Estimación de Coberturas de Intervalos de Confianza Bootstrap para la media de una población Exponencial mediante el software para análisis Estadístico, **R**

Abelardo E. Monsalve C.

UNIVERSIDAD CENTROCCIDENTAL "*LISANDRO ALVARADO*"

Decanato de Ciencias y Tecnología.

Barquisimeto, 2007

Estimación de Coberturas de Intervalos de Confianza Bootstrap para la media de una población Exponencial mediante el software para análisis Estadístico **R**

Por

Abelardo E. Monsalve C.

Trabajo de Ascenso presentado como requisito parcial para optar
a la categoría de Agregado en el escalafón del personal
docente e investigación de la UCLA.

UNIVERSIDAD CENTROCCIDENTAL "*LISANDRO ALVARADO*"
Decanato de Ciencias y Tecnología.

Barquisimeto, 2007

RESUMEN DEL TRABAJO DE ASCENSO PRESENTADO COMO REQUISITO PARCIAL PARA OPTAR A LA CATEGORÍA DE AGREGADO EN EL ESCALAFÓN DEL PERSONAL DOCENTE E INVESTIGACIÓN DE LA UCLA.

Estimación de Coberturas de Intervalos de Confianza Bootstrap para la media de una población Exponencial mediante el software para análisis Estadístico **R**

por

Abelardo Monsalve

UNIVERSIDAD CENTROCCIDENTAL "*LISANDRO ALVARADO*"

Decanato de Ciencias y Tecnología.

Barquisimeto, 2008

En la estadística, uno de los problemas comunes es el de obtener intervalos de confianza para un parámetro previamente estimado. En la teoría clásica esto puede llevarse a cabo a partir del método de los estadísticos pivotales, claro esta siempre y cuando se conozca la distribución de los mismos. En un contexto no paramétrico (planteamiento realista) los resultados que se obtienen son siempre aproximaciones las cuales, en muchos casos, no suelen ser fáciles salvo que el objetivo sea una media o una función suave de la media. En este trabajo se abordará la metodología Bootstrap como una salida practica para aproximar la distribución de dichos estadísticos pivotales , con la ventaja de que dicha aproximación es siempre posible (aún con el desconocimiento de la distribución de la población) y más sencilla, con un beneficio adicional, en algunos casos los intervalos construidos mediante técnicas bootstrap arrojan un menor error de recubrimiento que los de la teoría clásica. Dicho error de recubrimiento esta referido pues a la cobertura real del intervalo puesto que al estimar un intervalo con nivel de confianza (1-α) % lo obtenido es solo una aproximación y no representa dicho valor exacto. Se presentarán algunos resultados al aplicar esta metodología a un par de datos generados artificialmente, a decir, una distribución *exponencial* y una distribución *Ji- cuadrado*, para corroborar la eficacia y ventajas

de la metodología bootstrap, Seguidamente se lo aplicará a un conjunto de datos relativos al censo de la población de Venezuela para determinar un parámetro sencillo relacionado a la proporción de la población entre los dos últimos censos de 1990 y 2001.

Agradecimientos

Al Dr. José Manuel Prada por su apoyo y orientación en la consecusión de este trabajo.

Al Dr. Manuel Febrero, por su valiosa colaboración en la parte computacional de este trabajo

A la Universidad Centroccidental "Lisandro Alvarado" por permitirme formar parte del personal docente y brindarme la oportunidad de realizar estudios de Doctorado

A Dios y mi Familia, que me iluminan y orientan en cada uno de mis pasos

Índice general

Introducción

En los últimos años las técnicas estadísticas se han convertido en métodos analíticos de elección en ciencias biomédicas, la psicología, la educación, la economía, la teoría de la comunicación, la sociología, los estudios genéticos, la epidemiología, y otras muchas áreas.

Más recientemente, las ciencias tradicionales como la geología, la física, la astronomía han comenzado a utilizar cada vez más los métodos estadísticos como soporte para sus investigaciones.

El *bootstrap* es una técnica que tiene como ventaja el desconocimiento de hipótesis sobre el mecanismo que genera los datos y cuyo soporte se basa en el uso intensivo de herramientas computacionales, es por ello que el bootstrap debe su desarrollo y eficacia al potencial computacional que actualmente puede proveer los avances modernos de la computación, con la finalidad de simplificar los intrincados cálculos tradicionales de la teoría estadística.

Entre sus precursores teóricos se pueden mencionar Laplace (1810) y Chebyshev (final siglo XIX) con su trabajos acerca de la teoría limite. Las primeras contribuciones en este campo son debidas a Hubback (1878-1968) en su trabajo de muestreo espacial para ensayos agrícolas, ya que dio las primeras ideas acerca de esta metodología. Mahalanobis en la década de 1930 se encuentra entre los precursores del bootstrap por bloques. Ya más recientemente, Efron Bradley (Stanford University) y Hall P, a finales de los años '70 fusionan la potencia del Método de Monte Carlo con la resolución de problemas planteados de forma muy general.

El tradicional camino del aprendizaje de la estadística está bloqueado, en su mayoría, por una formidable pared de matemáticas. La visión, de parte de su creador, Efron Bradley ([10]), de la técnica *Bootstrap* evita dicha pared. El *bootstrap* es un método que basa su potencia en la herramienta computacional de tal manera que puede responder a muchas de las preguntas de la estadística real sin fórmulas.

Por ello y gracias a las computadoras modernas se pueden aplicar estas ideas con flexibilidad,

rapidez, facilidad, y con un mínimo de supuestos matemáticos.

El objetivo de este modesto trabajo es presentar, para aquellos que no conozcan y para quienes la conocen pero que no han hecho uso exhaustivo de sus bondades, la metodología *bootstrap* acompañado de su implementación como una herramienta eficaz a la hora de analizar y comprender complejos conjuntos de datos y más específicamente, lo que nos compete en este trabajo, determinar intervalos de confianza respecto de un parámetro de interés.

La palabra "comprender" es un elemento importante en la frase anterior. Efron Bradley y Tibshirani Robert en su libro ([10]) expresan, que su técnica no es un compendio de recetas de cocina de la estadística, sino que pretende dar al lector una buena comprensión intuitiva de la inferencia estadística.

Uno de los aspectos relevantes de la técnica *bootstrap* es que provee de una apreciación directa de la varianza, sesgo, media, y otros fenómenos probabilísticos de gran importancia en la inferencia estadística.

Ahora bien, *¿Qué significa que un intervalo de confianza contiene el verdadero valor con probabilidad 0,90?*

La respuesta habitual de la literatura parece tremendamente abstracta para la mayoría de las personas que se inician en el campo de la estadística. Los Intervalos de Confianza Bootstrap son construidos directamente de los conjuntos de datos reales, utilizando un simple algoritmo. Esto conlleva a un ahorro de un sin fin de rigurosos detalles matemáticos, que en algunos casos resulta casi imposible de resolver. La técnica bootstrap además evita los errores frecuentes que se producen al asociar medidas de confiabilidad a la estimación de un parámetro, cuya distribución es desconocida (y que en muchas ocasiones no es normal, es decir, una visión más realista) y para el que no se dispone de muestras grandes.

Como se ha mencionado anteriormente, en este trabajo se presentarán con detalles los algoritmos y su implementación en el software estadístico (libre, sin costo) R en su version 2.6.1 (2007-11-26) Copyright (C) 2007 The R Foundation for Statistical Computing ISBN 3-900051-07-0 (`http://www.r-project.org/`), de manera que cualquier persona que desee hacer uso de los mismos pueda de manera sencilla ejecutarlos. Así mismo se hará un estudio de simulación en el cual se determinará la eficacia de la técnica determinando las coberturas de intervalos de confianza alrededor de la media a partir de las aproximaciones bootstrap estándar y bootstrap del tipo percentil, así como también se presentarán las estimaciones de los intervalos de confianza

de Efron, Hall y Sesgo Corregido.

Para entender un poco la técnica la cual es ampliamente usada por los estadísticos, específicamente en el área de análisis no-paramétrico, y que recientemente se ha comenzado a usar por investigadores en otras areas un ejemplo se puede encontrar en ([1]) y ([13]) Ambos artículos aplican la técnica y presentan resultados satisfactorios en cada caso.

Es de destacar que este trabajo pretende mostrar que la técnica bootstrap es de fácil aplicación aún con conocimientos básicos de estadística y computación, y que los programas que permiten calcular los intervalos de confianza bootstrap, son fáciles de implementar en cualquier lenguaje de computación y en cualquier ordenador actualizado. Por otro lado, la teoría fundamenta la confiabilidad y eficiencia de los intervalos obtenidos mediante esta metodología.

En el capítulo 1 se presentará un breve resumen acerca de la teoría básica de la metodología bootstrap de manera que el lector se familiarice con la misma.

En el capítulo 2, se presentarán los aspectos teóricos y prácticos de los intervalos de confianza bootstrap, del tipo percentil , intervalos de confianza bootstrap de Efron, Hall y sesgo corregido destacando sus bondades e indicando además los pasos para su implementación en R, para cada caso en particular.

En el capítulo 3 se mostrará mediante tres ejemplos la eficacia de la metodología bootstrap para la obtención de intervalos de confianza. Los tres primeros ejemplos están referidos a dos poblaciones teóricas, a decir, la exponencial , una normal y una ji- cuadrado, en cada caso la población será con parámetro o parámetros conocidos. Se determinarán los intervalos de confianza del tipo bootstrap y se estimará además su cobertura. El objetivo que persigue estos ejemplos es dar confiabilidad al lector acerca de la potencialidad de la técnica y su implementación, la cual solo requiere un esfuerzo computacional. Seguidamente se aplicará a datos reales, con el fin de estimar aproximaciones de los intervalos de confianza del tipo bootstrap para la proporción de los valores medios de la población de Venezuela tomando los datos de los censos de 1990 y 2001. Con estos ejemplos se quiere mostrar la potencialidad de las aproximaciones del tipo bootstrap frente a los tradicionales intervalos de confianza.

Capítulo 1

La Metodología Bootstrap

En este capítulo se dará la idea intuitiva del *Bootstrap*, cuyo término se deriva de la frase "*pull oneself up by one's bootstrap*" , es decir, "salir adelante sin ayuda", para más detalles ver ([10]).

El bootstrap es uno de los métodos estadísticos denominados de computación intensiva, introducido por Efron(1979) ([10]). La metodología bootstrap permite efectuar inferencias estadísticas sin necesidad de suposiciones previas acerca de la distribución, que en muchos casos suelen ser de difícil justificación.

1.1. El Principio "Plug-In"

En estadística suele discutirse los términos de *parámetro* y *estadístico*. Un parámetro es una función de la función de distribución de probabilidad F. Un estadístico es una función de la muestra x. Así por ejemplo, la $var(x)$ es un parámetro de F, mientras que $\hat{var}(x)$ es un estadístico basado en la muestra x. La notación

$$\theta = \theta(F)$$

, con F función de Distribución de probabilidad, enfatiza que el valor θ del parámetro es obtenido aplicando algún procedimiento numérico $\theta(\cdot)$ a la distribución F. Por ejemplo, si F es la distribución probabilidad en la recta real, la esperanza puede pensarse como el parámetro

$$\theta = \theta(F) = E_F(x).$$

Aquí $\theta(F)$ determina θ mediante el valor promedio de x ponderado de acuerdo a F. Aún si F es desconocida, la forma de $\theta(F)$ nos da indicios acerca del funcional que tiene como entrada F y salida θ.

Pues bien, el "*principio plug-in*" es un método simple de estimación de parámetros a partir de la muestra. El "*estimador plug-in*" del parámetro $\theta = \theta(F)$ está definido por

$$\hat{\theta} = \theta(\hat{F})$$

En otras palabras, se estima la función $\theta = \theta(F)$ de la distribución de probabilidad F mediante la misma función de la distribución empírica $\hat{F}$, $\hat{\theta} = \theta(\hat{F})$.

Un *estimador plug-in* de la esperanza $\theta = E_F(x)$ es

$$\hat{\theta} = E_{\hat{F}}(x) = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$$

¿Cuán bueno es el *principio plug-in*?

Por lo general es bueno solo cuando la información disponible acerca de F proviene de la muestra x. Bajo estas circunstancias $\hat{\theta} = \theta(\hat{F})$ no puede ser mejorado como un estimador de $\theta = t(F)$, al menos en el sentido de la teoría estadística asintótica ($n \to \infty$).

1.2. El estimador Bootstrap

Si $X = (X_1, X_2, \ldots, X_n)$ es una muestra aleatoria simple genérica de tamaño n con función de distribución común denotada por $F(x) = P(X \leq x)$, simbólicamente

$$X_1, X_2, \ldots, X_n \sim F(x) \quad \text{o bien} \quad X_i \sim F(x), \quad i = 1, 2, \ldots, n.$$

Cuando se desconoce el valor de un parámetro θ de una población, y en consecuencia, se desea utilizar un estimador $\theta = \theta(X_1, X_2, \ldots, X_n)$ del mismo, es importante conocer la precisión de dicho estimador. Esta es una de las primeras aplicaciones del bootstrap.

Si en una realización muestral del vector aleatorio $X = (X_1, X_2, \ldots, X_n)$ se observa $X_1 = x_1$, $X_2 = x_2$, $\ldots$, $X_n = x_n$ se denominará al vector de componentes $(x_1, x_2, \ldots, x_n)$ la muestra original. Se puede decir que en la metodología bootstrap los datos observados en la muestra original $x = (x_1, x_2, \ldots, x_n)$ asumen el papel de la verdadera distribución desconocida $F(x)$, quedando ésta sustituida por su estimación $F_n(x)$, la cual suele ser frecuentemente la distribución empírica de $(x_1, x_2, \ldots, x_n)$ que asigna peso $1/n$ a cada x_i

$$F_n(x) = \frac{\text{número de } (x_i \leq x)}{n} = \frac{1}{n}\sum_{i=1}^{n} I(X_i \leq x)$$

con $I(A)$ la función indicadora:

$$I(A) \begin{cases} 1, & \text{si} x \in A \\ \\ 0, & \text{si} x \notin A \end{cases}$$

En adelante, denotaremos a las funciones de distribución poblacional $F(x)$ y $F_n(x)$ simplemente por F y F_n respectivamente.

Si $X^* = (X_1^*, X_2^*, \ldots, X_n^*)$ es una muestra aleatoria simple genérica de F_n, es decir, cada X_i^*, $1 \leq i \leq n$ de esta muestra es obtenido independientemente (con reemplazamiento) de la muestra original $x = (x_1, x_2, \ldots, x_n)$ de forma que $X_i^* = x_j$ para algún $1 \leq j \leq n$, al conjunto $X^* = (X_1^*, X_2^*, \ldots, X_n^*)$ se le denomina muestra bootstrap o remuestra bootstrap, La notación $X^{*(b)}$ indica que nos referimos a la b-ésima remuestra bootstrap obtenida de la muestra original, la cual de forma genérica, podemos denotar así

$$X^{*(b)} = (X_1^{*(b)}, X_2^{*(b)}, \ldots, X_n^{*(b)})$$

Es claro que para obtener el número total de posibles remuestras bootstrap (n^n), el tiempo requerido de ordenador puede ser considerable, en la práctica no es necesario extraer el total de las remuestras ya que en muchas ocasiones se logra la convergencia con un número aproximado de 1000 remuestras, o incluso menos.

Para resumir lo anterior se puede decir que el proceso bootstrap, y en particular, la construcción del estimador bootstrap de la desviación típica de un estimador, consta de las siguientes etapas:

1. Se extrae una muestra bootstrap $X^* = (X_1^*, X_2^*, \ldots, X_n^*)$ a partir de la muestra original $x = (x_1, x_2, \ldots, x_n)$ tal como se describió anteriormente.

2. Se aplica la función que define al estadístico de interés, $\hat{\theta}$ a la muestra construida en la etapa precedente, con lo que se obtiene

$$\hat{\theta}^{*(b)} = \hat{\theta}(X_1^{*(b)}, X_2^{*(b)}, \ldots, X_n^{*(b)})$$

3. Se repiten los pasos 1 y 2 precedentes, B veces

4. Se construye una distribución de probabilidad a partir de los B valores $\hat{\theta}^{*(b)}$, asignando una frecuencia relativa $1/B$ a cada punto $\hat{\theta}^{*(1)}$, $\hat{\theta}^{*(2)}$, ...,$\hat{\theta}^{*(B)}$. Esta distribución $G^*(\hat{\theta}^*)$,

es el estimador bootstrap de la distribución muestral exacta de $\hat{\theta}$,$G(\hat{\theta})$. En la metodología bootstrap se utiliza $G^*(\hat{\theta}^*)$ para efectuar inferencias sobre $\hat{\theta}$, ya que la distribución exacta $G(\hat{\theta})$ suele ser desconocida.

5. Se construye el estimador bootstrap de la desviación estándar (típica) del estimador $\hat{\theta}(X_1, X_2, \ldots, X_n)$:

$$\hat{\sigma}^*_{\hat{\theta}} = \left\{ \frac{\sum_{b=1}^{B} \left(\hat{\theta}^{*(b)} - \hat{\theta}^*(\bullet) \right)^2}{B-1} \right\}^{(1/2)}$$

con $\hat{\theta}^*(\bullet) = \dfrac{\sum_{b=1}^{B} \hat{\theta}^{*(b)}}{B}$

Pues bien, una vez estudiado las suposiciones básicas de la técnica bootstrap se está en condiciones de afrontar, la construcción de intervalos de confianza del tipo bootstrap.

Capítulo 2

Intervalos de Confianza Bootstrap

En este capítulo se presentarán formalmente los intervalos de confianza en estudio y algunos aspectos teóricos relevantes de los mismos.

En la teoría clásica la construcción de intervalos de confianza de nivel $(1-\alpha)$ requiere el conocimiento de la distribución de estadísticos pivotales. En un contexto no paramétrico (planteamiento realista) los resultados que se obtienen son siempre aproximaciones las cuales, no siempre suelen ser fáciles salvo que el objetivo sea una media o una función suave de la media. El planteamiento clásico se basa en resultados del tipo

$$\sqrt{n}\left(\frac{\overline{X}-\mu}{\sigma}\right) \simeq N(0,1) \tag{2.1}$$

Sin embargo, *El Bootstrap* ofrece una salida práctica para aproximar la distribución de dichos estadísticos pivotales, con la ventaja de que dicha aproximación es siempre posible y más fácil, con un beneficio adicional, en algunos casos los intervalos construidos mediante técnicas bootstrap arrojan un menor error de recubrimiento que los de la teoría clásica.

Pues bien, en lo que sigue, el objetivo se centrará en obtener, en particular, intervalos de confianza para la media.

2.1. Intervalos de Confianza Normales (Teoría Clásica)

Se comenzará presentando la técnica clásica para obtener intervalos de confianza normales. Sea $\vec{X} = (X_1, X_2, \ldots, X_n)$ una muestra aleatoria simple (m.a.s) de tamaño n de F- función de distribución desconocida. Entonces, para construir aproximaciones de los intervalos de confianza para la media μ con nivel de significación $(1-\alpha)$, se procede de la siguiente manera:

Caso Unilateral: tomando en cuenta (2.1)

$$P\left(\sqrt{n}\frac{\bar{X}-\mu}{\sigma} \leq z_\alpha\right) \simeq \alpha, \quad \text{con } z_\alpha = \Phi^{-1}(\alpha)$$

por lo tanto

$$P\left(\mu \in \left(-\infty, \bar{X} - \frac{\sigma}{\sqrt{n}} z_\alpha\right)\right) \simeq 1-\alpha$$

con lo cual el *intervalo unilateral* de confianza para la media de nivel $\simeq (1-\alpha)$ y que se denotará por J_0 es

$$\boxed{J_0 = \left(-\infty, \bar{x} - \frac{\sigma}{\sqrt{n}} z_\alpha\right)} \tag{2.2}$$

con *Error de Cobertura* denotado por $E_{cob\,J_0}$

$$E_{cob\,J_0} = P\left(\mu \in J_0\right) = 1-\alpha+0(n^{-1}) \tag{2.3}$$

Se ha denotado por $\bar{x}$, la media muestral y por s la desviación estándar (típica) muestral, definidas habitualmente como:

$$\bar{x} = \frac{\sum_{i=1}^n x_i}{n} \quad \text{y} \quad s^2 = \frac{\sum_{i=1}^n (x_i - \bar{x})^2}{n-1}$$

Caso Bilateral: de la misma manera, tomando (2.1)

$$P\left(-z_{1-\frac{\alpha}{2}} \leq \sqrt{n}\frac{\bar{X}-\mu}{\sigma} \leq z_{1-\frac{\alpha}{2}}\right) \simeq \alpha,$$

por lo tanto

$$P\left(\mu \in \left(\bar{X} \mp \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right)\right) \simeq 1-\alpha$$

obteniéndose el *intervalo bilateral* de confianza para la media de nivel $\simeq (1-\alpha)$ y que se denotará por I_0,

$$\boxed{I_0 = \left(\bar{x} - \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}, \bar{x} + \frac{\sigma}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right)} \tag{2.4}$$

En ambos casos (2.2) y (2.4), de no conocerse el valor de la desviación σ, ésta se sustituye por su correspondiente estimador $\widehat{\sigma} = s$, en tal caso J_0 e I_0 quedan expresados como:

$$\boxed{J_0 = \left(-\infty, \bar{x} - \frac{s}{\sqrt{n}} z_\alpha\right)} \tag{2.5}$$

and

$$\boxed{I_0 = \left(\bar{x} - \frac{s}{\sqrt{n}} z_{1-\frac{\alpha}{2}}, \bar{x} + \frac{s}{\sqrt{n}} z_{1-\frac{\alpha}{2}}\right)} \tag{2.6}$$

Es muy sencillo calcular estos intervalos de confianza normales para la media, usando el software estadístico R. El lector puede intentarlo siguiendo estos pasos

- Se genera una muestra aleatoria simple de tamaño $n = 100$, $x = (x_1, x_2, \ldots, x_n)$ de la población exponencial, por ejemplo, con $EX = \mu = 100$.
- Se estima su media y desviación correspondiente, usando las respectivas funciones predefinidas por R,
- y luego se calcula el intervalo de confianza normal para el parámetro en cuestión partir de (2.2) y (2.4) o bien (2.5) y (2.6) según sea el caso. (ver apéndice (1))

2.2. Intervalos de Confianza Bootstrap Unilaterales

De manera similar a como se hizo en la sección anterior, se presentará lo más relevante de la metodología bootstrap, necesaria para determinar las aproximaciones de los intervalos de confianza.

Sea $\vec{X} = (X_1, X_2, \ldots, X_n)$ una (m.a.s) de tamaño n de F. El objetivo es construir intervalos de confianza unilaterales de nivel $(1 - \alpha)$ para $\theta = \theta(F)$ a partir de $\hat{\theta} = \theta(\vec{X})$ con varianza $\frac{\sigma^2}{n}$.

Antes de continuar, es pertinente definir un par de conceptos:

- Se dice que $Y_n = O_p(b_n)$ si y sólo si para todo ϵ existe $c_\epsilon$tal que

$$P\left(\frac{|\, Y_n \,|}{|\, b_n \,|} \leq C_\epsilon\right) > 1 - \epsilon$$

  si $n \geq n_0(\epsilon, c_\epsilon)$

- Se dice que $Y_n = o_p(b_n)$ si y sólo si

$$\frac{Y_n}{b_n} \to 0, \quad (n \to \infty)$$

A continuación se presenta la metodología para obtener los intervalos del tipo bootstrap.

1. Método Percentil: Si x_α es el cuantil de orden α de $\sqrt{n}(\hat{\theta} - \theta)$ bajo F entonces

$$\alpha = P_{_F}\left(\sqrt{n}(\hat{\theta} - \theta) \leq x_\alpha\right) \quad \Rightarrow \quad P_{_F}\left(\theta \leq \hat{\theta} - \frac{x_\alpha}{\sqrt{n}}\right) = 1 - \alpha$$

por lo tanto:

$$\boxed{J = \left(-\infty, \hat{\theta} - \frac{x_\alpha}{\sqrt{n}}\right)} \quad \text{I.C unilateral para } \theta \text{ nivel } (1-\alpha)$$

Lo anterior solo es posible si se conoce la distribución de $\sqrt{n}(\hat{\theta} - \theta)$. Pues bien, el *método percentil* aproxima la distribución de $\sqrt{n}(\hat{\theta} - \theta)$ bajo F por la de $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ bajo $\hat{F}$ (Distribución Bootstrap). Por lo tanto si x_α es el cuantil de orden α de $\sqrt{n}(\hat{\theta}^* - \hat{\theta})$ entonces

$$\boxed{J_1 = \left(-\infty, \hat{\theta} - \frac{x^*_\alpha}{\sqrt{n}}\right)} \quad \text{I.C unilateral para } \theta \text{ nivel } (1-\alpha) \tag{2.7}$$

con *Error de Cobertura del Método Percentil* denotado por

$$E_{cob\, J_1} = P\left(\mu \in J_1\right) = 1 - \alpha + 0_p(n^{-1/2}) \tag{2.8}$$

Para ilustrar como se procedería en R, suponga que se quiere construir el intervalo unilateral para la media $\theta = \mu$ de una población X exponencial, por ejemplo, para ello se procede de la siguiente manera:

- Se genera una muestra aleatoria simple de tamaño $n = 100$, $x = (x_1, x_2, \ldots, x_n)$ de la población exponencial, con $EX = \mu = 100$.
- Se generan $B = 1000$ remuestras bootstrap $x^* = (x^*_1, x^*_2, \ldots, x^*_n)$ uniformes a partir de la muestra original y se calcula para cada remuestra, $R^* = \sqrt{n}\,(\bar{x}^* - ,\bar{x})$. Para este paso se puede usar la función predefinida por R, llamada ***resampling*** o bien la subrutina particular ***remuestreo.R***
- Se aproxima por Monte Carlo el valor del cuantil x_α esto es, $\widehat{x^*_\alpha}^{MC} = R^{*([B\alpha])}$, con $\alpha = 0{,}10$.
- Finalmente se calcula el intervalo, J_1 a partir de (2.7) (ver apéndice (2))

2. Método Percentil-t : Si x_α es el cuantil de orden α de $\sqrt{n}\dfrac{(\hat{\theta} - \theta)}{\hat{\sigma}}$ bajo F, entonces

$$\alpha = P_{_F}\left(\sqrt{n}\frac{(\hat{\theta} - \theta)}{\hat{\sigma}} \leq x_\alpha\right) \quad \Rightarrow \quad P_{_F}\left(\theta \leq \hat{\theta} - \frac{\hat{\sigma}}{\sqrt{n}}x_\alpha\right) = 1 - \alpha$$

por lo tanto:

$$\boxed{J = \left(-\infty, \hat{\theta} - \frac{\hat{\sigma}}{\sqrt{n}}x_\alpha\right)} \quad \text{I.C unilateral para } \theta \text{ nivel } (1-\alpha)$$

En este caso el *método percentil-t* aproxima la distribución de $\sqrt{n}\dfrac{(\hat{\theta}-\theta)}{\hat{\sigma}}$ bajo F por la de $\sqrt{n}\dfrac{(\hat{\theta}^*-\hat{\theta})}{\hat{\sigma}^*}$ bajo $\hat{F}$ (Distribución Bootstrap). Por lo tanto si x_α es el cuantil de orden α de dicha distribución, entonces

$$\boxed{J_2 = \left(-\infty, \hat{\theta} - \frac{\hat{\sigma}}{\sqrt{n}} x^*_\alpha\right)} \quad \text{I.C unilateral para } \theta \text{ nivel } (1-\alpha) \tag{2.9}$$

con *Error de Cobertura del Método Percentil-t* denotado por

$$E_{cob\, J_2} = P\left(\mu \in J_2\right) = 1 - \alpha + 0_p(n^{-1}) \tag{2.10}$$

Para el método *percentil-t* se procede de manera similar, solo que, en lugar de calcular $R^* = \sqrt{n}\left(\bar{x}^* - \bar{x}\right)$ se calcula $R^* = \sqrt{n}\dfrac{(\bar{x}^* - \bar{x})}{\hat{\sigma}^*}$, con lo cual se obtiene a partir de (2.9),(ver apéndice (2))

2.3. Intervalos de Confianza Bootstrap Bilaterales

Sea $\vec{X} = (X_1, X_2, \ldots, X_n$ una (m.a.s) de tamaño n de F. El objetivo es construir intervalos de confianza bilaterales de nivel $(1-\alpha)$ para $\theta = \theta(F)$ a partir de $\hat{\theta} = \theta(\vec{X})$ con varianza $\frac{\sigma^2}{n}$.

1. Método Percentil: Se procede de manera análoga a como se hizo en el caso unilateral. En este caso

$$P_{_F}\left(x_{_{\alpha/2}} \leq \sqrt{n}(\hat{\theta}-\theta) \leq x_{_{1-\alpha/2}}\right) = 1-\alpha$$

por lo tanto:

$$\boxed{I = \left(\hat{\theta} - \frac{x_{_{1-\alpha/2}}}{\sqrt{n}}\ ,\ \hat{\theta} - \frac{x_{_{\alpha/2}}}{\sqrt{n}}\right)} \quad \text{I.C bilateral para } \theta \text{ nivel } (1-\alpha)$$

aplicando el *método percentil* se obtiene

$$\boxed{I_1 = \left(\hat{\theta} - \frac{x^*_{_{1-\alpha/2}}}{\sqrt{n}}, \hat{\theta} - \frac{x^*_{_{\alpha/2}}}{\sqrt{n}}\right)} \quad \text{I.C bilateral para } \theta \text{ nivel } (1-\alpha) \tag{2.11}$$

donde $x^*_{_{1-\alpha/2}}$ y $x^*_{_{\alpha/2}}$ son los cuantiles de orden $1-\alpha/2$ y $\alpha/2$ respectivamente de la distribución de $\sqrt{n}\left(\hat{\theta}^* - \hat{\theta}\right)$. El *Error de Cobertura del Método Percentil*

$$E_{cob\, I_1} = P\left(\mu \in I_1\right) = 1 - \alpha + 0_p(n^{-1/2}) \tag{2.12}$$

El procedimiento para construir estos intervalos de confianza usando R es similar al usado para determinar J_1, con una simple modificación, que en este caso se calculan los dos extremos, inferior y superior:

- Se Genera una muestra aleatoria simple de tamaño $n = 100$, $x = (x_1, x_2, \ldots, x_n)$ de la población exponencial, con $EX = \mu = 100$.
- Se generan $B = 1000$ remuestras bootstrap $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ uniformes a partir de la muestra original y se calcula para cada remuestra, $R^* = \sqrt{n}\,(\bar{x}^* - \bar{x})$. Para este paso se usará la función predefinida por R, llamada ***resampling*** o bien la subrutina particular ***remuestreo.R***
- Se aproxima por Monte Carlo el valor de los cuantiles $x_{\alpha/2}$ y $x_{1-\alpha/2}$, esto es,

$$\widehat{x^*}^{MC}_{1-\alpha/2} = R^{*([B(1-\alpha/2)])} \quad \text{y} \quad \widehat{x^*}^{MC}_{\alpha/2} = R^{*([B\alpha/2])} \quad \text{con} \quad \alpha = 0{,}10,$$

  para

$$R^* = \sqrt{n}\,(\bar{x}^* - \bar{x})$$

- Finalmente se calcula el intervalo, I_1 a partir de (2.11)(ver apéndice (3))

2. Método Percentil-t : En este caso

$$P_{_F}\left(x_{\alpha/2} \leq \sqrt{n}\frac{(\hat{\theta} - \theta)}{\hat{\sigma}} \leq x_{1-\alpha/2}\right) = 1 - \alpha$$

por lo tanto:

$$\boxed{I = \left(\hat{\theta} - \frac{\hat{\sigma}}{\sqrt{n}} x_{1-\alpha/2}\ ,\ \hat{\theta} - \frac{\hat{\sigma}}{\sqrt{n}} x_{\alpha/2}\right)} \quad \text{I.C bilateral para } \theta \text{ nivel } (1-\alpha)$$

aplicando el *método percentil-t* se obtiene

$$\boxed{I_2 = \left(\hat{\theta} - \frac{\hat{\sigma}}{\sqrt{n}} x^*_{1-\alpha/2}\ ,\ \hat{\theta} - \frac{\hat{\sigma}}{\sqrt{n}} x^*_{\alpha/2}\right)} \quad \text{I.C bilateral para } \theta \text{ nivel } (1-\alpha) \tag{2.13}$$

donde $x^*_{1-\alpha/2}$ y $x^*_{\alpha/2}$ los cuantiles de orden $1-\alpha/2$ y $\alpha/2$ respectivamente de la distribución de $\sqrt{n}\dfrac{\hat{\theta}^* - \hat{\theta}}{\hat{\sigma}^*}$. El *Error de Cobertura del Método Percentil-t*

$$E_{cob\,I_2} = P\,(\mu \in I_2) = 1 - \alpha + 0_p(n^{-1}) \tag{2.14}$$

Para calcular el intervalo de confianza bootstrap por el *método percentil-t*, se aproxima por Monte Carlo el valor de los cuantiles $x_{\alpha/2}$ y $x_{1-\alpha/2}$, esto es,

$$\widehat{x^*}^{MC}_{1-\alpha/2} = R^{*([B(1-\alpha/2)])} \quad \text{y} \quad \widehat{x^*}^{MC}_{\alpha/2} = R^{*([B\alpha/2])} \quad \text{con} \quad \alpha = 0{,}10.$$

para

$$R^* = \sqrt{n}\frac{(\bar{x}^* - \bar{x})}{\hat{\sigma}^*}$$

Finalmente se calcula el intervalo, I_2 a partir de (2.13)

3. Método Percentil-t Simétrizado: Es análogo al método percentil-t, pero difiere en la forma de estimar los cuantiles. Conociéndose la distribución de $\sqrt{n}\frac{(\hat{\theta}-\theta)}{\hat{\sigma}}$ bajo F, y siendo $y_{1-\alpha}$ tal que

$$P\left(\sqrt{n}\left|\frac{(\hat{\theta}-\theta)}{\hat{\sigma}}\right| \leq y_{1-\alpha}\right) = 1-\alpha$$

entonces el intervalo de confianza de nivel alpha para θ es

$$\boxed{I = \left(\hat{\theta} - \frac{\hat{\sigma}}{\sqrt{n}}y_{1-\alpha}\ ,\ \hat{\theta} + \frac{\hat{\sigma}}{\sqrt{n}}y_{1-\alpha}\right)}$$

El método percentil-t simétrizado aproxima a la distribución de $\sqrt{n}\frac{(\hat{\theta}-\theta)}{\hat{\sigma}}$ bajo F, por la de $\sqrt{n}\frac{(\hat{\theta}^*-\theta)}{\hat{\sigma}^*}$ bajo $\hat{F}$, con lo cual se obtiene el intervalo de confianza bilateral de nivel $(1-\alpha)$ para θ

$$\boxed{I_3 = \left(\hat{\theta} - \frac{\hat{\sigma}}{\sqrt{n}}y^*_{1-\alpha}\ ,\ \hat{\theta} + \frac{\hat{\sigma}}{\sqrt{n}}y^*_{1-\alpha}\right)} \quad \text{I.C bilateral para } \theta \text{ nivel } (1-\alpha) \tag{2.15}$$

donde $y^*_{1-\alpha}$ es el cuantile de orden $1-\alpha$ de la distribución de $\sqrt{n}\left|\frac{\hat{\theta}^*-\hat{\theta}}{\hat{\sigma}^*}\right|$. El *Error de Cobertura del Método Percentil-t*

$$E_{cob\,I_3} = P\left(\mu \in I_3\right) = 1 - \alpha + 0_p(n^{-3/2}) \tag{2.16}$$

Finalmente, para calcular el intervalo de confianza bootstrap por el *método percentil-t simetrizar*, se aproxima por Monte Carlo el valor del cuantil $y_{1-\alpha}$, esto es,

$$\widehat{y^*}^{MC}_{1-\alpha} = R^{*([B(1-\alpha)])} \quad \text{con} \quad \alpha = 0{,}10. \quad \text{para} \quad R^* = \sqrt{n}\left|\frac{(\bar{x}^* - \bar{x})}{\hat{\sigma}^*}\right|$$

y el intervalo I_3 se obtiene aplicando (2.15)

Observación 1

a) *El* método percentil *invierte su esfuerzo en calcular una corrección para la escala de* θ*, por lo que no proporciona una corrección efectiva para el sesgo.*

b) *El* método percentil-t*, invierte su esfuerzo en corregir el sesgo ya que el estadístico lo esta por su escala, por ello este método tiene un mejor comportamiento.*

Los intervalos que se han calculado usando R, se pueden obtener con una única subrutina, de nombre ***bootstrapP9.R***, cuyos datos de entrada, son:

- ***muestra***: dato de entrada que contendrá la muestra aleatoria simple para la cual se le aplicará los métodos antes mencionados.
- ***alpha***: valor para determinar el nivel de confianza deseado.
- ***B***: El número de remuestras deseado para obtener los intervalos de confianza bootstrap.

Siguiendo la misma idea antes presentada en cada una de las construcciones de J_1, J_2, I_1, I_2 e I_3, se resume el calculo a:

- Generar una muestra aleatoria simple de tamaño $n = 100$, $x = (x_1, x_2, \ldots, x_n)$ de la población exponencial, con $EX = \mu = 100$.
- Generar $B = 1000$ remuestras bootstrap $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ uniformes a partir de la muestra original y se calcula para cada remuestra, calculando

$$R^* = \sqrt{n}\,(\bar{x}^* - \bar{x})\,, \quad R_2^* = \sqrt{n}\frac{(\bar{x}^* - \bar{x})}{\hat{\sigma}^*}, \quad R_3^* = \sqrt{n}\left|\frac{(\bar{x}^* - \bar{x})}{\hat{\sigma}^*}\right|$$

  . Para este paso se usará la función predefinida por R, llamada ***sample*** o bien la subrutina particular ***remuestreo.R***

- Se aproxima por Monte Carlo el valor de los cuantiles x_{α}, $y_{1-\alpha}$, $x_{\alpha/2}$ y $x_{1-\alpha/2}$, esto es,

$$\widehat{x_{\alpha}^*}^{MC} = R_1^{*([B\alpha])}, \quad \widehat{y^*}_{1-\alpha}^{MC} = R_3^{*([B(1-\alpha)])}$$

$$\widehat{x^*}_{1-\alpha/2}^{MC} = R_2^{*([B(1-\alpha/2)])} \quad \text{y} \quad \widehat{x^*}_{\alpha/2}^{MC} = R_2^{*([B\alpha/2])} \quad \text{con} \quad \alpha = 0{,}10,$$

- Finalmente se calculan los intervalos, J_0, J_1, J_2, I_0, I_1, I_2 e I_3 a partir de (2.2),(2.7),(2.9),(2.4),(2.11),(2.13),(2.15), (ver apéndice (3.3))

  Si se aplica la metodología antes descrita a una muestra de tamaño n proveniente de una distribución exponencial de parámetro $a = EX = 100$, se obtienen los intervalos que se muestran en la figura (2.1)

Intervalos	**Límite Inferior**	**Limite Superior**
Intervalo Normal Estándar Unilateral	$-\text{Inf}$	108.22
Aproximación Percentil Unilateral	$-\text{Inf}$	107.63
Aproximación Percentil-t Unilateral	$-\text{Inf}$	110.17
Intervalo Normal Estándar Bilateral	76.75	112.13
Aproximación Percentil Bilateral	77.02	111.70
Aproximación Percentil-t Bilateral	78.93	115.27
Aproximación Percentil-t Simetrizado	76.21	112.67

Cuadro 2.1: Intervalos de Confianza Bootstrap para la media $EX = \mu = 100$ de una población exponencial.

2.4. Otros Tipos de Intervalos de Confianza Bootstrap

1. Método Bootstrap Estándar: El método mas simple para obtener intervalos de confianza bootstrap, es el *método bootstrap estándar*. El principio es similar al usado para obtener intervalos de confianza normales para la media. Aquí si un estimador $\hat{\theta}$ esta distribuido según una ley normal con media θ y desviación estándar σ, entonces con probabilidad de $(1-\alpha)$

$$\hat{\theta} - z_{\alpha/2}\sigma < \theta < \hat{\theta} + z_{\alpha/2}\sigma$$

   Para construir el intervalo de confianza bootstrap estándar, σ es estimada a partir de las remuestras bootstrap, es decir, se sustituye por su estimación bootstrap, $\hat{\sigma}_{boot}$, así pues el intervalo que se obtiene será

$$\boxed{I_{BE} = \left(\hat{\theta} - z_{\alpha/2}\hat{\sigma}_{boot}\ ,\ \hat{\theta} + z_{\alpha/2}\hat{\sigma}_{boot}\right)} \tag{2.17}$$

Para ilustrar como se procedería en R, suponga que se quiere construir el intervalo bootstrap estándar para $\theta = \mu$ una población X exponencial, por ejemplo, para ello se procede de la siguiente manera:

- Se Genera una muestra aleatoria simple de tamaño $n = 100$, $x = (x_1, x_2, \ldots, x_n)$ de la población exponencial, con $EX = \mu = 100$.
- Se generan $B = 1000$ remuestras bootstrap $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ uniformes a partir de la muestra original y se calcula para cada remuestra, $R^* = \bar{x}^*$.
- Finalmente se calcula el intervalo, IBS a partir de (2.17)

2. Primer Método Percentil (Efron, 1979):

   Este intervalo de confianza del tipo percentil, viene dado por la ecuación

$$P\left(\hat{\theta}_I < \theta < \hat{\theta}_S\right)$$

   donde $\hat{\theta}_I$ es el valor en la distribución bootstrap que es excedido con probabilidad $1 - \alpha/2$ y $\hat{\theta}_S$ es el valor que es excedido con probabilidad $\alpha/2$, es decir son los cuantiles $1 - \alpha/2$ y $\alpha/2$ respectivamente de las remuestras bootstrap del parámetro de interés. Por lo tanto el intervalo de confianza percentil de Efron de nivel $(1 - \alpha)$ es

$$\boxed{\left(\hat{\theta}^*_{1-\alpha/2} < \theta < \hat{\theta}^*_{\alpha/2}\right)} \tag{2.18}$$

3. Segundo Método Percentil (Hall, 1992):

   Este intervalo de confianza del tipo percentil, viene dado por la ecuación

$$P\left(2\hat{\theta} - \hat{\theta}_S < \theta < 2\hat{\theta} + \hat{\theta}_I\right)$$

   donde $\hat{\theta}_I$ es el valor en la distribución bootstrap que es excedido con probabilidad $1 - \alpha/2$ y $\hat{\theta}_S$ es el valor que es excedido con probabilidad $\alpha/2$, es decir son los cuantiles $1 - \alpha/2$ y $\alpha/2$ respectivamente de las remuestras bootstrap del parámetro de interés. Entonces el intervalo de confianza percentil de Hall de nivel $(1 - \alpha)$ es

$$\boxed{\left(2\hat{\theta} - \hat{\theta}^*_{\alpha/2} < \theta < 2\hat{\theta} + \hat{\theta}^*_{1-\alpha/2}\right)} \tag{2.19}$$

   Para ilustrar como se procedería en R, suponga que se quiere construir el intervalo bootstrap estándar para $\theta = \mu$ una población X exponencial, por ejemplo, para ello se procede de la siguiente manera:

- Se Genera una muestra aleatoria simple de tamaño $n = 100$, $x = (x_1, x_2, \ldots, x_n)$ de la población exponencial, con $EX = \mu = 100$.
- Se generan $B = 1000$ remuestras bootstrap $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ uniformes a partir de la muestra original y se calcula para cada remuestra, $R^* = \bar{x}^*$.
- Se calculan los cuantiles

$$\widehat{x^*}^{MC}_{1-\alpha/2} = R^{*([B(1-\alpha/2)])} \quad \text{y} \quad \widehat{x^*}^{MC}_{\alpha/2} = R^{*([B\alpha/2])} \quad \text{con} \quad \alpha = 0{,}10,$$

- Finalmente se calculan los intervalos de *Efron* y *Hall*, a partir de (2.18) y (2.19),

4. Método Percentil por Sesgo Corregido:

El supuesto que se requiere para construir intervalos de confianza del tipo percentil por sesgo corregido para un parámetro θ, es que exista una función monótona creciente de un estimador $\hat{\theta}$ tal que al evaluar dicho estimador $f(\hat{\theta})$ sus valores sigan una distribución normal con media $(f(\theta) - z_0)$ y desviación estándar $\sigma = 1$, de tal manera que

$$P\left(-z_{\alpha/2} < f(\hat{\theta}) - f(\theta) + z_0 < z_{\alpha/2}\right) = 1 - \alpha$$

. Considerando que $f(\hat{\theta})$ tiene distribución normal de media $f(\theta) - z_0$ entonces para Z con distribución $N(0, 1)$

$$P\{f(\hat{\theta}) > f(\theta)\} = P(Z > z_0), \quad \Rightarrow \quad P(\hat{\theta} > \theta) = P(Z > z_0)$$

de manera que $z_0 = \Phi(p)$, donde P es la proporción de veces que los estimados bootstrap $\hat{\theta}^*$ son excedidos por θ. El siguiente paso es determinar los limites del intervalo de confianza. Para ello determinamos:

$$p_I = \Phi(2z_0 - z_{\alpha/2}), \quad mboxy \quad p_S = \Phi(2z_0 + z_{\alpha/2})$$

luego, el intervalo de confianza viene expresado por

$$\boxed{\left(\hat{\theta}_I^* \ , \ \hat{\theta}_S^*\right)} \tag{2.20}$$

donde $\hat{\theta}_I^*$ es el valor que excede una proporción p_I de todos los valores de la distribución bootstrap de los estimados $\hat{\theta}^*$, y $\hat{\theta}_S^*$ es el valor que excede una proporción p_S de todos los valores de la distribución bootstrap de los estimados $\hat{\theta}^*$.

A continuación se enumeran los pasos a seguir para obtener este intervalo bootstrap para $\theta = \mu$ de una población exponencial.

Intervalo de Confianza Bootstrap-Percentil por Sesgo Corregido:

a)

b) Se Genera una muestra aleatoria simple de tamaño $n = 100$, $x = (x_1, x_2, \ldots, x_n)$ de la población exponencial, con $EX = \mu = 100$.

c) Se generan $B = 1000$ remuestras bootstrap $x^* = (x_1^*, x_2^*, \ldots, x_n^*)$ uniformes a partir de la muestra original y se calcula para cada remuestra, $R^* = \bar{x}^*$.

d) Calcular la proporción de veces p que $R^* = \bar{x}^*$ supera o exceden a $\hat{\mu}$, el estimado de μ

e) Calcular z_0 el valor de la distribución Normal estándar que es superado con probabilidad p.

f) Calcular $p_I = \Phi(2z_0 - z_{\alpha/2})$

g) Calcular $p_S = \Phi(2z_0 + z_{\alpha/2})$

h) Se Calculan los valores de los extremos, $\hat{\theta}_I^* = R^{*([B(p_I)])}$, y $\hat{\theta}_S^* = R^{*([B(p_S)])}$

Los intervalos antes mencionados y presentados por separado pueden calcularse mediante la función ***bootstrapP9=function(muestra,alpha,B)*** la cual aplicada a una muestra exponencial de parámetro $EX = 100 = a$ obteniéndose

	Lower limit	Upper limit
Intervalo Bootstrap Estándar	67.65	108.67
Aproximación Percentil de Efron	69.11	110.34
Aproximación Percentil Hall	65.97	107.21
Aproximación Percentil Sesgo Corregido	67.38	107.68

El código en R de la función ***bootstrapP9=function(muestra,alpha,B)*** se puede obtener en (ver apéndice (3.3)) de este trabajo.

Capítulo 3

Simulaciones y Resultados

En este capítulo se determinarán las aproximaciones bootstrap y normales de los intervalos de confianza para dos poblaciones particulares, la exponencial y ji-cuadrado. Además se estudiará la cobertura de los intervalos de confianza antes mencionados. Específicamente se centrará el estudio en los intervalos bilaterales I_0, I_1, I_2 e I_3.

Por otro lado, se presentará un estudio sencillo respecto de la proporción de los promedios de la población Venezolana correspondiente a los censos 1990 y 2000, con la finalidad de aplicar las aproximaciones bootstrap de los intervalos de confianza de los diversos tipos tratados en este trabajo para dicha proporción.

3.1. Aplicación a una población Exponencial

En esta sección se determinarán los intervalos de confianza bootstrap y normales para la media de una población con distribución exponencial, así mismo estimaremos su cobertura real.

Sea $X = (X_1, X_2, \ldots, X_n)$ una muestra aleatoria simple de una población con distribución Exponencial de media $a = EX = 100$. Aquí el parámetro de interés es $\theta = \mu = EX$, para el cual se determinarán los intervalos. Una vez obtenidos los intervalos se estimará la cobertura real de los mismos.

Para ello, se procede como sigue:

1. Se generan $N = 1000$, muestras independientes de tamaño n de una población exponencial con parámetro $a = EX = 100$.

2. Para cada muestra se generan $B = 1000$ remuestras de tamaño n y a partir de éstas se calculan los intervalos de confianza expuestos en los capítulos anteriores.

3. Para cada muestra y los intervalos determinados para las mismas, se verifica si el valor $a = 100$ esta contenido en los intervalos que se van determinando, de pertenecer al intervalo se asigna a un contador el valor 1 de no estar entonces el valor es 0.

4. Finalmente se promedia el número de $1'$s que se han obtenido con lo cual tenemos un estimado de la cobertura para una realización. Esquemáticamente es como sigue:

Ecuaciones que se usarán para determinar los intervalos para cada muestra j, de tamaño n, con $j = 1, \ldots, N$

$$\sqrt{n}\frac{\bar{X} - a}{S} \simeq N(0,1) \longrightarrow I_0 = \left(\bar{x} \pm s\frac{z_{1-\alpha/2}}{\sqrt{n}}\right)$$

$$\sqrt{n}(\bar{X} - a) \simeq \sqrt{n}(\bar{X}^* - \bar{X}) \longrightarrow I_1 = \left(\bar{x} - \frac{x^*_{1-\alpha/2}}{\sqrt{n}}, \bar{x} + \frac{x^*_{\alpha/2}}{\sqrt{n}}\right)$$

$$\sqrt{n}\frac{\bar{X} - a}{S} \simeq \sqrt{n}\frac{\bar{X}^* - \bar{X}}{S^*} \longrightarrow I_2 = \left(\bar{x} - s\frac{x^*_{1-\alpha/2}}{\sqrt{n}}, \bar{x} + s\frac{x^*_{\alpha/2}}{\sqrt{n}}\right)$$

$$\sqrt{n}\left|\frac{\bar{X} - a}{S}\right| \simeq \sqrt{n}\left|\frac{\bar{X}^* - \bar{X}}{S^*}\right| \longrightarrow I_3 = \left(\bar{x} \pm s\frac{y^*_{1-\alpha}}{\sqrt{n}}\right)$$

$$(x_1, x_2, \ldots, x_n)_1 \quad \begin{cases} (x_1^*, x_2^*, \ldots, x_n^*)_1^1, \\ \vdots \\ (x_1^*, x_2^*, \ldots, x_n^*)_1^B, \end{cases} \longrightarrow I_i, \ \text{i=0,1,2,3} \begin{cases} 1, & \text{si } a = 100 \in I_i, \ \text{i=0,1,2,3} \\ 0, & \text{si } a = 100 \notin I_i, \ \text{i=0,1,2,3} \end{cases}$$

$$(x_1, x_2, \ldots, x_n)_2 \quad \begin{cases} (x_1^*, x_2^*, \ldots, x_n^*)_2^1, \\ \vdots \\ (x_1^*, x_2^*, \ldots, x_n^*)_2^B, \end{cases} \longrightarrow I_i, \ \text{i=0,1,2,3} \begin{cases} 1, & \text{si } a = 100 \in I_i, \ \text{i=0,1,2,3} \\ 0, & \text{si } a = 100 \notin I_i, \ \text{i=0,1,2,3} \end{cases}$$

$$\vdots$$

$$(x_1, x_2, \ldots, x_n)_N \quad \begin{cases} (x_1^*, x_2^*, \ldots, x_n)_N^1, \\ \vdots \\ (x_1^*, x_2^*, \ldots, x_n)_N^B, \end{cases} \longrightarrow I_i, \ \text{i=0,1,2,3} \begin{cases} 1, & \text{si } a = 100 \in I_i, \ \text{i=0,1,2,3} \\ 0, & \text{si } a = 100 \notin I_i, \ \text{i=0,1,2,3} \end{cases}$$

$$cob_{I_i} \frac{\sum_{j=1}^{N} 1}{N}, \ i = 0, 1, 2, 3, 4$$

Para obtener una realización de los estimados de la cobertura, aplicamos la función ***coberturas.R*** (ver apéndice (3.3)), la cual se se encarga de ejecutar todos los pasos antes especificados, dicha función se ejecuta de la siguiente manera

coberturas(B=1000,n=100,alpha=0.10,N=1000,valuet=100),

(por ejemplo). Aquí *valuet* se refiere al valor verdadero del parámetro de interés.

A continuación se aplicará dicha función a distintas combinaciones de B, n y N con la finalidad de dar una idea acerca de la importancia del número de remuestras necesarias y de los tamaños de la muestra.

- Para un tamaño de muestra $n = 10$ (pequeño) y un número de remuestras $B = 100$

(pequeño) con un valor de $N = 1000$ (grande) y $alpha = 0{,}10$ (para un intervalo de confianza 1-α %)-

```
> coberturas(B = 100, n = 10, alpha = 0.1, N = 1000, valuet = 100)
```

Con lo cual se obtienen las siguientes coberturas:

```
        J0    J1    J2    I0    I1    I2    I3  IBN   EPA   HPA   BPC
B=100 0.762 0.749 0.836 0.799 0.759 0.852 0.853 0.79 0.781 0.759 0.786
```

- Para un tamaño de muestra $n = 10$ (pequeño) y un número de remuestras $B = 1000$ (suficientemente grande), con un valor de $N = 1000$ (suficientemente grande) y $alpha = 0{,}10$

```
> coberturas(B = 1000, n = 10, alpha = 0.1, N = 1000, valuet = 100)
```

aumentándose el número de remuestras se obtiene:

```
          J0    J1    J2    I0   I1    I2    I3   IBN   EPA  HPA   BPC
B=1000 0.762 0.787 0.867 0.799 0.79 0.891 0.878 0.807 0.817 0.79 0.828
```

se puede observar una ligera mejora en algunos casos

- Para un tamaño de muestra $n = 100$ (grande) y un número de remuestras $B = 100$ (pequeño) con un valor de $N = 1000$ (grande) y $alpha = 0{,}10$

```
> coberturas(B = 100, n = 100, alpha = 0.1, N = 1000, valuet = 100)
```

```
         J0    J1    J2    I0    I1    I2    I3   IBN   EPA   HPA   BPC
B=100 0.873 0.862 0.891 0.901 0.855 0.866 0.883 0.881 0.873 0.855 0.865
```

- Para un tamaño de muestra $n = 100$ (grande) y un número de remuestras $B = 1000$ (suficientemente grande) con un valor de $N = 1000$ (grande) y $alpha = 0{,}10$

```
> coberturas(B = 1000, n = 100, alpha = 0.1, N = 1000, valuet = 100)
```

se observa una mejoría notable, ya que las coberturas se aproximan a la cobertura teórica

```
          J0    J1    J2    I0    I1    I2    I3    IBN   EPA   HPA BPC
B=1000 0.873 0.869 0.897 0.901 0.892 0.906 0.906 0.901 0.902 0.892 0.9
```

Se puede apreciar en las ejecuciones anteriores como incide el tamaño de la muestra n y el número de remuestras B. Cuando éstas son mayores, el valor de la cobertura tiende a mejorar, en contraposición a cuando sus valores son menores. La literatura sugiere valores de $B = 1000$ como mínimo para obtener una buena aproximación a la cobertura.

En las siguientes ejecuciones se repetirá el proceso anterior un aproximado de $M = 100$ veces para aproximar la cobertura real.

```
> E =Estudio(M=100,B=c(100),n=100,alpha=0.1,N=1000,valuet=100)
```

```
> F = Estudio(M=100,B=c(1000),n=100,alpha=0.1,N=1000,valuet=100)
```

Los resultados son mostrados en la tabla (3.1) Se puede observar que para $B = 100$ y $B = 1000$ con $n = 100$ en ambos casos se tiene una diferencia significativa en las estimaciones de cobertura. Es importante destacar la eficiencia de la aproximación Bootstrap del tipo Percentil -t Simétrizado, el cual recoge significativamente la asimetría (ya que es un método que invierte su esfuerzo en corregir el sesgo además de la escala) de la muestra de la población exponencial y cuya cobertura se destaca entre el resto de las aproximaciones. Se puede ver además que en cualquier caso las aproximaciones bootstrap percentil-t presentan una mejor aproximación, pues como ya se ha mencionado en secciones anteriores, corrigen el efecto de asimetría presente en este tipo de población.

Intervalos		B=100	B=1000
Intervalo Normal Estándar Unilateral	(J0)	0.87119	0.87119
Aproximación Percentil Unilateral	(J1)	0.85993	0.86517
Aproximación Percentil-t Unilateral	(J2)	0.88970	0.89608
Intervalo Normal Estándar Bilateral	(I0)	0.88989	0.88989
Aproximación Percentil Bilateral	(I1)	0.86921	0.88344
Aproximación Percentil-t Bilateral	(I2)	0.88280	0.89734
Aproximación Percentil-t Simétrizado	(I3)	0.89205	0.89746
Intervalo Bootstrap Estándar	(IBN)	0.88389	0.88698
Aproximación Percentil de Efron	(EPA)	0.87063	0.88621
Aproximación Percentil Hall	(HPA)	0.86921	0.88344
Aproximación Percentil Sesgo Corregido	(BPC)	0.86729	0.88713

Cuadro 3.1: Intervalos de Confianza Normales y Bootstrap para la media $EX = \mu = 100$ de una población exponencial.

3.2. Aplicación para poblaciones Normal y Ji-Cuadrado

Para dar una visión mejor de los comentarios anteriores, se aplicará el mismo procedimiento a dos nuevas poblaciones. La primera población será generada siguiendo una distribución normal y la segunda una distribución, Ji-cuadrado modificada.

1. En primer lugar para una población Normal $X \in N(0,1)$, esto es, $EX = \mu = 0$, parámetro de interés. Se procede de manera similar al ejemplo anterior:

   a) Para $n = 10$ y B = 100 y 1000 se obtiene,

   ```
   > coberturas(B=c(100,1000),n=10,alpha=0.10,N=1000,valuet=0)
   ```

   ```
            J0    J1    J2   I0   I1    I2    I3    IBN   EPA  HPA   BPC
   B=100  0.857 0.853 0.860 0.82 0.813 0.857 0.878 0.818 0.806 0.813 0.806
   B=1000 0.857 0.859 0.867 0.82 0.817 0.869 0.878 0.821 0.828 0.817 0.827
   ```

   b) Para $n = 100$ y $B = 100$ y 1000 se obtiene,

   ```
   > coberturas(B=c(100,1000),n=100,alpha=0.10,N=1000,valuet=0)
   ```

```
           J0    J1    J2   I0    I1    I2    I3    IBN    EPA   HPA   BPC
B=100   0.905 0.900 0.901 0.9 0.878 0.887 0.893 0.898 0.878 0.878 0.880
B=1000 0.905 0.907 0.914 0.9 0.900 0.903 0.899 0.898 0.890 0.900 0.894
```

En este caso, se puede observar que para un valor de $n = 10$ (pequeño) se evidencia una aproximación pobre de la cobertura en una gran parte de los intervalos de confianza, aún si se aumenta el número de remuestras. Sin embargo, cuando se aumenta el tamaño de la muestra, $n = 100$, los resultados presentan una mejoría. La población en estudio es una población Normal (distribución con densidad simétrica), por ello, la estimación de la cobertura en el caso del intervalo bilateral I_0 (intervalo normal estándar bilateral) es evidentemente la mejor. No así para el resto de los intervalos de los que se obtienen resultados importantes, pero no mejores que I_0, es de destacar el intervalo bilateral del tipo percentil I_3, aunque su aproximación es aceptable, éste aporta mejores resultados para aquellas poblaciones con una marcada densidad asimétrica, pues corrige el sesgo y la escala, lo cual no es el caso.

2. En segundo lugar para una población $X \in a\frac{\chi^2_\kappa}{\kappa}$ con $EX = a = 2$ y $\kappa = 2$

   En esta ocasión solo se aplicará el estudio para $n = 100$ y $B = 100$ y 1000 obteniéndose

```
> coberturas(B=c(100,1000),n=100,alpha=0.10,N=1000,valuet=0)
```

```
           J0    J1    J2   I0    I1    I2    I3   IBN    EPA   HPA    BPC
B=100   0.872 0.864 0.889 0.883 0.854 0.875 0.882 0.877 0.865 0.854 0.874
B=1000 0.872 0.871 0.896 0.883 0.880 0.902 0.893 0.884 0.885 0.880 0.887
```

Para esta población claramente con densidad levemente asimétrica se corrobora la eficacia de los métodos de aproximación percentil-t y percentil-t Simétrizado, frente al resto de los métodos y aproximaciones.

3.3. Aplicación a los datos de censos de Venezuela

Finalmente aplicaremos los intervalos del tipo bootstrap a un ejemplo sencillo, tomando datos reales extraídos de la página web del Instituto Nacional de Estadística. Los datos son referidos a la población, por estados, correspondientes a los censos 1990 y 2001 y se muestra en la tabla

(3.2), en el caso del censo 2001 no se incluye la población empadronada en el Censo de Comunidades Indígenas. fuente (`http://www.ine.gov.ve/censo/fichascenso/nacional_II.asp`). Cada par corresponde a un estado de Venezuela, a los cuales denotaremos por u al censo 2001 y por v al censo 1990. Una gráfica de los datos se muestra en la figura (3.1) Se está interesado

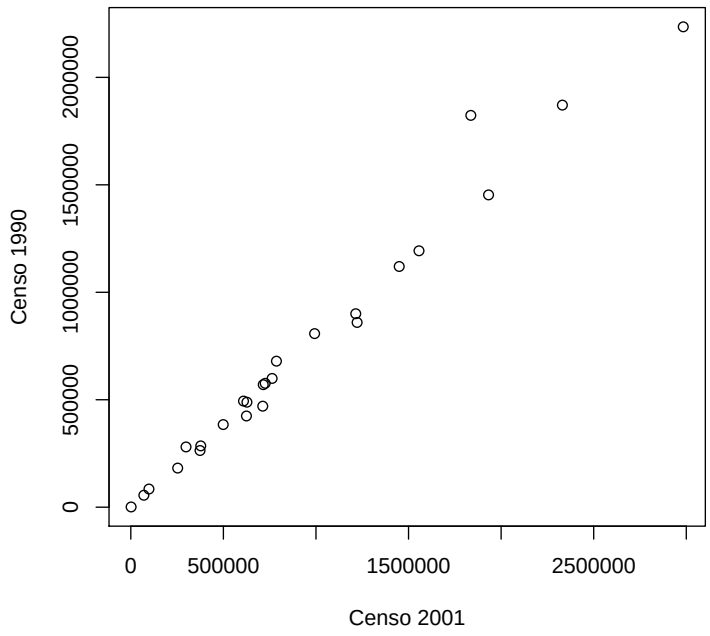

Figura 3.1: Población de Venezuela por Estados, Censos 1990 y 2001

en la proporción de las medias de los censos de 2001 y 1990, ya que esto permitirá estimar el total de población de Venezuela en 2001 a partir de la de 1990. Es claro que el margen de años entre que se aplica el censo es amplio, sin embargo, se tratará de obtener un estimado para ver si esto influye de manera significativa.

Si las poblaciones de los estados conforman una muestra aleatoria con (U, V) denotando el par de valores de la población para un estado seleccionado aleatoriamente, entonces la población total de 2001 es el producto del total de la población de 1990 y la proporción de los valores esperados $\theta = E(U)/E(V)$. Por lo tanto, el parámetro de interés es dicha proporción o razón.

En este caso no es obvio un modelo paramétrico para la distribución conjunta de (U, V), de manera que es natural estimar θ por su análogo empírico, esto es, $T = \bar{U}/\bar{V}$, la proporción o razón

Estado	**Censo 2001**	**Censo 1990**
Distrito Capital	1836286	1823222
Amazonas	70464	55717
Anzoátegui	1222225	859758
Apure	377756	285412
Aragua	1449616	1120132
Barinas	624508	424491
Bolívar	1214846	900310
Carabobo	1932168	1453232
Cojedes	253105	182066
Delta Amacuro	97987	84564
Falcón	763188	599185
Guárico	627086	488623
Lara	1556415	1193161
Mérida	715268	570215
Miranda	2330872	1871093
Monagas	712626	470157
Nueva Esparta	373851	263748
Portuguesa	725740	576435
Sucre	786483	679595
Táchira	992669	807712
Trujillo	608563	493912
Vargas	298109	280439
Yaracuy	499049	384536
Zulia	2983679	2235305
Dependencias Federales	1651	1123
Total	**23054210**	**18104143**

Cuadro 3.2: Población de Venezuela por Estados, Censos 1990 y 2001

de las medias muestrales. Bajo incertidumbre del modelo, se dispondrá solo de la información que los datos aporten. Por ello, es pertinente un análisis no-paramétrico (bootstrap) para aproximar intervalos de confianza respecto del parámetro de interés.

Por conveniencia se considerará un subconjunto de los datos en la tabla (3.2), que comprende diez datos (estados) seleccionados aleatoriamente:

cada remuestra simulada, se obtiene seleccionando aleatoriamente un dado j^* del conjunto $1, ldots, n$ con igual probabilidad y tomando $(u^*, v^*) = (u_{j*}, v_{j*})$.

A continuación mostramos los intervalos de confianza, calculados con la función (ver apéndice (3.3))

```
> bootstrapP9_Ratio(muestra=,alpha=,B=)
```

Obteniéndose

```
$`media de la muestra`
[1] 1.351170


$`Desviación estándar`
[1] 0.09766815


$`Sesgo estimado`
[1] 0.004020150


$`Media Bootstrap`
[1] 1.355190


$`Desviación Bootstrap`
[1] 0.02646236
```

Se observa que de los intervalos obtenidos el que más o los que más aproximan la relación de que el producto de la población total del censo 1990 y la razón o proporción de los promedios obtenida es la población del Censo 2001, son los intervalos del tipo percentil, claro está aquí se refleja además que el margen en que se aplican los censos incide en los resultados.

Intervalos	Límite Inferior	Limite Superior
Intervalo Normal Estándar Unilateral	-Inf	1.390751
Aproximación Percentil Unilateral	-Inf	1.379003
Aproximación Percentil-t Unilateral	-Inf	1.381717
Intervalo Normal Estándar Bilateral	1.300368	1.401972
Aproximación Percentil Bilateral	1.300080	1.388345
Aproximación Percentil-t Bilateral	1.292295	1.393690
Aproximación Percentil-t Simétrizado	1.301197	1.401143
Intervalo Bootstrap Estándar	1.307643	1.394696
Aproximación Percentil de Efron	1.313994	1.402259
Aproximación Percentil Hall	1.300080	1.388345
Aproximación Percentil Sesgo Corregido	1.310027	1.391580

Es claro que los intervalos de confianza del tipo normal no difieren en mayor grado de los del tipo bootstrap. El aporte que estos intervalos brindan es que para poblaciones cuyas distribuciones presenten características asimétricas, la metodología bootstrap mejora sustancialmente las estimaciones.

Actualmente la Metodología Bootstrap, no solo es aplicada para determinar intervalos de confianza. sus aplicaciones son de vital importancia en la estadística no-paramétrica, como herramienta indispensable para la consecución de muchos de sus resultados.

Una aplicación que se esta implementando, con el fin de presentar un futuro trabajo de investigación, está referida a estimar funciones mediante Regresión Polinómica Local. Una vez obtenidas estas funciones, el objetivo es determinar bandas de confianza del tipo bootstrap para dichas funciones en su soporte. Estas estimaciones están enmarcadas dentro de un proyecto que comprende otras más. El objetivo final, es aplicar test de Bondad de ajuste para determinar el mejor modelo posible para el estudio de los Tipos de Interés, es decir, para modelos dependientes del tiempo, dados por la ecuación diferencial estocástica:

$$dR_t = \mu(t, r_t)dt + \sigma(t, r_t)dW_t$$

con W_t un movimiento browniano estándar, donde $\mu(t, r_t)$ y $\sigma(t, r_t)$ son las funciones *drift* y *difusión* (o función de volatilidad) del proceso r_t, las cuales se estiman mediante dos procedi-

mientos que se especifican en ([11]). Entre los procedimientos se detalla además como obtener las bandas de confianza mediante metodología bootstrap, los cuales son en principio similares a como se hacen para un parámetro, en este caso para cada $t = t_0$ se estima por ejemplo la función $\mu(t_0, r_{t_0})$ y se calcula un intervalo de confianza para la misma, este procedimiento se aplica a cada punto $t = t_0$ en un dominio determinado, que en este caso es referido al período en años, meses, o días en estudio.

Pues bien, el objetivo final es poder aplicar estas técnicas y la metodología bootstrap en toda su amplitud, a los tipos de interés de nuestros índices económicos aportados por el Banco Central de Venezuela, dichos tipos de interés se refieren a los que se negocian en operaciones interbancarias.

Este es un proyecto que está en ejecución, y que presentaremos con todos los detalles como un avance con mayor complejidad de este primer trabajo con fines introductorios.

Apéndice

a continuación se presentan una serie de estratos de códigos que pretenden dar a entender como funcionan individualmente las rutinas para determinar cada intervalo por separado o el conjunto total de intervalos.

Ejemplo 1 *Este ejemplo muestra una secuencia de instrucciones (dadas en el capítulo 2) que permiten determinar, a través de la linea de comando del software* ***R****, los intervalos de confianza normales, para una distribución exponencial con parámetro dado.*

```
> alpha = 0.1
> muestra = rexp(n = 50, rate = 1/100)
> n = length(muestra)
> media = mean(muestra)
> desv = sd(muestra)
> aux = desv * qnorm(1 - (alpha/2))/sqrt(n)
> J0 = c(-Inf, media - desv * qnorm(alpha)/sqrt(n))
> I0 = c(media - aux, media + aux)
> J0

[1]     -Inf 146.0798

> I0

[1]  95.33155 152.38004
```

Ejemplo 2 *Este conjunto de instrucciones permiten determinar, a través de la linea de comando del software* ***R****, los intervalos de confianza bootstrap de tipo percentil unilaterales, para una distribución exponencial con parámetro dado.*

```
> B = 1000
> alpha = 0.1
> muestra = rexp(n = 50, rate = 1/100)
> n = length(muestra)
> media = mean(muestra)
> replicasI1 = numeric(B)
> for (i in 1:B)
{
+     remuestra = sample(muestra, n, replace = TRUE)
+     replicasI1[i] = sqrt(n) * (mean(remuestra) - media)
+
}
> J1 = c(-Inf, media - quantile(replicasI1,probs=alpha)/sqrt(n))
> J1

[1]     -Inf 120.7546

> B = 1000
> alpha = 0.1
> muestra = rexp(n = 50, rate = 1/100)
> n = length(muestra)
> media = mean(muestra)
> desv = sd(muestra)
> replicasI2 = numeric(B)
> for (i in 1:B)
{
+     remuestra = sample(muestra, n, replace = TRUE)
+     replicasI2[i] = sqrt(n) * (mean(remuestra) - media)/sd(remuestra)
+
}
> J2 = c(-Inf, media - desv * quantile(reporderI2,probs=alpha)/sqrt(n))
> J2
```

```
[1]     -Inf 120.5689
```

Ejemplo 3 *Este conjunto de instrucciones permiten determinar, a través de la linea de comando del software* ***R****, los intervalos de confianza bootstrap de tipo percentil bilaterales, Efron, Hall y sesgo Corregido, para una distribución exponencial con parámetro dado.*

```
> B = 1000
> alpha = 0.1
> muestra = rexp(n = 50, rate = 1/100)
> n = length(muestra)
> media = mean(muestra)
> replicasI1 = numeric(B)
> for (i in 1:B)
{
+     remuestra = sample(muestra, n, replace = TRUE)
+     replicasI1[i] = sqrt(n) * (mean(remuestra) - media)
+
}
> I1=(media-quantile(replicasI1,probs=1-(alpha/2))/sqrt(n)
+   ,media-quantile(replicasI1,probs=(alpha/2))/sqrt(n))
> I1

[1]  71.68004 117.84065

> B = 1000
> alpha = 0.1
> muestra = rexp(n = 50, rate = 1/100)
> n = length(muestra)
> media = mean(muestra)
> desv = sd(muestra)
> replicasI2 = numeric(B)
> for (i in 1:B)
{
```

```
+     remuestra = sample(muestra, n, replace = TRUE)
+     replicasI2[i] = sqrt(n) * (mean(remuestra) - media)/sd(remuestra)
+
}
> I2=c(media-desv*quantile(replicasI2,probs=1-(alpha/2))/sqrt(n)
+   ,media-desv*quantile(replicasI2,probs=alpha/2)/sqrt(n))
> I2

[1]  74.50119 125.85660

> B = 1000
> alpha = 0.1
> muestra = rexp(n = 50, rate = 1/100)
> n = length(muestra)
> media = mean(muestra)
> desv = sd(muestra)
> replicasI3 = numeric(B)
> for (i in 1:B)
{
+     remuestra = sample(muestra, n, replace = TRUE)
+     replicasI3[i] = sqrt(n) * abs((mean(remuestra) -media)/sd(remuestra))
+
}
>I3=c(media-desv*quantile(replicasI3,probs=1-alpha)/sqrt(n)
+    ,media+desv*quantile(replicasI3,probs=1-alpha)/sqrt(n))
> I3

[1]  80.23012 119.30597

> muestra = rexp(100, rate = 1/100)
> Intervalos = bootstrapP9(muestra = muestra, alpha = 0.1, B = 1000)
> Intervalos$Intervalos
```

```
$Intervalos
```

	Límite Inferior	Limite Superior
Intervalo Normal Estándar Unilateral	-Inf	108.2193
Aproximación Percentil Unilateral	-Inf	107.6273
Aproximación Percentil-t Unilateral	-Inf	110.1668
Intervalo Normal Estándar Bilateral	76.74753	112.1264
Aproximación Percentil Bilateral	77.01824	111.6986
Aproximación Percentil-t Bilateral	78.92752	115.2725
Aproximación Percentil-t Simetrizado	76.20746	112.6664

```
> B = 1000
> alpha = 0.1
> muestra = rexp(n = 50, rate = 1/100)
> n = length(muestra)
> media = mean(muestra)
> replicas = numeric(B)
> for (i in 1:B)
{
+     remuestra = sample(muestra, n, replace = TRUE)
+     replicas[i] = (mean(remuestra))
+
}
> mboot = (1/B) * sum(replicas)
> sdboot = sqrt(sum((replicas - mboot)^2)/B)
> IBN = c(media - sdboot * qnorm(alpha/2, lower.tail = FALSE),
+     media + sdboot * qnorm(alpha/2, lower.tail = FALSE))
> IBN

[1]  87.82217 134.01518

> B=1000
> alpha=0.10
> muestra=rexp(n=50,rate=1/100)
```

```
> n=length(muestra)
> media=mean(muestra)
> replicas=numeric(B)
> for (i in 1:B)
{
+
+       remuestra=sample(muestra,n,replace=TRUE)
+       replicas[i]=(mean(remuestra))
+
}
> mboot=(1/B)*sum(replicas)
> sdboot=sqrt(sum((replicas-mboot)^2)/B)
> EPA=(quantile(replicas,probs=c((alpha/2),1-alpha/2)))
> HPA=c(2*media-quantile(replicas,probs=c((1-alpha/2),alpha/2)))
> EPA

[1]  71.92332    124.8568

> HPA

[1]  71.17636    124.1098

> muestra = rexp(50, rate = 1/100)
> n = length(muestra)
> media = (1/n) * sum(muestra)
> desv = sqrt(sum((muestra - media)^2)/(n - 1))
> replicas = numeric(B)
> propor = numeric(B)
> for (i in 1:B)
{
+     remuestra = sample(muestra, n, replace = TRUE)
+     replicas[i] = mean(remuestra)
+     if (replicas[i] > media)
```

```
      {
        + propor[i] = 1
        +
       }
+
}
> p=mean(propor)
> z0=qnorm(1-p)
> a=pnorm(2*z0+qnorm(alpha/2,lower.tail=FALSE))
> b=pnorm(2*z0-qnorm(alpha/2,lower.tail=FALSE))
> BPC=c(quantile(replicas,probs=b),quantile(replicas,probs=a))
> BPC

[1]  79.94284 121.48035

> muestra = rexp(50, rate = 1/100)
> InterBoot = bootstrapP9(muestra = muestra, alpha = 0.1, B = 1000)
> InterBoot$`Intervalos Bootstrap`

$`Intervalos Bootstrap`
                                    Lower limit Upper limit
Intervalo Bootstrap Estándar           67.64637    108.6652
Aproximación Percentil de Efron        69.10577    110.3421
Aproximación Percentil Hall            65.96955    107.2058
Aproximación Percentil Sesgo Corregido 67.37747    107.6817
```

A continuación se presentan los códigos de las funciones usadas para determinar el conjunto de intervalos descritos en el capítulo 2

```
> bootstrapP9 = function(muestra, alpha, B)
{
  n = length(muestra)
  media = mean(muestra)
```

```
desv = sd(muestra)
replicasI1 = numeric(B)
replicasI2 = numeric(B)
replicasI3 = numeric(B)
J0 = c(-Inf, media - desv * qnorm(alpha)/sqrt(n))
I0 = c(media - desv * (qnorm(1 - (alpha/2))/sqrt(n)),
media + desv * (qnorm(1 - (alpha/2))/sqrt(n)))
for (i in 1:B)
{
  remuestra = sample(muestra, n, replace = TRUE)
  replicas[i]=mean(remuestra)
  replicasI1[i] = sqrt(n) * (mean(remuestra) - media)
  replicasI2[i] = sqrt(n) * (mean(remuestra) - media)/sd(remuestra)
  replicasI3[i] = sqrt(n) * abs((mean(remuestra) - media)/sd(remuestra))
  if (replicas[i]>media)
  {
    propor[i]=1
   }
}
mboot=(1/B)*sum(replicas)
sdboot=sqrt(sum((replicas-mboot)^2)/B)
sesgo=mboot-media

J1=c(-Inf,media-quantile(replicasI1,probs=alpha)/sqrt(n))
I1=(media-quantile(replicasI1,probs=1-(alpha/2))/sqrt(n),
       media-quantile(replicasI1,probs=(alpha/2)))


J2=c(-Inf,media-desv*quantile(replicasI2,probs=alpha)/sqrt(n))
I2=c(media-desv*quantile(replicasI2,probs=1-(alpha/2))/sqrt(n),
       media-desv*quantile(replicasI2,probs=alpha/2)/sqrt(n))
```

```
I3=c(media-desv*quantile(replicasI3,probs=1-alpha)/sqrt(n),
       media+desv*quantile(replicasI3,probs=1-alpha)/sqrt(n))

IBN=c(media-sdboot*(qnorm(alpha/2,lower.tail=FALSE)),
       media+sdboot*(qnorm(alpha/2,lower.tail=FALSE)))

EPA=(quantile(replicas,probs=c((alpha/2),1-alpha/2)))

HPA=c(2*media-quantile(replicas,probs=c((1-alpha/2),alpha/2)))

p=mean(propor)
z0=qnorm(1-p)
a=pnorm(2*z0+qnorm(alpha/2,lower.tail=FALSE))
b=pnorm(2*z0-qnorm(alpha/2,lower.tail=FALSE))

BPC=c(quantile(replicas,probs=b),quantile(replicas,probs=a))


mres=matrix(c(J0,J1,J2,I0,I1,I2,I3),nrow=7,ncol=2,byrow=TRUE,
     dimnames=list(c("Intervalo Normal Estándar Unilateral",
     "Aproximación Percentil Unilateral",
     "Aproximación Percentil-t Unilateral",
     "Intervalo Normal Estándar Bilateral",
     "Aproximación Percentil Bilateral",
     "Aproximación Percentil-t Bilateral",
     "Aproximación Percentil-t Simetrizado"),
     c("Límite Inferior","Limite Superior")))

mresboot=matrix(c(IBN,EPA,HPA,BPC),nrow=4,ncol=2,byrow=TRUE,
```

```
        dimnames=list(c("Intervalo Bootstrap Estándar",
        "Aproximación Percentil de Efron",
        "Aproximación Percentil Hall",
        "Aproximación Percentil Sesgo Corregido"),
        c("Lower limit","Upper limit")))
  return(list("media de la muestra"=media,"Intervalos"=mres,
        "Intervalos Bootstrap"=mresboot,
        "Desviación estándar"=desv,
        "Sesgo estimado"=sesgo,"Media Bootstrap"=
        mboot,"Desviación Bootstrap"=sdboot))

}


coberturas=function(B,n,alpha,N,valuet)
{
    p=length(B)
    m=length(n)
    ns=paste("n=",n,sep="")
    Bs=paste("B=",B,sep="")

    cJ0=matrix(0,nrow=p,ncol=m,dimnames=list(Bs,ns))
    cJ1=matrix(0,nrow=p,ncol=m,dimnames=list(Bs,ns))
    cJ2=matrix(0,nrow=p,ncol=m,dimnames=list(Bs,ns))
    cI0=matrix(0,nrow=p,ncol=m,dimnames=list(Bs,ns))
    cI1=matrix(0,nrow=p,ncol=m,dimnames=list(Bs,ns))
    cI2=matrix(0,nrow=p,ncol=m,dimnames=list(Bs,ns))
    cI3=matrix(0,nrow=p,ncol=m,dimnames=list(Bs,ns))
    cIBN=matrix(0,nrow=p,ncol=m,dimnames=list(Bs,ns))
    cEPA=matrix(0,nrow=p,ncol=m,dimnames=list(Bs,ns))
    cHPA=matrix(0,nrow=p,ncol=m,dimnames=list(Bs,ns))
```

```
cBPC=matrix(0,nrow=p,ncol=m,dimnames=list(Bs,ns))

for (k in 1:m)
{
   muestra=matrix(0,nrow=N,ncol=n[k])
   for (i in 1:N)
   {
     muestra[i,]=rexp(n[k],rate=1/100)
     #2*rchisq(n[k],df=2,ncp=0)/2 ,rexp(n[k],rate=1/100)
   }

   for (l in 1:p)
   {
        coberturaJ0=0
        coberturaJ1=0
        coberturaJ2=0
        coberturaI0=0
        coberturaI1=0
        coberturaI2=0
        coberturaI3=0
        coberturaEPA=0
        coberturaIBN=0
        coberturaHPA=0
        coberturaBPC=0

        for (j in 1:N)
        {

         c=bootstrapP9(muestra=muestra[j,],alpha,B[l])
         cob=c$"Intervalos"
         cobBoot=c$"Intervalos Bootstrap"
```

```
if (valuet<cob[1,2])
        {  coberturaJ0=coberturaJ0+1 }

if (valuet<cob[2,2])
        {  coberturaJ1=coberturaJ1+1 }

if (valuet<cob[3,2])
        {  coberturaJ2=coberturaJ2+1 }

if ((cob[4,1]<valuet)&&(valuet<cob[4,2]))
        {  coberturaI0=coberturaI0+1 }

if ((cob[5,1]<valuet)&&(valuet<cob[5,2]))
        {  coberturaI1=coberturaI1+1 }

if ((cob[6,1]<valuet)&&(valuet<cob[6,2]))
        {  coberturaI2=coberturaI2+1 }

if ((cob[7,1]<valuet)&&(valuet<cob[7,2]))
        {  coberturaI3=coberturaI3+1 }

if ((cobBoot[1,1]<valuet)&&(valuet<cobBoot[1,2]))
        {  coberturaIBN=coberturaIBN+1 }

if ((cobBoot[2,1]<valuet)&&(valuet<cobBoot[2,2]))
        {  coberturaEPA=coberturaEPA+1 }

if ((cobBoot[3,1]<valuet)&&(valuet<cobBoot[3,2]))
        {  coberturaHPA=coberturaHPA+1 }
```

```
            if ((cobBoot[4,1]<valuet)&&(valuet<cobBoot[4,2]))
                    {  coberturaBPC=coberturaBPC+1 }

          }
          cJ0[l,k]=(coberturaJ0)/N
          cJ1[l,k]=(coberturaJ1)/N
          cJ2[l,k]=(coberturaJ2)/N
          cI0[l,k]=(coberturaI0)/N
          cI1[l,k]=(coberturaI1)/N
          cI2[l,k]=(coberturaI2)/N
          cI3[l,k]=(coberturaI3)/N
          cIBN[l,k]=(coberturaIBN)/N
          cEPA[l,k]=(coberturaEPA)/N
          cHPA[l,k]=(coberturaHPA)/N
          cBPC[l,k]=(coberturaBPC)/N
       }
    }
X=matrix(c(cJ0,cJ1,cJ2,cI0,cI1,cI2,cI3,cIBN,cEPA,cHPA,cBPC),
nrow=p,ncol=11,byrow=F,dimnames=list(c(Bs),
c("J0","J1","J2","I0","I1","I2","I3","IBN","EPA","HPA","BPC")))
return(X)

Estudio=function(M,B,n,alpha,N,valuet)
{

    m=length(B)
    t0=numeric(m)
    t1=numeric(m)
    t2=numeric(m)
    t3=numeric(m)
```

```
t4=numeric(m)
t5=numeric(m)
t6=numeric(m)
t7=numeric(m)
t8=numeric(m)
t9=numeric(m)
t10=numeric(m)


ns=paste("n=",n,sep="")
Bs=paste("B=",B,sep="")

for (j in 1:M)
{
    c=coberturas(B=B,n=n,alpha=alpha,N=N,valuet=valuet)

    for (k in 1:m)
    {
        t0[k]=t0[k]+c[k,1]

        t1[k]=t1[k]+c[k,2]

        t2[k]=t2[k]+c[k,3]

        t3[k]=t3[k]+c[k,4]

        t4[k]=t4[k]+c[k,5]

        t5[k]=t5[k]+c[k,6]

        t6[k]=t6[k]+c[k,7]
```

```
        t7[k]=t7[k]+c[k,8]

        t8[k]=t8[k]+c[k,9]

        t9[k]=t9[k]+c[k,10]

        t10[k]=t10[k]+c[k,11]

    }

}
t0=t0/M;
t1=t1/M;
t2=t2/M;
t3=t3/M;
t4=t4/M;
t5=t5/M;
t6=t6/M;
t7=t7/M;
t8=t8/M;
t9=t9/M;
t10=t10/M;
total=matrix(c(t0,t1,t2,t3,t4,t5,t6,t7,t8,t9,t10),
nrow=11,ncol=m,byrow=TRUE,
dimnames=list(c("Intervalo Normal Estandar Unilateral",
"Aproximación Percentil Unilateral",
"Aproximación Percentil-t Unilateral",
"Intervalo Normal Estandar Bilateral",
"Aproximación Percentil Bilateral",
"Aproximación Percentil-t Bilateral",
```

```
    "Aproximación Percentil-t Simetrizado",
    "Intervalo Bootstrap Estandar",
    "Aproximación Percentil de Efron",
    "Aproximación Percentil Hall",
    "Aproximación Percentil Sesgo Corregido"),c(Bs)))
    F=cbind(n=n, X=total)
    return(total)
}
```

A continuación se presentan los códigos de las funciones usadas para determinar el conjunto de intervalos para la proporción de los promedios de la población de Venezuela en los censos 2001 y 1990

```
ratio <- function(d){
     mean(d$V1)/mean(d$V2)}
va<- function(d){sum((d$V1/d$V2-ratio(d))^2)/(9)}

bootstrapP9_Ratio=function(muestra,alpha,B)
{   media=ratio(muestra) #(1/n)*sum(muestra)
    desv=sqrt(va(muestra))
    replicasI1=numeric(B)
    replicasI2=numeric(B)
    replicasI3=numeric(B)
    replicas=numeric(B)
    propor=numeric(B)
    imuestra=numeric(B)
    n=dim(muestra)[1]
    ##Calculo de Intervalos de confianza Normales para la media

        ## I.C Unilateral Aproximación Normal
```

```
        J0=c(-Inf,media-desv*qnorm(alpha)/sqrt(n))

    ## I.C Bilateral Aproximación Normal

        aux=desv*qnorm(1-alpha/2)/sqrt(n)

        I0=c(media-aux,media+aux)


##Calculo de Intervalos de confianza Bootstrap

    ## Aproximación Percentil

        for (i in 1:B)
        { j=seq(1,n,by=1)
          k=sample(j,n,replace=TRUE)
          remuestra=muestra[k,]
          replicas[i]=ratio(remuestra)
          replicasI1[i]=sqrt(n)*(ratio(remuestra)-media)
          replicasI2[i]=sqrt(n)*(ratio(remuestra)-media)/sqrt(va(remuestra))
          replicasI3[i]=sqrt(n)*abs((ratio(remuestra)-media)/sqrt(va(remuestra)))
          if (replicas[i]>media)
          { propor[i]=1  }

        }
        mboot=(1/B)*sum(replicas)
        sdboot=sqrt(sum((replicas-mboot)^2)/B)
        sesgo=mboot-media

        ## I.C Unilateral
```

```
        J1=c(-Inf,media-quantile(replicasI1,probs=alpha)/sqrt(n))

    ## I.C Bilaterales:

        I1=c(media-quantile(replicasI1,probs=1-alpha/2)/sqrt(n),
        media-quantile(replicasI1,probs=alpha/2)/sqrt(n))

## Aproximación Percentil-t

    ## I.C Unilateral

        J2=c(-Inf,media-desv*quantile(replicasI2,probs=alpha)/sqrt(n))

    ## I.C Bilaterales:

        I2=c(media-desv*quantile(replicasI2,probs=1-alpha/2)/sqrt(n),
        media-desv*quantile(replicasI2,probs=alpha/2)/sqrt(n))

## Aproximación Percentil-t Simétrizado

    ## I.C Bilaterales:

        aux=desv*quantile(replicasI3,probs=1-alpha)/sqrt(n)

        I3=c(media-aux,media+aux)

## Aproximación Bootstrap Estándar

        aux=sdboot*qnorm(alpha/2,lower.tail=FALSE)

        IBN=c(media-aux,media+aux)
```

```
## Aproximación Percentil

      EPA=(quantile(replicas,probs=c((alpha/2),1-alpha/2)))

## Aproximación Percentil de Hall (1992)

      HPA=c(2*media-quantile(replicas,probs=c(1-alpha/2,alpha/2)))


## Aproximación Percentil con Corrección de Sesgo.

    p=mean(propor)
    z0=qnorm(1-p)
    b=pnorm(2*z0+qnorm(alpha/2,lower.tail=FALSE))
    a=pnorm(2*z0-qnorm(alpha/2,lower.tail=FALSE))

    BPC=c(quantile(sort(replicas),probs=a),quantile(sort(replicas),probs=b))


mres=matrix(c(J0,J1,J2,I0,I1,I2,I3),
nrow=7,ncol=2,byrow=TRUE,dimnames=list(c("Intervalo Normal Estándar Unilateral",
"Aproximación Percentil Unilateral",
"Aproximación Percentil-t Unilateral",
"Intervalo Normal Estándar Bilateral",
"Aproximación Percentil Bilateral",
"Aproximación Percentil-t Bilateral",
"Aproximación Percentil-t Simétrizado"),
c("Límite Inferior","Limite Superior")))
mresboot=matrix(c(IBN,EPA,HPA,BPC),
nrow=4,ncol=2,byrow=TRUE,
```

```
    dimnames=list(c("Intervalo Bootstrap Estándar",
    "Aproximación Percentil de Efron",
    "Aproximación Percentil Hall","Aproximación Percentil Sesgo Corregido"),
    c("Lower limit","Upper limit")))

    return(list("media de la muestra"=media,
    "Intervalos"=mres,"Intervalos Bootstrap"=mresboot,
    "Desviación estándar"=desv,"Sesgo estimado"=sesgo,
    "Media Bootstrap"=mboot,"Desviación Bootstrap"=sdboot))

}
```

Conclusiones

A través de la simulación estadística y el uso del software R se ha podido corroborar la teoría respecto de las aproximaciones de intervalos de confianza Bootstrap. A partir de las rutinas implementadas en R se pudo obtener en un principio los intervalos de confianza Bootstrap del tipo Percentil para luego comparar su comportamiento entre los distintos métodos percentil, entre ellos: percentil-t,percentil-t simétrizado y los intervalos de confianza normales de la teoría clásica. Se evidenció, para el caso en estudio, una población exponencial, una normal y una ji-cuadrado, las bondades de los métodos del tipo percentil, especialmente el caso de los intervalos del tipo percentil los cuales muestran ser eficientes dadas sus características en cuanto a corrección de escala y sesgo. Por otro lado, para un escenario, que comprende la población de Venezuela registrada en los dos últimos censos, 1990 y 2001, se hizo un estudio con el objeto de determinar intervalos de confianza para la proporción de las medias de las poblaciones de dichos censos con la finalidad de obtener un estimado de la población de 2001, a partir del estimado de la razón o proporción de los censos y de la población total del censo 1990.

En cuanto el Alcance de este trabajo, tal como se mencionó en el capítulo 3, el objetivo es dar a conocer, de manera sencilla, mediante las aplicaciones antes mencionadas la metodología Bootstrap, la cual no requiere un conocimiento teórico profundo, más allá de los conceptos básicos de la estadística y que actualmente es de vital importancia sobre todo en el análisis noparamétrico aplicado en diversas areas, tales como Finanzas, Estadística Espacial, Investigación en medio ambiente, ciencias biológicas ([5]), etc.

Es de destacar que el capítulo 3 y los programas en R que se presentan en el apéndice están referidos a probar la eficacia de los mismos, de manera que para los futuros interesados los programas sean útiles para poder reproducir esta técnica en sus investigaciones.

La metodología bootstrap ha recobrado importancia producto del avance de la tecnología, en cuanto a los procesadores computacionales, pues hace más viable su aplicación, hecho importante

ya que para los pioneros de esta metodología se hacia prácticamente imposible dado el gran número de operaciones que requiere la misma, y que hoy en día ha sido subsanado.

Queda por demás invitar a quienes estén interesados en ratificar sus resultados obtenidos con otras técnicas o bien para generar y (o mejorar) resultados nuevos (previos), a usar los algoritmos del capítulo dos y tres, según sea el caso y corroborar de ser necesario las coberturas de los intervalos que determinen en sus estudios y así dar a conocer entre sus colegas investigadores de las bondades de esta técnica y su forma práctica de aplicación.

En la actualidad la metodología bootstrap va más allá de la simple determinación de intervalos de confianza, es intensamente usado en análisis de series de tiempo aplicadas a las finanzas tal y como fue comentado en el capítulo 3, en análisis de regresión, en estimación de funciones, etc, dada su potencialidad y eficacia además de los satisfactorios resultados que arroja cada vez que se es aplicada esta metodología.

Bibliografía

[1] Alonso A., *Intervalos de Confianza Bootstrap para indicadores en regresión logística*, *MedULA*,**10**, 1-4,(2001).

[2] Bai C., y Olsen R., *Discussion of "Theoretical comparison of bootstrap confidence intervals", by P. Hall*, *Ann. Statist*,**16**, 953-956, (1988).

[3] Beran R., *Bootstrap Methods in statistics*, *Jber. d. Dt. Math Verein*,**86**, 14-30, (1984.)

[4] Buckland S.T., *Monte Carlo methods for confidence interval estimation using the bootstrap technique*, *Bull. appl. Statist*,**10**, 194-212, (1983).

[5] Bryan F. J. Manly, *Randomization, Boostrap and Monte Carlo Methods in Biology*, *Chapman& Hall*, Secon Edition, 1997, pags(35-61).

[6] Davison A.C., Hinkley D.V., y Schechtman E.,*Efficient bootstrap simulations*, *Biometrika*,**73** 555-566, (1986).

[7] Davison A.C., Hinkley D.V., *Bootstrap Methods and their applications*, *Cambridge University Press*, 1997.

[8] DiCiccio T.J., y Tibshirani R., *Bootstrap confidence intervals and bootstrap approximations*, *J. Americ. Statist Assoc.*,**82**,163-170, (1987).

[9] Efron B., Tibshirani R., *Bootstrap measures for standard errors, confidence intervals, and other measures of statistical accuracy*, *statistical Science*,**1**, 54-77,(1986.)

[10] Efron B., Tibshirani R., *An Introduction to the Bootstrap*, *Chapman& Hall*,1993.

[11] Fan J., Jiang J., Zhang C., y Zhou Z., *Time-dependent difussion models for term structure dynamics*, Statitice Sinice (2003), no.13, 965–992.

[12] Hall P., *On the Bootstrap and confidence intervals*, *Ann. Statist.*, 14, 1431-1452, (1986)a.

[13] Play L., y Matteucci S., *Intervalos de Confianza Bootstrap del Índice de biodiversidad de Shannon*, *Revista de la Facultad de Agronomía (LUZ)*, 18,222-234,(2001).

[14] Prada José M., *Técnicas de Remuestreo*, *Departamento de Estadística e Investigación 0perativa*, USC, 2006.

[15] Sing K., *On the asymptotic accuracy of Efron's bootstrap*, *Ann. Statist*,**9**,**6**, 1187-1195, (1981).