Tema 4: Inferencia Estadística I. Estimación Puntual.

Estadística. Grado en Ciencias del Mar

Concepto de inferencia estadística

La **inferencia estadística** es el proceso mediante el cual se extienden o generalizan a una población las conclusiones o resultados obtenidos a partir de la información proporcionada por una muestra de la misma.

Objetivos de la inferencia estadística

1. **Estimación de parámetros:** obtener valores aproximados de los parámetros que caracterizan el comportamiento de las variables de interés en la población.
2. **Contraste de hipótesis:** decidir sobre la validez o no de hipótesis relativas a alguna característica de la población.

Población y muestra

- En la práctica, la inferencia se realiza sobre **variables** (peso, talla, temperatura, concentración, velocidad, ...) que se miden en los elementos que componen la población.

- Por tanto cuando hablamos de caracterizar una población, en realidad nos referimos a caracterizar las variables de interés; y caracterizar una variable significa conocer en qué forma se reparten o distribuyen sus valores; en otras palabras, *conocer su distribución de probabilidad*.

- Para ello utilizamos la información que aporta una *muestra aleatoria*, definida como un conjunto de observaciones *independientes* $X_1, X_2, \ldots, X_n$ de la variable de interés.

¿De verdad una muestra informa sobre una población?

- **Función de distribución empírica:** Dada una muestra aleatoria $X_1, \ldots, X_n$, se define la función de distribución empírica:

$$\hat{F}_n(x) = \text{proporción de valores menores o iguales que } x \text{ en la muestra}$$

- **Teorema de Glivenko-Cantelli:** Sea $X_1, \ldots, X_n$ una muestra aleatoria de una variable aleatoria X con función de distribución $F(x)$, y sea $\hat{F}_n(x)$ la función de distribución empírica de la muestra. Entonces para cualquier valor x se verifica, a medida que $n \to \infty$:

$$E\left[\left(\hat{F}_n(x) - F(x)\right)^2\right] \to 0$$

- Por tanto a medida que aumenta el tamaño de la muestra, su distribución empírica se va asemejando cada vez más a la distribución de la variable de interés $\Rightarrow$ Efectivamente la muestra *informa* sobre la población.

Inferencia estadística paramétrica

- Cuando la variable X sobre la que deseamos realizar inferencias tiene una función de distribución caracterizada por un vector de parámetros $\theta = (\theta_1, \ldots, \theta_k)$, nuestro primer problema suele ser determinar un valor aproximado de θ.
- El proceso por el cuál se obtiene dicho valor aproximado se llama *estimación*. Un *estimador puntual* es una función de la muestra que produce valores próximos al parámetro que se desea conocer.
- ¿Cómo se construye un estimador?
    - Método de analogía
    - Método de los momentos
    - Método de máxima verosimilitud

¿Cómo se construye un estimador?

- **Método de analogía:** El parámetro poblacional se estima mediante su análogo en la muestra: la media poblacional se estima mediante la media muestral, la proporción en la población mediante la proporción en la muestra, ...

- **Método de los momentos:** El parámetro se expresa como función de los momentos (media, varianza, ... de la población) y se estima mediante la misma función evaluada a partir de los momentos análogos (media, varianza, ...) de la muestra.

- **Ejemplo:** en la distribución $Gamma\,(\alpha, \beta)$ se tiene que $\mu = \frac{\alpha}{\beta}$ y $\sigma^2 = \frac{\alpha}{\beta^2}$. Por tanto:

$$\beta = \frac{\mu}{\sigma^2} \Rightarrow \hat{\beta} = \frac{\overline{x}}{s^2}$$

$$\alpha = \mu\beta \Rightarrow \hat{\alpha} = \overline{x}\frac{\overline{x}}{s^2} = \frac{\overline{x}^2}{s^2}$$

¿Cómo se construye un estimador?

- **Método de analogía:** El parámetro poblacional se estima mediante su análogo en la muestra: la media poblacional se estima mediante la media muestral, la proporción en la población mediante la proporción en la muestra, …
- **Método de los momentos:** El parámetro se expresa como función de los momentos (media, varianza, … de la población) y se estima mediante la misma función evaluada a partir de los momentos análogos (media, varianza, …) de la muestra.
- **Método de máxima verosimilitud:** El parámetro se estima mediante aquel valor que maximiza *a priori* la probabilidad de observar la muestra que se ha observado.

Los distintos procedimientos pueden dar lugar a distintos estimadores para un mismo parámetro

Ejemplo

El abdomen del cangrejo de mar común (Carcinus maenas) está integrado por siete segmentos dispuestos paralelamente. En los machos se suelen apreciar fusiones entre los segmentos 3, 4 y 5. Se considera la variable aleatoria X=”Número de segmentos fusionados”. Esta variable puede tomar los valores 0 (ninguna fusión), 1 (se fusionan los segmentos 3 y 4, ó el 4 y 5), y 2 (se fusionan los tres segmentos entre sí). A través de diversas consideraciones sobre la genética de esta población de cangrejos, se llega a la conclusión de que las probabilidades asociadas a esta variable son de la forma:

$$P(X=0)=\frac{a-1}{a(a+1)} \quad P(X=1)=\frac{a-1}{a+1} \quad P(X=2)=\frac{1}{a}, \quad a>1$$

En una muestra de 100 cangrejos se han encontrado 18 sin fusiones, 43 que presentan una fusión y 39 que presentan dos fusiones. Utilizar esta información para obtener un valor aproximado de a:

a) Por analogía

b) Por el método de los momentos.

c) Por el método de máxima verosimilitud.

Método de analogía

El **método de analogía** consiste en expresar el parámetro como función de alguna operación numérica realizada con valores de la población, y calcular el estimador analógico como el resultado de aplicar *esa misma función* a los valores medidos en la muestra.

Ejemplo

En el ejemplo de los cangrejos se tiene que:

$$P(X=2) = \frac{1}{a} \rightarrow a = \frac{1}{P(X=2)}$$

Por tanto, si p_2 es la **proporción de cangrejos con dos fusiones en la muestra**, el estimador analógico de a es:

$$\hat{a} = \frac{1}{p_2}$$

Método de analogía

En nuestra muestra de 100 cangrejos hay 18 sin fusiones, 43 con una fusión y 39 con dos fusiones. Por tanto:

$$p_0 = \frac{18}{100} = 0.18 \qquad p_1 = \frac{43}{100} = 0.43 \qquad p_2 = \frac{39}{100} = 0.39$$

y el valor estimado de a es:

$$\hat{a} = \frac{1}{0.39} = 2.5641$$

Nótese que un **estimador analógico** es una **función de la muestra**. Una vez que la función se aplica y se obtiene un valor, éste es el *valor estimado del parametro*.

Por tanto, **distintas muestras** darán lugar a **distintos valores estimados.**

Método de analogía

Nótese que también podíamos haber argumentado que como:

$$P\left(X=1\right)=\frac{a-1}{a+1}\Rightarrow(a+1)P\left(X=1\right)=a-1\Rightarrow$$

$$\Rightarrow P(X=1)+1=a-aP(X=1)$$

de donde:

$$a=\frac{1+P(X=1)}{1-P(X=1)}$$

y por tanto **otro** estimador analógico **del mismo** parámetro a es:

$$\hat{a}=\frac{1+p_1}{1-p_1}$$

que en nuestro caso vale:

$$\hat{a}=\frac{1+0.43}{1-0.43}=\frac{1.43}{0.57}=2.5088$$

Método de analogía

También podíamos haber despejado a de:

$$P(X=0) = \frac{a-1}{a(a+1)} \Rightarrow (a^2+a)P(X=0) = a-1 \Rightarrow$$

$$\Rightarrow P(X=0)a^2 - (1-P(X=0))\,a + 1 = 0 \Rightarrow$$

$$a = \frac{(1-P(X=0)) \pm \sqrt{(1-P(X=0))^2 - 4P(X=0)}}{2P(X=0)}$$

El estimador analógico sería entonces:

$$\hat{a} = \frac{(1-p_0) \pm \sqrt{(1-p_0)^2 - 4p_0}}{2p_0}$$

Esta ecuación no siempre tiene solución; y cuando la tiene produce dos valores de $\hat{a} > 1$

Método de analogía

En nuestro ejemplo, como $p_0 = 0.18$ tenemos:

$$\hat{a} = \frac{0.82 \pm \sqrt{0.82^2 - 4 \cdot 0.18}}{2 \cdot 0.18}$$

que no tiene solución porque el término dentro de la raíz es negativo.

Método de los momentos

- Se define el **momento** de orden k de una variable aleatoria X como:

$$\mu_k = E\left[X^k\right]$$

- Asimismo se define el **momento muestral** de orden k de una muestra $X_1, X_2, \ldots, X_n$ como:

$$m_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

El **método de los momentos** consiste en expresar el parámetro como función de uno o varios momentos μ_k de la variable aleatoria y estimarlo como la misma función evaluada sobre los momentos muestrales correspondientes:

$$\theta = f\left(\mu_1, \mu_2, \ldots, \mu_r\right) \Rightarrow \hat{\theta} = f\left(m_1, m_2, \ldots, m_r\right)$$

Método de los momentos

El momento de primer orden es la esperanza, que en nuestro ejemplo es:

$$\mu = E\left[X\right] = 0 \cdot P\left(X = 0\right) + 1 \cdot P\left(X = 1\right) + 2 \cdot P\left(X = 2\right) =$$

$$= \frac{a-1}{a+1} + 2\frac{1}{a} = \frac{a\left(a-1\right) + 2\left(a+1\right)}{a\left(a+1\right)} = \frac{a^2+a+2}{a^2+a}$$

Por tanto:

$$a^2 + a + 2 = \mu\left(a^2 + a\right) \Rightarrow \left(\mu - 1\right)a^2 + \left(\mu - 1\right)a - 2 = 0$$

y de aquí podemos despejar a:

$$a = \frac{-\left(\mu-1\right) \pm \sqrt{\left(\mu-1\right)^2 + 8\left(\mu-1\right)}}{2(\mu-1)} = \frac{-\left(\mu-1\right) \pm \sqrt{\left(\mu-1\right)\left(\mu+7\right)}}{2(\mu-1)}$$

Para que $a > 1$ tomamos la raíz positiva:

$$a = \frac{-\left(\mu-1\right) + \sqrt{\left(\mu-1\right)\left(\mu+7\right)}}{2(\mu-1)}$$

(nótese que $\mu = \frac{a^2+a+2}{a^2+a} > 1$)

Método de los momentos

El estimador por el método de los momentos de a se obtiene sustituyendo el momento μ (la esperanza) por su homólogo muestral $\bar{x}$ (la media muestral):

$$\hat{a} = \frac{-\left(\overline{x}-1\right)+\sqrt{\left(\overline{x}-1\right)\left(\overline{x}+7\right)}}{2(\overline{x}-1)}$$

En nuestro ejemplo, el número medio de fusiones en el abdomen de la muestra de cangrejos es:

$$\overline{x} = \frac{0\cdot 18+1\cdot 43+2\cdot 39}{100} = \frac{121}{100} = 1.21$$

y por tanto:

$$\hat{a} = \frac{-0.21+\sqrt{0.21\cdot 8.21}}{2\cdot 0.21} = 2.6262$$

Nótese que aunque $\mu > 1$, en alguna muestra podría ocurrir que $\bar{x} < 1$ y en tal caso el estimador anterior no podría calcularse.

Método de máxima verosimilitud

El **estimador de máxima verosimilitud** se obtiene como aquel valor del parámetro que maximiza *a priori* la probabilidad de observar la muestra que se ha observado *de facto*.

En nuestro ejemplo, si se toma una muestra aleatoria de n cangrejos, a priori la probabilidad de que n_0 no tengan fusiones, n_1 tengan una fusión y n_2 tengan dos fusiones sería:

$$L\left(a\right) = \frac{n!}{n_0!n_1!n_2!}\pi_0^{n_0}\pi_1^{n_1}\pi_2^{n_2} = \frac{n!}{n_0!n_1!n_2!}\left(\frac{a-1}{a\left(a+1\right)}\right)^{n_0}\left(\frac{a-1}{a+1}\right)^{n_1}\left(\frac{1}{a}\right)^{n_2}$$

Esta función se denomina **función de verosimilitud**.

Método de máxima verosimilitud

Para obtener el valor de a es el valor que maximiza esta probabilidad podríamos derivar respecto de a e igualar a 0. **¡Complicado!**

Si tenemos en cuenta que el lugar donde una función alcanza el máximo es el mismo que donde lo alcanza su logaritmo, **maximizar la función anterior es equivalente a maximizar su logaritmo**.

El logaritmo de la verosimilitud en nuestro ejemplo es:

$$
\begin{aligned}
l\,(a) = \log\left(\frac{n!}{n_0!n_1!n_2!}\left(\frac{a-1}{a\,(a+1)}\right)^{n_0}\left(\frac{a-1}{a+1}\right)^{n_1}\left(\frac{1}{a}\right)^{n_2}\right) = \\
= \log\left(\frac{n!}{n_0!n_1!n_2!}\right) + n_0\left[\log(a-1) - \log(a) - \log(a+1)\right] + \\
+n_1\left[\log(a-1) - \log(a+1)\right] - n_2\log(a)
\end{aligned}
$$

Esta función se denomina **log-verosimilitud** y su derivada es normalmente sencilla de calcular.

Método de máxima verosimilitud

Simplificando:

$$l\left(a\right)=\log\left(\frac{n!}{n_0!n_1!n_2!}\right)+\left(n_0+n_1\right)\log(a-1)-$$

$$-\left(n_0+n_2\right)\log(a)-\left(n_0+n_1\right)\log(a+1)$$

Ahora es fácil derivar, igualar a cero y despejar:

$$l'\left(a\right)=\frac{n_0+n_1}{a-1}-\frac{n_0+n_2}{a}-\frac{n_0+n_1}{a+1}=0$$

$$\left(n_0+n_1\right)\left(a^2+a\right)-\left(n_0+n_2\right)\left(a^2-1\right)-\left(n_0+n_1\right)\left(a^2-a\right)=0$$

$$-\left(n_0+n_2\right)a^2+2\left(n_0+n_1\right)a+\left(n_0+n_2\right)=0$$

$$a=\frac{-2\left(n_0+n_1\right)\pm\sqrt{4(n_0+n_1)^2+4(n_0+n_2)^2}}{-2\left(n_0+n_2\right)}$$

Método de máxima verosimilitud

Simplificando, el estimador de máxima verosimilitud (MV) es:

$$\hat{a} = \frac{n_0 + n_1}{n_0 + n_2} \pm \sqrt{1 + \left(\frac{n_0 + n_1}{n_0 + n_2}\right)^2}$$

Como debe ocurrir que $\hat{a} > 1$, tomamos solamente la raíz positiva:

$$\hat{a} = \frac{n_0 + n_1}{n_0 + n_2} + \sqrt{1 + \left(\frac{n_0 + n_1}{n_0 + n_2}\right)^2}$$

En nuestro ejemplo $n_0 = 18$, $n_1 = 43$ y $n_2 = 39$. Por tanto:

$$\hat{a} = \frac{18 + 43}{18 + 39} + \sqrt{1 + \left(\frac{18 + 43}{18 + 39}\right)^2} = \frac{61}{57} + \sqrt{1 + \left(\frac{61}{57}\right)^2} = 2.5349$$

Método de máxima verosimilitud

En el caso particular de que $n_1 = n$, entonces $n_0 + n_2 = 0$, y el estimador anterior resulta ser $\hat{a} = \infty$

Este resultado era esperable, pues si $n_1 = n$ ello significa que todos los cangrejos tenían una única fusión y la función de verosimilitud quedaría reducida a:

$$L(a) = P(X = 1)^{n_1} = \left(\frac{a-1}{a+1}\right)^{n_1}$$

Es fácil comprobar que esta función es estrictamente creciente para $a > 1$; por tanto su máximo se alcanza para $\hat{a} = \infty$, lo que implica que:

$$\hat{P}\left(X = 1\right) = \lim_{a \to \infty} \frac{a-1}{a+1} = 1$$

es decir, si todos los cangrejos observados tienen una fusión, nuestra mejor estimación es que la probabilidad de tener una sola fusión es 1.

Método de máxima verosimilitud

PROBLEMA: ¿Cuál es el mejor estimador de todos los que hemos obtenido?

- Los estimadores analógicos no utilizan toda la información de la muestra: nuestros tres estimadores, o utilizan sólo p_0, o sólo p_1 o sólo p_2
- El estimador por el método de los momentos no es calculable si $\bar{x} < 1$
- En nuestro ejemplo el estimador MV utiliza toda la información de la muestra y puede calcularse siempre.

En general se suele preferir el método de máxima verosimilitud porque tiene varias propiedades que lo hacen particularmente interesante.

Propiedades de los estimadores MV:

Los estimadores de máxima verosimilitud son preferibles a los estimadores obtenidos por analogía o por el método de los momentos (en algunos casos los estimadores obtenidos por los distintos métodos coinciden, aunque no ocurre así en general), ya que gozan de mejores propiedades:

- **Consistencia:** los estimadores MV son consistentes, esto es, a medida que aumenta el tamaño de la muestra es más probable que el valor del estimador esté cada vez más próximo al valor del parámetro.

- **Eficiencia:** Si $\hat{\theta}$ es un estimador de un parámetro θ, el error cuadrático medio se define como $ECM\left[\hat{\theta}\right] = E\left[\left(\hat{\theta} - \theta\right)^2\right]$. A medida que aumenta el tamaño de muestra, los estimadores MV tienen el menor error cuadrático medio de entre los estimadores posibles (en otras palabras, los estimadores MV tienden a producir, en promedio, valores más próximos al verdadero valor del parámetro θ que otros estimadores).

- **Normalidad asintótica:** a medida que aumenta el tamaño de la muestra, los estimadores MV tienden a tener distribución normal. Esta propiedad permite construir intervalos de confianza.