Capítulo 1

Muestreo probabilístico. Estimadores

Como hemos definido anteriormente, el muestreo es el proceso que nos permite la extracción de una muestra a partir de una población. Dentro de este muestreo vamos a tratar en concreto con el denominado *muestreo probabilístico*. Bajo este tipo de muestreo, la probabilidad de seleccionar un elemento de la población se conoce o puede calcularse.

El objetivo fundamental de cualquier estudio de muestreo consiste en realizar inferencias sobre una población de interés. Estas inferencias se basan en la información contenida en una muestra seleccionada de la población. Generalmente estos estudios se centran en la investigación de ciertas características de la población dependientes de una variable objetivo. Estas características se denominan *parámetros poblacionales*. La estimación de estos parámetros se realizará mediante una función de los valores contenidos en la muestra. Esta función se denomina *estimador* y, por utilizar muestreo probabilístico, es una variable aleatoria.

1.1. Diseño muestral

Para un tamaño de muestra fijo, n, existen numerosos diseños o procedimientos de muestreo para obtener las n observaciones en la muestra. En esta sección se abordará el fundamento matemático que lleva consigo la extracción de muestras dentro de una población. Para ello, revisaremos también ciertos conceptos de cálculo de probabilidades que necesitaremos a lo largo del capítulo.

1.1.1. Espacio muestral

Antes de comenzar con la exposición de esta sección, repasaremos algunos conceptos básicos del cálculo de probabilidades.

Entendemos por *experimento* la observación de un fenómeno físico. De cada realización (ensayo o prueba) de dicho experimento se obtiene un resultado. Por ejemplo, la hora de salida del sol, la velocidad con la que un objeto cae, la temperatura con la que el agua se evapora. En la vida real existen experimentos cuyo resultado puede predecirse con exactitud si se conocen las condiciones en las que se desarrolla. Por ejemplo, con una presión atmosférica de 760 mm si se calienta agua a 100 grados centígrados entonces se transforma en vapor. A estos experimentos se les denomina *determinísticos*. Por el contrario existen experimentos cuyo resultado final no se puede predecir con exactitud aunque podemos afirmar algo con respecto a la frecuencia con que se producen. Por ejemplo, el lanzamiento de una moneda, la selección de una carta en la baraja, ... Si arrojamos una moneda un gran número de veces y esta moneda no está trucada aproximadamente la mitad de las veces se obtendría cara y la otra mitad cruz. Y cuanto mayor sea el número de lanzamientos, más proxima a 0.5 será la relación de caras cruces obtenidas respecto al número de lanzamientos efectuados. Estos son los denominados *experimentos aleatorios*.

El primer paso para construir un modelo matemático de un experimento aleatorio se basa en definir el conjunto de todos los posibles resultados asociados a dicho experimento y definir un conjunto con todos ellos. A este conjunto se le denomina *espacio muestral*. Se llama suceso elemental a cada uno de los posibles resultados de un experimento aleatorio, siempre que estos resultados sean mutuamente excluyentes y equiprobables (no pueden aparecer dos a la vez y la probabilidad de cada uno de ellos es la misma). Si Ω es el espacio muestral asociado a un experimento aleatorio llamamos *suceso* a cualquier subconjunto s tal que $s \subset \Omega$.

Si consideramos como fenómeno o experimento aleatorio la extracción aleatoria de muestras dentro de una población por un procedimiento o método de muestreo dado, podemos considerar como sucesos las muestras obtenidas. En lo sucesivo, denotaremos por S al conjunto formado por todas las muestras extraídas mediante un procedimiento de muestreo determinado.

Ejemplo 1 *Definir el conjunto S, es decir, el conjunto formado por todas las muestras que resultan de la extracción sin reemplazamiento de dos unidades de una población de 4 elementos sin tener en cuenta el orden de los elementos.*

Considerando la población $\Omega = \{u_1, u_2, u_3, u_4\}$, el conjunto S viene dado por

$$S = \{(u_1, u_2), (u_1, u_3), (u_1, u_4), (u_2, u_3), (u_2, u_4), (u_3, u_4)\}$$

Como comentábamos en el tema anterior, mediante muestreo probabilístico es posible asignar a cada muestra una probabilidad conocida de ser seleccionada de manera que podemos construir una función P definida en el conjunto de todas las muestras S y que toma valores en el intervalo $[0, 1]$,

$$P(\cdot) : S \longrightarrow [0, 1] \tag{1.1}$$

tal que

- $P(s) \geq 0, \quad \forall s \in S$
- $\sum_{s \in S} P(s) = 1.$

En numerosas ocasiones, al par $(S, P(\cdot))$ se le denomina *diseño muestral*.

Ejemplo 2 *Volvamos al ejemplo 1 y supongamos que todas las muestras son equiprobables (es decir, tienen la misma probabilidad de ser seleccionadas). Obtener el diseño muestral asociado a este experimento.*

En este caso, la función P viene dada por

$$P(\cdot) : S \longrightarrow [0, 1]$$

donde $S = \{(u_1, u_2), (u_1, u_3), (u_1, u_4), (u_2, u_3), (u_2, u_4), (u_3, u_4)\}$ *y*

$$P(s) = \frac{1}{6}, \quad s \in S.$$

Ejemplo 3 *En una población de 3 unidades $\Omega = \{u_1, u_2, u_3\}$ se extraen muestras de tamaño 1 mediante el siguiente método de muestreo: Se extrae al azar 1 bola de una urna que contiene 6 bolas (tres con el número 1, dos con el número 2 y una con el número 3), y se extrae de la población la unidad que tenga el mismo número que la bola extraída. Determinar el diseño muestral para este procedimiento de muestreo.*

Solución *Como se extrae una única bola de la urna, cada una de las muestras serían las unidades $S = \{u_1\}$, $\{u_2\}$ y $\{u_3\}$ y la probabilidad de selección de cada muestra viene dada por*

$$P(\{u_1\}) = \frac{3}{6}, \quad P(\{u_2\}) = \frac{2}{6}, \quad P(\{u_3\}) = \frac{1}{6}.$$

Ejemplo 4 *En una población de 3 unidades numeradas $\Omega = \{u_1, u_2, u_3\}$ se extraen muestras de tamaño 2 mediante el siguiente método de muestreo: Se extraen al azar 2 bolas de una urna que contiene 6 bolas (tres con el número 1, dos con el número 2 y una con el número 3) considerando la extracción de las bolas en una urna con reposición, y se extraen de la población las dos unidades que tengan los mismos números que las dos bolas extraídas. Determinar el diseño muestral para este procedimiento de muestreo considerando muestras no ordenadas.*

Solución *En este caso extraemos dos unidades de la población y además dicha extracción es sin reemplazamiento. Como el muestreo es con reposición, el conjunto de todas las muestras es igual a*

$$S = \{\{u_1, u_1\}, \{u_2, u_2\}, \{u_3, u_3\}, \{u_1, u_2\}, \{u_1, u_3\}, \{u_2, u_3\}\},$$

y las probabilidades de selección de cada muestra vienen dadas por

$$P(\{u_1,u_1\})=\frac{3}{6}\frac{3}{6},\quad P(\{u_2,u_2\})=\frac{2}{6}\frac{2}{6},\quad P(\{u_3,u_3\})=\frac{1}{6}\frac{1}{6}$$
$$P(\{u_1,u_2\})=2\frac{3}{6}\frac{2}{6},\quad P(\{u_2,u_3\})=2\frac{2}{6}\frac{1}{6},\quad P(\{u_1,u_3\})=2\frac{3}{6}\frac{1}{6}$$

1.1.2. Probabilidades de inclusión

Dada una población $\Omega=\{u_1,u_2,\ldots,u_N\}$ y dado un diseño muestral determinado $(S,P(\cdot))$ asociado a dicha población, para cualquier muestra $s\in S$, un elemento cualquiera de la población u_k $k\in\{1,\ldots,N\}$ puede pertenecer o no a la muestra. Para representar la pertenencia o no a la muestra, se define la variable *indicador de pertenencia a la muestra* como la siguiente aplicación

$$I_k : S \longrightarrow \{0,1\}$$

de manera que

$$I_k(s)=1,\quad \text{si}\quad u_k\in s\quad I_k(s)=0\quad \text{si}\quad u_k\notin s\quad \forall s\in S,\quad \forall u_k\in\Omega,$$

por lo tanto, I_k es una variable aleatoria definida sobre S y su distribución de probabilidad viene dada por

$$P[I_k=1]=\sum\nolimits_{s\in S;u_k\in s}P(s)=\pi_k\quad P[I_k=0]=1-\pi_k.$$

De esta forma, π_k es la probabilidad de que el elemento u_k esté en la muestra y se denomina *probabilidad de inclusión de primer orden.* Además, como la variable I_k es de tipo Bernouilli, su esperanza y su varianza vienen dadas por

$$E[I_k]=\pi_k,\quad Var(I_k)=\pi_k(1-\pi_k).$$

Ejemplo 5 *Calcular las probabilidades de inclusión de primer orden para cada una de los elementos de la población dados en los ejemplos 3 y 4*

Solución. *Calculamos la probabilidad de inclusión de primer orden para el Ejemplo 3. En este caso, tenemos que calcular π_1, π_2 y π_3 donde π_k, $k=1,2,3$ es la probabilidad de que el elemento u_k esté en una muestra determinada. En este caso,*

$$\begin{aligned}\pi_1 &= \sum_{s\in S;u_1\in S}P(s)=P(\{u_1\})=\frac{3}{6}\\ \pi_2 &= \sum_{s\in S;u_2\in S}P(s)=P(\{u_2\})=\frac{2}{6}\\ \pi_3 &= \sum_{s\in S;u_3\in S}P(s)=P(\{u_3\})=\frac{1}{6}.\end{aligned}$$

En este caso, tenemos que $\pi_1 + \pi_2 + \pi_3 = 1$ pero no siempre sucede así. Para el Ejemplo 4, tenemos que

$$\begin{aligned}
\pi_1 &= \sum_{s \in S; u_1 \in S} P(s) = P(\{u_1, u_1\}) + P(\{u_1, u_2\}) + P(\{u_1, u_3\}) = \frac{27}{36} \\
\pi_2 &= \sum_{s \in S; u_2 \in S} P(s) = P(\{u_2, u_2\}) + P(\{u_1, u_2\}) + P(\{u_2, u_3\}) = \frac{20}{36} \\
\pi_3 &= \sum_{s \in S; u_3 \in S} P(s) = P(\{u_3, u_3\}) + P(\{u_1, u_3\}) + P(\{u_2, u_3\}) = \frac{11}{36}.
\end{aligned}$$

Para las probabilidades obtenidas en este ejemplo, notar que $\pi_1 + \pi_2 + \pi_3 \neq 1$.

De manera análoga a las propiedades de primer orden, dadas dos unidades distintas u_i y u_j de la población Ω, $i \neq j$, se define la variable indicador de pertenencia a la muestra, I_{ij}, como

$$I_{ij} : S \longrightarrow \{0, 1\}$$

de manera que

$$I_{ij}(s) = 1, \quad \text{si} \quad u_i, u_j \in s \quad I_{ij}(s) = 0 \quad \text{en otro caso} \quad \forall s \in S, \quad \forall u_i, u_j \in \Omega,$$

por lo tanto I_{ij} es una variable aleatoria definida sobre S que toma los valores 0 y 1 con la siguiente probabilidad

$$\begin{aligned}
P[I_{ij} = 1] = \sum_{s \in S; u_i, u_j \in S} P(s) = \pi_{ij} \\
P[I_{ij} = 0] = 1 - \pi_{ij},
\end{aligned}$$

o lo que es lo mismo, sigue una distribución de Bernouilli de parámetro π_{ij}. A las probabilidades π_{ij} se le denomina *probabilidad de inclusión de segundo orden*

A partir de las probabilidades de inclusión de primer y segundo orden, podemos definir la *matriz del diseño muestral* como la siguiente matriz simétrica

$$\pi = (\pi_{ij})_{1 \leq i,j \leq N},$$

donde $\pi_{ii} = \pi_i$, $\forall i$. La importancia de esta matriz radica en el hecho de que la mayoría de los estimadores de los parámetros usuales y de sus varianzas son funciones de los elementos de esta matriz.

Ejemplo 6 *Calcular las probabilidades de inclusión de segundo orden para cada una de los elementos de la población dados en los ejemplos 3 y 4*

Solución. *Para el Ejemplo 3, las probabilidades de inclusión de segundo orden vienen dadas por*

$$\pi_{ij} = 0, \quad \forall i, j \in \{1, 2, 3\}.$$

En el caso del Ejemplo 4, las probabilidades de inclusión de segundo orden vienen dadas por

$$\begin{aligned}\pi_{12} &= P(\{u_1,u_2\}) = 2\frac{3}{6}\frac{2}{6}\\ \pi_{13} &= P(\{u_1,u_3\}) = 2\frac{3}{6}\frac{1}{6}\\ \pi_{23} &= P(\{u_2,u_3\}) = 2\frac{2}{6}\frac{1}{6}.\end{aligned}$$

Ejemplo 7 *Para la población $\Omega = \{u_1,u_2,u_3\}$ consideramos el siguiente proceso de selección de muestras de tamaño 2. Se extraen una primera unidad con probabilidades iguales de selección entre las 3 unidades y si ésta resulta ser u_1 se extrae la segunda unidad entre las dos restantes también con probabilidades iguales; pero si la primera no es u_1, la segunda se extrae de las tres que componen la población asignando doble probabilidad a u_1 que a cada una de las otras dos. Hallar el espacio muestral y las probabilidades asociadas a cada una de las muestras. Obtener las probabilidades de primera inclusión para todas las unidades que forman la población y las probabilidades de inclusión de segundo orden.*

Solución. *El conjunto de todas las muestras asociadas a este procedimiento de muestreo vienen dadas por*

$$S = \{(u_1,u_3),(u_1,u_2),(u_2,u_2),(u_2,u_1),(u_2,u_3),(u_3,u_1),(u_3,u_2),(u_3,u_3)\},$$

y las probabilidades asociadas a cada muestra vienen dadas por

$$\begin{aligned}P(\{u_1,u_3\}) &= P(\{u_1,u_2\}) = \frac{1}{6}\\ P(\{u_2,u_2\}) &= P(\{u_2,u_3\}) = P(\{u_3,u_2\}) = P(\{u_3,u_3\}) = \frac{1}{12}\\ P(\{u_2,u_1\}) &= P(\{u_3,u_1\}) = \frac{2}{12}.\end{aligned}$$

Las probabilidades de primera inclusión para cada una de las unidades vienen dadas como

$$\begin{aligned}\pi_1 &= P(\{u_1,u_3\}) + P(\{u_1,u_2\}) + P(\{u_2,u_1\}) + P(\{u_3,u_1\}) = \frac{8}{12}\\ \pi_2 &= P(\{u_1,u_2\}) + P(\{u_2,u_2\}) + P(\{u_2,u_3\}) + P(\{u_3,u_2\}) + P(\{u_2,u_1\}) = \frac{7}{12}\\ \pi_3 &= P(\{u_1,u_3\}) + P(\{u_2,u_3\}) + P(\{u_3,u_2\}) + P(\{u_3,u_3\}) + P(\{u_3,u_1\}) = \frac{7}{12}.\end{aligned}$$

En el caso de las probabilidades de segunda inclusión vienen dadas por

$$\pi_{ij} = P(\{u_i,u_j\})$$

1.2. Estadísticos y estimadores

El objetivo fundamental de cualquier estudio de muestreo consiste en realizar inferencias sobre la población de interés. Esta inferencia se basa en la información contenida en la muestra seleccionada de la población utilizando un procedimiento de muestreo determinado.

Recordemos que denotamos por $\Omega = \{u_1, u_2, \ldots, u_N\}$ al conjunto de N unidades que forman la población de estudio y por $s = \{u_1, u_2, \ldots, u_n\}$ al subconjunto de n unidades que forman la muestra s, seleccionada del espacio muestral Ω según un determinado procedimiento de muestreo.

Generalmente, el investigador pretende la estimación de ciertas características de la población y que dependen de la variable de estudio.

Llamamos X a la variable o característica de estudio medida sobre cada uno de los elementos de la población $\{X_1, X_2, \ldots, X_N\}$ donde X_i representa el valor de la característica X sobre el elemento i-ésimo de la población. En la mayoría de las ocasiones estamos interesados en ciertas funciones de los elementos $\{X_1, X_2, \ldots, X_N\}$ como son

- Total de la característica X sobre todos los elementos de la población

$$\sum_{i=1}^{N} X_i,$$

- Media aritmética de los valores de X sobre todos los elementos de la población

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

Estas funciones se denominan *parámetros poblacionales*.

Definición 1.1 *Cualquier función de las observaciones de la variable objetivo sobre los elementos de la población se denomina parámetro poblacional o simplemente parámetro y lo representamos por θ.*

A partir de los datos observados de la variable X sobre las unidades de la muestra $\{X_1, X_2, \ldots, X_n\}$, podemos construir funciones matemáticas que vamos a denominar *estadísticos* y que nos van a ayudar a estimar el valor del parámetro poblacional desconocido.

Definición 1.2 *Una función de las observaciones de la variable objetivo sobre los elementos de una muestra se denomina estadístico. Si este estadístico se utiliza para estimar un parámetro, se denomina estimador.*

El valor particular que el estimador toma para una muestra dada se denomina *estimación* o *estimación puntual*. De manera formal, el estimador puede expresarse como

$$\begin{array}{rcl} \widehat{\theta} : S & \longrightarrow & R \\ s = (u_1, u_2, \ldots, u_n) & \longrightarrow & \widehat{\theta}(X_1, X_2, \ldots, X_n) = t \end{array}$$

Como ejemplo de estimadores que se utilizan normalmente para aproximar los valores poblacionales dados anteriormente se utilizan el total de los valores o la media de los valores de una realización muestral que se expresarían respectivamente como

$$\begin{aligned} \widehat{\theta}(u_1, u_2, \ldots, u_n) &= \widehat{X} = \sum_{i=1}^{n} X_i \\ \widehat{\theta}(u_1, u_2, \ldots, u_n) &= \widehat{\widehat{X}} = \frac{\sum_{i=1}^{n} X_i}{n}. \end{aligned}$$

Por lo tanto, las *estimaciones* son números que resumen información sobre la muestra y los *parámetros* son números que resumen información sobre la población.

Ejemplo 8 *Para medir la variable X nivel de precipitación atmosférica en una determinada región disponemos de un marco de 4 zonas climáticas de la misma cuyos niveles de precipitación actual son de 6, 4, 3, y 8 decenas de litros por metro cuadrado. Se trata de estimar en decenas de litros por metro cuadrado el nivel actual medio de precipitación atmosférica en la región extrayendo muestras de tamaño 2 sin reposición y sin tener en cuenta el orden de colocación de sus elementos utilizando el estimador media aritmética de la muestra. Se supone que todas las zonas climáticas tienen la misma probabilidad de ser seleccionadas. Especificar el diseño muestral y las estimaciones correspondientes para cada muestra.*

Solución. *Para este ejemplo se tiene una población de 4 unidades $\Omega = \{u_1, u_2, u_3, u_4\}$ que corresponde a cada una de las regiones que conforma la población. Denotamos por X la variable "nivel de precipitación" y sea X_i el valor que toma la variable X sobre la unidad u_i con $i = 1, 2, 3, 4$. El procedimiento de muestreo considerado consiste en extraer muestras de tamaño 2 sin reposición y sin tener en cuenta el orden de los elementos. Por lo tanto, el diseño muestral para este estudio de muestreo viene dado por $(S, P(\cdot))$ donde*

$$S = \{(u_1, u_2), (u_1, u_3), (u_1, u_4), (u_2, u_3), (u_2, u_4), (u_3, u_4)\},$$

y $P(s) = 1/6$ para todo $s \in S$. Para estimar el parámetro poblacional

$$\bar{X} = \frac{1}{4}\sum_{i=1}^{4} X_i,$$

el problema considera el estimador media muestral

$$\widehat{\widehat{X}} = \frac{1}{2}\sum_{i=1}^{2} X_i.$$

Los valores que toma este estimador sobre cada una de las muestras se representa en la siguiente tabla

S	$p(s)$	$X(s)$	$\widehat{\bar{X}}$
(u_1, u_2)	*1/6*	*(6,4)*	*5*
(u_1, u_3)	*1/6*	*(6,3)*	*4.5*
(u_1, u_4)	*1/6*	*(6,8)*	*7*
(u_2, u_3)	*1/6*	*(4,3)*	*3.5*
(u_2, u_4)	*1/6*	*(4,8)*	*6*
(u_3, u_4)	*1/6*	*(3,8)*	*5.5*

Las estimaciones basadas en observaciones muestrales difieren de muestra a muestra y también del valor del parámetro poblacional. El estimador, por lo tanto, es una variable aleatoria. Esto nos lleva a introducir el concepto de *distribución en el muestreo del estimador.*

Definición 1.3 *Dada una población y un procedimiento de muestro, fijado el tamaño muestral, el conjunto de valores posibles de un estimador con su probabilidad de ocurrencia se denomina distribución muestral de dicho estimador.*

Un estimador puede representarse como la siguiente aplicación

$$\begin{array}{rcl} \widehat{\theta} : S & \longrightarrow & R \\ (u_1, u_2, \ldots, u_n) & \longrightarrow & \widehat{\theta}(X_1, X_2, \ldots, X_n) = t \end{array}$$

Formalmente, la distribución del estimador en el muestreo puede representarse como sigue. Sea

$$T = \{t \in R / \, \exists s = (u_1, u_2, \ldots, u_n) \in S; \, \widehat{\theta}(X_1, X_2, \ldots, X_n) = t\}.$$

El conjunto $T \in R$ constituye el conjunto de valores del estimador. Ahora vamos a definir las probabilidades de que el estimador tome estas valores como sigue

$$P^T(\widehat{\theta}(X_1, X_2, \ldots, X_n) = t) \sum_{\{s \in S / \widehat{\theta}(s) = t\}} P(s).$$

Al par $\{T, P^T\}$ formado por el conjunto de todos los posibles valores del estimador y por las probabilidades de que el estimador tome esos valores se le denomina *distribución del estimador en el muestreo.*

Ejemplo 9 *Considerando el Ejemplo 8, obtener la distribución en el muestreo del estimador media muestral para ese procedimiento de muestreo.*

Solución. *Los valores que toma el estimador son 5, 4.5, 7, 3.5, 6, 5.5 con la siguiente probabilidad*

$$P[\widehat{\bar{X}} = t] = \frac{1}{6}, \quad t = 5, 4.5, 7, 3.5, 6, 5.5.$$

Ejemplo 10 *En una población de $N = 10$ unidades se encuentran éstas formando 4 subconjuntos $A(i)$ $i = 1, 2, 3, 4$. Los valores de una característica X medida sobre los elementos de la población se presenta en la siguiente tabla adjunta.*

$A(i)$	$A(1)$	$A(2)$	$A(3)$	$A(4)$
X	$1, 2, 3$	$4, 6$	$9, 11$	$2, 2, 5$

Se considera un procedimiento de muestreo que consiste en elegir cada subconjunto $A(i)$ con probabilidades proporcionales a sus tamaños. Se considera el estimador $\widehat{T}_1$ "media aritmética de los muestra" para estimar la media poblacional y se considera el estimador $\widehat{T}_2$ "total de los elementos de la muestra" para estimar el total poblacional. Especificar el diseño muestral y las distribuciones de probabilidades en el muestreo de los estimadores $\widehat{T}_1$ y $\widehat{T}_2$.

Solución. *En este estudio estadístico, el procedimiento de muestreo consiste en elegir cada subconjunto con probabilidades proporcionales al tamaño de cada subconjunto. Luego la muestra seleccionada será alguno de los subconjuntos $A(i)$ con $i = 1, 2, 3, 4$, es decir*

$$S = \{A(1), A(2), A(3), A(4)\},$$

estos subconjuntos no tienen el mismo tamaño y la probabilidad de seleccionar cada una de estas muestras depende de ese tamaño. Sea $n(A(i))$ el número de elementos de $A(i)$, es decir,

$$n(A(1)) = 3,\ n(A(2)) = 2, n(A(3)) = 2, n(A(4)) = 3.$$

La probabilidad de elección de cada uno de estos bloques es proporcional a su número de elementos, luego,

$$P(A(1)) = 3x,\ P(A(2)) = 2x,\ P(A(3)) = 2x,\ P(A(4)) = 3x,$$

y la suma de las probabilidades de cada una de las muestras es 1, entonces

$$\sum_{i=1}^{4} P(A(i)) = 1 \Rightarrow x = \frac{1}{10},$$

y finalmente

$$P(A(1)) = \frac{3}{10}, \quad P(A(2)) = \frac{2}{10}, \quad P(A(3)) = \frac{2}{10}, \quad P(A(4)) = \frac{3}{10}.$$

Consideramos los estimadores $\widehat{T}_1$ y $\widehat{T}_2$, para cada una de las muestras obtenidas utilizando este procedimiento de muestreo el valor de este estimador es igual a

S	$P(s)$	$X(s)$	$\widehat{T}_1$	$\widehat{T}_2$
$A(1)$	$3/10$	$(1,2,3)$	2	6
$A(2)$	$2/10$	$(4,6)$	5	10
$A(3)$	$2/10$	$(9,11)$	10	20
$A(4)$	$3/10$	$(2,2,5)$	3	9

y la distribución de probabilidad del estimador en el muestreo es igual a

$$\begin{aligned} P[\widehat{T}_1 = 2] &= \frac{3}{10}, \quad P[\widehat{T}_1 = 5] = \frac{2}{10}, \quad P[\widehat{T}_1 = 10] = \frac{2}{10}, \quad P[\widehat{T}_1 = 3] = \frac{3}{10} \\ P[\widehat{T}_2 = 6] &= \frac{3}{10}, \quad P[\widehat{T}_2 = 10] = \frac{2}{10}, \quad P[\widehat{T}_2 = 20] = \frac{2}{10}, \quad P[\widehat{T}_2 = 9] = \frac{3}{10}. \end{aligned}$$

Hasta ahora los parámetros poblacionales que hemos considerado son

- Total poblacional para la variable X, X,

$$\theta = \theta(X_1, X_2, \ldots, X_N) = \sum_{i=1}^{N} X_i$$

- Media poblacional para la característica X,

$$\theta = \theta(X_1, X_2, \ldots, X_N) = \sum_{i=1}^{N} X_i/N$$

Sin embargo, estos parámetros poblacionales tienen sentido definirlos si la variable de estudio X es cuantitativa (cuantificable numéricamente) como puede ser el peso, la altura, ingresos, etc. Sin embargo, para variables que no sean cuantitativas, sino que sean cualitativas, también pueden definirse parámetros poblacionales análogos. Para ello, al contrario de lo que sucede con variables cuantitativas, no medimos el valor que toma la variable X sobre cada unidad de la población sino lo que se hace es analizar sobre cada unidad de la población su pertenencia o no a una determinada clase.

Si para cada unidad u_i $i = 1, 2, \ldots, N$ de la población definimos la característica A_i que toma el valor 1 si la unidad u_i pertenece a la clase A y que toma el valor 0 si la unidad u_i no pertenece a la clase A podemos definir el total de elementos de la población que pertenecen a la clase A (total de la clase) y la proporción de elementos de la población que pertenecen a la clase A (proporción de clase) del siguiente modo:

- Total de clase,

$$\theta = \theta(A_1, A_2, \ldots, A_N) = \sum_{i=1}^{N} A_i = A$$

- Proporción de clase,

$$\theta = \theta(A_1, A_2, \ldots, A_N) = \sum_{i=1}^{N} A_i/N = P.$$

Ejemplo 11 *Consideramos una población de 15 personas y queremos estimar el total de personas y la proporción de personas que tienen los ojos verdes. Se alinean estas 15 personas y se les anota el color de sus ojos que vienen representados por el siguiente vector*

$$C, A, N, N, A, A, V, V, A, V, N, C, C, C, A,$$

donde C es el color de ojos castaños, A el color de ojos azules, N, color de ojos negros y V el color de ojos verdes. Para realizar la estimación seleccionamos muestras de tamaño 5 utilizando un muestreo sistemático y consideramos como estimador del total de personas que tiene los ojos verdes en la población el total de personas en la muestra con los ojos verdes y como estimador de la proporción de personas que tienen los ojos verdes en la población, la proporción de personas en la muestra que tienen los ojos verdes. Si todas las muestras son equiprobables, obtener el diseño muestral para este estudio y la distribución de probabilidad de los estimadores.

Solución. *El diseño muestral y los valores que los estimadores toman en cada muestra vienen dados en el siguiente cuadro*

S	$p(s)$	$X(s)$	$\widehat{A}$	$\widehat{P}$
$(u_1, u_4, u_7, u_{10}, u_{13})$	1/3	(C,N,V,V,C)	2	2/5
$(u_2, u_5, u_8, u_{11}, u_{14})$	1/3	(A,A,V,N,C)	1	1/5
$(u_3, u_6, u_9, u_{12}, u_{15})$	1/3	(N,A,A,C,A)	0	0

Para cada parámetro poblacional es posible construir numerosos estimadores. Sin embargo es deseable que estos estimadores verifiquen una serie de propiedades. En la siguiente sección analizaremos alguna de estas propiedades.

1.2.1. Propiedades de los estimadores

El estimador $\widehat{\theta}$ de un parámetro poblacional θ es una variable aleatoria ya que, aunque depende unívocamente de los valores de la muestra observados $(X = x_i)$, la elección de la muestra es un proceso aleatorio. De un buen estimador se espera que verifique dos propiedades importantes. Una de estas propiedades se denomina *insesgadez*. La otra propiedad es que los valores del estimador no se alejen del verdadero valor del parámetro poblacional.

Antes de analizar estas propiedades, nos vamos a centrar en el análisis de ciertas características de centralización y dispersión de dichos estimadores, particularmente su esperanza, su varianza y sus momentos así como otras medidas relativas a su precisión.

Definición 1.4 *Sea $\widehat{\theta}$ un estimador del parámetro poblacional θ. Se define la esperanza de dicho estimador como*

$$E[\widehat{\theta}] = \sum_{t \in R} tP(\widehat{\theta} = t).$$

Es decir, la esperanza matemática de un estimador coincide con la esperanza de la variable aleatoria $\widehat{\theta}$.

Definición 1.5 *Sea $\widehat{\theta}$ un estimador del parámetro poblacional θ. El estimador $\widehat{\theta}$ es insesgado para el parámetro poblacional θ si $E[\widehat{\theta}] = \theta$.*

En el caso en el que $E[\widehat{\theta}]$ no es igual al valor del parámetro poblacional θ, el estimador $\widehat{\theta}$ se dice que es sesgado con respecto a θ.

Definición 1.6 *Si para un estimador $\widehat{\theta}$, $E[\widehat{\theta}] \neq \theta$, el estimador $\widehat{\theta}$ se denomina estimador sesgado de θ. La magnitud de este sesgo en $\widehat{\theta}$ viene dado por*

$$B(\widehat{\theta}) = E[\widehat{\theta}] - \theta.$$

El cociente

$$RB(\widehat{\theta}) = \frac{B(\widehat{\theta})}{\theta},$$

se denomina *sesgo relativo* del estimador $\widehat{\theta}$.

Ejemplo 12 *Considerando el Ejemplo 11, calcular la esperanza de los estimadores $\widehat{P}$ y $\widehat{A}$.*

Solución. *Tenemos que la distribución de los estimadores $\widehat{P}$ y $\widehat{A}$ viene dada por*

$$P[\widehat{A} = 2] = P[\widehat{A} = 1] = P[\widehat{A} = 0] = \frac{1}{3}$$

$$P[\widehat{P} = 2/5] = P[\widehat{P} = 1/5] = P[\widehat{P} = 0] = \frac{1}{3},$$

luego la esperanza para cada uno de estos estimadores es igual a

$$E[\widehat{A}] = 1, \quad E[\widehat{P}] = \frac{1}{5}.$$

Los valores poblacionales en este caso son $A = 3$ y $P = 1/5$ de manera que el estimador $\widehat{P}$ es insesgado y el estimador $\widehat{A}$ es sesgado. El sesgo de estos estimadores es igual a

$$B(\widehat{P}) = 0, \quad B(\widehat{A}) = E[\widehat{A}] - A = 1 - 3 = -2.$$

Los sesgos relativos para ambos estimadores vienen dados por

$$RB(\widehat{P}) = 0, \quad RB(\widehat{A}) = -\frac{2}{3}.$$

Además de la insesgadez, una propiedad importante que deben de verificar los estimadores es que tengan una varianza pequeña.

Definición 1.7 *Sea $\widehat{\theta}$ un estimador del parámetro poblacional θ. Se define la* varianza *de $\widehat{\theta}$, y se denota por $Var(\widehat{\theta})$, a la siguiente expresión*

$$\begin{aligned} Var(\widehat{\theta}) &= E(\widehat{\theta} - E(\widehat{\theta}))^2 \\ &= \sum_{t \in R} (t - E(\widehat{\theta}))^2 P(\widehat{\theta} = t) \\ &= E(\widehat{\theta}^2) - (E(\widehat{\theta}))^2. \end{aligned}$$

Es decir, la varianza es una medida que cuantifica la concentración de las estimaciones alrededor de su valor medio.

Definición 1.8 *Sea $\widehat{\theta}$ un estimador del parámetro poblacional θ. Se define el* error de muestreo *o* error de estimación *del estimador $\widehat{\theta}$ como su desviación típica, es decir, la raíz cuadrada de su varianza. Su expresión es la siguiente*

$$\sigma(\widehat{\theta}) = +\sqrt{V(\widehat{\theta})}.$$

Ejemplo 13 *Obtener los errores de estimación asociados a los estimadores $\widehat{P}$ y $\widehat{A}$ del ejercicio 11.*

Solución. *La varianza para el estimador $\widehat{A}$ viene dado por*

$$\begin{aligned} Var(\widehat{A}) &= \sum_{x} x^2 P[\widehat{A} = x] - (E[\widehat{A}])^2 \\ &= 2^2 \frac{1}{3} + 1^2 \frac{1}{3} - 1 \\ &= \frac{2}{3} \end{aligned}$$

luego el error de estimación asociado a $\widehat{A}$ viene dado por

$$\sigma(\widehat{A}) = \sqrt{2/3} = 0{,}8165.$$

En el caso del estimador $\widehat{P}$, su varianza viene dada por

$$\begin{aligned} Var(\widehat{P}) &= \sum_{x} x^2 P[\widehat{P} = x] - (E[\widehat{P}])^2 \\ &= \left(\frac{2}{5}\right)^2 \frac{1}{3} + \left(\frac{1}{5}\right)^2 \frac{1}{3} - \left(\frac{1}{5}\right)^2 \\ &= \frac{2}{75}, \end{aligned}$$

luego el error de estimación asociado a $\widehat{P}$ viene dado por

$$\sigma(\widehat{P}) = \sqrt{2/75} = 0{,}1633.$$

En numerosas ocasiones, la varianza y el error del estimador $\widehat{\theta}$ no son prácticos debido a que sus valores dependen de los valores de la variable en estudio

para todos los elementos de la población y generalmente estos datos no están disponibles. Para tener una idea de la magnitud del error involucrado en los valores de $\widehat{\theta}$, necesitamos estimar $V(\widehat{\theta})$ y $\sigma(\widehat{\theta})$ a partir de los datos muestrales. Sus estimadores se denotan por $\widehat{V}(\widehat{\theta})$ y $\widehat{\sigma}(\widehat{\theta})$. El término $\widehat{\sigma}(\widehat{\theta})$, denominado *estimación del error estándar del estimador* $\widehat{\theta}$, es la raíz cuadrada positiva de $\widehat{V}(\widehat{\theta})$. De este modo,

$$\widehat{\sigma}(\widehat{\theta}) = +\sqrt{\widehat{V}(\widehat{\theta})}.$$

Definición 1.9 *Sea $\widehat{\theta}$ un estimador del parámetro poblacional θ. Se define el* error relativo de muestreo *del estimador $\widehat{\theta}$ como el cociente entre su desviación típica y su valor esperado y la expresión viene dada por*

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

A diferencia del error de estimación o error de muestreo, es una medida adimensional lo que nos va a permitir comparar estimadores entre sí sin tener en cuenta las unidades de medida.

Ejemplo 14 *Obtener los errores relativos de estimación asociados a los estimadores $\widehat{P}$ y $\widehat{A}$ del ejercicio 11.*

Solución. *En este caso,*

$$CV(\widehat{A}) = \frac{\sigma(\widehat{A})}{E[\widehat{A}]} = 0{,}8165, \quad CV(\widehat{P}) = \frac{\sigma(\widehat{P})}{E[\widehat{P}]} = 0{,}8165$$

En el caso en que el estimador sea sesgado, se utiliza el denominado como *error cuadrático medio* para medir la variabilidad del estimador.

Definición 1.10 *Sea $\widehat{\theta}$ un estimador del parámetro poblacional θ. El error cuadrático medio mide la divergencia de los valores del estimador con respecto al verdadero valor del parámetro, es decir,*

$$\begin{aligned} ECM(\widehat{\theta}) &= E[(\widehat{\theta} - \theta)^2] \\ &= \sum_{t \in R} (t - \theta)^2 P(\widehat{\theta} = t). \end{aligned}$$

Ejemplo 15 *Obtener el error cuadrático medio asociados a los estimadores $\widehat{P}$ y $\widehat{A}$ del ejercicio 11.*

Solución. *En el caso del estimador $\widehat{A}$, se tiene que*

$$\begin{aligned} ECM(\widehat{A}) &= E[(\widehat{A} - A)^2] \\ &= (2-3)^2 \frac{1}{3} + (1-3)^2 \frac{1}{3} + (0-3)^2 \frac{1}{3} \\ &= \frac{14}{3}. \end{aligned}$$

Para el estimador $\widehat{P}$,

$$\begin{aligned} ECM(\widehat{P}) &= E[(\widehat{P}-P)^2] \\ &= \left(\frac{2}{5}-\frac{1}{5}\right)^2\frac{1}{3}+\left(\frac{1}{5}-\frac{1}{5}\right)^2\frac{1}{3}+\left(0-\frac{1}{5}\right)^2\frac{1}{3} \\ &= \frac{2}{75}. \end{aligned}$$

La raíz cuadrada positiva del error cuadrático medio se denomina *raíz del error cuadrático medio*. El error cuadrático medio y la varianza muestral se relacionan mediante la siguiente expresión

$$ECM(\widehat{\theta}) = \sigma(\widehat{\theta})^2 + B(\widehat{\theta})^2,$$

donde $B(\widehat{\theta})$ es el sesgo del estimador $\widehat{\theta}$. De este modo, para un estimador insesgado, el error cuadrático medio y la varianza de un estimador son equivalentes.

Ejemplo 16 *Comprobar la relación existente entre el error cuadrático medio, la varianza y el sesgo para los estimadores* $\widehat{A}$ *y* $\widehat{P}$.

Solución. *Como el estimador* $\widehat{P}$ *es insesgado, se tiene que su varianza coincide con su error cuadrático medio, es decir,*

$$ECM(\widehat{P}) = \sigma^2(\widehat{P}).$$

El estimador $\widehat{A}$ *es sesgado, luego*

$$ECM(\widehat{A}) = \sigma^2(\widehat{A}) + B(\widehat{A}).$$

La introducción del error cuadrático medio, nos permite hablar de eficiencia relativa de un parámetro.

Definición 1.11 *Si* $\widehat{\theta}_1$ *y* $\widehat{\theta}_2$ *son dos estimadores del estimador* θ*, la eficiencia relativa del estimador* $\widehat{\theta}_2$ *con respecto al estimador* $\widehat{\theta}_1$*, se define como*

$$ER = \frac{ECM(\widehat{\theta}_1)}{ECM(\widehat{\theta}_2}. \tag{1.2}$$

Ejemplo 17 *Se tiene una urna con tres bolas numerados con los valores* $\{1,2,3\}$ *respectivamente. La probabilidad de obtener cada una de las bolas es la misma*

1. *Determinar todas las posibles muestras de tamaño 2 que se pueden obtener con reemplazo y teniendo en cuenta la ordenación de los elementos en la muestra.*

2. *Determinar la distribución en el muestreo de la variable media muestral.*

3. *Calcular la esperanza, el sesgo, la varianza y el error cuadrático medio de la media muestral.*

Ejemplo 18 *Para medir la variable "nivel de concentración de sustancias tóxicas en suspensión" en una determinada zona industrial, se dispone de un marco de 4 zonas industriales cuyos niveles de concentración de sustancias tóxicas en suspensión actuales son 6, 4, 3 y 8 gramos/m^3. Estimar el nivel medio de concentración de sustancias tóxicas en suspensión extrayendo muestras de tamaño 2 sin reposición y sin tener en cuenta el orden de los elementos de la muestra. Considerando como posibles estimadores para este parámetro las medias aritmética, armónica y geométrica se pide*

- *Especificar el espacio muestral, las probabilidades asociadas a cada muestra y la distribución en el muestreo de los tres estimadores.*
- *Identificar los estimadores insesgados y calcular el error cuadrático medio asociado a cada uno y justificar qué estimador es el más adecuado para estimar el nivel medio de concentración de sustancias tóxicas en suspensión.*

Ejemplo 19 *Crespo 1 Para la población $A = \{A_1, A_2, A_3, A_4, A_5\}$ consideramos el siguiente proceso de selección de muestras de tamaño 3. De una urna con 3 bolas numeradas del 1 al 3 se extraen al azar y sin reposición 2 bolas. A continuación de otra urna con 2 bolas numeradas con el 4 y el 5 se extrae una bola. Se pide:*

- *Espacio muestral asociado a este experimento de muestreo y probabilidades de las muestras. Consideremos el estimador por analogía $\widehat{\theta}$ "suma de los subíndices de unidades de las muestras para estimar la característica poblacional" para estimar la característica poblacional θ "suma de los subíndices de las unidades de la población". Calcular la precisión del estimador.*
- *Se considera el estimador por analogía $\widehat{\overline{\theta}}$ "media de los subíndices de unidades de las muestras" para estimar la característica poblacional $\overline{\theta}$ "media de los subíndices de las unidades de la población". Calcular la precisión de este estimador, ¿qué estimador es mejor?*

1.3. Estimación por intervalos de confianza

Hasta ahora hemos utilizado la *estimación puntual* para estimar el valor de un parámetro desconocido. Es decir, hemos estimado el valor del parámetro poblacional θ mediante un único valor obtenido de los datos de la muestra. En esta sección analizaremos otro tipo de estimación que denominaremos *estimación por intervalos de confianza*. Una estimación por intervalos de confianza es una regla o procedimiento que nos permite calcular, basados en los datos de la muestra,

un intervalo dentro del cual se espera que esté el parámetro poblacional con una determinada probabilidad.

Para hallar los intervalos de confianza de un parámetro poblacional θ se partirá de un estimador puntual $\widehat{\theta}$ de dicho parámetro, generalmente insesgado; a partir de él se construirá el intervalo $(\widehat{\theta}-\epsilon, \widehat{\theta}+\epsilon)$, de amplitud 2ϵ, imponiendo que la probabilidad de que el parámetro poblacional desconocido θ se encuentre en dicho intervalo sea $1-\alpha$ con $0 < \alpha < 1$.

$$P[\widehat{\theta} - \epsilon \leq \theta \leq \widehat{\theta} + \epsilon] = 1 - \alpha. \tag{1.3}$$

Al término ϵ se le conoce como *término de error* para dicho intervalo de confianza y nos indica el margen de error o precisión de la estimación de θ. A la diferencia $1-\alpha$ se le denomina *nivel de confianza*.

Los valores más utilizados de α son 0.1, 0.05 y 0.01 lo que corresponde con niveles de confianza (o coeficiente de confianza) del 90 %, 95 % y 99 % respectivamente.

El término nivel de confianza, por ejemplo 89 %, se refiere a que si consideramos un número elevado de muestras y para cada una de ellas construimos dicho intervalo de confianza para un parámetro poblacional desconocido θ, tendremos que θ se encuentra en al menos el 89 % de los intervalos construidos.

1.3.1. Intervalos de confianza para estimadores insesgados

Se trata de estimar el parámetro poblacional θ mediante un intervalo de confianza basado en el estimador $\widehat{\theta}$ insesgado para θ $(E(\widehat{\theta})) = \theta)$. Para estimadores insesgados es necesario distinguir entre el caso en que la distribución del estimador pueda aproximarse mediante una distribución normal y en el caso en que dicha distribución no puede asegurarse que sea normal.

El estimador $\widehat{\theta}$ tiene una distribución normal

Supongamos que

$$\widehat{\theta} \longrightarrow N(E(\widehat{\theta}), Var(\widehat{\theta})).$$

En este caso, se tiene que

$$Z = \frac{\widehat{\theta} - E[\widehat{\theta}]}{\sigma(\widehat{\theta})} = \frac{\widehat{\theta} - \theta}{\sigma(\widehat{\theta})},$$

se distribuye como una $N(0,1)$. La varianza del estimador puede conocerse de algún estudio anterior o de una muestra piloto. Por lo tanto,

$$P[-z_{1-\alpha/2} \leq Z \leq z_{1-\alpha/2}] = 1 - \alpha,$$

donde $z_{1-\alpha/2} = F^{-1}_{N(0,1)}(1-\alpha/2)$ donde z_u denota el cuantil u de la distribución normal estándar, es decir, el valor que verifica

$$P(Z \leq z_u) = u,$$

o dicho de otra manera, el valor que deja a su izquierda un área igual a u debajo de la curva de la densidad normal estándar. Por lo tanto,

$$P\left[-z_{1-\alpha/2} \leq \frac{\widehat{\theta} - E[\widehat{\theta}]}{\sigma(\widehat{\theta})} \leq z_{1-\alpha/2}\right] = 1-\alpha,$$

y por ser $\widehat{\theta}$ un estimador insesgado del parámetro poblacional θ, se tiene que

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

En este caso, un intervalo de confianza para un coeficiente de confianza dado $1-\alpha$ viene dado por

$$[\widehat{\theta} - z_{1-\alpha/2}\sigma(\widehat{\theta}), \widehat{\theta} + z_{1-\alpha/2}\sigma(\widehat{\theta})].$$

En particular usaremos de manera repetida los cuantiles siguientes: $z_{0,95}$, $z_{0,975}$ y $z_{0,995}$. Estos cuantiles se utilizarán para calcular intervalos de confianza con un nivel de confianza al 90 %, al 95 % y al 99 % respectivamente. Mirando en la tabla de la normal, encontramos que $z_{0,95} = 1{,}64$, $z_{0,975} = 1{,}96$ y $z_{0,995} = 2{,}56$. Es decir, los intervalos

$$[\widehat{\theta} - 1{,}64\sigma(\widehat{\theta}), \widehat{\theta} + 1{,}64\sigma(\widehat{\theta})], \quad [\widehat{\theta} - 1{,}96\sigma(\widehat{\theta}), \widehat{\theta} + 1{,}96\sigma(\widehat{\theta})]$$
$$[\widehat{\theta} - 2{,}56\sigma(\widehat{\theta}), \widehat{\theta} + 2{,}56\sigma(\widehat{\theta})],$$

representan los intervalos de confianza con niveles de confianza de 90 %, 95 % y 99 % respectivamente para el estimador $\widehat{\theta}$.

Como hemos dicho, el intervalo de confianza al $100(1-\alpha)$ % para θ es

$$[\widehat{\theta} - z_{1-\alpha/2}\sigma(\widehat{\theta}), \widehat{\theta} + z_{1-\alpha/2}\sigma(\widehat{\theta})],$$

o bien se suele expresar como

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

Al término $z_{1-\alpha/2}\sigma(\widehat{\theta})$ se le denomina *término de error*. Notar que los extremos de este intervalo dependen de la muestra escogida.

Comentarios importantes:

- La construcción del intervalo de confianza está basada en la hipótesis de que la distribución de $\widehat{\theta}$ es normal. Sin embargo, si la distribución de $\widehat{\theta}$ no es normal, el intervalo no es válido, es decir, que no podemos garantizar que la confianza especificada sea cierta. Sin embargo, si la muestra es grande, posibilita que los intervalos sean aproximadamente válidos (la confianza no será exacta pero casi). ¿A partir de cuántas observaciones consideraremos una muestra como grande? Se suele considerar $n \geq 30$.

- Es usual que no se conozca el valor de $\sigma(\widehat{\theta})$ porque en sus cálculos intervienen datos poblacionales no conocidos, pero se utiliza en su lugar su estimación $\widehat{\sigma}(\widehat{\theta})$ que depende únicamente de datos muestrales.

Ejemplo 20 *Para el ejemplo 17, construir un intervalo de confianza al* 95 % *para el estimador media aritmética suponiendo que dicho estimador sigue una distribución normal.*

Solución. *En este caso, la esperanza del estimador* $E[\widehat{\bar{X}}]$ *y la varianza de dicho estimador* $\sigma^2(\widehat{\bar{X}})$ *vienen dados por*

$$E[\widehat{\bar{X}}] = 2, \quad \sigma^2(\widehat{\bar{X}}) = \frac{1}{3}, \quad \sigma(\widehat{\bar{X}}) = 0{,}5774.$$

Luego, el intervalo de confianza para el estimador $\widehat{X}$ *basado en el estimador* $\widehat{\bar{X}}$ *viene dado por*

$$(\widehat{\bar{X}} - z_{1-\alpha/2}\sigma(\widehat{\bar{X}}), \widehat{\bar{X}} + z_{1-\alpha/2}\sigma(\widehat{\bar{X}})),$$

luego el intervalo de confianza resulta ser

$$(\widehat{\bar{X}} - 1{,}96 \cdot 0{,}5774, \widehat{\bar{X}} + 1{,}96 \cdot 0{,}5774).$$

Para cada una de las muestras obtenidas se obtienen diferentes intervalos de confianza. Por ejemplo, consideramos la muestra formada por los elementos (u_1, u_1)*, de manera que para esa muestra particular el estimador media muestral es igual a* $\widehat{\bar{X}} = 1$ *y el intervalo de confianza es*

$$(1 - 1{,}96 \cdot 0{,}5774, 1 + 1{,}96 \cdot 0{,}5774) = (-0{,}1317, 2{,}1317).$$

Para la muestra formada por los elementos (u_1, u_2)*, el intervalo de confianza es*

$$(1{,}5 - 1{,}96 \cdot 0{,}5774, 1{,}5 + 1{,}96 \cdot 0{,}5774) = (0{,}3683, 2{,}6317).$$

Si realmente es dudoso que el estimador $\widehat{\theta}$ siga una distribución normal, puede utilizarse la distribución t de Student con n-1 grados de libertad para calcular el intervalo de confianza para θ. En este caso el intervalo de confianza viene dado por

$$[\widehat{\theta} - t_{n-1,1-\alpha/2}\sigma(\widehat{\theta}), \widehat{\theta} + t_{n-1,1-\alpha/2}\sigma(\widehat{\theta})],$$

donde $t_{1-\alpha/2} = F^{-1}_{t_{n-1}}(1 - \alpha/2)$.

Ejemplo 21 *Para el ejemplo 17, construir un intervalo de confianza al* 95 % *para el estimador media aritmética de la media poblacional suponiendo que dicho estimador no sigue una distribución normal.*

Solución. *En este caso, el intervalo de confianza viene dado por*

$$(\widehat{\bar{X}} - t_{n-1,1-\alpha/2}\sigma(\widehat{\bar{X}}), \widehat{\bar{X}} + t_{n-1,1-\alpha/2}\sigma(\widehat{\bar{X}})),$$

luego el intervalo de confianza resulta ser

$$(\widehat{\bar{X}} - 12{,}7062{\cdot}0{,}5774, \widehat{\bar{X}} + 12{,}7062{\cdot}0{,}5774).$$

Para cada una de las muestras obtenidas se obtienen diferentes intervalos de confianza. Por ejemplo, consideramos la muestra formada por los elementos (u_1, u_1)*, de manera que para esa muestra particular el estimador media muestral es igual a* $\widehat{\bar{X}} = 1$ *y el intervalo de confianza es*

$$(1 - 12{,}7062{\cdot}0{,}5774, 1 + 12{,}7062{\cdot}0{,}5774) = (-6{,}3366, 8{,}3366).$$

Para la muestra formada por los elementos (u_1, u_2)*, el intervalo de confianza es*

$$(1{,}5 - 12{,}7062{\cdot}0{,}5774, 1{,}5 + 12{,}7062{\cdot}0{,}5774) = (-5{,}8366, 8{,}8366).$$

El estimador $\widehat{\theta}$ no es normal

En el caso en el que la distribución del estimador $\widehat{\theta}$ no siga una distribución normal, puede encontrarse un intervalo de confianza utilizando la desigualdad de Chebyshev. Esta desigualdad afirma que para cualquier variable aleatoria X con media y varianza finita, se verifica la siguiente propiedad

$$P\{|X - E(X)| \leq K\sigma(X)\} \geq 1 - \frac{1}{K^2}, \quad \forall K \geq 0.$$

Si tomamos como variable aleatoria el estimador $\widehat{\theta}$, y si además $\widehat{\theta}$ es un estimador insesgado para el parámetro poblacional θ, se tiene que

$$P\left[|\widehat{\theta} - \theta| \leq K\sqrt{V(\widehat{\theta})}\right] = P\left[\widehat{\theta} - K\sqrt{V(\widehat{\theta})} \leq \theta \leq \widehat{\theta} + K\sqrt{V(\widehat{\theta})}\right] \geq 1 - \frac{1}{K^2}$$

Para un nivel de confianza $1 - \alpha$ deseado, basta con considerar $K = \sqrt{1/\alpha}$ y por lo tanto el intervalo de confianza viene dado por

$$\left(\widehat{\theta} - \frac{\sigma(\widehat{\theta})}{\sqrt{\alpha}}, \widehat{\theta} + \frac{\sigma(\widehat{\theta})}{\sqrt{\alpha}}\right)$$

Este intervalo suele ser más ancho que el obtenido cuando la distribución de $\widehat{\theta}$ es normal. A medida que $\widehat{\theta}$ se aleja más de la normalidad, la anchura de este intervalo es mucho mayor respecto del obtenido para normalidad. Ya sabemos que una estimación por intervalos es tanto mejor cuanto más reducido sea el intervalo de confianza correspondiente

Ejemplo 22 *Para el ejemplo 17, construir un intervalo de confianza al* 95 % *para el estimador media aritmética utilizando la desigualdad de Chebyshev.*

Solución. *En este caso particular, el intervalo de confianza al* 95 % *viene dado por*

$$\left(\widehat{\theta} - \frac{\sigma(\widehat{\theta})}{\sqrt{\alpha}}, \widehat{\theta} + \frac{\sigma(\widehat{\theta})}{\sqrt{\alpha}}\right),$$

luego utilizando los datos del problema, el intervalo de confianza es igual a

$$\left(\widehat{\theta} - \frac{0{,}5774}{\sqrt{0{,}05}}, \widehat{\theta} + \frac{0{,}5774}{\sqrt{0{,}05}}\right).$$

Si utilizamos la muestra (u_1, u_1)*, el intervalo de confianza resulta ser*

$$\left(1 - \frac{0{,}5774}{\sqrt{0{,}05}}, 1 + \frac{0{,}5774}{\sqrt{0{,}05}}\right) = (-1{,}5822, 3{,}5822).$$

1.3.2. Intervalos de confianza en estimadores sesgados

Ahora vamos a considerar el caso en que el estimador $\widehat{\theta}$ es sesgado para θ, es decir, existe un sesgo $B(\widehat{\theta}) = E(\widehat{\theta}) - \theta$. Ahora partimos de que, por el Teorema Central del Límite, y para un tamaño de muestra suficientemente grande, se tiene que

$$\frac{\widehat{\theta} - E(\widehat{\theta})}{\sigma(\widehat{\theta})} \to N(0, 1).$$

Por lo tanto, para un nivel α de significación podemos calcular

$$z_{\alpha/2} = F^{-1}_{N(0,1)}(1 - \alpha/2),$$

tal que

$$P\left\{-z_{\alpha/2} \leq \frac{\widehat{\theta} - E(\widehat{\theta})}{\sigma(\widehat{\theta})} \leq z_{\alpha/2}\right\} = 1 - \alpha.$$

Ahora, como $E(\widehat{\theta}) = B(\widehat{\theta}) + \theta$, se tiene que

$$P\left\{-z_{\alpha/2} \leq \frac{\widehat{\theta} - B(\widehat{\theta}) - \theta}{\sigma(\widehat{\theta})} \leq z_{\alpha/2}\right\} = 1 - \alpha,$$

de esta manera resulta que

$$P\left\{-z_{\alpha/2} \leq \frac{\widehat{\theta} - \theta}{\sigma(\widehat{\theta})} - \frac{B(\widehat{\theta})}{\sigma(\widehat{\theta})} \leq z_{\alpha/2}\right\} = 1 - \alpha.$$

En caso de que

$$\left|\frac{B(\widehat{\theta})}{\sigma(\widehat{\theta})}\right| < \frac{1}{10}, \tag{1.4}$$

ya sabemos que la influencia del sesgo es despreciable, con lo que el intervalo de confianza es el mismo que para el caso del estimador insesgado. Pero si no se cumple (1.4), el sesgo de $\widehat{\theta}$ influye en el intervalo de confianza. Operando en los conjuntos, se tiene que

$$\begin{aligned} &\{-\sigma(\widehat{\theta})z_{\alpha/2} \leq \widehat{\theta} - B(\widehat{\theta}) - \theta \leq \sigma(\widehat{\theta})z_{\alpha/2}\} \\ &= \{-\sigma(\widehat{\theta})z_{\alpha/2} - \widehat{\theta} + B(\widehat{\theta}) \leq -\theta \leq \sigma(\widehat{\theta})z_{\alpha/2} - \widehat{\theta} + B(\widehat{\theta})\}, \end{aligned}$$

de manera que el conjunto anterior puede expresarse como

$$\{\widehat{\theta} - B(\widehat{\theta}) - z_{\alpha/2}\sigma(\widehat{\theta}) \leq \theta \leq \widehat{\theta} - B(\widehat{\theta}) + z_{\alpha/2}\sigma(\widehat{\theta})\},$$

con lo que el intervalo de confianza para θ basado en el estimador $\widehat{\theta}$ en presencia del sesgo no despreciable $B(\widehat{\theta}) = E(\widehat{\theta}) - \theta$ es el siguiente

$$[\widehat{\theta} - B(\widehat{\theta}) - z_{\alpha/2}\sigma(\widehat{\theta}), \widehat{\theta} - B(\widehat{\theta}) + z_{\alpha/2}\sigma(\widehat{\theta})]$$

El intervalo de confianza para θ basado en el estimador $\widehat{\theta}$ en presencia del sesgo no despreciable es el siguiente

$$[\widehat{\theta} - z_{1-\alpha/2}\sigma(\widehat{\theta}) - |B(\widehat{\theta})|, \widehat{\theta} + z_{1-\alpha/2}\sigma(\widehat{\theta}) - |B(\widehat{\theta})|].$$

1.4. Determinación de estimadores insesgados

Supongamos que tenemos definida una característica X en una población $\Omega = \{u_1, u_2, \ldots, u_N\}$ que toma el valor numérico X_i sobre la unidad u_i, $i = 1, 2, \ldots, N$ dando lugar al conjunto de valores $\{X_1, X_2, \ldots, X_N\}$. Consideramos ahora una cierta función θ de los N valores X_i que suele denominarse *parámetro poblacional*. Seleccionamos una muestra s de unidades $s = \{u_1, u_2, \ldots, u_n\}$ mediante un procedimiento de muestreo dado y consideramos los valores $\{X_1, X_2, \ldots, X_n\}$ que toma la característica X en estudio sobre los elementos de la muestra. A partir de estos valores estimamos puntualmente el parámetro poblacional θ mediante un estimador $\widehat{\theta}$, con $\widehat{\theta}(s)$ basada en los valores X_i $i = 1, 2, \ldots, n$ que toma la característica X sobre las unidades de la muestra s.

Los parámetros poblacionales que trataremos en este curso son los siguientes:

- Total poblacional

$$X = \theta(X_1, X_2, \ldots, X_N) = \sum_{i=1}^{N} X_i$$

- Media poblacional

$$\bar{X} = \theta(X_1, X_2, \ldots, X_N) = \sum_{i=1}^{N} X_i/N$$

- Total de clase

$$A = \theta(A_1, A_2, \ldots, A_N) = \sum_{i=1}^{N} A_i$$

- Proporción de clase

$$P = \theta(A_1, A_2, \ldots, A_N) = \sum_{i=1}^{N} A_i/N$$

Vemos que en general, un parámetro poblacional θ puede expresarse como una suma de elementos Y_i

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

donde
$Y_i = X_i$ para el total poblacional
$Y_i = X_i/N$ para la media poblacional
$Y_i = A_i$ para el total de clase
$Y_i = A_i/N$ para la proporción de clase.

Ahora bien, ¿Qué forma debe de tener el estimador? En general, se ha demostrado que las mejores propiedades de estos estimadores suelen presentarlas los estimadores lineales insesgados en los valores de la muestra s es decir,

$$\widehat{\theta} = \sum_{i=1}^{n} \alpha_i Y_i, \tag{1.5}$$

y dependiendo de que el muestreo a realizar sea con reposición o sin reposición obtendremos unos valores determinados α_i. A los valores α_i se les denomina *pesos* o *factores de elevación*.

Consideremos una población de tamaño N y una muestra s de tamaño n. En este esquema de selección, cada unidad u_i de la población puede pertenecer a la muestra una sola vez. Ahora bien, para cada $i = 1, 2, \ldots, N$ consideremos la variable aleatoria e_i definida de la siguiente forma

$$e_i = \begin{cases} 1 & \text{si } u_i \in s \text{ con probabilidad } \pi_i \\ 0 & \text{si } u_i \notin s \text{ con probabilidad } 1 - \pi_i \end{cases}$$

De esta forma estamos considerando una variable aleatoria definida en función de la probabilidad π_i de que la i-ésima unidad de la población pertenezca a la muestra. La varianza y la esperanza de e_i viene dada por

$$E[e_i] = \pi_i, \quad Var[e_i] = \pi_i(1 - \pi_i).$$

Utilizando el estimador $\widehat{\theta}$ dado en (1.5), para que $\widehat{\theta}$ sea un estimador insesgado de θ, se tiene que cumplir que

$$E\left(\sum_{i=1}^{n} \alpha_i Y_i\right) = E\left(\sum_{i=1}^{N} \alpha_i Y_i e_i\right) = \sum_{i=1}^{N} \alpha_i Y_i \pi_i,$$

y por tanto los valores α_i que determinan la expresión del estimador se obtienen de la siguiente igualdad

$$\sum_{i=1}^{N} \alpha_i Y_i \pi_i = \sum_{i=1}^{N} Y_i \Longleftrightarrow 1 = \alpha_i \pi_i \Longleftrightarrow \alpha_i = \frac{1}{\pi_i}, \quad \forall i,$$

donde π_i es la probabilidad de que un individuo i esté en la muestra de tamaño n y por tanto la expresión del estimador lineal e insesgado para θ es

$$\widehat{\theta}_{HT} = \sum_{i=1}^{n} \frac{1}{\pi_i} Y_i. \tag{1.6}$$

Este tipo de estimadores se conoce como *estimadores de Horvitz-Thompson* y fueron propuestos en 1952. Aplicamos estos estimadores de Horvitz-Thompson a los parámetros poblacionales θ

$$\theta = \sum_{i=1}^{N} X_i \Rightarrow \widehat{\theta}_{HT} = \sum_{i=1}^{n} \frac{X_i}{\pi_i}$$

$$\theta = \sum_{i=1}^{N} \frac{X_i}{N} \Rightarrow \widehat{\theta}_{HT} = \sum_{i=1}^{n} \frac{X_i}{N\pi_i}$$

$$\theta = \sum_{i=1}^{N} A_i \Rightarrow \widehat{\theta}_{HT} = \sum_{i=1}^{n} \frac{A_i}{\pi_i}$$

$$\theta = \sum_{i=1}^{N} \frac{A_i}{N} \Rightarrow \widehat{\theta}_{HT} = \sum_{i=1}^{n} \frac{A_i}{N\pi_i}$$

Ahora, para cada $i, j = 1, 2, \ldots, N$ con $i \neq j$ consideramos la variable aleatoria producto $e_i e_j$, es decir,

$$e_i e_j = \begin{cases} 1 & \text{si } (u_i, u_j) \in s \text{ con probabilidad } \pi_{ij} \\ 0 & \text{si } (u_i, u_j) \notin s \text{ con probabilidad } 1 - \pi_{ij} \end{cases}$$

de manera que

$$E[e_i e_j] = \pi_{ij}, \quad Cov[e_i e_j] = \pi_{ij} - \pi_i \pi_j.$$

En este caso, vamos a obtener la varianza del estimador dado en (1.6).

$$\begin{aligned} Var(\widehat{\theta}_{HT}) &= V\left(\sum_{i=1}^{n} \frac{1}{\pi_i} Y_i\right) = V\left(\sum_{i=1}^{N} \frac{1}{\pi_i} Y_i e_i\right) \\ &= \sum_{i=1}^{N} V\left(\frac{Y_i}{\pi_i} e_i\right) + 2\sum_{i=1}^{N}\sum_{j>i}^{N} Cov\left(\frac{Y_i}{\pi_i} e_i, \frac{Y_j}{\pi_j} e_j\right) \\ &= \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i^2} V(e_i) + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j} Cov(e_i, e_j) \\ &= \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i}(1 - \pi_i) + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}(\pi_{ij} - \pi_i \pi_j) \end{aligned}$$

Como la expresión de la varianza del estimador de Horvitz-Thompson extiende sus índices hasta el valor N, y dados que los únicos datos serán generalmente los muestrales, será necesario estimar dicha varianza de forma que dependa únicamente de los valores muestrales (índices de la suma hasta n). La raíz cuadrada de esta estimación de la varianza se utiliza como error de muestreo del estimador de Horvitz-Thompson. Un estimador insesgado para $V(\widehat{\theta}_{HT})$ viene dada por

$$\widehat{V}(\widehat{\theta}_{HT}) = \sum_{i=1}^{n} \frac{Y_i^2}{\pi_i^2}(1-\pi_i) + 2\sum_{i=1}^{n}\sum_{j>i}^{n} \frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}} \tag{1.7}$$

Para comprobar la insesgadez del estimador dado en (1.7), es necesario probar que

$$E[\widehat{V}(\widehat{\theta}_{HT})] = V(\widehat{\theta}_{HT}).$$

Para ello, se tiene que

$$\begin{aligned} E[\widehat{V}(\widehat{\theta}_{HT})] &= E\left(\sum_{i=1}^{n} \frac{Y_i^2}{\pi_i^2}(1-\pi_i)\right) + 2E\left(\sum_{i=1}^{n}\sum_{j>i}^{n} \frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}\right) \\ &= E\left(\sum_{i=1}^{N} \frac{Y_i^2}{\pi_i^2}(1-\pi_i)e_i\right) + 2E\left(\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}\frac{(\pi_{ij}-\pi_i\pi_j)}{\pi_{ij}}e_ie_j\right) \\ &= \sum_{i=1}^{N} \frac{Y_i^2}{\pi_i}(1-\pi_i) + 2\sum_{i=1}^{N}\sum_{j>i}^{N} \frac{Y_i}{\pi_i}\frac{Y_j}{\pi_j}(\pi_{ij}-\pi_i\pi_j). \end{aligned}$$

Luego $\widehat{V}(\widehat{\theta}_{HT})$ es un estimador insesgado para $V(\widehat{\theta}_{HT})$.

Ejemplo 23 *Consideramos un procedimiento de muestreo que consiste en obtener 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 la muestra no interviene, es decir, muestras con los mismos elementos colocados en distinto orden se consideran iguales. Calcular el estimador de Horvitz-Thompson para los parámetros poblacionales total poblacional, media poblacional, total de clase y proporción de clase.*