CAPÍTULO 5: NOCIONES DE MUESTREO

Por: Enrique Wabo

1 GENERALIDADES

1.1 INTRODUCCIÓN

La Teoría del Muestreo tiene por objeto el desarrollo de procedimientos que permitan obtener información confiable sobre un conjunto de objetos, observando sólo una parte de esos objetos. Por ejemplo:

a) pretendemos conocer una característica de los alumnos de esta Facultad, y para ello sólo observamos una fracción de todos los alumnos;

b) pretendemos conocer una característica del ganado existente en una provincia y para ello sólo observamos una fracción de todos los animales.

c) Y si pretendemos conocer el volumen de madera de los árboles existentes en un rodal, sólo observamos una fracción de todos los árboles.

1.2 TÉRMINOS BÁSICOS

Cada objeto en el cual tenemos interés recibe el nombre de **Elemento**; el conjunto de todos los elementos en que tenemos interés recibe el nombre de **Universo**. Por lo tanto, cada objeto en el cual tenemos interés es un elemento del Universo.

La información que pretendemos corresponde a una propiedad de esos elementos, la que expresamos mediante **Números**. Los números que expresan la propiedad de interés puede ser un valor único para todos los elementos, en cuyo caso se dice que es una **Constante**.

De no ser así, los números van a variar de un elemento a otro, aunque algunos pueden repetirse; estos números toman la forma de una variable, que llamamos **Variable de Interés**, que generalmente es indicada con la letra Y.

Así, cada elemento del Universo está asociado con un valor de esa variable Y, de manera que si el universo posee N elementos habrá N valores Y disponibles. El conjunto de los N valores correspondientes a los N elementos del universo recibe el nombre de **Población**. En el Cuadro 1 se representa la relación entre elemento, universo y población.

Cuadro 1

UNIVERSO	*POBLACIÓN*
Elemento 1	Valor de la variable en el elemento 1
Elemento 2	Valor de la variable en el elemento 2
Elemento 3	Valor de la variable en el elemento 3
...	...
Elemento N	Valor de la variable en el elemento N

Puede decirse que disponemos de un conjunto de pares ordenados, donde el primer componente del par es el elemento y el segundo componente es el valor de la variable Y de interés asociado a ese elemento:

$$\left\{ \text{(Elemento, Y)} \right\}$$

La información buscada acerca de Y puede obtenerse por dos caminos. Uno de ellos es mediante el relevamiento al 100 %, que es cuando la información proviene de observar todos los elementos del Universo.

El otro, es mediante la observación de una parte del universo denominada **Muestra**; la operación por la cual se obtiene la muestra se denomina **Muestreo**. Es decir, que de todos los pares ordenados existentes en el universo sólo una parte va a estar presente en la muestra. Por eso, el término Muestra representa tanto a la muestra de elementos como a la muestra de valores.

Algunos especialistas rechazan el uso del término Universo y en su lugar usan la expresión **Población de Elementos** y lo que habíamos llamado simplemente Población ahora es denominada población de valores. Las equivalencias son:

Universo (de elementos) = Población de elementos

Población (de valores) = Población de valores

Muestra de elementos o valores (para ambos casos)

Un Universo y la Población de valores asociada deben estar acotados en el tiempo. Por ejemplo, el universo de alumnos de este año contendrá elementos diferentes al que tendrá dentro de cinco años, y es de esperar que también varíen los valores de la variable de interés. En otras palabras, la Población en el Tiempo I será diferente a la Población en el Tiempo II.

La cantidad de elementos del universo y, por extensión, la cantidad de valores de la variable de interés Y, se indican con la letra mayúscula N; el tamaño de la muestra se indica con la misma letra en minúscula: n. La relación n/N se conoce como Intensidad de Muestreo, y representa la proporción del universo que participó en la muestra; la expresión (1-n/N) es la intensidad de no muestreo, que representa la proporción del universo que no participó en la muestra. Así:

N = tamaño de la población

n = tamaño de la muestra

n/N = intensidad de muestreo

1– n/N = intensidad de no muestreo

1.3 CONSTANTES DE LA POBLACIÓN

Supongamos que contamos con los N valores de una población, que aplicamos alguna fórmula a esos valores, y que ese cálculo lo repetimos una y otra vez.

Cada vez que repetimos el cálculo obtenemos como resultado el mismo valor que obtuvimos en el cálculo anterior; es decir, que ese valor es lo que hemos definido como una Constante. Cualquier valor que presenta este comportamiento recibe el nombre de **Constante Poblacional**; algunos lo llaman **Parámetro**. Así, la constante poblacional es un valor constante para una población, en un momento dado.

Por ejemplo, supongamos que determinamos el promedio de todos los valores de la variable Y de interés y que obtenemos el valor 24; si repetimos el promedio volvemos a obtener 24 y lo mismo ocurrirá cada vez que repitamos el cálculo. En consecuencia, la media de los N valores Y es una constante poblacional.

Las constantes poblacionales en las que vamos a tener interés son básicamente tres:

a) el total de la población:

$$Y = \sum_{1}^{N} y_i = \overline{Y} \times N \qquad (1)$$

b) el promedio por elemento de una población, o promedio de Y por elemento:

$$\overline{Y} = \frac{\sum_{1}^{N} y_i}{N} = \frac{Y}{N} \qquad (2)$$

c) la razón entre dos variables de un universo, o promedio de Y por unidad de X.

$$R = \frac{\sum_{1}^{N} y}{\sum_{1}^{N} x} = \frac{\overline{Y}}{\overline{X}} \qquad (3)$$

Como ejemplo de un valor total tenemos la producción total de un cultivo de trigo o el volumen total presente en una plantación forestal.

Como ejemplo de una media por elemento tenemos la altura media por árbol o el peso medio por novillo.

Como ejemplo de media de Y por unidad de X tenemos la proporción de área basal de una especie en relación al área basal de todas las especies presentes en el rodal.

Por el momento nos vamos a concentrar en valores totales de la población y la media por elemento de la población; la razón se verá más adelante.

1.4 RAZONES DEL MUESTREO

Es evidente que nuestro interés es conocer una o más constantes poblacionales de una variable de interés. Por lo tanto, si pudiésemos observar los N valores de la población, podríamos determinar esas constantes sin inconvenientes y el problema estaría resuelto. Sin embargo, hay situaciones en las que el relevamiento al 100 % no se puede llevar a cabo, cuyas razones veremos en el siguiente punto.

Las principales razones para llevar a cabo un muestreo son:

Costos:

El relevamiento al 100 % es demasiado costoso y no se puede financiar. Por ejemplo, queremos conocer el peso medio por novillo en la provincia de Buenos Aires; pesar cada animal existente en la provincia sería una tarea excesivamente costosa en relación con la información que se obtendría.

Tiempo disponible:

El relevamiento al 100% demandaría un tiempo excesivo. Ello impide obtener información a corto plazo; o puede ocurrir que, con el paso del tiempo, la población final difiera de la original en la que se tenía interés.

Población infinita:

La población es infinita y cualquiera fuese la cantidad de elementos observados siempre será una muestra. Es el caso de una máquina que fabrica un producto; cualquiera fuese la cantidad de objetos medidos siempre habrá más objetos disponibles.

Impedimentos técnicos:

Puede ocurrir que sea técnicamente imposible hacer un relevamiento al 100 %. Es el caso de un biólogo que quiere determinar el número de lombrices por unidad de superficie en una propiedad; para ello debería dar vuelta la tierra de toda la propiedad, lo que carece de sentido. Otro caso es cuando los datos provienen de un método destructivo.

La solución a todos estos inconvenientes es trabajar con sólo una parte de la población: una Muestra. La muestra puede seleccionarse con distintos propósitos, que podemos reunir en tres grupos, que son:

- Cuando el muestreo es hecho para investigación; el resultado final suele ser la evaluación de una hipótesis. Los métodos de muestreo están incluidos dentro del Diseño de Experimentos.

- Cuando el muestreo es hecho para determinar valores totales de un ítem. Los métodos de muestreo están incluidos dentro del Diseño de Inventarios.

- Cuando el muestreo se hace para determinar si cierta operación se está llevando a cabo correctamente. Los métodos de muestreo están incluidos dentro del Control de Calidad.

Podemos preguntar si un relevamiento al 100 % es mejor que un relevamiento por muestreo. La respuesta es no. En primer lugar, debemos tener en cuenta la relación costo/beneficio; es decir, plantear si la mayor inversión en un relevamiento al 100% compensa lo que perderíamos por no usar toda la información.

En segundo lugar, hay que tener en cuenta que las mediciones requeridas estarán sujetas a errores de medición. Trabajar con una muestra nos permite concentrarnos en las tareas y reducir la posibilidad de esos errores: medir menos pero mejor.

Otra pregunta que podemos hacer es si podemos conocer el valor de las constantes poblacionales a partir de una muestra. La respuesta es no. Por un lado, es muy difícil que ello ocurra; por otro lado, si ocurriera no lo sabríamos. Por eso, nuestro objetivo es alcanzar un valor lo más cercano posible al valor de la constante pretendida. Por eso cambiamos la pregunta anterior por esta otra: ¿nos podemos acercar al valor de la constante poblacional que nos interesa? La respuesta es sí; mediante el uso de unas funciones especiales llamadas Estimadores que nos permiten obtener una estimación de la constante de interés.

Y aquí surge otra pregunta: el valor que hemos obtenido por aplicación del estimador, ¿qué tan cerca está de la constante desconocida? La respuesta es que no lo sabemos; y que sólo podemos responder en términos de probabilidades: hay una probabilidad del tanto % de que la media verdadera esté a una cierta distancia de la media estimada en la muestra.

1.5 ESTIMADORES Y ESTIMACIONES

Así como podemos hacer determinaciones con los valores de una población, también podemos hacerlo con los valores de una muestra; estas funciones reciben el nombre de Estadístico o Estadística.

Definimos como Estadística, a cualquier función de los datos de la muestra que no depende de constantes desconocidas. Así, Σy, $\Sigma \log(y)$ y $(N/n)\Sigma y$, son ejemplos de estadísticos. Si bien la cantidad de estadísticos disponibles es prácticamente infinita, sólo algunos nos son realmente útiles.

Nos interesa en particular un subconjunto de Estadísticos, que son aquellos que nos permiten obtener una buena estimación de la constante poblacional buscada, y que reciben el nombre de **Estimadores**. Un Estimador es una función de los datos de una muestra que no depende de constantes desconocidas, que nos permite una buena estimación de una constante poblacional.

Cuando el estimador de una constante poblacional se aplica a los datos de una muestra específica de las muchas disponibles, el valor obtenido es una **Estimación** de esa constante, la que corresponde a los datos de esa muestra.

Así, mientras el Estimador es una función única para cualquier muestra, la Estimación es el valor obtenido por aplicación del estimador en una muestra específica de todas las muestras posibles.
En este punto vamos a incorporar otra constante poblacional, que si bien no suele ser un objetivo específico del muestreo, nos será útil más adelante. Esta constante poblacional es la **Varianza por Elemento** de la variable Y:

$$\sigma^2 = \frac{\sum_{1}^{N} (y - \overline{Y})^2}{N} \qquad (4)$$

siendo $\overline{Y}$ la media de la población.

Resumiendo, las constantes poblacionales en que vamos a concentrar nuestro interés de aquí en más son: a) el total de Y en la población; b) la media de Y por elemento en la población; y c) la varianza de Y por elemento.

Existen diferentes métodos para definir estimadores: el método de los Momentos, el método de Máxima Verosimilitud, el método de Cuadrados Mínimos y otros.

Sin entrar en detalles, nos ocuparemos de los estimadores provistos por el Método de los Momentos, que son la copia muestral de la constante poblacional. Para indicar un estimador de una constante poblacional se utiliza el símbolo de la constante con un sombrerito encima; por ejemplo, si la constante es θ, la expresión $\hat{\theta}$ representa su estimador. Indicado de esta forma, el estimador no tiene una forma matemática específica; por ejemplo, tomemos por caso la media por elemento de la población $\overline{Y}$; para indicar un estimador ponemos $\hat{\overline{Y}}$, que puede ser la media aritmética de los n valores de la muestra, o la media geométrica de los mismos valores o la mediana; cada uno posee su forma algebraica específica.

En el Cuadro 2 se indican las tres constantes poblacionales que hemos tomado en cuenta y sus correspondientes estimadores "copias":

Cuadro 2

Constantes Poblacionales	Estimadores "COPIA" de la Muestra
$\overline{Y} = \frac{\sum_{1}^{N} y}{N}$ Media	$\overline{y} = \frac{\sum_{1}^{n} y}{n} = \hat{\overline{Y}}$
$Y = \overline{Y} \times N$ Total	$\hat{Y} = \overline{y} \times N$
$\sigma^2 = \frac{\sum_{1}^{N} (y - \overline{Y})^2}{N}$ Varianza	$\sigma_m^2 = \frac{\sum_{1}^{n} (y - \overline{y})^2}{n} = \hat{\sigma}^2$

En función de lo mencionado, queda claro que puede haber más de un estimador disponible para una constante poblacional. Esto lleva preguntarnos ¿qué criterios se deben tener en cuenta al momento de seleccionar un estimador? Eso lo veremos en el siguiente punto.

1.6 PROPIEDADES DESEABLES EN UN ESTIMADOR

Tres son las principales características a tener en cuenta al momento de elegir un estimador, que son: a) sesgo, b) precisión, y c) consistencia.

a) Sesgo

Para definir el concepto de SESGO debemos antes definir el concepto de valor esperado de un estimador. El **Valor Esperado de un Estimador** es el promedio de todos los valores posibles provenientes de la aplicación de ese estimador en odas las muestras distintas posibles. Si θ es la constante y $\hat{\theta}$ un estimador, $E(\hat{\theta})$ expresa el valor esperado o esperanza del estimador, que es igual a:

$$E(\hat{\theta}) = \frac{\sum_{i=1}^{M} \hat{\theta}_i}{M} \qquad (5)$$

donde M representa la cantidad de muestras distintas posibles y, por extensión, la cantidad de estimaciones posibles.

Es deseable que el estimador sea **Insesgado**. Esto es, que su valor esperado coincida con la constante que estima: $E(\hat{\theta}) = \theta$; en caso contrario el estimador es **Sesgado**.

b) Consistencia

Decimos que un estimador es consistente si a medida que $n \rightarrow N$ el estimador se acerca a la constante estimada. Así, en el límite para $n \rightarrow N$ el estimador coincide con el parámetro: $\hat{\theta} = \theta$. Un ejemplo es la media aritmética o media de la muestra; si n se lleva a N la media muestral $\bar{y}$ coincide con la media $\bar{Y}$ de la población.

Precisión

Para cada una de las M muestras distintas posibles hay un valor obtenido por aplicación del estimador; hay M estimaciones. Si promediamos estos M valores obtenemos el promedio de valores del estimador o valor esperado. La precisión de un estimador hace referencia a la dispersión de sus valores alrededor de esa media o valor esperado. A menor dispersión más preciso es el estimador y más confiable es el valor estimado a partir de una muestra específica. Esta dispersión se expresa como varianza del estimador. Esto es así, porque cuanto más preciso es el estimador más "parecidas" son las estimaciones, y es de esperar que la obtenida se encuentre cerca del valor de la constante poblacional estimada.

1.7 EL MUESTREO Y LA GENERACIÓN DE NUEVAS POBLACIONES

Definidos un universo y una variable de interés Y, queda definida la población de valores de la variable de interés Y. Uno podría pensar que ésta es la única población que está participando, pero no es así.

Ni bien definimos el tamaño de la muestra a usar aparecen nuevas poblaciones, ya no de la variable original Y. En realidad son las poblaciones de los valores estimados en todas las muestras distintas posibles por aplicación del estimador. Veamos un ejemplo.

Supongamos una población de tamaño N = 3 con los siguientes valores: 2, 5 y 8. Las constantes poblacionales que nos interesa son:

- el total de la población = 15
- la media por elemento = 5
- la varianza por elemento = 6

Supongamos que cada elemento puede participar sólo una vez en cada muestra (Muestreo Sin Reemplazo). Supongamos un tamaño de muestra de n= 2; hay tres muestras distintas posibles. Aplicamos los estimadores "copia" de totales, medias y varianzas a las distintas muestras posibles. Los resultados se indican en el siguiente cuadro:

Muestra	(Y_1, Y_2)	$\bar{y}$	$\hat{Y}$	σ_m^2
1	(2, 5)	3,5	10,5	2,25
2	(2, 8)	5,0	15,0	9,00
3	(5, 8)	6,5	19,5	2,25

Puede verse que ha aparecido una población de 3 valores para la media estimada, una población de tres valores para el total estimado y una población de tres valores para la varianza estimada. Cada una de estas poblaciones tiene su promedio o valor esperado, y su varianza.

La raíz cuadrada de cada una de estas varianzas recibe el nombre genérico de **Error Estándar**. Así, en el ejemplo, habría un error estándar para la media estimada, un error estándar para el total estimado y un error estándar para la varianza estimada.

El Error Estándar no es otra cosa que una desviación estándar, pero que expresa la variabilidad de los distintos valores estimados de una constante correspondientes a las distintas muestras posibles. No expresa la variabilidad de la variable original Y.

1.8 MUESTREO PROBABILISTA

Una muestra es un subconjunto de un universo y de una población, pero podemos preguntarnos ¿cualquier subconjunto de una población es una muestra? La respuesta es no. La primera condición es que la muestra sea **representativa** de la población de origen, ya que de no ser así ninguna información útil se puede obtener de ella. Se definen una serie de características a cumplir por la muestra para que se considere como un muestreo probabilista, que son las siguientes (Cochran, 1980):

- hay diferentes muestras distintas disponibles (M_1, M_2, M_3, ...);
- cada muestra posible tiene una probabilidad de selección conocida distinta de cero;
- por un proceso aleatorio se selecciona una de esas muestras;

- existe un método para obtener una única estimación a partir de la muestra seleccionada; y
- el estimador sigue una función de probabilidad conocida

Si se cumplen estas condiciones decimos que el muestreo es Probabilista.

Si bien cada elemento de la población debe tener una probabilidad conocida y distinta de cero de ser seleccionado, hay situaciones en las que esto no ocurre. Por ejemplo, cuando se selecciona una muestra "representativa". La condición de representatividad queda a criterio de quien hace la selección, de manera que jamás participarán de la muestra aquellos elementos que para el observador no reúnen esa condición.

Dentro de este esquema Probabilista, el diseño de muestreo básico es el conocido como **Muestreo Aleatorio Simple sin Reemplazo**. En el siguiente punto veremos este diseño en detalle.

2 MUESTREO ALEATORIO SIMPLE SIN REEMPLAZO (MAS)

2.1 CARACTERÍSTICAS

El diseño de muestreo es aleatorio porque los elementos de la muestra son seleccionados mediante un mecanismo de Azar. El más común consiste en numerar a todos los elementos del universo y luego llevar a cabo un sorteo hasta completar el tamaño de muestra predeterminado; los n elementos que han salido sorteados participan de la muestra. Una herramienta para esto son las Tablas de Números Aleatorios.

El diseño de muestreo es simple, porque no hay un tratamiento posterior de los elementos que fueron seleccionados, como ocurriría en un muestreo estratificado.

El diseño de muestreo es sin reemplazo, porque una vez que un elemento es seleccionado no es devuelto al universo y no puede volver a participar en la muestra. Así, en una muestra cada elemento puede participar sólo una vez.

Un dato que nos va a ser de utilidad es conocer el número M de muestras distintas de tamaño n que pueden construirse con un universo de tamaño N, si el muestreo es sin reemplazo. La respuesta la da el Análisis Combinatorio: el número buscado es la cantidad de combinaciones de N elementos tomados de n por vez. La fórmula de cálculo es:

$$M = \frac{N!}{n!\,(N-n)!} \qquad (6)$$

Por ejemplo, para una población de tamaño N = 50, un tamaño de muestra n = 5 y muestreo sin reemplazo, el número total de muestras distintas posibles es:

$$M = \frac{50!}{5!\;45!} = 2.118.760$$

2.2 EJEMPLO NUMÉRICO DE M.A.S.

2.2.1 Población de Referencia

Vamos a considerar la siguiente población de pesos en kilogramos:

Y_1	Y_2	Y_3	Y_4
3	7	8	10

El tamaño elegido de la muestra es n = 2. El número total de muestras distintas de tamaño 2 que se pueden formar, es:

$$M = \frac{4!}{2!\;2!} = \frac{24}{4} = 6$$

Las constantes de la población son:

Total : $Y = \Sigma y = 28$ kg

Media/elemento: $\overline{Y} = \frac{\Sigma y}{N} = \frac{28\text{ kg}}{4} = 7$ kg

Varianza/elemento: $\sigma^2 = \frac{\Sigma(yi - \overline{Y})^2}{N} = 6{,}5\text{ kg}^2$

2.2.2 Valores obtenidos y esperados de los estimadores

Los estimadores "copia" usados son los siguientes:

- para la media: $\overline{y} = \frac{\sum y}{n}$ (7)
- para el total: $\hat{Y} = \overline{y} \times N$ (8)
- para la varianza: $\sigma_m^2 = \frac{\sum (y - \overline{y})^2}{n}$ (9)

El Cuadro 3 indica las 6 distintas muestras posibles, y para cada muestra: a) los valores de Y asociados, b) el valor de la media estimada de Y, c) el total estimado de Y, y d) la varianza estimada de Y por elemento; usando estimadores "copia". En la penúltima fila se indican los valores esperados de los estimadores y en la última fila el valor de las correspondientes constantes poblacionales.

Cuadro 3

Nº	(Y_1, Y_2)	$\overline{y}$	$\hat{Y}$	σ_m^2
1	(3, 7)	5,0	20,0	4,00
2	(3, 8)	5,5	22,0	6,25

3	(3,10)	6,5	26,0	12,25
4	(7, 8)	7,5	30,0	0,25
5	(7,10)	8,5	34,0	2,25
6	(8,10)	9,0	36,0	1,00
Valor esperado del estimador:		**7,0**	**28,0**	**4,33**
Constante Poblacional:		**7,0**	**28,0**	**6,50**

Como puede verse en el Cuadro 3, los estimadores usados para determinar la media de la muestra y el total de la muestra son **Insesgados**. Si la media de la muestra es un estimador insesgado, es razonable que el total estimado a partir de esa media, ya que:

$$\hat{Y} = N\ \bar{y} \qquad (10)$$

Entonces, si $E(\bar{y}) = \bar{Y}$ se cumple que:

$$E(\hat{Y}) = E(N\ \bar{y}) = N\ E(\bar{y}) = N\bar{Y} = Y \qquad (11)$$

A diferencia del estimador de la media, el estimador usado para la varianza es **Sesgado**: la varianza por elemento de la población es 6,50 y el valor esperado del estimador es 4,33.

Resumiendo:

- La media aritmética $\bar{y}$ de los valores de la muestra es un estimador insesgado de la media de la población.
- El producto $\bar{y}\,N$ es un estimador INSESGADO del total de la población.
- La varianza de la muestra es un estimador SESGADO de la varianza por elemento de la población.

2.3 Constante Poblacional S_Y^2

A raíz del sesgo mostrado por la varianza "de la muestra, los especialistas buscaron otra alternativa para indicar la varianza de la población. Fue así que definieron la siguiente expresión:

$$\frac{\sum_{1}^{N} (y_i - \bar{Y})^2}{N - 1} \qquad (12)$$

Esta expresión no tiene un nombre propio, por lo que también se la denomina varianza, y es indicada con S^2. Para la población con la cual estamos trabajando, esta varianza toma el siguiente valor:

$$S^2 = 8{,}6667$$

Por cierto, estamos habilitados para usar la copia muestral como estimador de S^2. El estimador "copia" se indica con la misma letra pero en minúscula: s^2 .

Para entender las ventajas de usar este estimador repetiremos los cálculos indicados en el Cuadro 3, pero indicando s^2 en la última columna. Los resultados están en el Cuadro 4.

Cuadro 4

Nº	(Y_1, Y_2)	$\bar{y}$	$\hat{Y}$	s_y^2
1	(3, 7)	5,0	20,0	8,00
2	(3, 8)	5,5	22,0	12,50
3	(3,10)	6,5	26,0	24,50
4	(7, 8)	7,5	30,0	0,50
5	(7,10)	8,5	34,0	4,50
6	(8,10)	9,0	36,0	2,00
Valor esperado del estimador:		**7,0**	**28,0**	**8,667**
Constante Poblacional:		**7,0**	**28,0**	**8,667**

Como puede verse, el estimador copia s^2 es un estimador **insesgado** de la varianza S^2. Es por esta razón que en la Estadística se usa s^2 como indicador de la varianza de la muestra. Precisamente, de aquí en más usaremos las expresiones S^2 y s^2.

2.3.1 Varianzas en la muestra

En los Cuadros 3 y 4 vemos que aparecen nuevas poblaciones, distintas a la población original de los valores de la variable Y. Estas poblaciones nuevas son las correspondientes a los valores provenientes de aplicar los estimadores a todas las muestras distintas posibles. Y cada una posee su media y su varianza. La media es el valor esperado del estimador, ya indicados en los Cuadros 3 y 4. Ahora vamos a ver qué ocurre con las varianzas de estas poblaciones. En el Cuadro 5 se indican los valores de σ^2 y s^2 de los valores estimados en las distintas muestras posibles.

CUADRO 5

Muestra	$\bar{y}$	$\hat{Y}$	s_Y^2
1	5,0	20,0	8,00
2	5,5	22,0	12,50
3	6,5	26,0	24,50
4	7,5	30,0	0,50
5	8,5	34,0	4,50
6	9,0	36,0	2,00
σ^2	**2,1667**	**34,6667**	**65,7222**
s^2	**2,6000**	**41,6000**	**78,8667**

Como el estimador más utilizado es $\bar{y}$, la media de Y, su varianza para todas las muestras distintas posibles es la que en este momento nos interesa.

Por cierto, para la población del ejemplo es pequeño el número de muestras distintas que de ella se pueden obtener. Pero como ya vimos, si la población tuviera N = 50 unidades y el tamaño de la muestra fuese n = 5,

entonces hay disponibles 2.118.760 muestras sin reemplazo, lo que hace imposible calcular la varianza de la media usando el mecanismo aplicado. Esto nos lleva a pensar que la varianza del estimador de la media sólo puede determinarse cuando hay pocas muestras distintas posibles. Pero no es así.

Por suerte, los estudiosos nos dan la solución cuando nos dicen que no hace falta determinar todas las muestras posibles para conocer la varianza de la media, es suficiente con aplicar la siguiente fórmula:

$$\sigma_{\bar{y}}^2 = \frac{\sigma^2}{n} \times \left(\frac{N-n}{N-1}\right) \qquad (13)$$

Si aplicamos la fórmula con nuestros obtenemos el siguiente resultado:

$$\sigma_{\bar{y}}^2 = \frac{6{,}5}{2} \times \left(\frac{4-2}{4-1}\right) = \frac{6{,}5}{3} = \mathbf{2{,}1667}$$

El valor obtenido por aplicación de la fórmula coincide con el indicado en el Cuadro 5.

Pero dado que vamos a utilizar como expresión de la varianza a S^2 , vamos a expresar la varianza de la media como $S_{\bar{y}}^2$, y la fórmula anterior se convierte en:

$$S_{\bar{y}}^2 = \frac{S^2}{n} \times \left(\frac{N-n}{N}\right) = \frac{S^2}{n} \times \left(1-\frac{n}{N}\right) \qquad (14)$$

El estimador de esta varianza es:

$$s_{\bar{y}}^2 = \frac{s^2}{n} \times \left(\frac{N-n}{N}\right) = \frac{s^2}{n} \times \left(1-\frac{n}{N}\right) \qquad (15)$$

Reemplazando con los correspondientes valores, obtenemos:

$$\sigma_{\bar{y}}^2 = S_{\bar{y}}^2 = \frac{8{,}6667}{2} \times \left(1-\frac{2}{4}\right) = \mathbf{2{,}1667}$$

La raíz cuadrada de la varianza de la estimada recibe el nombre de **Error Estándar** del valor estimado. Así, la raíz cuadrada de la varianza de la media es el Error Estándar de la Media. Del mismo modo, la raíz cuadrada de la varianza del total estimado es el Error estándar del total estimado; etc.

En general, si un valor estimado es: $\theta = N\bar{z}$, entonces , su Error Estándar es:

$$S_\theta = N\, S_{\bar{z}}$$

Para el total estimado $N\,\bar{Y}$, su error estándar es:

$$S_{\hat{Y}} = N\, S_{\bar{y}} \qquad (16)$$

y su estimador es:

$$s_{\hat{Y}} = N\, s_{\bar{y}} \qquad (17)$$

Resumiendo, el indicador de la varianza poblacional que vamos a emplear es la expresión S^2, como estimador del mismo el término s^2, y como estimador del error estándar de la media vamos a usar:

$$s_{\bar{y}} = \frac{s}{\sqrt{n}} \times \sqrt{\left(1-\frac{n}{N}\right)} \qquad (17)$$

2.3.2 Intervalo de confianza

Si bien hay disponibles distintas muestras posibles, en la práctica vamos a operar con sólo una de ellas. Del mismo modo, de entre todas las medias distintas posibles nosotros vamos a detectar sólo una de ellas.

Acá surge una pregunta: la media estimada a partir de la muestra seleccionada ¿qué tan lejos o cerca está de la media verdadera? La respuesta es que no podemos saberlo y que solamente podemos hacer apreciaciones en términos de probabilidades.

Para poder hacer cálculos de probabilidades de la media estimada es necesario conocer el modelo de distribución de la media de la muestra. Aquí se presentan dos situaciones, que son:

1. Si la variable original Y sigue una distribución Normal, la media de la muestra de tamaño n sigue una distribución Normal, con media μ y varianza $S_{\bar{y}}^2$

2. Si la variable original Y no sigue una distribución Normal, el Teorema del Límite Central dice que a medida que el tamaño de la muestra aumenta la distribución de la media muestral tiende a una distribución normal con media μ y varianza $S_{\bar{y}}^2$.

El intervalo de confianza para una probabilidad P toma la siguiente forma:

$$IC(P) = \bar{y} \pm Z(p)\, s_{\bar{y}} \qquad (18)$$

donde Z(%) es el valor de la variable Normal estandarizada Z para un nivel de probabilidad p.

Cuando la varianza poblacional de Y es estimada a partir de la muestra, la variable Z es reemplazada por la variable "t" de Student para n-1 grados de libertad:

$$IC(P\%) = \bar{y} \pm t_{(n-1)}\, s_{\bar{y}} \qquad (19)$$

Como valor aproximado de "t" para una probabilidad del 95 % se puede emplear el valor 2 si la muestra es lo suficientemente grande. En este caso la expresión sería:

$$IC(95\%) = \bar{y} \pm (2)\, s_{\bar{y}} \qquad (20)$$

Volviendo al ejemplo, supongamos que la muestra seleccionado es la que hemos definido como muestra 4; en este caso, el intervalo de confianza para una probabilidad del 95 por ciento, es:

Media estimada = $\bar{y}$ = 7,50

Varianza estimada de Y = s^2 = 0,50

El Error estándar de la media se obtiene como:

$$s_{\bar{y}} = \frac{\sqrt{s2}}{\sqrt{n}}\sqrt{(1-\frac{n}{N})} = \frac{\sqrt{0,50}}{\sqrt{2}}\sqrt{1-\frac{2}{4}} = 0,35$$

$$C(95\%) = 7,5 \pm (2)\ 0,35 \begin{cases} \text{Límite Inferior (LI)} = 6,8 \\ \text{Límite Superior (LS)} = 8,2 \end{cases}$$

2.3.3 Cálculo del tamaño de la muestra para un error prefijado

Tomamos como punto de partida el intervalo de confianza: $\bar{y} \pm \text{Error}$, donde el error es el producto entre t(n-1) y el error estándar de la media: $t\, s_{\bar{y}}$. Si despejamos n, obtenemos la siguiente expresión, que permite calcular el tamaño de la muestra:

$$\hat{n} = \frac{1}{\frac{1}{N} + \frac{E^2}{t^2\ s^2}} = \frac{N\ t^2\ s^2}{t^2\ s^2 + N\ E^2} \qquad (21a)$$

Si la intensidad de muestreo es pequeña, digamos que si n/N ≤ 0,05, la fórmula (21a) se convierte en:

$$\hat{n} = \frac{t^2\ S^2}{E^2} \qquad (21b)$$

Si al error lo expresamos como por ciento de la media (E%), la fórmula es:

$$\hat{n} = \frac{1}{\frac{1}{N} + \frac{E_{\%}^2}{t^2\ CV_{\%}^2}} = \frac{N\ t^2\ CV_{\%}^2}{t^2\ CV_{\%}^2 + N\ E_{\%}^2} \qquad (22a)$$

donde CV% es el coeficiente de variación en por ciento. Si la intensidad de muestreo es pequeña, digamos que si n/N ≤ 0,05, la fórmula la fórmula (22a) se convierte en:

$$\hat{n} = \frac{t^2\ CV_{\%}^2}{E_{\%}^2} \qquad (22b)$$

Veamos un ejemplo con los siguientes datos:

N = 500 $\bar{y}$ = 200 s^2 = 3600

E = 20 E% = 10 CV% = 30

t = 2

Fórmula (21a):

$$\hat{n} = \frac{1}{\frac{1}{500} + \frac{400}{4 \times 3600}} = 33,58 \cong 34 \text{ unidades}$$

Fórmula (21b):

$$\hat{n} = \frac{4 \times 3600}{400} = 36 \text{ unidades}$$

Fórmula (22a):

$$\hat{n} = \frac{1}{\frac{1}{500} + \frac{100}{4 \times 900}} = 33,58 \cong 34 \text{ unidades}$$

Fórmula (22b):

$$\hat{n} = \frac{4 \times 900}{100} = 36 \text{ unidades}$$

3 ALGUNAS FÓRMULAS BÁSICAS

Una de las fórmulas comunes es la correspondiente a la Suma de Cuadrados de los desvíos respecto a la media. Cuando operamos con dos variables simultáneamente, tenemos:

- Suma de Cuadrados de los desvíos de X (SCDX)

$$SCDX = \sum (x - \bar{x})^2 = \sum x^2 - \frac{\left(\sum x\right)^2}{n} \qquad (23)$$

- Varianza estimada de X:

$$Sxx = Sx^2 = \frac{SCDX}{n-1}$$

- Suma de Cuadrados de los desvíos de Y (SCDY)

$$SCDY = \sum (y - \bar{y})^2 = \sum y^2 - \frac{\left(\sum y\right)^2}{n}$$

- Varianza estimada de Y:

$$Syy = Sy^2 = \frac{SCDY}{n-1}$$

- Suma de los productos cruzados (SPXY):

$$SPXY = \sum (x - \bar{x})(y - \bar{y}) = \sum xy - \frac{\sum x \sum y}{n}$$

- Covarianza entre X e Y:

$$Cov\ (X,Y) = Sxy = \frac{SPXY}{n-1}$$

Julio 2002