Capítulo 1

Muestreo Aleatorio Simple

Este método de muestreo proporciona un punto de partida para una exposición de los métodos de muestreo probabilístico no porque sea uno de los métodos de muestreo más utilizados sino porque constituyen la base de métodos de muestreo más complejos. Dependiendo si el muestreo es con reposición o sin reposición, podemos hablar de muestreo aleatorio simple con reposición o sin reposición respectivamente.

De manera formal, este diseño básico o técnica de muestreo se define de la siguiente manera

Definición 1.1 *Si se selecciona un tamaño de muestra n de una población de tamaño N de tal manera que cada muestra posible de tamaño n tenga la misma probabilidad de ser seleccionada, el procedimiento de muestreo se denomina muestreo aleatorio simple. A la muestra así obtenida se le denomina muestra aleatoria simple.*

Considerando muestreo aleatorio sin reposición, se obtiene la muestra unidad a unidad de forma aleatoria sin reposición a la población de las unidades previamente seleccionadas, teniendo presente además que el orden de colocación de los elementos en las muestras no interviene, es decir, muestras con los mismos elementos colocados en orden distinto se consideran iguales. De esta forma, las muestras con elementos repetidos son imposibles. Bajo muestreo aleatorio con reposición, las unidades seleccionadas son devueltas de nuevo a la población.

Expondremos una manera de seleccionar una muestra aleatoria simple utilizando un ejemplo concreto. Se pretende realizar un estudio sobre los hábitos de lectura en los estudiantes de Politécnica. Las alumnos que actualmente estudian en Politécnica son un total de 544 alumnos y se quiere extraer una muestra aleatoria simple de 65 alumnos. Una manera de extraer una muestra aleatoria simple consiste en asignar a cada alumnos un número del 1 al 544 asociando cada número a un único individuo. Una vez realizado esa asignación, se introducen 544 bolas numeradas en una urna (cada una con un número del 1 al 544), se mezclan cuidadosamente y de manera adecuada y entonces se seleccionan 65

bolas al azar. Si todo el proceso se realiza de manera adecuada, las bolas seleccionadas constituirían una muestra aleatoria simple de 65 estudiantes. Aunque es conceptualmente simple, este método es un poco trabajoso de ejecutar y depende de que las bolas se hayan mezclado de manera adecuada y que todas las bolas tengan el mismo peso y rozamiento.

Otra manera de seleccionar esta muestra aleatoria simple consistiría en utilizar una tabla de números aleatorios. Una tabla de números aleatorios es un conjunto de dígitos generado de modo que, normalmente, la tabla contendrá a cada uno de los diez dígitos $(0, 1, \ldots, 9)$, en proporciones aproximadamente iguales, sin mostrar tendencias en el patrón que se generan los dígitos. Por lo tanto, si se selecciona un número en un lugar aleatorio de la tabla, es igualmente probable que sea cualquiera de los dígitos entre el 0 y el 9. Estas tablas se construyen para asegurar que cada dígito, cada par de dígitos, cada tres dígitos, ... aparecen con la misma frecuencia. En el caso de extraer una muestra aleatoria simple, se elige un lugar para empezar a leer dichos números aleatorios. Después se selecciona una dirección (arriba, abajo, derecha e izquierda) y se van recogiendo dígitos de dos en dos hasta que se consiga el tamaño muestral adecuado. Utilizando este método, un elemento puede aparecer más de una vez. Si queremos extraer una muestra aleatoria simple sin reposición, la solución es ignorar los elementos repetidos.

Las ventajas que tiene este procedimiento de muestreo son las siguientes:

- Sencillo y de fácil comprensión.
- Cálculo rápido de medias y varianzas.
- Existen paquetes informáticos para analizar los datos

Por otra parte, las desventajas de este procedimiento de muestreo son:

- Requiere que se posea de antemano un listado completo de toda la población.
- Si trabajamos con muestras pequeñas, es posible que no representen a la población adecuadamente.

A continuación pasamos a describir este procedimiento de muestreo considerando muestreo sin reposición.

1.1. Diseño muestral

Vamos a analizar el diseño de este procedimiento de muestreo. Supongamos en todo momento que el tamaño de la población es N y el tamaño de la muestra es n.

1.1.1. Probabilidad de una muestra cualquiera

Dada la forma de definirse el procedimiento de selección de la muestra, el conjunto formado por todas las muestras S tiene un total de

$$C_{N,n} = \binom{N}{n},$$

muestras posibles, ya que estamos considerando muestras no ordenadas. Luego si todas las muestras son equiprobables, la probabilidad de cada muestra viene dada por

$$P(s) = \frac{1}{\binom{N}{n}}, \quad \forall s \in S$$

1.1.2. Probabilidad de primera inclusión

Calculemos la probabilidad que tiene cualquier unidad de la población de pertenecer a la muestra, o lo que es lo mismo, calcularemos π_i for $i = 1, 2, \ldots, N$. Por ello, consideramos el número de muestras posibles que se pueden formar con los elementos de la población y que contengan al elemento u_i. En este caso, el total de muestras que contienen a dicho elemento viene dado por

$$C_{N-1,n-1} = \binom{N-1}{n-1},$$

ya que en este caso se fija el elemento u_i y las muestras posibles resultan de las formas posibles de seleccionar de entre los $N-1$ elementos de la población restantes $n-1$ de ellos para la muestra (el elemento u_i ya pertenece a la muestra). Para $i = 1, 2, \ldots, N$, se tiene que

$$\begin{aligned} \pi_i &= P(u_i \in s) = \\ &= \frac{\text{Total de muestras que contienen a } u_i}{\text{Total de muestras}} = \frac{\binom{N-1}{n-1}}{\binom{N}{n}} = \frac{n}{N} \end{aligned}$$

1.1.3. Probabilidad de segunda inclusión

Vamos a calcular la probabilidad que tiene cualquier par de unidades de la población de pertenecer a una muestra determinada. Para ello, notemos que el número de muestras posibles que pueden formarse con los elementos de la población y que contengan al par (u_i, u_j) con $u_i \neq u_j$ es igual a

$$C_{N-2,n-2} = \binom{N-2}{n-2},$$

ya que en este caso se fija el par de elementos (u_i, u_j) y las muestras posibles resultan de las formas posibles de seleccionar de entre los N-2 elementos de la

población restantes $n-2$ de ellos para la muestra (los elementos u_i y u_j ya están fijos en la muestra). Tenemos entonces que

$$\begin{aligned}\pi_{ij} &= P((u_i,u_j)\in s) = \frac{\text{Casos favorables}}{\text{Casos posibles}} \\ &= \frac{\text{Total de muestras que contienen a } (u_i,u_j)}{\text{Total de muestras}} = \frac{n(n-1)}{N(N-1)}\end{aligned}$$

1.2. Estimadores lineales insesgados en muestreo aleatorio simple

Tal y como señalábamos en el capítulo anterior, si el parámetro poblacional tiene una expresión lineal del tipo

$$\theta = \sum_{i=1}^{N} Y_i,$$

entonces el estimador de Horvitz-Thompson para dicho parámetro poblacional viene dado por

$$\widehat{\theta}_{HT} = \sum_{i=1}^{n} \frac{Y_i}{\pi_i},$$

de modo que $E[\widehat{\theta}_{HT}] = \theta$ siendo π_i la probabilidad de que la unidad u_i pertenezca a una muestra determinada o probabilidad de primera inclusión. Partimos como es usual de una población formada por N elementos

$$\Omega = \{u_1, u_2, u_3, \ldots, u_N\},$$

en los que se estudiar una variable de interés X que toma los valores

$$X(\Omega) = \{X_1, X_2, X_3, \ldots, X_N\},$$

sobre cada elemento de la población. Para ello, se selecciona una muestra de tamaño n dada por

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

en los que la variable X toma los valores

$$X(s) = \{X_1, X_2, X_3, \ldots, X_n\},$$

sobre cada uno de los elementos de la muestra.

Como en muestreo aleatorio simple sin reposición la probabilidad de primera inclusión π_i viene dado por $\pi_i = n/N$, ya podemos especificar los estimadores lineales insesgados para los parámetros poblacionales más comunes a estimar. Tendremos que

- Total

$$\theta = X = \sum_{i=1}^{N} X_i \Longrightarrow Y_i = X_i \Longrightarrow \widehat{X} = \sum_{i=1}^{n} \frac{X_i}{\pi_i} = \sum_{i=1}^{n} \frac{X_i}{n/N} = \frac{N}{n} \sum_{i=1}^{n} X_i = N\widehat{\bar{X}}$$

- Media

$$\theta = \bar{X} = \sum_{i=1}^{N} X_i/N \Longrightarrow Y_i = X_i/N \Longrightarrow \widehat{\bar{X}} = \sum_{i=1}^{n} \frac{X_i}{Nn/N}$$

- Total de clase

$$\theta = A = \sum_{i=1}^{N} A_i \Longrightarrow Y_i = A_i \Longrightarrow \widehat{A} = \sum_{i=1}^{n} \frac{A_i}{n/N} = N\frac{1}{n} \sum_{i=1}^{n} A_i$$

- Proporción

$$\theta = P = \sum_{i=1}^{N} A_i/N \Longrightarrow Y_i = A_i/N \Longrightarrow \widehat{P} = \sum_{i=1}^{n} \frac{A_i/N}{n/N} = \frac{1}{n} \sum_{i=1}^{n} A_i$$

Evidentemente cualquier de estos estimadores nos indican muy poco acerca del parámetro poblacional a estimar a menos que sea posible evaluar la bondad del estimador. Por lo tanto, además de estimar los parámetros poblacionales, se desearía fijar un límite sobre el error de estimación. Mediante ciertos cálculos, es posible calcular la varianza del estimador de Horvitz-Thompson para cada uno de los estimadores. Las varianzas de los estimadores anteriores nos van a proporcionar los errores estándar de estimación y vienen dado por:

$$\begin{aligned} Var(\widehat{X}) &= N^2(1-f)\frac{S^2}{n} \\ Var(\widehat{\bar{X}}) &= (1-f)\frac{S^2}{n} \\ Var(\widehat{P}) &= \frac{N}{N-1}\frac{1}{n}(1-f)PQ \\ Var(\widehat{A}) &= \frac{N^3}{N-1}\frac{1}{n}(1-f)PQ \end{aligned}$$

Vamos a analizar las varianzas de los estimadores. En el caso del estimador del total y de la media poblacional dependen de S^2 que es la cuasi-varianza poblacional. Esta cuasi-varianza poblacional S^2 tiene la siguiente expresión

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{N} (X_i - \bar{X})^2,$$

donde $\bar{X}$ representa la media poblacional. Esta cuasi-varianza poblacional puede expresarse también de la siguiente manera.

$$\begin{aligned} S^2 &= \frac{1}{N-1}\left[\sum_{i=1}^{N}(X_i^2+(\bar{X}))^2 - 2X_i\bar{X}\right] \\ &= \frac{1}{N-1}\left[\sum_{i=1}^{N}X_i^2 + N(\bar{X})^2 - 2(\bar{X})^2N\right] \\ &= \frac{1}{N-1}\left[\sum_{i=1}^{N}X_i^2 - N(\bar{X})^2\right]. \end{aligned}$$

En el caso de la proporción y el total de clase, la varianza de dichos estimadores dependen de P y de Q. El parámetro P corresponde a la proporción de clase dentro de la población, es decir,

$$P = \frac{1}{N}\sum_{i=1}^{N}A_i,$$

y $Q = 1 - P$. Al parámetro f se le denomina *fracción de muestreo* y viene dado por

$$f = n/N,$$

y representa la fracción de la población N contenida en la muestra o la relación existente entre el tamaño de la muestra y el tamaño de la población. Siempre $n \leq N$; si $n = N$ entonces $f = 1$ y $1 - f = 0$. Por otra parte, si $n << N$, entonces $n/N \cong 0$ y por lo tanto $1 - f \approx 1$ y las varianzas de los estimadores serán mayores. A la diferencia $1 - f$ se le denomina *factor de corrección de población finita* y tiene en cuenta el hecho de que una estimación basada en una muestra de tamaño $n = 10$ de una población de $N = 20$ elementos, contiene más información acerca de la población que una muestra de $n = 10$ de una población de $N = 20000$ elementos.

Ejemplo 1 *Consideramos una población de 4 elementos dada por*

$$\Omega = \{u_1, u_2, u_3, u_4\},$$

sobre los que medimos una variable X obteniendo como resultados $\{8, 3, 4, 6\}$ en cada uno de los elementos de la población. Mediante muestreo aleatorio simple sin reposición, se extraen muestras de tamaño 2. Se pide:

- *Número de elementos del espacio muestral.*
- *Especificar dicho espacio muestral y determinar las probabilidades asociadas a las muestras.*
- *Hallar las distribuciones en el muestreo de los estimadores de la media y del total de X así como la varianza de los estimadores. Calcular la cuasivarianza de cada muestra.*

- *Comprobar la insesgadez de los estimadores y que se cumple*

$$Var(\widehat{\bar{X}}) = (1-f)\frac{S^2}{n}, \quad Var(\widehat{X}) = N^2(1-f)\frac{S^2}{n},$$

y además

$$E[\widehat{S}^2] = S^2.$$

Como vemos, la varianza de estos estimadores depende de una serie de parámetros poblaciones por lo que, en la mayoría de los casos prácticos, estos datos no estarán disponibles. En el caso en el que no dispongamos de estos valores poblacionales, utilizaremos estimaciones para estas varianzas. Las estimaciones son las siguientes:

$$\begin{aligned}
\widehat{V}(\widehat{X}) &= N^2(1-f)\frac{\widehat{S}^2}{n} \\
\widehat{V}(\widehat{\bar{X}}) &= (1-f)\frac{\widehat{S}^2}{n} \\
\widehat{V}(\widehat{P}) &= (1-f)\frac{1}{n-1}\widehat{P}\widehat{Q} \\
\widehat{V}(\widehat{A}) &= N^2(1-f)\frac{1}{n-1}\widehat{P}\widehat{Q}
\end{aligned}$$

donde la cantidad $\widehat{S}^2$ representa la cuasi-varianza muestral y que viene dada por

$$\widehat{S}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\bar{X}})^2, \tag{1.1}$$

siendo

$$\widehat{\bar{X}} = \frac{1}{n}\sum_{i=1}^{n}X_i,$$

and $\widehat{P}$ representa la probabilidad muestral, o lo que es lo mismo,

$$\widehat{P} = \frac{1}{n}\sum_{i=1}^{n}A_i,$$

luego $\widehat{Q} = 1 - \widehat{P}$.

Notar que $\widehat{S}^2$ dada en (1.1) puede expresarse como

$$\begin{aligned}
\widehat{S}^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \widehat{\bar{X}})^2 \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2 - n\widehat{\bar{X}}^2\right).
\end{aligned}$$

Ejemplo 2 *Una empresa industrial está interesada en el tiempo por semana que los científicos emplean para ciertas tareas triviales. Las hojas de control del tiempo de una muestra aleatoria simple de* $n = 50$ *empleados muestran que la cantidad promedio de tiempo empleado en esas tareas es de 10.31 horas, con una cuasi-varianza muestral* $S^2 = 2{,}25$*. La compañía emplea* $N = 750$ *científicos. Estimar el número total de horas por trabajador que se pierden por semana en tareas triviales y establecer el error de estimación asociada a dicha estimación.*

Solución. *La población se compone de* $N = 750$ *empleados, de los cuales se selecciona una muestra aleatoria de* $n = 50$ *hojas de control del tiempo. La cantidad promedio de tiempo que se pierde por los 50 empleados fue de* 10,31 *horas por semana. Luego la estimación del total de horas de trabajador que se pierden por semanas en tareas triviales* X *viene dada por*

$$\widehat{X} = N\widehat{\bar{X}} = 750(10{,}31) = 7732{,}5 \, \textit{horas.}$$

El error asociado a dicha estimación viene dada por

$$\sqrt{(\widehat{V}(\widehat{X}))} = \sqrt{(750)^2 \frac{2{,}25}{50} \frac{750 - 50}{750}} = 153{,}7 \, \textit{horas.}$$

Ejemplo 3 *Se seleccionó una muestra aleatoria simple de* $n = 100$ *estudiantes de último año de un IES con* $N = 300$ *estudiantes para estimar*

- *la fracción de estudiantes que han tenido trabajos a tiempo parcial durante su estancia en el instituto,*
- *la fracción de estudiantes del último año que asistirán a la universidad.*

Sean X_i *e* Y_i *(*$i = 1, 2, \ldots, 100$*) las respuestas del* i*-ésimo estudiante seleccionado. Se establecerá que* $X_i = 0$ *si el* i*-ésimo estudiante no ha tenido un trabajo a tiempo parcial durante su estancia en el colegio y* $X_i = 1$ *si lo ha tenido. Por otra parte,* $Y_i = 0$ *si el* i*-ésimo estudiante no piensa ir a la universidad y si* $Y_i = 1$ *si si piensa ir. Estimar* P_2 *la proporción de estudiantes de último año que piensa asistir a la universidad y* P_1 *la proporción de estudiantes de último año que ha tenido un trabajo a tiempo parcial considerando que*

$$\sum_{i=1}^{100} X_i = 15, \quad \sum_{i=1}^{100} Y_i = 65,$$

y determina además sus errores de muestreo

Ejemplo 4 *Una gran empresa constructora tiene 120 casas en diversas etapas de construcción. Para estimar la cantidad total (en miles de euros) que será registrada en el inventario de la construcción en proceso, se seleccionó una muestra aleatoria simple de 12 casas y se determinaron los costes acumulados en cada una de ellas. Los costos obtenidos para las 12 casas fueron los siguientes:*

$$35{,}5, 30{,}2, 28{,}9, 36{,}4, 29{,}8, 34{,}1, 32{,}6, 26{,}4, 38, 38{,}2, 32{,}2, 27{,}5.$$

- *Estimar los costes totales acumulados para las 120 casas y dar una estimación del error de muestreo. Dar un intervalo de confianza al* 95 % *para el coste total.*
- *Estimar la proporción de casas cuyos costes de construcción superan los 32.000 euros. Dar una estimación del error.*

Para estimar los costes totales acumulados para las 120 casas tenemos en cuenta que el estimador lineal insesgado del total de una característica X *sobre una población viene dado por:*

$$\widehat{X} = N\widehat{\bar{X}},$$

en este caso se tiene que $N = 120$ *y la media muestral será*

$$\widehat{\bar{X}} = \frac{1}{12}\sum_{i=1}^{12} X_i = 32{,}4833,$$

y por lo tanto

$$\widehat{X} = N\widehat{\bar{X}} = 120{\cdot}32{,}4833 = 3897{,}996 \cong 3898,$$

es decir, el coste total acumulado estimado para las 120 casas será de 3898 miles de euros.

Vamos a dar una estimación de la varianza de dicho estimador. Utilizando las fórmulas anteriores, se tiene que:

$$\widehat{V}(\widehat{X}) = N^2(1-f)\frac{\widehat{S}^2}{n}.$$

Calculamos la cuasivarianza muestral de los costes acumulados

$$\widehat{S}^2 = \frac{1}{11}\sum_{i=1}^{12}(X_i - \widehat{\bar{X}})^2 = \frac{\sum_{i=1}^{12} X_i^2 - n(\widehat{\bar{X}})^2}{n-1} = \frac{12839{,}36 - 12{\cdot}32{,}4833^2}{11} = 16'1233,$$

entonces

$$\widehat{V}(\widehat{X}) = 17410,$$

y la correspondiente estimación para el error de muestreo será

$$\sigma(\widehat{X}) = 131'958948$$

El intervalo de confianza al 95 % *viene dado por*

$$(\widehat{X} - z_{1-\alpha/2}\widehat{\sigma}_{\widehat{X}}, \widehat{X} + z_{1-\alpha/2}\widehat{\sigma}_{\widehat{X}}) = (3639'4, 4156'6).$$

De la muestra formada por 12 casas, únicamente los costes de construcción de 7 casas sobrepasan los 32000 euros, por lo tanto, la estimación de la proporción de casas que superan los 32000 euros es de

$$\widehat{P} = \frac{1}{n}\sum_{i=1}^{n} A_i = \frac{7}{12} = 0'5833,$$

o lo que es lo mismo el 58,33 % *de las casas sobrepasan los 32000 euros. Una estimación de la varianza del estimador* $\widehat{P}$ *viene dado por*

$$\widehat{V}(\widehat{P}) = (1-f)\frac{1}{n-1}\widehat{P}\widehat{Q} = 0'0199$$

de manera que el error asociado para dicha varianza es de

$$\widehat{e}(\widehat{P}) = 0'1411.$$

1.3. Tamaño de muestra fijado el error de estimación

Una cuestión muy importante en muestreo consiste en conocer el tamaño de muestra adecuado para cometer un determinado error de muestreo. En alguna etapa del diseño del procedimiento de muestreo, alguien debe tomar una decisión acerca del tamaño de la muestra que se seleccionará de la población. Como es natural, al aproximar las características poblacionales mediante estimadores basados en la muestra se comete un error, error que mide la representatividad de dicha muestra. Dependiendo del coste del muestreo, del presupuesto disponible y de otros muchos factores fijaremos un error de muestreo que en todo caso debe ser el mínimo posible. Dicho error de muestreo puede venir dado en términos absolutos, en términos relativos o sujeto adicionalmente a un coeficiente de confianza dado (sujeto a unos límites de tolerancia).

A continuación, calcularemos los tamaños de muestra necesarios para cometer un error de muestreo dado al estimar las características poblacionales más comunes mediante muestreo aleatorio simple sin reposición. Inicialmente distinguiremos entre el error común de muestreo $\epsilon = \sigma(\widehat{\theta})$ dado por la desviación típica del estimador y el error relativo de muestreo dado por el coeficiente de varianza del estimador,

$$e_r(\widehat{\theta}) = CV(\widehat{\theta}) = \frac{\sigma(\widehat{\theta})}{E(\widehat{\theta})}.$$

1.3.1. Tamaño de la muestra para un error de muestreo dado

El número de observaciones necesarias para estimar un parámetro poblacional θ a partir de un estimador $\widehat{\theta}$ con un error de estimación ϵ se encuentra resolviendo la siguiente expresión para n

$$\sqrt{Var(\widehat{\theta})} = \epsilon.$$

Analizaremos esta expresión para cada uno de los estimadores propuestos.

- Estimador de la media

$$\epsilon = \sigma(\widehat{\bar{X}}) = \sqrt{(1-f)\frac{S^2}{n}} \Longrightarrow \epsilon^2 = \left(1-\frac{n}{N}\right)\frac{S^2}{n} = \frac{S^2}{n} - \frac{S^2}{N}$$

$$\Longrightarrow \frac{S^2}{n} = \epsilon^2 + \frac{S^2}{N} \Longrightarrow n = \frac{S^2}{\epsilon^2 + \frac{S^2}{N}} = \frac{NS^2}{N\epsilon^2 + S^2}$$

Se observa que cuando $N \to \infty$ (fracción de muestreo n/N tendiendo a cero) el tamaño muestral $n \to S^2/\epsilon^2 = n_0$ (n es inversamente proporcional al error de muestreo). En una situación práctica, la solución para n presenta un problema debido a que en la mayoría de las ocasiones, la cuasi-varianza poblacional S^2 es desconocida. Puesto que la cuasi-varianza muestral $\widehat{S}^2$ suele estar disponible de algún experimento anterior, es posible obtener un tamaño de muestra aproximado al reemplazar S^2 por $\widehat{S}^2$ en la expresión anterior.

- Estimador del total

$$\epsilon = \sigma(\widehat{X}) = \sqrt{N^2(1-f)\frac{S^2}{n}} \Longrightarrow \epsilon^2 = N^2\left(1-\frac{n}{N}\right)\frac{S^2}{n} = N^2\frac{S^2}{n} - N^2\frac{S^2}{N}$$

$$\Longrightarrow N^2\frac{S^2}{n} = \epsilon^2 + \frac{N^2S^2}{N} \Longrightarrow n = \frac{N^2S^2}{\epsilon^2 + \frac{N^2S^2}{N}} = \frac{N^3S^2}{N\epsilon^2 + N^2S^2} = \frac{N^2S^2}{e^2 + NS^2}.$$

Análogamente al caso anterior, en la mayoría de las situaciones prácticas, la cuasi-varianza poblacional S^2 es desconocida. Si se tiene la cuasi-varianza muestral $\widehat{S}^2$ de algún experimento anterior, se reemplaza S^2 por $\widehat{S}^2$ en la expresión anterior.

- Estimador de la proporción

$$\epsilon = \sigma(\widehat{\bar{P}}) = \sqrt{\frac{N}{N-1}(1-f)\frac{PQ}{n}} \Longrightarrow \epsilon^2 = \frac{N}{N-1}\left(1-\frac{n}{N}\right)\frac{PQ}{n}$$

$$\Longrightarrow \frac{NPQ}{(N-1)n} = \epsilon^2 + \frac{PQ}{N-1} \Longrightarrow n = \frac{NPQ/(N-1)}{\epsilon^2 + \frac{PQ}{N-1}} = \frac{NPQ}{(N-1)\epsilon^2 + PQ}.$$

Se observa que cuando $N \to \infty$ (fracción de muestreo n/N tendiendo a cero) el tamaño muestral $n \to PQ/\epsilon^2 = n_0$ (n es inversamente proporcional al cuadrado del error de muestreo y directamente proporcional a la proporción poblacional P). Para la estimación de la proporción es muy interesante tener en cuenta que para poblaciones grandes o fracciones de muestreo pequeñas ($N \to \infty$), el valor máximo de n se obtiene para $P = Q = 1/2$. Para constatar este resultado sabemos que si $N \to \infty$ el tamaño muestral n tiende al valor $n_0 = PQ/e^2 = f(P)$, expresión que tenemos que maximizar en P. En este caso, el valor máximo de n para poblaciones grandes o fracciones de muestreo pequeñas se obtiene para

$P = Q = 1/2$. Por lo tanto, para un error prefijado se necesitarán tamaños de muestra más pequeños cuanto más próximo esté P a cero o a uno. Este resultado es muy importante en la práctica, ya que **cuando se estiman proporciones y no se conoce el valor de la proporción poblacional P ni se tiene una aproximación suya (proporcionada por una encuesta similar, por una encuesta piloto, por la misma encuesta realizada anteriormente o por cualquier otro método) entonces se toma** $P = 1/2$, con lo que estamos situándonos en el caso de máximo tamaño muestral para el error fijado, lo cual siempre es aceptable estadísticamente. La dificultad práctica puede ser que se obtenga un tamaño muestral n demasiado grande para el presupuesto de que se dispone.

- Estimador del total de clase

$$\epsilon = \sigma(\widehat{\bar{A}}) = \sqrt{\frac{N^3}{N-1}(1-f)\frac{PQ}{n}} \Longrightarrow \epsilon^2 = \frac{N^3}{N-1}\left(1-\frac{n}{N}\right)\frac{PQ}{n}$$

$$\Rightarrow e^2 = \frac{N^3PQ}{(N-1)n} - \frac{N^2PQ}{N-1} \Rightarrow n = \frac{N^3PQ}{(N-1)e^2 + N^2PQ}.$$

Dado que, en general, la varianza de los estimadores depende de parámetros poblacionales desconocidos, usaremos una estimación de la misma para determinar el tamaño muestral para un ϵ determinado. Estas varianzas estimadas las podemos obtener de estudios anteriores o encuestas piloto.

Ejemplo 5 *En el ejemplo 4, ¿cuál debería de ser el tamaño de muestra óptimo para estimar dicho coste total reduciendo el error de muestreo del primer apartado en un* 10 %*?*
Si deseamos reducir el error de muestreo del primer apartado en un 10 % *el máximo error que estamos dispuestos a admitir, considerando el anterior como* $\sigma(\widehat{X}) = 132$*, es un error de muestreo de* $\sigma(\widehat{X}) = 118{,}8$*. Impongamos esta condición para determinar el tamaño de muestra óptimo necesario para estimar el costo total con dicho error de muestreo, teniendo en cuenta que el la cuasivarianza muestral es de 16.1233. Sustituyendo*

$$n = \frac{N^2S^2}{\epsilon^2 + NS^2} = \frac{120^2 * 16{,}1233}{118{,}8^2 + 120 * 16{,}1233} = 14{,}4674 \cong 15 \textit{ casas}$$

1.3.2. Tamaño de muestra fijado el error relativo de muestreo

Análogamente, fijado el error relativo ϵ_r, el tamaño de muestra óptimo necesario se despeja de la ecuación siguiente:

$$\epsilon_r = CV(\widehat{\theta}) = \frac{\sigma(\widehat{\theta})}{E(\widehat{\theta})}.$$

Vamos a ver cuál es el tamaño de la muestra a seleccionar para cometer dicho error según las características poblacionales a estimar.

- Estimador de la media

$$e_r = CV(\hat{\bar{X}}) = \frac{\sigma(\hat{\bar{X}})}{E(\hat{\bar{X}})} = \frac{\sqrt{(1-f)\frac{S^2}{n}}}{\bar{X}}.$$

Elevando ambos miembros al cuadrado y despejando el valor de n, se tiene que

$$n = \frac{S^2}{e_r^2(\bar{X})^2 + S^2/N}. \tag{1.2}$$

Para poblaciones grandes ($N \to \infty$), se tiene que

$$n \to n_0 = \frac{S^2}{\bar{X}^2 e_r^2}.$$

Es decir, a medida que el valor de e_r aumenta, el valor de la muestra disminuye. Como en el caso del error absoluto, la solución para n presenta un problema debido a que en la mayoría de las ocasiones, la cuasi-varianza poblacional S^2 es desconocida. Puesto que la cuasivarianza muestral $\widehat{S}^2$ suele estar disponible de algún experimento aleatorio anterior, es posible obtener un tamaño de muestra aproximado al reemplazar S^2 por $\widehat{S}^2$.

- Estimador del total

$$e_r = CV(\widehat{X}) = \frac{\sigma(\widehat{X})}{E(\widehat{X})} = \frac{\sqrt{N^2(1-f)\frac{S^2}{n}}}{X}.$$

Elevando ambos miembros al cuadrado y despejando el valor de n, se tiene que

$$n = \frac{N^2 S^2}{e_r^2(\bar{X})^2 N^2 + S^2 N} = \frac{S^2}{\bar{X}^2 e_r^2 + S^2/N}. \tag{1.3}$$

Observamos que el tamaño de muestra necesario para cometer un error relativo de muestreo dado coincide para el estimador de la media (1.2) y del total (1.5).

- Estimador de la proporción

$$e_r = CV(\widehat{P}) = \frac{\sigma(\widehat{P})}{E(\widehat{P})} = \frac{\sqrt{\frac{N}{N-1}\frac{1}{n}(1-f)PQ}}{P}.$$

Elevando ambos miembros al cuadrado y despejando el valor de n, se tiene que

$$n = \frac{NQ}{(N-1)Pe_r^2 + Q} \tag{1.4}$$

Como antes, si $N \to \infty$, entonces el valor de la muestra tiende a $Q/(Pe_r^2)$. En la práctica, cuando se estiman proporciones y no se conoce el valor de

la proporción poblacional P ni se tiene una aproximación suya (proporcionada por una encuesta similar, por una encuesta piloto, por la misma encuesta realizada anteriormente, ni por ningún otro método) entonces se llama $P = 1/2$. Este caso $P = 1/2$ proporciona el caso de máximo tamaño muestral para el error fijado, lo cual es siempre aceptable estadísticamente. La dificultad práctica puede ser que se obtenga un tamaño muestral n demasiado grande para el presupuesto de que se dispone.

- Estimador del total de clase

$$e_r = CV(\widehat{A}) = \frac{\sigma(\widehat{A})}{E(\widehat{A})} = \frac{\sqrt{\frac{N^3}{N-1}\frac{1}{n}(1-f)PQ}}{A}.$$

Elevando ambos miembros al cuadrado y despejando el valor de n, se tiene que

$$n = \frac{NQ}{(N-1)Pe_r^2 + Q} \tag{1.5}$$

Observamos que el tamaño de muestra necesario para cometer un error relativo de muestreo dado coincide para el estimador de la proporción (1.4) y del total (1.5).

Ejemplo 6 *Volvamos al ejemplo 4. Una gran empresa constructora tiene 120 casas en diversas etapas de construcción. Para estimar la cantidad total (en miles de euros) que será registrada en el inventario de la construcción en proceso, se seleccionó una muestra aleatoria simple de 12 casas y se determinaron los costes acumulados en cada una de ellas. Los costos obtenidos para las 12 casas fueron los siguientes:*

$$35{,}5, 30{,}2, 28{,}9, 36{,}4, 29{,}8, 34{,}1, 32{,}6, 26{,}4, 38, 38{,}2, 32{,}2, 27{,}5.$$

- *Estimar los costes totales acumulados para las 120 casas y dar una estimación del error relativo de muestreo. ¿Cuál debería ser el tamaño muestral óptimo para reducir dicho error relativo en un* $10\,\%$*?*
- *Estimar la proporción de casas cuyos costes de construcción superan los 32 mil euros. Dar una estimación del error relativo de muestreo. ¿Cuál debería ser el tamaño muestral óptimo para reducir dicho error relativo en un* $10\,\%$*?*

Ejemplo 7 *Mediante muestreo aleatorio simple se trata de estimar la proporción de piezas correctas producidas en un proceso industrial en el que se fabrican un total de 8000 unidades. Una muestra piloto ha suministrado 1/5 de piezas defectuosas. Obtener el tamaño de muestra necesario para que el error de muestreo sea de 0.08 al estimar la proporción de piezas correctas producidas en el proceso de producción industrial. Hallar el tamaño de muestra necesario para*

que el error relativo de muestreo sea del 1,2 % *en la misma estimación. En el caso de que el error de muestreo sea de 0.08, se tiene que*

$$n = \frac{NPQ}{(N-1)\epsilon^2 + PQ} = \frac{8000{\cdot}4/5{\cdot}1/5}{7999{\cdot}0{,}08^2 + 1/5{\cdot}4/5} = 24{,}98 \cong 25\, piezas$$

Por otra parte, en el caso de que el error relativo de muestreo sea del 2 % *se tiene que*

$$n = \frac{NQ}{(N-1)Pe_r^2 + Q} = \frac{8000{\cdot}1/5}{7999{\cdot}0{,}2^2{\cdot}4/5 + 1/5} = 579{,}7774 \cong 580\, piezas$$

1.3.3. Tamaño de muestra para un error de muestreo y un coeficiente de confianza dados

En determinadas ocasiones, aparte de calcular el tamaño muestral para un error de muestreo dado, prefijamos un nivel de confianza adicional para el cálculo de dicho tamaño, con la finalidad de relajar en cierta forma el cálculo de n. De esta forma se halla n con un grado de tolerancia definido por el nivel de confianza.

Supongamos que estimamos el parámetro θ mediante el estimador insesgado $\widehat{\theta}$ cometiendo el error absoluto máximo admisible e_α para un coeficiente de confianza $1-\alpha$. Suponemos que el estimador $\widehat{\theta}$ sigue una distribución normal de media $E(\widehat{\theta}) = \theta$ y varianza $\sigma^2(\widehat{\theta})$. En este caso, se tiene que

$$P\left(|\widehat{\theta} - \theta| \leq e_\alpha\right) = 1-\alpha \Longrightarrow P(-e_\alpha \leq \widehat{\theta} - \theta \leq e_\alpha) = 1-\alpha$$

Por lo tanto,

$$P\left(-\frac{e_\alpha}{\sigma(\widehat{\theta})} \leq \frac{\widehat{\theta} - \theta}{\sigma(\widehat{\theta})} \leq \frac{e_\alpha}{\sigma(\widehat{\theta})}\right) = 1-\alpha,$$

de manera que

$$z_{1-\alpha/2} = \frac{e_\alpha}{\sigma(\widehat{\theta})} \Longrightarrow e_\alpha = z_{1-\alpha/2}\sigma(\widehat{\theta}).$$

De esta forma vemos que la identidad fundamental para obtener n según un error de muestreo dado cuando existe un coeficiente de confianza adicional dado es la siguiente

$$e_\alpha = z_{1-\alpha/2}\sigma(\widehat{\theta}).$$

Ejemplo 8 *Obtener los valores de n para un error de muestro* e_α *con coeficiente de confianza adicional* $1-\alpha$ *para los diferentes estimadores.*

- *Estimador de la media*

$$n = \frac{z_{1-\alpha/2}^2 NS^2}{Ne_\alpha^2 + z_{1-\alpha/2}^2 S^2}.$$

- *Estimador del total*

$$n = \frac{z^2_{1-\alpha/2} N^2 S^2}{e^2_\alpha + z^2_{1-\alpha/2} S^2 N}.$$

- *Estimador de la proporción*

$$n = \frac{z^2_{1-\alpha/2} NPQ}{(N-1)e^2_\alpha + z^2_{1-\alpha/2} PQ}.$$

- *Estimador del total de clase*

$$n = \frac{z^2_{1-\alpha/2} N^3 PQ}{(N-1)e^2_\alpha + z^2_{1-\alpha/2} N^2 PQ}.$$

1.3.4. Tamaño de muestra para un error relativo de muestreo y un coeficiente de confianza dados

En determinadas ocasiones, aparte de calcular el tamaño muestral para un error relativo de muestreo dado, prefijamos un nivel de confianza adicional para el cálculo de dicho tamaño, con la finalidad de relajar en cierta forma el cálculo de n. De esta forma se halla n con un grado de tolerancia definido por el nivel de confianza.

Supongamos que estimamos el parámetro θ mediante el estimador insesgado $\widehat{\theta}$ cometiendo un error relativo $e_{r,\alpha}$. Análogamente al caso anterior, consideramos que el estimador $\widehat{\theta}$ sigue una distribución normal de media $E(\widehat{\theta}) = \theta$ y varianza $\sigma^2(\widehat{\theta})$. Se tiene que

$$P\left(\left|\frac{\widehat{\theta}-\theta}{\theta}\right| \leq e_{r,\alpha}\right) = 1-\alpha \Longrightarrow P(-\theta e_{r,\alpha} \leq \widehat{\theta}-\theta \leq \theta e_{r,\alpha}) = 1-\alpha$$

Por lo tanto,

$$P\left(-\frac{\theta e_{r,\alpha}}{\sigma(\widehat{\theta})} \leq \frac{\widehat{\theta}-\theta}{\sigma(\widehat{\theta})} \leq \frac{\theta e_{r,\alpha}}{\sigma(\widehat{\theta})}\right) = 1-\alpha,$$

de donde se deduce que

$$\frac{\theta e_{r,\alpha}}{\sigma(\widehat{\theta})} = z_{1-\alpha/2}, \Longrightarrow e_{r,\alpha} = z_{1-\alpha/2} Cv(\widehat{\theta}).$$

De esta forma vemos que la identidad fundamental para obtener n según un error relativo de muestreo dado dado cuando existe un coeficiente de confianza adicional viene dado por

$$e_{r,\alpha} = z_{1-\alpha/2} Cv(\widehat{\theta}).$$

Ejemplo 9 *Obtener los valores de n para un error de muestro $e_{r,\alpha}$ con coeficiente de confianza adicional $1-\alpha$ para los diferentes estimadores.*

- *Estimador de la media*

$$n = \frac{z^2_{1-\alpha/2} C^2_{1,x}}{e^2_{r,\alpha} + z^2_{1-\alpha/2} C^2_{1,x}/N}, \quad C_{1,x} = \frac{S}{\bar{X}}$$

- *Estimador del total*

$$n = \frac{N z^2_{1-\alpha/2} C^2_{1,x}}{N e^2_{r,\alpha} + z^2_{1-\alpha/2} C^2_{1,x}}, \qquad C_{1,x} = \frac{S}{\bar{X}}$$

- *Estimador de la proporción*

$$n = \frac{NQ z^2_{1-\alpha/2}}{P(N-1) e^2_{r,\alpha} + z^2_{1-\alpha/2} Q}.$$

- *Estimador del total de clase*

$$n = \frac{NQ z^2_{1-\alpha/2}}{P(N-1) e^2_{r,\alpha} + z^2_{1-\alpha/2} Q}.$$