Cómputo Científico
Tarea IX

Joel Chacón Castillo

Guanajuato, México

1. Punto 1

Los siguientes 15 datos forman una muestra aleatoria de una distribución Gamma con parámetro de forma $\alpha = 3$ y un parámetro de escala $\beta = 2$ (la media es $\alpha\beta$ y la varianza es $\alpha\beta^2$, datos:
$\{14{,}18, 10{,}99, 3{,}38, 6{,}76, 5{,}56, 1{,}26, 4{,}05, 4{,}61, 1{,}78, 3{,}84, 4{,}69, 2{,}12, 2{,}39, 16{,}75, 4{,}19\}$.
Encuentre un intervalo de confianza para la media de la distribución.

Comentarios

Sea T un estadístico, estimador de θ. Suponiendo que se desea un intervalo de confianza para θ, además suponiendo que se conoce la distribución de $T - \theta$ y se pueden calcular sus cuantiles, es decir se puede conocer a y b tales que $P(a \leq T - \theta \leq b) = 1 - \alpha$, y por consiguiente $P(T - b \leq \theta \leq T - a) = 1 - \alpha$, entonces $(T - b, T - a)$ sería un intervalo del $100(1 - \alpha)\,\%$ de confianza para θ. En el entorno de bootstrap se aproxima la distribución $T-\theta$ mediante la distribución empítica $T^*-\theta^*$ por lo tanto una aproximación para a y b sería $t^*_{(B+1)(\alpha/2)} - \theta^*$ y $t^*_{(B+1)(1-\alpha/2)} - \theta^*$ respectivamente. Así, para las expresiones para las aproximaciones bootstrap de los límites de un intervalo de confianza son $t \hat{-} b = \theta^* - [t^*_{(B+1)(1-\alpha/2)} - \theta^*] = 2\theta^*_{(B+1)(1-\alpha/2)}$ y $t \hat{-} a = \theta^* - [t^*_{(B+1)(\alpha/2)} - \theta^*] = 2\theta^*_{(B+1)(\alpha/2)}$ A estos límites se les llama *límites de confianza bootstrap básicos*.

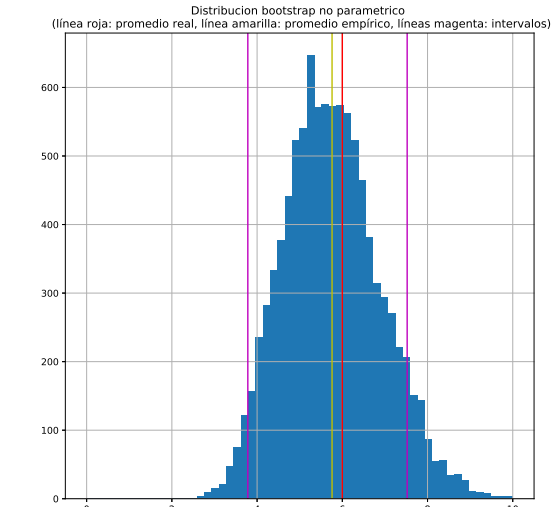
El *método de percentiles* es el más sencillo, sin embargo no se recomienda por no tener una justificaciones sólida. Suponiendo que B muestras bootstrap con los respectivos B valores de θ^*_i. Entonces, se tiene la ventaja de que los límites de confianza están en el mismo soporte que el parámetro de interés.

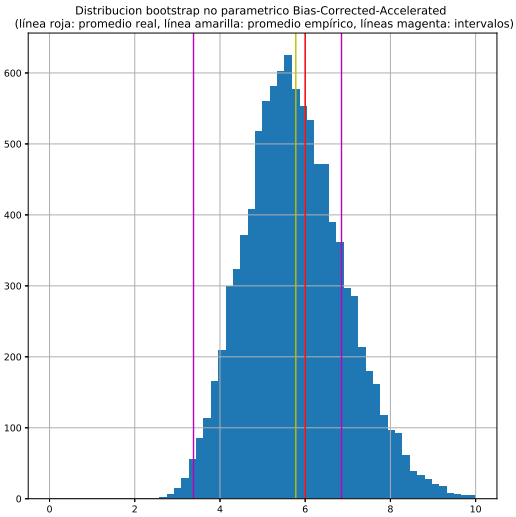
1.1. Codificación

Se programaron los tres métodos, en funciones separadas, donde se considera $B = 10000$, es decir se generan 10000 muestras con reemplazo. En la tabla 1 se presentan la media empírica θ^* y el intervalo de confianza al 90 %, la mejor media empírica pertenece al método paramétrico, esto se debe ya que se comenta que los datos provienen de una distribución gamma. Por otra parte, el método de sesgo corregido parece ser mejor que el método percentil, no obstane el intervalo de confianza es bastante similar. En la figura 1 se muestra la distribución de las muestras (se generan nuevas muestras en cada experimento), principalmente, se percibe que curva de dsitribución es muy similar a una dsitribución Gamma.

	Media teórica	**Media empírica**	θ_{α_1}	θ_{α_2}
No paramétrico - Percentil	6.0	5.7593	3.782	7.522
No paramétrico - Bias Corrected Accelerated	6.0	5.779	3.379	6.852
Percentil - Paramétrico (Gamma)	6.0	6.033	4.023	7.730

Cuadro 1: Cálculo de los intervalos considerando dos métodos no paramétricos bootstrap (percentil y por sesgo corregido) y un método paramétrico asumiendo una distribución Gamma.





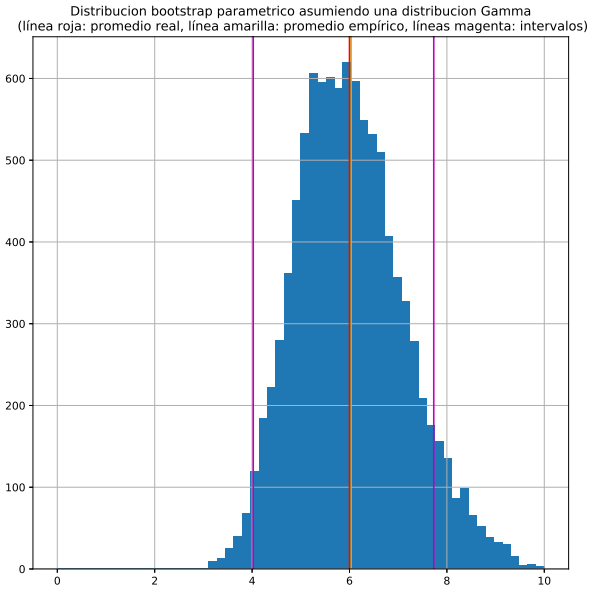


Figura 1: Distribución de las muestras generadas con cada uno de los tres métodos, la líneas verticales indican: rojo - media teórica, amarillo: media empírica, magenta: intervalos al 90 %.

.

2. Punto 2

En la librería *boot* de R accesar los datos *cd*4, los cuales son conteos de células $CD4$ en pacientes HIV-positivos al inicio y después de 1 año de tartamiento con un antiviral.

1. Construya un intervalo bootstrap para el coeficiente de correlación entre los conteos base y los conteos depués del tratamiento.
2. Calcule el coeficiente estimado de correlación corregido por sego usando el Jacknife.

Comentarios

Los intervalos BC_a (*bias corrected and accelerated*) son una mejora con respecto a los intervalos básico y de percentiles. Una mejora en el sentido de que son de *Segundo roden*, esto es, la probabilidad de cobertura de la forma c/n, mientras los otros intervalos son de *segundo orden* pues la cobertura es de la forma $a/sqrtn$. Éste método corrige por sesgo del plug-in, en forma automática. Los límites de confianza quedan en función de los cuantiles de θ^*, similar al método de percentiles, pero también dependen dos parámetros más: a =aceleración y z_0 =correción por sesgo. El intervalo BC_a con cobertura $1-2\alpha$, es $BC_a = (\theta_{\alpha_1}, \theta_{\alpha_2})$ donde $\alpha_1 = \Phi(z_0 + \frac{z_0+z_\alpha}{1-a(z_0+z_\alpha)})$ y $\alpha_2 = \Phi(z_0 + \frac{z_0+z_{1-\alpha}}{1-a(z_0+z_{1-\alpha})})$ Es importante notar que si $a = 0$ y z_0 entonces $\alpha_1 = \alpha$ y $\alpha_2 = 1-\alpha$ que resulta como el método de percentiles. Además z_0 es un cuantil asociado a la proporción de réplicas bootstrap menores que θ (estimador original): $z_0 = \Phi^{-1}(\frac{1}{B}\sum_{i=1} I(\theta^* < \theta))$, es decir z_0 es una medida del sesgo θ^*. si $z_0 = 0$ entonces la mitad de los θ^* está por debajo de θ.

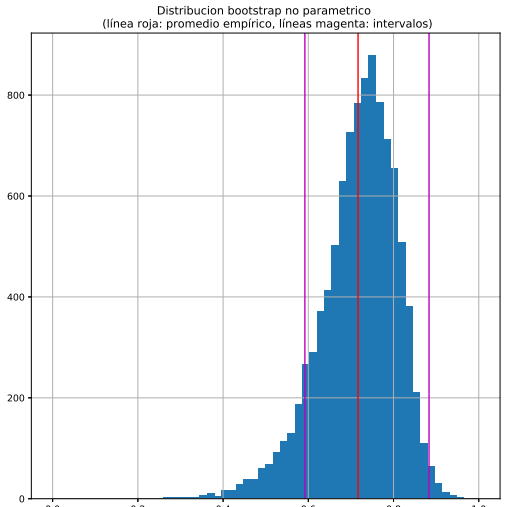
La aceleración (a) es estimada en términos de los valores de Jackknife del estimador θ. $\theta_{(i)}$ = valor de θ eliminando la i-ésima observación. y $\theta_{(.)}$ = promedio de θ_i (en los n datos), entonces a se define como: $a = \frac{\sum_{i=1}^n (\theta_{(.)} - \theta_{(i)})^3}{6(\sum_{i=1}^n (\theta_{(.)} - \theta_{(i)})^2)^{1,5}}$. El parámetro a es una tasa de cambio del error estándar de θ respecto al valor verdadero θ.

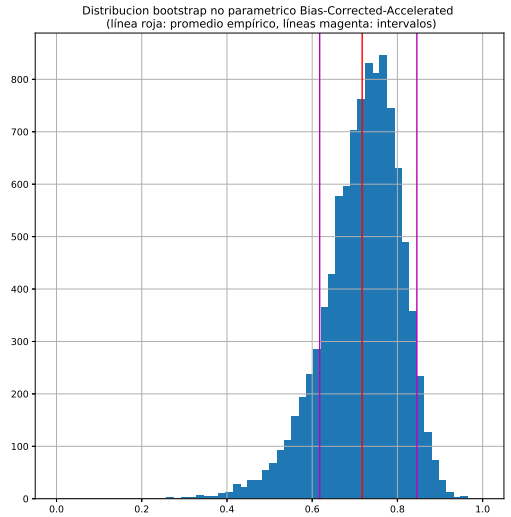
Se implementaron los mismos métodos que en el punto 1, además se configuró $B = 10000$, es decir el proceso de muestreo fue repetido 10000 veces. En la tabla 2 se presentan los resultados de la media emírica y el intérvalo de confianza al 90 %. Principalmente, se observa que los tres métodos obtuvieron una media empírica bastante similar. Por su parte el método de sesgo corregido proporciona un intervalo más cerrado en comparación a los otros métodos. Además se conoce que la distribución del coeficiente de pearson

tiene una distribución Hipergeométrica Gaussiana en [1] *Analytic posteriors for Pearson's correlation coefficient* por Fisher. Por lo tanto se puede mencionar que posiblemente el intervalo de confianza mas adecuado podría ser el paramétrico asumiendo una distribución normal.

	Media empírica	α_1	α_2
No paramétrico - Percentil	0.716	0.591	0.879
No paramétrico - Bias Corrected Accelerated	0.717	0.615	0.847
Percentil - Paramétrico (Normal)	0.713	0.557	0.930

Cuadro 2: Cálculo de los intervalos considerando dos métodos no paramétricos bootstrap (percentil y por sesgo corregido) y un método paramétrico asumiendo una distribución Gamma.





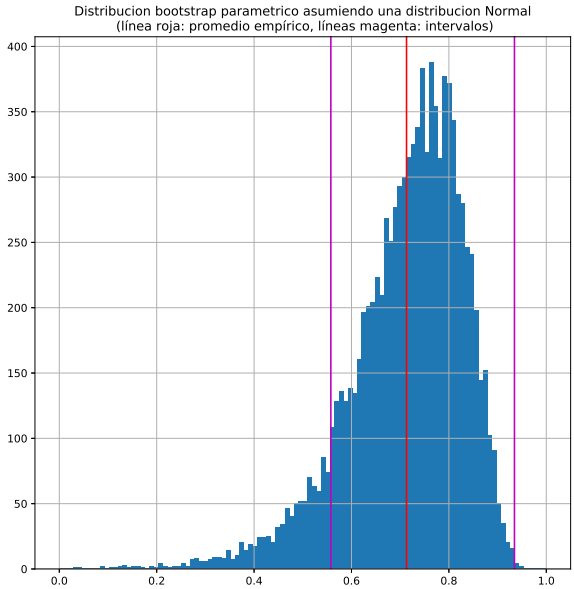


Figura 2: Distribución de las muestras generadas con cada uno de los tres métodos, la líneas verticales indican: rojo - media teórica, amarillo: media empírica, magenta: intervalos al 90 %.

Referencias

[1] A. Ly, M. Marsman, E.-J. Wagenmakers, Analytic posteriors for pearson's correlation coefficient, Statistica Neerlandica 72 (2018) 4–13.