Un apunte sobre estimadores

Manuel Delgado

Mayo 2025

1 La diferencia entre estadísticos y estimadores

Un estadístico $T(\overline{x})$ es cualquier función de los datos $\{x_i\}_{i=1,\ldots,n}$, que puede también depender de otros parámetros, como el número de mediciones, mientras estos parámetros sean conocidos.

Un estimador es un estadístico que uso para estimar algún parámetro de la distribución de probabilidad de los datos $\{x_i\}$. Es importante no utilizar el parámetro que quiero utilizar en la definición del estimador (o estadístico), ya que sino no lo puedo calcular.

2 Definición de verosimilitud (likelihood)

Supongamos que realizo n mediciones de una variable aleatoria $\{x_i\}_{i=1,\ldots,n}$ y por simplicidad supongamos que son variables aleatorias independientes e idénticamente distribuidas y cuya distribución de probabilidad tiene varios parámetros que agrupamos en $\overline{\theta}$. Es entonces que surge la pregunta: ¿Cuál es la probabilidad de que al medir obtuviéramos esos valores?

$$L(\overline{\theta}|\overline{x}) = \prod_{i=1}^{n} f_X(x_i, \overline{\theta})$$

A esta probabilidad la llamamos *verosimilitud* y es una función de los parámetros $\overline{\theta}$, ya que los x_i, las realizaciones de las variables aleatorias que medimos, son simplemente números. Es por esto que al integrar a la verosimilitud en todos los valores posibles de los parámetros, esta integral **no** es igual a 1.

Esta es una forma simple de escribir a la verosimilitud, pero si hubieramos medido variables aleatorias que vienen de distintas distribuciones, la verosimilitud igual se escribiría como una productoria. También podríamos tener variables que fueran dependientes y la verosimilitud se escribiría de la manera correspondiente a como estudiamos en la guía 4, aunque no nos centraremos en esos casos en lo que resta del curso.

Una pregunta que puede surgir más adelante es: ¿Qué ocurre si los datos vienen de una distribución de probabilidad que no es la que yo usé? Es decir: ¿Cómo me doy cuenta si mi modelo es correcto? Responder si nuestro modelo es correcto es algo que nunca podremos hacer, pero más adelante si vamos a aprender a como estudiar si nuestro modelo es incorrecto.

3 Propiedades de estimadores

Volvamos a hablar sobre estimadores.

$T(\overline{x})$ va a ser nuestro estimador y θ es el parámetro de la distribución que queremos estimar.

Consistencia

Decimos que un estimador es consistente si a medida que aumentamos el número de mediciones, o realizaciones de las variables aleatorias, el estimador se acerca cada vez más al parámetro. En otras palabras, podemos acercarnos al valor real del parámetro de la distribución tanto como querramos, solo es cuestión de medir. Lo importante también, es que nos acercamos en probabilidad, es decir, aumentamos las chances de que el estimador tenga un valor cercano al del parámetro.

Matemáticamente esto se representa como:

$$\lim_{n\to\infty} P(|T(\overline{x}) - \theta| > \epsilon) \to 0$$

donde ϵ es un número arbitrario que elegimos nosotros.

Una manera de comprobar si nuestro estimador es consistente es calculando la esperanza o varianza del estimador y chequeando si tiende al parámetro de la distribución o a 0, para $n \to \infty$, respectivamente.

Sesgo

Decimos que un estimador está sesgado si su esperanza no coincide con el parámetro, es decir:

$$E\left[T(\overline{x})\right] = \theta + b(\theta)$$

donde $b(\theta)$ es lo que llamamos sesgo (*bias* en inglés).

En el caso en que $b(\theta) = 0$ entonces decimos que el estimador es no sesgado.

Hay casos donde el estimador está sesgado, pero a medida que realizamos más mediciones el sesgo disminuye, es decir, si $\lim_{n\to\infty} b(\theta) \to 0$. Es entonces que decimos que el estimador es asintóticamente no sesgado.

Eficiencia

La eficiencia de cualquier estimador esta acotada inferiormente por la eficiencia del estimador de Cramer-Rao, cuyo estimador tiene mínima varianza (MV). La desigualdad de Cramer-Rao (generalizada para estimadores con sesgo) se escribe de la siguiente manera:

$$\mathrm{Var}\left[T(\overline{x})\right] \geq \frac{\left(\frac{\partial h}{\partial \theta} + \frac{\partial b}{\partial \theta}\right)^2}{E\left[-\frac{\partial^2 \ln L(\theta|\overline{x})}{\partial \theta^2}\right]}$$

donde $E\left[T(\overline{x})\right] = h(\theta) + b(\theta)$ donde de nuevo $b(\theta)$ es el sesgo del estimador.

En el caso en que se cumpla que el estimador es el de mínima varianza, se cumple la igualdad y vale la siguiente relación para el logaritmo de la verosimilitud (que si no me equivoco es para $b(\theta) = 0$):

$$\frac{\partial \ln L(\theta|\overline{x})}{\partial \theta} = A(\theta)\left[T(\overline{x}) - h(\theta)\right]$$

Que yo conozca, no existe un método sistemático para encontrar el estimador de mínima varianza, sino que es más probar si tu estimador cumple la relación de arriba.

Máxima verosimilitud

Lo que más nos gusta hacer a las personas que hacemos física, después de resolver el oscilador armónico o aproximar por Taylor, es maximizar funciones (o funcionales en el caso de la *acción* en Mecánica Clásica). Es entonces que vamos a asumir lo siguiente: lo que nosotros observamos, es lo más probable. Puede ser que hayamos tenido mala suerte y medimos cualquier cosa, pero igual vamos a asumir que lo que nos pasó, es lo más probable que podía pasar *¿Tanta mala suerte podemos tener?*.

Con esta esperanza es que buscamos maximizar la probabilidad de que haya pasado lo que nos pasó, es decir, la probabilidad de que hayamos medido los datos que medimos. Por eso es que proponemos maximizar la verosimilitud, para encontrar el parámetro θ que la maximiza. Se que estoy repitiendo muchas veces las mismas palabras, pero creo y espero que tengan sentido.

Dado que la verosimilitud está definida en muchas ocasiones como una productoria, a overlinees conviene maximizar el logaritmo natural de la verosimilitud, que dado que el logaritmo natural es una función monótonamente creciente, entonces tiene el mismo máximo que la verosimilitud.

Es entonces que el estimador de máxima verosimilitud nace de despejar $\hat{\theta}$ de la siguiente ecuación:

$$\left.\frac{\partial \ln L(\theta|\overline{x})}{\partial \theta}\right|_{\hat{\theta}} = 0$$

Suficiencia

Decimos que un estadístico es suficiente para un parámetro θ si tiene toda la información que puede obtener de los datos para poder decir algo sobre ese parámetro. Otra manera de decirlo es que un estadístico es suficiente si contiene toda la información que los datos proveen sobre los parámetros del modelo (*cita a wikipedia*). Es importante aclarar que el estadístico puede no estimar al parámetro y aún así ser suficiente. Es por eso que vengo usando la palabra estadístico y no estimador.

Otra cosa importante es que los estadísticos son suficientes para parámetros, no para distribuciones. Es decir, la siguiente pregunta no tiene sentido: ¿Cuál es un estadístico suficiente para la Gaussiana? Pero si tiene sentido preguntarse: ¿Cuál es un estadístico suficiente para el μ de la Gaussiana?

Como ejemplo para entender a que se refiere esto de tener toda la información de los datos, pensemos en el siguiente estadistico: $\sum_{i=1}^{n} x_i$. Este estadístico tiene toda la información necesaria de los datos para calcular el μ de la gaussiana (a excepción de la cantidad de mediciones), ya que no hace falta guardar cada una de las mediciones para decir algo sobre μ, con guardar la sumatoria de los datos alcanza.

Para bajarlo a la vida cotidiana, si voy a averiguar para hacer un doctorado en CITEDEF, lo que me importa es como me llevo con mi posible director, como es el lugar del trabajo y la gente, no si el dia que fui a consultar estaba lloviendo o me tuve que bajar del bondi porque chocó, esa no es información relevante para que yo pueda decir algo sobre lo que implica hacer un doctorado en CITEDEF.

La relación para comprobar si un estimador es suficiente es (y es un si y solo si):

$$L(\theta|\overline{x}) = g(T(\overline{x})|\theta)h(\overline{x})$$

Si calculamos el estimador de máxima verosimilitud, podemos ver que toda la información sobre los datos está contenida en el estadístico:

$$\left.\frac{\partial \ln L(\theta|\overline{x})}{\partial \theta}\right|_{\hat{\theta}} = \left.\frac{\partial g(T(\overline{x})|\theta)}{\partial \theta}\right|_{\hat{\theta}} = 0$$

Si $T(\overline{x})$ es un estadístico suficiente para θ entonces cualquier función del estadístico $f(T(\overline{x}))$ también es suficiente para θ.

Por último, si un estimador es eficiente, es decir tiene mínima varianza (Cramer-Rao), entonces es suficiente para el parámetro θ que estima. Esto tiene sentido, pues si tiene la mínima varianza, debe aprovechar toda la información de los datos para poder decir algo sobre θ con la menor de las incertezas posibles.

Condición de Darmois

Si la condición de Darmois se cumple entonces la distribución de probabilidad de la variable aleatoria X puede reescribirse como:

$$f_X(\overline{x}|\theta) = e^{B(\theta)C(x)+D(\theta)+E(x)}$$

A las distribuciones que cumplen esto se dice que pertenecen a la familia exponencial de distribuciones.

En caso de que se cumpla la distribución de Darmois, entonces el estadístico $T(\overline{x}) = \sum_{i=1}^{n} C(x_i)$ es un estadístico suficiente para θ. Un ejemplo simple de esto es el estadístico $\sum_{i=1}^{n} x_i$ que es suficiente para el parámetro λ (o la esperanza) de una distribución de Poisson.

4 Sobre la diferencia entre consistencia y sesgo

Abajo, en la figura 1, ejemplos que deberían representar distintos casos donde el estimador, es o no es consistente y es o no es sesgado:

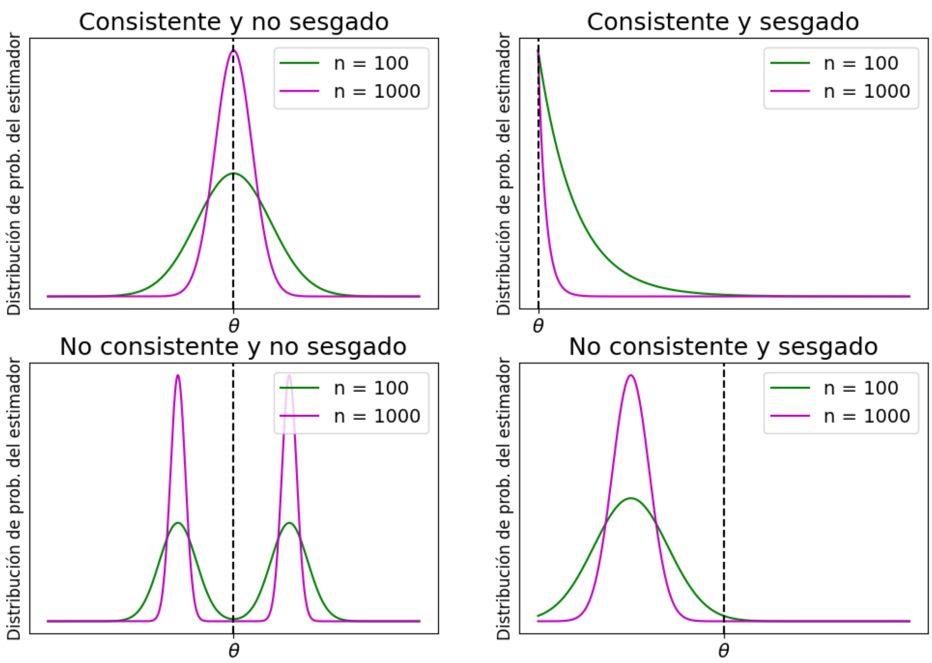

Figure 1: Distribuciones de probabilidad de varios estimadores para mostrar todas las combinaciones posibles entre sesgo y consistencia.

El caso de consistente y no sesgado es claro, a medida que aumento el número de mediciones el estimador se acerca más al valor real y en ambos casos la esperanza coincide con el parámetro.

El caso consistente y sesgado, vemos que me puedo acercar tanto como yo quiera al valor real de la distribución, pero en esperanza siempre me quedo un poquito a la derecha del valor real.

En el caso no consistente y no sesgado, la esperanza siempre coincide, pero al aumentar el número de mediciones el estimador empieza a oscilar alrededor del parámetro, sin poder acercarse cada vez más a este.

El caso no consistente y sesgado también es claro, en esperanza el estimador le erra completamente y a medida que mido más, no me acerco al parámetro.

Es importante que la cantidad de mediciones n, es un número representativo para mostrar como a medida que medimos más se vuelven más finas las distribuciones, no es que realicé n mediciones. Además este comportamiento podría no ocurrir para ciertos estimadores pero creo que ayuda a que nos hagamos una idea de que es lo que representan el sesgo y la consistencia.

Para cerrar y volverlo a decir: si un estimador es consistente, el estimador converge al parámetro a medida que aumentamos el número de mediciones, mientras que si un estimador es no sesgado, su esperanza coincide con el parámetro.

5 Reflexiones

Como se puede ver, el mundo de los estimadores es sorprendentemente amplio. Hay muchísimas condiciones, demasiados conceptos diferentes que representan cosas que nos parecen muy parecidas, pero son distintas y nos hace dudar un poco sobre todo lo que venimos haciendo a lo largo de la carrera y el poco respeto con el que hemos tratado en ocasiones a los datos.

Lo importante, creo yo, es aprender que todo esto existe y por más que no lo tuvimos en cuenta ayer y quizá no lo hagamos hoy en el día a día, nos hagamos una idea de qué cosas venimos haciendo bien y qué cosas venimos haciendo mal; un poco como viene pasando en toda la materia. Ojalá el día de mañana, sea el día en que podamos aprovechar todo esto al máximo (igual que un estadístico suficiente).

¡Espero que esto te haya servido! ¡Éxitos!