TEMA 3 – Introducción a la TEORÍA DE MUESTRAS

3.1 INTRODUCCIÓN

La **Inferencia estadística** es la parte de la Estadística que estudia las propiedades de una población analizando una muestra de esa población.

¿Por qué? - Sencillez y ahorro: Más barato analizar una muestra que una población
- Ciertos análisis destruyen las muestras

Dos técnicas inferenciales básicas:

- Estimación: A partir de unas funciones de los valores muestrales cuyo objetivo es estimar algún parámetro desconocido de la población .,,
  … dando un único valor (Estimación Puntual = Tema 4)
  … dando un intervalo (Estimación por Intervalos = Tema 5)
- Contraste de hipótesis: Compara los resultados observados con los resultados que se obtendrían bajo una hipótesis de trabajo preestablecida.
  ⇒ Si el resultado observado es coherente aceptamos la hipótesis,
  o si no es coherente la rechazamos. (Temas 6 y 7)

3.2 CONCEPTOS BÁSICOS

- Una **población** es el conjunto de elementos caracterizada mediante una variable aleatoria X (y ésta por su distribución de probabilidad, que puede ser desconocida totalmente o parcialmente).
  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.
  Una muestra de tamaño n contiene n elementos, que denotaremos por $(X_1,X_2,...X_n)$
  - Espacio muestral: Conjunto de todas las posibles muestras aleat. de tamaño n
  - La muestra es la inform. usada para conocer la distrib. de la var. X de la pobl.

- **Muestreo** es el procedimiento de elección de los elementos de la muestra
  - Muestreo probabilístico:
  Podemos conocer la probabilidad de extraer cada uno de los elementos de la muestra (y por tanto, la probabilidad de cada una de las muestras posibles)
  Podremos conocer, en términos de probabilidad, el error cometido al utilizar la muestra como representación de la población
  - Muestreo no probabilístico: Los elementos se seleccionan mediante un criterio concreto e intentando que sea representativa de la población.

- **Tipos de muestreo probabilístico**
  - muestreo aleat. simple (con y sin reposición),
  - muestreo sistemático (*se numeran los elementos y se toma un elemento cada K elem.*)
  - muestreo estratificado (* *la población se subdivide en estratos, * en cada estrato los individuos son uniformes entre sí, * a cada estrato se le asigna una cuota*)
  - muestreo por conglomerados o áreas (*se hacen grupos de elem. próximos geográf. entre ellos, por. ej. por provincias, y se estudian unas prov. que representen al país*),
  - muestreo por unidad monetaria, - muestreo polietápico, …

El más utilizado es el **muestreo aleatorio simple con reposición** (m.a.s.).

Consiste en extraer al azar y con la misma probabilidad un elemento entre todos los que conforman la población.

Se analiza y se devuelve a la pobl antes de extraer el sig. elemento de la muestra

La población no cambia en cada extracción y el elem. puede ser elegido de nuevo

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.

Se aplica principalmente cuando población tiene un número elevado de elementos o es infinita, por lo tanto, P(repetir un elemento)≈0

El **muestreo aleatorio simple sin reposición**

Consiste en extraer al azar y con la misma probabilidad un elemento entre todos los que conforman la población.

El elemeto extraído no se devuelve a la población

La población cambia en cada extracción

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.

Se utiliza principalmente cuando el número de elementos de la población no es elevado.

Así aseguro que la inform. de la muestra es distinta y representativa de la poblac.

- **<u>Estadístico</u>:** Cualquier función de los valores muestrales, siempre que no dependa de parámetros o constantes desconocidas. T=T(X_1,X_2,...X_n).

El estadístico es función de variables aleatorias, por lo tanto, es otra variable aleat. cuya distribución depende de la variable aleat. poblacional.

La distribución del estadístico en el muestreo y su posterior observación nos puede ayudar a inferir las características desconocidas de la población.

Sabremos en términos probabilísticos los errores de aproximación entre el valor del estadístico y el valor real del parámetro desconocido.

- **MEDIA MUESTRAL:** $$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

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

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

- **PROPORCIÓN MUESTRAL:** $$\hat{p} = \overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Se usa si la poblac. sigue una distrib. Bernouilli, con E(X)=p, Var(X)=p·(1-p)

3.3 CARACTERÍSTICAS DE UN ESTADÍSTICO EN EL MUESTREO

Sea X una var. aleat. **Poblac**., con **media = μ** y **Desviación típica = σ.**

Sea $\overline{X}$ = media muestral, entonces $E[\overline{X}] = \mu$

$$Var[\overline{X}] = \frac{\sigma^2}{n}$$

Si la población sigue una distr. Bernouilli entonces para las muestras:

$$E[\hat{p}] = p \qquad Var[\hat{p}] = \frac{p(1-p)}{n}$$

Sea **S^2** = Varianza muestral, entonces $E[S^2] = \frac{n-1}{n}\sigma^2$

Como $E[S^2] \neq \sigma^2$ se prefiere trabajar con la cuasivarianza muestral:

Sea **S_1^2** = Cuasivarianza muestral, entonces $E[S_1^2] = \sigma^2$

3.4 DISTRIBUCIÓN DE UN ESTADÍSTICO EN EL MUESTREO

Métodos para calcular las distribuciones de los estadísticos en el muestreo
(hay métodos exactos y métodos aproximados, sólo vemos algunos aproximados):

3.4.1 MÉTODO DE MONTE CARLO o MUESTREO ARTIFICIAL

Supuestamente conocida la distribución de X, es decir F(X), obtenemos valores del estadístico mediante la simulación de muestras aleatorias.

Podemos obtener tantos valores como queramos y, a partir de esos valores, podemos conocer las propiedades de la distribución del estadístico.

Paso.1: Se toma un valor u de U(0,1)

Paso.2: Tomo $X_i = F^{-1}(u)$. Lo repito n veces: Muestra de tamaño n

Paso.3: Tomando muchas muestras puedo intentar ver la distrib. del estadístico

3.4.2 MÉTODO ASINTÓTICO

Teorema Central del Límite: La suma de variables aleat. X_i indep. e idént. distrib tiene una distribución aprox. normal si $n \to \infty$

Ya que $\bar{X} = (\sum_{i=1}^{n} X_i) \cdot \frac{1}{n}$, con X_i var. aleat indep. e idént. distrib $\Rightarrow$

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

suficiente si n>30

3.5 DETERMINACIÓN DEL TAMAÑO MUESTRAL

*¿Cuánto ha de valer **n** (tamaño muestral) para "confiar" en los resultados obtenidos?*

Definimos: e: error máximo permitido

(1-α): nivel de confianza de forma que: **P(|T-θ|<e) = 1 - α**

La prob. de que el estadístico **T** *difiera del parám.* **θ** *en menos de* **e** *unidades es* **1 - α**

Generalmente será: **P($|\bar{X}-\mu|$<e) = 1- α**,

Con los datos anteriores debemos determinar el tamaño muestral necesario

A) No sabemos la distribución de $\bar{X}$.

Aplicaremos la desigualdad de Tchebycheff a $\bar{X}$: $P\left\{\left|\bar{X}-\mu\right| < k\frac{\sigma}{\sqrt{n}}\right\} \geq 1-\frac{1}{k^2}$

Fijado el nivel de confianza podremos despejar k = $\sqrt{\frac{1}{\alpha}}$

y fijado el error máximo permitido despejamos el tamaño muestral necesario:

$$n = \frac{k^2\sigma^2}{e^2}$$

B) Sabemos o aproximamos la distribución de $\bar{X}$ a una normal

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

$$P\left(\left|\frac{\bar{X}-\mu}{\frac{\sigma}{\sqrt{n}}}\right| < Z\alpha_{/2}\right) = 1-\alpha \quad \Rightarrow \quad P\left\{\left|\bar{X}-\mu\right| < z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right\} = 1-\alpha$$

Fijado **(1-α)** $\Rightarrow$ Calculo $z_{\alpha/2}$ de la distribución normal estándar

Fijado **e** $\Rightarrow$ Calculo $n = \frac{z^2_{\alpha/2}\sigma^2}{e^2}$

Otra cuestión relevante es el desconocimiento de σ^2 en la población. Alternativas:

- Estimar σ^2 por los valores obtenidos en estudios similares o anteriores.
- Estimar σ^2 mediante una muestra piloto, por ejemplo usando S^2
- Estimar σ^2 por el máximo valor posible si es que podemos acotarla.

P.ej.: Si la población sigue una distribución Bernoulli $\Rightarrow$ Var(X)=p·(1-p)

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

Normalmente a partir de un cierto tamaño n, apenas compensa aumentarlo

Como hay un coste por observación, a partir de cierto punto,no compensa elevar n, ya que eleva el coste del estudio, pero apenas produce mejoras en la precisión.