MUESTREO ESTADÍSTICO

Muestreo Aleatorio Estratificado (M.A.E)

SEMANA-7

Raúl Alberto Pérez

Universidad Nacional de Colombia, Escuela de Estadística, 2021-I

Postestratificación

En muchas investigaciones por muestreo es imposible o demasiado costoso estratificar la población previamente a la selección de la muestra.

Sin embargo, el presupuesto a menudo permite tomar muestras aleatoria simples sin reemplazo lo suficientemente grande como para poder clasificarla posteriormente en subgrupos o estratos de interés.

A la técnica anterior se le conoce como **postestratificación** y permite usar algunos resultados del muestreo estratificado para estimar los parámetros principales.

En este caso, la varianza de los estimadores es mayor que en el M.A.E pero menor que en el M.A.S si la clasificación ha sido correcta.

A medida que aumenta el tamaño de muestra n, los estimadores obtenidos se parecen más a los obtenidos mediante M.A.E con afijación proporcional.

La principal diferencia con el M.A.E es que los tamaños de muestra n_h son variables aleatorias. Pero estas vs.as tienen valor esperado dado por: $E[n_h] = n\left(\frac{N_h}{N}\right) = nP_h$, que corresponde al factor de asignación de la muestra global a los diferentes estratos en el M.A.E con afijación proporcional.

El estimador de la media poblacional cuando se ha llevado a cabo una postestratificación es:

$$\overline{y}_{post} = \frac{1}{N}\sum_{h=1}^{H} N_h\overline{y}_h = \sum_{h=1}^{H} W_h\overline{y}_h$$

con $\overline{y}_h$-es la media de la muestra que finalmente ha sido clasificada en el estrato h.

La varianza del estimador anterior es:

$$Var[\overline{y}_{post}] = \underbrace{\left(\frac{N-n}{N^2 n}\right)\sum_{h=1}^{H} N_h S_h^2}_{} + \underbrace{\frac{1}{n^2}\sum_{h=1}^{H}\left(1-\frac{N_h}{N}\right)S_h^2}_{}$$

El primer término de esta varianza, corresponde a la varianza de la media en el M.A.E con afijación proporcional.

Es decir, el primer término es:

$$Var[\overline{y}_{post}] = \frac{1}{N^2}\sum_{h=1}^{H} N_h^2\left(\frac{N_h-n_h}{N_h}\right)\frac{S_h^2}{n_h}, \quad \text{con: } n_h = n\,(N_h/N) = n w_h$$

El segundo término es el castigo derivado de la postestratificación, el cual disminuye rápidamente con el aumento del tamaño de la muestra.

Similarmente, para el total poblacional se tiene que:

$$\widehat{\tau}_{post} = N\overline{y}_{post}, \quad \text{y} \quad Var[\widehat{\tau}_{post}] = N^2 Var[\overline{y}_{post}]$$

Construcción y Número de Estratos

El establecimiento de los límites correspondientes a los diferentes estratos es un problema que han estudiado diferentes autores.

En general se han propuesto varios métodos, entre los cuales está el método de Acumulación de las Raíces de las Frecuencias*.

El método de las raíces de frecuencias, es en general el método más recomendado. El procedimiento considera los siguientes pasos:

1. Agrupar la variable de estratificación (Y) en un número suficientemente grande de clases, con intervalos de clases aproximadamente de igual tamaño.

*Dalenius & Hodges (1959). Minimum Variance Stratification. JASA. 54, 88-101

2. Determinar la distribución de frecuencia de Y, $f(y)$, para cada una de las clases, ie. se hallan las frecuencias absolutas de clase.

3. Acumular $\sqrt{f(y)}$.

4. Definir Q (ie. el ancho de los estratos), como el cociente entre la suma de todas las $\sqrt{f(y)}$ y el número de estratos escogidos, H.

5. Establecer los puntos de división de la variable, y que caracterizan a los estratos, como $Q, 2Q, \ldots, (H-1)Q$.

   Si los elementos originales están agrupados en clases pero no se conocen sus valores individuales, estos límites deben ajustarse adecuadamente a las clases existentes.

Ejemplo: Suponga que se desea hacer una investigación para estimar el promedio anual en ventas de 56 empresas, a partir de una muestra de $n = 15$ empresas.

Se encuentran datos disponibles de frecuencias en una clasificación por incrementos de $U\$50{,}000$. ¿Cómo determinar H=3 estratos?.

Los datos aparecen en la siguiente tabla:

Ingresos (miles de $U\$$)	Frecuencia $f(y)$	$\sqrt{\text{frecuencia}}$ $\sqrt{f(y)}$	$\sqrt{\text{frecuencia}}$ Acumulada $\sum\sqrt{f(y)}$
100-150	11	3.32	3.32
150-200	14	3.74	7.06
200-250	9	3	10.06
250-300	4	2	12.06
300-350	5	2.24	14.30
350-400	8	2.83	17.13
400-450	3	1.73	18.86
450-500	2	1.41	20.27

1. Se calcula el ancho del estrato usando la última columna, $\sqrt{\text{frecuencia}}$ acumulada: $Q = \frac{20{,}27}{3} = 6{,}76$

2. El primer intervalo tendrá límite superior el ingreso correspondiente a $6{,}76$ en la $\sqrt{\text{frec}}$ acumulada, es decir, $200{,}000$, que hace referencia al valor más cercano en la cuarta columna a 6.76 ($7{,}06$). El segundo intervalo irá desde $200{,}001$ hasta el límite superior correspondiente en la cuarta columna a $13{,}52$ ($2Q$), y en este caso el valor más cercano fue $14{,}30$, referente a $350{,}000$. Y el último intervalo irá desde $350{,}001$ hasta $500{,}000$.

3. Finalmente los estratos formados estarían dados por las empresas cuyos ingresos están en los siguientes estratos:
   E_1: Ingresos entre 100.000 y 200.000, con $N_1 = 25$
   E_2: Ingresos entre 200.001 y 350.000, con $N_2 = 18$
   E_3: Ingresos entre 350.001 y 500.000, con $N_3 = 13$

4. La forma óptima de muestrear el total de 15 empresas es mediante la realización del mismo tamaño de muestra en los tres estratos, es decir, $n_i = 5$, $i = 1, 2, 3$

Número Óptimo de Estratos

Para determinar el número óptimo de estratos, Cochran (1977), sugiere considerar como variables de estratificación aquella variable X sobre la cual se posea suficiente información y que esté altamente correlacionada con la variable de estudio Y.

Frecuentemente, se sugiere tomar una variable que corresponda a la misma que se desea estudiar pero de un censo anterior. A partir de esto, se considera la siguiente relación:

$$Var[\overline{y}_{est}] \approx \frac{\sigma^2}{n}\left[\frac{\rho^2}{H^2} + (1-\rho^2)\right]$$

donde, ρ-es el coeficiente de correlación lineal entre las variables X y Y.

Claramente, la varianza del estimador disminuye a medida que aumenta el número de estratos.

La siguiente tabla muestra el comportamiento de la $Var[\overline{y}_{est}]$ en función del número de estratos cuando $\rho = 0{,}8$

Nro. de Estratos-H	2	3	4	5	6
$Var[\overline{y}_{est}]$	$(0.520)\frac{S^2}{n}$	$(0.431)\frac{S^2}{n}$	$(0.400)\frac{S^2}{n}$	$(0.386)\frac{S^2}{n}$	$(0.378)\frac{S^2}{n}$

Generalmente, después de 6-estratos no hay mucha ganancia en la reducción de la varianza.

EJEMPLO:

Supongamos que se desea hacer un estudio con el fin de estimar ciertos parámetros en una región de un país que comprende 150 municipios.

Se desea conocer, por ejemplo, el número promedio de estudiantes por municipio, el número de habitantes con acceso a servicios públicos (agua, alcantarillado, luz, teléfono), el ingreso promedio, etc.

Adicionalmente también se desean estimaciones sobre los totales regionales para estas variables. Todas ellas pueden considerarse altamente correlacionadas con el número total de habitantes del municipio ($\rho \approx 0{,}8$, en todos los casos).

Se sugiere entonces estratificar los municipios de acuerdo al número de habitantes. Por lo tanto se necesita determinar no sólo el número de estratos sino también los límites correspondientes a cada uno de ellos.

Solución: El primer paso consiste en determinar el número de estratos H a considerar.

Para ello debe observarse el comportamiento de la $Var[\overline{y}_{est}]$ como una función del número de estratos, asumiendo constante los demás términos.

Lo práctico, con base en la tabla anterior, es establecer únicamente 6-estratos (inclusive 5-estratos también sería una alternativa válida).

La distribución de frecuencias del número de habitantes en los municipios de la región se da en la siguiente tabla:

Luego, se tiene que:

$$Q = \frac{\sum \sqrt{f(X)}}{H} = \frac{35.888}{6} = 5.981$$

Número de habitantes-X	f(X)	$\sqrt{f(X)}$	$\sum\sqrt{f(X)}$
0-4999	16	4.000	4.000
5000-9999	50	7.071	11.071
10000-14999	35	5.916	16.987
15000-19999	15	3.873	20.860
20000-24999	7	2.646	23.506
25000-29999	5	2.236	25.742
30000-34999	3	1.732	27.474
35000-39999	1	1.000	28.474
40000-44999	2	1.414	29.888
45000-49999	1	1.000	30.888
50000-54999	1	1.000	31.888
55000-59999	1	1.000	32.888
60000-64999	1	1.000	33.888
65000-69999	1	1.000	34.888
70000-74999	1	1.000	35.888

Luego los límites para los 6-estratos son:

Q $2Q$ $3Q$ $4Q$ $5Q$, Es decir:

5.981 11.962 17.943 23.924 29.905

Si la información disponible es únicamente la que aparece en la tabla anterior, entonces los límites para los estratos tendrán que ser ajustados de la forma que aparecen en la siguiente tabla y utilizando los datos de la primera tabla.

Identificación de Estrato H por número de habitantes	N_h
0-4999	16
5000-9999	50
10000-14999	35
15000-24999	22
25000-44999	11
45000-74999	6

Muestreo Doble Para la Estratificación

En lo anterior se ha supuesto que los

$$W_h = \frac{N_h}{N}\,, \quad \text{para} \quad h = 1, 2, \ldots, H$$

son constantes conocidas antes de iniciarse el muestreo.

En muchas situaciones esto no es el caso, incluso cuando parezca que el MAES es adecuado.

Por ejemplo, se puede desear estratificar una población de votantes a unas futuras elecciones según el género (hombre o mujer) o según el nivel educativo (primaria, secundaria, universitaria, etc.), pero no disponer de información para llevar a cabo la estratificación a partir de un ceso electoral.

La idea básica del Muestreo doble (muestreo en dos fases) es sencilla, pero complica el cálculo de las varianzas estimadas de los estimadores usados.

Se supone que la información preliminar, como por ejemplo, (el género, el nivel educativo, los kilómetros recorridos por un automóvil, etc.) en la que se basa la estratificación es fácil de obtener, mientras que la información detallada acerca de las variables de interés o estudio como por ejemplo, (las opiniones acerca de aspectos políticos, o acerca de la calidad de un automóvil, o acerca de la calidad de las instituciones públicas, etc.) no son fáciles de obtener.

En estos casos se puede tomar una muestra grande (muestra de la fase-1) para identificar los estratos y luego una muestra mucho menor (muestra de la fase-2) para recopilar los datos detallados de las variables de interés.

Por ejemplo, se podría llamar a muchos votantes para identificar el género, o el nivel de ingreso, o el nivel educativo, etc. (muestra de la fase-1) y luego entrevistar a unos pocos (muestra de la fase-2), con el fin de completar un cuestionario detallado con la información necesaria para las estimaciones de las variables de interés en el estudio.

Suponga que la muestra de la fase-1, de tamaño $n^{'}$, se utilizará para determinar qué elementos pertenecen a los distintos estratos. Luego,

$$w_h^{'} = \frac{n_h^{'}}{n^{'}}\,, \quad \text{para} \quad h =, 1, 2, \ldots, H$$

denota la proporción o porcentaje de elementos de la primera muestra que pertenecen al estrato h. Es decir, los $w_h^{'}$-son una estimación insesgada de los $W_h = \frac{N_h}{N}$.

En la segunda fase, se muestrean aleatoriamente, n_h-elementos de los $n_h^{'}$-elementos identificados pertenecientes al estrato h en la fase-1.

Las mediciones se obtienen a partir de estos n_h-elementos de la fase-2 en cada estrato, y $\bar{y}_h$ y S_h^2 se pueden calcular para cada estrato.

El estimador de la media poblacional μ cuando se ha llevado a cabo un doble muestreo para la estratificación es:

$$\overline{y}_{mdest} = \sum_{h=1}^{H} w_h^{'} \overline{y}_h = \sum_{h=1}^{H} \left(\frac{n_h^{'}}{n^{'}} \right) \overline{y}_h$$

con $\overline{y}_h$-es la media de la muestra que finalmente ha sido clasificada en el estrato h.

Si las fracciones de muestreo de la fase-2, ie. $\left(\frac{n_h}{N}\right)$, son todas pequeñas y N-es grande, una varianza aproximada de $\overline{y}_{mdest}$, está dada por:

$$Var[\overline{y}_{mdest}] = \left(\frac{n^{'}}{n^{'}-1}\right) \sum_{h=1}^{H} \left[\left(w_h^{'2} - \frac{w_h^{'}}{n^{'}}\right) \frac{S_h^2}{n_h} + \frac{w_h^{'}(\bar{y}_h - \bar{y}_{mdest})^2}{n^{'}}\right]$$

y si $n^{'}$-es tan grande que, $\frac{w_h^{'}}{n^{'}}$-es despreciable, entonces el estimador de la varianza anterior se reduce a:

$$\begin{aligned}\widehat{Var}[\overline{y}_{mdest}] &= \sum_{h=1}^{H} \left[\frac{w_h^{'2} S_h^2}{n_h} + \frac{w_h^{'}(\bar{y}_h - \bar{y}_{mdest})^2}{n^{'}}\right] \\ &= \sum_{h=1}^{H} \frac{w_h^{'2} S_h^2}{n_h} + \sum_{h=1}^{H} \frac{w_h^{'}(\bar{y}_h - \bar{y}_{mdest})^2}{n^{'}}\end{aligned}$$

EJEMPLO:

A partir de una lista de cantidades de inscripciones y cantidades de profesores de universidades de cierto país, se desea estimar el número promedio de inscripciones para un determinado semestre.

Las instituciones privadas suelen ser mas pequeñas que las públicas, por lo que se podría usar la estratificación entre privadas y públicas, sin embargo la lista disponible no está dividida de esta forma, aunque los datos están codificados de tal manera que se identifica el tipo de Universidad.

Por lo tanto el tipo de Universidad (privada o pública) se puede obtener rápidamente, mientras que los datos acerca del número de inscripciones y número de profesores son algo más complejos de obtener.

Se realizó una muestra sistemática de uno de cada diez estudiantes, para obtener información acerca del tipo de Universidad. Los resultados de esta muestra sistemática son los siguientes:

Privada	Pública	Total
$n_1^{'}$	$n_2^{'}$	$n^{'}$
84	57	141

Es decir que en este caso se tiene que:

$$w_1^{'} = \frac{n_1^{'}}{n^{'}} = \frac{84}{141} \quad \text{y} \quad w_1^{'} = \frac{n_2^{'}}{n^{'}} = \frac{57}{141}$$

Luego se tomaron sub-muestras de $n_1 = 11$-universidades privadas y de $n_2 = 12$-universidades públicas, dichas sub-muestras ofrecieron los siguientes datos del número de inscripciones y del número de profesores por universidad.

Las fracciones: n_h/N-se consideran pequeñas, ie. N-grande.

Privadas, $n_1 = 11$		Públicas, $n_2 = 12$	
Inscripciones	Profesorado	Inscripciones	Profesorado
1618	122	7332	452
1140	88	2356	131
1000	65	21879	996
1225	55	935	50
791	79	1293	106
1600	79	1293	106
746	40	8500	506
1701	75	6491	371
701	32	781	108
6918	428	7255	298
1050	110	2136	128
		5380	280

Estimar el número promedio de inscripciones por universidad para dicho país y establecer un LEE.

A partir de los datos se tiene que:

$$\overline{y}_{mdest} = \sum_{h=1}^{2} w_h^{'} \overline{y}_h = w_1^{'} \overline{y}_1 + w_2^{'} \overline{y}_2$$

$$= \sum_{h=1}^{2} \left(\frac{n_h^{'}}{n^{'}} \right) \overline{y}_h = \left(\frac{84}{141} \right) (1681) + \left(\frac{57}{141} \right) (5853)$$

$$= (0.60)(1681) + (0.40)(5853)$$

$$\overline{y}_{mdest} = 3349.8$$

Para estimar la varianza de $\overline{y}_{mdest}$ se procede como sigue:

$$\widehat{Var}[\bar{y}_{mdest}] = \sum_{h=1}^{2} \frac{w_h'^2 S_h^2}{n_h} + \sum_{h=1}^{2} \frac{w_h'(\bar{y}_h - \bar{y}_{mdest})^2}{n'}$$

$$= \left[\frac{1}{n_1}(w_1' S_1)^2 + \frac{1}{n_2}(w_2' S_2)^2\right] + \left(\frac{1}{n'}\left[w_1'(\bar{y}_1 - \bar{y}_{mdest}) + w_2'(\bar{y}_2 - \bar{y}_{mdest})\right]\right)$$

$$= \frac{1}{11}\Big[(0.60)(1773)\Big]^2 + \frac{1}{12}\Big[(0.40)(5763)\Big]^2$$

$$+ \left(\frac{1}{141}\left[(0.60)(1681\text{-}3349.8)^2 + (0.40)(5853\text{-}3349.8)^2\right]\right)$$

$$= 545708.05 \; + \; 29626.52$$

$$\widehat{Var}[\bar{y}_{mdest}] = 575334.57$$

y el LEE de μ estimada con $\overline{y}_{mdest} = 3349.8$ y un NC de aproximadamente el $95\,\%$ está dado por:

$$B = 2\sqrt{\widehat{Var}[\overline{y}_{mdest}]} = 2\sqrt{575334.57} = 2(758.5) = 1517$$

La segunda parte de la varianza (ie. la debida a la estimación de los verdaderos pesos de los estratos, ie. 29626.52), puede parecer grande, pero sólo representa el $5\,\%$ de la varianza total.

La estimación resultante del LEE de $\overline{y}_{mdest}$, sigue siendo bastante grande, debido a los pequeños tamaños de muestras y a la gran variación entre el número de inscripciones en las universidades, pero es mucho menor que el LEE asociado con un única muestra aleatoria de 23 universidades de la lista.

Verificar este dato.