UNIVERSIDAD DE SEVILLA
FACULTAD DE MATEMATICAS

MODELOS ALTERNATIVOS DE SIMULACION BOOTSTRAP

Memoria dirigida por:
Prof. Dr. D. Joaquín Muñoz García.
Prof. Dr. D. Antonio Pascual Acosta.

Memoria presentada por:
Rafael Pino Mejías.

UNIVERSIDAD DE SEVILLA
FACULTAD DE MATEMATICAS

MODELOS ALTERNATIVOS DE SIMULACION BOOTSTRAP

Visado en Sevilla
a 14 de Julio de 1992

Memoria dirigida por:
Prof. Dr. D. Joaquín Muñoz García.
Prof. Dr. D. Antonio Pascual Acosta.

Memoria presentada para
optar al Grado de Doctor en
Ciencias Matemáticas.
Sevilla, Julio de 1992.

Fdo.: Rafael Pino Mejías.

INDICE

INTRODUCCION 1

CAPITULO I. METODOS BOOTSTRAP. 9

1.0. INTRODUCCION 11

1.1. METODOS BOOTSTRAP. 12

1.1.1. Tipos de estimación bootstrap. 12

1.1.2. Método II de Efron. 19

1.2. METODOS BOOTSTRAP ALTERNATIVOS. 22

1.2.1. Problemática que presenta el método bootstrap. 22

1.2.2. Bootstrap paramétrico. 25

1.2.3. Bootstrap suavizado. 27

1.2.4. Bootstrap balanceado. 28

1.2.5. Cálculos bootstrap más eficientes. 30

1.2.5.1. Estimador del sesgo. 30

1.2.6. Método percentil con doble bootstrap. 32

CAPITULO II. GENERACION DE MUESTRAS BOOTSTRAP CON DETECCION DE MUESTRAS OUTLIERS. 36

2.0. INTRODUCCION. 38

2.1. VARIABILIDADES O ERRORES. 41

2.2. ESTRUCTURA ALGEBRAICA DEL CONJUNTO DE REALIZACIONES DE LAS MUESTRAS BOOTSTRAP. 43

2.2.1. Equivalencia entre el conjunto de realizaciones de las muestras bootstrap y el conjunto de las variaciones con repetición. 43

2.2.2. Composición de variaciones. 47

2.2.3. Relación de equivalencia $\mathfrak{R}$ sobre el conjunto de las variaciones con repetición. 51

2.2.4. Relación de equivalencia sobre el conjunto cociente inducido por $\mathfrak{R}$. . 58

2.3. ESTRUCTURA PROBABILISTICA. 62

2.4. GENERACION DE MUESTRAS BOOTSTRAP. 71

CAPITULO III. METODO DE SUAVIZACION BOOTSTRAP. 83

3.0. INTRODUCCION. 85

3.1. FUNCION DE DISTRIBUCION SUAVIZADA PARA DISTRIBUCIONES DEL TIPO I,II O III. 91

3.1.1. Función de Distribución empírica. 91

3.1.2. Función de Distribución suavizada para funciones de distribución F(x) del tipo I. 94

3.1.3. Función de Distribución suavizada para funciones de distribución F(x) del tipo II. 98

3.1.3. Función de Distribución suavizada para funciones de distribución F(x) del tipo III. 102

3.4. PROPIEDADES DE LAS FUNCIONES DE DISTRIBUCION SUAVIZADAS. 106

3.3. METODO BOOTSTRAP BASADO EN LAS FUNCIONES DE DISTRIBUCION SUAVIZADAS. 113

ANEXO I:

Programa utilizado para calcular la función de distribución y características de la variable número de puntos distintos de que consta una muestra bootstrap. 123

ANEXO II:

Subrutina utilizada para generar muestras bootstrap con detección de muestras bootstrap outliers. 127

ANEXO III:

Subrutina utilizada para la simulación de muestras bootstrap suavizadas. 132

ANEXO IV:

Descripción de los recursos software y hardware utilizados en las tabulaciones y simulaciones. 136

BIBLIOGRAFIA 138

1.2.2. Bootstrap paramétrico.

Ya en las primeras publicaciones donde Efron introduce el concepto de estimación Bootstrap distingue una variante de este método, que recibe el nombre de Bootstrap Paramétrico.

Si bien la función de distribución empírica puede considerarse como estimador no paramétrico de máxima verosimilitud de la función de distribución F(x), en ciertas situaciones donde se realiza un procedimiento estadístico bajo un entorno parámetrico dicha estimación podría mejorarse en el sentido de obtener como estimador de máxima verosimilitud otra función de distribución basada en la situación paramétrica.

En tal caso se podrá determinar una función de distribución de una determinada población teórica, como puede ser una normal multivariante.

Una vez calculada esa función de distribución, se puede realizar el proceso de estimación bootstrap, simulando muestras de tamaño n, pero no extraídas de la función de distribución empírica, sino de la función de distribución estimada. De este modo, la composición de las muestras bootstrap no se reduce a las observaciones de la muestra inicial.

Un caso donde se aplica el bootstrap paramétrico puede verse en Efron (1982) donde dicho autor aplica el bootstrap paramétrico a los datos de las Facultades de Derecho en Estados Unidos, siendo el estadístico utilizado el coeficiente de correlación lineal.

1.2.3. Bootstrap suavizado.

Si bien el bootstrap paramétrico ya induce un cierto grado de suavización en la función de distribución sobre la que se muestrea, pueden considerarse otros métodos más generales de lograr una función de distribución más suave.

Una primera posibilidad es la de utilizar una convolución de la función de distribución empírica con una función de distribución del tipo utilizado en el bootstrap paramétrico, logrando así un término medio entre el bootstrap paramétrico y el bootstrap habitual. Esto puede verse en Efron (1982).

Otra opción utilizada con cierta frecuencia en la literatura es la de utilizar funciones de distribución que se calculan a partir de estimaciones no paramétricas de la función de densidad, en particular empleando el método de la función núcleo. En Silverman (1987) se recoge un estudio sobre la posible conveniencia de utilizar el bootstrap suavizado.

De todas formas, estos métodos presentan el inconveniente de que en la práctica generalmente no está claro qué tipo de función de distribución suavizada se ha de utilizar. Además, una de las razones más importantes del uso del bootstrap radica en las propiedades de convergencia de la función de Distribución empírica $F_n(x)$ a $F(x)$. En el caso del bootstrap suavizado puede no estar nada claro tales convergencias.

1.2.4. Bootstrap balanceado.

Davison et al. (1986) propusieron un método para extraer las muestras bootstrap, dentro del contexto del Método II de Efron (1979), que en situaciones favorables reduce el error de simulación. Este método, llamado método balanceado, ó también método de permutaciones, se basa en lograr que cada observación de la muestra inicial aparezca exactamente B veces en el conjunto de las B muestras bootstrap simuladas. En concreto, el procedimiento a seguir es el siguiente:

1. Se realizan B copias de la muestra inicial $(x_1, x_2, \ldots, x_n)$ en un vector **V** de longitud nB. Por tanto

   $\mathbf{V}=(x_1, x_2, \ldots, x_n, x_1, x_2, \ldots, x_n, \ldots, x_1, x_2, \ldots, x_n)=$

   $=(V_1, V_2, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, V_{nB})$

2. Se realiza una permutación aleatoria de este vector **V** de longitud nB, obteniéndose otro vector **W**, de la misma longitud.

   $\mathbf{W}= (W_1, W_2, \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots, W_{nB})$

3. Se van construyendo las B muestras bootstrap leyendo bloques sucesivos de longitud n en el vector **W** permutado. Las B muestras bootstrap serían, por tanto,

   $\mathbf{X}^{*1} = (W_1, W_2, \ldots, W_n)$

   $\mathbf{X}^{*2} = (W_{n+1}, W_{n+2}, \ldots, W_{2n})$

   .....

   .....

   $\mathbf{X}^{*B} = (W_{(n-1)B+1}, W_{(n-1)B+2}, \ldots, W_{nB})$

A partir de aquí, las estimaciones bootstrap se realizan exactamente igual que con los métodos habituales de obtención de muestras bootstrap aplicando muestreo aleatorio con reemplazamiento.

Según los autores del artículo antes mencionado, este método de simulación reduce el error de estimación del sesgo, varianza y percentiles, de forma que incluso se puede reducir el número de simulaciones necesarios, sin que el tiempo de computación aumente excesivamente por el hecho de mantener la estructura balanceada del muestreo.

1.2.5. Cálculos bootstrap más eficientes.

Efron (1990) propone diversos métodos que permiten mejorar la eficiencia de las estimaciones bootstrap, si bien a diferencia de Davison et al. (1986) no se centra en la reducción del error de simulación sino en la disminución del número B de simulaciones necesarias para obtener estimaciones satisfactorias.

Se ilustra la exposición de este método en el caso de la estimación del sesgo.

1.2.5.1. ESTIMADOR DEL SESGO.

Se utiliza el mismo algoritmo de generación de muestras bootstrap, obteniendo así un conjunto de B pares

$$(X^*_b, T_b), \quad T_b = T(X^*_b), \quad b=1,2,\ldots,B$$

El estadístico $T=T(X_1,X_2,\ldots,X_n)$ se supondrá invariante ante permutaciones de la muestra. Sea $x_1,x_2,\ldots,x_n$ una realización de $X_1,X_2,\ldots,X_n$. Dada una muestra bootstrap $\mathbf{X}^* = (X^*_1,X^*_2,\ldots,X^*_n)$, se define

$$P_i = \frac{card\{X^*_j = x_i\}}{n}$$

Por tanto el vector $\mathbf{P}=(P_1,P_2,\ldots,P_n)$ es un vector de probabilidad cuya componente i-ésima contiene la proporción de elementos de la muestra bootstrap que coinciden con x_i. El vector **P** suele recibir el nombre de vector de remuestreo.

Se puede escribir $T(\mathbf{P})$ en lugar de $T(\mathbf{X}^*)$ dada la propiedad de invarianza ante permutaciones de la muestra. En particular, si se define el vector $\mathbf{P}_0=(1/n,1/n,\ldots,1/n)$, se cumple que $T(\mathbf{P}_0)=T_0$, el valor del estadístico T para la muestra inicialmente observada.

El vector de remuestreo $\mathbf{P}$ está formado por componentes cuyos valores son múltiplos enteros de 1/n. En estas técnicas se asume que $T(\mathbf{P})$ está definido como función continua de $\mathbf{P}=(P_1,P_2,\ldots,P_n)$ siendo $\mathbf{P}$ cualquier vector de probabilidad sobre n puntos.

Las muestras bootstrap $\mathbf{X}^*_1,\mathbf{X}^*_2,\ldots,\mathbf{X}^*_B$ pueden ser representadas por sus correspondientes vectores de remuestreo $\mathbf{P}_1,\mathbf{P}_2,\ldots,\mathbf{P}_B$, con $T(\mathbf{P}_b)=T_b$, para b=1,2,...,B. Se define

$$\overline{\mathbf{P}} = \frac{1}{B}\sum_{b=1}^{B}\mathbf{P}_b$$

como el vector media de los B vectores de remuestreo. Se define el estimador bootstrap eficiente $SESGO_B$ como

$$SESGO_B=\overline{T} - T(\overline{\mathbf{P}})$$

que puede producir importantes ahorros de cálculo frente al estimador habitual

$$sesgo_B = \overline{T} - T_0$$

La información presente en los vectores de probabilidad de remuestreo permiten también obtener nuevos estimadores del error estándar y de los percentiles. Dada la extensión que requiere su explicación, no se incluyen en esta memoria.

1.2.6. Método percentil con doble bootstrap.

Recientemente, Shi (1992) presenta otro método de estimaciones bootstrap basado en un bootstrap doble, en el cual se realiza muestreo bootstrap en la muestra inicial, y a continuación se realiza otro muestreo bootstrap sobre cada muestra bootstrap generada. En concreto, realiza una modificación del método habitual de estimación de percentiles que según este autor puede mejorar la proporción de cubrimiento del parámetro real.

Para ello razona de la forma siguiente:

Si se pudieran extraer muestras $\mathbf{X}=(X_1,X_2,\ldots,X_n)$ iid según $F(x)$ y se conociera el parámetro θ, se podría calcular un límite de confianza superior tomando

$$\overline{T}=\overline{T}(\boldsymbol{X},\beta)=\overline{G}^{-1}(\beta) \;/\; P\{\theta\leq\overline{T}(\boldsymbol{X},\beta)/_F\}=\alpha$$

siendo

$$\overline{G}(y) = P\{T^*\leq y/_{F_n}\}$$

es decir, la función de distribución bootstrap de T, muestreando en la función de distribución empírica correspondiente a la muestra $\mathbf{X}$.

Lo anterior se puede escribir según el siguiente sistema de ecuaciones:

$$P\{\theta \leq \overline{T}(x,\beta)/_{F}\} = \alpha$$

$$P\{T^{*} \leq \overline{T}(x,\beta)/_{F_n}\} = \beta$$

El anterior sistema se puede convertirse en otro sistema en donde F_n juegue el papel de F:

$$P\{T_0 \leq \overline{T}^{*}(X^{*},\overline{\beta})/_{F_n}\} = \alpha$$

$$P\{T^{**} \leq \overline{T}(x,\beta)/_{F_n^{*}}\} = \overline{\beta}$$

Que puede también escribirse como

$$P\{P[T^{**} \leq T_0/_{F_n^{*}}] \leq \overline{\beta}/_{F_n}\} = \alpha$$

En la anterior ecuación,

$$Q(\boldsymbol{X}^{*}) = P[T^{**} \leq T_0/_{F_n^{*}}]$$

es aleatoria, dependiendo de F_n.
Una vez calculado beta según dicha ecuación, se elegiría

$$\overline{T}(\boldsymbol{X},\overline{\beta}) = \overline{G}^{-1}(\overline{\beta}) \equiv \overline{T}_{([B+1]\overline{\beta})}$$

Nótese que Q, que depende de un nuevo bootstrap doble, en general puede ser estimado por un nuevo procedimiento de simulación Monte-Carlo.

Es decir,

$$Q(\boldsymbol{X}^*) = \frac{card\{T_i^{**} \leq T_0\}}{B'}$$

siendo B´ el número de simulaciones a utilizar en ese segundo procedimiento.

CAPITULO II

GENERACION DE MUESTRAS BOOTSTRAP CON DETECCION DE MUESTRAS OUTLIERS

2.0. INTRODUCCION. 38

2.1. VARIABILIDADES O ERRORES. 41

2.2. ESTRUCTURA ALGEBRAICA DEL CONJUNTO DE REALIZACIONES DE LAS MUESTRAS BOOTSTRAP. 43

2.2.1. Equivalencia entre el conjunto de realizaciones de las muestras bootstrap y el conjunto de las variaciones con repetición. 43

2.2.2. Composición de variaciones. 47

2.2.3. Relación de equivalencia $\mathfrak{R}$ sobre el conjunto de las variaciones con repetición. 51

2.2.4. Relación de equivalencia sobre el conjunto cociente inducido por $\mathfrak{R}$. . 58

2.3. ESTRUCTURA PROBABILISTICA. 62

2.4. GENERACION DE MUESTRAS BOOTSTRAP. 71

2.0. INTRODUCCION.

Como ya se señaló en el Capítulo I, entre los tres métodos de estimación bootstrap según Efron (1979), es el Método II, basado en el método de simulación, el que suele utilizarse con más frecuencia.

La implementación del correspondiente algoritmo en un sistema informático requiere diversas herramientas, dependiendo del tipo de estimador a utilizar, pero necesitará siempre, como mínimo, de un generador de números aleatorios.

La solución que se obtenga al aplicar el anterior mecanismo puede depender bastante de la calidad del generador empleado, que puede variar mucho entre sistemas distintos, o incluso entre varios generadores presentes en un mismo sistema.

La mala calidad del generador, u otras razones, como pudieran ser la mala inicialización de la semilla necesaria en los algoritmos de generación de números aleatorios, o tal vez el efecto de errores que pueden producirse en los cálculos de ciertos estadísticos, como pueden ser los de redondeo, pueden llevar a resultados inexactos, debido a la mala implementación o ejecución de una técnica, el bootstrap, que ya se ha visto en el Capítulo I que puede ser una gran alternativa en los problemas de estimación estadística.

De hecho, Davison et al. (1986) ponen de manifiesto la necesidad de considerar técnicas de reducción de la varianza en la simulación bootstrap, dado que "el error de simulación puede ser relativamente importante".

A ello debe unirse también el error debido al incremento de la varianza que puede conllevar realizar muestreo aleatorio con reemplazamiento.

Por estas razones, en este capítulo se presenta un método cuyo objetivo es mejorar la calidad de cualquier estimación bootstrap realizada mediante el Método II de Efron (1979), es decir, basada en los métodos de simulación (Monte-Carlo).

En concreto, en este capítulo se presentan un conjunto de técnicas que permitan eliminar alguno ó algunos de los errores antes mencionados. Estas técnicas se basarán en la determinación de las muestras bootstrap que se desvían excesivamente de la muestra original. A estas muestras, en paralelo a la teoría existente sobre las observaciones outliers, se las denominará "muestras bootstrap outliers".

De esta forma, se propone una variante del método II de Efron (1979), estableciendo un filtro en el mecanismo de generación de muestras bootstrap, eliminando aquellas muestras bootstrap que sean declaradas outliers.

2.1. VARIABILIDADES O ERRORES.

Como ya se ha indicado, en la generación de muestras bootstrap inciden entre otros los siguientes dos posibles errores: el error debido al proceso de simulación y el error producido en el proceso de muestreo con reemplazamiento.

Estos errores son los culpables de que las muestras que se generen puedan desviarse de la población base, cuya función de distribución es la función de distribución empírica.

Con intención de formalizar tales errores se presentan a continuación las siguientes definiciones de los tipos de errores o variabilidades que pueden presentarse en un proceso de estimación bootstrap de acuerdo con Muñoz García, J., Moreno-Rebollo, J.L., y Pascual-Acosta, A. (1990):

Variabilidad o error del medio: Es la variabilidad o error que se comete por el procedimiento de simulación en sí. Aquí se recogen los errores debidos a limitaciones o deficiencias en la implementación del proceso de simulación.

Variabilidad o error inherente: Es la variabilidad o error que se produce en el procedimiento bootstrap a causa de realizar el muestreo aleatorio con reemplazamiento, es decir, al recorrer la diversas muestras bootstrap posibles.

En base a estas variabilidades, se puede dar la siguiente definición de lo que se denominará muestras bootstrap outliers.

DEFINICION 2.1.1. Se llamará muestra bootstrap outlier a toda aquella muestra bootstrap generada según el Método II de Efron (1979), y que debido a la variabilidad o error inherente o bien a la variabilidad o error del medio se desvía marcadamente de la muestra original.

Los dos tipos de errores antes definidos estarán presentes en las muestras bootstrap generadas y serán difíciles o prácticamente imposibles distinguirlos, por lo que podría decirse que tal distinción sería materia de una conjetura, en paralelo a lo descrito por Anscombe (1960) para las observaciones outliers.

Con el fin de detectar las muestras bootstrap outliers que pueden generarse al aplicar el método II de Efron (1979), se va a analizar en el siguiente apartado la estructura algebraica que presenta el conjunto de todas las posibles muestras bootstrap, lo que llevará a posteriores estudios probabilísticos que conduzcan a criterios útiles que atenúen errores en los procesos estadísticos que se basan en el bootstrap.

2.2. ESTRUCTURA ALGEBRAICA DEL CONJUNTO DE REALIZACIONES DE LAS MUESTRAS BOOTSTRAP.

2.2.1. Equivalencia entre el conjunto de realizaciones de las muestras bootstrap y el conjunto de las variaciones con repetición.

La realización muestral $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ puede ser considerada como un conjunto finito al que se le puede aplicar la siguiente definición.

DEFINICION 2.2.1.1. (Kolmogorov-Fomin, 1975). Dos conjuntos M y N se dicen equivalentes si entre sus elementos se puede establecer una correspondencia biunívoca.

A partir de esta definición se puede deducir que dos conjuntos finitos son equivalentes si y solo si tienen el mismo número de elementos.

En particular, es posible aplicar esta definición a la muestra inicial de datos mediante una aplicación g biunívoca entre el conjunto de los elementos que forman el vector **x** y el conjunto $\mathbf{N}=\{1,2,\ldots,n\}$, definida de la forma siguiente:

DEFINICION 2.2.1.2. Dada una realización muestral $\mathbf{x}=(x_1, x_2, \ldots, x_n)$, se define la aplicación

$$g: \{x_1, x_2, \ldots, x_n\} \rightarrow \{1, 2, \ldots, n\}$$

de la siguiente forma:

$$g(x_i) = i \quad \forall\, i=1,2,\ldots,n$$

Por tanto, a cada componente de la muestra **x** se le asigna su índice i.

De esta definición puede deducirse la existencia de su inversa g^{-1}, que estará determinada de forma única. En efecto,

$$g^{-1}(i)=x_i,$$

i.e., la observación muestral correspondiente al lugar i-ésimo. Obsérvese que la anterior aplicación es válida cualquiera que sea la dimensión de los x_i.

De este modo, se consigue una transformación biunívoca entre los conjuntos $\{x_1,x_2,\ldots,x_n\}$ y $\{1,2,\ldots,n\}$

La aplicación g así definida es fundamental para todo el desarrollo siguiente, de forma que será posible trabajar directamente con el conjunto **N** de los primeros n números naturales, de tal forma que el análisis del conjunto de posibles muestras bootstrap correspondientes a la muestra original llevará a la misma estructura que el análisis del conjunto de todas las posibles realizaciones de muestras bootstrap sobre el conjunto $\{1,2,\ldots,n\}$.

A continuación se extiende la definición de esta aplicación a otra transformación **g** definida sobre el conjunto de las muestras bootstrap.

El conjunto de variaciones con repetición de los n elementos de **N** tomados de n en n puede ser representado como el producto cartesiano

$$\mathbf{N}^n=\mathbf{N} \times \mathbf{N} \times \ldots \times \mathbf{N}$$

DEFINICION 2.2.1.3. Sobre el conjunto de las muestras bootstrap se define la siguiente aplicación, con conjunto imagen $\mathbf{N}^n$:

$$g[(X^*_1, X^*_2, \ldots, X^*_n)] = (f(X^*_1), f(X^*_2), \ldots, f(X^*_n))$$

siendo f una función definida de forma que $f(X^*_i)=g(x_j)$, para aquel j tal que $X^*_i=x_j$.

Esta aplicación es biunívoca, definiéndose su inversa como sigue:

DEFINICION 2.2.1.4. Dada una variación con repetición $(i_1, i_2, \ldots, i_n)$ de los n elementos de **N**, se define su inversa g^{-1}

$$g^{-1}[(i_1, i_2, \ldots, i_n)] = (f^{-1}(i_1), f^{-1}(i_2), \ldots, f^{-1}(i_n))$$

siendo

$$f^{-1}(i_j) = x_{i_j}$$

LEMA 2.2.1.1. Existe una transformación biunívoca entre el conjunto de todas las posibles realizaciones de muestras bootstrap y el conjunto de las variaciones con repetición de n elementos tomados de n en n.

Demostración:

Dada una muestra bootstrap $\mathbf{X}^*=(X^*_1,X^*_2,\ldots,X^*_n)$, no hay más que aplicarle la transformación **g** definida en 2.2.1.3., para obtener una variación con repetición de $\mathbf{N}^n$.

En el sentido contrario, dada una variación con repetición $(i_1,i_2,\ldots,i_n)$, aplicándole la función inversa de g, g^{-1}, definida en 2.2.1.4, se transforma en una muestra bootstrap. ■

De este modo, el conjunto de posibles muestras bootstrap $\mathbf{X}^*$ se corresponde, aplicando previamente la transformación **g**, y de forma biunívoca, con el conjunto de las variaciones con repetición de n elementos, los primeros n números naturales, tomados de n en n, siendo n^n el número de tales variaciones, que coincide obviamente con el número de muestras bootstrap posibles.

En ese sentido, se puede considerar cada muestra bootstrap como un conjunto de n números, cada uno de ellos entre 1 y n, de forma que se identifica cada x_i por su índice i.

Todas estas variaciones con repetición (tantas como muestras bootstrap) son equiprobables bajo el Método II de Efron (1979), de acuerdo con el procedimiento de extracción, por lo que cada una de ellas tendrá la misma probabilidad de ser extraída, igual a $\frac{1}{n^n}$.

2.2.2. Composición de variaciones.

DEFINICION 2.2.2.1. Sobre el conjunto $\mathbf{N}^n$ se define una ley de composición interna

$$N^n \times N^n : \overset{\circ}{\rightarrow} N^n$$

de la forma siguiente:

Dados $\mathbf{z}=(z_1,z_2,\ldots,z_n)$ y $\mathbf{t}=(t_1,t_2,\ldots,t_n)$, pertenecientes a $\mathbf{N}^n$, se dice que el elemento $\mathbf{r}$ de $\mathbf{N}^n$ es igual a la composición de $\mathbf{z}$ y $\mathbf{t}$, es decir,

$$\mathbf{r}= \mathbf{z} \circ \mathbf{t},$$

si $r_i= \mathbf{z}(t_i)$ para $i=1,2,\ldots,n$, definiendo $\mathbf{z}(t_i)$ como el elemento del vector $\mathbf{z}$ que ocupa el lugar t_i de $\mathbf{z}$. Es decir,

$$\boldsymbol{z}(t_i) = z_{t_i}$$

La anterior definición se ilustra con el siguiente ejemplo: Sea $n=4$, $\mathbf{z}=(1,2,1,3)$ y $\mathbf{t}=(1,3,4,3)$. Se tiene entonces que $\mathbf{z} \circ \mathbf{t} = (1,1,3,1)$, pues:

$\mathbf{z}(t_1)= \mathbf{z}(1) = 1$ (elemento que aparece en el primer lugar de $\mathbf{z}$)
$\mathbf{z}(t_2)= \mathbf{z}(3) = 1$ (elemento que aparece en el tercer lugar de $\mathbf{z}$)
$\mathbf{z}(t_3)= \mathbf{z}(4) = 3$ (elemento que aparece en el cuarto lugar de $\mathbf{z}$)
$\mathbf{z}(t_4)= \mathbf{z}(3) = 1$ (elemento que aparece en el tercer lugar de $\mathbf{z}$)

TEOREMA 2.2.2.1. La ley $\circ$ así definida es efectivamente una ley de composición interna.

Demostración:

Es una ley de composición interna, ya que el resultado de aplicar $\circ$ a dos vectores $\mathbf{z}$, $\mathbf{t}$, es otro vector formado por n elementos de $\mathbf{N}$, y por tanto dicho resultado pertenece a $\mathbf{N}^n$. ■

TEOREMA 2.2.2.2. La ley de composición interna $\circ$ verifica las siguientes propiedades:

i) Es asociativa.

ii) Posee elemento neutro.

Demostración:

i) Es asociativa:

Sean $\mathbf{z}$, $\mathbf{t}$, $\mathbf{y}$ tres elementos de $\mathbf{N}^n$, se trata de demostrar que $\mathbf{r}=\mathbf{s}$, siendo $\mathbf{r}= \mathbf{z} \circ (\mathbf{t} \circ \mathbf{y})$, $\mathbf{s}= (\mathbf{z} \circ \mathbf{t}) \circ \mathbf{y}$.

En efecto:

$$r_i= \mathbf{z}\ ((\mathbf{t} \circ \mathbf{y})_i)= \mathbf{z}\ (\mathbf{t}\ (y_i))$$

La anterior expresión coincide con la componente de $\mathbf{z}$ cuya posición viene dada por la componente de $\mathbf{t}$ que aparece en el lugar y_i.

Por otra parte,

$$s_i = (\mathbf{z} \circ \mathbf{t})\ (\ y_i) = (\mathbf{z}(t_1),\mathbf{z}(t_2),\ldots,\mathbf{z}(t_n))(y_i)$$

Esta expresión será igual al elemento de $\mathbf{z} \circ \mathbf{t}$ que aparece en el lugar y_i, es decir, el elemento de $\mathbf{z}$ que aparece en la posición determinada por el elemento de $\mathbf{t}$ en el lugar y_i, que por tanto será igual a

$$\mathbf{z}\ (\mathbf{t}\ (y_i)) = r_i.$$

ii) El elemento neutro es la variación **e** = (1,2,...,n). Es decir,

$$\mathbf{e} \circ \mathbf{z} = \mathbf{z} \circ \mathbf{e} = \mathbf{z}$$

cualquiera que sea la variación con repetición **z**. En efecto:

$$(\mathbf{e} \circ \mathbf{z})_i = \mathbf{e}(z_i) = z_i$$

$$(\mathbf{z} \circ \mathbf{e})_i = \mathbf{z}(e_i) = \mathbf{z}(i) = z_i.$$

Al ser cierto para todo i=1,2,...,n, queda probada la propiedad ii). ■

Debe observarse, sin embargo, que no se verifica la propiedad conmutativa y carece de elemento simétrico.

Los siguientes contraejemplos permiten comprobar que no se verifican dichas propiedades:

Sea n=4, **z**=(1 2 2 3), **t**=(2 3 2 4).

z ° **t** = (2 2 2 3), **t** ° **z** = (2 3 3 2).

Carece de elemento simétrico:

Elemento simétrico a la derecha: Su existencia implicaría que para toda variación **z** debe existir un elemento **z'** tal que **z** ° **z'**=**e**. Si por ejemplo, n=4, **z** =(2 2 2 2), **z** ° **t** = **z**, distinto de **e**, cualquiera que sea **t**.

Elemento simétrico a la izquierda: De existir, para toda variación **z** debe existir otra variación **z'** tal que **z'** ° **z** = **e**. Si por ejemplo n=4, **z** = (2 2 2 2), cualquiera que fuese **z'** **z'** ° **z** sería un vector cuyas n componentes serían todas iguales a la segunda componente de **z'**, por lo que no se obtendría el vector nulo **e**.

A tenor de las anteriores propiedades, es inmediato el siguiente

TEOREMA 2.2.2.3. El conjunto de las variaciones con repetición de n elementos tomados de n en n, respecto la ley de composición interna $\circ$ definida, se dota de la estructura de semigrupo con elemento neutro.

Esa misma estructura la presenta el conjunto de las posibles muestras bootstrap, considerando la siguiente definición de la ley de composición interna.

DEFINICION 2.2.2.2. Dadas dos muestras bootstrap, X^{*1}, X^{*2}, se dice que la muestra bootstrap Y^* es la composición de ambas, y se escribe $Y^* = X^{*1} \circ X^{*2}$, si $g(Y^*) = g(X^{*1}) \circ g(X^{*2})$.

Es inmediato comprobar que se tiene así una ley de composición interna en el conjunto de las muestras bootstrap verificando las mismas propiedades que las que cumple $\circ$ en el conjunto de las variaciones con repetición. Por tanto, se cumple el siguiente teorema.

TEOREMA 2.2.2.4. El conjunto de las muestras bootstrap se dota de la estructura de semigrupo con elemento neutro respecto la ley de composición interna $\circ$.

Continuando con el estudio de las propiedades del conjunto $\mathbb{N}^n$, una vez que ha sido dotado de una ley de composición interna, se verá seguidamente que también puede definirse en él una relación de equivalencia.

2.2.3. Relación de equivalencia $\mathfrak{R}$ sobre el conjunto de las variaciones con repetición.

DEFINICION 2.2.3.1. Dados dos vectores $z, t \in N^n$, se define la siguiente relación:

$$z \mathfrak{R} t \Leftrightarrow F_n^*(g^{-1}(z)) \equiv F_n^*(g^{-1}(t))$$

Donde g^{-1} es la función definida en el Lema 2.2.1, mientras que $F^*_n()$ denota la función de distribución empírica asociada a la correspondiente muestra bootstrap.

Por tanto dos variaciones con repetición estarán relacionadas si y sólo si la función de distribución empírica es la misma para las respectivas muestras bootstrap correspondientes a ambos puntos de N^n, es decir, las muestras que se obtienen al aplicar g^{-1} a ambos puntos.

LEMA 2.2.3.1. $\mathfrak{R}$ es una relación de equivalencia.

Demostración:

i) Propiedad **reflexiva**. Cualquier vector z está relacionado consigo mismo, puesto que

$$F_n^*(g^{-1}(z)) \equiv F_n^*(g^{-1}(z)) \Rightarrow z \mathfrak{R} z$$

ii) Propiedad **simétrica.**

$$z \,\Re\, t \Rightarrow F_n^*(g^{-1}(z)) \equiv F_n^*(g^{-1}(t)) \Rightarrow t \Re z$$

iii) Propiedad **transitiva.**

$$z \,\Re\, t \leftarrow F_n^*(g^{-1}(z)) \equiv F_n^*(g^{-1}(t))$$

$$t \,\Re\, y \leftarrow F_n^*(g^{-1}(t)) \equiv F_n^*(g^{-1}(y))$$

$$\Rightarrow F_n^*(g^{-1}(z)) \equiv F_n^*(g^{-1}(y)) \Rightarrow z \,\Re\, y$$

Por tanto, queda demostrado que la relación así definida en $\mathbf{N}^n$ es una relación de equivalencia. ■

Esta relación de equivalencia establece en $\mathbf{N}^n$ una partición formada por todas las clases de equivalencia, cada una de las cuales estará constituida por aquellas variaciones de $\mathbf{N}^n$ equivalentes entre sí. Por tanto, los puntos de cada clase de equivalencia se caracterizan por corresponderse con muestras bootstrap que tienen todas ellas la misma función de distribución empírica. De este modo, aplicando $\mathbf{g}^{-1}$, se induce una partición en el conjunto de las muestras bootstrap posibles, de forma que cada elemento de dicha partición la constituyen todas aquellas muestras boostrap que generan una determinada función de distribución empírica.

El conjunto formado por las clases de equivalencia de $\mathbf{N}^n$ inducido por esta relación de equivalencia, es decir, el conjunto cociente, será denotado por $\boldsymbol{N}^n/\Re$.

DEFINICION 2.2.3.1. Se define el conjunto de funciones de Distribución **F** como aquel conjunto de funciones de distribución discretas que verifican las siguientes condiciones:

i) $W_G \subseteq N$, $\forall G \in F$, , siendo W_G el soporte de G.

ii) $\forall G \in F$ y $\forall i \in W_G$, $P_G[i] = \frac{k}{n}$ para algún $k \in \{0,1,2,\ldots,n\}$.

Por tanto, y según lo expuesto anteriormente, existe una aplicación biunívoca entre el conjunto cociente y el conjunto de funciones de distribución **F**:

TEOREMA 2.2.3.1. Existe una transformación biunívoca entre los conjuntos

$$N^n/\Re \text{ y } F$$

Dada la partición realizada en $\mathbf{N}^n$, es posible elegir un solo elemento de cada clase de equivalencia de forma que pueda considerarse su representante.

Definición 2.2.3.2. Dada una clase de equivalencia $C \in \mathbb{N}^n/\mathfrak{R}$ se define como su representante, y se notará $\mathbf{z}_C$, a aquella variación

$$z_C = (z_{C,1}, z_{C,2}, \ldots, z_{C,n}) \in C$$

tal que

$$z_{C,1} \leq z_{C,2} \leq \ldots \leq z_{C,n}$$

es decir, la que está ordenada de menor a mayor.

Esta elección está bien definida, pues al ser los elementos de cada clase de equivalencia vectores de dimensión finita n, siempre existirá un vector que cumpla la anterior propiedad.

Obsérvese que, dado el elemento $\mathbf{z}_C$ representante de una clase de equivalencia, cualquier otra variación **t** perteneciente a la misma clase es una permutación de los elementos de $\mathbf{z}_C$, por lo que todo estadístico T(**X**) invariante ante permutaciones de la muestra toma el mismo valor para cada una de las muestras bootstrap correspondientes a una misma clase de equivalencia de la partición de $\mathbf{N}^n$, por lo que se podría decir también que la anterior partición de clases de equivalencia del conjunto $\mathbf{N}^n$ induce una partición del conjunto imagen del estadístico T.

Nótese en todo caso que es posible que T tome el mismo valor en dos clases de equivalencia distintas, es decir dos funciones de distribución empírica distintas pueden originar el mismo valor de T.

En el siguiente Teorema se establece el cardinal del conjunto cociente $N^n/\mathfrak{R}$. En el Teorema 2.2.3.3. se determinará otra fórmula equivalente para dicho cardinal.

TEOREMA 2.2.3.2.: El número de clases de equivalencia que constituyen el conjunto cociente $N^n/\mathfrak{R}$ viene dado por:

$$\binom{n}{1}\binom{n-1}{0} + \binom{n}{2}\binom{n-1}{1} + \binom{n}{3}\binom{n-1}{2} + \ldots + \binom{n}{n-1}\binom{n-1}{n-2} + \binom{n}{n}\binom{n-1}{n-1}$$

Demostración:

Por la propia definición de la relación de equivalencia $\mathfrak{R}$, cada clase de equivalencia se caracteriza por tener asociada una función de distribución empírica que es la misma para todas las muestras bootstrap correspondientes a las variaciones integrantes de esa clase, y que distingue por tanto dos clases de equivalencias distintas. Por tanto, el número de clases equivalentes será igual al número de las distintas funciones de distribución empírica que pueden construirse con n

elementos, que podemos suponer además que son los primeros n números naturales 1,2,...n, según el comentario hecho en el párrafo previo al Teorema.

Ahora bien, cada una de esas funciones de distribución viene caracterizada en primer lugar por el número de puntos a los que asigna probabilidad no nula. De tal forma que podemos calcular el total de funciones de distribución de la forma siguiente: $C_1F_1+C_2F_2+\ldots+C_nF_n$, siendo C_k el número de combinaciones de n elementos tomados de k en k, es decir, el número combinatorio $\binom{n}{k}$.

F_i, por su parte, es igual al número de funciones de distribución empírica distintas que pueden obtenerse con k números fijados. De esta forma, cada componente del anterior sumatorio es igual al producto del número de funciones de distribución distintas del conjunto **F** que pueden obtenerse con k números multiplicado por el total de conjuntos posibles formado por k elementos, k fijo entre 1 y n.

Por otra parte, fijada una cualquiera de esas combinaciones de n elementos tomadas de k en k, el número de funciones de distribución utilizando esos k números (los k reciben probabilidad no nula) es igual al número de distribuciones distinguibles de n bolas indistinguibles en k celdas, de forma que ninguna queda vacía, el cual se sabe, por Feller (1973) que

es igual al número combinatorio $\binom{n-1}{k-1} = F_i$. Por tanto, el número total de clases de equivalencias es igual a

$$\sum_{k=1}^{n} \binom{n}{k} \binom{n-1}{k-1}$$

tal y como se quería demostrar ■.

La demostración empleada en el anterior Teorema ha permitido describir la estructura del conjunto de las clases de equivalencia. Existe otra forma de evaluar dicho cardinal, teniendo en cuenta que existen tantas clases de equivalencia como funciones de distribución en **F**. El conjunto **F** tiene un cardinal igual al número de combinaciones con repetición de n elementos tomados de n en n, que se sabe es igual a

$$\binom{2n-1}{n}$$

TEOREMA 2.2.3.3. El cardinal del conjunto cociente $\boldsymbol{N}^{n}/\Re$ también puede evaluarse mediante el número combinatorio $\binom{2n-1}{n}$.

2.2.4. Relación de equivalencia sobre el conjunto cociente inducido por $\mathfrak{R}$. .

Definición 2.2.4.1. Sobre el conjunto de clases de equivalencia $N^n/\mathfrak{R}$ se establece la siguiente aplicación h:

Dada una clase de equivalencia **C** perteneciente al conjunto cociente, y siendo z_C su representante elegido según la Definición 2.2.3.3. Se define entonces

$$h(C) = \sum_{i=1}^{n-1} S(z_{C,i+1} - z_{C,i})$$

Donde la función S(u) se define de la forma siguiente:

$$S(u) = \begin{cases} 1 \text{ si } u>0 \\ 0 \text{ si } u=0 \end{cases}$$

Obsérvese que no puede darse el caso <0, puesto que z_C es un representante de su clase de equivalencia elegido de forma que sus elementos están ordenados de forma monótona creciente.

El conjunto imagen de h es el conjunto {0,1,2,...,n-1}, de forma que h(**C**)+1 es igual al número de elementos distintos que forman la variación z_C, que obviamente toma valores entre 1 y n.

También se puede definir la función h conjugada, y que se representará por h^c, definida igualmente sobre el conjunto cociente $N^n/\mathfrak{R}$.

DEFINICION 2.2.4.2. En las condiciones de la Definición 2.2.4.1., se define la función conjugada de h, y se notará por h^c, a la siguiente función:

$$h^c(C) = \sum_{i=1}^{n-1} [1 - S(z_{C,i+1} - z_{C,i})] = (n-1) - h(C)$$

La imagen de h^c vuelve a ser el conjunto {0,1,2,...,n-1}, de forma que $h^c(C)+1$ es igual al número de elementos iguales de la variación z_C.

Es evidente que se cumple la siguiente propiedad:

$$h(C) + h^c(C) = n-1 \quad \forall\, C \in N^n/\Re$$

Utilizando la función h se puede definir en el conjunto cociente una nueva relación de equivalencia.

DEFINICION 2.2.4.3.

$$Dados\ C, D \in N^n/\Re,\quad C\ \Re^1\ D \Leftrightarrow h(C) = h(D)$$

TEOREMA 2.2.4.1. La relación $\Re^1$ es una relación de equivalencia:

Demostración:

i) Propiedad **reflexiva.**

$$h(C) = h(C) \rightarrow C \, \Re^1 \, C$$

ii) Propiedad **simétrica.**

$$C \, \Re^1 \, D \rightarrow h(C) = h(D) \rightarrow D \, \Re^1 \, C$$

iii) Propiedad **transitiva.**

$$C \, \Re^1 \, D \leftrightarrow h(C) = h(D)$$

$$D \, \Re^1 \, E \leftrightarrow h(D) = h(E)$$

$$\rightarrow h(C) = h(E) \rightarrow C \, \Re^1 \, E$$

■

De acuerdo con las definiciones previas, es inmediato el siguiente

TEOREMA 2.2.4.3.

$$Dados \; C, D \in N^n/\Re, \quad C \, \Re^1 \, D \leftrightarrow h^c(C) = h^c(D)$$

Por tanto a partir de esta relación de equivalencia se puede definir un nuevo conjunto cociente, que puede representarse por

$$\chi = (N^n/\Re)/\Re^1$$

y que está formado por tantas clases como posibles valores de h, es decir, n.

Resumiendo el proceso realizado en estos dos últimos apartados, se ha pasado en primer lugar del conjunto de todas las muestras bootstrap posibles al conjunto de clases de equivalencia caracterizadas por tener la misma función de distribución, es decir, al conjunto de posibles funciones de distribución.

A continuación, se han agrupado todas esas posibles funciones de distribución empíricas en clases de equivalencia, caracterizadas por el hecho de que la variable número de observaciones distintas toma el mismo valor en todos los integrantes de una de esas clases de equivalencia.

2.3. ESTRUCTURA PROBABILISTICA.

DEFINICION 2.3.1. Dada una muestra bootstrap $\mathbf{X}^*=(X^*_1,X^*_2,\ldots,X^*_n)$ se define la variable aleatoria $Y=h(\mathbf{C})+1$, siendo $\mathbf{C}$ la clase de equivalencia según $\mathfrak{R}$, correspondiente a $\mathbf{g}(\mathbf{X}^*)$.

Nótese que el carácter aleatorio de Y viene impuesto por la estructura probabilística de la muestra bootstrap.

Por tanto, dada una muestra bootstrap, Y toma un valor igual al número de puntos distintos de que consta dicha muestra bootstrap.

TEOREMA 2.3.1. En las condiciones de la Definición 2.3.1.,

$$Y=k \text{ sii } h(\mathbf{C})=k-1$$

La demostración es inmediata por la definición 2.2.4.1.

TEOREMA 2.3.2. El número de muestras bootstrap para las que la variable Y toma el valor k, viene dado según la siguiente expresión:

$$card\{Y=k\} = \binom{n}{k}\sum_{i=0}^{n}(-1)^i\binom{k}{i}(k-i)^n$$

Demostración:

El cálculo de ese número de muestras puede obtenerse como el producto de dos términos.

El primero vendrá dado por el número de subconjuntos posibles de tamaño k de un conjunto de n elementos, que se sabe es igual al número combinatorio

$$\binom{n}{k}$$

El segundo, fijados k elementos, vendrá dado por el conjunto de posibles muestras bootstrap de tamaño n, utilizando solo esos k elementos. Obsérvese que cada muestra bootstrap de ese tipo equivale a una distribución de n bolas distinguibles en k celdas. El número de distribuciones de n bolas distinguibles en k celdas, Feller (1973), viene dado por

$$\sum_{i=0}^{n} (-1)^i \binom{k}{i} (k-i)^n$$

Multiplicando ambos términos, se tiene la demostración.■

En la anterior demostración, fijadas k observaciones, el número de muestras bootstrap podría haberse evaluado por otra vía, estableciendo un sumatorio sobre todas las posibilidades, que sería

$$\sum_{i_1+\ldots+i_k=n,\ i_j\geq 1} \frac{n!}{i_1!\ldots i_k!}$$

TEOREMA 2.3.3. La probabilidad de que la variable Y tome el valor k viene dado según la siguiente expresión:

$$P\{Y=k\} = \frac{\binom{n}{k}\sum_{i=0}^{n}(-1)^i\binom{k}{i}(k-i)^n}{n^n}$$

Demostración:

Dado que todas las muestras bootstrap son equiprobables, solo hay que dividir el número de casos posibles, que es n^n, entre el número de casos favorables, establecido en el Teorema 2.4.3. ■

A partir de la función de probabilidad, es inmediato el cálculo de la función de distribución:

TEOREMA 2.3.4. La función de distribución correspondiente a la variable aleatoria Y viene dada según la siguiente expresión:

$$P\{Y\leq k\} = \frac{1}{n^n}\sum_{j=1}^{k}\binom{n}{j}\sum_{i=0}^{n}(-1)^i\binom{j}{i}(j-i)^n$$

De esta forma, además de establecer la distribución de probabilidad de la variable Y, teniendo en cuenta la equivalencia demostrada en el Teorema 2.3.1, se tiene también calculada la función de distribución de la función h definida en 2.2.4, puesto que según esa equivalencia

$$P\{h(.)=k-1\} = P\{Y=k\}$$

De este modo, se dispone de la función de distribución de la variable Y. En la tabla 1 se adjunta, para n entre 1 y 50, dicha función de distribución. En la tabla 2 se recoge, para cada valor de n, la media y desviación típica, así como estimadores de los percentiles. En concreto, fijado n, si G representa la función de distribución de Y, para cada valor de α se ha calculado el correspondiente percentil utilizando una definición de inversa de la función de distribución:

$$G^{-1}(\alpha) = inf\{x: G(x) \geq \alpha\}$$

En el Anexo I se encuentra un listado del programa utilizado para calcular dichas tablas.

TABLA 1

Función de distribución de la variable número k de puntos distintos en muestras bootstrap para n entre 1 y 20

k \ n	1	2	3	4	5	6	7	8	9	10
1	1.00000000	0.50000000	0.11111111	0.01562500	0.00160000	0.00012860	0.00000850	0.00000048	0.00000002	0.00000000
2		1.00000000	0.77777778	0.34375000	0.09760000	0.02006173	0.00322145	0.00042439	0.00004741	0.00000460
3			1.00000000	0.90625000	0.57760000	0.25154321	0.07997518	0.01977062	0.00398267	0.00067636
4				1.00000000	0.96160000	0.75308642	0.43696929	0.19010162	0.06463119	0.01786528
5					1.00000000	0.98456790	0.86536222	0.61067200	0.33591039	0.14646088
6						1.00000000	0.99388010	0.93030548	0.74897600	0.49160512
7							1.00000000	0.99759674	0.96534370	0.84722752
8								1.00000000	0.99906334	0.98330752
9									1.00000000	0.99963712
10										1.00000000

k \ n	11	12	13	14	15	16	17	18	19	20
1	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
2	0.00000039	0.00000003	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
3	0.00009929	0.00001284	0.00000148	0.00000016	0.00000001	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
4	0.00414518	0.00082762	0.00014497	0.00002262	0.00000318	0.00000041	0.00000005	0.00000001	0.00000000	0.00000000
5	0.05208823	0.01553115	0.00397365	0.00088904	0.00017663	0.00003156	0.00000512	0.00000076	0.00000010	0.00000001
6	0.26134905	0.11429621	0.04199813	0.01323241	0.00363868	0.00088639	0.00019369	0.00003836	0.00000695	0.00000116
7	0.63435580	0.39517731	0.20520284	0.09001983	0.03391191	0.01114433	0.00323979	0.00084331	0.00019858	0.00004268
8	0.91136958	0.75115406	0.53066492	0.31788900	0.16226740	0.07139348	0.02743026	0.00931982	0.00283221	0.00077753
9	0.99216527	0.95060150	0.83863378	0.65361583	0.44069040	0.25610686	0.12904473	0.05692967	0.02222643	0.00775919
10	0.99986009	0.99640054	0.97338547	0.89968462	0.75580903	0.56133198	0.36436167	0.20670203	0.10309913	0.04559527
11	1.00000000	0.99994628	0.99837578	0.98607047	0.93989810	0.83484043	0.67015361	0.47760850	0.30044785	0.16713207
12		1.00000000	0.99997944	0.99927822	0.99288966	0.96514030	0.89238410	0.76153555	0.58673531	0.40364973
13			1.00000000	0.99999215	0.99968345	0.99644874	0.98035172	0.93219247	0.83374156	0.68478321
14				1.00000000	0.99999701	0.99986276	0.99825992	0.98920333	0.95854662	0.88788172
15					1.00000000	0.99999887	0.99994109	0.99916168	0.99420024	0.97533943
16						1.00000000	0.99999957	0.99997494	0.99960215	0.99694718
17							1.00000000	0.99999984	0.99998942	0.99981372
18								1.00000000	0.99999994	0.99999557
19									1.00000000	0.99999998
20										1.00000000

TABLA 1 (Continuación)

Función de distribución del número k de puntos distintos en muestras bootstrap para n entre 21 y 30

k \ n	21	22	23	24	25	26	27	28	29	30
1	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
2	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
3	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
4	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
5	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
6	0.00000018	0.00000003	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
7	0.00000844	0.00000154	0.00000026	0.00000004	0.00000001	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
8	0.00019452	0.00004468	0.00000949	0.00000187	0.00000035	0.00000006	0.00000001	0.00000000	0.00000000	0.00000000
9	0.00244496	0.00070128	0.00018445	0.00004478	0.00001009	0.00000212	0.00000042	0.00000008	0.00000001	0.00000000
10	0.01803993	0.00644101	0.00209188	0.00062244	0.00017078	0.00004345	0.00001030	0.00000229	0.00000048	0.00000009
11	0.08268649	0.03664944	0.01466471	0.00533640	0.00177810	0.00054588	0.00015527	0.00004113	0.00001019	0.00000237
12	0.24734712	0.13537433	0.06653003	0.02954807	0.01193799	0.00441551	0.00150401	0.00047435	0.00013921	0.00003819
13	0.50866705	0.33946659	0.20344201	0.10983152	0.05367846	0.02388420	0.00973095	0.00365039	0.00126742	0.00040924
14	0.76771469	0.60851120	0.43759513	0.28444548	0.16724825	0.08924401	0.04341289	0.01934899	0.00794140	0.00301617
15	0.92666511	0.83426280	0.69796998	0.53547825	0.37416218	0.23768717	0.13746611	0.07261735	0.03518373	0.01570537
16	0.98568694	0.95335566	0.88524696	0.77404324	0.62762180	0.46742720	0.31836917	0.19820098	0.11298651	0.05916453
17	0.99842225	0.99187665	0.97108308	0.92273516	0.83580171	0.71001086	0.55906782	0.40530809	0.26983480	0.16501117
18	0.99991384	0.99919800	0.99548270	0.98249017	0.94930702	0.88390258	0.78036268	0.64468134	0.49400047	0.34949475
19	0.99999815	0.99996058	0.99959843	0.99753437	0.98962452	0.96753057	0.91999470	0.83799672	0.72110993	0.58012717
20	0.99999999	0.99999924	0.99998214	0.99980167	0.99867701	0.99397348	0.97966186	0.94617793	0.88347345	0.78658719
21	1.00000000	1.00000000	0.99999969	0.99999199	0.99990328	0.99930119	0.99656349	0.98752206	0.96459925	0.91815356
22		1.00000000	1.00000000	0.99999987	0.99999643	0.99995337	0.99963619	0.99807356	0.99249076	0.97720196
23			1.00000000	1.00000000	0.99999995	0.99999842	0.99997776	0.99981312	0.99893705	0.99556145
24				1.00000000	1.00000000	0.99999998	0.99999931	0.99998950	0.99990519	0.99942207
25					1.00000000	1.00000000	0.99999999	0.99999970	0.99999509	0.99995245
26						1.00000000	1.00000000	1.00000000	0.99999987	0.99999772
27							1.00000000	1.00000000	1.00000000	0.99999994
28								1.00000000	1.00000000	1.00000000
29									1.00000000	1.00000000
30										1.00000000

TABLA 1 (Continuación)

Función de distribución del número k de puntos distintos en muestras bootstrap para n entre 31 y 40

k \ n	31	32	33	34	35	36	37	38	39	40
1	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
2	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
3	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
4	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
5	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
6	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
7	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
8	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
9	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
10	0.00000002	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
11	0.00000052	0.00000011	0.00000002	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
12	0.00000983	0.00000238	0.00000055	0.00000012	0.00000002	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
13	0.00012343	0.00003491	0.00000929	0.00000234	0.00000056	0.00000013	0.00000003	0.00000001	0.00000000	0.00000000
14	0.00106495	0.00035105	0.00010846	0.00003152	0.00000865	0.00000224	0.00000055	0.00000013	0.00000003	0.00000001
15	0.00648804	0.00249128	0.00089280	0.00029976	0.00009462	0.00002817	0.00000794	0.00000212	0.00000054	0.00000013
16	0.02856680	0.01276967	0.00530597	0.00205733	0.00074713	0.00025500	0.00008205	0.00002497	0.00000720	0.00000198
17	0.09287746	0.04826106	0.02323207	0.01039846	0.00434326	0.00169882	0.00062432	0.00021625	0.00007080	0.00002197
18	0.22797184	0.13721115	0.07636319	0.03940988	0.01892105	0.00847903	0.00355824	0.00140278	0.00052110	0.00018292
19	0.43350141	0.29996645	0.19210224	0.11398734	0.06280182	0.03221432	0.01543015	0.00692233	0.00291739	0.00115838
20	0.66010651	0.51793017	0.37816348	0.25644859	0.16152922	0.09462573	0.05166426	0.02635692	0.01259818	0.00565765
21	0.84061954	0.73141512	0.59914389	0.45920718	0.32821077	0.21851646	0.13558035	0.07850921	0.04251553	0.02158301
22	0.94378858	0.88370629	0.79266993	0.67419355	0.53967271	0.40467210	0.28360357	0.18566855	0.11363135	0.06511006
23	0.98560557	0.96220349	0.91700163	0.84352035	0.74103808	0.61647460	0.48279207	0.35471105	0.24412385	0.15737650
24	0.99742013	0.99107847	0.97508819	0.94199644	0.88442470	0.79858603	0.68716108	0.55957539	0.42927458	0.30945674
25	0.99969007	0.99852381	0.99456591	0.98388690	0.96026467	0.91638395	0.84659775	0.75006571	0.63238796	0.50454908
26	0.99997641	0.99983590	0.99916765	0.99674378	0.98976152	0.97328947	0.94068883	0.88550258	0.80432339	0.69917536
27	0.99999895	0.99998841	0.99991416	0.99953711	0.99807858	0.99360257	0.98236415	0.95871405	0.91618429	0.84978156
28	0.99999998	0.99999952	0.99999436	0.99995559	0.99974588	0.99888253	0.99606548	0.98855225	0.97177259	0.93977638
29	1.00000000	0.99999999	0.99999978	0.99999728	0.99997727	0.99986218	0.99935891	0.99761621	0.99268807	0.98102808
30	1.00000000	1.00000000	1.00000000	0.99999990	0.99999870	0.99998848	0.99992611	0.99963693	0.99857609	0.99540074
31	1.00000000	1.00000000	1.00000000	1.00000000	0.99999996	0.99999938	0.99999422	0.99996081	0.99979687	0.99916082
32		1.00000000	1.00000000	1.00000000	1.00000000	0.99999998	0.99999971	0.99999712	0.99997943	0.99988766
33			1.00000000	1.00000000	1.00000000	1.00000000	0.99999999	0.99999986	0.99999858	0.99998931
34				1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	0.99999994	0.99999930
35					1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	0.99999997
36						1.00000000	1.00000000	1.00000000	1.00000000	1.00000000
37							1.00000000	1.00000000	1.00000000	1.00000000
38								1.00000000	1.00000000	1.00000000
39									1.00000000	1.00000000
40										1.00000000

TABLA 1 (Continuación)

Función de distribución del número k de puntos distintos en muestras bootstrap para n entre 41 y 50

k \ n	41	42	43	44	45	46	47	48	49	50
1	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
2	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
3	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
4	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
5	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
6	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
7	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
8	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
9	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
10	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
11	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
12	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
13	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
14	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
15	0.00000003	0.00000001	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
16	0.00000052	0.00000013	0.00000003	0.00000001	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000	0.00000000
17	0.00000648	0.00000182	0.00000049	0.00000013	0.00000003	0.00000001	0.00000000	0.00000000	0.00000000	0.00000000
18	0.00006084	0.00001922	0.00000578	0.00000166	0.00000045	0.00000012	0.00000003	0.00000001	0.00000000	0.00000000
19	0.00043453	0.00015439	0.00005209	0.00001672	0.00000512	0.00000150	0.00000042	0.00000011	0.00000003	0.00000001
20	0.00239370	0.00095665	0.00036207	0.00013008	0.00004446	0.00001449	0.00000451	0.00000134	0.00000038	0.00000011
21	0.01029701	0.00462866	0.00196533	0.00079015	0.00030150	0.00010942	0.00003785	0.00001251	0.00000395	0.00000120
22	0.03499850	0.01768779	0.00842440	0.00379027	0.00161464	0.00065272	0.00025093	0.00009193	0.00003215	0.00001076
23	0.09511739	0.05398050	0.02882025	0.01450624	0.00689852	0.00310635	0.00132730	0.00053928	0.00020876	0.00007714
24	0.20944166	0.13311773	0.07953674	0.04474278	0.02374065	0.01190513	0.00565367	0.00254779	0.00109169	0.00044563
25	0.37960969	0.26885521	0.17916465	0.11239567	0.06644961	0.03707960	0.01956285	0.00977666	0.00463698	0.00209116
26	0.57790835	0.45218719	0.33405399	0.23272201	0.15287374	0.09475112	0.05547438	0.03072530	0.01612554	0.00803353
27	0.75856741	0.64709149	0.52471532	0.40304120	0.29268105	0.20078737	0.13014793	0.07976784	0.04628251	0.02545802
28	0.88684786	0.80987755	0.71036610	0.59488574	0.47359416	0.35745440	0.25542679	0.17273130	0.11058061	0.06707403
29	0.95749289	0.91631407	0.85302225	0.76659973	0.66074892	0.54348506	0.42512933	0.31559763	0.22212865	0.14820979
30	0.98745488	0.97050536	0.93918778	0.88839206	0.81524863	0.72083659	0.61068010	0.49365289	0.37972269	0.27750455
31	0.99714884	0.99183239	0.97986582	0.95655146	0.91670491	0.85628442	0.77420954	0.67349174	0.56101923	0.44598641
32	0.99951169	0.99825681	0.99476052	0.98646857	0.96945801	0.93886572	0.89008352	0.82043955	0.73067062	0.62543251
33	0.99993855	0.99971928	0.99894816	0.99668599	0.99104081	0.97886310	0.95584786	0.91730334	0.85954251	0.78143629
34	0.99999449	0.99996673	0.99984047	0.99937323	0.99793190	0.99415222	0.98558940	0.96860374	0.93876383	0.89188271
35	0.99999966	0.99999719	0.99998217	0.99991032	0.99963095	0.99872589	0.99623487	0.99031536	0.97800545	0.95534674
36	0.99999999	0.99999984	0.99999858	0.99999053	0.99995011	0.99978516	0.99922463	0.99760730	0.99358053	0.98481172
37	1.00000000	0.99999999	0.99999992	0.99999928	0.99999502	0.99997252	0.99987629	0.99953365	0.99849839	0.99580076
38	1.00000000	1.00000000	1.00000000	0.99999996	0.99999964	0.99999740	0.99998501	0.99992950	0.99972264	0.99906886
39	1.00000000	1.00000000	1.00000000	1.00000000	0.99999998	0.99999982	0.99999865	0.99999190	0.99996022	0.99983680
40	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	0.99999999	0.99999991	0.99999931	0.99999566	0.99997776
41	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	0.99999996	0.99999965	0.99999769
42		1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	0.99999998	0.99999982
43			1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	0.99999999
44				1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000
45					1.00000000	1.00000000	1.00000000	1.00000000	1.00000000	1.00000000
46						1.00000000	1.00000000	1.00000000	1.00000000	1.00000000
47							1.00000000	1.00000000	1.00000000	1.00000000
48								1.00000000	1.00000000	1.00000000
49									1.00000000	1.00000000
50										1.00000000

TABLA 2

Medidas estadísticas de la distribución de la variable número de puntos distintos en m. bootstrap

			Percentiles																			
n	Media	D.típica	.01	.05	.10.	15	.20	.25	.30	.35	.40	.45	.50	.55	.60	.65	.70	.75	.80	.85	.90	.95
1	1.00000	0.00000	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
2	1.50000	0.50000	1	1	1	1	1	1	1	1	1	1	1	2	2	2	2	2	2	2	2	2
3	2.11111	0.56656	1	1	1	2	2	2	2	2	2	2	2	2	2	2	2	2	3	3	3	3
4	2.73438	0.64329	1	2	2	2	2	2	2	3	3	3	3	3	3	3	3	3	3	3	3	4
5	3.36160	0.71361	2	2	3	3	3	3	3	3	3	3	3	3	4	4	4	4	4	4	4	4
6	3.99061	0.77813	2	3	3	3	3	3	4	4	4	4	4	4	4	4	4	4	5	5	5	5
7	4.62058	0.83792	3	3	4	4	4	4	4	4	4	5	5	5	5	5	5	5	5	5	6	6
8	5.25113	0.89381	3	4	4	4	5	5	5	5	5	5	5	5	5	6	6	6	6	6	6	7
9	5.88205	0.94647	4	4	5	5	5	5	5	6	6	6	6	6	6	6	6	7	7	7	7	7
10	6.51322	0.99639	4	5	5	6	6	6	6	6	6	6	7	7	7	7	7	7	7	8	8	8
11	7.14457	1.04395	5	5	6	6	6	6	7	7	7	7	7	7	7	8	8	8	8	8	8	9
12	7.77605	1.08945	5	6	6	7	7	7	7	7	8	8	8	8	8	8	8	8	9	9	9	9
13	8.40764	1.13313	6	7	7	7	7	8	8	8	8	8	8	9	9	9	9	9	9	10	10	10
14	9.03931	1.17520	6	7	8	8	8	8	8	9	9	9	9	9	9	9	10	10	10	10	11	11
15	9.67103	1.21582	7	8	8	8	9	9	9	9	9	10	10	10	10	10	10	10	11	11	11	12
16	10.30281	1.25513	7	8	9	9	9	9	10	10	10	10	10	10	11	11	11	11	11	12	12	12
17	10.93463	1.29325	8	9	9	10	10	10	10	10	11	11	11	11	11	11	12	12	12	12	13	13
18	11.56649	1.33029	9	9	10	10	10	11	11	11	11	11	12	12	12	12	12	12	13	13	13	14
19	12.19837	1.36632	9	10	10	11	11	11	11	12	12	12	12	12	13	13	13	13	13	14	14	14
20	12.83028	1.40142	10	11	11	11	12	12	12	12	12	13	13	13	13	13	14	14	14	14	15	15
21	13.46221	1.43567	10	11	12	12	12	13	13	13	13	13	13	14	14	14	14	14	15	15	15	16
22	14.09416	1.46913	11	12	12	13	13	13	13	14	14	14	14	14	14	15	15	15	15	16	16	16
23	14.72612	1.50183	11	12	13	13	13	14	14	14	14	15	15	15	15	15	16	16	16	16	17	17
24	15.35809	1.53385	12	13	13	14	14	14	15	15	15	15	15	16	16	16	16	16	17	17	17	18
25	15.99008	1.56521	12	13	14	14	15	15	15	15	16	16	16	16	16	17	17	17	17	18	18	19
26	16.62208	1.59595	13	14	15	15	15	16	16	16	16	16	17	17	17	17	17	18	18	18	19	19
27	17.25409	1.62611	14	15	15	16	16	16	16	17	17	17	17	17	18	18	18	18	19	19	19	20
28	17.88610	1.65573	14	15	16	16	17	17	17	17	17	18	18	18	18	19	19	19	19	20	20	21
29	18.51813	1.68482	15	16	16	17	17	17	18	18	18	18	19	19	19	19	19	20	20	20	21	21
30	19.15015	1.71342	15	16	17	17	18	18	18	19	19	19	19	19	20	20	20	20	21	21	21	22
31	19.78219	1.74155	16	17	18	18	18	19	19	19	19	20	20	20	20	20	21	21	21	22	22	23
32	20.41423	1.76924	16	18	18	19	19	19	20	20	20	20	20	21	21	21	21	22	22	22	23	23
33	21.04628	1.79649	17	18	19	19	20	20	20	20	21	21	21	21	22	22	22	22	23	23	23	24
34	21.67833	1.82334	17	19	19	20	20	20	21	21	21	21	22	22	22	22	23	23	23	24	24	25
35	22.31038	1.84981	18	19	20	20	21	21	21	22	22	22	22	23	23	23	23	24	24	24	25	25
36	22.94244	1.87589	19	20	21	21	21	22	22	22	22	23	23	23	23	24	24	24	25	25	25	26
37	23.57450	1.90162	19	20	21	22	22	22	23	23	23	23	24	24	24	24	25	25	25	26	26	27
38	24.20656	1.92701	20	21	22	22	23	23	23	23	24	24	24	24	25	25	25	25	26	26	27	27
39	24.83863	1.95207	20	22	22	23	23	24	24	24	24	25	25	25	25	26	26	26	26	27	27	28
40	25.47070	1.97681	21	22	23	23	24	24	24	25	25	25	25	26	26	26	27	27	27	28	28	29
41	26.10278	2.00124	21	23	24	24	24	25	25	25	26	26	26	26	27	27	27	27	28	28	29	29
42	26.73485	2.02538	22	23	24	25	25	25	26	26	26	26	27	27	27	28	28	28	28	29	29	30
43	27.36693	2.04924	23	24	25	25	26	26	26	27	27	27	27	28	28	28	28	29	29	29	30	31
44	27.99901	2.07282	23	25	25	26	26	27	27	27	27	28	28	28	29	29	29	29	30	30	31	31
45	28.63109	2.09613	24	25	26	26	27	27	28	28	28	28	29	29	29	29	30	30	30	31	31	32
46	29.26317	2.11919	24	26	27	27	27	28	28	28	29	29	29	30	30	30	30	31	31	31	32	33
47	29.89525	2.14200	25	26	27	28	28	28	29	29	29	30	30	30	30	31	31	31	32	32	33	33
48	30.52734	2.16457	26	27	28	28	29	29	29	30	30	30	31	31	31	31	32	32	32	33	33	34
49	31.15943	2.18691	26	28	28	29	29	30	30	30	31	31	31	31	32	32	32	33	33	33	34	35
50	31.79152	2.20902	27	28	29	30	30	30	31	31	31	32	32	32	32	33	33	33	34	34	35	35

2.4. GENERACION DE MUESTRAS BOOTSTRAP.

Utilizando la función de distribución tabulada en el apartado 2.3, se propone el siguiente algoritmo de generación de muestras bootstrap, y que puede considerarse variante del método II de Efron (1979).

DEFINICION 2.4.1. Dada una realización muestral $x_1, x_2, \ldots, x_n$, se define el siguiente método de generación de muestras bootstrap de tamaño n, que se llamará "Generación de muestras bootstrap con detección de muestras outliers"

Para i=1 hasta B, repetir el siguiente procedimiento:

i) Calcular una muestra bootstrap $(X^*_1, X^*_2, \ldots, X^*_n)$ obtenida por muestreo con reemplazamiento de la muestra original.

ii) Calcular para dicha muestra el valor de la variable aleatoria Y, sea k su valor.

iii) Si $P[Y \leq k] < \alpha$, pasar al siguiente valor de i,

olvidando la muestra generada. En caso contrario, la muestra se considera válida.

De esta forma, no consideran las posibles B muestras bootstrap, sino solo un número $B_1 < B$, las correspondientes a aquellas muestras boostrap que se consideran, desde el punto de vista de la distribución de Y, que no se desvían excesivamente del resto de muestras bootstrap.

Las muestras que están siendo rechazadas, bajo tal criterio, se declaran "muestras bootstrap outliers", y no se tienen en cuenta en los análisis estadísticos a realizar.

El valor de α, en concordancia con los valores habitualmente empleados en los tests de significación, puede tomarse igual a 0.01 ó 0.05.

A la hora de la práctica, la comparación realizada en iii) puede hacerse de la forma siguiente:

$$Y \leq Y_\alpha$$

siendo Y_α el mayor valor tal que $P\{Y \leq Y_\alpha\} \leq \alpha$.

En la tabla 3 aparecen diversos valores de este punto crítico.

TABLA 3

Puntos críticos propuestos en el algoritmo de muestreo bootstrap con detección de muestras bootstrap outliers.

	Valor de α		
n	**.01**	**.05**	**.10**
1	-	-	-
2	-	-	-
3	-	-	-
4	1	1	1
5	1	1	2
6	1	2	2
7	2	2	3
8	2	3	3
9	3	3	4
10	3	4	4
11	4	4	5
12	4	5	5
13	5	6	6
14	5	6	7
15	6	7	7
16	6	7	8
17	7	8	8
18	8	8	9
19	8	9	9
20	9	10	10
21	9	10	11
22	10	11	11
23	10	11	12
24	11	12	12
25	11	12	13
26	12	13	14
27	13	13	14
28	13	14	15
29	14	15	15
30	15	15	16
31	15	16	17
32	15	17	17
33	16	17	18
34	16	17	18
35	17	18	19
36	18	20	20
37	18	19	20
38	19	20	21
39	19	21	21
40	20	21	22
41	20	22	23
42	21	22	23
43	22	23	24
44	22	24	24
45	23	24	25
46	23	25	26
47	24	25	26
48	25	26	27
49	25	27	27
50	26	27	28

En las tablas 4.1 a 4.6 aparecen diversas simulaciones donde se compara este método bootstrap con detección de muestras bootstrap outliers con el método II de Efron (1979). Los tiempos de CPU están medidos en segundos.

Se han utilizado los puntos críticos para $\alpha = 0.01,\ 0.05\ y\ 0.10$.

TABLA 4.1.

Comparación entre el bootstrap y el bootstrap con detección de muestras outliers, para los datos de las Facultades de Derecho, según Efron (1979). n=15, Estadístico: Coeficiente de correlación lineal. Puntos críticos: 6,7 y 8. Valor del coeficiente de correlación en la población completa: 0.761.

B	Método		Media	Desviación típica	CPU
100	Bootstrap		0.76958	0.16735	0.18
		6	0.75565	0.16241	0.15
	P.crítico	7	0.75083	0.16208	0.14
		8	0.75112	0.15941	0.15
500	Bootstrap		0.77071	0.12987	0.44
		6	0.76360	0.12915	0.41
	P.crítico	7	0.76312	0.12655	0.41
		8	0.76340	0.12356	0.30
1000	Bootstrap		0.77344	0.13984	0.97
		6	0.76930	0.13950	0.81
	P.crítico	7	0.76812	0.13756	0.86
		8	0.77078	0.13328	0.66
2000	Bootstrap		0.76923	0.13398	2.44
		6	0.76812	0.13240	2.60
	P.crítico	7	0.76827	0.13037	2.29
		8	0.76974	0.12580	1.95

Para estos mismos datos, con B=1000, se calculó un Intervalo de Confianza al 90%, método percentil, para dicho coeficiente de correlación lineal. Se obtuvo:

		Extr.inf.	Extr.Sup.
Bootstrap		0.53015	0.94639
	6	0.51103	0.95291
P.crítico	7	0.51363	0.95060
	8	0.51754	0.94978

TABLA 4.2.

Se analiza a continuación el conjunto de datos de Mardia et al. (1979), referidos a 88 estudiantes, para cada uno de los cuales se tiene cinco variables, dos referidas a exámenes con libro cerrado y tres con libro abierto. Se ha estudiado el mayor autovalor de la matriz de correlaciones muestrales. La última columna contiene el tiempo de CPU consumido.

B	Método		Media	Desviación típica	CPU
100	Bootstrap		4.69505	0.06578	0.86
		48	4.69578	0.06547	0.80
	P.crítico	49	4.69509	0.06475	0.79
		50	4.69406	0.06457	0.77
200	Bootstrap		4.69509	0.06566	1.46
		48	4.69601	0.06339	1.64
	P.crítico	49	4.69602	0.06266	1.46
		50	4.69556	0.06258	1.38
500	Bootstrap		4.69787	0.06667	3.39
		48	4.69883	0.06337	3.32
	P.crítico	49	4.69871	0.06333	3.40
		50	4.69869	0.06322	3.34
1000	Bootstrap		4.69641	0.06496	13.04
		48	4.69644	0.06288	12.98
	P.crítico	49	4.69677	0.06274	12.92
		50	4.69666	0.06265	12.92
2000	Bootstrap		4.69755	0.06419	27.98
		48	4.69506	0.06375	27.73
	P.crítico	49	4.69541	0.06353	27.70
		50	4.69579	0.06333	27.11

TABLA 4.3.

En la siguiente simulación, se han analizado 250 muestras de tamaño 10, extraídas de una Normal(0,1). Se utilizó para ello la rutina RNNOA de IMSL.

Para cada una de las 250 muestras, utilizando el bootstrap y el bootstrap con detección de muestras outliers, se calculó la media de las B ó B_1 medias, así como la longitud de un Intervalo de Confianza al 90%. Para cada intervalo de confianza se calculó también la cobertura del parámetro bajo estudio, 0 en este caso, entendiendo por cobertura proporción de muestras bootstrap para las que el Intervalo de Confianza contiene a dicho parámetro. A continuación, se calculó la media y desviación típica de tales cantidades, sobre esas 250 muestras.

		Media		Intervalo de confianza			
B	**Método**	**Media**	**D.T.**	**Long.**	**D.T.**	**Cobert.**	**CPU**
500	Boots.	-0.00366	0.30804	0.98990	16.03570	0.880	81.52
	3	-0.00405	0.30713	0.98517	15.95892	0.864	84.10
	P.C. 4	-0.00424	0.30724	0.97407	15.77863	0.864	83.39
	5	-0.00482	0.30690	0.92624	15.00003	0.832	70.39
1000	Boots.	-0.00339	0.30800	0.98604	15.96673	0.876	149.47
	3	-0.00427	0.30682	0.98706	15.98811	0.876	149.53
	P.C. 4	-0.00433	0.30702	0.97758	15.83341	0.868	148.06
	5	-0.00437	0.30092	0.92646	15.00101	0.848	117.90

TABLA 4.4.

Se repite el experimento de la tabla 4.3., con la única diferencia de que en este caso las 250 muestras de la normal N(0,1) son de tamaño 20.

		Media		Intervalo de confianza			
B	**Método**	**Media**	**D.T.**	**Long.**	**D.T.**	**Cobert.**	**CPU**
500	Boots.	-0.00035	0.22492	0.70561	11.30332	0.856	63.09
	9	-0.00019	0.22364	0.70111	11.23425	0.856	62.51
	P.C. 10	-0.00006	0.22370	0.69377	11.11815	0.848	61.73
	11	-0.00003	0.22406	0.67576	10.83174	0.840	53.73
1000	Boots.	-0.00095	0.22432	0.70467	11.28508	0.864	336.22
	9	-0.00117	0.22444	0.70243	11.25063	0.872	335.20
	P.C. 10	-0.00122	0.22447	0.69403	11.11607	0.864	310.71
	11	-0.00112	0.22475	0.67361	10.78660	0.844	285.78

TABLA 4.5.

Se repite el experimento de las tablas 4.3. y tablas 4.4, con la única diferencia de que en este caso la 250 muestras de la normal N(0,1) son de tamaño 25.

	Media			Intervalo de confianza			
B	**Método**	**Media**	**D.T.**	**Long.**	**D.T.**	**Cobert.**	**CPU**
500	Boots.	0.02285	0.20715	0.63890	10.17978	0.864	68.10
	12	0.02399	0.20857	0.63468	10.11110	0.856	69.43
	P.C. 13	0.02406	0.20880	0.62911	10.02197	0.856	67.47
	10	0.02419	0.20873	0.61275	9.76146	0.844	63.34
1000	Boots.	0.02406	0.20735	0.63992	10.18952	0.856	195.25
	12	0.02318	0.20776	0.64021	10.19986	0.864	201.25
	P.C. 13	0.02320	0.20780	0.63246	10.07708	0.860	186.72
	14	0.02306	0.20965	0.61743	9.83881	0.844	171.69

TABLA 4.6.

En este caso se realiza un análisis sobre 250 muestras de tamaño 20 de una Exponencial Negativa de media 1, generadas mediante la rutina de IMSL denominada RNEXP.

		Media		Intervalo de confianza			
B	**Método**	**Media**	**D.T.**	**Long.**	**D.T.**	**Cobert.**	**CPU**
500	Boots.	0.97407	0.22059	0.66266	10.84339	0.824	111.97
	9	0.97426	0.22098	0.65840	10.77758	0.820	113.75
	P.C. 10	0.97428	0.22091	0.65214	10.67619	0.820	107.67
	11	0.97453	0.22145	0.63375	10.38004	0.812	100.78
1000	Boots.	0.97433	0.22150	0.66418	10.88725	0.828	343.50
	9	0.97501	0.22229	0.66336	10.86795	0.820	320.85
	P.C. 10	0.97490	0.22208	0.65589	10.74635	0.820	311.09
	11	0.97502	0.22222	0.63701	10.44332	0.808	270.22

Comentario de las tablas anteriores:

TABLA 4.1.:

Se observa que la media del coeficiente de correlación, para el método bootstrap con detección de muestras outliers, se acerca más al valor verdadero del coeficiente de correlación lineal que el bootstrap utilizando el Método II de Efron. La desviación típica de las medias simuladas se hace inferior también. En general, es el método bootstrap con detección de muestras bootstrap outliers, empleando un punto crítico con $\alpha = 0.05$ el que da mejor resultado.

A medida que aumenta el número de simulaciones, la diferencia entre ambos métodos disminuye, lo que puede sugerir que el método bootstrap con detección de muestras outliers puede tener efectos más notorios con número de simulaciones no demasiado alto.

Nótese que los tiempos de CPU suelen ser inferior en relación al método II de Efron (1979), si bien existen diversas fluctuaciones debido al carácter multiusuario de los equipos donde se realizaron las pruebas.

TABLA 4.2.

Se observa que la media del mayor autovalor muestral da resultados parecidos, si bien el método bootstrap con detección de muestras bootstrap outliers presenta menor desviación típica.

Sí se produce una disminución en el tiempo de CPU, a favor del método bootstrap propuesto. Esta disminución es más efectiva con $\alpha = 0.05$ ó 0.01.

TABLAS 4.3.,4.4. Y 4.5.

El análisis de medias y desviaciones típicas ofrece conclusiones análogas a las tablas anteriores.

Respecto a los Intervalos de Confianza, se obtienen en general intervalos cuya longitud media es menor, con menor desviación típica y presentando una cobertura similar a la ofrecida por el bootstrap.

Los resultados más satisfactorios se obtienen con $\alpha = 0.01$ *y* 0.05. , ya que para 0.1 las coberturas disminuyen acusadamente.

TABLA 4.6.

En este caso, aunque los datos analizados se distribuyan según una exponencial, los resultados son totalmente análogos al caso de la distribución Normal.

CAPITULO III

METODO DE SUAVIZACION BOOTSTRAP.

3.0. INTRODUCCION. 85

3.1. FUNCION DE DISTRIBUCION SUAVIZADA PARA DISTRIBUCIONES DEL TIPO I,II O III. 91

3.1.1. Función de Distribución empírica. 91

3.1.2. Función de Distribución suavizada para funciones de distribución F(x) del tipo I. 94

3.1.3. Función de Distribución suavizada para funciones de distribución F(x) del tipo II. 98

3.1.3. Función de Distribución suavizada para funciones de distribución F(x) del tipo III. 102

3.4. PROPIEDADES DE LAS FUNCIONES DE DISTRIBUCION SUAVIZADAS. 106

3.3. METODO BOOTSTRAP BASADO EN LAS FUNCIONES DE DISTRIBUCION SUAVIZADAS. 113

3.0. INTRODUCCION.

En el capítulo II de esta memoria se ha estudiado la estructura del conjunto de las posibles muestras bootstrap generadas al implementar el método II de Efron (1979). En particular, se ha establecido la distribución de la variable aleatoria que contabiliza el número de observaciones repetidas de que consta una muestra bootstrap, a partir de la cual se ha propuesto un método orientado a la detección de muestras bootstrap outliers desde el punto de vista de dicha distribución.

Una de las características que fundamentan el proceso de muestreo bootstrap es el hecho de muestrear con reemplazamiento en una población de n elementos.

Así, todas las observaciones de una muestra bootstrap habrán salido del mismo conjunto inicial de valores, de forma que, citando a Silverman y Young (1987), "prácticamente casi cualquier muestra bootstrap generada por el método II de Efron (1979) contendrá valores repetidos".

De hecho, consultando la tabla I que aparece en el capítulo II de esta memoria, se puede ver que a partir de n=7 la probabilidad de que una muestra bootstrap conste de n observaciones distintas se hace inferior a 0.01, y a partir de n=15 es inferior a 0.00001.

Con el fin de evitar la problemática que plantea la repetición de valores, en las primeras publicaciones realizadas sobre el bootstrap se incorpora la posibilidad de modificar el mecanismo de simulación bootstrap, de forma que no se realice el muestreo a partir de la función de distribución empírica F_n, sino a partir de otra función de distribución que sea una "suavización" de $F_n(x)$. Esta posibilidad se hace especialmente atractiva en el caso de que la realización muestral inicial $\mathbf{x}=(\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_n)$ haya sido extraída de una población continua.

La función de Distribución suavizada puede construirse de diversas maneras, así Efron (1982) propone la convolución de la función de distribución empírica con una función de Distribución Normal, y en particular existen diversos estudios para el caso de que esa función de Distribución suavizada corresponda a una estimación no paramétrica del tipo de densidad utilizando el método del núcleo.

En Silverman y Young (1987) se analizan posibles ventajas de la utilización de funciones de distribución obtenidas a partir del uso de métodos de estimación no paramétrica basados en el método del núcleo. Para el caso de estadísticos que puedan considerarse funcionales lineales, o que tengan una buena aproximación lineal, llegan a la conclusión de que el error cuadrático medio del estimador correspondiente puede hacerse inferior al que se obtendría por el método bootstrap habitual tomando un valor adecuado del parámetro ventana de la correspondiente función núcleo.

El inconveniente del anterior método de suavización radica en la dificultad práctica que suele surgir, no solo a la hora de elegir función núcleo, sino en el método a seguir para calcular el valor de la ventana.

En este capítulo se proponen otros métodos de estimación bootstrap basado en funciones de distribución suavizadas, pero que no utilizan el método núcleo de estimación de funciones de densidad.

En concreto, estas funciones de distribución que se proponen se pueden considerar dentro del método general de estimación de la función de densidad por histogramas, y en concreto en el método de estimación por histogramas con particiones aleatorias.

Una ventaja respecto a la función núcleo radica en la no necesidad de estimación de un parámetro ventana, calculándose de forma inmediata la forma de la función de distribución suavizada a partir de las observaciones que componen la muestra inicial, lo que también resulta una ventaja frente a métodos más generales de estimación como puede ser el de la función núcleo variable, donde se estiman ventanas distintas para cada observación.

Los métodos de suavización que van a proponerse a continuación serán de aplicación al caso de que F(x) tenga un soporte del tipo (a,b) donde al menos uno de los extremos de dicho intervalo será finito.

Por tanto, los métodos propuestos podrán aplicarse a situaciones donde F(x) tenga un soporte de alguno de los siguientes tipos:

$$(a,b)\,,\ con\ a> -\infty\ y\ b< +\infty$$

$$(a,+\infty)\,,\ a > -\infty$$

$$(-\infty,b)\,,\ b < +\infty$$

Para formalizar las referencias a los anteriores tipos de función de distribución, se realizan las siguientes definiciones:

DEFINICION 3.0.1. Se dirá que una Función de Distribución continua es del tipo I si su soporte es de la forma

$$(a,b)\,,\ con\ a> -\infty\ y\ b< +\infty$$

DEFINICION 3.0.2. Se dirá que una Función de Distribución continua es del tipo II si su soporte es de la forma

$$(a,+\infty)\,,\ a > -\infty$$

DEFINICION 3.0.3. Se dirá que una Función de Distribución continua es del tipo III si su soporte es de la forma

$$(-\infty, b), \quad b < +\infty$$

Para cada una de esas tres posibles situaciones se propondrá una función de distribución suavizada, si bien las tres propuestas presentarán una gran analogía en su planteamiento y propiedades.

3.1. FUNCION DE DISTRIBUCION SUAVIZADA PARA DISTRIBUCIONES DEL TIPO I, II O III.

3.1.1. Función de distribución empírica.

Dadas $X_1, X_2, \ldots, X_n$, independientes e idénticamente distribuidas según $F(x)$ continua univariante, se define la función de distribución empírica, cualquiera que sea el soporte de $F(x)$:

$$F_n(x) = \begin{cases} 0 & \text{si } x < X_{(1)} \\ k/n & \text{si } X_{(k)} \leqslant x < X_{(k+1)} \\ 1 & \text{si } X_{(n)} \leqslant x \end{cases}$$

Se sabe que esta función de distribución empírica converge en probabilidad y casi seguro a la función de distribución poblacional $F(x)$, y por el Teorema de Glivenko-Cantelli, dicha convergencia se verifica también en sentido uniforme.

Esta es una de las razones fundamentales por las cuales se desarrolla la metodología bootstrap, sustituyendo la población inicial $F(x)$ por la población $F_n(x)$, y realizando a continuación el proceso de muestreo bootstrap en esta segunda población.

El método que se propone a continuación sustituye $F_n(x)$ por otra función de distribución, suavizada, cuya expresión será estudiada seguidamente.

La muestra original $X_1,X_2,\ldots,X_n$ pertenece a una población cuyo soporte completo se puede representar por un intervalo (a,b) donde al menos uno de los extremos es finito.

A la hora de la práctica, si bien es muy improbable que este soporte coincida con el que exhibe la muestra, en general sí puede ser conocido aunque sea de forma aproximada, por ejemplo a través de la experiencia de otras situaciones análogas, o el aporte de conocimientos de expertos en el caso de estudios en Biología, Sociología, o cualquier otro campo de investigación.

De esta forma, a partir de de la muestra $X_1,X_2,\ldots,X_n$, se construye una muestra "ampliada", cuya forma dependerá del tipo de F(x):

i) Para F(x) del tipo I: $(a,X_1,X_2,\ldots,X_n,b)$

ii) Para F(x) del tipo II: $(a,X_1,X_2,\ldots,X_n)$

iii) Para F(x) del tipo III: $(X_1,X_2,\ldots,X_n,b)$

Para cada una de tales muestras se puede representar la función de distribución empírica utilizando la siguiente función:

$$H_n(x)=\begin{cases} 0 & \text{si } X_{(0)} < x < X_{(1)} \\ k/n & \text{si } X_{(k)} \leqslant x < X_{(k+1)} \\ 1 & \text{si } X_{(n)} \leqslant x < X_{(n+1)} \end{cases}$$

Siendo

$$X_{(0)} = a, \quad X_{(n+1)} = b, , \forall\, n$$

Nótese que se asume la posibilidad de que $X_{(0)}$ ó $X_{(n+1)}$ sean

$$-\infty \text{ ó } +\infty$$

El método que aquí se propone va a obtener muestras bootstrap donde los valores posibles no van a ser solo los de la muestra inicial $(X_1, X_2, \ldots, X_n)$, sino que podrá obtenerse cualquier observación perteneciente a un intervalo que, dependiendo del tipo de soporte de F(x), será de alguna de las tres formas siguientes:

$$(a,b), \ (a, X_{(n)}) \text{ ó } (X_{(1)}, b)$$

De esta forma se evitará el problema de las observaciones repetidas en una muestra bootstrap, pues al realizar el muestreo bootstrap a partir de la función de distribución suavizada se va a cumplir que

$$P[X_i^* = X_j^*] = 0, \ \forall\, i,j=1,2,\ldots,n, \ i \neq j$$

De hecho, dado que la función de distribución suavizada va a ser de tipo continuo, para cualquier x,

$$P[X_i^* = x] = 0, \ \forall\, i=1,2,\ldots,n$$

3.1.2. Función de Distribución suavizada para funciones de distribución F(x) del tipo I.

En primer lugar, se calculan los puntos medios correspondientes a los diversos intervalos que se obtienen a partir de la muestra inicial, de forma que se tiene la siguiente

DEFINICION 3.1.2.1. Dada una muestra aleatoria simple extraída de una población con Función de Distribución F(x) del tipo I, se definen los puntos $Y_0, Y_1, \ldots, Y_n$ de la forma siguiente:

$$Y_i = \frac{X_{(i)} + X_{(i+1)}}{2}, \quad i = 1, 2, \ldots, n-1.$$

$$Y_0 = X_{(0)}, \quad Y_n = X_{(n+1)}$$

Se tiene así el conjunto de n+1 puntos formado por $Y_0, Y_1, \ldots, Y_n$, ordenados de menor a mayor, siendo posible definir los n intervalos que generan.

DEFINICION 3.1.2.2. Bajo las condiciones de la Definición 3.1.2., se definen n intervalos $I_1, \ldots, I_n$, siendo

$$I_j = [Y_{j-1}, Y_j), \quad j = 1, 2, \ldots, n$$

Estos intervalos verifican de forma evidente las siguientes propiedades:

$$\bigcap_{j=1}^{n} I_j = \varnothing, \quad \bigcup_{j=1}^{n} I_j = [X_{(0)}, X_{(n+1)})$$

Por tanto, el conjunto de los intervalos $\{I_j\}$ es una partición del soporte $(X_{(0)}, X_{(n+1)})$.

Una vez establecidas las anteriores definiciones, a continuación se define la función de Distribución $S_n(x)$ suavizada.

DEFINICION 3.1.2.3. Dada una muestra aleatoria simple $X_1, X_2, \ldots, X_n$ extraida de una población con función de distribución $F(x)$ univariante del tipo I, con soporte $(X_{(0)}, X_{(n+1)})$ y siendo los puntos Y_j e intervalos I_j los definidos en 3.1.2.1. y 3.1.2.2., se define la función de distribución $S_n(x)$ con soporte $(X_{(0)}, X_{(n+1)})$.

$$S_n(x) = \begin{cases} 0 & si\ x \le Y_0 \\ \frac{j-1}{n} + \frac{1}{n}\frac{x-Y_{j-1}}{Y_j-Y_{j-1}}, & si\ Y_{j-1} \le x < Y_j\ (x \in I_j) \\ 1 & si\ x \ge Y_n \end{cases}$$

Nótese que en los puntos medios Y_j se cumple que

$$S_n(Y_j) = \frac{j}{n} \quad , j=0,1,\ldots,n$$

Por tanto, la función de distribución empírica y la función de distribución suavizada coinciden en dichos puntos, i.e.,

$$S_n(Y_j) = F_n(Y_j)$$

TEOREMA 3.1.2.1. La función $S_n(x)$ definida en 3.1.2.1. es efectivamente una función de Distribución.

Demostración:

i) La función es 0 para $x < Y_0$, y vale 1 para $x > Y_n$. Por tanto,

$$\lim_{x\to-\infty} S_n(x) = 0, \quad \lim_{x\to+\infty} S_n(x) = 1$$

ii) Es una función monótona no decreciente. En efecto, pues dados dos valores x_1 y x_2, $x_1 < x_2$, puede ocurrir:

a) Pertenecen a intervalos distintos. Supongamos que

$$x_1 \in I_j, \; x_2 \in I_{j'}, \; j<j'$$

En tal caso,

$$S_n(x_1) = \frac{j-1}{n} + \frac{1}{n}\frac{x_1-Y_{j-1}}{Y_j-Y_{j-1}} \leq \frac{j-1}{n} + \frac{1}{n} = \frac{j}{n} \leq$$

$$\leq \frac{j'-1}{n} \leq \frac{j'-1}{n} + \frac{1}{n}\frac{x_2-Y_{j'-1}}{Y_{j'}-Y_{j'-1}} = S_n(x_2)$$

b) x_1 y x_2 pertenecen al mismo intervalo I_j.

$$S_n(x_1) = \frac{j-1}{n} + \frac{1}{n}\frac{x_1-Y_{j-1}}{Y_j-Y_{j-1}} \leq \frac{j-1}{n} + \frac{1}{n}\frac{x_2-Y_{j-1}}{Y_j-Y_{j-1}} = S_n(x_2)$$

iii) Es continua a la derecha.

Así es, ya que en el interior de cada intervalo I_j, la forma de la función de distribución coincide con una recta, que es función continua.

De hecho, es continua a la izquierda también: en el interior de cada intervalo, al ser una recta, lo es, y en los puntos Y_j,

$$S_n(Y_j) = \frac{j}{n} + \frac{1}{n}\frac{Y_j-Y_j}{Y_{j+1}-Y_j} = \frac{j}{n} = \lim_{x \to Y_j^-} \frac{j-1}{n} + \frac{1}{n}\frac{x-Y_{j-1}}{Y_j-Y_{j-1}}$$

La anterior continuidad se da también, de forma obvia, en los puntos extremos Y_0 e Y_{n+1}. ■

TEOREMA 3.1.2.2. La función de densidad correspondiente a la función de distribución $S_n(x)$ viene dada por la siguiente expresión:

$$f(x) = \frac{1}{n(Y_j-Y_{j-1})}, \quad si\ Y_{j-1} \leq x < Y_j$$

siendo 0 fuera del intervalo $[Y_0, Y_n]$.

Demostración:

Esta expresión es inmediata, simplemente derivando la función de distribución $S_n(x)$. ■

3.1.3. Función de Distribución suavizada para funciones de distribución F(x) del tipo II.

El método propuesto en el apartado anterior suponía que la función de Distribución F(x) de la cual se extrae la muestra inicial toma valores en un soporte finito (a,b).

En este apartado se define una función de distribución suavizada análoga a la anterior, pero con campo de aplicación al caso de que F(x) tenga un soporte del tipo $(a,+\infty)$.

Dada la muestra $(X_1,X_2,\ldots,X_n)$ se construye la muestra ordenada que incluya a X_0, y que será

$$(X_{(0)},X_{(1)},\ldots,X_{(n)})$$

siendo $X_{(0)}=a$.

DEFINICION 3.1.3.1. En las condiciones anteriores, se definen n intervalos $I_1,\ldots,I_n$, siendo

$$I_j = [X_{(j-1)},X_{(j)}),\ j=1,2,\ldots,n$$

Estos intervalos verifican de forma evidente las siguientes propiedades:

$$\bigcap_{j=1}^{n} I_j = \varnothing,\quad \bigcup_{j=1}^{n} I_j = [X_{(0)},X_{(n)})$$

Por tanto, el conjunto de los intervalos $\{I_j\}$ es una partición del soporte $(X_{(0)},X_{(n)})$.

Una vez establecidas las anteriores definiciones, a continuación se presenta la función de Distribución suavizada.

DEFINICION 3.1.3.2. Dada una muestra aleatoria simple $X_1,X_2,\ldots,X_n$ extraida de una población con función de distribución F(x) univariante del tipo II, y dado $X_{(0)}$ y los intervalos I_j según se ha señalado previamente, se define la función de distribución $L_n(x)$ sobre el soporte $(X_{(0)},X_{(n)})$.

$$L_n(x)=\begin{cases} 0 & si\ x \le X_{(0)} \\ \dfrac{j-1}{n}+\dfrac{1}{n}\dfrac{x-X_{(j-1)}}{X_{(j)}-X_{(j-1)}}, & si\ X_{(j-1)} \le x < X_{(j)}\ \ (x \in I_j) \\ 1 & si\ x \ge X_{(n)} \end{cases}$$

A partir de la definición, se observa que

$$L_n(X_{(j)}) = \frac{j}{n} = F_n(X_{(j)})$$

Por tanto, la función de distribución empírica y la función de distribución suavizada coincide en dichos puntos.

TEOREMA 3.1.3.1. La función $L_n(x)$ es efectivamente una función de Distribución.

Demostración:

i) La función es 0 para $x < X_0$, y vale 1 para $x > X_{(n)}$. Por tanto,

$$\lim_{x\to-\infty} L_n(x)=0, \quad \lim_{x\to+\infty} L_n(x)=1$$

ii) Es una función monótona no decreciente. En efecto, pues dadas dos valores x_1 y x_2, $x_1 < x_2$, puede ocurrir:

a) Pertenecen a intervalos distintos. Supongamos que

$$x_1 \in I_j, \; x_2 \in I_{j'}, \; j<j'$$

En tal caso,

$$L_n(x_1) = \frac{j-1}{n} + \frac{1}{n}\frac{x_1-X_{(j-1)}}{X_{(j)}-X_{(j-1)}} \leq \frac{j-1}{n}+\frac{1}{n} = \frac{j}{n} \leq$$

$$\leq \frac{j'-1}{n} \leq \frac{j'-1}{n} + \frac{1}{n}\frac{x_2-X_{(j'-1)}}{X_{(j')}-X_{(j'-1)}} = L_n(x_2)$$

b) x_1 y x_2 pertenecen al mismo intervalo I_j.

$$L_n(x_1) = \frac{j-1}{n}+\frac{1}{n}\frac{x_1-X_{(j-1)}}{X_{(j)}-X_{(j-1)}} \leq \frac{j-1}{n}+\frac{1}{n}\frac{x_2-X_{(j-1)}}{X_{(j)}-X_{(j-1)}} = L_n(x_2)$$

iii) Es continua a la derecha.

Así es, ya que en el interior de cada intervalo I_j, la forma de la función de distribución coincide con una recta, que es función continua.

De hecho, es continua a la izquierda también: en el interior de cada intervalo, al ser una recta, lo es, y en los puntos $X_{(j)}$,

$$L_n(X_{(j)}) = \frac{j}{n}+\frac{1}{n}\frac{X_{(j)}-X_{(j)}}{X_{(j+1)}-X_{(j)}} = \frac{j}{n}=\lim_{x\to X_{(j)}^-}\frac{j-1}{n}+\frac{1}{n}\frac{x-X_{(j-1)}}{X_{(j)}-X_{(j-1)}}$$

La anterior continuidad se da también, de forma obvia, en los puntos extremos X_0 y $X_{(n)}$. ■

TEOREMA 3.1.3.2. La función de densidad correspondiente a la función de distribución $L_n(x)$ viene dada por la siguiente expresión:

$$f(x) = \frac{1}{n(X_{(j)}-X_{(j-1)})}, \quad si\ X_{(j-1)}\le x<X_{(j)}$$

siendo 0 fuera del intervalo $(X_0,X_{(n)})$.

Demostración:

No hay más que derivar la expresión de la función de distribución $L_n(x)$. ■

3.1.4. Función de Distribución suavizada para funciones de distribución F(x) del tipo III.

En este apartado se supone que F(x) tiene un soporte del tipo $(-\infty, b)$.

Dada la muestra $(X_1, X_2, \ldots, X_n)$ se construye la muestra ordenada que incluya a X_{n+1}, y que será

$$(X_{(1)}, X_{(2)}, \ldots, X_{(n+1)})$$

siendo $X_{(n+1)}=b$.

DEFINICION 3.1.4.1. En las condiciones anteriores, se definen n intervalos $I_1, \ldots, I_n$, siendo

$$I_j = [X_{(j)}, X_{(j+1)}), \; j=1,2,\ldots,n$$

Estos intervalos verifican de forma evidente las siguientes propiedades:

$$\bigcap_{j=1}^{n} I_j = \varnothing, \quad \bigcup_{j=1}^{n} I_j = [X_{(1)}, X_{(n+1)})$$

Por tanto, el conjunto de los intervalos $\{I_j\}$ es una partición del soporte $(X_{(1)}, X_{n+1})$.

Una vez establecidas las anteriores definiciones, a continuación se presenta la función de Distribución suavizada.

DEFINICION 3.1.4.2. Dada una muestra aleatoria simple $X_1, X_2, \ldots, X_n$ extraida de una población con función de distribución F(x) univariante, y dado X_{n+1} y los intervalos I_j según se ha definido previamente, se define la función de distribución $U_n(x)$ sobre el soporte $(X_{(1)}, X_{(n+1)})$.

$$U_n(x) = \begin{cases} 0 & si\ x \le X_{(1)} \\ \frac{j-1}{n} + \frac{1}{n}\frac{x - X_{(j)}}{X_{(j+1)} - X_{(j)}}, & si\ X_{(j)} \le x < X_{(j+1)}\ \ (x \in I_j) \\ 1 & si\ x \ge X_{(n+1)} \end{cases}$$

TEOREMA 3.1.4.1. La función $U_n(x)$ es efectivamente una función de Distribución.

Demostración:

i) La función es 0 para $x < X_{(1)}$, y vale 1 para $x > X_{(n+1)}$. Por tanto,

$$\lim_{x \to -\infty} U_n(x) = 0, \quad \lim_{x \to +\infty} U_n(x) = 1$$

ii) Es una función monótona no decreciente. En efecto, pues dadas dos valores x_1 y x_2, $x_1 < x_2$, puede ocurrir:

a) Pertenecen a intervalos distintos. Supongamos que

$$x_1 \in I_j,\ x_2 \in I_{j'},\ j < j'$$

En tal caso,

$$U_n(x_1) = \frac{j-1}{n} + \frac{1}{n}\frac{x_1 - X_{(j)}}{X_{(j+1)} - X_{(j)}} \leq \frac{j-1}{n} + \frac{1}{n} = \frac{j}{n} \leq$$

$$\leq \frac{j'-1}{n} \leq \frac{j'-1}{n} + \frac{1}{n}\frac{x_2 - X_{(j')}}{X_{(j'+1)} - X_{(j')}} = U_n(x_2)$$

b) x_1 y x_2 pertenecen al mismo intervalo I_j.

$$U_n(x_1) = \frac{j-1}{n} + \frac{1}{n}\frac{x_1 - X_{(j)}}{X_{(j+1)} - X_{(j)}} \leq \frac{j-1}{n} + \frac{1}{n}\frac{x_2 - X_{(j)}}{X_{(j+1)} - X_{(j)}} = U_n(x_2)$$

iii) Es continua a la derecha.

Así es, ya que en el interior de cada intervalo I_j, la forma de la función de distribución coincide con una recta, que es función continua.

De hecho, es continua a la izquierda también: en el interior de cada intervalo, al ser una recta, lo es, y en los puntos $X_{(j)}$,

$$U_n(X_{(j)}) = \frac{j-1}{n} + \frac{1}{n}\frac{X_{(j)} - X_{(j)}}{X_{(j+1)} - X_{(j)}} = \frac{j-1}{n} = \lim_{x \to X_{(j)}^-} \frac{j-2}{n} + \frac{1}{n}\frac{x - X_{(j-1)}}{X_{(j)} - X_{(j-1)}}$$

La anterior continuidad se da también, de forma obvia, en los puntos extremos $X_{(1)}$ y $X_{(n+1)}$. ■

TEOREMA 3.1.4.2. La función de densidad correspondiente a la función de distribución $U_n(x)$ viene dada por la siguiente expresión:

$$f(x) = \frac{1}{n(X_{(j+1)} - X_{(j)})}, \quad si\ X_{(j)} \leq x < X_{(j+1)}$$

siendo 0 fuera del intervalo $(X_{(1)}, X_{(n+1)})$.

Demostración:

Al igual que en los casos anteriores, se obtiene simplemente derivando $U_n(x)$. ■

3.2. PROPIEDADES DE LAS FUNCIONES DE DISTRIBUCION SUAVIZADAS.

En este capítulo se estudian las propiedades que verifican las funciones de distribución suavizadas, observando que verifica las mismas convergencias a F(x) que la función de distribución empírica $F_n(x)$.

Notación: En lo que sigue, $R_n(x)$ denotará a cualquiera de las funciones de distribución suavizadas previamente definidas, i.e., $S_n(x)$, $L_n(x)$ o $U_n(x)$. De esa forma cualquier propiedad que se enuncie para $R_n(x)$ se entenderá que es de aplicación para las tres funciones de distribución suavizadas.

Nótese que la sucesión de funciones de distribución

$$\{R_n(x)\}_{n\geq 1}$$

puede considerarse una sucesión de variables aleatorias, por lo que le es de aplicación los conceptos de convergencias.

TEOREMA 3.2.1. Dada una m.a.s. $X_1, X_2, \ldots, X_n$ i.i.d. según F(x) del tipo I, II ó III, la diferencia entre las funciones de distribución $F_n(x)$ y $R_n(x)$ no supera 1/n en valor absoluto, i.e.,

$$|R_n(x) - F_n(x)| \leq \frac{1}{n}, \quad \forall\, x \in \mathbb{R}$$

Demostración:

a) $S_n(x)$:

Nótese en primer lugar, que si x es inferior a Y_0, ambas funciones de distribución toman el mismo valor, 0. Análogamente, si x es superior a Y_n, ambas toman el valor 1, por lo que se anula también la diferencia.

Supóngase pues, que x pertenece al intervalo I_j, es decir,

$$Y_{j-1} \leq x < Y_j$$

En este intervalo, la función de distribución empírica toma el siguiente valor:

$$F_n(x) = \frac{j-1}{n} \text{ si } x < X_{(j)}, \quad \frac{j}{n} \text{ si } x \geq X_{(j)}$$

En el primer caso,

$$|S_n(x) - F_n(x)| = |\frac{j-1}{n} + \frac{1}{n}\frac{x-Y_{j-1}}{Y_j-Y_{j-1}} - \frac{j-1}{n}| < \frac{1}{n}$$

En el segundo caso,

$$|S_n(x) - F_n(x)| = |\frac{j-1}{n} + \frac{1}{n}\frac{x-Y_{j-1}}{Y_j-Y_{j-1}} - \frac{j}{n}| = |\frac{1}{n}(\frac{x-Y_{j-1}}{Y_j-Y_{j-1}} - 1)| \leq \frac{1}{n}$$

b) $L_n(x)$:

Al igual que en el caso a), si x es inferior a X_0, ambas funciones de distribución toman el mismo valor, 0. Análogamente, si x es superior a $X_{(n)}$, ambas toman el valor 1, por lo que se anula también la diferencia.

Sea entonces x perteneciente al intervalo I_j, es decir,

$$X_{(j-1)} \leq x < X_{(j)}$$

En este intervalo, la función de distribución empírica toma el siguiente valor:

$$F_n(x) = \frac{j-1}{n}$$

$$|L_n(x) - F_n(x)| = |\frac{j-1}{n} + \frac{1}{n}\frac{x - X_{(j-1)}}{X_{(j)} - X_{(j-1)}} - \frac{(j-1)}{n}| < \frac{1}{n}$$

c) $U_n(x)$:

Análogamente al caso anterior, si x es inferior a $X_{(1)}$, ambas funciones de distribución toman el mismo valor, 0. Si x es superior a $X_{(n+1)}$, ambas toman el valor 1, por lo que se anula también la diferencia.

Considérese entonces x perteneciente al intervalo I_j, es decir,

$$X_{(j)} \leq x < X_{(j+1)}$$

En este intervalo, la función de distribución empírica toma el siguiente valor:

$$F_n(x) = \frac{j}{n}$$

$$|L_n(x) - F_n(x)| = |\frac{j-1}{n} + \frac{1}{n}\frac{x - X_{(j)}}{X_{(j+1)} - X_{(j)}} - \frac{j}{n}| < \frac{1}{n}$$

■

TEOREMA 3.2.2.

$$R_n(x) - F_n(x) \xrightarrow{P} 0, \quad n \to \infty$$

Demostración:

Se trata de demostrar que

$$\forall\, \varepsilon > 0, \lim_{n\to\infty} P[|R_n(x) - F_n(x)|] > \varepsilon] = 0$$

y puesto que

$$\forall\, \varepsilon > 0, \exists\, n_0 / \forall\, n > n_0, \frac{1}{n} < \varepsilon$$

se tendrá entonces que el suceso

$$\{ |R_n(x) - F(x)| > \varepsilon \}$$

tendrá probabilidad nula $\forall\, n > n_0$. ■

Los anteriores resultados van a permitir demostrar diversas propiedades de $R_n(x)$, en particular su convergencia a $F(x)$, función de distribución de la población original de la que se tomó la muestra $X_1, X_2, \ldots, X_n$.

TEOREMA 3.2.3.

$$R_n(x) \xrightarrow{P} F(x), \quad n \to \infty$$

Demostración:

Por la definición de convergencia en probabilidad, se trata de demostrar que

$$\lim_{n\to\infty} P[\,|R_n(x) - F(x)| > \varepsilon] = 0, \quad \forall\, \varepsilon > 0$$

En efecto:

$$P[|R_n(x) - F(x)| > \varepsilon] = P[|R_n(x) - F_n(x) + F_n(x) - F(x)| > \varepsilon] \leq$$

$$\leq P[|R_n(x) - F_n(x)| + |F_n(x) - F(x)| > \varepsilon] \leq$$

Para demostrar que el límite del primer término es 0, nótese que

$$\forall\, \varepsilon > 0\ \exists\ n_0 \ /\ \forall\, n > n_0,\ \frac{1}{n} < \frac{1}{n_0} < \frac{\varepsilon}{2}$$

por lo que la probabilidad del primer término se anula $\forall\, n > n_0$.

En cuanto al segundo término, debido al resultado conocido de que la función de distribución empírica converge en probabilidad a F(x), se tendrá que

$$\exists\ n_1 \ /\ \forall n > n_1,\ P[\,|F_n(x) - F(x)| > \frac{\varepsilon}{2}] < \frac{\varepsilon'}{2}$$

Tomando $N=\max(n_0, n_1)$, entonces

$$\forall\, \varepsilon' > 0,\ \exists\ N/\ \forall\, n > N, P[\,|R_n(x) - F(x)| > \varepsilon] < \varepsilon'$$

■

TEOREMA 3.2.4.

$$R_n(x) \xrightarrow{c.s.} F(x), \quad cuando\ n\to\infty$$

Demostración:

Por la condición necesaria y suficiente de la convergencia casi seguro, se trata de demostrar que

$$\forall\ \varepsilon>0,\ \lim_{m\to\infty} P[\sup_{m\geq n} |R_m(x)-F(x)|>\varepsilon]=0$$

$$P[\sup_{m\geq n} |R_m(x)-F(x)|>\varepsilon] = P[\sup_{m\geq n} |R_m(x)-F_m(x)+F_m(x)-F(x)|>\varepsilon] \leq$$

$$\leq P[\sup_{m\geq n} |R_m(x)-F_m(x)|>\frac{\varepsilon}{2}] + P[\sup_{m\geq n} |F_m(x)-F(x)|>\frac{\varepsilon}{2}] \leq$$

Aplicando el mismo razonamiento del Teorema anterior, para el primer sumando, y teniendo en cuenta que $F_n(x) \xrightarrow{c.s.} F(x)$,

se tendrá pues que

$$\forall\ \varepsilon' > 0,\ \exists\ N/\ \forall\ n>N,\ P[\sup_{m\geq n} |R_m(x)-F(x)|>\varepsilon] <\varepsilon'$$

■

TEOREMA 3.2.5. $R_n(x)$ converge uniformemente a F(x), i.e.,

$$\sup_{-\infty<x<\infty} |R_n(x) - F(x)| \xrightarrow{c.s} 0 \text{ cuando } n\to\infty$$

Demostración:

$$\forall\, x,\ |R_n(x) - F(x)| \le |R_n(x) - F_n(x)| + |F_n(x) - F(x)| \le$$

$$\le \frac{1}{n} + |F_n(x) - F(x)|$$

Por tanto,

$$\sup_{-\infty<x<+\infty} \{\, |R_n(x) - F(x)| \,\} \le \frac{1}{n} + \sup_{-\infty<x<+\infty} \{|F_n(x) - F(x)|\}$$

El primer término converge a 0. El segundo también, ya que la función de distribución empírica $F_n(x)$ converge uniformemente a F(x).■

TEOREMA 3.2.6.

$$R_n(x) \xrightarrow{L} N(F(x),\ \frac{F(x)\,(1-F(x))}{n}),\quad n\to\infty$$

Demostración:

Se sabe que que $F_n(x)$ converge en Ley a una distribución de este tipo, en concreto

$$N(F(x),\ \frac{F(x)\,(1-F(x))}{n})$$

Por otra parte, según el Teorema 3.2.2.,

$$R_n(x) - F_n(x) \xrightarrow{p} 0$$

De ambos resultados, se obtiene el Teorema. ■

3.3. METODO BOOTSTRAP BASADO EN LAS FUNCIONES DE DISTRIBUCION SUAVIZADAS.

Con vistas a la simulación de muestras bootstrap es conveniente calcular la inversa de las funciones de distribución suavizadas definidas anteriormente.

TEOREMA 3.3.1. Las funciones inversas de las funciones de distribución suavizadas $S_n(x)$, $L_n(x)$ y $U_n(x)$ definidas en el apartado 3.1., vienen dadas por las siguientes expresiones:

$$S_n^{-1}(u) = Y_{j-1} + n\left(u - \frac{j-1}{n}\right)(Y_j - Y_{j-1}),\ si\ \frac{j-1}{n} \le u < \frac{j}{n}$$

$$L_n^{-1}(u) = X_{(j-1)} + n\left(u - \frac{j-1}{n}\right)(X_{(j)} - X_{(j-1)}),\ si\ \frac{j-1}{n} \le u < \frac{j}{n}$$

$$U_n^{-1}(u) = X_{(j)} + n\left(u - \frac{j-1}{n}\right)(X_{(j+1)} - X_{(j)}),\ si\ \frac{j-1}{n} \le u < \frac{j}{n}$$

Demostración:

Obsérvese que en cada intervalo I_j, cualquiera de las funciones de distribución suavizada consideradas toma un recorrido de valores comprendidos entre

$$\frac{j-1}{n}\ y\ \frac{j}{n}.$$

Por tanto dado u perteneciente al intervalo (0,1), su imagen inversa estará en aquel intervalo I_j tal que

$$\frac{j-1}{n} \leq u < \frac{j}{n}$$

El siguiente paso es calcular la inversa de la función de distribución en tal intervalo. ■

Nótese que en la fórmula anterior el valor de j puede calcularse directamente como [nu]+1, siendo [] el operador parte entera, lo que facilita la implementación práctica del cálculo de la inversa.

A partir del anterior Teorema, se propone a continuación el siguiente algoritmo de extracción de muestras bootstrap, todas de tamaño n, a partir de la función de distribución suavizada correspondiente al tipo de F(x).

Este algoritmo se basa en el método II de Efron (1979) de estimación bootstrap, con la diferencia de haber sustituido la función de distribución empírica por la función de distribución suavizada.

Al igual que en el apartado 3.2., $R_n(x)$ denotará cualquiera de las tres funciones de distribución suavizadas previamente definidas.

DEFINICION 3.3.1. En las condiciones anteriores, se define el siguiente algoritmo de simulación de muestras bootstrap suavizadas, con función de Distribución $R_n(x)$

Paso 1. Obtener una muestra de tamaño n de la población $R_n(x)$, i.e.

$$X_1^*, X_2^*, \ldots, X_n^* \; iid \sim R_n(x)$$

Paso 2. Repetir el paso 1, de forma independiente, un número B de veces, obteniendo así B muestras bootstrap suavizadas.

La única diferencia con el Método II de Efron (1979) radica en la sutitución de $F_n(x)$ por la correspondiente función de distribución suavizada.

La implementación práctica de este algoritmo puede realizarse teniendo en cuenta la inversa de $R_n(x)$.

En concreto, se propone su implementación según el siguiente algoritmo equivalente.

Paso 1. Simular una observación u, de una variable aleatoria uniforme en (0,1). Calcular

$$X_1^* = R_n^{-1}(u)$$

Paso 2. Repetir el paso 1 un total de n veces, independientemente, de forma que se obtiene una muestra bootstrap

$$\boldsymbol{X}^{*1} = (X^*_1, X^*_2, \ldots, X^*_n)$$

Paso 3. Se repiten los pasos 1 y 2 un número B de veces, obteniéndose así B muestras bootstrap suavizadas:

$$\boldsymbol{X}^{*1}, \boldsymbol{X}^{*2}, \ldots, \boldsymbol{X}^{*B}$$

Para cada una de estas B muestras bootstrap suavizadas se calcularía el estadístico correspondiente y se realizaría el análisis oportuno.

3.4. SIMULACIONES.

En este apartado se analizan diversas simulaciones realizadas utilizando la función de distribución suavizada $S_n(x)$, en comparación con el Método II de Efron (1979) en el que se emplea la Función de Distribución Empírica $F_n(x)$.

En el Anexo III figura el listado de la subrutina empleada para la simulación de muestras bootstrap a partir de $S_n(x)$.

Las características técnicas de los recursos hardware y software empleados son las mismas que en el capítulo anterior, según se recoge en el Anexo IV. El tiempo de CPU se mide en segundos.

TABLA 5.1.

Se analiza a continuación una muestra de tamaño 54 que aparece en Efron (1986) relativa a datos sanguíneos. El estadístico analizado ha sido la media muestral. Se calcula un intervalo de confianza al 95%, según el método percentil. Se consideraron puntos extremos 0.0 y 9.6.

		Media		Intervalo de confianza		
B	**Método**	**Media**	**D.T.**	**Ex.inf.**	**Ex.sup.**	**CPU**
100	Boots.	2.32711	0.19638	1.96852	2.76667	0.19
	Suaviz.	2.30960	0.20396	1.92408	2.77446	0.23
200	Boots.	2.32156	0.19451	1.92407	2.75370	0.29
	Suaviz.	2.31926	0.19103	1.97547	2.71235	0.38
500	Boots.	2.31513	0.20174	1.96111	2.73889	0.59
	Suaviz.	2.32876	0.20108	1.96849	2.79160	0.66
1000	Boots.	2.31699	0.20611	1.94074	2.73704	1.30
	Suaviz.	2.33288	0.20636	1.93425	2.77279	1.42
2000	Boots.	2.32622	0.20852	1.94444	2.76111	3.50
	Suaviz.	2.32783	0.20261	1.93730	2.76815	3.65

TABLA 5.2.

Se analizan 150 muestras de tamaño 15 de una población BETA de parámetros 0.5 y 1, siendo la media, en este caso 1/3, la característica a estudiar utilizando la media aritmética muestral. Los extremos a considerar en la definición de $S_n(x)$ son 0,1.

MEDIA

B	Método	Media	D.T.	CPU
50	Bootstrap	0.3265	0.0739	4.12
	B.Suaviz.	0.3291	0.0725	4.39
100	Bootstrap	0.3267	0.0745	7.74
	B.Suaviz.	0.3294	0.0737	8.32
200	Bootstrap	0.3270	0.0737	16.26
	B.Suaviz.	0.3276	0.0736	17.48
500	Bootstrap	0.3274	0.0735	59.56
	B.Suaviz.	0.3288	0.0732	61.80

TABLA 5.3.

Se analizan 150 muestras de tamaño 15 de una población BETA de parámetros 3 y 2, siendo la media, en este caso 0.6, la característica a estudiar utilizando la media aritmética muestral. Los extremos a considerar en la definición de $S_n(x)$ son 0,1.

MEDIA

B	Método	Media	D.T.	CPU
50	Bootstrap	0.5837	0.0501	4.05
	B.Suaviz.	0.5994	0.0496	4.36
100	Bootstrap	0.5835	0.0501	7.41
	B.Suaviz.	0.5987	0.0492	8.10
200	Bootstrap	0.5827	0.0500	16.43
	B.Suaviz.	0.5986	0.0475	17.68
500	Bootstrap	0.5838	0.0500	59.46
	B.Suaviz.	0.5988	0.0732	61.60

TABLA 5.4.

Se analizan 150 muestras de tamaño 15 de una población BETA de parámetros 5 y 5, siendo la media, en este caso 0.5, la característica a estudiar utilizando la media aritmética muestral. Los extremos a considerar en la definición de $S_n(x)$ son 0,1.

MEDIA

B	Método	Media	D.T.	CPU
50	Bootstrap	0.4890	0.0373	3.87
	B.Suaviz.	0.5037	0.0372	4.52
100	Bootstrap	0.4951	0.0374	7.76
	B.Suaviz.	0.5040	0.0372	8.45
200	Bootstrap	0.4869	0.0368	16.90
	B.Suaviz.	0.5036	0.0352	18.09
500	Bootstrap	0.4950	0.0369	59.00
	B.Suaviz.	0.5041	0.0359	60.44

Comentarios sobre las simulaciones:

TABLA 5.1.

Se observan estimaciones análogas para la media, si bien el bootstrap suavizado presenta una desviación típica ligeramente inferior. El extremo superior de los Intervalos de Confianza que proporciona el bootstrap suavizado es ligeramente superior. El aumento de tiempo de CPU que requiere el bootstrap suavizado es poco notable.

TABLAS 5.1., 5.2., 5.3.:

El bootstrap suavizado se acerca más a la media poblacional, al tiempo que ofrece menor desviación típica. Tampoco se produce un aumento excesivo en los tiempos de CPU.

ANEXO I

Programa utilizado para calcular la función de distribución y características de la variable número de puntos distintos de que consta una muestra bootstrap.

```
(*** ESTE PROGRAMA, IMPLEMENTADO EN PASCAL, PERMITE EVALUAR LA FUNCION
DE DISTRIBUCION DE LA VARIABLE ALEATORIA NUMERO DE OBSERVACIONES
DISTINTAS EN UNA MUESTRA BOOTSTRAP, ADEMAS DE CALCULAR LAS MEDIDAS
ESTADISTICAS DE TAL VARIABLE. SE UTILIZO PARA CONFECCIONAR LAS TABLAS
1,2 Y 3 DEL CAPITULO II ***)


(*** En las siguientes declaraciones,
     numta es el número de tamaños muestrales a tabular,
     ncuan es el número de percentiles a calcular       ***)

program tabula(input,output,f,g);
const nmax=100; numta=10;blanco='            ';ncuan=20;
var n,i,j,k,numini:integer;
    x,s,pot,sumx,sumxcu:quadruple;
    q1,med,q3:array [1..nmax] of integer;
    v,media,sigma:array [1..nmax] of quadruple;
    tabla:array [1..nmax,1..nmax] of quadruple;
    signi:array[1..ncuan] of quadruple;
    cuanti: array[1..nmax,1..ncuan] of integer;
    nomfic:packed array[1..40] of char;
    f,g:text;


(*** Esta función calcula el número combinatorio n sobre k y pertenece
a la librería de subrutinas estadísticas y matemáticas IMSL ***)


function dbinom(n,k:integer):double;fortran;


(*** La siguiente función calcula la probabilidad de que una muestra
bootstrap tome un número de observaciones distintas igual a k, según
la fórmula visto en el Capítulo II ***)


function prob(n,k:integer):quadruple;
var i,j,menos:integer;

    p,sum,prod:quadruple;

begin
 p:=1;
 for i:=1 to n do p:=p*1.0/n;
 if (k=1) then
      prob:=n*p
 else
   begin
      sum:=0;
      menos:=-1;
      for i:=0 to k do
      begin
        prod:=1;
        for j:=1 to n do prod:=prod*(k-i);
```

```
        menos:=menos*(-1);
        sum:=sum+menos*dbinom(k,i)*prod;
      end;
        prob:=(dbinom(n,k)*p)*sum;
   end;
end; (*** fin de prob ***)

(*** El programa realiza la tabulación para todos los tamaños de
muestra comprendidos entre los valores de numini y
numini+numta-1 ***)

begin
 write('deme tamaño de muestra inicial: ');
 readln(numini);
 write('deme fichero salida: ');
 readln(nomfic);
 open(f,nomfic,new);
 rewrite(f);

(*** percentiles a evaluar ***)

 for i:=2 to ncuan do
     signi[i]:=(i-1)*0.05;
 signi[1]:=0.01;
k:=0;
for n:=numini to numini+numta-1 do
begin
 k:=k+1;
 for i:=1 to ncuan do cuanti[n,i]:=0;
 s:=0;
 sumx:=0;
 sumxcu:=0;
 for i:=1 to n do
   begin
     x:=prob(n,i);
        sumx:=sumx+i*x;
        sumxcu:=sumxcu+(i**2)*x;
        s:=s+x;
     for j:=1 to ncuan do
      if (cuanti[n,j] =0) then
          if (s >= signi[j]) then cuanti[n,j]:=i;
        tabla[i,k]:=s;
   end;

 media[k]:=sumx;
 sigma[k]:=sqrt(sumxcu-sumx**2);
end;

(*** escritura de la función de distribución ***)

write(f,'            Función de distribución del número k de puntos
distintos en');
writeln(f,'  muestras  bootstrap  para  n  entre  ',numini:3,'  y
',numini+numta-1:3);
writeln(f);
write(f,'   k ');
```

```
for i:=numini to numini+numta-1 do write(f,i:9,'    ');
writeln(f);

for i:=1 to numini+numta-1 do
begin write(f,i:4,' ');
      k:=0;
      for j:=numini to numini+numta-1 do
      begin k:=k+1;
        if (i > j) then
          write(f,blanco:12)
        else
          write(f,tabla[i,k]:12:8);
      end;
      writeln(f);
end;
  close(f);

(*** medidas estadísticas ***)

open (g,'distri.bin',new);
rewrite(g);
writeln(g,'Distribución del número de puntos distintos en m.
bootstrap');
write(g,'  n  ' , '    Media    ', 'Desv. tÝpica');
for i:=1 to ncuan do write(g,signi[i]:3:2);writeln(g);
for i:=numini to  numini+numta-1 do
begin
   write(g,i:3, media[i]:11:5,sigma[i]:12:5);
   for j:=1 to ncuan do write(g,cuanti[i,j]:4);
   writeln(g);
end;
close(g);
end.
```

ANEXO II

Subrutina utilizada para generar muestras bootstrap con detección de muestras bootstrap outliers.

```
SUBROUTINE BOOTDETECT

      INTEGER N,I,J,B,EST,POS,IN,METODO,DIMEN,K,OPCION,L,K1
      INTEGER IPER(B*N),OPC,B2,PCRIT,B1,ELEG(N)
      REAL MUESTRA(N),RESMEDIA,SOLUCION,RESULTADO,RES(B)
      REAL ALFA,RESVAR,H(B),NRES(N),P1,P2,MRAND(N)
      REAL VAR_REAL,VARSESGO,U(N),DATO(5),VR
      REAL E1,E2,S1,S2,S11,S22,VAR1,VAR2,RESM
      REAL A1,A2,VR1,P,PR(N),AUX1,AUX2,VART,BINDF,CERO
      REAL MRAND2(N), W(B),CARD
      CHARACTER*1  OP,FIC2,FIC
      EXTERNAL BINDF,ORDENA

C*************************************************************
C* Generación de MRAND, muestras aleatorias
c* METODO=1 permite realizar bootstrap balanceado
C*************************************************************

      IF (METODO.EQ.1) THEN
           DIMEN=B*N
           CALL RNPER(DIMEN,IPER)
      ENDIF


C** Bucle de B simulaciones


      B1=0
      DO J=1,B


      DO I=1,N
        ELEG(I)=0
      ENDDO

      IF (METODO.EQ.1) THEN
           DO I=1,N
             K=IPER((J-1)*N+I)
             POS=K-(K/N)*N +1
             MRAND(I)=MUESTRA(POS)
             ELEG(POS)=1
           END DO
      ENDIF

      IF (METODO.EQ.2) THEN
           DO I=1,N
                X=RNUNF()
                POS=INT(X*N)+1
                MRAND(I)=MUESTRA(POS)
                ELEG(POS)=1
           END DO
      ENDIF
```

```
C*************************************************************************
******
C**  BOOTSTRAP    CON    DETECCION    DE    MUESTRAS    OUTLIERS
C*************************************************************************
*****

      NCRITICO=0
           DO I=1,N
              IF (ELEG(I).EQ.1) THEN
                NCRITICO=NCRITICO+1
              ENDIF
           ENDDO


C** Si NCRITICO >= PCRIT entonces la muestra generada  es válida

      IF (NCRITICO.GT.PCRIT) THEN

C** B1 cuenta las simulaciones que se van a realizar , luego RES(B1),
C** siempre B1 < B


           B1 = B1 + 1

C* Cálculo del estadÝstico correspondiente a partir de la muestra
aleatoria
C* generada.


                CALL MEDIA(MRAND,N,RESULTADO)



C* En cada simulacion obtenemos el resultado del estadistico, se
C* guarda en RES


                RES(B1)=RESULTADO

      ENDIF

C**   FIN Si (NCRITICO.GE.PCRIT)


      END DO
C*Termina el bucle de la B simulaciones
```

```
C*****************************************************************
C**         CALCULO DE LOS RESULTADOS
C**
C** RESMEDIA - Media de las simulaciones
C** SOLUCION - Desviación típica de las simulaciones
C** VARSESGO - Sesgo
C*****************************************************************


      CALL MEDIASIMU(B1,RES,RESMEDIA)

      CALL DESVISIMU(B1,RES,RESMEDIA,SOLUCION)


C*****************************************************************
C*****                 INTERVALOS DE CONFIANZA
C*****************************************************************


      IF (OP.EQ. 'S' .OR. OP.EQ.'s') THEN

           CALL INTERVALO(RES,B1,N,SOLUCION,VR,EST,H,FIC2,E1,E2,
     +           MUESTRA,MRAND,U,NRES,A1,A2,OPC,OPCION,W)

      ENDIF



C*****************************************************************
C****                RESULTADOS  OBTENIDOS                    ****
C*****************************************************************

      IF (OPCION.EQ. 3)   THEN

      IF (FIC2.EQ.'N' .OR. FIC2.EQ.'n')  THEN


      PRINT 23
      PRINT 24
      PRINT 23


      PRINT 23
      PRINT 21,VR
      PRINT 23
      PRINT 15, RESMEDIA
      PRINT 20, SOLUCION
      PRINT 25, VARSESGO
      PRINT 23
      PRINT 26
      PRINT 23

      ENDIF
```

```
C*************************************************************
C****                  FORMATOS UTILIZADOS
C*************************************************************

5     FORMAT( 1X,'SESGO  **************', F2.0)
10    FORMAT(1X,'VARIANZA  ***',F14.5)
15    FORMAT(1X,'LA MEDIA (BOOTSTRAP): ',F14.5)
20    FORMAT(1X,'LA DESVIACION TIPICA (BOOTSTRAP): ',F14.5)
21    FORMAT(' VALOR ESTADISTICO MUESTRA INICIAL: ', F14.5)
23    FORMAT('0')
24    FORMAT('****************   RESULTADOS   ******************')
25    FORMAT(1X,'EL SESGO ES :',F14.5)
26    FORMAT('************************************************')
27    FORMAT('*+*************** VALORES TEORICOS ***************')

      END
```

ANEXO III

Subrutina utilizada para la simulación de muestras bootstrap suavizadas.

```
      SUBROUTINE BOOTSV

C*DECLARACIONES

      INTEGER N,I,J,B,EST,POS,IN,METODO,DIMEN,K,OPCION,L,K1
      REAL MUESTRA(N),RESMEDIA,SOLUCION,RESULTADO,RES(B)
      REAL ALFA,RESVAR,H(B),NRES(N),P1,P2,MRAND(N)
      REAL VAR_REAL,VARSESGO,AUX
      CHARACTER*1  OP,FIC2,FIC
      INTEGER IPER(B*N),OPC
      REAL E1,E2,U(N),DATO(4),VR
      REAL A1,A2,VR1,P,PR(N),AUX1,AUX2,VART,BINDF,CERO
      REAL MAUX(N+2),UU,V1,V2,JOTA,Y(N+2),W(B)

      EXTERNAL BINDF,ORDENA


C*************************************************************
****
C*** OPC - Método de Intervalo de Confianza
C*** METODO - Método de obtención de las muestras bootstrap

C*************************************************************
****



C*LLamada a la rutina de generación de n. aleatorios para cada
C*elemento del array MUESTRA, siendo la salida un array MRAND, con N
C*elementos

C*************************************************************
C***  BOOTSTRAP SUAVIZADO
C***
C***  A partir de una muestra inicial, se genera otra a la cual se
C***  le aplica, las B simulaciones.
C*************************************************************

C** Ordenación de la muestra inicial

      MAUX(1)=V1
      MAUX(N+2)=V2

      DO I=1,N
           MAUX(I+1)=MUESTRA(I)
      ENDDO

C** Ordenación de MAUX

      CALL ORDENA(MAUX,N+2)

C** Se genera el vector de Y(N+2) MAUX(1)= X0
C** y MAUX(n+2) = Xn+1

      Y(1)=MAUX(1)
```

```
      Y(N+2)=MAUX(N+2)
      DO I=2,N+1
           Y(I)=(MAUX(I+1) + MAUX(I+2))/2
      ENDDO


C* Bucle de las B simulaciones
      DO J=1,B

      DO I=1,N

C** Calcular U uniforme en (0,1)

                UU = RNUNF()

C** Obtención de j = nint(u*n) + 1

              JOTA = NINT(UU*N) +1

C** Cálculo de la fórmula

      MRAND(I)=Y(JOTA) + N*(UU - ((JOTA)/N))*(Y(JOTA+1)-Y(JOTA))

      ENDDO
C** I=1,N ****

                      CALL MEDIA(MRAND,N,RESULTADO)

C* En cada simulacion se obtiene el resultado del estadistico, se
C* guarda en RES

                RES(J)=RESULTADO

      END DO
C*Termina el bucle del n. de simulaciones,B


C*****************************************************************

      CALL MEDIASIMU(B,RES,RESMEDIA)

      CALL DESVISIMU(B,RES,RESMEDIA,SOLUCION)


C*****************************************************************
*******
C*****                INTERVALOS DE CONFIANZA
C*****************************************************************
*******



      CALL INTERVALO(RES,B,N,SOLUCION,VR,EST,H,FIC2,E1,E2,
```

```
C*****************************************************************
C************************** RESULTADOS ***************************
C*****************************************************************

      IF (OPCION.EQ. 3)    THEN

      IF (FIC2.EQ.'N' .OR. FIC2.EQ.'n')  THEN


23    FORMAT('0')
      PRINT 23
      PRINT *,'********** RESULTADOS **********************'
      PRINT 23


      PRINT *, ' LA MEDIA (BOOTSTRAP): ', RESMEDIA
      PRINT *, ' LA DESVIACION TIPICA (BOOT. SUAVIZADO): ',SOLUCION


      PRINT 23
      PRINT *,'*****************************************'
      PRINT 23

      ENDIF


      END
```

ANEXO IV

Descripción de los recursos software y hardware utilizados en las tabulaciones y simulaciones.

Los programas utilizados para el cálculo de las tabulaciones y simulaciones que aparecen en esta memoria se han realizado sobre el nodo SECLU0 de la Red Informática Científica de Andalucía. Este nodo consta de dos equipos, un VAX 8530 y un VAX 6410, bajo sistema operativo VMS, que comparten dispositivos de almacenamiento secundario, y que residen físicamente en el Centro de Informática Científica de Andalucía.

Los programas, escritos en PASCAL y FORTRAN, utilizan rutinas de la librerías estadística y matemática IMSL.

BIBLIOGRAFIA

Anscombe F.J. (1960). Rejection of outliers. Technometrics Vol. 2.

Bickel, P.J., & Freedman, D.A., (1981). Some Asymptotic Theory for the Bootstrap. The Annals of Statistics, Vol. 9, No. 6.

Boos,D.D., & Monahan, J.F. (1986). Bootstrap methods using prior information. Biometrika, Vol. 73, 1.

Davison, A.C., & Hinkley, D.V., & Schechtman, E. (1986). Efficient bootstrap simulation. Biometrika 73.

Davison, A.C., & Hinkley, D.V.. (1988). Saddlepont approximations in resampling methods. Biometrika, Vol. 75.

Efron, B. (1979). Bootstrap methods: Another look at the Jackknife. Ann. Statist. 7.

Efron, B. (1982). The Jackknife, the Bootstrap and Other Resampling Plans. Society for Industrial and Applied Mathematics. Philadelphia, Pensilvania.

Efron, B. & Tibshirani, R. (1986). Bootstrap Methods for Standar Errors, Confidence Intervals, and Other Measures of Statistical Accuracy. Statistical Science. Vol.1 , No.1.

Efron, B. (1990). More Efficient Bootstrap Computations. Journal of the American Statistical Association. Vol. 85, No. 409, Theory and Methods.

Feller, W., (1973). Introducción a la Teoría de Probabilidad y sus aplicaciones. Ed. Limusa-Wiley.

Kolmogorov, A.N., & Fomin, S.V. (1975). Elementos de la Teoría de Funciones y del Análisis Funcional. Ed. MIR.

Mardia,K.V., Kent, J.T., & Bibby,J.M. (1979). Multivariate Analysis. Academic, Nueva York.

Muñoz García, J., Moreno Rebollo, J.L., & Pascual Acosta, A. (1990). Outliers: A formal approach. International Statistical Review. Vol. 58, 3.

Shi, S. G. (1992). Accurate and effcient double-bootstrap confidence limit method. Computational Statistics & Data Analysis.

Silverman, B.W. y Young, G.A. (1987). The bootstrap: To smooth or not to smooth. Biometrika, Vol. 74, 3.

Stine, R. (1990). Modern Methods of Data Analysis. Editado por John Fox & J. Scott Long.

Tukey, J.W. (1986). Sunset Salvo. American Statistician, 40.

Wu, C.F.J. (1986). Jackknife, Bootstrap and Other Resampling Methods in Regression Analysis. The Annals of Statistics, Vol. 14, No. 4.