Universidad Nacional de Colombia

Facultad de ciencias
Departamento de estadística

CASO DE ESTUDIO I

Estadística Bayesiana
Prof. Juan Camilo Sosa Martínez

Autores

Ana Sofía Bello Dueñas
Estadística
abellod@unal.edu.co

Germán Camilo Vásquez Herrera
Estadística
gvasquez@unal.edu.co

Bogotá, 14 de Septiembre de 2023

Tiempo de reparación

Verizon es la principal compañía telefónica local (ILEC, incumbent local exchange carrier) para una gran área del este de Estados Unidos. Como tal, es responsable de brindar servicio de reparación a los clientes de otras empresas telefónicas de la competencia (CLEC, competing local exchange carrier) en esta región.

Verizon está sujeto a multas si los tiempos de reparación (el tiempo que lleva solucionar un problema) para los clientes de alguna CLEC son sustancialmente peores que los de los clientes de Verizon. El conjunto de datos Verizon.csv (Chihara and Hesterberg,2019, Sec. 1.3) contiene una muestra de los tiempos de reparación de $n_1 = 1664$ clientes de Verizon (ILEC) y $n_2 = 23$ clientes de la competencia (CLEC). De acuerdo con los datos, los tiempos medios de reparación son $\bar{x_1} = 8.41$ y $\bar{x_2} = 16.51$ horas para ILEC y CLEC, respectivamente.

El objetivo de este caso es determinar si la diferencia entre los tiempos promedio de reparación es lo suficientemente grande para declararse como significativa, y por tanto, ser tenida en cuenta como evidencia para llevar a cabo una intervensión y multar a Verizon.

Modelo

La distribución Exponencial es popular para modelar tiempos dado que este modelo permite producir distribuciones con diferentes tasas de decaimiento y variedades de sesgo (para más información acerca de este modelo probabilístico, ver por ejemplo Sosa et al. 2014, Sec. 10.4).

Así, considere modelos Exponenciales independientes de la forma

$$y_{k,i}|\lambda_k \overset{iid}{\sim} Exp(\lambda_k) \Longleftrightarrow p(y_{k,i}|\lambda_k) = \frac{1}{\lambda_k} exp\left(-\frac{y_{k,i}}{\lambda_k}\right)$$

para $i = 1, ..., n_k$ y $k = 1, 2$ (1: ILEC, 2: CLEC), donde $y_{k,i}$ es el tiempo de reparación (en horas) del individuo i en el grupo k, n_k es el tamaño de la muestra del grupo k, y finalmente, $\boldsymbol{y_k} = y_{k,1}, ..., y_{k,n_k}$ es el vector columna de observaciones correspondiente.

PARTE 1: Análisis Bayesiano

1. Ajuste los modelos Gamma-Inversa-Exponencial con $a_k = 3$ y $b_k = 17$ en cada grupo. A partir de las distribuciones posteriores obtenga la distribución posterior de η. Reporte la media, el coeficiente de variación y un intervalo de credibilidad al 95 % para η. Presente los resultados visual y tabularmente. Interprete los resultados obtenidos (máximo 100 palabras).

Solución. De acuerdo con los resultados preliminares, la distribución posterior de λ_k es

$$p(\lambda_k|\boldsymbol{y}) \ \propto \ \lambda^{n_k+a_k+1} \ exp \ \left(-\frac{s_k+b}{\lambda}\right)$$

$$\lambda_k|\boldsymbol{y} \overset{iid}{\sim} GI(a_k+n_k, \ s_k+b_k)$$

Con $s_k = \sum_{i=1}^{n_k} y_{i,k}$, para $k = 1, \ 2$.

Se hicieron dos simulaciones Monte Carlo de tamaño 10000, una para λ_1 (parámetro del grupo ILEC) y la otra para λ_2 (parámetro del grupo CLEC), esto con el objetivo de aproximar la distribución posterior de $\eta = \lambda_1 - \lambda_2$, con este procedimiento se obtuvieron los siguientes resultados

Media	**CV**	**2.5 %**	**97.5 %**
-7.461	0.432	-14.962	-2.361

Tabla 1: Resultados de las estimaciones sobre η.

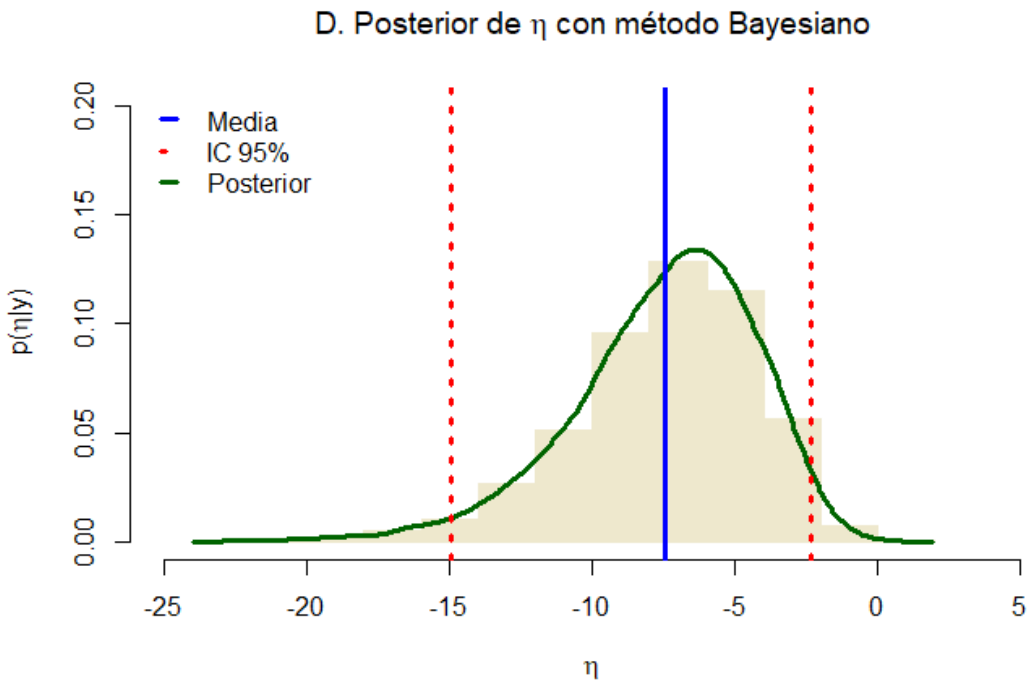

Figura 1: Distribución posterior de η.

Siguiendo estos resultados, dado que el coeficiente de variación es 43.2 %, que indica una variabilidad alta entre la diferencia promedio de tiempos de reparación, y que la media resultó ser -7.461, es decir, los tiempos promedios de reparación para CLEC son mayores que los de

ILEC, podemos concluir que hay evidencia para realizar una intervención y multar a la compañía telefónica Verizon.

2. Lleve a cabo un análisis de sensitividad. Para ello, considere los siguientes estados de información externos al conjunto de datos:

   - Distribución previa 1: $a_k = 3$ y $b_k = 17$, para $k = 1, 2$
   - Distribución previa 2: $a_k = 2$ y $b_k = 8.5$, para $k = 1, 2$
   - Distribución previa 3: $a_k = 3$ y $b_1 = 16.8$ y $b_2 = 33$, para $k = 1, 2$
   - Distribución previa 4: $a_k = 2$ y $b_1 = 8.4$ y $b_2 = 16.5$, para $k = 1, 2$

   En cada caso calcule la media y el coeficiente de variación a priori, y repetir el númeral anterior. Presente los resultados visual y tabularmente. Interprete los resultados obtenidos (máximo 100 palabras).

   Solución.

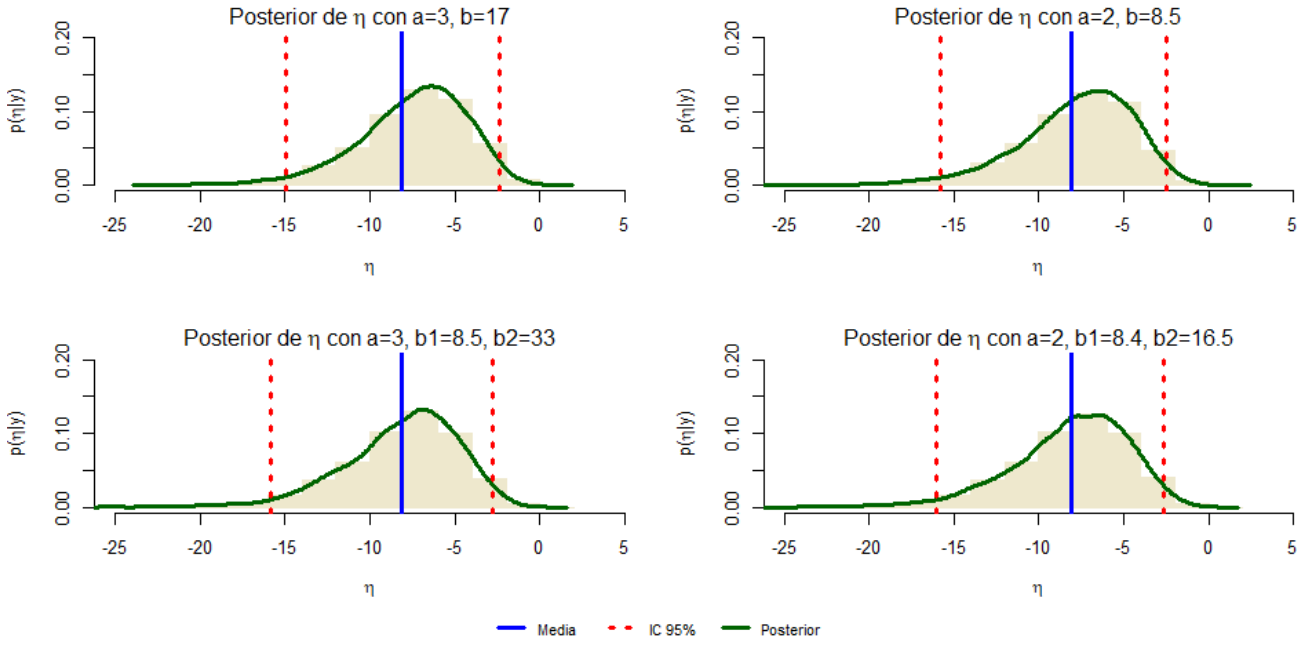

Figura 2: Resultados gráficos del análisis de sensitividad.

Tal y como se puede observar en la tabla 2 y en las gráficas de la figura 2, las estimaciones para la media, el intervalo de credibilidad y el coeficiente de variación no cambian mucho, lo que indica que el modelo no se deja guiar tanto por la información previa y en cambio deja que sean los datos observados quienes aporten más información.

La razón por la que hay "**NA**" en el coeficiente de variación a priori de las observaciones del grupo ILEC (*CV a priori 1*) y en el coeficiente de variación a priori de los datos del grupo CLEC (*CV a priori 2*) en las previas 1 y 2 es porque en ambos casos la información previa tiene un

	Previa 1	Previa 2	Previa 3	Previa 4
Media	-7.461	-7.779	-8.122	-8.1
CV	0.432	0.438	0.417	0.426
2.5 %	-14.962	-15.822	-15.809	-16.005
97.5 %	-2.361	-2.475	-2.788	-2.684
Media a priori 1	8.5	8.5	8.4	8.4
CV a priori 1	1	NA	1	NA
Media a priori 2	8.5	8.5	16.5	16.5
CV a priori 2	1	NA	1	NA

Tabla 2: Resultados del análisis de sensitividad.

hiperparámetro $a_k = 2$, de modo que la varianza no está definida. Por otro lado, en el escenario de la previa 1 y la previa 2, el coeficiente de variación es de 100 % lo que indica que esta información previa es poco informativa.

3. En cada población, evalúe la bondad de ajuste del modelo propuesto utilizando la distribución previa 1, utilizando como estadísticos de prueba la media y la desviación estándar. Presente los resultados visual y tabularmente. Interprete los resultados obtenidos (máximo 100 palabras).

   Nota: calcule los valores p predictivos posteriores y en cada grupo realice la visualización de las distribuciones predictivas de los estadísticos de prueba de manera conjunta (dispersograma con histogramas marginales).

   Solución. A continuación se presentan los valores de la media y la desviación estándar de las primeras 6 muestras (de un total de 10000) de cada grupo. Dichas muestras se obtuvieron a partir de un vector de 10000 λ's generados aleatoriamente de la distribución posterior Gamma inversa con hiperparámetros $a_k + n_k$ y $s_k + b_k$, para $k = 1, 2$.

	Media ILEC	SD ILEC	Media CLEC	SD CLEC
$\boldsymbol{\theta}^{(1)}$	8.391	8.132	13.598	14.234
$\boldsymbol{\theta}^{(2)}$	8.159	8.273	19.819	15.245
$\boldsymbol{\theta}^{(3)}$	8.429	8.409	14.45	13.502
$\boldsymbol{\theta}^{(4)}$	8.062	8.262	16.081	19.277
$\boldsymbol{\theta}^{(5)}$	8.13	7.72	20.49	21.253
$\boldsymbol{\theta}^{(6)}$	8.591	8.419	14.299	13.085

Tabla 3: Estadísticas de las primeras seis muestras de cada grupo.

En la siguiente tabla 4 se muestran las estadísticas observadas en la muestra original de ILEC y CLEC.

En la tabla 5 se muestran los valores de ***ppp*** obtenidos para la media y la desviación estándar de ILEC y CLEC.

Media ILEC	SD ILEC	Media CLEC	SD CLEC
8.412	14.690	16.509	19.503

Tabla 4: Estadísticas observadas en las muestras originales

	Media ILEC	SD ILEC	Media CLEC	SD CLEC
ppp	0.505	1	0.557	0.778

Tabla 5: Valores ppp para la media y la desviación estándar de cada grupo

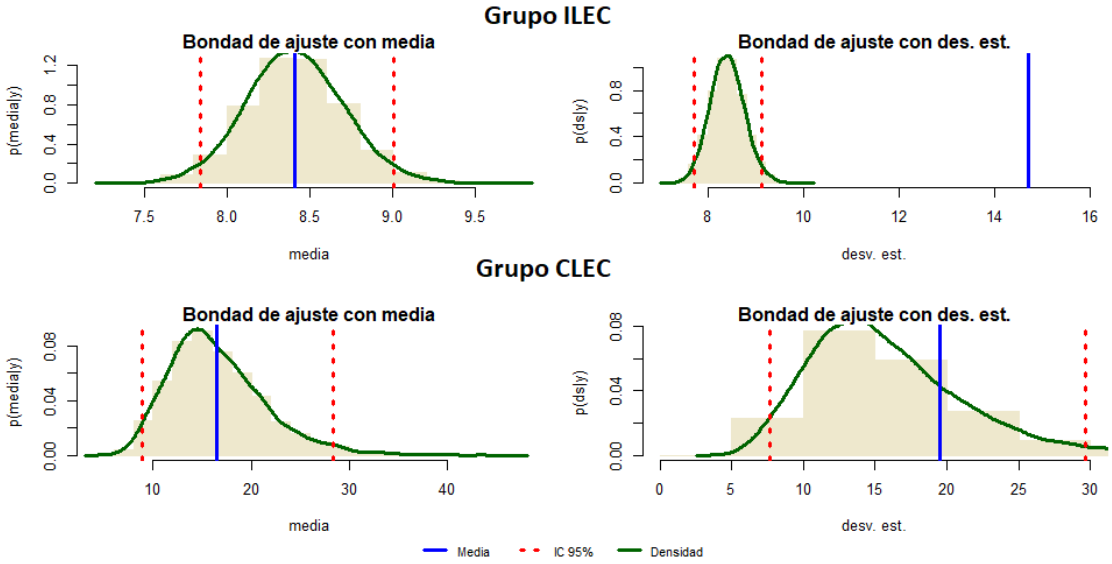

Figura 3: Bondad de ajuste con media y varianza.

Según lo que podemos observar en las gráficas de Figura 3 y los resultados de los valores ***ppp***, para la compañía ILEC, el modelo presenta un comportamiento no muy bueno ya que a pesar de que la media se encuentra dentro del intervalo de credibilidad del 95 %, la desviación estándar está muy lejos del límite superior del intervalo de credibilidad y se obtuvo un $ppp = 1$, es decir, que el modelo está subestimando la variabilidad de los tiempos de ILEC. Esto puede verse también en Figura 4.

Por otro lado, para el grupo CLEC, el modelo se comporta muy bien porque tanto la media como la desviación estándar se encuentran dentro de los intervalos de credibilidad del 95 %. Este resultado puede visualizarse en el dispersograma de Figura 5.

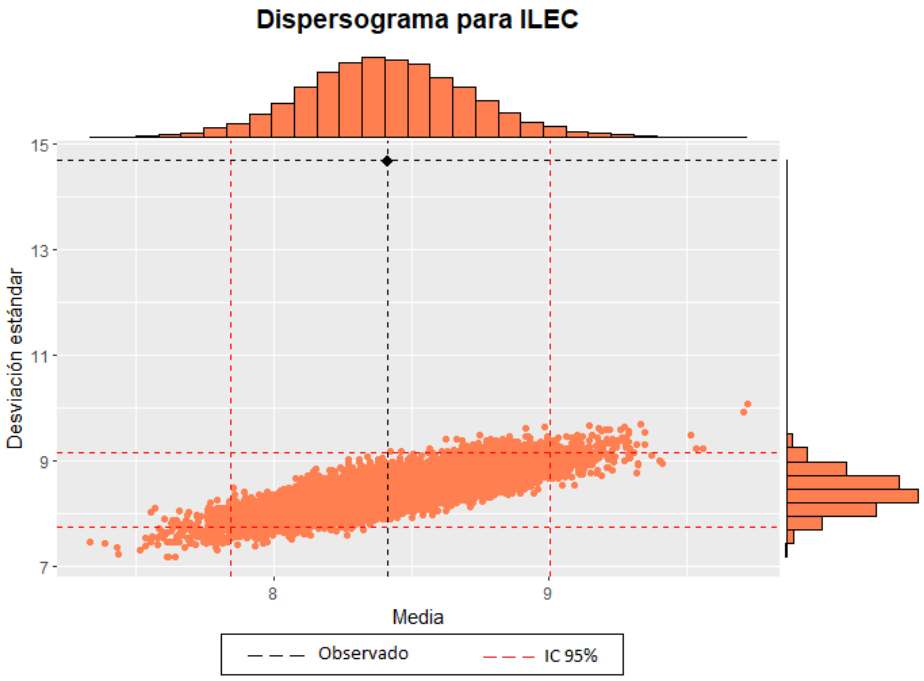

Figura 4: Dispersograma grupo ILEC

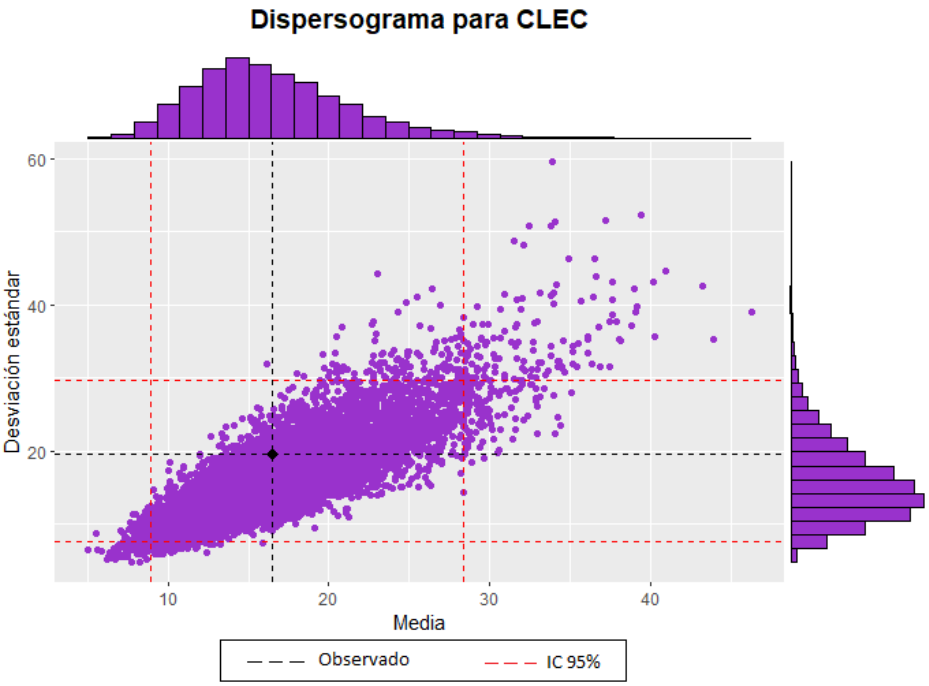

Figura 5: Dispersograma grupo CLEC

PARTE 2: Análisis frecuentista

Repita el númeral 1. de la PARTE 1 usando la Normalidad asintótica del MLE, *Bootstrap paramétrico* y *Bootstrap no paramétrico*. Presente los resultados visual y tabularmente. Interprete los resultados obtenidos (máximo 100 palabras).

Nota: asintóticamente se tiene que $\hat{\lambda}_{MLE} \approx N(\lambda, \hat{I}^{-1})$, donde $\hat{\lambda}_{MLE}$ es el MLE de λ y $\hat{I}$ es la información observada de Fisher.

Nota: cuando utilice *Bootstrap*, use una cantidad de remuestras adecuada y el método de los percentiles para calcular los intervalos de confianza.

Solución.

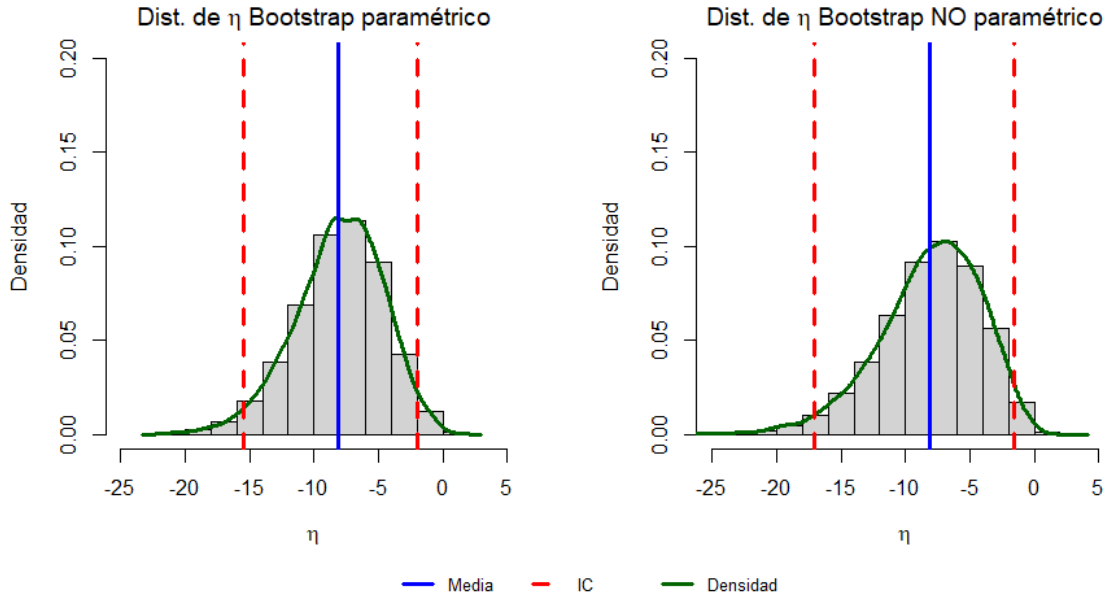

Figura 6: Distribución de *eta* con Bootstrap paramétrico y no paramétrico

	Media	**CV**	**2.5 %**	**97.5 %**
N. Asintótico	-8.097	0.426	-14.857	-1.338
Bootstrap P.	-8.049	0.429	-15.497	-1.949
Bootstrap no P.	-8.075	0.494	-17.055	-1.615

Tabla 6: Estimaciones sobre η con análisis frecuentista

Se observa que tanto en las estimaciones de la media como del coeficiente de variación son muy similares en los tres métodos frecuentistas, se ve tal vez una leve reducción del intervalo de confianza en los métodos que utilizan Bootstrap a comparación del análisis con normalidad asintótica.

En comparación con el método bayesiano, en realidad, el método frecuentista muestra que hay una

diferencia un poco más grande en los tiempos promedios de reparación que la diferencia estimada por métodos bayesianos. Los coeficientes de variación obtenidos con ambos métodos son muy similares.

El análisis frecuentista también nos lleva a concluir que el grupo ILEC debe ser multado ya que la estimación de la diferencia de los tiempos promedios (*aproximadamente* -8) indica que tarda más tiempo promedio en hacer las reparaciones para CLEC.

Cabe mencionar que la razón por la que no está el resultado gráfico para el análisis con normalidad asintótica es porque no se conoce el verdadero valor del parámetro η ($\hat{\eta}_{MLE} \sim N(\eta,\ I_1^{-1} + I_2^{-1})$). Además, para hacer las estimaciones sobre η para este método, se usa $\hat{\eta}_{MLE} = \hat{\lambda}_{1_{MLE}} - \hat{\lambda}_{2_{MLE}}$, que es MLE por la invarianza del estimador de máxima verosimilitud.

PARTE 3: Simulación

Simule 100000 muestras aleatorias de poblaciones Exponenciales bajo los siguientes escenarios:

- Escenario 1: $n_1 = 10$, $n_2 = 10$, $\lambda_1 = \bar{y}_1$ y $\lambda_2 = \bar{y}_2$.
- Escenario 2: $n_1 = 20$, $n_2 = 20$, $\lambda_1 = \bar{y}_1$ y $\lambda_2 = \bar{y}_2$.
- Escenario 3: $n_1 = 50$, $n_2 = 50$, $\lambda_1 = \bar{y}_1$ y $\lambda_2 = \bar{y}_2$.
- Escenario 4: $n_1 = 100$, $n_2 = 100$, $\lambda_1 = \bar{y}_1$ y $\lambda_2 = \bar{y}_2$.

donde $\bar{y}_k = \frac{1}{n_k}\sum_{i=1}^{n_k} y_{k,i}$ es la media muestral observada del grupo k. Observe que el valor verdadero de η en cada caso es $\eta = \lambda_1 - \lambda_2 = \bar{y}_1 - \bar{y}_2$.

Usando cada muestra, ajuste el modelo de manera tanto Bayesiana (usando la distribución previa 1) como frecuentista (usando la Normalidad asintótica, *Bootstrap* paramétrico, *Bootstrap* no paramétrico), y en cada caso calcule la proporción de veces que el intervalo de credibilidad/confianza al 95 % contiene el valor verdadero de η. Reporte los resultados tabularmente. Interprete los resultados obtenidos (máximo 100 palabras).

Solución.

	Bayesiana	**N. Asintótica**	***Bootstrap* P.**	***Bootstrap* no P.**
Escenario 1	0.93	0.945	0.938	0.892
Escenario 2	0.939	0.949	0.944	0.919
Escenario 3	0.943	0.948	0.946	0.935
Escenario 4	0.946	0.949	0.947	0.941

Tabla 7: Proporción de veces que el verdadero valor del parámetro está en el IC del 95 %.

Según los resultados obtenidos en Tabla 7, a medida que aumenta el tamaño de muestra, la proporción de veces que el verdadero valor del parámetro cae dentro del intervalo de credibilidad (o de confianza en métodos frecuentistas) aumenta. En general, las mayores proporciones se presentan al usar la normalidad asintótica, seguido del método frecuentista usando Bootstrap paramétrico, luego el método Bayesiano y, finalmente, las proporciones menores se obtienen con Bootstrap no paramétrico.

Referencias

Chihara, L. M. and Hesterberg, T. C. (2019). *Mathematical statistics with resampling and R.* John Wiley & Sons.

Sosa, J. C., Ospina, L. E., and Berdugo, E. P. (2014). *Estadística descriptiva y probabilidades.* U. Externado de Colombia.