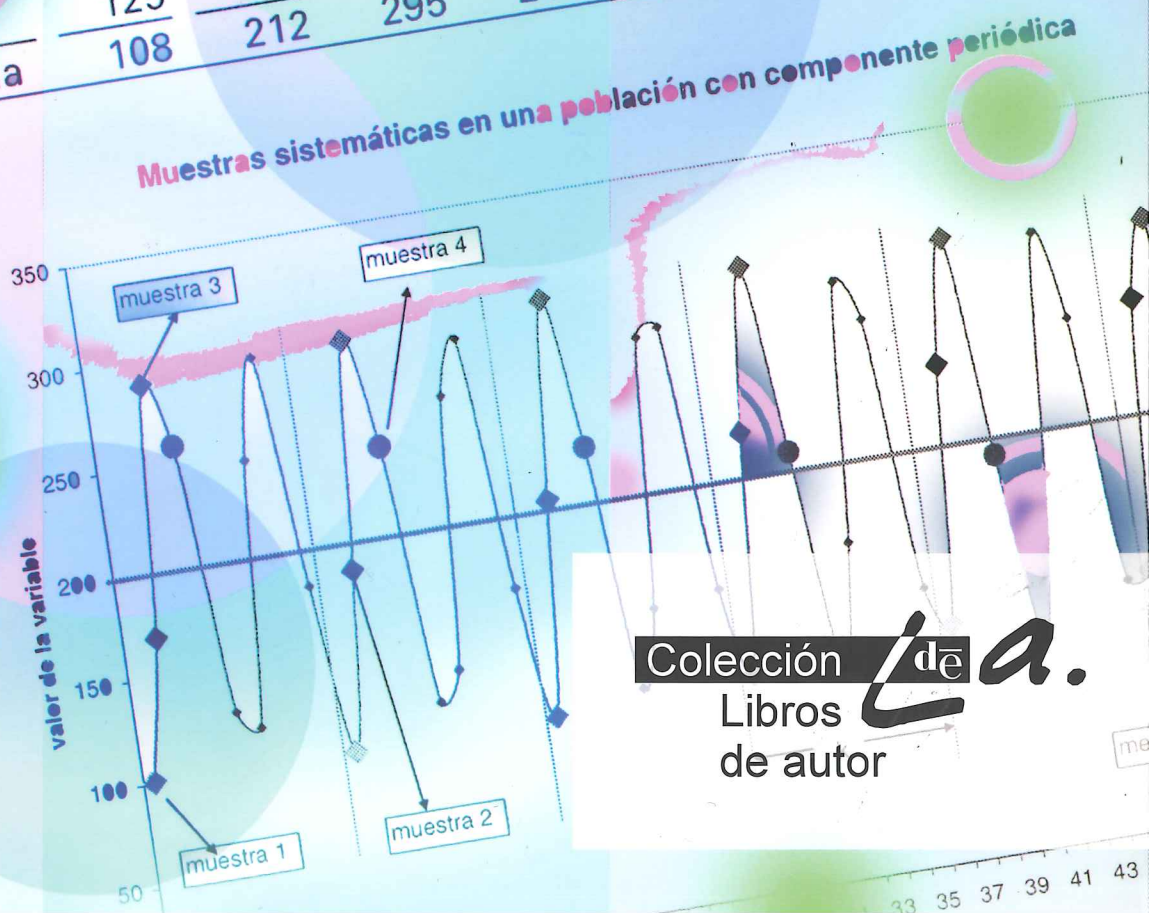
Julio Casado

Manual básico de Estadística

Muestra	1	2	3	4	5			
	101	169	289	258	128	119	246	294
	100	186	295	242	117	131	261	287
	101	204	299	226	108	145	274	277
	106	221	299	208	102	161	285	265
	114	238	297	191	100	178	293	250
	125	254	291	173	100	195	298	234
Media	108	212	295	216	109	155	276	268



Colección de la. Libros de autor

INE

Julio Casado

Manual básico de Estadística

INE

INSTITUTO NACIONAL DE ESTADISTICA

Colección de Libros de autor

Madrid, 2007

Ficha editorial

Título: Manual básico de Estadística

Nº INE: 181

NIPO: 605-07-048-X

Depósito Legal: NA-3441-2007

ISBN: 978-84-260-3741-1

Tarifa: 4

Edita: INE
Paseo de la Castellana, 183 - 28046 Madrid

Impreso en España/*Printed in Spain*
Gráficas Lizarra
Ctra. de Tafalla, km 1 - 31132 Villatuerta (Navarra)

Presentación

El *Manual básico de Estadística* se ha elaborado para que sea un instrumento de gran utilidad para todas aquellas personas que sin tener una formación estadística académica participan en los procesos de producción de la información estadística y tienen interés por conocer los conceptos y técnicas estadísticas básicas que se utilizan en los mismos.

Para el Instituto Nacional de Estadística va a constituir un elemento imprescindible para la formación del personal que se incorpora regularmente a las distintas labores del Instituto tanto en los Servicios Centrales como en las Delegaciones Provinciales. Las personas que anualmente se incorporan al Instituto son muchas, y su formación y las funciones que desempeñan son muy diferentes siendo todas y cada una de ellas de una gran importancia en el proceso de producción estadística.

También creemos que puede ser de gran utilidad para el personal que desarrolla tareas semejantes en los ministerios, comunidades autónomas, empresas públicas o privadas, ..., ya que cada día se incrementa la necesidad de información relativa a todos los aspectos de la realidad económica y social de un país.

El Manual describe de forma resumida, las principales fases que necesariamente hay que abordar en la realización de operaciones estadísticas, para centrarse después en una explicación sencilla de las principales definiciones y conceptos estadísticos, y describir con brevedad las técnicas de muestreo que suelen aplicarse. Su presentación minuciosa y ordenada permitirá al lector entender progresivamente las principales ideas que subyacen tras la terminología estadística.

La producción de información estadística conlleva un largo y complejo proceso en el que intervienen muchas personas que desarrollan distintas funciones para las

que se requiere distinto grado de formación y de especialización. La participación de todas ellas es igualmente importante para la consecución de productos de alta calidad.

Por todo ello, espero y deseo que este Manual sea ampliamente difundido y utilizado no solo por el personal del INE sino también por todas aquellas personas que participan en la producción de información estadística en cualquier ámbito institucional.

Carmen Alcaide Guindo
Presidenta del INE

Índice

Capítulo 1
Operaciones estadísticas

1. Introducción: datos, información y estadística

1. En el mundo actual la disponibilidad de información es absolutamente necesaria para la toma de decisiones. El Gobierno, las empresas, los individuos necesitan información que sirva de ayuda en la solución de problemas. Pero antes de tener información se necesitan datos. *La información es el resultado de obtener, clasificar, procesar y resumir datos.* De la misma forma que el petróleo crudo es la materia prima para obtener gasolina o la madera es materia prima para obtener papel, los datos son la materia prima de la que se obtiene información. *Podemos definir los datos como el conjunto de observaciones o hechos que, una vez recogidos, organizados y procesados, se transforman en información o conocimiento.* Los datos sólo tienen sentido si sirven para aportar información. *La información se define como un conjunto de datos que han sido ya organizados para tener un significado útil para la toma de decisiones.* Es decir, la información consiste en un grupo de datos que tienen una relevancia y un propósito.

2. Tanto los datos como la información no son necesariamente numéricos. La información que proporciona la prensa o la televisión a partir de hechos (datos) que ocurren en una ciudad, en un país o en cualquier parte del mundo, constituye un ejemplo de información no necesariamente numérica. Cuando la información es numérica nos introducimos en el mundo de la Estadística. *La Estadística es la ciencia que se refiere al tratamiento de datos numéricos.* Las estadísticas proporcionan información por medio de números, es decir, son datos numéricos que han sido organizados para tener un significado útil que sirva de ayuda en la toma de

decisiones. La información estadística proporciona datos numéricos que miden variables de interés sobre un conjunto de elementos o *población objetivo*. En estudios económicos y sociales la población objetivo, es decir, la población a la que se refiere la información, suele estar constituida por empresas, establecimientos, hogares e individuos.

3. Todos los días en los medios de comunicación pueden encontrarse noticias basadas en datos estadísticos: desempleo, precios de consumo, encuestas electorales, sondeos de opinión, audiencias de televisión, radio y otros medios, gastos de los hogares, inversión publicitaria, cotizaciones de bolsa, datos macroeconómicos, ... Además de estas estadísticas de trascendencia pública a través de los medios, se realizan otra multitud de estudios estadísticos por las empresas y otras entidades: aceptación de productos por los consumidores, hábitos de compra, grado de satisfacción de los clientes, imágenes de marcas (de partidos políticos), rendimiento de campañas de publicidad, hábitos de desplazamiento de los ciudadanos, ... *Los datos y estudios estadísticos son ya algo habitual en cualquier actividad económica, política o social.*

2. Usuarios de información estadística

1. Si nos preguntamos quién utiliza información estadística nos encontramos que cualquier ente social utiliza en mayor o menor medida datos estadísticos como información de ayuda para la toma de decisiones o, simplemente, para estar informado.

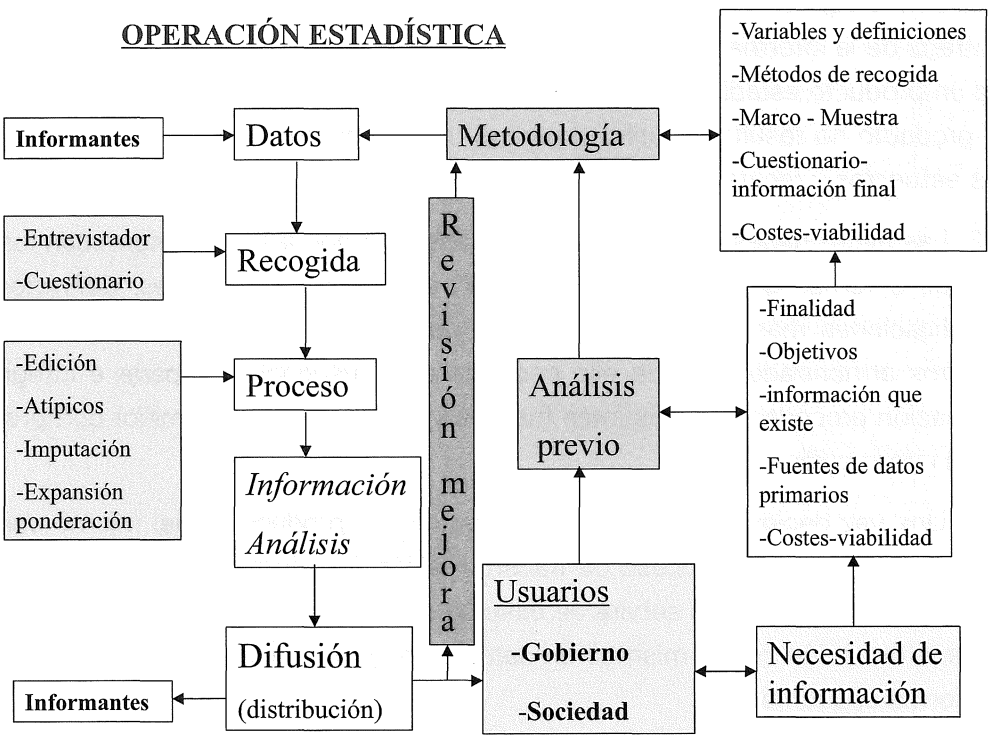
2. *El Gobierno*, tanto nacional como autonómico o local, necesita información sobre la población, la economía y otros asuntos que le permita tomar decisiones sobre localización de servicios, tasas impositivas, ayudas sociales, ... En forma inversa, los datos estadísticos también permiten a los ciudadanos disponer de información sobre la actuación y el desempeño de sus gobernantes.

3. *Las empresas* necesitan información sobre la evolución de sus productos y los de la competencia en el mercado, sobre la economía, la población y las tendencias sociales. Ello les permitirá tomar decisiones sobre políticas de marketing de sus productos, dónde abrir nuevas oficinas y locales, localización de almacenes y fábricas, ...

4. Otros *grupos y organizaciones sociales* necesitan información sobre una amplia variedad de aspectos socioeconómicos relacionados con la salud, nivel educativo, distribución de rentas, tendencias de voto político, empleo marginal, delincuencia, ... sobre las cuales basar sus líneas de acción.

5. Todos *los individuos* en algún momento necesitan información para evaluar la compra o alquiler de una vivienda, analizar la rentabilidad de sus ahorros o por simple curiosidad. Constantemente los medios de comunicación hacen públicos resultados estadísticos que permiten a los ciudadanos estar informados de la evolución de los precios, la economía, el mercado de trabajo, ...

3. La producción estadística

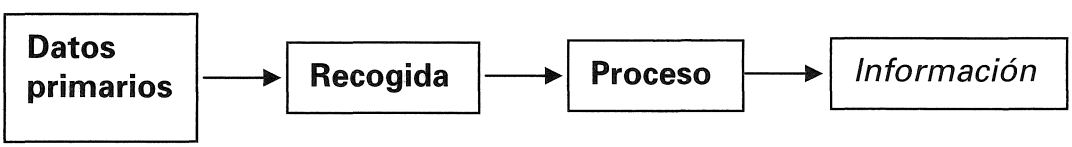


1. Llamaremos *operación estadística* o *estudio estadístico* al proceso por el cuál se obtiene información estadística. Un producto estadístico surge porque existe una necesidad de información, a partir de la cuál comienza el *diseño* del estudio en el que, a partir de las necesidades de los usuarios, se definen los objetivos del mismo, se analiza qué información existe, si hay necesidad de recogida de datos porque no hay información alternativa, qué datos habría que recoger, su viabilidad, posibles fuentes de datos primarios, los costes, hasta llegar a establecer una *metodología* que sirve ya de base para la recogida de datos, su proceso y la difusión de la información final a los usuarios. Debe establecerse también la utilidad del estudio, es decir, qué decisiones podrán tomarse y en qué forma sus resultados van a servir de ayuda en esas decisiones. En operaciones repetitivas las fases de análisis previos y metodología se convierten en una fase de análisis de procesos, de mejora de los mismos y de adaptación del producto a la realidad, siempre cambiante, del entorno socioeconómico.

2. Es importante que los objetivos se establezcan de una forma clara y concreta, incluyendo aspectos presupuestarios, calidad esperada y fechas de realización y de entrega de la información final. La orientación al usuario y a las utilidades prácticas de un producto estadístico es indispensable para acometer el diseño del mismo: si el producto no resulta de utilidad para el usuario se convierte en algo inservible y los esfuerzos y recursos dedicados al mismo habrán sido inútiles.

3. Las necesidades que se pretenden satisfacer son las que determinan los conceptos o temas que se quieren estudiar, así como las variables, definiciones y clasificaciones más adecuadas. Siempre que sea posible deben utilizarse instrumentos armonizados ya que ello permitirá a los usuarios comparar e integrar información procedente de distintas fuentes y les ayudará a una mejor comprensión de la información.

4. Una vez decidida la realización del estudio, la producción del mismo consiste esencialmente en un proceso de transformación de información numérica: *recogida* de la información que servirá de base para el estudio o datos primarios, *proceso* o transformación de los mismos en datos útiles y obtención de *resultados* finales. En forma esquemática:



Como puede verse la obtención de un producto estadístico no difiere conceptualmente de la de cualquier producto industrial, que requiere de un proceso de transformación de materias primas en un producto acabado. La diferencia radica en que tanto los datos primarios como el producto final son información numérica. En los capítulos que siguen se irán exponiendo los principales aspectos de la producción estadística.

5. El número se convierte en el elemento básico de la Estadística, cuya primera finalidad es el de resumir, simplificar y extraer de conjuntos, en general muy grandes, de números (datos primarios), sus características esenciales y transformarlos en un conjunto relativamente reducido de información numérica (datos finales) del que se puedan extraer conclusiones sobre las que basar la toma de decisiones, que es el objetivo final de la información estadística. Como ejemplo de esta síntesis de la información, para la elaboración del índice de precios de consumo el INE toma mensualmente información de más de 160.000 precios que finalmente quedan reducidos a un solo número que los resume todos. Esta labor de simplificación es fundamental para la comprensión de los fenómenos que se estudian. La mente humana no tiene capacidad para asimilar y extraer conclusiones de una lista de 160.000 números.

6. Debe indicarse, por otra parte, que los números estadísticos son el resultado de mediciones sobre las unidades que se desean estudiar y, como cualquier medida, está sujeta a errores. No debe pensarse que porque estemos tratando números, aparentemente exactos, éstos están libres de error. Precisamente una parte importante de la producción estadística debe dedicarse a la detección y depuración/corrección de errores, de forma que se llegue a datos finales que reflejen de forma fiel la realidad del fenómeno que estamos analizando.

Capítulo 2
Formas de tomar datos de una población

1. Censos y muestras

1. Para estudiar una población la primera posibilidad es obtener los datos necesarios de todas y cada una de las unidades que forman la población. Estaríamos entonces ante un estudio censal o *censo*. El censo se caracteriza por obtener información de todos los elementos de la población de interés, mientras que en el *muestreo* sólo se obtienen datos de una parte de la población que representa al conjunto de la misma.

2. La toma de muestras de una población es una práctica habitual y cotidiana en la vida diaria de las personas pese a que ni siquiera nos demos cuenta de ello. Las pruebas de sabor son ejemplos típicos: el cocinero que prueba un guiso está comprobando si los ingredientes están adecuadamente combinados para el gusto de los clientes; el cliente que ve la sopa humeante prueba un poco para apreciar su temperatura; el camarero nos ofrece una pequeña cantidad de la botella de vino que hemos elegido para que decidamos si nos gusta o no. Son ejemplos de tomas de pequeñas muestras para apreciar la calidad/gusto del guiso, sopa o vino, que se hacen sin necesidad de tener que comerse todo el guiso o beberse todo el vino, es decir, sin necesidad de seleccionar toda la población.

3. El objetivo de las técnicas de muestreo es proporcionar una serie de procedimientos que permitan conocer características o valores referidas al total de unidades de un conjunto, estudiando sólo una parte de las unidades del conjunto. Estas características que se desean conocer son las *variables de estudio*. La *población* o *universo* es el conjunto total de unidades de las que se desea información, mien-

tras que la *muestra* es la parte de unidades de la población sobre la que se mide la información, es decir, las variables de estudio. El *tamaño de la población* es el número *N* de unidades que forman la población y el *tamaño de muestra* es el número *n* de unidades seleccionadas para la muestra. También nos referiremos a las unidades de estudio como *unidades estadísticas* y a las variables de estudio como *variables estadísticas.*

4. Cuando se utiliza el muestreo para estudiar una población debe tenerse presente que, dependiendo de qué elementos entren en la muestra se obtendrán unos u otros resultados, es decir, la información sobre la población que se obtenga al seleccionar una muestra va a depender de la muestra seleccionada, lo que da origen al *error de muestreo* que se estudiará más adelante.

2. Ventajas e inconvenientes

1. Un censo presenta dos principales ventajas respecto al muestreo: la primera es que los resultados finales se basan en toda la población y no están, por tanto, sujetos a error de muestreo, y la segunda es el grado de detalle de la información final, la cuál puede proporcionarse para subgrupos y áreas geográficas de pequeño tamaño respecto a la población total. Pero presenta también inconvenientes: el *coste*, esto es, la operación censal puede ser muy costosa para grandes poblaciones; el *tiempo de realización*, que puede ser grande respecto a una operación por muestreo y, finalmente, los *errores* de recogida de datos y proceso que suelen aumentar con el número de elementos de los que se recogen datos.

2. En general hay tres principales ventajas en el muestreo respecto a la investigación total de la población o censo, que son recíprocos a los inconvenientes del censo:

– *Menor coste*, derivado de obtener información de una parte de la población.

– *Mayor rapidez* en la obtención de resultados por el mismo motivo anterior.

– *Mayor calidad.* Al reducirse el volumen de trabajo se puede emplear personal especialista, mejor preparado y entrenado. Igualmente los procesos de supervisión y proceso de datos están mejor controlados, lo que redunda en una mejor calidad de trabajo y una disminución de errores (no de muestreo) respecto al censo.

Los inconvenientes del muestreo se corresponden con las ventajas del censo: la información basada en una muestra está sujeta a error de muestreo y el grado de detalle de la información final está siempre limitado por el tamaño de muestra que la soporta por lo que no es posible llegar a los niveles de desglose de un censo.

3. También debe tenerse en cuenta que la realización de un censo resulta muchas veces imposible. El cocinero no puede comerse todo el guiso, dejaría a los clientes sin comer. En general siempre que obtener información implica la destrucción o consumo de la unidad de estudio, como son pruebas de resistencia o duración de materiales, no resulta posible (no es rentable) la realización de un censo. Otro ejemplo sobre la imposibilidad monetaria de un censo lo obtenemos de las quinielas de fútbol y loterías: cuando una persona gasta 7 euros en la lotería primitiva está seleccionando una muestra de 7 resultados de las 13.983.816 formas posibles de obtener los 6 aciertos (una apuesta = un euro); si quiere tener la seguridad absoluta de acertar tendría que apostar a toda la población de resultados posibles y gastarse la bonita cifra de casi catorce millones de euros, lo cuál sería bastante ruinoso.

4. Otro aspecto a considerar se relaciona con los informantes, es decir, las unidades que tendrán que aportar sus datos: el grado de molestia a las unidades informantes es menor en una muestra ya que sólo hay que dirigirse a las que se incluyen en la muestra y no a toda la población.

3. Los registros administrativos

1. Una tercera alternativa al censo y al muestreo como formas de plantear la recogida de datos, es la utilización de registros administrativos. Un *registro administrativo* es el resultado de las operaciones habituales de una organización: una empresa lleva un registro de cada uno de sus empleados con el nombre, sexo, edad, fecha de ingreso en la empresa, salario, ... Otros ejemplos típicos son registros de nacimientos, muertes, matrimonios, matriculación de vehículos, directorios telefónicos, registro de sociedades, registros tributarios,... La utilización de registros administrativos como fuente de recogida de datos primarios para una operación estadística resulta bastante atractiva pero no deja de tener algunos inconvenientes.

2. Las principales ventajas de utilizar registros administrativos son:

– *Error:* la información se obtiene de todos los elementos que componen el registro por lo que no hay error de muestreo, aunque podría considerarse la utilización de una muestra de sus elementos.

– *Simplicidad-coste:* la utilización de datos administrativos elimina la necesidad de diseñar un censo o una muestra y los costes asociados. Los costes de recogida de información son menores.

– *Molestias:* no hay molestias para los informantes ya que los datos que se precisan ya han sido facilitados. Esto es fundamental, ya que las necesidades de información son cada vez mayores y su primera consecuencia es el aumento de la carga de respuesta. Siempre que sea posible hay que utilizar registros administrativos como fuente de datos primarios.

– *Evolución temporal:* los registros se actualizan permanentemente y la recogida de información en forma periódica permite la realización de análisis de tendencia.

3. Algunos inconvenientes de los registros administrativos son:

– *Flexibilidad:* los datos primarios a recoger están limitados por la información administrativa que contenga el registro. Este inconveniente se vería reducido si siempre que se establezca un registro se pensara no sólo en su uso administrativo, sino también en su explotación estadística.

– *Cobertura:* la información que se obtenga quedará referida a la población contenida en el registro.

Capítulo 3
Métodos de recogida de datos

1. La recogida de datos

1. La *recogida de datos* es el proceso encaminado a conseguir que el informante facilite los datos primarios requeridos. Se incluye aquí la solicitud y obtención de los datos, la comprobación preliminar de su coherencia y completitud y el seguimiento y control del desarrollo de la operación de recogida, también llamada *operación de campo*.

2. La repercusión de las operaciones de recogida de datos en la calidad de la información final es fundamental, ya que los datos que se recogen constituyen la información primaria que posteriormente sirve para la elaboración de la información estadística. Por tanto, la calidad de los datos recogidos determina de una forma esencial la del producto final: sin datos primarios de calidad no hay información final de calidad.

3. Uno de los aspectos más importantes para conseguir el éxito de una operación estadística es el nivel de respuesta obtenido y la calidad de la información suministrada. En este contexto, los informantes constituyen el recurso más valioso para las instituciones que producen información estadística y por tanto se les deberá dar todo tipo de facilidades para que proporcionen los datos solicitados reduciendo al mínimo la carga que este trabajo supone.

4. En la recogida de datos debe disponerse de un sistema de control con el objetivo de dar información exacta del estado en que se encuentran los trabajos de recogida, de las incidencias que se estén produciendo y sus causas, de la cobertura de muestra que se alcanza y del cumplimiento del calendario, con información a

tiempo real para poder tomar las medidas adecuadas y resolver los problemas que se van presentando.

5. En lo posible, el seguimiento y control de los trabajos de campo debe hacerse a distintos niveles de desagregación geográfica y de tipos de unidades y es fundamental también hacerlo a nivel de técnico de encuesta, inspector y entrevistador.

2. Métodos de recogida

1. *Entrevista personal* (cara a cara): la recogida de datos se realiza por medio de entrevistadores bien entrenados que visitan a las unidades informantes y recogen los datos por medio de un cuestionario. Si los entrevistadores reciben el entrenamiento adecuado, éste es un buen método para asegurar mejores tasas de respuesta y una buena calidad en los datos recogidos. Sin embargo es un procedimiento costoso debido a los costes asociados al entrenamiento de los entrevistadores y a los viajes necesarios para localizar y entrevistar a los informantes.

2. *Entrevista personal con ayuda de ordenador (CAPI):* en lugar de utilizar un cuestionario sobre papel como antes, el cuestionario está integrado en una aplicación informática y el entrevistador realiza la entrevista ayudándose de un ordenador portátil, que marca el flujo de preguntas. Ahorra la posterior grabación de datos y permite que la aplicación informática incorpore una serie de controles de validación sobre las respuestas que mejoran la consistencia y calidad de los datos primarios.

3. *Entrevista telefónica:* la entrevista para la recogida de datos se realiza por teléfono. Es un método más rápido y menos costoso que la entrevista personal, pero por contra, es necesario que el entrevistado disponga de teléfono. La entrevista no debe ser excesivamente larga para evitar que el informante se canse y cuelgue el teléfono prematuramente.

4. *Entrevista telefónica con ayuda de ordenador (CATI):* es una entrevista telefónica como la anterior, pero en la que el entrevistador graba directamente las contestaciones en un ordenador. Ahorra la posterior grabación de datos, aunque puede ser costosa de implantar por la infraestructura informática que necesita. Los entrevistadores necesitan tener habilidad para la grabación. Como en el sistema CAPI, algunos controles de validación son simultáneos a la entrevista.

5. *Correo:* el cuestionario se envía por correo a los informantes, que los devuelven, una vez completados, también por correo. Es un método poco costoso de recogida de información y permite distribuir grandes cantidades de cuestionarios en poco tiempo. Da también la oportunidad de llegar a unidades con las que resulta muy difícil contactar. Para el informante es más cómodo: no tiene que contestar al momento y puede decidir cuando rellenar el cuestionario dentro del plazo establecido. La recogida de datos por correo requiere el uso de listas actualizadas de direcciones. Su mayor desventaja es que las tasas de respuesta son bastante menores que las proporcionadas por otros métodos y ello puede tener consecuencias en la calidad de los datos y en la fiabilidad de los resultados. Las personas con bajo nivel cultural pueden tener problemas para entender y rellenar el cuestionario.

6. *Entrega en mano:* los cuestionarios se entregan en mano a los informantes y, posteriormente, se vuelve a recogerlos. Normalmente se obtienen mejores tasas de respuesta que con el envío por correo, pero resulta bastante más costoso. Es un método bastante adecuado cuando se necesita información de varios o todos los miembros de un hogar. Como variante del método, se pueden entregar los cuestionarios en mano y los informantes los devuelven por correo una vez completados, lo que contribuye a reducir costes y aporta una mayor sensación de intimidad a los informantes.

7. En los dos últimos métodos, por correo y entrega en mano, es el propio informante quien rellena el cuestionario. Ello requiere un cuidadoso diseño del mismo, bien estructurado y con instrucciones claras para que sea sencillo y cómodo de rellenar. Pese a todo, es recomendable incluir en el cuestionario el nombre de una persona de contacto y un número de teléfono, de llamada gratuita, para resolver cualquier duda que se le pueda presentar al informante.

8. *Otros métodos* de recogida: entre otros métodos de recogida se puede citar la *observación directa*, que se utiliza con frecuencia en estudios de precios, con registros administrativos y cuando el informante no tiene que contestar a ningún cuestionario, sino simplemente aportar la documentación o ficheros a partir de los cuales es el propio entrevistador el que rellena el cuestionario o aporta los datos al ordenador. La elección de uno u otro método depende de distintos factores: complejidad y longitud del cuestionario, sensibilidad de los datos que se piden, costes de recogida, dispersión geográfica de los informantes y listas o marcos disponibles. Con frecuencia la mejor estrategia es una combinación de distintos métodos.

Por ejemplo, la recogida por correo es bastante más eficiente cuando se combina con llamadas telefónicas de seguimiento, e incluso visitas personales en los casos más reacios a responder.

Capítulo 4
El cuestionario y los entrevistadores

1. El cuestionario: concepto

1. El cuestionario es el conjunto de preguntas y cuestiones diseñadas para recoger los datos de los informantes. Un cuestionario puede ser directamente cumplimentado por el informante o administrado por un entrevistador. El cuestionario juega un papel fundamental en la recogida y calidad de los datos, por lo que su imagen y redacción de las preguntas influye en la obtención de datos válidos y fiables, en el comportamiento de los informantes y en el trabajo de los entrevistadores.

2. Un cuestionario bien diseñado facilita la respuesta y permite una recogida eficiente de datos con un mínimo de errores. Su diseño debe tener en cuenta los objetivos del estudio, el método de recogida de datos, el plan de proceso de datos y las características de la población a entrevistar. Debemos pensar también que la estructura y la apariencia o aspecto del cuestionario contribuyen a la imagen que se hará el informante del INE.

2. Aspectos a tener en cuenta en el diseño del cuestionario

1. El cuestionario debe ser atractivo y resultar fácil de cumplimentar, tanto para el informante como para el entrevistador, con el número de preguntas estrictamente necesario. El cuestionario debe tener un diseño atractivo que anime a los informantes a facilitar la información y contribuya a evitar el rechazo, es decir, debe estar totalmente orientado al informante y centrarse en los datos de interés para el

estudio, sin preguntas superfluas, siguiendo un orden lógico en las preguntas que facilite la respuesta. En lo posible, deben evitarse preguntas ya hechas en otro estudio y tampoco deben incluirse preguntas que no vayan a ser posteriormente objeto de explotación.

2. Utilizar palabras y conceptos claros y perfectamente entendibles por los informantes, cuidando esmeradamente la redacción de las preguntas y evitando que las mismas puedan tener una cierta respuesta inducida. Tener presente que la misma pregunta redactada de forma diferente puede dar resultados diferentes.

3. En la introducción/portada del cuestionario debe darse el nombre del estudio, explicar sus objetivos, razonar la importancia de colaborar y rellenar el cuestionario, informar de la confidencialidad de los datos y explicar la utilización de la información final. Todo ello en una forma concisa y entendible por los informantes.

4. Incluir las instrucciones necesarias en el propio cuestionario, fáciles de entender y de encontrar. Separar adecuadamente (título o nombre) las secciones del cuestionario y las preguntas; utilizar colores, sombreados, distintos tipos de letra, ilustraciones y símbolos que atraigan la atención, den mayor claridad y atractivo al cuestionario y guíen a los informantes o entrevistadores a través del mismo, facilitando su cumplimentación. Al final del cuestionario dejar un espacio para comentarios adicionales que quieran hacer los informantes y expresar la gratitud por su colaboración.

5. Deben hacerse pruebas reales a pequeña escala para evaluar el cuestionario. Estas pruebas servirán para comprobar el funcionamiento real del cuestionario ante los informantes, evaluar la redacción de las preguntas, su secuencia, si son o no entendibles, comprobar categorías de respuesta, longitud de la entrevista, problemas de negativa, ... Estas pruebas son importantes para detectar y corregir problemas del cuestionario. Debe pensarse siempre que un cuestionario mal diseñado puede invalidar totalmente los resultados de un estudio.

3. Los entrevistadores

1. El papel del entrevistador, como persona que recoge los datos de los informantes, es muy importante. El proceso de obtener datos a través de entrevista requiere ciertas cualidades y habilidades, sin las cuales la eficiencia y calidad de los datos puede reducirse sensiblemente. Algunas de ellas son:

– Utilización de coche y teléfono. Disponibilidad para viajar.

– Buena capacidad de comunicación y relación con personas.

– Un aspecto profesional que de confianza al entrevistado.

– Disponibilidad para trabajar tardes y fines de semana. Las entrevistas a hogares realizadas en horario normal de trabajo presentan altas tasas de ausencias. Es necesario realizarlas en horas en que el personal del hogar se encuentra en casa.

2. Los entrevistadores deben ser adecuadamente entrenados para la recogida de datos. La forma y el estilo en que el entrevistador se dirige a los informantes tiene una gran influencia en su reacción hacia la entrevista, en su disposición a colaborar en el estudio y en la imagen del INE que perciba el entrevistado. Por ello, además de cuidar su aspecto personal, los entrevistadores deben preocuparse de hacer una introducción adecuada antes de empezar las preguntas. En esta introducción se debe:

– Dar el nombre y la tarjeta de identificación, dejando tiempo para que el informante la compruebe.

– Explicar el estudio que se está realizando y por quién.

– Explicar los objetivos del estudio, por qué ha sido seleccionado para la muestra y la importancia de su colaboración.

– Informar de la confidencialidad de los datos aportados.

– La información contenida en la portada del cuestionario (ver 4.2.3) puede servir de guión al entrevistador en esta introducción.

Es importante además, que el entrevistador esté familiarizado con la técnica correcta de entrevista, que incluye:

– Habilidad para escuchar con atención.

– Mantener la entrevista en el mínimo de tiempo.

– Estimular el interés del informante.

– Hacer las preguntas ***tal y como están*** redactadas en el cuestionario.

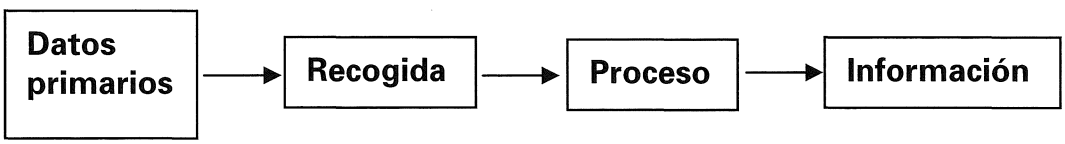
– NO SUGERIR ninguna respuesta al informante.

– Contestar adecuadamente a las preguntas del informante.

Capítulo 5
Proceso de datos

1. Concepto

1. Los datos primarios, los datos recogidos, son simplemente observaciones o hechos. Cuando los datos se organizan y presentan adecuadamente se hacen información. El *proceso de datos* consiste en transformar datos en información. En forma simplificada:



2. Desde hace años el proceso de datos se hace de forma rápida y fácil gracias al uso de ordenadores. El proceso de datos comprende los siguientes pasos:

– Codificación de datos.

– Grabación/entrada de datos.

– Edición/imputación de datos.

– Producción de información.

2. Codificación

1. Antes de que los datos primarios sean introducidos en el ordenador deben codificarse. La codificación significa etiquetar las respuestas de forma abreviada y única para cada respuesta (a menudo con simples códigos numéricos). La razón de codificar es que hace mucho más fácil la grabación y posterior manipulación de los datos. La codificación puede ser hecha por los propios entrevistadores o por personal de oficina.

2. Los cuestionarios suelen estar precodificados ya que la mayor parte de las preguntas son cerradas. Una *pregunta cerrada* implica que sólo se permiten un número predeterminado de respuestas posibles, las cuales, tienen ya fijado un código en el cuestionario. Una *pregunta abierta* significa que se permite cualquier respuesta y el entrevistador debe anotar literalmente la respuesta del informante. La codificación de estas preguntas es más difícil y suele hacerse seleccionando una muestra de respuestas y diseñando una estructura de códigos que captura y categoriza la mayoría de respuestas.

3. Grabación

1. El teclado de un ordenador es una de las herramientas actuales más conocidas para la introducción de datos en ordenador, al igual que hace años lo eran las tarjetas perforadas. Otros instrumentos que se utilizan para la entrada de datos son los lectores de código de barras, escáner, aparatos de lectura óptica y lápices de pantalla.

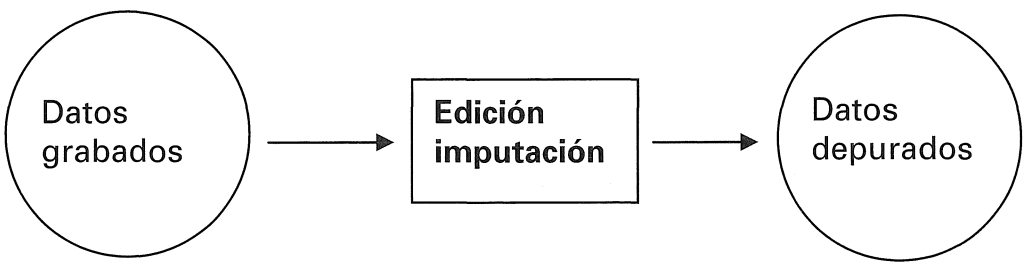
2. Mientras que hace años la grabación de datos solía hacerse de forma centralizada por personal especializado, el uso extensivo del ordenador ha hecho que la grabación de datos se realice de forma muy cercana a la recogida de los mismos.

3. Con los métodos de entrevista con ayuda de ordenador, bien sea entrevista personal (CAPI) o telefónica (CATI) son los propios entrevistadores los que realizan directamente la entrada de datos.

4. Edición (depuración) e imputación

1. Una vez grabados los datos debe realizarse un detallado chequeo de los mismos para detectar y corregir datos que faltan, datos inválidos o inconsistentes o

datos erróneos o potencialmente erróneos. De esta forma se eliminan problemas en los datos primarios que podrían originar información incorrecta. Este proceso se conoce como edición e imputación de datos. Su objetivo es obtener un fichero de datos primarios depurado y *limpio* que sirva de base para las tablas e información final. En forma esquemática:



2. Algunos de los chequeos que se realizan en este proceso son:

– *Validación:* asegura que los datos caen dentro de ciertos límites. Por ejemplo, que caracteres alfabéticos no aparezcan en un campo numérico, o que el mes del año no sea mayor de 12.

– *Consistencia:* chequea la consistencia lógica de los datos. Por ejemplo, una persona que ha contestado en una pregunta *nunca he estado casado* no puede tener como estado civil *divorciado*.

– *Atípicos (outliers):* un dato atípico es un valor extremo, es decir, un valor que se aparta significativamente de los demás.

3. La *imputación* es el proceso de resolver los problemas de datos faltantes, inconsistentes, no válidos o atípicos detectados en la edición de datos. Se realiza cambiando los datos originales por otros que aseguren un registro verosímil y coherente. Aunque se hayan ya resuelto problemas en los datos en la etapa de recogida mediante nuevos contactos con los informantes o inspección visual del cuestionario, siempre quedarán problemas en los datos que es necesario resolver mediante métodos de edición e imputación que produzcan un fichero de datos de entrada completo y consistente, es decir, depurado.

4. Ambas fases, depuración e imputación, deben proporcionar valiosa información sobre la calidad de la recogida de datos y para la mejora de futuras operaciones. Deben considerarse como parte de la recogida de datos y utilizarse para evaluar y obtener indicadores de calidad de los datos primarios. Es importante tam-

bién tener en cuenta que su única finalidad es la completitud y consistencia de datos primarios recogidos y no la modificación casi total de los mismos. La depuración e imputación no mejora la calidad de los datos primarios, sólo mejora su coherencia.

5. La contribución de la depuración/imputación a la reducción de errores de campo es limitada. Para reducir errores es mejor prestar atención a la fase de recogida de datos, en lugar de centrarse en tareas de limpieza al final de la misma. Mejor prevenir que curar.

5. Estimación - expansión

1. Con este proceso los datos primarios depurados se convierten en información. Para ello, a cada dato primario se le aplica el factor de ponderación/expansión necesario para referir los datos a la población objeto de estudio, obteniéndose los datos ponderados o expandidos. La posterior agregación de los datos ponderados para las distintas poblaciones/subpoblaciones sobre las que se deba facilitar información proporciona las tablas finales para su publicación en los distintos medios de difusión (papel, soporte magnético, internet u otros).

2. Este proceso de transformación de datos primarios en información se sintetiza en la fórmula

$$\text{Información} = \sum \text{dato primario} \bullet w$$

donde w es la ponderación o factor de expansión del dato primario depurado correspondiente a cada unidad elemental de estudio; w es el factor por el que se multiplica cada dato primario para hacer que la referencia del mismo sea la población objetivo. El símbolo sumatorio indica la agregación de los datos ponderados a las distintas poblaciones sobre las que deba referirse la información o producto final.

3. El proceso productivo resumido en la fórmula anterior es independiente del tipo de unidad informante y de que el estudio sea censal o muestral. Si el estudio es censal la ponderación *w* será la unidad, si elaboramos cualquier tipo de número índice el dato primario deberá referirse al conjunto poblacional aplicándole la ponderación adecuada, si el estudio es muestral *w* será el factor de expansión necesario para referir la información a la población total, si hablamos de muestras de empresas o de población la forma de obtener resultados finales es la misma, aunque los procedimientos muestrales y de recogida de información puedan diferir.

Capítulo 6
Trabajar con porcentajes

1. Porcentajes y números índices simples

1. La utilización de porcentajes es muy común al manejar datos estadísticos por lo que conviene repasar este concepto. Veamos algunos datos sobre la matriculación de turismos en España:

	Total	Importados	% importación
Matriculación de turismos			
1996	968.363	566.970	58,5%
1997	1.091.190	661.078	60,6%
1998	1.282.970	821.928	64,1%
1999	1.502.531	994.102	66,2%
2000	1.457.494	956.360	65,6%
Tasas de variación interanuales			
1997	12,7%	16,6%	
1998	17,6%	24,3%	
1999	17,1%	20,9%	
2000	-3,0%	-3,8%	
Índices de matriculación: 1996=100			
1996	100,0	100,0	
1997	112,7	116,6	
1998	132,5	145,0	
1999	155,2	175,3	
2000	150,5	168,7	

2. Podemos ver que en 1999 se han matriculado 219.561 turismos más (diferencia absoluta) que en el año anterior, lo que supone un incremento relativo de

$$\frac{1.502.531 - 1.282.970}{1.282.970} \times 100 = \frac{219.561}{1.282.970} = 17{,}1\%$$

es decir, estamos expresando que por cada 100 turismos matriculados en 1998, se han matriculado 117 en 1999. La cantidad por la que se divide en el denominador es la *base* del porcentaje. En caso de no multiplicar por 100 se obtiene el tanto por uno o variación relativa por unidad. Al mismo resultado se llega dividiendo directamente las matriculaciones de ambos años

$$\frac{1.502.531}{1.282.970} \times 100 = 117{,}1$$

y restando posteriormente 100. En este caso en que se esta comparando el cambio del valor de una variable (matriculaciones) entre dos años consecutivos, el porcentaje anual de cambio se llama también *tasa de variación interanual.*

3. Al ser cocientes entre dos cantidades los porcentajes son *directamente comparables*. Así, podemos ver que la matriculación de turismos de importación en 1999 se ha incrementado en 72.114 unidades, una cantidad lógicamente inferior a las 219.561 unidades de incremento total, pero en términos relativos el crecimiento de los importados es de

$$\frac{994.102 - 821.928}{821.928} = \frac{72.114}{821.928} = 20{,}9\%$$

que expresa un crecimiento porcentual claramente superior, casi cuatro puntos, al 17,1% de incremento en el total de matriculaciones.

4. Para el año 2000 se registran descensos de 45.037 unidades en el total de matriculaciones y de 37.742 en los de importación, que se traducen en porcentajes de

$$\frac{1.457.494 - 1.502.531}{1.502.531} \times 100 = -3{,}0\% \quad \text{y}$$

$$\frac{956.360 - 994.102}{994.102} \times 100 = -3{,}8\%$$

respectivamente. En otras palabras, por cada 100 turismos matriculados en 1999 se han matriculado 97 en 2000

$$\frac{1.457.494}{1.502.531} \times 100 = 97$$

5. La forma de cálculo de porcentajes que hemos visto corresponde a cambios que se producen en una misma variable: si una cierta variable pasa de tener el valor B al valor A, el cambio absoluto es A-B y el cambio porcentual es

$$\frac{A - B}{B} \times 100$$

en donde B es la *base* del porcentaje. En otras ocasiones el porcentaje refleja la *proporción* que supone una cantidad respecto a un total del que forma parte. Así en la tabla de matriculación de turismos podemos ver que en el año 1996 los turismos importados son 566.970 unidades que, respecto al total de 968.363 unidades matriculadas, suponen una proporción de

$$\frac{586.970}{968.363} \times 100 = 58,5\%$$

es decir, por cada 100 turismos matriculados en 1996, casi 59 fueron de importación. Suele expresarse diciendo que la participación de turismos de importación en el total matriculados era del 58,5% en el año 1996. En forma similar se han calculado el resto de datos de la columna *%importación* del cuadro del párrafo 1.

6. En la tabla de matriculación de turismos se observa que la participación de los importados en el total se incrementa todos los años con excepción de 2000 que disminuye ligeramente respecto a 1999. Esto es consecuencia de las mayores tasas de variación interanuales que registran los turismos de importación. Es decir, los turismos de importación han incrementado sus matriculaciones anuales a un ritmo mayor que el total y, como consecuencia, su participación ha pasado del 58,5% en 1996 al 65,6% en 2000. Justamente en 2000 la participación de los importados sufre un ligero retroceso respecto a 1999 (de 66,2 a 65,6) como consecuencia de experimentar un mayor descenso en las matriculaciones (-3,8% frente al –3,0% del total).

7. Los porcentajes están íntimamente ligados con los números índices. Un *número índice* es el cociente entre dos valores de una variable multiplicado por 100. En la tabla de matriculaciones figuran los turismos matriculados en cada año to-

mando como base 100 el año 1996, es decir, estamos dividiendo las matriculaciones de cada año por las correspondientes de 1996. Así, el índice de matriculaciones del total turismos en 2000 respecto a 1996 es

$$\frac{1.457.494}{968.363} \times 100 = 150,5$$

En forma similar se calculan el resto de los valores índices de la tabla. Las tasas de variación interanuales que hemos visto anteriormente pueden calcularse igualmente a partir de los respectivos índices de matriculación. Así, la variación porcentual del total de turismos matriculados en 1999 con respecto a 1998 es

$$\frac{155,2 - 132,5}{132,5} \times 100 = 17,1\%$$

8. Inversamente las tasas de variación interanuales pueden calcularse en forma de *índices de variación interanuales* dividiendo las matriculaciones de cada año entre las del año anterior y multiplicando por 100. Los datos que obtenemos son

Índices de variación interanual

	Total	Importados
1997	112,7	116,6
1998	117,6	124,3
1999	117,1	120,9
2000	97,0	96,2

que se corresponden con las tasas de variación interanuales vistas anteriormente en la tabla 1. Vemos pues, que los porcentajes son una forma de expresar los índices de cambio entre dos valores de una variable.

9. Al manejar porcentajes hay que tener siempre cierto cuidado. Debemos ser conscientes de que un porcentaje es, en el fondo, un cociente y hay que tener claro cuál es la base del porcentaje. Un error que se comete con cierta frecuencia es el de sumar porcentajes. Así, a partir de las tasas de variación interanuales del total de turismos matriculados, podría pensarse que la variación acumulada de las

matriculaciones del año 2000 respecto a 1996 es la suma de los porcentajes de cambio entre años consecutivos, es decir

$$12,7 + 17,6 + 17,1 - 3,0 = 44,4\%$$

Fácilmente se comprueba que el porcentaje correcto de variación entre ambos años es

$$\frac{1.457.494}{968.363} = 1,505 \rightarrow 50,5\%$$

del 50,5%. Al resultado correcto se llega también si se transforman los porcentajes de variación interanuales a índices de variación interanuales (en tanto por uno) y se multiplican, es decir

$$1,127 \times 1,176 \times 1,171 \times 0,973 = 1,505 \rightarrow 50,5\%$$

ya que esta operación es equivalente a

$$\frac{1.091.190}{968.363} \times \frac{1.282.970}{1.091.190} \times \frac{1.502.531}{1.282.970} \times \frac{1.457.494}{1.502.531} = \frac{1.457.494}{968.363} = 1,505 \rightarrow 50,5\%$$

La suma citada de porcentajes es una aproximación al verdadero valor sólo si los porcentajes que se suman están próximos a la unidad (±5%) y no se suman más allá de tres o cuatro porcentajes.

10. Se ha citado en el párrafo 6 el hecho de que los mayores crecimientos interanuales de los turismos de importación se reflejan en un incremento en su participación en el total. Por ejemplo en 1998 la matriculación de turismos de importación se incrementa en un 24,3% frente al 17,6% del total y su participación en el total de matriculaciones pasa del 60,6% en 1997 al 64,1% en 1998. La relación entre estos porcentajes es

$$60,6 \times \frac{1,243}{1,176} = 64,1$$

como fácilmente puede comprobarse al trabajar con las cifras absolutas

$$\frac{661.078}{1.091.190} \times \frac{821.928}{661.078} \times \frac{1.091.190}{1.282.970} = \frac{821.928}{1.282.970} = 0,641 \rightarrow 64,1\%$$

La relación 1,243/1,176 es el cociente entre los índices de crecimiento respecto al año anterior de los turismos de importación y del total y está expresando el mayor o menor crecimiento relativo de la importación respecto al total, siendo así el factor por el que hay que multiplicar la participación para obtener su nuevo valor.

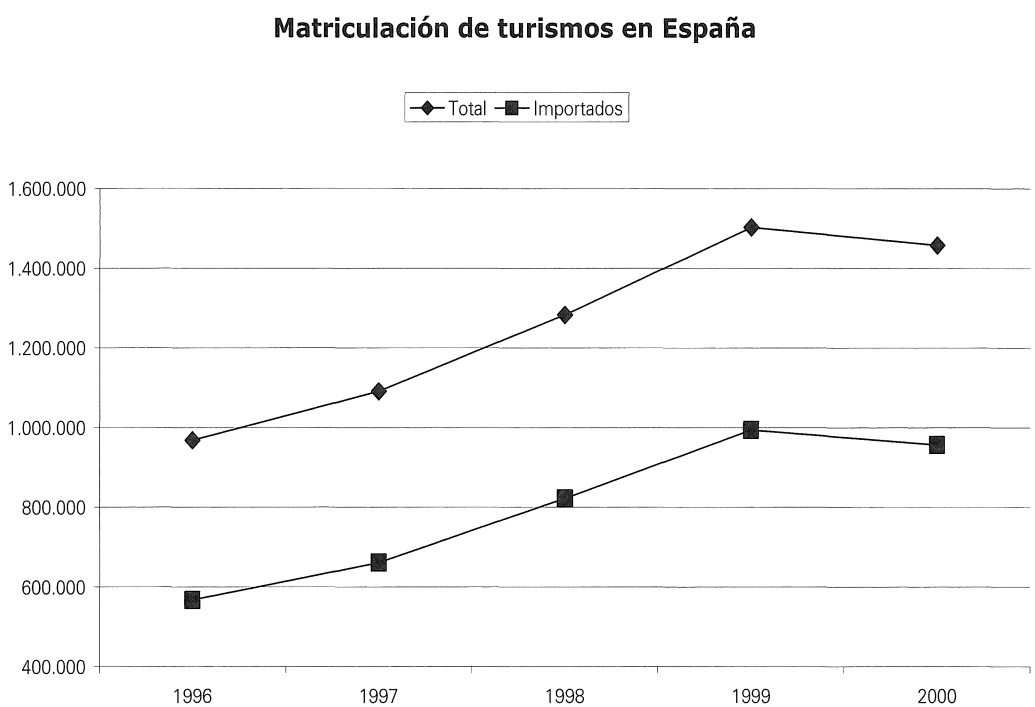
11. Supongamos que en 1998 el valor monetario de los turismos matriculados hubiera crecido un 30% respecto al año anterior, frente a un crecimiento en cantidad o número de unidades de un 17,6%. Sería un grave error deducir de estos datos que el crecimiento en precios ha sido de $\frac{30}{17,6} = 70,5\%$. Esto se puede comprobar fácilmente teniendo en cuenta que Valor = Precio*Cantidad, es decir, $V = PQ$. La relación entre los valores monetarios de ambos años es $\frac{V_2}{V_1} = \frac{P_2Q_2}{P_1Q_1} = \frac{P_2}{P_1}\frac{Q_2}{Q_1}$ y se deduce de aquí que la relación entre los precios de ambos años es $\frac{P_2}{P_1} = \frac{V_2/V_1}{Q_2/Q_1}$, por lo que la operación correcta para calcular el crecimiento de precios es $\frac{1,30}{1,176} = 10,5\%$.

12. El ejemplo del párrafo anterior muestra también que hay que ser cuidadosos en la interpretación de datos. El resultado muestra que aparentemente los turismos han experimentado un crecimiento en sus precios de un 10,5% en 1998 respecto al año anterior. Sin embargo, lo que refleja realmente el número, es que el coste unitario para el consumidor de un turismo matriculado en 1988 ha sufrido un incremento de un 10,5% respecto a 1997, y esta cifra incluye también un posible desplazamiento de los gustos de los compradores de coches hacia turismos de gama alta en detrimento de los de menor coste. Y este posible desplazamiento es bastante real si nos fijamos en la evolución tan positiva que se ha visto en los turismos de importación y le añadimos el que este tipo de vehículos son en su mayoría de gama alta.

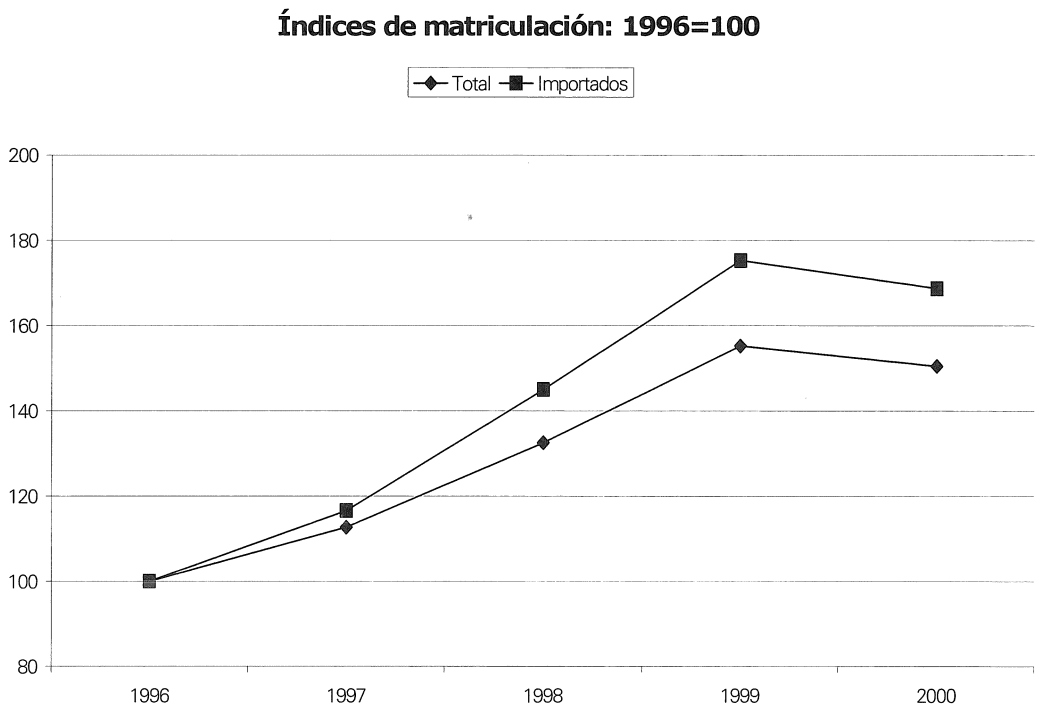
2. Primeros gráficos

1. La representación gráfica juega un importante papel en la descripción y comprensión de datos estadísticos por lo que interesa empezar a familiarizarse con los distintos tipos de gráficos que suelen utilizarse. De hecho, la información estadísti-

ca que publican los medios de comunicación suele ir acompañada de gráficos que resumen los aspectos más destacados de los datos o los aspectos que al medio en cuestión le interesa más destacar. A título introductorio se proporcionan algunos gráficos sencillos que presentan distintos rasgos de los datos de matriculaciones del apartado anterior.



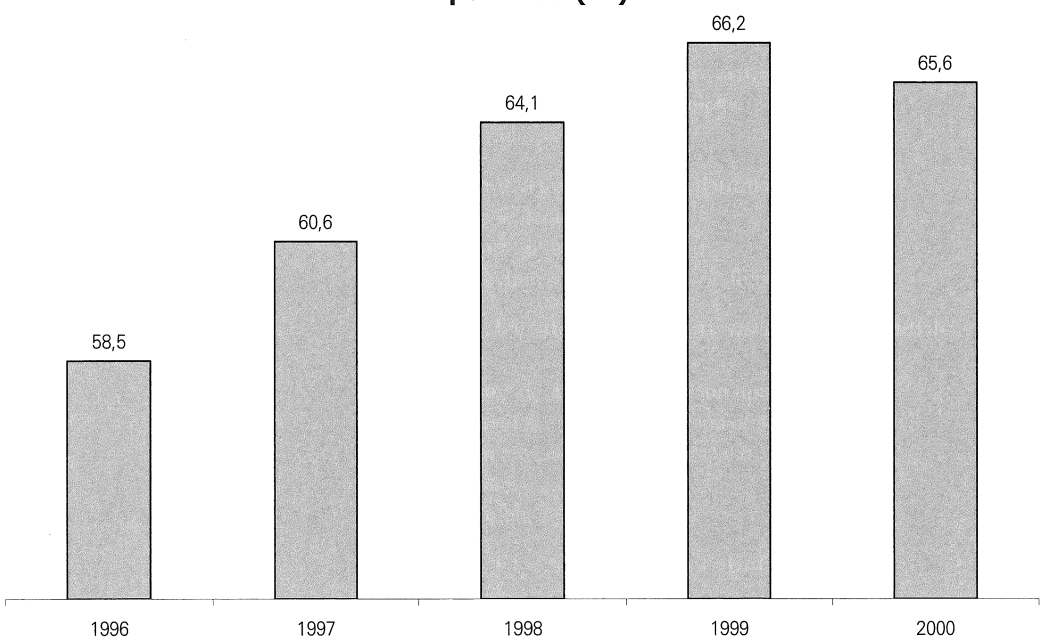
2. Este primer gráfico muestra la evolución de las matriculaciones tanto en el total como en turismos de importación en términos absolutos. La escala vertical (ordenadas) representa el número de turismos matriculados y en la horizontal se muestran los años considerados. El gráfico refleja la evolución positiva de las mismas y el frenazo que sufren en el año 2000. Al ser la escala vertical de valores absolutos, el gráfico refleja la cantidad de matriculaciones en cada año y las diferencias absolutas de un año a otro. Los superiores crecimientos relativos habidos en la matriculación de turismos de importación, se reflejan mejor en los dos gráficos siguientes que muestran los índices de matriculaciones y las variaciones anuales respectivamente:



3. Este gráfico con los índices de matriculaciones en base a 1996, muestra la evolución de las mismas en términos relativos y comparables y refleja la evolución más positiva registrada en la matriculación de turismos de importación que, en el gráfico anterior, con la escala de valores absolutos, no quedaba reflejada. El siguiente gráfico muestra en forma de columnas los porcentajes de variación anuales en las matriculaciones. En el mismo se incluye con cada columna del gráfico el valor que representa. Finalmente el último gráfico, también en forma de columnas, proporciona la participación alcanzada cada año por los turismos de importación. En él no se muestra el eje de ordenadas y cada columna figura con el valor que representa.

Matriculación de turismos. Tasas interanuales

Matriculación de turismos. Participación de turismos importados (%)



Capítulo 7
Variables y distribuciones

1. Unidades estadísticas y variables

1. Las *unidades estadísticas* son los *elementos* que componen el conjunto estudiado. Dicho conjunto debe estar definido con precisión y bien delimitado: es necesario conocer con exactitud si una cierta unidad pertenece o no al conjunto que se desea estudiar.

2. Cada unidad estadística puede ser descrita en relación a una o varias *características*. Así, si consideramos el personal del INE pueden estudiarse características como: sexo, edad, categoría profesional, estado civil, antigüedad, salario. En la matriculación de automóviles puede estudiarse si son o no de importación, marca y modelo, cilindrada, tipo de combustible que utilizan, provincia en que se matriculan. Estas posibles características a estudiar sobre las unidades se llaman *variables estadísticas* o simplemente variables.

3. Las variables pueden presentar distintas *modalidades o categorías*: la edad de una persona o su antigüedad en el trabajo puede tomar distintos valores, el combustible de un turismo puede ser gasolina súper, gasóleo, eurosuper, ... Las modalidades son las diferentes situaciones posibles de una variable y deben ser incompatibles y exhaustivas: cada elemento que se estudia debe presentar una y sólo una modalidad. En el caso de existir solamente dos modalidades se habla de característica o variable *dicotómica*.

4. Una variable es *cualitativa* o *de atributos* si sus diversas modalidades o categorías no son medibles: sexo, profesión, marca, tipo de combustible, ... Una variable es cuantitativa o *numérica* si sus modalidades son medibles, es decir, cada

modalidad está asociada a un número: edad, antigüedad en el trabajo, salario, ... En una variable cuantitativa su valor sobre cada unidad resulta de una medida numérica realizada sobre la misma, mientras que en una variable cualitativa su *valor* sobre cada unidad resulta de una clasificación que determina si la unidad pertenece o no a una modalidad, aunque ésta pueda identificarse con un código numérico. Por ejemplo el nivel profesional de un funcionario se expresa por un número pero es una variable cualitativa; la provincia de nacimiento es también una variable cualitativa aunque pueda expresarse por un código numérico asociado al nombre; la edad, la antigüedad en el empleo o el salario de una persona son variables cuantitativas. Es importante recalcar que las variables cualitativas no se miden numéricamente, sólo se ***clasifican***, mientras que las cuantitativas tienen una medida numérica.

5. Las variables cuantitativas pueden ser *discretas* o *continuas*, según que la medida de cada modalidad sea un número entero o un número real, respectivamente. El número de goles por partido de fútbol, el tamaño de una familia o el número de habitantes de los municipios españoles son variables discretas: ninguna familia está formada por 3,45 miembros y ningún partido de fútbol termina con 2,5 goles. La edad o el peso de una persona, el gasto en alimentación de una familia en un año o la hora de entrada al trabajo de los empleados de una empresa son variables continuas, que pueden tomar cualquier valor en un cierto intervalo: un empleado puede tener registrada su entrada al trabajo a las 8:32 horas, pero puede haber entrado en cualquier instante entre las 8:32 y las 8:33 horas. En la práctica las variables continuas se tratan como discretas al limitar el número de los decimales en su medición.

2. Distribuciones de frecuencia

1. Cuando sobre un conjunto de unidades estadísticas estudiamos una variable, cada unidad queda asociada al valor o categoría (clasificación) de la variable que le corresponde según que ésta sea cuantitativa o cualitativa respectivamente. El conjunto de unidades junto con el valor o categoría definido para cada unidad constituye una *distribución estadística*. Por ejemplo si en el colectivo de funcionarios del INE consideramos el grupo profesional (A, B, C o D) al que se pertenece, tendríamos la distribución de funcionarios por grupo:

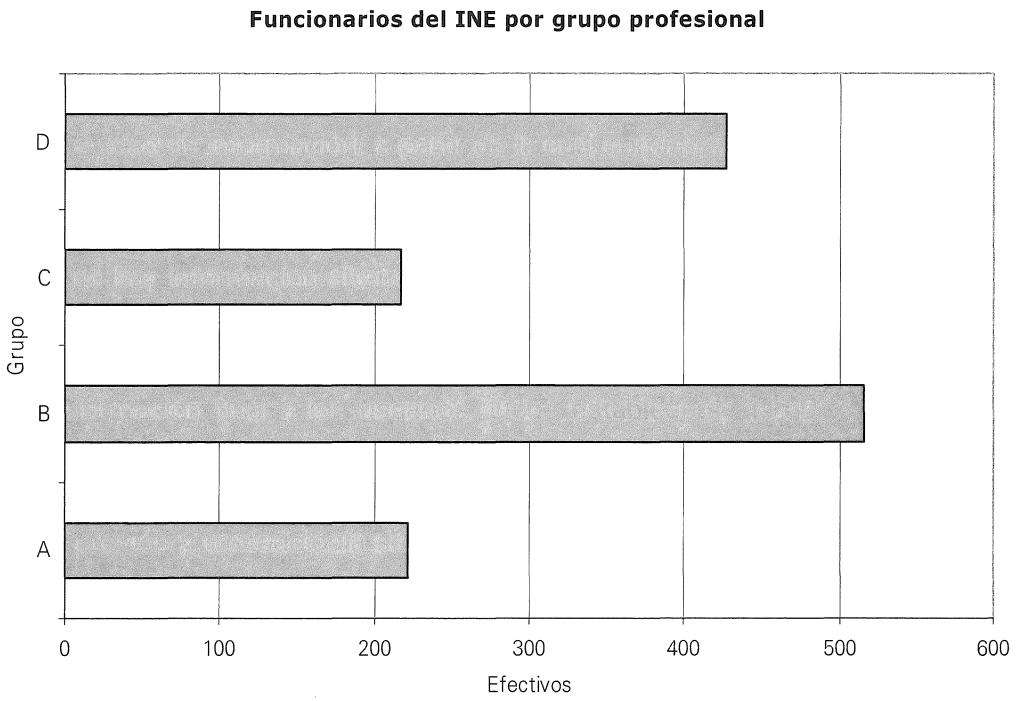
Funcionario	Grupo
1	D
2	B
3	C
4	C
5	A
6	B
...	...

Si ahora contamos el número de funcionarios en cada grupo profesional obtenemos la siguiente *tabla o distribución de frecuencias*:

Funcionarios INE por grupo profesional (31-12-2000)

Grupo	Efectivos	Porcentaje
A	222	16,1
B	516	37,3
C	217	15,7
D	427	30,9
Total	1.382	100,0

A los efectivos de cada grupo se les denominan *frecuencias absolutas*. Vemos, pues, que la frecuencia es el número de unidades que presentan la misma modalidad de la variable que se estudia. La proporción de cada categoría respecto al total se denomina *frecuencia relativa*. En forma gráfica:



2. Cuando tenemos una distribución cuantitativa pueden agruparse, de forma similar al caso cualitativo, las unidades que presentan un mismo valor de la variable para formar una tabla de frecuencias. Si la variable presenta un gran número de valores diferentes la agrupación de las unidades puede hacerse por *intervalos* como vemos en el siguiente ejemplo. Los datos corresponden al precio de venta de un mismo artículo tomado en 30 supermercados (euros):

6,20	5,45	9,60	4,80	7,75	6,00	6,00	4,15	5,40	5,60	6,55
4,75	6,05	5,20	5,20	5,65	6,65	8,75	7,70	8,60	3,35	4,35
5,90	4,25	6,35	4,30	6,75	9,90	4,20	3,60			

La ordenación de los datos proporciona una mejor panorámica de la distribución:

3,35	3,60	4,15	4,20	4,25	4,30	4,35	4,75	4,80	5,20	5,20
5,40	5,45	5,60	5,65	5,90	6,00	6,00	6,05	6,20	6,35	6,55
6,65	6,75	7,70	7,75	8,60	8,75	9,60	9,90			

Podemos agrupar los datos por intervalos de euro, es decir, precio menor o igual a 4 euros, mayor de 4 y menor o igual que 5, ..., mayor de 9, obteniéndose la siguiente tabla:

Precio	Frecuencia
4	2
5	7
6	9
7	6
8	2
9	2
10	2
Total	30

Los intervalos de precio que sirven para la agrupación de las unidades se llaman *intervalos de clase* o simplemente *clases*. Las clases se corresponden a las categorías de la variable. Obsérvese la forma de presentar las clases de precios en la tabla: para cada clase se da el límite superior y contiene las unidades con precio superior al límite de la clase anterior y menor o igual al límite de clase, así, la clase 7 contiene los supermercados con precio mayor de 6 (límite de clase anterior) y menor o igual que 7 euros. Cuando nos enfrentamos a un gran número de observaciones la agrupación en clases resulta imprescindible para poder apreciar la naturaleza general de la información. De hecho es bastante normal que al publicar resultados estadísticos nos encontremos con distribuciones de frecuencia agrupadas en clases. Así, la población ocupada por grupos de edad que proporciona la Encuesta de Población Activa en el 4º trimestre de 2002 es como sigue:

EPA. 4º trimestre 2002

Población ocupada según la edad	(miles)
Total	16.377,3
De 16 a 19 años	322,2
De 20 a 24 años	1.418,3
De 25 a 54 años	12.882,2
De 55 años y más	1.754,5

3. Cuando se divide una variable en intervalos de clase, la amplitud de cada clase va a depender del número de clases a considerar, que en general no debe ser superior a 15. Un número excesivo de clases puede originar ciertas irregularidades al tener frecuencias pequeñas por clase. Los límites o extremos de clase deben ser números redondos. La amplitud de clase puede ser constante o variable. Si la amplitud es constante ésta puede calcularse dividiendo la diferencia entre el mayor y menor valor por el número de clases y redondeando convenientemente. Si la amplitud de clase es variable, ésta será más pequeña en la parte de la variable que sea más frecuente y más grande en la parte con menor frecuencia. La frecuencia de clase dividida por su amplitud se llama *frecuencia media de clase*. Las clases suelen definirse de forma que las frecuencias de cada clase sean progresivamente crecientes para, después de un máximo volverse sucesivamente decrecientes. Es aconsejable también que las amplitudes de clase sean múltiplos de un número fijo (por ejemplo sí la amplitud menor es 25, que todas las amplitudes de clase sean múltiplos de 25).

3. Diagrama de barras e histograma de frecuencias

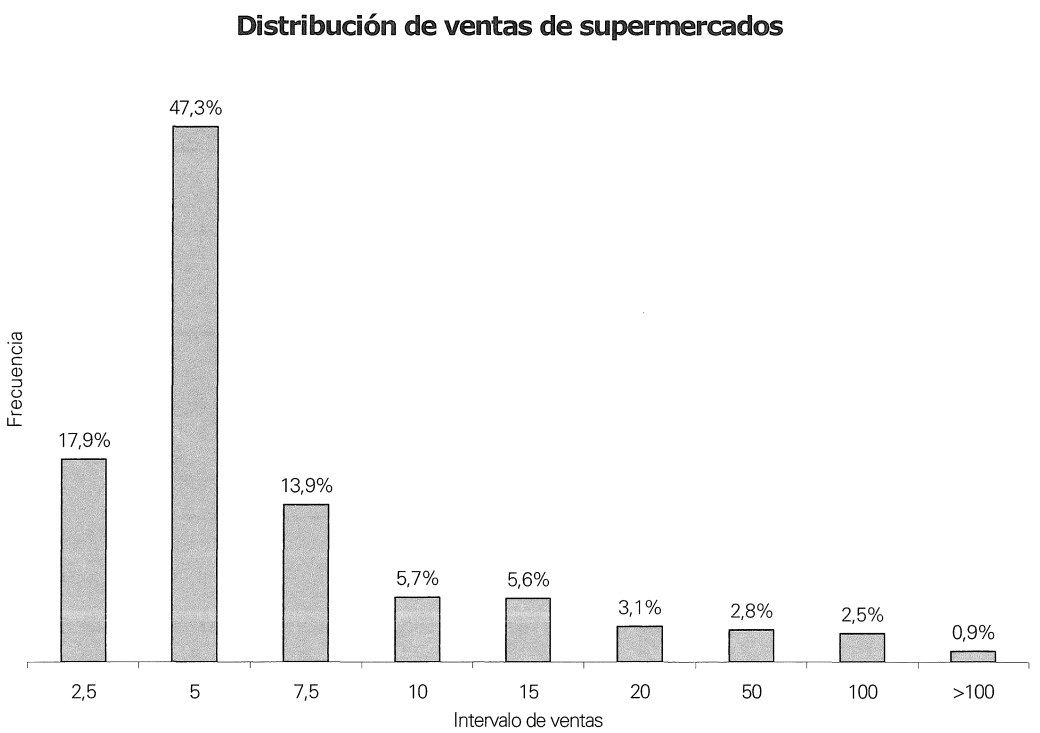
1. Vamos a considerar la distribución de ventas de un cierto año de una población de 2.960 supermercados con superficie de ventas igual o superior a 400 m^2. La distribución de frecuencias absolutas y porcentuales se dan en la tabla que sigue. Puede verse que los primeros intervalos de venta tienen amplitud igual a 2,5 millones de euros, siguen dos intervalos (10-15, 15-20) de amplitud igual a 5 millones, otro más (20-50) de amplitud 30 millones y uno de amplitud igual a 50 millones, quedando el último abierto y que incluye los establecimientos con ventas superiores a 100 millones de euros. Se proporciona también la distribución acumulativa de frecuencias, es decir, en cada intervalo de clase la frecuencia acumulada es la frecuencia de su clase más la de las clases inferiores. Así, vemos que el 84,9% de los supermercados presentan ventas iguales o inferiores a 10 millones de euros (17,9 + 47,3 + 13,9 + 5,7 = 84,9%).

Distribución de ventas de supermercados

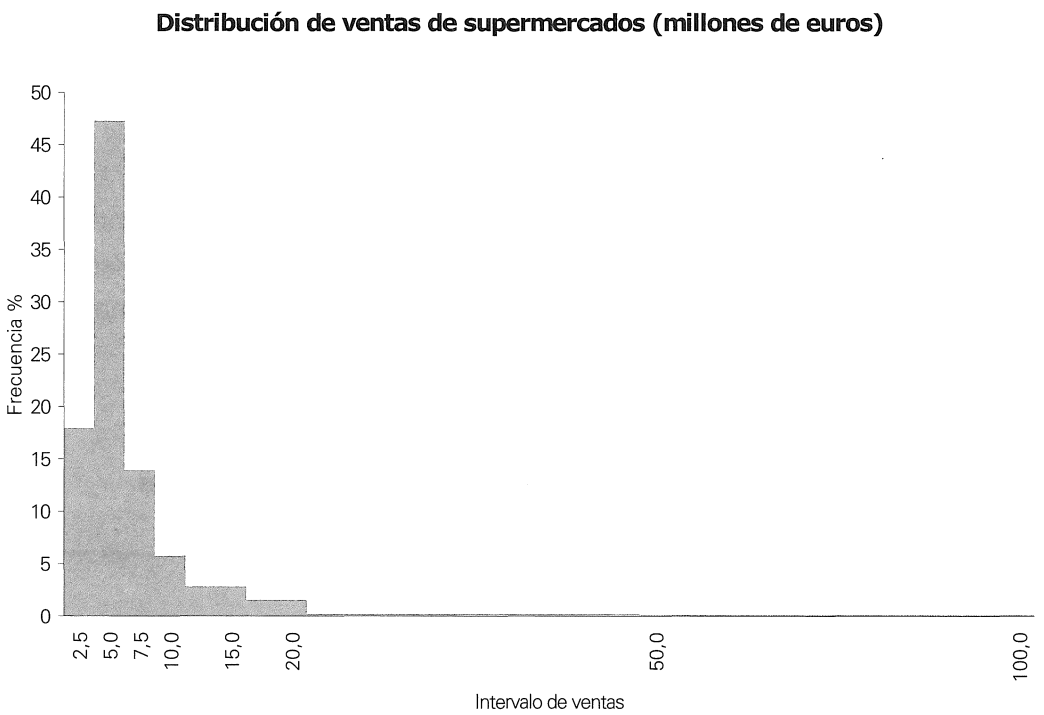
Millones de euros

Intervalo ventas	Número supermercados	Frecuencia relativa	Frecuencia acumulada
2,5	531	17,9	17,9
5,0	1.400	47,3	65,2
7,5	412	13,9	79,2
10,0	170	5,7	84,9
15,0	167	5,6	90,5
20,0	93	3,1	93,7
50,0	84	2,8	96,5
100,0	75	2,5	99,1
Mayor	28	0,9	100,0
Total	2.960	100,0	

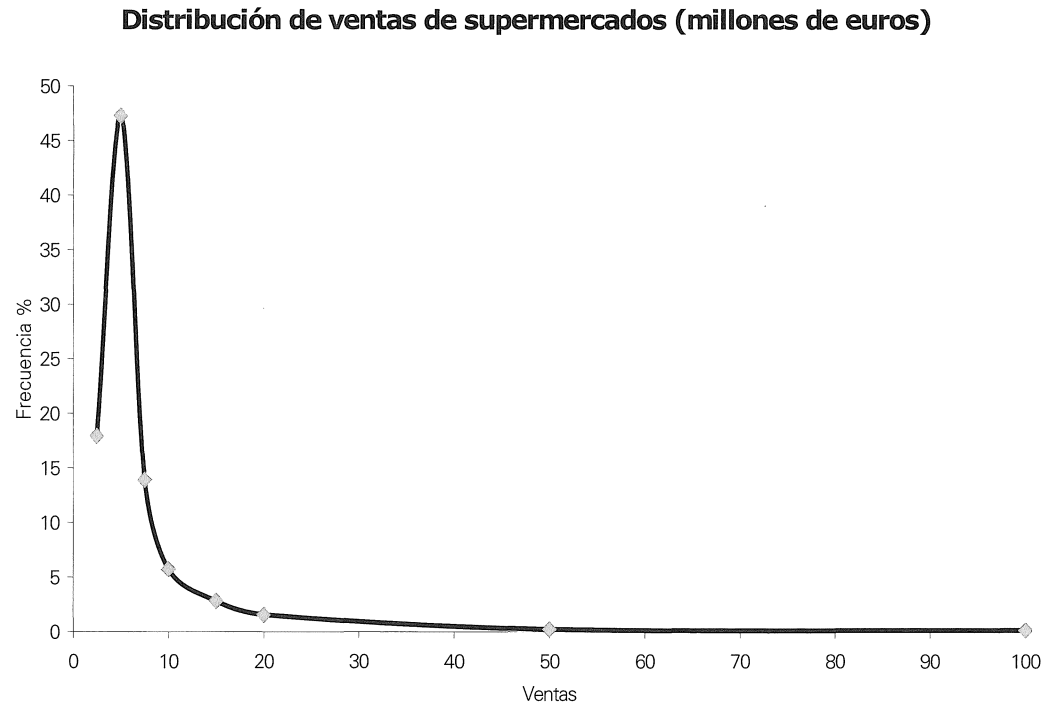
2. La distribución anterior es un ejemplo de distribución asimétrica, con altas frecuencias en valores moderados de la variable, pero con una cola de unidades en valores muy altos de la variable, en este caso las ventas. Con este tipo de distribuciones resulta prácticamente inevitable que los intervalos de clase tengan amplitud variable para poder llegar a un número razonable de intervalos. Obsérvese que todas las amplitudes de clase son múltiplos de la amplitud menor, así, la clase 20-50 equivale a 12 veces la amplitud menor. Si llevamos la distribución de frecuencias a un gráfico de barras como el que sigue, debe tenerse presente que el gráfico se caracteriza porque la altura de cada barra es proporcional al valor que representa y que las clases se representan en el eje X como modalidades, es decir, sin escala numérica:



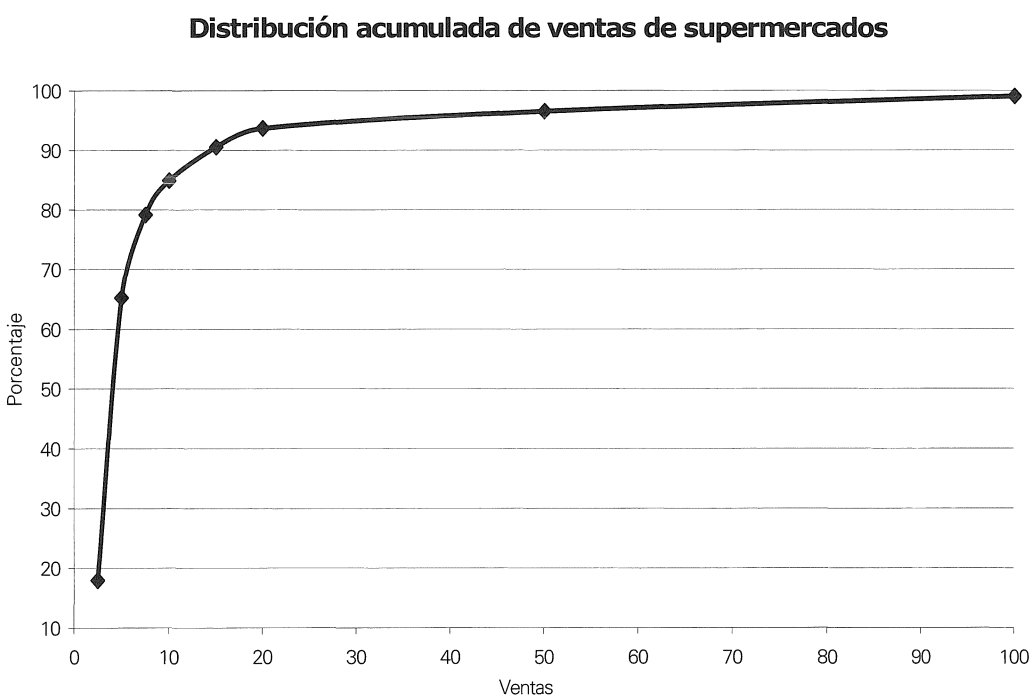
3. En distribuciones de frecuencia con amplitudes de clase variables, otra forma de representación gráfica es el *histograma de frecuencias*, en el que cada clase se representa por un rectángulo cuya base es la amplitud del intervalo y cuya superficie representa la frecuencia de clase. A continuación se presenta el histograma de la distribución de supermercados. Puede observarse como en los intervalos de amplitud 2,5 la altura de la columna del histograma es igual a la altura del gráfico de barras ya que se toma la amplitud menor como unidad para medir la superficie de la columna, pero en los intervalos 10-15 y 15-20, cuya amplitud es el doble, la altura de la columna es la mitad de la frecuencia respectiva, de forma que el área representa correctamente la frecuencia de cada clase, y análogamente sucede con el resto de intervalos. Así pues, la diferencia entre ambos tipos de gráficos es que la frecuencia, que se representa por la altura en el diagrama de barras, corresponde en el histograma al área de la columna. Ello implica introducir una escala numérica en el eje X.



4. El gráfico que sigue es otra forma de ver el histograma y permite apreciar el alto grado de asimetría de la distribución, que, en el diagrama de barras queda algo disimulado por la falta de escalado en el eje X.



5. En tabla de distribución de ventas de supermercados se había incluido una columna con la distribución acumulada de frecuencias. El gráfico de la distribución acumulada permite apreciar visualmente valores de la variable de estudio por debajo de los cuales se sitúa un cierto porcentaje de las unidades. Podemos apreciar en el gráfico que alrededor del 94% de supermercados presentan ventas de hasta 20 millones de euros y tan sólo un 6% supera ésa cifra. El gráfico acumulado resulta de utilidad para apreciar visualmente ciertos valores de la variable, como veremos en el siguiente capítulo.



Capítulo 8
Medidas descriptivas de una variable estadística

1. Los conceptos de media y desviación típica

1. Volvamos al conjunto de precios de venta de un artículo en 30 supermercados del capítulo anterior. Al ser un conjunto no muy amplio de elementos la simple inspección visual de los datos permite extraer alguna conclusión como que los precios varían entre 3,35 y 9,90 euros y que los valores más frecuentes parecen estar entre los 5 y 7 euros. Estas conclusiones son más evidentes cuando se ven los precios ordenados. Poca información más puede extraerse de la simple inspección visual de los datos individuales, y esto porque estamos ante un conjunto reducido de elementos. Si nos tuviéramos que enfrentar a una lista de unos pocos cientos o miles de precios, la inspección visual de los mismos no nos aportaría prácticamente ninguna información. *Se hace necesario clasificar, resumir y simplificar para poder comprender.* Las primeras dos medidas que resumen una distribución estadística son la media y la desviación típica.

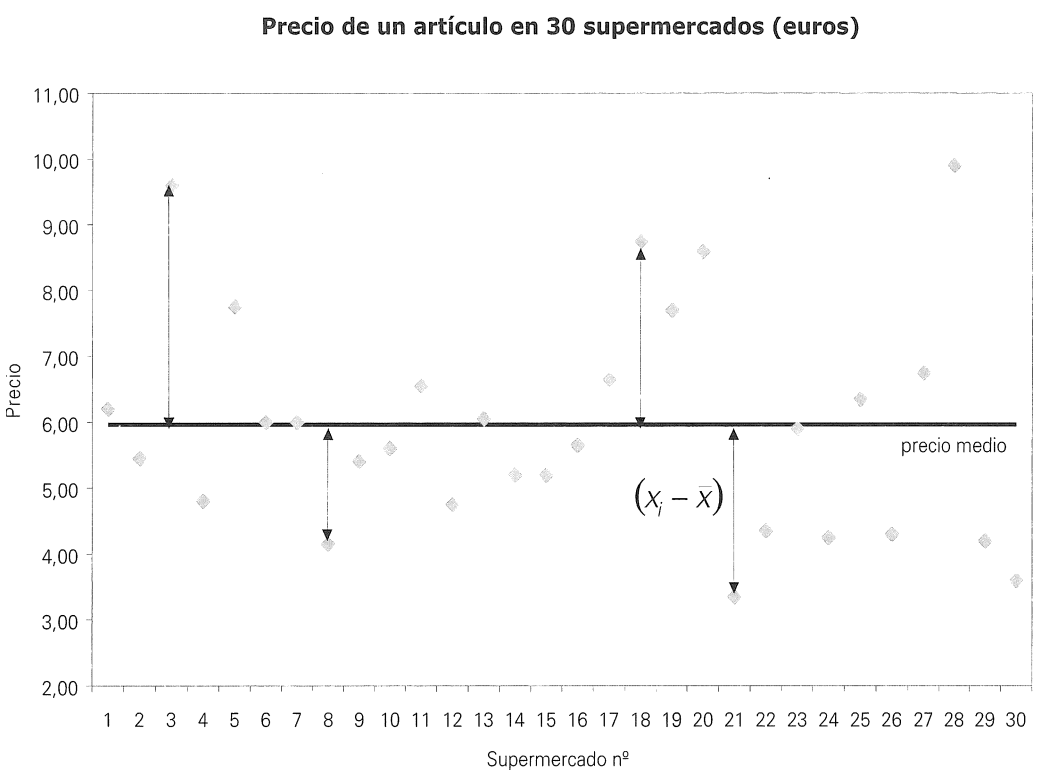
2. La *media aritmética* de una distribución es el resultado de sumar el valor de la variable de estudio de todas las unidades y dividir por el número de unidades. Es decir,

$$\overline{x} = \frac{\sum x}{n}$$

donde $\sum x$ representa la suma de todos los valores de la variable y n es el número de valores. En el caso de los precios de venta, obtenemos el precio medio sumando todos los precios obtenidos y dividiendo por el número de precios, es decir:

$$\text{precio medio} = \bar{x} = \frac{6{,}20 + 5{,}45 + \cdots + 4{,}20 + 3{,}60}{30} = 5{,}97 \text{ euros}$$

El siguiente gráfico representa los precios individuales obtenidos en cada supermercado y su valor promedio; puede observarse como la media se sitúa entre los valores individuales de la variable, es decir, la media es una *característica central* o medida centralizada de la distribución: *los valores individuales se distribuyen alrededor de la media.*



3. El gráfico refleja cómo los valores individuales de la variable varían alrededor de la media. La variabilidad de los datos es una característica inherente a la Estadística. La medida de la variabilidad o *dispersión* de los datos alrededor de la media la proporciona la *desviación típica*, cuyo cálculo se basa en las diferencias $(x_i - \bar{x})$ entre los valores individuales y la media, reflejadas en el gráfico por algunas flechas. La suma de estas diferencias es cero, por lo que el cálculo de la desviación típica se hace con el cuadrado de las diferencias y responde a la fórmula

$$\sigma = \sqrt{\frac{\sum (x - \bar{x})^2}{n}}$$

es decir, a) se calculan las diferencias $(x - \bar{x})$ individuales a la media y se elevan al cuadrado, b) se suman los cuadrados obtenidos y se divide por el número de unidades y c) se extrae la raíz cuadrada. Para los precios obtenemos

$$\sigma = \sqrt{\frac{(6{,}20 - 5{,}97)^2 + (5{,}45 - 5{,}97)^2 + \cdots + (3{,}60 - 5{,}97)^2}{30}} = 1{,}64 \text{ euros}$$

Si no se extrae la raíz cuadrada, el resultado que se obtiene se llama *varianza* y se denota por σ^2. La desviación típica y la varianza son *medidas de dispersión.* La desviación típica se llama también *desviación estándar.*

4. El significado de la desviación típica se entiende mejor si se tiene en cuenta la siguiente propiedad: en cualquier distribución estadística al menos el 75% de los valores individuales se encuentran comprendidos entre la media y más menos dos veces la desviación típica. En el caso de los precios, al menos el 75% de los mismos estarían comprendidos entre $5{,}97 - 2 \times 1{,}64 = 2{,}69$ euros y $5{,}97 + 2 \times 1{,}64 = 9{,}25$ euros; en concreto, 28 de los 30 precios (el 93%) están comprendidos entre los citados valores. Resulta evidente que cuanto mayor sea el porcentaje de valores comprendidos entre la media y más menos dos veces la desviación típica mayor es la concentración de valores alrededor de la media o, lo que es lo mismo, menor es la dispersión de los valores individuales.

5. La desviación típica se expresa en las mismas unidades que la media por lo que no es directamente comparable entre distintas distribuciones. Así, si tuviéramos la distribución de precios de un segundo artículo en los mismos 30 supermercados con un precio medio de 8,00 euros y la misma desviación típica, 1,64 euros, podemos pensar que este segundo artículo presenta la misma dispersión que el primero, lo cuál es cierto en términos absolutos, pero si expresamos la desviación típica en términos relativos a la media, es decir, como cociente a la media nos encontramos con que el primer artículo presenta una variación relativa de

$$\frac{1{,}64}{5{,}97} = 0{,}275 \rightarrow 27{,}5\%$$

mientras que para el segundo la variación relativa es

$$\frac{1{,}64}{8{,}00} = 0{,}205 \rightarrow 20{,}5\%$$

Es decir, el segundo artículo presenta menor dispersión en los precios.

El cociente entre la desviación típica y la media se llama *coeficiente de variación* (CV):

$$CV = \frac{\sigma}{\bar{x}}$$

y puede expresarse en porcentaje siendo una medida de dispersión comparable entre distintas distribuciones.

6. Tanto la media como la desviación estándar sólo son calculables con variables numéricas, ya que se requiere calcular la suma de los valores de la variable, y esto, obviamente, es posible si la variable toma valores numéricos. Con variables cualitativas pueden calcularse porcentajes: porcentaje de mujeres, porcentaje de turismos que utilizan combustible eurosuper. Así, en la tabla de funcionarios del INE de 7.2.1 puede verse que los funcionarios del grupo A son el 16,1% del total. Estos porcentajes, en distribuciones cualitativas se conocen como *proporciones* y hacen referencia a la fracción porcentual de la población que cumple una determinada propiedad.

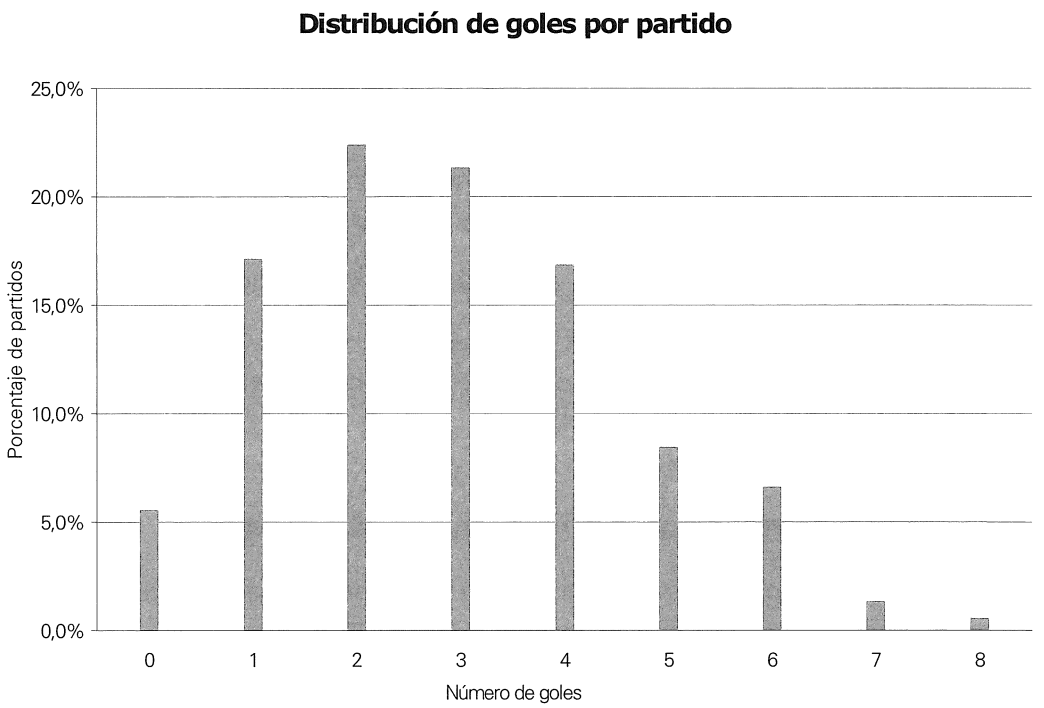
7. Si para una variable cualitativa estamos interesados en una determinada modalidad o categoría *C* de la variable, podemos asignar el valor 1 a las unidades que pertenecen a la categoría C y el valor 0 a las que no pertenecen. Siendo *c* el número de unidades que pertenecen a la clase en cuestión, la media de esta variable dicotómica es $\bar{x} = \frac{\sum x}{n} = \frac{c}{n} = p$, es decir, la proporción de unidades que pertenecen a la clase y la expresión de la varianza se convierte en $\sigma^2 = p(1-p)$.

2. Datos agrupados en una distribución de frecuencias

1. La siguiente tabla de frecuencias proporciona el número de goles marcados en cada uno de los 380 partidos de fútbol de la 1ª división de la liga española de la temporada 2000-2001:

Número de goles	Número de partidos	Porcentaje partidos
0	21	5,5
1	65	17,1
2	85	22,4
3	81	21,3
4	64	16,8
5	32	8,4
6	25	6,6
7	5	1,3
8	2	0,5
Total	380	100,0

La variable estadística que se considera es el total de goles marcados en cada partido, cuyo valor varía entre 0 y 8, dándose para cada valor de la variable su frecuencia absoluta (número de partidos en que se repite el valor) y su frecuencia relativa en porcentaje. El gráfico muestra el diagrama de barras de la distribución relativa de frecuencias:



2. La media y la desviación típica de la distribución se calculan de la forma ya vista. Para el cálculo de la media puede tenerse en cuenta que el valor 0 se repite 21 veces y habría que sumarlo 21 veces (0×21=0), el valor 1 se repite 65 veces y habría que sumarlo 65 veces (1×65=65), ..., y de forma análoga se repiten las diferencias a la media para el cálculo de la desviación típica. Ello nos lleva a

$$\text{media} = \overline{x} = \frac{0 \cdot 21 + 1 \cdot 65 + 2 \cdot 85 + \cdots + 7 \cdot 5 + 8 \cdot 2}{380} = 2{,}88 \text{ goles/partido}$$

$$\text{d. típ.} = \sigma = \sqrt{\frac{(0 - 2{,}88)^2 \cdot 21 + (1 - 2{,}88)^2 \cdot 65 + \cdots + (8 - 2{,}88)^2 \cdot 2}{380}} = 1{,}68 \text{ goles}$$

Es decir, la media y la desviación típica se calculan de acuerdo a las fórmulas

$$\overline{x} = \frac{\sum xf}{\sum f} = \frac{\sum xf}{n}$$

$$\sigma = \sqrt{\frac{\sum (x - \overline{x})^2 f}{\sum f}} = \sqrt{\frac{\sum (x - \overline{x})^2 f}{n}}$$

respectivamente, y en las que *f* indica la frecuencia con que se presenta cada valor. Aunque ambas fórmulas puedan aparentar ser diferentes a las vistas en el apartado anterior, son conceptualmente iguales y lo único que hacen es tomar en cuenta la frecuencia repetitiva de los distintos valores de la variable a la hora de sumar. En el cálculo pueden emplearse indistintamente las frecuencias absolutas o las relativas. En caso de utilizar las relativas debe tenerse presente que la suma de las frecuencias relativas ($\sum f$) es 1 o 100 según que se utilicen en proporción a la unidad o en porcentaje. El coeficiente de variación resulta de 1,68/2,88 = 58,3%, es decir, la distribución de goles por partido presenta mayor dispersión que la de precios vista en 8.1.5.

3. Datos agrupados en intervalos de clase

1. Vamos a ver como se realiza el cálculo de la media y la varianza cuando tenemos los datos agrupados en intervalos de clase. Tomaremos como referencia la vida útil de un conjunto de 60 bombillas:

Duración en horas de sesenta bombillas

620	650	660	675	717	732	747	753	753	755
766	776	781	787	788	792	803	807	811	811
817	817	822	823	827	828	829	831	832	833
841	844	847	852	857	863	867	869	872	878
880	881	889	890	891	897	907	918	923	933
947	958	970	980	1.030	1.050	1.056	1.076	1.082	1.088

Si calculamos la media y la desviación típica utilizando los 60 valores de la tabla de acuerdo a las fórmulas vistas en 8.1.2 y 8.1.3 obtenemos 851,32 horas para la media y 102,57 horas para la desviación estándar, con un coeficiente de variación del 12,05%. Realicemos ahora una agrupación de los valores en intervalos de clase. Teniendo en cuenta que la diferencia entre el mayor y el menor valor es 1.088-620=468, y de acuerdo a lo indicado en el párrafo 7.2.3, podemos formar clases de amplitud 50, obteniendo la siguiente distribución de frecuencias:

Intervalo clase	Frecuencia absoluta (*f*)
600	0
650	2
700	2
750	3
800	9
850	17
900	13
950	5
1.000	3
1.050	2
1.100	4
mayor	0
Total	60

Recuérdese la forma de presentar las clases por su límite superior, esto es, las 9 observaciones en 800 corresponden a las 9 bombillas cuya duración es mayor que 750 horas y menor o igual que 800 horas.

Cuando nos encontramos con una distribución de frecuencias por intervalos de clase, la información que tenemos es que 9 bombillas duran entre 750 y 800

horas, pero no conocemos la duración individual de cada una de las 9 bombillas incluidas en esta clase, y análogamente sucede con el resto de clases. La solución para calcular la media y la varianza es suponer que las bombillas incluidas en cada intervalo tienen una duración media igual al valor central del intervalo, que se denomina *marca de clase*. Es decir, suponemos que las 9 bombillas del intervalo 750-800 tienen una duración media de $750 + \frac{800 - 750}{2} = 775$ horas, y así para el resto de intervalos de clase. A partir de aquí el cálculo de la media y varianza se realiza en la forma vista en el ejemplo de los goles, utilizando la marca de clase como valor de la variable en cada intervalo. En la tabla que sigue se detallan los cálculos:

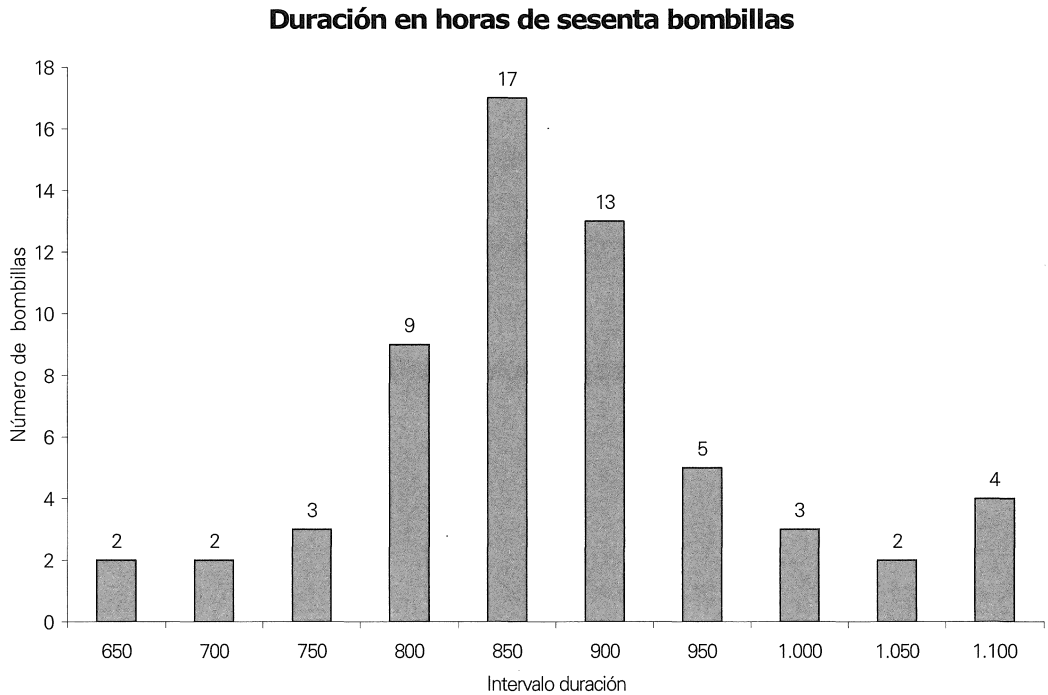
Distribución de frecuencias de la duración de bombillas y cálculo de la media y desviación estándar

Intervalo clase	Frecuencia absoluta (*f*)	Frecuencia relativa % acumulada	Marca de clase (*x*)	xf	$(x - \overline{x})^2 f$
600	0	,00%			
650	2	3,33%	625	1.250,00	102.001,39
700	2	6,67%	675	1.350,00	61.834,72
750	3	11,67%	725	2.175,00	47.502,08
800	9	26,67%	775	6.975,00	51.756,25
850	17	55,00%	825	14.025,00	11.345,14
900	13	76,67%	875	11.375,00	7.592,36
950	5	85,00%	925	4.625,00	27.503,47
1.000	3	90,00%	975	2.925,00	46.252,08
1.050	2	93,33%	1.025	2.050,00	60.668,06
1.100	4	100,00%	1.075	4.300,00	201.002,78
mayor	0	100,00%			
Total	60			51.050,00	617.458,33
		media /	varianza	850,83	10.290,97
			desv. est.		101,44
			CV		11,92%

La agrupación de valores en intervalos de clase produce un cierta pérdida de información, aunque permite apreciar mejor la naturaleza de los datos. Pensemos que si en lugar de los 60 valores del ejemplo tuviéramos 60.000 poco apreciaríamos mirándolos uno a uno. Debido a esta pérdida de información es necesario

introducir la suposición respecto a las marcas de clase para los cálculos, resultando que los valores obtenidos para la media y la desviación estándar son aproximaciones a los valores verdaderos de 851,32 y 102,57 respectivamente, que habíamos obtenido utilizando todos los valores, es decir, toda la información.

El siguiente gráfico refleja la distribución de frecuencias por clases de las bombillas:

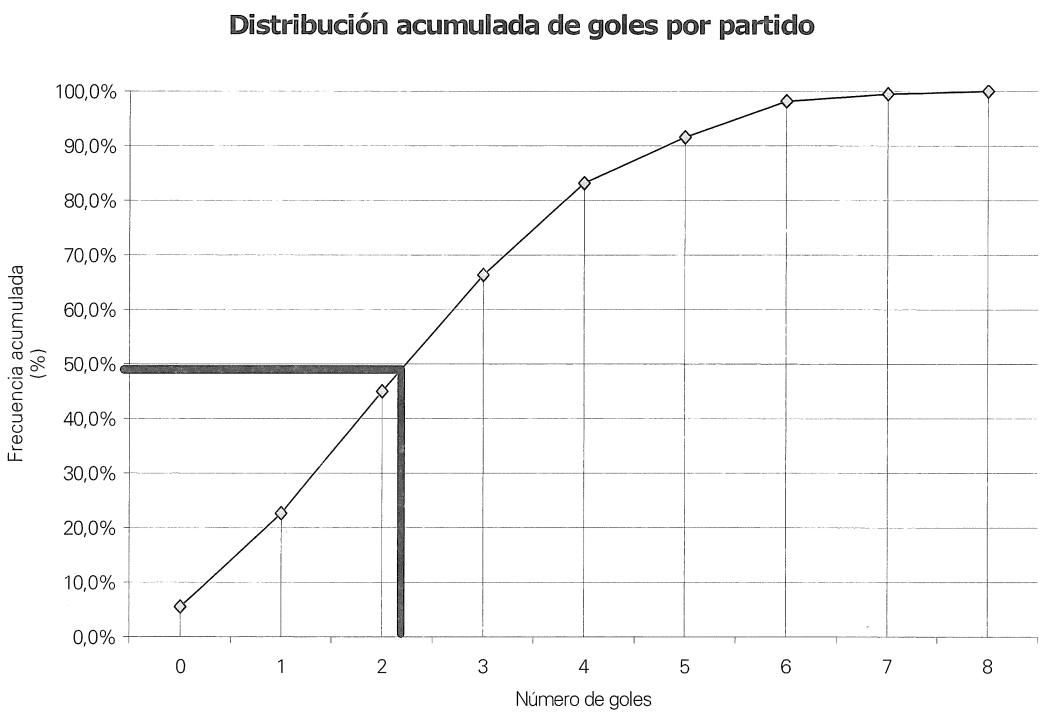


4. Otras medidas características

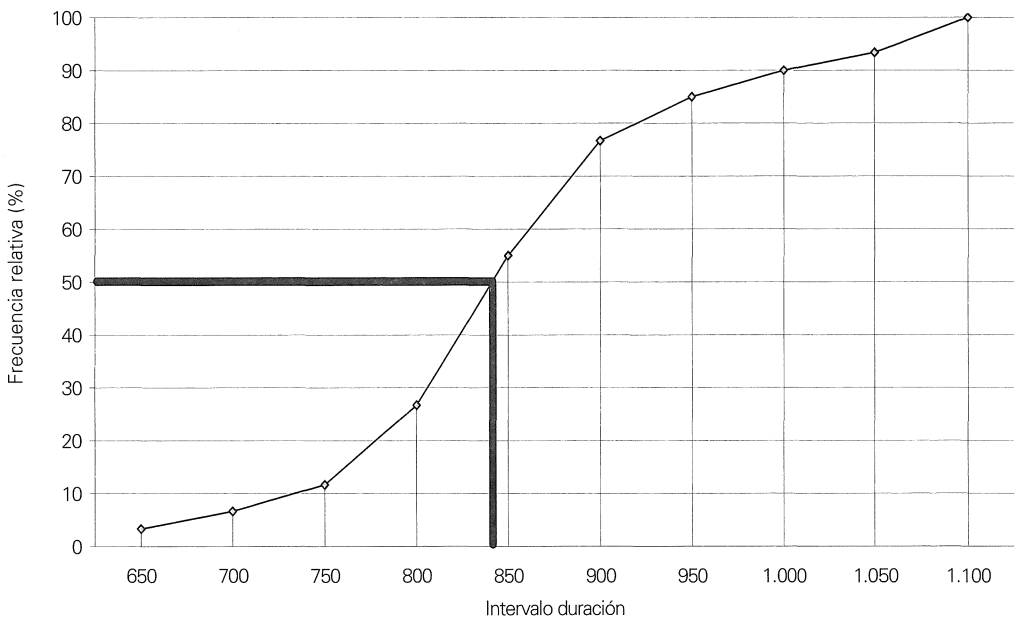
1. La *moda* es el valor más frecuente de la variable estadística, y se corresponde con el máximo del diagrama de barras. En la distribución de goles la moda son 2 goles por partido, que se presenta en 85 de los 380 partidos. En el caso de las bombillas, con una distribución en intervalos de clase, el intervalo modal es 800 – 850, en el que se encuentran 17 de las 60 bombillas. La moda representa, como la media, un valor central de la distribución y su determinación es visual a partir de la tabla de frecuencias o de su gráfico.

2. La *mediana* divide a la población en dos partes iguales o aproximadamente iguales en número de unidades de estudio o unidades estadísticas. La mitad de las unidades presentan valores de la variable menores o iguales que la mediana en

tanto que la otra mitad tiene valores iguales o mayores que la mediana. La determinación de la mediana requiere la ordenación de las unidades de la población por valores de la variable de estudio en forma creciente o decreciente. Si el número de unidades *n* es impar la mediana corresponde al valor de la unidad que ocupa el lugar $\frac{n-1}{2}+1$ en la ordenación, y si *n* es par la mediana se calcula como la media de los dos valores que ocupan la posición $\frac{n}{2}$ y $\frac{n}{2}+1$. En definitiva, si tenemos los siete valores (5, 5, 6, **8**, 8, 10, 10) la mediana es 8, pero si tenemos los seis valores (5, 5, **6**, **8**, 8, 10) la mediana es $\frac{6+8}{2}=7$. El gráfico de la distribución acumulativa de la frecuencia relativa permite apreciar de forma rápida la mediana o el intervalo mediano. En el caso de los goles vemos gráficamente como la mediana está comprendida entre dos y tres goles por partido:



Frecuencia relativa acumulada de la duración de sesenta bombillas

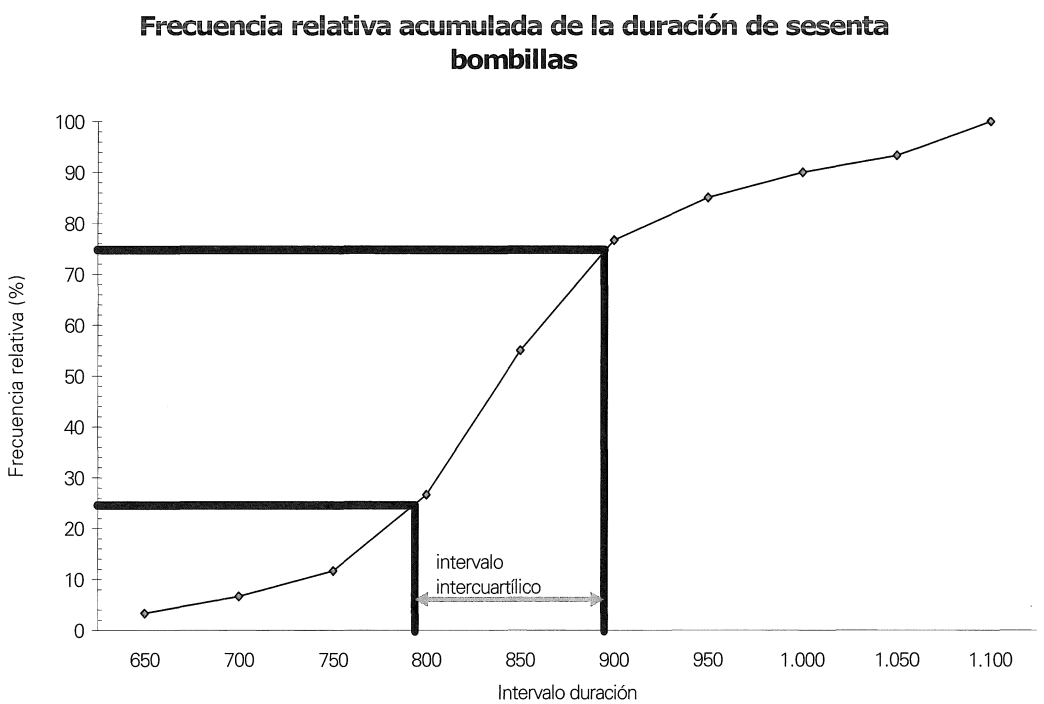


En la distribución de duración de bombillas por intervalos de clase, el intervalo en el que se incluye la mediana es el 800 – 850, bastante cerca de 850 horas. Si observamos la serie de sesenta valores los dos valores centrales (30 y 31) son 833 horas y 841 horas y la mediana sería la media de ambos (*n* es par), es decir, 837 horas. Vemos pues como el gráfico de frecuencias relativas acumuladas permite tener una apreciación inmediata de la mediana.

3. Tanto la moda como la mediana son, al igual que la media, características centrales de una distribución. A diferencia de la media, su cálculo no está basado en todos los valores de la distribución, por lo que posibles valores extremos no tienen incidencia en el valor de la moda o mediana, mientras que sí pueden influir de forma notable sobre la media. Así, la mediana de los 7 valores (5, 5, 6, 8, 8, 10,10) es 8 y su media es 7,42. Ahora, para los 7 valores (5, 5, 6, 8, 8, 10, 100), la mediana sigue siendo 8, pero la media es 20,3 por la influencia del valor 100, que es claramente un *valor atípico*. Cuando se tienen valores extremadamente altos o bajos en una distribución, es aconsejable calcular la media con todos los valores y realizar también su cálculo excluyendo los valores extremos para poder decidir cuál de los

valores obtenidos representa o *condensa* mejor al conjunto. El cálculo de la mediana puede ayudar a decidir qué valor medio es el adecuado.

4. En forma similar a la mediana que divide a la población en dos partes con el mismo número de unidades, a partir de la ordenación de las unidades se pueden determinar otros valores que caracterizan la ordenación. El *cuartil 1/4* sería el valor de la variable que deja por debajo al 25% de las unidades, el *cuartil 3/4* sería el valor de la variable que deja por debajo al 75% de las unidades, la *decila 1* corresponde al valor que deja por debajo al 10% de las unidades, la *decila 2* deja por debajo al 20% de unidades, ... A la diferencia *cuartil 3/4 cuartil 1/4* se le denomina *intervalo intercuartílico* y representa la zona central de la variable que incluye al 50% de sus valores.



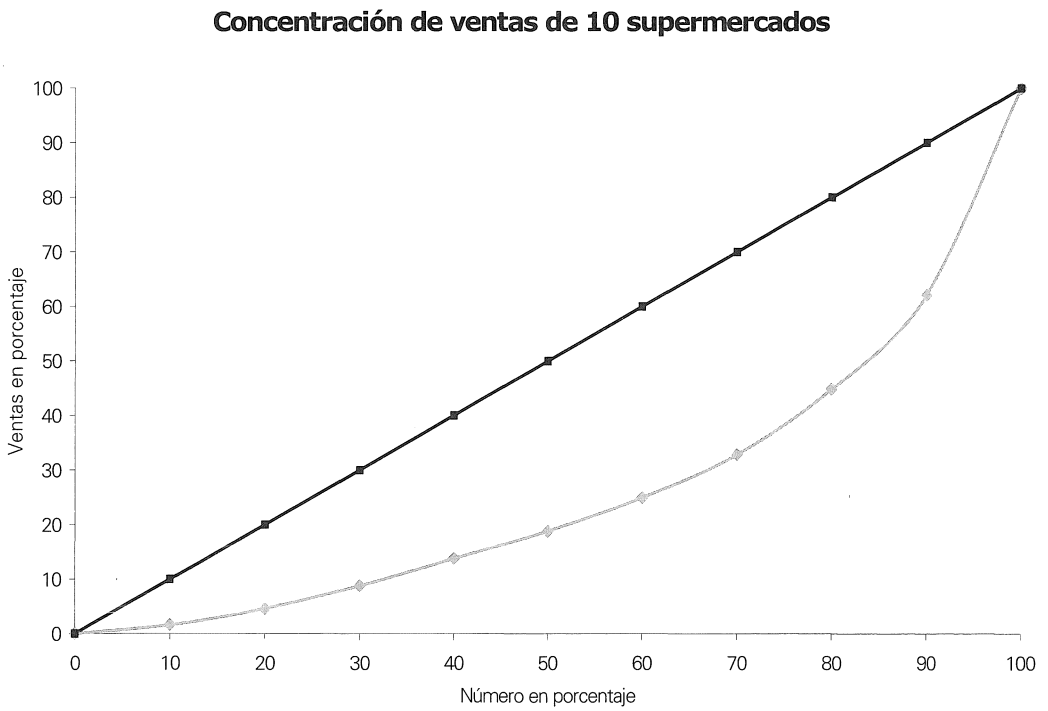
5. La curva de concentración

1. Vamos a considerar la siguiente tabla con las ventas anuales en millones de euros de 10 supermercados.

Ventas anuales en millones de euros de 10 supermercados

Superm. Número	%número	%número acum.	Ventas	%Ventas	%Ventas acum.
1	10	10	1,06	1,6	1,6
2	10	20	2,00	3,0	4,6
3	10	30	2,77	4,2	8,8
4	10	40	3,32	5,0	13,8
5	10	50	3,35	5,0	18,8
6	20	60	4,07	6,1	25,0
7	10	70	5,23	7,9	32,8
8	10	80	7,97	12,0	44,8
9	10	90	11,48	17,3	62,1
10	10	100	25,18	37,9	100,0
Total	10		66,42		

La primera columna contiene la identificación de cada supermercado, la segunda representa el porcentaje de cada observación (1/10), que se acumulan en la tercera columna, la cuarta columna contiene las ventas que realiza cada uno de los supermercados, las cuales se trasladan a porcentajes sobre el total de ventas en la quinta columna y se acumulan en la última columna. Obsérvese que los supermercados están ordenados por ventas de menor a mayor, de forma que podemos ver comparando la tercera y última columna que los tres supermercados de menor venta sólo realizan el 8,8% de las ventas totales de los 10 considerados, aunque sean el 30% de las observaciones. La tercera y última columna en forma gráfica son la *curva de concentración*:



2. La concentración refleja la medida en que unas pocas observaciones presentan una gran contribución al total de la variable de estudio. Puede verse en la tabla que el supermercado con mayores ventas hace el 37,9% de las ventas de los 10. Si los 10 supermercados vendieran exactamente lo mismo, cada uno de ellos contribuiría a las ventas totales con un 10%, lo mismo que contribuyen en número, hablaríamos de concentración nula (todos iguales) y la curva de concentración coincidiría con la diagonal del gráfico. En el otro extremo, si un solo supermercado realizara todas las ventas y los otros tuvieran venta nula, la concentración sería la máxima posible y la curva se confundiría con los dos lados del triángulo inferior derecho formados por el eje X y la ordenada del punto (100,100). De ahí que una medida de la concentración sea la proporción de superficie del triángulo inferior derecho que queda entre la diagonal y la curva de concentración: 0 si no hay concentración y 1 con máxima concentración.

3. Volvamos a la distribución de ventas de supermercados que vimos en el apartado 7.3, dónde se expusieron algunos gráficos descriptivos de la distribución. Las ventas tienen un valor mínimo de 0,9 millones y un máximo de 201 millones de euros, es decir, la variable presenta un *recorrido* (diferencia entre el mayor valor y

el menor valor) de casi 200 millones. Calculados con todos los valores individuales, algunos valores característicos son: venta media = 8,54 millones, desviación típica = 17,02 millones, coeficiente de variación = 199%, mediana = 3,91 millones. Vemos que es una distribución con alta variabilidad, como ya se desprendía de los gráficos descriptivos, y su alto grado de asimetría, visto también en los gráficos, se refleja en la fuerte diferencia entre la media y la mediana, influenciada la media por las altas ventas de los grandes supermercados.

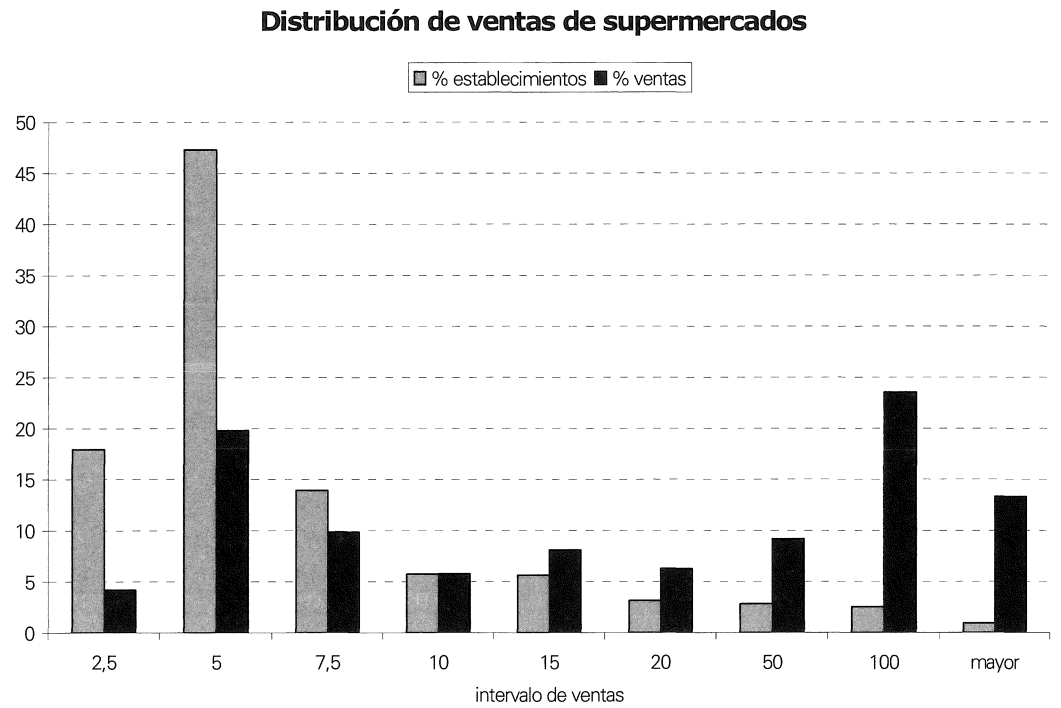
4. Vamos a añadir a la tabla de frecuencias del capítulo anterior las ventas que totalizan los supermercados incluidos en cada intervalo de clase, es decir, vamos a considerar la siguiente tabla:

Distribución de ventas de supermercados (millones de euros)

Intervalo ventas	Número superm.	Frecuencia relativa	Frecuencia acumulada	Ventas totales	%Ventas	%Ventas acumulado
2,5	531	17,9	17,9	1.058	4,2	4,2
5	1.400	47,3	65,2	5.001	19,8	24,0
7,5	412	13,9	79,2	2.489	9,9	33,8
10	170	5,7	84,9	1.459	5,8	39,6
15	167	5,6	90,5	2.046	8,1	47,7
20	93	3,1	93,7	1.585	6,3	54,0
50	84	2,8	96,5	2.320	9,2	63,2
100	75	2,5	99,1	5.946	23,5	86,7
mayor	28	0,9	100,0	3.361	13,3	100,0
Total	2.960	100,0		25.265	100,0	

Vemos que los 2.960 supermercados totalizan 25.265 millones de euros de ventas, de los cuales 5.001 millones (el 19,8%) corresponden a los 1.400 supermercados (el 47,3%) cuyas ventas individuales están comprendidas entre 2,5 y 5 millones de euros, mientras que el 0,9% de supermercados que venden por encima de 100 millones suponen el 13,3% de las ventas. También se ha añadido una última columna con las ventas acumuladas en forma porcentual, dónde se puede ver que el 90,5% de supermercados con ventas iguales o inferiores a 15 millones totalizan el 47,7% de las ventas, mientras que el 9,5% restante de establecimientos se reparten el otro 52,3% de las ventas. La tabla es totalmente similar a la del párrafo 1, pero resumiendo los datos por intervalos de clase, en lugar de considerar la lista ordenada de los 2.960 supermercados. Tablas como las anteriores que

recogen la ordenación de las unidades por valores de la variable y acumulan los porcentajes de unidades y los de la variable de estudio, se llama *tabla de concentración*. El siguiente gráfico de barras recoge los porcentajes de establecimientos (frecuencia) y de ventas que se registran en cada intervalo:

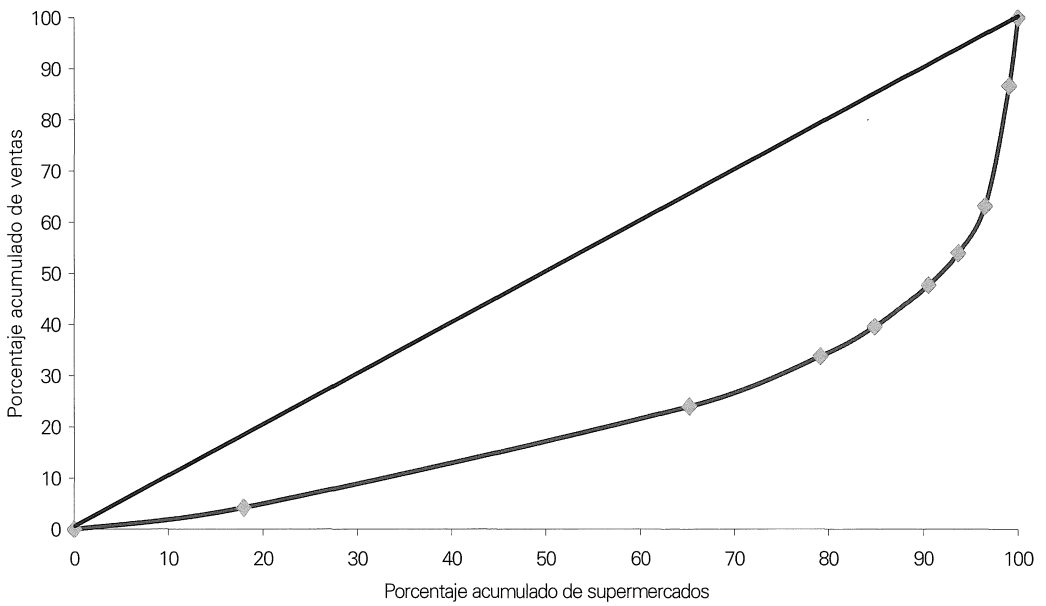


El gráfico permite apreciar las diferencias existentes entre la contribución o importancia numérica de los establecimientos en cada intervalo y su contribución a las ventas totales o importancia ponderada. Como se vio en el caso de los 10 supermercados, estas diferencias son la base de la *curva de concentración*.

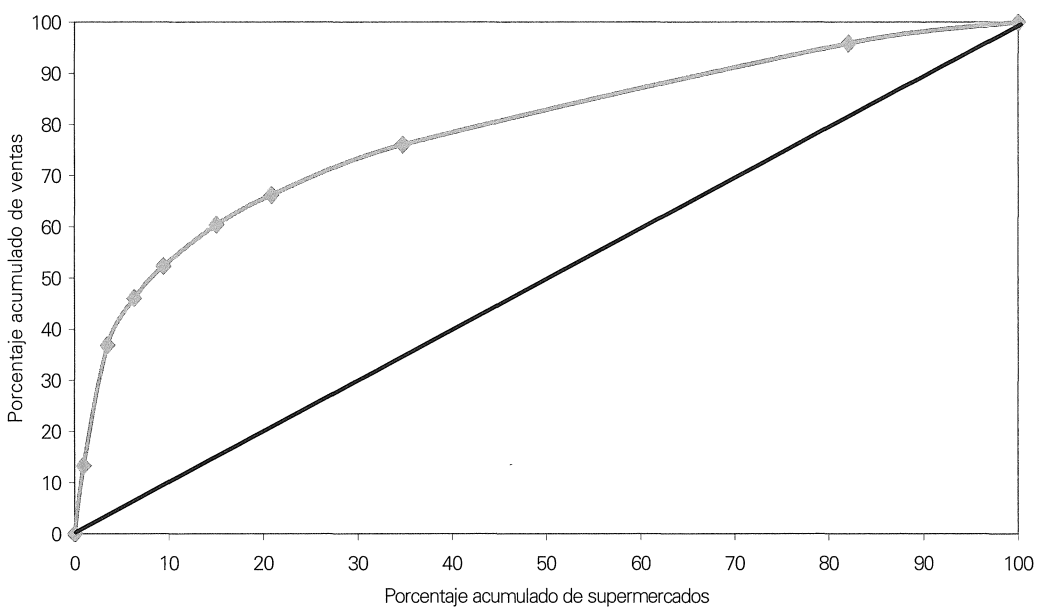
5. La cuarta columna de la tabla proporciona la frecuencia relativa acumulada, es decir, el porcentaje de observaciones (número de supermercados) con ventas menores o igual al límite correspondiente de clase, mientras que la última columna contiene el valor acumulado de la variable de estudio (ventas) en forma porcentual. Como se ha visto, ambas series de datos forman la curva de concentración. Según que la ordenación de nuestros datos sea ascendente (menor a mayor) o descendente (mayor a menor), la curva de concentración se situará por debajo o por encima de la diagonal y la lectura del gráfico será en el primer caso que el 20% de supermercados más pequeños no llega a realizar el 5% de las ventas totales, y en

el segundo caso que el 10% de supermercados más grandes realizan más del 50% de las ventas:

Curva de concentración de ventas (ordenación de menor a mayor)

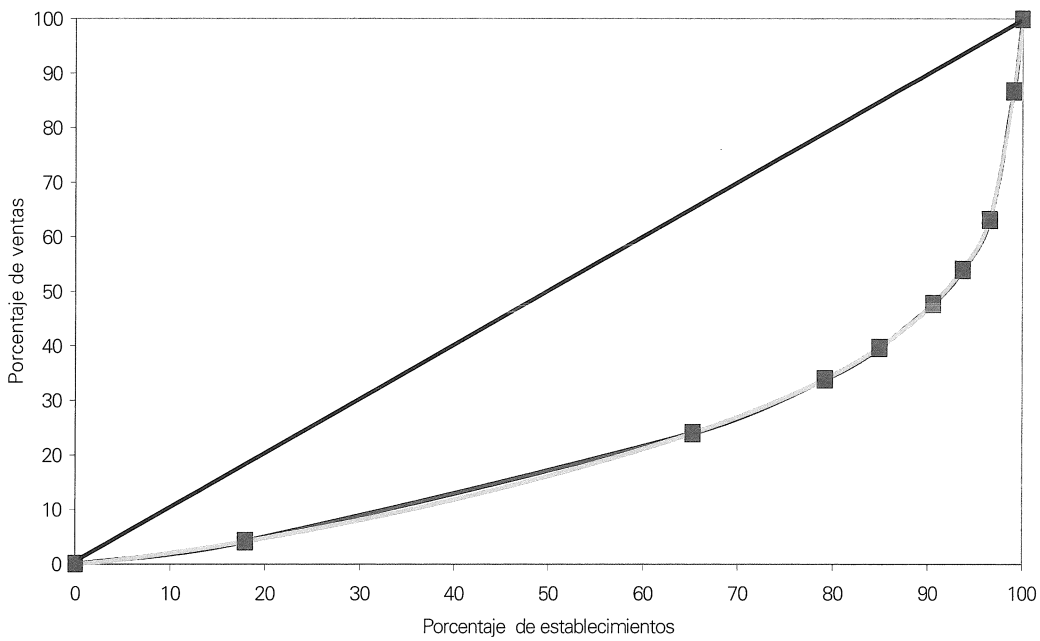


Curva de concentración de ventas (ordenación de mayor a menor)

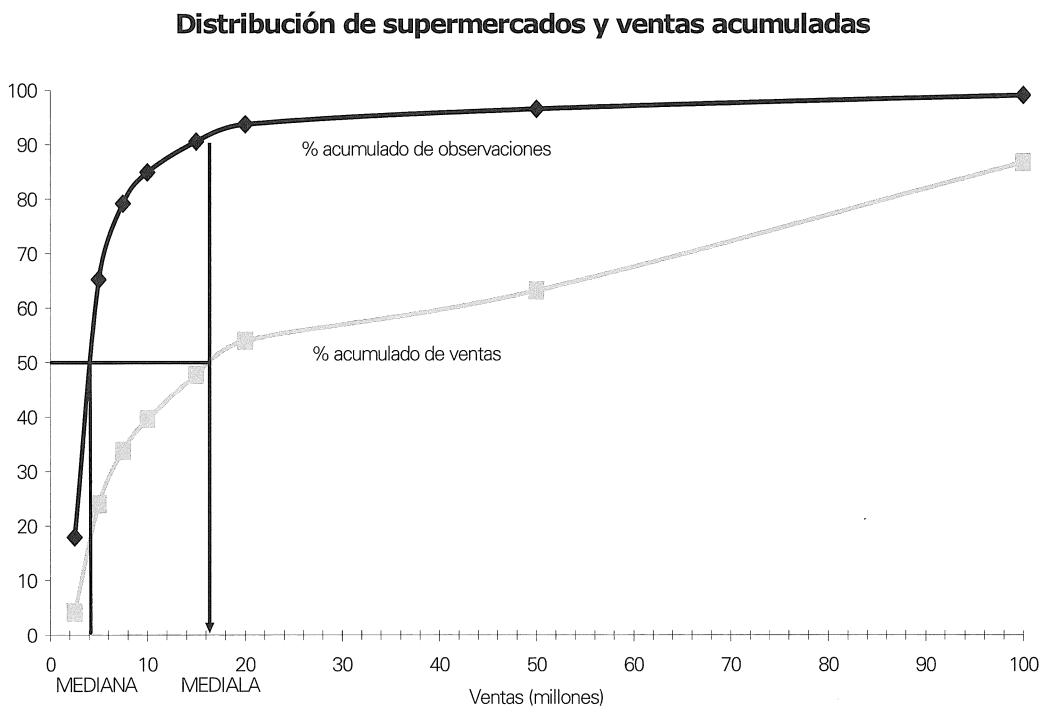


6. En el siguiente gráfico se proporciona la curva de concentración dibujada con cada una de las 2.960 observaciones, y también se han marcado los puntos con los que se ha dibujado la curva en base a la tabla con los intervalos de clase. Puede apreciarse el alto grado de aproximación que se obtiene con los intervalos de clase.

Curva de concentración de ventas con todas las observaciones



7. Habíamos definido la mediana como el valor de la variable que divide a las observaciones ordenadas en dos partes iguales en número de unidades. De forma análoga se define la *mediala* como el valor de la variable que divide a las observaciones ordenadas en dos partes, conteniendo cada una las unidades que realizan el 50% del total de la variable. En el caso de los supermercados, la mediala divide a los mismos en dos partes: los que presentan ventas inferiores a la mediala y totalizan el 50% de las ventas, y los que presentan valores superiores a la mediala y totalizan el otro 50%. En forma gráfica la mediana se obtenía de la distribución acumulativa de la frecuencia relativa y ahora la mediala se obtiene de la curva acumulativa porcentual de ventas, como se ilustra en el gráfico que sigue, en el que se aprecia que la mediala se sitúa en un valor algo superior a los 16 millones de euros. Puede verse también que los supermercados con ventas inferiores a la mediala son, en número, algo más del 90%.


Distribución de supermercados y ventas acumuladas
100
90
80
70
60
50
40
30
20
10
0
% acumulado de observaciones
% acumulado de ventas
0
10
20
30
40
50
60
70
80
90
100
MEDIANA
MEDIALA
Ventas (millones)

Capítulo 9
El concepto de probabilidad

1. Aleatoriedad y sucesos

1. La idea intuitiva de probabilidad como posibilidad de que ocurra algo es bastante común en todas las personas y en muchas ocasiones se tiene una apreciación de su valor. Cuando juegas un décimo a la lotería de Navidad sabes que la probabilidad de que te toque el premio gordo es muy pequeña, pero sabes también que puedes coger algo en los premios menores e incluso recuperar el dinero gastado si el número que juegas tiene su última cifra igual a la del primer premio. Lo mismo sucede con la primitiva, las quinielas o cualquier otro juego de azar. Al lanzar una moneda bien hecha sabemos que aproximadamente la mitad de las veces obtendremos cara. La posibilidad de obtener un as en una baraja de 40 cartas es de 4 sobre 40.

2. El concepto de probabilidad está ligado a los *fenómenos aleatorios*, de los cuales los juegos de azar son ejemplos tradicionales. Un fenómeno o experimento aleatorio se caracteriza porque podemos conocer de antemano todos los resultados posibles pero no tenemos la certeza del resultado concreto que se va a producir. Antes de un partido de fútbol entre dos equipos sabemos que ganará uno de ellos o empatarán, incluso podemos apostar una porra sobre posibles resultados en goles, pero hasta que no termina el partido no sabemos el resultado final. Cuando una tienda abre sus puertas al público no hay certeza sobre los ingresos por ventas que se van a realizar en el día, aunque por la experiencia anterior se disponga de información sobre un cierto intervalo en el que se espera que se muevan las ventas: sólo después de hacer caja se sabrá con certeza cuáles han

sido los ingresos del día. La *incertidumbre* sobre el resultado, aunque sea pequeña, es consustancial a la aleatoriedad y la probabilidad lo que hace es medir numéricamente el grado de incertidumbre.

3. Una propiedad fundamental de los fenómenos aleatorios es la *regularidad* de los resultados en series largas de observaciones. El ejemplo clásico de la moneda simétrica nos servirá para explicar su significado. Lancemos una moneda 10 veces y anotemos el número de caras que obtenemos. Repitamos la experiencia realizando ahora 15 tiradas de la moneda[1] anotando igualmente el número de caras obtenido y así sucesivamente aumentando en cada experiencia el número de lanzamientos.

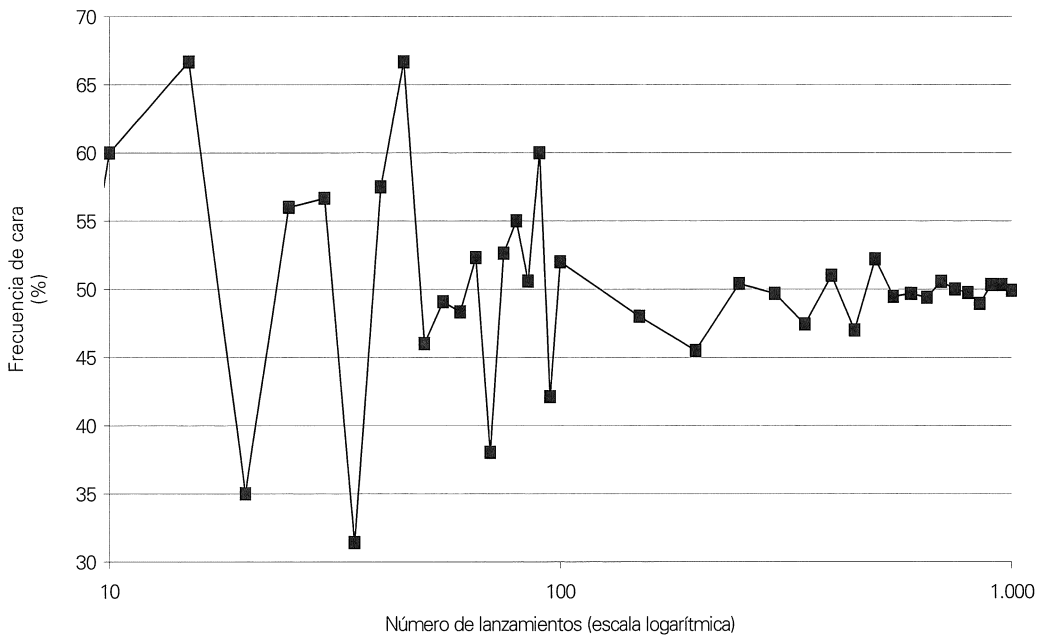
Los resultados que se obtienen serán similares a los de la siguiente tabla, reflejados también en el gráfico que sigue:

Frecuencia de cara en lanzamiento de una moneda

Nº de tiradas	Nº de caras	Frecuencia %	Nº de tiradas	Nº de caras	Frecuencia %
10	6	60,0	100	52	52,0
20	7	35,0	200	91	45,5
30	17	56,7	300	149	49,7
40	23	57,5	400	204	51,0
50	23	46,0	500	261	52,2
60	29	48,3	600	298	49,7
70	27	38,0	700	354	50,6
80	44	55,0	800	398	49,8
90	54	60,0	900	453	50,3
			1.000	499	49,9

[1] Ojo, no cinco tiradas más sino que lanzamos la moneda 15 veces de nuevo.

Frecuencia relativa de cara en lanzamientos de una moneda



La *frecuencia relativa* de obtener cara es el resultado de dividir el número de caras obtenido entre el total de lanzamientos realizados, expresado en este caso en forma porcentual, y es el mismo concepto de frecuencia relativa que hemos visto en 7.2 relativo a variables estadísticas. En el gráfico se observa claramente como el comportamiento de las series de lanzamientos es errática cuando el número de lanzamientos es pequeño, pero a medida que éste aumenta la frecuencia de cara se regulariza y tiende a estabilizarse en torno al 50%. Esta tendencia estabilizadora de la frecuencia relativa de un resultado en torno a un cierto valor cuando se repiten las observaciones en idénticas condiciones de experimentación, se denomina regularidad de las series estadísticas o *ley del azar* y constituye la base empírica del concepto de probabilidad. De hecho, a la vista del gráfico anterior cualquiera estaría de acuerdo en que la probabilidad de cara parece ser del 50%, aunque si alguien lanza la moneda 200 veces resulte que obtiene un 45,5% de caras. El 45,5% de caras obtenido en una sucesión particular de 200 lanzamientos es la frecuencia observada, mientras que el 50% de posibilidades de cara que asignamos al suceso {cara} es su probabilidad ideal o teórica. Al estudiar el muestreo volveremos sobre este asunto.

4. El conjunto de todos los resultados posibles de un experimento aleatorio se denomina *espacio muestral E*. A cada elemento (resultado posible) del espacio muestral le llamamos punto muestral o *suceso elemental.* Un *suceso S* es cualquier resultado de un experimento aleatorio. Al tirar una moneda una vez los resultados posibles son E = {cara, cruz} = {C, X}. Al lanzar un dado el espacio muestral (el conjunto de los resultados posibles) es $\mathbf{E} = \{1,2,3,4,5,6\}$. Además de los sucesos elementales podemos considerar otros como el suceso de *obtener par, de obtener mayor de 4, ...* Los sucesos elementales no se pueden descomponer en otros más sencillos. Al lanzar un dado el suceso *obtener par* puede representarse como $\mathbf{par} = \{2,4,6\}$, es decir, sucede si obtenemos un 2, un 4 o un 6. El suceso queda descompuesto en otros más elementales, es decir, es un *suceso compuesto*. Sin embargo el suceso *obtener 2* no podemos descomponerlo en otros más simples: o sucede o no sucede. Es un suceso elemental o punto muestral.

2. Probabilidad

1. La *probabilidad* de un suceso elemental es un número comprendido entre 0 y 1 que mide la posibilidad o verosimilitud de su ocurrencia, y tal que la suma de probabilidades de todos los sucesos elementales es igual a 1. Una probabilidad cercana a 0 corresponde a un resultado muy poco verosímil (la probabilidad 0 le correspondería a un suceso imposible de suceder), mientras que si la probabilidad es cercana a la unidad el resultado es muy posible que ocurra (la probabilidad 1 sería para un suceso seguro de ocurrir). Es frecuente expresar la probabilidad en forma porcentual. Al lanzar una moneda bien hecha la probabilidad de cara es 1/2, que también se expresa como del 50%. En el lanzamiento de un dado la probabilidad de cada uno de los 6 resultados elementales posibles es 1/6. La probabilidad de obtener el 7 de oros en una baraja de 40 cartas es de 1/40.

2. La probabilidad de un suceso cualquiera está dada por la suma de probabilidades de los sucesos elementales que lo definen. En el lanzamiento de un dado la probabilidad de obtener $\mathbf{par} = \{2,4,6\}$ será la probabilidad de obtener 2, más la de obtener 4, más la de obtener 6, esto es 3/6 = 1/2. La probabilidad de obtener una carta de oros (hay 10 oros en 40 cartas) será de 10/40 = 1/4. Resulta pues que conocida la probabilidad de los sucesos elementales, podemos conocer también la probabilidad de cualquier otro suceso. Como ya se ha dicho, la suma de probabilida-

des de todos los sucesos elementales es la unidad, y esto es así porque es seguro que uno de ellos ocurrirá.

3. En la práctica la asignación de probabilidades depende de la información disponible sobre el fenómeno considerado. En el lanzamiento de una moneda resulta casi intuitivo asignar 50% de probabilidad a cada uno de los dos posibles resultados y además puede comprobarse experimentalmente, como se ha visto anteriormente. Lo mismo sucede con otros casos relacionados con juegos de azar como lanzar un dado, juegos de cartas, loterías, ..., dónde los sucesos elementales tienen igual probabilidad. En caso de equiprobabilidad de sucesos elementales la probabilidad de un suceso coincide con la relación entre el número de sucesos elementales que lo definen y el número total de sucesos elementales posibles. Así en el caso de la baraja que hemos visto, tenemos 40 cartas posibles y 10 cartas favorables de oros, de forma que la probabilidad de oros es 1/4. La relación de casos favorables a casos posibles para determinar la probabilidad de un suceso se conoce como *regla de Laplace*, pero sólo es válida en caso de *equiprobabilidad* de sucesos elementales.

4. Si no hay equiprobabilidad de sucesos elementales, siempre es válida la regla de que la probabilidad de un suceso es la suma de probabilidades de los sucesos elementales que lo definen. Como se ha dicho, la asignación de probabilidades va a depender de la información disponible. Ante un partido de fútbol entre los equipos A y B se puede dar igual probabilidad de ganar a ambos si no tenemos ninguna otra información. Ahora bien, si sabemos que el equipo A es el primero de la liga, que el equipo B es uno de los últimos, que el partido se juega en el campo del A, el cuál ha ganado los últimos 10 partidos que ha jugado en su campo, mientras que el equipo B tan solo ha ganado 1 y empatado 2 de los últimos 10 partidos que ha jugado fuera de su campo, entonces nuestra asignación inicial de dar a ambos igual probabilidad de ganar cambia totalmente y nos decantamos claramente por la victoria de A, aunque no descartemos totalmente la posibilidad de que el equipo B pueda empatar e incluso ganar. Sin embargo, si diferentes personas, con la información anterior, tienen que materializar numéricamente la probabilidad de victoria del equipo A, seguramente nos encontraremos con diferentes valores.

5. La probabilidad es un concepto ideal y teórico, en muchas ocasiones imposible de cuantificar con exactitud matemática. En las encuestas suelen calcularse tasas y porcentajes que son estimaciones o aproximaciones de probabilidades. La

audiencia televisiva de un cierto partido de fútbol, es decir, el porcentaje de personas que lo han visto por TV, es una aproximación a la probabilidad de que una persona seleccionada al azar haya visto o no el partido. La tasa de paro que refleja la EPA es una aproximación a la probabilidad de que una persona activa se encuentre o no en paro. Las encuestas electorales aproximan la probabilidad de voto de los partidos políticos.

3. Independencia

1. Consideremos una urna con 7 bolas blancas y 3 negras. Si sacamos una bola la probabilidad de que sea blanca es de 7/10. Si ahora devolvemos la bola extraída a la urna y sacamos una segunda bola, la probabilidad de que esta segunda bola sea negra es de 3/10, ya que seguimos teniendo en la urna 7 blancas y 3 negras. Ahora bien, si la primera bola extraída no la devolvemos a la urna, nos quedan 9 bolas y la probabilidad de obtener negra en la segunda extracción va a depender del resultado de la primera bola sacada: si la primera bola ha sido blanca, quedan 6 blancas y 3 negras y la probabilidad de negra a la segunda es de 3/9; pero si la primera bola fue negra, quedan 7 blancas y 2 negras y ahora la probabilidad de negra será de 2/9. Al devolver la primera bola a la urna el resultado de la segunda extracción no está condicionado por la primera bola sacada y ambas extracciones son independientes. Es decir, *dos sucesos son independientes si la ocurrencia de uno no influye sobre la probabilidad de que ocurra el otro.*

2. En el lanzamiento de una moneda dos veces el resultado del segundo lanzamiento es independiente del que hayamos obtenido la primera vez, y lo mismo sucede con un tercer o cuarto lanzamiento, el resultado de cada uno no está influenciado por los anteriores y son por tanto independientes. El experimento de lanzar una moneda tres veces consecutivas es equivalente al de lanzar tres monedas una sola vez, por la independencia de sucesos.

Capítulo 10
Variables aleatorias

1. Concepto

1. Consideremos el número de caras que se pueden obtener al lanzar una moneda tres veces (o también al lanzar tres monedas una vez). Los posibles resultados y la probabilidad de cada uno se detallan a continuación:

Resultado	Número de caras	Probabilidad %
XXX	0	1/8=12,5
XXC	1	1/8=12,5
XCX	1	1/8=12,5
CXX	1	1/8=12,5
CCX	2	1/8=12,5
CXC	2	1/8=12,5
XCC	2	1/8=12,5
CCC	3	1/8=12,5

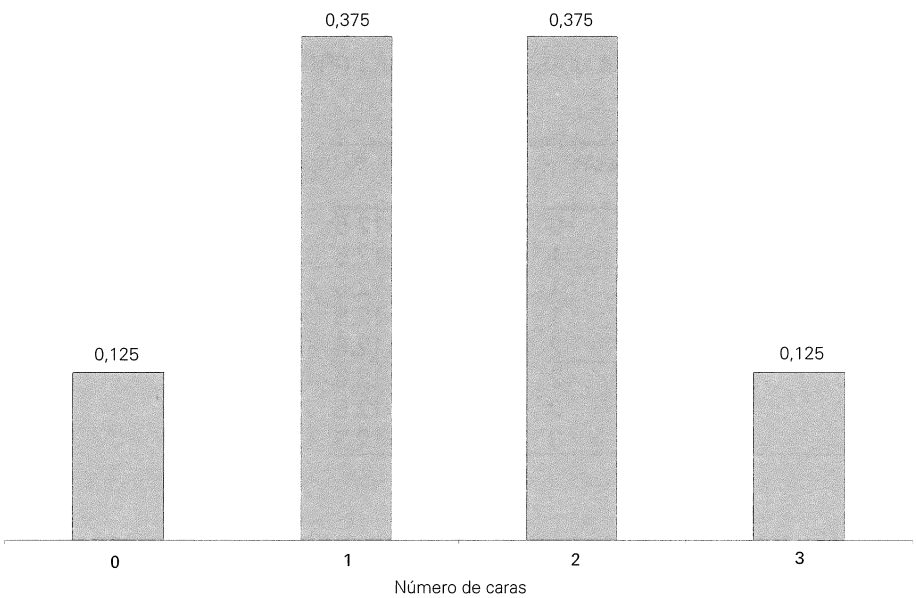
Como sólo estamos interesados en el número de caras podemos resumir los resultados en la tabla siguiente:

Número de caras	Probabilidad %
0	1/8=12,5
1	3/8=37,5
2	3/8=37,5
3	1/8=12,5

Estamos ante una *variable aleatoria*, el número de caras que se obtienen al lanzar una moneda tres veces, es decir, los valores numéricos de la variable están asociados al resultado de un experimento aleatorio. La probabilidad de cada resultado posible induce la probabilidad de que la variable aleatoria tome un valor cualquiera. Así, la probabilidad del valor 1 corresponde a la probabilidad de obtener cualquiera de los resultados XXC, XCX o CXX, esto es P(1) = P(XXC) + P(XCX) + P(CXX) = 1/8 + 1/8 + 1/8 = 3/8 = 0,375 = 37,5%.

2. El conjunto de posibles valores de una variable aleatoria junto con la probabilidad de cada valor, define una *distribución de probabilidad*. La tabla anterior refleja la distribución de probabilidad del número de caras que se obtienen al lanzar una moneda tres veces y que, gráficamente, presenta el siguiente aspecto:

Distribución de probabilidad del número de caras al lanzar una moneda tres veces



3. Igual que la suma de probabilidades de los sucesos elementales es la unidad, también la suma de probabilidades de todos los posibles valores de una variable aleatoria es la unidad.

2. Media y varianza de una variable aleatoria

1. Si nos fijamos en la distribución de probabilidad correspondiente al número de caras que se obtienen al lanzar una moneda tres veces,

Número de caras	Probabilidad
0	1/8
1	3/8
2	3/8
3	1/8

nos debe recordar a las distribuciones de frecuencias vistas en el capítulo anterior: estamos ante una variable que toma diferentes valores, pero ahora en lugar de frecuencias observadas tenemos probabilidades.

Podemos calcular la media, varianza y otras características (mediana, cuartiles, ...) de una variable aleatoria en forma análoga a la que hemos visto para una variable estadística, sustituyendo el papel de la frecuencia por la probabilidad como forma de ponderar cada valor de la variable. Así, para la variable aleatoria que estamos considerando, número de caras en tres lanzamientos, el valor medio es

$$0 \cdot \frac{1}{8} + 1 \cdot \frac{3}{8} + 2 \cdot \frac{3}{8} + 3 \cdot \frac{1}{8} = 1{,}5 \text{ caras}$$

y análogamente se procedería para calcular la varianza y la desviación típica o estándar.

2. A diferencia de las variables estadísticas en las que el valor medio y otras características son valores calculados con datos observados, en el caso de una variable aleatoria su valor medio es ideal o teórico en la misma medida que lo es la probabilidad asociada a cada valor de la variable aleatoria. El número medio de caras al lanzar tres monedas es el promedio que se espera obtener cuando se hace un número infinito de lanzamientos. Esta idea de media como valor esperado hace que a la media de una variable aleatoria se la llame también *esperanza matemática*. La variable aleatoria más sencilla es la que corresponde al lanzamiento de una moneda una vez.

Número de caras	Probabilidad
0	1/2
1	1/2

Su valor medio es 0*1/2 + 1*1/2 = 0,5 y el gráfico de 9.1.3, que sirvió para ilustrar la idea de probabilidad, refleja también el concepto de esperanza matemática de una variable aleatoria como el promedio esperado en un número infinito de repeticiones.

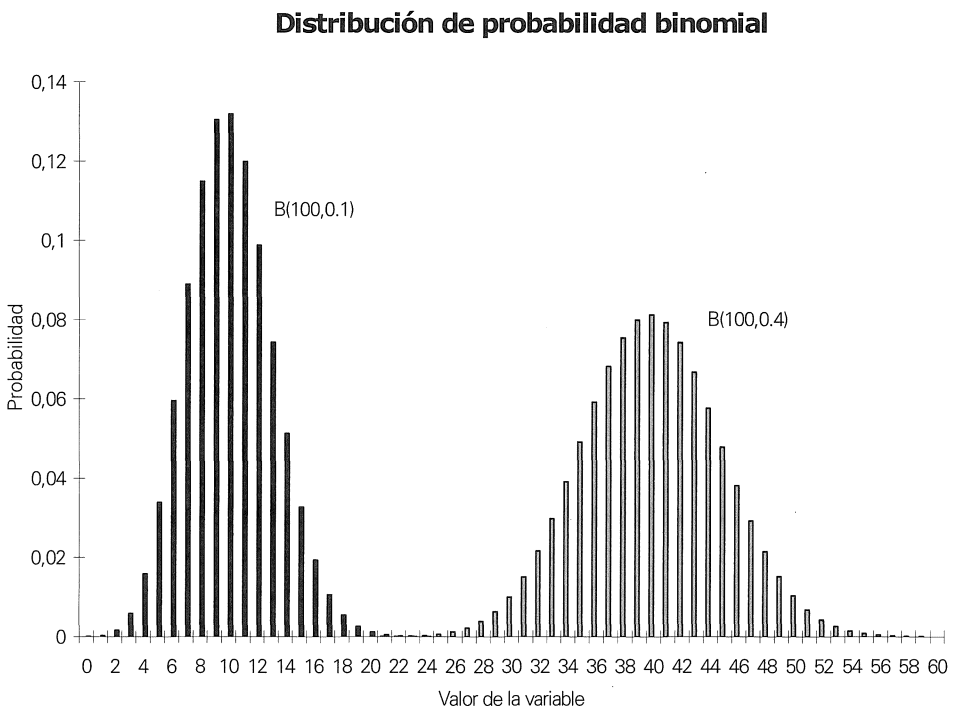
3. Algunas distribuciones de probabilidad

1. Al lanzar una moneda hay dos resultados posibles: cara y cruz. A pruebas aleatorias que dan lugar a dos posibles resultados se las denomina *pruebas éxito-fracaso* o pruebas SI-NO y son el origen de varias distribuciones de probabilidad. La más sencilla es la asociada a una sola prueba, como la que acabamos de ver en el lanzamiento de una moneda. Si llamamos *p* a la probabilidad de éxito (1/2 en el caso de la moneda), la distribución de probabilidad asociada a la variable aleatoria que da el número de éxitos en una prueba es

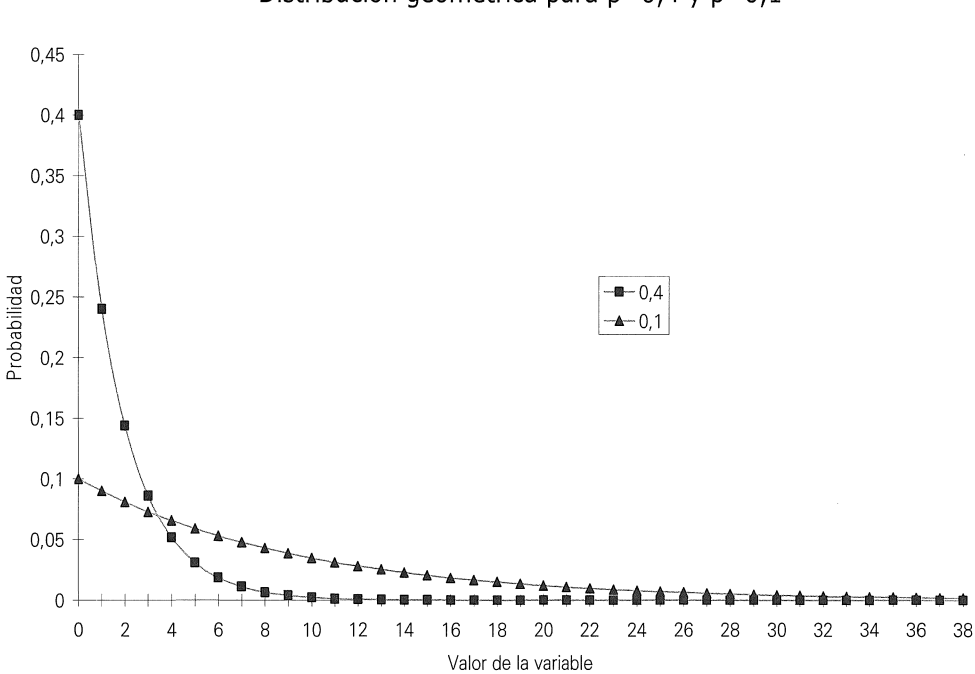
Número de éxitos	Probabilidad
0	1-p = q
1	p

cuyo valor medio es p y su varianza pq. Se denomina *distribución de Bernouilli*. Se corresponde con variables estadísticas dicotómicas en las que las unidades se clasifican según pertenezcan o no a una cierta clase, asignándose el valor 1 si pertenecen y el valor 0 si no pertenecen.

2. Si ahora consideramos la realización de n pruebas éxito – fracaso independientes como sería el lanzamiento de n monedas (p =1/2) o el lanzamiento sucesivo n veces de un dado, considerando la obtención de 6 como éxito (*p* = 1/6), obtenemos la *distribución binomial*, cuyo valor medio es np y la varianza npq. En el gráfico que sigue se presenta la distribución de probabilidad de dos binomiales con n=100 y p=0,1 en un caso y n =100, p = 0,4 en el otro.

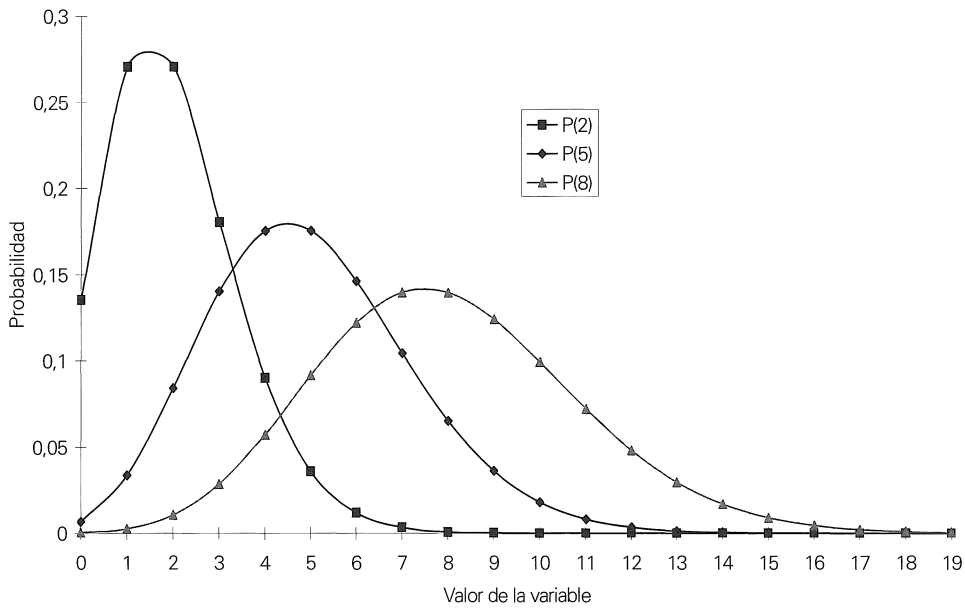


3. Asociada también a las pruebas éxito-fracaso tenemos la *distribución de probabilidad geométrica* que nos proporciona el número de fallos antes de obtener el primer éxito en pruebas sucesivas independientes (número de cruces obtenidas antes de la primera cara), con media q/p y varianza q/p^2.



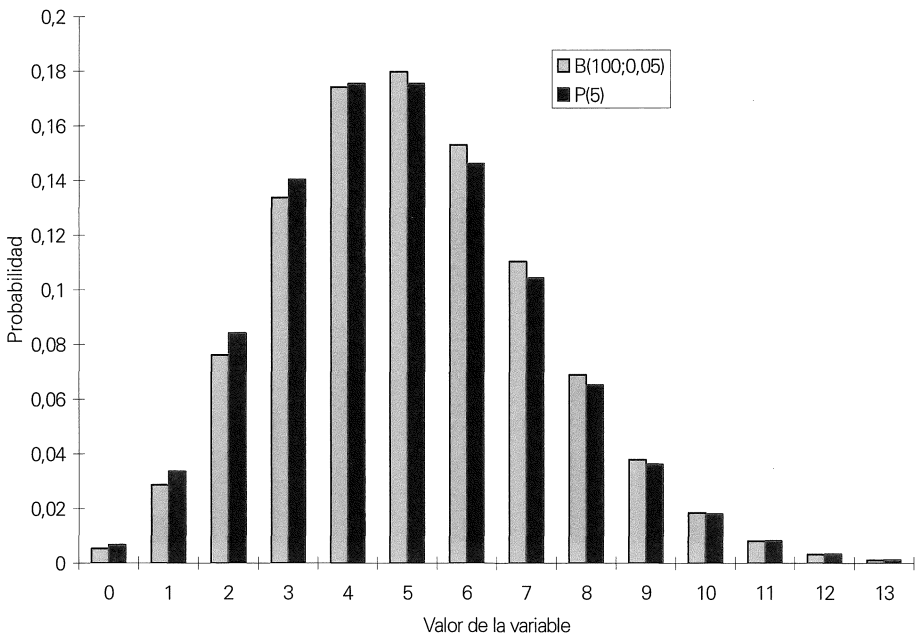
4. La *distribución de Poisson* se da en situaciones en que tenemos un gran número de experimentos independientes y una pequeña probabilidad de éxito en cada uno de ellos, preguntándonos cuál es la probabilidad de obtener x éxitos en n experimentos, siendo n muy grande. Por ello, a veces, se la llama ley de los sucesos raros. Tiene la particularidad de que la media y la varianza son iguales. Su distribución de probabilidad tiene el siguiente aspecto gráfico para valores de la media iguales a 2, 5 y 8:

Distribución de probabilidad de Poisson para distintos valores de la media

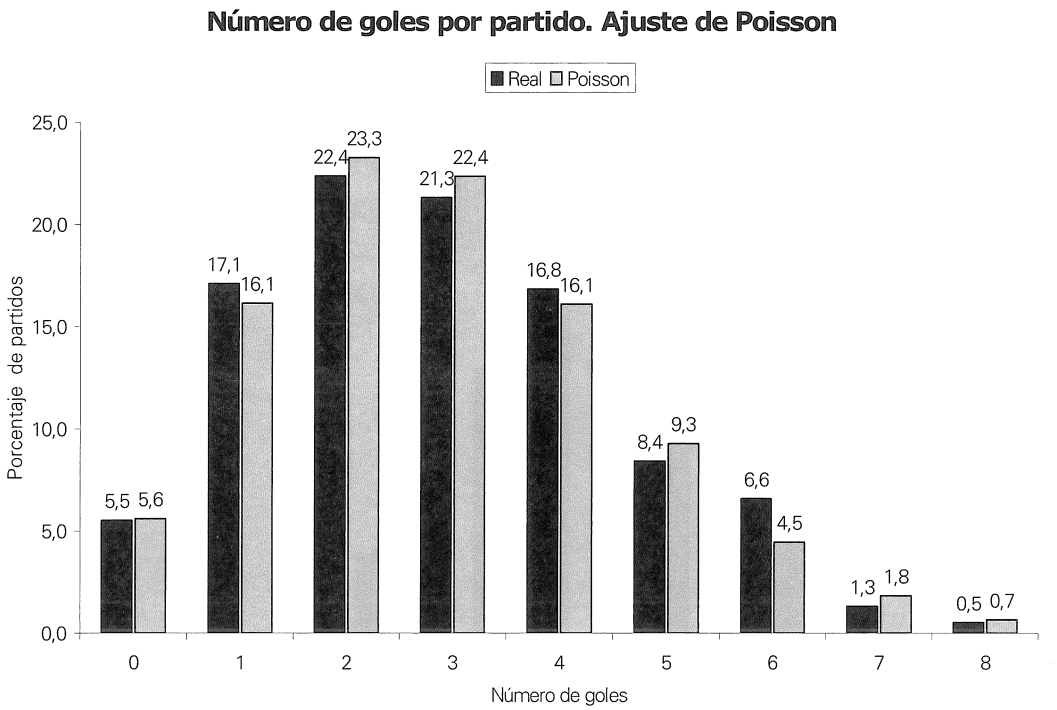


5. La distribución de Poisson corresponde al límite al que tiende la distribución binomial cuando n es grande y p pequeño como refleja el siguiente gráfico en el que se detalla la distribución de probabilidad binomial con n=100 y p=0,05 y una distribución de Poisson con media np=5:

Distribución de Poisson como límite de la Binomial



6. Las variables aleatorias y sus distribuciones de probabilidad asociadas son modelos teóricos que sirven para describir multitud de situaciones reales como ventas semanales de un cierto artículo en un supermercado, número de artículos defectuosos en un lote, número de vehículos que pasan por un punto determinado en una hora, distribución de precios de un artículo, ... En 8.2.1 habíamos considerado la distribución de frecuencias de goles marcados en cada uno de los 380 partidos de fútbol de la primera división de la liga española de la temporada 2000-2001. La media, 2,88 goles por partido y la varianza, 2,83 son muy similares y es un primer indicio de que la distribución de Poisson podría describir satisfactoriamente la distribución de frecuencias de goles por partido. Si calculamos las probabilidades teóricas de una distribución de Poisson de media 2,88 y comparamos con las frecuencias reales obtenidas obtenemos el siguiente gráfico:



El gráfico refleja como la distribución teórica de Poisson explica satisfactoriamente el número de goles por partido. Ello se debe a que la distribución de Poisson se produce con sucesos que ocurren aleatoriamente en el tiempo o en el espacio con baja probabilidad de éxito en intervalos cortos o espacios pequeños, como sucede con la probabilidad de gol en un intervalo pequeño de tiempo.

7. Un ejemplo clásico de aplicación de la ley de Poisson lo da la distribución de los impactos de bombas volantes en el área sur de Londres durante la segunda guerra mundial. El área total se dividió en pequeñas áreas y se contabilizó el número de impactos en cada área pequeña. La distribución del número de impactos por área mostró un buen ajuste de la distribución de Poisson, lo que indicaba que las bombas caían al azar, a pesar de que hubiera algunas áreas con 4 y hasta 5 impactos. Podemos considerar los lanzamientos de bombas como experimentos aleatorios independientes. Al dividir todo el área en que caen las bombas en sectores relativamente pequeños, la probabilidad de que una bomba alcance un determinado sector es pequeña.

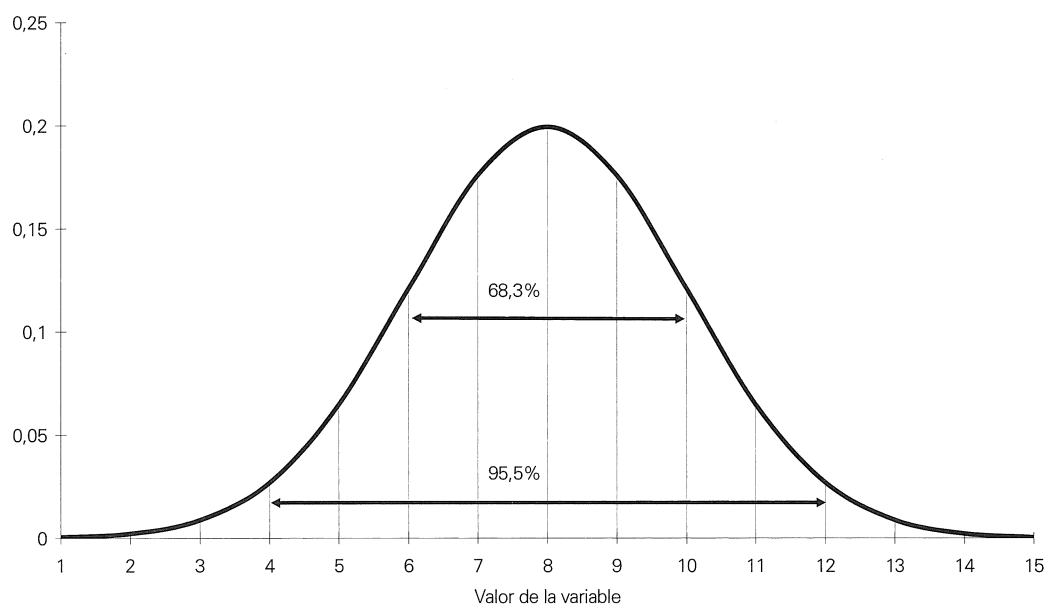
4. La distribución normal

1. Las variables aleatorias que hemos visto pueden tomar sólo los valores 0, 1, 2, ..., y son, por tanto, *variables discretas*. Pero al igual que sucede con las varia-

bles estadísticas existen variables aleatorias que pueden tomar cualquier valor dentro de un intervalo dado: son *variables aleatorias continuas*. Mientras que para una variable discreta cada valor posible de la misma tiene asociado una probabilidad concreta de que tome ese valor, en el caso continuo la variable aleatoria puede tomar los infinitos valores que existen en su intervalo de definición, resultando que la probabilidad de un valor particular es cero. Si hablamos del peso de todas las personas residentes en España, la probabilidad de encontrar una persona que pese exactamente 61,345768934123 kg será prácticamente nula; sin embargo encontraremos un porcentaje apreciable de personas cuyo peso esté comprendido entre 60 y 65 kg. Es decir, en el caso de variables continuas en lugar de hablar de función de probabilidad, hablamos de *densidad de probabilidad*: la masa unitaria de probabilidad se distribuye en el intervalo de definición de la variable de forma que en unas zonas la densidad de probabilidad es mayor que en otras. En el caso del peso de personas, encontraremos un mayor porcentaje de personas (mayor densidad de probabilidad) con peso entre 60 y 65 kg que entre 130 y 135 kg.

2. En la práctica, en un gran número de aplicaciones se encuentran distribuciones que se aproximan a la llamada *distribución normal de probabilidad*, como es el caso de errores de medida de magnitudes físicas y astronómicas y de un gran número de distribuciones demográficas y biológicas. La distribución normal queda definida por su valor medio y su desviación típica. El aspecto de su densidad de probabilidad es el del siguiente gráfico, que refleja una distribución normal de media 8 y desviación típica 2 (coeficiente de variación = 2/8 = 25%):

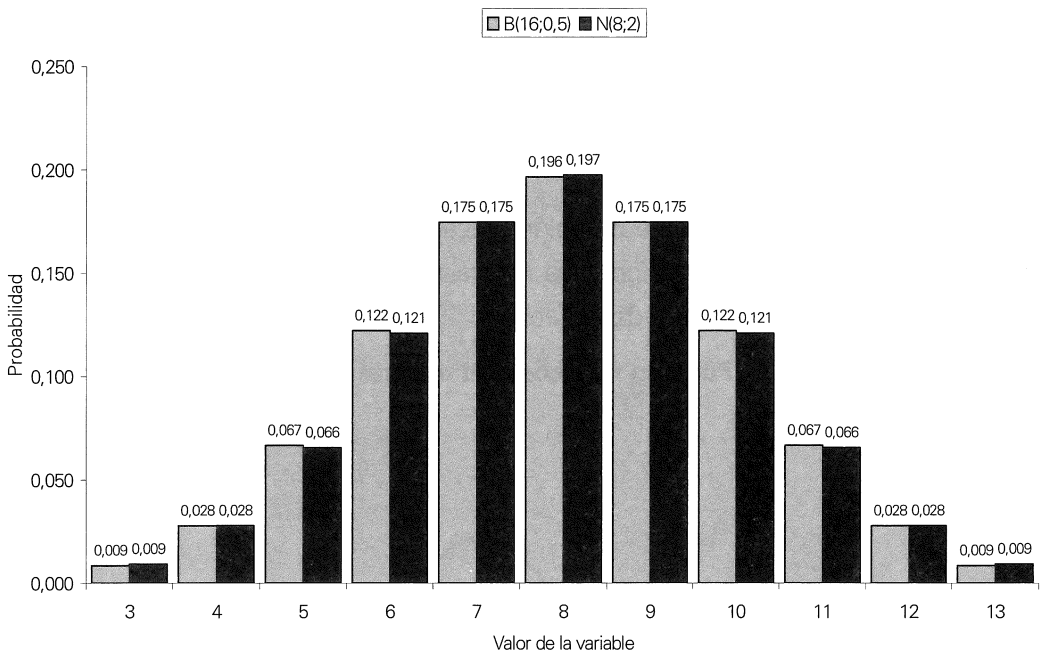
Función de densidad normal N(8;2)



3. Como puede verse, la mayor parte de la probabilidad se concentra en valores relativamente cercanos a la media. En concreto, en cualquier distribución normal el 68,3% de probabilidad está comprendida entre la media y más, menos una vez la desviación típica, mientras que entre la media y más, menos dos veces la desviación típica se encuentra el 95,5% de la probabilidad o de las observaciones. La probabilidad se distribuye en forma simétrica alrededor de la media.

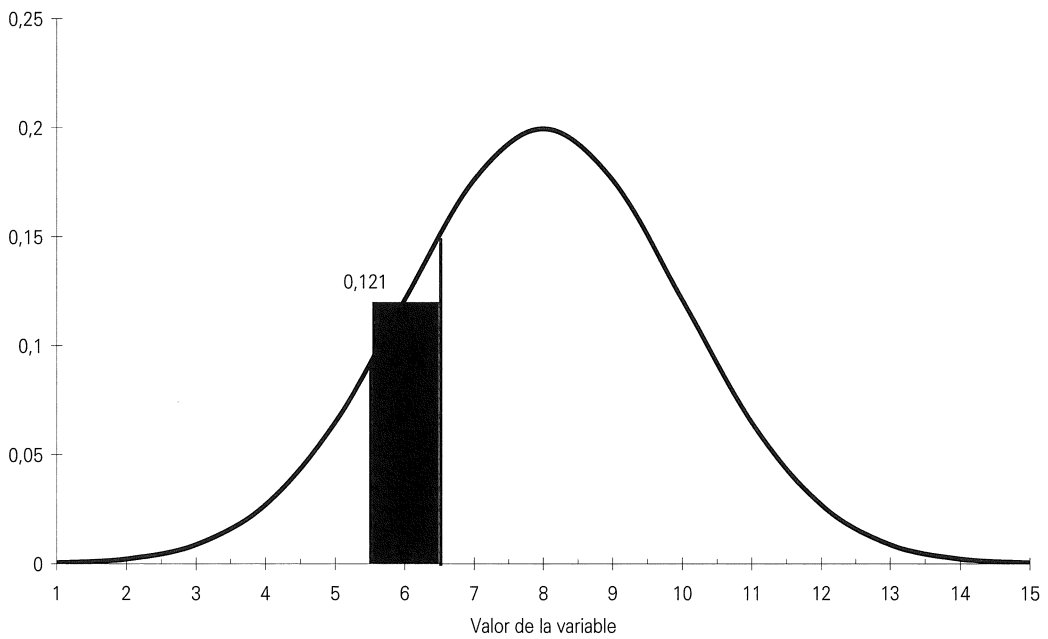
4. La gran importancia teórica de la distribución normal radica en que, bajo ciertas condiciones, son muchas las distribuciones que tienden hacia la normal. En la figura que sigue puede verse la aproximación entre una distribución binomial con n=16 y p=0,5 (media = 8, desviación típica = 2) y la normal de media 8 y desviación típica 2.

Aproximación normal de una distribución binomial n=16, p=0,5



En el caso de la normal la probabilidad del valor 6 se aproxima por el área bajo la curva normal entre 5,5 y 6,5, es decir, la probabilidad comprendida entre 5,5 y 6,5 como se refleja en el gráfico siguiente, y análogamente se aproximan el resto de probabilidades de cada valor.

Aproximación normal a la distribución binomial



5. La distribución normal juega un papel clave en la teoría de muestras y en el estudio de errores de muestreo como veremos en el siguiente capítulo.

Capítulo 11
Muestreo probabilístico y muestreo aleatorio simple

1. Introducción

1. Vimos en el capitulo 2 que una muestra es una parte de una población y que el objetivo que se tiene al estudiar una muestra es el conocimiento de características o valores de todo el conjunto poblacional del que se ha tomado la muestra, es decir, nuestro interés es el conocimiento de valores poblacionales. Un *valor poblacional* es una expresión que sintetiza los valores de la variable de estudio en las N unidades de la población completa. Así, si llamamos *x* a la variable, podemos estar interesados en conocer su total $X = \sum_{1}^{N} x$ o su promedio $\overline{X} = \frac{\sum_{1}^{N} x}{N}$ en toda la población, y también en conocer sus valores en distintas partes de la población. Si consideramos el sueldo anual de los empleados del INE, además de conocer el sueldo medio por empleado, podemos estar interesados en conocer el sueldo medio distinguiendo entre personal funcionario y laboral y dentro del personal funcionario podemos considerar el sueldo medio por grupo profesional; también podemos interesarnos por la proporción de funcionarios del grupo A respecto al total de trabajadores y el porcentaje que suponen sus sueldos respecto al total de remuneraciones del INE.

2. Un valor poblacional es un valor constante que depende sólo de los N valores x. De una población de tamaño N pueden obtenerse diferentes muestras de tamaño n. Si N es grande, por ejemplo personas residentes en España, y n es relativa-

mente pequeño, por ejemplo una muestra de 10.000 personas, el número de posibles muestras es prácticamente infinito. Cuando tomamos una muestra de n unidades para conocer un valor poblacional lo que obtenemos es una *estimación* del mismo que, dependiendo de qué elementos entren en la muestra, podrá tomar valores diferentes. Es decir, el valor estimado es un valor obtenido con los datos proporcionados por las unidades de la muestra y como tal es un valor único para cada muestra, pero depende de las n unidades concretas que se hayan seleccionado, varía de muestra a muestra: *muestras diferentes proporcionan estimaciones diferentes.*

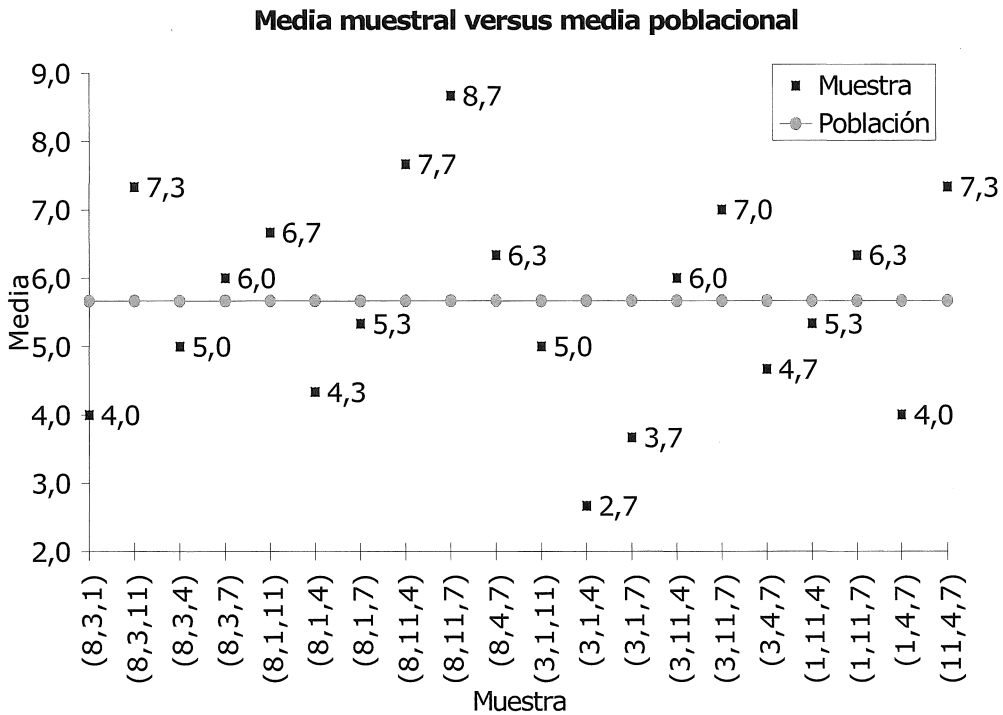
3. Si a cada muestra del conjunto de posibles muestras de tamaño n o *espacio muestral*, le asignamos una probabilidad no nula de ser seleccionada, de forma que la suma de probabilidades de todas las posibles muestras sea la unidad, tenemos un *muestreo probabilístico*. Al asignar probabilidades de selección a las posibles muestras, estamos convirtiendo el muestreo en un experimento aleatorio en el que cada muestra (y su estimación) es un suceso elemental de un espacio muestral, con una probabilidad conocida de suceder.

4. En la práctica es suficiente con asignar probabilidades de selección a las N unidades de la población. Es decir, en el muestreo probabilístico cada unidad de la población tiene una probabilidad conocida y no nula de ser seleccionada. Cuando todas las unidades de la población tienen la misma probabilidad de ser seleccionadas, también todas las posibles muestras son equiprobables y se habla de *muestreo aleatorio simple*. En este caso, siendo n el tamaño de la muestra y N el de la población, la probabilidad de seleccionar una unidad cualquiera de la población es $\frac{n}{N}$.

2. Variabilidad de muestreo. Error estándar

1. Trataremos de aclarar los conceptos anteriores con un ejemplo. Sea una población de N=6 elementos en los que la variable x, objeto de estudio, toma los valores $x = \{8, 3, 1, 11, 4, 7\}$. La media poblacional es $\overline{X} = \frac{8+3+1+11+4+7}{6} = 5{,}7$, con una desviación típica de $\sigma = \sqrt{\frac{(8-5{,}7)^2 + \cdots + (7-5{,}7)^2}{6}} = 3{,}35$ y un coeficiente de variación $CV = \frac{3{,}35}{5{,}7} = 59{,}1\%$. En una muestra aleatoria simple, la media mues-

tral es un estimador de la media poblacional, así, si nuestra muestra, de tamaño 3, estuviera formada por los valores (3,11,4) la media muestral sería $\bar{x} = 6{,}0$. Seleccionemos todas las muestras posibles de tamaño 3 calculando para cada una la media muestral. Los resultados se muestran en el siguiente gráfico:



2. Sobre el eje de abscisas se señalan los componentes de cada una de las posibles 20 muestras aleatorias de tamaño 3, todas equiprobables, es decir la probabilidad de tomar una muestra cualquiera es 1/20. En el eje de ordenadas se señala para cada una de las muestras la media muestral correspondiente. También se indica la media poblacional que es constante e igual a 5,7, de acuerdo al cálculo anterior.

3. El gráfico refleja cómo el valor poblacional (la media) es una constante pero su estimador (la media muestral) presenta valores diferentes según las unidades que componen la muestra, es decir, la estimación, varía de muestra a muestra. Puede observarse también como las distintas estimaciones se sitúan alrededor del verdadero valor que se quiere estimar.

4. Puesto que cada muestra en el ejemplo tiene una probabilidad de 1/20 de ser seleccionada, cada uno de los 20 valores muestrales tiene también una probabili-

dad de 1/20 de ser obtenido, es decir, denotando por $\bar{x}$ la media muestral resulta $P(\bar{x}=2{,}7)=P(\bar{x}=3{,}7)=\cdots=P(\bar{x}=8{,}7)=1/20$. Este conjunto de posibles valores del estimador junto con la probabilidad de obtener cada valor constituye la distribución en el muestreo del estimador. En base a esta distribución puede calcularse la probabilidad de que el estimador tome valores en un cierto intervalo; así, el intervalo (4,5; 6,5) comprende 9 de las 20 muestras. Es decir, la probabilidad de que la media muestral tome valores comprendidos entre 4,5 y 6,5 es de 9/20.

5. Siendo el estimador una variable aleatoria pueden estudiarse distintas características del mismo, como son su media o esperanza matemática, la varianza y su raíz cuadrada o desviación típica, y el coeficiente de variación, esto es, el cociente entre la desviación típica del estimador y su esperanza matemática. En particular, la desviación típica del estimador se llama *error de muestreo o error estándar.*

6. Sobre el ejemplo anterior fácilmente podemos comprobar que el promedio de las 20 estimaciones es 5,7 que coincide con la media poblacional. Esto no es casualidad, es debido a que en el muestreo aleatorio de unidades elementales la media muestral es un *estimador insesgado* de la media poblacional, es decir, la esperanza matemática del estimador coincide con el verdadero valor que se quiere estimar. En caso contrario el estimador se dice *sesgado* y a la diferencia entre la esperanza matemática o valor medio del estimador y el valor a estimar se le llama *sesgo.*

7. Calculemos a continuación la desviación típica del estimador en nuestro ejemplo. Recordemos que dado un conjunto de valores $x_1, x_2, \cdots, x_n$, la desviación típica se define como la raíz cuadrada de la varianza, es decir

$$\sigma = \sqrt{\frac{\sum_{1}^{n}(x_i - \bar{x})^2}{n}}$$

donde $\bar{x} = \frac{\sum x_i}{n}$ es el valor medio. En nuestro caso x_i son las 20 estimaciones del gráfico y $\bar{x}$ es su valor medio por lo que

$$\sigma = \sqrt{\frac{(2{,}7-5{,}7)^2+(3{,}7-5{,}7)^2+\cdots+(8{,}7-5{,}7)^2}{20}} = 1{,}5$$

Así pues, el error de muestreo en el ejemplo es 1,5 y nos da una medida de la variabilidad de las estimaciones individuales alrededor de su media. El coeficiente de variación de las estimaciones sería

$$CV = \frac{1,5}{5,7} = 0,264 \rightarrow 26,4\%$$

El coeficiente de variación del estimador se denomina *error de muestreo relativo*. Veremos posteriormente que no es necesario tomar todas las posibles muestras para calcular el error de muestreo, lo cuál en la práctica sería irrealizable. Los propios datos muestrales permiten obtener una estimación de la magnitud del error de muestreo.

3. El papel de la distribución normal en el muestreo probabilístico

1. En el muestreo probabilístico cada posible muestra y, por tanto, cada estimación, tiene una probabilidad conocida de ser seleccionada. Esto significa que estamos ante un conjunto de posibles valores estimados, cada uno con una cierta probabilidad de ser obtenido, es decir, estamos ante una distribución de probabilidad: la *distribución de muestreo*. Si conociéramos esta distribución podríamos determinar la probabilidad de obtener cada posible valor del estimador o, lo que es más interesante, la probabilidad de que el valor estimado en una muestra concreta esté comprendido en un cierto intervalo. En la práctica del muestreo no tenemos a nuestra disposición todas las muestras posibles, sino que estamos con una muestra particular, es decir, con una estimación concreta de entre todas las posibles y lo que nos interesa es el grado de aproximación de esa estimación al verdadero valor poblacional. No olvidemos que el único objetivo de una muestra es conocer características de la población de la que se ha extraído por lo que necesitamos saber también hasta que punto nuestros datos estimados reflejan la realidad.

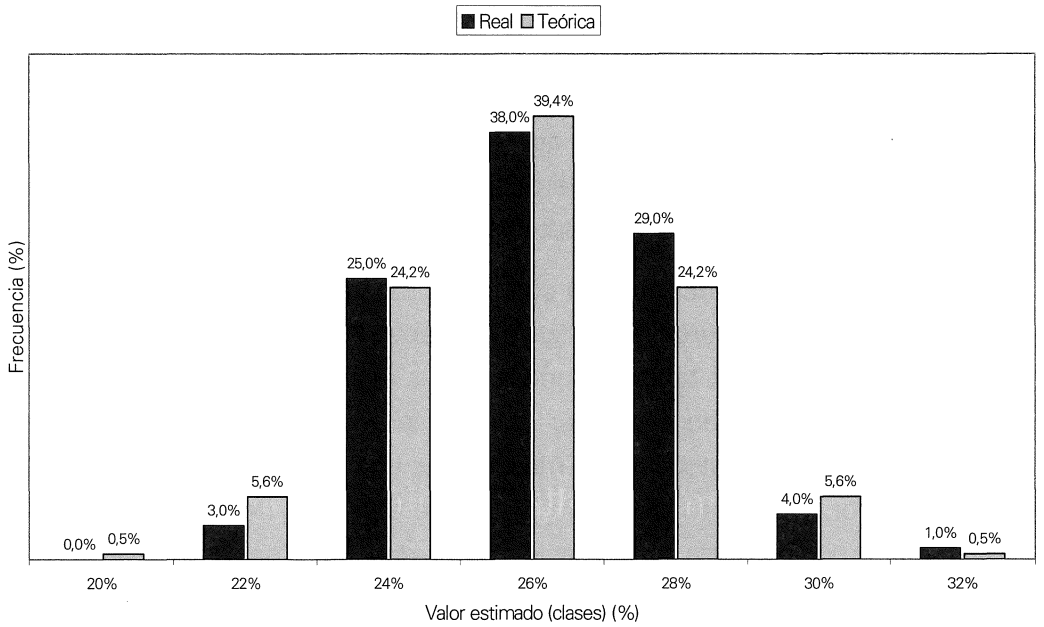
2. La distribución de probabilidad normal nos proporciona la respuesta ya que la distribución de estimaciones se ajusta a una distribución normal siempre que el tamaño de muestra sea grande. La desviación típica de esta distribución normal de estimaciones es el error estándar y su valor medio, en caso de estimaciones insesgadas, es el verdadero valor poblacional. Veamos su significado práctico. Supongamos que un cierto partido político obtiene un 25% de los votos en unas de-

terminadas elecciones y que estamos en situación de poder obtener distintas muestras aleatorias de las papeletas de votación que estimen el porcentaje de votos obtenidos. Esta situación puede simularse obteniendo muestras de una distribución de Bernouilli con probabilidad de éxito p=25%. Obtengamos 100 muestras aleatorias de tamaño n=500 y calculemos en cada muestra el porcentaje de votos obtenido por el partido en cuestión. Los resultados que se obtienen son similares a los de la siguiente tabla:

Intervalo estimación %	% de muestras
20	0,0
22	3,0
24	25,0
26	38,0
28	29,0
30	4,0
32	1,0

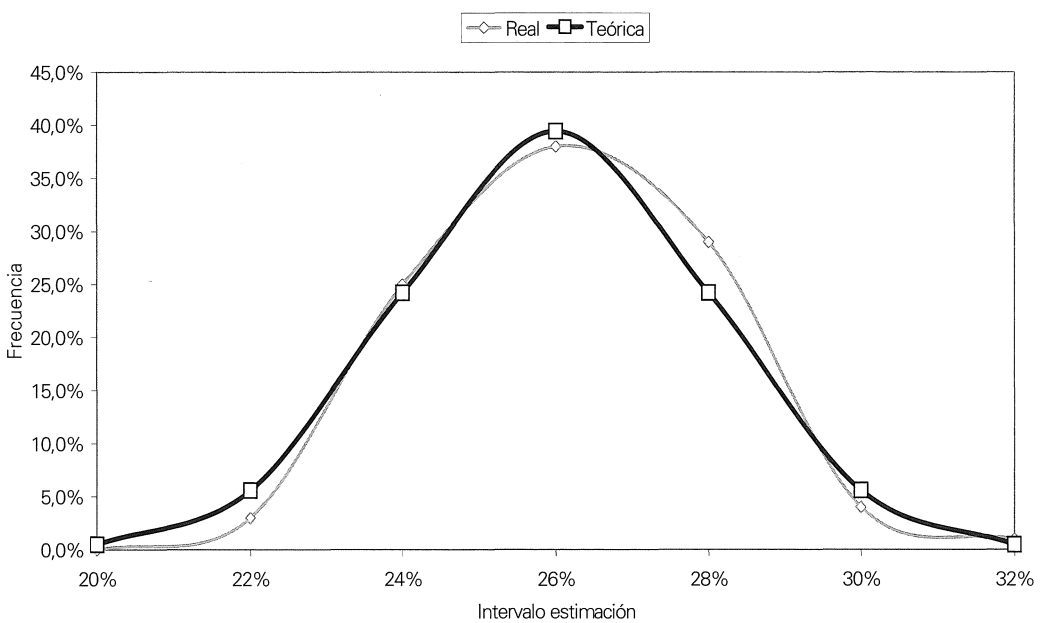
El 38% de muestras han dado estimaciones comprendidas entre 24% y 26%, con sólo un punto de diferencia respecto al valor verdadero de 25% y sólo 8 de las 100 muestras obtenidas arrojan estimaciones fuera del intervalo comprendido entre 22% y 28%. El error estándar de la proporción estimada es de 1,9 puntos porcentuales y según lo dicho, la distribución de estimaciones debería ajustarse a una distribución normal de media 25 y desviación típica de 1,9 puntos porcentuales, por lo que podemos calcular la probabilidad teórica de obtener estimaciones en cada uno de los intervalos de la tabla anterior. Los resultados se muestran en el siguiente gráfico. Al contemplarlo debe tenerse presente que estamos comparando la distribución de estimaciones de sólo 100 muestras aleatorias concretas con la que obtendríamos si dispusiéramos de todas las estimaciones de las casi infinitas muestras posibles de votantes de tamaño 500, por lo que es esperable que el ajuste no sea perfecto. Se trata simplemente de comprobar de forma experimental que la teoría de muestras parece ajustarse a la realidad práctica.

Distribución de estimaciones de una proporción p=25% en 100 muestras aleatorias de tamaño n=500



El mismo gráfico expresado en forma lineal da otra visión de la aproximación entre las frecuencias observadas y las proporcionadas teóricamente por la normal:

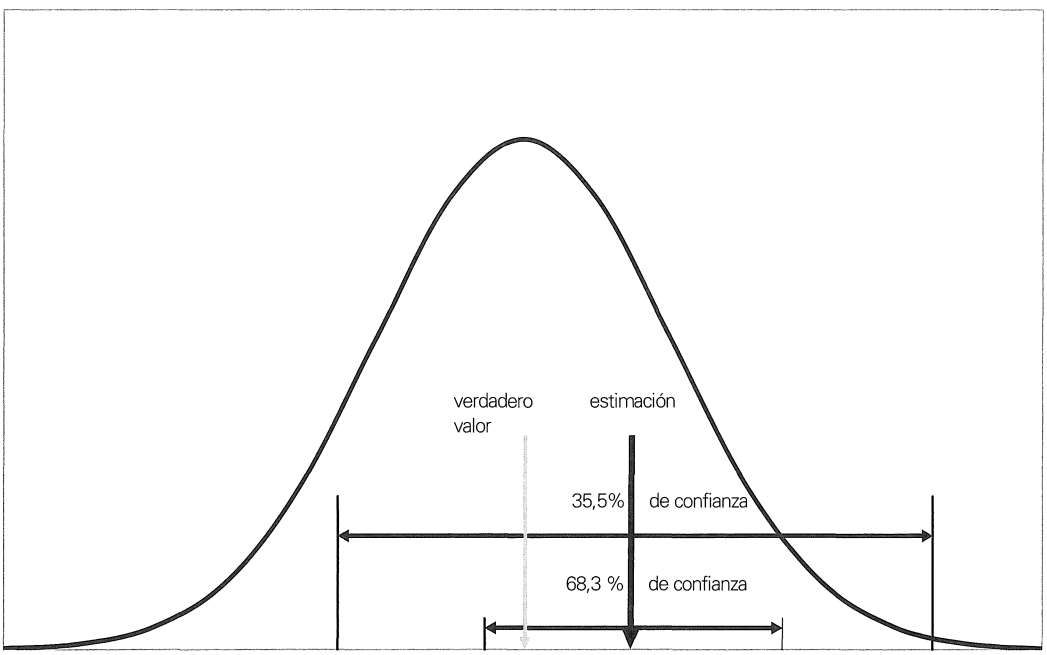
Distribución de estimaciones de una proporción p=25% en 100 muestras aleatorias de tamaño n=500



3. Resulta entonces que con estimadores insesgados, la distribución de posibles estimaciones es normal con media el verdadero valor y desviación típica igual al error de muestreo. Resulta, por tanto, que el 68,3% de las muestras posibles van a proporcionar estimaciones comprendidas entre el verdadero valor y más/menos una vez el error estándar, mientras que el 95,5% de las mismas proporcionan estimaciones comprendidas entre el verdadero valor y más/menos dos veces el error de muestreo, en virtud de las propiedades de la distribución normal.

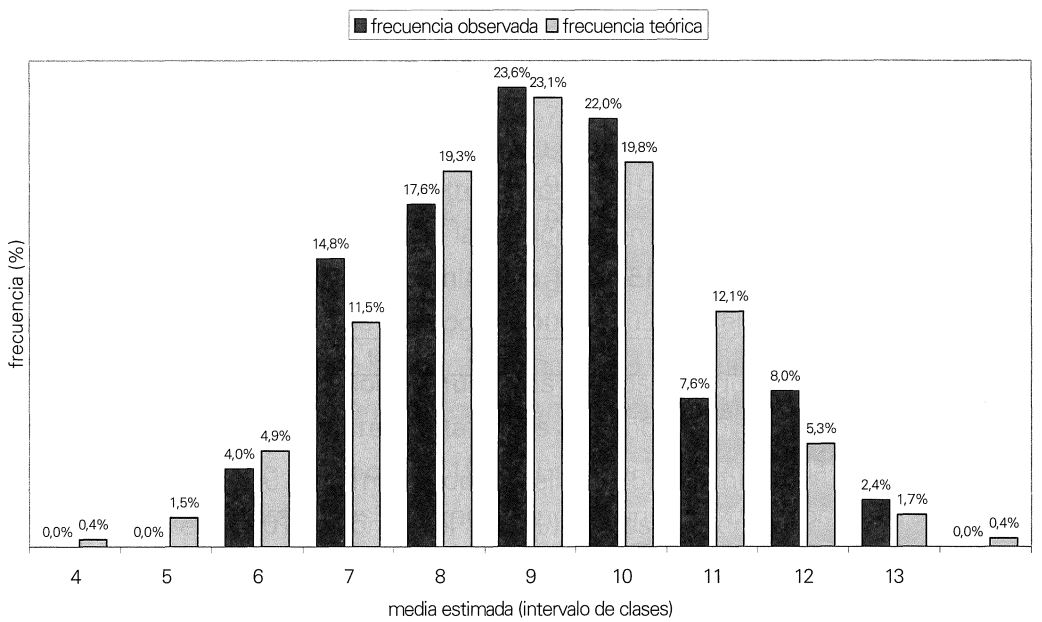
4. En la práctica nos enfrentamos con una muestra y una estimación concreta y no conocemos el valor verdadero y tampoco sabemos sí nuestra estimación queda por encima o por debajo: tan sólo conocemos la estimación y el error estándar. En virtud de lo anterior podemos tener una confianza del 68% de que el intervalo comprendido entre la estimación y más/menos una vez el error estándar cubra el verdadero valor, y podemos tener un 95,5% de confianza de que el valor verdadero esté comprendido entre la estimación obtenida y más/menos dos veces el error estándar. Ello se refleja en el siguiente gráfico, en el que los valores desconocidos aparecen en un tono gris.

Intervalos de confianza alrededor de una estimación



5. Insistimos en que la tendencia a que la distribución de estimaciones sea normal es independiente de la forma de la población de partida y de su grado de asimetría, siempre que el tamaño muestral sea grande. En el capítulo 7 se describió una población de 2.960 supermercados con un alto grado de asimetría. El gráfico que sigue muestra la distribución de estimaciones de la venta media que se ha obtenido con 250 muestras aleatorias de tamaño 100, comparadas con las que teóricamente deberían haberse obtenido de la correspondiente distribución normal. Puede verse que, pese a la asimetría de la distribución original, la distribución de estimaciones se acerca a la normal. La venta media real es de 8,54 millones de euros y puede verse en el gráfico que casi el 24% de estimaciones caen en el intervalo de 8 a 9 millones.

Distribución de medias muestrales de 250 muestras aleatorias (n=100)



4. Estimadores y error estándar en muestreo aleatorio simple

1. Ya hemos adelantado que la media muestral es un estimador insesgado de la media poblacional. Si estamos interesados en el total poblacional $X = \sum_{1}^{N} x = N\overline{X}$,

el estimador insesgado es el resultado de multiplicar la media muestral por el tamaño N de la población, es decir, $\hat{X} = N\bar{x} = \frac{N}{n}\sum_{1}^{n} x$. Obsérvese que la estimación del total implica multiplicar cada valor de la muestra por el factor $\frac{N}{n}$ denominado *factor de expansión*, *factor de elevación*, o simplemente *ponderación*, y que corresponde al inverso de la probabilidad de selección en la muestra de cada unidad de la población visto en el párrafo 11.1.4.

2. En el caso de variables dicotómicas, es decir, que toman el valor 1 si la unidad presenta la característica que se estudia y el valor 0 si no la presenta, la proporción poblacional es un valor medio, por lo que la proporción muestral p es el estimador insesgado de la proporción poblacional P y Np es el estimador insesgado del total de clase, esto es, del número total de elementos de la población que presentan la característica estudiada.

3. El error estándar del estimador de la media se corresponde con la fórmula $\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}}\frac{\sigma}{\sqrt{n}}$, donde σ es la desviación típica de la población. Así, en el ejemplo visto en 11.2 teníamos N=6, n=3 y $\sigma = 3{,}35$ por lo que el error estándar de la media es $\sigma_{\bar{x}} = \sqrt{\frac{6-3}{6-1}}\frac{3{,}35}{\sqrt{3}} = 1{,}5$, que coincide con el allí calculado sobre la base de todas las posibles muestras. En la práctica, no se conoce la varianza de la población σ^2 y el error estándar de la media se estima a partir de los propios datos muestrales mediante

$$\hat{\sigma}_{\bar{x}} = \sqrt{1-\frac{n}{N}}\cdot\frac{s}{\sqrt{n}}, \quad \text{donde} \quad s^2 = \frac{\sum_{1}^{n}(x_i-\bar{x})^2}{n-1}$$

También es muy frecuente en la práctica que el tamaño de muestra n sea pequeño en relación con el de la población N, con lo que el término $1-\frac{n}{N}$ es muy próximo a la unidad y en consecuencia $\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}}$. (Ver nota al final del capítulo sobre el uso de s^2).

El error estándar del estimador del total poblacional se estima multiplicando el error estándar de la media por el tamaño de la población, es decir,

$$\hat{\sigma}_{\hat{X}} = N\hat{\sigma}_{\bar{x}} = N\sqrt{1-\frac{n}{N}}\cdot\frac{s}{\sqrt{n}} \cong N\frac{s}{\sqrt{n}}$$

Suele ser muy habitual manejar los errores estándar en términos relativos, que se obtienen al dividir el error absoluto por el valor estimado:

$$ee_r = \frac{\hat{\sigma}_{\bar{x}}}{\bar{x}} = \frac{s}{\bar{x}\sqrt{n}} = \frac{cv}{\sqrt{n}}$$

donde $cv = \frac{s}{\bar{x}}$ es una estimación del coeficiente de variación de la población calculado con los datos muestrales. Fácilmente puede comprobarse que el error estándar relativo es igual para la media que para el total.

4. En el caso de proporciones el error estándar en términos absolutos, es decir, en puntos porcentuales, se aproxima por $\hat{\sigma}_p = \sqrt{\frac{p(1-p)}{n}}$. Al trabajar con errores absolutos en proporciones debe tenerse presente que, por ejemplo, 2 puntos de error para una proporción del 50% se convierte en un 4% (2/50) de error relativo, pero si la proporción es del 10% supone un 20% de error relativo. Para proporciones suele ser más aconsejable manejar el error en términos absolutos.

5. Veamos un ejemplo sencillo que aclare lo anterior. Supongamos una ciudad de 5.000 viviendas de las que seleccionamos una muestra aleatoria de 50 viviendas para estimar la población total de la ciudad. El número de residentes habituales que se obtiene en las viviendas de la muestra es el siguiente:

Número de residentes por vivienda

0	0	0	0	0	0	0	0	1	1
1	1	1	1	1	1	1	2	2	2
2	2	2	2	3	3	3	3	3	3
3	3	3	3	3	3	4	4	4	4
4	4	4	4	4	4	5	5	6	7

El valor cero indica que la vivienda no está ocupada. La suma de habitantes de las viviendas de la muestra es de 122, la media es de 2,44 personas por vivienda

(incluyendo las no ocupadas) y la desviación típica $s = \sqrt{\frac{\sum_{1}^{50}(x_i - \bar{x})^2}{n-1}}$ es 1,70 per

sonas. Puesto que N = 5.000 y n = 50, el factor de expansión de los datos muestrales es 5.000/50 = 100 y obtenemos la población estimada de la ciudad multiplicando los datos muestrales por 100:

$$\hat{H} = 100(0 + \cdots + 6 + 7) = 100 \cdot 122 = 12.200 \text{ habitantes}$$

Si ahora nos fijamos en la característica de que la vivienda esté o no ocupada, en la muestra tenemos 8 viviendas no ocupadas y 42 ocupadas, de forma que la estimación del número de viviendas ocupadas será $\hat{V}_o = 100 \cdot 42 = 4.200$, mientras que las restantes 800 viviendas no están habitadas, resultando la proporción estimada de viviendas sin habitar en 800/5.000 = 16%.

6. Calculemos ahora el error estándar de las estimaciones anteriores. A partir de los propios datos muestrales hemos obtenido una media de 2,44 y desviación típica s de 1,70 y, por tanto, el error de muestreo para la estimación 2,44 del número medio de habitantes por vivienda será de $\hat{\sigma}_{\bar{x}} = \frac{1,70}{\sqrt{50}} = 0,24$ habitantes y para

el total estimado (12.200) de habitantes de la ciudad, el error estándar será de $\hat{\sigma}_{\hat{x}} = 5.000 \cdot 0,24 = 1.200$ habitantes. De acuerdo a lo dicho en 11.3.4 podemos tener una confianza del 68,3% de que el verdadero número de habitantes de la ciudad este comprendido en el intervalo

(12.200-1.200; 12.200+1.200) = (11.000; 13.400)

y una confianza del 95,5% de que el verdadero valor esté comprendido en el intervalo

(12.200-2*1.200; 12.200+2*1.200) = (9.800; 14.600)

El error estándar en términos relativos es $ee_r = \frac{1.200}{12.200} = 9,8\%$, igual para la media que para el total.

El error de muestreo de la proporción de viviendas no habitadas (16%) será de

$$\hat{\sigma}_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0{,}16(1-0{,}16)}{50}} = 5{,}2 \text{ puntos}$$

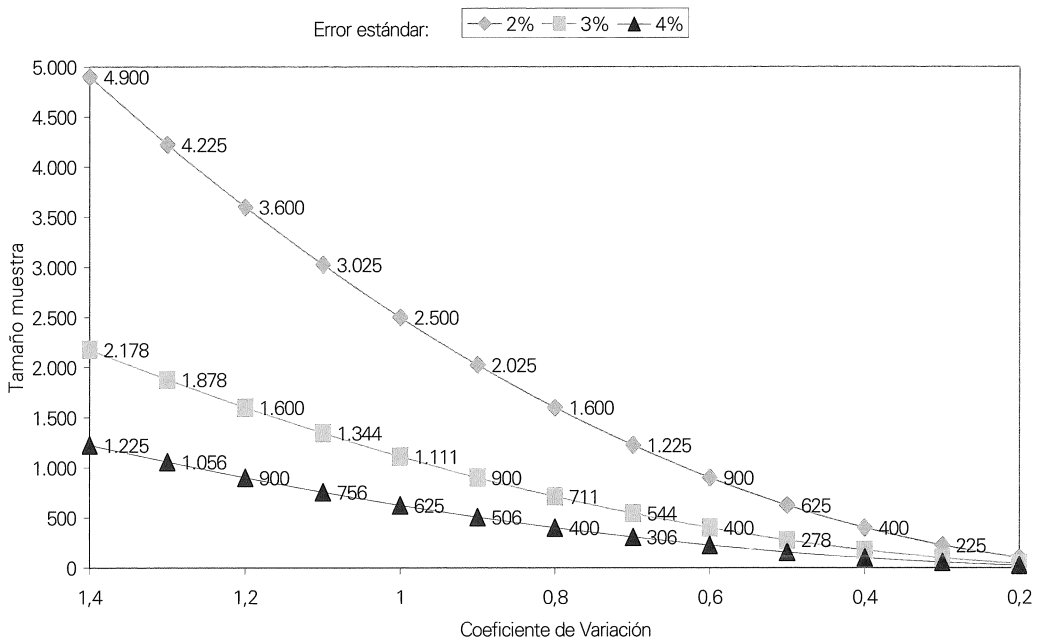
en términos absolutos y de 5,2/16 = 32,5% en términos relativos. Para la proporción de viviendas habitadas (84%) el error de muestreo es

$$s_p = \sqrt{\frac{p(1-p)}{n}} = \sqrt{\frac{0{,}84(1-0{,}84)}{50}} = 5{,}2 \text{ puntos}$$

igual en términos absolutos, pero en términos relativos es de 5,2/84 = 6,2%.

7. Si se quiere determinar el tamaño de muestra necesario para obtener un cierto nivel de error estándar, no hay más que despejar n en las fórmulas anteriores. Conviene notar que el error estándar es inversamente proporcional a la raíz cuadrada del tamaño de muestra. Esto significa, por ejemplo, que para reducir el error estándar a la mitad es necesario tomar un tamaño de muestra cuatro veces superior y si lo queremos reducir a la tercera parte necesitaremos una muestra nueve veces mayor. El siguiente gráfico relaciona el coeficiente de variación de la población, el error estándar y el tamaño de muestra. En el mismo puede comprobarse como para un coeficiente de variación poblacional del 80%, necesitamos una muestra de 400 unidades si deseamos un error de muestreo del 4%; pero si el error de muestreo deseado es del 2%, entonces la muestra necesaria se incrementa hasta 1.600 unidades. El gráfico ilustra también como el tamaño de muestra se incrementa con la variabilidad de la población.

Tamaño de muestra según CV de la población y error estándar



Nota sobre la utilización de s^2:

En 11.4.3 se ha introducido la expresión $s^2 = \frac{\sum_{1}^{n}(x_i - \bar{x})^2}{n-1}$, que se denomina *cuasivarianza*, y en la que la suma de cuadrados de las desviaciones a la media se divide por n-1 en lugar de n. También allí se indicaba que en muestreo aleatorio simple la varianza del estimador de la media (cuadrado del error estándar) es $\sigma^2_{\bar{x}} = \frac{N-n}{N-1}\frac{\sigma^2}{n}$, dónde σ^2 es la varianza poblacional, en general desconocida. Pues bien, la razón teórica para introducir la cuasivarianza s^2 es que el estimador insesgado de la varianza de la media $\sigma^2_{\bar{x}} = \frac{N-n}{N-1}\frac{\sigma^2}{n}$, es $\hat{\sigma}^2_{\bar{x}} = (1-\frac{n}{N})\frac{s^2}{n}$.

En la práctica, siendo n suficientemente grande resulta indiferente dividir la suma de cuadrados de las desviaciones a la media por n-1 o por n.

También para ser consistentes con lo anterior, la varianza estimada de la proporción muestral debería calcularse como $\hat{\sigma}^2_p = (1-\frac{n}{N})\frac{p(1-p)}{n-1}$. Sin embargo el factor (1-n/N) es muy pocas veces importante y con muestras de tamaño moderado o grande no importa que se trabaje con n ó n-1. De ahí que en 11.4.4 se haya utilizado la expresión $\hat{\sigma}_p = \sqrt{\frac{p(1-p)}{n}}$.

Capítulo 12
Población y marco. Muestreo en etapas

1. Unidades de muestreo y unidades elementales

1. Conviene distinguir entre *unidad elemental* y *unidad de muestreo*. La unidad elemental o unidad de estudio es todo elemento o individuo miembro de la población objetivo o población de referencia. Las variables objeto de estudio o variables estadísticas se miden sobre las unidades elementales o unidades estadísticas. Las unidades de muestreo son aquellas que forman parte del proceso de selección de la muestra. La unidad de muestreo puede coincidir con la unidad elemental, en cuyo caso hablamos de muestreo de unidades elementales, o puede referirse a un conjunto de unidades elementales, que se denomina *conglomerado*.

2. Por ejemplo en la Encuesta Industrial de Empresas las unidades de estudio son las empresas y se toma una muestra de las mismas a partir de un listado o directorio de empresas (DIRCE). En este caso la unidad elemental y la unidad de muestreo coinciden. Sin embargo, en la Encuesta de Población Activa (EPA), las unidades de estudio son las personas residentes en viviendas familiares, pero no se toma directamente una muestra de personas: primero se selecciona una muestra de secciones censales[1] y, posteriormente, dentro de cada sección seleccionada

1 Una sección censal es una porción de territorio bien delimitado. Las secciones censales cubren todo el territorio nacional y no se solapan entre sí. España está dividida en unas 34.000 secciones, con un tamaño medio que no llega a 1.500 habitantes.

se toma una muestra de viviendas. Tenemos dos unidades de muestreo: las secciones, en primer lugar, y las viviendas de las secciones de la muestra, en segundo lugar. Las unidades de estudio son las personas que habitan las viviendas seleccionadas. La muestra de la EPA es un ejemplo de *muestreo en dos etapas*.

2. Marco de muestreo

1. El concepto de población establecido anteriormente como conjunto de unidades de las que se desea información, se refiere a la población objetivo o *población de referencia* y constituye un modelo ideal. En la práctica, la muestra se selecciona a partir de un material soporte, denominado *marco*, que coincide en mayor o menor grado con la población objetivo. En sentido estricto, el marco de muestreo se define como la lista de unidades de muestreo a partir de la cual se selecciona la muestra.

Si hablamos por ejemplo de la Encuesta Industrial de Empresas 2002, la población de referencia son las empresas industriales que han tenido actividad económica durante el año 2002 y el marco de muestreo es la lista del DIRCE. Pero, como suele suceder con cualquier listado, su actualización no es perfecta y nos encontramos que en la realidad existen empresas que han tenido actividad económica en el año que no figuran en el DIRCE, y a la inversa, tenemos empresas en el DIRCE que no existen o no han tenido actividad económica en el año. Resulta evidente que la coincidencia entre el marco y la población objetivo debe ser máxima. La *cobertura* de un marco es la medida en que dicho marco incluye a todos los elementos de la población objetivo.

2. En sentido amplio, el marco de muestreo comprende no sólo listas de unidades de muestreo, sino que incluye todo el material e información previa que disponemos sobre la población y su agrupación en unidades de muestreo. En el caso de las secciones censales forman parte del marco, además de la identificación y delimitación geográfica de las mismas, la información que tengamos de ellas sobre tamaño en términos de habitantes, hogares y viviendas, los callejeros de sección, los planos, la lista de viviendas de cada sección, ... En el DIRCE el marco incluye toda la información disponible sobre las empresas, como actividad económica, número de empleados, facturación, año de constitución, holding al que pertenece, personas de contacto, ...

3. Muestreo en etapas

1. Es frecuente que el muestreo de unidades elementales no sea utilizado en la práctica por la imposibilidad práctica en muchas ocasiones de obtener una lista de unidades elementales en la cuál basar la selección de la muestra, y también porque la selección de unidades elementales proporciona en general una muestra muy esparcida de unidades a entrevistar con el consiguiente incremento de coste y tiempo.

2. Para evitar estos inconvenientes surge, de forma natural, el *muestreo de conglomerados*, agrupando las unidades elementales próximas en un conglomerado que se constituye en la nueva unidad de muestreo, más grande que la unidad elemental, como es el caso de las secciones en la muestra de la EPA. Los conglomerados deben estar perfectamente definidos, lo cuál significa que no haya solapamiento entre ellos (una unidad elemental pertenece sólo a un conglomerado) y que el conjunto de todos los conglomerados contiene a la población objeto de estudio.

3. La agrupación de unidades elementales en unidades de muestreo más amplias tiene ventajas e inconvenientes. Entre las ventajas podemos citar el ahorro de coste y tiempo, y la mayor facilidad de preparar listas (sólo se necesitan para los conglomerados de la muestra). De los inconvenientes hay que destacar la menor precisión derivada de una mayor homogeneidad de las unidades elementales dentro de un conglomerado respecto a la característica de estudio. Por ejemplo en una cierta sección de la EPA puede haber una cierta tendencia a mayor o menor desempleo dependiendo de en qué zona o barrio de una cierta ciudad se localice. En un muestreo sobre intención de voto en unas elecciones es sabido que determinadas zonas tienen una acusada fidelidad a un determinado partido político.

4. Si en el proceso de muestreo investigamos todas las unidades elementales contenidas en los conglomerados seleccionados en la muestra, el muestreo se denomina *en una etapa* o monoetápico. Ahora bien, para evitar el inconveniente apuntado (homogeneidad dentro del conglomerado) podemos investigar no todas las unidades elementales del conglomerado, sino seleccionar a su vez una muestra probabilística de las mismas. Estaríamos así ante un *muestreo en dos etapas*, como hemos visto en la EPA: las unidades de primera etapa o unidades primarias de muestreo serían los conglomerados y las unidades de segunda etapa serían las unidades elementales.

5. Este proceso puede generalizarse llevándonos así al *muestreo multietápico o polietápico*. Obsérvese que en muestreo por etapas se definen distintas unidades de muestreo y que la *lista* de unidades de muestreo en una etapa dada, sólo es necesario disponerla para las unidades seleccionadas en la etapa inmediatamente anterior. Se constituye así una jerarquía entre las distintas unidades de muestreo de acuerdo a las etapas del proceso.

Capítulo 13
Muestreo estratificado

1. Definición y objetivos

1. El muestreo estratificado consiste en : 1º) Dividir la población de N unidades en un cierto número de subpoblaciones llamadas *estratos*, de forma que las unidades que componen cada estrato sean lo más homogéneas posibles en cuanto a la variable objeto de estudio. Cada unidad de la población ha de pertenecer a uno y sólo uno de los estratos formados. El número de unidades que pertenecen a un estrato dado es el tamaño del estrato. 2º) Seleccionar una muestra probabilística en cada estrato. La muestra de cada estrato es independiente de la muestra de cualquier otro estrato. Si la muestra dentro de cada estrato es una muestra aleatoria simple (probabilidades iguales) tenemos el muestreo aleatorio estratificado.

2. Los principales *objetivos* del muestreo estratificado son:

a) Ganancia en precisión respecto al muestreo no estratificado. Es el objetivo fundamental y en poblaciones muy asimétricas pueden conseguirse excelentes resultados como veremos en un ejemplo.

b) Posibilidad de obtener estimadores separados para cada estrato o agrupación de estratos, lo que proporciona una información más rica y detallada.

c) Más eficacia en la organización administrativa, al poder considerar como variables de estratificación provincias o regiones geográficas, que permiten una mayor descentralización de la organización de Campo y de tareas administrativas.

d) Los problemas de muestreo pueden diferir marcadamente en diferentes partes de la población. Al ser el proceso de muestreo independiente en cada estrato,

pueden aplicarse métodos diferentes de muestreo por estrato de acuerdo a la información de que se disponga.

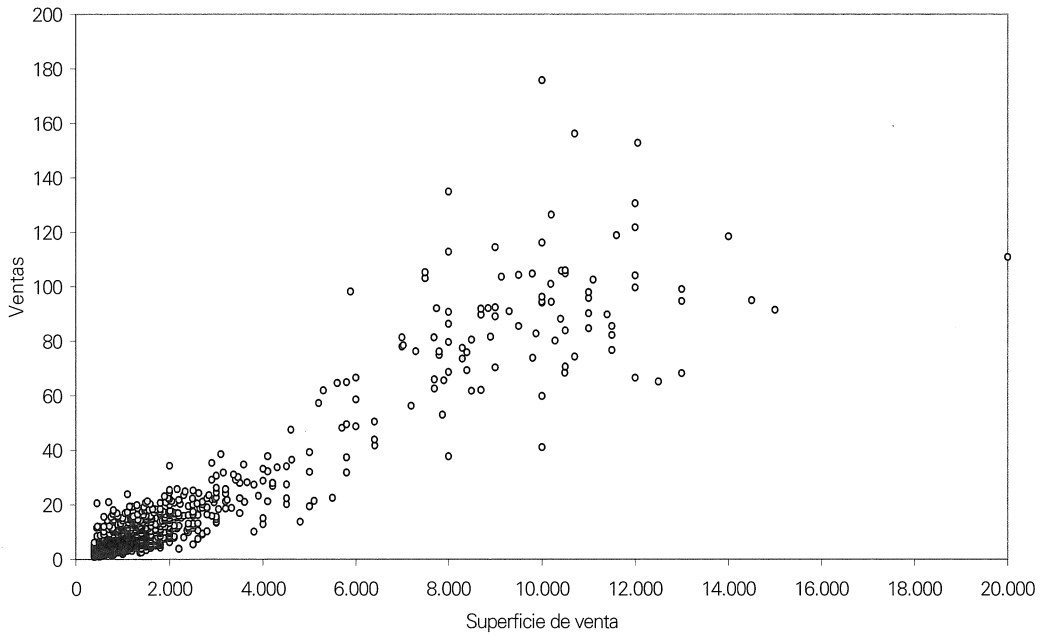
3. Respecto a las variables o criterios de estratificación, su número y el número de estratos, dependen de los objetivos concretos de cada caso, de la información disponible y de la estructura de la población; las variables utilizadas en la estratificación, deberán estar correlacionadas con las variables objeto de investigación, aunque también pueden incluirse criterios *administrativos* (regiones geográficas, tamaño de municipio, sectores de actividad) o necesidades de desglose o clasificación de la información final, por ejemplo una encuesta a funcionarios en la que se necesita información separada por ministerio en el que se trabaja: el ministerio se convertiría en una variable de estratificación. En general, un número moderado de variables de estratificación y de estratos es suficiente para obtener ganancias de precisión, ganancia que suele ser decreciente al aumentar el número de estratos.

2. Un ejemplo: ventas de supermercados

1. Volvamos a la población de 2.960 supermercados con superficie de venta igual o superior a 400 m^2, que hemos visto en capítulos anteriores. Recordemos que las ventas totales son de 25.265 millones de euros con una venta media de 8,54 millones por supermercado y un coeficiente de variación de 199%. Esto significa que una muestra aleatoria de 100 supermercados, seleccionada para estimar las ventas totales, tendría un error de muestreo de $\frac{199}{\sqrt{100}} = 19{,}9\%$.

2. Sabemos que la superficie de venta está correlacionada con las ventas según puede verse en el gráfico.

Superficie y ventas de supermercados



Podemos utilizar la superficie como variable para formar tres estratos: entre 400 y 999 metros cuadrados, de 1.000 a 2.499 m^2 y 2.500 m^2 y más. Los datos poblacionales se resumen en la siguiente tabla:

Datos poblacionales de supermercados

	Nº de establecimientos	Venta total (millones)	Venta Media (millones)	Desviación típica	Coeficiente variación
Total población	**2.960**	**25.265**	**8,54**	**17,02**	**199**
Estrato 1					
<1.000 m^2	2.148	8.099	3,77	2,11	56
Estrato 2					
1.000-2.500 m^2	617	5.617	9,10	4,96	54
Estrato 3					
≥ 2.500 m^2	195	11.549	59,23	38,08	64

Lo interesante del cuadro es comprobar cómo el coeficiente de variación de la población, de 199%, se reduce drásticamente con la estratificación: en lugar de considerar la población total, pasamos a considerar tres subpoblaciones (estratos) con coeficientes de variación de 56%, 54% y 64%, respectivamente, es decir mucho más homogéneas. Resulta claro que al seleccionar ahora una muestra en cada estrato, el error de muestreo se reducirá sensiblemente.

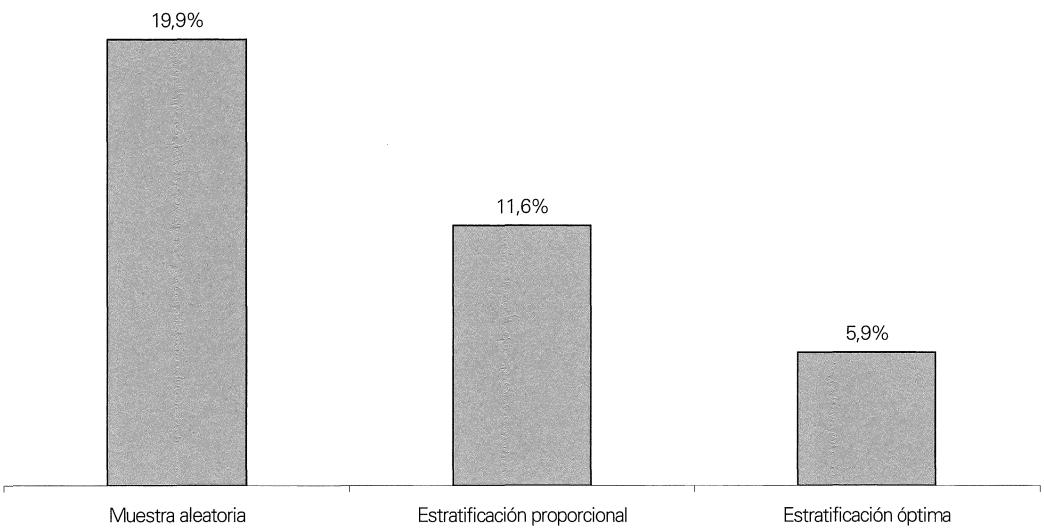
3. Afijación

1. En muestreo estratificado surge el problema de cómo distribuir la muestra total entre los estratos, que se conoce como *afijación* de la muestra. En principio la muestra puede distribuirse de cualquier forma, según el buen juicio del diseñador de la misma. Pero lo normal es que se utilice alguno de los dos siguientes criterios.

2. La *afijación proporcional* distribuye la muestra total en proporción al número de unidades de cada estrato. Así, en una muestra de 100 supermercados repartidos proporcionalmente al número de supermercados de cada estrato, resultaría en 72 supermercados en el estrato 1, 21 en el estrato 2 y 7 en el estrato de los más grandes. El error de muestreo que arrojaría esta muestra para la estimación de las ventas totales sería del 11,6%, en lugar del 20% obtenido con una muestra aleatoria, reflejando el efecto de la disminución de la variabilidad poblacional producida por la estratificación.

3. La *afijación óptima* distribuye la muestra total entre los estratos de forma que se minimice el error de muestreo. Para ello tiene en cuenta no sólo el número de unidades de cada estrato, sino también la desviación típica de cada uno. En el caso de los supermercados puede verse en el cuadro que la desviación típica en cada estrato varía entre 2,11 en los pequeños y 38,08 en los más grandes y la consecuencia es que la afijación óptima *tira* de la muestra hacia los grandes. La anterior muestra de 100 supermercados repartida de forma óptima supondría muestrear 30, 20 y 50 supermercados en los estratos 1, 2 y 3, respectivamente, es decir, la muestra se va hacia los estratos con mayor desviación típica. El error de muestreo que se obtendría con esta distribución muestral es del 5,9%. Vemos, pues, que la estratificación y la forma de distribuir la muestra entre estratos puede producir importantes ganancias en precisión:

Comparación de errores de muestreo. Población de supermercados. Tamaño de muestra n = 100



4. La estratificación, si es eficiente, produce también importantes ahorros de coste ya que permite trabajar con tamaños de muestra menores. En el gráfico vemos que el error estándar con afijación proporcional es casi la mitad que el de una muestra aleatoria de igual tamaño. Para alcanzar esta precisión con muestreo aleatorio se necesitaría una muestra casi cuatro veces mayor, lo que implicaría un aumento de costes de similares proporciones.

Capítulo 14
Estimador de razón

1. El estimador de razón trata de mejorar la precisión de un estimador utilizando la información que se posee, para la población investigada, de una variable auxiliar que se supone *correlacionada* con la variable de estudio. Sea Y_i la variable de estudio y sea X_i la variable auxiliar conocida para el universo o población en estudio, es decir, conocemos el total poblacional X.

2. Supongamos que se desea estimar la producción de trigo de una población de N=5.000 explotaciones agrarias mediante una muestra aleatoria de n=100 explotaciones, y poseemos información sobre la superficie cultivada:

explotación	prod. trigo (Y_i)	superf. cultivada (X_i)
1	Y_1	X_1
2	Y_2	X_2
.....	.....	.....
n=100	Y_n	X_n
total muestral	y=5.100 t	x=1.700 ha

El estimador insesgado lineal de la producción de trigo es

$$\hat{Y} = \frac{N}{n}\sum_{1}^{n} Y_i = \frac{N}{n} y = 50 * 5.100 = 255.000 \text{ t}$$

Puesto que poseemos información de la superficie cultivada X_i y conocemos su total poblacional X=100.000 ha, podemos, además, estimarlo con los datos de la muestra

$$\hat{X} = \frac{N}{n}\sum_{1}^{n} X_i = \frac{N}{n}x = 50 * 1.700 = 85.000 \text{ ha}$$

El cociente $\frac{X}{\hat{X}} = \frac{\overline{X}}{\overline{x}}$ constituye una cierta medida de la representatividad de la muestra: si $\frac{X}{\hat{X}} > 1$, como es nuestro caso en que $\frac{X}{\hat{X}} = 1{,}176$, indicaría que en la muestra hay una mayor representación de explotaciones pequeñas, mientras que si $\frac{X}{\hat{X}} < 1$, tendríamos una mayor representación de explotaciones grandes. Habiendo correlación entre ambas variables parece lógico utilizar la desviación $\frac{X}{\hat{X}}$, cometida en la estimación de la variable conocida para corregir la estimación de Y. Esto nos lleva al estimador

$$\hat{Y}_R = \hat{Y}\frac{X}{\hat{X}} = \frac{\hat{Y}}{\hat{X}}X = \hat{R}X$$

En muestro ejemplo

$$\hat{Y}_R = 255.000 \text{ t}\,\frac{100.000 \text{ ha}}{85.000 \text{ ha}} = \frac{255.000 \text{ t}}{85.000 \text{ ha}}100.000 \text{ ha} = 3\ {}^{t}/_{ha}\ 100.000 \text{ ha} = 300.000 \text{ t}$$

$\hat{R} = \frac{\hat{Y}}{\hat{X}} = \frac{\overline{y}}{\overline{x}}$, se llama *estimador de razón*, $\hat{Y}_R$ es el *estimador del total por el método de razón*. $\hat{Y}_R$ lo podemos escribir como

$$\hat{Y}_R = \frac{N}{n}y\frac{X}{\frac{N}{n}x} = \frac{X}{x}\sum_{1}^{n} Y_i = \frac{100.000 \text{ ha}}{1.700 \text{ ha}}5.100 \text{ t} = 300.000 \text{ t}$$

es decir, el estimador del total por razón equivale a la expansión de los datos muestrales mediante el factor X/x , relación entre el valor poblacional y el valor muestral de la variable auxiliar X_i , en lugar de utilizar la expansión N/n de número o expansión simple, directa o de diseño. Al factor X/x le llamamos factor-X. El cociente entre ambos factores coincide con la medida de representatividad muestral $\frac{X}{\hat{X}}$, ya que

$$\frac{X/x}{N/n} = \frac{X}{\frac{N}{n}x} = \frac{X}{\hat{X}} = \frac{\overline{X}}{\overline{x}}$$

La media $\overline{Y}$ se estima por $\hat{\overline{Y}}_R = \frac{\hat{Y}_R}{N} = \hat{R}\frac{X}{N} = \hat{R}\overline{X}$.

3. Otros ejemplos del método de razón los encontramos cuando se estiman ventas o valores de producción de una población de empresas o establecimientos utilizando los datos de personal empleado como variable auxiliar. En general, cuando la razón $\frac{Y_i}{X_i}$ es similar en todas las unidades de la población, la razón muestral $\frac{y}{x}$ varía poco de muestra a muestra y el estimador de razón es de gran precisión.

En el ejemplo de explotaciones agrícolas, si el rendimiento o producción de trigo por hectárea no varía mucho entre todas las explotaciones de la población, la estimación por razón puede ser considerablemente mejor que la de simple expansión.

4. Tanto el muestreo estratificado como el estimador de razón son ejemplos de procedimientos que utilizan de forma muy eficiente la información auxiliar disponible sobre la población objeto de estudio para mejorar las estimaciones y disminuir tamaños de muestra y costes. Lo habitual en la práctica es que ambos procedimientos se utilicen de forma simultánea, es decir, se realiza un muestreo estratificado y se utiliza estimador de razón en cada estrato, combinándose de forma adecuada las estimaciones posteriormente.

5. Como un ejemplo más de utilización del método de razón podemos citar la Encuesta de Población Activa (EPA). De acuerdo al procedimiento de selección de la muestra, brevemente descrito en 12.1.2, el factor de expansión de diseño o directo a utilizar sería el cociente entre el número de viviendas en la población y en la muestra. Sin embargo, la EPA utiliza como variable auxiliar en sus estimaciones los datos de número de habitantes o población residente en viviendas familiares, conocidos a través del padrón y de las proyecciones de población. En consecuencia la expansión en la EPA se basa en la relación de la población total residente en viviendas familiares a la población residente en las viviendas de la muestra, calculada en cada estrato.

Capítulo 15
Muestreo sistemático

1. Sea una población $\{u_1, u_2, \ldots, u_N\}$. La selección sistemática de una muestra de n unidades se realiza en la siguiente forma: sea $k = N/n$ (suponemos N divisible por n), tomamos un número i al azar $1 \leq i \leq k$ con probabilidad $1/k$ y la muestra sistemática queda formada por las n unidades $\{u_i, u_{i+k}, u_{i+2k}, \ldots, u_{i+(n-1)k}\}$. Como vemos, la selección de la primera unidad determina la muestra completa y existen $k = N/n$ muestras posibles. Las k muestras posibles son equiprobables (prob. = $1/k$) y la probabilidad de que la unidad u_i esté en la muestra es $1/k = n/N$. La media muestral es el estimador insesgado de la media poblacional.

2. Por ejemplo si N=60 y n=10 tenemos (k=6) las siguientes seis muestras posibles, dónde se indica el valor de la variable X en estudio en cada unidad seleccionada:

Muestra

1	2	3	4	5	6
X_1	X_2	X_3	X_4	X_5	X_6
X_7	X_8	X_9	X_{10}	X_{11}	X_{12}
.....	.....	.....	.....	.....	.....
X_{49}	X_{50}	X_{51}	X_{52}	X_{53}	X_{54}
X_{55}	X_{56}	X_{57}	X_{58}	X_{59}	X_{60}

3. El muestreo sistemático es de fácil aplicación práctica y asegura además que la muestra se extiende a toda la población. El comportamiento del muestreo sistemático respecto al muestreo aleatorio simple, depende en gran medida de las propiedades de la población. En poblaciones en las cuales la numeración de las unidades puede considerarse al azar respecto a la característica que se mide, cabría esperar que el muestreo sistemático fuera equivalente al muestreo aleatorio simple y que tuviera un error de muestreo similar e incluso menor por su efecto distribuidor de la muestra.

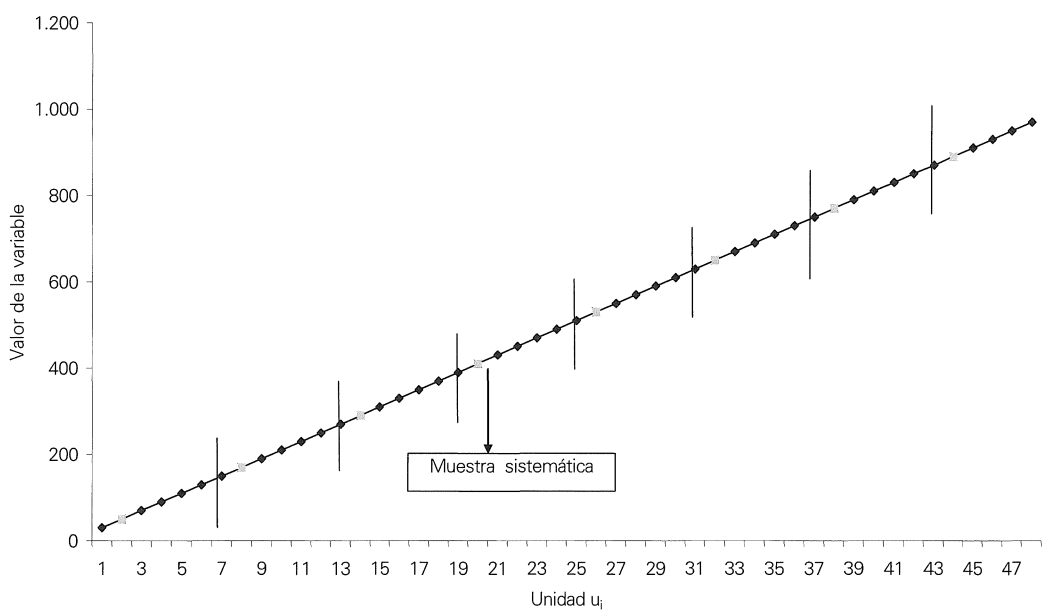
4. Podría considerarse la población dividida en n estratos, los cuales consisten de las primeras k unidades, las segundas k unidades, ..., es decir, al contemplar el cuadro de muestras posibles en horizontal, cada fila sería un estrato. La muestra sistemática correspondería a una muestra estratificada con una unidad por estrato, por lo que sería razonable esperar una mayor precisión respecto al muestreo aleatorio simple.

5. El cuadro que sigue presenta las seis muestras sistemáticas de tamaño 8 de una población de 48 unidades, junto con la media estimada por cada muestra.

Muestra						
	1	2	3	4	5	6
	30	50	70	90	110	130
	150	170	190	210	230	250
	270	290	310	330	350	370
	390	410	430	450	470	490
	510	530	550	570	590	610
	630	650	670	690	710	730
	750	770	790	810	830	850
	870	890	910	930	950	970
Media	450	470	490	510	530	550

La población está ordenada por el valor de la variable, cuya media poblacional es igual a 500, y se puede apreciar horizontalmente el efecto de estratificación aludido. En el gráfico puede observarse la tendencia lineal de los valores poblacionales:

Muestreo sistemático en una población con tendencia lineal

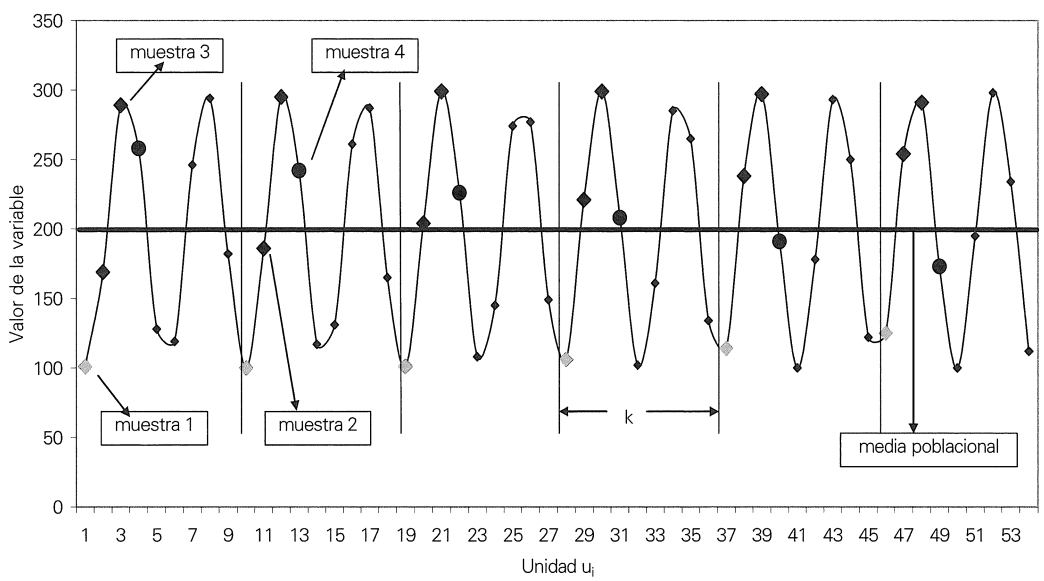


Se destaca también en el gráfico una de las posibles muestras sistemáticas. Intuitivamente se ve que la muestra sistemática es más efectiva que la muestra aleatoria simple ya que asegura presencia en la muestra de todas las zonas de tendencia, pero es menos efectiva que la muestra estratificada ya que si la muestra sistemática es muy baja en un estrato, es muy baja en todos, mientras que la estratificación da oportunidad a seleccionar cualquier unidad dentro de cada estrato. El comportamiento de la muestra sistemática podría mejorarse usando una muestra centralmente ubicada. De hecho, las muestras 3 y 4 son las que presentan estimaciones más próximas al valor poblacional.

6. En poblaciones cuya ordenación tiene una componente periódica hay que ser especialmente cuidadosos en el uso del muestreo sistemático. El cuadro y gráfico que siguen presenta los datos de una población con valor medio $\overline{X} = 198$ y las distintas muestras sistemáticas:

	Muestra								
	1	2	3	4	5	6	7	8	9
	101	169	289	258	128	119	246	294	182
	100	186	295	242	117	131	261	287	165
	101	204	299	226	108	145	274	277	149
	106	221	299	208	102	161	285	265	134
	114	238	297	191	100	178	293	250	122
	125	254	291	173	100	195	298	234	112
Media	108	212	295	216	109	155	276	268	144

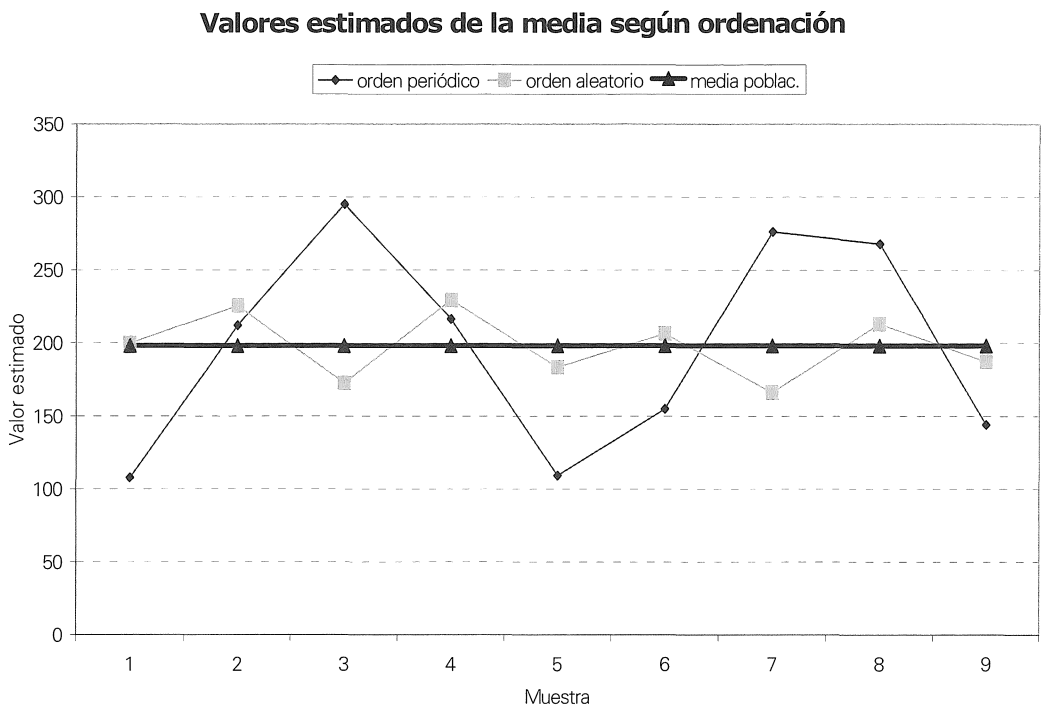
Muestras sistemáticas en una población con componente periódica



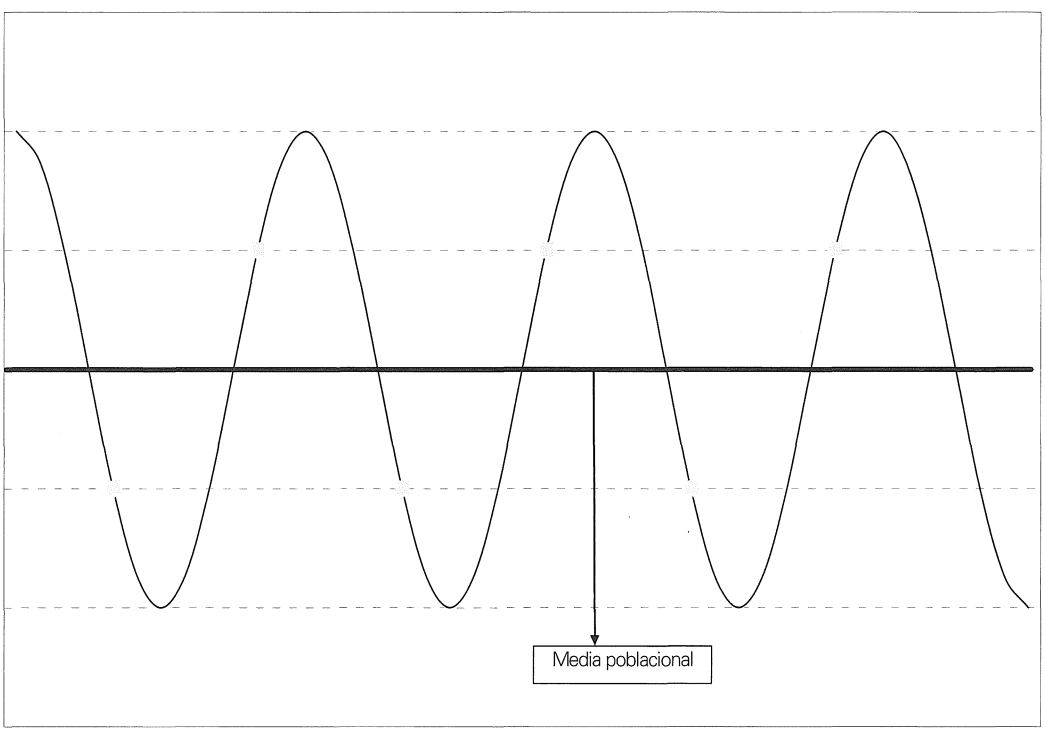
Debido a la periodicidad en la ordenación existen muestras que repiten los valores más pequeños o los más grandes y se alejan sensiblemente de la realidad. Esto sucede porque el valor k que determina la composición de la muestra sistemática coincide o es múltiplo entero del periodo que marca la ordenación de la población. Sin embargo, si reordenamos de forma aleatoria la población el problema puede evitarse. El cuadro que sigue presenta los mismos datos poblacionales ordenados aleatoriamente, junto a las estimaciones de cada muestra sistemática.

	Muestra								
	1	2	3	4	5	6	7	8	9
	195	246	261	242	173	295	102	293	100
	234	169	100	117	178	131	100	186	299
	112	287	114	277	165	297	226	204	250
	299	254	101	294	274	145	134	122	125
	101	289	161	238	191	106	285	182	128
	258	108	298	208	119	265	149	291	221
media	200	226	173	229	183	207	166	213	187

El gráfico proporciona las medias estimadas por cada muestra sistemática cuando la ordenación tiene una componente periódica o es aleatoria: la variabilidad de las estimaciones (error de muestreo) con ordenación aleatoria es sensiblemente menor.



7. Poblaciones con tendencia más o menos periódica se encuentran en la práctica con relativa frecuencia y no siempre es posible proceder a una ordenación aleatoria de la misma. Ejemplos son el flujo de tránsito por un punto de una carretera durante las 24 horas del día y las ventas de una tienda durante los días de la semana. Para estimar un promedio sobre un periodo de tiempo, una muestra sistemática diaria a las 6 de la tarde o cada martes, no sería obviamente juicioso. La estrategia correcta es girar la muestra sobre la curva periódica, por ejemplo, viendo que cada día de la semana y cada segmento horario esté igualmente representado, en el caso de las ventas de una tienda. La figura que sigue presenta una muestra sistemática que proporciona una media muestral exactamente igual a la media poblacional.



Capítulo 16
El efecto del diseño

1. En 11.4.3 se vio que la varianza de la media en un muestreo aleatorio simple se estima mediante $v_{as}(\bar{x}) = (1-\frac{n}{N})\frac{s^2}{n}$. Cuando el diseño muestral se hace más complejo (muestreo estratificado y/o por conglomerados, uso del estimador de razón), la fórmula para estimar la varianza de la media resulta también más compleja y suele recurrirse en la práctica a relacionar la varianza del estimador de la media con el diseño actual y la que se obtendría con un muestreo aleatorio simple con el mismo número de unidades elementales en la muestra. En 13.3.3 ya se utilizó implícitamente esta relación para comparar la precisión del muestreo estratificado respecto al aleatorio simple. Se define el efecto del diseño (Efd) como la razón de la varianza real del estimador de una muestra dada a la varianza que tendría el estimador en una muestra aleatoria simple con el mismo número de unidades elementales:

$$Efd = \frac{v(\bar{x})}{v_{as}(\bar{x})}$$

De dónde

$$v(\bar{x}) = v_{as}(\bar{x}) \cdot Efd = (1-\frac{n}{N})\frac{s^2}{n} \cdot Efd \cong \frac{s^2}{n} \cdot Efd = \frac{s_d^2}{n}$$

con

$$s_d^2 = s^2 \cdot Efd$$

En términos relativos tendríamos $\frac{v(\bar{x})}{\bar{x}^2} = \frac{s^2}{\bar{x}^2 n} \cdot Efd = \frac{cv^2}{n} . Efd = \frac{cv_d^2}{n}$

2. s_d^2 puede considerarse como una varianza por elemento que incorpora todas las complejidades del diseño muestral y, por tanto, varía con cualquier cambio que se haga en el diseño de la muestra, mientras que s^2 se refiere a la varianza por elemento sin considerar el diseño muestral. De la misma forma cv sería el coeficiente de variación sin incluir el diseño muestral y cv_d se referiría a un coeficiente de variación que incorpora toda la complejidad del diseño. El *Efd* suele ser menor a la unidad en muestreo estratificado, expresando la reducción en la varianza debida a la estratificación, mientras que en muestreo de conglomerados será mayor que uno debido al similar comportamiento de las unidades dentro de cada conglomerado, que hace perder eficiencia al muestreo. Una ventaja de introducir la idea del efecto del diseño es que permite manejar, en diseños de muestreo complejos, los conceptos de error estándar y tamaño de muestra con las fórmulas sencillas del muestro aleatorio simple.

3. Al planear una nueva encuesta es frecuente que no se disponga de datos concretos sobre la variabilidad poblacional s^2 por elemento. Si el diseño muestral va a incorporar estratificación, el uso de conglomerados y/o estimadores del tipo de razón, tampoco se conocerá a priori la incidencia del diseño muestral en los errores de muestreo. En ocasiones será necesario realizar algún tipo de estudio piloto previo que permita tener estimaciones de los datos necesarios para realizar un diseño eficiente. En otras ocasiones podrá utilizarse la experiencia de otras encuestas para establecer alguna aproximación de los valores s^2 o s_d^2 o los correspondientes coeficientes de variación cv, cv_d sobre los que basar el diseño muestral y establecer los tamaños de muestra.

Capítulo 17
Otros aspectos del muestreo

1. Muestreo en dos fases

1. Habrá ocasiones en que el conocimiento previo que se dispone del universo objeto de estudio es muy limitado e insuficiente para proceder a una estratificación eficiente o para la utilización de estimadores del tipo de razón que nos permitan importantes reducciones del error estándar. En estos casos puede ser conveniente la realización de una primera muestra, relativamente amplia, con el objeto de estimar aquellas características básicas que nos sirvan para la utilización posterior de muestreo estratificado o de estimadores de razón. Una vez determinadas las características del universo que sean de interés, se selecciona en una segunda fase una submuestra de la primera sobre la que ya se estudian propiamente las variables objeto de estudio. Este proceso se conoce como *muestreo doble o muestreo en dos fases*. El proceso se justifica si la información obtenida en la primera fase permite una reducción de muestra en la segunda fase que compense costes.

2. La muestra correspondiente a la primera fase se denomina también *muestra censal*, muestra maestra o censo muestral. Estas denominaciones indican un primer proceso de muestreo sustitutivo de un censo completo, es decir, cuyo fin es conocer características poblacionales, incluso el propio tamaño del universo N, necesarios para el posterior diseño de la muestra. Este procedimiento censal en base a una muestra no debe sorprender: es práctica habitual en grandes operaciones censales proporcionar resultados basados en una muestra de los cuestionarios censales en lugar de utilizar la información completa del censo total. La muestra en

segunda fase puede denominarse muestra principal o *muestra de estudio*, ya que es la muestra sobre la que se miden las variables objeto de estudio.

2. Muchas variables de estudio

1. Cuando se estudia la teoría de muestras siempre se habla de la variable de estudio X. Sin embargo cuando se selecciona una muestra van a ser muchas variables X las que se estudien en cada unidad muestral, lo que significa que la muestra va a proporcionar multitud de estimaciones cada una con su propio nivel de error estándar, es decir, no puede hablarse de la calidad global de una muestra, sino que cada estimación que proporcione, tendrá su propio error de muestreo. Previamente habrá que haber definido un tamaño de muestra en función de un cierto error estándar. Si quisiéramos el mismo nivel de error estándar para cada variable en estudio resultarían tamaños de muestra diferentes para cada una, lo cuál, desde un punto de vista práctico no tiene sentido. Lo normal será que entre las variables a estudiar haya unas pocas de mayor importancia y sean éstas las que predominen en la determinación del tamaño de muestra, llegándose a una solución de compromiso. Un problema similar surge al establecer la distribución óptima de una muestra estratificada para distintas variables a estudiar: cada variable nos puede proporcionar afijaciones diferentes y debe llegarse a una solución única.

2. El concepto de error de muestreo surge porque al tomar cientos o miles de muestras independientes de una población para estimar un parámetro, las estimaciones presentan una variabilidad aleatoria que puede aproximarse por la distribución normal. En una forma análoga se puede pensar que cuando una muestra proporciona cientos, miles de estimaciones se pueden aplicar las propiedades de la distribución normal y pensar que, por ejemplo, un 5% de las estimaciones quedan fuera de su intervalo de confianza ($\pm$ 2 veces el error estándar), es decir, alejadas de la realidad, sin que pueda saberse cuales son: es el analista de los resultados el que con su conocimiento y experiencia puede separar, quizá no totalmente, aquellos datos que reflejen la realidad de aquellos otros que pueden ser debidos a variaciones extremas de muestreo o a sesgos introducidos en la muestra, no importantes para muchas de las variables investigadas pero que sí lo son para otras.

3. Muestreo repetido de la misma población

1. En la actualidad es práctica común la de utilizar muestras para recoger series de datos sobre la misma población que se publican a intervalos regulares de tiempo. Ejemplos de ello los tenemos en las encuestas de población activa o de fuerza de trabajo que realizan los países desarrollados, los paneles de audiencia de televisión, muestras continuas de hogares o de tiendas para medir el consumo, ...

2. Cuando la misma población se muestrea repetidamente en el tiempo, estamos en una posición ideal para obtener estimadores realistas de costes y varianzas y, en consecuencia, para aplicar técnicas que conducen a una utilización óptima del muestreo. Una cuestión importante en muestreo repetido es con qué frecuencia y de qué manera debe cambiarse la muestra a lo largo del tiempo. Podemos optar entre las siguientes alternativas:

a) Utilizar la misma muestra, llamada *panel*, en cada repetición del muestreo o periodo.

b) Mantener en cada periodo una proporción π_c de muestra común con el periodo anterior, renovando el resto de la muestra.

c) Utilizar en cada periodo muestras independientes.

3. Hay muchas consideraciones que afectan a la decisión. Los entrevistados pueden negarse a dar la misma información una y otra vez. Los que responden pueden influirse por la información que reciben durante las entrevistas lo que contribuye a introducir paulatinamente sesgos en la muestra y, en este caso, suele decirse que la muestra se contamina con el tiempo. Otras veces puede haber mejor cooperación en segunda y sucesivas tomas de información. Si conseguir la colaboración de una unidad muestral implica un coste relativamente alto respecto a la toma de información puede ser aconsejable utilizar la misma muestra o una alta proporción de muestra común.

4. Con los datos de muestras sucesivas de la misma población hay tres clases de cantidades a estimar y, en cada caso, la política de renovación de la muestra es diferente si deseamos maximizar la precisión:

– Si deseamos estimar el cambio de un periodo al siguiente o de un año al mismo periodo del año anterior, es mejor retener la misma muestra.

– Para estimar el valor promedio sobre varios periodos, es mejor tomar muestras independientes en cada periodo.

– Si nuestro interés se centra en el valor para el periodo más reciente, entonces se obtiene la misma precisión conservando la misma muestra o cambiándola en cada periodo; el cambio parcial de parte de la muestra puede ser mejor que cualquiera de estas alternativas.

5. Lo anterior es consecuencia de la correlación positiva entre las medidas de la misma unidad en dos periodos consecutivos. Al mantener la muestra constante en periodos consecutivos, existe una alta correlación entre los datos de las unidades muestrales en ambas ocasiones, lo que hace que los errores en las estimaciones tiendan a permanecer en la misma dirección (es decir, si el error es +2,5% en el primer periodo, puede ser +1,5% en el siguiente, pero difícilmente será -3%), lo que hace que los cambios se midan con menor error absoluto que las estimaciones individuales de cada periodo.

6. En muestreo repetido de la misma población puede tener total sentido la dedicación de parte de los recursos a lo que anteriormente se ha indicado como primera fase del muestreo o censo muestral ya que su coste se amortiza sobre varias realizaciones de la muestra objetivo. En estudios periódicos en el tiempo esta primera fase censal se vuelve imprescindible si el universo o población que se pretende estudiar cambia en el tiempo y no se dispone de información sobre su evolución: en estos casos resulta necesario realizar estudios censales periódicos (cada cinco, dos años, o de forma continua) para preservar de sesgos a la muestra de estudio. Lógicamente, la muestra de estudio, aunque se pretenda constante en el tiempo, estará afectada por la propia evolución de la población y será necesario introducir cambios paulatinos en la misma para su adaptación al carácter cambiante y evolutivo de la misma.

7. Cuando se muestrean poblaciones con un alto grado de asimetría ya se vio la importancia del muestreo estratificado para la precisión. En estos casos la varianza por estrato suele aumentar con el valor de la variable de estudio (tamaño de la unidad) de forma que la afijación óptima es la única garantía para que el factor de expansión de las unidades grande o muy grandes se mantenga dentro de límites razonables. Pensemos que en cualquier proceso de muestreo, el total poblacional se estima aplicando a cada unidad muestral un factor de expansión F_i, de forma

que el total estimado es $\hat{X} = \sum_{1}^{n} X_i F_i$. La cantidad $\frac{X_i F_i}{\hat{X}}$ es la contribución de la i-ésima unidad muestral a la estimación y es la misma para la estimación del total que para la media. Con muestreo aleatorio o con afijación proporcional F_i es igual para todas las unidades muestrales y la contribución depende del valor X_i: valores muy altos van a resultar en contribuciones muy altas y estimaciones con alto error de muestreo y, por tanto, poco fiables. Resulta intuitivo que cuanto mayor es X_i menor debe ser F_i con el fin de preservar a la estimación final de contribuciones extremas debidas a una sola o unas pocas unidades: no parecería muy fiable una estimación obtenida con una muestra de 100 unidades (100 sumandos), de las cuales una sola de ellas represente el 80% del total estimado, cuando cada sumando en promedio contribuya con un 1%. La afijación óptima es la única garantía para evitar estos problemas.

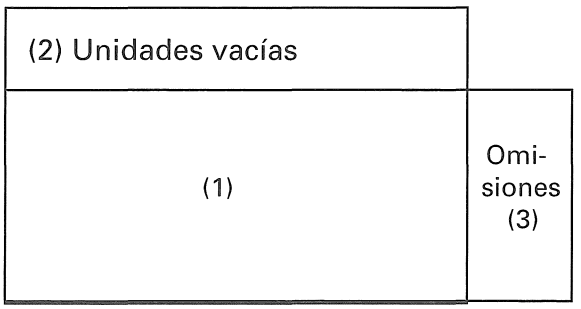
Capítulo 18
Errores ajenos al muestreo

1. Introducción

1. Hasta ahora hemos supuesto que 1) la población marco coincide con la población objetivo, 2) que la muestra real alcanzada se corresponde con la muestra inicialmente planificada y seleccionada probabilísticamente y 3) que la información obtenida en cada unidad muestral es correcta. En estas condiciones la única fuente de error del estimador es el error de muestreo que es la variación aleatoria que se presenta cuando se miden n de las unidades en lugar de la población completa N. Lamentablemente esta situación ideal no se da con frecuencia en la práctica y debemos asumir la presencia de otros errores, que se presentan cuando no se cumple cualquiera de los tres supuestos mencionados y que se agrupan bajo el nombre de *errores ajenos al muestreo*.

2. Errores de cobertura

1. Cuando la población marco no coincide con la población objetivo tenemos los llamados *errores de cobertura*. Recordemos que la población marco es la población que sirve de base para la selección de la muestra. Podemos pensar en un listado del que se selecciona la muestra: puede haber unidades de la población objetivo no contenidas en el listado (omisiones) o puede haber unidades en el listado que no se corresponden con la población objetivo (unidades vacías), incluso el listado puede contener unidades duplicadas:



(1)+(2) = población marco
(1)+(3) = población objetivo

2. Con la muestra seleccionada de la población marco podremos estimar la proporción de unidades (1) y hacer que los resultados estimados se refieran al universo (1), parte coincidente entre la población marco y la población objetivo, pero no a la parte (3), conjunto de unidades omitidas en el listado. Una solución para disminuir errores de cobertura puede ser la utilización de varios listados. No obstante, si las proporciones (2) y (3) son altas será necesario utilizar conjuntamente una muestra de la lista junto con otro procedimiento de selección, por ejemplo áreas, que nos permita acceder a la parte (3). Una muestra en primera fase nos puede servir para determinar estimaciones de (1) y (3) y por tanto de la población objetivo.

3. Los problemas de cobertura no son exclusivos de la utilización de listas. Pensemos en un muestreo por áreas en una ciudad en el que se parte de planos o mapas incompletos: manzanas, urbanizaciones o barrios de reciente construcción pueden quedar omitidos del marco.

3. Falta de respuesta

1. Cuando la muestra real alcanzada no se corresponde con la muestra inicialmente planificada, es decir, no se obtiene información en todas las unidades de la muestra, decimos que existe *falta de respuesta* o no respuesta. Aparte la no respuesta por unidades omitidas en el marco, ya mencionada, la falta de respuesta puede agruparse en dos principales tipos:

a) *No localizado* o falta de contacto, que puede ser debido a:

a1) Ausencia temporal durante las horas de entrevista (no-en-casa). Es conocido que familias en las cuales ambos padres trabajan y las familias sin niños son más difíciles de alcanzar que familias con niños pequeños o con personas jubiladas.

a2) Viaje, vacaciones.

a3) Enfermedad.

a4) Problemas de lenguaje.

a5) Movilidad geográfica: cambio de dirección o domicilio, cambio de ciudad.

a6) Falta de motivación o experiencia en el entrevistador para contactar con el entrevistado. Está comprobado que las tasas de no respuesta varían por entrevistador.

a7) Barrio o vecindad *difícil*.

b) *Negativa* a colaborar, debido a:

b1) Falta de tiempo.

b2) Falta de motivación o de interés por el tema de la encuesta.

b3) No desea que el entrevistador conozca sus respuestas u opiniones.

b4) No desea estar *registrado*.

b5) Cansancio de las entrevistas.

b6) Cuestionario demasiado largo, preguntas complicadas, preguntas que rozan la intimidad.

b7) Los *hueso duro*. Personas que cerradamente rechazan ser entrevistadas o están sistemáticamente fuera de casa durante el tiempo disponible para el trabajo de campo.

b8) Falta de habilidad del entrevistador para conseguir la colaboración. Vale aquí el comentario de a6): hay entrevistadores que consiguen mejores tasa de respuesta que otros.

b9) La colaboración es, finalmente, voluntaria: *busque a otro que yo no puedo ahora.*

2. A estos dos grupos de no respuesta puede añadirse la falta de respuesta parcial: el entrevistado no responde a parte de las preguntas porque no tiene la información o, simplemente, no está dispuesto a facilitarla.

3. Para evaluar los efectos de la falta de respuesta conviene pensar en la población dividida en dos estratos: en el primero se incluyen todas las unidades para las cuales se obtendrían mediciones si caen en la muestra y en el segundo se incluyen las unidades para las que no se obtendrían mediciones. La muestra no proporciona información del estrato 2, lo cuál no sería un problema si se pudiera suponer que las características que se miden en el muestreo son las mismas en el estrato 2 que en el estrato1. Desde el momento que esto no sea así estaremos en presencia de un sesgo causado por la falta de respuesta. El problema es que al no disponer de información del estrato que no responde el tamaño del sesgo es desconocido.

4. La falta de respuesta no debe ignorarse o pensar que se corrige sustituyendo en la muestra a los que no responde por otros que sí colaboren, ya que ello no va eliminar el sesgo, simplemente nos mantiene el tamaño de muestra. Por el contrario hay que ser conscientes de que la no respuesta va a ocurrir y asignar, en lo posible, algunos recursos y disponer de algunas estrategias para reducir su proporción. Algunos procedimientos para reducir la no respuesta son:

1) Cartas y llamadas telefónicas por adelantado.

2) Dar algún incentivo por la colaboración.

3) Programar visitas repetidas puede ser de gran efectividad para reducir los no-en-casa.

4) Mejora de los procedimientos de recogida de información. Si la información se recoge por entrevista personal el entrenamiento del entrevistador es fundamental: la interacción positiva entrevistador-entrevistado es básica para el éxito de la entrevista, lo cuál puede requerir que el entrevistador disponga de distintas estrategias para afrontar la entrevista en función de ciertas características observables de los encuestados. Preservar la intimidad del entrevistado puede favorecer el dejarle el cuestionario para que lo rellene y envíe posteriormente por correo, aunque se haya tenido un primer contacto personal para obtener la colaboración. Otro aspecto a tener en cuenta es que cuanto más activa (más tiempo requiere) sea la colaboración de la unidad muestral menor es su disposición a colaborar: pensemos en

un panel de audiencia de TV en el que el hogar debe rellenar y enviar por correo un largo y tedioso cuestionario sobre qué ha visto cada día en relación con la instalación de un audímetro conectado al televisor que registra y transmite lo que el televisor emite en cada momento; la colaboración del hogar en el caso del audímetro es mucho más pasiva (menos molestia), lo cuál favorece la colaboración.

5. En la práctica y a pesar de las medidas que se tomen será imposible, en general, reducir la no respuesta a cero por lo que se hace imprescindible su medición y control. Un primer aspecto en este sentido es cuantificar la tasa de no respuesta según distintas causas. Ello puede ayudar para reducir las tasas de no respuesta en encuestas posteriores. En ocasiones será posible recoger ciertas características observables de las unidades no respuesta que puedan ser utilizadas posteriormente en procedimientos de ajuste para remover los sesgos de no respuesta en las estimaciones finales.

6. Normalmente, además de las variables que hayan servido para la estratificación del universo se dispone de información poblacional de otras características que pueden servir para controlar la *microrrepresentatividad* final de la muestra obtenida, comparando los valores poblacionales de estas variables conocidas con los estimados por la muestra. Este control de microrrepresentatividad es fundamental en presencia de falta de respuesta y nos puede ayudar a determinar ciertas características del estrato de no respuesta Las desviaciones que se producen pueden utilizarse para modificar los factores de expansión originales de cada unidad muestral, en un proceso iterativo, hasta conseguir que los valores *estimados* coincidan con los conocidos en el Universo para las distintas variables incluidas en el proceso. Este proceso iterativo de ajuste en los factores originales de expansión se conoce también como *equilibraje* de la muestra y puede contribuir a remover sesgos introducidos en la muestra final, en la medida en que las variables objeto de investigación puedan estar correlacionadas con las variables que intervienen en el proceso de equilibraje.

4. Errores de medida

1. Un tercer tipo de error no de muestreo se produce por *errores de medición* y errores que se introducen en la producción de los resultados de una encuesta. Estos errores suceden cuando el valor medido X^* (o el utilizado para la estimación)

no se corresponde con el valor real X. Se conocen también por *errores de respuesta* y pueden ser varias las causas que los producen:

1) Instrumentos de medición (cuestionario–entrevistador) inadecuados o sujetos a error.

2) Fallos de memoria. El entrevistado responde lo que él cree que hizo, pero no lo qué realmente hizo.

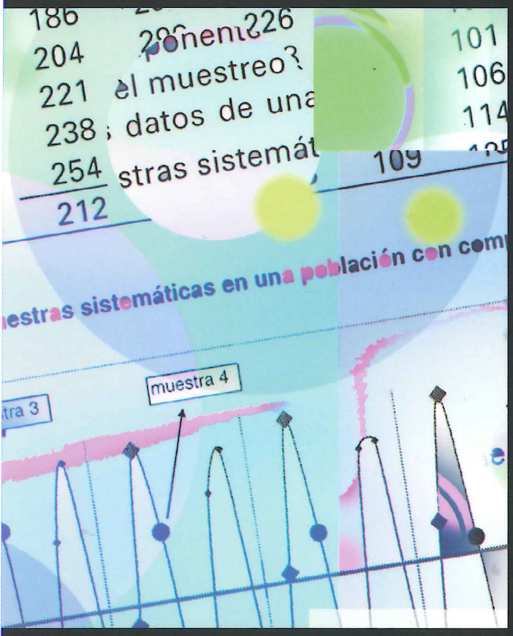
3) El entrevistado da una respuesta falsa, bien inducido por el entrevistador (quizá por el cuestionario), o bien porque no desea que *su verdad* quede registrada (*qué dirán...)*

4) Olvido. Por ejemplo en un panel de hogares el hogar colaborador olvida anotar algunas compras en el diario o en un panel de audímetros una persona olvida identificarse.

5) Falta de información. El informante no dispone de toda la información para contestar y da una respuesta aproximada.

6) Errores de codificación y grabación que introducen en el proceso un valor erróneo con independencia de que el valor original fuera correcto o no.

2. Como comentario final hay que decir que al planear un estudio por muestreo debe prestarse especial atención a los errores no de muestreo que pueden presentarse en cualquier fase del trabajo y, si son importantes, incluso invalidar los resultados. Por otra parte detectarlos y cuantificarlos no es tarea fácil. Sólo la anticipación y el análisis cuidadoso de cada paso en el proceso de muestreo y de los resultados pueden ayudar. Los errores de muestreo desde el momento que pueden ser evaluados y estimados dejan de tener importancia. El error de muestreo se constituye en una medida de la calidad del diseño teórico de la muestra pero no mide la calidad real, afectada por los errores no de muestreo.



"El *Manual básico de Estadística* se ha elaborado para que sea un instrumento de gran utilidad para todas aquellas personas que sin tener una formación estadística académica, participan en los procesos de producción de la información estadística y tienen interés por conocer los conceptos y técnicas estadísticas básicas que se utilizan en los mismos.

El *Manual* describe de forma resumida, las principales fases que necesariamente hay que abordar en la realización de operaciones estadísticas, para centrarse después en una explicación sencilla de las principales definiciones y conceptos estadísticos, y describir con brevedad las técnicas de muestreo que suelen aplicarse. Su presentación minuciosa y ordenada permitirá al lector entender progresivamente las principales ideas que subyacen tras la terminología estadística".

Colección Libros de autor.

ISBN 978-84-260-3741-1

INSTITUTO NACIONAL DE ESTADÍSTICA