Capítulo 6

Introducción a la Inferencia Estadística

6.1. Introducción

El principal objetivo de la Estadística es inferir o estimar características de una población que no es completamente observable (o no interesa observarla en su totalidad) a través del análisis de una parte de ella a la que llamamos *muestra.* Las razones por las que generalmente se trabaja con muestras son principalmente:

- Económicas.
- Tiempo: si la población es muy grande llevaría tanto tiempo analizarla que incluso la característica de interés podría variar en ese período. Por ejemplo, la tasa de paro.
- Destrucción: la medición de cierta característica podría llevar a la destrucción del individuo. Por ejemplo, al estudiar la supervivencia de ciertos animales a un tratamiento.

Lo que se hace entonces es analizar la muestra y extrapolar conclusiones desde la muestra a la población. Ahora bien, para considerar válidas en la población las conclusiones obtenidas en la muestra, ésta ha de representar bien a la población (*representativa*). Por lo tanto, la selección de la muestra es de suma importancia, y para ello hay diversos métodos (métodos de muestreo). Cuando se intuye que la característica en estudio puede presentar valores homogéneos

en la población, una forma de obtener una muestra representativa es eligiéndola al azar. A este método de selección de la muestra se le llama *muestreo aleatorio simple* y es el más sencillo.

La Inferencia Estadística se puede clasificar en *inferencia paramétrica* e *inferencia no paramétrica.* La inferencia paramétrica tiene lugar cuando se conoce la distribución de la variable de estudio en la población, y el interés recae sobre los parámetros desconocidos de la misma. La inferencia no paramétrica tiene lugar si no se conoce la distribución y sólo se suponen propiedades generales de la misma. Nosotros nos centramos en la inferencia paramétrica, y nuestro objetivo será inferir o estimar parámetros poblacionales a partir de la información que nos proporciona una muestra.

Supongamos que estudiamos una variable X en una población y sabemos que presenta una distribución F_θ, donde θ es el parámetro de la distribución y es desconocido. Los problemas de inferencia que pueden darse son: de *estimación*, en los que se busca un valor (estimación puntual) para θ o un conjunto de valores posibles para el mismo (estimación por intervalos de confianza), y de *contraste,* cuyo objetivo es comprobar si es cierta o falsa cierta hipótesis formulada sobre el parámetro θ. En el Tema 7 se estudia la estimación puntual y por intervalos de confianza, y en Tema 8 estudiaremos problemas de contraste de hipótesis.

Ejemplo: Supongamos que queremos estudiar el tiempo de fallo de una población de cierto tipo de componentes. Intuimos (por estudios anteriores por ejemplo) que el tiempo de fallo X sigue una distribución Exponencial, $X \to Exp(\lambda)$, con λ desconocido, ya que no observamos el tiempo de fallo de todos los componentes de la población. Tendremos que estimar su valor en base a la información que proporciona una muestra. Dado que $E(X) = 1/\lambda$, y parece lógico estimar la media poblacional con la media muestral $\overline{x}$, tenemos que $\hat{\lambda} = 1/\overline{x}$.

6.2. Muestra aleatoria simple. Estadísticos muestrales

Sea X la variable aleatoria de interés en la población, con función de probabilidad o densidad $f(x;\theta)$, donde θ denota el parámetro o parámetros desconocidos. *Una muestra aleatoria simple (m.a.s.)* de tamaño n es un conjunto de variables $X_1, ..., X_n$ tales que:

- $X_1, ..., X_n$ son independientes
- $X_1, ..., X_n$ son idénticamente distribuidas, con la misma distribución que la variable poblacional X.

Nota: una vez observada la variable sobre los n individuos de la muestra, tendremos n valores u observaciones $x_1, ..., x_n$.

Un estadístico es una función de las variables aleatorias de la muestra, en la cual no aparecen parámetros desconocidos. Un *estadístico* es por lo tanto una variable aleatoria, y lo denotamos por $T(X_1, ..., X_n)$. El valor que toma el estadístico una vez observada la muestra es $T(x_1, ..., x_n)$. Al ser los estadísticos variable aleatorias, presentarán distribuciones de probabilidad, a las que llamamos *distribuciones de muestreo*. Si un estadístico lo usamos para estimar un parámetro desconocido de la población (por ejemplo la media μ, varianza σ^2, etc.) lo llamaremos *estimador* de ese parámetro. Al valor que toma una vez observada la muestra se le llama *estimación puntual* del parámetro. Para cada parámetro habrá que encontrar "el mejor estimador", para cometer en la estimación el menor error posible. El error de estimación depende fundamentalmente de la variabilidad poblacional y del tamaño de la muestra.

Ejemplos de estadísticos son los siguientes:

- Media muestral:
$$\bar{X} = \frac{X_1 + ... + X_n}{n}$$
- Varianza muestral:
$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$$

6.3. Distribuciones de muestreo

6.3.1. Media muestral

- Sea $X_1, ..., X_n$ una m.a.s. de una población X con $E(X) = \mu$ y $Var(X) = \sigma^2$. El estadístico *media muestral* hemos visto que se define como
$$\bar{X} = \frac{X_1 + ... + X_n}{n},$$
Se puede comprobar que:
$$E(\bar{X}) = \mu \text{ y } Var(\bar{X}) = \frac{\sigma^2}{n}$$
El Teorema Central del Límite según vimos establece que:
$$\bar{X} = \frac{X_1 + ... + X_n}{n} \underset{(n\to\infty)}{\to} N\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$$

- Sea $X_1, ..., X_n$ una m.a.s. de una población X con distribución $N(\mu, \sigma)$. Entonces,

$$\bar{X} = \frac{X_1 + ... + X_n}{n} \to N\left(\mu, \frac{\sigma}{\sqrt{n}}\right),$$

al ser combinación lineal de variables normales e independientes.

6.3.2. Varianza muestral

- Sea $X_1, ..., X_n$ una m.a.s. de una población X con $E(X) = \mu$ y $Var(X) = \sigma^2$. El estadístico *varianza muestral* se define como

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \mu)^2}{n-1}$$

- Sea $X_1, ..., X_n$ una m.a.s. de una población X con distribución $N(\mu, \sigma)$. Entonces:

$$\frac{(n-1)S^2}{\sigma^2} \to \chi^2_{n-1}$$

y $\bar{X}$ y S^2 son independientes.

6.3.3. Diferencia de medias muestrales

Sea $X_1, ..., X_{n_1}$ una m.a.s de una población X, e $Y_1, ..., Y_{n_2}$ una m.a.s. de una población Y. Suponemos que las poblaciones X e Y son independientes y con distribuciones normales $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$ respectivamente.

Se pueden presentar los siguientes casos:

(a) σ_1^2, σ_2^2 conocidas:

$$\bar{X} - \bar{Y} \to N\left(\mu_1 - \mu_2, \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}\right),$$

o equivalentemente

$$Z = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n2}}} \to N(0,1)$$

(b) $\sigma_1^2 = \sigma_2^2 = \sigma^2$ desconocidas:

$$T = \frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n2}}} \to t_{n_1+n_2-2},$$

siendo

$$S_p = \sqrt{\frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}}$$

y S_1^2 y S_2^2 las varianzas muestrales de X e Y respectivamente.

6.3.4. Cociente de varianzas muestrales

Sea $X_1, ..., X_{n_1}$ una m.a.s de una población X, e $Y_1, ..., Y_{n_2}$ una m.a.s. de una población Y. Suponemos que las poblaciones X e Y son independientes y con distribuciones normales $N(\mu_1, \sigma_1^2)$ y $N(\mu_2, \sigma_2^2)$ respectivamente.

Entonces,

$$F = \frac{\frac{S_1^2}{\sigma_1^2}}{\frac{S_2^2}{\sigma_2^2}} \rightarrow F_{n_1-1, n_2-1}$$

Estudiamos además la distribución de una proporción muestral y de la diferencia de dos proporciones muestrales.

6.3.5. Proporción muestral

Sea $X_1, ..., X_n$ una m.a.s. de una población X. Sea p la proporción de individuos en la población que presentan una determinada característica, y $\hat{p}$ la proporción muestral. Entonces,

$$\hat{p} \rightarrow N(p, \sqrt{\frac{p(1-p)}{n}})$$

Nota: El número de individuos que presentan la característica en la muestra sigue una distribución $B(n,p)$, que con n suficientemente grande se puede aproximar a una $N(np, \sqrt{np(1-p)})$. Por lo tanto, la proporción muestral sigue también una distribución Normal con los parámetros arriba indicados.

6.3.6. Diferencia de proporciones muestrales

Sea $X_1, ..., X_{n_1}$ una m.a.s de una población X, e $Y_1, ..., Y_{n_2}$ una m.a.s. de una población Y. Suponemos que las poblaciones X e Y son independientes. Denotamos por p_1 y p_2 las proporciones poblacionales y por $\hat{p_1}$ y $\hat{p_2}$ las correspondientes proporciones muestrales.

Entonces:

$$\hat{p_1} - \hat{p_2} \to N\left(p_1 - p_2, \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_1}}\right)$$

Por lo tanto:

$$Z = \frac{\hat{p_1} - \hat{p_2} - (p_1 - p_2)}{\sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_1}}} \to N(0,1)$$

6.4. Ejercicios

1. Una cementera elabora un tipo de cemento que tiene un contenido medio de aditivo B542 de 100mg/kg con una desviación típica de 10 mg/kg. Suponemos que la distribución es Normal. Calcula la probabilidad de que al tomar una muestra de 20kg de la producción diaria el contenido de aditivo sea, en media, menor de 95 mg/kg.

2. En una industria se fabrican unos cables cuya resistencia sigue una distribución Normal de media 200 ohmios y desviación típica de 15 ohmios. Se toma una muestra de 15 cables.

   a) ¿Qué probabilidad hay de que la media muestral sea menor que 195 ohmios?.

   b) ¿Qué tamaño de la muestra se debe tomar para garantizar una duración media de la muestra superior a 195 ohmios con una probabilidad mayor o igual que el 95 %.

3. Se toma una muestra de 25 observaciones de una población Normal que tiene una varianza $\sigma^2 = 10$. ¿Cuál es la probabilidad de que la varianza muestral sea mayor que 16?.

4. La vida eficaz de un componente sigue una distribución Normal de media 5000 horas y desviación típica de 40 horas. Nos proponen un nuevo componente y nos garantizan una vida media de 5050 horas y desviación típica de 30 horas. Decidimos hacer una prueba y tomamos 25 componentes de cada grupo. Decidimos cambiar de proveedor si la diferencia de duración es, en media, al menos de 25 horas. Si el nuevo proveedor está en lo cierto, ¿qué probabilidad tiene de que le compremos sus componentes?.

5. Si S_1^2 y S_2^2 son las varianzas muestrales de m.a.s. independientes de tamaños $n_1 = 10$ y $n_2 = 20$ tomadas de poblaciones normales que tienen las mismas varianzas, calcular la probabilidad de que el cociente de varianzas muestrales S_1^2 / S_2^2 sea menor que 2.42.

6. El resultado de una encuesta de opiniones fue que el 59 % de la población española piensa que la situación económica es buena o muy buena. Supongamos, extrapolando los resultados del sondeo a la población entera que la proporción de todos los españoles con esta opinión es efectivamente 0.59.

   a) Muchos de los sondeos tienen un *margen de error* de orden ± 3 puntos. ¿Cuál es la probabilidad de que una muestra aleatoria de 300 españoles presente una proporción muestral que no se aleje en más de 0.03 de la proporción auténtica $p = 0{,}59$?.

   b) Contesta a la pregunta anterior para una muestra de 600 individuos y otra de 1200. ¿Cuál es el efecto de aumentar el tamaño muestral?.

7. En condiciones normales, una máquina produce piezas con una tasa de defectuosas del 1 %. Para comprobar que la máquina sigue bien ajustada, se escogen al azar cada día 100 piezas en la producción y se les somete a un test. ¿Cuál es la probabilidad de que, si la máquina está bien ajustada, haya en una de esas muestras más del 2 % de piezas defectuosas?.