

OPTIMIZACIÓN DE TAMAÑOS DE MUESTRA EN DIFERENTES MUESTREOS Y ESTIMACIÓN EN MAS Y MAE: *CALnYES*

María Teresa Cabero Morán

Dpto. de Estadística, Facultad de Ciencias. Universidad de Salamanca
mateca@usal.es

RESUMEN: *CALnYES* es un programa informático que calcula tamaños de muestra en muestreo aleatorio simple, estratificado, muestreo por conglomerados, por conglomerados mezclado con estratificado, con las distintas posibilidades de afijaciones por estratos. Además, incluye el cálculo de tamaños de muestra en muestreo directo e inverso para la estimación de tamaños de población. Se suma una segunda parte en la que la aplicación realiza la estimación de medias y proporciones en muestreo aleatorio simple y estratificado. Presenta la estimación puntual, la varianza estimada y el error, además del intervalo, calculado a partir de la confianza elegida por el usuario. También, la aplicación ofrece las posibilidades necesarias para los distintos casos en los que se haya de calcular el tamaño de muestra (muestreo con o sin reposición, totales, proporciones o medias) mediante la elección del error con el que se quiera trabajar. Además ofrece la posibilidad de escoger la afijación a seguir en la asignación de muestras por estratos en muestreo estratificado (uniforme, proporcional u óptima con error o coste mínimos).

CALnYES presenta una forma sencilla, rápida y cómoda de calcular el tamaño de muestra y estimar medias, totales y proporciones en muestreo aleatorio simple y estratificado. La optimización está asegurada sin tener que invertir mucho tiempo con cálculos y fórmulas.

PALABRAS CLAVE: aplicaciones informáticas, tamaño de la muestra, estimación.

ABSTRACT: *CALnYES* is a software that calculates sample size for simple random sampling, stratified, one phase cluster sampling, stratified cluster sampling for different allocations. Besides, it includes the sample size calculation for direct and inverse sampling for the population size estimation. A second part of the software estimates means and proportions for simple random sampling and stratified. It develops the punctual estimation, its estimate variance and its error, and the interval, calculated from the confidence level chosen by the user. Also, the application offers the necessary possibilities for the different cases in which the user would have to calculate the sample size (simple random sampling with replacement or without replacement, totals, proportions or means) choosing the error for operating. It has the possibility to choose the allocation of samples in the stratum for stratified random sampling (uniform, proportional or optimum with minimum error or cost). *CALnYES* is an easy, fast and comfortable way to calculate the sample size and to estimate means, totals and proportions in simple and stratified random sampling. The optimization is assured without doing many calculations and formulas.

KEY WORDS: software, sample size, estimation

Recibido: 23 de octubre 2008

Revisado: 15 de enero 2009

Aceptado: 6 de febrero 2009

1. Introducción

El tipo de muestra que se ha de escoger en cada estudio dependerá de las propiedades que se conozcan de la población con la que se ha de trabajar. Una vez que el investigador elija el tipo de muestreo y el error que no desea sobrepasar obtendrá el número de observaciones óptimo mediante *CALnYES*, lo que proporcionará el número de encuestas a realizar.

Los orígenes de *CALnYES* ('CÁLculo de n Y ESTimaciones') se remontan al programa realizado por el autor *MASYE*, una aplicación que calcula tamaños de muestra en muestreo aleatorio simple (MAS) y estratificado (MAE).

MASYE dio lugar a la creación de otra versión *MASYE2* en la cual el número de tipos de muestreo a escoger se veía incrementado (muestreo por conglomerados, por conglomerados mezclado con estratificado, con las distintas posibilidades de afijaciones por estratos). Además, se incluía el cálculo de tamaños de muestra en muestreo directo e inverso para la estimación de tamaños de poblaciones.

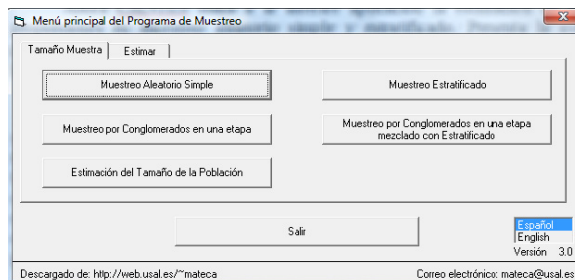
La aplicación ofrece las opciones necesarias para los distintos casos en los que se haya de calcular el tamaño de muestra (muestreo con o sin reposición, totales, proporciones o medias) mediante la elección del error con el que se quiere trabajar y la afijación a seguir en la asignación de muestras por estratos en muestreo estratificado (uniforme, proporcional u óptima con error o coste mínimos).

CALnYES consta de dos partes importantes: cálculo y optimización de tamaños de muestra, la primera; y, una segunda de estimación puntual y por intervalos en muestreo aleatorio simple y estratificado.

Así, una vez recogidos los datos, cuando se haya llevado a cabo la encuesta o estudio, la segunda parte de *CALnYES* entra en funcionamiento. Introduciendo la información recogida, calculará las deseadas estimaciones. Recaltar, que en este trabajo, éstas son en muestreo aleatorio simple y estratificado. Mientras que el cálculo del tamaño de la muestra es para muestreo aleatorio simple, muestreo aleatorio estratificado, muestreo por conglomerados, por conglomerados mezclado con estratificado; además de cálculo de tamaños de muestra en muestreo directo e inverso para la estimación de tamaños de poblaciones.

El programa presenta la siguiente pantalla de inicio:

Fig. 1.
Menú principal de CALnYES



Entrando en la opción deseada, la aplicación pide al usuario todos los datos necesarios.

Al escoger las opciones deseadas *CALnYES* proporciona las expresiones matemáticas que utiliza en los cálculos va a realizar en cada momento, tanto para hallar el tamaño de la muestra como para hacer estimaciones.

Escogiendo la confianza deseada por el usuario, el programa calcula automáticamente 'k' o cuantil de la distribución normal estándar correspondiente, visible en cada pantalla.

Los tipos de muestreo que van a entrar en juego son los que a continuación se presentan en los que se irá detallando e ilustrando la forma de introducir los datos en *CALnYES*.

2. Tamaño de muestra en muestra en muestreo aleatorio simple

Recuérdese que se desea extraer una muestra aleatoria simple de tamaño n , X_1, \dots, X_n , de una población de tamaño N . Ésta puede ser con o sin reemplazamiento.

El tamaño de muestra depende del parámetro que se quiera estimar, del nivel de confianza y del error máximo a asumir. Dicho parámetro será la media o el total (variables cuantitativas) o la proporción de individuos que cumplen una cierta característica (variables cualitativas).

Téngase en cuenta que si no se tiene información sobre la proporción de la población, se hace el caso más desfavorable: $p = q = 0,5$, pues es el que proporcionará el máximo tamaño de muestra a tomar.

Como ejemplo:

Se quiere determinar la proporción de consumidores que están a favor de un producto mediante una encuesta en la que se realizará dicha pregunta. Con este fin, se ha de calcular el tamaño de muestra para estimar dicha proporción de personas a favor con una confianza aproximada del 95'5%, y un error máximo del 10%. Hay 2.000 consumidores.

Por el supuesto se necesita utilizar muestreo aleatorio simple. Además, se desea calcular el tamaño de muestra para estimar una proporción y, como no se dice lo contrario, el muestreo es sin reposición. No se conoce ningún dato acerca de la proporción, así que se supone que $p = 50\%$.

En *CALnYES* después de elegir la pestaña de 'Tamaño muestra' y 'Muestreo Aleatorio Simple', opciones: **Muestreo sin reposición** → **Estimación para la proporción** → **Confianza: 95,5; Error: 0,10; Tamaño de la población: 2000; Proporción (p): 50%** y 'Calcular Tamaño de Muestra' se obtienen los resultados que se presentan en la Fig. 2.

Fig. 2.
Entrada de datos y resultados en CALnYES para MAS

Se necesita una muestra de 96 personas para que el error no sobrepase el 10% ('Tamaño de la muestra'). Según las fórmulas de cálculo presentadas el tamaño de la muestra debería ser de 95,66066733 ('Tamaño Calculado'), el cual se redondea por exceso, para obtener el número de observaciones definitivo que se han de utilizar en el estudio (96).

3. Estimación de medias, totales y proporciones en mas

Una vez recopilados los datos en la población, se pasa a estimar su media, su total o su proporción.

Por ejemplo:

Un psicólogo desea estimar el tiempo promedio que necesita un niño para terminar una tarea sencilla. También desea saber el tiempo total para terminar la tarea entre todos los niños, todo ello estableciendo límites al 95% para los errores de estimación. La consulta tiene 98 niños, y selecciona una muestra de 8, a los que se les toma el tiempo en minutos, y se obtienen los siguientes resultados:

4,2 5,1 7,9 3,8 5,3 4,6 5,1 4,1

Se trata de estimar una media y un total en un muestreo aleatorio simple, puesto que no se dice lo contrario, sin reposición, después de obtener los datos para una muestra de 8 individuos. No se sabe nada acerca de la varianza de la población.

En CALnYES después de elegir la pestaña de 'Estimar' y 'Muestreo Aleatorio Simple', opciones: **Muestreo sin reposición** → **Estimación para la media** →

Tamaño de la Población (N): 98; Confianza: 95; N° de Filas: 8; N° de Columnas: 1 → **Ajustar Celdas** → **(Meter datos)** → **Calcular**, aparece la pantalla de la figura 3.1. Se muestran las fórmulas que está aplicando en cada momento. Sin más que ir eligiendo las opciones van cambiando las expresiones.

En la parte derecha de la pantalla se ha de introducir los datos recogidos en las variables cuantitativas para estimar la media y el total. Indicar que éstos podrían estar en una hoja de cálculo de Excel y podrían ser movidos (copiados) aquí. Da igual que se coloquen en una columna, o en el caso de ser muchos, en varias. La aplicación los contará y dará el tamaño de la muestra, además, de calcular la estimación puntual y por intervalos, según la confianza elegida.

Fig. 3.1.

Entrada de datos y resultados en CALnYES para estimar la media en MAS

Muestreo Aleatorio Simple (estimación)

Muestreo...
 CDN Reposición (o población infinita)
 SIN Reposición

Tamaño de la Población (N) 98
 Confianza 95 %

N° de Filas 8 N° de Columnas 1

Estimación para...
 ... la media
 ... el total (N conocida)
 ... proporción

¿Varianza Conocida?

	A
1	4,2
2	5,1
3	7,9
4	3,8
5	5,3
6	4,6
7	5,1
8	5,1

Tamaño de la Muestra (n) 8

Media 5,01
 Cuasivarianza 1,65
 Valor calculado de 'k' 1,96
 Varianza de la Media 0,19
 Error 0,85
 Límites Intervalo Confianza 4,16 5,87

$$\hat{\mu} = \bar{X}, \text{Var}(\hat{\mu}) = \left(1 - \frac{n}{N}\right) \cdot \frac{s_c^2}{n}$$

$$e = k \cdot \sqrt{\text{Var}(\hat{\mu})}$$

Se obtiene una media muestral o estimación puntual de la media de $\hat{\mu} = \bar{X} = 5,01$ minutos. La cuasivarianza $s_c^2 = 1,65 \text{ min}^2$ para esa muestra. Por otra parte, la varianza estimada de la estimación de la media $\text{Var}(\hat{\mu}) = 0,19 \text{ min}^2$, lo que resulta un error de estimación de $e = 0,85 \text{ min}$. Así, el intervalo de confianza es $[4,16; 5,87]$ minutos. El tiempo medio ocupado por cada niño de los 98 que acuden a la consulta está entre 4,16 y 5,87 minutos.

Para obtener el total no hay más que cambiar **Estimación para el total (N conocida)** → **Calcular**, conservando los demás datos anteriores.

Fig. 3.2.

Entrada de datos y resultados en CALnYES para estimar el total en MAS

Muestreo Aleatorio Simple (estimación)

Muestreo...
 CON Reposición (o población infinita)
 SIN Reposición

Tamaño de la Población (N) 98
 Confianza 95 %

Estimación para...
 ... la media
 ... el total (N conocida)
 ... proporción

¿Varianza Conocida?

Nº de Filas 8 Nº de Columnas 1 Ajustar Celdas

	A
1	4,2
2	5,1
3	7,9
4	3,8
5	5,3
6	4,6
7	5,1
8	...

Tamaño de la Muestra (n) 8

Total 491,23
 Cuasivarianza 1,65
 Valor calculado de 'k' 1,96
 Varianza del Total 1.822,08
 Error 83,66
 Límites Intervalo Confianza 407,56 574,89

Calcular Salir

$$\hat{t} = N \cdot \bar{X}, \text{Var}(\hat{t}) = N^2 \cdot \left(1 - \frac{n}{N}\right) \cdot \frac{s_c^2}{n},$$

$$e = k \cdot \sqrt{\text{Var}(\hat{t})}$$

Se observa una estimación puntual del total de $\hat{t} = 491,23$ minutos. La cuasivarianza $s_c^2 = 1,65 \text{ min}^2$ para esa muestra, que es la que ya aparecía en el caso de la media. Por otra parte, la varianza estimada de la estimación del total $\text{Var}(\hat{t}) = 1822,08 \text{ min}^2$, lo que resulta un error de estimación de $e = 83,66$ min. Así, el intervalo de confianza es $[407,56; 574,89]$ minutos. El tiempo total ocupado por los 98 niños está entre 407,56 y 574,89 minutos.

4. Tamaño de muestra en muestreo aleatorio estratificado

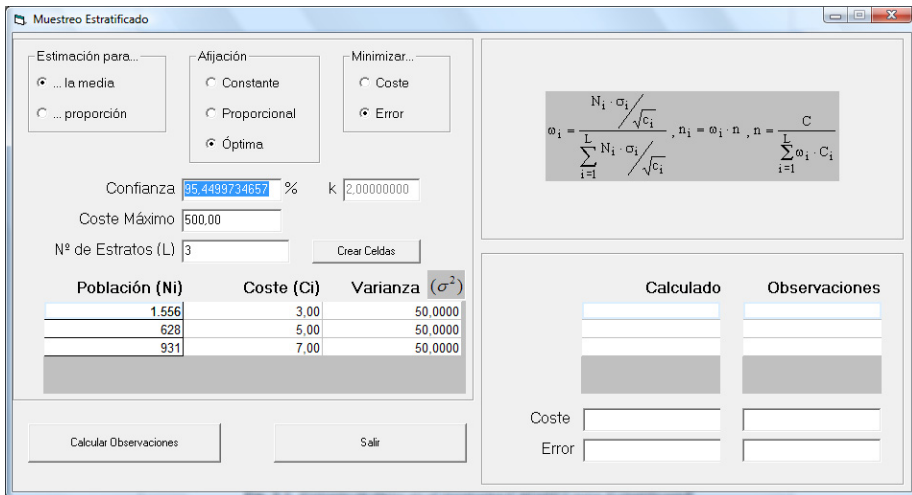
CALnYES, en este caso puede calcular el tamaño de la muestra para estimar una media o proporción. Cuando se trate de un total bastaría con utilizar el caso de media y de proporción correspondiente y ajustar el error.

Recuérdese que en el muestreo aleatorio estratificado la población está dividida en 'L' estratos de tamaños N_1, \dots, N_L , donde $N = N_1 + \dots + N_L$, es el tamaño de la población. Para cada uno se extraen muestras de tamaños n_1, \dots, n_L , respectivamente, donde $n = n_1 + \dots + n_L$, es el tamaño de la muestra total. A la forma de hacer ese reparto se le conoce por 'afijación'.

CALnYES utiliza tres tipos de afijación: uniforme, proporcional y óptima (minimizar coste o error).

La Fig. 4. presenta la pantalla para calcular un tamaño de muestra en un Muestreo Aleatorio Estratificado, una vez seleccionadas las pestañas de ‘Tamaño Muestra’ y posteriormente ‘Muestreo Aleatorio Estratificado’. De nuevo, la aplicación realizará los cálculos matemáticos aplicables según las opciones preferidas obteniendo los tamaños de muestra calculados en cada estrato. Primero se calcula la afijación la cual proporciona un tamaño ‘n’ de muestra total. A partir de este valor es de donde se obtienen los diferentes ‘n_i’, los que serán números reales (‘Calculado’), tras lo que se ejecuta un proceso de redondeo mediante optimización que busca reducir el error/coste según se haya elegido (‘Observaciones’). Al mismo tiempo, informa del error cometido con esa afijación y el coste.

Fig. 4.
Entrada de datos en CALnYES para MAE



En el próximo punto (5) se verá un ejemplo conjunto de los apartados 4 y 5.

5. Estimación de medias y proporciones en mae

Se ha visto en el apartado anterior que el programa calcula el número de observaciones necesarias para que con un muestreo aleatorio estratificado se estime una media o una proporción, según la afijación elegida.

Una vez recogidos los datos CALnYES estima la media o la proporción en la población de la que se han extraído.

Por ejemplo:

Una ladera de una montaña en el Valle del Jerte está dividida en tres bancales con cerezos. Cada uno tiene propiedades diferentes en cuanto a la temperatu-

ra, la presión, la humedad, por lo que se supone una producción de fruta diferente. Los árboles se plantan en líneas de 100 cerezos y hay 8, 6 y 3 líneas, respectivamente, del bancal más bajo al más alto. Se quiere estimar la producción media por cerezo con un 95% de confianza en este año para lo que se coge uno de cada 20 cerezos en cada bancal.

Los costes de muestreo de cada cerezo son mayores al aumentar la altura, de 5, 11 y 17€, respectivamente.

Se desea saber cuántos árboles deberían escoger al año siguiente en cada bancal para que el error al estimar la producción media sea mínimo, si dispone de 1.500€ y cuál es ese error.

La población está dividida en tres grupos según el bancal, así que, el muestreo a utilizar es estratificado con número de estratos 3. Se pretende encontrar una estimación de la producción media por cerezo.

El tamaño de la población de cada bancal (estrato) es de:

Bancal	N_i
1	$100 \cdot 8 = 800$
2	$100 \cdot 6 = 600$
3	$100 \cdot 3 = 300$

Y, puesto que se elige uno de cada veinte cerezos, el tamaño de muestra:

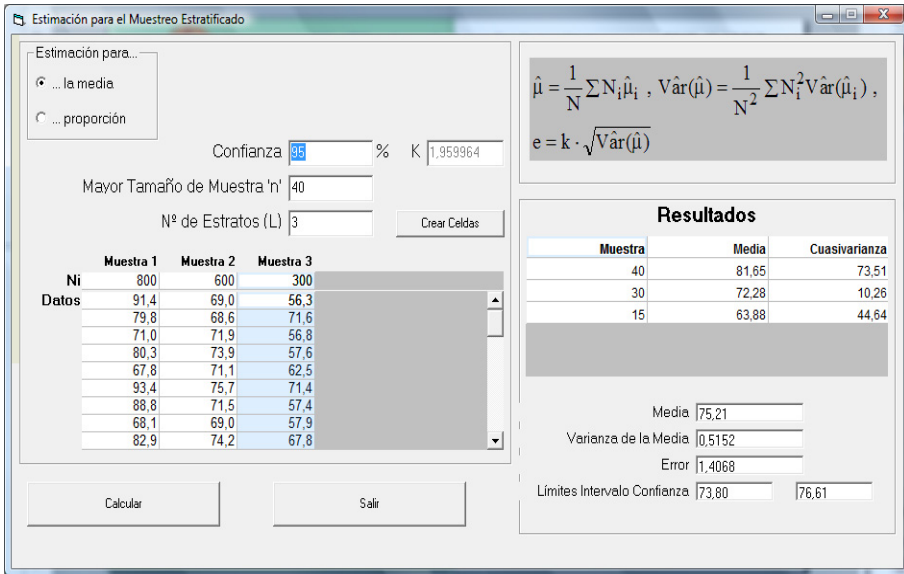
Bancal	n_i
1	$800 \div 20 = 40$
2	$600 \div 20 = 30$
3	$300 \div 20 = 15$

Se elige 'Estimar' y 'Muestreo Estratificado'. Opciones: **Estimación para la media** → **Confianza: 95**; → **Mayor Tamaño de Muestra 'n': 40**; **Nº de Estratos (L): 3**; → **Crear celdas** → **N_i (en fila): 800; 600; 300** **Datos (tecleados o copiados de Excel u otras aplicaciones)**; → **Calcular**.

Se han de dar los datos recogidos en cada una de las muestras para cada uno de los estratos, cada una de ellas ocupará una columna. También se ha de especificar el máximo tamaño de muestra, así como el número de estratos.

Fig. 5.1.

Entrada de datos y resultados en CALnYES para estimación en MAE



Se obtiene una estimación puntual de la media de producción por cerezo en toda la ladera es de 75,21 kgr. La varianza estimada de la estimación de la media es $\text{Var}(\hat{\mu}) = 0,5152$, lo que resulta un error de estimación de $e = 1,4068$. Así, la producción media en el bancal es está entre 73,80 y 76,61 kgr con un 95% de confianza.

La estimación puntual de la media en el primer estrato es $\hat{\mu}_1 = 81,65$ kgr por cerezo en el primer bancal, en el segundo $\hat{\mu}_2 = 72,28$ kgr y en el tercero, $\hat{\mu}_3 = 63,88$, con respectivas cuasivarianzas en las muestras de $s_{c1}^2 = 73,51$, $s_{c2}^2 = 10,26$ y $s_{c3}^2 = 44,64$ (columnas de 'Media' y 'Cuasivarianza').

Ahora se desea hallar el tamaño de muestra con un presupuesto de 1.500€ Se ha de considerar, por tanto, afijación óptima para minimizar el error con un coste dado.

El coste de estudiar un árbol en cada bancal es de 5, 11 y 17€, respectivamente.

Basándose en el estudio anterior como prueba ‘piloto’, se utiliza como estimación de las varianzas las cuasivarianzas anteriores: $\sigma_1^2 = 73,51$, $\sigma_2^2 = 10,25$ y $\sigma_3^2 = 44,64$. Se usa ‘Tamaño de Muestra’ y ‘Muestreo Estratificado’.

Opciones: **Estimación para la media** → **Afijación: Óptima** → **Confianza: 95**; **Coste Máximo: 1500**; **Nº de Estratos (L): 3**; → **Crear celdas** → **Población (Ni): (en columna) 800; 600; 300**; **Coste (Ci): (en columna): 5; 11; 17**; **Varianza (σ_i^2): (en columna): 73,51; 10,26; 44,64** → **Calcular Valores**. Se obtienen los resultados:

Fig. 5.2.

*Entrada de datos y resultados en
CALnYES para tamaño de muestra en MAE*

Estimación para...
 ... la media
 ... proporción

Afijación
 Constante
 Proporcional
 Óptima

Minimizar...
 Coste
 Error

Confianza 95,0000 % k 1,95996400

Coste Máximo 1.500,00

Nº de Estratos (L) 3

Población (Ni)	Coste (Ci)	Varianza (σ_i^2)
800	5,00	73,5100
600	11,00	10,2600
300	17,00	44,6400

$$\omega_i = \frac{N_i \cdot \sigma_i / \sqrt{c_i}}{\sum_{i=1}^L N_i \cdot \sigma_i / \sqrt{c_i}}, n_i = \omega_i \cdot n, n = \frac{C}{\sum_{i=1}^L \omega_i \cdot C_i}$$

Calculado	Observaciones
153,4969	153
28,9968	28
24,3265	25
Coste 1.500,00	1.498,00
Error 0,8327	0,8336

Se han de tomar muestras de tamaños respectivos en cada estrato de 153, 28 y 25 cerezos ($n = 153 + 28 + 25 = 206$, la muestra total) para que no gastar más de 1500 euros y cometer el mínimo error posible. El coste total para obtener esta muestra será de 1.498€ (‘Coste’ en la columna de ‘Observaciones’). El error de esta asignación es de 0,8336 (‘Error’ en la columna de ‘Observaciones’).

Señalar que si se ha de cambiar alguno de los otros valores la aplicación no lo permite a menos que pinchemos en el botón ‘Cambiar valores’. Mientras tanto, el programa permite cambiar el número de observaciones y automáticamente calcula su coste y error.

Si hubiese sido el caso de una proporción, se habrían de dar los casos favorables en cada una de las muestras para cada uno de los estratos, así como el tamaño de cada muestra y de cada población, respectivamente. Se ha de especificar el número de estratos.

Los resultados son: la proporción muestral, el producto de cada proporción por su complementaria ($p \cdot q$) y la varianza estimada de la estimación de la proporción (‘Varianza Prop.’) en cada uno de los estratos. Además, presenta la proporción global estimada (‘Proporción’), la estimación de la varianza de la proporción global estimada (‘Varianza de la Proporción’), el error de estimación y el intervalo de confianza para la proporción en la población.

6. Muestreo por conglomerados

Recuérdese que se consideran N conglomerados de tamaños respectivos M_1, \dots, M_N , donde $M_1 + \dots + M_N = M$ (número de individuos de la población). Se extrae una muestra de ‘ n ’ conglomerados. Se llama \bar{M} al tamaño medio de conglomerado.

Por ejemplo:

Un director de marketing quiere estimar el coste de publicidad promedio mensual por los anuncios en ciertas empresas. El director no puede obtener un coste de cada anuncio, pero puede obtener la cantidad total gastada en las 96 empresas y el número de anuncios en cada fábrica. Entonces decide usar un muestreo tomando cada industria como un conglomerado. Quiere saber cuántos conglomerados debe seleccionar en la muestra si quiere que el límite para el error de estimación sea menor que 2 euros en el coste medio por anuncio. Sabe de datos obtenidos en el mes anterior que la varianza es de 845,56 y que hay 710 anuncios en total.

La población está dividida en conglomerados (cada empresa).

En el supuesto sí se conoce M pues se da el total de anuncios ($M = 710$). En la situación contraria se ha de saber el tamaño promedio de conglomerado, ‘ \bar{M} ’, que si no se conoce se estima por una muestra ‘piloto’ previa.

En *CALnYES* se toma ‘Tamaño muestra’ y ‘Muestreo por Conglomerados en una Etapa’. Opciones: **Estimación para la media** → **Confianza: 95,45; Error: 2; N° Conglomerados en la Población (N): 96; Varianza () : 845,56** → **N° de elementos Totales (M): 710** → **Calcular Tamaño de Muestra.**

Se obtienen los siguientes resultados:

Fig. 6.

Entrada de datos y resultados en CALnYES para MpC

Muestreo por Conglomerados en una etapa

Estimación para...

... la media

... el total

... proporción

Confianza 95,4500 %

Error 2,0000

Nº Conglomerados en la Población (N) 96

Varianza (σ^2) 845,5600

Nº de elementos Totales (M) 710,0000

Tamaño Medio de Conglomerado (\bar{M}) 1,2340

Recuerde!

$$\hat{\sigma}^2 = \frac{\sum (\tau_i - \hat{\mu} \cdot M_i)^2}{n - 1}$$

Valor calculado de 'k' 2,00000000

Tamaño Calculado 13,31459138

Tamaño de la Muestra 14

Calcular Tamaño de Muestra

Salir

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

$$n_{\infty} = \frac{k^2 \cdot \sigma^2}{e^2 \cdot \bar{M}^2}$$

$$n = \frac{1}{\frac{1}{n_{\infty}} + \frac{1}{N}}$$

Se necesita una muestra de 14 empresas para que el error no sobrepase los 2€ ('Tamaño de la muestra'). Según las fórmulas de cálculo presentadas el tamaño de la muestra debería ser de 13,31459138 ('Tamaño Calculado'), el cual se redondea por exceso, para obtener el número de observaciones (conglomerados) definitivo que se han de utilizar en el estudio.

7. Muestreo por conglomerados en una etapa mezclado con estratificado

Recuérdese que se llama N_i al número de conglomerados del estrato ' i ', $i = 1, \dots, L$ y n_i al número de conglomerados seleccionados en la muestra del estrato ' i '. Ahora, M_{ij} será el número de individuos del estrato ' i ', $i = 1, \dots, L$ del conglomerado ' j ', $j = 1, \dots, N_i$. Se verifica que $\sum_{i,j} M_{ij} = M$, el número de individuos totales.

Se extrae una muestra en cada estrato de tamaño n_i conglomerados, $i = 1, \dots, L$, donde $n_1 + \dots + n_L = n$. Así se han de hallar cada uno de los valores n_i .

Aquí, al igual que en el muestreo por conglomerados, no se ha de distinguir el caso de si se desea estimar una media o una proporción. Se recuerda que sólo varía a la hora de calcular la varianza a partir de la prueba 'piloto' primaria.

El programa utiliza tres tipos de afijación: uniforme, proporcional y óptima (coste o error mínimo).

Por ejemplo:

Una empresa hispano-francesa de moda quiere conocer el tiempo medio que tarda en confeccionar un vestido de novia. Dispone de 30 talleres en España y 45 en Francia. Para ello quiere utilizar unas muestras de talleres considerando que existen dos estratos (España y Francia). Se desea hallar el tamaño de una muestra en cada país para hacer un muestreo en toda la empresa sin que el error sobrepase media hora. Partiendo de que el coste por observación en España es de 10€ y 30€ en Francia, se desea saber cuál es el coste total. Se sabe de estudios anteriores que la varianza en España es de 22.396,52 y el mismo dato para Francia es 61.724,24. No se conoce el número total de trajes que se confeccionan en ambos países pero sí que la media de trajes confeccionados por taller es de 122,5.

La población está dividida en talleres de los que se desea extraer una muestra. A su vez se compone de dos estratos configurados por los dos países (España y Francia). De esta manera, el muestreo a utilizar es por conglomerados mezclado con estratificado. Se pretende encontrar un número de talleres a estudiar en cada país (estrato), cuya suma será el tamaño de muestra total, para un error máximo de 0,5 horas.

Se sabe que hay $N_1 = 30$ conglomerados en España y $N_2 = 45$ en Francia y se tienen sendas varianzas $\sigma_1^2 = 22396,52$ y $\sigma_2^2 = 61724,24$.

No se conoce el valor de M , pues no hay datos sobre el número total de trajes, pero sí que la media de vestidos confeccionados por taller es de $122,5 = \bar{M}$. Se utiliza la afijación óptima con un error dado para estimar una media (tiempo medio); es decir, lo que se quiere es minimizar el coste.

Introduciendo los datos como se ilustra en la figura 7. de la siguiente manera:

‘Tamaño muestra’ y ‘Muestreo por Conglomerados en una etapa mezclado con Estratificado’. Opciones: **Afijación: Óptima → Minimizar: Coste → Confianza: 95,45; Error Máximo: 0,5; Tamaño Medio de Conglomerado (\bar{M}): 122,5; N° de Estratos (L): 2; → Crear celdas → Población (Ni): (en columna) 30; 45; Coste (Ci): (en columna) 10; 30; Varianza (σ^2): (en columna) 22396,52; 61724,24 → Calcular Observaciones;** se obtienen los resultados:

Se han de tomar sendas muestras de tamaños 13 y 18 ($n = 13 + 18 = 31$, la muestra total) para que no se sobrepase un error máximo de 0,5 horas. El error real al extraer esa muestra es de 0,4950 (‘Error’ en la columna de ‘Observaciones’). El coste total es de 670€ (‘Coste’ en la columna ‘Observaciones’).

Según las fórmulas de cálculo presentadas el tamaño de la muestra debería ser en cada estrato de 12,2744; 17,6469 (‘Calculado’), y, puesto que han de ser números naturales, se redondean optimizando para que sin sobrepasar el error sea de coste más bajo, que es el número de observaciones definitivo que se han de utilizar en el estudio.

Fig. 7.

*Entrada de datos y resultados en CALnYES
para MpC mezclado con estratificado*

Afijación:
 Constante Óptima
 Proporcional

Minimizar...:
 Coste
 Error

Confianza: 95,4500 % k: 2,00000000

Error Máximo: 0,50

Tamaño de la Población (M):

Tamaño Medio de Conglomerado (M̄): 122,5000

Nº de Estratos (L): 2

Población (Ni)	Coste (Ci)	Varianza (σ ²)
30	10,00	22.396,5200
45	30,00	61.724,2400

Recuerde...:

$$\hat{\sigma}^2 = \frac{\sum (\xi_i - \hat{\mu} \cdot M_i)^2}{n - 1}$$

$$\omega_i = \frac{N_i \cdot c_i / \sqrt{c_i}}{\sum_{i=1}^L N_i \cdot c_i / \sqrt{c_i}}, n_i = \omega_i \cdot n, n = \frac{\sum_{i=1}^L N_i^2 \cdot c_i^2 / \omega_i}{\frac{M^2 \cdot N^2 \cdot e^2}{k^2} + \sum_{i=1}^L N_i \cdot c_i^2}$$

	Calculado	Observaciones
Coste	12,2744	13
Error	17,6469	18

Coste: 652,15 670,00
Error: 0,5062 0,4950

Al igual que en el muestreo estratificado se podrían cambiar las asignaciones muestrales y el coste y el error serán recalculados automáticamente.

Añadir que si se pretende cambiar algún dato para realizar nuevos cálculos, se utiliza 'Cambiar Valores'.

8. Tamaño de la población (n)

Como es sabido, ya no se busca estimar un parámetro poblacional, sino el tamaño de la población. Muchas veces éste se desconoce a la hora de hacer un estudio y haría falta precisar algunas estimaciones; otras veces, lo que interesa es el valor en sí mismo.

Se utilizarán dos métodos para la estimación: muestreo directo y muestreo inverso.

Recuérdese que:

- Muestreo directo: se elige una muestra de tamaño 't', se marcan las unidades muestrales y, posteriormente, se extrae una segunda muestra de tamaño 'n' de entre las que 's' están marcadas.
- Muestreo inverso: se elige una muestra de tamaño 't', se marcan las unidades muestrales y se extrae una muestra posteriormente de tamaño 'n' hasta encontrar 's' unidades marcadas.

Como ejemplo:

En estudio de mercado se desea saber los consumidores potenciales de un producto por Internet. Se quiere calcular los que se debería encontrar inicialmente para ‘marcar’ y poder estimar el total con un error del 10%, si en la segunda muestra se encontrasen 75. De estudios anteriores se tiene que el número estimado de consumidores fue de 750.

Se desea estimar un tamaño de población de consumidores. Según el estudio deseado, el método a elegir es muestreo directo ($n = 75$, es fijo) y no se sabe cuántos se van a encontrar inicialmente (t ?, también va a ser fijo).

Puesto que se asume un error del 10% ha de ser relativo, pues se está hablando de unidades (consumidores), así que hay que calcular el error absoluto y la varianza, que es el dato que solicita CALnYES:

$$e = e_r \cdot N = 0,10 \cdot 750 = 75, \text{ y, como } e = 2 \cdot \sqrt{\text{Var}(\hat{N})},$$

$$\text{Var}(\hat{N}) = 1406,25.$$

Se elige ‘Tamaño muestra’ y ‘Estimación del Tamaño de la Población’. Opciones: **Tipo de Muestreo: Directo** → **Valor Conocido: ‘n’ conocida; Tamaño de la población (N): 750; Varianza (\hat{N}): 1406,25; Valor de ‘n’: 75** → **Calcular**.

Fig. 8.
*Entrada de datos y resultados en CALnYES
para cálculo de n en la estimación de N*

Se necesita una primera muestra de 632 consumidores para que el error no sobrepase el 10% (‘Valor de t ’). Éstas se ‘marcarán’ y después se extraerá otra muestra de 75 y se contarán las que estén señaladas.

9. Características técnicas del programa *calnyes*

La aplicación está desarrollada con Visual Basic versión 6.0 utilizando objetos y librerías de Windows.

Funciona bajo cualquier versión de este sistema operativo a partir de Windows 98.

El programa es gratuito y se puede descargar en la página de Internet con dirección <http://web.usal.es/~mateca/otrasDescargas.htm>. Mediante un correo electrónico a mateca@usal.es se solicita una clave necesaria para la descomprimirlo.

En la página web existen dos versiones: el programa solo y la versión completa que incluye todos los objetos necesarios para la instalación. De esta forma, alguien que realice la instalación por primera vez, debería bajarse la versión completa mientras que para las actualizaciones sólo sería necesario descargar el programa y sustituirlo por el anterior.

Es posible que el programa funcione directamente sin realizar la instalación completa, cuando se ejecute en ordenadores que tengan instalado Microsoft Office.

Por último indicar que la aplicación tiene una fecha de caducidad para obligar a los usuarios a instalar las últimas versiones que, en todo caso, serán gratuitas.

10. Otras características del programa

CALnYES está diseñado para admitir múltiples idiomas (actualmente en español e inglés). En el menú inicial podemos seleccionarlo (véase Fig. 1.). La inclusión de nuevos idiomas se reduce a la traducción de una tabla de rótulos y mensajes.

Para ayudar en la comprensión de los cálculos la aplicación proporciona en todo momento las expresiones matemáticas que está utilizando, tanto para hallar el tamaño de la muestra como para hacer estimaciones. Éstas variarán según se vayan eligiendo unas u otras opciones.

Para que no pueda haber confusiones, si después de haber hallado resultados se modifica un dato inicial, el programa pone dichos resultados en color rojo. Esto no evita el que se puedan cambiar los datos en el momento que se desee y recalcular.

Añadir que en la opción de ‘Muestreo Estratificado’ o ‘Muestreo por Conglomerados Mezclado con Estratificado’, se ha de usar el botón ‘Cambiar Valores’ para que lo anterior sea posible.

Comentar que *CALnYES* ha pasado varios controles de calidad para comprobar que todos y cada uno de los resultados son fiables al cien por cien en todas y cada una de las opciones hasta el momento posibles.

CALnYES es parte de un gran proyecto en el que en el presente ya se está trabajando.

11. Conclusiones

CALnYES es un programa informático que presenta de una forma sencilla, rápida y cómoda el cálculo de tamaños de muestra en muestreo aleatorio simple,

estratificado, por conglomerados, por conglomerados mezclado con estratificado, con las distintas posibilidades de afijaciones por estratos. Además, incluye el cálculo de tamaños de muestra en muestreo directo e inverso para la estimación de tamaños de población. Y realiza la estimación puntual, la varianza estimada, error e intervalo de confianza en medias, totales y proporciones en muestreo aleatorio simple y estratificado.

La aplicación evita al usuario tener que realizar cálculos complejos, no hay más que introducir los datos en cada opción correspondiente.

Cuando sea necesario usar una gran cantidad de datos o estén en forma de tabla (cálculo del tamaño de muestra y estimaciones en estratificados), el programa utiliza objetos de tipo Excel que permite al usuario copiar y pegar los datos de otras aplicaciones (Excel, Access, etc.) y ahorrar trabajo en teclearlos.

El programa presenta en pantalla las fórmulas que está utilizando en cada momento.

Además incorpora unos algoritmos de optimización que aseguran que los tamaños de muestra hallados son los mejores posibles.

Cabe destacar, que aún así, en los casos de muestreo estratificado es posible cambiar los tamaños de muestra en cada estrato y *CALnYES* recalcula automáticamente error y costes.

Se ha pretendido que esta aplicación además de ser una herramienta para la investigación mediante técnicas que utilicen el muestreo, en particular, las encuestas, también sea didáctica y pueda ser usada en clases de diversas titulaciones y asignaturas.

12. Bibliografía

- Azorín, F. y Sánchez Crespo, J. L. (1994). *Métodos y aplicaciones de muestreo*. Alianza Editorial. Madrid
- Cabero Morán, M. T. (2007). *MASYE: un programa para el cálculo de los tamaños de muestra en muestreo aleatorio simple y estratificado*. Libro de Actas del X Congreso de Metodología de las CC. Sociales y de la Salud. Barcelona.
- Cabero Morán, M. T. (2007). *Cálculo y optimización de tamaños de muestras*. Libro de Actas del Congreso Informática en Salud. La Habana (Cuba).
- Cabero Morán, M. T., García Martín, M., Prieto García, M. M. y Mecoleta Finó, S. (2007) *MASYE2: optimización de tamaños de muestra en diferentes tipos de muestreo y en la estimación del tamaño de la población*. Libro de Actas del XXX Congreso Nacional de Estadística e I.O. Valladolid
- García, P.A., González, A. y Maldonado, J.A. (1999): *Problemas en el Diseño y Validación de Cuestionarios: tratamiento con QUESTPOT v.1.2*. Estadística Española Vol. 41, Núm. 144, 1999, págs. 19 – 46.
- Scheaffer, R. (1987). *Elementos de Muestreo*. Grupo Editorial Iberoamericana. Méjico D.F. (Orig. 1986)