Clase 6: Sesgo y error cuadrático medio

Matías Carrasco

6 de octubre de 2019

Índice

1. Propiedades de los estimadores

Como vimos, pueden haber varios estimadores razonables para un mismo parámetro θ. Es por esto que nos gustaría disponer de ciertos criterios que nos permitan elegir un estimador entre varios.

Sin embargo, esto es como cuando vamos a comprar una computadora nueva, existen varias características que pueden indicar que una computadora es mejor que otra, pero difícilmente nos podamos poner de acuerdo en un solo criterio que sirva para decir cuál es la mejor computadora. Con los estimadores pasa exactamente los mismo, dependiendo del contexto, ciertas personas priorizan algunos criterios sobre otros.

Para entender las dos nociones que discutiremos a seguir, supongamos que queremos saber qué hora es ahora mismo. Obviamente es imposible saberlo, pero podemos estimarla usando un reloj. Si disponemos de varios relojes, ¿qué reloj es mejor? Seguramente estés de acuerdo con los siguientes criterios:

- *Exactitud*: ¿está nuestro reloj en hora? ¿o está atrasado/adelantado?
- *Precisión*: ¿marca nuestro reloj los segundos? ¿o se limita a marcar los minutos?

Cuando se trata de un estimador, la primer propiedad refiere al *sesgo* y la segunda al *error cuadŕatico medio*[1].

2. Sesgo de un estimador

Un estimador debe estar "próximo" en algún sentido al valor verdadero del parámetro desconocido. De manera formal, se dice que T es un estimador *insesgado* de θ si el valor esperado de T es igual a θ.

[1] A veces al error cuadrático medio se lo llama error estándar.

Sesgo

El estimador T es un estimador insesgado del parámetro θ si

$$\mathbf{E}(T) = \theta.$$

Si el estimador es no es insesgado, entonces la diferencia

$$\text{Sesgo}(\theta) = \mathbf{E}(T) - \theta$$

es conocida como sesgo del estimador T.

Muchas veces nos interesan propiedades asintóticas de los estimadores. Es decir, propiedades para valores grandes de n. Así, aunque un estimador puede ser sesgado, puede que el sesgo tienda a cero a medida que n crece.

Sesgo asintótico

Un estimador T es un estimador asintóticamente insesgado del parámetro θ si

$$\lim_{n \to +\infty} \mathbf{E}(T) = \theta.$$

Es decir, si $\text{Sesgo}(\theta) \to 0$ cuando $n \to +\infty$.

Veamos un par de ejemplos. Si utilizamos el promedio $\overline{X}_n$ como estimador de μ la esperanza de X, entonces

$$\mathbf{E}\left(\overline{X}_n\right) = \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}(X_i) = \frac{\mu + \cdots + \mu}{n} = \mu.$$

Luego, $\overline{X}_n$ es un estimador insesgado de μ.

¿Qué pasa con la varianza? Veamos si Σ_n^2 es un estimador insesgado de σ^2. Calculamos

$$\begin{aligned}
\mathbf{E}\left(\Sigma_n^2\right) &= \mathbf{E}\left(\frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2\right) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(\left(X_i - \overline{X}_n\right)^2\right) \\
&= \frac{1}{n}\sum_{i=1}^{n}\mathbf{E}\left(X_i^2\right) - 2\mathbf{E}\left(X_i\overline{X}_n\right) + \mathbf{E}\left(\overline{X}_n^2\right)
\end{aligned}$$

Calculemos cada término por separado. Por un lado

$$\begin{aligned}
\mathbf{E}\left(X_i\overline{X}_n\right) &= \frac{1}{n}\sum_{j=1}^{n}\mathbf{E}(X_iX_j) = \frac{\mathbf{E}\left(X_i^2\right)}{n} + \frac{1}{n}\sum_{i\neq j}\mathbf{E}(X_i)\mathbf{E}(X_j) \\
&= \frac{\mu^2 + \sigma^2}{n} + \frac{n-1}{n}\mu^2 = \mu^2 + \frac{\sigma^2}{n}.
\end{aligned}$$

Por otro

$$\mathbf{E}\left(\overline{X}_n^2\right) = \frac{1}{n^2}\sum_{i,j=1}^{n}\mathbf{E}(X_iX_j) = \frac{\mu^2 + \sigma^2}{n} + \frac{n-1}{n}\mu^2 = \mu^2 + \frac{\sigma^2}{n}.$$

Juntando ambos cálculos obtenemos

$$\mathbf{E}\left(\Sigma_n^2\right) = \mu^2 + \sigma^2 - \mu^2 - \frac{\sigma^2}{n} = \frac{n-1}{n}\sigma^2.$$

Vemos así que Σ_n^2 es un estimador sesgado de σ^2. El sesgo de Σ_n^2 es

$$\text{Sesgo}\left(\Sigma_n^2\right) = -\frac{\sigma^2}{n} \to 0$$

cuando $n \to +\infty$. Entonces, Σ_n^2 es un estimador asintóticamente insesgado de σ^2.

Es por esto que muchas veces se define

$$S_n^2 = \frac{n}{n-1}\Sigma_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \overline{X}_n\right)^2$$

como estimador de σ^2, pues este sí es insesgado. Nosotros usaremos de todos modos ambos estimadores, S_n^2 y Σ_n^2, pues el sesgo es pequeño para muestras grandes.

3. Error cuadrático medio de un estimador

Supongamos que T_1 y T_2 son estimadores insesgados de θ. Esto indica que la distribución de cada estimador está centrada en el verdadero valor θ. Sin embargo, las varianzas de estas distribuciones pueden ser diferentes. La Figura 1 representa esta situación. Puesto que T_1 tiene una varianza más pequeña que T_2, es más probable que el estimador T_1 produzca una estimación más cercana al verdadero valor θ. Cuando se elige uno de entre varios estimadores, un principio útil es seleccionar el estimador que tenga la menor varianza.

A veces es necesario utilizar un estimador sesgado, como por ejemplo S_n^2. En tales casos la cantidad que mide la precisión del estimador es el error cuadrático medio, que es el cuadrado esperado de la diferencia entre T y θ.

Error cuadrático medio

El error cuadrático medio de un estimador T del parámetro θ está definido como

$$\text{ECM}(T) = \mathbf{E}\left((T-\theta)^2\right).$$

El error cuadrático medio puede reescribirse de la siguiente manera:

$$\begin{aligned}\text{ECM}(T) &= \mathbf{E}\left((T - \mathbf{E}(T))^2\right) + (\theta - \mathbf{E}(T))^2 = \mathbf{var}(T) + \text{Sesgo}(T)^2 \\ &= (\text{varianza}) + (\text{sesgo})^2\end{aligned}$$

Esto es, el error cuadrático medio de T es igual a la varianza del estimador más el cuadrádo del sesgo. Si T es un estimador insesgado de θ, el error cuadrático medio de T es igual a la varianza de T.

El error cuadrático medio es un criterio importante para comparar dos estimadores. Sean T_1 y T_2 do estimadores del parámetro θ, y $\text{ECM}(T_1)$ y $\text{ECM}(T_2)$ los errores cuadráticos medios de T_1 y T_2. Entonces, la *eficiencia relativa* de T_2 con respecto a T_1 se define como

$$\frac{\text{ECM}(T_1)}{\text{ECM}(T_2)}.$$

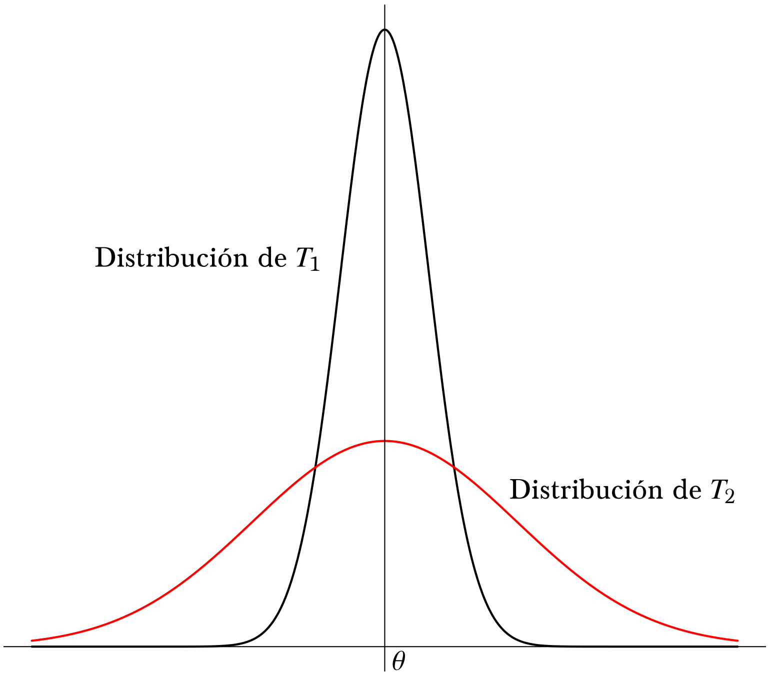

Figura 1: Distribuciones de muestreo de dos estimadores insesgados T_1 y T_2.

Si la eficiencia relativa es menor que uno, entonces puede concluirse que T_1 es un estimador más eficiente de θ que T_2, en el sentido de que tiene un error cuadrático medio más pequeño.

Por ejemplo, supongamos que queremos estimar la media μ de una población. Se tiene un muestreo aleatorio de n observaciones $X_1,\ldots,X_n$ y se quiere comparar dos estimadores posibles de μ: la media muestral $\overline{X}_n$ y una observación de la muestra, por ejemplo X_i.

Notar que $\overline{X}_n$ y X_i son ambos estimadores insesgados de μ; en consecuencia el error cuadrático medio de ambos estimadores es simplemente la varianza. Para la media muestral, se tiene

$$\mathrm{ECM}\left(\overline{X}_n\right) = \frac{\sum_{i=1}^{n} \mathbf{var}\left(X_i\right)}{n^2} = \frac{\sigma^2}{n}.$$

Para X_i la varianza es $\mathbf{var}\left(X_i\right) = \sigma^2$. Por consiguiente, la eficiencia relativa de X_i con respecto a $\overline{X}_n$ es

$$\frac{\mathrm{ECM}\left(T_1\right)}{\mathrm{ECM}\left(T_2\right)} = \frac{\sigma^2/n}{\sigma^2} = \frac{1}{n}.$$

Puesto que $1/n < 1$ si $n \geq 2$, podemos concluir que la media muestral es un mejor estimador de μ que una sola observación X_i.

Estimador consistente

Otra manera de medir la proximidad de un estimador T al parámetro θ es en términos de la *consistencia*. Denotemos el estimador como T_n para enfatizar que depende de un muestreo aleatorio de tamaño.

Consistencia

Si T_n es un estimador de θ basado en un muestreo aleatorio de n observaciones, entonces T_n es consistente para θ si

$$\lim_{n\to+\infty} \mathbf{P}(|T_n - \theta| < \varepsilon) = 1.$$

Esto se suele escribir $T_n \xrightarrow{P} \theta$ en donde la letra P recuerda que la convergencia es con probabilidad alta.

Es así que la consistencia es una propiedad de un muestreo grande, describe el comportamiento límite de T_n a medida que n tiende a infinito. Ver la Figura 2, en la cual se muestra la distribución de un estimador consistente a medida que n crece. Notar que la definición es equivalente a que

$$\lim_{n\to+\infty} \mathbf{P}(|T_n - \theta| \geq \varepsilon) = 0.$$

Estas probabilidades se muestran en la figura en rojo.

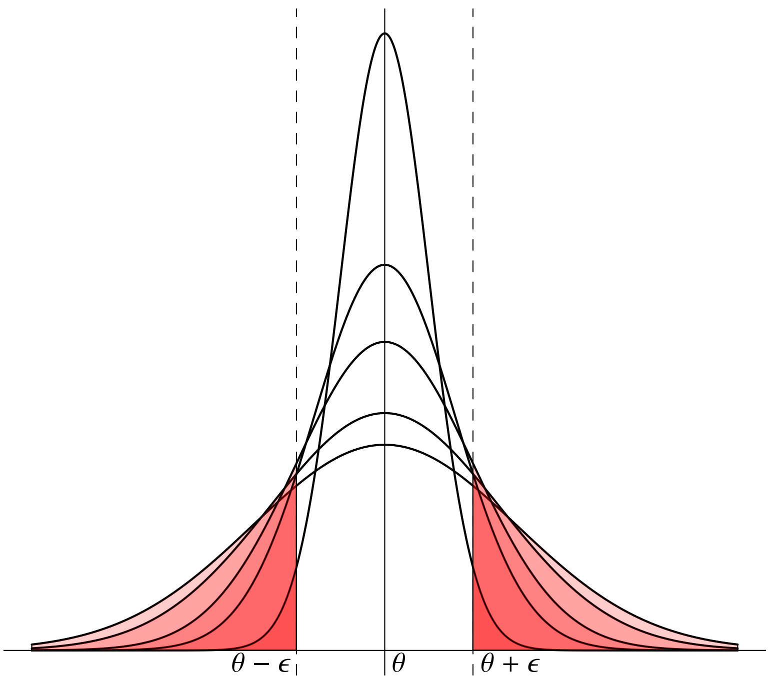

Figura 2: Distribución de un estimador consistente de θ para varios valores de n. Las áreas marcadas en rojo tienden a cero a medida que n crece.

Como veremos en seguida, una forma de mostrar que un estimador es consistente es probar que su error cuadrático medio tiende a cero cuando n tiende a infinito.

Recordar que la desigualdad de Chebychev establece una cota superior para las colas de una distribución de una variable que tiene varianza.

Sea T_n es un estimador de θ. Apliquemos la desigualdad de Chebyshev:

$$\mathbf{P}(|T_n-\theta|\geq\varepsilon)\leq\frac{\mathbf{E}\left((T_n-\theta)^2\right)}{\varepsilon^2}.$$

Notar que por definición $\mathrm{ECM}(T_n)=\mathbf{E}\left((T_n-\theta)^2\right)$. Hemos probado entonces la desigualdad siguiente:

Desigualdad de Chebychev para estimadores

Sea T_n un estimador del parámetro θ basado en un muestreo de tamaño n. Entonces, para todo $\varepsilon>0$ vale que

$$\mathbf{P}(|T_n-\theta|\geq\varepsilon)\leq\frac{\mathrm{ECM}(T_n)}{\varepsilon^2}.$$

El siguiente corolario es inmediato a partir de la desigualdad de Chebychev.

Criterio de consistencia

Sea T_n un estimador del parámetro θ basado en un muestreo de tamaño n. Si el error cuadrático medio de T_n tiende a cero cuando n tiende a infinito, entonces T_n es consistente.

Notar que un estimador con error cuadrático medio que tiende a cero es asintóticamente insesgado. De hecho, la desigualdad de Cauchy-Schwarz nos dice que

$$|\mathbf{E}(T_n)-\theta|\leq\mathbf{E}(|T_n-\theta|)\leq\sqrt{\mathbf{E}((T_n-\theta)^2)}\to 0$$

cuando $n\to+\infty$.

Consistencia del promedio y varianza muestral

Una consecuencia del criterio de consistencia de la sección anterior es que el promedio $\overline{X}_n$ es un estimador consistente de $\mu=\mathbf{E}(X)$ la esperanza de la distribución de la población.

Ley de los Grandes Números

El promedio muestral $\overline{X}_n$ es un estimador consistente de $\mu=\mathbf{E}(X)$.

La demostración es de nuevo muy simple. Por el criterio de consistencia que probamos antes, nos basta con probar que el error cuadrático medio de $\overline{X}_n$ tiende a cero cuando $n\to+\infty$. Sabemos del cálculo que hicimos más arriba que

$$\mathrm{ECM}\left(\overline{X}_n\right)=\frac{\sigma^2}{n}\to 0$$

cuando $n\to+\infty$, lo cual prueba nuestra afirmación.

¿Qué ocurre con Σ_n^2 o S_n^2? ¿Es un estimador consistente de σ^2? La respuesta es sí, pero para probarlo precisamos un argumento que nos permita aplicar una función continua a un estimador consistente.

Consistencia y continuidad
Si T_n es un estimador consistente del parámetro θ, y $g : \mathbb{R} \to \mathbb{R}$ es una función continua, entonces $g(T_n)$ es un estimador consistente de $g(\theta)$.

La prueba se basa en las definiciones de consistencia y de continuidad. Fijemos un $\varepsilon > 0$, y tomemos $\delta > 0$ tal que si $|x - \theta| < \delta$ entonces $|g(x) - g(\theta)| < \varepsilon$. El $\delta > 0$ sabemos que existe pues g es continua.

Esto implica que si $|T_n - \theta| < \delta$, entonces $|g(T_n) - g(\theta)| < \varepsilon$. Dicho de otro modo

$$\mathbf{P}\left(|T_n - \theta| < \delta\right) \leq \mathbf{P}\left(|g(T_n) - g(\theta)| < \varepsilon\right) \leq 1.$$

Si hacemos n tender a infinito, el lado izquierdo de la ecuación anterior tiende a 1, por lo que $\mathbf{P}\left(|g(T_n) - g(\theta)| > \varepsilon\right) \to 1$ cuando $n \to +\infty$. Esto prueba que $g(T_n)$ es un estimador consistente de $g(\theta)$.

Apliquemos este resultado a Σ_n^2. Por un lado, un cálculo directo muestra que

$$\Sigma_n^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \overline{X}_n^2.$$

El primer término del lado derecho converge, con probabilidad alta, a $\mathbf{E}\left(X^2\right)$ por la LGN aplicada a la sucesión i.i.d. de variables $X_1^2, \ldots, X_n^2$. Para ver la convergencia del segundo miembro aplicamos el teorema anterior con $g(x) = x^2$. Es decir, como $\overline{X}_n$ es un estimador consistente de $\mathbf{E}(X)$, entonces $\overline{X}_n^2$ es un estimador consistente de $\mathbf{E}(X)^2$. Entonces Σ_n^2 es un estimador consistente de $\mathbf{E}\left(X^2\right) - \mathbf{E}(X)^2 = \sigma^2$. Lo mismo vale para S_n^2.

¿Qué estimador elegir?

Un fabricante produce componentes eléctricos que tienen un tiempo de vida útil que se modela mediante una variable aleatoria X con distribución exponencial de parámetro λ. Para estimar λ se proponen los siguientes métodos:

1. Hacer un muestreo de n componentes y medir sus tiempos de vida útil $X_1, \ldots, X_n$ con total exactitud, observado en tiempo continuo a cada uno de ellos. Esta opción puede ser bastante costosa.

2. Observar los componentes una vez al día, de modo que si al comenzar el día un determinado componente está roto, pero estaba sano el día anterior, solo se puede deducir que se rompió en el lapso de 24hrs que transcurrieron entre las observaciones. En este caso se mide Y_i el redondeo al mayor entero más cercano de X_i. Esta opción es, sin dudas, menos costosa que la anterior.

3. Una opción intermedia respecto al costo es observar con total exactitud los componentes, pero solamente hasta que la mitad de ellos hayan fallado. Esto equivale a medir el tiempo de vida medio τ.

Consideremos el caso 1. Vamos a aplicar el método de los momentos para definir un estimador $\hat{\lambda}_1$ de λ. Como la esperanza $\mathbf{E}(X) = 1/\lambda$, vemos fácilmente que

$$\hat{\lambda}_1 = \frac{1}{\overline{X}_n}.$$

Notar que $\hat{\lambda}_1$ se puede calcular en el caso 1, pues estamos suponiendo que se miden con total exactitud las $X_i's$.

¿Es $\hat{\lambda}_1$ un estimador insesgado de λ? Para responder debemos calcular su distribución.

Comencemos por calcular la densidad de $Z = X_1 + X_2$. Recordando la fórmula de convolución para densidades, tenemos que

$$p(z) = \int_{-\infty}^{+\infty} p(x)p(z-x)dx = \lambda^2 \int_0^z e^{-\lambda x} e^{-\lambda(z-x)}dx = \lambda^2 z e^{-\lambda z}, \quad (z > 0).$$

En general, se puede probar por inducción que la densidad de la suma $X_1 + \cdots X_n$ es

$$p_n(x) = \frac{\lambda^n}{(n-1)!} x^{n-1} e^{-\lambda x}, \quad (x > 0).$$

Esta es la densidad de la distribución Gamma(n, λ), pero eso no es relevante. Usando la fórmula de cambio de variable se ve fácilmente que la densidad de $\hat{\lambda}_1$ es

$$p(y) = \frac{(n\lambda)^n}{(n-1)!} \frac{1}{y^{n+1}} e^{-n\lambda/y}, \quad (y > 0).$$

Siendo duchos con las integrales se puede probar que

$$\mathbf{E}\left(\hat{\lambda}_1\right) = \frac{n}{n-1}\lambda, \quad \mathbf{var}\left(\hat{\lambda}_1\right) = \frac{n^2}{(n-1)^2(n-2)}\lambda^2.$$

En particular $\hat{\lambda}_1$ no es insesgado, pero si es asintóticamente insesgado, ya que

$$\text{Sesgo}\left(\hat{\lambda}_1\right) = \frac{\lambda}{n-1} \to 0 \quad (n \to \infty).$$

Además, como la varianza de $\hat{\lambda}_1$ también tiende a cero, deducimos que ECM $\left(\hat{\lambda}_1\right)$ también tiende a cero cuando $n \to \infty$. En particular $\hat{\lambda}_1$ es consistente, aunque esto lo podríamos haber deducido directamente de la LGN y de que la función $g(x) = 1/x$ es continua para $x > 0$.

Veamos ahora el caso 2. Llamemos $Y_i = \lceil X_i \rceil$ el menor de los enteros mayores que X_i. Entonces lo que se mide en este caso es Y_i, y no X_i.

¿Cuál es la distribución de Y_i? Como Y_i es discreta, debemos calcular su f.p.p.. Para cada $k \geq 1$, tenemos que

$$\mathbf{P}(Y_i = k) = \mathbf{P}(k-1 < X_i \leq k) = \int_{k-1}^{k} \lambda e^{-\lambda x} dx = e^{-\lambda(k-1)} - e^{-\lambda k}.$$

Si llamamos $p = 1 - e^{-\lambda}$, podemos escribir la probabilidad anterior como

$$\mathbf{P}(Y_i = k) = p(1-p)^{k-1}, \quad (k \geq 1).$$

Es decir, Y_i tiene distribución geométrica de parámetro $p = 1 - e^{-\lambda}$. De aquí se puede despejar λ en función de p:

$$\lambda = -\ln(1-p).$$

Como la esperanza de una geométrica es $1/p$, el estimador de momentos de p es también $1/\overline{Y}_n$. Entonces un estimador razonable para λ es

$$\hat{\lambda}_2 = -\ln\left(1 - \frac{1}{\overline{Y}_n}\right).$$

A diferencia del caso anterior, las cuentas son bastante más complicadas para $\hat{\lambda}_2$. Por la ley de los grandes números y la continuidad de la función $g(y) = -\ln\left(1 - \frac{1}{y}\right)$ deducimos que $\hat{\lambda}_2$ es un estimador consistente de λ. Sin embargo es bastante más difícil calcular el sesgo.

La f.p.p. de $\overline{Y}_n$ es relativamente simple, y está dada por

$$\mathbf{P}\left(\overline{Y}_n = k/n\right) = \binom{k-1}{n-1} p^n (1-p)^{k-n}.$$

Luego, la fórmula para la esperanza de $\hat{\lambda}_2$ es

$$\mathbf{E}\left(\hat{\lambda}_2\right) = -\sum_{k=n}^{\infty} \ln\left(1 - \frac{n}{k}\right) \binom{k-1}{n-1} p^n (1-p)^{k-n}.$$

Pero esta suma es prácticamente imposible de calcular analíticamente. Sin embargo se puede probar (aplicando una desigualdad conocida que no vimos en el curso; la desigualdad de Jensen) que

$$\mathbf{E}\left(\hat{\lambda}_2\right) > \lambda,$$

por lo que $\hat{\lambda}_2$ no es insesgado. ¿Pero qué podemos decir del sesgo asintótico? Por ahora ¡no mucho! (más adelante veremos un método llamado *el método delta* que sirve para este caso). Una posibilidad es usar técnicas de análisis numérico, o incluso simular el experimento para varios valores de λ. Pero ambas opciones escapan al contenido del curso. Aunque sea una opción más barata, es más cara desde el punto de vista del análisis estadístico.

Veamos ahora la tercera opción. El tiempo de vida medio corresponde a estimar la mediana M de X usando la mediana muestral. Esto quiere decir que si ordenamos las variables

$$X_1^* < X_2^* < \cdots < X_n^*,$$

entonces $\tau = X^*_{\lfloor \frac{n+1}{2} \rfloor}$, y es de esperar que esté cerca de M. Como M es el valor que cumple

$$\mathbf{P}(X > M) = 1/2,$$

es fácil calcularla: $M = \ln(2)/\lambda$.

Si medimos experimentalmente el tiempo de vida medio τ, entonces

$$\hat{\lambda}_3 = \frac{\ln(2)}{\tau}$$

es un estimador razonable para λ.

Para facilitar las cuentas vamos a servirnos del siguiente truco. Cada X_i tiene distribución exponencial de parámetro λ, por lo que podemos escribir $X_i = Z_i/\lambda$ con Z_i exponencial de parámetro 1. De este modo τ se escribe τ_1/λ en donde $\tau_1 = Z^*_{\lfloor \frac{n+1}{2} \rfloor}$ es la mediana muestral de las $Z_i's$.

La ventaja de esto es que $\hat{\lambda}_3 = \lambda \frac{\ln(2)}{\tau_1}$, y por lo tanto

$$\mathbf{E}\left(\hat{\lambda}_3\right) = \lambda \ln(2) \mathbf{E}\left(\frac{1}{\tau_1}\right), \quad \mathbf{var}\left(\hat{\lambda}_3\right) = \lambda^2 (\ln(2))^2 \mathbf{var}\left(\frac{1}{\tau_1}\right),$$

con el aditivo de que la esperanza y la varianza de $1/\tau_1$ no dependen de λ. Si bien este cálculo sigue siendo difícil, se puede proceder con la ayuda de aproximaciones numéricas. Se puede ver que $\hat{\lambda}_3$ tiene sesgo, pero es asintóticamente insesgado. Esta opción también es intermedia desde el punto de vista del análisis estadístico.