TEMA 3

INTRODUCCIÓN A LA TEORÍA DE MUESTRAS

3.1 INTRODUCCIÓN

Con este tema comienza la INFERENCIA ESTADÍSTICA, que es la parte que se ocupa de confrontar los resultados que se derivan de un modelo estadístico con sus análogos en una realidad observable, con el objetivo de poder ajustar a dicha realidad un modelo adecuado que la explique. Para ello se parte de una cierta población sobre la que está definido un modelo estadístico que no es completamente conocido y se extrae una muestra mediante algún método aleatorio, con cuya información se pretende inferir alguna cuestión desconocida del modelo planteado.

De este modo, en el Cálculo de Probabilidades se partía de un modelo estadístico y la cuestión era ¿qué podemos afirmar sobre los resultados?; en la Inferencia Estadística, se parte de un conjunto de resultados observados y nos preguntamos ¿qué podemos afirmar sobre el modelo estadístico?

En este curso vamos a abordar dos técnicas inferenciales básicas: estimación y contraste de hipótesis. La primera consiste en construir unas funciones de los valores muestrales cuyo objetivo es estimar (aproximar) las diversas características desconocidas de la variable poblacional. Un concepto ligado a la estimación es la precisión, es decir, la mayor o menor concentración de los posibles valores del estimador respecto del parámetro que se pretende estimar. Esta forma de trabajar se conoce como Estimación Puntual. Otra forma de actuar consiste en construir un intervalo que contenga al verdadero valor del parámetro con una probabilidad fijada de antemano; en tal caso se habla de Estimación por Intervalos de Confianza.

La segunda técnica inferencial es el contraste de hipótesis, que se basa en comparar los resultados observados con los resultados que se obtendrían bajo una hipótesis de trabajo preestablecida. Si el resultado observado es coherente con lo esperado entonces no podemos rechazar la hipótesis. Sin embargo, si el resultado es anómalo con respecto a la hipótesis planteada entonces deberemos rechazarla y modificar nuestras creencias sobre la población. En definitiva, un contraste de hipótesis

es una regla de decisión acerca de una hipótesis, por lo tanto, deberemos tener en cuenta las posibles equivocaciones al tomar la decisión y sus consecuencias, cuantificando las garantías del proceso.

3.2 CONCEPTOS BÁSICOS

Una **población** es un conjunto de elementos con una o más características en común. La población se caracteriza mediante una variable aleatoria X y ésta por su distribución de probabilidad que especifica el comportamiento aleatorio de la población. Esta distribución puede ser desconocida totalmente o parcialmente. En cualquier caso, el objetivo del estudio será conocer la distribución de probabilidad de la variable X.

Una **muestra** es un subconjunto finito de los elementos de una población. Si la muestra contiene n elementos, que denotaremos por $(X_1,X_2,...X_n)$, entonces diremos que la muestra tiene tamaño n. Llamaremos espacio muestral al conjunto de todas las posibles muestras aleatorias de tamaño n. La muestra será el conjunto de información que utilizaremos para conocer la distribución de la variable X definida en la población.

Se llama **muestreo** al procedimiento de elección de los elementos que formarán la muestra y puede ser probabilístico o no. El muestreo es probabilístico si podemos conocer la probabilidad de extraer cada uno de los elementos de la muestra (por tanto, la probabilidad de cada una de las muestras posibles) mientras que en el caso de muestreo no probabilístico, los elementos se seleccionan mediante un criterio concreto e intentando que sea representativa de la población. Cuando el muestreo es probabilístico entonces podremos conocer, en términos de probabilidad, el error que se comete al utilizar la muestra como representación de la población y podemos generalizar los resultados de la muestra a la población.

3.2.1 Tipos de muestreo probabilístico

Existen varios tipos de muestreo probabilísticos como son el muestreo aleatorio simple (con y sin reposición), muestreo sistemático, muestreo estratificado, muestreo por conglomerados, muestreo por unidad monetaria, muestreo polietápico, … En este punto solo vamos a considerar el más importante que es el muestreo aleatorio simple.

El método de muestreo más utilizado es el **muestreo aleatorio simple con reposición** (m.a.s.). Éste consiste en extraer al azar y con la misma probabilidad entre

todos los elementos que conforman la población. Se observan las características necesarias y se devuelve a la población antes de extraer el siguiente elemento que integrará la muestra. De esta forma, la población no cambia en cada extracción y el elemento puede ser elegido de nuevo. Por lo tanto, la muestra aleatoria extraída (X_1,X_2,...X_n) está formada por n variables aleatorias que son independientes e idénticamente distribuidas según X que es la variable poblacional de interés.

Este tipo de muestreo se aplica principalmente cuando la población tiene un número elevado de elementos o es infinita, por lo tanto, la probabilidad de repetir uno es prácticamente cero.

El **muestreo aleatorio simple sin reposición** consiste en extrae al azar y con la misma probabilidad entre todos los elementos de la población. Se observan las características necesarias y no se devuelve a la población, es decir, se elimina para extraer el siguiente elemento muestral. De esta forma, la población cambia en cada extracción y el elemento solo puede ser seleccionado una vez, siendo cada elemento muestral un resultado dependiente de los observados anteriormente. Por lo tanto, la muestra aleatoria extraída (X_1,X_2,...X_n) está formada por n variables aleatorias que son dependientes e idénticamente distribuidas según X que es la variable poblacional de interés.

Este tipo de muestreo se utiliza principalmente cuando la población es finita, es decir, cuando el número de elementos de la población no es elevado. De esta forma, aseguramos que toda la información de la muestra es distinta y representativa de la población.

El **muestreo aleatorio sistemático** consiste en dividir la población en subconjuntos de tamaño k, siendo k el cociente entre el tamaño poblacional y el tamaño muestral: k=N/n. De esta forma basta seleccionar un elemento en cada subconjunto para tener una muestra aleatoria de tamaño n. Se procede eligiendo al azar un elemento en el primer conjunto y se repite la misma elección en todos los demás subconjuntos. Por ejemplo, si se elige el tercer elemento del primer subconjunto, se elegirán los terceros elementos del resto de subconjuntos. Es un método muy sencillo porque solo necesita una única elección al azar y los resultados son comparables a los muestreos aleatorios simples. Solo es desaconsejable cuando la población presenta algún tipo de periodicidad

porque al ser un muestreo periódico (de k en k) podemos incorporar alguna homogeneidad que no se da realmente.

El **muestreo aleatorio estratificado** consiste en dividir la población en subconjuntos que llamamos estratos. Los elementos poblacionales de un estrato son semejantes entre sí con respecto a la variable de estudio y diferentes a otros estratos. Para realizar esta división es necesaria cierta información de la población. Los estratos más comunes son el género (hombre/mujer), grupos de edad (jóvenes, adultos, mayores), nivel de estudios, estado civil, … A continuación se selecciona dentro de cada estrato un número de elementos utilizando un muestreo aleatorio simple y la unión de todas las muestras de cada estrato será la muestra total. De esta forma nos aseguramos que en la muestra existen todo tipo de comportamientos de la variable que se refleja en la construcción de los estratos. La distribución de la muestra entre los diferentes estratos se llama afijación y, habitualmente, hay tres formas de calcularla:

1. Afijación uniforme: Se seleccionan el mismo número de elementos de cada estrato, es decir, si n es el tamaño muestral necesario y L es el número de estratos entonces de cada estrato se toman al azar n/L elementos.
2. Afijación proporcional: Se selecciona proporcionalmente al tamaño de cada estrato, es decir, si un estrato representa el 10% de la población entonces se toma el 10% de la muestra en dicho estrato; si otro estrato supone un 30% de la población se toma un 30% de la muestra en ese estrato y así sucesivamente.
3. Afijación de mínima varianza: Se selecciona en cada estrato en función de su variabilidad, es decir, de los estratos con mayor varianza se toman más elementos muestrales y de los estratos con menor variabilidad se seleccionan menos elementos muestrales. Además, el tamaño muestral en cada estrato también dependerá del número de elementos poblaciones que hay en el estrato. De esta forma la varianza del estadístico es mínima lo que da mayor precisión a los resultados.

El **muestreo aleatorio por conglomerados** se emplea cuando tenemos acceso a conjuntos de dos o más elementos poblacionales. Por ejemplo, demarcaciones

geográficas, unidades de trabajo, departamentos, lotes de productos, … son conglomerados naturales que nos permiten acceder posteriormente a los elementos individuales: pueblos, empleados, trabajadores, piezas concretas, … El muestreo se realiza seleccionando al azar conglomerados hasta completar el tamaño muestral necesario, es decir, se eligen dos, tres, cuatro o más conglomerados y se estudian todos los elementos que forman parte de esos conglomerados. En este tipo de muestreo se supone que dentro de cada conglomerado tenemos elementos heterogéneos respecto a la variable en estudio y los diferentes conglomerados son similares entre sí (justo lo contrario que en el muestreo aleatorio estratificado).

Por último, cabe recordar que el proceso de muestreo se puede realizar en varias etapas dependiendo de cómo accedemos a la información sobre los elementos poblacionales. Si tenemos conglomerados inicialmente y decidimos aplicar otro tipo de muestreo dentro de cada conglomerado para tener una muestra más representativa entonces estaríamos formulando un **muestreo bietápico**. Por ejemplo, queremos estudiar los resultados de selectividad. Inicialmente, la información viene dada por tribunales (conglomerados) y dentro de cada uno podríamos seleccionar una muestra de estudiantes mediante muestreo aleatorio simple. Este proceso se puede generalizar dando lugar a un **muestreo trietápico** o, en general, **muestreo polietápico**. Por ejemplo, se seleccionan comarcas, dentro de cada comarca en una segunda etapa seleccionamos pueblos, dentro de cada pueblo en una tercera etapa seleccionamos comercios y, por último, en cada comercio medimos las ventas de unos productos concretos.

3.2.2 Estadístico

Un **estadístico** es cualquier función de los valores muestrales, siempre que no dependa de parámetros o constantes desconocidas. Se emplea para resumir y comprender la información contenida en la muestra. Cuando se fija la muestra tendremos un valor numérico para el estadístico. Así pues, podemos representar un estadístico como: $T=T(X_1,X_2,...X_n)$. El estadístico se construye como una función de variables aleatorias, por lo tanto, es otra variable aleatoria y su distribución dependerá de la propia variable aleatoria poblacional. Esto implica que el conocimiento de la distribución del estadístico en el muestreo y su posterior observación nos puede ayudar a inferir las características desconocidas de la población. Además, sabremos en términos

probabilísticos los correspondientes errores de aproximación entre el valor del estadístico y el valor real del parámetro desconocido.

Los estadísticos más usuales son aquellos que ya hemos utilizado en la Estadística Descriptiva:

La **MEDIA MUESTRAL** es la media aritmética de los valores muestrales observados: $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$,e indica la posición central de la muestra.

La **VARIANZA MUESTRAL** es la varianza de los valores muestrales observados: $S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$, e indica la variabilidad o dispersión de la muestra.

Si la variable aleatoria poblacional sigue una distribución Bernoulli entonces su media es el parámetro p, que indica la proporción poblacional de elementos que tienen una cierta característica. En este caso concreto, el estadístico media muestral se denomina **PROPORCIÓN MUESTRAL**: $\hat{p} = \frac{1}{n}\sum_{i=1}^{n} X_i$, e indica la proporción de elementos muestrales que poseen la característica de interés.

3.3 CARACTERÍSTICAS DE UN ESTADÍSTICO EN EL MUESTREO

Como hemos dicho anteriormente, un estadístico es una variable aleatoria, lo cual implica que tendrá asociada una función de distribución y sus correspondientes características: media, varianza, ... Éstas estarán relacionadas con la correspondiente distribución poblacional y sus características, por lo tanto, el estudio de la distribución de un estadístico nos puede permitir conocer como nos aproximaremos a la distribución poblacional o a sus propiedades.

En primer lugar, abordaremos el estudio de las características (media y varianza) de un estadístico en el muestreo y, posteriormente, plantearemos los métodos para el cálculo de la distribución.

Sea X una variable aleatoria poblacional cuya media es μ y cuya desviación típica es σ. Se extrae una m.a.s. de tamaño n y con los valores observados se construye el estadístico media muestral. Comenzamos por plantear las características de este estadístico. La esperanza de la media muestral coincide con la media poblacional:

$$E\left[\overline{X}\right] = \mu$$

Esto implica que la media muestral proporcionará valores aleatorios que se sitúan alrededor de la media poblacional, unas veces por debajo y otras veces por encima.

La varianza de la media muestral es el cociente entre la varianza poblacional y el tamaño muestral:

$$V\left[\overline{X}\right] = \sigma^2 \big/ n$$

Esto implica que la dispersión es directamente proporcional a la variabilidad poblacional e inversamente proporcional al tamaño muestral, permitiendo que al aumentar el tamaño muestral nos podamos aproximar más al verdadero valor de la media poblacional.

Los resultados anteriores son igualmente válidos para la proporción muestral sustituyendo la media y la varianza poblacional por sus respectivos valores según la distribución Bernoulli:

$$E\left[\hat{p}\right] = p \quad y \quad V\left[\hat{p}\right] = \frac{p(1-p)}{n}$$

En cuanto a la varianza muestral sólo vamos a calcular su esperanza. Ésta es la varianza poblacional multiplicada por uno menos el inverso del tamaño muestral:

$$E\left[S^2\right] = \sigma^2 \frac{n-1}{n} = \sigma^2\left(1-\frac{1}{n}\right) = \frac{n-1}{n}\sigma^2$$

Este resultado nos indica que la varianza muestral proporciona valores alrededor de un punto inferior al deseado, es decir, que dicho estadístico tiende a infravalorar a la varianza poblacional. Por lo tanto, deberíamos modificar la varianza muestral para conseguir un estadístico cuya esperanza sea la propia varianza poblacional. Este nuevo estadístico es la **CUASIVARIANZA MUESTRAL**, que se define como:

$$S_1^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$$

Este estadístico verifica que su valor esperado coincide con el parámetro varianza poblacional: $E\left[S_1^2\right] = \sigma^2$. Por lo tanto, la cuasivarianza muestral proporcionará valores alrededor de la varianza poblacional, unas veces por encima y otras por debajo.

3.4 DISTRIBUCIÓN DE UN ESTADÍSTICO EN EL MUESTREO

En este epígrafe vamos a presentar los métodos para calcular las distribuciones de los estadísticos en el muestreo. Se pueden clasificar como métodos exactos o métodos aproximados, según sea exacta o aproximada la distribución del estadístico analizado. En principio presentaremos los métodos exactos, aunque en este curso no nos preocuparemos de ellos, porque necesitan desarrollos teóricos complicados. Dentro de los Métodos exactos, se pueden distinguir los Métodos que utilizan el cambio de variable como herramienta matemática, o los Métodos que utilizan la Función Generatriz de Momentos (No vista en este curso).

3.4.0. Métodos exactos

Consiste en obtener la distribución del estadístico mediante un cambio de variable y utilizando herramientas del Cálculo de Probabilidades. Se conocen pocos casos en los que este método se puede aplicar. Algunos de los que utilizaremos son los siguientes:

a) Si $X_1, \ldots, X_n$ es una m.a.s. de $X \sim Bi(n,p)$ entonces $n\hat{p}_n = S_n \sim Bi(n,p)$

b) Si $X_1, \ldots, X_n$ es una m.a.s. de $X \sim Poisson(\lambda)$ entonces $nx_n = S_n \sim Poisson(n\lambda)$

c) Si $X_1, \ldots, X_n$ es una m.a.s. de $X \sim N(\mu,\sigma)$ entonces $S_n \sim N(n\mu, n\sigma)$ y $\overline{X_n} \sim N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$

3.4 DISTRIBUCIÓN DE UN ESTADÍSTICO EN EL MUESTREO

En este epígrafe vamos a presentar los métodos para calcular las distribuciones de los estadísticos en el muestreo. Se pueden clasificar como métodos exactos o métodos aproximados, según sea exacta o aproximada la distribución del estadístico analizado. En este curso solo nos preocuparemos de los métodos aproximados que son más sencillos porque no necesitan de desarrollos teóricos complicados.

3.4.1 MÉTODO DE MONTE CARLO o MUESTREO ARTIFICIAL

Este método consiste en la aproximación de la distribución mediante el estudio empírico de los valores del estadístico: histogramas, medidas numéricas, ... Se basa en la obtención de valores del estadístico mediante la simulación de muestras aleatorias. Cada muestra aleatoria simulada proporciona un valor del estadístico. Podemos obtener tantos valores como queramos y, a partir de esos valores, podemos dibujar un histograma que aproxime su función de cuantía o función de densidad y podemos calcular las medidas numéricas (media, varianza, asimetría, curtosis, ...) que nos permitan conocer las propiedades de la distribución para su aproximación a un modelo estadístico.

3.4.2 MÉTODO ASINTÓTICO

Este método consiste en aproximar la distribución del estadístico cuando el tamaño muestral es elevado. Se basa en el Teorema Central del Límite que afirma que la suma de variables aleatorias X_i independientes e idénticamente distribuidas con media y varianza finitas tiene una distribución aproximadamente normal cuando n tiende a infinito:

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sigma\sqrt{n}} \approx N(0,1) \quad \text{cuando } n \to \infty$$

Este resultado nos permite conocer la distribución aproximada de la media muestral, puesto que se expresa como una suma de variables aleatorias independientes e idénticamente distribuidas que son los valores muestrales:

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \approx N(0,1) \quad \text{cuando } n \to \infty$$

3.5 DETERMINACIÓN DEL TAMAÑO MUESTRAL

Una cuestión importante que no hemos abordado todavía es el tamaño muestral, es decir, la cantidad de valores que tenemos que observar para "confiar" en los resultados que obtengamos. En primer lugar, recordamos que un estadístico es una variable aleatoria y, por lo tanto, sus resultados debemos valorarlos en términos probabilísticos. Para ello será necesario definir previamente dos términos: el error máximo permitido (e) y el nivel de confianza (1-α). El primero indica cuál es la máxima diferencia admitida entre el estadístico y el parámetro que aproximamos; y el segundo representa la probabilidad con la que aseguramos el resultado sobre el error. La expresión que utilizaremos será del siguiente tipo:

$$P\{|T-\theta| < e\} = 1-\alpha$$

Con una probabilidad del (1-α)100%, el estadístico T no difiere en más de e unidades del verdadero valor θ, el parámetro de interés.

Generalmente, los estudios se centran como resultado principal en la media de la variable poblacional. Por este motivo, el estadístico que utilizaremos es la media muestral y el parámetro será μ=E[X]. A partir de la ecuación anterior, nos planteamos la determinación del tamaño muestral necesario y, en este punto, vamos a suponer que la media muestral sigue una distribución normal, bien porque la población sigue una v.a. normal o bien porque el tamaño muestral va a ser elevado y entonces el Teorema Central del Límite nos avalaría nuestro resultado.

Sabemos o aproximamos la distribución de la media muestral a una normal

Si suponemos que $\overline{X} \sim N\left(\mu, \sigma/\sqrt{n}\right)$ entonces podemos afirmar que:

$$P\left\{|\overline{X}-\mu| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha$$

Fijado el nivel de confianza podremos calcular el valor del percentil $z_{\alpha/2}$ de la distribución normal estándar y fijado el error máximo permitido despejamos el tamaño muestral necesario:

$$n = \frac{z_{\alpha/2}^2 \sigma^2}{e^2}$$

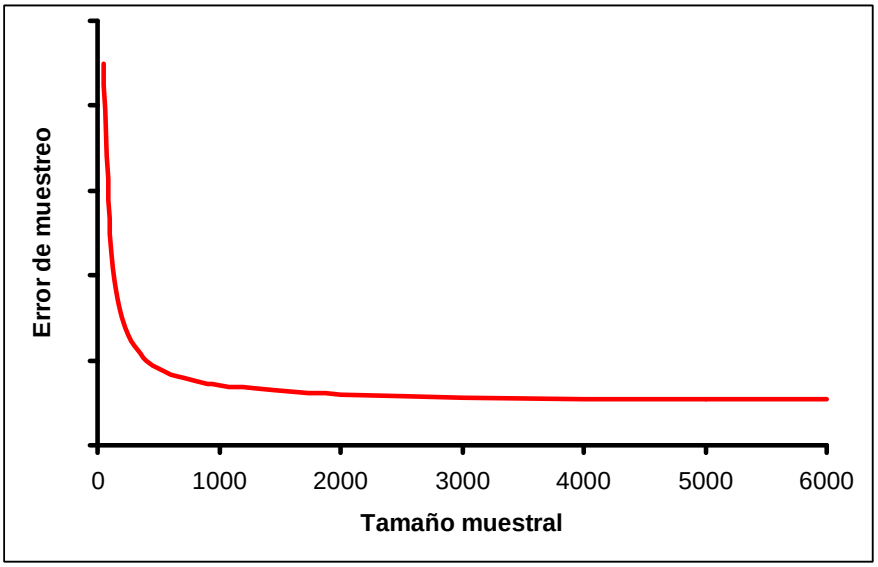
El tamaño muestral es directamente proporcional al nivel de confianza e inversamente proporcional al error máximo permitido. Así pues, un aumento de la confianza produce un incremento en el tamaño muestral pero un aumento en el error estaría disminuyendo el tamaño muestral necesario.

Otra cuestión relevante es el desconocimiento de la varianza poblacional σ^2 en la mayoría de las situaciones. Para solucionar este problema existen diferentes alternativas:

- Estimar σ^2 por los valores obtenidos en estudios similares o anteriores.
- Estimar σ^2 mediante una muestra piloto.
- Estimar σ^2 por el máximo valor posible si podemos acotarla por el conocimiento sobre el problema planteado. Este último caso es el habitualmente usado cuando la variable poblacional sigue una distribución Bernoulli. En concreto, la máxima varianza de una Bernoulli es 0,25 que se alcanza cuando p=0,5 (el máximo desconocimiento posible en una dicotómica es asignar el 50% de probabilidad a cada una de las dos categorías posibles).

$$\max_p V[X] = \max_p p(1-p) = \frac{1}{4} \text{ que se alcanza cuando } p = \frac{1}{2}$$

Por otro lado, hay que advertir que el incremento del tamaño muestral a partir de un límite no reporta una mejora en la precisión de la estimación. Así por ejemplo, podemos observar el siguiente gráfico donde se representa el error de muestreo en función del tamaño muestral, y se advierte que el error de muestreo no disminuye significativamente al aumentar el tamaño muestral a partir de un nivel.



Además, hay que tener en cuenta que en el proceso de muestreo hay un coste por observación, así que elevar el tamaño muestral supone un incremento en el coste del estudio pero éste no va asociado a un beneficio real en las inferencias que vayamos a plantear.