Tema 58

Introducción a la Teoría de Muestras

58.1 Conceptos básicos

Los primeros conceptos fundamentales a considerar en Teoría de Muestras son los de Población y Muestra.

Se denomina población o universo (según el diccionario de términos estadísticos de Kendall y Buckland, 1971) a cualquier colección, finita o infinita de individuos o elementos (unidades de las que en general deseamos obtener cierta información).

A veces, se ha dado distinto significado a los términos universo y población, indicando por universo un conjunto de elementos y por población el conjunto de números obtenidos midiendo cierta característica de los mismos. Así, de un mismo universo, pueden obtenerse varias poblaciones que, a su vez, pueden ser de distinta dimensión.

Se da el nombre de censo a la enumeración y anotación de ciertas características de todos los elementos de la población.

En las poblaciones es posible medir o contar en cada unidad una o varias características, o clasificar sus unidades de acuerdo a ellas. A partir de los resultados podríamos llegar al conocimiento de valores como la media, el total de la clase, la proporción, etc. a los que denominaremos parámetros o características poblacionales.

La población que se intenta investigar se denomina población objetivo, y puede considerarse como un modelo ideal cuya contrapartida en el mundo real estará formada por otro conjunto en el que existirán omisiones, duplicaciones o unidades extrañas.

Por otra parte, quizás no pueda obtenerse información de algunas unidades por diferentes motivos como inaccesibilidad, ausencias o negativas a colaborar, o bien por las propias limitaciones de los instrumentos de medida. Todo ello hace que el conjunto que realmente es objeto de la investigación difiera de la

población objetivo, y será denominado población investigada.

En muchos casos puede no ser posible o conveniente obtener información de todas las unidades de la población, por lo que el estudio se reducirá a una parte de la misma.

Se denomina muestra a una parte de la población o a un subconjunto de un conjunto de unidades, que se obtiene con el fin de investigar las propiedades de la población o conjunto de procedencia. Se desea, por tanto, que la muestra sea lo más representativa posible de los elementos de la población en el sentido de que proporcione buena información sobre ésta.

Para elegir las unidades que constituyen la muestra, es necesario disponer de un conjunto real de unidades que se ajuste lo mejor posible al conjunto ideal que constituye la población objetivo. Es lo que denominamos población marco.

La formación del marco (de la lista de unidades a muestrear) es una de las operaciones más delicadas.

Cuando se extrae una muestra los datos obtenidos a partir de ella nos permiten inferir unos valores aproximados de la población. A estos valores aproximados se les denomina estimaciones, las cuales vendrán afectadas por un error que denominaremos error debido al muestreo. Cuanto menor sea éste, más precisas serán las estimaciones.

58.2 Tipos de muestreo

Se denomina muestreo al procedimiento mediante el cual se obtienen una o más muestras de una población.

La selección de la muestra es el objetivo de todo tipo de muestreo, los cuales según la "*Encyclopedy of Statistical*" Science, pueden descomponerse en dos grandes bloques:

58.2.1 Muestreo probabilístico

Aquél en que se conoce, o puede calcularse de antemano la probabilidad de seleccionar cada una de las posibles muestras en la población. Es decir, la selección de una muestra es considerada como un experimento aleatorio.

Cochran, Mosteller y Tukey (1954) llaman a un muestreo "semiprobabilístico superior" cuando se conoce la probabilidad de extracción de una cierta parte de la población pero no de un elemento dentro de ella. Por ejemplo, la selección aleatoria de manzanas en una ciudad, y no se conoce la probabilidad de tomar como unidad muestral cada una de las viviendas en las manzanas.

Por el contrario, el muestreo "semiprobabilístico inferior" es aquél en que se conoce la probabilidad de selección de los elementos dentro de una parte de la población, pero no la probabilidad de selección de dicha parte. Por ejemplo, se seleccionan municipios de un país de forma arbitraria, y dentro de éstos, se hace una selección aleatoria de viviendas.

Es claro que para que el muestreo sea probabilístico ha de ser probabilístico superior e inferior.

Esto se puede expresar simbólicamente de la siguiente forma:

Sea $m = \{u_{i1}, ..., u_{in}\}$ una muestra cualquiera de tamaño n extraída de la población $U = \{u_1, ..., u_N\}$, y sea M el conjunto de todas las muestras posibles de $\dot{U}$. El método de muestreo es probabilístico cuando se tiene definida una función de probabilidad $P(m)$ verificando:

$$0 \leq P(m) \leq 1 \qquad \text{y} \qquad \sum_{m \in M} P(m) = 1$$

Pasamos ahora a hacer una descripción berve de algunos métodos de muestreo probabilístico:

1. Muestreo aleatorio sin reemplazamiento (aleatorio simple)

Se toman las unidades que han de formar la muestra, una a una y sin reposición. Si la población es finita, la probabilidad de extracción de un elemento dependerá de los que hayan sido ya extraidos. Todas las muestras son equiprobables. Se denomina también "muestreo irrestrictamente aleatorio".

2. Muestreo aleatorio con reemplazamiento

Se toman las unidades una a una, con reemplazamiento, y de forma que en cada extracción, todas tienen la misma probabilidad de ser elegidas. Todas las muestras son equiprobables.

Formalmente coincide con el muestreo en poblaciones infinitas, ya que al devolver a la población cada elemento, ésta es inagotable y el resultado de la extracción de un elemento es independiente de los anteriores a él.

3. Muestreo estratificado

Se divide la población en varias subpoblaciones (generalmente de elementos con características comunes). La selección de la muestra se hace tomando elementos de cada una de las poblaciones, denominadas "estratos". Cuando la selección de las unidades para formar la muestra es aleatoria en cada estrato se denomina Muestreo Estratificado Aleatorio.

4. Muestreo por conglomerados o áreas

Consiste en sustituir las unidades elementales de la población por unidades de elementos agrupados y elegir algunas de éstas últimas para formar la muestra.

En este tipo de muestreo, como la selección es de grupos de unidades o "conglomerados", interesa que cada uno represente lo mejor posible a la poblacion; o sea, constituido por unidades heterogéneas.

5. Muestreo bietápico

Es una modificación del anterior: se realizan los conglomerados y se seleccionan algunos de ellos. A continuación, se efectua una selección de unidades dentro de cada uno (no forman parte en la muestra todos los elementos de cada conglomerado).

6. Muestreo polietápico

Es una generalización del anterior a más de dos etapas.

7. Muestreo sistemático

Ordenados los elementos, se elige aleatoriamente uno entre los que ocupan los lugares 1 al k, ambos incluidos. A continuación se toman los siguientes de k en k.

8. Muestreo doble o bifásico

Se toma una muestra, generalmente grande, de forma rápida, sencilla y poco costosa. Para ello se considera alguna variable auxiliar relacionada con la que es objeto de nuestro estudio. A partir de esta muestra se tomará ya la definitiva.

Por ejemplo, si se quiere hacer un estudio sobre presupuestos familiares, pueden recogerse primero datos de domicilios y alquileres de un gran número de familias y, dentro de ellas, elegir una submuestra para los presupuestos.

9. Muestreo múltiple o polifásico

Generalización del anterior a más fases.

10. Muestreo repetido

Se realiza generalmente en poblaciones dinámicas, obteniendo muestras a intervalos regulares de tiempo, para estudiar la evolución de la población.

11. Muestreo sucesional (secuencial)

Suele emplearse en modelos de inspección para aceptación o rechazo. No se fija de antemano el número de unidades que formarán parte de la muestra, sino que se va examinando cada unidad y a continuación se decide si seguir el muestreo o conformarse con la información ya obtenida.

58.2.2 Muestreos no probabilísticos

Este tipo de muestreos se caracterizan por desconocer las probabilidades de selección de las muestras o cualquiera de las condiciones que se les suele exigir a las muestras probabilísticas.

Algunos métodos de este tipo son:

1. Muestreo intencional u opinático

En este tipo de muestreo, es la persona que selecciona la muestra la que procura que ésta sea representativa. La representatividad depende, por tanto, de su intención u opinión, y la evaluación de la representatividad es totalmente subjetiva. Este muestreo carece, por tanto, de una base teórica consistente (según la escuela frecuentista); sin embargo, su uso es bastante generalizado, por ejemplo en el Muestreo por Cuotas, de gran importancia en los estudios de opinión y análisis de mercados (ver Moser y Stuart, 1952-53) [Supongamos que el diseño de la encuesta ha seguido los principios del muestreo probabilístico hasta llegar al momento de seleccionar las personas que han de ser entrevistadas. Esta etapa final consiste, en el muestreo por cuotas, en asignar a cada entrevistador un número de entrevistas a personas en un determinado grupo de edad, sexo, nivel económico, lugar y posiblemente otras características sociológicas o económicas. Sujeto a estas restricciones se deja en libertad al entrevistador para que éste elija a las personas que cumplan dichos requisitos. El margen de libertad dado al entrevistador puede introducir sesgos en el proceso de selección, que en general no podrán ser detectados. El desconocimiento de las probabilidades de selección hace que no puedan estimarse los errores debidos al muestreo.

2. Muestreo sin norma

En este tipo de muestreo se toma la muestra de cualquier manera, por razones de comodidad, o a capricho. Si la población es homogénea, la muestra obtenida puede ser representativa.

A veces, la uniformidad puede sustituirse por una buena mezcla de las unidades elementales antes de tomar las muestras (por ejemplo, barajar naipes o girar bolas en un bombo).

3. Muestreo por cuotas

En una primera etapa se descompone la población en grupos de elementos (excluentes y exhaustivos) con características comunes definidas previamente (cada grupo definido reprsentará un determinado porcentaje de la población; a dicho porcentaje es a lo que se suele denominar cuota). Las características que se fijen dependerán fundamentalmente de la encuesta que se desee realizar. La segunda etapa consiste en elegir una muestra que refleje exactamente la descomposición en cuotas que se ha fijado en la población. La elección de los elementos que compondrán la muestra queda al "arbitrio" del entrevistador con la restricción al menos mínima de las cuotas.

58.2.3 Muestreos mixtos

En la práctica, de acuerdo con las características del campo en que se esté efectuando el muestreo, es frecuente el uso de métodos mixtos (probabilísticos y no probabilísticos) y diseños complejos. Así, es muy frecuente empezar clasificando

la población en estratos, dividir después cada estrato en áres o conglomerados, para posteriormente establecer cierto número de etapas.

58.3 Conveniencias y limitaciones

El objetivo del muestreo es la inferencia de las conclusiones obtenidas en la muestra a la población. Puesto que la inferencia siempre supone un riesgo, es útil indicar en qué casos conviene obtener muestras en lugar de hacer un estudio exhaustivo de la población.

En todo caso, la decisión óptima consiste en emplear los recursos mínimos para obtener una determinada información, o bien obtener la máxima información utilizando unos recursos prefijados.

Puede decirse que se trata de "minimizar la pérdida total" en la que se incluyen los recursos empleados (económicos, tiempo, esfuerzo, etc.) y el "error" y la probabilidad de cometerlo, entendiendo por error alguna medida de la desviación entre los valores desconocidos y los estimados

58.3.1 Conveniencias

Hay que tomar muestras en las siguientes situaciones, entre otras:

a) Cuando la población sea tan grande que el censo exceda de las posibilidades del investigador.

b) Cuando la población sea suficientemente uniforme para que cualquier muestra dé una buena representación.

c) Cuando el proceso de medida o investigacion sea destructivo (de los caractéres de cada elemento) como ocurre al consumir un articulo para juzgar su calidad o al determinar una dosis letal o un punto de ruptura.

En las situaciones a) y c) es claro que, por grande que sea el riesgo, será inferior a la certeza de un coste prohibitivo. En la situación b) se admite que el riesgo es despreciable.

Hay situaciones menos extremas que se aproximan a las anteriores:

- Sin ser estrictamente imposible el estudio de toda la población, puede serlo prácticamente por su gran número de elementos.
- La población es aceptablemente homogénea.
- Los elementos investigados no llegan a inutilizarse, pero sí llega a disminuir su valor.

Otras razones que pueden hacer ventajoso el muestreo son:

1. Economía: Es evidente que si en vez de examinar la población entera examinamos sólo una parte, el coste será inferior. Además pueden aplicarse los recursos no utilizados para la obtención de otras informaciones de interés.

En un sentido más amplio de "economía" podíamos incluir aquí la posibilidad de una mayor rapidez en la obtención de los resultados: al disponer de ciertos

recursos es ventajoso obtener, mediante muestras, información más rápida, lo que aumentará su utilidad, sobre todo en fenómenos dinámicos o evolutivos. Además no sólo hay que considerar el coste absoluto, sino también el relativo, esto es, el coste en relación a la cantidad de información obtenida.

2. Calidad: Es posible cuidar más la precisión de cada observación al tener que hacer menor número de ellas

58.3.2 Limitaciones

Es obvio que no será posible utilizar muestras cuando se necesite información de cada uno de los elementos poblacionales. Otra limitación es la dificultad que supone el empleo de un instrumento delicado y complejo como es la Teoría de Muestras.

El muestreo exige menos trabajo material, pero más refinamiento y preparación. Requiere, no sólamente una base adecuada en los diseñadores, sino también una cierta preparación de los entrevistadores, inspectores y supervisores. Todo esto hace que el coste por unidad sea mayor en las muestras que en los censos.

58.4 Consideraciones para determinar el tamaño de la muestra

Los principales pasos involucrados en la elección del tamaño muestral son los siguientes:

(1) Debe existir algún enunciado respecto a lo que se espera de la muestra. Este puede darse en términos de límites de error deseado o bien en términos de alguna decisión o acción que debe tomarse una vez que se conocen los resultados de la muestra.

(2) Se debe encontrar una ecuación que relacione n con la precisión deseada de la muestra. La ecuación variará según el contenido del enunciado de precisión y el tipo de muestreo propuesto. Una de las ventajas del muestreo probabilístico es que permite la elaboración de esta ecuación.

(3) Esta ecuación tendrá como parámetros ciertas propiedades desconocidas de la población, que deben estimarse para obtener resultados específicos.

(4) Con frecuencia sucede que los datos se estipulan para ciertas subdivisiones mayores de la población y que los límites de error deseados se establecen para cada subdivisión. De ser así, se hace un cálculo separado para el valor n en cada subdivisión y el n total se encuentra por adición.

(5) Generalmente se mide más de un atributo o característica. En ocasiones, el número es grande. Si se estipula un grado de precisión para cada característica, los cálculos conducirán a una serie de valores n uno por cada atributo. Por lo tanto, debe encontrarse un método para reconciliar estos valores.

(6) Debe apreciarse el valor elegido de "n" para que sea consistente con los recursos de muestreo disponibles. Esto exige una estimación del costo, trabajo, tiempo y materiales que se necesitan para obtener una muestra del tamaño propuesto. En ocasiones es claro que n debe reducirse drásticamente, y entonces es necesario tomar una decisión difícil, que es de proceder con una muestra mucho más pequeña, lo que reduce la precisión, o bien abandonar los esfuerzos hasta contar con meyores recursos.

58.5 Estimación y tamaño muestral

Las técnicas de muestreo deberían conseguir muestras o subconjuntos de la población a estudiar que sean "miniaturas de la población". Es decir, cuando se vaya a realizar una encuesta por muestreo, la muestra ha de representar de una forma lo más aproximada posible a la población que se desea estudiar. En este sentido, aparecen un conjunto de técnicas que tienen como objetivo delimitar como se estructura la población con el fin de que la muestra tenga una representatividad adecuada.

Consideramos una población finita, $\mathcal{U}$, formada por N elementos

$$\mathcal{U} = \{u_1, ..., u_N\}$$

y representamos por Y a la variable objeto de estudio sobre la población, denotando por Y_i al valor de la variable sobre el elemento u_i de la población. Así pues, la variable objeto de estudio, Y, tomará sobre los elementos de la población unos valores determinados $\{Y_1, ..., Y_i, ..., Y_N\}$.

El objetivo de un estudio mediante muestreo es el de estimar alguna característica de la población, expresada en términos de $\{Y_1, ..., Y_i, ..., Y_N\}$, a partir de la observación de la variable objeto de estudio en sólo una parte de la población, denominada muestra, $m = \{u_{i1}, ..., u_{in}\}$. Al número de individuos que componen la muestra "n" se le denomina tamaño muestral y denominamos fracción de muestreo a:

$$f = \frac{n}{N}$$

Los distintos métodos de muestreo se caracterizan por el procedimiento utilizado para la selección de la muestra m.

La característica poblacional objeto de interés recibe el nombre genérico de parámetro poblacional y la representeremos por $\theta = \theta(Y_1, ..., Y_N)$. Los parámet-

ros poblacionales más usuales son:

$$\text{Total poblacional:} \qquad T(Y) = \sum_{i=1}^{N} Y_i$$

$$\text{Media poblacional:} \qquad \overline{Y} = \tfrac{1}{N} \sum_{i=1}^{N} Y_i$$

Otros parámetros, que desempeñan un papel importante en el muestreo son los siguientes:

$$\text{Varianza poblacional:} \qquad \sigma_Y^2 = \frac{1}{N} \sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2$$

$$\text{Cuasi-varianza poblacional:} \qquad \sigma_{cY}^2 = \frac{1}{N-1} \sum_{i=1}^{N} \left(Y_i - \overline{Y}\right)^2$$

$$\text{Coeficiente de variación:} \qquad CV(Y) = \frac{\sigma_Y}{\overline{Y}}$$

En el caso en que la variable objeto de estudio sea una variable tipo $0-1$ ($Y_i = 0$ ausencia, $Y_i = 1$ presencia (de la cualidad objeto de estudio en los elementos de la población) en cada elemento u_i), $T(Y)$ representa el número de elementos de la población que presentan la cualidad objeto de estudio, y la media poblacional, $\overline{Y}$, la proporción de individuos en la población que presentan la cualidad, y la notaremos por P. En este caso, se verifica que $\sigma_Y^2 = P(1-P)$.

A partir de la muestra seleccionada, $m = \{u_{i1}, ..., u_{in}\}$, y de los valores de la variable objeto de estudio sobre las unidades muestrales $\{Y_{i1}, ..., Y_{in}\}$ podemos definir la versión muestral de los parámetros anteriormente considerados:

$$\text{Total muestral:} \qquad t(Y) = \sum_{j=1}^{n} Y_{ij} = \sum_{u_i \in m} Y_i$$

$$\text{Media muestral:} \qquad \overline{y(m)} = \frac{1}{n} \sum_{u_i \in m} Y_i$$

$$\text{Varianza muestral:} \qquad S_Y^2 = \frac{1}{n} \sum_{u_i \in m} \left(Y_i - \overline{y(m)}\right)^2$$

$$\text{Cuasi-varianza muestral:} \qquad S_{cY}^2 = \frac{1}{n-1} \sum_{u_i \in m} \left(Y_i - \overline{y(m)}\right)^2$$

58.5.1 Estimación

Comentaremos brevemente la estimación puntual y por intervalos de confianza.

1. Estimación puntual

Una vez seleccionada una muestra $m = \{u_{i1}, ..., u_{in}\}$, de acuerdo al diseño muestral establecido (es decir, de acuerdo al espacio muestral y al muestreo elegido) y observada la variable objeto de estudio sobre los elementos muestrales $\{Y_{i1}, ..., Y_{in}\}$, el problema que se plantea es el de utilizar la información recogida con objeto de estimar o inferir el valor del parámetro poblacional, desconocido, objeto de interés $\theta = \theta(Y_1, ..., Y_N)$.

La estimación del parámetro se realizará mediante una función de los valores muestrales observados, que denotaremos por $\widehat{\theta}(Y_1, ..., Y_N)$ o simplemente por $\widehat{\theta}$ y que denominaremos estimador de θ.

Esta forma de inferencia se denomina inferencia puntual, ya que proporciona un único valor como estimación del parámetro objeto de estudio.

El valor del estimador en la muestra seleccionada m, $\widehat{\theta}(Y_{i1}, ..., Y_{in})$, que también denotaremos por $\widehat{\theta}(m)$, se denomina estimación del parámetro, proporcionándonos, en general, cada muestra una estimación distinta de θ. Por tanto, sería deseable que "en media" el valor de las estimaciones coincidiése con el verdadero valor del parámetro, θ. Esta idea nos lleva a definir el concepto de estimador insesgado.

Definición 1 *Sea $\mathcal{M}$ un espacio muestral y $P(\cdot)$ la probabilidad asociada a un tipo de muestreo. Sea $\widehat{\theta}$ un estimador del parámetro desconocido θ. Se*

Se define la media o esperanza del estimador, y la representaremos por $E\left[\widehat{\theta}\right]$, como:

$$E\left[\widehat{\theta}\right] = \sum_{m \in \mathcal{M}} \widehat{\theta}(m) P(m)$$

Se dirá que un estimador $\widehat{\theta}$ es insesgado del parámetro poblacional θ sii se verifica:

$$E\left[\widehat{\theta}\right] = \theta(Y_1, ..., Y_N) \qquad \forall (Y_1, ..., Y_N)$$

Una característica de interés a la hora de evaluar la "calidad" o precisión de un estimador insesgado $\widehat{\theta}$ es su varianza, que nos informa sobre la dispersión de los distintos valores que puede tomar el estimador en torno a su media.

Definición 2 *Se define la varianza de $\widehat{\theta}$, representada por $\sigma^2\left(\widehat{\theta}\right)$ como sigue:*

$$\sigma^2\left(\widehat{\theta}\right) = \sum_{m \in \mathcal{M}} \left(\widehat{\theta}(m) - E\left[\widehat{\theta}\right]\right)^2 P(m)$$

Si el estimador $\widehat{\theta}$ es insesgado, su varianza nos permite realizar una evaluación de la estimación puntual en el sentido de que si la varianza es relativamente pequeña "la dispersión de la estimación puntual $\widehat{\theta}(m)$ respecto del parámetro desconocido $\left|\widehat{\theta}(m) - \theta\right|$, aunque sea desconocida, se espera que sea pequeña".

2. Estimación por intervalos

Otra forma de realizar inferencia es proporcionando un conjunto de valores, $[S_1(\cdot), S_2(\cdot)]$, admisibles para el parámetro, siendo $S_1(\cdot)$ y $S_2(\cdot)$ dos funciones definidas sobre los valores muestrales observados, de forma que θ se encuentre en ese intervalo con una determinada probabilidad, $1-\alpha$, fijada de antemano.

Definición 3 *Dado un espacio muestral $\mathcal{M}$ y un tipo de muestreo (al cual le asociamos la probabilidad $P(\cdot)$) se dirá que $[S_1(\cdot), S_2(\cdot)]$ es un intervalo de confianza para el parámetro θ al nivel $1-\alpha$ si se verifica la siguiente propiedad:*

$$P[S_1(\cdot) \leq \theta \leq S_2(\cdot)] \geq 1-\alpha \qquad \forall (Y_1, ..., Y_N) \qquad (1)$$

Es importante realizar una correcta interpretación de los resultados empíricos obtenidos al construir un intervalo de confianza. Cuando a partir del diseño muestral utilizado seleccionamos una muestra concreta m, no se verifica que

$$P[S_1(m) \leq \theta \leq S_2(m)] \geq 1-\alpha$$

La afirmación $S_1(m) \leq \theta \leq S_2(m)$ es cierta o falsa, aunque no sabemos en que situación nos encontramos ya que desconocemos el verdadero valor del parámetro θ.

Una interpretación frecuentista de (1) nos indica que "si aplicando el diseño muestral $(\mathcal{M}, P(\cdot))$ obtenemos 100 muestras $m_1, ..., m_{100}$ y para cada una de ellas costruimos el correspondiente intervalo

$$[S_1(m_1), S_2(m_2)] \qquad i = 1, ..., 100$$

cabe esperar que aproximadamente al $100(1-\alpha)\,\%$ de los intervalos pertenezca el parámetro θ desconocido".

Como caso particular, si el parámetro de interés es θ y $\widehat{\theta}$ es un estimador insesgado, bajo ciertas hipótesis, bastante generales, se puede afirmar que la distribución de

$$\frac{\widehat{\theta}-\theta}{\sqrt{\sigma^2\left(\widehat{\theta}\right)}}$$

se aproxima a una $N(0,1)$.

Bajo estas condiciones se verifica, de forma aproximada que

$$P\left[\left|\frac{\widehat{\theta}-\theta}{\sqrt{\sigma^2\left(\widehat{\theta}\right)}}\right| \leq z_{1-\frac{\alpha}{2}}\right] = 1-\alpha$$

siendo $z_{1-\frac{\alpha}{2}}$ el percentil de orden $1-\frac{\alpha}{2}$ de la distribución $N(0,1)$. De la expresión anterior se deduce que

$$P\left[\left|\widehat{\theta}-\theta\right| \leq z_{1-\frac{\alpha}{2}}\sqrt{\sigma^2\left(\widehat{\theta}\right)}\right] = 1-\alpha \qquad (2)$$

es decir, que el intervalo de confianza para θ a nivel de significación $1-\alpha$ es:

$$\left]\widehat{\theta}-z_{1-\frac{\alpha}{2}}\sqrt{\sigma^2\left(\widehat{\theta}\right)}, \widehat{\theta}+z_{1-\frac{\alpha}{2}}\sqrt{\sigma^2\left(\widehat{\theta}\right)}\right[$$

58.5.2 Determinación del tamaño muestral

Hasta el momento se ha considerado que el tamaño muestral n era un valor fijado a priori. Sin embargo, el tamaño muestral puede determinarse fijando un error máximo admisible δ y un nivel de confianza $1-\alpha$, para que la precisión de la estimación alcance un determinado nivel. La forma de evaluar la precisión depende del tipo de parámetro objeto de estudio.

Si el parámetro es dimensional, por ejemplo el total o la media poblacional, la precisión se evalúa exigiendo que

$$P\left[\left|\frac{\widehat{\theta}}{\theta}-1\right| \leq \delta\right] \geq 1-\alpha$$

o equivalentemente

$$P\left[\left|\widehat{\theta}-\theta\right| \leq \delta\,|\theta|\right] \geq 1-\alpha$$

Si el parámetro es adimensional, por ejemplo la proporción, la precisión se evalúa mediante la expresión:

$$P\left[\left|\widehat{P}-P\right| \leq \delta\right] \geq 1-\alpha \qquad (3)$$

Comparando (3) con la expresión (2) recogida para el intervalo de confianza, se obtiene que

$$\delta\,|\theta| = z_{1-\frac{\alpha}{2}}\,\sigma\left(\widehat{\theta}\right) \qquad (4)$$

y puesto que la varianza del estimador, $\sigma^2\left(\widehat{\theta}\right)$, depende del tamaño muestral n, la expresión anterior relaciona n con el error máximo admisible δ y con el nivel de confianza $1-\alpha$.

Es también de destacar que (4) depende de los parámetros desconocidos θ y $\sigma\left(\widehat{\theta}\right)$, por lo que desde un punto de vista práctico, con el objeto de determinar el tamaño muestral n, una vez fijados δ y $1-\alpha$, se ha de utilizar alguna estrategia para evitar la dependencia de n de parámetros desconocidos.

58.5.3 Determinación del tamaño muestral en el muestreo aleatorio simple

Tal como lo hemos descrito el muestreo aleatorio simple $MAS(N,n)$ consta de un espacio muestral $\mathcal{M}$, formado por todos los posibles subconjuntos de n elementos de la población $U = \{u_1, ..., u_N\}$, en el que todas las muestras tienen la misma probabilidad de ser seleccionadas.

El número de muestras posibles de tamaño n es $\binom{N}{n}$, y el espacio muestral lo podemos representar por

$$\mathcal{M} = \left\{ m_1, ..., m_{\binom{N}{n}} \right\}$$

siendo $P(m) = \frac{1}{\binom{N}{n}} \qquad \forall m \in \mathcal{M}$.

1. Estimación puntual en el MAS(N,n)

Los estimadores a utilizar dependen del parámetro objeto de interés y del diseño muestral que se utilice. Para el $MAS(N,n)$ los estimadores más usuales, sus varianzas y los estimadores de las mismas son los siguientes:

Parámetro: θ	Estimador $\widehat{\theta}$	Varianza de $\widehat{\theta}$ $\sigma^2\left(\widehat{\theta}\right)$	Estimador de $\sigma^2\left(\widehat{\theta}\right)$ $\widehat{\sigma}^2\left(\widehat{\theta}\right)$
Total poblacional: $T(Y)$	$N\overline{y(m)}$	$N^2 \frac{1-f}{n} \sigma_c^2$	$N^2 \frac{1-f}{n} S_c^2$
Media poblacional: $\overline{Y}$	$\overline{y(m)}$	$\frac{1-f}{n} \sigma_c^2$	$\frac{1-f}{n} S_c^2$
Proporción muestral: P	p	$\frac{1-f}{n} \frac{N}{N-1} p(1-p)$	$\frac{1-f}{n-1} p(1-p)$

donde $f = \frac{n}{N}$.

De las expresiones de las varianzas de los estimadores, se obtiene que al incrementar el tamaño muestral, n, disminuye la varianza del estimador, y por tanto, aumenta la precisión de éste.

Conviene hacer notar también que los estimadores recogidos en la tabla anterior son insesgados.

2. Intervalos de confianza

De forma inmediata, a partir de la tabla anterior construimos la siguiente:

Parámetro: θ	Intervalo de confianza a nivel de significación $1-\alpha$ $\widehat{\theta} \mp z_{1-\frac{\alpha}{2}} \widehat{\sigma}\left(\widehat{\theta}\right)$
Total poblacional: $T(Y)$	$N\overline{y(m)} \mp z_{1-\frac{\alpha}{2}} \sqrt{N^2 \frac{1-f}{n} S_c^2}$
Media poblacional: $\overline{Y}$	$\overline{y(m)} \mp z_{1-\frac{\alpha}{2}} \sqrt{\frac{1-f}{n} S_c^2}$
Proporción muestral: P	$p \mp z_{1-\frac{\alpha}{2}} \sqrt{\frac{1-f}{n-1} p(1-p)}$

3. Determinación del tamaño muestral en un caso particular

Para ilustrar el procedimiento descrito anteriormente, y dada su importancia práctica, a continuación se determina la expresión del tamaño muestral, n, para unos valores prefijados de δ y $1-\alpha$, cuando el parámetro objeto de estudio es la proporción poblacional P.

Un estimador insesgado de P para el $MAS(N,n)$ es la proporción muestral p, y como se trata de un parámetro adimensional, el objetivo es determinar n, de forma que

$$P\left[|P-p| \leq \delta\right] \geq 1-\alpha$$

Teniendo en cuenta la primera tabla que hemos dado resulta:

$$P\left[|P-p| \leq z_{1-\frac{\alpha}{2}} \sqrt{\frac{1-f}{n} \frac{N}{N-1} P(1-P)}\right] = 1-\alpha$$

Como $0 \leq P \leq 1$ entonces $P(1-P) \leq \frac{1}{4}$, de donde se deduce que

$$P\left[|P-p| \leq z_{1-\frac{\alpha}{2}} \sqrt{\frac{1-f}{n} \frac{N}{N-1} \frac{1}{4}}\right] \geq 1-\alpha$$

y por tanto

$$\delta = z_{1-\frac{\alpha}{2}} \sqrt{\frac{1-f}{n} \frac{N}{N-1} \frac{1}{4}}$$

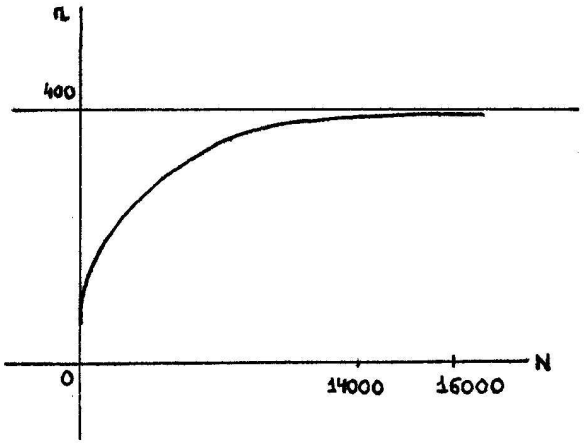
Así,

$$n = \frac{n_0}{1+\frac{n_0-1}{N}} \qquad \text{con} \qquad n_0 = \left(\frac{z_{1-\frac{\alpha}{2}}}{2\delta}\right)^2$$

Como consecuencia de lo anterior tenemos:

- Se verifica que $n \leq n_0$
- La disminución de δ lleva asociado un crecimiento de n_0
- El aumento del nivel de confianza, $1-\alpha$, lleva asociado un incremento del percentil $z_{1-\frac{\alpha}{2}}$, y por tanto, un incremento de n_0.
- Si $1-\alpha = 0.95 \Rightarrow z_{1-\frac{\alpha}{2}} = z_{0.975} = 1.96 \Rightarrow n_0 \simeq \frac{1}{\delta^2}$
- A partir deltamaño poblacional $N = (n_0 - 1)^2$, el tamaño muestral permanece constante e igual a n_0.

Para terminar vamos a representar gráficamente N frente a n : $\delta = 0.05$, $1-\alpha = 0.95$, $n_0 = 400$



A partir del tamaño poblacional $N = (n_0 - 1)^2 = 159201$ el tamaño muestral es constante $n = n_0$, y por tanto, a partir de este N el tamaño muestral no se ve afectado por el tamaño poblacional.