REPUBLICA ARGENTINA

SERVICIO ESTADISTICO NACIONAL

CURSO DE

TEORIA Y PRACTICA DE LA MUESTRA

Profesor: Sr. SIGFRIDO MAZZA

agosto - diciembre 1955

CENTRO DE CAPACITACION TECNICA PARA

FUNCIONARIOS ESTADISTICOS

M.t.p.
1a. a 5a. Conferencia
año: 1955
Prof. Sigfrido Mazza

MUESTREO SIMPLE AL AZAR

En las aplicaciones prácticas de la Teoría de las Muestras, debe tratarse, en general, con "poblaciones" o "universos" finitos que constan (o se definen como estando constituídos) de un número N de individuos o unidades.

Si las unidades se distinguen entre sí (si se las ha numerado o se ha asociado a cada una un símbolo que la particularice), el número de grupos distintos de n unidades que pueden formarse con los elementos de la población es:

$$\binom{N}{n} = C_n^N = \frac{N(N-1)....(N-n+1)}{n!} = \frac{N!}{n!\,(N-n)!} \qquad (1)$$

Esta fórmula, como se sabe, dá el número de grupos (de combinaciones) diferentes que pueden formarse eligiendo n de entre N objetos distintos dados, cuando se consideran como diferentes dos grupos que difieren en al menos uno de los elementos que lo integran.

Así, si se tiene una población constituída por 5 unidades (N = 5) que se distinguen asignándoles respectivamente los números 1,2, 3,4,5, los grupos posibles de extensión n = 3 serán en número de:

$$\binom{5}{3} = \frac{5 \times 4 \times 3}{3 \times 2 \times 1} = 10$$

a saber:

123 124 125 134 135 145

234 235 245 345

Cada uno de los grupos de n tomados de los N constituye una "muestra" posible obtenida de la población.

En el caso del ejemplo, cuando decimos que se ha obtenido la "muestra" 123, significamos que ella está constituída por las unidades de la población que tienen asignados esos números, tomados en el orden dado o en cualquier otro que pueda obtenerse permitiéndolos de todos los modos posibles entre sí.

Afortunadamente no es nunca necesario calcular ni escribir el número de muestras diferentes que son posibles; sólo importa, para fines teóricos, conocer ese número expresado simbólicamente en la forma dada en (1)

El muestreo simple al azar es aquel método de selección de la muestra de extensión n de la población de N que asigna a cada una de las $\binom{N}{n}$ que son posibles la misma probabilidad de ser elegida.

Como la suma de estas probabilidades todas iguales debe ser 1, se sigue que la probabilidad de selección es:

$$1/\binom{N}{n} = \frac{n!}{N(N-1) \ldots\ldots (N-n+1)} \qquad (2)$$

Cuál es el modelo operatorio equivalente al muestreo simple al azar? Supongamos tener una urna que contiene N bolillas prácticamente idénticas numeradas de 1 a N, de la cual, a ciegas y mezclando cuidadosamente el contenido al cabo de cada operación, se realizan n extracciones sucesivas, sin reponer la bolilla obtenida en cada caso. El resultado de este experimento será la obtención de un grupo de n bolillas (la muestra) que presentarán los números

$$y_1 \quad y_2 \ldots\ldots \quad y_n$$

representando con y_1 el número obtenido en la primera prueba, que puede ser uno cualquiera entre 1 y N; y_2, el obtenido en la segunda prueba, que será distinto del y_1, etc.etc. En cada paso de este experimento, todas las bolillas que permanecen en la urna tienen la misma probabilidad de ser extraídas en la siguiente tirada para entrar a formar parte de la muestra. Esta probabilidad es:

$$1/(N-k+1)$$

siendo el denominador el número de bolillas contenidas en la urna al momento de realizar la k-ésima extracción. Se sigue de aquí que, la probabilidad de obtener n bolillas determinadas en un cierto orden también determinado, es:

$$\frac{1}{N} \cdot \frac{1}{(N-1)} \cdot \frac{1}{(N-2)} \cdots\cdots \frac{1}{N-n+1}$$

y la probabilidad de tener las n bolillas determinadas cualquiera sea el orden en que se presenten es:

$$\frac{n!}{N(N-1) \cdots\cdots (N-n+1)}$$

es decir, la (2).

Es interesante observar que, en el muestreo simple al azar, la probabilidad de que una bolilla determinada (una unidad específica de la población que se somete al muestreo) sea obtenida en una cierta extracción (p.e. la r - ésima), es la misma que la probabilidad de obtenerla en la primera, a saber, 1/N. En efecto, la probabilidad de obtener una cierta bolilla en la r- ésima prueba es igual a la probabilidad de que no haya aparecido en ninguna de las r - 1 pruebas precedentes, multiplicada por la probabilidad de que lo haga en la r - ésima, o sea:

$$\frac{N-1}{N} \cdot \frac{N-2}{N-1} \cdots\cdots \frac{N-r+1}{N-r+2} \times \frac{1}{N-r+1} = \frac{1}{N}$$

De aquí se sigue inmediatamente que: en el muestreo simple al azar, la probabilidad de que una determinada unidad de la población aparezca incluída en la muestra es la misma para todas las unidades que integran dicha población e igual a n/N. En efecto, la probabilidad de que una unidad determinada aparezca incluída en la muestra es la suma de las probabilidades de que se la obtenga en la 1a., o en la 2a.,..... o en la n-ésima extracción.

El anterior resultado se ha usado a veces para caracterizar al "muestreo simple al azar". Sin embargo, esta propiedad no es pecífica totalmente a este método de selección pues, como veremos más adelante, en el método denominado de "muestreo sistemático", también cada unidad de la población tiene la misma probabilidad de ser incluída en la muestra, aún cuando cada una de ellas no tiene la misma probabilidad de selección en la primera prueba, como ocurre en el muestreo simple al azar.

PROCEDIMIENTO PRACTICO PARA LA SELECCION DE UNA MUESTRA SIMPLE AL AZAR. USO DE LAS TABLAS DE NUMEROS AL AZAR.

Si bien el modelo de la "urna ideal" es apropiado y útil para la descripción de cómo se opera en el muestreo simple al azar - aunque utilizable -no es ciertamente práctico como mecanismo real para la selección de las unidades de la población que constituirán la muestra a estudiarse.

El procedimiento práctico para la selección de una muestra simple al azar de extensión N de una población que consta de N unidades, es la utilización de una cualquiera de las varias tablas de números al azar que se han publicado (Tippet, Babington, Smith , Fisher y Yates), para lo cual:

1º.- Se asocia a cada una de las N unidades de la población uno de los números de 1 a N, es decir, se listan las unidades que integran la población y se las numera consecutivamente.

2º.- Se eligen N números diferentes de la tabla de números al azar

3º.- Se toman como constituyentes de la muestra aquellas N unidades de la población que tienen asignados números iguales a los obtenidos en el 2º paso.

A continuación se reproduce parte de la tabla de números al azar preparada por Tippet

PARTE DE UNA TABLA DE NUMEROS AL AZAR

(Tomada de L.H.C. Tippet: "Random Sampling Numbers")

	COLUMNAS							
Fila	1-4	5-8	9-12	13-16	17-20	21-24	25-28	29-32
1	1089	8719	9542	2259	0384	2346	4624	9295
2	9385	7902	9902	1726	8340	1105	6299	4638
3	6984	8660	3893	6772	5700	4528	7800	0102
4	0052	1007	1679	2548	9281	5075	8037	3648
5	5736	9249	6809	5204	4007	3703	1079	1855
6	1501	5988	7421	2937	8019	1257	1672	1800
7	5372	6212	2583	3099	5642	7137	4343	2082
8	4677	3280	0729	8112	6422	7283	3769	5997
9	7856	2562	8457	8974	4460	1778	5312	2827
10	4936	2701	2630	4252	5596	8646	2854	1221
11	7315	0454	8995	0492	0604	4924	2990	6230
12	5098	8457	6082	6667	5286	6000	8136	1074
13	0741	5030	0842	4320	2403	1913	2734	7151
14	7357	8286	3267	4637	4620	9960	5325	7420
15	1310	3324	3790	0126	5070	8155	2370	0864
16	8726	2655	1527	1052	5195	1896	6978	8490
17	0599	9464	4486	0104	5588	1734	0009	3787
18	3442	4177	6554	3555	7385	3374	3676	8068
19	0793	2908	4638	6076	1506	1331	5304	6074
20	6390	3064	1555	7487	6841	8040	8147	7143
21	7278	7667	1427	7533	2770	2463	4716	6754
22	4658	9377	0004	3916	9183	9306	5340	9969
23	4898	8746	9396	8608	7086	0580	7298	2247
24	6954	6766	9663	9853	7387	7880	2665	3593
25	5911	3213	2825	3825	9158	4214	8591	0001
26	3272	9820	9278	1527	8392	0728	0286	0527
27	5512	8981	9041	1027	5373	8104	5991	4979
28	7360	5698	4785	2755	7856	7315	9094	2996
29	7691	7658	4800	5153	9189	3122	5380	9624
30	8197	3686	5965	3476	3083	2913	3336	9860
31	2179	2056	7926	1467	8548	9364	9185	0333
32	4335	7878	0526	9327	6831	5494	5131	8603
33	2996	3220	6922	0539	6142	3181	0784	9024
34	7390	1400	9149	7909	1531	8771	9584	1176
35	3146	6354	4403	9799	7302	4241	8118	6640
36	1143	5988	0064	6823	3093	7331	7034	5027
37	6867	9449	0808	8093	0941	9466	2956	8634
38	0423	0308	5871	5691	7660	8541	2660	0588
39	1220	0792	0484	0278	9416	2526	5932	9357
40	0602	4920	5730	2019	8514	9184	8172	5935

Supongamos tener una población que consta de 350 unidades de muestreo, de la que se desea extraer una muestra simple al azar de 10. Para hacer la selección, entramos en un punto cualquiera de la tabla, p.e. la fila 11 y la columna 12, y tomamos para incluir en la muestra aquellas unidades de la población que tienen asignados números iguales a aquellos menores que 351 que se leen en la tabla, a partir de la fila 11 hacia abajo, y están formados por las cifras que pertenecen a las columnas 13, 14 y 15. Procediendo así, encontramos los números:

49 12 105 10 152 102 275 347 146 53

Las unidades de la población que tienen asignados estos números constituirán nuestra muestra.

Siguiendo el anterior procedimiento, se han tenido que leer 23 números de la tabla para obtener los 10 requeridos. Hay pues un cierto "desaprovechamiento" de la tabla, que puede obviarse de varias maneras. Una de ellas consiste en asignar a cada unidad de la población 2 números en lugar de 1, de modo que todo número de la tabla entre 001 y 700 inclusive nos dirige a una unidad de la población. Hay varias maneras de hacer esta asignación:

1º	El par de números	Equivale a
	001 - 002	1
	003 - 004	2
	005 - 006	3
	⋮	⋮
	699 - 700	350

2º	El par de números	Equivale a
	001 - 351	1
	002 - 352	2
	003 - 353	3
	⋮	⋮
	350 - 700	350

Las unidades integrantes de la muestra de 10 seleccionados por uno y otro de los caminos anteriores son:

1º) 25 333 216 232 6 53 5 128 304 196

2º) 49 316 82 113 12 105 10 5 257 41

En ambos casos se han tenido que leer solo 12 números de la tabla para tener los 10 requeridos.

Suele ocurrir a veces que la población total está subdividida en grupos cada uno de los cuales contiene N_1, N_2, N_3, ... N_n, unidades de muestreo (por cierto $\sum N_i = N$). Por ejemplo las viviendas

están agrupadas en manzanas en una ciudad, los departamentos o partidos agrupados en provincias, los abonados telefónicos se agrupan en las páginas de la guía telefónica, etc.

Tratándose de obtener una muestra simple al azar puede utilizarse un procedimiento que no requiere numerar todas las unidades.

Para ejemplificar supongamos una población que consta de 371 unidades que se distribuye en 14 grupos, cada uno de los cuales consta del número de unidades anotado en la segunda columna del cuadro siguiente:

Grupo Nº	Unidades	Acumulado
1	25	25
2	17	42
3	5	47
4	59	106
5	64	170
6	22	192
7	38	230
8	16	246
9	21	267
10	12	279
11	14	293
12	38	331
13	17	348
14	23	371
	371	

Debiendo elegir una muestra al azar de 10 unidades, se construye la columna 3 acumulando las unidades contenidas en cada grupo, y se toman de una tabla de números al azar 10 números menores que 372 ; supongamos que se encuentran los siguientes:

29 72 128 96 326 199 202 58 117 33

Las unidades 29 y 33 caen en el grupo 2
" " 58, 72 y 96 " " " " 4
" " 117 y 128 " " " " 5
" " 199 y 202 " " " " 7
" " 326 " " " " 12

Resulta así que solo será necesario numerar parte de las unidades de 5 de los 14 grupos. Del 2º grupo, entran en la muestra las unidades nº 4 y nº 8 (29 - 25 = 4; 33 - 25 = 8); del 4º grupo entran en la muestra las unidades nº :

58 - 47 = 11
72 - 47 = 25
96 - 47 = 49

etc. etc.

DEFINICIONES Y NOTACION

Toda vez que se realiza un censo completo o una investigación o relevamiento mediante una muestra, se lo hace para tener información acerca de las propiedades o características de una población o universo que se ha elegido y delimitado para constituir una entidad cuyo estudio se estima que puede ofrecer toda o una parte de la evidencia requerida para orientar una decisión o un curso de acción o incrementar el conocimiento. Tanto en el caso de un censo completo como en el de la muestra, la información buscada acerca de las características de la población, se obtiene a partir de la medición, y el registro, bien para cada unidad de la población censada, bien para cada una que ha sido incluída en la muestra, de ciertas "características" tales como edad, estado civil, número de habitaciones, superficie dedicada a un cierto cultivo, número de explotaciones agrícolas, etc, etc. según que las "unidades" que integran el universo objeto de estudio sean individuos, viviendas familiares, explotaciones agrícolas, áreas geográficas, etc.

Sea N el número de unidades que constituyen la población, el "valor" que una cierta caractorística tiene en cada una de ellas se indicará con

$$a_1 \quad a_2 \quad a_3 \ldots\ldots\ldots\ldots a_N$$

siendo a_i el "valor" de la característica considerada en la i-ésima unidad de la población en el orden en que las unidades están listadas en el "padrón" .

En lo que sigue usaremos la siguiente notación:

Valor Total de la característica en la población (Total en la población)

$$A = a_1 + a_2 + \ldots + a_N = \sum_1^N a_i$$

Valor medio de la característica por unidad en la población (valor medio en la población)

$$\bar{A} = \mu_a = A/N = \frac{1}{N} \sum_1^N a_i$$

Se tiene evidentemente que

$$A = N\mu = N\bar{A}$$

Variancia en la población:

$$S^2 = \frac{1}{N-1} \sum_1^N (a_i - \mu_a)^2 = \frac{1}{N-1} \sum_1^N (a_i - \bar{A})^2$$

La expresión de la variancia en la población mediante S^2 (usando N como divisor de la suma de los cuadrados de los desvíos con respecto a la media μ) en lugar de la que se usa corrientemente:

$$\sigma^2 = \frac{1}{N} \sum_{1}^{N} (a_i - \mu)^2$$

tiene la ventaja de que permite escribir en forma más simple los resultados, los que, por cierto, son equivalentes en una u otra notación.

A lo largo de estas lecciones usaremos siempre la denominación de "variancia" para referirnos a S^2, en lugar de la de "cuadrado medio" que suele dársele, usando la expresión "variancia sigma cuadrado" cuando nos referimos a σ^2

En general, en una encuesta o investigación, son varias las características que interesan y que se miden en cada uno de los individuos seleccionados, de modo que tendrán que considerarse los valores de las "características" o "variables" a, b, c, ...- Cada una de estas variables toma, para el i-ésimo individuo de la población, los valores

$$a_i \; , \; b_i , c_i , \; \ldots$$

respectivamente. Si las unidades de la población son, p.e., explotaciones agrícolas, pueden interesar las siguientes características:

a) Superficie total
b) " dedicada al cultivo de cereales
c) Número de cabezas de ganado vacuno, etc., etc.

Toda vez que tengan que considerarse varias características a, b, c, ... en los individuos de una misma población, indicaremos los respectivos valores totales con

$$A \; , \; B \; , \; C \; , \; \ldots$$

los valores medios con

$$\bar{A} \; , \; \bar{B} \; , \; \bar{C} \; , \; \ldots$$

y las variancias con

$$S_a^2 \; , \; S_b^2 \; , \; S_c^2 \; , \; \ldots \; .-$$

Extraída una muestra de extensión n, el valor de la característica considerada en las unidades incluídas en ella se indicarán con:

$$y_1 \quad y_2 \quad \ldots\ldots \quad y_n$$

siendo entonces y_i, el valor de dicha característica en la unidad obtenida en la i-ésima prueba o extracción. y_i, es una variable que puede tomar los valores a_1 a_2 ..... a_N que sean distintos de los tomados por y_1 , y_2 , ... y_{i-1}

El total en la muestra es:

$$y = y_1 + y_2 + \ldots\ldots + y_n = \sum_1^n y_i$$

La media de la muestra es:

$$\bar{y} = y/n = \frac{1}{n} \sum_1^n y_i$$

La variancia en la muestra es:

$$s^2 = \frac{1}{n-1} \sum_1^n (y_i - \bar{y})^2$$

Usando la terminología introducida por R.A. Fisher, los números:

$$A \quad \mu \quad s^2$$

se denominan "parámetros"(de la población). Los correspondientes valores

$$y \qquad \bar{y} \qquad s^2$$

que se calculan con los datos de una muestra, se denominan "estadísticas". A partir de una muestra se calculan "estadísticas" con las que se pretende estimar "parámetros" de la población de donde aquella ha sido obtenida.

Si es t una "estadística" mediante la cual se estima un cierto "parámetro" T, pondremos

$$\text{Est } (T) = \hat{T} = t$$

con lo que significamos que la "estimación de T" está dada por t.

En lo que sigue mostraremos que:

$$\text{Est } (\mu) = \bar{y}$$

$$\text{Est } (A) = N\bar{y} = \frac{N}{n} y$$

$$\text{Est } (s^2) = s^2$$

El factor N/n por el cual se multiplica el total y de la muestra para tener la estimación del total A de la población se denomina "factor de

expansión" , y su inversa n/N es la "tasa" o "fracción" de muestreo.

Importa observar que un "parámetro" es una constante, mientras que una "estadística" es una variable aleatoria función de las observaciones. Así, la media de la muestra

$$\bar{y} = y_1 + y_2 + \ldots\ldots + y_n$$

es una función de las n variables y_1 , y_2 .... y_n.

En el caso del muestreo simple al azar, hemos visto que, de una población de N unidades pueden obtenerse $\binom{N}{n}$ muestras diferentes. Para cada una de ellas puede calcularse la media y_i ($i = 1,2,\ldots \binom{N}{n}$), y, salvo el caso de que todos valores a_i en la población sean iguales entre sí, se tendrán valores distintos para $\bar{y}$, cada uno de los cuales se dará con una cierta frecuencia. Otro tanto ocurrirá para cualquier otra "estadística"; lo que permite decir que las "estadísticas" son variables aleatorias que tienen una cierta distribución que depende de los parámetros de la población y del tamaño de la muestra. El que una determinada "estadística" sea admisible o preferible para la estimación de un parámetro depende de las propiedades de su distribución. Así, en general, una "estadística" que se distribuya de modo tal que su valor medio es igual al valor del parámetro θ que con ella se pretende estimar será utilizable para estimar θ y de varias que tengan la misma propiedad, será preferible aquella que tenga la menor variabilidad.

Para aclarar lo que antecede, como así también para facilitar la comprensión de algunos conceptos que desarrollaremos más adelante, consideremos una población en la que una cierta característica de las unidades que la integran tiene los siguientes valores:

a_1	a_2	a_3	a_4	a_5	a_6
3	6	2	4	4	8

Para esta población se tiene:

$$N = 6 \qquad A = 27 \qquad \mu = 4,5 \qquad S^2 = 4,7$$

A continuación damos la lista de los valores en cada una de las 20 muestras distintas de extensión 3 que pueden obtenerse de la población y también los valores de $\bar{y}$ o $\hat{\theta}$ para cada una de ellas.

Muestra Nº		Suma y	Media $\bar{y}$
1	362	11	3.666
2	364	13	4.333
3	364	13	4.333
4	368	17	5.666
5	324	9	3.
6	324	9	3.
7	328	13	4.333
8	344	11	3.666
9	348	15	5.
10	348	15	5.
11	624	12	4.
12	624	12	4.
13	628	16	5.333
14	644	14	4.666
15	648	18	6.
16	648	18	6.
17	244	10	3.333
18	248	14	4.666
19	248	14	4.666
20	448	16	5.333
		270	

Para calcular la media de las medias de las muestras, observamos que

$$\frac{1}{20} \sum_{j=1}^{20} \bar{y}_j = \frac{1}{20} \sum_{j=1}^{20} y_j / 3 = \frac{1}{60} \sum_{j=1}^{20} y_j$$

de modo que ella es $270 / 60 = 4.5$, es decir, coincide con la media μ de la población. Esto puede expresarse diciendo que la "esperanza matemática" de las medias de las muestras es igual a la media de la población.

Veamos ahora cómo compara la esperanza matemática de la varian cia de las muestras con la variancia de la población.

Recordamos que la variancia de la muestra viene dada por

$$s^2 = \frac{1}{n-1} \sum_{1}^{n} (y_i - \bar{y})^2$$

o bien, expresada en una forma más conveniente para el cálculo:

$$s^2 = \frac{1}{n-1} \left[\sum_{1}^{n} y_i^2 - \frac{\left(\sum y_i \right)^2}{n} \right]$$

En el cuadro siguiente se dan los valores requeridos para el cálculo de la variancia para cada una de las 20 muestras.

Muestra Nº	y_1^2	y_2^2	y_3^2	(1) $\sum y_i^2$	(2) $\left(\sum y_i\right)^2/3$
1	9	36	4	49	121/3
2	9	36	16	61	169/3
3	9	36	16	61	169/3
4	9	36	64	109	289/3
5	9	4	16	29	81/3
6	9	4	16	29	81/3
7	9	4	64	77	169/3
8	9	16	16	41	121/3
9	9	16	64	89	225/3
10	9	16	64	89	225/3
11	36	4	16	56	144/3
12	36	4	16	56	144/3
13	36	4	64	104	256/3
14	36	16	16	68	196/3
15	36	16	64	116	324/3
16	36	16	64	116	324/3
17	4	16	16	36	100/3
18	4	16	64	84	196/3
19	4	16	64	84	196/3
20	16	16	64	96	256/3
				1450	1262

Siendo C_n^N el número de muestras diferentes, el valor medio de las variancias de las muestras será:

$$\frac{1}{C_n^N}\sum_{1}^{C_n^N}\left\{\frac{1}{n-1}\left[\sum_{1}^{n} y_i^2 - \frac{\left(\sum_{1}^{n} y_i\right)^2}{n}\right]\right\} = \frac{1}{n-1}\left\{\frac{1}{C_n^N}\sum_{1}^{C_n^N}\left(\sum_{1}^{n} y_i^2\right) - \frac{1}{C_n^N}\sum_{1}^{C_n^N}\left[\frac{\left(\sum_{1}^{n} y_i\right)^2}{n}\right]\right\}$$

o sea, 1/(n-1) de la diferencia de los promedios de los valores anotados en las columnas (1) y (2) del cuadro anterior. Se tiene así:

$$\frac{1}{2}\left[\frac{1450}{20} - \frac{1262}{20}\right] = 4.7$$

lo que nos dice que la esperanza matemática de las variancias de las muestras es igual a la variancia de la población.

Hemos visto que las medias de las muestras tienen una distribución cuya media coincide con la de la población; una pregunta que ahora cabe es: cuál es la variancia de la distribución de las medias?

En nuestro ejemplo: $N = 6$, $n = 3$, $S^2 = A.^7$, de modo que:

$$\text{Var}(\bar{y}) = \frac{6-3}{6} \cdot \frac{A.^7}{3} = .783 .$$

En el ejemplo que antecede hemos hallado que, el valor medio sobre todas las muestras, tanto de la media como de la variancia de las mismas, coincide con los respectivos valores μ y S^2 de la población.

Cuando ocurre que una "estadística" es tal que para todo n se verifica que su valor medio (o esperanza matemática) es igual al valor del parámetro de la población que con ella se pretende estimar, se dice que dicha "estadística" da, o es, una estimación "no viciada" (unbiased) del parámetro. En otros términos, si

$$t = t(y_1\ y_2\ \ldots\ y_n)$$

es una función de las n observaciones obtenidas en una muestra, y se verifica que

$$E(t) = \theta \text{ para todo } n$$

entonces t es una estimación "no viciada" del parámetro θ

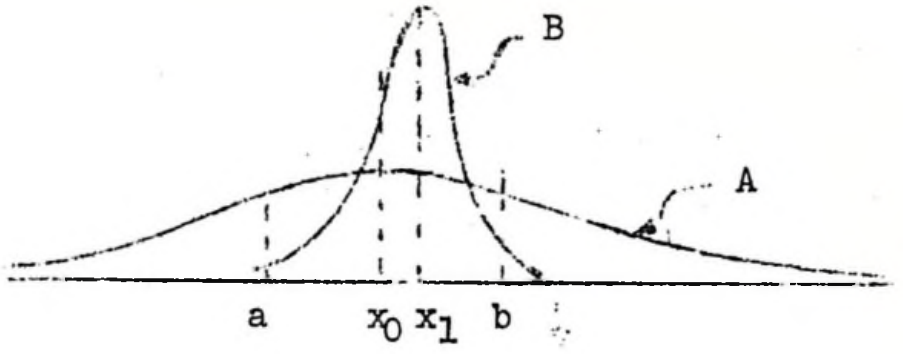
Si

$$E(t) \neq \theta \text{ , la diferencia}$$
$$E(t) - \theta = \delta$$

es el "error sistemático de estimación" o el "vicio" nacido de la utilización de una "función estimadora viciada" de θ .

En la práctica, si bien es preferible utilizar estimaciones "no viciadas", no está totalmente excluído el uso de estimaciones que, dando lugar a un "error sistemático de estimación" pequeño (que puede tender a 0 cuando el tamaño de la muestra crece) aseguran una mayor precisión en la estimación.

En efecto, lo que interesa en la práctica es el riesgo de error de una estimación calculada a partir de una muestra particular, más que el error promedio de las estimaciones obtenidas de todas las muestras posibles.
Supongamos que se tienen dos "estimadores" A y B, el primero de los cuales es "no viciado" y sí el segundo, cuyas respectivas distribuciones son las que se muestran en la figura siguiente:



El valor medio de A coincide con el verdadero valor X_o de la población, pero sólo una proporción relativamente pequeña del área total que representa la distribución del "estimador" está comprendida entre los límites del intervalo (a , b) que se ha adoptado como standard para definir como "aceptable" el error de estimación. El "estimador" B es "viciado" , puesto que su valor medio X_1 difiere de X_o ($X_1 - X_o$ es el "vicio" o " error sistemático de estimación "); pero con una alta frecuencia ofrece estimaciones que son próximas a X_o , de modo que es menor el riesgo de errar cuando se lo use para estimar X_o .
Más adelante consideraremos métodos de estimación que, aunque "viciados" , son , en algunos casos, de utilización preferible en tanto que tienen un error standard mucho menor que los "no viciados" .

De una "estadística" t diremos que ofrece una estimación "consistente" del parámetro θ , cuando para $n = N$ se tiene

$$t(y_1 \; y_2 \ldots\ldots y_N) = \theta$$

Es evidente que la media y la variancia de la muestra son "consistente" en el sentido de la definición.

En el ejemplo numérico de más arriba hemos mostrado que:

a) La esperanza matemática de la media de las muestras es igual a la media de la población.

b) La esperanza matemática de la variancia s^2 de las muestras es igual a la variancia S^2 de la población.

y hemos adelantado que

c) La variancia $V(\bar{y})$ de las medias de las muestras de extensión n está dada por

$$V(\bar{y}) = \frac{N - n}{N} \cdot \frac{s^2}{n}$$

DEMOSTRACIÓN DE a)

En el cálculo de las Probabilidades se demuestra que si es

$$Z = a_1 x_1 + a_2 x_2 + \ldots\ldots + a_n x_n$$

donde las a_i son constantes y las x_i variables aleatorias cualesquiera, se tiene:

$$EZ = a_1 E x_1 + a_2 E x_2 + \ldots\ldots + a_n E x_n = \sum_1^n a_i E x_i$$

Se sigue de aquí inmediatamente que, siendo

$$\bar{y} = \frac{1}{n} \sum_1^n y_i$$

será

$$E(\bar{y}) = \frac{1}{n} \sum_1^n E y_i = \frac{1}{n} \sum_1^n \mu_i$$

poniendo $E y_i = \mu_i$

En el caso particular del muestreo simple al azar $\bar{y}$ es la media de la muestra $y_1 y_2 \ldots\ldots y_n$, y se tiene

$$E y_1 = E y_2 = \ldots\ldots = E y_n = \mu$$

siendo μ la media

$$\frac{1}{N} \sum_1^n a_i$$

de la población de N unidades $a_1, a_2, \ldots\ldots a_N$, de donde la muestra ha sido extraída.

Se ha visto, en efecto, que la probabilidad de que un individuo de la población sea incluído en la muestra en la k-ésima extracción es $1/N$, y como ese individuo puede tener uno cualquiera de los valores $a_1, a_2, \ldots\ldots a_N$, resulta que

$$E(y_n) = \frac{1}{N} \sum_1^N a_i = \mu$$

En resumen, se tiene que en el muestreo simple al azar

$$E(\bar{y}) = \mu$$

lo que nos dice que la media de la muestra es una estimación "no viciada" de la media de la población.

DEMOSTRACION DE c)

La variancia de las medias de las muestras es:

$$V(\bar{y}) = E(\bar{y} - \mu)^2 = E\,\bar{y}^2 - \mu^2 \qquad (1)$$

ahora bien

$$E\,\bar{y}^2 = E\left\{\frac{1}{M}(y_1 + y_2 + \dots + y_n)\right\}^2 = \frac{1}{M^2} E\left[\sum_1^n y_i^2 + \sum_{(i+j)1}^{n} y_i\, y_j\right] =$$

$$= \frac{1}{n^2}\left[\sum_1^n E\, y_i^2 + \sum_{i \neq j = 1}^{n} E\, y_i\, y_j\right] \qquad (2)$$

Tenemos aquí que calcular

1) $E\, y_i^2$ 2) $E\, y_i\, y_j$

1) Sabemos que y_j es una variable aleatoria que toma los valores a_1 , a_2 ..... a_N con probabilidades $1/N$, y por lo tanto

$$E\, y_i^2 = \frac{1}{N} \sum_1^N a_i^2$$

El 2º miembro puede escribirse:

$$\frac{1}{N} \sum_1^N (a_i - \mu + \mu)^2 = \frac{1}{N} \sum_1^N \left\{(a_i - \mu)^2 + 2\mu(a_i - \mu) + \mu^2\right\} =$$

$$= \frac{1}{N} \sum_1^N (a_i - \mu)^2 + 2\mu \frac{1}{N} \sum_1^N (a_i - \mu) + \mu^2$$

y como

$$\frac{1}{N} \sum_1^N (a_i - \mu)^2 = \sigma^2$$

$$\frac{1}{N} \sum_1^N (a_i - \mu) = 0$$

resulta

$$E\, y_i^2 = \sigma^2 + \mu^2 \qquad (3)$$

2) $E\, y_i y_j$ es la esperanza matemática de una variable aleatoria producto de otras dos (y_i e y_j). Esta variable aleatoria toma

M + n I a V

los N(N-1) valores $a_i\ a_j$ que se obtienen apareando de todos los modos posibles dos elementos distintos del conjunto $a_1, a_2, \ldots\ldots a_N$, y cada uno de ellos con la probabilidad $1/N(N-1)$, de modo que

$$E(y_i\ y_j) = \frac{1}{N(N-1)} \sum_{i+j=1}^{N} \sum a_i\ a_j$$

El 2º miembro puede escribirse:

$$\frac{1}{N(N-1)} \sum\sum (a_i - \mu + \mu)(a_j - \mu + \mu) =$$

$$= \frac{1}{N(N-1)} \sum\sum \left[(a_i - \mu)(a_j - \mu) - \mu(a_i - \mu) - \mu(a_j - \mu) + \mu^2\right] =$$

$$= \frac{1}{N(N-1)} \sum\sum (a_i - \mu)(a_j - \mu) + \mu^2$$

puesto que

$$\mu \sum\sum (a_i - \mu) = \mu \sum\sum (a_j - \mu) = 0$$

resulta pues

$$E\ (y_i\ y_j) = \frac{1}{N(N-1)} \sum\sum (a_i - \mu)(a_j - \mu) + \mu^2$$

El primer término del 2º miembro es la "COVARIANCIA" de la población que suele indicarse con $\sigma_{i\ j}$

Para hallar el valor de la covariancia observemos que

$$\frac{1}{N(N-1)} \sum\sum (a_i - \mu)(a_j - \mu) = \frac{1}{N(N-1)} \left[\sum_1^N (a_i - \mu)\right]^2 + \frac{1}{N(N-1)} \sum_1^N (a_i - \mu)^2$$

pero

$$\sum (a_i - \mu) = 0 \quad y \quad \sum_1^N (a_i - \mu)^2 = N\sigma^2$$

de modo que

$$\sigma_{i\ j} = -\frac{\sigma^2}{N-1}$$

y se tiene finalmente

$$E(y_i\ y_j) = \mu^2 - \frac{\sigma^2}{N-1} \qquad (4)$$

Reemplazando en (2) los valores de $E\ y_i$ y $E\ y_i y_j$ hallados en (3) y (4) tomemos:

M.t.p. I a V

$$E\,\bar{y}^2 = \frac{1}{n^2}\left[\sum_{1}^{N}(\sigma^2+\mu^2)+\sum_{i\neq j=1}^{n}\sum\left(\mu^2-\frac{\sigma^2}{N-1}\right)\right]$$

$$= \frac{1}{n^2}\left[n(\sigma^2+\mu^2)+n(n-1)\left(\mu^2-\frac{\sigma^2}{N-1}\right)\right]$$

de donde, después de simples transformaciones algebraicas, se llega a:

$$E\,\bar{y}^2 = \frac{N-n}{N-1}\cdot\frac{\sigma^2}{n}+\mu^2 \qquad (5)$$

y poniendo este resultado en la (1)

$$V(\bar{y}) = \frac{N-n}{N-1}\cdot\frac{\sigma^2}{n} \qquad (6)$$

Recordando nuestra definición de S^2

$$S^2 = \frac{N}{N-1}\sigma^2$$

de (6) se obtiene el resultado buscado:

$$\boxed{V(\bar{y}) = \frac{N-n}{N}\cdot\frac{S^2}{n}} \qquad (7)$$

que más adelante nos detendremos a analizar

DEMOSTRACION DE b

$$E\,s^2 = S^2$$

siendo

$$s = \frac{1}{n-1}\sum_{1}^{n}\left(y_i-\bar{y}\right)^2$$

$$E\,S^2 = \frac{1}{n-1}\,E\sum_{1}^{n}\left(y_i-\bar{y}\right)^2$$

$$= \frac{1}{n-1}\,E\left\{\sum_{1}^{n}y_i^2-n\bar{y}^2\right\}$$

$$= \frac{1}{n-1}\left\{\sum_{1}^{n}E\,y_i^2-n\,E\,\bar{y}^2\right\}$$

más arriba hemos hallado que

$$E\, y_i^{\,2} = \sigma^2 + \mu^2$$

$$E\, \bar{y}^{\,2} = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{} + \mu^2$$

$$E\, s^{\,2} = \frac{1}{n-1} \left\{ n\left(\sigma^2 + \mu^2\right) - \frac{N-n}{N-1}\sigma^2 - n\mu^2 \right\}$$

$$= \frac{1}{n-1} \left\{ n\,\sigma^2 + n\mu^2 - \frac{N-n}{N-1}\,\sigma^2 - n\mu^2 \right\}$$

$$= \frac{1}{n-1} \left\{ \frac{N\,n - n - N + n}{N-1}\,\sigma^2 \right\}$$

obteniéndose finalmente:

$$E\,(S^2) = \frac{N}{N-1}\,\sigma^2 = S^2$$

EL CASO DE PROPORCIONES O PORCENTAJES

Antes de seguir adelante derivaremos las fórmulas que corresponden a las dadas arriba en el caso en que la totalidad de los individuos que integran la población se clasifican en dos grupos según que posean o no una determinada característica (p.e varones o mujeres, propietarios o no propietarios. etc.).-

Si atribuimos el valor 1 a todo individuo que posea la ca racterística "C" considerada y el 0 al que no la posee (posee la característica "no - C"), nuestras a_i serán solo ceros o unos y el total

$$A = \sum_{1}^{N} a_i$$

de la población será igual al número de unidades "C";

$$N - A$$

Será el número de las que poseen la características "no - C".

La media μ de la población será

$$\mu = A/N = P$$

Si representamos en P la fracción de unidades "C" en la población.

Puesto que $A = NP$

$$N - A = N(1 - P) = NQ$$

Q será la fracción de unidades "no - C".

La varianza σ^2 de la población se obtiene inmediatamente observando que en su expresión

$$\frac{1}{N} \sum_{1}^{N} (a_i - \mu)^2$$

la suma consta de A sumandos en los cuales $a_i = 1$ y N - A sumandos en los cuales $a_i = 0$, de modo que

$$\sigma^2 = \left[A\,(1-P)^2 + (N-A)\,P^2 \right]$$

$$= P\,(1-P)^2 + (1-P)\,P^2$$

de donde resulta finalmente

$$\sigma^2 = P\,Q$$

Puesto que, por definición

$$S^2 = \frac{N}{N-1}\sigma^2$$

resulta

$$S^2 = \frac{N}{N-1} PQ$$

Pasando a considerar una muestra de extensión n; se tendrá que:

$$y = \sum_{1}^{n} y_i$$

no es otra cosa que el número de unidades "C" incluídas en la muestra y :

$$\bar{y} = y/n = p$$

la fracción de unidades "C" en la misma.

Si se pone $q = 1 - p$, por el mismo camino seguido más arriba, se obtendrá

$$s^2 = \frac{n}{n-1}\, p.q$$

Por cierto que también en este caso se tendrá:

$$E(\bar{y}) = E(p) = P$$

$$E(s^2) = S^2 = \frac{N}{N-1}\, PQ$$

teniéndose para la variancia de p:

$$V(p) = \frac{N-n}{N} \cdot \frac{S^2}{n} = \frac{N-n}{N-1} \cdot \frac{PQ}{n}$$

Variancia relativa- Coeficiente de variación

La variancia

$$s^2 = \frac{1}{N-1} \sum_{1}^{N} (a_i - \mu)^2 = \frac{N}{N-1}\, \sigma^2$$

de la población está definida en términos de los desvíos absolutos $a_i - \mu$. Si en lugar de estos, se consideran los desvíos relativos

$$(a_i - \mu) / \mu$$

la expresión

$$\frac{1}{N-1} \sum_{1}^{N} \left(\frac{a_i - \mu}{\mu} \right)^2 \frac{s^2}{\mu^2} = \frac{N}{N-1}\, \frac{\sigma^2}{\mu^2}$$

lo que se denomina la "variancia relativa" de la característica a en la población, y su raíz cuadrada es el "coeficiente de variación":

$$C.V_a = \frac{S}{\mu} = \sqrt{\frac{N}{N-1}} \cdot \frac{\sigma}{\mu}$$

Que no es otra cosa que el desvío standard medido en unidades de la media.

En general, la Variancia de la variable "relativa"

$$\frac{Z - E(z)}{E(z)}$$

(cuyo valor medio es cero), se denominará "Variancia relativa" de Z, y su raíz cuadrada será el "coeficiente de variación" de la misma.

La "Variancia relativa" de la media $\bar{y}$ de muestras simples al azar de una población de N unidades cuya media y variancia son μ y S^2 respectivamente será :

$$\frac{V(\bar{y})}{\left[E(\bar{y})\right]^2} = \left(\frac{1}{n} - \frac{1}{N}\right)\frac{S^2}{\mu^2}$$

y el "coeficiente de variación":

$$C.V = \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)}.C.V_a$$

Puesto que para la estimación del total de la población se tiene:

$$y = N\bar{y} \qquad V(y) = N^2 V(\bar{y})$$

se sigue que:

$$\frac{V(y)}{\left[E(y)\right]^2} = \frac{N^2 V(\bar{y})}{N^2\left[E(\bar{y})\right]^2} = \frac{V(\bar{y})}{\left[E(\bar{y})\right]^2}$$

es decir, las estimaciones del total y de la media tienen la misma Variancia relativa y, por ende, el mismo coeficiente de variación.-

Cuando se trata de porcentajes, hemos visto que

$$\sigma^2 = P\,Q$$

de modo que $$C.V = \sqrt{Q/P}$$

y el coeficiente de variación de la estimación p de P, basado en una muestra de extensión n, será :

$$C.V \sqrt{\frac{N-n}{N-1} \cdot \frac{1}{n} \cdot \frac{Q}{P}}$$

En las fórmulas que dan $V(\bar{y})$ y $V(p)$ aparecen los factores

$$\frac{N-n}{N} \quad y \quad \frac{N-n}{N-1}$$

Ahora, se sabe que, para muestras de extensión n de una población infinita, la variancia de la media $\bar{y}$, está dada por σ^2/n de modo que la única diferencia cuando se trata de poblaciones finitas de N unidades es la introducción de los factores arriba mencionados. Estos factores, que reducen la variancia de la estimación, se denominan "factores de finitud". Con ésta denominación abreviamos la más correcta, pero larga, de "factores de corrección que surgen por tratarse de una población finita".-

Si la tasa de muestreo n/N es pequeña, el factor de finitud es próximo a 1 y el tamaño de la población no tiene mayor efecto sobre la variancia de la estimación . En la práctica el factor de finitud puede ignorarse toda vez que la tasa de muestreo sea inferior al 5%, y en algunos casos aún cuando llegue hasta un 10%, El efecto de la no consideración del factor de finitud será sobreestimar la varriancia de la estimación.-

El "error standard"- Precisión de las estimaciones.

La variancia de una característica a en una población de N individuos, ha sido definida mediante:

$$S^2 = \frac{1}{N-1} \sum_{1}^{N} (a_i - \mu)^2 =$$

$$= \frac{N}{N-1} \sigma^2$$

donde μ es el valor medio de dicha característica. La raiz cuadrada de la anterior expresión, es el "desvío standard".-

Hemos visto, por otra parte, que la variancia de la media $\bar{y}$ de una muestra simple al azar de extensión n de esa población es:

$$V(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 = \frac{N-n}{N-1} \cdot \frac{\sigma^2}{n} \qquad (1)$$

y su raiz cuadrada es el"error standard" de la estimación $\bar{y}$ de μ, que indicaremos con $\sigma_{\bar{y}}$.-

El "error standard" no es otra cosa que el"desvío standard" de la distribución de una "estadística".- En lo sucesivo usaremos la denominación de "error standard" para referirnos a la raiz cuadrada de la variancia de una estimación, reservando lo de "desvío standard" para la correspondiente expresión referida a la población. En efecto, el "desvío standard" es una constante que expresa la variabilidad de los valores de la característica que interesa en la población al momento en que se la estudia, mientras que el "error standard" expresa la variabilidad de una estimación construída a partir de una muestra extraída de la población y depende, como veremos más adelante, del esquema de muestreo, y del particular método de estimación empleado, que determinan la forma de la función que lo relaciona con ciertas propiedades de la población.-

La denominación de "error standard" sugiere la idea de una medida o norma para la evolución de un "error", que en este caso es la discrepancia entre una estimación y el valor estimado-siempre y cuando la primera tenga al segundo como valor medio o "esperanza matemática" (•)- de modo tal que en general, está dado por:

$$E\left[\left(z - E(z)\right)^2\right]$$

limitándonos por ahora al caso de la estimación de la media μ mediante la media por elemento en una muestra simple al azar de extensión n, puede lo anterior justificarse al considerar que, para valores fijos de N, S y n, es prácticamente cierto que el error o la discrepancia entre la estimación y el valor estimado no superará a 3 veces el error standard.-

Por practicamente cierto se entiende que es pequeño el porcentaje de casos en que, la repetida aplicación de la misma operación de muestreo, dará resultados que no satisfacen la desigualdad indicada.-

(•) El valor medio del cuadrado de los desvíos con respecto al valor estimado de una estimación z cuya esperanza matemática es distinta de dicho valor estimado, suele denominarse "error medio cuadrático".- La diferencia entre el valor estimado y la esperanza matemática de la estimación, es el "vicio" o "error sistemático de estimación".-

La anterior afirmación se funda en el teorema de "Tchebycheff" que dice que: "Si una variable aleatoria X tiene una distribución cualquiera de media μ y varianza σ^2 la probabilidad de que se verifique la desigualdad

$$|X - \mu| \geq t\sigma \qquad t > 0$$

es menor o igual que $1/t^2$.

M.t.p.

Ahora, puesto que la media $\bar{y}$ tiene media μ y variancia

$$\sigma^2_{\bar{y}} = \left(\frac{1}{n} - \frac{1}{N}\right) S^2$$

resulta que para cualquier $t > 0$ se tendrá:

$$P_r \left\{ |\bar{y} - \mu| \geq t\, \sigma_{\bar{y}} \right\} \leq 1/t^2$$

de modo que, tomando p.e. $t = 3$, resultará que la probabilidad de desvíos mayores que 3 veces el error standard es $\leq 1/9$.-
Por otra parte, se demuestra en el cálculo de las probabilidades que la distribución de los medios de muestras de extensión n extraídas de una población infinita cualquiera, tiende a la normal cuando n crece indefinidamente.- Ahora, dados los tamaños de las poblaciones y de las muestras que de ella se extraen, con que se trabajó en la práctica y cuando no hay en la población individuos en lo que la característica considerada toma valores inusitados que desvían notablemente del conjunto de los restantes, es admisible la utilización de las frecuencias con que se dan los desvíos de una variable que se distribuye normalmente para aplicarlos a las desviaciones de la media con respecto al valor que ella estima en muestras simples al azar de poblaciones finitas.-

Admitido esto, puede afirmarse que en no menos del 95% de los casos la estimación estará comprendida en el intervalo $\mu \pm 2\sigma_{\bar{y}}$ y, en el menos el 99.7% en el intervalo $\mu \pm 3\sigma_{\bar{y}}$.-

Es así como el "error standard" sirve como patrón de la precisión, en tanto que permite fijar límites, que serán superados con probabilidades conocidas, a las desviaciones de una estimación (en nuestro caso de la media) con respecto al valor estimado.-

La posibilidad de evaluar objetivamente la precisión de una estimación, es característica esencial del muestreo aleatorio, que lo

distingue de cualquier otro método de investigación parcial de una población destinada a ofrecer datos para la construcción de estimaciones.

La precisión de la estimación de la media en el muestreo simple al azar, puede también expresarse en términos del coeficiente de variación , el qué, como hemos visto viene dado por:

$$C.V\,\bar{y} = \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)} . C.Va$$

cuyos múltiplos determinan intervalos con centro en la media μ dentro de los cuales estará el "error relativo" con probabilidades iguales a las dadas más arriba.

La fórmula que da el error standard de la media $\bar{y}$ en el muestreo simple al azar de una cierta población, muestra que él depende únicamente del tamaño n de la muestra, por lo que, en este esquema de muestreo, el aumento de la precisión de la estimación sólo puede lograrse incrementando el número de unidades incluídas en la muestra.

Más adelante veremos cómo, manteniendo el mismo esquema de muestreo, puede construirse una estimación de la media de la población que tiene, en ciertas condiciones, una precisión mucho mayor, para el mismo tamaño de la muestra, que la estimación obtenida mediante la media por elemento en la muestra (V. " estimación por cociente ").

La posibilidad de esta otra estimación requiere el conocimiento previo de los valores, para cada uno de los individuos incluídos en la muestra y para la población total, de otra característica b relacionada de un cierto modo con la a .

Si es

$$\bar{y} = \frac{1}{n} \sum_{1}^{n} y_i$$

la media de una muestra simple al azar de extensión n de una población cuya media es μ, y si indicamos con $\sigma_{\bar{y}}$ el error stan-

dard de la estimación, hemos visto que es posible afirmar que en más del 99% de los casos se tendrá:

$$|\bar{y} - \mu| < 3\sigma_{\bar{y}}$$

o sea $$-3\sigma_{\bar{y}} < \bar{y} - \mu < 3\sigma_{\bar{y}}$$

o bién $$\bar{y} - 3\sigma_{\bar{y}} < \mu < \bar{y} + 3\sigma_{\bar{y}}$$

lo que nos dice que-supuesto conocido el error standard $\sigma_{\bar{y}}$ -con probabilidad superior a .99, el intervalo de amplitud $6\sigma_{\bar{y}}$ centrado en la media de la muestra, comprenderá el verdadero valor μ que se pretende estimar.

El intervalo

$$\bar{y} \pm 3\sigma_{\bar{y}}$$

Es el "intervalo de confianza" de la estimación, para un coeficiente de confianza del 99%.- El intervalo $\bar{y} \pm 2\sigma_{\bar{y}}$, es el "intervalo de confianza" para un "coeficiente de confianza" del 95%.-

Todo lo que antecede vale también para el caso de la estimación del total de la población o de un porcentaje, estando entonces los respectivos intervalos de confianza dados por:

$$N\bar{y} \pm k\sigma_{N\bar{y}} \qquad p \pm k\sigma_{p}$$

donde k será igual a 2 ó a 3, según que se considere un coeficiente de confianza del 95 o del 99 por ciento.-

Puesto que los errores standard $\sigma_{\bar{y}}$, $\sigma_{N\bar{y}}$, σ_{p} , dependen del tamaño n de la muestra, se vé que la precisión de la estimación puede hacerse tan grande, o bién la amplitud del intervalo de confianza tan pequeño como se quiera-para un coeficiente de confianza pre-

determinado-eligiendo n suficientemente grande.-

Hemos supuesto hasta ahora que el error standard de la estimación era conocido, lo que supone que es conocida la variancia de los valores de la caractarística que se estudia en la población.- Ahora, no es este precisamente el caso común en la práctica y mas adelante mostraremos cómo la muestra misma ofrece los medios para estimar el error standard de . modo tal que la evaluación de la precisión de una estimación se obtiene a partir única. y oxclusivamente, de los datos de la muestra.-

M.t.p.

DETERMINACION DEL TAMAÑO DE LA MUESTRA

Una cuestión que de inmediato se plantea cuando se pla
ea una encuesta mediante una muestra es: Cuál es el tamaño que ha de
ener la muestra? Así planteada, la pregunta es incompleta y para que
ueda intentarse una respuesta es necesario agregar: ..."para estimar
l valor requerido de la población con tal o cual precisión y tal o
ual grado de confianza de que la estimación está comprendida dentro
el margen de error admitido.

Los principales pasos en la determinación del tamaño
e la muestra son:

1º) El establecimiento de lo que se espera de la muestra y esto en
érminos del límite de error admisible, o bien en términos de alguna
ecisión o acción que ha de tomarse una vez conocido el resultado de
a muestra.

2º) La determinación de la ecuación que liga el tamaño n de la mues
ra con la precisión deseada.

3º) La estimación del valor de aquellos parámetros de la población
ue figuren como tales en la anterior ecuación.

4º) La apreciación del valor que se obtenga para n desde el punto
e vista de su consistencia con los recursos disponibles para la eje-
ución de la operación de muestreo. Esto supone una estimación del
osto, trabajo, tiempo, materiales, personal y otros recursos necesa-
ios para tener una muestra del tamaño requerido.

Suponiendo que se trata de estimar la media de los va-
ores de una cierta característica a de una población de N unidades,
i es d el margen de error elegido, lo que se requiere es que el tama
o n de la muestra sea tal que ella ofrezca una estimación $\bar{y}$ de μ
al que sea pequeña la probabilidad de que se tenga:

$$|\bar{y} - \mu| \geq d$$

La probabilidad

$$\Pr\left\{|\bar{y} - \mu| \geq d\right\} = \alpha$$

epresenta el riezgo de que la estimación discrepe del valor estimado
n más de d, y el valor que para ella se elija depende del grado de
onfianza que se desea asociar a la predicción de que el error de es-
imación es inferior al límite d prefijado.
n general, se eligen para α los valores 5 % ó 1 %.

Hemos visto que el error standard de la media de una
uestra simple al azar de extensión n es:

$$\sigma_{\bar{y}} = \sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)} \cdot S$$

iendo S el desvío standard de la población. Expresando d en términos
el error standard tendremos, siendo t una cierta constante positiva:

$$d = t\sqrt{\left(\frac{1}{n} - \frac{1}{N}\right)} \cdot S$$

xpresión que resuelta con respecto a n dá:

$$n = \frac{(\frac{tS}{d})^2}{1 + \frac{1}{N}(\frac{tS}{d})^2}$$

Poniendo

$$(\frac{tS}{d})^2 = n_o$$

queda finalmente

$$n = \frac{n_o}{1 + n_o/n}$$

Si el tamaño N de la población es grande, bastará tomar n_o como tamaño de la muestra.

Hemos visto más arriba, que salvo en el caso de que haya en la población elementos en los cuales la característica estudiada tome valores inusitados, puede afirmarse que, con probabilidades inferiores a .05 y .01, respectivamente, los desvíos de la estimación superarán a 2 y 3 veces el error standard, de modo que, si en la fórmula que dá no se toma t = 2 ó 3, el valor que se obtenga para n asegurará que la muestra ofrecerá una estimación tal que el riesgo de un error superior a d será menor que 5 % ó 1 %, respectivamente.

Se sigue de aquí que, fijado el riesgo de un error superior a d en un 5 % o en 1 %, se trata, en última instancia de determinar n de modo tal que el error standard

$$S\sqrt{(\frac{1}{n} - \frac{1}{N})}$$

de la estimación sea igual a 1/2 o a 1/3, respectivamente, de d.

Ejemplo: En una empresa que cuenta con 13.000 obreros se desea estimar mediante una muestra el salario promedio horario con un error inferior a M$n. 0,10 y un riesgo de 1 %. Se sabe que el desvío standard de los salarios horarios es m$n. 1,25. Cuál es el tamaño que debe tomarse?

Siendo d = 0,10 t = 3 S = 1,25

se tiene

$$n_o = (\frac{3 \times 1,25}{,10})^2 = 1.406,25 \doteq 1.407$$

de modo que

$$n = \frac{1.407}{1 + \frac{1.407}{13.000}} \doteq 1.270$$

Si observamos que d/t no es sino el error standard $\sigma_{\bar{y}}$ de la media, puede ponerse

$$n_o = (\frac{tS}{d})^2 = (\frac{S}{\sigma_{\bar{y}}})^2$$

ic que muestra que n no es sino el cuadrado del cociente del desvío standard de la población y el error standard de estimación requerido. De la fórmula anterior se sigue inmediatamente que

$$n_o = \frac{(C.V_a)^2}{(C.V.\bar{y})^2}$$

donde $C.V_a$ y $C.V_{\bar{y}}$ son, respectivamente, los coeficientes de variación de la población y de la estimación, de modo que n puede expresarse:

$$n = \frac{1}{\frac{1}{N} + \frac{(C.V_{\bar{y}})^2}{(C.V_a)^2}}$$

entendiéndose que $C.V_{\bar{y}}$ es el coeficiente de variación deseado para la estimación.

Fijado el límite de error relativo

$$d' = (\bar{y} - \mu)/\mu$$

el valor que en la anterior fórmula se tomará para $C.V_{\bar{y}}$ será igual a $d'/2$ ó $d'/3$, según que sea 5 % ó 1 % el riesgo que se adopte.

Para N grande, la anterior fórmula se reduce a:

$$n = \frac{(C.V_a)^2}{(C.V_{\bar{y}})^2} = n_o$$

Ejemplo: Se desea estimar con un error relativo del 10 % (y un riesgo de 5 %) el ingreso medio mensual en una población de 100.000 familias. En base a estudios previos se sabe que la variancia relativa de los ingresos familiares mensuales es $(C.V_a)^2 = 1.5$. Cuál es el tamaño de la muestra que debe tomarse?

Siendo el riesgo 5 %, el coeficiente de variación $C.V_{\bar{y}}$ deseado es 5 %, de modo que (por ser N tan grande) se tendrá:

$$n = \frac{1.5}{.0025} = 600.-$$

Más arriba hemos visto que el error standard de la estimación p de la proporción P de individuos que en una población de N tienen una cierta característica "C" es:

$$\sigma_p = \sqrt{\frac{N-n}{N-1}} \cdot \sqrt{\frac{PQ}{n}}$$

de aquí se obtiene inmediatamente que para un error dado d y un cierto $t > o$, el tamaño n de la muestra requerido es:

$$n = \frac{\frac{t^2 PQ}{d^2}}{1 + \frac{1}{N}\left(\frac{t^2 PQ}{d^2} - 1\right)}$$

o bien poniendo

$$n_o = \frac{t^2 pq}{d^2}$$

$$n = \frac{n_o}{1 + \frac{n_o - 1}{N}} \doteq \frac{n_o}{1 + \frac{n_o}{N}}$$

Como en el caso anterior, se tomará t = 2 ó 3 según que se adopte un riesgo de 5 % ó de 1 %.

En la fórmula que dá n figura el desvío standard, el coeficiente de variación o la proporción P de la población, que son parámetros ciertamente desconocidos al momento en que debe calcularse el tamaño de la muestra a extraerse. La posibilidad de la utilización de la fórmula requiere pues que de algún modo se haga una estimación de uno u otro de dichos parámetros. Como puede llegarse a tener tal estimación depende de las circunstancias particulares de cada caso; así, puede quizá disponerse de información obtenida en estudios previos acerca de la misma característica o de otra semejante o asociada, o bien, puede ser necesario recurrir a una investigación piloto para lograr la información requerida para esa estimación. Cualquiera que sea el procedimiento debe tenerse presente que una estimación de dichos parámetros muy alejada de su verdadero valor puede significar que se tenga, ya sea un valor de n demasiado reducido para alcanzar la precisión deseada, ya un valor demasiado grande, lo que significa un exceso innecesario de costo. Por cierto que una vez tenida la muestra, puede objetivamente juzgarse el grado de precisión alcanzado realmente.

En el caso de tratarse de la estimación de un porcentaje, puede tomarse 0,50 como valor de P en la fórmula, en cuyo caso se tiene el valor máximo de n con el que ciertamente se satisfará la condición de precisión impuesta. Este procedimiento es aplicable toda vez que no sea particularmente costoso comparado con cualquier otro procedimiento que permita lograr otra estimación de P más próxima a su valor real en la población.-

EL ERROR STANDARD. SU ESTIMACION A PARTIR DE LA MUESTRA

En todas las anteiores consideraciones acerca del error standard de una estimación, se ha supuesto conocido el valor del desvío standard en la población, lo que por cierto no ocurre generalmente en la práctica. Resulta así que, para evaluar dicho error standard es necesario recurrir a una estimación del parámetro de la población del cual depende, la que, como veremos, puede obtenerse a partir de los datos mismos ofrecidos por la muestra.

Hemos visto más arriba que, siendo

$$s^2 = \frac{1}{n-1} \sum_{1}^{n} (y_i - \bar{y})^2$$

la variancia de los valores en una muestra simple al azar de una cierta población el valor medio de la misma para todas las muestras posibles, es decir, la "esperanza matemática" de s^2, verifica

$$E(s^2) = S^2$$

lo que significa que (además de "consistente") es s^2 una estimación "no viciada" de la variancia de la población. Se sigue de aquí que

$$v(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) s^2$$

será una estimación "no viciada" de la variancia $V(\bar{y})$ de las medias de las muestras. Otro tanto ocurre con las estimaciones

$$v(y) = N^2 \left(\frac{1}{n} - \frac{1}{N}\right) s^2 \qquad v(p) = \frac{N-n}{N-1} \cdot \frac{pq}{n}$$

de las variancias del total y y de la fracción p.

Importa observar que el desvío standard s de la muestra no es una estimación "no viciada" de S, de modo que el error standard estimado estará afectado de un "error sistemático". Este error es, en general, pequeño y tiende a cero cuando el tamaño de la muestra crece, y en la mayoría de las aplicaciones no tiene importancia alguna.

La variancia relativa S^2/μ^2 se estima mediante,

$$(c.v.)^2 = s^2/\bar{x}^2$$

Esta estimación, si bien es "consistente", es "viciada" pero como en el caso de s, esta circunstancia no tiene importancia prác

tica si el tamaño de la muestra es razonablemente grande.

Tanto s^2, como s y c.v., calculados con los valores obtenidos en la muestra, son variables de muestra a muestra, y su coeficiente de variación en todas las muestras posibles medirá la precisión con que estiman el correspondiente parámetro de la población.

A continuación se dan, para cada una de las 20 muestras posibles de la simple población hipotética de la pág. 11, los valores de s^2, s y c.v.

Muestra Nº	Media: $\bar{y}$	s^2	s	c.v.
1	3.666	4.333	2.129	.581
2	4.333	2.333	1.528	.353
3	4.333	2.333	1.528	.353
4	5.666	6.333	2.517	.444
5	3.	1.	1.	.333
6	3.	1.	1.	.333
7	4.333	10.333	3.217	.742
8	3.666	.333	.577	.157
9	5.	7.	2.646	.529
10	5.	7.	2.646	.529
11	4.	4.	2.	.500
12	4.	4.	2.	.500
13	5.333	9.333	3.055	.573
14	4.666	1.333	1.137	.244
15	6.	4.	2.	.333
16	6.	4.	2.	.333
17	3.333	1.333	1.137	.341
18	4.666	9.333	3.055	.655
19	4.666	9.333	3.055	.655
20	5.333	1.333	1.137	.213
Total		93.996	39.364	8.701
Promedio		4,7	1.968	.435
Población		4.7	2.168	.482

En las dos últimas líneas del cuadro se comparan los valores medios sobre todas las muestras de s^2, s y c.v. con los valores correspondientes de los parámetros S^2, S y C.V. de la población que aquellos estiman, lo que pone de manifiesto el "vicio" de que están afectados los promedios de s y c.v.

Para obtener la expresión de la variancia relativa de s^2 en muestras simples al azar de extensión n, se parte de su definición:

$$C.V._{s^2}^2 = \frac{E\,(s^2 - Es^2)^2}{(E\,s^2)^2} = \frac{Es^4 - (Es^2)^2}{(E\,s^2)^2}$$

y todo se reduce a calcular $E\,s^4$.

M.t.n.

Si es μ la media de la población, y se pone

$$z_i = y_i - \mu$$

$$\bar{z} = \bar{y} - \mu$$

se tiene:

$$Es^4 = E\left\{\frac{\sum_1^n (z_i - \bar{z})^2}{n-1}\right\}^2 = \frac{1}{(n-1)^2} E\left[\left(\sum_1^n z_i^2\right)^2 - 2n\bar{z}\sum_1^n z_i^2 + n^2 z^4\right] \quad (1)$$

Desarrollando cada uno de los términos que figuran entre corchetes en el último miembro, resulta:

$$\left(\sum_1^n x_i^2\right)^2 = \sum_1^n z_i^4 + \sum_{i\neq j}^n z_i^2 z_j^2$$

$$\bar{z}^2 \sum_1^n z_i^2 = \frac{1}{n^2}\left[\sum_1^n z_i^4 + 2\sum_{i\neq j}^n z_i^3 z_j + \sum_{i\neq j}^n z_i^2 z_j^2 + \sum z_i^2 z_j z_k\right.$$

$$\bar{z}^4 = \frac{1}{n^4}\left[\sum_1^n z_i^4 + 4\sum_{i\neq j}^n z_i^3 x_j + 3\sum_{i\neq j}^n z_i^2 z_j^2 + 6\sum_{i\neq j\neq k}^n z_i^2 z_j z_k + \sum_{i\neq j\neq k\neq m} z_i z_j z_k z_m\right.$$

requiriendo la obtención del resultado buscado la evaluación de las siguientes esperanzas matemáticas

$$Ez_i^4 \qquad Ez_i^2 z_j^2 \qquad Ez_i^3 z_j \qquad Ez_i^2 z_j z_k \qquad Ez_i z_j z_k z_m$$

Estas esperanzas matemáticas deben calcularse teniendo presente que, tratándose de un esquema de muestreo "sin reposición", los desvíos $z_1, z_2, \ldots z_n$ no son independientes. La obtención de los valores de esas esperanzas no es difícil, pero si larga y tediosa, de modo que nos limitaremos a dar el resultado final:

M.t.p.

$$V_{s^2}^2 = \frac{(N-1)^2}{N^2(n-1)^2}\left\{\frac{(n-1)^2}{n} - \frac{n-1}{n(N-1)}\left[(n-2)(n-3)-(n-1)\right] -\right.$$

$$\left. - \frac{4(n-1)(n-2)(n-3)}{n(N-1)(N-2)} - \frac{6(n-1)(n-2)(n-3)}{n(N-1)(N-2)(N-3)}\right\}\beta +$$

$$+ \frac{(N-1)^2}{N^2(n-1)^2}\left\{\frac{(n-1)N}{n(N-1)}\left[(n-1)^2+2\right] + \frac{2(n-1)(n-2)(n-3)N}{n(N-1)(N-2)}\right.$$

$$\left. + \frac{3(n-1)(n-2)(n-3)N}{n(N-1)(N-2)(N-3)} - \frac{N^2(n-1)^2}{(N-1)^2}\right\}$$

donde

$$\beta = \mu_4/\sigma^4 \qquad y \qquad \mu_4 = \frac{1}{N}\sum_1^N (a_i-\mu)^4$$

Si se supone que la población sometida al muestreo es infinita o, lo que es equivalente, que el esquema de muestreo es "con reposición", el problema se simplifica pues entonces $z_1, z_2, \ldots\ldots z_n$, son independientes y se tiene:

$$Ez_i^2 z_j^2 = Ez_i^2 Ez_j^2 \quad (i \neq j)$$

$$Ez_i^3 z_j = Ez_i^3 Ez_j$$

$$Ez_i^2 z_j z_k = Ez_i^2 Ez_j Ez_k$$

$$Ez_i z_j z_k z_m = Ez_i Ez_j Ez_k Ez_m$$

y son todas nulas, con excepción de la primera, dado que $Ez_h = 0$

Se tiene entonces en este caso:

$$E\left(\sum_1^n z_i^2\right)^2 = n\mu_4 + n(n-1)\sigma^4$$

$$E\left(\bar{z}^2 \sum_1^n z_i^2\right) = \frac{1}{n^2}\left[n\mu_4 + n(n-1)\sigma^4\right] = \frac{\mu_4}{n} + \frac{n-1}{n}\sigma^4$$

$$E\bar{z}^2 = \frac{1}{n^4}\left[n\mu_4 + 3n(n-1)\sigma^4\right] = \frac{\mu_4}{n} + \frac{3(n-1)}{n^3}\sigma^4$$

M. t. n.

siendo

$$\mu_4 = Ez_i^4 \qquad \sigma^2 = En_i^2 = Ez_i^2$$

Reemplazando los anteriores valores en la (1), y restando σ^4, luego de simples operaciones algebraicas se obtiene:

$$Es^4 - \sigma^4 = \frac{\mu_4}{n} - \frac{\sigma^4}{n-1} \cdot \frac{n-3}{n}$$

o bien, finalmente :

$$C.V._{s^2}^2 = \frac{Es^4 - \sigma^4}{4} = \frac{1}{n}\left(\beta - \frac{n-3}{n-1}\right) \qquad (2)$$

si n es relativamente grande, puede tomarse (n-3)/(n-1) como siendo igual a 1, y se tiene la expresión aproximada:

$$C.V._{s^2}^2 \pm \frac{1}{n}(\beta - 1) \qquad (3)$$

Para determinar la esperanza matemática y la variancia de s, pongamos:

$$s^2 = \sigma^2 + z$$

donde z es una variable aleatoria que verifica:

$$E(z) = 0 \qquad E(z^2) = V(s^2)$$

Podemos escribir

$$s = (\sigma^2 + z)^{1/2} = \sigma\left(1 + \frac{z}{\sigma^2}\right)^2$$

de donde tomando los primeros términos del desarrollo en serie del último miembro, se tiene la siguiente aproximación :

$$s = \sigma\left[1 + \frac{1}{2} \cdot \frac{z}{\sigma^2} + \frac{(1/2)(-1/2)}{2!} \cdot \left(\frac{z}{\sigma^2}\right)^2\right]$$

cuya esperanza matemática es:

$$E(s) = \sigma\left[1 - \frac{1}{8} \cdot \frac{V(s^2)}{\sigma^4}\right] \qquad (4)$$

M.t.p.

Este resultado pone de manifiesto que el valor medio de s subestima al verdadero valor σ, y también que este error será insignificante si el tamaño de la muestra es suficientemente grande.

Partiendo de la (4) es fácil hallar la expresión de la variancia de la estimación s. En efecto, puesto que por definición es :

$$V(s) = E\left[s-E(s)\right]^2 = E(s^2) - \left[E(s)\right]^2$$

se tiene:

$$V(s) = \sigma^2 - \sigma^2\left[1-\frac{1}{8}\cdot\frac{V(s^2)}{\sigma^2}\right]^2 = \sigma^2\left[1-1+\frac{1}{4}\cdot\frac{V(s^2)}{\sigma^4} - \frac{1}{64}\left(\frac{V(s^2)}{\sigma^4}\right)^2\right]$$

de donde resulta como expresión aproximada de V(s) :

$$V(s) \doteq \frac{V(s^2)}{4\sigma^2} \qquad (5)$$

De la anterior se obtiene como expresión aproximada del coeficiente de variación :

$$C.V_s \doteq \sqrt{\frac{\beta - 1}{4\ n}} \qquad (6)$$

o bien :

$$C.V_s \doteq \sqrt{\frac{\beta - \frac{n-3}{n-1}}{4n}} \qquad (7)$$

que da una mejor aproximación aplicable cuando n es pequeño o cuando β es próxima a 1.

Los resultados precedentes muestran que la precisión de la estimación s basada en una muestra de extensión n depende del valor del parámetro β de la población, de modo tal que si se desea tener una estimación de precisión determinada, a saber, con un coeficiente de variación del 10% p.e., se requiere algún conocimiento previo acerca del valor del momento reducido de 4º orden de la poblacón, con lo cual podrá entonces determinarse el tamaño n de la muestra que debe tomarse para alcanzar esa precisión.

Si se conoce β, a partir de la (6) puede calcularse n que vendrá dado por :

$$n \doteq \frac{\beta - 1}{4\ C.V_s^2} \qquad (8)$$

de modo que si de desea un coeficiente de variación del 10 % para que

M.t.p.

s pueda considerarse como siendo una estimación fidedigna de σ, deberá tomarse $C.V_s^2 = .01$, de modo que será :

$$n \geq \frac{\beta - 1}{.04}$$

Es sabido que para una población normal el momento reducido de 4º orden es igual a 3, de modo tal que una muestra de 50 unidades ofrecerá una estimación cuyo coeficiente de variación es igual a 10 %.

Ahora bien, desgraciadamente, en la práctica es muy poco frecuente encontrar poblaciones en las que los valores de las características que interesan se distribuyen normalmente, en particular en las poblaciones que se enfrentan en los estudios demográficos, económicos o sociales, en las que, en general, los valores de β superan notablemante al que es característico de la distribución normal.

La distribución de los Departamentos de las Provincias y Territorios Nacionales de la República Argentina según el número de sus habitantes a la fecha del 4º Censo, tiene un $\beta = 20.01$.

Valores más altos tienen las siguientes poblaciones (.)

Explotaciones Agrícolas en los E.E.U.U. - 1940

(según su extensión en acres)

Extensión (acres)			Por ciento de explotaciones
hasta		10	8,3
10	-	29	16.6
30	-	49	12,6
50	-	69	8,4
70	-	99	12.8
100	-	139	11.3
140	-	179	10,2
180	-	219	4,6
220	-	259	3,4
260	-	379	5.3
380	-	999	4.9
1,000	-	4.999	1.4
5,000	y	más	.2

$\mu = 192$
$\sigma = 869$
$C.V. = 4,53$
$\beta = 603$

(.) Mencionadas por Hansen, Hurwitz y Madow en "Sample Survey Methods and Theory", Vol. I, pp. 142 y 143.

M.t.p.

Comercios minoristas independientes en E.E.U.U. - 1939

(según monto de las ventas).

Monto de las ventas (miles)			Por ciento de comercios
Hasta		50	93.36
50	-	99	4.11
100	-	299	2.03
300	-	499	.28
500	-	999	.16
1.000	-	4.999	.05
5.000	y	más	.01

μ = 22.348

σ = 153.067

C.V. = 6.85

β = 5.905

Poblaciones con valores grandes de β no son la excepción, y es evidente que el muestreo simple al azar será un método particularmente inefi ciente teniendo en cuenta el tamaño de la muestra que se requiere para tener estimaciones lo suficientemente precisas para que sean de alguna utilidad. Otros son los métodos de muestreo que en estos casos deben a plicarse.

Cuando lo que interesa en una población es la proporción de individuos que poseen una cierta característica "C", es posible dar algunas reglas para determinar el tamaño de la muestra requerida para estimar la variancia con un grado determinado de fidedignidad.

Un razonamiento análogo al que llevó a determinar la variancia en una población en la que la fracción P de individuos poseían una cierta característica "C", permite demostrar que

$$\mu_4 = \frac{1}{N}\sum_{1}^{N}(y_i - P)^4 = P(1-P)^4 + Q(0-P)^4 = PQ - 3P^2Q^2$$

de donde se sigue, puesto que, como hemos visto

$$\sigma^2 = PQ$$

que

$$\beta = \frac{1}{PQ} - 3$$

M.t.p.

de modo que reemplazando este valor en (7) tenemos :

$$C.V._s \doteq \sqrt{\frac{1}{4n}\left(\frac{1}{PQ} - \frac{4n-6}{n-1}\right)} \qquad (9)$$

como expresión del coeficiente de variación de s calculado a partir de una muestra simple al azar de extensión n, extraída con reposición de la población.

Imponiendo a $C.V._s$ la condición de ser igual a .1, si se toma en (9) n = 60 se tiene:

$$.01 = \frac{1}{240}\left[\frac{1}{PQ} - \frac{240-6}{59}\right] \qquad (10)$$

de donde se sigue facilmente :

$$P^2 - P + .1571 = 0$$

ecuación de 2º grado que dá para P los siguientes valores :

$$P_1 = .20 \qquad P_2 = .80$$

de modo que $C.V._s$ será menor que .1 para cualquier valor de PQ superior a .16 (siendo n 60). En efecto, siendo a una fracción positiva menor que .60, si se tiene :

$$P = .20 + a$$

$$Q = 1-(.20+a) = .80-a$$

resulta $$PQ = .16 + .60a - a^2$$

lo que muestra que $PQ > .16$. Ahora, el 2º miembro de la (10) decrece cuando PQ crece, de modo que si la igualdad (10) se verifica para $PQ=.16$, el 2º miembro será menor que .01 para cualquier $PQ>.16$.

Se concluye de aquí que, $C.V._s$ será menor que 10 % para una muestra de extensión n = 60 toda vez que se tenga

$$.20 \leq P \leq .80$$

Por otra parte, si P = .20, la proporción p en una muestra de extensión n = 60 estará comprendida en el intervalo

$$.20 - 2\sqrt{\frac{.16}{60}} = .10 \quad , \qquad .20 + 2\sqrt{\frac{.16}{60}} = .30$$

y si P = .80, p estará en el intervalo

$$.80 - 2\sqrt{\frac{.16}{60}} = .70 \; , \qquad .80 + 2\sqrt{\frac{.16}{60}} = .90$$

Estos resultados nos permiten afirmar que, si en una muestra de extensión n = 60 la fracción p está comprendida entre .30 y .70, el valor P en la población, con alta probabilidad estará comprendido entre .20 y .80.

De aquí se sigue que una muestra de 60 unidades será suficiente para estimar el desvío standard de una proporción p que está comprendida entre 30 y 70 %, con un coeficiente de variación inferior a 10 %.

Una segunda regla, dada por Hansen, Hurwitz y Madow, y que no demostraremos aquí, dice que: Ya sean o no, p ó q, menores que 30 %, si np y nq son ambos mayores que 35, el coeficiente de variación de s será menor que 10 %.

Cuando las características cuyo estudio interesa son variables cuantitativas tales como: salarios, monto de ventas, extensión de explotaciones agrícolas, ingresos familiares, etc., no es posible dar reglas semejantes a las de más arriba, y todo depende del conocimiento que de algún modo se pueda lograr acerca de la población que se pretende estudiar. Sí puede afirmarse, en general que, si el tamaño de la muestra es superior a 500 unidades, la estimación del error standard será suficientemente precisa, siempre y cuando la población no sea notablemente asimétrica o tenga alguna otra peculiaridad que signifique un particularmente alto valor para β.-

Cuadro Nº

POBLACION HIPOTETICA DE N = 15 INDIVIDUOS

(Característica: a Ingresos mensuales)

Individuo	Ingresos
Nº	$
1	$a_1 = 1.000$
2	$a_2 = 2.000$
3	$a_3 = 1.500$
4	$a_4 = 3.000$
5	$a_5 = 2.500$
6	$a_6 = 5.500$
7	$a_7 = 4.000$
8	$a_8 = 5.000$
9	$a_9 = 8.500$
10	$a_{10} = 2.500$
11	$a_{11} = 7.000$
12	$a_{12} = 8.000$
13	$a_{13} = 2.000$
14	$a_{14} = 3.500$
15	$a_{15} = 1.500$

Total de la población: A = $ 57.000

Ingreso medio por individuo: $\bar{x}$ = $ 3.833,33

Desvío standard: σ_a = $ 2,394,35

Coeficiente de variación: C V_a = .625

DISTRIBUCION DE LOS VALORES MEDIOS DE TODAS LAS MUESTRAS POSIBLES DE EXTENSIONES n = 1, 2, 3, 4 Y 5 QUE PUEDEN OBTENERSE DE LA POBLACION DEL CUADRO Nº

Col. 1 Ingreso medio estimado (miles m$n.)	Número de muestras, de cada uno de los siguientes tamaños, que dan ingresos medios estimados comprendidos en los distintos intervalos de la Col. 1				
	n = 1	n = 2	n = 3	n = 4	n = 5
.75 - 1.25	1	-	-	-	-
1.25 - 1.75	2	5	12	6	5
1.75 - 2.25	2	12	37	60	87
2.25 - 2.75	2	13	52	130	240
2.75 - 3.25	1	11	58	178	439
3.25 - 3.75	1	11	60	243	609
3.75 - 4.25	1	9	72	248	627
4.25 - 4.75	-	9	58	203	493
4.75 - 5.25	1	11	42	150	311
5.25 - 5.75	1	10	29	86	138
5.75 - 6.25	-	5	19	41	45
6.25 - 6.75	-	3	9	15	8
6.75 - 7.25	1	3	5	4	1
7.25 - 7.75	-	1	2	1	-
7.75 - 8.25	1	1	-	-	-
8.25 - 8.75	1	1	-	-	-
Total: C_n^{15}	15	105	455	1.365	3.003
Promedio:	3.833	3.833	3.833	3.833	3.833

DEPARTAMENTOS DE LAS PROVINCIAS Y TERRITORIOS NACIONALES SEGUN SU POBLACION A LA FECHA DEL IV CENSO GENERAL DE LA NACION (1)

Nº de habitantes	Departamentos	Nº de habitantes	Departamentos
- 5.000	92	110.000 - 120,000	2
5.000 - 10.000	90	120.000 - 130.000	5
10.000 - 20.000	106	130.000 - 140.000	-
20.000 - 30.000	54	140.000 - 150.000	4
30.000 - 40.000	33	150.000 - 200.000	-
40.000 - 50.000	23	200.000 - 250.000	3
50.000 - 60.000	16	250.000 - 300.000	2
60.000 - 70.000	12	300.000 - 350.000	1
70.000 - 80.000	3	350.000 - 400.000	1
80.000 - 90.000	6	400.000 - 450.000	-
90.000 - 100.000	6	450.000 - 500.000	-
100.000 - 110.000	1	500.000 - 550.000	1
			461

Fuente: IV Censo General de la Nación (1947) - Resultados Generales del Censo de Población - Informe C1 - Dirección General del Servicio Estadístico Nacional.

$\mu = 27.267$

$\sigma = 40.670$

C.V. 1.49

$\beta = 20.01$

Capítulo II

MUESTRAS DE UNA POBLACION DIVIDIDA EN ESTRATOS

1) Descripción.- El esquema de muestreo de una población dividida en estratos ("muestreo estratificado") supone que la población de N unidades ha sido clasificada en un cierto número k de grupos cada uno de los cuales consta de N_1 , N_2 , N_3 ..... N_k unidades y tales que

$$\sum_{h=1}^{K} N_h = N.$$

Los grupos o subpoblaciones, que en su conjunto forman la población total, se denominan "estratos". La máxima ventaja de este esquema de muestreo se logra cuando las N_h son conocidas.-

Esquemáticamente, la situación puede representarse del modo siguiente:

Estrato Nº	1	2	..	h	...	k
	a_{11}	a_{21}	...	a_{h1}	....	a_{k1}
	a_{12}	a_{22}		a_{h2}		a_{k2}
	.	.	.	.		.
	.	.		.		.
	.	.		.		.
	.	.		.		.
	a_{1N_1}	a_{2N_2}	...	a_{hN_h}		a_{kN_k}
Población en el estrato :	N_1	N_2	..	N_h	...	N_k
Total :	A_1	A_2	..	A_h	...	A_k
Media :	μ_1	μ_2	..	μ_h	...	μ_K
Variancia:	s_1^2	s_2^2	..	s_h^2	..	s_k^2

$$\sum_{h=1}^{k} N_h = N$$

a_{hj} es el valor de una cierta característica para el j-ésimo individuo del h-ésimo estrato. Es evidente que :

$$A_h = \sum_{j=1}^{N_h} a_{hj}$$

$$\mu_h = \frac{1}{N_h} \sum_{1}^{N_h} a_{hj} = A_h / N_h.$$

$$s_h^2 = \frac{1}{N_h - 1} \sum_{1}^{N_h} (a_{hj} - \mu_h)^2$$

Si indicamos con A el valor total de la característica que se considera en la población, y con μ la media por unidad en la población total, tendremos :

$$A = \sum_{h=1}^{k} A_h = \sum_{1}^{k} \sum_{1}^{N_h} a_{hj}$$

$$\mu = \frac{1}{N} \sum_{1}^{k} N_h \mu_h = A/N.$$

El objeto del muestreo es estimar uno u otro de los anteriores valores para la población.

En el "muestreo estratificado simple al azar".- que es el único del que nos ocuparemos- la muestra, de extensión total n se forma extrayendo una muestra simple al azar de cada uno de los estratos, teniéndose así k muestras - independientes - de extensión $n_1, n_2 \ldots$

Indicando con y_{hj} el valor de la característica estudiada para la unidad incluída en la muestra en la j-ésima extracción del h-ésimo estrato, el esquema, análogo al de más arriba, para la muestra será :

Muestra del estrato.	1	2	..	h	..	k
	y_{11}	y_{21}	....	y_{h1}	...	y_{k1}
	y_{12}	y_{22}		y_{h2}		y_{k2}
	:	:		:		:
	$Y1n_1$	$Y2n_2$		Yhn_h		Ykn_k
Extensión de la muestra	n_1	n_2	..	n_h	...	n_k
Total :	y_1	y_2	..	y_h	...	y_k
Media :	$\bar{y}_1$	$\bar{y}_2$	..	$\bar{y}_h$	...	$\bar{y}_k$
Variancia:	s_1^2	s_2^2	..	s_h^2	..	s_k^2

$$\sum_{1}^{k} n_h = n$$

Por definición es :

$$y_h = \sum_{j=1}^{n_h} y_{hj}$$

$$\bar{y}_h = \frac{1}{n_h} \sum_{1}^{n_h} y_{hj} = y_h / n_h$$

$$s_h^2 = \frac{1}{n_h - 1} \sum_{j=1}^{n_h} (y_{hj} - \bar{y}_h)^2$$

La media de la muestra total de extensión n es

$$\bar{y} = \frac{1}{n} \sum_{h=1}^{k} n_h \bar{y}_h$$

En el muestreo estratificado simple, el número total de muestras distintas que es posible obtener, es igual al producto de tantos factores como estratos, siendo cada uno de dichos factores igual al número de muestras simples al azar que es posible extraer de cada estrato, es decir :

$$\binom{N_1}{n_1} \cdot \binom{N2}{n_2} \cdot \ldots \binom{N_k}{n_k} \qquad (2)$$

El j-ésimo individuo del i-ésimo estrato figurará en :

$$\binom{N_1}{n_1} \cdot \binom{N_2}{n_2} \ldots \binom{N_{i-1}}{n_{i-1}} \binom{N_i-1}{n_i-1} \binom{N_{i+1}}{n_{i+1}} \ldots \binom{N_k}{n_k} \qquad (b)$$

muestras, y por lo tanto, la probabilidad de que él sea incluido en una muestra de extensión $n = n_1 + n_2 + \ldots + n_k$, es el cociente de (b) sobre (a), que dá :

$$n_i / N_i$$

lo que nos dice que, en el muestreo estratificado no todos los individuos de la población tienen la misma probabilidad de ser incluidos en la muestra. Esa probabilidad depende del estrato en que el individuo está ubicado, y se tiene entonces un esquema de muestreo aleatorio con probabilidades variables. Lo antedicho no vale en el caso en que, usándose una tasa de muestreo constante en todos los estratos, se tiene :

$$\frac{n_1}{N_1} = \frac{n_2}{N_2} = \ldots = \frac{n_k}{N_k} = \text{const.}$$

de modo que

$$\sum n_i = \text{const.} \times \sum N_i$$

o sea :

$$n = \text{const.} \times N$$

Resulta así que la tasa de muestreo constante es n/N y todos los individuos tienen la misma probabilidad de inclusión en la muestra.

La estratificación de la población previa a la extracción de la muestra es una técnica de uso corriente, siendo las siguientes algunas de las razones para que así sea :

1º) Si se desea tener información de precisión conocida acerca de ciertas subdivisiones de la población, es aconsejable tratar cada subdivisión como constituyendo ella misma una "población".

2º) Razones de conveniencia de organización o administración pueden hacer necesario o aconsejable el uso de la estratificación, p.e. en el caso en qué el organismo que ejecuta la investigación está estructurado en forma tal que resulta más eficaz y económico distribuir la ejecución y supervisión de la operación en diversos centros que cubren cada una un sector parcial de la población total.

3º) Distintas partes de la población total pueden presentar problemas de muestreo notablemente diferentes, de modo que resulta aconsejable considerar cada una de ellas como constituyendo un estrato p.e. zonas urbanas y rurales.

4º) Finalmente, la estratificación puede llevar a lograr una mayor precisión en la estimación de los parámetros de la población. Mas adelante veremos cómo, estratos internamente homogéneos y heterogéneos entre sí, a través de una adjudicación apropiada de la muestra total entre los diferentes estratos, lleva a una apreciable reducción en la variancia de la estimación comparada con la que corresponde al muestreo simple al azar.

Ejemplo:

Supongamos que los N = 15 individuos que constituyen la población hipotética del cuadro Nº I, se han agrupado en 3 estratos del modo siguiente :

							Total	Media
Estrato Nº 1 :	a_{1h} - 1	1,5	1,5				$A_1 = 4$	$\mu_1 = 1,33$
Estrato Nº 2 :	a_{2h} - 2	3	2,5	2,5	2	3,5	$A_2 = 15,5$	$\mu_2 = 2,58$
Estrato Nº 3 :	a_{3h} - 5,5	4	5	8,5	7	8	$A_3 = 38$	$\mu_3 = 6,33$
							$A = 57,5$	$\mu = 3,83$

El número de muestras diferentes de extensión n = 5 que pueden extraerse cuando la adjudicación es:

$$n_1 = 1 \qquad n_2 = 2 \qquad n_3 = 3$$

está dado por

$$\binom{3}{1} \times \binom{6}{2} \times \binom{6}{2} = 675$$

No es difícil construir la lista de las 675 muestras y calcular el valor medio para cada una de ellas, el que, por tratarse de una adjudicación que fija una tasa de muestreo constante igual a 1/3 en todos los estratos, ofrece, como veremos, una estimación "no - viciada" de la media de la población.

En el cuadro siguiente se dá la distribución de las 675 medias obtenidas, como asi también los de las 3003 que son posibles tomando muestras simples al azar de la misma extensión.

Ingreso medio estimado	Frecuencia de las medias de muestras de extensión Nº 5	
Miles de $	Muestreo simple al azar	Muestreo estratificado Adjud. $n_1 = 1, n_2 = 2, n_3 = 2$
1.25 -1.75	5	-
1.75 -2.25	87	-
2.25 -2.75	240	-

2.75 - 3.25	439	74
3.25 - 3.75	609	205
3.75 - 4.25	627	275
4.25 - 4.75	493	114
4.75 - 5.25	311	7
5.25 - 5.75	138	-
5.75 - 6.25	45	-
6.25 - 6.75	8	-
6.75 - 7.25	1	-
Total	3.003	.675

Se pone aquí de manifiesto cuan grande es la ganancia en precisión de la estimación comparada con la ofrecida por el muestreo simple al azar de la misma población, lograda mediante la particular estratificación y adjudicación utilizadas.

Mas adelante veremos cuales son los puntos que deben tenerse presentes para la formación de los estratos y para la adjudicación de la muestra para que sea posible lograr una real ventaja en precisión de la estimación con respecto a la que podría obtenerse a partir de una muestra simple al azar de la misma población.-

Estimaciones y sus propiedades

En el muestreo estratificado, la estimación del valor medio por unidad en la población, está dada por:

$$\bar{y}_e = \frac{1}{N} \sum_{h=1}^{k} N_h \bar{y}_h = \sum_{h=1}^{k} p_h \bar{y}_h \qquad (1)$$

indicando N_h/N con p_h

es decir, por la media aritmética ponderada de las medias de la muestra extraídas de cada estrato.

Recordando que la media de la muestra total viene dada por:

$$\bar{y} = \frac{1}{n} \sum_{1}^{k} n_h \bar{y}_h$$

se sigue que $\bar{y}_e$ será diferente de $\bar{y}$ salvo en el caso de que

$$\frac{N_h}{N} = \frac{n_h}{n}$$

o bien

$$\frac{n_h}{N_h} = \frac{n}{N} = \text{const.}$$

Cuando la tasa de muestreo es la misma en todo estrato e igual a la tasa de muestreo de la población, se dice que se ha he-

cho una " adjudicación proporcional " de las n_h , con lo que se obtiene una " muestra autoponderada ".

Pasando ahora a las propiedades de la estimación (1), es evidente en primer término que ella es "consistente" pues si $n = N$, siendo entonces necesariamente $\bar{y}_h = \mu_h$ resulta $\bar{y}_e = \mu$

La estimación es también " no viciada ", puesto que, siendo la muestra que se extrae de cada estrato una muestra simple al azar , es $E(\bar{y}_h) = \mu_h$ y, por ende:

$$E(\bar{y}_e) = \frac{1}{N} \sum_{1}^{k} N_h E(\bar{y}_h) = \mu$$

Puede demostrarse también que la estimación (1) es lo que se denomina " la mejor estimación lineal no viciada ", significando esto que, de todas las estimaciones de la forma:

$$t = \sum_{h=1}^{k} \sum_{j=1} a_{hj} \; y_{hj}$$

tales que $E(t) = \mu$ que es la que tiene la variancia mínima.

Variancia de la estimación $\bar{y}_e$

Un conocido teorema del C. de las Probabilidades nos enseña que si

$$z = a_1 x_1 + a_2 x_2 + \ldots\ldots\ldots + a_k x_k$$

donde las a_i son constantes cualquiera y las x_i variables aleatorias independientes, se tiene:

$$V(z) = a_1^2 \quad V(x_1) + a_2^2 \; V(x_2) + \ldots + a_k^2 \; V(x_k)$$

Aplicando este teorema a la variable aleatoria

$$\bar{y}_e = \sum_1^k p_h \bar{y}_h$$

se tiene:

$$V(\bar{y}_e) = \sum_1^k p_h^2 \; V(\bar{y}_h)$$

Ahora bien, como de cada estrato se extrae una muestra simple al azar, es

$$V(\bar{y}_h) = \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S_h^2$$

donde S_h^2 es la variancia en el h-ésimo estrato en la población. De lo anterior resulta pues que:

$$V(\bar{y}_e) = \sum_{h=1}^k \left(\frac{1}{n_h} - \frac{1}{N_h} \right) p_h^2 \quad S_h^2 \qquad (2)$$

o bien, recordando que $p_h = N_h / N$:

$$V(\bar{y}_e) = \frac{1}{N^2} \sum_{h=1}^{k} N_h (N_h - n_h) \frac{S_h^2}{n_h} \qquad (2')$$

a) Si n_h es pequeño con respecto a N_h para todo $h = 1, 2 \ldots k$, o sea si la tasa de muestreo n_h / N_h en cada estrato es pequeña, la (2) puede escribirse :

$$V(\bar{y}_e) = \sum_{h=1}^{k} p_h^2 \frac{S_h^2}{n_h} \qquad (3)$$

b) En el caso de adjudicación proporcional de la muestra en los diferentes estratos, es decir, cuando

$$n_h = n \frac{N_h}{N} = np_h$$

se ve fácilmente que la variancia de $\bar{y}_e$ puede expresarse del modo siguiente:

$$V(\bar{y}_e) = (\frac{1}{n} - \frac{1}{N}) \sum_{h=1}^{k} p_h S_h^2 \qquad (4)$$

c) Si la adjudicación es proporcional y la variancia en todos los

estratos es la misma ($S_1^2 = S_2^2 = \ldots\ldots = S_k^2 = S_w^2$),

se tiene

$$V(\bar{y}_e) = \left(\frac{1}{n} - \frac{1}{N}\right) S_w^2 \qquad (5)$$

representando S_w^2 la variancia dentro de cada estrato, común a todos ellos.

De todo lo que antecede es inmediato que

$$y_e = N\bar{y}_e = \sum_{h=1}^{k} N_h \bar{y}_h$$

es una estimación consistente y " no viciada " del total A de la población y que su variancia es:

$$V(y_e) = \sum_{h=1}^{k} N_h (N_h - n_h) \frac{S_h^2}{n_h} \qquad (6)$$

Adjudicación óptima de la muestra en los diferentes estratos.

Dado un cierto tamaño total $n = \sum n_k$ de la muestra a extraerse de la población, hemos visto que la variancia de la estimación de la media es:

$$V(\bar{y}_e) = \sum_{h=1}^{k} \left(\frac{1}{n_h} - \frac{1}{N_h} \right) p_h^2 S_h^2$$

siendo ella entonces una función de las n_h, es decir, del número de unidades que del total n se extraen al azar de cada estrato. Una pregunta que de inmediato surge es: Cuál es la adjudicación óptima de la muestra en los diferentes estratos ? o sea cómo deben tomarse los valores de n_1, n_2 ... n_k para que, cumpliéndose $\sum n_h = n$, $V(\bar{y}_e)$ sea un mínimo ?.

Para resolver este problema construyamos la función

$$F(n_1, n_2, \ldots n_k) = V(\bar{y}_e) + an$$

de las k variables $n_1, n_2 \ldots n_k$, en lo que a es una constante que por el momento queda indeterminada y $n = \sum n_h$ está predeterminada, de modo que an es una constante para cualquier sistema de valores positivos de $n_1, n_2 \ldots n_h$ cuya suma sea n.

La F es:

$$F = \sum_{h=1}^{k} \left(\frac{1}{n_h} - \frac{1}{N_h} \right) p_h^2 S_h^2 + a \sum_{h=1}^{k} n_h$$

$$= \sum_{h=1}^{k} \left[p_h^2 \, S_h^2 \Big/ n_h + a n_h - p_h^2 \, S_h^2 \Big/ N_h \right]$$

$$= \sum_{h=1}^{k} \left[p_h^2 \, S_h^2 \Big/ n_h + a n_h - 2 p_h \, S_h \sqrt{a} + 2 p_h \, S_h \sqrt{a} - p_h^2 \, S_h^2 \Big/ N_h \right]$$

$$= \sum_{h=1}^{k} \left(p_h \, S_h \Big/ \sqrt{n_h} - \sqrt{a n_h} \right)^2 + 2\sqrt{a} \sum_{h=1}^{k} p_h \, S_h - \sum_{h=1}^{k} p_h^2 \, S_h^2 \Big/ N_h$$

lo que muestra que F (y por tanto $V(\bar{y}_e)$ será un mínimo cuando las n_h sean tales que:

$$\frac{p_h \, S_h}{\sqrt{n_h}} = \sqrt{a n_h} \quad \text{para } h = 1, 2 \ldots\ldots k$$

es decir:

$$n_h = \frac{p_h \, S_h}{\sqrt{a}} = \frac{N_h \, S_h}{N \sqrt{a}}$$

lo que nos dice que la adjudicación óptima (o de Tschuprow-Neyman) es aquella que a cada estrato asigna una muestra de extensión proporcio-

nal a $N_h S_h$. Para determinar el valor de la constante a basta observar que

$$\sum_{h=1}^{k} n_h = n = \frac{1}{N\sqrt{a}} \sum_{h=1}^{k} N_h S_h$$

$$\therefore \quad \frac{1}{\sqrt{a}} = \frac{n N}{\sum N_h S_h}$$

de modo que resulta

$$n_h = n \frac{N_h S_h}{\sum N_h S_h} \qquad (7)$$

de modo que la adjudicación óptima se logra usando una tasa de muestreo variable de estrato a estrato, proporcional al desvío standard S_h .

Puede ocurrir que la aplicación de la fórmula (7) lleve a adjudicar a uno o más estratos una muestra de extensión superior a los respectivos N_h . En tal caso se incluye en la muestra el 100% de la población de dichos estratos y se completa la muestra de n adjudicando la diferencia a los estratos restantes de acuerdo con dicha fórmula. Por cierto que en este caso sólo contribuyen a la variancia de la estimación aquellos estratos en los que la tasa de muestreo es inferior a 100 %.

Sustituyendo en $V(\bar{y}_e)$ el valor de n_h por el hallado en (7) se tiene:

$$V_{op}(\bar{y}_e) = \frac{1}{n}\left(\sum_{k=1}^{k} p_h S_n\right)^2 - \frac{1}{N}\sum_{h=1}^{k} p_h S_h^2$$

como valor de la variancia mínima obtenida por la adjudicación óptima de la muestra.

Adjudicación óptima en el caso de costos variables.

En el párrafo anterior la determinación de la adjudicación se ha hecho sin tomar en cuenta posibles diferencias en el costo del muestreo en los distintos estratos. Introduciendo consideraciones de costos diferenciales, puede plantearse el problema de asignar las n_h de modo de lograr una estimación de precisión prefijada a un costo total mínimo o, vice-versa, una precisión máxima para un costo total prefijado.

En general, puede considerarse que el costo total C de una muestra de una población dividida en estratos está representado por una " función de costo " de la forma siguiente:

$$C = a + \sum_{1}^{k} n_h c_h$$

donde a representa la parte " fija " del costo, independiente del tamaño total de la muestra y de n_h ; c_h representa el costo unitario por observación en el h-ésimo estrato.

En caso de que el costo unitario por observación sea el mismo en todos los estratos, se tendrá:

$$C = a + nc$$

En lo que sigue dejaremos de lado el elemento fijo del costo, tomando como expresión de la función de costo:

$$C = \sum_{h=1}^{k} n_h c_h$$

Como lo hicimos más arriba escribimos la función

$$F = V(\bar{y}_e) + b \sum_{1}^{k} n_h c_h$$

donde b es una constante por el momento indeterminada.

Siguiendo un camino análogo al del párrafo anterior llegaremos a tener para F :

$$F = \sum_{h=1}^{k} \left(\frac{p_h S_h}{\sqrt{n_h}} - \sqrt{b c_h n_h} \right)^2 + \text{término indep. de } n_h$$

lo que muestra que $V(\bar{y}_e)$ será un mínimo para un valor fijo de C, o bien C será un mínimo para un valor fijado de $V(\bar{y}_e)$ cuando las n_h sean tales que:

$$\frac{p_h S_h}{\sqrt{n_h}} = \sqrt{b c_h n_h} \qquad \text{para } h = 1, 2 \ldots\ldots k$$

o sea, cuando

$$n_h = \frac{p_h S_h}{\sqrt{b c_h}} = \frac{N_h S_h}{N \sqrt{b c_h}} \qquad h = 1, 2 \ldots . k \qquad (8)$$

lo que indica que el tamaño de la muestra a tomarse en un estrato aumentará proporcionalmente a:

a) el tamaño N_h del estrato

b) la variabilidad de la población en el mismo

c) el menor costo unitario por observación en dicho estrato.

Si es C_o el monto del presupuesto destinado para una encuesta, para determinar los valores de n_h que aseguran una precisión máxima, para un n también prefijado, reemplazando en :

$$C_o = \sum_{h=1}^{k} n_h c_h$$

n_h por su valor hallado en (8), se tiene:

$$C_o = \frac{1}{\sqrt{b}} \sum_{h=1}^{k} \sqrt{c_h}\, p_h S_h$$

de donde resulta:

$$\frac{1}{\sqrt{b}} = \frac{C_o}{\sum \sqrt{c_h}\, p_h S_h}$$

Llevando a (8) el valor hallado para $1 / \sqrt{b}$, se tiene finalmente

$$n_h = \frac{p_h S_h C_o}{\sqrt{c_h} \sum \sqrt{c_h}\, p_h S_h} \qquad (9)$$

que da la adjudicación óptima para un costo total prefijado.
Si los c_h fueran iguales a c para todo h, se tendría:

$$C_o = nc$$

y la (9) se reduciría a :

$$n_h = n \frac{p_h S_h}{\sum p_h S_h}$$

que es el mismo resultado hallado en el párrafo anterior.
Para determinar la adjudicación que asegure un costo mínimo para tener una variancia V_o prefijada, partimos de :

$$V_o = \sum_1^k \left(\frac{1}{n_h} - \frac{1}{N_h} \right) p_h^2 \, S_h^2$$

$$\therefore \quad \sum_1^k \frac{1}{n_h} p_h^2 S_h^2 = V_o + \sum_1^k \frac{1}{N_h} p_h^2 S_h^2$$

Ahora, como

$$\frac{1}{n_h} = \frac{\sqrt{b c_h}}{p_h S_h}$$

y $$p_h = \frac{N_h}{N}$$

resulta:

$$\sqrt{b} \sum_1^k \sqrt{c_h} \, p_h S_h = V_o + \frac{1}{N} \sum_1^k p_h S_h^2$$

$$\therefore \quad \frac{1}{\sqrt{b}} = \frac{\sum_{1}^{k} \sqrt{c_h}\, p_h\, S_h}{V_o + \frac{1}{N} \sum_{1}^{k} p_h\, S_h^2}$$

de modo que, reemplazando en la (8) queda:

$$n_h = \frac{p_h\, S_h}{\sqrt{c_h}} \cdot \frac{\sum \sqrt{c_h}\, p_h\, S_h}{V_o + \frac{1}{N} \sum p_h\, S_h^2} \qquad (10)$$

Si, pues, se conoce el costo unitario por observación en cada estrato, y se desea lograr una precisión prefijada para la estimación, la adjudicación de las n_h en los diferentes estratos de acuerdo con la (10), asegura un costo total mínimo.-

Si $c_h = c$ para $h = 1, 2 \ldots k$, la (10) se reduce a:

$$n_h = p_h\, S_h \frac{\sum p_h\, S_h}{V_o + \frac{1}{N} \sum p_h\, S_h^2}$$

y, sumando para h de 1 a k:

$$n = \frac{\left(\sum p_h\, S_h \right)^2}{V_o + \frac{1}{N} \sum p_h\, S_h^2} \qquad (11)$$

teniéndose así el tamaño mínimo de la muestra requerida para estimar la media de la población con una precisión prefijada. La (11) podría haberse obtenido despejando n en la expresión que da $V_{op}(\bar{y}_e)$ y escribiendo V_o en lugar de $V_{op}(\bar{y}_e)$.

Comparación de las V en el muestreo estratificado y en el muestreo simple al azar.

Supuesto que se extrae una muestra simple al azar de extensión n de una población de N unidades, sabemos que la varian cia de la estimación de la media de la población viene da da por:

$$V(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) \cdot S^2 \qquad (11)$$

Si la población ha sido estratificada, formándose k es tratos, con N_h unidades por estrato, hemos visto que según que se haya hecho una adjudicación proporcional o una adjudicación óptima de la muestra en los diversos estratos, la variancia de la estimación de la media viene dada por:

$$V_p(\bar{y}_e) = \left(\frac{1}{n} - \frac{1}{N}\right) \sum_{1}^{k} p_h S_h^2 \qquad (12)$$

$$V_{op}(\bar{y}_e) = \frac{1}{n}\left(\sum_{1}^{k} p_h S_h\right)^2 - \frac{1}{N}\sum_{1}^{k} p_h S_h^2 \qquad (13)$$

La comparación de V, V_p, y V_{op} es particularmente inte resante, pues pone de manifiesto el camino por el cual puede lograrse una mayor ventaja de la estratificación.

a) Comparación de V_p en V.

Restando la (12) de la (11), se tiene:

$$V(\bar{y}) - V_p(\bar{y}_e) = \left(\frac{1}{n} - \frac{1}{N}\right)\left[S^2 - \sum_{1}^{k} p_h S_h^2\right] \qquad (14)$$

lo que nos lleva a comparar S^2 con $\sum p_h S_h^2$

Ahora , S^2 es la variancia en la población total y está dada por:

$$S^2 = \frac{1}{N-1} \sum_{h=1}^{k} \sum_{j=1}^{n_h} (a_{hj} - \mu)^2$$

y podemos escribir:

$$(N-1)S^2 = \sum_1^k \sum_1^{n_h} (a_{hj} - \mu_h + \mu_h - \mu)^2$$

$$= \sum_1^k \sum_1^{n_h} (a_{hj} - \mu_h)^2 + \sum_1^k N_h (\mu_h - \mu)^2$$

puesto que $\sum \sum (a_{hj} - \mu_h)(\mu_1 - \mu) = 0$). Pero

$$\sum_1^{N_h} (a_{hj} - \mu_h)^2 = (N_h - 1)S_h^2$$

De modo que resulta

$$(N-1)\,S^2 = \sum_1^k (N_h - 1)S_h^2 + \sum_1^k N_h (\mu_h - \mu)^2$$

dividiendo ahora por N queda:

$$\frac{N-1}{N}\,S^2 = \sum_1^k \frac{N_h - 1}{N_h} \cdot \frac{N_h}{N}\,S_h^2 + \sum_1^k \frac{N_h}{N}(\mu_h - \mu)^2$$

Si N y N_h son suficientemente grandes, $(N-1)N$ y $(N_h - 1)/N_h$ serán próximos a la unidad y podremos escribir (aproximadamente)

$$S^2 = \sum_1^k p_h S_h^2 + \sum_1^k p_h (\mu_h - \mu)^2$$

de donde resulta :

$$S^2 - \sum_1^k p_h S_h^2 = \sum_1^k p_h (\mu_h - \mu)^2$$

y reemplazando este resultado en la (14), queda finalmente:

$$V(\bar{y}) - V_p\ (\bar{y}_e) = \left(\frac{1}{n} - \frac{1}{N} \right) \sum_1^k p_h (\mu_h - \mu)^2 \qquad (15)$$

lo que nos dice que, cuanto más difieran entre sí las medidas de los estratos, mayor será la ganancia en la precisión de la estimación basada en una muestra estratificada con adjudicación proporcional con respecto a la obtenida mediante una muestra simple al azar, de la misma extensión total.

b) Comparación de V_p con V_{op}.-

Recordando las expresiones que dan V_p y V_{op}, tenemos:

$$V_p - V_{op} = \left(\frac{1}{n} - \frac{1}{N} \right) \sum p_h S_h^2 - \frac{1}{n} \left(\sum p_h S_h \right)^2 + \frac{1}{n} \sum p_h S_h^2 =$$

$$= \frac{1}{n} \left[\sum p_h S_h^2 - \left(\sum p_h S_h \right)^2 \right]$$

Si indicamos $\sum p_h S_h$ con $\bar{S}$, la anterior se escribirá:

$$V_p - V_{op} = \frac{1}{n} \left[\sum p_h S_h^2 - \bar{S}^2 \right]$$

Ahora el segundo miembro es idénticamente igual a:

$$\frac{1}{n} \sum p_h (S_h - \bar{S})^2$$

de modo que se tiene finalmente:

$$V_p - V_{op} = \frac{1}{n} \sum p_h (S_h - \bar{S})^2 \qquad (16)$$

lo que nos dice que, la adjudicación óptima dará comparativamente con la proporcional, una precisión tanto mayor cuanto más difieran entre sí los diferentes estratos en su variabilidad.

c) Comparación de V_{op} con V.

Si en la (15) reemplazamos el valor de $V_p\ (\bar{y}_e)$ por el que tiene en la (16), resulta :

$$V(\bar{y}) - \frac{1}{n} \sum_{1}^{k} p_h (S_h - S)^2 - V_{op} (\bar{y}_e) = \left(\frac{1}{n} - \frac{1}{N} \right) \sum_{1}^{k} p_h (\mu_h - \mu)^2$$

o sea:

$$V(\bar{y}) - V_{op} (\bar{y}_e) = \left(\frac{1}{n} - \frac{1}{N} \right) \sum_{1}^{k} p_h (\mu_h - \mu)^2 + \frac{1}{n} \sum_{1}^{k} p_h (S_h - S)^2 \quad (17)$$

La interpretación de este resultado es inmediata.

EJEMPLO: En el cuadro siguiente se da la distribución, según el volumen en pesos de sus ventas anuales, de los 2.529 establecimientos textiles registrados en el Censo Industrial de 1948 en cinco intervalos que se consideraran como constituyendo otros tantos estratos.

Estrato Nº	Volumen de ventas (miles de $)	Número de establecimientos N_h	Ventas totales (miles de $) A_h	Promedio de ventas (miles de $) U_h	Desvío Standard (miles de $) S_h (.)
1	100	952	32.980	34.6	32.8
2	100 - 500	712	178.545	250.8	120.-
3	500 - 1.000	296	215.814	729.1	352.2
4	1.000 - 10.000	507	1.395.898	2.753.2	1.566.-
5	10.000 - 100.000	62	1.374.964	22.176.8	18.350.-
	TOTALES:	2.529	3.198.201	1.264.6	4.680.-

Con los datos de este cuadro calcularemos los errores standard de las estimaciones del volumen total de ventas que se tendrían tomando: bien una muestra simple al azar de establecimientos, bien muestras estratificadas de la misma extensión, y esto, de acuerdo con diversos modos de adjudicación.

Supondremos que el tamaño de la muestra a tomarse es igual aproximadamente del 20 % de la población, es decir, n = 506.

El error standard de la estimación del total basado en una muestra simple al azar de la extensión fijada, sería:

$$\sigma_{N_y} = NS \sqrt{\frac{1}{n} - \frac{1}{N}} = 2529 \times 4680 \times .0397 = 469.878 \$$$

que significa un coeficiente de variación de 14.3 %.

Una muestra estratificada con adjudicación proporcional requeriría tomar muestras de las siguientes extensiones:

$n_1 = 190$ $n_2 = 142$ $n_3 = 59$ $n_4 = 101$ $n_5 = 12$ $(n = 504)$

teniéndose entonces para el error standard de la estimación del total

$$\sigma_{p\,(N_y)} = 297.157 \$ \qquad (C.V. = 9.3\ \%)$$

(.).- Los S_h anotados no son reales, sino son valores -que quizá pueden aproximarse a los reales- que se han tomado únicamente para los fines del ejemplo.

M.t.p.

Lo que significa una reducción del coeficiente de variación de, aproximadamente, un 35%.-

En el cuadro siguiente se dan los valores requeridos para calcular ese error standard, como así también los que se necesitan para calcular los n_h de la adjudicación óptima.-

Estrato Nº	$p_h S_h$	$p_h S_h^2$
1	12.35	405.-
2	33.78	4.053.60
3	41.24	14.525.65
4	313.98	491.697.38
5	449.58	8.249.701.25
total	850.93	8.760.382.88

Hemos visto más arriba que para la adjudicación óptima debe tomarse:

$$n_h = n \frac{p_h S_h}{\sum p_h S_h}$$

En el caso actual se tiene:

$$n_1 = 506 \times .01451 \doteq 7$$
$$n_2 = 506 \times .03970 \doteq 20$$
$$n_3 = 506 \times .04847 \doteq 25$$
$$n_4 = 506 \times .36898 \doteq 187$$
$$n_5 = 506 \times .52834 \doteq 267$$

y nos encontramos con el caso en que la fórmula de la adjudicación óptima requiere tomar, de un determinado estrato, una muestra de extensión superior al número de unidades que hay en él.- Ya hemos indicado como se resuelve esta situación.- Deben incluirse en la muestra las 62 unidades del 5º estrato, distribuyéndose las 444 restantes de acuerdo a lo que resulte de la adjudicación óptima en los otros 4 estratos.- Recalculando los valores de $p_h S_h$ y $p_h S_h / \sum p_h S_h$ se tiene ahora:

$$n_1 = 444 \times .03077 \doteq 14$$
$$n_2 = 444 \times .08416 \doteq 37$$
$$n_3 = 444 \times .10275 \doteq 46$$
$$n_4 = 444 \times .78232 \doteq 347$$
$$n_5 =$$

Con esta adjudicación, el error standard de la estimación del total resulta ser:

$$\sigma_{op}(N\bar{y}) = 30.837\ \$$$

de modo tal que se ha logrado una estimación que tiene un coeficiente de variación inferior a 1% (.8%), y que por cierto, compara ventajosamente con las correspondientes al muestreo simple al azar o al estratificado con adjudicación proporcional.

A continuación se resumen los anteriores resultados, agregando otro que discutiremos más adelante.-

Estrato Nº	Muestreo simple al azar	Adjud. proporcional n_h	Adjud. proporcional n_h/N_h	Adjud. óptima n_h	Adjud. óptima n_h/N_h	Adjud. prop. al tamaño n_h	Adjud. prop. al tamaño n_h/N_h
1		190	19.96%	14	1.47%	8	.84%
2		142	19.94%	37	5.19%	43	6.04%
3		59	19.93%	46	15.54%	53	17.90%
4		101	19.92%	347	68.44%	340	67.06%
5		12	19.35%	62	100.-%	62	100.-%
n	506	504	19.94%	506		506	
$W_{N\bar{y}}$	14.3%	9.3%		.8%		1%	

MUESTREO ESTRATIFICADO PARA LA ESTIMACION DE PROPORCIONES

Supongamos que se trata de estimar la proporción de unidades que en una población de N posee una cierta característica "C", estando la población en cuestión dividida en k estratos de N_h ($h = 1,2..k$) unidades cada uno.-

Sea

$$P_h = A_h/N_h$$

la proporción de unidades "C" en el h-ésimo estrato (A_h es el número de unidades "C" en dicho estrato). Si se extrae una muestra simple al azar de extensión n_h del h-ésimo estrato, indicaremos con

$$p_h = a_h/n_h$$

La proporción de unidades "C" en esa muestra. La extensión total de la muestra es de $n = \sum_1^k n_h$, y $\sum_1^k a_h/n$, la proporción de unidades "C" es la misma.-

M. t. p.

La estimación de la proporción de unidades "C" en la población total está dada por:

$$p_e = \frac{1}{N} \sum_{1}^{k} N_h p_h \qquad (1)$$

es decir por la media aritmética ponderada de las proporciones de unidades "C" en las muestras obtenidas de los k estratos.-

Esta fórmula, así como todas las siguientes, pueden derivarse directamente de las obtenidas en los párrafos procedentes si se asigna el valor 1 a toda unidad que posee la característica "C" y el valor 0, a toda unidad "no-C", con lo que resulta que P_h no es sino el valor medio por unidad en el estrato h y p_h el valor medio por unidad en la muestra extraída de él.-

Teniendo esto en cuenta, y recordando que, tratándose de proporciones (pag. 22), es

$$S^2 = \frac{N}{N-1} P.Q \qquad (Q = 1-P)$$

se sigue de inmediato (de (2) ó (2'), pag. II.9 y II. 10

$$V(p_e) = \sum_{1}^{k} \left(\frac{1}{n_h} - \frac{1}{N_h} \right) . \left(\frac{N_h}{N} \right) \frac{N_h}{N_h - 1} P_h Q_h \qquad (2)$$

o bien

$$V(p_e) = \frac{1}{N^2} \sum_{1}^{k} \frac{N_h^2 (N_h - n_h)}{N_h - 1} . \frac{P_h Q_h}{n_h} \qquad (3)$$

Si N_h es grande, con lo que $N_h / (N_h - 1)$ puede tomarse como igual a la unidad, se tiene poniendo (como lo haremos en lo que sigue) $N_h / = w_h$ (.):

$$V(p_e) = \sum_{1}^{k} w_h^2 \frac{P_h Q_h}{n_h} \qquad (4)$$

En el caso de adjudicación proporcional ($w_h = n/N$):

$$V(p_e) = \left(\frac{1}{n} - \frac{1}{N} \right) \sum_{1}^{k} w_h P_h Q_h \qquad (5)$$

Para un tamaño fijo n de la muestra total, la adjudicación óptima será aquella que asigne al estrato h una muestra de extensión

---

(.) Anteriormente usabamos p_h para expresar el cociente N_h/N; usamos ahora w_h para evitar confusión con la p_h que representa la proporción de unidades "C" en la muestra del estrato h.

$$n_h = n \frac{w_h \sqrt{P_h Q_h}}{\sum w_h \sqrt{P_h Q_h}} \qquad (6)$$

supuesto que $N_h/(N_h - 1) \simeq 1$

El valor de la variancia de la estimación en el caso de una adjudicación óptima es:

$$V_{op}(p_e) = \frac{1}{n}\left(\sum_1^k w_h \sqrt{P_h Q_h}\right)^2 - \frac{1}{N}\sum_1^k w_h P_h Q_h$$

Por otra parte, de la (5), se sigue que la variancia de la estimación en el caso de una adjudicación proporcional, es:

$$V_p(p_e) = \frac{1}{n}\sum_1^k w_h P_h Q_h - \frac{1}{N}\sum_1^k w_h P_h Q_h$$

de modo que resulta:

$$V_p(p_e) - V_{op}(p_e) = \frac{1}{n}\left[\sum_1^k w_h P_h Q_h - \left(\sum_1^k w_h \sqrt{P_h Q_h}\right)^2\right]$$

de donde se sigue facilmente una conclusión análoga a la obtenida de la (16) (pag. II.22).-

Las estimaciones $v(p_e)$ de $V(p_e)$ para adjudicación proporcional u óptima se obtienen reemplazando P_h y Q_h por los correspondientes p_h y q_h obtenidos en la muestra, lo que, si bien no dá estimaciones "no viciadas", es apropiado para la determinación de los límites de confianza.-

M. t. p.

EVALUACION, EN BASE A LA MUESTRA, DE LA GANANCIA OBTENIDA POR LA ESTRATIFICACION

Una vez obtenida la muestra de una población dividida en estratos, resulta de interés, para orientar investigaciones futuras, evaluar la ganancia en la precisión alcanzada, comparada con la que se hubiera obtenido extrayendo de la misma población una muestra simple al azar de idéntica extensión.

Sabemos que la estimación de la media de la población está dada por

$$\bar{y}_e = \frac{1}{N} \sum_{1}^{k} N_h \bar{y}_h$$

y que la variancia de la misma es

$$V(\bar{y}_e) = \sum_{1}^{k} \left(\frac{1}{n_h} - \frac{1}{N_h}\right) p_h^2 S_h^2$$

Una estimación "no viciada" de esta variancia se obtiene reemplazando S^2 por s^2, es decir, sustituyendo la variancia desconocida de la población de cada estrato, por la variancia de las unidades de la muestra extraída del mismo, para tener,

$$v(\bar{y}_e) = \sum_{1}^{k} \left(\frac{1}{n_h} - \frac{1}{N_h}\right) p_h^2 s_h^2$$

Ahora bien la variancia de la media $\bar{y}$ de una muestra simple al azar de extensión n está dada por:

$$V(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right) S^2 \qquad (1)$$

donde S^2 es la variancia de la población. Para hacer la comparación se necesita tener una estimación $v(\bar{y})$ de $V(\bar{y})$, la que no es posible obtener reemplazando S^2 por s^2 en (1) calculada a partir de los datos del muestreo estratificado. El problema consiste, pues, en hallar una estimación "no viciada" de S^2 (y por ende de $V(\bar{y})$ basada en los resultados dados por la muestra extraída de la población dividida en estratos.

Sabemos que:

$$S^2 = \frac{1}{N-1} \sum_{1}^{k} (N_h - 1) S_h^2 + \frac{N}{N-1} \sum_{1}^{k} P_h (\mu_h - \mu)^2 \qquad (2)$$

donde μ_h y μ son, respectivamente, la media de la población en el h-ésimo estrato y la media de la población total. Como s_h^2 es una estimación "no viciada" de S_h^2, la obtención de una estimación de S^2 que tenga la misma propiedad, requiere que, en base a los datos ofrecidos por la muestra estratificada, se construya una función de las observaciones, que indicaremos en $F(y_{hj})$, tal que:

$$E\left[F(y_{hj})\right] = \sum_1^k P_h(\mu_h - \mu)^2 = \sum_1^k P_h \mu_h^2 - \mu^2 \qquad (3)$$

Una vez hallada esa función se tendrá, en efecto:

$$\frac{1}{N-1} E\left\{\sum_1^k (N_h - 1)s_h^2 + NF(y_{hj})\right\} =$$

$$= \frac{1}{N-1}\left\{\sum_1^k (N_h - 1) S_h^2 + N\sum_1^k P_h (\mu_h - \mu)^2\right\} = S^2 \qquad (4)$$

Para construir esa función $F(y_{hj})$ observamos que la media de una muestra extraída del h-ésimo estrato puede expresarse del modo siguiente:

$$\bar{y}_h = \mu_h + z_h \qquad (a)$$

siendo z_h (que no es sino la desviación de la media de la muestra con respecto a la media del estrato) una variable que verifica:

$$E(z_h) = 0 \qquad E(z_h^2) = \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2$$

Multiplicando ambos miembros de la (a) por p_h y sumando con respecto a h de 1 a k, se tiene:

$$\sum_1^k p_h \bar{y}_h = \sum_1^k p_h \mu_h + \sum_1^k p_h z_h$$

o sea

$$\bar{y}_e = \mu + \sum_1^k p_h z_h \qquad (b)$$

puesto que

$$\bar{y}_e = \frac{1}{N} \sum_1^k N_h \bar{y}_h$$

$$\mu = \frac{1}{N} \sum_1^k N_h \mu_h$$

Elevando al cuadrado la (a) y la (b) queda:

$$\bar{y}_h^2 = \mu_h^2 + z_h^2 + 2\mu_h z_h \qquad (c)$$

$$\bar{y}_e^2 = \mu^2 + \sum p_h^2 z_h^2 + \sum_{h \neq j} \sum p_h p_j z_h z_j + 2\mu \sum p_h z_h \qquad (d)$$

teniéndose para sus respectivas esperanzas matemáticas:

$$E(\bar{y}_h^2) = \mu_h^2 + \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2 \qquad (e)$$

$$E(\bar{y}_e^2) = \mu^2 + \sum_1^k \left(\frac{1}{n_h} - \frac{1}{N_h}\right) p_h^2 S_h^2 \qquad (f)$$

dado que, como se indicó más arriba:

$$E(z_h) = 0 \qquad E(z_h^2) = \left(\frac{1}{n_h} - \frac{1}{N_h}\right) S_h^2$$

y $$E(z_h z_j) = E(z_h) \cdot E(z_j)$$

M.t.p.

puesto que los variables aleatorios z_h son independientes, en tanto que lo son las muestras extraídas de los distintos estratos.

Multiplicando ambos miembros de la (e) por p_h y sumando con respecto a h de 1 a k, resulta:

$$E\left(\sum p_h \bar{y}_h^2\right) = \sum_1^k p_h \mu_h^2 + \sum_1^k \left(\frac{1}{n_h} - \frac{1}{N_h}\right) p_h S_h^2 \quad (g)$$

Ahora los 2dos. términos de los 2dos. miembros de (f) y (g) no son sino las esperanzas matemáticas de las expresiones que se obtienen reemplazando S_h^2 por s_h^2, de modo que podemos escribir:

$$\mu^2 = E\left\{\bar{y}_e^2 - \sum_1^k \frac{1}{n_h} - \frac{1}{N_h} \; p_h^2 \; s_h^2\right\}$$

$$\sum_1^k p_h \; {}^2_h = E\left\{\sum_1^k p_h \bar{y}_h^2 - \sum_1^k \left(\frac{1}{n_h} \quad \frac{1}{N_h}\right) p_h \; s_h^2\right\}$$

Restando la 1ra. de la 2da. de estas igualdades, se obtiene, después de simples transformaciones:

$$\sum_1^k p_h(\mu_h - \mu)^2 = E\left\{\sum_1^k p_h(\bar{y}_h - \bar{y}_e)^2 - \sum_1^k \left(\frac{1}{n_h} - \frac{1}{N_h}\right) p_h(1-p_h)s_h^2\right\} \quad (h)$$

lo que muestra que la expresión entre corchetes es precisamente la función $F(y_{hj})$ que se buscaba. Reemplazando en el primer miembro de la (4) $F(y_{hj})$ por su valor hallado en (h), tendremos una estimación "no - viciada" de S^2 dada por:

$$Est(S^2) = \frac{1}{N-1}\sum_1^k (N_h - 1)s_h^2 + \frac{N}{N-1}\left\{\sum_1^k p_h(\bar{y}_h - \bar{y}_e)^2 - \right.$$

$$\left. - \sum_1^k \left(\frac{1}{n_h} - \frac{1}{N_h}\right) p_h(1-p_h)s_h^2\right\}$$

M.t.p.

o bien:

$$Est(S^2) = \sum_1^k p_h s_h^2 + \frac{N}{N-1}\left\{\sum_1^k p_h(\bar{y}_h-\bar{y}_e)^2 - \sum_1^k p_h(1-p_h)s_h^2/n_h\right\} \quad (4)$$

Si en (1) reemplazamos ^{2} por su estimación dada en (4), se tendrá una estimación "no viciada" de $V(\bar{y})$, basada en los resultados de la muestra estratificada, dada por:

$$v(\bar{y}) = \left(\frac{1}{n} - \frac{1}{N}\right)\left\{\sum_1^k p_h s_h^2 + \frac{N}{N-1}\left\{\sum_1^k p_h(\bar{y}_h-\bar{y}_e)^2 - \sum_1^k p_h(1-p_h)s_h^2/n_h\right\}\right\} \quad (5)$$

y la estimación de la diferencia entre $V(\bar{y})$ y $V(\bar{y}_e)$, será:

$$V(\bar{y}) - v(\bar{y}_e) = \frac{1}{n}\sum_1^k p_h s_h^2 - \sum_1^k p_h^2 s_h^2/n_h + \frac{N-n}{(N-1)n}\left\{\sum_1^k p_h(\bar{y}_h-\bar{y}_e)^2 - \sum_1^k p_h(1-p_h)s_h^2/n_h\right\} \quad (6)$$

Si la población de la cual se ha extraído la muestra es grande, de modo tal que $N/(N-1) \neq 1$, de la (6) se obtiene:

$$v(\bar{y}) - v(\bar{y}_e) = \frac{1}{n}\sum_1^k p_h s_h^2 - \sum_1^k p_h^2 s_h^2/n_h + \frac{N-n}{N_n}\left\{\sum_1^k p_h(\bar{y}_h-\bar{y}_e)^2 - \sum_1^k p_h(1-p_h)s_h^2/n_h\right\} \quad (7)$$

Relacionando la (6) ó la (7) con la

$$v(\bar{y}_e) = \sum_1^k \left(\frac{1}{n_h} - \frac{1}{N_h}\right) p_h^2 s_h^2$$

se tiene una estimación que puede expresarse como porcentaje de la ganancia obtenida por la estratificación

En el caso de una adjudicación proporcional de la muestra en los diferentes estratos, los dos primeros términos del segundo miembro de la (7) resultan iguales, de manera que queda entonces:

$$v(\bar{y}) - v_p(\bar{y}_e) = \frac{N-n}{N_n}\left\{\frac{1}{n}\sum_1^k n_h(\bar{y}_h-\bar{y}_e)^2 - \frac{1}{n}\sum_1^k\left(1-\frac{n_h}{n}\right)s_h^2\right\} \qquad (8)$$

La (5) puede escribirse, suponiendo $N/(N-1) \simeq 1$

$$v(\bar{y}) = \frac{N-n}{N_n}\left\{\sum_1^k\left(p_h - \frac{p_h}{n_h} + \frac{p_h^2}{n_h}\right)s_h^2 + \sum_1^k p_h(\bar{y}_h-\bar{y}_e)^2\right. \qquad (9)$$

Ahora, si las S_h^2 son iguales en todos los estratos, es decir

$$s_h^2 = s_w^2 \qquad (h = 1,2 \ldots\ldots k)$$

donde s_w^2 indica la variancia común "dentro" de los estratos, una estimación"no viciada" viene dada por

$$s_w^2 = \frac{1}{n-k}\sum_1^k\sum_1^{n_h}(y_{hi} - \bar{y}_h)^2$$

reemplazando este valor en (9), resulta:

$$v(\bar{y}) = \frac{N-n}{N_n^2}\left\{(n-k+1)s_w^2 + \sum_1^k n_h(\bar{y}_h - \bar{y}_e)^2\right.$$

si se tiene en cuenta que $n_h = np_h$

Si finalmente se pone

$$\sum_1^k n_h(\bar{y}_h - \bar{y}_e)^2 = (k-1)\,\bar{n}\,s_b^2$$

donde $\bar{n} = n/k$, se tiene

$$v(\bar{y}) = \frac{N-n}{N}\,\frac{1}{n^2}\left\{(n-k+1)\,s_w^2 + (k-1)\,\bar{n}\,s_b^2\right\} \qquad (10)$$

M.tp.

Las cantidades s_w^2 y $\bar{n}\, s_b^2$ se denominan, respectivamente, cuadrados medios "dentro" y "entre" estratos, y se calculan fácilmente construyendo el siguiente cuadro de análisis de la variancia :

Fuente de variación	Grados de libertad	Suma de cuadrados	Cuadrados medios
Entre estratos. . . .	k - 1	$\sum n_h(\bar{y}_h - \bar{y}_e)^2$	$\bar{n}\ s_b^2$
Dentro de los estratos	n - k	$\sum\sum(y_{hi} - \bar{y}_h)^2$	s_w^2
Total . . .	n - 1	$\sum\sum(y_{hi} - \bar{y}_e)^2$	s^2

por otra parte, reemplazando s_h^2 por s_w^2 en la expresión que dá la variancia estimada en el caso de adjudicación proporcional, queda : :

$$v_p(\bar{y}_e) = \frac{N - n}{N} \quad \frac{1}{n} \quad s_w^2 \qquad (11)$$

De las (10) y (11) se obtiene fácilmente la expresión que dá la ganancia relativa en precisión lograda por la estratificación.-

M.t.p.

$$\frac{k - 1}{n} \left\{ \frac{\bar{n} s_b^2}{s_w^2} - 1 \right\}$$

DIFICULTADES PRACTICAS PARA LA DETERMINACION DE LA ADJUDICACION OPTIMA

Se ha visto que, para un mismo tamaño global n de la muestra a extraerse de una población dividida en estratos, la adjudicación óptima es la que lleva a la más alta precisión de la estimación de la característica estudiada y que, la diferencia en la precisión lograda mediante esa adjudicación comparada con la que ofrece una adjudicación proporcional es tanto más grande cuanto más difieren entre sí los estratos en su variabilidad interna. Siendo esto así y dejando de lado el hecho de que la adjudicación óptima, al requerir el uso de tasas de muestreo diferentes de estrato a estrato, supone una mayor complicación en los cálculos de las estimaciones y sus variancias, para tomar en cuenta sólo la eficiencia, será preferible en general la adjudicación óptima a la proporcional.

Ahora bien, en la práctica, dos son las dificultades que se hacen presentes cuando se trata de usar la adjudicación de Neyman-Tchuprow:

1º.- El desconocimiento del valor de S_h en cada estrato, al cual es proporcional el tamaño n_h de la muestra que de él ha de extraerse;

2º.- El hecho de que, en general, interesan en una investigación no sólo una, sino varias características de la población, de cada una de las cuales la muestra ha de ofrecer una estimación.

Cuando se estudia una sola característica, o varias que en cada estrato tienen la misma variabilidad o variabilidades que difieren entre sí por un factor constante, la primera dificultad se resuelve utilizando para la determinación de los n_h, estimaciones de S_h de las S_h desconocidas. Al proceder así, no se tendrá una adjudicación óptima, pero sí quizá una que se le aproxime y que resulte también ventajosa con respecto a la proporcional. Como regla práctica, la adjudicación óptima, o una que se le aproxime, se considerará preferible a la proporcional, si la ganancia esperada de precisión -calculada con anticipación a la extracción de la muestra- supera el 20 %.

Si se dispone de información previa, proveniente de otra u otras muestras o de censos anteriores, puede tomarse para la adjudicación de la muestra actual el valor de S_h en aquella oportunidad, siempre y cuando se pueda razonablemente admitir que hay una cierta estabilidad en la variabilidad de los valores de la característica objeto del estudio dentro de cada estrato. Importa destacar que no es necesario que los valores de S_h sean actualmente los mismos que se tenían en la ocasión previa para que sean utilizables para los fines de la adjudicación en la ocasión actual; basta que se mantenga su magnitud relativa de estrato a estrato. Supongamos, en efecto, que en una investigación previa se tiene que los S_h en los diferentes estratos están entre sí como los números:

$$c_1 : c_2 : c_3 ; \ldots : c_k$$

y que es admisible pensar que los desvíos standard S_h desconocidos actuales guardan entre sí las mismas relaciones. Si se atribuye a S_1 un

walor caulquiera positivo H, se tendrá:

$$S_1 = H \qquad S_2 = c_2 H \quad \ldots \quad S_k = c_k H$$

de modo que

$$\frac{1}{N} \sum N_h S_h = \frac{H}{N} \sum (1 + c_2 + \ldots + c_k) \; N_h = \frac{H}{N} C \sum N_h = HC$$

indicando con C la suma $1 + c_2 + \ldots + c_k$. La adjudicación óptima (o una aproximación) se tendrá entonces tomando

$$n_h = n \frac{N_h S_h}{\sum N_h S_h} = n \frac{N_h c_h H}{HC} = n \frac{N_h c_h}{C}$$

El resultado anterior no es sino una consecuencia del hecho de que, por su forma, para la aplicación de la fórmula de la adjudicación basta conocer números $cS_1, cS_2, \ldots, cS_k$ que difieran de los respectivos S_h en una constante multiplicativa cualquiera c

Puede afirmarse que, si de algún modo se han obtenido estimaciones relativamente razonables de los S_h, o de las relaciones que entre sí guardan, podrá lograrse una adjudicación que dará una variancia de la estimación que difiere poco de la que se obtendría por una adjudicación óptima, dado que dicha variancia es poco sensible a las desviaciones de los valores n_h con respecto a aquellos que la hacen un mínimo. Un medio práctico para lograr una adjudicación que se aproxima a la óptima cuando las unidades de la población difieren más o menos notablemente en "tamaño" y tienen al mismo tiempo una relativa estabilidad, de modo que el "tamaño" en una cierta fecha ofrece una estimación razonable de su "tamaño" relativo en otra subsiguiente, es aquel que, cuando se trata de estimar características estrechamente correlacionadas con aquella que se ha adoptado para definir el "tamaño", partiendo del supuesto de que las S_h de cada estrato son aproximadamente proporcionales al "tamaño" medio de los elementos en cada uno de ellos, adjudica la muestra de modo tal que las tasas de muestreos son proporcionales a dicho "tamaño" medio.

Por "tamaño" de una unidad se entiende el valor que una cierta característica toma para la unidad en cuestión. Así, p.e., si se trata de establecimientos industriales, su "tamaño" puede estar dado por la producción anual en unidades físicas o en pesos, por el número de obreros ocupados, el monto de los salarios pagados, el capital empleado, etc.; tratándose de explotaciones agrícolas, establecimientos comerciales, o familias, el "tamaño" puede ser dado por los ingresos anuales, etc.

Ahora bien, en los casos mencionados más arriba y en otros muchos que podrían citarse, ocurre que hay una marcada estabilidad del "tamaño" en el tiempo, por lo que el "tamaño" de las unidades en una fecha ofrece una medida bastante aproximada de su "tamaño" relativo en otra posterior.

Sea t_{hj} el valor que la característica con la que se define el "tamaño" toma en la j-ésima unidad del h-ésimo estrato; el tamaño medio por unidad en dicho estrato será:

$$\bar{T}_h = \frac{1}{N_h} \sum_{j=1}^{k} t_{hj}$$

Admitiendo que S_h es proporcional a $\bar{T}_h$, es decir

$$S_h / \bar{T}_h = c \text{ (constante) para } h = 1, 2, \ldots, k$$

reemplazando S_h por $\bar{T}_h$ en la fórmula

$$n_h = n \frac{N_h S_h}{\sum N_h S_h}$$

que dá la adjudicación óptima, se tiene:

$$n_h = n \frac{N_h \bar{T}_h}{T}$$

donde T representa la suma $\sum N_h \bar{T}_h$. De la anterior se sigue que

$$n_h / N_h = \bar{T}_h \quad (n/T)$$

es decir, la tasa de muestreo es, para cada estrato, proporcional al "tamaño" medio por unidad en el mismo. (*)

El supuesto básico del resultado precedente es que el coeficiente de variación de la característica que se estudia es aproximadamente el mismo en todos los estratos, de modo tal que es:

$$S_h / \mu_h = \text{constante para } h = 1, 2, \ldots, k$$

Ahora, si los valores de la característica que interesa son aproximadamente proporcionales al "tamaño" de modo tal que en media es

$$\mu_h \doteq c\bar{T}_h \quad (c = \text{constante})$$

(*).- Obsérvese que el anterior razonamiento es el mismo que se hizo más arriba cuando se señaló que para la adjudicación óptima basta conocer un conjunto de números proporcionales a los S_h.

será

$$S_h / \bar{T}_h \doteq \text{constante}$$

de donde se sigue la regla de adjudicación mencionada.

En el ejemplo de los establecimientos textiles registrados en el Censo Industrial de 1948, se ha calculado la adjudicación que se obtendría a partir del "tamaño" dado por el monto promedio de las ventas en pesos, y se ve en el cuadro comparativo final cuan poco difiere ella de la adjudicación óptima y cuan pequeña es la diferencia en el coeficiente de variación de la estimación del total al que se llega mediante uno y otro procedimiento de adjudicación. Vale la pena observar que esta notable aproximación se ha obtenido aún cuando los coeficientes de variación de los diversos estratos no son iguales entre sí. En efecto, se tiene:

Estrato Nº	C.V.
1	94,8 %
2	47,8 %
3	48,3 %
4	56,8 %
5	82,7 %

Este resultado es general en la práctica. Si las unidades difieren notablemente en "tamaño" de estrato a estrato (tal como ocurre en el ejemplo), para los fines de la adjudicación, los coeficientes de variabilidad pueden considerarse, como suficientemente próximos entre sí cuando la máxima variación relativa entre ellos alcanza el valor 2 o aún 3.

Es evidente que en un caso práctico, el "tamaño" medio de las unidades de cada estrato que servirá de base para la determinación de la adjudicación, es desconocido al momento de diseñarse la muestra.

La adjudicación se hará en base al "tamaño" medio de alguna fecha anterior si, como se ha indicado, puede admitirse que el mismo ofrece una estimación razonable del "tamaño" relativo actual.

En el caso del ejemplo precedente, el "tamaño" dado por los resultados del relevamiento censal de 1948, podría haberse utilizado para determinar la adjudicación de una muestra en una fecha posterior destinada a estimar, no solamente el monto total de ventas de pesos, sino también, el número de obreros ocupados, el monto de salarios pagados, valor de las materias primas o energía eléctrica consumida, o cualquier otra característica más o menos estrechamente relacionada con el monto de las ventas.

Si el supuesto de la aproximada constancia del coeficiente de variación en los varios estratos no es acertada, puede ocurrir que la aplicación de la regla lleve a una adjudicación por la cual algún estrato resulte muy pobremente representado (en dicho estrato el coeficiente de variación es notablemente mayor que en los restantes, lo que

generalmente suele ocurrir en aquellos cuyo "tamaño" medio es pequeño). En este caso la extensión de la muestra que este método de adjudicación le asigna será la tercera o la cuarta parte de la que le atribuiría la adjudicación óptima. Cuando esto ocurre, debe incrementarse el tamaño de la muestra a tomarse de dicho estrato, lo que puede lograrse asociando a él un "tamaño" doble o triple del que corresponde, antes de hacer la determinación de la adjudicación.

En lo que a la segunda dificultad señalada más arriba se refiere, a saber, la determinación de la adjudicación de la muestra en el caso en que se requiere la estimación de varias características de la población estratificada, no es, en algunos casos, realmente seria, como puede mostrarse si se tienen en cuenta algunas de las consideraciones precedentes. En efecto, se ha señalado que, en muchos casos, interesan en un estudio diversas características de los elementos de la población que tienden a variar en forma paralela de unidad a unidad, tal como ocurre p.e. con el volumen de las ventas, personal ocupado, salarios y sueldos pagados, etc., etc., en establecimientos industriales o comerciales o bien, los ingresos mensuales, alquiler pagado, etc., en las familias, de modo que, en los diversos estratos, los desvíos standard S_h de cada una de estas características tienden a ser aproximadamente proporcionales, lo que hace que la adjudicación de la muestra que es o se aproxima al óptimo para la estimación de cada una de ellas sea también razonablemente buena para la estimación de las otras.

No ocurre así, en general, cuando se trata de estimar proporciones en la población total. En este caso, la adjudicación óptima para la estimación del porcentaje de individuos de la población que poseen un cierto atributo, frecuentemente no lo es para la estimación de la proporción de los que poseen otro, de modo que, cuando se trata de obtener estimaciones de diversos atributos la adjudicación proporcional es con frecuencia la preferible.

Resulta de aquí que, cuando en un estudio se requieren estimaciones de características "contínuas" y de atributos, es forzoso un análisis previo que, a partir de los requerimientos de precisión impuestos, establezca en primer término la adjudicación que es óptima, o que se le aproxima, para las características "contínuas" (monto de ventas, ingresos, salarios, consumo de energía eléctrica, etc.) y ajuste luego el tamaño de la muestra a extraerse de alguno o algunos de los estratos de manera que puedan obtenerse estimaciones de porcentajes que satisfagan esos requerimientos de precisión.

Si, cuando se ha estratificado una población de acuerdo al "tamaño" de las unidades, de lo que se trata es de tener estimaciones de porcentajes separadamente en los diversos estratos con el objeto de hacer comparaciones, la adjudicación que es óptima para las características "contínuas", resulta ser también la apropiada para dicho objeto. En efecto, la adjudicación óptima tiende a incrementar el tamaño de la muestra a tomarse de aquellos estratos en los que se agrupan los relativamente pocos individuos de la población en los que la característica "tamaño" toma los valores más altos, al mismo tiempo que reduce el tamaño de la muestra a tomarse de aquellos otros estratos que agrupan a los más numerosos individuos "pequeños". Y precisamente el "número" de individuos en la muestra tiene más importan

cia que la "proporción" de los mismos en la muestra de una cierta población finita en la precisión de la estimación de un porcentaje.

Se tiene pues que, cuando se trata de estimar características "contínuas" para la población y porcentajes en los diversos estratos de una población estratificada de acuerdo al "tamaño" de los individuos, la adjudicación óptima es la apropiada.

Una vez más debe señalarse que no es posible dar reglas prácticas de aplicación general, sino tan solo apuntar hacia consideraciones que es necesario tener presentes al momento del diseño de una muestra. El objetivo es, en todos los casos, alcanzar la máxima eficiencia dado un cierto conjunto de medios y recursos, y, fundamentalmente, de conocimientos -ofrecidos por censos o muestras previas- acerca de las características de la población a estudiarse. El ingenio y la habilidad del muestrista, ayudados por la teoría y la experiencia, son los ingredientes principales, para, en cada caso particular, aprovechar eficientemente la información y los recursos disponibles, mediante un diseño apropiado de la muestra : método de muestreo y de estimación.

CASO DE N_h DESCONOCIDO

Puede ocurrir que para una cierta estratificación, conveniente o deseable, se desconozca el total N_h de la población en cada estrato y que deba recurrirse a valores estimados N_h'.

Así p.e., tratándose de una investigación en la que las unidades son explotaciones agrícolas estratificadas por región geográfica, puede ocurrir que se desconozca el número actual de explotaciones en cada estrato, teniéndose sí, el número que había en una fecha anterior en la que se realizó un censo. Si se usa este número, o bien uno derivado de una muestra dirigida precisamente a estimar las N_h, ocurrirá que, en las estimaciones, en lugar de las ponderaciones exactas p_h se utilizarán las p_h' estimadas, de modo que se tendrá

$$\bar{y}'_e = \sum_1^k p'_h \bar{y}_h \qquad (1)$$

en lugar de

$$\bar{y}_e = \sum_1^k p_h \bar{y}_h \qquad (2)$$

como estimación de la media de la población.

Es evidente que la (1) es una estimación "viciada", siendo el "error medio" (E.M.)

. . .

M.t.p.

$$E\left[\sum_{1}^{k} (p_h - p'_h)\,\bar{y}_h\right] = \sum_{1}^{k} (p_h - p'_h)\,\mu_h \qquad (3)$$

Si indicamos con μ el valor medio de la población, y es

$$E\left[\sum_{1}^{k} p'_h\,\bar{y}_h\right] = \mu'$$

puesto que

$$E\,(\bar{y}'_e - \mu)^2 = E\,(\bar{y}'_e - \mu')^2 + (\mu' - \mu)^2$$

se vé que la pérdida de precisión en la estimación depende de la magnitud $(\mu' - \mu)^2$ o sea de

$$\left\{\sum_{1}^{k} (p'_h - p_h)\,\mu_h\right\}^2$$

que es independiente del tamaño n de la muestra, de manera que la precisión no se aumenta acrecentando n, pudiendo ciertamente ocurrir que para una cierta extensión de la muestra, el muestreo simple al azar permite una estimación más precisa que el muestreo estratificado.

Lo que antecede no ofrece medio alguno para decidir, en un determinado caso, si se ha de estratificar o no cuando para la estimación se han de usar ponderaciones que se saben erróneas, pues no se puede predecir la magnitud del "error".

Una salida la ofrece el método de "muestreo doble" de Neyman por el cual las ponderaciones se estiman a partir de una muestra simple al azar preliminar, y la característica que se estudia, se estima en base a una muestra extraída de la muestra preliminar.

Se demuestra que si los pesos p'_h se estiman tomando una muestra simple al azar grande de extensión n' y la característica que se estudia en base a una muestra estratificada de extensión total $n < n'$ entonces la variancia de la estimación se incrementa aproximadamente en

$$\frac{\sum_{1}^{k} p'_h\,(\mu_h - \mu)^2}{n'}$$

si se supone que las tasas de muestreo son pequeñas.

. . .

M.t.p.

C A P I T U L O I I I

ESTIMACION POR COCIENTE

1.- En los capítulos anteriores, las estimaciones de la media o del total de la población se construyeron a partir de "la media por elemento" de la muestra, ya fuera éste una muestra simple al azar, ya una muestra análoga extraída de cada uno de los estratos en que la población se había descompuesto.

Nos ocuparemos ahora de un procedimiento de estimación mediante el cual, por la utilización de alguna información complementaria apropiada acerca de las características de los elementos de la población, es posible -en ciertas condiciones- obtener estimaciones más precisas que las derivadas de la media de la muestra.

Este método es el de la "estimación por cociente" y su más ventajosa aplicación supone que para cada unidad incluída en la muestra, aparte del valor y_i que en ella toma la característica que se estudia, se conoce el valor x_i de otra, y tambien el total de esta última para toda la población.

Es evidente, intuitivamente, que para que la información complementaria ofrecida por la x_i y su total en la población sea útil como ayuda para hacer una estimación más precisa que la que sin ella se obtendría, los valores de la variable complementaria en la población deben estar más o menos estrechamente correlacionados (positivamente) con los de la variable que interesa en la investigación.

2.- NOTACION Y DEFINICIONES

Sean

$$a_1 \quad a_2 \quad \ldots\ldots \quad a_N$$

los valores que la característica que se estudia (y que llamaremos "la variable a") toma para cada una de las N unidades de la población; y sean

$$b_1 \quad b_2 \quad \ldots\ldots \quad b_N$$

los valores de otra característica ("la viariable b") para las mismas unidades.

Por definición es

$$A = \sum_{1}^{N} a_i = N\bar{A} = N\mu_a$$

$$B = \sum_{1}^{N} b_1 = N\bar{B} = N\mu_b$$

Ya hemos adelantado más arriba que el valor de B se supone conocido.

La relación entre los totales (o las medias) de los valores a_i y b_i, es:

$$R = A / B = \mu_a / \mu_b$$

Si se toma de la población una muestra simple al azar de extensión n obteniéndose:

$$y_1 \quad y_2 \quad \ldots\ldots \quad y_n$$

y para cada una de las unidades incluídas en la muestra se evalúa el correspondiente valor de la variable complementaria, se tendrán los n valores.

$$x_1 \quad x_2 \ldots\ldots\ldots \quad x_n$$

Los correspondientes totales de la muestra serán

$$\sum_{1}^{n} y_i = y = n\bar{y}$$

$$\sum_{1}^{n} x_i = x = n\bar{x}$$

El cociente

$$\hat{R} = \frac{\sum y_i}{\sum x_i} = \frac{y}{x} = \frac{\bar{y}}{\bar{x}}$$

se usará como estimación dol R de la población.

Siendo el total B conocido,

$$\hat{A}_R = \frac{y}{x} B = \frac{\bar{y}}{\bar{x}} B = \hat{R}B \qquad (1)$$

es la "estimación por cociente" del total A de la población.

La estimación de la media de la población será:

$$\hat{\mu}_a = \hat{R}\,\bar{B}$$

Ejemplo 1.- Sea la "variable a" la extensión total dedicada al cultivo de un cierto cereal, y la "variable b" el área total de N explotaciones agrícolas. Si se extrae una muestra de n explotaciones, y_i y x_i serán los valores respectivos en las unidades en ella incluídos. Supuesta conocida el área total B de las N explotaciones, la (1) dará el área total estimada dedicada al cultivo de dicho cereal en la población.

Ejemplo 2.- En el ejemplo anterior, la "variable b" podría ser el área dedicada al cultivo del mismo cereal en alguna fecha anterior (p.e. en el último censo) y $\hat{R}$ daría entonces el número de hectáreas cultivadas hoy por cada una cultivada al tiempo del último censo.

Muchos otros ejemplos de aplicación de la "estimación por cociente" pueden encontrarse en el libro de E. Denning: "Some theory of sampling". pag. 170.

3.- PROPIEDADES DE LA ESTIMACION

Es evidente que la estimacion $\hat{A}_R$ (o $\hat{R}$, o $\hat{\mu}_a$) es "consistento".

Ella es, en general, "viciada", pero el error sistemático de estimación es pequeño comparado con el error standard de estimación apenas la extensión de la muestra sea relativamente grande (p.e. > 30). Esta "desventaja" con respecto a la estimación basada en la media por elemento, queda más que compensada por su mayor precisión.

Como en el cociente que expresa $\hat{R}$, tanto el numerador como el denominador son variables aleatorias, no ha sido posible dar una fórmula exacta de la variancia de la estimación valedera en general para cualquier población. Puede, si, darse una expresión aproximada, cuyo error tiende a cero con n creciente.

La diferencia

$$\hat{A}_R - A = \frac{\bar{y}}{\bar{x}}\,B - A$$

puede escribirse:

$$B\left(\frac{\bar{y}}{\bar{x}} - \frac{A}{B}\right) = \frac{N\bar{B}b}{\bar{x}}\,(\bar{y} - R\,\bar{x})$$

Ahora, si la muestra es grande, $\bar{x}$ será próxima a $\bar{B}_b$, y el

cociente $\overline{B}_{b/\overline{X}}$ cercano a 1.

Aceptada esta aproximación, se tiene:

$$\hat{A}_R - A \doteq N(\overline{y} - R\overline{x})$$

de donde se sigue

$$E(\hat{A}_R - A) \doteq N\,E(\overline{y} - R\overline{x}) = 0$$

es decir, tomamos a $\hat{A}_R$ como siendo una estimación "no viciada" del total A de la población, y podemos escribir:

$$V(\hat{A}_R) = N^2\, V(\overline{y} - R\overline{x}) \qquad (2)$$

Ahora bien:

$$\overline{y} - R\overline{x} = \frac{1}{n}\sum_{1}^{n}(y_i - Rx_i) = \frac{1}{n}\sum_{1}^{n} u_i = \overline{u}$$

(habiendo puesto $u_i = y_i - Rx_i$) de manera que:

$$V(\hat{A}_R) = N^2 V(\overline{u})$$

Pero

$$v(\overline{u}) = \frac{N-n}{N_n}\cdot\frac{1}{N-1}\sum_{1}^{N}(u_i - \overline{U})^2 \qquad (3)$$

donde $\overline{U}$ es la media de la población de valores

$$U_i = a_i - Rb_i\ ,$$

es decir:

$$\overline{U} = \frac{1}{N}\sum_{1}^{N} U_i = \frac{1}{N}\sum_{1}^{N}(a_i - Rb_i) = \overline{A}_a - R\overline{B}_b = 0$$

pues $R = a\,\bar{A}/\bar{B}_b$ por definición. La (3) es pues:

$$V(\bar{u}) = \frac{N-n}{Nn} \cdot \frac{1}{N-1} \sum_1^N U_i^2$$

$$= \frac{N-n}{Nn} \cdot \frac{1}{N-1} \sum_1^N (a_i - Rb_i)^2$$

y por lo tanto queda, como expresión aproximada de la variancia de $\hat{A}_R$:

$$V(\hat{A}_R) = \frac{N}{n} \cdot \frac{N-n}{N-1} \sum_1^N (a_i - Rb_i)^2 \qquad (4)$$

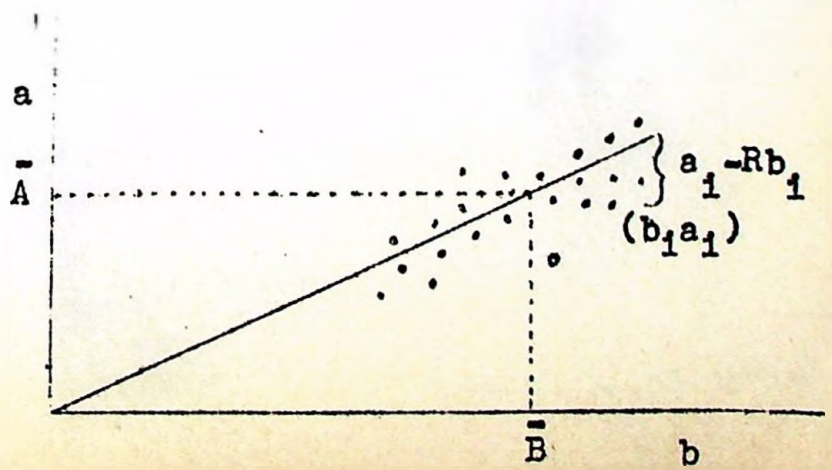
Se sigue de aquí, sin dificultad, que

$$V(\hat{\mu}_a) = \frac{1}{n} \frac{N-n}{N(N-1)} \sum_1^N (a_i - Rb_i)^2 \qquad (5)$$

$$V(\hat{R}) = \frac{N}{n} \cdot \frac{N-n}{N-1} \frac{1}{B^2} \sum_1^N (a_i - Rb_i)^2 \qquad (6)$$

Si, usando un sistema de ejes rectangulares, se representa cada elemento de la población por un punto que tiene coordenadas $(b_i a_i)$ la recta que pasa por el origen de coordenadas y por el punto $(\bar{B}, \bar{A})$, tiene una pendiente igual a R (ver fig.).

Las variancias (4), (5) y (6), escritas más arriba dependen de:



$$\frac{1}{N-1}\sum_{1}^{N}(a_i - Rb_i)^2 \qquad (a)$$

que no es sino el cuadrado medio de los desvíos verticales de los distintos puntos con respecto a la recta de pendiente R. Merece observarse que la (a) desempeña un papel semejante al de S^2 en el caso del muestreo simple al azar con estimación basada en la media por elemento.

Observando que la (a) puede escribirse

$$\frac{1}{N-1}\sum_{1}^{N} b_i^2(r_i - R)^2$$

se ve que las variancias (4), (5) y (6), pueden expresarse también en términos de los cuadrados de las desviaciones con respecto a R de los cocientes individuales $r_i = a_i/b_i$.

Otra forma de expresión de las anteriores variancias que, en tanto que pone de manifiesto más claramente los parámetros de las poblaciones constituídas por los valores a_i y b_i, de los que dependen, resulta más sugestiva, se obtiene por el siguiente camino.

Podemos escribir:

$$\sum_{1}^{N}(a_i - Rb_i)^2 = \sum_{1}^{N}(a_i - \bar{A} + \bar{A}_a - Rb_i)^2 = \sum_{1}^{N}\left[a_i - \bar{A}_a - R(b_i - \bar{B}_b)\right]^2 =$$

$$= \sum_{1}^{N}(a_i - \bar{A}_a)^2 + R^2\sum_{1}^{N}(b_i - \bar{B}_b)^2 - 2R\sum_{1}^{N}(a_i - \bar{A}_a)(b_i - \bar{B}_b)$$

Dividiendo por N-1, y recordando que :

$$\rho S_a S_b = \frac{1}{N-1}\sum_{1}^{N}(a_i - \bar{A}_a)(b_i - \bar{B}_b)$$

donde ρ indica el coeficiente de correlación entre la "variable a" y la "variable b", resulta :

$$V(\hat{A}_R) = \frac{N(N-n)}{n}\left[S_a^2 + R^2 S_b^2 - 2\rho R S_a S_b\right] \qquad (7)$$

de donde, sacando μ_a^2 como factor común del paréntesis, se obtiene: (.)

$$\frac{1}{N-1}\sum_{1}^{N}(a_i - Rb_i)^2 \qquad (a)$$

que no es sino el cuadrado medio de los desvíos verticales de los distintos puntos con respecto a la recta de pendiente R. Merece observarse que la (a) desempeña un papel semejante al de S^2 en el caso del muestreo simple al azar con estimación basada en la media por elemento.

Observando que la (a) puede escribirse

$$\frac{1}{N-1}\sum_{1}^{N} b_i^2(r_i - R)^2$$

se ve que las variancias (4), (5) y (6), pueden expresarse también en términos de los cuadrados de las desviaciones con respecto a R de los cocientes individuales $r_i = a_i/b_i$.

Otra forma de expresión de las anteriores variancias que, en tanto que pone de manifiesto más claramente los parámetros de las poblaciones constituídas por los valores a_i y b_i, de los que dependen, resulta más sugestiva, se obtiene por el siguiente camino.

Podemos escribir:

$$\sum_{1}^{N}(a_i - Rb_i)^2 = \sum_{1}^{N}(a_i - \bar{A} + \bar{A}_a - Rb_i)^2 = \sum_{1}^{N}\left[a_i - \bar{A}_a - R(b_i - \bar{B}_b)\right]^2 =$$

$$= \sum_{1}^{N}(a_i - \bar{A}_a)^2 + R^2\sum_{1}^{N}(b_i - \bar{B}_b)^2 - 2R\sum_{1}^{N}(a_i - \bar{A}_a)(b_i - \bar{B}_b)$$

Dividiendo por N-1, y recordando que :

$$\rho S_a S_b = \frac{1}{N-1}\sum_{1}^{N}(a_i - \bar{A}_a)(b_i - \bar{B}_b)$$

donde ρ indica el coeficiente de correlación entre la "variable a" y la "variable b", resulta :

$$V(\hat{A}_R) = \frac{N(N-n)}{n}\left[S_a^2 + R^2 S_b^2 - 2\rho R S_a S_b\right] \qquad (7)$$

de donde, sacando μ_a^2 como factor común del paréntesis, se obtiene: (.)

$$V(\hat{A}_R) = \frac{N(N-n)}{n} \mu_a^2 \left[C_a^2 + C_b^2 - 2\rho C_a C_b \right]$$

o bién, puesto que $N\mu_a^2 = A^2/N$, y poniendo $\rho C_a C_b = C_{ab}$:

$$V(\hat{A}_R) = \frac{1}{n} \cdot \frac{N-n}{N} A^2 \left[C_a^2 + C_b^2 - 2 C_{ab} \right] \qquad (8)$$

con lo que se obtiene la variancia de la estimación del total en función de los cuadrados de los coeficientes de variación de las "variables a y b" y del coeficiente de covariación C_{ab} de las mismas.

Las expresiones para $V(\hat{\mu}_a)$ y $V(\hat{R})$ son :

$$V(\hat{\mu}_a) = \frac{1}{n} \cdot \frac{N-n}{N} \bar{A}_a^2 \left[C_a^2 + C_b^2 - 2 C_{ab} \right] \qquad (9)$$

$$V(\hat{R}) = \frac{1}{n} \frac{N-n}{N} R^2 \left[C_a^2 + C_b^2 - 2 C_{ab} \right] \qquad (10)$$

(.).- Para mayor simplicidad y claridad en las fórmulas, usaremos la letra C para indicar el "coeficiente de variación", en lugar del símbolo C.V. que hemos usado en los capítulos anteriores.-

Evaluación aproximada del "vicio" de la estimación $\hat{A}_R$

Para obtener la expresión aproximada de la variancia de $\hat{A}_R$, se partió de la identidad

$$\hat{A}_R - A = \frac{N\bar{B}}{\bar{x}}(\bar{y} - R\,\bar{x})$$

para tomar luego $\bar{B}/\bar{x}$ como siendo igual a 1, teniendo entonces

$$N\,(\bar{y} - R\,\bar{x})$$

como primera aproximación de la diferencia $\hat{A}_R - A$, cuya esperanza matemática resultaba así ser igual a cero. Para poder tener una expresión aproximada del "vicio" de la estimación, es necesario llevar aquella aproximación un paso más adelante, lo que puede hacerse observando que es:

$$\frac{\bar{B}}{\bar{x}} = \frac{1}{1 - \frac{\bar{x} - \bar{B}}{\bar{B}}} = \left(1 + \frac{\bar{x} - \bar{B}}{\bar{B}}\right)^{-1}$$

Para un tamaño suficientemente grande de la muestra, será ciertamente $(\bar{x} - \bar{B})/\bar{B} < 1$, y el desarrollo del último término será convergente. Tomando solamente la parte lineal de dicho desarrollo:

$$1 - \frac{\bar{x} - \bar{B}}{\bar{B}}$$

se tiene como 2a. aproximación de $\hat{A}_R - A$:

$$N\,(\bar{y} - R\,\bar{x})\,\left(1 - \frac{\bar{x} - \bar{B}}{\bar{B}}\right) =$$

$$= N\left(\bar{y} - R\,\bar{x} - \frac{\bar{y}(\bar{x} - \bar{B})}{\bar{B}} + R\,\frac{\bar{x}(\bar{x} - \bar{B})}{\bar{B}}\right)$$

M t.p.

de modo que:

$$E\,(\hat{A}_R - A) = \frac{N}{\bar{B}}\left\{ R\,E\left[\bar{x}\,(\bar{x}-\bar{B})\right] - E\left[\bar{y}\,(\bar{x}-\bar{B})\right]\right\}$$

puesto que:

$$E\,(\bar{y} - R\,\bar{x}) = 0.$$

Ahora

$$E\left[\bar{x}\,(\bar{x}-\bar{B})\right] = E\,(\bar{x}-\bar{B})^2 = \left(\frac{1}{n} - \frac{1}{N}\right)\sigma_b^2$$

$$E\left[\bar{y}(\bar{x}-\bar{B})\right] = E\,(\bar{y}-\bar{A})(\bar{x}-\bar{B}) = \left(\frac{1}{n} - \frac{1}{N}\right)\rho\,\sigma_a\,\sigma_b$$

de modo que resulta finalmente:

$$E\,(\hat{A}_R - A) = E(\hat{A}_R) - A = \frac{N}{\bar{B}}\left(\frac{1}{n} - \frac{1}{N}\right)\left[R\sigma_b^2 - \rho\,\sigma_a\,\sigma_b\right] \qquad (11)$$

lo que muestra que el"vicio" de la estimación (que puede ser positivo o negativo) es tal que:

1º) Tiende a cero cuando el tamaño n de la muestra crece.

2º) Es nulo cuando

$$R\sigma_b^2 = \rho\,\sigma_a\,\sigma_b$$

o sea, cuando

$$\bar{A} = \rho\,\frac{\sigma_a}{\sigma_b}\,\bar{B}$$

lo que significa que la línea de regresión de los pares ($a_i b_i$) en la población es una recta que pasa por el origen.

La (11) puede también escribirse:

M.t.p.

$$E(\hat{A}_R) - A = N \frac{\bar{S}_b}{\bar{B}} \left(\frac{1}{n} - \frac{1}{N}\right) \left[R S_b - \rho S_a\right] \qquad (12)$$

$$= N C_{\bar{x}} \sqrt{\frac{1}{n} - \frac{1}{N}} \left[R S_b - \rho S_a\right]$$

puesto que:

$$\sqrt{\frac{1}{n} - \frac{1}{N}} \cdot \frac{\bar{S}_b}{\bar{B}} = C_{\bar{x}}$$

Por otra parte, el error standard de $\tilde{A}_R$ es:

$$\sigma_{\tilde{A}_R} = N \sqrt{\frac{1}{n} - \frac{1}{N}} \sqrt{(S_a^2 + R^2 S_b^2 - 2\rho R S_a S_b)} \qquad (13)$$

de modo que, comparando el valor absoluto de (12) con (13)

Se tiene:

$$\frac{\left| E(\tilde{A}_R) - A \right|}{\sigma_{\tilde{A}_R}} = C_{\bar{x}} \cdot \frac{\left| R S_b - \rho S_a \right|}{\sqrt{S_a^2 + R^2 S_b^2 - 2\rho R S_a S_b}}$$

El segundo factor del 2º término es menor o igual que 1, de manera que puede afirmarse que:

$$\frac{\left| E(\tilde{A}_R) - A \right|}{\sigma_{\tilde{A}_R}} \leq C_{\bar{x}}$$

de modo que la relación entre el vicio de la estimación y su error standard puede hacerse tan pequeña como se quiera tomando una muestra de tamaño tal que el coeficiente de variación de $\bar{x}$ sea suficientemente pequeño. Así, si el tamaño de la muestra es tal que el coeficiente de variación de $\bar{x}$ es 0-1, la relación del "vicio" de error standard de la estimación será $\leq$ 0-1, valor éste que puede considerarse como insignificante para el "vicio" de una estimación en la práctica.

M.t.p.

COMPARACION CON LA VARIANCIA DE LA ESTIMACION BASADA EN LA MEDIA POR ELEMENTO.

Sabemos que, si se toma una muestra simple al azar de extensión n de una población de N unidades y variancia S_a^2, la variancia de la estimación $\hat{A}$ es:

$$V(\hat{A}) = \frac{N(N-n)}{n} S_a^2$$

Restando de ésta la $V(\hat{A}_R)$ dada en (7), queda:

$$V(\hat{A}) - V(\hat{A}_R) = \frac{N(N-n)}{n} \left[2 \rho R S_a S_b - R^2 S_b^2 \right]$$

de donde se sigue que la diferencia será positiva toda vez que verifique:

$$2 \rho R S_a S_b > R^2 S_b^2$$

o sea:

$$\rho > \frac{1}{2} \frac{R S_b}{S_a}$$

o bien, puesto que $R = \mu_a / \mu_b$:

$$\rho > \frac{1}{2} \cdot \frac{C_b}{C_a} \qquad (14)$$

El precedente resultado muestra que la diferencia $V(\hat{A}) - V(\hat{A}_R)$ será tanto más grande cuando más alta sea la correlación entre la "variable a" y la "variable b" complementaria, siempre y cuando se verifique simultáneamente, que $C_b < 2 C_a$

TAMAÑO DE LA MUESTRA

Si en la fórmula (10) que da la variancia $V(\hat{R})$ pone-

mos

$$C_a^2 + C^2{}_b - 2\,C_{ab} = F,\ \text{queda:}$$

$$\frac{\hat{V}(R)}{R^2} = \left(\frac{1}{n} - \frac{1}{N}\right) F \qquad (15)$$

de donde resulta fácil obtener una estimación del tamaño n de la muestra requerida para alcanzar una precisión predeterminada. En efecto, de la (12) se obtiene:

$$n = \frac{F}{\frac{\hat{V}(\hat{R})}{R^2} + \frac{F}{N}}$$

o bien, teniendo en cuenta que $\hat{V}(\hat{R})/R^2 = C^2_{\hat{R}}$:

$$n = \frac{F/\,C^2_{\hat{R}}}{1 + \frac{1}{N} F/\,C^2_{\hat{R}}}$$

y, poniendo $F/\,C^2_{\hat{R}} = n_0$:

$$n = \frac{n_0}{1 + \frac{1}{N}\, n_0}$$

Para aplicar esta fórmula es necesario tener una estimación previa de F, es decir, de los coeficientes de variación de ambas variables y del coeficiente de correlación ρ.

ESTIMACION DE LA VARIANSIA A PARTIR DE LA MUESTRA

Homos visto que

$$V(\hat{A}_R) = \frac{N}{n} \cdot \frac{N-n}{N-1} \sum_{1}^{N} (a_i - Rb_i)^2$$

Para tener la estimación $v(\hat{A}_R)$ de $V(\hat{A}_R)$, se sustituye la sumatoria del segundo miembro por

$$\sum_{1}^{n} (y_i - \hat{R}\, x_i)^2$$

construída con los datos obtenidos en la muestra, Desarrollando el cuadrado del binomio se obtiene la siguiente expresión para $v(\hat{A}_R)$ que resulta particularmente apropiada para el cálculo:

$$v(\hat{A}_R) = \frac{N}{n} \cdot \frac{N-n}{N-1} \left[\sum_{1}^{n} y_i^2 + \hat{R}^2 \sum_{1}^{n} x_i^2 - 2\hat{R} \sum_{1}^{n} x_i\, y_i \right] \qquad (16)$$

La raíz cuadrada de $v(\hat{A}_R)$ dá el error standard de estimación $s(\hat{A}_R)$ que se utilizará para calcular el intervalo de confianza de la estimación.

ESTIMACION POR COCIENTE EN EL CASO DE PROPORCIONES

Consideremos una población que consta de N unidades de muestreo, cada una de las cuales encierra ó está constituída por un cierto número de elementos (p.e. familias que constan de individuos, manzanas de una ciudad en las que hay viviendas familiares, etc.) Supongamos que los elementos que comprende cada unidad de muestreo se clasifican en dos categorías: "C" y "no C".

Sean

$$b_1\ b_2\ \dots\ b_N$$

los elementos comprendidos en cada una de las N unidades, y

$$a_1\ a_2\ \dots\ a_N \qquad (a_i \leq b_i)$$

aquellos que entre los anteriores poseen la característica "C"

Si se toma una muestra simple al azar de unidades de extensión n,

$$\sum_1^n y_i \qquad\qquad \sum_1^n x_i$$

serán, respectivamente, el número total de elementos "C" y el número total de elementos "C" y "no C" en la muestra. El cociente

$$\sum_1^n y_i \ \Big/ \ \sum_1^n x_i = p$$

es la proporción de elementos "C" en la muestra, y es una estimación de la proporción correspondiente

$$\sum_1^N a_i \ \Big/ \ \sum_1^N b_i = P$$

en la población

Se tiene aquí un caso de "estimación por cociente", desempeñando P el mismo papel que R y p el mismo que $\hat{R}$

La variancia de la estimación p, viene dada por:

$$V(p) = \frac{N}{n} \cdot \frac{N-n}{N-1} \cdot \frac{1}{B^2} \sum_{1}^{N} (a_i - Pb_i)^2$$

donde $B = \sum_{1}^{N} b_i$. La estimación v(p) de V(p) es:

$$v(p) = \frac{N}{n} \cdot \frac{N-n}{N-1} \cdot \frac{1}{B^2} \sum_{1}^{n} (y_i - px_i)^2$$

o bien, si se desconoce $B = N \mu_b$

$$v(p) = \frac{N-n}{Nn(n-1)\bar{x}^2} \sum_{1}^{n} (y_i - px_i)^2$$

El esquema de muestreo que hemos supuesto para el anterior desarrollo, en el que constando la población de unidades constituídas por elementos, la muestra selecciona unidades cuyos elementos se enumeran totalmente, es un ejemplo de lo que se denomina "muestreo por grupos" (6 "por racimos", 6 "por conglomerados") del que nos ocuparemos más adelante.

El punto importante a notar aquí, es la diferencia entre la fórmula de la variancia que corresponde a este esquema y la que correspondería utilizar en el caso de que se hubiera tomado una muestra simple al azar de elementos de la población de extensión igual al número de elementos obtenidos en el muestreo por grupos.

En cualquiera de los dos casos, la estimación de P estaría dada por la misma expresión; la variancia será en cambio muy diferente, salvo el caso en que los diversos grupos sean ellos mismos muestras al azar de una superpoblación de elementos.

ESTIMACION POR COCIENTE EN EL MUESTREO ESTRATIFICADO

Recordamos que, en el muestreo estratificado, la estimación del total en la población de la característica objeto de estudio, se obtenía haciendo la suma, para todos los estratos, de las correspondientes estimaciones, (para cada uno de ellos) basadas en la media por elemento en la muestra, es decir:

$$\hat{A} = \sum_{h=1}^{k} N_h \bar{y}_h$$

Analogamente, un procedimiento para tener la estimación por cociente del total A de la población, consiste en hacer la estimación por cociente del total en cada estrato y sumar los resultados, de modo tal que

$$\hat{A}_{RS} = \sum_{h=1}^{k} \frac{y_h}{x_h} B_h = \sum_{h=1}^{k} \frac{\bar{y}_h}{\bar{x}_h} B_h$$

donde B_h es el total de la característica b en el estrato h, que se supone conocido.

El segundo subíndice en $\hat{A}_{RS}$ expresa que la estimación por cociente se ha basado en una estimación del mismo tipo por cada estrato separadamente.

No ofrece ninguna dificultad obtener la variancia $V(\hat{A}_R)$ en este caso. Ella viene dada por:

$$V(\hat{A}_{RS}) = \sum_{h=1}^{k} \frac{N_h(N_h - n_h)}{n_h} \left(S_{ah}^2 + R_h^2 S_{bh}^2 - 2 R_h \rho_h S_{ah} S_{bh} \right)$$

indicando en todos los casos el subíndice h que el símbolo al que acompaña, se refiere al valor que el mismo expresa para el h-ésimo estrato.

Dijimos más arriba que la estimación por cociente era "viciada" y que la fórmula para la variancia de la estimación era aproximada. Estos dos hechos merecen ser tenidos muy especialmente en cuenta en este caso. En primer término, la fórmula de la variancia dada más arriba requiere para su validez que el tamaño de la muestra extraída de cada estrato sea suficientemente grande como para que sea válida la aplicación de la $V(\hat{A}_{hR})$.

Otra fórmula de estimación por cociente es la que denominaremos de "estimación por cociente total", en la que el factor que multiplica al total B de la variable complementaria b en toda la población, es el cociente del total de la "variable a" y el de la "variable b" en la muestra, 6, en otros términos, el cociente de las estimaciones de los totales A y B obtenidos a partir de la muestra estratificada.

Así, con los datos de la muestra se calculan

$$\hat{A} = \sum_{h=1}^{k} N_h \bar{y}_h \qquad \hat{B} = \sum_{h=1}^{k} N_h \bar{x}_h$$

y con estos:

$$\hat{A}_{RT} = \frac{\sum N_h \bar{y}_h}{\sum N_h \bar{x}_h} B = \frac{\bar{y}_e}{\bar{x}_e} B$$

(usamos el segundo subíndice T para expresar que se trata de una estimación por cociente total).

Usando esta fórmula para la estimación, no se requiere el conocimiento del total B_h de la "variable b" en cada estrato.

La derivación de la fórmula aproximada de la variancia de la estimación no ofrece, tampoco en este caso, ninguna dificultad. En efecto, siguiendo un camino análogo al que nos llevó a obtener al comienzo de este capítulo la fórmula de $V(\hat{A}_R)$, se llega a tener

$$V(\hat{A}_{RT}) = \sum_{h=1}^{k} \frac{N_h (N_h - n_h)}{n_h} \left[S_{ah}^2 + R^2 S_{bh}^2 - 2R \rho_h S_{ah} S_{bh} \right]$$

La validez de esta fórmula requiere que el tamaño total n de la muestra sea suficientemente grande. Si se compara $V(\hat{A}_{RS})$ con $V(\hat{A}_{RT})$ se concluye que, en general, cuando $V(\hat{A}_{RS})$ es aplicable por ser suficientemente grandes las n_h, $\hat{A}_{RS}$ es más precisa que $\hat{A}_{RT}$. Por otra parte, si el tamaño de la muestra en cada estrato es relativamente pequeño, será preferible la estimación $\hat{A}_{RT}$. Estas no constituyen "reglas" definitivas, y en este caso, como en otros muchos la decisión dependerá de la fisonomía del problema práctico que se debe encarar.

CAPITULO IV

MUESTREO SIMPLE GRUPOS (I)

En los esquemas de muestreo que hemos considerado hasta ahora, cada uno de los elementos o unidades de la población era también la unidad de muestreo. Por la operación del muestreo se seleccionaban elementos de la población en los que se evaluaba o medía la o las características objeto de la investigación. Las unidades que constituían la población podían ser simples o complejas (p.e.personas individuales o familias) y su definición como tales dependía enteramente del propósito de la investigación (p.e. los elementos de la población serán las personas individuales, si el propósito es investigar el ingreso medio por persona; serán las familias (grupos de personas) si el objeto es investigar la distribución de los ingresos o gastos familiares), y de ingún modo estaba determinada por el esquema de muestreo.

En el muestreo "por grupos" (o "conglomerados", o "racimos") tambien son las características de las unidades o elementos de una población las que interesan, pero, para los fines de la operación de muestreo, ellas se agrupan para constituir "unidades primarias de muestreo" cuyas características no hacen al objeto del estudio, pero sí al plan de muestreo. Por ejemplo, para una investigación en la que interesan ciertos características de las explotaciones agrícolas de una región, para los fines del muestreo, se divide dicha región en áreas geográficas cada una de las cuales encierra varias explotaciones, y esto de modo tal que cada explotación de la región se asocia una y solo una de las áreas en que se la ha subdividido. Esas áreas, que encierran un grupo de explotaciones, son las "unidades primarias de muestreo". En los esquemas que vamos a considerar, se llega a los elementos de la población, por intermedio de las unidades primarias que los agrupan.

En lalgunos casos, tomada una muestra simple al azar de unidades primarias, o las características que son el objeto del estudio, se evalúan para todos los elementos que forman parte de las unidades primarias que aparecen incluídas en la muestra. Este es el esquema de "muestreo simple por grupos, en una etapa". En otros casos, cada unidad primaria incluída en la muestra, es ella misma objeto de muestreo, tomándose de ella una muestra simple al azar de los elementos que comprende. Este es el esquema de "muestreo simple por grupos, en dos etapas" o de "submuestreo". Es claro que no hay nada que limite, al menos teóricamente, el número de las etapas sucesivas de submuestreo que llevan de las unidades primarias en las que inicialmente se agrupan los elementos de la población, a los elementos mismos portadores de las caracteríasticas que con el objeto de la investigación.

Varias son las dificulatades de orden práctico y económico que limitarían el campo de aplicación de la técnica de las muestras que se resuelven con los esquemas mencionados, los que ofrecen métodos científicos de muestreo aplicables en casos en que los esquemas simples desarrollados en los capítulos anteriores resultarían, si no impracticables, prohibitivamente onerosos.

Así, p.e., es poco menos que imposible extraer una muestra simple al azar de personas de la población de un país, o aún de un ciudad, no se dispone de una lista completa y actualizada de todos los individuos que la componen. En cambio, generalmente puede disponerse, o es posible construir sin gran dificultad, una lista de "radios", "manzanas" u otras áreas que pueden constituir unidades primarias de muestreo apropiadas.

Por otra parte, aún cuando se disponga de un padrón que registra a todos los elementos de la población, la captación de la información requerida de elementos dispersos sobre una extensa área, (p.e. explotaciones agrícolas o viviendas elegidas al azar en el área de una provincia o de una ciudad) resultaría antieconómica considerada la incidencia del factor "costo de desplazamiemto" en el costo total de la investigación. Si bien una muestra al azar de elementos de la población resultaría más eficiente desde el punto de vista del tamaño, debe tenerse en cuenta que no siempre es esta la componente preponderante en el costo y que, una muestra de tamaño algo mayor de elementos agrupados en conglomerados (p.e. explotaciones agrícolas contiguas, viviendas en una manzana, etc.) distribuídas en área total, puede ofrecer estimaciones de la precisión requerida, a un costo total notablemente menor.

Cuando los grupos o conglomerados de elementos que constituirán las unidades primarias de muestreo, se forman asociando unívocamente cada unidad de la población o una y una sola de las áreas parciales en que se subdivide el área que encierra la totalidad de dichos elementos, el esquema que aquí estudiamos es el de "muestreo por áreas", reservándose la denominación genérica de "muestreo por grupos" a aquellos casos en que el medio por el cual se determina la unidad primaria no es un área.

En lo que sigue nos ocuparemos del caso más simple del muestreo en una sola etapa de una población costituída por N unidades primarias o grupos cada una de las cuales contiene: 1º) un mismo número de elementos, 2º) diferente número de elementos de la población.

1º) MUESTREO SIMPLE POR GRUPOS DE UNA ETAPA - UNIDADES PRIMARIA DE MUESTREO ONTENIENDO TODOS EL MISMO NUMERO M DE ELEMENTOS

NOTACION: sea:

a_{ij} ($i = 1 \ldots N$, $j = 1 \ldots M$) el valor que la característica que se estudia toma en el j-ésimo elemento del i -é-simo grupo.

$$\bar{a}_i = \frac{1}{M} \sum_{1}^{M} a_{ij}$$ el valor medio por elemento en el i-ésimo grupo

$$\bar{a} = \frac{1}{N} \sum_{1}^{N} \bar{a}_i$$ la media aritmética de los valores medios de los N grupos.

Puesto que todos los grupos constan del mismo número de elementos, la media $\bar{a}$ sera tambien la media por elemento en la población. En efecto:

$$\bar{a} = \frac{1}{N} \sum_{1}^{N} \bar{a}_i = \frac{1}{N} \sum_{1}^{N} \left(\frac{1}{M} \sum_{1}^{M} a_{ij} \right) = \frac{1}{NM} \sum_{1}^{N} \sum_{1}^{M} a_{ij}$$

En la figura de la página siguiente se da una representación gráfica que puede ayudar a interpretar estos conceptos y los desarrollos que siguen.

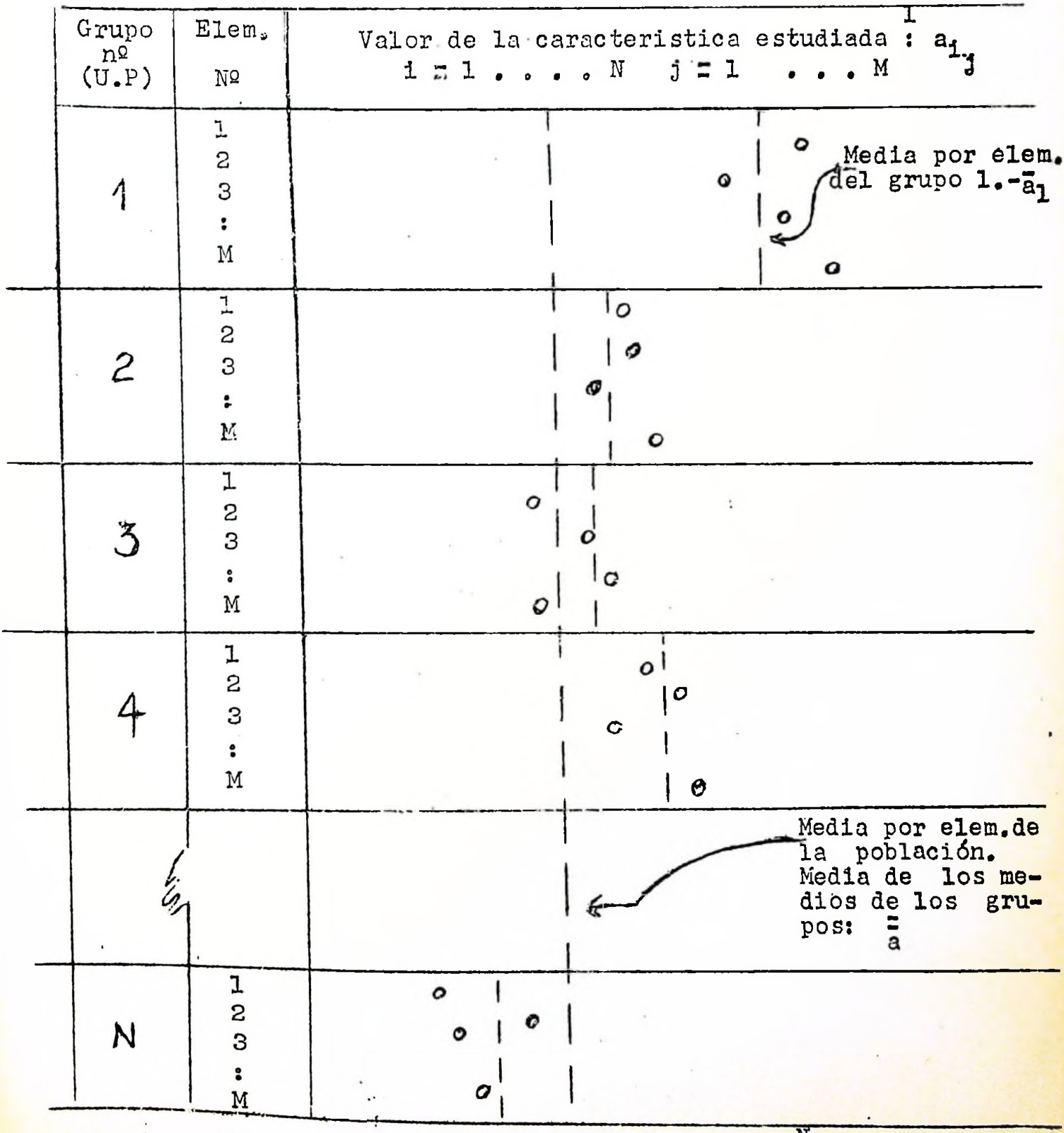
Supongamos que se toma una muestra simple al azar de n grupos. Puesto que el muestreo se realiza en una sola etapa, la característica que se estudia se evalúa en todos los M elementos que constituyen cada grupo incluído en la muestra, de modo que a cada grupo en la muestra le corresponde un valor bien definido de la media o el total de la característica en cuestión, o del porcentaje de elementos que poseen un cierto atributo. Si indicamos con $\bar{y}_h$ (h = 1 ... n) la media por elemento en el h-ésimo grupo incluído en la muestra, se tendrá que la misma ofrece n valores

$$\bar{y}_1 \quad \bar{y}_2 \quad \dots \quad \bar{y}_n$$

que constituye una muestra simple al azar de extensión n extraída de una población de N unidades cuyos elementos tienen valores.

POBLACION DE N.M. ELEMENTOS FORMADA POR

N UNIDADES PRIMARIAS DE M ELEMENTOS



$$\bar{a}_i = \frac{1}{M} \sum_{j=1}^{M} a_{ij} \qquad \bar{\bar{a}} = \frac{1}{NM} \sum_{i} \sum_{j} a_{ij} = \frac{1}{N} \sum_{i=1}^{N} \bar{a}_i$$

$$\bar{a}_1 \quad \bar{a}_2 \cdots \bar{a}_N$$

Lo mismo puede decirse para el caso de totales o porcentajes.

ESTIMACIONES Y SU VARIANCIA

Es evidente que la media aritmética de los valores medios de los grupos que integran la muestra:

$$\bar{\bar{y}} = \frac{1}{n} \sum_{1}^{n} \bar{y}_h \qquad (1)$$

constituye una estimación "consistente" de la media $\bar{\bar{a}}$.- Ella es también "no viciada" pues:

$$E(\bar{\bar{y}}) = \frac{1}{n} \sum_{1}^{n} E(\bar{y}_h) = \bar{\bar{a}}$$

Cuando, pues, todos los grupos están integrados por el mismo número de elementos, y el muestreo se realiza en una sola etapa la obtención de las estimaciones resulta particularmente simple, puesto que, como ya hemos indicado más arriba, el problema puede interpretarse como siendo el del muestreo simple al azar de una población de N unidades en las que una cierta característica toma los valores.

$$\bar{a}_1 \quad \bar{a}_2 \cdots \bar{a}_N$$

o bien

$$M\bar{a}_1 \quad M\bar{a}_2 \cdots M\bar{a}_N$$

ó

$$p_1 \quad p_2 \cdots p_N$$

según que se trate de estimar el valor medio, o el total o el porcentaje en la población.

Teniendo en cuenta esta interpretación resulta inmediato que la variancia de la estimación $\bar{\bar{y}}$ vendrá dada por:

$$V(\bar{\bar{y}}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 \qquad (2)$$

indicando s_b^2 la variancia de las $\bar{a}_i$ en la población, es decir:

$$s_b^2 = \frac{1}{N-1} \sum_1^N (\bar{a}_i - \bar{\bar{a}})^2$$

Ejemplo: Sea una población de 30 elementos que se han clasificado en 6 grupos de 5 elementos cada uno, y para los que una cierta característica toma los valores dados en el cuadro siguiente:

Grupo Nº	a_{ij}					Total	$\bar{a}_i$	$\bar{a}_i^2$
1	3	7	6	12	4	32	6.4	40.96
2	8	6	5	6	11	36	7.2	51.84
3	13	15	6	3	4	41	8.2	67.24
4	9	12	6	12	5	44	8.8	77.44
5	14	20	5	8	14	61	12.2	148.84
6	2	7	11	3	10	33	6.6	43.56
						247	$\bar{\bar{a}}$ 8.23	429.88

El número total de muestras posibles de extensión n = 3 de esta población de 6 grupos es 20. A continuación se dan los totales de cada una de las muestras posibles y la correspondiente estimación que ofrecen del total de la población:

Muestra Nº	Total de la muestra	Total estimado de la población	Muestra Nº	Total de la muestra	Total estimado de la población
1	109	218	11	121	242
2	112	224	12	138	276
3	129	258	13	110	220
4	101	202	14	141	282
5	117	234	15	113	226
6	134	268	16	130	260
7	106	212	17	146	292
8	137	274	18	118	236
9	109	218	19	135	270
10	126	252	20	138	276
					Total 4.940

El valor medio sobre todas las muestras posibles de las estimaciones del total de la población, coincide ciertamente con el total 247 que se estima

La variancia S^2_b es:

$$S^2_b = \frac{1}{N-1}\left\{\sum_1^N a_1^2 - Na^{=2}\right\} = \frac{1}{5}\left\{429.88-406.40\right\} = 4.696$$

de modo que la variancia de la estimación de la media es:

$$V(\bar{y}) = \left(\frac{1}{3} - \frac{1}{6}\right) 4.696 = .7.827$$

y la del total:

$$V(N\bar{M}\bar{y}) = \underline{704.43}$$

Supongamos ahora que, en lugar de una muestra de grupos, se hubiera tomado una muestra simple al azar de elementos de tamaño igual al número de elementos incluídos en el muestra de grupos. La variancia de la estimación del total de la población sería entonces:

$$V(N\bar{M}\bar{y}) = (NM)^2 \left(\frac{1}{nM} - \frac{1}{NM}\right)S^2$$

siendo ahora S^2 la variancia de los elementos en la población. Se tendría entonces:

$$V(N\bar{M}\bar{y}) = (30)^2\left(\frac{1}{15} - \frac{1}{30}\right)18.97 = 569.10$$

Se ve aquí que, en este caso, la variancia de la estimación del total basada en una muestra simple al azar de elementos, es menor que la de la estimación basada en una muestra de grupos que reunen el mismo número total de elementos.

EFICIENCIA RELATIVA DEL MUESTREO POR GRUPOS

La eficiencia relativa de las estimaciones obtenidas a partir de dos métodos de estimación o de muestreo diferentes cualesquiera, está dada por el cociente de las inversas de sus respectivas variancias, de modo que, siendo en el caso que se está considerando:

$$V(\bar{\bar{y}}) = \left(\frac{1}{n} - \frac{1}{N}\right) s_b^2 \qquad V(\bar{y}) = \left(\frac{1}{nM} - \frac{1}{NM}\right) s^2$$

se tendrá:

$$E.R. = \frac{1/V(\bar{\bar{y}})}{1/V(\bar{y})} = \frac{V(\bar{y})}{V(\bar{\bar{y}})} = \frac{s^2}{MS_b^2}$$

En el ejemplo de más arriba se tiene:

$$E.R. = \frac{17.22}{5x4.696} = 73.^3 \ \%$$

Partiendo de la identidad algebraica

$$\sum_1^N \sum_1^M (a_{ij} - \bar{\bar{a}})^2 = \sum_1^N \sum_1^M (a_{ij} - \bar{a}_i)^2 + M \sum_1^N (\bar{a}_i - \bar{\bar{a}})^2$$

puede construirse el siguiente cuadro de "análisis de la variancia" de la población dividida en grupos:

Fuente de variación	Sumas de cuadrados	Cuadrados medios
Entre grupos	$M \sum_1^N (\bar{a}_i - \bar{\bar{a}})^2$	$\frac{M}{N-1} \sum_1^N (a_i - \bar{\bar{a}})^2 = MS_b^2$
Dentro de los grupos	$\sum_1^N \sum_1^M (a_{ij} - \bar{a}_i)^2$	$\frac{1}{N(M-1)} \sum_1^N \sum_1^M (a_{ij} - \bar{a}_i)^2 = S_w^2$
Total	$\sum_1^N \sum_1^M (a_{ij} - \bar{\bar{a}})^2$	$\frac{1}{NM-1} \sum_1^N \sum_1^M (a_{ij} - \bar{\bar{a}})^2 = s^2$

Los "cuadrados medios" dados en la última columna se obtienen dividiendo las correspondientes "sumas de cuadrados" por sus respectivos "grados de libertad".

En el cuadro se ve que la eficiencia relativa del muestreo por grupos con respecto al muestreo simple al azar de elementos está dada por la relación del "cuadrado medio entre grupos" al "cuadrado medio total" (que no es sino el número que hemos llamado "varian cia S^2").

Supongamos tener NM elementos que se agruparán en N grupos de M elementos cada uno. Podemos imaginar que la operación consiste en tomar uno a uno los elementos para repartirlos en N recipientes de manera que en cada uno de estos haya el mismo número M de unidades. Es fácil ver que el número de ordenamientos diferentes que pueden obtenerse por el procedimiento indicado es:

$$K = \frac{(NM)!}{N!\,(M!)^N}$$

En efecto, sea K el número de ordenamientos diferentes; si se permutan entre sí los N grupos que constituyen cada uno de ellos y dentro de cada grupo se permutan entre sí los M elementos que en él hay, se obtendrán

$$K \times N! \times (M!)^N$$

ordenamientos de los NM elementos, los que no son otra cosa que las $(NM)!$ permutaciones de esos elementos.

Si para formar los N grupos se van eligiendo los elementos uno a uno al azar, el resultado es equivalente a la elección al azar de uno entre los

$$\frac{(NM)!}{N!\,(M!)^N}$$

agrupamientos posibles.

Siendo

$$S_b^2 = \frac{1}{N-1} \sum_1^N (\bar{a}_i - \bar{\bar{a}})^2$$

$$S_w^{-2} = \frac{1}{N(M-1)} \sum_1^N \sum_1^M (a_{ij} - \bar{a}_i)^2 = \frac{1}{N} \sum_1^N S_i^2$$

donde $$S_i^2 = \frac{1}{M-1} \sum_{1}^{M} (a_{ij} - \bar{a}_i)^2$$

se desea saber cuales son:

$$E\,(S_b^2) \quad y \quad E\,(\bar{S}_w^2)$$

Puesto que uno de los K agrupamientos contiene todos los elementos de la población, todos ellos tienen la misma media $\bar{\bar{a}}$, y entonces:

$$E\,(S_b^2) = \frac{1}{N-1} \left\{ \sum_{1}^{N} E\,(\bar{a}_i)^2 - N\bar{\bar{a}}^2 \right\}$$

y $$E\,(\bar{S}_w) = \frac{1}{N(M-1)} \left\{ \sum_{1}^{N} \sum_{1}^{M} E(a_{ij}^2) - M \sum_{1}^{N} E(\bar{a}_i^2) \right\}$$

Ahora $$E\,(\bar{a}_i^2) = E(\bar{a}_i - \bar{\bar{a}})^2 + \bar{\bar{a}}^2$$

$$E\,(a_{ij}^2) = E\,(a_{ij} - \bar{\bar{a}})^2 + \bar{\bar{a}}^2$$

y como $\bar{a}_i$ y a_{ij} son, respectivamente, las medias de muestras de tamaño M y 1 extraídas al azar de una población de NM elementos, resulta:

$$E\,(\bar{a}_i - \bar{\bar{a}})^2 = \left(\frac{1}{M} - \frac{1}{NM}\right) S^2$$

$$E(a_{ij} - \bar{\bar{a}})^2 = \left(1 - \frac{1}{NM}\right) S^2$$

teniéndose entonces:

$$E(S_b^2) = \frac{1}{N-1} \left\{ \sum_{1}^{N} \left[\bar{\bar{a}}^2 + \frac{1}{M} \left(1 - \frac{1}{N}\right) S^2 \right] - N\bar{\bar{a}}^2 \right\} =$$

$$= \frac{1}{N-1} \left\{ N\bar{\bar{a}}^2 + \frac{N}{M} \cdot \frac{N-1}{N} S^2 - N\bar{\bar{a}}^2 \right\} = S^2 / M$$

y $$E(\bar{S}_w^2) = \frac{1}{N\,(M-1)} \left\{ \sum_{1}^{N} \sum_{1}^{M} \left[\bar{\bar{a}}^2 + \left(1 - \frac{1}{NM}\right) S^2 \right] - M \sum_{1}^{N} \left[\frac{1}{M} \left(1 - \frac{1}{N}\right) S^2 + \bar{\bar{a}}^2 \right] \right\} =$$

$$= \frac{1}{N(M-1)} \left\{ NM\bar{a}^2 + (NM-1)\, S^2 - N\frac{N-1}{N} S^2 - NM\bar{a}^2 \right\} = S^2$$

Los anteriores resultados muestran que, si los grupos son muestras al azar de una población de NM elementos, se verifica que

$$E\,(MS_b^2) = E(\bar{S}_w^2) = S^2$$

de modo que entonces el muestreo por grupos tiene, en media, la misma eficiencia que el muestreo de elementos.

EVALUACION DE LA EFICIENCIA RELATIVA A PARTIR DE LOS RESULTADOS DE LA MUESTRA

Las consideraciones precedentes sirvieron para poner de manifiesto de qué características de la población dividida en grupos dependía la eficiencia relativa.- En la práctica no se conocen los valores S_b^2 y $\bar{S}_w^2$ y S^2 disponiéndose solamente de los correspondientes valores calculados en la muestra.- Si, pues, interesa evaluar la eficiencia relativa, será necesario ver como puede ello estimarse utilizando los resultados ofrecidos por la muestra.-

Esta evaluación puede hacerse construyendo un cuadro de "análisis de la variancia" para la muestra, semejante al construído más arriba para la población total.- Se tiene así:

Fuente de variación	Suma de cuadrados	Cuadrados medios
Entre grupos	$M\sum_{1}^{n}(\bar{y}_i-\bar{\bar{y}})^2$	$\frac{M}{n-1}\sum_{1}^{n}(y_i-\bar{\bar{y}})^2 = Ms_b^2$
Dentro de los grupos	$\sum_{1}^{n}\sum_{1}^{M}(y_{ij}-\bar{y}_i)^2$	$\frac{1}{n(M-1)}\sum_{1}^{n}\sum_{1}^{M}(y_{ij}-\bar{y}_i)^2 = \bar{s}_w^2$
Total	$\sum_{1}^{n}\sum_{1}^{M}(y_{ij}-\bar{\bar{y}})^2$	$\frac{1}{NM-1}\sum_{1}^{n}\sum_{1}^{M}(y_{ij}-\bar{\bar{y}})^2 = s^2$

Puesto que s_b^2 es la variancia de la muestra

$$\bar{y}_1 \quad \bar{y}_2 \cdots \bar{y}_n$$

obtenida por el muestreo simple el azar de la población de N unidades

$$\bar{a}_1 \quad \bar{a}_2 \cdots \bar{a}_n$$

será:

$$E(s_b^2) = S_b^2$$

A su vez

$$E(s_w^{-2}) = S_w^{-2}$$

puesto que

$$s_w^{-2} \frac{1}{N} s_i^2$$

y

$$s_1^2 \quad s_2^2 \quad \ldots \quad s_N^2$$

Por su parte, s^2 no es una estimación "no viciada" de S^2 puesto que los elementos y_{ij} a partir de los cuales se calcula s^2 no constituyen una muestra simple al azar de elementos de la población.- Es fácil, sin embargo, obtener una estimación "no viciada" de S^2, para lo cual basta observar que:

$$(NM-1)S^2 = (N-1)MS_b^2 + N(M-1)\bar{S}_w^2 \qquad (a)$$

de modo que, siendo la exoresión

$$(N-1)Ms_b^2 \quad N(M-1)s_w^{-2}$$

una estimación "no viciada" del segundo miembro de la (a) (por serlo s_b^2 y s_w^{-2} de S_b^2 y $\bar{S}_w^2$ respectivamente), se tendrá que

$$\frac{(N-1)Ms_b^2 \quad N(M-1)s_w^{-2}}{NM-1}$$

dará una estimación "no viciada" de la variancia S^2 de los elementos de la población.

Resulta de aquí que la eficiencia relativa estimada a partir de los datos obtenidos en la muestra, vendrá dada por:

$$\hat{E.R.} = \frac{1}{NM-1} \cdot \frac{(N-1)Ms_b^2 \quad N(M-1)s_w^{-2}}{Ms_b^2}$$

Si el número N de grupos es suficientemente grande como para que puedan tomarse $(N-1)/N = 1$, se tendrá como valor aproximado de la eficiencia relativa estimada:

M.t.p.

$$\hat{E}.R. = \frac{1}{M} + \left(\frac{M-1}{M}\right) \frac{s_w^{-2}}{Ms_b^2}$$

y si, M es grande, de modo que $(M-1)/M = 1$:

$$E.R. \quad \frac{s_w^{-2}}{Ms_b^2}$$

Método práctico para la ejecución de los cálculos requeridos para el análisis de la variancia.-

Para el "análisis de la variancia" de la muestra deben calcularse las siguientes sumas de cuadrados:

$$a) \quad M \sum_{1}^{n} (\bar{y}_i - \bar{\bar{y}})^2$$

$$b) \quad \sum_{1}^{n} \sum_{1}^{M} (y_{ij} - \bar{y}_i)^2$$

$$c) \quad \sum_{1}^{n} \sum_{1}^{M} (y_{ij} - \bar{\bar{y}})^2$$

Los cálculos se simplifican grandemente si se pone:

$$\bar{y}_i = \frac{1}{M} \sum_{1}^{M} y_{ij} = \frac{1}{M} T_i$$

$$\bar{\bar{y}} = \frac{1}{nM} \sum_{1}^{n} \sum_{1}^{M} y_{ij} = T$$

de modo que se tiene

$$M \sum_{1}^{n} (\bar{y}_i - \bar{\bar{y}})^2 = M \sum_{1}^{n} \left(\frac{1}{M} T_i - T\right)^2 = \frac{1}{M} \sum_{1}^{n} T_i^2 - \frac{1}{nM} T^2$$

M.t.p.

$$\sum_{1}^{n}\sum_{1}^{M}(y_{ij}-\bar{y}_i)^2 = \sum_{1}^{n}\sum_{1}^{M}(y_{ij}-\frac{1}{M}T_i)^2 = \sum_{1}^{n}\sum_{1}^{M}y_{ij}^2 - \frac{1}{M}\sum_{1}^{n}T_i^2 \qquad (1)$$

$$\sum_{1}^{n}\sum_{1}^{M}(y_{ij}-\bar{\bar{y}})^2 = \sum_{1}^{n}\sum_{1}^{M}(y_{ij}-T^2) = \sum_{1}^{n}\sum_{1}^{M}y_{ij}^2 - \frac{1}{nM}T^2$$

Resulta así que solo se requiere calcular

$$T_i = \sum_{1}^{M}y_{ij}, \quad T = \sum_{1}^{n}T_i \quad \sum_{1}^{n}\sum_{1}^{M}y_{ij}^2$$

para tener las (1) que, divididas por sus respectivos grados de libertad (n-1, n(M-1) y nM-1), dan los cuadrados medios requeridos.-

Ejemplo:- Una población ha sido subdividida en 120 grupos, cada uno de los cuales consta de 10 elementos. Se ha tomado una muestra simple al azar de 20 grupos, obteniendose los resultados que se registran en el cuadro siguiente.- Cual es la estimación de la media de la característica estudiada en la población y cual la eficiencia relativa de la misma comparada con la que se hubiera obtenido extrayendo de la población una muestra simple al azar de 200 elementos?.

Grupo Nº	Valores observados: y_{ij}										T_i	T_i^2
1	13.2	17.9	24.3	12.5	14.7	14.3	17.6	19.8	20.-	15.5	169.8	28.832.04
2	9.3	6.7	12.5	13.9	14.6	7.9	6.5	12.6	11.1	12.-	107.1	11.470,41
3	23.5	20.9	14.6	17.6	15.5	15.-	21.-	23.2	21.9	20.-	193.2	37.326.24
4	8.-	7.2	6.5	10.9	7.-	6.8	9.2	11.3	6.5	8.-	81.4	6.625.96
5	21.-	23.2	24.9	30.-	19.6	21.3	19.8	21.4	19.-	23.5	223.7	50.041.69
6	7.9	10.3	12.4	11.3	10.9	10.-	7.9	14.3	16.5	9.6	111.1	12.343.21
7	16.-	19.-	19.5	21.6	25.5	18.2	21.-	22.0	19.5	17.3	199.6	39.840,16
8	21.-	26.5	28.2	19.6	14.3	19.5	23.-	19.2	21.5	23.3	216.1	46.699.21
9	6.-	9.6	8.3	7.1	12.5	9.8	10.1	14.2	13.9	6.5	98.-	9.604.-
10	8.-	8.3	9.5	10.4	6.5	4.3	9.6	10.4	7.2	11.-	85.2	7.259.04
11	16.5	16.2	19.3	18.5	19.1	20.2	21.1	19.8	20.-	18.5	189.2	35.796.64
12	32.5	29.7	23.6	31.2	27.8	19.7	23.6	29.7	32.-	27.3	277.1	76.784.41
13	16.3	13.2	14.6	14.5	19.1	20.3	21.5	19.3	18.7	16.2	173.7	30.171.69
14	6.4	3.9	7.3	8.2	8.7	9.3	6.5	7.3	12.1	10.2	79.9	6.384.01
15	10.5	11.3	13.1	9.6	12.3	14.3	7.8	14.5	9.2	7.6	110.2	12.144.04
16	37.6	32.9	26.8	30.2	27.6	31.2	36.-	30.3	27.8	26.9	307.3	94.433.29
17	19.-	19.6	21.3	16.3	14.9	15.-	23.6	17.2	21.5	19.3	187.7	35.231.29
18	9.3	6.8	9.1	10.3	6.7	8.6	10.2	9.2	8.6	11.2	90.-	8.100.-
19	16.7	15.-	15.3	14.-	19.-	23.5	12.4	19.6	21.2	13.9	170.6	29.104.36
20	28.3	29.2	31.3	25.5	28.4	23.2	29.7	31.5	26.2	33.9	287.2	82.483.84
											3.358.1	660.675.53

Se tiene en este caso: N = 120, M = 10, n = 20

Siendo $\sum_{1}^{20} T_i$ 3358.1, se tendrá: $\bar{\bar{y}} = 16.79$

como valor estimado de la media de la población.-

Los valores que se requieren para el análisis de la varian
cia son:

M.t.p.

$$T = 3.358.1$$

$$T^2 = 11.276.835.61$$

$$\sum \sum y_{ij}^2 = 67.638.59$$

con los cuales se calculan las sumas de cuadrados

a) = 9.683.38

b) = 1.571.04

c) = 11.254.42

(En la práctica se calcularán solo 2 de estas, para tener la 3ra.)

Se tienen así todos los elementos para la construcción del cuadro de análisis de la variancia:

Fuente de variación	Suma de cuadrados	Grados de libertad	Cuadrados medios
Entre grupos	9.683.38	19	514.91 = Ms_b^2
Dentro de los grupos	1.571.04	180	8.73 = $\bar{s}_w^2$
Total	11.254.42	199	

Siendo N = 120 grande ($\frac{N}{N-1}$ = 991), se usará para calcular la eficiencia relativa, la fórmula:

$$\hat{E}.R. = \frac{1}{M} + \left(\frac{M-1}{M}\right) \cdot \frac{s_w^{-2}}{Ms_b^2}$$

teniéndose:

$$\hat{E}.R. = 11.5\ \%$$

lo que muestra que, dejando de lado consideraciones de costo u otras que puedan hacer aconsejable la utilización de grupos del tamaño elegido como unidades de muestreo, para atender solamente a la precisión de la estimación, es ciertamente preferible el muestreo por elementos Este, comparativamente pobre resultado, pone de manifiesto que en el proceso de manipulación de la poblacion para la formación de los grupos no se ha logrado hacer que ellos sean suficientemente heterogéneos internamente al mismo tiempo que semejantes entre sí, de modo que a una variabilidad grande dentro de los grupos se asocia una variabili-

dad pequeña entre grupos, que es lo que hace que la eficiencia relativa del muestreo por grupos crezca.-

M.t.p.

EVALUACION DE LA EFICACIA RELATIVA EN TERMINOS DEL COEFICIENTE DE CORRELACION "INTRA - CLASE"

Una más clara apreciación de los factores de los que depende la eficiencia relativa del muestreo por grupos en una etapa se logra expresándola en términos del coeficiente de correlación "Intra-clase" de los elementos que integran los diversos grupos.

Por definición, el coeficiente de correlación "Intra-clase" es:

$$\rho = \frac{E\,(a_{ij} - \bar{a})\,(a_{ik} - \bar{a})}{E\,(a_{ij} - \bar{a})^2} \qquad \begin{matrix} i = 1,2 \ldots N \\ j \neq k = 1,2 \ldots M \end{matrix} \qquad (1)$$

En el numerador, para cada valor del subindice i-que toma los valores de 1 a N =, los subindices j y k toman todos los valores de 1 a M, pero en cada producto es siempre $j \neq k$, de modo que, teniéndose N. grupos de M elementos cada uno, hay en el numerador NM (M-1) sumandos, siendo entonces

$$E(a_{ij} - \bar{a})\,(a_{ik} - \bar{a}) = \frac{1}{NM(M-1)} \sum_{i=1}^{N} \sum_{j \neq k = 1}^{M} (a_{ij} - \bar{a})\,(a_{ik} - \bar{a})$$

El numerador de la (1) puede escribirse:

$$E(a_{ij} - \bar{a}_i + \bar{a}_i - \bar{a})\,(a_{ik} - \bar{a}_i + \bar{a}_i - \bar{a}) = E(a_{ij} - \bar{a}_i)\,(a_{ik} - \bar{a}_i) + E(\bar{a}_i - \bar{a})^2 \qquad (2)$$

si se tiene en cuenta que:

$$E(a_{ij} - \bar{a}_i)\,(\bar{a}_i - \bar{a}) = E(a_{ik} - \bar{a}_i)\,(\bar{a}_i - \bar{a}) = 0$$

Para calcular la esperanza matemática que figura en el 1er. término del 2º miembro de la (2), consideremos en primer lugar que i es fijo, de modo que debe calcularse:

$$E(a_{ij} - \bar{a}_i)\,(a_{ik} - \bar{a}) = \frac{1}{M(M-1)} \sum_{j \neq k = 1}^{M} (a_{ij} - \bar{a}_i)\,(a_{ik} - \bar{a}_i) =$$

$$= \frac{1}{M(M-1)} \left\{ \left[\sum (a_{ij} - \bar{a}_i) \right]^2 - \sum (a_{ij} - \bar{a}_i)^2 \right\} = -\frac{S_i^2}{M}$$

Tomando ahora la esperanza matemática para i variando de 1 á N, queda:

$$E(a_{ij} - \bar{a}_i)\ (a_{ik} - \bar{a}_i) = -\frac{1}{M} E(S_i^2) = -\frac{1}{M} \cdot \frac{1}{N} \sum_1^N S_i^2 = -\frac{\bar{S}_w^2}{M}$$

Ahora, el 2º término del 2º miembro de la (2) y el denominador de la (1), son respectivamente:

$$E(\bar{a}_i - \bar{\bar{a}})^2 = \frac{N-1}{N} S_b^2$$

$$E(a_{ij} - \bar{\bar{a}})^2 = \frac{NM-1}{NM} S^2$$

de modo que queda finalmente:

$$\rho = \frac{M(N-1)S_b^2 - N\bar{S}_w^2}{(NM-1)S^2} \qquad (3)$$

De esto obtenemos:

$$(NM-1)S^2 \rho = M(N-1)S_b^2 - N\bar{S}_w^2 \qquad (4)$$

Ya hemos visto más arriba que:

$$(NM-1)S^2 = M(N-1)S_b^2 + N(M-1)\bar{S}_w^2 \qquad (5)$$

Multiplicando ambos miembros de la (4) por M-1 y sumando al resultado la (5), se sigue que:

$$S_b^2 = \frac{NM-1}{M(N-1)} \cdot \frac{S^2}{M}\left[1 + (M-1)\rho\right] \qquad (6)$$

A su vez, se obtiene tambien fácilmente de la (4) y (5) que:

$$\bar{S}_w^2 = \frac{NM-1}{NM} S^2 (1-\rho) \qquad (7)$$

Puesto que los primeros miembros de las (6) y (7) son esencialmente positivos, se vé inmediatamente que el coeficiente de correlación "intra-clase" verifica la desigualdad:

$$-\frac{1}{M-1} \leq \rho \leq 1$$

que es una propiedad que lo distingue del coeficiente de correlación "entre-clase" cuyo valor está en el intervalo $-1, +1$

De la (6) se sigue inmediatamente que:

$$\frac{S^2}{MS_b^2} = \frac{M(N-1)}{NM-1} \cdot \frac{1}{\left[1 + (M-1)\rho\right]} \qquad (8)$$

de modo que se tiene así expresada la eficiencia relativa en función de ρ.-

La (8) muestra que la E.R. decrece cuando P crece, es decir, cuanto más alto es el grado de inter-relación de los elementos que constituyen los diversos grupos, o, en otros términos cuando mayor es la "redundancia" de la información ofrecida por los valores de la característica estudiada en los individuos de cada grupo.

Como ya indicáramos anteriormente, cuando los grupos que constituirán las unidades de muestreo están formados por elementos contiguos de la población, tales como las explotaciones agrícolas encerradas en una cierta área rural, las viviendas ubicadas en una manzana o en un grupo de manzanas vecinas en una ciudad, etc., se encontrará que el coeficiente de correlación intra-clase para una gran variedad de características que pueden ser objeto de estudio es positivo, de modo que la eficiencia relativa será inferior a 100 %. Por otra parte, si el coeficiente de correlación intra-clase es negativo, la precisión del muestreo por grupos será mayor que la del muestreo simple al azar de elementos, y tanto más cuanto más próximo sea aquél a su valor mínimo $-1/(M-1)$.

En la práctica, el mayor o menor grado de libertad de que se dispone para la formación de los grupos al hacer el diseño de la muestra, depende del particular problema que se tiene entre manos. Debe tenerse presente que, formando grupos grandes se aumenta la variabilidad interna, pero también se pierde eficiencia. Por otra parte, un número mayor de grupos más pequeños, puede quizás resultar más eficiente, pero se pierde, en parte, algo de la ventaja de menor costo que aconseja la utilización de los grupos como unidades de muestreo en lugar de los elementos de la población. La norma básica a seguir es: grupos de la máxima heterogeneidad interna y de la máxima semejanza entre sí, y, en segundo término, el máximo número de grupos del menor tamaño posible que pueda lograrse teniendo en cuenta los medios de que se dispone para la ejecución de la operación de muestreo.

CAPITULO V

MUESTREO SIMPLE POR GRUPOS - (II)

SUB, MUESTREO EN 2 ETAPAS

Pasaremos ahora a considerar el muestreo por grupos en 2 etapas en el que, supuesto que se tienen N grupos o unidades primarias de muestreo, se llega a los elementos de la población que han de ofrecer la información para la confección de las estimaciones, mediante una doble operación de muestreo : 1º una muestra de unidades primarias, 2º una muestra de elmentos extraída de cada unidad obtenida en la primera muestra. Supondremos que en ambos pasos de obtiene la muestra mediante una operación de muestreo simple al azar y, que las N unidades primarias en las que se agrupan los elementos de la población constan todas del mismo número M de elementos. La operación de muestreo selecciona n unidades primarias, y de cada una de ellas m unidades secundarias, de modo que el tamaño total de la muestra es n.m.-

Sea a_{ij} el valor que la característica estudiada toma en el j-ésimo individuo de la i-ésima unidad primaria. Si es $\bar{\bar{a}}$ la media por elemento en la población de NM elementos y $\bar{a}_i$ la media por elemento en la i-ésima unidad primaria, puede escribirse:

$$a_{ij} = \bar{\bar{a}} + (\bar{a}_i - \bar{\bar{a}}) + (a_{ij} - \bar{a}_i)$$

o bien

$$a_{ij} = \bar{\bar{a}} + t_i + u_{ij}$$

si se pone

$$t_i = \bar{a}_i - \bar{\bar{a}} \qquad u_{ij} = a_{ij} - \bar{a}_i$$

Supongamos ahora que se toma una muestra simple al azar de n unidades primarias, y de cada una de ellas una muestra análoga de extensión m. Si se indica con y_{ij} el valor de la característica estudiada en el j-ésimo elemento incluído en la muestra extraída del i-ésimo grupo obtenido en la primera etapa de muestreo, tendremos:

$$y_{ij} = \bar{\bar{a}} + t'_i + u'_{ij}$$

donde t'_i y u'_{ij} indican variables aleatorias que pueden tomar los valores

$$t_1 \quad t_2 \dots t_N$$

$$u_{i1} \quad u_{i2} \dots u_{iM} \qquad i = 1, N$$

respectivamente, con probabilidades $1/N$ y $1/M$.

Es evidente que

$$E(t'_i) = 0 \qquad E(t'_i)^2 = S_b^2$$

siendo

$$S_b^2 = \frac{1}{N-1} \sum_1^N (\bar{a}_i - \bar{\bar{a}})^2$$

y que

$$E_j(u'_{ij}) = 0 \qquad E_j(u'_{ij})^2 = S_i^2$$

donde E_j indica que la E.M. se calcula dentro de un grupo, es decir con respecto a j para un subíndice i fijo.

S_i^2 es:

$$S_i^2 = \frac{1}{M-1} \sum_1^M (a_{ij} - \bar{a}_i)^2$$

La media por elemento en la muestra es:

$$\bar{\bar{y}} = \frac{1}{nm} \sum_1^n \sum_1^m y_{ij}$$

y constituye una estimación consistente y "no viciada" de $\bar{\bar{a}}$. En efecto

$$\sum_1^n \sum_1^m y_{ij} = nm\bar{\bar{a}} + m\sum_1^n t'_i + \sum_1^n \sum_1^m u'_{ij}$$

de donde se sigue inmediatamente que

$$E\left(\sum_{1}^{n}\sum_{1}^{m} y_{ij}\right) = nm\bar{a}$$

puesto que

$$E\left(\sum_{1}^{n} t_i'\right) = 0 \qquad E\left(\sum_{1}^{n}\sum_{1}^{m} u'_{ij}\right) = 0$$

Para determinar la variancia de la estimación partimos de:

$$\bar{y} - \bar{a} = \frac{1}{n}\sum_{1}^{n} t_i' + \frac{1}{nm}\sum_{1}^{n}\sum_{1}^{m} u'_{ij}$$

o bien, poniendo:

$$\frac{1}{n}\sum_{1}^{n} t_i' = \bar{t}' \qquad \frac{1}{m}\sum_{1}^{n} u'_{ij} = \bar{u}_i'$$

$$\bar{y} - \bar{a} = t' + \frac{1}{n}\sum_{1}^{n} \bar{u}_i' \qquad (1)$$

Aquí, $\bar{t}'$ es la media aritmética de los n desvíos de las medias de los n grupos otenidos en la primera etapa de muestreo con respecto a la media de la población, y u_i' es la media aritmética de los m desvíos de los valores y_{ij} obtenidos en la muestra extraída del i-ésimo grupo, con respecto a la media del mismo.

De la (1) se sigue que:

$$E(\bar{y} - \bar{a})^2 = E\left(\bar{t}' + \frac{1}{n}\sum_{1}^{n} \bar{u}_i'\right)^2 =$$

$$= E(\bar{t}')^2 + \frac{1}{n^2} E\left(\sum \bar{u}_i'\right)^2 + \frac{2}{n} E\left(\bar{t}' \cdot \sum u_i'\right) = E(\bar{t}')^2 + \frac{1}{n^2} E\left(\sum \bar{u}_i\right)^2 \qquad (2)$$

puesto que $E(\bar{t}' \cdot \sum \bar{u}_i') = 0$

Ahora

$$E(\bar{t}')^2 = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2$$

puesto que $\bar{t}'$ es la media aritmética de los valores obtenidos en una muestra simple al azar de extensión n extraída de una población de N valores t que tienen media 0 y variancia S_b^2.

Por su parte

$$\frac{1}{n^2} E\left(\sum \bar{u}_i'\right)^2 = \frac{1}{n^2} E\left\{\sum \bar{u}_i'^2 + \sum\sum_{i \neq j} \bar{u}'_i \, \bar{u}'_j\right\}$$

La E.M. de la doble sumatoria es nula puesto que las muestras extraídas de dos grupos obtenidos en la primera etapa de muestreo son independientes. Para calcular,

$$\frac{1}{n^2} E\left\{\sum \bar{u}_i'^2\right\}$$

debe tenerse en cuenta que para cada i fijo $\bar{u}_i'$ es una variable aleatoria, de manera que la E debe calcularse primeramente para un conjunto de n grupos de dados y luego para todas las muestras posibles de n unidades primarias, de manera que:

$$\frac{1}{n^2} E\left\{\sum \bar{u}_i'^2\right\} = \frac{1}{n^2} E_i\left\{\sum E_i \, \bar{u}_i'^2 = \frac{1}{n^2} E_i\left\{\sum \left(\frac{1}{m} - \frac{1}{M}\right) S_i^2\right\}\right.$$

puesto que para cada i es $\bar{u}_i'$ la media aritmética de una muestra simple al azar de extensión m_i extraída de la población de M elementos cuyos valores tienen media 0 y variancia S_i^2

Ahora

$$\frac{1}{n^2} E_i\left\{\sum \left(\frac{1}{m} - \frac{1}{M}\right) S_i^2\right\} = \frac{1}{n^2} \left(\frac{1}{m} - \frac{1}{M}\right) n \sum E_i (S_i)^2 =$$

$$\frac{1}{n} \left(\frac{1}{m} - \frac{1}{M}\right) \frac{1}{N} \sum_1^N S_i^2 = \frac{1}{n} \left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2 =$$

donde

$$\bar{S}_w^2 = \frac{1}{N} \sum_1^N S_i^2 = \frac{1}{N(M-1)} \sum_1^N \sum_1^M (a_{ij} - \bar{a}_i)^2$$

Llevando los valores hallados a la (2), queda finalmente:

$$V_2\ (\bar{\bar{y}}) = \left(\frac{1}{n} - \frac{1}{n}\right) S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2 \qquad (3)$$

donde se usa el símbolo $V_2\ (\bar{\bar{y}})$ para indicar que se trata de una estimación basada en muestreo en dos etapas.

Si en lugar de tomar una muestra de extensión m de cada una de las n unidades primarias obtenidas en la primera etapa de muestreo, se analizan todos los elementos que las integran, es decir, solo hay una etapa de muestreo, siendo m = M, la (3) se reduce a

$$\left(\frac{1}{n} - \frac{1}{N}\right) S_b^2$$

que es la variancia de la estimación de la media por elemento que se obtuvo en el capítulo anterior para el caso del muestreo por grupos de una etapa.

A su vez, si la muestra de unidades primarias es tal que agota las N en que se agrupan los elementos de la población, y de cada uno de ellas se toma una muestra simple al azar de extensión m, se vé que se está frente al caso de muestreo de una población estratificada con adjudicación proporcional. Al ser n = N, el primer sumando de (3) se anula y queda:

$$\frac{1}{n}\left(\frac{1}{m} - \frac{1}{n}\right) \bar{S}_w^2$$

que es la forma a la que se reduce la variancia de la estimación de la media en el muestreo estratificado con adjudicación proporcional cuando el número de estratos es N, el tamaño de la población NM, el tamaño de cada estrato M y la muestra extraída de cada uno tiene tamaño m.

Si el número de unidades primarias de muestreo y el de elementos dentro de cada una es suficientemente grande como para que pueda considerarse que

$$\frac{N - n}{N} \quad y \quad \frac{M - n}{M}$$

son iguales a la unidad, la (3) se reduce a:

$$V_2\ (\bar{\bar{y}}) \quad \frac{S_b^2}{n} + \frac{S_w^2}{nm}$$

Comparación de $V_2(\bar{\bar{y}})$ con $V(\bar{y})$ y $V_1(\bar{\bar{y}})$

Supuesto que se tiene una población de N.M. elementos distribuídos en N conglomerados de M elementos cada uno, resulta de interés comparar la variancia de la estimación de la media resultante de la extracción de una muestra de tamaño global n.m. en 2 etapas, con la que se obtendría si:

1º) Se extrae de la población una muestra simple al azar de la misma extensión n.m.

2º) Una muestra en 1 etapa de la población clasificada en el mismo número de grupos, la que, debiendo constar de n.m elementos, requiere la selección de n.m/M grupos, en cada uno de los cuales la característica estudiada se evaluará en todos los M elementos que lo integran.

De estas comparaciones surgirá en qué casos, es decir, para qué valores de alguna característica de la población, el muestreo en 2 etapas ofrecerá una estimación más precisa.

1º) Comparación con $V(\bar{y})$

La variancia de la estimación de la media basada en una muestra simple al azar es:

$$V(\bar{y}) = \left(\frac{1}{nm} - \frac{1}{NM}\right) S^2$$

donde S^2 es la variancia de los elementos en la población.

Ahora, si en la expresión que dá $V_2(\bar{\bar{y}})$, sustituímos S_b^2 y S_w^2 por sus valores dados en términos de S^2 y del coeficiente de correlación intraclase(v. pág), se tendrá:

$$V_2(\bar{\bar{y}}) = \frac{NM-1}{NM}\cdot\frac{S^2}{nm}\left\{\frac{M-N}{M}(1-\rho) + \frac{N-n}{N-1}\cdot\frac{m}{M}\left[1-(M-1)\rho\right]\right\} +$$

$$+ \frac{NM-1}{NM}\cdot\frac{S^2}{nm}\left\{1-\frac{m}{M}\left(1-\frac{N-n}{N-1}\right)+\rho\left[\frac{N-n}{N-1}\cdot\frac{m}{M}(M-1)-\frac{M-m}{M}\right]\right\}$$

Si la tasa de muestreo n/M es pequeña y tomamos $(NM-1)/NM \doteq 1$, puede escribirse aproximadamente:

$$V_2(\bar{\bar{y}}) \doteq \frac{S^2}{nm}\left\{1+\rho\left(\frac{N-n}{N-1}m-1\right)\right\}$$

y puesto que $1/NM$ se supone prácticamente 0, de modo que

$$V_2(\bar{\bar{y}}) \doteq \frac{S^2}{nm}$$

resulta que la eficiencia relativa del muestreo en 2 etapas con respecto al muestreo simple al azar viene dado por:

$$E.R. = \frac{V(\bar{y})}{V_2(\bar{y})} = \frac{1}{1 + \rho\left(\frac{N-n}{N-1}\right)(m-1)}$$

que es la expresión que se tendría para la eficiencia relativa del muestreo por grupos en 1 etapa. Si cada grupo constara de m(N-n)/(N-1) elementos. Si el coeficiente de correlación intra-clase es negativo, el muestreo por grupos en 2 etapas es más eficiente que el muestreo simple al azar, de elementos, ocurriendo lo contrario cuando él es positivo.

2º) Comparación con $V_1(\bar{y})$

Cuando en el muestreo por grupos en 1 etapa se seleccionan nm/M unidades primarias, la variancia de la estimación de la media es:

$$V_1(\bar{y}) = \left(\frac{M}{nm} - \frac{1}{N}\right) S_b^2$$

Si restamos de ésta la $V_2(\bar{y})$ queda:

$$V_1(\bar{y}) - V_2(\bar{y}) = \frac{1}{n}\left\{\left(\frac{M}{m} - 1\right) S_b^2 - \left(\frac{1}{m} - \frac{1}{M}\right) S_w^2\right\} =$$

$$= \frac{1}{n}\left\{\left(\frac{M}{n} - 1\right) S_b^2 - \frac{1}{n}\left(\frac{M}{m} - 1\right) S_w^2\right\} =$$

$$\frac{1}{n}\left(\frac{M}{m} - 1\right)\left(S_b^2 - \frac{1}{M} S_w^2\right)$$

Ahora hemos visto más arriba (pág.) que, para N grande, es:

$$S_b^2 - \frac{1}{M} S_w^2 \doteq \rho S^2$$

de modo que podemos escribir:

$$V_1(\bar{y}) - V_2(\bar{y}) \doteq \frac{1}{n}\left(\frac{M}{m} - 1\right) \rho S^2$$

Si $\rho > 0$, la diferencia entre $V_1(\bar{y})$ y $V_2(\bar{y})$ será positiva, y tanto más grande cuanto más pequeño sea m con respecto a M, es decir, cuanto más pequeña sea la tasa de muestreo m/M dentro de cada grupo. En otros términos, si el coeficiente de correlación es positivo, para un determinado tamaño n.m de la muestra global, es preferible el muestreo por grupos en 2 etapas que selecciona un número n de grupos y m elementos de cada uno de ellos, al muestreo en 1 etapa que selecciona nm/M grupos y evalúa la característica en los M elementos que cada uno encierra. Si en cambio, el coeficiente de correlación intra-clases es negativo, el muestreo en 1 etapa da una estimación más precisa.

Estimación de la variancia a partir de los datos de la muestra:

Como en general no se conocen los valores de los parámetros S_b^2 y S_w^2 de la población dividida en grupos, para evaluar la precisión de la estimación de la media $\bar{y}$ es necesario calcular una estimación de $V_2(\bar{y})$. Para tener esta estimación, que se indicará con $v_2(\bar{y})$, partimos de las expresiones

$$s_b^2 = \frac{1}{n-1} \sum_{1}^{n} (\bar{y}_i - \bar{\bar{y}})^2$$

$$s_i^2 \quad \frac{1}{m-1} \quad \sum_{1}^{m} (y_{ij} - \bar{y}_i)^2$$

construídas con los datos ofrecidos por la muestra, y calculamos sus respectivas esperanzas matemáticas con las que obtendremos expresiones que permitirán formar estimaciones "no viciadas" de S_b^2 y S_w^2.

Por de pronto observemos que, puesto que dentro de cada grupo seleccionado en la primera etapa del muestreo, se toma una muestra simple al azar de elementos, se verificará que

$$E(s_i^2) = S_i^2$$

y por lo tanto

$$E(\bar{s}_w^2) = E\left\{ \frac{1}{n} \sum_{1}^{n} s_i^2 \right\} = \bar{S}_w^2$$

Para calcular la esperanza matemática de s_b^2, escribimos

$$s_b^2 = \frac{1}{n-1} \left\{ \sum_{1}^{n} \bar{y}_i^2 - n\bar{\bar{y}}^2 \right\}$$

de manera que será:

.t.p.V

$$(n-1)\ E\ (\ s_b^2\ )\ E\ \left\{ \sum_1^n \bar{y}_i^2 - nE\ (\bar{\bar{y}})^2 \right\} \qquad (1)$$

Ahora $$E\left(\sum \bar{y}_i^2\right) = E_i \left\{ \sum E_j\ (\bar{y}_i^2) = E_i \quad E_j(\bar{y}_i - \bar{a}_i)^2 \quad \bar{a}_i^2 \right.$$

y como para cada i fijo $\bar{y}_i$ es la media aritmética de una muestra simple al azar de extensión m extraída de una población de M elementos cuya media y variancia son, respectivamente, $\bar{a}_i$ y S_i^2, se tiene que el último miembro de la igualdad anterior es:

$$E_i \left\{ \sum \left(\frac{1}{m} - \frac{1}{M} \right) S_i^2 + \bar{a}_i^2 \right\} = \left(\frac{1}{m} - \frac{1}{M} \right) \sum E_i(S_i^2) + nE_i(\bar{a}_i^2) =$$

$$= n\left(\frac{1}{m} - \frac{1}{M}\right)\ \bar{S}_w^2 + \frac{n}{N}\ \sum_1^N \bar{a}_i^2 \qquad (2)$$

Puesto que

$$(N-1)\ S_a^2 = \sum_1^N (\bar{a}_i - \bar{\bar{a}})^2 = \sum_1^N \bar{a}_i^2 - N\bar{\bar{a}}^2$$

se tiene

$$\frac{n}{N} \sum_1^N \bar{a}_i^2 = n\left(1 - \frac{1}{N}\right) S_b^2 + n\bar{\bar{a}}^2$$

de modo que la (2) puede escribirse:

$$n\left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2 + n\left(1 - \frac{1}{N}\right) S_b^2 + n\bar{\bar{a}}^2 \qquad (3)$$

Por otra parte

$$V_2\ (\bar{\bar{y}}) = E\ (\bar{\bar{y}}^2) - \bar{\bar{a}}^2 = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2$$

de modo que

$$nE\ (\bar{\bar{y}})^2 = n\left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2 + n\bar{\bar{a}}^2 \qquad (4)$$

Reemplazando los 2 términos el 2º miembro de la (1) por sus valores hallados en (3) y (4), se tiene:

$$(n-1)E(s_b^2) = n\left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2 + n\left(1 - \frac{1}{N}\right) S_b^2 + n\bar{\bar{a}}^2 - n\left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 - \left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2 - n\bar{\bar{a}}^2 =$$

$$= (n-1) \left\{ S_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2 \right\}$$

de modo que resulta finalmente:

$$E\,(s_b^2) = S_b^2 + \left(\frac{1}{m} - \frac{1}{M}\right)\,\bar{S}_w^2 \qquad (5)$$

de donde se sigue que la estimación de S_b^2 vendrá dado por:

$$\overline{S}_b^2 = s_b^2 - \left(\frac{1}{m} - \frac{1}{M}\right)\,\bar{s}_w^2$$

$\overline{S}_b^2$ y $\overline{S}_w^2$ Los valores requeridosspara calcular las estimaciones se obtienen a partir del cuadro de análisisde variancia de la muestra.

Si en la expresión que dá $V_2\,(\bar{\bar{y}})$ reemplazamos S_b^2 y $\bar{S}_w^2$ por sus estimaciones basadas en la muestra, tendremos la estimación buscada de la variancia.

$$v_2\,(\bar{\bar{y}}) = \left(\frac{1}{n} - \frac{1}{N}\right)\,s_b^2 + \frac{1}{N}\,\left(\frac{1}{m} - \frac{1}{M}\right)\,\bar{s}_w^2$$

CAPITULO V

MUESTREO SIMPLE POR GRUPOS - (II)

SUB, MUESTREO EN 2 ETAPAS

Pasaremos ahora a considerar el muestreo por grupos en 2 etapas en el que, supuesto que se tienen N grupos o unidades primarias de muestreo, se llega a los elementos de la población que han de ofrecer la información para la confección de las estimaciones, mediante una doble operación de muestreo : 1º una muestra de unidades primarias, 2º una muestra de elmentos extraída de cada unidad obtenida en la primera muestra. Supondremos que en ambos pasos de obtiene la muestra mediante una operación de muestreo simple al azar y, que las N unidades primarias en las que se agrupan los elementos de la población constan todas del mismo número M de elementos. La operación de muestreo selecciona n unidades primarias, y de cada una de ellas m unidades secundarias, de modo que el tamaño total de la muestra es n.m.-

Sea a_{ij} el valor que la característica estudiada toma en el j-ésimo individuo de la i-ésima unidad primaria. Si es $\bar{\bar{a}}$ la media por elemento en la población de NM elementos y $\bar{a}_i$ la media por elemento en la i-ésima unidad primaria, puede escribirse:

$$a_{ij} = \bar{\bar{a}} + (\bar{a}_i - \bar{\bar{a}}) + (a_{ij} - \bar{a}_i)$$

o bien

$$a_{ij} = \bar{\bar{a}} + t_i + u_{ij}$$

si se pone

$$t_i = \bar{a}_i - \bar{\bar{a}} \qquad u_{ij} = a_{ij} - \bar{a}_i$$

Supongamos ahora que se toma una muestra simple al azar de n unidades primarias, y de cada una de ellas una muestra análoga de extensión m. Si se indica con y_{ij} el valor de la característica estudiada en el j-ésimo elemento incluído en la muestra extraída del i-ésimo grupo obtenido en la primera etapa de muestreo, tendremos:

$$y_{ij} = \bar{\bar{a}} + t'_i + u'_{ij}$$

donde t'_i y u'_{ij} indican variables aleatorias que pueden tomar los valores

$$t_1 \quad t_2 \cdots t_N$$

$$u_{ij} \quad u_{i2} \cdots u_{iM} \qquad i = 1,N$$

respectivamente, con probabilidades 1/N y 1/M.

Es evidente que

$$E(t'_i) = 0 \qquad E(t'_i)^2 = S_b^2$$

siendo

$$S_b^2 = \frac{1}{N-1} \sum_1^N (\bar{a}_i - \bar{\bar{a}})^2$$

y que

$$E_j(u'_{ij}) = 0 \qquad E_j(u'_{ij})^2 = S_i^2$$

donde E_j indica que la E.M. se calcula dentro de un grupo, es decir con respecto a j para un subíndice i fijo.

S_i^2 es:

$$S_i^2 = \frac{1}{M-1} \sum_1^M (a_{ij} - \bar{a}_i)^2$$

La media por elemento en la muestra es:

$$\bar{\bar{y}} = \frac{1}{nm} \sum_1^n \sum_1^m y_{ij}$$

y constituye una estimación consistente y "no viciada" de $\bar{\bar{a}}$. En efecto

$$\sum_1^n \sum_1^m y_{ij} = nm\bar{\bar{a}} + m\sum_1^n t'_i + \sum_1^n \sum_1^m u'_{ij}$$

M

de donde se sigue inmediatamente que

$$E\left(\sum_{1}^{n}\sum_{1}^{m} y_{ij}\right) = nm\bar{a}$$

puesto que

$$E\left(\sum_{1}^{n} t_i'\right) = 0 \qquad E\left(\sum_{1}^{n}\sum_{1}^{m} u_{ij}'\right) = 0$$

Para determinar la variancia de la estimación partimos de:

$$\bar{y} - \bar{a} = \frac{1}{n}\sum_{1}^{n} t_i' + \frac{1}{nm}\sum_{1}^{n}\sum_{1}^{m} u'_{ij}$$

o bien, poniendo:

$$\frac{1}{n}\sum_{1}^{n} t_i' = \bar{t}' \qquad \frac{1}{m}\sum_{1}^{n} u_{ij}' = \bar{u}_i'$$

$$\bar{y} - \bar{a} = t' + \frac{1}{n}\sum_{1}^{n} \bar{u}_i' \qquad (1)$$

Aquí, $\bar{t}'$ es la media aritmética de los n desvíos de las medias de los n grupos otenidos en la primera etapa de muestreo con respecto a la media de la población, y u_i' es la media aritmética de los m desvíos de los valores y_{ij} obtenidos en la muestra extraída del i-ésimo grupo, con respecto a la media del mismo.

De la (1) se sigue que:

$$E(\bar{y} - \bar{a})^2 = E\left(\bar{t}' + \frac{1}{n}\sum_{1}^{n} \bar{u}_i'\right)^2 =$$

$$= E(\bar{t}')^2 + \frac{1}{n^2} E\left(\sum \bar{u}_i'\right)^2 + \frac{2}{n} E\left(\bar{t}' \cdot \sum u_i'\right) = E(\bar{t}')^2 + \frac{1}{n^2} E\left(\sum \bar{u}_i\right)^2 \qquad (2)$$

puesto que $E(\bar{t}' \cdot \sum \bar{u}_i') = 0$

Ahora

$$E(\bar{t}')^2 = \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2$$

puesto que $\bar{t}'$ es la media aritmética de los valores obtenidos en una muestra simple al azar de extensión n extraída de una población de N valores t que tienen media 0 y variancia S_b^2.

Por su parte

$$\frac{1}{n^2} E\left(\sum \bar{u}_i'\right)^2 = \frac{1}{n^2} E\left\{\sum \bar{u}_i'^2 + \sum\sum_{i \neq j} \bar{u}'_i \, \bar{u}'_j\right\}$$

La E.M. de la doble sumatoria es nula puesto que las muestras extraídas de dos grupos obtenidos en la primera etapa de muestreo son independientes. Para calcular,

$$\frac{1}{n^2} E\left\{\sum \bar{u}_i'^2\right\}$$

debe tenerse en cuenta que para cada i fijo $\bar{u}_i'$ es una variable aleatoria, de manera que la E debe calcularse primeramente para un conjunto de n grupos de dados y luego para todas las muestras posibles de n unidades primarias, de manera que:

$$\frac{1}{n^2} E\left\{\sum \bar{u}_i'^2\right\} = \frac{1}{n^2} E_i\left\{\sum E_i \, \bar{u}_i'^2 = \frac{1}{n^2} E_i\left\{\sum \left(\frac{1}{m} - \frac{1}{M}\right) S_i^2\right\}\right.$$

puesto que para cada i es $\bar{u}_i'$ la media aritmética de una muestra simple al azar de extensión m extraída de la población de M elementos cuyos valores tienen media 0 y variancia S_i^2

Ahora

$$\frac{1}{n^2} E_i\left\{\sum \left(\frac{1}{m} - \frac{1}{M}\right) S_i^2\right\} = \frac{1}{n^2} \left(\frac{1}{m} - \frac{1}{M}\right) n \sum E_i \, (S_i)^2 =$$

$$\frac{1}{n} \left(\frac{1}{m} - \frac{1}{M}\right) \frac{1}{N} \sum_1^N S_i^2 = \frac{1}{n} \left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2 =$$

donde

$$\bar{S}_w^2 = \frac{1}{N} \sum_1^N S_i^2 = \frac{1}{N(M-1)} \sum_1^N \sum_1^M (a_{ij} - \bar{a}_i)^2$$

Llevando los valores hallados a la (2), queda finalmente:

$$V_2\ (\bar{\bar{y}}) = \left(\frac{1}{n} - \frac{1}{n}\right) S_b^2 + \frac{1}{n}\left(\frac{1}{m} - \frac{1}{M}\right) \bar{S}_w^2 \qquad (3)$$

donde se usa el símbolo $V_2\ (\bar{\bar{y}})$ para indicar que se trata de una estimación basada en muestreo en dos etapas.

Si en lugar de tomar una muestra de extensión m de cada una de las n unidades primarias obtenidas en la primera etapa de muestreo, se analizan todos los elementos que las integran, es decir, solo hay una etapa de muestreo, siendo m = M, la (3) se reduce a

$$\left(\frac{1}{n} - \frac{1}{N}\right) S_b^2$$

que es la variancia de la estimación de la media por elemento que se obtuvo en el capítulo anterior para el caso del muestreo por grupos de una etapa.

A su vez, si la muestra de unidades primarias es tal que agota las N en que se agrupan los elementos de la población, y de cada uno de ellas se toma una muestra simple al azar de extensión m, se vé que se está frente al caso de muestreo de una población estratificada con adjudicación proporcional. Al ser n = N, el primer sumando de (3) se anula y queda:

$$\frac{1}{n}\left(\frac{1}{m} - \frac{1}{n}\right) \bar{S}_w^2$$

que es la forma a la que se reduce la variancia de la estimación de la media en el muestreo estratificado con adjudicación proporcional cuando el número de estratos es N, el tamaño de la población NM, el tamaño de cada estrato M y la muestra extraída de cada uno tiene tamaño m.

Si el número de unidades primarias de muestreo y el de elementos dentro de cada una es suficientemente grande como para que pueda considerarse que

$$\frac{N - n}{N} \quad y \quad \frac{M - n}{M}$$

son iguales a la unidad, la (3) se reduce a:

$$V_2\ (\bar{\bar{y}}) \quad \frac{S_b^2}{n} + \frac{S_w^2}{nm}$$

Comparación de $V_2(\bar{\bar{y}})$ con $V(\bar{y})$ y $V_1(\bar{\bar{y}})$

Supuesto que se tiene una población de N.M. elementos distribuídos en N conglomerados de M elementos cada uno, resulta de interés comparar la variancia de la estimación de la media resultante de la extracción de una muestra de tamaño global n.m. en 2 etapas, con la que se obtendría si:

1º) Se extrae de la población una muestra simple al azar de la misma extensión n.m.

2º) Una muestra en 1 etapa de la población clasificada en el mismo número de grupos, la que, debiendo constar de n.m elementos, requiere la selección de n.m/M grupos, en cada uno de los cuales la característica estudiada se evaluará en todos los M elementos que lo integran.

De estas comparaciones surgirá en qué casos, es decir, para qué valores de alguna característica de la población, el muestreo en 2 etapas ofrecerá una estimación más precisa.

1º) Comparación con $V(\bar{y})$

La variancia de la estimación de la media basada en una muestra simple al azar es:

$$V(\bar{y}) = \left(\frac{1}{nm} - \frac{1}{NM}\right) S^2$$

donde S^2 es la variancia de los elementos en la población.

Ahora, si en la expresión que dá $V_2(\bar{\bar{y}})$, sustituímos S_b^2 y S_w^2 por sus valores dados en términos de S^2 y del coeficiente de correlación intraclase(v. pág), se tendrá:

$$V_2(\bar{\bar{y}}) = \frac{NM-1}{NM} \cdot \frac{S^2}{nm} \left\{ \frac{M-N}{M}(1-\rho) + \frac{N-n}{N-1} \cdot \frac{m}{M} \left[1-(M-1)\rho \right] \right\} =$$

$$= \frac{NM-1}{NM} \cdot \frac{S^2}{nm} \left\{ 1 - \frac{m}{M}\left(1 - \frac{N-n}{N-1}\right) + \rho \left[\frac{N-n}{N-1} \cdot \frac{m}{M}(M-1) - \frac{M-m}{M} \right] \right\}$$

Si la tasa de muestreo n/M es pequeña y tomamos $(NM-1)/NM \doteq 1$, puede escribirse aproximadamente:

$$V_2(\bar{\bar{y}}) \doteq \frac{S^2}{nm}\left[1 + \rho\left(\frac{N-n}{N-1} m - 1\right) \right]$$

y puesto que $1/NM$ se supone prácticamente 0, de modo que

$$V_2(\bar{\bar{y}}) \doteq \frac{S^2}{nm}$$

resulta que la eficiencia relativa del muestreo en 2 etapas con respecto al muestreo simple al azar viene dado por:

$$E.R. = \frac{V(\bar{y})}{V_2(\bar{y})} = \frac{1}{1 + \rho\left(\frac{N-n}{N-1}\right)(m-1)}$$

que es la expresión que se tendría para la eficiencia relativa del muestreo por grupos en 1 etapa. Si cada grupo constara de m(N-n)/(N-1) elementos. Si el coeficiente de correlación intra-clase es negativo, el muestreo por grupos en 2 etapas es más eficiente que el muestreo simple al azar, de elementos, ocurriendo lo contrario cuando él es positivo.

2º) Comparación con $V_1(\bar{y})$

Cuando en el muestreo por grupos en 1 etapa se seleccionan nm/M unidades primarias, la variancia de la estimación de la media es:

$$V_1(\bar{y}) = \left(\frac{M}{nm} - \frac{1}{N}\right) S_b^2$$

Si restamos de ésta la $V_2(\bar{y})$ queda:

$$V_1(\bar{y}) - V_2(\bar{y}) = \frac{1}{n}\left\{\left(\frac{M}{m} - 1\right) S_b^2 - \left(\frac{1}{m} - \frac{1}{M}\right) S_w^2\right\} =$$

$$= \frac{1}{n}\left\{\left(\frac{M}{n} - 1\right) S_b^2 - \frac{1}{n}\left(\frac{M}{m} - 1\right) S_w^2\right\} =$$

$$\frac{1}{n}\left(\frac{M}{m} - 1\right)\left(S_b^2 - \frac{1}{M} S_w^2\right)$$

Ahora hemos visto más arriba (pág.) que, para N grande, es:

$$S_b^2 - \frac{1}{M} S_w^2 \approx \rho S^2$$

de modo que podemos escribir:

$$V_1(\bar{y}) - V_2(\bar{y}) = \frac{1}{n}\left(\frac{M}{m} - 1\right) \rho S^2$$

Si $\rho > 0$, la diferencia entre $V_1(\bar{y})$ y $V_2(\bar{y})$ será positiva, y tanto más grande cuanto más pequeño sea m con respecto a M, es decir, cuanto más pequeña sea la tasa de muestreo m/M dentro de cada grupo. En otros términos, si el coeficiente de correlación es positivo, para un determinado tamaño n,m de la muestra global, es preferible el muestreo por grupos en 2 etapas que selecciona un número n de grupos y m elementos de cada uno de ellos, al muestreo en 1 etapa que selecciona nm/M grupos y evalúa la característica en los M elementos que cada uno encierra. Si en cambio, el coeficiente de correlación intra-clases es negativo, el muestreo en 1 etapa da una estimación más precisa.

Estimación de la variancia a partir de los datos de la muestra:

Como en general no se conocen los valores de los parámetros S_b^2 y S_w^2 de la población dividida en grupos, para evaluar la precisión de la estimación de la media $\bar{y}$ es necesario calcular una estimación de $V_2(\bar{y})$. Para tener esta estimación, que se indicará con $v_2(\bar{y})$, partimos de las expresiones

$$s_b^2 = \frac{1}{n-1} \sum_{1}^{n} (\bar{y}_i - \bar{\bar{y}})^2$$

$$s_i^2 \quad \frac{1}{m-1} \sum_{1}^{m} (y_{ij} - \bar{y}_i)^2$$

construídas con los datos ofrecidos por la muestra, y calculamos sus respectivas esperanzas matemáticas con las que obtendremos expresiones que permitirán formar estimaciones "no viciadas" de S_b^2 y S_w^2.

Por de pronto observemos que, puesto que dentro de cada grupo seleccionado en la primera etapa del muestreo, se toma una muestra simple al azar de elementos, se verificará que

$$E(s_i^2) = S_i^2$$

y por lo tanto

$$E(s_w^2) = E\left\{ \frac{1}{n} \sum_{1}^{n} s_i^2 \right\} = \bar{S}_w^2$$

Para calcular la esperanza matemática de s_b^2, escribimos

$$s_b^2 = \frac{1}{n-1} \left\{ \sum_{1}^{n} \bar{y}_i^2 - n\bar{\bar{y}}^2 \right\}$$

de manera que será:

.t.p.V

$$(n-1)\ E\left(s_b^2\right)\ E\left\{\sum_1^n \bar{y}_1^2 - nE\,(\bar{\bar{y}})^2\right\} \qquad (1)$$

Ahora $$E\left(\sum \bar{y}_i^2\right) = E_i\left\{\sum E_j\,(\bar{y}_i^2) = E_i \cdot E_j(\bar{y}_i - \bar{a}_i)^2 \quad \bar{a}_i^2\right.$$

y como para cada i fijo $\bar{y}_i$ es la media aritmética de una muestra simple al azar de extensión m extraída de una población de M elementos cuya media y variancia son, respectivamente, $\bar{a}_i$ y S_i^2, se tiene que el último miembro de la igualdad anterior es:

$$E_i\left\{\sum\left(\frac{1}{m}-\frac{1}{M}\right)S_i^2 + \bar{a}_i^2\right\} = \left(\frac{1}{m}-\frac{1}{M}\right)\sum E_i(S_i^2) + nE_i(\bar{a}_i^2) =$$

$$= n\left(\frac{1}{m}-\frac{1}{M}\right)\ \bar{S}_w^2 + \frac{n}{N}\ \sum_1^N \bar{a}_i^2 \qquad (2)$$

Puesto que

$$(N-1)\,S_a^2 = \sum_1^N (\bar{a}_i - \bar{\bar{a}})^2 = \sum_1^N \bar{a}_i^2 - N\bar{\bar{a}}^2$$

se tiene

$$\frac{n}{N}\sum_1^N \bar{a}_i^2 = n\left(1-\frac{1}{N}\right)S_b^2 + n\bar{\bar{a}}^2$$

de modo que la (2) puede escribirse:

$$n\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + n\left(1-\frac{1}{N}\right)S_b^2 + n\bar{\bar{a}}^2 \qquad (3)$$

Por otra parte

$$V_2\,(\bar{\bar{y}}) = E\,(\bar{\bar{y}}^2) - \bar{\bar{a}}^2 = \left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \frac{1}{n}\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2$$

de modo que

$$nE\,(\bar{\bar{y}})^2 = n\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + n\bar{\bar{a}}^2 \qquad (4)$$

Reemplazando los 2 términos el 2º miembro de la (1) por sus valores hallados en (3) y (4), se tiene:

$$(n-1)E(s_b^2) = n\left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 + n\left(1-\frac{1}{N}\right)S_b^2 + n\bar{\bar{a}}^2 - n\left(\frac{1}{n}-\frac{1}{N}\right)S_b^2 - \left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2 - n\bar{\bar{a}}^2 =$$

$$= (n-1)\left[S_b^2 + \left(\frac{1}{m}-\frac{1}{M}\right)\bar{S}_w^2\right]$$

de modo que resulta finalmente:

$$E\ (s_b^2) = S_b^2 + (\frac{1}{m} - \frac{1}{M})\ \bar{S}_w^2 \qquad (5)$$

de donde se sigue que la estimación de S_b^2 vendrá dado por:

$$''S_b^2 = s_b^2 - (\frac{1}{m} - \frac{1}{M})\ \bar{s}_w^2$$

Los valores requeridosspara calcular las estimaciones $''S_b^2$ y $''\bar{S}_w^2$ se obtienen a partir del cuadro de análisisde variancia de la muestra.

Si en la expresión que dá $V_2\ (\bar{\bar{y}})$ reemplazamos S_b^2 y $\bar{S}_w^2$ por sus estimaciones basadas en la muestra, tendremos la estimación buscada de la variancia.

$$v_2\ (\bar{\bar{y}}) = \frac{1}{n} - \frac{1}{N})\ s_b^2 + \frac{1}{N}\ (\frac{1}{m} - \frac{1}{M})\ \bar{s}_w^2$$

2º) Unidades primarias de tamaños diferentes (n = 1)

Supondremos ahora que la población que ha de someterse al muestreo consta de un cierto número N de grupos o unidades primarias y que es M_i (i = 1 ... N) el número de elementos en cada uno de ellos.

Consideremos, en primer término, el caso en que la primera etapa de muestreo consite en la selección de una única unidad primaria (n = 1) de la cual se elegirá una muestra simple al azar de m_i elementos, que constituirán la muestra en base a la cual se estimarán los parámetros que interesan de la población que consta de $M_o = \sum M_i$ elementos.

En los desarrollos subsiguientes consideraremos que se trata de estimar el valor medio por elemento de una cierta característica

Para ejemplificar tomaremos la siguiente población hipotética:

Grupo Nº	Valores a_{ij}	M_i	$M_i \bar{a}_i$	$\bar{a}_i$	s_i^2
1	3 1 5	3	9	3	4
2	4 2 3 6 5 4	6	24	4	2
3	2 6 3 3 9 8 5 4	8	40	5	6.29
		$M_o = 17$	73	$\bar{\bar{a}}_i = 4$	

$$\bar{\bar{a}} = \frac{\sum M_i \bar{a}_i}{\sum M_i} = 4.29$$

En el caso en que cada grupo consta de un número diferente de elemontos, varias son las estimaciones que pueden construírse. A continuación analizaremos algunas de ellas.

1º) Estimación basada en la media de la muestra.

Supuesto que se ha elegido un grupo de entre la N y que se ha extraído del mismo una muestra de m_i elementos, una estimación simple de la media por elemento de la población, es la media de los elementos obtenidos en la muestra. Se tendrá así

$$\hat{\bar{a}}_I = \bar{y}_i$$

donde

$$\bar{y}_i = \frac{1}{m_i}\,(y_{i1} + y_{i2} + \cdots\ y_{im_i})$$

$\bar{y}_i$ es, como se ve fácilmente una variable aleatoria que toma

$$C_{m_1}^{M_1} + C_{m_2}^{M_2} + \cdots C_{m_N}^{M_N}$$

valores, cada uno de ellos con una cierta probabilidad

$P_{(ij)}$ $(i=1,2, \ldots, N;\ j = 1,2, \ldots, C_{m_i}^{M_i})$. Ahora:

$$p_{(ij)} = p_{(i)}\ \ p_{(j/i)}$$

donde p(i) es la probabilidad de elegir, en la primera operación de muestreo, el i-ésimo grupo y $p_{(j/i)}$ es la probabilidad condicional de obtener el j-ésimo valor de $\bar{y}_i$ en el grupo en cuestión.

Puesto que en el caso que estamos considerando, la elección del único grupo se hace asignando a todos ellos la misma probabilidad, será

$$p_i = 1/N \qquad i = 1, \ldots, N$$

y como la muestra a extraerse del grupo obtenido es una muestra simple al azar, será:

$$p_{(j/i)} = 1/C_{m_i}^{M_i}$$

Para calcular la E. M. de la estimación, se tendrá que evaluar

$$E\,(\bar{y}_i) = \sum_1^N \sum_1^{M_i} p_{(ij)}\,\bar{y}_i = \sum_i^N p_i \sum_i^{M_i} p_{(j/i)}\,\bar{y}_i$$

Pero $\sum p_{(j/i)}\,\bar{y}_i$ no es otra cosa que la E.M. de las medias de muestras de extensión m_i extraídas por una operación de muestreo simple al azar de una población de M_i elementos cuya media es $\bar{a}_i$ Resulta así que

$$E\,(\bar{y}_i)\ \ \frac{1}{N}\ \ \sum \bar{a}_i = \bar{\bar{a}}_i$$

donde $\bar{\bar{a}}_i$ indica la media aritmética de las medias de los N grupos. El anterior resultado muestra que la estimación que hemos construído es "viciada", puesto que

$$E\underset{I}{(\bar{\bar{a}})} = \bar{\bar{a}}_i \neq \bar{\bar{a}}$$

Para tener el error medio cuadrático de esta estimación debemos calcular

$$E(\bar{y}_i - \bar{\bar{a}})^2 = E(\bar{y}_i - \bar{a}_i + \bar{a}_i - \bar{\bar{a}}_i + \bar{\bar{a}}_i - \bar{\bar{a}})^2 = \tag{1}$$

$$= E(\bar{y}_i - \bar{a}_i)^2 + E(\bar{a}_i - \bar{\bar{a}}_i)^2 + (\bar{\bar{a}}_i - \bar{\bar{a}})^2$$

puesto que puede mostrarse fácilmente que las E.M. de los productos

$(\bar{y}_i - \bar{a}_i)(\bar{a}_i - \bar{\bar{a}}_i)$, etc. son nulas.

Ahora:

$$E(\bar{y}_i - \bar{a}_i)^2 = \sum_1^N \sum_1^{M_i} p_{(ij)} (\bar{y}_i - \bar{a}_i)^2 = \sum_1^N p_{(i)} \left[\sum_1^{M_i} p_{(j/i)} (\bar{y}_i - \bar{a}_i)^2 \right]$$

pero la suma entre corchetes no es otra cosa que la E.M. de los cuadrados de los desvíos de la media de una muestra simple al azar con respecto a la media de la población de donde se la ha extraído, y por lo tanto ella vale:

$$\left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2$$

de modo que queda:

$$E(\bar{y}_i - \bar{a}_i)^2 = \sum_1^N p_{(i)} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2 = \frac{1}{N} \sum_1^N \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2 \tag{2}$$

Por su parte:

$$E(\bar{a}_i - \bar{\bar{a}}_i)^2 = \sum_1^N p_{(i)} (\bar{a}_i - \bar{\bar{a}}_i)^2 = \frac{1}{N} \sum_1^N (\bar{a}_i - \bar{\bar{a}}_i)^2 \tag{3}$$

puesto que cada grupo se elige con probabilidad $1/N$.-

Llevando los resultados (2) y (3) a la (1), se tiene:

$$\hat{V}(\bar{\bar{a}}_I) = \frac{1}{N}\sum_{1}^{N}\left(\frac{1}{\bar{m}_i} - \frac{1}{M_i}\right) s_i^2 + \frac{1}{N}\sum_{1}^{M}(\bar{\bar{a}}_i - \bar{\bar{a}}_i)^2 + (\bar{\bar{a}}_i - \bar{\bar{a}})^2 \qquad (4)$$

El error medio cuadrático de la estimación, -que hemos indicado con $V_{(\bar{\bar{a}}_I)}$ como si se tratara de una variancia-, está constituído por tres componentes: 1º) variancia dentro de los grupos, dada por el primer sumando del 2º miembro; 2º) variancia entre grupos, dada por el segundo sumando, y 3º) el cuadrado del "vicio" o "error sistemático" $\bar{\bar{a}}_i - \bar{\bar{a}}$. Se ve que el tamaño de la muestra solo afecta a la primera componente.

En el caso de la población hipotética de más arriba, y supuesto que el tamaño m de la muestra es siempre igual a 2, cualquiera sea el grupo obtenido en la primera etapa del muestreo, se tiene:

$$\frac{1}{N}\sum_{1}^{N}\left(\frac{1}{m_i} - \frac{1}{M_i}\right) s_i^2 = 2.02$$

$$\frac{1}{N}\sum_{1}^{N}(\bar{\bar{a}}_i - \bar{\bar{a}}_i)^2 = .66$$

$$(\bar{\bar{a}}_i - \bar{\bar{a}})^2 = .08$$

de modo que:

$$\hat{V}(\bar{\bar{a}}_I) = \underline{2.76}$$

2º Estimación basada en el total $M_i \bar{y}_i$.-

Sea $\bar{y}_i$ la media de la muestra extraída del i-ésimo grupo obtenido en la primera etapa de muestreo; $M_i \bar{y}_i$ es una estimación, no viciada del total de dicho grupo, y $NM_i\bar{y}_i$ una estimación, también "no viciada" del total de la población. Se sigue de aquí que:

M.t.p. V

$$\hat{\bar{a}}_{II} = \frac{N M_i \bar{y}_i}{M_o} = \frac{M_i y_i}{\bar{M}} \qquad (\bar{M} = \frac{M_o}{N})$$

será una estimación "no viciada" de $\bar{a}$.- La afirmación anterior puede probarse del siguiente modo:

$$E(\hat{\bar{a}}_{II}) = \frac{1}{M} \sum_1^N \sum_1^{M_i} p_{(ij)} M_i \bar{y}_i = \frac{1}{M} \sum_1^N p_i M_i \left[\sum_1^{M_i} p_{(j/i)} \bar{y}_i \right] =$$

$$\frac{1}{\bar{M}} \sum_1^N \frac{1}{N} M_i \bar{a}_i = \frac{1}{M_o} \sum_1^N M_i \bar{a}_i = \bar{\bar{a}}$$

Para hallar la variancia de la estimación debe calcularse:

$$E\left(\frac{M_i \bar{y}_i}{\bar{M}} - \bar{\bar{a}}\right)^2 = E\left(\frac{M_i \bar{y}_i}{\bar{M}} - \frac{M_i \bar{a}_i}{\bar{M}} + \frac{M_i \bar{a}_i}{\bar{M}} - \bar{\bar{a}}\right)^2 =$$

$$= E\left\{ \frac{M_i^2}{M_2} (\bar{y}_i - \bar{a}_i)^2 \right\} + E\left\{ \frac{N^2}{M_o^2} (M_i \bar{a}_i - \frac{1}{N} \sum M_i \bar{a}_i)^2 \right\} \qquad (5)$$

(la E.M. del doble producto es nula).-

El primer sumando de (5) es:

$$\frac{1}{\bar{M}^2} \sum_1^N p_{(i)} M_i \left[\sum_1^{M_i} p_{(j/i)} (\bar{y}_i - \bar{a}_i)^2 \right] = \frac{N}{M_o^2} \sum_1^N M_i \left(\frac{1}{M_i} \frac{1}{M_i}\right) S_i^2 \qquad (6)$$

el 2º sumando de (5) es:

$$\frac{N^2}{M_o^2} \sum_1^N p_i (M_i \bar{a}_i - \frac{1}{N} \sum M_i \bar{a}_i)^2 = \frac{N}{M_o^2} \sum_1^N (M_i \bar{a}_i - \frac{1}{N} \sum M_i \bar{a}_i)^2 \qquad (7)$$

y llevando estos valores (6) y (7) a la (5), queda:

$$V(\bar{a}_{II}) = \frac{N}{M_o^2} \sum_1^N \frac{M_i - m_i}{m_i} S_i^2 + \frac{N}{M_o^2} \sum_1^N (M_i \bar{a}_i - \frac{1}{N} \sum M_i \bar{a}_i)^2 \qquad (8)$$

Merece observarse que la componente "entre grupos" de la variancia depende de los cuadrados de los desvíos de los totales $M_i \bar{a}_i$ de los distintos grupos con respecto a la media aritmética de esos totales, lo que hace que, en general, la variancia de la estimación $\bar{a}_{II}$ sea grande y, en particular, mayor que la de la estimación "viciada" $\bar{a}_I$.

En el caso de la población de más arriba las componentes "dentro" y "entre" grupos son, respectivamente:

.32 4.99

de modo que $V(\bar{a}_{II}) = 5.31$

3º) Selección del grupo con probabilidad proporcional a su tamaño.

En los dos casos anteriores la selección de la unidad primaria de donde se extraía la muestra de elementos se hacía de modo tal que todos tenían la misma probabilidad 1/N de ser escogidos en la primera etapa del muestreo. Supondremos ahora que esa selección se hace de modo tal que cada unidad primaria tenga una probabilidad de ser escogida proporcional a su tamaño M_i.

Para mostrar cómo se hace esa selección, con los datos de la población hipotética de la pág. construímos el siguiente cuadro:

Grupo Nº	M_i	Acum.	Intervalo
1	3	3	1 - 3
2	6	9	4 - 9
3	8	17	10 - 17

Elegido un número al azar 1 y 17, se toma para extraer la muestra de elementos aquella unidad primaria que tiene asociado el intervalo de números en el que se encuentra aquel. Así, si se ha encontrado el número 8, la unidad primaria escogida para la songunda etapa del muestreo es la Nº 2.

Como antes, una vez tenido el grupo, se extrae de él una muestra simple al azar de m_1 elementos.

Para estimar la media por elemento en la población tomaremos la:

$$\bar{a}_{III} = \bar{y}_1$$

es decir, la media de la muestra de elementos extraída del grupo que se ha obtenido en la primera etapa de muestreo en la que se hizo la selección asignando a cada uno una probabilidad M_1/M_0.

La esperanza matemática de esta estimación es:

$$E(\bar{a}_{III}) = \sum_{1}^{N}\sum_{1}^{M_1} P(1j)\,\bar{y}_1 = \sum_{1}^{N} P(1)\left(\sum_{1}^{M_1} P(j/1)\,\bar{y}_1\right)$$

Aquí, como más arriba es $P(j/1) = 1/C_{n_1}^{M_1}$, pero $P(1) = M_1/M_0$, de modo que:

$$E(\bar{a}_{III}) = \frac{1}{M_0}\sum_{1}^{N} M_1\bar{a}_1 = \bar{a}$$

lo que muestra que la estimación es "no viciada".

La varianza de la estimación será:

$$E(\bar{y}_1 - \bar{a})^2 = E(\bar{y}_1 - \bar{a}_1)^2 \quad E(\bar{a}_1 - \bar{a})^2$$

(puesto que $E(\bar{y}_1 - \bar{a}_1)(\bar{a}_1 - \bar{a}) = 0$) y se tendrá:

$$V(\bar{a}_{III}) = \sum_{1}^{N}\sum_{1}^{N_1} p_{1j}(\bar{y}_1 - \bar{a}_1)^2 \text{ y } \sum_{1}^{N} P(1)(\bar{a}_1 - \bar{a})^2 = \frac{1}{M_0}\sum_{1}^{N} M_1\left(\frac{1}{m_1}\frac{1}{n_1}\right)S_1^2$$

$$+ \frac{1}{M_0}\sum_{1}^{N} M_1\,(\bar{a}_1 - \bar{a})^2$$

En el caso del ejemplo se tiene:

$$V(\bar{a}_{III}) = 2{,}14$$

4º) Selección de la unidad primaria con probabilidad proporcional al tamaño estimado

En la práctica es el caso más común el que no se tenga una información exacta y actualizada del número de elementos que integran cada unidad primaria que permita determinar la correcta probabilidad proporcional al tamaño que se asociará a cada una para su selección en la primera etapa de muestreo. Tal es el caso, p.e., en que las unidades primarias están constituídas por manzanas o grupos de manzanas de una ciudad, siendo los elementos las viviendas en ellas ubicadas. En un caso como éste es seguro que al momento de diseñar una muestra no se cuenta con un registro exacto del número de viviendas de cada unidad, pero sí con alguna información que ofrezca una estimación de ese número que podrá utilizarse para determinar una probabilidad de selección que indicaremos con P_i ($\sum P_i = 1$).

Una primera estimación de $\bar{a}$ que es posible construir cuando la selección de las unidades primarias se hace con probabilidades P_i, es:

$$\hat{\bar{a}}_{IV} = \frac{M_i \bar{y}_i}{P_i M_o}$$

Esta estimación es "no viciada" puesto que:

$$E(\hat{\bar{a}}_{IV}) = \sum_1^N P_i \frac{M_i}{P_i M_o} \left\{ \sum_1^{M_i} p_{(j/i)} \bar{y}_i \right\} = \frac{1}{M_o} \sum_1^N M_i \bar{a} = \bar{a}$$

El valor de m_i se determina generalmente tomando

$$m_i = k \frac{M_i}{P_i}$$

siendo k una constante que se describe como siendo la "tasa global de muestreo" y esto debido a que, siendo

$$E(m_i) = \sum P_i m_i = k \sum M_i = k M_o$$

se tiene que

$$k = \frac{E(m_i)}{M_o}$$

es decir, k expresa la relación entre el valor esperado del tamaño de la muestra a tomarse del grupo elegido en la primera etapa de muestreo, al número total de elementos en la población.

La variancia de la estimación se obtiene a partir de

$$E(\hat{\bar{a}}_{IV} - \bar{a})^2 = E\left\{ \frac{M_i \bar{y}}{p_i M_o} - \frac{M_i \bar{a}_i}{P_i M_o} \quad \frac{M_i a_i}{P_i M_o} - \bar{a} \right\}^2 =$$

$$E\left\{ \frac{M_i}{P_i M_o} \quad (\bar{y}_i - \bar{a}_i) \right\}^2 + E\left\{ \frac{1}{M_o} \quad \left(\frac{M_i \bar{a}_i}{P_i} - \sum M_i \bar{a}_i \right) \right\}^2 =$$

$$\frac{1}{M_o^2} \sum_{1}^{N} \frac{M_i^2}{P_i} \left(\frac{1}{m_i} - \frac{1}{M_i} \right) S_i^2 + \frac{1}{M_o^2} \sum_{1}^{N} P_i \left(\frac{M_i \bar{a}_i}{P_i} - \sum M_i \bar{a}_i \right)^2$$

Si en la anterior ponemos:

$$\frac{m_i P_i}{M_i} = k$$

e indicamos el valor total de la característica a en cada grupo por A_i de modo que

$$M_i \bar{a}_i = A_i \qquad \sum M_i \bar{a}_i = A$$

se tiene:

$$V(\hat{\bar{a}}_{IV}) = \frac{1}{M_o^2} \sum_{1}^{N} \frac{(M_i - m_i)}{K} S_i^2 + \frac{1}{M_o^2} \sum_{1}^{N} P_i \left(\frac{A_i}{P_i} - A \right)^2$$

Una segunda estimación sería:

$$\hat{\bar{a}}_V = \bar{y}_i$$

Que es "viciada", y cuyo error medio cuadrático se obtiene calculando:

$$E(\hat{\bar{a}}_V - \bar{a}) = E(\bar{y}_i - \bar{a}_i + \bar{a}_i - \bar{a}_1 + \bar{a}_1 - \bar{a})^2 =$$

$$= E(\bar{y}_i - \bar{a}_i)^2 + E(\bar{a}_i - \bar{\bar{a}}_i)^2 + (\bar{\bar{a}}_i - \bar{\bar{a}})^2$$

Fácilmente, siguiendo la misma línea de razonamiento utilizado repetidamente más arriba, se encuentra que:

$$E(\bar{y}_i - \bar{a}_i)^2 = \sum_1^N P_i \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2$$

$$E(a_i - \bar{\bar{a}}_i)^2 = \sum_1^N P_i (\bar{a}_i - \bar{\bar{a}}_i)^2$$

de modo que

$$V(\hat{\bar{a}}_V) = \sum_1^N P_i \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2 + \sum_1^N P_i (\bar{a}_i - \bar{\bar{a}}_i)^2 + (\bar{\bar{a}}_i - \bar{\bar{a}})^2$$

En el cuadro siguiente se resumen todos los resultados anteriores.

MUESTREO POR GRUPOS EN 2 ETAPAS

UNIDADES PRIMARIAS DE DIFERENTE TAMAÑO

1a. etapa: $n = 1$ 2a. etapa: m_i

Estimaciones de la media por elemento en la población

Nº	Prob. unid.	selección primaria	Estimación	Varianza
I		1/N	$\bar{y}_i$ viciada	$\frac{1}{N}\sum_1^N \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2 + \frac{1}{N}\sum_1^N (\bar{a}_i - \bar{\bar{a}}_i)^2 + (\bar{\bar{a}}_i - \bar{\bar{a}})^2$
II		1/N	$M_i \bar{y}_i / \bar{M}$ no viciada	$\frac{N}{M_o^2}\sum_1^N M_i \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2 + \frac{N}{M_o}\sum_1^N \frac{1}{N} M_i \bar{a}_i$

Nº	Prob. selección Unid. primaria	Estimación	Variancia
III	M_i/M_o	$\bar{y}_i$ no viciada	$\frac{1}{M_o}\sum_1^N M_i\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_i^2+\frac{1}{M_o}\sum_1^N(\bar{a}_i-\bar{\bar{a}})^2$
IV	P_i	$M_i\bar{y}_i/P_iM_o$ no vic.	$\frac{1}{M^2}\sum_1^N\frac{M_i-m_i}{k}S_i^2+\frac{1}{M^2}\sum_1^N P_i\left(\frac{A_i}{P_i}-A\right)^2$
V	P_i	$\bar{y}_i$ viciada	$\sum P_i\left(\frac{1}{m_i}-\frac{1}{M_i}\right)S_i^2+\sum P_i(\bar{a}_i-\bar{\bar{a}}_i)^2+(\bar{\bar{a}}_i-\bar{\bar{a}})^2$

3º) Unidades primarias de tamaños diferentes.- $n > 1$.-

Supongamos ahora que, en la primera etapa del muestreo se seleccionen no ya una, sino n unidades primarias, de cada una de las cuales se extrae en una segunda operación, una muestra simple al azar de m_i elementos.- La muestra de unidades primarias es también una muestra simple al azar, de modo que la probabilidad de que una cualquiera de ellas sea incluída en la muestra es n/N.-

Pueden también, en este caso, construirse varias estimaciones de la media por elemento en la población.

a) Media aritmética de las medias de las muestras de los grupos.-

Esta viene dada por:

$$\hat{\bar{a}}_I = \frac{1}{n} \sum_{1}^{n} \bar{y}_i \qquad (1)$$

donde $\bar{y}_i$ es la media de la muestra de m_i elementos, extraída del i-ésimo grupo obtenido en la primera etapa del muestreo.-

La E.M. de esta estimación es:

$$E(\hat{\bar{a}}_I) = \frac{1}{n} \sum_{1}^{n} E\,\bar{y}_i = \frac{1}{n} \sum_{1}^{n} \left\{ \sum_{1}^{N} \sum_{1}^{M_i} p(ij)\bar{y}_i \right\} =$$

$$= \frac{1}{n} \sum_{1}^{n} \left\{ \sum_{1}^{N} P_i \left(\sum_{1}^{M_i} p(j/i)\bar{y}_i \right) \right\}$$

Pero

$$p_i = \frac{1}{N} \qquad y \qquad \sum_{1}^{M_i} p(j/i)\bar{y}_i = \bar{a}_i$$

de modo que

$$E(\hat{\bar{a}}_I) = \frac{1}{N} \sum_{1}^{N} \bar{a}_i = \bar{\bar{a}} \neq \bar{a}$$

lo que muestra que la estimación es "viciada".-

Para determinar el error medio cuadrático, partimos de la identidad algebraica:

$$\frac{1}{n}\sum_{1}^{n} \bar{y}_i - \bar{\bar{a}} = \frac{1}{n}\sum_{1}^{n} \bar{y}_i - \frac{1}{n}\sum_{1}^{n} \bar{a}_i - \frac{1}{n}\sum_{1}^{n} \bar{a}_i - \frac{1}{N}\sum_{1}^{N} \bar{a}_i +$$

$$+ \frac{1}{N}\sum_{1}^{N} \bar{a}_i - \bar{\bar{a}}$$

lo que, teniendo en cuenta que

$$\frac{1}{N}\sum_{1}^{N} \bar{a}_i = \bar{\bar{a}}_i$$

y poniendo

$$\frac{1}{n}\sum_{1}^{n} \bar{a}_i = \bar{\bar{a}}_{i\,(n)}$$

puede también escribirse:

$$\frac{1}{n}\sum_{1}^{n} \bar{y}_i - \bar{\bar{a}} = \frac{1}{n}\sum_{1}^{n} (\bar{y}_i - \bar{a}_i) + (\bar{\bar{a}}_{i(n)} - \bar{\bar{a}}_i) + (\bar{\bar{a}}_i - \bar{\bar{a}})$$

y la E.M. que se debe calcular es:

$$E(\hat{\bar{\bar{a}}}_I - \bar{\bar{a}})^2 = \frac{1}{n^2}\left\{ E\left[\sum_{1}^{n} (\bar{y}_i - \bar{a}_i)\right]^2 + E(\bar{\bar{a}}_{i\,(n)} - \bar{a}_i)^2 + \right.$$

$$\left. + (\bar{\bar{a}}_i - \bar{\bar{a}})^2 \right. \qquad (2)$$

puesto que las E.M. de los dobles productos son todas nulas.-

El primer sumando del 2º miembro de la (2) es

$$\frac{1}{n^2} E\left[\sum_{1}^{n} (\bar{y}_i - \bar{a}_i)^2 + \sum_{i \neq j}\sum (\bar{y}_i - \bar{a}_i)(\bar{y}_j - \bar{a}_j)\right] = \frac{1}{n^2} E\sum_{1}^{n} (\bar{y}_i - \bar{a}_i)^2$$

puesto que, siendo independientes las muestras que se extraen de grupos diferentes, la E.M. de la suma doble es nula.-

Ahora:

$$\frac{1}{n^2} E \sum_{1}^{n} (\bar{y}_i - \bar{a}_i)^2 = \frac{1}{n} \sum_{1}^{n} E(\bar{y}_i - \bar{a}_i)^2 =$$

$$= \frac{1}{n^2} \sum_{1}^{n} \left\{ \sum_{1}^{N} \sum_{1}^{M_i} p(ij)(\bar{y}_i - \bar{a}_i)^2 \right\} =$$

$$= \frac{1}{n^2} \sum_{1}^{n} \left\{ \sum_{1}^{N} p(i) \left[\sum_{1}^{M_i} p(j/i)(\bar{y}_i - \bar{a}_i)^2 \right] \right\}$$

y como

$$p(i) = \frac{1}{N} \sum_{1}^{M_i} p(j/i)(\bar{y}_i - \bar{a}_i)^2 = \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2$$

resulta finalmente

$$\frac{1}{n^2} E \left[\sum_{1}^{n} (\bar{y}_i - \bar{a}_i) \right]^2 = \frac{1}{nM} \sum_{1}^{N} \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2 \qquad (3)$$

Por su parte, el 2º sumando de la (2), puesto que se trata de la esperanza matemática de los cuadrados de los desvíos de la media de una muestra simple al azar de extensión n con respecto a la media de la población es:

$$\left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 \qquad (4)$$

donde

$$S_b^2 = \frac{1}{N-1} \sum_{1}^{N} (\bar{a}_i - \bar{\bar{a}})^2$$

Reemplazando en (2) los valores hallados en (3) y (4) se tiene

$$V(\hat{\bar{a}}_I) = \frac{1}{nN} \sum_1^N \left(\frac{1}{m_i} - \frac{1}{M_i}\right) S_i^2 \quad \left(\frac{1}{n} - \frac{1}{N}\right) S_b^2 + (\bar{a}_i - \bar{\bar{a}})^2 \qquad (5)$$

b) Estimación basada en $M_i \bar{y}_i$.-

Esta segunda estimación es:

$$\hat{\bar{a}}_{II} = \frac{N}{nM_o} \sum_1^n M_i \bar{y}_i$$

donde, como antes, $\bar{y}_i$ es la media de la muestra de m_i elementos extraída del i-ésimo grupo obtenido en la muestra de n unidades primarias de entre las N en que se agrupan los elementos de la población.-

La E.M. de la estimación es:

$$E(\hat{\bar{a}}_{II}) = \frac{N}{nM_o} \sum_1^n E(M_i \bar{y}_i)$$

Ahora,

$$E(M_i \bar{y}_i) = \sum_1^N \sum_1^{M_i} p(ij) M_i \bar{y}_i = \sum_1^N p(i) M_i \sum_1^{M_i} p(j/i) \bar{y}_i =$$

$$= \frac{1}{N} \sum_1^N M_i \bar{a}_i = \frac{1}{N} M_o \bar{\bar{a}}$$

de modo que

$$E(\hat{\bar{a}}_{II}) = \bar{\bar{a}}$$

lo que muestra que tenemos ahora una estimación "no viciada".-

Para calcular la variancia partimos de:

$$\frac{N}{nM_o} \sum_1^n M_i \bar{y}_i - \bar{\bar{a}} = \frac{N}{nM_o} \sum_1^n M_i \bar{y}_i - \frac{N}{nM_o} \sum_1^n M_i \bar{a}_i + \frac{N}{nM_o} \sum_1^n M_i \bar{a}_i - \bar{\bar{a}} =$$

$$= \frac{N}{n M_o} \sum_{1}^{n} M_i(\bar{y}_i - \bar{a}_i) + \frac{N}{M_o}\left[\frac{1}{n}\sum_{1}^{n} M_i\bar{a}_i - \frac{1}{N}\sum_{1}^{N} M_i\bar{a}_i\right] \qquad (6)$$

Aquí, $\frac{1}{N}\sum_{1}^{N} M_i\bar{a}_i$ es la media por unidad primaria en la población, y $\frac{1}{n}\sum_{1}^{n} M_i\bar{a}_i$ la media análoga en la muestra.- Si indicamos estos dos valores con A_i y $\bar{A}_i\,(n)$, respectivamente, la (6) puede escribirse

$$\hat{\bar{a}}'_{II} - \bar{a} = \frac{N}{nM_o}\sum_{1}^{n} M_i(\bar{y}_i - \bar{a}_i) + \frac{N}{M_o}\left[\bar{A}_{i(n)} - \bar{A}_i\right]$$

y la E.M. que tenemos que calcular es:

$$E(\hat{\bar{a}}'_{II} - \bar{a})^2 = \frac{N^2}{n^2 M_o^2} E\left\{\sum_{1}^{n} M_i(\bar{y}_i - \bar{a}_i)^2 + \frac{N^2}{M_o^2} E\left[\bar{A}_{i(n)} - \bar{A}_i\right]^2\right. \qquad (7)$$

Siguiendo un camino análogo al usado repetidamente más arriba, se encuentra inmediatamente que:

$$E\left\{\sum_{1}^{n} M_i(\bar{y}_i - \bar{a}_i)\right\}^2 = n\frac{1}{N}\sum_{1}^{N} M_i^2\left(\frac{1}{m_i} - \frac{1}{M_i}\right)s_i^2$$

La E.M. en el 2do. sumando en la (7) es:

$$\frac{N^2}{M_o^2} E\left[\bar{A}_i(n) - \bar{A}_i\right]^2 = \frac{N^2}{M_o^2}\left(\frac{1}{n} - \frac{1}{N}\right)s'^2_b$$

donde

$$S_b^2 = \frac{1}{N-1} \sum_{1}^{N} \left(M_i \bar{a}_i - \frac{1}{N} \sum_{1}^{N} M_i \bar{a}_i \right)^2$$

de modo que se tiene finalmente:

$$\hat{V}(\bar{a}_{II}) = \frac{1}{nN} \sum_{1}^{N} \frac{M_i^2}{\bar{M}^2} \left(\frac{1}{m_i} - \frac{1}{M_i} \right) s_i^2 + \frac{1}{\bar{M}^2} \left(\frac{1}{n} - \frac{1}{N} \right) s_b^2 \qquad (8)$$

(donde $\bar{M} = M_o/N$, tamaño medio de los grupos).-

c) Si las unidades primarias se seleccionan con probabilidades proporcionales a su tamaño, y la extensión m de la muestra que se extrae de cada uno de los grupos obtenidos en esa primera etapa de muestreo es constante, la media de la muestra es una estimación "no viciada" de la media de la población.-

En efecto, si las unidades primarias se eligen "con reposición" y es r_i el número de veces que la i-ésima unidad se presenta en una muestra específica, y puesto que

$$E(r_i) = nP_i$$

donde P_i es la probabilidad de selección de la i-ésima unidad primaria, se tiene:

$$E(\bar{a}_{II}) = \frac{1}{n} E \sum_{1}^{N} r_i \bar{v}_i = \frac{1}{n} \sum_{1}^{N} \frac{m M_i \bar{a}_i}{M} = \bar{a}$$

Cuando la selección de las unidades primarias se hace con probabilidades proporcionales al tamaño estimado, una estimación "no viciada", cualesquiera sean los valores que se asignen a los m_i es:

$$\bar{\bar{y}}_{IV} = \frac{1}{nM_o} \sum_{1}^{n} \frac{M_i}{P_i} \bar{y}_i$$

Se verifica, en efecto, fácilmente, que:

$$E(\bar{\bar{y}}_{IV}) = \frac{1}{nM_o} E\left(\sum_{1}^{N} \frac{r_i M_i}{P_i} \bar{y}_i \right) = \frac{1}{nM_o} \sum_{1}^{N} nM_i \bar{a}_i = \bar{\bar{a}}$$

Si se toma

$$m_i = \frac{KM_i}{P_i}$$

la anterior estimación es "auto-ponderada", puesto que se reduce a:

$$\bar{\bar{y}}_{IV} = \frac{1}{nkM_o} \sum_{1}^{n} m_i \bar{y}_i = \frac{y}{nkM_o}$$

siendo $y = \sum_{1}^{n} m_i \bar{y}_i$

Vimos más arriba cuando se consideró el caso $n = 1$, que k era la "tasa global de muestreo". Si $n > 1$, esta tasa es nk, puesto que se tiene:

$$k = \frac{m_i P_i}{M_i} = \frac{\sum_{1}^{N} m_i P_i}{M_o}$$

Pero el número medio de elementos en la muestra es

$$E\left(\sum_{1}^{n} m_i \right) = E\left(\sum_{1}^{N} r_i m_i \right) = n \sum_{1}^{N} P_i m_i$$

de modo que resulta

$$nk = \frac{1}{\bar{M}} \times \text{(Número medio de elementos en la muestra)}$$

La "tasa global de muestreo", que es la relación del número de elementos en la muestra al número de elementos en la población, debe distinguirse tanto de la tasa de muestreo n/N de unidades primarias, como de la tasa m_i/M_i de sub-muestreo de elementos en cada grupo.

Cuando $n > 1$, varias son las extensiones posibles de la estimación que más arriba hemos indicado con $\hat{\bar{a}}_V$ (siendo P_i la probabilidad de selección de la i-ésima unidad primaria); una de ellas es la media ponderada de la muestra:

$$\hat{\bar{a}}_V = \left(\sum_1^n \frac{M_i \bar{y}_i}{P_i} \right) \Big/ \left(\sum_1^n \frac{M_i}{P_i} \right)$$

El numerador es una estimación "no viciada" de

$$n \sum_1^N M_i \bar{a}_i$$

y el denominador una estimación también "no viciada" de

$$n \sum_1^N M_i$$

Si $m_i = kM_i/P_i$ como en la IV, esta estimación se convierte en la media simple no ponderada de la muestra: $\bar{y}$.-

Para obtener la variancia de algunas de las estimaciones precedentes, partiremos de la expresión de una estimación por cociente dada por:

$$\hat{R} = \left(\sum_{1}^{n} \frac{M_i \bar{y}_i}{P_i} \right) \Big/ \left(\sum_{1}^{n} \frac{M_i \bar{x}_i}{P_i} \right)$$

la cual, si x_{ij} se toma como variable auxiliar que es igual a 1 para todo elemento de la población -de modo que $\bar{x}_i = 1$ - se reduce a $\hat{\bar{a}}'_V$, la que, a su vez, para particulares valores de P_i, se reduce a otras de las estimaciones consideradas más arriba.

La $\hat{R}$ puede ser de utilidad en sí misma en el caso en que X_{ij} es el valor de la característica que se estudia en un censo previo, en cuyo caso $\hat{R}$ se utilizaría para hacer la estimación de valores medios o totales de dicha característica usando de esa información previa.

No entraremos en el proceso de derivación de la variancia de $\hat{R}$, limitándonos a dar el resultado.

Si se extrae una muestra con reposición de n unidades primarias, siendo P_i la probabilidad de selección asignada a la i-ésima unidad, y de cada una de las unidades elegidas se toma una sub-muestra simple al azar de extensión m_i, y se construye la estimación:

$$\hat{R} \quad \left(\sum_{1}^{n} \frac{M_i}{P_i} \bar{y}_i \right) \Big/ \left(\sum_{1}^{n} \frac{M_i}{P_i} \bar{x}_i \right)$$

si el tamaño de la muestra es grande, resulta:

$$V(\hat{R}) \doteq \frac{1}{nB^2} \sum_{1}^{N} \left[\frac{1}{P_i} (A_i - RB_i)^2 + \frac{M_i (M_i - m_i)}{P_i} \frac{S^2_{di}}{m_i} \right]$$

donde

$$S^2_{di} = \frac{1}{(M_i - 1)} \sum_{1}^{M_i} \left[(a_{ij} - Rb_{ij}) - (\bar{a}_i - R\bar{b}_i) \right]^2$$

Si en la anterior fórmula se toma $m_i = KM_i/P_i$, de modo que $\hat{R}$ se reduce a

$\sum m_i \bar{y}_i / \sum m_i \bar{x}_i$, resulta:

$$V(\hat{R}) \doteq \frac{1}{nB_o^2} \sum_1^M \left[\frac{1}{P_i} (A_i - RB_i)^2 + \frac{M_i - m_i}{k} S_{di}^2 \right] \qquad a)$$

De esta expresión pueden obtenerse las fórmulas de la variancia aplicables a las estimaciones de subíndices V y III respectivamente, según el valor que se asigna a b_{ij} y a P_i.-

Para el mismo método de selección descripto anteriormente, la variancia de la estimación

$$\bar{\bar{y}}_{IV} = \frac{1}{nM_o} \sum_1^n \frac{M_i}{P_i} \bar{y}_i$$

viene dada por

$$V(\bar{\bar{y}}_{IV}) = \frac{1}{nM_o^2} \left[\frac{1}{P_i} (A_i - P_i A)^2 + \frac{M_i (M_i - m_i)}{P_i} \frac{S_i^2}{m_i} \right] \qquad b)$$

donde A es el valor total de la característica estudiada en la población.-

Si se toma $m_i = kM_i/P_i$, de modo que

$$\bar{\bar{y}}_{IV} = \sum_1^n m_i \bar{y}_i / nkM_o$$

resulta:

$$V(\bar{\bar{y}}_{IV}) = \frac{1}{nM_o^2} \sum_1^N \frac{1}{P_i} (A_i - P_i A)^2 + \frac{M_i - m_i}{k} S_i^2 \Bigg]$$

La estimación de la variancia a) calculada a partir de los datos de la muestra viene dada por:

$$v(\hat{R}) = \frac{1}{n(n-1) B^2} \sum_1^n \left[\frac{M_i}{P_i m_i} (y_i - Rx_i) \right]^2$$

donde $\hat{R}$ es la estimación de R basada en los datos ofrecidos por la muestra, y

$$\hat{B} = \frac{1}{m} \sum_1^n \frac{M_i \bar{x}_i}{P_i}$$

es una estimación "no viciada" del total de b_{ij} en la población.-

A su vez, la estimación de $V(\bar{\bar{y}}_{IV})$ viene dada por:

$$v(\bar{\bar{y}}_{IV}) = \frac{1}{n(N-1)\,M_o^2}\sum_{1}^{n}(\hat{Y}_i - \hat{\bar{Y}}_n)^2$$

donde

$$\hat{Y} = \frac{M_i\,\bar{y}_i}{P_i} \qquad y \qquad \hat{\bar{Y}}_n = \frac{1}{n}\sum_{1}^{n}\hat{Y}_i = M_o\,\bar{\bar{y}}_{IV}$$

Probabilidades óptimas de selección.-

Un problema que de modo natural se plantea en el caso del muestreo en dos etapas con probabilidades variables de selección es: ¿dados los tamaños de las unidades primarias —o buenas estimaciones de las mismas-qué probabilidades de selección deben adjudicarse a esas unidades para hacer mínima la variancia para un costo total dado?.-

Este problema fué estudiado y resuelto por Hanson y Hurwitz.- A pesar de su interés, no nos detendremos en él.- Nos interesa sólo mencionarlo destacando que la respuesta depende de que se tenga suficiente información previa de ciertas características de la población a estudiar y estimaciones más o menos correctas acerca de los diferentes elementos que intervienen en la formación del costo total de la operación de muestreo.

Sub-muestreo estratificado.-

El muestreo en dos etapas puede aplicarse evidentemente al caso en que la población haya sido previamente clasificada en un cierto número de estratos.- Este caso no tiene ninguna dificultad conceptual; si la tiene de notación.-

Supondremos que la población se ha dividido en k estratos, cada uno de los cuales consta de N_h $(h=1,2,\ldots, k)$ unidades primarias de muestreo, siendo

$$\sum_{1}^{k} N_h = N$$

Ahora, una unidad primaria en un cierto estrato, p.e, la i-ésima del h-ésimo estrato consta de

M_{hi} unidades secundarias o elementos.

El valor de la característica estudiada en el j-ésimo elemento de la i-ésima unidad primaria del h-ésimo estrato es:

$$a_{hij} \qquad \begin{matrix} h = 1 \dots k \\ i = 1 \dots N_h \\ j = 1 \dots M_{hi} \end{matrix}$$

El número total de elementos en el h-ésimo estrato es

$$M_h = \sum_{1}^{N_h} M_{hi} = N_h \bar{M}_h$$

En lo que respecta a la muestra, indicaremos con:

n_k el número de unidades primarias del h-ésimo estrato incluídas en la muestra ($\sum n_h = n$)

m_{hi} el número de elementos de la i-ésima unidad primaria del h-ésimo estrato obtenidas en el 2a. etapa del muestreo

($\sum m_{hi} = m_h$)

m_h es el número de elementos de la muestra provenientes del h-ésimo estrato.

El valor total de la característica en la población es:

$$\sum_{1}^{k} \sum_{1}^{N_h} \sum_{1}^{M_{hi}} a_{hij}$$

Medias.- El valor medio de los valores de los elementos en la i-ésima unidad primaria del h-ésimo estrato (que son en número de M_{hi}) es:

$$\bar{a}_{hi} = \frac{1}{M_{hi}} \sum_{1}^{M_{hi}} a_{hij}$$

En el h-ésimo estrato, el total de la característica estudiada es:

$$\sum_{1}^{N_h} \sum_{1}^{M_{hi}} a_{hij}$$

y como hay en él $\sum_{1}^{N_h} M_{hi} = M_h$ unidades primarias, se tiene:

$$\bar{a}_h = \frac{1}{M_h} \sum_{1}^{N_h} \sum_{1}^{M_{hi}} a_{hij} = \frac{1}{M_h} \sum_{1}^{N_h} M_{hi} \bar{a}_{hi}$$

que es la medias por elemento en el h-ésimo estrato.

El número total de elementos en la población es

$$\sum_{1}^{k} \sum_{1}^{N_h} M_{hi}$$

de modo que la media por elemento en la población es:

$$\frac{\sum\sum\sum a_{hij}}{\sum\sum M_{hi}} = \frac{\sum\sum\sum a_{hij}}{\sum M_h} = \frac{\sum M_h \bar{a}_h}{\sum M_h} = \bar{\bar{a}}$$

lo que puede indicarse con

$$\sum p_h \bar{a}_h$$

si se pone

$$p_h = M_h / \sum M_h$$

Para la muestra usaremos la siguiente notación:

y_{hij} valor del j-ésimo elemento obtenido de la h-ésima unidad primaria en el k-ésimo estrato.

$\sum_{1}^{m_{hi}} y_{hij}/m_{hi} = \bar{y}_{hi}$ media por elemento en la muestra del i-ésimo grupo del h-ésimo estrato.

$\frac{M_{hi}}{m_{hi}} \sum y_{hij}$ total estimado de la característica en el i-ésimo grupo del h-ésimo estrato.

$\sum_{1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum y_{hij}$ total estimado en los n_h grupos en la muestra de unidades primarias del h-ésimo estrato.

$\frac{N_h}{n_h} \sum_{1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum y_{hij}$

total estimado de la característica en el h-ésimo estrato.

$\bar{y}_h = \frac{1}{M_h} \cdot \frac{N_h}{n_h} \sum_{1}^{n_h} \frac{M_{hi}}{m_{hi}} \sum_{1}^{m_h} y_{hij}$

promedio por elemento en el h-ésimo estrato.

Este valor desempeña en el caso que estamos considerando el mismo papel que $\bar{y}_h$ en el muestreo estratificado. Recordemos que en aquel caso la estimación de la media de la población venía dada por

$$\bar{y}_e = \frac{1}{N}\sum N_h \bar{y}_h$$

donde N_h y N eran, respectivamente, el número de elementos en el h-ésimo estrato y en la población total.

En el caso actual, será pués

$$\bar{\bar{y}}_e = \frac{\sum M_h \bar{y}_h}{\sum M_h} = \sum_1^k p_h \bar{y}_h$$

la estimación de la media por elemento en la población.

Recordemos ahora que, si es

$$z = \sum c_i x_i$$

es decir, la variable aleatoria z es una combinación lineal de las variables aleatorias x_i que tienen variancia $V(x_i)$, se tiene:

$$V(z) = \sum c_i^2 V(x_i)$$

Este fué el teorema que aplicamos para hallar la variancia de la estimación en el muestreo simple estratificado.

En el caso actual será

$$V(\bar{\bar{y}}_e) = \sum p_h^2 \cdot V(\bar{y}_h)$$

donde $V(\bar{y}_h)$ es la variancia de la media por elemento obtenida en una muestra de m_{hi} elementos extraídos de n_h unidades primarias obtenidas en la muestra proveniente del h-ésimo estrato, es decir, la variancia de la estimación de la medida por elemento en el muestreo por grupos en 2 etapas.

Recordamos que mas arriba estudiamos la estimación

$$\frac{N\sum_1^n M_i \bar{y}_i}{n\sum M_i}$$

Ahora tenemos

$$\frac{N_h \sum M_{hi} \bar{y}_{hi}}{n_h \sum M_{hi}}$$

de modo que resulta

$$V(\bar{\bar{y}}_e) = \sum_{1}^{k} P_h^2 \left\{ \left(\frac{1}{n_h} - \frac{1}{N_h} \right) S'^2_{hb} + \frac{1}{n_h N_h} \sum_{1}^{N_h} \frac{M_{hi}^2}{\bar{M}_h^2} \left(\frac{1}{m_{hi}} - \frac{1}{M_{hi}} \right) S^2_{hi} \right\}$$

donde

$$S'^2_{hb} = \frac{1}{N_h - 1} \sum \left(\frac{M_{hi}}{\bar{M}_h} \bar{a}_{hi} - \bar{\bar{a}}_h \right)^2 \qquad S^2_{hi} = \frac{1}{N_{hi} - 1} \sum (a_{hij} - \bar{a}_{hi})^2$$