CAPÍTULO 6

CONCEPTOS DE OBJETO Y DE UNIDAD DE ANÁLISIS. POBLACIÓN Y MUESTRA

6.1. Objeto y unidad de análisis

En el apartado 1.3 se subrayó la importancia que la herramienta intelectual y práctica de la matriz tiene en el "método de la asociación", es decir, en la versión del enfoque estándar que ha dominado en las ciencias sociales.

En estas ciencias se usan varios tipos de matrices, pero el tipo dominante en la fase de la recolección es la llamada "MATRIZ DE DATOS", o matriz "casos por variables". Se decía en el apartado 1.3 que la matriz no es nada más que un cruce entre un haz de vectores paralelos horizontales y un haz de vectores paralelos verticales. En el tipo llamado "matriz de datos", los vectores horizontales se refieren a objetos y los vectores verticales a propiedades de estos objetos.

El término 'OBJETO' se entiende en un sentido gnoseológico, como posible objeto del pensamiento (cualquier cosa en la que se piense). Por lo tanto, los objetos en las filas de una matriz de datos pueden ser individuos (humanos o no), familias, grupos, instituciones, provincias, Estados, eventos, etcétera. Pero, en una matriz dada, todo los objetos en las filas deben ser del mismo tipo: en efecto, no se podría construir una matriz que llevase simultáneamente, por ejemplo, seres humanos y Estados en sus filas, porque los vectores relativos no podrían ser paralelos: es decir, tener referentes del mismo tipo y la misma secuencia de propiedades en las columnas: las propiedades que se pueden referir a individuos no se pueden referir a Estados, y viceversa.

El tipo de objetos que están en las filas determina el tipo de propiedades que pueden estar en las columnas. La propiedad "sexo" no se puede referir a una institución, una provincia o un Estado, como la propiedad "número de ciudadanos adultos" no se puede referir a un individuo.

El tipo de objeto acerca del cual se buscan informaciones en una investigación se llama "UNIDAD DE ANÁLISIS". Esta expresión tiene un referente abstracto, puede ser "ama de casa argentina adulta", pero no puede ser "la señora Ramírez". Las unidades más frecuentemente usadas en la investigación social son el individuo, la familia, el grupo, la empresa, el distrito electoral, el municipio, la provincia, el Estado. También pueden ser unidades eventos, como la elección, la guerra, etcétera.

En una investigación se debe definir no sólo la unidad, sino también el ámbito espacio-temporal que interesa. La necesidad de delimitar el ámbito espacial es obvia: no es difícil darse cuenta de que es cosa diferente estudiar las amas de casa de una ciudad, de una provincia, de un Estado o de un continente. Menos obvia es la necesidad de delimitar el ámbito temporal: como en sociología y ciencia política el sondeo es una herramienta privilegiada de recolección de informaciones, habitualmente se da por sentado que el ámbito temporal sea un genérico presente.[1] Una falta seria de la investigación social es, en efecto, desconocer la perspectiva diacrónica y no aprovechar ni siquiera las oportunidades de analizar los sondeos pasados que se guardan en los archivos de datos.[2]

6.2. Población y muestra

Una vez determinadas la unidad de análisis y el ámbito espacio-temporal, el conjunto de los ejemplares de esa unidad que se encuentran en dicho ámbito es llamado POBLACIÓN.[3] Cada ejemplar de esta población puede devenir un CASO, es decir, el referente de una fila de la matriz.

Cuando la unidad es una provincia y el ámbito un Estado dado en un período dado, o la unidad es un Estado y el ámbito un continente dado en un período dado, la población no es numerosa, y habitualmente se recolectan informaciones acerca de todos sus miembros (es decir, todos los ejemplares de esta unidad dentro del ámbito espacio-temporal). Este procedimiento se llama ENUMERACIÓN COMPLETA.

Pero cuando los miembros de la población son muchos (como en las encuestas sobre individuos adultos de una nación) recoger informaciones sobre todos cuesta un gran esfuerzo e inversión de tiempo y recursos, y sólo se hace raramente, por agencias oficiales del Estado, y para fines que trascienden la investigación social (piénsese en un censo poblacional).

Descontando a esos casos, se presenta el problema de elegir un pequeño subconjunto de estos miembros de la población para investigarlos con un menor gasto de recursos, convirtiéndolos en casos de una matriz de datos. Este

[1] La responsabilidad de esta falta de profundidad diacrónica no cae únicamente sobre el sondeo, porque también los psicólogos —que prefieren usar otras herramientas de recolección además del sondeo— tienen una orientación marcadamente sincrónica.

[2] En los Estados Unidos y Europa del Norte hay muchos archivos de este tipo, donde los investigadores depositan los datos resultantes de sus encuestas para permitir a otros investigadores practicar "análisis secundario" de sus datos. En la Argentina, el INDEC proporciona al público encuestas como la EPH (Encuesta Permanente de Hogares), y muchos estudiantes e investigadores la usan en sus trabajos. Más allá de este ejemplo, sin embargo, en la Argentina —como en muchos otros países— no hay sólidas tradiciones de análisis secundario.

[3] Nótese que en el lenguaje técnico de las ciencias sociales este término tiene un sentido más amplio del que posee en el discurso ordinario. Este sentido fue desarrollado por Malthus (1747) y tiene una gran importancia, no sólo en la estadística sino también en la teoría de la evolución de Darwin.

problema se aborda con una herramienta clásica de las ciencias sociales: el muestreo.

Una MUESTRA es *cualquier* subconjunto, amplísimo o limitadísimo, de miembros de una población que se investiga con el fin de extender a toda la población las conclusiones resultantes del análisis de las informaciones relativas al subconjunto.[4] Esta extrapolación (de los resultados del análisis) de la muestra a la población entera es llamada INFERENCIA ESTADÍSTICA, y tiene reglas precisas que veremos pronto.

Antes cabe resaltar tres malas costumbres muy difundidas en las ciencias sociales actuales. La primera es una costumbre terminológica, y consiste en el hecho de que a menudo se habla de inferencia de la muestra *al universo*. Este uso es impropio, porque el universo es por supuesto infinito mientras que cualquier población sólo puede ser finita. Como pasa a menudo, el uso terminológico impropio no acontece por casualidad, sino porque permite extender a las encuestas de las ciencias sociales fórmulas matemáticas asentadas en supuestos que sólo son legítimos para conjuntos infinitos, es decir universos.

La segunda mala costumbre consiste en el hábito de extender la inferencia más allá de la población de la cual se extrajo la muestra. Un caso clamoroso en la literatura de las ciencias sociales es la renombrada "relación Kinsey". Kinsey era un psicólogo norteamericano que tenía un consultorio privado en una ciudad de Minnesota. Basándose en las declaraciones de sus pacientes (que pueden considerarse una muestra no aleatoria[5] de habitantes de aquella ciudad) él publicó con sus colaboradores dos volúmenes titulados *Conducta sexual del varón humano* (1948) y *Conducta sexual de la hembra humana* (1953). Los sociólogos no están para nada exentos de esta mala costumbre: por ejemplo, el clásico *The People's Choice* (El pueblo elige) está basado en una muestra de 600 electores del condado de Erie en Ohio (Lazarsfeld *et al.* 1944).

6.3. Muestras aleatorias y no aleatorias

La tercera mala costumbre necesita una discusión más profunda. Todas las empresas que producen y venden sondeos —y no sólo ellas— declaran generalmente que sus muestras son "aleatorias y representativas", estos dos términos se usan ritualmente, sin corresponder a ninguna calidad precisa de las muestras mismas ni de los procedimientos que se utilizaron para su extracción. En los pasajes siguientes se dará un significado a estos términos.

Una muestra se dice aleatoria cuando todos los miembros de la población de la que se extrae tienen la misma probabilidad de ser extraídos y entrar en la

[4] En este capítulo se habla del muestreo que se realiza en el marco de investigaciones de tipo estándar (véanse los apartados 2.3 y 2.4). Otras formas de muestreo serán presentadas en el apartado 12.5.

[5] Se aclararán pronto los criterios según los cuales una muestra puede considerarse aleatoria.

muestra. Si se sale a la calle Y de la ciudad X y se entrevista a los primeros cien sujetos que pasan, ésta no es una muestra aleatoria de habitantes de la ciudad X, porque muchos de ellos no pasan jamás por la calle Y, otros pasan raramente y otros diariamente. Por consiguiente, las probabilidades de ser extraídos no son iguales. Para construir una muestra aleatoria de esta población se debe elaborar un catálogo completo de sus miembros y extraer algunos de ellos con una tabla de números aleatorios o con otro procedimiento que garantice efectivamente la misma probabilidad de ser extraído a cada miembro de la población.

Por lo tanto, la naturaleza aleatoria de una muestra depende integralmente del procedimiento de extracción y no tiene nada que ver con su resultado: si se extraen 30 bolillas negras y ninguna blanca de una bolsa, sin mirar de reojo dentro de la bolsa y sin hacer las bolillas distinguibles al tacto, la muestra es perfectamente aleatoria a pesar del resultado, y cualquiera sea la distribución de los colores en las bolillas de la bolsa.

Algunos estadísticos hablan de "muestra aleatoria simple" cuando los miembros de una población tienen la misma probabilidad de ser extraídos, y de muestra aleatoria sin más cuando cada miembro de la población tiene una probabilidad conocida y no nula de ser extraído. Pero esta distinción ofrece una cobertura científica a procedimientos que tienen poco que ver con la ciencia y mucho con los presupuestos de las empresas comerciales de sondeos. Imaginemos una investigación de ámbito nacional. Si la muestra fuese extraída con procedimiento aleatorio, podrían entrar en ella muchos habitantes de áreas remotas de alta montaña o de otros lugares de difícil acceso. Los entrevistadores de las empresas deberían perder tiempo y gastar dinero para contactar a estos sujetos, y para cada una de estas entrevistas la empresa gastaría un múltiplo del valor medio de las entrevistas en las ciudades principales, donde seguramente habitan entrevistadores de su red nacional.

Para reducir esos gastos, las empresas dividen el territorio nacional en áreas, asignando a cada una un número dado de entrevistas. A las áreas remotas se les asigna un número mínimo de entrevistas, de manera tal que todos sus habitantes tengan una probabilidad extremadamente baja, pero conocida y no nula, de entrar en la muestra —y así la ortodoxia estadística es preservada.

Después de ahorrar dinero con esos procedimientos, las empresas tratan de remediar su obvia consecuencia (subrepresentación de todas las áreas periféricas del país) con otro procedimiento discutible: la ponderación. Supongamos que a un área de alta montaña le fueran asignadas 3 entrevistas en lugar de las 27 que le corresponderían según su proporción sobre la población nacional. Debido a que el cociente 27/3 es 9, la ponderación consiste en multiplicar por 9 cada montañés entrevistado, es decir contar sus respuestas 9 veces en todos los análisis que involucren la variable en cuestión.

Se produce así una doble proyección: cada montañés (y cada habitante de áreas remotas) es proyectado en 5, 10, o más de sus clones ficticios. El paso siguiente es proyectar esta mezcla de individuos y de clones a toda la población nacional. En ambos casos se manifiesta un supuesto atomista, en el sentido de que se descuidan no sólo la especificidad de cada individuo, sino también la influencia que el contexto de relaciones sociales en las que cada individuo está inserto ejerce en su personalidad.[6]

Por otro lado, las muestras telefónicas, incluso si han sido extraídas aleatoriamente de una guía de teléfonos, no cumplen ni siquiera con los criterios más tolerantes para ser consideradas aleatorias. Esto se debe al obvio motivo de que todos los miembros de una población que no tienen acceso a un teléfono no tienen ninguna posibilidad de ser incluidos en la muestra. Y por el motivo más sutil de que los miembros de familias numerosas, y/o los que se encuentran raramente en la casa en los horarios en que se acostumbra llamar, tienen *a priori* una probabilidad reducida, y se desconoce el grado de esta reducción para cada categoría.

Hay otros problemas vinculados con el concepto de extracción aleatoria. Un procedimiento que garantice a cada miembro de una población la misma probabilidad de ser extraído es una condición necesaria y suficiente para lograr una muestra aleatoria de dicha población cuando sus miembros son objetos inanimados, que no pueden rechazar ser extraídos y ser examinados. Pero —como todos los entrevistadores saben bien— las poblaciones de seres humanos no cumplen con estos requisitos. A menudo no se encuentran en su casa, ni contestan el teléfono, muchos de los que se encuentran se niegan a la entrevista porque sospechan objetivos comerciales o fiscales, o porque no desean dedicar tiempo a una actividad de la cual no entienden el espíritu ni las finalidades.

Si las probabilidades de no encontrarse en casa o de negarse a la entrevista fuesen igualmente distribuidas en las varias capas de una población humana, estos inconvenientes sólo reducirían las dimensiones de una muestra sin perjudicar su naturaleza aleatoria. Pero casi un siglo de experiencia con los sondeos ha mostrado que los jóvenes y los adultos ocupados tienen mucho menos probabilidades de encontrarse en casa que las amas de casa y los jubilados. Además, adultos empleados y viejos muestran una mayor propensión a negarse a la entrevista. Por lo tanto, estas categorías tienen *a priori* una menor probabilidad de ser efectivamente entrevistadas.

Por consiguiente, aun una muestra perfectamente aleatoria al momento de la extracción se vuelve casi siempre no aleatoria cuando se trata de transformar cada sujeto en un caso de la matriz: es sabido que las amas de casa tienen una mayor probabilidad de ocupar filas en una matriz de datos, seguidas por los adultos desempleados, los jubilados y los jóvenes —en este orden. Los adultos empleados son los que tienen *a priori* una menor probabilidad de convertirse en casos de una matriz de datos.

[6] Por este motivo, muchos investigadores de una escuela muy sensible a la influencia del contexto de relaciones sociales en las elecciones individuales, como el Bureau of Applied Social Research de la Universidad de Columbia, se han mostrado reacios a generalizar más allá del ámbito de sus específicas investigaciones. Véase Martire (2006: cap. 3).

6.4. ¿Representativo de qué?

El otro término fetiche que cabe examinar es 'representativo'. En los textos estándar se leen definiciones como la siguiente: "[una muestra es] representativa si reproduce —en escala reducida— la población objeto del estudio (para permitir la generalización de los resultados obtenidos en la muestra a la población total)" (Corbetta 2003: 159).

Reproducir en escala reducida un diseño o una hoja escrita es algo que una fotocopiadora hace rápida y fácilmente, reduciendo en la misma proporción en la copia las distancias entre cada pareja de puntos en el original. Pero ¿cómo se puede lograr el mismo resultado con poblaciones de seres humanos? Éstas son caracterizadas no sólo por distancias físicas entre sus miembros, sino por muchas otras propiedades. La analogía con la fotocopiadora no puede ser tan simple, pero es útil. En primer lugar, porque resalta el hecho de que la representatividad se juzga confrontando características del original con su análogo en la copia. Esta comparación se puede hacer sólo cuando la copia fue ya producida. Por consiguiente, mientras para juzgar si una muestra es aleatoria debemos mirar el procedimiento con el que se extrae, independientemente del resultado, para juzgar si una muestra es representativa debemos mirar el resultado, independientemente del procedimiento.

Como se decía, para juzgar si la reproducción en escala reducida de un diseño fue correcta, se confrontan parejas de distancias entre puntos. Pero las poblaciones tienen propiedades más importantes que la distancia entre sus miembros. ¿Cómo se confrontan estas propiedades de la población con las de la muestra para juzgar si la última es representativa?

Para responder tenemos que adelantar el concepto de distribución de una propiedad, que va a ser tratado en el capítulo siguiente. Las propiedades de una población pueden ser constantes (como por ejemplo, el sexo en un convento de monjas) o variar. Varían si diferentes miembros de la población tienen diferentes estados en ellas: continuando en el ejemplo, afuera del convento el sexo varía porque algunos individuos son masculinos y otros son femeninos. En este caso, el sexo tiene una distribución, que se puede expresar en cifras absolutas (5.312 mujeres y 4.893 hombres en el pueblo X) o en porcentajes.

Ya tenemos un resultado de la analogía: si se extrae una muestra de los habitantes de ese pueblo, se entiende que ésta no puede ser considerada representativa de la población (es decir, una reproducción de esta última en escala reducida) si los porcentajes de hombres y mujeres son sensiblemente diferentes de los mismos porcentajes en la población.

Pero pronto se plantean dos interrogantes:

a) ¿Cuán diferente es "sensiblemente" diferente?
b) ¿Qué pasa con las otras propiedades distintas del sexo?

Para el primer interrogante los textos de estadística no proporcionan una respuesta. Y eso no es casual: sería un poco ridículo establecer que si hay una diferencia de un punto entre los mismos porcentajes en la población y en la muestra, la última es representativa, y que si la diferencia excede un punto, la muestra no lo es.[7] Esta consideración nos permite sacar una primera conclusión: mientras que para juzgar si una muestra es aleatoria tenemos una pauta clara (misma probabilidad de entrar en la muestra para todos los miembros de la población), para juzgar la representatividad debemos recurrir a consideraciones difusas y subjetivas. Al responder el segundo interrogante veremos que esta difusión y subjetividad caracterizan todo el campo semántico de la representatividad.

Supongamos que la comparación entre las distribuciones de los sexos en la población y en la muestra nos lleve a la conclusión de que son bastante parecidas y, por lo tanto, la muestra es representativa en lo que concierne al sexo, ¿de eso se puede sacar la conclusión de que la muestra también es representativa en lo que concierne a cualquier otra propiedad de la población?

La respuesta es ¡obviamente, no! Y de ésta se siguen algunas importantes consecuencias:

1) La representatividad tiene que ser controlada y eventualmente afirmada para cada propiedad por separado.
2) Debido a que la representatividad se evalúa comparando la distribución de una propiedad en la muestra con la distribución de la misma propiedad en la población, únicamente se puede evaluar para las propiedades cuya distribución en la población es conocida —es decir, sólo para las propiedades que se relevan con un censo poblacional. Para toda otra propiedad, incluyendo opiniones, actitudes, valores, etcétera, la representatividad no se puede mínimamente controlar, y por lo tanto no puede ser afirmada.
3) La oración "esta muestra es representativa" (sin alguna calificación), que a menudo se escucha, no tiene sentido si no se le agrega "en lo que concierne a la(s) propiedad(es) X (Y, Z)". Y además no tiene ninguna credibilidad si no se le agregan tablas que comparan las distribuciones de esas propiedades en la población y en la muestra. Una práctica que —en rigor de verdad— se encuentra muy raramente cumplida en los relatos de sondeos que afirman que la muestra es representativa.

Ahora que se clarificó el significado propio de los dos términos fetiche, se pueden contestar dos preguntas. Empezamos con la más básica: ¿extraer de forma aleatoria una muestra garantiza automáticamente que ésta sea representativa?

La respuesta es, obviamente, negativa, y la hemos adelantado ya, cuando se dijo (véase más arriba en este capítulo) que se pueden extraer de forma perfec-

[7] Cabe resaltar, sin embargo, que en la estadística inferencial hay un sinnúmero de estos umbrales rígidos para discriminar, por ejemplo, si el promedio de la distribución de una variable en una muestra es "significativamente" diferente del promedio de la distribución de la misma variable en la población.

tamente aleatoria (es decir, sin alterar de modo alguno la igual probabilidad de cada bolilla de ser extraída) 30 bolillas negras y ninguna blanca de una bolsa que contiene 30 bolillas negras y 30 blancas (arriba no se especificó el contenido de la bolsa). La muestra es aleatoria, pero ciertamente no es representativa de la población con respecto de la propiedad "color". Paradójicamente, la única manera de garantizar la extracción de un número igual (cualquiera) de bolillas negras y de bolillas blancas sería mirar en la bolsa cuando se extraen —violando abiertamente los requisitos de una extracción aleatoria.

Si uno sale a la calle y entrevista los primeros 50 varones y las primeras 50 mujeres que encuentra, la muestra será *grosso modo* representativa —respecto del sexo— de la población de la ciudad, de la provincia, del Estado, del continente y del mundo. Pero violará abiertamente los requisitos de una extracción aleatoria.

Estos ejemplos pueden naturalmente ser generalizados —con las necesarias adaptaciones— a cualquier otra población, propiedad y forma de extracción. La pregunta consiguiente no puede sino ser: "Si una extracción aleatoria no garantiza representatividad en alguna propiedad, y ésta puede garantizarse en algunas propiedades únicamente con una extracción que dista de ser aleatoria, ¿por qué sería preferible extraer muestras aleatorias?"

La respuesta es que la extracción aleatoria sólo da una garantía negativa, es decir, la de no introducir sesgos de magnitud y dirección desconocida y no controlable en la distribución de las propiedades en la muestra con respecto de las mismas distribuciones en la población. Si salimos a la calle y entrevistamos los primeros 50 varones y las primeras 50 mujeres, la distribución del sexo en esa muestra será *grosso modo* representativa de la población de la provincia, del Estado, etcétera. Pero las distribuciones de muchas otras propiedades relevantes para una investigación social serán ciertamente sesgadas, y a menudo fuertemente sesgadas (piénsese en la distribución del lugar de residencia, de la profesión, de la edad).[8]

Se dijo más arriba que el azar puede producir distribuciones tan sesgadas como 30 bolillas del mismo color de una bolsa con bolillas equidistribuidas en dos (o más) colores. Pero sesgos tan fuertes son extremadamente raros: la mayoría de las extracciones producirán muestras con sesgos pequeños, o incluso ningún sesgo.

Concluyendo, se puede afirmar que la extracción aleatoria no da ninguna garantía absoluta de que la distribución de una propiedad cualquiera en la muestra sea representativa de la distribución de la misma propiedad en la población. Sólo da:

- Una garantía absoluta de que los sesgos (en el sentido de diferencias entre estas dos distribuciones) no sean introducidos por el investigador, sino producidos por el azar.

[8] Serán inevitablemente subrepresentadas todas las ocupaciones que no permiten pasar mucho tiempo en la calle, y los grupos de edad en los que —por una razón u otra— no se pasa mucho tiempo en la calle: los infantes, los viejos, etcétera.

- Una garantía razonable de que esos sesgos sean de magnitud limitada. Si una muestra de alcance nacional es extraída de forma aleatoria, hay una probabilidad prácticamente nula de que todos sus integrantes circulen por el mismo barrio, a diferencia de la muestra que resultaría de salir a la calle a entrevistar a los que pasan con la única preocupación de garantizar la representatividad con respecto al sexo.

6.5. Cómo garantizar una (limitada) representatividad de una muestra aleatoria

¿No hay alguna manera de asegurar algo más que esta garantía negativa con una muestra aleatoria?

Sí la hay, pero con respecto a un número limitado de propiedades, bajo algunas condiciones (que veremos pronto), y sólo con un tipo particular de extracción (denominada SISTEMÁTICA) que vamos a describir.

Supongamos que se tenga un catálogo exhaustivo de los miembros de una población, listado con un orden que no tenga alguna relación con las propiedades que interesan al investigador (por ejemplo, un orden alfabético de apellidos). Supongamos que esta población tiene 80.000 miembros (por ejemplo, los habitantes adultos de una ciudad media) y que queremos extraer una muestra de 400 individuos. Para lograr una muestra aleatoria sistemática se debe:

1) dividir el listado en 400 segmentos, cada cual comprendiendo 200 miembros, número que resulta de dividir el tamaño de la población por el tamaño de la muestra (80.000/400 = 200) y que se denomina "intervalo de muestreo";
2) extraer de forma aleatoria un número de 1 a 200: supongamos que sea el 78.

La muestra sistemática será formada por el 78° individuo de cada segmento, es decir por los individuos que llevan los números 78, 278, 478, y así sucesivamente hasta el número 79.878.

Veamos ahora cómo se puede lograr que esta particular muestra del ejemplo sea no sólo aleatoria sino también representativa de la población de la ciudad con respecto a dos propiedades: el sexo de los habitantes y su barrio de residencia (supongamos que los barrios sean 5, con un promedio de 16.000 habitantes adultos). La condición necesaria y suficiente para conseguir este objetivo es ordenar el listado por barrio, y dentro de cada barrio dividirlo por sexo. Es decir, empezar el listado con todos los varones (adultos) del barrio A, seguidos por todas las mujeres del barrio A, por todos los varones del barrio B, por todas la mujeres del barrio B, etcétera.[9]

Imaginemos que los varones del barrio A sean 9.000. Extrayendo el individuo que lleva el número 78 (o cualquier otro número) en cada segmento de 200

[9] Ese método fue propuesto por vez primera por uno de los autores (Marradi 1997).

individuos, vamos a extraer en la muestra 45 varones del barrio A, es decir el 11,2% de los individuos de la muestra. Esto es exactamente el porcentaje de varones que viven en el barrio A sobre todos los adultos de la ciudad. Se puede fácilmente comprobar que el mismo mecanismo funciona para todos los segmentos en que hemos dividido la población. El resultado es que tenemos una muestra extraída de forma aleatoria que es representativa de la población de la ciudad con respecto al sexo y al barrio de residencia.

Naturalmente, no se puede decir nada de su representatividad con respecto a todas las otras propiedades —es decir, infinitas menos dos. Lo que podemos hacer, si queremos, es segmentar la población teniendo en cuenta también una tercera propiedad (por ejemplo, el nivel de instrucción), e incluso una cuarta (por ejemplo, el grupo de edad). Pero cada vez que se considera una nueva propiedad el número de segmentos a ordenar en secuencia crece en proporción geométrica: si con el sexo y cinco barrios de residencia los segmentos eran 10, considerando cuatro niveles de instrucción ellos serán 40, y considerando también seis grupos de edad, serán 240. Con una población de 80.000 individuos, el tamaño promedio de cada subgrupo (por ejemplo, varones del barrio A con instrucción baja y menores de 25 años) va a ser de 333 individuos. Reduciéndose la diferencia entre el número de individuos que hay en cada segmento (en el ejemplo, 200) y el número de individuos en cada subgrupo, se acrecienta la probabilidad de que la extracción aleatoria sistemática atribuya a un subgrupo un porcentaje de extraídos no exactamente correspondiente a su porcentaje en la población (aun si las diferencias no pueden matemáticamente exceder una unidad en cifras absolutas).

El principal inconveniente de este método es, por lo tanto, la complicación procedimental vinculada con el manejo de subgrupos formados por el producto lógico de categorías en diferentes propiedades. Pero en el caso de que se atribuya gran importancia a la representatividad de la muestra con respecto a pocas propiedades, el método que se ilustró es la manera más simple[10] de conseguirla sin violar la naturaleza aleatoria de la muestra.

[10] Una forma alternativa es el así llamado "muestreo estratificado". Se definen subestratos con un criterio cualquiera y luego se sortean submuestras dentro de cada subestrato. Por otro lado, a menudo se usa un muestreo estratificado "no proporcional" para sobrerrepresentar o subrepresentar un segmento particular de la población estudiada.

CAPÍTULO 7

CONCEPTOS DE PROPIEDADES. VARIABLES, FIDELIDAD Y FIABILIDAD

7.1. La definición operativa

En las columnas de una matriz de datos se encuentran las variables. Una variable es un vector de signos que representan los estados de los casos en las propiedades que interesan. Habitualmente, pero no necesariamente, estos signos son números.[1]

A veces la relación entre un estado en una propiedad y el signo que lo representa es directa e intuitiva: si la entrevistada es Morena Ruiz y tiene 18 años, tal estado en la propiedad "edad" será representado por el mismo 18 en la celda ubicada en el cruce entre el vector-fila relativo a Morena Ruiz y el vector-columna representando la edad. Para saber que este 18 es el número de años de Morena Ruiz, tenemos que saber cuál es el titular del vector-fila y qué propiedad se representa en el vector-columna. Sólo si tenemos estas informaciones, este 18 se convierte de un mero número en un dato —y lo mismo para todos los otros números en la matriz. Pues sólo se puede hablar de una matriz de datos si todos los números (y, más generalmente, los signos) que se ven son interpretados, es decir, son DATOS.

Continuando con Morena Ruiz, es probable que estemos interesados no sólo en su edad, sino también en su título de estudio (en el caso: licenciada). Pero no hay cifras que representen directamente este título (ni los otros títulos).

Por lo tanto, para poner esta información en las celdas relativas necesitamos una convención que conecte el título de estudio "licenciada" a un número particular, y lo mismo para cada otro título que deseemos registrar. Una convención del género se llama PLAN DE CODIFICACIÓN.

Un plan de codificación de la propiedad "título de estudio" podría ser:

[1] Se ponen números para facilitar el análisis estadístico de las relaciones entre variables. Pero este hábito conlleva el riesgo de que se hagan operaciones estadísticas sobre números que sólo son tales en apariencia.