TEMA 5. Muestreo y distribuciones en el muestreo

Nuestro objetivo fundamental es saber qué modelo va a seguir la población, y para ello haremos uso de la información que obtengamos de una parte de esa población llamada muestra.

Hacemos así introducción a la inferencia estadística: Que pretende estimar los parámetros de una población (su medida, su varianza, etc.) a partir de la información contenida en una o varias muestras.

En este capítulo 5 estudiaremos una serie de conceptos básicos, y que serán fundamentales para el posterior desarrollo de la inferencia estadística.

5.1 Al finalizar el tema el alumno debe conocer.....

- ✓ Importancia de la inferencia estadística.
- ✓ Conceptos fundamentales de la inferencia estadística como: Población, muestra, parámetro poblacional, estadístico muestral, estimación.
- ✓ Características de la muestra.
- ✓ Función de distribución empírica.
- ✓ Características de la distribución de probabilidad de estadísticos muestrales.
- ✓ Características de la distribución de probabilidad de estadísticos muestrales en poblaciones normales.
- ✓ Distribución de la proporción muestral.

5.2 Importancia de la inferencia estadística.

El análisis exploratorio de datos busca descubrir y resumir la información que contienen los datos utilizando gráficos y resúmenes numéricos. Las conclusiones que obtenemos del análisis de datos se refieren a los datos concretos que examinamos. A menudo, queremos extender estas conclusiones a algún grupo mayor de individuos y debemos tener en cuenta que si nuestros datos no representan de una manera adecuada a este grupo, las conclusiones que obtenemos a partir de los datos no pueden extenderse al grupo mayor.

En los próximos apartados dedicaremos gran parte de nuestra atención a analizar problemas que tienen por objeto averiguar las propiedades de un grupo de individuos, a partir de la información proporcionada por un subconjunto relativamente pequeño de los mismos:

Llamaremos población a un grupo entero de individuos, objetos o medidas que tienen una característica común observable y muestra al subconjunto de individuos, objetos o medidas de la población, cuyas características han sido observadas.

En inferencia estadística se razona desde lo particular, la muestra, hasta lo general, la población, y se trata por tanto de un razonamiento inductivo. Por el contrario, en el cálculo de probabilidades que hemos visto anteriormente se razonaba de un modo deductivo, puesto que se partía del conocimiento total de la población.

Algunos ejemplos de poblaciones pueden ser:

- La renta de todas las familias que viven en la Comunidad Autónoma de Madrid.
- Los dividendos anuales obtenidos por cada uno de los valores negociados en la Bolsa de Madrid.
- El coste anual que les supone a todas las empresas del sector industrial la mano de obra que utilizan en el proceso productivo.
- Los errores que aparecen en la Contabilidad de una empresa a lo largo de un año.

La principal razón por la que se observa una muestra en lugar de la población completa, es el hecho de que es generalmente imposible estudiar a todos los miembros de una población dada. Porque la población, tal como se ha definido, tiene un número infinito de elementos o es tan grande que un análisis exhaustivo de la misma realizando censos exige la movilización de muchos recursos humanos y suele ser muy costosa.

Ante tal situación, existen métodos alternativos cuyo coste económico y tiempo se reducen considerablemente. Estos métodos están constituidos por las muestras cuya finalidad es construir modelos reducidos de la población total y usarlas para estimar

el valor de los parámetros, con resultados extrapolables al universo del que proceden.

Conviene recordar el Instituto Nacional de Estadística organismo autónomo de carácter administrativo adscrito al Ministerio de Economía y Hacienda, para elaborar sus trabajos (sirva como ejemplo las estadísticas sanitarias, la encuesta de población activa (EPA, etc.) recurre a muestras.

Evidentemente deseamos seleccionar estas muestras de modo que sean representativas de la población de donde provienen. Imaginemos que la Consejería de Economía y Hacienda de la Comunidad Autónoma de Castilla-La Mancha quiere tener información sobre los ingresos por hogar en la Comunidad, resultaría poco aconsejable que se estudiaran sólo los hogares de la provincia de Toledo. Es poco probable que este grupo de hogares represente adecuadamente a toda la población. Una forma de evitar este tipo de problemas es utilizar un proceso de selección de la muestra basado en el principio de aleatorización.

5.3 Muestra aleatoria simple.

Una muestra aleatoria simple de tamaño " n " consiste en n objetos (individuos) de una población de N objetos (individuos), escogidos de manera que cualquier conjunto de n objetos de la población tenga la misma oportunidad de convertirse en la muestra realmente seleccionada.
Una muestra aleatoria simple no sólo da a cada individuo la misma oportunidad de ser escogido evitando por tanto el sesgo en la selección (diremos que el diseño de un estudio es sesgado si favorece sistemáticamente ciertos resultados), sino que también da a cada posible muestra la misma oportunidad de ser escogida.

Puede pensarse en el proceso de muestreo aleatorio simple de la forma siguiente:
Supongamos que los N miembros de la población se introducen en un enorme sombrero y se mezclan concienzudamente. Una muestra aleatoria simple se obtiene extrayendo a n de ellos. En la práctica no es necesario hacerlo de este modo, los programas estadísticos pueden escoger una muestra aleatoria simple casi de forma instantánea de una lista de individuos de una población, o también se puede

aleatorizar utilizando una tabla de dígitos aleatorios.

Por el momento nos limitaremos a muestras que hayan sido seleccionadas mediante esquemas de muestreo aleatorio simple. Sin embargo, debemos aclarar que este no es el único procedimiento que existe para elegir individuos de una población, y que, en determinadas circunstancias, pueden resultar preferibles esquemas de muestreo alternativos.

5.4 Parámetros poblacionales y estadísticos muestrales.

Generalmente diremos que los parámetros poblacionales son las características numéricas de la población. En la estadística clásica un parámetro se puede considerar como una constante fija cuyo valor se desconoce. Uno de los problemas más comunes en la estadística inferencial es estudiar una población con una función de distribución $F(x,\theta)$ donde la forma de la función de distribución es conocida pero depende de un parámetro θ desconocido, ya que si fuese conocido tendríamos totalmente especificada la función de distribución.

Un estadístico es una variable aleatoria, que es función de las observaciones muestrales y no contiene ningún valor o parámetro desconocido. Continuando con la población de función de distribución $F(x,\theta)$, y considerando una muestra aleatoria simple $(x_1,\cdots,x_n)$ constituida por n variables aleatorias independientes e idénticamente distribuidas podemos definir como estadísticos:

$$\hat{e}_1 = \frac{x_1 + \cdots + x_n}{n} \qquad \hat{e}_2 = \frac{x_1^2 + \cdots + x_n^2}{n} \qquad \cdots \text{etc.}$$

Por tanto, parámetro y estadístico son conceptos muy diferentes. Un parámetro es una constante que describe la población y cuando se conoce queda determinado el modelo probabilístico y un estadístico es una variable aleatoria cuyo valor depende de las observaciones muestrales.
Veamos el siguiente cuadro:

Dada una población finita de tamaño N, y una muestra aleatoria simple de tamaño n, $(x_1,\cdots,x_n)$, obtenida de la población de partida, tenemos:

	Parámetros poblacionales	Estadísticos muestrales
Media	$\mu = \frac{\sum_{i=1}^{N} x_i}{N}$	$\overline{x} = \frac{\sum_{i=1}^{n} x_i}{n}$
Varianza	$\delta^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$	$S^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{x})^2}{n-1}$
Proporción	$p = \frac{\textit{número de éxitos en } N \textit{ pruebas}}{N}$	$\hat{p}_x = \frac{\textit{número de éxitos en } n \textit{ pruebas}}{n}$

Si la población de partida no es finita utilizaremos la misma notación para designar a estos parámetros poblacionales, pero estos no pueden ser calculados a partir de muestras finitas, sino que tendremos que recurrir al cálculo de valores esperados de variables aleatorias de tipo continuo.

La estadística inferencial o inductiva consiste en utilizar un estadístico para llegar a una conclusión o inferencia sobre el parámetro poblacional correspondiente. Por ejemplo, podemos calcular la media aritmética de una muestra, recurriendo al estadístico $\overline{x}$, y utilizarlo como estimación de la media aritmética de la población μ; el estadístico se utiliza como estimador del parámetro.

5.5 Función de distribución empírica.

La función de distribución empírica tiene las mismas propiedades que la función de distribución de la variable aleatoria, lo que implica que cuando el tamaño de la muestra crece, la gráfica de la función de distribución empírica se aproxima bastante a la de la función de distribución de la población, con lo que puede utilizarse como estimador de la misma.

$$F_n(x) = \frac{N(x)}{n}$$

Siendo $N(x)$ el número de valores observados menores o iguales que x.

5.6 Distribución muestral de estadísticos.

Es importante recordar que en el análisis estadístico tiene mucha importancia la información obtenida de una muestra representativa de la población: un director de una empresa elige una muestra representativa para determinar el grado de satisfacción que presentan los usuarios de su producto, un partido político selecciona una muestra de ciudadanos para analizar si su programa político producirá los resultados deseados, etc en estos casos los resultados obtenidos sólo son estimaciones de lo que ocurre en toda la población. El valor del estadístico es aleatorio porque depende de los elementos elegidos en la muestra seleccionada y, por lo tanto, el estadístico tiene una distribución de probabilidad la cual llamamos Distribución Muestral del Estadístico. Esta distribución dependerá del tamaño de la muestra, luego podemos decir que existe diferencia entre la distribución de la población de la cual se ha tomado la muestra y la distribución de alguna función de esa muestra.

La distribución muestral de un estadístico se puede obtener tomando todas las posibles muestras de la población de un tamaño fijado "n" calculando el valor del estadístico para cada una de las muestras y construyendo la distribución de estos valores. Como para cualquier distribución, las dos medidas fundamentales son la media y la desviación típica, también denominada error típico.
En esta asignatura estudiaremos las distribuciones muestrales del los estadísticos media, varianza y proporción muestral, pues son de bastante utilidad en diferentes aplicaciones estadísticas.

- **Distribución muestral del estadístico media muestral.**

- <u>Cuando se conoce la varianza poblacional</u>:
  Si tenemos una muestra aleatoria de tamaño *n* procedente de una población con distribución normal $N(\mu,\sigma)$ entonces la distribución del estadístico media muestral será:

$$\overline{X} = \frac{\sum x}{n} \to N\left(\mu, \sigma/\sqrt{n}\right)$$, si tipificamos $$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \to N(0,1)$$

Si la distribución no es normal, basta que $n \geq 30$ para que la media muestral siga una distribución normal, por el Teorema Central del Límite.

- Cuando no se conoce la varianza poblacional

  Calculamos la varianza muestral y la usamos como varianza poblacional (cuasivarianza)

$$n \geq 30 \quad Z = \frac{\overline{X} - \mu}{s/\sqrt{n}} \rightarrow N(0,1)$$

$$n < 30 \quad T = \frac{\overline{X} - \mu}{s/\sqrt{n}} \rightarrow t_{n-1}$$

- **Distribución muestral del estadístico varianza muestral.**

A partir de la definición de la cuasivarianza muestral, obtenemos:

- Cuando no se conoce la media poblacional.

$$S^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

$$(n-1)S^2 = \sum (x_i - \bar{x})^2 \qquad \textit{dividimos todo por } \sigma^2$$

$$\frac{(n-1)S^2}{\sigma^2} = \frac{\sum (x_i - \bar{x})^2}{\sigma^2} \Rightarrow \chi^2_{n-1}$$

- Cuando se conoce la media poblacional.

$$\frac{\sum (x_i - \mu)^2}{\sigma^2} \rightarrow \chi^2_n$$

Estos estadísticos son independientes. Recordar que los grados de libertad, son los números de variables que integran, en este caso, el estadístico.

- **Distribución muestral del estadístico proporción muestral**

Sea una muestra aleatoria simple de tamaño n procedente de una población con distribución B(1, p)

$$P_X = \frac{X}{n} \to N\left(p, \sqrt{\frac{pq}{n}}\right)$$

$$E(P_X) = p$$

$$V(P_x) = \frac{pq}{n}$$

Si tipificamos:

$$Z = \frac{P_X - p}{\sqrt{\frac{pq}{n}}} \to N(0,1)$$

5.7 Resumen y preguntas frecuentes

$$N(\mu,\sigma)\begin{cases}\mu\begin{cases}Si\ \sigma^2 \to \overline{X} = \dfrac{\sum x_i}{n} \to N\left(\mu, \dfrac{\sigma}{\sqrt{n}}\right) \Rightarrow Z = \dfrac{\overline{X}-\mu}{\sigma/\sqrt{n}} \to N(0,1) \\ No\ \sigma^2\begin{cases} n \geq 30 \to \overline{X} = \dfrac{\sum x_i}{n} \to N\left(\mu, \dfrac{s}{\sqrt{n}}\right) \Rightarrow Z = \dfrac{\overline{X}-\mu}{s/\sqrt{n}} \to N(0,1) \\ n < 30 \to \overline{X} = \dfrac{\sum x_i}{n} \to N\left(\mu, \dfrac{s}{\sqrt{n}}\right) \Rightarrow T = \dfrac{\overline{X}-\mu}{s/\sqrt{n}} \to t_{n-1} \end{cases}\end{cases} \\ \sigma^2\begin{cases} No\ \mu \to \dfrac{(n-1)S^2}{\sigma^2} = \dfrac{\sum (x_i - \bar{x})^2}{\sigma^2} \to \chi^2_{n-1} \\ Si\ \mu \to \dfrac{\sum (x_i - \mu)^2}{\sigma^2} \to \chi^2_n \end{cases}\end{cases}$$

- Explique cual es el objetivo básico de la inferencia estadística.
- Explique si existe alguna diferencia entre Población y muestra. Ponga un ejemplo.
- ¿Cuándo podemos decir que una muestra es válida para realizar inferencia?
- Concepto de muestra aleatoria simple.
- ¿Qué es un parámetro poblacional? Ponga algún ejemplo.
- ¿Qué es un estadístico? Ponga algún ejemplo.
- ¿Qué utilidad tiene la Función de distribución empírica?
- Explique qué es la distribución de probabilidad de una población y la distribución de probabilidad de un estadístico muestral. ¿se puede afirmar que es lo mismo?

- ¿Cómo se obtiene la distribución muestral de un estadístico?
- ¿Cuál es la media y varianza del estadístico media muestral?¿y de la varianza muestral?
- ¿Qué distribución sigue la media muestral cuando se conoce la varianza poblacional?¿y cuando no se conoce la varianza poblacional?
- ¿Qué distribución sigue la varianza muestral?
- ¿Qué distribución sigue la proporción muestral?