natorias más elementales. En la siguiente seccion vamos a ocuparnos de estudiar la forma en que la teoría de las probabilidades se utiliza en el proceso de obtención de muestras aleatorias, mientras que en los siguientes capítulos nos ocuparemos del papel que juega la teoría de las probabilidades en el proceso de inducción.

4.3. Aspectos generales del muestreo en la investigación sociológica

La teoría del muestreo es el estudio de las relaciones existentes entre una población y las muestras extraídas de la misma. Se denomina «población» a un conjunto de casos o unidades que tienen en común una serie determinada de características —por ejemplo, el tener un trabajo remunerado determina la población laboral, o el hecho de residir en el medio rural determina la población rural—, y sobre la que se desea obtener cierta información. Dicha información puede consistir en la proporción de viviendas con cuartos de baño, el número de personas que opinan de un modo determinado o la proporción de posibles votantes en las próximas elecciones. Estos valores que se pretenden conocer, y que se expresarán mediante medidas de frecuencia, tendencia central o variación, tales como proporciones, razones, medias, desviaciones típicas, etc., se les denomina *valores verdaderos* (Sánchez-Crespo, 1971, pág. 11).

Normalmente, no se pueden calcular directamente tales valores porque las poblaciones no resultan directamente asequibles. De este modo, hay que recurrir al *muestreo*, que es un procedimiento por el que se infieren los valores verdaderos de una población a través de la experiencia obtenida con un grupo que contiene un número menor de casos que la población. Una *muestra* será el grupo de elementos seleccionados con la intención de estimar los valores verdaderos de la población. El investigador debe de preocuparse de que el número y el tipo de objetos incluidos en la muestra sean lo suficientemente representativos de la población total como para permitir hacer generalizaciones seguras acerca de la población. En otras palabras, los procedimientos de muestreo son unos medios para desarrollar una adecuada validez externa.

Diversas son las ventajas que ofrece el uso de las muestras para estimar valores de una población. En términos generales, se puede afirmar que el muestreo permite una reducción considerable de los costes materiales del estudio, una mayor rapidez en la obtención de la información y el logro de unos datos más comprensivos. Aunque esto último puede parecer aparentemente contradictorio, lo cierto es que a veces un buen plan de muestreo ofrece mejores estimaciones de los valores de una población que el propio Censo de Población. Este hecho ha sido, no obstante, destacado muchas veces por los propios estadísticos que elaboran los censos nacionales de población, ya que un proyecto de

tal magnitud produce más *errores no muestrales* y de mayor cuantía que el propio *error de muestreo* que se origina al estimar los parámetros de la población por medio de la muestra *.

Dadas las ventajas del muestreo, el buen muestreo no es practicable sin una clara conceptualización de lo que se está muestreando. Tal como afirma Smith (1975, pág. 106), existen muestras en busca de universos y universos en busca de muestras. Muchos problemas pueden eliminarse si previamente se conceptualizan claramente los objetos que han de servir como base para las generalizaciones del investigador. Las nociones de «universo general» y «universo de trabajo» son claves para entender este problema (Sjoberg y Nett, 1968, pág. 130). El universo general es la población abstracta y teórica a la que el investigador desea generalizar sus resultados, mientras que el universo de trabajo es la operacionalización concreta de ese universo general del que se va a obtener la muestra. Supongamos que deseamos estudiar el mercado de los ejecutivos en las grandes empresas españolas (universo general). Se puede operacionalizar, por ejemplo, a través de los listados de ejecutivos que están trabajando en una fecha determinada en las cien mayores empresas españolas (universo de trabajo).

Es importante realizar esta distinción entre ambos tipos de universo porque, en las investigaciones sociológicas, rara vez se tiene la oportunidad de obtener muestras directamente en los universos generales. Los temas de auténtico interés sociológico rara vez se pueden enmarcar en listados concretos, de los que se pueda obtener una muestra precisa con todos los requerimientos que demanda el cálculo de probabilidades. Los «pequeños grupos», la «conducta desviada», la «interacción en lugares públicos», la «despersonalización del trabajo burocrático», son fenómenos que difícilmente pueden ser estudiados siguiendo estrictos diseños muestrales.

Con todo, los diseños muestrales son necesarios si se desea que la investigación sociológica ofrezca resultados científicos. Todos sabemos que muchas personas tienden a realizar afirmaciones generales muy amplias, a partir del conocimiento de casos muy particulares. Esto es lo que Smith llama «muestras en busca de universos». Las muestras sesgadas se producen, precisamente, porque el investigador o la persona que hace la selección muestral se deja llevar, inconscientemente, por sus preferencias al elegir los casos. Esta es la razón por la que ha de evitarse que los entrevistadores tengan libertad para elegir la última unidad muestral.

Otro aspecto irónico de la investigación social es la existencia de

* En España, el Instituto Nacional de Estadística ha diseñado una Encuesta General de Población, de tipo contínuo y que proporciona estimaciones independientes bimensuales sobre las familias españolas. Para algunas características, tales como presupuestos familiares, gastos de consumo, nivel cultural, equipamiento, estimación del paro, etc., la E. G. B. ofrece las estimaciones más precisas de que se dispone.

universos generales teóricamente interesantes, pero que son relativamente abstractos o inaccesibles desde un punto de vista muestral (Smith, *op. cit.*, pág. 110). La mayor parte de los universos relacionales o interactivos son de este tipo, al igual que muchas organizaciones sociales, tales como burocracias, asociaciones voluntarias y comunidades. Pese a tales dificultades y problemas, el sociólogo debe de esforzarse por emplear diseños muestrales aleatorios, siempre que ello le sea posible, aunque, en último término, todo ello dependa de las facilidades materiales —dinero, tiempo, equipo— de que se disponga y del grado de exactitud deseado.

4.3.1. *Tipos de muestreo*

Para algunos, el único muestreo científicamente relevante es el *muestreo de probabilidad* o *muestreo aleatorio*. Pero, por todo lo que hemos dicho anteriormente, no siempre resulta posible en la investigación sociológica obtener una muestra probabilística; de ahí que con frecuencia el sociólogo tiene que recurrir a diseños muestrales arbitrarios para lograr algún tipo de resultado. Ahora bien, siempre que sea posible, se ha de preferir el muestreo aleatorio, ya que sólo en una muestra de este tipo se puede calcular un intervalo de confianza dentro del que se encuentran, con un nivel de probabilidad dado, los parámetros del universo.

La característica que distingue a una muestra probabilística es que cada individuo debe tener una *probabilidad conocida* de poder ser incluido en la muestra. De esta manera, se pueden realizar legítimamente inferencias estadísticas. Si las probabilidades se desconocen, no se podrá utilizar la inferencia estadística. Con el muestreo no probabilístico se puede llegar a obtener una muestra muy representativa, pero no se podrá evaluar a partir de ella los márgenes de error.

Desgraciadamente, no siempre es posible satisfacer las condiciones que exige un muestreo probabilístico, sobre todo la que hace referencia a la necesidad de disponer de un listado completo de las unidades del universo de trabajo. Así, por ejemplo, si un investigador deseara estudiar cualquier tipo de conducta desviada, como, por ejemplo, la cleptomanía, el abuso de drogas, etc., iba a ser completamente imposible obtener una lista completa de tales conductas, dado el carácter semioculto de las mismas. En tal caso hay que recurrir al *muestreo no probabilístico*, en el que generalmente se desconoce la probabilidad de selección que tiene cada unidad. El principal problema que tienen las muestras no probabilísticas es que rara vez se puede saber cuán representativa es la muestra del universo de trabajo.

4.3.2. *Muestreo aleatorio simple*

El muestreo probabilístico más sencillo es el que se denomina *muestreo aleatorio simple*. Para obtener una muestra aleatoria simple se parte de un conjunto listado de elementos de la población y, entonces, se seleccionan aleatoriamente N elementos para formar con ellos la muestra. La selección aleatoria se lleva a cabo de tal manera que: 1) cada elemento en la población tenga idéntica probabilidad de ser incluido en la muestra, y 2) cada posible combinación de N elementos tenga la misma probabilidad de constituir la muestra. Obsérvese que una selección aleatoria o al azar no significa una selección hecha de cualquier modo o casualmente; más bien significa un proceso de selección que da a cada elemento en la población la misma oportunidad de aparecer en la muestra.

Así, si el número de elementos que constituyen la muestra es M, la probabilidad de cada elemento individual en la muestra debe ser $1/M$. Si, por ejemplo, se desea extraer una muestra aleatoria simple de los 650 alumnos que componen un curso introductorio en una Facultad de Medicina, el proceso de selección debe ser tal que cada uno de los alumnos tenga una probabilidad de 1/650 de ser incluido en la muestra.

Además, la probabilidad de que cada alumno sea incluido en la muestra debe ser independiente de la probabilidad que tenga cualquier otro alumno de ser incluido. De este modo se podrá conseguir que cada combinación posible de N elementos tenga idéntica oportunidad de constituir la muestra. Cuando se cumple esta condición, y de acuerdo con la regla de la multiplicación de probabilidades, de una población de tamaño M se podrán extraer M^n posibles muestras aleatorias simples de tamaño n. Así, de la población formada por los 650 alumnos del curso introductorio de Medicina, si decidiéramos extraer una muestra de 100 alumnos, existirían $(650)^{100}$ posibles muestras de las que realizar la selección, lo que representa, ciertamente, una cifra enorme.

Supongamos que hemos decidido extraer una muestra de tamaño 100 del referido curso. Para hacerlo, ordenaríamos en primer lugar los 650 alumnos desde el número 1 al número 650. La selección aleatoria de los 100 componentes de la muestra se puede realizar fácilmente con la ayuda de la tabla A del apéndice. Detengámonos un momento a explicar cómo se forma y cómo se utiliza *una tabla de números aleatorios*.

Los *números aleatorios* son un conjunto de cifras del 0 al 9 cuya ordenación es totalmente casual, no respondiendo a plan alguno. Este conjunto de cifras cumple la propiedad de que:

$$P(a)=1/10;\ \text{siendo}\ a=0,\ 1,\ 2,\ \ldots,\ 9$$

En la tabla 1 se ha reproducido una parte de la tabla A de números aleatorios que se incluye en el apéndice. Las cifras están agrupadas en bloques de 5×2, con el fin de facilitar su presentación y lectura

TABLA 1

Reproducción de una parte de una tabla de números aleatorios

10	09	73	25	33	76	52	01	35	86	34	67	35	48	76	80	95	90	91	17	39	29	27	49	45
37	54	20	48	05	64	80	47	42	96	24	80	52	40	37	20	63	61	04	02	00	82	29	16	65
08	42	26	89	53	19	64	50	93	03	23	20	90	25	60	15	95	33	47	64	35	08	03	36	06
99	01	90	25	29	09	37	67	07	15	38	31	13	11	65	88	67	67	43	07	04	43	62	76	59
12	80	79	99	70	80	15	73	61	47	64	03	23	66	53	98	95	11	68	77	12	17	17	68	33
66	06	57	47	17	34	07	27	68	50	36	69	73	61	70	65	81	33	98	85	11	19	92	01	70
31	06	01	08	05	45	57	18	24	06	35	30	34	26	14	86	79	90	74	39	23	40	30	97	32
85	26	97	76	02	02	05	16	56	92	68	66	57	48	18	73	05	38	52	47	18	62	38	85	79
63	57	33	21	35	05	32	54	70	48	90	55	35	75	48	28	46	82	87	09	83	49	12	56	24
73	79	64	57	53	03	52	96	47	78	35	80	83	42	82	60	93	52	03	44	35	27	38	84	35

Tal como señala Doménech (1977, págs. 51 y sigs.), la construcción de una tabla de números aleatorios es, teóricamente, muy simple. A partir de una urna o bombo de lotería que contenga 10 bolas idénticas, numeradas del 0 al 9 —con lo que todas ellas tienen la misma probabilidad de ser extraídas—, se extrae una bola y se anota esta primera cifra aleatoria. Se vuelve a introducir la bola en la urna, se mezclan nuevamente y se realiza una nueva extracción, y así sucesivamente.

De esta forma se ha construido la tabla 1, y su utilización en estadística permite que intervenga el azar en una serie de operaciones, sin necesidad de recurrir cada vez a una urna con bolas o a un bombo de lotería. La extracción de una muestra en una población finita se puede hacer ahora con más facilidad.

Así, supongamos una población de 100 individuos, de la que queremos extraer una muestra al azar de $n=10$ individuos. Los individuos de esta población están numerados del 00 al 99. Se toman bloques de dos cifras en la tabla de números aleatorios, con lo que tendremos números al azar comprendidos entre 00 y 99. La muestra estará formada por los 10 primeros individuos cuyo número venga dado por la tabla de números aleatorios.

Siguiendo las filas de los números contenidos en la tabla 1, los 10 primeros números seleccionados serán:

10-9-73-25-33-76-52-1-35-86

con lo que, buscando los correspondientes números en la lista de los 100 individuos, se tendría una muestra de 10 individuos seleccionados aleatoriamente.

Volviendo al ejemplo anterior de la población de estudiantes de medicina, de la que se deseaba obtener una muestra de 100 alumnos, el procedimiento de selección mediante la tabla de números aleatorios será idéntico. Se listarán los 650 alumnos que forman la población y, a continuación, se seleccionarán los 100 primeros números menores de 650 que aparecen en la tabla de números aleatorios. De esta forma se habrá conseguido una muestra de 100 individuos seleccionados aleatoriamente.

Por lo que se refiere a la selección de los números que constituyen la muestra, conviene hacer notar lo siguiente. Si se toman los números que se van seleccionando, aunque alguno de ellos salga más de una vez, diremos que se trata de una *muestra con reemplazamiento*. Si, por el contrario, seleccionamos los números de forma que aparezcan una sola vez, no seleccionando, pues, los que hayan aparecido previamente, diremos que hemos obtenido una *muestra sin reemplazamiento*.

En este segundo caso, aunque sí se cumple la primera condición del muestreo aleatorio, esto es, que cada elemento de la población tenga idéntica probabilidad de ser incluido en la muestra, no se cumple, sin embargo, la segunda condición, que, como se recordará, hace referencia a la equiprobabilidad de cada posible muestra de ser elegida.

En efecto, cuando el muestreo es con reemplazamiento, el número posible de muestras és M^n. Pero cuando el muestreo es sin reemplazamiento, el número de posibles muestras de tamaño n viene restringido por el requisito de que cada caso esté presente tan sólo una vez en cada muestra. De este modo, el número de posibles muestras ya no es M^n, sino que viene dado por las combinaciones de M elementos tomados de n en n, que es:

$$\binom{M}{n}=\frac{M!}{(M-n)!\,n!}$$

El número de muestras de tamaño $n=100$ sin reemplazamiento que se podría extraer de la población $M=650$ alumnos sería:

$$\binom{650}{100}=\frac{650!}{(650-100)!\,100!}=\frac{650!}{550!\,100!}$$

lo que no deja de ser también una cifra astronómica, aunque menor que 650^{100}.

Aunque técnicamente existan, como vemos, diferencias entre las muestras con y sin reemplazamiento, en la práctica el error que se produce al utilizar las segundas en lugar de las primeras es mínimo cuando el tamaño de n es relativamente pequeño en relación a M. Además, el sociólogo pocas veces recurre a las muestras aleatorias simples, no sólo por la posibilidad de extraer el mismo caso más de una vez, sino también porque la mayor parte de las veces no dispone del listado ordenado de las unidades que componen el universo de trabajo.

Ahora bien, aunque en la práctica de la investigación pocas veces se utiliza el muestreo aleatorio simple, tiene gran interés estadístico por ser la técnica muestral básica de la estadística inferencial, y a partir de

la cual se han derivado la mayoría de las teorías y técnicas estadísticas originales. Además, el muestreo aleatorio simple sirve como modelo a partir del que se han derivado el resto de las técnicas muestrales aleatorias.

4.3.3. *Estimadores y errores de muestreo* *

Supongamos que, siguiendo el procedimiento aleatorio simple, se ha obtenido una muestra de n unidades. Se dirá que la expresión:

$$p=\frac{a}{n}=\frac{\Sigma_1 A_i}{n} \qquad [4.15]$$

es un *estimador de la proporción p*. A_i representa una variable cualquiera asignada a cada unidad de la población, tal como personas que poseen coche, familias de consumo alto, personas de ideología de izquierdas, etc. El sumatorio de todos los A_i representa, en los ejemplos anteriores, el total de personas que poseen coche, el total de familias de consumo alto o el total de personas de ideología de izquierda. Es lo que se denomina total de clase.

Dado que el estimador p ha sido calculado en base a las n unidades de la muestra, en lugar de las N unidades que constituyen la población, su valor estará afectado por un error que se denomina *error de muestreo*. Lo que se pretende al extraer una buena muestra es que el error de muestreo sea lo más pequeño posible, para que así el estimador sea tanto más preciso. Sánchez-Crespo explica de este modo el concepto de error de muestreo. Cada muestra de tamaño n que se extraiga de la población N dará una proporción p diferente de la anterior. Como el número de muestras sin reemplazamiento que se pueden obtener es $\binom{N}{n}$, éste será también el número de los posibles estimadores de p. Pues bien, el error de muestreo es la desviación típica de todos esos posibles valores de p (Sánchez-Crespo, *op. cit.*, pág. 35).

La estimación del error de muestreo se realiza utilizando los valores de la muestra, por medio de la fórmula:

$$\text{Error de muestreo}=s=\sqrt{\frac{N-n}{N}\cdot\frac{pq}{n-1}} \qquad [4.16]$$

en donde $q=1-p$, y s es un estimador de la desviación típica de p.

En esta fórmula, el factor $\frac{N-n}{N}$ se puede escribir como $1-\frac{n}{N}=1-f$, siendo f una probabilidad llamada *fracción de muestreo*, ya que representa el cociente entre el tamaño de muestra n y el tamaño de la población N. Cuando el valor de n es muy pequeño en relación al de N, f también es muy pequeño, y todo el factor $\frac{N-n}{N}$ puede considerarse igual a la unidad, con lo que la anterior fórmula queda de la forma:

$$\text{Error de muestreo}=\sqrt{\frac{pq}{n-1}}=\sqrt{\frac{pq}{n}} \qquad [4.17]$$

A partir de la estimación del error de muestreo se pueden determinar los *intervalos de confianza*, que son intervalos del tipo:

$$(p-zs,\ p+zs) \qquad [4.18]$$

Se denominan de este modo por el hecho de que el valor que se trata de estimar se encuentra dentro del citado intervalo con una «confianza», medida en términos de probabilidad, determinada por el valor que tome z. Así, si suponemos que el estimador p se distribuye normalmente, para $z=2{,}81$ la citada confianza alcanzará el 995 por 1.000 —ya que, en una distribución normal, la probabilidad de que la variable aleatoria sea distinta de su media en $\pm 2{,}81$ veces la desviación típica es 0,005—. Dicho en otras palabras, de cada mil muestras que se extrajeran mediante idéntico procedimiento, sólo en cinco de ellas el intervalo de confianza no cubriría el valor de p. El intervalo de confianza será tanto más pequeño cuanto mayor sea el tamaño muestral n.

Si lo que se pretende es calcular el total de la clase a que hace referencia la variable a, se utilizará el estimador:

$$a=N\cdot p \qquad [4.19]$$

cuyo error de muestreo puede estimarse por la fórmula:

$$s_a=N\cdot s=N\sqrt{\frac{N-n}{N}\cdot\frac{p\cdot q}{n-1}}$$

Veamos, a través de un ejemplo, la utilización de estas fórmulas de estimación de la proporción p y de establecimiento del intervalo de confianza. Supongamos que en la población española, que en 1980 era de alrededor de 37 millones de habitantes, se ha obtenido una muestra aleatoria de 10.000. La población activa en la muestra es de 4.000, y de éstos se encuentran en paro 450. A partir de estos datos se desea estimar el porcentaje de la población activa, el correspondiente error de muestreo y el intervalo de confianza, con un riesgo del 3 por 1.000. También se desea estimar el número de personas activas que se encuentran en situación de desempleo.

* Puede resultar conveniente estudiar los apartados 4.3.3 y 4.3.4 después de haber estudiado los modelos inferenciales en los capítulos 5 y 6.

El porcentaje estimado será:

$$p=\frac{4.000}{10.000}=0{,}40\ (40\ \%)$$

El error de muestreo aproximado será, utilizando la fórmula [4.17]:

$$s=\sqrt{\frac{pq}{n}}=\sqrt{\frac{0{,}40\cdot 0{,}60}{10.000}}=0{,}0049$$

con lo que el intervalo de confianza será, recordando la fórmula [4.18]:

$$(0{,}40-2{,}97\cdot 0{,}0049;\ 0{,}40+2{,}97\cdot 0{,}0049)=(0{,}385;\ 0{,}415)$$

que, dicho en otras palabras, puede expresarse diciendo que el porcentaje de la población activa está comprendido entre el 38,5 y el 41,5 por 100, con una probabilidad del 997 por 1.000.

Por lo que se refiere a la estimación del número de personas en situación de paro, será, teniendo en cuenta la fórmula [4.19] y que $p=\frac{450}{10.000}$:

$$a=37.000.000\times\frac{450}{10.000}=1.665.000 \text{ personas en paro}$$

4.3.4. *Determinación del tamaño de la muestra*

Sabemos, a través de la teoría de las muestras, que un número suficientemente grande de casos tomados aleatoriamente de un universo o población presenta, con casi toda seguridad, los mismos caracteres que el universo o población. Tanto por la ley del cálculo de probabilidades que rige la teoría de las muestras como por el propio sentido común, sabemos que cuanto mayor sea el número de elementos considerados más seguro será el resultado. Las respuestas de 50 personas elegidas al azar en una gran ciudad, aunque hayan sido escogidas con toda la cautela posible, no pueden ser representativas de las actitudes políticas de toda la población. Pero quizá no sea necesario, por otro lado, elegir a 50.000 personas para conocer con bastante exactitud la distribución de tales actitudes. Además, la selección de una muestra de 50.000 puede estar fuera de la capacidad económica y material de cualquier investigador. Este ha de encontrar un equilibrio entre los márgenes de exactitud que pretende obtener de los resultados de la muestra y el coste de la misma.

En general, se puede afirmar que se ha de utilizar la muestra que mejor represente el universo de trabajo con los medios materiales y económicos de que dispone el investigador. Unas veces será suficiente seleccionar 500 unidades para obtener una buena representación del universo de trabajo, y otras veces será necesario recurrir a muestras de hasta 30.000 unidades para alcanzar los márgenes de precisión deseados. Así, por ejemplo, si deseamos conocer el grado de satisfacción que ha provocado la retransmisión televisada de un programa habitual, tal como un partido de fútbol de liga, será precisa una muestra de alrededor de 1.000 personas. Con este tamaño, ya es posible conocer con bastante aproximación el grado de satisfacción entre los telespectadores de la retransmisión deportiva. Pero si lo que deseamos es hacer una predicción ajustada de la intención de voto municipal en España, país con una alta diversidad cultural y, por tanto, política, será preciso obtener muestras muy amplias en cada una de las regiones, lo que dará un tamaño muestral nacional muy elevado, probablemente mayor de 20.000.

En el siguiente cuadro hemos elaborado un cuadro con el tamaño de las muestras empleadas en la década 1970-1980 por el Centro de Investigaciones Sociológicas (hasta 1976 llamado Instituto de la Opinión Pública), y que es la institución española que realiza más encuestas de carácter sociopolítico.

Tamaño de las muestras en 176 encuestas realizadas por el Centro de Investigaciones Sociológicas en el período 1970-1980

Tamaño de las muestras	*N*	*%*
Menos de 500 unidades	14	8
500 - 1.000	17	10
1.001 - 1.500	88	50
1.501 - 2.000	15	9
2.001 - 2.500	16	9
2.501 - 5.000	7	4
5.001 - 10.000	8	5
10.001 - 15.000	2	1
15.001 - 20.000	1	*
20.001 - 25.000	3	2
25.001 - 30.000	5	3
Total	176	

FUENTE: Banco de Datos del CIS. Elaboración propia.

La distribución que se incluye en este cuadro pone de manifiesto que de las 176 encuestas realizadas por el CIS en la década 1970-1980, el 50 por 100 se hizo sobre muestras cuyo tamaño está comprendido entre 1.001 y 1.500 unidades, siendo la moda o valor más frecuente 1.200.

Las muestras de tamaño superior a 10.000 son minoritarias, sólo el 6 por 100, lo que pone de manifiesto que, por su elevado coste y gran complejidad, se utilizan en ocasiones excepcionales, como pueda ser en vísperas de elecciones generales, para conocer con precisión la intención de voto de la población española.

Igualmente son minoritarias las encuestas cuyos tamaños muestrales son inferiores a 1.000, tan sólo el 18 por 100, tratándose por lo general de estudios específicos realizados sobre poblaciones concretas, lo que no suele requerir un elevado tamaño muestral.

Volvamos ahora al caso del muestreo aleatorio simple y veamos como se determina el tamaño de una muestra, con el fin de obtener una precisión dada. Sabemos que el hecho de que el intervalo de confianza $p+zs$ contenga el valor p que tratamos de estimar, con un cierto nivel de probabilidad, equivale a decir que la diferencia en valor absoluto entre P y su estimación muestral p es menor o igual que $z \cdot s = E$, siendo E una cota de error absoluto especificada (para más detalle, ver Sánchez-Crespo, *op. cit.*, pág. 38, y Sánchez-Crespo, 1967).

De este modo se puede determinar el tamaño n de la muestra para estimar la proporción P —unidades con cierta característica—, de forma tal que la estimación p no difiera de P en más de la cota de error E con una probabilidad predeterminada.

Haciendo $n \simeq n-1$, ya que, para tamaños altos de n, la sustracción de una unidad no va a alterar prácticamente el valor de n, tendremos que:

$$E^2 = z^2 s^2 = z^2 \frac{N-n}{N} \cdot \frac{pq}{n}$$

con lo que:

$$NnE^2 = z^2 (N-n)\, pq = z^2 Npq - z^2 npq$$

y, por tanto:

$$NnE^2 + z^2 npq = z^2 Npq$$

despejando n, queda:

$$n = \frac{z^2 Npq}{NE^2 + z^2 pq} \qquad [4.20]$$

A partir, pues, del conocimiento del error absoluto prefijado, el margen de probabilidad deseado y el valor de p, es posible determinar el tamaño n de la muestra en una población de tamaño N conocido.

Veamos su aplicación a través de un ejemplo. Deseamos conocer el número de personas de todas las edades que sería necesario incluir en una muestra nacional para estimar la tasa de actividad en España, con un error absoluto de $E=0{,}03$ y una probabilidad del 95,5 por 100. El valor censal de p es del 0,40 por 100, según datos del último censo.

Los datos de que disponemos son los siguientes:

$$z^* = 2;\ E = 0{,}03;\ N = 37.000.000;\ p = 0{,}40 \text{ y } q = 0{,}60$$

con lo que, sustituyendo en la fórmula [4.20], tenemos:

$$n = \frac{z^2 Npq}{NE^2 + z^2 pq} = \frac{4 \cdot 37 \cdot 10^6\ (0{,}40)\ (0{,}60)}{37 \cdot 10^6\ (0{,}03)^2 + 4\ (0{,}40)\ (0{,}60)} = 1.066$$

Es decir, el tamaño de la muestra que se necesitaría es $n=1.066$ personas.

Puede parecer, para el no conocedor de la teoría de las muestras, que los universos de mayor tamaño han de requerir muestras igualmente de mayor tamaño. Pero ésta es una idea que hay que desechar de inmediato, ya que, ante todo, conviene aclarar *que el número de casos n a considerar en una muestra no depende de las dimensiones N del universo*. Es decir, no debe creerse que n constituye una cuota fija proporcional al universo, cosa que a veces parece desprenderse cuando en una publicación se indica el tamaño de la muestra por medio de la fracción del muestreo, es decir, basándose en el cociente entre el número de unidades elementales de la muestra y el de las que constituyen la población. Así, a veces, se suele hablar de una muestra del 5, del 1, del 10 por 100, etcétera. Pero debe quedar claro que n no depende del tamaño de N. Veamos su demostración matemática.

Elevando al cuadrado la fórmula [4.16], en la que se han sustituido los valores a estimar por los correspondientes en la población, y haciendo $n \simeq n-1$, queda que:

$$s^2 = \frac{N-n}{N} \cdot \frac{pq}{n}$$

y dividiendo por p^2:

$$C^2(p) = \frac{N-n}{nN} \cdot \frac{q}{p}$$

en donde C es un estimador del coeficiente de variación de p. Pues bien, despejando n, en esta fórmula queda que:

$$n = \frac{q}{pC^2(p) + \dfrac{q}{N}}$$

y dado que $\dfrac{q}{N}$ se puede considerar un valor aproximadamente igual

---

* z es igual a 2 porque la probabilidad dada es del 95,5 por 100, y sabemos que en una curva normal, se encuentra a ± 2 veces la desviación típica el 95,5 por 100 de todas las posibles muestras.

a 0, ya que q es menor que la unidad y N es un número elevado, queda que:

$$n=\frac{q}{pC^2(p)} \qquad [4.21]$$

con lo que queda claro que en la determinación de n no interviene el valor de N.

Colocándonos en el caso más desfavorable, esto es, que la proporción de casos favorables y desfavorables sea el 50 por 100, $p=q=1/2$, y fijando una precisión del 10 por 100, se obtiene que

$$\frac{1}{100}=\frac{N-n}{nN}$$

de donde:

$$n=\frac{100\,N}{N+100}$$

y, dando valores a N, se obtienen los siguientes valores de n, para un nivel de confianza del 95,5 por 100:

N	n	$f=n/N$
2.000	95	0,047
3.000	97	0,032
5.000	98	0,020
10.000	99	0,010
50.000	100	0,002
100.000	100	0,001
1.000.000	100	0,0001
3.000.000	100	0,00003
30.000.000	100	0,000003

FUENTE: J. L. SÁNCHEZ-CRESPO, *Principios elementales del muestreo*, Madrid, 1971, página 43.

Vemos, pues, por medio de esta tabla, que se puede necesitar prácticamente idéntica muestra para proporcionar datos de una pequeña ciudad de 50.000 habitantes que de una nación de 30 millones.

Para diferentes márgenes de error y de intervalo de confianza, y para valores fijos de p y q, se han construido tablas prontuarias que ofrecen la amplitud de la muestra para el caso de *poblaciones finitas*, no *muy grandes*. En la siguiente tabla aparecen los tamaños muestrales que se necesitan para márgenes de error que van del 1 al 10 por 100, en la hipótesis, más desfavorable, de $p=50$ por 100, y con un margen de confianza del 95,5 por 100.

El uso de esta tabla es bien sencillo. Si se quieren estudiar ciertas características, tales como intención de voto, ideología, etc., de una comunidad de 20.000 personas, y se establece como validez de los resultados un margen de error del 2 por 100 y un nivel de confianza del 95,5 por 100, la muestra deberá estar constituida por 2.222 personas, tal como se puede obtener mirando en la celdilla en la que se cruzan el valor $n=20.000$ de las filas y el valor ± 2 de las columnas.

Tabla para la determinación de una muestra sacada de una población finita, para márgenes de error de 1, 2, 3, 4, 5, 10 por 100, en la hipótesis de p=50 por 100. Nivel de confianza del 95,5 por 100

Amplitud de la población	Amplitud de la muestra para márgenes de error abajo indicados					
	± 1 %	± 2 %	± 3 %	± 4 %	± 5 %	± 10 %
500					222	94
1.000				385	286	83
1.500			638	441	316	91
2.000			714	476	333	95
2.500		1.250	769	500	345	96
3.000		1.364	811	517	353	97
3.500		1.458	843	530	359	97
4.000		1.538	870	541	364	98
4.500		1.607	891	549	367	98
5.000		1.667	909	556	370	98
6.000		1.765	938	566	375	98
7.000		1.842	949	574	378	99
8.000		1.905	976	580	381	99
9.000		1.957	989	584	383	99
10.000	5.000	2.000	1.000	588	385	99
15.000	6.000	2.143	1.034	600	390	99
20.00	6.667	2.222	1.053	606	392	100
25.000	7.143	2.273	1.064	610	394	100
50.000	8.333	2.381	1.087	617	397	100
100.000	9.091	2.439	1.099	621	398	100
∞	10.000	2.500	1.111	625	400	100

p = proporción (en porcentajes) de los elementos portadores del carácter considerado. Si p es < 50 por 100 la muestra necesaria es más pequeña.

FUENTE: G. TAGLIACARNE, *Técnica y práctica de las Investigaciones de Mercado*, 1962, página 156.

4.3.5. *Otros tipos de muestreo probabilístico*

En la práctica de la complejidad de la investigación sociológica, no suele ser corriente que el sociólogo disponga de una lista actualizada de las «unidades elementales» sobre las que va a realizar su investigación, sean obreros, votantes, familias, viviendas, etc. Incluso a veces, cuando tal lista existe pero es de ámbito geográfico disperso, la extracción aleatoria simple puede producir una muestra cuyas unidades se

encuentran repartidas de tal modo que haga prohibitivo el coste de desplazamiento de los entrevistadores que han de conectar tales unidades.

Por esa razón, y de forma general, se hace necesario recurrir a una muestra de grupos de unidades elementales, denominados *conglomerados* (en inglés, *clusters*). Cuando es posible determinar los límites geográficos de los conglomerados, y así resulta de interés al investigador, el muestreo se denomina de *áreas*.

Cuando en la muestra de conglomerados se conecta con todas las unidades elementales que los forman, se dice entonces que el muestreo es en *una sola etapa* o sin submuestreo. A veces resulta de mayor interés, para reducir costos e incrementar la precisión, preparar una lista de unidades elementales dentro de cada conglomerado, a partir de la cual se obtiene una muestra de éstas. En tal caso, el muestreo se denomina *bietápico* o con submuestreo. Esta forma de proceder puede generalizarse fácilmente a un número mayor de etapas: en cada una de éstas existe un tipo de unidades de muestreo, denominándose primarias a las de la primera etapa, secundarias a las de la segunda, etc. Esta forma de muestreo se denomina *polietápico* o en varias muestras, y en él es necesario establecer una jerarquía de unidades de muestreo. Más adelante ofreceremos un ejemplo real de muestreo polietápico, pero antes introduzcamos un concepto fundamental en el diseño muestral, el de estratificación de la muestra.

En una *muestra estratificada* se dividen primeramente todos los individuos en grupos o categorías y, a continuación, se seleccionan muestras independientes dentro de cada grupo o estrato. Los estratos se deben definir de tal manera que cada individuo aparezca en sólo un estrato. Cuando las fracciones muestrales para cada estrato son idénticas se tiene el *muestreo estratificado proporcional*, y cuando son de tamaños diferentes se tiene el *muestreo estratificado desproporcional*.

Varios son los objetivos que se pueden perseguir al estratificar una muestra. Sánchez-Crespo cita los siguientes: 1) ofrecer estimaciones separadas para ciertas subpoblaciones; 2) agrupar unidades de muestreo homogéneas entre sí en estratos, con objeto de mejorar la precisión de las estimaciones globales, y 3) utilizar métodos diferentes de muestreo en los distintos estratos (Sánchez-Crespo, 1971, pág. 17).

Cuando se calculan estimaciones de la media y de la desviación típica a partir de muestras estratificadas, es preciso calcular los correspondientes valores para cada uno de los estratos y, a continuación, se ponderan de acuerdo con el tamaño relativo del estrato en la población. Así, si W_i representa el peso o ponderación del estrato i en la población y si establecemos que $\Sigma W_i = 1$, con lo que se consigue reducir los pesos a proporciones, se puede establecer la fórmula para estimar la media de la población como sigue:

$$\overline{X} = \sum_{i=1}^{K} W_i \overline{X}_i \qquad [4.22]$$

en donde $\overline{X}_i$ son las medias muestrales en cada uno de los K estratos. Veamos a través de un ejemplo sencillo la utilización de las ponderaciones en la determinación de los parámetros en una muestra estratificada. Supongamos que hemos tomado datos de tres comarcas en una provincia y los valores obtenidos son los siguientes:

	COMARCA			
	1	*2*	*3*	*Total*
Tamaño comarca	20.000	30.000	50.000	100.000
Peso W_i	0,20	0,30	0,50	$W_i = 1$
Tamaño muestra	100	100	100	$n = 300$
Media muestral $\overline{X}_i$	1.500	2.000	3.000	

La media muestral $\overline{X}_i$ hace referencia a la media de una característica determinada que se esté investigando. Pues bien, los datos que aparecen en el cuadro anterior ponen de manifiesto que se ha obtenido una muestra desproporcional, ya que se tienen fracciones muestrales diferentes para cada estrato, esto es: $\frac{100}{20.000}$ en la comarca 1; $\frac{100}{30.000}$ en la comarca 2, y $\frac{100}{50.000}$ en la comarca 3. Supongamos también que dentro de cada estrato se ha realizado un muestreo aleatorio simple y que las muestras son independientes entre sí. La media estimada será, aplicando la fórmula [4.22], la siguiente:

$$\overline{X} = 0{,}20\ (1.500) + 0{,}30\ (2.000) + 0{,}50\ (3.000) = 300 + 600 + 1.500 = 2.400$$

Otra propiedad interesante del muestreo estratificado es que puede demostrarse que cualquier estrato de una muestra aleatoria simple de una población es, en sí misma, una muestra aleatoria simple del correspondiente estrato de la población. Dicho en otras palabras, el procedimiento de obtención en primer lugar de una muestra aleatoria simple y después dividirla en estratos es equivalente al procedimiento de obtener una muestra aleatoria estratificada, utilizando como fracción de muestreo dentro de cada estrato la proporción de ese estrato que había en la muestra aleatoria simple (Sellitz *et al.*, 1961, págs. 580 y sigs.).

El procedimiento que usualmente se sigue, pues, en la obtención de muestras estratificadas es el de dividir la población objeto de estudio en grupos que llamamos estratos y, a continuación, se obtiene una muestra de cada estrato. Algunas veces, sin embargo, resulta conveniente di-

vidir la población en un número más amplio de grupos, llamados conglomerados (o *clusters*), y realizar el muestreo entre los conglomerados. Así, por ejemplo, se puede dividir una ciudad en unos cuantos cientos de secciones censales y, entonces, seleccionar aleatoriamente 50 secciones para la muestra. Este tipo de diseño muestral se denomina *muestreo de conglomerados*, y se utiliza frecuentemente en las encuestas sociológicas, con el objeto de reducir el coste en la fase de recolección de datos. Para ello se seleccionan conglomerados lo más heterogéneos posible, pero que sean lo suficientemente pequeños como para reducir los costes de desplazamientos de los entrevistadores.

En el muestreo de conglomerados no se seleccionan las unidades finales directamente. En un proceso claramente polietápico se obtienen muestras de conglomerados. En el diseño más simple en este tipo de muestreo se puede utilizar una selección aleatoria entre los conglomerados y, a continuación, se selecciona cada unidad individual perteneciente a los conglomerados incluidos en la muestra de conglomerados. Tal diseño se denomina a veces muestra de conglomerados en una sola etapa, ya que, de hecho, sólo se selecciona una muestra. En un diseño polietápico, por otro lado, las cosas pueden ser más complicadas. Así, por ejemplo, se puede obtener en primer lugar una muestra de secciones censales dentro de una ciudad. A continuación se puede obtener una muestra aleatoria simple de manzanas en cada sección. En una tercera etapa se puede instruir al entrevistador para que seleccione determinada enésima vivienda en cada manzana y que entreviste a un miembro, seleccionado al azar, de la familia que resida en dicha vivienda. En este caso vemos, pues, que el proceso aleatorio se introduce varias veces.

Veamos a través de un ejemplo la complejidad de un diseño muestral polietápico utilizado para realizar una encuesta sobre actitudes regionales de la población española (Jiménez Blanco *et al.*, 1977, páginas 15 y sigs.). Se asignó una cuota provincial mínima de 100 entrevistas, que para las provincias más pobladas podían ser hasta 400 entrevistas. De este modo, el tamaño muestral para todo el territorio nacional fue de 6.500, lo que permitió obtener una muestra cuyos resultados iban a tener un error máximo admisible del 10 por 100, con un nivel de significación del 95 por 100.

El reparto intraprovincial de las entrevistas se realizó del siguiente modo. Como el tema de estudio era la problemática regional, se estimó que el criterio más acertado para la estratificación de la muestra sería el del tamaño del municipio, que, al combinarse con cada provincia, asegura una mayor homogeneidad de la población en cada contexto. De acuerdo con este criterio, se establecieron para cada provincia los siguientes seis estratos:

1. Areas metropolitanas.
2. Municipios cuya población de hecho es de 100.000 o más habitantes.

3. Municipios cuya población de hecho oscila entre 50.000 y 99.999 habitantes.
4. Municipios cuya población de hecho está comprendida entre 10.000 y 49.000 habitantes.
5. Municipios cuya población de hecho está comprendida entre 3.000 y 9.999 habitantes.
6. Municipios cuya población de hecho es menor de 3.000 habitantes.

Para cada provincia se establecieron los porcentajes que del total provincial representan cada uno de estos estratos y, en base a estas proposiciones, se repartió el total de entrevistas asignadas a éste entre aquellos estratos.

La elección de los puntos o unidades últimas de muestreo se llevó a cabo en las siguientes etapas: *a)* Elección de municipios —se realizó una elección con probabilidad proporcional al número de habitantes—. *b)* Elección de entidades singulares de población —se eligió, con probabilidad proporcional al número de habitantes, una entidad de población entre todas las que componían cada municipio—. *c)* Elección de la ruta —cada entidad de población se dividió en sectores, y en cada uno de ellos se eligió al azar un origen de ruta a seguir por entrevistador y, mediante una tabla de números aleatorios, se seleccionaron los portales con entrevistas a realizar—. *d)* Elección del hogar —una vez efectuada la elección del portal se censaron todos los hogares del mismo mediante una nueva serie de números aleatorios, con lo que se determinó el hogar a entrevistar. *e)* Elección de la persona a entrevistar —se realizó mediante una combinación del número de personas de la familia mayores de dieciocho años (sujetos de la entrevista), el número del cuestionario a aplicar y una tabla de números aleatorios.

Vemos cómo en las numerosas etapas del muestreo el azar interviene constantemente, con lo que se asegura el carácter probabilístico de las sucesivas elecciones y se evita la introducción de sesgos, tanto por parte del investigador que diseña la muestra como del entrevistador que elige las unidades últimas. Obsérvese también que el tamaño muestral se elige de forma apriorística —cosa que se hace comúnmente en las encuestas sociológicas—, en función de las disponibilidades de tiempo y dinero con que se cuenta para hacer la investigación, y posteriormente se distribuye la muestra polietápicamente y de acuerdo con estrictos criterios aleatorios.

4.3.6. *Muestreo no probabilístico*

Existen técnicas muestrales que no implican el criterio de aleatoriedad y probabilidad en la selección de las unidades muestrales. Se utilizan algunas veces tales técnicas porque tienen unos costes más bajos en la recolección de datos, o porque al utilizarlas se evitan los proble-

mas que a menudo se presentan al extraer muestras al azar. La máxima desventaja de las muestras no probabilísticas es que no permiten la obtención de una estimación válida de los márgenes de error, y, en tal sentido, el sociólogo debe tratar de evitar, siempre que ello sea posible, su utilización.

Entre las técnicas no probabilísticas destacan las siguientes:

a) *Muestras accidentales.* En un muestreo accidental se toman simplemente los casos que vienen a mano, continuando el proceso hasta que la muestra adquiere un tamaño precisado. Esto es lo que hacen los periodistas de radio y televisión cuando tratan de «pulsar la opinión pública y, con sus micrófonos y cámaras, se dirigen a las primeras personas que encuentran en la calle y se dejan entrevistar.

En un muestreo accidental no hay forma de conocer los sesgos que se introducen al entrevistar, por ejemplo, personas atípicas o casos extremos, y lo único que puede desearse al proceder de este modo es que la equivocación no sea excesiva.

b) *Muestras sistemáticas.* Una muestra sistemática se consigue extrayendo de una lista cada enésimo caso; por ejemplo, extrayendo cada décima unidad. Este tipo de muestreo es no probabilístico, ya que si, por ejemplo, seleccionáramos en una cola de personas cada diez de ellas, las personas que ocupan los puestos 10, 20, 30, etc., tienen una probabilidad de 1,00 de ser incluidas, mientras que el resto de las personas de la cola tienen una probabilidad cero.

c) *Muestras de cuota.* Es quizá el tipo de muestreo más popular y más utilizado por los analistas de mercados y de opinión pública. El tipo de técnica muestral por medio de cuotas goza de tanta aceptación porque es un medio barato, rápido y conveniente de obtener datos. Una muestra por cuotas se obtiene al especificar las características deseadas de los sujetos que se desea entrevistar, y entonces se deja en libertad al entrevistador para que encuentre y entreviste una cuota de personas que posean las referidas características. Obviamente, el procedimiento es no probabilístico, ya que se deja en libertad a los investigadores para que alcancen la cuota prefijada de entrevistas de la forma que les sea más conveniente.

d) *Muestras intencionadas.* La hipótesis básica del muestreo intencionado (en inglés, *purposive sampling*) es que, con un buen juicio y una estrategia adecuada, se pueden decidir fácilmente los casos a ser incluidos en la muestra. Una estrategia corriente es tomar casos que se juzgan como típicos de la población, suponiendo que los errores de juicio en la selección tenderán a compensarse entre sí.

Ahora bien, sin una comprobación de otro tipo, no es posible saber si los casos «típicos» lo son en realidad; además, cuando se producen cambios es preciso, además, saber cómo afectan al caso «típico».

Los sociólogos y antropólogos que estudian comunidades rurales, o los sociólogos que estudian establecimientos e instituciones sociales concretos, siguen de algún modo un muestreo intencionado, ya que, en último término, se suelen apoyar en sus respectivos conocimientos subjetivos, y no en un criterio objetivo, contrastable y riguroso, como el cálculo de probabilidades, para elegir sus casos de estudio.

4.4. Terminología

Se recomienda la memorización y comprensión del significado de cada uno de los términos y conceptos siguientes:

— Parámetros.
— Indicadores estadísticos; estadístico.
— Probabilidad matemática u objetiva.
— Probabilidad real o personalista.
— Probabilidad *a priori.*
— Probabilidad empírica; probabilidad verdadera.
— Adición de probabilidades.
— Producto de probabilidades.
— Probabilidad condicional.
— Sucesos dependientes; sucesos independientes.
— Proceso; proceso estocástico.
— Cadenas de Markov.
— Variaciones.
— Permutaciones.
— Combinaciones.
— Teoría del muestreo.
— Valores verdaderos.
— Muestreo; muestra.
— Errores no muestrales.
— Error de muestreo.
— Universo general; universo de trabajo.
— Muestreo de probabilidad o muestreo aleatorio.
— Muestreo no probabilístico.
— Muestreo aleatorio simple.
— Números aleatorios; tabla de números aleatorios.
— Muestreo con reemplazamiento; muestreo sin reemplazamiento.
— Estimadores.
— Errores de muestreo.
— Fracción de muestreo.
— Intervalos de confianza.
— Tamaño de la muestra.
— Muestreo de conglomerados; muestreo de áreas.
— Muestreo polietápico.
— Muestreo estratificado proporcional y desproporcional.