5. MÉTODOS DE REMUESTREO

Para obtener una red neuronal que realice una tarea deseada de manera correcta dicha red pasará por dos fases: uno de entrenamiento y otro de validación. Para ambas fases se utilizará unos patrones que generalmente se dividirán en dos grupos: uno para el entrenamiento y otro para la validación. Sin embargo, si el número de patrones es insuficiente se podría entrenar a la red con todos los patrones disponibles y verificar la validez de dicha red mediante otros métodos.

El método que se empleará será el método de *Bootstrap* que es un método de análisis de datos por remuestreo. Este método fue descrito por Bradley Efron en 1979. Los métodos de análisis de datos por remuestreo son métodos costosos computacionalmente. En estas técnicas el cálculo de un estimador se realiza múltiples veces sobre muestras de datos (remuestreados) de una muestra aleatoria tomada previamente.

Normalmente para poder estimar unos parámetros que se desean conocer de una población se necesita realizar previamente unas suposiciones y crear un modelo teórico que puede no ajustarse a las características reales de la población bajo estudio. Las técnicas de remuestreo sólo se basan en los datos disponibles y en el conocimiento de cómo dichos datos fueron recolectados. Así, el análisis de datos por remuestreo se puede considerar como basado en el diseño y no basado en modelo.

En los siguientes apartados se introducirá los métodos de remuestreo y en particular el método de *bootstrap.*

5.1. ESTIMACIÓN DE LA DISTRIBUCIÓN DE MUESTREO

El objetivo del método *bootstrap*, y de la estadística en general, es estudiar una cualidad o característica que posee una población. En nuestro caso la población será las respuestas de salida de la red neuronal y la característica será el error cometido por dicha red neuronal.

La población tendrá una distribución a la que realizaremos una o más inferencias. La distribución de población lo denotaremos *X*. La población estará formada por *N* casos y tendrá asociado algún parámetro θ que se busca estimar. Para ello, tomamos una muestra aleatoria de *n* casos de la población, *x*, y hallaremos una estimación del parámetro, *t*.

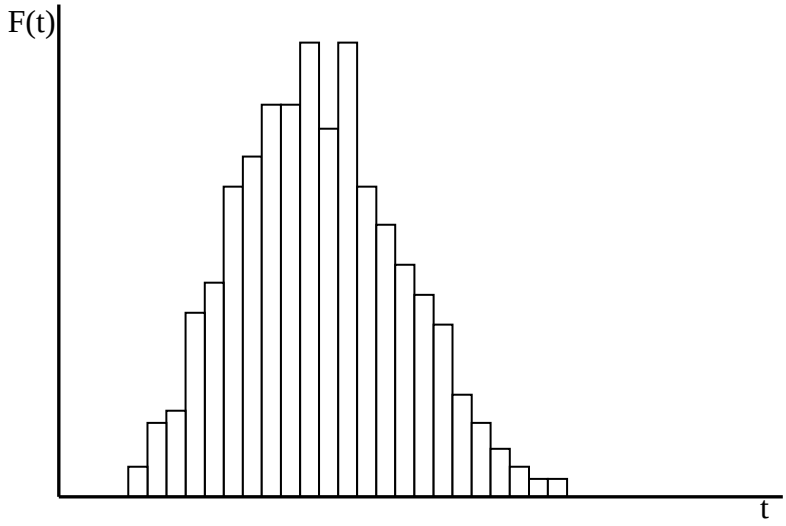

Figura 5.1: distribución de muestreo.

La distribución de muestreo de un estimador, $F(t)$, está formada por los valores del estimador de muchas muestras de la población. Para tener una distribución de

muestreo completa debemos de tener los valores de las estimaciones de todas las posibles muestras de tamaño *n* de la población. En total habrá *M* muestras de este tamaño con:

$$M = \frac{N!}{n!(N-n)!} \qquad \textit{Ec.(5.1)}$$

Sin embargo, solamente se ha tomado una única muestra. Esto se ve reflejado en la Figura 5.2.

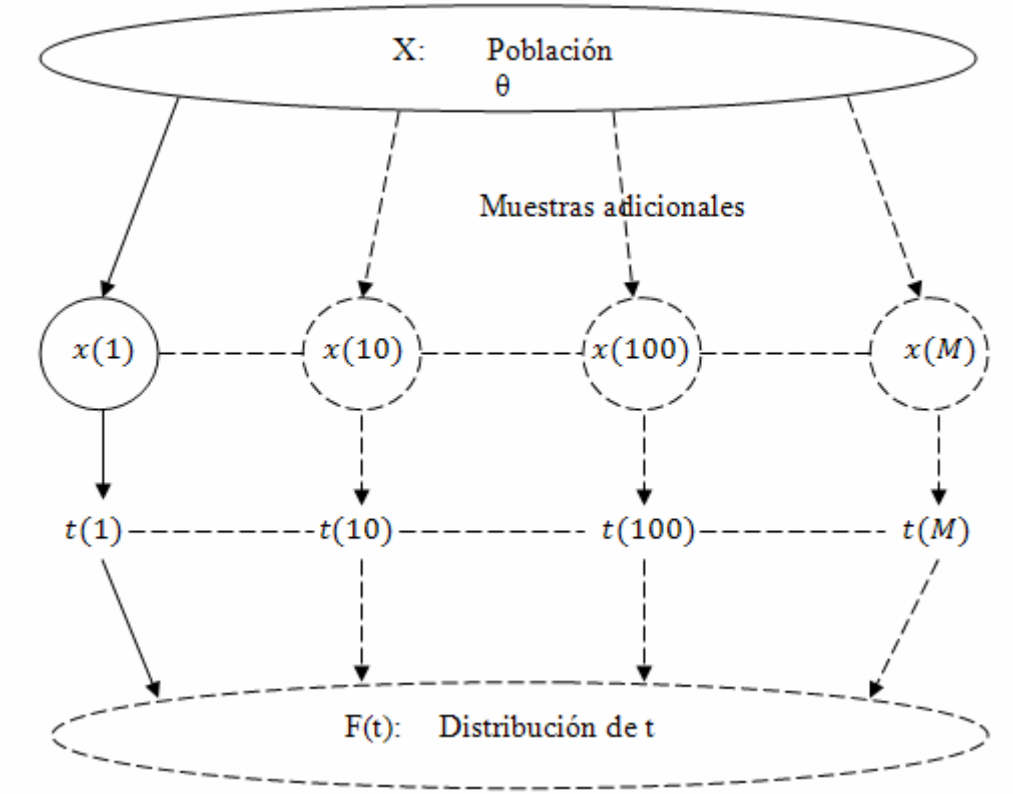

Figura 5.2: Población y distribución de muestreo real.

En este caso lo único que conocemos es el tamaño de la población *N* que vamos a estudiar, las muestras tomadas a dicha población y su tamaño, *x* y *n* respectivamente, y el estimador calculado a partir de esta muestra, *t*.

Para evitar tomar tantas muestras se recurre a la estadística tradicional. La teoría indica si la población tiene una distribución normal la distribución de muestreo también es normal. Si la población no tiene una población normal se aplica el Teorema Central del Límite si la muestra es grande y tendremos una distribución de muestreo normal. Si la población tiene una media μ y una desviación estándar σ, la distribución de muestreo de la población tendrá una media μ y una desviación estándar $\sigma/\sqrt{n}$ Sin embargo, esto es sólo un modelo donde pequeñas desviaciones puede invalidarla. Otro problema es el cálculo σ^2 que es distinta para cada modelo y puede ser complicado hallarlo e incluso imposible.

Si la distribución de población no es conocida se puede recurrir a los métodos de remuestreo. En un entorno *bootstrap* tendremos, al igual que en el primer caso, una población $\hat{X}$ de tamaño *N*. La población *bootstrap* se diferenciará de una población real por el símbolo ^. A diferencia de la población *X*, $\hat{X}$ es completamente conocido. Así mismo, el parámetro $\hat{\theta}$ también es conocido y se puede hallar a partir de la población.

Como $\hat{X}$ es completamente conocida, se puede tomar tantas muestras aleatorias de tamaño *n* que se desee, x^*, y hallar la estimación del parámetro buscado, t^*, para cada muestra. Esto se muestra en la Figura 5.3. En teoría esto se podría realizar para todas las *M* posibles muestras de tamaño *n* que se puede obtener a partir de $\hat{X}$ y hallar la distribución de muestreo de t^*. Sin embargo, en la práctica *M* es demasiado grande y se toma *B* muestras de tamaño *n* siendo *B* un número suficientemente elevado.

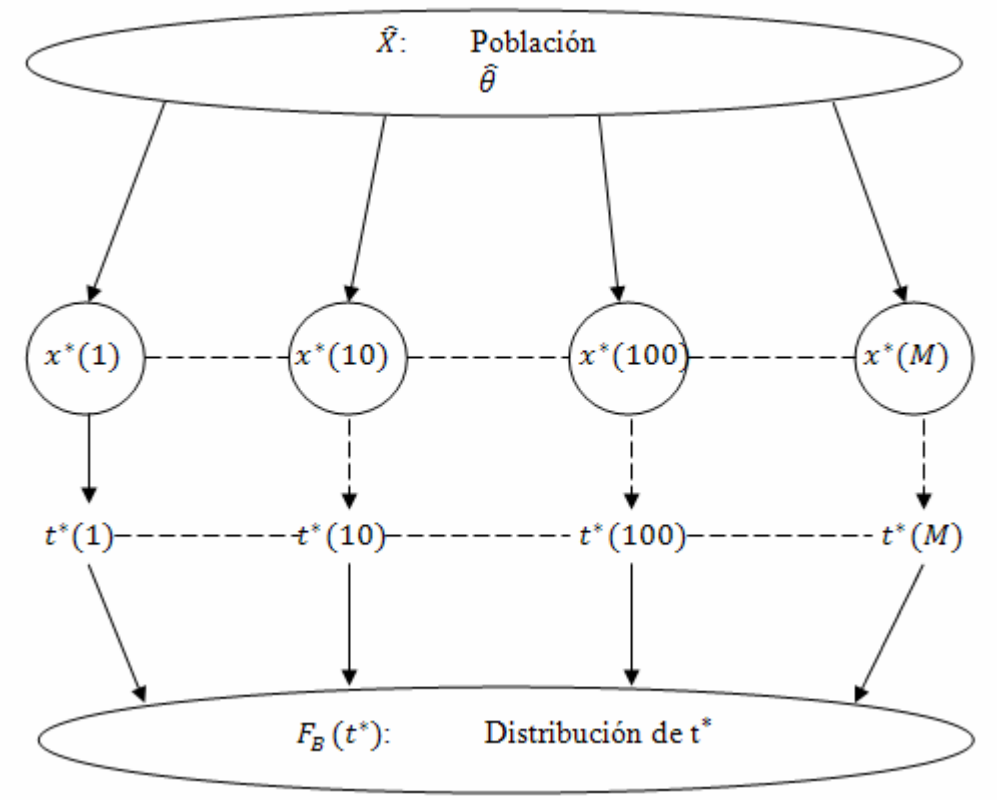

Figura 5.3: Población y distribución de muestreo Bootstrap.

5.1.1. Estimación de la distribución de población

Para realizar nuestras estimaciones mediante el método de *bootstrap*, necesitamos conocer $\hat{X}$. $\hat{X}$ será una aproximación numérica de *X*, y las distintas formas de obtenerlo se explicará en los siguientes apartados. Todos los métodos hallan $\hat{X}$ a partir de la única muestra aleatoria de tamaño *n* conocida de *X*, esto es, *x*. Existen cuatro métodos distintos para obtener $\hat{X}$ dependiendo de cuánto se conozca de *X* y de cuánto se suponga.

5.1.1.1. Estimación no paramétrica

De los cuatro métodos de estimación, normalmente se emplea la estimación no paramétrica. En esta estimación no se realiza ninguna suposición sobre la forma de la distribución de la población, de ahí el nombre de no paramétrica. Se intenta estudiar a la población *X* a partir de las muestras tomadas aleatoriamente. Así, lo único que se conoce de la población *X* es:

- Es de tamaño *N*.
- La muestra aleatoria *x* de tamaño *n*.

Al no conocer nada más de la distribución de la población se asume que los $(N-1)$ casos que no han sido muestreados y no forman parte del conjunto *x* tienen una distribución igual a la población completa *X*.

Existen tres formas para formar la población no paramétrica $\hat{X}$ según el tamaño de la población, *N*, y el tamaño de la muestra, *n*.

Poblaciones grandes

Si el tamaño de la población es al menos veinte veces mayor que el tamaño de la muestra, se puede considerar que la población es grande. La forma para obtener la muestra *x* es mediante el muestreo sin reposición. En el caso de poblaciones grandes, la distribución de población no se ve afectado por el método de muestreo. Así, al formar $\hat{X}$ podemos simular el muestreo sin reposición muestreando con reposición. De esta forma, $c=N/n$ no necesita ser un entero.

Poblaciones pequeñas, N/n entero

La población no paramétrica $\hat{X}$ estará formada por *c* copias de *x* siendo *c:*

$$c = \frac{N}{n}$$

Esto sólo es posible si *c* es un número entero, en caso contrario el método se complica algo más.

Poblaciones pequeñas, N/n no entero

Si *c=N/n* no es entero se utiliza un algoritmo propuesto por Booth, Butler y Hall (1994). Sea *C* la parte entera de *c=N/n*. Formaremos $\hat{X}$ con *C* copias de *x*. Sin embargo, aún necesitamos *k=N-Cn* casos más para que la población estimada $\hat{X}$ sea de tamaño *N*. Estos *k* casos se tomarán de manera aleatoria de *x*. Tras tomar *m* muestras bootstrap de $\hat{X}$, se cambian esos *k* casos por otros *k* casos aleatorios tomados de *x* para balancear la aparición de todos los casos de *x* en la población estimada $\hat{X}$. Un valor típico para *m* sería *m=100*.

5.1.1.2. Estimación paramétrica

En esta estimación, al igual que en el caso anterior, conocemos:

- El tamaño de la población *X*, *N*.
- La muestra aleatoria *x* de tamaño *n*.

Además de lo anterior, también conocemos:

- La forma paramétrica de la población *X*, esto es, se pude expresar matemáticamente la distribución de la población.

Con esto, la $\hat{X}$ paramétrica se obtendría generando *N* casos siguiendo la forma de la distribución de la población, por ejemplo una normal, poisson… Si estas

distribuciones necesitan parámetros que no conocemos, éstas se obtendrían a partir de las muestras aleatorias tomadas *x*.

5.1.1.3. Estimación suavizada

Esta estimación está entre los dos descritos anteriormente. En la estimación no paramétrica lo único conocido de la población era su tamaño y la muestra aleatoria tomada y se formaba la población $\hat{X}$ con múltiples copias de la muestra *x*.

Sin embargo, la población $\hat{X}$ estará formada por *N* casos de los cuales tan sólo tendremos como mucho *n* distintos valores. Se podría pensar que si se tomara otra muestra obtendríamos otros valores diferentes que los obtenidos en la primera muestra, pero no tenemos datos suficientes como para suponer una distribución para formar $\hat{X}$ con el método paramétrico.

Este problema se soluciona utilizando una distribución suavizada de la muestra tomada *x*. Esto es, se deja que valores cercanos a los obtenidos en *x* contribuyan a la formación de $\hat{X}$. Existen muchas técnicas de suavizado entre las que elegir además de cuánto se desea suavizar la distribución.

5.1.1.4. Estimación basada en modelos

La última forma de estimar la población *X* requiere que se conozca un modelo para obtener los casos que componen a la población. Un ejemplo sería la distribución de la población en el análisis de varianza donde el valor del i-ésimo miembro del j-ésimo grupo de tratamiento es:

$$y_{ij} = \mu_j + \varepsilon_{ij} \qquad \textit{Ec.(5.2)}$$

Donde μ_j es la media de la población donde todos han sido expuestos al tratamiento j-ésimo, y ε_{ij} es un error aleatorio. La distribución del error se asume normal de media cero y con una varianza σ^2. En este modelo son los errores los que son muestreados en vez de las muestras propiamente dichas.

En el estudio de la varianza, la distribución de población de errores, se puede estimar paramétricamente o de forma no paramétrica. También se puede suavizar la distribución del error estimado.

5.2. DISTRIBUCIÓN DE MUESTREO BOOTSTRAP

Nuestro objetivo es el estudio de una población *X* formado por *N* casos. Esta población tendrá una distribución para los valores de un atributo que estará caracterizado por un parámetro θ que se desea estimar. La estimación de θ, *t*, tendrá una distribución de muestreo que se formará tomando todas las *M* muestras de tamaño *n* de la población posible. Cada muestra tendrá una distribución de los valores del atributo del que se hallará el estimador. La colección de los *M* estimadores formará la distribución de muestreo $F(t \mid X, n)$.

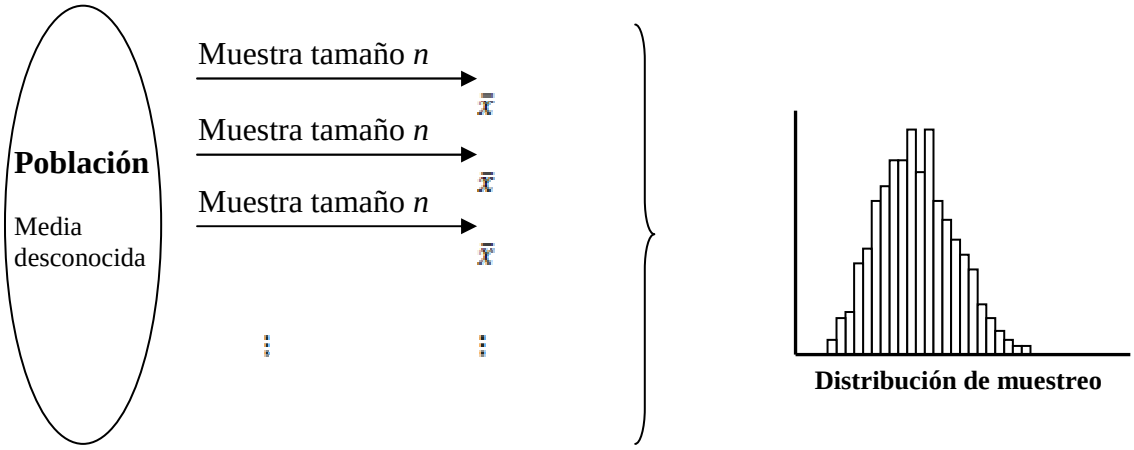

Figura 5.4: Distribución de muestreo de la media.

La distribución de muestreo $F(t \mid X, n)$, nos indicaría cómo de preciso es *t* como estimador de θ. Esto es, a partir de la distribución de muestreo podríamos conocer el sesgo o bias, que se hallaría de la siguiente manera:

$$Bias(t \mid X, n, \theta) = Media(t \mid X, n) - \theta \qquad Ec.(5.3)$$

la desviación estándar:

$$SE(t \mid X, n) = \sqrt{(1/M)\sum_{i=1}^{M}[t_i - Media(t \mid X, n)]^2} \qquad Ec.(5.4)$$

y el error cuadrático medio:

$$RMS(t \mid X, n, \theta) = \sqrt{(1/M)\sum_{i=1}^{M}(t_i - \theta)^2} \qquad Ec.(5.5)$$

Debido al tamaño de la población, generalmente es imposible hallar la distribución de muestreo. Esto se soluciona con el método *bootstrap*. Se sustituye la población *X* por una estimación no paramétrica $\hat{X}$ descrita en el apartado anterior. Esta población no será más que copias de la única muestra de la población real del que disponemos. La población $\hat{X}$ tendrá asociado un parámetro $\hat{\theta}$ y se calculará su estimación *t** y la distribución de muestreo $F_B(t^* \mid \hat{X}, n)$ de la misma manera se calcularía a la población *X* si fuese posible.

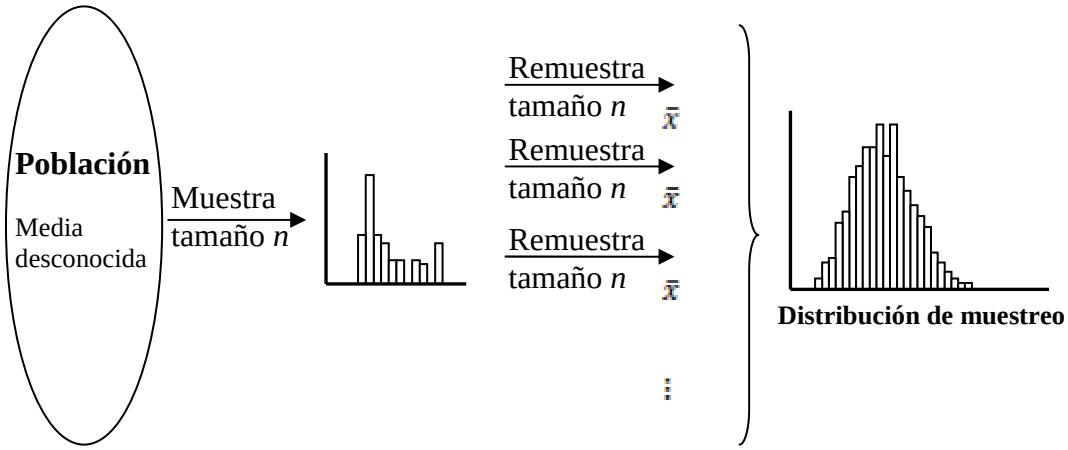

Figura 5.5: Distribución de muestreo de la media mediante Boostrap.

Si $\hat{X}$ está formado sólo por copias de una única muestra de tamaño *n* de la población real, ¿cuánto nos puede decir de la población real la distribución de muestreo obtenido a partir de la población estimada? Al igual que en la población real, también se puede calcular el sesgo, la desviación estándar y el error cuadrático medio a la población estimada. Esto se verá en los siguientes apartados. Sin embargo, se verá distintos métodos para tomar las muestras a la población y hallar la distribución de muestreo.

5.2.1. Distribución de muestreo bootstrap de Monte Carlo

Para tener una distribución de muestreo completa de un estimador *t** se necesitas todas las *M* posibles muestras de tamaño *n* que se pueda obtener da la población estimada $\hat{X}$ de tamaño *N*. Sin embargo, *M* suele ser un número demasiado elevado para llevar este método a la práctica. Generalmente se aproxima la distribución de muestreo *bootstrap*. Esta aproximación consistirá en *B* muestras aleatorias en vez de *M*. *B* deberá ser un número suficientemente elevado, como por ejemplo 2000 o 5000, para obtener una buena aproximación de la distribución de muestreo.

Esta aproximación se llama Monte Carlo. Aquí tanto las muestras como la secuencia de los casos que aparecen en dichas muestras son aleatorias. A continuación veremos cómo se podría generar muestras aleatorias con secuencias aleatorias.

5.2.1.1. Muestras aleatorias sin reposición

Para crear una muestra aleatoria sin reposición de tamaño *n* de una población de tamaño *N* se sigue los siguientes pasos:

1. Se asigna un número aleatorio a cada uno de los *N* casos que conforman a la población.
2. Se ordena los *N* casos.
3. Se selecciona para la muestra los que tienen los *n* números más pequeños.

5.2.1.2. Muestras aleatorias con reposición

Cuando la población es al menos 20 veces mayor que la muestra aleatoria, utilizar muestras aleatorias con reposición es una buena aproximación al muestreo sin reposición.

Para crear una muestra aleatoria con reposición de tamaño *n* de una población de tamaño *N* se sigue los siguientes pasos:

1. Se asigna un número aleatorio a cada uno de los *N* casos que conforman a la población.
2. Se ordena los *N* casos.
3. Se selecciona para la muestra el caso con el número más pequeño asignado.
4. Repetir los pasos de 1 al 3 hasta completar la muestra de tamaño *n*.

En este algoritmo, como la población *bootstrap* es una estimación no paramétrica de la población real, $N = n$ dado que en una estimación no paramétrica $\hat{X}$ está formado por múltiples copias de una única muestra aleatoria de la población *X*.

5.2.2. Estimación bootstrap del error estándar

Cuando deseamos conocer una población a partir de una muestra se cometerá un error en la representación, ya que la muestra no será una representación exacta de dicha población, y este error será más acusado cuanto menor sea la muestra. A pesar de que estaremos introduciendo un error de muestreo al estimar la población, este error se puede calcular con precisión. Este error será el error estándar y se halla de la siguiente manera:

$$\overline{SE}_{Boot}(t \mid X, n) = \sqrt{[1/(B-1)]\sum_{b=1}^{B}\left[t_b^* - (1/B)\sum_{b=1}^{B} t_b^*\right]^2} \qquad Ec.(5.6)$$

Como $\overline{SE}$ ha sido calculado de una distribución de muestreo bootstrap, esta $\overline{SE}$ es sólo una estimación de la *SE* real.

El número de muestras aleatorias de tamaño *n* suficiente para el cálculo de la *SE* es en torno a la 100 muestras. Sin embargo, para una mayor estabilidad de la estimación se recomienda un número mayor, como por ejemplo unos 500.

5.2.3. Estimación bootstrap del bias

El sesgo o *bias* se introduce en el momento de obtener las muestras por diversos motivos, como por ejemplo, favorecer algunos elementos de la población, excluir determinados grupos de la población, o si la muestra no se ha tomado de manera

aleatoria. En este caso la muestra no será una buena representación de la población y las estimaciones realizadas pueden ser erróneas.

El *bias* se puede definir como la diferencia entre la media del parámetro buscado o estadístico y el valor estimado y se calcula con la siguiente ecuación:

$$\overline{Bias}_{Boot}(t \mid X, n, \theta) = (1/B)\sum_{b=1}^{B} t_b^* - \theta \qquad Ec.(5.7)$$

En general, el bias sólo es un problema cuando es mayor que la cuarta parte del *SE*. En caso contrario puede ser ignorado. Esto se justifica si vemos el error cuadrático medio, que se define como:

$$RMS(t \mid X, n, \theta) = \sqrt{(1/M)\sum_{i=1}^{M}(t_i - \theta)^2} \qquad Ec.(5.8)$$

El error cuadrático medio está relacionado con error estándar y el bias de la siguiente manera:

$$RMS(t \mid X, n, \theta) = \sqrt{SE(t \mid X, n)^2 + Bias(t \mid X, n, \theta)^2} \qquad Ec.(5.9)$$

Si $Bias(t|X, n, \theta) = 0.25\ SE$, entonces el error cuadrático medio será:

$$RMS = \sqrt{SE^2 + (0.25SE)^2} = 1.0308SE \qquad Ec.(5.10)$$

Esto es, el RMS sólo es un 3% mayor que el *SE*, y por tanto, el *SE* se puede considerar como una buena medida de la exactitud de la estimación.