



# **Documentación de apoyo a la preparación de la oposición al Cuerpo General Administrativo de la Administración del Estado, especialidad Estadística**

## **Módulo 3: Probabilidad y Muestreo**



## ÍNDICE


<b>Unidad 9. Fenómenos aleatorios. Conceptos de probabilidad. Propiedades. Probabilidad Condicionada. Independencia de sucesos. Teorema de la Probabilidad Total y Teorema de Bayes.</b>	<b>5</b>
9.1 Fenómenos aleatorios	5
9.2 Conceptos de probabilidad	8
9.2.1 Propiedades	9
9.3 Probabilidad Condicionada	10
9.4 Independencia de sucesos	11
9.5 Teorema de la Probabilidad Total	12
9.6 Teorema de Bayes	13
 <b>Unidad 10. Variables aleatorias. Variables discretas, función de probabilidad. Distribución Binomial. Variables continuas, función de densidad. La distribución Uniforme. La distribución Normal.</b>	 <b>14</b>
10.1. Variables aleatorias	14
10.2. Variables discretas	15
10.2.1 Función de probabilidad	15
10.2.2 Distribución Binomial	17
10.3. Variables continuas	19
10.3.1 Función de densidad	19
10.3.2 La distribución Uniforme	21
10.3.3 La distribución Normal	22



<b>Unidad 11. Formas de investigar una población: censos, encuestas, registros administrativos y big data: ventajas e inconvenientes. ....</b>	<b>25</b>
11.1 Introducción.....	25
11.2 Formas de investigar una población .....	25
11.2.1 Censos y Encuestas. Ventajas e inconvenientes.....	26
11.2.2 Registros administrativos. Definición. Ventajas e inconvenientes.....	28
11.2.3 Big Data. Ventajas e inconvenientes .....	30
 <b>Unidad 12. Conceptos de población, marco y muestra. Tipos de muestreo. Concepto de estimador y sus principales propiedades. Errores de muestreo. Errores ajenos al muestreo: de cobertura, de falta de respuesta y de medida. ..</b>	<b>32</b>
12.1 Conceptos población y muestra.....	32
12.2 Marco. Definición y principales características .....	33
12.3 Tipos de muestreo .....	34
12.4 Concepto de estimador.....	35
12.5 Errores de muestreo e insesgadez .....	36
12.5.1 Sesgo en las estimaciones.....	37
12.5.1 Error de muestreo: Precisión.....	37
12.6 Errores ajenos al muestreo.....	38
12.6.1 Errores de cobertura.....	39
12.6.2 Falta de respuesta.....	40
12.6.3 Errores de medida .....	42



<b>Unidad 13. Muestreo aleatorio simple: estimadores del total y de la media y sus errores de muestreo. Tamaño de la muestra. Otros métodos de muestreo probabilístico: estratificado, por etapas y sistemático: concepto y razones principales para su aplicación. ....</b>	<b>44</b>
13.1 Muestreo aleatorio simple.....	44
13.1.1 Estimadores del total y de la media y sus errores de muestreo.....	45
13.1.2 Tamaño de la muestra.....	47
13.2 Otros métodos de muestreo probabilístico.....	48
13.2.1 Muestreo estratificado.....	49
13.2.2 Muestreo de Conglomerados sin submuestreo.....	50
13.2.3 Muestreo por etapas: Muestreo de Conglomerados con submuestreo .....	51
13.2.4 Muestreo Sistemático .....	52



## Unidad 9. Fenómenos aleatorios. Conceptos de probabilidad. Propiedades. Probabilidad Condicionada. Independencia de sucesos. Teorema de la Probabilidad Total y Teorema de Bayes.

### 9.1 Fenómenos aleatorios

Cualquier disciplina científica tiene, entre sus principales fines, ser capaz de establecer leyes que describan los fenómenos observados que pertenecen a su parcela de estudio. Cuando conociendo todos los factores de un experimento o fenómenos somos capaces de predecir con exactitud el resultado del mencionado experimento, estamos ante lo que se conoce con el nombre de **fenómenos deterministas**. Por ejemplo, si lanzamos una pelota de fútbol desde una altura determinada en condiciones de vacío es posible saber el tiempo que tardará en llegar al suelo, etc...

Pero en otras ocasiones y aparentemente bajo el mismo conjunto de condiciones iniciales, no es posible predecir el resultado exacto de lo que va a ocurrir en una experiencia concreta. Ejemplos de ello son:

1. Si se lanza una moneda equilibrada al aire, sabemos que caerá pero, no sabemos a priori, si caerá del lado de la cara o de la cruz.
2. Si se colocan diversas bolas numeradas, de las mismas dimensiones, en una bolsa y se extrae una de ellas, es imposible predecir cuál será la bola numerada seleccionada.
3. Si nos preguntan por cuál será el equipo vencedor del próximo mundial de fútbol, podemos decir que equipos son favoritos por distintos factores pero no sabemos exactamente quién ganará a priori.

Con esto queremos decir que hay ocasiones en las que podemos conocer con certeza el resultado concreto que se va a producir pero, en otras muchas, no siempre es posible determinar que va a suceder con exactitud. A esos experimentos, en los que no somos capaces de prever el resultado, se les denomina **fenómenos aleatorios**. El objeto de estudio del Cálculo de Probabilidades son precisamente los fenómenos aleatorios.



A continuación vamos a enunciar una serie de propiedades que caracterizan a los fenómenos aleatorios:

1. Todos los posibles resultados asociados al experimento se conocen previamente.
2. Un resultado concreto no es predecible dadas unas mismas condiciones.
3. Existe una regularidad estadística, es decir, si se dan las mismas condiciones en que se produce el fenómeno, cuando se repite muchas veces se observa que cada resultado se obtiene en un porcentaje estable de casos. Por ejemplo, si lanzo una moneda tres veces, las tres me podría salir cara (es decir un 100%), pero si la lanzo 1.000 veces, entonces saldrá cara entorno al 50% de las veces y, este porcentaje se mantiene si se va incrementando el número de tiradas.

Así pues, nuestra intención es describir lo que puede suceder cuando se produce un fenómeno a través de probabilidades. Es decir, asignar una probabilidad a cada uno de los acontecimientos posibles, que informe de la frecuencia con que hay que esperar que se presente cada uno, después de numerosas observaciones del fenómeno.

### **Espacio Muestral**

Dado un experimento aleatorio, se llama espacio muestral al conjunto de todos los resultados posibles de la experiencia aleatoria. En adelante lo designaremos por  $\Omega$ .

Obsérvese que asociados a un mismo experimento aleatorio pueden considerarse diferentes espacios muestrales; así en el lanzamiento de un dado, un posible espacio muestral podría ser  $\Omega_1 = \{\text{par}; \text{impar}\}$ , otro  $\Omega_2 = \{\text{menor que } 4; \text{mayor o igual que } 4\}$ , y otro  $\Omega_3 = \{1; 2; 3; 4; 5; 6\}$ . Todo depende de lo que queramos estudiar en nuestro experimento aleatorio y de la facilidad que tengamos en asignar la probabilidad a los sucesos del espacio muestral que consideremos.

### **Sucesos**

Se llama suceso a cualquier subconjunto del espacio muestral. En el caso del lanzamiento del dado un posible suceso será "que salga tres". Otro posible suceso podría ser: "que salga par". O, por ejemplo, "que salga mayor que 4".

Como vemos, se pueden considerar sucesos que incluyen a otros sucesos ("que salga par" contiene a los sucesos "que salga dos", "que salga cuatro" y "que salga seis"). Por lo tanto, podemos definir dos tipos de sucesos:



- **Sucesos elementales:** sucesos formados por un solo elemento del espacio muestral (por ejemplo, que salga dos).
- **Sucesos compuestos:** sucesos formados por más de un elemento del espacio muestral (por ejemplo, que salga par).

### Suceso Imposible y Suceso seguro

- Suceso imposible es aquel que no puede ocurrir nunca. Se usa el símbolo de conjunto vacío  $\phi$  para denotarlo.
- Suceso seguro, está formado por todos los elementos del espacio muestral, luego es aquel suceso que ocurre siempre. Es lo contrario del suceso imposible.

### Operaciones con sucesos

Sea  $\Omega$  el espacio muestral correspondiente a una experiencia aleatoria. Sean  $A$  y  $B$  dos sucesos cualesquiera de  $\Omega$ . Se definen las siguientes operaciones con sucesos:

- **Unión de los sucesos  $A$  y  $B$ :** se denota  $A \cup B$ , representa otro suceso que resulta de unir los sucesos elementales de  $A$  con los de  $B$ . El suceso  $A \cup B$  sucede siempre que el resultado pertenezca a  $A$  a  $B$ , o a ambos simultáneamente.
- **Intersección de los sucesos  $A$  y  $B$ :** se denota  $A \cap B$ , es también un nuevo suceso formado por todos los sucesos que son a la vez de  $A$  y  $B$ . Si la intersección de dos sucesos  $A$  y  $B$  es vacía,  $A \cap B = \phi$ , ambos sucesos no pueden ocurrir simultáneamente; se dice entonces que  $A$  y  $B$  son incompatibles.
- **Diferencia de sucesos  $A - B$ :** es un nuevo suceso formado por los sucesos de  $A$  que no son de  $B$ .
- **Complementario de cualquier suceso  $A$ :** se denota  $A^c = \Omega - A$ , es el suceso contrario de  $A$  y significa que el resultado del fenómeno no pertenece al suceso  $A$ . También se le suele designar con  $\bar{A}$ .
- Dos sucesos  $A$  y  $B$  pueden estar relacionados de modo que siempre que ocurre  $A$ , ocurre  $B$ ; lo que equivale a que se verifique la relación  $A \subset B$ ; esto es, cualquier elemento de  $A$  pertenece también a  $B$ . Se verifica por tanto que  $A \cap B = A$  y que  $A \cup B = B$ .
- Dado cualquier suceso  $A$ , siempre se cumple que  $\phi \subset A \subset \Omega$ .



- El complementario de la unión es la intersección de los complementarios,  
$$(A \cup B)^c = A^c \cap B^c$$
- El complementario de la intersección es la unión de los complementarios  
$$(A \cap B)^c = A^c \cup B^c$$

**Ejercicio 9.1:** En el conjunto de los números naturales menores de 12 (excluido el 0) se definen los subconjuntos  $A = \{\text{números menores de 6}\}$  y  $B = \{\text{números primos menores de 12}\}$ . Determina los elementos que forman los siguientes conjuntos.

- a)  $A \cup B$
- b)  $A \cap B$
- c)  $A^c$
- d)  $(A \cap B)^c$
- e)  $A^c \cap B^c$

*Nota: el 1 no se considera número primo*

**Solución 9.1:**


En primer lugar el subconjunto A está formado por los siguientes elementos  $A = \{1, 2, 3, 4, 5\}$  y el  $B = \{2, 3, 5, 7, 11\}$ . Teniendo en cuenta esto pasamos a resolver los distintos apartados:

- a)  $A \cup B = \{1, 2, 3, 4, 5, 7, 11\}$
- b)  $A \cap B = \{2, 3, 5\}$
- c)  $A^c = \{6, 7, 8, 9, 10, 11\}$
- d)  $(A \cap B)^c = \{1, 4, 6, 7, 8, 9, 10, 11\}$
- e)  $A^c \cap B^c = \{6, 8, 9\}$

## 9.2 Conceptos de probabilidad

El concepto de probabilidad admite varias definiciones, según el punto de partida que se tome. La definición que adoptaremos será la definición axiomática que planteó el matemático ruso Kolmogorov.

La definición axiomática supone definir la probabilidad como cualquier asignación de un número a cada posible suceso del espacio muestral, el cual representa el grado de credibilidad de que vaya a ocurrir dicho suceso y, además, debe verificarse una serie de axiomas.



**Definición:** Sea  $\Omega$  un espacio muestral. Sea  $S$  el conjunto de todos los sucesos de  $\Omega$ . Llamaremos probabilidad a toda aplicación  $P$  definida entre  $S$  y el conjunto de los números reales  $\mathbb{R}$

$$P: S \rightarrow \mathbb{R}$$

que verifica los axiomas siguientes:

- **Axioma 1:** para cualquier suceso  $A$  del espacio muestral, su probabilidad es un número no negativo,  $P(A) \geq 0$
- **Axioma 2:** la probabilidad del espacio muestral es 1.  $P(\Omega) = 1$
- **Axioma 3:** Si dos sucesos son incompatibles, la probabilidad de su unión es igual a la suma de probabilidades.

$$A \cap B = \phi \Rightarrow P(A \cup B) = P(A) + P(B)$$

### 9.2.1 Propiedades

Las propiedades que se enuncian a continuación se pueden demostrar fácilmente haciendo uso de los axiomas anteriores.

1. Para cualquier suceso  $A$  y  $A^c$ , su *complementario*,  $P(A^c) = 1 - P(A)$

**Demostración:** Por ser sucesos complementarios,  $A \cap A^c = \phi$ . Y por la misma razón,  $A \cup A^c = \Omega$

Según el *axioma 3*,  $A \cap A^c = \phi \Rightarrow P(A \cup A^c) = P(A) + P(A^c)$

Según el *axioma 2*,  $P(\Omega) = 1$ ,  $P(\Omega) = P(A \cup A^c) = P(A) + P(A^c) = 1$

Y de esta última igualdad se deduce de manera inmediata que  $P(A^c) = 1 - P(A)$

2.  $P(\phi) = 0$

**Demostración:** El suceso contrario de  $\phi$ , es el denominado suceso seguro,  $\Omega$ . Así teniendo en cuenta la *propiedad 1*,  $P(\phi) = 1 - P(\phi^c) = 1 - P(\Omega)$ . Y aplicando el *axioma 2*  $P(\phi) = 1 - P(\Omega) = 1 - 1 = 0$

3. Si  $A \subset B \subset \Omega$  entonces  $P(B - A) = P(B) - P(A)$

**Demostración:** Si  $A \subset B$  entonces  $B = A \cup (B - A)$ . Los sucesos  $A$  y  $B - A$  son incompatibles,  $A \cap (B - A) = \phi$ . Por tanto, teniendo en cuenta el axioma 3

$$P(B) = P(A \cup (B - A)) = P(A) + P(B - A)$$

4. Si  $A \subset B \subset \Omega$  entonces  $P(A) \leq P(B)$
5. Si  $A$  y  $B$  son sucesos compatibles,  $A \cap B \neq \phi$ , entonces la probabilidad de la unión

$$P(A \cup B) = P(A) + P(B) - P(B \cap A)$$

### Regla de Laplace

Aunque hemos definido la probabilidad por medio de la definición axiomática, la regla de Laplace, es tal vez la primera definición de probabilidad que se estableció. El problema que plantea es que esta regla sólo resulta válida cuando todos los sucesos elementales del espacio muestral tienen la misma probabilidad, es decir, son equiprobables. Por ello, no es una definición que podamos generalizar a todas las experiencias aleatorias.

**Definición:** Sea un espacio muestral  $\Omega = \{a_1, a_2, \dots, a_n\}$ . Y sea  $A$  un suceso de  $\Omega$ . Si todos los sucesos elementales son equiprobables, la Regla de Laplace dice:

La probabilidad del suceso  $A$  es igual al cociente entre el número de resultados favorables y el número de resultados posibles. Así,

$$P(A) = \frac{\text{Número de casos favorables al suceso } A}{\text{Número de casos posibles}(n)}$$

Esta regla es aplicable en juegos de azar tales como el lanzamiento de un dado o una moneda equilibrada,...

## 9.3 Probabilidad Condicionada

Supongamos que disponemos de una bolsa con bolas numeradas del 1 al 12. Si se extrae una bola al azar y se pregunta por la probabilidad de que el número de dicha bola sea par diríamos que es

$$P(\text{ser par}) = \frac{6}{12} = \frac{1}{2}$$

Ahora bien, si como ayuda para identificar el número de la bola, una vez extraída nos dicen que es un número menor que 6, la nueva información obliga a modificar la asignación de probabilidades

$$P(\text{ser par} / \text{n}^\circ \text{menor que seis}) = \frac{2}{5}$$

No es sorprendente que se obtenga un resultado distinto, puesto que la información que se parte en ambos casos es diferente. En los dos casos lo que difiere es la información disponible, no el suceso cuya probabilidad se quiere evaluar.

**Definición:** Sean dos sucesos  $A$  y  $B$  y tal que  $P(B) > 0$ , se denomina probabilidad de  $A$  condicionada por  $B$  a:

$$P(A/B) = \frac{P(A \cap B)}{P(B)}$$

**Ejercicio 9.2:** En un experimento aleatorio se sabe que  $P(A) = 0,7$ ,  $P(B) = 0,35$  y  $P(A \cup B) = 0,75$ . Se pide calcular:

- a)  $P(A \cap B)$
- b)  $P(A/B)$
- c)  $P(A/A \cap B)$

**Solución 9.2:**

$$\text{a) } P(A \cap B) = P(A) + P(B) - P(A \cup B) = 0,7 + 0,35 - 0,75 = 0,3$$

$$\text{b) } P(A/B) = \frac{P(A \cap B)}{P(B)} = \frac{0,3}{0,35} = 0,86$$

$$\text{c) } P(A/A \cap B) = \frac{P(A \cap (A \cap B))}{P(A \cap B)} = \frac{P(A \cap B)}{P(A \cap B)} = 1$$

## 9.4 Independencia de sucesos

No siempre la probabilidad condicionada de un suceso es distinta de la inicial. Para ilustrar esto, volvamos al ejemplo introducido en el apartado de probabilidad condicionada. Supongamos que la bola seleccionada pertenece al suceso  $C = \text{"n}^\circ \text{ menor que siete"}$ . Y nos preguntamos por la probabilidad del suceso  $A = \text{"ser par"}$

$$P(A) = \frac{6}{12} = \frac{1}{2}$$

$$P(A/D) = \frac{3}{6} = \frac{1}{2}$$

Podemos concluir que los sucesos "ser par" y "nº menor que siete" son independientes, en el sentido que saber que la bola seleccionada tiene un número inferior a siete no añade información sobre si es par o no la bola.

**Definición:** Sean dos sucesos  $A$  y  $B$  y tal que  $P(B) > 0$ , se dice que son independientes cuando:

$$P(A/B) = P(A)$$

O equivalentemente

$$A \text{ y } B \text{ independientes} \Rightarrow P(A \cap B) = P(A)P(B)$$

## 9.5 Teorema de la Probabilidad Total

**Teorema:** Sean  $n$  sucesos  $A_1, A_2, \dots, A_n$  incompatibles dos a dos y tales que  $A_1 \cup A_2 \cup \dots \cup A_n = \Omega$ . Entonces para cualquier suceso  $B$  se verifica que:

$$P(B) = P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + \dots + P(A_n)P(B/A_n)$$

**Ejercicio 9.3:** Se tienen dos urnas de bolas. La primera contiene 6 bolas verdes y 2 negras, la segunda, 3 verdes y 5 negras. Lanzamos un dado, si sale 1 o 2 extraemos una bola de la primera urna, si sale 3, 4, 5 o 6 extraemos una bola de la segunda urna. Hallar la probabilidad de que la bola extraída sea verde.

**Solución 9.3:**

- $A_1$ : "Seleccionar Primera urna", si en el dado sale  $\{1,2\}$   $P(A_1)=2/6=1/3$
- $A_2$ : "Seleccionar Segunda urna, Si en el dado sale  $\{3,4,5,6\}$   $P(A_2)=4/6=2/3$
- $B$  "seleccionar bola verde"  $P(B/A_1)=6/8$  y  $P(B/A_2)=3/8$

$$P(B) = P(A_1)P\left(\frac{B}{A_1}\right) + P(A_2)P\left(\frac{B}{A_2}\right) = \frac{1}{3} \cdot \frac{6}{8} + \frac{2}{3} \cdot \frac{3}{8} = \frac{12}{24} = \frac{1}{2} = 0,5$$

## 9.6 Teorema de Bayes

Consideremos ahora, bajo las mismas condiciones del teorema anterior, que estamos interesados en conocer la probabilidad de que ocurrido el suceso B la causa que lo haya producido sea la  $A_i$ . Expresado analíticamente, queremos calcular  $P(A_i/B)$ .

Es decir, el Teorema de Bayes estudia lo que suele llamarse probabilidades a posteriori.

**Teorema:** Sean  $n$  sucesos  $A_1, A_2, \dots, A_n$  con probabilidades mayores que cero, incompatibles dos a dos y tales que  $A_1 \cup A_2 \cup \dots \cup A_n = \Omega$ . Sea  $B$  un suceso con  $P(B) > 0$  para el que se conocen las probabilidades  $P\left(\frac{B}{A_1}\right), \dots, P\left(\frac{B}{A_n}\right)$ . Entonces

$$P(A_i/B) = \frac{P(A_i)P(B/A_i)}{P(A_1)P(B/A_1) + P(A_2)P(B/A_2) + \dots + P(A_n)P(B/A_n)}$$

Para cualquier  $i$  que toma valores entre 1 y  $n$

**Ejercicio 9.4:** En una fábrica tenemos 3 máquinas. La  $A_1$  produce 200 piezas, con el 4% de defectuosas, la  $A_2$  produce 300 con el 5% de defectuosas y la  $A_3$  produce 400 con el 2% de defectuosas, al final del día se toma una pieza al azar. 1) ¿Cuál es la probabilidad de que sea defectuosa? 2) ¿Cuál es la probabilidad de que siendo defectuosa proceda de la máquina  $A_1$ ?

### Solución 9.4:

Denotamos por  $A_i$ ="piezas de la máquina  $i$ " y por  $D$ ="ser pieza defectuosa"

- Probabilidades a priori:  $P(A_1) = \frac{200}{900} = \frac{2}{9}; P(A_2) = \frac{3}{9}; P(A_3) = \frac{4}{9}$
- Probabilidades Condicionadas:  $P(D/A_1) = 0,04; P(D/A_2) = 0,05; P(D/A_3) = 0,02$
- Teorema de la probabilidad total:  $P(D) = \frac{2}{9}0,04 + \frac{3}{9}0,05 + \frac{4}{9}0,02 = 0,0344$
- Teorema de Bayes  $P(A_1/D) = \frac{\frac{2}{9}0,04}{0,0344} = 0,2584$

## Unidad 10. Variables aleatorias. Variables discretas, función de probabilidad. Distribución Binomial. Variables continuas, función de densidad. La distribución Uniforme. La distribución Normal.

### 10.1. Variables aleatorias

Consideremos el experimento aleatorio consistente en lanzar una moneda en tres ocasiones. El espacio muestral de este experimento sería el siguiente, teniendo en cuenta que denotamos por "X" si en el lanzamiento se ha obtenido cruz y "C" si se ha obtenido cara:

$$\Omega = \{XXX, XXC, XCX, CXX, CCX, CXC, XCC, CCC\}$$


A cada uno de los resultados posibles le podemos asignar un valor numérico, por ejemplo, el número de caras obtenidas.

Resultado	Número Caras
XXX	0
XXC	1
XCX	1
CXX	1
CCX	2
CXC	2
XCC	2
CCC	3

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. A continuación se introduce la definición formal de lo que es una variable aleatoria.

**Definición:** Una variable aleatoria,  $X$ , es cualquier función que a cada posible resultado de un experimento aleatorio le asigna un número real

$$\begin{aligned} X: \Omega &\rightarrow \mathbb{R} \\ w &\rightarrow X(w) \end{aligned}$$



Donde  $\Omega$  es el espacio muestral asociado a un experimento aleatorio y  $w$  es el resultado de una realización del mismo, se lleva a cabo un proceso de medida y se obtiene un número que denotamos por  $X(w)$ .

Antes de pasar a la clasificación de las variables aleatorias, señalemos que es posible asignar diferentes variables aleatorias a un mismo experimento. Así, por ejemplo, en nuestro ejemplo del lanzamiento en tres ocasiones de una moneda podemos considerar, además, la variable que hace corresponder a cada resultado el número de cruces obtenidas, y otra variable distinta sería la que asigna a cada resultado el valor uno si contiene alguna cruz y cero en caso contrario ...

Según los valores que pueden tomar las variables aleatorias se clasifican en:

- **Variables discretas:** toman un número finito o a lo sumo un número infinito numerable de valores (contable).
- **Variables continuas:** toman una cantidad infinita no numerable de valores.

## 10.2. Variables discretas

Acabamos de ver en el apartado anterior cuando decimos que una variable aleatoria es discreta.

Como ejemplos de variables aleatorias discretas podríamos mencionar:

- El ejemplo con el que comenzábamos la unidad, la variable aleatoria  $X$  que asigna el número de caras en el experimento de lanzar tres veces una moneda equilibrada. Dicha variable puede tomar los valores:  $\{0, 1, 2, 3\}$ .
- En un supermercado podemos definir  $X$  como la variable que mide el número de clientes que visitan el supermercado por día. Dicha variable toma valores en los números naturales unión el cero,  $\mathbb{N} \cup \{0\} = \{0, 1, 2, 3, \dots\}$ .
- En una fábrica, es posible definir una variable aleatoria  $X$ , como el número de trabajadores que sufren un accidente laboral en un año.

### 10.2.1 Función de probabilidad

**Definición:** Se llama función de probabilidad de una variable aleatoria discreta  $X$  al resultado de asignar a cada valor de la variable  $x_i$ , su probabilidad  $p_i$ .



- Cada  $p_i$  es un número comprendido entre 0 y 1:  $0 \leq p_i \leq 1$
- La suma de todos los  $p_i$  es 1:  $p_1 + p_2 + \dots + p_n = \sum_{i=1}^n p_i = 1$

Para definir los parámetros de la media o esperanza matemática y la varianza de una variable discreta se procede como se establece a continuación:

- Esperanza  $\mu = E[X] = x_1p_1 + x_2p_2 + \dots + x_np_n = \sum_{i=1}^n x_ip_i$
- Varianza  $\sigma^2 = V[X] = E[(X - \mu)^2] = (x_1^2p_1 + x_2^2p_2 + \dots + x_n^2p_n) - \mu^2 = \sum_{i=1}^n x_i^2p_i - \mu^2$
- Desviación Típica:  $\sigma = \sqrt{\sigma^2} = \sqrt{(\sum_{i=1}^n x_i^2p_i - \mu^2)}$

**Definición:** La función de distribución es la que asocia a cada número real la probabilidad de que la variable aleatoria tome valores menores o iguales que ese número.

$$F(k) = P[X \leq k] = \sum_{x_i \leq k} p_i$$

Esta función tiene las siguientes propiedades:

1. Es monótona no decreciente ; es decir  $F(x_1) \leq F(x_2)$  siempre que sea  $x_1 < x_2$
2.  $P(x_1 < X \leq x_2) = F(x_2) - F(x_1)$
3. Es continua por la derecha
4.  $F(-\infty) = 0, F(\infty) = 1$

**Ejercicio 10.1:** Con la variable aleatoria  $X$ , cuya función de probabilidad viene dada en la tabla siguiente

$x_i$	1	2	3	4
$p_i$	0,2	a	0,5	0,1

- a) Determine el valor a para que sea función de probabilidad de la variable aleatoria  $X$ .
- b) Calcule la esperanza y la varianza
- c)  $P[X \leq 2,5]$

### Solución 10.1:

- $0,2 + a + 0,5 + 0,1 = 1 \Rightarrow a = 0,2$
- $\mu = E[X] = 1 * 0,2 + 2 * 0,2 + 3 * 0,5 + 4 * 0,1 = 2,5$
- $\sigma^2 = V[X] = 1^2 * 0,2 + 2^2 * 0,2 + 3^2 * 0,5 + 4^2 * 0,1 - 2,5^2 = 0,85$
- $P[X \leq 2,5] = 0,2 + 0,2 = 0,4$

### 10.2.2 Distribución Binomial

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 una moneda), y  $q = 1 - p$  a la de fracaso, 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 Bernoulli. Se corresponde con variables estadísticas dicotómicas en las que las unidades se clasifican según pertenezcan o no a una cierta clase, asignándole el valor 1 si pertenecen y el valor 0 si no pertenecen.

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,  $npq$ .

**Definición:** Partimos de una experiencia dicotómica en la que  $p$  es la probabilidad de éxito. La repetimos  $n$  veces y observamos el número,  $X$ , de éxitos que se consiguen (la variable  $X$ , es discreta porque toma valores  $\{0, 1, \dots, n\}$ ).

La distribución de probabilidad de  $X$  se llama **distribución binomial**  $B(n, p)$ .

La probabilidad de obtener  $k$  éxitos es:

$$P[X = k] = \binom{n}{k} p^k q^{n-k}$$

Se obtiene así la siguiente **distribución de probabilidad**:

Variable X	0	..	k	..	n
Probabilidad $P[X = k]$	$\binom{n}{0} p^0 q^n$		$\binom{n}{k} p^k q^{n-k}$		$\binom{n}{n} p^n q^0$

Los parámetros de esta distribución son:

- **Esperanza:**  $E[X] = \mu = np$
- **Varianza :**  $V[X] = \sigma^2 = npq$

**Ejercicio 10.2:** En una binomial  $B(8; 0,2)$ . Calcular  $P[X = 0]$ ,  $P[X \neq 0]$ ,  $P[X = 2]$ . Así como la media y la varianza.

**Solución 10.2:**

- $P[X = 0] = 0,8^8 = 0,168$  (probabilidad de "ningún éxito")
- $P[X \neq 0] = 1 - 0,8^8 = 0,832$  (probabilidad de "algún éxito")
- $P[X = 2] = \binom{8}{2} 0,2^2 0,8^6 = 0,294$
- **Esperanza:**  $E[X] = np = 8 * 0,2 = 1,6$
- **Varianza :**  $\sigma^2 = npq = 8 * 0,2 * 0,8 = 1,28$



**Ejercicio 10.3:** En un municipio el 30% de los habitantes está a favor de la creación de unas instalaciones deportivas municipales y en contra el resto. Elegidas 10 personas al azar, comprueba si la variable aleatoria que expresa el número de personas de la población favorable a la construcción del polideportivo sigue una distribución binomial. En caso afirmativo, señala los parámetros de la distribución.

**Solución 10.3:** Se trata de una distribución binomial de parámetros  $n = 10$  y

$p = 0,3$ , es decir,  $B(10; 0,3)$ .

$p$  representa la probabilidad de estar a favor de la creación de las instalaciones deportivas municipales

### 10.3. Variables continuas

Como vimos al principio de la unidad una variable aleatoria es continua cuando puede tomar cualquier valor dentro de un intervalo. Mientras que para una variable discreta cada valor posible de la misma tiene asociado una probabilidad concreta que tome ese valor, en el caso continuo la variable aleatoria puede tomar los infinitos valores que existen en un intervalo de definición, resultando que la probabilidad de un valor particular es 0. Debemos entonces buscar una alternativa para describir las probabilidades asociadas a este tipo de variables.

Al igual que hicimos con las variables aleatorias discretas vamos a introducir algunos ejemplos de variables continuas.

- La variable aleatoria que asigna a cada uno de los animales de una granja su peso. Esta variable aleatoria puede tomar, en principio, cualquier valor dentro de un intervalo del conjunto de los números reales  $\mathbb{R}$ .
- La distancia que recorre un vehículo en un día
- El volumen de una piscina

#### 10.3.1 Función de densidad

Si hablamos del peso de todas las personas residentes en España la probabilidad de que pese 62,345679725 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 función 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 las personas, encontraremos un mayor porcentaje de personas (mayor densidad de probabilidad) con peso entre 60 y 65 Kg que entre 130 y 135.

**Definición:** Una función  $f: \mathbb{R} \rightarrow \mathbb{R}$  se llama **función de densidad** sobre  $\mathbb{R}$  si cumple:

1.  $f(x)$  es no negativa  $f(x) \geq 0$  para todo  $x$  que toma valores en los números reales.
2.  $\int_{-\infty}^{\infty} f(x)dx = 1$

Asimismo notemos que si una función  $f: \mathbb{R} \rightarrow \mathbb{R}$  es una función de densidad

$$F(x) = P\{X \leq x\} = \int_{-\infty}^x f(x)dx$$

$F$  es la **función de distribución** y se cumple que  $F'(x) = f(x)$

Para obtener los parámetros de la media y la varianza de una variable continua se procede como sigue:

- **Esperanza**  $\mu = E[X] = \int_{-\infty}^{\infty} xf(x)dx$
- **Varianza**  $\sigma^2 = V[X] = E[(X - \mu)^2] = \int_{-\infty}^{\infty} x^2 f(x)dx - \mu^2$

**Ejercicio 10.4:** Calcular  $k$  para que la función

$$f(x) = \begin{cases} k, & x \in [1,4] \\ 0, & \text{en otro caso} \end{cases}$$

Sea una función de densidad. Hallar la probabilidad  $P[1,6 \leq X \leq 5,2]$

**Solución 10.4:**

$$\int_{-\infty}^{\infty} f(x)dx = \int_1^4 kdx = k(4 - 1) = 3k = 1 \Rightarrow k = \frac{1}{3}$$

$$P[1,6 \leq X \leq 5,2] = \int_{1,6}^4 \frac{1}{3} dx = \frac{1}{3}(4 - 1,6) = 0,8$$

### 10.3.2 La distribución Uniforme

**Definición:** Una variable aleatoria  $X$  se distribuye según una distribución uniforme o rectangular en el intervalo real  $(a,b)$  y la denotamos por  $X \equiv U(a,b)$ , si su función de densidad es

$$f(x) = \begin{cases} 0 & \text{si } x \leq a \\ \frac{1}{b-a} & \text{si } a < x < b \\ 0 & \text{si } x \geq b \end{cases}$$

Comprobamos que  $f(x)$  es función de densidad:

- 1)  $f(x) \geq 0 \quad \forall x \in (a, b)$
- 2)  $\int_{-\infty}^{\infty} f(x) dx = \int_a^b \frac{1}{b-a} dx = \frac{1}{b-a} (b-a) = 1$

Por tanto es función de densidad

- **Función de Distribución**

$$F_X(x) = P\{X \leq x\} = \int_{-\infty}^x f(x) dx$$

$$F(x) = P\{X \leq x\} = \begin{cases} 0 & \text{si } x \leq a \\ \frac{x-a}{b-a} & \text{si } a < x < b \\ 1 & \text{si } x \geq b \end{cases}$$

Los parámetros de esta distribución son:

- **Esperanza:**  $E[X] = \mu = \frac{b+a}{2}$
- **Varianza :**  $V[X] = \sigma^2 = \frac{(b-a)^2}{12}$

### 10.3.3 La distribución Normal

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.

La gran importancia teórica de la distribución normal radica en que, bajo ciertas condiciones, son muchas las distribuciones que tienden hacia la normal.

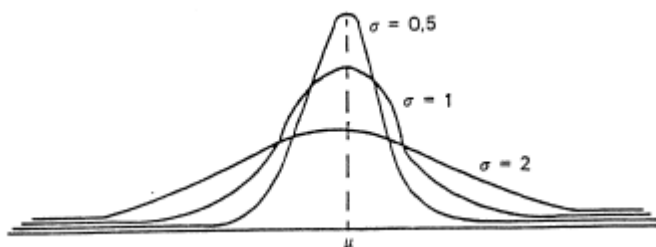
**Definición:** Una variable aleatoria  $X$  se dice que sigue una **distribución normal** con parámetros  $\mu$  y  $\sigma$  ( $\mu, \sigma \in \mathbb{R}$ ,  $\sigma > 0$ ) si su función de densidad es:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \quad \forall x \in \mathbb{R}$$

Si  $X$  está normalmente distribuida lo denotaremos de la siguiente forma  $X \equiv N(\mu, \sigma)$

La función de densidad de una variable normal tiene las siguientes propiedades

- Dicha función tiene la representación siguiente



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 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 de las observaciones. La probabilidad se distribuye en forma simétrica alrededor de la media.

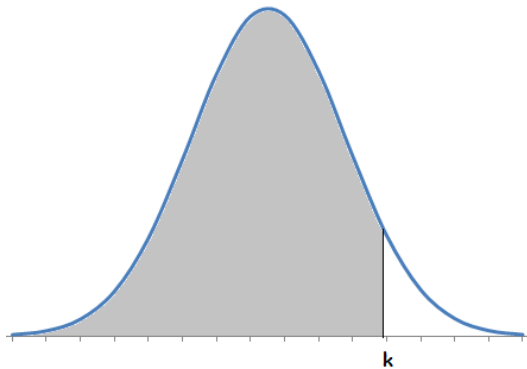
- Es continua en toda la recta real
- Es simétrica respecto de  $\mu$
- Nunca toma el valor 0, pero tiene al eje horizontal como asíntota horizontal

- Es estrictamente creciente si  $x < \mu$  y estrictamente decreciente si  $x > \mu$ .
- Máximo: hacer máxima  $f(x)$  equivale a hacer mínimo  $e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ , que es cuando  $x = \mu$
- Tiene dos puntos de inflexión que son  $\mu \pm \sigma$

La distribución normal más importante, es la  $N(0,1)$ , que suele designarse por  $Z$ , y que recibe el nombre de **distribución normal estándar**. Su función de densidad es

$$\phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \quad \forall z \in \mathbb{R}$$

Las probabilidades de la distribución normal estándar  $P[Z \leq k]$  se encuentran tabuladas y representan el área debajo de la campana de Gauss para los valores menores que  $k$ .



Para el resto de distribuciones normales, que no son la estándar, el cálculo de probabilidad de que ocurra un suceso entre  $(-\infty, k]$ , precisa resolver la siguiente integral:

$$P[X \leq k] = \int_{-\infty}^k \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} dx$$

Para evitar calcular esta integral para cada valor de la media  $\mu$  y para cada valor de la desviación estándar  $\sigma$ , se hace una sustitución o cambio de variable a la distribución  $N(0,1)$  para posteriormente hacer uso de los valores que se encuentran tabulados con referencia a la distribución normal tipificada  $N(0,1)$ . El resultado que se muestra a continuación pone de manifiesto lo señalado.

**Proposición:** Si  $X$  es  $N(\mu, \sigma)$ , para calcular la probabilidad  $P[X \leq k]$  se realiza de la siguiente manera



$$P[X \leq k] = P\left[Z \leq \frac{k - \mu}{\sigma}\right]$$

El cambio  $Z = \frac{(x-\mu)}{\sigma}$  se llama tipificación de la variable, donde Z es una variable  $N(0,1)$ .

**Ejercicio 10.5:** Si consideramos una variable X que se distribuye como  $N(6,4)$ . Se pide calcular las siguientes probabilidades:

- $P[X \leq 3]$
- $P[X \geq 12]$
- $P[5 \leq X \leq 8]$

**Solución 10.5:**

- $P[X \leq 3] = P\left[Z \leq \frac{3-6}{4}\right] = P[Z \leq -0,75] = 1 - P[Z \leq 0,75] = 0,2266$
- $P[X \geq 12] = P\left[Z \geq \frac{12-6}{4}\right] = P[Z \geq 1,5] = 1 - P[Z \leq 1,5] = 0,0668$
- $P[5 \leq X \leq 8] = P[-0,25 \leq Z \leq 0,5] = P[Z \leq 0,5] - P[Z \leq -0,25] = 0,2902$



## Unidad 11. Formas de investigar una población: censos, encuestas, registros administrativos y big data: ventajas e inconvenientes

### 11.1 Introducción

En el proceso de toma de decisiones por parte de una persona o una organización se hace necesario el conocimiento, **la información**. Por ejemplo, el Gobierno (tanto nacional, como autonómico o local) necesita información para establecer sus diferentes políticas tanto económicas como de tipo social y hacer un seguimiento de las mismas.

Para obtener información se necesitan **datos**. Podemos definir los datos como el conjunto de observaciones o hechos que, una vez recogidos, organizados y procesados se transforman en información o conocimiento. Tanto los datos como la información, no son necesariamente numéricos. Cuando la información es numérica nos introducimos en el mundo de la Estadística.

Se entiende por **operación estadística** el proceso por el cual se obtiene información estadística. Cualquier estudio estadístico pasa por una fase de diseño en la que a partir de las necesidades de los usuarios se establece la necesidad de la información a obtener y su viabilidad, se definen los objetivos del estudio, la metodología a seguir para alcanzarlos, posibles fuentes de datos, los costes, y el resto de especificaciones necesarias para la obtención del producto final.

La Ley de la Función Estadística Pública regula la actividad estadística para fines estatales y encomienda al INE y a otros organismos estatales que realizan producción estadística **la realización de las operaciones estadísticas** de interés estatal (censos demográficos y económicos, cuentas nacionales, estadísticas demográficas y sociales, indicadores económicos y sociales, coordinación y mantenimiento de los directorios de empresas,...).

### 11.2 Formas de investigar una población

Cuando se lleva a cabo una investigación estadística existe un conjunto de elementos (personas, viviendas, empresas,...) o unidades del que se desea obtener



información relativa a sus características. A este conjunto de elementos lo denotamos con el nombre de **población**.

La forma de obtener la información original de las unidades de análisis que componen la población a estudiar puede ser efectuada a través de distintos métodos: **censo, encuesta por muestreo, registro administrativo y big data**.

En lo sucesivo vamos a describir cada uno de los métodos anteriormente mencionados, exponiendo las ventajas e inconvenientes de cada uno de ellos.

### 11.2.1 Censos y Encuestas. Ventajas e inconvenientes

La primera posibilidad para el estudio de una población sería la realización de lo que se conoce con el nombre de **censo**. Hablamos de censo, cuando se toman los datos necesarios de manera exhaustiva de todas y cada una de las unidades que conforman la población objeto de estudio.

A la hora de realizar un censo debemos garantizar que se cumplan una serie de características esenciales:

- **Universalidad**: debe registrar de forma separada a todos los elementos que conforman la población. Resulta clave, asimismo, que se definan de manera clara los criterios que establecen cuando una unidad pertenece o no a la población.
- **Simultaneidad**: la información recogida debe estar referida al mismo periodo de tiempo, es decir, tener el mismo período de referencia para todas y cada una de las unidades.
- **Periodicidad**: cuando se realiza una operación censal para el estudio de una población de forma sistemática, se suele llevar a cabo teniendo en cuenta una periodicidad temporal, con el fin de facilitar la comparabilidad de los datos obtenidos. Por ejemplo, en el caso de los Censos de Población y Viviendas en España se vienen realizando cada diez años.

La realización de un censo no siempre es viable por diversas razones, por ejemplo cuando obtener información implica la destrucción o consumo de la unidad de estudio. Estos inconvenientes llevan a plantearse el estudio de la población tomando sólo una parte de los elementos de la misma. Así pues, definimos **encuesta por muestreo** como el proceso mediante el cual se obtienen conclusiones de la población a partir de la



información proporcionada por una parte de ella (muestra). El desarrollo de las técnicas de muestreo ha permitido el gran incremento de las investigaciones mediante encuestas.

Por otro lado, 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 en otra de las unidades de este módulo de probabilidad y muestreo.

### Ventajas e inconvenientes

Un censo presenta dos principales **ventajas** respecto al muestreo:

1. **Sin errores de muestreo**: los resultados finales se basan en toda la población, y como consecuencia no están sujetos al error de muestreo.
2. **Elevada desagregación de los resultados**: al realizarse una enumeración completa de la población, es posible dar un grado de detalle de la información final incluso para subgrupos y áreas geográficas de pequeño tamaño respecto a la población total.

Pero presenta también **inconvenientes**:

1. **Coste elevado**: la operación censal puede ser muy costosa para grandes poblaciones.
2. **El tiempo de realización**, que puede ser grande respecto a una operación por muestreo.
3. **Errores ajenos al muestreo**: los errores de recogida de datos y proceso que suelen aumentar con el número de elementos de los que se recogen datos.
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 todas las unidades de la población.

Por otro lado, en lo que se refiere a las encuestas por muestreo 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:



1. **Menor coste:** si los datos se obtienen de una pequeña parte de la población, los gastos son menores que los de realizar una investigación exhaustiva, aun cuando el coste por unidad pueda ser superior.
2. **Resultados rápidos:** la obtención de los resultados se realiza con mayor rapidez derivado de obtener información de una parte de la población.
3. **Errores ajenos al muestreo:** al reducirse el volumen de trabajo se puede emplear personal especializado, mejor preparado y entrenado. Igualmente los procesos de supervisión y proceso de datos están mejor controlados, lo que redundará en una mejor calidad del trabajo y una disminución de los errores (no de muestreo) respecto al censo.
4. **Menos limitaciones en las características a investigar:** el uso de equipos más especializados hace que el muestreo tenga más posibilidades y flexibilidad respecto a la información a obtener.

Los inconvenientes de las encuestas son:

1. **Errores de muestreo:** las encuestas están sujetas a los errores de muestreo, pero se disponen de métodos para medirlo y controlarlo.
2. **Menor desagregación de los resultados:** el grado de detalle de la información final está siempre limitado por el tamaño de la muestra que la soporta por lo que no es posible llegar a los niveles de desglose de un censo.

### 11.2.2 Registros administrativos. Definición. Ventajas e inconvenientes.

Una tercera alternativa al censo y al muestreo como formas de investigación de una población, es la utilización de **registros administrativos**.

Un registro administrativo es un conjunto de unidades y datos derivados de las mismas que son el resultado de las operaciones habituales de una organización y que se recogen en principio con el fin de llevar a cabo diversas actividades no estadísticas.

Algunos ejemplos típicos serían los registros de nacimientos, defunciones, matrimonios, matriculación de vehículos, directorios telefónicos, registro de sociedades, registros tributarios,.....

Las principales **ventajas** de utilizar registros administrativos son:



1. **No hay error de muestreo:** 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.
2. **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 disminuyen de manera considerable.
3. **Reducción de la carga de respuesta:** 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.
4. **Evolución temporal:** los registros se actualizan permanentemente y la recogida de información en forma periódica permite la realización de análisis de tendencias.
5. **Información más desagregada:** el uso de registros puede permitir por ejemplo obtener la información por zonas geográficas o actividad económica.

Algunos **inconvenientes** de los registros administrativos son:

-**Limitación a la información contenida en el registro:** los datos primarios a recoger están limitados por la información administrativa que contenga el registro. La población objetivo de una fuente administrativa puede ser diferente de la población objetivo para el fin estadístico. 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.

-**Necesidad de adaptación de las definiciones estadísticas** Asimismo, las definiciones de las variables estadísticas deben ajustarse a las disponibles en los registros administrativos.

Como ejemplo de operaciones estadísticas que realiza el INE procedentes de la explotación de registros administrativos tenemos las Estadísticas del Movimiento Natural de la Población (MNP) y el próximo Censo de Población de 2021 que se obtendrá con la explotación de la información de registros administrativos.



### 11.2.3 Big Data. Ventajas e inconvenientes

**Big Data** es un término que se utiliza en la actualidad para hacer referencia a la disponibilidad exponencialmente creciente de datos digitales de muy diversa naturaleza que las nuevas tecnologías de la información y de las comunicaciones han producido en los últimos años y que previsiblemente seguirá con la misma tendencia ascendente en el futuro próximo.

La caracterización más habitual de este fenómeno proviene del análisis original de Laney, que, sin emplear el término Big Data, caracterizaba el fenómeno por 'las tres V': aquella fuente de información de gran Volumen de datos generados a gran Velocidad y con gran Variedad en sus estructuras.

El gran volumen viene determinado porque las fuentes de datos utilizados frecuentemente superan el orden de los terabyte ( $10^{12}$  bytes) y pueden llegar a órdenes de magnitud de petabytes ( $10^{15}$  bytes). La gran variedad hace referencia a la cantidad de estructuras de datos incluyéndose también casos de datos no estructurados o datos de entrada que pueden ser imágenes, textos, audios, ... y la gran velocidad hace referencia a aquella con la que se generan y analizan los datos.

Los dos pasos a considerar en el Big Data son el tratamiento de la información y su análisis.

Para el tratamiento de la información se utilizan infraestructuras que permiten el procesamiento paralelo, por lo que se generan ficheros distribuidos sobre grupos de ordenadores usándose tecnologías específicas.

Una forma de acceso a la información es mediante web scraping. Consiste en rastrear sitios web buscando la información relevante (por ejemplo, precios de artículos que se venden por internet).

Respecto al análisis de la información, se utilizan técnicas como el Data Mining o el Machine Learning y métodos automáticos de construcción de modelos.

Algunas de las fuentes de información de Big Data y sus potenciales usos estadísticos son:

- Datos de posicionamiento de la telefonía móvil



Algunos de sus potenciales usos estadísticos son para el estudio de la movilidad de la población, en las encuestas turísticas para el conocimiento del número de turistas y sus desgloses por nacionalidad o para el conocimiento de movimientos migratorios.

- Datos de pago mediante medios electrónicos (tarjetas bancarias, etc)

Tienen utilidad potencial en las encuestas de presupuestos familiares, gastos de los turistas o comercio al por menor.

- Datos de cámaras en las vías de circulación

Su utilidad potencial está en las encuestas de turismo para conocer las entradas y destinos de los turistas extranjeros.

- Datos de alquiler de apartamentos turísticos en plataformas de alquiler en internet

Tienen utilidad potencial para la estimación del número de turistas que utilizan este tipo de alojamiento.

- Datos de movimientos societarios en empresas

De potencial utilización en los directorios de empresas para detectar cambios en las estructuras de los grupos empresariales.

- Datos de compras por internet

Pueden utilizarse potencialmente para la elaboración de indicadores de precios.

- Datos con información de usos del suelo a partir de satélites

Pueden utilizarse en las encuestas agrícolas.

En la actualidad los institutos de estadística están trabajando con la finalidad de incorporar estas fuentes de información en los procesos de producción estadística. La utilización prevista no sólo es directa, sino que puede utilizarse para potenciar la información obtenida por otras vías, como por ejemplo, encuestas, combinando los datos de ambas fuentes y permitiendo mayores desgloses tanto en el tiempo como en el espacio en base a la utilización de modelos.



## Unidad 12. Conceptos de población, marco y muestra. Tipos de muestreo. Concepto de estimador y sus principales propiedades. Errores de muestreo. Errores ajenos al muestreo: de cobertura, de falta de respuesta y de medida.

El objetivo de esta unidad es realizar una descripción de los aspectos y conceptos generales de una encuesta por muestreo.

### 12.1 Conceptos población y muestra

El concepto de **población objetivo** corresponde al conjunto de unidades de las que se desea obtener una información. La **unidad elemental** o unidad de investigación, es todo elemento de la población sobre el que se realiza la medición.

Así pues, si deseamos conocer el volumen de negocios de las empresas del territorio nacional, la población objetivo sería la relación de empresas del territorio nacional. Y si lo que queremos es estudiar el salario de los habitantes de un municipio, la población objeto de estudio sería cada uno de los habitantes de ese municipio mayores de 16 años, que es la edad legal para trabajar en España.

Se llama **muestra** a una “pequeña” parte de la población objeto de estudio, a partir de la cual se pretende el conocimiento de características o valores de todo el conjunto (población) del que se ha tomado la muestra. Los datos obtenidos a partir de la muestra se denominan **estimaciones** y nos permiten inferir las características de la población que es en lo que estamos interesados.

Por otro lado, las **unidades de muestreo** son las que se utilizan en la 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.

Por ejemplo, si se está interesado en estudiar a los individuos pero solo se dispone de una lista de viviendas. La vivienda es la unidad de muestreo y las personas de la vivienda la unidad elemental o de investigación. En este caso la unidad de muestreo y la unidad elemental no coincidirían.



## 12.2 Marco. Definición y principales características

Se denomina **marco de muestreo** a la relación de unidades a partir de la cual se selecciona la muestra. En un *sentido amplio*, el marco se considera a la lista o relación de todas las unidades de la población objetivo con la información necesaria para la identificación de las mismas, junto con información complementaria de cada una de las unidades que se puede utilizar para la mejora del diseño muestral. Esta información adicional sirve, entre otros fines, para la formación de estratos o en el tratamiento de la falta de respuesta.

Dentro de sus características principales destacaríamos por un lado que debe ser un fiel reflejo de la población objetivo y, por otro, con el fin de facilitar los trabajos de la recogida de la información, las unidades que lo forman deben estar perfectamente identificadas para posibilitar la localización, en el caso de ser seleccionadas en la muestra.

En la práctica, el marco de selección puede coincidir en mayor o menor grado con la población objetivo, ya que este puede presentar algunos defectos tales como:

1. **La existencia de unidades duplicadas.** Por ello se debe realizar una depuración previa para evitar la inclusión de unidades repetidas en el marco.
2. **Unidades omitidas.** Todos los elementos de la población objetivo deben de estar en el marco. La ausencia en el marco de una parte de la población investigada proporciona una subestimación de las características investigadas
3. **Unidades incluidas que no pertenecen a la población objetivo.** La existencia de este tipo de unidades no introduce sesgos pero aumenta la variabilidad de la estimación.

Es importante señalar que la formación del marco puede tener un impacto importante en el coste de la operación. Asimismo, generalmente se recurre a formar los marcos a partir de otras fuentes ya existentes y, se realizan actualizaciones periódicas y continuas del mismo, para garantizar que la población muestreada sea idéntica a la población objeto de estudio.



## 12.3 Tipos de muestreo

Al procedimiento mediante el cual se selecciona la muestra se le denomina **muestreo**. Ahora bien, la pregunta sería cómo se puede seleccionar esta muestra de tal forma que el subconjunto de la población sea lo más representativo de ésta.

Aunque hay muchos tipos de obtener una muestra que garantizan la representatividad, existen formas básicas de clasificar todos ellos:

- **Muestreo Probabilístico.** Está basado en la estadística matemática y se define como el tipo de muestreo en el que se conoce a priori la probabilidad que tiene cada una de las posibles muestras de ser seleccionada.

Existen dos elementos definitorios de este tipo de muestreo:


- Las unidades son seleccionadas de manera aleatoria, de esta forma estamos garantizando que todas las unidades pueden ser seleccionadas y en igualdad de condiciones y no solo las que pertenecen a un determinado grupo.
- Todas las unidades muestrales tienen probabilidad de selección no nula, podemos aplicar los métodos estadísticos que permiten, no solo dar estimaciones, sino también el cálculo de los errores de muestreo.

El principal problema de hacer un muestreo de este tipo es que para poder controlar las probabilidades de selección de la muestra, es necesario conocer de forma teórica todas las posibles muestras. Y eso conlleva disponer de una relación de todas las unidades de la población. Por ejemplo, las empresas de investigación por encuestas dirigidas a los hogares no disponen de un directorio completo con todos los hogares del territorio que desean investigar, por tanto en estos casos no pueden hacer uso de técnicas de muestreo probabilístico.

Ejemplos de este tipo de muestreo son el muestreo aleatorio simple, muestreo estratificado, muestreo de conglomerados, etc.

El muestreo probabilístico es el más utilizado en los Institutos Nacionales de Estadística para las encuestas oficiales.

- **Muestreo No Probabilístico.** En este tipo de muestreo, la selección de la muestra no está sometida a criterios probabilísticos. Algunos métodos no probabilísticos pueden llegar a tener una fase de selección aleatoria, pero en alguna otra fase hay decisiones que son determinísticas, es decir en algún momento hay una decisión no aleatoria para



decidir las unidades que finalmente se van a entrevistar. Esto puede producir lo que más adelante llamaremos sesgo.

Ahora bien, este tipo de muestreo es frecuentemente utilizado puesto que resulta una alternativa más rápida a la hora de plantear el diseño y más barata puesto que las entrevistas se pueden concentrar en unas zonas geográficas reduciendo costes de los trabajos de la recogida de la información. Asimismo, no es necesario disponer de un marco de unidades para la selección.

Sin embargo dentro de las desventajas tendríamos que, a diferencia del muestreo probabilístico, no es posible determinar el margen de error real de la estimación, puesto que no se conoce la probabilidad de selección de la unidad.

Dentro de los tipos de muestreo no probabilístico, los más utilizados son el muestreo opinático en el que la persona que selecciona la muestra procura que esta sea representativa (selección de unidades tipo) y el muestreo por cuotas en el que la muestra se selecciona en un número proporcional al de los que cumplen unas determinadas características de la población. En general, las cuotas más utilizadas son las de edad y sexo, por ser normalmente la información disponible de la población. Se utiliza en las encuestas de opinión.

## 12.4 Concepto de estimador

El estimador es la **expresión matemática** que nos permite inferir las características de la población a partir de los datos de una muestra. El valor que toma el estimador en una determinada muestra o, valor inferido, se conoce como estimación.

La utilización de métodos de estimación adecuados permite obtener estimaciones consistentes en ausencia de casos de no respuesta.

El estimador de Horvitz-Thompson es el estimador básico que sirve para realizar la estimación de parámetros poblacionales lineales como el total o la media poblacional. Dicho estimador pondera cada unidad de la muestra con la inversa de la probabilidad de selección.

Así el estimador de Horvitz-Thompson para estimar el total poblacional

$$X = \sum_{i=1}^N X_i$$

Viene dado por la expresión siguiente:

$$\hat{X} = \sum_{i=1}^n \frac{X_i}{\pi_i}$$


suponiendo que hemos obtenido una muestra de tamaño  $n$ , mediante un diseño muestral probabilístico donde  $\pi_i$  es la probabilidad de selección de la unidad muestral.

## 12.5 Errores de muestreo e insesgadez

Como hemos dicho, estimar mediante una muestra no nos garantiza obtener el verdadero valor del parámetro. Es más, si hacemos la misma encuesta al mismo número de unidades muestrales, pero cambiando cualquiera de éstas, lo más probable es que la estimación sea diferente; es decir, cada valor que estimamos está asociado a la muestra que elijamos.

Imaginemos que calculamos el tiempo medio que tardamos en llegar al trabajo desde casa. Para ello tenemos los datos de los últimos 20 días y tomamos una muestra de 5 de ellos para estimarlo. Supongamos que nos da una estimación de 20,4 minutos. Ahora tomamos otros 5 y nos da una media de 20,2 minutos. Luego otros 5 y tenemos 19,6 minutos... Así podríamos obtener todas las estimaciones de los 20 días agrupados de 5 en 5. Se puede calcular que hay 15.504 posibles combinaciones y, si cada combinación puede generar una estimación, implica que podría haber 15.504 posibles estimaciones del tiempo medio que se tarda. Con estas cifras parece que es tal la variabilidad de estimaciones que nos invita a pensar en lo poco fiable que sería el valor que obtuviésemos. Pero supongamos dos casos extremos:

1. Imaginemos que todos los valores que obtenemos están en un intervalo entre 19,4 y 20,8. ¿Tenemos con esto una idea bastante buena de cuanto se tarda de media en ir al trabajo? Desde luego cualquiera de esas estimaciones se encuentra bastante cerca de las otras pues todas están en un intervalo muy próximo, con lo que quedarme con cualquier valor es suficiente para hacerme una idea del tiempo que me lleva ir a trabajar.
2. En cambio supongamos que las estimaciones son mucho más variables, por ejemplo una de ellas fuese 30,5 minutos, otra 10,4 minutos, otra 7,5, otra 45



minutos... Y así hasta ver las 15.504 muestras (y por tanto 15.504 estimaciones) se reparten entre un intervalo muy amplio de entre 7 minutos y 50 minutos. Es evidente que aquí cualquier valor parece que puede ser muy poco representativo del que sería el valor medio.

La realidad es que nosotros no disponemos de todas las posibles muestras, sino solo de una: la que seleccionamos para la encuesta. Si nuestro caso fuera el primero del ejemplo anterior, no nos importaría mucho cual fuese la seleccionada, pues todas las posibles muestras proporcionan estimaciones muy cercanas y cualquiera nos daría una idea bastante buena del tiempo medio. Pero si fuese el segundo caso, donde las posibles estimaciones son extremadamente variables, la que elijamos podría distanciarse mucho del verdadero valor del tiempo medio.

### 12.5.1 Sesgo en las estimaciones

Una de las características importantes de la estimación es la conveniencia de que esta sea insesgada. Esto quiere decir que la media de todas las posibles estimaciones debe coincidir con el valor del parámetro poblacional que buscamos. En caso de que no sea así, diremos que la estimación tiene un sesgo.

Si denotamos por  $\hat{Y}$  al estimador de una característica poblacional que se desea estimar, el valor esperado de  $\hat{Y}$  viene dado por la expresión

$$E[\hat{Y}] = \sum_{i=1}^k p_i \hat{Y}_i$$

Donde denotamos por  $k$  al número de posibles muestras (y por lo tanto de posibles estimaciones). Cada una de las muestras nos da un valor de estimación  $\hat{Y}_i$ , y por  $p_i$  denotamos a la probabilidad de selección de dicha muestra.

El sesgo del estimador es:

$$\text{Sesgo}[\hat{Y}] = E[\hat{Y}] - Y$$

Por ello, cuando el sesgo es nulo hablamos de estimador insesgado.

### 12.5.1 Error de muestreo: Precisión

Además de pedir que el estimador sea insesgado, también se desea que dicho estimador sea preciso, es decir, que el error de muestreo asociado sea mínimo. El error



de muestreo es la penalización por observar sólo una parte de la población y no la totalidad de la misma. Si bien, si el muestreo es probabilístico, podemos saber la probabilidad de cada muestra y, por tanto, dar una medida de este error.

La expresión sería la siguiente: si denotamos por  $k$  al número de posibles muestras (y por lo tanto de posibles estimaciones) y cada una nos da un valor de estimación  $\hat{Y}_i$ , con una probabilidad de ser seleccionada dicha muestra  $p_i$ . Entonces el error de muestreo viene dado por la expresión:

$$EM = \sqrt{V[\hat{Y}]} = \sqrt{\sum_{i=1}^k (\hat{Y}_i - E[\hat{Y}])^2 p_i}$$

Un estimador puede tener un error de muestreo muy bajo, pero ser sesgado. Para estas ocasiones utilizamos el error cuadrático medio (ECM) que es la combinación de ambas medidas, error de muestreo y sesgo.

$$ECM = E[(\hat{Y} - Y)^2] = V[\hat{Y}] + (\text{Sesgo}[\hat{Y}])^2$$

El error cuadrático medio nos proporciona una medida de lo distante que está el conjunto de las posibles estimaciones. A este concepto se le llama acuracidad. Si un estimador proporciona estimaciones que distan poco del valor objetivo y, además, la media de todas ellas coincide con dicho valor entonces el estimador es muy acurado. Si en cambio, hay un alto error de muestreo o un nivel de sesgo muy elevado o ambas cosas a la vez, entonces el estimador se dice poco acurado.

En muchas ocasiones el sesgo disminuye a medida que la muestra aumenta, considerando que es despreciable si se dispone de un número considerable de unidades muestrales. En ese caso o, si el estimador es insesgado, la bondad del estimador (la acuracidad) dependerá solo del error de muestreo.

## 12.6 Errores ajenos al muestreo

Hasta ahora hemos supuesto que la población marco coincide con la población objetivo, que la muestra real alcanzada se corresponde con la muestra inicialmente planificada y seleccionada probabilísticamente y 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.

Como comentario hay que decir que al plantear un estudio por muestreo debe prestarse especial atención a los errores que no son de muestreo que pueden presentarse en cualquier fase del trabajo y, si son importantes, incluso pueden 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, de recogida y de procesamiento pueden ayudar a mitigar el sesgo. La calidad final de una encuesta considera tanto los errores de muestreo como los ajenos al muestreo.

### 12.6.1 Errores de cobertura


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, puede verse como una lista de unidades de muestreo.

Podemos pensar en un marco o 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, erróneamente incluidas), incluso el listado puede contener unidades duplicadas. Veamos con un poco más de detalle estos problemas que podemos encontrar en las unidades del marco:

- **Unidad vacía:** es aquella que estando incluida en el marco o lista utilizada no contenía ninguna unidad perteneciente a la población objetivo que se desea investigar. Generalmente se produce cuando, por no estar actualizado, incluye unidades que han dejado de pertenecer a la población objetivo.

Ejemplo: Una vivienda deshabitada en una encuesta de hogares que utiliza un marco de viviendas para estimar las características de sus habitantes constituye una unidad vacía.

- **Unidad erróneamente incluida:** es aquella que estando incluida en el marco no es en realidad una unidad perteneciente al marco que se desea muestrear (no



es una unidad de la población objetivo). Generalmente se produce cuando la población objetivo es una subpoblación de la población marco.

Ejemplo: en una encuesta dirigida al sector de la industria, si se incluyen unidades del sector servicios, estas unidades serían erróneamente incluidas.

La presencia de unidades erróneamente incluidas o vacías se produce principalmente:

- Por no estar actualizado el marco, puesto que incluye unidades que han dejado de pertenecer al colectivo que se desea muestrear.
- Porque la población que se desea muestrear es una subpoblación de la cubierta por el marco.
- Por errores en la asignación de ciertas características en el marco.
- **Unidad repetida:** es aquella que aparece más de una vez en la población marco. La presencia de estas alteran las probabilidades de selección que en vez de ser iguales para todas las unidades son proporcionales al número de veces que se repite en el marco. Es obvio que este problema se soluciona si se depura el marco disponible eliminando las identificaciones repetidas.

Ejemplos:

- En encuestas de hogares que utilizan un marco de viviendas, los hogares que tengan más de una vivienda en el marco, son unidades repetidas.
- Si el marco se ha configurado reuniendo diferentes fuentes, es posible que alguna unidad esté repetida

### 12.6.2 Falta de respuesta

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 que hay no respuesta**. 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 entrevistar que familias con niños pequeños o con personas jubiladas.
  - a2) Viaje, vacaciones.



- a3) Enfermedad
- a4) Problema 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.
- a7) Barrio o vecindad difícil.
- b) **Negativa a colaborar**, debido a:
  - b1) Falta de tiempo.
  - b2) Falta de motivación o interés por el tema de la encuesta
  - b3) No desea que el entrevistador conozca sus respuestas u opiniones.
  - b4) Cansancio de las entrevistas.
  - b5) Cuestionario demasiado largo, preguntas complicadas, preguntas que rozan la intimidad.
  - b6) Personas que rechazan ser entrevistadas o están sistemáticamente fuera de casa durante el tiempo disponible para el trabajo de campo.
  - b7) Dificultad del entrevistador para conseguir la colaboración.

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.

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 obtendrán 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 segundo estrato, 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 estrato 1. 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.

La falta de respuesta no debe ignorarse o pensar que se corrige sustituyendo en la muestra a los que no responden por otros que si colaboran, ya que ello no va a eliminar



el sesgo, simplemente mantendrá el tamaño de la muestra teórica. 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:

- Cartas y llamadas telefónicas por adelantado.
- Programar visitas repetidas puede ser de gran efectividad para reducir los no-en-casa.
- Incentivar la colaboración.

La mejora de los procedimientos de recogida de información es también un procedimiento para reducir la falta de respuesta. El entrenamiento del entrevistador es fundamental: la interacción positiva entrevistador-entrevistado es básica para el éxito de la entrevista, lo cual puede requerir que el entrevistador disponga de estrategias para afrontar la entrevista en función de ciertas características observables de los encuestados. En otros casos en los que el tema tratado es especialmente sensible, con el fin de preservar la intimidad del entrevistado, se pueden utilizar procedimientos de recogida como la autocumplimentación. Otro aspecto a tener en cuenta es que cuanto más tiempo requiera la colaboración de la unidad muestral menor es la disposición a participar: 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 que 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; el trabajo a realizar por el hogar en el caso del audímetro es menor, lo que favorece la colaboración.

En la práctica y a pesar de las medidas que se toman 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 la tasa de no respuesta en encuestas posteriores. En ocasiones será posible recoger ciertas características observables de las unidades que no han colaborado y que puedan ser utilizadas posteriormente en procedimientos de ajuste para mitigar los sesgos de no respuesta en las estimaciones finales.


### 12.6.3 Errores de medida

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 de  $X'$  (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 lo producen:

- **Instrumentos de medición** (cuestionario-entrevistador) inadecuados y sujetos a error.
- **Fallos de memoria:** el entrevistado responde lo que el cree que hizo, pero no lo que realmente hizo.
- **El entrevistado da una respuesta falsa**, bien inducido por el entrevistador (quizá por el cuestionario), o bien porque no desea informar sobre su situación (qué dirán...)
- **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.
- **Falta de información:** el informante no dispone de toda la información para contestar y da una respuesta aproximada.
- **Errores de codificación y grabación** que introducen en el proceso un valor erróneo con independencia de que el valor original fuera correcto o no.



## Unidad 13. Muestreo aleatorio simple: estimadores del total y de la media y sus errores de muestreo. Tamaño de la muestra. Otros métodos de muestreo probabilístico: estratificado, por etapas y sistemático: concepto y razones principales para su aplicación.

En la unidad anterior habíamos visto dos grandes grupos de muestreo: probabilístico y no probabilístico. Como habíamos visto la principal ventaja del muestreo probabilístico es que de esta manera se permite evaluar el posible error de muestreo y con ellos se obtiene un margen de error para las estimaciones, además de que garantiza el que todas las unidades del marco de selección puedan ser seleccionadas.

En esta unidad se van a detallar algunos de los principales tipos de muestreo probabilístico.

### 13.1 Muestreo aleatorio simple

El **muestreo aleatorio simple** es, dentro de los muestreos probabilísticos, el de diseño más sencillo. Es un método de selección en una sola etapa en el que se asigna la misma probabilidad de selección a cada una de las posibles muestras.

Como consecuencia, cada unidad en la muestra tiene la misma probabilidad de ser elegida. Si tenemos una población de  $N$  unidades y seleccionamos una muestra de tamaño  $n$ , entonces la probabilidad de que una unidad de la población sea seleccionada es la misma para todas las unidades de la población:  $n/N$ .

Hacemos notar que la selección de las unidades se puede hacer con o sin reemplazo. El muestreo con reemplazamiento permite seleccionar una unidad más de una vez. El muestreo sin reemplazamiento significa que una vez que se ha seleccionado una unidad, no se puede volver a seleccionar. El muestreo aleatorio simple con reemplazamiento y sin reemplazamiento son prácticamente idénticos si el tamaño de la muestra es una fracción muy pequeña del tamaño de la población. Esto se debe a que la posibilidad de que la misma unidad aparezca más de una vez en la muestra es pequeña. Generalmente, el muestreo sin reemplazamiento produce resultados más precisos y es operativamente más conveniente. En lo que sigue de esta unidad, asumimos que el muestreo es sin reemplazamiento.



A modo de ejemplo consideremos una población de cuatro unidades que denotamos por  $\{U_1, U_2, U_3, U_4\}$  y supongamos que se selecciona una muestra de tamaño dos mediante muestreo aleatorio simple. Las muestras posibles serían 6 (las combinaciones de las cuatro unidades de 2 en 2):

$$\{U_1, U_2\}, \{U_1, U_3\}, \{U_1, U_4\}, \{U_2, U_3\}, \{U_2, U_4\}, \{U_3, U_4\}$$

Pues si tenemos que asignar la misma probabilidad a cada una de las posibles muestras, y como la suma de todas las probabilidades debe ser 1, entonces a cada una de las posibles muestras se le asigna  $1/6$ . Y la probabilidad de que una unidad sea seleccionada es  $n/N = 2/4 = 0.5$ .

Podríamos señalar una serie de aspectos positivos de este tipo de muestreo probabilístico:

- Resulta sencillo de implementar y no requiere de información auxiliar en el marco de selección, solo un marco con toda la población objetivo de la encuesta y la información necesaria para el contacto de las unidades seleccionadas.
- Fácil programación de la selección de la muestra, los factores de elevación y los errores de muestreo.
- Cálculo sencillo del tamaño de la muestra necesario para una precisión dada

Entre los inconvenientes cabe destacar:

- La no utilización de información auxiliar en el diseño, aunque esta esté disponible en el marco, impide mejorar la representatividad de la muestra.
- Puede tener asociado un coste elevado si se realizan entrevistas personales, ya que la muestra puede estar ampliamente distribuida geográficamente.
- Cabe la posibilidad de extraer una muestra "mala". Dado que todas las muestras de tamaño  $n$  tienen la misma probabilidad de ser incluidas en la muestra, es posible extraer una muestra que no esté bien distribuida respecto a alguna característica importante como por ejemplo la Comunidad Autónoma de residencia y que no represente la variedad de la población a estudio.

### 13.1.1 Estimadores del total y de la media y sus errores de muestreo

En el muestreo aleatorio simple, la expresión para el estimador del total poblacional sería

$$\hat{X} = \frac{N}{n} \sum_{i=1}^n X_i$$

Al cociente  $N/n$  se le llama **factor de elevación** e indica el total de unidades en la población que están representadas por una unidad de la muestra. Su inverso  $n/N$  se llama **fracción de muestreo** y se designa por  $f$ . Este estimador es lineal e insesgado.

Por otro lado, el estimador de la media poblacional de una característica es el siguiente

$$\hat{X} = \frac{1}{n} \sum_{i=1}^n X_i = \bar{x}$$

Donde  $\bar{x}$  es la media muestral. Así pues, la media muestral es un estimador insesgado de la media poblacional.

Al igual que tenemos expresiones para los estimadores en el muestreo aleatorio simple, también se pueden obtener las fórmulas del **error muestral** de cada uno de ellos.

Es importante analizar las expresiones del error, pues debemos asegurarnos que éste sea lo menos grande posible. Si vemos que podemos actuar de forma que este error se minimice, sabríamos que hacer para conseguir unas buenas estimaciones. Para ello, nos vamos a centrar en las características de una de las fórmulas del error, el de la media, pues son extrapolables al resto.

Se puede calcular que el error de muestreo(EM) de la media es:

$$\sigma_{\bar{x}} = \sqrt{\frac{N-n}{N-1}} \frac{\sigma}{\sqrt{n}}$$

A la vista de la expresión el error depende del tamaño poblacional  $N$ , del tamaño muestral  $n$  y  $\sigma$  que es la desviación típica de la población.

En la práctica, no se conoce la varianza de la población  $\sigma^2$  y el error estándar de la media se estima a partir de los propios datos muestrales mediante

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

También es muy frecuente en la práctica que el tamaño de la 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  $\hat{\sigma}_{\bar{x}} \approx \frac{s}{\sqrt{n}}$ .



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,

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

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

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

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.

### 13.1.2 Tamaño de la muestra

En este apartado vamos a establecer cómo determinar el tamaño muestral cuando estamos ante un muestreo aleatorio simple.

De los dos factores que intervienen en el error del muestreo, el único que es “controlable” es el tamaño muestral ( $n$ ), puesto que la varianza poblacional  $\sigma^2$  está determinada por la variable objeto de estudio. Pero en cambio sí que podemos diseñar una muestra con un tamaño u otro diferente. Así pues este factor es variable y si queremos que nuestro error de muestreo sea pequeño, entonces deberemos tomar una muestra del tamaño adecuado; como hemos visto en el apartado anterior el error de muestreo depende del inverso de la raíz cuadrada del tamaño muestral.

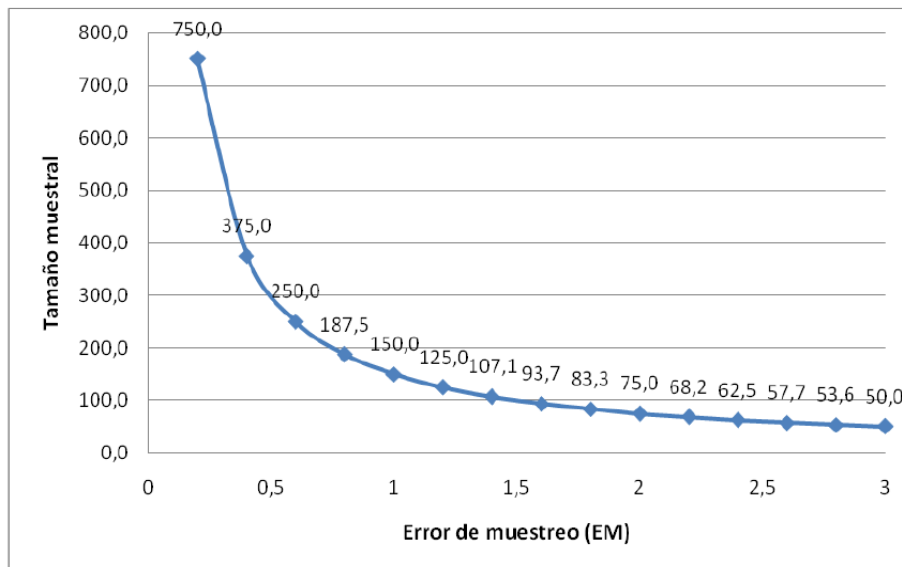
Entonces, ¿por qué no aumentar siempre el tamaño de muestra de las encuestas a niveles muy altos para asegurarnos de este modo que tenemos unas estimaciones muy precisas? Pues por dos motivos fundamentalmente. El primero es que el principal coste de una encuesta reside en la recogida de la información a través de un cuestionario dirigido a un informante. Por ello, a mayor número de entrevistas, más costosa es la operación estadística. Por lo tanto, el límite presupuestario se convierte en un límite para el tamaño muestral. Y el segundo motivo es que el aumento de la muestra no se traslada directamente en disminución de error muestral, es decir, duplicar el tamaño de la muestra, no reduce a la mitad el error muestral, sino en menor medida, tan sólo en aproximadamente un 29%. Así pues, si se quiere determinar el tamaño de la muestra



para obtener un cierto nivel de error estándar, no hay más que despejar el tamaño muestral  $n$ , de las fórmulas de los errores que introdujimos anteriormente. Por ejemplo, de la fórmula del error de muestreo de la media se puede despejar y obtener el tamaño muestral en función de la varianza poblacional, y el error muestral máximo que se desea obtener. Despejando en dicha fórmula se obtiene que el tamaño muestral sería:

$$n = \frac{\sigma^2 N}{EM^2(N - 1) + \sigma^2}$$

Un buen recurso a utilizar cuando se nos presenta el problema de establecer un tamaño muestral es realizar una gráfica que relacione el tamaño muestral necesario según el error de muestreo relativo para un  $\sigma^2$  y un tamaño poblacional  $N$  dados. Así para determinar el tamaño muestral se selecciona el Error de Muestreo (EM) máximo asumible y se observa en la gráfica cual es el tamaño mínimo necesario asociado.



## 13.2 Otros métodos de muestreo probabilístico

Acabamos de ver el tipo de muestreo probabilístico básico: el muestreo aleatorio simple. Junto con él existen otros tipos de muestreos probabilísticos, donde la muestra se obtiene mediante diseños más complejos, aunque siempre manteniendo el principio de aleatoriedad y de asignación de probabilidad de selección de la muestra. Los motivos para recurrir a este tipo de muestreos son varios. Por un lado, obtener diseños más eficientes, es decir, con menor error de muestreo para el mismo tamaño muestral, asimismo obtener estimaciones en subgrupos poblacionales y controlar los errores de



muestreo de dichas subpoblaciones. No menos importante: permiten adecuar el tipo de muestreo a la información y marcos disponibles.

A continuación, vamos a describir de forma básica la filosofía de los más utilizados y que resultan necesarios para la comprensión de los diseños muestrales de las encuestas que realiza el INE.

### 13.2.1 Muestreo estratificado

En algunas ocasiones se dispone de información adicional que puede ser de utilidad en el diseño muestral para la mejora de éste. En el caso de que la variable objeto de estudio tome valores diferenciados en distintos subgrupos de la población, parece lógico pensar que sería más apropiado tomar una muestra independiente en cada uno de esos subgrupos. Por ejemplo, al investigar el número de horas que dedica una persona a formación y estudio, podría dividirse la población en grupos de edad y obtener una muestra en cada uno de los subgrupos definidos

Formalizando la idea introducida con anterioridad, el **muestreo estratificado** consiste en:

1. Dividir la población de  $N$  unidades en un cierto número de subgrupos a los que denominaremos estratos, de forma que las unidades que componen cada estrato sean lo más homogéneas posibles y que cada estrato no se solape, es decir, 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.
2. Seleccionar una muestra probabilística de cada estrato. La muestra de cada estrato es independiente de la muestra de cualquier otro estrato. Si la muestra de cada estrato es una muestra aleatoria simple (probabilidades iguales) tenemos el muestreo aleatorio estratificado.

*Nota: Si la variable en estudio realmente depende de los estratos formados, la reducción del error podría ser importante, pero si no es así, entonces no será efectiva. Por este motivo se dice que el muestreo estratificado es mejor cuando más homogéneo son los grupos dentro de ellos y más heterogéneos entre sí.*

En el muestreo estratificado surge el problema de cómo distribuir la muestra. A continuación se presentan algunos criterios:



- **Uniforme:** consiste en asignar el mismo tamaño muestral para todos los estratos. Se utiliza cuando los tamaños de los estratos, su variabilidad y coste de investigar una unidad, son similares.
- **Proporcional:** distribuye la muestra total en proporción al número de unidades de cada estrato. Cada estrato aparece representado en la muestra con el mismo peso que presenta en la población total.
- **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. Vemos que la estratificación y la forma de distribuir la muestra entre estratos puede producir importantes ganancias en precisión

Cabe señalar que los principales motivos para usar la estratificación son:

- Mejora la representatividad de la muestra, en lo que se refiere a las variables utilizadas en la estratificación.
- Como tomamos una muestra aleatoria representativa en cada subgrupo poblacional (estrato), nos permite obtener estimaciones para cada uno de estos subgrupos. Y agregando estas estimaciones tendremos resultados para los agregados de dichos subgrupos y para el total poblacional.
- 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 los trabajos de campo y de tareas administrativas.

Una exigencia para su aplicación es la necesidad de disponer de información auxiliar en el marco que permita definir los estratos.

### 13.2.2 Muestreo de Conglomerados sin submuestreo

En este caso la población compuesta por  $N$  unidades elementales, se divide también en  $M$  grupos de unidades mayores que se denominan conglomerados. Este tipo de muestreo consiste en la selección de muestras constituidas por todas las unidades elementales de  $m$  conglomerados elegidos entre los  $M$  disponibles.

Hay una serie de razones principales para la aplicación del **muestreo de conglomerados**:



- Facilita la elaboración del marco, puesto que sólo sería preciso tener en primera instancia un buen marco actualizado de los conglomerados.
- Disminuye el coste de los trabajos de campo debido a menor dispersión de la muestra sobre todo cuando la entrevista es personal
- Si la selección de los conglomerados se realiza con probabilidad proporcional al tamaño medido éste por una variable correlacionada con la variable objetivo, mejora la precisión de las estimaciones.

El principal inconveniente es que generalmente disminuye la precisión de las estimaciones debido al llamado efecto conglomerado. El motivo es que la representatividad de la muestra puede ser menor, para el mismo número de unidades elementales investigadas. Si las unidades que conforman cada conglomerado son todas uniformes dentro de él, no aporta mucha información preguntar a varios o a todos las unidades elementales del conglomerado. Sería preferible coger entonces una muestra aleatoria simple. En cambio si dentro de cada conglomerado hay bastante heterogeneidad y los conglomerados son homogéneos entre sí, entonces el muestreo por conglomerados proporciona el mismo error o menor que un muestreo aleatorio simple.


### 13.2.3 Muestreo por etapas: Muestreo de Conglomerados con submuestreo

El **muestreo de conglomerados con submuestreo** consiste en:

1. En la primera etapa se selecciona una muestra de  $m$  conglomerados de los  $M$  que constituyen la población. A las unidades de muestreo de esta etapa se las denomina unidades primarias o de primera etapa.
2. En la segunda etapa se selecciona de manera independiente en cada unidad primaria, una muestra de unidades elementales.

Dentro de las ventajas de este tipo de muestreo podemos mencionar:

- Resulta una estrategia más eficiente que el muestreo de conglomerados sin submuestreo cuando las unidades dentro de cada conglomerado son homogéneas
- Reduce el coste y tiempo en desplazamientos de los entrevistadores, cuando las entrevistas son personales.
- No es necesario disponer de un marco de unidades de la población objetivo. Basta con disponer de un buen marco para cada una de las etapas de la selección. Por



ejemplo, en las encuestas dirigidas a los hogares no es necesario disponer de un marco actualizado de todas las viviendas de la población, sino que basta con un listado de las secciones censales (unidades de primera etapa) y luego un listado de viviendas únicamente de las secciones seleccionadas en la primera etapa.

Como desventajas citamos:

- En general es menos eficiente que el muestreo aleatorio simple, aunque puede ser mejor que el muestreo de conglomerados sin submuestreo.
- Las fórmulas para el cálculo de las estimaciones y de la varianza son más complejas, en comparación con las de los diseños en una etapa.

#### 13.2.4 Muestreo Sistemático

El **muestreo sistemático** se refiere a un conjunto de procedimientos para seleccionar muestras de forma rápida y sencilla. La unidad de muestreo puede ser tanto la unidad elemental como el conglomerado.

Este muestreo parte de un marco donde están todas las unidades elementales. Supongamos que hay  $N$  unidades en dicho marco y se quiere recoger una muestra de tamaño  $n$ . Determinamos lo que se conoce con el nombre de **periodo**  $k = N/n$  (es dividir todo el marco en  $n$  grupos de tamaño  $k$ ) y se selecciona de forma aleatoria un número entre  $1, 2, 3, \dots, k$ . Ese número aleatorio determina la primera unidad de la muestra. El resto se van seleccionando avanzando de  $k$  en  $k$  posiciones en la lista que forma el marco. De esta manera se obtiene una muestra aleatoria con solo un arranque aleatorio. Por lo expuesto se observa que la selección de la primera unidad determina la muestra completa y existen  $k = N/n$  muestras posibles, todas ellas equiprobables ( $prob = 1/k$ ).

Para ilustrar este tipo de método de selección, suponemos que se dispone de una población de  $N=54$  unidades y de la que se quiere obtener una muestra de tamaño  $n=9$  unidades. El período sería ( $k = 54/9 = 6$ ). El siguiente paso sería seleccionar de manera aleatoria un número entre 1 y  $k = 6$ . Si el número seleccionado fuera el 2, las unidades muestrales seleccionadas, serían aquellas que en la lista de selección ocuparan las posiciones 2, 8, 14, 20, 26, 32, 38, 44 y 50. En este ejemplo existen seis distintas posibles muestras.

El muestreo sistemático es de fácil aplicación práctica y asegura además que la muestra se extiende a toda la población. El comportamiento de 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 tuviese un error de muestreo similar e incluso menor por su efecto distribuidor de la muestra.

Por otro lado, para el uso de muestreo sistemático no se precisa de información auxiliar disponible en el marco.

En poblaciones cuya ordenación tiene una componente periódica, hay que ser especialmente cuidadoso en el uso de muestreo sistemático ya que no proporciona necesariamente una muestra representativa. Por ejemplo, si en un marco de selección que recoge la población de hombres y de mujeres, estos están alternados y  $k$  es par, la muestra sistemática sólo tendrá hombres o mujeres, lo que no es una muestra representativa de la población estudiada.