CAPÍTULO VII: Estimación de parámetros

Existen dos formas de hacer inferencia acerca de un parámetro de una población: estimar el valor de un parámetro desconocido o tomar una decisión acerca de un valor hipotético del parámetro. Por ejemplo, se puede desear estimar el número medio μ de trabajos presentados cada hora a un centro de procesamiento de datos o se puede querer decidir si la media μ excede o no cierto valor μ_0. Los métodos de estimación serán tratados en este capítulo, mientras que el método para tomar una decisión acerca de uno ó más parámetros de una población, denominado **prueba de hipótesis**, se desarrollará en el próximo capítulo.

Una variable aleatoria o población X está caracterizada por su función de densidad de probabilidad. Esta función de densidad de probabilidad, siempre que esté determinada con todo detalle, permite calcular cualquier parámetro poblacional θ, es decir, aquella constante que informa de manera sintética de una propiedad relevante o característica de la población o variable aleatoria, tal como el valor medio μ, o la varianza σ^2, parámetros clásicos de centralización y de dispersión, respectivamente, o la proporción binomial *p* de éxitos en una población.

Si el valor de un parámetro θ es desconocido, los **estimadores** que se puedan construir permitirán la **estimación** de tal parámetro. A tal efecto, se entiende por **estimador** cualquier variable aleatoria que se defina a partir de la sucesión de variables aleatorias que integran una muestra extraída al azar de una población. Se debe valorar en un estimador su capacidad de extraer "al máximo" la información contenida en la muestra, ya que redundará en la calidad y precisión de las estimaciones.

Un estimador $\Theta(X_1,X_2,...,X_n)$ (o simplemente Θ) de un parámetro θ es una regla o fórmula que muestra como utilizar las observaciones de una muestra para calcular un solo número (un punto) ó un intervalo que nos sirva como estimación del valor del parámetro θ.

Definición:
Un **estimador puntual** es una regla o fórmula que expresa cómo calcular una estimación numérica con base en las determinaciones de contenidas en una muestra. El número que resulta del cálculo es una **estimación puntual**.

Al considerar un estimador Θ de un parámetro poblacional θ, la realización de una muestra aleatoria de tamaño *n*, X_1, X_2 ,..., X_n, suministra *n* datos, valores u observaciones, x_1, x_2 ,..., x_n, que determinan una **estimación puntual** del parámetro desconocido. Es decir:

$\hat{\theta}$, estimación puntual de θ, corresponde al valor del estimador Θ en x_1, x_2 ,..., x_n

Si se pretende, por ejemplo, estimar puntualmente el valor medio μ con el estimador media muestral, se extrae una muestra aleatoria de la población y se observa el valor de la variable X en los n individuos de la muestra. En tal caso, los *n* datos obtenidos x_1, x_2 ,..., x_n, permiten calcular lo deseado. Es decir:

$\bar{x}$, estimación puntual de μ, que corresponde al valor del estimador $\bar{X}$ en x_1, x_2 ,..., x_n: $\bar{x} = \dfrac{x_1 + x_2 + \cdots + x_n}{n}$.

De forma similar, la varianza muestral, S^2, es una estimador puntual de σ^2, es decir $\hat{\sigma}^2 = S^2$.

Definición:
Un **estimador de intervalo** es una regla, casi siempre expresada como una fórmula, que permite calcular dos puntos a partir de los datos de una muestra. El objetivo es formar un intervalo que contenga a con un grado de confianza elevado.

Estimación por Intervalo

La **estimación por intervalos** de un parámetro θ consiste en la determinación de un intervalo, que contendrá el parámetro con un nivel de confianza $1-\alpha$, $0< \alpha <1$, fijado por el experimentador. Par ello se requerirá lo siguiente:

- Una muestra aleatoria $X_1, X_2, ..., X_n$ de tamaño *n* extraída de la población X.
- Un estimador Θ del parámetro poblacional θ, con distribución o ley de probabilidad conocida.
- El nivel de confianza $1-\alpha$, establecido a priori por el experimentador (los valores más utilizados son 0.95, 0.90 y 0.99).

A continuación se expondrá, de forma más general, el procedimiento de obtención de un intervalo de confianza para un parámetro θ de una población univariante X. Si **E(θ)** es el **estadístico** adecuado para θ,con distribución de probabilidad conocida**,** y $1-\alpha$ es el nivel de confianza adoptado, se podrá que:

$$P(e_{\alpha/2} < E(\theta) < e_{1-\alpha/2}) = 1-\alpha$$

donde $e_{\alpha/2}$ y $e_{1-\alpha/2}$ son los valores críticos que verifican respectivamente: $P(E \le e_{\alpha/2}) = \alpha/2$ y $P(E \le e_{1-\alpha/2}) = 1-\alpha/2$.

Esta expresión probabilística es equivalente a:

$$P\left(\Theta\left(X_1, X_2, \ldots, X_n, e_{\alpha/2}\right) < \theta < \Theta\left(X_1, X_2, \ldots, X_n, e_{1-\alpha/2}\right)\right) = 1-\alpha$$

donde $\Theta\left(X_1, X_2, \ldots, X_n, e_{\alpha/2}\right)$ y $\Theta\left(X_1, X_2, \ldots, X_n, e_{1-\alpha/2}\right)$ son términos aleatorios que dependen de la muestra, del estadístico **E(θ)** y del nivel de confianza.

Por ello, con una probabilidad $1-\alpha$ el intervalo aleatorio

$$\left(\Theta\left(X_1, X_2, \ldots, X_n, e_{\alpha/2}\right), \Theta\left(X_1, X_2, \ldots, X_n, e_{1-\alpha/2}\right)\right)$$

contendrá el parámetro θ.

Si se extrae una muestra de tamaño *n* y con los datos u observaciones, $x_1, x_2, ..., x_n$, se calculan los extremos del intervalo, se dispondrá del **intervalo de confianza** $\left(\hat{\theta}_i, \hat{\theta}_s\right)$ **para el parámetro** θ donde los extremos se obtienen evaluando los datos u observaciones de la muestra, es decir:

$$\hat{\theta}_i = \Theta\left(x_1, x_2, \ldots, x_n, e_{\alpha/2}\right) \text{ y } \hat{\theta}_s = \Theta\left(x_1, x_2, \ldots, x_n, e_{1-\alpha/2}\right)$$

que contendrá dicho parámetro con un nivel de confianza 1-α o expresado en porcentaje de (1-α) 100 %..

Intervalo de confianza para la media de una población normal

Sea X tiene una distribución N(μ,σ^2), donde μ es el parámetro desconocido. Sea X_1, X_2, ..., X_n una muestra aleatoria de la v.a. X y sea $\overline{X}$ el estimador puntual media muestral.

Caso I: varianza σ^2 conocida

Dado que X tiene una distribución N(μ,σ^2) con σ^2 conocida entonces la v.a $\overline{X} \sim N(\mu,\sigma^2/n)$ de donde resulta que $Z = (\overline{X} - \mu) / \sqrt{\sigma^2/n} \sim N(0, 1)$. Note que aunque Z depende de μ, su distribución de probabilidad no.

Dado 1-α, el nivel de confianza elegido para construir el intervalo, es posible encontrar z_0 tal que:

$$P(-z_0 < Z < z_0) = 1 - \alpha$$

Si se reemplaza la v.a. Z por su expresión, está igualdad resulta: $P\left(-z_0 < \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} < z_0\right) = 1 - \alpha$. Si se despeja el parámetro μ en la expresión dentro del paréntesis se obtiene:

$$P\left(\overline{X} - z_0 * \sigma/\sqrt{n} < \mu < \overline{X} + z_0 * \sigma/\sqrt{n}\right) = 1 - \alpha.$$

Obsérvese que:

$$\Theta\left(X_1, X_2, \ldots, X_n, e_{\alpha/2}\right) = \overline{X} - z_0 * \sigma/\sqrt{n} \quad \text{donde} \quad -z_0 = z_{\alpha/2}$$

y

$$\Theta\left(X_1, X_2, \ldots, X_n, e_{1-\alpha/2}\right) = \overline{X} + z_0 * \sigma/\sqrt{n} \quad \text{donde} \quad z_0 = z_{1-\alpha/2}.$$

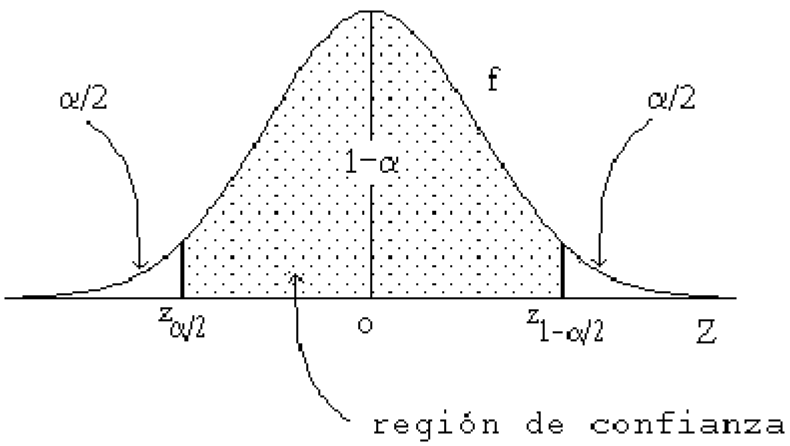
Por lo tanto, con una probabilidad 1-α el intervalo aleatorio

$$\left(\overline{X} - z_0 * \sigma/\sqrt{n}, \overline{X} + z_0 * \sigma/\sqrt{n}\right)$$

contendrá el parámetro μ.

Si se extrae una muestra de tamaño *n* y con los datos u observaciones, $x_1, x_2, \ldots, x_n$, se calculan los extremos del intervalo, se dispondrá del **intervalo de confianza** $(\bar{x} - z_0 * \sigma/\sqrt{n}, \bar{x} + z_0 * \sigma/\sqrt{n})$ **para el parámetro** μ para un nivel de confianza de $(1-\alpha)$ 100 %.

La proposición probabilística $P\left(\bar{X} - z_0 * \sigma/\sqrt{n} < \mu < \bar{X} + z_0 * \sigma/\sqrt{n}\right) = 1 - \alpha$, debe interpretarse *muy cuidadosamente.* Es incorrecto decir que la probabilidad de que el parámetro μ se encuentre contenido en el intervalo $(\bar{x} - z_0 * \sigma/\sqrt{n}, \bar{x} + z_0 * \sigma/\sqrt{n})$ es igual a $1-\alpha$. Esto es, no puede asociarse ningún valor de probabilidad a la proposición $\bar{x} - z_0 * \sigma/\sqrt{n} < \mu < \bar{x} + z_0 * \sigma/\sqrt{n}$ debido a que ésta contiene sólo constantes. Si se permite asignar a dicha proposición un grado de confianza igual a la probabilidad del intervalo aleatorio $\left(\bar{X} - z_0 * \sigma/\sqrt{n}, \bar{X} + z_0 * \sigma/\sqrt{n}\right)$.



Caso II: varianza σ^2 desconocida

Dado que X tiene una distribución $N(\mu,\sigma^2)$ con σ^2 desconocida, no es posible utilizar la expresión del intervalo deducido anteriormente. Se requiere de un intervalo donde no figure σ^2. El valor de σ^2 puede estimarse mediante el estimador puntual varianza muestral S^2.

Como X está normalmente distribuida, el estadístico $T = \dfrac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}}$ tiene una distribución t de Student con n-1 grados de libertad.

Además, dado $1-\alpha$, el nivel de confianza elegido para construir el intervalo, es posible encontrar t_0 tal que:

$$P(-t_0 < T < t_0) = 1 - \alpha$$

Si se reemplaza la v.a. T por su expresión, está igualdad resulta: $P\left(-t_0 < \dfrac{\overline{X}-\mu}{S/\sqrt{n}} < t_0\right) = 1-\alpha$. Si se despeja el parámetro μ en la expresión dentro del paréntesis se obtiene:

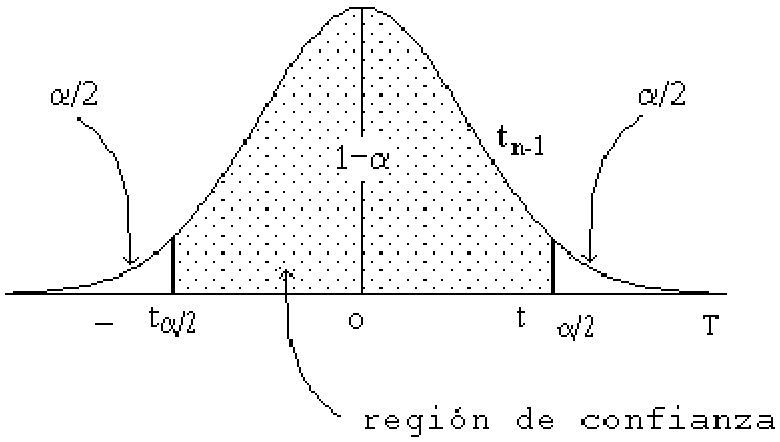
$$P\left(\overline{X} - t_0 * S/\sqrt{n} < \mu < \overline{X} + t_0 * S/\sqrt{n}\right) = 1-\alpha.$$

Por lo tanto, con una probabilidad 1-α el intervalo aleatorio

$$\left(\overline{X} - t_0 * S/\sqrt{n}, \overline{X} + t_0 * S/\sqrt{n}\right)$$

contendrá el parámetro μ.

Si se extrae una muestra de tamaño *n* y con los datos u observaciones, $x_1, x_2, ..., x_n$, se calculan los extremos del intervalo, se dispondrá del **intervalo de confianza** $(\overline{x} - t_0 * s/\sqrt{n}, \overline{x} + t_0 * s/\sqrt{n})$ **para el parámetro** μ para un nivel de confianza de (1-α) 100 %.



Observación importante:

Los intervalos presentados anteriormente se construyeron sobre la base de una v.a. X ~ N(μ,σ²) con varianza σ² conocida o desconocida.

- Si la v.a. X **no está normalmente** distribuida pero σ^2 conocida y el tamaño de la muestra es $n \geq 30$, el teorema central del límite garantiza de todos modos la distribución normal estándar del estadístico $Z = (\overline{X}-\mu) / \sqrt{\sigma^2/n}$ y por lo tanto el intervalo correspondiente a esta situación sería el presentado en el caso I.
- Si la v.a. X **no está normalmente** distribuida pero σ^2 desconocida y el tamaño de la muestra es $n \geq 30$, la varianza muestral s^2 se acercará al verdadero valor de σ^2 y entonces sigue siendo aplicable el teorema central de límite proporcionando el siguiente intervalo:

$(\bar{x} - z_0 * s/\sqrt{n}, \bar{x} + z_0 * s/\sqrt{n})$. Este intervalo se conoce como **intervalo de confianza de muestra grande**.

Intervalo de confianza para la varianza de una población normal

Sea X tiene una distribución $N(\mu,\sigma^2)$, donde μ es el parámetro desconocido. Sea $X_1, X_2, ..., X_n$ una muestra aleatoria de la v.a. X y sea S^2 el estimador puntual varianza muestral.

Se sabe que si la v.a. $X \sim N(\mu,\sigma^2)$, entonces la v.a. $\frac{(n-1)S^2}{\sigma^2}$ es una variable χ^2 con n - 1 grados de libertad.

Dado $1-\alpha$, el nivel de confianza elegido para construir el intervalo, es posible encontrar dos valores de la v.a. χ^2_{n-1} tales que:

$$P(\chi^2_{\alpha/2,n-1} < \chi^2_{n-1} \leq \chi^2_{1-\alpha/2,n-1}) = 1-\alpha$$

donde $\chi^2_{\alpha/2,n-1}$ y $\chi^2_{1-\alpha/2,n-1}$ son valores de la distribución χ^2_{n-1}, con áreas de $\alpha/2$ y $1-\alpha/2$, respectivamente a la izquierda.

Al sustituir la variable χ^2_{n-1} por la expresión $\frac{(n-1)S^2}{\sigma^2}$ resulta:

$$P(\chi^2_{\alpha/2,n-1} < \frac{(n-1)S^2}{\sigma^2} \leq \chi^2_{1-\alpha/2,n-1}) = 1-\alpha$$

Al dividir cada término de la desigualdad por $(n-1)\ S^2$, y a continuación invertir cada término (cambiando por lo tanto el sentido de las desigualdades) se obtiene:

$$P(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}} < \sigma^2 \leq \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}) = 1-\alpha\ .$$

A partir de esta expresión resulta el intervalo aleatorio para estimar σ^2:

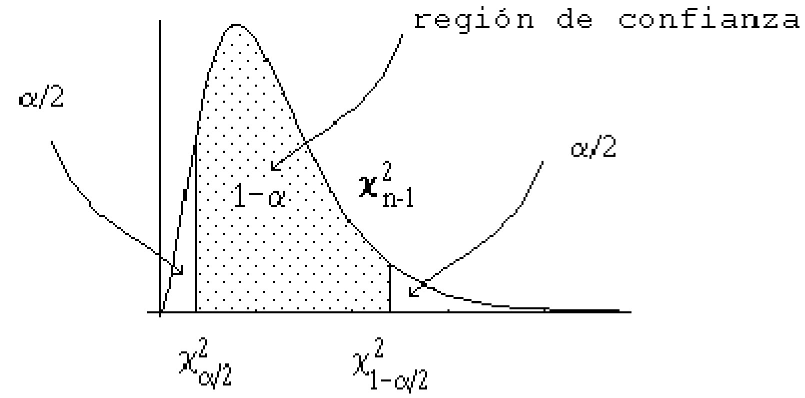
$$(\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}})$$

Si selecciona una muestra aleatoria particular de tamaño n y se calcula la varianza muestral s^2, el intervalo de confianza del $(1-\alpha)$ 100 % para σ^2 está dado por:

$$(\frac{(n-1)s^2}{\chi^2_{1-\alpha/2,n-1}}, \frac{(n-1)s^2}{\chi^2_{\alpha/2,n-1}})$$

Obsérvese que se obtiene un intervalo de confianza del $(1-\alpha)$ 100 % para σ, aplicando la raíz cuadrada a cada punto extremo del intervalo de confianza construido para σ^2:

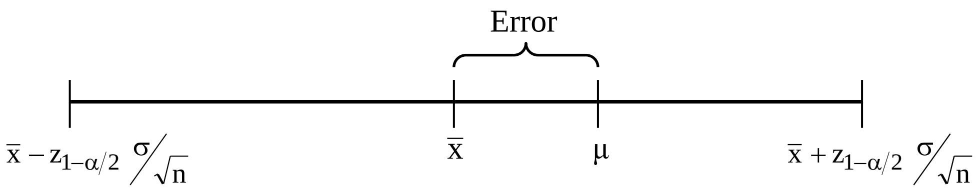
$$\left(\sqrt{\frac{(n-1)S^2}{\chi^2_{1-\alpha/2,n-1}}}, \sqrt{\frac{(n-1)S^2}{\chi^2_{\alpha/2,n-1}}} \right)$$



Tamaño de Muestra

Tamaño de muestra para estimar la media de una población normal

El intervalo de confianza (1-α)*100% proporciona una precisión de la exactitud de la estimación puntual. Si μ es realmente el valor central del intervalo, entonces $\bar{x}$ estima a μ sin error. La mayor parte de las veces, sin embargo, $\bar{x}$ no será exactamente igual a μ y la estimación puntual es errónea. El tamaño de este error será el valor absoluto de la diferencia entre μ y $\bar{x}$, $|\bar{x}-\mu|$, y se puede tener una confianza del (1-α)*100% de que **esta diferencia no excederá una cantidad específica e**.



Con frecuencia se desea saber que tan grande debe ser una muestra para asegurar que el error en la estimación de μ sea menor que una cantidad específica **e**.

Para estimar el tamaño de muestra necesario, de manera tal que con una probabilidad 1-α, la media muestral $\bar{X}$ se aleje de la media poblacional μ en a lo sumo **e** unidades se recurre a la siguiente expresión

$$P\left(-z_0 < \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} < z_0 \right) = 1-\alpha$$

Multiplicando los miembros de la desigualdad por $\sigma/\sqrt{n}$ resulta

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

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

donde **e** = $z_{1-\alpha/2} * \sigma/\sqrt{n}$. A partir de la expresión de **e** es posible obtener el tamaño de muestra *n*.

Entonces, si $\bar{x}$ se utiliza como una estimación puntual de μ, se puede tener una confianza del (1-α)*100% de que el error no excederá una cantidad específica e cuando el tamaño de la muestra es:

$$n = \left(\frac{z_{1-\alpha/2}\ \sigma}{e}\right)^2$$

$|\bar{x} - \mu|$ se conoce con el nombre de **aproximación o error de estimación**, con que se construyó el intervalo de confianza para μ y es tal que $|\bar{x} - \mu| \le$ **e**.

Ejemplo:
Como parte de una encuesta del Departamento de Energía de Estados Unidos (DOE), se seleccionarán al azar familias estadounidenses y se las interrogará acerca de la cantidad de dinero que gastaron el año anterior en combustible o gas para calefacción. Al DOE le interesa especialmente la cantidad media μ gastada el año anterior en combustible para calefacción Si el DOE quiere que la estimación de μ sea correcta en no más 10 dólares con un coeficiente de confianza de 0.95, ¿cuántas familias deberá incluir en la muestra? La cantidad de dinero gastada anualmente en combustible o gas para calefacción es una v.a. distribuida normalmente con desvío estándar σ = 130 dólares.

Según el enunciado 1-α = 0.95, $\alpha = 0.05$, $1 - \alpha/2 = 1 - 0.05/2 = 0.975$, σ = 130 dólares y **e** = 10 dólares.
Por lo tanto,

$n = \left(\frac{z_{1-\alpha/2}\ \sigma}{e}\right)^2 = \left(\frac{z_{0.975}\ 130}{10}\right)^2 = \left(\frac{1.96 * 130}{10}\right)^2 = (25.48)^2 = 649.23$, aproximadamente igual a 650 familias.

Se requiere un tamaño de muestra de por lo menos 650 familias para obtener una estimación $\bar{x}$ que difiera de μ en una cantidad que no exceda 10 dólares, con una confianza del 95%.

__________________________________Fin Capítulo VI ____________________________________