UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

COLEGIO DE CIENCIAS Y HUMANIDADES

UNIDAD ACADEMICA DE LOS CICLOS

PROFESIONALES Y POSGRADO

INSTITUTO DE MATEMATICAS
APLICADAS.
~~AVANZADAS~~ Y SISTEMAS.

ALGUNOS DESARROLLOS
DE LA
TEORIA DEL MUESTREO

QUE PARA OBTENER EL GRADO DE

MAESTRO EN ESTADISTICA E INVESTIGACION
DE OPERACIONES.

P R E S E N T A

BLANCA ROSA PEREZ SALVADOR.

JULIO 1983.

INDICE.

INTRODUCCION.

Debido a la importancia que tiene el muestreo, entre las técnicas estadísticas, consideramos de interés profundizar en algunos de sus conceptos. Algunos tópicos especializados no estan, a nuestro juicio, exhaustivamente tratados en los textos usuales y desconocemos si existen reportes en la literatura que los estudia. Se tomaron algunos de estos temas y por medio de desarrollos algebraicos, se obtuvieron resultados que esperamos ayuden para su comprensión, y así determinar que esquema de muestreo es más conveniente usar en un caso dado, al aumentar el número de opciones disponibles.

El trabajo consta de 3 capítulos.

En el capítulo I, se desarrolla la expresión de la varian- varianza del estimador bajo un esquema de muestreo secuencial p.p.t. sin reemplazo. Después se encuentra una expresión para la varianza del estimador de razón insesgado en una población particular, finalmente se compara esta varianza con la varianza del estimador p.p.t. con reemplazo. Estos resultados, hasta donde sabemos, no se reportan en la literatura; sólo Des Raj lo hace para $n = 2$.

En el capítulo II se encuentra la expresión de la varianza del estimador de regresión. Des Raj sólo presenta una expresión asintótica de dicha varianza. En este trabajo se encontró la expresión exacta para cualquier n, aunque en una población particular. También en este capítulo se exploró la posibilidad de tener un estimador de regresión doble. Dicho estimador se dá como una aportación en este trabajo, al igual que la expresión de su sesgo que resultó susceptible de ser corregido.

Finalmente, en el capítulo III se propone un estimador de la varianza para el estimador del total en un muestreo polietápico, donde las unidades primarias de muestreo (u.p.m.) se seleccionan por p.p.t. con reem-

plazo y el submuestreo se realiza sólo una vez, para obtener el estimador del total como una suma ponderada de cada subtotal por el número de veces que cada u.p.m. aparece en la muestra. Este estimador, al igual que el estimador de regresión doble, es una aportación de este trabajo.

RESUMEN POR CAPITULOS.

CAPITULO I.

1. Varianza en Muestreo Secuencial

Aquí obtenemos una expresión y damos una cota para la varianza del estimador de la forma

$$\hat{Y} = y_1 + y_2 + \ldots + y_{n-1} + y_n \frac{1 - \sum_{\cdot}^{n-1} p_i}{p_n} \qquad \text{en}$$

donde cada unidad en la muestra es seleccionada en forma secuencial, sin reemplazo y con probabilidad proporcional al tamaño de las unidades restantes.

2. Varianza para el Estimador de Razón Insesgado

En este caso se presenta la forma general de la varianza para el estimador de razón insesgado. Des Raj desarrolla éste y el caso anterior para $n = 2$.

CAPITULO II.

1. Media y Varianza del estimador de Regresión

En esta parte se encuentra la expresión para la varianza del estimador de regresión en una población particular. Este estimador resultó ser insesgado.

2. Estimador de Regresion Doble

Aquí se analizan las propiedades y se dan expresiones para el estimador de regresión con dos medidas auxiliares conocidas. Esto es una generalización de lo tratado para el estimador de regresión usual.

CAPITULO III.

1. Estimador de la Varianza del Estimador de $\hat{\bar{Y}}$ en el esquema C (Des Raj)

Se propone un estimador insesgado de la varianza en un muestreo polietápico, donde las unidades primarias de muestreo se seleccionan con reemplazo y el estimador se obtiene calculando para estas u.p.m. el submuestreo sólo una vez, ponderando al estimador resultante por las veces que aparece cada u.p.m. en la muestra de primera etapa.

NOTACION.

U_i unidad i-ésima de la población que puede ser: un individuo, una familia, una hectarea etc.

N tamaño de la población.

n tamaño de la muestra.

Y_i característica de interés medida en la unidad i de la población.

X_i característica auxiliar medida en la unidad i de la población.

X total de la variable X (suma de X_i, i de 1 a N).

Y total de la variable Y.

$\bar{X}$ promedio de la variable X, $\bar{X} = X/N$.

$\bar{Y}$ promedio de la variable Y.

$p_i = X_i/X$ probabilidad de seleccionar la unidad i por p.p.t.

$p_i' = X_i/(X - \sum_{j=1}^{k} X_{i_j})$ $\quad i_j \neq i \quad j = 1, \ldots, k.$

y_i medida de la característica Y en la i-ésima unidad muestral.

Sx_i suma sobre los n valores obtenidos en la muestra.

$\Sigma X_i = \Sigma x_i$ suma sobre los N valores de la población.

A_k conjunto aleatorio cuyos elementos son las primeras k unidades seleccionadas en la muestra.

$E_{A_k}(\,.\,) = \Sigma\,(\,.\,)P(\,A_k$ es parte de la muestra))

$\Sigma_{A_k} y_i y_j \ldots y_t$ suma sobre los indices tales que $u_i, u_j, \ldots, u_t \in A_k$ y que $i < j < \ldots < t$.

$E_{u_i}(\,.\,) = \Sigma\,(\,.\,)P(U_i$ es elemento de la muestra).

$$\sum_{i<j} = \sum_{i=1}^{N-1} \sum_{j=i+1}^{N}$$

$\acute{\Sigma}$ sumatoria sobre todas las posibles muestras de tamaño n.

$\tilde{\Sigma}$ sumatoria sobre todas las posibles muestras manteniendo constante el número de unidades en cada partición de la población.

$\ddot{\Sigma}$ sumatoria sobre los elementos dentro de una partición.

$$E_1(E_2(E_3(X))) = E_Y(E_Z(E_X(X \mid Y,Z))).$$

CAPITULO I

Desarrollos para la varianza del estimador del total bajo dos esquemas de muestreo.

1. VARIANZA EN MUESTREO SECUENCIAL

Este primer desarrollo es una generalización del caso desarrollado en la sección III.24 de teoría del muestreo de Des Raj[1], para n = 2.

Cuando n = 2, se seleccionan dos unidades de una población con el siguiente esquema: La primera selección se hace con probabilidad p_i basada en X_i (i = 1, 2,..., N), y la segunda selección se hace con probabilidad proporcional a las unidades restantes. En esta situación la probabilidad de que se seleccione U_i cuando se sabe que U_j es la primera unidad seleccionada, está dada por $p_i/(1 - p_j)$. Con estos valores se dan los estimadores:

$$t_1 = \frac{y_1}{p_1} \quad y \quad t_2 = y_1 + y_2 \frac{1 - p_1}{p_2}$$

donde y_1 y y_2 corresponden a los valores de la primera y segunda selección.

Ahora bien,

$$E(t_1) = \Sigma\left(\frac{y_i}{p_i}\right) p_i = \Sigma y_i = Y \qquad y,$$

$$E(t_2) = E_1E(t_2|t_1) = E\left(y_1 + \sum_{j\neq 1} y_j \frac{(1-p_1)}{p_j} \quad \frac{p_j}{1-p_1}\right) =$$

$$= E\left(y_1 + \sum_{j\neq 1} y_i\right) = E(Y) = Y \qquad (1.1.1)$$

de esto se deduce que t_1 y t_2 son estimadores insesgados de Y.

Para calcular la varianza de estos estimadores utilizamos el resultado III.25 del Des Raj; y para t_1 tenemos.

$$V(t_1) = \Sigma\left(\frac{y_i}{p_i} - Y\right)^2 p_i = \Sigma\left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2 x_i x_j \qquad (1.1.2)$$

y para t_2

$$V(t_2) = E_1V_2(t_2) + V_1E_2(t_2).$$

pero como $E_2(t_2) = E(t_2|t_1) = Y$ entonces

$$V_2E_2(t_2) = 0 \quad y \quad V(t_2) = E_1V_2(t_2). \qquad (1.1.3)$$

Ahora calculamos $V_2(t_2)$.

$$V_2(t_2) = \sum_{j\neq 1}\left(y_1 + y_j\frac{1-p_1}{p_j} - \Sigma y_i\right)^2 p'_j = \sum_{j\neq 1}\left(\frac{y_j}{p'_j} - \sum_{i\neq 1} y_i\right)^2 p'_j$$

$$= \sum_{\substack{i,j\neq 1\\ i<j}} x_i x_j\left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2 \qquad (1.1.4)$$

sustituimos (1.1.4) en (1.1.3) y nos queda:

$$V(t_2) = E_1\left(\sum_{i,j\neq 1} x_i x_j\left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2\right) = \sum_{k=1}^{n}\left\{\sum_{i,j\neq k} x_i x_j\left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2\right\}p_k$$

$$= \sum_{i,j}\sum_{\substack{k\neq i\\ k\neq j}} x_i x_j\left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2 p_k = \sum_{i,j} x_i x_j\left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2(1-p_j-p_i)$$

(1.1.5)

Comparando (1.1.2) y (1.1.5) se deduce que $V(t_2) < V(t_1)$.

Por otro lado, $E(t_1t_2) = E_1(t_1E(t_2|t_1)) = Y^2$. lo que implica que $Cov(t_1,t_2) = 0$, esto es t_1 y t_2 no están correlacionadas y por lo tanto el estimador

$t = (t_1 + t_2)/2$ es insesgado y su varianza es:

$$V(t) = \frac{V(t_1) + V(t_2)}{4} < \frac{V(t_1)}{2} \qquad (1.1.6)$$

La generalización se da cuando n es cualquier valor y cada unidad en la muestra es seleccionada secuencialmente sin reemplazo y con probabilidad proporcional al tamaño de las unidades restantes, esto es:

u_1 se obtiene con probabilidad X_1/X

u_2 se obtiene con probabilidad $X_2/X-X_1$

$\vdots$

u_n se obtiene con probabilidad $X_n/(X - \sum_{i=1}^{n-1} X_i)$

así $P(u_{(1)},u_{(2)},\ldots,u_{(n)}) = \frac{X_1}{X}\,\frac{X_2}{X - X_1}\cdots\frac{X_n}{X - \sum_{i=1}^{n-1} X_i}$

(muestra con un orden dado).

Generalizando, los estimadores de Y son:

$$t_1 = \frac{y_1}{p_1}\,,\quad t_2 = y_1 + y_2\frac{1 - p_1}{p_2}\,,\ldots,\quad t_n = \sum_{i=1}^{n-1} y_i + y_n\left(\frac{1 - \sum_{i=1}^{n-1} p_i}{p_n}\right)$$

Para facilitar el manipuleo algebraico, en el cálculo de los momentos de estos estimadores, introducimos los siguientes conjuntos aleatorios.

$$A_n = \{u_1, u_2, u_3, \ldots, u_n\}$$

$$A_{n-1} = \{u_1, u_2, u_3, \ldots, u_{n-1}\}$$

$$\vdots$$

$$A_1 = \{u_1\}$$

la suma es sobre las permutaciones de A_i.

Ayudandonos de estos conjuntos calculamos la varianza de t_n

$$V_{A_n}(t_n) = E_{A_{n-1}\,u_n}(V(t_n|A_{n-1})) + V_{A_{n-1}\,u_n}(E(t_n|A_{n-1}))$$

para el segundo sumando tenemos

$$E(t_n|A_{n-1}) = \sum_{A_{n-1}} Y_i + \sum_{A^c_{n-1}} Y_k \left(\frac{1 - \sum_{A_{n-1}} p_i}{p_k}\right) \frac{p_k}{1 - \sum_{A_{n-1}} p_i} = \sum_{i=1}^{N} Y_i = Y$$

de lo que resulta que

$$V(E(t_n|A_{n-1})) = V(Y) = 0$$

y de aqui la varianza queda como:

$$V(t_n) = E(V(t_n|A_{n-1})) \qquad \text{valor que procedemos a calcular}$$

$$V(t_n|A_{n-1}) = E\left(\left(\sum_{A_{n-1}} Y_i + Y_n\left(\frac{1 - \sum_{A_{n-1}} p_i}{p_n}\right) - \sum_{A_{n-1}} Y_i - \sum_{A^c_{n-1}} Y_i\right)^2 |A_{n-1}\right) =$$

$$= E\left(\frac{Y_n}{p_n} - \sum_{A^c_{n-1}} Y_i\right)^2 = \sum_{i<j} x_i x_j \left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2 . \quad u_i, u_j \in A^c_{n-1}$$

Utilizando este resultado,

$$E_{u_{n-1}}\left(\sum_{A_{n-1}^c} x_i x_j \left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2 \mid A_{n-2}\right) = \sum_{A_{n-2}^c}\left(\sum_{A_{n-2}^c - \{u_k\}} x_i x_j \left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2\right) \frac{p_k}{1 - \sum_{A_{n-2}} p_i} \qquad (1.1.7)$$

observése que,

$$V(t_n) = E_{A_{n-1}}\left(\sum_{\substack{A_{n-1}^c \\ i<j}} x_i x_j \left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2\right) = E_{A_{n-2}}\left(E_{u_{n-1}}\left(\sum x_i x_j \left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2 \mid A_{n-2}\right)\right)$$

$$= \sum_{u_k \in A_{n-2}^c} \frac{1}{1 - \sum_{A_{n-2}} p_i}\left(\sum_{A_{n-2}^c - \{u_k\}} x_i x_j \left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2 p_k\right) =$$

$$= \frac{1}{1 - \sum_{A_{n-2}} p_i} \sum_{A_{n-2}^c} x_i x_j \left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2 \sum_{A_{n-2}^c - \{u_i\, u_j\}} p_k$$

$$= \frac{1}{1 - \sum_{A_{n-2}} p_i} \sum_{A_{n-2}^c} x_i x_j \left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2 \left(1 - \sum_{A_{n-2}} p_s - p_i - p_j\right).$$

$$= \sum_{A_{n-2}^c} x_i x_j \left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2 - \frac{1}{1 - \sum_{A_{n-2}} p_i} \sum_{A_{n-2}^c} x_i x_j \left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2 (p_i + p_j)$$

finalmente si la expresión anterior se sustituye en (1.1.7) resulta

$$V(t_n) = E_{A_{n-2}}\left(\sum_{\substack{A_{n-2}^c \\ i<j}} x_i x_j \left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2\right) - E_{A_{n-2}}\left(\frac{1}{1 - \sum_{A_{n-2}} p_i} \sum_{\substack{A_{n-2}^c \\ i<j}} x_i x_j \left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2 (p_i + p_j)\right)$$

cuyo primer sumando no es otra cosa que $V(t_{n-1})$, quedando,

$$V(t_n) = V(t_{n-1}) - E\left(\frac{1}{1 - \sum_{A_{n-2}} p_i} \sum_{\substack{A_{n-2}^c \\ i<j}} x_i x_j \left(\frac{y_i}{x_i} - \frac{y_j}{x_j}\right)^2 (p_i + p_j)\right) \qquad (1.1.8)$$

pero el término que se está restando es positivo, de aquí se sigue que:

$$V(t_n) < V(t_{n-1}) \quad \text{para} \quad n \geq 2.$$

Ahora calcularemos la covarianza de t_i y t_j $\quad i \neq j$.

$$E(t_i\ t_j) = E(t_i\ E(t_j|A_i)) \text{ con } i < j$$

$$E(t_j|A_i) = E(\bar{t}_j\ |A_i,\ A_{j-1}-A_i) = E(t_j|A_{j-1}) = y$$

lo que implica que $E(t_i\ t_j) = E(t_i\ Y) = Y^2$, e igual que antes t_i y t_j no están correlacionadas y el estimador

$$t = \frac{t_1 + t_2 + \ldots t_n}{n}$$ es insesgado y su varianza

$$V(t) = \frac{V(t_1) + V(t_2) + \ldots + V(t_n)}{n^2}$$

y dado que $V(t_1) > V(t_2) > \ldots > V(t_n)$ se sigue finalmente que

$$V(t) < \frac{V(t_1)}{n}$$

Por último, este estimador es mejor que el muestreo con igual probabilidad si $V(t_1) < \frac{N-n}{N} S_y^2$.

2. VARIANZA PARA EL ESTIMADOR DE RAZON INSESGADO Y PRECISION RELATIVA DE ESTE ESTIMADOR

Este segundo desarrollo es la generalización de lo presentado en la sección V.11 del Des Raj, para n = 2.

El esquema de muestreo utilizado en esta parte es sin reemplazo, seleccionando la primera unidad en la muestra con probabilidad proporcional al tamaño y las n-1 unidades restantes con igual probabilidad. Este esquema se conoce como muestreo con probabilidad proporcional al tamaño agregado (ppta).

El estimador es $\hat{Y} = X \frac{Sy_i}{Sx_i}$ el cual resulta ser insesgado y con varianza igual a

$$V(\hat{y}) = (N')^{-1} X\Sigma \frac{(Sy_i)^2}{Sx_i} - Y^2 \qquad (1.2.1)$$

$$N' = \binom{N-1}{n-1}$$

Para entender mejor el contenido de esta sección se presenta los desarrollos hechos por Des Raj para n = 2.

En este caso se tiene a y_1 y y_2 en la muestra, y_1 seleccionada con p.p.t. y y_2 con probabilidad $1/(N-1)$.

El estimador es:

$$\hat{Y} = X \frac{y_1 + y_2}{x_1 + x_2}$$ y la varianza de ésto es

$$V(\hat{Y}) = \frac{1}{N-1} X\Sigma'_{i<j} \frac{(y_i+y_j)^2}{x_i+x_j} - Y^2 = X\Sigma'_{i<j}(y_i+y_j)\frac{\left(\frac{y_i+y_j}{x_i+x_j} - R\right)}{N-1}$$

utilizando el hecho que $Y = X \cdot R$.

Supondremos que la población está particionada en m conjuntos ajenos y no vacíos, con cardinalidad N_i, i de 1 a m, y con la propiedad que si dos elementos Y_i y Y_j están en un mismo conjunto de la partición, entonces $X_i = X_j$, en cambio si Y_i y Y_j están en conjuntos diferentes entonces $X_i \neq X_j$, con esto tenemos una clase de equivalencia.

También supondremos que para cada conjunto en la partición, X y Y se relacionan con el siguiente modelo lineal.

$$Y_{im} = RX_i + e_{im}$$

con $\sum_m e_{im} = E(e_{im}|i) = 0$

y $\sum_m e_{im}^2 = V(e_{im}|i) = a\, N_i\, X_i^g$

Así entonces, la varianza del estimador es

$$V_{ppta} = (N-1)^{-1} X \sum \frac{(e_{im} + e_{jm})^2}{x_i + x_j} =$$

$$= [2(N-1)]^{-1} aX \left[2 \sum_i \sum_{j>i} N_i N_j \frac{x_i^g + x_j^g}{x_i + x_j} + \sum N_i (N_i - 2) x_i^{g-1}\right]$$

y la varianza del estimador ppt con reemplazo en la misma población es

$$V_{ppt} = aX\sum \frac{N_i x_i^{g-1}}{2} = [2(N-1)]^{-1} aX \left[\sum_i \sum_{i<j} N_i N_j (x_i^{g-1} + x_j^{g-1}) + \sum_i N_i (N_i - 1) x_i^{g-1}\right]$$

de este modo la cantidad $\Delta = V_{ppt} - V_{ppta}$ está dada por

$$= [2(N-1)]^{-1} aX \left[\sum N_i x_i^{g-1} - \sum\sum \frac{N_i N_j (x_i - x_j)(x_i^{g-1} - x_j^{g-1})}{x_i + x_j}\right]$$

analizando la cantidad Δ se puede ver directamente que si $g = 0$ ó $g = 1$, entonces dicha cantidad es mayor que cero.

Para $g = 2$, se tiene que el muestreo ppt será superior siempre que

$$\sum_{j>i}\sum \frac{N_i N_j (x_i - x_j)^2}{x_i + x_j} > X$$

para $g = 3$ $\Delta = (aX/2)[N/(N-1)][\bar{x}^2 - (u-1)V(x)]$. . En este caso el muestreo ppt producirá una varianza menor si $\bar{X}^2 - (N-1)V(x) < 0$.

Ahora compararemos con un muestreo de igual probabilidad,

$$V_{ep} = \frac{N(N-2)}{2}\left(R^2 S_x^2 + a \sum \frac{N_i x_i^g}{N-1}\right)$$

por lo tanto cuando $g = 1$, el muestreo ppta es superior al de igual probabilidad, pero sólo un poco mejor que el muestreo ppt. La generalización a cualquier valor de n se efectuará a partir del estimador y su varianza los cuales son:

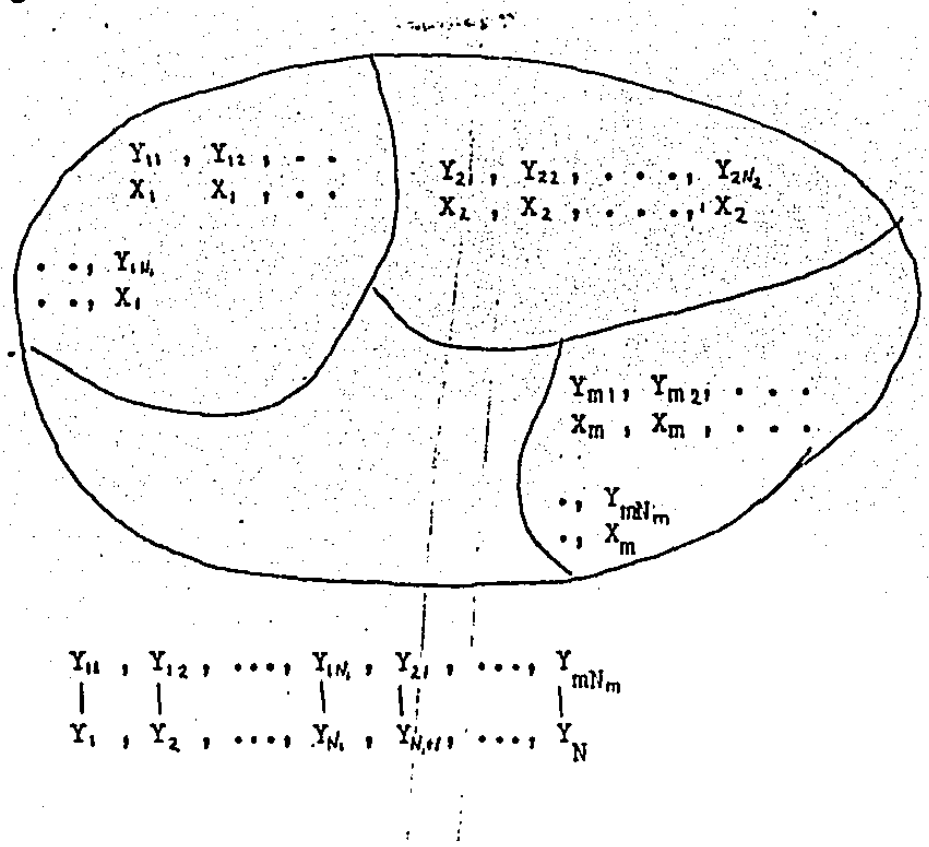
$$\hat{Y} = X \frac{Sy_i}{Sx_i}$$

$$V(\hat{Y}) = (N^1)^{-1} X \sum' \frac{(Sy_i)^2}{Sx_i} - Y^2. \qquad (1.2.1)$$

Como puede apreciarse, en la expresión de la varianza se tiene como sumandos términos con denominador diferente de acuerdo a la muestra de la cual provienen. Para tener denominadores comunes haremos las siguientes suposiciones.

Primero: la población presenta una partición de m subconjuntos definidos por una medida común, X_i y con cardinalidad N_i. Esto es, analizaremos el estimador en una población particular, la que se esquematiza en la si-

guiente figura.



Segundo: X y Y dentro de cada suboconjunto se relacionan mediante el modelo lineal siguiente:

$$Y_{im} = RX_i + e_{im} \equiv Y_j = RX_j + e_j \ , \quad j = \sum_{k=0}^{i-1} N_k + m \ ; \quad N_o = 0$$

con $\quad E(e_{im}|X_i) = 0 \quad$ y $\quad E(e_{im}^2|X_i) = aN_iX_i^g$

Entonces dentro de este contexto calcularemos la varianza de Y, primeramente encontraremos la expresión para $(Sy_i)^2$.

$$(Sy_i)^2 = (R\ S\ x_i + Se_i)^2 = R(Sx_i)^2 + 2R\ Sx_i\ Se_i + (Se_i)^2 =$$

$$\frac{(Sy_i)^2}{Sx_i} = R^2Sx_i + 2R\ Se_i + \frac{(Se_i)^2}{Sx_i}$$

luego sumamos sobre todas las posibles muestras.

$$\acute{\Sigma}\frac{(Sy_i)^2}{Sx_i} = R^2\ \acute{\Sigma}(Sx_i) + 2R\ \acute{\Sigma}(Se_i) + \acute{\Sigma}\frac{(Se_i)^2}{Sx_i}$$

$$= R^2XN' + 0 + \acute{\Sigma}\frac{(Se_i)^2}{Sx_i} = \frac{Y^2N'}{X} + \acute{\Sigma}\frac{(Se_i)^2}{Sx_i}$$

finalmente utilizando este resultado tenemos que:

$$V(\hat{Y}) = \frac{X(R^2XN') + X\ \acute{\Sigma}\frac{(Se_i)^2}{Sx_i}}{N'} - Y^2$$

$$= (N')^{-1}X\ \acute{\Sigma}\frac{(Se_i)^2}{Sx_i} \qquad (1.2.3)$$

Ahora analizaremos la sumatoria $\acute{\Sigma}\frac{(Se_i)^2}{Sx_i}$ donde se puede factorizar los términos con igual denominador, o sea, los términos provenientes de las muestras con igual número de unidades en cada elemento de la partición. Tomemos, por ejemplo, las muestras con n_1 elementos del grupo i_1, n_2 del grupo $i_2,\ldots,n_k$ elementos del grupo i_k, $n_1 + n_2 + \ldots + n_k = n$, siendo de la forma $y_{i_1,1}, y_{i_1,2}, y_{i_1,3},\ldots, y_{i_1,n_1},\ldots, y_{i_2,n_1},\ldots, y_{i_k n_k}$

El número de muestras diferentes con estas características es,

$$\binom{N_{i_1}}{n_{i_1}}\binom{N_{i_2}}{n_{i_2}}\cdots\binom{N_{i_k}}{n_{n_k}}$$

y un mismo elemento Y_{i_r} aparece en $\prod_{j\neq r}\binom{N_{i_j}}{n_{i_j}}\cdot\frac{n_{i_r}}{N_{i_r}}$ muestras diferentes.

El denominador común en estos sumandos es Sn_ix_i, y el numerador

correspondiente es

$$\sum''(Se_{im})^2 = \sum''(Se_{im}^2 + 2Se_{im}\, e_{jm'}) = \pi\binom{N_{i_j}}{n_{i_j}} \sum''_{m}(\Sigma e_{i_r m}^2)\binom{n_{i_r}}{N_{i_r}} + 2\sum''Se_{i_r m} e_{j_r m'} =$$

$$= \pi_{j}\binom{N_{i_j}}{n_{i_j}}\, a \sum'' x_{i_r}^{g} n_{i_r} + 2\sum''Se_{i_r m}\, e_{j_r m'} \qquad (1.2.4)$$

Por comodidad, cambiaremos la notación en el segundo sumando, i_r será i e i_j será j. Así analizaremos dos casos, $i = j$ o $i \neq j$.

Para $i \neq j$ tenemos:

$$\sum''_{i\neq j} S\, e_{im}\, e_{jm'} = \Sigma e_{im} \binom{N_i - 1}{n_j - 1} \sum_{m} e_{jm'} = 0$$

y para $i = j$ tenemos

$$\sum''_{m\neq m'} S\, e_{im}\, e_{im'} = \Sigma e_{im} \binom{N_i - 2}{n_i - 2} \sum_{m\neq m'} e_{im'} = \Sigma e_{im}\,(-e_{im}) \binom{N_i - 2}{n_i - 2}$$

$$= -\binom{N_i - 2}{n_i - 2}\, \Sigma e_{im}^2 = -\binom{N_i - 2}{n_i - 2}\, aN_i x_i^g = -\binom{N_i}{n_i} \frac{n_i(n_i - 1)}{N_i - 1}\, a x_i^g$$

sustituyendo en (1.2.4) y factorizando queda

$$\sum''(Se_i)^2 = \pi\binom{N_i}{n_i} \sum''' a x_i \left(n_i - \frac{n_i(n_i - 1)}{N_i - 1}\right) = \pi\binom{N_j}{n_j} \sum''' a_i x_i^g \frac{n_i(N_i - n_i)}{N_i - 1}$$

para finalmente llegar a la expresión de la varianza

$$V(\hat{Y}) = a(N')^{-1} X \sum'' \pi \binom{N_j}{n_i} \sum''' \frac{n_i(N_i - n_i)}{N_i - 1}\, x_i^g \cdot (\sum''' n_i x_i)^{-1}$$

$$= a\binom{N-1}{n-1}^{-1} X \sum_i'' \pi_j \binom{N_j}{n_j} \sum''' \frac{n_i(N_i-n_i)}{N_i-1} x_i^g \left(\sum' n_i x_i\right)^{-1}$$

(1.2.5)

Este término lo podemos comparar con la varianza de otros estimadores, por ejemplo, con la del estimador ppt con reemplazo la cual es:

$$Vppt = \frac{1}{n} \Sigma\, p_i \left(\frac{y_i}{p_i} - Y\right)^2 = \frac{1}{n} \Sigma\, p_i \left(\frac{RX_i + e_{im}}{p_i} - Y\right)^2 =$$

$$= \frac{1}{n} \Sigma\, p_i \left(RX_i + \frac{e_{im}}{p_i} - Y\right)^2 = \frac{1}{n} \Sigma\, p_i \left(\frac{e_{im}}{p_i}\right)^2 = \frac{X}{n} \Sigma\, \frac{(e_{im})^2}{X_i} =$$

$$= \frac{(N')^{-1} aX}{n} \sum_j'' \pi_j \binom{N_j}{n_j} \sum''' n_i x_i^{g-1} \quad . \qquad (1.2.6)$$

y para $g = 1$ tenemos:

$$Vppta = a(N')^{-1} X \frac{1}{n} \sum_i'' \pi \binom{N_j}{n_j} \sum''' \frac{n_i(N_i-n_i)}{N_i-1} x_i \left(\sum''' n_i x_i\right)^{-1}$$

$$Vppt = a(N')^{-1} X \frac{1}{n} \sum_i'' \pi \binom{N_j}{n_j} \sum''' n_i \left(\frac{\sum''' n_i x_i}{\sum''' n_i x_i}\right)$$

finalmente, tomando la diferencia resulta

$$Vppta - Vppt = \frac{a(N')^{-1} X}{n} \sum_i'' \pi \binom{N_j}{n_j} \frac{\sum''' n_i \left(\frac{N_i - n_i}{N_i - 1} - 1\right) x_i}{\sum n_i x_i}$$

$$= a(N')^{-1} X \sum'' \pi \binom{N_j}{n_j} \sum''' n_i \frac{(N_i - n_i - N_i + 1)x_i}{N_i - 1} \Big/ \sum n_i x_i$$

$$= a(N')^{-1} X \sum'' \pi \binom{N_j}{n_j} \sum''' \left(- \frac{n_i (n_i - 1)}{N_i - 1} \right) x_i / \sum n_i x_i \leq 0$$

Con lo que concluimos que Vppta es menor que Vppt cuando $g = 1$.

CAPITULO II.

Estimadores de Regresión.

1. SESGO Y VARIANZA DEL ESTIMADOR DE REGRESION

La referencia de este desarrollo se encuentra en la sección V.14 de Teoría de Muestreo de Des Raj.

Como antecedente, incluiremos lo realizado por Des Raj.

El estimador de regresión es:

$$\hat{\bar{Y}} = \bar{y} - \hat{b}(\bar{x}-\bar{X}) \qquad (2.1.1)$$

$$\text{donde } \hat{b} = \frac{S(y_i-\bar{y})(x_i-\bar{x})}{S(x_i-\bar{x})^2}$$

Este estimador es no lineal y en general es sesgado. Ademas presenta gran dificultad para obtener una expresión simple de sus momentos, debido al denominador del segundo sumando.

Des Raj da una aproximación de la varianza de este estimador cuando n es grande, aproximación que aquí también presentamos.

Se recordará que el coeficiente de regresión b de la muestra converge en probabilidad hacia un valor finito, a saber B, el coeficiente de regresión de toda la población; así la variable aleatoria $\sqrt{n}\ (\hat{b}-B)(\bar{x}-\bar{X})$ converge hacia cero. Por lo tanto, la distribución limite de

$\sqrt{n}(\bar{y}-b(\bar{x}-\bar{X})-\bar{Y}) = \sqrt{n}[\bar{y}-B(\bar{x}-\bar{X})-\bar{Y}-(b-B)(\bar{x}-\bar{X})]$ será la misma que la de $\sqrt{n}[\bar{y}-B(\bar{x}-\bar{X})-\bar{Y}]$. De este modo, para n grande la varianza del estima-

dor de regresión es:

$$V(\hat{\bar{y}}) = \frac{S_y^2 + B^2S_x^2 - 2B\rho S_xS_y}{n} = \frac{S_y^2(1 - \rho^2)}{n}$$

puesto que $B = \rho S_y/S_x$.

De aquí se puede ver que para muestras grandes el estimador de regresión es mejor que la media usual, pero se debe tener presente que su uso se justifica sólo cuando el costo de una muestra grande no excede a las ventajas obtenidas por el cálculo.

En general no se puede dar una expresión simple para $E(\hat{\bar{Y}})$, por los denominadores de $\hat{b}$, sin embargo podemos particularizar y suponer que se tiene una población con las mismas características que se dieron en el primer capítulo, esto es, una población particionada en m subconjunto donde X y Y se relacionan linealmente dentro de cada elemento de la partición.

Ahora calculamos la media de este estimador,

$$\hat{\bar{Y}} = b\bar{x} + a + \bar{e} - (b + \frac{S(e_i - \bar{e})(x_i - \bar{x})}{S(x_i - \bar{x})^2}(\bar{x} - \bar{X})$$

$$E(\hat{\bar{Y}}) = b\bar{X} + a + E_1\frac{1}{N}E_2\bar{e} + bE(\bar{x} - \bar{X}) + \frac{E_1(\bar{x} - \bar{X})\ \Sigma E_2(e_i - \bar{e})(x_i - \bar{X})}{S(x_i - \bar{x})^2}$$

(2.1.3)

$$= b\bar{X} + a.$$

De aquí resulta que $\hat{\bar{Y}}$ es insesgado y su varianza es:

$$V(\hat{\bar{Y}}) = E_1V_2(\hat{\bar{Y}}) + V_1E_2(\hat{\bar{Y}})$$

pero tenemos que:

$$E_2(\hat{\bar{Y}}) = b\bar{x} + a + E_2(\bar{e}) - (b + \frac{S(x_i-\bar{x})(E_2(e_i-\bar{e})}{S(x_i-\bar{x})^2})(\bar{x}-\bar{X})$$

$$= b\bar{x} + a - b(\bar{x}-\bar{X}) = a + bX$$

por lo que $V_1(E_2(\hat{\bar{Y}})) = 0$.

Al anularse el segundo sumando de la varianza, ésta nos queda como:

$$V(\hat{\bar{Y}}) = E_1V_2(\hat{\bar{Y}}) = E_1(E_2(\hat{\bar{Y}}^2) - (E_2(\hat{\bar{Y}}))^2) \qquad (2.1.4)$$

Ahora sólo hay que encontrar el valor de cada uno de los dos términos en la expresión

$$\hat{\bar{Y}}^2 = \bar{y}^2 - 2\bar{y}\hat{b}(\bar{x}-\bar{X}) + \hat{b}^2(\bar{x}-\bar{X})^2$$

y cada sumando nos da

$$\bar{y}^2 = (b\bar{x} + a + \bar{e})^2 = (b\bar{x} + a)^2 + 2(b\bar{x} + a)\bar{e} + \bar{e}^2$$

$$2\bar{y}\hat{b}(\bar{x}-\bar{X}) = 2(\bar{x}-\bar{X})(b\bar{x} + a + \bar{e})(b + \frac{Sx_ie_i - n\bar{x}\bar{e}}{S(x_i-\bar{x})^2})$$

$$= 2(\bar{x}-\bar{X})[b(b\bar{x} + a) + b\bar{e} + (b\bar{x} + a)\frac{Sx_ie_i - n\bar{x}\bar{e}}{S(x_i-\bar{x})^2} + \frac{Sx_ie_i\bar{e} - n\bar{x}\bar{e}^2}{S(x_i-\bar{x})^2}]$$

$$\hat{b}^2 = b^2 + 2b\frac{Sx_ie_i - n\bar{x}\bar{e}}{S(x_i-\bar{x})^2} + \frac{(Sx_ie_i)^2 - 2n\bar{x}\bar{e}\,Sx_ie_i + n^2\bar{x}^2\bar{e}^2}{(S(x_i-\bar{x})^2)^2}$$

Para obtener $E_2(\hat{\bar{Y}})$ se necesita a $E_2(\bar{e})$ y $E_2(\bar{e}^2)$. La esperanza condicional se calcula dadas las distintas celdas de la partición inducida por $X_1, \ldots, X_m$

Tomemos una muestra con las siguientes medidas $X_{i_1}, X_{i_2}, X_{i_3}, \ldots, X_{i_r}$ y $n_{i_1}, n_{i_2}, \ldots, n_{i_r}$ unidades de cada medida respectivamente.

$$n = \sum_{j=1}^{r} n_{ij} \quad ; \text{ y calculemos los valores esperados requeridos}$$

Primeramente tenemos:

$$\bar{e} = \frac{1}{n} (Se_{i_1 j} + Se_{i_2 j} + \ldots + Se_{i_r j}) =$$

$$= \frac{1}{n} (n_{i_1}\bar{e}_{i_1} + n_{i_2}\bar{e}_{i_2} + \ldots + n_{i_r}\bar{e}_{i_r})$$

cuyo valor esperado es:

$$E_2(\bar{e}) = \frac{1}{n} E_2\left(\sum_k E_3\left(\sum_{j_k}^{n_k} e_{i_k j_k}\right)\right) =$$

$$= \frac{1}{n} E_2\left(\sum_{k=1}^{r} \frac{\binom{N_{i_k}-1}{n_{i_k}-1}}{\binom{N_{i_k}}{n_{i_k}}} \sum^{N_{i_k}} e_{i_k j_k}\right) = 0 \qquad (2.1.6)$$

Enseguida encontramos el otro valor esperado

$$E_2(\bar{e}^2) = \frac{1}{n^2}E_2S(\overset{n_k}{S} e_{i_k j_k})^2 + \frac{1}{n^2}E_2 \overset{''}{S} E_3(Se_{i_k j_k})(Se_{i_k j_k})$$

$$= \frac{1}{n^2}E_2\overset{r}{S} E_3(\overset{n_r}{S} e_{i_k j_k})^2 + \frac{1}{n^2}E_2\overset{''}{S} E_3(Se_{i_k j_k})(Se_{i_k j_k}) \qquad (2.1.7)$$

$$E_3(\overset{n_k}{\underset{j_k}{S}} e_{i_k j_k})^2 = E_3(S\, e^2_{i_k j_k} + S_{i_k j_k}\, e_{i_k j_k}) = \frac{\binom{N_{i_k}-1}{n_{i_k}-1}}{\binom{N_{i_k}}{n_{i_k}}} \sum_{j=1}^{N_{i_k}} e_{i_k j} +$$

$$+ \frac{\binom{N_{i_k}-2}{n_{i_k}-2}}{\binom{N_{i_k}}{n_{i_k}}} \sum_{j=1}^{N_{i_k}} e^2_{i_k j} = \frac{\binom{N_{i_k}-1}{n_{i_k}-1}}{\binom{N_{i_k}}{n_{i_k}}} N_{i_k}\, \sigma^2_{i_k} = \frac{n_{i_k}(N_{i_k}-n_{i_k})}{N_{i_k}-1}\, \sigma^2_{i_k}$$

$$E_3((\overset{n_{ir}}{S} e_{i_k j})(Se_{i_{k'} j}) \mid i_k \neq i_{k'}, n_{i_k}, n_{i_{k'}}) = \frac{1}{N'} \Sigma(Se_{i_k j})\Sigma(Se_{i_{k'} j}) =$$

$$(2.1.8)$$

$$= 0.$$

Sustituyendo (2.1.8) en (2.1.7) tenemos

$$E_2(\bar{e}^2) = \frac{1}{n}\sum^{r} \frac{n_{i_k}(N_{i_k}-n_{i_k})}{N_{i_k}-1}\, \sigma^2_{i_k} \qquad (2.1.9)$$

y con los resultados (2.1.6) y (2.1.9) podemos calcular los valores siguientes:

$$E_2(\bar{y}^2) = (bx + a)^2 + \frac{1}{n^2}\Sigma \frac{n_{i_k}(N_{i_k} - n_{i_k})}{N_{i_k}-1}\, \sigma^2_{i_k}$$

$$E_2(\bar{y}\hat{b}(\bar{x}-\bar{X})) = (\bar{x}-\bar{X})(b(b\bar{x}+a)) + \frac{E_2(S\, n_{i_k}\, x_{i_k}\, e_{i_k}\, \bar{e} - n\bar{x}\bar{e}^2)}{S(x_i - \bar{x})^2}$$

$$= (\bar{x}-\bar{X})(b(b\bar{x}+a) + \frac{S\, x_{i_k} \frac{n_{i_k}(N_{i_k}-n_{i_k})}{N_{i_k}-1} - n\bar{x}\, S\frac{n_{i_k}(N_{i_k}-n_{i_k})}{N_{i_k}-1}}{S\,(x_i-\bar{x})^2}$$

y

$$E_2(\hat{b}^2) = b^2 + (S\frac{n_{i_k}(N_{i_k}-n_{i_k})}{N_{i_k}-1}\, x_{i_k}^2\, \sigma_{i_k}^2 - 2\bar{x}\, S\, x_{i_k}\frac{n_{i_k}(N_{i_k}-n_{i_k})}{N_{i_k}-1} +$$

$$+ \bar{x}^2\, S\frac{n_{i_k}(N_{i_k}-n_{i_k})}{N_{i_k}-1}\, \sigma_{i_k}^2)\, \frac{1}{S(x_i-\bar{x})^2} = \qquad (2.1.10)$$

$$b^2 + (S(x_{i_k}^2 - 2\bar{x}x_{i_k} + \bar{x})\, \frac{n_{i_k}(N_{i_k}-n_{i_k})}{N_{i_k}-1}\, \sigma_{i_k})\, \frac{1}{(S(x_i-\bar{x})^2)^2}$$

$$= b^2 + (S(x_{i_k}-\bar{x})^2\, \frac{n_{i_k}(N_{i_k}-n_{i_k})}{N_{i_k}-1}\, \sigma_{i_k})\, \frac{1}{(S(x_i-\bar{x})^2)^2}$$

Y con estos tres últimos resultados ya podemos encontrar el valor de

$$E_2(\hat{\bar{Y}}^2) = (b\bar{x}+a)^2 + \frac{1}{n^2}\, S\frac{n_{i_k}(N_{i_k}-n_{i_k})}{N_{i_k}-1} -$$

(2.1.11)

$$-2(\bar{x}-\bar{X})\left(b(b\bar{x}+a) + \frac{1}{n}\left(S(x_{i_k}-\bar{x})\frac{n_{i_k}(N_{i_k}-n_{i_k})}{N_{i_k}-1}\sigma^2_{i_k}\right)\right)\frac{1}{S(x_i-\bar{x})^2}$$

$$+(\bar{x}-\bar{X})^2\left(b^2+\left((S(x_i-\bar{x})^2\frac{n_{i_k}(N_{i_k}-n_{i_k})}{N_{i_k}-1}\sigma^2_{i_k}\right)\right)\frac{1}{S(x_i-\bar{x})^2}$$

$$= (bx+a)^2 - 2(\bar{x}-\bar{X})b(b\bar{x}+a) + (\bar{x}-\bar{X})^2b^2 +$$

$$\left(\frac{1}{n} - \frac{2(\bar{x}-\bar{X})\frac{1}{n}(x_i-\bar{x})}{S(x_i-\bar{x})^2} + \frac{(\bar{x}-\bar{X})^2(x_i-\bar{x})^2}{(S(x_i-\bar{x})^2)^2}\right)\frac{n_{i_k}(N_{i_k}-n_{i_k})}{N_{i_k}-1}$$

$$= (b\bar{x}+a-b\bar{x}+\bar{X}b)^2 + S\left(\frac{1}{n} - \frac{(x_i-\bar{x})(\bar{x}-X)}{S(x_i-\bar{x})^2}\right)\frac{n_{i_k}(N_{i_k}-n_{i_k})}{N_{i_k}-1}\sigma^2_{i_k}$$

$$= (b\bar{X}+a)^2 + S\left(\frac{1}{n} - \frac{(\bar{x}_i-\bar{X})(x_i-\bar{x})}{S(x_i-\bar{x})^2}\right)\frac{n_{i_k}(N_{i_k}-n_{i_k})}{N_{i_k}-1}\sigma^2_{i_k} \qquad (2.1.11)$$

de todo esto se concluye que la varianza es igual a:

$$V(\hat{\bar{Y}}) = E_1V_2(\hat{\bar{Y}}) = \frac{1}{\binom{N}{n}}\sum\prod_{k=1}^{r}\binom{N_{i_k}}{n_{i_k}}''\Sigma\left(\frac{1}{n} - \frac{(\bar{x}-\bar{X})(x_i-\bar{x})}{S(x_i-\bar{x})^2}\right)\frac{n_{i_k}(N_{i_k}-n_{i_k})}{N_{i_k}-1}\sigma^2_{i_k}$$

(2.1.12)

Donde si $n < N_i$ para alguna i entonces no se garantiza que $\hat{b}$ existe.

2. ESTIMADOR DE REGRESION DOBLE

Este desarrollo es una generalización del estimador tratado anteriormente, aqui consideraremos dos medidas auxiliares conocidas, X_1 y X_2, para todas y cada una de las unidades de la población.reo.

X_1, X_2 y Y se relaciónan linealmente mediante el siguiente modelo,

$$Y = a + b_1X_1 + b_2X_2 + e \qquad (2.2.1)$$

Así en lugar de estimar a Y directamente, se estimará el valor de a, $a = Y - b_1X_1 - b_2X_2$ y después se corregirá con los valores de b_1X_1 y b_2X_2, lo que da como estimador del promedio a:

$$\hat{\bar{Y}} = \bar{y} - b_1(\bar{x}_1 - \bar{X}_1) - b_2(\bar{x}_2 - \bar{X}_2) \qquad (2.2.2)$$

Si b_1 y b_2 son constantes, el estimador es insesgado. Para calcular su varianza nos auxiliaremos de la variable $u_i = y_i - b_1x_{i1} - b_2x_{i2}$ asi:

$$V(\hat{\bar{Y}}) = V(\bar{u}) = \frac{1}{n}(1-f)S_u^2 \quad ; \quad f = \frac{n}{N} \qquad (2.2.3)$$

Ahora calculamos S_u^2

$$S_u^2 = (N-1)^{-1}\Sigma(u_i-\bar{u})^2 = (N-1)^{-1}\Sigma(y_i-\bar{Y}-b_1(x_{i1}-\bar{X}_1) - b_2(x_{i2}-X_2))$$

$$= (N-1)^{-1}(\Sigma(y_i-\bar{Y})^2 + b_1^2\Sigma(x_{i1} - \bar{X}_1)^2 + \Sigma b_2^2(x_{i2} - X_2)^2 -$$

$$-2b_1\Sigma(y_i-\bar{Y})(x_{i1} - \bar{X}_1) - 2b_2\Sigma(y_i-\bar{Y})(x_{i2} - \bar{X}_2) +$$

$$+2b_1b_2\Sigma(x_{i1} - \bar{X}_1)(x_{i2} - \bar{X}_2))$$

Este resultado lo sustituimos en (2.2.3) quedando:

$$V(\hat{\bar{Y}}) = \frac{1}{n}(1-f)(S_y^2 + b_1^2 S_{x_1}^2 + b_2^2 S_{x_2}^2 - 2b_1\rho_{x_1y} S_{x_1} S_y - 2b_2\rho_{x_2y} S_{x_2} S_y +$$

$$+2b_1 b_2 \rho_{x_1x_2} S_{x_1} S_{x_2}) \qquad (2.2.4)$$

Para determinar en esta expresión los valores de b_1 y b_2 que minimizan la varianza se procede a la derivación, igualación a cero y finalmente a la resolución del sistema de ecuaciones resultante quedando.

$$b_k = \frac{S_y \rho_{xky} - \rho_{x_1x_2} \rho_{xjy}}{S_{x_k} \quad 1 - \rho_{x_1x_2}^2} \qquad k \neq j$$

Derivando por segunda vez la misma expresión obtenemos los elementos matriciales siguientes,

$$\frac{\partial^2 V(\hat{\bar{Y}})}{\partial x_i \partial x_j} = \begin{cases} S_{x_i}^2 & \text{si} \quad i \neq j \ . \\ 0 & \text{si} \quad i \neq j \ . \end{cases}$$

esto es, los elementos de una matriz diagonal positiva definida. Lo que implica que los valores encontrados de b_1 y b_2 son un mínimo de $V(\bar{Y})$. Este estimador será mejor que el promedio muestral cuando:

$$b_1^2 S_{x_1}^2 + b_2^2 S_{x_2}^2 + 2b_1 b_2 \rho_{x_1x_2} S_{x_1} S_{x_2} - 2b_1 \rho_{x_1y} S_{x_1} S_y - 2b\rho_{x_2y} S_{x_2} S_y < 0$$

De aquí, se puede ver, que se necesita conocer el valor de S_y^2 y el de ρ_{x_iy} para calcular a b_1 y b_2. Cuando desconocemos ambos va-

lores, se tiene la alternativa de estimar a b_1 y b_2. Así proponemos los estimadores siguientes, basados en los estimadores de regresión.

$$b_i = \frac{Sy_j(x_{ji} - \bar{X}_i)\ S_{x_k}/S_{x_i} - \rho_{x_1x_2}\ Sy_i(x_{j_k} - \bar{X}_k)}{S_{x_1} S_{x_2}(1-\rho^2_{x_1x_2})} \qquad (2.2.5)$$

Para encontrar los valores esperados de $\hat{b}_k(\bar{x}_k - \bar{X}_k)$ es necesario calcular $E(Sy_i(x_{ik} - \bar{X}_k)(S(x_{ij} - \bar{X}_j)) = E(Su_i\ Sv_i)$ entonces:

$$E(Su_i Sv_i) = \frac{n}{N} \Sigma u_i v_i + \frac{n(n-1)}{N(N-1)} \sum_{i \neq j} u_i v_j$$

$$= \frac{n(N-n)}{N(N-1)} \Sigma u_i v_i + \frac{n(n-1)}{N(N-1)} \Sigma u_i\ \Sigma v_i$$

$$= \frac{n(N-n)}{N(N-1)} \Sigma u_i v_i .$$

Utilizando estos resultados llegamos a que:

$$E(\hat{\bar{Y}}) = \bar{Y} - \frac{N-n}{N(N-1)} \frac{\Sigma y_i(x_{i_1} - \bar{X}_1)^2 S_{x_2}/S_{x_1} + \Sigma y_i(x_{i_2} - \bar{X}_2)^2 S_{x_1}/S_{x_2} - 2\rho_{x_1x_2} \Sigma y_i(x_{i_1} - \bar{X}_1)(x_{i_2} - \bar{X}_2)}{S_{x_1} S_{x_2}(1-\rho^2_{x_1x_2})} \qquad (2.2.6)$$

Como se ve, este estimador es sesgado, pero a partir de él podemos tener un estimador insesgado sumando el término

$$\frac{N-n}{n(N-1)} \frac{C(x_1, x_2, y)}{S_{x_1} S_{x_2}(1-\rho^2_{x_1x_2})}$$

donde

$$C(x_1,x_2,y) = S_{y_i}(x_{i1}-\bar{X}_1)^2 S_{x_2}/S_{x_1} + S_{y_i}(x_{i2} - \bar{X}_2)^2 -$$

$$-2\rho_{x_1x_2}\, S_{y_i}(x_{i1}-X_1)(x_{i2}-X_2)$$

ya que

$$E(Sy_i(x_{i_k}-\bar{X}_k)(x_{ij}-\bar{X}_j)) = \frac{n}{N}\Sigma y_i(x_{i_k}-\bar{X}_k)(x_{ij}-\bar{X}_j).$$

Entonces el estimador

$$\hat{\bar{Y}}' = \hat{\bar{Y}} + \frac{N-n}{n(N-1)}\,\frac{C(x_1,x_2,y)}{S_{X_1}S_{X_2}(1-\rho_{x_1x_2})} =$$

$$= \bar{y} - C'(x_1,x_2,y)$$

es insesgado.

Este estimador será mejor que el de la media usual si

$$2E(C'(x_1,x_2,y)\cdot(\bar{y} - \bar{Y})) + E(C'(x_1,x_2,y)^2) < 0.$$

CAPITULO III.

1. ESTIMADOR DE LA VARIANZA DE $\hat{Y}$ EN EL ESQUEMA C (Des Raj).

En este último desarrollo, se encuentra un estimador insesgado de la varianza del estimador del total, de un esquema de muestreo polietápico donde las unidades primarias de muestreo se selecionan por ppt con reemplazo y el submuestreo de segunda étapa se realiza sólo una vez a cada upm en la muestra y el estimador final se cálcula ponderando el estimador de segunda étapa por el número de veces que cada upm aparecio en la muestra.

El estimador del total es entonces:

$$\hat{Y} = \frac{1}{n} \Sigma \frac{\lambda_i T_i}{p_i} \quad . \qquad (3.1.1)$$

Donde; λ_i es el número de veces que aparece en la muestra la i-ésima upm $\lambda_i = 0,1,2,\ldots,n$ y $\Sigma\lambda_i = n$.

Hipótesis que cumple el modelo:

T_i es un estimador insesgado de Y_i con las siguientes características

$$E_2(T_i) = Y_i \qquad y$$

$$V_2(T_i) = V(T_i|i) = \sigma_i^2 \qquad (3.1.2)$$

Se conoce un estimador insesgado de σ_i^2 para toda i.

Así

$$E(\lambda_i) = np_i, \quad E(\lambda_i^2) = n(n-1)p_i + np_i, \quad y$$

$$E(\lambda_i\lambda_j) = n(n-1)p_ip_j \tag{3.1.3}$$

Utilizando esto, podemos calcular el valor esperado del estimador,

$$E(\hat{Y}) = \frac{1}{n}\sum E_1\left(\frac{\lambda_i}{p_i} E_2(T|i)\right) = \frac{1}{n}\sum E\left(\frac{\lambda_i}{p_i} Y_i\right) = \frac{1}{n}\sum \frac{np_iY_i}{p_i} =$$

$$= \sum Y_i = Y \tag{3.1.4}$$

y la varianza

$$V(\hat{Y}) = E_1\left(\frac{1}{n^2}\sum \lambda_i^2 \frac{\sigma_i}{p_i}\right) + V_1\left(\frac{1}{n}\sum \lambda_i \frac{Y_i}{p_i}\right) =$$

$$= \frac{1}{n}\sum \frac{\sigma_i^2}{p_i} + \frac{n-1}{n}\sum \sigma_i^2 + \frac{1}{n}\sum p_i\left(\frac{Y_i}{p_i} - \bar{Y}\right)^2 =$$

$$= \frac{1}{n}\sum \frac{\sigma_i}{p_i} + \frac{n-1}{n}\sum \sigma_i^2 + \frac{1}{n}\left(\sum \frac{Y_i^2}{p_i} - Y^2\right) \tag{3.1.5}$$

Ahora vamos a probar que el siguiente término es un estimador insesgado de la varianza.

$$\hat{V}(\hat{Y}) = \frac{1}{n(n-1)}\sum \lambda_i\left(\frac{T_i}{p_i} - \frac{1}{n}\sum \lambda_i \frac{T_i}{p_i}\right)^2 + \frac{1}{n}\sum\frac{\lambda_i\hat{\sigma}_i^2}{p_i} =$$

$$= \frac{1}{n(n-1)}\left(\sum \lambda_i \frac{T_i^2}{p_i} - n\left(\frac{1}{n}\sum \frac{\lambda_iT_i}{p_i}\right)^2\right) + \frac{1}{n}\sum \frac{\lambda_i\sigma_i^2}{p} \tag{3.1.6}$$

Sabiendo que T_i y T_j son independientes para $i \neq j$, encontramos que:

$$E(\Sigma\ \lambda_i \frac{T_i^2}{p_i}) = \Sigma\ \frac{1}{p_i^2}\ E\ \lambda_i E_2 T_i^2 = \Sigma\ \frac{1}{p^2}\ E_1\lambda_i(Y_i^2 + \sigma_i^2) =$$

$$= \Sigma\ \frac{np_i}{p_i^2}\ (Y_i^2 + \sigma_i^2) = n\ \Sigma\ \frac{Y_i^2 + \sigma_i^2}{p_i} \ldots . \qquad (3.1.7)$$

$$E(\frac{1}{n}\ \Sigma\ \frac{\lambda_i T_i^2}{p_i^2}) = \frac{1}{n}\ \Sigma\ \frac{1}{p_i^2}\ E_1(\lambda_i^2 E_2 T_i^2) = \frac{1}{n}\ \Sigma\ \frac{1}{p_i^2}(n(n-1)p_i^2 + np_i)(\sigma_i^2 + Y_i^2) =$$

$$= (n-1)\ \Sigma\ (\sigma_i^2 + Y_i^2)(1 + \frac{1}{p_i}) \qquad (3.1.8)$$

$$E(\frac{2}{n}\ \Sigma\ \lambda_i\lambda_j\ \frac{T_iT_j}{p_ip_j}) = \frac{2}{n}\ \Sigma\ E_1\frac{\lambda_i\lambda_j}{p_ip_j}\ E_2(T_iT_j) = \frac{2}{n}\ \Sigma\ \frac{n(n-1)p_ip_j}{p_ip_j}\ Y_iY_j =$$

$$= 2(n-1)\ \Sigma\ Y_iY_j . \qquad (3.1.9)$$

Y por último

$$E(\Sigma\ \frac{\lambda_i\sigma_i^2}{p_i}) = \Sigma\ E_1\frac{\lambda_i}{p_i}\ E_2\hat{\sigma}_i^2 = \Sigma\ E_1\frac{\lambda_i}{p_i}\ \sigma_i^2 = n\ \Sigma\ \sigma_i^2$$

De todo esto tenemos:

$$E(\hat{V}(\hat{Y})) = \frac{1}{n(n-1)}\ (n\ \Sigma\ (\frac{Y_i^2 + \sigma_i^2}{p_i}) - (n-1)\ \Sigma\ \{\sigma_i^2 + Y_i^2) -$$

$$- \Sigma\ (\frac{\sigma_i^2 + Y_i^2}{p_i}) - 2(n-1)\ \Sigma\ Y_iY_j) + \Sigma\ \sigma_i^2 =$$

$$= \frac{1}{n(n-1)} \left((n-1) \sum \frac{Y_i^2 + \sigma_i^2}{p_i} - (n-1)\left(\sum Y_i^2 + 2 \sum Y_i Y_j\right) - \right.$$

$$\left. - (n-1) \sum \sigma_i^2 \right) + \sum \sigma_i^2 =$$

$$= \frac{1}{n} \left(\sum \frac{Y_i^2}{p_i} - Y^2 \right) + \frac{1}{n} \sum \frac{\sigma_i^2}{p_i} - \frac{1}{n} \sum \sigma_i^2 + \sum \sigma_i^2 =$$

$$= \frac{1}{n} \left(\sum \frac{Y_i^2}{p_i} - Y^2 \right) + \frac{1}{n} \sum \frac{\sigma_i^2}{p_i} + \frac{n-1}{n} \sum \sigma_i^2 = V(\hat{Y})$$

(3.1.9)

con esto se demuestra que $\hat{V}(\hat{Y})$ es insesgada.

Comentarios Finales.

Los resultados obtenidos en general son satisfactorios.

En el capítulo I, se lograron las generalizaciones deseadas.

En el capítulo II, se dió el estimador de regresión doble quedando como problema abierto el estudio de los estimadores de regresión múltiple, estos estimadores parecen ser de expresiones algebraicas complejas.

En el capítulo III, se encontró el estimador insesgado de la varianza de $\hat{\bar{Y}}$, con esto se logró lo que se pretendia.

BIBLIOGRAFIA

- DES RAJ. Samply Tehory (1968) Mc Graw-Hill, Inc. Nueva York.

- DES RAJ. Teoría del Muestreo. Traducción al Español (1980) Fondo de la Cultura Económica. Av. de la Universidad 975. México 12, D.F.