manual metodologia 22/5/08 15:36 Página 1

La investigación en Ciencias Sociales: lógicas, métodos y técnicas para abordar la realidad social

Horacio Chitarroni (coordinador)
Stella Maris Aguirre
Mariana Colotta
Valeria Coniglio
Lucía Destro
Verónica Diyarian
Viviana Escanes
Cecilia Maestro

manual metodologia 22/5/08 15:36 Página 225

10. LA SELECCIÓN DE LA EVIDENCIA EMPÍRICA.
UNIVERSOS Y MUESTRAS.

Horacio Chitarroni

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 (Chein, 1980).

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 (Mayntz et al, 1975; Kish, 1993).

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.

Las muestras pueden obtenerse de diferentes manera, en función de los propósitos a los que sirven. Pero existe una gruesa distinción entre dos tipos claramente diferentes:

Las muestra probabilísticas o aleatorias, que son aquellas en cuya selección entran procedimientos de azar. A ellas nos referiremos en primer término.

Las muestras no probabilísticas, o sea aquellas en las que los elementos que las componen son elegidos por procedimientos que no incluyen el azar. Estas últimas se tratarán en una sección posterior de este capítulo.

manual metodologia 22/5/08 15:36 Página 226

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. 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 subregistro 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 (Kish, 1993; Slonim, 1974).

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 prome-

manual metodologia 22/5/08 15:36 Página 227

dio 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 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 97 fósforos (porque una de ellas contenía 96 y la otra 98), ¿en cuánto nos estaremos equivocando al decir que hay 970 en todo el *pack*?. Pues bien, las técnicas de muestreo aleatorio tienden a procurar estos objetivos.

3. Las muestras probabilísticas o aleatorias

No es lo menos frecuente que –como en el ejemplo de los fósforos– necesitemos 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.

Como ya se dijo, en el muestreo probabilístico los elementos de la población que integrarán la muestra son seleccionados por procedimientos de azar. Y, al menos en el momento inicial, todos ellos tienen alguna probabilidad de ser seleccionados, que no siempre ni

manual metodologia 22/5/08 15:36 Página 228

necesariamente ha de ser igual para todos. Esta probabilidad puede ser conocida y cuantificada (García Ferrando, 1985). Para que esto sea posible, en el momento de seleccionar la muestra debieran estar presentes –de manera real o virtual– todos los elementos que componen la población.

Para plantearlo de un modo simple, debiéramos tener, por ejemplo, un listado que contuviera todos los elementos que componen el universo, para poder elegir por sorteo cierta cantidad de ellos. Este listado, que se denomina marco muestral, no siempre existe ni es fácilmente accesible (Kish, 1993). Ya veremos que algunas formas de muestreo están indicadas cuando no tenemos un listado de esta clase: por ejemplo las muestras de conglomerados polietápicos, a las que se refiere el punto *3.4.5.*

Y aun cuando el listado exista, suele suceder que no incluya la totalidad de los elementos: por ejemplo, por estar desactualizado. Esto sucedería su usáramos el padrón electoral como marco para una muestra de personas en edad de votar: algunos de los que han alcanzado la edad no están, todavía, incorporados, en tanto que otros que sí figuran han fallecido y aún no se les dio de baja. Se trataría de un marco muestral defectuoso: faltan unos y sobran otros. Si quisiéramos disponer de una lista de todas las viviendas de una ciudad, para obtener una muestra de ellas, podríamos acudir a un registro del catastro municipal. Pero sucedería lo mismo: algunas de las viviendas que figuran habrán sido demolidas, en tanto que se habrán construido otras que, tal vez, no estén todavía incorporadas. Casi todos los marcos muestrales tienen ciertos defectos o imperfecciones, que deben tenerse en cuenta al analizar los resultados derivados del muestreo (Kish, 1993).

A veces, es posible introducir algunas correcciones al respecto: valiéndonos del último ejemplo, podríamos tener conocimiento de que en los barrios más pobres de la ciudad se construyen viviendas precarias que no siempre se inscriben en el registro catastral. De modo que el marco muestral no las registra adecuadamente y –por eso– tampoco lo haría la muestra. Si así fuera, podríamos precavernos otorgando a las pocas viviendas de esa clase que sí están en el listado una probabilidad de integrar la muestra mayor que al resto: ya veremos que hay diseños muestrales que contemplan esta posibilidad: es el caso de las muestras estratificadas, que se consideran en el punto *3.4.3.*

Por cierto, nunca podemos estar seguros de que lo que averiguamos en una muestra será 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 (Sierra Bravo, 1995). 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 supe-

manual metodologia 22/5/08 15:36 Página 229

rior 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 (Padua, 1979).

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.

3.1 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 1/2, 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 51%). 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

manual metodologia 22/5/08 15:36 Página 230

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 vamos 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.

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 garantiza 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

manual metodologia 22/5/08 15:36 Página 231

bolsa, mientras que se volverán aún más raros los montoncitos "70 y 30" (Noelle, 1970).

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

3.2. Error muestral y probabilidad (precisión y confianza)

Cuanto se ha dicho en el punto anterior nos conduce a dos nociones que son muy importantes en el muestreo: la precisión o margen de error con que podemos hacer una estimación y la confianza o probabilidad con que podemos afirmar que esa estimación es correcta, pues nunca estaremos seguros del todo (Chein, 1980).

La primera de estas dos ideas remite al tamaño del error de muestreo. En el ejemplo del punto anterior –la bolsa con bolillas negras y blancas– sacaríamos un puñado de bolillas de la bolsa con el propósito de averiguar, por ejemplo, la proporción de bolillas blancas (o negras) que ella contiene. Ya vimos que tendríamos un buen motivo para hacerlo así: si volcáramos toda la bolsa en el suelo con el objeto de llevar a cabo un conteo total, además de tomarnos muchísimo trabajo y tardar largo rato, seguramente nos equivocaríamos. Pero también vimos que este procedimiento tendría otra consecuencia: si tomamos diez bolillas nada nos garantizará que la mezcla de colores se asemeje mucho a la que hay dentro del saco: podría diferir mucho. En cambio, cuando aumentábamos la cantidad de bolillas extraídas, podíamos confiar en que la mezcla de colores se iba aproximando crecientemente a la que existe en la bolsa. Ya vimos que si en la bolsa hubiera mitad de cada color, una extracción de diez bolillas podría arrojar cuatro blancas y seis negras o inclusive siete y tres. Con lo que, a partir de tan solo diez casos, haríamos una estimación de las proporciones existentes sujeta a un error muy grande. En cambio, hemos sugerido que si sacamos un centenar o un millar, aunque tampoco obtengamos exactamente la misma mezcla de colores que hay en el total, tendremos una mucho más parecida: jamás obtendríamos 600 y 400 si las proporciones reales son mitad de cada color. Es decir, con una muestra más grande se consigue aumentar la *precisión* o, en otras palabras, achicar el error de estimación.

Pero también afirmamos que por grande que fuera la muestra, nunca podríamos garantizar una estimación exacta: nada nos asegurará que, sacando un millón, obtengamos 500 mil de cada color. Hasta lo consideraríamos una casualidad si así ocurriera. Es decir, siempre hay un error de estimación, pero disminuye a medida que crece el tamaño de las muestras. Hay procedimientos matemáticos –que no abordaremos aquí– aptos para determinar cuántos casos necesitamos si es que no queremos superar un cierto margen de error. Por ejemplo, si queremos estimar la proporción de bolillas negras sin equivocarnos por más de cierta cantidad de puntos porcentuales. Por ejemplo, que si hubiera 50% de bolillas negras en la bolsa, la muestra no me pueda engañar en mucho: digamos que contenga entre 47% y

manual metodologia 22/5/08 15:36 Página 232

53%. O, dicho de otra manera, que si en la muestra tengo una mezcla de 45% blancas y 55% negras, pueda confiar en que en realidad debe haber entre 42% y 48% de bolillas blancas y –por lo tanto– entre 52% y 58% de bolillas negras

¿Qué pueda confiar? Pues sí, ¿pero cuánta confianza podría depositar en que es así? Nunca tendré entera certeza. En el mismo ejemplo brindado en el punto precedente, habíamos afirmado que si obtuviéramos muchos puñados de bolillas en lugar de uno solo, ellos diferirían entre sí: presentarían diferentes mezclas de colores. Pero a medida que fueran más grandes, muy pocos de ellos tendrían mezclas muy apartadas de la que realmente existe en la bolsa y, por lo tanto, también tenderían a parecerse entre sí.

Para ponerlo de otra forma, si los puñaditos fueran pequeños mostrarían una gran diversidad entre ellos y, además, muchos tendrían "mezclas raras" y muy diferentes de la realmente existente en la bolsa. Entre las pequeñas muestras de tan solo diez elementos, tal vez menos de la mitad de ellas contuvieran bolillas negras y blancas por mitades, en tanto que las otras podrían tener cualquier mezcla. De manera que si seleccionara uno solo de estos montoncitos, tendría una probabilidad mayor a 50% de incurrir en un error muy grueso, al obtener una muestra muy poco parecida a la población. En cambio con puñados grandes –por ejemplo, de un centenar de bolillas– la mayoría contendrían una proporción cercana a la de la bolsa: por ejemplo entre 47% y 52% de bolillas blancas. Habría unos pocos algo más apartados: por ejemplo con 55% o 45%. Y serían muy, pero muy escasos, los que guardarían una distancia mayor: tal vez tan solo uno o dos de cada cien montoncitos. Si esto último fuera así, entonces, al obtener una muestra de tamaño igual a cien, podríamos tener una elevada *confianza* en no equivocarnos en más de cinco puntos: digamos de 98%. Quiere decir que, con una muestra más grande también aumentamos la confianza en nuestras estimaciones, así como mejorábamos su precisión.

3.3 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. 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. Ello exige la aplicación de una serie de fórmulas y cálculos que no consideraremos aquí[4], ya que estamos abor-

manual metodologia 22/5/08 15:36 Página 233

dando la cuestión del muestreo de un modo conceptual. Sin embargo, procuraremos dar una idea acerca de los principales factores que se tienen en cuenta al determinar el tamaño de las muestras.

Uno de ellos es la precisión con que queremos estimar: al decir cuál será el porcentaje de votantes que sufragarán por cierto partido en las próximas elecciones, ¿aceptaremos equivocarnos por un punto, por dos o por tres?

Otro es la confianza que pretendemos asignar a nuestras estimaciones: esta confianza también se puede cuantificar, puesto que es posible decidir, antes de elegir la muestra, si queremos afirmar algo con una seguridad de 99% o nos conformaremos con una de 95%.

Un tercer elemento estriba en una característica de la población: su heterogeneidad: 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... Efectivamente, Si en la bolsa de nuestro ejemplo hubiera sólo bolillas blancas –o negras– bastaría con sacar una muestra de tan solo un elemento para describir el conjunto. Si las hubiera de dos colores, acaso una treintena nos brindaran una razonable aproximación. En cambio, si la policromía llegara a una decena de colores, sin duda necesitaríamos una muestra mucho más grande. La heterogeneidad de las poblaciones también es cuantificable mediante una medida estadística: se trata del desvío estándar, que no trataremos aquí[5].

En términos generales, son estos tres elementos mencionados –el error que estamos dispuestos a aceptar, la seguridad que pretendemos otorgar a nuestras estimaciones y una estimación de la heterogeneidad de la población– los que se combinan en una fórmula matemática que permite determinar el tamaño de la muestra. En cambio, como ya fue señalado más arriba, muy poco tiene que ver el tamaño de la población. Baste señalar que si la bolsa tuviera millones de fichas, pero todas de un único color, seguiría bastándonos con una.

Pero lo habitual –sobre todo en las disciplinas sociales– es que los fenómenos no estén distribuidos de un modo tan uniforme: ya se sabe, por ejemplo, que las condiciones de vida o las opiniones de las personas suelen variar considerablemente. Por lo tanto, cuando se pretenden estimaciones dotadas de razonable precisión y confianza, lo habitual es que se requieran muestras de –al menos– varios centenares de casos.

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 tuviéra-

manual metodologia 22/5/08 15:36 Página 234

mos una variable de tramos de edades 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.

3.4. Tipos de muestras probabilísticas

Bajo estos fundamentos, existen varias maneras de obtener muestras probabilísticas. La opción por unas u otras depende tanto de los propósitos de la investigación que se desea llevar a cabo, como de las características de la población a indagar y de los recursos y posibilidades con que cuenta el investigador. Los cuatro tipos básicos de muestras de probabilidad son:

- las muestras al azar simple
- las muestras al azar sistemático
- las muestras estratificadas
- las muestras por conglomerados o *clusters*

3.4.1. Muestras al azar simple

En las muestras de azar simple, hemos de contar con un listado completo de los N elementos que componen la población (Kish, 1993). Estos elementos deben ser numerados y, a continuación, se eligen mediante una tabla de números al azar, un bolillero o cualquier otro medio aleatorio, n elementos que integrarán la muestra. El procedimiento es, pues, similar al que aplica quien saca papelitos numerados de una galera. A condición de que sea con reposición –es decir que cada uno de los elementos que salen sorteados se vuelvan a "meter dentro de la galera" – realmente cada elemento de la población tiene una probabilidad igual a los demás de integrar la muestra. 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 (Blalock, 1986).

En otros términos, si de un total de 3 mil escuelas de una jurisdicción cualquiera se quisiera seleccionar al azar simple 300, con el propósito de hacer una inspección sobre su estado edilicio, la probabilidad de selección de la primera escuela sería 1/3000. Si una vez que fue elegida la tacháramos del listado, para no volver a seleccionarla, al hacer la segunda selección la probabilidad sería algo mayor: 1/2999.

Las muestras al azar simple constituyen una suerte de ideal que muchas veces no puede

manual metodologia 22/5/08 15:36 Página 235

alcanzarse (Kish, 1993). En primer lugar, porque frecuentemente no tenemos un listado lo suficientemente completo de los elementos componentes de la población, que nos permita numerarlos y seleccionarlos con tanta sencillez. Pero también, a veces resultarían costosas y poco prácticas. Piénsese en el mismo ejemplo de las escuelas: ellas podrían estar dispersas a lo largo y ancho de un territorio muy vasto y los inspectores gastarían una enormidad de tiempo y dinero trasladándose de una a otra: luego veremos que hay formas más prácticas de resolver problemas como este. A la inversa, podría muy bien suceder que no saliera sorteada ninguna escuela situada en un lejano confín del territorio, del cual lo ignoramos todo y, por eso mismo, sería importante inspeccionar al menos alguna de sus escuelas ¿Cómo garantizar que el azar la haga aparecer? También hay soluciones para ello: ya lo veremos.

La figura 1 esquematiza un muestreo al azar simple de n = 4 elementos sobre una población total de N = 20. La fracción de muestreo sería, pues, 0,2.

Figura 1: muestra al azar simple

n = 4

1
8
11
19

N= 20

1	6	**11**	16
2	7	12	17
3	**8**	13	18
4	9	14	**19**
5	10	15	20

$f = n/N = 4/20 = 0,2$

3.4.2. Muestras al azar sistemático

El muestreo al azar sistemático requiere, al igual que el de azar simple, de un listado o marco muestral. El paso inicial consiste en obtener un cociente entre N (cantidad de elementos que componen la población) y n (cantidad de ellos que hemos de incluir en la muestra). En el ejemplo de las escuelas este cociente sería 3000/300 =10. Luego, se determina un primer número al azar entre 1 y N (en este caso, entre 1 y 3000). Supongamos que salga el 115: pues bien, la primera escuela elegida será la n° 115 en el listado. A partir de ella elegiremos las otras 299 con intervalo igual a N/n, es decir igual a 10. De manera que hecha la primera selección, le seguirá la 125, luego la 135, etc. Puesto que llegaríamos al final del listado antes de haber agotado la selección, continuaríamos contando desde el principio.

El muestreo al azar sistemático asegura dispersión a lo largo de toda la población, pero

manual metodologia 22/5/08 15:36 Página 236

podría acarrear algún inconveniente en caso de que el listado guardara un orden predeterminado y no azaroso (Blalock, 1986). Si en el ejemplo anterior el marco muestral estuviera ordenado en forma decreciente por el tamaño de estos establecimientos y la primera elección al azar recayera sobre una escuela situada al comienzo del listado, tenderíamos a elegir escuelas algo más grandes que si comenzáramos a elegir desde un punto más avanzado.

La figura 2 esquematiza un muestreo al azar sistemático, donde –como en el caso anterior– se han seleccionado cuatro elementos sobre veinte. La primera elección recayó sobre el elemento n° 2 y de allí en mas se realizaron las siguientes selecciones con un intervalo de 20/4 = 5.

Figura 2: muestra al azar sistemático

n = 4

2
7
12
17

N= 20

1	6	11	16
2	**7**	**12**	**17**
3	8	13	18
4	9	14	19
5	10	15	20

f = 4/20 = 0,2

N/n = 20/4 = 5

3.4.3 Las muestras estratificadas

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 separa el universo en *estratos* o *subuniversos*. Para ello debemos disponer un marco muestral, vale decir, un listado de los elementos que componen la población, que contenga la información suficiente como para hacer tal división.

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 separadas para cada carrera (o de una sola lista donde para cada alumno figurara la carrera que cursa) 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 Bue-

manual metodologia 22/5/08 15:36 Página 237

nos 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. Obviamente, pues, se estratifica con variables cuya distribución en el universo es conocida.

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 (Chein, 1980). 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 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 (N_1 = 90.000) en tanto que el restante sólo alberga al 10% (N_2 = 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?[6] 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:

manual metodologia 22/5/08 15:36 Página 238

$f_1 = n_1/N_1 = 900/90.000 = 0,01$
$f_2 = n_2/N_2 = 100/10.000 = 0,01$

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, +-3%) y c) el grado de heterogeneidad en la distribución poblacional de la variable que queremos estimar. Los primeros dos factores dependen de decisiones que se toman a priori, pero el tercero nos resulta impuesto: la distribución de los ingresos familiares o la de los jefes de hogar por sexo es como es. Si la población es muy heterogénea, ya lo hemos visto, necesitaremos una muestra más grande.

Sin embargo, podemos hacer algo al respecto: supongamos que necesitamos estimar el ingreso promedio de las familias que habitan en una ciudad. Supongamos también que esta ciudad presenta tres áreas o estratos geográficos bien diferenciados y 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, asumiremos 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 (Chein, 1980).

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 sen-

manual metodologia 22/5/08 15:36 Página 239

tido: 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 (Mayntz et al, 1975). Pero, ¿cuánto más o cuánto menos? Existe una fórmula estadística que permite resolver la cuestión[7].

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 desvío estándar de la población (Chein, 1980; Mayntz et al, 1975). En este caso, la estratificación habrá de redundar en una estimación sujeta a un menor error.

La ponderación posterior en el muestreo no proporcional

A esta altura, el lector 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 en desmedro de otras, 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 promedios 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.

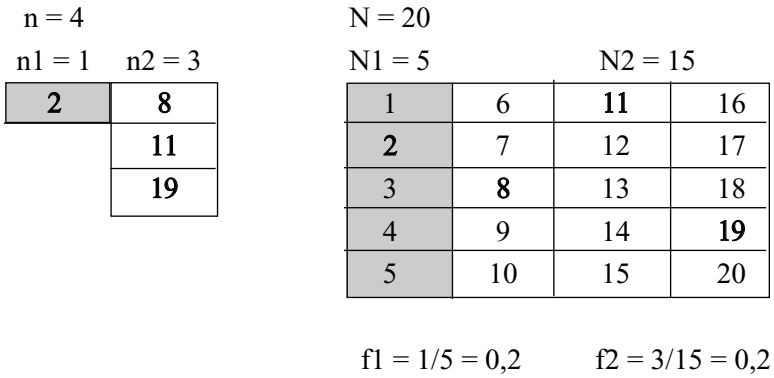
Pero si las fracciones de muestreo empleadas en un muestreo estratificado difieren, entonces los casos de cada estrato deberán ser expandidos aplicando la inversa de la fracción muestral empleada en cada uno de dichos estratos. Ello permite restituir a cada uno el peso que realmente le corresponde.

La figura 3 reproduce el esquema de una muestra estratificada proporcional. Las fracciones de muestreo son iguales en ambos estratos. La muestra consta de cuatro elementos:

manual metodologia 22/5/08 15:36 Página 240

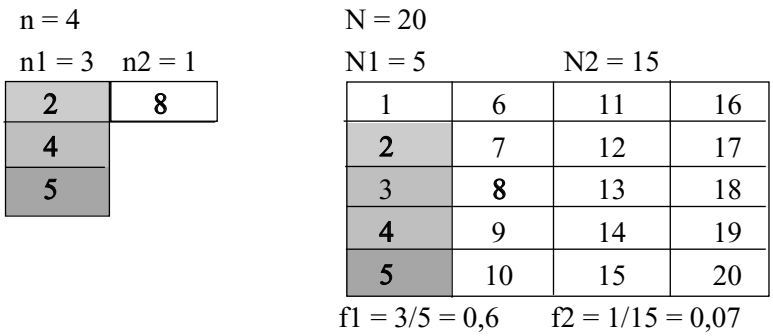
en el estrato 1 (grisado) se seleccionó un solo elemento sobre cinco, en tanto que en el estrato dos se seleccionaron tres sobre quince.

Figura 3: muestra estratificada proporcional



La figura 4, en cambio, esquematiza un muestreo estratificado no proporcional. En el estrato 1, que presenta gran heterogeneidad, se seleccionaron tres elementos sobre cinco, de manera que la fracción muestral es muy grande: 0,60. En tanto que en el otro estrato, enteramente homogéneo (puesto que todas las casillas son de igual color), bastó con un solo elemento sobre quince. La fracción de muestro es allí 0,07.

Figura 4: muestra estratificada no proporcional



3.4.4. 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

manual metodologia 22/5/08 15:36 Página 241

ocuparemos de otro procedimiento de muestreo, donde el propósito es, en alguna medida, inverso (Blalock, 1986): 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 registros 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" (Blalock, 1986) .

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

3.4.5 Conglomerados polietápicos o de fases múltiples

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 en las disciplinas sociales. Se trata del muestreo

manual metodologia 22/5/08 15:36 Página 242

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[8].

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,

manual metodologia 22/5/08 15:36 Página 243

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[9] sugieren una solución sencilla para un problema complejo: determinado el tamaño muestral como si se tratara de una muestra al azar simple, debiera obtenerse una 50% más grande si se tratara de una muestra de etapas múltiples.

4. Muestreo no probabilístico

Tal como se lo indicó en los parágrafos anteriores, el muestreo probabilístico provee una serie de ventajas, puesto que el azar resuelve muchos problemas. Si queremos tener razonables garantías de que la muestra se asemeja al universo, lo mejor es dejarlo todo en sus manos.

Pero a veces sencillamente los propósitos del muestreo son diferentes. O bien no contamos con los recursos necesarios como para aspirar a una muestra de azar: ya vimos que para que sean eficaces tienen que ser grandes. Pero una muestra grande exige –en las ciencias sociales– un amplio operativo que demanda considerable logística. Y podría ocurrir que el presupuesto no alcanzara sino para un centenar de casos. Es aquí donde entran las muestras que no se obtienen mediante procedimientos de azar, sino seleccionando cierto tipo de casos elegidos voluntariamente.

4.1. Muestreo coincidental

El llamado muestreo coincidental –o también *accidental*– es, seguramente, el más objetable desde el punto de vista de las ciencias sociales, aunque muchas veces no queda otro remedio que emplearlo. Se trata de seleccionar –por ejemplo para realizar una encuesta– a las personas que pasan casualmente por el sitio donde se ha ubicado el encuestador: generalmente un punto por donde pasa mucha gente, tal como una estación ferroviaria o una esquina céntrica.

Podría ser que el encuestador se autoimponga una pauta, tal como seleccionar a uno de cada diez. Aunque lo más usual es que aborde a los que parecen menos apurados o le inspiran más confianza. Pero lo que importa es que –sea como fuere– quienes tienen alguna oportunidad de ser seleccionados son, tan solo, aquellos que pasan cerca. La población a la que se supone representan no está presente ni en forma física ni virtual. Son seleccionados de manera fortuita y casual: pero casualidad no es equivalente a azar. Por lo tanto, no cabe aquí hablar de errores de muestreo ni de probabilidades de acertar o de equivocarse. Esta clase de muestreo no forma parte, pues, de los procederes científicos.

manual metodologia 22/5/08 15:36 Página 244

Y sin embargo, a veces no queda otro remedio... Cuando se realizan sondeos de opinión mediante encuestas domiciliarias, se emplean muestras de conglomerados de múltiples etapas, tal como se describió en el punto *3.5.1.* Ese procedimiento selecciona al azar los puntos muestrales: pero sucede que hay lugares que son poco accesibles, ya sea por razones de seguridad –como las villas de emergencia– o bien porque existen restricciones como sucede en los barrios privados. En estos casos es frecuente que los encuestadores se sitúen en la periferia y aborden a las personas que entran o salen. Lo óptimo –hay que decirlo– es enemigo de lo bueno: es preferible hacerlo así, vulnerando el azar, a dejar de lado segmentos de población que, por sus peculiares características, pueden presumirse muy diferentes al resto.

4.2. Muestreo por cuotas

El muestreo por cuotas es tenido por la mejor alternativa para sustituir al muestreo probabilístico en procura de obtener una razonable réplica "en pequeño" de la población, cuando no se cuenta con la posibilidad de llevar a cabo un muestreo al azar (Mayntz et al, 1975). Guarda alguna semejanza con el muestreo estratificado –que ya se ha explicado– pero no debe ser confundido con él (Blalock, 1986).

Como en el caso del muestreo estratificado, es necesario conocer la distribución en el universo de ciertas variables importantes que nos permitan construir las *cuotas.* Por ejemplo, si tuviéramos que hacer una encuesta de opinión en la Ciudad de Buenos Aires dirigida a personas de 15 y más años, podríamos tener en cuenta la distribución de la población por sexo y edad. Esta distribución podríamos conocerla fácilmente apelando a los datos del último censo:

Cuadro 4.2.1: Población de la Ciudad de Buenos Aires de 15 años y más por grupos de edad y sexo

Edades	Varones	Mujeres	Total	Varones	Mujeres	Total
15-24	203.592	211.029	414.621	9%	9%	18%
25-39	294.381	321.007	615.388	13%	14%	27%
40-59	297.297	370.405	667.702	13%	16%	29%
60 y más	225.281	384.185	609.466	10%	17%	26%
Total	1.020.551	1.286.626	2.307.177	44%	56%	100%

Fuente: INDEC-Censo Nacional de Población y Vivienda 2001

Trasladados a porcentajes, estos datos nos dicen que en la ciudad había un 56% de mujeres y un 44% de varones. Pero además, un 18% de esta población tenía entre 15 y 24 años

manual metodologia 22/5/08 15:36 Página 245

y se repartía por mitades entre ambos sexos. Supongamos que hubiéramos previsto una muestra de 100 casos: pues entonces determinaríamos entrevistar a 44 hombres y 56 mujeres. Tanto en uno como en otro caso, habría entre ellos 9 jóvenes de 15 a 24 años. Al pasar al siguiente tramo de edad, en cambio, tomaríamos 13 varones y 14 mujeres. De igual modo procederíamos con cada grupo, hasta llegar a las personas de más de 60 años: allí seleccionaríamos 10 varones y 17 mujeres.

Una vez determinada esta composición de la muestra, repartiríamos los casos definidos en función de sexo y edad entre los encuestadores: ellos debieran buscar libremente a los entrevistados, con la sola condición de que respondieran a estas especificaciones.

Por este procedimiento, no estaríamos eligiendo al azar ni otorgando a todos los habitantes de la ciudad alguna probabilidad de integrar la muestra, porque los encuestadores elegirían a su entero arbitrio, probablemente entre sus conocidos. Sería, en tal sentido, una situación semejante a la del muestreo coincidental. Pero nos aseguraríamos de que la muestra sería una réplica en pequeño de la población, al menos en las variables tenidas en cuenta al definir las cuotas.

Nada impide, por cierto, que en un muestreo coincidental se tengan en cuenta cuotas: cierta cantidad de mujeres y varones, de determinadas edades.

4.3. Muestreo intencional

Por fin, el caso del muestreo intencional se aparta bastante de la lógica tenida en cuenta hasta aquí. En este tipo de muestreo no buscamos en modo alguno que la muestra se parezca a la población. Por el contrario, nos interesamos exclusivamente en cierto tipo de casos, que nos resulten relevantes desde el punto de vista teórico (Padua, 1979).

En una investigación sobre la percepción acerca de los planes sociales que proveen ingresos a personas sin ocupación y con menores de edad a cargo[10], podría interesarnos especialmente la opinión de ciertos grupos extremos: personas de muy bajos recursos –que reciban y que no reciban estos subsidios pero que puedan considerarse como beneficiarias potenciales de ellos– y personas de nivel socioeconómico muy elevado. O bien, por el contrario, personas de ingresos bajos, pero no tan bajos como para recibir un subsidio público, bajo la presunción de que estas últimas serán las que se muestren menos favorables a tal tipo de programas, al considerarse víctimas de una injusticia.

Si así fuera, elegiríamos precisamente a personas pertenecientes a estos grupos sociales: sería poco útil una muestra probabilística que, en cambio, estaría integrada por muchos sujetos poco útiles a los efectos de nuestra investigación (Blalock, 1986).

manual metodologia 22/5/08 15:36 Página 246

5. Un plan de muestreo: algunas pautas para decidir

Tal como se ha explicado en las páginas precedentes, el plan de muestreo que se elija es una decisión estrechamente ligada con los propósitos de la investigación, pero también con las características de la población a indagar y con los recursos con que cuenta el investigador. Estos últimos se refieren tanto a la disponibilidad de marcos muestrales e información sobre la población como a recursos económicos, humanos y de tiempo.

A los efectos de tomar una decisión, conviene formularse las siguientes preguntas:

- *¿Queremos que la muestra se parezca a la población?* Ya hemos dicho que no siempre –aunque sí frecuentemente– es este el propósito del muestreo. Si lo fuera hemos de pensar en una muestra probabilística. Pero si no lo fuera la desecharíamos de plano y optaríamos por una muestra intencional, más económica y adecuada a nuestros propósitos

- *¿Tenemos tiempo y recursos para una muestra de azar?* Una muestra probabilística o de azar requiere, por lo general, varios centenares de casos: si luego hay que llevar a cabo una encuesta hará falta una gran cantidad de encuestadores o bien mucho tiempo, o una combinación de ambas cosas. Además, si su diseño es complicado debe ser llevada a cabo por un especialista en muestreo ¿Cuenta el equipo de investigación con recursos humanos y materiales suficientes? Si no fuera así, mejor sería optar por una razonable muestra no probabilística por cuotas.

- *¿Tenemos un listado razonablemente actualizado y completo?* Ya se dijo que las muestras probabilísticas o al azar requieren la presencia física o virtual de todos los elementos que componen la población. En las disciplinas sociales, es frecuente que estos elementos sean o bien personas o grupos de ellas –como las familias– o instituciones de distinto tipo, como empresas o escuelas. La población de una ciudad, por ejemplo, está materialmente presente en el territorio de ella pero no hay un listado actualizado, conteniendo las direcciones, que permita numerarlos y sortearlos. Esto sí que sucede, en cambio, con las escuelas –al menos con las de gestión pública– que figuran en los listados del Ministerio de Educación, en formato digital. O con los Hospitales. Si tenemos un listado, entonces se puede hacer una muestra al azar simple o sistemático. Y también se puede hacer una estratificación del universo o población, si el listado cuenta con información relevante para ello: por ejemplo, con los registros de las escuelas, antes de elegir al azar podríamos estratificar por nivel (primario, secundario, terciario), por regiones del país o por barrios de una ciudad. Si tal listado no está disponible, en cambio, la selección al azar simple o sistemático no será posible. Es el caso de las personas o las

manual metodologia 22/5/08 15:36 Página 247

familias: sería preciso apelar a una muestra de conglomerados de varias etapas, que nos permita seleccionar fracciones del vasto territorio donde están dispersas, para llegar a ellas.

- *¿Están los elementos del universo agrupados en unas unidades mayores, que faciliten su captura?* Si sucede así, hemos de aprovecharnos de esta ventaja. Esto ocurriría, por ejemplo, si quisiéramos entrevistar empleados bancarios, maestras de escuela o policías: no necesitaríamos contar con listados de personas, porque ellas podrían ser halladas en sus lugares de trabajo. Bastaría con los listados –mucho más pequeños y manejables– de sucursales bancarias, escuelas o dependencias policiales. Seleccionando muestras al azar simple o sistemático de estas instituciones, podríamos luego entrevistar a una parte o a todos los sujetos que trabajan en ellas: sería una muestra de conglomerados.

- *¿Se trata de una población concentrada o dispersa?* Puede ocurrir que contemos con un listado o que no contemos con él. Que la población esté concentrada en unidades mayores o que no lo esté. Pero hay otra condición que influye sobre las decisiones muestrales: la dispersión geográfica. Supongamos que queremos entrevistar directores de 400 escuelas de todo el país: les hallaríamos dentro de los establecimientos escolares. Y puesto que existe un registro de estos establecimientos, nada impediría que se eligieran al azar simple. Pero podría resultar de ello una dispersión enorme, que obligaría a los encuestadores a viajar sin descanso ni sosiego de la Puna a los confines más australes de la Patagonia y desde Cuyo a la Mesopotamia cruzando la región Pampeana y atravesando también cada una de estas regiones porque ninguna escuela queda próxima a otra. Las demoras y los gastos serían, a no dudar, excesivos. En estos casos, aun habiendo un listado, sería más sensato obtener una muestra de conglomerados seleccionando áreas geográficas distribuidas en todo el territorio nacional y no demasiado amplias. De este modo, las escuelas de cada una de estas áreas quedarían razonablemente próximas entre sí. Cada encuestador se ocuparía de unas pocas áreas de una misma región, con lo que la tarea se facilitaría considerablemente.

manual metodologia 22/5/08 15:36 Página 248

Bibliografía

- BLALOCK, H. (1986), *Estadística Social.* México: Fondo de Cultura Económica.
- CHEIN, I. (1980). "Introducción al muestreo": Apéndice "A" en Selltiz et. al. *Métodos de investigación en las relaciones sociales*. Madrid: Rialp.
- GARCIA FERRANDO, M. (1985), *Socioestadística.* Madrid: Alianza Universidad.
- KISH, L. (1993). "Procedimientos de muestreo". En Festinger, L. y Katz, D. *Los métodos de investigación en las ciencias sociales*. México: Paidós.
- MAYNTZ, R., HOLM, K. y HÜBNER, P.(1975). *Introducción a los métodos de la sociología empírica*. Madrid: Alianza Universidad.
- NOELLE, E. (1970). *Encuestas en la sociedad de masas*. Madrid: Alianza Editorial.
- PADUA, J. (1987). *Técnicas de investigación aplicadas a las ciencias sociales*. México: El Colegio de México/Fondo de Cultura Económica.
- SIERRA BRAVO, R. (1995). *Técnicas de investigación social. Teoría y ejercicios*. Madrid: Paraninfo.
- SLONIM, M. (1974). *Muestreo*. Buenos Aires: Editorial Americana.

Notas

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.

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.

4 Los más básicos pueden encontrarse en cualquier manual de estadística. Por ejemplo: BLALOCK, Hubert (1986), *Estadística Social*, Fondo de Cultura Económica, México. O bien: GARCIA FERRANDO, Manuel (1985), *Socioestadística*, Alianza Universidad, Madrid. Existen, también, textos abocados exclusivamente a problemas de muestreo aleatorio.

5 Aunque sí en un capítulo posterior, donde se brindan algunos fundamentos básicos de estadística descriptiva.

6 La distribución de los casos muestrales entre los estratos se denomina *afijación.*

7 Se trata de la llamada fórmula de *Neymann*, de afijación óptima, que no veremos aquí. Esta fórmula permite que la fracción de muestreo sea directamente proporcional a la heterogeneidad de los estratos, medida por sus desvíos estándar: una medida estadística de la que nos ocupamos en un capítulo posterior.

8 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 el encuestador no fuera expulsado por personal de vigilancia...

9 Seguramente basados en la experiencia.

10

manual metodologia 22/5/08 15:36 Página 249

11. LA OBTENCIÓN DE LA EVIDENCIA EMPÍRICA.
LAS ENCUESTAS: DISEÑO DE CUESTIONARIOS. ANEXO. LA CODIFICACIÓN DE LOS DATOS

Mariana Colotta

1. Encuestas y cuestionarios

¿Quién alguna vez no participó de una encuesta?, ¿a quién alguna vez no lo interceptaron en alguna esquina muy concurrida para preguntarle a quién iba a votar?, ¿quién alguna vez no recibió una llamada telefónica en la que se lo interrogaba sobre el programa televisivo que estaba viendo en ese momento? ó ¿quién alguna vez no se despertó de una siesta con un timbrazo de una encuestadora presentada en la puerta de su casa, preguntándole sobre qué productos de limpieza usa en su hogar? Sin lugar a dudas en algún momento de nuestra vida hemos sido interpelados por esta técnica de recolección de datos que ha brillado y sigue haciéndolo desde los años setenta.

Ahora bien, se mencionan indistintamente el concepto de encuesta y cuestionario, pero ¿son lo mismo? Remitiéndonos a Johan Galtung (1978), presentamos en términos generales el *método de la encuesta*, como el método para llenar matrices de datos. Sin embargo desde un punto de vista más específico lo acotamos como aquel método, que valiéndose de dos técnicas de recolección de datos, una oral, la entrevista y otra escrita, el cuestionario, se centra en los individuos como unidades de análisis de los cuales considera sus actitudes, sus comportamientos y algunas variables contextuales o sociodemográficas. En resumidas cuentas, la encuesta es el método y el cuestionario es la técnica.

¿Por qué nos pareció importante dedicarle un capítulo de este libro al tema de los famosos cuestionarios? Porque si bien es conocida la difusión y aplicación del cuestionario como técnica de recolección de datos, su planificación, elaboración y redacción es una tarea compleja, por lo que consideramos valioso reflexionar a lo largo de este capítulo sobre las cuestiones varias que lo atraviesan hasta llegar a su elaboración final.

Por un lado, subyacen a su diseño temas tales como planteamiento de hipótesis, formulación de objetivos de investigación, especificación de las preguntas y variables; a la par de otras cuestiones más operativas, como el orden y disposición de las preguntas, longitud, aspectos formales, filtros, instrucciones y precodificación.

Por otra parte, si bien no se entrará en detalle en este capítulo, la selección de la muestra en la que se aplicará este instrumento debe ser realizada teniendo en cuenta dos pregun-

manual metodologia 22/5/08 15:36 Página 250

tas: ¿es relevante? y ¿es practicable? La primera cuestión tendrá que ver con la importancia de su aplicación en consonancia con el objetivo de la investigación, mientras que la practibilidad nos remitirá al dinero, tiempo, recursos humanos y materiales que deben ser invertidos en la investigación y pesan a la hora de tomar decisiones.

2. Las preguntas

Las preguntas son el elemento básico del cuestionario. Estas hacen referencia a la estructura formal del mismo; son la expresión manifiesta, normalmente en forma de interrogación, mediante la cual se recaba información de diverso tipo: edad, sexo, nivel de estudios alcanzado, características ocupacionales, preferencias ideológicas de la persona entrevistada, eventualmente, características de su hogar, etc., como también en qué provincia o distrito ha sido realizado el trabajo de campo.

Generalmente las preguntas coinciden con las variables, pero puede ocurrir que:

a) haya variables que no requieran pregunta (*ejemplo: sexo que se constata por observación)*

b) se pueden generar distintas variables a partir de las preguntas en que se piden dos o más respuestas y cada respuesta se convierte en una variable diferente

c) o bien, se puede construir una única variable a partir de las respuestas a diferentes preguntas. (*ejemplo: las preguntas de la EPH acerca del trabajo en la semana de referencia, la remuneración recibida y la cantidad de horas, junto con las búsquedas, en caso de no haber trabajado, permiten la construcción de la variable "condición de actividad": ocupado, desocupado, inactivo. Aunque también se incorporan a la matriz individualmente).*

Atendiendo a la función, finalidad y tipo de análisis de las preguntas éstas pueden ser de distinto tipo.

- Preguntas *abiertas*
- Preguntas *cerradas*
- Preguntas *semicerradas*

En las *preguntas abiertas* no se establece ningún tipo de respuesta. El entrevistado se expresa con sus propias palabras. Generalmente se utiliza en preguntas exploratorias o cuando no se puede presumir la reacción u opinión del entrevistado.

No es bueno abusar de las preguntas abiertas, debido al trabajo de codificación posterior que requieren. Otra recomendación es la de dejar espacio en el cuestionario para la buena transcripción de la respuesta de manera textual, clara y legible. Ejemplo de pregunta abierta: *¿Qué noticias de las ocurridas en el mundo le ha llamado más la atención en el último mes?*

manual metodologia 22/5/08 15:36 Página 251

Entre sus desventajas Jorge Padua (1987) menciona la dificultad por parte del entrevistado de contestar cuando no tiene la respuesta prevista, o los pormenores para clasificar las respuestas en el libro de códigos y la necesidad de entrevistadores bien entrenados. Sin embargo, entre sus ventajas, sobresale la buena comprensión de los motivos del entrevistado[1], así como un mayor grado de compromiso por parte de éste. Asimismo, propician un mejor contacto entre entrevistado-entrevistador e incrementan la motivación para la entrevista.

En las *preguntas cerradas,* la persona entrevistada no puede salirse de las categorías impuestas en el cuestionario. Son más rápidas y cómodas en la etapa de recolección y evitan la codificación a posteriori. Uno de los problemas que presentan es que las alternativas de respuesta deben ser exhaustivas y el investigador puede pasar por alto alguna respuesta importante, por lo que se recurre con frecuencia a añadir una última categoría residual: "otros". Ejemplo de pregunta cerrada : *¿A qué partido político está afiliado? PJ, UCR, Recrear, ARI, Otros*[2].

En esta dicotomía de preguntas abiertas vs cerradas, es bueno considerar las *preguntas semicerradas***,** pues además de la lista cerrada de respuestas , el entrevistado puede añadir espontáneamente algo que no está previsto en el cuestionario.

Ejemplo: *¿ Qué tipo de noticias le interesan más a ud.?*

De política internacional
De política nacional
Económicas y laborales
Culturales
De sucesos
Deportivas
Otras ¿ Cuáles?.....

Entre las ventajas mencionadas por Padua (1987), respecto de las preguntas cerradas, se mencionan la facilidad para registrarlas, interpretarlas, codificarlas y analizadas, sin necesidad de recurrir a entrevistadores entrenados. En cuanto a las desventajas, se menciona el hecho de que las preguntas cerradas impiden una clasificación muy fina, ya que las respuestas que se encuentran justamente en el borde de dos alternativas tienden a ser forzadas en algunas de las dos categorías. Con ello, el entrevistado podría verse impulsado a encasillar su respuesta en una alternativa preestablecida que, eventualmente, podría no representar cabalmente su opinión[3]. Una posibilidad para contrarrestar este tipo de sesgos en las respuestas es la de darle al entrevistado la posibilidad de seleccionar más de una alternativa; incluso puede limitarse la elección a un número fijo de alternativas (por ejemplo, las tres principales) para evitar que se marquen todas y evitar que la pregunta pierda su poder discriminatorio.

manual metodologia 22/5/08 15:36 Página 252

Dentro de las preguntas cerradas, el mencionado autor distingue la de múltiples respuestas, llamada vulgarmente *de cafetería,* porque el cliente puede elegir entre diferentes alternativas, tal como si lo hiciera entre platos o bebidas. La ventaja de este tipo de preguntas consiste en que admiten más posibilidades de respuesta, dando una mayor alternativa de elección y otorgando una mayor información.

Otra clasificación posible de las preguntas es aquella que las divide en:

- Preguntas dicotómicas
- Preguntas categorizadas o de escala ordinal
- Preguntas de escala numérica
- Preguntas de valorización

En las *preguntas dicotómicas* la persona entrevistada tiene que elegir entre dos alternativas posibles de respuesta: *Si-No* ; *De acuerdo-En desacuerdo*; *A favor-En contra*; *Aprueba-desaprueba*.

Las *preguntas categorizadas o de escala ordinal* no piden una respuesta tan terminante como las dicotómicas. Pueden resultar más fáciles de responder para el entrevistado puesto que tiene varias opciones entre las cuales elegir. Las opciones de respuesta deben ir siempre incluidas en el enunciado de la pregunta, pero si esto no es posible el entrevistador, debe leer las respuestas o mostrar una tarjeta que enumere todas las alternativas, tal como se verá más adelante.

Ejemplo: ¿ Cuál es su opinión respecto a legalización del aborto?

Muy de acuerdo
Bastante de acuerdo
De acuerdo
Ni de acuerdo ni en desacuerdo
En desacuerdo
Bastante en desacuerdo
Muy en desacuerdo

Las *preguntas cuantitativas* son aquellas cuya respuesta es un número, por ejemplo la edad expresada en años cumplidos, la antigüedad en meses o años en el trabajo o los ingresos percibidos en cierto período. Dentro de estas se ubica también la variante de las de *escalas numéricas* donde se le pide al entrevistado que se ubique en un continuo del 1 al 10, como los dos extremos de la escala.

Ejemplo*: En esta tarjeta hay un serie de casillas que van del 1 que significa izquierda al 10 que significa derecha ¿ En qué casilla se colocaría Ud. según sus ideas políticas?*

Izquierda	1	2	3	4	5	6	7	8	9	10	Derecha

manual metodologia 22/5/08 15:36 Página 253

Similares a las de escala son las *preguntas de valoración*, pero en este caso los extremos son una valoración mínima (00) y una valoración máxima(10). Suelen usarse para la valoración de imagen de líderes políticos o instituciones.

Otro criterio clasificatorio distingue a las preguntas directas frente a las indirectas. En el primer caso se aborda directa y sencillamente lo que se quiere saber:

¿ Ha trabajado ud. la semana pasada?

En el caso de las *indirectas* se esconde la información que se busca junto a otra información que no es de real de interés para el estudio. Esta estrategia presupone una mayor fiabilidad a las respuestas de ciertas cuestiones:

Ejemplo*: Le voy a leer una serie de actividades y quisiera que me dijera si las realizó o no Ud. la semana pasada*

Ir al cine
Pasear
Hacer deportes
Realizar tareas domésticas
Trabajar
Estudiar
Ver televisión

Asimismo, es posible diferenciar las preguntas de respuesta única frente a las que admiten respuesta múltiple. En las primeras la elección del entrevistado se remite a una única opción:

Ejemplo*: En su ocupación usted es...*

Patrón o empleador
Trabajador por cuenta propia
Obrero o empleado
Trabajador familiar sin remuneración

En el caso de las *respuestas múltiples* la posibilidad de elección se amplía. En algunos casos, se le pide al entrevistado que de un listado de respuestas se limite a dar un máximo de dos o tres respuestas; o por el contrario permitirle contestar libremente. También se le puede pedir que las jerarquice de acuerdo a su importancia:

Ejemplo*: En su tiempo libre usted suele...*

Ir al cine
Pasear
Hacer deportes
Reunirse con amigos
Leer
Ver televisión

manual metodologia 22/5/08 15:36 Página 254

Además de la clasificación antes presentada, podemos mencionar :

Las *multipreguntas o preguntas en batería*:

Estas preguntas permiten utilizar la misma formulación para preguntar sobre aspectos diferentes, recordando que cada uno de los ítems es una variable por sí misma, por lo que cada uno deberá llevar las mismas opciones de respuesta. El inconveniente que presenta esta modalidad de preguntas es que el abuso en el uso de las mismas pueden llegar no sólo a hacer interminable el cuestionario sino a estimular al entrevistado a contestar de la misma forma para cada ítem a modo de latiguillo.

Ejemplo: ¿ Podría decirme si a Ud. actualmente las cosas le van muy bien, bien, regular, mal o muy mal respecto a ...?

	Muy bien	Bien	Regular	Mal	Muy mal	Ns/Nc
Sus relaciones afectivas						
Sus relaciones familiares						
Su situación económica						

Preguntas filtro:

Tienen la función de evitar que contesten aquellas personas a quienes no va destinada la pregunta. Se realizan previamente a otra pregunta o grupo de preguntas para no abordar con las siguientes a aquellos que no le compete el tema. Una desventaja en el abuso de uso de este tipo de preguntas, es que reducen la muestra sensiblemente hasta el extremo de hacer imposible el análisis estadístico de la información que recogen. Ejemplo: *filtro de mujeres madres de niños de 0-5 años.*

Ejemplo 1: P 8 ¿Trabajó durante la semana pasada?

si (pasa a pregunta nro.9)

no (pasa a pregunta nro. XX)

P 9. ¿Cuántas horas trabajó en la semana?

Ejemplo 2: P 5 ¿Está ud. afiliado a un partido político?

si (pasa a pregunta nro.6)

no (pasa a pregunta nro. XX)

P 6. ¿A qué partido político está Ud. afiliado?

manual metodologia 22/5/08 15:36 Página 255

Preguntas de control o identificación:

Son las preguntas sociodemográficas y recaban información sobre la identificación de la muestra; las variables de identificación del entrevistador y control de trabajo de campo y las preguntas o variables que se refieren a las características básicas del entrevistado: sexo, edad, ocupación, estudios, estado civil, ingresos etc. Frente a las resistencias que presentan los entrevistados a la hora de responder sobre determinados detalles de su realidad sociodemográfica es bueno argumentarles que el cuestionario es anónimo y que este tipo de preguntas no apuntan a la identificación personal del entrevistado sino de aspectos o dimensiones que permiten agrupar a los entrevistados por categorías sociodemográficas.

Obviamente, ello sucede así en las encuestas de opinión; en las encuestas sociodemográficas, como las encuestas de hogares que realizan regularmente casi todos los países,[4] son estas preguntas las que tienen el principal protagonismo. En estos casos un único cuestionario puede, eventualmente, recabar información acerca de todos los miembros de un grupo familiar en lugar de (o además de) aplicarse un cuestionario individual a cada persona. Cuando se emplean estos cuestionarios dirigidos al conjunto del hogar, en ellos se suele incluir la información referida a las características de la vivienda y los datos sociodemográficos básicos de cada integrante, mientras que pueden utilizarse cuestionarios individuales para indagar con mayor detalle ciertos aspectos, tales como las características ocupacionales o migratorias.

Cuando se opta por esta clase de cuestionarios colectivos, resulta muy útil incluir un cuadro o grilla del hogar, semejante a esta:

N° de miembro	Relación *1 Jefe* *2 cónyuge* *3 hijo/a* *4 otro fam.* *5 otro no fam.*	Sexo *1 varón* *2 mujer*	Edad (años cumplidos)	Asistencia escolar *1 asiste* *2 no asiste pero asistió* *3 nunca asistió*	Cobertura salud *1 prepaga* *2 obra social* *3 ambas* *4 hospital público*
1	1	1	46	2	2
2	2	2	38	2	2
3	3	2	14	1	2
4	3	1	11	1	2

En el caso de que se aplicara, complementariamente, un cuestionario individual a cada miembro del hogar –o a cada miembro adulto– estos cuestionarios individuales se identificarían por el número de miembro, que aparece consignado en la primera columna de la grilla.

manual metodologia 22/5/08 15:36 Página 256

3. Reglas básicas para la formulación de preguntas

Veamos ahora un par de consejos a la hora de redactar las preguntas del cuestionario. La primera regla que se puede mencionar tiene que ver con la cuestión de que cada pregunta *sea relevante para los objetivos de la investigación.* No por ser más largo un cuestionario es mejor, sino que cada pregunta que lo compone debe tener una razón de ser; si la misma no tiene una utilidad visible se debe prescindir de ella[5].

Una segunda regla, puede ser aquella que nos dice que las preguntas deben ser formuladas de forma que la *respuesta sea directa e inequívoca sobre el punto de información deseado.* Para ello, se pueden introducir en el propio enunciado de la pregunta las opciones de respuesta o bien el entrevistador leerá o mostrará a través de una tarjeta la lista de categorías que componen las respuestas, según ya se lo ha expuesto más arriba.

Una tercera regla implica que se debe utilizar un *lenguaje sencillo*, a fin de que las preguntas sean comprendidos por todos, independientemente del nivel educativo y social de las personas entrevistadas. Por otro lado, las preguntas *nunca deben estar sesgadas*, es decir no deben estar formuladas de manera que inciten a una determinada respuesta. Ejemplo: *La mayoría de los argentinos cree en Dios, ¿comparte Ud. esta creencia?*

Una cuarta regla considera que las preguntas deben ser *claras y entendidas de la misma manera por todos los entrevistados.* Ocasionalmente surgen preguntas que no logran obtener la misma interpretación por parte de los entrevistados y deben ser evitadas.

En las preguntas que se piden frecuencias o cantidades, el investigador debe establecer sus propios criterios y resolver la ambigüedad a base de respuestas concretas.

Ejemplo:

¿Con qué frecuencia suele Ud. ir al cine?, se le pueden dar las categorías de respuesta: *mucho, bastante, poco, muy poco o nunca*, pero para evitar subjetividades (el mucho o poco es relativo según los parámetros subjetivos de cada persona), se recomienda formularla de la siguiente manera:

¿ Con qué frecuencia suele Ud. ir al cine?

Al menos una vez a la semana
Al menos una vez al mes
1 menos una vez al trimestre
Al menos una vez al año
Nunca o casi nunca

Otras sugerencias que surgen de la bibliografía sobre el tema tienen que ver con:

manual metodologia 22/5/08 15:36 Página 257

• *Evitar las palabras abstractas*, como "tipo", "clase" y "especie": *Ejemplo:*
¿ Qué tipo de religión practica?

• Cada pregunta debe contener *una sola idea.* No hacer preguntas que incluyan dos preguntas, dado que se puede estar de acuerdo en una parte de la frase y no en la otra. *Ejemplo: ¿ Está Ud. de acuerdo con que el Gobierno debería de gastar más en educación y menos en defensa?*

• Las preguntas deben ser *cortas y concisas*; de esta manera se gana rapidez y fluidez en la entrevista.

• Las preguntas deben ser específicas y *hacer referencia al hecho concreto por el que se quiere preguntar*. Por ejemplo en la pregunta: *¿Votó Ud. en las últimas elecciones?* el entrevistado puede contestar refiriéndose a elecciones nacionales, provinciales o municipales; por lo que la correcta formulación debería ser:
¿Votó Ud. en las última elecciones nacionales del 2007?

• La redacción de las preguntas debe hacerse de forma *personal y directa*, con el fin de que los entrevistados sientan que interesan sus opiniones y su caso concreto.

• Las preguntas deben ser *formuladas de forma positiva*, empezar por una negación o plantearla de forma negativa puede dar lugar a dudas en el sentido de la respuesta.

• *Evitar palabras técnicas, abreviaturas o siglas.* Si es necesario formular varias preguntas referidas a la misma institución es necesario poner en la primera el nombre entero y a continuación y entre paréntesis las siglas.

• Se deben *evitar preguntas que obliguen a cálculos o esfuerzos de memoria*, dado que son difíciles de contestar y pueden llevar a la respuesta rápida para salir del paso.

4. Categorías de respuesta

Las categorías u opciones de respuesta de las preguntas también tienen una serie de reglas básicas a seguir:

• *Las categorías tienen que ser exhaustivas,* es decir, deben abarcar todas las respuestas que puedan darse de forma que un entrevistado no se encuentre en la imposibilidad de contestar por no encontrar la categoría adecuada a su caso o a su opinión. Muchas veces esto se resuelve incorporando la categoría residual *otros*, en el cual debería caer solamente una pequeña proporción respuestas, de lo contrario se debe reformular el sistema de categorías planteado.

• *Las categorías tienen que ser excluyentes,* el entrevistado no puede encuadrarse en dos categorías distintas, porque las dos corresponden con un solo caso. Es un error común en las preguntas de intervalo, que el intervalo anterior termine en la misma can-

manual metodologia 22/5/08 15:36 Página 258

tidad que comienza el siguiente:

Ejemplo manera incorrecta: ¿ Cuál es su edad? 0-5 años; 5-10 años; 11-15 años; 15-20 años.

Ejemplo manera correcta: : ¿ Cuál es su edad? 0-5 años; 6-10 años; 11-15 años; 16-20 años.

• *Las categorías de respuesta No sabe/ No contesta,* facilitan la recogida de datos y la codificación, pero nunca se le deben leer al entrevistado.

5. El uso de tarjetas

En los cuestionarios con frecuencia hay que recurrir a *tarjetas complementarias* para que la persona entrevistada visualice las opciones de respuesta. Esta estrategia se lleva a cabo porque el entrevistado nunca debe tener en su mano el cuestionario. Por lo tanto cuando sea necesario que vea las categorías de respuesta, se le muestra una tarjeta en la que aparece la parte del cuestionario que interesa que el entrevistado tenga a la vista.

En estas tarjetas, no debe aparecer la formulación de la pregunta ni las categorías de respuesta *"No sabe"*, *" No contesta"*. Por otro lado, en los cuestionarios se debe indicar, al final de la pregunta, que el encuestador debe mostrar la tarjeta que corresponda:

(ENTREVISTADOR: mostrar tarjeta A).

El empleo de tarjetas es siempre preferible cuando las alternativas de respuesta son muy numerosas, puesto que si se las lee al entrevistado, éste podría olvidar algunas y tender a seleccionar las leídas en último término. Por lo demás, suele variarse el orden en forma aleatoria en las tarjetas, porque ello evita el sesgo o tendencia sistemática a seleccionar siempre las primeras (o bien las últimas).

6. El encolumnado del cuestionario

Cada variable volcada en el cuestionario lleva aparejada un número o números que va entre paréntesis y que significan el lugar que ese dato ha de ocupar en una matriz de datos.

El cuestionario se encolumna para poder interpretar estos datos y saber qué variable o pregunta corresponde a cada columna o cada grupo de columnas

Ejemplo:

Estado civil (col.17)
Soltero 1
Casado 2
Separado 3

manual metodologia 22/5/08 15:36 Página 259

Divorciado 4
Viudo 5

Como se ve en el ejemplo, por un lado en el cuestionario aparecen el número de columna que indica la ubicación de la variable estado civil en la matriz de datos y por el otro existen códigos sobre los distintos valores que pueden tomar esas variables. El libro de códigos acompaña todo estudio para informar sobre los valores o códigos pertinentes para cada pregunta.

El encolumnado del cuestionario se realiza siguiendo los siguientes criterios: cada variable o pregunta del cuestionario deberá identificarse de forma inequívoca con una o varias columnas donde se transcriben los datos para luego poder utilizarlos y analizarlos. En cada columna sólo puede transcribirse un dígito, por lo tanto cada variable o pregunta debe constar de tantas columnas como dígitos sean necesarios para poder recoger los datos requeridos. El número de la columna se suele poner siempre entre ( ), para no confundirlo. Cada cuestionario comienza con la columna número 1 y se continúa numerando correlativamente hasta el número de la última columna requerida. Las preguntas abiertas y semicerradas llevan dos columnas si se prevé que pueden dar lugar a múltiples respuestas

Ejemplos :

Sexo (c.1)
1 Masculino
2 Femenino

¿Qué opinión le merece la pena de muerte?(c2-c3)

¿A qué religión pertenece? (c4-c5)
1 Catolicismo
2 Protestantismo
3 Judaísmo
4 Budismo
5 Ninguna
8 Otras (especificar:……..)
9 Ns/Nc

7. Orden lógico del cuestionario

El orden lógico de un cuestionario está relacionado con los momentos por los que irá transitando la persona encuestada a medida que se desarrolla la encuesta.

manual metodologia 22/5/08 15:36 Página 260

La bibliografía especializada en el tema sugiere que el cuestionario debe comenzar con una introducción que explique los objetivos y advierta el anonimato de las respuestas. Se debe empezar por preguntas fáciles y generales que susciten el interés de la persona entrevistada y que no lo predispongan en contra de la entrevista.

Las preguntas más importantes deben ir hacia la mitad del cuestionario; mientras que las sensibles y las sociodemográficas se sugiere colocarlas al final, cuando ya se ha ganado la confianza del entrevistado, a excepción de aquellas preguntas filtro[6]. La entrevista debe transcurrir suave y con lógica de un tema a otro y avanzando de las preguntas más generales a las más específicas.

Johan Galtung (1978) nos dice que el problema del orden del cuestionario, es el problema de la dinámica de la recolección de datos vista como un proceso en el cual se va desarrollando la interacción entre dos partes: el entrevistador y el entrevistado. Este autor sugiere el siguiente sistema de fases parsoniano aplicado a los cuestionarios, el cual atraviesa como en todo proceso de interacción normal, fases que van desde la adaptación, logro de la meta, integración y latencia.

..." Se introduce el instrumento con algunas preguntas que no plantean problemas, por ejemplo, del tipo de variables de base, y se emplea un enfoque gradual al área central del problema. A través de este método se introduce gradualmente al sujeto no solo al problema sino también al rol del sujeto sometido al instrumento..." (Galtung,1978)

Fase/ Término	**Definición general**	**En términos de instrumento**
1-Adaptación	Exploración, indagación, calentamiento	Preguntas introductorias, fáciles
2-Logro de la meta	Realizar la tarea, la fase principal	Las preguntas principales: complicadas, emotivas
3- Integración	Relajación de la tensión, enfriamiento	Preguntas fáciles: oportunidad para expresar el sentimiento
4- Latencia	Suspensión de la interacción	Espera del próximo cuestionario

Otros autores, antes citados, entre ellos Padua (1987), centran el debate respecto al orden lógico del cuestionario, partiendo de la motivación del entrevistado. La entrevista es vista como un proceso que refleja la motivación del entrevistado dependiendo del objetivo del estudio, el tipo de preguntas, la disponibilidad del sujeto y la duración de la entrevista entre otros. Consideran dos factores de importancia en lo concerniente al cuestionario como unidad, que afectarán directamente la motivación: el orden de las preguntas y el tamaño del cuestionario.

manual metodologia 22/5/08 15:36 Página 261

Entre las recomendaciones que se sugieren respecto del orden del cuestionario, se puede mencionar que antes que nada se apuntará a lograr que el entrevistado se relaje y tenga confianza en el entrevistador. Para ello lo mejor será recurrir a preguntas sobre áreas de especial interés para el entrevistado como deportes y otras actividades del tiempo libre y algunas preguntas sobre información de los medios masivos de comunicación, diarias, revistas o televisión.

En cuanto al tamaño del cuestionario, un cuestionario demasiado corto lleva a una pérdida de información y no le da el tiempo suficiente al encuestado de compenetrarse del problema. Por otro lado, un cuestionario demasiado largo no es aconsejable, ya que podría ser difícil mantener en un nivel adecuado el interés del entrevistado.

8. Aspectos formales del cuestionario

En la bibliografía sobre el tema, también nos encontramos con una serie de convenciones sobre aspectos formales de los cuestionarios; entre ellos se sugiere que las preguntas deben ser numeradas para poder ser identificadas. Si bien cada investigador puede decidir la fórmula a seguir, recurriendo a números o letras, el objetivo es mantener un orden lógico que pueda diferenciar las diversas preguntas.

Las posibilidades de respuesta también debe llevar un código numérico que identifique las respuestas de otras de forma inequívoca. En el caso de las escalas, éste número o código tiene un valor por sí mismo para su tratamiento estadístico posterior.

Ejemplo:

5 Muy de acuerdo
4 Algo de acuerdo
3 Ni de acuerdo Ni desacuerdo
2 Algo en desacuerdo
1 Muy en desacuerdo

Por el contrario, en otro tipo de preguntas, los códigos significan sólo la traducción de las categorías a números, para facilitar su manejo informático a la hora de ingresarlos a la matriz de datos.

Ejemplo: ¿ Qué sentimiento le inspira a Ud. la política?

1 Entusiasmo
2 Compromiso
3 Interés
4 Indiferencia
5 Aburrimiento

manual metodologia 22/5/08 15:36 Página 262

6 Desconfianza
8 No sabe
9 No contesta

Como se ve en este ejemplo los códigos 8 y 9 en general mantienen la rutina de estandarizar las respuesta de *No Sabe y No Contesta* respectivamente, dejando en blanco los valores intermedios si no requieren ser utilizados. En caso de que las opciones de respuesta superen los dos dígitos, el primer código empleado será el 01 hasta n y los números de las respuestas estandar serán 98 *No Sabe* y 99 *No Contesta*.

Las *preguntas filtro*, antes mencionadas, son aquellas que según la respuesta dada por el entrevistado, requieren que se haga o no la pregunta o serie de preguntas siguientes.

P10 ¿ En particular, el Presidente del Gobierno, le inspira ahora más o menos confianza que cuando accedió a la Presidencia?
1 Más (Pase a P10b)
2 Menos (Pase a P10c)
3 Igual
8 Ns
9 Nc

P10b ¿ Por qué le inspira ahora más confianza?

_____

P10c ¿ Por qué le inspira ahora menos confianza?

_____

Al final de filtros muy largos o complicados, se recomienda poner una nota delante de la pregunta que debe aplicarse a toda la muestra, indicando que a partir de ese punto el cuestionario se sigue preguntando a todos los entrevistados. *Por ejemplo:* ***A TODOS***

Cuando el cuestionario es administrado por un entrevistador, las instrucciones escritas en el instrumento deben ser mínimas ya que todas las pertinentes a cada una de las preguntas deben ser aprendidas en el proceso de capacitación. En el caso de investigaciones extensas, se recomienda un *manual para entrevistadores/encuestadores*, en el que se especifiquen con el mayor detalle características de las preguntas, instrucciones especiales, alternativas, etc.

manual metodologia 22/5/08 15:36 Página 263

9. La prueba piloto

Otro punto a considerar es el de la prueba piloto, siempre infaltable antes de la aplicación del instrumento de recolección de datos de manera definitiva, pero ponderada de manera diferente por los distintos autores que escriben sobre ella.

Para Galtung (1978) de ninguna manera la prueba piloto debe ser considerada abstractamente. No es sólo una prueba del instrumento sino de todo el proceso de recolección de datos y de los primeros pasos del análisis. Es una prueba de la practicabilidad y de la relevancia, por lo que la muestra para esta instancia debe cubrir la amplitud de sujetos posibles y los extremos sociales y actitudinales.

La prueba piloto no es sólo una prueba de los sujetos, sino también lo es de los entrevistadores, observadores o experimentadores. *(Ejemplo: está mal que el director del proyecto haga la prueba si luego él no participa del campo*).

En la bibliografía sobre el tema, se sugiere focalizar los resultados de la piloto en un triple análisis:

- Análisis de los NS/NC viendo su frecuencia y distribución.
- Análisis de respuestas: centrándose en los resultados obtenidos para cada variable y para cada unidad de análisis. *Ejemplo, analizar la distribución de frecuencia de la variable nivel educativo, para saber que niveles de escolaridad concentran la mayor cantidad de los casos, detectando si se trata de universitarios o secundarios completos.*
- *Por otro lado, también el análisis se puede realizar en el sentido de las unidades de análisis, es decir deteniéndose en el comportamiento de ciertos casos a lo largo de las diferentes variables medidas.* Frente a problemas detectados en la piloto, lo que se puede hacer es lo siguiente: arreglar la entrevista, reformular items y dispersar las preguntas demasiado semejantes para que no se presenten juntas.
- Análisis de correlación: si se predice que una o varias variables inciden en el comportamiento de otra u otras variables y a partir de la piloto esta relación no se hace evidente, la tentación puede surgir de buscar otros indicadores o seleccionar otras variables.

Otros autores, entre ellos, Azofra (1999) diferencian la prueba piloto del "pretest". El " pretest "se aplica a una muestra reducida de personas, aproximadamente de 30 a 50 personas con características similares a la muestra real, pero no necesariamente elegidas al azar. Los objetivos principales de esta instancia son: comprobar que todas las preguntas se entiendan y se interpretan de la misma manera; confirmar que las categorías de respuesta sean exhaustivas y, de lo contrario, añadir todas las que el encuestador no previó; cerrar alguna pregunta que se decidió dejar abierta; observar que el orden y la disposición de las preguntas es el adecuado y otorga fluidez al desarrollo de la entrevista; y medir la longitud del cuestionario.

manual metodologia 22/5/08 15:36 Página 264

Por el contrario, en la prueba piloto, la muestra pretesteada se puede decir que es una submuestra de la real. Los entrevistados representan a la población a estudiar y deben ser elegidos al azar. Además de los distintos objetivos enumerados para el "pretest", tiene la finalidad de medir la validez de la encuesta definitiva o de parte de ella, de forma que si un investigador aborda un proyecto muy amplio, sin saber si los resultados que va a obtener van a merecer la pena en cuanto a tiempo, esfuerzo y dinero, los resultados obtenidos de la encuesta piloto le orientan a la realización definitiva del proyecto.

Asimismo, el pretest permite, en ocasiones, estimar ciertos parámetros[7] de la población que resultan necesarios para calcular el tamaño de la muestra.

10. Tipos de cuestionarios

Entre los diferentes tipos de cuestionarios podemos mencionar los autoadministrados, los enviados por correo y por entrevista personal, telefónica, domiciliaria o casual.

La encuesta telefónica al igual que las otras entrevistas personales, requieren de un entrevistador y un entrevistado. Sin embargo, existen una serie de diferencias que las distancian unas de otras.

Primeramente el cuestionario aplicado por teléfono debe ser corto, 15 ó 20 preguntas como máximo, dado que la falta de contacto personal frente al cansancio del entrevistado, pueden llevar a cortar la comunicación.

También las categorías de respuesta deben ser cortas y pocas. La lectura de una lista larga de categorías, implica el peligro que el entrevistado conteste las últimas por olvido de las primeras. Si las frases de las categorías son largas, el entrevistado se puede perder y contestar cualquier cosa con tal que no se le repita la pregunta.

Las preguntas abiertas no se recomiendan en las encuestas telefónicas, dado que resulta molesto para el entrevistado la espera al otro lado del teléfono del tiempo necesario para la transcripción de su respuesta. Este tipo de encuestas telefónicas están de moda porque son menos costosas que la entrevista personal domiciliaria y permiten mayor rapidez en la obtención de datos y en los resultados.

Otra técnica empleada en algunas investigaciones es la del cuestionario autoaplicado o autoadministrado. Este tipo de cuestionario se implementa en diferentes modalidades: una de ellas es la de la aplicación a un grupo de personas con la supervisión de un entrevistador que explica la forma de rellenarlo y que resuelve cualquier duda que se presente durante el tiempo que los encuestados se toman para completarlo.

La dificultad mayor se encuentra en los cuestionarios por correo, en los que el entrevistado debe rellenar las preguntas sin ningún tipo de ayuda, lo que obliga a que el formato y

manual metodologia 22/5/08 15:36 Página 265

diseño del cuestionario deba ser confeccionado minuciosamente. Este tipo de cuestionarios sufre fuertes limitaciones en cuanto a las altas tasas de no respuesta que generan[8].

El cuestionario por correo debe ir acompañado de una carta de presentación capaz de despertar el interés del que la lea, destacando la importancia de su colaboración y garantizando el anonimato de sus respuestas, aún sabiendo que su nombre y dirección son conocidos. Se sugiere que el cuestionario sea corto, para que no suponga un gran esfuerzo para contestarlo. Conviene manejarse con un formato de cuestionario atractivo, ligero, y cómodo de contestar para no aumentar el número de desertores.

11. Limitaciones del método de la encuesta y los cuestionarios

Varias, son las limitaciones citadas en la bibliografía acerca del método de las encuestas en general y de los cuestionarios en su acepción más restringida, debido a que:

- Los indicadores y las variables utilizadas para caracterizarlos se escogen con el propósito específico de un proyecto de investigación.
- Para obtener las respuestas *se necesita el conocimiento del individuo e incluso su cooperación.*
- El cuestionario es un instrumento donde se recogen las respuestas de las personas entrevistadas; siendo a su vez manejado por un grupo numeroso de personas que son los entrevistadores. *Ambos colectivos pueden ocasionar sesgos importantes en la investigación* si por ejemplo a los entrevistados se los aborda con preguntas que no sean claras, concisas, ordenadas y no sesgadas, y a los entrevistadores no se les marcan las normas precisas y concretas en el propio cuestionario quedando bajo su libre interpretación.
- *El método de encuestas es demasiado individualista:* por esa tendencia a tratar al individuo como una unidad social. Sólo el individuo puede ser entrevistado y sólo a él se le puede aplicar una encuesta[9]. Por otro lado, se introduce también el individualismo con el modelo probabilístico de muestreo ; en la medida que el individuo es arrancado de su contexto social y se lo ve como muestra de una sociedad formada por una sola persona que debe ser comparada con otras sociedades de una persona.
- *El método de la encuesta es demasiado democrático:* porque se le pregunta a todo el mundo y existe siempre más de una alternativa de respuesta.
- *El método de encuesta es demasiado estático*: porque produce reacciones verbales manifiestas en un punto determinado del tiempo y nada más. La encuesta no es buen instrumento para detectar cambios bruscos y es inadecuado en sociedades que hay cambios repentinos y de gran magnitud[10].

manual metodologia 22/5/08 15:36 Página 266

- *El método de encuesta está restringido a un rango medio de posición social:* pues presupone algún tipo de interacción verbal y llega a los extremos de la estructura social en que el tipo de interacción verbal es practicable. Pueden quedar excluidos extremos bajos o muy altos: analfabetos, ancianos, no participantes, destituidos, vagabundos. Las elites, usualmente, también suelen quedar fuera porque no quieren demostrar sus posiciones públicamente[11].
- El encuestado ideal presenta las siguientes características: está bien socializado y disciplinado, acostumbrado a los exámenes, a escuchar y contestar honesta y claramente. Encadenado con lo anterior se puede ver como el método de encuestas trabaja solamente a través de distancias relativamente cortas, porque es adecuado para el endo-grupo y no para el exo-grupo. Presupone interacción social si no hay conflicto. El método de la encuesta es el ideal en aquellas situaciones en las que el dato a recabar es proporcionado por unidades de análisis que no ponen resistencia a ser indagadas por el investigador social, a quien consideran dentro de su grupo social o de pertenencia y que consideran a las encuestas y a los resultados que de ellas se publican como una práctica habitual de la época que les toca vivir.

12. Ventajas del método de la encuesta

Sin embargo, más allá de las limitaciones mencionadas, el método de la encuesta y los cuestionarios se siguen aplicando ampliamente. Entre las razones del éxito como instrumentos de recolección de datos, sobresalen en la bibliografía los siguientes:

- Permiten obtener datos teoréticamente relevantes.
- Los resultados arrojados por las encuestas son "susceptibles de tratamiento estadístico", lo que posibilita la utilización del análisis de correlación y multivariado.
- Presuponen la utilización de las herramientas de los test estadísticos de hipótesis acerca de la posibilidad de generalizar desde las muestras a los universos.
- Son más rápidos y menos costosos de aplicar que otras técnicas de recolección datos.
- Favorecen la comparabilidad de los resultados obtenidos en distintas muestras y unidades de análisis aplicando el mismo instrumento de recolección de datos.
- Resultan adecuados para obtener información proveniente de poblaciones muy numerosas.

ANEXO: EL PROCESO DE CODIFICACIÓN

Una de las tareas por las que indefectiblemente se debe atravesar en toda investigación

manual metodologia 22/5/08 15:36 Página 267

es la codificación. Una vez concluido el trabajo de campo y a fin de volcar la información recabada en una matriz de datos, es menester pasar por esta tareas, rutinarias y mecánicas pero no por ello menos necesarias e importantes.

Resumidamente, podemos decir que la codificación consiste en asignar a cada respuesta obtenida en nuestro trabajo de campo un código numérico.

El procedimiento de codificación como veremos más adelante, varía según se efectúe sobre preguntas cerradas, semicerradas o abiertas; de acuerdo al tipo de pregunta, la codificación tendrá características particulares. La codificación como parte del proceso de investigación se comienza a definir desde el mismo momento que se diseña el cuestionario, por ello nos pareció oportuno incluirlo como anexo al capítulo correspondiente a este instrumento de recolección de datos.

1. Algunas definiciones de codificación

Según Maurice Duverger (1961) la codificación consiste en asignar un número a cada categoría de respuesta. Y conforme a los tiempos por los cuales éste autor escribía, agrega que este número determinará el lugar correspondiente a la perforación sobre la ficha que sirve de base a los escrutinios mecanográficos; sin embargo en la actualidad el sistema de tarjetas ha sido abolido y reemplazado por procedimientos computarizados.

Por su parte, Selltiz (1962) y otros autores ponen el acento en el criterio subjetivo del codificador. Para ellos, la codificación se encuadra en el procedimiento técnico por el cual los datos originales son transformados en símbolos (generalmente números) de modo tal que ellos pueden ser tabulados y contados. Dicha transformación no es automática, sino que envuelve un juicio de parte del codificador.

2. Codificación y categorización como operaciones diferentes pero simultáneas

Si bien la categorización y la codificación de los datos se realiza simultáneamente, a nivel conceptual se trata de dos aspectos diferentes.

La *categorización* de los datos se realiza antes que la *codificación*, y se da en la medida en que se consigna la respuesta del sujeto dentro de las categorías generales de las variables

(por ejemplo, si se trata de la variable estado civil, las categorías posibles son soltero, casado, viudo, divorciado, etc), con independencia del número de código que les corresponda.

Ahora bien, en la medida que se coloca un número a cada una de las categorías de respuestas posibles, ya hablamos de codificación, dado que el número indica el símbolo en

manual metodologia 22/5/08 15:36 Página 268

que la codificación corresponde a cada respuesta (conforme al ejemplo anterior: código 1 si se trata de un individuo soltero; 2 si es casado; 3 si es viudo , 4 si es divorciado, etc.)

3. Distintas clases de preguntas y codificación

La codificación se realiza sobre la base de un código y éste es construido de manera distinta según sea el tipo de preguntas que integren el cuestionario. Tal como mencionamos en el capítulo de cuestionarios, podemos hablar de dos grandes grupos: las preguntas abiertas, en las que el entrevistado puede contestar libremente; por lo tanto el número de respuestas que pueden darse, resulta teóricamente ilimitado; y por otro lado, las preguntas cerradas que obligan al entrevistado a optar por una serie de posibles respuestas previamente consignadas. En las preguntas abiertas el proceso de codificación se realizará una vez concluido el trabajo de campo, en la mayoría de los casos; mientras que en las cerradas, hablaremos de precodificación y la tarea del codificador será mucho más sencilla.

¿Qué es un código? En términos generales, puede decirse que el código es el conjunto de símbolos correspondientes a cada una de las categorías en que son divididas las respuestas obtenidas. El establecimiento de un libro de códigos implica entonces la clasificación de la pregunta, el ordenamiento de sus respuestas y la traducción a un símbolo numérico. A fin de realizar esta clasificación, el codificador debe considerar que la misma ha de ser completa, lo suficientemente clara y precisa, no tener un carácter demasiado teórico y adaptarse a las necesidades más generales en lo posible.

4. Codificación de preguntas cerradas

En las preguntas cerradas no se admite otra posibilidad de respuesta más que las predeterminadas en el cuestionario. La tarea del codificador consiste en transcribir el código circulado por el encuestador al casillero correspondiente. En general, está situado en el margen derecho de la pregunta y en él aparece un número impreso que corresponde al número de pregunta.

En este tipo de preguntas, el recurso de colocar junto a cada respuesta el código correspondiente tiene una finalidad práctica: ahorra esfuerzo, evitando la duplicación innecesaria de tareas y liberando al codificador de buscar cuál es el número correspondiente a cada respuesta de este tipo.

Ejemplos

P.15 ¿ Cómo calificaría en términos generales la gestión del Presidente Kirchner?

manual metodologia 22/5/08 15:36 Página 269

1- Muy buena
2- Buena
3- Regular
→ 4- Mala
0- No sabe /No contesta

15	4

P.16 ¿Cuál es el máximo nivel de educación formal alcanzado por el principal sostén del hogar?

1-Primaria incompleta o completa
2-Secundaria incompleta
3-Secundaria completa
→ 4-Terciaria o universitaria incompleta
5-Terciaria o universitaria completa
0- No sabe /No contesta

16	4

5. Codificación de preguntas semicerradas

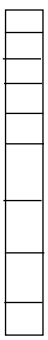
Son las preguntas que si bien tienen una cantidad de códigos predeterminados admiten otras posibilidades de respuesta además de las enunciadas en el cuestionario.

En el caso de que se encuentre una respuesta de las predeterminadas circulada por el encuestador, el codificador debe transcribir el código al casillero correspondiente, al igual que en las preguntas cerradas.

En el caso de que el encuestador haya anotado alguna respuesta no contenida dentro de las predeterminadas, el codificador deberá otorgarle un nuevo código a esa respuesta y especificarla en el libro de códigos, del cual hablaremos luego. El codificador debe ser cuidadoso en no superponer el nuevo código con algún otro existente.

P10¿ Cuál es su hobby preferido?

Ver televisión
Escuchar radio
Ver videos
Practicar deportes
Asistir a recitales
Ir al cine
Ir a Pubs/discoteques
Viajar
Leer
Otros (especificar :*Navegar en Internet* ...)



10	12

manual metodologia 22/5/08 15:36 Página 270

Como el entrevistado respondió " navegar en Internet" que es una respuesta ausente en la lista de las opciones predeterminadas, el codificador debe darle a esta respuesta un código nuevo. Los códigos nuevos se enumerarán a partir del 10 en adelante y el número que efectivamente se asignará dependerá de la cantidad de nuevas respuestas que hayan salido con anterioridad. Si una respuesta aparece con poca frecuencia no ameritará la apertura de un nuevo código y se ingresará en la base de datos con el código correspondiente al "otros".

6. Codificación de preguntas abiertas

Las preguntas abiertas, son preguntas que no tienen respuestas predeterminadas. En este caso, la cantidad de códigos dependerá de la cantidad de respuesta que hayan dado los entrevistados.

La tarea del codificador consistirá en clasificar cada respuesta con un código y marcar ese código en el casillero de codificación que se encuentra generalmente, al margen derecho de la pregunta. Para codificar las preguntas abiertas, el codificador requerirá de un libro de códigos que explicaremos a continuación.

El libro de códigos es el instrumento que el codificador utilizar para clasificar las respuestas que no hayan sido predeterminadas en el cuestionario. Cada pregunta contará para sí con una hoja de códigos, en la cual se van anotando cada una de las respuestas que vayan saliendo del cuestionario.

En general, el código (0) ó (00) corresponde para clasificar las no respuestas, por lo que cada hoja de códigos comenzará con el código (1) o (01)

Ejemplo

3- ¿ Por qué se incorporaron nuevas tecnologías de información y comunicación en esta dependencia estatal?

3	01

.....*Para agilizar procesos y hacerlos más eficientes*.....

Supongamos que la respuesta del primer cuestionario que estamos codificando sea " para agilizar procesos y hacerlos más eficientes", como lo muestra el ejemplo, a esta respuesta le corresponde el código 01. Este código se utilizará siempre que la respuesta de cualquier entrevistado haga referencia a eficiencia de procesos o agilización de los mismos.

Así cada vez que surja una nueva respuesta será clasificada con el código que le sucede en forma inmediata.

manual metodologia 22/5/08 15:36 Página 271

Veamos un ejemplo de una hoja de códigos: en la parte superior se transcribe la pregunta del cuestionario y posteriormente se van enumerando códigos enunciados por los entrevistados.

3- ¿ Por qué se incorporaron nuevas tecnologías de información y comunicación en esta dependencia estatal?
1. Agilización de procesos / eficientizar
2. Transparencia
3. Requerimientos legales/ obligatoriedad
4. Despapelizar/ digitalizar documentos
5. Control
6. Base de datos
7. Comunicación
8. Obtener información para decisiones

La codificación de preguntas abiertas, presupone un proceso que lleva a la elaboración definitiva del libro de códigos. En primer lugar se establece un " precódigo " fundado en un análisis hipotético de los tipos de respuestas posibles. Este precódigo se aplica después a una muestra de entrevistas lo que permite rectificar experimentalmente sus categorías y establecer el código propiamente dicho, que , por lo general, es sometido nuevamente a una verificación, en la que varios codificadores los aplican a una muestra de cuestionarios.

En cuanto a recomendaciones a tomar en cuenta a la hora de codificar preguntas abiertas, podemos mencionar:

- Es importante familiarizarse no sólo con el código, sino además con las diferentes categorías en que se dividió cada pregunta, de modo tal que no se generen dudas acerca de la codificación de las respuestas.
- Deben codificarse las ideas y no las palabras.
- Se debe considerar la categoría denominada "otras respuestas" en cada pregunta, para prever la probabilidad de que una respuesta no pueda ser codificada dentro de las otras categorías establecidas. Sin embargo, si el número de respuestas que entran en esta categoría es muy elevado, será necesario crear nuevas categorías y revisar la división de categorías original.

7. Codificación de respuestas múltiples

En reiteradas ocasiones las preguntas tienen posibilidad de múltiples respuestas, es decir que para una pregunta, el entrevistado puede elegir más de una opción al responder.

manual metodologia 22/5/08 15:36 Página 272

La cantidad de respuestas que se pueden codificar está determinada por la cantidad de casilleros de codificación que incluya la pregunta.

Ejemplo: Respuesta múltiple en pregunta cerrada

P30 ¿Qué aspectos mejoraría del barrio al que pertenece? – Menciones tres en orden de importancia-c.25-27

1-Servicios de salud 1ro.....05
2- Sistema educativo 2do.....03
3- Policía 3ero.....01
4-Transporte
5-Seguridad del barrio
9-Otros
10- No sabe/ No contesta

25	05
26	03
27	01

Se asignaron 3 columnas para esta pregunta por ser múltiple

Ejemplo: Respuesta múltiple en pregunta abierta

P31 ¿Si Ud. necesita ayuda financiera, a quien se la pide? – Menciones tres respuestas en orden de importancia-c.35-37

1ero
.....*Pareja*.....
2do
..... *Vecino*.....
3ero
.....*Amigo*.....

35	03
36	05
37	01

Se asignaron 3 columnas para esta pregunta por ser múltiple

8. Compatibilidad de las preguntas

Decimos que dos o más preguntas son compatibles entre sí cuando las respuestas que dan los entrevistados son similares o equivalentes. En este caso las respuestas se pueden codificar con la misma hoja de códigos.

Por ejemplo, si preguntamos: P.9 ¿Cuál es el problema de la ciudad de Buenos Aires? Y

manual metodologia 22/5/08 15:36 Página 273

a continuación preguntamos también: P10 ¿Cuál es el principal problema de su barrio?

Las respuestas enunciadas por los encuestados seguramente serán similares para ambas preguntas podrán codificarse utilizando la misma hoja del libro de códigos. En la hoja del libro de códigos se deberá aclarar que comprende las menciones de ambas preguntas.

9. Rango de una pregunta

El rango máximo de una pregunta es la cantidad máxima de códigos que admite esa pregunta. En el caso de preguntas cerradas el rango de la variable se encuentra determinado por la cantidad de respuestas posibles asentadas en el cuestionario.

Ejemplo:

P.15 ¿ Cómo calificaría en términos generales la gestión del Presidente Kirchner?

1-Muy buena
2-Buena
3-Regular
→ 4-Mala →
0-No sabe /No contesta

15	4

Esta pregunta tiene 4 rangos -muy buena, buena, regular y mala- porque el código no sabe /no contesta no es contemplado como valor de la variable. En el caso de las preguntas semicerradas y abiertas no es posible predeterminar las categorías de respuestas. A fines de facilitar la codificación se establece un número determinado de cantidad de respuestas permitidas materializadas en la cantidad de dígitos que admite el casillero de codificación.

Lo más frecuente es que las preguntas de este tipo tengan rango de 9 (hasta 9 respuestas posibles) o rango 99 (hasta 99 respuestas posibles). Las preguntas abiertas, son las que normalmente tienen rango 99 debido a que se desconoce a priori las diversas y posibles respuestas dadas por los entrevistados.

10. La codificación como tarea grupal

En general, la codificación es una tarea que se realiza en grupo, por lo que es imprescindible un buen funcionamiento del equipo de trabajo, especialmente al codificar las preguntas abiertas.

manual metodologia 22/5/08 15:36 Página 274

En la medida que los codificadores trabajen en el mismo lugar y en el mismo momento, se evita que se repitan códigos o que se utilicen diferentes criterios para codificar una misma respuesta; de esta manera se garantiza que cada código que se abra figure en el libro de códigos de cada uno de los codificadores.

Cada codificador tiene un libro de códigos y además existe un libro de códigos consolidado que lleva el encargado de la codificación. Cada uno de los codificadores antes de abrir un nuevo código debe consultarlo con aquel que lleve el consolidado a fin de lograr un único criterio.

11. A modo de conclusión

Terminado el campo, comienza la fase de codificación. Los codificadores corrigen errores, cierran preguntas abiertas, controlan el trabajo del encuestador y consiguen que el cuestionario quede en condiciones para ser grabado en un sistema informático que permita la tabulación y obtención de resultados estadísticos. Por otra parte, los codificadores deben elaborar un libro de códigos que refleje el significado de los códigos y todas aquellas cuestiones que hayan podido surgir en la codificación de los cuestionarios.

Bibliografía

- AZOFRA, M. J. (1999). *Cuestionarios* en Cuadernos Metodológicos Nro. 26. CIS (Centro de Investigaciones Sociológicas). Madrid.
- CALI, M. L. (1999). Apuntes de cátedra de "Metodología y Técnicas de Investigación en Ciencias Sociales II" USAL. Buenos Aires.
- GALTUNG J. (1978) T*eoría y Método de la Investigación Social* . Tomo I. Buenos Aires: EUDEBA.
- PADUA, J. (1987) *Técnicas de Investigación aplicadas a las ciencias sociales*. México: Fondo de Cultura Económica.

Notas

1 Lo que incrementaría su validez.

2 Obviamente, esta pregunta estaría precedida por otra acerca de si el entrevistado está afiliado a algún partido político. Sólo sería formulada a quienes respondieran afirmativamente.

3 Lo que afectaría su fiabilidad y validez.

4 El caso de la EPH que lleva a cabo el INDEC en las principales ciudades de la Argentina.

5 No debe olvidarse que alargar innecesariamente el cuestionario incrementa los costos, tanto del trabajo de campo (los encuestadores cobran, habitualmente, en relación con la duración de la entrevista) como de la pos-

manual metodologia 22/5/08 15:36 Página 275

terior carga de datos.

6 Una vez más, cabe señalar que esto sucede así en las encuestas de opinión: no en las encuestas de hogares o de condiciones de vida, en las que las preguntas sociodemográficas integran el cuerpo principal y las preguntas de opinión –cuando las hay– cumplen un papel subordinado.

7 Se trata de ciertas características que se describen a través de medidas estadísticas, tales como la varianza. En otro lugar de este texto se volverá sobre este tema.

8 La no respuesta, además de achicar la muestra, supone un serio problema si existen razones para sospechar que las personas que no contestan se diferencian de los que sí lo hacen en algunos aspectos relevantes: su nivel socioeconómico, su educación o sus opiniones o actitudes, etc.

9 Ello se amplía al grupo conviviente en las encuestas de hogares y censos.

10 Lo cual resulta parcialmente compensado cuando las encuestas sobre los mismos temas se repiten con cierta asiduidad y regularidad. La mayoría de los países incorporan en sus programas estadísticos encuestas regulares sobre una variedad de temas, que permiten seguir la evolución tendencial de muchos fenómenos.

11 Respecto de las elites una manera de reemplazar el cuestionario es usar datos proporcionados por la elite, pero no solicitados por el sociólogo, tales como discursos, artículos, etc.

manual metodologia 22/5/08 15:36 Página 313

13. EL ANÁLISIS DE LOS DATOS: TÉCNICAS DE ANÁLISIS CUANTITATIVO.
ANÁLISIS DESCRIPTIVO.
LA LÓGICA DE LA COMPARACIÓN. CRUCES DE VARIABLES.

Horacio Chitarroni

1. Las herramientas estadísticas

En este capítulo se expondrán algunas técnicas que posibilitan el análisis de los datos cuantitativos, ya provengan éstos de recolecciones propias (usualmente cuando hemos aplicado una encuesta) o bien se trate de encuestas realizadas por otros o de bases de datos ya construidas y disponibles en formato digital. Es decir, cuando disponemos de información referida a un conjunto relativamente amplio de variables para una cantidad apreciable de casos.

En ninguna de estas oportunidades es posible servirse directamente de la matriz de datos en el análisis. Ella contiene la totalidad de la información, totalmente desagregada. Y, obviamente, no se puede leer una matriz de esta clase ni proceder a comparar caso a caso. Por otra parte, cuando se dispone de información cuantitativa no nos interesamos por los casos individuales: ellos pierden importancia y lo que cuenta son las características agregadas. Es decir, el conjunto de todos los casos, o bien subconjuntos de aquellos que tienen algo en común. Pero los individuos por sí mismos carecen de toda relevancia: podría decirse que se trata de la *desaparición estadística del caso.*

En los análisis más elementales, nos interesamos en las distintas variables consideradas de a una: ello ocurre cuando solo pretendemos describir: ¿cómo es la estructura de edades de la población argentina? Pero así como avanzamos en descripciones más complejas o cuando pretendemos explicar o predecir ciertos fenómenos, entonces lo común es que debamos prestar atención al comportamiento de varias variables a la vez: dos o más. Por ejemplo, ¿difiere la estructura de edades entre varones y mujeres? (presumiblemente debiera ser así, porque las mujeres son más longevas). El sexo, ¿influye sobre la probabilidad de terminar el ciclo secundario? (muchos piensan que sí, porque los varones suelen incorporarse antes al mercado de trabajo y, al hacerlo, a veces abandonan los estudios).

¿Cómo podemos hacerlo? Necesitaremos de algunas herramientas que nos permitan resumir la información. La estadística, en sus usos más básicos, provee de tales herramientas. La estadística es una rama de la matemática que ha cobrado considerable desarrollo en

manual metodologia 22/5/08 15:36 Página 314

tiempos más o menos recientes: en particular, desde el siglo XIX y, más fuertemente, en el siglo XX. Se trata de una disciplina auxiliar, aplicable a un sinnúmero de problemas en casi todas las ciencias fácticas. Su relación con las ciencias sociales fue temprana, porque justamente, gran parte de su desarrollo se debió a los esfuerzos de los Estados modernos por traducir cuantitativamente sus necesidades y recursos. En los últimos años su potencial se incrementó en mucho merced a la difusión de los ordenadores personales y algunas disciplinas –particularmente la economía– han avanzado en la construcción de modelos estadísticos muy sofisticados para describir, explicar y predecir sus fenómenos. Otras ciencias sociales hacen un uso más modesto y limitado de las técnicas estadísticas, en especial por la dificultad de medición que presentan algunas de las variables que estudian. Sin embargo, no pueden en modo alguno prescindir de ellas en sus análisis, aunque no son, por cierto, las únicas: como ya está dicho no toda la evidencia empírica puede cuantificarse.

2. La descripción de la realidad social

En términos metodológicos, la descripción es lógicamente previa a la explicación. Ya se ha visto que, al proponer objetivos de investigación, los descriptivos preceden a los explicativos. Primero hemos de conocer qué cosas suceden, para luego conjeturar –y eventualmente comprobar– por qué razones suceden.

En esta parte, nos ocuparemos brevemente de los modos más adecuados de describir. Por supuesto que los propósitos descriptivos, en las ciencias sociales, pueden referirse a una enorme variedad de fenómenos. Podríamos querer describir cosas tan variadas como los equilibrios de fuerzas entre diferentes grupos sociales o políticos, la estructura demográfica de una población, sus hábitos de consumo o de empleo del tiempo libre, los desniveles de desarrollo entre países de una región o entre jurisdicciones de un mismo país y muchas cosas más.

A veces, las descripciones hacen uso de una variable por vez –la estructura educativa de la población en general–, mientras que en otras ocasiones se apela también a comparaciones: las estructuras educativas de los varones y las mujeres; en este segundo caso, la variable sexo nos permite separar entre diferentes subgrupos a describir. Sin embargo, el uso de al menos dos variables a la vez también es pertinente cuando se asumen propósitos explicativos: los procedimientos correspondientes serán examinados más adelante.

3. Tratamiento univariado de datos categóricos

Ya hemos visto que las variables admiten una gruesa clasificación entre aquellas cuyos

manual metodologia 22/5/08 15:36 Página 315

valores son numéricos (como las edades o los ingresos de las personas o las tasas de mortalidad infantil o el PBI de los países) y aquellas otras que sólo agrupan –y eventualmente ordenan– a los casos en categorías cualitativas.

Esta segunda clase de variables se denominan genéricamente variables *categóricas*: repárese que en vez de *valores* –término reservado a las magnitudes numéricas– tienen *categorías*. Las variables categóricas pueden ser clasificadas, a su vez, en *nominales* y *ordinales* en función de su nivel o capacidad de medición (según las categorías asuman un orden arbitrario o natural).

La pertenencia regional de los países (OCDE, América Latina, Sudeste Asiático, Europa Oriental, Africa, Oriente Medio) o el lugar de origen de los migrantes limítrofes que habitan en Argentina (Bolivia, Brasil, Chile, Paraguay, Uruguay), o bien la pertenencia al sistema público o privado de los establecimientos educativos son variables nominales, que sólo agrupan o clasifican a los objetos o unidades de análisis de las cuales se predican (países, migrantes internos o escuelas, respectivamente).

En cambio el máximo nivel educativo alcanzado por las personas (sin instruccción, primario incompleto, primario completo, secundario incompleto, etc.), el grado de desarrollo humano de los países según la clasificación del Programa de las Naciones Unidas para el Desarrollo (alto, medio alto, medio y bajo) o el nivel a que pertenecen los establecimientos escolares dependientes del Gobierno de la Ciudad de Buenos Aires (preescolar, primario, secundario, terciario) son variables ordinales, porque sus categorías tienen un orden fijo que se respeta al mencionarlas, ya sea de mayor a menor o a la inversa. Los países de origen o las regiones, en cambio, se mencionan en un orden arbitrario o meramente convencional (como puede ser el alfabético), que no establece jerarquía alguna.

3.1. Las distribuciones de frecuencias

En esta parte nos ocuparemos del modo de describir datos categóricos. La herramienta estadística elemental para hacerlo son las tablas o cuadros[1] que presentan distribuciones de frecuencias de una variable, como la que vemos a continuación. Ella muestra la población ocupada de los principales centros urbanos de la Argentina, según su categoría ocupacional, que es una variable que describe la relación de las personas con los medios de producción: se puede ser propietario de un negocio o empresa y emplear personal (patrón o empleador), se puede estar trabajando en relación de dependencia por un sueldo (obrero o empleado) o se puede trabajar en forma autónoma sin contratar personal (cuentapropista). También hay personas que trabajan sin recibir remuneración (por ejemplo, en un negocio familiar). Esta es una mera clasificación que no contempla ningún orden: se trata de una variable de las que hemos llamado nominales.

manual metodologia 22/5/08 15:36 Página 316

Tabla 3.1.1: Población ocupada de 10 años y más según categoría ocupacional. Total de aglomerados urbanos.

Categoría ocupacional	Total	%
Patrón o empleador	357.158	3,5
Trabajador por cuenta propia	2.155.463	21,1
Obrero o empleado	7.524.365	73,8
Trabajador sin remuneración	163.177	1,6
Total	10.200.163	100,0

Nota: los beneficiarios de planes de empleo se consideran como ocupados asalariados
Fuente: elaboración propia en base a EPH-INDEC (IV trimestre de 2003)

Por fuera del cuadro: título, notas y fuentes.

Veamos qué contiene esta tabla y qué información nos brinda. Primero nos ocuparemos de la información que aparece por fuera –antes y después– del cuadro.

- En primer lugar, el cuadro o tabla lleva un número. Es usual que así sea: los cuadros van –generalmente– insertos en un texto que alude al contenido de los mismos y ello requiere que se los identifique con precisión: "*Como lo muestra el cuadro 3.1.1*" –dirá el texto. Los cuadros pueden numerarse a partir del 1, en forma sucesiva y en todo el documento. Pero si éste es extenso y está dividido en partes o capítulos, conviene reiniciar la numeración al interior de cada uno de estos capítulos, tal como se hace en este ejemplo: así, el primer cuadro del apartado 3.1 se numera 3.1.1, en tanto que el siguiente llevaría la numeración 3.1.2. La razón para hacerlo así es que, en caso de insertar un nuevo cuadro que no había sido previsto, evitaremos tener que "correr" la numeración de todos los que le suceden en el texto.
- En segundo término, la tabla lleva un título. El título menciona dos elementos que ya vimos al ocuparnos de la estructura del dato: las unidades de cuenta del cuadro, que en general coinciden con las unidades de análisis (sabemos que se trata de la población ocupada de diez años y más que habita en las zonas urbanas, por lo que cada uno de estos ocupados es una unidad de análisis) y la variable que clasifica a estas unidades (que es la categoría ocupacional).
- En tercer término, al pie del cuadro aparece información adicional, consignada en una tipografía más reducida. Primero tenemos una nota aclaratoria: ella nos informa que las personas que son beneficiarias de programas sociales de provisión de empleos aparecen contadas entre los ocupados y se las ha considerado como trabajadores en relación de

manual metodologia 22/5/08 15:36 Página 317

dependencia (obreros o empleados). Este es el lugar donde se suele incluir cualquier aclaración que parezca necesario agregar en una tabla.

- Inmediatamente después –y esto es lo último que suele consignarse al pie de un cuadro– aparece la fuente de la información. Se nos indica que los datos fueron elaborados por el autor a partir de la base de datos de la Encuesta Permanente de Hogares (EPH) que realiza el INDEC. Y también sabemos que provienen de la onda correspondiente al cuarto trimestre de 2003. En este caso, el autor ha hecho algo con la información secundaria de la EPH: seguramente ha procesado la base de datos para obtener estas cifras. Si, en cambio, se hubiera limitado a copiar información difundida por el INDEC, por ejemplo en forma digital, la indicación de la fuente se limitaría a consignar: EPH-INDEC. Pero si la tabla fuera tomada de la publicación de otro autor –o inclusive de una publicación del INDEC o de otro organismo– la fuente debiera hacer justicia a la obra de la cual se la extrajo:

Fuente: Beccaria y López, *Sin Trabajo*, UNICEF / Losada, Buenos Aires, 1997.
Fuente: "Sintesis. Situación y evolución social", INDEC, Nº 1, 1992.

También podría ocurrir que el cuadro hubiera sido tomado del texto de otro autor, quien, a su vez, lo tomó de una tercera fuente. En este caso se indicaría así:

Fuente: CEPAL–PNUD, reproducido por Bustelo, Eduardo, "La producción del Estado de malestar. Ajuste y política social en América Latina" en Minujín, A. et. al. (1995). *Cuesta abajo. Los nuevos pobres: efectos de la crisis en la sociedad Argentina.* Buenos Aires: UNICEF/ Losada.

Por dentro del cuadro: frecuencias absolutas y relativas

En segunda instancia, veamos lo sustancial: lo que está dentro del cuadro.

- La primera columna enuncia las categorías de la variable. Cada una de las unidades o casos aparece clasificado en una de estas categorías.
- La segunda columna contiene las cantidades de casos o unidades que caen en cada categoría. Se las denomina frecuencias absolutas. Y la suma de las mismas equivale al total de casos[2]. Qué podemos decir acerca de la distribución de los casos en las distintas categorías: evidentemente, la más numerosa es la de los asalariados, que cuenta con 7,5 millones, en tanto que son muy pocos los trabajadores no remunerados: apenas 163 mil. Pero la interpretación de cifras grandes resulta un tanto trabajosa y oscura. Por eso, conviene hacer con ellas algo muy sencillo, cuyo resultado aparece en la tercera columna.

manual metodologia 22/5/08 15:36 Página 318

- La tercera columna contiene lo que se llama frecuencias relativas o porcentuales. Para hacer esta conversión a porcentajes, el total de casos (alrededor de 10,2 millones) se equipara a 100. Y la frecuencia de cada categoría se divide por el total y se multiplica por 100:

357.158 / 10.200.163 * 100 = **3,5**
2.155.463 / 10.200.163 * 100 = **21,1**
7.524.365 / 10.200.163 * 100 = **73,8**
163.177 / 10.200.163 * 100 = **1,6**

En realidad, cuando se emplean procesadores estadísticos[3] no hace falta hacer estas cuentas, porque las tablas ya aparecen con esta columna de frecuencias relativas. Si ahora prestamos alguna atención a estas cifras, veremos que el análisis de la distribución resulta bastante más claro: apreciaremos fácilmente que casi tres cuartas partes de los ocupados son asalariados, en tanto que alrededor de un quinto trabaja por propia cuenta y los empleadores son apenas 3,5 de cada 100. Es decir, disponemos de una descripción acerca de cómo se relaciona la gente con el mercado de trabajo. Las cifras expresadas en porcentajes son, en ese sentido, mucho más clara e inmediatamente interpretables. Tal como lo ha señalado Zeisel (1962), permiten reducir todos los números a una escala fácilmente manejable –ya que en general todos los porcentajes son menores a 100– y transforman uno de los números –el total– en la cifra 100.

En muchas ocasiones, se opta por suprimir la columna que contiene las frecuencias absolutas, mostrando sólo los porcentajes, con lo que la tabla anterior quedaría así:

Tabla 3.1.2: Población ocupada de 10 años y más según categoría ocupacional (en %). Total de aglomerados urbanos.

Categoría ocupacional	%
Patrón o empleador	3,5
Trabajador por cuenta propia	21,1
Obrero o empleado	73,8
Trabajador sin remuneración	1,6
Total	100,0 (10.200.163)

Nota: los beneficiarios de planes de empleo se consideran como ocupados asalariados
Fuente: elaboración propia en base a EPH-INDEC (IV trimestre de 2003)

manual metodologia 22/5/08 15:36 Página 319

Adviértase que, en este caso, se agrega en el título de la tabla y entre paréntesis la aclaración de que se trata de cifras expresadas en porcentajes. Asimismo, en la última fila y debajo del 100% figura, también entre paréntesis, el total sobre el que se obtuvieron tales porcentajes. Esto último es una precaución que no siempre se toma, pero que es conveniente y aconsejable. Alguien podría necesitar estimar la cantidad absoluta de trabajadores en alguna categoría ocupacional o en todas: a partir del total es muy fácil reconstruir esas frecuencias absolutas mediante el sencillo expediente de multiplicar dicho total por el porcentaje correspondiente y dividir por 100. Por ejemplo, para obtener la cantidad de cuentapropistas:

21,1 * 10.200.163 / 100 = 2.155.463

Pero además, cuando tenemos muestras pequeñas –como suele ocurrir en las encuestas de opinión– es importante saber cuántos casos hay en cada celda de la tabla. Ello se debe a que los porcentajes muy reducidos, correspondientes a celdillas con frecuencias muy bajas, pueden estar sujetos a errores de muestreo excesivos[4].

Será oportuno señalar, en este punto, una restricción referida al trabajo con porcentajes. Aunque no existe un acuerdo total respecto de cuál es el número mínimo de casos con el que se ha de contar, se suele señalar que no deben obtenerse porcentajes sobre cantidades muy pequeñas. Blalock (1986) aconseja no menos de 50 casos como base para los porcentajes. En tanto que Galtung ha sugerido un mínimo de 20 casos, basándose en la regla práctica de que las cifras porcentuales no estén demasiado sujetas a fluctuaciones ante cualquier pequeño cambio en las cifras absolutas: "…puede argumentarse que si el cambio de una unidad provoca una diferencia mayor que 5% el análisis llega a ser muy vulnerable, y esto conduce a una base mínima de 20" (Galtung, 1978: 64).

Efectivamente, podría ocurrir que tuviéramos los siguientes datos, obtenidos sobre muy pocos casos:

Cuadro 3.1.3: Argentina: jurisdicciones con bajo nivel de desarrollo humano por pertenencia regional

Jurisdicciones	Región
Corrientes	NEA
Chaco	NEA
Formosa	NEA
La Rioja	NOA
Santiago del Estero	NOA

manual metodologia 22/5/08 15:36 Página 320

Jujuy	NOA
Misiones	NEA
San Juan	Cuyo

Fuente: PNUD

En este cuadro tendría poco sentido decir que, entre las jurisdicciones de la Argentina con bajo desarrollo humano el 50% pertenecen al NEA, el 37,5% al NOA y el 12,5% a la región Cuyo. Con una distribución de ocho casos, es más razonable y claro señalar que, de ocho, cuatro son provincias de la región Nordeste, tres del Noroeste y una de Cuyo.

3.2. El caso de las variables ordinales: frecuencias acumuladas

La tabla Nº1 mostraba la distribución de una variable cuyo nivel de medición era nominal. En este caso, todo el contenido posible aparece en las tres columnas que ya hemos visto. Pero veremos ahora el caso de una variable ordinal, que nos permitirá hacer algo más:

Tabla 3.2.1: Población de 14 años y más según máximo nivel educativo alcanzado (en %). Total de aglomerados urbanos.

	%	% acumulado (A)	% acumulado (D)
Hasta primaria incompleta	10,8	10,8	100,0
Hasta secundaria incompleta	49,2	60,0	89,2
Hasta superior incompleta	29,8	89,8	40,0
Superior completa	10,2	100,0	10,2
Total	100,0		

Fuente: elaboración propia en base a EPH-INDEC (Octubre de 2002)

En la tabla 2 de nuestro ejemplo, tenemos la distribución de frecuencias de una variable que, además de clasificar, permite ordenar los casos. Hemos omitido aquí la columna de frecuencias absolutas, mostrando directamente los porcentajes. Veamos: casi 11% de la población urbana de 14 y más años no logró terminar la educación básica: algunos de ellos, eventualmente, siquiera la habrán comenzado. Pero casi la mitad logró, al menos, terminar ese nivel y tal vez cursar algunos años de la educación media, aunque sin culminarla. Y menos de un tercio terminó el ciclo secundario, incluyendo a los que han cursado algunos años de educación superior (terciaria o universitaria), sin llegar a completarla. Esto último es un privilegio limitado a uno de cada diez (que es lo mismo que decir diez de cada cien).

manual metodologia 22/5/08 15:36 Página 321

Pero además, tenemos una columna adicional que sólo es útil cuando las variables son, al menos, ordinales: la columna de frecuencias acumuladas. Acumular frecuencias significa adicionar, al porcentaje contenido en cada celda, el contenido en la celda anterior. Por ejemplo: 49,2 + 10,8 = 60,0.

En realidad, al acumular frecuencias puede optarse por hacerlo en orden ascendente o descendente. Veamos los efectos: la columna de frecuencias acumuladas ascendentes nos dice que seis de cada diez (es decir, 60%) personas de 14 y más años no llegaron a completar el nivel medio. Y que casi 90% no poseen un diploma de la educación superior. Al revés, si acumulamos de forma descendente, diremos que sólo 40% alcanzaron a completar la educación media. Y que casi 90% completaron, al menos, la educación básica.

Cuantas más categorías posea la variable cuya distribución se muestra, tanto más útil será emplear las frecuencias acumuladas en el análisis. Pero es evidente que ellas no tendrían sentido alguno si se tratara de un mero criterio clasificatorio: es razonable decir que en la categoría "hasta superior incompleta" tenemos acumulado el 90% de la población; pero no lo sería afirmar que hasta la categoría "cuentapropista" tenemos acumulada la cuarta parte de los ocupados, porque el orden de las categorías es, en este caso, arbitrario y no admite tal acumulación.

3.3. Proporciones y razones: haciendo "hablar" a los números

De haber omitido la multiplicación por 100 tendríamos proporciones del total de casos en lugar de porcentajes. En este caso el total sería representado por la unidad y la frecuencia relativa de cada categoría sería una proporción de ese total, con lo que la penúltima columna del cuadro hubiera contenido esas proporciones:

Tabla 3.3.1: Población de 10 años y más según condición de actividad. Total de aglomerados urbanos.

Condición de actividad	Total	Proporción	%
Ocupado	9.113.667	0,47	47,1
Desocupado	1.556.716	0,08	8,1
Inactivo	8.664.919	0,45	44,8
Total	19.335.302	1,00	100,0

Nota: los beneficiarios de planes de empleo que desempeñan contraprestación laboral se consideran como ocupados

Fuente: elaboración propia en base a EPH-INDEC (semestre I de 2004)

manual metodologia 22/5/08 15:36 Página 322

Como se deduce de lo anterior, los porcentajes no son sino proporciones traducidas a otra escala. Por lo regular, sin embargo, suelen usarse los porcentajes con preferencia a las proporciones por los motivos de simplicidad e interpretabilidad ya apuntados. Y sólo se muestran estos últimos porque sería inútil incluir ambas cosas: hemos sombreado deliberadamente la penúltima columna, para indicar que, si bien es un paso previo para llegar al porcentaje, no se la muestra en el cuadro.

En algunas ocasiones, si el total sobre el que se calcula la proporción es muy grande en relación con la categoría de casos que queremos destacar, esta traducción suele hacerse multiplicando por mil. Esto pasa con algunas tasas que se emplean asiduamente en demografía. Tal es el caso de la tasa de mortalidad infantil, un indicador que se construye dividiendo la cantidad de niños que fallecen antes de haber cumplido un años de edad en un período dado (normalmente un año), por la cantidad de nacidos vivos en ese mismo lapso. Ocurre que, cada año nacen muchos niños y, por fortuna, los que mueren antes del año son relativamente pocos. Por ejemplo, durante 2003, en Argentina nacieron 697.952 niños y en el mismo lapso murieron 11.494 de ellos. Si quisiéramos indicar la proporción de fallecidos sobre los nacidos, este cociente arrojaría una cifra muy pequeña: 0,016. Aun si la multiplicáramos por cien, deberíamos decir que falleció el 1,6% de los niños. En estos casos, cuando el numerador del cociente es muy grande en relación con el denominador, se acostumbra multiplicar por mil: afirmaremos, pues, que la tasa de mortalidad infantil alcanza a 16,5 por mil (16,5%o). Y en otros casos, se llega a multiplicar por diez mil o cien mil. Estas tasas se tornan, entonces, más fácilmente inteligibles.

Además de examinar las distribuciones de frecuencias y apreciar qué categorías concentran una mayor proporción de casos, es posible emplear algunos sencillos recursos adicionales para extraer más información de los números: se trata de "hacerlos hablar", revelando lo que no nos dicen en una primera mirada.

Volvamos a la tabla 1, que distribuía a la población ocupada por categoría ocupacional: es posible – y de gran utilidad – calcular unos sencillos cocientes entre las frecuencias de las diferentes categorías. Estas frecuencias se denominan *razones.* Por ejemplo, podríamos hacerlo dividiendo la cantidad de asalariados sobre la cantidad de empleadores. Ello nos daría una idea acerca del tamaño medio de los establecimientos que ocupan personal:

Obrero o empleado / patrón o empleador = 7.524.365 / 357.158 = 21,1

manual metodologia 22/5/08 15:36 Página 323

También podemos hacerlo con las frecuencias relativas, puesto que el resultado será el mismo:

73,8 / 3,5 = 21,1

Podemos decir que hay algo más de veinte asalariados por cada empleador. O, lo que es lo mismo, que cada empleador ocupa, en promedio, una veintena de empleados u obreros.

También podemos invertir estos cocientes. Si dividimos a los empleadores por los asalariados tendremos:

3,5 / 73,8 = 0,05

Diríamos que por cada salariado, tenemos 0,05 empleadores. Esto resulta muy poco claro (porque las personas no son divisibles). Pero se clarifica bastante afirmando que por cada 100 asalariados hay 5 empleadores, que es lo mismo pero suena mucho mejor.

También podemos querer saber cuál es la relación entre la gente que trabaja por un salario y la que, en cambio, lo hace por su cuenta (contratando personal o sin hacerlo). Si esta fuera nuestra inquietud, bien podríamos sumar las frecuencias de empleadores y trabajadores por cuenta propia (21,1 + 3,5 : 24,6) y dividir este resultado por la proporción de asalariados:

24,6 / 73,8 = 0,33

En este caso, diremos que tenemos 33 personas que trabajan en forma autónoma por cada 100 que lo hacen en relación de dependencia. Y de haber obtenido el cociente al revés:

73,8 / 24,6 = 3

Podemos afirmar, sin temor a equivocarnos, que hay tres personas que trabajan en relación de dependencia por cada una que lo hace por cuenta propia o como patrón.

Estas razones son de gran utilidad. Constituyen indicadores sintéticos de la situación del mercado laboral (en este caso) en un país y en un momento dado. Y podrían servir muy adecuadamente para hacer una comparación entre diferentes países (en un diseño de corte transversal) o bien a lo largo del tiempo (en un diseño longitudinal). Por ejemplo, que sea alta la última razón calculada (entre trabajadores asalariados y los que no lo son) suele ser un indicador de desarrollo del mercado de trabajo

Ya se trata de proporciones, porcentajes o razones, tendremos cifras decimales. Cabe, pues, preguntarse con cuántos decimales deben calcularse los porcentajes y proporciones. No hay, acerca de esto, una respuesta absoluta pero, en términos generales, puede admitirse que en el caso de los porcentajes basta con escribir un solo decimal, en tanto que en las proporciones y las razones emplearemos dos[5].

manual metodologia 22/5/08 15:36 Página 324

Sintetizando lo dicho, las proporciones son cocientes entre la frecuencia de una categoría y el total de casos, en tanto que las razones se calculan entre frecuencias de diferentes categorías de una variable. Muchos indicadores de uso corriente en temas tales como educación, mercado de trabajo o demografía (a más de muchos otros) se construyen mediante razones o proporciones. Veamos algunos ejemplos:

- Tasa de natalidad: es un cociente entre el total de personas nacidas en el curso de un año dado en cierta jurisdicción (país, provincia, etc.) y el total de personas que habitaban en dicha jurisdicción en ese mismo año[6]. Es, por lo tanto, una *proporción* calculada entre una parte de los habitantes –los nacidos en ese año– y la totalidad de ellos.
- Tasa de mortalidad: es un cociente entre el total de personas fallecidas en el curso de un año dado en cierta jurisdicción (país, provincia, etc.) y el total de personas que habitaban en dicha jurisdicción en ese mismo año[7]. Es, por lo tanto, una *proporción* calculada entre una parte de los habitantes –los que murieron en ese año– y la totalidad de ellos.
- Tasa de mortalidad infantil: es un cociente entre el total de niños fallecidos antes de cumplir un año de edad, en el curso de un año dado y en cierta jurisdicción (provincia, país, etc.) y el total de los niños nacidos en esa misma jurisdicción y en el mismo año. Es, por lo tanto, una *razón* calculada entre una parte de los niños (los que fallecieron en cierto lapso sin haber cumplido un año) y otra parte (los nacidos en el mismo período)[8].
- Razón de masculinidad: es un cociente entre el total de varones y el total de mujeres que habitan en una jurisdicción (provincia, país, etc.) en un cierto momento. Es, por lo tanto, una *razón* calculada entre una parte de los habitantes (los de sexo masculino) y otra parte de ellos (los de sexo femenino).
- Tasa de asistencia escolar: es un cociente entre el total de los niños de cierta edad que están asistiendo a la escuela en una jurisdicción (provincia, país, etc.) en un cierto momento, y el total de niños de esa misma edad que habitan en esa jurisdicción. Es, pues, una *proporció*n entre una parte de los niños (los que van a la escuela) y el total de ellos.
- Tasa de abandono escolar: es un cociente entre los niños que abandonaron la escuela en un año dado, en cierta jurisdicción (provincia, país, etc.) y el total de los que se encontraban asistiendo a la escuela en ese mismo año y jurisdicción. Se trata de una *proporción* entre una parte de los niños inscriptos en la escuela (los que la abandonaron) y la totalidad de ellos.
- Tasa de actividad: es un cociente entre las personas que están trabajando más las que están buscando trabajo sobre el total de habitantes de una jurisdicción (provincia, país,

manual metodologia 22/5/08 15:36 Página 325

etc.) en un momento dado. Se trata, pues, de una *proporción* calculada entre una parte (los que trabajan o procuran hacerlo) y el total de la población.

- Tasa de empleo: es un cociente entre las personas que están trabajando y el total de habitantes de una jurisdicción (provincia, país, etc.) en un momento dado. Se trata, pues, de una *proporción* calculada entre una parte (los que trabajan) y el total de la población.

- Tasa de desempleo: es un cociente entre las personas que están buscando trabajo y las que trabajan o buscan trabajo, en una jurisdicción (provincia, país, etc.) en un momento dado. Consiste en una *razón* calculada entre una parte de los que están en disposición de trabajar (los que intentan trabajar sin conseguirlo) y la totalidad de ellos (los que están intentando más los que ya lo lograron).

3.4. Respuestas múltiples: ¿cómo porcentualizar?

Hasta aquí, ha sido el caso de que las categorías de las variables son mutuamente excluyentes. Es decir, un caso determinado no puede caer al mismo tiempo en dos de estos casilleros. Por lo común, es así.

Sin embargo, existen algunas excepciones a esta regla, como a muchas otras... A veces, en los cuestionarios se incluyen preguntas que admiten respuestas múltiples (así lo vimos en el capítulo 11): en estos casos, cada encuestado puede ser clasificado en diferentes categorías de una misma variable (puesto que cada pregunta es una variable). Por ejemplo, a una pregunta como la siguiente:

¿Cuáles de las actividades que se enumeran a continuación suele realizar en su tiempo libre?:

1. Leer
2. Ver televisión
3. Escuchar radio
4. Escuchar música
5. Salir con amigos
6. Hacer deporte
7. Otras

...cada interrogado podría responder marcando una, dos o muchas de las alternativas propuestas: por ejemplo, tres. El resultado sería que habría más respuestas que respondentes ¿Cómo hemos de tratar, desde el punto de vista estadístico, estos casos? La especificidad consiste en que en las tablas de frecuencias de las preguntas de respuestas múltiples, puede optarse por dos alternativas: o bien se obtienen porcentajes sobre el total de respues-

manual metodologia 22/5/08 15:36 Página 326

tas y no de encuestados, o bien se lo hace sobre el total de encuestados, en cuyo caso los porcentajes suman más de 100. Veamos :

Tabla 3.4.1. Uso del tiempo libre

Uso del tiempo libre	Respuestas		Encuestados	
	Total	%	Total	%
Leer	65	5,4	65	16,3
Ver televisión	392	32,7	392	98,0
Escuchar radio	205	17,1	205	51,3
Escuchar música	116	9,7	116	29,0
Salir con amigos	206	17,2	206	51,5
Hacer deportes	112	9,3	112	28,0
Otras	104	8,7	104	26,0
Total	1200	100	400	300,0

Fuente: datos ficticios.

Cada una de las dos columnas nos dice algo distinto. La primera nos muestra la distribución de las respuestas: lo que más frecuentemente hace la gente con su tiempo libre es ver televisión. Y le siguen en importancia escuchar la radio o salir con amigos, actividades que se disputan el segundo lugar en las preferencias. En cambio, es relativamente poco frecuente dedicarse a la lectura, que obtuvo poco más de 5% de las menciones. Se diría que se trata de una distribución de preferencias.

La segunda, en cambio, nos proporciona información acerca de cuánta gente se dedica a una u otra actividad en sus ratos libres. Sobre las 400 personas interrogadas, casi todas (98%) ven televisión. La mitad acostumbra salir con amigos y una proporción similar (aunque, por cierto, no sabemos si se trata de los mismos) escucha radio. Sólo 16% suele ocupar su tiempo leyendo. Si sumamos estos porcentajes, la suma excede de 100%, sencillamente porque las personas no hacen una sola cosa: las mismas que ven televisión en algún momento, practican un deporte en otro. En este caso, como se ha pedido a cada encuestado que señalara hasta tres cosas, suman exactamente 300.

4. Variables cuantitativas: construcción de intervalos de clase. Reducción de categorías

Pasaremos ahora a considerar el tratamiento estadístico básico de las variables numéri-

manual metodologia 22/5/08 15:36 Página 327

cas o cuantitativas. Ellas habilitan el uso de instrumentos estadísticos más poderosos, porque además de trabajar con las cantidades de casos –las frecuencias– podemos hacerlo con los valores que ellos asumen en las variables, que también se expresan en números.

4.1. Construcción de intervalos de clase

El tratamiento más elemental apunta a examinar y describir las distribuciones de un modo similar al que empleamos con las variables categóricas. Supongamos que se trata de analizar la distribución de un conjunto de los trabajadores de una empresa según su nivel de ingresos o según los años de antigüedad en el puesto de trabajo. En ambos casos, seguramente contaremos con el dato puntual: el sueldo que percibe cada uno y el tiempo que lleva en la empresa. Sin embargo, a los efectos de analizar esta información nos serviría de muy poco mirar un listado que contuviera –acaso– dos centenares de filas. Peor aun si tuviésemos que considerar la distribución por edades de la población de un país con varios millones de habitantes. Ni siquiera sería una adecuada solución agrupar a todos los niños de menos de un año, a los de uno, a los de dos, etc. Lo más razonable será construir tramos o categorías con estas variables cuantitativas, que nos permitan agrupar los casos.

Por ejemplo, en el caso de la distribución de edades, los censos suelen agrupar a la población por grupos o quinquenales, es decir, formando tramos de cinco años:

Tabla 4.1.1. Población total por grupos quinquenales de edad.

Tramos de edad	Total	%
De 0 a 4 años	3.349.278	9,2
De 5 a 9 años	3.471.217	9,6
De 10 a 14 años	3.427.200	9,5
De 15 a 19 años	3.188.304	8,8
Etc.	...	

Fuente: INDEC – CNPV 2001

Estos tramos, que reciben también el nombre de *intervalos de clase* o *categorías*[10] son, en este caso, de igual tamaño y se ha cuidado que sean exhaustivos (toda persona puede ser ubicada en uno de ellos) y excluyentes (a cada persona le corresponde sólo uno de estos tramos). Puesto que se trata de los años cumplidos, quien cumpla cinco años al día siguiente de ser interrogado, caerá en el primer tramo. En tanto que quien los haya cumplido ya, será

manual metodologia 22/5/08 15:36 Página 328

clasificado en el segundo. Se observará, asimismo, que todos los intervalos tienen igual amplitud o tamaño, pues abarcan cinco años.

En cambio, en el caso de los ingresos, tendría poco sentido hacer tramos o intervalos de tan escasa amplitud: ni siquiera sería razonable construirlos todos de igual tamaño. No hay un criterio único, pero un agrupamiento posible sería:

Tabla 4.1.2. Población ocupada según tramos de ingresos de la ocupación principal. Total de aglomerados urbanos.

Tramos de ingresos	Total	%	% acumulado
Hasta 150	1.015.408	18,2	18,2
De 151 a 300	874.280	15,6	33,8
De 301 a 500	1.292.900	23,1	56,9
De 501 a 800	1.299.023	23,2	80,1
De 801 a 1200	634.727	11,4	91,5
De 1201 a 1700	241.303	4,3	95,8
De 1701 a 2300	120.476	2,2	98
De 2301 a 3000	59.189	1,1	99,1
Más de 3000	51.033	0,9	100
Total	5.588.339	100,0	

Fuente: elaboración propia en base a EPH-INDEC (semestre I de 2004)

Se observará que, en este caso, la amplitud de los tramos es creciente. Aunque no existe, al respecto, ninguna regla fija, hay motivos para hacerlo así. Una primera razón es que a medida que se progresa en la escala las frecuencias tienden a disminuir: por encima de los tres mil pesos sólo se ubica menos de 1% del total, de manera que establecer tramos por encima de esa cifra resultaría ocioso. Otro motivo estriba en que, a medida que aumenta el ingreso, las diferencias pequeñas pierden importancia: 150 pesos son una suma significativa en las escalas más bajas, ya que quien gana 300 duplica el ingreso del que sólo percibe 150. Pero para el que percibe un ingreso de 2300, 150 apenas implicarían una diferencia de 6,5%, casi desdeñable.

Otra peculiaridad es que –además de que difiere el tamaño de los intervalos– el último de ellos ha quedado abierto: puesto que hay pocos trabajadores con remuneraciones por encima de 3 mil pesos, no tiene sentido establecer más tramos hasta llegar al límite superior[11].

manual metodologia 22/5/08 15:36 Página 329

De paso, el examen de esta tabla nos sirve para aplicar algunos de los instrumentos de análisis que ya hemos visto antes: por ejemplo, el examen de las *frecuencias relativas* nos dice que entre 301 y 800 pesos se agrupa casi la mitad de los trabajadores.

En cuanto a las *frecuencias acumuladas*, ellas nos dicen que casi 60% de los ocupados obtienen un ingreso mensual que no supera los $500 por mes, en tanto que ocho de cada diez no sobrepasan los $800.

Otra herramienta útil pueden ser las *razones*: por cada ocupado que percibe una remuneración de 3 mil o más hay casi veinte que no sobrepasan los $150.

4.2. La reducción de categorías: cuando es preciso simplificar

Siempre conviene, inicialmente, disponer de los datos con el máximo desagregado, porque ello nos brinda información más detallada. Así, cuando en una encuesta se pregunta por la edad o los ingresos, es preferible obtener el dato puntual: los años cumplidos o el ingreso percibido en pesos. Aunque en este último caso, a veces, se opta por pedir a los encuestados que se ubiquen en un intervalo, debido a que la gente suele ser reticente para declarar sus ingresos puntuales, puesto que lo siente como una intrusión. Esto mismo ocurre cuando es presumible que las personas sólo cuenten con el dato aproximado. Por ejemplo, las encuestas sobre ocupación suelen preguntar por la cantidad de personas que trabajan en el establecimiento en el cual se desempeña el encuestado: es presumible que éste lo ignore, por lo que le será más fácil ubicar a la empresa donde trabaja en un intervalo. Por ejemplo:

Tabla 4.2.1. Ocupados en relación de dependencia según tamaño del establecimiento. Total urbano.

Tamaño del establecimiento	Total	%	% acumulado
Un ocupado	501.317	8,9	8,9
De dos a cinco ocupados	1.356.656	24,0	32,8
De seis a 15 ocupados	1.078.382	19,0	51,9
De 16 a 25 ocupados	502.889	8,9	60,7
De 26 a 50 ocupados	632.661	11,2	71,9
De 51 a 100 ocupados	591.425	10,4	82,4
De 101 a 500 ocupados	657.443	11,6	94,0
Más de 500 ocupados	341.146	6,0	100,0
Total	5.661.919	100,0	

Fuente: elaboración propia en base a EPH-INDEC (onda mayo de 2003)

manual metodologia 22/5/08 15:36 Página 330

Pero ya sea que tengamos el dato puntual o una clasificación bastante desagregada –con muchos intervalos o categorías– es posible que a los efectos del análisis no necesitamos tanto detalle, que hasta podría oscurecer la interpretación de los datos. Inclusive, si los casos no fueran muchos, una demasía de desagregado dejaría muy pocos en cada celda[12].

Un modo de sortear este problema consiste en la fusión de intervalos o categorías, realizando lo que Barton (1973) denomina "reducción de un espacio de propiedades por intermedio de la simplificación de las dimensiones". En el caso de las edades, podríamos adicionar de a dos los tramos, construyendo intervalos de diez años de amplitud, con lo que su número quedaría reducido a la mitad. En el de los ingresos, podríamos proceder de similar manera, haciendo que el primero llegara hasta 300, el segundo hasta 800, etc. Está claro que al hacerlo así, los nuevos intervalos tendrían unas frecuencias que resultarían de sumar las de los que fueron unificados.

Muchas veces, estas reducciones dimensionales responden a precisos propósitos del análisis. Supongamos que el objetivo fuera analizar la carga demográfica que tiene un país en un momento dado. Esto es, la relación entre las personas que están en edades activas –en condiciones potenciales de trabajar– y las que, en cambio, deben ser sostenidas porque aun no han alcanzado la edad necesaria (los niños) o porque ya la han sobrepasado (los ancianos). Si este fuera el caso, bastaría con reducir los tramos de edad a sólo tres:

Tabla 4.2.2. Población total por grupos de edad.

Tramos de edad	Total	%
De 0 a 14 años	10.247.695	28,3
De 15 a 64 años	22.424.815	61,8
De 65 y más años	3.587.620	9,9
Total	36.260.130	100,0

Fuente: INDEC - Censo Nacional de Población y Vivienda 2001

Una herramienta muy útil para medir la "carga demográfica" sería calcular una razón que relacionara a los "pasivos" (niños más ancianos) con los "potencialmente activos" (las personas en edad de trabajar, ubicadas en el tramo del medio). Esta razón se usa con frecuencia y se denomina *índice de dependencia potencial*:

IDP = (población de 0 a 14 + población de 65 y más)/población de 15 a 64

En el caso de Argentina, en 2001, esta relación arrojaba 0,62. En los países de Europa Continental este índice es elevado, debido a la alta proporción de población anciana.

manual metodologia 22/5/08 15:36 Página 331

En cambio, en muchas naciones de Africa o América Latina el índice se ve incrementado por la presencia de niños, debido a la elevada fecundidad.

Pero si estuviéramos haciendo un análisis sobre el sistema educativo, entonces haríamos un agrupamiento bien diferente:

Tabla 4.2.3. Población total por grupos de edad.

Tramos de edad	Total	%
De 0 a 4 años	3.349.278	9,2
De 5 años	714.495	2,0
De 6 a 14 años	6.183.922	17,1
De 15 a 17 años	1.921.972	5,3
De 18 a 24 años	4.465.671	12,3
De 25 años y más	19.624.792	54,1
Total	36.260.130	100,0

Fuente: INDEC - Censo Nacional de Población y Vivienda 2001

Los tramos o intervalos segundo, tercero y cuarto abarcan las edades cubiertas por la educación obligatoria en Argentina. El primero (cinco años) corresponde al ciclo preescolar o inicial. El segundo (seis a 14 años) corresponde a la Educación General Básica y el tercero (15 a 17 años) al Polimodal[13].

En el caso del tamaño de las empresas, también podríamos recurrir a una reducción basada en propósitos de análisis. Por ejemplo, podríamos querer separar a los ocupados según se desempeñen en microempresas, empresas medianas y empresas grandes. Un modo posible de hacerlo sería:

Tabla 4.2.4. Ocupados en relación de dependencia según tamaño del establecimiento. Total urbano.

Tamaño del establecimiento	Total	%	% acumulado
Hasta cinco ocupados	1.857.973	32,8	32,8
De seis a 100 ocupados	2.805.357	49,6	82,36
Más de 100 ocupados	998.589	17,6	100,0
Total	5.661.919	100,0	

Fuente: elaboración propia en base a EPH-INDEC (onda mayo de 2003)

manual metodologia 22/5/08 15:36 Página 332

5. El uso de las medidas estadísticas descriptivas (tendencia central, posición y dispersión)

La estadística descriptiva provee un conjunto de medidas muy simples, que no tienen otro propósito que resumir ciertas características centrales de un conjunto de datos u observaciones. Estos descriptivos no se predican de cada uno de los individuos que integran la población, sino del conjunto. Así, cuando calculamos el promedio de edad o de ingresos de un conjunto de individuos (mediante el sencillo expediente de sumar las edades –o los ingresos– de todos los componentes de la población y dividir luego por el total de quienes la integran) nos resulta una cifra que, probablemente, no corresponda a la edad o al ingreso de ninguno en particular: es un atributo del conjunto. El promedio de edad de un anciano de 80 años y un joven de 20 es igual a 50. En este caso, diríamos que ese promedio nos dice muy poco acerca de esa pequeña población constituida por sólo dos casos. Pero cuando las distribuciones son más numerosas, las medidas de este tipo suelen ser de mayor utilidad.

Entre estas medidas *de resumen* podemos establecer tres grandes conjuntos: las de tendencia central, las de orden o posición y las de dispersión.

5.1. Medidas de tendencia central

Las medidas de tendencia central más simples y utilizadas son la media aritmética, la mediana y el modo o moda. Comenzaremos por esta última, que es la más sencilla.

El modo

El modo es el valor de variable que reúne la mayor frecuencia, el que más veces se repite en una distribución. Así, en un curso de ingreso de la universidad, es probable que el modo de la variable edad sea 18 años, puesto que se supone que la mayor parte de los cursantes son personas que han terminado el año anterior el ciclo medio. Habrá, por supuesto, personas que tengan 19, 20 o más años (incluso, excepcionalmente, podría haber alguien que todavía no hubiera cumplido 18, pero serán menos). Igualmente, si distribuimos a los perceptores de jubilaciones por el importe del beneficio previsional, observaremos que el valor que más se repite es $ 300, que corresponde a la jubilación mínima. Aunque más de la mitad ganan más que eso, se trata de cifras dispersas, que se repiten menos. Entonces, $300 es el modo o valor modal. En la notación estadística el modo se indica con Mo.

Obviamente, hay distribuciones que carecen de modo (podría ser que ningún valor se repitiera más que los otros). Y puede haber distribuciones que tengan más de un modo –a las que denominaremos bimodales si tienen dos o polimodales si tienen más–. En rigor,

manual metodologia 22/5/08 15:36 Página 333

aunque estrictamente hubiera un solo valor modal, si existiera dos o más valores de variable donde tienden a agruparse los casos de un modo significativo, se alude igualmente a distribuciones bi o polimodales.

También puede suceder que haya un único valor modal (o más de uno) pero que resulte poco significativo en términos de describir la distribución: si imaginamos la distribución de ingresos de una muestra de trabajadores informales, podría ocurrir que estos fueran todos bajos pero muy variables: entre 200 y 400 pesos mensuales. Podría suceder, sin embargo, que hubiera unos pocos de ellos que ganaran 800 pesos (dos, por ejemplo): esta cifra sería el modo, pero nos diría bien poco acerca de las características centrales de esta distribución de ingresos.

Obviamente, el modo es inestable en las distribuciones pequeñas (de pocos casos). Porque aumenta la probabilidad de que todos los valores sean diferentes entre sí, de manera que con que uno solo de ellos se repita, quedaría determinado el modo. Si por casualidad hubiera sido seleccionado en la muestra otro sujeto, tendríamos también otro valor modal, eventualmente muy diferente, sin que las características generales de la distribución hubieran cambiado demasiado.

Con la finalidad de prevenir una confusión recurrente, convendrá enfatizar una vez más que el modo es el valor de la variable que presenta la frecuencia más elevada y no dicha frecuencia.

La mediana

La mediana es, en cambio, el valor de la variable que divide la distribución en dos mitades, dejando por debajo y por encima igual cantidad de frecuencias. De manera tal que el 50% de los casos asumen un valor inferior a la mediana, mientras que el 50% restante tiene un valor de variable superior. Habitualmente en la notación estadística se la indica con Md.

Así, la mediana es fácil de determinar en una distribución con cantidad impar de casos, porque vendrá dada por el valor del caso intermedio. Si tenemos la distribución de las tasas de desempleo de cinco ciudades de la Argentina relevadas por la Encuesta Permanente de Hogares (EPH), bastará ordenarlos en orden creciente o decreciente y resultará claro que la mediana será la tasa de la Ciudad de Buenos Aires (13,5%), que deja dos ciudades por encima y dos con valores inferiores. No hay modo en esta distribución, puesto que ningún puntaje se repite.

manual metodologia 22/5/08 15:36 Página 334

Tabla 5.1.1. Jurisdicciones seleccionadas: tasas de desempleo

Aglomerado	Tasa de desempleo (%)
Corrientes	19,7
Gran Resistencia	17,4
Ciudad de Buenos Aires	13,5
Gran Mendoza	11,5
Santa Cruz	3,0

Fuente: EPH-INDEC (Onda octubre de 2002)

Pero si tuviéramos un número par de observaciones (por ejemplo, si eliminamos Santa Cruz), ya no tendríamos ese caso intermedio. En tal situación, el valor de la mediana se obtiene promediando los dos valores centrales, que en este caso serían los correspondientes a Gran Resistencia y la Ciudad de Buenos Aires:
Md = (17,4 + 13,5)/2 = 15,5

Y observaríamos que este valor ya no corresponde a ninguna de las observaciones en particular: no hay ninguna ciudad con una tasa de desempleo de 15,5%. Es, ni más ni menos, una medida descriptiva que caracteriza al conjunto: una medida que resume la *tendencia central*.

Obviamente, cuando tenemos frecuencias agrupadas (por ejemplo, la población distribuida por edades agrupadas en intervalos, cada uno de ellos con su respectiva frecuencia) ya no es tan sencillo obtener la mediana. En tales casos, ella se estima mediante una fórmula de la que es innecesario ocuparnos aquí[14].

La media aritmética

La media aritmética no es otra cosa que el promedio, que todos estamos habituados a calcular en un sinnúmero de oportunidades, en nuestra vida cotidiana. Para calcularla basta, pues, sumar las puntuaciones que asumen en la variable cada uno de los casos y dividir esta suma por la cantidad total de casos. Si lo referimos a la distribución de las tasas de desempleo de las cinco ciudades de la tabla anterior, la media aritmética (que se representa usualmente con el símbolo $\overline{X}$[15]) será:

$$\overline{X} = (19,7+17,4+13,5+11,5+3)/5 = 13$$

Este tampoco es, pues, un valor que corresponda a ningún caso en especial, sino un atributo del conjunto: nuevamente describe la *tendencia central* de la distribución. Como en el

manual metodologia 22/5/08 15:36 Página 335

caso de la mediana, en caso de tener valores agrupados en intervalos, deberíamos usar una fórmula apenas más complicada, que tampoco interesa tratar aquí.

¿Qué pasará si suprimimos uno de los valores de la distribución (Santa Cruz, al igual que en la oportunidad anterior)?: la media será ahora 15,5. Ha aumentado, pues, en dos puntos y medio. Esta variación es mayor que la que experimentó la mediana al efectuar la misma supresión ¿Por qué?: porque al ser un promedio, en cuyo cálculo intervienen la totalidad de los valores, resulta sensible a los puntajes extremos (muy altos o muy bajos). Cuanto más extremos sean, más influyen en la media: ya sea "tironeándola" para arriba o para abajo. En las situaciones extremas, si unos pocos casos asumen en una variable valores muy divergentes del resto (mucho más altos o mucho más bajos), la media puede resultar una medida distorsiva, que describa engañosamente la distribución[16].

Le mediana resulta, en estos casos, una medida más equilibrada, porque en su cálculo no intervienen los valores más extremos. Las distribuciones de salarios –como las de ingresos en general– tienen usualmente esta característica: hay una pequeña proporción de la fuerza de trabajo altamente remunerada, mientras que la mayoría tiende a ganar poco. El resultado de las distribuciones así sesgadas es que la mediana asume, generalmente, un valor más bajo que el de la media y más representativo. Por ejemplo, según los datos emergentes de la EPH que lleva a cabo el INDEC, en el Gran Buenos Aires la media de ingresos laborales era, hacia fines del año 2002, de $ 612, en tanto que la mediana se situaba en $ 400: la mitad de los trabajadores obtenía una remuneración que no excedía de esa cifra. Y el modo alcanzaba –tan solo– a $ 150[17]. Pero además, un análisis más desagregado permitía determinar que menos del 30% de los ocupados gozaba de una remuneración igual o superior al promedio.

Un aspecto en el que debe repararse es que, si bien –tal como hemos visto– las medidas de tendencia central pueden no coincidir con los valores que asumen concretamente los casos que integran la distribución, están siempre expresadas en valores de variable: años si se trata de las edades, puntos porcentuales de la PEA si hablamos de las tasas de desempleo o pesos si estamos considerando los ingresos.

5.2. Medidas de orden o posición

El segundo conjunto de medidas de resumen (que intersecta con el anterior) es el correspondiente a las llamadas medidas de posición u orden. Nuestra ya conocida mediana es la más simple y básica. Ella fija –a la manera del fiel de la balanza– el punto intermedio: aquel valor –ya fue dicho– por debajo y por encima del cual tenemos la mitad de los casos.

A similitud de ella, tenemos los cuartiles: estos son tres y dividen la distribución en cuatro segmentos, cada uno de los cuales reúne al 25% de los casos. Obviamente, el valor del

manual metodologia 22/5/08 15:36 Página 336

segundo cuartil coincide siempre con el valor de la mediana. Aunque, en rigor, se denominan cuartiles estos puntajes de corte, el término se hace extensivo a los cuatro segmentos que quedan así delimitados:

25%	25%	25%	25%
	C1	C2 = Md	C3

Por debajo del cuartil 1 tendremos, entonces, el 25% inferior de la distribución: la cuarta parte peor remunerada del conjunto de los trabajadores, si se tratara de los ingresos laborales. Y por encima del cuartil 3, se ubicará el 25% mejor pago. Para seguir empleando el ejemplo de los ingresos laborales en el Gran Buenos Aires, en octubre de 2002 el valor del primer cuartil era $ 300, mientras que el del tercer cuartil era $900. Esto implica, también, que si el 25% peor remunerado ganaba por debajo de $300, el siguiente 25% obtenía un ingreso situado entre esa cifra y $400 (el valor de la mediana que, ya lo hemos visto, equivale al segundo cuartil). Y entre la mediana y el tercer cuartil tendríamos a los trabajadores que recibían remuneraciones situadas entre 400 y 900 pesos.

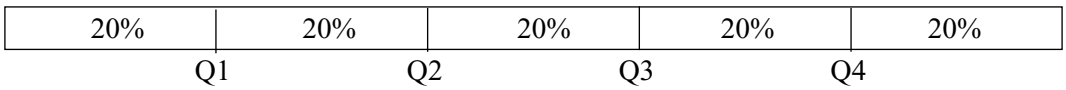
Por supuesto, que no todos los trabajadores que quedan en el segmento inferior (ni en cualquiera otro) ganaban una remuneración uniforme: lo único que sabemos es que no sobrepasaban los $300, pero esta afligente condición los afectaría de manera diferencial: ya vimos que muchos de ellos percibían $150 (el valor modal): seguramente habría quienes ganaban aún menos y quienes, en cambio, se situaban muy próximos al valor tope. Esto mismo pasará en cada uno de los otros segmentos. Si queremos indagar con mayor detalle, podríamos muy bien combinar estas medidas de posición con las de tendencia central ya vistas, calculando la media, la mediana y aún el modo al interior de cada uno de los segmentos, lo que nos daría una idea acerca de su conformación interna.

Por ejemplo, le media del 25% inferior (del primer cuartil, para usar el término en su sentido ampliado) era de $ 153 y la mediana y el modo coincidían en 150. En tanto, para el 25% superior (el cuarto cuartil, diríamos) estas medidas eran, respectivamente, $1800 la media, $1300 la mediana y $1000 el modo. Esto ya habilita una comparación que ofrece cierto interés con lo que vimos antes, respecto del segmento inferior: al interior de éste las tres medidas arrojaban prácticamente los mismos valores. Ello sugiere una gran homogeneidad en las remuneraciones de los trabajadores más pobres, cosa que no sucede entre los mejor posicionados: entre estos últimos, la media era considerablemente superior a la mediana y al modo: sin duda que influía, dentro de este grupo, un conjunto de personas que obtenían unos ingresos muy elevados, presionando hacia arriba sobre el promedio.

Con una lógica análoga a la de los cuartiles, los quintiles, que son cuatro, dividen la distribución en cinco segmentos, cada uno de los cuales contiene el 20% de los casos totales.

manual metodologia 22/5/08 15:36 Página 337

También aquí los segmentos toman, por extensión, el nombre de quintiles y son muy frecuentemente empleados para estratificar la población en función de sus ingresos (ya se trate de los laborales o los ingresos familiares totales). Obviamente, con los quintiles es posible llevar a cabo el mismo tipo de análisis que con los cuartiles, combinando muy productivamente las capacidades descriptivas de las medidas de posición y las de tendencia central.



Y es fácil ahora comprender que la distribución puede ser dividida en más fragmentos que incluyan igual proporción de casos: los nueve deciles la fragmentan en diez partes iguales, con el 10% de los casos cada una (donde el 5° decil coincide con la mediana y con el segundo cuartil). Y los 99 percentiles hacen lo mismo, pero estableciendo 100 subconjuntos con el 1% de los casos cada uno de ellos (el percentil 50 coincide, claro está, con la mediana, con el decil quinto y con el segundo cuartil). Menos usados, pero también eventualmente útiles, son los ventiles; son diecinueve y segmentan la distribución en veinte grupos con el 5% de los casos en cada uno de ellos.

5.3. Medidas de dispersión o heterogeneidad

El tercer subconjunto de medidas de resumen apunta a otro concepto. Se trata de captar la heterogeneidad o desigualdad entre las puntuaciones que asumen los casos en una variable: técnicamente, se denomina a esto *dispersión.* La idea es simple: la distribución de los ingresos podría ser muy homogénea (todos ganan cifras muy próximas entre sí) o muy heterogénea o desigual (algunos ganan mucho, otros muy poco, en tanto que otros perciben remuneraciones intermedias).

En la primera situación, todos tendrán puntuaciones muy semejantes a la media: por ejemplo, en una escuela de educación básica habrá poca heterogeneidad en las edades: la mayor parte de los alumnos tendrá entre seis y 14 años (aunque habrá algunos repitentes, que tengan algo más). Si la media se sitúa en 11 años –por ejemplo– casi nadie se apartará de ella en más de cuatro años (hacia arriba o hacia abajo). En cambio, si se trata de los concurrentes a una sala de cine, podrá haber desde niños o adolescentes hasta ancianos, de manera que sea cual fuere la edad promedio, ella surgirá de una mezcla variada y habrá muchos casos "despegados" hacia arriba y hacia abajo. Lo mismo ocurriría si relevamos los ingresos de los hogares en una villa de emergencia, en un barrio exclusivo de clases altas y en un barrio de configuración heterogénea. En el primer caso todos tendrán ingresos bajos

manual metodologia 22/5/08 15:36 Página 338

(y el promedio será bajo) y en el segundo –por el contrario– todos los tendrán altos (y también lo será el promedio). En cambio, en la tercera situación podremos hallar una heterogeneidad mucho mayor: desde familias que habiten en casas tomadas hasta otras que lo hagan en edificios de departamentos caros. De manera que la media aritmética se situaría, seguramente, en una cifra intermedia que resultará de promediar ingresos bajos, medios y altos.

La construcción de una medida de dispersión

¿Cómo mensurar estos grados de heterogeneidad a través de una única medida sintética? Hemos designado esta propiedad como dispersión: ¿dispersión en torno a qué? En los ejemplos precedentes, se advierte que cuando hablamos de dispersión, aludimos a la proximidad o alejamiento relativo con respecto al valor promedio. En consecuencia, un modo posible de construir una medida que resuma esta variabilidad sería calcular las diferencias entre las puntuaciones de cada caso (que se indican con x en la notación estadística) y la media aritmética y, eventualmente, sumar esas diferencias o residuos. Esta suma de diferencias resultara mínima cuando todos los valores se aproximan al promedio y máxima cuanto más se *dispersan* en torno a él: esperaríamos que fuera relativamente pequeña en la villa de emergencia o el country y más grande en el barrio urbano de población diversa.

Ahora bien, puesto que algunos casos se alejan del promedio hacia arriba y otros hacia abajo, habrá de ocurrir que si hacemos la suma algebraica de los residuos, éstos se compensarán y el resultado será igual a cero, tal como se aprecia en la tabla siguiente, que retoma el ejemplo de las tasas de desempleo en cinco aglomerados urbanos:

Tabla 5.3.1

Aglomerado	X	$X - \overline{X}$
Corrientes	19,7	6,68
Gran Resistencia	17,4	4,38
Ciudad de Buenos Aires	13,5	0,48
Gran Mendoza	11,5	-1,52
Santa Cruz	3,0	-10,02
Media aritmética	13,02	$\Sigma = 0$

Fuente: EPH-INDEC (Onda octubre de 2002)

Para sortear este inconveniente, podríamos obviar el signo y hacer una suma de los valores absolutos y luego promediarla (vale decir, dividirla por el número total de observaciones, que se simboliza con n[18]). Esto sí nos proveería una primera medida de la magnitud de la dispersión, aunque bastante rudimentaria: esta medida –el promedio de los desvíos absolutos de cada puntuación a la media– se denomina desviación media.

manual metodologia 22/5/08 15:36 Página 339

$DM = \Sigma\ [x - \bar{x}] / n$

(los corchetes indican que se trata de la suma de los valores absolutos, prescindiendo de los signos).

La varianza

Hay otro modo de sortear el problema de la suma cero: si se elevan al cuadrado los residuos a la media, todos ellos se tornan positivos y pueden sumarse. Esa suma, que se denomina suma cuadrática o suma de cuadrados, tiene una gran importancia en estadística porque sirve como insumo en la construcción de muchas otras medidas. En lo inmediato nos servirá, mediante el simple expediente de promediarla, para construir una medida de dispersión más precisa y útil que la ya vista, denominada varianza. La varianza se indica con la letra griega sigma elevada al cuadrado[19]: s^2

Tabla 5.3.2

Aglomerado	X	$X - \bar{X}$	$(X - \bar{X})^2$
Corrientes	19,7	6,68	44,62
Gran Resistencia	17,4	4,38	19,18
Ciudad de Buenos Aires	13,5	0,48	0,23
Gran Mendoza	11,5	-1,52	2,31
Santa Cruz	3,0	-10,02	100,40
Media aritmética	13,02 Σ	$\Sigma = 0$	$\Sigma = 166,75$

Fuente: EPH-INDEC (Onda octubre de 2002)

$\sigma^2 = \Sigma\ (X - \bar{X})^2/n = 166,75/5 = 33,34$

Esta sencilla fórmula sufre (como las de la mediana y la media aritmética) ciertas complicaciones si no disponemos de los datos totalmente desagregados, sino dispuestos en intervalos de clase y con las frecuencias agrupadas. Una vez más, es innecesario ocuparse aquí de tales cuestiones.

La varianza tiene importancia conceptual y es, asimismo, la base de otras medidas estadísticas que se derivan de ella. Pero su empleo directo es limitado, aunque intuitivamente se puede comprender que una varianza grande (lo que requiere que la suma de cuadrados también lo sea) se corresponde con mucha dispersión y una varianza pequeña, inversamente, con escasa dispersión (o mucha homogeneidad, que es lo mismo).

manual metodologia 22/5/08 15:36 Página 340

El desvío estándar

A partir de la varianza se deriva directamente otra medida de dispersión de mucho uso e importancia: el desvío estándar o desviación típica. No es más que la raíz cuadrada de la varianza (o, lo que es análogo, la varianza es el cuadrado del desvío estándar). Se indica, en consecuencia, con los mismos símbolos que la varianza (σ o s) pero sin elevar al cuadrado.

$$\sigma = \sqrt{\Sigma (X - \overline{X})^2/n} = \sqrt{166,75/5} = 5,77$$

Al igual que la varianza, el desvío estándar tiene múltiples aplicaciones estadísticas y muchas veces se lo emplea en forma directa cuando se debe comparar la variabilidad entre dos distribuciones distintas. Por ejemplo, ¿las tasas de desempleo de los aglomerados urbanos que releva la EPH eran, hace diez años atrás, más homogéneas o más heterogéneas que en la actualidad? Si limitamos el ejemplo a los cinco aglomerados considerados en el ejemplo que venimos utilizando, podríamos calcular fácilmente el desvío estándar de las tasas de desempleo de octubre de 1992. Veámoslo:

Aglomerado	X	$X - \overline{X}$	$(X - \overline{X})^2$
Corrientes	3,6	-0,78	0,6084
Gran Resistencia	5,4	1,02	1,0404
Ciudad de Buenos Aires	4,8	0,42	0,1764
Gran Mendoza	4,4	0,02	0,0004
Santa Cruz	3,7	-0,68	0,4624
Media aritmética	4,38	$\Sigma = 0$	Σ=2,288

Fuente: EPH-INDEC (Onda octubre de 2002)

$$\sigma = \sqrt{\Sigma (X - \overline{X})^2/n} = \sqrt{2,288/5} = 0,68$$

Según se aprecia, el desvío estándar arrojaba un valor mucho menor en 1992, sugiriendo una mayor homogeneidad en las puntuaciones. No obstante, es preciso no olvidar que el desvío estándar (al igual que la varianza y las medidas de tendencia central y posición) está expresado en puntuaciones de la variable. Y tratándose de las tasas de desempleo, los valores absolutos de estas puntuaciones eran mucho más elevados en 2002 que en 1992 (cuando la desocupación todavía se mantenía en su registro histórico, próximo al 5%).

Si se observa la fórmula, no resulta difícil entender que, si las puntuaciones son pequeñas también lo será la media. Y, en consecuencia, los residuos a la media no podrán ser –en términos absolutos– muy grandes, como tampoco sus cuadrados, ni el promedio de la

manual metodologia 22/5/08 15:36 Página 341

sumatoria de estos últimos. De manera que el resultado final habrá de ser –necesariamente– una cifra chica. En cambio, si se trata de números grandes sucederá lo inverso y el resultado final tenderá a ser elevado.

Esto quiere decir que el valor del desvío estándar no solo refleja el grado de heterogeneidad u homogeneidad de la distribución, sino que también está determinado por la magnitud absoluta de las puntuaciones. Ello hace que no sea una medida adecuada para comparar distintas distribuciones toda vez que los puntajes originales sean manifiestamente diferentes, como aquí ocurre. No sabemos cuánto del aumento del desvío estándar obedece a que las tasas de estos aglomerados urbanos son, efectivamente, más desiguales entre sí y cuánto a que todas ellas han tendido a aumentar: en promedio se habían triplicado (4,38 en 1992 y 13,02 en 2002).

El coeficiente de variación

Es preciso, pues, contar con una medida que refleje puramente el aumento de la heterogeneidad, no influenciado por el incremento absolutos de las puntuaciones. Ello no es nada difícil si se divide el valor del desvío estándar por el valor de la media aritmética, con lo cual se logrará estandarizarlo: ya no estará expresado en puntuaciones de la variable, sino en "medias aritméticas" de cada una de las distribuciones que comparamos. Esta medida se denomina coeficiente de variación y resulta un buen descriptivo del grado de heterogeneidad de una distribución.

$$CV = \sigma / \overline{X}$$

Si ahora calculamos ambos coeficientes de variación obtenemos valores de 0,16 para 1992 y 0,44 para 2002. Esto nos dice que el grado de desigualdad o heterogeneidad entre las tasas de desempleo de estos aglomerados se ha triplicado. La comparación directa de los desvíos estándar nos hubiera conducido a afirmar, en cambio, que la heterogeneidad creció más de ocho veces, lo cual sería erróneo.

6. Trascendiendo la descripción univariada: las hipótesis y el cruce de variables

Las técnicas estadísticas expuestas hasta ahora sirven para resumir y examinar los datos, con el propósito de que nos sugieran ideas. Pero frecuentemente, acudimos a ellos con ideas previas y con el propósito de verificar si estas ideas son acertadas o erróneas. Claro está que –según quedó dicho en capítulos anteriores– estas dos alternativas no son en modo alguno excluyentes y muchas veces se entremezclan y superponen. Por ejemplo, es muy frecuente que al refutar alguna idea preconcebida, los datos nos induzcan a imaginar otra nueva.

manual metodologia 22/5/08 15:36 Página 342

6.1. Las hipótesis y los datos

Las hipótesis, ya lo hemos visto, son conjeturas acerca de posibles relaciones entre variables. Estas conjeturas procuran responder a preguntas de investigación.

- Suponemos, por ejemplo, que las tasas de desempleo crecerán a medida que disminuye el nivel educativo (aquí, la pregunta sería: ¿por qué algunas personas suelen estar desempleadas con más frecuencia que otras?).
- O pensamos que esto mismo influye en la probabilidad de obtener un trabajo en el sector formal de la economía: creemos que esta probabilidad será más alta si se tienen más calificaciones educativas (la pregunta correspondiente será: ¿qué es lo que hace que algunas personas encuentren mejores trabajos que otras?)
- O bien, conjeturamos que la distribución de la población ocupada por rama de actividad diferirá según se trate de hombres o mujeres (en este caso, hay una pregunta que apunta a propósitos descriptivos: ¿cuáles son los trabajos más frecuentes de los varones y las mujeres?; ¿difieren entre sí?).
- O que la distribución por edades de los ocupados puede ser diferente entre varones y mujeres: tal vez, los varones comienzan a trabajar antes y permanecen hasta una edad más avanzada, porque se jubilan más tarde, además de que algunas mujeres dejan de trabajar mientras tienen hijos pequeños (aquí nos preguntamos: ¿son ciertas estas presunciones?).

Todas estas suposiciones son hipotéticas: no son ocurrencias caprichosas o meros acertijos. Hay indicios o razones para pensar así. Pero no estamos seguros de que lo que creemos sea verdad y hay que ponerlo a prueba.

El propósito de cruzar las variables en los cuadros o *tablas de contingencia* es, frecuentemente, verificar en qué medida son ciertas estas presunciones. Visto así, los cuadros que cruzan variables vienen a ser la expresión operacional de las hipótesis, que nos permitirán decidir acerca de ellas.

Existe, pues, una estrecha relación –que una vez más debe ser resaltada– entre datos y teorías. Lo que nos lleva a cruzar ciertas variables y no otras son nuestras conjeturas, desprendidas de la teoría. Lo que procuramos advertir al examinar y analizar los cuadros es si tales proposiciones basadas en la teoría son o no acertadas.

Aunque eventualmente pueda procederse en forma inductiva y esperar que los mismos cuadros nos sugieran relaciones, muy raramente se hacen cruces azarosos, de "todo con todo".

6.2. Comparaciones o explicaciones

Cuando suponemos estas relaciones entre variables, existe un matiz de diferencia. A veces, simplemente, pensamos que dos subgrupos o subuniversos pueden diferir en algo y

manual metodologia 22/5/08 15:36 Página 343

queremos compararlos. Es el caso de la distribución por edades de varones y mujeres ocupados. No es que creemos que el sexo induce variaciones en la edad. Simplemente, conjeturamos que existen razones que pueden hacer que estos grupos difieran. Otro tanto sucede si hacemos esta misma comparación entre las poblaciones de una provincia con comportamientos migratorios expulsores (como Corrientes) y otra que suele recibir flujos migratorios (como Santa Cruz o Tierra del Fuego): en estos casos, esperaremos que en la primera sobreabunden niños y personas mayores, pero exista una proporción baja de adultos jóvenes. Al revés sucederá en la provincia receptora, porque generalmente migran las personas en edades económicamente activas. Cuando es así, decimos que usamos una **variable de corte o de comparación** (que sería el sexo en el primer caso o la jurisdicción en el segundo), para examinar la distribución diferencial de la variable de distribución (la edad).

En cambio, otras veces, ponemos más énfasis en el vínculo de influencia o determinación de una variable sobre la distribución de la otra. Creemos que una "hace variar" a la otra. Pensamos, por caso, que el nivel educativo modifica o hace variar la probabilidad de obtener cierta clase de empleos: a medida que se es más educado aumenta la probabilidad de obtener empleos en el sector formal. En estos casos, hablamos de **variable independiente**. Y, de alguna manera, estamos pensando en un vínculo causal. Aunque debe quedar claro que la mera asociación empírica, comprobable por métodos estadísticos, no es suficiente para probar causalidad (concepto al que nos referiremos en el próximo capítulo), aunque sí necesaria: es, apenas, un requisito.

6.3. Status lógico de las variables

En las oportunidades en que cruzamos variables, ya sea con propósitos de comparar entre subgrupos o de explorar relaciones de determinación, debemos decidir cuál de las variables opera como variable de corte o bien como variable independiente. A los efectos de estas decisiones, aunque no hay unas reglas fijas, existen criterios orientadores. El primero es el status lógico de las variables:

Variables de base o atributos sociodemográficos	→	Variables de actitud u opinión	→	Variables de conducta
Nivel socioeconómico		*Ideología*		*Voto*

En general, hemos de suponer que las variables de base o sociodemográficas (el sexo, la edad, el nivel educativo, el nivel socioeconómico) anteceden a las variables actitudinales (como podrían serlo la ideología política, la opinión sobre un tópico particular o la percep-

manual metodologia 22/5/08 15:36 Página 344

ción acerca de la situación económica). En tanto que estas predisposiciones serán antecedentes respecto de las conductas (como la emisión del voto, la decisión de concurrir o no a una concentración o las decisiones de ahorrar dinero en el banco o invertirlo en cambiar de automóvil). El planteo de Bourdieu acerca del modo en que las posiciones sociales objetivas que la gente ocupa (*campus*) se transforma en subjetividad (*habitus*), que a su vez influye sobre la conducta, permite sustentar esta secuencia.

La dirección de dicha secuencia no es, sin embargo, fija e inalterable. Podría ocurrir que cierta actitud (por ejemplo, una fuerte tendencia a la competencia con los demás) impulsara a alguien a ascender socialmente, determinando –a lo largo del tiempo– su nivel socioeconómico. El planteo de Max Weber en *La Etica Protestante y el Espíritu del Capitalismo* va en este sentido.

En ocasiones, sin embargo, necesitamos relacionar variables de base entre sí. Por ejemplo, el nivel educativo y la situación laboral. En estos casos, un primer criterio para establecer la secuencia que atribuimos a la relación entre las variables, consiste en distinguir entre status adscriptos y adquiridos. La edad y el sexo son –en principio– status adscriptos, que tenemos adheridos. Permanecen fijos o se modifican en forma independiente de nuestra voluntad[20]. El nivel socioeconómico de la familia de origen, de igual modo, es algo que recibimos al nacer. En cambio, el nivel educativo o la ocupación propia son adquisiciones, como también podría decirse que lo es el nivel socioeconómico de un sujeto, aunque obviamente estará influenciado por el de su hogar de origen: nadie duda que quienes provienen de familias de clase alta tienen mejores oportunidades para adquirir educación y, luego, para acceder a mejores empleos[21].

En segunda instancia –por ejemplo si se trata de relacionar dos status adquiridos– puede apelarse a la idea de la prelación temporal. En general, tendemos a pensar que existen etapas relativamente secuenciadas en la vida de las personas: primero se adquiere la educación formal y luego se ingresa al mundo del trabajo; desde este punto de vista la educación sería variable independiente con respecto a la ocupación. Pero también esto puede invertirse: puede ser que los jóvenes que deben incorporarse tempranamente al mundo del trabajo, interrumpan, por este motivo, sus estudios.

De igual modo, si hemos de relacionar el nivel educativo alcanzado por las mujeres con la edad que tienen al iniciar su vida reproductiva (se sabe que en general, así como aumenta el nivel educativo tiende a diferirse el inicio de la fecundidad), tenderemos a pensar que la duración de la etapa formativa es la variable independiente: las mujeres que prolongan su escolaridad suelen diferir, por esa razón, tanto el inicio de su vida conyugal como reproductiva. Pero también puede suceder que la maternidad precoz –en las adolescentes– precipite la interrupción de los estudios.

manual metodologia 22/5/08 15:36 Página 345

En fin, que hay apenas unas pocas reglas orientadoras pero no fijas para determinar cómo hemos de suponer las relaciones entre variables. En lo esencial, ello depende del marco teórico que adoptemos.

6.4. Construcción de tablas de contingencia: posiciones de las variables

Es usual, por convención que no siempre se respeta, ubicar la variable de corte o independiente en las columnas de la tabla, en tanto que suele ubicarse la otra en las filas. La disposición de la tabla 6.4.1 usa el sexo como variable de corte para examinar en forma comparada las distribuciones por edades de varones y mujeres.

Tabla 6.4.1. Población ocupada por tramos de edad, según sexo
Gran Buenos Aires

Edades	Varones	Mujeres	Total
Hasta 24			
De 25 a 34			
De 35 a 44			
De 45 a 49			
De 50 a 54			
De 55 a 59			
60 y más			
Total			

Fuente: elaboración propia en base a EPH-INDEC (onda de ...)

Ocasionalmente, se deja de lado esta regla tradicional por razones prácticas y estéticas: si la variable independiente tiene muchas categorías, será más razonable colocarla en las filas, pues es preferible tener un cuadro largo y no uno tan ancho que no quepa en la página: la razón fundamental es que, en la mayor parte de los textos, las páginas son más largas que anchas (aunque los procesadores de texto permiten introducir una página *apaisada*, si es preciso). Por ejemplo, en el cuadro que sigue a este párrafo no podríamos incluir más columnas sin recurrir a una página horizontal. Este recurso incluye siempre alguna molestia para el lector, que debe cambiar de posición el texto.

manual metodologia 22/5/08 15:36 Página 346

Tabla 6.4.2. Sector de ocupación según nivel educativo
Total de ocupados. Gran Buenos Aires

Sector de ocupación	Hasta primaria incompleta	Hasta secundaria incompleta	Hasta superior incompleta	Superior completa	Total
Sector formal					
Sector informal					
Total					

Nota: cualquier nota aclaratoria se coloca antes de la fuente
Fuente: elaboración propia en base a EPH-INDEC (onda de...)

6.5. Elementos presentes en los cuadros o tablas de contingencia

El contenido de las celdas

En una tabla, las frecuencias de las celdas se denominan frecuencias *condicionales*: ellas suponen una clasificación bivariada de las unidades de análisis, puesto que se trata de la cantidad de casos que comparten, simultáneamente, una cierta categoría en cada una de las dos variables. En la tabla 6.4.1, por ejemplo, la celda sombreada contendrá la cantidad de personas que comparten la doble condición de ser de sexo masculino y no exceder de 24 años de edad.

En tanto que los totales de las columnas reciben, usualmente, el nombre de *subtotales:* se trata de la distribución univariada de todas las unidades de análisis según la variable independiente o de corte. En la tabla 6.4.1, al pie de la primera columna tendremos la cantidad total de varones.

Por su parte, los totales de fila (por la variable de distribución o dependiente) se suelen llamar *marginales*: en este caso constituyen la distribución univariada de la muestra o población según la variable situada en las filas. En la tabla 6.4.1, al final de la primera fila tendremos la cantidad total de personas de hasta 24 años.

Por fin, la celda situada en la esquina de abajo y a la derecha, contiene el total de casos incluidos en el cuadro. La frecuencia contenida en esta celdilla resulta tanto de la sumatoria de la última columna como de la última fila. Y también, por supuesto, de la suma de las frecuencias condicionales de todas las celdas interiores.

Titulado, notas y fuentes

Ya hemos visto algunas reglas referidas a los elementos que son externos a los cuadros –titulado, notas y fuente– al tratar las tablas más simples, que muestran la distribución de una sola variable. La totalidad de esas normas es aplicable, asimismo, a los cuadros que tienen dos variables, como también a los que incluyen más de dos (de los que nos ocuparemos oportunamente).

manual metodologia 22/5/08 15:36 Página 347

Pero es preciso agregar que, en estos casos, hay que mencionar en el título ambas variables involucradas ¿En qué orden hacerlo?: lo usual es mencionar primero la variable de distribución y en segundo término la variable independiente o de corte. Los títulos de las tablas 6.4.1 y 6.4.2 proporcionan sendos ejemplos de lo dicho.

6.6. Análisis de tablas de contingencia: tres modos de obtener los porcentajes

El modo más usual de analizar una tabla consiste en expresar las frecuencias de las celdas en porcentajes. Frente a esta tarea, caben tres alternativas:

- obtener porcentajes sobre los subtotales (por las columnas)
- obtener porcentajes sobre los marginales (por filas)
- obtener porcentajes sobre el total de casos (sobre la celda inferior de la derecha = N)

Esta última alternativa ofrece una utilidad escasa y es poco usual. Porque lo único que haría es mostrarnos las mismas frecuencias absolutas, pero convirtiendo a 100 el total de casos, sin proporcionar información adicional. Sólo nos diría, por ejemplo, qué proporción del total de ocupados son mujeres de 25 a 34 años. Esto no respondería a nuestra conjetura inicial (que diferirían las edades de los ocupados, entre mujeres y varones, lo que exigiría una comparación). No sería, sin embargo, inútil saberlo si nuestro propósito fuese otro: saber qué proporción de trabajadoras están en edad de criar hijos pequeños, a los efectos de revisar la normativa acerca de guarderías a cargo de los empleadores.

6.7. La lectura porcentual de las tablas de contingencia: la "influencia de la variable independiente" y la "descripción de perfil" de un grupo.

Aquí nos referiremos, principalmente, a las dos primeras modalidades de análisis: la influencia de la variable independiente y la comparación de perfiles.

La influencia de la variable independiente

Cuando queremos ver cómo una variable "hace variar" a otra; por ejemplo, mostramos cómo la incidencia de la informalidad varía según el nivel educativo (o, lo que es lo mismo, que si comparamos trabajadores con diferente nivel educativo, variará la proporción de informales entre unos y otros niveles):

manual metodologia 22/5/08 15:36 Página 348

Tabla 6.7.1. Población ocupada por sector de ocupación, según nivel educativo (en %) Gran Buenos Aires

Sector de ocupación	Primaria incompleta	Primaria completa	Secundaria incompleta	Secundaria completa	Superior incompleta	Superior completa	Total
Formal	66,0	68,5	72,1	79,7	80,4	87,8	75,5
Informal	34,0	31,5	27,9	20,3	19,6	12,2	24,5
Total	100,0	100,0	100,0	100,0	100,0	100,0	100,0

Fuente: elaboración propia en base a EPH-INDEC (onda de ...)

Aquí, diremos que "la proporción de informales disminuye el aumentar el nivel educativo: más de un tercio de los que no han completado la primaria lo son. Esa proporción baja a uno de cada cinco cuando se ha completado el nivel medio. En cambio, poco más de uno de cada diez se desempeña en ese sector cuando se trata de profesionales". Otro modo de decirlo es que, "entre los de menor nivel educativo, la proporción de informales casi triplica a la que se observa entre los universitarios".

No es cierto, de ningún modo, que un 34% de los informales tengan baja educación. Ni que un 12% de ellos sean universitarios. No hablamos aquí de los informales sino de los de bajo o alto nivel educativo. Los comparamos y miramos cómo incide en ellos la informalidad. Queremos ver cómo esta varía según el nivel educativo. Qué cosas la determinan.

Adviértase que al "leer" el cuadro, no ha sido necesario mencionar la totalidad de las cifras: ¿para qué hacerlo si ya están en el cuadro? Sólo ponemos énfasis en aquellas que sirven mejor para apuntalar nuestras observaciones.

En este caso, seguramente, la conjetura que motivó la construcción de esta tabla era que las personas con más educación tienen mejores oportunidades para insertarse en la economía formal. Sabemos que ésta suele emplear tecnologías más avanzadas y, por eso, requiere más calificaciones. En cambio, pensamos que probablemente los menos educados deberán conformarse con empleos de menor calidad, seguramente menos remunerados, en empresas pequeñas o bien por cuenta propia: las posiciones típicas de la llamada economía informal tal como la definía el Programa Regional de Empleo para América Latina y el Caribe (PREALC): "...la franja de actividades de baja productividad en que se inserta el excedente de población incapaz de ser absorbido por las ocupaciones generadas en el sector moderno de la economía urbana" (PREALC/OIT, 1978). Esto es lo que queremos ver. Y, en consecuencia, al mirar los números procuramos apelar a las cifras y a las comparaciones entre ellas que permitan afianzar o refutar nuestra suposición. ¿Qué esperaríamos si ella fuera cierta?: pues encontrar porcentajes de informales más bajos en las categorías de alta educación. O, lo que es lo mismo, porcentajes de trabajadores formales más altos en

manual metodologia 22/5/08 15:36 Página 349

ellas y decrecientes así como disminuye el nivel educativo. Pues eso es, precisamente, lo que muestra el cuadro 6.7.1 y lo que permite afianzar la hipótesis.

La comparación de perfiles (o de composición

Es el caso en que queremos ver cómo se compone un subgrupo, por ejemplo, los asalariados precarios, por sexo (o por nivel educativo): aquí, queremos describir ese subgrupo mostrando que tiene cierta especificidad con respecto a otros subgrupos o al grupo total.

Tabla 6.7.2. Distribución de los asalariados por nivel educativo según condición de registro (en %)
Gran Buenos Aires

Condición de registro	Primaria incompleta	Primaria completa	Secundaria incompleta	Secundaria completa	Superior incompleta	Superior completa	Total
Registrado	4,2	18,5	15,3	23,2	18,1	20,8	100,0
No registrado	9,7	29,4	23,8	15,9	12,9	8,4	100,0
Total	6,6	23,2	19,0	20,0	15,8	15,4	100,0

Fuente: elaboración propia en base a EPH-INDEC (onda de ...)

¿Qué pregunta estaría tratando de responder quien construyera esta tabla, obteniendo los porcentajes de este modo? Veamos la probable línea de razonamiento seguida. Se plantea una discusión acerca del trabajo en negro que, como se sabe, se extendió mucho en la Argentina en los años noventa. Una de las posturas en debate sostiene que quedan relegados al trabajo en negro los trabajadores que reúnen pocas calificaciones educativas. En cambio, la otra sostiene que esta forma de trabajo se ha extendido tanto, que también incluye una elevada proporción de trabajadores con alta educación ¿Será así?. Una de las primeras cosas que sería menester hacer para ponerlo a prueba sería ver cuál es la composición de los asalariados no registrados por nivel educativo. Y, seguramente, interesaría compararlos con los que trabajan en regla. Tal lo que muestra la tabla 6.7.2.

Aquí estamos comparando diferentes distribuciones. En este caso podremos afirmar que las personas que no han completado la primaria están sobrerrepresentadas entre los precarios: casi uno de cada diez asalariados precarios está en esa situación, cuando esta proporción es de 4% entre los no precarios y de 7% en la población total. Hay, pues, una diferencia de más de tres puntos porcentuales (no de 3%). Hay más personas de baja educación entre los asalariados precarios: es una especificidad de ese grupo. En cambio, solo 8% de los precarios son universitarios, cuando la proporción de universitarios es el doble en el total de asalariados y casi el triple entre los asalariados registrados. Acá resaltamos la

manual metodologia 22/5/08 15:36 Página 350

especificidad de estos que nos interesan (los trabajadores en negro). Pero no basta decir que uno de cada diez tiene baja educación: hay que ver si eso es distinto que entre los asalariados totales o entre los registrados.

En cambio, no podemos decir que un diez por ciento de los de baja educación son precarios. Eso no es verdad. No hablamos de "los de baja educación": hablamos de los precarios. Describimos la precariedad: destacamos sus diferencias con la no precariedad, sus notas distintivas.

Efectivamente, diríamos, pues, que el trabajo precario selecciona con preferencia a personas de pocas calificaciones educativas, aunque su expansión ha terminado por hacer que también incluya a algunos trabajadores mejor calificados: inclusive, un 8% de universitarios.

6.8. Para presentar los cuadros de incidencia: si nos interesa una sola categoría

En el primero de los dos cuadros (6.7.1), sería posible mostrar una sola de las filas: la de los registrados o la de los no registrados, porque ambas suman 100, de modo que una es –por fuerza– el complemento de la otra. Esto es posible cuando queremos mostrar la incidencia de algo.

Tabla 6.8.1. Población ocupada: incidencia de la informalidad según nivel educativo (en %)

Gran Buenos Aires

Sector de ocupación	Primaria incompleta	Primaria completa	Secundaria incompleta	Secundaria completa	Superior incompleta	Superior completa	Total
Informal	34,0	31,5	27,9	20,3	19,6	12,2	24,5

Fuente: elaboración propia en base a EPH-INDEC (onda de ...)

Sería, pues, ocioso e innecesario sobrecargar el cuadro con la otra fila, que nos exhibiera los porcentajes complementarios correspondientes a los trabajadores del sector formal. De haber querido resaltar la formalidad, hubiéramos procedido al revés, mostrando solamente los porcentajes de trabajadores formales en cada nivel educativo.

6.9. Algunas propiedades de la asociación entre variables: forma, sentido e intensidad

La idea de asociación

Tanto cuando vimos que las personas de baja educación tenían una mayor proclividad a insertarse en el sector informal (cuadro 6.7.1) como cuando apreciamos que las ocupaciones precarias reclutan con mayor asiduidad personas de bajas calificaciones educativas

manual metodologia 22/5/08 15:36 Página 351

(cuadro 6.7.2), lo que estamos diciendo es que las variables se *asocian* entre si. Más precisamente, esta *asociación* existe entre cierta categoría de una de las variables con cierta categoría de la otra: quien cae en la categoría "Primaria incompleta" tiende a caer en la categoría "informal" (en el primer cuadro) o en la categoría "No registrado" (en el caso del segundo).

Al revés, podríamos decir que –en el caso del primero de los cuadros– la categoría "Superior completa" se asocia con la categoría "Formal", así como en el segundo cuadro se asocia con "Registrado".

Este concepto de *asociación estadística* –que trataremos muy someramente aquí– es sumamente importante y debe ser bien comprendido ¿Por qué decimos que existe esta asociación entre ciertos valores de una de las variables y ciertos valores de la otra? ¿Cómo lo sabemos?

En realidad, son los porcentajes los que nos permiten descubrirlo. Pensemos en el primero de los cuadros (el 6.7.1). Supongamos que el hecho de trabajar en un empleo formal o informal fuese enteramente fortuito y azaroso. Que no dependiera para nada de los atributos educativos con que cuentan los individuos, sino más bien de su suerte o de cualquiera otra condición que nos es desconocida. Si esto sucediera así, las proporciones de trabajadores formales e informales serían aproximadamente las mismas en todas las categorías educativas. Y por lo tanto, no diferirían de las que presenta el conjunto total de la población, cuando no se tiene en cuenta el nivel educativo.

Pero en ese cuadro no sucede así: mientras que la población total se distribuye en 75% de formales y 25% de informales, aproximadamente, en las distintas categorías no ocurre otro tanto. La proporción de informales, que es una cuarta parte en el total, se eleva a más de un tercio entre los menos educados. En tanto que la proporción de trabajadores formales, que ronda las tres cuartas partes en el conjunto total, se eleva "en forma inesperada" a casi 88% si se trata de los más educados. **En esos aumentos de los porcentajes estriba, justamente, lo que llamamos asociación.** En otros términos, si ciertos valores de las variables se asocian, entonces la celda correspondiente presentará un porcentaje más elevado que el que se observa en la distribución marginal, situada en la última columna.

Esta misma lógica podemos, también, aplicarla al segundo cuadro (el 6.7.2). En este caso, bajo el supuesto de que los trabajos registrados no discriminaran por nivel educativo, habríamos de esperar que entre quienes los ocupan, la distribución por niveles educativos fuera semejante a la del total. Por ejemplo, si en el conjunto total de casos hay algo menos de 7% de personas que no completaron la educación básica, entonces el porcentaje debiera ser similar a ese tanto entre asalariados registrados como entre los no registrados. Y lo mismo debiera suceder con los de alta educación (superior completa), que alcanzan a 15% en el total: también debieran rondar esa proporción entre registrados y no registrados.

manual metodologia 22/5/08 15:36 Página 352

Nuevamente, no es esto lo que muestra la tabla: la proporción de trabajadores con baja educación (primaria incompleta) sube a casi 10% entre los no registrados. A la vez, la de sujetos con educación superior aumenta a casi 21% entre los registrados. **Nuevamente, estas elevaciones de los porcentajes con respecto a la distribución que –en este caso– vemos en los subtotales, es lo define la existencia de asociación**.

¿Y por qué en uno de los casos comparamos con la distribución marginal, en tanto que en el otro lo hicimos con la subtotal? Pues bien, esto depende del sentido en que hemos porcentualizado: si los porcentajes han sido obtenidos al interior de cada columna, entonces comparamos con la última columna: en ella tenemos una distribución no afectada por el nivel educativo. En cambio, cuando obtenemos porcentajes por filas, entonces hemos de comparar con la última fila: en ella tenemos una distribución no desagregada por la condición de registro de los asalariados.

El concepto opuesto al de asociación es el de **independencia** entre las variables: cuando la distribución entre formales e informales es indiferente al nivel educativo, o bien cuando la distribución por nivel educativo no difiere entre trabajadores registrados y no registrados, entonces afirmaremos que las variables son independientes entre sí. No están asociadas.

Intensidad de la asociación: la diferencia porcentual

Pero tan pronto como se comprende el concepto de asociación y se constata su eventual existencia, surge una segunda cuestión. Dos variables que se asocian pueden hacerlo muy estrechamente o bien débilmente. Un ejemplo en el que, deliberadamente, ambas variables son dicotomías[22], permitirá exponer esta cuestión de la intensidad o fuerza de la asociación.

Condición de actividad	Varones	Mujeres	Total	D%
Ocupados	85%	70%	80%	15
Desocupados	15%	30%	20%	-15
Total	100%	100%	100%	

En esta tabla, cuyos datos son ficticios, la proporción de ocupados, que es de 80% en el total, crece a 85% entre los varones y desciende a 70% entre las mujeres. Diríamos, pues, que el sexo se asocia con la condición de actividad. ¿Poco o mucho? Imaginemos, para responder a esta pregunta, dos situaciones extremas:

Condición de actividad	Varones	Mujeres	Total	D%
Ocupados	80%	80%	80%	0
Desocupados	20%	20%	20%	0
Total	100%	100%	100%	

manual metodologia 22/5/08 15:36 Página 353

En la tabla que antecede, las variables son perfectamente independientes, porque la proporción de ocupados –y la de desocupados– es exactamente igual entre varones y mujeres y, por lo tanto, igual a la que aparece en el total. Pero en la tabla que sigue aparece la situación opuesta: las variables están todo lo asociadas que podrían estarlo:

Condición de actividad	Varones	Mujeres	Total	D%
Ocupados	100%	0%	60%	100
Desocupados	0%	100%	40%	-100
Total	100%	100%	100%	

Efectivamente: mientras que la totalidad de los varones están ocupados, todas las mujeres están desocupadas.

Ahora podemos explicarnos la presencia de la columna que se agregó a la derecha de cada una de las tablas de contingencia. Si al porcentaje de la primera columna (varones) le restamos el de la segunda (mujeres) obtenemos una medida rudimentaria del grado de asociación, comúnmente llamada *diferencia porcentual*. Esta medida, como se aprecia, asume valor cero en la segunda tabla (cuando las variables son independientes) y valor 100 en la tercera (cuando están totalmente asociadas). En cambio, en la primera tabla, alcanza un valor de 15 *puntos porcentuales*[23]. Tenemos pues, una medida que varía en un rango de cero a 100: por comparación con estos límites teóricos, diríamos que el valor 15 obtenido en la primera de las tres tablas, corresponde a una débil asociación. La fuerza o intensidad de la asociación crece así como esta diferencia se aproxima a 100 y decrece cuando tiende a cero.

La diferencia porcentual, pues, es una medida de asociación estandarizada entre cero y cien, muy útil cuando las variables son dicotómicas

Claro está que esto es así de sencillo cuando se trata de cuadros que cruzan variables de sólo dos categorías. Pero si son más –como es habitual– las cosas se complican. En primer lugar, en el ejemplo que brindamos la diferencia porcentual asume un único valor, igual aunque de signo inverso, en ambas filas. Esto es así solamente en caso de dicotomías, pero en cambio, cuando los cuadros tienen más filas, tendremos tantas diferencias como filas, todas ellas distintas entre sí[24]:Ya no hay, pues, una única medida de la intensidad de la asociación que caracterice a todo el cuadro.

Condición de actividad	Varones	Mujeres	Total	D%
Ocupados	64%	41%	51%	23
Desocupados	10%	8%	9%	2
Inactivos	26%	51%	40%	-25
Total	100%	100%	100%	

manual metodologia 22/5/08 15:36 Página 354

Y si se agregan una o más columnas, entonces los problemas se acentúan:

	Tramos de edad				D%
	14 a 24 años	25 a 59 años	60 y más años	Total	
Ocupados	30%	71%	23%	51%	¿?
Desocupados	14%	9%	3%	9%	¿?
Inactivos	56%	20%	74%	40%	¿?
Total	100%	100%	100%	100%	

El problema estriba en que ahora, al tener más columnas, podemos obtener –para cada una de las filas– más de una diferencia porcentual. Podríamos optar por calcularlas entre las columnas extremas (14 a 24 y 60 y más) o bien entre cualquiera de ellas y la columna intermedia (35 a 59) (Mora y Araujo, 1965). Todas expresarían cosas diferentes: limitándonos a la primera fila, la diferencia entre columnas extremas mostraría la mayor tendencia a la ocupación de los jóvenes con respecto a las personas mayores, mientras que cualquiera de las obtenidas con la columna del medio nos mostraría la predominancia de ocupación entre las personas de las edades centrales con respecto a jóvenes y viejos. Pero no tendríamos una única medida de la fuerza de la relación entre las variables.

No hay ninguna solución satisfactoria a este problema, aunque eventualmente se ha sugerido que lo más adecuado es emplear la que arroje un valor más elevado[25]. La diferencia porcentual, como medida sintética de la asociación en las variables, sólo funciona adecuadamente con las dicotomías: cuando los cuadros son más grandes la estadística dispone de un amplio conjunto de medidas de fuerza de la asociación para variables categóricas, cuyo estudio excede los límites de este texto[26].

Puesto que más arriba se ha definido la asociación entre las variables como el alejamiento de las frecuencias condicionales –expresadas en los porcentajes de las celdas– con respecto a las que esperaríamos hallar en el caso de independencia de las variables –expresadas en las distribuciones marginales– sería lícito preguntarse por qué no tratar de construir una medida de la fuerza de la asociación que tomara en cuenta ese alejamiento. Por cierto que esta idea fue concebida mucho tiempo atrás y es el fundamento de una prueba estadística de uso muy corriente[27].

Si las variables son ordinales: el sentido de la asociación

Cuando las variables cruzadas en las tablas de contingencia son nominales sólo podemos apreciar el modo en que sus valores se asocian (“cuáles con cuáles”, por así decirlo) y si esta asociación es intensa o débil. Pero si tenemos dos variables ordinales (también

manual metodologia 22/5/08 15:36 Página 355

podría tratarse de variables cuantitativas en las que se han construido intervalos de clase), entonces aparece un nuevo atributo de la asociación: el *sentido* o dirección de la misma. Veamos:

Ingreso	Calificación de la tarea			
	Alta	Media	Baja	Total
Alto	70%	20%	5%	30%
Medio	40%	65%	35%	55%
Bajo	10%	15%	60%	15%
Total	100%	100%	100%	100%

En este cuadro, es posible apreciar que cuanto más alta es la calificación de la tarea, más elevado es el ingreso. En la columna correspondiente a los trabajadores de calificación alta, es donde hallamos un mayor porcentaje de ingresos altos. Entre los trabajadores de calificación media, prevalecen los de ingresos medios. Finalmente, entre los de baja calificación sobreabundan los que perciben bajos ingresos. Adviértase que aparecen cargadas las celdillas del cuadro que van desde el extremo superior izquierdo al extremo inferior derecho: se trata de la denominada *diagonal positiva*. Cuando las frecuencias tienden a concentrarse en esta diagonal, podemos decir que existe una relación positiva entre las variables: a medida que crecen los valores de una de ella, también tienden a hacerlo los de la otra[28].

En el cuadro que sigue, se aprecia una relación diferente:

Fecundidad	Nivel educativo			
	Alta	Media	Baja	Total
Alta	1%	9%	40%	30%
Media	35%	61%	35%	45%
Baja	64%	30%	25%	25%
Total	100%	100%	100%	100%

A medida que aumenta su nivel educativo, las mujeres tienen menor cantidad de hijos. Efectivamente, la baja fecundidad predomina entre las de mejor nivel educativo y, a la inversa, entre las menos instruidas hay un porcentaje alto de mujeres con fecundidad elevada. Aquí, aparecen sombreadas las celdillas que conforman la diagonal negativa del cuadro, que va desde la esquina situada arriba y a la derecha hasta la ubicada abajo y a la izquierda. Diremos que hay una relación negativa entre las variables: así como crece la educación tiende a descender la fecundidad.

manual metodologia 22/5/08 15:36 Página 356

Esta propiedad de las relaciones entre variables –cabe decirlo una vez más– sólo puede predicarse cuando ellas son, al menos, de nivel ordinal.

6.10. Cuando la variable dependiente es cuantitativa: resumiendo la información

Hasta ahora, nos hemos ocupado de casos en que ambas variables cruzadas eran categóricas, pero no siempre sucede así. Por ejemplo, una vez que vimos cómo se distribuyen las personas que están ocupadas por su categoría ocupacional (empleadores, trabajadores por cuenta propia y asalariados), podríamos preguntarnos qué ingresos obtienen unos y otros por su trabajo. Es cierto que se pueden construir tramos de ingresos y cruzar dicha variable con la categoría ocupacional, Pero también se pueden analizar los ingresos aplicando a ellos el instrumental estadístico básico que es posible emplear con variables cuantitativas. Vale decir, medidas de tendencia central, de orden y de dispersión. En este caso, sería pertinente mostrar una tabla como la que sigue:

Tabla 6.9.1. Población ocupada por categoría ocupacional: indicadores de ingresos seleccionados (en pesos corrientes)
Total urbano

Categoría ocupacional	Patrón	Cuenta propia	Obrero o empleado	Total
Media	1.489	461	580	584
Mediana	1.000	300	425	400
Primer cuartil	500	150	200	200
Tercer cuartil	2.000	550	700	700
Desvío estándar	1.662	602	724	764
Coeficiente de variación	1.1	1.3	1.2	1.3

Fuente: elaboración propia en base a EPH-INDEC (IV trimestre de 2003)

Este cuadro tiene la información de ingresos resumida en las celdas y nos permite apreciar, por ejemplo, que el ingreso medio de los empleadores es equivalente a 2,6 veces el que perciben los asalariados. Pero estos aparecen favorecidos en relación con los autónomos, pues obtiene alrededor de 26% más. Para averiguar esto bastará con aplicar el procedimiento de calcular razones, ya aprendido:

Patrón/asalariado = 1489/580= 2,6
Asalariado/cuenta propia = 580/461= 1,26

También podríamos afirmar, considerando ahora el ingreso mediano, que la mitad de los asalariados ganan menos de $425. Esta cifra equivale a 42% de la mediana de los

manual metodologia 22/5/08 15:36 Página 357

empleadores. Pero la mediana de los trabajadores por cuenta propia representa, a su vez, 70% de la que corresponde a los asalariados.

Asalariados/empleadores*100 = 425/1000*100 = 42,5
Cuenta propia/asalariados*100 = 300/425*100 = 70,6

Parecería, pues, que ocupar una posición asalariada es más conveniente que desempeñarse por cuenta propia. Pero quienes cuentan con suficiente capital como para convertirse en patrones consiguen, asimismo, mejores ingresos.

Todo esto lo podríamos hacer también con los respectivos cuartiles. Igualmente podríamos haber elegido comparar a los empleadores con los cuentapropistas, a efectos de ver cuánta ventaja reporta disponer de una mayor dotación de capital a quien quiere desempeñarse por su cuenta. El cuadro ofrece las cifras, y el usuario podría disponer de ellas para realizar las comparaciones que más oportunas le parecieran. Pero si quisiéramos excusarle de la obligación de hacer cuentas, también sería factible agregar algunas de estas comparaciones en una columna adicional. Por ejemplo, podemos quitar la columna de total, para no sobrecargar el cuadro[29] y colocar otras que contengan las razones o los porcentajes que hemos estado calculando:

Tabla 6.9.2. Población ocupada por categoría ocupacional: indicadores de ingresos seleccionados (en pesos corrientes)
Total urbano

Categoría ocupacional	Patrón (1)	Cuenta propia (2)	Obrero o empleado (3)	(1)/(3)	(3)/(2)
Media	1.489	461	580	2,6	1,3
Mediana	1.000	300	425	2,4	1,4
Primer cuartil	500	150	200	2,5	1,3
Tercer cuartil	2.000	550	700	2,9	1,3
Desvío estándar	1.662	602	724	2,3	1,2
Coeficiente de variación	1,1	1,3	1,2	0,9	0,9

Fuente: elaboración propia en base a EPH-INDEC (IV trimestre de 2003)

Todavía, sería posible hacer algunas comparaciones en sentido vertical. Por ejemplo, los cocientes entre la media y la mediana en cada categoría ocupacional. Ya hemos visto antes que, cuando la media supera en mucho a la mediana, hemos de pensar que ello se debe a la presencia de algunos casos con valores muy elevados, en tanto que la mayoría se sitúan hacia los valores más bajos. Si lo hiciéramos, comprobaríamos que la media representa 1,5

manual metodologia 22/5/08 15:36 Página 358

veces el valor de la mediana entre patrones y trabajadores por cuenta propia, pero esa relación desciende a 1,4 para el caso de los asalariados.

6.10. Cuando las variables cambian de sujeto

Es posible pensar a las variables como unos predicados que se refieren a ciertos sujetos: las unidades de análisis.

Cuando una encuesta pregunta a la gente si está trabajando o está desocupada, indaga su edad, su ingreso laboral o su nivel educativo, cualquiera de estos atributos (trabajar, tener 35 años, ganar mil pesos al mes y haber completado el nivel secundario) son propiedades del sujeto interrogado. Pero el porcentaje total de personas que trabajan en relación con la población total de un país, el promedio de edad de los ocupados, el promedio de ingresos que obtienen o el porcentaje de ellos que han egresado de la educación media ya no corresponden a cada uno de estos sujetos, sino que son medidas que describen al conjunto. Pueden atribuirse, por ejemplo, al país, a la provincia o a la ciudad que habitan.

Por otra parte, algunos de estos atributos eran, al ser predicados de los individuos, variables categóricas: tal el caso de la ocupación o el nivel educativo. Sin embargo, al convertirse en medidas resumidas referidas a un conjunto mayor –provincia, país, etc. – tales como porcentajes o promedios, todos ellos se han convertido en variables cuantitativas. En estos casos, podemos comparar a estas entidades entre sí a través de estas nuevas variables:

Tabla 6.10.1. Tasas de desempleo y proporciones de población bajo la línea de pobreza (en %). Aglomerados urbanos seleccionados.

Aglomerado	Desempleo	Pobreza
Bahía Blanca	14,6	40,1
Comodoro Rivadavia	11,2	30,4
Concordia	14,9	73,4
Corrientes	13,7	73,5
Gran Catamarca	16,2	63,1
Gran Córdoba	16,5	54,7
Gran La Plata	12,9	41,9
Gran Mendoza	9,3	56,5
Total urbano	15,6	54,8

Fuente: elaboración propia en base a EPH-INDEC (mayo de 2003)

Podríamos, por ejemplo, señalar que pese a que las tasas de desempleo de Concordia y Corrientes son menores a las que exhiben Catamarca o Córdoba, la proporción de pobres

manual metodologia 22/5/08 15:36 Página 359

es mucho mayor en las dos primeras ciudades. Seguramente, la mayor importancia del empleo industrial en las dos ciudades mencionadas en primer término hace que los ingresos de quienes trabajan sean, en promedio, más altos. En cambio, en las otras dos abunda el trabajo informal, muy mal remunerado. Por eso, aunque haya menos desocupados, el porcentaje de pobres es mayor. Otra conjetura posible sería que los trabajos ofrecidos en Concordia y Corrientes son de tan mala calidad, que la gente ni siquiera los busca y se mantiene al margen del mercado laboral: ello determina que haya menos desempleados. Otras ciudades como Bahía Blanca y La Plata o Comodoro Rivadavia muestran situaciones comparativamente mejores en ambos aspectos: menos desempleo y menos pobreza. En especial, la ciudad patagónica.

6.11. Cuadros que muestran evolución: el tiempo como variable

En ciertas ocasiones, algunas variables han sido medidas repetidamente a lo largo del tiempo: ya hemos hablado, en un capítulo anterior, de los diseños de investigación longitudinales o diacrónicos. Supongamos que, en vez de medir y comparar la tasa de desempleo o el porcentaje de población pobre entre diferentes países o entre diferentes ciudades (como se lo hizo en el cuadro 6.10.1) se hubieran medido estas variables para una sola ciudad, provincia o país en distintas oportunidades a lo largo de cierto período: por ejemplo, una vez cada año. En vez de tener varias ciudades cuyo desempleo comparamos entre sí (La Plata, Bahía Blanca, Catamarca, etc.) tendremos varias *Catamarcas* desplegadas en el tiempo. Cada una de ellas generará una observación en respuesta a la variable medida, como si fuese una unidad de análisis diferente. Eso es lo que muestra la tabla que sigue (6.11.1).

Tabla 6.11.1. Evolución del desempleo 1995 – 2003 (en %)
Gran Catamarca

1995	1996	1997	1998	1999	2000	2001	2002	2003
12,4	16,5	14,8	11,4	10,7	19,6	22,3	25,5	16,2

Fuente: EPH-INDEC (ondas de mayo de cada año)

Si miramos el cuadro podremos ver que la tasa de desocupación, tras algunas oscilaciones, siguió un curso ascendente en Catamarca a partir de 1999. En 2002 más que duplicaba la observada al comienzo de la serie. En 2003 muestra una importante mejoría: al pasar de 25,5% a 16,2% retrocede más de nueve puntos porcentuales.

6.12. Interpretando el cuadro: no se leen números sino que se habla de sujetos o actores sociales.

Como se ha visto, elaboramos cuadros que cruzan variables con el propósito de ver si

manual metodologia 22/5/08 15:36 Página 360

algunas presunciones que abrigamos son ciertas o, por el contrario, son erróneas. Los cuadros contienen números: sin embargo, la interpretación que hacemos de dichos cuadros, aunque basada en el examen de esos números, nos lleva más allá de ellos: nos conduce a hacer afirmaciones respecto de entidades reales que están por detrás de las cifras. Los números no hacen sino traducir de un modo peculiar ciertas propiedades y relaciones propias de estas entidades reales. Por eso, la interpretación de un cuadro no consiste en leer y repetir cifras. Consiste, justamente, en afirmar cosas respecto de las entidades a las que se refieren los cuadros. Estas afirmaciones, claro está, no son arbitrarias, sino que se sustentan en las cifras.

Ha de tenerse, pues, en cuenta que:

Detrás de los números hay gente, países, riqueza –mejor o peor repartida–, pobreza, trabajo, desempleo, etc.

Raramente elegimos en forma azarosa examinar un dato u otro, ni cruzamos las variables por capricho. Algún supuesto tenemos cuando decidimos examinar la distribución de ciertas variables y cruzarlas con otras. Si así fuera, al mirar los datos hemos de preguntarnos: ¿era cierto lo que suponíamos?

Y aún cuando no hubiéramos abrigado presunciones: ¿qué nos sugiere lo que vemos? ¿Era esperable o inesperado?

Tanto en el caso en que vemos refutadas nuestras hipótesis como cuando nos encontramos con un dato inesperado, hemos de preguntarnos por qué y tratar de imaginar una respuesta.

Esta respuesta será, necesariamente, una nueva suposición, que a su vez nos conducirá a examinar nuevos datos para confirmarla. Tal la secuencia que sigue el análisis, en procura de construir conocimiento empíricamente sustentado (Errandonea, 2003).

Relacionar siempre; *suponer* siempre; *interrogar* siempre: así se avanza en el análisis.

7. Gráficos en lugar de cuadros: otro modo de presentar la información

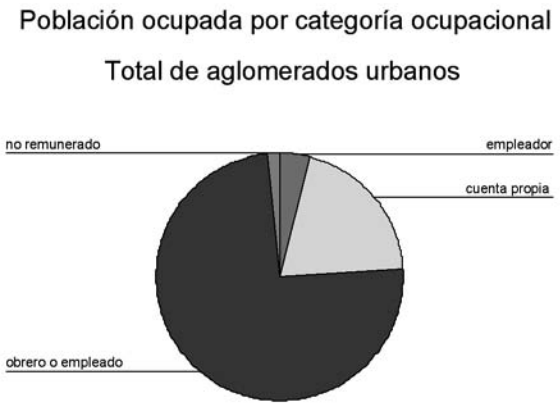
Gran parte de la información que puede presentarse en las tablas –tanto las que muestran la distribución de una sola variable como las que cruzan variables– se puede presentar en forma gráfica. Los gráficos, insertos en un texto como sustituto eventual de las tablas o cuadros, suelen hacerlo menos monótono. Además, para las personas menos habituadas a vérselas con números, suelen ser más fáciles de interpretar. Por fin, hay cierto tipo de datos –por ejemplo los que corresponden a series temporales– donde los gráficos pueden suplir con ventaja a las tablas, brindando una visión más sintética.

En los gráficos rigen las mismas normas que en las tablas en lo que atañe a la titulación, fuentes y notas. Al igual que en la numeración correlativa, bien a lo largo de todo el texto o al interior de cada capítulo.

manual metodologia 22/5/08 15:36 Página 361

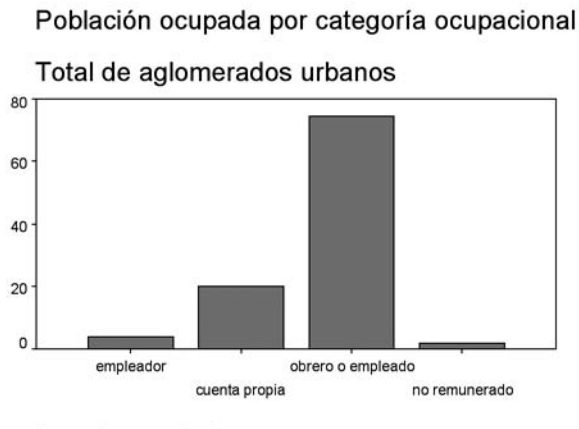
A continuación se brindan algunos ejemplos de representación gráfica de los datos.

7.1. El caso de una sola variable categórica: gráficas de sector circular o tortas



En este primer caso el gráfico muestra la distribución de una sola variable categórica: se trata de la distribución de la población ocupada según su categoría ocupacional. En este tipo de gráfico la superficie total del círculo representa el total de la población, en tanto que cada sector circular muestra una parte de ella. Se advierte sin dificultad que los asalariados sobrepasan el 70% del total, mientras que le siguen los trabajadores por cuenta propia, que son algo menos de una cuarta parte.

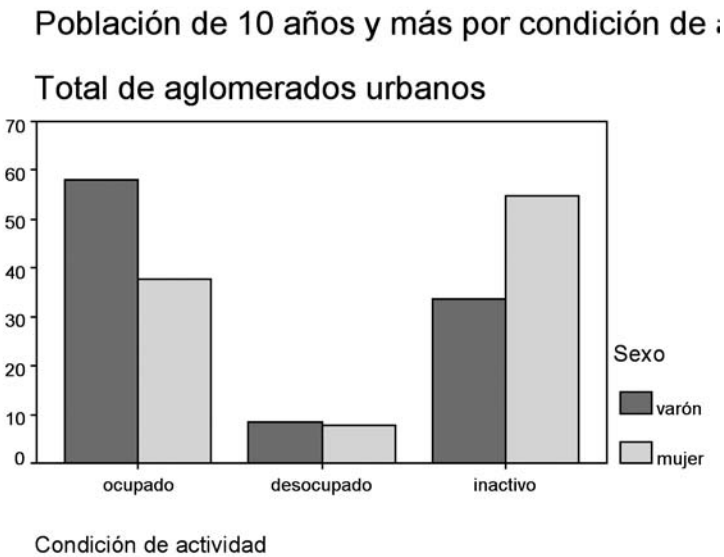
Un modo alternativo de representar una variable categórica –en este caso la misma– es mediante barras simples situadas sobre un par de ejes. En el eje horizontal –denominado eje de categorías– tenemos precisamente las categorías de la variable, en tanto que el eje vertical o de frecuencias contiene a estas últimas, ya expresadas en porcentajes como es el caso en este ejemplo, o bien en absolutos.



manual metodologia 22/5/08 15:36 Página 362

7.2. El caso de dos variables categóricas: barras agrupadas o apiladas

Puede darse el caso de que se quiera representar la distribución de una variable categórica –por ejemplo la condición de actividad: ocupados, desocupados e inactivos– pero desagregada por una segunda variable, también de categorías, como lo es el sexo. Es decir, algo similar a lo que contendría una tabla de contingencia. En este caso tenemos dos alternativas de representación gráfica. La primera son las barras agrupadas. Y su lógica es semejante a la de las barras simples: tendremos frecuencias –más habitualmente relativas o porcentuales– en el eje vertical y categorías en el horizontal. En este la condición de actividad se representa para cada sexo por separado. Se aprecia que entre los varones, la proporción de ocupados se aproxima al 60% mientras que la de inactivos es algo más de un tercio. Entre las mujeres, a la inversa, la inactividad llega al 55% en tanto que la proporción de ocupadas no alcanza a 40%.

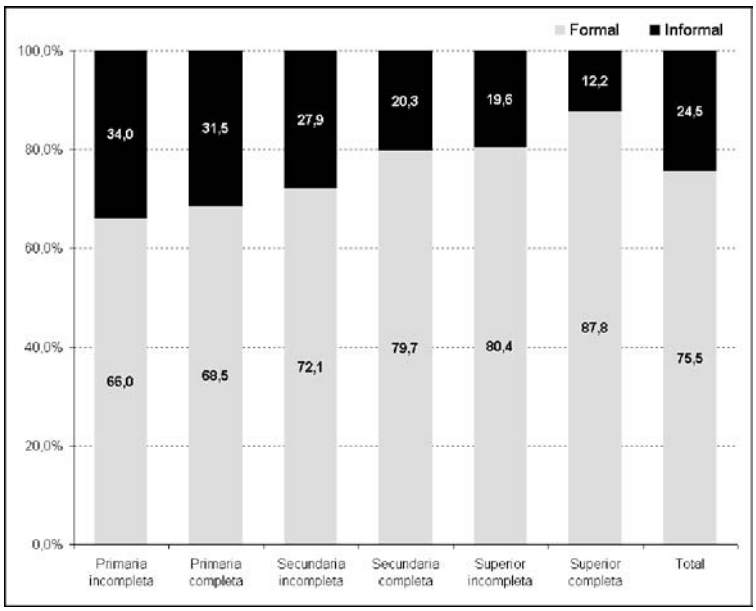


Fuente: EPH-INDEC (1° semestre de 2004)

En este segundo gráfico, en cambio, se muestra la distribución de los trabajadores que se desempeñan en el sector formal e informal, pero para cada nivel educativo por separado. El eje horizontal de la gráfica contiene las categorías de la variable de corte (el nivel educativo). El eje vertical es el eje de frecuencias, porque muestra las frecuencias relativas o porcentuales. En este caso, cada barra representa el 100% de los ocupados correspondientes a cada nivel educativo. Y dentro de cada uno de estos grupos podemos apreciar qué proporción corresponde a trabajadores del sector formal e informal.

manual metodologia 22/5/08 15:36 Página 363

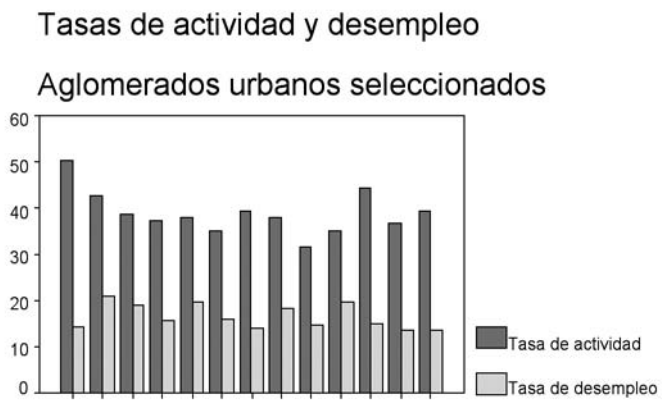
Población ocupada por sector de inserción laboral según nivel educativo



Fuente: EPH-INDEC (onda mayo de 2003)

7.3. Pocas unidades y una variable cuantitativa. Barras

En principio, este gráfico puede parecer semejante al anterior. Sin embargo, el contenido de los ejes ha cambiado. En el eje vertical, donde antes teníamos frecuencias, ahora tenemos valores de variables cuantitativas: la tasa de actividad y la tasa de desempleo. Y ambas variables están medidas de casos que aparecen situados en el eje horizontal: algunas ciudades de la Argentina.

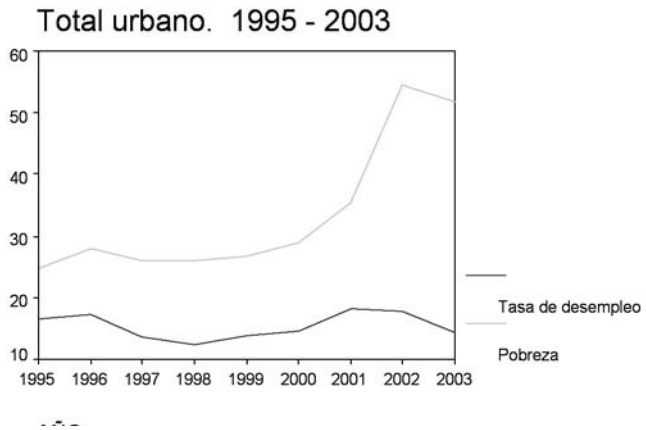


manual metodologia 22/5/08 15:36 Página 364

7.4. El caso de las series temporales: líneas

Por fin, este gráfico representa los valores de una variable cuantitativa que ha sido medida a lo largo del tiempo. Sobre el eje horizontal se han representado los años o puntos de medición, en tanto que el vertical contiene los valores de la variable, que son la tasa de desempleo y la incidencia de la pobreza.

Evolución de la tasa de desempleo y la pobreza



8. La lógica de la descripción

El análisis descriptivo es, por así decirlo, el primer modo de aproximación a los fenómenos que estudia la ciencia[30]. La descripción es el propósito más módico y elemental, aunque no por eso el menos importante: no es posible pretender explicar o predecir fenómenos que aun no han sido adecuadamente descriptos.

La descripción, cabe decirlo una vez más, no es ingenua ni teóricamente neutral: como no puede darse cuenta de la infinita complejidad de la realidad, fuerza es abstraer ciertos aspectos y centrarnos en ellos, desdeñando multitud de otras cosas. Cuáles aspectos serán privilegiados y qué otros dejados de lado es una decisión teórica: en muchas ocasiones no habrá un total consenso en torno a esto. La objetividad de la descripción no consiste, pues, en reflejar la naturaleza a la manera de un espejo sino, en todo caso, en dar cuenta de qué aspectos se han privilegiado y por qué razones. Y, en segunda instancia, en presentar de un modo claro e interpretable la evidencia seleccionada relativa a estos aspectos.

Efectivamente, la descripción no procede en forma arbitraria, sino que contempla siempre un plan. Y el plan debe, asimismo, guardar ciertos preceptos lógicos:

De lo general a lo particular:

- Primero suele cuantificarse el peso de la subpoblación bajo análisis en relación con la

manual metodologia 22/5/08 15:36 Página 365

población mayor de la que forma parte: por ejemplo, en un estudio sobre los jóvenes de 15 a 29 años, qué proporción del total de la población del país representa ese grupo. Inclusive, sería pertinente a este nivel tan general, comparar el peso que los jóvenes tienen en Argentina con el que registran en otros países de la región o del mundo.

- En una segunda instancia, podríamos interesarnos en la evolución de esta gravitación a lo largo de las últimas décadas: ¿aumenta o disminuye el peso de la población joven sobre la población total? Esta tendencia es similar o diferente a la que se observa en otros países?
- Luego, se aborda la descripción del grupo específico en sus aspectos más generales: la distribución de los jóvenes por tramos de edad, sexo, nivel educativo, etc., seguramente en comparación con los no jóvenes: si no comparamos, no sabemos si el perfil es igual o distinto, si tienen alguna especificidad.
- Luego se pasaría a la descripción de aspectos más específicos: por ejemplo, la situación ocupacional (los que trabajan, buscan trabajo o están inactivos). Y enseguida, para el subgrupo que trabaja, podría abordarse su forma de inserción en el mercado de trabajo (asalariados, autónomos), cuántas horas dedican a la actividad laboral, qué tipos de tareas desempeñan, qué ingresos obtienen. También en este caso, seguramente se compararía con los trabajadores de otras edades, para advertir en qué se diferencian los jóvenes.
- Y luego, al interior de los trabajadores jóvenes, podrían indagarse las diferencias ocupacionales entre varones y mujeres o entre diferentes franjas de edad o de nivel educativo.

De los atributos estructurales o adscriptos a lo adquirido:

Supongamos que tenemos que hacer una descripción general de las condiciones de vida de la población de un país o de una región, para lo cual contamos con datos de una vasta encuesta sobre condiciones de vida[31]. En este caso, comenzaremos por describir las condiciones más estructurales de la población, a través de sus atributos demográficos, para luego ir pasando a aspectos más específicos:

- primero la distribución por sexo y edad
- luego las condiciones de vivienda y servicios (agua, electricidad, gas, cloacas)
- después el acceso a la educación y a la cobertura de salud
- a continuación la relación con el mercado de trabajo
- luego los ingresos obtenidos por los hogares y su distribución
- después, en qué se gastan esos ingresos (los hábitos de consumo)
- posteriormente el uso del tiempo libre
- luego las opiniones políticas, preocupaciones y percepciones

manual metodologia 22/5/08 15:36 Página 366

Evidentemente hay una lógica detrás de esta secuencia: en primer lugar uno se inserta en la sociedad con cierta identidad básica conferida por el sexo y la edad. Luego, ocupa un cierto espacio físico y tiene o no tiene acceso a ciertos servicios de infraestructura urbana, así como a otros servicios básicos, tales como los de salud y educación. Después se inserta en la vida económica a través de un trabajo, que le permite obtener ingresos. Luego, dispone de cierta manera de estos ingresos para atender sus consumos. Posteriormente, hace algún empleo de su tiempo libre. Por fin, sustenta ciertas opiniones y tiene ciertas percepciones acerca del mundo que le rodea. Adviértase que, en este caso, la lógica de la descripción comienza por los aspectos más estructurales y culmina en los más subjetivos.

9. Por fin: ¿será verdad...?

Quien haya mirado con alguna atención los cuadros insertos en este capítulo habrá advertido que, en la generalidad de los casos, se trata de datos que provienen de la Encuesta Permanente de Hogares (EPH) que se lleva a cabo regularmente en las áreas urbanas más importantes de Argentina. Pues bien: se trata de una encuesta por muestreo: en consecuencia, todas las características que apreciamos se refieren a las personas encuestadas y no a la población total. Sin embargo, estamos dispuestos a creer que la distribución de la población total no difiere en demasía: el capítulo sobre el muestreo aleatorio nos ha ofrecido razones para que confiemos en ello.

Ahora bien, no sólo se trata de las distribuciones de algún atributo –como la categoría ocupacional de los ocupados– o del promedio de sus ingresos, sino también de las relaciones entre variables que apreciamos en las tablas de contingencia ¿Será verdad que ciertas variables se asocian también en la población, o podrá tratarse de un resultado casual y engañoso que apreciamos en la muestra? Es de suma importancia para el análisis poder responder a tal pregunta.

La estadística también cuenta con instrumentos capaces de brindar una respuesta. Estos instrumentos no serán expuestos en detalle aquí[32]. Sin embargo, se abordará brevemente la cuestión desde una perspectiva conceptual.

En el punto 6.9 de este capítulo, al tratar el tema de la asociación entre variables y su detección a través de la lectura porcentual de los cuadros, se afirmó que si las variables fuesen independientes, los porcentajes observados en las celdas interiores, al compararlos columna a columna (de una categoría a otra de la variable independiente) no debieran diferir entre sí, como tampoco debieran apartarse de la distribución marginal. El apartamiento de esta última distribución, mediante la elevación de un porcentaje cualquiera, podía interpretarse como un indicio de asociación entre ciertas categorías de las variables involucradas.

manual metodologia 22/5/08 15:36 Página 367

Pues bien, la idea es que, en cualquier cuadro, podrían reconstruirse sin dificultad las frecuencias que se esperaría hallar en las celdas interiores si la distribución porcentual resultara igual a la del marginal. Haciendo este ejercicio, se logra una tabla "gemela" que contiene las frecuencias distribuidas en las celdas tal como debieran estarlo en caso de perfecta independencia de las variables: es decir, con iguales distribuciones de Y para cada categoría de X. Luego, bastaría comparar cada celda de la tabla "gemela" con la de la tabla real: como si superpusiéramos una tabla sobre la otra. Una alta coincidencia entre ambas sugeriría que las variables no están relacionadas, porque la tabla "gemela" ha sido construida deliberadamente para simular tal situación. Por el contrario, las diferencias entre ambas indicarían un apartamiento de la situación de independencia de las variables, que sería mayor cuanto más distintas fueran las distribuciones.

Ese grado de apartamiento puede cuantificarse mediante una fórmula y, asimismo, puede ser comparado con un cierto parámetro para saber si existe o no una alta probabilidad de que la asociación que apreciamos en una muestra no exista en la población de la que fue extraída. Ello forma parte de la estadística inferencial.

Tabla original:			
	X1	X2	Total
Y1	35	5	40
Y2	35	25	60
Total	70	30	100

Tabla "gemela":			
	X1	X2	Total
Y1	28	12	40
Y2	42	18	60
Total	70	30	100

Obsérvese que en la tabla "gemela" se han llenado las celdas interiores –las de las columnas X1 y X2– respetando las mismas proporciones de 40% y 60% de la columna final. En la tabla original, sin embargo, las frecuencias de las celdas eran diferentes: más altas en algunos casos y más bajas en otros. Ello sugeriría, pues, que las variables no son enteramente independientes.

manual metodologia 22/5/08 15:36 Página 368

10. Las tipologías: construcción y substrucción de un espacio de propiedades

Todos los cuadros vistos hasta aquí tienen una lógica subyacente: cruzan variables de tal manera que en las celdas quedan ubicados los casos que comparten ciertas combinaciones de valores o categorías en dichas variables. Es decir, estas celdas albergan frecuencias o cantidades de unidades de análisis que tienen ciertos atributos en común.

Existe otro camino para la tabulación cruzada, que es –al mismo tiempo– una modalidad operacional y una vía de generación de conceptos: un recurso heurístico. Se trata de las tipologías o espacios de propiedades (Barton, 1973) que quedan determinados por el cruce de dos o más variables categóricas. Un ejemplo facilitará la comprensión de esta idea.

La pobreza puede medirse de dos maneras: una de ellas es el llamado *método directo*, empleado en Argentina desde los años ochenta y que apunta a la pobreza estructural: la padecen los hogares con Necesidades Básicas Insatisfechas. Según este criterio, tienen NBI los hogares que:

- cuenten con –al menos– un niño de 6 a 12 años de edad que no concurra a un establecimiento educativo
- ocupen una vivienda inadecuada (una casilla, una pieza de inquilinato, un local no construido para fines habitacionales o una casa sin agua o fabricada con materiales inadecuados)
- ocupen una vivienda que carezca de retrete en el baño o no tenga baño
- se encuentren hacinados, es decir que haya más de tres personas por habitación en la vivienda (se excluyen el baño y la cocina)
- o bien tengan baja capacidad de susbsistencia: un jefe de hogar con escasa educación y cuatro o más personas por cada perceptor de ingresos.

Como puede apreciarse, se trata de un índice sencillo, que clasifica como "pobres estructurales" a los hogares que cumplen al menos una de estas condiciones.

Por otro lado, en la década pasada comenzó a medirse la pobreza por ingresos –el *método indirecto*– , comparando los ingresos monetarios de que disponen los hogares con los que requerirían para adquirir una canasta básica de bienes y servicios previamente establecida en forma normativa. El valor de esta canasta varía según el tamaño y composición de las familias y se denomina línea de pobreza (LP). De acuerdo a este criterio, son pobres los hogares cuyos ingresos no alcanzan a cubrir el valor de la LP (que se va actualizando conforme se modifican los precios).

Ambos variables son dicotómicas (clasifican a los hogares en pobres y no pobres) y pueden cruzarse entre sí:

manual metodologia 22/5/08 15:36 Página 369

Ingresos respecto a la LP	Condición de NBI	
	Con NBI	Sin NBI
Inferiores	*Estructurales*	*Empobrecidos o declinantes*
Superiores	*Inerciales o emergentes*	*No pobres*

El cruce de ambas variables da lugar a cuatro casilleros a los que se puede asignar significado teórico y "ponerles nombres". Así, aquellos hogares en los que coinciden ambos atributos serían los pobres estructurales: siempre han sido pobres. En cambio, los que no tienen carencias estructurales pero sí insuficientes ingresos son los que se han empobrecido, probablemente porque se han degradado sus condiciones de empleo: han conocido un pasado de mayor prosperidad. Puede haber hogares que presenten rasgos de pobreza estructural y, sin embargo, hayan aumentado sus ingresos hasta superar la LP: estas son familias que mejoraron su situación, pero conservan atributos de sus pasadas carencias. Finalmente están los hogares que no son pobres por ninguno de los dos criterios. El cruce de las dos variables iniciales ha dado lugar, pues, a la creación de una nueva variable –de cuatro categorías– que permitirá clasificar a los hogares, analizar su distribución y sus características diferenciales. Se trata de una *tipología.*

Las tipologías suelen ser muy útiles, porque al reagrupar los casos con un criterio nuevo muchas veces permiten echar una mirada diferente sobre ellos, descubrir rasgos antes no conocidos, etc.

También es posible proceder al revés, partiendo de conceptos ya existentes y procurando imaginar que dimensiones subyacen a ellos. Esta operación se denomina *substrucción* de un espacio de propiedades (Barton, 1973). Implica una reflexión y una suerte de deconstrucción del concepto: algo así como "mirar qué tiene detrás".

Vamos a intentar un ejercicio para ver qué podría haber detrás de la *precariedad ocupacional*: supone partir de un concepto usado un tanto difusamente, refinarlo y darle contenido empírico, a la vez que ponerlo a prueba como criterio clasificatorio. Por ejemplo, podemos pensar que hay dos atributos detrás de una inserción laboral precaria: la falta de la protección legal que brinda el estar registrado en la seguridad social y la falta de estabilidad en el puesto de trabajo. Podemos cruzar ambos criterios:

	Asalariados	
	Estables *No precarios*	Inestables *Precariedad temporal o coyuntural*
Registrado en la seguridad social		
No registrado en la seguridad social	*Precariedad consolidada*	*Precariedad inestable o típica*

manual metodologia 22/5/08 15:36 Página 370

En principio el ejercicio posibilita distinguir tres situaciones posibles al interior de la precariedad laboral. Antes, apenas si disponíamos del concepto de precariedad, empleado en forma difusa. Nos hemos forzado a definirlo con mayor precisión y, además, hemos creado una variable nueva, de la que antes no disponíamos.

Las tipologías tienen, pues, un valor de elaboración teórica. Por cierto que, dada la relativa falta de consenso que existe en torno a los conceptos de las ciencias sociales, frente a unos mismos conceptos iniciales podrían realizarse diferentes substrucciones. Que, a su vez, darían lugar a distintos desarrollo teóricos y análisis empíricos.

Bibliografía

- BARTON, A. (1973). "El concepto de espacio de propiedades en la investigación social". En *Conceptos y variables en la investigación social.* Buenos Aires: Ediciones Nueva Visión.
- BLALOCK, H. (1986), *Estadística Social.* México: Fondo de Cultura Económica.
- ERRANDONEA, A. (2003). "Algunas reflexiones en defensa de la construcción empírica del conocimiento sociológico". En Lago Martínez, S., Gómez Rojas, G. y Mauro, M. *En torno de las metodologías: abordajes cualitativos y cuantitativos.* Buenos Aires: Editorial Proa XXI.
- GALTUNG, Johan (1978). *Teoría y Método de la Investigación Social.* Tomo I. Buenos Aires: EUDEBA.
- GARCIA FERRANDO, M. (1985), *Socioestadística.* Madrid: Alianza Universidad.
- MORA y ARAUJO, M. (1965). "Recomendaciones para la lectura y análisis de cuadros". Ficha N° 514. Elementos de Metodología y Técnicas de la Investigación Social. Servicio de Documentación de Sociología. Facultad de Filosofía y Letras. Universidad de Buenos Aires.
- PREALC/OIT (1978). *El sector informal: funcionamiento y políticas.* Santiago de Chile: PREALC/OIT.
- ZEIZEL, H. (1962. *Dígalo con números.* México: Fondo de Cultura Económica.

Notas

1 Empleamos estas expresiones en forma indistinta.

2 Cabe señalar que la EPH es una encuesta por muestreo que se realiza a varios miles de personas. Sin embargo, la base cuenta con un factor de expansión que permite hacer una extrapolación al total de la población de las áreas relevadas: es por eso que en el cuadro aparecen 10,2 millones de ocupados, aunque no se haya entrevistado a tal cantidad de personas.

3 Como por ejemplo el SPSS.

manual metodologia 22/5/08 15:36 Página 371

4 Sobre este punto se proporcionarán más detalles en el capítulo correspondiente a muestreo.
5 Nótese que en la tabla 3, al calcular la proporción con sólo dos decimales, se produce un redondeo: entre los inactivos, si agregáramos un tercer dígito luego de la coma tendríamos 0,448, exactamente equivalente a 44,8%.
6 Puesto que la cantidad de habitantes va cambiando a lo largo del año, se toma la que se estima que había al promediar el mismo.
7 Idem nota anterior.
8 Se trata de una razón y no de una proporción, porque el numerador no es una parte enteramente contenida en el denominador: los niños que fallecen en el curso de cierto año pueden haber nacido el año anterior. Todos quedan incluidos, eso sí, en un conjunto mayor: los niños que no sobrepasan el año de edad.
10 De hecho, se ha categorizado una variable cuantitativa, convirtiéndola en categórica.
11 Aunque este, eventualmente, podría ser una cifra mucho más elevada: por ejemplo 25 mil o más.
12 Este problema se acentúa cuando se usan datos provenientes de muestras, porque entonces las estimaciones se tornan inseguras. Abordaremos este problema, en forma conceptual, en un capítulo posterior.
13 Este era el ordenamiento del sistema educativo vigente en gran parte del territorio de la Argentina desde la década del noventa, a partir de la Ley Federal de Educación. Algunas jurisdicciones no habían adherido a ella, pero por lo que los niveles de educación diferían.
14 Se la puede hallar en cualquier manual de estadística y, por otra parte, el cada vez más amplio acceso a datos desagregados que pueden ser tratados con programas de computación, hace que su utilización sea infrecuente.
15 Cuando se trabaja con datos muestrales se emplea X para indicar la media muestral, en tanto que se usa la letra griega Mu para la media poblacional.
16 Los viejos profesores de estadística se suelen solazar en repetir el cuento de los dos sujetos que disponían, para alimentarse, de tres pollos. Uno era dueño de tres y el otro de ninguno, Sin embargo, el promedio indicaría que poseían un pollo y medio cada uno. También mencionan la historia de un hombre que murió ahogado en un río cuya profundidad era de 50 centímetros (en promedio...).
17 En la reiteración de este valor influía decisivamente el Plan Jefas y Jefes de Hogar, que proporcionaba un ingreso de $ 150 y contemplaba, en más de 70% de los casos, una contraprestación laboral.
18 Es costumbre emplear la n minúscula para indicar el número total de casos de la muestra y la N mayúscula cuando se trata del total poblacional.
19 Aunque, cuando se trabaja con muestras, se suele reservar para la varianza poblacional, empleándose una s2 para el caso de la muestra.
20 El cambio de identidad sexual, conducta posible, es de todas formas dificultoso y relativamente poco frecuente.
21 El capital social, como lo ha señalado Bourdieu, contribuye decisivamente a los logros de las personas.
22 Es decir, sólo tienen dos categorías.
23 Las diferencias porcentuales se expresan en *puntos porcentuales* y no en *porcentajes*. Hemos de decir que la tasa ocupación de los varones supera a la de las mujeres en 15 puntos porcentuales, no en 15 por ciento.
24 Debe destacarse, sin embargo, que en los cuadros de dos columnas, la suma algebraica de las diferencias porcentuales es igual a cero.
25 En este caso, sería la que podríamos obtener en la última fila, para los inactivos, entre las personas de 60 y más años y las de 25 a 59, que alcanza un valor de 54.
26 Se trata de las llamadas medidas de contingencia para variables nominales y ordinales, y pueden consultarse en cualquier texto de estadística. Por ejemplo: Blalock, H. (1986), o bien: Garcia Ferrando, M. (1985).
27 Se trata de la prueba de Chi Cuadrado, que también puede consultarse en los textos mencionados en la nota anterior.
28 Por supuesto que esto sucede así toda vez que, al construir la tabla, tomemos la precaución de ordenar los valores de las variables en forma convergente: alto, medio y bajo en ambas.
29 Los cuadros con exceso de cifras pueden volverse confusos y difíciles de interpretar, sobre todo para personas no muy habituadas a lidiar con ellos.

manual metodologia 22/5/08 15:36 Página 372

30 Salvo en los casos en que se trata de ámbitos tan vírgenes y desconocidos, que requieren una primera aproximación exploratoria, tendiente a identificar algunos rasgos o aspectos relevantes a fin de describirlos luego.
31 Por ejemplo, en la Argentina se llevó a cabo en 1997 y en 2001 la Encuesta de Condiciones de Vida y Acceso a Programas Sociales del SIEMPRO (Sistema de Información, Monitoreo y Evaluación de Programas Sociales). Y el Gobierno de la Ciudad de Buenos Aires viene realizando desde 2002 una Encuesta Anual de Hogares.
32 Se trata de pruebas de estadística inferencial o test de hipótesis, tales como la prueba de Chi Cuadrado –entre otras– que pueden verse en cualquier manual de estadística.