定之下，$\text{var}({}_k\hat{Y}_u)$ 之附件期望值爲

$$E[\text{var}({}_k\hat{Y}_u)\,|\,x_i\text{'s}]=\left(\frac{S_e}{kN}\right)^2\sum_{i=1}^{m+1}t_i^2+\left(\frac{B}{kN}\right)^2\sum_{i=1}^{m+1}t_i^2s_{x_i}^2$$

將上式再對 x_i 作期望時，我們可得 $\text{var}({}_k\hat{Y}_u)$ 之期望值。現在先求 t_i^2 之期望值於下：

$$E(t_1^2)=\frac{knC}{6(k-1)}(8k-7)+\left[\frac{3k(N+1)}{2(n+1)}\right]^2=E(t_{m+1}^2)$$

$$E(t_i^2)=\frac{k}{3(k-1)}\left\{(n(2k-1)-3k^2+5k-1)C+3k(k-1)\left(\frac{N+1}{n+1}\right)^2\right\} \text{ for } i=2,\cdots,m$$

式中 $C=(N+1)^2(N-n)/(n+1)^2(n+2)$。將上面之期望值總計並以 k^2N^2 除之，則得

$$\frac{1}{k^2N^2}\sum_{i=1}^{m+1}E(t_i^2)=\frac{1}{n}\left(1+\frac{2k-1}{3k(k-1)}\right)$$

上式僅保留 n^{-1} 級各項。

利用 (26) 及 (27) 兩近似公式繼續求 $t_i^2s_{x_i}^2$ 之期望值：

$$E(t_i^2s_{x_i}^2)=E\left[t_i^2\left\{\frac{1}{k}\sum_{j=1}^{k}x_{k\,i+j}^2-\frac{2}{k(k-1)}\sum_{l<h}x_{k\,i+j}x_{k\,i+h}\right\}\right]$$

式中

$$E(t_i^2x_{k\,i+j}^2)=\text{Var}(t_ix_{k\,i+j})+[E(t_ix_{k\,i+j})]^2$$

$$E(t_i^2x_{k\,i+j}x_{k\,i+h})=\text{Cov}(t_ix_{k\,i+j},\ t_ix_{k\,i+h})+[E(t_ix_{k\,i+j})][E(t_ix_{k\,i+h})]$$

經過冗長的演算後簡化而得：

$$E(t_i^2s_{x_i}^2)=\frac{nk^4(N+1)^2(N-n)}{12(n+1)^4(n+2)}$$

因此

$$\left(\frac{B}{kN}\right)^2\sum_{i=1}^{m+1}E(t_i^2s_{x_i}^2)\fallingdotseq\frac{k}{n^3}\rho^2S_y^2$$

故 $\mathrm{var}({}_k\hat{Y}_u)$ 之期望值如下：

$$E[\mathrm{var}({}_k\hat{Y}_u)]=E[E\{\mathrm{var}({}_k\hat{Y}_u)|x_i's\}]$$

$$=\frac{S_y^2}{n}\left(1+\frac{2k-1}{3k(k-1)}\right)+\frac{k}{n^3}\rho^2S_y^2$$

將上式與 (29) 式之 $\mathrm{Var}({}_k\hat{Y}_u)$ 比較結果可知到 n^{-1} 級項爲止兩式相等。因此，假如 n 相當大而 k 不太大時，我們可利用 (32) 式作爲 ${}_k\hat{Y}_u$ 之變方的推算式。

七、參 考 文 獻

(1) Cochran, W.G. (1963): Sampling Techniques. Second edition. John Wiley & Sons, Inc., New York.

(2) Hansen, M.H., Hurwitz, W.N. and Madow, W.G. (1953): Sample Survey Methods and Theory. Vol. 2. John Wiley & Sons, Inc., New York.

(3) Hartley, H.O. and Ross, A. (1954): Unbiased ratio estimator. Nature 174:270-271.

(4) Mickey, R. (1959): Some finite population unbiased ratio and regression estimators. Journal of Ameriban Statistical Association 54:594-612.

(5) Wey, I.T. (1966): Estimation of the Mean Using the Rank Statistics of An Auxiliary Variable. Unpublished Ph.D. thesis. Library, Iowa State University, Ames, Iowa, U.S.A.

(6) Wilks, S.S. (1962): Mathematical Statistics. John Wiley & Sons, Inc., New York.

(7) Williams, W.H. (1961): Generating unbiased ra io and regression estimators. Biometrics 17: 267-274.

(8) Williams, W.H. (1962): On two methods of unbiased estimation with auxiliary variates. Journal of American Statistica lAssociation 57:184-186.

Estimation of the Mean by Pseudo-post-stratification of an Auxiliary Variate

by

ING-TZER WEY

Summary

In a finite population consisting of N pairs of two variates, we arrange the pairs in increasing order of magnitude with respect to the auxiliary z-variates and denote the ordered pairs by $(z_{(i)}, y_{(i)})$. Note that if there are equally valued variates in the population, they can be arranged in order at random. Suppose that a simple random sample of size n is drawn without replacement from the population. Let $z_{(i)}$ denote the i-th order statistic of the z-variate and $y_{(i)}$ denote the observed value associated with the order statistic $z_{(i)}$. Further, let x_i denote the rank of the z-value in the population to which the i-th order statistic $z_{(i)}$ in the sample is equal. Since the sampling distribution of x_i is the same as that of $z_{(i)}$, being hypergeometric distribution, and the random variate x_i is simpler to handle than the random variate $z_{(i)}$, we have employed the pairs $(x_i, y_{(i)})$ instead of $(z_{(i)}, y_{(i)})$ to estimate the population mean $\bar{Y}$.

An unbiased estimator ${}_k\hat{\bar{Y}}_u$ for the population mean $\bar{Y}$ derived in this study is expressed in Equation (22). This estimator is obtained by averaging k conditionally unbiased estimators which are each derived by fixing a different set of paired observations by which the population is divided into m strata with respect to the x_i-variates after selection of the sample of size n and then applying the method of stratified estimation.

In some population, e.g. where the regression of y on x is bell-shaped, it is numerically shown that the unbiased estimator ${}_k\hat{\bar{Y}}_u$ is more precise than the unbiased ratio estimator derived by Hartley and Ross (1954). The unbiased estimator ${}_k\hat{\bar{Y}}_u$ is also compared with the ordinary post-stratified estimator and it is shown that the former is more precise than the latter under the assumption of a linear model and equal stratum sizes. The comparison was based on the approximation to order of n^{-1}. An estimator for the variance of ${}_k\hat{\bar{Y}}_u$ is derived and the result is displayed in Equation (32). An optimum number k of conditionally unbiased estimators contained in ${}_k\hat{\bar{Y}}_u$ is given in Equation (30) under the assumption of a linear model.