Aula Estadística

Nociones Básicas de Muestreo

Autor: Jorge Rubén Lorenzo.
Profesor Auxiliar Cátedra de Estadística y Sistemas de Información.

Escuela de Ciencias de la Educación.

Facultad de Filosofía y Humanidades

Universidad Nacional de Córdoba.

Resumen: En esta monografía se discuten nociones básicas sobre muestreo para un curso inicial de estadística. Comenzamos presentando de manera sucinta, algunas nociones fundamentales de la temática y luego se discuten los lineamientos teóricos de los principales métodos de muestreo probabilístico. Se ensaya mediante ejemplos sencillos, la aplicación de estas técnicas de muestreo. Luego se introducen nociones elementales para el cálculo del tamaño muestral en estudios de casos y controles. Continuando con la temática, se ejemplifica la manera de calcular el tamaño muestral en estudios de estimación de una proporción y una media en la población. Finalmente, se extiende el ejemplo para el caso del cálculo del tamaño muestral en prueba de hipótesis. Si bien los ejemplos dados son sencillos y no profundizan en la temática, resultan fáciles de aplicar en el aula y el mismo material sirve para la discusión en clase de procedimientos de estadística inferencial.

Palabras Claves: muestreo probabilístico, población, muestra, cálculo del tamaño muestral.

Comentario: Esta monografía fue escrita como una forma de divulgación de los resultados de un proyecto de investigación sobre didáctica de la estadística, el cual está vigente desde el año 2006 y radicado en el Centro de Investigaciones de la Facultad de Filosofía y Humanidades de la Universidad Nacional de Córdoba, y dirigido por el autor.

Destinatario: se consideran posibles destinatarios de esta obra principalmente aquellas personas que actualmente están investigando en la didáctica específica en este dominio del conocimiento. Sin embargo, también puede ser de utilidad para docentes de nivel medio, especialmente aquellos interesados en encontrar recursos educativos para el dictado de materias que tengan a la estadística como

eje central. Otros destinatarios posibles son alumnos de grado y posgrado interesados en ampliar su conocimiento sobre el tema de muestreo luego de haber tomado algún curso básico de estadística.

Nivel educativo: nivel medio y superior.

1. Introducción

Imaginemos que una persona está preparando arroz para el almuerzo, lo ha puesto a hervir y al cabo de un tiempo desea saber si está cocido. La persona tomará una cuchara, y sacará de la cacerola una cucharada de arroz, si al probarlo determina que el arroz está listo, terminará la cocción, pero si el arroz que ha probado aún no está listo, dejará la cacerola en hervor otro tiempo. Sigamos con la imaginación; esa misma persona luego de comer, mira por la ventana y percibe que el cielo está nublado y denso, entonces toma un paraguas y se dispone a salir. Aunque no lo parezca, este imaginario personaje está tomando decisiones en base a la teoría del muestreo. En efecto, para determinar si el arroz en cocción está listo o no, tomó una parte del mismo (la cucharada) para determinar el estado de cocción del arroz contenido en toda la cacerola. Del mismo modo, la porción del cielo que miró a través de la ventana, le permitió realizar una inferencia del estado del cielo en un espacio mucho mayor.

Estos sencillos ejemplos nos ilustran sobre la importancia del muestreo: una fracción, nos informa sobre la constitución del todo. La sabiduría popular ha sabido captar la importancia de este tema en la frase “para muestra, sobra un botón”, y en estadística los botones son de importancia fundamental ya que muchas veces se puede conocer características de las poblaciones solo a través de muestras. Si la muestra ha sido bien seleccionada, ésta proporcionará muy valiosa información de la población ya que en lo esencial contendrá, si no todas, sus principales características.

2. Algo de terminología

Para adentrarnos en la teoría del muestreo tendremos que definir algunos términos básicos, estos son:

a) **Población**: conjunto de individuos, elementos u objetos de los cuales se quiere obtener información.

b) **Unidades de muestreo**: número de elementos de la población, no solapados, que se van a estudiar. Así, las unidades de muestreo son todo individuo, elemento u objeto de los cuales se extraerá información. En general las unidades de muestreo, son denominadas también unidades de análisis.

c) **Muestra**: conjunto de unidades de muestreo extraídas de la población mediante un procedimiento definido.

La manera en que hemos presentado las definiciones refleja el procedimiento de muestreo. En primer lugar debemos establecer los límites de la población de donde se va a extraer la muestra, en este sentido, la población definida no debería ser tan extensa que resulte infinita, ni tan acotada que pueda ser abarcada en su totalidad, y por tanto no se requiera una muestra. Si dijéramos "todas las mujeres del mundo" habríamos definido teóricamente una población, pero ésta es infinita y prácticamente inabarcable. Ahora bien, si dijéramos todos los médicos residentes de un hospital rural, también habríamos definido una población, pero en este caso resulta finita y abarcable.

Una vez delimitada la población, las unidades de análisis quedan por defecto también definidas, y se procede al muestreo de esas unidades.

Surgen entonces otras cuestiones a considerar que son: a) el método de selección de la muestra, b) el tamaño de la muestra, c) el sesgo de la muestra y d) el error muestral. Una vez conocida la población, existen diferentes manera de recoger una muestra, a esto se refiere el método de muestreo. En términos generales existen tres tipos de muestreos, uno de ellos conocido como **muestro probabilístico**, en donde cada muestra tiene la misma probabilidad de ser recolectada; esto es, todos los individuos de la población tienen la misma chance de integrar la muestra durante el proceso de selección. Otro tipo de muestreo, se conoce como **muestreo intencional**, en donde la representatividad de la muestra está dada por el propio investigador, quien selecciona los individuos que mayor información puedan dar sobre la población en estudio. Este tipo de muestreo está

sujeto a lo que el investigador considera representativo, dado que él es quien decide cómo debe quedar conformada la muestra; sin embargo, no debe entenderse por ello que existe error en la constitución de la muestra, dado que en ocasiones solo interesa recolectar información de una parte puntual de la población, siempre acorde a los objetivos de la investigación. Finalmente, existe otro tipo de muestreo llamado **muestreo accidental**, en donde no se tiene en cuenta una norma o método para la recolección de las unidades de análisis. En este caso, el investigador incluye en su muestra las unidades de análisis de que dispone en un momento dado. Aunque este es el método de muestreo más falible, es comúnmente usado dado que es el más fácil de implementar especialmente cuando la intención no es la inferencia, sino la puesta a prueba de instrumentos de recolección de datos.

El tamaño muestral es otro aspecto importante dado que depende del tamaño y la composición de la población. La muestra resulta representativa en la medida en que aproxima con cierta fidelidad las características de la población, y en este punto es necesario que no se confunda tamaño con calidad de la muestra. Si bien es cierto que las muestras pequeñas suelen adolecer de falta de representatividad, también es cierto que las muestras muy grandes pueden estar sesgadas si la técnica de muestro no ha sido la adecuada. Muchas veces, partiendo de un muestreo probabilístico en una población conocida, es posible estimar el tamaño de la muestra con precisión.

El sesgo de la muestra puede darse en dos sentidos generales, uno de los cuales hemos mostrado ya, y que se refiere a que la selección no considere igualmente a todas las unidades de análisis. En este sentido, el muestreo intencional y el accidental caen en este tipo de sesgo, pero deberá recordarse que a veces es el único tipo de muestreo que el investigador quiere o puede implementar. Otro tipo de sesgo se ha denominado *sesgo de respuesta* por afectar principalmente a encuestas; se da siempre que el encuestado no quiera responder o que la pregunta del cuestionario no capte la dimensión significativa de la respuesta.

Por último, el error muestral es un estadístico que se calcula a partir de la muestra, y es una medida que indica cuánto se aparta ésta de la población. En estadística esto suele denominarse error estándar. Cada vez que se estima una

media o una proporción a partir de una muestra, suele acompañarse de su correspondiente error estándar de la media o la proporción, según sea el caso.

3. Una aproximación al muestreo probabilístico

El muestreo probabilístico o aleatorio, se define como el procedimiento mediante el cual, en la muestra seleccionada, todos los miembros de la población han tenido igual chance de ser seleccionados; esto equivale a decir que todas las unidades de análisis de esa población tienen la misma posibilidad de ser elegidas para formar la muestra.

La definición dada se refiere al **muestreo aleatorio simple con reposición**, es decir cada vez que se toma una unidad de análisis y se efectúa la medición de la variable de interés, dicha unidad vuelve a la población y tiene la misma chance que las demás para volver a ser seleccionada. Un procedimiento de estas características, hace inagotable la selección de elementos de la población. Ahora bien, si se selecciona un elemento de la muestra y se mide la variable de interés, pero no se lo repone en la población, el procedimiento de muestreo se denomina **muestreo aleatorio simple sin reposición**; es decir, algunas veces el procedimiento de muestreo no necesariamente requerirá retornar la unidad de análisis al conjunto de la población.

Cuando la población es infinita o tan grande que pueda considerarse como tal, no habrá diferencia entre ambos métodos. Ahora bien, si la fracción de muestreo es mayor que 0.1, existirán diferencias en las conclusiones que puedan sacarse de la muestra por el procedimiento elegido. La fracción de muestreo se refiere a la proporción entre el tamaño de la muestra (n) y el tamaño de la población (N), esto es: *fracción de muestreo*= n/N. Si esta fracción es igual al valor de 0.1, quiere decir que la muestra contiene el 10% de la población. Evidentemente, el problema consiste en obtener una muestra no sesgada cuando el tamaño de la población es finito y no muy grande. La fracción de muestreo es una medida de la probabilidad para ser seleccionada, asignada a cada unidad de análisis. Si N disminuye en cada selección de una unidad (es decir, el individuo seleccionado no se repone en la población), hay que tener en cuenta que el resto de las unidades tendrán una mayor probabilidad de ser elegidas. Como se dijo, si N es grande, las probabilidades posteriores a cada selección de la unidad de análisis, no variarán sustancialmente, pero si N es pequeño esa variación puede

perjudicar el procedimiento de muestreo. Por lo tanto, el valor crítico 0.1 en la fracción de muestreo, es un elemento de juicio para optar por un muestreo aleatorio simple con o sin reposición. La siguiente tabla ilustra lo dicho en los anteriores párrafos. Tomemos dos hipotéticas poblaciones, una de tamaño 100 y otra de tamaño 1000. A medida que el tamaño de la muestra aumenta, la probabilidad de ser seleccionado aumenta también; así cuando el tamaño de la muestra contiene 20 unidades de análisis, en la población de tamaño 100, el elemento a ser muestreado tiene una chance del 20% de ser seleccionado, mientras que en la población de tamaño 1000 solo tiene el 2%. En este punto es dado observar que en la primera población se ha pasado el límite crítico de la fracción de muestreo de 0,1. Cuando la muestra alcanza el tamaño 30, las probabilidades de una unidad de análisis de ser seleccionada en la muestra de tamaño 100 es ya del 30%, muy superior a las chances que tiene una unidad de análisis si la muestra fuera de tamaño 10. Nótese que en la muestra de tamaño 1000, las chance en esta instancia son solo del 3%. Esta es la razón por la cual, al trabajar con poblaciones pequeñas, es siempre necesario considerar el muestreo con reposición.

	n=10	n=20	n=30
N=100	10%	20%	30%
N=1000	1%	2%	3%

Si todos los elementos de la población pueden ser listados y ordenados aleatoriamente, el muestreo aleatorio simple sin reposición puede llevarse a cabo por un procedimiento llamado **muestro aleatorio sistemático**. En este caso, conociendo la fracción de muestreo, se procede a establecer el intervalo mediante el cual se seleccionará una unidad de análisis de la lista (siempre que ésta se halle aleatorizada al momento de realizar la muestra). Por ejemplo, si se desea hacer una selección aleatoria de 10 escuelas y se cuenta que existen 260 en total, el *factor de elevación* será 260/10= 26; es decir cada escuela seleccionada representa a 26 escuelas. Si todas las escuelas se hallan ordenadas al azar en una lista, se seleccionará el *valor de arranque* del muestreo de manera aleatoria.

El valor de arranque será un número cualquiera entre 1 y 26; entonces, sea x el valor seleccionado, ese será la primera escuela que conforme la muestra. La segunda escuela, será la que en la lista ocupe el valor x+26, la tercera escuela será la que ocupe el lugar x+2*26 y así sucesivamente. En otras palabras, el factor de elevación N/n=v indica también el total de conjuntos que pueden formarse en la población, entonces siendo x un valor aleatorio, el total de los elementos de la muestra queda definido como $x+v$; $x+2v$; $x+3v$... Si del factor de muestreo resulta que v es un número no entero, se lo redondea al entero menor, con lo que algunas muestras tendrán n-1, lo cual no introduce ninguna perturbación en el muestreo si n >50.

Si la población es infinita o muy grande podremos seleccionar unidades de análisis para la muestra sin la necesidad de reponerlos, pero deberíamos preguntarnos qué tan grande debería ser la muestra para que pudiera representar todas las características de la población. Para evitar el problema de recolectar una muestra demasiado grande, se puede utilizar una variante del muestreo aleatorio conocida como **muestreo estratificado**. Como su nombre lo indica, se trata de identificar estratos dentro la población y luego proceder a realizar un muestreo aleatorio simple en cada uno de ellos. Por lo tanto, sea N el tamaño de la población, cada estrato quedará definido como N_1, N_2...N_k, de modo que el tamaño de la población sea igual a la suma de los tamaños de cada estrato:

$$N= N_1+N_2,+...N_k$$

Por el mismo procedimiento, la muestra total quedará definida por la suma de las muestras obtenidas en cada uno de los estratos; entonces sea n_1, n_2...n_k el tamaño de la muestra en cada uno de los estratos, el tamaño de la muestra total será igual a:

$$n= n_1+n_2+...n_k$$

El muestreo estratificado es una variante que maximiza la información extraída de una población muy grande, ya que es menos costosa y garantiza la precisión

de los estimadores, para una o más características de la población. La dificultad radica en que los estratos deben ser adecuadamente definidos, de modo que no se cuenten las mismas unidades de análisis en más de un estrato; es decir, que no exista solapamiento entre ellos. En general cuanto más diferentes son los estratos entre sí, y cuanto más homogéneos son éstos en su interior, mejores las estimaciones que se obtienen.

Existen varias maneras de seleccionar una muestra de tamaño n en cada uno de los estratos. La más común de todas es tomar la muestra de tamaño proporcional al tamaño del estrato. Otra manera de hacerlo es considerando la varianza de la variable en estudio; en algunos casos ésta es conocida a partir de estudios previos y es posible aplicar algunas fórmulas de cálculo para obtener el tamaño muestral, pero la situación usual es que la varianza se desconoce y es necesario estimarla a partir de la muestra. Por último, se pueden seleccionar muestras iguales para cada estrato, lo cual resulta practicable si cada estrato tiene aproximadamente el mismo tamaño.

Si la población es de fácil acceso, será fácil también implementar un procedimiento de muestreo, pero si la población es de difícil acceso lo será también el muestreo. Esto último es frecuente cuando se trabaja con personas y se debe obtener un panorama general de la población. En tales casos, suelen utilizarse un tipo de muestreo conocido como **muestreo por conglomerados**. Los conglomerados son zonas o áreas en los que se ha dividido la población, y deben ser lo más representativos posible de ésta, y además deben ser homogéneos entre sí. La homogeneidad del conglomerado es un verdadero desafío para este tipo de muestreo, dado que si éstos no son homogéneos la representatividad de la muestra se deteriora. Por ello, es común que los investigadores definan un conglomerado, que a su vez contenga otros conglomerados. Entonces, el muestreo puede seleccionar en primera etapa un número dado de conglomerados y en segunda etapa, pueden obtener una muestra de algunos conglomerados contenidos en los primeros. Por lo dicho, este tipo de muestreo se conoce como **muestreo por conglomerados polietápicos**. Los conglomerados son fácilmente comprensibles cuando se los define como áreas geográficas, pero ciertas variables de constructo pueden ser tratadas como conglomerados.

4. Un ejemplo aplicado de técnicas de muestreo

Intentaremos mostrar la aplicación de los conceptos que hemos descripto anteriormente, en un sencillo ejemplo. Para ello supondremos que los directivos de una institución desean conocer aspectos claves de su alumnado para mejorar su oferta educativa. En una primera aproximación se decide trabajar sobre una muestra representativa del alumnado comprendido en el segundo y tercer ciclo de EGB, para lo cual se procede a describirlo en términos de población. A tal fin se construye la siguiente tabla:

Tabla 1: Población en función de la distribución de matrícula en EGB 2 y EGB 3

	Nivel	Total Alumnos
EGB 2	4	56
	5	67
	6	57
EGB 3	7	89
	8	91
	9	87
Total		**447**

Se tiene entonces que la población total de alumnos objeto de estudio es: N=447. Ahora bien, por razones de costo, la escuela puede realizar el estudio tomando una muestra de solo 50 alumnos en total, lo cual obliga a los investigadores a pensar en que esos 50 alumnos seleccionados deben representar lo más fielmente posible el total de la población escolar de EGB 2 y 3. Conociendo el tamaño de la población (N) y de la muestra (n), se calcula la fracción de muestreo como:

$$f = \frac{n}{N} = \frac{50}{447} = 0.11$$

Luego, se calcula el factor de elevación como:

$$v = \frac{N}{n} = \frac{447}{50} = 8.94$$

Entonces se tiene que la muestra es el 11% aproximadamente del total de la población, y cada uno de los alumnos seleccionados para conformarla representará aproximadamente a 9 alumnos. Puesto que la fracción de muestreo es mayor que 0.1, el muestreo deberá hacerse con reposición.

Definida ya la muestra, deben escogerse las variables que serán objeto de estudio. En este caso, los investigadores deciden trabajar sobre dos mediciones de logro y dos mediciones de la condición del estudiante. Estas variables son:

Variable 1: Rendimiento en Matemáticas,

Variable 2: Rendimiento en Lengua,

Estas variables son de naturaleza métrica en tanto sus valores se obtienen de una medición con un instrumento estandarizado.

Variable 3: Cobro de Beca,

Variable 4: Uso del Comedor,

Estas variables son nominales y dicotómicas, en tanto se clasifican como SI en el caso de que el escolar este cobrando una beca de estudios o almuerce diariamente en el comedor escolar; se clasifica como NO en caso contrario. Para codificar estas variables se usan los valores 1 (SI) y 0 (NO).

Las dos primeras variables están afectadas por el nivel de escolaridad, en tanto se considera que las exigencias de una evaluación para un escolar en el nivel 4, no puede ser la misma que para un alumno en el nivel 9. Es decir, el logro deberá medirse con exámenes equivalentes pero no iguales. Por esta razón se decide tomar a cada uno de los niveles de escolaridad como un estrato, de modo que se recortan seis estratos correspondientes a los niveles 4 al 9, cuya composición en cantidad de alumnos está expresada en la tabla 1. Se desprende de ello que el muestreo más apropiado será el aleatorio estratificado. Por tanto, el tamaño muestral al interior de cada estrato se obtiene de la siguiente manera:

$$n * \left(\frac{N_j}{N}\right)$$

Donde:

n= tamaño de la muestra a seleccionar,

N_j= tamaño del estrato,

N= tamaño de la población.

Tabla 2: Tamaños muestrales por estrato

n_1	50*(56/447)=6.26	n_1=6	Nivel 4
n_2	50*(67/447)=7.49	n_2=8	Nivel 5
n_3	50*(57/447)=6.37	n_3=6	Nivel 6
n_4	50*(89/447)=9.95	n_4=10	Nivel 7
n_5	50*(91/447)=10.17	n_5=10	Nivel 8
n_6	50*(87/447)=9.73	n_6=10	Nivel 9
Total		n_{1-6}=50	

Nótese que en el procedimiento de cálculo, los decimales se redondearon al número entero superior cada vez que la parte decimal era mayor que 0.5, a excepción de n_2, el cual se redondeó al entero superior aún cuando su parte decimal fue de 0.49. Esto se aplicó para no obtener una muestra de tamaño total menor a 50.

Una vez obtenidos los tamaños muestrales se procede a efectuar la medición de las variables rendimiento en matemáticas y lengua; luego se calcula la media y la varianza en cada uno de los estratos:

Tabla 3: rendimiento en matemáticas

Matemáticas					
Nivel 4	Nivel 5	Nivel 6	Nivel 7	Nivel 8	Nivel 9
9	5	6	6	5	9
7	2	5	6	4	8
5	9	9	8	8	5
9	8	9	9	9	4
4	4	5	6	9	6
6	3	2	5	8	3
	6		4	5	9
	6		5	3	8
			8	6	4
			2	7	5

Tabla 3 continuación: rendimiento en lengua

Lengua					
Nivel 4	Nivel 5	Nivel 6	Nivel 7	Nivel 8	Nivel 9
4	9	6	9	9	10
7	2	10	6	4	3
10	8	9	9	9	5
9	10	10	10	10	5
10	4	5	10	9	7
6	10	5	5	5	10
	6		3	5	9
	7		8	8	7
			8	9	7
			7	5	9

Con los valores de la tabla 3, es posible calcular la media y la varianza de las variables rendimiento en matemáticas y lengua, tal como se muestra en la tabla 4.

Tabla 4: promedio y varianza en Matemáticas y Lengua

	Matemáticas	
	Media	Varianza
Nivel 4	6.66	4.26
Nivel 5	5.37	6.69
Nivel 6	6	7.2
Nivel 7	5.9	4.32
Nivel 8	6.4	4.48
Nivel 9	6.1	4.98

Tabla 4: Continuación

	Lengua	
	Media	Varianza
Nivel 4	7,66	5,86
Nivel 5	7	8,28
Nivel 6	7,5	5,9
Nivel 7	7,5	5,16
Nivel 8	7,3	5,12
Nivel 9	7,2	5,51

Ahora estamos en condiciones de estimar media y varianza para el total de la población para las dos variables en estudio. Entonces, si el procedimiento de muestreo ha sido aleatorio estratificado, sabemos que la media poblacional se calcula mediante la siguiente fórmula:

$$\bar{X} = \sum_{\mathrm{i}=1}^{\mathrm{k}} \frac{N_h}{\mathrm{N}} \bar{x}h$$

Donde:

$\bar{X}$ = estimación de la media poblacional.

Nh= tamaño del estrato h.

N= tamaño poblacional.

$\bar{x}h$ = media muestral de la variable x en el estrato h.

La varianza poblacional se calcula con la siguiente fórmula:

$$V(x) = \sum_{i=1}^{k} w_h^2 \, (1 - fh) \frac{S_h^2}{n_h}$$

Donde:

Wh= fracción de muestreo.

fh= fracción de muestreo dentro del estrato.

$\frac{S_h^2}{n_h}$ Varianza para el estrato h, sobre su tamaño

Aunque los cálculos pueden obtenerse fácilmente en una hoja de cálculo, se presentará una tabla que ejemplifica los procedimientos paso a paso. Nótese de antemano que la fracción de muestreo está presente en el cálculo de la estimación de la media y la varianza.

Tabla 5: cálculo de componentes de la fórmula de media y varianza

Estrato	$\frac{N_h}{N}$	$\left(\frac{N_h}{N}\right)^2$	*fh*	1- *fh*
Nivel 4	56/447=0.125	0,015625	6/56=0,1071	0,8929
Nivel 5	67/447=0.149	0,022201	8/67=0,1194	0.8806
Nivel 6	57/447=0.127	0,016129	6/57=0,1052	0.8948
Nivel 7	89/447=0.199	0,039601	10/89=0,1123	0.8877
Nivel 8	91/447=0.203	0,041209	10/91=0,1098	0.8902
Nivel 9	87/447=0.194	0,037636	10/87=0,1149	0.8851

Para el caso de la variable rendimiento en matemáticas se tiene la siguiente tabla de cálculo:

Tabla 6: Estimación del rendimiento en matemáticas

Estrato	$\frac{N_h}{N} * \bar{x}h$
Nivel 4	0.125*6.66= 0.8325
Nivel 5	0.149*5.37= 0.80013
Nivel 6	0.127*6= 0.762
Nivel 7	0.199*5.9= 0.1741
Nivel 8	0.203*6.4= 1.2992
Nivel 9	0.194*6.1= 1.1834
Total	**5.05113**

Tabla 6 continuación: Estimación del rendimiento en lengua

Estrato	$\frac{N_h}{N} * \bar{x}h$
Nivel 4	0.125*7.66= 0.9575
Nivel 5	0.149*7= 1,043
Nivel 6	0.127*7.5= 0,9525
Nivel 7	0.199*7.5= 1,4925
Nivel 8	0.203*7.3= 1,4819
Nivel 9	0.194*7.2= 1,3968
Total	**7,3242**

En las tablas anteriores se ha estimado la media de la población para la asignatura matemáticas, que resulta igual a 5,05113; y para lengua que resulta igual a 7,3242. Resta ahora estimar las varianzas de ambas variables.

Tabla 7: Estimación de la varianza en matemáticas

w_h^2 * 1- fh	$(S\_h\text{^}2)/n\_h$	$(w_h^2$ * 1- $fh)$ * $(S\_h\text{^}2)/n\_h$
0,015625 * 0,8929= 0,01395	4,26/6=0,71	0,0099045
0,022201 * 0,8806= 0,01955	6,69/8=0,84	0,016422
0,016129 * 0,8948= 0,01443	7,2/6=1,2	0,017316
0,039601 * 0,8877= 0,03515	4,32/10=0,432	0,0151848
0,041209 * 0,8902= 0,03668	4,48/10=0,448	0,01643264
0,037636 * 0,8851= 0,03331	4,98/10=0,498	0,01658838
		Total **0,092**

Tabla 7 continuación: Estimación de la varianza en lengua

w_h^2 * 1- fh	$(S\_h\text{^}2)/n\_h$	$(w_h^2$ * 1- $fh)$ * $(S\_h\text{^}2)/n\_h$
0,015625 * 0,8929= 0,01395	5,86/6=0,97667	0,0136245465
0,022201 * 0,8806= 0,01955	8,28/8=1,035	0,02023425
0,016129 * 0,8948= 0,01443	5,9/6=0,9833	0,014189019
0,039601 * 0,8877= 0,03515	5,16/10=0,516	0,0181374
0,041209 * 0,8902= 0,03668	5,12/10=0,512	0,01878016
0,037636 * 0,8851= 0,03331	5,51/10=0,551	0,01835381
		Total **0,1033**

Hemos resuelto ya el problema de la estimación de los parámetros media y varianza de las variables rendimiento en matemática y lengua, las cuales resultan ser:

Matemáticas		Lengua	
Media	Varianza	Media	Varianza
5.05113	0,092	7,3242	0,1033

Nos resta ahora estimar los parámetros de las variables 3 y 4: Cobro de Beca - Uso del Comedor. Hay dos cuestiones que diferencian el tipo de muestreo de estas variables respecto de las anteriores. La primera es –como ya se dijo- que estas variables son dicotómicas y se clasifican como 1 y 0; la segunda cuestión es que dividir en estratos no sería el procedimiento adecuado. En la conformación por niveles, sabemos que los niveles 4, 5, y 6 corresponden a EGB 2, pero los niveles 7, 8 y 9, reciben el nombre en la ciudad de Córdoba de Ciclo Básico

Unificado (C.B.U.). Es factible tratar a estos niveles como dos grupos diferentes y homogéneos en su constitución interna, con lo cual una técnica de muestreo por conglomerados resultaría apropiada. Los datos recogidos diferencian el Grupo 1= EGB 2 y Grupo 2= CBU. Por tratarse de una variable que puede registrarse fácilmente, los directivos decidieron trabajar sobre una muestra de 60 alumnos para obtener el mismo tamaño muestral en cada conglomerado. Así, para el Grupo 1 n=30 y para el Grupo 2 n=30. La siguiente tabla resume los resultados obtenidos:

Tabla 8: distribución de las variables cobro de Beca y Uso de comedor

(total de grupo n=30)	Cobro de Beca (SI)		Uso de Comedor (SI)	
Grupo 1	26	0,87	22	0.73
Grupo 2	12	0,4	23	0.76

Sobre la base de la muestra se puede estimar la proporción poblacional de aquellos escolares que cobran beca estudiantil y usan el comedor. Para ello se aplica la siguiente fórmula de cálculo:

$$\widehat{P_B} = \frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} M_i}$$

$$\widehat{P_C} = \frac{\sum_{i=1}^{n} A_i}{\sum_{i=1}^{n} M_i}$$

Donde:

A_i= total de la variable que toma valores de 0 ó 1 en el conglomerado i.

M_i= Tamaño del conglomerado i.

$$\widehat{P_B} = \frac{26 + 12}{30 + 30} = \frac{38}{60} = 0{,}633$$

$$\widehat{P_C} = \frac{22 + 23}{30 + 30} = \frac{45}{60} = 0.75$$

Al comienzo del ejemplo se dijo que los directivos de una institución desean conocer aspectos claves de su alumnado para mejorar su oferta educativa en términos de calidad. Ahora, cuentan con importante información sobre las

variables analizadas, tomadas de una muestra representativa de la población escolar:

Rendimiento en matemáticas	Rendimiento en lengua	Proporción de escolares con becas	Proporción de escolares que usan el comedor
μ=5.05113 σ^2=0,092	μ=7,3242 σ^2=0,1033	ρ=0.63	ρ=0.75

5. Estudios de casos y controles

Los diseños de casos y controles son comunes en estudios clínicos, que tienen por objeto identificar factores de riesgo en la población, directamente relacionados a una enfermedad. Sin embargo, esta metodología se ha ampliado hacia otros campos de investigación, tale como la psicología, la economía y la educación. El ejemplo que desarrollaremos aquí estará centrado en estudios clínicos dado que en esa área son más fácilmente comprensibles.

Si bien, los estudios de cohortes siempre resultan los más adecuados, en ocasiones estudios de casos y controles ofrecen ventajas comparativas al ser éstos más económicos y demandar menos tiempo. La principal diferencia entre ambos tipos de estudios radica en el método empleado para seleccionar la muestra. En los estudios de casos y controles, se identifica un grupo de personas con una condición o enfermedad (casos), y se los compara con un grupo apropiado que no tenga esa condición o enfermedad (controles). En investigación clínica, la condición a la que se hace referencia es una enfermedad, pero bien podría ser otra variable diferente; por ejemplo, la condición podría ser la situación de empleo categorizada como trabajador o desocupado.

Una vez seleccionados los casos y los controles, se verifica la presencia de un factor que se supone está asociado a la enfermedad, de lo que resulta que, si la frecuencia de exposición al factor es mayor en el grupo de casos que de controles, podremos decir que el factor predispone a la enfermedad. Al contrario, si la frecuencia de exposición al factor es mayor en el grupo de controles, diremos que ese factor previene la enfermedad.

En estudios de este tipo, la distribución de sujetos se presenta en una tabla 2x2 como la siguiente:

Disposición de sujetos en un estudio de casos y controles			
FACTOR	Casos	Controles	
Expuestos	a	b	a + b
No expuestos	c	d	c + d
	a + c	b + d	Total

Así se tiene que existen

1) casos que están expuestos (a)
2) casos que no están expuestos (c)
3) controles que están expuestos (b)
4) controles que no están expuestos (d)

Como medida de la frecuencia de exposición entre los casos se puede utilizar el cociente:

$$Eca. = \frac{P_1}{1 - P_1}$$

P_1 denota la probabilidad de exposición entre los casos, que es el cociente entre los casos expuestos y no expuestos. De la tabla se deduce que tal probabilidad resulta de:

$$Eca. = \frac{a/(a+c)}{c/(a+c)} = \frac{a}{c}$$

De modo similar se valora la frecuencia de exposición entre los controles, que resulta en:

$$Econ. = \frac{P_2}{1 - P_2}$$

$$Econ. = \frac{b/(b+d)}{d/(b+d)} = \frac{b}{d}$$

La medida más utilizada para cuantificar la relación entre la exposición al factor y la presencia de la condición es la razón de ventaja o más conocida como *odds ratio* (OR), por su denominación anglosajona.

$$OR = \frac{P_1(1 - P_2)}{P_2(1 - P_1)} = \frac{a/c}{b/d} = \frac{a * d}{b * c}$$

Si el valor de OR es igual a 1 la presencia o ausencia del factor no se asocia a la enfermedad (casos). Si el valor es menor que 1 la presencia del factor tiene un efecto preventivo en la aparición de la enfermedad; al contrario si, es mayor que 1 la exposición aumenta las probabilidades que los expuestos al factor sean quienes tienen la enfermedad. De cualquier modo, las estimaciones del OR se calculan dentro de un intervalo de confianza del 95%, para poder rechazar la hipótesis de no asociación entre exposición al factor y condición.

Se puede observar que si se construye una tabla como la mostrada más atrás, el OR se obtiene de la multiplicación cruzada de las celdas, lo cual lo hace un indicador muy fácil de obtener y, bajo la hipótesis adecuada, resulta en una estimación robusta de la tasa de incidencia de un factor. El verdadero desafío para utilizar esta técnica es contar con una muestra adecuada; para determinar el tamaño de la muestra es necesario contar con los siguientes datos:

1) La magnitud de la diferencia que se quiera detectar que resulte estadísticamente significativa y que tenga interés clínico. Para ello se necesita conocer:

1.a) la aproximación al OR que se desea estimar, cuyo parámetro se designa como w.

1.b) la frecuencia de exposición entre los casos, cuyo parámetro es P_1.

1.c) la frecuencia de exposición ente los controles, cuyo parámetro es P_2.

2) El riesgo de cometer un erro tipo I, cuyo parámetro es α. En general, se establece una seguridad del 95%, que resulta en un valor de α=0.05, para prevenir este error.

3) El poder estadístico del estudio, cuyo parámetro es 1-β. Esto representa la probabilidad de cometer un error tipo II. Es común en este caso utilizar un valor de β=0.2, que resulta en un poder estadístico del 80%.

Entonces, si se plantea una hipótesis bilateral, el cálculo del tamaño muestral se obtiene de aplicar la ecuación 1:

$$n = \frac{\left[\left[Z_{1-\alpha/2}\sqrt{2P(1-P)} + Z_{1-\beta}\sqrt{P_1(1-P_1) + P_2(1-P_2)}\right]\right]^2}{(P_1 - P_2)^2}$$

El valor de P es igual a $P_1+P_2/2$; y los valores $Z_{1-\alpha/2}$ y $Z_{1-\beta}$ se obtienen de la distribución normal estándar, en función de la seguridad y poder estadístico escogidos para el estudio. En este caso, sabiendo que α=0.05 y β=0.2, los parámetros referidos tienen un valor de:

$Z_{1-\alpha/2} = 1.96$

$Z_{1-\beta} = 0.84$

La ecuación 1 se aplica si se quiere obtener una muestra con la misma cantidad de casos y controles; pero si se sabe que la relación entre ambos no implica la misma proporción, la ecuación 1 se modifica teniendo en cuenta ese dato. Así, donde $c=m/n$, denota la cantidad de controles por cada caso, la ecuación 1 se modifica como lo muestra la ecuación 2:

$$n = \frac{\left[\left[Z_{1-\alpha/2}\sqrt{(c+1)P(1-P)} + Z_{1-\beta}\sqrt{cP_1(1-P_1) + P_2(1-P_2)}\right]\right]^2}{c(P_1 - P_2)^2}$$

6. Un ejemplo aplicado para el cálculo muestral en un estudio de casos y controles.

En una población rural se ha detectado la aparición de una alergia eruptiva que afecta a una fracción importante de niños. Se sabe también que esta alergia puede ser causada por un virus o por la exposición a agentes alergénicos ambientales, tales como plaguicidas. El departamento de salud se propone determinar si existe relación entre la aparición de la enfermedad y el hecho de vivir en cercanías de campos sembrados y tratados con plaguicidas.

De la información con que se cuenta hasta ahora, se deduce que un 30% de los controles viven en cercanías de campos tratados con plaguicidas. A los fines del estudio, se considera importante un OR de 3 como diferencia entre los grupos. Con estos datos pueden establecerse los siguientes parámetros para determinar el tamaño de la muestra necesario para el estudio de casos y controles:

a) Frecuencia de exposición entre los controles: 30%

b) OR previsto como clínicamente importante: 3

c) Nivel de seguridad contra el ET I: 95%

d) Poder estadístico (ET II): 80%

Se estima ahora la frecuencia de exposición entre los casos, mediante la aplicación de la siguiente ecuación:

$$P_1 = \frac{wp_2}{(1 - P_2) + wP_2} = \frac{3 * 0.30}{(1 - 0.30) + 3 * 30} = 0.56$$

Se estima entonces que un 56% de los casos, viven en cercanías de campos tratados con plaguicidas. Se necesita ahora conocer el valor de P, para aplicar la ecuación 1. Este valor resulta de 0.30+0.56/2=0.43. Reemplazando los valores en la ecuación 1 se tiene:

$$n = \frac{\left[\left[1.96\sqrt{2 * 0.56(1 - 0.56)} + 0.84\sqrt{0.56\,(1 - 0.56) + 0.30(1 - 0.30)}\right]\right]^2}{(0.56 - 0.30)^2} = 55.8 \cong 56$$

Tenemos entonces que, por las características del estudio se necesitan estudiar 112 individuos, esto es 56 casos y 56 controles, para poder detectar como significativo un OR de 3.

Supongamos ahora que para el estudio se decide incluir dos controles por caso, así en el estudio *c*=2. Bajo estas circunstancias, el cálculo del tamaño muestral se obtiene aplicando la ecuación 2 como sigue:

$$n = \frac{\left[\left[1.96\sqrt{(2 + 1)0.56(1 - 0.56)} + 0.84\sqrt{2 * 0.56\,(1 - 0.56) + 0.30(1 - 0.30)}\right]\right]^2}{2(0.56 - 0.30)^2} = 38.27 \cong 38$$

Siendo $c=m/n$, la cantidad de controles por caso, y sabiendo que la relación es de dos controles por caso, la muestra deberá contener aproximadamente 38 casos y 76 controles.

7. Determinación del tamaño muestral en estudios para determinar parámetros

Sabemos que uno de los principales objetivos del muestreo es contar con una fracción de la población, que permita conocer cómo se comporta ésta en su totalidad. Por ello, la determinación del tamaño muestral es un requisito fundamental para estudios que tienen como objetivo determinar un parámetro en una población. En los ejemplos que siguen se presentarán los procedimientos para estimar una proporción y una media en la población.

7.1 Cálculo del tamaño muestral para estimar una proporción

Si se desea estimar una proporción, es necesario contar con algunos datos previos para el cálculo del tamaño de la muestra, estos datos son:

a) El nivel de confianza de la estimación, que se calcula como 1-α, siendo α la probabilidad asociada al error tipo I. Generalmente se toman niveles de confianza de 95% o 99%, cuyos valores estandarizados son 1.96 y 2.58 respectivamente.

b) La precisión que se desea obtener en el estudio.

c) Un valor aproximado del parámetro a estimar. Generalmente, estudios previos informan sobre estimaciones realizadas con anterioridad, o bien ofrecen alguna aproximación de cuál podría ser el valor del parámetro. Si no se cuenta con esa información previa, se utiliza un valor de 0.5 (50%), que maximiza el tamaño muestral.

Ejemplo: se sabe por estimaciones previas que la prevalencia de la depresión en la población adulta es del 5%. Se desea realiza un estudio para determinar cuál es la proporción de personas afectadas en el presente por ese trastorno, ¿cuál es el tamaño óptimo de la muestra en este caso? Se definen los datos para este problema de la siguiente manera:

a) Nivel de confianza de la estimación: 95%

b) Precisión para el estudio 3%

c) Un valor aproximado del parámetro a estimar 5%

Con estos datos se puede aplicar la siguiente ecuación para estimar el tamaño de la muestra:

$$n = \frac{Z_\alpha^2 * p * q}{d^2}$$

Los símbolos de la fórmula, tienen en este ejemplo, los siguientes valores:

$Z_\alpha^2 = (1.96)^2 = 3.841$

$p = proporción\ esperada,$ 5% (0.05)

$q = (1 - p) = 1 - 0.05 = 0.95$

$d = nivel\ de\ presición,$ 3%

Reemplazando los valores en la fórmula se tiene:

$$n = \frac{1.96^2 * 0.05 * 0.95}{0.03^2} = 202.66 \cong 203$$

Es decir, para este estudio se requiere una muestra de 203 casos.

Ahora bien, suponiendo que la depresión sea un trastorno difícil de diagnosticar y que no se tengan datos realistas para hacer una estimación previa de su incidencia en la población, deberíamos tomar una proporción esperada de casos igual a 0.5. Si se reemplaza el valor en la ecuación se tiene que:

$$n = \frac{1.96^2 * 0.5 * 0.5}{0.03^2} = 1066.94 \cong 1067$$

Resulta evidente la ventaja de contar con estimaciones de la incidencia de una enfermedad en la población, dada la ventaja a la hora de estimar el tamaño de una muestra representativa.

Supongamos ahora que se conoce el tamaño de la población en donde se quiere estimar el parámetro; en tal caso, la ecuación presentada anteriormente se modifica de la siguiente forma:

$$n = \frac{N * Z_\alpha^2 * p * q}{d^2 * (N - 1) + Z_\alpha^2 * p * q}$$

Volviendo al planteo original del problema, supongamos ahora que la población de individuos adultos es igual a 17500, entonces N=17500. Los valores anteriores no se modifican, entonces se tiene:

$Z_\alpha^2 = (1.96)^2 = 3.841$

$p = proporción\ esperada,$ 5% (0.05)

$q = (1 - p) = 1 - 0.05 = 0.95$

$d = nivel\ de\ presición, 3\%$

Reemplazando en la ecuación se tiene:

$$n = \frac{17500 * 1.96^2 * 0.05 * 0.95}{0.03^2(17500 - 1) + 1.96^2 * 0.05 * 0.95} = 200.44 \cong 200$$

Ahora sabemos que necesitamos exactamente 200 casos para conformar la muestra.

7.2 Cálculo del tamaño muestral para estimar una media

Para poder calcular el tamaño necesario de una muestra para estimar una media, deberíamos contar con los siguientes datos:

a) El nivel de confianza de la estimación, que se calcula como 1-α, siendo α la probabilidad asociada al error tipo I (Z_α^2).

b) La precisión con la que se desea estimar el parámetro, que es la amplitud del intervalo de confianza necesario (d^2).

c) Una estimación previa de la varianza de la distribución de la variable (S^2).

Si se conocen esos datos, se puede aplicar la siguiente ecuación para calcular el tamaño de la muestra necesario.

$$n = \frac{Z_\alpha^2 * S^2}{d^2}$$

Ejemplo: se desea conocer la estatura promedio en una población con una seguridad del 95% ($Z_\alpha^2 = 1.962^2$), y una precisión de ± 5 cm. Se tiene información que la varianza de la estatura en esa población es de 110 cm. Reemplazando los valores en la fórmula se tiene que:

$$n = \frac{1.962^2 * 110}{5^2} = 16.9 \cong 17$$

Es decir, necesitamos solo 17 individuos para estimar la media de la estatura.

Como en el caso anterior, la población bajo estudio puede ser finita y entonces la ecuación anterior se modifica de la siguiente manera:

$$n = \frac{N * Z_\alpha^2 * S^2}{d^2 * (N-1) + Z_\alpha^2 * S^2}$$

Digamos entonces que la población a estudiar consta de 22.000 individuos, por lo cual el cálculo del tamaño de la muestra se efectúa con la siguiente modificación.

$$n = \frac{22.000 * 1.962^2 * 110}{5^2(22.000-1) + 1.962^2 * 110} = 16.88 \cong 17$$

8. Determinación del tamaño muestral en estudios de contrastes de hipótesis

Para el contraste de hipótesis, generalmente se pretende comprobar si las medias o proporciones de dos muestras son estadísticamente diferentes. Para determinar el tamaño muestral se precisa conocer:

a) La magnitud de la diferencia a detectar que tenga interés o sea relevante.

b) Una aproximación a los parámetros de la variable en estudio.

c) Seguridad para el estudio, esto es el riesgo de cometer un error tipo I (α).

d) Poder estadístico del estudio, o riesgo de cometer un error tipo II (1-β).

e) Definir si la hipótesis será unidireccional o bidireccional, siendo esta última más conservadora en cuanto a la predicción.

8.1 Comparar dos proporciones

Para el cálculo del la comparación de dos proporciones se usa la ecuación vista ya en el cálculo de la muestra para estudios de casos y controles, esto es:

$$n = \frac{\left[\left[Z_{1-\alpha/2}\sqrt{2P(1-P)} + Z_{1-\beta}\sqrt{P_1(1-P_1) + P_2(1-P_2)}\right]\right]^2}{(P_1 - P_2)^2}$$

Los parámetros de la ecuación son los mismos, excepto que ahora, P_1 es la proporción en el grupo de referencia o control y P_2 es la proporción en grupo nuevo o tratamiento. P es la media de las dos proporciones $P_1 + P_2/2$.

Para la comparación de dos medias se utiliza una variación de la fórmula vista en el apartado 7, esta es:

$$n = \frac{2\,(Z_\alpha + Z_\beta)^2 * S^2}{d^2}$$

Aquí, la varianza (S^2) es la que corresponde a la del grupo control o de referencia, mientras que d es el valor mínimo de la diferencia que se desea detectar y que resulta importante.

Ejemplo de comparación de dos proporciones: supongamos que se dispone de un tratamiento (T1) para el alivio del dolor, que tiene una eficacia del 70%. Se sabe que existe un nuevo tratamiento (T2), con el cual se desea comparar el anterior, estableciendo que el nuevo tratamiento para ser puesto en práctica, debe asegurar una eficacia del 90% en el alivio del dolor. Para el estudio se fija un nivel de confianza del 95% (α=0.05) y un poder estadístico del 80% (1-β=0.84). Aplicando la ecuación se tiene que el tamaño de la muestra es de:

P=0.7+0.9/2=0.8

$$n = \frac{\left[1.64 * \sqrt{2 * 0.8\,(1-0.8) + 0.84\sqrt{0.7\,(1-0.7) + 0.9\,(1-0.9)}}\right]^2}{[0.7-0.9]^2} = 48$$

Para realizar la comparación, en cada grupo se precisan 48 sujetos.

Ejemplo de comparación de dos medias: un nuevo método de enseñanza de la escritura, se considera eficaz si logra disminuir en 15 la cantidad de errores de escritura, respecto al método habitual. Se sabe que la desviación estándar de los sujetos que aprendieron con el método habitual, es de 16. Se trabaja sobre una confianza del 95% y un poder estadístico del 80%. El cálculo del tamaño muestral se obtiene con la aplicación de la ecuación:

$$n = \frac{2(1.64 + 0.84)^2 * 16^2}{15^2} = 13.99 \cong 14$$

Se necesitan 14 sujetos por grupo.

8.2 Tamaño muestral ajustado a las pérdidas

En todo estudio se hace necesario ajustar el tamaño muestral a las posibles pérdidas. Este ajuste se realiza sobre la cantidad de sujetos que conformarán la muestra mediante la siguiente ecuación.

$$n_a = n * \left(\frac{1}{1-R}\right)$$

Así por ejemplo, si en el estudio anterior se espera tener un 15% de pérdidas de sujetos, el ajuste del tamaño muestral estará dado por:

$$n_a = 14 * \left(\frac{1}{1-0.15}\right) = 16.47 \cong 17$$

Habría que incluir entonces 17 sujetos por grupo.

Anexo

Tablas con valores de Z_α y Z_β más frecuentemente utilizados

Z_α		
α	Hipótesis unilateral	Hipótesis bilateral
0.2	0.842	1.282
0.15	1.036	1.440
0.1	1.282	1.645
0.05	1.645	1.960
0.025	1.960	2.240
0.01	2.326	2.576
Potencia		
β	$1-\beta$	Z_β
0.01	0.99	2.326
0.05	0.95	1.645
0.10	0.90	1.282
0.15	0.85	1.036
0.20	0.80	0.842
0.25	0.75	0.674
0.30	0.70	0.524
0.35	0.65	0.385
0.40	0.60	0.253
0.45	0.55	0.126
0.50	0.50	0.00

Bibliografía

Lagarea Barreiro, P. y Puerto Albandoz, J. P. (2001). Población y muestra. Técnicas de muestreo. *Management Mathematics for European Schools*, 94342 – CP. Pag. 1 -20.

Pértegas Díaz, S. y Pita Fernandez S. (2002). Cálculo del tamaño muestral en estudios de casos y controles. Cátedra de Atención Primaria, 9: 148 – 150.