IDICSO
Instituto de Investigación en Ciencias Sociales
Facultad de Ciencias Sociales
Universidad del Salvador

MATERIAL DEL ÁREA
EMPLEO Y POBLACIÓN

Material AEPHC8 Junio 2004

Elementos básicos de muestreo aleatorio.

HORACIO CHITARRONI

http://www.salvador.edu.ar/csoc/idicso

Hipólito Yrigoyen 2441 – C1089AAU Ciudad de Buenos Aires – República Argentina

TABLA DE CONTENIDOS

Notas sobre el autor

HORACIO CHITARRONI

- Lic. en Sociología, Universidad Nacional de Buenos Aires (UBA).
- Profesor de Enseñanza Secundaria, Normal y Especial en Sociología, UBA.
- Docente Titular, Facultad de Ciencias Sociales, Universidad del Salvador (USAL).
- Docente de la Maestría en Ciencias Sociales del Trabajo, Facultad de Ciencias Sociales, UBA.
- Coordinador e Investigador Principal, Área Empleo y Población, IDICSO, USAL.
- Consultor del Consejo Nacional de Coordinación de Políticas Sociales – SIEMPRO (Sistema de Evaluación, Seguimiento y Monitoreo de Programas Sociales).

Dirigir comentarios a la siguiente casilla de correo electrónico:
Lic. Horacio Chitarroni: hchitarroni@siempro.gov.ar

Departamento de Comunicación y Tecnología del IDICSO: idicso@yahoo.com.ar

1. ¿Qué cosa es una muestra?

Una población es un conjunto de elementos definidos por ciertas especificaciones: por ejemplo, las personas que votaron en las últimas elecciones, los habitantes de la Ciudad de Buenos Aires de 14 y más años, los hogares de la Ciudad de Paraná, los establecimientos escolares de educación básica de la provincia de Corrientes o las ONGs dedicadas a la temática de género.

Al interior de una población, pueden distinguirse subpoblaciones o estratos que nos interese considerar en forma separada para el análisis: por ejemplo, dentro de las personas que votaron en la última elección pueden separarse a las mujeres de los varones. O bien, entre las escuelas de Corrientes pueden distinguirse las públicas de las privadas. Ahora bien, si sólo me intereso por las escuelas públicas, entonces la población está ceñida a ellas: ya no son un estrato o subpoblación de un conjunto mayor, sino que ellas mismas constituyen toda la población.

En muchas ocasiones, las poblaciones son excesivamente grandes para ser indagadas en su totalidad (un relevamiento total de una población se denomina censo) o bien, aun pudiendo serlo, no resultaría conveniente ni se justificaría hacerlo así.

En estos casos, se recurre a una muestra: un subconjunto de los elementos que componen la población (casi siempre una pequeña proporción de ellos) obtenidos bajo ciertos recaudos de manera que satisfagan nuestras necesidades. No es lo menos frecuente que estas necesidades apunten a estimar ciertas características del conjunto total (la población) a través de características del subconjunto (la muestra). Esto, en términos estadísticos, se denomina hacer una inferencia. Y – como lo sugiere el sentido común – tendrá sentido hacerlo si podemos confiar en que la muestra se parece aceptablemente a la población: los recaudos que se toman para obtener una muestra probabilística se encaminan a garantizar esto último.

Por cierto, nunca podemos estar seguros de que lo que averiguamos en la muestra es exactamente así en la población: por ejemplo, si en una muestra de población de una cierta ciudad la proporción de mujeres es de 52,3%, no hay garantía alguna de que esta proporción sea la misma en el total de la población. Más aun, casi seguramente no debe ser exactamente la misma. Y como prueba de ello, si obtuviéramos una nueva muestra, seguramente hallaríamos una proporción levemente distinta, aunque es de esperar que no *demasiado* distinta: por ejemplo, 52,1 o 52,5 (o cualquiera otra, no muy superior ni muy inferior).

La proporción de mujeres (o cualquier otra medida, tal como el promedio de edad o de ingresos) que encontramos en la muestra se denomina un *estimador*, puesto que pretende *estimar* la verdadera medida poblacional: esta última se designa como *parámetro*. Vale decir, que las muestras proporcionan *estimadores* de los *parámetros* poblacionales.

El cociente entre la cantidad de elementos incluidos en una muestra – el tamaño muestral – y el total de la población o universo, se denomina fracción de muestreo y, generalmente, se trata de un número muy pequeño.

2. ¿Por qué usar muestras?

¿Por qué usar una muestra? De lo dicho más arriba se infiere que apelamos a ellas toda vez que los universos que queremos indagar son demasiado vastos, dispersos o numerosos, de manera que demandaría mucho esfuerzo indagarlos en su totalidad: resultaría muy costoso, llevaría demasiado tiempo o ambas cosas a la vez. Pero esto parece sugerir que hemos de preferir las muestras como un mal menor, y que si pudiéramos, sería siempre mejor relevar la totalidad del universo: es decir, llevar a cabo un censo.

En realidad, no es siempre ni necesariamente así. En todo caso, podría serlo si se tratara de averiguar cosas muy simples, pero si las características que queremos relevar requirieran de un observador especializado, jamás dispondríamos de suficientes personas convenientemente adiestradas como para llevar a cabo este relevamiento exhaustivo. Esta es la razón por la cual los censos de población, habitualmente, recogen una información muy escueta: al punto que mucha gente se pregunta por las razones de tanto esfuerzo, en vista de que se va a obtener un producto tan magro. En realidad, los censos implican un esfuerzo extensivo: se obtiene una información escasa y simple de *casi* la totalidad de los elementos que componen el universo. En realidad, ningún censo logra la exhaustividad completa. Hay regularmente y por distintas razones, un subrregistro de casos, ya que algunas personas se niegan a ser censadas, en tanto que otras pueden estar en lugares inaccesibles o invisibles para los censistas[1]. También hay, con menor frecuencia, dobles conteos. Por eso, aunque – como se ha dicho en el punto anterior – jamás conoceremos con exactitud el valor del parámetro a través de una muestra, tampoco lo lograríamos seguramente mediante un censo. Y si la información a recoger es compleja y exige pericia por parte del encuestador, entonces es muy probable que la estimación muestral resulte en definitiva más confiable y cercana al verdadero parámetro que la que se obtendría de un procedimiento censal.

Un ejemplo sencillo puede facilitar la comprensión de esta cuestión: supongamos que quisiéramos saber cuántos fósforos hay en un *pack* que contiene diez cajas. Cada caja anuncia tener 100 fósforos, pero sabemos que esta cantidad puede no ser exacta. ¿abriríamos todas las cajas, volcando en el piso su contenido para, luego, sentarnos a contar los fósforos uno por uno?. Supongamos que así lo hiciéramos: en el primer conteo acaso obtendríamos algo menos de mil: digamos 994. Obstinadamente, contaríamos de nuevo: ¿llegaríamos al mismo resultado?. Es improbable. Si contáramos una y otra vez, seguramente, además de quedar muy cansados y aburridos, obtendríamos resultados diferentes sin acertar a saber cuál sería el correcto. También es verdad que ninguno de los resultados se alejaría demasiado: a veces 994, a veces 1002, en otra ocasión 999, etc. Podríamos sacar un promedio entre los distintos resultados o aún aceptar cualquiera de ellos sin temor a errar mucho. Pero, siendo así, ¿no hubiera sido mejor contar los fósforos de un par de cajas y obtener de ellas el promedio, para luego multiplicarlo por diez, con

[1] Por ejemplo, las personas que viven en la vía pública – los "sin techo" – son muy difíciles de incluir en los censos, aunque el que se llevó a cabo en la Argentina, en 2001, hizo un esfuerzo en tal sentido. El autor participó en el diseño y la dirección de un relevamiento de esta población realizado en la Ciudad de Buenos Aires en 1997, que estuvo lejos de ser exhaustivo.

mucho menor esfuerzo...?. Sin duda, casi todos convendrán en que este último procedimiento es el más práctico.

Ahora bien, en nuestro ejemplo se trataba de una cosa muy simple: tanto como contar. Pero supongamos que tuviéramos que medir el largo o – peor aún – el diámetro de la cabeza de cada fósforo para obtener el valor promedio: siguiendo la misma lógica, seguramente, sería mejor hacer esto muy cuidadosamente con una limitada cantidad de fósforos, sacar el promedio y – luego – suponer que el promedio general no se situará muy lejano del así obtenido. Puesto que si hubiéramos pretendido hacer esta medición, no ya para las diez cajas sino para el centenar de fósforos que contiene cada una, nos expondríamos a incurrir en múltiples errores de medición que nos alejarían, en definitiva, del valor verdadero en mayor medida que la estimación.

En nuestro primer ejemplo, las dos cajas de fósforos que hemos contado serían una muestra del universo compuesto por el *pack* de diez cajas. En el segundo ejemplo, el puñado de fósforos que sometimos a medición sería una muestra del millar de fósforos contenidos en el *pack.* En definitiva, y una vez más, es casi imposible conocer el verdadero valor del parámetro: hemos de contentarnos con estimarlo y, para ello, conviene hacerlo asegurándonos de que hemos seleccionado la muestra de tal manera de que el error en que se incurre al estimar sea el menor posible. Y por añadidura, es muy deseable poder conocer el tamaño probable de ese error: vale decir, si el conteo realizado a partir de dos cajas arroja una cantidad promedio de 997 fósforos (porque una de ellas contenía 996 y la otra 998), ¿en cuánto nos estaremos equivocando al decir que hay 997 en todo el *pack*?. Pues bien, las técnicas de muestreo aleatorio tienden a procurar estos objetivos.

3. ¿Por qué una muestra más grande es mejor? (*¡pero no tanto...!*)

Un primer principio del muestro aleatorio, que el sentido común tiende a aceptar de buen grado, dice que cuanto mayor es el tamaño de la muestra, tanto mejor. Hasta cierto punto, esto es verdad. Y seguramente, el sentido común nos inclinará a pensar también que si el universo es muy vasto, tanto más grande habrá de ser la muestra. Pues bien: un segundo principio que, en cambio, nos haría fruncir el ceño con cierta desconfianza, afirma que esto último no es cierto: los universos más grandes no requieren tamaños muestrales mayores.

En definitiva, podremos resumir diciendo que una muestra más grande tiende a ser mejor, pero que el tamaño muestral es independiente de la magnitud del universo.

Un pequeño ejemplo nos persuadirá de que es así. Resulta muy claro que, cuando arrojamos una moneda al aire – puesto que tiene dos lados – la probabilidad de obtener cara es de ½, es decir, 0,50[2]. Sin embargo, si tiramos diez veces la moneda, sólo por casualidad obtendremos cinco caras: podremos obtener seis, siete, tres o cuatro. Inclusive, también podrá pasar que obtengamos ocho o dos, etc. Bastará con hacer la prueba para comprobarlo[3]. Muy bien, ¿qué esperaremos que suceda si tiramos la moneda un centenar de veces?: seguramente no obtendremos 50 caras, pero es muy poco probable que obtengamos setenta o treinta; aún sesenta o cuarenta. Y si la tiramos mil veces: allí tampoco tendremos 500 caras, pero seguramente será imposible obtener 400 o 600: saldrá una proporción de caras cercana a 50% (por ejemplo, 49% o 50%). Y si obtuviéramos, por ejemplo, 700 caras, pensaríamos seriamente que se trata de una moneda tramposa. Vale decir, a medida que aumentamos la cantidad de tiradas (que equivaldría al tamaño muestral), el resultado se va aproximando cada vez más a la probabilidad teórica. Digamos que el error será paulatinamente más pequeño: sin embargo, no hay ningún número de tiradas – por elevado que sea – que nos asegure obtener 50% de caras. Después de cierto número de tiradas, el error tiende a estabilizarse y desciende sólo marginalmente con los sucesivos aumentos, de manera que no tiene sentido seguir incrementando desmedidamente las tiradas: lo mismo sucede con los tamaños de las muestras. Por otra parte, el ejemplo es útil para advertir otra cosa: el resultado se fue aproximando a la probabilidad teórica independientemente de la cantidad posible de tiradas, que es infinita. Esta cantidad infinita de tiradas posibles sería el equivalente del universo.

En el caso de las muestras, las cosas son similares: la diferencia estriba en que en nuestro ejemplo nosotros conocemos anticipadamente la probabilidad teórica, a la que nos va aproximando a medida que crece nuestra "muestra de tiradas de moneda". En cambio, en las situaciones usuales del muestreo estamos tratando de estimar algo que desconocemos: un parámetro del universo, a partir de los resultados de una muestra de casos. Y también aquí, a medida que agregamos casos a la muestra, el resultado tiende a aproximarse al parámetro, independientemente de cuál sea el tamaño del universo, que bien podría ser infinito sin que esto dejara de ser cierto.

[2] Esto, siempre que estemos dispuestos a desdeñar la probabilidad de que caiga de canto...
[3] El autor acaba de hacerlo: obtuvo cara en siete oportunidades sobre diez tiradas.

Un segundo ejemplo resultará de utilidad. Tenemos una enorme bolsa que contiene bollillas negras y blancas, por mitades. Supongamos que extraemos de ella sólo dos bolillas: nada garantizará que saldrá una de cada color. Bien podremos extraer dos negras o dos blancas. Si aumentamos la cantidad de bollillas extraídas a diez, entonces tampoco acertaremos a extraer cinco de cada color, pero nos sorprendería extraer las diez de un mismo color. Si la muestra obtenida fuera de cien (o de mil) bolillas, las proporciones se irían equilibrando paulatinamente, aproximándose a las que tenemos en la bolsa. Al punto que si desconociéramos estas últimas y, en una extracción de un centenar de bolillas obtuviéramos 48 blancas y 52 negras, no vacilaríamos en pensar que en la bolsa debe haber, aproximadamente, la mitad de cada color. Y a estos efectos sería totalmente indiferente el tamaño de la bolsa, que podría contener mil, 10 mil, 100 mil o cien millones de bolillas.

Este mismo ejemplo, complicándolo un poco, nos servirá para introducir otra noción. Pensemos ahora que en vez de tomar una sola muestra de diez bolillas extraemos diez montoncitos, cada uno de diez. Estas muestras presentarán mezclas bastante diferentes entre sí: las habrá con distintas combinaciones de colores. Acaso haya alguna que tenga, exactamente, cinco bolillas de cada tonalidad. Y es concebible (aunque poco probable) encontrar algún montoncito monocolor. Ahora, hagamos el supuesto de que obtenemos otras diez muestritas pero de mayor tamaño: de cien bolillas cada una. ¿Qué sucederá?: pues sucederá que, si bien persistirá la heterogeneidad, la mayoría de los montoncitos se aproximará a las proporciones de bolillas blancas y negras existente en la bolsa (50 y 50), en tanto que pocos se alejarán significativamente de esas proporciones: será muy raro encontrar ahora mezclas de 70 y 30. Y si repitiéramos la operación, pero ahora con muestras de 300 bolillas, entonces se acentuará la tendencia al agrupamiento en torno a las proporciones de la bolsa, mientras que se volverán aún más raros los montoncitos “70 y 30”.

En definitiva, que cuanto más grande sea la muestra, podremos confiar más en que se parece a la población.

4. Los fundamentos y la distribución de muestreo (*créase o no...*)

En el punto anterior procuramos persuadir al lector de que estas cosas suceden como suceden por la vía del sentido común y la comprensión intuitiva. Ha llegado, sin embargo, el momento de apelar a razones más científicas. Nos ocuparemos, pues, de la distribución muestral.

La teoría del muestreo aleatorio se sustenta en el llamado *teorema del límite central*, que afirma lo siguiente:

Teorema del límite central

Si de una población cuya distribución es normal, con media = μ y varianza = σ^2 se obtuvieran sucesivas muestras aleatorias de tamaño = n, las medias de estas muestras formarían una distribución normal, cuya media sería igual a μ (la misma media de la población) y cuya varianza sería igual a la varianza poblacional dividida por el tamaño de las muestras, es decir σ^2 / n. Por lo tanto, el desvío estándar de esta distribución de medias muestrales, que recibe el nombre genérico de distribución muestral o distribución de muestreo, será igual a la raíz cuadrada de esa misma expresión: $\sigma / \sqrt{n}$.

Analicemos con algún cuidado lo que afirma este teorema.

- en primer lugar pide que la población en cuestión sea normal. ¿Qué quiere decir esto?. Supongamos que se tratara de estimar el ingreso promedio de los hogares argentinos. Para que pudiésemos valernos de este teorema con tal propósito, se requeriría que dicha variable siguiera una distribución normal (lo cual no es un inconveniente menor, porque muchísimas variables – y particularmente los ingresos – distan de distribuirse normalmente).
- en segundo lugar se dice que, si se cumpliera esa condición y nos pusiéramos a obtener una gran cantidad de muestras diferentes de esa población (todas de igual tamaño) y luego nos tomáramos el trabajo de calcular las medias de ingresos de cada muestra (que, obviamente, serían diferentes entre sí), entonces estas medias de ingresos conformarían a su vez una distribución *normal* denominada *distribución de muestreo.*
- y, casi mágicamente, esta nueva distribución de medias de muestras, tendría una media que resultaría igual a μ, la media de la población de la que se extrajeron todas estas muestras. Para decirlo de otro modo, las medias muestrales se distribuirían normalmente en torno a la media de la población.
- Por otra parte, la varianza de esta distribución que hemos llamado distribución de muestreo sería igual a la varianza de la población, σ^2 dividida por el tamaño de las muestras. Esto involucra dos consecuencias: a) quiere decir que la varianza de la distribución de muestreo es proporcional a la de la población (vale decir, cuanto más heterogénea sea la población tanto más lo será la distribución de muestreo) y b) podremos afirmar que la varianza de la distribución de muestreo es inversamente proporcional al tamaño de las muestras (si las muestras son grandes, la distribución de

sus medias será más homogénea, es decir, estas medias se parecerán más entre sí y – por cierto – se parecerán más a la media de la población).

Ahora bien, si – como lo hemos señalado – en la mayoría de las oportunidades nos interesamos por estimar variables que no tienen una distribución normal en la población, ¿de qué nos servirá toda esta *cháchara*?.[4] Pues bien, existe una extensión del mencionado teorema, que nos exonera del requisito de la normalidad. Se trata de la llamada *ley de los grandes números*, que dice así:

Ley de los grandes números

Si de una población cualquiera, con media = μ y varianza = σ2 se obtuvieran sucesivas muestras aleatorias de tamaño = n, a medida que se incrementa el tamaño muestral, las medias de estas muestras tendrán a formar una distribución normal, cuya media sería igual a μ (la misma media de la población) y cuya varianza sería igual a la varianza poblacional dividida por el tamaño de las muestras, es decir σ2 / n. Por lo tanto, el desvío estándar de esta distribución de medias muestrales, que recibe el nombre genérico de distribución muestral o distribución de muestreo, será igual a la raíz cuadrada de esa misma expresión: σ / √n.

Dicho de otro modo, que aunque la población no sea normal (aunque las variables que pretendemos estimar tengan, en la población, una distribución apartada de la normalidad), si las muestras son suficientemente grandes (digamos, n >= 100), entonces todas estas cosas se producirán de todos modos. Las medias muestrales se distribuirán normalmente en torno a la media de la población.

Ahora bien, si podemos convencernos de que la distribución de muestreo será normal, entonces también convendrá recordar ciertas propiedades del modelo de la curva normal, ya estudiadas oportunamente. Sabemos que se trata de una curva simétrica (donde la media y la mediana coinciden) y que aproximadamente dos desvíos a cada lado de la media quedan comprendidos el 95% de los casos.[5] En otros términos, que menos del 5% de los casos pueden alejarse del valor promedio (hacia arriba o hacia abajo) en más de dos desvíos estándar. Si aplicamos esto a una distribución de muestreo, podríamos decir que menos del 5% de las medias muestrales (siempre y cuando las muestras sean de tamaño suficiente) podrían distar de la media poblacional más de dos desvíos estándar de la

[4] El término *cháchara*, que el diccionario define como palabrerío insustancial, ha caído en desuso. Es una suerte de antigualla lingüística. Sin embargo, hay una anécdota histórica referida a él. A comienzos de los años 80 el gobierno del presidente Alfonsín convocó a un plebiscito para aceptar o rechazar un acuerdo de límites con la República de Chile, que había sido motivo de larga controversia. El partido gobernante – el radicalismo – impulsaba la aceptación, en tanto que una parte de la oposición justicialista lo rechazaba. En torno a la cuestión, tuvo lugar un debate público y televisado entre el entonces canciller – Licenciado Dante Caputo – y un senador justicialista, el doctor Vicente Saadi. El debate adquirió algunos ribetes humorísticos, puesto que – por ineficiencia de sus asesores o por propia confusión – el veterano dirigente justicialista entremezcló sus apuntes y extravió el rumbo de su argumentación. Entonces, ofuscado, acusó a su oponente de que sus palabras eran "pura cháchara". Quedó famoso... Esto, por cierto, nada tiene que ver con el muestreo aleatorio, pero el episodio me resulta divertido y cedí al arbitrio de rememorarlo.

[5] En rigor, 1,96 desvíos estándar a cada lado de la media contienen el 95%. Si nos alejamos dos desvíos llegamos a 95,5%.

distribución de muestreo. O bien, al revés, que el 95% de las medias muestrales no se alejan de la media de la población en más de dos desvíos.

Inmediatamente podremos dar un nuevo paso: nadie se toma el trabajo, en la realidad, de obtener una larga serie de muestras de igual magnitud. Habitualmente, solo obtenemos una, de un tamaño dado, que podemos imaginar como un caso cualquiera de la distribución de muestreo integrada por infinitas muestras de ese mismo tamaño. Ahora bien, si calculamos la media (por ejemplo, la media de los ingresos familiares) de los elementos de la muestra, podremos tener confianza en que esa es una de las muestras integrantes del 95% central y, por lo tanto, la media de la población no debe distar de ella más de dos desvíos estándar, hacia arriba o hacia abajo. Con esta confianza, si pudiéramos estimar el valor de ese desvío estándar y lo sumáramos y restáramos dos veces a la media muestral, quedaría determinado un rango (que denominaremos un intervalo de confianza) dentro del cual, con algo más de 95% de probabilidad, debe estar ubicada la media de la población.

Esta distancia entre el estimador muestral y el parámetro poblacional está dada, pues, por el desvío estándar de la distribución muestral y se denomina error de muestreo. Se indica con e. Puesto que, como hemos visto, el desvío estándar de la distribución muestral es tanto mayor cuanto más heterogéneo es el universo, las estimaciones muestrales están sujetas a un error más grande si la dispersión es mucha. Inversamente, dado que ese desvío es más pequeño a medida que el tamaño de la muestra aumenta (porque ese tamaño está en el denominador del cociente), con muestras más grandes tendremos un intervalo de confianza menor y estimaremos con mayor precisión.

Por supuesto que nunca tenemos la certeza de que el verdadero valor del parámetro cae en el intervalo de confianza así construido. Aun cuando contemos dos desvíos a cada lado, habrá una probabilidad de casi 5% de que nuestra muestra no pertenezca al área central, sino a uno de los extremos: las colitas de la distribución normal albergan casi 2,5% de muestras situadas a más de dos desvíos hacia arriba y otro tanto situadas a más de dos desvíos hacia abajo, con respecto a la media de la población. Y si decidiéramos movernos tres desvíos a cada lado en lugar de dos ensancharíamos el intervalo de confianza, pero tampoco podríamos eliminar el riesgo de equivocarnos, aunque lo reduciríamos.[6]

En definitiva, podemos escribir:

$$e = \sigma / \sqrt{n} \quad (1)$$

$$\mu = X + - (z * e) \quad (2)$$

O bien, lo que es lo mismo:

[6] Hay aproximadamente 0,3% de muestras cuyas medias distan en más de tres desvíos de la media de la población. La certidumbre total nunca se alcanza porque la curva normal es infinita y asintótica: nunca corta a la abscisa.

$$\mu = X + - (z * \sigma / \sqrt{n}) \quad (3)$$

El término z, que aparece en la segunda y tercera ecuación como multiplicador de e nos indica el nivel de confianza, es decir la probabilidad con que queremos hacer la estimación: z = 2 para 95,5%, z = 1,96 para 95%, etc. Obviamente, cuando z aumenta obtenemos más seguridad pero se ensancha el intervalo de confianza, por lo que cedemos en precisión. Lo usual es adoptar un valor de z de 1,96 o 2.

¿Y si pretendiéramos mantener el nivel de confianza pero aumentar, a la vez, la precisión?: La ecuación 3 deja claro que el modo de lograrlo será incrementar n: vale decir obtener una muestra de mayor tamaño. Ser pretencioso sale caro...

Claro que en el lector sagaz persistirá una duda. La ecuación 3 incluye varios datos conocidos: la media de la muestra, el valor de z (que depende de la seguridad que queremos otorgar a las estimaciones) y el tamaño de la muestra. Pero nos falta algo: se trata del desvío estándar de la población. Obviamente, lo desconocemos: pero el problema no es grave. Hemos de sustituirlo por el desvío estándar de la muestra, que es un estimador del anterior.[7]

La influencia del tamaño poblacional

En el punto 3 hemos enfatizado que el tamaño de la población no incide en las cuestiones propias del muestreo. Y, en particular, que debe desecharse la idea de que el tamaño de la muestra debe guardar alguna proporcionalidad con el del universo. Esto es efectivamente así cuando se trata de lo que llamamos poblaciones infinitas, que por supuesto no lo son, pero a los efectos del muestreo es como si lo fueran, tal como lo sugieren los ejemplos de la tirada de monedas y la bolsa de fichas brindados en el punto que antecede. Se consideran tales aquellas cuyo tamaño excede de 100 mil casos. Por oposición las que no sobrepasan ese tamaño se denominan poblaciones finitas. Cuando este es el caso, la muestra puede representar una proporción importante en relación con el universo, vale decir, la fracción de muestreo (f = n/N) es relativamente grande.

Cuando así ocurre, la fórmula para la determinación del error muestral lleva agregado un factor de corrección:

$$(N - n) / (N - 1) \quad (4)$$

Si el tamaño de la muestra (n) representa una proporción importante del tamaño poblacional (N), este factor de corrección resultaría ser un número apreciablemente menor que la unidad, con lo que, multiplicado por el error de muestreo opera reduciéndolo. Esto

[7] En rigor, hay una distribución de las varianzas muestrales, cuya media no es exactamente la varianza de la población sino dicha varianza multiplicada por n -1 / n. Puesto que este factor se aproxima más a uno cuanto mayor es la muestra, a medida que el tamaño muestral aumenta las varianzas muestrales tienden a distribuirse normalmente en torno a la varianza poblacional.

suena bastante lógico: si fuera posible tomar una muestra que contuviera la totalidad de los elementos del universo (N = n), obviamente este factor sería igual a cero, por lo que el error de muestreo resultaría anulado. Pero toda vez que la muestra es una proporción muy pequeña del universo, el factor es muy cercano a 1, por lo que resulta desdeñable. Por caso, si tomáramos una muestra de 1000 elementos de un universo de 100 mil (f = 0,01) arrojaría un valor de 0,99.

El caso de las proporciones

Todas las fórmulas que hemos expuesto, así como el ejemplo empleado, suponen que se está procurando estimar una variable cuantitativa, como lo es el ingreso de los hogares, que tiene una media y una varianza. Sin embargo es muy frecuente que las estimaciones más importantes se refieran a variables categóricas y de lo que se trata es de estimar proporciones o porcentajes. Por ejemplo, cuando las encuestas de hogares procuran estimar la tasa de desempleo o cuando los sondeos electorales quieren averiguar el porcentaje de intención de voto por un cierto candidato. En estos casos se trata de saber cuál sería el porcentaje de desocupados en el Gran Buenos Aires, en mayo de 2003, si la EPH realizada en esa fecha arrojó una estimación de 16,4%. O bien, cuál será la verdadera intención de voto por el candidato demócrata John Kerry en el electorado norteamericano, si una encuesta reciente le adjudica 42%.[8]

Cuando de esto se trata, en rigor, las cosas no son mucho más complicadas que en el caso de una variable cuantitativa. Si concebimos aquello cuya proporción pretende estimarse como un atributo dicotómico (vale decir, desempleados/no desempeados; votantes de Kerry/no votantes de Kerry) y adjudicamos valores uno y cero a ambas categorías de la dicotomía,[9] entonces la variable en cuestión se convierte, de hecho, en una variable de intervalos iguales: al haber sólo un intervalo, no puede haber desigualdad. Y la proporción de valores "uno" equivale a la media aritmética de esa variable. Por otra parte, en estos casos, el producto de la proporción que queremos estimar (que se indica con p) por su complementario (que indicamos con q = 1 – p) es equivalente a la varianza de la variable y expresa el grado de homogeneidad o heterogeneidad de la distribución (donde todos los tienen igual valor, p*q = 0).[10]

De esta manera, las fórmulas que anteceden resultan fácilmente adaptables. Por ejemplo, la ecuación 3 quedaría así:

[8] Según ya lo hemos visto, en ambos casos, no obtendríamos una estimación puntual sino un intervalo de confianza dentro del cual se situará la verdadera proporción, con cierta probabilidad.

[9] En estos ejemplos, otorgaríamos el valor uno a los desempleados y a los votantes de Kerry, respectivamente, en tanto que puntuaríamos con cero a quienes no estuvieran desocupados y a los que manifestaran una intención de voto distinta.

[10] Quien quiera persuadirse de ello podrá simular una distribución, por ejemplo, de cinco casos, donde tres de ellos tengan valor 1 y los dos restantes valor cero. La media aritmética sería 3/5 = 0,60. En tanto que la proporción de "unos" sería la misma: 0,60 o 60% si queremos expresarlo en porcentaje. Asimismo, el lector desconfiado podrá calcular la varianza de esta distribución de pocos casos, aplicando las fórmulas convencionales, para comprobar que resultará igual al producto de 0,60 por su complementario 0,40, es decir, 0,24.

$$P = p + - (z * \sqrt{P*Q /n)} \quad (5)$$

En esta ecuación, la P mayúscula es la proporción existente en la población (siendo Q su complementario), en tanto que la p minúscula es la que hemos hallado en la muestra. Como en el caso anterior, tampoco conocemos el valor de P*Q en la población, pero podemos sustituirlo por su estimador, el p*q muestral.

Unos ejemplos y unos ejercicios

A los efectos de clarificar todo lo dicho, brindamos un par de ejemplos muy sencillos. Supongamos que hemos obtenido una muestra de 600 hogares para estimar – entre otras cosas – el ingreso familiar promedio en la Ciudad de Buenos Aires. La muestra arroja una media de $1.458 y un desvío estándar de $1.547,8. ¿cuál sería el intervalo de confianza para la media poblacional, para el 95%? (z = 1,96).

Aplicamos la ecuación 3:

$$\mu = X + - (z * \sigma /\sqrt{n)} = 1.458 + - (1{,}96 + 1.547{,}8 / \sqrt{600}) = 1.458 + - 123{,}8$$

De tal manera, con 95% de probabilidad, diremos que en la población el ingreso medio de las familias variaría aproximadamente entre $ 1334 y $ 1582, que resultan de restar y sumar a la media de la muestra $123.

En la misma muestra podemos determinar que el 36.5% de los hogares de la Ciudad de Buenos Aires tenía por jefa a una mujer. ¿Entre qué límites se hallará el verdadero porcentaje de hogares encabezados por mujeres en la población?

En este caso, se trata de la ecuación 5:

$$P = p + - (z * \sqrt{P*Q /n)} = 36{,}5 + - (1{,}96 * \sqrt{36{,}5*63{,}5/600}) = 36{,}5 + - 3{,}8$$

Con lo cual, el porcentaje oscilaría aproximadamente entre 33% y 40%, con 95% de probabilidad.

5. El tamaño de la muestra (*¿lo adivinamos...?*)

Tal como se ha visto, un factor decisivo, tanto en la precisión como en el grado de seguridad de las estimaciones es el tamaño de la muestra[11]. Esto conduce a pensar que dicho tamaño no se determina en forma caprichosa o azarosa. De lo contrario podría suceder que, tras obtener trabajosamente una muestra de 500 casos, advirtiéramos que resulta insuficiente para nuestras necesidades de precisión. En tal caso, ¿qué haríamos?. ¿Obtendríamos otra muestra más grande, para probar si así resulta adecuada?. Este método de "ensayo y error" resultaría exasperante y antieconómico. Más prudente sería tratar de determinar, antes de extraer la muestra, cuál es el tamaño requerido. Así se procede en la realidad y el tamaño muestral puede ser determinado con fundamento estadístico. Veamos el modo de hacerlo.

Si retornamos a la ecuación 3, veremos que por simple pasaje de términos es posible escribir:

$$\mu - X = (z * \sigma / \sqrt{n}) \quad (6)$$

Y, a partir de la ecuación 5:

$$P - p = (z * \sqrt{P*Q / n}) \quad (7)$$

Luego, se trata, simplemente, de despejar n, de manera que...

$$n = z^2 * \sigma^2 / (\mu - X)^2 \quad (8)$$

$$n = z^2 * p*q / (P - p)^2 \quad (9)$$

Las ecuaciones 8 y 9 permiten, pues, determinar el tamaño muestral a partir de ciertos datos. En el numerador de ambas aparece el término z, vale decir, el nivel de confianza que queremos otorgar a las estimaciones. Obviamente, al pretender mayor confianza requerirá una muestra más grande. Asimismo, tenemos en el numerador la varianza poblacional (o bien el producto p*q que, ya lo hemos visto, es su equivalente en el caso de los atributos dicotómicos). Esto confirma una idea que ya habíamos anticipado: los universos más heterogéneos requieren muestras más grandes. Ello es lógico, puesto que es más difícil captar la diversidad que la uniformidad: si todos los botones son de igual color, para muestra basta un botón... Finalmente, ¿qué cosa tenemos en el denominador del cociente?. Lo que allí aparece – las diferencias *(P-p)* o *(μ - X), según el caso* – son los máximos errores de estimación que estamos dispuestos a tolerar. Vale decir, cuando queremos estimar la media de los ingresos familiares, ¿cuántos pesos de error estamos

[11] Una mirada a la fórmula del error muestral nos persuadirá de ello, porque el n aparece en el denominador del cociente.

dispuestos a aceptar? Si pretendemos una estimación con un error no mayor de 50 pesos, entonces ese será el valor que colocaremos (debidamente elevado al cuadrado) en el denominador. Igualmente, ¿por cuántos puntos porcentuales admitiremos equivocarnos al estimar la tasa de desempleo?. Si no queremos un error de más de dos puntos porcentuales, entonces P-p será igual a dos. Obviamente, una estimación más precisa – un error más pequeño – pedirá más casos muestrales.

Persiste, sin embargo, un pequeño problema: las fórmulas incluyen las expresiones de la varianza (o el producto p*q) en la población, que usualmente desconocemos. En el punto anterior, hemos visto que reemplazábamos esto por sus estimadores, vale decir, por los valores hallados en la muestra. Sin embargo, ahora no podemos proceder así, porque si estamos tratando de definir el tamaño de la muestra, ello quiere decir que aún no contamos con ella. ¿Qué haremos?. Existen aquí varias alternativas:

- en primer lugar, es probable que exista un antecedente cercano de un estudio similar. Por ejemplo, las encuestas de hogares se realizan periódicamente en las mismas ciudades y los sondeos electorales se replican cada pocos días al aproximarse la fecha de los comicios. Si así fuera, ese estudio nos proveería una razonable estimación de la varianza (o de p*q).
- en segunda instancia, si no hay tales antecedentes, casi todas las encuestas van precedidas de una prueba piloto o *pretest* destinado a poner a prueba el cuestionario. Para estas pruebas se suele emplear una muestra pequeña (50 o 100 casos) tomada de la misma población que se estudiará. Si tomamos el recaudo de elegir, deliberadamente, los casos del pretest introduciendo deliberada heterogeneidad[12], podríamos confiar en que la muestra del pretest tendría una varianza mayor que la de la población. Y si fuera así, podríamos emplear esa varianza para la fijación del tamaño muestral sin temor alguno a "quedarnos cortos". En todo caso, el perjuicio sería económico, porque podríamos gastar dinero en una muestra más grande que lo necesario.
- Por fin, queda otra alternativa que nos cubre de cualquier contingencia, aunque no resulta menos gravosa. El producto p * q alcanza su máxima expresión cuando la distribución se divide por mitades: 50% de los casos con valor cero y 50% con valor uno. Cuando esto ocurre, p * q = 50 * 50 = 2.500. Si nos basamos en este supuesto de máxima heterogeneidad de la distribución poblacional, entonces todo el riesgo estribará en extraer una muestra de tamaño excesivo, pero jamás insuficiente. Si no tenemos idea alguna acerca de la varianza del universo, podemos emplear este criterio conservador para fijar el tamaño muestral.

Unos ejemplos y unos ejercicios

Vamos a obtener una muestra que nos permita estimar el ingreso promedio de los hogares de la Ciudad de Buenos Aires con un error no mayor de + - 100 pesos y un nivel de

[12] Por ejemplo, en una encuesta de opinión podría incluirse en el pretest una mezcla deliberadamente heterogénea de personas de niveles socioeconómicos diversos, bajo el supuesto de que ello incrementaría la heterogeneidad de las respuestas.

confianza de 95%. Disponemos, a partir de la última onda de la EPH, de una estimación de la varianza de esta variable, que asciende a 2.395.685.

$n = z^2 * \sigma^2 / (\mu - X)^2 = 1{,}96^2 * 2.395.685/100^2 = 920$

Supongamos que, en la misma encuesta, se quiera estimar el porcentaje de hogares con jefatura femenina. En este caso, queremos que el error no exceda de + - 3 puntos porcentuales. No tenemos idea del valor de p * q en la población y decidimos usar el criterio de máxima heterogeneidad, para no correr riesgos. En este caso:

$n = z^2 * p*q / (\mu - X)^2 = 1{,}96^2 * 2.500 / 2^2 = 1.067$

Según se aprecia, los tamaños requeridos no serían muy discrepantes. En este caso, debiéramos optar por el mayor de ambos, que nos alcanzaría para la segunda estimación y sobraría para la primera.

El caso de los universos finitos

Ya vimos, al ocuparnos del error de estimación, que cuando trabajamos con lo que se denomina universos finitos (menos de 100 mil casos), el error se achica a medida que crece la fracción muestral (es decir, cuando la muestra representa una proporción importante del total de elementos incluidos en el universo). En estos casos, así como incluíamos un factor de corrección en la estimación del error, también hemos de modificar la fórmula destinada a determinar el tamaño de la muestra. La fórmula e emplear en estos casos sería:

$$\frac{Z^2 \text{ x } PQ \text{ x}(N-1)}{E^2 \text{ x } [(N-1) + (Z^2 \text{ x } PQ)]}$$

(esta fórmula tiene agregado el factor de corrección)

Donde:

Z: nivel de significación (para 95% = 1,96)

PQ: estimador de la varianza poblacional (para estimación de proporciones)

N: tamaño de la población

E: error aceptado en la estimación de un parámetro poblacional

Regla empírica de las celdas

Más allá de la determinación del tamaño de la muestra con fundamentos estadísticos, que hemos expuesto en esta última parte, existe un criterio más empírico que complementa al anterior. Consiste en contemplar un mínimo de casos muestrales por celda, que hagan posible un análisis razonable. En tal sentido, suele afirmarse que, teniendo en cuenta el cuadro de mayores dimensiones que se proyecte realizar, se contemple un mínimo de 20 casos muestrales promedio por cada celda, más un adicional de 20%. De suerte que, si tuviéramos una variable de tramos etarios de cinco categorías y una variable de nivel

educativo que tenga siete, el cruce de ambas generaría un cuadro de cuarenta celdas. Si multiplicamos cuarenta celdas por veinte casos obtenemos 800. Si a ello le sumamos un 20% adicional llegaríamos a un tamaño muestral de 960 casos mínimos.

6. El caso de las muestras estratificadas

Todas las fórmulas que hemos expuesto sirven, en principio, para el procedimiento más básico de muestreo aleatorio: la muestra al azar simple con reposición, donde todos los elementos del universo tienen igual probabilidad de ser seleccionados[13]. Pero hay otras modalidades de muestreo al azar, de las que hemos de ocuparnos. Una de ellas es el muestreo *estratificado*, donde, antes de seleccionar la muestra, se subdivide el universo en *estratos* o *subuniversos*. Para ello, debemos disponer de lo que se denomina un marco muestral: vale decir, un listado de los elementos que componen la población.

Por ejemplo, si fuera necesario seleccionar una muestra de alumnos de la facultad de ciencias sociales, con el propósito de entrevistarlos y averiguar sus opiniones y expectativas, podríamos decidir estratificar por carrera. De manera que elegiríamos submuestras separadas de alumnos de cada carrera: bastaría para ello, con disponer de las listas de alumnos y seleccionaríamos al azar simple o sistemático de cada una de estas listas.

En otro caso, si fuera preciso seleccionar una muestra de escuelas de la Ciudad de Buenos Aires para evaluar el rendimiento de los alumnos, antes de obtenerla podríamos estratificar la ciudad con criterios geográficos: por ejemplo, un estrato integrado por los barrios de la zona sur, otro por los barrios de la zona norte y un tercero – más grande – con los barrios centrales. En este caso, confeccionaríamos listados separados de las escuelas situadas en los distritos escolares de las tres zonas y, sobre estos listados obtendríamos las respectivas submuestras, seleccionando escuelas al azar.

Hay que notar que uno no elige arbitrariamente la variable estratificadora: en el primer caso, probablemente elegimos la carrera porque suponemos que las opiniones de los alumnos pueden diferenciarse significativamente en función de este criterio. En el segundo ejemplo, hemos elegido estratificar por zonas porque sabemos que estas zonas difieren en su conformación socieconómica y es probable que esto influya en los rendimientos escolares de los alumnos.

Por qué estratificamos

¿Cuáles son las razones por las que se suele apelar a la estratificación?. Un buen motivo puede ser asegurarnos de que en nuestra muestra habrá suficientes elementos de cada uno de estos estratos. Si bien el azar suele garantizar que así suceda, podría ocurrir que alguno de los estratos tenga un escaso peso en la población y, sin embargo, importe desde el punto de vista del análisis. Por ejemplo, alguna de las carreras que se dictan en la facultad podría tener un alumnado muy escaso: correríamos el riesgo de que en la muestra no

[13] Una muestra con reposición implica que, cada vez que un elemento es seleccionado, se lo vuelve a incluir antes de realizar la próxima selección (de modo que, teóricamente, podría volver a ser seleccionado). Si no se lo hiciera así, en el caso de que fuéramos a elegir *n* casos sobre una población de 1.000, el primer elemento elegido habría tenido una probabilidad de ser seleccionado de 1/1.000 = 0,001. Pero el segundo tendría una probabilidad algo mayor: 1/999, mientras que el tercero sería elegido con probabilidad 1/998, etc. Aunque en la práctica no suele emplearse muestreo con reposición, las variaciones son, sin embargo, desdeñables.

apareciera ningún alumno o bien que fueran seleccionados muy pocos. Lo mismo podría suceder si alguno de los estratos definidos al interior de la ciudad fuera muy pequeño.

Por esta misma razón, puede ocurrir que sea necesario otorgarle a los estratos un peso diferente en la muestra del que tienen en la población. A esto se lo denomina muestreo estratificado no proporcional. Vale decir, supongamos que hemos dividido una población que consta de 100 mil elementos (N = 100.000) en dos estratos. El primero contiene el 90% de los casos (N1 = 90.000) en tanto que el restante sólo alberga al 10% (N2 = 10.000). Si debemos obtener una muestra de mil casos (n = 1.000), la fracción de muestro general sería:

f = n/N = 1.000/100.000 = 0,01

Supongamos también que decidimos obtener una muestra estratificada. ¿Cómo distribuimos los casos entre los estratos?[14]. La primera alternativa consistiría en obtener una muestra estratificada proporcional, manteniendo iguales fracciones de muestreo. En ese caso, seleccionaríamos 900 casos del primer estrato y sólo 100 del segundo:

$f_1 = n_1/N_1 = 900/90.000 = 0,01$

$f_2 = n_2/N_2 = 100/10.000 = 0,01$

Pero tal vez pensemos que sólo 100 casos son pocos para llevar a cabo el análisis. Entonces, podríamos asignar los casos a los estratos en forma no proporcional: por ejemplo, la mitad a cada estrato. Si ello fuera así, tendríamos fracciones muestrales diferentes: mayor en el estrato pequeño:

$f_1 = n_1/N_1 = 500/90.000 = 0,0055$

$f_2 = n_2/N_2 = 500/10.000 = 0,05$

Pero hay, también, una segunda razón – algo más complicada – que podría inducirnos a emplear el muestreo no proporcional. Hemos visto que al determinar el tamaño de la muestra juegan tres factores: a) la seguridad que queremos conferir a nuestras estimaciones (por ejemplo, 95%), b) el error de estimación que estamos dispuestos a tolerar (por ejemplo, +-$10 o +- 3%) y c) el grado de heterogeneidad en la distribución poblacional de la variable que queremos estimar, que se expresa en la varianza o en el producto p*q. Los primeros dos factores dependen de decisiones del muestrista, pero el tercero nos es impuesto: la distribución de los ingresos familiares o de los jefes de hogar por sexo es como es. Si la población es muy heterogénea, ya lo hemos visto, nos pedirá una muestra muy grande.

Sin embargo, podemos hacer algo al respecto: supongamos que sabemos que los tres estratos de la ciudad son heterogéneos entre sí pero relativamente homogéneos al interior. Vale decir, la mayor parte de lo hogares del estrato norte poseen ingresos altos y relativamente parecidos, en tanto que en el estrato sur hay hogares homogéneamente pobres. Finalmente, en el estrato intermedio, hay más heterogeneidad, porque convive una mezcla de hogares de diferente nivel socioeconómico. De suceder así, no habría motivos para que adjudicáramos iguales fracciones muestrales, respetando en la muestra el mismo peso que los estratos tienen en la población (a los efectos de este ejemplo, para simplificar,

[14] La distribución de los casos muestrales entre los estratos se denomina *afijación*.

supondremos que en la población los tres estratos tienen igual tamaño). Un uso óptimo de los casos muestrales, que obtuviera de ellos el máximo rendimiento posible, sugeriría asignar muchos al estrato heterogéneo y pocos a los más homogéneos.

Para una mejor comprensión de esto, podemos volver a nuestro ejemplo inicial de la gran bolsa que contenía bolillas negras y blancas. Supongamos que, al interior de la bolsa grande, hubiese tres bolsas pequeñas, que contuvieran un tercio del total. En las dos primeras hay fichas de un solo color, mientras que en la última las bolillas están mezcladas: si tuviéramos que estimar qué proporción de fichas de cada color hay en total extrayendo una muestra de treinta fichas, ¿tomaríamos diez de cada bolsa?. Evidentemente, no tendría sentido: en las bolsas monocromáticas bastaría con una sola ficha para saber de qué color son las demás. En cambio, sería bueno destinar las 28 restantes a la bolsa mezclada.

Para decirlo de una vez, allí donde exista una gran diversidad, difícil de captar, destinaremos más casos. Donde hay una realidad homogénea, en cambio, necesitaremos menos. Pero, ¿cuánto más o cuánto menos?. La fórmula de *Neymann*, de afijación óptima, nos resolverá la cuestión:

$$n_e = (N_e * \sigma_e) / \sum(N_e * \sigma_e) * n \quad (10)$$

Donde:

n_e*: tamaño de un estrato en la muestra*

N_e*: tamaño de un estrato en la población*

σ_e*: desvío estándar de un estrato en la población*

n *= tamaño muestral total*

Con esta fórmula se logra que el tamaño muestral de cada estrato quede determinado, a la vez, por el peso que el estrato tiene en la población y por su grado de heterogeneidad. Como podemos ver, la fórmula de Neymann incluye el desvío estándar de cada uno los estratos en la población (que, en el caso de atributos, se reemplaza por la raíz cuadrada de p*q). Lo normal es que se carezca de esta información pero, al menos, deberemos contar con una razonable estimación. Si no la tuviéramos (es decir, si en realidad ignoramos el grado de homogeneidad o heterogeneidad de los estratos), entonces no podremos apelar a la asignación óptima de los casos muestrales y lo más sensato será, seguramente, estratificar en forma proporcional.

Ejemplo:

Teniendo en cuenta el tamaño de los estratos, si asignáramos proporcionalmente los casos muestrales, otorgaríamos 200 al 1°, 300 al 2° y 500 al 3°. Sin embargo, el estrato de mayor tamaño es el más homogéneo, al tiempo que el intermedio resulta ser el de mayor heterogeneidad interna. Aplicando la fórmula de la afijación óptima, resultan unos tamaños muestrales algo diferentes:

Estratos	N_e	$\sqrt{p*q_e}$	$N_e*(p*q)_e$
1	200.000	42	8.485.281
2	300.000	50	15.000.000
3	500.000	40	20.000.000
	1.000.000		43.485.281

n total = 1.000
Así, para los tres estratos:
n_1 = 8.485.281/43.485.281*1.000= 195
n_2 = 15.000.000/43.485.281*1.000= 345
n_3 = 20.000.000/43.485.281*1.000= 460

Como consecuencia de la aplicación de este procedimiento, los estratos más homogéneos resultan subrepresentados en la muestra, en tanto que hemos asignado relativamente más casos al más heterogéneo.

El error en el muestreo estratificado

Aun cuando no se conozcan las dispersiones de los estratos, si tenemos razones para suponer que ellos son internamente más homogéneos que la población considerada en su conjunto, valdrá la pena estratificar, porque ello reducirá el error de estimación que – como ya lo hemos visto – crece a medida que aumenta el error estándar de la población. En este caso, la estratificación habrá de redundar en una estimación sujeta a un menor error. Toda vez que se haya aplicado el muestreo estratificado – proporcional o no proporcional – la estimación del error de muestreo habrá de realizarse aplicando la siguiente fórmula:

$$e = \sqrt{\sum(N^2_e / N^2) * (\sigma^2_e / n_e)} \quad (11)$$

Donde:

e: error muestral

N^2_e: tamaño poblacional de los estratos

N^2: tamaño de la población

σ^2_e: varianza de los estratos en la población (se sustituye por los correspondientes estimadores muestrales y por p*q en caso de atributos)

n_e: tamaño muestral de los estratos

Ejemplo:

Nos valemos nuevamente de los datos del ejemplo anterior:

Estratos	N^2	(1) N^2_e/N^2	$p*q_e$	n_e	(2) $p*q_e/n_e$	(1) * (2)
1	40.000.000.000	0,04	1.800	195	9,2	0,4
2	90.000.000.000	0,09	2.500	345	7,2	0,7
3	250.000.000.000	0,25	1.600	460	3,5	0,9
∑	1.000.000.000.000					1,9

El error de muestreo sería la raíz cuadrada de la sumatoria que aparece en la última columna de la tabla = 1,4. Vale decir, el error de muestreo sería +- 1,4 puntos porcentuales, en la estimación de un atributo.

La ponderación posterior en el muestreo no proporcional

A esta altura, alguien podría haber advertido – con gran sagacidad – que, toda vez que hubiéramos alterado en la muestra el peso que realmente tienen los estratos en la población, estaríamos distorsionando los datos. Por ejemplo, nuestra muestra de universitarios podrá incluir una proporción exagerada de alumnos de cierta carrera o la de escuelas podría concentrar un exceso de las situadas en ciertas zonas de la ciudad, ya que utilizamos fracciones de muestreo diferentes.

Efectivamente, así sería. Pero la solución a este problema no es difícil y se logra a través de los factores de expansión: ¿en qué consisten? En primer lugar, muchas veces, además de estimar promedio o porcentajes, nos interesa conocer la cantidad aproximada de elementos que, en la población, cumplen con ciertas condiciones. Si hubiéramos empleado una muestra al azar simple, con una única fracción de muestreo (por ejemplo, f = 1000 / 300.000 = 0,0033), entonces bastaría con multiplicar todos los resultados por la inversa de dicha fracción: 1/f = 300. Así, por caso, 50 mujeres en situación de inactividad halladas en la muestra equivaldrían a 15 mil mujeres inactivas en la población.

Pero si las fracciones de muestreo empleadas en un muestreo estratificado difieren, entonces los casos de cada estrato deberán ser expandidos aplicando la inversa de la fracción muestral empleada en cada uno de dichos estratos. En la práctica, lo que se hace usualmente es crear en la base de datos (por ejemplo, en una base de datos de SPSS) una variable de expansión, que sea igual a la inversa de la fracción de muestreo del estrato a que corresponde cada caso. Previamente a realizar el procesamiento de los datos, se pondera el archivo mediante esta variable[15].

[15] El programa SPSS dispone de un comando que permite ponderar el archivo.

7. El muestreo por conglomerados

Hemos visto que, en el caso de la estratificación, lo ideal es dar con una variable capaz de subdividir el universo en partes (estratos) que sean internamente homogéneos pero diferentes entre sí. Hecho esto, se seleccionan al azar casos dentro de cada estrato. Ahora nos ocuparemos de otro procedimiento de muestreo, donde el propósito es, en alguna medida, inverso: se trata del muestreo de conglomerados.

Vamos a suponer que deseamos conocer algunas características de los alumnos que cursan el ciclo primario en la Ciudad de Buenos Aires, para lo cual necesitamos aplicarles un cuestionario. Podríamos apelar a los registro de la Secretaría de Educación de esa jurisdicción y extraer una muestra seleccionando al azar simple o sistemático a los alumnos: en los mismos registros (que serían nuestro marco muestral) encontraríamos la información necesaria para localizarlos. Sin embargo, esto resultaría muy trabajoso y es posible aplicar un procedimiento más práctico. En nuestro universo, los elementos (los alumnos) están naturalmente agrupados en unos conjuntos o conglomerados (las escuelas). Y resultaría muy sencillo contar con un listado de escuelas. ¿Por qué no seleccionar al azar (simple o sistemático) cierta cantidad de estos conglomerados?: por ejemplo, una décima parte de ellos?

Una vez seleccionadas las escuelas, sería posible aplicar la encuesta a la totalidad de los alumnos, en cada una de ellas. Habríamos llegado a los alumnos a través de los conglomerados que los agrupan.

Esta es una muestra por conglomerados de etapa única: hemos seleccionado al azar sólo una vez. La eficacia de este tipo de muestras depende de dos factores. En primer lugar, de la relación m/M, donde m es la cantidad de conglomerados seleccionados y M es la cantidad existente en el universo. Cuanto mayor es esta relación, menor será el error de muestreo: obviamente, si seleccionáramos la totalidad de los conglomerados no habría error alguno. En segundo término, la muestra será tanto mejor cuanto más se parezcan los conglomerados entre sí: si fueran muy semejantes unos a otros, perderíamos muy poco al seleccionar sólo algunos para incluir en la muestra. Otra vez, vale emplear un razonamiento "por el absurdo": si todos los conglomerados fueran idénticos entre sí, bastaría con quedarse con uno solo. De manera que, al revés de lo que ocurría con los estratos, aquí el ideal consistiría en que hubiera una gran homogeneidad interconglomerados (similares entre sí) y una amplia heterogeneidad intraconglomerados (que toda la diversidad del universo quedara representada al interior de cada uno). En otros términos, que cada conglomerado fuera "un universo en pequeño".

El error de muestreo en el caso de los conglomerados de etapa única se estima mediante la siguiente fórmula:

$$\mathbf{e} = \sqrt{\sigma^2_m / m * (1 - m/M)} \quad (12)$$

Donde:

e: error muestral

σ^2_m : varianza de las medias (o de las proporciones) entre los conglomerados

m = cantidad de conglomerados en la muestra

M = cantidad de conglomerados en el universo

Según se advierte, si los conglomerados difieren mucho entre sí (con lo que aumentará la varianza de sus medias), crece el error de muestreo. En cambio, disminuye cuanto mayor es la razón m/M (cuantos más conglomerados integran la muestra).

Ejemplo:

Supongamos que elegimos 10 escuelas (m) sobre un total de 50 (M). En cada una de estas escuelas aplicamos una encuesta que permite determinar que hay los siguientes porcentajes de alumnos provenientes de hogares en situación de pobreza, acreedores a becas:

22%, 8%, 34%, 40%, 12%, 15%, 25%, 52%, 64%, 39%.

Si calculamos la varianza de esta variable (de los diez porcentajes), obtenemos un valor de 327,4 (para lo cual, calcularíamos antes la media de estos porcentajes, que es 31,1, restaríamos esta media a cada porcentaje, elevaríamos los residuos al cuadrado y dividiríamos ese valor por diez). Luego, sustituyendo en la fórmula 12:

e = √327,4 / 10 * (1 – 10 /50) = 5,1

En definitiva, el error de le estimación del porcentaje de alumnos pobres en la población sería +- 5,1 puntos porcentuales.

Conglomerados polietápicos

Hemos dejado para la parte final un tipo de muestreo que es, a la vez, el más complejo y uno de los más frecuentemente utilizados. Se trata del muestreo por conglomerados de múltiples etapas o polietápico. Es el tipo de muestreo del que suelen valerse las encuestas de hogares, así como las de opinión y los sondeos electorales.

Supongamos que queremos preguntar a las personas de 18 años y más (habilitadas para votar) que residen en la Ciudad de Buenos Aires su opinión acerca del desempeño del gobierno de la jurisdicción (o sobre cualquier otro tópico). ¿De dónde sacaríamos el marco muestral, es decir un listado con los datos de todos los habitantes de la ciudad. No existe: no podríamos disponer de tal listado. Pero podríamos tratar de dar con las personas dentro de los conglomerados que los agrupan: los hogares. Sin embargo, tampoco tenemos un listado de hogares: ni siquiera uno actualizado de viviendas. ¿Qué podemos hacer?

El territorio de cualquier ciudad está naturalmente dividido en jurisdicciones administrativas. Por ejemplo, las fracciones censales, que son grandes jurisdicciones geográficas al interior de la ciudad. Pues bien, podría seleccionarse al azar algunas de estas fracciones. A su vez, las fracciones están divididas en áreas menores, que se denominan radios censales. En un segundo paso o etapa, sería posible seleccionar al azar cierta cantidad de radios al interior de cada una de las fracciones que "sobrevivieron" al primer sorteo. Finalmente, tendríamos algunos radios de ciertas fracciones. Y dentro de estos radios, tendremos manzanas, que apelando a la cartografía (o a una buena guía

Filcar, en el peor de los casos...) podrían ser numeradas y seleccionadas al azar. Estas manzanas sobrevivientes a los tres sorteos se denominan, habitualmente, puntos muestra. Dependiendo del total de hogares que queremos seleccionar (es decir del *n* muestral), suele determinarse previamente cuántos puntos muestra hemos de requerir. Por ejemplo, si nuestra muestra total constara de 600 personas (no más de una por vivienda), se podría seleccionar un centenar de puntos muestra, de manera que se encuestarían unas seis viviendas por cada manzana.

A esta altura, ya estaríamos cerca de las personas, puesto que las hallaríamos dentro de sus viviendas. Para seleccionar seis viviendas en cada manzana, podríamos recorrer previamente las manzanas elegidas, para hacer un conteo. Supongamos que en una manzana identificáramos aproximadamente 60 viviendas: pues bien, en ese caso instruiríamos al encuestador para que, partiendo de cierta esquina predeterminada y avanzando en el sentido de las agujas del reloj, fuera tocando el timbre en una de cada diez.[16]

En las encuestas domiciliarias debe preverse un porcentaje considerable de rechazos (personas que se niegan a ser entrevistadas). Generalmente, este margen de rechazos se conoce por experiencia y puede ser estimado. Para compensar, es posible seleccionar más puntos muestra de los necesarios, a los efectos de los posibles reemplazos.

No escapará a la sagacidad del lector que al proceder de este modo no hemos hecho una sino varias selecciones al azar. En cada una de ellas habrá una distribución de muestreo y, por lo tanto, estará presente el correspondiente error. Puesto que los errores se adicionan, el error final resultará considerablemente mayor. Las fórmulas que permiten estimarlo son sumamente complejas (existe más de un procedimiento) y desbordan con mucho esta breve exposición. Algunos textos, sin embargo[17] sugieren una solución sencilla para un problema complejo: determinado el tamaño muestral como si se tratara de una muestra al azar simple, debiéramos multiplicarlo por 1,5 si fuéramos a obtener una muestra de etapas múltiples.

Selección con probabilidad proporcional al tamaño

Ya llegando al final, hemos de señalar que este tipo de muestreo vulnera totalmente el principio de igualdad de probabilidades de ser seleccionado. Por de pronto, quienes habiten en una fracción o en un radio que no resulta elegido, ya no podrán caer en la muestra. Pero hay algo más grave. Supongamos que elegimos al azar simple o sistemático las fracciones censales. Obviamente, estas no tienen el mismo tamaño: habrá fracciones que concentren mucha más población que otras. Pues bien, una vez elegidas, quien viva en una fracción muy densa, por ejemplo con mil viviendas, tendrá una menor probabilidad de ser elegido que el que habita en una fracción donde sólo hay 200 viviendas. En el

[16] Por cierto que los edificios de departamentos suponen un problema adicional, aunque no difícil de solucionar en teoría: es posible indicar una pauta a seguir en esos casos. Siempre y cuando no fuéramos expulsados por personal de vigilancia...
[17] Seguramente basados en la experiencia.

primer caso la probabilidad es de 1/1000 = 0,001, mientras que en el segundo es de 1/200 = 0,005.

Esta menor probabilidad de ser elegido puede compensarse en la etapa previa, al seleccionar las fracciones, empleando el sistema de selección proporcional al tamaño. De esta manera, la fracción más densa tendrá mayor probabilidad de caer en la muestra que la de menor tamaño, compensando así la "desventaja" que afrontarán sus habitantes. La tabla que sigue ilustra el modo en que se seleccionan:

Fracción	Tamaño	%	% acum.	Intervalo de selección
F1	700	38,9	38,9	1 - 39
F2	500	27,8	66,7	40 -67
F3	400	22,2	88,9	68 - 89
F4	200	11,1	100,0	90 - 100
Total	1800	100,0		

Si tuviéramos que elegir una fracción entre estas cuatro, seleccionaríamos números al azar entre 1 y 100. Si sale del 1 al 39, elegimos la primera, si sale del 40 al 67, la segunda, etc. Obviamente, la primera tendrá mayor probabilidad de ser seleccionada.

BREVE HISTORIA DEL IDICSO

Los orígenes del IDICSO se remontan a 1970, cuando se crea el "Proyecto de Estudio sobre la Ciencia Latinoamericana (ECLA)" que, por una Resolución Rectoral (21/MAY/1973), adquiere rango de Instituto en 1973. Desde ese entonces y hasta 1981, se desarrolla una ininterrumpida labor de investigación, capacitación y asistencia técnica en la que se destacan: estudios acerca de la relación entre el sistema científico-tecnológico y el sector productivo, estudios acerca de la productividad de las organizaciones científicas y evaluación de proyectos, estudios sobre política y planificación científico tecnológica y estudios sobre innovación y cambio tecnológico en empresas. Las actividades de investigación en esta etapa se reflejan en la nómina de publicaciones de la "Serie ECLA" (SECLA). Este instituto pasa a depender orgánica y funcionalmente de la Facultad de Ciencias Sociales a partir del 19 de Noviembre de 1981, cambiando su denominación por la de Instituto de Investigación en Ciencias Sociales (IDICSO) el 28 de Junio de 1982.

Los fundamentos de la creación del IDICSO se encuentran en la necesidad de:

- Desarrollar la investigación pura y aplicada en Ciencias Sociales.
- Contribuir a través de la investigación científica al conocimiento y solución de los problemas de la sociedad contemporánea.
- Favorecer la labor interdisciplinaria en el campo de las Ciencias Sociales.
- Vincular efectivamente la actividad docente con la de investigación en el ámbito de la facultad, promoviendo la formación como investigadores, tanto de docentes como de alumnos.
- Realizar actividades de investigación aplicada y de asistencia técnica que permitan establecer lazos con la comunidad.

A partir de 1983 y hasta 1987 se desarrollan actividades de investigación y extensión en relación con la temática de la integración latinoamericana como consecuencia de la incorporación al IDICSO del Instituto de Hispanoamérica perteneciente a la Universidad del Salvador. Asimismo, en este período el IDICSO desarrolló una intensa labor en la docencia de post-grado, particularmente en los Doctorados en Ciencia Política y en Relaciones Internacionales que se dictan en la Facultad de Ciencias Sociales. Desde 1989 y hasta el año 2001, se suman investigaciones en otras áreas de la Sociología y la Ciencia Política que se reflejan en las series "Papeles" (SPI) e "Investigaciones" (SII) del IDICSO. Asimismo, se llevan a cabo actividades de asesoramiento y consultoría con organismos públicos y privados. Sumándose a partir del año 2003 la "Serie Documentos de Trabajo" (SDTI).

La investigación constituye un componente indispensable de la actividad universitaria. En la presente etapa, el IDICSO se propone no sólo continuar con las líneas de investigación existentes sino también incorporar otras con el propósito de dar cuenta de la diversidad disciplinaria, teórica y metodológica de la Facultad de Ciencias Sociales. En este sentido, las áreas de investigación del IDICSO constituyen ámbitos de articulación de la docencia y la investigación así como de realización de tesis de grado y post-grado. En su carácter de Instituto de Investigación de la Facultad de Ciencias Sociales de la Universidad del Salvador, el IDICSO atiende asimismo demandas institucionales de organismos públicos, privados y del tercer sector en proyectos de investigación y asistencia técnica.

ÁREAS DE INVESTIGACIÓN DEL IDICSO		
❑ Desarrollo Social Local y Regional	❑ Organizaciones No Gubernamentales y Políticas Públicas	❑ Empleo y Población
❑ Recursos Energéticos y Planificación	❑ Relaciones Internacionales de América Latina	❑ Relaciones Internacionales de Asia y el Pacífico
❑ Gobernabilidad y Reforma Política	❑ Historia Cultural y Social Contemporánea	❑ Historia de las Relaciones Internacionales en el Mundo Antiguo y Medieval
❑ Sociedad, Estado y Religión en América Latina	❑ Relaciones Iglesia-Estados	❑ Migraciones y Derechos Humanos

IDICSO

Instituto de Investigación en Ciencias Sociales
Facultad de Ciencias Sociales
Universidad del Salvador

SERIE MATERIALES DE ÁREAS DE INVESTIGACIÓN DEL IDICSO
Edición y corrección: ***Ricardo De Dicco*****, Departamento de Comunicación y Tecnología del IDICSO**
Tel/Fax: (++5411) 4952-1403
Email: idicso@yahoo.com.ar
Sitio Web: http://www.salvador.edu.ar/csoc/idicso

Hipólito Yrigoyen 2441
C1089AAU Ciudad de Buenos Aires
República Argentina