ESTIMACIÓN PUNTUAL

1 Introducción

En este tema estudiaremos métodos para *estimar puntualmente parámetros* de algunas distribuciones estadísticas, las cuales constituyen elementos básicos de los sistemas inteligentes estocásticos. Dichas estimaciones se efectuarán a partir de los resultados obtenidos en una *muestra aleatoria simple* extraida de la distribución de probabilidad de referencia.

En primer lugar se introducen dos propiedades importantes a exigir a un buen estimador: la *insesgadez* y la *eficiencia*. Esta última surge ante la necesidad de escoger entre un número infinito de estimadores insesgados para un mismo parámetro.

Una vez estudiadas las propiedades que resultan de interés para un estimador, se presentará un método denominado de *máxima verosimilitud*, a partir del cual se pueden obtener estimadores que aunque no se garantiza que tengan las propiedades de insesgadez y eficiencia, si que van a verificar la interesante propiedad de que son asintóticamente insesgados y eficientes.

A lo largo del tema se ilustran los distintos conceptos a partir de las distribuciones *binomial* y *normal*.

2 Muestreo aleatorio simple

En las ciencias experimentales, entre las que se encuadra la Inteligencia Artificial, cuando se pretende estudiar un determinado problema –supongamos el comportamiento de un sistema inteligente estocástico– en la mayoría de las ocasiones resulta imposible establecer un modelo del mismo de manera analítica, y se hace por tanto necesario el experimentar con el sistema, para de esta manera recoger datos y a continuación utilizar dichos datos para especificar el modelo.

Veamos un ejemplo que puede servir para clarificar el párrafo anterior. Supongamos que, para resolver una instanciación del problema del agente viajero de la que previamente conocemos la solución, hemos implementado un heurístico estocástico de búsqueda del tipo de los que se estudiarán en la Parte II de la asignatura. Nuestro interés radica en conocer, y a ser posible modelar, el comportamiento de dicho heurístico estocástico de búsqueda.

Dada la estocasticidad del heurístico cada vez que lo ejecutamos podemos obtener una solución distinta, y supongamos además que con la condición de parada implementada, el tiempo de ejecución varía de un experimento a otro. Finalmente supongamos que, dada una ejecución, nos interesa poder estimar por una parte la probabilidad con la que se va a alcanzar el óptimo y por otra parte el tiempo que va a tardar en converger dicho heurístico.

Con ambos objetivos en mente, hemos ejecutado 20 veces el heurístico, habiéndose recogido los datos correspondientes a cada ejecución en la tabla 1.1.

Tabla 1.1: Resultados obtenidos en las 20 ejecuciones del heurístico estocástico de búsqueda.

Ejecución	$X =$ Éxito	$Y =$ Tiempo
1	si	34 sg.
2	no	29 sg.
3	no	31 sg.
4	si	33 sg.
5	si	28 sg.
6	si	33 sg.
7	no	27 sg.
8	no	35 sg.
9	si	31 sg.
10	no	29 sg.
11	no	27 sg.
12	no	30 sg.
13	si	35 sg.
14	si	29 sg.
15	no	38 sg.
16	si	29 sg.
17	si	33 sg.
18	no	30 sg.
19	no	25 sg.
20	no	14 sg.

La variable $X =$ Éxito toma 2 posibles valores. En dicha variable estamos interesados en estimar la probabilidad de cualquiera de los dos valores. La variable $Y =$ Tiempo se mueve en un dominio que –utópicamente– es continuo, y si admitimos un modelo de distribución normal para la misma, podemos estar interesados en estimar los dos parámetros de dicha distribución –la esperanza matemática y la varianza–. Volveremos sobre esta cuestión de las estimaciones en el siguiente apartado.

Lo reseñable en este punto es que a partir de la información contenida en la tabla 1.1 relativa a una *muestra aleatoria simple* de *tamaño* 20 extraida de dos variables aleatorias X e Y, tratamos de estimar unos parámetros. Hemos utilizado el término muestreo aleatorio simple para indicar que los resultados de los 20 experimentos pueden contemplarse como realizaciones de 20 *variables aleatorias independientes* e *idénticamente distribuidas.* La independencia hace alusión a que el resultado a obtener en una determinada ejecución no está condicionado por el del obtenido en cualquier otra ejecución, mientras que el término idénticamente distribuidas está refiriéndose a que el modelo probabilístico que rige el comportamiento de cada una de las ejecuciones es el mismo.

Veamos el concepto de una manera más precisa.

Definición 1.1
Dada una variable aleatoria X con función de cuantía (densidad), $P(X = x)$ $(f_X(x))$ conocida, una *muestra aleatoria simple* de tamaño n, $x_1, \ldots, x_n$, extraida de X, corresponde a la realización de n variables aleatorias, $X_1, \ldots, X_n$, independientes e idénticamente distribuidas, cuya ley de probabilidad coincide con la de X.

Suele ser habitual utilizar el término población para designar a la variable aleatoria X. Esta terminología tiene una clara reminiscencia sociológica, ya que en esta ciencia experimental es práctica habitual la extracción de muestras de una determinada población que se trata de estudiar.

Conviene aclarar que aunque existen otros tipos de muestreo –sistemático, estratificado, por conglomerados, polietápico, ...— no resultan de interés para el contenido de esta asignatura.

3 Estimador puntual

Siguiendo el ejemplo introducido en el punto anterior, supongamos que estamos interesados en estimar para la variable X la probabilidad, p, de que tome el valor si. Igualmente estamos interesados en estimar para la variable Y –recordemos que hemos supuesto que sigue un modelo de probabilidad normal– la esperanza matemática, μ, y la varianza poblacional, σ^2.

Estimadores intuitivos para dichos parámetros son respectivamente, la proporción de éxito, la media aritmética y la varianza muestral. A lo largo de este tema veremos que dichos estimadores presentan buenas propiedades.

DEFINICIÓN 1.2

Dada una variable aleatoria, X, a la que denominamos *población*, con ley de probabilidad conocida, $P(X = x; \boldsymbol{\theta})$ (o $f_X(x; \boldsymbol{\theta})$ en caso de que X sea continua), en la cual se desconoce el valor del parámetro k dimensional $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$, y dada una muestra aleatoria simple de tamaño n, $x_1, \ldots, x_n$, extraida de X, un *estimador* para el parámetro θ_j $(j = 1, \ldots, k)$ es una función $\theta_j^*(X_1, \ldots, X_n)$ construida a partir de las variables aleatorias independientes e idénticamente distribuidas asociadas a la muestra aleatoria simple. El valor de la función $\theta_j^*(X_1, \ldots, X_n)$ en la muestra aleatoria simple, es decir $\theta_j^*(x_1, \ldots, x_n)$, es la *estimación* del parámetro θ_j $(j = 1, \ldots, k)$.

EJEMPLO 1.1

En relación con la variable X introducida en el apartado anterior, conocemos que sigue una ley de probabilidad de Bernouilli con parámetro p desconocido. Es decir $X \to B(1, p)$, de ahí que:

$$P(X = x; \boldsymbol{\theta}) = P(X = x; p) \begin{cases} \text{p} & \text{si } x = 1 \\ \text{1 - p} & \text{si } x = 0. \end{cases} \tag{1}$$

En este caso el parámetro $\boldsymbol{\theta}$ es de dimensión uno y con su conocimiento tendríamos totalmente especificada la ley de probabilidad.[1]

Un estimador intuitivo para el parámetro p, puede ser la frecuencia relativa de éxito, es decir:

$$\boldsymbol{\theta}^*(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i. \tag{2}$$

Teniendo en cuenta la muestra aleatoria simple introducida en la Tabla 1.1, tenemos que una estimación para p es:

$$\boldsymbol{\theta}^*(x_1, \ldots, x_n) = \frac{1}{20}(1 + 0 + 0 + 1 + 1 + 1 + \ldots + 1 + 0 + 0 + 0) = \frac{9}{20}.$$

EJEMPLO 1.2

Si admitimos que la variable Y midiendo el tiempo de ejecución del heurístico estocástico de búsqueda sigue una ley de distribución normal, es decir: $Y \to \mathcal{N}(y; \mu, \sigma)$, con parámetros μ y σ desconocidos, tendremos que:

$$f_Y(y; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{y-\mu}{\sigma})^2} \tag{3}$$

con $y \in \Re, \mu \in \Re, \sigma \in \Re^+$.

En este caso el parámetro a estimar $\boldsymbol{\theta} = (\mu, \sigma)$ es bidimensional. Estimadores intuitivos para μ y σ son respectivamente la media muestral y la desviación típica muestral. Es decir:

$$\theta_1^*(X_1, \ldots, X_n) = \frac{1}{n} \sum_{i=1}^{n} X_i = \overline{X} \tag{4}$$

[1] Nótese que la ley de probabilidad de Bernouilli puede tambien expresarse como: $P(X = x; p) = p^x(1 - p)^{(1-x)}$ con $x = 0, 1$.

$$\theta_2^*(X_1,\ldots,X_n)=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2}=S_n \tag{5}$$

Teniendo en cuenta los datos de la tabla 1.1, las estimaciones correspondientes para ambos parámetros son:

$$\theta_1^*(x_1,\ldots,x_n)=\frac{1}{20}(34+29+\ldots+14)=30 \tag{6}$$

$$\theta_2^*(x_1,\ldots,x_n)=\sqrt{\frac{1}{20}(34-30)^2+(29-30)^2+\ldots+(14-30)^2}=\simeq 4.83 \tag{7}$$

4 Estimador insesgado

Un estimador puede verse como una variable aleatoria, ya que se define como una función de variables aleatorias. Las estimaciones son en concreto los valores que toma dicha variable aleatoria. Quiere esto decir que las estimaciones para un determinado parámetro dependen de la muestra aleatoria simple que se haya extraido de la población, y por tanto van a variar. Sería deseable que a pesar de esa variabilidad, por término medio los valores de dichas estimaciones se acerquen al valor del parámetro que se pretende estimar. Esta idea es la que subyace en el concepto de insesgadez de un estimador.

Definición 1.3
Un estimador $\theta(X_1,\ldots,X_n)$ para un parámetro θ se dice *insesgado* para dicho parámetro si se verifica:

$$\mathsf{E}[\theta^*(X_1,\ldots,X_n)]=\theta. \tag{8}$$

Es decir un estimador es insesgado para un parámetro cuando el valor esperado de dicho estimador coincide con el parámetro a estimar. Nótese que por abuso de notación hemos suprimido los subíndices tanto del estimador como del parámetro. Mientras no induzca a error esta será la notación habitual.

En el caso en que se verifique:

$$\mathsf{E}[\theta^*(X_1,\ldots,X_n)]=\theta+b(\theta) \tag{9}$$

con $b(\theta)\neq 0$, diremos que el estimador $\theta^*(X_1,\ldots,X_n)$ es sesgado para el parámetro θ. A la cantidad $b(\theta)$ se le denomina *sesgo*.

Teorema 1.1
La media aritmética, $\overline{X}=\frac{1}{n}\sum_{i=1}^{n}X_i$, es un estimador insesgado para el parámetro poblacional esperanza matemática, $\mathsf{E}(X)$.
Demostración: $\mathsf{E}(\overline{X})=\mathsf{E}(\frac{1}{n}\sum_{i=1}^{n}X_i)=\frac{1}{n}\sum_{i=1}^{n}\mathsf{E}(X_i)=\frac{1}{n}\sum_{i=1}^{n}\mathsf{E}(X)=\frac{1}{n}n\mathsf{E}(X)=\mathsf{E}(X)$.

Nótese que el resultado es válido independientemente de la ley de probabilidad de la variable aleatoria X.

Teorema 1.2
La cuasivarianza muestral, $S_{n-1}^2=\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2$, es un estimador insesgado para la varianza poblacional, $VarX$.
Demostración:

$$\mathsf{E}(S_{n-1}^2)=\mathsf{E}(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2)=\frac{1}{n-1}\mathsf{E}[(\sum_{i=1}^{n}(X_i-\overline{X})^2)]=\frac{1}{n-1}\mathsf{E}[(\sum_{i=1}^{n}(X_i-\mathsf{E}(X)-\overline{X}+\mathsf{E}(X))^2)]=$$

$$\frac{1}{n-1}\mathsf{E}[(\sum_{i=1}^{n}[(X_i-\mathsf{E}(X))-(\overline{X}-\mathsf{E}(X))]^2)]=$$

$$\frac{1}{n-1}\sum_{i=1}^{n}[\mathsf{E}(X_i-\mathsf{E}(X))^2+\mathsf{E}(\overline{X}-\mathsf{E}(\overline{X}))^2-2\mathsf{E}[(X_i-\mathsf{E}(X_i))(\overline{X}-\mathsf{E}(\overline{X}))]].(10)$$

Calculemos el valor de cada uno de los tres sumandos de la expresión anterior.

Al ser $X_1,\dots X_n$ variables aleatorias independientes e idénticamente distribuidas que la población X, tenemos que el primer término vale:

$$\mathsf{E}(X_i-\mathsf{E}(X))^2=\mathsf{E}(X-\mathsf{E}(X))^2=VarX. \tag{11}$$

Desarrollando el segundo término obtenemos:

$$\mathsf{E}(\overline{X}-\mathsf{E}(\overline{X}))^2=Var(\overline{X})=Var[\frac{1}{n}(X_1+\ldots+X_n)]. \tag{12}$$

Utilizando las propiedades de la varianza, la anterior expresión, resulta ser:

$$\mathsf{E}(\overline{X}-\mathsf{E}(\overline{X}))^2=\frac{1}{n^2}(VarX_1+\ldots+VarX_n)=\frac{1}{n^2}nVarX=\frac{VarX}{n}. \tag{13}$$

El desarrollo del tercer término nos conduce a:

$$\mathsf{E}[(X_i-\mathsf{E}(X_i))(\overline{X}-\mathsf{E}(\overline{X}))]=\mathsf{E}[(X_i-\mathsf{E}(X_i))(\frac{1}{n}[(X_1-\mathsf{E}(X))+\ldots+(X_n-\mathsf{E}(X))])]=$$

$$\frac{1}{n}\mathsf{E}[(X_i-\mathsf{E}(X_i))(X_1-\mathsf{E}(X_1))+\ldots+(X_i-\mathsf{E}(X_i))(X_i-\mathsf{E}(X_i))+\ldots+(X_i-\mathsf{E}(X_i))(X_n-\mathsf{E}(X_n))]=$$

$$\frac{1}{n}(0+\ldots+VarX_i+\ldots+0)=\frac{VarX}{n}. \tag{14}$$

Llevando los resultados obtenidos a la ecuación (1.10), obtenemos:

$$\mathsf{E}(S_{n-1}^2)=\frac{1}{n-1}\sum_{i=1}^{n}[VarX+\frac{VarX}{n}-2\frac{VarX}{n}]=$$

$$\frac{1}{n-1}\sum_{i=1}^{n}(\frac{nVarX+VarX-2VarX}{n})=\frac{1}{n-1}\sum_{i=1}^{n}\frac{(n-1)VarX}{n}=VarX. \tag{15}$$

Una lectura del teorema anterior en términos de la varianza muestral, $S_n^2=\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2$, viene a decir que:

$$\mathsf{E}(S_n^2)=\mathsf{E}(\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2)=\mathsf{E}(\frac{n-1}{n}\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2)=$$

$$\frac{n-1}{n}\mathsf{E}(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\overline{X})^2)=\frac{n-1}{n}VarX. \tag{16}$$

Es decir la varianza muestral no es un estimador insesgado para la varianza poblacional, sin embargo, si el tamaño de la muestra, n, fuese muy grande el valor esperado para S_n^2 tendería hacia $VarX$. Esta idea es la que subyace en el concepto de insesgadez asintótica que definimos a continuación.

Definición 1.4
Un estimador $\theta^*(X_1,\dots,X_n)$ se dice que es *asintóticamente insesgado* para un parámetro θ, si se verifica:

$$\lim_{n\to\infty}\mathsf{E}[\theta^*(X_1,\dots,X_n)]=\theta. \tag{17}$$

A continuación demostraremos un teorema cuyo significado en la práctica es que una vez que se tienen dos estimadores insesgados para un mismo parámetro, se es capaz de obtener un número infinito de estimadores insesgados para dicho parámetro.

TEOREMA 1.3
Cualquier combinación lineal de dos estimadores insesgados para un mismo parámetro es tambien un estimador insesgado para dicho parámetro.

DEMOSTRACIÓN: Sean $\theta_1^*(X_1,\ldots,X_n)$ y $\theta_2^*(X_1,\ldots,X_n)$ dos estimadores insesgados para el parámetro θ. Es decir: $\mathsf{E}[\theta_1^*(X_1,\ldots,X_n)]=\theta$ y $\mathsf{E}[\theta_2^*(X_1,\ldots,X_n)]=\theta$.

Sean $\lambda_1,\lambda_2\in\Re$ tales que $\lambda_1+\lambda_2=1$. Consideremos el estimador:

$$\theta^*(X_1,\ldots,X_n)=\lambda_1\theta_1^*(X_1,\ldots,X_n)+\lambda_2\theta_2^*(X_1,\ldots,X_n) \tag{18}$$

construido como combinación lineal de los dos estimadores insesgados. Veamos que $\theta^*(X_1,\ldots,X_n)$ es tambien un estimador insesgado para θ.

$$\mathsf{E}[\theta^*(X_1,\ldots,X_n)]=\mathsf{E}[\lambda_1\theta_1^*(X_1,\ldots,X_n)+\lambda_2\theta_2^*(X_1,\ldots,X_n)]=$$

$$\lambda_1\mathsf{E}[\theta_1^*(X_1,\ldots,X_n)]+\lambda_2\mathsf{E}[\theta_2^*(X_1,\ldots,X_n)]=\lambda_1\theta+\lambda_2\theta=\theta. \tag{19}$$

El teorema anterior nos plantea un problema de elección del mejor estimador de entre un número infinito de estimadores insesgados. Es evidente que independientemente del criterio que adoptemos para juzgar a los estimadores insesgados, no podemos aplicarlo a todos y cada uno de ellos (si existen dos estimadores insesgados, el teorema afirma que existe un número infinito de estimadores insesgados). Veremos en el apartado siguiente como abordar este problema de la elección a partir del concepto de eficiencia de un estimador.

5 Estimador eficiente

El concepto de *eficiencia de un estimador* está relacionado con el de su varianza. La idea es que entre dos estimadores cuyo valor medio coincida con el parámetro a estimar es preferible áquel cuya variabilidad sea menor, ya que de esta manera se conseguirán valores más cercanos al del parámetro que se quiere estimar.

Por tanto el estimador eficiente para un determinado parámetro será aquel que siendo insesgado tenga menor varianza dentro de los estimadores insesgados.

DEFINICIÓN 1.5
$\theta^*(X_1,\ldots,X_n)$ es un estimador eficiente para el parámetro θ, si se verifican las siguientes dos condiciones:

$$(i)\mathsf{E}[\theta^*(X_1,\ldots,X_n)]=\theta \tag{20}$$

$$(ii)Var[\theta^*(X_1,\ldots,X_n)]\leq Var[\theta^{'}(X_1,\ldots,X_n)] \tag{21}$$

para cualquier estimador $\theta^{'}(X_1,\ldots,X_n)$ que cumpla $\mathsf{E}[\theta^{'}(X_1,\ldots,X_n)]=\theta$.

A pesar de tener definido de manera precisa el concepto de estimador eficiente, el problema de determinar si un estimador insesgado es eficiente o no lo es, no parece tarea fácil, si lo abordamos directamente, es decir calculando la varianza de todos y cada uno de los infinitos estimadores insesgados que podemos construir.

Sin embargo a partir del teorema que enunciaremos (sin demostrar) a continuación, tendremos una consición suficiente para la determinación de la eficiencia de un estimador.

Teorema 1.4
Sean $X_1, \ldots, X_n$ un conjunto de variables independientes e idénticamente distribuidas que la variable X, de la cual se sabe que sigue una ley de probabilidad $P(X = x; \theta)$ (o $f_X(x; \theta)$ en caso de que X sea continua), con parámetro θ desconocido.

Sea $\theta^*(X_1, \ldots, X_n)$ un estimador para θ, verificando $\mathsf{E}[\theta^*(X_1, \ldots, X_n)] = \theta + b(\theta)$.

Se verifica la siguiente desigualdad:

$$Var(\theta^*(X_1, \ldots, X_n) \geq \frac{(1 + b'(\theta))^2}{n\mathsf{E}[(\frac{\partial \ln P(X=x;\theta)}{\partial \theta})^2]}. \tag{22}$$

El resultado anterior proporciona una cota –conocida como cota o frontera de Cramer-Rao– para la varianza de cualquier estimador. Se desprende del mismo que si el estimador es insesgado y alcanza la cota de Cramer-Rao, es decir si su varianza es la menor posible, entonces el estimador es eficiente. Sin embargo si el estimador no alcanzase dicha cota, el teorema no garantiza que dicho estimador no sea eficiente. Por tanto el teorema nos proporciona una condición suficiente para que el estimador sea eficiente.

A continuación veremos dos ejemplos de aplicación del resultado anterior.

Ejemplo 1.3
Dada una variable aleatoria X siguiendo una ley de distribución de Bernouilli de parámetro p desconocido, vamos a comprobar que el estimador que mide la frecuencia relativa de éxito obtenida en una muestra aleatoria simple de tamaño de n, es un estimador eficiente para p. Nótese que el estimador coincide con el propuesto en el ejemplo 1.1.

Tenemos que $X \to B(1, p)$. Por tanto $P(X = x; p) = p^x(1-p)^{(1-x)}$ con $x = 0, 1$. Recordemos que $\mathsf{E}(X) = p$ y $Var(X) = p(1-p)$. El estimador propuesto $\theta^*(X_1, \ldots, X_n)$ se ha definido como:

$$\theta^*(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i. \tag{23}$$

Veamos que dicho estimador es insesgado. Utilizando las propiedades de la esperanza matemática, tenemos:

$$\mathsf{E}[\theta^*(X_1, \ldots, X_n)] = \mathsf{E}[\frac{1}{n}\sum_{i=1}^{n} X_i] = \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}(X_i) = \frac{1}{n} n\ p = p. \tag{24}$$

Para ver si la varianza del estimador alcanza la cota de Cramer-Rao, calcularemos en primer lugar dicha varianza:

$$Var[\theta^*(X_1, \ldots, X_n)] = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2} Var(\sum_{i=1}^{n} X_i) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}. \tag{25}$$

La cota de Cramer-Rao, cuya expresión genérica es:

$$\frac{(1 + b'(\theta))^2}{n\mathsf{E}[(\frac{\partial \ln P(X=x;\theta)}{\partial \theta})^2]} \tag{26}$$

en este ejemplo se convierte en:

$$\frac{1}{n\mathsf{E}[(\frac{\partial \ln(p^x(1-p)^{(1-x)})}{\partial p})^2]} = \frac{1}{n\mathsf{E}[(\frac{\partial (x \ln p + (1-x)\ln(1-p))}{\partial p})^2]} = \frac{1}{n\mathsf{E}[(\frac{x}{p} - \frac{(1-x)}{(1-p)})^2]} =$$

$$\frac{1}{n\mathsf{E}[(\frac{(X - pX - p + pX)}{p(1-p)})^2]} = \frac{1}{n\mathsf{E}[(\frac{(X-p)}{p(1-p)})^2]} = \frac{p^2(1-p)^2}{nVarX} = \frac{p^2(1-p)^2}{np(1-p)} = \frac{p(1-p)}{n}. \tag{27}$$

Debido a la igualdad entre las expresiones obtenidas en las ecuaciones (1.25) y (1.27), podemos afirmar que el estimador $\frac{1}{n}\sum_{i=1}^{n} X_i$ es eficiente para p.

En el siguiente ejemplo vamos a generalizar el resultado al caso en que la variable aleatoria población sigue un modelo binomial.

Ejemplo 1.4
Dada una variable aleatoria X siguiendo una ley de distribución binomial de parámetros (m, p) con p desconocido, vamos a comprobar que el estimador que mide la frecuencia relativa de éxito obtenida a partir de una muestrea aleatoria simple de tamaño n, $x_1, \ldots, x_n$, es un estimador eficiente para p.

Antes de proceder con los cálculos de la esperanza matemática y la varianza, tratemos de ilustrar con un simple supuesto lo que se plantea en este ejemplo. Siguiendo con el supuesto del heurístico estocástico de búsqueda, supongamos ahora que efectuamos 50 repeticiones del experimento que se ilustraba en la tabla 1.1, que consistía en lanzar el heurístico 20 veces anotando si éste era capaz o no de encontrar el óptimo. Una vez anotado para cada una de las 50 repeticiones el número de éxitos (de los 20 en total posibles) encontrados, parece intuitivo que la suma del número de éxitos dividido por 1000 puede considerarse una buena estimación para la probabilidad de éxito.

Usando notación matemática, tenemos que $X \to B(m, p)$, con p desconocido. Es decir:

$$P(X = x; p) = \binom{m}{x} p^x (1-p)^{(m-x)} \tag{28}$$

con $x = 0, 1, \ldots, m$.

Recordemos que para la distribución binomial se tiene $\mathsf{E}(X) = mp$ y $Var(X) = mp(1-p)$.

Queremos comprobar que el estimador $\theta^*(X_1, \ldots, X_n) = \frac{1}{mn}\sum_{i=1}^{n} X_i$ es eficiente para el parámetro p.

En primer lugar nos planteamos por su insesgadez:

$$\mathsf{E}[\theta^*(X_1, \ldots, X_n)] = \mathsf{E}[\frac{1}{mn}\sum_{i=1}^{n} X_i] = \frac{1}{mn}\sum_{i=1}^{n} \mathsf{E}(X_i) = \frac{1}{mn}\sum_{i=1}^{n} mp = \frac{mnp}{mn} = p. \tag{29}$$

Por tanto el estimador propuesto es insesgado.

La varianza del estimador será:

$$Var[\theta^*(X_1, \ldots, X_n)] = Var[\frac{1}{mn}\sum_{i=1}^{n} X_i] = \frac{1}{m^2n^2}\sum_{i=1}^{n} Var(X_i) = \frac{nmp(1-p)}{m^2n^2} = \frac{p(1-p)}{mn}. \tag{30}$$

Veamos la expresión de la cota de Cramer-Rao en este caso:

$$\frac{1}{n\mathsf{E}[(\frac{\partial \ln(\binom{m}{x} p^x(1-p)^{(m-x)})}{\partial p})^2]} = \frac{1}{n\mathsf{E}[(\frac{\partial(\ln\binom{m}{x} + x\ln p + (m-x)\ln(1-p))}{\partial p})^2]} = \frac{1}{n\mathsf{E}[(\frac{x}{p} - \frac{m-x}{1-p})^2]} =$$

$$\frac{1}{n\mathsf{E}[(\frac{X-pX-mp+pX}{p(1-p)})^2]} = \frac{p^2(1-p)^2}{n\mathsf{E}[(X-mp)^2]} = \frac{p^2(1-p)^2}{nmp(1-p)} = \frac{p(1-p)}{mn}. \tag{31}$$

Del resultado anterior se desprende que el estimador $\frac{1}{mn}\sum_{i=1}^{n} X_i$ es eficiente para el parámetro p.

Veamos un último ejemplo, relativo en este caso a la distribución normal.

EJEMPLO 1.5
Dada una variable aleatoria X siguiendo una ley de distribución normal de parámetros (μ, σ) con μ desconocido, vamos a comprobar que el estimador media aritmética obtenido a partir de una muestrea aleatoria simple de tamaño n, $x_1, \ldots, x_n$, es un estimador eficiente para μ.

Nótese que este ejemplo guarda relación con la modelización asumida para la variable Tiempo del apartado 1.2. Tenemos que $X \to \mathcal{N}(x; \mu, \sigma)$. Es decir:

$$f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2} \tag{32}$$

con $x \in \Re, \mu \in \Re, \sigma \in \Re^+$. Recordemos que $\mathsf{E}(X) = \mu$ y $VarX = \sigma^2$.

El estimador propuesto es: $\theta^*(X_1, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^n X_i$. Comprobemos en primer lugar que se trata de un estimador insesgado.

$$\mathsf{E}[\theta^*(X_1, \ldots, X_n)] = \mathsf{E}[\frac{1}{n}\sum_{i=1}^n X_i] = \frac{1}{n}\sum_{i=1}^n \mathsf{E}(X_i) = \frac{1}{n}\sum_{i=1}^n \mu = \frac{n\mu}{n} = \mu. \tag{33}$$

La varianza del estimador se calculará de la siguiente manera:

$$Var[\theta^*(X_1, \ldots, X_n)] = Var[\frac{1}{n}\sum_{i=1}^n X_i] = \frac{1}{n^2}\sum_{i=1}^n Var(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}. \tag{34}$$

La cota de Cramer-Rao se establecerá a partir de los siguientes cálculos:

$$\frac{1}{n\mathsf{E}[(\frac{\partial}{\partial\mu}\ln(\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2})^2]} = \frac{1}{n\mathsf{E}[(\frac{\partial}{\partial\mu}(-\ln(\sqrt{2\pi}\sigma)-\frac{1}{2}(\frac{x-\mu}{\sigma})^2)]} =$$

$$\frac{1}{n\mathsf{E}[(\frac{x-\mu}{\sigma})^2)]} = \frac{\sigma^4}{n\sigma^2} = \frac{\sigma^2}{n}. \tag{35}$$

El resultado anterior confirma que la varianza del estimador alcanza la cota de Cramer-Rao y por tanto el estimador es eficiente.

6 La función de verosimilitud

Supongamos que hemos lanzado una moneda –de la que desconocemos con que probabilidad, p, sale cara (valor que será representado como 1)– al aire 10 veces, habiendo obtenido los siguientes valores:

$$x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 1, x_5 = 0, x_6 = 0, x_7 = 0, x_8 = 1, x_9 = 0, x_{10} = 1$$

los cuales constituyen la muestra aleatoria simple de tamaño 10.

Nos preguntamos por el valor del parámetro p que hace que los datos sean más creíbles o más verosímiles. En principio cualquier valor de p ha podido generar una muestra aleatoria simple como la anterior, pero no parece muy creíble que la moneda estuviese tan sesgada como para sostener que el valor de p fuese de 0.99 o incluso de 0.001.

Para calcular la verosimilitud de la muestra cuando el valor de p es 0.1, debemos de calcular la probabilidad de generar la anterior muestra aleatoria simple en el caso de que los datos se hayan generado a partir de un modelo de Bernouilli de parámetro $p = 0.1$. Es decir, en tal caso la verosimilitud de la muestra será:

$$P(B(1; 0.1) = 0) \cdot P(B(1; 0.1) = 1) \cdot P(B(1; 0.1) = 1) \cdot P(B(1; 0.1) = 1) \cdot P(B(1; 0.1) = 0) \cdot$$

$$P(B(1; 0.1) = 0) \cdot P(B(1; 0.1) = 0) \cdot P(B(1; 0.1) = 1) \cdot P(B(1; 0.1) = 0) \cdot P(B(1; 0.1) = 1) =$$

$$(0.1)^0 \cdot (0.9)^1 \cdot (0.1)^1 \cdot (0.9)^0 \cdot (0.1)^1 \cdot (0.9)^0 \cdot (0.1)^1 \cdot (0.9)^0 \cdot (0.1)^0 \cdot (0.9)^1 \cdot (0.1)^0 \cdot (0.9)^1 \cdot$$

$$(0.1)^0 \cdot (0.9)^1 \cdot (0.1)^1 \cdot (0.9)^0 \cdot (0.1)^0 \cdot (0.9)^1 \cdot (0.1)^1 \cdot (0.9)^0 \cdot = (0.1)^5 \cdot (0.9)^5 \simeq 0.0000059049.$$

La tabla 1.2 recoge algunos valores de la verosimilitud para distintos valores del parámetro.

Tabla 1.2: Verosimilitudes asociadas a la muestra aleatoria simple: $x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 1, x_5 = 0, x_6 = 0, x_7 = 0, x_8 = 1, x_9 = 0, x_{10} = 1$ extraida de una distribución de Bernouilli de parámetro p, según distintos valores de p.

Parámetro p	Verosimilitud
$p = 0.1$	$(0.1)^5 \cdot (0.9)^5$
$p = 0.2$	$(0.2)^5 \cdot (0.8)^5$
$p = 0.3$	$(0.3)^5 \cdot (0.7)^5$
$p = 0.4$	$(0.4)^5 \cdot (0.6)^5$
$p = 0.5$	$(0.5)^5 \cdot (0.5)^5$
$p = 0.6$	$(0.6)^5 \cdot (0.4)^5$
$p = 0.7$	$(0.7)^5 \cdot (0.3)^5$
$p = 0.8$	$(0.8)^5 \cdot (0.2)^5$
$p = 0.9$	$(0.9)^5 \cdot (0.1)^5$

Como veremos en el apartado siguiente un método de estimación de un parámetro va a consistir en la búsqueda del valor del parámetro que maximice la verosimilitud. No adelantemos acontecimientos y previamente definamos lo que se entiende por función de verosimilitud.

Definición 1.6
Dada una variable aleatoria, X, con ley de probabilidad conocida, $P(X = x; \boldsymbol{\theta})$ (o $f_X(x; \boldsymbol{\theta})$ en caso de que X sea continua), y dependiente de un parámetro k dimensional $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_k)$. Sea una muestra aleatoria simple de tamaño n, $x_1, \ldots, x_n$, realización de las variables aleatorias $X_1, \ldots, X_n$ las cuales son independientes e idénticamente distribuidas que X. La *función de verosimilitud* asociada a la muestra aleatoria simple, $x_1, \ldots, x_n$ dependiente del parámetro $\boldsymbol{\theta}$, se denota por $L(x_1, \ldots, x_n; \boldsymbol{\theta})$ y se define como:

$$L(x_1, \ldots, x_n; \boldsymbol{\theta}) = \prod_{i=1}^{n} P(X_i = x_i; \theta_1, \ldots, \theta_k) \tag{36}$$

En el caso de que la variable aleatoria X sea continua bastará con sustituir $P(X_i = x_i; \theta_1, \ldots, \theta_k)$ por $f_{X_i}(x_i; \theta_1, \ldots, \theta_k)$.

La función de verosimilitud se interpreta como la probabilidad (o densidad) de obtener la muestra aleatoria simple en cuestión cuando el valor del parámetro se ha fijado a $\boldsymbol{\theta}$).

Tanto en este tema –en el que los sistemas estocásticos considerados se reducen a modelos probabilísticos univariantes– como en temas posteriores –en los que el dominio se modelizará por medio de modelos probabilísticos multivariantes– la función de verosimilitud constituye una buena medida de la bondad de un modelo propuesto.

Veamos a continuación la expresión de la función de verosimilitud para las muestras aleatorias simples extraidas de distribuciones binomiales o normales.

Ejemplo 1.6
En el caso de una distribución binomial de parámetros (m, p), la función de verosimilitud asociada a una muestra aleatoria simple de tamaño n es:

$$L(x_1, \ldots x_n; p) = \prod_{i=1}^{n} \binom{m}{x_i} p^{x_i} (1-p)^{(m-x_i)} = (\prod_{i=1}^{n} \binom{m}{x_i}) p^{\sum_{i=1}^{n} x_i} (1-p)^{nm - \sum_{i=1}^{n} x_i}. \tag{37}$$

Ejemplo 1.7
En el caso de una distribución normal de parámetros (μ, σ), la función de verosimilitud asociada a una muestra aleatoria simple de tamaño n es:

$$L(x_1, \dots x_n; \mu, \sigma) = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}(\frac{x_i-\mu}{\sigma})^2} = (\frac{1}{\sqrt{2\pi}\sigma})^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} \quad (38)$$

7 Estimador máximo verosímil

Hasta el presente, en este tema hemos introducido el concepto de estimador puntual, hemos definido dos propiedades que cabe exigirle a los buenos estimadores puntuales, pero no tenemos por el momento un método con el que obtener estimadores. De hecho en los apartados anteriores cuando necesitábamos introducir algún estimador puntual hemos apelado a la intuición.

En este apartado vamos a presentar un método para obtener estimadores puntuales. El método se fundamenta en la función de verosimilitud asociada a una muestra aleatoria simple, y en esencia viene a decir que un estimador puntual para un parámetro se obtiene por medio de la expresión del parámetro que maximiza la función de verosimilitud.

Definición 1.7
Dada una variable aleatoria X con ley de probabilidad conocida y dependiente de un parámetro $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ desconocido. Supongamos que extraemos de X una muestra aleatoria simple, $x_1, \dots, x_n$. El *estimador máximo verosímil*, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$ se obtiene al maximizar la función de verosimilitud, $L(x_1, \dots, x_n; \theta_1, \dots, \theta_k)$.

A nivel práctico, y teniendo en cuenta que el logaritmo neperiano es una función monótona creciente, obtendremos el estimador máximo verosímil, $\hat{\boldsymbol{\theta}} = (\hat{\theta}_1, \dots, \hat{\theta}_k)$ para el parámetro $\boldsymbol{\theta} = (\theta_1, \dots, \theta_k)$ como solución del siguiente sistema de k ecuaciones y k incógnitas:

$$\begin{cases} \frac{\partial}{\partial\theta_1} \ln L(x_1, \dots, x_n; \theta_1, \dots, \theta_k) & = 0 \\ \vdots & \\ \frac{\partial}{\partial\theta_k} \ln L(x_1, \dots, x_n; \theta_1, \dots, \theta_k) & = 0. \end{cases} \quad (39)$$

Recurrimos de nuevo a las distribuciones binomial y normal para ilustrar con dos ejemplos el método descrito.

Ejemplo 1.8
Dada una muestra aleatoria simple de tamaño n extraida de una distribución binomial de parámetros (m, p), siendo p desconocido, veamos como obtener el estimador máximo verosímil para p.
Tal y como se ha obtenido en el ejemplo 1.6, la función de verosimiltud es:

$$L(x_1, \dots x_n; p) = (\prod_{i=1}^{n} \tbinom{m}{x_i})(p^{\sum_{i=1}^{n} x_i})(1-p)^{nm-\sum_{i=1}^{n} x_i}. (40)$$

Calcularemos en primer lugar el logaritmo neperiano de dicha función de verosimilitud:

$$\ln L(x_1, \dots x_n; p) = \sum_{i=1}^{n} \ln \tbinom{m}{x_i} + \sum_{i=1}^{n} x_i \ln p + (mn - \sum_{i=1}^{n} x_i) \ln(1-p). (41)$$

Al ser el parámetro a estimar p, unidimensional, el sistema genérico de k ecuaciones y k incógnitas, se reduce a la siguiente ecuación:

$$\frac{\partial}{\partial p} \ln L(x_1, \dots x_n; p) = \frac{\sum_{i=1}^{n} x_i}{p} + \frac{(mn - \sum_{i=1}^{n} x_i)(-1)}{1-p} = 0 \quad (42)$$

Si $p \neq 0$ y $p \neq 1$, la anterior ecuación es equivalente a:

$$\sum_{i=1}^{n} x_i - p\sum_{i=1}^{n} x_i - pmn + p\sum_{i=1}^{n} x_i = 0 \tag{43}$$

De ahí obtenemos que:

$$\hat{p} = \frac{\sum_{i=1}^{n} X_i}{mn}. \tag{44}$$

Nótese que este estimador máximo verosímil para el parámetro p coincide con el que en el ejemplo 1.4 hemos demostrado que es eficiente para p.

Ejemplo 1.9
Dada una muestra aleatoria simple de tamaño n extraida de una distribución normal de parámetros (μ, σ), ambos desconocidos, veamos como obtener los estimadores máximo verosímiles para ambos parámetros.

Usando la expresión para la función de verosimilitud asociada a la muestra obtenida en el ejemplo 1.7, tenemos que:

$$L(x_1, \ldots x_n; \mu, \sigma) = (\frac{1}{\sqrt{2\pi}\sigma})^n e^{-\frac{1}{2\sigma^2}(\sum_{i=1}^{n}(x_i-\mu)^2)}. \tag{45}$$

Calculamos a continuación el logaritmo neperiano de la anterior expresión y obtenemos:

$$\ln L(x_1, \ldots x_n; \mu, \sigma) = -n\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2. \tag{46}$$

En este caso el parámetro $\boldsymbol{\theta} = (\mu, \sigma)$ es bidimensional, de ahí que los estimadores máximo verosímiles para μ y σ –que se denotarán respectivamente por $\hat{\mu}$ y $\hat{\sigma}$– se obtendrán como solución del siguiente sistema de 2 ecuaciones y 2 incógnitas:

$$\left\{ \begin{array}{l} \frac{\partial}{\partial\mu}(-n\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}(\sum_{i=1}^{n}(x_i-\mu)^2) = 0 \\ \frac{\partial}{\partial\sigma}(-n\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}(\sum_{i=1}^{n}(x_i-\mu)^2) = 0 \end{array} \right. \tag{47}$$

Calculando las derivadas parciales, se tiene:

$$\left\{ \begin{array}{l} \frac{\sum_{i=1}^{n}(x_i-\mu)}{\sigma^2} = 0 \\ \frac{-n}{\sigma} + \frac{(\sum_{i=1}^{n}(x_i-\mu)^2)}{\sigma^3} = 0. \end{array} \right. \tag{48}$$

De la primera ecuación obtenemos –siempre que $\sigma \neq 0$– que:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} X_i = \overline{X}. \tag{49}$$

Sustituyendo este valor en la segunda ecuación obtenemos:

$$-n\sigma^2 + \sum_{i=1}^{n}(X_i - \overline{X})^2 = 0. \tag{50}$$

De ahí que el estimador máximo verosímil para σ sea:

$$\hat{\sigma} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}. \tag{51}$$

Veamos a continuación, obviando la demostración, dos resultados relativos a los estimadores máximo verosímiles.

Teorema 1.5
El estimador máximo verosímil es *invariante* frente a una transformación de los parámetros.

Es decir que si $\hat{\theta}(X_1, \ldots, X_n)$ es el estimador máximo verosímil para el parámetro θ, entonces el estimador máximo verosímil para $g(\theta)$ es $g(\hat{\theta}(X_1, \ldots, X_n))$.

Para ilustrar este teorema consideremos el ejemplo anterior, si bien ahora los dos parámetros sobre los que se quiere obtener los estimadores máximo verosímiles son μ y σ^2. Obviamente el estimador máximo verosímil para μ no variará. Por lo que respecta al estimador máximo verosímil para σ^2, al aplicar el teorema anterior con $g(\theta) = \theta^2$, tenemos que:

$$\hat{\sigma^2} = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2. \tag{52}$$

Puede comprobarse que se llega al mismo resultado si directamente se tratase de resolver el siguiente sistema de 2 ecuaciones y 2 incógnitas:

$$\begin{cases} \frac{\partial}{\partial\mu}(-n\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}(\sum_{i=1}^{n}(x_i - \mu)^2) = 0 \\ \frac{\partial}{\partial\sigma^2}(-n\ln(\sqrt{2\pi}\sigma) - \frac{1}{2\sigma^2}(\sum_{i=1}^{n}(x_i - \mu)^2) = 0. \end{cases} \tag{53}$$

El siguiente resultado se relaciona con las propiedades de los estimadores máximo verosímiles.

Teorema 1.6
Los estimadores máximo verosímiles son *asintóticamente insesgados y eficientes.*

Quiere esto decir que el método de estimación máximo verosímil proporciona unos estimadores que tienen unas buenas propiedades siempre y cuando el tamaño de la muestra sea lo suficientemente grande.

8 Notas bibliográficas

La teoría de la estimación fué fundada por Fisher en una serie de trabajos fundamentales. Así por ejemplo, el concepto de eficiencia se introdujo en Fisher (1921) y Fisher (1925), mientras que el método de máxima verosimilitud –si bien había sido utilizado por Gauss (1880) en casos particulares– fué establecido en toda su generalidad en Fisher (1912) y posteriormente en Fisher (1934).

Sin embargo el método de estimación puntual de parámetros más antiguo que se haya propuesto se debe a Pearson (1894) y es el denominado *método de los momentos*, que no ha sido tratado en este tema. Este método consiste en igualar un número conveniente de momentos muestrales a los correspondientes momentos poblacionales, que son funciones de los parámetros desconocidos. Considerando tantos momentos como parámetros haya que estimar, y resolviendo las ecuaciones resultantes respecto a dichos parámetros, se obtienen estimaciones de éstos.

Referencias que proporcionan una visión global del estado del arte en cuanto a la teoría de muestreo son Cochran (1977) y Kish (1965).

Una visión del muestreo más enfocada a la Inteligencia Artificial puede encontrarse en los libros de Mitchell (1997) y Cohen (1995).

9 Recursos en internet

- http://www-history.mcs.st-and.ac.uk/history/Mathematicians/Gauss.html
  Biografía de Johan Carl Friedrich Gauss
- http://www-history.mcs.st-and.ac.uk/history/Mathematicians/Pearson.html
  Biografía de Karl Pearson

- http://www-history.mcs.st-and.ac.uk/history/Mathematicians/Fisher.html
  Biografía de Sir Ronald Aylmer Fisher
- http://europa.eu.int/comm/eurostat
  Oficina Estadística de la Comunidad Económica Europea
- http://www.ine.es
  Instituto Nacional de Estadística
- http://www.eustat.es
  Instituto de Estadística de Euskadi
- http://www.spss.com
  Statistical Package for the Social Sciences
- http://www.sas.com
  Statistical Analysis Software
- http://www.intellektik.informatik.tu-darmstadt.de/ tom/IJCAI01/ijcai2001-empai.html
  Workshop on Empirical Methods in Artificial Intelligence dentro del International Joint Conference on Artificial Intelligence 2001

Ejercicios

1. Designemos por $S_1^2, S_2^2, ..., S_k^2$, las k varianzas muestrales basadas en muestras independientes de tamaños respectivos $n_1, n_2, ..., n_k$. Demostrar que una estimación insesgada para la varianza poblacional viene dada por
$$\frac{n_1 S_1^2 + ... + n_k S_k^2}{n_1 + ... + n_k - k}.$$
2. Dada una variable aleatoria X con distribución uniforme en el intervalo $(0, \theta)$, demostrar que:
   - el estimador $2\overline{X}$ es insesgado para θ
   - el estimador $(1 + \frac{1}{n})max(X_1, \ldots, X_n)$ es insesgado para θ.
3. Dada una variable aleatoria siguiendo una distribución de probabilidad de Poisson de parámetro λ desconocido. Supongamos que se extrae una muestra aleatoria de tamaño n de dicha población. Demostrar que la media muestral constituye un estimador eficiente para λ.
   Nota: La distribución de Poisson de parámetro λ es: $P(X = x; \lambda) = e^{-\lambda} \frac{\lambda^x}{x!}$ con $\lambda > 0; x = 0, 1, \ldots$
4. Sea A el evento "alcanzar el óptimo global en un problema de optimización combinatorial con un heurístico estocástico concreto".

   Sea p la probabilidad del evento A, y considérese la variable aleatoria X midiendo el número de ensayos independientes necesarios hasta que ocurra A.

   (i) Encontrar la estimación de máxima verosímilitud de p, correspondiente a un solo valor de X.

   (ii) Demostrar que utilizando una muestra de tamaño n, $x_1, x_2, ..., x_n$ el estimador máximo verosímil de p es $\hat{p} = 1/\overline{X}$.

   (iii) En relación con el apartado (i) supongamos que la primera vez que se alcanza el óptimo global es en la quinta ejecución del algoritmo estocástico de búsqueda. Calcular la estimación de máxima verosimilitud para el parámetro p.

   (iv) Hemos repetido 10 veces el experimento de medir el número de ensayos necesarios hasta alcanzar el óptimo global, habiendo obtenido los siguientes valores:

   $x_1 = 5, x_2 = 4, x_3 = 7, x_4 = 6, x_5 = 8, x_6 = 9, x_7 = 4, x_8 = 5, x_9 = 5, x_{10} = 6.$

   Calcular la estimación máximo verosímil para el parámetro p.

5. Una empresa dedicada a la venta de ordenadores ha modelado la variable aleatoria que mide el número de veces que un ordenador de una determinada marca se estropea en un periodo de 10 años. Según dichos estudios el modelo probabilístico subyacente sigue una ley de distribución de Poisson. Encuéntrese la estimación máximo verosímil de la probabilidad de que un ordenador de la marca anterior no se estropee durante un periodo de 10 años, empleando la siguiente muestra de tamaño 112.

No. veces ordenador estropeado	0	1	2	3	4	5
No. de ordenadores	44	42	21	9	4	2

6. Siguiendo con el problema de encontrar el óptimo global por medio de un heurístico estocástico, planteamos el experimento consistente en lanzar 100 pruebas cada una de ellas con 10 ejecuciones del algoritmo, e ir contando el número de veces que en cada una de las 100 pruebas se alcanza el óptimo. Los resultados se muestran en la siguiente tabla:

No. veces se alcanza el óptimo	0	1	2	3	4	5	6	7	8	9	10
No. de pruebas	0	1	6	7	23	26	21	12	3	1	0

Obtener el estimador máximo verosímil para la probabilidad de que dicho heurístico estocástico alcanze el óptimo global.

7. Un lago contiene N peces. Mediante un experimento con red se obtuvieron x peces, los cuales se marcaron y fueron arrojados de nuevo al agua. En un segundo experimento se capturaron y peces, encontrándose que z de ellos estaban marcados. Demostrar que el estimador máximo verosímil de N viene dado por $\hat{N} = xy/z$.

Referencias

1. P. R. Cohen (1995). *Empirical Methods for Artificial Intelligence.* Kluwer Academic Publishers.

2. W. G. Cochran (1977). *Sampling Techniques.* Wiley.

3. R. A. Fisher (1912). On an absolute criterion for fitting frequency curves. *Mess. of Mathematics*, 41, pp. 155.

4. R. A. Fisher (1921). On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society, A*, 222, pp. 309.

5. R. A. Fisher (1925). Theory of statistical estimation. *Proceedings of the Cambridge Philosophical Society*, 22, pp. 700.

6. R. A. Fisher (1934). Two new properties of mathematical likelihood. *Proceedings of the Royal Society, A*, pp. 285.

7. C. F. Gauss (1880). *Werke*, Vol. 4. Gotinga.

8. L. Kish (1965). *Survey Sampling.* Wiley.

9. T. M. Mitchell (1997). *Machine Learning.* McGraw-Hill.

10. K. Pearson (1894). Contribution to the mathematical theory of evolution. *Philosophical Transactions of the Royal Society, A*, 185, pp. 71.