Capítulo 4

Inferencia Estadística I: Estimación Puntual.

4.1. Introducción.

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. Este proceso de inferencia puede perseguir dos objetivos diferentes:

1. *Estimación de parámetros:* utilizar los datos de la muestra para obtener valores aproximados de los parámetros que caracterizan el comportamiento de las variables de interés en la población.

2. *Contraste de hipótesis:* utilizar la información de la muestra para decidir sobre la validez o no de hipótesis relativas a alguna característica de la población.

Dado que la muestra sólo proporciona información parcial sobre la población, los métodos de inferencia estadística se apoyan en el cálculo de probabilidades para cuantificar los márgenes de error probables o para evaluar el riesgo de incurrir en decisiones incorrectas.

Obviamente el desarrollo de los procedimientos de inferencia requiere disponer de una muestra lo suficientemente representativa de la población. En este capítulo presentaremos algunos conceptos elementales sobre muestreo, para a continuación ocuparnos del problema de la estimación de parámetros: qué es un estimador, qué características debe tener y cómo se puede construir un estimador adecuado para un parámetro de interés.

Objetivos.

Al finalizar este capítulo, el alumno deberá:

1. Conocer y comprender los conceptos de población y muestra aleatoria.
2. Entender el significado de la inferencia estadística y distinguir entre inferencia paramétrica e inferencia no paramétrica.
3. Conocer y manejar el concepto de estimador puntual, así como entender el significado de las propiedades de sesgo, varianza y consistencia de un estimador
4. Conocer y ser capaz de aplicar los distintos métodos de obtención de estimadores: momentos, máxima verosimilitud y mínimos cuadrados.
5. Ser capaz de interpretar el significado de los parámetros estimados.
6. Ser capaz de valorar el grado de ajuste conseguido mediante el modelo paramétrico estimado.

4.2. Población y muestra aleatoria.

En la introducción de este capítulo hemos definido la *inferencia estadística* como 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. Conviene, por tanto, precisar el significado de los términos *población* y *muestra.*

La definición habitual de *población* es la de conjunto formado por *todos* los sujetos u objetos que comparten una o varias características comunes, y sobre los que se desea obtener información. Desde esta perspectiva podemos hablar, por ejemplo, de la población formada por todos los seres humanos que habitan la Tierra, de la población de hormigas de la isla de Gran Canaria, o de la población de delfines mulares hembra del Atántico. Esta definición, sin embargo, presenta dificultades en muchos casos: ¿cuál es la población si el objetivo de nuestro estudio es caracterizar la temperatura del magma volcánico? ¿Y si nuestro objetivo es estudiar la velocidad de una corriente marina? En otro contexto, si deseamos saber si un tratamiento médico es efectivo contra determinada enfermedad, parece lógico considerar como población el conjunto de personas susceptibles de recibir el tratamiento; pero este conjunto incluye tanto aquellos que padecen la enfermedad actualmente, como aquellos que la padecerán en el futuro y a los que podría aplicárseles el tratamiento.

Vemos, pues, que hay poblaciones tangibles (personas, delfines u hormigas), conceptuales (los estados físicos del magma o los comportamientos dinámicos de la corriente marina) e incluso hipotéticas (los sujetos que en el futuro podrían contraer una enfermedad). En cualquier caso, cuando se estudia una población, el objetivo no es, propiamente, el conjunto de sujetos, objetos u entes conceptuales que puedan formar esa población en un instante concreto, sino determinadas *características* que medimos sobre ellos, y que se traducen en *variables aleatorias,* toda vez que sus valores no son conocidos a priori. En este sentido, desde un punto de vista práctico, caracterizar una *población* es equivalente a *conocer la distribución de probabilidad* $\mathbb{P}$ *de la variable aleatoria* X que se mide sobre la misma: la temperatura del magma, la velocidad de la corriente o la variable binaria $1-0$ que indica si un paciente se cura o no.

Normalmente, la población completa no suele ser accesible (por su tamaño, por cuestiones de coste o tiempo, o simplemente porque la población es hipotética), por lo que su estudio habrá de realizarse a partir de sólo una parte de la misma. Se denomina *muestra* a un subconjunto de la población. Para que la información proporcionada por una muestra pueda emplearse aceptablemente para obtener conclusiones sobre la población es necesario:

- Que la muestra sea *representativa*, esto es, que refleje de la mejor manera posible las características de la población. Si una muestra no fuese representativa, es obvio que lo que se pueda deducir de ella no podrá extenderse a la población; en particular la estimación de parámetros en tales condiciones podría estar fuertemente sesgada y los contrastes de hipótesis podrían conducir a decisiones erróneas con mayor frecuencia de lo previsto.

- Que la muestra tenga un tamaño suficiente. En general, cuanto mayor sea el tamaño, más información proporcionará. El tamaño adecuado de la muestra depende de cuál sea el problema que nos planteamos (estimación de parámetros o contraste de hipótesis), de las características de la población (en general, a mayor heterogeneidad de la población con respecto a la variable de interés, mayor habrá de ser el tamaño de la muestra) y de la magnitud de los errores que estamos dispuestos a cometer en nuestro proceso de inferencia.

Como hemos señalado más arriba, habitualmente nuestro interés se centra en el estudio de alguna variable aleatoria X que se mide sobre la población. El comportamiento de dicha variable aleatoria X queda caracterizado por su *distribución de probabilidad* $\mathbb{P}$. En este contexto, definimos una *muestra aleatoria* de tamaño n de una distribución de probabilidad $\mathbb{P}$ como *un conjunto de variables aleatorias* $X_1, \ldots, X_n$ *independientes y con la misma distribución* $\mathbb{P}$. En la práctica, la obtención de una muestra aleatoria se traduce en seleccionar

al azar y de manera independiente n elementos de la población y medir el valor de X en cada uno de ellos. Así, si X es la velocidad de la corriente marina en un punto, $X_1, \ldots, X_n$ serían n observaciones independientes de dicha velocidad en ese punto; si X es la variable binaria $1-0$ que representa la curación (o no) de una enfermedad tras aplicar un tratamiento, $X_1, \ldots, X_n$ sería el efecto del tratamiento en un conjunto de n pacientes elegidos de manera independiente en la misma población.

Podemos preguntarnos de qué manera y hasta qué punto una muestra aleatoria $X_1, \ldots, X_n$ de observaciones de una variable aleatoria X nos informa sobre la distribución de probabilidad de X (evidentemente, si la muestra no contuviese información a este respecto, no tendría sentido el muestreo). Para responder a esta pregunta definimos la función de *distribución empírica* de la muestra como:

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

siendo $I(X_i \leq x)$ uno o cero según ocurra o no el suceso $\{X_i \leq x\}$ (por tanto, $\hat{F}_n(x)$ es la proporción de veces que en la muestra se han observado valores menores o iguales que x). El teorema de *Glivenko-Cantelli,* que enunciamos a continuación, prueba que a medida que el tamaño de muestra n se incrementa, la función de distribución empírica $\hat{F}_n(x)$ se va aproximando cada vez más a la función de distribución acumulativa $F(x)$ de la variable X.

Teorema 4.1. ***(de Glivenko-Cantelli)*** *Sea $X_1, \ldots, X_n$ una muestra aleatoria de una variable aleatoria X con función de distribución acumulativa $F(x)$, y sea $\hat{F}_n(x)$ la función de distribución empírica de la muestra. Entonces para cualquier valor x se verifica, a medida que $n \to \infty$:*

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

Demostración. Es inmediato observar que, para cada x, la variable $I(X_i \leq x)$ sigue una distribución de Bernoulli de parámetro $F(x)$, cualquiera que sea i. Por tanto, tal como vimos en el capítulo anterior, $E[I(X_i \leq x)] = F(x)$ y $\text{var}(I(X_i \leq x)) = F(x)(1 - F(x))$. Aplicando ahora las propiedades de la esperanza y la varianza de una suma de variables aleatorias independientes:

$$E\left[\hat{F}_n(x)\right] = \frac{1}{n}\sum_{i=1}^{n} E[I(X_i \leq x)] = F(x)$$

$$\text{var}\left(\hat{F}_n(x)\right) = \frac{1}{n^2}\sum_{i=1}^{n}\text{var}\left(I\left(X_i \leq x\right)\right) = \frac{1}{n}F(x)\left(1-F(x)\right)$$

Por tanto:

$$E\left[\left(\hat{F}_n(t) - F(t)\right)^2\right] = \text{var}\left(\hat{F}_n(t)\right) = \frac{1}{n}F(t)\left(1-F(t)\right) \to 0$$

cuando $n \to \infty$. □

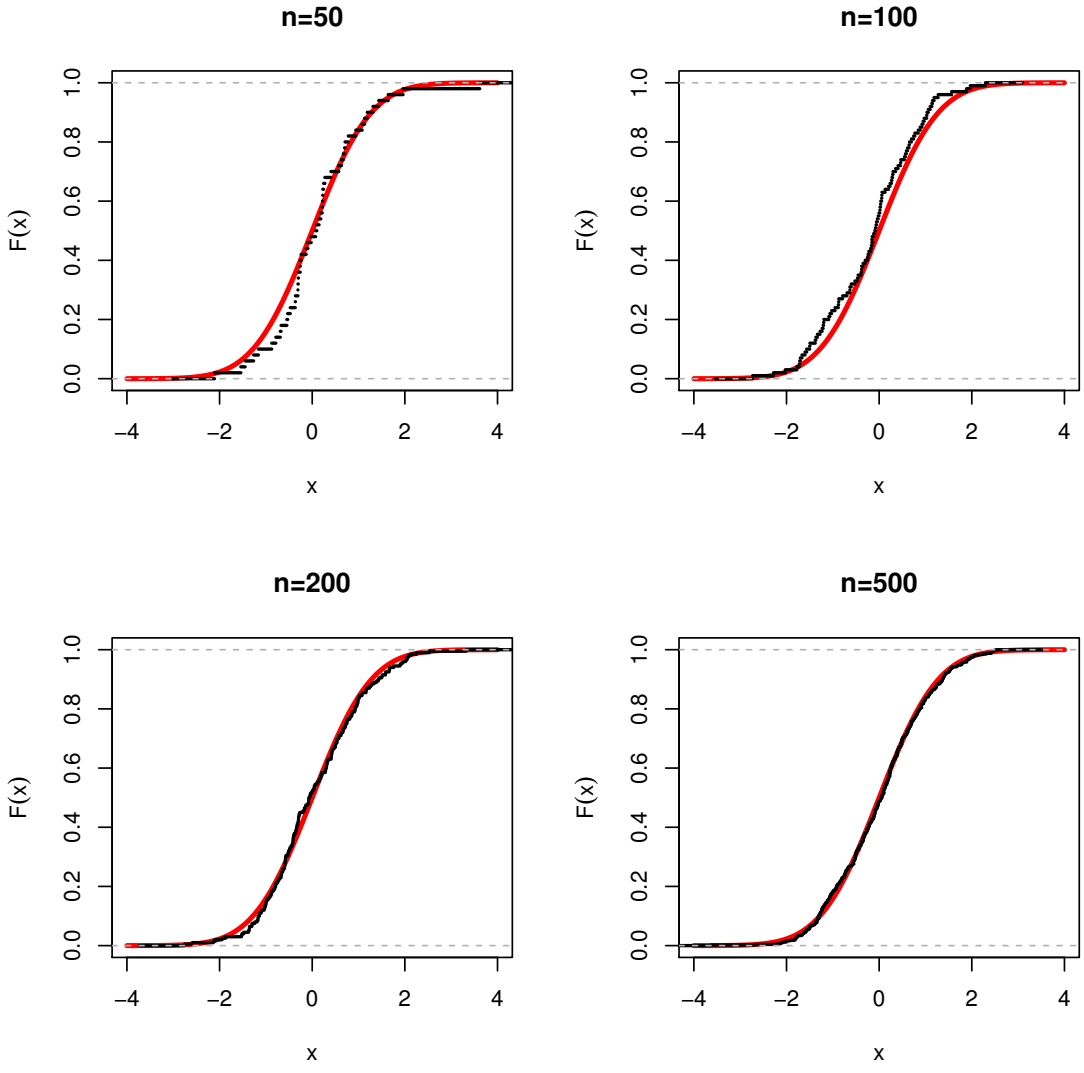

Figura 4.1: Efecto del Teorema de Glivenko-Cantelli: a medida que aumenta el tamaño de la muestra, la función de distribución empírica de la muestra, $\hat{F}_n(x)$, se aproxima cada vez más a la función de distribución acumulativa teórica $F(x)$ de la variable aleatoria.

Así pues, el teorema de Glivenko-Cantelli garantiza que el muestreo aleatorio produce muestras representativas de la variable de interés que, con el tamaño adecuado, permiten aproximar razonablemente la función de distribución acumulativa de dicha variable. Por esta razón este teorema suele conocerse también como **teorema fundamental de la estadística.**

En la figura 4.1 se muestran superpuestas la función de distribución acumulativa de la distribución normal de parámetros $\mu = 0$ y $\sigma = 1$ y la distribución empírica obtenida para muestras aleatorias de tamaños respectivos 50, 100, 200 y 500. Puede apreciarse que a medida que aumenta el tamaño muestral, la función empírica tiende a confundirse con la teórica.

4.3. Inferencia paramétrica vs. inferencia no paramétrica.

Como sabemos, el comportamiento de una variable aleatoria X queda caracterizado mediante su función de distribución acumulativa $F(x)$. Cuando el investigador toma una muestra aleatoria $X_1, X_2, \ldots, X_n$ de esta variable, puede encontrarse en alguno de los siguientes escenarios:

1. Conoce la expresión funcional de $F(x)$, pero no conoce los valores de los parámetros que la caracterizan, y que denotaremos por $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Esto es lo que sucede, por ejemplo, si se sabe (o se sospecha) que los datos proceden de una distribución exponencial (de la que no se conoce el valor del parámetro λ), de una Weibull (de la que no se sabe lo que valen κ y η), de una Normal (de la que no se conocen μ y σ), ...
2. No sabe nada de $F(x)$ salvo, quizás, si es continua o escalonada.

El primer escenario corresponde a la así llamada *inferencia paramétrica*. Cualquier afirmación, en términos de probabilidad, sobre las características de la variable X requiere obtener alguna aproximación del valor del parámetro Θ, proceso que se conoce con el nombre de *estimación*. El segundo escenario corresponde a un problema de *inferencia no paramétrica*. Como veremos, en el primer caso los contrastes de hipótesis se establecen en términos de Θ; en el segundo caso se establecen en términos de características más generales usualmente relacionadas con la forma de $F(x)$.

Señalemos por último que, dado que en la práctica una de las situaciones más habituales es asumir que $F(x)$ corresponde a la distribución normal, es habitual denominar inferencia paramétrica a la inferencia basada en dicha distribución.

4.4. Estimación.

En el capítulo anterior hemos visto una colección de distribuciones de probabilidad que permiten modelar el comportamiento de numerosas variables aleatorias que aparecen en las aplicaciones prácticas: el peso o la longitud de un pez de determinada especie, la altura de ola en una zona costera, el número de nidos de tortuga en una playa, el tiempo entre ocurrencias de un fenómeno meteorológico, etc. Este proceso de modelización requiere ajustar de algún modo los parámetros característicos de la distribución de probabilidad a emplear. Así, por ejemplo, si modelamos la longitud de los peces de una especie mediante una distribución normal, ¿cuáles son los valores de μ y σ adecuados?; si modelamos la altura de ola mediante una distribución de Weibull, ¿cuáles son los valores de los parámetros de localización y escala?; si se modela el número de nidos de tortuga en una playa mediante la distribución de Poisson, ¿cuál es el valor de λ?.

La obtención del valor aproximado de un parámetro se denomina *estimación.* La estimación es *puntual* si proporciona un único valor aproximado para dicho parámetro; es *por intervalo* si proporciona un intervalo que, con cierta confianza, contiene al parámetro.

4.4.1. Definiciones básicas

Estadístico: Dada una muestra aleatoria $X_1, X_2, \ldots, X_n$ se llama *estadístico* a cualquier función de sus valores.

Estimador: Dado un parámetro θ característico de una población, y una muestra aleatoria $X_1, X_2, \ldots, X_n$ de la misma, se llama *estimador* de ϑ a cualquier estadístico $\hat{\theta} = \hat{\theta}\,(X_1, X_2, ..., X_n)$ cuyos valores se aproximen a θ.

Si bien los estimadores muchas veces pueden construirse de forma natural –estimar la esperanza de una variable mediante la media de una muestra aleatoria de la misma, estimar una proporción poblacional mediante la proporción equivalente en la muestra– existen diversos métodos, que veremos en la sección 4.4.3, que permiten construir estimadores en casos más generales, y además con buenas propiedades.

Nótese de la definición anterior que *un estimador es una variable aleatoria*: no puede predecirse su valor mientras no se haya obtenido la muestra. Por tanto, un estimador habrá de caracterizarse en términos de una distribución de probabilidad sobre sus posibles valores.

Como distintas muestras producirán distintos valores del estimador $\hat{\theta}$, es de esperar que algunos de estos valores estén más próximos al valor de θ y otros estén más alejados. Por tanto ¿cuando podemos considerar que $\hat{\theta} = \hat{\theta}\,(X_1, X_2, ..., X_n)$ produce valores próximos a θ?

Como veremos a continuación, la respuesta a esta pregunta está estrechamente relacionada con la distribución de probabilidad de $\hat{\theta}$.

4.4.2. Propiedades deseables de un estimador.

4.4.2.1. Exactitud:

Dado que el estimador puede tomar muchos valores diferentes (según cual sea la muestra que se obtenga), una manera de medir la proximidad entre el estimador y el parámetro es mediante la distancia entre el valor esperado del estimador y el valor del parámetro. Dicha distancia recibe el nombre de *sesgo* del estimador:

$$\text{Sesgo}(\hat{\theta}) = E[\hat{\theta}] - \theta$$

Cuando el sesgo del estimador es cero (en cuyo caso $E[\hat{\theta}] = \theta$), el estimador es *exacto* (también se le suele llamar *insesgado* o *centrado*). En caso contrario el estimador es *sesgado*. En general resulta deseable que un estimador sea insesgado. Un sesgo positivo en el estimador significa que sus valores, en media, están por encima del parámetro que pretende estimar y por tanto tiende a sobreestimarlo. De modo similar, los estimadores con sesgo negativo tienden a subestimar el parámetro.

Ejemplo 4.1. La media muestral es un estimador centrado de la media poblacional. En efecto:

$$E\left[\bar{X}\right] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} X_i\right] = \frac{1}{n}\sum_{i=1}^{n} E\left[X_i\right] = \frac{1}{n}n\mu = \mu$$

Ejemplo 4.2. La varianza muestral es un estimador sesgado de la varianza poblacional. En efecto, la varianza muestral se define como:

$$S^2 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

Para calcular la esperanza de S^2 observemos en primer lugar que:

$$\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2=\sum_{i=1}^{n}\left(X_i-\mu+\mu-\bar{X}\right)^2=\sum_{i=1}^{n}\left((X_i-\mu)-\left(\bar{X}-\mu\right)\right)^2=$$
$$=\sum_{i=1}^{n}\left((X_i-\mu)^2-2\left(X_i-\mu\right)\left(\bar{X}-\mu\right)+\left(\bar{X}-\mu\right)^2\right)=$$
$$=\sum_{i=1}^{n}\left(X_i-\mu\right)^2-2\left(\bar{X}-\mu\right)\sum_{i=1}^{n}\left(X_i-\mu\right)+\sum_{i=1}^{n}\left(\bar{X}-\mu\right)^2=$$

$$=\sum_{i=1}^{n}\left(X_i-\mu\right)^2-2\left(\bar{X}-\mu\right)n\left(\bar{X}-\mu\right)+n\left(\bar{X}-\mu\right)^2=$$
$$=\sum_{i=1}^{n}\left(X_i-\mu\right)^2-n\left(\bar{X}-\mu\right)^2$$

Se tiene:

$$E\left[\sum_{i=1}^{n}\left(X_i-\mu\right)^2\right]=\sum_{i=1}^{n}E\left[\left(X_i-\mu\right)^2\right]=n\sigma^2$$

Por ser las X_i independientes:

$$E\left[\left(\bar{X}-\mu\right)^2\right]=var\left(\overline{X}\right)=var\left(\frac{1}{n}\sum_{i=1}^{n}X_i\right)=\frac{1}{n^2}\,var\left(\sum_{i=1}^{n}X_i\right)=$$
$$=\frac{1}{n^2}\sum_{i=1}^{n}var\left(X_i\right)=\frac{1}{n^2}n\sigma^2=\frac{\sigma^2}{n}$$

Por tanto:

$$E\left[S^2\right]=\frac{1}{n}E\left[\sum_{i=1}^{n}\left(X_i-\mu\right)^2-n\left(\bar{X}-\mu\right)^2\right]=\frac{1}{n}\left(n\sigma^2-\sigma^2\right)=\frac{n-1}{n}\sigma^2$$

Así pues:

$$\text{Sesgo}(S^2)=E\left[S^2\right]-\sigma^2=\frac{n-1}{n}\sigma^2-\sigma^2=-\frac{1}{n}\sigma^2$$

de donde se sigue que la varianza muestral subestima la varianza poblacional (si bien es cierto que a medida que el tamaño de la muestra n aumenta, el sesgo se hace más pequeño).

Ejemplo 4.3. La cuasivarianza muestral, definida como

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2$$

sí es un estimador centrado de la varianza poblacional. En efecto:

$$E\left[s^2\right] = E\left[\frac{1}{n-1}\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] = \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i - \bar{X}\right)^2\right] =$$
$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left(X_i - \mu\right)^2 - n\left(\bar{X} - \mu\right)^2\right] = \frac{1}{n-1}\left[n\sigma^2 - \sigma^2\right] = \sigma^2$$

Por esta razón, como estimador de la varianza poblacional, en la práctica se prefiere la cuasivarianza muestral.

Ejemplo 4.4. Si X es una variable aleatoria de Bernoulli de parámetro p, la proporción muestral de éxitos $\hat{p}$ es un estimador insesgado de la proporción poblacional p. En efecto, la proporción muestral de éxitos al observar una muestra aleatoria de tamaño n es:

$$\hat{p} = \frac{\text{Número de éxitos}}{\text{Número de Observaciones}} = \frac{N_E}{n}$$

Como X es de Bernoulli, el número N_E de éxitos en n pruebas independientes sigue una distribución $B(n,p)$, y por tanto:

$$E\left[\hat{p}\right] = E\left[\frac{N_E}{n}\right] = \frac{1}{n}E\left[N_E\right] = \frac{1}{n}n\cdot p = p$$

4.4.2.2. Precisión.

Tal como hemos visto, un estimador es una variable aleatoria cuyo valor cambia con la muestra. Si el estimador es centrado, ello indica que el centro de la distribución de valores del estimador coincide con el parámetro que se pretende estimar. Si embargo esto no nos informa de si dicha distribución tiene mucha o poca dispersión en torno al parámetro. Si la dispersión es grande, significa que habrá muestras que darán lugar a estimaciones muy alejadas del valor del parámetro. Si la dispersión es pequeña, aún en la peor de las muestras posibles, la estimación obtenida estará próxima al valor del parámetro. Por tanto, si se dispone de

varios estimadores centrados del mismo parámetro, será preferible (producirá estimaciones más precisas del parámetro) aquél que tenga la menor dispersión. Dado que la dispersión se mide mediante la varianza del estimador[1], el mejor estimador centrado será el de menor varianza (en caso de existir).

La desviación típica del estimador recibe el nombre de *error estándar*. Se suele denotar como

$$\sigma_{\hat{\theta}} = \sqrt{var\left(\hat{\theta}\right)}$$

Puede demostrarse que la media muestral, la cuasivarianza muestral y la proporción muestral son estimadores insesgados y de mínima varianza de sus parámetros respectivos.

4.4.2.3. Menor Error Cuadrático Medio.

Se define el *error cuadrático medio* (ECM) de un estimador $\hat{\theta}$ para un parámetro θ, como:

$$ECM\left[\hat{\theta}\right] = E\left[\left(\hat{\theta} - \theta\right)^2\right] = \left(\text{Sesgo}\left(\hat{\theta}\right)\right)^2 + var\left(\hat{\theta}\right)$$

El ECM constituye una medida conjunta (de hecho es la suma) del sesgo y la varianza de un estimador. Es deseable que el error cuadrático medio de un estimador sea pequeño. El ECM es una medida que resulta útil cuando se debe elegir entre varios estimadores del mismo parámetro con características muy diferentes de sesgo y varianza. Así por ejemplo, puede ser más útil un estimador ligeramente sesgado pero con muy poca varianza (tal que, aunque sesgadas, todas las estimaciones están próximas al parámetro), que uno centrado pero con varianza mucho mayor (que puede dar lugar a muchas estimaciones muy alejadas del parámetro).

4.4.2.4. Consistencia de un estimador.

Un estimador $\hat{\theta}$ de un parámetro θ es consistente si verifica que:

$$\lim_{n\to\infty} P\left(\left|\hat{\theta} - \theta\right| \leq \varepsilon\right) = 1 \quad \forall \varepsilon > 0$$

[1] O de manera equivalente, mediante la desviación típica. La desviación típica de un estimador recibe el nombre de *error estándar*.

lo que significa que 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. En general es deseable que los estimadores que utilicemos sean consistentes.

Puede demostrarse que la media muestral, la varianza muestral y la proporción muestral son estimadores consistentes de sus parámetros respectivos. Por ejemplo, para probar que la media muestral es un estimador consistente de la media poblacional basta tener en cuenta que $E\left[\overline{X}\right] = \mu$ y $var\left(\overline{X}\right) = \frac{\sigma^2}{n}$. De acuerdo con el teorema de Chebyshev, para cualquier valor de $k \geq 1$ se tiene:

$$P\left(\left|\overline{X} - \mu\right| > k\frac{\sigma}{\sqrt{n}}\right) \leq \frac{1}{k^2}$$

Eligiendo entonces $\varepsilon = k\frac{\sigma}{\sqrt{n}}$ (esto es, $k = \frac{\varepsilon\sqrt{n}}{\sigma}$) se tiene que

$$P\left(\left|\overline{X} - \mu\right| > \varepsilon\right) \leq \frac{1}{n}\left(\frac{\sigma}{\varepsilon}\right)^2$$

por lo que cuando $n \to \infty$ resulta $P\left(\left|\overline{X} - \mu\right| > \varepsilon\right) \to 0$, o lo que es lo mismo

$$P\left(\left|\overline{X} - \mu\right| \leq \varepsilon\right) \to 1$$

lo que prueba que la media muestral $\overline{X}$ es un estimador consistente de la media poblacional μ. Ello además vuelve a justificar, como ya hemos visto anteriormente, que el concepto de esperanza de una variable aleatoria puede identificarse con el de media aritmética para grandes valores de n.

4.4.3. Métodos de obtención de estimadores puntuales.

En esta sección abordamos el problema de cómo pueden obtenerse funciones cuyos valores se aproximen al de un parámetro desconocido de cierta distribución de probabilidad. Tres son los métodos que se emplean habitualmente para ello: el método de los momentos, el método de máxima verosimilitud y el método de los mínimos cuadrados.

4.4.3.1. Método de los momentos.

Recordemos que dada una variable aleatoria X, se define el momento de orden k respecto al origen como:

$$\mu_k = E\left[X^k\right] = \begin{cases} \sum\limits_{x_i \in E} x_i^k P\left(X = x_i\right) & \text{si } X\text{es discreta} \\ \int_{-\infty}^{\infty} x^k f\left(x\right) dx & \text{si } X\text{es continua} \end{cases}$$

Ya hemos visto en varias ocasiones que $\mu = \mu_1$ y $\sigma^2 = \mu_2 - \mu_1^2$. De la misma forma que la esperanza y la varianza se pueden poner en función de los momentos, en general si una variable aleatoria X depende de unos parámetros desconocidos $\theta_1, \theta_2, \ldots, \theta_k$, muchas veces será posible expresar estos parámetros como funciones de algunos momentos de la variable, esto es, $\theta_j = g_j(\mu_1, \mu_2, ...)$, $j = 1, 2, ..., k$. El método de los momentos consiste en determinar estas funciones, estimar los momentos correspondientes mediante sus análogos muestrales:

$$\hat{\mu}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \hat{\mu}_2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2, \ ..., \ \hat{\mu}_k = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

y por último estimar los θ_j, mediante las funciones anteriores evaluadas en los momentos muestrales: $\hat{\theta}_j = g_j(\hat{\mu}_1, \hat{\mu}_2, ...)$, $j = 1, 2, ..., k$

Este método tiene su fundamento en el hecho de que los momentos muestrales son estimadores insesgados de los momentos poblacionales. Asimismo ya hemos visto que si se toma una muestra aleatoria, a medida que aumenta su tamaño su distribución empírica se va pareciendo cada vez más a la distribución de probabilidad de la variable observada. Intuitivamente ello nos indica que los momentos muestrales se van a ir pareciendo cada vez más a los poblacionales a medida que aumenta el tamaño de la muestra.

Ejemplo 4.5. Supongamos que se desea estimar el parámetro p de una variable Bernoulli $b(p)$. Sabemos que

$$E[X] = p$$

Por lo que p puede expresarse en términos de los momentos simplemente como

$$p = E[X] = \mu_1$$

Para estimar p, simplemente sustituimos μ_1 en esta ecuación por su estimador $\hat{\mu}_1 = \overline{X}$ con lo que como estimador de p se obtiene:

$$\hat{p} = \hat{\mu}_1 = \overline{X}$$

Nótese que al ser $X \approx b(p)$, la variable X sólo toma los valores 1 (éxito) o 0 (fracaso), por lo que la media aritmética de n observaciones de X es:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{\text{Nº de éxitos en } n\text{pruebas}}{n}$$

esto es, la proporción de éxitos en la muestra.

Ejemplo 4.6. Se desea estimar el parámetro p de una variable $Geo(p)$. En este caso, sabemos que:

$$\mu_1 = E[X] = \frac{1-p}{p}$$

De aquí despejamos p:

$$p\mu_1 = 1 - p \Rightarrow p(1+\mu_1) = 1 \Rightarrow p = \frac{1}{1+\mu_1}$$

El estimador por el método de los momentos se obtiene sustituyendo el momento poblacional por el correspondiente momento muestral. Por tanto:

$$\hat{p} = \frac{1}{1+\hat{\mu}_1} = \frac{1}{1+\overline{X}}$$

Ejemplo 4.7. Se desea estimar el número de ardillas N que hay en un bosque. Para ello se capturan inicialmente N_M ardillas, que son marcadas y devueltas al bosque. A continuación y durante n días se procede del modo siguiente: se recorre el bosque durante un periodo de tiempo fijo y se van contando las ardillas que se avistan hasta encontrar una ardilla marcada. Sea X_i el número de ardillas no marcadas que se han avistado el día i. Para estimar N por el método de los momentos basta observar que $X_i \approx Geo(p)$ siendo $p = \frac{N_M}{N}$. Por tanto

$$N = \frac{N_M}{p}$$

En el ejemplo anterior ya hemos visto que el estimador de p es $\hat{p} = \frac{1}{1+\overline{X}}$. Por tanto el estimador del número de ardillas en el bosque será:

$$\hat{N} = \frac{N_M}{\hat{p}} = N_M\left(1+\overline{X}\right)$$

siendo $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.

Ejemplo 4.8. Si $X \approx N(\mu, \sigma)$ y se desea estimar μ y σ por el método de los momentos, basta observar que como:

$$\mu = E[X] = \mu_1, \quad \sigma^2 = E\left[X^2\right] - (E[X])^2 = \mu_2 - \mu_1^2$$

los estimadores serán:

$$\hat{\mu} = \hat{\mu}_1 = \overline{X}$$
$$\hat{\sigma}^2 = \hat{\mu}_2 - \hat{\mu}_1^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - \left(\overline{X}\right)^2$$

Ejemplo 4.9. Si $X \approx \mathcal{G}(\kappa, \eta)$, para estimar los parámetros κ y η por el método de los momentos, recordemos que

$$\mu = \kappa \cdot \eta, \quad \sigma^2 = \kappa \cdot \eta^2$$

Teniendo en cuenta que $\mu_1 = \mu$ y $\sigma^2 = \mu_2 - \mu_1^2$, resulta:

$$\kappa \cdot \eta = \mu_1$$
$$\kappa \cdot \eta^2 = \mu_2 - \mu_1^2$$

Para expresar κ y η en función de los momentos poblacionales, dividimos el segundo término entre el primero y obtenemos:

$$\eta = \frac{\mu_2}{\mu_1} - \mu_1$$

Sustituimos este valor en el primer término y despejamos κ:

$$\kappa = \frac{\mu_1}{\eta} = \frac{\mu_1^2}{\mu_2 - \mu_1^2}$$

Los estimadores por el método de los momentos se obtienen entonces sustituyendo en estas expresiones los momentos poblacionales por los muestrales:

$$\hat{\eta} = \frac{1}{n\overline{X}}\sum_{i=i}^{n} X_i^2 - \overline{X}$$
$$\hat{\kappa} = \frac{\overline{X}^2}{\frac{1}{n}\sum_{i=i}^{n} X_i^2 - \overline{X}^2}$$

Ejemplo 4.10. Si $X \approx W(\kappa, \eta)$, para estimar κ y η por el método de los momentos, al igual que en el caso anterior bastará con tener en cuenta que su esperanza y varianza son:

$$\mu = \eta \cdot \Gamma\left(1 + \frac{1}{\kappa}\right), \quad \sigma^2 = \eta^2\left[\Gamma\left(1 + \frac{2}{\kappa}\right) - \left(\Gamma\left(1 + \frac{1}{\kappa}\right)\right)^2\right]$$

y por tanto:

$$\eta \cdot \Gamma\left(1+\frac{1}{\kappa}\right) = \mu_1$$
$$\eta^2\left[\Gamma\left(1+\frac{2}{\kappa}\right) - \left(\Gamma\left(1+\frac{1}{\kappa}\right)\right)^2\right] = \mu_2 - \mu_1^2$$

Si dividimos el segundo término por el cuadrado del primero nos queda una ecuación en κ :

$$\frac{\Gamma\left(1+\frac{2}{\kappa}\right)}{\left[\Gamma\left(1+\frac{1}{\kappa}\right)\right]^2} = \frac{\mu_2}{\mu_1^2}$$

El estimador de κ se obtiene resolviendo esta ecuación sustituyendo μ_1 y μ_2 por los correspondientes momentos muestrales:

$$\frac{\Gamma\left(1+\frac{2}{\hat{\kappa}}\right)}{\left[\Gamma\left(1+\frac{1}{\hat{\kappa}}\right)\right]^2} = \frac{\frac{1}{n}\sum_{i=i}^{n} X_i^2}{\left(\overline{X}\right)^2} \tag{4.1}$$

Obviamente no es posible despejar de aquí el valor de $\hat{\kappa}$ explícitamente, pero es posible construir un algoritmo numérico que resuelva el problema. Una vez obtenido $\hat{\kappa}$, el valor de $\hat{\eta}$ se obtiene de la ecuación $\eta \cdot \Gamma\left(1+\frac{1}{\kappa}\right) = \mu_1$ mediante:

$$\hat{\eta} = \frac{\overline{X}}{\Gamma\left(1+\frac{1}{\hat{\kappa}}\right)} \tag{4.2}$$

Utilización de R para estimar los parámetros de la distribución de Weibull por el método de los momentos. Veamos como podemos utilizar R para resolver numéricamente la ecuación 4.1 y así obtener $\hat{\kappa}$ y $\hat{\eta}$. Para ello supongamos que se desea ajustar una distribución de Weibull a la siguiente muestra de alturas de ola, correspondiente a 30 olas elegidas al azar entre las registradas en una escollera durante un periodo de marea alta:

```
olas = c(2.1, 2.82, 4.2, 6.34, 2.4, 3.1, 2.15, 2.73, 3.12, 2.41, 4.59, 2.81, 2.61,
    3.81, 3.13, 3.06, 5.85, 3.57, 2.64, 4.08, 3.38, 1.88, 1.94, 3.24, 1.98, 3.29,
    0.21, 2.68, 1.74, 4.25)
```

La figura 4.2 muestra el histograma correspondiente a estos datos.

En primer lugar observemos que a partir de la ecuación 4.1, si llamamos:

$$h\left(\hat{\kappa}\right) = \frac{\Gamma\left(1+\frac{2}{\hat{\kappa}}\right)}{\left[\Gamma\left(1+\frac{1}{\hat{\kappa}}\right)\right]^2} - \frac{\frac{1}{n}\sum_{i=i}^{n} X_i^2}{\left(\overline{X}\right)^2}$$

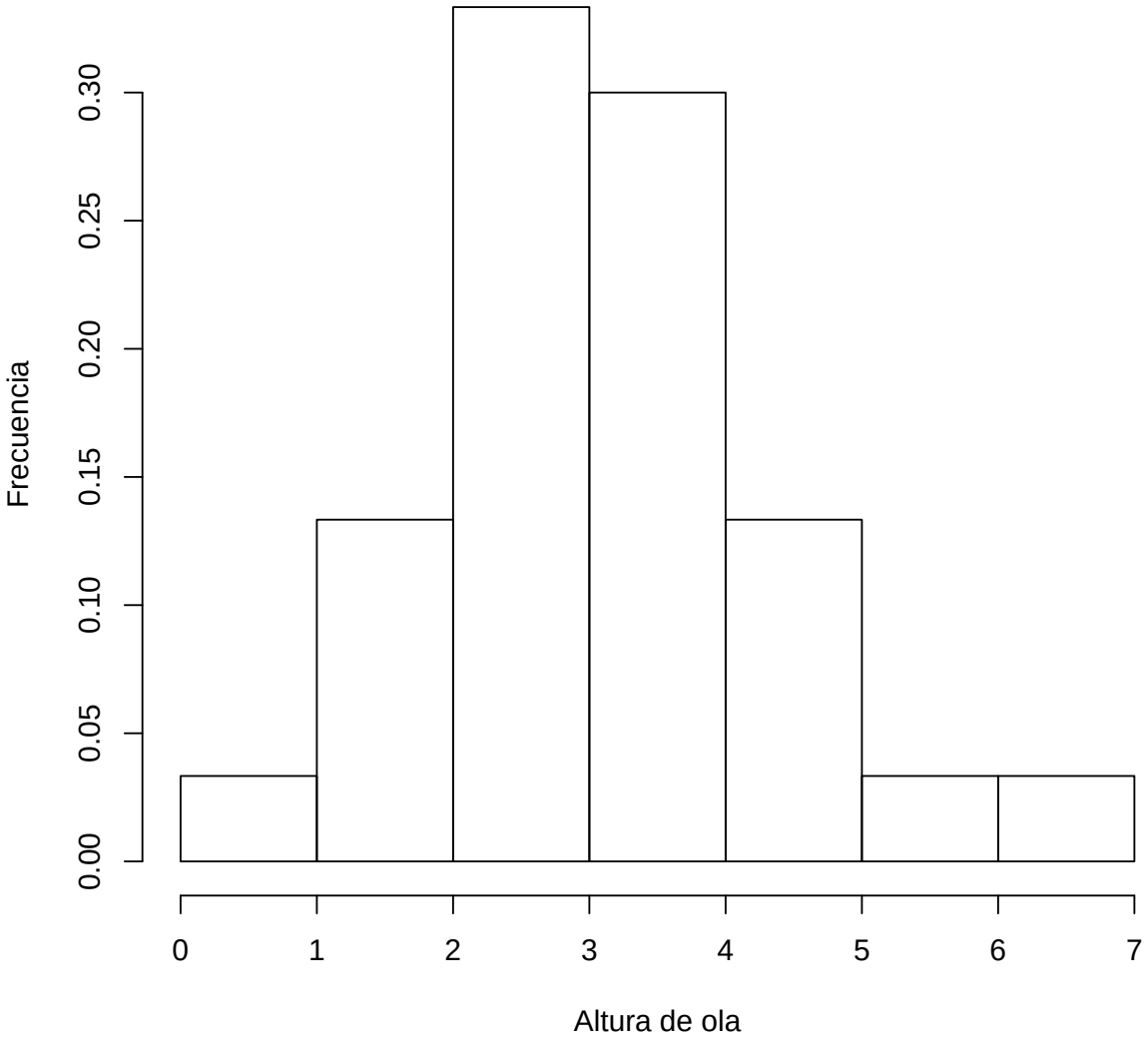

Figura 4.2: Histograma de alturas de ola registradas durante la marea alta en una escollera.

entonces el estimador por el método de los momentos de κ es el valor $\hat{\kappa}$ tal que $h(\hat{\kappa}) = 0$. Por tanto $\hat{\kappa}$ es una raiz de la función h, que puede obtenerse utilizando R mediante la función `uniroot()` que ejecuta un algoritmo de bisección. Ello significa que si proporcionamos un intervalo $[a, b]$ tal que $\text{signo}(h(a)) \neq \text{signo}(h(b))$, `uniroot()` es capaz de encontrar el punto dentro de ese intervalo en el que la función h se anula. Para ello, en primer lugar implementamos la función $h(k)$:

```
h = function(k, x) {
    n = length(x)
    m2 = sum(x^2)/n
    m1 = mean(x)
    return(gamma(1 + 2/k)/gamma(1 + 1/k)^2 - m2/m1^2)
}
```

Nótese que hemos hecho depender la función h no sólo de κ, sino también de la muestra `x` (aquí `x` es un vector que contiene todos los valores de la muestra). Ello permite que dentro de esta función se puedan calcular los momentos de la muestra, necesarios para obtener $h\left(\hat{\kappa}\right)$. Comprobamos que esta función cambia de signo en los extremos del intervalo $[1, 10]$:

```
h(1, olas)

## [1] 0.849

h(10, olas)

## [1] -0.1365
```

lo que indica que esta función tiene una raiz en dicho intervalo. Para obtener esta raiz utilizamos la función *uniroot()*, que nos proporciona el estimador $\hat{\kappa}$ buscado:

```
kappa = uniroot(h, interval = c(1, 10), x = olas)$root
kappa

## [1] 2.785
```

Por último sustituimos este valor en la ecuación 4.2, lo que nos permite obtener $\hat{\eta}$:

```
eta = mean(olas)/gamma(1 + 1/kappa)
eta

## [1] 3.449
```

4.4.3.2. Método de la máxima verosimilitud.

Sea X una variable aleatoria cuya distribución de probabilidad depende uno o varios parámetros desconocidos $\theta_1, \theta_2, \ldots, \theta_k$, y sea $f_\Theta(x)$ su función de probabilidad o de densidad (según que X sea discreta o continua), siendo $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Se desea estimar Θ, y supongamos que para ello disponemos de una muestra aleatoria $(X_1, X_2, \ldots, X_n)$ que ha producido los valores $(x_1, x_2, \ldots, x_n)$. El *método de la máxima verosimilitud* consiste en tomar como estimador de Θ aquel valor que asigna mayor probabilidad al conjunto de valores observado. La idea detrás de este método es que si la muestra aleatoria ha producido los valores $(x_1, x_2, \ldots, x_n)$ es porque debía ser *muy probable* que estos valores se observasen; por tanto los valores que resultan *verosímiles* para Θ son aquellos que hacen que sea muy probable observar $(x_1, x_2, \ldots, x_n)$; y el *más verosímil* es el que maximiza dicha probabilidad.

De un modo más formal, se define la *función de verosimilitud* como:

$$\begin{aligned} L(\Theta) =& L((\theta_1, \theta_2, \ldots, \theta_k) | x_1, x_2, \ldots, x_n) = \\ =& f(x_1, x_2, \ldots, x_n | \Theta = (\theta_1, \theta_2, \ldots, \theta_k)) = f_\Theta(x_1, x_2, \ldots, x_n) \end{aligned}$$

Esta función representa la probabilidad (o densidad) conjunta de las variables $X_1, X_2, \ldots, X_n$ en el punto $(x_1, x_2, \ldots, x_n)$ cuando el valor del parámetro es $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$. Como $(X_1, X_2, \ldots, X_n)$ una muestra aleatoria, ello significa que las X_i son independientes y con la misma distribución y por tanto su función de probabilidad (o densidad) conjunta es el producto de las funciones de probabilidad (o densidad) de cada variable. Por tanto:

- Si $X_1, X_2, \ldots, X_n$ son variables discretas :

$$L(\Theta) = f_\Theta(x_1, x_2, \ldots, x_n) = P_\Theta(X_1 = x_1) P_\Theta(X_2 = x_2) \cdots P_\Theta(X_n = x_n)$$

  siendo P_Θ la función de probabilidad de las X_i.

- Si $X_1, X_2, \ldots, X_n$ son variables continuas :

$$L(\Theta) = f_\Theta(x_1, x_2, \ldots, x_n) = f_\Theta(x_1) f_\Theta(x_2) \cdots f_\Theta(x_n)$$

  siendo $f_\Theta(x)$ la función de densidad de las X_i.

El *estimador de máxima verosimilitud (estimador MV)* es entonces el valor del parámetro $\Theta = (\theta_1, \theta_2, \ldots, \theta_k)$ que maximiza esta función:

$$\hat{\Theta} = \arg\max L(\Theta)$$

Este valor puede obtenerse la mayoría de las veces derivando $L(\Theta)$ respecto a cada θ_i, igualando a cero y despejando las θ_i:

$$\frac{\partial}{\partial \theta_i} L(\theta_1, \theta_2, \ldots, \theta_k) = 0,\ i = 1, 2, \ldots, k$$

Notemos que como $L(\Theta)$ es un producto de n términos que dependen de Θ, la obtención de su derivada es en general un proceso complicado (recuérdese como se calcula la derivada de un producto). Por ello, para obtener el máximo de $L(\Theta)$ suele utilizarse en su lugar la log-verosimilitud:

$$\ell(\Theta) = \log(L(\Theta)) = \begin{cases} \sum_{i=1}^{n} \log(P_\Theta(X_i = x_i)) & \text{si las } X_i\text{son discretas.} \\ \sum_{i=1}^{n} \log(f_\Theta(x_i)) & \text{si las } X_i\text{son continuas.} \end{cases}$$

Por ser el logaritmo una función monótona, el máximo de $L(\Theta)$ coincide con el máximo de su logaritmo $\ell(\Theta)$, esto es,

$$\hat{\Theta} = \arg\max\ L(\Theta) = \arg\max \ell(\Theta)$$

siendo la derivada de $\ell(\Theta)$ mucho más sencilla de calcular (ya que la derivada de una suma de términos es simplemente la suma de las derivadas). Por tanto, en la práctica los estimadores de máxima verosimilitud se obtendrán en la mayoría de las ocasiones resolviendo:

$$\frac{\partial}{\partial \theta_i} \ell(\theta_1, \theta_2, \ldots, \theta_k) = 0,\ i = 1, 2, \ldots, k$$

Propiedades de los estimadores de máxima verosimilitud.

Los estimadores de máxima verosimilitud son preferibles a los estimadores obtenidos por el método de los momentos (en algunos casos los estimadores obtenidos por ambos 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:* a medida que aumenta el tamaño de muestra, los estimadores MV tienen el menor error cuadrático medio de entre los estimadores posibles.
- *Normalidad asintótica:* a medida que aumenta el tamaño de la muestra, los estimadores MV tienden a tener distribución normal.

Ejemplo 4.11. Supongamos que $X \approx exp\left(\frac{1}{\theta}\right)$. En este caso

$$f_\theta(x) = \frac{1}{\theta} e^{-\frac{1}{\theta}x},\ x \geq 0$$

Dada una muestra $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$ de esta variable, la función de verosimilitud es:

$$L(\theta) = f_\theta(x_1) \cdot f_\theta(x_2) \cdot \ldots \cdot f_\theta(x_n) = \frac{1}{\theta} e^{-\frac{x_1}{\theta}} \cdot \frac{1}{\theta} e^{-\frac{x_2}{\theta}} \cdot \ldots \cdot \frac{1}{\theta} e^{-\frac{x_n}{\theta}} = \left(\frac{1}{\theta}\right)^n e^{-\frac{1}{\theta}(\sum x_i)}$$

Calculando su logaritmo obtenemos la log-verosimilitud:

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

Derivamos e igualamos a 0:

$$\ell'(\theta) = 0 \Rightarrow -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i = 0 \Rightarrow \hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i = \overline{x}$$

(en el último paso le hemos añadido el símbolo ˆ a θ para indicar que es un estimador). Podemos confirmar que es un máximo hallando la derivada segunda $\ell''(\theta)$ y comprobando que $\ell''(\overline{x}) < 0$.

Ejemplo 4.12. 5. Supongamos que se desea estimar el parámetro p de una variable de Bernoulli, $X \approx Be(p)$ por el método de máxima verosimilitud. Si se ha observado la muestra $X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n$, (donde los x_i son 1 ó 0 según que se obtenga éxito o fracaso), la función de verosimilitud asociada es:

$$\begin{aligned} L(p) &= P(X_1 = x_1) P(X_2 = x_2) \ldots P(X_n = x_n) = \\ &= p^{x_1}(1-p)^{1-x_1} p^{x_2}(1-p)^{1-x_2} \ldots p^{x_n}(1-p)^{1-x_n} = p^{\sum_{i=1}^n x_i}(1-p)^{n-\sum_{i=1}^n x_i} \end{aligned}$$

La log-verosimilitud será entonces:

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

Derivamos respecto a p e igualamos a 0:

$$\frac{\partial}{\partial p}\ell\left(p\right)=\left(\sum_{i=1}^{n}x_i\right)\frac{1}{p}-\left(n-\sum_{i=1}^{n}x_i\right)\frac{1}{1-p}=0$$

Despejamos p:

$$\begin{aligned}\left(\sum_{i=1}^{n}x_i\right)\frac{1}{p}&=\left(n-\sum_{i=1}^{n}x_i\right)\frac{1}{1-p}\\ \left(\sum_{i=1}^{n}x_i\right)\left(1-p\right)&=\left(n-\sum_{i=1}^{n}x_i\right)p\\ \sum_{i=1}^{n}x_i&=np\\ \hat{p}=\frac{1}{n}\sum_{i=1}^{n}x_i&=\frac{\text{Número de éxitos}}{n}\end{aligned}$$

Como vemos, en este caso hemos obtenido el mismo estimador que por el método de los momentos, si bien en general no tiene por qué ocurrir así.

Ejemplo 4.13. (*modelo de regresión lineal)* Se dispone de n observaciones de dos variables $\{(X_i,Y_i)\,,\,i=1,\ldots,n\}$, siendo las Y_i independientes y tales que, para cada i, $Y_i\approx N\left(\beta_0+\beta_1X_i,\sigma\right)$, con β_0, β_1 y σ parámetros desconocidos. Así, en este modelo se asume que para cada valor fijo $X=x$, la Y sigue una distribución normal con esperanza $E\left[Y\,|X=x\right]=\beta_0+\beta_1x$ y varianza σ^2. Dicho de otra forma, los valores medios de Y siguen la recta $y=\beta_0+\beta_1x$; y los valores individuales de Y se distribuyen alrededor de esta recta, centrados en ella, y con varianza constante σ^2. La figura 4.3 ilustra esta situación.

Este modelo resulta en la práctica adecuado para representar la relación entre muchas variables: talla (X) y peso (Y) de los sujetos adultos de una especie; velocidad del viento (X) y altura de ola (Y); concentración de un compuesto químico (X) y absorbancia medida espectroscópicamente(Y); ...

Para estimar los parámetros β_0, β_1 y σ por máxima verosimilitud debemos determinar primero la función de verosimilitud. Como $Y_i\approx N\left(\beta_0+\beta_1X_i,\sigma\right)$, tenemos que

$$f_{\beta_0,\beta_1,\sigma}\left(y_i\,|X=x_i\right)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{y_i-\left(\beta_0+\beta_1x_i\right)}{\sigma}\right)^2\right)$$

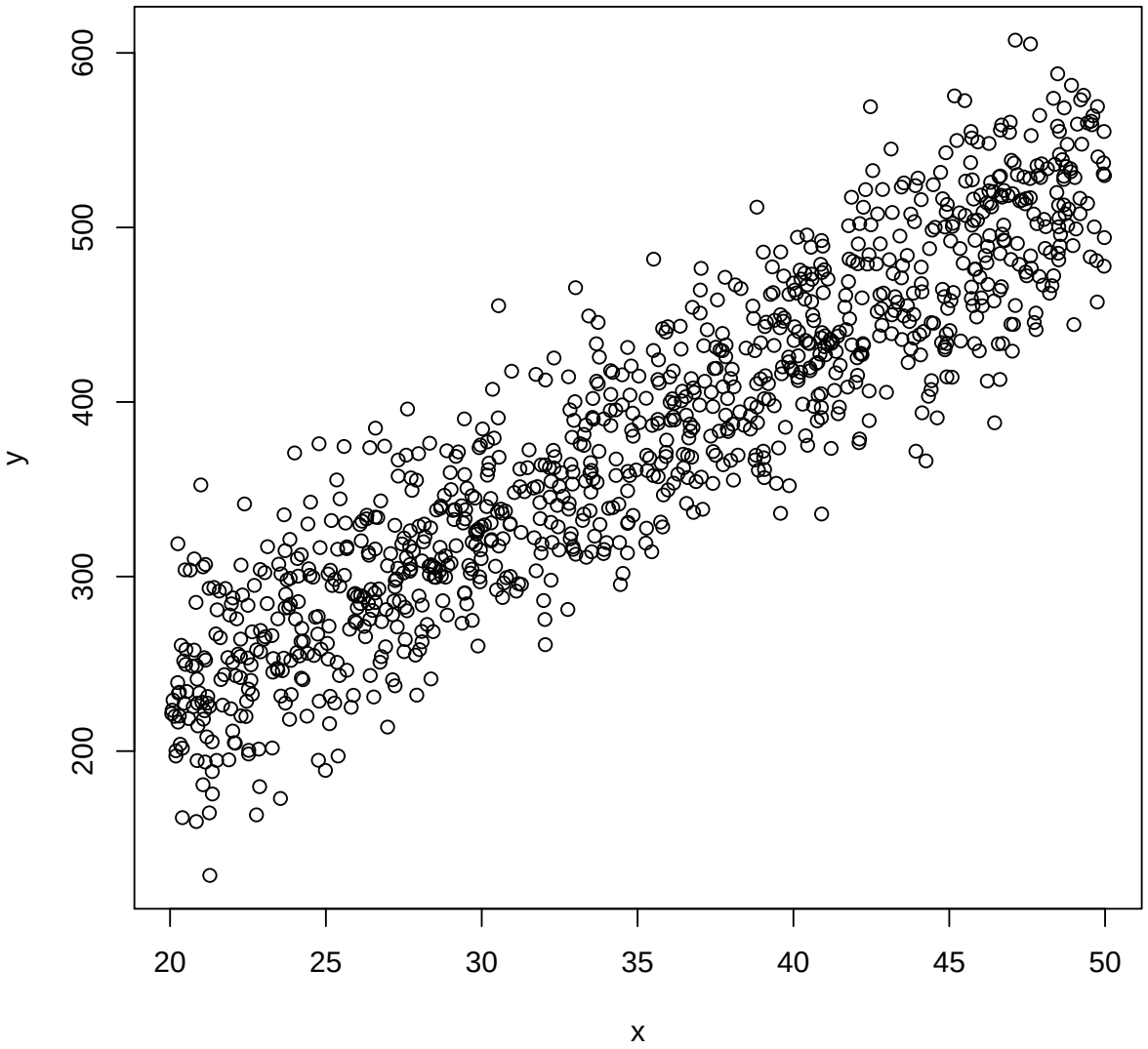

Figura 4.3: Nube de puntos que sigue un modelo de regresión lineal: $Y_i \approx N(\beta_0 + \beta_1 X_i, \sigma)$

Por tanto la función de verosimilitud será:

$$L(\beta_0, \beta_1, \sigma) = \prod_{i=1}^{n} f_{\beta_0,\beta_1,\sigma}(y_i) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_{i=i}^{n}\left(\frac{y_i - (\beta_0 + \beta_1 x_i)}{\sigma}\right)^2\right)$$

y la log-verosimilitud:

$$\ell\left(\beta_0,\beta_1,\sigma\right)=-n\log\left(\sigma\right)-n\log\left(\sqrt{2\pi}\right)-\frac{1}{2\sigma^2}\sum_{i=i}^{n}\left(y_i-\left(\beta_0+\beta_1x_i\right)\right)^2$$

Para obtener los valores de β_0, β_1 y σ que maximizan esta expresión, derivamos e igualamos a 0:

$$\frac{\partial}{\partial\beta_0}\ell\left(\beta_0,\beta_1,\sigma\right)=\frac{1}{\sigma^2}\sum_{i=i}^{n}\left(y_i-\left(\beta_0+\beta_1x_i\right)\right)=0\Rightarrow\sum_{i=i}^{n}\left(y_i-\left(\beta_0+\beta_1x_i\right)\right)=0$$
$$\frac{\partial}{\partial\beta_1}\ell\left(\beta_0,\beta_1,\sigma\right)=\frac{1}{\sigma^2}\sum_{i=i}^{n}\left(y_i-\left(\beta_0+\beta_1x_i\right)\right)x_i=0\Rightarrow\sum_{i=i}^{n}\left(y_i-\left(\beta_0+\beta_1x_i\right)\right)x_i=0$$
$$\frac{\partial}{\partial\sigma}\ell\left(\beta_0,\beta_1,\sigma\right)=-\frac{n}{\sigma}+\frac{1}{\sigma^3}\sum_{i=i}^{n}\left(y_i-\left(\beta_0+\beta_1x_i\right)\right)^2=0\Rightarrow\sum_{i=i}^{n}\left(y_i-\left(\beta_0+\beta_1x_i\right)\right)^2=n\sigma^2$$

De la primera ecuación se obtiene:

$$\sum_{i=1}^{n}\left(y_i-\beta_0-\beta_1x_i\right)=0\Rightarrow\sum_{i=1}^{n}y_i-\sum_{i=1}^{n}\beta_0-\sum_{i=1}^{n}\beta_1x_i=0\Rightarrow\sum_{i=1}^{n}y_i-n\beta_0-\beta_1\sum_{i=1}^{n}x_i=0\Rightarrow$$

$$\Rightarrow\beta_0=\frac{\sum_{i=1}^{n}y_i}{n}-\beta_1\frac{\sum_{i=1}^{n}x_i}{n}\Rightarrow\beta_0=\bar{y}-\beta_1\bar{x}\tag{4.3}$$

Sustituyendo en la segunda ecuación:

$$\sum\left(y_i-\beta_0-\beta_1x_i\right)x_i=0\ \ \Rightarrow\ \ \sum_{i=1}^{n}\left(y_i-\left(\bar{y}-\beta_1\bar{x}\right)-\beta_1x_i\right)x_i=0\Rightarrow$$

$$\sum_{i=1}^{n}\left(y_i-\overline{y}\right)x_i-\beta_1\sum_{i=1}^{n}\left(x_i-\overline{x}\right)x_i=0\ \ \Rightarrow\ \ \beta_1=\frac{\sum_{i=1}^{n}\left(y_i-\overline{y}\right)x_i}{\sum_{i=1}^{n}\left(x_i-\overline{x}\right)x_i}=\frac{\sum_{i=1}^{n}y_ix_i-\overline{y}\sum_{i=1}^{n}x_i}{\sum_{i=1}^{n}x_i^2-\overline{x}\sum_{i=1}^{n}x_i}\Rightarrow$$

$$\Rightarrow\hat{\beta}_1=\frac{\sum_{i=1}^{n}y_ix_i-n\overline{x}\,\overline{y}}{\sum_{i=1}^{n}x_i^2-n\left(\overline{x}\right)^2}\tag{4.4}$$

Por último, de la tercera ecuación se obtiene:

$$\sigma^2=\frac{1}{n}\sum_{i=i}^{n}\left(y_i-\left(\beta_0+\beta_1x_i\right)\right)^2$$

Sustituyendo β_0 por $\overline{y} - \beta_1 \overline{x}$, tras operar y simplificar, queda:

$$\sigma^2 = \frac{1}{n}\left[\sum_{i=1}^{n}(y_i - \overline{y})^2 - \beta_1^2 \sum_{i=1}^{n}(x_i - \overline{x})^2\right] \tag{4.5}$$

De esta forma, tras obtener el estimador $\hat{\beta}_1$ utilizando la ecuación 4.4, el estimador $\hat{\beta}_0$ se obtiene sustituyendo $\hat{\beta}_1$ en 4.3 y el estimador $\hat{\sigma}$ sustituyendo $\hat{\beta}_1$ en la ecuación 4.5.

Ejemplo 4.14. Supongamos ahora que tomamos una muestra de n observaciones de una variable con distribución de Weibull de parámetros κ y η. Para estimar estos parámetros por máxima verosimilitud, obtenemos primero la función de verosimilitud:

$$\begin{aligned} L(\kappa, \eta) &= \prod_{i=1}^{n}\left[\frac{\kappa}{\eta}\left(\frac{x_i}{\eta}\right)^{\kappa-1}\exp\left(-(x_i/\eta)^{\kappa}\right)\right] = \\ &= \left(\frac{\kappa}{\eta^{\kappa}}\right)^{n}\left(\prod_{i=1}^{n} x_i\right)^{\kappa-1}\exp\left(-\sum_{i=i}^{n}(x_i/\eta)^{\kappa}\right) \end{aligned} \tag{4.6}$$

La log-verosimilitud es entonces:

$$\ell(\kappa, \eta) = n\log(\kappa) - n\kappa\log(\eta) + (\kappa - 1)\sum_{i=1}^{n}\log(x_i) - \sum_{i=i}^{n}(x_i/\eta)^{\kappa}$$

Para determinar los valores de κ y η que maximizan esta expresión, calculamos las derivadas parciales e igualamos a 0:

$$\begin{aligned} \frac{\partial \ell(\kappa, \eta)}{\partial \kappa} &= \frac{n}{\kappa} - n\log(\eta) + \sum_{i=1}^{n}\log(x_i) - \sum_{i=i}^{n}(x_i/\eta)^{\kappa}\log(x_i/\eta) = 0 \\ \frac{\partial \ell(\kappa, \eta)}{\partial \eta} &= -\frac{n\kappa}{\eta} + \frac{\kappa}{\eta}\sum_{i=i}^{n}(x_i/\eta)^{\kappa} = 0 \end{aligned}$$

De la segunda ecuación se obtiene:

$$\frac{1}{\eta^{\kappa}}\sum_{i=1}^{n} x_i^{\kappa} = n \Rightarrow \eta = \left(\frac{1}{n}\sum_{i=1}^{n} x_i^{\kappa}\right)^{1/\kappa} \tag{4.7}$$

Reordenamos la primera ecuación:

$$\frac{n}{\kappa} - n\log(\eta) + \sum_{i=1}^{n}\log(x_i) - \frac{1}{\eta^\kappa}\sum_{i=i}^{n} x_i^\kappa\left(\log(x_i) - \log(\eta)\right) = 0$$

$$\frac{n}{\kappa} - n\log(\eta) + \sum_{i=1}^{n}\log(x_i) - \frac{1}{\eta^\kappa}\sum_{i=i}^{n} x_i^\kappa\log(x_i) + \frac{\log(\eta)}{\eta^\kappa}\sum_{i=i}^{n} x_i^\kappa = 0$$

$$\frac{n}{\kappa} + \log(\eta)\left(\frac{1}{\eta^\kappa}\sum_{i=i}^{n} x_i^\kappa - n\right) + \sum_{i=1}^{n}\log(x_i) - \frac{1}{\eta^\kappa}\sum_{i=i}^{n} x_i^\kappa\log(x_i) = 0$$

y sustituimos el valor de η:

$$\frac{n}{\kappa} + \frac{1}{\kappa}\log\left(\frac{1}{n}\sum_{i=1}^{n} x_i^\kappa\right)\left(\frac{1}{\frac{1}{n}\sum_{i=1}^{n} x_i^\kappa}\sum_{i=i}^{n} x_i^\kappa - n\right) + \sum_{i=1}^{n}\log(x_i) - \frac{n}{\sum_{i=1}^{n} x_i^\kappa}\sum_{i=i}^{n} x_i^\kappa\log(x_i) = 0$$

de donde, tras simplificar, se obtiene:

$$\kappa = \left[\frac{\sum_{i=1}^{n} x_i^\kappa\log(x_i)}{\sum_{i=1}^{n} x_i^\kappa} - \frac{\sum_{i=1}^{n}\log(x_i)}{n}\right]^{-1} \tag{4.8}$$

Esta última ecuación no tiene una solución explícita, debiendo resolverse numéricamente. Una vez que se obtenga de esta manera el valor estimado de $\hat{\kappa}$, se sustituye en la ecuación 4.7 obteniéndose así el estimador máximo verosímil $\hat{\eta}$.

Utilización de R para la estimación de parámetros por el método de máxima verosimilitud.

Como hemos visto en este último ejemplo, la estimación de parámetros por el método de máxima verosimilitud puede ser costosa debido a los cálculos que se deben realizar. Además como también ha ocurrido en este ejemplo, el método no tiene por qué proporcionar soluciones explícitas para los parámetros, por lo que finalmente deben aplicarse métodos numéricos para su obtención. Si bien podríamos proceder con la ecuación 4.8 de modo similar a como ya hicimos para obtener los estimadores por el método de los momentos (definiendo una función que cambie de signo en los extremos y utilizar *uniroot()*), presentamos a continuación un método más general que utiliza la función *optim()* de R para obtener directamente los valores de los parámetros que maximizan la log-verosimilitud.

Para ello es preciso definir primero una función que calcule la log-verosimilitud. En el caso de la distribución de Weibull, la ecuación 4.6 nos da su log-verosimilitud. Su implementación en R es muy sencilla:

```
logver = function(parms, x) {
    k = parms[1]
    eta = parms[2]
    n = length(x)
    lv = n * log(k) - n * k * log(eta) + (k - 1) * sum(log(x)) - sum((x/eta)^k)
    return(lv)
}
```

Como vemos, *logver()* depende de dos vectores: *parms*, que contiene los parámetros de la distribución, y *x* que contiene los valores observados en la muestra. Para obtener ahora los valores de los parámetros que maximizan la log-verosimilitud, utilizaremos *optim()* con los siguientes argumentos:

- *par*: valores iniciales de los parámetros, con los que el algoritmo inicia la búsqueda del óptimo. En este caso usaremos *c(1,1)* (valor que hemos tomado de forma arbitraria). En la siguiente sección presentamos un método que permite obtener estos valores iniciales.
- *logver*: la función a optimizar, en este caso la log-verosimilitud.
- *x=olas*: argumentos adicionales de la función a optimizar, en este caso, los datos muestrales de alturas de ola.
- *control=list(fnscale=-1)*: con esto indicamos que lo que se pretende es *maximizar* la función (por defecto, *optim()* trata de minimizar).

Así pues, la llamada a la función *optim()* se realiza de la forma siguiente:

```
optim(par = c(1, 1), logver, x = olas, control = list(fnscale = -1))$par

## [1] 2.622 3.427
```

La función nos devuelve los valores de los parámetros que maximizan la log-verosimilitud, en el mismo orden en que se definen en la función *logver*, esto es, primero $\hat{\kappa}$ y luego $\hat{\eta}$. Como podemos ver, los valores son ligeramente diferentes a los obtenidos en el ejemplo 4.10 por el

método de los momentos, aún habiendo utilizado los mismos datos. Como hemos señalado más arriba, en general el método de los momentos y el método de máxima verosimilitud no producen exactamente los mismos valores estimados para los parámetros, siendo preferibles los estimadores MV por gozar de mejores propiedades.

Señalemos por último que R implementa las funciones de densidad de muchas distribuciones de probabilidad habituales en la práctica. Ello permite definir la función de log-verosimilitud de una manera alternativa muy simple, teniendo en cuenta que $\ell(\Theta) = \sum_{i=1}^{n} \log(f_\Theta(x_i))$. A modo de ejemplo, en el caso particular de la distribución de Weibull, su función de densidad en R es $f_{\kappa,\eta}(x) =$*dweibull(x,k,eta)*, por lo que la función de log-verosimilitud puede definirse como:

```
logver = function(parms, x) {
    k = parms[1]
    eta = parms[2]
    lv = sum(log(dweibull(x, k, eta)))
    return(lv)
}
```

lo que nos ahorraría tener que escribir explícitamente la función de log-verosimilitud tal como hicimos en la implementación anterior de *logver()*.

Para simplificar aún más las cosas, la librería MASS cuenta con una función específica para el cálculo de estimadores de máxima verosimilitud, la función *fitdistr()*. Para estimar los parámetros de la distribución de Weibull para estos datos simplemente utilizaríamos:

```
library(MASS)
fitdistr(olas, "weibull")
```

```
##    shape    scale
##   2.6214   3.4261
##  (0.3584) (0.2505)
```

Los valores que se muestran entre paréntesis son estimaciones de los errores estándar para el estimador de cada parámetro. Las pequeñas diferencias numéricas que se observan con la solución anterior se deben simplemente a errores de redondeo asociados a los distintos algoritmos de optimización empleados. La función *fitdistr()* reconoce las distribuciones *beta, cauchy, chi-squared, exponential, f, gamma, geometric, log-normal, lognormal,*

logistic, negative binomial, normal, Poisson, t y *weibull*. Si quisiéramos ajustar los parámetros de alguna otra distribución, deberemos implementar una función con la densidad correspondiente (o utilizar el método desarrollado más arriba).

4.4.3.3. Método de los mínimos cuadrados

En el contexto de la estimación de parámetros de una distribución de probabilidad, este método se traduce en localizar los parámetros de la distribución que minimicen los cuadrados de las distancias entre la función de distribución empírica de los datos y la función de distribución teórica correspondiente a dichos parámetros. En la práctica, este método es poco preciso, pero permite obtener estimaciones iniciales de los parámetros que luego se emplean como valores iniciales para la estimación de máxima verosimilitud, tal como hemos visto en la sección anterior.

Para aplicar este método, igual que en los casos anteriores suponemos que se cuenta con una muestra de n observaciones independientes $E = \{x_1, x_2, \ldots, x_n\}$ de una variable aleatoria X con función de distribución acumulativa $F_\Theta(x)$, y que esos valores están ordenados de menor a mayor. Sea $N(x_i)$ el número de observaciones cuyo valor es menor o igual que x_i (obviamente si todas las x_i son distintas, entonces $N(x_i) = i$). Las frecuencias relativas acumuladas $\hat{F}(x_i) = N(x_i)/n$, constituyen una aproximación de la función de distribución $F_\Theta(x)$ de la variable X. Esta aproximación, no obstante, da lugar a que para el valor más alto observado, x_n, se tenga $\hat{F}(x_n) = 1$, lo que de algún modo impone la restricción de que el valor más alto posible es precisamente x_n; ahora bien, que x_n sea el valor más alto observado en esta muestra particular no significa que sea el valor más alto que pueda observarse en general. Para evitar este problema pueden emplearse diversas alternativas, siendo las más frecuentes las siguientes:

$$(a)\ \hat{F}(x_i) = \frac{N(x_i)}{n+1} \qquad (b)\ \hat{F}(x_i) = \frac{N(x_i) - 0{,}5}{n} \qquad (c)\ \hat{F}(x_i) = \frac{N(x_i) - 0{,}3}{N(x_i) + 0{,}4}$$

El método de mínimos cuadrados consiste entonces en encontrar el valor de Θ que minimiza la suma de las diferencias al al cuadrado:

$$SC(\Theta) = \sum_{x_i \in E} \left(\hat{F}(x_i) - F_\Theta(x_i) \right)^2$$

Por tanto el *estimador de mínimos cuadrados* (*estimador MC*) es:

$$\hat{\Theta} = \arg\min SC(\Theta)$$

Ejemplo 4.15. Utilizaremos de nuevo los datos de alturas de ola del ejemplo 4.10, para estimar por mínimos cuadrados los parámetros κ y η de la distribución de Weibull que presumiblemente ha generado esos datos. Para ello consideraremos la estimación *(a)* anterior de la distribución empírica. Asimismo, la función de distribución acumulativa de Weibull que ya hemos visto en el capítulo anterior es de la forma $F_{\kappa,\eta}(x) = 1 - \exp\left(-(t/\eta)^{\kappa}\right)$. Debemos hallar entonces los valores de κ y η que minimizan:

$$SC(\kappa,\eta) = \sum_{i=1}^{n}\left(\hat{F}(x_i) - F_{\kappa,\eta}(x_i)\right)^2 = \sum_{i=1}^{n}\left(\frac{N(x_i)}{n+1} - 1 + \exp\left(-\left(\frac{t}{\eta}\right)^{\kappa}\right)\right)^2$$

Si bien podemos tratar de resolver este problema directamente (derivando con respecto a ambos parámetros, igualando a 0 y resolviendo las ecuaciones resultantes), es más sencillo linealizar el modelo de Weibull. Para ello observemos que:

$$\begin{aligned} 1 - F_{\kappa,\eta}(x) = \exp\left(-\left(\frac{t}{\eta}\right)^{\kappa}\right) \Rightarrow \ln\left(1 - F_{\kappa,\eta}(x)\right) = -\left(\frac{t}{\eta}\right)^{\kappa} \Rightarrow \\ \Rightarrow \ln\left(-\ln\left(1 - F_{\kappa,\eta}(x)\right)\right) = \kappa \ln\left(\frac{x}{\eta}\right) \Rightarrow \\ \Rightarrow \ln\left(-\ln\left(1 - F_{\kappa,\eta}(x)\right)\right) = \kappa \ln(x) - \kappa \ln(\eta) \end{aligned}$$

Esta última ecuación es lineal; llamando:

$$y = \ln\left(-\ln\left(1 - F_{\kappa,\eta}(x)\right)\right); \qquad t = \ln(x); \qquad \theta = -\kappa \ln(\eta)$$

podemos reescribir la ecuación anterior de la forma $y = \kappa t + \theta$. Para estimar entonces κ y η a partir de una muestra ordenada de valores $(x_1, x_2, \ldots, x_n)$ llamaremos:

$$\begin{aligned} \hat{y}_i &= \ln\left(-\ln\left(1 - \hat{F}(x_i)\right)\right) = \ln\left(-\ln\left(1 - \frac{N(x_i)}{n+1}\right)\right) \\ t_i &= \ln(x_i) \end{aligned}$$

y la suma de cuadrados a minimizar será:

$$SC(\kappa,\theta) = \sum_{i=1}^{n}\left(\hat{y}_i - (\kappa t_i + \theta)\right)^2$$

que corresponde a la suma de cuadrados de las distancias entre las observaciones $\hat{y}_i$ y los valores predichos por la recta $y = \kappa t + \theta$. Para obtener los valores de κ y θ que minimizan

$SC(\kappa, \theta)$, derivamos e igualamos a cero:

$$\begin{array}{rcll} \frac{\partial SC(\kappa,\delta)}{\partial \theta} & = & -2\sum_{i=1}^{n}(\hat{y}_i - (\kappa t_i + \theta)) = 0 & \Rightarrow \quad \sum_{i=1}^{n}(\hat{y}_i - (\kappa t_i + \theta)) = 0 \\ \frac{\partial SC(\kappa,\delta)}{\partial \kappa} & = & -2\sum_{i=1}^{n}(\hat{y}_i - (\kappa t_i + \theta))\, t_i = 0 & \Rightarrow \quad \sum_{i=1}^{n}(\hat{y}_i - (\kappa t_i + \theta))\, t_i = 0 \end{array}$$

Estas ecuaciones son análogas a las que ya resolvimos en el ejemplo 4.13 cuando obtuvimos los parámetros de un modelo de regresión lineal por el método de máxima verosimilitud. Por tanto la solución se obtiene del mismo modo, resultando:

$$\hat{\kappa} = \frac{\sum_{i=1}^{n}\hat{y}_i t_i - n\bar{t}\,\bar{\hat{y}}}{\sum_{i=1}^{n} t_i^2 - n\left(\bar{t}\right)^2}$$
$$\hat{\theta} = \bar{\hat{y}} - \hat{\kappa}\bar{t}$$

Por último, como $\theta = -\kappa \ln(\eta)$, se tiene que $\eta = \exp(-\theta/\kappa)$, por lo que $\hat{\eta} = \exp\left(-\hat{\theta}/\hat{\kappa}\right)$.

Podemos utilizar R para realizar esta estimación:

```
x = sort(olas)
Fxi = cumsum(table(x))/(length(x) + 1)
yi = log(-log(1 - Fxi))
ti = log(x)
parms = coef(lm(yi ~ ti))
names(parms) = NULL
kappa = parms[2]
eta = exp(-parms[1]/kappa)
kappa

## [1] 1.689

eta

## [1] 3.78
```

Hemos aprovechado que R cuenta con la función *lm()* que calcula la recta de mínimos cuadrados para predecir *yi* en función de *ti*; asimismo, hemos utilizado la función *coef()* para extraer los coeficientes de esa recta. Tal como puede apreciarse, los valores estimados $\hat{\kappa}$ y $\hat{\eta}$ se

alejan de los que ya obtuvimos por los métodos de los momentos y de máxima verosimilitud pues, como ya se ha dicho, el método de los mínimos cuadrados no es excesivamente preciso. Ahora bien, para utilizar el método de los momentos debimos proporcionar a R un intervalo de búsqueda; y para usar máxima verosimilitud debimos proporcionar unos valores iniciales de los parámetros. Para el método de mínimos cuadrados sólo hemos necesitado los datos. Por tanto, aunque los valores estimados proporcionados por este método no sean muy buenos, pueden utilizarse como valores iniciales para aplicar a continuación el método de máxima verosimilitud.

4.5. Estimación paramétrica con datos censurados.

En ocasiones los datos disponibles para un estudio contienen mediciones incompletas de la variable de interés. Por ejemplo:

1. Se estudia el tiempo que dura la presencia de un contaminante en el entorno costero. Se han realizado 18 ensayos, consistentes en expulsar una cantidad fija del contaminante a través de un emisario submarino y registrar durante cuantos días se detecta en la zona de emisión. Los ensayos duran como mucho una semana y en tres de ellos, al término del ensayo el contaminante aún era detectable. Si X es el número de días que dura la presencia del contaminante, de las 18 observaciones hay tres en las que no se conoce el valor exacto de X, sino sólo que $X \geq 7$.

2. Se dispone de un aparato para medir la altura de ola. Tras sufrir una avería, para las olas de más de 6 metros el aparato registra siempre el valor 6. Si se han observado las alturas de 100 olas y en 12 de ellas el valor registrado es 6, ello quiere decir que en esas 12 observaciones es $X \geq 6$ (siendo X la altura de ola).

3. Se dispone de un aparato para medir la concentración de CO_2 disuelto en el agua de mar. La sensibilidad del aparato es tal que si la concentración está por debajo del valor u, se registra un cero. Por tanto, si el valor 0 se ha registrado k veces durante un periodo de observación, ello significa que en realidad ha habido k valores para los que $X \leq u$ (siendo X la concentración de CO_2).

Cuando se dan estas circunstancias, los datos se dicen *censurados: no se conoce su valor exacto, pero sí que están por debajo (censura por la izquierda) o por encima (censura por la derecha) de cierto valor.* Si se desea estimar los parámetros de las distribuciones de probabilidad de variables como las citadas, sería incorrecto considerar los valores censurados como si fuesen los valores realmente observados en la variable. En el tercero de los ejemplos, si

quisiéramos estimar la concentración media de CO_2 disuelto y considerásemos que los ceros que da el aparato son reales, cuando en realidad son producto de su falta de sensibilidad, es evidente que subestimaríamos la concentración media de CO_2 en la zona de interés.

En presencia de datos censurados, el único método que produce estimaciones fiables es el método de máxima verosimilitud, ya que es posible incluir la presencia de la censura en la función de verosimilitud:

- Si los datos presentan censura por la derecha (como los de los ejemplos 1 y 2 anteriores): sean $x_1, x_2, \ldots, x_r$ las observaciones completas, y $x_{r+1}, x_{r+2}, \ldots, x_n$ las observaciones censuradas (esto es, sólo se sabe que $X_{r+1} \geq x_{r+1}, X_{r+2} \geq x_{r+2}, \ldots, X_n \geq x_n$). La verosimilitud en este caso es:

$$L(\Theta) = f_\Theta(x_1) f_\Theta(x_2) \ldots f_\Theta(x_r) S_\Theta(x_r) S_\Theta(x_{r+2}) \ldots S_\Theta(x_n)$$

  siendo $S_\Theta(x) = 1 - F_\Theta(x) = P(X \geq x)$ la llamada *función de supervivencia de* X.

- Si los datos presentan censura por la izquierda (como los del ejemplo 3 anterior): sean $x_1, x_2, \ldots, x_r$ las observaciones completas, y $x_{r+1}, x_{r+2}, \ldots, x_n$ las observaciones censuradas (esto es, sólo se sabe que $X_{r+1} \leq x_{r+1}, X_{r+2} \leq x_{r+2}, \ldots, X_n \leq x_n$). La verosimilitud en este caso es:

$$L(\Theta) = f_\Theta(x_1) f_\Theta(x_2) \ldots f_\Theta(x_r) F_\Theta(x_r) F_\Theta(x_{r+2}) \ldots F_\Theta(x_n)$$

  siendo $F_\Theta(x) = P(X \leq x)$ la función de distribución acumulativa de X.

Una vez definida la función de verosimilitud con datos censurados, el resto del proceso de estimación es análogo al método de máxima verosimilitud ya visto: derivar la log-verosimilitud con respecto a cada uno de los parámetros, igualar a cero cada derivada y resolver el sistema de ecuaciones resultante.

El lector puede comprobar, a modo de ejemplo, que si $X \approx W(\kappa, \eta)$, los estimadores MV de κ y η en presencia de censura por la derecha se obtienen a partir de:

$$\hat{\kappa} = \left(\frac{\sum_{i=1}^{n} x_i^{\hat{\kappa}} \log(x_i)}{\sum_{i=1}^{n} x_i^{\hat{\kappa}}} - \frac{\sum_{i=1}^{r} \log(x_i)}{r} \right)^{-1}$$

$$\hat{\eta} = \left(\frac{1}{r} \sum_{i=1}^{n} (x_i)^{\kappa} \right)^{1/\kappa}$$