Estimación por intervalos

Introducción. Precisión de las estimaciones

Como vimos en el tema correspondiente, la estimación puntual aborda el problema de conocer el valor de un parámetro desconocido asignándole un único valor, a partir de los datos de una muestra. Dentro de los métodos de construcción de estimadores, el método de máxima verosimilitud resultaba especialmente atractivo tanto por su concepto, asignar al parámetro el valor que hace más plausible la muestra que hemos obtenido, como por sus propiedades, consistencia y ser asintóticamente Normal y eficiente. Sin embargo las propiedades de los estimadores hacen referencia a su comportamiento probabilístico, es decir a lo que ocurre cuando se extrae un gran número de muestras, y no resuelven el problema de determinar la precisión de una estimación concreta.

Clarifiquemos este problema de la precisión de una estimación con un ejemplo. Supongamos que el C.I. de los alumnos de 4º de E.G.B. sigue una distribución Normal y queremos conocer su media a partir de los datos de una muestra de 10 alumnos que ha arrojado los siguientes valores:

88, 95, 111, 82, 119, 105, 91, 100, 102, 98.

Como el estimador de máxima verosimilitud para la media de una población Normal es la media muestral, calculamos la media de es muestra, lo cual nos proporciona el valor 99,1 como estimación para el cociente intelectual medio de los alumnos de 4º de E.G.B.. Será muy difícil que el valor 99.1 coincida con el verdadero valor de la media de la población, es más es prácticamente imposible que coincida con el valor exacto. Es decir, cuando realizamos una estimación estamos casi seguros de que cometemos un error, pero no es eso lo peor, lo más grave es que desconocemos la magnitud de nuestro error, no sabemos si nos equivocamos en una unidad en cinco o en diez.

Se comprende que en estas condiciones, aunque la estimación proporciona información acerca del valor del parámetro resuelve poco nuestra incertidumbre y difícilmente nos permitirá tomar una decisión transcendente acerca de la población. Para paliar estas dificultades de la estimación puntual, se recurre a la estimación por intervalos, la cual, a partir de los datos de la muestra, proporciona un intervalo de valores dentro del cual se encuentra con una determinada "probabilidad" que llamaremos nivel de confianza, el verdadero valor del parámetro. De esta forma considerando el nivel de confianza y la amplitud del intervalo podremos evaluar la magnitud máxima de nuestro error y el riesgo que conllevan nuestras decisiones basadas en la información de la muestra.

Por ejemplo, si por los procedimientos que veremos posteriormente, calculásemos un intervalo de confianza para la media del Cociente Intelectual de los alumnos de 4º de E.G.B., a un nivel de confianza del 95%, a partir de los datos de la muestra anterior obtendríamos que dicha media se encuentra comprendida entre 91,26 y 106,94, lo que nos indica que a ese nivel de confianza, nuestra estimación inicial de 99,1 puede desviarse, por exceso o por defecto, hasta 7,84 puntos de la verdadera media

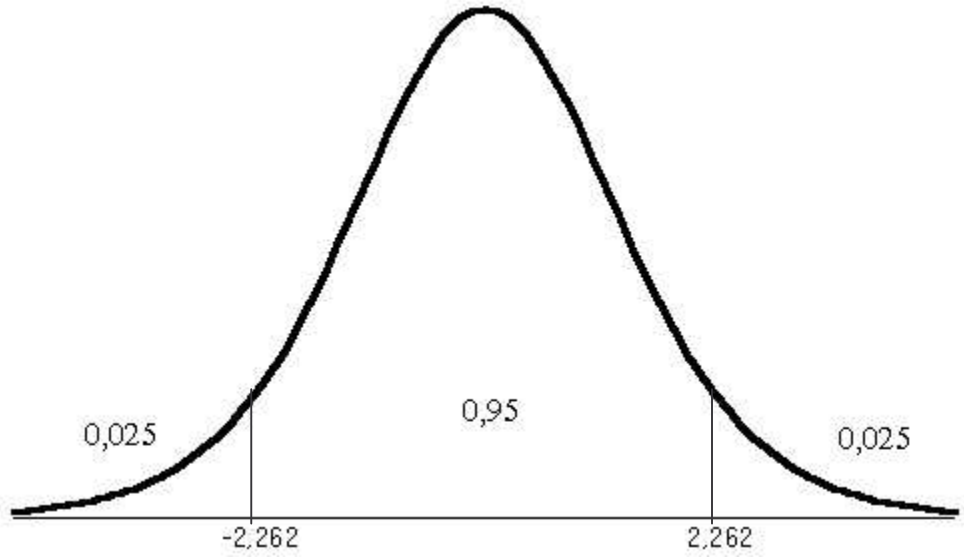
de la población. Error que puede ser excesivo o no, dependiendo del grado de precisión que se requiera para cumplir nuestros objetivos.

Ahora bien, si a partir de una muestra de 500 individuos hubiésemos obtenido la misma media de 99,1 y la misma dispersión, nuestra estimación puntual no variaría, pero si la precisión de la misma. En efecto calculando a partir de esa muestra de 500 alumnos el intervalo de confianza para la media de la población, con igual nivel de confianza, se obtiene que la media de la población debe de estar comprendida entre 98,18 y 100,02 lo que nos muestra que con un nivel de confianza del 95% la media de la población diferirá de la estimación realizada, a lo sumo, en 0,92 unidades. Vemos así que los intervalos de confianza proporcionan un rango de valores plausibles para el verdadero valor del parámetro que tiene en consideración la precisión de las estimaciones obtenidas con la información que contiene la muestra.

Veamos ahora como hemos podido determinar, a partir de los datos de la muestra, el intervalo de confianza para la media de la población y que significado tienen sus elementos. Como el C.I. seguía en la población una distribución Normal, en virtud de los resultados del tema anterior sabemos que el cociente de Student seguirá una distribución t de Student con 9 grados de libertad:

$$\frac{\bar{x}-\mu}{\frac{S}{\sqrt{9}}} \quad \rightarrow \quad t_9$$

entonces fijado arbitrariamente el nivel de confianza del 95%, buscamos en las tablas de la t de Student dos números, tales que la probabilidad de que el estadístico tome un valor comprendido entre ambos sea 0,95.



Obtenemos así los valores -2,262 y 2,262 que verifican:

$$\Pr ob\left(-2{,}262 < \frac{\bar{x}-\mu}{\frac{S}{3}} \leq 2{,}262 \right) \quad = \quad 0{,}95$$

Si ahora multiplicamos todos los miembros de la desigualdad por un número positivo, como es S/3, el sentido de la desigualdad no varía y la probabilidad de que se verifique es la misma, por tanto podemos escribir:

$$\Pr ob\left(-2{,}262 \cdot \frac{S}{3} < \bar{x} - \mu \leq 2{,}262 \cdot \frac{S}{3} \right) \quad = \quad 0{,}95$$

restando a todos los miembros de la desigualdad un mismo número esta no se modificará y la probabilidad de que se cumpla tampoco:

$$\Pr ob\left(-\bar{x} - 2{,}262 \cdot \frac{S}{3} < -\mu \leq -\bar{x} + 2{,}262 \cdot \frac{S}{3} \right) \quad = \quad 0{,}95$$

si multiplicamos por -1, el sentido de las desigualdades se invertirá aunque permanecerá el valor de la probabilidad, luego:

$$\Pr ob\left(\bar{x} + 2{,}262 \cdot \frac{S}{3} \geq \mu > \bar{x} - 2{,}262 \cdot \frac{S}{3} \right) \quad = \quad 0{,}95$$

reordenando, en la forma habitual, la desigualdad queda:

$$\Pr ob\left(\bar{x} - 2{,}262 \cdot \frac{S}{3} < \mu \leq \bar{x} + 2{,}262 \cdot \frac{S}{3} \right) \quad = \quad 0{,}95$$

Lo que nos indica que la probabilidad de que la media de la población esté comprendida entre ambos extremos es 0,95. Debe de notarse que el valor de μ es fijo y lo que depende de los valores muestrales y por consiguiente es aleatorio son los extremos del intervalo, en este sentido el nivel de confianza del 95% nos indica que si extraemos un gran número de muestras de tamaño 10 de la población, aproximadamente, en noventa y cinco de cada cien de ellas el intervalo:

$$\left(\bar{x} - 2{,}262 \frac{S}{3} \;,, \; \bar{x} + 2{,}262 \frac{S}{3} \right)$$

cubrirá el verdadero valor del parámetro y sólo en cinco de cada cien, más o menos, estará fuera de él. Finalmente, si sustituimos en la expresión anterior la media y desviación típica muestral por los valores obtenidos en nuestra muestra, obtendremos la estimación por intervalo indicada antes:

$$\left(99{,}1 - 2{,}262 \cdot \frac{10{,}4}{3} \;,, \; 99{,}1 + 2{,}262 \cdot \frac{10{,}4}{3} \right) \quad = \quad \left(91{,}26 \;,, \; 106{,}94 \right)$$

Es muy importante señalar que una vez fijados los valores de la muestra, los extremos aleatorios del intervalo han pasado a ser unos números fijos y determinados, por consiguiente, no cabe ahora hablar de la probabilidad de que ese intervalo contenga el verdadero valor del parámetro, lo contendrá o no. Por eso se habla de nivel de

confianza del intervalo, como una medida de la fiabilidad del proceso seguido que nos garantiza que aproximadamente, en el 95% de las ocasiones los valores que encontremos para los extremos del intervalo incluirán entre ellos el valor del parámetro.

Intervalos de Confianza. Definiciones

Una vez planteado como la estimación por intervalos surge para abordar el problema de acotar la precisión de las estimaciones vamos a dar las definiciones de los elementos que han ido apareciendo en la anterior introducción.

Estimación por intervalos

Es el procedimiento de la Inferencia Estadística que asigna al parámetro o parámetros de una población, un intervalo de valores donde se encontrará este con una "probabilidad" previamente fijada y que habitualmente será alta. Esa "probabilidad" a la que hemos hecho referencia, se denomina *nivel de confianza* y los valores mas empleados para ella son: 0,9, 0,95 y 0,99.

Intervalo de Confianza

Es un intervalo de extremos aleatorios que con un nivel de confianza determinado, contiene el verdadero valor del parámetro.

Nivel de confianza

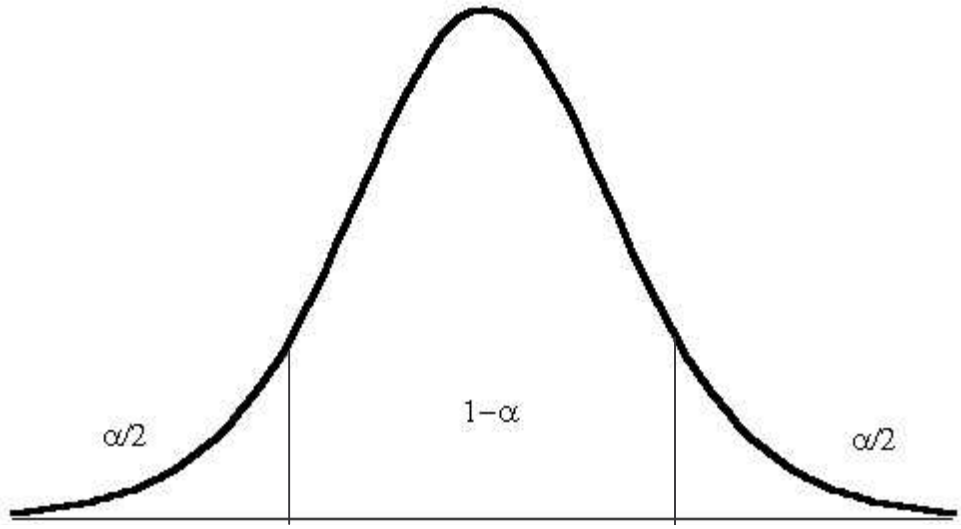
Es la probabilidad "a priori" de que el intervalo contenga el verdadero valor del parámetro. Decimos que es una probabilidad "a priori" porque es la probabilidad que tenemos antes de extraer la muestra de capturar dentro del intervalo al valor del parámetro. Una vez obtenidos los valores de una muestra concreta y sustituidos en la expresión tendremos una estimación que contendrá o no al parámetro.

Intervalo de Confianza para la media

Una vez presentados y definidos los elementos básicos de la estimación por intervalos de confianza, vamos a deducir una expresión general para el intervalo de confianza de la media de una población Norma. Sea X una variable aleatoria que sigue una distribución Normal de media μ y desviación típica σ, de esta población extraemos una muestra aleatoria de tamaño n. En virtud del teorema de Fisher el cociente de Student seguirá una distribución t de Student con n-1 grados de libertad:

$$\frac{\bar{x}-\mu}{\frac{S}{\sqrt{n-1}}} \rightarrow t_{n-1}$$

Por consiguiente, fijado un nivel de confianza 1-α, podemos encontrar dos valores tales que la probabilidad del cociente anterior de encontrarse entre ambos sea precisamente 1-α.



Es decir:

$$\Pr ob\left(-t_{1-\alpha/2} < \frac{\bar{x}-\mu}{\frac{S}{\sqrt{n-1}}} \le t_{1-\alpha/2} \right) \quad = \quad 1-\alpha$$

como puede verse, se han elegido los valores que dejan por debajo y por encima, respectivamente, una probabilidad de $\alpha/2$. Esta no es la única elección posible, pero es la mejor, en el sentido de que proporciona de entre todos los intervalos de nivel de confianza $1-\alpha$ el de menor amplitud. Haciendo operaciones, de forma semejante a como hicimos en el ejemplo, obtenemos:

$$\Pr ob\left(-t_{1-\alpha/2} \frac{S}{\sqrt{n-1}} < \bar{x}-\mu \le t_{1-\alpha/2} \frac{S}{\sqrt{n-1}} \right) \quad = \quad 1-\alpha$$

restando ahora la media muestral

$$\Pr ob\left(-\bar{x} - t_{1-\alpha/2} \frac{S}{\sqrt{n-1}} < -\mu \le -\bar{x} + t_{1-\alpha/2} \frac{S}{\sqrt{n-1}} \right) \quad = \quad 1-\alpha$$

cambiando de signo

$$\Pr ob\left(\bar{x} + t_{1-\alpha/2} \frac{S}{\sqrt{n-1}} \ge \mu > \bar{x} - t_{1-\alpha/2} \frac{S}{\sqrt{n-1}} \right) \quad = \quad 1-\alpha$$

reordenando

$$\Pr ob\left(\bar{x} - t_{1-\alpha/2} \frac{S}{\sqrt{n-1}} < \mu \le \bar{x} + t_{1-\alpha/2} \frac{S}{\sqrt{n-1}} \right) \quad = \quad 1-\alpha$$

por consiguiente, la siguiente expresión proporciona un intervalo de confianza para la media de la población a un nivel de confianza 1-α:

$$\left(\bar{x} - t_{1-\alpha/2} \frac{S}{\sqrt{n-1}} \;,, \; \bar{x} + t_{1-\alpha/2} \frac{S}{\sqrt{n-1}} \right)$$

En lo sucesivo no tendremos más que sustituir los valores muestrales, en la expresión anterior, para obtener las correspondientes estimaciones.

Método de la probabilidad inversa

La anterior deducción de la expresión del intervalo de confianza ha seguido el procedimiento, llamado método de la probabilidad inversa. Este procedimiento de deducción es más sencillo e intuitivo que el método general de Neyman, pero requiere que exista un estadístico que cumpla las siguiente condiciones:

Ser una función continua y monótona del parámetro.

Que su distribución no dependa del parámetro.

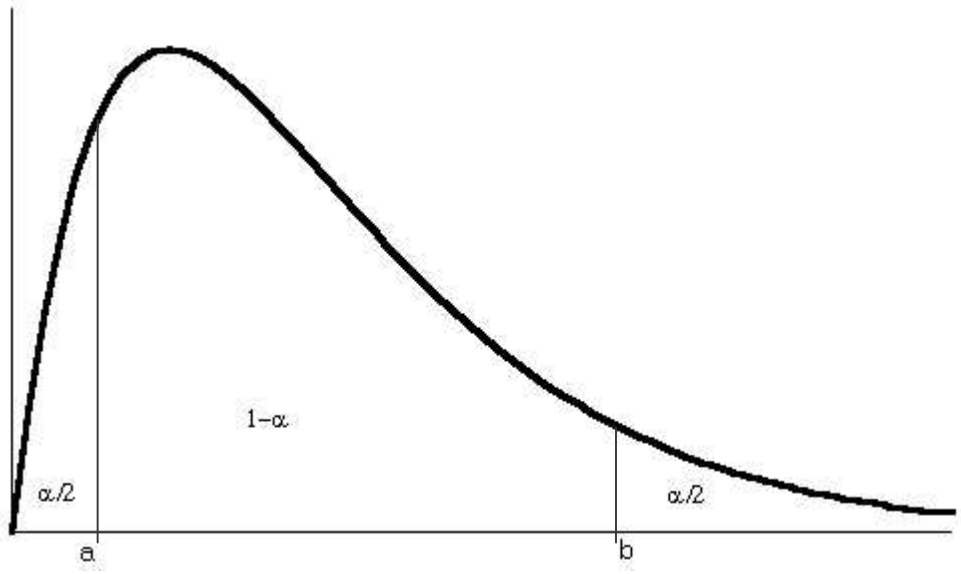
Ambas condiciones las cumple el cociente de Student y por ello hemos podido deducir el intervalo. Todos los intervalos que presentamos en este tema pueden deducirse por este método, no obstante en aquellas situaciones en que no sea posible determinar un estadístico con las condiciones señaladas deberá recurrirse al procedimiento general.

Intervalo para la varianza

Sea nuevamente una variable aleatoria X que sigue una distribución Normal de media μ y desviación típica σ, de la cual se extrae una muestra aleatoria simple de tamaño n. Por el teorema de Fisher tenemos que el estadístico:

$$\frac{nS^2}{\sigma^2} \quad \rightarrow \quad \chi^2_{n-1}$$

sigue una distribución Ji-cuadrado con n-1 grados de libertad. Por consiguiente en las tablas de la distribución Ji-cuadrado, podré determinar dos números *a* y *b,* tales que la probabilidad de que el estadístico se encuentre entre ambos sea igual al nivel de confianza fijado 1-α.



Por las mismas razones que en el caso anterior, estos números serán aquellos que dejen por debajo y por encima, respectivamente, una probabilidad de α/2. Por tanto, podemos escribir:

$$\Pr ob\left(a < \frac{nS^2}{\sigma^2} \leq b \right) \quad = \quad 1 - \alpha$$

dividiendo todos los miembros de la desigualdad por un número positivo las desigualdades no cambian de sentido y la probabilidad de que se cumplan es la misma, por consiguiente:

$$\Pr ob\left(\frac{a}{nS^2} < \frac{1}{\sigma^2} \leq \frac{b}{nS^2} \right) \quad = \quad 1 - \alpha$$

si invertimos los términos de las desigualdades, estas cambiarán de sentido y tendremos:

$$\Pr ob\left(\frac{nS^2}{a} \geq \sigma^2 > \frac{nS^2}{b} \right) \quad = \quad 1 - \alpha$$

reordenando, obtenemos finalmente:

$$\Pr ob\left(\frac{nS^2}{b} < \sigma^2 \leq \frac{nS^2}{a} \right) \quad = \quad 1 - \alpha$$

por lo que la expresión del intervalo de confianza para la varianza será:

$$\left(\frac{nS^2}{b} \,,, \frac{nS^2}{a} \right)$$

Intervalo de confianza para una proporción

Sea una población, en la que una determinada proporción de individuos P presentan una característica. De esta población se extrae una muestra aleatoria simple de tamaño n, en la cual una proporción p presentan la característica en cuestión. La proporción muestral sigue aproximadamente una distribución Normal:

$$p \quad \rightarrow \quad N\left(P, \sqrt{\frac{P \cdot Q}{n}}\right)$$

por consiguiente el estadístico:

$$\frac{p - P}{\sqrt{\frac{P \cdot Q}{n}}} \quad \rightarrow \quad N(0,1)$$

sigue una distribución Normal tipificada. Entonces fijado un nivel de confianza 1-α podemos encontrar en las tablas de la Normal dos números que verifiquen:

$$\Pr ob\left(-\lambda_{1-\alpha/2} < \frac{p - P}{\sqrt{\frac{P \cdot Q}{n}}} \leq \lambda_{1-\alpha/2}\right) \quad = \quad 1 - \alpha$$

operando de forma análoga, aunque algo más laboriosa, a los casos anteriores se obtiene la siguiente expresión del intervalo de confianza para una proporción:

$$\left(\frac{n}{n+\lambda^2}\left(p + \frac{\lambda^2}{2n} - \lambda\sqrt{\frac{p \cdot q}{n} + \frac{\lambda^2}{4n^2}}\right) \,,, \frac{n}{n+\lambda^2}\left(p + \frac{\lambda^2}{2n} + \lambda\sqrt{\frac{p \cdot q}{n} + \frac{\lambda^2}{4n^2}}\right)\right)$$

despreciando los términos menos significativos se obtiene la siguiente expresión aproximada:

$$\left(p - \lambda_{1-\alpha/2}\sqrt{\frac{p \cdot q}{n}} \,,, \, p + \lambda_{1-\alpha/2}\sqrt{\frac{p \cdot q}{n}}\right)$$

que es la empleada habitualmente.