Índice

Capítulo 1

Estimación por punto

En este capítulo se abordan problemas de estimación por punto, cuyos desarrollos teóricos se estudian en CB-capítulo 5 ó EII-capítulo 2. Entre ellos destacan los de la determinación de estimadores mediante el *método de la máxima verosimilitud*, así como los relacionados con la distribución en el muestreo de los estimadores utilizados en las situaciones más habituales, tales como el cálculo de determinadas probabilidades en las que aquellos están implicados, o la determinación del tamaño de la muestra para una precisión dada.

Problema 1.1

En un estudio sobre el efecto de la contaminación industrial en los alrededores de una gran ciudad, se eligieron al azar 10 huevos de pelícano de la isla de Anacapa situada frente a la ciudad californiana de Los Ángeles, observándose en ellos la concentración, en partes por millón, de bifemil policlorado PCB, un agente contaminante industrial. Los resultados obtenidos fueron los siguientes:

260 , 270 , 166 , 175 , 204 , 225 , 220 , 185 , 235 , 250

Suponiendo que la concentración del contaminante en estudio sigue una distribución normal de media μ, se pide:
a) **Determinar la estimación de máxima verosimilitud de μ.**
b) **Calcular la probabilidad de que μ y su estimador de máxima verosimilitud difieran, en valor absoluto, menos de 10 partes por millón.**

Si llamamos X a la variable aleatoria *concentración, en partes por millón, de PCB*, el enunciado del problema nos indica que se puede suponer para X una distribución normal $N(\mu, \sigma)$.

a) El estimador de máxima verosimilitud para μ, en esta situación de variable aleatoria normal de varianza desconocida, viene determinado en CB-ejemplo 5.4 ó EII-ejemplo 2.2, resultando ser la media muestral,

$$\hat{\mu} = \overline{x}$$

por lo que la estimación de máxima verosimilitud buscada será

$$\hat{\mu} = \overline{x} = \frac{260 + \cdots + 250}{10} = \frac{2190}{10} = 219.$$

b) La probabilidad pedida es

$$P\{|\overline{x} - \mu| < 10\}$$

para lo que necesitamos conocer la distribución en el muestreo de la media muestral, en la situación en la que nos movemos aquí de una población normal de varianza desconocida. En esta situación (véase CB-sección 5.4 ó EII-sección 2.4) la distribución de la media muestral (tipificada) es una t de Student:

$$\frac{\overline{x} - \mu}{S/\sqrt{n}} \rightsquigarrow t_{n-1}$$

en donde S es la cuasidesviación típica muestral,

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{x})^2}.$$

La probabilidad pedida será, por tanto,

$$\begin{aligned} P\{|\overline{x} - \mu| < 10\} &= P\left\{\frac{|\overline{x} - \mu|}{S/\sqrt{n}} < \frac{10}{S/\sqrt{n}}\right\} = P\left\{|t_9| < \frac{10}{35'99/\sqrt{10}}\right\} \\ &= P\left\{|t_9| < 0'88\right\} = 1 - 2 \cdot P\left\{t_9 > 0'88\right\} \approx 1 - 2 \cdot 0'2 = 0'6 \end{aligned}$$

en donde la probabilidad

$$P\{t_9 > 0'88\} \approx P\{t_9 > 0'883\} = 0'2$$

se obtiene de la Tabla 5 de la distribución t de Student.

Problema 1.2

Se quiere dar una estimación de máxima verosimilitud de la probabilidad p de sufrir una avería grave, que lleve a su sustitución, en la bomba del agua de un vehículo de una marca, modelo y año de fabricación determinados.
Para ello se eligieron al azar diez automóviles de la marca modelo y año en análisis y se anotó si habían tenido o no alguna avería grave en su bomba del agua desde su fabricación hace seis años. Los resultados obtenidos fueron los siguientes:

Automóvil número	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
¿Tuvo avería grave?	**SÍ**	**NO**	**NO**	**SÍ**	**NO**	**SÍ**	**SÍ**	**NO**	**NO**	**NO**

Se pide:
a) **Modelizar el problema planteado indicando la interpretación del parámetro p en el modelo que haya establecido.**
b) **Determinar la estimación de máxima verosimilitud de p.**
c) **Supuesto que contamos con una muestra de 101 automóviles del mismo modelo, marca y año que los que son objeto de estudio, calcular, aproximadamente, la probabilidad de que p y su estimador de máxima verosimilitud difieran, en valor absoluto, en menos de $0'1$.**

---

a) El problema se puede modelizar mediante una variable aleatoria dicotómica X que tome el valor 1 si el coche ha cambiado su bomba del agua y cero si no la ha cambiado. Denotando por p la probabilidad de que X tome el valor 1; es decir, la probabilidad de que cambie la bomba del agua, podemos modelizar X mediante una variable de Bernoulli $B(1,p)$ (la distribución binomial $B(1,p)$ recibe el nombre de distribución de Bernoulli), en donde p es la probabilidad de *éxito*: "haber cambiado la bomba del agua".

b) La función de masa de X es

$$p(x) = p^x\,(1-p)^{1-x} \qquad\qquad x = 0, 1$$

con lo que la función de verosimilitud de la muestra será

$$L(p) = \prod_{i=1}^{n} p^{x_i}\,(1-p)^{1-x_i}$$

de logaritmo

$$\log L(p) = \sum_{i=1}^{n} \left[x_i \log p + (1-x_i)\log(1-p)\right].$$

Su derivada igualada a cero —ecuación de verosimilitud— será

$$\frac{d}{dp}\log L(p) = \sum_{i=1}^{n}\left[\frac{x_i}{p} - (1-x_i)\frac{1}{1-p}\right] = 0$$

es decir,

$$\frac{\sum_{i=1}^{n} x_i}{p} - \frac{n - \sum_{i=1}^{n} x_i}{1-p} = 0$$

o bien,

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n}$$

es decir, la proporción muestral. La estimación de máxima verosimilitud será, ahora, el cociente entre los *éxitos* de la muestra y el tamaño de ésta; es decir,

$$\hat{p} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{4}{10} = 0'4.$$

c) La probabilidad pedida es

$$P\left\{|\hat{p} - p| < 0'1\right\}$$

para lo que necesitamos conocer la distribución en el muestreo de $\hat{p}$. Como contamos con una muestra de tamaño suficientemente grande —$n > 100$—, podemos aproximar la distribución de $\hat{p}$ mediante una normal (CB-sección 5.5 ó EII-sección 2.5) de la forma

$$\frac{\hat{p} - p}{\sqrt{\frac{p \cdot (1-p)}{n}}} \approx N(0,1)$$

con lo que la probabilidad pedida será,

$$P\left\{|\hat{p} - p| < 0'1\right\} = P\left\{\frac{|\hat{p} - p|}{\sqrt{\frac{p \cdot (1-p)}{n}}} < \frac{0'1}{\sqrt{\frac{p \cdot (1-p)}{n}}}\right\} \approx P\left\{|Z| < \frac{0'1}{\sqrt{\frac{p \cdot (1-p)}{101}}}\right\}$$

siendo Z una variable aleatoria con distribución $N(0,1)$. Como no conocemos p, para el cálculo de la probabilidad anterior utilizaremos como estimación suya la obtenida en el apartado *b)*, con lo que la probabilidad pedida será, aproximadamente igual a

$$P\left\{|Z| < 0'1/\sqrt{0'4 \cdot 0'6/101}\right\} = P\left\{|Z| < 2'05\right\} = 1 - 2 \cdot P\left\{Z > 2'05\right\} = 1 - 2 \cdot 0'0202 = 0'9596$$

en donde la probabilidad

$$P\left\{Z > 2'05\right\} = 0'0202$$

se ha obtenido de la Tabla 3 de la distribución $N(0,1)$.

Problema 1.3

Se sabe que el número X de clientes que acuden a un determinado servicio informático es una variable aleatoria discreta con función de masa o probabilidad

$$p_\theta(x) = \frac{(\log \theta)^{x-1}}{\theta\,(x-1)!} \qquad x = 1, 2, 3, ...$$

siendo $\theta > 1$ un parámetro desconocido. Utilizando una muestra aleatoria simple de X de tamaño n, determinar el estimador $\widehat{\theta}$ de máxima verosimilitud de θ.

Determinando previamente la distribución de $Y = X - 1$ y si en una muestra previa de tamaño $n = 101$ se obtuvo una media muestral $\overline{x} = 20$, calcular aproximadamente la probabilidad

$$P\{\widehat{\theta} > e\}$$

La función de verosimilitud de la muestra es

$$L(\theta) = p_\theta(x_1, ..., x_n) = \prod_{i=1}^{n} p_\theta(x_i) = \frac{(\log\theta)^{\sum_{i=1}^n x_i - n}}{\theta^n \prod_{i=1}^n (x_i - 1)!}$$

la cual tiene por logaritmo

$$\log L(\theta) = \left(\sum_{i=1}^{n} x_i - n\right) \log\log\theta - n\,\log\theta - \log\prod_{i=1}^{n}(x_i - 1)!$$

cuya derivada igualada a cero es

$$\frac{\partial}{\partial\theta}\log L(\theta) = \left(\sum_{i=1}^{n} x_i - n\right)\frac{1}{\log\theta}\,\frac{1}{\theta} - \frac{n}{\theta} = 0$$

de donde despejando se obtiene como estimador máximo-verosímil para θ,

$$\widehat{\theta} = e^{\overline{x}-1}.$$

La función de masa de $Y = X - 1$ será

$$p_\theta(y) = P\{Y = y\} = P\{X - 1 = y\} = P\{X = y + 1\} = \frac{(\log\theta)^y}{\theta\, y!}$$

si $y = 0, 1, 2, ...$, que es la función de masa de una distribución de Poisson de parámetro $\log\theta$.

La probabilidad pedida será

$$P\{\widehat{\theta} > e\} = P\{e^{\overline{x}-1} > e\} = P\left\{\frac{1}{n}\sum_{i=1}^{n}(X_i - 1) > 1\right\} = P\{\overline{y} > 1\}$$

Como, según hemos visto, las $Y_i = X_i - 1$ siguen una distribución $\mathcal{P}(\log\theta)$ y el tamaño muestral es suficientemente grande ($n = 101$), la media muestral $\overline{y}$ sigue aproximadamente una distribución normal (véase CB-sección 5.5 ó EII-sección 2.5)

$$\frac{\overline{y} - \log\theta}{\sqrt{\overline{y}/n}} \approx N(0,1)$$

con lo que, tipificando, se obtiene que la probabilidad pedida es aproximadamente

$$P\left\{Z > -41'5\right\} \approx 1$$

con $Z \rightsquigarrow N(0,1)$.

Problema 1.4

Por razones aún desconocidas, el porcentaje, p, de esquizofrénicos en todos los países es, de forma invariable, del $1\,\%$. Determinar el tamaño de muestra necesario para que el porcentaje de esa muestra difiera en términos absolutos de p en menos de $0'003$ con probabilidad $0'9$, suponiendo que dicho tamaño muestral va a resultar grande.

Nos piden el tamaño de muestra necesario (véase el ejemplo 5.8 de CB) para que se verifique la igualdad

$$P\{|\hat{p} - p| < 0'003\} = 0'9 \qquad [1.1]$$

suponiendo que es

$$\hat{p} \approx N\left(p\,,\,\sqrt{p(1-p)/n}\,\right) \equiv N\left(0'01\,,\,\sqrt{0'01\cdot 0'99/n}\,\right)$$

al ser el tamaño muestral suficientemente grande.

Tipificando en [1.1] se obtiene que es

$$P\{|Z| < 0'003\frac{\sqrt{n}}{\sqrt{p(1-p)}}\} = 0'9.$$

con $Z \rightsquigarrow N(0,1)$. Como por otro lado es

$$P\{|Z| < 1'645\} = 0'9$$

será

$$0'003\frac{\sqrt{n}}{\sqrt{p(1-p)}} = 1'645$$

obteniéndose de ahí el valor $n = 2976'6$; es decir, son necesarios $n = 2977$ individuos para alcanzar la precisión deseada.

Problema 1.5

Se sometió a 9 personas a un curso intensivo de dudosa eficacia, de informática, anotándose el nivel de conocimientos de estos nueve alumnos antes del comienzo del curso, X, y una vez finalizado éste, Y. Los resultados obtenidos por los nueve estudiantes fueron los siguientes:

X_i	7	6	5	3	6	2	6	5	7
Y_i	8	6	4	6	7	6	5	6	7

Admitiendo para X e Y distribuciones normales de igual media, calcular la probabilidad de que repitiendo el curso con una nueva muestra también de 9 alumnos, se obtuviera una diferencia de medias muestrales mayor que la obtenida en ésta (es decir, se mejoraran los resultados del curso realizado), suponiendo que, en esa nueva muestra, la cuasivarianza muestral será la misma que en el experimento realizado.

El enunciado nos dice que puede admitirse para X e Y las distribuciones, $X \rightsquigarrow N(\mu, \sigma_1)$ e $Y \rightsquigarrow N(\mu, \sigma_2)$. Claramente éste es un experimento de Datos Apareados puesto que las calificaciones se obtienen en los mismos nueve individuos. (Véase el ejemplo 5.13 de CB).

La variable diferencia $D = Y - X$ (mejora de conocimientos) seguirá una distribución $D \rightsquigarrow N(0, \sigma_d)$ y la media muestral de las diferencias (es decir, la diferencia de medias muestrales), $\overline{d} = \overline{y} - \overline{x}$, una distribución

$$\frac{\overline{d} - 0}{S_d/\sqrt{n}} \rightsquigarrow t_{n-1}.$$

Como de los datos del enunciado se obtiene, para la variable diferencia D, una media muestral igual a $0'89$ y una cuasivarianza muestral de $S_d^2 = 2'86$, la probabilidad que nos piden es que para un nuevo curso,

$$P\{\overline{d} > 0'89\} = P\left\{t_8 > \frac{0'89}{\sqrt{2'86/9}}\right\} = P\left\{t_8 > 1'58\right\} = 0'08$$

valor obtenido por interpolación lineal a partir de los datos de la tabla de la t de Student.

Problema 1.6

El tiempo en días que tarda un ordenador en quedar inutilizado por un determinado virus informático es una variable aleatoria X con la siguiente función de densidad:

$$f_\theta(x) = \frac{1}{2}\,\theta^3\,x^2\,e^{-\theta x} \qquad\qquad x > 0$$

siendo θ un parámetro desconocido. En 5 ordenadores elegidos al azar, el virus en estudio dejó inutilizado el ordenador al cabo de 15, 20, 10, 13 y 12 días. Determinar la estimación de máxima verosimilitud del parámetro θ.

Para determinar el estimador de máxima verosimilitud (CB-sección 5.2 ó EII-sección 2.2) lo primero que deberemos construir es la función de verosimilitud, la cual es

$$f_\theta(x_1,...,x_n) = \frac{1}{2^n}\,\theta^{3n}\,\prod_{i=1}^{n} x_i^2\,\exp\left\{-\theta\sum_{i=1}^{n} x_i\right\} \qquad \text{si} \quad x_1,...,x_n > 0$$

de donde será

$$\log f_\theta(x_1,...,x_n) = -n\log 2 + 3n\log\theta + \log\prod_{i=1}^{n} x_i^2 - \theta\sum_{i=1}^{n} x_i$$

obteniéndose de la ecuación de verosimilitud

$$\frac{\partial}{\partial\theta}\log f_\theta(x_1,...,x_n) = \frac{3n}{\theta} - \sum_{i=1}^{n} x_i = 0$$

el estimador de máxima verosimilitud para θ

$$\hat{\theta} = \frac{3n}{\sum_{i=1}^{n} X_i}.$$

De los datos del enunciado se obtiene que la estimación de máxima verosimilitud (es decir, el valor del estimador de máxima verosimilitud para la muestra observada) es

$$\hat{\theta} = \frac{3n}{\sum_{i=1}^{n} X_i} = \frac{3\cdot 5}{15+20+10+13+12} = \frac{15}{70} = 0'2143.$$

Problema 1.7

Se cree que el tiempo de vida útil de una determinada componente electrónica incluida en los ordenadores es una variable aleatoria X con función de densidad

$$f_\theta(x) = \theta^2\,x\,e^{-\theta x} \qquad\qquad x > 0$$

dependiente de un parámetro θ. Elegida una muestra aleatoria simple de X se obtuvieron los siguientes diez valores

$$1\ ,\ 1'2\ ,\ 2\ ,\ 0'9\ ,\ 2'4\ ,\ 1'7\ ,\ 2'1\ ,\ 2'5\ ,\ 1'8\ ,\ 3'4$$

Se pide:

a) **Determinar la estimación de máxima verosimilitud del parámetro θ.**

b) **Si nos dicen que en una muestra de tamaño $n = 121$ de esta variable, se obtuvo una cuasidesviación típica muestral igual a 10, ¿cuál será la probabilidad (aproximada) de que la media de la muestra y de la población difieran en más de 2 unidades?**

a) El estimador de máxima verosimilitud (CB-sección 5.2 ó EII-sección 2.2) es el valor del parámetro que hace máxima a la función de verosimilitud, que, en este caso, es igual a

$$L(\theta) = f_\theta(x_1, ..., x_n) = \prod_{i=1}^{n} f_\theta(x_i) = \prod_{i=1}^{n} \theta^2 \, e^{-\theta x_i} \, x_i = \theta^{2n} \, \exp\{-\theta \sum_{i=1}^{n} x_i\} \prod_{i=1}^{n} x_i$$

si $x_1, ..., x_n > 0$.

Como el máximo de una función y de su logaritmo se alcanzan en el mismo valor de la variable, dado que la función de verosimilitud es de tipo exponencial, nos resultará más simple determinar el máximo para el logaritmo de la función de verosimilitud,

$$\log L(\theta) = 2n \log \theta - \theta \sum_{i=1}^{n} x_i + \log \prod_{i=1}^{n} x_i$$

Por el tipo de función que tenemos que maximizar, la obtención del máximo resultará más simple si igualamos la derivada a cero en la ecuación anterior, obteniendo la ecuación de verosimilitud

$$\frac{\partial}{\partial \theta} \log L(\theta) = \frac{2n}{\theta} - \sum_{i=1}^{n} x_i = 0$$

de donde, despejando θ, obtenemos el estimador de máxima verosimilitud para θ

$$\hat{\theta} = \frac{2n}{\sum_{i=1}^{n} X_i}.$$

De los datos del enunciado se obtiene que la estimación de máxima verosimilitud (es decir, el valor del estimador de máxima verosimilitud para la muestra observada) es

$$\hat{\theta} = \frac{2n}{\sum_{i=1}^{n} X_i} = \frac{2}{\overline{x}} = \frac{2}{1'9} = 1'053.$$

b) Estamos en un caso de estimación de la media μ de una población no normal y tamaños de muestra suficientemente grandes ($n > 100$) (CB-sección 5.5 ó EII-sección 2.5) en el que la distribución de la media muestral se puede aproximar por un normal $N(0, 1)$

$$\frac{\overline{x} - \mu}{S/\sqrt{n}} \approx N(0, 1).$$

La probabilidad que nos piden es

$$P\{|\overline{x}-\mu|>2\}$$

con lo que tipificando, para obtener un suceso equivalente al anterior en donde aparezca una $Z \rightsquigarrow N(0,1)$ cuya probabilidad será fácilmente calculable utilizando las tablas de la normal estándar, será,

$$\begin{aligned} P\{|\overline{x}-\mu|>2\} &\approx P\left\{\frac{|\overline{x}-\mu|}{S/\sqrt{n}} > \frac{2}{S/\sqrt{n}}\right\} \\ &= P\left\{|Z| > \frac{2}{10/11}\right\} = 2\cdot P\left\{Z>2'2\right\} = 2\cdot 0'0139 = 0'0278. \end{aligned}$$

Problema 1.8

Estudios anteriores han demostrado que puede admitirse, en una determinada región geográfica, una distribución de Poisson de parámetro θ para el número de hembras de un insecto. Si puede admitirse que es $\theta=1$, calcular el número mínimo de veces, n, que debe de muestrearse en la región en cuestión para que la diferencia entre el número medio de hembras del insecto en la muestra y el valor supuesto para θ difieran en una o menos de una unidad, con probabilidad mayor o igual a $0'95$.
(Observación: n será pequeño.)

El enunciado nos dice que puede admitirse para la variable X =*número de hembras de un insecto*, una distribución de Poisson $\mathcal{P}(1)$ y nos pide que determinemos el menor valor de n para el que

$$P\{|\overline{x}-1|\leq 1\}\geq 0'95.$$

Es decir, que determinemos el valor de n tal que

$$P\left\{0\leq \sum_{i=1}^{n} X_i \leq 2n\right\}\geq 0'95$$

o bien

$$P\left\{\sum_{i=1}^{n} X_i > 2n\right\}\leq 0'05$$

siendo $\sum_{i=1}^{n} X_i \rightsquigarrow \mathcal{P}(n)$.

Si fuera $n=1$, de las tablas de la distribución de Poisson, se obtiene que

$$P\{W_1>2\} = 0'0613+0'0153+0'0031+0'0005+0'0001 = 0'0803 > 0'05$$

con $W_1 \rightsquigarrow \mathcal{P}(1)$, por lo que debemos aumentar el tamaño de la muestra.

Si fuera $n = 2$, sería

$$P\{W_2 > 4\} = 0'0361 + 0'0120 + 0'0034 + 0'0009 + 0'0002 = 0'0526 > 0'05$$

con $W_2 \rightsquigarrow \mathcal{P}(2)$, por lo que debemos aumentar el tamaño de la muestra.

Si fuera $n = 3$, sería

$$P\{W_3 > 6\} = 0'0216 + 0'0081 + 0'0027 + 0'0008 + 0'0002 = 0'0334 < 0'05$$

con $W_3 \rightsquigarrow \mathcal{P}(3)$, por lo que el tamaño mínimo con el que obtener la precisión deseada será $n = 3$, ya que, si fuéramos aumentando el valor de n, es decir, la cola anterior, la probabilidad cola iría disminuyendo.

Problema 1.9

Se sabe que el tiempo de supervivencia a un tipo de cáncer, en ratas de laboratorio, sigue una distribución de probabilidad dada por la siguiente función de densidad

$$f_\theta(x) = 0'05 \exp\left\{-0'05(x-\theta)\right\} \qquad , \qquad x > \theta.$$

Si los tiempos de supervivenvia de 17 ratas afectadas de la enfermedad en estudio fueron

$$188, 188, 190, 192, 206, 209, 213, 216, 220, 227, 230, 234, 246, 265, 304, 216, 244$$

determinar la estimación de máxima verosimilitud del parámetro θ.

Determinaremos primero el estimador de máxima verosimilitud del parámetro (CB-sección 5.2 ó EII-sección 2.2) calculando después el valor de éste para los valores de la muestra.

La función de verosimilitud de la muestra será

$$L(\theta) = f_\theta(x_1, ..., x_n) = \prod_{i=1}^{n} f_\theta(x_i) = 0'05^n \exp\left\{-0'05 \sum_{i=1}^{n}(x_i - \theta)\right\}$$

si $x_1, ..., x_n > \theta$.

Como siempre, el método de la máxima verosimilitud se basa en asignar a θ el valor que maximice la función $L(\theta)$; el problema es que ahora θ aparece en el recorrido de la variable, es decir, que $L(\theta)$ toma un valor distinto de cero si $\theta < x_1, ..., x_n$ y si algún x_i es tal que $x_i \leq \theta$ será $L(\theta) = 0$. En la estimación de θ habrá que tener también en cuenta, por tanto, el recorrido de $L(\theta)$.

La función