V II Muestreo por Conglomerados

Dr. Jesús Mellado

Por algunas razones naturales, los elementos muestrales se encuentran formando grupos, como por ejemlo, las personas que viven en colonias de una ciudad, lo elementos de una caja de una línea de producción, los clubes de personas, las áreas arboladas de un terreno, etc.

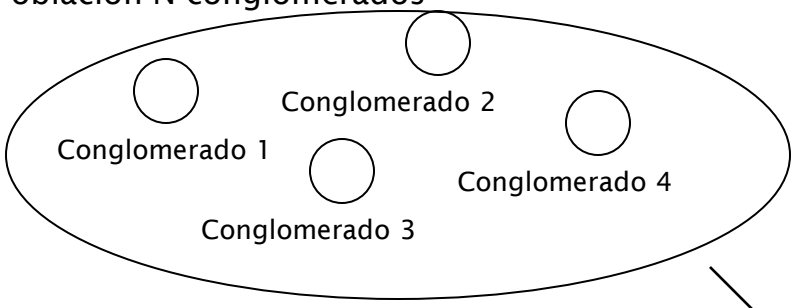
Como el objetivo del muestreo es levantar la mayor cantidad de información al menor costo, en este tipo de casos lo más económico el encuestar a un elemento muestral y a todos sus vecinos, así se ahorran los costos de un traslado del encuestador.

A este modelo de muestreo se llama “Por conglomerados”, ya que una vez seleccionado un elemento para la muestra, se incluyen también a todos los elementos que estén alrededor de él.

A diferencia del muestreo estratificado, este muestreo no requiere que los elementos tengan características homogéneas.

Características

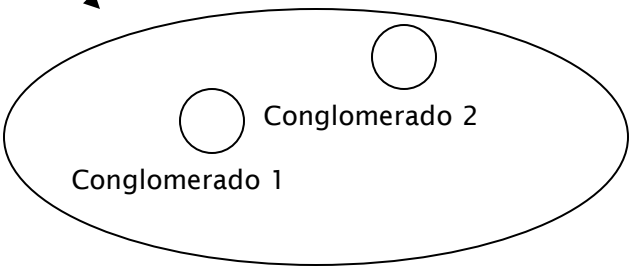
Población N conglomerados



Se tienen ***N*** conglomerados

De los ***N*** conglomerados se obtiene una muestra de ***n*** conglomerados

Muestra ***n*** conglomerados



Conglomerado 1 ○ Tamaño m_1

Conglomerado 2 ○ Tamaño m_2

Conglomerado 3 ○ Tamaño m_3

Conglomerado 4 ○ Tamaño m_4

El promedio del tamaño de los conglomerados de la muestra se calcula de la siguiente manera:

$$\bar{m} = \frac{1}{n}\sum_{i=1}^{n} m_i$$

Departamento de Estadística y Cálculo

El tamaño de toda la población se calcula de la siguiente manera:

$$M = \sum_{i=1}^{N} m_i$$

Nótese que los parámetros marcados con "M" mayúscula se refieren a toda la población.

El tamaño promedio de los conglomerados de toda la población se calcula de la siguiente manera:

$$\bar{M} = \frac{M}{N}$$

Selección de la muestra.

Si los conglomerados son evidentes, se sigue un proceso aleatorio para su selección, de lo contrario se selecciona aleatoriamente los elementos y muestrear y después se identifica su conglomerado.

En cada conglomerado se obtiene una suma de la variable que se va a medir (en este método se trabaja con la suma más que con la media)

A la suma de la variable de cada conglomerado se llamará y_i

Estimación de la media

Una vez seleccionados los conglomerados a muestrear, se obtiene de cada uno su tamaño (m_i) y la suma de la variaable que se desea analizar (y_i). Nótece que es la suma de las variables, no la media.

Conglomerado	m_i	y_i
1	32	125
2	28	136
3	25	145
4	27	134
5	26	135
6	30	128
	168	803

Después se suma cada una de las columnas y se aplica la siguiente ecuación:

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

Como los valores de las sumatorias ya está calculado en la tabla, solamente se sustituyen los valores:

$$y = \frac{803}{168} = 4.77$$

Estimación de la varianza de la media

Para el cálculo de la varianza de la media es conveniente agregar dos columnas a la tabla, en la primera se multiplica la media general por el tamaño de cada conglomerado; en la siguiente columna se resta el total de cada conglomerado menos el la columna anterior y se eleva al cuadrado. La columna se suma.

Conglomerado	m_i	y_i	$\tilde{y}m_i$	$(y_i - \tilde{y}m_i)^2$
1	32	125	152.95	781.336
2	28	136	133.83	4.694
3	25	145	119.49	650.554
4	27	134	129.05	24.467
5	26	135	124.27	115.051
6	30	128	143.39	236.940
	168	803		1813.042

La varianza se calcula con la siguiente ecuación:

$$V(y)=\left(\frac{N-n}{Nn\left(\frac{M}{N}\right)^2}\right)\frac{\sum_{i=1}^{n}(y_i-ym_i)^2}{n-1}$$

Si N=81 conglomerados y M=2268 elementos en la población. Nótece que se la sumatoria ya está calculada en la tabla anterior.

$$V(y)=\left(\frac{81-6}{81(6)\left(\frac{2268}{81}\right)^2}\right)\frac{1813.04}{6-1}=0.0713$$

Intervalo de confianza de la media

El intervalo de confianza para la media es la siguiente:

$$y-2\sqrt{V(y)}<\mu<y+2\sqrt{V(y)}$$

$$4.77-2\sqrt{0.071}<\mu<4.77+2\sqrt{0.071}$$

$$4.24<\mu<5.31$$

Tamaño de la muestra para estimar la media

Para realizar los cálculos es necesario encontrar la varianza del total en la muestra con la siguiente ecuación:

$$s_c^2=\frac{\sum_{i=1}^{n}(y_i-ym_i)^2}{n-1} \qquad s_c^2=362.61$$

Se determina el error máximo que se permite en los resultados. A este valor se le llamará B. Las ecuaciones para encontrar el tamaño de la muestra son las siguientes:

$$D=\frac{B^2M^{-2}}{4} \qquad n=\frac{Ns_c^2}{N^2D+s_c^2}$$

Si B=0.4

$$D=\frac{(0.4)^2(2268/81)^2}{4}=31.36$$

El resultado es el número de conglomerados que se deben muestrear. El resultado se redondea al entero superior

$$n=\frac{(81)(362.61)}{(81)31.36+362.61}=10.11 \qquad n = 11$$

Ejemplo

Con el fin de determinar si es conveniente instalar una productora de yogurt en cierto poblado, se desea conocer el consumo mensual por persona al mes. De un total de 120 conglomerados detectados se establecieron 8 conglomerados con los resultados que se muestran. Estimar la media, su intervalo de confianza y el tamaño adecuado de la muestra si el error máximo es 0.2. El total de de personas estimado es de 10,000.

Conglomerad	Personas	Suma litros
1	85	78
2	65	59
3	78	70
4	79	70
5	56	52
6	82	73
7	74	66
8	82	72

Conglomerado	m_i	y_i	$\tilde{y}m_i$	$(y_i - \tilde{y}m_i)^2$
1	85	78	406.28	107767.602
2	65	59	310.68	63345.100
3	78	70	267.67	39072.111
4	79	70	391.94	103645.670
5	56	52	353.7	91024.327
6	82	73	391.94	101723.027
7	74	66	353.7	82772.660
8	82	72	391.94	102361.908
	601	540		691712.405

La varianza

$$V(y) = \left(\frac{N-n}{Nn\left(\frac{M}{N}\right)^2} \right) \frac{\sum_{i=1}^{n} (y_i - ym_i)^2}{n-1}$$

$$V(y) = \left(\frac{120-8}{120(8)\left(\frac{10000}{120}\right)^2} \right) \frac{691{,}712.4}{8-1} = 1.66$$

La media

$$y = \frac{\sum_{i=1}^{n} y_i}{\sum_{i=1}^{n} m_i} \qquad y = \frac{540}{601} = 0.89$$

La varianza es alta porque es un estimador sesgado para muestras menores a 20 conglomerados

El intervalo de confianza

$$y - 2\sqrt{V(y)} < \mu < y + 2\sqrt{V(y)}$$

$$0.89 - 2\sqrt{1.66} < \mu < 0.89 + 2\sqrt{1.66}$$

$$0 < \mu < 3.47$$

Tamaño de la muestra

$$s_c^2 = \frac{\sum_{i=1}^{n} (y_i - ym_i)^2}{n-1} \qquad s_c^2 = 138{,}342$$

$$D = \frac{B^2 M^2}{4} \qquad n = \frac{N s_c^2}{N^2 D + s_c^2}$$

$$D = \frac{(0.2)^2 (10000/120)^2}{4} = 69.44$$

$$n = \frac{(120)(138{,}342)}{(120)69.44 + 138{,}342} = 113.18 \qquad n = 114$$

Estimación del total

Para estimar el total de una variable de toda la población se puede llenar la tabla que se muestra, donde cada renglón corresponde a cada estrato, en la primera columna se ubica el tamaño de ese estrato (N_i), en la segunda columna el tamaño de la muestra para ese estrato (n_i), en la tercera columna la media calculada para cada estrato (y) y en la cuarta columna se realiza la multiplicación $N_i y_i$.

Estrato	N_i	n_i	y_i	$N_i y_i$
1	1190	12	32	38080
2	926	10	25	23150
3	825	9	26	21450
4	1350	14	27	36450
N =	4291		suma	119130

Se calcula el valor de N, que es la suma del tamaño de cada estrato.

Se calcula la suma de la última columna, el valor resultante es el total.

La ecuación es como se muestra:

$$y = \sum_{i=1}^{L} N_i y_i$$

Estimación de la varianza del total

La varianza del total permitirá establecer el intervalo de confianza.

Para calcular la varianza del total se debe calcular la varianza de cada estrato con las siguientes fórmulas.

$$s_i^2 = \frac{\sum_{j=1}^{n_i} (y_{i,j} - y_i)^2}{n_i - 1}$$

O bien

$$s_i^2 = \frac{\sum_{j=1}^{n_i} x^2{}_{i,j} - \frac{\left(\sum_{j=1}^{n_i} x_{i,j}\right)^2}{n_i}}{n_i - 1}$$

Donde *i* es el número de estrato y *j* es cada uno de las observaciones de cada estrato

La varianza poblacional (S_i^2) se puede agregar a la tabla de la media, para failitar los cálculos siguientes:

Estrato	N_i	n_i	y_i	$N_i y_i$	s_i^2
1	1190	12	32	38080	12
2	926	10	25	23150	13
3	825	9	26	21450	13
4	1350	14	27	36450	14
N =	4291		suma	119130	

Una vez que se obtiene la varianza muestral de cada estrato se calcula la varianza del total de cada estrato con la siguiente fórmula. Utilizando las columnas de la tabla anterior se pueden facilitar los cálculos.

$$V(\hat{t}_i) = \frac{s_i^2}{n_i}\left(\frac{N_i - n_i}{N_i}\right)$$

Estrato	N_i	n_i	y_i	$N_i y_i$	s_i^2	$V(t_i)$
1	1190	12	32	38080	12	0.99
2	926	10	25	23150	13	1.29
3	825	9	26	21450	13	1.43
4	1350	14	27	36450	14	0.99
N =	4291		suma	119130		

Para seguir con los cálculos es necesario multiplicar cada varianza del total por N_i^2 y ubicar el resultado en una nueva columna, sumar la columna. El resultado es la varianza del total de toda la muestra.

Estrato	N_i	n_i	y_i	$N_i y_i$	s_i^2	$V(t_i)$	$N^2V(y_i)$
1	1190	12	32	38080	12	0.99	1401820.0
2	926	10	25	23150	13	1.29	1102680.8
3	825	9	26	21450	13	1.43	972400.0
4	1350	14	27	36450	14	0.99	1803600.0
N =	4291		suma	119130		V(t)	5280500.8

Intervalo de confianza del total

El intervalo de confianza para el total es la siguiente:

$$\hat{\tau} - 2\sqrt{V(\hat{\tau})} < \tau < \hat{\tau} + 2\sqrt{V(\hat{\tau})}$$

Si t = 119130 y V(y) = 5,280,500; entonces el intervalo de confianza será el siguiente:

$$119{,}130 - 2\sqrt{5{,}280{,}500.8} < \tau < 119{,}130 + 2\sqrt{5.280{,}500.8}$$

$$114{,}534 < \tau < 123{,}725$$

Tamaño de la muestra para estimar el total

Para encontrar el tamaño de la muestra es necesario asignar a cada estrato un valor w_i, que será la proporción de datos que corresponden al estrato. La sumatoria de los valores w_i debe ser 1.

Algunas veces cada valor wi se calcula con la ecuación $w_i = N_i/N$

Los cálculos se facilitan si se crea la tabla que se muestra a la derecha, donde se muestra el tamaño de cada estrato, su varianza muestral y el valor w_i asignado.

Estrato	N_i	s_i^2	w_i
1	1190	12	0.3
2	926	13	0.2
3	825	13	0.2
4	1350	14	0.3
N =	4291		

Para realizar los cálculos es necesario agregar una columna para calcular $N_i^2 s_i^2/w_i$ (columna 1 al cuadrado por la columna 2 entre la columna 3) y sumar cada uno de los renglones.

Estrato	N_i	s_i^2	w_i	$N_i^2 s_i^2/w_i$
1	1190	12	0.3	56644000
2	926	13	0.2	55735940
3	825	13	0.2	44240625
4	1350	14	0.3	85050000
N =	4291			241670565

También es necesario agregar una columna para agregar $N_i s_i^2$ (columna 1 por columna 2) y sumar los valores de la columna.

Estrato	N_i	s_i^2	w_i	$N_i^2 s_i^2/w_i$	$N_i s_i^2$
1	1190	12	0.3	56644000	14280
2	926	13	0.2	55735940	12038
3	825	13	0.2	44240625	10725
4	1350	14	0.3	85050000	18900
N =	4291			241670565	55943

El paso siguiente es definir el error máximo que se desea para la media, a ese valor se le llamará ***B***, así por ejemplo, si el total es 119,130 y se desea un error máximo de 5,000, ***B***=5,000

$$D = \frac{B^2}{4N^2}$$

Se define el valor D con la ecuación que se muestra a la derecha.

$$D = \frac{(5{,}000)^2}{4(4{,}291)^2} = 0.339$$

Por último, se calcula el valor de n (tamaño de la muestra) utilizando la ecuación que se muestra. El valor del numerador ya se tiene calculado en la cuarta columna de la tabla previamente creada, y la segunda parte del denominador de igual manera ya se tiene calculado en la quinta columna de la tabla .

$$n = \frac{\sum_{i=1}^{L} N_i^2 s_i^2 / w_i}{N^2 D + \sum N_i s_i^2}$$

Dado que las observaviones no pueden ser parciales, el valor de n se aumenta al entero siguiente superior. n=39

$$n = \frac{241670565}{(4291^2)0.339 + 55943} = 38.32$$

Ejemplo En una zona se desea estimar el peso total de la producción de papa de tres parcelas. Las parcelas están repartidas en tres ranchos con diferentes condiciones climáticas, así que se planea un muestreo estratificado. En el primer rancho se muestrearon 10 plantas de 900, en el segundo rancho 12 plantas de 1100 y en el tercero 12 de 1050. Con los datos que se muestran a continuación encontrar el total con su intervalo de confianza al 95% de seguridad y con el tamaño de muestra para tener un error máximo de 250 kilos (datos ficticios).

Rancho 1	2	2.5	2	2.5	3	2	3	3	2.5	3		
Rancho 2	3	3.5	4	4	3.5	3.5	4	2.5	3	3.5	3.5	4
Rancho 3	2	2.5	3	3.5	2.5	3	2.5	3.5	2	3	3	3.5

Estrato	N_i	n_i	y_i	$N_i y_i$
1	900	10	2.55	2295
2	1100	12	3.50	3850
3	1050	12	2.83	2975
N =	3050		suma	9120

Después de llenar la tabla se sabe que el total es 9120 kilos.

Estrato	N_i	n_i	y_i	$N_i y_i$	s_i^2	$V(y_i)$	$N^2V(y_i)$
1	900	10	2.55	2295	0.19	0.019	15352.5
2	1100	12	3.50	3850	0.23	0.019	22666.7
3	1050	12	2.83	2975	0.29	0.024	26146.6
N =	3050		suma	9120		suma	64165.8

También se puede concluir que la varianza del total es 68165.8

Al aplicar la ecuación para el intervalo de confianza

$$8613 < \tau < 9626$$

Para el tamaño de la muestra:

Estrato	N_i	s_i^2	w_i	$N_i^2 s_i^2 / w_i$	$N_i s_i^2$
1	900	0.19	0.3	526125	172.5
2	1100	0.23	0.4	762500	250
3	1050	0.29	0.3	921932	302.27
				2210557	724.77

B=250

D=0.0017

$$n = \frac{2{,}210{,}557}{(3050^2)0.0017 + 724.77} = 135.2$$

El tamaño de la muestra debe ser 135, lo que significa que se requieren 101 mas observaciones para llegar a la exactitud requerida.

Dr. Jesús Mellado Bosque

Estimación de una proporción

Para estimar una proporción de una variable de toda la población se puede llenar la tabla que se muestra, donde cada renglón corresponde a cada estrato, en la primera columna se ubica el tamaño de ese estrato (N_i), en la segunda columna el tamaño de la muestra para ese estrato (n_i), en la tercera columna la proporción calculada para cada estrato (p_i) y en la cuarta columna se realiza la multiplicación $N_i p_i$.

Se calcula el valor de N, que es la suma de los tamaño sde cada estrato.

Se calcula la suma de la última columna y se divide entre N, el resultado es la proporción de toda la población.

La fórmula es como se muestra:

Estrato	N_i	n_i	p_i	Nip_i
1	1190	12	0.26	309.4
2	926	10	0.24	222.24
3	825	9	0.26	214.5
4	1350	14	0.2	270
N =	4291		suma	1016.14
			p	0.24

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{L} N_i \hat{p}_i$$

Estimación de la varianza de la proporción

La varianza de la proporción permitirá establecer el intervalo de confianza para la proporción.

Para calcular la varianza de la proporción se debe calcular la varianza de cada estrato multiplicando p_iq_i, donde q_i es 1-p_i

La varianza se puede agregar a la tabla de la media, para failitar los cálculos ṣiguientes

Estrato	N_i	n_i	p_i	Nip_i	p_iq_i
1	1190	12	0.26	309.40	0.19
2	926	10	0.24	222.24	0.18
3	825	9	0.26	214.50	0.19
4	1350	14	0.2	270.00	0.16
N =	4291		suma	1016.14	
			p	0.24	

Una vez que se obtiene la varianza muestral de cada estrato se calcula la varianza de la proporción de cada estrato con la siguiente ecuación. Utilizando la columna de la tabla anterior se pueden facilitar los cálculos.

$$V(y_i) = \frac{p_i q_i}{n_i}\left(\frac{N_i - n_i}{N_i}\right)$$

Estrato	N_i	n_i	p_i	Nip_i	p_iq_i	$V(y_i)$
1	1190	12	0.26	309.40	0.19	0.02
2	926	10	0.24	222.24	0.18	0.02
3	825	9	0.26	214.50	0.19	0.02
4	1350	14	0.2	270.00	0.16	0.01
N =	4291		suma	1016.14		
			p	0.24		

Para seguir con los cálculos es necesario multiplicar cada varianza de la media por N_i^2 y ubicar el resultado en una nueva columna, sumar la columna y luego dividir la suma entre $1/N^2$. El resultado es la varianza de la media de toda la muestra.

Estrato	N_i	n_i	p_i	Nip_i	p_iq_i	$V(y_i)$	$N^2V(y_i)$
1	1190	12	0.26	309.40	0.19	0.02	22475.847
2	926	10	0.24	222.24	0.18	0.02	15471.460
3	825	9	0.26	214.50	0.19	0.02	14391.520
4	1350	14	0.2	270.00	0.16	0.01	20612.571
N =	4291		suma	1016.14		suma	72951.399
			p	0.24		V(p)	0.004

Intervalo de confianza de la proporción

El intervalo de confianza para la proporción es la siguiente:

$$\hat{p} - 2\sqrt{V(\hat{p})} < p < \hat{p} + 2\sqrt{V(p)}$$

Si p = 0.24 y V(p) = 0.004; entonces el intervalo de confianza será el siguiente:

$$0.24 - 2\sqrt{0.004} < p < 0.24 + 2\sqrt{0.004}$$

$$0.1109 < \mu < 0.3627$$

Tamaño de la muestra para estimar la proporción

Para encontrar el tamaño de la muestra es necesario asignar a cada estrato un valor w_i, que será la proporción de datos que corresponden al estrato. La sumatoria de los valores w_i debe ser 1.

Algunas veces cada valor wi se calcula con la ecuación $w_i = N_i/N$

Los cálculos se facilitan si se crea la tabla que se muestra a la derecha, donde se muestra el tamaño de cada estrato, su varianza muestral y el valor w_i asignado.

Estrato	N_i	s_i^2	w_i
1	1190	12	0.3
2	926	13	0.2
3	825	13	0.2
4	1350	14	0.3
N =	4291		

Para realizar los cálculos es necesario agregar una columna para calcular $N_i^2p_iq_i/w_i$ (columna 1 al cuadrado por la columna 2 entre la columna 3) y sumar cada uno de los renglones.

Estrato	N_i	p_iq_i	w_i	$N_i^2s_i^2/w_i$
1	1190	0.19	0.3	908192
2	926	0.18	0.2	782018
3	825	0.19	0.2	654761
4	1350	0.16	0.3	972000
N =	4291			3316971

También es necesario agregar una columna para agregar $N_is_i^2$ (columna 1 por columna 2) y sumar los valores de la columna.

Estrato	N_i	p_iq_i	w_i	$N_i^2s_i^2/w_i$	$N_is_i^2$
1	1190	0.19	0.3	908192	229
2	926	0.18	0.2	782018	169
3	825	0.19	0.2	654761	159
4	1350	0.16	0.3	972000	216
N =	4291			3316971	773

El paso siguiente es definir el error máximo que se desea para la proporción, a ese valor se le llamará ***B***, así por ejemplo, si la media es 0.24 y se desea un error máximo de 0.1, ***B***=0.1;

$$D = \frac{B^2}{4}$$

Se define el valor D con la ecuación que se muestra a la derecha.

$$D = \frac{(0.1)^2}{4} = 0.0025$$

Por último, se calcula el valor de n (tamaño de la muestra) utilizando la ecuación que se muestra. El valor del numerador ya se tiene calculado en la cuarta columna de la tabla previamente creada, y la segunda parte del denominador de igual manera ya se tiene calculado en la quinta columna de la tabla .

$$n = \frac{\sum_{i=1}^{L} N_i^2 p_i q_i / w_i}{N^2 D + \sum N_i p_i q_i}$$

Dado que las observaviones no pueden ser parciales, el valor de n se aumenta al entero siguiente superior. n=71

$$n = \frac{3{,}316{,}971}{(4291^2)0.0025 + 773} = 70.86$$

Ejemplo En una una planta productora de botes de yogurt se desea saber qué proporción de los botes no tienen el PH recomendado. La producción se lleva a cabo a través de tres máquinas, así que se decidió realizar la prueba por estratos. En la primera máquina, de una produccipon de 1200 botes se muestrearon 14; en la segunda máquina, de 1300 botes se muestrearon 15 y en la tercera máquina, de 1200 botes se muestrearon 14. Cada vez que en bote tiene un PH diferente se marca con un 1.

Encontrar el estimador de la proporción con su intervaloo de confianza al 95% y el tamaño de la muestra necesario para tener un error máximo de 0.1 (datos ficticios).

Máquina 1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	
Máquina 2	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0
Máquina 3	0	0	1	0	1	0	0	0	0	0	0	0	1	0	

Estrato	N_i	n_i	p_i	$N_i o_i$
1	1200	14	0.143	171.429
2	1300	15	0.133	173.333
3	1200	14	0.214	257.143
N =	3700		suma	601.90
			p	0.16

Después de llenar la tabla se sabe que la proporción general es 0.16

Estrato	N_i	n_i	p_i	$N_i o_i$	$p_i q_i$	$V(p_i)$	$N^2 V(p_i)$
1	1200	14	0.143	171.429	0.12	0.009	12447.8
2	1300	15	0.133	173.333	0.12	0.008	12869.0
3	1200	14	0.214	257.143	0.17	0.012	17115.7
N =	3700		suma	601.90		suma	42432.6
			p	0.16		V(p)	0.0031

También se puede concluir que la varianza de la proporción es 0.0031

Al aplicar la ecuación para el intervalo de confianza $0.0513 < p < 0.274$

Para el tamaño de la muestra:

Estrato	N_i	p_iq_i	w_i	$N_i^2 p_i q_i / w_i$	$N_i p_i q_i$
1	1200	0.12	0.3	543673	147
2	1300	0.12	0.4	555822	150
3	1200	0.17	0.3	747551	202
				1847047	499

B=0.1

D=0.063

$$n = \frac{1{,}847{,}047}{(3700^2)0.0025 + 499} = 53.19$$

El tamaño de la muestra debe ser 54, pero como en la muestra original fueron 43 observaciones es necesario muestrear 11 mas.