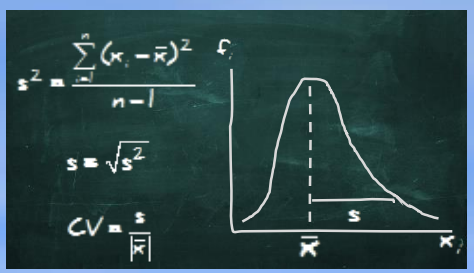
METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli



UAB
Universitat Autònoma de Barcelona

METODOLOGÍA DE LA INVESTIGACIÓN SOCIAL CUANTITATIVA

Pedro López-Roldán
Sandra Fachelli

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Pedro López-Roldán
Centre d'Estudis Sociològics sobre la Vida Quotidiana i el Treball (http://quit.uab.cat)
Institut d'Estudis del Treball (http://iet.uab.cat/)
Departament de Sociologia. Universitat Autònoma de Barcelona
pedro.lopez.roldan@uab.cat

Sandra Fachelli
Departament de Sociologia i Anàlisi de les Organitzacions
Universitat de Barcelona
Grup de Recerca en Eduació i Treball (http://grupsderecerca.uab.cat/gret)
Departament de Sociologia. Universitat Autònoma de Barcelona
sandra.fachelli@ub.edu

Edición digital: http://ddd.uab.cat/record/129382

1ª edición, febrero de 2015

Edifici B · Campus de la UAB · 08193 Bellaterra
(Cerdanyola del Vallés) · Barcelona · España
Tel. +34 93 581 1676

Índice general

Metodología de la Investigación Social Cuantitativa

Pedro López-Roldán
Sandra Fachelli

PARTE II. PRODUCCIÓN

Capítulo II.4
El diseño de la muestra

Bellaterra (Cerdanyola del Vallès) | Barcelona
Dipòsit Digital de Documents
Universitat Autònoma de Barcelona

Cómo citar este capítulo:

López-Roldán, P.; Fachelli, S. (2017). El diseño de la muestra. En P. López-Roldán y S. Fachelli, *Metodología de la Investigación Social Cuantitativa*. Bellaterra. (Cerdanyola del Vallès): Dipòsit Digital de Documents, Universitat Autònoma de Barcelona. Capítulo II.4. https://ddd.uab.cat/record/185163

Capítulo acabado de redactar en agosto de 2017

Contenido

PARTE II
Capítulo 4

El diseño de la muestra

El diseño de la muestra corresponde a una tarea específica, de implicaciones metodológicas y requerimientos técnicos, destinada a elegir una representación adecuada de unidades de nuestra población objeto de estudio. Una muestra no es más que la elección de una parte de un todo que es la población. Nos referiremos fundamentalmente a muestreo estadístico, por tanto, al diseño y la obtención de una muestra estadísticamente representativa de la población que se inscribe en un proceso de investigación de carácter cuantitativo donde la teoría del muestreo y de probabilidades son elementos importantes definitorios. No obstante, comentaremos en el final del capítulo algunas estrategias de muestreo no probabilístico o cualitativo que nos completará la visión de diferentes alternativas que se plantean ante la necesidad de elegir a los informantes de la investigación y, más allá, de las unidades sobre las que se define el objeto de estudio.

Por las exigencias técnicas asociadas al diseño de la muestra este capítulo II.4 se complementa con el capítulo III.4 dedicado a la inferencia estadística. Como allí destacamos, cuando se trabaja con información estadística se diferencia entre el aspecto descriptivo de los datos del inferencial. Es decir, mediante la descripción damos cuenta de un fenómeno a través de cálculos, medidas, datos que reflejan cuantitativamente un determinado aspecto de la realidad estudiada, mientras que la mediante la inferencia establece en qué medida esos cálculos, medidas o datos pueden extrapolarse al conjunto de la población a partir de la muestra determinando el margen de error que podemos cometer al hacerlo. Por tanto, la inferencia estadística no ayuda a establecer el grado de validez externa de los resultados de nuestra investigación.

En nuestra exposición combinaremos aspectos explicativos con un discurso comprensible con otros comentarios más técnicos que requieren de un bagaje conceptual estadístico que se podrá asimilar con posterioridad si no se dispone. Ello no impedirá la lectura ni la aplicación de los principales aspectos involucrados en el diseño de la muestra, en particular, los destinados a la determinación del tamaño de la muestra o del margen de error.

Introduciremos al lector/a, en un primer apartado, recorriendo algunos de los principales conceptos implicados en el muestreo estadístico. Veremos en particular una

clasificación de los diferentes tipos de muestreo que se explicarán con mayor detalle en el resto del capítulo. Así, empezaremos por el tipo de muestreo fundamental: el muestreo aleatorio simple, que, además de constituir una primera estrategia, sirve de base del razonamiento estadístico y para dar cuenta del resto de métodos de muestreo: el muestreo sistemático, el muestreo estratificado, el muestreo de conglomerados o el muestreo polietápico. Finalizaremos el capítulo dando cuenta del muestreo no probabilístico y presentando algunos ejemplos de diseños de muestras.

1. El muestreo estadístico. Definiciones y conceptos básicos

El objetivo general de todo muestreo es llegar a conocer determinadas características de una población, a partir de una selección de unidades de ésta, con el menor coste posible en dinero, tiempo y trabajo[1]. Mediante las técnicas estadísticas, las leyes probabilísticas y los diseños muestrales basados en varios métodos de muestreo, podemos aproximarnos al conocimiento de estas características sin necesidad de tener que obtener la información exhaustiva de toda la población, como en el censo, garantizando la representatividad y sabiendo que cometeremos un determinado error estadístico, que se puede determinar de antemano en cada caso, por el hecho de tener una parte del todo.

Presentamos seguidamente una relación de conceptos asociados al diseño de muestras estadísticas.

1.1. Muestra

Una muestra estadística es una parte o subconjunto de unidades representativas de un conjunto llamado población o universo, seleccionadas de forma aleatoria, y que se somete a observación científica con el objetivo de obtener resultados válidos para el universo total investigado, dentro de unos límites de error y de probabilidad de que se pueden determinar en cada caso. Denotaremos al tamaño de la muestra mediante n.

Se pueden establecer las siguientes condiciones de las muestras:

1. Que comprendan parte del universo y no la totalidad.
2. Que el tamaño de la muestra sea estadísticamente proporcionado a la magnitud del universo.
3. Que se dé una ausencia de distorsión en la elección de los elementos de la muestra con el fin de evitar la introducción de sesgos que desvirtúen la representatividad.
4. Que sea posible poner a prueba hipótesis sustantivas de relaciones entre variables.
5. Que sea posible poner a prueba hipótesis de generalización, de la muestra en el universo, es decir, que sea representativa con un cierto grado de incertidumbre. En este sentido las muestras se dice que son probabilísticas y cualquier cálculo con los datos muestrales son estimaciones de características o parámetros poblacionales.

[1] Es posible diseñar muestras introduciendo funciones de coste intentando alcanzar la máxima precisión dada la restricción de unos costos determinados. Kish (1972) propuso una función general de cuatro factores de gastos fijos y variables como aproximación sencilla y aproximada de los costes reales finales. En el texto no trataremos esta perspectiva específica. No obstante, el diseño real de cualquier muestra siempre se enfrenta con el dilema entre precisión y costes, lo que se traduce, en particular, en la determinación del número de encuestas a realizar.

Mediante la estadística descriptiva alcanzamos el objetivo de describir la información estadística facilitando la tarea de análisis e interpretación de los datos mediante diversos cálculos (como los estadísticos descriptivos de la distribución de una variable y de las relaciones entre ellas), y cumplimos así la condición 4.

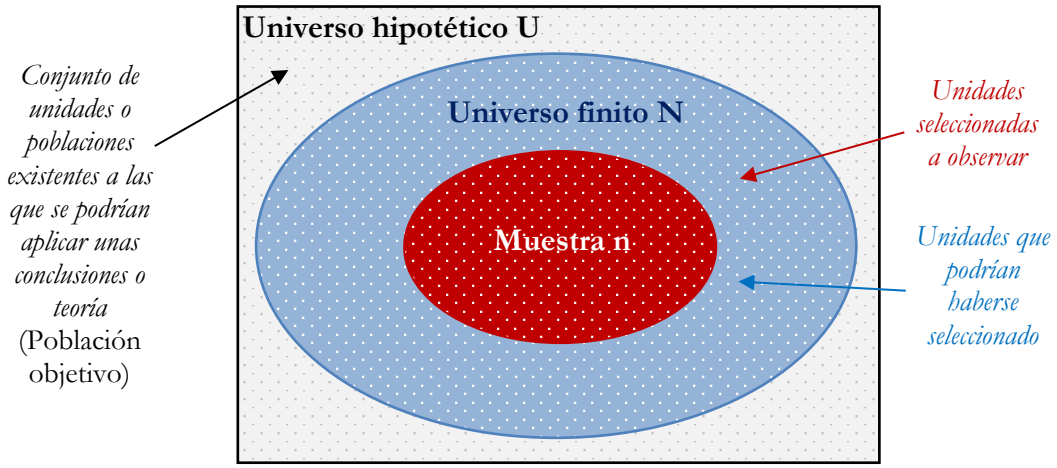
Mediante la estadística inferencial fundamentamos los principios estadísticos que nos permiten alcanzar el objetivo de obtener generalizaciones estadísticas de la población a partir de la muestra, y cumplimos así con la condición 5. La estadística inferencial (o estadística inductiva) nos hablará de la significación de los cálculos en base al proceso de estimación de los parámetros poblacionales a partir de los estadísticos muestrales, fundamentándose en la Teoría de Probabilidades y la Teoría del Muestreo, que nos proporcionan los procedimientos por los que esta generalización o inferencia será posible.

1.2. Universo o Población

Universo o Población son expresiones equivalentes para referirse al conjunto total de elementos que constituyen el ámbito de interés analítico y sobre el que queremos inferir las conclusiones de nuestro análisis, conclusiones de naturaleza estadística y también sustantiva o teórica. En particular se habla de población marco o universo finito, al conjunto preciso de unidades del que se extrae la muestra, y universo hipotético o población objetivo, el conjunto poblacional al que se pueden extrapolar los resultados. Denotaremos al tamaño de la población mediante N.

En el Gráfico II.4.1 se representan los conceptos de universo (hipotético y finito) y de muestra.

Gráfico II.4.1. Representación de una muestra respecto de su población



Al hablar de poblaciones se establece la distinción entre una población finita y una infinita. Desde el punto de vista del muestreo, la distinción se basa en la importancia relativa que tiene el tamaño de la muestra n en relación al tamaño de población N. Si el tamaño de la muestra es muy pequeño respecto a la de la población (habitualmente se admite que represente menos del 5%) se suele considerar infinita la población. En cambio, si la muestra necesaria es considerable en relación a la población (por encima

del 10% se suele considerar necesario, y entre un 5% y un 10% recomendable) se considera finita la población y se han de utilizar factores de corrección de población finita. Igualmente se considera que una población finita a toda población formada por menos de 100.000 unidades, e infinita a aquella que tiene 100.000 o más.

Como ejemplos de poblaciones podemos imaginar un sinfín de posibilidades. Será habitual considerar a las personas, en particular, personas residentes en una ciudad, comarca, provincia, región, país, continente, etc. Pueden ser asimismo consumidores/as, televidentes, internautas, fabricantes, jóvenes, estudiantes, trabajadores/as, etc. Pero también nuestra población puede venir definida a partir de considerar como unidades a los edificios o las viviendas-hogares, a las zonas geográficas desde un punto de vista topográfico (par el muestreo por teledetección), a las empresas, a los productos fabricados, a los productos vendidos-consumidos. O pueden ser asociaciones e instituciones. Igualmente, los artículos de revista o de prensa, las imágenes, las películas o los programas de televisión pude definir unidades de una población.

Se utiliza la expresión subpoblación para referirse a un subconjunto de la población. Por ejemplo, la subpoblación de mujeres de la población catalana. Este subgrupo en términos de la muestra se llamaría submuestra.

1.3. Fracción de muestreo

La fracción de muestreo indica simplemente el porcentaje que representa la muestra sobre la población.

$$f = \frac{n}{N} \times 100 \qquad \textit{Ecuación 1}$$

Por ejemplo, una muestra de n=2.000 personas sobre una población de N=4.000.000 da lugar a una fracción de muestreo del 0,05%.

1.4. Coeficiente de elevación

El coeficiente de elevación indica el número de veces que está contenida la muestra en la población, es decir, por qué valor debemos multiplicar cada unidad de la muestra para tener los efectivos de la población.

$$f^{-1} = 1 \Big/ \frac{n}{N} = \frac{N}{n} \qquad \textit{Ecuación 2}$$

En el ejemplo anterior el coeficiente de elevación sería de 2.000.

En ocasiones los resultados que se obtienen en una muestra de un tamaño determinado se expresan y se presentan en los informes y estudios en términos de los valores que corresponden al tamaño de la población. Así lo hace, por ejemplo, en la Encuesta de Población Activa y, en general, en la *Labor Force Survey* de Eurostat. Esta operación se conoce como elevación del dato muestral a la población.

Esta operación tiene el interés de reproducirnos las dimensiones poblacionales de nuestros datos. Ahora bien, conviene tener muy presente que los análisis descriptivos si bien pueden ser equivalentes tratando datos muestrales referidos al número de casos de la muestra o tratando datos elevados referidos a todos los casos de la población, no es así con respecto a los razonamientos o análisis inferenciales. Uno de los parámetros fundamentales de la estadística inferencial es justamente el tamaño de la muestra, sobre él se aplican las distintas fórmulas que nos ayudan a establecer la significación de nuestros resultados, por ello, no podemos reemplazar el tamaño muestral por el tamaño poblacional pues se entendería que la muestra es enormemente grande.

1.5. Equilibrio de la muestra

Junto a la operación de elevación, en ocasiones, se plantea otra operación que es la ponderación de la muestra. La ponderación obedece a necesidades diversas. Una de ellas es la que se deriva del equilibrio de la muestra. Cuando se obtiene una muestra tras un trabajo de campo, a posteriori, cabe plantearse su validación y preguntarse hasta qué punto la muestra refleja la población de la que se ha extraído, hasta qué punto es una buena representación. En este sentido se habla de verificar el equilibrio de la muestra mediante la determinación de las posibles discrepancias entre la población y la muestra. No obstante, este criterio cabe planteárselo, primero, en el diseño de la muestra, en el proceso de recogida de información, a continuación, y, finalmente, una vez obtenida la muestra.

Se trata de un análisis de validación que depende de la información poblacional disponible y del conocimiento que exista del fenómeno en cuestión para contrastar un resultado muestral concreto con relación a un conocimiento general existente o esperable. Cuando se observa que nuestros datos difieren lo esperado cabe plantearse cómo ajustar ese desequilibrio. Por ejemplo, nuestra muestra puede caracterizarse por tener un porcentaje excesivo de varones; sabiendo que en la población son aproximadamente el 48%, una muestra que arroje un porcentaje de varones del 45% estará infravalorando el peso y los rasgos de la población masculina. Lo mismo se puede dar cuando registramos más personas adultas mayores, menos personas con ingresos extremos (muy ricas o muy pobres), o menos inmigrantes. Para remediar el sesgo que se produce se procede al equilibrio de la muestra mediante la ponderación de la muestra. Se trata de una ponderación a posteriori, una vez se ha obtenido la muestra, para resolver los posibles desajustes en relación al diseño con el fin de equilibrarla y dar un mayor peso a los perfiles de la muestra que quedaron infravalorados o sobrevalorados.

En otras ocasiones, la ponderación puede ser también necesaria como resultado deliberado de un diseño muestral que, para garantizar la presencia de determinadas características o por el procedimiento de elección de las unidades, contempla la necesaria ponderación de la muestra una vez ésta se ha obtenido con el fin de restablecer el criterio de estricta aleatoriedad y proporcionalidad.

Tanto en el caso de la elevación como de la ponderación, las operaciones de elevar y ponderar requieren la construcción de una variable específica de elevación o de ponderación que se aplica en el tratamiento estadístico de los datos. Y puede suceder

que tanto la elevación como la ponderación se unan en una sola variable que actúa simultáneamente modificando nuestros datos en los dos sentidos.

1.6. Unidad de la muestra

Cada uno de los elementos de una muestra o de la población se denomina unidad o individuo (ya sea una persona o no). El listado de todas las unidades de donde se extrae la muestra constituye la base o el marco de la muestra, también llamado técnicamente como espacio muestral.

El marco de la muestra puede ser el censo, es decir, la relación exhaustiva de todas las unidades poblacionales: de habitantes, electoral, de viviendas, de empresas, de escuelas, de hospitales, de secciones censales; o un archivo de instituciones, asociaciones, empresas, usuarios-clientes; podría ser un mapa, el listado de teléfonos, un catálogo, todos los artículos de un diario en un período dado, los internautas de un portal, etc.

Siempre que sea posibles es conveniente disponer o construir el marco muestral con todas las unidades poblacionales, para así identificar inequívocamente a las unidades y no distorsionar una extracción aleatoria de la muestra. No obstante, no siempre es posible disponer del necesario marco muestral. En la investigación de mercados suele ser una preocupación recurrente ante la dificultad o imposibilidad de acceder a bases de datos con el listado de unidades de una determinada población, dado el carácter privado de esas organizaciones y por no disponer del acceso a determinada información de carácter personal de la ciudadanía y al amparo de la ley de protección de datos.

La constitución del marco muestral es una operación delicada que debe ser controlada a fin de determinar su corrección o detectar posibles errores, por omisiones o de cobertura. Cuando no es posible disponer de este marco se proponen procedimientos alternativos como las llamadas rutas aleatorias en el caso del muestreo por cuotas o la autoselección de las unidades (por ejemplo, los visitantes de una página web). Es muy importante se conscientes de las limitaciones y de los importantes sesgos que pueden producir una incorrecta selección de las unidades al no ser un buen reflejo de la población.

Se habla de marco de la muestra en sentido restringido cuando consideramos exclusivamente el listado de las unidades de donde se extraerá la muestra. Si consideramos además información complementaria de estas unidades para mejorar el diseño de la muestra se habla de marco de la muestra en sentido amplio.

La unidad identifica a la unidad estadística única que figura numerada e individualizada, pues cada unidad de muestreo debe ser única en relación a cada elemento de la población, que aparece en un registro, si existe, o que se puede construir, condición que teóricamente debería darse siempre con el fin de respetar los principios de la teoría de probabilidades y del muestreo. El conjunto de las unidades estadísticas o de muestreo determinan la base sobre la que se generalizan los resultados y sobre la que cabe plantear la representatividad en sentido estricto.

Pero puede suceder que la unidad estadística (de análisis) no coincida con la unidad muestral al considerar tipos de unidades distintas. Se llaman unidades simples (individuos) a la unidad elemental de muestreo, y compuestas a los grupos de unidades no solapadas: familias (hogares), zonas geográficas y, en general, conglomerados distintos. Por ejemplo, podemos realizar un muestreo de hogares entrevistando en una encuesta a la persona principal del hogar y obtener, a través de ella, información de todos los miembros del mismo; así, los individuos serían las unidades de análisis. En este sentido, también es posible establecer una jerarquía de unidades de muestreo, desde un primer nivel en que consideraríamos unidades elementales, un segundo nivel de grupos de unidades elementales, y sucesivamente agrupamientos de nivel superior. En general es recomendable que la unidad muestral esté lo más desagregada posible.

Se denomina talla de la muestra al número de individuos o elementos que comprende la unidad de la muestra, determinando el nivel de agregación muestral. El marco de la muestra puede ser de unidades elementales o de unidades compuestas, o incluso considerar en un mismo diseño de muestreo y en diferentes etapas marcos múltiples con diferentes niveles de agregación. En este sentido una muestra es la selección de un conjunto de unidades de uno más marcos muestrales.

1.7. Parámetros poblacionales y estadísticos muestrales

Inicialmente las características que quieren ser estudiadas están definidas por los objetivos de la investigación y por una reflexión teórica que ayudan a precisar y definir el fenómeno de interés de estudio. Así, por ejemplo, podríamos tener interés en conocer el comportamiento electoral de una población en las próximas elecciones, o bien saber las características de los potenciales usuarios de un servicio social del ayuntamiento, o bien determinar la audiencia potencial de un canal de televisión.

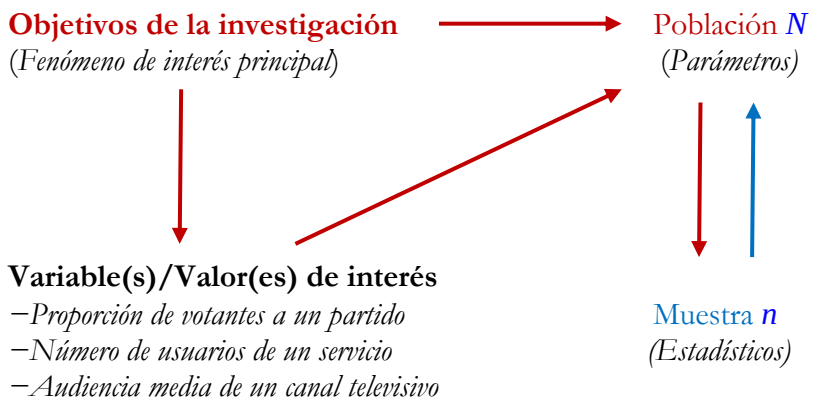
Si la información que disponemos hace referencia a toda la población, como es el caso de los datos censales o, por ejemplo, en el caso de disponer del escrutinio de una consulta electoral o de la información de todos los asociados/as de una organización, entonces tenemos un conocimiento exhaustivo y cierto de las características que queremos conocer: el porcentaje de la población ocupada o desocupada, la distribución de la población según su condición socioeconómica, el número de jóvenes emancipados, el porcentaje de votos que obtiene cada partido político en las elecciones, etc.

Esta información de toda la población constituye un valor cierto, es el resultado de contabilizar una característica de todos y cada uno de los individuos de esta población. Sin embargo, no es lo más habitual disponer de información exhaustiva de cada una de las unidades poblacionales, al contrario, la investigación social se nutre de estudios y datos que son el resultado de la selección de una parte de la población, una muestra estadística, fundamentalmente por el elevado coste económico, de tiempo y recursos materiales que supone llegar a tener información de cada unidad de la población. El objetivo final es el mismo, obtener información de determinadas características de la población, pero se opta por la extracción de una muestra aleatoria de un conjunto de unidades de esta población con las que se estima aquella característica poblacional.

Cuando procedemos de esta forma, decimos que hacemos estimaciones estadísticas (calculamos estadísticos muestrales, como una media o un porcentaje) de las características poblacionales (los mismos cálculos referidos a toda la población que se denominan parámetros poblacionales). Al ser estimaciones, los resultados o las medidas que obtenemos están afectadas por un determinado margen de error, por el hecho de disponer tan sólo de una parte de la población (la muestra) con la que se quiere conocer un rasgo para todo el conjunto poblacional. Las estimaciones más el margen de error constituyen el intervalo de confianza, como veremos más adelante y fundamentaremos en la tercera parte de este manual.

Nos referimos a conceptos que descansan en razonamientos estadísticos de inferencia. Esquemáticamente los expresamos mediante la representación del Gráfico II.4.2. Los parámetros poblacionales son los valores ciertos que caracterizan a la población y que se suelen expresar en términos de algún cálculo aritmético de una característica o variable que afecta a toda la población. Estos parámetros o características poblacionales se derivan de los objetivos y de la problemática de la investigación, y deben ser lo más precisos y centrales posible para la investigación. Se pueden concretar en la necesidad de conocer de forma precisa un parámetro concreto (porcentaje de intención de voto a un partido político) o bien la investigación está interesada en múltiples relaciones entre variables en las que no hay una única característica de interés (estudio de las condiciones de vida de la población).

Gráfico II.4.2. Relación entre parámetros poblacionales y estadísticos muestrales



En muestreo algunos de los parámetros más habituales son los siguientes[2]:

- μ: Parámetro que hace alusión a las medias de variables cuantitativas (por ejemplo, media de visitas diarias a un centro comercial, audiencia media de un medio de comunicación, el salario o los ingresos medios de un grupo social).
- P: Parámetro que indica porcentaje, proporciones o frecuencias relativas de variables cuantitativas o cualitativas (por ejemplo, proporción de los que ven cada día la televisión, porcentaje favorable a la despenalización del aborto, porcentaje de parados, de votantes a una opción política).

[2] Se utilizará como criterio general la identificación de las características, valores o parámetros poblacionales con letras mayúsculas o letras griegas. Las características, valores o estadísticos muestrales identificarán con letras minúsculas o letras latinas.

- X: Parámetro que indica la frecuencia absoluta o el total de casos de variables cuantitativas o cualitativas (por ejemplo, número de visitantes de un museo, número de personas que practican algún tipo de reciclaje o número de compradores de un segmento de mercado). Con un total se verifica que $X = N \times \mu$ o que $X = N \times P$.
- R: Parámetro referido a razones, es decir, cocientes entre dos cantidades (por ejemplo, la renta per cápita o el ingreso adulto equivalente).
- σ^2: Parámetro de variabilidad asociado a cada uno de los anteriores como es la varianza, un dato fundamental para determinar el error que se comete en las estimaciones muestrales[3].

La Tabla II.4.1 reproduce estos parámetros poblacionales expresados matemáticamente con su fórmula correspondiente, así como la de los estadísticos muestrales[4].

Tabla II.4.1. Principales parámetros poblacionales y estadísticos muestrales

	Parámetros poblacionales	Estadísticos muestrales
Media	$\mu = \frac{\sum_{i=1}^{N} X_i}{N}$	$\hat{\mu} = \bar{x} = \frac{\sum_{i=1}^{n} X_i}{n}$
Varianza	$\sigma^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N}$	$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Varianza ajustada	$S^2 = \frac{\sum_{i=1}^{N} (X_i - \mu)^2}{N-1}$	$\hat{S}^2 = s^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n-1}$
Proporción	$P = \frac{C}{N}$	$\hat{P} = p = \frac{c}{n}$
Porcentaje	$P = \frac{C}{N} \times 100$	$\hat{P} = p = \frac{c}{n} \times 100$

[3] La varianza nos indica con un valor cuál es el promedio (por ejemplo, de ingresos) de la diferencia entre los ingresos medios y los ingresos de los individuos que están tanto por encima de la media (los más ricos) como por debajo de la media (los más pobres). Entre unos y otros surge un valor promedio que resume el grado de variabilidad de la característica estudiada, en este caso, de los ingresos. Cuanto más variable es un fenómeno, menos homogéneo es el comportamiento social, y mayor será el error que podemos cometer al estimar un determinado valor en nuestra muestra. Por el contrario, si el fenómeno es poco variable, imaginemos por un momento que todos ingresáramos lo mismo, entonces la varianza sería cero, nadie es más rico ni más pobre, y a la hora de estimar los ingresos medios de una población no habría forma de equivocarse. Es más, ni siquiera sería necesario preguntarle a mucha gente, con una persona sería suficiente, sabiendo sus ingresos sabremos los de toda la población sin cometer ningún error. Cuando los ingresos varían mucho entre la población siempre existirá la posibilidad de cometer algún error a la hora de estimar el ingreso medio, pues nunca tendremos una representación exhaustiva de todos los posibles ingresos. Así, la varianza se convierte en un indicador del error que comentemos al realizar nuestras estimaciones, cuanto más variable ser un fenómeno mayor será el margen de error. Y ello, como veremos, se traducirá en la necesidad de tener un tamaño muestral mayor o menor.

[4] El lector/a que se aproxime por primera vez a estas expresiones necesita indagar en el conocimiento de los fundamentos de la estadística descriptiva e inferencial. En este capítulo realizaremos una aproximación básica e instrumental para el objetivo del diseño de la muestra. En el apartado III del Manual MISC, en particular en los capítulos III.3 y III.4, se explican los conceptos estadísticos de base de la formulación que aparece aquí.

Los estadísticos muestrales son los valores computados en la muestra que estiman los parámetros poblacionales. Es decir, se corresponden con los mismos conceptos, pero expresados y obtenidos con los datos de la muestra. De nuevo, pues, tenemos: la media muestral, la proporción o porcentaje muestral, el total muestral y la varianza muestral. La estimación de estos estadísticos en la muestra implica, para afirmar con una parte una característica del todo, tener asociado un nivel de error. Este error se puede determinar matemáticamente a partir de considerar las estimaciones como variables aleatorias que varían de una muestra a otra, siguiendo una forma de distribución conocida. Es decir, que cada muestra que podamos extraer de una misma población de tamaño n da lugar a diferentes valores que estiman el verdadero valor poblacional. El conjunto de todos estos posibles valores de un estadístico, por ejemplo, una media o una proporción, configura una distribución muestral del estadístico considerado, que tiene unas características matemáticas conocidas que nos ayudarán a determinar el error cometido en la estimación[5].

1.8. Error muestral

El error muestral e de las estimaciones nos mide el grado de exactitud o de precisión con el que inferimos de la muestra a población. Este valor, como sugerimos anteriormente, vendrá determinado por la variabilidad del estadístico, es decir, que el error se cuantifica mediante las varianzas del estadístico considerado en cada caso. Como veremos esta variabilidad se denominará error típico del estadístico, y se corresponde con un cálculo que depende de la varianza y del tamaño de la muestra, además de otra característica que es el nivel de confianza.

El error muestral se define como la divergencia entre los valores de los estadísticos obtenidos en la muestra de una variable y los existentes en la población. Si consideráramos el estadístico de la media, el error muestral sería la diferencia entre el valor muestral de la media ($\bar{x}$) y el valor poblacional de la media (μ), es decir, la diferencia: $e = \bar{x} - \mu$. Esta es una definición de lo real pero no conocido, ya que el parámetro poblacional es desconocido, justamente realizamos un estudio por muestreo para saberlo, situación aparentemente contradictoria. Para solucionar este interrogante se razona en términos probabilísticos a partir del conocimiento de la forma de la distribución del estadístico como tendremos ocasión de explicar con más detalle en el capítulo III.4.

En este sentido, cualquier afirmación para dar cuenta del valor de un parámetro poblacional desconocido a partir del valor del estadístico que lo estima, tendrá un grado de desviación, de variación al alza o de variación a la baja respecto del verdadero valor poblacional, esa variación es la que mide el error muestral o margen de error. De ahí se deriva una relación básica del muestreo y de las estimaciones estadísticas:

Valor poblacional = **Valor estimado en la muestra** ± **Error de muestreo**
(**Parámetro**) (**Estadístico**)

[5] Cuando procedemos a realizar estimaciones a través de los estadísticos formalmente se exige que sean no sesgadas, es decir, que el estimador por sí mismo no produzca desviaciones, que el valor esperado en la muestra se corresponda con el valor poblacional. Además, las estimaciones deben ser consistentes, es decir, que cuando reemplazamos el tamaño n de la muestra por el tamaño N de la población se reproduzca el parámetro poblacional considerado.

Cuando a una estimación puntual de un estadístico le sumamos y restamos el error de muestreo configuramos un intervalo de posibles valores y realizamos una estimación por intervalos de confianza. Más adelante veremos cómo calcular o aproximar el margen de error asociado a cada estimación. Pero veamos un ejemplo para clarificar qué significa asumir un margen de error y determinar un intervalo de confianza. Situémonos en el contexto del ejemplo de las elecciones al rectorado y supongamos que la candidatura A tiene una estimación puntual del 25% de los votos y que el margen de error asociado a la estimación del porcentaje fuera del 3%. Esto significaría que el valor poblacional se situaría en un intervalo que resulta de sumar y restar este margen error a la estimación del porcentaje. Así, afirmaríamos que la candidatura A obtendrá unos resultados que se situarán en una horquilla definida por:

$$25\% \pm 3\%$$

es decir, en un intervalo entre el

$$22\% \text{ y el } 28\%.$$

y que escribimos así (22%,28%).

Cualquier valor puntual estimado es una aproximación al valor poblacional desconocido afectado por el error muestral, por lo tanto, esto no significa que el porcentaje sea estrictamente del 25% en nuestro ejemplo, sino que sólo podemos establecer un intervalo de valores, entre un mínimo y un máximo, en cuyo interior se situará, probablemente el parámetro poblacional. Y decimos probablemente porque, como veremos, el afirmar sobre el intervalo se establece con un grado de probabilidad llamado nivel de confianza, habitualmente del 95%, lo que significa que en el 5% de los casos nos podemos equivocar, si bien es un nivel suficientemente bajo de incertidumbre que garantiza la bondad de nuestras estimaciones. En consecuencia, tenemos dos fuentes de incertidumbre, la del error de la muestra que hemos tomado (el candidato puede obtener entre el 22% y el 28% de los votos) y la probabilidad de equivocarnos en esa estimación, valorada en un 5%, de que el valor del parámetro no se encuentre en el intervalo que hemos definido.

El margen de error se puede expresar en términos absolutos o relativos. El error absoluto, el que se suele presentar en los estudios, es el que acabamos de comentar respecto de un porcentaje, y que se puede aplicar igualmente en el caso de una media, en este caso se expresaría en las unidades de medida propias de la variable. El error relativo es adimensional y se expresa como tanto por ciento de error sin la unidad de medida y facilita la comparación entre muestras distintas. El error relativo de una proporción es:

$$er_p = \frac{error\ absoluto}{proporción} \times 100 \qquad \textit{Ecuación 3}$$

El error relativo de una media es:

$$er_{\bar{x}} = \frac{error\ absoluto}{media} \times 100 \qquad \textit{Ecuación 4}$$

Por ejemplo, una proporción de 0,25 (o porcentaje del 25%) con un error absoluto del 0,03 (o del 3%) tiene un error relativo de 0,12 (o del 12%).

El error muestral es un aspecto clave de los estudios por muestreo que debemos considerar en dos momentos de la investigación:

a) antes de proceder a obtener la muestra y de proceder a la recogida de información, en el que hay que determinar a priori qué nivel de error estadístico estamos dispuestos a asumir, junto a la decisión del tamaño de la muestra, y
b) después de obtener la muestra, para determinar el error asociado a cada estimación o prueba estadística de hipótesis que podamos plantear.

En este sentido, y como situación frecuente de particular interés en la investigación, conviene tener presente la necesidad, que hay que prever como objetivo del diseño muestral para garantizar un error muestral suficiente, qué tipo de análisis se desea realizar: si son estimaciones globales del conjunto poblacional o bien se desean realizar análisis específicos y autónomos de colectivos sociales o zonas geográficas. Considerar estas subpoblaciones de forma separada de la muestra total exige trabajar con submuestras que deben tener un tamaño suficiente, pues los errores muestrales aumentan al reducirse el número de casos y, en consecuencia, se pierde validez en los resultados. En este sentido, pues, será posible realizar ese tipo de análisis siempre que se haya previsto en el diseño un tamaño muestral suficiente que garantice errores muestrales aceptables.

¿Cuánto es un error muestral aceptable? Si lo expresamos en términos porcentuales, teniendo en cuenta en particular que los intervalos se construyen sumando y restando el valor del error, sería deseable alcanzar niveles del 2%. No siempre es posible alcanzar ese margen de error pues para ello se requiere un tamaño muestral relativamente elevado y, por ello, costes económicos no siempre asumibles. En la medida de lo posible convendría no superar el 3%, es decir, un ± 3%, lo que implica intervalos de 6 puntos porcentuales. Por otra parte, el error tiene implicaciones directas sobre el tamaño de la muestra y, por tanto, sobre los costes. ¿Cuál sería el tamaño mínimo aceptable? La respuesta es relativa, depende del tipo de estudio, de la naturaleza del fenómeno estudiado, de la facilidad de acceso a las unidades, etc. Podemos señalar, no obstante, en un contexto de estudios por encuesta, la recomendación de no bajar nunca de las 800 entrevistas; como veremos más adelante, bajo determinados parámetros que se tienen en cuenta, este número de individuos se correspondería con un 3,5% de error muestral, un valor que podemos considerar alto, por tanto, relativamente impreciso.

1.9. Intervalo de confianza, nivel de confianza y nivel de significación

Hemos comentado que las estimaciones de un parámetro de la población se construyen a partir de intervalos de confianza cuando se considera el nivel de error que se comete. Teniendo en cuenta el estadístico de la muestra y el error típico las estimaciones se expresan en términos de desviación o variación que contempla el intervalo. Estas estimaciones por intervalo implican una afirmación probabilística que determina el nivel de confianza de una conclusión estadística. Así el nivel de confianza se define como la probabilidad de obtener el valor poblacional a partir de la muestra.

Si decimos que el nivel de confianza es del 95% (o 0,95, en tanto por uno), estamos diciendo que para un tamaño muestral y desviación obtenidas, el parámetro poblacional se situará el 95% de las veces (o de las muestras) en un intervalo de valores que determina el valor del estadístico (o estimación puntual) sumándole y restándole

el error muestral e. En el caso de la media, este es un intervalo definido por $\bar{x} \pm e$, es decir, $(\bar{x} - e, \bar{x} + e)$. Si el estadístico sigue una distribución muestral estadística normal[6], considerar un nivel de confianza del 95% implica afirmar que el 95% de los casos se sitúa entre los valores típicos −1,96 y 1,96, llamados límites de confianza.

El intervalo de confianza es, por tanto, el conjunto de valores donde se encontrará el valor del parámetro poblacional con una probabilidad dada y un margen de error, y nos da una indicación de la exactitud de nuestra estimación.

De forma complementaria se entiende por nivel de significación a la probabilidad de obtener un valor extremo del estadístico que estima el parámetro poblacional. Si el nivel de confianza es del 95%, el de significación es del 5% (o 0,05, en tanto por uno).

1.10. Error sistemático

Tan importante como el error estadístico, o más, es el error sistemático, es decir, toda aquella fuente de error no estadística que puede derivarse resultado de cualquier actuación incorrecta a lo largo de todo el proceso de investigación, en particular, por ejemplo, como resultado de un error de selección de la muestra o como resultado de una inadecuada medida. Desde este punto de vista, hay que tener presente otra relación fundamental:

Error total = Error muestral ± Error sistemático

En la investigación social es tan importante o más el error sistemático que el error muestral. El error muestral es un error estadístico, aleatorio, que nos proporciona una medida cuantitativa básica al poner en relación la muestra y la población, variable como veremos según el tamaño de la muestra, la variabilidad de lo que estimamos, o incluso del método de muestreo. En general, disponiendo de un tamaño muestral suficiente obtendremos un error muestral aceptable que nos garantizará una fiabilidad o precisión adecuadas de los resultados de un estudio. Pero por muy reducido que sea este error muestral de nada nos servirá si introducimos otras fuentes de error de tipo no estadístico o no estrictamente muestrales, de tipo sistemático.

Las fuentes de error sistemático se refieren a errores de medida y se pueden introducir en todo el proceso de investigación. Son fuentes de error sistemático:

- La elección de indicadores de los conceptos no adecuados.
- La inadecuada selección de la población y de las unidades. En particular, se denomina error de cobertura al error que se deriva por el hecho de no disponer de listas completas, actualizadas y correctas de las unidades poblacionales a partir de las cuales efectuar la extracción aleatoria de las unidades que compondrán la muestra. El caso de autoselección de las personas que forman la muestra es una situación que induce fácilmente a este tipo de error.
- La recogida de datos por encuesta se enfrenta con el problema del error de no respuesta, que se produce cuando no ha sido posible el contacto con la persona

[6] Es un concepto que se explicará en el tema III.4 y hace referencia a un comportamiento teórico de los estadísticos que permite conocer su comportamiento en cualquier tipo de muestra y determinar tanto el nivel de error como los intervalos de confianza. Una de las distribuciones teóricas más importantes es la normal.

seleccionada (a veces resultado de considerar que es demasiado costoso llegar a contactar) o bien se produce el rechazo de la persona a realizar la entrevista.
- Derivada del anterior se produce la incorrecta sustitución de los rechazos al realizar la entrevista.
- La pérdida de datos.
- La incorrecta consignación de las respuestas, la incorrecta codificación o registro no controlado de los datos en soporte informático.
- Las preguntas mal formuladas (sesgadas) o formuladas en un orden, con determinadas palabras, sobre determinadas características de la vida personal, etc.
- La ausencia de información por no respuesta a las preguntas del cuestionario.
- La inadecuada respuesta del entrevistado (por no recuerdo, incorrecta compresión, respuesta estereotipada o socialmente deseable, respuestas engañosas).
- Las perturbaciones o sesgos introducidos por el entrevistador/a, derivadas de una formación insuficiente, de unas condiciones de trabajo, de provocar un efecto de error no deseado en formular las preguntas, etc.

Alcanzar un bajo nivel de error muestral y de error sistemático son dos objetivos que se plantean en toda investigación por muestreo. Habrá que valorar ambos tipos de errores estableciendo un equilibrio óptimo en función de los recursos y del contexto de cada investigación.

1.11. Tipos de muestreo

El diseño de una muestra puede seguir estrategias distintas que identifican diferentes tipos de muestreo. Una primera distinción fundamental es la que los clasifica en función de si son probabilísticos o no. Nuestro mayor interés en este texto es con relación a los primeros. Daremos cuenta de ellos en los apartados siguientes, pero presentamos simplemente la relación de tipos de muestreo.

a) Muestreo probabilístico. Aquel muestreo en que, de forma estricta, todas las unidades de la población tienen una probabilidad conocida de ser incluidas en la muestra, y, por lo tanto, también se conoce la probabilidad de obtener cada una de las muestras mediante un procedimiento de aleatorización. En esta categoría se encuentran:
   - Muestreo aleatorio simple
   - Muestreo sistemático
   - Muestreo estratificado
   - Muestreo por conglomerados
   - Muestreo polietápico

   En el muestreo probabilístico, además, se puede considerar una distinción cuando los diferentes elementos de la muestra o bien son idénticos o similares y tienen la misma probabilidad de ser elegidas, o bien esta probabilidad es diferente. Así se diferencia entre muestreo con probabilidades iguales y muestreo con probabilidades desiguales. También el muestreo tiene implicaciones en función de si se trata de muestreo con reposición o sin reposición (o reemplazamiento).

b) Muestreo no probabilístico. En este caso no se conocen las probabilidades de cada unidad de muestreo de pertenecer a la muestra. Cabe contemplar en esta categoría:
   - Muestreo por cuotas
   - Muestreo casual o incidental
   - Muestreo de conveniencia
   - Muestreo intencional o razonado
   - Muestreo de bola de nieve

1.12. Tareas implicadas en el diseño de una muestra

En el diseño de una muestra cabe contemplar las siguientes tareas relevantes:

1) Ante todo, la definición de los objetivos de investigación, la construcción de un modelo de análisis y determinación de los recursos disponibles.
2) La precisión de las variables y parámetros poblacionales a medir o que expresan el centro de interés de la investigación.
3) La delimitación de la población objetivo y de la población marco. Definición de las unidades de muestreo y medidas. Delimitación espacial y temporal.
4) Constituir la base de la muestra o marco de muestreo, cuando sea posible.
5) Elección del tipo de muestreo.
6) Determinar el tamaño de la muestra, con un nivel de confianza y error muestral dados.
7) Extracción aleatoria de la muestra.
8) Organización del desarrollo del trabajo de campo, recogida de información y seguimiento.
9) Validación y ajuste de la muestra.

2. Métodos de muestreo probabilístico

2.1. Muestreo aleatorio simple

El muestreo aleatorio simple (MAS) es el tipo de muestreo más sencillo, pero fundamental pues constituye la técnica muestral básica de la estadística inferencial de donde se derivan las demás y con la que se comparan los demás métodos.

Una muestra aleatoria simple se define como aquella donde las unidades se seleccionan o extraen aleatoriamente cumpliendo estas condiciones:

1) Cada unidad de la población tiene idéntica probabilidad de ser incluida en la muestra: $\text{Prob}(U_i) = \frac{1}{N}$
2) Cada combinación de unidades, es decir, cada muestra posible de tamaño n que se puede seleccionar tiene igual probabilidad de constituirse (condición de

equiprobabilidad). En total existen múltiples combinaciones de posibles muestras donde se seleccionan n casos entre una población de N[7].

De estas dos condiciones se deriva que la probabilidad de una unidad cualquiera U_i de pertenecer a una muestra será n/N. En este caso hemos supuesto una extracción de unidades sin reemplazamiento. Cuando existe reemplazamiento una misma unidad puede salir más de una vez en la selección y, en cada caso, todas las unidades tienen una probabilidad $1/N$ de aparecer en cada extracción.

Una muestra aleatoria, extraída al azar, no significa una muestra extraída casualmente o de forma azarosa, sino que se establecen determinadas condiciones para considerarla extracción al azar. Garantizar estas condiciones es de suma importancia.

Para extraer una muestra aleatoria tradicionalmente se disponía de tablas en papel con una relación de números dispuestos al azar y que determinaban las unidades a extraer. Hoy en día el software estadístico genera los números aleatorios. El procedimiento de extracción consiste en cuatro pasos simples:

1. Se numeran de 1 a N correlativamente todos los individuos de la población en la base de datos de la muestra.
2. Se determina el tamaño de la muestra n.
3. Se generan los n números aleatorios.
4. Los n números se extraen del listado de la base de datos y constituyen la muestra.

En el contexto del MAS podemos realizar las estimaciones de los estadísticos muestrales que vimos en la Tabla II.4.1. En capítulo III.4 de detallan los fundamentos de la utilización de la estadística descriptiva e inferencial en el contexto del muestreo. Aquí recuperaremos algunas ideas generales que nos servirán para plantear dos tipos de tareas implicadas en el análisis de la información cuantitativa obtenida por muestreo: la determinación del tamaño de la muestra y la determinación del error asociado a una estimación una vez se ha obtenido la muestra. La primera tarea es el aspecto más importante del diseño de muestras, la segunda, aplicando razonamientos similares, es una tarea propia del análisis posterior de los datos que veremos reiteradamente a lo largo del texto; aquí simplemente se apuntarán algunas ideas generales de introducción.

2.1.1. *Determinación del tamaño de la muestra*

El objetivo de una muestra está en alcanzar la mayor representatividad o precisión posible en la estimación de los parámetros poblacionales. En otras palabras, y refiriéndonos en general al diseño de encuestas por muestreo, se trata de reducir el

[7] En total existen múltiples combinaciones de posibles muestras donde se seleccionan n casos entre una población de N. El número de posibles muestras es C_N^n, es decir, combinaciones de N elementos tomados de n en n, siendo, por tanto, la probabilidad de cada una de ellas: $\frac{1}{C_N^n} = \frac{N!}{n!(N-n)!}$.

error muestral dadas unas limitaciones de tiempo, dinero y trabajo, lo que se traduce siempre en la decisión de fijar el número de unidades de la muestra: cuántas encuestas se realizarán.

Por la ley de regularidad estadística se sabe que, a partir de un determinado número de unidades, los valores tienden a estabilizarse, nuevos elementos en la muestra aumentan cada vez en menor cuantía la fiabilidad y disminuyen poco el error, hecho que hace innecesario el aumento del tamaño a partir de un determinado momento. Se trata por tanto de encontrar el equilibrio entre la fiabilidad deseada, los objetivos de la investigación y los costes en tiempo, dinero y trabajo.

En la determinación del tamaño muestral se conjugan estos cuatro elementos que intervendrán en la fórmula de cálculo:

1. La amplitud del universo, diferenciando dos situaciones: si la población es finita (si tienes menos de 100.000 individuos) o es infinita (a partir de 100.000 individuos). Con poblaciones finitas el tamaño muestral tiende a ser cada vez más sensible al tamaño poblacional por lo que es necesario introducir un factor de corrección por finitud e igual a $\sqrt{\frac{N-n}{N-1}}$. Cuando la población es muy numerosa la diferencia con respecto al número de muestra es muy pequeña y el numerador de ese factor tiene a igualarse al denominador, por lo que el factor tiende a ser 1; y tiende a ser un valor cada vez mayor a medida que población y muestra se aproximan en número de casos.
2. El nivel de confianza adoptado. Como comentamos en el primer apartado el nivel de confianza establece la probabilidad o confiabilidad de nuestros resultados, los cuales se elaboran y razonan en términos probabilísticos. El criterio habitual que se sigue es considerar un nivel de confianza del 95,5%, lo que implica considerar 2σ, es decir, un valor de dos unidades de desviación a partir de la media en la distribución normal, que se puede expresar también diciendo que $z=2$ (valor tipificado de la distribución normal). Este es el criterio que fija el valor de la fórmula de determinación de la muestra que veremos seguidamente, donde simplemente substituiremos el valor de z por un 2. También se puede elegir el 95%, en ese caso el número de unidades sigma de desviación sería de $1{,}96\sigma$, es decir $z=1{,}96$. Si quisiéramos mayor confianza, por ejemplo, del 99,7%, el número de unidades de desviación sería de 3σ, o $z=3$.

3. El error muestral e asociado al estadístico elegido de estimación. Veremos sobre todo estimaciones de medias y proporciones (o porcentajes) y en cada caso los cálculos tendrán sus especificidades[8].
4. La varianza (o desviación típica) de la población: σ^2. Cuando se estiman medias es la fórmula habitual de cálculo de la varianza de una variable, y puede ser un dato disponible o estimado. Cuando se estiman proporciones el valor de la varianza es

[8] En general el error muestral e se obtiene multiplicando el número de unidades de desviación de la distribución muestral del estadístico, el valor z en el caso de la distribución normal, por el error típico del estadístico. En el caso de la media: $e = z \cdot \sigma_{\bar{x}}$ y en el caso de la proporción $e = z \cdot \sigma_p$.

igual a: $P \times (1-P) = P \times Q$, es decir, la proporción (o porcentaje) P asociado a una estimación (por ejemplo, la proporción de desempleados) multiplicado por Q que es el complementario de P (la proporción de los que no están desempleados), que es $1-P$ ($100-P$ si fueran porcentajes)[9].

En el contexto del muestreo aleatorio simple las fórmulas de determinación del tamaño de la muestra n, teniendo en cuenta si se estima una media o una proporción, y teniendo en cuenta si se estudia una población finita o infinita, se presentan en la Tabla II.4.2. En las fórmulas aparecen los símbolos siguientes:

z^2: el número de unidades de desviación que indica el nivel de confianza adoptado, elevado al cuadrado.
σ^2: la varianza de la variable cuantitativa sobre la que se calcula la media.
e^2: el error muestral considerado, elevado al cuadrado.
N: el tamaño de la población.
P: la proporción (o porcentaje) de individuos que tienen una característica.
Q: la proporción (o porcentaje) de individuos que no tienen la característica.

Tabla II.4.2. Fórmulas de determinación del tamaño de la muestra según el tipo de población y el parámetro estimado, dado un error muestral

Tamaño en función del error		**Población**	
		Infinita	Finita
Parámetro	Media	$n = \frac{z^2 \times \sigma^2}{e^2}$	$n = \frac{z^2 \times \sigma^2 \times N}{(N-1) \times e^2 + z^2 \times \sigma^2}$
	Proporción	$n = \frac{z^2 \times P \times Q}{e^2}$	$n = \frac{z^2 \times P \times Q \times N}{(N-1) \times e^2 + z^2 \times P \times Q}$

Como se puede observar en las fórmulas, los valores del tamaño y del error están en relación inversa (en el numerador y el denominador) y son dos valores de la fórmula con los que se puede "jugar" hasta encontrar el equilibrio entre precisión y costes: podemos fijar el tamaño y ver el error asociado, o, inversamente, fijar el error y determinar el tamaño asociado. El valor de las unidades de desviación z lo fijamos por convención en niveles del 95,5%, es decir, con z=2 (o bien z=1,96).

En todos los casos, finalmente, debemos especificar el valor de la varianza, ya sea σ^2 en el caso de una media o $P \times Q$ en el caso de una proporción. Pero estos valores hacen referencia a la varianza poblacional, lo que nos plantea el siguiente dilema: por un lado, estoy planteando es un estudio para conocer las características relacionadas con una determinada variable y, por otro lado, para determinar el tamaño de la muestra necesito

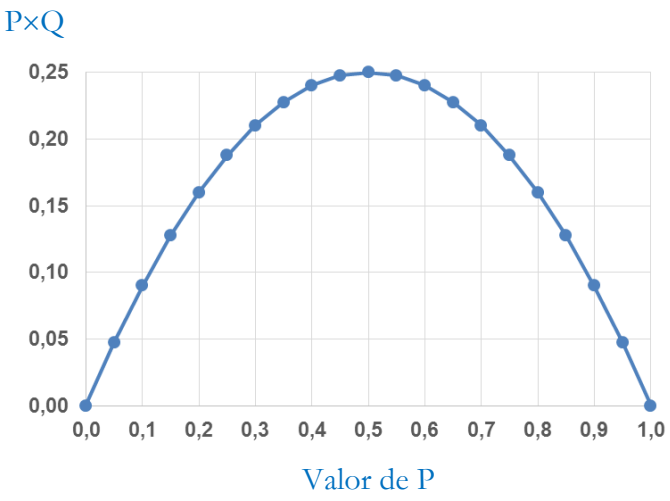
[9] La explicación y demostración de este resultado se puede consultar en el capítulo III.4. Surge de considerar a una proporción (o porcentaje) como una variable dicotómica codificada con 0 y 1. Los dos valores serán: 1 si tiene la característica (estar desempleado) y 0 si no la tiene (no estar desempleado). La media de una variable como esta es P y su varianza es el producto PQ.

conocer primero los datos poblacionales de la variable que quiero estudiar. En consecuencia, si disponemos de esa información para qué necesitamos realizar el estudio. Bien, la cuestión merece dos comentarios. Por una parte, efectivamente, no tendría sentido; de hecho, desconocemos casi siempre el dato poblacional, lo que nos obliga a buscarlo o a estimarlo, aproximándonos a él a través de información externa como un censo, a través de investigaciones anteriores o *adhoc*, considerando variables parecidas muy correlacionadas con la de interés, realizando conjeturas o mediante pruebas piloto. Por otra parte, aún si conociéramos el dato de la varianza de la variable, los estudios no se limitan simplemente a conocer el comportamiento de una variable, sino que se recoge información múltiple con un número elevado de variables distintas que justifican el estudio. En esos casos la variable de la que se precisa conocer la varianza actúa de variable central o sintética de la naturaleza del estudio realizado y permite diseñar la muestra teniendo en cuenta sus características.

Este dilema se plantea tanto para la estimación de medias como de proporciones. En el caso de las medias no queda más remedio que proceder como hemos indicado. Pero en el caso de las proporciones existe una solución muy práctica que resuelve el problema del conocimiento previo de la varianza. La solución surge del comportamiento de los valores posibles de la varianza de una proporción, del producto de P por Q. Sabemos que este producto alcanza un valor máximo, por tanto, proporciona un error máximo. Para cualquier par P y Q, siempre se cumple que: $P \times Q \leq 1/4$, es decir, siempre es menor o igual como máximo a 0,25, situación que se produce cuando $P=0,5$ y $Q=0,5$. Este resultado se puede ilustrar gráficamente a partir del cálculo de todos los posibles valores del producto de P por Q (Gráfico II.4.3). El máximo valor de todas las combinaciones posibles de P y Q se alcanza cuando $P=0,5$ y $Q=0,5$, cuyo producto $P \times Q=0,25$, todos los demás valores son inferiores, es decir, la varianza, y el error que comporta, siempre son inferiores. Si en lugar de proporciones utilizamos los porcentajes el valor máximo sería de 2.500, es decir, con $P=Q=50\%$.

Gráfico II.4.3. Determinación del valor máximo de P×Q

P	Q	P×Q
0,00	1,00	0,0000
0,05	0,95	0,0475
0,10	0,90	0,0900
0,15	0,85	0,1275
0,20	0,80	0,1600
0,25	0,75	0,1875
0,30	0,70	0,2100
0,35	0,65	0,2275
0,40	0,60	0,2400
0,45	0,55	0,2475
0,50	0,50	0,2500
0,55	0,45	0,2475
0,60	0,40	0,2400
0,65	0,35	0,2275
0,70	0,30	0,2100
0,75	0,25	0,1875
0,80	0,20	0,1600
0,85	0,15	0,1275
0,90	0,10	0,0900
0,95	0,05	0,0475
1,00	0,00	0,0000



Cuando se considera esta situación extrema se dice que estamos en la situación de máxima incertidumbre o de máxima indeterminación. Este supuesto es el que se asume también habitualmente en los sondeos electorales y encuestas sociológicas en el momento de determinar el tamaño de la muestra.

En las fórmulas de la Tabla II.4.2. se ha considerado el tamaño de las muestras en función del resto de parámetros de la fórmula. Alternativamente, y más práctico, se puede calcular el error que se comete dados el resto de parámetros de la fórmula, fijando en particular el tamaño de la muestra y viendo el error asociado. Las fórmulas son, entonces, las de la Tabla II.4.3.

Tabla II.4.3. Fórmulas de determinación del error de la muestra según el tipo de población y el parámetro estimado, dado un tamaño muestral

Error en función del tamaño		**Población**	
		Infinita	Finita
Parámetro	Media	$e = z \times \sqrt{\frac{\sigma^2}{n}}$	$e = z \times \sqrt{\frac{\sigma^2}{n} \times \frac{N-n}{N-1}}$
	Proporción	$e = z \times \sqrt{\frac{P \times Q}{n}}$	$e = z \times \sqrt{\frac{P \times Q}{n} \times \frac{N-n}{N-1}}$

Aplicaremos lo visto para calcular el tamaño de la muestra que habitualmente se plantea en los sondeos de opinión y electorales. Un sondeo elaborado por el Instituto Opina con motivo de las elecciones al Parlamento de Cataluña de 2003, publicado en el diario El País, donde se adjuntaba la ficha técnica siguiente:

ELECCIONES EN CATALUÑA *Sondeo de Opina para EL PAÍS*

FICHA TÉCNICA

La encuesta ha sido realizada por el INSTITUTO OPINA. **Realización del trabajo de campo:** 31 de octubre, 1 y 2 de noviembre de 2003. **Ámbito geográfico:** Cataluña. **Recogida de la información:** mediante entrevista telefónica. **Universo de análisis:** población mayor de 18 años residente en hogares con teléfono. **Error muestral:** El margen de error para el total de la muestra es de ±2,14% para un margen de confianza del 95% y bajo el supuesto de máxima indeterminación (p=q=50%). **Procedimiento de muestreo:** selección polietápica del entrevistado: unidades primarias de muestreo (MUNICIPIOS) seleccionadas de forma aleatoria proporcional para cada provincia. unidades secundarias (HOGARES) mediante la selección aleatoria de números de teléfono. Unidades últimas (INDIVIDUOS) según cuotas cruzadas de SEXO, EDAD y RECUERDO DE VOTO GENERALES 2000.

Tamaño de la muestra: 2.100 entrevistas. 900 en la provincia de Barcelona (290 en Barcelona ciudad, 256 en el área metropolitana, 354 en el resto de la provincia). **400 en la provincia de Girona** (53 en Girona ciudad, 347 en el resto de la provincia). **400 en la provincia de Lleida** (126 en Lleida ciudad, 274 en el resto de la provincia). **400 en la provincia de Tarragona** (78 en Tarragona ciudad, 322 en el resto de la provincia).

Ponderación: el total de la muestra se ha ponderado para otorgar a cada una de las provincias su peso real dentro del conjunto de la población de Cataluña. **De 900 en la provincia de Barcelona a 1.594** (514 en Barcelona ciudad, 453 en el área metropolitana, 627 en el resto de la provincia). **De 400 en la provincia de Girona a 185** (25 en Girona ciudad, 160 en el resto de la provincia). **De 400 en la provincia de Lleida a 122** (38 en Lleida ciudad, 84 en el resto de la provincia). **De 400 en la provincia de Tarragona a 199** (39 en Tarragona ciudad, 160 en el resto de la provincia).

Se puede leer que se ha realizado una encuesta con una muestra de 2.100 personas mayores de 18 años en Cataluña. El margen de error ha sido del 2,14% para un nivel de confianza del 95% (z=1,96) y bajo el supuesto de máxima indeterminación (P=Q=50%). Con estos datos, el tamaño de la muestra es el

resultado de aplicar la fórmula del cálculo del tamaño muestral de una población infinita pues los electores superan los 100.000 individuos:

$$n = \frac{z^2 \times P \times Q}{e^2} = \frac{1{,}96^2 \times 0{,}5 \times 0{,}5}{0{,}0214^2} \cong 2100$$

El mismo resultado se obtiene si hacemos los cálculos en porcentajes:

$$n = \frac{z^2 \times P \times Q}{e^2} = \frac{1{,}96^2 \times 50 \times 50}{2{,}14^2} \cong 2100$$

Los valores obtenidos con decimales siempre se redondean al entero superior.

2.1.2. *Tablas para la determinación del tamaño de la muestra*

A continuación, aparecen varias tablas destinadas a facilitar la determinación del tamaño de la muestra sin tener que realizar manualmente los cálculos a través de las expresiones que hemos visto. Se presentan las tablas de tamaño de la muestra en función del error muestral, tanto para poblaciones finitas como infinitas, y también del error en función del tamaño de la muestra.

Estas tablas de cálculo del tamaño y del error se adjuntan en un archivo Excel que se puede encontrar en la página web del manual con el nombre FórmulasMAS.xlsx. Igualmente se incluye una primera pestaña con todas las fórmulas de la Tabla II.4.2 y Tabla II.4.3**¡Error! No se encuentra el origen de la referencia.** que facilitan el cálculo automático del tamaño y del error introduciendo los valores de los diferentes parámetros en el caso de la estimación de proporciones:

Muestreo aleatorio simple (MAS)
Fórmulas para determinar el tamaño de la muestra
Estimación de proporciones (porcentajes)

	Poblaciones finitas $n=\frac{z^2 \cdot PQ \cdot N}{(N-1)\cdot e^2 + z^2 \cdot PQ}$ $e = z \cdot \sqrt{\frac{PQ}{n} \cdot \frac{N-n}{N-1}}$			Poblaciones infinitas $n=\frac{z^2 \cdot PQ}{e^2}$ $e = z \cdot \sqrt{\frac{PQ}{n}}$	
Nivel de confianza	z	2		z	2
Error muestral (%)	e	2,00		e	3,00
Tamaño de la población	N	10.000		n	1.111
Tamaño de la muestra	n	2.000		P	50
Porcentaje	P	50		$Q=100-P$	50
Porcentaje complementario	$Q=100-P$	50			
Si el error es del: 2,00	n?	2.000	3,00	n?	1.111
Si el tamaño de muestra es de: 2.000	e?	2,00	1.111	e?	3,00

Número de unidades de la muestra (n), con una población finita (N<100.000), para un nivel de confianza del 95,5% (z=2) y bajo el supuesto de máxima indeterminación (P=Q=50%)

$$n = \frac{z^2 \cdot PQ \cdot N}{(N-1) \cdot e^2 + z^2 \cdot PQ}$$

	Margen de error (e)										
Población (N)	**1,0**	**1,5**	**2,0**	**2,5**	**3,0**	**3,5**	**4,0**	**4,5**	**5,0**	**7,5**	**10,0**
250	244	237	227	216	204	191	179	166	154	104	71
500	476	449	417	381	345	310	278	248	222	131	83
1.000	909	816	714	615	526	449	385	331	286	151	91
1.500	1.304	1.121	938	774	638	529	441	372	316	159	94
2.000	1.667	1.379	1.111	889	714	580	476	396	333	163	95
2.500	2.000	1.600	1.250	976	769	615	500	412	345	166	96
3.000	2.308	1.791	1.364	1.043	811	642	517	424	353	168	97
3.500	2.593	1.958	1.458	1.098	843	662	530	433	359	169	97
4.000	2.857	2.105	1.538	1.143	870	678	541	440	364	170	98
4.500	3.103	2.236	1.607	1.180	891	691	549	445	367	171	98
5.000	3.333	2.353	1.667	1.212	909	702	556	449	370	172	98
6.000	3.750	2.553	1.765	1.263	938	719	566	456	375	173	98
7.000	4.118	2.718	1.842	1.302	959	731	574	461	378	173	99
8.000	4.444	2.857	1.905	1.333	976	741	580	465	381	174	99
9.000	4.737	2.975	1.957	1.358	989	748	584	468	383	174	99
10.000	5.000	3.077	2.000	1.379	1.000	755	588	471	385	175	99
15.000	6.000	3.429	2.143	1.446	1.034	774	600	478	390	176	99
20.000	6.667	3.636	2.222	1.481	1.053	784	606	482	392	176	100
25.000	7.143	3.774	2.273	1.504	1.064	791	610	484	394	177	100
30.000	7.500	3.871	2.308	1.519	1.071	795	612	486	395	177	100
35.000	7.778	3.944	2.333	1.530	1.077	798	614	487	395	177	100
40.000	8.000	4.000	2.353	1.538	1.081	800	615	488	396	177	100
45.000	8.182	4.045	2.368	1.545	1.084	802	616	488	396	177	100
50.000	8.333	4.082	2.381	1.550	1.087	803	617	489	397	177	100
100.000	9.091	4.255	2.439	1.575	1.099	810	621	491	398	177	100

Número de unidades de la muestra (n), con una población finita (N<100.000), para un nivel de confianza del 99,7% (z=3) y bajo el supuesto de máxima indeterminación (P=Q=50%)

$$n = \frac{z^2 \cdot PQ \cdot N}{(N-1) \cdot e^2 + z^2 \cdot PQ}$$

	Margen de error (e)										
Población (N)	**1,0**	**1,5**	**2,0**	**2,5**	**3,0**	**3,5**	**4,0**	**4,5**	**5,0**	**7,5**	**10,0**
250	247	244	239	234	227	220	212	204	196	154	118
500	489	476	459	439	417	393	369	345	321	222	155
1.000	957	909	849	783	714	647	584	526	474	286	184
1.500	1.406	1.304	1.184	1.059	938	826	726	638	563	316	196
2.000	1.837	1.667	1.475	1.286	1.111	957	826	714	621	333	202
2.500	2.250	2.000	1.731	1.475	1.250	1.059	900	769	662	345	206
3.000	2.647	2.308	1.957	1.636	1.364	1.139	957	811	692	353	209
3.500	3.029	2.593	2.158	1.775	1.458	1.205	1.003	843	716	359	211
4.000	3.396	2.857	2.338	1.895	1.538	1.259	1.040	870	735	364	213
4.500	3.750	3.103	2.500	2.000	1.607	1.304	1.071	891	750	367	214
5.000	4.091	3.333	2.647	2.093	1.667	1.343	1.098	909	763	370	215
6.000	4.737	3.750	2.903	2.250	1.765	1.406	1.139	938	783	375	217
7.000	5.339	4.118	3.119	2.377	1.842	1.455	1.171	959	797	378	218
8.000	5.902	4.444	3.303	2.483	1.905	1.494	1.196	976	809	381	219
9.000	6.429	4.737	3.462	2.571	1.957	1.525	1.216	989	818	383	220
10.000	6.923	5.000	3.600	2.647	2.000	1.552	1.233	1.000	826	385	220
15.000	9.000	6.000	4.091	2.903	2.143	1.636	1.286	1.034	849	390	222
20.000	10.588	6.667	4.390	3.051	2.222	1.682	1.314	1.053	861	392	222
25.000	11.842	7.143	4.592	3.147	2.273	1.711	1.331	1.064	869	394	223
30.000	12.857	7.500	4.737	3.214	2.308	1.731	1.343	1.071	874	395	223
35.000	13.696	7.778	4.846	3.264	2.333	1.745	1.352	1.077	877	395	224
40.000	14.400	8.000	4.932	3.303	2.353	1.756	1.358	1.081	880	396	224
45.000	15.000	8.182	5.000	3.333	2.368	1.765	1.364	1.084	882	396	224
50.000	15.517	8.333	5.056	3.358	2.381	1.772	1.368	1.087	884	397	224
100.000	18.367	9.091	5.325	3.475	2.439	1.804	1.387	1.099	892	398	224

Número de unidades de la muestra (n), con una población infinita (N≥100.000), para un nivel de confianza del 95,5% (z=2) y diferentes valores de P y Q.

$$n = \frac{z^2 \cdot PQ}{e^2}$$

	Valors de P i Q en % (P + Q = 100)													
Margen de error (e)	P= 1 / Q= 99	2 / 98	3 / 97	4 / 96	5 / 95	10 / 90	15 / 85	20 / 80	25 / 75	30 / 70	35 / 65	40 / 60	45 / 55	50 / 50
0,1	39.600	78.400	116.400	153.600	190.000	360.000	510.000	640.000	750.000	840.000	910.000	960.000	990.000	1.000.000
0,2	9.900	19.600	29.100	38.400	47.500	90.000	127.500	160.000	187.500	210.000	227.500	240.000	247.500	250.000
0,3	4.400	8.711	12.933	17.067	21.111	40.000	56.667	71.111	83.333	93.333	101.111	106.667	110.000	111.111
0,4	2.475	4.900	7.275	9.600	11.875	22.500	31.875	40.000	46.875	52.500	56.875	60.000	61.875	62.500
0,5	1.584	3.136	4.656	6.144	7.600	14.400	20.400	25.600	30.000	33.600	36.400	38.400	39.600	40.000
0,6	1.100	2.178	3.233	4.267	5.278	10.000	14.167	17.778	20.833	23.333	25.278	26.667	27.500	27.778
0,7	808	1.600	2.376	3.135	3.878	7.347	10.408	13.061	15.306	17.143	18.571	19.592	20.204	20.408
0,8	619	1.225	1.819	2.400	2.969	5.625	7.969	10.000	11.719	13.125	14.219	15.000	15.469	15.625
0,9	489	968	1.437	1.896	2.346	4.444	6.296	7.901	9.259	10.370	11.235	11.852	12.222	12.346
1,0	396	784	1.164	1.536	1.900	3.600	5.100	6.400	7.500	8.400	9.100	9.600	9.900	10.000
1,5	176	348	517	683	844	1.600	2.267	2.844	3.333	3.733	4.044	4.267	4.400	4.444
2,0	99	196	291	384	475	900	1.275	1.600	1.875	2.100	2.275	2.400	2.475	2.500
2,5	63	125	186	246	304	576	816	1.024	1.200	1.344	1.456	1.536	1.584	1.600
3,0	44	87	129	171	211	400	567	711	833	933	1.011	1.067	1.100	1.111
3,5	32	64	95	125	155	294	416	522	612	686	743	784	808	816
4,0	25	49	73	96	119	225	319	400	469	525	569	600	619	625
4,5	20	39	57	76	94	178	252	316	370	415	449	474	489	494
5,0	16	31	47	61	76	144	204	256	300	336	364	384	396	400
6,0	11	22	32	43	53	100	142	178	208	233	253	267	275	278
7,0	8	16	24	31	39	73	104	131	153	171	186	196	202	204
8,0	6	12	18	24	30	56	80	100	117	131	142	150	155	156
9,0	5	10	14	19	23	44	63	79	93	104	112	119	122	123
10,0	4	8	12	15	19	36	51	64	75	84	91	96	99	100
15,0	2	3	5	7	8	16	23	28	33	37	40	43	44	44
20,0	1	2	3	4	5	9	13	16	19	21	23	24	25	25
25,0	1	1	2	2	3	6	8	10	12	13	15	15	16	16
30,0	0	1	1	2	2	4	6	7	8	9	10	11	11	11
35,0	0	1	1	1	2	3	4	5	6	7	7	8	8	8
40,0	0	0	1	1	1	2	3	4	5	5	6	6	6	6

Número de unidades de la muestra (n), con una población infinita (N≥100.000), para un nivel de confianza del 99,7% (z=3) y diferentes valores de P y Q.

$$n = \frac{z^2 \cdot PQ}{e^2}$$

	Valors de P i Q en % (P + Q = 100)													
Margen de error (e)	P= 1 / Q= 99	2 / 98	3 / 97	4 / 96	5 / 95	10 / 90	15 / 85	20 / 80	25 / 75	30 / 70	35 / 65	40 / 60	45 / 55	50 / 50
0,1	89.100	176.400	261.900	345.600	427.500	810.000	1.147.500	1.440.000	1.687.500	1.890.000	2.047.500	2.160.000	2.227.500	2.250.000
0,2	22.275	44.100	65.475	86.400	106.875	202.500	286.875	360.000	421.875	472.500	511.875	540.000	556.875	562.500
0,3	9.900	19.600	29.100	38.400	47.500	90.000	127.500	160.000	187.500	210.000	227.500	240.000	247.500	250.000
0,4	5.569	11.025	16.369	21.600	26.719	50.625	71.719	90.000	105.469	118.125	127.969	135.000	139.219	140.625
0,5	3.564	7.056	10.476	13.824	17.100	32.400	45.900	57.600	67.500	75.600	81.900	86.400	89.100	90.000
0,6	2.475	4.900	7.275	9.600	11.875	22.500	31.875	40.000	46.875	52.500	56.875	60.000	61.875	62.500
0,7	1.818	3.600	5.345	7.053	8.724	16.531	23.418	29.388	34.439	38.571	41.786	44.082	45.459	45.918
0,8	1.392	2.756	4.092	5.400	6.680	12.656	17.930	22.500	26.367	29.531	31.992	33.750	34.805	35.156
0,9	1.100	2.178	3.233	4.267	5.278	10.000	14.167	17.778	20.833	23.333	25.278	26.667	27.500	27.778
1,0	891	1.764	2.619	3.456	4.275	8.100	11.475	14.400	16.875	18.900	20.475	21.600	22.275	22.500
1,5	396	784	1.164	1.536	1.900	3.600	5.100	6.400	7.500	8.400	9.100	9.600	9.900	10.000
2,0	223	441	655	864	1.069	2.025	2.869	3.600	4.219	4.725	5.119	5.400	5.569	5.625
2,5	143	282	419	553	684	1.296	1.836	2.304	2.700	3.024	3.276	3.456	3.564	3.600
3,0	99	196	291	384	475	900	1.275	1.600	1.875	2.100	2.275	2.400	2.475	2.500
3,5	73	144	214	282	349	661	937	1.176	1.378	1.543	1.671	1.763	1.818	1.837
4,0	56	110	164	216	267	506	717	900	1.055	1.181	1.280	1.350	1.392	1.406
4,5	44	87	129	171	211	400	567	711	833	933	1.011	1.067	1.100	1.111
5,0	36	71	105	138	171	324	459	576	675	756	819	864	891	900
6,0	25	49	73	96	119	225	319	400	469	525	569	600	619	625
7,0	18	36	53	71	87	165	234	294	344	386	418	441	455	459
8,0	14	28	41	54	67	127	179	225	264	295	320	338	348	352
9,0	11	22	32	43	53	100	142	178	208	233	253	267	275	278
10,0	9	18	26	35	43	81	115	144	169	189	205	216	223	225
15,0	4	8	12	15	19	36	51	64	75	84	91	96	99	100
20,0	2	4	7	9	11	20	29	36	42	47	51	54	56	56
25,0	1	3	4	6	7	13	18	23	27	30	33	35	36	36
30,0	1	2	3	4	5	9	13	16	19	21	23	24	25	25
35,0	1	1	2	3	3	7	9	12	14	15	17	18	18	18
40,0	1	1	2	2	3	5	7	9	11	12	13	14	14	14

Determinación del margen de error (e), según el tamaño de la muestra (n) y diferentes valores de P y Q , para un nivel de confianza del 95,5% (z=2) con una población infinita (N≥100.000)

$$e = z \cdot \sqrt{\frac{PQ}{n}}$$

	Valors de P i Q en % (P + Q = 100)													
Tamaño de la	P=	1	2	3	4	5	10	15	20	25	30	35	40	45	50
muestra (n)	Q=	99	98	97	96	95	90	85	80	75	70	65	60	55	50
25		4,0	5,6	6,8	7,8	8,7	12,0	14,3	16,0	17,3	18,3	19,1	19,6	19,9	20,0
50		2,8	4,0	4,8	5,5	6,2	8,5	10,1	11,3	12,2	13,0	13,5	13,9	14,1	14,1
75		2,3	3,2	3,9	4,5	5,0	6,9	8,2	9,2	10,0	10,6	11,0	11,3	11,5	11,5
100		2,0	2,8	3,4	3,9	4,4	6,0	7,1	8,0	8,7	9,2	9,5	9,8	9,9	10,0
150		1,6	2,3	2,8	3,2	3,6	4,9	5,8	6,5	7,1	7,5	7,8	8,0	8,1	8,2
200		1,4	2,0	2,4	2,8	3,1	4,2	5,0	5,7	6,1	6,5	6,7	6,9	7,0	7,1
250		1,3	1,8	2,2	2,5	2,8	3,8	4,5	5,1	5,5	5,8	6,0	6,2	6,3	6,3
300		1,1	1,6	2,0	2,3	2,5	3,5	4,1	4,6	5,0	5,3	5,5	5,7	5,7	5,8
400		1,0	1,4	1,7	2,0	2,2	3,0	3,6	4,0	4,3	4,6	4,8	4,9	5,0	5,0
500		0,9	1,3	1,5	1,8	1,9	2,7	3,2	3,6	3,9	4,1	4,3	4,4	4,4	4,5
600		0,8	1,1	1,4	1,6	1,8	2,4	2,9	3,3	3,5	3,7	3,9	4,0	4,1	4,1
800		0,7	1,0	1,2	1,4	1,5	2,1	2,5	2,8	3,1	3,2	3,4	3,5	3,5	3,5
1.000		0,6	0,9	1,1	1,2	1,4	1,9	2,3	2,5	2,7	2,9	3,0	3,1	3,1	3,2
1.200		0,6	0,8	1,0	1,1	1,3	1,7	2,1	2,3	2,5	2,6	2,8	2,8	2,9	2,9
1.500		0,5	0,7	0,9	1,0	1,1	1,5	1,8	2,1	2,2	2,4	2,5	2,5	2,6	2,6
2.000		0,4	0,6	0,8	0,9	1,0	1,3	1,6	1,8	1,9	2,0	2,1	2,2	2,2	2,2
2.500		0,4	0,6	0,7	0,8	0,9	1,2	1,4	1,6	1,7	1,8	1,9	2,0	2,0	2,0
3.000		0,4	0,5	0,6	0,7	0,8	1,1	1,3	1,5	1,6	1,7	1,7	1,8	1,8	1,8
4.000		0,3	0,4	0,5	0,6	0,7	0,9	1,1	1,3	1,4	1,4	1,5	1,5	1,6	1,6
5.000		0,3	0,4	0,5	0,6	0,6	0,8	1,0	1,1	1,2	1,3	1,3	1,4	1,4	1,4
7.500		0,2	0,3	0,4	0,5	0,5	0,7	0,8	0,9	1,0	1,1	1,1	1,1	1,1	1,2
10.000		0,2	0,3	0,3	0,4	0,4	0,6	0,7	0,8	0,9	0,9	1,0	1,0	1,0	1,0
15.000		0,2	0,2	0,3	0,3	0,4	0,5	0,6	0,7	0,7	0,7	0,8	0,8	0,8	0,8
25.000		0,1	0,2	0,2	0,2	0,3	0,4	0,5	0,5	0,5	0,6	0,6	0,6	0,6	0,6
50.000		0,1	0,1	0,2	0,2	0,2	0,3	0,3	0,4	0,4	0,4	0,4	0,4	0,4	0,4

Determinación del margen de error (e), según el tamaño de la muestra (n) y diferentes valores de P y Q , para un nivel de confianza del 99,7% (z=3) con una población infinita (N≥100.000)

$$e = z \cdot \sqrt{\frac{PQ}{n}}$$

	Valors de P i Q en % (P + Q = 100)													
Tamaño de la	P=	1	2	3	4	5	10	15	20	25	30	35	40	45	50
muestra (n)	Q=	99	98	97	96	95	90	85	80	75	70	65	60	55	50
25		6,0	8,4	10,2	11,8	13,1	18,0	21,4	24,0	26,0	27,5	28,6	29,4	29,8	30,0
50		4,2	5,9	7,2	8,3	9,2	12,7	15,1	17,0	18,4	19,4	20,2	20,8	21,1	21,2
75		3,4	4,8	5,9	6,8	7,5	10,4	12,4	13,9	15,0	15,9	16,5	17,0	17,2	17,3
100		3,0	4,2	5,1	5,9	6,5	9,0	10,7	12,0	13,0	13,7	14,3	14,7	14,9	15,0
150		2,4	3,4	4,2	4,8	5,3	7,3	8,7	9,8	10,6	11,2	11,7	12,0	12,2	12,2
200		2,1	3,0	3,6	4,2	4,6	6,4	7,6	8,5	9,2	9,7	10,1	10,4	10,6	10,6
250		1,9	2,7	3,2	3,7	4,1	5,7	6,8	7,6	8,2	8,7	9,0	9,3	9,4	9,5
300		1,7	2,4	3,0	3,4	3,8	5,2	6,2	6,9	7,5	7,9	8,3	8,5	8,6	8,7
400		1,5	2,1	2,6	2,9	3,3	4,5	5,4	6,0	6,5	6,9	7,2	7,3	7,5	7,5
500		1,3	1,9	2,3	2,6	2,9	4,0	4,8	5,4	5,8	6,1	6,4	6,6	6,7	6,7
600		1,2	1,7	2,1	2,4	2,7	3,7	4,4	4,9	5,3	5,6	5,8	6,0	6,1	6,1
800		1,1	1,5	1,8	2,1	2,3	3,2	3,8	4,2	4,6	4,9	5,1	5,2	5,3	5,3
1.000		0,9	1,3	1,6	1,9	2,1	2,8	3,4	3,8	4,1	4,3	4,5	4,6	4,7	4,7
1.200		0,9	1,2	1,5	1,7	1,9	2,6	3,1	3,5	3,8	4,0	4,1	4,2	4,3	4,3
1.500		0,8	1,1	1,3	1,5	1,7	2,3	2,8	3,1	3,4	3,5	3,7	3,8	3,9	3,9
2.000		0,7	0,9	1,1	1,3	1,5	2,0	2,4	2,7	2,9	3,1	3,2	3,3	3,3	3,4
2.500		0,6	0,8	1,0	1,2	1,3	1,8	2,1	2,4	2,6	2,7	2,9	2,9	3,0	3,0
3.000		0,5	0,8	0,9	1,1	1,2	1,6	2,0	2,2	2,4	2,5	2,6	2,7	2,7	2,7
4.000		0,5	0,7	0,8	0,9	1,0	1,4	1,7	1,9	2,1	2,2	2,3	2,3	2,4	2,4
5.000		0,4	0,6	0,7	0,8	0,9	1,3	1,5	1,7	1,8	1,9	2,0	2,1	2,1	2,1
7.500		0,3	0,5	0,6	0,7	0,8	1,0	1,2	1,4	1,5	1,6	1,7	1,7	1,7	1,7
10.000		0,3	0,4	0,5	0,6	0,7	0,9	1,1	1,2	1,3	1,4	1,4	1,5	1,5	1,5
15.000		0,2	0,3	0,4	0,5	0,5	0,7	0,9	1,0	1,1	1,1	1,2	1,2	1,2	1,2
25.000		0,2	0,3	0,3	0,4	0,4	0,6	0,7	0,8	0,8	0,9	0,9	0,9	0,9	0,9
50.000		0,1	0,2	0,2	0,3	0,3	0,4	0,5	0,5	0,6	0,6	0,6	0,7	0,7	0,7

2.1.3. *Ejercicios de cálculo del tamaño de la muestra*

Ejercicio 1

En un estudio sobre el sindicato CC.OO. se quiere obtener una muestra aleatoria de la población afiliada a la que administrar un cuestionario para estudiar su perfil y la valoración de la acción sindical. ¿Cuál debe ser el tamaño de la muestra? Sabemos que a diciembre de 1998 el número de afiliados/as era de 123.440 personas.

Tipo de población: Infinita
Nivel de error que queremos asumir: 3%
Varianza: Supuesto de máxima indeterminación, P=Q=50%
Nivel de confianza: 95,5%

Cálculo de *n*: $$n = \frac{z^2 \cdot PQ}{e^2} = \frac{2^2 \cdot 50 \cdot 50}{3^2} = \frac{2^2}{4 \cdot 0,03^2} = 1111,11 = 1112$$

Ejercicio 2

En un estudio sobre mercado de trabajo y empresa en la Región Metropolitana de Barcelona en el contexto de los Juegos Olímpicos[10] *se considera el análisis de diversas características de los centros de trabajo y su comportamiento frente al mercado de trabajo. Se quiere recoger información por encuesta en relación a una población de 24.141 empresas de más de 10 trabajadores/as. ¿Cuántas hay que seleccionar para obtener una muestra representativa?*

Tipo de población: Finita
Nivel de error que queremos asumir: 3,9%
Varianza: Supuesto de máxima incertidumbre, P=Q=50%
Nivel de confianza: 95,5%

Cálculo de *n*: $$n = \frac{2^2 \cdot 50 \cdot 50 \cdot 24141}{(24141-1) \cdot 3,9^2 + 2^2 \cdot 50 \cdot 50} = 640$$

Ejercicio 3

En un segmento de mercado con una población identificada de 1000 hogares y hábitos de consumo similares se quiere realizar un muestreo aleatorio simple para conocer el número medio de unidades de consumo anual de un cierto producto de consumo cultural. Decidir un nivel de error muestral y calcular el tamaño de la muestra necesaria para realizar este estudio sabiendo que según las estimaciones de otros estudios anteriores la varianza poblacional del número de unidades de consumo se sitúa sobre un valor de 100. A continuación determinar el tamaño de la muestra si la población fuera de 200.000 hogares.

Tipo de población: Finita
Nivel de error que queremos asumir: por ejemplo, 0,5 o 1 unidad de consumo
Varianza: 100
Nivel de confianza: 95,5%

[10] Grupo QUIT (1997). *Economía, Trabajo y Empresa. Sobre el impacto económico y laboral de los Juegos Olímpicos de 1992*. Madrid: Consejo Económico y Social.

Cálculo de n: $$n = \frac{z^2 \cdot \sigma^2 \cdot N}{(N-1) \cdot e^2 + z^2 \cdot \sigma^2} = \frac{2^2 \cdot 100 \cdot 1000}{(1000-1) \cdot 0,5^2 + 2^2 \cdot 100} = 615,6 = 616$$

$$n = \frac{z^2 \cdot \sigma^2 \cdot N}{(N-1) \cdot e^2 + z^2 \cdot \sigma^2} = \frac{2^2 \cdot 100 \cdot 1000}{(1000-1) \cdot 1^2 + 2^2 \cdot 100} = 285,91 = 286$$

Si N=200.000, cálculo de n: $$n = \frac{z^2 \cdot \sigma^2}{e^2} = \frac{2^2 \cdot 100}{0,5^2} = 1.600$$

$$n = \frac{z^2 \cdot \sigma^2}{e^2} = \frac{2^2 \cdot 100}{1^2} = 400$$

Ejercicio 4

Un sondeo electoral realizó las estimaciones de intención de voto a partir de una muestra de la población de 2100 entrevistas. ¿Qué error muestral se está considerando?

Tipo de población: Infinita
Tamaño de la muestra: 2.100
Varianza: Supuesto de máxima incertidumbre, P=Q=50%
Nivel de confianza: 95%

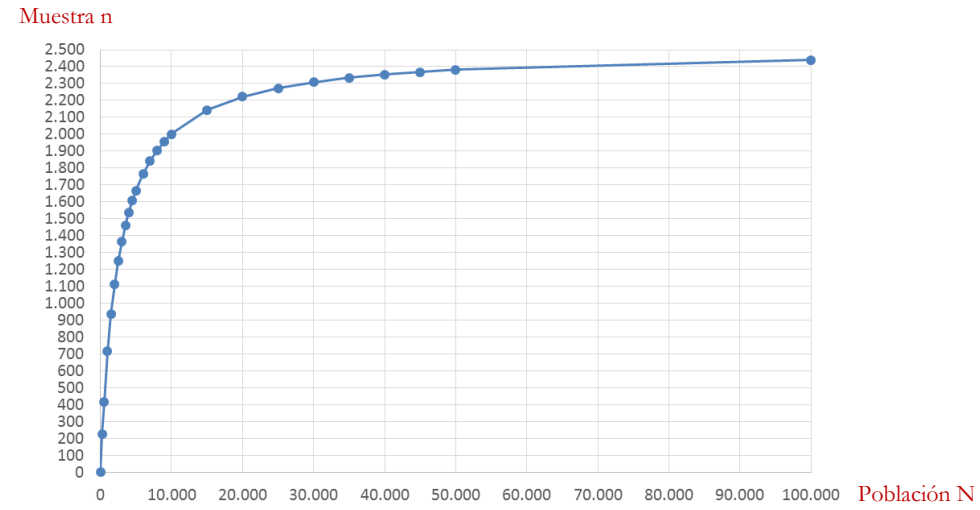
Cálculo de e: $$e = z \cdot \sqrt{\frac{PQ}{n}} = 1,96 \cdot \sqrt{\frac{50 \cdot 50}{2100}} = 2,14\%$$

2.1.4. *Relación entre el tamaño de la muestra, de la población y del error*

Finalmente incluimos un apartado con un análisis ilustrativo de las relaciones que dan entre el tamaño de la muestra y el tamaño de la población, así como entre el tamaño de la muestra y el error muestral.

En el primer caso una representación gráfica del tamaño de la muestra en función del tamaño muestral (Gráfico II.4.4) dibuja una curva que pasa por el origen de coordenadas con un comportamiento asintótico que nos pone de manifiesto una relación directa y un comportamiento creciente (a medida que aumenta el tamaño de la población necesitamos más muestra para un determinado nivel de error muestral) pero no es un comportamiento de proporcionalidad, de crecimiento lineal. Al inicio, con poblaciones pequeñas, se requiere un número importante de elementos de la muestra, y a medida que crece el tamaño poblacional los elementos muestrales requeridos crecen notablemente, pero a partir de un determinado N los incrementos son cada vez más reducidos, hasta llegar a un momento en que los aumentos de N no producen aumentos en el tamaño de la muestra. A partir de entonces la población es infinita.

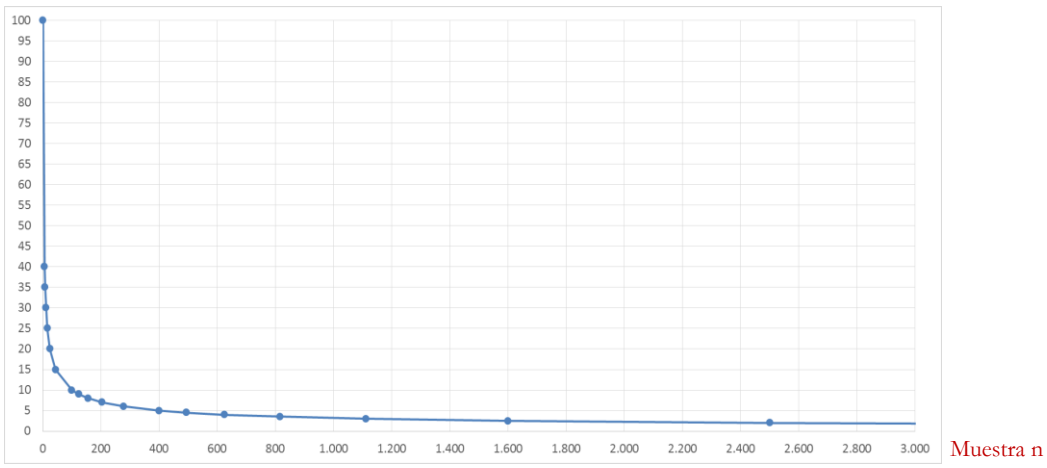
Gráfico II.4.4. Relación entre el tamaño de la muestra y el tamaño de la población
Poblaciones finitas, e=2%, z=2, P=Q=50%



En particular, al pasar de una población 1 a 1000 la muestra necesaria pasa de 1 a 714, mientras que de 1000 a 2000 tan sólo necesitamos aumentar la muestra en 397, y de 2000 a 3000, el aumento es de 253, progresivamente la necesidad de nuevos elementos en la muestra disminuye hasta estabilizarse a partir de 100.000 individuos.

Un comportamiento similar se da entre el error muestral y el tamaño de la muestra en el sentido de que tampoco se trata de una relación proporcional lineal (Gráfico II.4.5). Error y tamaño tienen una relación inversa, a medida que aumenta el tamaño muestral se reduce el error. Pero el tamaño *n* es inversamente proporcional al cuadrado del error de muestreo, lo que implica que cualquier pequeño incremento de unidades en la muestra, cuando éstas son pequeñas, producen una reducción muy importante del error cometido, pero llega un momento en que las reducciones son cada vez más pequeñas de modo que reducciones mínimas del error suponen incrementos muy elevados del tamaño de la muestra.

Gráfico II.4.5. Relación entre el tamaño de la muestra y el margen de error.
Población infinita, z=2, P=Q=50%



En particular, al pasar de un tamaño de 1 a 100 se reduce el error del 100% al 10%, una reducción de 90 puntos porcentuales, pero al pasar de 100 a 200, la reducción es tan solo del 3%.

2.1.5. *El error asociado a una estimación*

Para finalizar este apartado apuntamos someramente uno de los aspectos que permanentemente se plantea en el análisis de los datos derivados de una muestra estadística: el error que se deriva de una estimación puntual en la muestra obtenida. Como hemos destacado anteriormente, toda afirmación estadística expresada, por ejemplo, en términos de un porcentaje o de una media responde al aspecto descriptivo de los resultados de un estudio cuantitativo por muestreo. Este dato se acompaña de un elemento adicional en términos inferenciales para dar cuenta del error asociado al estadístico en cuestión, así como para plantear distintas cuestiones sobre su significación. Sería el caso, por ejemplo, de estar interesados en conocer la media de los ingresos de una población y de plantearse si esa media es igual entre varones o mujeres, o bien difiere significativamente entre ambos sexos. Una prueba de hipótesis estadística establecerá, basándose en el error asociado a esas estimaciones, si podemos concluir la igualdad o la significación de la diferencia.

Pero consideremos el caso sencillo de estimar una sola media o un solo porcentaje, y planteemos la cuestión del error asociado a estas estimaciones. Por ejemplo, supongamos desconocida la media de ingresos μ de la población de un municipio. Una estimación de este valor se puede obtener tomando una muestra aleatoria de sus habitantes. Pongamos que obtenemos una muestra de $n=1.000$ personas a las que les preguntamos por sus ingresos, y con las 1.000 respuestas u observaciones calculamos la media de los ingresos. Supongamos que los ingresos medios obtenidos fueran $\bar{x}=1.500€$, esta sería la estimación puntual de la media poblacional. Tengamos presente que en otra posible muestra de 1.000 personas hubiera dado unos ingresos medios diferentes, por ejemplo, de 1.600€. De hecho, cada posible muestra generaría puntualmente un resultado distinto, lo que nos está indicado la existencia de una variabilidad y de que estamos cometiendo un error de estimación. Este es un aspecto crucial que nos conducirá a hablar de estimaciones por intervalos teniendo en cuenta el error que se comete. Por el hecho de disponer de una parte de la población, aunque sea representativa e incluso numerosa, cada estimación del valor verdadero (y desconocido) del parámetro poblacional a partir del estadístico muestral tiene un error respecto del valor real y desconocido del parámetro. Nuestro resultado que puede estar más cercano (¡si tenemos buena suerte con la muestra!) o menos cercano (¡si tenemos mala suerte!), pero difícilmente coincidirá con el verdadero valor de la media. El margen de error e es la diferencia máxima entre la estimación puntual obtenida en la muestra y el valor verdadero del parámetro poblacional y nos mide el grado de exactitud o de precisión con el que inferimos de la muestra a la población. Este valor vendrá determinado por la variabilidad del estadístico, es decir, que el error se cuantifica mediante las varianzas del estadístico considerado en cada caso. Como veremos esta variabilidad se denomina error típico del estadístico, y se corresponde con un cálculo que depende de la varianza y del tamaño de la muestra, además de otra

característica como el nivel de confianza, de forma similar a como hemos visto en el cálculo del tamaño de la muestra.

Pero el error muestral se define simplemente como la divergencia entre los valores de los estadísticos obtenidos en la muestra de una variable y los existentes en la población. En el caso del estadístico de la media, el error muestral sería la diferencia entre el valor muestral de la media, $\bar{x}$, y el valor poblacional de la media, μ, es decir, $\bar{x} - \mu$. Esta es una definición de lo real pero no conocido, ya que el parámetro poblacional es desconocido y es lo que se quiere saber. Para solucionar este interrogante, razonaremos en términos probabilísticos a partir del conocimiento de la forma de la distribución del estadístico. En este sentido, cualquier afirmación para dar cuenta del valor de un parámetro poblacional desconocido a partir del valor del estadístico que lo estima, tendrá un grado de desviación, de variación al alza o de variación a la baja, que es el que mide el error de muestreo. De aquí se deriva una relación básica del muestreo y de las estimaciones estadísticas:

Valor poblacional = **Valor estimado en la muestra** ± **Error de muestreo**
(**Parámetro**) (**Estadístico**)

De esta forma, cada estadístico ($\bar{x}$, p, etc.) debemos entenderlo como si fuera una variable, en el sentido de que puede tomar varios valores según la muestra que consideramos (de todas las muestras posibles que podemos tomar de la población). Así, todos los posibles valores que estiman el único valor verdadero del parámetro forman lo que se llama una distribución muestral del estadístico. En el caso de una media y una proporción es distribución muestral teórica es la normal, la cual nos ayudará a determinar, con una certeza suficiente, el margen de error que cometemos al hacer estimaciones.

Si, como ejemplo, obtenemos, tras preguntar a una muestra de 1000 personas, $n=1000$, que la intención de voto a un determinado partido en las próximas elecciones es del 30% de los votos, $p=30$, siendo $q=100-p=70$, el error asociado se calcula a partir del producto del valor de la distribución normal z y el error típico de la estimación de la proporción s_p, es decir:

$$e = z \times s_p$$

Asumiendo, como es habitual, un nivel de confianza del 95,5%, es decir, $z=2$, y siendo el error típico de la estimación es igual a $\sqrt{p \times q / n}$, por tanto, igual a $\sqrt{30 \times 70 / 1000} = 1,45$, se obtiene que el error asociado a la estimación es $e = z \times s_p = 2 \times 1,45 = 2,9$, es decir, del 2,9%.

Cuando a la estimación puntual le añadimos el margen de error, sumando y restando el error de muestreo, realizamos la estimación por intervalos de confianza. Es decir, que el valor poblacional se situará en un intervalo definido por:

$$30\% \pm 2,9\%$$

es decir, en un intervalo entre el

27,1% y el 32,9%.

y que escribimos así (27,1%,32,9%).

Cualquier valor puntual estimado es una aproximación al valor poblacional desconocido afectado por el error muestral, por lo tanto, esto no significa que el porcentaje sea estrictamente del 30% en nuestro ejemplo, sino que sólo podemos establecer un intervalo de valores, entre un mínimo y un máximo, en cuyo interior se situará, probablemente el parámetro poblacional. Y decimos probablemente porque el intervalo se establece con un grado de probabilidad o nivel de confianza, del 95%. De esta forma funciona el razonamiento estadístico en todos los casos a partir de muestras aleatorias representativas de una población.

2.2. Muestreo sistemático

El muestreo sistemático (MS) es muy similar al muestreo aleatorio simple, se diferencia simplemente en el método de selección de las unidades muestrales, más mecánico y elemental, buscando ahorrar esfuerzos en la extracción.

El método consiste en lo siguiente:

1. Se disponen en un listado numerado de las unidades del universo.
2. Se elige al azar la primera unidad de la muestra, r, mientras sea inferior al coeficiente de elevación $k = f^{-1} = \frac{N}{n}$, con $1 \leq r \leq k$.
3. Los restantes elementos de la muestra se obtienen sumando al primero el coeficiente de elevación k hasta obtener el tamaño total n. Se trata de una progresión de secuencia $r, r+k, r+2k, r+3k,..., r+(n-1)k$.

Por ejemplo, si consideramos la extracción de una muestra aleatoria de n=1.000 individuos de la población de clientes de una empresa formada por N=10.000 unidades, el coeficiente de elevación será k =10.000/1.000=10. Elegimos al azar el primer individuo entre 1 y 10, por ejemplo, el 3, entonces todos los elementos de la muestra serán las unidades sucesivas numeradas con: 3, 13, 23, 33, 43, ... hasta llegar a la unidad 9.993, el último elemento de los 1.000 que forman la muestra.

Cuando no se dispone de un marco de la muestra numerado, se puede recurrir a algún listado disponible. Tradicionalmente se hacía por ejemplo utilizando la "guía de teléfonos" con la relación de todos los números fijos existentes. Se trata de arbitrar un mecanismo de extracción sistemática de números de teléfono con una secuencia dada (cada un cierto número de páginas) desde el comienzo del listado hasta el final. El censo electoral suele ser un marco de muestreo habitual para obtener una muestra aleatoria que se puede utilizar para la extracción sistemática. La selección aleatoria de una de cada 10 personas o de una cada 5 minutos que sale de un colegio electoral (de un museo, de un avión, etc.) es un procedimiento que nos remite a la idea de una extracción sistemática sin marco de muestreo.

Si se garantiza que cada elemento y cada muestra posible de la población tiene idénticas probabilidades de ser elegido, las estimaciones de los parámetros siguen las mismas características que en el MAS y la técnica será igual de precisa. Pero, además, podría ser un procedimiento más eficiente, es decir, puede aportar estimaciones más precisas que el MAS en función de la tendencia derivada de la ordenación de las unidades poblacionales. Si la ordenación de las unidades presenta una tendencia continua en relación a la variable estudiada, entonces se pueden esperar mejores estimaciones. Si, por ejemplo, queremos estudiar alguna característica de las empresas y ordenamos la población según el número de trabajadores/as, un muestreo sistemático nos garantiza la representación en la muestra tanto de las más pequeñas (las más numerosas, con diferencia, sobre todo en el caso del tejido empresarial español) como de las más grandes (las menos numerosas), situación que no necesariamente se consigue con el muestreo aleatorio simple.

Pero de la misma forma que se obtiene una representatividad y se pueden obtener mejores estimaciones, también es posible obtener peores resultados. En el muestreo sistemático estamos suponiendo la aleatoriedad, lo que implica o exige analizar la disposición de las unidades poblacionales para que no sigan ningún tipo ordenación cíclica que sesga sistemáticamente la muestra en relación a los objetivos de la investigación y del muestreo. Si al hacer la extracción está presente una tendencia periódica no controlada que no representa las características poblacionales estudiadas el resultado será una estimación no muy precisa. Es el caso del estudio de las prácticas de consumo o de los gastos de los hogares a lo largo del año cuando se aplica un muestreo sistemático a los días de todo el año, nos encontraremos con el inicio y el final de mes que implican comportamientos de gasto muy diferentes y por lo tanto los perfiles agregados de los hogares están medidos de forma diferente y no representativa.

2.3. Muestreo aleatorio estratificado

El muestreo aleatorio estratificado (MAE) es un método de muestreo que tiene la gran ventaja de permitir mejorar la precisión de las estimaciones en relación al muestreo aleatorio simple, es decir, disminuye el error muestral y de las estimaciones para un mismo tamaño de muestra, o bien reduce el tamaño para un mismo margen de error.

En el MAE se parte de la consideración de que la población no es homogénea según los objetivos de la investigación, y se trata de dividir a la población en categorías o grupos que tienen un interés analítico en función de una o más características o variables criterio de la población que definen la heterogeneidad. Estos grupos son las subpoblaciones que definen los estratos. Desde el punto de vista de la investigación resulta fundamental que estas características se respeten de forma más o menos estricta en la muestra y así garantizar su representatividad en términos muestrales, a la vez que, con esta forma de proceder, como hemos dicho, se consigue un ganancia en la reducción de la variabilidad y del error de las estimaciones que se hacen para toda la población.

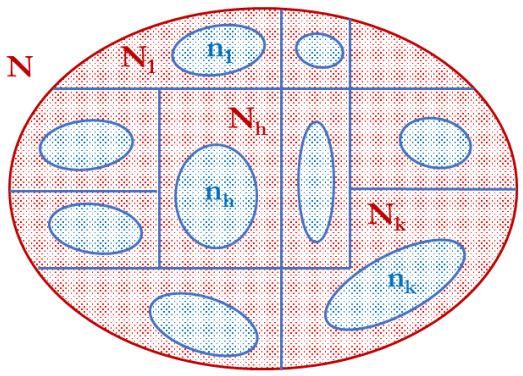
Este tipo de muestreo, por otro lado, nos permite plantear y nos facilita la necesidad de efectuar estudios o estimaciones de características poblacionales a partir de la definición de subpoblaciones que se hacen corresponder con los estratos. Estas subpoblaciones objeto de estudio específico resultan representativas en sus respectivas

submuestras con un grado de error aceptable que el MAE favorece, siempre que el tamaño de la muestra sea suficientemente grande. En todo caso, el número de muestra necesario siempre será menor, para el mismo nivel de error muestral, que en el muestreo aleatorio simple. En consecuencia, se trata de un método que reporta un menor coste y es más eficiente también en este sentido.

Pero, la cuestión crucial en este tipo de muestreo, que limita en la práctica su aplicación generalizada, es la exigencia de disponer de la información de las variables criterio para todas las unidades de la población para clasificarlas en estratos poblacionales. El método estratificado implica, por un lado, definir y construir los estratos poblacionales *h* donde se agrupan todas las unidades (un total de *k* estratos) de forma que todas sean asignadas a un estrato (criterio de exhaustividad, de forma que la unión de todos los estratos es igual a al número de unidades de la población) y cada una pertenezca a un solo estrato (criterio de exclusividad, la intersección entre dos estratos cualesquiera es nula). A continuación, se extrae una muestra aleatoria independiente en cada uno de estos estratos (n_h). El número de unidades que se deben extraer de cada estrato viene determinado por el criterio de reparto que se aplique en función de un tamaño de muestra que es fijada, operación llamada de afijación de la muestra. Posteriormente, reunimos toda la información para obtener estimaciones globales de toda la población.

En el Gráfico II.4.6 adjunto se representa la forma de proceder de este método de muestreo.

Gráfico II.4.6. Representación del muestreo aleatorio estratificado



N: Tamaño de la población
N_h: Tamaño de cada estrato poblacional *h*
n_h: Tamaño de la muestra de cada estrato *h*
n: Tamaño de la muestra total (suma de los n_h)

con $h=1...k$ estratos

El tipo de muestreo que se realiza en cada estrato habitualmente es el mismo, y es característico proceder a obtener una muestra aleatoria simple dentro de cada estrato. Pero también puede ser diferente, utilizando diferentes estrategias en cada estrato para optimizar los recursos económicos, por razones derivadas de las características particulares que definen el estrato o simplemente por la imposibilidad de poder obtener un muestreo aleatorio en alguno de ellos.

La estratificación implica pues conocer unas características poblacionales para cada unidad que son objeto de interés central para la investigación. Pero suele suceder que estas características de interés directo no son conocidas o no están disponibles. Y si se conocieran la pregunta es inmediata, ¿por qué hacer una muestra? Como señalamos en

un momento anterior, habitualmente los estudios no se centran en la necesidad de obtener observaciones de una sola característica, a pesar de que haya una que sea de interés central de la investigación. En estos casos en los que no disponemos de la variable principal de estratificación la solución es recurrir a otro tipo de información para estratificar siempre que sean determinantes del fenómeno estudiado y consecuentemente tengan una elevada correlación con los criterios de estratificación deseados.

Las variables que actúan como criterio clasificatorio se pueden reducir a una única variable o pueden considerarse varias de ellas simultáneamente. Veamos algunos ejemplos de variables de estratificación. Es habitual estratificar unidades que son personas mediante el sexo, la edad, la categoría socioeconómica, el número de habitantes del municipio donde residen (tamaño del hábitat). Cuando se hacen estudios de familias es posible estratificar según el tamaño de los hogares. Si estudiamos la población universitaria un criterio de estratificación es considerar la facultad en la que estudian. Si analizamos el comportamiento de los consumidores es recomendable dividir la población en segmentos de mercado. Cuando se consideran estudios de empresas es frecuente emplear como criterio de estratificación el tamaño de la empresa, utilizando como indicador o bien el número de trabajadores o bien el nivel de facturación, de la mismo forma que se podría utilizar el sector de actividad, o una combinación de sector y tamaño de la empresa. Si se quiere hacer un estudio de la sanidad, a partir de los hospitales, se estratifica por el número de pacientes. Si la investigación es educativa y consideran las escuelas un criterio es considerar el número de alumnos del centro.

El método de estratificación comprende dos etapas principales: la determinación de los estratos y la afijación de la muestra.

1. La determinación de los estratos de la población a partir de una o varias características conocidas a priori de esta (corte de la muestra) implica considerar finalmente una variable clasificatoria categórica, de tipo nominal u ordinal. La obtención de esta clasificación cuando se hace a partir de un único criterio clasificatorio es inmediata, hay que considerar directamente sus categorías (distritos del municipio, comunidad autónoma, sexo, ...) o proceder a una agrupación de sus valores (números de trabajadores en intervalos, grupos de edad, niveles de estudios, etc.). Si la clasificación tiene en cuenta varios criterios clasificatorios simultáneamente (usamos por tanto una tipología con diferentes tipos que definen los estratos) se puede generar por varios procedimientos. Lo más sencillo es considerar el cruce de las diversas variables y considerar tantos estratos como combinaciones de valores resulten. Alternativamente, cuando el número de variables es elevado, se puede utilizar alguna técnica de análisis multivariable clasificatoria que nos ayude a generar una tipología de la población en estratos.

   El número idóneo de estratos no se puede determinar apriorísticamente, depende de varios factores donde hay que considerar las particularidades de cada diseño y los objetivos de investigación. Si consideramos el procedimiento de combinación de valores de diferentes variables, el número de estratos se multiplica rápidamente cuando el número de variables aumenta y sus valores son numerosos. No obstante, por estudios empíricos y teóricos se aconseja multiplicar el número de estratos,

hasta cierto límite en que nuevos estratos no reportan una ganancia significativa de precisión. La consideración de una diversidad de estratos tiene que ver y se justifica con el comportamiento de la variabilidad de los estratos.

Cuando el objetivo de la estratificación es considerar grupos de unidades homogéneas para dar cuenta de la heterogeneidad del fenómeno estudiado, lo que estamos diciendo es que se consideran grupos en el interior de los cuales la variabilidad es reducida, y la fuente de variación se da entre los diferentes grupos. Una variabilidad pequeña en el interior de un estrato se traduce inmediatamente en una reducción del tamaño de muestra necesaria; al límite, si todas las unidades fueran idénticas, la variabilidad interna sería nula y sólo nos haría falta una unidad para dar cuenta de las características del estrato. Llevando este razonamiento al extremo se trataría de considerar tantos estratos como unidades tiene la población, pero evidentemente ya no estaríamos hablando de muestreo sino de un censo. Pero el razonamiento nos sugiere la conveniencia de considerar una diversidad de estratos ya que de esta forma conseguimos reducir la variabilidad interna.

Este razonamiento es bien conocido en estadística, y se expresa en la descomposición de la varianza total de una variable (cantidad constante) en dos partes, la varianza interna o intra grupos y la variabilidad externa o entre grupos. El análisis de varianza se basa en esta distinción y algunos métodos de clasificación automática hacen uso igualmente de estos conceptos. Cuando se consigue que la variable o variables de estratificación hagan mínima la variabilidad intra grupos (se formen estratos homogéneos), lo que significa al mismo que tiempo hacer máxima la variabilidad entre grupos (los grupos son heterogéneos entre ellos), el resultado es una ganancia en eficiencia del muestreo.

2. La afijación de la muestra consiste en repartir un tamaño de muestra determinada previamente entre los diferentes estratos. Con la cuota de muestra que corresponde a cada estrato se procede a efectuar la extracción aleatoria de la muestra del estrato, a través de un muestreo aleatorio simple, por ejemplo. La afijación de la muestra se puede operar de tres formas diferentes:
   a) Afijación igual. Simplemente asignamos la misma muestra a cada uno de los estratos: $n_h = \frac{n}{k}$. Pero suele ser un criterio poco eficiente.
   b) Afijación proporcional. En este caso la afijación distribuye el tamaño de muestra total n proporcionalmente en función del tamaño poblacional de cada estrato: $n_h = \frac{N_h}{N} \cdot n$, de forma que cuando mayor sea el estrato más cuota de muestra le toca y logrando así que cada unidad de la muestra represente el mismo número de unidades de la población. Cuando cada unidad de la muestra tiene el mismo peso y representa al mismo número de unidades de la población la fracción de muestreo es la misma en todos los estratos, $\frac{n_h}{N} = \frac{n}{N}$, y de la muestra se dice que es autoponderada; de esta manera el cálculo, por ejemplo, de una media se puede hacer a partir del conjunto de unidades de todos los estratos.

c) Afijación no proporcional u óptima. La distribución de la muestra tiene en cuenta además del tamaño del estrato, la variabilidad de éste. La afijación llamada óptima de Neyman, considera diferentes fracciones de muestreo según este doble criterio que se expresa mediante las fórmulas:

   – Para la estimación de medias: $n_h = \frac{N_h \times \sigma_h}{\sum_{h=1}^{k} N_h \times \sigma_h} \times n$

   – Para la estimación de proporciones: $n_h = \frac{N_h \times \sqrt{P_h \times Q_h}}{\sum_{h=1}^{k} N_h \times \sqrt{P_h \times Q_h}} \times n$

   Así cuanto mayor sea el estrato mayor cuota de muestra le corresponde, y cuando más variable (heterogéneo) internamente sea el estrato mayor cuota de muestra necesita también. Este doble criterio plantea alterar la estricta proporcionalidad, lo que lleva a considerar la ponderación posterior para restablecerla. De este modo se consigue, primero, la presencia en la muestra de fenómenos menos frecuentes, más atípicos y variables, y, segundo, devolver su peso específico en términos de proporción poblacional. La afijación óptima puede basarse también en un criterio adicional de coste, minimizando este para una varianza dada o minimizando la varianza para un coste dado. El coste, por otro lado, puede ser igual para cada estrato o con costes desiguales para cada estrato. En el caso de la afijación óptima de Neyman se supone que los costes de los estratos son aproximadamente iguales.

El muestreo estratificado siempre supone una ganancia de precisión sobre el MAS, mejor cuanto mayor sea la diferencia entre los estratos, más correlación se dé entre las variables estudiadas y la variable de estratificación y mejor sea la información o la medida de los estratos. Por otra parte, se demuestra que el muestreo estratificado con afijación óptima es más eficiente que el muestreo estratificado con afijación proporcional, las estimaciones están afectadas de un menor error, y su eficiencia será mayor cuando las diferencias de las estimaciones de los estadísticos entre los estratos sean más grandes.

En esta exposición hemos considerado siempre el caso de la estratificación a priori, pero también existe la posibilidad de operar estratificaciones a posteriori, por ejemplo, a partir de la extracción de una muestra aleatoria simple y estratificando según determinantes criterios poblacionales aplicados a la muestra. Aunque de hecho se trata más bien de un ajuste o equilibrio de la muestra hecha después de su obtención.

2.4. Muestreo por conglomerados

El muestreo por conglomerados (MC), llamado también muestreo de *cluster*, de conjuntos, de racimos o de grapas. A diferencia del muestreo aleatorio simple o del estratificado, en este tipo de muestreo las unidades no son simples sino compuestas o complejas (con un tamaño mayor de 1), es decir, se clasifica a la población en grupos, llamados conglomerados, cada uno de los cuales incluye la vez otras unidades más simples o desagregados. Por ejemplo, las unidades agregadas o conglomerados

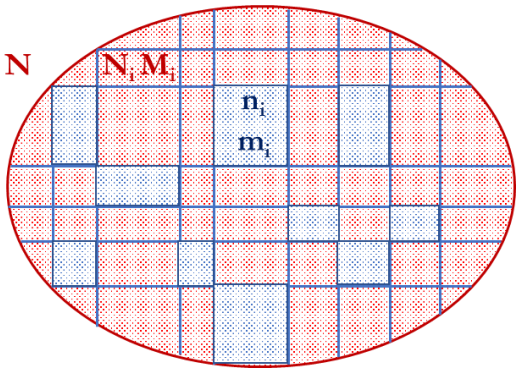
(unidades de muestreo primarias) podrían ser las escuelas de educación primaria de un territorio, y las unidades más simples de cada conglomerado (unidades de muestreo secundarias) el profesorado de estos centros. La estrategia de muestreo consiste en hacer una extracción aleatoria inicial de algunas de las unidades primarias o conglomerados (mediante un muestreo aleatorio simple u otro tipo de diseño) y luego tomar todas las unidades secundarias de cada conglomerado escogido. De esta forma, los elementos individuales de la población sólo pueden participar en la muestra si pertenecen a un conglomerado incluido en la muestra. La unidad de muestreo primaria no es la unidad de observación, sino la secundaria, y hay que tener presente el tamaño de ambos tipos de unidades.

Si esquematizamos el procedimiento del muestreo por conglomerados, se trata de:

1. Dividir a la población N en M conglomerados, con N_i elementos cada uno. Cada uno de los M_i conglomerados pueden ser iguales o de diferente tamaño.
2. Se escogen aleatoriamente m conglomerados de M. La selección se puede hacer con idéntica probabilidad para cada conglomerado o con probabilidad proporcional a su tamaño.
3. De cada conglomerado elegido M_i se toman todos los elementos que lo forman, n_i.

Gráficamente se podría representar de la forma siguiente:

Gráfico II.4.7. Representación del muestreo por conglomerados



M: Número de conglomerados de la población
N_i: Tamaño de unidades elementales de cada conglomerado poblacional M_i $(i= 1...M)$
N: Tamaño de la población
m: Número de conglomerados de la muestra
n_i: Tamaño de unidades elementales de cada conglomerado de la muestra m_i $(i=1...m)$
n: Tamaño de la muestra

Frente al muestreo aleatorio estratificado que pretende establecer también grupos poblacionales (estratos) homogéneos, en el muestreo por conglomerados lo que se busca es precisamente lo contrario, constituir grupos o unidades compuestas que sean lo más heterogéneas internamente y lo más homogéneas entre ellas. Porque lo que se pretende idealmente es que cada conglomerado sea una representación, a escala reducida, de todo el universo considerado; cada conglomerado actúa como un pequeño universo poblacional con toda la diversidad de lo que se estudia, y con la elección de unos cuantos de ellos representamos toda la población con la gran ventaja de concentrar todas las unidades más simples de la muestra en estos conglomerados. Cuando esta concentración se realiza sobre la base de conglomerados que corresponden a divisiones (administrativas) del territorio es evidente que la elección de

unos cuantos conglomerados reduce el número de desplazamientos en el espacio, reduciendo así los costes del muestreo.

De hecho, esta es la clave principal del recurso a esta estrategia de muestreo. La dispersión geográfica de las unidades en el territorio eleva notablemente los gastos de la investigación en tiempo y recursos, de forma que la distribución de la población en áreas geográficas de las que se escogen unas cuantas disminuye la necesidad de desplazamientos. Por este motivo también se identifica y designa al muestreo por conglomerados como muestreo por áreas.

Pero esta no es la única razón que justifica el uso de este método. La necesidad de disponer de un marco de muestreo de todas las unidades poblacionales se encuentra a menudo con las dificultades prácticas para su obtención o confección, lo que impide hacer viable el planteamiento de un muestreo estrictamente aleatorio. Con el muestreo por conglomerados lo que se consigue es reducir el problema de elaborar un listado de las unidades poblacionales para solamente aquellos conglomerados que se han seleccionado inicialmente. Por tanto, no se necesita un marco muy específico de las unidades secundarias, y la constitución de un marco inicial de conglomerados o unidades primarias es económico y fácil de conseguir.

Sin embargo, la estrategia del muestreo por conglomerados resulta de las más ineficientes o imprecisas. En general para el mismo tamaño de muestra es menos eficiente que el MAS. Siempre será más representativa y precisa una muestra aleatoria simple del profesorado de educación primaria extraída del conjunto de docentes que la muestra del mismo tamaño extraída a partir de una selección previa de colegios a partir de los cuales se toma en todo su profesorado. En este sentido se debería compensar el mayor error en las estimaciones con un aumento del tamaño muestral.

La eficiencia del muestreo por conglomerados aumenta cuando aumenta la heterogeneidad interna de los conglomerados y, en consecuencia, aumenta la homogeneidad entre ellos. Pero conseguir este grado de semejanza entre los conglomerados para representar fielmente las características de la población no es fácil ni habitual. También aumenta la eficiencia cuando el tamaño de los conglomerados disminuye, por lo que es conveniente escoger conglomerados de tamaño no muy grande. Es recomendable igualmente que el tamaño de los conglomerados sea similar ya que la precisión de las estimaciones aumenta.

Por otra parte, como inconvenientes adicionales, hay que señalar que exige tratamientos estadísticos complejos y que pierde parte de su carácter probabilístico dado que los conglomerados no seleccionados determinan una probabilidad 0 de las unidades finales de pertenecer a la muestra y de 1 de los seleccionados.

La situación que hemos presentado aquí hasta ahora corresponde a lo que se denomina muestreo monoetápico de conglomerados o de conglomerados en una etapa. Pero también se puede plantear un muestreo en sucesivas etapas (escalones) en la que se consideran unidades terciarias, cuartas, etc. En la primera etapa se realiza una selección aleatoria de los conglomerados considerados como unidades primarias, después cada conglomerado seleccionado se divide también en conglomerados y se extrae nuevamente de forma aleatoria otra muestra de conglomerados como unidades

secundarias. Si tomamos todas las unidades terciarias de estos conglomerados tendremos un muestreo por conglomerados bietápico, también llamado muestreo por conglomerados con submuestreo. Si consideramos más conglomerados el muestreo será progresivamente de tres o más etapas, denominándose muestreo polietápico por conglomerados. En cada extracción aleatoria se pueden seguir varias estrategias de muestreo.

Como ejemplos de conglomerados podríamos considerar cualquier división territorial que, partiendo de la unidad elemental final, el individuo, considerara sucesivamente unidades agregadas o conglomerados en términos de hogares, manzanas, secciones censales, barrios, distritos, municipios, comarcas, provincias, regiones. Las urnas electorales representan igualmente conglomerados del comportamiento del voto. Los departamentos universitarios son unidades agregadas de su profesorado que se pueden considerar en un muestreo por conglomerados. En un estudio para medir la satisfacción de los usuarios/as del transporte público se pueden considerar los como conglomerados. Los centros penitenciarios, los asilos, las escuelas, los centros hospitalarios, los museos, etc. uelen ser consideradas como unidades primarias naturales, es decir, un tipo de unidad agregada propia de la organización social, política, económica o cultural de una sociedad.

En relación a la determinación del número de conglomerados y su tamaño se pueden hacer las siguientes observaciones como condiciones de eficiencia. Por un lado, el número de conglomerados debe ser suficientemente grande para que actúe la ley de los grandes números y se produzcan compensaciones entre los conglomerados extraídos. Por otro, los conglomerados deben tener un tamaño moderado pero suficiente, es decir, deben garantizar la suficiente variabilidad (o la máxima heterogeneidad) en su interior sin constituir conglomerados excesivamente grandes. Además, el tamaño de los conglomerados debe ser lo más parecido posible. En consecuencia, los conglomerados entre sí deben ser lo más parecidos posible.

2.5. El muestreo polietápico

El muestreo polietápico, también llamado submuestreo, en varias etapas, bietápico, trietápico, etc. más que un tipo de muestreo es un diseño muestral que combina más de un tipo de muestreo. El método consiste en designar las unidades en cascada, por sucesivas etapas. En la primera etapa se escogen las unidades primarias. En la segunda se escogen dentro de cada unidad primaria de la muestra, una muestra de unidades secundarias, y así sucesivamente con unidades posteriores, combinando diversas estrategias de muestreo en cada etapa.

El objetivo es repartir óptimamente la muestra entre las unidades para simplificar el trabajo de tratamiento de la base de la muestra y permitir la concentración geográfica de las unidades a observar, aunque suele ser menos eficiente en beneficio de una reducción de los costes. Cuando las unidades primarias son muy heterogéneas respecto a las características del estudio conviene definir unidades primarias grandes y aumentar el número de unidades secundarias. Cuando las unidades primarias son homogéneas conviene aumentar el número de unidades primarias y reducir las unidades secundarias.

El tamaño de las unidades puede ser igual (sería lo deseable) o diferente (es lo habitual). Para elegir a las unidades se debe garantizar que las de la última etapa tengan igual probabilidad de pertenecer a la muestra (muestreo autoponderado) y, por tanto, sean representativas de la población. La probabilidad de selección de una unidad final está determinada por la probabilidad acumulativa de cada etapa o nivel.

3. El muestreo no probabilístico

Hay dos tipos básicos de muestras, o de selecciones muestrales: el muestreo probabilístico, lo que hemos tenido ocasión de ver hasta ahora, y el muestreo no probabilístico o no aleatorio.

En todos los métodos destinados a obtener una muestra aleatoria, los elementos de la población se seleccionan de tal manera que cada uno tiene una probabilidad no nula y conocida de ser incluido en la muestra. La muestra de la población pues se debe seleccionar por métodos aleatorios que aseguren la extracción independientemente de cualquier juicio subjetivo, o no más allá de los criterios definidos en la investigación que orientan el diseño de la muestra. Esto significa que podemos elaborar conclusiones de inferencia estadística desde los datos de la muestra en los datos de la población. Siempre que se planteen estudios extensivos, en particular por encuesta, la muestra aleatoria es la más conveniente y la más representativa de la población y en consecuencia es la que menos sesgos comportará. En las muestras perfectamente aleatorias las diferencias entre los datos de la muestra y los atribuibles a la población son únicamente debidos a los errores muestrales, siempre que los errores de medida y otros errores sistemáticos no introduzcan un grado de invalidez y pese a que una muestra perfectamente aleatoria es un ideal difícilmente alcanzable.

Por su parte, las muestras no probabilísticas se seleccionan en base a la apreciación de los investigadores/as en función de determinados objetivos analíticos propios y particulares. En ellas algunas unidades de la base de sondeo tienen una probabilidad diferente y desconocida de salir a la muestra en relación a otras unidades. Por tanto, las muestras no probabilísticas se fundamentan en el criterio de selección del propio investigador/a según los objetivos de la investigación y con un juicio y decisiones objetivadas que juega una función clave para determinar qué unidades han formar parte de la muestra. Este criterio puede estar fundamentado sobre la conveniencia y la facilidad, aunque justificado, o sobre la base de una norma sistemáticamente empleada.

De hecho, las muestras no probabilísticas se utilizan en muchas investigaciones, ya sean extensivas y para producir información de tipo cuantitativo como en estudios de orientación cualitativa. La estrategia no probabilística se plantea en particular:

- Cuando un muestreo probabilístico es prohibitivo económicamente.
- Cuando es difícil de implementar, en particular, si no se dispone de un marco de muestreo adecuado por varias razones: la población a analizar está muy dispersa, o tiene (se le ha impuesto) algún estigma social que no permite su libre identificación: drogadictos, enfermos de SIDA, prostitución, alcohólicos, criminales, inmigrantes, élites económicas y sociales, etc.
- Cuando no son necesarias representaciones precisas: en los primeros estadios de la investigación para explorar determinados conceptos, testar el cuestionario, etc.,

donde no se está interesado en extender la representatividad a toda la población, como por ejemplo en ensayos, para la construcción de escalas, para generar hipótesis, en análisis exploratorios, para construir modelos, etc.
- Cuando corresponde a un diseño de análisis cualitativo en el que la selección de las unidades informantes responde a un criterio intensivo y calificado o analítico para dar cuenta en profundidad de un rasgo, vivencia o dinámica social.

Veremos seguidamente algunos tipos de muestreo no probabilístico.

3.1. Muestreo por cuotas

Es un tipo de muestreo muy generalizado en las investigaciones aplicadas por encuesta porque ha demostrado su eficiencia en relación a los diseños más cuidadosos de las muestras probabilísticas, y de hecho se utiliza a menudo como si se tratara de un método probabilístico. Pero en la medida en que no parte de una extracción aleatoria sobre la base de un marco de muestreo no puede equipararse. En su lugar, esta estrategia de muestreo deja finalmente la elección de la unidad muestral al criterio de los encuestadores/as a partir de determinadas indicaciones o perfiles que debe respetar, siguiendo, en particular, estrategias pseudo-aleatorias con las llamadas rutas aleatorias.

El procedimiento es sencillo y se puede resumir en las siguientes tareas básicas:

1. Elección de las variables criterio y de sus valores correspondientes para establecer las cuotas poblacionales y de muestreo.
2. Cruce de las variables con sus valores y determinación, a partir de la tabla correspondiente, de las cuotas que corresponde a cada casilla.
3. Atribución a cada casilla o cuota del peso que le corresponde en términos de la proporción que estas características tienen en la población total.
4. Atribución de la cuota o números de elementos de la muestra correspondiente a cada perfil.
5. Orientaciones e instrucciones a los encuestadores/as para elegir convenientemente a los entrevistados/as.

En primer lugar, pues, se divide la población en subgrupos según los valores de las variables que se consideran son más influyentes en el modelo de análisis que se lleva a cabo. Estas variables suelen ser las denominadas estructurales, variables independientes básicas (sociodemográficas) relacionadas con los objetivos de la investigación como, por ejemplo, el sexo para definir cuotas de hombres y mujeres, la edad para determinar subgrupos de personas más jóvenes o mayores, el nivel educativo alcanzado para configurar cuotas educativo-culturales, etc. Evidentemente el número de variables a retener y el de los valores de cada una de ellas no debe ser elevado ya que la recogida de datos se complicaría enormemente. Si consideramos varias de estas variables se configura una matriz con forma de tabla de contingencia donde cada casilla, resultado del cruce de los valores de estas variables, configura una cuota de características de la población (conocidas habitualmente a partir de datos censales) cuya proporción se debe respetar en los elementos seleccionados finalmente en la muestra seleccionada. Si todos los elementos están bien ponderados la muestra debería ser una razonable representación de la población.

Imaginemos que queremos realizar un estudio utilizando el muestreo por cuotas teniendo en cuenta el nivel de estudios, la edad y el sexo de la población ocupada española. En la Tabla II.4.4 aparecen las cuotas poblacionales que deberán respetarse en la extracción de la muestra. Contiene los porcentajes sobre el total de la población ocupada de 16 y más años, 17.514.455 personas según los datos del Censo de Población de 2011, de cada combinación de categorías de las tres variables.

Tabla II.4.4. Cuotas poblacionales según Estudios, Edad y Sexo. Porcentaje sobre el total de la población de 15 y más años.

Cuotas poblacionales (%)		Estudios		
		Primarios	Secundarios	Terciarios
Varones	16-29	0,93	2,79	2,66
	31-49	5,18	20,98	8,33
	50 o más	1,54	8,87	3,40
Mujeres	16-29	0,43	1,61	2,20
	31-49	4,25	14,60	6,01
	50 o más	2,65	10,63	2,93

Fuente: INE, Censo de Población 2011

La Tabla II.4.5 aplica esas cuotas a un tamaño de muestra, en este caso sobre 1.000 individuos.

Tabla II.4.5. Cuotas muestrales según Estudios, Edad y Sexo.

Cuotas muestrales (encuestas)		Estudios			**Total**
		Primarios	Secundarios	Terciarios	
Varones	16-29	9	28	27	**64**
	31-49	52	210	83	**345**
	50 o más	15	89	34	**138**
Mujeres	16-29	4	16	22	**42**
	31-49	42	146	60	**249**
	50 o más	26	106	29	**162**
Total		**142**	**521**	**201**	**1.000**

Fuente: INE, Censo de Población 2011

A partir de estas cuotas los encuestadores reciben instrucciones específicas para recoger el número de cuestionarios que aseguran las proporciones poblacionales en términos de cuotas de muestra. Así, por ejemplo, un encuestador en concreto puede tener el encargo de encuestar a 50 personas en un determinado barrio, 23 deben ser mujeres y 26 varones; los entrevistados de cada sexo se distribuirán en tres grupos de edades y tres grupos de estudios de la forma que recoge la Tabla II.4.6.

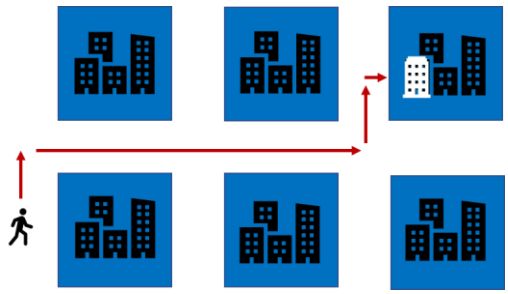
Tabla II.4.6. Asignación de encuestas al encuestador/a

		Estudios			Total
		Primarios	Secundarios	Terciarios	**Total**
Varones	16-29	1	1	1	**4**
	31-49	3	9	4	**17**
	50 o más	1	4	2	**7**

Mujeres	16-29	1	1	1	**3**
	31-49	2	7	3	**12**
	50 o más	1	5	1	**8**
Total		**7**	**26**	**10**	**50**

Los encuestadores/as tienen que identificar los individuos de las cuotas simplemente preguntando al tomar contacto con la persona que se seleccione. Las personas se pueden seleccionar a partir de una ruta aleatoria, a la salida de algún tipo de establecimiento o edificio (centro comercial, colegio electoral, etc.), entre los peatones en lugares determinados de la ciudad, etc.

En el caso de las rutas aleatorias los elementos de la muestra se seleccionan empezando la ruta en un punto elegido al azar dentro de un área geográfica. Desde él el encuestador/a, a partir de un criterio de aleatorización predefinido, va seleccionando viviendas y a los individuos que en ellas habitan. Por ejemplo, desde el punto de arranque el encuestador/a camina dos manzanas a la derecha, gira a la izquierda y en la acera de enfrente selecciona de forma aleatoria una vivienda de un edificio y una persona que se corresponda a las cuotas asignadas.



Con las distintas rutas elegidas se trata de alcanzar la mayor cobertura geográfica.

Con esta estrategia se busca implementar un procedimiento de selección semialeatorio o semiprobabilístico pues no garantiza las condiciones exigibles a una muestra estrictamente aleatoria, en particular, se desconoce la probabilidad de ser elegido un individuo y no se garantiza que todos tengan la misma probabilidad de ser seleccionados. De ello se deriva que los razonamientos inferenciales o probabilísticos no se puedan plantear en sentido estricto. Aun así, se suelen aplicar los mismos conceptos, formulaciones y conclusiones estadísticas, si bien contendrán un grado de incertidumbre mayor que en muchas ocasiones generará resultados carentes de la suficiente fiabilidad.

El método tiene bastantes similitudes con el muestreo aleatorio estratificado, de tipo proporcional. También se trata de dividir la población en subpoblaciones, y aquí las cuotas actúan a modo de estratos de características poblacionales que son conocidas, pero con una diferencia importante: la muestra no se extrae de forma probabilística, sino que la selección de los individuos depende finalmente del juicio de los entrevistadores/as. El muestreo por cuotas representa una buena alternativa cuando el criterio de economía de costes prima a la investigación, pero, al no tratarse de una muestra aleatoria nunca podemos saber los errores que se cometen, y el nivel de

precisión supuesto depende de la idoneidad del modelo de que justifica los criterios de selección de las unidades.

Pero además se pueden introducir sesgos específicos: a veces no se dispone de la información suficiente para atribuir la cuota en cada casilla; pese a que las cuotas sean correctas se pueden dar sesgos en la selección de las unidades de la muestra tendiendo a escoger personas fácilmente accesibles o condicionadas por criterios subjetivos del entrevistador (dejar las personas que viven en pisos altos sin ascensor, pisos de más difícil acceso o apariencia de abandono, escoger personas afines al entrevistador/a, buscar a las personas en determinados momentos del día, etc.) que el procedimiento de rutas aleatorias con una hoja de ruta bien detallado intenta corregir; puede resultar difícil completar determinados perfiles muy específicos definidos por una cuota (por ejemplo, "mujer de categoría profesional alta de 30 a 45 años residente en un barrio determinado"). Finalmente puede suceder que la muestra realmente mantenga las proporciones poblacionales definidas por las variables que configuran las cuotas, pero en relación a otras variables estudiadas no se dé una verdadera representatividad.

Para mejorar la eficiencia en la utilización de este tipo de muestreo habitualmente se combina con otras estrategias en un diseño polietápico, por ejemplo, estratificando geográficamente en una primera etapa (o más etapas) y escogiendo las unidades finales con las cuotas definidas en la segunda etapa (o ulterior etapa).

3.2. Muestra razonada

La muestra razonada se realiza sobre la base del conocimiento que el investigador/a tiene de la población, de los elementos que la componen y de las líneas maestras de la investigación. Entonces los casos son juzgados de interés para la investigación, pero no son seleccionados aleatoriamente. Veamos algunos casos:

1. Para realizar un pre-test, por ejemplo, para avalar el diseño de un cuestionario.
2. A veces se desea estudiar una amplia población en la que muchos miembros de un determinado subconjunto de la misma son fácilmente identificables pero la enumeración de todos ellos sería prácticamente imposible. Por ejemplo, si se quiere estudiar el liderazgo de un movimiento social amplio y extendido muchos de los líderes son fácilmente localizables dada su visibilidad, actuación o representatividad, pero hacer una muestra de todos ellos puede resultar imposible; estudiar una parte de los más visibles se pueden tener los datos suficientes para el propósito del estudio.
3. Aplicado a los sondeos electorales se puede escoger, en base a resultados precedentes, una muestra que sea lo más representativa de los resultados globales y que sirva en sucesivas elecciones y sondeos para evaluar su representatividad.

Otras muestras que tienen casi el mismo fundamento son las siguientes:

– La muestra de un caso típico. Aplicable cuando se tienen limitaciones de tiempo y presupuesto. Se seleccionan pocos casos típicos que sean los más habituales en relación a la población que se desea estudiar. El juicio del investigador/a es decisivo, pero también otros resultados de estudios previos que pueden orientar

en esta selección. El sesgo de este tipo de selección es evidente, pero de él se es suficientemente consciente cuando se realiza este tipo de muestra.

- La muestra de un caso crítico. Es similar al caso típico, pero en lugar de escoger entre los casos más normalizados se toman los individuos entre los diferentes grupos de la población que mediante su heterogeneidad se parezcan lo más posible a la población, según también la experiencia y los conocimientos del investigador/a.
- La muestra más similar o disimilar. Se trata de muestras sobre todo para análisis comparativos entre casos, países, comunidades, etc. y se dispone de un número limitado de casos.

3.3. Muestra de conveniencia

Es un tipo de muestreo en el que las unidades están disponibles y son fáciles de localizar, tienen un carácter de representatividad de la población que se quiere analizar, pero se hace una selección conveniente de varias unidades con el objetivo de constituir grupos reducidos y controlados en el contexto de diseños de tipo experimental. Así se configuran grupos de control y experimentales a partir de la elección aleatoria de unos cuantos individuos que los forman.

3.4. Muestra de bola de nieve

En la muestra de bola de nieve a partir de un núcleo básico de muestra de pocos casos que reúne una serie de características de interés para el estudio, la muestra se edifica progresivamente y se va ampliando a partir de la relación existente o de la elección que un mismo miembro o unidad de la muestra inicial realiza en sugerir o vincularse con otros miembros, a medida que se incorpora un nuevo elemento relacionado con el anterior se construye un red de relaciones de las unidades muestrales que va creciendo, literalmente, como un bola de nieve. Proceder de esta manera se justifica muchas veces ante la imposibilidad de reconocer o localizar los individuos de una población específica como pueden ser los consumidores de drogas, los trabajadores/as ilegales, los miembros de comunidades marginales, de élites de poder, etc.

4. Ejemplos de diseños muestrales

En este último aparatado del capítulo presentaremos brevemente algunos ejemplos de diseños muestrales elaborados en el contexto de la producción de información a través de encuestas sociales.

4.1. Sondeo electoral

Los sondeos electorales son un ejemplo característico de investigación social aplicada que de forma cotidiana nos informa sobre el pulso político del momento. Tomaremos como ejemplo el Barómetro Electoral de la empresa Metroscopia de julio de 2017. En su página web (http://www.metroscopia.org), con fecha de 21 de agosto, tenía publicados estos resultados:

Tabla II.4.7. Resultados del sondeo electoral de Metroscopia

		Intención de voto declarada	Recuerdo de voto declarado	Resultado real 26J sobre censo
Unidos Podemos	UP / Podemos	15.2	18.4	9.3
	IU	1.6		
	En Comú Podem	0.5	0.7	2.4
	Compromís – EUPV	0.3	0.4	1.9
	En Marea – Anova - EU	0.1	0.5	1.0
		17.7	20.0	14.6
PP		17.2	22.6	22.9
PSOE		15.7	17.1	15.7
Ciudadanos		15.3	12.1	9.0
Otros y blanco		10.7	9.5	7.6
Total		76.6	81.3	69.8
Abstención		*23.4*	*18.7*	*30.2*

La **Intención Directa de Voto** es **la respuesta espontánea** dada por cada persona entrevistada al preguntarle por el posible sentido de su voto en unas hipotéticas inminentes elecciones.

Corresponden a la estimación del resultado electoral que se obtiene a partir de un trabajo de campo por muestreo cuyas características se presentan en forma de ficha técnica:

> FICHA TÉCNICA: El sondeo se ha efectuado mediante entrevistas telefónicas a una muestra nacional de personas residentes en España, mayores de 18 años y con derecho a votar en elecciones generales. Se han completado 1.532 entrevistas a través de llamadas a teléfonos móviles seleccionados de forma aleatoria a partir de un generador automático de números telefónicos. Posteriormente se han calibrado los datos a partir de una ponderación múltiple por sexo, edad, hábitat y región (Comunidad Autónoma). La eficiencia de la ponderación es del 66.4%, de modo que la muestra efectiva equivale a 1.017 entrevistas. El error de muestreo, para un nivel de confianza del 95.5% (que es el habitualmente adoptado) y asumiendo los principios del muestreo aleatorio simple, en la hipótesis más desfavorable de máxima indeterminación (p=q=50%), es de ±2.6 puntos (tras la ponderación, ±3.1 puntos). La recogida de información y el tratamiento de la misma han sido llevados a cabo íntegramente en Metroscopia. Fecha de realización del trabajo de campo: entre los días 27 y 28 de junio de 2017.

En este caso el muestreo de personas se realiza a través de una selección aleatoria de números de teléfono y, como se explicita, ajustando la muestra según cuotas de sexo, edad, hábitat (tamaño del lugar de residencia) y Comunidad Autónoma. Se trata pues de un muestreo aleatorio simple de individuos seleccionados a partir de un marco muestral, como es la base de datos de teléfonos, que seguramente genera algún sesgo de selección que se busca corregir con esa ponderación de cuotas poblacionales conocidas. De esa forma el tamaño inicial de la muestra de 1.532 entrevistas se ve reducido hasta 1.017. Considerando un nivel de confianza del 95,5% (z=2), en el supuesto de máxima indeterminación (P=Q=50%), tratándose de una población infinita, el error muestral es igual a 3,14%:

$$e = z \times \sqrt{\frac{P \times Q}{n}} = 2 \times \sqrt{\frac{50 \times 50}{1.017}} = 3{,}14$$

4.2. Barómetro del Centro de Investigaciones Sociológicas

Los Barómetros del Centro de Investigaciones Sociológicas (http://www.cis.es) se realizan con una periodicidad mensual con el objetivo medir el estado de la opinión pública española del momento y donde se recoge igual igualmente información social y demográfica para el análisis. Son encuestas con un cuestionario estandarizado donde se entrevistan alrededor de 2.500 personas dentro del territorio nacional.

Los barómetros se caracterizan por contener un bloque de preguntas fijas con las que se elaboran los "indicadores del barómetro". Además, cada barómetro contiene otro bloque de preguntas variable dedicado a un tema de interés político o social. Los meses de enero, abril, julio y octubre los barómetros incluyen un conjunto de preguntas fijas sobre actitudes políticas a partir de las que el CIS calcula y publica la estimación de voto.

Los resultados de los barómetros mensuales se hacen públicos a través de la web del CIS inicialmente en formato de "avance de resultados" y tras la finalización del resto de procesos técnicos, incluida la anonimización, los datos del estudio pasan a formar parte del Banco de Datos del CIS, cuyos microdatos junto al resto de la documentación asociada pueden descargarse de la página web.

Entre las características del barómetro se encuentran las siguientes (CIS, 2017):

- Encuesta personal realizada en hogares.
- Ámbito nacional.
- El universo es la población española de más de 18 años.
- El tamaño de la muestra es de 2.500 unidades últimas o individuos.
- Se emplea la afijación proporcional.
- El procedimiento de muestreo es polietápico, estratificado por conglomerados, con selección de las unidades primarias de muestreo (municipios) y de las unidades secundarias (secciones) de forma aleatoria proporcional, y de las unidades últimas (individuos) por rutas aleatorias y cuotas de sexo y edad. Los estratos se han formado por el cruce de las 17 comunidades autónomas con el tamaño de hábitat dividido en siete categorías: menor o igual a 2.000 habitantes; de 2.001 a 10.000; de 10.001 a 50.000; de 50.001 a 100.000; de 100.001 a 400.000; de 400.001 a 1.000.000; y más de 1.000.000 de habitantes.
- El error de muestreo, bajo el supuesto de muestreo aleatorio simple, se sitúa en ±1,9% para el conjunto de la muestra, para un nivel de confianza del 95,5% (dos sigmas) y la situación más desfavorable (P=Q).
- El trabajo de campo se realiza en la primera quincena de cada mes, con la excepción del mes de agosto, que no se realiza.

Si tomamos el estudio nº 3179 correspondiente al barómetro del mes de junio de 2017, la ficha técnica que publica el CIS presenta la información siguiente:

ESTUDIO CIS Nº 3179
BARÓMETRO DE JUNIO
FICHA TÉCNICA

Ámbito: Nacional.
Universo: Población española de ambos sexos de 18 años y más.
Tamaño de la muestra:
Diseñada: 2.500 entrevistas.
Realizada: 2.492 entrevistas.
Afijación: Proporcional.
Ponderación: No procede.
Puntos de Muestreo: 253 municipios y 47 provincias.
Procedimiento de muestreo: Polietápico, estratificado por conglomerados, con selección de las unidades primarias de muestreo (municipios) y de las unidades secundarias (secciones) de forma aleatoria proporcional, y de las unidades últimas (individuos) por rutas aleatorias y cuotas de sexo y edad. Los estratos se han formado por el cruce de las 17 comunidades autónomas, con el tamaño de hábitat, dividido en 7 categorías: menor o igual a 2.000 habitantes; de 2.001 a 10.000; de 10.001 a 50.000; de 50.001 a 100.000; de 100.001 a 400.000; de 400.001 a 1.000.000, y más de 1.000.000 de habitantes. Los cuestionarios se han aplicado mediante entrevista personal en los domicilios.
Error muestral: Para un nivel de confianza del 95,5% (dos sigmas), y P=Q, el error real es de ±2,0% para el conjunto de la muestra y en el supuesto de muestreo aleatorio simple.
Fecha de realización: Del 1 al 9 de junio de 2017.

Como en el ejemplo anterior, considerando las condiciones del muestreo aleatorio simple y semejantes supuestos, el tamaño de la muestra de 2.492 entrevistas se obtiene adoptando un nivel de confianza del 95,5% (z=2), en el supuesto de máxima indeterminación (P=Q=50%), tratándose una población infinita y con un error muestral fijado del 2%, así calculado[11]:

$$n = \frac{z^2 \times P \times Q}{e^2} = \frac{2^2 \times 50 \times 50}{2,0016^2} = 2.496$$

4.3. La Muestra Continua de Vidas Laborales (MCVL)

La Muestra Continua de Vidas Laborales (MCVL) es un tipo particular de muestra pues se tratar de muestrear los registros administrativos de la Seguridad Social[12]. La MCVL es una fuente de información que, desde el año 2004, nos ofrece datos relativos de la distribución de la población de un determinado año según diversas características laborales registradas administrativamente a través de la Seguridad Social, así como de su historia laboral desde que existen registros informatizados.

Los microdatos de la MCVL se obtienen a partir de una muestra diseñada tomando como población de referencia una definición amplia: «Todas las personas que han estado en situación de afiliado en alta, o recibiendo alguna pensión contributiva de la Seguridad Social en algún momento del año de referencia» (MTIN, 2006). Dos situaciones, cotizantes o pensionistas, que, por las características de construcción de la muestra, pueden darse de forma sucesiva o simultáneamente. Así, desde el punto de

[11] De hecho, el error real es del 2,0016% que a dos decimales es el 2,00%.
[12] Se puede consultar en la dirección siguiente del Ministerio: http://www.seg-social.es/Internet_1/Estadistica/Est/Muestra_Continua_de_Vidas_Laborales/index.htm.

vista del ámbito temporal, se consideran a todos los individuos que han estado en relación con la Seguridad Social en algún momento del año, no en una fecha fija, para facilitar así la presencia en la muestra de personas que trabajan regularmente pero que entran y salen de manera continuada de una situación de alta laboral.
Por tanto, para datos globales no se trata tanto de la población activa, como de los perceptores de ingresos , donde se distinguen tres colectivos básicos: los empleados que están en alta laboral (trabajan y cotizan por cuenta propia o ajena); los cotizantes para pensiones que no trabajan (convenio especial, incapacidad transitoria y los perceptores de prestaciones de desempleo no contributivas); y los perceptores de pensiones contributivas (jubilación, incapacidad permanente), incluyendo las generadas por el Seguro Obligatorio de Vejez e Invalidez (SOVI) y las pensiones de supervivencia (viudedad y orfandad) . Un cuarto colectivo se incluye en la muestra al contemplar las personas que perciben un subsidio de desempleo. En conjunto pues se trata de personas activas (ocupados y desempleados) y no activas que mantienen una relación administrativa con la Seguridad Social.

La población de referencia es de unos 30.000.000 de personas. Sobre esta población el tamaño de la muestra se fija en el 4%, un total aproximado de 1.200.000 individuos, un tamaño al que se asocia un factor de elevación de 25 y un error muestral muy bajo: para datos globales, una muestra de esa magnitud, si consideráramos un nivel de confianza del 95,5%, en el supuesto de una estimación porcentual asumiendo que P=Q=50%, arroja un error de tan solo el 0,09%.

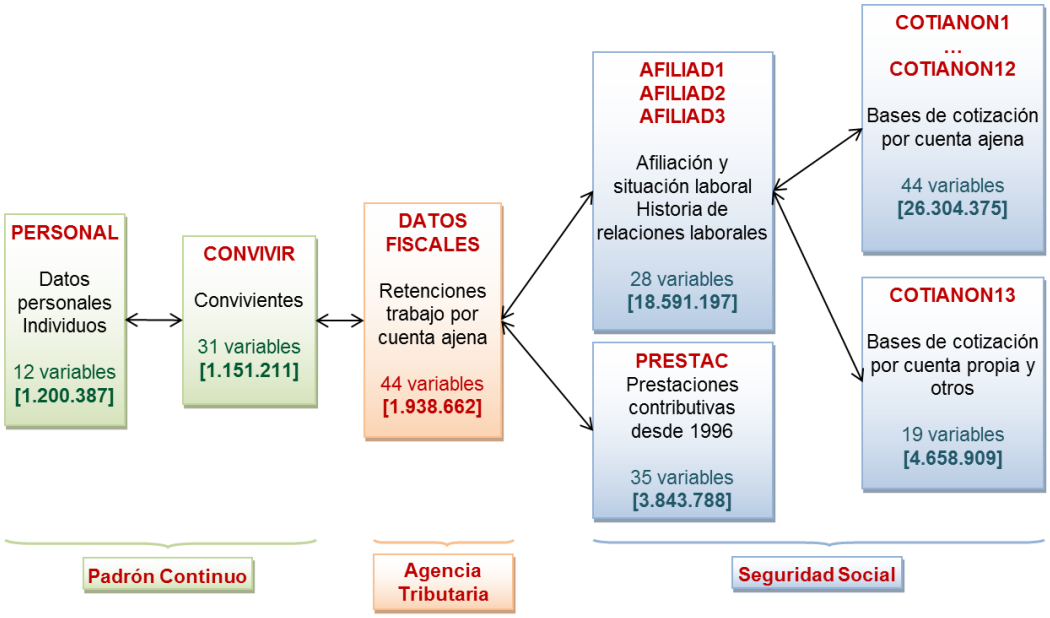
El tipo de muestreo es aleatorio simple, manteniendo una estructura de panel desde la primera extracción aleatoria del año 2004. De esta forma las unidades se extraen manteniendo los códigos de la primera que aseguran que siempre serán seleccionadas las mismas personas, en la medida que continúen manteniendo relación con la Seguridad Social, y renovándose de forma representativa con nuevas incorporaciones.

Cabe destacar en consecuencia que la muestra es representativa de la situación de cada año, sobre la base de una misma muestra panel inicial, pero no de las cohortes anteriores al año de referencia. Sí lo es de las trayectorias pasadas en relación a la situación actual, entendiendo que los diferentes episodios son acontecimientos propios de cada individuo en su vida laboral hasta el año de referencia.

La información de la Seguridad Social se completa con datos del Padrón Municipal Continuo y con datos fiscales de la Agencia Tributaria. Así, los datos de la MCVL se organizan en torno a la persona física, con una variable de identificación que permite vincular todos los ficheros como se muestra en el Gráfico II.4.8., entre paréntesis aparece el número de registros de cada base de datos.

Remitimos al lector/a interesado a las referencias de la literatura sobre la muestra (MTIN, 2006; Durán, 2007; García Pérez, 2008; Lapuerta, 2010, López-Roldán, 2011; Miguélez y López-Roldán, 2014).

Gráfico II.4.8. Estructura de los ficheros de la MCVL (2011)



4.4. Encuesta de Condiciones de Vida (ECV)

Un diseño mucho más complejo es el de La Encuesta de Condiciones de Vida (ECV) del Instituto Nacional de Estadística (http://www.ine.es; INE, 2013), que se realiza desde el año 2004 siguiendo criterios armonizados para todos los países de la Unión Europea (EU-SILC, http://ec.europa.eu/eurostat/web/income-and-living-conditions) . La ECV tuvo su antecesora en el Panel de Hogares de la Unión Europea (PHOGUE), durante el periodo 1994-2001, de características y objetivos similares. Su objetivo fundamental es disponer de una fuente de datos de referencia sobre estadísticas comparativas de la distribución de ingresos y la exclusión social en el ámbito europeo que permita el estudio de la pobreza y desigualdad, el seguimiento de la cohesión social en el territorio de su ámbito, el estudio de las necesidades de la población y del impacto de las políticas sociales y económicas sobre los hogares y las personas, así como para el diseño de nuevas políticas.

Se trata de una encuesta de periodicidad anual dirigida a la población que reside en viviendas familiares principales tomando como período de referencia de los ingresos el año natural anterior a la entrevista. También proporciona información longitudinal ya que es una encuesta de panel rotante en la que las personas entrevistadas colaboran cuatro años seguidos, por ello la muestra la forman cuatro submuestras independientes y cada año se renueva la muestra en uno de los paneles. De esta forma se puede conocer la evolución de las variables investigadas a lo largo del tiempo.

La ECV está diseñada para obtener información sobre ingresos de los hogares privados y en general sobre su situación económica, pobreza, carencias, protección social e igualdad de trato, empleo y actividad, jubilaciones, pensiones y situación socioeconómica de las personas mayores, vivienda y costes asociados a la misma,

desarrollo regional, nivel de formación, salud y efectos de ambos sobre la condición socioeconómica.

La población objeto de estudio son las personas residentes en España miembros de hogares privados que residen en viviendas familiares principales, así como dichos hogares. Aunque todas las personas forman parte de la población objetivo no todas las personas son investigadas exhaustivamente, sólo los que tengan 16 años o más a 31 de diciembre del año anterior al de la entrevista. No se incluyen las personas que viven en hogares colectivos. Las unidades estadísticas consideradas son los hogares y las personas miembros de dichos hogares.

El ámbito geográfico es todo el territorio español y se publican resultados tanto para el total nacional como a nivel de comunidad autónoma.

Para cubrir los objetivos de la encuesta de proporcionar estimaciones con un grado de fiabilidad aceptable en el ámbito nacional y de Comunidad Autónoma, se selecciona una muestra de 16.000 viviendas distribuidas en 2.000 secciones censales[13]. La muestra se distribuye entre Comunidades Autónomas asignando una parte uniformemente y otra proporcionalmente al tamaño de la Comunidad. La parte uniforme es aproximadamente el 40% de las secciones. En cada sección se seleccionan ocho viviendas titulares, así como otras ocho viviendas para sustituir las incidencias habidas en las viviendas titulares. El tamaño muestral es de alrededor de 13.000 hogares y 35.000 personas.

El tipo de muestreo es bietápico con estratificación en las unidades de primera etapa. Las unidades de muestreo de primera etapa son las secciones censales y las de segunda etapa son las viviendas familiares principales. Dentro de ellas no se realiza submuestreo alguno, investigándose a todos los hogares que tienen su residencia habitual en las mismas. Se consideran así dos unidades básicas de observación y análisis: los hogares privados que residen en las viviendas familiares principales seleccionadas en la muestra, y las personas miembros de dichos hogares. Las secciones se seleccionan dentro de cada estrato con probabilidad proporcional a su tamaño. Las viviendas, en cada sección, con igual probabilidad mediante muestreo sistemático con arranque aleatorio. Este procedimiento conduce a muestras autoponderadas en cada estrato.

El marco utilizado para la selección de la muestra es un marco de áreas formado por la relación de secciones censales utilizadas en el Padrón Municipal de habitantes. Para las unidades de segunda etapa se utiliza la relación de viviendas familiares principales en cada una de las secciones seleccionadas para la muestra. En cada Comunidad Autónoma las unidades de primera etapa se agrupan en estratos de acuerdo con el tamaño del municipio al que pertenece la sección. Se consideran los siguientes estratos:

- Estrato 0: Municipios de más de 500.000 habitantes.
- Estrato 1: Municipio capital de provincia (excepto los anteriores).
- Estrato 2: Municipios con más de 100.000 habitantes (excepto los anteriores).
- Estrato 3: Municipios de 50.000 a 100.000 habitantes (excepto los anteriores).

[13] En la encuesta de 2016, en Cataluña se ha realizado una ampliación de muestra en colaboración con el Instituto de Estadística de Cataluña (Idescat), incorporando 224 secciones censales adicionales. Se puede consultar la web del Idescat: https://www.idescat.cat/pub/?id=ecv&lang=es, y del Instituto de Estudios Regionales y Metropolitanos de Barcelona: https://iermb.uab.cat/es/encuestas/cohesion-social-urbana.

- Estrato 4: Municipios de 20.000 a 50.000 habitantes (excepto los anteriores).
- Estrato 5: Municipios de 10.000 a 20.000 habitantes.
- Estrato 6: Municipios con menos de 10.000 habitantes.

La recogida de información se realiza mediante entrevista personal si bien los datos relativos a ingresos del hogar se construyen combinando la información proporcionada por el informante con registros administrativos.

Respecto al ámbito temporal, los periodos de referencia en el cuestionario son diferentes en función del apartado. Se considera la semana de referencia, definida como la semana inmediatamente anterior (de lunes a domingo) a la de la entrevista según el calendario. El momento actual en preguntas relacionadas con la actividad. El año natural anterior a la realización de la encuesta relacionadas con las rentas percibidas el último año. Y el periodo biográfico de la persona, diferente para cada una. En algunas preguntas aisladas se recoge información referida a otros momentos particulares del tiempo.

En relación con el error de muestreo, al tratarse de un diseño complejo con información diversa es difícil dar un dato de referencia. El último dato disponible del error estándar de la tasa de riesgo de pobreza para el total nacional (encuesta de 2015) ronda el 2,7%[14].

Los errores ajenos al muestreo derivados de las tasas de no respuesta son los siguientes:

- Proporción de contacto con la vivienda: 0,99.
- Proporción de hogares entrevistados aceptados para la base de datos: 0,73.
- Tasa de no respuesta del hogar 27,7%.
- Proporción de personas entrevistadas en hogares aceptados para la base de datos: 0,98.
- Tasa de no respuesta individual: 1,61%
- Tasa de no respuesta individual total 28,88%

Otros aspectos más precisos y técnicos como el factor de elevación o el factor por grupos de edad, así como los conceptos y definiciones utilizados en la encuesta, otros aspectos metodológicos y el propio cuestionario se pueden consultar en la página web del INE (INE,2013).

4.5. Diseño de una muestra estratificada

En el contexto de la Encuesta de Condiciones de Vida y Hábitos de la Población de Cataluña, que en su primera edición en 1985 recibió el nombre de "Enquesta Metropolitana de Barcelona", se aplicó un diseño muestral estratificado que se aplicó durante 5 ediciones de la encuesta (1985, 1990, 1995, 2000 y 2006). En la actualidad la encuesta se ha integrado en la Encuesta de Condiciones de Vida que acabamos de comentar.

[14] Para obtener más información sobre los errores de muestreo se pueden consultar los informes de calidad para Eurostat: http://ec.europa.eu/eurostat/web/income-and-living-conditions/quality/eu-and-national-quality-reports.

El tipo de diseño implicaba la construcción de los estratos poblacionales a partir de cada uno de los cuales extraer una muestra aleatoria. Para ello se siguió una metodología de interés destinada a la construcción de una tipología de estratos sociales a partir de los datos censales y mediante técnicas de análisis multivariable donde se combinan el análisis factorial de componentes principales y el análisis de clasificación. No daremos cuenta en estas páginas de las características de ese diseño. Remitimos al lector/a las publicaciones derivadas de este trabajo de diseño muestral (López-Roldán, P.; Lozares, C.; Domínguez, M., 2000; López-Roldán, P.; Lozares, C., 2007; López-Roldán, P.; Lozares, C., 2008) y a los capítulos III.11 de análisis factorial y III.12 de análisis de clasificación donde se utiliza la construcción de la muestra estratificada como ejemplo de uso de esas técnicas.

5. Bibliografía

Abad, A.; Servín, L. A. (1984). *Introducción al muestreo*. México: Limusa.

Azorín, F.; Sánchez-Crespo, J. L. (1986). *Métodos y aplicaciones del muestreo*. Madrid: Alianza. Alianza Universidad Textos, 99.

Babbie, E. R. (1990). *Survey Research Methods*. Belmont, CA: Wadsworth Publishing Company.

Bardina, X.; Farré, M.; López-Roldán, P. (2005). *Estadística: un curs introductori per a estudiants de ciències socials i humanes. Volum 2: Descriptiva i exploratòria bivariant*. Bellaterra (Barcelona): Universitat Autònoma de Barcelona. Col·lecció Materials, 166.

Bailey, K. (1987). Survey Sampling. En: *Methods of Social Research*. New York: Free Press, 79-103.

Blalock, H. M. (1981). *Estadística Social*. México: Fondo de Cultura Económica.

Casas, J. M. et al. (2006). *Ejercicios de inferencia estadística y muestreo: para economía y administración de empresas*. Madrid: Pirámide.

CIS (2017). *Nota de investigación sobre la metodología general de los barómetros mensuales del Centro de Investigaciones Sociológicas*. Madrid: CIS. http://www.cis.es/cis/export/sites/default/-Archivos/NotasdeInvestigacion/NI004_MetodologiaBarometros_Informe.pdf

Clairin, R. (2001). *Manual de muestreo*. Madrid: La Muralla.

Cochran, W. G. (1971). *Técnicas de muestreo*. México: CECSA.

Durán, A. (2007). La Muestra Continua de Vidas Laborales de la Seguridad Social. *Revista del Ministerio de Trabajo y Asuntos Sociales,* VI, 231-240. http://www.mtin.es/es/publica/pub_electronicas/destacadas/revista/numeros/ExtraSS07/Est09.pdf

Emmel, N. (2013). *Sampling and choosing cases in qualitative research: a realist approach*. London: Sage.

García Ferrando, M. (1994). *Socioestadística. Introducción a la estadística en sociología*. 2a edició amp. Madrid: Alianza. Alianza Universidad Textos, 96.

García Pérez, J. I. (2008). La muestra continua de vidas laborales: una guía de uso para el análisis de transiciones. *Revista de Economía Aplicada*, *XVI*, E-1, 5-28. http://www.revecap.com/revista/numeros/e1/pdf/garcia.pdf

Gondar, J. E. (2003). *Muestreo aplicado al marketing*. Madrid: Data Mining Institute.

Grosbas, J.-M. (1987). *Méthodes Statistiques des Sondages*. Paris: Economica.

Grais, B. (1977). *Méthodes statistiques*. Paris: Dunod.

Hopkins, K. D.; Hopkins, B. R.; Glass, G. V. (1997). *Estadística Básica para las ciencias sociales y del comportamiento*. 3ª. edición. Naucalpan de Juárez: Prentice-Hall Hispanoamericana.

INE (2013). *Encuesta de Condiciones de Vida. Metodología*. Madrid: INE. http://www.ine.es/daco/daco42/condivi/ecv_metodo.pdf

Jacquart, H. (1988). *Qui? quoi? comment? ou, la pratique des sondages*. Paris: Eyrolles.

Kalton, G. (1983). *Introduction to Survey Sample*. Beverly Hills: Sage Publications. Sage University Papers series on Quantitative Applications in Social Sciences, 07-035.

Kish, L. (1972). *Muestreo de encuesta*. México: Trillas.

Levy, P. S.; Lemeshow, S. (1999). *Sampling of populations: methods and applications*. New York: John Wiley & Sons, 3rd ed.

Lapuerta, I. (2010): «Claves para el trabajo con la muestra continua de vidas laborales». *DemoSoc Working Paper*, 37. Barcelona. Disponible en: <http://repositori.upf.edu/bitstream/handle/10230/6337/DEMOSOC37.pdf?sequence=1>.

López-Roldán, P. (2011). La Muestra Continua de Vidas Laborales: posibilidades y limitaciones. Aplicación al estudio de la ocupación de la población inmigrante. *Metodología de Encuestas*, 13, 7-32. http://casus.usal.es/pkp/index.php/MdE/article/view/1010

López-Roldán, P.; Lozares, C.; Domínguez, M. (2000). La construcció d'una mostra estratificada a partir de dades censals. *Qüestiió. Quaderns d'Estadística i Investigació Operativa*, 24, 1, 111-136. http://www.raco.cat/index.php/Questiio/article/view/143995/195695

López-Roldán, P.; Lozares, C. (2007). Implicaciones sociológicas en la construcción de una muestra estratificada. *Empiria*, 14: 87-108. http://e-spacio.uned.es/fez/eserv.php?pid=bibliuned:Empiria-2007-14-0001&dsID=Documento.pdf

López-Roldán, P.; Lozares, C. (2008). La construcción de la muestra. En: *El trabajo de campo de la Encuesta de condiciones de vida y hábitos de la población de Cataluña, 2006*, editado por el Institut d'Estudis Regionals i Metropolitans de Barcelona, Barcelona: IERMB, 17-39. https://iermb.uab.cat/wp-content/uploads/2015/11/Metodologies-i-recerques-1.pdf

Lohr, S. L. (1999). *Sampling: Design and Analysis*. Duxbury Press, Pacific Grove, CA.

Lozares, C.; López-Roldán, P. (1991). El muestreo estratificado por análisis multivariado. En: *El pluralismo metodológico en la investigación social: ensayos típicos*, editado por M. Latiesa. Granada: Universidad de Granada, 107-160.

Neyman, J. (1934). On two different aspects of the representative method: the method of stratified sampling and the method of purposive selection. *Journal of the Royal Statistical Society*, 97, 558-606.

Mayntz, R.; Holm, K.; Hübner, P. (1985). *Introducción a los métodos de la sociología empírica*. Madrid: Alianza.

Miguélez, F; López-Roldán, P. (Coord.) (2014). *Crisis, empleo e inmigración en España. Un análisis de las trayectorias laborales*. Bellaterra (Barcelona): Universitat Autònoma de Barcelona.

Ministerio de Trabajo y Asuntos Sociales (2006). *La muestra continua de vidas laborales*. Madrid: Ministerio de Trabajo y Asuntos Sociales. Colección Informes y Estudios, Serie Seguridad Social, 24.

Mulvey, J. M. (1983). Multivariated Stratified Sampling by Optimization. Management Science, 29, 6, 715-724.

Padua, J. et al. (1979). *Técnicas de investigación aplicadas a las ciencias sociales*. México: Fondo de Cultura Económica.

Pérez, C. (2005). *Muestreo Estadístico. Conceptos y problemas resueltos*. Madrid: Pearson.

Pérez, C. (2010). *Técnicas de muestreo estadístico*. Madrid : Garceta.

Pfefferman, D.; Rao, C. R. (2009). *Sample Surveys: Design, Methods and Applications*. Amsterdam: Elsevier.

Pulido San Román, A. (1987). *Estadística y Técnicas de Investigación Social*. Madrid: Pirámide.

Rodríguez Monge, A. (1991): Una aproximación a la teoría de muestras. En: *El pluralismo metodológico en la investigación social: ensayos típicos*, editado por M. Latiesa. Granada: Universidad de Granada. Biblioteca de Ciencias Políticas y Sociología. Serie Monografías, 2, 227-256.

Rodríguez Osuna, J. (1989). La muestra: teoría y aplicación. En: *El análisis de la realidad social*. 2a edició. Madrid: Alianza. Alianza Universidad Textos, 105, 287-320.

Rodríguez Osuna, J. (1991). *Métodos de muestreo*. Madrid: Centro de Investigaciones Sociológicas.

Rodríguez Osuna, J. (2005). *Métodos de muestreo: casos prácticos*. Madrid: Centro de Investigaciones Sociológicas.

Sánchez Crespo, J. L. (1971). *Principios elementales del muestreo y estimación de proporciones*. Madrid: INE.

Sánchez Crespo, J. L. (1976). *Muestreo de poblaciones finitas aplicado al diseño de encuestas*. Madrid: Instituto Nacional de Estadística.

Sánchez Carrión, J. J. (1999). *Manual de análisis estadístico de los datos*. Madrid: Alianza. Manuales 055.

Saris, W. E. (2007). *Design, evaluation, and analysis of questionnaires for survey research*. Hoboken, N.J.: Wiley-Interscience.

Scheaffer, R.L.; Mendenhall, W.; Ott, L. (2007). *Elementos de muestreo*. México: Grupo Editorial Iberoamérica.