Tema 4 **MUESTREO Y ESTIMACIÓN**

4.1 MUESTREO

Conceptos previos

Ventajas y limitaciones del muestreo

Tipos de muestreo

Métodos de muestreo probabilístico

Métodos de muestreo no probabilístico

4.2 INTRODUCCIÓN A LA TEORÍA DE LA ESTIMACIÓN

Estimación puntual y error de muestreo

Estimación por intervalos de confianza

Ejemplo de estimación de una media poblacional

Ejemplo de estimación de un porcentaje poblacional

MUESTREO y **ESTIMACIÓN**

Resumen: Son muchas las ocasiones en que desde la dirección de una empresa se necesita recurrir a encuestas para conocer parámetros o estadísticos de una población.

La **Teoría de Muestras** propone una serie de métodos para conseguir que la muestra seleccionada sea *representativa* de toda la población.

Por otra parte, y conocidos los datos de la muestra, la **Teoría de la Estimación** permite encontrar valores o intervalos de confianza para los parámetros poblacionales: media, porcentaje, etc que son de interés para la gestión empresarial.

4.1 **MUESTREO**

La Teoría de Muestras es una rama de la Estadística Descriptiva que se apoya en los siguientes **conceptos previos**

Población es un conjunto homogéneo de individuos o unidades experimentales de los que nos interesa analizar una o varias características medibles o no medibles. En la mayor parte de los casos, no va a ser posible disponer de todos los datos poblacionales por una o más de las siguientes causas:

1. Es imposible acceder a todos los individuos
2. La población es infinita (ejemplo: medidas sobre un objeto)
3. No se dispone de suficientes recursos económicos o de tiempo
4. El individuo se destruye en la prueba

En estas ocasiones, en lugar de hacer un **censo**, (un estudio exhaustivo de todos los individuos de la población), seleccionaremos un conjunto de elementos representativos que llamaremos **muestra**, siendo la **unidad muestral** cada uno de los individuos o fuentes de información, es decir, cada uno de los posibles componentes de la muestra.

Cuando la muestra está bien escogida, de los datos de la **encuesta** podemos obtener una información similar a la del censo con mayor rapidez y menor coste. Esto justifica que en la práctica el análisis de poblaciones grandes se haga preferentemente mediante muestreo.

La clave de un procedimiento de muestreo es garantizar que la muestra sea *representativa* de la población. Por tanto, cualquier información respecto a las diferencias entre sus elementos debe tenerse en cuenta para seleccionar la muestra.

Dos condiciones añadidas a la representatividad son:

- Que los elementos se elijan aleatoriamente y
- Que la muestra tenga el tamaño adecuado.

Ventajas y limitaciones del muestreo

La utilización de muestras presenta las siguientes ventajas:

1. Economía y rapidez
2. Posibilidad de encuestar a grandes poblaciones y núcleos urbanos
3. Calidad, por la posibilidad de cuidar más la precisión de la observación o medida de cada elemento

Por otra parte, los inconvenientes más importantes son:

1. Riesgo de que la muestra no se elija debidamente y resulte sesgada.
2. Necesidad de disponer de información adecuada de toda la población (listas, censos, ficheros, mapas, etc)
3. La utilización de una herramienta matemática compleja como es la Teoría de muestras.

Tipos de muestreo

Suelen considerarse tres tipos de muestreo:

Probabilístico

El principio de selección de elementos en una muestra aleatoria es el mismo del reparto de una baraja, esto es; todos los objetos de la población tienen la misma probabilidad de ser seleccionados para formar la muestra. A esta probabilidad se le llama razón o **fracción de muestreo** (sampling ratio), y es igual al número de elementos de la muestra dividido por el número de elementos de la población (f = n/N) o (f = 100n/N), expresado en porcentajes. Su inverso es el **coeficiente de elevación.**

El que la fracción de muestreo supere el valor del 5% es deseable, ya que las estimaciones sobre los datos de estas muestras son más precisas, como más adelante veremos.

Intencional u opinático

Es el investigador quien con su criterio o intención selecciona la muestra procurando que la representatividad responda a sus propios intereses. Su base teórica es insatisfactoria, pero se usa con cierta frecuencia sobre todo en el llamado muestreo por cuotas.

Sin norma, circunstancial o errático:

En el que se toma la muestra de cualquier forma por razones de comodidad, economía u otras circunstancias sin atender a ningún criterio científico, simplemente se toma una parte de la población. Es un método muy usado en el comercio, donde se supone que, por ejemplo, un trozo de tela o un vaso de vino representan eficazmente los artículos completos.

De estos tres tipos de muestreo, el ideal es el probabilístico, pero no siempre es fácil cumplir todos los requisitos que exige y es frecuenta utilizar **muestras mixtas** entre probabilísticas y opináticas o intencionales.

Los métodos o procedimientos de muestreo probabilístico son:

Muestreo aleatorio simple

Es el muestreo probabilístico básico aplicable a poblaciones homogéneas desde el punto de vista de la característica a estudiar. Todas las muestras y todos los elementos tienen la misma probabilidad de ser seleccionados, y la elección de un individuo no influye en la del siguiente.

La forma de llevarlo a cabo es asignar un número a cada uno de los individuos de la población y, seguidamente, se van eligiendo al azar los componentes de la muestra mediante cualquier procedimiento aleatorio (tabla nº 5 de números aleatorios del final de este volumen, programa de ordenador, etc)

Dependiendo de que las extracciones de hagan con o sin reemplazamiento, aparecen estos dos procedimientos de muestreo.

Muestreo sistemático

Similar al anterior, consiste en elaborar un listado de todos los elementos que incluye la población. La diferencia estriba en el método para la selección de los casos que se realiza por un procedimiento que representa un gran ahorro de tiempo. Para obtener una muestra de tamaño n, se ordenan previamente los individuos de la población N, después se elige uno al azar entre los k primeros, k es el entero más próximo a N/n, y a continuación, de k en k, se eligen todos los demás hasta completar la muestra. Si por ejemplo el primer elemento es el número h los siguientes son h+k, h+2k, y así hasta completar la muestra

Muestreo estratificado

Este tipo de muestreo es el adecuado cuando la población no es homogénea pero puede ser dividida en categorías, estratos o grupos homogéneos a los fines del estudio.

La ventaja es que mediante este tipo de muestreo se logra una muestra final más representativa de la población.

Dentro de cada estrato el procedimiento de selección se logra mediante las técnicas del muestreo aleatorio simple. Se puede estratificar por sexo, nivel socioeconómico, características geográficas, grupos de edades, etc. Se divide la población total en clases homogéneas (estratos)

El problema de dividir la muestra entre los diferentes estratos y decidir el tamaño muestral de cada uno se llama afijación de la muestra, que puede ser:

- **Afijación simple**: que consiste en dar el mismo tamaño muestral a todos los estratos. Este método es, en general poco recomendable.
- **Afijación proporcional:** los tamaños muestrales son proporcionales a la población de cada estrato.
- **Afijación óptima**: que consiste en dar tamaños muestrales proporcionalmente al tamaño y a la variabilidad de cada estrato si fuese conocida.

Muestreo por conglomerados

Se utiliza cuando la población está formada por conjuntos o conglomerados que tienen la particularidad de que cada uno de ellos representa muy bien a la población. Se selecciona un número reducido de conglomerados para representar a la población en el análisis según presupuesto.

Ejemplo: España se divide en tantos conglomerados como provincias tiene, una provincia es un conglomerado que representa a la población española.

Cuando el muestreo aleatorio simple puede producir una muestra cuyas unidades se encuentran diseminadas de tal modo que haga prohibitivo el costo de desplazamiento de los encuestadores se utiliza esta técnica.

En resumen; las ideas de *estratificación* y de *conglomerado* son opuestas:

La estratificación funciona tanto mejor cuanto mayor sean las diferencias entre los estratos y más homogéneos sean éstos internamente; los conglomerados funcionan si hay muy pocas diferencias entre ellos y son muy heterogéneos internamente (incluyen toda la variabilidad de la población de cada uno)

Muestreo polietápico

Consiste en la realización del muestreo en dos o más etapas. Se procede a la división de la población en grupos, a los cuales se aplica un muestreo aleatorio simple, aplicando posteriormente muestreos aleatorios sobre los componentes de los grupos que forman la primera muestra obtenida.

Este proceso se puede repetir sucesivas veces hasta conseguir muestras formadas por los elementos muestrales que interesa investigar.

Los métodos de muestreo no probabilístico más utilizados son:

Muestreo por cuotas

Cuando no existe o no se puede formar una base de la muestra pero se conoce la composición en estratos de la población (en tantos por ciento) y la encuesta se realiza por entrevista, se utiliza este método que consiste en asignar a cada entrevistador un

número de entrevistas a realizar, indicándole la que corresponde a cada estrato, pero dejando a su elección las unidades de la muestra (cuotas). En este método, el tamaño de la muestra debe ser un 50% superior al de un método probabilístico, debido a que el error que supone la muestra es mayor, al tratarse de un sistema imperfecto.

Presenta los siguientes inconvenientes:

- No pueden aplicarse fórmulas estadísticas
- Existe el peligro de que los entrevistadores elijan personas cercanas a ellos (comodidad)

Muestreo por rutas aleatorias

Es un tipo de muestreo de carácter semiprobabilístico, al combinar ciertos aspectos del muestreo probabilístico y no probabilístico. Sólo se puede emplear para seleccionar muestras dentro de las ciudades.

Consiste en elegir, aleatoriamente, dentro de una ciudad, diferentes puntos geográficos desde los que se va a partir para realizar la encuesta. El número de puntos de partida que suelen tomarse es variable, aunque se suele utilizar entre el 5 y el 10 por cien de los elementos que componen la muestra, de forma que cada ruta tenga de 10 a 20 entrevistas respectivamente.

Cada entrevistador tiene que realizar el número de encuestas fijado para cada ruta. Para ello, recibe unas instrucciones muy precisas y concretas sobre lo que tiene que hacer para realizar las encuestas, de manera que todos los entrevistadores actúen con la misma metodología.

Muestreo “bola de nieve”

Recomendado para el estudio de casos de interés especial en individuos que son difíciles de identificar, Son ejemplos: clientes de productos o servicios exclusivos, drogodependientes, etc.

La técnica consiste en localizar algunos individuos típicos, los cuales conducen a otros y así sucesivamente va creciendo la “bola de nieve” hasta conseguir una muestra adecuada.

4.2 **INTRODUCCIÓN A LA TEORÍA DE LA ESTIMACIÓN.**

Definidos los conceptos de población y muestra, variables cualitativas y cuantitativas, parámetros estadísticos y la información que aportan, es inmediato aceptar la importancia que tiene en todo proceso de investigación el conocimiento de los parámetros poblacionales:

- **Media**, precio medio al que se vende nuestro producto en 98 establecimientos
- **Varianza**, variabilidad de cierta variable de interés para el control de calidad
- **Proporción** o **porcentaje**, porcentaje de reclamaciones a nuestro servicio
- **Lambda de Poisson**, media de artículos defectuosos por lote, etc.

Como ya sabemos, en la mayor parte de los casos, no va a ser posible disponer de todos los datos poblacionales. Es por ello que trataremos de aproximarnos al conocimiento de dichos parámetros poblacionales de la forma más rigurosa posible.

La **Teoría de la Estimación** es una rama de la Inferencia Estadística que permite encontrar valores aproximados de estos parámetros desconocidos a partir de los datos muestrales. Esta estimación puede ser:

1. **Puntual**, que consiste en dar un valor para el parámetro desconocido, acompañado del error de muestreo, o bien
2. Por **intervalos de confianza**, que consiste en dar un intervalo dentro del cual se tiene un determinado nivel de confianza en que se encuentre el parámetro poblacional desconocido

La Teoría de la Estimación introduce los siguientes conceptos:

A partir de los datos muestrales se obtienen los **estadísticos o estimadores** de los parámetros normalmente por el **método analógico**, que propone el mismo procedimiento de cálculo con los n datos de la muestra que el que aplicaríamos caso de conocer los N datos poblacionales.

Ejemplo, puesto que la media poblacional es: $\mu = \sum_{i=1}^{N} x_i / N$, entonces su estimador es la media muestral $\bar{x} = \sum_{i=1}^{n} x_i / n = \hat{\mu}$. La elección de los mejores estimadores se basa en exigirles una serie de requisitos o condiciones como son:

- Que sean centrados o **insesgados,** (ejemplo: E($\bar{x}$) = μ, que nos garantiza que la media muestral tiende a la poblacional al aumentar n).
- Que su "variabilidad " sea la menor posible, (**mínima varianza**).
- Que sea mínimo su **error cuadrático medio**, (ECM = sesgo^2+varianza), etc.

El paso siguiente es encontrar la **distribución en el muestreo** de los estimadores seleccionados; en ello reside el fundamento matemático de la Teoría de la Estimación, que tiene como referencias básicas:

El **teorema de Fisher**, que demuestra que si una variable aleatoria es normal de media μ y desviación típica σ; X $\rightarrow$ N (μ, σ), se cumple:

1.- μ y σ son variables aleatorias independientes.

2.- El conjunto de las medias muestrales de tamaño n se ajustan, a su vez, a una nueva distribución normal N (μ, σ / $\sqrt{n}$)

3.- $n s^2 / \sigma^2 = (n-1)\ S^2 / \sigma^2 \cong \chi^2_{n-1}$

El **teorema central del límite**, que demuestra que aunque la variable aleatoria X no se ajuste a una distribución normal, si las muestras son de suficiente tamaño (n > 30), las medias muestrales son normales N (μ,σ / $\sqrt{n}$).

Gosset, trabajando con **muestras pequeñas** en la fábrica de cervezas Guinness propone que para n < 30, las medias muestrales se ajustan a la distribución t de Student que él mismo definió.

Finalmente las **aproximaciones** de la distribución binomial y de la distribución de Poisson a la normal permiten hacer estimaciones fiables para los parámetros

poblacionales (P proporción ó 100 x P porcentaje) y λ de Poisson si se cumplen las condiciones para que tales aproximaciones sean válidas y que ya vimos en el tema 2.

Otro concepto básico para hacer estimaciones es la **fracción de muestreo** f = n / N o bien f = 100 n / N (en %).

- Si f ≤ 5 % hablaremos de **población infinita** o muy grande en comparación con la muestra por no alcanzar al 5 % de los individuos.
- Si f > 5 % diremos que se trata de una **población finita**; para cada caso cambian las fórmulas a emplear.

El **error de muestreo** que acompaña a la estimación por punto, (ejemplo: $\sigma / \sqrt{n}$ ó $S / \sqrt{n}$ al estimar la media), es tanto menor cuanto mayor sea el tamaño muestral n y, consiguientemente, el coste del muestreo.

Estimación puntual

Consiste en proponer un valor para el parámetro poblacional desconocido, acompañado del error de muestreo.

Un ejemplo de enorme interés para toda empresa es la estimación del I.P.C. que cada mes hace el Instituto Nacional de Estadística por la gran cantidad de variables económicas en las que repercute. También las estimaciones que hacen las empresas dedicadas a estudios sociológicos acerca del porcentaje de votantes que es de esperar tengan los diferentes partidos políticos en los días previos a los procesos electorales, etc.

Las empresas estarán también interesadas en estimaciones sobre la **media** del precio al que se venden los productos que comercializan en distintos establecimientos, **proporción o porcentaje** de clientes que presentan reclamaciones a los servicios que prestan, media de artículos defectuosos en cualquier proceso productivo, **variabilidad** de precios, medidas efectuadas para el control de calidad, etc.

Estimación por intervalos de confianza

Para los mismos supuestos de la estimación puntual, dado que en muchos casos los estimadores propuestos presentan poca variabilidad alrededor del parámetro poblacional, cabe esperar que el valor del estadístico calculado para una muestra concreta, esté próximo al verdadero valor del parámetro que se desea estimar.

Es por ello que a partir de este valor del estadístico y de su distribución, puedan proponerse los límites inferior y superior de un intervalo, que con un cierto nivel de confianza contenga al verdadero valor del parámetro.

El **nivel de confianza**, normalmente 95%, 95.4% (para error de muestreo de dos desviaciones típicas, o dos sigmas) ó 99 %, hace referencia al grado de probabilidad o de "seguridad " que tememos de que contenga el verdadero valor del parámetro.

El cálculo del **tamaño de la muestra** necesario para poder estimar un parámetro desconocido, es un problema interesante y, en todo caso, previo al de la estimación.

Dependiendo del nivel de confianza que asignemos al intervalo y del máximo error que estemos dispuestos a admitir, (semiamplitud del intervalo), la correspondiente fórmula o consulta en la correspondiente tabla (ejemplo la del final del tema), nos permitirá saber cuántos datos muestrales necesitamos y, todo ello, tras encontrar el punto de equilibrio óptimo entre la precisión que exijamos a la estimación y el coste que estemos dispuestos a asumir para el muestreo.

Ejemplo de estimación de una media poblacional

El producto que comercializamos se vende en 660 establecimientos. Al objeto de estimar su precio medio efectuamos un muestreo aleatorio simple que nos proporciona una muestra piloto de n = 10 establecimientos de la que se obtiene una desviación típica de los precios de 0.80 €. Se pretende:

1. Calcular el **tamaño muestral** necesario para estimar el precio medio con un error máximo de 0.30 € y con un nivel de confianza del 95 %.

Solución: El tamaño muestral n se obtiene mediante

$$\mathbf{n} = z^2_{\frac{\alpha}{2}} \sigma^2 / e^2 = 1.96^2\, 0.8^2 / 0.3^2 = 27.32 = \mathbf{28} \text{ establecimientos}$$

(del 95 % de nivel de confianza se deduce el nivel de significación $\alpha = 0.05$, y tras consultar la tabla 1 de la normal tipificada N (0 , 1) se obtiene $z_{\frac{\alpha}{2}} = 1.96$).

nota: en el cálculo de tamaños muestrales, el redondeo debe hacerse al entero siguiente.

Elegidos al azar los 18 establecimientos necesarios para completar la muestra, se obtienen los datos relativos al precio de venta:

(3.45 / 3.50 / 3.20 / . . . . . /4.10 euros)

A partir de los cuales, podemos calcular: n = 28, $\bar{x}$ = 3.60 €, y S = 0.82 €

2. **Estimación puntual** del precio medio al que se vende el producto en los 660 establecimientos:

En primer lugar debemos calcular la fracción de muestreo:

f = 100 n / N = 100 x 28 / 660 = 4 % < 5 %, población infinita (muestra muy pequeña comparada con la población), entonces el error de muestreo es:

S / $\sqrt{n}$ = 2.052 x 0.82 / $\sqrt{28}$ = 0.15 €

Al ser la esperanza matemática de la media muestral la media poblacional, E ($\bar{x}$) = μ y la media poblacional estará cerca de 3.60 €, podemos concluir:

Analizada la muestra de 28 establecimientos, el precio medio al que se vende el producto estará próximo a 3.60 €, con un error de muestreo de +/- 0.15 €.

3. **Estimación por intervalos** de 95 y 99 % de confianza del precio medio:

Al ser n = 28 < 30, muestra pequeña:$\mu \in (\bar{x} - t_{\frac{\alpha}{2}} S / \sqrt{n} \; ; \; \bar{x} + t_{\frac{\alpha}{2}} S / \sqrt{n})$.

consultada la tabla 2 de la t de Student se obtiene:

$t_{\frac{\alpha}{2}=0.025}$ (n-1 = 27 g.l.) = 2.052

$t_{\frac{\alpha}{2}=0.005}$ (n-1 = 27 g.l.) = 2.771

Intervalo de 95 % de confianza

$\mu \in$ 3.60 +/- 2.052 x 0.82 / $\sqrt{28}$ = (3.28 ; 3.92 €)

Intervalo de 99 % de confianza

$\mu \in$ 3.60 +/- 2.771 x 0.82 / $\sqrt{28}$ = (3.17 ; 4.03 €)

El segundo intervalo es menos preciso al habérsele exigido mayor nivel de confianza.

Conclusión: ***Con la muestra analizada y al 99 % de confianza, el precio medio al que se vende el producto en los 660 establecimientos está entre 3.17 y 4.03 €***

Ejemplo de estimación de un porcentaje poblacional

i) **Cálculo de tamaños muestrales**

Nuestra empresa se propone hacer una investigación de mercado para determinar el porcentaje de televidentes de un total de 150.000 de cierto programa que vio nuestro anuncio. ¿Qué número de personas deberá ser preguntado si no queremos aceptar un error superior al 1%? Distinguir los casos:

1. Se dispone del formulario y no se tiene muestra piloto

2. Una muestra piloto indicó que tal porcentaje no supera, en ningún caso el 30%

Solución:

1. Con la ayuda del formulario sin información previa:

$$n = \frac{z_{\frac{\alpha}{2}}^2 PQ}{e^2} = \frac{1.96^2 \times 50 \times 50}{1^2} = \mathbf{9.604}$$

Con este tamaño muestral la fracción de muestreo es:

9604/150000 = 0.064(6.4%) > 5%, población finita, por lo que debemos aplicar:

$$n = \frac{n_i}{1 + \frac{n_i}{N}} = \frac{9604}{1 + 9604/150000} = 9.026{,}09 = \mathbf{9.027}$$

2. Con la información de la muestra piloto:

$$n = \frac{z_{\frac{\alpha}{2}}^2 PQ}{e^2} = \frac{1.96^2 \times 30 \times 70}{1^2} = 8.067{,}36 = \mathbf{8.068}$$

Con este tamaño muestral la fracción de muestreo es:

8068/150000 = 0.0538, (5.38%) > 5%, población finita, por lo que debemos aplicar:

$$n = \frac{8068}{1 + 8068/150000} = 7.656{,}20 = \mathbf{7.657}$$

Obsérvese la reducción del coste del muestreo que se consigue con esta información

ii) **Estimación puntual y por intervalos de confianza**

Nuestra empresa tiene una cartera de 720 clientes. Elegidos 120 al azar, 75 de ellos manifestaron estar muy satisfechos con la atención recibida de nuestros empleados.

Se pide:

1. Fracción de muestreo e interpretación

2. Estimación puntual y por intervalos del 95 y 99% de confianza del porcentaje de clientes muy satisfechos

Solución:

1. La fracción de muestreo es para este caso: f = 100 n/N = 100 120/720 = **16.67%**, que al superar al 5% nos indica que **la población es finita**.

2. **Estación puntual**

Es de esperar que el porcentaje poblacional esté próximo al muestral

P = 75 / 120 = 0.625 = **62.5%**, con un error de muestreo:

$$\sigma_{\overleftarrow{p}} = \sqrt{\frac{PQ}{n}}\ \sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{62{,}5\times 37{,}5}{120}}\ \sqrt{\frac{720-120}{719}} = \pm \mathbf{4.04\%}$$

Estimación por intervalos de confianza

$$p \in (\ \hat{p} \pm z_{\frac{\alpha}{2}}\sigma_p) = (62.5 \pm z_{\alpha}\ 4.05)$$

$z_{\alpha/2}$ se obtiene de la tabla 1 de la normal N (0, 1), que da los valores 1.96 para el 95%, y 2.576 para el 99% de confianza respectivamente. Y sustituyendo valores se concluye:

Es de esperar que el porcentaje de clientes muy satisfechos esté entre el 54.56 y 70.44% al 95% de confianza, y entre el 52.07 y 72.93% al 99% de confianza.

Al aumentar el nivel de confianza disminuye la precisión del intervalo.